Springer Series in Optical Sciences 160

Peter Seitz
Albert J.P. Theuwissen *Editors*

Single-Photon Imaging

Springer

Springer Series in
OPTICAL SCIENCES 160

founded by H.K.V. Lotsch

Springer Series in
OPTICAL SCIENCES

The Springer Series in Optical Sciences, under the leadership of Editor-in-Chief *William T. Rhodes*, Georgia Institute of Technology, USA, provides an expanding selection of research monographs in all major areas of optics: lasers and quantum optics, ultrafast phenomena, optical spectroscopy techniques, optoelectronics, quantum information, information optics, applied laser technology, industrial applications, and other topics of contemporary interest.
With this broad coverage of topics, the series is of use to all research scientists and engineers who need up-to-date reference books.
The editors encourage prospective authors to correspond with them in advance of submitting a manuscript. Submission of manuscripts should be made to the Editor-in-Chief or one of the Editors. See also www.springer.com/series/624

For further volumes:
http://www.springer.com/series/624

Peter Seitz
Albert J.P. Theuwissen
Editors

Single-Photon Imaging

With 250 Figures

Springer

Editors

Peter Seitz
CSEM SA
Bahnhofstraße 1, 7302 Landquart, Switzerland
E-mail: peter.seitz@csem.ch

Albert J.P. Theuwissen
Delft University of Technology
Mekelweg 4, 2628 CD Delft, The Netherlands
E-mail: a.j.p.theuwissen@tudelft.nl

Springer Series in Optical Sciences ISSN 0342-4111 e-ISSN 1556-1534
ISBN 978-3-642-18442-0 e-ISBN 978-3-642-18443-7
DOI 10.1007/978-3-642-18443-7
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011934863

Cover design: eStudio Calamar Steinen

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Dark clouds hung over physics toward the end of the nineteenth century, when physicists began to appreciate that their comprehension of the nature of light was critically incomplete. The classical description of light as an electromagnetic wave satisfying the beautiful equations of Maxwell obviously failed to explain significant optical effects: How is radiation absorbed by matter? How can light with such strange, narrow spectra be emitted by gases or solid materials? How can the nature of blackbody radiation be explained? In a radical step, Albert Einstein and Max Planck provided the key to this impasse by introducing the revolutionary notion that the energy states of the electromagnetic field are not continuous but rather quantized – they successfully imagined the photon. So, finally the clouds parted, opening vistas into the strange world of quantum physics.

A natural consequence of the concept of a photon is the existence of an ultimate detection limit of electromagnetic radiation. Once you can sense each individual photon (possibly gaining also information about its energy and polarization state), you know all about incident radiation that can be known. For this reason, the holy grail of photosensing is the spatially resolved detection of light with this ultimate precision, single photon imaging. The aim of this book is to provide a comprehensive and systematic overview of all relevant approaches currently in use to realize practical single photon imagers. In all of these devices, three major tasks have to be accomplished: (1) incoming photons must enter the detector, where they are converted into electronic charge; (2) this photogenerated charge must be collected and possibly amplified at the same time; and (3) the collected charge must be detected with suitable electronic circuitry.

In all these steps, one has to fight thermally generated noise: The photogeneration process competes with dark noise charge generation in the conversion layer; in the photocharge collection and amplification process, signal charges must be handled while avoiding the detrimental effects of thermally generated charge carriers; finally, the first stages of any electronic charge detection circuitry suffers from thermally generated Johnson noise in the channel of transistors or in resistors. Depending on the boundary conditions of a photodetection problem – for example, the

photosensitive area, the response time, the mean detection rate, the exposure time, the frame rate, the spectral distribution of the radiation, the operating temperature, and the power consumption – a different technological approach will come out as an optimum. For this reason, the present book provides a theoretical and practical framework, where researchers and practitioners will find in condensed form all relevant information to resolve their particular single photon imaging solution.

In Chap. 1, relevant fundamental concepts for treating noise phenomena in optoelectronics are summarized, and a rigorous definition of the precise meaning of "single photon imaging" is given. State-of-the-art semiconductor technology especially suited for ultra-low-noise image sensing is presented in Chap. 2. The use of photocathodes in vacuum for single photon imaging is treated in several chapters: in Chap. 3, the charge multiplication processes is implemented with avalanche photodiode (APD) arrays; in Chap. 4, the photoelectrons are accelerated to a high voltage, and their bombardment of semiconductor imagers causes a large number of secondary electrons being created in the image sensor; in Chap. 5, a suitable geometry of several electrodes, each multiplying the incident electron packets by a factor, provides for photocharge multiplication of up to factor of one million. It is also possible to exploit the avalanche effect in semiconductors, without having to use vacuum devices. In Chap. 6, the avalanche effect is used in so-called electron-multiplying charge-coupled devices (EMCCDs), while Chap. 7 describes CMOS compatible semiconductor image sensors for single-photon avalanche detection (SPADs). In synchronous applications, where one samples the images at regular times while accumulating photogenerated charges between samples, it is possible to realize single photon CMOS imagers through systematic bandpass filtering, exploiting the parallelisms possible in CMOS imagers; this approach to single photon imaging is described in detail in Chap. 8, and the complementary Chap. 9 treats suitable architectures for the implementation of such single photon CMOS imagers. If one is not constrained to use standard CMOS processes, an interesting class of structures, called double-gate transistors and charge modulating devices (CMDs), make it possible to sense individual electrons with very high conversion gains of several $100\,\mu V$ per electron, as described in Chap. 10. The case of high-energy photons (UV and X-ray radiation) arriving at arbitrary times is treated in Chap. 11, showing the way to efficient, energy-sensitive X-ray single photon imagers implemented with standard CMOS processes. Each of the last three chapters describes an important practical application in which single photon imaging is a key capability: in Chap. 12, optical time-of-flight range imaging is covered, with which complete 3D images of a scene can be acquired with millimeter resolution in real time. Astronomical and aerospace applications in which single photon imagers are essential are presented in Chap. 13. Finally, Chap. 14 describes a highly relevant application of gated ultra-low-noise imagers in the life sciences, namely very sensitive and highly specific pharmaceutical and medical diagnostics through time-resolved fluorescence imaging.

No panacea exists yet for the practical and economical solution of the many single photon imaging problems in the world, ranging from fundamental scientific research to the availability of cell phone cameras with which brilliant pictures can

be taken also under extreme low-light conditions. Finding a solution still requires skillfully elaborating a good technological compromise. If the authors of this book have achieved their goal of providing a useful and powerful tool to many engineers and researchers in the wide field of image science, then our ambition has been fulfilled and the efforts of all involved colleagues have been worthwhile.

We would like to express our sincere thanks to the authors of the various chapters for their kindness and willingness to contribute to this book, for their hard work required in actually carrying through with the promise, and for their determination to meet all the deadlines revising and updating their chapters, with the goal to provide the most valuable and up-to-date contributions.

Landquart *Peter Seitz*
Delft *Albert Theuwissen*
March 2011

Contents

Contributors

Verle Aebi Intevac Photonics, Inc., 3560 Bassett Street, Santa Clara, CA 95054, USA, vaebi@intevac.com

Hiroaki Aihara Department of Physics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
and
Institute for the Physics and Mathematics of the Universe (IPMU), The University of Tokyo, 5-1-5 Kashiwa-no-ha, Kashiwa-shi, Chiba 277-8568, Japan, aihara@phys.s.u-tokyo.ac.jp

Edoardo Charbon Technical University Delft, Mekelweg 4, 2628 CD Delft, The Netherlands, e.charbon@tudelft.nl

Kenneth Costello Intevac Photonics, Inc., 3560 Bassett Street, Santa Clara, CA 95054, USA, kcostello@intevac.com

Christof Fattinger F. Hoffmann-La Roche Ltd., Pharmaceutical Research and Early Development, Discovery Technologies, Grenzacherstrasse 124, 4070 Basel, Switzerland, christof.fattinger@roche.com

Matthew W. Fishburn Technical University Delft, Mekelweg 4, 2628 CD Delft, The Netherlands, m.w.fishburn@tudelft.nl

Boyd Fowler Fairchild Imaging, 1801 McCarthy Blvd., Milpitas, CA 95035, USA, boyd.fowler@fcimg.com

Shoji Kawahito Research Institute of Electronics, Shizuoka University, 3-5-1, Johoku, Naka-ku, Hamamatsu 432-8011, Japan, kawahito@idl.rie.shizuoka.ac.jp

Christian Lotto Heliotis AG, D4 Platz, 6039 Root Längenbold, Switzerland
and
CSEM SA, Photonics Division, Technopark, CH-8005 Zurich, Switzerland, christian.lotto@a3.epfl.ch

Yoshiyuki Matsunaga Image Sensor Business Unit, Semiconductor Company, Panasonic Co., Ltd., 1 Kotari-yakemachi, Nagaokakyo City, Kyoto 617-8520, Japan, matsunaga.yoshiyuki001@jp.panasonic.com

Pierre Magnan ISAE, 10 Av. E. Belin, 31055 Toulouse Cedex, France, Pierre.Magnan@isae.fr

R. Daniel McGrath Aptina Imaging, San Jose, CA 95134, USA, dmcgrath@ieee.org

Francis Müller F. Hoffmann-La Roche Ltd., Pharmaceutical Research and Early Development, Discovery Technologies, Grenzacherstrasse 124, 4070 Basel, Switzerland, francis.mueller@roche.com

Gert Nützel PHOTONIS Technologies S.A., Axis Business Park E, 18 Avenue de, Pythagore, 33700 Mérignac, France, g.nutzel@photonis.com

Mark Stanford Robbins e2v technologies Ltd, 106 Waterhouse Lane, Chelmsford, Essex CM1 2QU, UK, mark.robbins@physics.org

Peter Seitz CSEM SA, Nanomedicine Division, Bahnhofstrasse 1, 7302 Landquart, Switzerland
and
EPFL, STI-IMT-NE, Institute of Microengineering, Rue A.-L. Breguet 2, 2000 Neuchâtel, Switzerland, peter.seitz@csem.ch, peter.seitz@epfl.ch

Andrea Simoni Fondazione Bruno Kessler, Via Sommarive 18, 38123, Trento, Italy, simoni@fbk.eu

David Stoppa Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy, stoppa@fbk.eu

Albert J.P. Theuwissen Harvest Imaging and Delft University of Technology, Kleine Schoolstraat 9, 3960 Bree, Belgium, albert@harvestimaging.com, a.j.p.theuwissen@tudelft.nl

Chapter 1
Fundamentals of Noise in Optoelectronics

Peter Seitz

Abstract Electromagnetic radiation can be described as a stream of individual photons. In solid-state detectors (e.g., photocathodes or semiconductor photosensors), each photon of sufficient energy creates one or several mobile charge carriers which can be subsequently detected with sensitive electronic circuits, possibly after charge packet multiplication employing the avalanche effect. Two types of noise limit the resolution with which individual photons can be detected: (1) The number of detectable photons or photoelectrons shows a statistical variation, which is often well-described as a Poisson distribution. (2) Photogeneration of mobile charge carriers competes with thermal generation, and thermal noise is compromising the generation of photocharges ("dark current") as well as the performance of electronic charge detection circuits ("Johnson noise" and "random telegraph signal noise"). It is shown that the laws of physics and the performance of today's semiconductor fabrication processes allow the detection of individual photons and photocharges in image sensors at room temperature and at video rate.

1.1 Introduction

At the end of the nineteenth century, it was realized that the classical description of light as an electromagnetic wave satisfying Maxwell's equation fails to properly explain important optical phenomena. In particular, the interaction with matter, such as the emission or absorption of light, requires an improved theoretical framework. The breakthrough came with Einstein's hypothesis that light consists of quanta of energy [1], the so-called photons. As a consequence, the measurement of properties of electromagnetic radiation has an ultimate limit, imposed by the description of light as a stream of individual photons. It is only natural, therefore, that scientists and engineers alike want to perform their measurements of light to this ultimate precision, the single photon.

P. Seitz and A.J.P. Theuwissen (eds.), *Single-Photon Imaging*,
Springer Series in Optical Sciences 160, DOI 10.1007/978-3-642-18443-7_1,

As we will work out in this contribution, the main obstruction to achieving this goal is, ironically, the presence of electromagnetic radiation. This is, of course, the ubiquitous blackbody radiation surrounding us whenever the temperature of our environment is not identical to zero. The goal of the present work is to understand how the coupling between this omnipresent "temperature bath" and matter influences our measurements of light, elucidating under which circumstances single-photon sensing of light is possible in practice.

In a first part, basic properties of quantized systems are recalled, in particular essential properties of the Poisson distribution. In the second part, the noise properties of most common light sources are investigated, revealing that under most practical conditions, the photons of these light sources have essentially a Poisson distribution. In the third part, the energy band model of solid state matter is employed to explain the principles and fundamental limitations of photosensing with semiconductors. In the fourth part, noise sources in relevant electronic components and circuits are studied, to determine the ultimate, temperature-dependent limits of the electronic detection of charge. Finally, all the material is brought together for a concise summary of the physical and technological limits of the detection of light, explaining under which circumstances single-photon imaging is possible.

1.2 Quantization of Electromagnetic Radiation, Electrical Charge, and Energy States in Bound Systems

The photoelectric effect – the emission of electrons from solid matter as a result of the absorption of energy from electromagnetic radiation – seems to be compelling evidence for the existence of photons. Actually, this is not the case: Although it is true that the atoms in solid matter are absorbing energy from a light beam in quantized packets, this can also be understood in a semiclassical picture in which light is described as a classical electromagnetic wave and only the atoms are treated as quantized objects [2]. Careful analysis along similar lines can also show that the individual pulses detected with "single-photon counting devices" are not conclusive evidence for the existence of photons.

It is an astonishing fact that there are only relatively few optical phenomena that cannot be explained with a semiclassical theory [2]. Of course, only the full quantum optical approach in the framework of quantum electrodynamics (QED) is completely consistent, both with itself and with the complete body of experimental evidence [3]. QED mathematically describes the interaction between light and matter by specifying how electrically charged particles interact through the exchange of photons.

Despite this surprising fact that the photon concept is not really required to understand photodetectors, the simplicity, intuitiveness and basic correctness of the corpuscular photon picture makes it so attractive that it is adopted by most researchers in the field.

As a consequence, we consider light to be a stream of photons, individual particles of zero rest mass traveling at the speed of light $c = 2.9979 \times 10^8\,\mathrm{m/s}$ in vacuum and at retarded speeds in matter. The wavelike nature of the photon is reflected in the fact that a frequency ν and a corresponding wavelength λ can be associated with it, so that the energy E of a photon is given by:

$$E = h\nu = \frac{hc}{\lambda} \tag{1.1}$$

with Planck's constant $h = 6.6262 \times 10^{-34}\,\mathrm{Js}$.

Each atom of matter consists of a system of several electrons bound to the protons in a nucleus by the attractive Coulomb force. The electron is a subatomic elementary particle carrying a negative charge of $-q$, with the unit charge $q = 1.6022 \times 10^{-19}\,\mathrm{C}$. In effect, the electrical charge of any particle is always found to be a multiple of this unit charge, making electrical charge also a quantized physical property.

Finally, according to quantum mechanics, the bound system of an atom does not have stable states with arbitrary energies. Rather, only a discrete set of stable states, each with its proper energy level, is allowed [4]. It is concluded that also the energy states of a bound system are quantized.

1.3 Basic Properties of the Poisson Distribution

In any kind of counting problem, the Poisson distribution arises as an almost inevitable consequence of statistical independence [5]. For this reason, the Poisson distribution is clearly the most important probability law in photon or electron counting problems. The typical question asked in such a problem is the following: What is the probability $p_N(k)$ that a number $k \geq 0$ of events is observed during a fixed observation period T if these events occur with a known average rate N/T and, independent of the time since the last event? The quantity N describes, therefore, the average number of events observed during the time T. The sought probability distribution $p_N(k)$ is the Poissonian:

$$p_N(k) = \frac{N^k \mathrm{e}^{-N}}{k!}. \tag{1.2}$$

In Fig. 1.1 the Poisson distributions for the three small expectation values $N = 1$, $N = 4$, and $N = 9$ are depicted.

The Poisson distribution has a few surprising properties, for example the fact that if the average number of observed events is $N = 1$, the probability that no events are observed ($k = 0$) is the same as the probability that exactly one event occurs ($k = 1$). And because $p_1(1) = 0.368$, we observe exactly one event only in about a third of the cases.

Another important property of the Poisson distribution is its relation to Bernoulli trials. A Bernoulli trial is an experiment whose outcome is random and can only

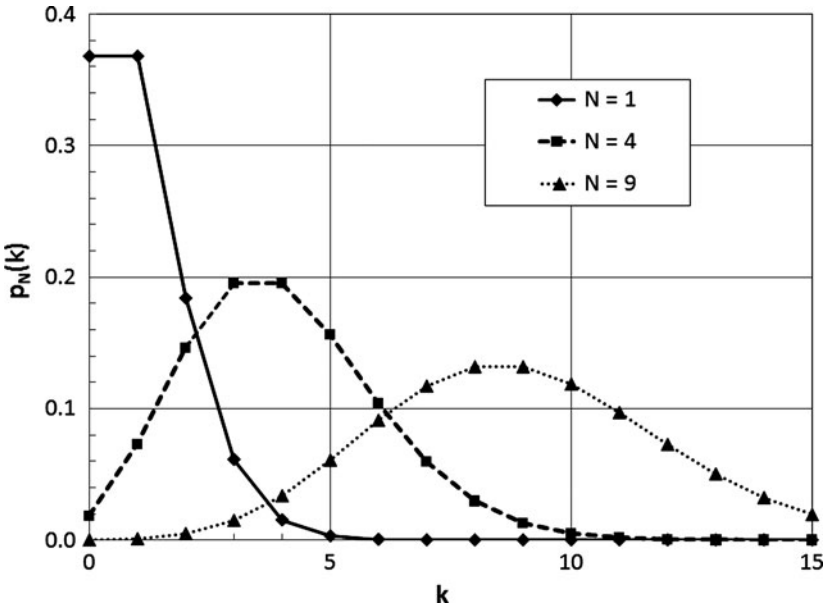

Fig. 1.1 Poisson distribution $p_N(k)$ for three different cases of small expectation value N. These are discrete distributions, and the connecting lines serve only as visual aids

be of two possible outcomes, either "success" or "failure." If consecutive Bernoulli trials, a sequence which is called "binomial selection," are statistically independent of each other with a fixed success probability η, the so-called *binomial selection theorem* holds [5]:

> Binomial selection of a Poisson process yields a Poisson process, and the mean M of the output of the selection process is the mean N of the input times the success probability η, $M = \eta \cdot N$.

The complete physical process of the detection of photons can be described mathematically as a cascade of binomial selection processes. If we can assure that the input to this whole chain of events is a Poisson process then we are certain that the output of the whole detection process is also a Poisson process. In Sect. 1.4 the different types of interaction processes between electromagnetic radiation and matter are investigated, leading to the insight that all these processes are, indeed, cascades of binomial selection processes. In Sect. 1.5, the photon emission properties of the most common light sources are examined, and under many practical conditions, the emitted photon streams shows, indeed, a Poisson distribution.

As a consequence, the probabilistic description of the photodetection process becomes very simple in most practical cases: Wherever in an experiment one samples and inspects photon streams or photogenerated charges, they are Poisson distributed!

1.4 Interaction of Radiation and Matter

The actual detection of electromagnetic radiation is a complicated process, involving many types of interactions between light and matter, such as reflection, refraction, diffraction, scattering, absorption, and electronic conversion. Each of these individual processes, however, is by itself a binomial selection process or a sequence of binomial selection processes.

As illustrated in Fig. 1.2, the most common interactions between light and matter include the following processes:

- Absorption, either by a neutral-density or a color filter, Fig. 1.2a, involves binomial selection of incident photons into transmitted and absorbed photons. If the incident photons are Poisson distributed then both the transmitted and the absorbed photons are also Poisson distributed.
- Beam-splitting or reflection, Fig. 1.2b involves binomial selection of incident photons into transmitted and reflected beams. If the incident photons are Poisson distributed then both the transmitted and the reflected photons are also Poisson distributed.
- Diffraction, Fig. 1.2c, consists of a cascade of binomial selection processes. If the incident photons are Poisson distributed then the photons in each diffracted beam are also Poisson distributed.
- Scattering, Fig. 1.2d, also consists of a cascade of binomial selection processes. If the incident photons are Poisson distributed then the photons observed under any scattering direction are also Poisson distributed.
- Generation of photo-charge pairs in a semiconductor, Fig. 1.2e, involves the interaction of incident photons with the atoms of the solid. If the energy of an incident photon is sufficiently high, the photon can create with a certain probability an electron–hole pair. If the incident photons are Poisson distributed then both the photo-generated charges and the transmitted photons are Poisson distributed.

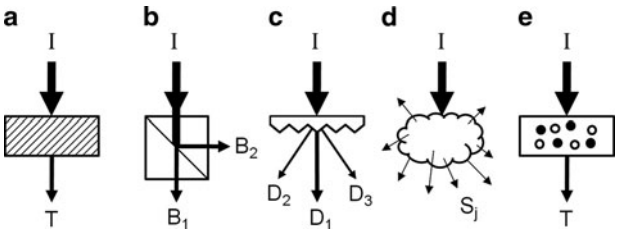

Fig. 1.2 Schematic illustration of the various types of interaction processes occurring between incident electromagnetic radiation I and matter. (**a**) A neutral-density or color filter produces the transmitted reduced intensity T. (**b**) A beam-splitter sends incident photons arbitrarily into one of two beams B_1 and B_2. The intensities B_1 and B_2 can be different. (**c**) A diffraction filter produces several diffracted beams D_j, whose intensities are typically not the same. (**d**) A scattering object produces scattered light S, which can usually be observed in all directions. (**e**) The detection of incident photons with a semiconductor creates charge pairs (photogenerated electrons and holes)

1.5 Noise Properties of Light Sources

In the quantum picture, an electromagnetic wave corresponds to a stream of photons. Depending on the detailed nature of the photon generation process, the statistical distribution of these photons will vary. These statistical properties are summarized later for the four most common light sources employed in practice, coherent light sources such as single-mode lasers, thermal (blackbody radiator) light sources, partially coherent light sources such as discharge lamps and light emitting diodes (LEDs).

1.5.1 Coherent Light (Single-Mode Lasers)

In classical physics, light is described as an electromagnetic wave satisfying Maxwell's equations. In this model, any type of wave in free space can be represented by a linear superposition of plane waves, so-called modes, of the form [6]:

$$U(x, y, z, t) = \mathrm{e}^{-\mathrm{i}\omega t}\ \mathrm{e}^{\mathrm{i}(ux+vy+wz)} \tag{1.3}$$

with the wave vector $\mathbf{k} = (u,\ v,\ w)$, the spatial coordinate $\mathbf{r} = (x,\ y,\ z)$, time t, and the angular frequency $\omega = |\mathbf{k}|c$. Such a monochromatic beam of light with constant power is called coherent light. The light emitted by a single-mode laser operating well above threshold is a good physical approximation to such a perfectly coherent light source.

Assuming that the emission of photons produced in a coherent light source is the effect of individual, independent emission processes, it can be shown that the emitted number of photons n is a random variable which has a Poisson distribution [2]. As a consequence, the variance s_N of this Poisson distribution is equal to the mean number N of the emitted photons, corresponding to the expectation value $N = < n >$ of n

$$s_N = N. \tag{1.4}$$

1.5.2 Thermal (Incandescent) Light Sources

Light emitted from atoms, molecules, and solids, under condition of thermal equilibrium and in the absence of other external energy sources, is known as thermal light or blackbody radiation. The temperature-dependent spectral energy density of a thermal light source is determined by Planck's blackbody radiation law [6]. If only a single mode of such a blackbody radiation field is considered, the resulting probability distribution of the photon number is not a Poisson but a Bose–Einstein distribution [6], with a variance s_N given by:

$$s_N = N + N^2. \tag{1.5}$$

In practice, it is very rare that only one mode of a blackbody radiation source is of importance, and in most experimental cases the statistical properties of multimode thermal light must be considered. Assuming that a thermal light source contains M independent thermal modes of similar frequencies, it can be shown [7] that the photon number variance s_N of this multimode thermal light source is given by:

$$s_N = N + \frac{N^2}{M}. \tag{1.6}$$

The number M of thermal modes present in an actual blackbody radiation source is usually very large. When light from a blackbody radiation cavity with volume V is filtered with a narrow-band filter of bandwidth $\Delta\lambda$ around a central wavelength λ, the resulting narrow-band thermal radiation field contains a number M of thermal modes of

$$M = V\frac{8\pi\,\Delta\lambda}{\lambda^4} \tag{1.7}$$

which follows directly from the mode density in a three-dimensional resonator [6]. As an example, consider an incandescent light source with a cavity volume (filament part) of $V = 1\,\text{mm}^3$, filtered through a narrow-bandwidth filter with $\Delta\lambda = 1\,\text{nm}$ around the central wavelength $\lambda = 600\,\text{nm}$. According to (1.7), this thermal radiation field contains the large number of $M = 1.9 \times 10^8$ thermal modes.

Consequentially, in most practical cases thermal light sources contain such a large number of modes that the statistics of the emitted photon numbers are effectively described by a Poisson distribution. The multimode variance given in (1.6) is then practically equal to the Poisson case (1.4).

As a concluding note, it should be mentioned that even photons from a single thermal mode may approach a Poisson distribution if the detection times are long enough: If the time interval chosen for the individual observations of photon numbers is much larger than the coherence time of the thermal field (the "memory" scale of the field), the photon number statistics are again approaching a Poisson distribution, as has been calculated in detail in [8].

1.5.3 Partially Coherent Light (Discharge Lamps)

The light from a single spectral line of a discharge lamp has classical intensity fluctuations on a time scale determined by the radiation's coherence time τ [2]. This type of light source is, therefore, partially coherent. The intensity fluctuations will give rise to greater fluctuations in photon number than for a source with constant power such as a perfectly coherent source. For this reason, partially coherent light cannot be Poisson distributed. In [7] a semiclassical treatment of the counting statistics of a fluctuating field is given, and it is shown that the variance s_N of the photon number may be expressed as:

$$s_N = N + < \Delta W(t)^2 >, \tag{1.8}$$

where $W(t)$ denotes the averaged photon count rate during the detection time interval t. If the detection time t is much longer than the coherence time τ then the intensity fluctuations on time-scales comparable to τ will become inconsequential, and the intensity maybe taken as effectively constant. In this case again, the variance of the time-averaged partially coherent light is reduced to the Poisson case (1.4). Typical coherence times of discharge lamps are of the order of 1–100 ns.

1.5.4 Light Emitting Diodes

Semiconductor materials can emit light as a result of electron–hole recombination. A convenient way of achieving this is to inject electrons and holes into the space charge region of a pn-junction by applying a forward bias to it [6]. The resulting recombination radiation is known as injection electroluminescence, and the pn-junction light source is termed LED.

Because of the high conversion efficiency of modern LEDs, the photon statistics of these solid state light sources are strongly influenced by the charge carrier statistics of the injected current. If a stable electronic current source with low internal resistance is employed, the charge carriers in the current I are Poisson distributed, and the variance s_I of the current's so-called shot-noise is given by:

$$s_I = 2\,q\,I\,B, \tag{1.9}$$

where B denotes the bandwidth of the measurement circuit. As a consequence, the photons emitted by the LED are also Poisson distributed, and the variance is given again by (1.4).

It is not difficult, though, to devise a source of current with a sub-Poisson charge carrier distribution. Consider a stable voltage source in series with a resistor R. Because of the thermal coupling of the resistor to its environment with temperature T, the voltage across the resistor's terminals fluctuates statistically with the so-called Johnson noise. The variance s_V of this Johnson noise is given by:

$$s_V = 4kT\,R\,B \tag{1.10}$$

with Boltzmann's constant $k = 1.3807 \times 10^{-23}\,\mathrm{J/K}$. This voltage noise across a resistor corresponds to a current noise through it, with a variance s_I expressed in terms of the current I and the voltage V across the resistor

$$s_I = \frac{4kT\,I\,B}{V}. \tag{1.11}$$

Comparison of (1.9) and (1.11) reveals that as soon as the voltage across the resistor exceeds a value of $2kT/q$, corresponding to about 52 mV at room temperature, the statistics of the charge carriers in the resistor become sub-Poissonian. Therefore,

instead of driving a high-efficiency LED with a stable current source for the generation of Poisson distributed electroluminescence light, it can be operated with a stable voltage source through a series resistor, resulting in sub-Poissonian statistics of the emitted light [9].

1.6 The Meaning of "Single-Photon Imaging"

In the previous sections we have seen how the statistics of an incident stream of photons are influenced by the various optical and optoelectronic components in the beam path. At the end of this chain, the photons are interacting in an image sensor, where they create electronic charges or charge-pairs; as detailed earlier, these are often Poisson distributed. The final task is electronically to detect and convert these charges into voltage signals, which can then be further processed. Unfortunately, this electronic charge detection is also a statistical process, often with zero mean, as will be discussed in more detail in Sect. 1.9. If the variance of the photons interacting with the image sensor is denoted by s_N and the variance of the electronic photocharge detection process is s_D, the total variance s of the image sensor signal is given by:

$$s = s_N + s_D. \tag{1.12}$$

The consequences of this noisy electronic photocharge detection and process are graphically illustrated in Fig. 1.3 for an image sensor under low-intensity illumination (mean number of interacting photons < 10).

Round symbols in Fig. 1.3 indicate the positions where light sources are imaged onto the image sensor. During the observation time, the light sources may emit a (small) number of photons which then interact with the image sensor (full circles), or no photons interact with the image sensor (open circles), either because

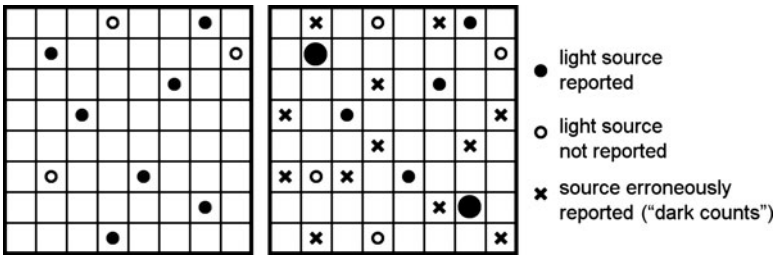

Fig. 1.3 Graphical illustration of the consequences of the noisy electronic charge conversion process taking place in the image sensor and its associated electronics. *Round symbols*: position of light sources. *Full circles*: interaction of photons with the sensor. *Open circles*: no photons detectable. *Crosses*: photons are erroneously reported due to the noise of the electronic charge detection process. *Left*: ideal detection (no electronic noise). *Right*: Real detection (broadened distribution and dark counts)

none are emitted or none are detected by the image sensor. The left-hand picture shows the distribution of photons interacting with the image sensor. The right-hand picture illustrates the effects of the subsequent electronic photocharge detection and conversion process.

The first effect is the broadening of the probability distribution, as seen already in (1.12). As a consequence, the reported number of photons becomes more imprecise, such as interacting photons that may now not be detected any more (previously full circles on the left become open circles on the right); or a number of interacting photons which are reported too high (larger full circles on the right).

The second effect, however, is much more disconcerting in practice: As the electronic noise is acting on all pixels, photons maybe reported even in locations where no light sources are imaged onto the image sensor, so-called "dark counts." This is indicated with crosses on the right-hand picture. If the low-intensity light sources are only sparsely distributed over the image sensor, it can become impossible to identify them if the electronic noise of the charge detection process is too large. Based on this assertion, the following practical definition of the general topic of this book, "single-photon imaging" is proposed:

> Single-photon imaging is the detection of two-dimensional patterns of low-intensity light, i.e. mean photon numbers in the pixels of less than 10, where the electronic photocharge detection process contributes such little noise that the probability of erroneously reporting a photon where there is none is appreciably smaller than the probability of having at least one photon in a pixel.

To determine for each particular case how much variance can be tolerated in the electronic photocharge detection process, it is assumed that the electronic charge detection noise is normally distributed with zero mean and standard deviation σ_D

$$p_D(x) = \frac{1}{\sqrt{2\pi}\,\sigma_D}\ \mathrm{e}^{-\frac{x^2}{2\sigma_D^2}}. \tag{1.13}$$

It can be argued that the electronic charge detection process consists of several independent stochastic contributions, and because of the central limit theorem the compound distribution is approaching the Gaussian (1.13) [5]. Note that the output of the image sensor is usually a voltage, but the coordinate x in (1.13) is scaled such that $x = 1$ corresponds to the detection of exactly one photocharge. The probability p of (erroneously) reporting one or more photocharges in a pixel is then given by:

$$p = \int_{0.5}^{\infty} p_D(x)\,\mathrm{d}x = \frac{1}{2}\mathrm{erfc}\left(\frac{1}{\sqrt{8}\,\sigma_D}\right), \tag{1.14}$$

where erfc is the complementary error function, $\mathrm{erfc}(x) = 1 - \mathrm{erf}(x)$ [10]. For a handy estimation of which σ_D is required to reach a given probability p, Winitzki's approximation of the error function erf (x) is employed, which can easily be inverted algebraically [11]. The following estimate for σ_D results

$$\sigma_D \cong \frac{1}{\sqrt{8}} \sqrt{\frac{2a}{ab - \frac{4}{\pi} + \sqrt{\left(ab - \frac{4}{\pi}\right)^2 + 4ab}}}, \tag{1.15}$$

where $a = 0.147$ and b is the following function of the probability p

$$b = -\ln\left(1 - (1 - 2p)^2\right). \tag{1.16}$$

As an example, if it should be assured that in less than every tenth pixel, $p < 0.1$, a photon is erroneously reported then $\sigma_D < 0.39$ must be achieved. And for $p < 0.01$, $\sigma_D < 0.215$ is necessary.

1.7 Energy Band Model of Solid State Matter

Quantum theory explains the discrete energy levels of the stable states of bound electron-core systems. As a consequence, isolated atoms, ions, and molecules show only discrete energy levels and the interaction with light can only occur if the energy of interacting photons corresponds to a transition between these discrete energy states [4]. In solids, however, the distances between the cores are so small that their bound electrons interact with each other, and the problem must be treated as a many-body system instead of a single electron-core system.

As a consequence, the energy levels of solids are not single lines any more but they are broadened. Since the electrons close to the core are well shielded from the fields of the neighboring cores, the lower energy levels of a solid are not broadened, and they rather correspond to those of the isolated atoms. In contrast, the energies of the higher-lying discrete atomic levels split into closely spaced discrete levels and effectively form bands [6]. This is illustrated in Fig. 1.4a.

The highest partially occupied energy band of a solid is called conduction band, and the energy band just below it is called the valence band. The difference E_g between the highest energy of the valence band E_V and the lowest energy of the conduction band E_C is called the bandgap energy:

$$E_\mathrm{g} = E_\mathrm{C} - E_\mathrm{V}. \tag{1.17}$$

If the conduction band of a solid is partially filled at zero temperature, $T = 0$, the solid is a metal; it conducts current well at any temperature. If, however, the conduction band is completely empty at zero temperature then the solid is a semiconductor; it cannot conduct current at zero temperature because there are no energy states available in the valence band which would correspond to electrons moving freely in the solid. Insulators are just special cases of semiconductors whose bandgap energy E_g is larger than about 5 eV. Actually, solids appearing as “insulators” at moderate temperature are effectively semiconductors at elevated

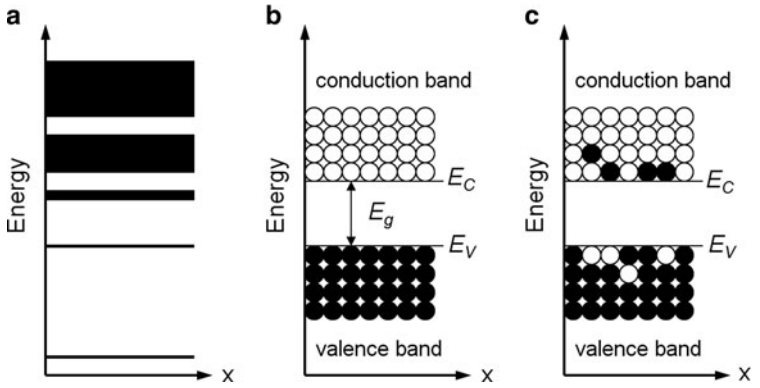

Fig. 1.4 Schematic illustration of the energy distribution of bound electron-core states in solids. (**a**) Broadening of the discrete energy levels of an isolated electron-core system into bands for solid-state material. (**b**) Occupancy of valence and conduction bands in a semiconductor at $T = 0$. (**c**) Occupancy of valence and conduction bands at $T > 0$. *Open circle*: vacant state at this energy level; *filled circle*: occupied state

temperatures. A good example is diamond, with $E_g = 5.45\,\text{eV}$ [13], which is considered to be an insulator at room temperature but diamond is increasingly made use of for the fabrication of semiconductor circuits operating at a high temperatures of 400 °C and above [14].

1.8 Detection of Electromagnetic Radiation with Semiconductors

The schematic illustration in Fig. 1.4 indicates that the excitation of an electron from the valence band into the conduction band, i.e., the generation of an electron–hole pair, requires an interaction with an energy of at least the bandgap energy E_g. If this interaction energy is provided by an incident photon, which is absorbed in the event, then the created electron–hole pair contributes to the measurement signal. However, if the interaction energy is provided by thermal energy from the environment (a phonon) then the created electron–hole pair contributes to the noise impairing the measurement signal.

1.8.1 Quantum Efficiency and Band Structure

The efficient generation of a photon out of an electron–hole pair requires a semiconductor with a direct bandgap, i.e., the emission process does not require the concurrent presence of an additional phonon, to satisfy the simultaneous conservation of energy and momentum. Fortunately, it is not necessary to fulfill

this requirement for the detection process, because the basic events can occur sequentially: First, an incident photon of sufficient energy excites an electron from the valence band into the conduction band by a so-called vertical transition, i.e., no momentum transfer occurs [8]. It is only afterward that the excited electron moves to the bottom of the conductance band through the fast release of one or several phonons (thermalization), and the created hole moves to the top of the valence band through similar thermalization.

As a consequence, direct- and indirect-bandgap semiconductors are equally efficient in the conversion of incident photons of sufficient energy into electron–hole pairs: Quantum efficiencies of close to 100% can be realized in practice in an intermediate energy range, i.e., almost all incident photons in this energy range will create electron–hole pairs. If the energy of the incident electron is lower than the bandgap energy then the semiconductor is essentially transparent to the incident electromagnetic radiation; if the energy of the incident electron is much larger than the bandgap energy then the incident light is already absorbed in the covering layers of a device, and no or few photons can reach the bulk of the semiconductor and interact there. As a consequence, the photodetection quantum efficiency of a semiconductor drops off toward the infrared and the ultraviolet part of the spectrum.

1.8.2 Thermal Equilibrium and Nonequilibrium Carrier Concentrations

As mentioned earlier, the energy required to create an electron–hole pair can also be provided thermally by the absorption of phonons of sufficient energy from the environment. The higher the temperature, the more phonons are available for the thermal creation of electron–hole pairs. In thermal equilibrium, a pure (undoped) semiconductor has an equal concentration n_i of mobile electrons and holes, the so-called intrinsic carrier concentration, given by:

$$n_\mathrm{i} = N_0 \, \mathrm{e}^{-\frac{E_\mathrm{g}}{2kT}} \propto T^{\frac{3}{2}} \, \mathrm{e}^{-\frac{E_\mathrm{g}}{2kT}}, \tag{1.18}$$

where the factor N_0 depends only on $T^{3/2}$ and on the effective masses of the electrons in the conduction band and of the holes in the valence band. These values are a function of the exact details of the form of conduction and valence band, and therefore, they are highly specific for each semiconductor. As an example, the intrinsic carrier concentration for silicon, with a bandgap energy of $E_\mathrm{g} = 1.12\,\mathrm{eV}$, is $n_\mathrm{i} = 1.45 \times 10^{10}\,\mathrm{cm}^{-3}$ at $T = 300\,\mathrm{K}$ [12].

It is very important to note that the foregoing is only true for thermal equilibrium. As the presence of mobile charge carriers in a doped or intrinsic semiconductor in thermal equilibrium makes it impossible to distinguish between this "background" charge and the photogenerated charge, high-sensitivity photosensor devices make all use of fully depleted volumes of semiconductors. In these so-called space charge regions all mobile charge carriers have been removed by an electric field. This is

usually accomplished either with a diode structure (a pn-junction) or an MOS (metal oxide semiconductor) capacitance [12]. In both cases, the background charge is reduced to zero in the depletion region, and the energy distribution of the states looks like Fig. 1.4b, corresponding to the case of zero temperature. As a consequence, the energy distribution of states in the depletion region is *not* in thermal equilibrium, and care must be taken when applying models and equations that only hold true for thermal equilibrium.

1.8.3 Dark Current

As the space charge region of a photodetector is swept clean of all mobile charge carriers, any charge present there must either be a part of the photogenerated signal or it must be thermally produced and belongs to the noise. The thermally produced charge carriers move under the influence of the electric field in the space charge region, resulting in the so-called dark current. This dark current has two components, as illustrated in Fig. 1.5 in the case of an MOS structure.

If the charge carriers are generated within the space charge region, this part of the dark current is called generation current. If charge carriers are thermally generated in the bulk of the semiconductor, it is possible that they move by diffusion to the edge of the space charge region, where they are swept across by the electric field and contribute to the dark current. This part of the dark current is called diffusion current. As indicated in Fig. 1.5, only charge carriers thermally produced at a distance of less than L, the diffusion length, from the space charge region can contribute to the diffusion current.

The generation dark current density j_{gen} in a space charge region of width w, which depends on the voltage V_R with which the photodetector is biased, is calculated according to

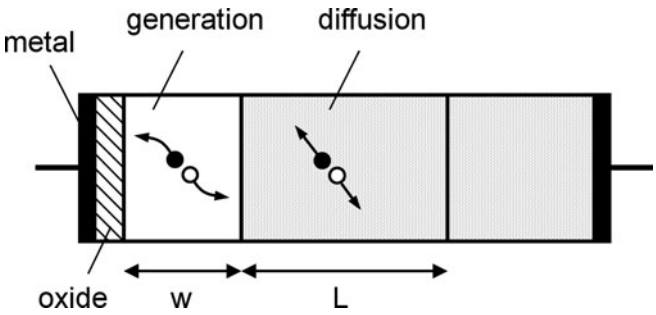

Fig. 1.5 Schematic illustration of the effects contributing to the dark current in an MOS (metal-oxide-semiconductor) structure: Generation current (space charge region width w) and diffusion current (diffusion length L) in the semiconductor

$$j_{\text{gen}} = \frac{q\, n_i}{2\,\tau}\, w(V_R) \tag{1.19}$$

with the generation lifetime τ of electron–hole pairs [12]. In direct bandgap semiconductors, the lifetime τ is of the order of tens of nanoseconds, and in good-quality indirect bandgap semiconductors, it can be as high as tens of milliseconds [6].

The diffusion dark current density j_{diff} in the bulk of a semiconductor with doping concentration N is given by:

$$j_{\text{diff}} = q\, n_i^2 \frac{D}{NL} \tag{1.20}$$

with the diffusivity D, which is related to the diffusion length L by:

$$L = \sqrt{D\tau}. \tag{1.21}$$

In silicon, it is often the generation current that dominates at room temperature, and the diffusion current becomes important only at elevated temperatures [12]. If only the generation current must be considered, the temperature dependence of the dark current density can be calculated by combining (1.18) and (1.19):

$$j_{\text{dark}} \propto T^{\frac{3}{2}}\, \mathrm{e}^{-\frac{E_{\text{g}}}{2kT}}\, w(V_R). \tag{1.22}$$

Modern semiconductor technology makes it possible to reach amazingly low dark current densities, even at room temperature. The lowest value reported to date is $j_{\text{dark}} = 0.15\,\text{pA cm}^{-2}$ at $T = 300\,\text{K}$ in a CCD image sensor [15]. This corresponds to less than one thermally generated electron per second in a pixel of size $10 \times 10\,\mu\text{m}^2$.

1.8.4 Avalanche Effect and Excess Noise Factor

Until now we have always assumed that the number of charge carriers rests constant once created. There is a physical effect, however, which allows to multiply charge packets by arbitrary factors m. This so-called avalanche effect is illustrated in Fig. 1.6. An electron (1) is accelerated in an electric field until it has acquired sufficient energy E_k larger than the bandgap E_{g}, so that this energy is released in the creation of a secondary electron (2) and a hole (2*). Both electrons are accelerated now until they have acquired sufficient energy for the creation of two pairs of electrons (3) and holes (3*). Of course, the same holds true for the holes, which themselves can create additional electron–hole pairs. This is repeated and leads to a cascade of charge pair creation, effectively multiplying the number N of input charges by a factor m, resulting in a larger charge packet $M = m \times N$. The avalanche multiplication factor m depends strongly on the actual value of the electric field [12].

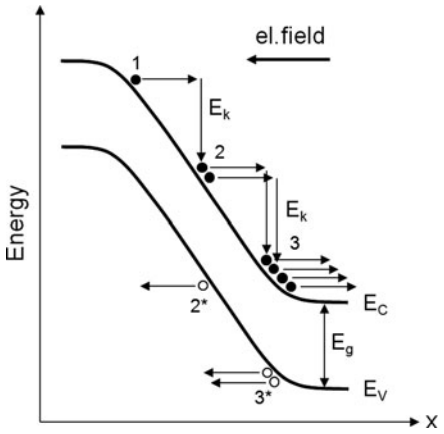

Fig. 1.6 Schematic illustration of the avalanche effect: An electron (1) is accelerated in an electric field until it gains sufficient energy $E_k > E_g$ to create an electron–hole pair (2) and (2*). This is repeated and leads to a cascade of charge pair creation, effectively multiplying the input charge by a field-dependent factor

As the avalanche effect is by itself a statistical process, it is not surprising that avalanche multiplication changes the statistics of the charge packets. If the variance of the original charge packet is denoted by s_N then the variance s_M of the multiplied charge packet is given by:

$$s_\mathrm{M} = m^2 \, s_N \, F, \tag{1.23}$$

where F denotes the excess noise factor of the avalanche multiplication [5]. Note that even for completely noise-free multiplication ($F = 1$), the multiplication of a Poisson input cannot produce a Poisson output: The mean increases linearly with the multiplication factor, while, according to (1.23), the variance increases with the square of this factor.

1.9 Electronic Detection of Charge

The last step in the photosensing chain consists of the precise electronic detection of photogenerated charge packets. Obviously, the electronic charge detection circuits should add only insignificant amounts of noise, so that also very small charge packets, down to a single unit charge, can be reliably detected. Although this sounds like a straightforward, simple task, in reality this is the biggest obstacle today for solid-state single-photon imaging, apart from the technology-dependent dark current discussed in Sect. 1.8.3.

The fundamental reason for the fact that electronic circuits and components are noisy is the interaction of the free electrons with their thermal environment. In the conducting materials, which are used for the construction of electronic elements, the motion of the electrons has a random component, because of their nonzero kinetic energy. According to the law of equipartition [16] the average kinetic energy E_k of each free electron is given by:

$$E_\mathrm{k} = \frac{3}{2} k \, T. \tag{1.24}$$

The random microscopic motion of the free electrons, also called diffusion, has macroscopic electrical effects in the electronic components usually employed for the realization of electronic circuits.

1.9.1 Basic Components of Electronics and their Noise Properties

The three basic components of which most electronic circuits are composed are the resistance, the capacitance, and the transistor, as illustrated in Fig. 1.7. As most photosensors are fabricated today with technology related to Complementary Metal Oxide Semiconductor (CMOS) processes, we will assume for the following that the transistors are of the metal-oxide field-effect (MOS-FET) type [12]. Also, the effects of inductance will be ignored, because they are usually negligible in circuits employed for low-noise electronic charge detection.

In a resistor, illustrated in Fig. 1.7a, diffusion of free electrons results in a random current contribution with zero mean through the device, which in turn causes a fluctuating voltage with zero mean across the resistor's terminals. The variance s_V of the noise voltage and the variance s_I of the noise current, the so-called Johnson noise, as already introduced in (1.10), are given by:

$$s_V = 4kT\,BR; \quad s_I = \frac{4\,kT\,B}{R}, \tag{1.25}$$

where B denotes the bandwidth of the measuring circuit [17]. Note that, the Johnson noise described by (1.25) is only an approximation. The resistor noise is only white (frequency-independent) if the measurement frequencies are below kT/h, corresponding to about 6 THz at room temperature. The quantum mechanically correct spectral noise density distribution shows a drop-off at high frequencies, which is required for a finite total energy contained in the noise [17].

In an ideal capacitance, schematically illustrated in Fig. 1.7b, every "stored" electron on one electrode is compensated by a positive mirror charge on the other electrode, so that charge neutrality is observed. Although, it is not possible to localize such a charge pair on a capacitance, the total number of charge pairs is constant, and, as a consequence, no current noise is created in an ideal capacitance.

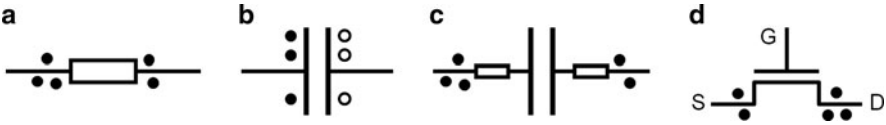

Fig. 1.7 Schematic illustration of the three basic components of electronic circuits, realized with CMOS processes. (**a**) Resistance, (**b**) Ideal capacitance, (**c**) Capacitance with resistive leads, (**d**) Tranistor of the MOS-FET type

However, the leads of a capacitance are not perfect conductors but they are rather resistors, as schematically illustrated in Fig. 1.7c. As described earlier, these resistors are a source of Johnson noise, and this noise spectrum is filtered by the low-pass filter represented by the RC circuit. As the effective bandwidth B of an RC filter is of the order of $1/(R \times C)$, it is immediately concluded from (1.25) that the voltage noise variance s_{Vcap} across a capacitance is proportional to kT/C, and therefore independent of the actual resistance value R. Detailed calculation (integration of the spectral noise density with the appropriate filter function of the RC low-pass filter) yields the following voltage noise variance, also called kTC noise:

$$s_{\mathrm{Vcap}} = \frac{kT}{C}. \tag{1.26}$$

The third key component in an electronic circuit is the transistor. As the predominant semiconductor technology today for the implementation of image sensors is of the CMOS type, the employed transistors are field-effect transistors (FETs). Such a FET is schematically illustrated in Fig. 1.7d, and it consists of a gate electrode G whose voltage V_{G} is modulating the current I_{DS} flowing from drain D to source S. If the drain–source voltage V_{DS} is not too large then the FET is operating in its linear region as a programmable resistance, for which

$$I_{\mathrm{DS}} = 2K(V_{\mathrm{G}} - V_{\mathrm{T}})V_{\mathrm{DS}}. \tag{1.27}$$

With the geometry- and material-dependent device constant K and the threshold voltage V_{T}, see for example [18]. The value of the drain–source resistance R_{DS} is therefore given by:

$$R_{\mathrm{DS}} = \frac{1}{2K(V_{\mathrm{G}} - V_{\mathrm{T}})}. \tag{1.28}$$

As a consequence of the existence of this resistance, the drain–source region of a FET in its linear region exhibits Johnson noise with a variance $s_{I\mathrm{DS}}$ given by (1.25):

$$s_{I\mathrm{DS}} = \frac{4kTB}{R_{\mathrm{DS}}}. \tag{1.29}$$

This implies that the current across the FET's drain–source terminals is fluctuating statistically, even when the gate voltage is kept absolutely stable.

If the drain–source voltage is large, the FET is operating in its saturation region, and it effectively behaves as a programmable current source [18], where the drain–source current I_{DS} depends quadratically on the gate voltage V_{G}

$$I_{\mathrm{DS}} = K(V_{\mathrm{G}} - V_{\mathrm{T}})^2. \tag{1.30}$$

A simple model of the noise properties of a FET in saturation can be derived by assuming that the source–drain region of a transistor consists of two parts: A conducting channel with tapered shape extends from the source toward the

drain region, and it disappears (it "pinches off") close to the FET's drain. The short section between the pinch-off point and the drain consists of a completely depleted semiconductor region. The drain–source current is due to charge carriers that flow down the conducting channel and are injected at the pinch-off point into the depletion region near the drain.

As there are no free charge carriers in a depletion region, the noise properties of a FET in saturation are described by the Johnson noise generated in the conducting channel. Its resistance is approximated by the resistance R_{DS} of the FET in its linear region, as given by (1.28), and the current noise of a FET in its saturation region is consequentially that of a resistance with this value.

Although the physical origin of the current noise of a FET in saturation is the Johnson noise created in the conducting channel, an alternate useful model for the FET's noise performance assumes that the drain–source region is noise-free and that a hypothetical voltage noise source exists at the gate of the FET; this is the so-called "input-referred" voltage noise with variance s_{VG}. It can be estimated using the transconductance g_m, i.e., the differential change of the drain–source current with gate voltage at constant drain–source voltage

$$g_m = \left.\frac{\partial I_{\mathrm{DS}}}{\partial V_{\mathrm{G}}}\right|_{V_{\mathrm{DS}}=\mathrm{const}} = 2K(V_{\mathrm{G}} - V_{\mathrm{T}}) = \frac{1}{R_{\mathrm{DS}}}, \tag{1.31}$$

where use has been made of (1.28) and (1.30). The variance s_{VG} can now be calculated with (1.29) and (1.31):

$$s_{V\mathrm{G}} = \left(\frac{\partial I_{\mathrm{DS}}}{\partial V_{\mathrm{G}}}\right)^{-2} s_{I\mathrm{DS}} = \frac{s_{I\mathrm{DS}}}{g_m^2} = \frac{4kTB}{g_m}. \tag{1.32}$$

As mentioned, this is a simplified model for the input-referred channel noise in a FET. A more detailed calculation results in the well-known Klaassen-Prins equation for the input-referred channel noise variance in MOS transistors [19]:

$$s_{V\mathrm{G}} = \frac{4kT\,B\,\alpha}{g_m}. \tag{1.33}$$

With a parameter α that depends on the operation regime of the FET. In saturation $\alpha = 2/3$ [19].

As a numerical example, consider a MOS-FET in saturation with a transconductance g_m of $1/g_m = 1\,\mathrm{k\Omega}$, operating at $T = 300\,\mathrm{K}$ and measurement bandwidth $B = 20\,\mathrm{MHz}$. The input-referred root-mean-square voltage noise σ_{VG}, calculated as the square root of the variance s_{VG} in (1.33), is $\sigma_{\mathrm{VG}} = 14.86\,\mu\mathrm{V}$.

The statistical variation of the drain–source current in a MOS-FET's channel is only well described by Johnson noise for high frequencies, as illustrated in Fig. 1.8, where the current noise spectral density $S_I(f)$ is shown schematically as a function of temporal frequency f. At lower frequencies, $S_I(f)$ is proportional to $1/f^{\beta}$, where β is a parameter close to 1 in a wide frequency range [20]. For this reason, the

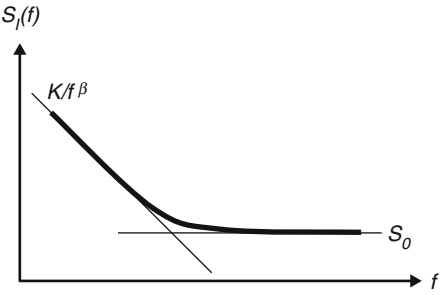

Fig. 1.8 Schematic illustration of the current noise spectral density $S_I(f)$ in a MOS-FET. At lower frequencies, $1/f$ noise dominates, while the noise spectrum is white at higher frequencies, due to its origin as Johnson noise in the FET's channel

low-frequency part of $S_I(f)$ is called "1/f noise." Typical transition frequencies from 1/f to Johnson noise are between 10 and a few 100 kHz in CMOS transistors.

It is recognized today that the physical origin of 1/f noise in MOS-FETs is the capture and emission of charge carriers from the transistor's channel by traps in the SiO_2 gate oxide [20]. This trapping–detrapping effect causes discrete modulations of the transistor's source–drain conductance called random telegraph signals (RTS) [21]. The superposition of even few RTS already leads to 1/f noise in ordinary MOS-FETs. It should be noted, though, that in deep submicron MOS-FETs, RTS becomes apparent because of the involvement of only a very small number of oxide traps.

1.9.2 Basic Circuits for Electronic Charge Detection

The most common approach to the sensitive electronic detection of charge is to place this charge on the gate of a transistor and to exploit the corresponding change in the transistor channel's electrical properties. In practice, widespread use is made of the source-follower configuration illustrated in Fig. 1.9a, because it combines high dynamic range, excellent sensitivity, and good linearity, while requiring only a small silicon floor space.

Most often, the measurement transistor is loaded with a current source, realized with a MOS-FET in saturation biased with an appropriate voltage V_L [21]. The output voltage V_{SF} then tracks the input voltage V_{in} with good fidelity according to

$$V_{SF} = V_{\text{in}} - V_{\text{T}}, \tag{1.34}$$

where V_{T} denotes the threshold voltage of the measurement MOS-FET.

An alternate circuit for the measurement of small amounts of charge is shown in Fig. 1.9b. This current-sink inverter, also called common-source amplifier, consists of an n-MOS load and a p-MOS measurement transistor, both operating in saturation [21]. It has the advantage over the much more popular source follower in Fig. 1.9a that the small-signal behavior shows voltage amplification, $\Delta V_{\text{inv}} = A \times \Delta V_{\text{in}}$, with an amplification factor A of the order of 10 [21]. This reduces the noise contribution of the downstream circuits, albeit at the cost of reduced dynamic range [22]. As in

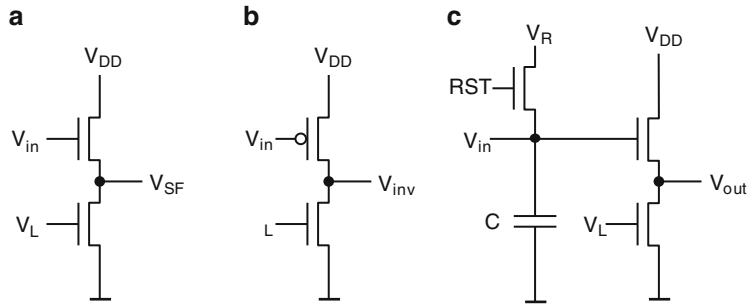

Fig. 1.9 Basic circuits for the electronic detection of charge. (**a**) Source follower with load transistor, (**b**) Current-sink inverter, (**c**) Source follower with reset transistor RST and effective input capacitance C

both cases, the measurement transistor is operating in saturation, the noise properties are comparable, and the white-noise part is adequately described by the Klaassen-Prins equation (1.33).

A more complete pixel circuit, based on a source follower for the detection of photogenerated charge Q on the gate of the measurement transistor, is illustrated in Fig. 1.9c, including a reset transistor (with reset signal RST) and the effective capacitance C at the gate of the measurement MOS-FET. The load transistor can be designed such that its noise contribution is negligible compared to the noise of the measurement transistor. The overall charge measurement noise in Fig. 1.9c is then dominated by the noise of the reset transistor, the so-called reset noise. Using correlated multiple sampling techniques, as described for example in [23], this reset noise is effectively eliminated, and due to the high-pass nature of this filtering, the 1/f part of the noise spectral density is also removed. The remaining noise is white and well described by the Klaassen-Prins equation (1.33). Therefore and as $V_{\text{in}} = C/Q$, the root-mean-square noise σ_Q of the charge measurement process is given by:

$$\sigma_Q = C\sqrt{\frac{4kT\,B\,\alpha}{g_m}}. \tag{1.35}$$

As a numerical example for the noise limitations of electronic charge measurement using this source-follower based detection approach, we assume an effective input capacitance C of 50 fF and the same figures as above ($1/g_m = 1\,\text{k}\Omega$, $T = 300\,\text{K}$, $B = 20\,\text{MHz}$). This results in a Johnson-noise limited charge measurement resolution of $\sigma_Q = 4.6$ electrons.

1.9.3 Conclusions for Single-Electron Charge Detection

As explained in Sect. 1.6, the notion of "single-photon electronic imaging" implies that a photocharge detection noise σ_Q of 0.2–0.4 is required, depending on the

actual application. However, as estimated earlier based on (1.35), good solid-state image sensors operating at video frame rates and room temperature achieve typical photocharge detection noise levels σ_Q of 5–10 electrons. It is concluded that an improvement of more than a factor of 10 in the charge detection noise is required. Over the past 20 years, many approaches have been proposed to attain this, and the most important ones are described in this book.

As the dominant noise source in the electronic detection of photocharge is the reset noise in (1.26), it is of foremost importance to reduce or remove it. Fortunately, the correlated multiple sampling (CMS) techniques described in [23] are capable of eliminating reset noise practically completely. Viable circuits for implementing these CMS techniques are presented in Chap. 9 of this book.

As the CMS techniques represent efficient low-pass filters, another important source of noise is eliminated at the same time, the 1/f noise illustrated in Fig. 1.8. It is essential, though, that the measurement band is well above the transition frequency from 1/f noise to Johnson noise, implying that the measurement frequency should not be lower than the typical MOS-FET transition frequencies of 10 to a few 100 kHz.

The remaining source of noise is, therefore, Johnson noise in the channel of the MOS-FET employed for the electronic detection of photocharge, which is at the heart of the charge noise formula (1.35).

The most effectual way to reduce the photodetection charge noise is to lower the effective detection capacitance C. An obvious possibility is to employ minimum-size transistors in deep-submicron semiconductor processes, and this has led to capacitances C of only a few fF. Even smaller values of C can be achieved with special transistor types such as double-gate MOS-FETs or charge-modulating devices (CMDs), as described in Chap. 10 of this book. Capacitances of < 1 fF and single-electron detection noise have been obtained in this way.

Another possibility is to reduce the operation temperature. This is not very effective, however, because the absolute temperature appears under the square root of (1.35). The real benefit of lowering the temperature is reduction of the dark current (1.22) and associated noise, as described in Chap. 2 of this book.

Although it is feasible, in principle, to increase the transconductance g_m in (1.35) by modifying the geometry of the detection MOS-FET, increasing g_m implies increasing the gate capacitance of the transistor, and this more than offsets all improvements achievable in this way.

Finally, reduction of the measurement bandwidth B is a practical and highly successful approach to single-electron photocharge detection, as detailed in Chap. 8 of this book. Although, it is true that reducing the output bandwidth of a conventional image sensor such as a CCD will necessarily decrease also the imager's frame-rate; it is possible to adapt the architecture of the image sensor to resolve this problem. One possibility is to provide the image sensor with a multitude of output channels, and each is operated at reduced bandwidth. A more effective way, which is particularly practical with CMOS image sensors, is to reduce the effective bandwidth of each column without compromising the overall frame rate. It is in this fashion that the subelectron readout results described in [22] or in Chap. 8 have been

obtained. Then again, one might think of intrapixel bandwidth reduction techniques but it appears today that this would not bring much improvement over the column-wise band-pass filtering approaches.

Because of the practical difficulties of achieving subelectron electronic detection noise, a successful alternative is to employ physical amplification mechanisms for the production of more than one charge per interacting photon. A good example of this approach is the employment of the avalanche effect described in Sect. 1.8.4. The advantage of this approach is that it can be integrated monolithically on the same chip and even in each pixel, in particular for the realization of monolithic SPAD imagers, single-photon avalanche photodetectors detailed in Chap. 7 of this book.

There exist also a significant number of alternate approaches making use of physical multiplication effects of photocharge. The most important of these techniques are described in Chaps. 3–6 and 12 of this book.

1.10 Summary: Physical Limits of the Detection of Light

The deliberations of this introductory chapter leave no doubt about the fact that not only does there already exist a multitude of effective single-photon imaging solutions but also more and more monolithic techniques are being developed with a performance approaching that of existing hybrid solutions. As a consequence, single-photon image sensors at lower cost and with enhanced performance are becoming increasingly available, opening single-photon electronic imaging to many more technical applications and even to the consumer market. For this reason, it is interesting to consider what the physical limits are of the detection of light with single-photon resolution.

1.10.1 Sensitive Wavelength Range

As long as the energy of an incident photon is in a range allowing the photon to interact with the detector material, mobile charges are being generated which can subsequently be detected with an electronic charge detection circuit. For high-energy photons (ultraviolet or X-ray region), more than one unit charge is mobilized per photon, and single-photon detection is easy to accomplish, as explained in Chap. 11. If the energy is too high then the detector becomes effectively transparent to the incident radiation. If, on the other hand, the energy of an incident photon is too small – lower than the bandgap in a semiconductor – then the detector also becomes transparent. As an example, silicon is an efficient detector material for single-photon imaging for the wavelength range between 0.1 nm ($E = 12.4\,\mathrm{keV}$) and 1,000 nm ($E = 1.24\,\mathrm{eV}$).

If a detector material is employed requiring less energy E_g for the creation of mobile charges – for example a semiconductor with smaller bandgap – then infrared

radiation with longer wavelength can be detected. However, as the dark current increases exponentially with the energy $E_g = hc/\lambda_g$ as described by (1.22), single-photon detection is severely hampered for lower E_g. As an example, the exponential factor in the dark current formula (1.22) is 7,300 times larger for germanium ($\lambda_g = 1{,}560$ nm) than for silicon ($\lambda_g = 1{,}120$ nm).

It is concluded that uncooled, room-temperature single-photon electronic imaging is restricted to detector materials with λ_g below about 1,300 nm.

1.10.2 Dark Current and Quantum Efficiency

If a detector needs to exhibit high quantum efficiency then the active thickness of the detector should be comparable to the interaction depth of the photons. If the energy of the incident photons is approaching E_g then the interaction depth becomes quite large, and the active detector thickness should increase correspondingly. However, (1.22) shows that the dark current density increases with the active detector thickness w.

It is concluded that an increase in quantum efficiency is only possible at the expense of increased dark current, and this trade-off is particularly difficult to make if the energy of the incident photons is close to E_g.

1.10.3 Electronic Charge Detection

A key finding of this chapter is the fact that room-temperature electronic detection of charge packets with a resolution of better than one electron r.m.s. is clearly practicable. If this readout noise is less than about 0.2–0.4 electrons r.m.s. one can truly speak of single-electron charge detection.

A combination of such a single-electron charge detection circuit with a suitable photodetector material offering a quantum efficiency of close to 100%, a sufficiently low dark current density and a geometrical fill factor close to unity represents a photosensor with single-photon resolution. Several types of such single-photon electronic image sensors realized using hybrid techniques are described in the chapters of this book.

As the ubiquitous silicon technology allows the realization of on-chip and intrapixel charge detection circuits with less than one electron r.m.s. noise and as silicon offers a quantum efficiency close to 100% over the visible and near infrared spectral range, pixel architectures providing for a geometrical fill factor close to 100% would make it possible to fabricate monolithic and, therefore, cost-effective single-photon electronic imagers. The current development of backside-illuminated CMOS-based image sensor technology, as described in Chap. 2, is exactly this missing link for affordable single-photon imagers, opening single-photon electronic image sensing for wide-spread use in many technical applications and even for the consumer market.

References

1. A. Einstein, Über einen die Erzeugung und Verwandlung des Lichts betreffenden heuristischen Gesichtspunkt, Annalen der Physik, **17**, 132 (1905)
2. M. Fox, *Quantum Photonics – An Introduction* (Oxford University Press, New York, 2006)
3. R.P. Feynman, *Quantum Electrodynamics* (Westview Press, Boulder CO, 1998)
4. D.A.B. Miller, *Quantum Mechanics for Scientists and Engineers* (Cambridge University Press, New York, 2008)
5. H.H. Barrett, K.J. Myers, *Foundations of Image Science* (Wiley, New York, 2004)
6. B.E.A. Saleh, M.C. Teich, *Fundamentals of Photonics* (Wiley, New York, 1991)
7. L. Mandel, E. Wolf, *Optical Coherence and Quantum Optics* (Cambridge University Press, New York, 1995)
8. B.E.A. Saleh, *Photoelectron Statistics* (Springer, Berlin, 1978)
9. F. Wölfl et al., Improved photon-number squeezing in light-emitting diodes, J. Mod. Opt. **45**, 1147 (1998)
10. M. Abramowitz, I.A. Stegun (eds.), *Handbook of Mathematical Functions* (Dover Publications, New York, 1965)
11. S. Winitzki, Uniform approximations for transcendental functions, in *Proceedings of ICCSA-2003, Lecture Notes in Computer Science*, **2667/2003** (Springer, Berlin, 2003), p. 962
12. S.M. Sze, K.K. Ng, *Physics of Semiconductor Devices*, 3rd edn. (Wiley InterScience, New York, 2006)
13. C. Kittel, *Introduction to Solid State Physics*, 8th edn. (Wiley, New York, 2004)
14. E. Kohn, A. Denisenko, Concepts for diamond electronics, Thin Solid Films, **515**, 4333 (2007)
15. E.W. Bogaart et al., Very low dark current CCD image sensor, IEEE Trans. Electr.Dev. **56**, 2642 (2009)
16. F. Reif, *Fundamentals of Statistical and Thermal Physics* (Waveland Press Inc., Long Grove Ill., 2008)
17. H. Nyquist, Thermal agitation of electric charge in conductors, Phys. Rev. **32**, 110 (1928)
18. A.S. Sedra, K.C. Smith, *Microelectronic Circuits*, 6th edn. (Oxford University Press, New York, 2010)
19. F. M. Klaassen, J. Prins, Thermal noise of MOS transistors, Philips Res. Rep. **22**, 505 (1967)
20. C. Jakobson, I. Bloom, Y. Nemirovsky, 1/f noise in CMOS transistors for analog applications from subthreshold to saturation, Solid State Elect. **42**, 1807 (1998)
21. P.E. Allen, D.R. Holberg, *CMOS Analog Circuit Design*, 2nd edn. (Oxford University Press, New York, 2002)
22. Ch. Lotto, P. Seitz, Synchronous and asynchronous detection of ultra-low light levels, in *Proceedings of the 2009 International Image Sensor Workshop*, Bergen, Norway, 26–28 June 2009
23. G.R. Hopkinson, D.H. Lumb, Noise reduction techniques for CCD image sensors, J. Phys. E: Sci. Instrum. **15**, 1214 (1982)

Chapter 2
Image Sensor Technology

R. Daniel McGrath

Abstract Single photon imaging is an extension of solid state imaging, making use of devices that benefit from the outstanding electro-optical properties of silicon and which have the potential, like CCD and CIS devices before, to benefit from the integration made possible with silicon process technology. This chapter provides an introduction to the functionality of integrated image sensors. It presents the typical process flow and a discussion of where the technology used to build CMOS logic and mixed signal parts is useful to provide the special performance and unique features expected in imaging. As a conclusion, process enhancements including special process modules and design rules for high-performance image sensors are discussed.

This chapter also addresses the question to which degree single photon imaging devices can be an extension of the CIS model to achieve mass commercialization. The goal is to build valuable imaging products with an existing fabrication tool set, derived from a conventional CMOS process with some process enhancements using existing tools. This implies that the commercially interesting pixel array design can be achieved using special design rules, and that the process enhancements will support the operating voltages and clock rates required for single photon detection.

2.1 Program and a Brief History of Solid-State Image Sensors

Single-photon imaging is an extension of solid-state imaging where the sensitivity of the photodetection process is systematically increased until the arrival of individual photons can be detected.

There were attempts in the 1960s thru 1980s to realize solid-state image sensors with developing transistor technology and designs. Some foreshadowed the later metal-oxide-semiconductor (MOS) image sensors [1] and led to specialized sensors used for noncommercial imaging applications. Despite attempts at commercialization, the performance was not satisfactory due to process mismatch and defects. Attempts to overcome this with special transistors, for example, charge modulation

P. Seitz and A.J.P. Theuwissen (eds.), *Single-Photon Imaging*,
Springer Series in Optical Sciences 160, DOI 10.1007/978-3-642-18443-7_2,

devices (CMDs) and static induction transistors (SITs), did not overcome these limitations.

Charge-coupled devices (CCDs) emerged as the first widely commercialized imaging technology [2]. The CCD is a massively parallel device. As such it could be free to deviate from the design practices used for transistors. Instead of device geometries being defined by the active mask, that is openings in the field oxide, and a single gate pattern, the devices are defined by implant geometries and one or more gate patterns. It was necessary to develop processes providing specialized functionality, for example, buried channel potential profiles and vertical overflow drains. The process tools are the same used with mainstream silicon manufacture, but can have the goal change from achieving minimum geometry to achieving precise alignment. Operating voltages at multiple levels spanning a wider range from −10 V (used to provide low dark current while holding charge) to +20 V (used to provide a global clear of the array) are required. In addition, special care was needed for certain processes to control defects. The result was a specialized process using a combination of process modules used for logic and mixed-signal processes and of process modules optimized for imaging, for example, low defects. As in other process enhancements, the tool set lagged by generations behind the cutting edge logic process.

With the steady improvements in complementary metal-oxide-semiconductor (CMOS) technology since the 1970s, it was proposed in the 1990s [3] that CMOS image sensor (CIS) designs using the existing processes could meet image sensor performance requirements [4], not only achieving CCD performance within CMOS fabrication facilities, but also surpassing this in some aspects including data rate, integration with circuitry and cost. In reality, CIS was able to provide required performance only with process enhancements including the mixing of process tool sets between process technologies, that is adding more advanced interconnect technology. The result is that CIS products for mass markets can be built in CMOS process fabrication facilities where an investment has been made to develop and support process enhancements, though this requires implementation of special design rules to optimize the pixel array.

The expectation for single-photon imaging devices is that they will extend the CIS model. That is, that they can be built based on an existing fabrication tool set, derived from a conventional CMOS process with some process enhancements using existing tools. The pixel design will use special design rules. The enhanced process will support operating voltages that are obtainable from the process tools.

2.2 Anatomy of an Image Sensor

The function of an image sensor is to collect photons, convert them into electrical signal, and process this electrical signal (Fig. 2.1). The optical collection and the conversion are functions of the pixel. The processing is done using analog and digital CMOS circuitry downstream. The more sophisticated nuances of image sensors, those that provide the quality required for a photograph or a video, are not an issue in single-photon image sensors. Rather what is required is an

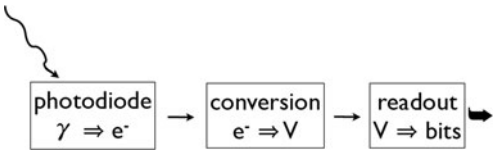

Fig. 2.1 Image sensor function

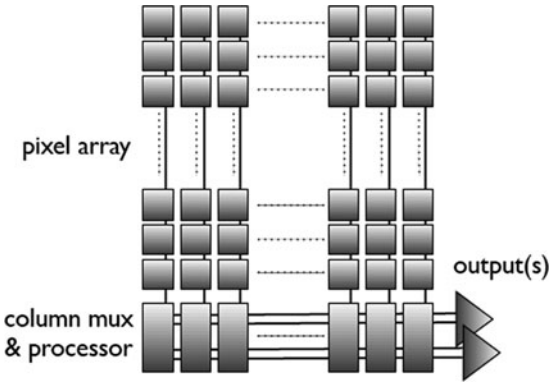

Fig. 2.2 Image sensor array architecture

understanding of the application-specific functionality for single-photon imaging, so there is only a need to understand the straightforward architecture. For instance, the single-photon imaging application does not require the number of pixels or the minimum-sized pixels of typical image sensors. But the basic functionality of light capture, conversion and readout is that of other image sensors.

The typical architecture of a commercial image sensor is a two-dimensional pixel array with a one-dimensional array of analog column readout circuitry, which is multiplexed onto one or more output channels (Fig. 2.2). The pixel array has local wiring plus drive, bus and signal lines, which form the optical aperture of the pixel.

For a CCD, the busing is done by extending the polysilicon gates, and the readout is done by shift registers so that interconnects can be fabricated with one layer of metal. While the number of metal layers is small, the gates must control the charge transfer regions leading to an issue of the gates absorbing visible light. The conversion from photons to charge is done either in the shift register, that is, a frame transfer pixel, where photons are absorbed in the silicon of the CCD register itself or in a photodiode from which charge can be transferred into the CCD register, that is, an interline pixel. Readout of the pixel array is through charge domain transfers with transport of charge packets realized under the influence of potential profiles. The photocharge is shifted from where it is generated into a local maximum where it is held and then through the manipulation of gate potentials through successive potential wells that are sequentially created (Fig. 2.3). The column circuit provides charge domain multiplexing into a horizontal shift register. At the end of each channel horizontal register, there is a diode capacitor, which converts charge to voltage. The only transistors, typically three to five, on the image sensor are involved in the buffering of this voltage, using a floating diffusion capacitor to convert charge

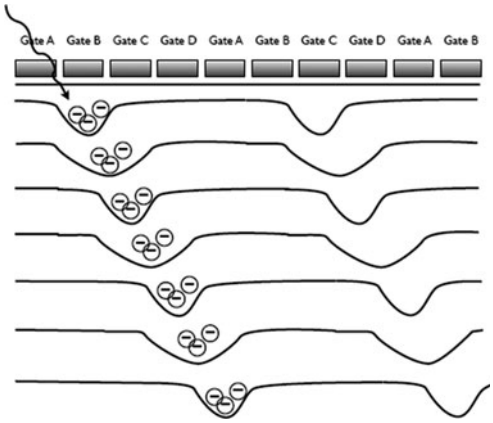

Fig. 2.3 CCD register operation

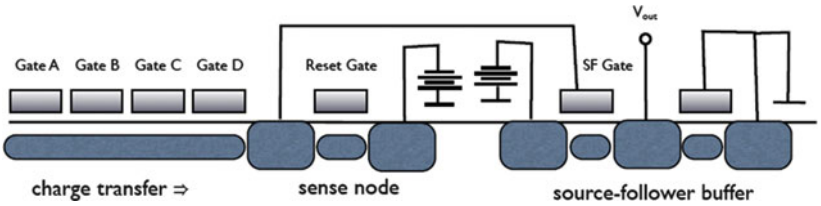

Fig. 2.4 CCD register operation

into voltage and providing low impedance drive for the signal output from the die (Fig. 2.4).

While a CCD is a MOS structure, it is more a device than a circuit. A series of process enhancements have been required to get excellent performance for imaging. To minimize light lost to the control gates, process enhancement (Figs. 2.5 and 2.6) include the following: (a) thinning the device to a silicon thickness of 10 μm to <3 μm so that photocharge can be generated and collected from photons illuminating the back surface; (b) creating photodiodes separate from the shift registers with control gates to transfer readout; and (c) opening subpixel-sized holes in the gate structures next to the shift registers with pinning layers to maintain the potential within the charge gathering and charge transfer depletion regions in the silicon. To increase charge readout efficiency and lower noise, doping has been developed to create buried channels in the shift registers and the readout transistors, so that the signal charge does not interact with the silicon–SiO_2 interface traps. To control regions created with no gate to increase optical response, shallow JFET-like pinning layers have been added, enabling the creation of interline transfer pixels (Fig. 2.7). The pixel-specific process enhancements lead to a process, which has little overlap with the process needed for building CMOS mixed-signal circuitry.

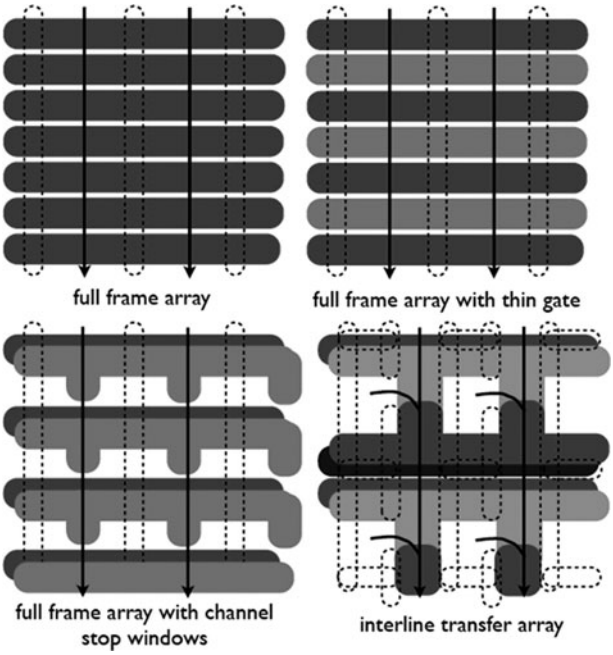

Fig. 2.5 CCD array options (*top view*)

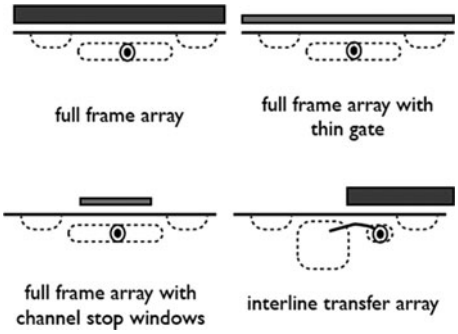

Fig. 2.6 CCD array options (*side view* across channel)

For a CIS, illumination is typically from the front side of the die, so that the optical aperture is formed by horizontal busing on one metal layer and vertical bussing on a second metal layer [5]. Depending on whether the local interconnect can be integrated with one of these bus layers, the pixel may require two or three metal layers. (The other circuitry integrated on the chip may require the use of more metal layers.) Photons are converted to charge in a photodiode and charge is converted on a diode capacitor, which may be the photodiode, "3T pixel," or a floating node onto which the charge is transferred by a gate, "4T pixel" (Fig. 2.8).

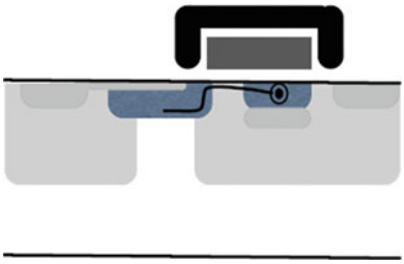

Fig. 2.7 CCD interline-transfer pinned-photodiode (*side view*)

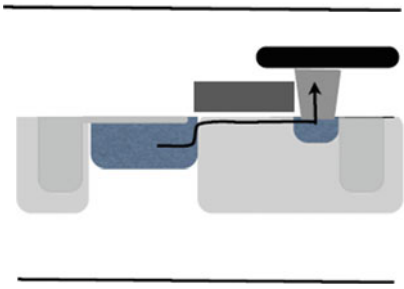

Fig. 2.8 CIS pinned-photodiode (*side-view*)

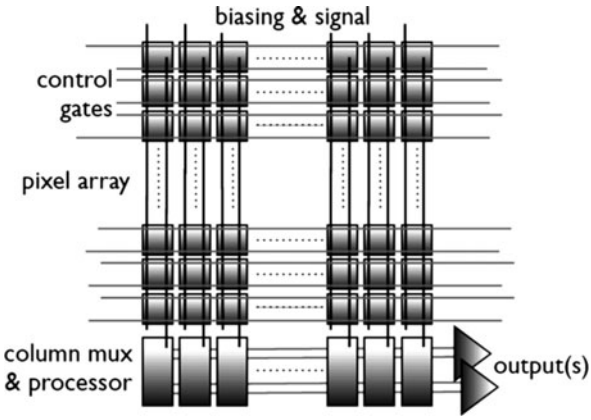

Fig. 2.9 *x–y* Nature of image sensor array with analog column processors

The one-dimensional column circuit array samples the readout from a row of pixels and places the signal onto channel buses, which feed into a gain stage followed by an analog-to-digital (A/D) conversion circuit for each channel (Fig. 2.9). An alternative architecture is to distribute the gain and A/D function into the column circuitry so that the channels are digital buses (Fig. 2.10) [6].

There are other possible architectures, for instance the CMD [7] or SIT [8] devices, where the photons are converted into charge, which is held on either a lateral or a vertical diffusion that modulates the current in a channel in the silicon. The motivation for this architecture is that it results in higher conversion gain due to the junction gate field-effect transistor JFET-like structure. These pixels have design challenges, linked with the implementation of reset function (e.g., memory effects due to incomplete reset), to the optical aperture (e.g., structures that limit a front-side aperture) and readout (e.g., lack of process control in readout threshold voltage).

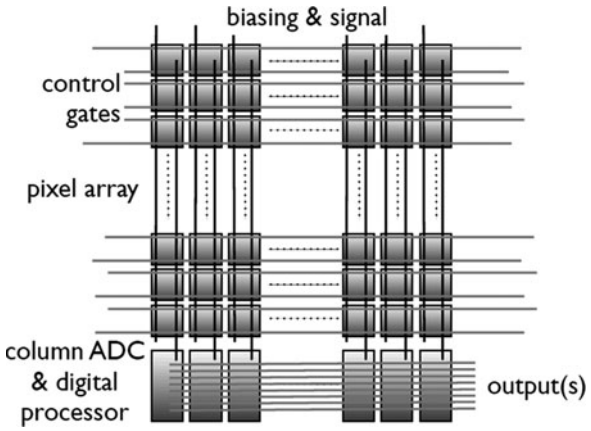

Fig. 2.10 *x–y* Nature of image sensor array with digital column processors

2.3 Operation

The basic operation of an image sensor relies on a one-dimensional or two-dimensional array of pixels. The pixels in this array, typically laid out in a regular array of rows and columns, must provide a means to collect the light and convert it into electrons or holes. Then through some of elements in the pixels, in the array, in the column, and in a pipeline after the array, the signal is conveyed to the chip output.

In silicon image sensor arrays, the typical sensing element is a volume of the silicon itself. Photons are absorbed by creating electron–hole pairs across an indirect bandgap of 1.12 eV. This creation of photocarriers requires interaction with the lattice phonons to conserve momentum and energy. The bandgap is the right energy for photon absorption from the ultraviolet through the visible into the very near-infrared with one photon creating one pair of photocarriers. However, the absorption coefficient of silicon varies significantly between blue light absorbed very close to the silicon surface and red light where complete absorption requires several micron thickness.

There is also photoresponse due to more energetic photons, such as X-rays and gamma rays, which create multiple photocarriers per absorption in a single location and from charged particles such as beta rays, energetic electrons, high-energy particles, and cosmic rays, which create a trail of electron–hole pairs and often an end event consisting of a rather localized cloud of electron–hole pairs.

In a block of silicon that is not electrically biased, the photocarrier creation would be balanced by an increase in recombination, providing net current neutrality. For an image sensor pixel, it is necessary to take advantage of the semiconductor behavior of silicon to create a region out of equilibrium (e.g., a depletion region completely drained of mobile charge carriers), so that the photocarriers can be collected before they recombine. It is also necessary to have the low defect density

present in semiconductor grade silicon with its associated long carrier lifetime, so that the probability of recombining is negligible during this transport phase.

Despite the high quality of single crystal silicon, there is background charge generation (i.e., dark current) independent of the illumination level. This is due to defects in the middle of the energy band from metal contamination, crystal defects and from the interface states between the surface of the silicon and insulators such as the gate oxide, the field oxide, or oxide-filled trenches. A key enabler for MOS devices is that the oxide–silicon interface is of very high quality, but natural mismatch between the amorphous oxide ring structure and the silicon lattice still contains a layer of defects. A dark-generated carrier and a photocarrier cannot be distinguished during image sensor operation, so dark current limits low light performance.

Dark current generation is a critical performance limitation for single-photon imaging because each defect-generated charge will trigger a false event in the imager. Covering the topic sufficiently is beyond the scope of this article, but it is worth providing a short overview. Image sensor pixels are built around regions in the silicon biased out of equilibrium in deep depletion so as to hold charge isolated in their band structure, for example, the photodiode or the floating diffusion used for charge conversion. These are subject to nonphotonic charge generation. Defects, such as metallic contaminates, located within a deep-depletion region can provide mid-bandgap energy states that act as isolated stepping stones, stepping stones isolated in both location and energy, for charge transfer from the valence band to the conduction band, forming dark current defects that are often isolated to a single pixel and very bright. Interface states at silicon–insulator interfaces within the depletion region provide a "U-shaped" continuum of such stepping stones with the minimum at mid-energy gap, but with the ones with energies near the center of the band gap being efficient producers with a distribution dark current sites with lower level generation rates. These generally are distributed with statistics across the pixels providing a noisy background in an image that can go from a near uniform distribution to a Milky-Way-like starry night. Defects located outside the deep-depletion region can also cause the neighboring silicon to be perturbed from equilibrium resulting in diffusion current thermalized to the bandgap energy. The current from these "bulk" defects provides a background distribution that contributes to all pixels. (Defects within the deep-depletion region and defects outside it can be distinguished by the activation energy of their generation rate: depletion traps are characterized as having energy near one half of the bandgap energy, while diffusion defects are characterized by having energy near the bandgap).

Dark current generation can be enhanced by compressive-stress-induced bandgap narrowing, by perturbations of the defect potential profile due to high fields, and by increased defect sites due to a combination of increase in crystalline defects, for example, dislocation planes and implant end-of-range damage. Dark current can be decreased by inducing defect sites, such as crystal defects or oxygen-precipitates away from the critical nonequilibrium depletion region, where metallic defects can diffuse and become trapped without generating defect charge. Dark current can

also be decreased by minimizing damage within the depletion region or within the photocharge region within the diffusion paths of the depletion region. Thermal processes can anneal damage. Final thermal anneals with hydrogen or deuterium provided by the ambient or from hydrogen-rich dielectric layers can decrease electrically active interface states, decreasing dark current from these. Device design can minimize the contact in space and over operation for the depletion region with interfaces and enhancements due to electric fields and stress.

This current of photocarriers can be sensed and read out in various ways as follows:

(a) The photocurrent can be integrated in a potential well. The charge can then be transferred by using gate voltages to modify the potential of the well and of adjacent wells to transfer the signal charge through the array to an output node.
(b) The photocurrent can create a voltage drop across a resistor or a subthreshold MOSFET with the signal being sampled during readout rather than being integrated.
(c) The photocurrent can be integrated onto an isolated capacitive node created by space charge depletion regions, which are in a nonequilibrium state due to the biasing sequence of the operation.

CCD's use Option (a). CIS and other MOS sensors use Options (b) or (c) [9]. The result in all cases is to transform the photogenerated charge into a voltage, which is a function of the illumination level. Options (a) and (c) can provide a near-linear transfer function, while Option (b) can be used when a nonlinear performance (e.g., a logarithmic response) is desired.

The arrangement of the pixels as an array makes it natural to read out the pixels in a row-by-row order. A repeated pattern of horizontal bus lines connect single pixels or shared pixels (i.e., multiple photodiodes sharing a single readout) and provide clock signals to provide charge transfer, reset, and pixel select functions.

At the top and/or bottom of the columns of the array are output multiplexers built using charge domain or using MOS circuitry. Biasing and signal readout providing connections to the pixels within a column are provided by vertical bus lines.

The simplest readout is a nested pattern of row timing with the period of the row read and of column timing with the period of the output signal readout. In a typical sequence, a row of pixels is reset to clear the pixels in it of any initial charge by reading them out, by using a gate to force the charge to substrate or by using a gate to connect them to set voltage. The signal current or the integrated signal charge is read out, either buffered to give it sufficient drive impedance or directly digitized.

2.4 Image Sensor Devices

Silicon image sensors are made up of a combination of MOSFETs and diodes, but making an effective pixel requires extending it to new devices and to novel operation for existing devices.

The basic semiconductor device is a diode made by doping two adjoining areas in the semiconductor n-type and p-type [10]. Diodes can be reversed biased to provide isolation between the two regions of different doping to allow the separation of circuit elements and to create capacitors. Typically, the diode is formed with one of the regions having higher doping than the other to create an asymmetric region depleted of free charge to provide manufacturing control.

A second semiconductor device is the MOS capacitor, created between a heavily doped region in the silicon and a polysilicon gate separated by a thin silicon oxide. For example, a gate oxide capacitor designed to work at 3.3 V might have dielectric thickness of 7 nm, or a sandwich of silicon nitride between two layers of silicon oxide, for example, 100 nm for a 10 V process down to 7 nm for a 3.3 V process.

The MOSFET used in digital logic consists of a MOS capacitor placed between two diodes. The gate is patterned on top of a thin silicon oxide or oxide nitride layer so as to modulate a channel along the insulator silicon interface so as to act as a switch [11]. Heavily doped, very shallow diffusions are implanted at each end to provide the contacts for the carriers. Special processes such as implanting silicon or other species may be used to amorphize the silicon to keep the implant from becoming deeper through channeling, but these may cause a problem in an imaging device if it results in residual defects. An opposite type of doping in the form of a deep well and of a shallow threshold voltage adjust are formed under the channel to control the operating voltages. Engineering is done to lower the resistance and to set the work function of the gate through doping and the formation of a silicide layer. The potential profiles at the ends of the channel controlled by the gates are adjusted through the sidewall spacers formed with deposition and etching of silicon oxides and nitride layers. The potential under the spacer is engineered for speed and for reliability through lightly doped drain diffusions, LDD, as extensions of the diffusions and "halo" diffusions of the opposite polarity.

The MOSFETs in analog or mixed signal circuits has the same structure, but perform as transfer functions, either acting as switches to connect and isolate nodes or acting as analog components providing precise outputs depending on the voltage applied.

The CCD in its simplest form, a full-frame transfer array, is a single large integrated device built around registers of abutting MOS capacitors requiring at least one adjacent diode to provide a bias and a sink for its photocarriers and its dark current charge [2]. A single reverse-biased diode next to a MOS capacitor can create a nonequilibrium condition under the capacitor where the mobile charge near the surface is drained away to create a depleted region. The surface potential of silicon under the capacitor gate has a potential that depends on the diode and upon the gate voltage. In the same way, a series of capacitors can be held in nonequilibrium by a single diode with each capacitor being able to control the potential under its gate. By applying different voltages to the various gates, local potential maximums can be created where isolated packets of charge generated by light or by dark current can be collected. By a sequence of biases on the capacitor gates, the maximum of the potentials can be shifted from capacitor to capacitor until it reaches the diode. The capacitor gates can be arranged in more complicated arrays with the charge packets

being moved synchronously in parallel arrays of shift registers and with parallel arrays being multiplexed into single serial arrays.

The CCD is commonly read out using a floating diffusion connected to the gate of a MOSFET. The floating diffusion is a diode inserted between two capacitor stages. On one side, there is a stage which acts as a switch to set a bias of the isolated diode, either by providing a barrier with its channel so that excess charge above this potential is drained off to a biased final diode or by acting as a switch to create a channel for charge to the biased final diode to force the floating node to this voltage. The gate of the MOSFET follows the bias on the floating node and modulates its channel potential to dynamically set the voltage of its source diffusion to create a classic source-follower circuit.

To address problems related to signal charge being corrupted by traps at the surface of the capacitors either capturing signal charge or creating dark current, a buried channel capacitor is formed where the surface is counter doped so that the potential maximum where charge is held is shifted from the interface with the gate to a potential channel deeper in the silicon. This keeps the signal charge away from interface traps to prevent signal charge from being trapped; it allows the potential well to be clocked so that for part of its duty cycle the surface is accumulated with free carriers to turn off the generation of dark current by defects there.

While logic and analog MOSFETs can be operated between a low voltage, typically ground, and a high voltage, Vdd, to get the benefit of a buried channel requires operating outside this range. This is because the buried channel has a threshold voltage that is shifted as in a depletion device with the channel being on at ground.

The need for control gates over all of the registers limits the responsivity of CCDs when used with front-side illumination. Photons need to enter through the gates or through the gaps between the gates. A solution to this used in interline transfer CCDs is the pinned photodiode, which consists of a diode doping located between a shallow, highly doped layer and a deep well doping with these two doping regions electrically connected. The result is a back-to-back diode controlled by the surface doping, so that it does not need to be covered by an obstructing gate. The pinned photodiode is read out by an adjoining gate providing a barrier to a shift register, which with a high gate bias provides a potential path. The pinned photodiode potential is a buried channel.

Barrier devices are created to provide means to control blooming in CCDs (i.e., the photocharge from one photodiode that has become full spilling into other photodiodes, creating image artifacts due to signal in the wrong location in an image) and to provide a means to flush unwanted charge. These are areas of lower potential along the shift register, where charge filling up a potential well will preferentially flow rather than spill into other pixels in a shift register. These can be static barriers to a laterally adjoining biased diffusion, gate-controlled stages with a lower potential to laterally adjoining biased diffusions or vertical barriers over the well to an oppositely doped substrate. The overflow will occur either because the stage is filling with charge or because the channel-to-barrier is collapsed by

lowering the channel potential through gate control or, in the case of the vertical barrier, a combination of gate and substrate control.

The CIS pixel typically consists of photodiodes, diffusions used as floating nodes to hold charge and MOSFETs providing the three functions of switch, charge transfers device and analog buffer. While there are overlaps in the functions of the devices, the technology differences between CCDs and CISs make for differences in the device design.

The MOS vertical, parasitic, capacitor can be used for a photodiode and as a component of a floating diffusion capacitance. The top plate is a shallow heavy doping, the bottom is the well doping and the dielectric is the depleted silicon in between. Unlike in the peripheral circuitry, this diode's upper surface would not be silicided to let photons in for a photodiode and to avoid the extra dark current associated with silicide–silicon interfaces. This photodiode can be directly contacted by a contact so it can be wired to the gate of a buffer FET or a diffusion of a reset FET.

A pinned photodiode like the ones in an interline CCD is commonly used for CIS photodiodes due to its ability to be totally emptied of signal charge during reset and its low dark current [12]. It can be built adjacent to an FET to provide readout and reset and antiblooming. Unlike the CCD version, it cannot be reset to substrate and antiblooming is only possible laterally through FET structures. Its long wavelength response can be enhanced by chaining doping to create a deep diffusion volume.

The pixel has MOSFETs that are each customized for their function. In the same manner as FETs in the input–output circuits on the die, they are built with thicker gate oxide (e.g., $\sim$ 7 nm required for 3.3 V signals) because the voltages required for pixel operation does not scale as fast as logic. Beyond that, there are three functions required with their own requirements as follows [13]:

- Buffer FET such as that used in source-followers need to have a threshold shift doping to set the gate-to-source voltage drop to be a medium value, a value as low as possible without requiring increased gate length. They need to provide good transconductance with small current.
- Barrier devices as in CCDs are used to reset and maintain the voltage in floating diffusions. They are switches where having a zero or depletion threshold voltage allows operation within the normal voltage range.
- The transfer FET used to read out the pinned photodiode is unique in that it is a charge transfer device, which transfers charge from a buried channel device to a surface device. It provides a low potential channel from photodiode to floating diffusion to prevent lag, a high off-barrier for well capacity and a potential profile that minimizes dark current and directs what is generated away from the photodiode into the floating diffusion. The result of this is that it needs to be viewed as two FETs in series sharing a single gate, that is (1) a vertical FET at the photodiode end controlling a potential barrier away from the surface, which sets the photodiode charge handling capacity when "off" and allows complete

transfer when it is "on" and (2) a conventional surface channel FET controlling a horizontal channel connecting to the floating diffusion.

A typical CIS pixel only uses either n-MOSFET or p-MOSFET devices due to the need for tight packing of devices to enable small pixels [14, 15]. The full CMOS process is needed to enable low power, fast addressing and readout by allowing on-chip circuits.

The need for the development of new devices within the process enhancement can also extend beyond the pixel array to peripheral circuits required to provide the best performance for a pixel array. Examples include having both n-MOS and p-MOS devices with floating bases (achieved through the use of deep dopants to isolate the local well), having capacitors purposely made to achieve linear transfer functions to provide double sampling to shift time differences into voltage differential signals, and having dense memory to allow on-chip line stores for correlated signal correction.

2.5 Image Sensor Process Technology

The goal of mass market and high-performance image sensors is to build them in silicon integrated technology, so that they can benefit from the manufacturing infrastructure, expertise, and cost efficiencies existing because of the large global production for silicon integrated circuits. At the top level, this means building a planar technology on silicon wafers in a clean room environment [16, 17]. At the tool set level, it means wherever possible using the standard equipment set for manufacture and test. And at the process level, it means wherever possible using standard process modules and materials.

The top level flow is that wafers are sawed from single-crystal silicon ingots (boules) of semiconductor-grade purity pulled from a silicon seed crystal. While it is possible to use this wafer directly for building the integrated circuit, typically for image sensors, an additional thin epitaxial layer is grown in a reactor. The substrates and the epitaxial layers are doped as part of the growth process either n- or p-type. The doping type and the concentration of the epitaxial layer and the substrate can independently be varied. The wafers are introduced into a cleanroom environment and follow procedures to maintain cleanliness as handled by machines and people. Because the planar process is going to be used to create fine geometries precisely aligned, it is important that the wafers be flat and that the particle density is extremely small.

The basic tool sets silicon processing includes the following:

- Photolithography to create patterns to define structures in and on the silicon.
- Developing to remove organics patterned by photolithography.
- Thermal oxidation to create a high-quality oxide–silicon interface.
- Plasma deposition of insulating films used in building capacitors and MOSFETs, used to electrically isolate interconnect metal layers and used to provide device passivation.

- Spin-on coating to create insulating films used for creating optical elements such as color filters and microlenses.
- Chemical vapor deposition of polysilicon to create gates for MOSFET devices.
- Silicidation to chemically create a shunt layer on silicon diffusions and polysilicon.
- Plasma etching to create trenches or pattern overlayers.
- Metal deposition of aluminum, tungsten and copper and of boundary and antireflective layers for metals.
- Chemical mechanical polishing to planarize a surface (e.g., to achieve the flatness required by the depth of field of the photolithography optics) or to remove excess metal from a damascene process (i.e., a process where metal is defined by etching patterns in the underlying insulating layer and then polishing the metal back to the dielectric surface).
- Thermal treatments, either furnace or rapid thermal annealing using a pulsed heat source (e.g., for implant activation and reduction in crystalline defects).

A typical MOS process flow (Fig. 2.11) starts with a wafer with a lightly doped device layer (e.g., a doping that is low enough to result in a large depletion zone for charge collection but high enough to be controlled during the subsequent process), either substrate or epitaxial. Shallow trench isolations (STIs) are patterned, etched, and filled to provide isolation between devices and in the process of defining the dimensions of the devices in the circuit, in particular the channel width. A pair of photolithography steps divide the wafer into regions that are to get a chain of implants to create n-wells and regions that are to get a chain of implants to create p-wells. This establishes which devices are to be p-type and which are to be n-type. Photolithography and implants can be done at this point to set the threshold voltage of the MOSFETs. Gate oxides are thermally grown for each of the voltage ranges to be fabricated. Polysilicon gates are deposited, doped, patterned, and etched, providing the other critical dimension of a MOSFET, the channel length. The engineering of the MOSFET is then done through a combination of sidewall formation, halo and lightly doped drain implants and source–drain implants to control the channel ends and through silicidation to lower gate and drain resistances. Thermal processes are needed to activate the implanted doping that is to make sure the dopants substitute for silicon on the crystal lattice sites. After this there are

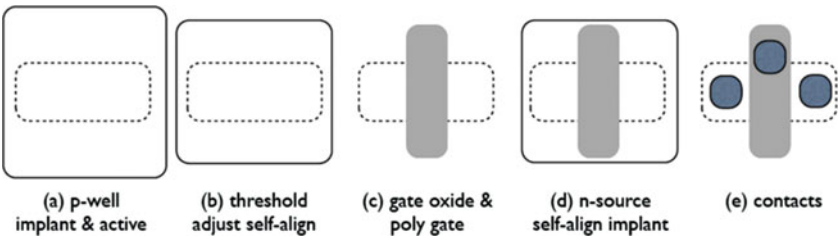

Fig. 2.11 Image sensor array architecture

a series of insulator depositions, contact and via patterning and filling and metal filling. At various critical points in the process, chemical–mechanical polishing is done to provide a flat surface to allow precision photolithography in following steps. At the end of the process, there is a final silicon nitride layer deposited for circuit reliability followed by a patterning and etching of this layer to create bond pads.

The basic device in a MOS process is the FET. It is formed in a well of the appropriate doping. Its critical dimensions are determined a geometry created by the STI pattern and by a geometry formed by the gate pattern. The issue that drives a MOSFET process is small size, consistent performance, device speed, and drive capability.

In going to "nanometer" processes (i.e., processes with feature sizes much less than 1 μm), there have been a couple of major refinements to the MOSFET process. One is a need for shallow doping distributions requiring precise control of thermal cycles, leading to the shift from furnace to rapid-thermal anneals. A second is a shift from aluminum to copper for metallization with the shift to a damascene process, where the metal pattern is established by etching trenches in the isolation and etching off metal not in the trench and where blanket protection layers are formed at each metal level.

CCD technology originated in the same time period when CMOS technology was being developed, before the semiconductor industry became dominated by it. CCD technology development was not constrained in either layout or the development of process modules. In contrast to the MOSFET device, the basic CCD device is made up of charge transfer channel created from gates from a series of layers of polysilicon, each one defining the length of a stage, running across a channel defined by channel-stop implants. There is only one well doping, which may be patterned at the pixel level to provide an antiblooming barrier to the substrate and which should be as high as feasible to improve the photodiode collection depth. And there is a counter-doping to relocate the channel from the surface to the substrate. There can be additional doping to direct charge flow within a single transfer stage and facilitate the transfer between phases. Getting good transfer in this buried channel means that electric fringing fields are important in setting the potentials for charge handling wells and channels created deeper in the silicon away from the gate insulator interface so there is benefit in continuing to use thicker oxide (e.g., there is a need to smooth the lateral potential profiles to avoid creating local potential wells that trap charge). Getting good optical response lead to the use of pinning implants to control buried channel potential in regions without gates. It also can lead to the development of a thinning process so that the final device can be illuminate from the back surface (albeit with the resulting loss of longer wavelengths from green into the near infrared). Silicides are not used on the gates. The massively parallel nature of a CCD pixel array means only one or two layers of metal are necessary.

A CCD process flow (Fig. 2.12) starts with a wafer with a lightly doped device layer, either substrate or epitaxial. It will have a specification for metal contamination levels and will have gone through some form of processing, such as oxygen precipitation to provide gettering of residual contaminants. Local oxidation isolation (LOCOS) or trench isolation may be used, but with doping in the trench

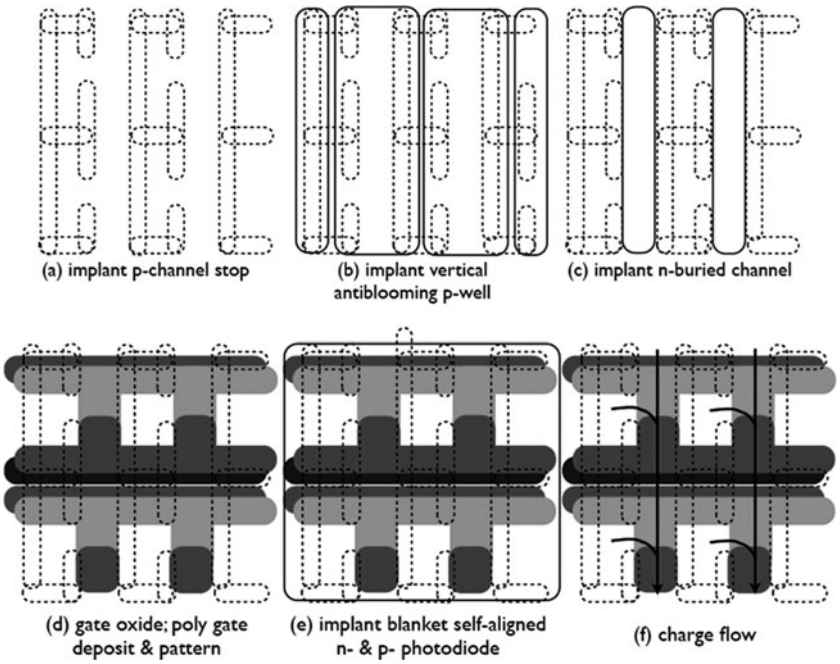

Fig. 2.12 CCD simplified process

walls providing passivation of the trench walls. Alternatively, there may be no trenches at all with the isolation being formed by just the doping between devices. Photolithography defines an in-pixel-patterned well that can be diffused into the wafer and laterally to contribute to the channel potential and provide isolation. Photolithography and implants can be done at this point to create the buried channel and to adjust the channel potentials in the transfer stages. Thick gate oxide is thermally grown for each of the voltage ranges to be fabricated. Multiple layers of polysilicon gates are deposited, doped, patterned, and etched, providing the other critical dimension of a transfer stage, the channel length. Thermal processes are needed to activate the implanted doping, that is to make sure the dopants substitute for silicon on the crystal lattice sites. After this, there is an insulator deposition, contact and via patterning and filling and metal filling. At the end of the process, there is a sequence of passivation deposition and post-metal anneal in a hydrogen environment to move hydrogen to reduce defect concentration by reacting with dark-current-generating dangling bonds at silicon–oxide interfaces. There is a patterning and etching of this layer to create bond pads. Following this is a sequence to form color filters and microlenses to provide color separation and increased responsivity.

To increase the responsivity, backside thinning to a few micrometers through the use of silicon-on-insulator or physical and chemical etching to allow backside illumination.

The issues that drive CCDs are optical response, near perfect charge transfer, and charge generation from defects. Defects can be due to defects at silicon–oxide interfaces and to metal contamination in the silicon.

While early on the claim for CIS devices was that they could be built using a standard CMOS process, the issues that drive the CIS technology today is to maintain CMOS functionality for peripheral circuitry while building a pixel array that meets the same challenges as CCDs, the optical response and the low defect levels. The approach for the CIS process then is to start with a CMOS process and add to it process enhancements that provide high-performance pixels, bearing in mind that the enhancements must leave the devices built in the base process unchanged in performance.

A CIS process flow (Figs. 2.13 and 2.14) starts with a wafer with a lightly doped device layer, either substrate or epitaxial. The wafer needs to have extremely low contamination and should go through gettering processes to reduce the contamination level further. Shallow trenches are patterned, etched, and filled to provide isolation between devices and in the process defining the dimensions of the devices in the circuit, in particular the channel width. The shallow trench process includes doping to passivate the upper corner, the sidewall and the bottom of the trench. A photolithography step divides the wafer into regions that are to get a chain of implants to create n-wells, regions that are to get a chain of implants to create p-wells, and the region in the pixel array, where the well doping is patterned at a

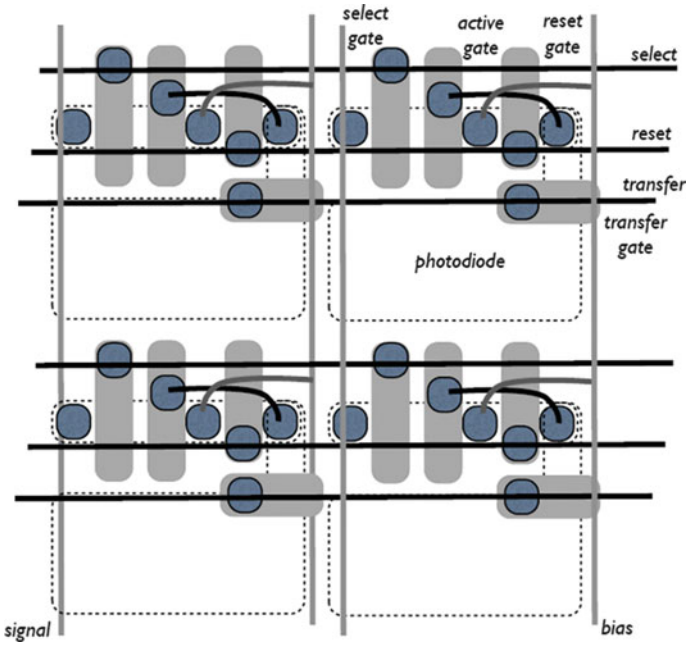

Fig. 2.13 CIS nonshared pinned-photodiode pixel

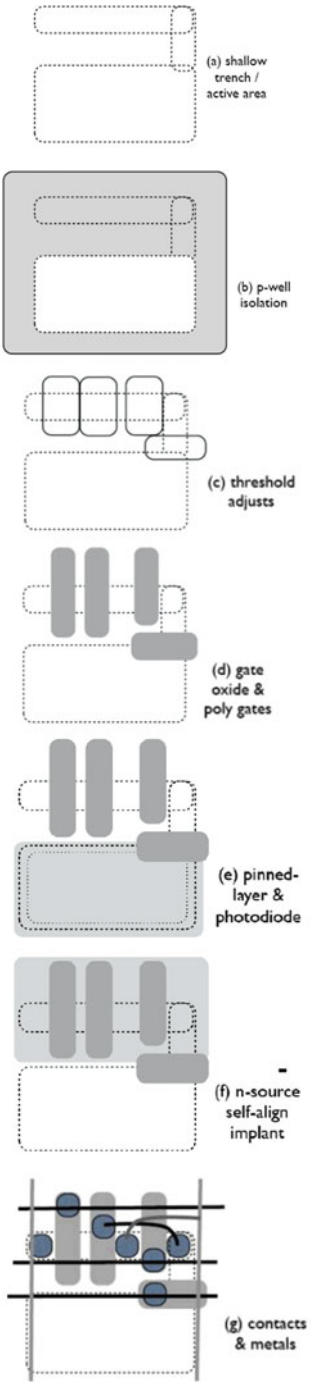

Fig. 2.14 Simplified CIS process

pixel scale with areas such as the photodiode left native. This establishes which devices are to be p-type, which are to be n-type, and which will be special devices for the pixel. Photolithography and implants can be done at this point to set the threshold voltage of the MOSFETs. Gate oxides are thermally grown for each of the voltage ranges to be fabricated with the higher voltage devices being used in the pixel array. While the base CMOS process for commercial image sensors has scaled to the 130 and 90 nm technology node, the operation of many pixels is based on devices and voltages from the 350 nm technology node. Polysilicon gates are deposited, doped, patterned, and etched, providing the other critical dimension of a MOSFET, the channel length. The engineering of the MOSFET is then done through a combination of sidewall formation, halo and lightly doped drain implants, and source–drain implants to control the channel ends. Additional depositions, photolithography and etches are used to engineer the charge transfer from the photodiode through the transfer gate and to the floating diffusion. Silicidation to lower gate and drain resistances are done selectively, excluded from the diffusions in the pixel array because they block light over the photodiode and because of its association with increased defect generation. Additional photolithography and implantation is used to create the pinned photodiode, forming a shallow doping for the pinning layer and providing a chain of implants to engineer the photodiode to extend the photodiode diffusion to extend deep into the silicon with a large aspect ratio. Thermal processes are needed to activate the implanted doping, that is to make sure the dopants substitute for silicon on the crystal lattice sites. After this, there are a series of insulator depositions, contact and via patterning and filling and metal filling. At various critical points in the process, chemical–mechanic polishing is done to provide a flat surface to allow precision photolithography in following steps. Additional process steps may be done to create a lightpipe (i.e., an optical means of guiding photons between the front-side interconnect structure) to enhance optical acceptance, particularly in processes with the copper metallization with its additional dielectric layers that would otherwise create a dielectric stack resulting in interference effects. At the end of the process, there is a sequence of passivation deposition and postmetallization anneal in a hydrogen environment to move hydrogen to reduce defect concentrations at interfaces. There is a patterning and etching of this layer to create bond pads. Following this is a sequence to form color filters and microlenses to provide color separation and increased responsivity.

To increase the responsivity, backside thinning through the use of silicon-on-insulator or physical and chemical etching to allow backside illumination [18]. This requires wafer bonding to a carrier wafer, a scheme to make electrical contact from the backside or the die edge and a process to allow alignment of backside processes such as metal light shield, color filters, and microlenses.

To work fully within the CMOS process technology flow, the process extension should not only include an enhanced process flow, but also the supporting physical design kit files that describe the enhanced process. These include updated design rules, device symbols, SPICE models, and layout-versus-schematic verification files.

2.6 Outlook for a Single Photon Process Technology

Image sensor process technology provides an approach to achieve commercial single photon image detectors: demonstrate that a pixel exists, understand what is critical and then look how these can be added as enhancements to a CMOS process flow.

Possible pixel device directions to achieve single photon counting through process enhancements include the following:

- Increased conversion gain by providing a floating diffusion with decreased capacitance.
- Refinement of the single photon avalanche detector to produce controlled multiplication with high yield.

These and other approaches will be discussed in the following chapters.

Process enhancements for a single photon detector will have similarities to those for an image sensor, but there are enough differences that the process flow should start from the basic logic process and not from the image sensor process. While there are similarities, for instance, both need high optical acceptance, good conversion gain, and low density of bright defects, there are critical performance parameters not required for the single photon pixel, including, for example:

- Linearity is essential for image sensors integrating tone scale, but is not needed when the pixel is detecting whether a single event happens. A single photon pixel may not require that the floating diffusion have linear capacitance or that the capacitance between pixels is well matched.
- Small size is essential for high-resolution photographic imaging, but not needed for applications with larger pixels providing higher responsivity.
- Well capacity need only be large enough to capture the photocarriers of a single photon event.

On the other hand, there are critical performance parameters required for a single photon pixel not required in an image sensor process flow as follows:

- Multiplication uniformity to make sure that pixels all have similar detection threshold.
- Almost 100% optical fill factor and quantum efficiency as every photon needs to be counted.
- High speed readout to allow image synthesis from multiple frames in a reasonable time.

In summary, the approach used for CIS process technology development can provide a model for single photon image sensors. Start with a standard process flow. Look to add as process enhancements the steps required to create the single photon pixel. In doing this, it is important to use the base process modules where possible and to add only process modules as are compatible with the base process flow so as to preserve their functionality consistent with their device models. The twin goals are

to create an optimized pixel and to preserve the CMOS performance that allows the reuse of the CMOS design flow.

References

1. P.J.W. Noble, Self-scanned silicon image detector arrays, IEEE Trans. Electron Devices **15**, 202–209 (1968)
2. A.J.P. Theuwissen, *Solid-State Imaging with Charge-Coupled Devices* (Kluwer Academic, Dordrecht, 1996)
3. O. Yadid-Pecht, R. Ginosar, Y. Shacham-Diamand, A random access photodiode array for intelligent image capture, IEEE Trans. Electron Devices **38**, 202, 1772–1780 (1991)
4. E.R. Fossum, CMOS image sensors: electronic camera-on-a-chip, IEEE Trans. Electron Devices **44**, 1689–1698 (1997)
5. M. Mitsuyoshi et al., 1/4-Inch 2-Mpixel MOS image sensor with 1.75 transistors/pixel, IEEE J. Solid-State Circuits **39**, 2426–2430 (2004)
6. J-H. Park et al., A high-speed low-noise CMOS image sensor with 13-b column-parallel single-ended cyclic ADCs, IEEE Trans. Electron Devices **56**, 2414–2422 (2009)
7. J. Hynecek, BCMD – an improved photosite structure for high density image sensors, IEEE Trans. Electron Devices **38**, 1011–1020 (1991)
8. J. Nishizawa et al., Static induction transistor image sensors. IEEE Trans. Electron Devices **26**, 1970–1977 (1979)
9. H-Y. Cheng et al., An integrating wide dynamic-range image sensor with a logarithmic response, IEEE Trans. Electron Devices **56**, 2423–2428 (2009)
10. C. Wolfe et al., *Physical Properties of Semiconductors* (Prentice Hall, Englewood Cliffs, NJ, 1989)
11. N. Richard, *Tauber, Silicon Processing for the VLSI Era*, vol. 3, The Sub-Micron MOSFET (Lattice Press, CA, 1995)
12. S.K. Mendis et al., CMOS active pixel image sensors for highly integrated imaging systems, IEEE J. Solid-State Circuits **32**, 187–197 (1997)
13. H. Takahashi et al., A 3.9-μm pixel pitch VGA format 10-b digital output CMOS image sensor with 1.5 transistor/pixel, IEEE J. Solid-State Circuits **39**, 2417–2425 (2004)
14. K. Itonaga et al., 0.9 μm Pitch pixel CMOS image sensor design methodology, IEDM Digest, 555 (2009)
15. E. Stevens et al., Low-crosstalk and low-dark-current CMOS image-sensor technology using a hole-based detector, Proc. IEEE ISSCC, 60 (2008)
16. R.N. Tauber, *Silicon Processing for the VLSI Era*, vol. 1, Process Technology, 2nd edn. (Lattice Press, CA, 2000)
17. R.N. Tauber, *Silicon Processing for the VLSI Era*, vol. 4, Deep Sub-Micron Process Technology (Lattice Press, CA, 2002)
18. P. Coudrain et al., Investigation of a sequential three-dimensional process for back-illuminated CMOS image sensors with miniaturized pixels, IEEE Trans. Electron Devices **56**, 2403–2413 (2009)

Chapter 3
Hybrid Avalanche Photodiode Array Imaging

Hiroaki Aihara

Abstract A hybrid avalanche photodiode (APD) array is a vacuum tube containing a photocathode and an array of avalanche photodiodes. It is a hybrid device that combines a traditional phototube technology and an advanced semiconductor technology. A photon produces a photoelectron with quantum efficiency at the photocathode. Unlike a phototube with dynodes, multiplication of the photoelectron is provided by a bombardment of the accelerated photoelectron into the avalanche photodiode resulting in a number of electron–hole pairs and a subsequent avalanche multiplication of the secondary electrons at the pn junction of the reverse-biased diode. The resulting total gain ranging from 10^4 to 10^5 is large enough to retain a single-photon sensitivity by using low-noise amplifiers. Segmentation of the pn junction of the diode provides the position information of an incident photoelectron and enables imaging of an incident photon. We report the recent progress on R&D of a single-pixel large format hybrid APD and a multipixel hybrid APD array.

3.1 Introduction

Although the concept of a hybrid avalanche photodiode (APD) was introduced quite a while ago [1, 2], it is only in recent years that a device that has a single-photon imaging capability becomes available. It is primarily because a hybrid APD requires an advanced semiconductor technology. A hybrid APD array is a vacuum tube consisting of a photocathode and an array of APDs.[1] It combines a traditional phototube technology and a semiconductor technology and therefore is called hybrid. A photon produces a photoelectron with quantum efficiency (typically 20–30% for visible light) at the photocathode. Unlike a phototube (PMT) with multiple dynodes, multiplication of the photoelectron is provided by two

[1] A hybrid photon detector consisting of a photocathode and a silicon pixel array not operated in avalanche mode has also been developed [3].

P. Seitz and A.J.P. Theuwissen (eds.), *Single-Photon Imaging*,
Springer Series in Optical Sciences 160, DOI 10.1007/978-3-642-18443-7_3,

and only two processes, a bombardment of the accelerated photoelectron into the APD resulting in a number of electron–hole pairs and a subsequent avalanche multiplication of the secondary electrons at the pn junction of the reverse-biased APD. The total gain ranging from 10^4 to 10^5 is large enough to retain a single-photon sensitivity by using a low-noise preamplifier. Segmenting the pn junction of the diode into pixels provides the position information of an incident photoelectron and, therefore, enables imaging of an incident photon.

The hybrid APD is simpler in mechanical construction and faster in signal processing than a PMT, but it has a smaller gain and requires a low-noise preamplifier. Because it utilizes a photocathode to convert photons to electrons, it is best suited when a large photosensitive coverage (with limited cost) is required despite the fact its quantum efficiency is lower than a photodiode on which a photon is directly injected.

3.2 Principle of Hybrid APD Operation

Figure 3.1 illustrates the principle of operation of a hybrid APD. A photoelectron originated from the photocathode is accelerated by the electric field created by applying a negative high voltage on the photocathode. The photoelectron directly

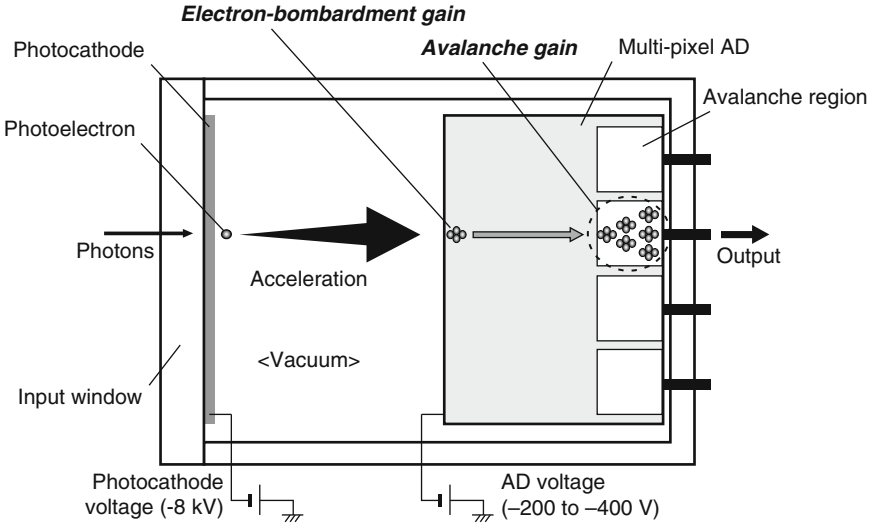

Fig. 3.1 Hybrid avalanche photodiode: a photon produces a photoelectron with quantum efficiency at the photocathode. Multiplication of the photoelectron is provided by two processes, a bombardment of the accelerated photoelectron into the avalanche photodiode resulting in a number of electron–hole pairs and an avalanche multiplication of the secondary electrons at the pn junction of the reverse-biased avalanche photodiode. Segmenting the pn junction of the diode into pixels provides the position information of an incident photoelectron and enables imaging of an incident photon

injects into the APD, in which the electron loses its energy by ionizing the silicon and creates electron–hole pairs. The number of electron–hole pairs is approximately given by the voltage difference between the photocathode and the APD divided by the specific ionization energy in silicon, which is ∼3.6 eV. If, for example, −8 kV is applied on the photocathode and the APD is placed on (or near) ground, the number of electron–hole pairs produced by the bombardment of an accelerated photoelectron is $8{,}000/3.6 \sim 2{,}000$. This electron bombardment (EB) multiplication is the first gain in the electron multiplication process in the hybrid APD and can be compared with the first gain (2–5) of the multidynode system of the traditional PMT. A large first gain (without amplifying noise) is the key to achieve good signal-to-noise ratio, similar to any multistage amplifier system.

Each secondary electron drifts toward the pn junction following the electric field created within the APD by a reverse bias voltage. At and in close proximity of the pn junction, where the electric field reaches $\sim 10^5$ V/cm, each secondary electron undergoes avalanche process and results in further multiplication of secondary electrons. The avalanche gain obtained in the currently available APD ranges from 10 to 100, yielding the total gain of hybrid APD of 10^4–10^5. This gain is about a factor of 10–100 smaller than that of a typical PMT. This smaller gain can, however, be offset by a low-noise preamplifier by taking advantage of the intrinsically low noise signal from the hybrid APD. Because the electron mobility[2] and the avalanche process within the semiconductor are very fast, careful design of photoelectron flight path from the photocathode to the surface of the APD yields very fast signal with practically negligible transit-time spread. A fast, single-photon sensitive device with large photocathode coverage can be realized.

3.3 Single-pixel Large Format Hybrid APD

3.3.1 Device Description

One of the applications of the hybrid APD is a single-pixel large format photodetector, shown in Fig. 3.2. A 13-inch diameter hybrid APD has been developed by Hamamatsu Photonics (HPK) [4]. Its photocathode has effective sensitive area of 300 mm in diameter. The voltage difference between the photocathode and the APD is designed to be a maximum of 20 kV to accelerate a photoelectron to produce ∼4,500 electron–hole pairs upon bombardment into APD. The size of APD is 5 mm in diameter and the electric field within the vacuum tube is designed to function as an electrostatic lens to focus photoelectrons on the APD. A further multiplication $(10 - 100)$ of the secondary electrons is provided by the avalanche process in the

[2]The carrier mobilities in silicon are 1,350 and 480 $\mathrm{cm}^2/(\mathrm{V\,s})$ for electrons and holes, respectively. Because the electron mobility is about three times larger, the electron signal is predominantly used in silicon devices.

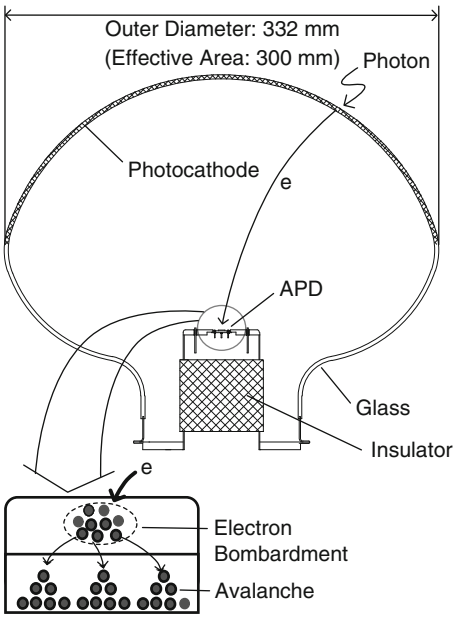

Fig. 3.2 Single-pixel large format hybrid APD and principle of operation. A 13-inch diameter hybrid APD has been developed by HPK. A photoelectron is accelerated by a maximum of 20 kV voltage difference and is bombarded into a 5 mm diameter APD. The bombardment gain is ∼4,500 at 20 kV and the additional gain of 10 − 100 is obtained by avalanche multiplication in the APD. The total gain of the device rages from 10^4 to 10^5

APD. Because of the long flight path of the photoelectron a single-pixel large format APD must be shielded from the magnetic field.

Figure 3.3 shows the cross section of a typical APD [5]. The pn junction is fabricated on the front surface and photoelectrons are irradiated from the back (or bulk-side) surface. In order for the secondary electrons generated by the bombardment of photoelectrons to reach the avalanche region before recombination, the substrate must be fully depleted. For the depletion layer of ∼100 μm thickness and a substrate with a high resistivity of ∼4 kΩcm, the reverse-bias voltage of ∼300 V is required. The maximum electric field at the pn junction can reach 3.8×10^5 V/cm for the p type impurity of $\sim 1.8 \times 10^{16}$ cm^3 of the deeply diffused p layer facing the n^+ layer.

This device has been realized by overcoming technical difficulties related to the photocathode surface activation process with alkali vapor. The hybrid APD in Fig. 3.2 works in a positive high voltage mode. The photocathode is connected to the ground and the APD is on the positive high voltage with the insulator to separate them. The output of APD is AC-coupled to a preamplifier via high-voltage decoupling capacitors. The alkali vapor could stain the surface of the insulator and cause the nonuniform surface electric field resulting in local discharge and flashing.

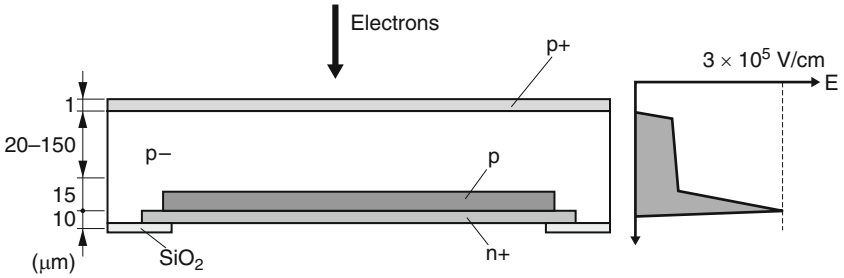

Fig. 3.3 Cross section of a typical APD. Here, p^+ (n^+) stands for a high p (n) type impurity density of the order of 10^{19} cm^{-3}, while p^- a low p type impurity density of the order of 10^{13} cm^{-3}

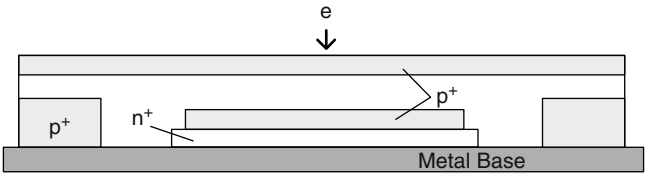

Fig. 3.4 Cross section of a back-illuminated APD, improved for the better isolation between p and n areas of the APD surface

Applying a high-resistivity coating on the surface is found to mitigate the problem. Another difficulty is the increase of the APD leakage current after activation process. The alkali vapor could contaminate the surface of APD. A back-illuminated APD shown in Fig. 3.4 has been developed to improve the isolation between p and n areas of the APD surface. This back-illuminated structure is also found to improve a long-term stability because the structure eliminates possible damage of p–n isolation due to electron bombardment.

3.3.2 Performance

Figure 3.5 shows EB gain in units of the number of electrons calculated from measurements of photocathode current and diode output current. A reverse-bias voltage on the diode during the measurement is kept at 30 V, which is low enough to disable avalanche multiplication. Above some threshold voltage, the EB gain is proportional to the electron energy equivalent to the applied HV between the photocathode and the diode. This threshold is due to energy lost by the photoelectron in the thin contact (and dead) layer of the diode surface. At 20 kV the EB gain was ∼4,500. The avalanche gain and the leakage current are shown in Fig. 3.6. The abrupt increase of the leakage current at ∼400 V indicates the breakdown of the diode. The maximum obtainable gain was ∼50 for a reverse-bias voltage of ∼390 V. The total gain of this hybrid APD is a product of the two gains and is $\sim 2.5 \times 10^5$.

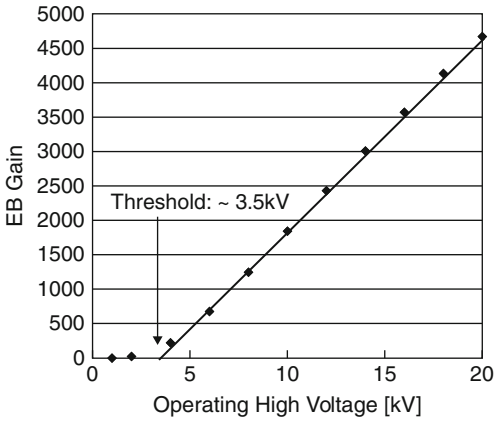

Fig. 3.5 Electron bombardment (EB) gain as a function of the high voltage applied between photocathode and APD

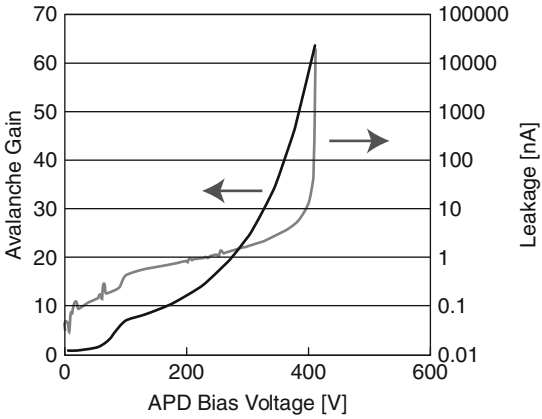

Fig. 3.6 Avalanche gain and leakage current (nA) as a function of the reverse bias voltage applied on APD

Figure 3.7 shows a pulse-height spectrum obtained with a picosecond light pulse with a wavelength of 441 nm [6]. A single photon peak is clearly seen. The signal-to-noise ratio for one photoelectron signal is ∼10. The signal is found to be linear up to ∼3,000 photoelectrons. Figure 3.8 shows a transit time spread (TTS) for single photon signals. It is ∼440 ps at FWHM (or $\sigma \sim 190$ ps), significantly better than PMT.

A hybrid APD is free from an afterpulse that is common to a high-gain PMT. The afterpulse is caused by an electron escaping the dynodes of PMT and subsequently ionizing the residual gas within the vacuum tube. If the ion hits the photocathode, the photocathode will emit an electron that will start delayed multiplication process, creating the afterpulse. Because an electron cannot escape from the APD, the hybrid APD will not suffer from an afterpulse. A hybrid APD could, however, suffer from increased frequency of dark noise, a signal current without incident photons. It is caused by a local discharge on the surface of the insulator that must stand a large HV, if the surface treatment is less than perfect. This HV insulation and the associated dark noise remain to be an issue for hybrid APDs.

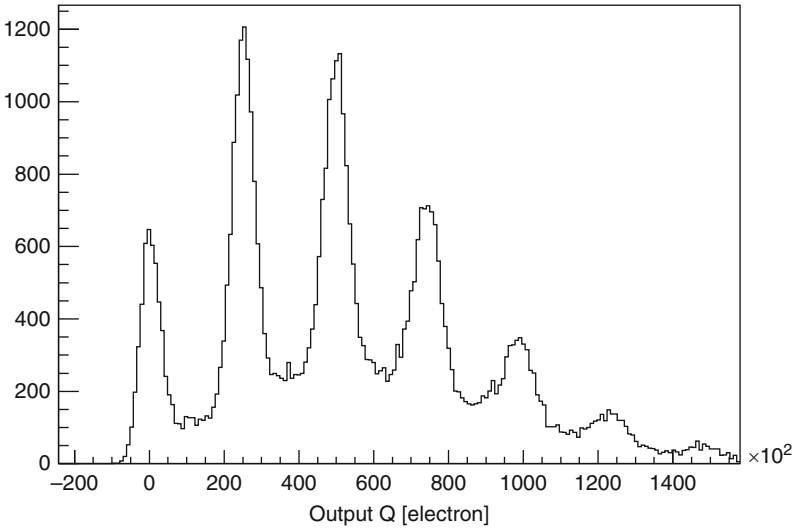

Fig. 3.7 Pulse height spectrum obtained with a single-pixel 13-inch hybrid APD. Single photon peak and peaks corresponding up to six photoelectrons are seen. A peak at around 0 electrons is a pedestal

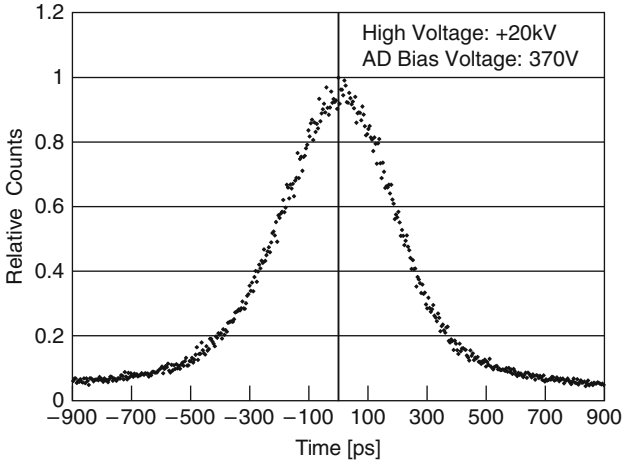

Fig. 3.8 Transit time spread for single-photon signals obtained with a single-pixel 13-inch hybrid APD. The spread is ∼440 ps at FWHM (or $\sigma \sim 190$ ps)

3.3.3 Application

A single-pixel large format hybrid APD finds its way to applications where a large photocathode coverage and an improved signal response, such as a single photon sensitivity and a fast signal response with minimal TTS, are required. Because a

hybrid APD is simpler in mechanical construction than PMT, its fabrication cost is expected to be significantly reduced when produced in mass. Therefore, hybrid APDs can be deployed in a large scale detector in which the total cost is driven by the cost of a photosensor. A large-format hybrid APD is considered as an alternative to a large-format PMT [7]. Super Kamiokande, a large underground water Cherenkov detector shown in Fig. 3.9, employs about 11,000 20-inch PMTs to detector feeble Cherenkov lights produced by neutrino-induced reactions and possible nucleon decays in water. The photocathode of PMTs covers 40% of the inner surface of the detector volume. Even a larger detector with a total mass of about one megaton has been envisioned for the next generation neutrino detector. There are currently at least three proposals to construct such a detector, Hyper Kamiokande in Japan, Water Cherenkov detectors at DUSEL in US, and MEMPHYS in Europe. The current design of Hyper Kamiokande, shown in Fig. 3.10, calls for two identical large caverns, each containing a cylindrical detector of 43 m in diameter and 250 m in length. Each detector holds more than 550k ton water and is to be instrumented with 50,000 20-inch Super-Kamiokande type PMTs. The total photocathode coverage is 20% of the water surface and the total cost of PMTs amounts to more than $350 M. A large-format hybrid APD is a viable alternative that is superior in performance and lower in cost. This large-scale deployment of hybrid APDs, if happened, would certainly boost the use of hybrid APDs in a variety of other applications.

3.4 Multipixel Hybrid APD Array

A multipixel hybrid APD array is a device in which the pn junction of APD is segmented into pixels and provides position information of the incident photon.

3.4.1 Device Description

In a multipixel APD [5] both the deeply diffused p layer and the n^+ layer of the APD are divided into pixels as shown in Fig. 3.11. Each pixel works as an individual

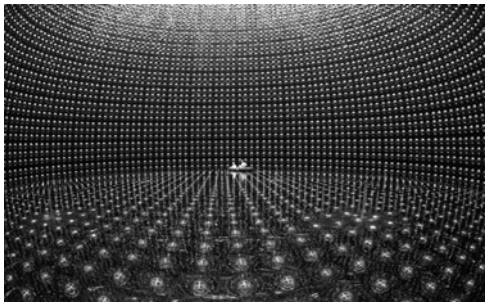

Fig. 3.9 The inside of Super Kamiokande. The tank is partially filled with water. There are about 11,000 20-inch PMTs covering 40% of the inner surface of the detector

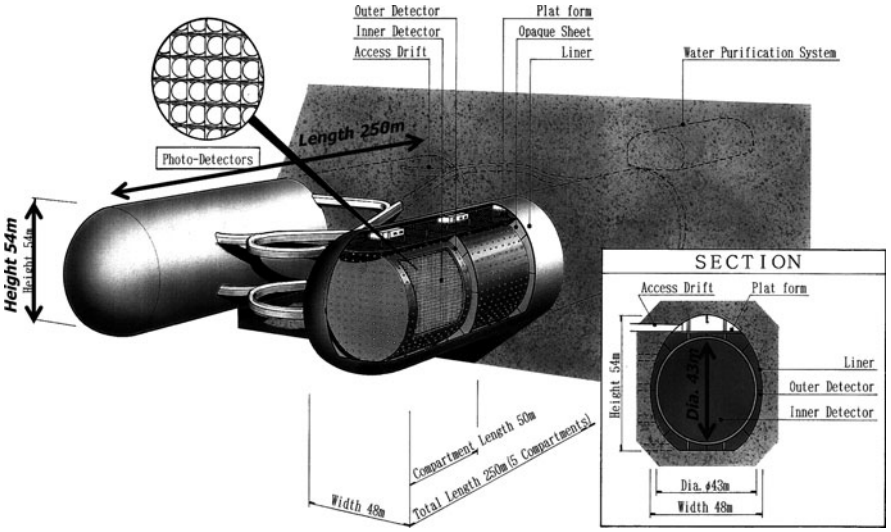

Fig. 3.10 Concepual design of Hyper Kamiokande. The total volume of ∼1 Mton is about 20 times larger than that of Super Kamiokande. It consists of two identical cylindrical detectors of 43 m in diameter and 250 m long. To cover 20% of the inner surface of the detector about 100,000 20-inch PMTs are needed

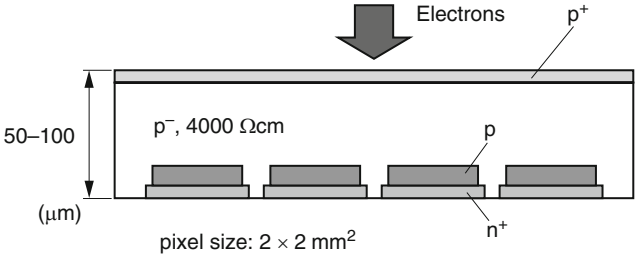

Fig. 3.11 Cross section of a mutipixel APD. Both the deeply diffused p layer and the n^+ layer of APD are divided into pixels

small APD. The multipixel APD is bump-bonded to a multilayer interface board, which is, in turn, mounted on the ceramic stem as shown in Fig. 3.12. A multipixel APD developed by HPK, shown in Fig. 3.13, has the effective area of $16 \times 16\,\mathrm{mm}^2$ segmented into 64 (8×8) pixels, each $2 \times 2\,\mathrm{mm}^2$ large. The multipixel APD inevitably has a low-gain region between neighboring pixels caused by an edge effect of the electric field between the pixels. The width of the low-gain region of this APD was found to be ∼150μm. The presence of the low-gain boundary region ultimately limits the segmentation of the APD, because the packing fraction (defined by the effective sensitive area divided by the total surface area) becomes too small. A pixel size of 0.5×0.5 mm can be achieved, resulting in a 1,024 (32×32) pixel APD with a total effective area of $16 \times 16\,\mathrm{mm}^2$.

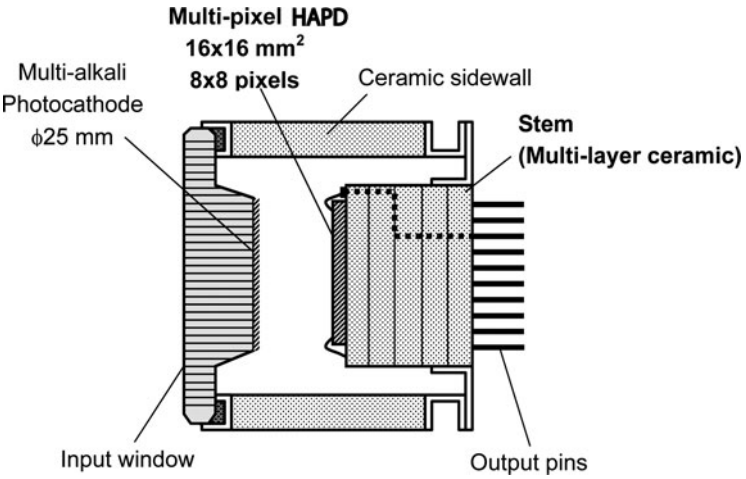

Fig. 3.12 Cross section of a multipixel hybrid APD. The APD is bump-bonded to a multilayer ceramic interface board, which is mounted on the ceramic stem. A multilayer ceramic board connects each bonding pad in the vacuum to the corresponding output pin in the atmosphere

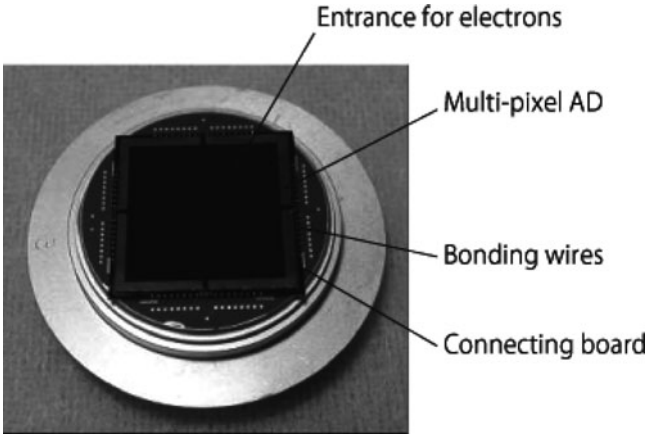

Fig. 3.13 A 64-pixel APD with the effective sensitive area of $16 \times 16\,\mathrm{mm}^2$ developed by HPK

In a proximity-focused hybrid APD shown in Fig. 3.14 the electric field between the photocathode and the APD is planar and uniform. In addition, the flight path of photoelectrons can be made very short. It has, therefore, a good immunity to the magnetic field as long as the magnetic field is parallel (or near parallel) to the electric field.[3] This is why a proximity-focused hybrid APD is often considered as a single-photon sensitive large photocathode sensor in strong magnetic field environment.

[3]Note that electrons drift in the direction parallel to $\boldsymbol{E} \times \boldsymbol{B}$.

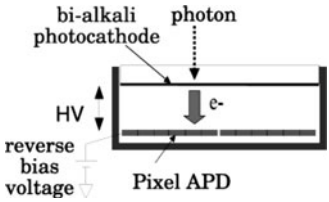

Fig. 3.14 Cross section of a proximity-focused hybrid APD array

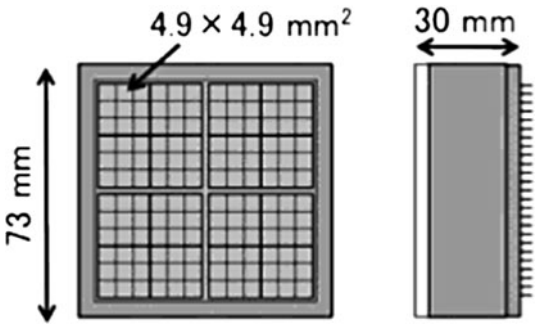

Fig. 3.15 Structure of a 144-pixel hybrid APD array. It contains four APDs, placed in a 2 × 2 array, each segmented to thirty six $4.9 \times 4.9\,\mathrm{mm}^2$ pixels. This 144-pixel hybrid APD array has the outer dimensions of $73 \times 73\,\mathrm{mm}^2$, ∼65% of which corresponds to the effective sensitive area

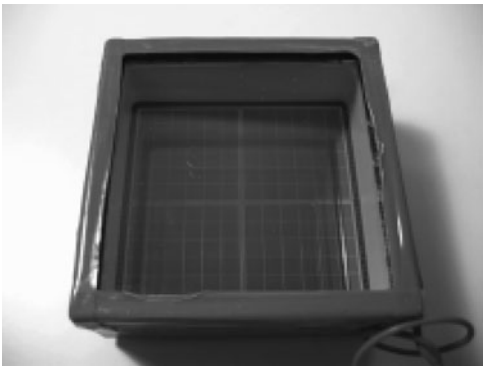

Fig. 3.16 A 144-channel hybrid APD array developed by HPK

Figures 3.15 and 3.16 show the structure of a large hybrid APD array being developed by HPK [8]. It contains four APDs, placed in a 2 × 2 array, each segmented to thirty six $4.9 \times 4.9\,\mathrm{mm}^2$ pixels. This 144-pixel hybrid APD array has the outer dimensions of $73 \times 73\,\mathrm{mm}^2$. Approximately 65% of the total surface area corresponds to the effective sensitive area. It operates in a negative HV mode, i.e., typically −8.5 kV is applied to the photocathode. The reverse-bias voltage applied to the APD is ∼300 V. The distance between the photocathode and the APD is only ∼30 mm. Each pixel has the detector capacitance of ∼80 pF. Each of 144-pixel channels is connected to a preamplifier, a shaper and a variable gain amplifier followed by a comparator for digitization and a shift register for a pipeline readout, all implemented on a single ASIC chip. The expected position resolution

of a multipixel hybrid APD array is given by a formula of $\sigma = \text{pixel width}/\sqrt{12}$. For the 144-pixel hybrid APD array, the pixel width is 4.9 mm and, therefore, the position resolution of $\sim$1.4 mm or better is expected.

3.4.2 Performance

The pulse height distribution of one channel in a prototype of the 144-pixel hybrid APD array for an LED light is shown in Fig. 3.17. A high voltage of -8.5 kV was applied to the photocathode and the reverse bias voltage of $\sim$330 V was applied to each of four APDs. A clear single-photoelectron peak is seen and the signal-to-noise ratio is calculated to be $\sim$13. A technical challenge of a hybrid APD array is to ensure the uniformity in gain and noise over the entire sensitive area. The uniformity of gain and signal-to-noise ratio of the 144-channel prototype was investigated. Figure 3.18 shows the total gain and the signal-to-noise ratio for each of the 36 channels in one of the four APD chips of the hybrid APD array. The nine channels located at the boundary adjacent to the neighboring chip and one broken channel are excluded in Fig. 3.18. The boundary channels suffer crosstalks from the neighboring APD. The remaining 26 channels have a similar gain ($\pm$15% or better) and signal-to-noise ratio ($>$10).

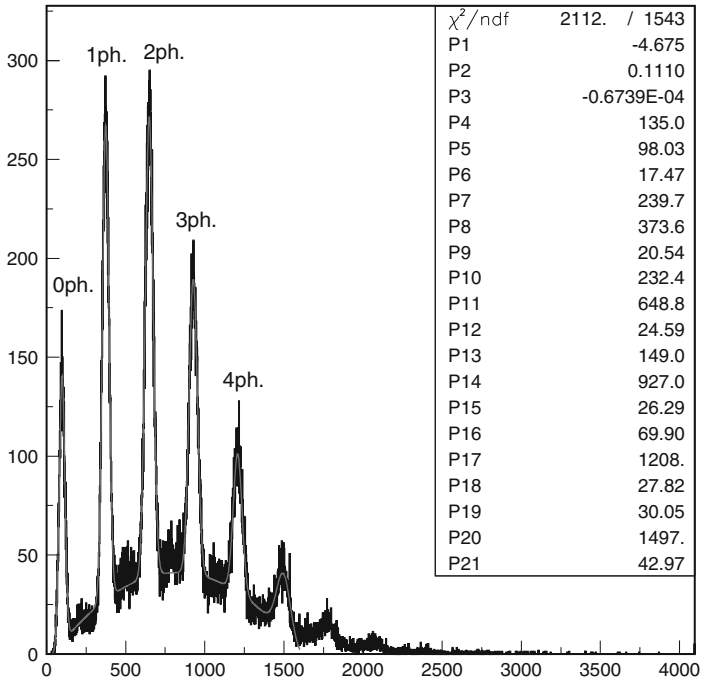

Fig. 3.17 Pulse height distribution of a single channel of the 144-channel hybrid APD array

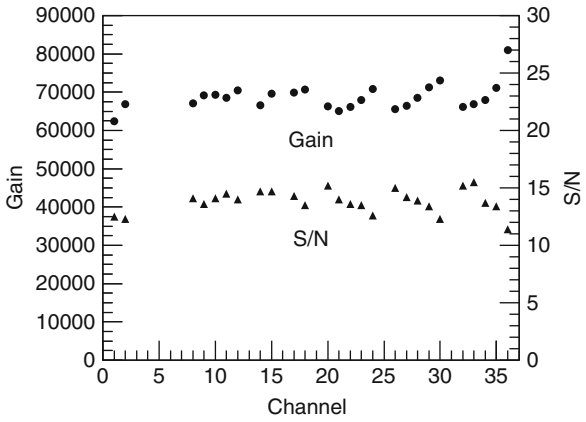

Fig. 3.18 Total gain and signal-to-noise ratio (S/N) for 26 channels in one of four APDs of the 144-channel hybrid APD array

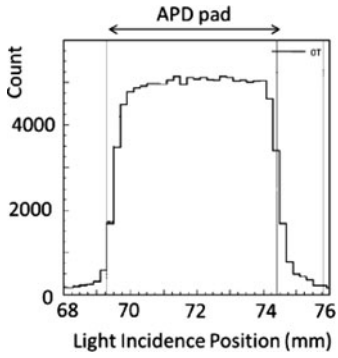

Fig. 3.19 Signal response obtained by scanning a focused pulsed laser light across a 4.9 mm wide pixel (or pad) of one of four APDs in the 144-channel hybrid APD array

Figure 3.19 shows a signal response obtained by scanning a focused pulsed laser light (with a spot size of ∼0.2 mm) across a 4.9 mm wide pixel (or pad) of the APD. The pixel shape is clearly reconstructed. A crosstalk to the 0.6 mm apart neighboring pixel is measured to be less than 4%.

3.4.3 Application

A ring imaging Cherenkov detector (RICH) has been used as a particle identification device in many high-energy experiments. Recently, Belle experiment at the KEKB collider considers a RICH with aerogel radiator, shown in Fig. 3.20, in the forward region where the space is limited and the strong magnetic field is present [9]. The photosensor of this RICH must be position-sensitive with a granularity of $\sim 5 \times 5\,\text{mm}^2$. It must have large effective area to collect as many photons as possible. It needs to be immune to 1.5 T magnetic field perpendicular to the photon detector plane. In addition, it must cover the total of $4\,\text{m}^2$, resulting in a total number of

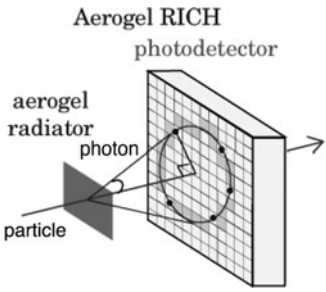

Fig. 3.20 Concept of a ring imaging cherenkov detector (RICH) with aerogel radiator

segmentations (or channels) of $\sim 10^5$. The 144-channel hybrid APD array has been developed by HPK for this application.

3.5 Conclusions and Remaining Issues

A hybrid APD array offers excellent solutions for applications where a single photon sensitivity, a fast signal response with small to negligible transit time spread, and a wide area photocathode coverage are required. Its mechanical construction is simple and the prospect for low-cost mass-production is promising.

We note that a high-gain, low-noise electronics is key for successful deployment of a hybrid APD array (either single-pixel or mutipixel). In particular, when a total number of channels required is as large as 10^5 it is necessary to fabricate an ASIC that meets the requirements for the APD array readout. The ASIC typically includes a preamplifier, a shaper, and a digitization circuit often with pipeline readout capability [8,10]. Taking advantage of advanced semiconductor and integrated circuit technologies, a hybrid APD array is becoming a viable alternative to traditional PMTs.

References

1. N. Sclar, Electron Device Conference, Washington DC, October, 1957
2. R. DeSalvo, Nucl. Instrum. Meth. A **387**, 92 (1997)
3. M. Moritz et al., IEEE Trans. Nucl. Sci. **NS-51**, 1060 (2004)
4. Y. Kawai et al., Nucl. Instrum. Meth. A **579**, 42 (2007)
5. M. Suyama, Ph.D. Dissertation, The Graduate University for Advanced Studies, Japan, KEK report 2002-16, 2003
6. H. Nakayama et al., Nucl. Instrum. Meth. A **567**, 172 (2006)
7. H. Aihara, in *Proceedings of Workshop for European Strategy for Future Neutrino Physics*, CERN, 1-3 October 2009, http://indico.cern.ch/getFile.py/access?contribId=33&sessionId=5&resId=0&materialId=0&confId=69984
8. S. Nishida et al., Nucl. Instrum. Meth. A **595**, 150 (2008)
9. T. Abe et al., Belle II Technical Design Report, KEK report 2010-1, 2010
10. T. Abe et al., Nucl. Instrum. Meth. A **623**, 279 (2010)

Chapter 4
Electron Bombarded Semiconductor Image Sensors

Verle Aebi and Kenneth Costello

Abstract The low noise electron bombarded semiconductor gain process is now enabling new classes of single photon sensitive, photocathode based, image sensors. These imagers form two broad classes of devices: low pixel count imagers for high temporal bandwidth photon counting applications and high pixel count imagers for single photon sensitive staring applications. The first class of devices has demonstrated imaging photon counting imagers operating at bandwidths on the order of 1 GHz and able to distinguish multiple photon events. In the case of the second class of devices, single photon sensitivity in a megapixel format is obtained by integrating modern CMOS image sensors with a photocathode in an electron bombarded configuration. An overview of both classes of devices is presented in this chapter and is shown to approach the characteristics desired in a perfect detector having 100% quantum efficiency, infinite gain and bandwidth, and no excess gain noise.

4.1 Introduction

Single-photon detection has been possible for many years using electron multiplication in vacuum as described in detail in Chap. 5 of this volume and references therein. In this approach to single-photon detection, a photocathode is used to generate photoelectrons that, after emission into vacuum, are accelerated to high energy by an applied voltage and impact an electron multiplier that provides relatively low-noise amplification. This low-noise amplification increases the single-photon output signal to a level sufficiently above the system electronics noise floor to enable detection of single photons. Electron multipliers widely used in the past are the dynode chains in photomultiplier tubes or a microchannel plate as used in an image intensifier tube [1, 2].

More recently, photon detectors have been developed where the electron multiplication results from direct electron bombardment of a semiconductor anode by a photoelectron emitted by the photocathode. The photoelectron is accelerated to high

P. Seitz and A.J.P. Theuwissen (eds.), *Single-Photon Imaging*,
Springer Series in Optical Sciences 160, DOI 10.1007/978-3-642-18443-7_4,

energy by the voltage applied between the photocathode and anode of the device and gain is achieved by generation of electron–hole pairs as this energy is dissipated in the semiconductor. The electron bombarded semiconductor (EBS) gain process has the advantage of even lower excess noise factor than that obtained with a dynode chain or with a microchannel plate [3, 4]. The EBS gain process is also very linear compared to previous electron multipliers enabling very high dynamic range.

Two broad classes of single or low photon number image sensors have been realized based on use of the EBS gain process. Members of the first class are termed hybrid photomultipliers or photodetectors (HPDs) or intensified photodiodes (IPDs). These devices readout all pixels in parallel and maintain the photon time of arrival information as well as the spatial location of the photon. The second class of EBS image sensors consists of electron bombarded CCD (EBCCD) or electron bombarded CMOS (EBCMOS) image sensors. This class of image sensor does not preserve the photon time of arrival, but only outputs the integrated signal on a per pixel per frame basis.

Any type of photocathode can be utilized with EBS image sensors. Classical positive affinity photocathodes have been used for many years in a variety of photomultiplier tubes (PMT) and image intensifier devices [5]. More recently, modern negative electron affinity (NEA), III–V semiconductor, photocathodes have been developed to enable high quantum efficiency in the visible and near infra-red spectral bands [6]. Typical NEA photocathode quantum efficiency curves obtained with GaAs and GaAsP III–V semiconductor materials are illustrated in Fig. 4.1

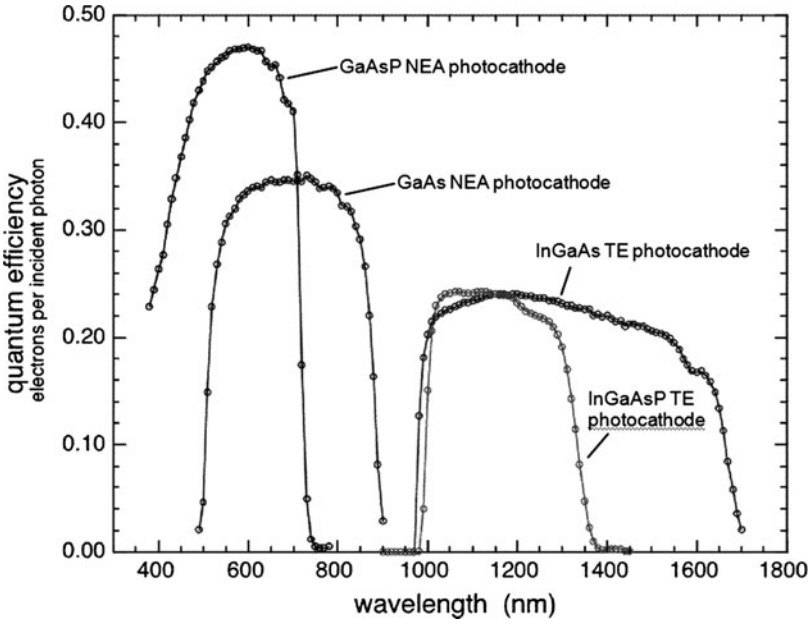

Fig. 4.1 Quantum efficiency curves for GaAs and GaAsP NEA photocathodes and InGaAs and InGaAsP TE photocathodes as used in EBS imagers

[2, 7]. Figure 4.1 also illustrates quantum efficiency achieved with transferred electron (TE) photocathodes based on various alloys of InGaAs and InGaAsP lattice matched to InP in the 950–1,700 nm spectral range [8]. Much of the work on EBS image sensors has focused on the use of NEA and TE photocathodes to obtain the highest quantum efficiency and photon detection performance.

Both classes of EBS image sensors have important applications. The HPD class of detectors have demonstrated low pixel counts (on the order of 10×10 pixel arrays), but have very good timing resolution, low after pulsing, high dynamic range, and the ability to discriminate multiple photons. The EBCCD and EBCMOS image sensors have demonstrated pixel counts as high as 2 megapixel. This has enabled very high-resolution, single, and low photon number imaging.

4.2 Electron Bombarded Semiconductor Gain Process

EBS gain occurs when a high energy electron impacts a semiconductor, where the electron energy is dissipated by a series of scattering collisions. Electron–hole pairs are generated in this process. The generated electrons follow random paths in the semiconductor with some electrons backscattering out of the semiconductor. The electrons can backscatter both elastically and inelastically from the semiconductor target. The backscatter coefficient increases with average atomic number of the target and decreases with increased landing energy [9]. Approximately 30% of 2 keV electrons are backscattered from a silicon target with approximately 85% of the total electron energy deposited in the target [10, 11].

The EBS gain process can be characterized by the average energy required to generate an electron–hole pair upon electron bombardment of a semiconductor target. Typical values are 3.64 eV for silicon and 4.4 eV for GaAs [6, 12]. The generated electrons are then collected by a reverse biased diode for an IPD or EBCMOS device or in a potential well for a CCD. Generated electrons are lost by recombination with holes in the bulk or at the backside surface. Efficient passivation of the backside surface is critical to achieving high EBS gain at low landing energies for a backside thinned CMOS or CCD image sensor.

A typical EBS gain versus electron landing energy curve for a backside bombarded CCD is presented in the paper by Aebi et al. and similar results are presented in the paper by Suyama et al. for a HPD imaging array [13, 14]. EBS gain of 180 at 2 keV electron landing energy is obtained by Aebi and an EBS gain of 2,100 at 9 keV electron landing energy is reported by Suyama. More recently, Nikzad has achieved EBS gains as high as 50 at an electron landing voltage of 600 V through MBE deposition of a delta-doped boron passivation layer on backside thinned CCD devices [15].

The EBS gain process is very low noise as the gain mechanism is essentially deterministic and no additional carrier multiplication occurs for the secondary electrons and holes generated by the EBS process. The noise statistics of this process is characterized by the Fano factor, F [16]. The noise fluctuations are found to

be less than would be expected for a gain process governed by Poisson statistics. The variance for an EBS gain process with average single photoelectron gain G_e is given by:

$$s_e = G_e F.$$

For silicon, the Fano factor is approximately 0.115 and for GaAs 0.10, both substantially smaller than a value of 1 that would result for a purely Poisson process [17].

4.3 Hybrid Photomultiplier EBS Image Sensors

Hybrid photomultiplier detectors (HPDs and IPDs) were first demonstrated using an electron bombarded silicon PIN diode in the 1960s [18]. More recently, the electron bombarded PIN or Schottky diode has been replaced with an avalanche photodiode (APD) operated in the linear gain regime to enable higher overall sensor gain without increasing the photocathode-to-anode bias voltage [19]. These devices are close to meeting the definition of a perfect detector: high quantum efficiency; high single photoelectron gain; very low excess noise factor; and high bandwidth with large associated active area.

4.3.1 Hybrid Photomultiplier Gain and Noise Analysis

Schematically the IPD can be thought of as a device where the photon absorbing layer is separated from the gain region of the device by a vacuum gap as opposed to an APD, where the optical absorbing semiconductor layer is in direct contact with the high field, avalanche gain, region of the device. Figure 4.2 illustrates the IPD structure.

The large excess noise factor of the APD gain stage has minimal impact on the excess noise factor of the IPD due to the high gain and extremely low excess noise factor of the EBS gain process. The combined noise factor, K_f, of this two-stage gain process can be evaluated using Friis's formula:

$$K_f^{IPD} = K_f^e + \frac{K_f^{apd} - 1}{G_e},$$

where K_f^e is the EBS noise factor, K_f^{apd} is the APD noise factor, and G_e is the electron bombarded gain. For an IPD incorporating a GaAs APD anode, the electron bombarded gain and noise factor for a detected photoelectron have been determined to be 1×10^3 and 1.003 at 8 kV photocathode-to-anode differential voltage [20]. Assuming the GaAs APD has a noise factor equivalent to a gain, M, of 12; the total IPD gain, G, and noise factor are:

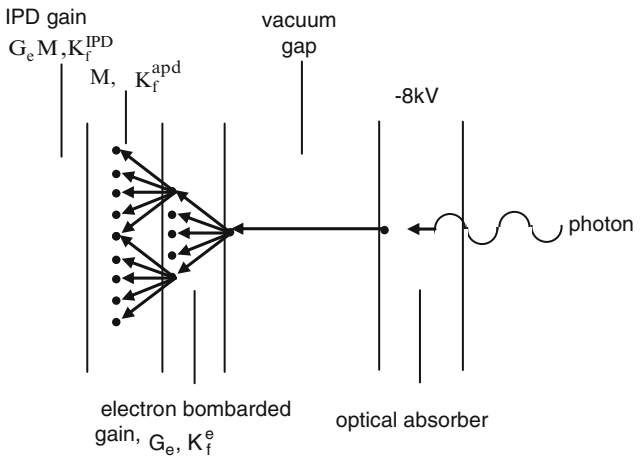

Fig. 4.2 Output gain, G_eM, and noise factor, K_f^{IPD}, are graphically represented for the IPD. The gain is derived from a first-stage electron bombarded gain followed by a second-stage APD avalanche gain

$$G = G_eM = 1.2 \times 10^4,$$
$$K_f^{IPD} = 1.014.$$

Single photoelectron gain and associated excess noise factor is sufficient to achieve single-photon detection at high electrical bandwidth. LaRue has conducted a detailed noise analysis of the IPD [20]. Figure 4.3 is an example of multiple photon discrimination achieved with an IPD at 1,300 nm utilizing an InGaAsP TE photocathode.

4.3.2 Hybrid Photomultiplier Time Response

IPD and HPD detectors have also demonstrated high bandwidth and excellent timing characteristics. An analysis of the impulse response for the IPD has been conducted by LaRue to optimize frequency response as a function of the EBS diode design parameters [20, 21]. High bandwidth has been demonstrated for the InGaAsP TE photocathode IPD with single-photon response rise time of 200 ps and associated pulse width of 500 ps with timing jitter of 70 ps [22]. Timing resolution of $<$100 ps has also been demonstrated for a high-quantum efficiency GaAsP NEA photocathode HPD [14].

4.3.3 Hybrid Photomultiplier Imagers

Imaging electron optics can be used in a hybrid PMT to obtain many parallel channels with performance similar to that of a single channel device. Early work

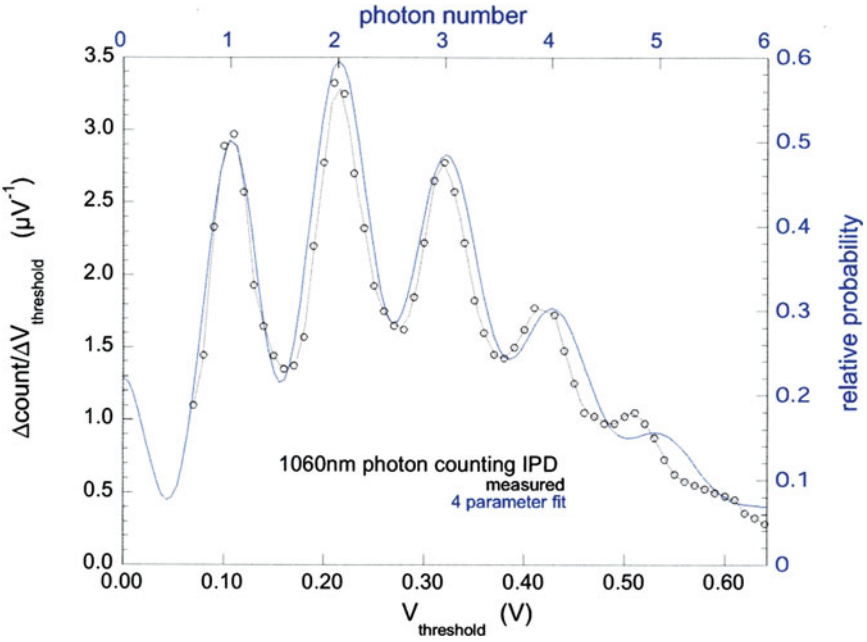

Fig. 4.3 This multiple photon pulse height distribution was measured at room temperature using a 1,300 nm optical impulse source. Discrimination of up to six photons is evident. The peaks of the envelope follow Poisson statistics

in this area led to the development of the Digicon [23]. The Digicon utilized magnetically focused electron optics with a 38-pixel linear silicon diode anode array. A proximity focused electron optic 8×8 APD diode array HPD has been more recently demonstrated [14]. In this configuration proximity focus electron optics were used as in an image intensifier tube [2]. Advantages of the proximity focus design include ability to operate in magnetic fields up to 1.5 T. Electrostatically focused IPDs have been designed and built that utilize imaging electron optics with a crossover and an electron bombarded APD anode array.

The SIMION 7.0 electron optics simulation of the electrostatically focused IPD imaging design is shown in Fig. 4.4. This design utilizes two simple baffles of identical aperture placed midway between the TE photocathode and the APD anode array. The aperture diameter and position with respect to the photocathode are optimized, so that the point-spread function at the anode is minimized in diameter. The electron trajectories between the cathode and anode show that there is a crossover between the focusing baffles so that the image is inverting. In this design, the magnification, M, is 0.63; however, increasing the distance between the anode and the last focusing baffle can easily change the magnification.

With this magnification, a $100\,\mu m^2$ APD anode pixel is mapped back to the photocathode creating a 100/0.63 or $158\,\mu m^2$ active cathode pixel. An important

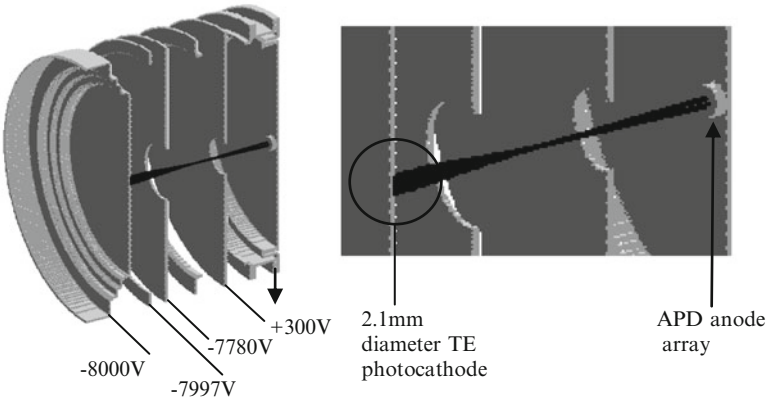

Fig. 4.4 Cross-section of imaging IPD design showing electron trajectories. Positive ions generated at the anode by impinging electrons are prevented from bombarding the photocathode by the ion trap electrode that is biased at +300 V

feature of this design is that the second aperture also operates as an ion trap. With the application of +300 V to this electrode, positive ions generated at the anode by impinging electrons from the cathode are repelled before they can move back through this aperture and damage the photocathode. This eliminates after-pulsing further improving the IPD noise factor.

An additional feature of this imaging design is that a smaller area APD anode pixel can be used relative to pixel size at the photocathode to increase bandwidth beyond 1 GHz. In this design, the APD anode geometry is always mapped back to the photocathode regardless of size as long as the photocathode is sufficiently large to intercept the mapping. In this manner, bandwidth is improved for a large pixel size relative to an all solid-state detector technology, where parasitic capacitance increases in proportion to pixel area. In the case of the IPD, the photocathode pixel size does not contribute to the parasitic capacitance of the anode.

4.4 EBCCD and EBCMOS EBS Image Sensors

One of the initial approaches demonstrated for electron bombarded silicon imagers was based on the use of CCDs and EBCCD image sensors were demonstrated shortly after the invention of the CCD [24, 25]. A schematic of a proximity focused electron optic design for this type of sensor is shown in Fig. 4.5. In these devices, the CCD is backside thinned to enable direct electron bombardment of the backside illuminated CCD by the photoelectrons emitted from the photocathode. A back illuminated configuration is required as the photoelectron range at reasonable bombardment energies is short and will not penetrate the typical polysilicon gate

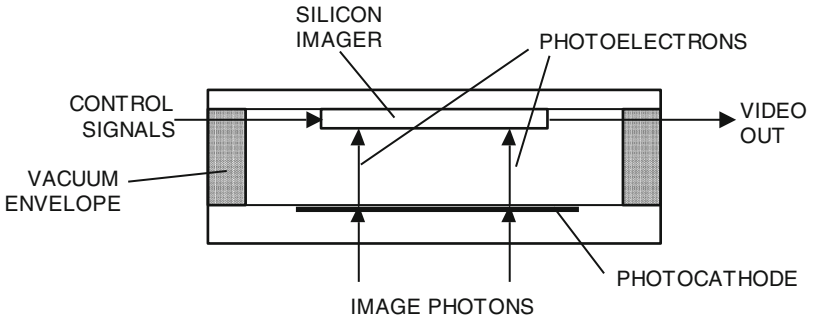

Fig. 4.5 Schematic cross-section of an EBCCD or EBCMOS image sensor

structure on the front side of a CCD [11]. The EBS gain process results in a low excess noise factor similar to that observed for the HPD or IPD and enables photocathode referred noise floors of less than one photoelectron. The generated electrons are collected in the CCD potential wells for subsequent readout. Significant advantages have been demonstrated for low light level imaging applications with the EBCCD configuration relative to other low light level imaging technologies such as EMCCDs or image intensified cameras [11, 13, 15, 26–28].

Recently, CMOS image sensors have gained acceptance for a wide variety of image sensor applications. CMOS image sensors have advantages over CCD image sensors as they are compatible with mainstream CMOS semiconductor fabrication technology and do not require dedicated semiconductor processing lines as is the case for CCDs. CMOS image sensors also enable integration of higher levels of on-chip functionality relative to a CCD image sensor. Typical capabilities integrated on a CMOS image sensor include digital sequencers to enable readout of arbitrary regions of interest, on-chip analog-to-digital converters, noise cancellation circuits, and circuitry to increased dynamic range [29]. These new capabilities have been combined with the low-noise EBS gain process in the EBCMOS image sensor [30, 31].

Ultimate performance of an EBCMOS sensor will be determined to a large extent by the CMOS imager architecture and design. First, it is essential that the CMOS imaging area has 100% fill factor (no dead area). Any reduction in active area will result in lost photoelectrons. This is equivalent to a reduction in photocathode quantum efficiency or sensitivity. At the lowest light levels, low light level camera performance is dictated by photon statistics. It is essential for the imager to detect the maximum number of photons for adequate low light level resolution and performance. Second, the CMOS imager architecture must maximize integration of the image photons with close to 100% duty cycle. This requirement when combined with high fill factor enables the collected signal to be maximized for good low light level performance.

One hundred percent fill factor can be achieved in a CMOS imager with a properly designed backside illuminated pixel design. A standard CMOS imager

cannot be used in a frontside illuminated electron bombarded mode since the metal and dielectric stack (typically 4–5 μm thick for a modern submicron CMOS process) will block the electrons from reaching the silicon at moderate acceleration voltages (2 kV typical). In addition, fill factor would be restricted to the optical fill factor of the CMOS imager pixel, which can be less than 50%. In a backside thinned CMOS imager, the chip is flip-chip bonded onto a carrier substrate and the silicon substrate is removed by mechanical and chemical thinning. The free silicon surface is then passivated to reduce carrier recombination at the surface. A properly designed pixel will allow a majority of the generated charge to be collected by the photodiode in the pixel regardless of photoelectron impact position in the pixel. This enables high single photoelectron signal-to-noise ratio (SNR) to be obtained with 2 keV electron energy EB gains on the order of 200. SNRs above one are achieved for a single photoelectron if the on-chip electronics noise (pixel referenced) is less than the EB gain. This is easily the case with modern CMOS image sensor technology, where total temporal noise floors of <10 electrons r.m.s. are possible.

An electronic shutter approach to enable integration of the image signal at close to a 100% duty cycle is implemented in the CMOS image chip. The above approach of a properly designed backside thinned CMOS imager with a high duty cycle shutter architecture enables essentially full utilization of the available signal from the photocathode. This is essential for high performance low light level imaging, where the ultimate performance is determined by the photocathode quantum efficiency and the associated signal limited shot noise.

Acknowledgements We thank the support and collaboration of many coworkers over the years at Varian, Inc., Stanford University, and Intevac; without their support, this work would not have been possible. We are especially grateful to the insights offered by R.L. Bell, W.E. Spicer, and R.L. LaRue.

References

1. R.W. Engstrom, *Photomultiplier Handbook* (RCA Corporation, New York, 1980)
2. E.J. Bender, Present Image-Intensifier Tube Structures, *Electro-Optical Imaging: System Performance and Modelling* (SPIE Press, Bellingham, Washington, 2000), pp. 5-1–5-96
3. H. Melchior, M.B. Fisher, F.R. Arams, Photodetectors for optical communication systems, Proc. IEEE **58**(10), 1466–1486 (1970)
4. R.L. Bell, Noise figure of the MCP image intensifier tube, IEEE Trans. Electron Devices **22**, 821–829 (1975)
5. A.H. Sommer, *Photoemissive Materials: Preparation, Properties, and Uses* (Robert E. Krieger Publishing Company, Huntington, New York, 1980)
6. R.L. Bell, *Negative Electron Affinity Devices* (Clarendon, Oxford, 1973)
7. J.P. Edgecumbe, V.W. Aebi, G.A. Davis, A GaAsP photocathode with 40% QE at 550 nm, Electron Tubes Image Intensifiers (SPIE, San Jose, CA, 1992), vol. 1655, pp. 204–211
8. K.A. Costello, V.W. Aebi, G.A. Davis, R.A. LaRue, R.E. Weiss, Transferred electron photocathode with *greater than 20% quantum efficiency beyond 1 micron*, Photodetectors and Power Meters II SPIE **2550**, 177–188 (1995)

9. K.F.J. Heinrich, *Electron Beam X-Ray Microanalysis* (Von Nostrand Reinhold, New York, 1981)
10. E.H. Darlington, V.E. Coslett, Backscattering of 0.5–10 keV electrons from solid targets, J. Phys. D Appl. Phys. **5**, 1969–1980 (1972)
11. M.K. Ravel, A.L. Reinheimer, Backside-thinned CCDs for keV electron detection, Charge-Coupled Devices Solid State Opt. Sensors II SPIE **1447**, 109–122 (1991)
12. J.R. Fiebiger, R.S. Muller, Pair-production energies in silicon and germanium bombarded with low energy electrons, J. Appl. Phys. **43**, 3202–3207 (1972)
13. V.W. Aebi, K.A. Costello, J.P. Edgecumbe, J.J. Boyle, W.L. Robbins, R. Bell, D. Burt, A. Harris, I. Palmer, P. Pool, Gallium arsenide electron bombarded CCD technology, Image Intensifiers and Applications; and Characteristics and Consequences of Space Debris and Near-Earth Objects SPIE **3434**, 37–44 (1998)
14. M. Suyama, A. Fukasawa, J. Haba, T. Iijima, S. Iwata, M. Sakuda, T. Sumiyoshi, F. Takasaki, M. Tanaka, T. Tsuboyama, Y. Yamada, Development of a multi-pixel hyprid photodetector with high quantum efficiency and gain, IEEE Trans. Nucl. Sci. **51**(3), 1056–1059 (2004)
15. S. Nikzad, Q. Yu, A.L. Smith, T.J. Jones, T.A. Tombrello, S.T. Elliot, Direct detection and imaging of low-energy electrons with delta-doped charge-coupled devices, Appl. Phys. Lett. **73**(23), 3417–3419 (1998)
16. U. Fano, Ionization yield of radiations. II. The fluctuations of the number of ions, Phys. Rev. **72**, 26–29 (1947)
17. R. Alig, S. Bloom, C. Struck, Scattering by ionization and phonon emission in semiconductors, Phys. Rev. B **22**, 5565–5582 (1980)
18. R. Kalibjian, A phototube using a diode as a multiplier element, IEEE Trans. Nucl. Sci. **NS-12**, 367–369 (1965)
19. P. Cushman, R. Russack, A photomultiplier tube incorporating an avalanche photodiode, Nucl. Instrum. Methods Phys. Res. A **333**, 381–390 (1993)
20. R.A. LaRue, K.A. Costello, G.A. Davis, J.P. Edgecumbe, V.W. Aebi, Photon counting III–V hybrid photomultipliers using transmission mode photocathodes, IEEE Trans. Electron Devices **44**, 672–678 (1997)
21. R.A. LaRue, J.P. Edgecumbe, G.A. Davis, S. Gospe, V. Aebi, High quantum efficiency photomultiplier with fast time response, Int. Symp. Optics, Imaging and Instrumentation, Photocathodes and Power Meters SPIE **2022**, 64–73 (1993)
22. W.H. Farr, J. Gin, D. Nguyen, GigaHertz bandwidth photon counting, SPIE **7320** (2009)
23. E.A. Beaver, C.E. McIlwain, A digital multichannel photometer, Rev. Sci. Instrum. **42**, 1321–1324 (1971)
24. S.R. Shortes, W.W. Chan, W.C. Rhines, J.B. Barton, D.R. Collins, Development of a thinned, backside-illuminated charge-coupled device imager, IEDM Tech. Dig. 415 (1973)
25. G.A. Antcliffe, C.G. Roberts, J.B. Barton, K.K. Pharr, D.W. Mueller, D.R. Collins, Operation of CCD's with stationary and moving electron-beam input, IEEE Trans. Electron Devices **22**, 857–861 (1975)
26. G.M. Williams, A.L. Reinheimer, V.W. Aebi, K.A. Costello, Electron-bombarded back-illuminated CCD sensors for low light level imaging applications, SPIE **2415**, 211–235 (1995)
27. M. Suyama, A. Kageyama, I. Mizuno, K. Kinoshita, M. Muramatsu, K. Yamamoto, An electron bombardment CCD tube, SPIE **3173**, 422–429 (1997)
28. C.B. Johnson, Review of electron bombarded CCD cameras, SPIE **3434**, 45–53 (1998)
29. E.R. Fossum, CMOS image sensors: electronic camera-on-a-chip, IEEE Trans. Electron Devices **44**, 1689–1698 (1997)
30. V.W. Aebi, J.J. Boyle, Electron bombarded active pixel sensor, US Patent No. 6,285,018 B1 (2001)
31. V.W. Aebi, K.A. Costello, P.W. Arcuni, P. Genis, S.J. Gustafson, EBAPS®: next generation, low power, digital night vision, Presented at the OPTRO 2005 International Symposium, Paris, France, 10 May 2005 (European Space Agency ISSN 1022–6656, 2005)

Chapter 5
Single-Photon Imaging Using Electron Multiplication in Vacuum

Gert Nützel

Abstract Some decades ago single photon detection and imaging used to be the terrain of photocathode based devices only. The most prominent devices of this family are photomultiplier tubes (PMT) and image intensifiers. During the past decades these devices lost ground to advanced solid state devices like APDs, low noise CMOS imagers, and EMCCDs.

Can one expect solid state devices to totally replace vacuum based ones like digital photography replaced the film based one?

To answer this question, a deeper understanding of technologies as well as the way the key performance parameters are defined and measured is required. This chapter describes the working principles of photocathodes in general, image intensifiers, and PMTs. It also compares performance characteristics of state-of-the-art state solid state detectors with vacuum based devices. This comparison is not always obvious because of the differences in measurements and definitions employed. Therefore, this chapter also offers the key to compare the performance equally.

5.1 Introduction

Some decades ago, single-photon detection was the domain of devices based on photocathodes and vacuum transport of electrons only. Image intensifiers (I2) were the standard device for imaging applications and the photomultiplier tube (PMT) was the standard device for timing applications (Figs. 5.1 and 5.2).

Despite the impressive development of semiconductor technologies, vacuum devices, mainly image intensifiers and PMTs, still play an important role in sectors where single photons are imaged or detected. This chapter provides an overview of this technology, including a performance comparison with state-of-the-art solid-state sensors.

Will the soldier of the future use low-noise complementary metal oxide semiconductor (CMOS) sensors to see in the night? Will future PET systems use avalanche

P. Seitz and A.J.P. Theuwissen (eds.), *Single-Photon Imaging*,
Springer Series in Optical Sciences 160, DOI 10.1007/978-3-642-18443-7_5,

Fig. 5.1 An image intensifier with integrated power supply

Fig. 5.2 An example of a photomultiplier tube (PMT)

photodetectors (APDs) for their detectors? To compare these different technologies, the system architect has to bridge two differences.

The first difference is in the apple-to-apple category. The terminology and standards of the two worlds do differ. To describe temporal noise, the CMOS community uses terms such as quantum efficiency, dark current density, and read noise, whereas scientists working with I2 refer to equivalent background illumination (EBI), signal-to-noise ratio (SNR) (measured according to Mil-Specs), and cathode sensitivity. These differences can be bridged and brought back to the SNR at given circumstances. This chapter provides the physical background and the tools to do that.

Then there is the difference in apple-to-pear category. What is more important: the small size of a CMOS sensor or the low power consumption of an I2? This chapter will list the advantages and disadvantages of the different technologies, and it is up to the system architect to select the appropriate detection technology for his or her application.

5.2 The Photocathode

5.2.1 The Working Principle of Photocathodes

Image Intensifiers as well as PMTs are based on photocathodes. Their function is to convert as many incoming photons as possible into photoelectrons emitted into the vacuum [1]. The basic principle is simple and described by the following three-step process:

Step 1: Absorption of the photon and creation of an electron–hole pair
Step 2: Transport of the electron to the vacuum interface
Step 3: Emission of the electron into the vacuum

These steps are illustrated in Fig. 5.3.

The absorption of the photons in condensed matter is very well understood and can be described precisely by optical theory. The index of refraction (n) and the extinction ratio (k) of the photocathode material play a predominant role. The long wavelength limit of absorption is given by the bandgap of the photocathode material. Photocathodes are transparent for photons with an energy above the bandgap. The cutoff wavelength λ_{co} is given by:

$$\lambda_{co} = \frac{hc}{E_g} \quad \text{or} \quad \lambda_{co}(\text{nm}) = \frac{1{,}240}{E_g(\text{eV})}. \tag{5.1}$$

Below this wavelength, a certain fraction of the photons is absorbed. This fraction depends on the thickness and the optical properties of the photocathode materials.

Figure 5.4 shows the absorption lengths of S25 and GaAs photocathodes. These are widely used broadband photocathodes. The absorption length is the thickness of the material absorbing a fraction of $1/e \approx 63\%$.

The absorption length for UV/blue is much shorter than the absorption length for red/NIR. Many photocathodes are semitransparent for the longer wavelength. In that case, the interference of the waves directly reflected from the glass cathode interface and from the cathode vacuum interface plays a role and must be accounted for in a

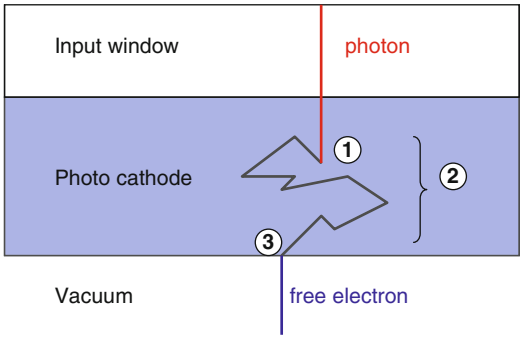

Fig. 5.3 The three-step process of photoemission in a photocathode

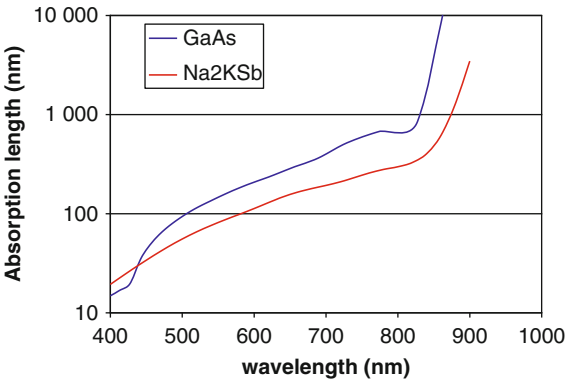

Fig. 5.4 Absorption lengths of the most common photocathode materials

complete optical model. The created photoelectron will have an excess energy equal to the energy of its generating photon minus the bandgap energy.

The second step involves the transport mechanism of the photoelectron from the location of its creation toward the vacuum interface. If there is no electrical field in the photocathode, the only mechanism available is diffusion. After a random walk, electrons will end up at the surface. During its random walk, an electron will lose part or all of its excess energy. The dominant energy loss mechanism is phonon interaction. Many hot electrons will become thermalized and adopt a Fermi–Dirac distribution on top of the conductance band. All practical cathodes will have regions of band bending at their glass and vacuum interfaces. In these regions, there will be an electrical force acting on the electrons.

When this force is pointed toward the vacuum, the diffusion process is helped by the drift. When it is pointing toward the glass interface, it creates losses. A second enemy of high-quantum efficiency is the recombination of electron–hole pairs. The longer the net transport path length to the vacuum interface, the higher the chance that the electron–hole pair recombines. The number of traps, such as crystal impurities or grain boundaries, plays an important role. The better the crystal structure and the fewer impurities, the higher the chance that the electron will survive the transport step 2 and reach the vacuum interface. As a consequence, thin cathodes and electrical fields toward the vacuum interface help to reduce the gross travel length.

Step 3 deals with the electron emission process into the vacuum. The most important parameter of this step is the electron affinity. This is the energy level of the vacuum compared to the energy level of the conductance band. Most materials have a positive electron affinity (PEA). This means that an electron requires a sufficient amount of energy above the conductance band to escape to the vacuum. This required energy can be thermal. A fraction of the thermalized electrons having a Fermi–Dirac distribution will have enough energy to escape. In photocathodes with PEA, the predominant source of energy is the excess energy at the moment of creation. Consider the following example: a photon with a wavelength of 400 nm has an energy of 3 eV. Assuming this photon creates a photoelectron in a photocathode

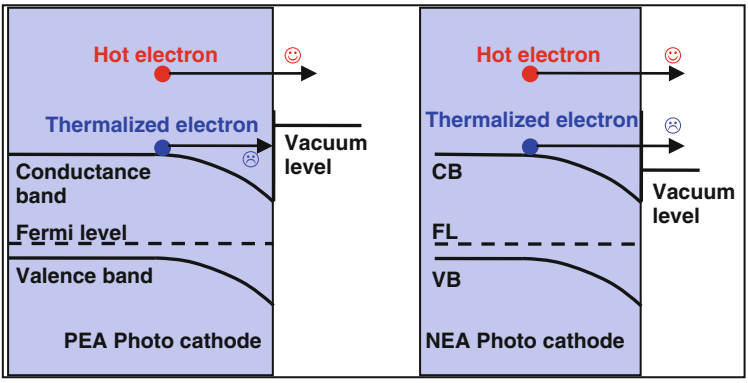

Fig. 5.5 The emission process in PEA and NEA photocathodes

with a band gap of 1.4 eV, the photoelectron has an excess energy of 1.6 eV. Assume further a PEA of 0.4 eV, then this electron can lose 1.2 eV during its random walk and still has enough energy to escape to the vacuum.

Some materials, among them the more efficient photocathodes, show negative electron affinity (NEA). In this case, electrons gain energy when escaping from the bulk to the vacuum. These NEA photocathodes need a specially prepared surface layer involving Cesium or Rubidium to achieve NEA. It is commonly accepted that NEA photocathodes do have a band-bending region followed by a tunnel barrier as can be seen in Fig. 5.5. These two phenomena have the consequence that even in NEA photocathodes electrons that arrived at the vacuum interface have an escape probability well below 100%.

The total quantum efficiency is now given by the multiplication of the efficiencies of the individual three steps.

There are very many different types of photocathodes reported and described in literature. This chapter limits itself to the most commonly used types, divided into two categories: multialkali and III–V photocathodes.

5.2.2 Multialkali Photocathodes

Multialkali photocathodes are based on alkali metals: sodium (Na) and/or potassium (K), combined with Antimony (Sb) for the bulk and Cesium (Cs) or in some case Rubidium (Rb) for the vacuum exit layer.

Most-common bulk structures are:

Bialkali: Na_3Sb or K_3SB
S20: Na_2KSb (typically 60 nm thick)
S25: Na_2KSb (typically 170 nm thick)

The reason why these materials are historically used for photocathodes includes the following:

- They can be deposited as polycrystalline semiconducting layers with reasonably high carrier lifetimes.
- They are very good absorbers for photons from 200 up to 850 nm.
- The required Cs for the vacuum interface is chemically closely related to the bulk building blocks Na and K.
- A good chemical bond between Cs and the bulk can be easily achieved.

The alkali metals are very strong oxidizers. The smallest amount of oxygen or water totally burns the photocathode. Therefore, they have to be kept in a very good vacuum (10^{-9} mbar or better). The smallest vacuum leak will end the life of the photocathode.

The manufacturing process of multialkali photocathodes always starts with the creation of a very good vacuum. Dispensers are then used to create an atmosphere with background pressures of Potassium and Sodium. Then an Antimony flux is created, either pulsating or continuous. In other processes, the windows are pre-evaporated with Antimony and this metal reacts with the alkali gasses in the atmosphere. All processes use an elevated temperature to enhance the crystallinity of the material.

After the required thickness is achieved, the process finishes in all cases by creating a Cesium (or Rubidium) pressure to create a Cesium-rich top layer. The exact photocathode recipes are very well kept trade secrets of the individual manufacturers.

All devices requiring a photocathode are sealed in the same vacuum environment as when the cathode was created. A common way to seal PMTs is to pinch off the copper vacuum tube connecting the glass enclosure to the manufacturing vacuum system. Image intensifiers are sealed by soldering the photocathode on a flange of the I2 enclosure.

Multialkali photocathodes have two weak points. The active layer is directly deposited on the window. This causes a high recombination rate on the cathode–window interface. This electron sink becomes even worse by the common downward band bending of conduction and valence band near the window. This is illustrated in Fig. 5.6.

The second drawback is the polycrystalline structure with its many recombination centers. Therefore, the mean free path length of an electron is relatively short. Fortunately, this is partially compensated by the very good optical absorption.

An S20 photocathode has PEA. As is illustrated in Fig. 5.5, only hot electrons are able to escape. Because electrons lose their energy during the transport step (2), PEA cathodes must be thin; typically 50–70 nm. The PEA also explains the cutoff wavelength well below 850 nm. The energy of the photons at the cutoff wavelength is equal to the bandgap minus the electron affinity. The advantage of the PEA surface is that not only thermalized photoelectrons but also thermally generated electrons (the dark current) cannot escape. Therefore, the dark current of PEA cathodes is extremely low, typically $1 \times 10^{-16}\,\mathrm{A/cm^2}$ for S20 and even $2 \times 10^{-18}\,\mathrm{A/cm^2}$

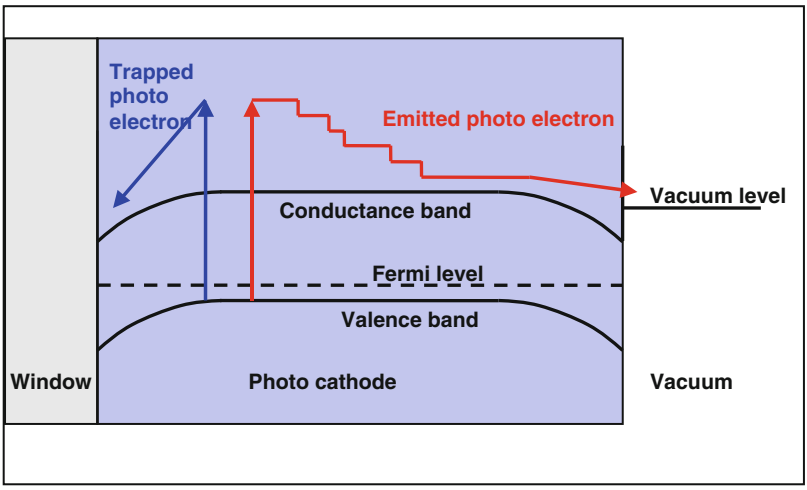

Fig. 5.6 The emission and loss processes in an S25 photocathode

for bialkali. An S25 has NEA and therefore a wavelength sensitivity up to the bandgap-related absorption cutoff. Because thermalized electrons can escape in an NEA structure, S25 cathodes can be much thicker, typically 170 nm. The cutoff wavelength is now directly determined by the bandgap. A drawback is that NEA photocathodes have also a higher dark current, typically 7×10^{-15} A/cm^2.

5.2.3 III–V Photocathodes

The most common III–V photocathode is the one based on GaAs. The bandgap of GaAs is 1.4 eV, equal to the bandgap of the S25 photocathode. The advantage of GaAs is that wafers of this material are readily available as a very pure monocrystalline semiconducting material. The main applications of GaAs are high speed electronics and lasers. Several suppliers can manufacture doped structures and hetero junctions on customer request.

One major drawback of S25 cathodes, namely the electron sink at the window interface, is solved by a window layer. This is a lattice matched layer of AlGaAs with higher band gap, which is positioned between the window and the active GaAs layer. The hetero-junction between AlGaAs and GaAs creates an upward bending of the conduction band and therefore an electrostatic force on the electrons in the direction of the vacuum interface. This is illustrated in Fig. 5.7.

Instead of an interface being an electron sink, now the interface is an electron mirror that reflects electrons back to the direction of the vacuum interface. The disadvantage is a loss of sensitivity for photons below 550 nm, which are absorbed by the AlGaAs window layer.

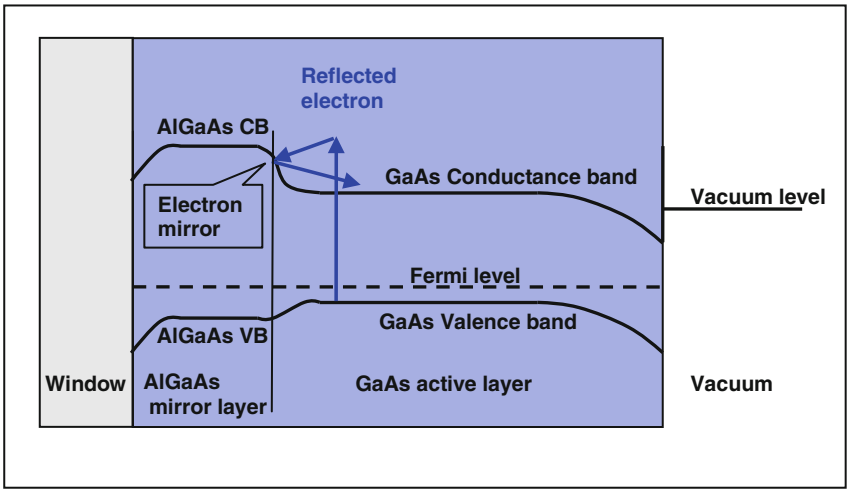

Fig. 5.7 The GaAs photocathode structure with the AlGaAs window layer

The GaAs photocathode bulk does not need to be fabricated in ultra-high vacuum and kept in vacuum forever afterward. The bulk of III–V photocathodes can be fabricated, in general, on the manufacturing premises of GaAs crystal growers and transported under ambient conditions subsequently.

For all practical purposes, the GaAs bulk needs NEA. To achieve this, exit layers using Cesium or Rubidium are used. Unfortunately, these materials do not chemically bond to the GaAs bulk. These alkali emission enhancers have to be applied to a very clean surface in very good vacuum ($<10^{10}$ mbar). The Cesiated GaAs surface is even more sensitive to oxidation than an S25 photocathode. The biggest drawback of III–V cathodes is, therefore, their extreme sensitivity to overexposure and especially to ion feedback. Figure 5.8 illustrates this in an example of an overilluminated GaAs photocathode. Where the Cesium has disappeared, the QE of the photocathode drops practically to zero.

5.3 Image Intensifiers

5.3.1 Working Principle

The first image intensifiers (Generation 1, or Gen 1 for short) had two main functions:

- Conversion of photons into electrons by the photocathode.
- Conversion of photoelectrons into multiple photons by the phosphor screen.

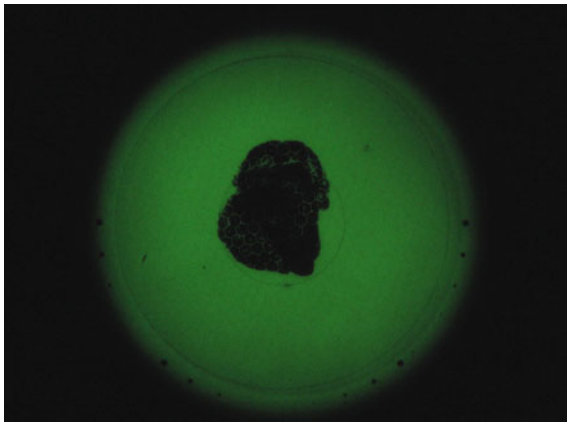

Fig. 5.8 An over-illuminated GaAs photocathode

Fig. 5.9 The principle of an image intensifier. To the *left*, the input window with photocathode is shown, in the *middle* the MCP and to the *right* an anode

The gain was provided by the phosphor, so high accelerating voltages were used (typically 12–20 kV) to achieve adequate gain. Focusing is achieved by electrostatic lenses. Nowadays, Gen 1 intensifiers are only manufactured for the replacement market or for some niche applications. Current image intensifiers (Gen 2/3) have one additional function: multiplication of electrons by a multichannel plate (MCP). Gen 2/3 use proximity focusing. The image sharpness is achieved by placing all components closely together (Fig. 5.9). The typical gap between photocathode and MCP is 0.1–0.2 mm, and between MCP and phosphor screen it is 0.5 mm. The difference between Gen 2 and Gen 3 image intensifiers is the type of the photocathode: Gen 2 uses multialkali and Gen 3 uses GaAs-based photocathodes. All other features and performance figures are very comparable.

The three active components are held in position by the enclosure: a stack of metal and ceramic rings. The enclosure also ensures the ultra-high vacuum inside the I2, and it provides the voltage feed-throughs (Fig. 5.10). Finally, every I2 requires a power supply, most of which are delivered as a “wrap-around” version.

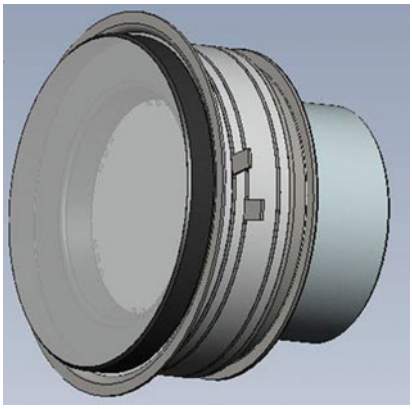

Fig. 5.10 An assembled image intensifier. The *dark gray rings* are the metal rings to electrically connect the components, the *light gray rings* are the ceramic isolators

Fig. 5.11 Soldier equipped with night vision, seen through a night vision device

5.3.2 Applications

The most important application for image intensifiers is night vision. For this application, image intensifiers are supplied in standardized packaging including a potted wrap-around power supply. End-users are military and other law enforcing agencies (Fig. 5.11). Apart from this, there is also a smaller and growing market for leisure, such as tracking, yachting, or hunting.

Another smaller but important market can be found in scientific and industrial applications. Requirements leading to the use of an image intensifier as a detector include very low dark current, low power consumption, or fast gating capability. There is a long list of niche application, where I2 make valuable contributions: astronomy, bioluminescence, cell biology, chemiluminescence, corona imaging, engine combustion analysis, fluorescence lifetime imaging, high-speed cameras, medical diagnostics, microscopy, missile warning systems, semiconductor manufacturing, spectroscopy, time-resolved imaging, and X-ray imaging.

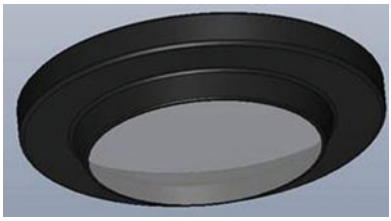

Fig. 5.12 The input window with its antiveiling glare (AVG) coating

5.3.3 The Components of an Image Intensifier

5.3.3.1 Input Window and Photocathode

The input windows of Gen 2/3 image intensifiers are traditionally 5.5 mm thick and made from Corning 7056 glass or an equivalent. The window has a ledge to bring the active surface close to the MCP while creating space for the MCP front contact and clamp mechanism. Optical reflections from the tilted surfaces are avoided by applying a black (antiveiling glare, AVG) coating (Fig. 5.12).

When designing optics for a high-resolution image intensifier, one must consider the aberrations introduced by the relatively thick planar input window. Objectives designed for focus in air will yield large chromatic and spherical aberrations, especially when they have a small aperture (f-number). The photocathode as described in Sect. 5.2 is evaporated on the window (Gen 2) or bonded to the window (Gen 3).

5.3.3.2 Multichannel Plate

The MCP has the function to multiply the emitted photoelectrons while maintaining the spatial information. This is achieved by using a thin glass plate with many holes. A typical MCP has about six million channels over an active surface of 18 mm diameter. Every channel has a diameter of 6 μm and a length-to-diameter ratio of 50 (Fig. 5.13).

The MCP is manufactured by drawing single fibers, stacking them, redrawing the stacked fibers, and restacking them. This is followed by a fusion process to seamlessly fuse all cladding fibers. Then a saw cuts thin slices of the "boule." These slices are ground to the correct thickness and polished. The next step etches the core fibers. and the matrix of fused cladding glass with holes remains. The cladding glass is made of a composition containing lead (Pb) and alkali metals. A high temperature reduction process makes the wall of the holes slightly conductive, and the glass possesses a reasonably high secondary emission yield (SEY). Finally, the front and back end electrodes are evaporated on the MCP, usually using a Nickel–Chrome alloy. The back end electrode penetrates a few channel diameters into the channel. This prevents the creation of many secondary electrons close to the channel output and therefore provides a better focusing. The channels are not perpendicular to the

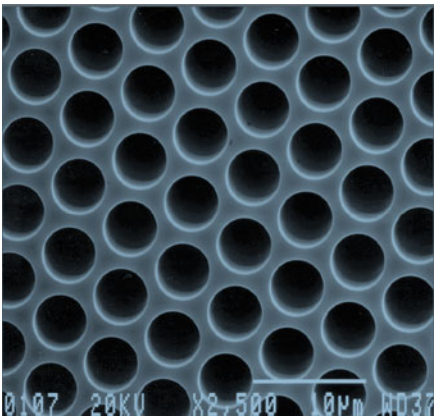

Fig. 5.13 An SEM picture of an MCP

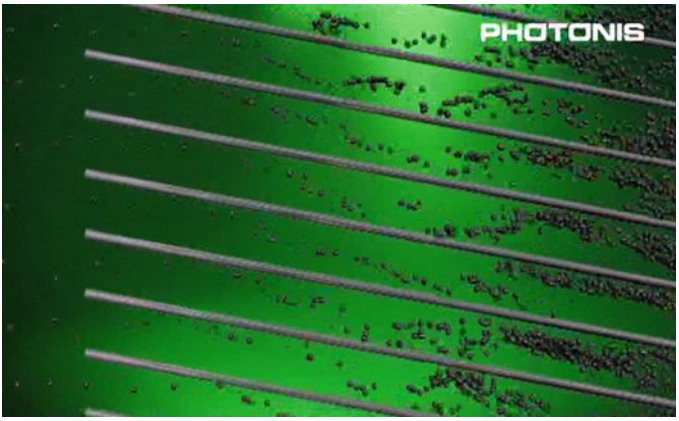

Fig. 5.14 Visualization of the electron gain mechanism within the MCP channels

surfaces but they rather have an angle of a few degrees (the so-called bias angle). This angle reduces ion and light feedback.

The gain is achieved by applying a voltage of approximately 1 kV across the electrodes. A typical MCP has a resistance of 100 MΩ. A voltage of 1 kV results in a typical current of 10 μA. This is called the strip current. This current distributes the voltage evenly along the channel. When the input of the channel is hit by a photoelectron, it will create a few secondary electrons. The field inside the channel will accelerate these electrons in the direction of the output. After the secondary electrons have gained sufficient kinetic energy, they will collide again with the wall, creating more secondary electrons. This process repeats itself like an avalanche until a cloud of electrons leaves the channel.

The gain mechanism is illustrated in Fig. 5.14, see also reference [2].

Simulations show that typically 20 wall interactions occur until the output side of the MCP is reached. The Gain G can therefore be expressed as:

$$G = \eta_1 \cdot \eta_2^{20}, \tag{5.2}$$

where η_1 denotes the SEY of the first impact of the primary photoelectron and η_2 is the SEY of the ensuing impacts in the channel.

The gain adjustment is achieved by the MCP voltage. The secondary electron yield of every step will change with the voltage. This and the factor η_2^{20} let the MCP gain triple for 6% extra SEY. A typical rule of thumb is that gain doubles for every 50 V additional MCP voltage. The gain mechanism through charge multiplication of many small gain steps has a consequence for the statistics of gain. From event to event, there are large fluctuations, which can be explained by Poisson statistics of the first few gain steps. The table below shows the statistics for $\eta = 1.5$. In 22% of the cases, the event is extinguished because there is no secondary electron emitted. Also, the gain variation is illustrated: when for an individual event, the first impact creates four secondaries, the expected gain of that event will be four times higher than when only one secondary is created.

Poisson statistics for $\eta = 1.5$	
Number of secondaries	Frequency (%)
0	22
1	33
2	25
3	13
4	5
5	1

An MCP will saturate when its emission current is around 10% of the strip current, so for the typical strip current of 10 μA, the maximum emission current is 1 μA.

MCPs used for image intensifiers with GaAs photocathode have an ion barrier at the input. This is very thin film of SiO_2 or Al_2O_3 that blocks ions and prevents them from damaging the sensitive Cs layer on the photocathode. Unfortunately, this layer also blocks a fraction of the electrons, so it reduces also the obtained quantum efficiency.

5.3.3.3 Phosphor Screen and Output Window

Traditionally, many output windows for image intensifiers consist of fiberoptic face plates. These are bundles of fibers with fused claddings. A typical diameter of the individual fiber is 4–6 μm. There are several advantages using fiber optics: The image is available at short focal length without any objects potentially vignetting the beam. Second, the image plane can be given a curvature by polishing the fiber. Third, an image rotation can be realized by twisting the fiber bundle after heating it.

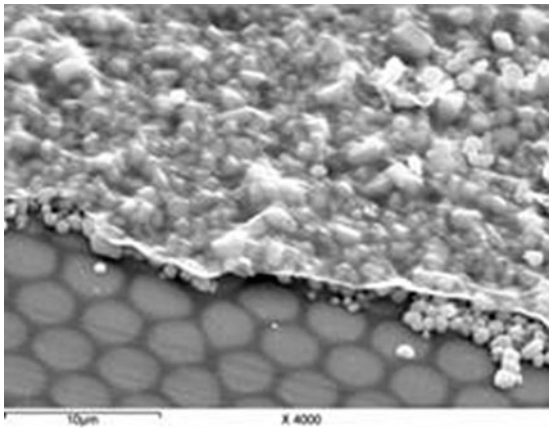

Fig. 5.15 SEM photo of a phosphor screen

This is advantageous when the fiber is used in a single axis optical system. The objective turns the image upside down, while the ocular maintains the orientation. The "twister" provides the required additional image rotation in a very short building length. Also for digital applications, the fiber has an advantage: A CMOS or CCD chip can be directly optically coupled to the fiber.

The phosphor is deposited on the vacuum side of this output window. Commonly used phosphors include P43 (GdO2S:Tb) and P22 (ZnS:Cu,Al). Either a settling method or a brushing technique is used to apply the phosphor. On top of the phosphor, a very thin Aluminum layer is applied. The function of this layer is to increase the efficiency of the screen, to prevent light feedback to the photocathode, and finally to conduct the deposited charge back to the power supply. Figure 5.15 shows an SEM photo with the phosphor powder, the Aluminum backing layer, and the fiber faceplate clearly visible.

5.3.3.4 Enclosure and Power Supply

The three active components (the input window with the photocathode, the MCP and the output window with the phosphor screen) are held in their position by the enclosure. This is a stack of isolating ceramic rings with metal feed-throughs. The output window is glass fritted to its flange and welded to the rest of the enclosure. The MCP is mechanically supported on the output face and mechanically clamped on the input side. The input window can only be connected after photocathode deposition and activation. It is either soldered (Gen 2) or cold sealed (Gen 3) inside the cathode deposition or activation equipment.

Typical required voltages are 200–1,000 V for the photocathode, 600–1,000 V for the MCP and 4–6 kV for the anode. For most applications, these voltages are supplied by an integrated power supply (Fig. 5.16). There are a number of provisions to control and protect the image intensifiers. The gain is controlled primarily by the

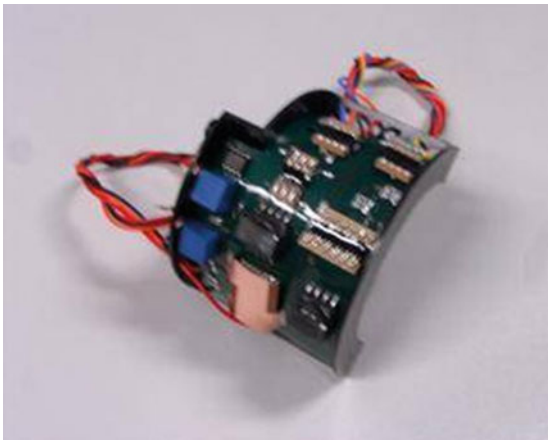

Fig. 5.16 A power supply unit (PSU) providing photocathode, MCP, and anode voltages

MCP voltage. The maximum gain is usually set by the factory but some versions of power supply units (PSUs) have externally controlled gain. When a certain preset anode current is reached, an electronic circuit will reduce the MCP voltage to cap the anode current, hence the output brightness. This is called the "maximum output brightness (MOB)" control. Some PSUs have an autogating function. This is an additional control of the gain. An alternating negative ("on") and positive ("off") voltage to the cathode reduces the duty cycle and limits the cathode current. Autogating enables the use of the photocathode at relatively high light levels without damaging it.

Finally, the photocathode is protected by a BSP circuit. This limits the current through the photocathode to protect it.

5.3.4 Performance Characteristics

Two regimes are distinguished to describe the imaging performance of an image intensifier [3]. At high light levels, there are enough photons available to form a practically noise-free image. The image quality is determined by the image output contrast. The contrast will be dependent on the spatial frequency, and it is quantified with the modulation transfer function (MTF).

At a level of around 1 mlux photocathode illumination, the noise created by the limited number of photons (shot noise) starts to reduce the imaging performance. The lower limit of illumination level is given by the dark current. The lowest practical light level depends on the application. For night vision applications, the lowest light level of practical use is so-called "heavily overcast star light." This light level translates into a sensor illumination of 10 μlux. This is still two orders of magnitude above a typical photocathode dark current corresponding to 0.15 μlux.

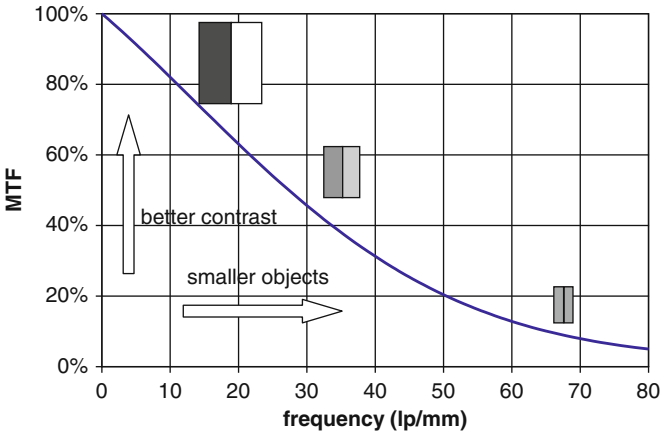

Fig. 5.17 A typical MTF of an image intensifier

5.3.4.1 High Light Level Performance

The best way to describe the image performance of an image intensifier at high light levels is the MTF. The MTF is defined as the contrast at the output when a sinusoidal signal with 100% contrast is presented at the input. The output contrast is given as a function of spatial frequency, usually from 0 to 100 lp/mm (line-pairs per mm). Figure 5.17 shows a typical MTF of an image intensifier. The following components will define the total MTF of the image intensifier:

- *The cathode gap*: Electrons leaving the photocathode have an "escape energy." This energy can be transversal, creating a spot size on the MCP with a diameter D_c given by:

$$D_c = 2 \cdot z_c \sqrt{\frac{E_0}{E_c}}, \tag{5.3}$$

  where z_c is the gap between cathode and MCP, E_0 is the escape energy, and E_c is the cathode–MCP voltage.

Assuming an escape energy of 100 meV, a cathode–MCP acceleration voltage of 200 V and a cathode distance of 200 V, a diameter of the electron spot on the MCP of around 18 μm will result.

- *The MCP channel and fiber diameter*: The information of the exact position of the photoelectron within one MCP channel is lost. The same holds for the output fiber. No detail smaller than the MCP channel diameter or the output fiber can be successfully imaged.
- *The anode gap*: The MTF of the anode gap is much more complicated than the one of the cathode gap. The electrons leaving the MCP channel have, on the one hand, a very wide energy distribution. On the other hand, the angular distribution

is not Lambertian. Electrons leaving the channel are more or less collimated. The length of the end spoiling plays a role in this collimation process. To complicate things even further, the changes of axial field strength at the position of the end spoiling and also at the exit of the MCP leads to a lensing effect. As a consequence, the MTF of the anode gap cannot be adequately described analytically. It needs to be simulated by Monte Carlo simulations or it must be practically measured.

- *The phosphor screen*: There are a number of factors defining the MTF of the phosphor layer, the most important one being the thickness of the layer. But also the grain size of the phosphor grains, the reflectivity of the Aluminum backing layer, and the light scattering within the layer are significant.

5.3.4.2 Low Light Level Performance

From 1 mlux downward, the noise performance of the image intensifier plays an increasingly important role [4]. Image intensifiers are shot noise limited until extremely low light levels of below 1 μlux. So unlike the case of low-light CMOS and CCD imagers, dark current and read-noise are insignificant. Figure 5.18 is a magnified low-light-level image of the phosphor screen when the photocathode is illuminated with 10 μlux. Every spot is the result of a single photoelectron.

In this shot noise regime, the noisiness of the image is determined by three factors:

1. *The quantum efficiency of the photocathode*. Quantum efficiencies of broadband photocathodes are around 15% between 350 and 850 nm for S25 photocathodes and 40% between 600 and 850 nm for GaAs photocathodes. S20 and bialkali photocathodes are more blue sensitive and can achieve a QE of up to 50% at 450 nm.

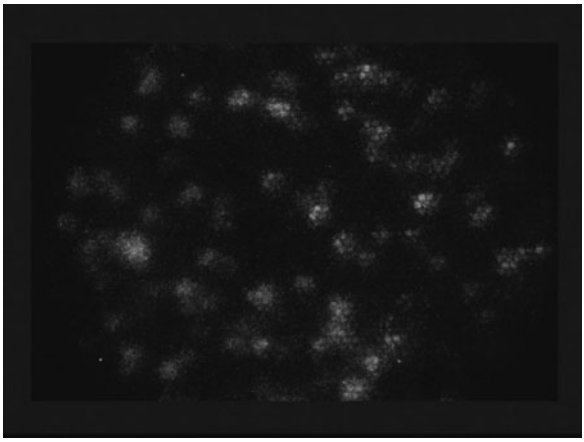

Fig. 5.18 An image of the phosphor screen with the image intensifier illuminated with 10 μlux

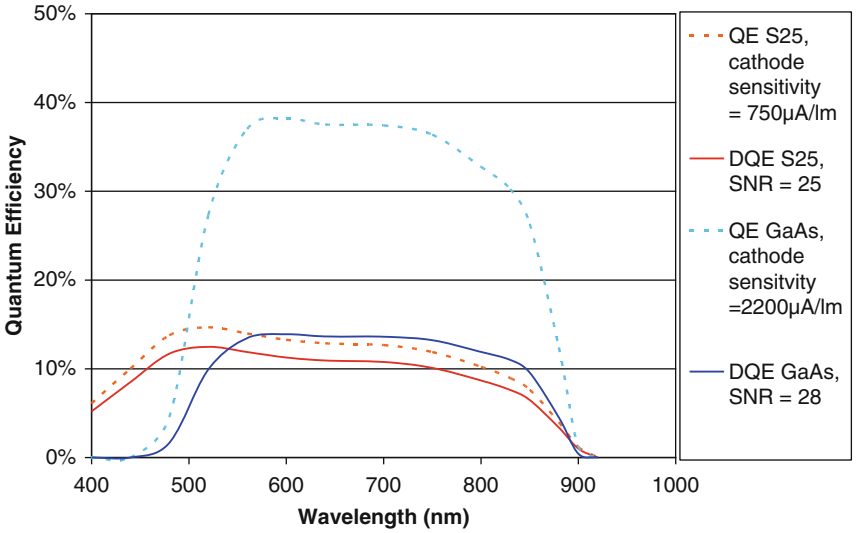

Fig. 5.19 Quantum efficiencies (QE) and detected quantum efficiencies (DQE) of S25 and GaAs photocathodes

2. *The collection efficiency of the MCP*. Some photoelectrons do not create a detectable effect on the phosphor. They are annihilated by the MCP cladding or by the Poisson statistics of the first interaction with the MCP. This explains a drawback of GaAs (Gen 3) photocathodes. The ion barrier on top of the MCP, necessary to protect the sensitive Cs film on the GaAs photocathode, also annihilates many photoelectrons, resulting in a low collection efficiency. The detected quantum efficiency (DQE) is the product of QE and collection efficiency. The DQE of both S25 and GaAs photocathodes is around 12%, see Fig. 5.19. The noise at low light levels show Poisson statistics, so its root mean square deviation is equal to the square of the average number of events. Therefore, a good measure of the DQE is the SNR at low light levels. Figure 5.19 shows typical values and also illustrates that one only benefits slightly from the higher QE of GaAs photocathodes because the DQE is very similar.
3. *The gain noise*. The photoelectrons that are not annihilated will create an avalanche of electrons in the MCP channel. The total gain is achieved by many steps of small gain. Poisson statistics determines the spread of the gain per event. This spread is large and causes extra noise in the output signal.

Image intensifier specifications usually define the SNR, a number between 15 and 35. This refers to a specific test where the intensifier is illuminated through an aperture of 200 µm with a light intensity of 108 µlux. The output signal is filtered with a 10 Hz low-pass filter. The measured SNR of this signal is referred to as the SNR of the image intensifier.

The dark current is usually expressed as an EBI, defined as the light level (in μlux) required to achieve the same output as the thermal dark current of the photocathode. A state-of-the-art EBI for S25 and GaAs photo cathodes is 0.10 μlux, which is about 1% of the minimum light level of 10 μlux of relevance in the field.

5.3.4.3 Image Artifacts

Image intensifiers have some typical artifacts. First of all, there are the haloes, round brighter areas around bright spots. There are three main causes for haloes. The most disturbing halo is the one created by backscattered electrons on the MCP. The diameter of this halo is equal to about four times the cathode gap. GaAs-based tubes suffer the most from this halo because electrons also backscatter from the ion barrier and the cathode gap. Also, the halo is larger due to the larger cathode gap needed because of the higher cathode voltage used. The second halo is the optical halo created by the light transmitted through the photocathode, reflected by the MCP and illuminating the photocathode from the backside. This is more prominent with S25 cathodes since they are more transparent. The third halo is the electron backscatter halo created by the aluminum of the screen (Fig. 5.20).

A second typical artifacts are honeycomb or square patterned structures created by either the drawing process of the fiber or of the MCP. In most cases, they can be found on the border of a stacked bundle of fibers as a byproduct of the manufacturing process. These patterns have usually a very low contrast but because the human eye is very sensitive to these types of patterns, they are still visible. The honeycomb structures are often nicknamed “chicken wire” (Fig. 5.21).

A third artifact is the appearance of black spots. There are several causes for black spots. They can be damages or particles on the photocathode, the MCP, the phosphor screen, or the fiber. The phenomena is comparable to dead pixels in CMOS or CCD imagers. Unlike in digital devices, these artifacts cannot be corrected for in a direct view system (Fig. 5.22).

Fig. 5.20 Example of a halo

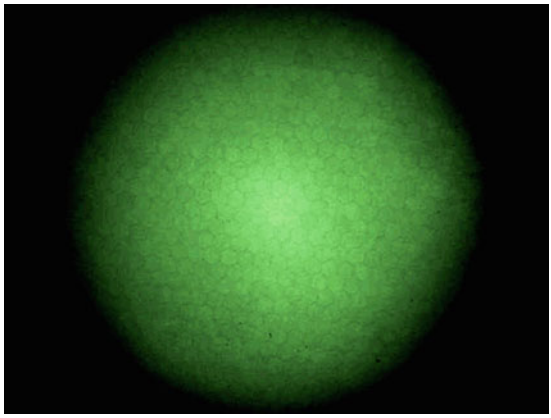

Fig. 5.21 Example of chicken wire

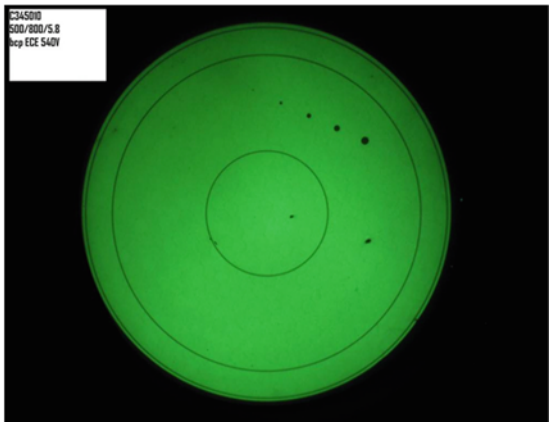

Fig. 5.22 Example of black spots

5.3.4.4 Comparison with Solid State Imagers

It is not straightforward to compare an image intensifier with a CMOS low light level detector. First of all, an image intensifier is a complete system from "photon-to-photon." Comparing apples with apples – one apple is the image intensifier and the other apple a hybrid system consisting of CMOS sensor, a camera, and the display.

Then there are many differences between definitions and terminology used. The I2 community talks about cathode sensitivity in $\mu A/lm$, resolution in lp/mm and dark current in EBI. CMOS people talk about quantum efficiency, pixel size, and count and dark current density in pA/cm^2, respectively. The third obstacle in the comparison lies in conceptual differences; an image intensifier just has neither pixel structure nor a frame rate.

To provide a practical basis for the comparison, a reference is chosen, namely a system based on an SXGA($1{,}280 \times 1{,}024$ pixel) CMOS image sensor with a pixel

pitch of $10 \times 10\,\mu$m. The diagonal of such a sensor is 16.5 mm, close to the 18 mm active diameter of most image intensifiers. We further assume that the CMOS-based camera has a frame rate of 60 Hz.

Resolution: A round I2 fiber with 18 mm active diameter with 4 μm fibers with a pitch of 5 μm represents ten million fibers. This is not a completely appropriate comparison with the CMOS pixels because other I2 components limit the maximum contrast of two neighboring fibers as given by the MTF curve. A better comparison is to convert the limiting resolution to the number of equivalent pixels. A state-of-the-art image intensifier has a limiting resolution of 72 lp/mm. To achieve this with a pixelated structure, one needs at least 144 pixels/mm. With an 18 mm round surface, this translates to a total of 5.3 million pixels. So the equivalent pixel count of an image intensifier is 5.3 M pixels. This is four times more than our CMOS reference.

Frame rate: The image intensifier does not work with individually exposed frames, so the concept of a frame rate is not applicable. The time-limiting factor is the phosphor. The often used P43 phosphor (GdO2S:Tb) has a decay time of 1 ms, so one could say that the equivalent frame rate is 1,000 Hz. This is 17 times more than the reference CMOS sensor.

DQE: The DQE of S25 or GaAs is around 13%. The QE of solid-state imagers is usually substantially higher; up to 80% is reported. When it comes to low-noise CMOS structures, designers choose usually the pinned photodiode and avoid deep pixel structures. With these noise-reducing measures, QE is sacrificed, especially in the NIR wavelength range. A fair average DQE in the range of 700–850 nm of a low-noise CMOS imager is 40%, about three times higher than that of an image intensifier.

Dark current and read noise: Dark current of solid-state imagers is usually expressed in $\mathrm{pA/cm^2}$ and read noise in pixel equivalent electrons. Recent CMOS developments such as correlated double sampling, pinned photodiodes, and bandwidth engineering reduce the read noise to values of below one electron at room temperature, as described in several chapters of this book. On the contrary, dark current is much more related to the CMOS technology node and the semiconductor foundry. State-of-the-art-values are CMOS image sensors with pinned photodiode offering about $10\,\mathrm{pA/cm^2}$ at room temperature. The dark current of an image intensifier is expressed in EBI, which can be converted to a current density value through:

$$I_{\mathrm{dark}} = PR \cdot \mathrm{EBI} \cdot S, \tag{5.4}$$

where PR is the photoresponse, and S is the imager's surface. Assuming $PR = 700\,\mu\mathrm{A/lm}$, a typical EBI of 0.1 μlux and a surface S of $1\,\mathrm{cm^2}$, results in a typical dark current of an image intensifier of $7\,\mathrm{fA/cm^2}$, which is more than four orders of magnitude lower than the dark current of a CMOS sensor. Higher temperatures, through ambient or through internal power dissipation, will significantly increase the dark current.

Based on these numbers, a fair comparison of signal and noise values for a state-of-the-art CMOS imager and an image intensifier at an illumination level of 10 μlux

at the face plate (I2) becomes possible. Compare the data for one $10 \times 10\,\mu m^2$ pixel with the same surface of the image intensifier at a frame rate of 60 Hz.

	I2	CMOS
White sensitivity ($\mu A/lm$)	750	3,000
QE at 800 nm	15%	35%
DQE at 800 nm	13%	35%
Signal (e/frame/pix)	0.07	0.19
Shot noise (e/frame/pix)	0.26	0.43
Dark current (e/frame/pix)	0.0007	10.00
Dark current noise (e/fr/pix)	0.026	3.16
Read noise (e/fr/pix)	<0.01	2.00
Total noise (electrons/frame/pix)	0.27	3.77

This example shows that image intensifiers are still unsurpassed in noise performance when it comes to very low light levels. Their dark current is so low that in practical circumstances only the signal shot noise contributes to the SNR. Despite the impressive reduction in read noise of CMOS imagers in recent years, there is still one order of magnitude to bridge until they become usable at the lowest required light levels. When it comes to dark current noise, CMOS imagers will require heavy cooling to achieve satisfactory levels. This will come with a power penalty. The practical application will determine whether this is an important drawback or can be tolerated.

Power consumption: The power consumption of an image intensifier is 30–60 mW. An uncooled CMOS-based camera system will typically consume 1 W. When cooling is required to reduce dark current of a CMOS imager, this comes with a heavy power penalty. Cooling powers of 5–10 W have to be taken into account.

5.3.5 Special Image Intensifiers

5.3.5.1 Fast Gating Image Intensifiers

Image Intensifiers are gated by a reversed potential between photocathode and MCP. A standard image intensifier can achieve a gating time of 20 ns. Specialized image intensifiers contain means to decrease gating times to less than 1 ns. This can be achieved with conductive layers below the photocathode and optimized construction of MCP suspension and cathode mounting, as well as matching impedance to the supply lines.

5.3.5.2 High Gain Intensifiers

Standard image intensifiers exhibit a gain of typically 10^4. The gain can be increased by increasing the MCP voltage. The maximum gain is limited by the light and ion feedback and also by the maximum voltage a single MCP can tolerate. There are special image intensifiers available for very high gain of up to 10^6. This is achieved by using stacks of two ("chevron stacked") or three ("Z stacked") MCPs. These types of image intensifiers are usually applied for single-photon imaging.

5.3.5.3 Image Intensifiers with Electronic Output

When an image intensifier is not used for direct viewing but the image is required in an electronic format, then an electronic output can be provided. The traditional way to achieve this is to couple a CMOS or CCD imager directly on the fiber output. These devices are called ICCD or ICMOS (I for "intensified"), respectively. An alternative solution is to make the CMOS sensor charge-sensitive and replace the phosphor screen by this chip. Such a device is called MCPCMOS.

Yet, another solution is to electron bombard the back side of the CMOS or CCD sensor directly with the photoelectrons. This solution is called EBCMOS and is described in detail in Chap. 4 of this book.

There are also solutions for photoncounting with fast timing requirements. With these solutions, individual events are marked with an (x, y) coordinate and with a time-stamp. The anode behind the Chevron or Z stacked MCPs is, for example, a resistive plate. Siegmund published also a solution with cross strip anodes [5]. The very fast transient time of such a device makes timing measurements with a resolution of better than 100 ps possible. By using centroiding of single events, a spatial resolution of about 3 μm is possible.

5.3.5.4 Open MCP Detectors

Open MCP detectors are usually high-gain solutions, possibly with electronic readout anodes. These detectors have no input window; hence they can only be operated in vacuum. The MCP as such is sensitive to electrons and ions. Special layers on the MCP can make these detectors sensitive to hard UV or X-ray radiation.

5.4 Photomultiplier Tube

PMTs are used to convert photons into measurable electronic pulses [6]. PMTs are nonimaging devices; they just provide a single spatial channel. In many applications, such as in nuclear medicine and astrophysics, many PMTs are combined to effectively form an imaging detector. PMTs are the traditional alternatives to solid-state APDs described in Chap. 7 of this book.

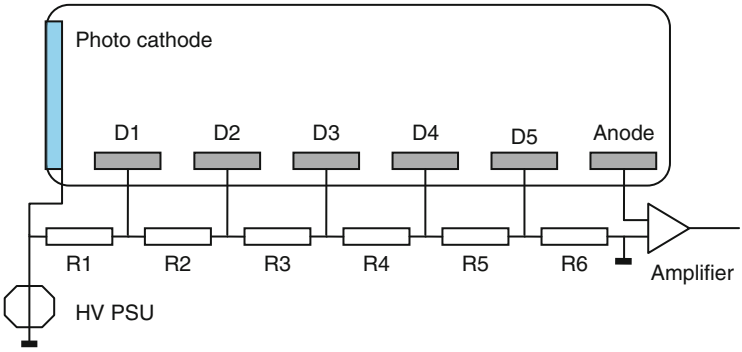

Fig. 5.23 The principle of a PMT

5.4.1 Working Principle

PMTs are vacuum devices using photocathodes to convert photons into free electrons. Electrical fields are used to accelerate these electrons toward a set of dynodes. Dynodes multiply the single electron to form electron packets, which are collected by the anode. The anode current is made available externally and is connected to fast charge amplifiers. Figure 5.23 illustrated the working principle of a PMT.

5.4.2 Applications

Applications are divided into applications using a scintillator and nonscintillator applications. The difference is that an event in a nonscintillator application consists of a single photoelectron. Here, the QE of the photocathode plays a very important role: only the QE fraction of events can be detected.

Many PMT applications are scintillator based. They are used to detect highly energetic particles such as gamma photons or positrons. The energy of these particles will create many photons in the scintillator. Even with QE $\ll$ 100% every event will be detected by the PMT and the resulting pulse is the sum of the pulses of many individual photoelectrons created at the same time. In this case, a dark current pulse has a much lower energy and can easily be suppressed by thresholding.

The largest PMT market for scintillator-based applications is nuclear medicine with PET and SPECT cameras (Fig. 5.24). Well-known industrial applications include nondestructive testing, thickness, and density measurement and oil well logging. PMTs are also used extensively in high energy physics and as cosmic ray detectors.

PMTs without scintillators are used for analytical applications such as luminescence detection and spectroscopy. Many scientific Cherenkov experiments use PMTs to analyze the Cherenkov radiation from either high energy physics or astrophysics experiments.

Fig. 5.24 A modern PET scanner using PMTs

5.4.3 The Components of a PMT

5.4.3.1 Tube and Photocathode

Many PMTs have a photocathode with a high blue sensitivity. In this case, photocathodes are unusually optimized for the specific wavelength of the scintillator. The working principle of photocathodes has been explained in Sect. 5.2. In most cases, the photocathode is manufactured inside the PMT tube. Antimony is preevaporated on the input window and the PMT contains dispensers for the alkali metals.

5.4.3.2 Dynodes and Anode and Power Supply

The surface of dynodes has to be made of a material with a high secondary yield. The material mainly used is Antimony (Sb), which is activated by the alkali metals used to form the photocathode. The electro-optics has to be designed in such a way that the photoelectrons are accelerated to the first dynode, where they can secondary electrons by impact ionization. The secondaries created at the first dynode are accelerated to the second dynode to again create secondaries there.

Primary design considerations include the following optimizations:

- The secondary electrons must all be accelerated to the next dynode in the chain and not skip one.
- The time required for the electron to travel from the photocathode to the first dynode or from one dynode to the next should be independent from the location of creation, as far as possible.
- Ion feedback must be prevented and their effects must be minimized.

Depending on the application, there are different dynode configurations.

Type of dynode	Main characteristics
Venetian blind dynodes	Good collection efficiency, good gain stability but mediocre timing characteristics
Box dynodes	Good collection efficiency, bad timing characteristics
Linear focusing dynodes	Fast timing but compromised collection efficiency
Circular cage dynodes	High compactness
Mesh dynodes	Optimized for operation in magnetic field
Foil dynodes	Optimized for operation in magnetic field

The anode has the function to collect all the secondary electrons emitted by the last dynode. The geometry must be optimized to minimize space charge effects for highly linear response. The anode should also match the characteristic impedance of the output connection. The best solution is usually found using a grid between the second-last and last dynode.

5.4.3.3 Special PMTs

There are a number of special PMTs worthwhile mentioning here: Multipixel PMTs, MCP-based PMTs, and hybrid PMTs.

Multipixel PMTs are devices where multiple dynode and anode structures are positioned in a single enclosure. The multiple channels then also include the same photocathode. Main advantage is the better sensitive area ratio.

MCP based PMTs: There are also PMTs, which do not use a dynode structure but MCP stacks as an electron multiplier. There are several advantages. First, because of the small dimensions and high field strengths, the MCP-based PMTs are very fast and have small timing differences. Second, they are relatively insensitive to magnetic fields. Third, it is very easy to incorporate many pixels in one device. Figure 5.25 shows an example of such an MCP-based PMT, the Planacon of Photonis Inc. with an 8×8 anode array, see also reference [7].

Hybrid PMT: A third special type PMT is the hybrid PMT [8]. The gain of such a PMT is achieved by hurling the photoelectron with high energy into a Silicon PIN diode. The fast electron will create several electron hole pairs that result in a measurable charge pulse. This type of detector is successfully applied in the LHC-B Rich detector of CERN, as illustrated in Fig. 5.26.

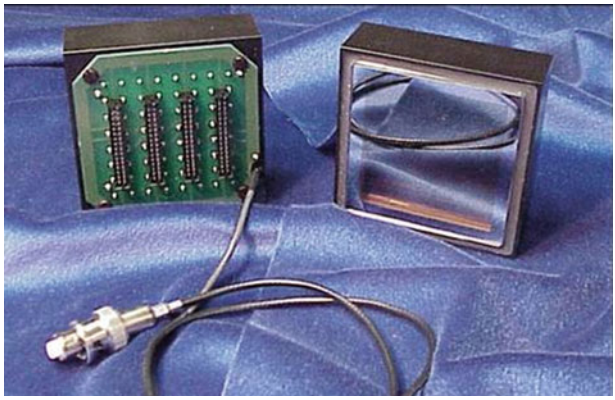

Fig. 5.25 Planacon, an MCP-based PMT

Fig. 5.26 CERN's LHCb-Rich detector using hybrid PMTs

5.4.4 Performance Characteristics

5.4.4.1 Quantum Efficiency and Dark Current

The achievable quantum efficiency of the photocathode depends very much on the required wavelength. PMTs used with scintillators can be optimized for the emissive wavelength of the scintillator. These S20 or bi-alkali photocathodes have a QE of up to 45%. In the red and NIR wavelength range, GaAs photocathodes are superior.

As explained in Sect. 5.2, the dark current of PEA photocathodes is extremely low. A bialkali photocathode has a typical dark current density of $2 \times 10^{-18}\,\mathrm{A/cm^2}$ and an S20 photocathode of $10^{-16}\,\mathrm{A/cm^2}$. The bialkali dark current density corresponds to about 15 counts/($\mathrm{cm^2\,s}$).

5.4.4.2 Gain and Dynamic Range

The gain (M) is defined as the number of electrons collected by the anode per incident photoelectron. It depends on the secondary yield of the dynode (η_d) and the number of dynodes n_d as:

$$G = \eta_d^{n_d}. \tag{5.5}$$

Taking for example $\eta_d = 4$ and $n_d = 10$, then $G = 4^{10} = 10^6$ results. This gain equation also shows the importance of a high secondary yield of the dynodes. For a material having $\eta_d = 2.5$, the gain of the same 10-stage PMT would drop two orders of magnitude to 9,500.

The gain achieved by many smaller gain steps has the drawback of gain variation. The distribution showing the relative frequency of pulse height is called single electron spectrum (SES). From event to event, there is a spread of gain caused by the first few gain steps. This phenomena is identical to the one described for the MCP gain earlier in this chapter. To achieve a constant gain per event, it is crucial to use a first dynode with extra high secondary yield. The best known material is diamond with a secondary yield of about 10. For multiphoton events, the gain characteristics become more favorable; the resulting event can be thought to be built up of more individual events. Therefore, the variation decreases with the square root of the number of photoelectrons.

5.4.4.3 Timing Characteristics

Single photon: A single photoelectron will create a pulse as illustrated in Fig. 5.27. It takes the transient time (tr) after the initial photoelectron until the pulse arrives at the anode.

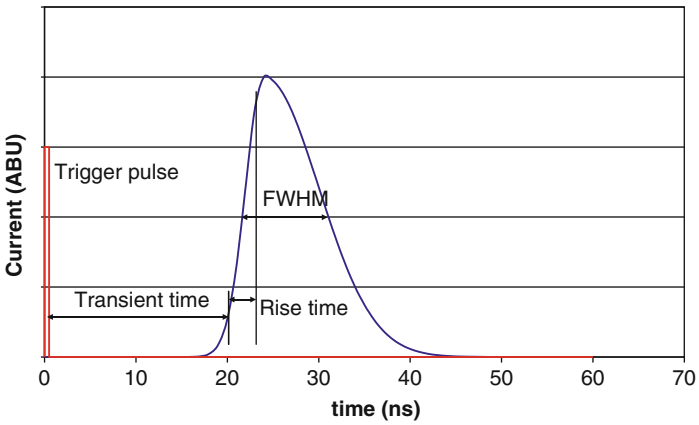

Fig. 5.27 Timing characteristics of a typical PMT pulse. The rise time and FWHM are chosen larger than in reality for reasons of clarity of the drawing

The transient time is caused by the travel time of the electron from photocathode to the first dynode and then from dynode to dynode. The pulse-to-pulse variation of the transient time is called jitter. The transient time is also dependent on the exact position on the anode. This is described by the "center to edge transient time difference." The rise time and pulse width are caused by the different electron trajectories as well as by the energy distribution of secondary electrons.

Characteristics	Type value (ns)
Transient time	20–40
Jitter	0.2–0.3
Center to edge difference	1–2
Rise time	0.8
FWHM	4–5

Multiple photon events: If a scintillator creates many photoelectrons from the same interaction event, the created pulse becomes the sum of many individual pulses. This has no influence on the average transient time. The time jitter and gain variation is reduced with a factor of the square root of the number of electrons. The time jitter of the combined single-photon events is now called "timing resolution." The contributors of the timing resolution are the timing jitter and the center-to-edge timing difference. A burst of photons will impact all over the photocathode, some at the edge and some at the center.

The gain variation of the summed pulse is called energy resolution. Both timing resolution and energy resolution will improve with the square root of the number of photoelectrons. These parameters will benefit, therefore, from a high DQE of the photocathode and the number of photons created by the scintillator (Fig. 5.28).

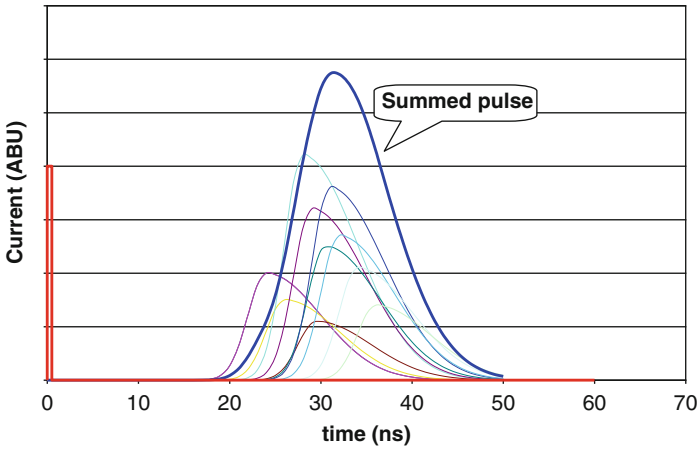

Fig. 5.28 Principle of the multiphoton events. The sum changes the shape of the pulse but reduces jitter and gain variation. For didactic purposes, the various effects are not drawn on a realistic scale

5.4.4.4 Comparison with Solid-State Detectors

Avalanche diodes and arrays of parallel-connected APDs, a so-called silicon-photomultiplier (SiPM), are competing with PMTs in several areas. Because their main characteristics are rather different, the choice will depend on the specific application.

Advantages of APDs are the faster timing and lower jitter, higher DQE, better energy resolution, coverage of longer wavelength, and smaller dimensions. The advantages of PMTs are several orders of magnitude lower dark current, higher gain, larger sensitive areas, lower cost per cm^2, and lower temperature sensitivity.

5.5 Conclusions and Outlook

Image intensifiers and PMTs are clearly not dinosaurs that will be wiped out any time soon by solid-state devices. The unique feature of the photocathode is its extremely low dark current. This dark current is from four orders of magnitude (S25, GaAs) to eight orders of magnitude (multialkali) lower than that of CMOS imagers. Also, the very low power consumption of image intensifiers and the possibility to make large area detectors in a price-competitive way for PMTs are factors keeping these vacuum-based devices attractive. The demands of the individual application will determine the best choice of detector, whether it should either be solid-state or photocathode based.

References

1. I.P. Csorba, *Image Tubes* (Howard W. Sams & Co, Inc., Indianapolis, IN, 1985)
2. E.H. Eberhardt, Gain model for micro channel plates, Appl. Opt. **18**, 9 (1979)
3. *RCA Electro-Optics Handbook*, Technical Series EOH-11, *Open Libraries* OL5246811M, www.burle.com/sgi/byteserver.pl/pdf/electro_optics.pdf (1974)
4. T.H. Hoenderken, C.W. Hagen, J.E. Barth, P. Kruit, G.O. Nutzel, Influence of the microchannel plate and anode gap parameters on the spatial resolution of an image intensifier, J. Vac. Sci. Technol. B Microelectronics Nanometer Struct. **19**, 843 (2001)
5. O. Siegmund, J. Vallerga, B. Welsh, A. Tremsin, J. McPhate, Imaging photon counting detectors for high time resolution astronomy, High Time Resolution Astrophysics, Astrophysics and Space Science Library **351**, 327 (2008)
6. S.O. Flyckt, C. Marmonier *Photomultiplier Tubes – Principles and Applications* (Photonis Technologies S.A., France, 2002), www.burle.com/sgi/byteserver.pl/pdf/photo.pdf
7. S. Korpara, I. Adachic, R. Dolenecb, P. Križan, R. Pestotnikb, A. Stanovnike, Photonics MCP PMT as a light sensor for the Belle II RICH, Nucl. Instrum. Methods Phys. Res. A Accelerators, Spectrometers, Detectors and Associated Equipment (in print, 2011)
8. G. Aglieri Rinellaa, T. Gysa, D. Piedigrossia, A. Van Lysebettena, Performance studies of pixel hybrid photon detectors for the LHCb RICH counters, Nucl. Phys. B Proc. Suppl. **150**, 285 (2006)

Chapter 6
Electron-Multiplying Charge Coupled Devices – EMCCDs

Mark Stanford Robbins

Abstract Charge coupled devices utilising electron multiplication (EMCCDs) have been available commercially for almost a decade. Currently they are the image sensors of choice for a wide range of applications requiring ultra high sensitivity whilst maintaining high data rates. Even at low data rates, these can yield benefits for single photon imaging. This chapter reviews the technology behind the EMCCD that is particularly relevant for very low light applications. Characteristics of the EMCCD devices are discussed, including the detailed behaviour of the multiplication gain. The various noise sources are examined, including the noise introduced by the multiplication process and that originating from the generation of background signal. The conventional use of signal to noise as a figure of merit for an EMCCD device is reviewed and it is shown that this can mask the benefit of the EMCCD, especially when the background signals are low. Photon counting techniques have been found to improve performance under some circumstances and the different approaches employed to date are covered. Device structures for optimising the signal generation are discussed and show that devices capable of almost ideal sensitivity are a reality.

6.1 Introduction

Charge coupled devices (CCDs) utilising electron multiplication (EMCCDs) have been available commercially for almost a decade. Currently, they are the image sensors of choice for a wide range of applications requiring ultra high sensitivity, whilst maintaining high data rates. Even at low data rates, these can yield benefits for single photon imaging. This chapter reviews the technology behind the EMCCD that is particularly relevant for very low light applications, including the characteristics of the multiplication gain and noise.

P. Seitz and A.J.P. Theuwissen (eds.), *Single-Photon Imaging*,
Springer Series in Optical Sciences 160, DOI 10.1007/978-3-642-18443-7_6,

6.2 Harnessing Impact Ionisation for Ultra Sensitive CCD Imaging

Madan et al. published the first experimental observation of charge multiplication in CCDs as early as 1983 [1]. The main thrust of this work was to investigate the maximum practical clock amplitudes that could be applied to conventional CCDs, but it was noted also that the avalanche multiplication observed may be used to measure small signals. It was not until 1988 that workers at MIT published an attempt at measuring the noise of the multiplication process [2]. Although improvement in post-detection noise was reported, difficulties were encountered with spurious charge generation. Hynecek also showed charge multiplication from a regular virtual phase device [3] and the term charge carrier multiplier (CCM) was coined to describe the use of a transfer barrier to establish the high field before charge transfer takes place. The year 2001 saw the first papers describing practical devices using a multiplication register from e2v technologies [4] and Texas Instruments [5]. Early comparison of EMCCDs with image intensified CCDs (ICCDs) [6] showed the fully solid-state technology as a creditable alternative to the intensified device for non-gated applications, especially as the EMCCD also possesses the benefits of the conventional CCD, such as higher quantum efficiency and better resolution. Since these early developments a wide range of applications have benefitted from the characteristics of the EMCCD, from surveillance applications including low light colour imaging [7, 8] through to high speed, single photon imaging in the life sciences [9] and adaptive optics applications [10] and Lucky Imaging [11] in astronomy, for example. The EMCCD has also found application in the imaging of single, high energy photons, for example in SPECT systems [12, 13], and the analysis of the light generated by individual photons interacting in a scintillator has been of particular interest [14, 15].

6.3 The Electron Multiplying CCD Concept

This section provides an overview of the operation of EMCCD devices starting with a review of the limitations imposed on conventional devices by the charge to voltage conversion process.

6.3.1 Output Amplifier Noise

Photon-generated electrons have to be converted to a signal suitable for transfer through a conventional video chain. This is achieved by a charge to voltage conversion process. The on-chip circuitry generally consists of one or more AC or DC coupled source-follower stages. A typical, single stage circuit is shown in Fig. 6.1. Prior to the transfer of signal to the output node, the potential of the node is

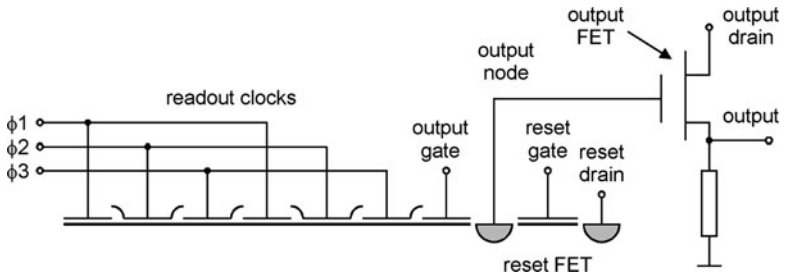

Fig. 6.1 A schematic of a typical single stage CCD output circuit

reset to the DC level on the reset drain by turning the reset FET on. The reset FET is then turned off and signal charge is transferred to the node. Johnson noise from the channel of the reset FET in the on state introduces uncertainty in the potential to which the node is reset. This is known as kTC noise and can be eliminated using analogue or digital signal processing techniques such as correlated double sampling (CDS). CDS involves external circuitry to sample the output after reset and again with signal present. Subtraction of the values removes the common reset component. In a well-designed system, once kTC noise is eliminated the noise from the charge to voltage conversion process is dominated by the $1/f$ and white noise from the first-stage source-follower. A full analysis of the effect of various sampling approaches is beyond the scope of this chapter but is covered in [16]. As a useful approximation for clamp and sample CDS with an optimum pre-sampling filter, an effective noise bandwidth can be defined as being between $f_c/2$ and $3f_c/2$, where f_c is the CCD read-out frequency. Integrating the transistor noise between these frequencies gives the noise equivalent signal (r.m.s. electrons) as:

$$\text{NES} \approx 2e_w f_0^{1/2}(1 + f_c/f_0)^{1/2}/R,$$

where e_w is the white noise spectral density ($\mu V/Hz^{0.5}$), f_c is the clocking frequency (Hz), f_0 is the frequency (Hz) where the spectral density of the $1/f$ noise is equal to the white noise and R is the output circuit responsivity (μV/electron). The white noise and the responsivity both depend on the size of the transistor. Taking these dependencies into account for an optimal design gives us [17]:

$$\text{NES} \propto (1 + f_c/f_0)^{1/2}/\sqrt{R}, \tag{6.1}$$

where the constant of proportionality is technology dependent but independent of transistor size. Although this relationship is only approximate, it does agree with practical measurements as shown in Fig. 6.2.

The lowest noise floor achievable with these devices is around two electrons r.m.s. but for this the read-out rate has to be below a few tens of kilohertz [17]. As the read-out rate increases so does the noise. For applications requiring video

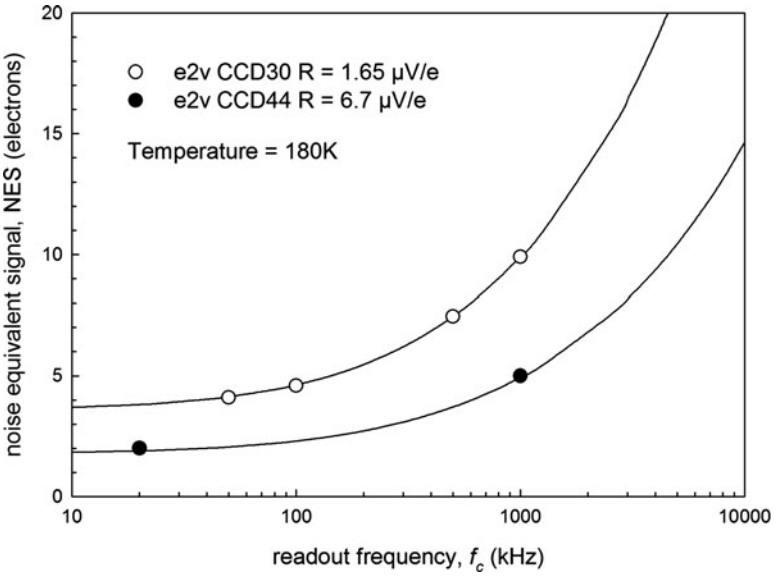

Fig. 6.2 The noise equivalent signal from two scientific CCDs having different output circuit responsivities. Clamp and sample CDS is used with optimum pre-sample filtering. The points are measured data and the *solid lines* are calculated using (6.1)

frame rates using a single output, for example, the pixel rates may have to be above 10 MHz, leading to noise in excess of ten electrons, even for technologies optimised for speed. Various techniques can be used to reduce the noise, for example the use of multiple nondestructive samples, but these tend to reduce the frame rates achievable. Another option is to use multiple outputs to reduce the data rate per output for a given frame rate, and thus reduce the noise. Unfortunately, the noise reduction attainable is insufficient and true single photon imaging using conventional CCD structures is not practical.

6.3.2 The Use of Multiplication Gain

To enable single photons to be detected, the signal has to be raised above the noise of the video chain, including the on-chip and off-chip components. This is achieved in the EMCCD by increasing the signal through impact ionisation. The imaging structure of the device is the same as a conventional CCD. An additional serial register, the multiplication register, is included before the charge to voltage conversion circuitry, as shown schematically in Fig. 6.3. The multiplication register is made up of a number of multiplication elements, each element providing a small amount of gain. A schematic view of a multiplication element is shown in

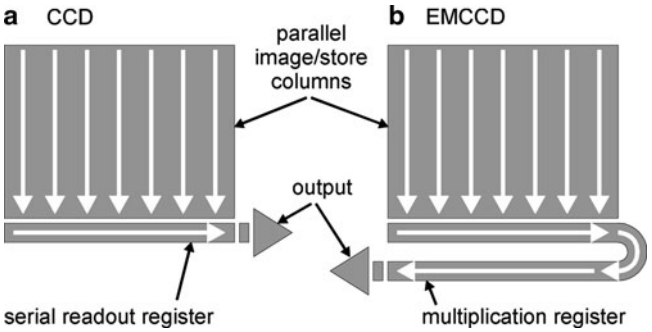

Fig. 6.3 A schematic comparison between the layout of a conventional and an electron multiplying CCD showing the addition of the multiplication register

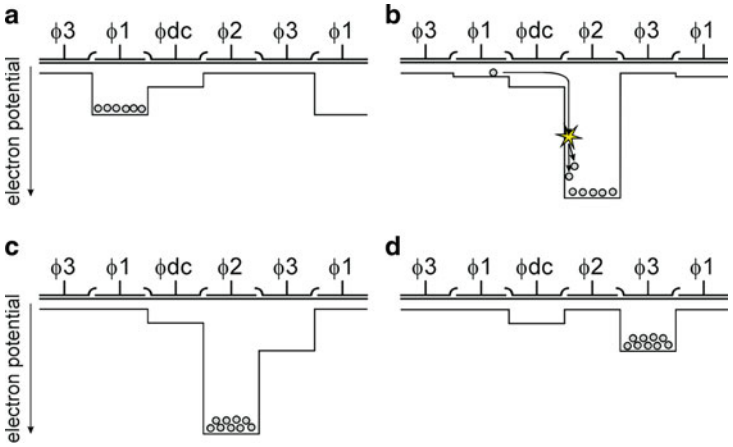

Fig. 6.4 A schematic of signal charge being transferred through a multiplication element, comprising three clocked phases, $\phi 1$, $\phi 2$ and $\phi 3$ and a dc biased phase, ϕdc

Fig. 6.4. Signal charge enters the element through the first phase, $\phi 1$. A high field is established between the high voltage phase, $\phi 2$, and an adjacent phase held at a constant DC level, ϕdc, by the application of a few tens of volts potential difference. The charge under the first phase is then clocked over the DC phase, through the high field and into the HV phase where it will be clocked out, and into next element, via $\phi 3$. Typically, 400–600 multiplication elements are employed in series. A useful 1,000× mean gain can be achieved with 464 elements, each providing a 1.5% probability of an impact ionisation event occurring as every electron passes through. This multiplication gain can be applied at high data rates and operation up to a clock frequency of 35 MHz has been demonstrated [18].

The probability of an impact ionisation event occurring is dependent on the field experienced by the signal electrons and thus the bias applied between the high voltage phase and the DC phase. The relationship between gain and potential

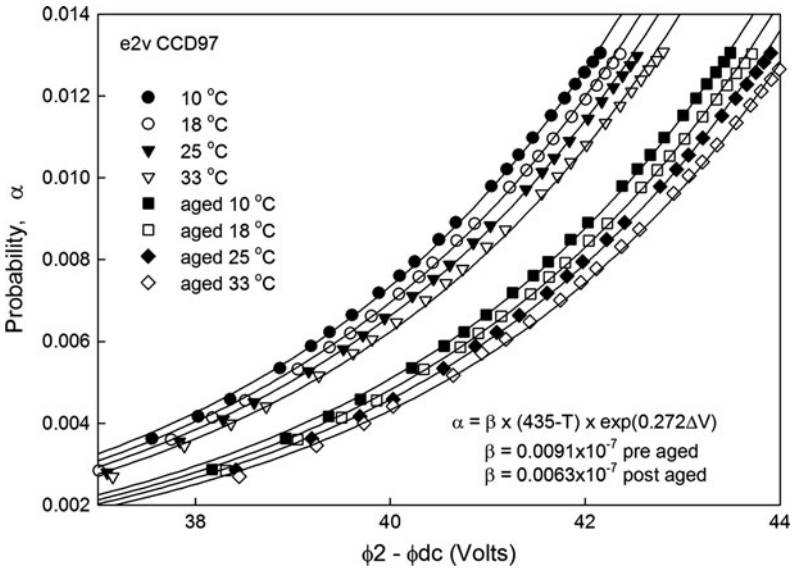

Fig. 6.5 Typical gain characteristics of an e2v CCD97 showing the probability of an impact ionization event as a function of clock level and temperature. The aged curves are taken following operation with high gain whilst experiencing very high levels of illumination

is dependent on the architecture and technology used. However, devices currently available produce a mean gain per multiplication element given empirically by the following relationship:

$$g = 1 + \alpha \approx 1 + \beta(B - T)\exp(C\Delta V)),$$

where α is the probability of an impact ionisation event occurring, T is the temperature, ΔV is the potential difference between the high voltage and DC phase at the time of charge transfer, β is a function of time and accounts for any gain change due to charging of the gate dielectric, for example, and B and C are constants. The total mean gain, G is given by:

$$G = g^n \approx \exp(n\beta(B - T)\exp(C\Delta V)),$$

where n is the total number of elements in the multiplication register. As an example, the measured value of α for an e2v CCD97 device is presented in Fig. 6.5. Here, B is 435 K and C is 0.272 V^{-1}. The TC285 device from Texas instruments has a similar value of B but C is around 0.49 V^{-1}. Thus, the same gain as in the e2v device can be achieved but with lower clock amplitudes.

6.3.3 Noise and Signal-to-Noise Ratio

The noise on the output signal from an EMCCD includes the output amplifier noise, σ_A, discussed in Sect. 6.3.1, the Poissonian shot noise on the photon generated signal, shot noise on the dark signal and other spurious signal sources, all modified by the multiplication gain process. Referring the r.m.s. noise equivalent signal to the input of the multiplication register, and just considering dark signal generated in the image and store regions of the device, gives us:

$$\text{NES} = \sqrt{\left(F^2(S+D) + \frac{\sigma_A^2}{G^2}\right)},$$

where S is the mean number of photon-generated signal electrons and D is the mean number of dark signal electrons. As the multiplication gain is increased, the output amplifier noise can become insignificant, leaving the shot noise on the signal and dark signal, which is increased by the multiplication process. The increase is described by the excess noise factor, F. Various attempts to measure F have been published. The first measurement presented gave a noise factor close to unity [3], which would mean that an EMCCD operates almost as the ideal photon imager. Unfortunately, it appears that this was in error and a later measurement [19] presented $F^2 \sim 2.2$. The devices characterised in [20] behaved as ideal staircase avalanche photodiodes, the noise factor being caused by the stochastic nature of the gain process alone. In this case, the noise factor squared was found to be given by:

$$F^2 = 2(G-1)G^{-(1+1/n)} + 1/G. \tag{6.2}$$

For a high gain and large number of multiplication stages, this reduces to a value $F^2 = 2$. The experimental data presented in [20] together with the calculated values using (6.2) are shown in Fig. 6.6. Similar experimental results have been obtained subsequently [7, 8, 21] and measurements at different operating temperatures [7, 18, 20] have shown no significant temperature dependence.

The signal-to-noise ratio from an EMCCD running at high gain will be greater than from a conventional CCD only when the signal level is lower than the square of the CCD amplifier noise. Hence, it has been suggested that in some applications, where the signal levels are above this value, a low noise conventional device may be a better choice than the EMCCD [22]. This is clear for applications where the contrast in the scene being viewed is low. For high contrast scenes, for example where there is little background signal, the use of signal over noise on the signal as a figure of merit masks the benefit of the EMCCD, as the noise on the background can be much lower than the noise on the signal. This has been shown by a direct practical comparison between the CCD and EMCCD over a range of signal levels, presented by Tulloch [23]. A comparison was made by imaging a test chart using an e2v CCD97 having both an EMCCD output and a conventional low noise scientific

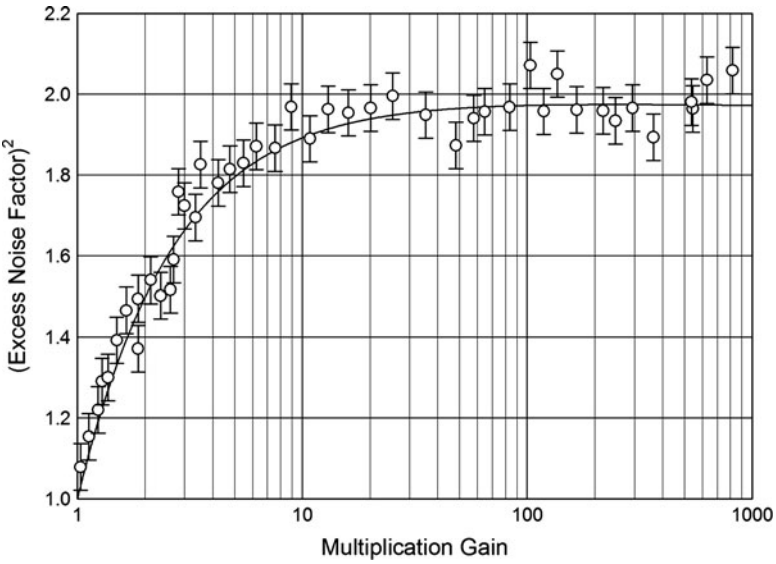

Fig. 6.6 The excess noise factor as a function of multiplication gain for a 536 multiplication element device characterised in [20]. The *solid curve* is calculated using (6.2)

output with an amplifier noise of 3.3 r.m.s. electrons. If the signal to noise is used as a measure, it would be expected that the CCD will outperform the EMCCD at signals above around 11 electrons. However, qualitatively it was observed that the level of detail revealed using the conventional output was equal to that from the EMCCD output only once the mean signal in the bar pattern reached around 22 electrons/pixel for zero background signal. At signal levels lower than this, the use of the EMCCD output was beneficial.

This is illustrated also in the results of simulations shown in Fig. 6.7, together with the effect of background signal. Here, the amplifier noise from a theoretical CCD was set at 1 r.m.s. electron and the mean signal in the bars above background was one electron. At a mean background signal of one electron, the detail in the image is about the same, whether a CCD or EMCCD is used. However, at zero background signal the benefit of the charge multiplication is observed clearly.

6.3.4 Output Signal Distributions

The noise factor gives a measure of the increase in width of the signal distribution from the multiplication process but it does not describe the shape of the output distribution. For this, a more detailed numerical or analytical approach is required. Approximate analytical equations for the output distributions assuming

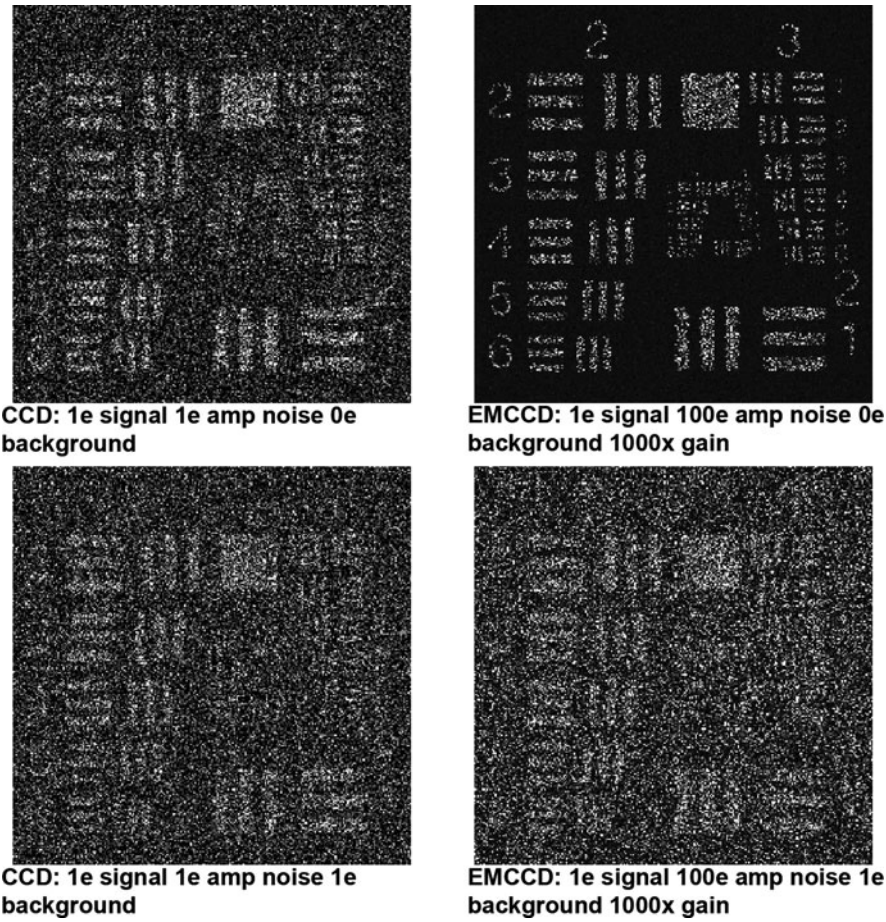

Fig. 6.7 The result of Monte Carlo simulations comparing an ideal CCD with an EMCCD for a mean signal of one electron on a background of zero and one electron

zero amplifier noise have been presented by Basden et al. [24]. For a small integer number of input electrons, n, the output distribution is well described by:

$$P(x) = \frac{x^{n-1}\exp(-x/G)}{G^n(n-1)!},$$

where x is the output signal. If the input signal is distributed according to Poisson statistics with mean μ, the output distribution becomes:

$$P(x,\mu) = \sum_{n=1}^{\infty} \frac{\exp(-\mu - x/G)\mu^n (x/G)^{n-1}}{G(n-1)!n!}. \tag{6.3}$$

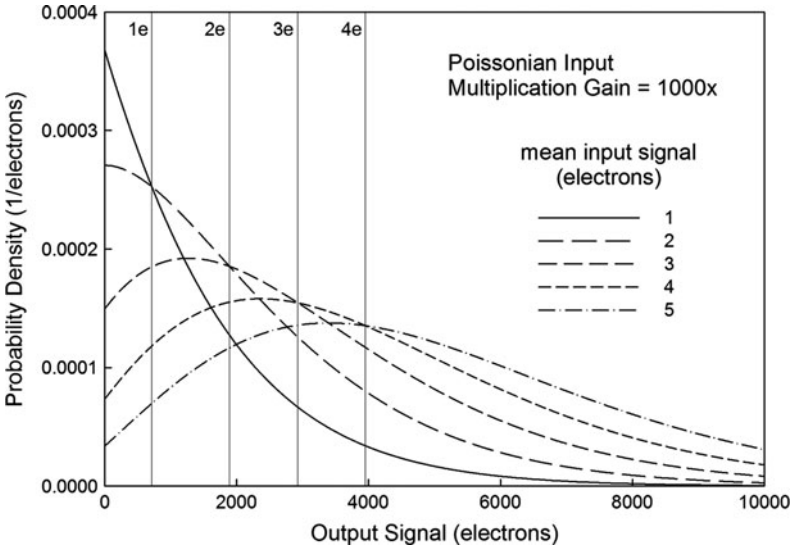

Fig. 6.8 The output signal distribution calculated using (6.3). The *vertical lines* indicate the thresholds used for the multiple-threshold photon counting scheme (Sect. 6.4)

This is plotted in Fig. 6.8 which shows that there is significant uncertainty in determining the exact input signal size for a measured output.

To illustrate this further, Fig. 6.9 shows the result of a simulation comparing conventional devices with the EMCCD. Here, a mean signal of one electron has been generated and is distributed according to Poisson statistics. A theoretical conventional device with an amplifier noise of 0.2 r.m.s. electrons reveals this distribution and the individual electrons appearing at the input can be distinguished. The output signal is virtually lost in the noise if the amplifier noise is increased to 1 r.m.s. electron. An EMCCD with a mean gain of 1,000× raises the input signal above the amplifier noise, even with an amplifier noise of 200 r.m.s. electrons used in this example. Although the effective input referred amplifier noise has been reduced to 0.2 r.m.s. electrons, it is clear the multiplication process makes the evaluation of the exact number of electrons at the input uncertain. However, the mean signal levels can be determined and, at low fluxes, the width of the distribution of output signals can be narrowed using photon counting techniques, described in Sect. 6.4.

6.4 Photon Counting with the EMCCD

The signal-to-noise ratio of an ideal shot noise limited imager, having zero amplifier noise, is simply equal to the square root of the number of electrons generated. The EMCCD noise factor reduces this ratio by $2^{1/2}$ which, from a signal to noise viewpoint, is equivalent to halving the signal size in the ideal detector. However, a

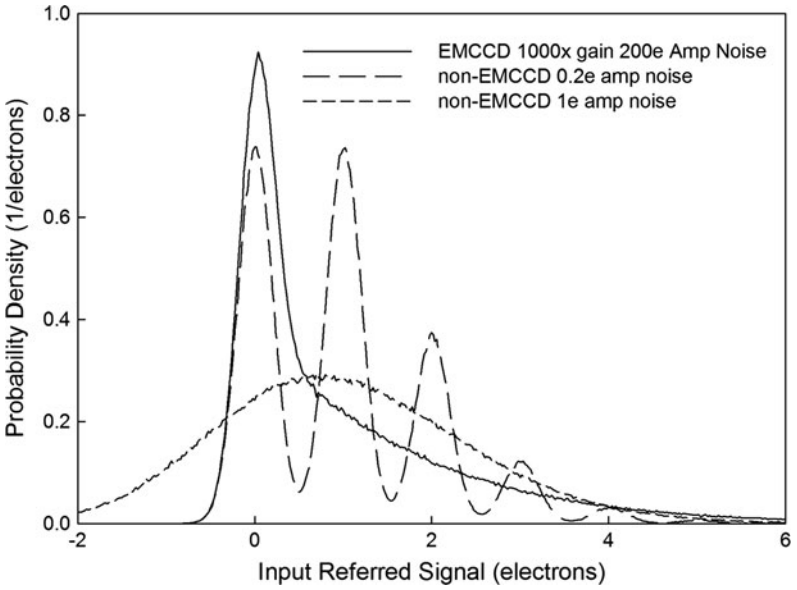

Fig. 6.9 The results of a Monte Carlo simulation of the output signal distributions for a mean input signal of one electron. The EMCCD here has an effective input referred amplifier noise of 0.2 r.m.s. electrons

reduction in the noise factor towards the shot noise limit can be achieved by the use of photon counting techniques and can be useful in photometric applications.

The simple, single-threshold scheme has been investigated by Wen et al. [25] for possible use in astronomy. Here, a single threshold on the output is set, above which the output is counted as originating from a single electron. The threshold is chosen to be high enough so false hits are not triggered by the amplifier noise. They demonstrated that the EMCCD can function as a photon counter in their low flux application if the background signal is sufficiently low. This photon counting technique had been studied theoretically by Basden et al. [24], who also investigated a second method utilising multiple thresholds to decide on the most likely input signal. The thresholds proposed are shown in Fig. 6.8.

The result of this processing on the excess noise factor is shown in Fig. 6.10. The noise factor for the single threshold scheme drops at mean signal levels above about 0.1 electrons/pixel due to the increasingly significant possibility of generating more than one electron in a pixel. Below this signal level, the single-threshold scheme virtually eliminates the noise associated with the multiplication process. The multiple-threshold scheme does have some benefit at signal levels below about ten electrons but has no effect at higher signals. The use of multiple thresholds requires a correction to the calculated photon flux. A simple correction was proposed by Basden but a Bayesian estimation technique [26] has been proposed to improve the photometric accuracy.

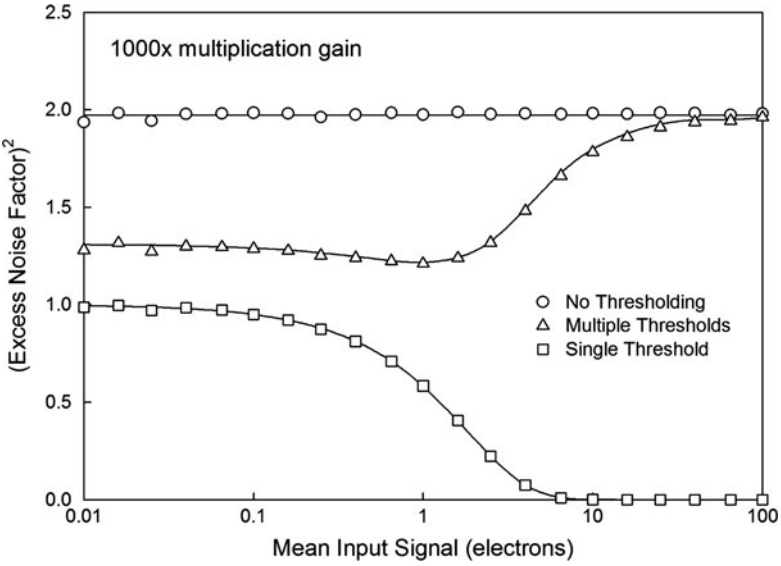

Fig. 6.10 The excess noise factor calculated following various photon counting schemes applied to a Monte Carlo simulation of the EMCCD. The *solid curves* result from an appropriate numerical integration of (6.3)

6.5 Background Signal Generation

Signal generated in the absence of illumination ultimately limits the sensitivity of the EMCCD. With appropriate optimisation, the background signals can be reduced to extremely low levels, as illustrated in Fig. 6.11. The background signal consists of the conventional dark signal together with spurious signal generated through the clocking process. The details of these contributions are discussed in this section.

6.5.1 Dark Signal

The generation of dark signal in CCDs has been covered elsewhere, for example by Janesick [27]. Features particularly relevant to EMCCDs are summarised here. Dark signal originates from the diffusion of electrons from the neutral semiconductor regions, the thermal generation from the bulk silicon in the depleted space charge regions and thermal generation from the silicon surface at the silicon/silicon dioxide interfaces. Thermally generated components dominate below room temperature, and the highest density of so-called generation centres occurs at the silicon surface. The surface dark signal originating from this source is in the order of 1 nA/cm^2 at 20°C and has a temperature dependence approximately given by:

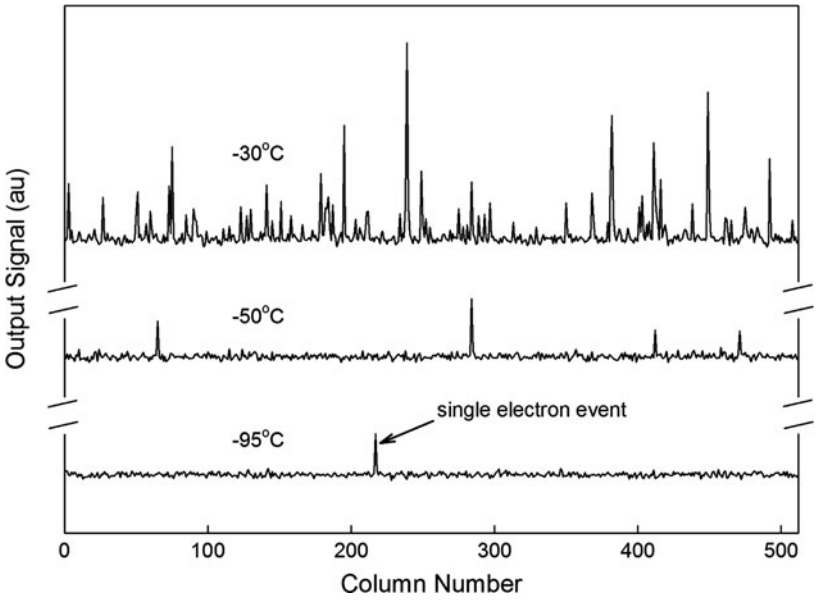

Fig. 6.11 The dark output signal from a row of an e2v CCD97 512 × 512 pixel device operating at 35 frames/s. Note the single electron events are visible above the very low noise background. (Data courtesy of Andor Technology)

$$I_{\text{surface}} \propto T^3 \exp(-E_g/2kT),$$

where T is the temperature, E_g is the silicon band gap and k is the Boltzmann constant. A component of the total dark signal originates from the bulk silicon. The observed mean generation rate from the bulk is a combination of the generation rates from the different species of defects present. For e2v devices, at temperatures above about −50°C, an effective combined bulk dark signal temperature dependence of

$$I_{\text{bulk}} \propto T^2 \exp(-0.78\,\text{eV/kT})$$

is observed. The magnitude of this bulk dark signal is typically over two orders of magnitude lower than the surface dark signal at room temperature. Typical bulk and surface dark signal densities as a function of temperature are shown in Fig. 6.12. The total dark signal can be reduced significantly by cooling the device. However, unless the frame rates are very high, or very deep cooling is available, the surface dark signal has to be suppressed for single photon imaging. This is achieved by flooding the surface with holes by increasing the substrate to gate potential beyond the pinning point, as described by Saks [28]. In this situation, the device is said to operate in inverted mode and the surface dark signal is eliminated, leaving the bulk component. This pinning of the entire surface removes the signal localisation otherwise achievable through the gate biasing. Therefore,

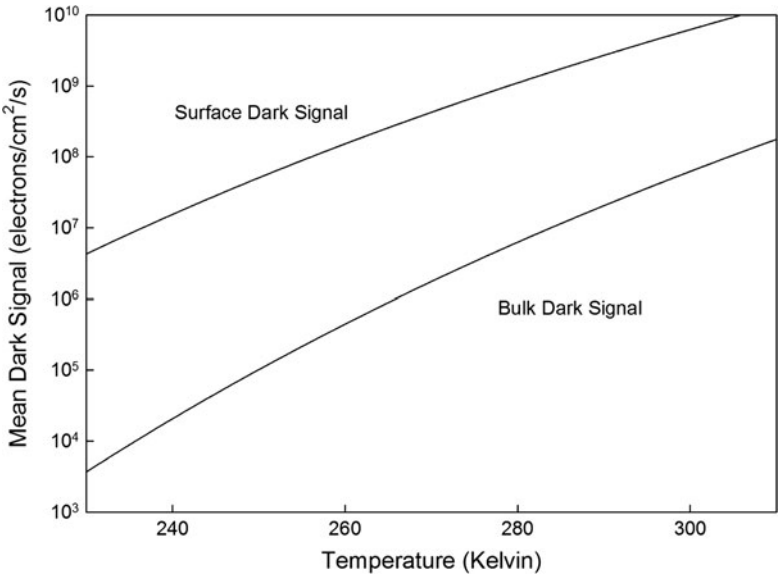

Fig. 6.12 Typical temperature dependencies of bulk and surface dark signal

devices intended to operate in inverted mode need additional implants to ensure that charge can be confined during integration. An inverted mode device, having a pixel size of $16\,\mu\text{m} \times 16\,\mu\text{m}$, will generate around 100 electrons/pixel/s at room temperature, reducing to around 1 electron/pixel/s if the device is cooled to −20°C.

It is clear that inverted mode operation can be of great benefit to EMCCDs. However, free holes present at the surface of the multiplication register would be multiplied together with signal electrons, thus producing extreme levels of unwanted background. In practice, the image and storage areas tend to be operated in inverted mode to suppress the surface dark signal from these regions, and the multiplication register functions with the surface depleted of mobile charge. Therefore, as the signal is clocked through each stage of the multiplication register, dark signal will be acquired from the surface and the bulk. However, it is only the dark signal collected from the first elements of the multiplication register that will experience the full gain applied to the signal. Dark signal collected from multiplication elements close to the output amplifier will experience very little gain. If N dark signal electrons are generated within a multiplication register element during a time $1/f$, where f is the readout frequency, then the mean input referred dark signal is given by:

$$S_{\text{in}} = \frac{N}{G}\left(\frac{G-1}{g-1}\right) \approx \frac{nN}{\ln(G)}$$

assuming a uniform gain per stage along the register. A multiplication register, having 536 multiplication elements of size $16\,\mu\text{m} \times 40\,\mu\text{m}$, operating at a frequency

of 15 MHz and 1,000× gain, will generate an effective mean input referred dark signal of around 0.2 electrons/pixel at room temperature. This is insignificant compared with the image section dark signal unless the integration times are very short, which may be the case in some applications.

6.5.2 Statistics of Dark Signal Generation

The dark signal is distributed according to Poisson statistics, modified by the multiplication process as described in Sect. 6.6. If the chance of getting more than one dark signal electron in a pixel is insignificant, the output signal distribution for dark signal generated outside of the multiplication register is given by the probability density:

$$P(x) = \frac{S_{\text{dark}}}{G} \exp\left(-\frac{x}{G}\right) \quad x > 0, \tag{6.4}$$

where S_{dark} is the mean number of dark signal electrons in the pixel. This describes the dark signal distribution when it is dominated by the image and store components and can be used to calibrate the gain using background signal alone if required. The dark signal from the multiplication register has a different distribution. Noting there is a greater contribution from the dark signal generated furthest from the output, and combining the distributions along the register gives:

$$P(x) \approx \frac{S_{\text{in}}}{n} \ln(G) \sum_{i=1}^{n} G^{-i/n} \exp(-xG^{-i/n}) \quad x > 0. \tag{6.5}$$

Equations (6.4) and (6.5) are plotted in Fig. 6.13 together with the results of a Monte Carlo simulation. The same mean input referred dark signal is used in both cases. The distribution of dark signal from the multiplication register is seen to be narrower than that from other areas of the device, indicating that the dark signal from the multiplication register can have less of an impact on the imaging performance compared with that implied by the mean dark signal value alone. A more detailed treatment of the statistics of signal generated within the multiplication register has been given in [29].

6.5.3 Spurious Charge Generation

Once the dark signal has been reduced to insignificant levels by cooling, for example, the ultimate sensitivity of the EMCCD is limited by other background signal components. In an optimally designed device, this spurious background is dominated by signal that is generated by holes undergoing impact ionisation. This occurs in the image and store regions of the device where the surface is taken in and out of inversion during the charge transfer process and is known generally as

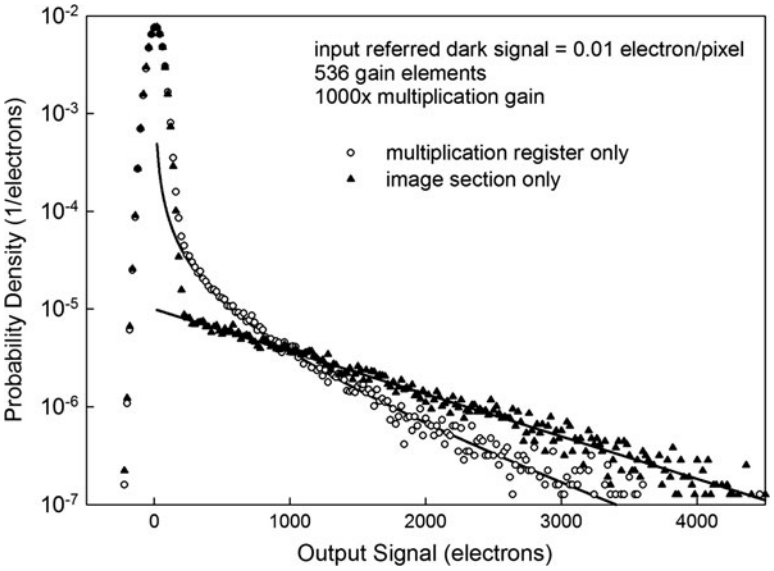

Fig. 6.13 Calculated dark signal distributions for an input referred dark signal of 0.01 electrons/pixel. The points result from Monte Carlo simulations including a 50 electron amplifier noise. The *solid curves* are calculated by the application of (6.4) and (6.5)

clock-induced charge. Other components can exist, including clock-induced charge from the multiplication register, although this can be confused with multiplication register dark signal [25]. Also, there will be a very low level of signal generated through the absorption of photons emitted by the multiplication process itself. Reports from early devices [12] showed the total spurious background signal to be as high as 3–4 electrons/pixel and would therefore make them unsuitable for single photon imaging. However, measurements made on the first devices aimed at scientific applications gave a significantly lower mean value of 1 spurious electron generated every 100–200 transfers [6, 30]. The level of clock-induced charge has been found to be virtually independent of temperature [31], contrary to the description of spurious charge in [27]. Careful optimisation of the clocks, by the control of the waveform shapes and maximising the parallel clocking rate for example, has yielded a spurious background signal as low as 0.001 electrons/pixel/frame measured on a 512 × 512 pixel e2v CCD97 device [32].

6.6 Improving the Efficiency of Signal Generation

The EMCCD technology overcomes the limitations imposed by the charge to voltage conversion process and various techniques can be used to reduce the unwanted background signal. To optimise the sensitivity of the devices, attention

must be turned also to improving the ability to generate signal electrons from the incident photons. EMCCDs benefit from the years of development applied to increasing the efficiency of signal generation in CCDs. As with CCDs, almost all of the pixel area of the EMCCD can be sensitive to the illumination, depending on the architecture employed. The efficiency of signal generation is dependent on the absorption and reflection characteristics of the incident photons throughout the structure. Light has to pass through, and is therefore partially absorbed by, the gate structure of conventional front illuminated devices before it can interact with the underlying silicon. The quantum efficiency (QE) is a measure of the efficiency of converting the incident photons to signal electrons. Typically, the peak QE for front illuminated devices is around 40–50%. Even with this non-ideal conversion, front illuminated EMCCDs have been found to outperform Gen II image intensifiers [33]. In some designs, a fraction of the electrode structure is removed, creating open electrode devices, thus reducing the unwanted absorption. Virtual phase devices also have improved quantum efficiency as there is no gate structure over the virtual phase gate electrode and a peak quantum efficiency of 65% for the virtual phase TI TC285 has been reported [18].

The most significant improvement in QE is obtained by eliminating the effect of absorption in the electrode layers entirely, through the use of back illumination. Here, input photons are incident on the back surface, with the device mounted face down on a substrate. The device is thinned to a thickness of a few tens of microns and treated to ensure almost all photoelectrons generated within the active depth are collected as signal. Thus, devices with QE values greater than 90% are obtained. The use of back illuminated EMCCDs for low light microscopy applications has been discussed by Coats et al. [9]. Sensitivity in the red and near infrared regions of the spectrum can be improved by increasing the thickness of the active silicon. Field-free regions have to be minimised to maintain resolution. This is achieved through the use of higher resistivity material than normally used to create a deeper depletion region. Deep depleted EMCCD devices, manufactured on 1,500 Ωcm silicon, are being evaluated as part of a programme to develop sensors optimised for astronomical adaptive optics applications [10]. The spectral performance of a range of EMCCD technologies is shown in Fig. 6.14. Future developments will see efficiencies increasing further over a broad spectral range.

6.7 Concluding Comments

In this chapter, the features of EMCCD sensors have been described. These make them ideally suited to many single-photon imaging applications and developments are ongoing to improve the sensitivity and operational characteristics further. The introduction of the EMCCD has made high-resolution and high-speed single photon imaging a reality. Commercial off the shelf devices are well established and many new applications are emerging requiring custom design and novel architectures.

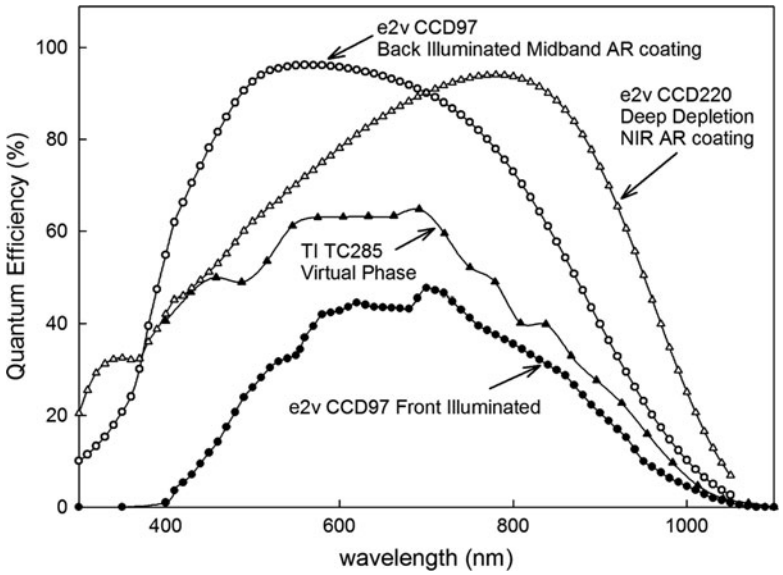

Fig. 6.14 The quantum efficiency obtainable for a range of EMCCD technologies

Acknowledgements The content of this chapter has benefited from the support of many people within the High Performance Imaging Solutions Group at e2v technologies. Special appreciation is due to David Burt for his valuable suggestions and enlightening discussions.

References

1. S.K. Madan, B. Bhaumik, J.M. Vasi, Experimental observation of avalanche multiplication in charge-coupled devices, IEEE Trans. Electron Devices **30**, 694–699 (1983)
2. S.A. Gajar, B.E. Burke, Charge amplification by impact ionization in charge-coupled devices, IEEE Trans. Electron Devices **35**, 2435–2436 (1988)
3. J. Hynecek, CCM – a new low-noise charge carrier multiplier suitable for detection of charge in small pixel CCD image sensors, IEEE Trans. Electron Devices **39**, 1972–1975 (1992)
4. P. Jerram et al., The LLLCCD: low light imaging without the need for an intensifier, Proc. SPIE **4306**, 178–186 (2001)
5. J. Hynecek, Impactron – a new solid state image intensifier, IEEE Trans. Electron Devices **48**, 2238–2241 (2001)
6. D.J. Denvir, E. Conroy, Electron multiplying CCDs, Proc. SPIE **4877**, 55–68 (2003)
7. I. Kobayashi, H. Shibuya, T. Tachibana, T. Nishiwaki, S. Kashima, J. Hynecek, Design and performance of charge multiplying color FIT CCD image sensor, Proc. SPIE **5577**, 62761F-1–62761F-9 (2006)
8. G.B. Heim, J. Burkepile, W.W. Frame, Low-light-level EMCCD color camera, Proc. SPIE **6209**, 62090F-1–62090F-11 (2006)
9. C.G. Coates, D.J. Denvir, E. Conroy, N.G. McHale, K. Thornbury, M. Hollywood, Back-illuminated electron multiplying technology: the world's most sensitive CCD for ultralow-light microscopy, Proc. SPIE **4962**, 319–328 (2003)

10. P. Feautrier et al., The L3Vision CCD220 with its OCam test camera for AO applications in Europe, Proc. SPIE **7021**, 70210C-1–70210C-12 (2008)
11. N.M. Law, C.D. Mackay, J.E. Baldwin, Lucky imaging: high angular resolution imaging in the visible from the ground, Astron. Astrophys. **446**(2), 739–745 (2006)
12. T.C. Soesbe, M.A. Lewis, N.V. Slavine, P.P. Antich, Development of an EMCCD based gamma camera for preclinical SPECT imaging, IEEE Trans. Nucl. Sci. **54**(5), 1516–1524 (2007)
13. J.W.T. Heemskerk, A.H. Westra, P.M. Linotte, K.M. Ligtvoet, W. Zbijewski, F.J. Beekman, Front-illuminated versus back-illuminated photon-counting CCD-based gamma camera: important consequences for spatial resolution and energy resolution, Phys. Med. Biol. **52**, N149–N162 (2007)
14. B.W. Miller, H. Bradford Barber, H.H. Barrett, I. Shestakova, B. Singh, V.V. Nagarkar, Single-photon spatial and energy resolution enhancement of a columnar CsI(Tl)/EMCCD gamma-camera using maximum-likelihood estimation, Proc. SPIE **6142**, 6142T1–6142T9 (2006)
15. D. Hall, A. Holland, High-resolution x-ray and γ-ray imaging using a scintillator-coupled electron-multiplying CCD, Proc. SPIE **7449** (2009)
16. G.R. Hopkinson, D.H. Lumb, Noise reduction techniques for CCD image sensors, J. Phys. E Sci. Instrum. **15**, 1214–1222 (1982)
17. D.J. Burt, Extending the performance limits of CCD image sensors, GEC J. Res. **12**(3), 130–140 (1995)
18. S. Ohta, H. Shibuya, I. Kobayashi, T. Tachibana, T. Nishiwaki, J. Hynecek, Characterization results of 1 k × 1 k charge-multiplying CCD image sensor, Proc. SPIE **5301**, 99–108 (2004)
19. J. Hynecek, T. Nishiwaki, Excess noise and other important characteristics of low light level imaging using charge multiplying CCDs, IEEE Trans. Electron Devices **50**, 239–245 (2003)
20. M.S. Robbins, B.J. Hadwen, The noise performance of electron multiplying charge-coupled devices, IEEE Trans. Electron Devices **50**, 1227–1232 (2003)
21. M.J. DeWeert, J.B. Cole, A.W. Sparks, A. Acker, Photon transfer methods and results for electron multiplication CCDs, Proc. SPIE **5558**, 248–259 (2004)
22. A. O'Grady, A comparison of EMCCD, CCD and emerging technologies optimized for low light spectroscopy applications, Proc. SPIE **6093** (2006)
23. S. Tulloch, Modelling the suitability of EMCCDs for spectroscopic applications, Proc. SPIE **7021**, 70212C-1–70212C-10 (2008)
24. A.G. Basden, C.A. Haniff, C.D. Mackay, Photon counting strategies with low-light-level CCDs. Mon. Notices R. Astron. Soc. **345**, 985–991 (2003)
25. Y. Wen, B.J. Rauscher, R.G. Baker, M.C. Clampin, P. Fochie, S.R. Heap, G. Hilton, P. Jorden, D. Linder, B. Mott, P. Pool, A. Waczynski, B. Woodgate, Individual photon counting using e2v L3 CCDs for low background astronomical spectroscopy, Proc. SPIE **6276**, 62761H-1–62761H-8 (2006)
26. E. Lantz, J.-L. Blanchet, L. Furfaro, F. Devaux, Multi-imaging and Bayesian estimation for photon counting with EMCCDs, Mon. Notices R. Astron. Soc. **386**, 2262–2270 (2008)
27. J.R. Janesick: *Scientific Charge Coupled Devices*, (SPIE Press monograph, 2001, ISBN 0819436984)
28. N.S. Saks, A technique for suppressing dark current generated by interface states in buried channel CCD imagers, Electron Device Lett. IEEE **1**(7), 131–133 (1980)
29. T. Plakhotnik, A. Chennu, A.V. Zvyagin, Statistics of single-electron signals in electron-multiplying charge-coupled devices, IEEE Trans. Electron Devices **53**, 618–62 (2006)
30. L. Zhang, L. Neves, J.S. Lundeen, I.A. Walmsley, A characterization of the single-photon sensitivity of an electron multiplying charge-coupled device J. Phys. B At. Mol. Opt. Phys. **42**(11) 114011 (10 pp) (2009)
31. e2v Technologies Low light technical note 4, dark signal and clock-induced charge in L3Vision CCD sensors
32. O. Daigle, S. Blais-Ouellette, Photon counting with an EMCCD, Proc. SPIE **7536**, 753606-1–753606-10 (2010)
33. W. Zhang, Q. Chen, Signal-to-noise ratio performance comparison of electron multiplying CCD and intensified CCD detectors, IEEE Int. Conf. Image Anal. Signal Process. **2009**, 337–341 (2009)

Chapter 7
Monolithic Single-Photon Avalanche Diodes: SPADs

Edoardo Charbon and Matthew W. Fishburn

Abstract The art of creating monolithic single-photon photodetectors is a mix of design skills and device physics knowledge, and it requires an understanding of the mechanisms underlying single-photon detection in highly complex integrated systems. This chapter begins with the fundamentals of avalanching, the basics for integration of avalanche photodiodes operating in Geiger-mode, and the issues associated with large arrays. We outline the techniques that made it possible to integrate single-photon detectors in standard CMOS processes, and to design compact ancillary electronics for operating and reading pixels based on these devices. Finally, we present a description of several readout architectures designed for massive arrays of single-photon detectors. A discussion of future trends in the context of the most advanced applications in various fields of research concludes this chapter.

7.1 A Brief Historical Perspective

Solid-state single-photon detectors have existed for decades and, while several flavors of solid-state detectors exist in various technologies and ranges of operation, from cryogenic to room temperature detectors, silicon avalanche photodiodes (APDs) have emerged as the most versatile and easy to use among them [1]. A class of APDs operating above breakdown, in so-called Geiger-mode and known as single-photon avalanche diodes (SPADs) or Geiger-mode APDs (GAPDs), is of particular interest due to their amenability to integration in planar silicon processes in combination with conventional digital and analog circuitries. The first SPADs implemented in a planar technology have emerged relatively recently [2, 3]. But, while the physics of solid-state SPADs is well understood [4], it is only with the advent of devices integrated in conventional complementary metal–oxide–semiconductor (CMOS) processes [5] that the evolution onto smaller and smaller feature sizes has rapidly advanced to the point that it has now become possible to

P. Seitz and A.J.P. Theuwissen (eds.), *Single-Photon Imaging*,
Springer Series in Optical Sciences 160, DOI 10.1007/978-3-642-18443-7_7,

envision large imaging systems based on SPADs. It is on these systems that this chapter will focus.

7.2 Fundamental Mechanisms

The models covered in the following sections are based on [1, 4, 6].

7.2.1 SPAD Structure and Operation

A SPAD is essentially a p–n junction biased above breakdown and equipped with avalanche quenching and recharge mechanisms. When a photon is incident on the diode, an avalanche may be triggered. The quenching circuitries stop the avalanche thus preventing the destruction of the device, while the recharge circuitries prepare the SPAD for the next detection. When many engineers think of a p–n junction's I–V curve, they think of the steady-state curve shown in Fig. 7.1. In this curve, there are three main regions: forward, reverse, and breakdown. Breakdown occurs when the electric field is strong enough that accelerated carriers cause ionization, that is the creation of mobile electron–hole pairs. The breakdown region can be further subdivided into the linear and Geiger-mode regions. In the linear region, the electrical field is strong enough to cause significant ionization through free

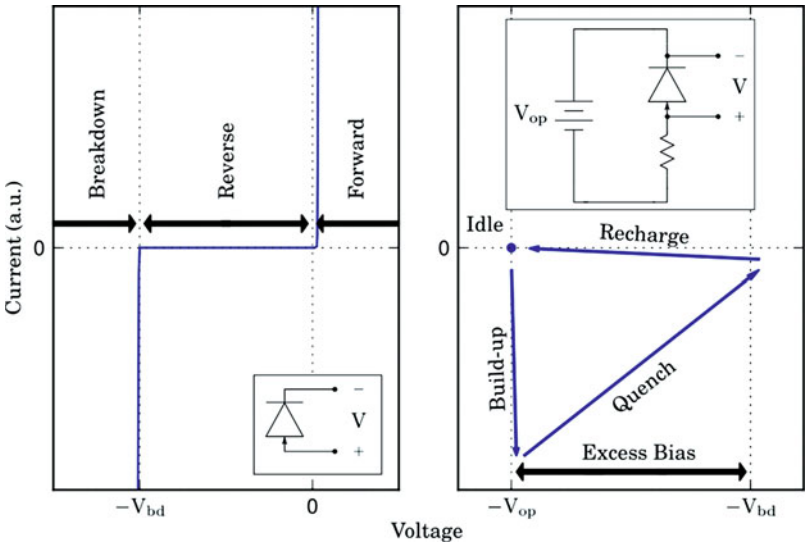

Fig. 7.1 I–V characteristics shown for a p–n junction (*left*) and a SPAD in a standard passive quenching scheme (*right*)

electrons, but not through holes of the same energy. The region is named for the expected linear gain per injected electron. In Geiger-mode, named after its similarity to the operation of a Geiger counter, the electric field is so strong that both electrons and holes cause significant ionization. In Geiger-mode, the current is limited only by the space-charge phenomenon, which is the electric field drop from the charge of carriers currently moving across the depletion region of the diode.

Quantitatively, to be in Geiger-mode, the expected number of created carriers from ionization must exceed one. For a one-dimensional junction ranging from z_0 to z_1, and with $\bar{\alpha}$ (units cm^{-1}) representing the mean ionization rate in a semiconductor with equal electron and hole ionization coefficients, the condition to be in Geiger-mode is:

$$\int_{z_0}^{z_1} \bar{\alpha} \, \mathrm{d}z > 1. \tag{7.1}$$

The voltage sufficient to meet this condition for a particular p–n junction is called the breakdown voltage, and the excess bias is the voltage applied above this voltage. In this chapter, the breakdown voltage will be considered positive, not negative. Readers unfamiliar with this terminology should be wary of the signs of all voltages. In the Geiger-mode regime, a single carrier in the depletion region can create an enormous current. The location of the carrier in the depletion region can change the chance that it triggers as avalanche, as many diodes have a nonuniform electric field. In particular, one-sided p–n junctions will have a linear electric field. In diodes with a nonuniform electric field, the portion of the depletion region where at least 95% of the ionization takes place is termed the multiplication region, also called the avalanche region in some texts [6]. This spatial region will not be called the avalanche region, as this term is used in this text for regions of the diode under avalanche. Carriers generated in the depletion region but not the multiplication region will drift either toward or away from the multiplication region depending on their charge and the p–n junction's doping levels. Due to the larger ionization coefficients of electrons compared to holes in many semiconductors, it is almost always preferable to have more electrons than holes drift toward the multiplication region. The avalanche's originating carrier can be the result of an electron–hole pair caused by a single photon, hence the term SPAD [6].

If a SPAD is operating at the steady-state value on the p–n junction's I–V curve, in practice it will heat up and be destroyed. However, the I–V curve of a SPAD coupled to quenching circuitry, also called ballast circuitry, has very different characteristics from a p–n junction's I–V curve. This section will cover the dynamics of an avalanche assuming a large resistance is placed in series with the diode. This is known as passive quenching, and produces the I–V curve shown in Fig. 7.1 when the diode is under low light. This I–V curve is not a true steady-state curve as the location on the I–V curve depends on whether an avalanche is occurring.

The I–V curve of the SPAD and ballast circuitry is dependent on the amount of light impinging on the SPAD. Under moderate to no light, the curve has four phases, each with distinct dominating factors: idle, buildup, quenching and spread, and recharge. The first phase, an unstable, idle state, is a quasi-steady-state condition

that occurs when no free carriers are near the p–n junction itself. The second phase, avalanche buildup, is the result of positive feedback from ionization until negative feedback from the space-charge phenomena balances carrier creation from ionization. The third phase, spread and quenching, is the result of the multiplication- or photon-assisted spread of the avalanche region and global, negative feedback from the quenching circuitry. In the final phase, recharge, the quenching circuitry restores the diode to the conditions in the first phase. A basis for the triggering sources, noise and light, is also examined in detail.

When a SPAD is placed under very intense light, the negative feedback in the recharge phase does not have sufficient time to restore the diode to its idle state. This creates a steady-state value close to the value on a p–n junction's I–V curve. In SPAD-based systems, especially communication systems, it may be necessary to ensure that the SPADs are not in this state.

7.2.2 Idle State and Avalanche Buildup

A SPAD is idle when there are no free carriers in the depletion region and the excess bias is as desired. The large charge separation has created a very high electric field in the middle portion of the SPAD, but without any free carriers ionization cannot occur. The charge separation also contributes a parasitic capacitance effect, though other capacitances from the guard ring and transistor gates may have a larger effect on the parasitic capacitance than the depletion region's charge separation. The SPAD will remain in this state until an event, such as the interaction with an incident photon, injects an electron–hole pair into or near the depletion region.

When carriers are injected into the depletion region, there are several possibilities. The first possibility is that all carriers drift out of the depletion region without causing ionization. The carriers might even drift through portions of the multiplication region without ionizing. The second possibility is that ionizations of the initial or ionization-generated carriers do occur, but all carriers are swept out of the depletion region without reaching sufficient number to trigger an avalanche. If a photon caused the initial pair's creation, then this photon will not be detected. To give a nuanced view, there is always the possibility that any number of ionization-generated carriers could fail to trigger an avalanche; however once the number of carriers reaches a certain threshold, this border case's probability is so small that it can be ignored. Finally, enough ionization can occur that other processes will begin to govern the avalanche dynamics. The buildup phase of an avalanche occurs from when the initial electron–hole pair is injected to when the space-charge phenomena causes the excess bias across a particular area of the diode to drop to nearly zero. The drop in excess bias will create a balance between carrier generation from ionization and carrier loss from drift. Due to parasitic capacitance effects, in general the applied voltage across the diode does not change greatly during the buildup phase.

The initial ionization occurs in a relatively small area. There is always the possibility that diffusion will cause carriers to move substantial distances; however,

the probability of this is so low that it can be ignored in practice. Based on the local conditions of positive feedback from ionization and negative feedback from carrier drift, the number of free carriers c in a small portion of the diode is approximated analytically by equation:

$$\frac{\mathrm{d}c}{\mathrm{d}t} = c\left(\frac{1}{\tau_\mathrm{p}} - \frac{1}{\tau_\mathrm{n}}\right), \tag{7.2}$$

where τ_p relates to the positive feedback from ionization, while τ_n relates to the negative feedback from carrier drift. Both terms will also change as the excess bias varies due to the space-charge phenomena. Multiple ways of deriving τ_p and τ_n exist: the analysis here is based after [1], as the use of the same ionization coefficients for electrons and holes simplifies the derivation without essentially changing the results. For an analysis with different ionization rates for electrons and holes, we refer to [4].

For a small portion of the diode, assuming that electron and hole concentrations can be written as a function of depth, $n(z,t)$ and $p(z,t)$, respectively, with carriers traveling at the saturation velocity v_s, the current density $j_\mathrm{e}(z,t)$ from electrons is given by:

$$j_\mathrm{e}(z,t) = -(q \cdot v_\mathrm{s} \cdot n(z,t)), \tag{7.3}$$

where q is the electron charge. In a small time Δt, the increase in electron concentration due to ionization will be the distance each carrier travels (units cm) multiplied by the ionization rate (units cm^{-1}) and the carrier concentration (units cm^{-3}), which is $(v_\mathrm{s}\Delta t) \cdot (\bar{\alpha}) \cdot (n(z,t) + p(z,t))$. This expression is the basis for the carrier generation term in the continuity equation. The continuity equation for electrons, which governs the change in electron concentration, will be

$$\frac{\mathrm{d}n(z,t)}{\mathrm{d}t} = \frac{1}{q}\frac{\mathrm{d}j_\mathrm{e}(z,t)}{\mathrm{d}z} + v_\mathrm{s}\bar{\alpha}(n(z,t)+p(z,t)) = -v_\mathrm{s}\frac{\mathrm{d}n(z,t)}{\mathrm{d}z} + v_\mathrm{s}\bar{\alpha}(n(z,t)+p(z,t)), \tag{7.4}$$

with a use of (7.3) in simplifying the expression. Similarly, for holes

$$\frac{\mathrm{d}p(z,t)}{\mathrm{d}t} = v_\mathrm{s}\frac{\mathrm{d}p(z,t)}{\mathrm{d}z} + v_\mathrm{s}\bar{\alpha}(n(z,t) + p(z,t)). \tag{7.5}$$

Since the total carrier concentration is a sum of the electron and hole concentrations, $c(z,t) = n(z,t) + p(z,t)$, (7.4) and (7.5) can be combined to yield:

$$\frac{\mathrm{d}c(z,t)}{\mathrm{d}t} = v_\mathrm{s}\left(\frac{\mathrm{d}p(z,t)}{\mathrm{d}z} - \frac{\mathrm{d}n(z,t)}{\mathrm{d}z}\right) + 2\bar{\alpha}v_\mathrm{s}c(z,t). \tag{7.6}$$

For a p–n junction with its depletion region's p-doped edge at z_0 and the depletion region's n-doped edge at z_1 with total depletion region width $z_\mathrm{d} = z_1 - z_0$, the expression for the carrier concentration can be integrated in the spatial dimension z with the following boundary conditions: there is a negligible electron concentration at the edge of the depletion region on the p-doped side of the junction, $n(z_0,t) = 0$; there is a negligible hole concentration at the edge of the depletion region on

the n-doped side of the junction, $p(z_1, t) = 0$; and there are identical hole and electron concentrations at opposite sides of the diode, which are equal to the carrier concentration, $n(z_1, t) = p(z_0, t) = c(t) = 0$. Thus, the carrier concentration will change over time as:

$$\begin{aligned}(z_1 - z_0)\frac{\mathrm{d}c(t)}{\mathrm{d}t} &= -2c(t)v_\mathrm{s} + 2\bar{\alpha}v_\mathrm{s}c(t)(z_1 - z_0) \\ \frac{\mathrm{d}c(t)}{\mathrm{d}t} &= 2c(t)\left(v_\mathrm{s}\bar{\alpha} - \frac{v_\mathrm{s}}{z_\mathrm{d}}\right)\end{aligned} \quad . \tag{7.7}$$

The form of (7.2) is identical to (7.7). If the variables α and z_d were constant, then (7.7) would absurdly imply that the carrier concentration would build forever. These variables are actually functions of the excess bias, which is related to the carrier concentration by the space-charge phenomena. The depletion region width z_d will change based on the applied bias. We will assume that the excess bias is small enough that a voltage drop from V_op to V_bd does not change this distance, which is a valid assumption for the standard operating conditions of most planar silicon diodes. Because the space-charge resistance times unit area AR_sc is a function of depletion region width, this expression may be taken as constant, too.

While the depletion region width does not change appreciably, the same cannot be said of the ionization rate. The mean ionization rate as a function of the electric field is:

$$\bar{\alpha}(\xi) = A_\alpha \exp\left(-\left(\alpha/|\xi|\right)^m\right), \tag{7.8}$$

where the variables m and A_α (not to be confused with the area, A) are constants that depend on the material, and in silicon they have been measured as 1 and $10^7\,\mathrm{cm}^{-1}$ [7]. The variable α can be found by solving for equality in (7.1), the breakdown condition. The electric field in a p–n junction will be a function of the applied voltage V, turning ξ into $\xi(V)$. For current $I(t) = j(t)A = c(t)v_\mathrm{s}qA$, the applied voltage will vary as $V = V_\mathrm{op} - I \quad R_\mathrm{SC} = V_\mathrm{op} - c(t)qv_\mathrm{s}(AR_\mathrm{sc})$. Thus, the ionization rate will be:

$$\bar{\alpha}(\xi(V)) = A_\alpha \cdot \exp\left(-\left(\alpha/|\xi(V)|\right)^m\right) = A_\alpha \cdot \exp\left(-\left(\alpha/\left|\xi(V_\mathrm{op} - c(t)qv_\mathrm{s}(AR_\mathrm{sc}))\right|\right)^m\right). \tag{7.9}$$

In a one-sided junction, the electric field will vary as the square root of the applied voltage, and this expression can be further reduced to:

$$\bar{\alpha}(c(t)) = z_\mathrm{d}A_\alpha^2 \cdot \exp\left(-\left(\frac{V_\mathrm{bd}}{V_\mathrm{op} - c(t)qv_\mathrm{s}(AR_\mathrm{sc})}\right)^{1/2}\right). \tag{7.10}$$

Equations (7.7) and (7.10) give sufficient information to generate a differential equation governing the behavior of $c(t)$ for the initial buildup stage in a small portion of the diode, with boundary conditions given by the time of the electron–hole pair injection. For a single-photon generating an electron–hole pair at time 0, the boundary conditions are $c(t) = 0\,\mathrm{cm}^{-3}$ for $t < 0$ and $c(0) = 2\,\mathrm{cm}^{-3}$.

The given analysis assumes deterministic generation rates. In reality ionization is a statistical process, with uncertainty around the deterministic "average" solution. The ionization statistics during the buildup phase are known to contribute significant jitter to the timing uncertainty of the photon's arrival. Increasing the voltage threshold for determining the avalanche onset will increase the timing uncertainty, with the best jitter achieved with either a low voltage threshold or current-based approach [2, 8].

Unless the avalanche onset can be determined with some other technique, the statistics of ionization will set a limit on the timing uncertainty achievable with SPADs. Higher breakdown voltages, lower doping and smaller parasitic capacitances will lower this jitter.

7.2.3 Quench, Spread, and Recharge

Two modes for quenching and recharge exist: active and passive. In active mode, active circuitries are used to control the process. In passive mode, the avalanche current is passively controlling the process by way of a ballast resistive device. The literature on the subject is extensive, and it is beyond the scope of this chapter to describe it further [9]. We describe hereafter only the mechanisms at device level underlying quenching.

During the initial phase of the avalanche, positive feedback in a local portion of the diode rapidly increases the carrier concentration until the space-charge phenomena indirectly stops the process. After this initial phase, the excess bias across the diode begins to decrease as the parasitic capacitance and quenching circuitry increase in applied voltage. Unlike in the buildup phase, when the applied voltage across the SPAD did not change greatly, in this phase the applied voltage will drop below the breakdown voltage, quenching the avalanche. Depending upon the SPAD's characteristics, the spread of the avalanche in the planar directions may or may not play an important role in quenching. The avalanche spreads via both multiplication-assisted and photon-assisted processes. Photon-assisted spreading occurs from photons generated by recombination [9]. Multiplication-assisted spreading occurs from carrier diffusion in the planar directions [10]. Photon-assisted spreading is known to be dominant in large diodes, with multiplication-assisted processes dominating in smaller diodes.

Multiplication-assisted spreading is a process that combines positive feedback with diffusion, being governed by the equation:

$$\frac{\partial c(x,y)}{\partial t} = D\nabla^2 c(x,y) + \frac{c(x,y)}{\tau(x,y)}. \tag{7.11}$$

In (7.11), $c(x,y)$ is the free carrier concentration at a particular position, D is the diffusion coefficient of carriers in the planar direction, and τ is the same time constant presented for the buildup phase in (7.7) earlier. This differential equation can be solved numerically with an FEM model of a SPAD, such as the one shown

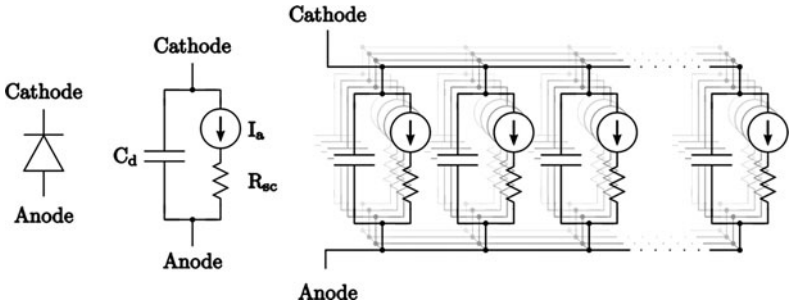

Fig. 7.2 Diode models of a single diode (*center*) or a single diode split into multiple diodes (*right*) are shown for a p–n+ junction. The current source and resistance should be exchanged for a p+–n diode, with the direction of the current source remaining the same

in Fig. 7.2. The spreading speed of the avalanche is experimentally known to be $v_s = 2\sqrt{D/\tau}$. In a planar silicon diode with D estimated at 15 cm^2/s and $\tau = 10$ ps, the spreading velocity is 2.5×10^6 cm/s. This is slightly below 25% of the saturation velocity of carriers in silicon at a similar temperature, though the τ value is a bit small compared to the expected value in many diodes. As the applied voltage decreases due to interactions with the quenching circuitry, the τ value will also increase, causing the spreading speed to be reduced.

In a small planar diode, the avalanche will spread across the entire diode before quenching begins to occur. Whether or not a diode falls into this category will also depend on the parasitic capacitance, as large capacitances will allow more spreading before the avalanche can be sensed. In diodes that have the avalanche spread across the entire diode before sizeable quenching occurs, (7.7) and (7.10) with V as a function of time can also be used to describe the quenching of the avalanche. In (7.10), the variable V must be modified to account for changes in the applied voltage across the diode, including the effects of the parasitic capacitance and quenching circuitry.

For larger diodes, the quenching and spread cannot be decoupled. The spread and quenching will occur at the same time, and the observed voltage waveform may even vary as a function of initiation position or number of simultaneous initiations, giving rise to position-sensitive and photon-number-resolving SPADs. These diodes can be modeled as a finite number of smaller diodes, with the carrier concentration within a particular diode governed by (7.7), (7.10), and (7.11). Eventually, the avalanche will need to be completely quenched. The avalanche can be considered to have quenched when the number of free carriers in the multiplication region reaches some small number. An inductance-like component, which is caused by the inertia given to charges in the depletion region, aids in ensuring that the applied voltage eventually falls below the breakdown voltage. At this point in time, the excess bias across the SPAD is nearly zero or negative, and ideally all free carriers have been swept out of the depletion region.

The SPAD's state needs to be restored to the idle state, so the SPAD is sensitive to single charge carriers again. The restoration occurs during the recharge phase. Recharge is achieved by increasing the excess bias across the SPAD back to the

desired value. In a passively quenched SPAD, the recharge phase is a simple RC circuit, with the quenching resistance in series to the parasitic capacitance. In practice, however, many passively quenched circuits use a transistor operating as a current source, meaning that the excess bias voltage is better approximated as a linear function during this time. In actively quenched circuits, the recharge behavior will depend on the quenching circuit. In addition, there will be a variation in the parasitic capacitance because the width of charge separation in the p–n junction changes; however, this variation is usually small enough to be neglected.

It might be confusing that the current into the diode is the same sign during an avalanche and recharge phase, since the voltage across the SPAD increases and decreases with the same direction of current. However, different effects are dominating during different phases, so the voltage change cannot be associated with current flowing in a particular direction. During recharge, the current causes a buildup of charge separation to create an unstable system, whereas during an avalanche the current is that unstable system's release of energy.

During recharge, the SPAD will be initially insensitive to injected carriers, but it will become increasingly sensitive as the excess bias changes. Carriers injected during recharge can be problematic for silicon SPADs. If no carriers cause an avalanche, the recharge stage should be complete and the SPAD will remain idle until another carrier is injected, when all the phases will repeat.

7.2.4 Example Waveforms

This section presents example waveforms from an FEM model of a planar silicon SPAD in a standard quenching circuit, as shown in Figs. 7.2 and 7.3. A $2 \times 40\,\mu m^2$ diode with a breakdown voltage of approximately 19 V was simulated at an excess bias of 3 V. The SPAD was split into $0.1 \times 0.1\,\mu m^2$ diodes, making a total of 20 by 400 simulated diodes. The concentration of carriers within each diode was governed by (7.10), (7.7), and (7.11). The constants used were a diffusion coefficient of $17 \times 10^8\,\mu m^2/s$, a space-charge resistance-unit area of $12\,k\Omega\,\mu m^2$, a quenching resistance of $300\,k\Omega$, and a total parasitic capacitance of 300 fF. The parasitic capacitance was assumed to be the sum of 195 fF from the guard ring, 5 fF from

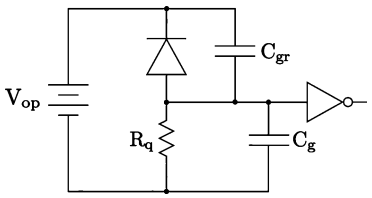

Fig. 7.3 Standard SPAD circuit with a voltage source, SPAD, quenching resistance, capacitors, and inverter is shown. The capacitors represent the parasitic capacitances from the guard ring and transistor gate, and are approximately 10–100 s of fF. The resistor varies from approximately 100 kΩ to 1 MΩ. During simulation, the SPAD is normally replaced by a model shown in Fig. 7.2

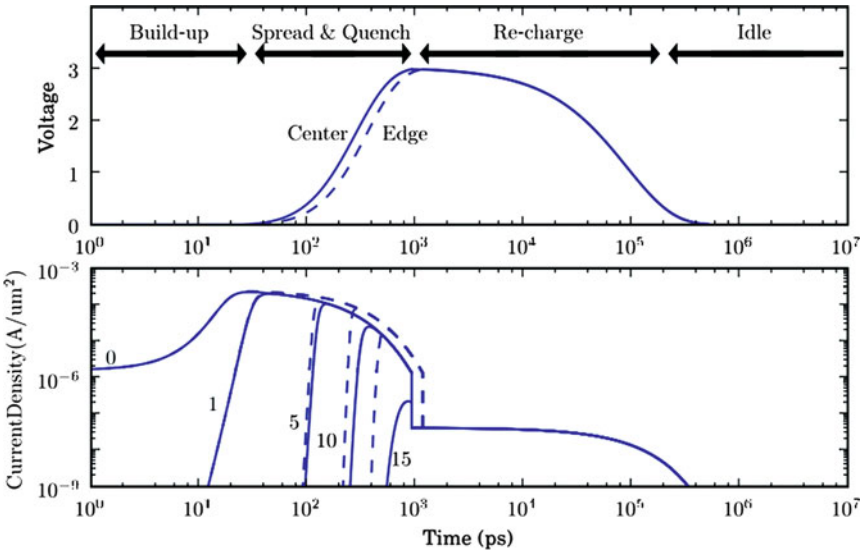

Fig. 7.4 Voltage drop (*top*) and current density (*bottom*) vs. time in the various phases of an avalanche in log scale. Avalanches triggered in the center (*solid line*) and in the edge (*dashed line*) of the diode are shown. The superimposed lines represent current densities 0, 1, 5, 10, and 15 μm away from initiation, respectively

the thresholder's transistor gate, and 100 fF from the diode capacitance. The initial current densities were assumed to be 0 $A/\mu m^2$ for all element diodes, except for one diode with a current density of 1.6×10^{-6} $A/\mu m^2/s$, corresponding to the injection of a single charge carrier. Photon-assisted spreading was ignored, as this diode is a relatively small planar diode that is unlikely to have photon-assisted spreading.

The voltage drop across the SPAD is shown in Fig. 7.4 as a function of time. The voltage drop remains relatively small during the buildup phase, with a sizeable increase only during the quench phase of the avalanche. Once the voltage drop is the same value as the excess bias, ionization is assumed to have stopped. This occurs after approximately 1 ns in this diode, while the voltage returns to zero during the recharge phase.

The current densities for the element diodes, also shown in Fig. 7.4 as a function of time, increase rapidly when the avalanche spreads to that particular region of the diode. The saturated value of current density is limited by the space-charge resistance; once a particular portion of the diode is space-charge limited, the voltage drop across the current source should be the breakdown voltage. Because the element diodes' capacitances are small compared to the total capacitance, nearly all of the ionization current exits the element diodes. After ionization ceases, a small current will continue to exit the diodes, as the voltage across the diode capacitance must also be recharged. The direction of the current is the same in both cases.

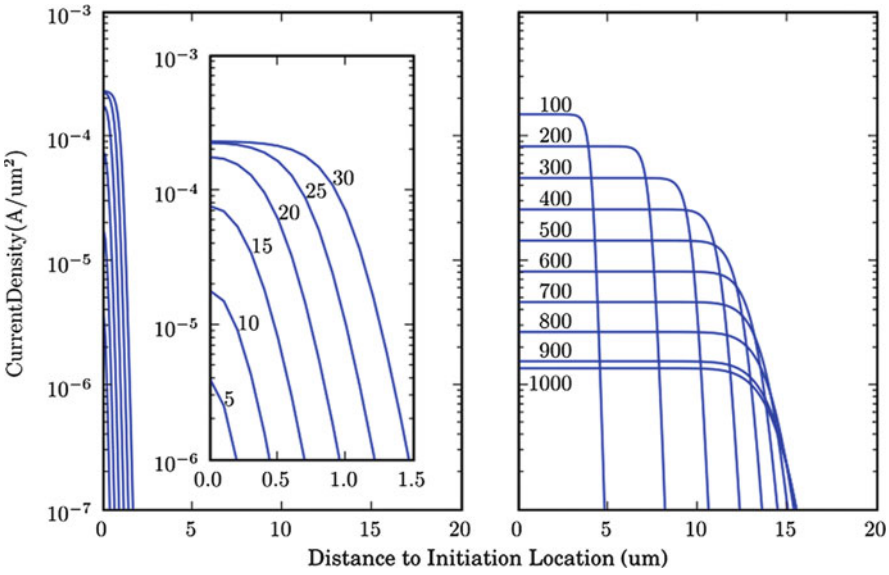

Fig. 7.5 Current density after triggering: at 5 ps intervals during buildup (*left*); at 100 ps intervals during spread and quench phases (*right*). The avalanche is triggered in the center. Due to symmetry, the current density in the junction varies as a function of position to the avalanche initiation triggering position

Figure 7.4 presents both an avalanche triggered in the center of the diode and on the edge of the diode. As the diode initially sources less current when it is triggered on the edge, the voltage drop is smaller and the avalanche spreads more quickly. The difference in quenching time in this diode is approximately 200 ps; connecting multiple thresholders and a picosecond-resolution time-to-digital converter (TDC[1]) to this diode would yield position sensitivity [14]. The current densities as a function of distance are shown during the buildup and quenching/spreading phases in Fig. 7.5. In this particular diode, the buildup phase will last approximately 35 ps, and the avalanche will have spread almost 1.5 μm in all directions, with space-charge limited current density in a circular area with a radius of approximately 1 μm. The spreading and quenching phase will last around 1 ns.

The spreading speed of the avalanche will slow as the voltage across the quenching resistance increases, as the later current density curves show. After ionization ceases and recharge begins, the current density across the diode will be uniform due to the recharge of the diode capacitance. In reality all of the avalanche phases are based on statistical processes, and the presented, deterministic curves only capture the essence of the expected case. Many factors, including ionization noise, photon-assisted spreading, and environment variations, will change

[1] A TDC is a sort of chronometer capable of discerning pulse position with high precision.

variable governing the ionization and spread. However, this example is useful for the standard operating conditions of many SPADs and establishes a baseline for functionality.

7.2.5 Pulse-Shaping

While many circuit architectures are possible for pulse-shaping, the majority of published work contains a single comparator connected to the SPAD. In CMOS-based circuits, this comparator is often a minimum-sized inverter or a thresholder. The comparator will output a digital-level signal about the state of the SPAD. For a SPAD with a simple resistance for a quenching circuit, the falling edge of the output inverter's waveform will signal an avalanche's onset.

While a single sample of the waveform is sufficient for counting single photons and obtaining times-of-arrival, there is additional information in the output waveform that can be obtained. As described earlier, the spread of the avalanche can change the output waveform. For example, if a large-area SPAD receives an enormous number of photons simultaneously, the voltage across the quenching resistance will rise much more quickly than if a single photon was received. In a similar fashion, if a SPAD has a rectangular active area with one side much longer than the other, initial carriers injected at the edge of the major axis can create a slower voltage rise than when injected in the middle. For specialized applications requiring these types of information, circuitry often samples the output waveform multiple times. This is most commonly done with a pair of thresholders with their outputs connected to some sort of TDC.

SPADs that rely on multiple samplings of the output waveform are generally named after their target application. Compensated SPADs are used to compensate for changes in the SPAD's output waveform for the number of impinging photons [11]. As previously described, when a large number of photons are incident on a large detector in a small time window, the current in the SPAD will increase much more quickly than otherwise. For jitter-sensitive applications, such as 3D imaging, this effect can create time-of-arrival (TOA) differences in the 100 s of picoseconds. Compensated SPADs are a subclass of photon-number-resolving SPADs, which sample the output waveform to gain information about the number of impinging photons [12].

Position-sensitive SPADs have been shown that use the rise time of the waveform to obtain information about the initiation location of the avalanche [13]. There are a number of effects that make these detectors less useful and more difficult to use than one might imagine. The avalanche spread slows as the avalanche quenches, creating different position sensitivity in the diode's middle than on the edges. In a symmetric diode, only the position from the center may be available. For example, in an elliptical diode with a minor axis that is short compared to the major axis, photons triggered at either edge of the major axis will produce the same output waveform.

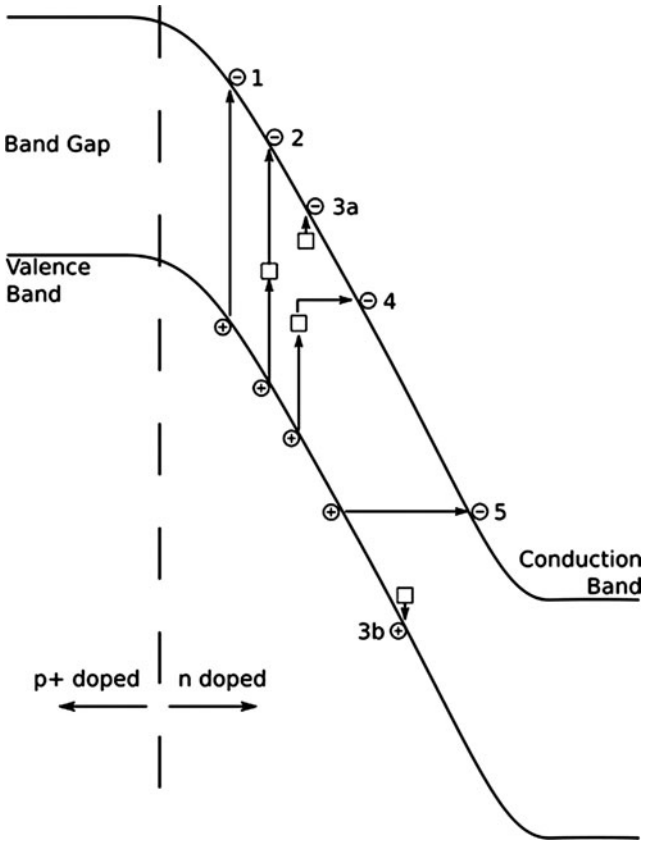

Fig. 7.6 Band diagram of a SPAD junction. *Squares* represent traps caused by lattice defects. The sources of dark counts are: thermal noise (1), trap-assisted noise (2), afterpulsing (3a and 3b), trap-assisted tunneling (4), tunneling-assisted (5) noise. After [1]

7.2.6 Uncorrelated Noise: Dark Counts

Dark counts are a source of intrinsic noise in SPADs, and are generally characterized in terms of the dark count rate (DCR), that is the number of avalanches occurring in a SPAD when there is no light impinging on the active area. DCR sources can be roughly classified into thermal noise, trap-assisted sources, and tunneling-assisted sources, though sources such as trap-assisted tunneling may fall into both categories. A graphical representation of the noise's relation to the energy bands is shown in Fig. 7.6.

The amount of tunneling-assisted noise is strongly dependent on the doping of the junction and excess bias, and only weakly dependent on the temperature. This type of noise has been shown to be problematic when working with deep submicron CMOS processes, though its effects have been mitigated at the 130 nm node by using

existing or modifying doping profiles of implants as appropriate [14]. Whether or not the effects will be severe in SPADs in more advanced CMOS processes remains to be seen, though progress has been made in CMOS processes with a feature size of 90nm [15]. SPADs can be constructed with very little tunneling-assisted noise in older CMOS processes, especially in 0.35 μm and 0.8 μm processes [16, 17].

Trap-assisted noise depends strongly on both the CMOS process and the temperature. Traps are lattice defects in silicon that can hold and release carriers. Traps are most prevalent near the surface, and can also be problematic near isolation layers, such as shallow trench isolation (STI). STI-based guard rings can have prohibitive amounts of trap-assisted noise due to their introduction of a trap-laden surface next to the depletion region. For diodes without enormous amounts of tunneling-assisted noise, the trap-assisted noise will limit the noise level above a certain temperature, and below this temperature the tunneling-assisted noise will dominate.

The DCR in SPADs is known to have a long tail; that is, a few noisy pixels source most of the dark counts [2]. Care must be given when examining the median and mean DCR, as the median can hide the poor performance of these noisy pixels. Whether or not a SPAD is "too noisy" depends on the target application and the SPAD construction. Photon-starved applications require lower noise rates. It is nearly impossible to predict DCR in SPADs without prior fabrication. The concentration of traps, for example, can vary significantly from process to process, as it depends on the doping levels and annealing steps in the CMOS process. Some environments, such as radioactive ones, will cause the number of traps to increase over time [18]. However, traps can also be used in the characterization and stabilization of DCR. Overall, noise rates in the single-digit MHz per square millimeter range are quite common for CMOS-based SPADs at room temperature [19]. At a noise rate of 5 MHz per square millimeter, a 10 μm diameter circular SPAD would have an expected DCR that is just under 400 Hz.

7.2.7 Correlated Noise: Afterpulsing and Other Time Uncertainties

There are three main sources of correlated noise: afterpulsing, electronic crosstalk, and optical crosstalk. Additionally, variations in dead time under high photon fluxes is covered in this section, as it causes signal distortion that is correlated with the count rate.

7.2.7.1 Afterpulsing

During the later portion of the recharge phase, sufficient excess bias exists such that an injected carrier can cause another avalanche. This is problematic in silicon detectors, as lattice defects can hold carriers after bursts of current and release them on the order of nanoseconds [20]. This unwanted, correlated activity is termed afterpulsing, and can have a significant effect in the first 50 ns or so following an

avalanche [21]. The effect is highly temperature dependent, with longer dead times required for the same afterpulsing probability at very cold temperatures [22].

The recharge time thus becomes a trade-off between afterpulsing and detector dead time. If the dead time is set too low, then after-pulsing becomes problematic. If the dead time is set too high, then the detector spends too much time inactive from recharge. The ideal dead time will vary with application, detector size, and temperature. Using a quenching transistor instead of a quenching resistor allows the dead time to be changed through a bias voltage to the transistor's gate, which is a common technique in CMOS-based SPADs [18].

7.2.7.2 Optical and Electrical Crosstalk

A large number of free carriers are generated during an avalanche. Some of these carriers will recombine and create additional photons that can cause optical crosstalk. This effect is most significant in large diodes with small separation distances. Large diodes have larger parasitic capacitances, increasing the number of carriers involved in an avalanche that increases the chance of generating a photon by recombination. Smaller separation distances imply that any recombination-generated photon is less likely to be absorbed before impinging on the next SPAD. Optical crosstalk probabilities higher than 20% have been measured in tightly packed SPAD arrays [23]. Some silicon photomultipliers include optical isolation of SPADs to lower this probability [24]. At the other end of the range, a CMOS chip with pixels containing a well-isolated SPAD and logic will have negligible crosstalk, as has been observed [5].

Further reduction of the number of avalanching carriers may be best achieved by reducing the active area of a SPAD, and thus its capacitance at a cost of lower fill factor if the pixel pitch is kept constant. Figure 7.7 shows the impact of a pixel near saturation onto the neighboring pixels at their nominal DCR values, indicating immeasurable crosstalk in the array structure of the design in [16].

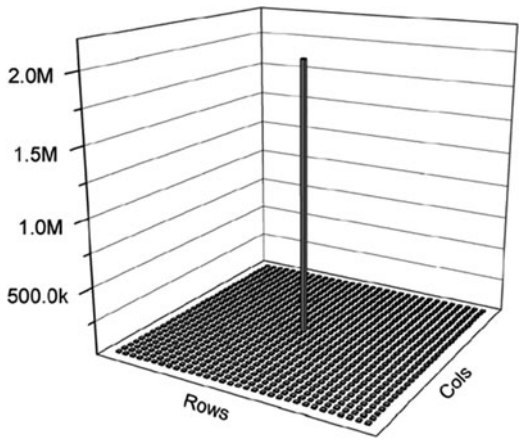

Fig. 7.7 Crosstalk measured in a 32 × 32 array of low-pitch passively recharged pixels, when the center pixel is selectively illuminated near saturation with an optical fiber [16]

When multiple p + –n junctions are located in the same n-well, or several p–n+ junctions sit on the substrate, the hot carriers exiting the depletion region can bounce around the substrate, eventually being injected into the depletion region of another diode. This effect is electrical crosstalk. Electrical crosstalk has been found to be negligible in CMOS-based arrays when each SPAD is isolated in its own well [5].

Crosstalk can be measured experimentally by observing the correlation between events in two diodes; it is quite difficult to separate observation of optical and electrical crosstalk. To estimate the optical crosstalk probabilities, three things are needed. First, the geometries of the two SPADs under question are required. Second, the photon emission rate per charge carrier must be known, along with the number of carriers involved in the avalanche. Finally, the spectrum of emitted photons must also be known.

7.2.7.3 Charge Pile-Up

As mentioned earlier, avalanches can occur during the recharge phase, causing distortions in the dead time and also count rate. This effect is commonly called charge pile-up. Pile-up causes a much faster saturation of counts than would be expected. With a dead time of 1 μs, a saturated count rate that is less than 10% of the maximum 1 MHz value has been observed, though the count rate can be linearized to closely match the expected value [14]. The effect will significantly distort the SNR and needs to be considered when the SPAD is operating with a count rate that is a sizeable fraction of the maximum count rate. The statistics of this effect have been studied, with possible compensation.

7.2.8 Sensitivity: Photon Detection Probability

If light creates the avalanche-initiating electron–hole pair, the initial carrier depth is a strong function of the photon's energy, and hence its wavelength. A photon's absorption depth in silicon is an exponential process when the photon energy is larger than the band-gap. The large difference in absorption depth in silicon, shown in Fig. 7.8, is caused by silicon's indirect band-gap. In indirect band-gap materials, phonons must also impart some energy for electron–hole pair generation. This is not the case for direct band-gap materials. The photon detection probability (PDP) analysis for silicon is easily extendable to other materials if the mean absorption depth in that material is known as a function of the incident photon energy.

Due to the fixed location of the depletion and multiplication regions in most silicon SPADs, there is an ideal mean absorption depth for photons. Light with a mean absorption depth that is too small, such as ultraviolet (UV) light, will create carriers too close to the surface. These carriers will probably recombine before reaching the multiplication region. Light with a large mean absorption depth, such as infrared light, either will create electron–hole pairs too deep in the silicon, or

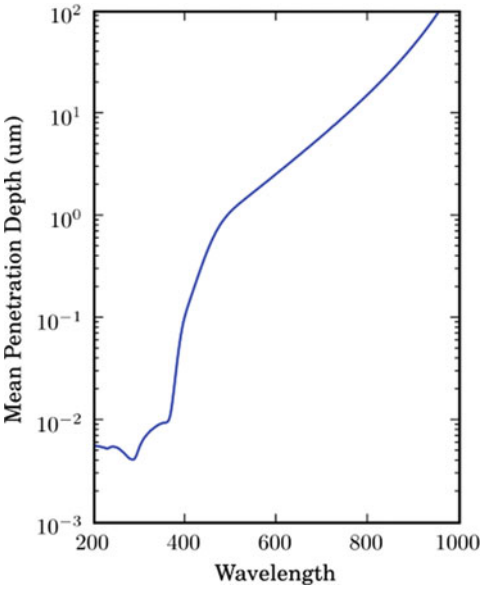

Fig. 7.8 Mean penetration depth vs. photon wavelength in silicon [53]

will not create the pairs at all. For planar silicon detectors, the ideal wavelength is usually blue light with a wavelength in the 400–500 nm range. Other materials, such as SiC or InGaAs/InP, shift the optimal wavelength into the UV or infrared.

The PDP can be calculated if the avalanche probability as a function of electron–hole pair creation is known. This section will present such a calculation based on a model derived from [25]. This function will be denoted as $p(z)$. Combining $p(z)$ with the probability of electron–hole pair generation at that depth by a specific wavelength of light will yield the PDP for that specific wavelength. Calculating $p(z)$ in the multiplication region is straightforward in a numerical manner when the ionization coefficients as a function of depth are known. The ionization coefficients of silicon have been experimentally measured as a function of the electric field. The electric field can be calculated if the doping levels are known, and for a well-based SPAD in silicon the doping levels can normally be calculated from the breakdown voltage. Thus, the breakdown voltage, in tandem with assumptions about the diode construction and measured results from the literature, gives sufficient information to calculate the PDP.

Once the ionization coefficients are known, $p(z)$ is calculated as follows for a p+–n SPAD with the same ionization coefficients of electrons and holes, though the extensions to an p–n+ SPAD or a SPAD with different electron and hole ionization coefficients is straightforward. Let $p_e(z)$ and $p_h(z)$ be the respective probabilities that an electron and hole at a depth z trigger an avalanche. For an electron–hole pair, $p(z)$ is the probability that the hole or the electron generate the avalanche, with $p(z) = p_e(z) + p_h(z) - p_e(z)p_h(z)$.

The value of $p(z)$ is related to the value of $p(z + dz)$ by the probability that the electron does or does not ionize and generate an electron–hole pair between

z and $z + \mathrm{d}z$. If "A" signifies a ionization-generated carrier causing an avalanche and "B" the initial carrier causing an avalanche, then P[avalanche] = P[A or B] = P[A] + P[$\bar{\text{A}}$] · P[B|$\bar{\text{A}}$]. The electron will ionize between z and $z + \mathrm{d}z$ with probability $\alpha(z) \cdot \mathrm{d}z$, so the probability of an avalanche caused by a generated carrier is P[A] = $\alpha(z) \cdot p(z) \cdot \mathrm{d}z$. Note that the generated carriers are assumed to be created at a depth of z, although if they were created instead at $z + \mathrm{d}z$ the difference would vanish. Statement B is independent of A, as ionization will not significantly alter the electric field, thus implying that P[$\bar{\text{A}}$] · P[B|$\bar{\text{A}}$] = $(1 - \alpha(z) \cdot p(z) \cdot \mathrm{d}z)\, p_{\mathrm{e}}(z + \mathrm{d}z)$. Thus, the equation relating $p_{\mathrm{e}}(z)$ to $p_{\mathrm{e}}(z + \mathrm{d}z)$ is:

$$p_{\mathrm{e}}(z) = \alpha(z) \cdot p(z) \cdot \mathrm{d}z + (1 - \alpha(z) \cdot p(z) \cdot \mathrm{d}z)\, p_{\mathrm{e}}(z + \mathrm{d}z). \tag{7.12}$$

A differential equation for $\mathrm{d}p_{\mathrm{e}}(z)/\mathrm{d}z$ can be found by rearranging the terms above and taking the limit as $\mathrm{d}z$ goes to zero, as

$$\frac{\mathrm{d}p_{\mathrm{e}}(z)}{\mathrm{d}z} = \alpha(z) \cdot p(z) \cdot (p_{\mathrm{e}}(z) - 1). \tag{7.13}$$

Similarly, for holes

$$\frac{\mathrm{d}p_{\mathrm{h}}(z)}{\mathrm{d}z} = -\alpha(z) \cdot p(z) \cdot (p_{\mathrm{h}}(z) - 1). \tag{7.14}$$

Assuming the depletion region starts at z_0 in the p+-doped side and ends at z_1 in the n-doped side, the boundary conditions to the solution of (7.13) and (7.14) are $p_{\mathrm{h}}(z_0) = 0$ and $p_{\mathrm{e}}(z_1) = 0$. To estimate the equations' solutions, $p_{\mathrm{e}}(z_1) = 0$ can be swept from 0 to 1, with numerical analysis yielding $p_{\mathrm{e}}(z)$, $p_{\mathrm{h}}(z)$, and $p(z)$ for z between z_0 and z_1. For the initial value of $p_{\mathrm{e}}(z)$, which gives $p_{\mathrm{e}}(z_1) = 0$, the boundary conditions will be met and values for all p functions will be known between z_0 and z_1.

If the electron–hole pair is generated outside of the depletion region, then the triggering probability will vary depending on the generation location. For example, carriers generated in the well will increase the PDP, but at the expense of distortions in the timing jitter [18]. In this instance, it will be assumed that $p(z)$ varies linearly between 0 at the surface and $p(z_0)$, and that $p(z)$ is 0 everywhere else. There is an additional effect that merits mention. There is material between the air and the surface of the SPAD, commonly called the optical stack. Materials such as silicon dioxide do not have uniform transmittance curves, and extremely accurate PDP calculations should include the effects of the optical stack.

Finally, since the triggering probability as a function of electron–hole pair generation depth is known, and the probability density function of pair generation at that depth is known for any particular wavelength, the PDP at a particular wavelength is the integral of the product of the triggering probability function with the PDF for that particular wavelength. Theoretical PDP curves are shown in Fig. 7.9 for a p+–n junction at a depth of 200 nm with a breakdown voltage of approximately 18.9 V. Ionization coefficients are taken from [26].

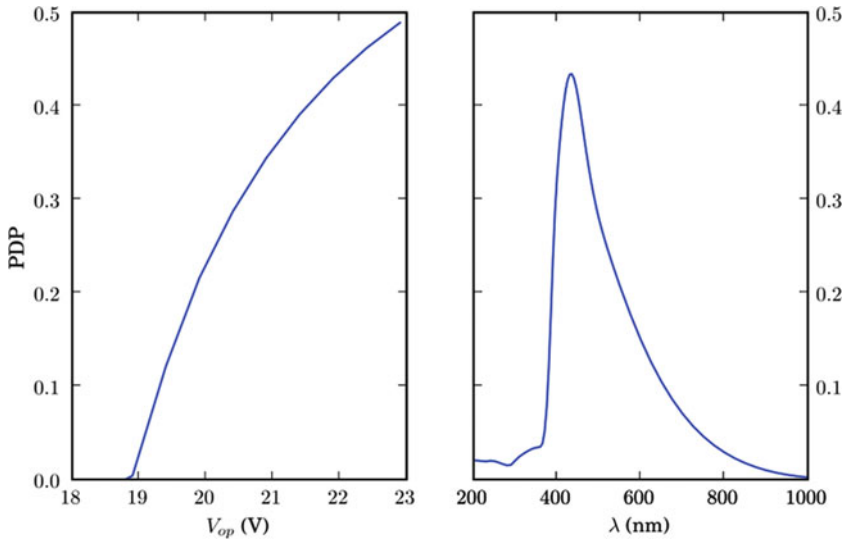

Fig. 7.9 Modeled photon detection probability vs. applied voltage for impinging light with a wavelength of 425 nm (*left*) and vs. wavelength at an excess bias voltage of 3 V (*right*)

7.2.9 Wavelength Discrimination

When the depth of a SPAD's junction is increased, sensitivity to blue light will decrease more quickly than sensitivity to red light. This technique can be used for wavelength discrimination between junctions at different depths, and SPADs with different junction characteristics on a single chip have been realized [27]. In addition, modulating the excess bias will change the avalanche probability as a function of electron–hole generation depth, and allow some portion of wavelength information to be extracted from a single junction [28]. However, the first method is strongly dependent on the CMOS process having the correct layers for two different types of SPADs, and it can only discriminate a small number of wavelengths, while the second technique is not very sensitive due to the small fluctuations in PDP ratio. Of course coupling with optical filters or prisms can create wavelength sensitivity in a particular system that utilizes SPADs. Optical pass-band filters are especially common in 3D imaging, as they allow the rejection of background light.

7.3 Fabricating Monolithic SPADs

7.3.1 Vertical Versus Planar SPADs

There exist two main implementation styles for both APDs and SPADs. In the first style, known as reach-through APD (RAPD), one builds a p + –π–p–n structure [3].

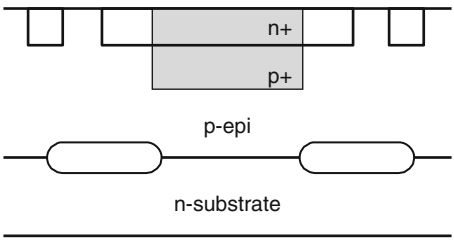

Fig. 7.10 Cross-section of an early planar SPAD

When reverse biased, the depletion region extends from the cathode to the anode. Thus, the multiplication region is deep in the p/n+ junction. Due to the depth of the multiplication region, this device is indicated for absorption of red and NIR photons up 1.1 μm (for silicon). Since the photoelectrons drift until they reach the multiplication region, a larger timing uncertainty is generally observed.

The second known implementation style is compatible with planar CMOS processes, and it involves shallow or medium depth p or n layers to form high-voltage pn junctions. Cova and others have investigated devices designed in this style since the 1970s, yielding a number of structures [2]. All these structures have in common a p–n junction and a zone designed to prevent premature edge breakdown (PEB). An example of the early structures is shown in Fig. 7.10. In [31], n + /p+ enrichment in p-substrate was used, while PEB was prevented by confining p+ enrichment to the center of the APD.

More recently, many authors have developed devices using dedicated planar and nonplanar processes, achieving superior performance in terms of sensitivity and noise. A good example is the work of Kindt [32]. The main disadvantage of using dedicated processes is generally the lack of libraries that can support complex functionalities and deep-submicron feature sizes, thus limiting array sizes. An interesting alternative is the use of a hybrid approach whereby the APD array and ancillary electronics are implemented in two different processes, each optimized for performance and speed, respectively [33]. If the ancillary electronics is implemented in CMOS, high degrees of miniaturization are possible. The price to pay is increased fabrication complexity.

7.3.2 Implementation in Planar Processes

7.3.2.1 Premature Edge Breakdown

Implementing a SPAD in a planar process first involves finding a way to prevent PEB. PEB consists of spurious, preferential avalanching on the periphery of the SPAD active area, thus causing a drastic reduction of sensitivity. Several techniques exist to implement PEB prevention. In essence, the techniques have in common the reduction of the electric field at the edges and/or the increase of the breakdown

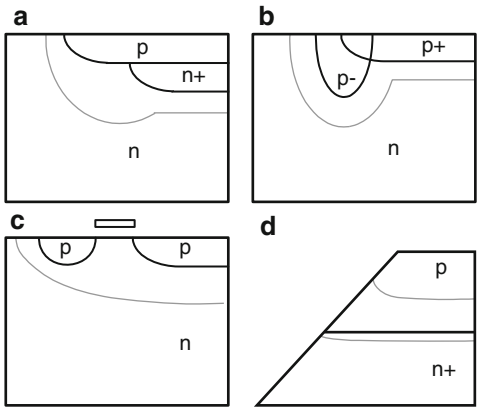

Fig. 7.11 Premature edge breakdown prevention mechanisms in planar and semiplanar processes

voltage locally, so as to maximize the probability that the avalanche is initiated in the center of the multiplication region. This is the region where the critical electric field for impact ionization is reached and, possibly, exceeded.

In Fig. 7.11, four of the most used structures are shown. In (a), the n+ layer maximizes the electric field in the middle of the diode. In (b), the lightly doped p− implant reduces the electric field at the edge of the p+ implant. In (c), a floating p implant locally increases the breakdown voltage. With a polysilicon gate, one can further extend the depletion region (gray line in the figure). Finally, in a process with trenches it is possible to decrease the electric field using the geometry of solution (d). When trenches are used, one needs to adopt techniques to prevent traps accumulated in the trench during fabrication from inducing PEB. An effective technique proposed in [16] consists of using several layers of doped semiconductor material with decreasing doping levels from the trench to the multiplication region. The purpose is to achieve short mean-free paths close to the trench, thereby forcing carriers generated there to recombine before reaching the multiplication region.

The structures of Fig. 7.11(a)–(c) are indicated in a number of CMOS technologies, while a trench-based structure (d) is mostly appropriate in deep-submicron CMOS technologies, where deep and medium tubs are not available without a major change in the fabrication process. In the remainder of the chapter, we focus our attention on schemes (b) and (d) because they require, in general no modifications to the process and thus enable the design of large SPAD array chips in standard CMOS technologies.

7.3.2.2 Quenching and Recharge Implementation

Upon design of a p–n junction capable of withstanding relatively high voltages and whereby PEB is practically prevented, the device may be biased above breakdown into Geiger mode. As discussed earlier, in this mode of operation the avalanche must be quenched to prevent destruction of the device. There exist a variety of avalanche

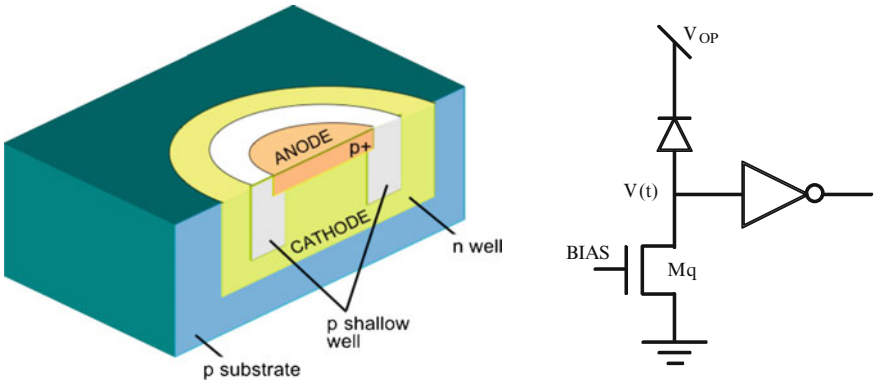

Fig. 7.12 SPAD cross-section in a conventional CMOS process (*left*); passive quench and recharge circuitries, as well as pulse shaping (*right*)

quenching techniques, active and passive, that perform this task. In active methods, the avalanche is detected, generally using a thresholder as described earlier, and stopped by acting on the bias. In passive methods, the p–n junction bias is self-adjusted, for example, by a ballast resistor. Recharge can also be active and passive. In active recharge, the bias across the diode is reestablished by a switch activated by an avalanche detector. In passive recharge, the ballast resistor performs this task.

Figure 7.12 shows the cross-section of a SPAD and simple circuitry to perform passive quenching and recharging. Upon photon detection, buildup and spread force a current through the device and the ballast resistor; this in turn develops a voltage drop that significantly raises voltage $V(t)$. Upon quenching, recharge begins bringing the device back to the idle state. This process generates a current pulse that is converted to a digital voltage level by means of a pulse shaping circuitry, also shown in the figure. The pulse shaper is also acting as an impedance adapter to drive the load of the column readout often employed in a SPAD matrix.

The ballast device may be implemented as a resistor or as an active element acting as a nonlinear resistor (M_q in Fig. 7.12). Assuming that the latter can be approximated as a linear element, recharge voltage V(t) yields:

$$V(t) \approx V_E e^{-\frac{t-t_0}{RC}}, \; t \geq t_0, \tag{7.15}$$

where R and C are estimated overall resistance and capacitance to the ground at the cathode, and V_E is the excess bias voltage, that is the voltage above breakdown at which the SPAD is biased. Time t_0 is the instant at which the quenching is complete. Upon proper selection of the bias (the signal BIAS in Fig. 7.12), M_q may also be operated in active mode, thus causing a controlled recharge. When controlled recharge is applied, V(t) becomes:

$$V(t) \approx V_E - \frac{I_R}{C}(t - t_0), \; t_0 \leq t \leq t_0 + \frac{CV_E}{I_R}, \tag{7.16}$$

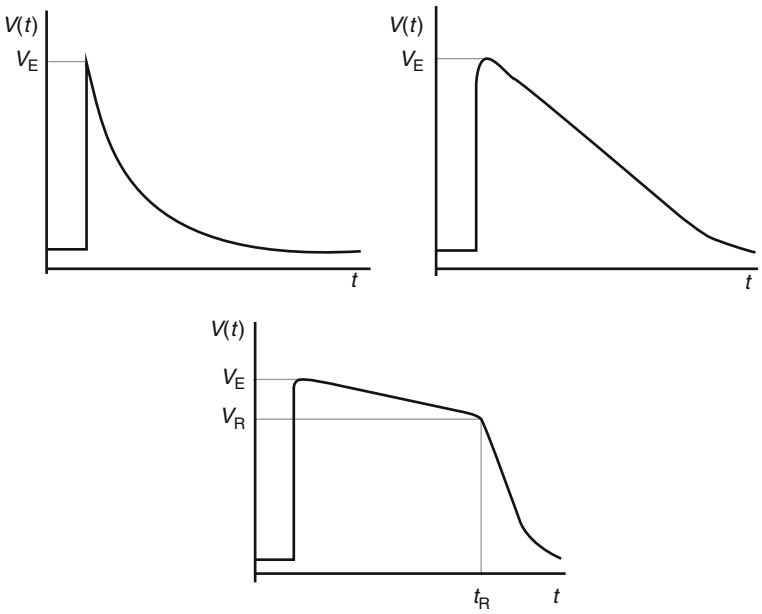

Fig. 7.13 SPAD recharge mechanisms: passive (*left*), single-slope active (*right*), and double-slope active (*center*)

where I_R is the current discharged by M_q of Fig. 7.12. In reality, the transistor may not be operated in strong inversion throughout the recharge, and thus (7.16) is only an approximation. The accuracy of this approximation is discussed in the literature.

The advantage of using an active recharge is a better control of the detection cycle and in particular of the overall time spent during quenching and recharge, known collectively as dead time. Furthermore, active recharge can also be performed in multislope mode to allow for a precise control of dead time over larger SPAD arrays, thus improving overall detection uniformity, especially in high illumination regimes. Figure 7.13 shows three typical recharge profiles. Passive recharge is the most commonly used technique, while active recharge is used in many devices whereby the recharge process has to respond to specific requirements. Figure 7.13 also shows two types of active recharge known as single- and double-slope active recharge. Single-slope recharge is simple to implement requiring only one bias per SPAD. In double-slope recharge [34], the SPAD's dead time is effectively controlled by time t_R at which the second slope is activated. If the voltage V_R achieved at this point still disables the avalanche, then it is guaranteed that the device is still in dead time regime. Thus, the dead time can be triggered by one properly timed control signal, and thus it is independent of R.

Dead time is an important parameter, as it determines the maximum count rate of a detector and thus the saturation intensity. A variety of active quenching and recharge circuits can be found in the literature whereby the differentiating factors

Table 7.1 Comparison of CMOS SPAD performance for a variety of published devices

Measurement	0.8 μm [17]	0.35 μm [20]	130 nm [16]	130 nm [42]	Unit
Timing jitter (@ 637 nm)	82[a]	80	125	200	ps
DCR (@ 300 K)	350 (V_E: 5.0 V)	750 (V_E: 3.3 V)	220 (V_E: 2.0 V)	1,221 (V_E: 0.6 V)	Hz
Active area	38	78	58	50	μm^2
Mean DCR per active area	9.2	9.6	3.8	24.4	$Hz/\mu m^2$
Breakdown (V_{BD})	25.5	17.4	12.8	14.4	V
Dead time	<40	40	100	100	ns
PDP @ 460 nm	26 (V_E: 5.0 V)	40 (V_E: 3.3 V)	26 (V_E: 2.0 V)	17.5 (V_E: 0.6 V)	%
EM spectrum (PDP > 1%)	380–900	350–1,000	380–900	350–1,000	nm
Technology	CMOS	CMOS	CMOS	CMOS	–

[a]Recently improved measurement

are complexity as well as dead time programmability and stability. One of the most important considerations in the selection of the best possible recharge mechanism is simplicity, which is of particular importance when it comes to miniaturization.

7.3.3 SPAD Nonidealities

Individual SPADs are characterized by a number of performance measures discussed earlier. In this section, these measures are summarized in terms of sensitivity, measured as *PDP*, noise performance, measured a rate of spurious pulses due to thermal events or *DCR*. Other parameters include *timing jitter*, also known somewhat inappropriately as *timing resolution*, *afterpulsing probability*, and, as mentioned earlier, *dead time*. These parameters have appeared in the literature for individual SPADs implemented in a variety of CMOS processes [15–17, 19, 33–39, 48]. Some performance parameters found in individual SPADs are described in Table 7.1 for four different implementations in CMOS submicron and deep-submicron processes achieved by the researchers of our group.

As mentioned earlier, the physical process underlying afterpulsing has been thoroughly researched in the literature. More details on afterpulsing characterization can be found in [16, 21, 43] for the devices described in this work.

7.3.4 SPAD Array Nonidealities

When implemented in an array, other performance measures become relevant to the quality of the imager. Besides dead time uniformity, timing jitter uniformity and PDP uniformity, as well as DCR uniformity and crosstalk have to be accounted for and properly characterized. Figure 7.14 shows the dead time and PDP uniformities achieved in a 32 × 32 pixel array implemented in a 0.8 μm CMOS process [17].

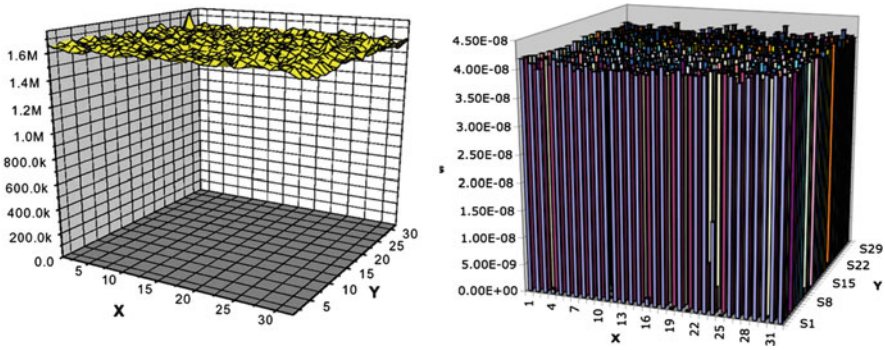

Fig. 7.14 PDP (*left*) and dead time uniformity (*right*) in a 32 × 32 array of low-pitch passively recharged pixels [17]

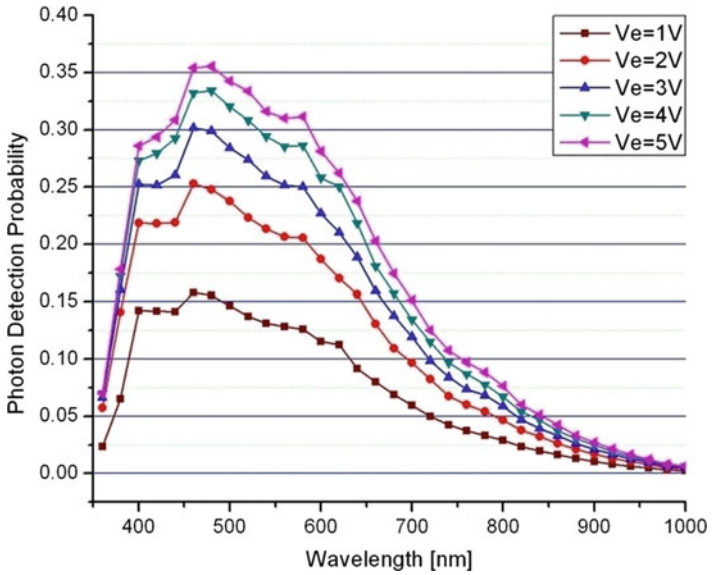

Fig. 7.15 PDP in a SPAD implemented in 130 nm CMOS process as a function of wavelength and excess bias [16]

PDP of course will also be a function of the input wavelength as derived in the earlier sections. Due to the band-gap of silicon, in CMOS SPAD implementations, the sensitivity range is mostly in the visible spectrum, with somewhat reduced near infrared and near UV PDP.

Figure 7.15 shows a plot of PDP as a function of wavelength for various values of excess bias voltage V_E for an implementation of SPADs in 130 nm [16].

Figure 7.16 shows the timing jitter on a SPAD implemented in 0.35 μm CMOS technology. In the inset of the figure, the uniformity of such jitter can be seen in

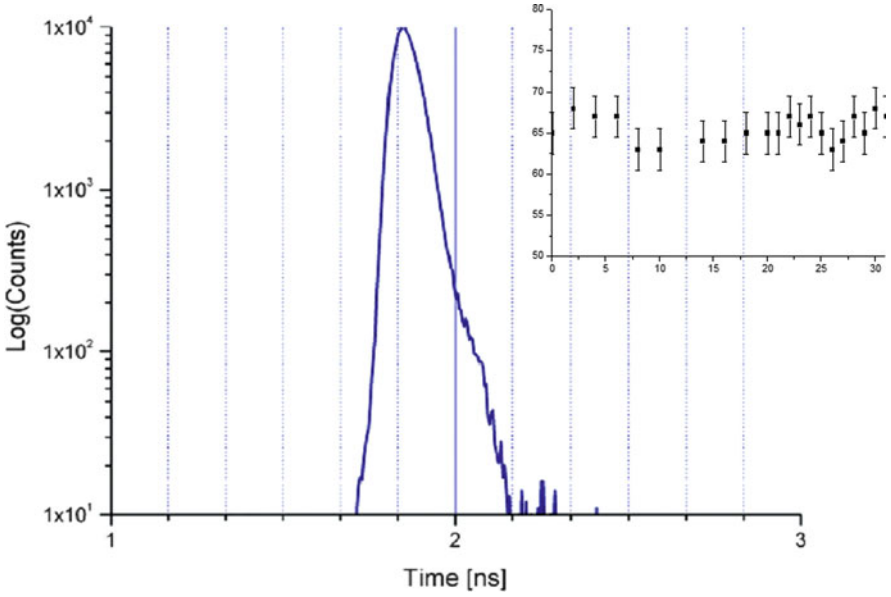

Fig. 7.16 Timing jitter performance reported in [36]. The *inset* shows the FWHM timing jitter over a 2 mm sized array of SPADs

an array of pixels integrated on the same chip measured by exposing the chip to a cone of light from a pulsed laser source. In this case, a femtosecond Ti:Sapphire laser source doubled to achieve a wavelength 488 nm was used. The uniformity in time is also very important, especially in photon-starved applications whereby long measurements may be needed to reach good accuracy.

Crosstalk issues were discussed earlier and a plot of measured crosstalk in a monolithic SPAD array implemented in 0.8 μm CMOS process is shown in Fig. 7.7. Afterpulsing is well reported in the literature. Afterpulsing in submicron CMOS SPADs is characterized in detail [20,36]. The variation of afterpulsing across a large array is generally contained due to the lithographically well-matched geometries that ensure well-matched capacitances among SPAD pixels and thus a tight control on the number of charges involved in avalanches.

7.4 Architecting SPAD Arrays

7.4.1 Basic Architectures

SPADs are dynamic devices generating a digital pulse upon detection of a photon. Unlike conventional diodes, they cannot hold a charge proportional to the overall photon count. Photopulses must be counted in situ or read outside the image sensor

and counted externally. Due to their reaction speed and low timing uncertainty, SPADs are most appropriate for photon TOA evaluation. However, even this operation must be performed upon photon detection. To address this problem, researchers have adopted a number of architectures that take advantage of the low propagation delay or high level of miniaturization achievable in standard submicron and deep-submicron CMOS technologies.

The currently available architectures are differentiated based, respectively, on processing and readout mechanisms: (1) on-chip, whereby processing is shared among all pixels and random-access or sequential readout is used; (2) in-column, whereby processing is shared among clusters of pixels, for example, columns, through event-driven mechanisms, and cluster-based readout is used; (3) in-pixel, whereby processing is pixel-specific and the result is saved locally; parallel or random-access readout is subsequently performed.

When any kind of sharing is used, trade-offs between pixel utilization, column/cluster size, and detection bandwidth are generally the result. In these cases, understanding application specifications is key to an appropriate use of the available techniques. We present a few examples hereafter as exemplifications of various architectures.

7.4.2 On-Chip Architecture

An example of an on-chip architecture is the design proposed in [17]; it comprises a matrix of 32×32 pixels, each with an independent SPAD, a quenching mechanism, pulse shaping, and column access circuitry. The readout scheme was based on random access whereby all time-sensitive operations had to be performed off-chip, and an overall jitter as low as 70 ps was measured on a pixel while the entire array was operating. An IRF of 79 ps was measured in [44]. The block diagram of the imager and the pixel schematic are shown in Fig. 7.17. The micrograph of the chip is shown in Fig. 7.18.

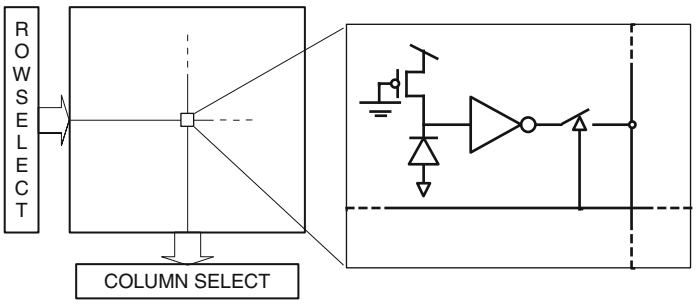

Fig. 7.17 Block diagram and pixel schematic of the 32×32 SPAD array with random access readout

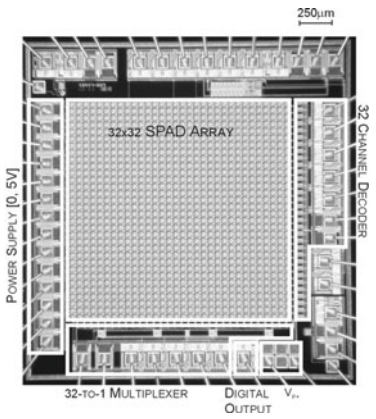

Fig. 7.18 32 × 32 SPAD array with random access readout. The chip was implemented in 0.8 μm CMOS technology

7.4.3 In-Column Architecture

Among in-column architectures, at least two variants were devised to address the readout bottleneck. The first, known as event-driven variant, consists of using the column as a bus that is addressed every time a photon is detected. The address of the row where the photon was detected is sent to the bottom of the column where the TOA is evaluated, either off chip [21, 36] or on chip [19, 45]. The second approach, known as latchless pipeline [46], uses the column as a timing-preserving delay line. Every photon triggers a pulse that is injected onto the pipeline at a precise location that corresponds to the physical place where the pixel is situated. The row information is thus encoded in the timing of the pulse arrival at the end of the pipeline, thus it can be sequentially reconstructed by a single TDC at the bottom of the column. The TDC will also detect the exact TOA of the photon within a predefined window of time.

Figure 7.19 shows a schematic of the injection mechanism at each pixel. The avalanche current is sensed and converted into a digital voltage pulse, as before, by a properly designed inverter. The L to H transition at the inverter's output pulls down node "X" through transistor T_{PD} and resistor R_{PU}, provided that gating transistor T_G is enabled by signal "GATE." The anode of the diode is intentionally set to a negative voltage, as before. T_q was sized for a dead time τ_{DT} of 40 ns and by choosing a gating window τ_G that satisfies inequality $\tau_G < \tau_D < \tau_{DT}$. When there is no activity on the preceding delay line, signal "VIN_j" is at logic level L; hence, the gate of source-degenerated transistor T_{PP} is L, thus the impedance at node "X" is dominated by the impedance at the drain of T_{PP}. When a photon is detected, a pulse is originated at this point and it is propagated toward the remainder of the delay line. When there is activity on the delay line, a logic transition L to H on "VIN_j" occurs, thus causing "X" to become a low impedance node. During this time, any photon detection in this stage will have no effect on traveling pulses but it will inject spurious pulses onto the line when it is at logic level L, hence the need for gated SPAD operation. To avoid ghost pulses, an appropriately sized NMOS was added to the cathode of the

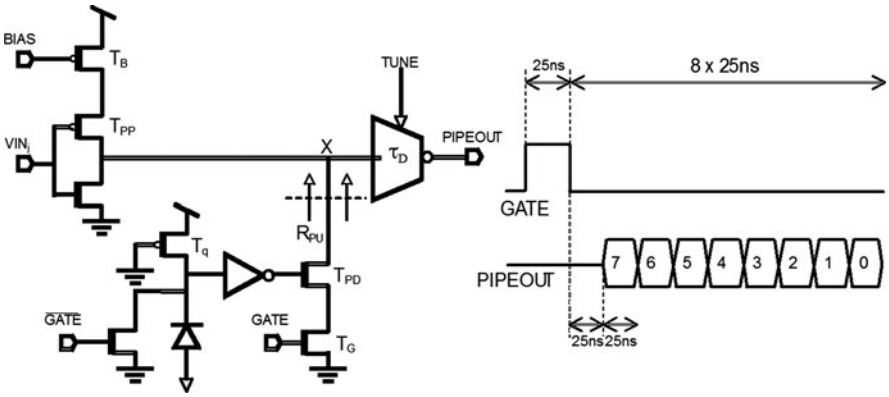

Fig. 7.19 Schematic diagram of the latchless pipelined readout

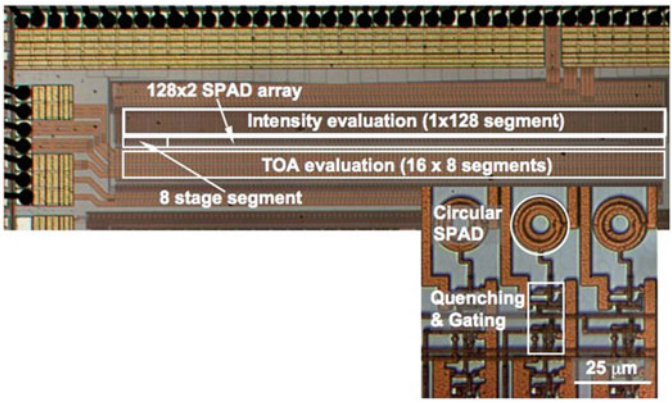

Fig. 7.20 Integrated version of the latchless pipeline implemented in 0.35 μm CMOS technology [46]

diode. A simplified timing diagram to operate the eight-stage delay line is shown in the inset of the figure. Controls "BIAS" (transistor T_B) and "TUNE" are used for coarse- and fine-tuning of the delay line, respectively. The goal is to compensate for technological variations and temperature.

A chip implementing the concept for a 128×2 SPAD array is shown in Fig. 7.20; the architecture was implemented in 0.35 μm CMOS. The chip also includes a single SPAD line for 8-bit time-uncorrelated photon counting.

7.4.4 In-Pixel Architecture

To the best of our knowledge, the first CMOS designs implementing massively parallel in-pixel single-photon processing were RADHARD2 [21] and SPSD [34].

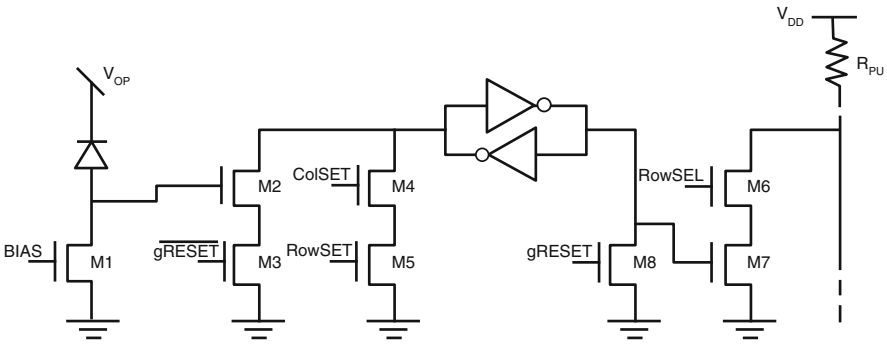

Fig. 7.21 Schematic diagram of the RADHARD2 pixel with an embedded counter. The content of the counter is read out using a simple pull-down transistor, and it may be set and reset using appropriate controls

RADHARD2 comprises a matrix of 32 × 32 pixels, each equipped with an independent counter. The content of the counters is read out in rolling shutter mode at very high speed, thus use of the chip in moderate time-resolution imaging, such as fluorescence correlated spectroscopy (FCS), is possible on a much larger pixel scale than earlier attempts [47].

In this design, the avalanche voltage is sensed by M2 that forces the latch to logic level "1." Transistor M7 acts as pull-down of the column line that is kept high by resistor R_{PU}, while M6 is the row selection switch, controlled by "RowSEL." When the column is pulled down, a buffer (not shown in the schematic) controls a pad and the output of the chip for that column is interpreted as a photon detected in the previous interval of time. Transistors M4 and M5 are controlled, respectively, by column line ("ColSET") and row line ("RowSET") to force the static memory of a specific pixel to logic level "1," irrespective of the SPAD state, for testing purposes. M8 is used to operate a global or row-based reset via signal "gRESET," whereas M3 prevents memory conflicts in case of a SPAD firing during reset. SPAD quenching and recharge are performed by transistor M1 that can be adjusted globally via signal "BIAS," so as to select a proper trade-off between dead time and afterpulsing probability. The pixel comprises ten NMOS and only two PMOS transistors, thus enabling minimization of the NWELL surface, hence ensuring a pitch of 30 μm (Fig. 7.21).

The chip micrograph is shown in Fig. 7.22. The inset also shows a detail of the pixels and their column data readout interconnect and row-wise control lines. To construct images with multibit gray levels, a high-frequency readout was put in place capable of reading an entire 1-bit frame in 2.88 μs [21].

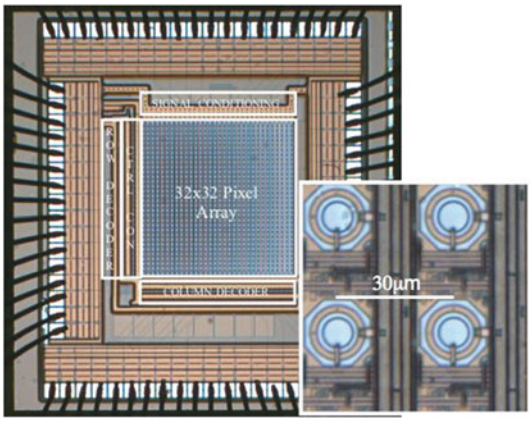

Fig. 7.22 Photomicrograph of RADHARD2, a 32 × 32 parallel-counting pixel array implemented in 0.35 μm CMOS technology. The *inset* shows a zoom of 4 × 4 pixels [21]

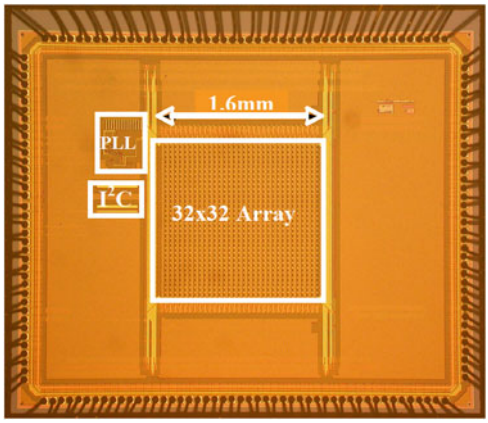

Fig. 7.23 Photomicrograph of MEGAFRAME, a 32 × 32 pixel array, capable of performing one million TOA evaluations per pixel per second at 119 ps time resolution

7.5 Trends in Monolithic Array Designs

More recently, with the implementation of the first SPADs in 130 nm CMOS technologies [37, 42] and 90 nm [48], it has been possible to integrate more functionality on pixel. The pixels of the array in the MEGAFRAME project for example comprise a multibit counter and a picosecond resolution TDC [49, 51] or time-to-amplitude converter (TAC) [50]. One of the early implementations of the MEGAFRAME concept comprises an array of 32 × 32 pixels *each* of which is capable of performing TOA measurements with picosecond resolution and digital photon counting; it was conceived to operate both in time-correlated single-photon counting (TCSPC) and in time-uncorrelated photon counting (TUPC) modes. Figure 7.23 shows a photomicrograph of the chip.

The MEGAFRAME potential has been demonstrated in a number of applications involving time-resolved imaging, such as fluorescence lifetime imaging microscopy (FLIM) [49–51], operating in TCSPC mode, thanks to its high timing resolution and

readout speed. TUPC has also shown to be effective in FCS applications, whereas other chips, such as RADHARD2, have been employed successfully [52].

More recently, the MEGAFRAME project has evolved toward higher lateral and time resolutions with the design of a new sensor comprising 160×128 pixels capable of 55 ps time resolution [52, 53]. The design includes a phase-lock loop stabilized 10-bit TDC array with an IRF of 140 ps (FWHM).

7.6 Conclusions

In this chapter, we have introduced the fundamental mechanisms at the basis of monolithic SPADs with a particular focus on devices implemented in a planar CMOS process. We have outlined the challenges of implementing SPADs in deep-submicron CMOS technologies. Finally, we have covered the issues related to large SPAD arrays and the development of high-performance devices.

Acknowledgements The authors are grateful to current and former graduate students and postdoctoral fellows of the AQUA group and the MEGAFRAME project that made this research possible. Special thanks go to Lucio Carrara, Marek Gersbach, Cristiano Niclass, and Maximilian Sergio who were responsible for the designs outlined here, as well as Fausto Borghetti, Claudio Favi, Robert Henderson, Mohammad Karami, Theo Kluter, Estelle Labonne, Yuki Maruyama, Justin Richardson, David Stoppa, and Richard Walker who codesigned the chips. The authors acknowledge Giordano Beretta, Claudio Bruschini, Dmitri Boiko, Neil Gunther, Lindsay Grant, David Li, and Luciano Sbaiz for useful discussions.

References

1. A. Rochas, *Single-Photon Avalanche Diodes in CMOS Technology*, Ph.D. dissertation, EPFL, Lausanne, Switzerland, 2003
2. S. Cova, A. Longoni, A. Andreoni, Towards picosecond resolution with single-photon avalanche diodes, Rev. Sci. Instr. **52**(3), 408–412 (1981)
3. R.J. McIntyre, Recent developments in silicon avalanche photodiodes, Measurement **3**(4), 146–152 (1985)
4. A. Spinelli, A.L. Lacaita, Physics and numerical simulation of single photon avalanche diodes, IEEE Trans. Electron Devices **44**(11), 1931–1943 (1997)
5. A. Rochas et al., Single photon detector fabricated in a complementary metal–oxide–semiconductor high-voltage technology, Rev. Sci. Instr. **74**(7), 3263–3270 (2003)
6. S.M. Sze, Physics of Semiconductor Devices. 2nd edn. Wiley-Interscience (1981).
7. C.A. Lee, R.A. Logan, R.L. Batdorf, J.J. Kleimack, W. Wiegmann, Ionization rates of holes and electrons in silicon. Phys. Rev. **134**, A761–A773 (1964)
8. A. Gulinatti, P. Maccagnani, I. Rech, M. Ghioni, S. Cova, 35 ps Time resolution at room temperature with large area single photon avalanche diodes, Electron. Lett. **41**(5), 272–274 (2005)
9. S. Cova, M. Ghioni, A. Lacaita, C. Samori, F. Zappa, Avalanche photodiodes and quenching circuits for single-photon detection, Appl. Opt. **35**(12), 1956–1976 (1996)
10. A. Lacaita, S. Cova, A. Spinelli, F. Zappa, Photon-assisted avalanche spreading in reach-through photodiodes, Appl. Phys. Lett. **62**(6), 606–608 (1993)

11. A. Lacaita, M. Mastrapasqua, M. Ghioni, S. Vanoli, Observation of avalanche propagation by multiplication assisted diffusion in p-n junctions, Appl. Phys. Lett. **57**(5), 489–491 (1990)
12. G. Kirchner, F. Koidl, Compensation of SPAD time-walk effects, J. Opt. A Pure Appl. Opt. **1**(2), 163 (1999)
13. J. Blazej, I. Prochazka, Avalanche photodiode output pulse rise-time study, SPIE Photon Counting Appl. Quantum Opt. Quantum Inf. Transf. Process. II **7355**(1), 73550M (2009)
14. G. Ripamonti, S. Cova, M. Ghioni, M. Mastrapasqua, S. Vanoli, (PS)2: a new semiconductor device for positron-sensitive picosecond detection of single optical photons, Nucl. Instrum. Methods Phys. Res. A **310**(1–2), 184–188 (1991)
15. R.K. Henderson, E.A.G. Webster, R. Walker, J.A. Richardson, L.A. Grant, A 3×3, 5 μm pitch, 3-transistor single photon avalanche diode array with integrated 11 V bias generation in 90 nm CMOS technology. Electron Devices Meeting (IEDM), 2010 IEEE International, 14.2.1–14.2.4 (2010)
16. M. Gersbach, J. Richardson, E. Mazaleyrat, S. Hardillier, C. Niclass, R. Henderson, L. Grant, E. Charbon, A low-noise single-photon detector implemented in a 130 nm CMOS imaging process, Solid-State Electron. **53**(7), 803–808 (2009)
17. C. Niclass, A. Rochas, P.A. Besse, E. Charbon, Design and characterization of a CMOS 3-D image sensor based on single photon avalanche diodes, IEEE J. Solid-State Circuits **40**(9), 1847–1854 (2005)
18. G. Ripamonti, S. Cova, Carrier diffusion effects in the time-response of a fast photodiode, Solid-state Electron. **28**(9), 925–931 (1985)
19. S. Tisa, F. Zappa, I. Labanca, On-chip detection and counting of single-photons, in *IEEE International Electron Device Meeting* (2005), pp. 815–818
20. C. Niclass, M. Sergio, E. Charbon, A single photon avalanche diode array fabricated in deep-submicron CMOS technology, IEEE Des. Autom. Test Eur. 1–6 (2006)
21. L. Carrara, C. Niclass, N. Scheidegger, H. Shea, E. Charbon, A gamma, X-ray and high energy proton radiation-tolerant CMOS image sensor for space applications, *IEEE Int. Solid-State Circuits Conference* (2009), pp. 40–41
22. S. Cova, A. Lacaita, G. Ripamonti, Trapping phenomena in avalanche photodiodes on nanosecond scale, Electron Device Lett. **12**(12), 685–687 (1991)
23. M. Ghioni, A. Gulinatti, I. Rech, F. Zappa, S. Cova, Progress in silicon single-photon avalanche diodes, IEEE J. Sel. Top. Quantum Electron. **13**(4), 852–862 (2007)
24. I. Rech, I. Labanca, G. Armellini, A. Gulinatti, M. Ghioni, S. Cova, Operation of silicon single photon avalanche diodes at cryogenic temperature, Rev. Sci. Instrum. **78**(6), 063105 (2007)
25. W. Oldham, R. Samuelson, P. Antognetti, Triggering phenomena in avalanche diodes, Electron Devices. IEEE Trans. **19**(9) 1056–1060, Sep. 1
26. P. Buzhan, B. Dolgoshein, L. Filatov, A. Ilyin, V. Kaplin, A. Karakash, S. Klemin, R. Mirzoyan, A. Otte, E. Popova, V. Sosnovtsev, M. Teshima, Large area silicon photomultipliers: performance and applications, Nucl. Instrum. Methods Phys. Res. A **567**(1), 78 – 82 (2006)
27. H. Miyamoto, M. Teshima, B. Dolgosheim, R. Mirzoyan, J. Nincovic, H. Krawczynski, SiPM development and application for astroparticle physics experiments, in *31st International Cosmic Ray Conference*, 2009
28. W. Grant, Electron and hole ionization rates in epitaxial silicon at high electric fields, Solid-State Electron. **16**(10), 1189–1203 (1973)
29. H. Finkelstein, M. Hsu, S. Esener, Dual-junction single-photon avalanche diode, Electron. Lett. **43**(22), (2007)
30. E. Charbon, C. Niclass, Controlling spectral response of photodetector for an image sensor, US Patent 7,683,308 B2, 23 Mar 2010
31. F. Zappa et al., Integrated array of avalanche photodiodes for single-photon counting, IEEE ESSDERC 600–603, (1997)
32. W.J. Kindt, *Geiger Mode Avalanche Photodiode Arrays for Spatially Resolved Single Photon Counting*, Ph.D. thesis, Delft University Press, ISBN 90–407–1845–8, 1999

33. B. Aull et al., Geiger-mode avalanche photodiodes for three-dimensional imaging, Lincoln Lab. J. **13**(2), 335–50 (2002)
34. C. Niclass, C. Favi, T. Kluter, F. Monnier, E. Charbon, Single-photon synchronous detection, IEEE J. Solid-State Circuits **44**(7), 1977–1989 (2009)
35. H. Finkelstein, M.J. Hsu, S.C. Esener STI-bounded single-photon avalanche diode in a deep-submicrometer CMOS technology, IEEE Electron Device Lett. **27**, 887 (2006)
36. C. Niclass, M. Sergio, E. Charbon, A single photon avalanche diode array fabricated in 0.35 μm CMOS and based on an event-driven readout for TCSPC experiments, in *SPIE Optics East*, Boston, 2006
37. C. Niclass, M. Gersbach, R.K. Henderson, L. Grant, E. Charbon, A single photon avalanche diode implemented in 130 nm CMOS technology, IEEE J. Sel. Top. Quantum Electron. **13**(4), 863–869 (2007)
38. D. Stoppa, L. Pacheri, M. Scandiuzzo, L. Gonzo, G.-F. Della Betta, A. Simoni, A CMOS 3-D imager based on single photon avalanche diode, IEEE Trans. Circuits Syst. **54**(1), 4–12 (2007)
39. L. Pancheri, D. Stoppa, Low-noise CMOS single-photon avalanche diodes with 32 ns dead time, in *IEEE European Solid-State Device Conference*, 2007
40. N. Faramarzpour, M.J. Deen, S. Shirani, Q. Fang, Fully integrated single photon avalanche diode detector in standard CMOS 0.18-um technology, IEEE Trans. Electron Devices **55**(3), 760–767 (2008)
41. C. Niclass, M. Sergio, E. Charbon, A CMOS 64 × 48 single photon avalanche diode array with event-driven readout, in *IEEE European Solid-State Circuit Conference*, 2006
42. J. Richardson, L. Grant, R. Henderson, A low-dark count single-photon avalanche diode structure compatible with standard nanometer scale CMOS technology, in *International Image Sensor Workshop*, 2009
43. M. Gersbach, Y. Maruyama, E. Labonne, J. Richardson, R. Walker, L. Grant, R.K. Henderson, F. Borghetti, D. Stoppa, E. Charbon, A parallel 32×32 time-to-digital converter array fabricated in a 130 nm imaging CMOS technology, in *IEEE European Solid-State Device Conference*, 2009
44. M. Gersbach, D.L. Boiko, C. Niclass, C. Petersen, E. Charbon, Fast fluorescence dynamics in nonratiometric calcium indicators, Opt. Lett. **34**(3), 362–364 (2009)
45. C. Niclass, C. Favi, T. Kluter, M. Gersbach, E. Charbon, A 128 × 128 single-photon image sensor with column-level 10-bit time-to-digital converter array, IEEE J. Solid-State Circuits **43**(12), 2977–2989 (2008)
46. M. Sergio, C. Niclass, E. Charbon, A 128 × 2 CMOS single photon streak camera with timing-preserving latchless pipeline readout, in *IEEE Intl. Solid-State Circuits Conference* (2007), pp. 120–121
47. A. Rochas, M. Gösch, A. Serov, R.S. Popovic, T. Lasser, R. Rigler, First fully integrated 2-D array of single-photon detectors in standard CMOS technology, IEEE Photon. Technol. Lett. **15**(7) (2003)
48. M. Karami, M. Gersbach, E. Charbon, A new single-photon avalanche diode in 90 nm standard CMOS technology, in *SPIE Optics+Photonics, NanoScience Engineering, Single-Photon Imaging*, 2010
49. J. Richardson, R. Walker, L. Grant, D. Stoppa, F. Borghetti, E. Charbon, M. Gersbach, R.K. Henderson, A 32 × 32 50ps resolution 10 bit time to digital converter array in 130 nm CMOS for time correlated imaging, in *IEEE Custom Integrated Circuits Conference*, 2009
50. D. Stoppa, F. Borghetti, J. Richardson, R. Walker, L. Grant, R.K. Henderson, M. Gersbach, E. Charbon, A 32 × 32-pixel array with in-pixel photon counting and arrival time measurement in the analog domain, in *IEEE European Solid-State Device Conference*, 2009
51. M. Gersbach, R. Trimananda, Y. Maruyama, M. Fishburn, D. Stoppa, J. Richardson, R.K. Henderson, E. Charbon, High frame-rate TCSPC-FLIM readout system using a SPAD-based image sensor, in *SPIE Optics+Photonics, NanoScience Engineering, Single-Photon Imaging*, 2010

52. C. Veerappan, J. Richardson, R. Walker, D.U. Li, M.W. Fishburn, Y. Maruyama, D. Stoppa, F. Borghetti, M. Gersbach, R.K. Henderson, E. Charbon, A 160 × 128 Single-Photon Image Sensor with on-Pixel 55ps 10bit Time-to-Digital Converter, *IEEE Intl. Solid-State Circuits Conference (ISSCC)*, Feb. 2011
53. R. Hull, *Properties of Crystalline Silicon*. The Institution of Engineering and Technology (1999)

Chapter 8
Single Photon CMOS Imaging Through Noise Minimization

Boyd Fowler

Abstract This chapter presents the theory and circuitry necessary to build CMOS image sensors with single photon detection capability. The chapter begins with the basic theory of CMOS image sensor photon detection. Then a discussion about additive noise systems and the sources of noise in CMOS image sensors is presented. Signal amplification and bandwidth control in low-noise CMOS image sensors are also discussed in detail. Then two complete architectures for photon counting are presented. Finally, various methods for optimizing the performance of low-noise CMOS image sensors are reviewed.

8.1 Introduction

Noise sets the fundamental detection limit on CMOS image sensors. In order to overcome this barrier and measure individual photons, CMOS image sensors must either multiply the incoming photons or photo-generated carriers or reduce the read-out noise. Devices such as EMCCDs and SPADs, discussed in the previous chapters, multiply photo-generated carriers to achieve photon-counting performance. In this chapter we focus on CMOS image sensors (CIS) that reduce the additive noise components to a level where individual photo-generated carriers can be detected.

Readout noise can be separated into two components, fixed pattern and temporal. Fixed pattern noise is a pixel to pixel, i.e., spatial, variation in either offset or gain. Fixed pattern noise does not change as a function of time. Temporal noise is the frame to frame variation of a single pixel without any illumination. The distribution of temporal noise also typically varies from pixel to pixel. In this chapter we will concentrate on temporal noise.

Temporal noise is added to each circuit in the signal path starting at the photo-detector. Each additive noise component further degrades the signal, making it more difficult to achieve single photon-counting performance. Although, many circuit

P. Seitz and A.J.P. Theuwissen (eds.), *Single-Photon Imaging*,
Springer Series in Optical Sciences 160, DOI 10.1007/978-3-642-18443-7_8,

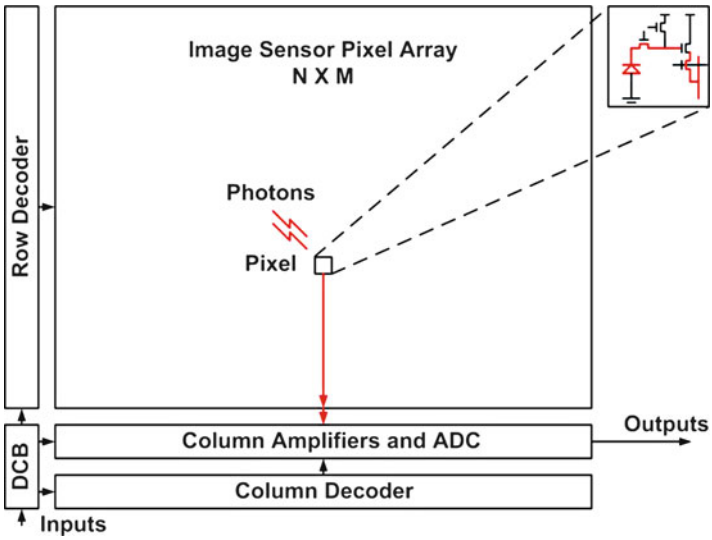

Fig. 8.1 CMOS image sensor block diagram

techniques can be used to minimize these noise components they to do not come without a cost. Noise reduction always requires tradeoffs of silicon area, speed, power dissipation, design complexity, and cost or risk. No one solution fits the needs of all sensors or applications. Throughout the remainder of this chapter we will explore these tradeoffs.

There is a large body of literature which presents specific circuits and methods for reducing temporal noise in CMOS image sensors [1–13]. In this paper, we construct a general framework for understanding these results. All of these circuits and methods use a capacitive detector to measure the amount of photo-generated charge in a specific pixel. The voltage generated by the capacitive detector is then amplified and transmitted across the chip to an analog to digital converter (ADC). The ADC could be located either on or off chip. Throughout the remainder of the chapter, we will focus on the signal path from photo-detection to the ADC. Low-noise ADC techniques will be covered in another chapter. A block diagram of a typical CIS is shown in Fig. 8.1 with the low noise signal path highlighted in red. DCB is an abbreviation for digital control block. The goal of this chapter is to explain how various CIS architectures can achieve single photon-detection.

This chapter is organized as follows: Sect. 8.2 describes the basic theory of the CIS photon detection, additive noise systems, and the noise sources in CIS. Section 8.3 discusses the roles of signal amplification and bandwidth control on the low-noise CIS. Section 8.4 describes two architectures for photo-counting CIS. Section 8.5 describes methods for optimizing the performance of low-noise CIS to achieve single photon-detection, and the final section provides summary conclusions.

8.2 Theory

In a CIS, photons are converted to electron–hole pairs through the photo-electric effect, and the resulting photo-generated carriers are integrated by the photo-detectors in each pixel. Charge is transferred from the photo-detector to a capacitor in each pixel and converted to a voltage. Typically, this voltage is then amplified, filtered, digitized, multiplexed, and finally read out of the sensor.

The probability of detecting a photon is the product of three factors: the quantum efficiency (QE) of the detector, the probability that a photo-generated carrier will be collected by the photo-detector where it was generated, and the probability of detecting the carrier after it is collected by the photo-detector.

In this section we formulate a model for QE and MTF to understand the probability of collecting a photon and the probability that a photo-carrier is collected by the right photo-detector. Then we derive the probability that a collected photo-carrier is detected by the sensor. Then we discuss the performance of systems with additive noise sources, and finally, we review the primary noise sources in CIS.

8.2.1 QE and MTF

The two primary optical factors that determine the low light performance of a CIS are quantum efficiency (QE) and modulation transfer function (MTF) [14]. The QE of the sensor is defined as the number of photo-generated carriers divided by the number of incident photons of a specific wavelength (λ), and the MTF is defined as the Fourier transform of the point spread function at a specific wavelength is MTF(f_x, λ), where f_x is spatial frequency. The point spread function measures the sensor's response to an illumination spot that is much smaller than one pixel, i.e., the spatial impulse response. Both QE and MTF can be modeled using the same methodology. In Fig. 8.2, we show a 2-D device model for a front-side illuminated CIS. This model is intended to represent a standard pinned photo-diode pixel, where the depletion region is buried below the surface to minimize dark current. The optical stack in our model only has a planar dielectric, but in many CIS it also includes a micro-lens, a color filter, and many different dielectrics. Figure 8.3 shows the linear systems model that will be used to derive QE and MTF of a CIS. $F(x)$ is the monochromatic photon flux at position x the surface of the sensor, $j_{\text{ph}}(x)$ is current density generated in the silicon, $i_{\text{ph}}(x)$ is the current collected by the photo-detector, and x_{p} is the pixel pitch. The impulse response of this system is

$$h(x) = \mathcal{T}d(x) * x_{\text{p}} \sqcap \left(\frac{x}{x_{\text{p}}}\right), \tag{8.1}$$

where $*$ is the convolution operator and

$$\sqcap(x) = \begin{cases} 1 & \text{if } |x| < 1/2, \\ 0 & \text{otherwise.} \end{cases} \tag{8.2}$$

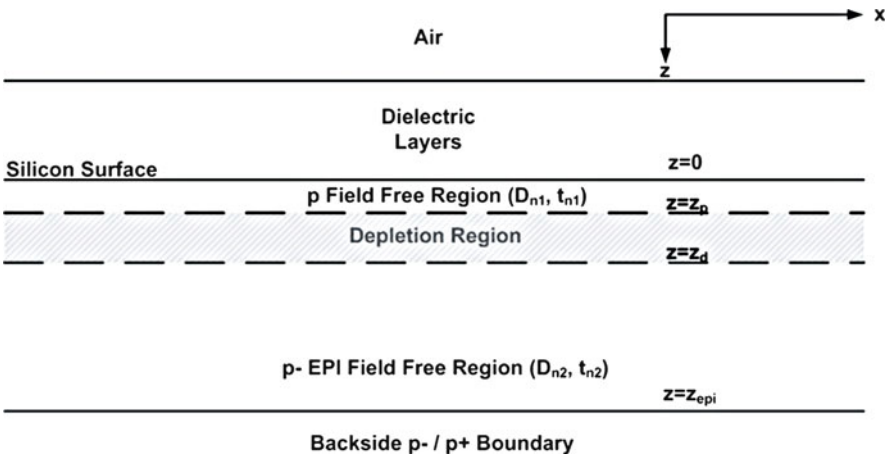

Fig. 8.2 Front-side illuminated CIS device model

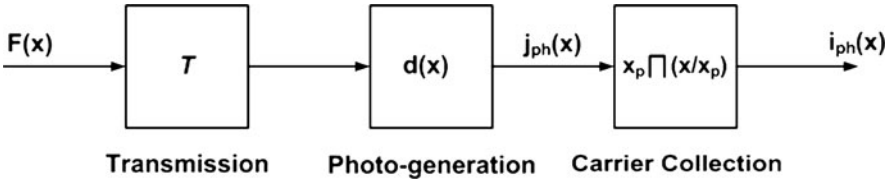

Fig. 8.3 QE and MTF linear system model

$\mathcal{T}$ is the transmission of light through the dielectric stack, $d(x)$ the diffusion of carriers in the substrate, and $x_\mathrm{p} \sqcap \left(\frac{x}{x_\mathrm{p}}\right)$ is the integration of photo-generated carriers in the pixel. The Fourier transform [15] of $h(x)$ is

$$H(f_x) = \mathcal{T} D(f_x) x_\mathrm{p}^2 \mathrm{sinc}(x_\mathrm{p} f_x), \tag{8.3}$$

where $D(f_x)$ is the Fourier transform of $d(x)$, and

$$\mathrm{sinc}(f_x) = \frac{\sin(\pi f_x)}{\pi f_x}. \tag{8.4}$$

Note that

$$\mathcal{T} D(0) = q\eta \tag{8.5}$$

and therefore the generalized QE of the sensor is

$$\eta = \mathcal{T} \frac{D(0)}{q}. \tag{8.6}$$

It can be shown that the transmitted energy of an electromagnetic wave normal to a sensor surface, modeled by a stack of three dielectrics, is

$$\mathcal{T} = 1 - \left(\frac{n_1 n_3 - n_2^2}{n_1 n_3 + n_2^2}\right)^2, \tag{8.7}$$

where n_1, n_2, and n_3 are the indices of refraction for the first, second, and third dielectric materials. For example, at 550 nm the real part of the index of refraction of air, silicon dioxide, and silicon are 1, 1.455, and 4.0844, respectively, therefore a front-side illuminated sensor will transmit about 82% of the normal incident photons. When

$$n_2 = \sqrt{n_1 n_3} \tag{8.8}$$

100% of the incident energy is transmitted to the silicon surface, i.e., $\mathcal{T} = 1$. Note that QE is also a function of incident illumination angle. A derivation of (8.7) and a complete discussion of the propagation of electro-magnetic waves through dielectrics is given in [16].

To find $D(f_x)$ we define

$$j_{\text{ph}}(x) = j_{\text{ph}}^{sc}(x) + j_{\text{ph}}^{n}(x), \tag{8.9}$$

where $j_{\text{ph}}^{sc}(x)$ is the photo-current density due to generation in the depletion region and j_{ph}^{n} is the photo-current density due to generation in the quasi-neutral p regions. Assuming a 1-D photon flux $F(x)$ at $z = 0$ then

$$D(f_x) = \frac{\mathscr{F}[j_{\text{ph}}(x)]}{\mathcal{T} F(f_x)}, \tag{8.10}$$

where $\mathscr{F}[j_{\text{ph}}(x)]$ is the Fourier transform of $j_{\text{ph}}(x)$. The generation rate at (x, z) is

$$g(x, z) = \chi \mathrm{e}^{-\chi z} F(x), \tag{8.11}$$

where χ is the absorption coefficient of silicon at wavelength λ.[1] Assuming that all of the photo-charge generated in the depletion region is collected the photo-detector then

$$j_{\text{ph}}^{sc}(x) = q \int_{z_p}^{z_d} g(x, z)\mathrm{d}z = qF(x)(\mathrm{e}^{-\chi z_p} - \mathrm{e}^{-\chi z_d}), \tag{8.12}$$

and

$$\mathscr{F}[j_{\text{ph}}^{sc}(x)] = q(\mathrm{e}^{-\chi z_p} - \mathrm{e}^{-\chi z_d}) F(f_x). \tag{8.13}$$

[1] Note that $\chi = \frac{4\pi}{\lambda}\kappa$, where κ is the imaginary part of the index of refraction, i.e., $\hat{n} = n(1 + \mathrm{i}\kappa)$.

Now we can determine $j^n_{\rm ph}(x)$ by noting that

$$j^n_{\rm ph}(x) = -qD_{n1}\frac{\partial n_p(x,z)}{\partial z}|_{z=z_p} + qD_{n2}\frac{\partial n_p(x,z)}{\partial z}|_{z=z_d}, \tag{8.14}$$

where D_{n1} is the electron diffusion constant in the quasi-neutral p region above the photo-detector, D_{n2} the electron diffusion constant in the quasi-neutral p− EPI region below the photo-detector, and $n_p(x,z)$ is the photo-generated electron concentration at (x,z). Note that, if we take the Fourier transform of (8.14) with respect to x we find

$$\mathscr{F}[j^n_{\rm ph}(x)] = -qD_{n1}\frac{\partial N_p(f_x,z)}{\partial z}|_{z=z_p} + qD_{n2}\frac{\partial N_p(f_x,z)}{\partial z}|_{z=z_d}. \tag{8.15}$$

To find $n_p(x,z)$ we need to solve the 2-D continuity equation [17] in steady state above and below the depleted region, i.e.,

$$0 = D_n\left(\frac{\partial^2 n_p}{\partial x^2} + \frac{\partial^2 n_p}{\partial z^2}\right) - \frac{n_p}{\tau_n} + g(x,z), \tag{8.16}$$

where τ_n is the recombination time of electrons in the quasi-neutral region. We have assumed that there are no electric fields in the quasi-neutral region. At the edges of the depletion region the boundary conditions are

$$n_p(x,z_p) = n_p(x,z_d) = 0, \tag{8.17}$$

and at the edges of the silicon they are

$$\frac{\partial n_p(x,0)}{\partial z} = \frac{\partial n_p(x,z_{\rm epi})}{\partial z} = 0. \tag{8.18}$$

Taking the Fourier transform of (8.16) with respect to x we find

$$0 = D_n\left((\mathrm{i}2\pi f_x)^2 N_p(f_x,z) + \frac{\partial^2 N_p}{\partial z^2}\right) - \frac{N_p}{\tau_n} + G(f_x,z), \tag{8.19}$$

where

$$G(f_x,z) = \mathscr{F}[g(x,z)] = \chi \mathrm{e}^{-\chi z}F(f_x). \tag{8.20}$$

Now we have a second-order differential equation with constant coefficients for N_p. Let

$$L_f^2 = \frac{L_n^2}{1+(2\pi f_x L_n)^2}, \tag{8.21}$$

where $L_n = \sqrt{D_n \tau_n}$ is the diffusion length of electrons in the quasi-neutral p region. Then substituting into (8.19) we obtain

$$-\frac{\partial^2 N_p}{\partial z^2} + \frac{N_p}{L_f^2} = \frac{G(f_x, z)}{D_n}. \tag{8.22}$$

The solution for $N_p(f_x, z)$ with respect to z has the form:

$$N_p(f_x, z) = c_1 \mathrm{e}^{-\frac{z}{L_f}} + c_2 \mathrm{e}^{-\frac{z}{L_f}} + c_3 G(f_x, z), \tag{8.23}$$

and

$$c_3 = \frac{L_f^2}{D_n(1 - (\chi L_f)^2)}. \tag{8.24}$$

c_1 and c_2 are determined by the boundary conditions in (8.17) and (8.18). Then the solution for N_p is plugged into (8.15) and we have a solution for $D(f_x)$. The MTF of the system is

$$\mathrm{MTF}(f_x) = \frac{D(f_x)}{D(0)} \mathrm{sinc}(x_\mathrm{p} f_x) \ \forall \ |f_x| < \frac{1}{2x_\mathrm{p}}, \tag{8.25}$$

where $D(f_x)/D(0)$ is the diffusion MTF and $\mathrm{sinc}(x_\mathrm{p} f_x)$ is the geometric MTF.

Finally, the probability that a photo-generated carrier will be collected by the photo-detector where it was generated is

$$P\{\text{photo-carrier collected by correct detector}\} = \int_{-x_\mathrm{p}/2}^{x_\mathrm{p}/2} \mathscr{F}^{-1}[\mathrm{MTF}(f_x)]\mathrm{d}x, \tag{8.26}$$

where $\mathscr{F}^{-1}[\mathrm{MTF}(f_x)]$ is the inverse Fourier transform of the MTF, i.e., the point spread function.

MTF and QE are always tradeoffs, higher QE causes lower MTF and visa versa. Figures 8.4 and 8.5 show simulated QE and MTF for a front-side illuminated CIS with 100% fill factor 6.5μm pixels assuming various p− EPI thicknesses. Thicker EPI increases the chances of absorbing more photons, and thinner EPI reduces the chance of the photo-generated carriers diffusing to adjacent pixels. A wider depletion region increases the MTF without degrading the QE. These simulations are based on the analysis methodology described in this section.

In this model we have assumed that the width of the depletion region is independent of the total integrated charge and that the photo-electron concentration is zero, i.e., $n_p(x, z) = 0$, in the depletion region. Neither of these assumptions is correct, but they do not significantly affect the model except when the depletion region is almost as wide as the EPI thickness. In a typical CIS the depletion width is less than 1 μm and the EPI thickness is greater than 5 μm. The same QE/MTF model described for a front-side illuminated sensor can be modified for

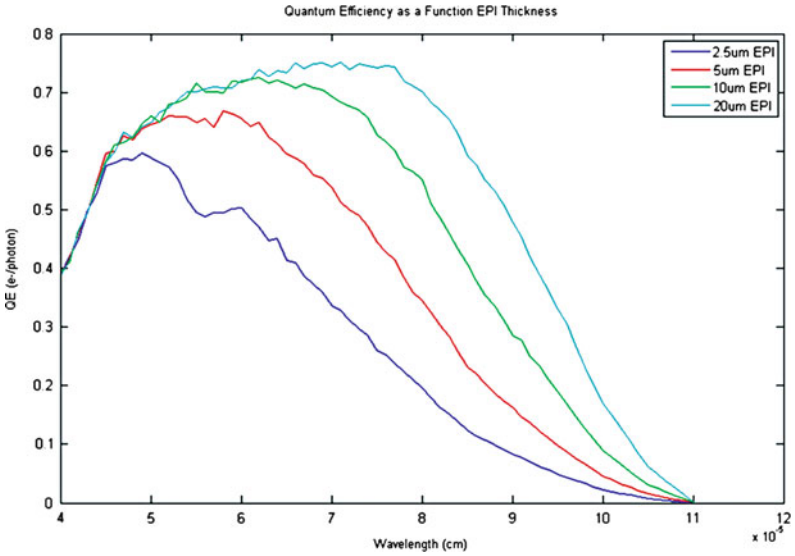

Fig. 8.4 QE vs. EPI thickness

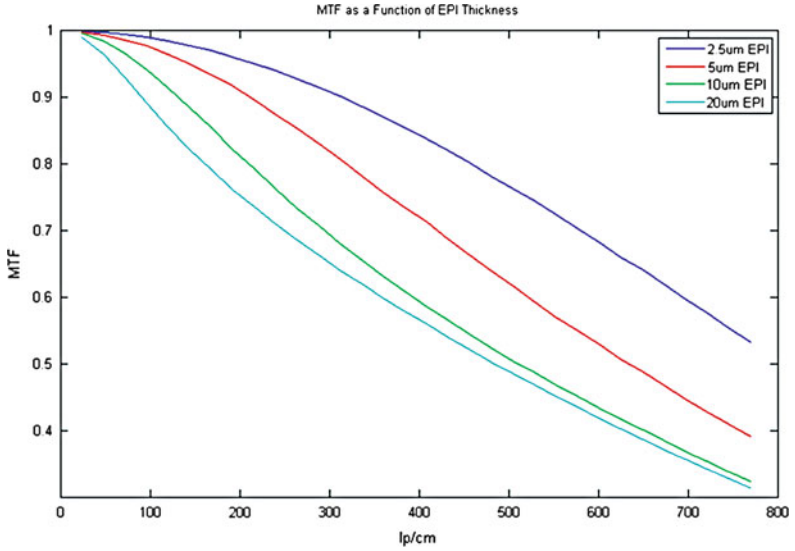

Fig. 8.5 MTF vs. EPI thickness at 600 nm

a back-side illuminated sensor by moving the depleting region and then resolving for $D(f_x)$. It is also necessary to increase the complexity of the electro-magnetic wave transmission model to account for wave propagation in a multilayer stratified medium. This can be done using a characteristic matrix methodology with complex indices of refraction as described in [16].

8.2.2 Photo-carrier Detection Probability

Both the collection of photo-generated carriers and the noise from the readout electronics can be modeled as stochastic processes. Moreover, the collection of the photo-generated carriers can be modeled as a Poisson process with parameter N_{ph}, where N_{ph}/t is the carrier arrival rate. The integral of the dark carriers can also be modeled as a Poisson process with parameter N_{dc}, where the average dark current is qN_{dc}/t. The read noise can be modeled as a zero mean Gaussian process with variance σ_{rn}^2.

We will use the following simple linear model to investigate how noise affects photo-carrier detection probability,

$$\mathbf{y}(t) = \int_0^t ((\mathbf{i}_{\text{ph}} + \mathbf{i}_{\text{dc}})/q)\mathrm{d}t + \mathbf{n}(t), \tag{8.27}$$

where $\mathbf{y}(t)$ is the measured pixel value in units of electrons at time t, $\mathbf{i}_{\text{ph}}$ the photo-carrier current, $\mathbf{i}_{\text{dc}}$ the dark current of the sensor, $\mathbf{n}(t)$ the read noise, and t is the sensor integration time. Note that $y(t)$ is real valued because $\mathbf{n}(t)$ is real valued. To simplify notation let the signal be defined as:

$$\mathbf{s}(t) = \int_0^t (\mathbf{i}_{\text{ph}}/q)\mathrm{d}t. \tag{8.28}$$

The probability density function of the sum of two independent random variables is the convolution of their probability density functions. Therefore, the probability density function of $\mathbf{y}(t)$ given that l photo-carriers are collected by the detector, i.e., $\mathbf{s}(t) = l$, is

$$f_y(y|\mathbf{s}(t) = l) = \sum_{k=0}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_{\text{rn}}} \mathrm{e}^{\frac{(y-l-k)^2}{2\sigma_{\text{rn}}^2}} \frac{(N_{\text{dc}})^k \mathrm{e}^{-N_{\text{dc}}}}{k!} \quad \forall \ |y| < \infty, \tag{8.29}$$

where k is the number of dark current carriers.

Determining if a photo-carrier has been collected can be posed as a detection or a hypothesis testing problem [18]. Moreover, given an observation of $y(t)$ we are trying to determine when we should decide that $\mathbf{s}(t) = 1$ or $\mathbf{s}(t) = 0$. One method for making this decision is based on the maximum likelihood, i.e., decide on the state of $s(t)$ based on the highest probability between the two possibilities. Using (8.29) we can construct the maximum likelihood estimator for detecting a single photo-carrier. Figure 8.6 shows the probability density functions for $\mathbf{s}(t) = 1$ and $\mathbf{s}(t) = 0$, with $N_{\text{dc}}/t = 1\text{e-/pixel/s}$, $\sigma_{\text{rn}}^2 = 0.04\text{e-}^2$, and $t = 1/30\,\text{s}$. The maximum likelihood decision rule for determining if $\mathbf{s}(t) = 1$, assuming equal probability that either zero or one electron–hole pair is collected by the photo-detector, is

$$\frac{f_y(y|\mathbf{s}(t) = 1)}{f_y(y|\mathbf{s}(t) = 0)} \geq 1. \tag{8.30}$$

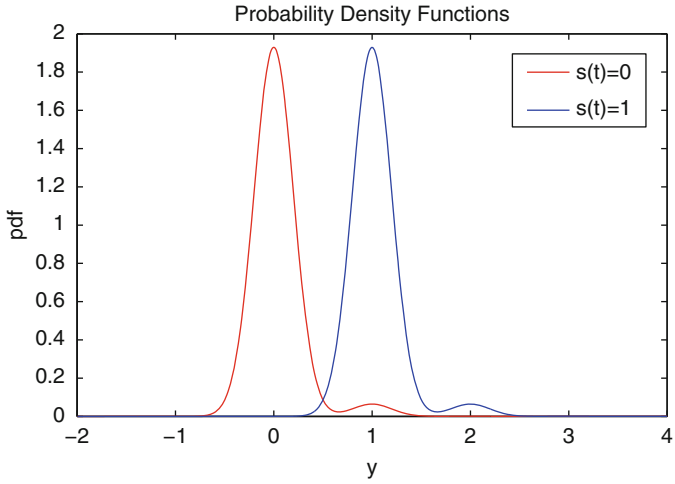

Fig. 8.6 Probability density functions

To determine the probability of detection we define

$$y \in \Omega \text{ when } \frac{f_y(y|\mathbf{s}(t)=1)}{f_y(y|\mathbf{s}(t)=0)} \geq 1, \tag{8.31}$$

i.e., Ω is the set of y values where the detector concludes that a photo-carrier has been collected. Therefore, the probability of photo-carrier detection is

$$\begin{aligned} P\{\text{photo-carrier detection}\} = &\frac{1}{2}\int_{y\in\Omega} f_y(y|\mathbf{s}(t)=1)\mathrm{d}y \\ &+\frac{1}{2}\int_{y\notin\Omega} f_y(y|\mathbf{s}(t)=0)\mathrm{d}y. \end{aligned} \tag{8.32}$$

Figure 8.7 shows the probability of photo-carrier detection as a function σ_{rn}^2, and Figure 8.8 shows the probability of photo-carrier detection as a function of the dark current rate N_{dc}/t. If the probability of detecting a photo-carrier needs be greater than 90% then σ_{rn} must be less than 0.38e− RMS, with $N_{\text{dc}} = 3.3 \times 10^{-6}$e−/pixel, and N_{dc}/t must be less than 6.7e-/pixel/s with $\sigma_{\text{rn}} = 0.05\text{e}^{-2}$ and $t = 1/30$ s.

8.2.3 Additive Temporal Noise Systems

A more comprehensive noise model for a CIS is shown in Fig. 8.9. In this model we have included additive noise components, linear gain stages, and frequency domain filters. It can be shown that the input referred noise power is

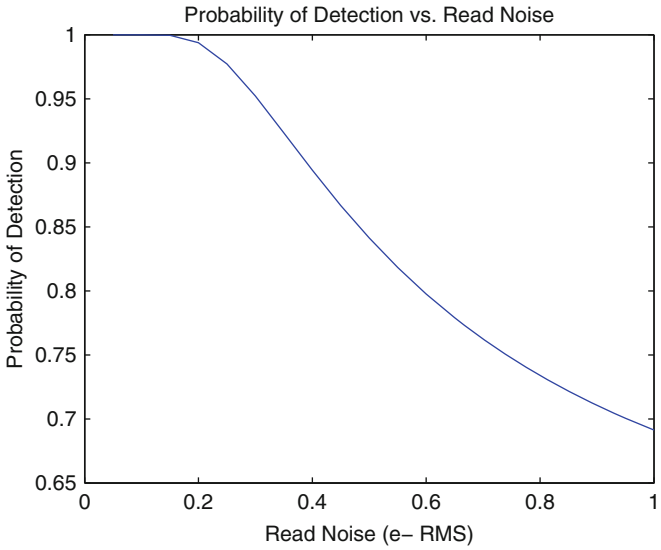

Fig. 8.7 Probability of detecting a single photo-carrier as a function of additive noise with $N_{\mathrm{dc}}/t = 10^{-4}$e−/s and $t = 1/30$ s

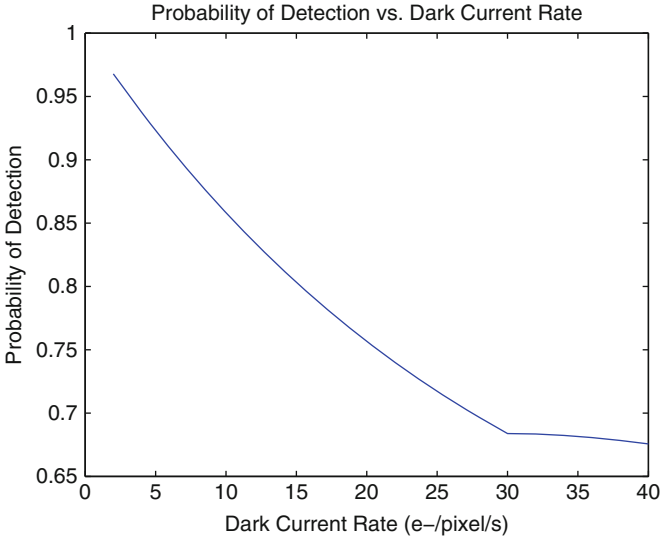

Fig. 8.8 Probability of photo-carrier detection as a function of dark current rate with $\sigma_{\mathrm{rn}} = 0.05$e− and $t = 1/30$ s

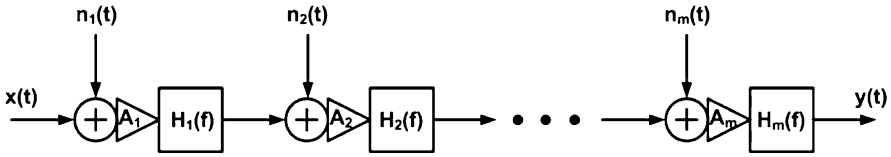

Fig. 8.9 CIS noise model

$$\sigma^2_{\text{input}} = \int_{-\infty}^{\infty} S_{n_1}(f)|H_1(f)|^2 \mathrm{d}f + \sum_{k=2}^{m} \left[\frac{\int_{-\infty}^{\infty} S_{n_k}(f)(\prod_{l=1}^{m} |H_l(f)|^2)\mathrm{d}f}{\prod_{l=1}^{k-1} A_l} \right], \tag{8.33}$$

where S_{n_k} is the power spectral density of $\mathbf{n}_k$, A_l the gain of amplifier l, H_l the filter transfer function at stage k, and m is the number of gain stages in the sensor. From (8.33) it is clear that noise from each stage in the readout chain, except the first, can be minimized by increasing the amplifier gains and/or reducing the filter bandwidths. As the noise spectrum has both high- and low-frequency components (thermal and $1/f$ noise) it is necessary for the overall filter transfer function to be bandpass. This is typically done by using correlated double sampling (CDS) to high pass filter the noise and then a series of low pass filters to attenuate the high frequency noise components. Optimizing these design parameters is critical in the development of a low-noise CIS. In most CIS designs it will be necessary to tradeoff amplifier gain with full well capacity and filter bandwidth with frame rate. Therefore, noise performance must be traded off with other parameters in the system. This will be discussed in more detail later in the final section of this Chapter. Our goal is to optimize the gain and bandwidth such that only noise from the first stage will affect the overall sensor performance.

To optimize the noise performance of a CIS it is important to understand the noise sources and how they are affected by other design parameters. In the following two Subsections we will discuss both uncorrelated and correlated noise sources in a CIS.

8.2.4 Uncorrelated Temporal Noise Sources

8.2.4.1 Thermal and Shot Noise

The physical basis for these noise sources in CMOS circuits derives from classical Boltzmann thermodynamics, i.e., the uncertainty in a system at thermal equilibrium is $\frac{1}{2}kT$ per degree of freedom, where k is Boltzmann's constant and T is absolute temperature. A complete description of noise in semiconductor devices can be found in [19]. We will assume that all of the noise sources analyzed in this chapter can be modeled as zero mean wide sense stationary random processes [18, 20]. It can be

shown that the thermal noise voltage, V_{n}, in series with a resistor r has (two-sided) power spectral density:

$$S_{V_{\mathrm{n}}}(f) = 2kTr \quad \forall \quad -\infty < f < \infty. \tag{8.34}$$

The derivation of this result was first given by Nyquist [21]. Note that the channel noise of an MOS transistor, operated *above threshold* (gate voltage such that the drift current is dominant, as opposed to *subthreshold* where diffusion current dominates) in the linear region, has the same power spectral density as (8.34) with r being the channel resistance of the MOS transistor. The gate referred noise power spectral density of an MOS transistor, operated above threshold in the saturated region, is

$$S_{V_{\mathrm{g}}}(f) = \frac{\alpha 2kT}{g_m} \quad \forall \quad -\infty < f < \infty, \tag{8.35}$$

where g_m is the transconductance of the MOS transistor and α is a constant based on the process and transistor size.[2] The channel of an MOS transistor operated in subthreshold can be modeled by two independent random diffusion processes which yield a forward and a reverse shot noise current in the channel. When an MOS transistor in subthreshold is operated below saturation, i.e., the drain to source voltage is much less than the thermal voltage kT/q, both forward and reverse diffusion currents are present and the resulting drain current noise power spectral density is

$$S_{I_{\mathrm{d}}}(f) = 2qi_{\mathrm{d}} \quad \forall \quad -\infty < f < \infty, \tag{8.36}$$

where q is the charge of an electron and i_{d} is the average drain current. As the transistor approaches saturation the reverse current goes to zero and the noise power is reduced by $1/2$, i.e.,

$$S_{I_{\mathrm{d}}}(f) = qi_{\mathrm{d}} \quad \forall \quad -\infty < f < \infty. \tag{8.37}$$

A more detailed discussion of white noise in subthreshold MOS transistors is given by Sarpeshkar, et al., in [22].[3] Note that the reverse bias current shot noise in a diode has the same power spectral density as (8.37).

8.2.4.2 Flicker and Random Telegraph Signal Noise

Both of these noise sources, also known as 1/f and RTS noise, respectively, cause low-frequency fluctuations in electronic devices. These fluctuations are clearly

[2] α is typically set to $2/3$ for long channel devices.

[3] It may be helpful to point out that equation (1) in [22] contains a typographical error which the authors corrected before publication, but which was not included in the published version. Equation (1) should read as follows: $I_f \propto \mathrm{e}^{-\frac{V_s}{U_T}}, I_r \propto \mathrm{e}^{-\frac{V_d}{U_T}}, I = I_f - I_r, I \propto \mathrm{e}^{-\frac{V_s}{U_T}}\left(1 - \mathrm{e}^{-\frac{V_d - V_s}{U_T}}\right) = I_{\mathrm{sat}}\left(1 - \mathrm{e}^{-\frac{V_{ds}}{U_T}}\right)$.

observed in the drain currents of both NMOS and PMOS FETs. They are also apparent in both the time and frequency domains. There are two well accepted theories for explaining the physical origin of 1/f and RTS noise in CMOS circuits. The first is the McWhorter model [23] and the second the is Hooge model [24]. The McWhorter model hypothesis is that the number of carriers in the MOSFET channel fluctuate randomly because of charge trapping and detrapping. The Hooge model hypothesis is that the mobility in the MOSFET channel fluctuates randomly because of interactions between the carriers and the silicon lattice. In most silicon MOSFETs 1/f noise can be attributed to a combination of carrier number and mobility fluctuation. RTS noise in MOSFETs is generally attributed to carrier number fluctuation due to trapping and detrapping of carriers at the silicon–silicon dioxide interface.

Van der Ziel [19] showed that these low-frequency fluctuations can be mathematically modeled as a summation of Lorentzian random processes with an exponentially distributed set of time constants. The key modeling difference between 1/f and RTS noise is the number of terms in the spectral sum of Lorentzian processes. Moreover, the power spectral density of 1/f noise can be modeled by:

$$S_{1/f}(f) = \sum_{l=1}^{m} \frac{2b_l\tau_l}{1+(2\pi f\tau_l)^2} \quad \forall \quad -\infty < f < \infty, \tag{8.38}$$

where m, b_l, and τ_l are constants. The power spectral density of RTS noise can be modeled by:

$$S_{\mathrm{RTS}}(f) = \frac{2b_1\tau_1}{1+(2\pi f\tau_1)^2} \quad \forall \quad -\infty < f < \infty. \tag{8.39}$$

These models are both wide sense stationary and match measured data.

MOSFET 1/f noise is most often simulated using a simplified compact device model [25–27]. In SPICE the NLEV3 gate referred noise voltage spectrum is

$$S_{v_g}(f) = \frac{\mathrm{KF}}{C_{\mathrm{ox}}WLf^{\mathrm{AF}}} \quad \forall \quad 0 < f < \infty, \tag{8.40}$$

where KF is the a bias dependent constant, C_{ox} the gate oxide capacitance per unit area, W and L are the transistor width and length, respectively, and AF is a constant. Clearly, this model is not wide sense stationary and has infinite total power. To correct for these shortcomings we will combine the standard SPICE model and the Van der Ziel sum of Lorentzians model for simulations and optimization in this chapter, i.e.,

$$S_{v_g}(f) = \frac{\mathrm{KF}}{C_{\mathrm{ox}}WL}\sum_{l=1}^{m} \frac{b_l\tau_l}{1+(2\pi f\tau_l)^2} \quad \forall \quad -\infty < f < \infty. \tag{8.41}$$

Note that both 1/f and RTS noises vary wildly from process to process and among individual MOSFETs. The only way to properly model 1/f and RTS noise in a specific process is to measure many MOSFETs under many different bias conditions and then use the ensemble averages to estimate the various parameters in (8.41) [28]. Note that b_l and τ_l can depend on the size and bias of the of MOSFET.

8.2.4.3 Charge Transfer Noise

There are at least two possible causes for this phenomenon depending on the CIS process and the sensor design. The first is incomplete charge transfer between the photo-detector and the floating diffusion node, and the second is charge trapping and de-trapping at the Si–SiO^2 interface.[4] Incomplete charge transfer results from variations in thermionic emission and/or quantum mechanical tunneling from either a potential barrier or a potential pocket between the photo-detector and the floating diffusion node. Charge trapping also causes variations in the charge transferred to the floating diffusion node.

In [29] it is shown that variations in charge transfer are very small in sensors with low lag.[5] If lag is significant then in steady state the dark current will fill the photo-detector at the same rate as carriers are emitted over the barrier, and therefore only the dark current will be seen as noise. Under transient conditions, i.e., when the illumination level changes rapidly between light and dark, there will be an additional transfer noise due to the uncertainty in the number of electrons emitted from the photo-detector during each frame. Additional information on thermionic emission from potential wells is discussed in [30, 31].

Charge trapping and detrapping at the Si–SiO_2 interface is often the dominant source of charge transfer noise in CIS. Specifically, as charge is transferred from the photo-detector to the floating diffusion node it comes in contact with the Si–SiO^2 interface near the floating diffusion node. In addition, while the transfer gate voltage is held high electrons from the floating diffusion can interact with surface traps in the transfer channel. Both of these mechanisms enable charge trapping and/or de-trapping during charge transfer. In [29] it is shown that this noise can be modeled as a sum of Lorentzian processes just like 1/f noise. It can also be shown that the variance after CDS, given a single trap frequency ν, is

$$\sigma_{ct}^2 = \frac{\mu_N}{2}(1 - \mathrm{e}^{-2\nu t}), \tag{8.42}$$

[4] It is also possible that charge trapping occurs in the bulk, but typically the density of bulk traps is very low in modern CMOS processes.

[5] It may be helpful to point out that equations (1) and (4) in [29] contain typographical errors which the authors would like to point out. Equation (1) should read $P_r(N+1, t+\mathrm{d}t|N, t) = \frac{1}{1+\mathrm{e}^{\frac{q^2N}{k\theta c_{pd}}}} \frac{(N_0-N)v_{\mathrm{th}}\mathrm{d}t}{2l_{pd}}$ and equation (4) should read $\frac{1}{\tau_N} = \frac{1}{1+\mathrm{e}^{\frac{q^2N}{k\theta c_{pd}}}} \frac{(N_0-N)v_{\mathrm{th}}}{2l_{pd}}$.

where μ_N is the average number of traps in the transfer gate and t is the on time of the transfer gate. If ν is exponentially distributed between ν_1 and ν_2 and t is small then the variance after CDS is approximately

$$\sigma_{ct}^2 \approx \mu_N \frac{t(\nu_2 - \nu_1)}{ln(\frac{\nu_2}{\nu_1})}. \tag{8.43}$$

Note that σ_{ct}^2 is in units of e-^{2}. For example, if $\mu_N = 1$, $\nu_1 = 10$ transitions/s, $\nu_2 = 10^7$ transitions/s, and $t = 1\,\mu$s then $\sigma_{ct}^2 = 0.72$ e-^{2}. As this noise occurs at the input of the sensor it is critical to minimize the charge transfer time and/or the density of traps in the transfer gate.

8.2.4.4 Capacitive Reset Noise

This noise source, also called kTC reset noise, is just another manifestation of thermal noise when a capacitor is reset to a fixed voltage through a resistor. Although this phenomenon can be thought of in many ways, the simplest is again based on thermodynamics, i.e., the energy of a system at thermal equilibrium is equal to $kT/2$ per degree of freedom in the system. A capacitor has one degree of freedom, i.e. the voltage across it, and therefore

$$\frac{cv^2}{2} = \frac{kT}{2} \Rightarrow \sigma_v^2 = \frac{kT}{c}. \tag{8.44}$$

Under nonequilibrium conditions capacitive reset noise can be reduced below kT/c and many methods have been developed including soft reset [3,4] and active reset [5, 32]. Soft reset can achieve a factor of 2 in noise power reduction and active reset can achieve more than a factor of 10 in noise power reduction. These techniques can be used to directly reduce reset noise in CIS, but they are even more useful when performed in conjunction with CDS.

8.2.5 Correlated Temporal Noise Sources

Correlated noise is more objectionable than uncorrelated noise in imagery because it is easy to detect at very low levels. For example, correlated noise sources must be about ten times lower, RMS, than the uncorrelated noise sources in the sensor to be unobservable. Correlated noise is perceived as a set of vertical and/or horizontal lines that move from frame to frame. These lines are clearly visible in both still and video imagery under low illumination levels. Clearly, it is very important to understand the causes of correlated noise and how to reduce it.

Correlated noise is not a fundamental physical process associated with the sensor. It is typically caused by design flaws either at the system or the sensor level. At the

system level these flaws include level poor power supply regulation or bias control, coupling between power supplies, and ground bounce. At the sensor level correlated noise is caused by poor power supply rejection of the analog circuitry, weak bias signals to the analog circuitry, weak power or ground distribution networks, and general bad design practices.

8.3 Amplification and Bandwidth Control

Photon counting performance in a CIS is achieved through a combination of amplification, bandwidth control, device optimization, and good design practices. There are three possible places in the readout hierarchy where amplification can be performed, the pixel, the column, and the output of the chip. To achieve the lowest noise performance signal amplification needs to be implemented as close to the photo-detector as possible. In this section we will focus on pixel and column level amplification in a CIS.

In addition to amplification, the signal bandwidth needs to be controlled to minimize unwanted high- and low-frequency noise sources. In most sensor designs the signal bandwidth is limited by the row readout rate, which is typically between 100s of nanoseconds and 100s of microseconds. In this section we discuss how the signal bandwidth can be controlled in a CIS.

8.3.1 Amplification

Amplification at the pixel level is always the best way to improve sensor low light performance, but often it is difficult to implement because of other constraints such as fill factor and/or quantum efficiency. Pixel-level amplification in MOS image sensors has been around since the late 1960s [33, 34], but it only became practical in the early 1990s because of CMOS process scaling [35]. In the literature there are three well described techniques for implementing pixel level amplification. These include source follower buffering, floating diffusion capacitance minimization, and capacitive trans-impedance amplification [36]. The ubiquitous technique of adding a source follower transistor to the pixel, i.e., the active pixel sensor (APS), allows the floating diffusion node to be isolated from the bit line capacitance. This increases the gain because the effective conversion capacitance for the photo-generated charge is much smaller. Moreover, the voltage sensed by the column amplifier is q/c_{fd}, and if c_{fd} is reduced by a factor of 10–100 then the pixel gain is much higher in an APS than in a passive pixel sensor (PPS), i.e., a pixel with only one transistor that connects the photo-detector to the bit line [37]. Figure 8.10 shows a 1T passive pixel with a surface diode, a 4T APS with a pinned photo-diode, and a 4T capacitive trans-impedance amplifier (CTIA) pixel with a pinned photo-diode.

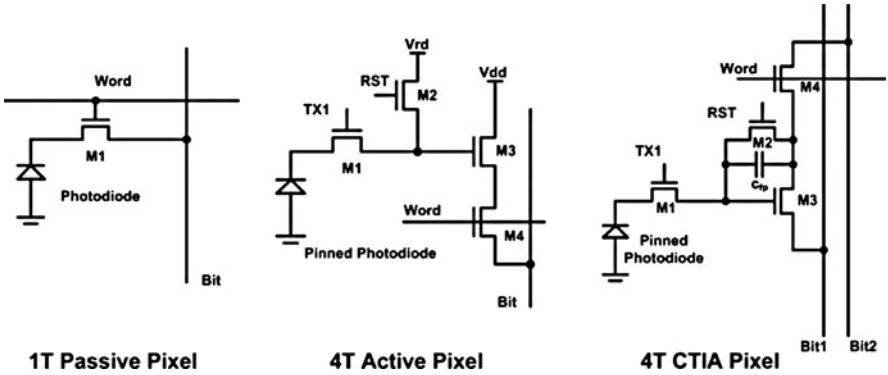

Fig. 8.10 Pixel architectures

After the addition of a source follower transistor, the next step is to increase the voltage gain at the pixel by minimizing the floating diffusion capacitance. As the voltage at the input of the source follower is approximately

$$v_{\text{fd}} = q/c_{\text{fd}}, \tag{8.45}$$

where c_{fd} is the floating diffusion capacitance of the pixel, the input referred noise can be made insignificant if c_{fd} is made small enough. For example, if $c_{\text{fd}} = 1.6 \times 10^{-19}$ F then each electron will cause a 1 V change at the input of the source follower. The continued scaling of CIS processes, and their associated parasitic capacitances, allows for smaller values of c_{fd} and higher gain at the pixel. For example, modern cell phone camera sensors often have a floating diffusion capacitance of less then 1 fF. Unfortunately, smaller floating diffusion capacitance also leads to lower full well capacity. A sensor with a 1 fF floating diffusion capacitance will typically have a full well capacity of <6,250 electrons, assuming a maximum 1 volt swing at the input of the source follower transistor.

Another method for increasing gain at the pixel is by using a capacitive trans-impedance amplifier (CTIA) as shown in Fig. 8.10. In this circuit the feedback capacitance c_{fp} controls the gain. The voltage at the output of the amplifier is approximately q/c_{fp}, assuming the voltage gain of transistor M3 is much greater than the ratio of c_{fd}/c_{fp}.

The column amplifier is the next stage in the signal amplification chain. Column amplifiers have been used since the early 1990s to improve the performance of CIS but only in 2003 were they used to achieve a read noise floor of 2e− RMS [8]. There are many different column level amplifier architectures but only a few are commonly used. Figure 8.11 shows three commonly used column amplifier circuits. The first is a single stage capacitive column amplifier, the second is a two-stage capacitive column amplifier, and finally the third is a pair of single stage capacitive

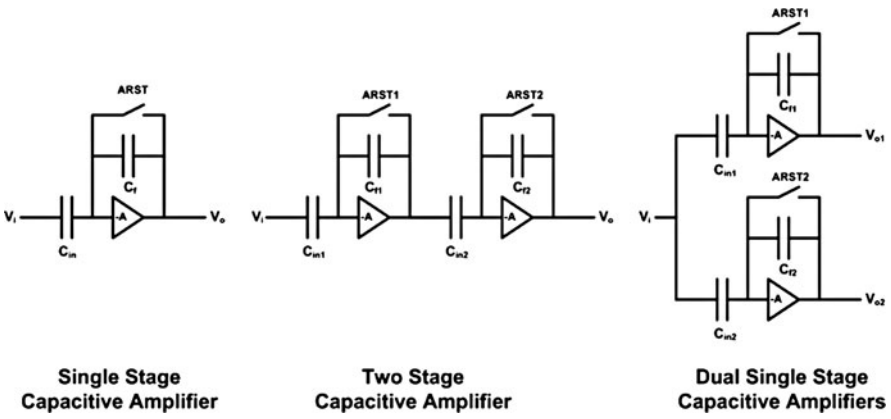

Fig. 8.11 Column amplifier architectures

amplifiers [12]. The single stage amplifier is the most common architecture for CIS. It is composed of a either a single ended MOS amplifier or a differential MOS amplifier, with at least two capacitors and a reset switch across the feedback capacitor c_f. Although, the single stage column amplifier typically has the lowest noise it often suffers from limited gain bandwidth and higher power dissipation than a two-stage amplifier.

The two stage capacitive amplifier is often used in sensors that require higher gain and/or bandwidth than a single stage amplifier can achieve. It is composed of two, lower gain, single stage capacitive amplifiers. This architecture usually achieves the lowest power dissipation for a given bandwidth, but it requires more silicon area. Its noise performance can be optimized to be similar to a single stage capacitive amplifier, but in general the lowest noise performance is achieved by using the fewest active components.

The pair of single stage column amplifiers is used in situations where high- and low-gain amplification is required simultaneously to maximize the sensors dynamic range. It is composed of two single stage capacitive amplifiers operated in parallel. One amplifier is high-gain and other is low gain. The high gain amplifier reduces noise, and the low-gain amplifier maximizes the effective full well capacity of the sensor. It requires almost twice the power dissipation of a standard single stage capacitive column amplifier.

In Fig. 8.12 three typical single ended MOS amplifier architectures are shown. The main tradeoffs between the different amplifiers are open loop voltage gain and output swing. Moreover, the simple two transistor amplifier has the lowest gain but the largest output swing, the NMOS cascode amplifier has moderate gain and moderate output swing, and the complete cascode amplifier has the highest gain but the smallest output swing. In Fig. 8.13 three typical differential MOS amplifier architectures are shown. They have the same trade-offs as the single

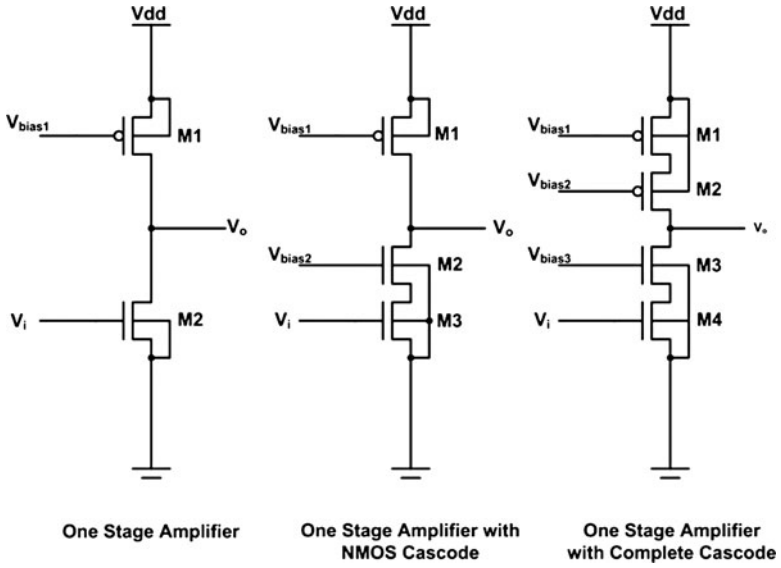

Fig. 8.12 Singled ended MOS amplifier architectures

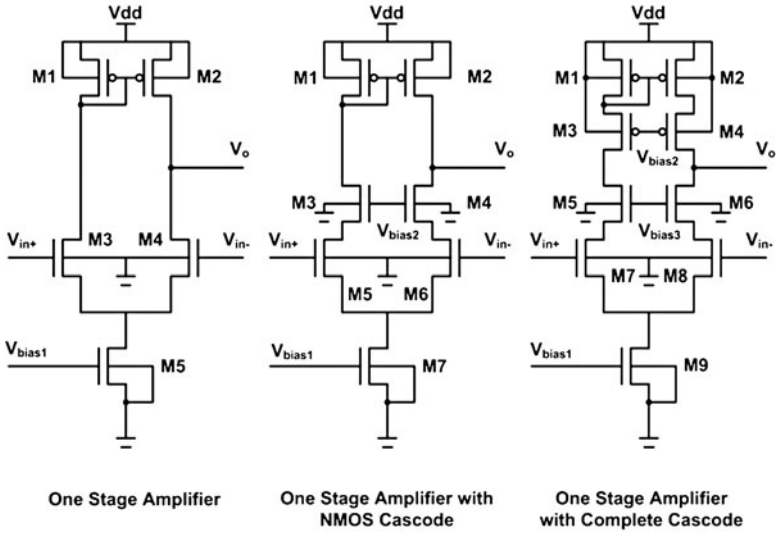

Fig. 8.13 Differential MOS amplifier architectures

ended amplifiers. The differential amplifiers have better power supply rejection than single ended amplifiers and can operate in either inverting or noninverting modes. Single ended amplifiers always have lower noise and larger swing than differential amplifiers for the same amount of silicon area and/or power dissipation.

In most CIS, signal amplification is distributed between the pixel and the column. The exact architecture is a balance among all of the design requirements including read noise, dark current, fill factor, quantum efficiency, full well capacity, die size, and/or cost.

8.3.2 Bandwidth Control

As shown in (8.33), both amplification and bandwidth control are necessary to minimize noise in a CIS. The bandwidth of the signal path is controlled at the high end by low pass filtering in the amplifier chain and at the low end by correlated double sampling.

Although it is possible to construct a multipole low pass filter in the signal path, it is not typically done because of the added circuit complexity, and limited noise improvement.[6] Therefore, most designs place a simple single pole low pass filter at the output of column amplifier just before correlated double sampling. The frequency domain transfer function of the column amplifier can be modeled by

$$H_{\text{amp}}(f) = \frac{1}{1 + (\mathrm{i} f / f_0)}, \tag{8.46}$$

where f is frequency in Hz, $\mathrm{i} = \sqrt{-1}$, and f_0 is the bandwidth of the low pass filter. f_0 is usually set between 2 and 10 times the row readout rate to minimize the high frequency thermal noise in the system. Moreover, if f_0 is set too low then the column amplifier will not settle during the sampling period, and if f_0 is too high thermal noise will not be minimized in the system. To enable multiple line rates some designs allow f_0 to be programmable by adjusting the size of a capacitor at the output of the column amplifier [13].

Correlated double sampling (CDS) is the process of measuring the reset voltage on the floating diffusion node and then transferring charge to the floating diffusion node and subtracting the two samples. This process removes both the kTC noise generated by resetting the floating diffusion node and the capacitors in the column amplifier. The CDS process can be modeled in the time domain by

$$a(t) = y(t) - \xi y(t - \Delta t), \tag{8.47}$$

where $a(t)$ is the CDS data, $y(t)$ the column level amplifier output, Δt the difference in time between the CDS samples, and ξ is the CDS efficiency factor. ξ must be very close to 1 otherwise the CDS operation will not remove all of the kTC noise. For example, if $kT = 4.16 \times 10^{-21}\,\mathrm{C}^2/\mathrm{F}$, $c = 4\,fF$, and $\xi = 0.98$ the

[6] The maximum reduction in noise power by using a multipole filter vs. a single pole filter is only $\pi/2$.

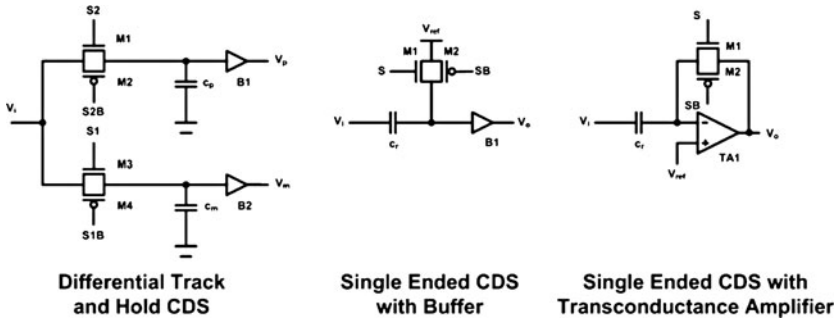

Fig. 8.14 CDS circuits

residual kTC noise will be 0.51e− RMS. It can be shown that the frequency domain transfer function for the CDS operation described in (8.47) is

$$H_{\text{cds}}(f) = (1 - \xi \mathrm{e}^{-2\pi \mathrm{i} f \Delta t}). \tag{8.48}$$

Moreover, CDS reduces kTC noise power by a factor of $(1 - \xi)^2$ and also high pass filters the 1/f and RTS noise from the MOS transistors in the signal path. Note that CDS increases the thermal noise power by a factor of two.

CDS can be performed in either the analog or the digital domain, and sometimes it is done in both. Low-noise CIS typically uses simple track and hold circuits to perform analog CDS. Figure 8.14 shows three typical CDS circuits. The first is a differential track and hold circuit. It tracks the output of the column amplifier, while S1 and S2 are high, and then stores the amplifier output when each switch is turned off. The difference between V_p and V_m is the CDS signal. This circuit adds

$$\frac{kT}{c_p} + \frac{kT}{c_m} + 2\sigma_{\text{B}}^2 \tag{8.49}$$

additional voltage noise power to the output signal, where σ_{B}^2 is the input referred noise of the unity gain buffer. The differential circuit topology improves the common mode noise rejection, but the unity gain buffers add a fixed offset to the signal. The second CDS circuit is single ended, and samples the reset level of the column amplifier while S is high and then subtracts this value from the signal when S is low. This circuit adds

$$\frac{kT}{c_r} + \sigma_{\text{B}}^2 \tag{8.50}$$

additional noise to the signal. The unity gain buffer adds a fixed offset to the signal, but this offset is on average $\sqrt{2}$ less than in the differential CDS circuit. The final CDS circuit is also single ended and operates similar to the last circuit, but

it removes the transconductance amplifier offset. Note that the transconductance amplifier is often used as the first stage of a comparator in a column level ADC.

The complete transfer function for the column amplifier and CDS is

$$H(f) = \frac{1 - \xi e^{-2\pi i f \Delta t}}{1 + (i f / f_0)}. \tag{8.51}$$

This transfer function is bandpass and by controlling f_0 and Δt the noise bandwidth can be optimized, and the total read noise of the sensor can be minimized.

8.4 Architectures

In this section we present two CIS architectures for low light applications. The first architecture uses a source follower to read out the pixel and the second uses a CTIA. For each architecture, the operation is explained and then gain and noise characteristics are derived. In the remainder of this section the transconductance, body effect conductance, and output conductance of the MOS transistor MX will be denoted g_{mx}, g_{mbx}, and g_{ox}, respectively. In addition, the gate to source capacitance, gate to drain capacitance, and channel noise current of the MOS transistor MX will be denoted c_{gsx}, c_{gdx}, and I_{nx}, respectively.

8.4.1 4T Pixel with Pinned Photodiode Column Level Amplification and CDS

The most common low-noise CIS architecture is the 4T pinned photodiode pixel with column level amplification and CDS. A schematic of the circuit we will analyze in this section is shown in Fig. 8.15. It contains a 4T pixel with a pinned photodiode, a single stage capacitive column amplifier, and a CDS sampling circuit. The output of the CDS circuit can be digitized either on or off chip.

The operation of the circuit starts with enabling the row select transistor M4 of the first row in the sensor. Then the floating diffusion node is reset through transistor M2 to Vrd while the column amplifier is also held in reset. After reset is complete RST is lowered turning off M2 and storing kTC reset noise from the floating diffusion node onto the column amplifier across the input capacitor c_{in}. Then ARST falls bringing the column amplifier out of reset. After the column amplifier settles the reset signal is sampled onto c_m by S1. Note that this also stores the kTC reset noise from the column amplifier onto c_m. Then TX1 is pulsed high and charge from the pinned photo-diode is transferred to the floating diffusion node. After the column amplifier settles again the data signal is sampled onto c_p by S2. This completes the read out of the first row of the sensor. Now the second row is read out in a similar fashion, and each row in sequence is read out until the frame is complete. This mode of operation is called rolling shutter.

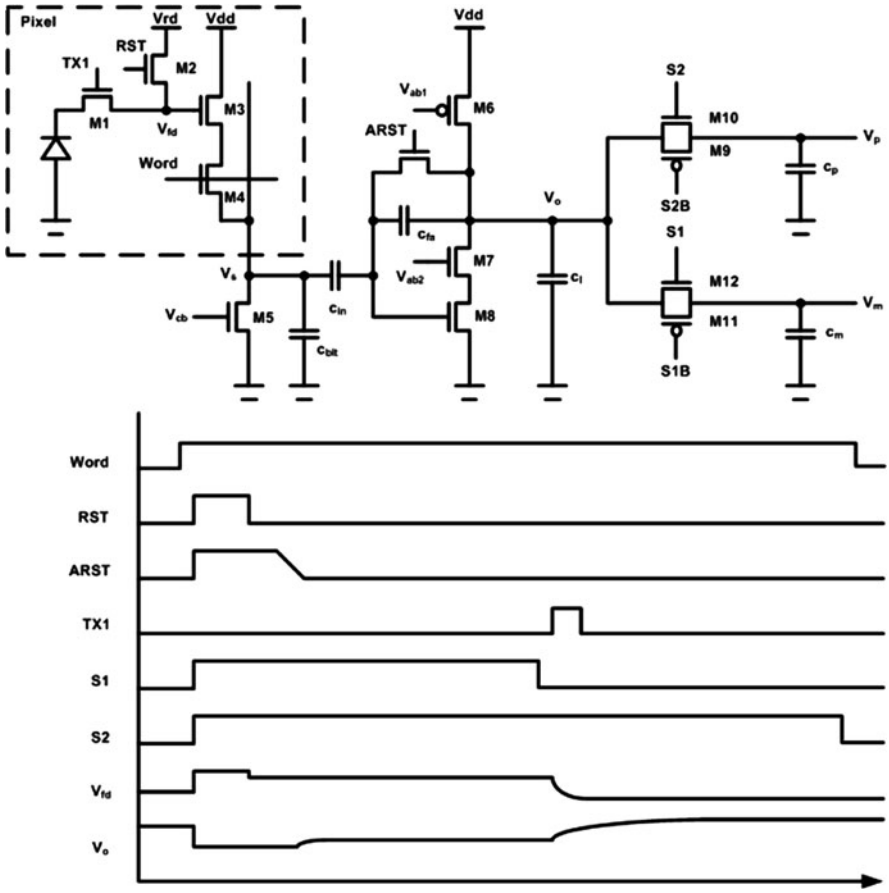

Fig. 8.15 4T pinned photo-diode pixel with single stage capacitive column amplifier and CDS, and timing waveforms

The noise performance of this circuit can be evaluated using a linear systems model assuming the noise amplitude is small in comparison to the swing of the circuit and all of the noise sources are wide sense stationary. Figure 8.16 shows a small signal model of the 4T pixel and column amplifier. In this model we have assumed that the resistance of transistors M4, M9, M10, M11, M12 and the bit line are insignificant. Assuming that the gain of the column amplifier is high then its input, V_x, is a virtual ground. Therefore, we can independently analyze the pixel and column circuitry.[7] It can be shown that the transfer function between the source follower channel noise I_{n3} and its source voltage V_s is

[7]Note that this assumption is necessary to limit the noise bandwidth, because we have assumed that several resistive components in the circuit can be neglected.

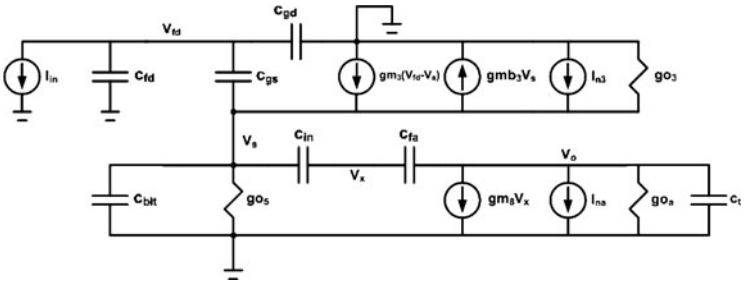

Fig. 8.16 Small signal model for 4T pinned photo-diode pixel with single stage capacitive column amplifier, and CDS

$$\frac{V_{\rm s}}{I_{\rm n3}}(f) = \frac{1}{(j2\pi f)(c_{gs3}+c_b)+g_{o3}+g_{o5}+g_{mb3}+g_{m3}(1-\beta_1)}, \tag{8.52}$$

where $c_b \approx c_{\rm bit}+c_{\rm in}$, and $\beta_1 = \frac{c_{gs3}}{c_{\rm fd}+c_{gs3}+c_{gd3}}$. The transfer function between input current $I_{\rm in}$ and source voltage $V_{\rm s}$ is

$$\frac{V_{\rm s}}{I_{\rm in}}(f) = \frac{-1}{({\rm j}2\pi f)(c_{\rm fd}+c_{gs3}+c_{gd3})} \times \frac{g_{m3}}{((\rm j2\pi f)(c_{gs3}+c_b)+g_{o3}+g_{o5}+g_{mb3}+g_{m3}(1-\beta_1))}. \tag{8.53}$$

Note that at low frequencies

$$\frac{V_{\rm s}}{I_{\rm in}}(f) \approx \frac{-A_{\rm sf}}{({\rm j}2\pi f)(c_{\rm fd}+c_{gd3}+c_{gs3}(1-A_{\rm sf}))}, \tag{8.54}$$

where

$$A_{\rm sf} = \frac{g_{m3}}{g_{m3}+g_{mb3}+g_{o3}+g_{o5}} \tag{8.55}$$

is the DC source follower voltage gain. Therefore, the source follower acts like a charge integrator with a gain of

$$\frac{qA_{\rm sf}}{c_{\rm fd}+c_{gd3}+c_{gs3}(1-A_{\rm sf})}. \tag{8.56}$$

It can also be shown that the column amplifier transfer function is

$$\frac{V_{\rm o}}{V_{\rm s}}(f) = \frac{\beta_2(({\rm j}2\pi f)c_{fa}-g_{m8})}{(j2\pi f)(c_t+c_{fa}\beta_2)+g_{m8}(1-\beta_2)+g_{oa}}, \tag{8.57}$$

where $\beta_2 = c_{\rm in}/(c_{\rm in} + c_{fa})$, $c_t = c_l + c_m + c_p/2$, and g_{oa} is the output conductance of the column amplifier. The transfer function between the column amplifier noise and the output is

$$\frac{V_{\rm o}}{I_{\rm na}}(f) = \frac{1}{({\rm j}2\pi f)(c_t + c_{fa}\beta_2) + g_{m8}(1-\beta_2) + g_{oa}}. \tag{8.58}$$

The complete signal transfer function is just the product of the source follower and the column amplifier transfer functions, i.e.,

$$\frac{V_{\rm o}}{I_i}(f) = \frac{V_{\rm s}}{I_i}(f)\frac{V_{\rm o}}{V_{\rm s}}(f). \tag{8.59}$$

At low frequencies when $g_{m3} >> g_{mb3} >> g_{o3} \geq g_{o5}$ and $g_{m8} >> g_{o3}$ then the conversion gain cg of the sensor is

$$cg \approx \frac{qA_{\rm sf}}{(c_{\rm fd} + c_{gd3} + c_{gs3}(1 - A_{\rm sf}))}\frac{c_{\rm in}}{c_{fa}}. \tag{8.60}$$

Note that cg is in units of volts per electron. For example, if $c_{\rm fd} = 0.8\,{\rm fF}$, $c_{gd} = 0.2\,{\rm fF}$, $c_{\rm in} = 1\,{\rm pF}$, $c_{fa} = 0.1\,{\rm pF}$, $c_t = 5\,{\rm pF}$, $g_{m3} = 86\,\mu{\rm S}$, $g_{mb3} = 21.5\,\mu{\rm S}$, $g_{o3} = 0.86\,\mu{\rm S}$, $g_{o5} = 0.1\,\mu{\rm S}$, $g_{oa} = 0.01\,\mu{\rm S}$ then the sensor conversion gain is approximately 0.78 mV per electron. The complete output referred noise power including CDS is

$$\sigma_n^2 = \int_{-\infty}^{\infty}\left[\left|\frac{V_{\rm s}}{I_{\rm n3}}(f)\frac{V_{\rm o}}{V_{\rm s}}(f)H_{\rm cds}(f)\right|^2 S_{I_{\rm n3}}(f) + \left|\frac{V_{\rm o}}{I_{\rm na}}(f)H_{\rm cds}(f)\right|^2 S_{I_{\rm na}}(f)\right]{\rm d}f + \frac{kT}{c_p} + \frac{kT}{c_m}. \tag{8.61}$$

Finally the input referred noise, in units of electrons RMS, is σ_n/cg. Extending the previous example, if we assume a white noise limited sensor, i.e., $S_{I_{\rm n3}}(f) = \frac{4}{3}kTg_{m3}$ and $S_{I_{\rm na}}(f) = \frac{4}{3}kTg_{\rm m8}$, $g_{m8} = 860\,\mu{\rm S}$, $g_{oa} = 0.01\,\mu{\rm S}$, $\Delta t = 10\,\mu{\rm s}$, and $\xi = 1$ then $\sigma_n^2 = 92.4n\ {\rm V}^2$ and the input referred sensor noise is 0.39e− RMS.

8.4.2 4T CTIA Pixel with Pinned Photo Diode Column Level Amplification and CDS

A 4T CTIA pixel integrates an additional gain stage next to the photo-detector, theoretically enabling lower noise operation than a standard 4T CIS. A schematic of this circuit is shown in Fig. 8.17. It contains a 4T CTIA pixel with a pinned

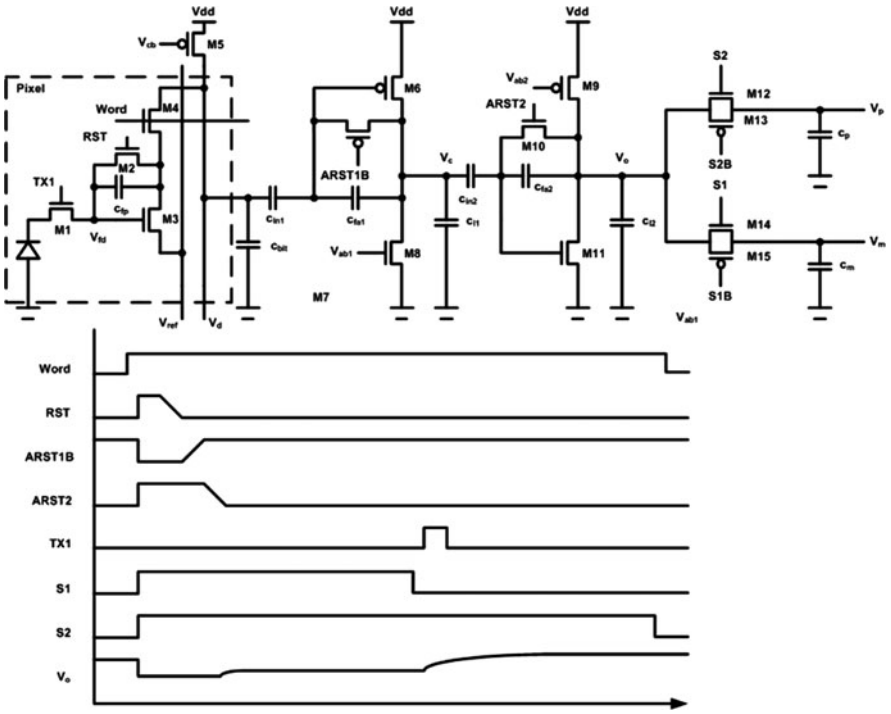

Fig. 8.17 4T CTIA pinned photo-diode pixel with single stage capacitive column amplifier and CDS, and timing waveforms

photo-diode, a two-stage capacitive column amplifier and a CDS sampling circuit. The output of the CDS circuit can be digitized either on or off chip.

The operation of the circuit starts with enabling the row select transistor M4 of the first row in the sensor. Then the floating diffusion node is reset through transistor M2 to a voltage controlled by the threshold of M3 and the bias voltage V_{cb} while the column amplifier is also held in reset. After reset is complete RST is lowered turning off M2 and storing kTC reset noise onto c_{fp}. The kTC reset noise stored on c_{fp} is sampled by the column amplifier on the input capacitor c_{in1}. Then ARST1B rises and the kTC reset noise from the first stage column amplifier is stored on the second stage column amplifier's input capacitance c_{in2}. Then ARST2 falls, bringing the column amplifier completely out of reset. Note that the reset sequence for the column is necessary to minimize noise and maximize voltage swing. After the column amplifier settles the reset signal is sampled onto c_m by S1. This stores the residual kTC reset noise from the column amplifier onto c_m. Then TX1 is pulsed high and charge from the pinned photo-diode is transferred to the input of the CTIA. After the CTIA and the column amplifier settle again, the data signal is sampled onto c_p by S2. This completes the readout of the first row of the sensor. Now the second row is read out in a similar fashion, and each row in sequence is read out until the frame is complete.

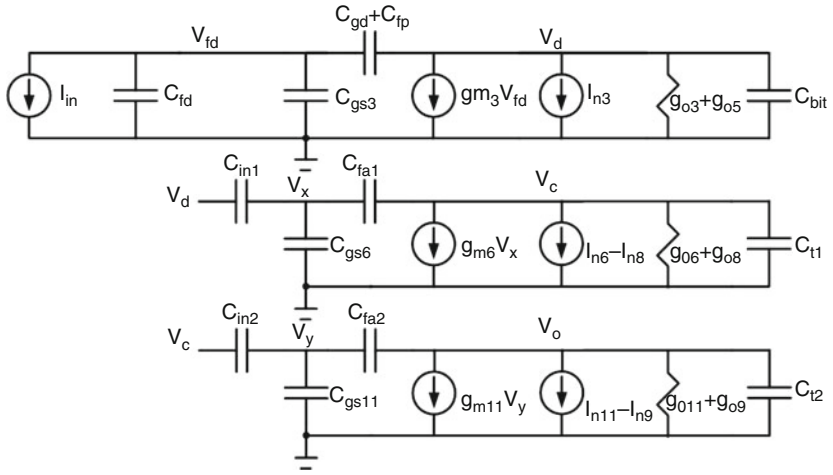

Fig. 8.18 Small signal model for 4T CTIA pinned photo-diode pixel with two stage capacitive column amplifier and CDS

Again, the noise performance of this circuit can be evaluated using a linear systems model assuming the noise amplitude is small in comparison to the swing of the circuit and all of the noise sources are wide sense stationary. Figure 8.18 shows a small signal model of the 4T pixel and column amplifier. In this model we have assumed that the resistance of transistors M4, M12, M13, M14, M15 and the bit line are insignificant. Assuming the gain of the column amplifiers are high then the inputs, i.e., V_x and V_y, are virtual grounds. Therefore, we can independently analyze the pixel and column circuitry. It can be shown that the transfer function between the source follower channel noise I_{n3} and its drain voltage V_{d} is

$$\frac{V_{\mathrm{d}}}{I_{\mathrm{n3}}}(f) = \frac{1}{(\mathrm{j}2\pi f)} \times \frac{1}{(\gamma_1 g_{m3} + g_{o3} + g_{o5} + (c_{\mathrm{bit}} + c_{\mathrm{in1}} + (c_{fp} + c_{gd3})(1-\gamma_1))(\mathrm{j}2\pi f))}. \tag{8.62}$$

The transfer function between in the input current and the CTIA drain voltage is

$$\frac{V_{\mathrm{d}}}{I_{\mathrm{in}}}(f) = \frac{g_{m3} - (c_{fp} + c_{gd3})(\mathrm{j}2\pi f)}{(j\,2\pi f)(c_{gd3} + c_{fp} + c_{\mathrm{fd}} + c_{gs3})} \times \frac{1}{(\gamma_1 g_{m3} + g_{o3} + g_{o5} + (c_{\mathrm{bit}} + c_{\mathrm{in1}} + (c_{fp} + c_{gd3})(1-\gamma_1))(\mathrm{j}2\pi f))}, \tag{8.63}$$

where

$$\gamma_1 = \frac{c_{fp} + c_{gd}}{c_{\mathrm{fd}} + c_{gs3} + c_{fp} + c_{gd3}}. \tag{8.64}$$

$$\frac{V_c}{V_{\mathrm{d}}}(f) = \frac{g_{m6} - c_{fa1}(\mathrm{j}2\pi f)}{(\mathrm{j}2\pi f)(\gamma_2 g_{m6} + g_{o6} + g_{o8} + (c_{t1} + c_{fa1}(1-\gamma_2))(\mathrm{j}2\pi f))} \tag{8.65}$$

where

$$\gamma_2 = \frac{c_{fa1}}{c_{fa1} + c_{\mathrm{in1}}}. \tag{8.66}$$

The noise transfer function between the output of the first stage of the column amplifier and node V_c is

$$\frac{V_c}{I_{n6}}(f) = \frac{1}{(\mathrm{j}2\pi f)(\gamma_2 g_{m6} + g_{o6} + g_{o8} + (c_{t1} + c_{fa1}(1-\gamma_2))(\mathrm{j}2\pi f))}. \tag{8.67}$$

The voltage transfer function between node V_c and V_{o} is

$$\frac{V_{\mathrm{o}}}{V_c}(f) = \frac{g_{m11} - c_{fa2}(\mathrm{j}2\pi f)}{(\mathrm{j}2\pi f)(\gamma_3 g_{m11} + g_{o11} + g_{o9} + (c_{t2} + c_{fa2}(1-\gamma_3))(\mathrm{j}2\pi f))} \tag{8.68}$$

where

$$\gamma_3 = \frac{c_{fa2}}{c_{fa2} + c_{\mathrm{in2}}}. \tag{8.69}$$

Finally, the noise transfer function between the last stage of the column amplifier and V_{o} is

$$\frac{V_{\mathrm{o}}}{I_{n11}}(f) = \frac{1}{(\mathrm{j}2\pi f)(\gamma_3 g_{m11} + g_{o11} + g_{o9} + (c_{t2} + c_{fa2}(1-\gamma_3))(\mathrm{j}2\pi f))}. \tag{8.70}$$

At low frequencies, if $g_{m3} >> g_{o3} + g_{o5}$ then the sensor conversion gain is

$$cg \approx \frac{q}{c_{fp} + c_{gd3}} \frac{c_{\mathrm{in1}}}{c_{fa1}} \frac{c_{\mathrm{in2}}}{c_{fa2}} \tag{8.71}$$

For example, if $c_{fp} = 0.8\,\mathrm{fF}$, $c_{gd3} = 0.2\,\mathrm{fF}$, $c_{\mathrm{in1}} = 1\,\mathrm{pF}$, $c_{fa1} = 0.33\,\mathrm{pF}$, $c_{\mathrm{in2}} = 1\,\mathrm{pF}$, and $c_{cfa2} = 0.33\,\mathrm{pF}$ then the sensor conversion gain is $\sim$1.5 mV per electron. The complete output referred noise power including CDS is

$$\sigma_n^2 = \int_{-\infty}^{\infty} \left[\left| \frac{V_{\mathrm{d}}}{I_{\mathrm{n3}}}(f) \frac{V_c}{V_{\mathrm{d}}}(f) \frac{V_{\mathrm{o}}}{V_c}(f) H_{\mathrm{cds}}(f) \right|^2 S_{I_{\mathrm{n3}}}(f) \right.$$
$$\left. + \left| \frac{V_c}{I_{n6}}(f) \frac{V_{\mathrm{o}}}{V_c}(f) H_{\mathrm{cds}}(f) \right|^2 (S_{I_{n6}}(f) + S_{I_{n8}}) \right] \mathrm{d}f$$

$$+\int_{-\infty}^{\infty}\left|\frac{V_{\rm o}}{I_{n11}}(f)H_{\rm cds}(f)\right|^2(S_{I_{n11}}(f)+S_{I_{n9}}(f))\mathrm{d}f$$
$$+\frac{kT}{c_p}+\frac{kT}{c_m}. \tag{8.72}$$

Extending the previous example, if we assume a white noise limited sensor, i.e., $S_{I_{\rm n3}}(f)=\frac{4}{3}kTg_{m3}$, $S_{I_{n6}}(f)=\frac{4}{3}kTg_{m6}$, $S_{I_{n11}}(f)=\frac{4}{3}kTg_{m11}$, $S_{I_{n8}}(f)=\frac{4}{3}kTg_{m8}$ and $S_{I_{n9}}(f)=\frac{4}{3}kTg_{m9}$, and $g_{m3}=86\,\mu$S, $g_{o3}=0.86\,\mu$S, $g_{o5}=0.1\,\mu$S, $g_{m6}=860\,\mu$S, $g_{o6}=0.43\,\mu$S, $g_{m8}=86\,\mu$S, $g_{o8}=0.043\,\mu$S, $g_{m9}=86\,\mu$S, $g_{o9}=0.043\,\mu$S, $g_{m11}=860\,\mu$S, $g_{o11}=0.43\,\mu$S, $c_{\rm fd}=0.8\,$fF, $c_{\rm bit}=2\,$pF, $c_{l1}=1\,$pF, $c_{l2}=20\,$pF, $c_m=1\,$pF, $c_p=1\,$pF, $\Delta t=10\,\mu$s, and $\xi=1$ then $\sigma_n^2=178n\,\mathrm{V}^2$ and the input referred sensor noise is 0.29e− RMS.

8.4.3 Architecture Comparison

Although the read noise performance of both architectures is similar, does either have a clear theoretical advantage over the other? To answer this question we will compare the relative input referred noise performance of the source follower and CTIA pixel circuits. The column amplifier and CDS circuitry are not included in this analysis because in principle they can be designed such that they will not affect the sensor's read noise performance. The relative noise performance of the source follower pixel is

$$\frac{|\frac{V_{\rm s}}{I_{\rm n3}}(0)|^2}{|\frac{V_{\rm s}}{I_{\rm in}}(0)|^2}, \tag{8.73}$$

and the relative noise performance of the CTIA pixel is

$$\frac{|\frac{V_{\rm d}}{I_{\rm n3}}(0)|^2}{|\frac{V_{\rm d}}{I_{\rm in}}(0)|^2}. \tag{8.74}$$

If we assume the same transistor size for M3 in both circuits and we divide (8.74) by (8.73) we find

$$\frac{|\frac{V_{\rm s}}{I_{\rm n3}}(0)\frac{V_{\rm d}}{I_{\rm in}}(0)|^2}{|\frac{V_{\rm s}}{I_{\rm in}}(0)\frac{V_{\rm d}}{I_{\rm n3}}(0)|^2}=\left(1+\frac{c_{fp}}{c_{gs}+c_{gd}+c_{\rm fd}}\right)^2>1. \tag{8.75}$$

This implies that the CTIA pixel has a slight disadvantage over the source follower pixel because (8.75) is always greater than 1. This difference is small but it shows that any additional capacitance in the pixel can reduce the input referred read noise. On the other hand, if the noise in the column amplifiers is significant then this small advantage will be lost to the higher gain of the CTIA pixel.

8.5 Low-Noise CMOS Image Sensor Optimization

In the pursuit of photon counting it is important to optimize both the electrical and optical properties of the detector. In general, the optimization process for low light CIS is a nested pair of loops. The inner loop is a sequence of analysis steps switching between electrical and optical, and the outer loop is a sequence of analysis, implementation, and measurements steps. In this section we discuss the inner loop.

8.5.1 Electrical

The key electrical parameters that must be optimized in a low light CIS are read noise, dark current, full well capacity, signal to noise ratio, and linearity. Read noise is defined as the electrical noise in the sensor when the integration time and input illumination are set to zero. Therefore, shot noise from photons and/or dark current is not included in read noise. Dark current is carrier leakage from the photo-detector, the transfer gate, or the floating diffusion node that is measured by the sensor. Dark current is measured at a fixed temperature in units of electrons per pixel per second. Dark current is a strong function of temperature and this fact can be used to determine its origin and to minimize it [38, 39]. Full well capacity is the number of the carriers that can be collected by the photo-detector and read out by the sensor without saturating. Signal to noise ratio is defined as the number of collected signal carriers divided by the total input referred RMS noise, i.e.,

$$\mathrm{SNR} = \frac{N_{\mathrm{ph}}}{\sqrt{N_{\mathrm{ph}} + N_{\mathrm{dc}} + \sigma_{\mathrm{rn}}^2}}. \tag{8.76}$$

Linearity is how close the sensor response, as a function of total collected photons, matches a straight line least squares fit. Linearity is critical in systems that are trying to quantitatively measure the number of collected electrons.

The first step in electrical performance optimization is the selection of an appropriate readout architecture, such as the two we discussed in the previous section. This selection is a function of the CIS process technology, the pixel size, and the application of the sensor. It is likely that several architectures will need to be investigated in parallel to determine the best choice for a specific sensor design. The next step in the design is determining the gain and bandwidth of each stage in the readout chain. Typically, the gain at the pixel level must be in the range of 30–300 μV/e$-$ and the gain at the column must be between about 5 V/V and 30 V/V. The gain should be adjusted such that noise from all of the circuitry other than the first transistor connected to the float diffusion is $<1/10$ of the total read noise power. In general the higher the gain the lower noise, but there are diminishing returns and higher gain also leads to smaller full well capacity and peak SNR. Moreover, since

the voltage swing V_{swing} in a CIS is limited by the power supply voltage, typically between 1.2 and 3.3 Volts in most modern CIS processes, the full well capacity is bounded by power supply voltage divided by the conversion gain of the sensor. This gives us an upper bound on the conversion gain of the sensor, i.e.,

$$cg = V_{\text{swing}}/\text{FWC} \leq Vdd/\text{FWC}. \tag{8.77}$$

As the peak SNR of the system is then bounded by the square root of the full well capacity this gives us another way to bound the conversion gain, i.e.,

$$cg \leq V_{\text{swing}}/\text{SNR}^2. \tag{8.78}$$

The bandwidth of the final low pass filter before CDS must be about 2-10 times the bandwidth of the row readout rate and the period between CDS samples cannot be more than the line rate. This bounds the noise bandwidth of the sensor readout chain.

Read noise performance of the sensor architecture can be minimized using many different methods, but in this section we will use nonlinear programming techniques [40]. Most math software packages include routines to perform nonlinear programming. For example, Octave, a free software package available on the web, includes the routine sqp. In this chapter we will use Octave and sqp to demonstrate the effectiveness of this methodology. Although it is possible to optimize the performance of the sensor across all of the circuit parameters, i.e., lengths, widths, and bias points of each MOS transistor, capacitor sizes, and CDS sampling period, it is typically more effective to reduce the search space and focus on the pixel electronics. Assuming that the column amplifier has already been designed and the read noise is limited by the pixel readout transistor, then the sensor read noise can be optimized by adjusting the size and bias conditions of this transistor. For example, using (8.61) divided by (8.60) squared as the performance function to minimize, with a set of appropriate constraints, we can optimize the input referred read noise of the sensor using sqp. In (8.61) we assume that the output referred current noise power spectral density of M3 is

$$S_{I_{\text{n3}}}(f) = \frac{4}{3}kTg_{m3} + \frac{\text{KF}g_{m3}^2}{C_{\text{ox}}W_3L_3}\sum_{l=1}^{m}\frac{2b_l\tau_l}{1+(2\pi\tau_l f)^2}, \tag{8.79}$$

where $m = 11$, $2\pi\tau_l = 10^{\frac{1-l}{2}}$ s, $b_l = 1$, $\text{KF} = 2.45 \times 10^{-26}\,\text{V}^2\,\text{F/s}$, and W_3 and L_3 are the width and length transistor M3, respectively . The transconductance of M3 is

$$g_{m3} = \sqrt{2\frac{W_3}{L_3}C_{\text{ox}}\mu_n i_{\text{d3}}}, \tag{8.80}$$

where $C_{\text{ox}} = 4.9 \times 10^{-7}\,\text{F/cm}^2$ is the oxide capacitance of transistor M3, $\mu_n = 510\,\text{cm}^2/(\text{s–V})$ the electron mobility of transistor M3, and i_{d3} is the drain current

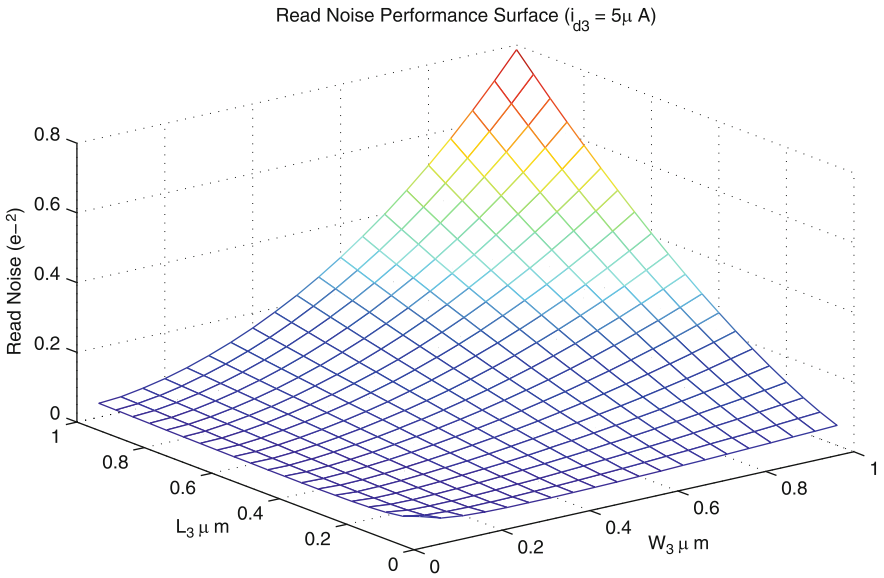

Fig. 8.19 Read noise performance surface as a function the length and width of transistor M3

of M3. We set $c_{gs3} = \frac{2C_{\text{ox}}W_3L_3}{3}$, $c_{gd} = C_{\text{ox}}W_3 10^{-5}$, $kT = 4.16 \times 10^{-21}$ J, $g_{mb3} = 0.2g_{m3}$ S, $g_{o3} = i_{\text{d3}}10^{-4}/L_3$ S, $g_{o5} = i_{\text{d3}}10^{-2}$ S, and $c_{\text{fd}} = 10^{-15}$ F. The noise from the column amplifier is

$$S_{I_{\text{na}}}(f) = 1.1\frac{4}{3}kTgm_8, \tag{8.81}$$

where $gm_8 = 1.6 \times 10^{-3}$ S. We set $c_{\text{bit}} = 2$ pF, $c_m = 10$ pF, $c_p = 10$ pF, $c_l = 20$ pF, $g_{oa} = 4 \times 10^{-7}$S, $c_{\text{in}} = 1$ pF, $c_{fa} = 40$ fF, $\Delta t = 25\,\mu$s, and $\xi = 1$. We assume the following constraints on the optimization parameters, $1\,\mu\text{A} \leq i_{\text{d}} \leq 5\,\mu\text{A}$, $0.3\,\mu\text{m} \leq W_3 \leq 2\,\mu\text{m}$, and $0.35\,\mu\text{m} \leq L_3 \leq 2\,\mu\text{m}$. Finally, we set the initial values for $i_{\text{d}} = 2.5\,\mu\text{A}$, $W_3 = 1.5\,\mu\text{m}$, and $L_3 = 0.5\,\mu\text{m}$. After six iterations sqp returns a minimum read noise of 0.17e− RMS with parameters, $i_{\text{d}} = 5\,\mu\text{A}$, $W_3 = 0.434\,\mu\text{m}$, and $L_3 = 0.35\,\mu\text{m}$. Using these circuit parameters the sensor conversion gain is 2.5mV/e−, and assuming a 1.5V swing at the output of column amplifier the full well capacity is 600e−. Therefore, the sensor's peak signal to ratio is <25. Figure 8.19 shows the read noise performance surface as a function of W_3 and L_3 assuming that $i_{\text{d3}} = 5\,\mu\text{A}$. The total optimization time for this process on an Intel 2 Duo processor running at 2.4 GHz was about 1 min. The number of parameters and constraints can be increased to explore a wider set of possible solutions, but at the expense of increased search time.

After optimizing the read noise using the previously described method, the circuit performance should be confirmed using SPICE simulation. After a few iterations, read noise optimization using sqp and SPICE simulation to verify circuit

performance, the design will converge.[8] The final step in the electrical design process is to check for problems that could cause correlated noise in the sensor. Typically, any power supply or bias that is distributed to the pixel and/or column circuitry in parallel needs to be double checked. In addition, parasitic coupling between power supplies and/or bias lines and the floating diffusion node needs to be checked. Where possible it is always helpful to route power and bias signals to the pixels vertically because this reduces correlated noise.

Assuming that the circuitry has been optimized, three key device parameters must be investigated to determine if they will degrade the low light performance of the sensor. These parameters are photo-detector lag, transfer noise, and dark current. Lag is amount of charge left in the photo-detector after the charge transfer process. Lag is caused by a potential barrier between the photo-detector and the floating diffusion. The potential barrier is the result of poor transfer gate design, insufficient voltage on the transfer gate, or insufficient transfer time. To mitigate the last two causes it is important that the transfer gate voltage and transfer gate period can be adjusted in the final sensor design. Transfer gate noise caused by traps can be mitigated by pinning the transfer gate surface, i.e. forcing holes to the surface in an NMOS gate, when the gate is not in use. This will also minimize any dark current generated in the transfer gate. Note that this typically requires a negative voltage on the transfer gate.

Dark current is usually the most significant device related problem for low-light sensors. Reduction in dark current can be achieved by optimizing pixel layout, but this is typically only a second-order effect. The primary causes for dark current, discussed in the first chapter, are a function of the CIS process. Significant reductions in dark current can only be achieved by improving the CIS process or reducing the operation temperature of the sensor. Therefore, in most applications the sensor should be designed to operate at the lowest temperature supported by the SPICE models (usually around −40°C).

8.5.2 Optical

The QE can be optimized on a front-side illuminated sensor using a microlens and a thick EPI layer, but it is difficult to achieve more than about 80%. Dielectric reflections in the optical stack always limit the performance of front-side illuminated sensors. In addition to reflection in the dielectric stack, front-side illuminated sensors can also be limited by poorly designed micro-lenses, scattering, diffraction, and/or improperly placed metal wires in the pixel. Optical design of a frontside illuminated pixel is typically done using ray tracing techniques [41,42] for pixels larger

[8] At this point in the section the reader might ask why all of the optimization is not done in SPICE. The answer to this question is that the hand analysis gives the designer much better intuition about the circuit, and it is much faster than SPICE for searching the design space.

than about 3.3 μm×3.3 μm, and using finite difference time domain electromagnetic wave simulation [43, 44] for pixels less than 3.3 μm×3.3 μm. Design and analysis of the optical stack and the EPI thickness are always required for a given sensor to achieve the highest possible QE.

Backside illuminated sensors can achieve close to 100% QE, but this only happens over a limited range of wavelengths. To achieve this performance an antireflective dielectric coating that matches the index of refraction of air to silicon needs to be deposited on the sensor. A typical antireflective coating, such as MgF_2 and/or HfO, on a 10 μm thick EPI backside illuminated sensor can achieve a peak QE between 90% and 95%. As the dielectric transmission $\mathcal{T}$ is a function of wavelength and incident illumination angle, this implies that the sensor QE is also a function of these parameters. Although backside illuminated sensors offer the highest possible QE they typically have much lower MTF than a frontside illuminated device, almost a factor of 2 sometimes, except when the EPI is fully depleted [45]. Fully depleted sensors have much higher dark current than devices that are only partially depleted.

MTF is optimized by reducing the EPI thickness and/or using a wider photodetector depletion layer. As shown in Sect. 8.2, QE and MTF are a tradeoff that must be evaluated for each application.

8.6 Conclusion

At this point in the chapter the reader should ask "Do any CMOS image sensors manufactured today achieve photon counting sensitivity?" The answer is no. Both the read noise and the dark current are typically too high to achieve photon counting sensitivity at room temperature. As the total noise power of a sensor is the sum of the read noise power and dark current shot noise power, both of these components must be very low to achieve photon counting.

The lower limit on read noise is typically determined by 1/f and RTS noise sources. Both of these noise sources are caused by imperfections in the CMOS sensor fabrication process. These imperfections typically occur at the silicon or silicon dioxide interface. Although CMOS image sensor fabrication processes are continuously improving it will likely be several more years before thermal noise dominates the sensor performance.

In addition to read noise, dark current must also be reduced such that it does not contribute any significant noise to the sensor. Dark current is caused by both process imperfections, i.e., Shockley Hall Read leakage (SHR), and fundamental device physics, i.e., diffusion leakage. Just as with 1/f and RTS noise, SHR leakage current is continuously being improved by better and better process control, but diffusion-based leakage current can only be controlled by temperature and the diffusion constant of the minority carriers in the substrate. For example, researchers have tried to use holes instead of electrons to reduce diffusion-based leakage current, i.e., hole accumulation or PMOS pinned photodiodes [46, 47].

Single photon counting sensitivity in CMOS image sensors will be achieved in the next few years. The cell phone camera market is the cash engine driving development of new process technology for CMOS image sensors, and applications such as night vision and 3D imaging are driving the demand for this technology.

In the pursuit of single photon imaging, we have presented a theoretical framework for understanding, modeling, and optimizing low-noise CIS. This includes both the electrical and optical parts of the sensor. We hope that these results will contribute to the understanding, evaluation, and implementation of low-noise CIS.

References

1. O. Yadid-Pecht, B. Mansoorian, E.R. Fossum, B. Pain, Optimization of noise and responsitivity in CMOS active pixel sensor for detection of ultra low light levels, in *Proceedings of SPIE*, vol. 3019, San Jose, January 1997
2. I. Inoue et al., Low dark current pinned photo-diode for CMOS image sensors, in *1999 IEEE Workshop on CCDs and AIS*, Nagano, June 1999
3. B. Pain et al., Analysis and enhancement of low-light-level performance of photodiode-type CMOS active pixel imagers operated with sub-threshold reset, in *1999 IEEE Workshop on CCDs and AIS*, Nagano, June 1999
4. H. Tian, B. Fowler, A. El Gamal, Analysis of temporal noise in CMOS APS, in *Proceedings of SPIE*, vol. 3649, San Jose, January 1999
5. B.A. Fowler, M.D. Godfrey, J. Balicki, J. Canfield, Low-noise readout using active reset for CMOS APS, in *Proceedings of SPIE*, vol. 3965, San Jose, January 2000, pp. 126–135
6. W. Loose et al., 2/3in CMOS image sensor for high definition television, in *2001 IEEE Workshop on CCDs and AIS*, Lake Tahoe, Nevada, June 2001
7. B. Pain et al., Reset noise suppression in two-dimensional CMOS photodiode pixels through column-based feedback-reset, in *2002 IEDM*, 2002, pp. 809–811
8. A. Krymski et al., A 2 e- Noise 1.3Megapixel CMOS sensor, in *2003 IEEE Workshop on CCDs and AIS*, Elmau, Germany, May 2003
9. K. Lee, E. Yoon, A CMOS image sensor with reset level control using current source for noise suppression, in *2004 ISSCC Digest of Technical Papers*, San Francisco, CA, February 2004, pp. 114–115
10. Y. Chen, S. Kleinfelder, CMOS active pixel sensor achieving 90 dB dynamic range with column-level active reset, in *Proceedings of SPIE*, vol. 5301, San Jose, January 2004, pp. 438–449
11. L. Kozlowski et al., A progressive 1920×1080 imaging system–on–chip for HDTV cameras, in *ISSCC Digest of Technical Papers*, San Francisco, February 2005
12. B. Fowler, X. Liu, S. Mims, J. Balicki, W. Li, H. Do, P. Vu, Wide dynamic range low light level CMOS image sensor, in *2009 Internation Image Sensor Workshop*, Bergen, Norway, June 2009
13. B. Fowler et al., A 5.5Mpixel 100 Frames/sec wide dynamic range CMOS image sensor for scientific applications, in *Proceedings of SPIE*, vol. 7536, San Jose, January 2010
14. E.G. Stevens, J.P. Lavine, An analytical aperture and two-layer carrier diffusion MTF and quantum efficiency model for solid-state image sensors. IEEE Trans. Electr. Dev. **41**(10), (1994)
15. R.N. Bracewell, *The Fourier Transform and its Applications* (McGraw Hill, NY, USA, 1986)
16. M. Born, E. Wolf, *Principles of Optics* (Pergamon Press, New York, 1986)
17. R. Muller, T. Kamins, *Device Electronics for Integrated Circuits* (Wiley, NY, USA, 1986)
18. K.S. Shanmugan, A.M. Breipohl, *Random Signals Detection, Estimation and Data Analysis* (Wiley, New York, 1988)
19. A. Van der Ziel, *Noise in Solid State Devices* (Wiley, New York, 1986)
20. A. Papoulis, *Probability Random Variables and Stochastic Processes* (McGraw Hill, New York, 1984)

21. H. Nyquist, Thermal agitation of electric charge in conductors. Phys. Rev. **32**, 110–113 (1928)
22. R. Sarpeshkar, T. Delbruck, C.A. Mead, White noise in MOS transistors and resistors. IEEE Circuits Devices Mag. **9**(6), 23–29 (1993)
23. A.L. McWhorter, 1/f Noise and related surface effects in Germanium, *Lincoln Labs Report* No. 80, May 1955
24. F.N. Hooge, A.M.H. Hoppenbrouwers, 1/f noise in continuous gold films. Physica, **45**, 386–392 (1969)
25. F.M. Klaassen, Characterization of low 1/f noise in MOS transistors. IEEE Trans. Electr. Dev. **18**(10), 887–891 (1971)
26. H. Mikoshiba, 1/f noise in n-channel silicon-gate MOS transistors. IEEE Trans. Electr. Dev. **29**, 965 (1982)
27. K.K. Hung et al., A unified model for the Flicker noise in metal-oxide-semiconductor field-effect transistors. IEEE Trans. Electr. Dev. **37**(3), 654–665 (1990)
28. B. Fowler, S. Mims, B. Frymire, 1/f noise measurement in CMOS image sensors, in *Proceedings of SPIE*, February 2005, pp. 25–33
29. B. Fowler, X. Liu, Charge transfer noise in image sensors, in *2007 Internation Image Sensor Workshop*, Ogunquit, Maine USA, June 2007
30. S. Kawai et al., Thermionic-emission based barrier height analysis for precise estimation of charge handling capacity in CCD registers. IEEE Trans. Electr. Dev. **44**(10), 1588–1592 (1997)
31. J. Janesick, *Scientific Charge Coupled Devices* (SPIE Press, New York, 2001)
32. B. Fowler, M.G.D. Godfrey, S. Mims, Reset noise reduction in capactive sensors. IEEE Trans. Circuits Syst. I **53**(8), 1658–1669 (2006)
33. S.G. Chamberlain, Photosensitivity and scanning of silicon image detector arrays. IEEE J. Solid State Circuits **4**(6), 333–342 (1969)
34. S.G. Chamberlain, V.K. Aggarwal, Photosensitivity and characterization of a solid-state integrating photodetector. IEEE J. Solid State Circuits **7**(2), 202–204 (1972)
35. S. Mendis et al., Progress in CMOS active pixel image sensors, in *Proceedings of SPIE*, San Jose, February 1994, pp. 19–29
36. J. Yang et al., A 3MPixel low-noise flexible architecture CMOS image sensor, in *ISSCC Digest of Technical Papers*, San Francisco, February 2006
37. G.P. Weckler, Operation of p-n junction photodetectors in a photon flux integration mode. IEEE J. Solid State Circuits **2**, 65–73 (1967)
38. R. Widenhorn, M.M. Blouke, A. Weber, A. Rest, E. Bodegom, Temperature dependence of dark current in a CCD, in *Proceedings of SPIE*, vol. 4669, San Jose, January 2002
39. N.V. Loukianova et al., Leakage current modeling of test structures for characterization of dark current in CMOS image sensors. IEEE Trans. Electr. Dev. **50**(1), 77–83 (2003)
40. D.A. Pierre (ed.), *Optimization Theory with Applications* (Dover, NY, USA, 1986)
41. P.B. Catrysse, X. Liu, A.E. Gamal, QE reduction due to pixel vignetting CMOS image sensors, in *Proceedings of SPIE*, vol. 3965, San Jose, January 2000
42. G. Agranov, V. Berezin, R.H. Tsai, Crosstalk and micorlens study in a color CMOS image sensor. IEEE Trans. Electr. Dev. **50**(1), 4–11 (2003)
43. C.C. Fesenmaier, Y. Huo, P.B. Catrysse, Optical confinement methods for continued scaling of CMOS image sensor pixels. Opt. Lett. **16**(25) (2008)
44. F. Hirigoyen et al., Finite-difference time domain based electro-optical methodologies to improve CMOS image sensors pixel performances, in *Proceedings of SPIE*, vol. 7723, Brussels, Belgium, May 2010
45. D.E. Groom et al., Point-spread function in depleted and partially depleted CCDs, in *Proceedings 4th ESO Workshop on Optical Detectors for Astronomy*, Garching, Germany, September 1999
46. K. Yonemoto, H. Sumi, A CMOS image sensor with simple fixed-pattern-noise reduction technology and a hole accumulation diode. IEEE J. Solid State Circuits **35**(12), 2038–2043 (2000)
47. E. Stevens et al., Low-crosstalk and low-dark-current CMOS image-sensor technology using a hole-based detector, in *ISSCC Digest of Technical Papers*, San Francisco, February 2008

Chapter 9
Architectures for Low-noise CMOS Electronic Imaging

Shoji Kawahito

Abstract This chapter discusses various types of signal readout architectures for CMOS image sensors, implementing ultra-low-noise conversion of photo-generated charge packets into digital output values. It is based on a detailed analysis of the different noise sources in a CMOS imager, the noise responses of column noise cancelling circuits using correlated double sampling (CDS) and correlated multiple sampling (CMS) techniques and a noiseless signal readout technique using a precise digitizer. Finally, a practical example for the design of a CMOS image sensor with single-photon resolution is presented, and the technological requirements for meeting the condition for room-temperature readout noise of significantly less than 1 electron are discussed.

9.1 Introduction

For single photon imaging, efforts toward achieving both high quantum efficiency which is ideally unity for visible light and noiseless signal reading will be necessary. The materials, geometrical structures, and optical characteristics of above and inside of photo detectors determines the quantum efficiency. In CMOS imagers, on the other hand, many different types of noise are superimposed on the photo signal. The total noise, which is usually expressed as an input-referred noise, or expressed as the number of noise electrons, depends on the pixel devices and circuits, entire signal readout architectures and types of signal processing and noise cancelling methods. This chapter focuses on the low-noise signal readout architectures and circuits. Fundamentals on the low-noise signal readout in CMOS image sensors are described in Chap. 8. Chapter 9 discusses types of signal readout architectures from photo-electron to digital, the noise responses of the column noise canceling circuits using correlated double sampling (CDS) and correlated multiple sampling (CMS) techniques, a noiseless signal readout technique using a precise digitizer, and finally a possibility of a CMOS single photon image sensor and a technological requirement for meeting the condition for the single photon counting.

P. Seitz and A.J.P. Theuwissen (eds.), *Single-Photon Imaging*,
Springer Series in Optical Sciences 160, DOI 10.1007/978-3-642-18443-7_9,

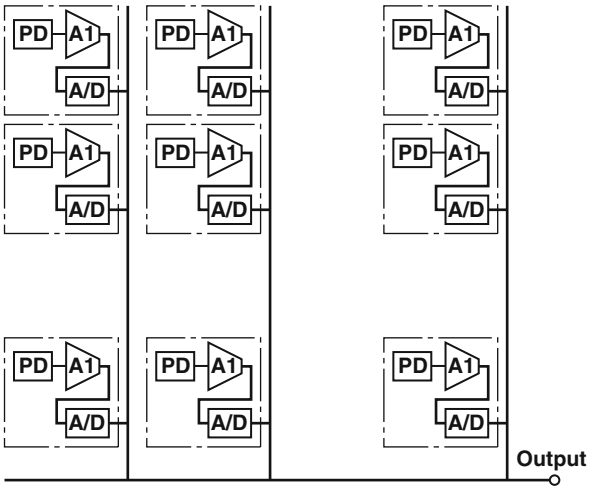

Fig. 9.1 Signal readout architecture Type I (pixel parallel ADC)

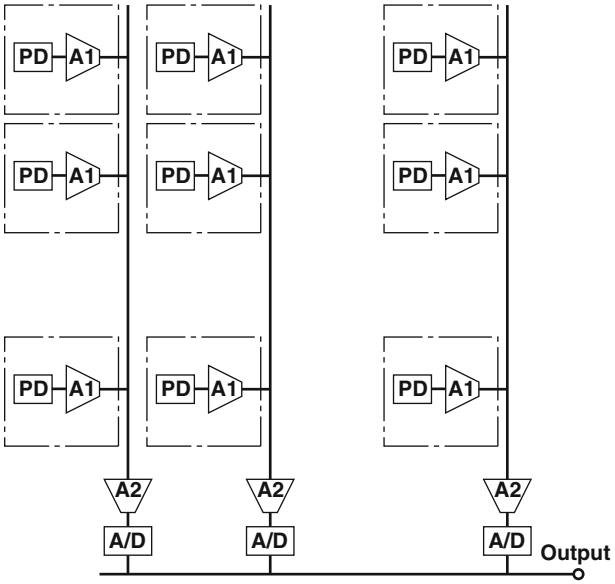

Fig. 9.2 Signal readout architecture Type II (column parallel ADC)

9.2 Signal Readout Architectures

Signal readout architectures of active pixel CMOS imagers are classified into three types as shown in Figs. 9.1(Type I), 9.2 (Type II), and 9.3 (Type III) where analog-to-digital converters (ADCs) are located at pixels, columns, and imager outputs, respectively.

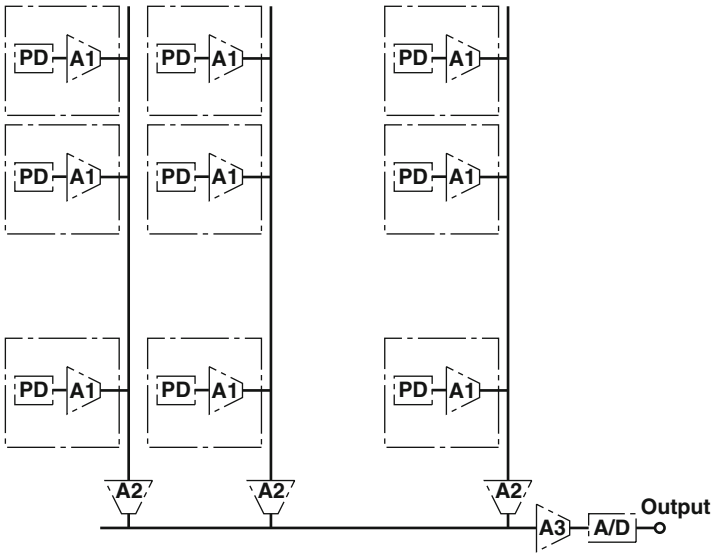

Fig. 9.3 Signal readout architecture Type III (serial ADC)

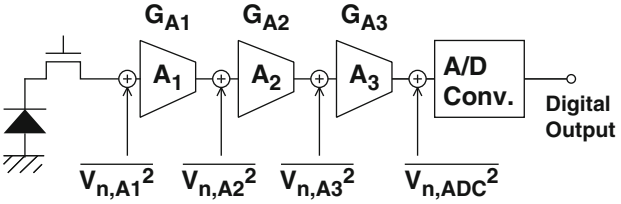

Fig. 9.4 Signal chain for the Type III

A diagram to explain the total input referred noise of the Type-III is shown in Fig. 9.4. From this diagram, the input referred noise of this mostly analog implementation is given by

$$\overline{v_{n,\text{in}}^2} = \overline{v_{n,A1}^2} + \frac{1}{G_{A1}^2}\overline{v_{n,A2}^2} + \frac{1}{G_{A1}^2 G_{A2}^2}\overline{v_{n,A3}^2} + \frac{1}{G_{A1}^2 G_{A2}^2 G_{A3}^2}\overline{v_{n,\text{ADC}}^2} \tag{9.1}$$

where $\overline{v_{n,A1}^2}$, $\overline{v_{n,A2}^2}$, and $\overline{v_{n,A3}^2}$ are input-referred noise power (mean square voltage) of an in-pixel amplifier A1, column amplifier A2 and output amplifier A3, respectively, and G_{A1}, G_{A2}, and G_{A3} are gains of the A1, A2, and A3, respectively.

In CMOS image sensors, a source follower is used for the in-pixel amplifier whose gain G_{A1} is ideally unity. Usually, the output amplifier is a unity gain buffer. Assuming $G_{A1} = G_{A3} = 1$, (9.1) is simplified to

$$\overline{v_{n,\text{in}}^2} = \overline{v_{n,A1}^2} + \overline{v_{n,A2}^2} + \frac{1}{G_{A2}^2}\overline{v_{n,A3}^2} + \frac{1}{G_{A2}^2}\overline{v_{n,\text{ADC}}^2} \tag{9.2}$$

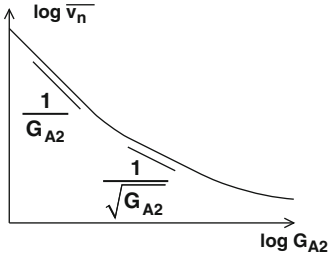

Fig. 9.5 Input-referred noise of the Type III as a function of the column amplifier gain

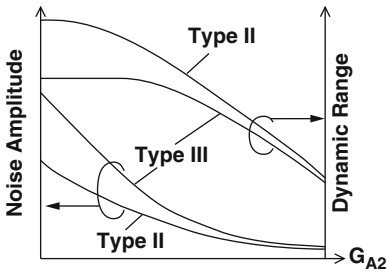

Fig. 9.6 Dynamic range and input-referred noise as a function of the column amplifier gain for Types II and III

Figure 9.5 shows a typical curve for the input-referred noise of the Type III as a function of the column amplifier gain. Wideband output amplifiers and A/D converters have usually large noise level because of their wideband thermal noise. In low-gain region, the gain has a significant noise reduction effect to the wideband output amplifier and A/D converter by a factor of G_{A2}. With a relatively large gain, the noises added at the output of the column amplifier is sufficiently suppressed by the gain, and the resulting dominant noise components are due to the thermal noises of the in-pixel amplifier and column amplifier. As to be explained in (9.4), $\overline{v_{n,A1}^2}$ and $\overline{v_{n,A2}^2}$ are not independent of G_{A2} and have a component which is inversely proportional to G_{A2} because of a bandwidth limiting effect. In this region, the input-referred noise is reduced by $1/\sqrt{G_{A2}}$. For a very large gain of the column amplifier, however, the noise reduction effect of G_{A2} is limited by a noise component such as a 1/f noise of the in-pixel source follower which does not much depend on the gain.

An advantage of the readout architecture using a column-parallel ADC is to eliminate the wideband output amplifier and wideband A/D converter. If the column ADC can be designed to be sufficiently low noise, a relatively low input-referred noise can be realized with an only low column amplifier gain. This leads to a better dynamic range in the Type II at low-gain region as shown in Fig. 9.6 when compared with that of the Type III.

In the Type I where an ADC is located in each pixel, the number of noise sources can be minimized. In this sense, the Type I is attractive for implementing very low-noise imagers. Because of the limitation of pixel area, however, only a very simple amplifier and ADC can be implemented. An efficient in-pixel ADC architecture has been proposed using multiple channel bit serial (MCBS) circuits [1]. By sharing an ADC for four pixels, the number of transistors per pixel in the MCBS is only

six transistors. A difficulty of very low noise design of the pixel level ADC is the bandwidth limiting of the in-pixel amplifier and A/D converter. To implement these circuits in a practical pixel size, only a very small capacitance is available. The development of a new circuit architecture for reducing both thermal noise spectrum and noise bandwidth without large capacitance is necessary.

9.3 Correlated Samplings and their Noise Responses

9.3.1 Correlated Double Sampling and Correlated Multiple Sampling

As explained in the previous section, a large gain in the column amplifiers has a significant effect for reducing the readout noise of CMOS imagers. On the other hand, column readout circuits need to have a function of cancelling fixed pattern and kTC or reset noises generated in the active pixels. To accommodate the noise cancelling to the amplification, two types of sampling schemes can be used. One is CDS with pre-amplification and the other is CMS.

The timing diagrams and analog circuit implementations are shown in Figs. 9.7 and 9.8, respectively.

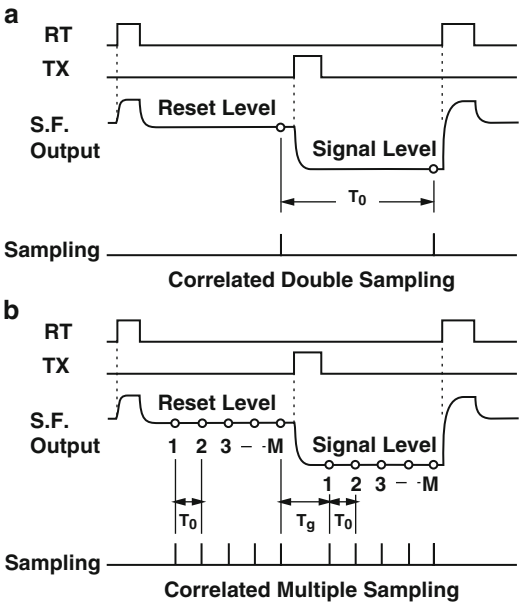

Fig. 9.7 Correlated Double Sampling (CDS) and Correlated Multiple Sampling (CMS)

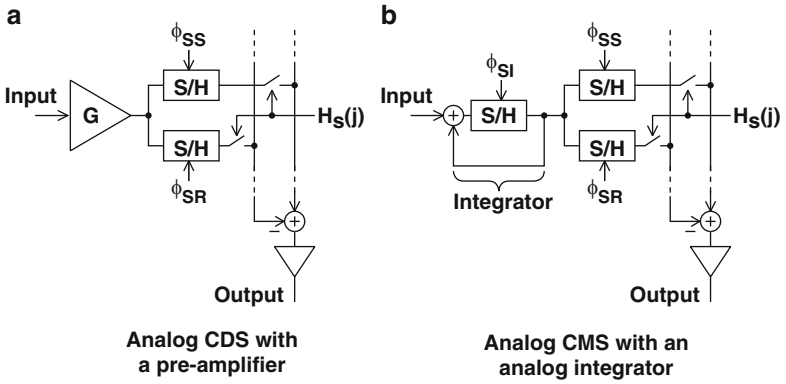

Fig. 9.8 Analog CDS and CMS circuits

In the CDS with pre-amplification, the output V_{out} as a function of time t is expressed as:

$$V_{\text{out}}(t) = G\left(V_{\text{in}}(t) - V_{\text{in}}(t - T_0)\right), \tag{9.3}$$

where n is an integer, T_0 time difference of the two samplings, and G is a gain of the pre-amplifier.

In the CMS, reset and signal levels of the pixel output are sampled M times with an interval of T_0 and the sampled signals are added up in an integrator. In Fig. 9.8b, the integrator is expressed as a sample-and-hold (S/H) circuit with a feedback to the input. In Fig. 9.7, T_{g} is a gap between the two groups of samplings for the reset and photo signal. If $T_{\text{g}} = N_g T_0$, where N_g is an integer, the CMS process can be written as:

$$V_{\text{out}}(nT_0) = \sum_{k=0}^{M-1} \left\{V_{\text{in}}((n-k)T_0) - V_{\text{in}}((n-k-M-N_g+1)T_0)\right\}. \tag{9.4}$$

The CDS and CMS can be entirely or in part performed in digital domain. Figure 9.9 depicts a digital CDS (Fig. 9.9a), a digital CMS with an analog integration and digital differencing (Fig. 9.9b), and a fully digital CMS (Fig. 9.9c).

The fundamental property of the digital-domain processing to the temporal noise reduction effect is not much different from the analog-domain counterpart. The significant effect of the digital domain processing is found in cancelling column-wise fixed-pattern noise (FPN). The column-wise FPN is also called vertical FPN (VFPN) because the VFPN appears as vertical lines on the reproduced image. The sources of the VFPN are mainly due to analog circuits such as offsets of amplifiers and charge injections of switches. The use of differencing in digital domain almost perfectly cancels the VFPN because the analog signal paths for the reset and signal levels are the same. CMOS imagers employing the digital CDS have very small VFPNs below $0.5e^{-}$[2–4].

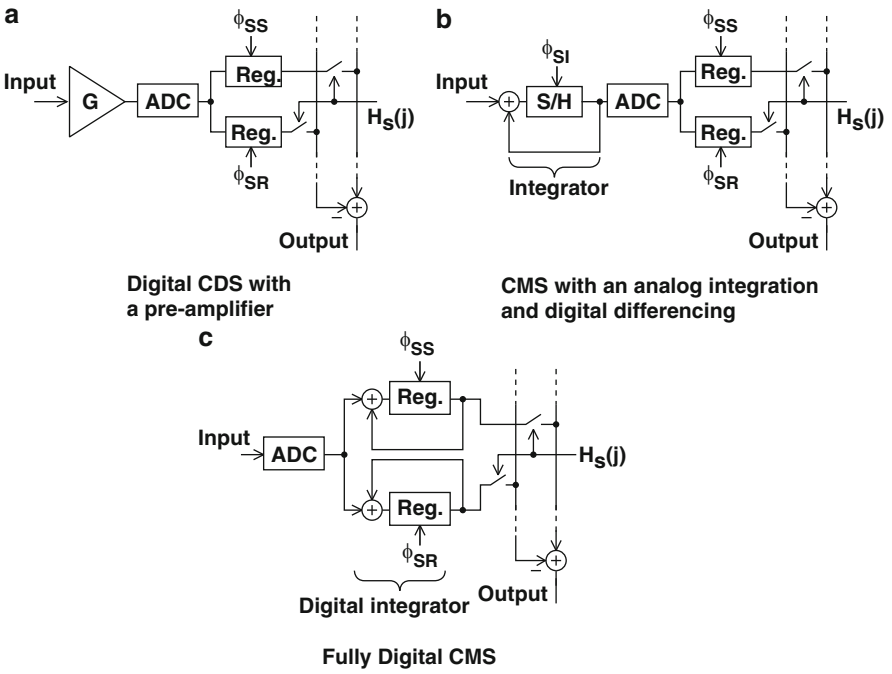

Fig. 9.9 Digital CDS and CMS circuits

9.3.2 Response of CDS and CMS to Thermal and 1/f Noises

9.3.2.1 Modeling of CDS

A model of noise canceling circuits is shown in Fig. 9.10. The current flowing in RC circuits is given by:

$$i = \frac{1}{R}(v_s(t) - v_{\mathrm{LP}}(t)) = C\frac{\mathrm{d}v_{\mathrm{LP}}(t)}{\mathrm{d}t} \tag{9.5}$$

The function of the CDS is expressed as:

$$v_{\mathrm{CDS}}(t) = v_{\mathrm{LP}}(t) - v_{\mathrm{LP}}(t - T_0). \tag{9.6}$$

Using Laplace transform, the transfer function of the CDS is given by [5]

$$V_{\mathrm{CDS}}(s) = V_{\mathrm{LP}}(s) - V_{\mathrm{LP}}(s)\mathrm{e}^{-sT_0} = \frac{1}{1 + sCR}(1 - \mathrm{e}^{-sT_0})V_s(s). \tag{9.7}$$

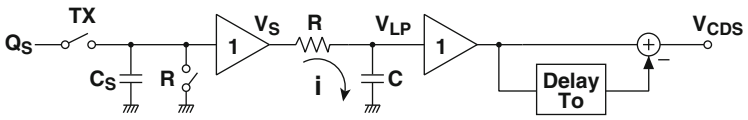

Fig. 9.10 Model of CDS circuits

9.3.2.2 Response of CDS to Thermal and 1/f Noises

The noise spectrum including the thermal and 1/f noises is expressed as:

$$S_n(f) = S_{n,\mathrm{th}} + \frac{N_f}{f}, \tag{9.8}$$

where $S_{n,\mathrm{th}}$ is the thermal noise spectrum and N_f is a 1/f noise coefficient. Using a transfer function derived from (9.7), the mean square noise power is calculated with

$$\overline{v_{n,\mathrm{CDS}}^2} = \int_{-\infty}^{\infty} \left(S_{n,\mathrm{th}} + \frac{N_f}{f}\right) \frac{1}{1+(\omega/\omega_\mathrm{c})^2} |1 - \mathrm{e}^{-j\omega T_0}|^2 \mathrm{d}f \tag{9.9}$$

$$= \int_{-\infty}^{\infty} \left(S_{n,\mathrm{th}} + \frac{N_f}{f}\right) \frac{1}{1+(\omega/\omega_\mathrm{c})^2} 4\sin^2(\omega T_0/2) \mathrm{d}f, \tag{9.10}$$

where $\omega_\mathrm{c} = 1/CR$ is a cutoff angular frequency of the CDS circuits. With $x = \omega T_0/2$, $x_c = \omega_\mathrm{c} T_0/2$, the above equation is expressed as:

$$\overline{v_{n,\mathrm{CDS}}^2} = \frac{S_{n,\mathrm{th}}}{\pi T_0} \int_0^{\infty} \frac{4\sin^2 x}{1+(x/x_c)^2} \mathrm{d}x + N_f \int_0^{\infty} \frac{1}{x} \frac{4\sin^2 x}{1+(x/x_c)^2} \mathrm{d}x. \tag{9.11}$$

The first and second terms of the left-hand side of the above equation correspond to the thermal ($\overline{v_{n,\mathrm{th,CDS}}^2}$) and 1/f ($\overline{v_{n,f,\mathrm{CDS}}^2}$) noises, respectively. The integral of the first term is calculated as [6]:

$$\int_0^{\infty} \frac{4\sin^2 x}{1+(x/x_c)^2} \mathrm{d}x = \frac{\pi x_c}{4} \left(1 - \exp(-2x_c)\right). \tag{9.12}$$

If a sufficient settling of the CDS circuit is assumed, e.g., $\omega_\mathrm{c} T_0 > 5$, the mean square thermal noise after the CDS operation is given by:

$$\overline{v_{n,\mathrm{th,CDS}}^2} = \frac{S_{n,\mathrm{th}}\omega_\mathrm{c}}{2} \left(1 - \exp(-2x_c)\right) \simeq \frac{S_{n,\mathrm{th}}\omega_\mathrm{c}}{2}. \tag{9.13}$$

As the mean square thermal noise of a RC circuit is expressed as $S_{n,\mathrm{th}}\omega_\mathrm{c}/4$, the CDS operation has an effect of doubling the thermal noise at the input.

This integral of the second term of the left-hand side of (9.11) is expressed as [6–8]:

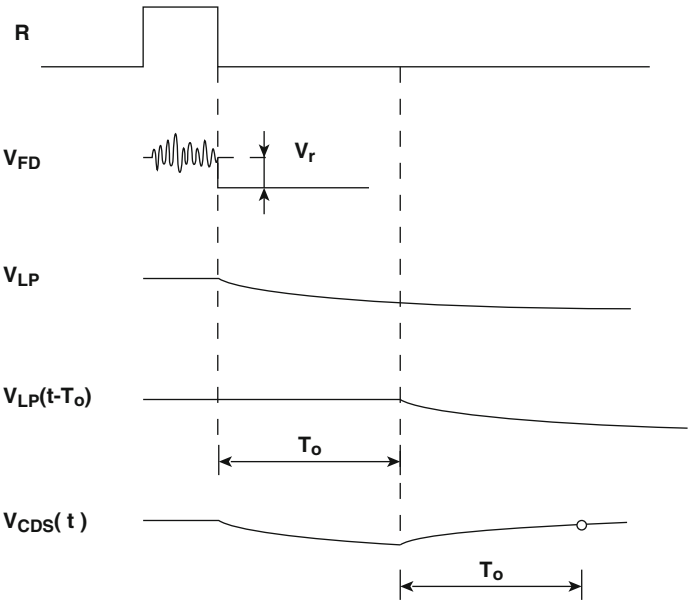

Fig. 9.11 Waveform of the CDS

$$\int_0^\infty \frac{1}{x}\frac{4\sin^2 x}{1+(x/x_c)^2}\mathrm{d}x = 2\left\{\gamma + \ln 2x_c + \int_0^\infty \frac{x\cos x}{x^2+(2x_c)^2}\mathrm{d}x\right\}, \tag{9.14}$$

where γ is Euler constant and $\gamma = 0.577215....$ Assuming $\omega_c T_0$ is > 5 for sufficient settling, the third term of (9.14) can be ignored, and the following useful formula for the response of the CDS to 1/f noise is obtained.

$$\overline{v_{n,f,\mathrm{CDS}}^2} \simeq 2N_f(\gamma + \ln \omega_c T_0). \tag{9.15}$$

Figure 9.12 shows the comparison between (9.14) and (9.15). For $\omega_c T_0 > 1$, (9.15) is a good approximation of (9.14).

9.3.2.3 Response of CMS Circuit to Thermal and 1/f Noises

From (9.4), the transfer function of the CMS can be calculated with the Z-transform as:

$$H_{\mathrm{CMS}}(z) = \frac{(1-z^{-M})(1-z^{-M-N_g+1})}{1-z^{-1}}, \tag{9.16}$$

where $z = \exp(j\omega T_0)$. The noise power (mean square noise voltage) after the CMS process can be calculated with

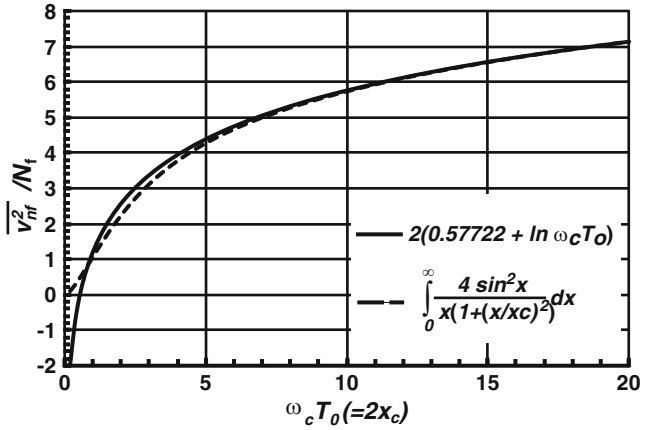

Fig. 9.12 Response of CDS circuit to 1/f noise (Comparison of (9.14) and (9.15))

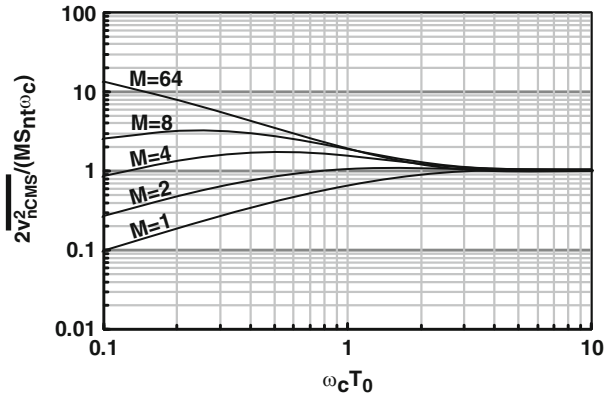

Fig. 9.13 Response of CMS to thermal noise

$$\overline{v_{n,\mathrm{CMS}}^2} = \int_0^\infty S_n(f) \frac{1}{1+(\omega/\omega_c)^2} \left| H_{\mathrm{CMS}}(\mathrm{e}^{j\omega T_0}) \right|^2 \mathrm{d}f. \tag{9.17}$$

If $N_g = 1$, it is expressed as:

$$\overline{v_{n,\mathrm{CMS}}^2} = \int_{-\infty}^\infty S_n(f) \frac{1}{1+(\omega/\omega_c)^2} \frac{4\sin^4(\omega M T_0/2)}{\sin^2(\omega T_0/2)} \mathrm{d}f. \tag{9.18}$$

Figure 9.13 shows numerically calculated results of the response of the CMS to thermal noise using (9.18) and $S_n(f) = S_{n,\mathrm{th}}$. The noise power after the CMS process is normalized by $M S_{n,\mathrm{th}} \omega_c / 2$.

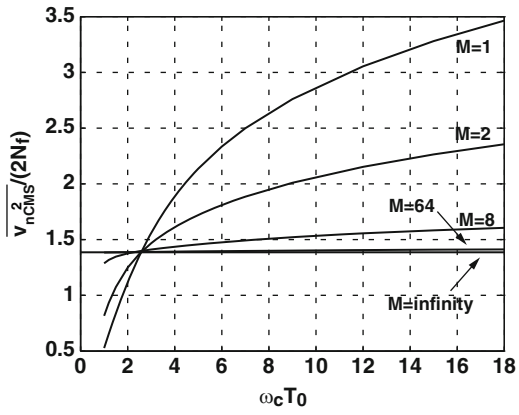

Fig. 9.14 Response of CDS and CMS to 1/f noise

For $\omega_c T_0 >> 1$, the thermal noise power with the CMS is approximated as

$$\overline{v_{n,\mathrm{th,CMS}}^2} \simeq \frac{M S_{\mathrm{nt}} \omega_c}{2} \tag{9.19}$$

The input referred thermal noise power of the CMS, given by $\overline{v_{n,\mathrm{th,CMS}}^2}/M^2$, is inversely proportional to M.

The response of the CMS to the 1/f noise is numerically calculated with (9.18) and $S_n(f) = N_f/f$. Figure 9.14 shows calculation results of the response of the CMS to 1/f noise [9]. The noise power after the CMS process is normalized with $2N_f$, where N_f is the spectral density of the 1/f noise at unit frequency. N_g of 1 is assumed. In Fig. 9.14, M of 1 corresponds to the CDS. ω_c is the inverse of the time constant of the sampling circuits in the CMS. Because of the slewing behavior of the pixel source follower for a large signal and sufficient settling for precise CDS operation, $\omega_c T_0$ must be chosen to be large enough, e.g., $\omega_c T_0 > 5$.

For a large M the normalized noise power converges to $\ln 4 \simeq 1.39$. The CMS with the large M has a higher noise reduction effect to 1/f noise than that of the CDS, e.g., the noise power of the CMS is half of the CDS if $\omega_c T_0 = 10$.

9.4 Noise in Active-pixel CMOS Image Sensors Using Column CMS Circuits

The column CDS and CMS circuits can be implemented using a switched-capacitor (SC) amplifier and SC integrator, respectively. A typical column CDS amplifier is shown in Fig. 8.15 in Chap. 8. Detailed noise analysis of the readout circuits including in-pixel source follower and high-gain column amplifiers are given in Chap. 8. In this section, the noise reduction effect of the CMS circuits using the SC integrator is analyzed.

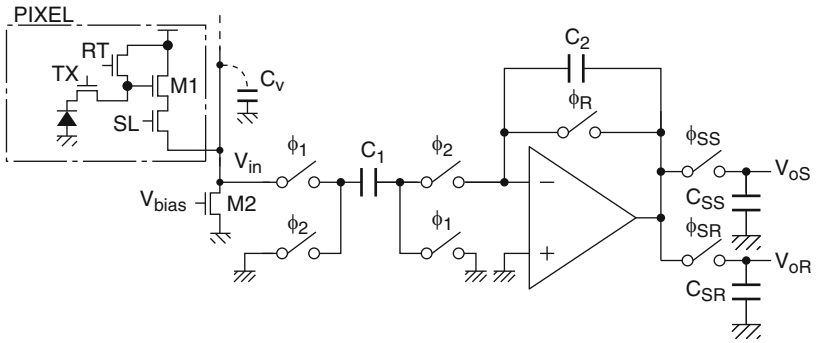

Fig. 9.15 Column CMS circuit with an SC integrator

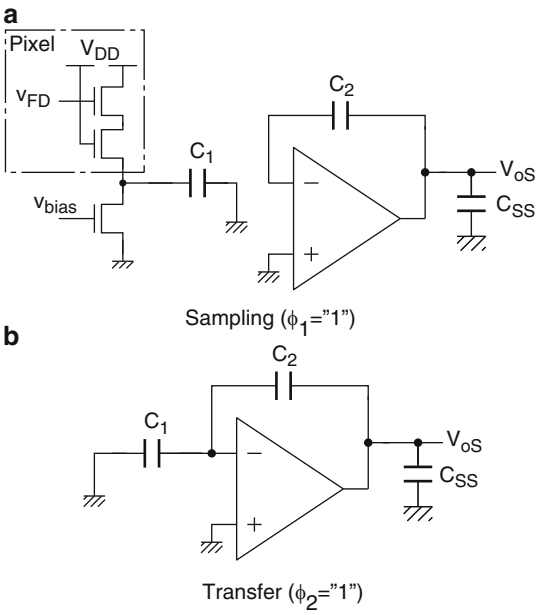

Fig. 9.16 Operation of the CMS

The column CMS circuit using an SC integrator is shown in Fig. 9.15 together with an 4Tr. active pixel circuit. The CMS circuit has two phases as shown in Fig. 9.16. In the sampling phase ($\phi_1 =$ "1"), the pixel output is sampled at the capacitor C_1, and the sampled charge is transferred to the feedback capacitor C_2 in the charge transfer phase. This operation is repeated for M times if a gain of M is needed. The outputs of the SC integrator for the reset and signal levels of the pixel output are sampled and held at capacitors C_{SS} and C_{SR}, respectively.

Figure 9.17 shows the equivalent circuit of the in-pixel source follower with a capacitor connected at the output which includes C_1 of the CMS circuits and other parasitic capacitances. The gate input of the source follower, which is connected

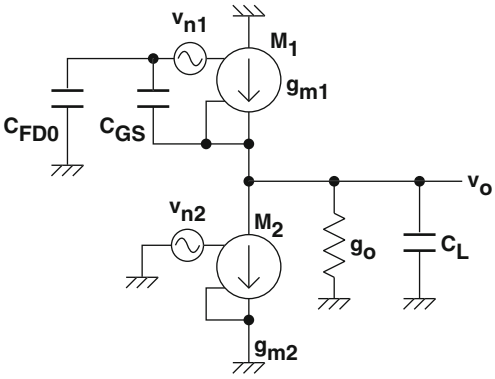

Fig. 9.17 Equivalent circuit of the source follower with the CMS in sampling phase

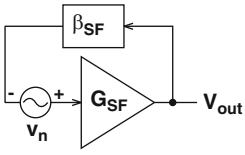

Fig. 9.18 Noise gain factor of a source follower

to a floating diffusion, has a very small capacitance C_{FD0}. Because of this, gate-to-source capacitance C_{GS} positively feeds the output signal back to the input by a factor given by:

$$\beta_{\mathrm{SF}} = \frac{C_{\mathrm{GS}}}{C_{\mathrm{FD0}} + C_{\mathrm{GS}}} \tag{9.20}$$

An equivalent circuit of the source follower with the positive feedback is shown in Fig. 9.18. This shows that internal noise of the source follower is amplified by the gain of

$$G_{n,\mathrm{SF}} = \frac{G_{\mathrm{SF}}}{1 - G_{\mathrm{SF}}\beta_{\mathrm{SF}}} \tag{9.21}$$

where G_{SF} is a source follower gain given by:

$$G_{\mathrm{SF}} = \frac{g_{m1}}{g_{m1} + g_o}. \tag{9.22}$$

The output conductance g_o contains the drain conductance of M1 and M2 and a mutual conductance of M1 due to a body-bias effect (see Sect. 8.4 for more details).

Another effect that increase the thermal noise of the in-pixel source follower is an excess noise factor. Thermal noise spectrum of an MOS transistor with a submicrometer channel length is given by:

$$S_{n,\mathrm{th,MOS}} \cong 4k_{\mathrm{B}}T\frac{\xi}{g_m}, \tag{9.23}$$

where ξ is the excess noise factor [10], and g_m the transconductance of the MOS transistor. The excess noise factor is a function of the channel length (L), and it takes $\sim 2/3$ for $L > 2\,\mu m$, and increases as the channel length decreases.

Thermal noise spectrum of the source follower $S_{n,\mathrm{th,SF}}$ is given by:

$$S_{n,\mathrm{th,SF}} = G_{n\mathrm{SF}}^2 4k_\mathrm{B}T\frac{\xi_\mathrm{SF}}{g_{m1}}, \tag{9.24}$$

where ξ_SF is an excess noise factor of the source follower given by:

$$\xi_\mathrm{SF} = \xi_1 + \frac{g_{m2}}{g_{m1}}\xi_2, \tag{9.25}$$

where ξ_1 and ξ_2 are excess noise factors and g_{m1} and g_{m2} are transconductances of M1 and M2, respectively.

The 1/f noise spectrum of the source follower is expressed as:

$$S_{n,f,\mathrm{SF}}(f) = G_{n,\mathrm{SF}}^2\left(1 + \frac{N_{f2}g_{m2}^2}{N_{f1}g_{m1}^2}\right)\frac{N_{f1}}{f} \simeq G_{n,\mathrm{SF}}^2\frac{N_{f1}}{f}, \tag{9.26}$$

where N_{f1} and N_{f2} are the spectral density at unit frequency of M1 and M2, respectively, and N_{f2} is much smaller than N_{f1} because the transistor M_2 is common for each column and the size can be chosen much larger than that of M_1.

In the sampling phase of the CMS, the thermal noise power of the pixel source follower sampled in C_1 is given by:

$$\overline{v_{n,\mathrm{th,SF}}^2} = \frac{S_{n,\mathrm{th,SF}}\omega_{\mathrm{c,SF}}}{4} \tag{9.27}$$

where $\omega_{\mathrm{c,SF}}$ is the angular cutoff frequency of the source follower and it is given by:

$$\omega_{\mathrm{c,SF}} = \frac{g_{m1}}{(C_\mathrm{v} + C_1)G_{n,\mathrm{SF}}}, \tag{9.28}$$

where C_v is the parasitic capacitance of the vertical signal line. Using (9.24) and (9.28), the thermal noise power of the pixel source follower can also be expressed as:

$$\overline{v_{n,\mathrm{th,SF}}^2} = \frac{k_\mathrm{B}T\xi_\mathrm{SF}G_{n,\mathrm{SF}}}{C_\mathrm{v} + C_1}. \tag{9.29}$$

In the transfer phase, the noise charge $C_1\overline{v_{n,\mathrm{th,SF}}}$ is transferred to C_2, and noise voltage $(C_1/C_2)\overline{v_{n,\mathrm{th,SF}}}$ appears at the integrator output. The process of sampling and transfer is repeated M times for integration. To perform the CMS, the integrator outputs for the reset and signal levels are sampled and held at C_SR and C_SS and the difference of the two outputs are taken. This doubles the above noise power and the total output-referred thermal noise power is given by $2M\overline{v_{n,\mathrm{th,SF}}^2} = MS_{n,\mathrm{th,SF}}\omega_{\mathrm{c,SF}}/2$. This is identical to (9.19) for calculating the thermal noise after the CMS process using transfer functions if $\omega_\mathrm{c}T_0 > 5$.

In transfer phase, the noise power due to an amplifier ($\overline{v_{n,\mathrm{th,A}}^2}$) is added. This can be calculated if the amplifier circuits and values of capacitances (C_1, C_2, C_SR, and

C_{SS}) are known [11]. The total noise power referred to the integrator input or the source follower output is given by:

$$\overline{v_{n,\text{th,in}}^2} = 2\frac{\overline{v_{n,\text{th,SF}}^2} + \overline{v_{n,\text{th,A}}^2}}{M}. \tag{9.30}$$

For 1/f noise, the transistor $M1$ of the in-pixel source follower is the dominant noise source because of the very small size. The input referred noise power if $M >> 1$ is given by:

$$\overline{v_{n,f,\text{in}}^2} = 2\ln 4 G_{n,\text{SF}}^2 N_{f1} \tag{9.31}$$

The total noise includes the thermal and 1/f noises of the in-pixel source follower and column CMS circuits and the other noises such as an ADC noise which is added at the output of the CMS circuits. The total noise as a function of the number of samplings M, which has a same effect of the gain in the column amplifier(G_{A2}), behaves as shown in Fig. 9.5. For a very large value of M, the total input referred noise is dominated by 1/f noise of the source follower because the other noise can be sufficiently small by the large M.

The input-referred noise as the number of noise electrons is expressed as the r.m.s. noise at the source follower output divided by the charge-to-voltage conversion gain of the pixel source follower G_c given by:

$$G_c = \frac{qG_{\text{SF}}}{C_{\text{FD0}} + C_{\text{GS}}(1 - G_{\text{SF}})} \tag{9.32}$$

Example 9.1. For $C_v = 1\text{pF}$, $C_1 = 1\text{pF}$, $C_2 = 1\text{pF}$, $G_{\text{SF}} = 0.85$, $C_{\text{FD0}} = 2\text{fF}$, $C_{\text{GS}} = 2\text{fF}$, $\xi_1 = 1$, $\xi_2 = 0.7$, and $g_{m2}/g_{m1} = 0.5$,

$$\overline{v_{n,\text{th,SF}}} = 64.3\,\mu V_{\text{rms}}$$

and for $M = 32$,

$$\overline{v_{n,\text{th,in}}} = 16.1\,\mu V_{\text{rms}}$$

if the noise added at the transfer phase is ignored. As the conversion gain G_c is calculated to be $59.2\,\mu\text{V/e}^-$, the input-referred thermal noise using the CMS with M of 32 is 0.27 e^-.

9.5 Possibility of Single Photon Detection

9.5.1 Single Photon Detection Using Quantization

In active pixel CMOS image sensors, photo electrons generated by the absorption of photons are detected by a floating diffusion amplifier using a small capacitance in the pixel, and the resulting voltage signal is read out through a source follower

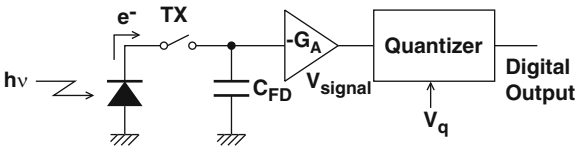

Fig. 9.19 Electron counting

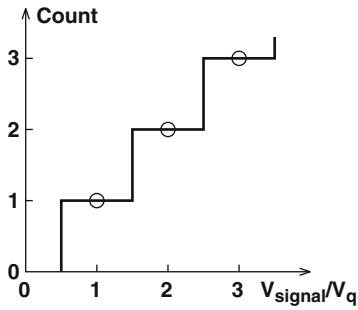

Fig. 9.20 Quantization

and the following readout circuitry. If there is no influence of noise, the signal level read through the circuitry must be discrete, and if a circuitry like a analog-to-digital converter exactly measure the voltage as a digital number, noiseless detection of photo signal electron is possible [12, 13].

The principle of electron counting is shown in Fig. 9.19. An electron is transferred to the floating diffusion (FD), a voltage step q/C_{FD} will appear, where q is elementary charge, and C is the FD capacitance. Since the electron has negative charge, the voltage step is negative. If the voltage is amplified by a gain of $-G_{\mathrm{A}}$ and n electrons are transferred to the FD, the signal voltage that appears at the amplifier output is given by:

$$V_{\mathrm{signal}} = nG_{\mathrm{A}}\frac{q}{C_{\mathrm{FD}}}. \tag{9.33}$$

The amplifier output is quantized by a precise A-to-D converter, or quantizer. The quantization step V_{q} is chosen exactly as qG_A/C_{FD}. The threshold level of the quantizer is chosen as $(n-0.5)V_{\mathrm{q}}$, (n = 0, 1, 2, ...) as shown in Fig. 9.20. Note that, the horizontal axis of Fig. 9.20 is normalized by V_{q} and it shows the number of signal electrons.

If there is no influence of noise, the quantizer outputs an exact digital code that equals to the number of signal electrons. The existing noise during the signal detection causes the miscounting of the number of electrons to be measured. If the noise is a white noise like a thermal noise, the influence of the noise in electron counting can be easily analyzed.

The probability density function of the thermal noise amplitude $p_{n,\mathrm{th}}(x)$ normalized by V_{q} follows a Gaussian distribution, and is given by:

$$p_{n,\mathrm{th}}(x) = \frac{1}{\sqrt{2\pi}(\sigma_n/V_{\mathrm{q}})}\exp\left(-\frac{x^2}{2(\sigma_n/V_{\mathrm{q}})^2}\right), \tag{9.34}$$

where σ_n is the r.m.s. amplitude of the thermal noise.

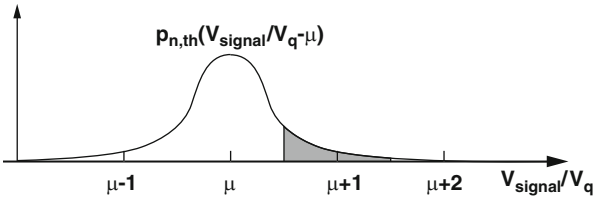

Fig. 9.21 Influence of noise in electron counting

If the number of signal electrons is μ, the noise power caused by the miscounting to $\mu + 1$ is V_q^2. From Fig. 9.21, the possibility of miscounting is obviously given by:

$$\int_{0.5}^{1.5} p_{n,\mathrm{th}}(x)\mathrm{d}x. \tag{9.35}$$

A miscounting to $\mu - 1$ also has a noise power increase of V_q^2. The miscounting to $\mu + 2$ leads to a noise power increase of $(2V_q)^2$. For calculating the total noise power due to the miscounting, note that the number of signal electrons cannot be negative.

The total noise power after quantization can be calculated as:

$$P_n(\mu) = 2\sum_{j=1}^{\mu-1}\int_{j-0.5}^{j+0.5} p_{n,\mathrm{th}}(x)(jV_q)^2\mathrm{d}x + \sum_{j=\mu}^{\infty}\int_{j-0.5}^{j+0.5} p_{n,\mathrm{th}}(x)(j^2+\mu^2)V_q^2\mathrm{d}x. \tag{9.36}$$

The thermal noise power before quantization P_{no} is given by:

$$P_{no} = \int_{-\infty}^{\infty} \sigma_n^2 p_{n,\mathrm{th}}(x)\mathrm{d}x = \sigma_n^2 \tag{9.37}$$

Figure 9.22 shows the calculation results using (9.36) for μ of 1,3,5, and 7. The noise power after quantization is normalized by the noise power before quantization P_{n0}. The horizontal axis is the number of noise electrons N_{ne}, that is the input-referred noise amplitude before quantization is divided by the conversion gain of an electron charge to voltage. If μ is large enough and the number of noise electrons is much larger than unity, the quantization does not contribute to the noise reduction. For small μ, the quantization reduces the noise power by half for $N_{\mathrm{ne}} >> 1$ because of the fact that the number of signal electrons cannot be negative. For $N_{\mathrm{ne}} \simeq 0.4$, the quantization rather increases the noise power. In this region, the miscounting caused by the noise can be larger than the average noise of $\simeq$0.4. For $N_{\mathrm{ne}} < 0.4$, the noise power with quantization abruptly decreases as N_{ne} decreases. For $N_{\mathrm{ne}} < 0.1$, the noise power after quantization can be considered to be zero. In this region, noiseless detection of signal electrons or true photo electron counting is possible.

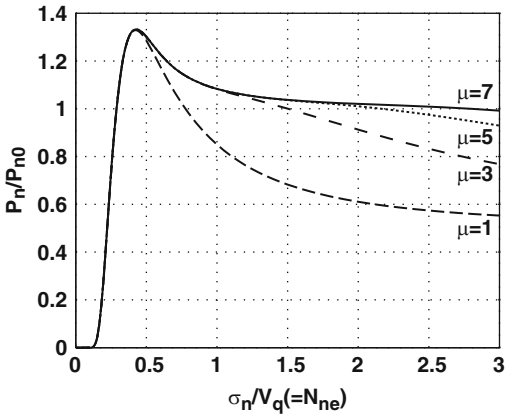

Fig. 9.22 Noise power with quantization (P_n / P_{n0})

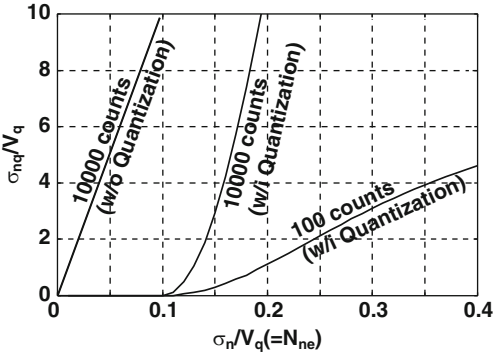

Fig. 9.23 Effect of quantization in digital photo electron counting

For $N_{\text{ne}} < 0.1$, the possibility of miscounting becomes very small because of the small noise amplitude.

This photo electron counting technique is useful for a noise-free digital signal accumulation. Figure 9.23 shows the equivalent number of noise electrons with quantization ($\sqrt{P_n} / V_{\text{q}} = \sigma_{nq} / V_{\text{q}}$) and digital domain accumulation of 100 and 10,000 times. The accumulated noise in digital domain accumulation for 10,000 times without quantization is also shown in Fig. 9.23. Without quantization, the noise increases to ten electrons for the digital accumulation of 10,000 times even if the readout noise is 0.1 electron, while the noise remains zero for the digital accumulation of 10,000 times if the quantization technique is applied for the 0.1-electron readout noise.

9.5.2 Condition for Single Photon Detection

For noiseless electron counting using quantization, the readout noise must be $<0.1\,\text{e}^-$. This section discusses what condition is necessary to achieve 0.1e^- noise level using the column CMS.

If the thermal and 1/f noises equally contribute to the total noise and the two noise components are independent to each other, both the thermal noise and the 1/f noise must be smaller than $0.07e^-$. As shown in **Ex. 9.1**, the thermal noise can be sufficiently small by using a large number of samplings in the CMS. To reduce the noise to be less than $0.07e^-$, $C_{FD0} = C_{GS}$ of 0.5fF is required assuming the other parameters are same as in **Ex. 9.1**. This small capacitance with the source follower gain of 0.85 corresponds to the conversion gain of $237\,\mu V/e^-$, which is not surprisingly large. High conversion gain pixels in CMOS imagers such as $200\,\mu V/e^-$ [14] and $190\,\mu V/e^-$ [15] have been reported.

From (9.31) and (9.32) the number of noise electrons generated by the 1/f noise can be expressed as:

$$N_{n,f} = \frac{\overline{v_{n,f,\mathrm{in}}}}{G_c} = \frac{2G_{n,\mathrm{SF}}\sqrt{\ln 2N_f}}{G_c} \tag{9.38}$$

The 1/f noise in MOS transistors originates from the channel conductivity fluctuations caused by a channel carrier number fluctuation or channel carrier mobility fluctuation due to traps in the gate oxide. In n-channel MOS transistors often used in active pixels in CMOS imagers, the carrier number fluctuation is believed to be the dominant mechanism. The spectral density at unit frequency due to the channel carrier number fluctuation is expressed as [16]:

$$N_f = \frac{q^2 N_{ot}}{C_{ox}^2 A_G}, \tag{9.39}$$

where $A_G = LW$ is the channel area of the MOS transistor and $N_{ot}[\mathrm{cm}^{-2}]$ is the effective trap density. More general and physically exact models exist to explain the mechanism of the 1/f noise, which is beyond the scope of this chapter [17].

In order to minimize the number of noise electrons, the size of the transistor must be optimized. A large A_G reduces the 1/f noise spectral density, while increasing the noise gain factor. The minimum can be found for the condition C_{GS} equals to C_{FD0}. This is obvious because $G_{n,\mathrm{SF}}/G_c$ in (9.38) is expressed as $(C_{FD0} + C_{GS})/q$ from (9.21), (9.20), and (9.32) and N_f is inversely proportional to C_{GS} if the gate to source capacitance of the MOS transistor operating in saturation region is expressed as $(2/3)C_{ox}A_G$. Then the $N_{n,f}$ is expressed as:

$$N_{n,f} \simeq \frac{8}{3}\sqrt{\ln 2N_{ot}A_G} \simeq 2.2\sqrt{N_{ot}A_G}. \tag{9.40}$$

Example 9.2. For $N_{ot} = 10^8\mathrm{cm}^{-2}$ and $A_G = 0.5\,\mu\mathrm{m}^2$,

$$N_{n,f} = 1.6e^-$$

Equation (9.40) suggests us an interesting strategy to a low-noise design of CMOS imagers. The use of a small size transistor leads to a low input- referred

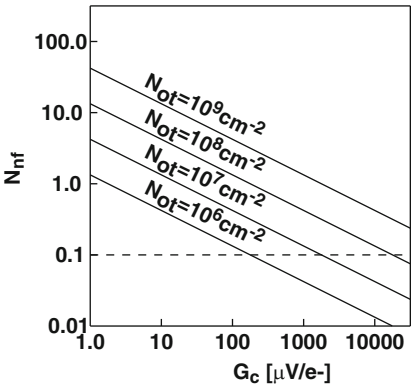

Fig. 9.24 Input-referred 1/f noise as a function of conversion gain

1/f noise. The 1/f noise is believed to be reduced by increasing the transistor size as (9.39) predicts. The input-referred 1/f noise of the pixel source follower as the number of noise electrons, however, can be reduced by choosing a small transistor size and meeting the condition $C_{\mathrm{FD0}} = C_{\mathrm{GS}}$, because of the increased conversion gain. Using (9.32) in (9.40), the noise expressed as a function of the conversion gain is given by:

$$N_{n,f} \simeq 2.72\sqrt{\frac{G_{\mathrm{SF}}}{2-G_{\mathrm{SF}}}\frac{qN_{ot}}{C_{ox}G_{\mathrm{c}}}}. \tag{9.41}$$

Figure 9.24 depicts the relationship between the $N_{n,f}$ and G_{c} with N_{ot} as a parameter. $C_{ox} = 5fF/\mu\mathrm{m}^2$ and $G_{\mathrm{SF}} = 0.85$ are assumed.

In the present CMOS technology, N_{ot} takes values in the range of 10^7–$10^8\mathrm{cm}^{-2}$. If $N_{ot} = 10^7$, G_{c} of $> 3.5\,\mathrm{mV/e}^-$ is required for achieving $N_{n,f}$ of smaller than $0.07\mathrm{e}^-$.

A CMOS single-photon imaging or a noiseless photoelectron counting is theoretically possible. However, a charge detector with very high conversion gain (more than 20 times higher than the state of the art), or a process technology with very low N_{ot} (more than 20 times smaller than the state of the art) is required.

It should be noted that earlier discussion is based on the mean value of noise over the pixels of CMOS imagers, and it can be reduced by the small transistor size and the resulting high conversion gain. The use of a small-size transistor in each pixel, on the other hand, may increase the occurrence of anomalously large random telegraph signal (RTS) noises [18]. The development of a strategy for reducing the pixel-wise mean random noise while suppressing the anomalously large RTS noise is necessary for true CMOS single-photon imaging.

References

1. D.X.D. Yang, B. Fowler, A. El Gamal, A Nyquist rate pixel-level ADC for CMOS imager sensors. IEEE J. Solid-State Circuits, **34**(3), 348–356 (1999)
2. S. Yoshihara, Y. Nitta, M. Kikuchi, K. Koseki, Y. Ito, Y. Inada, S. Kuramochi, H. Wakabayashi, M. Okano, H. Kuriyama, J. Inutsuka, A. Tajima, T. Nakajima, Y. Kudoh, F. Koga, Y. Kasagi,

S. Watanabe, T. Nomoto, A 1/1.8-inch 6.4 Mpixel 60 frames/s CMOS image sensor with seamless mode change. IEEE J. Solid-State Circuits, **41**(12), 2998–3006 (2006)
3. S. Matsuo, T.J. Bales, M. Shoda, S. Osawa, K. Kawamura, A. Andersson, M. Haque, H. Honda, B. Almond, Y. Mo, J. Gleason, T. Chow, I. Takayanagi, 8.9-Megapixel video image sensor with 14-b column-parallel SA-ADC. IEEE Trans. Electr. Dev. **56**(11), 2380–2389 (2009)
4. J.H. Park, S. Aoyama, T. Watanabe, K. Isobe, S. Kawahito, A high-speed low noise CMOS image sensor with 13-b column-parallel single-ended cyclic ADCs. IEEE Trans. Electr. Dev. **56**(11), 2414–2422 (2009)
5. H.M. Wey, W. Guggenburl, An improved correlated double sampling circuits for low-noise charge coupled devices. IEEE Trans. Circuits Syst. **37**(12), 1559–1565 (1990)
6. M. Abramowitz, I.A. Stegun, *Handobook of Mathematical Functions* (Dover Publications, NY, USA, 1970)
7. R.J. Kansy, Response of a correlated double sampling circuit to 1/f noise. IEEE J. Solid-State Circuits, **SC-15**(3) (1980)
8. G.R. Hopkinson, D.H. Lumb, Noise reduction techniques for CCD image sensors. J. Phys. E: Sci. Instrum. **15**, 1215–1223 (1982)
9. N. Kawai, S. Kawahito, Effectiveness of a correlated multiple sampling differential averager for 1/f noise. IEICE Express Lett. **2**(13), 379–383 (2005)
10. J.S. Goo, C.H. Choi, A. Abramo, J.G. Ahn, Z. Yu, T.H. Lee, R.W. Dutton, Physical origin of the excess thermal noise in short channel MOSFETs. IEEE Electr. Dev. Lett. **22**(2), 101–103 (2001)
11. S. Kawahito, S. Itoh, Noise calculation model and analysis of high-gain readout circuits for CMOS image sensors. Proc. SPIE, **6816**, 68160D (2008)
12. S. Kawahito, S. Itoh, A Photon count imaging using a extremely small capacitor and a high-precision low-noise quantizer. Proc. SPIE-IS&T, **5017**, 68–75 (2003)
13. N. Kawai, S. Kawahito, A low-noise signal readout circuit using double-stage noise canceling architecture, in *Proceedings of IEEE Workshop on Charge-Coupled Devices and Advanced Image Sensors*, Karuizawa, June 2005, pp. 27–30
14. S. Adachi, W. Lee, N. Akahane, H. Oshikubo, K. Mizobuchi, S. Sugawa, A 200-$\mu V/e^-$ CMOS image sensor with 100-ke full well capacity. IEEE J. Solid-State Circuits, **43**(4), 823–830 (2008)
15. G. Agranov, R. Mauritzson, J. Ladd, A. Dokoutchaev, X. Fan, X. Li, Z. Yin, R. Johnson, V. Lenchenkov, S. Nagaraja, W. Gazeley, J. Bai, H. Lee, A. D'Anna, G. De-Amicis, Pixel continues to shrink, pixel development for novel CMOS image sensors, in *Proceedings of the 2009 International Image Sensor Workshop*, Bergen, Norway, 22–28 June 2009, pp. 58–61
16. Y. Nomirovsky, I. Brouk, C.G. Jackobson, 1/f noise in CMOS transistors for analog applications. IEEE Trans. Electr. Dev. **48**(5), 921–927 (2001)
17. M. Haartman, M. Oestling, *Low-frequency Noise in Advanced MOS Devices* (Springer, Berlin, Heidelberg, New York, 2007)
18. K.M. Findlater et al., Source follower noise limitations in CMOS active pixel sensors. Proc. SPIE, **5251**, 187–195 (2004)

Chapter 10
Low-Noise Electronic Imaging with Double-Gate FETs and Charge-Modulation Devices

Yoshiyuki Matsunaga

Abstract Virtually complete charge transfer can be achieved with charge coupled devices (CCDs) in the buried channel configuration because the buried channel CCD introduces no transfer noise. However, conventional charge detectors incorporated into CCDs have a comparably large noise, which consists of thermal noise and 1/f noise from MOS transistor in the output charge detectors.

Highly sensitive double gate charge detectors are presented in this chapter to show the ability of CCD low noise performance of less than 1 electron rms at room temperature and at video frequency.

The concept of double gate charge detection can alternatively be implemented in a pixel incorporating a charge modulation device (CMD), reaching also an effective readout noise level of less than 1 electron rms. In practice, the CMD shows larger fixed-pattern noise behavior, necessitating supplementary signal processing efforts.

10.1 Introduction

Charge coupled devices (CCDs) have achieved virtually complete charge transfer by using the buried-channel configuration, resulting in charge transfer inefficiencies of less than 10^{-6}. Also, the buried-channel CCD operation introduces no transfer noise. However, conventional charge detection circuits employed in CCDs have a comparatively large noise, which consists of thermal noise and $1/f$ noise from metal oxide semiconductor (MOS) transistors in the charge detectors [1]. The detector noise sources degrade the CCD performance significantly.

In this chapter, highly sensitive double-gate charge detector devices are presented to exploit the ability of CCD lownoise performance which is one of the most important characteristics for image sensor applications. Low leakage current CCD sensors with a low-noise charge detector will ideally be almost noiseless photosensors

In addition, a new charge-modulation image pixel is presented for applications of the double-gate detectors.

P. Seitz and A.J.P. Theuwissen (eds.), *Single-Photon Imaging*,
Springer Series in Optical Sciences 160, DOI 10.1007/978-3-642-18443-7_10,

10.2 Double-Gate FET Charge Detector

The primary parameter dominating the noise performance of a charge detector is the effective capacitance at the gate of the first transistor, as explained for example in the first chapter of this book. For this reason, a successful strategy for the realization of ultra-low-noise photosensors is the design of charge detectors with very low sensing capacitance, preferentially below 1 fF. The double-gate FET structure is very useful for this purpose because of both its high sensitivity due to low sensing capacitance and its low noise due to complete charge transfer in the charge detector. The double-gate FET can be implemented either as a floating well or as a floating surface device.

10.2.1 Floating Well Type

A new high-sensitivity CCD output amplifier called "double-gate floating well detector" has been fabricated and evaluated. At room temperature, the new detector achieves a charge/voltage conversion gain of 76 μV/electron and a noise equivalent signal of 1.2 electrons rms over a 3.58 MHz video bandwidth.

This detector consists of an n-type buried sensing channel and a P-type bulk channel detection transistor crossing underneath the sensing channel, representing the doublegate FET. This structure is very useful for a CCD image sensor which handles a small signal charge of a few thousand photogenerated electrons or holes.

10.2.1.1 Device Structure

Almost all charge detectors for CCD image sensors are floating diffusion detectors. Figure 10.1 shows the top view of the detector incorporated into the end of a CCD register. Signal charges are clocked along the CCD register and transferred out into a CCD sensing channel. Sensing channel potential is modulated by signal charges, and a modulated potential is sensed by the detection MOS transistor's gate. The output voltage signals are buffered by double stage source followers. The detection sensitivity is determined by the conversion gain for signal charges to the output voltages, that is, by the effective sensing capacitance.

However, the sensing capacitance of the conventional detector cannot be sufficiently small for the following reasons: First, the sensing channel has a comparatively large area because it inevitably has a contact hole. Second, the detection transistor gate capacitance must be added to the sensing capacitance. Third, the detector suffers from large kTC noise, corresponding to charge fluctuation in the reset gate OFF period due to reset gate channel thermal noise in the ON period.

The newly developed charge detector has low sensing capacitance and a structure that is free from kTC noise. The principle is a combination of an n-type buried sensing channel and a P-channel detection transistor across it, as shown in Fig. 10.2

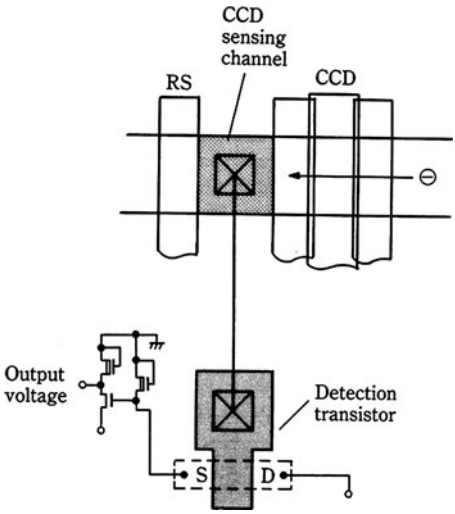

Fig. 10.1 Plane view of conventional floating diffusion charge detector

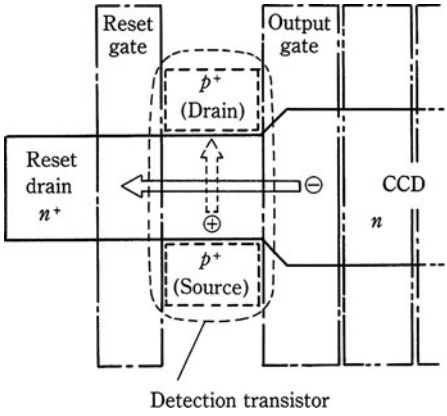

Fig. 10.2 Plain view of new floating well detector

[2, 3]. Figure 10.3 shows a cross-sectional view of the detection transistor across the sensing channel. A shallow p-well with a sensing channel in it and source/drain p+ regions are formed on an n-type substrate. The hole current in the P-channel detection transistor is made to flow not along the silicon–silicon dioxide interface but rather underneath the sensing channel. As a result, suppression of $1/f$ noise which is caused by silicon–silicon dioxide traps can be expected. Charge packets clocked along the CCD channel are transferred into the sensing channel. The charge packets directly modulate the channel conductance in the detection transistor.

The sensing capacitance for the detection transistor can be smaller than that for the widely used "floating diffusion detector." The small sensing capacitance has been achieved by the smaller sensing region, in which the sensing channel and the P-type channel for the detection transistor are crossed. Moreover, a "double-gate structure" is introduced for smaller sensing capacitance. In this structure, a 1,000 nm

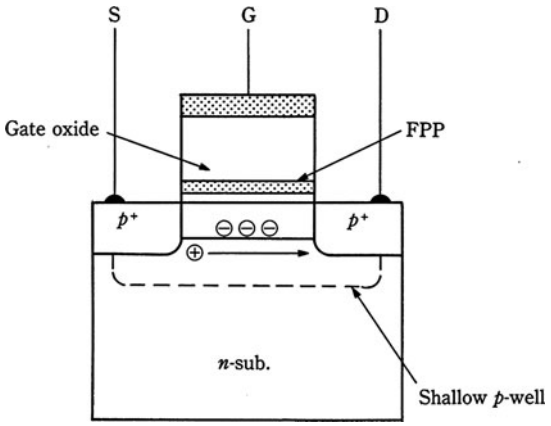

Fig. 10.3 Cross-section of floating well detection transistor

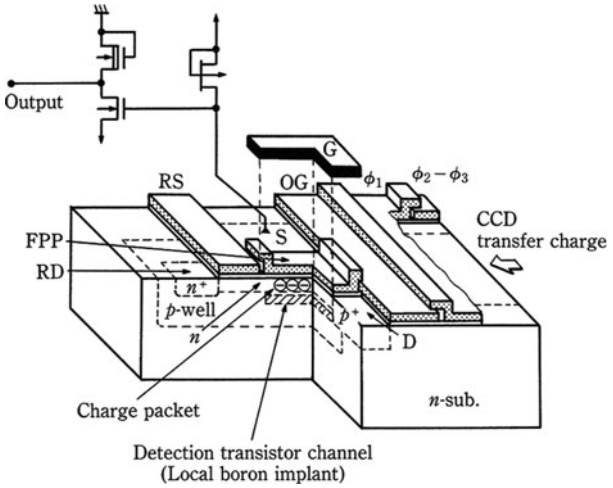

Fig. 10.4 Practical floating well charge detector fabricated in shallow p-well on n-type substrate

thick oxide is employed for the detection transistor's gate oxide. A flat potential plate (FPP), which is floating electrically, is inserted under the detection transistor gate (G). The FPP suppresses the two-dimensional effect which causes a potential pocket in the sensing channel. The transfer loss induced by a potential pocket results in transfer noise in the detector.

Figure 10.4 shows a practical output amplifier for detecting the charge (electron) packets transferred in a three-phase linear CCD which is fabricated in a shallow p-well on the n-type substrate of 45 Ωcm resistivity. The potential control gate (G) area with the thick gate oxide layer measures $5 \times 5\,\mu m^2$, and the thin gate layer under the FPP is 93 nm thick. To form the P-type channel for the detection transistor, a local deep boron implantation the sensing channel is required. Figure 10.5 shows

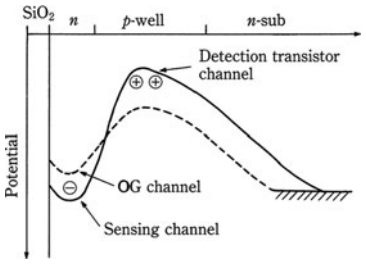

Fig. 10.5 Potential profiles along the direction perpendicular to device surface

the potential profiles along the direction perpendicular to the surface. The solid line and the dotted line are those under the FPP and under the output gate (OG), respectively. The abrupt potential profile under the sensing channel is essential for the signal charge in the sensing channel to modulate the detection transistor channel conductance effectively. The source and drain regions of the P-channel detection transistor are made by self-alignment with the output gate (OG), reset gate (RS) and the FPP. The source is connected to the P-channel load transistor made by a junction-type FET incorporated into the same chip. The junction FET is superior for suppressing $1/f$ noise compared to a MOS-FET [4]

A charge (electron) packet transferred into the sensing channel through the output gate (OG) is converted to a voltage signal, using the firststage source-follower which consists of the detection transistor and the load transistor. An output signal is generated using the second stage source-follower constructed with n-channel MOS transistors. To remove the charge (electron) packet from the sensing channel region into reset drain (RD), a positive pulse is applied to the reset gate (RS). Then the sensing channel is completely depleted in a so-called "transfer operation." Thus, the detector has no reset noise [5–7].

10.2.1.2 Device Simulation

A two-dimensional device simulator was utilized for designing the detector. In a conventional detector, an effective sensing capacitance is determined approximately by the gate oxide capacitance of a detection transistor [2,3]. For high sensitivity, low sensing capacitance has been achieved by employing a 1 μm thick gate oxide for the detection transistor. Figure 10.6a shows a simulated potential profile for electrons in the detecting section. A potential pocket is observed at the edge of the output gate (OG). The potential pocket causes not only charge (electron) transfer loss but also transfer noise. To eliminate the pocket, the FPP has been inserted under the potential control gate (G). The potential profile in the new detector with the FPP is shown in Fig. 10.6b. The potential pocket vanishes, which enables the new structure to obtain reduced transfer loss sufficient for the realization of a low-noise detector.

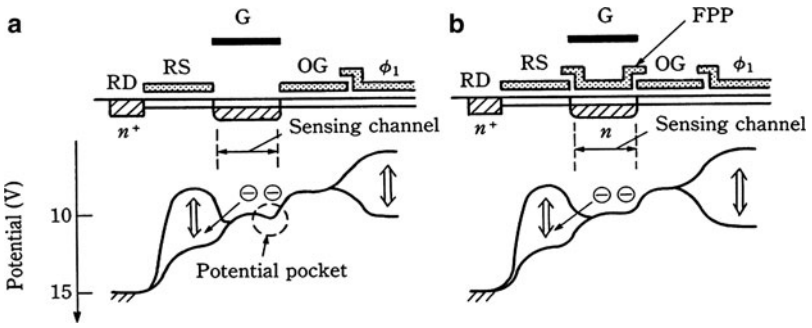

Fig. 10.6 Potential profiles along the direction perpendicular to device surface. (**a**) Detection transistor with only 1 μm thick gate oxide (**b**) Detection transistor with FPP

10.2.1.3 Evaluation

The detector was evaluated at a 7.16 MHz clock frequency, corresponding to twice the NTSC color subcarrier frequency. The signal duty defined by the signal output period in the clock cycle was approximately 70%, which was determined by a 40 ns positive pulse supplied to the reset gate (RS). The charge/voltage conversion gain of the detector was measured by injecting a series of charges in a packet into the CCD from an injection diode having a capacitance of about 1 fF. The size of the charge packet was calculated from the reset drain current. The resultant output amplitude was observed through an oscilloscope. The output signal bandwidth is 20 MHz under 7 pF capacitance load. The results of the measurements are shown in Fig. 10.7 The relation between injection charge quantity and output voltage is linear over a range of charge packets from 40 to 2,000 electrons/packet, with a responsivity of 76 μV/electron. The output noise of the new detector was measured using a spectrum analyzer. Figure 10.8 shows the noise spectra of the detector, both in reset operation and in transfer operation. In the reset operation, residual charge in the sensing channel includes reset noise (kTC noise), charge due to thermal noise generated in the channel region of the reset gate (RS). Conversely, in the transfer operation, no noise charge exists, because the signal charge in the sensing channel is completely transferred into the reset drain by applying a positive pulse to the reset gate. The $1/f$ noise below 2 MHz, which is observed in the noise spectrum for the transfer operation, is generated by the detection transistor. The overall RMS noise value in the transfer operation, within the frequency range of 10 kHz–3.58 MHz, was 64 μV rms at room temperature. The value for the "floating surface detector" with the FPP structure, which is fabricated on the same substrate for comparison and is described below, was 84 μV rms under the same measuring conditions. The difference in the noise values is caused by the $1/f$ noise due to silicon–silicon dioxide interface traps. As shown in Fig. 10.7, the new detector with a charge/voltage conversion gain as high as 76 μV/electron, a noise level as low as 64 μV rms and a 70% signal duty ratio, achieves an effective readout noise as low as 1.2 electrons rms.

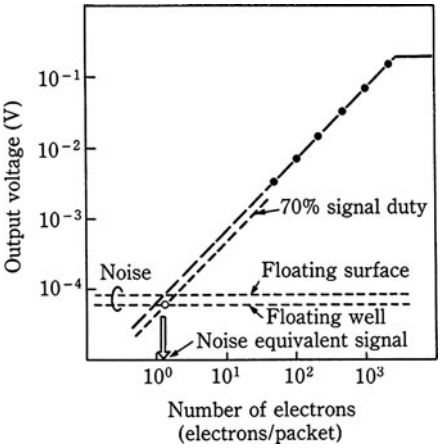

Fig. 10.7 Charge/voltage conversion characteristics and noise levels

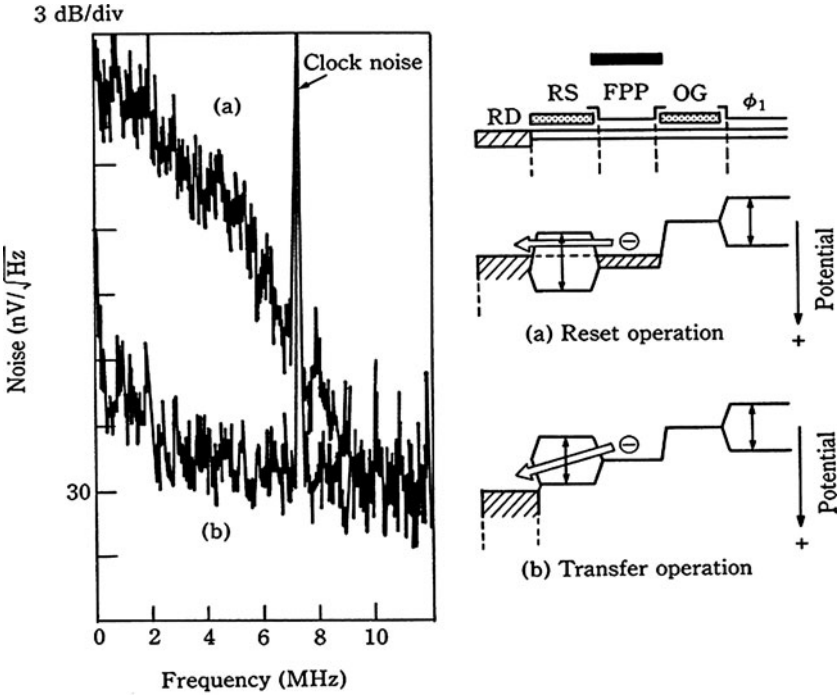

Fig. 10.8 Noise spectra in reset operation and in transfer operation

10.2.2 Floating Surface Type

An alternative low-noise and wide dynamic range charge detector called "double-gate floating surface detector" differs from the floating well type in that the P-channel FET current flows at the silicon–silicon dioxide interface. An advantage of this type is its high conversion gain due to the small sensing area compared to the floating well type. The device achieves a charge-to-voltage conversion gain of 220 μV/electron and a dynamic range of 79 dB at 3.58 MHz video bandwidth and at room temperature. This detector shows an effective readout noise of only 0.5 electrons rms, so that it can be used as an electron counting detector and the evaluation can be done by directly observing the discrete levels corresponding to the number of signal electrons. The 79 dB dynamic range is sufficient for a CCD image sensor detector.

10.2.2.1 Device Structure

The developed charge detector has low sensing capacitance and a structure which is inherently free of kTC noise. The principle is a combination of an n-type buried sensing channel and a P-channel surface detection transistor across it [5–7]. Figure 10.9 shows a cross-sectional view of the detection transistor across the sensing channel. A shallow p-well with a sensing channel in it and source/drain p+ regions are formed on an n-type substrate. The shallow p-well under a sensing channel is completely depleted. The P-type channel for the detection transistor lies at the silicon–silicon dioxide interface. The charge packets modulate the channel conductance in the P-channel detection transistor.

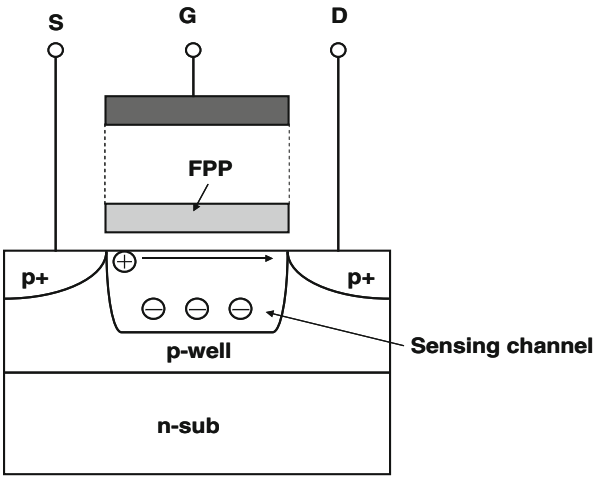

Fig. 10.9 Crosssection of floating surface detection transistor

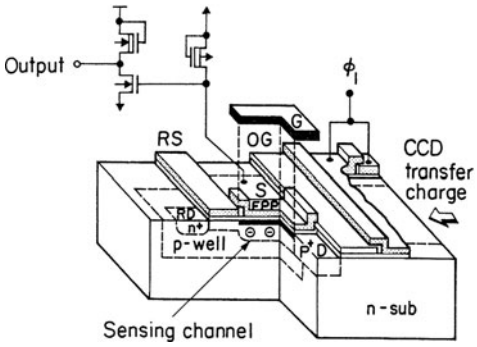

Fig. 10.10 Practical floating surface charge detector structure

Figure 10.10 shows a practical output amplifier for detecting the charge (electron) packets transferred in a three-phase linear CCD, which is fabricated in a shallow p-well on the n-type substrate of 45 Ωcm resistivity. The potential control gate G area with the thick gate oxide layer measures $4 \times 4\,\mu m^2$ which is smaller than that of the floating well type. The thin gate layer under the FPP is 120 nm thick which is slight thicker than that of the floating well type. To form the P-type channel for the detection transistor, a local deep phosphorus implant on the sensing channel is required [3, 7]. A charge (electron) packet is transferred into the sensing channel through the output gate (OG) at the fall time of a $\phi 1$ clock voltage applied on the last transfer electrode of the CCD. Subsequently, it is converted to a voltage signal using the first-stage source-follower, which consists of the detection transistor and a load transistor. The drain of the detection transistor is grounded, and the source voltage of the load transistor is 15 V. The voltage applied to the potential control gate G is adjusted between 30 and −80 V. An output signal is generated using the second-stage source-follower constructed with n-channel MOS transistors. To remove the charge (electron) packet from the sensing channel region into the reset drain (RD), a positive pulse is applied to the reset gate (RS). This completely depletes the sensing channel in a so-called "transfer operation." As a consequence, the detector has no reset (kTC) noise [5, 6].

10.2.2.2 Device Simulation

A two-dimensional device simulator, handling both signal electrons in a bulk potential well and hole current along the silicon–silicon dioxide interface simultaneously, has been newly developed to analyze the floating surface detector. The simulations were done along the signal electron transfer direction. Figure 10.11a,b shows the potential distributions for no signal charge and for 3600-electron signal charge, respectively. The potential level at $y = 0$ in Fig. 10.11 indicates the surface potential. The first maximum potential level along the y axis indicates the buried-channel

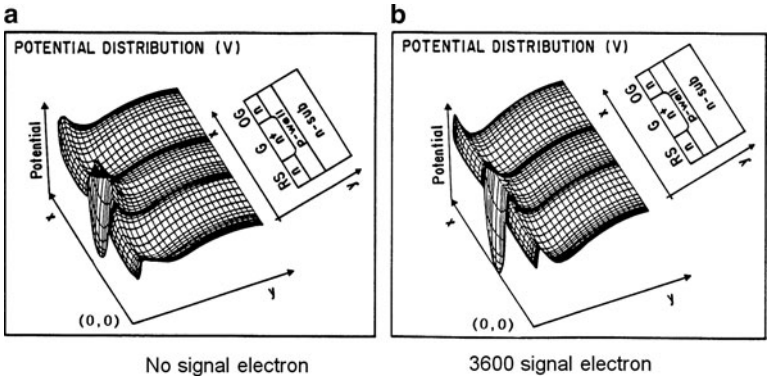

Fig. 10.11 Two-dimensional potential distribution at detection section. (**a**) No signal electron. (**b**) 3,600 signal electron

potential. As shown in Fig. 10.4b, the surface potential in the detector section is lowered by signal charge, compared to that for no signal charge. The sensing capacitance is calculated with the amount of signal charge and the difference in the surface potentials between Fig. 10.11a,b which corresponds to the charge/voltage conversion gain of the detector. The resultant sensing capacitance is 0.22 fF/μm per FPP length.

10.2.2.3 Evaluation

The device is evaluated at a 7.16-MHz clock frequency, which is twice the NTSC color subcarrier frequency. The signal duty, which is the signal output period to the clock cycle time ratio, is approximately 75%, which was determined by a 35 ns positive reset pulse. The measured charge-to-voltage conversion gain of this detector is shown in Fig. 10.12 as a solid line. It was measured by injecting a series of charge packets into the sensing channel. The conversion gain is linear over a charge packet range from 20 to 4,500 electrons/packet, and it is as high as 220 μV/electron. The measured conversion gain is higher than that obtained from the simulated sensing capacitance value by about 20%. A three-dimensional device simulation may be needed for more accurate estimation, because the difference between the measurement and the calculation can be due to potential distribution perpendicular to the signal electron transfer direction.

The output noise of the new detector was measured using a spectrum analyzer. Figure 10.13 shows noise spectra for the device, both in the reset operation and in the transfer operation. In the reset operation, residual charge in the sensing channel includes reset noise (kTC noise) charge due to the thermal noise generated in the reset gate (RS). Conversely, in the transfer operation, the reset noise is completely suppressed, because no residual charge exists in the sensing channel. The overall

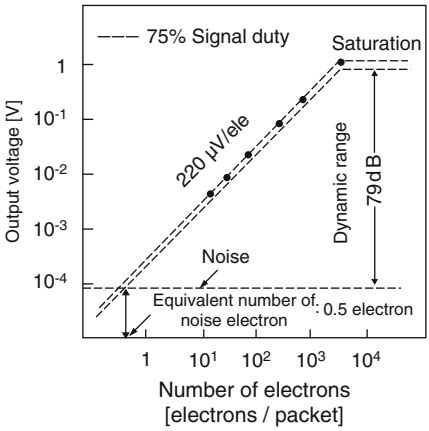

Fig. 10.12 Charge/voltage conversion characteristics

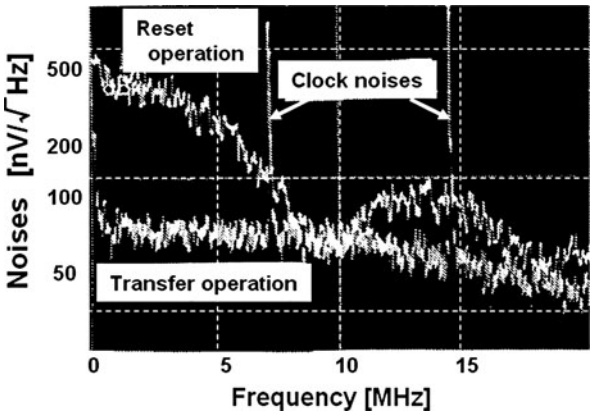

Fig. 10.13 Noise spectra of floating surface detector in reset operation and in transfer operation

rms noise value, within the frequency range of 10 kHz–3.58 MHz, is 84 μV rms at room temperature.

The noise level of 84 μV rms and a signal duty ratio of 75% result in an effective readout noise of as low as 0.5 electron. Thus, the dynamic range is as wide as 79 dB.

In the signal range below 20 electrons, charge generation by photoelectric conversion effect in the CCD with weak incident light through a neutral density (ND) filter is induced to evaluate the detector, as shown in Fig. 10.14. Injection packet size in the sensing channel is known in advance by this method. A signal packet of ten electrons is made by inserting a 10^{-2} ND filter in the light path. The light intensity is set so as to generate 1,000 electrons/packet, which is measured using the current meter in the figure. Figure 10.15a shows waveforms of the RS input pulse, a $\phi 1$ input pulse, and the resultant output signal. Figure 10.15b is a

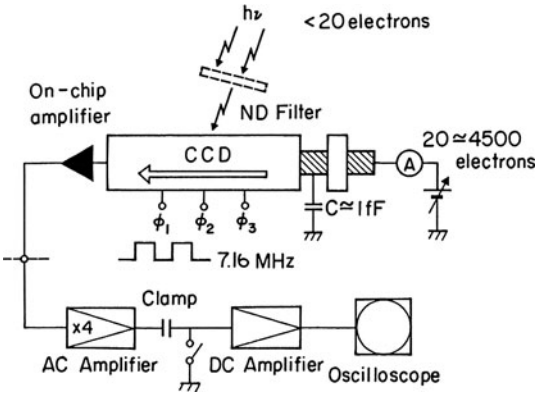

Fig. 10.14 Evaluation circuit

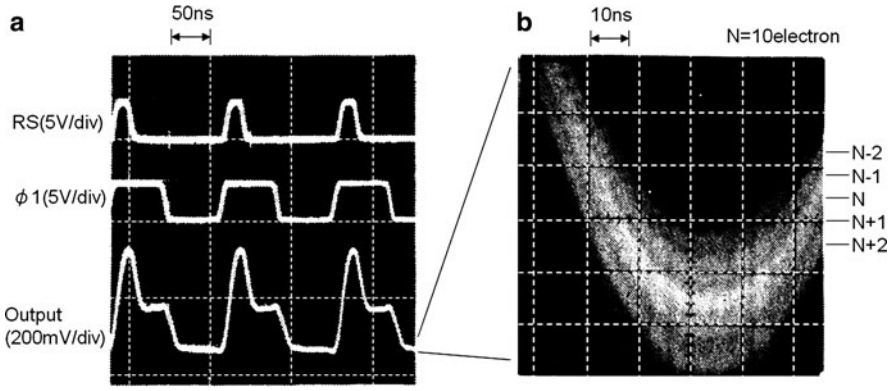

Fig. 10.15 Charge/voltage conversion characteristics. (**a**) RS, φ1 input, and resultant signal output. (**b**) Amplified output voltage around ten signal electrons

photograph of the signal voltage amplified by an external circuit when the average number of electrons in a signal packet is 10. The amplified signal is processed through a 3.58 MHz low-pass filter to remove worthless high-frequency noise. Discrete voltage levels corresponding to around ten electrons are observed. Output levels, which correspond to other than ten electrons, are caused by fluctuation of the incident light. The charge-to-voltage conversion characteristics obtained by observing the discrete levels are shown in Fig. 10.16. The conversion gain above 10 electrons in Fig. 10.16 is approximately the same as that in Fig. 10.12. The conversion ratio under ten electrons shows less linearity. However, high responsivity is still maintained in the small signal region.

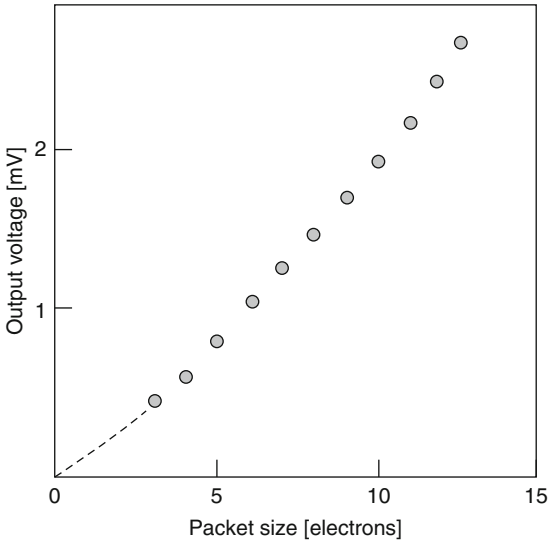

Fig. 10.16 Charge/voltage conversion characteristics obtained by discrete level observation

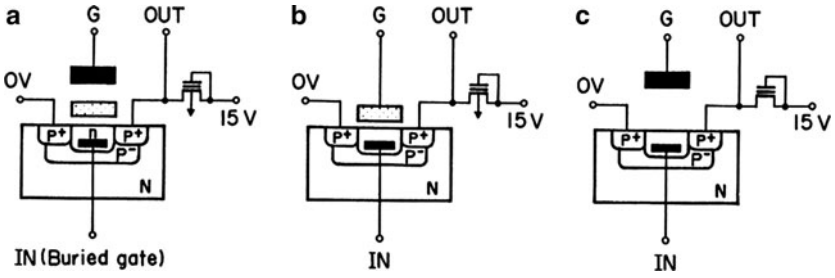

Fig. 10.17 Charge/voltage conversion characteristics. (**a**) Double-gate transistor. (**b**)Thin-gate-oxide transistor (**c**) Thick-gate-oxide transistor

10.2.2.4 Discussion

The detection transistor is the P-channel surface transistor with the sensing channel as the buried gate. Here, the gate electrode structures in the transistor are discussed. Figure 10.17a–c shows source-follower diagrams using the double-gate transistor, a thin-gate-oxide transistor, and a thick-gate-oxide transistor, respectively. The oxide thicknesses of the transistors are 120 and 1,000 nm, respectively. The same voltage as the RD voltage is applied to the buried gate by switching the neighboring RS gate on. Figure 10.18 shows the buried-gate voltage versus the output voltage of the source followers. The source-follower gains are obtained by the slopes of curves (a)–(c) in the figure. They are 0.78, 0.31, and 0.49, respectively. In the buried-gate voltage range of 13–15 V, the output voltages of types (a) and (b) are saturated.

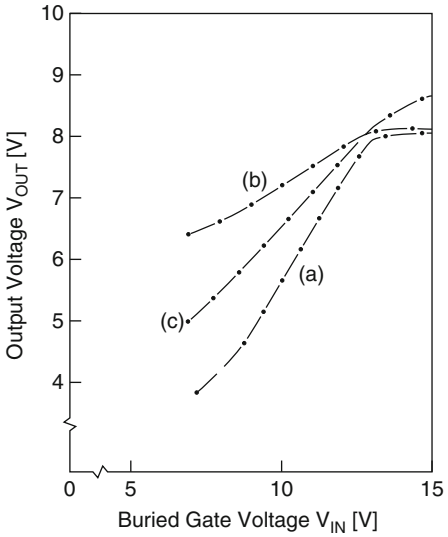

Fig. 10.18 Source-follower characteristics

This shows that the sensing channels are completely depleted. On the contrary, the saturation is not observed in type (c). The source-follower gain of type (c) is smaller than that of type (a) in spite of the thick gate oxide, since the P-type channel is also not flat and the transistor characteristics are degraded.

The sourcefollower gain g1 is represented as $C_{BC}/(C_{BC} + C_{GC})$, where C_{BC} and C_{GC} are buried-gate transistor channel capacitance and gate oxide capacitance, respectively. However, the charge-to-voltage conversion gain $g2$ is proportional to $(1/C_{BC} + 1/C_{GC}) \times C_{BC}/(C_{BC} + C_{GC}) = 1/C_{GC}$, because the depletion-layer depth under the sensing channel reaches to 15 μm, and the capacitance between the buried gate and the substrate is negligibly small. So, the relationship between $g1$ and $g2$ is described by the following equation:

$$g2 = \{g1/(1 - g1)\}C_{BC}. \tag{10.1}$$

Equation (10.1) shows that the charge-to-voltage conversion gain $g2$ strongly depends on the source-follower gain $g1$, $g2$ increases dramatically if $g1$ is close to 1. Especially, the charge-to-voltage conversion gain rapidly decreases in the small-signal region. The conversion gain increases seven times by using the double-gate structure compared to the thin-gate-oxide transistor. The conversion gain ratio of the double-gate to thick-gate-oxide transistors is 220:31 = 3.55:0.50, in good accordance with the $g1/(1 - g1)$ ratio of 3.55:0.45. It is clear from the above discussion that the double-gate structure can realize superior high conversion gain detectors.

10.3 CCD Image Sensor with Double-Gate FET Charge Detector

The pixel size of the CCD image sensor has been made smaller to reduce chip size for small optical assembly and to realize megapixel sensors for HDTV cameras. It is not acceptable, however, to compromise the sensor performance for a smaller cell size, and in particular, the sensitivity and the dynamic range remain high [1]. In contrast, the buried-channel charge-coupled device achieves complete charge transfer with no transfer noise. Therefore, almost all solid-state image sensors utilize low-noise buried-channel CCD for scanning circuits.

In the following we describe a 1/3-in interline transfer CCD area image sensor with 94-dB dynamic range and 1.0-electron rms noise floating surface detector. With capacitance-coupled negative-feedback configuration, the saturation level of the detector is successfu1ly extended by a factor of 5, maintaining the superior low-noise characteristics. Moreover, $1/f$ noise in the output signal, which is a very conspicuous streak noise in the reproduced image in spite of its low-noise power, is successfully suppressed with a novel so-called alternate gain inversion (AGI) signal processing technique.

10.3.1 Sensor Construction

Figure 10.19 shows photographs of the fabricated device chip and the detection transistor. The chip size is 7.1 mm (H) × 5.1 mm (V), and the image size is 4.8 mm (H) × 3.6 mm (V), which is suitable for a 1/3-in lens format. The device is the most popular interline transfer CCD configuration at present, which consists of photodiodes, vertical CCDs (V-CCDs), and a horizontal CDD (H-CCD). For low leakage current noise, the buried photodiode structure, which has a high concentration p^+-layer on the n-type storage region, is adopted [7, 8]. The charge detector at the end of the H-CCD is a double-gate floating surface detector with an effective readout noise of 1.0 electron rms. A dynamic range of 94 dB is obtained by returning the inverted voltage signal onto a feedback gate. The saturation number of electrons is extended to 5×10^4 electrons, which corresponds to the saturation charge of the photodiode.

The device is fabricated in a shallow p-well on an n-type substrate of 47 Ωcm resistivity with 1.0 μm fine pattern design rule. Two ion-implantation processes are added to the conventional CCD fabrication process. One is a high-energy phosphorous implantation for the sensing channel, and the other is a source/drain boron implantation for the P-channel detection transistor.

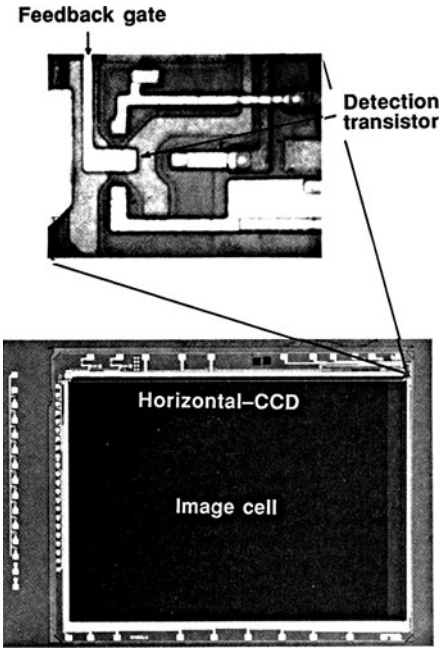

Fig. 10.19 Photographs of sensor chip and enlarged detector

10.3.2 Feedback Charge Detector

Figure 10.20 shows the detector structure. The signal charge (electron) packets are clocked along the H-CCD, and transferred into the detection transistor signal sensing region under the FPP. The detected charges are moved out into a reset drain (RD) by applying a positive pulse on a reset gate (RS). The signal charge in the sensing region modulates the P-channel conductance in the detection transistor. An output signal is generated using two source-follower stages, constructed with n-channel MOS transistors. An inverted signal voltage is applied to the feedback gate of the detection transistor. The feedback amplifier is an inverter and its noise is designed to be much smaller than that of the source followers in order not to add extra noise by the feedback loop. The capacitance-coupled feedback loop allows the dynamic range to be extended. The oxide thickness between the feedback gate and the FPP is approximately 1.0 μm.

Figure 10.21 shows the increase in the saturation charge by the negative feedback operation. Signal charge Qs is converted into voltage signal Vs by the very low sensing capacitance Ce in the case of the open loop as shown in Fig. 10.21a. On the contrary, in the feedback operation, the effective sensing capacitance increases n times by the inverted signal voltage, which is applied to the feedback gate, so that the output signal voltage is compressed, as shown in Fig. 10.21b, and the saturation charge increases by a factor of n, as shown in Fig. 10.21c. The parameter n is equal to the well-known feedback formula $(1 + kA)$, where k and A are the feedback

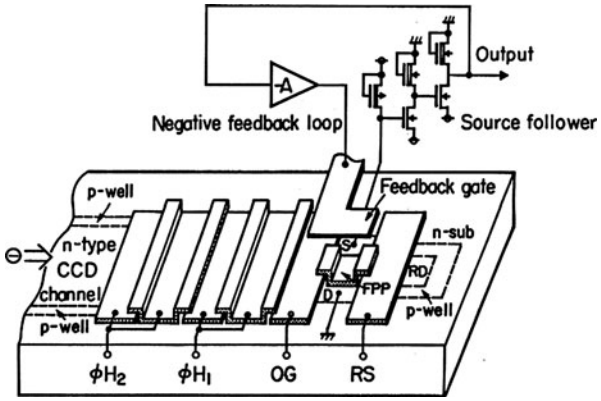

Fig. 10.20 Detector structure

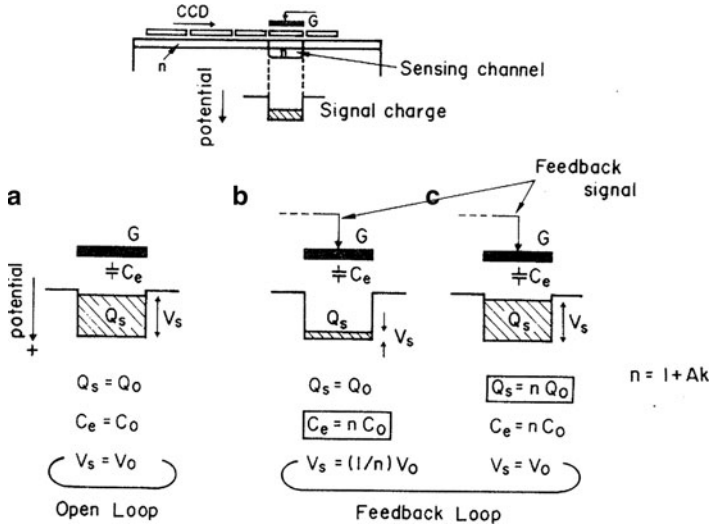

Fig. 10.21 Potential diagrams of sensing channel. (**a**) Saturation condition in open loop. (**b**) Compressed condition by negative feedback. (**c**) Saturation condition in negative feedback

coefficient and the feedback amplifier gain, respectively. In this detector, k is equal to $Cg/(Cg + Cc)$, where Cg and Cc are the gate capacitance of the detection transistor and the diffusion capacitance of the n-type sensing channel under the FPP, respectively. It is derived from the partition ratio of the Cg and Cc series capacitances.

Noiseless signal charge is converted into the voltage signal S, while the detector noise is N in the open-loop charge detector. By applying the inverted output signal onto the feedback gate, not only the voltage signal but also the detection noise is compressed, and they result in S' and N'. Signal-to-noise ratio S'/N' is equal to

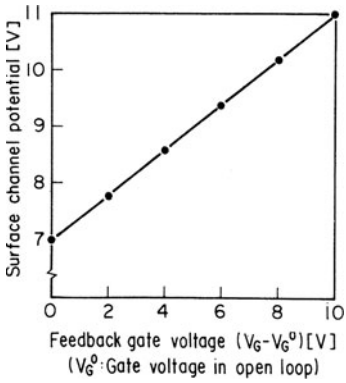

Fig. 10.22 Surface channel modulation of detection transistor

S/N, as well known from the negative-feedback theory. Thus, the superior noise performance of the detector is maintained.

10.3.3 Evaluation

Figure 10.22 shows the surface-channel potential of the detection transistor versus the feedback gate voltage. The surface-channel potential is measured using the test device of the detection transistor fabricated on the same chip as the image sensor. The feedback coefficient k, which corresponds to the slope of the line in the figure, is approximately 0.4. In the practical detector, n is determined to be 5, taking the photodiode saturation into consideration. Accordingly, the feedback amplifier gain A is set to 10. The equivalent input noise voltage of the feedback amplifier is approximately 10 μV rms at 3.58 MHz bandwidth which is significantly smaller than the 80 μV rms of the source followers.

The H-CCD (ϕH1, ϕH2) and the reset gate RS were driven at a 7.16 MHz clock frequency, which is twice the NTSC color subcarrier frequency. Figure 10.23 shows the RS and ϕH1 input pulses, and the output voltage signals. Figure 10.23a,b shows the output voltage signal of the saturation level without the feedback loop and the compressed voltage signal by the negative feedback, respectively. The waveform X in the figure is not the clock noise, but a feedthrough voltage of a part of the inverted signal from the feedback gate. Figure 10.23c is the saturation signal with the feedback loop. The noise voltage is also suppressed 14 dB. Figure 10.24 shows the measured charge-to-voltage conversion characteristics and the noise levels of the detector. The charge-to-voltage conversion gain is linear within 5% over a range of charge packets from 200 to 10,000 electrons/packet. The charge/voltage conversion gain, the noise voltage, and the saturation voltage of the detector without the feedback loop are 100 μV/electron, 80 μV rms, and 1.0 V, respectively. The saturation charge of 10,000 electrons/packet and the noise equivalent signal of 1.0 electron rms are obtained from these values, taking a signal duty T_s/T_0 into consideration, where T_s and T_0 are the effective output signal period and the clock

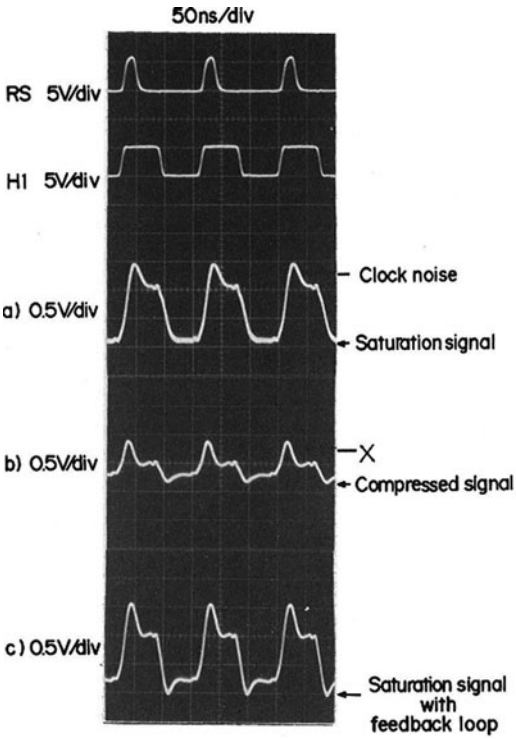

Fig. 10.23 Input pulses and output signal waveforms of detector. (**a**) Saturation signal without feedback loop. (**b**) Compressed signal with negative feedback. (**c**) Saturation signal with negative feedback loop

cycle of the horizontal CCD driving pulse, respectively. With the capacitance-coupled feedback loop, the saturation charge level is increased by a factor of 5, and the dynamic range is improved to 94 dB, maintaining the effective readout noise of 1.0 electron rms.

The characteristics of the image sensor with the new detector are shown in Table 10.1. The very low dark current is realized by the buried photodiode structure and the pinned operation of the V-CCD. In the pinned operation, the surface emission under the transfer electrodes, on which the low-level voltage of transfer pulse is applied, is suppressed because holes are accumulated there.

10.3.4 Signal Processing

The output signal includes the large $1/f$ noise generated by the MOS transistors of the on-chip detector. Figure 10.25 shows the noise spectrum of the on-chip detector without the feedback loop. The $1/f$ noise generates a very conspicuous streak noise in the reproduced image in spite of its small noise power. The $1/f$ noise should therefore be suppressed to realize a low-noise image sensor. A noise reduction signal processing technique called alternate gain inversion (AGI) is proposed to solve the problem. This processing is able to suppress not only the $1/f$ noise but also the kTC

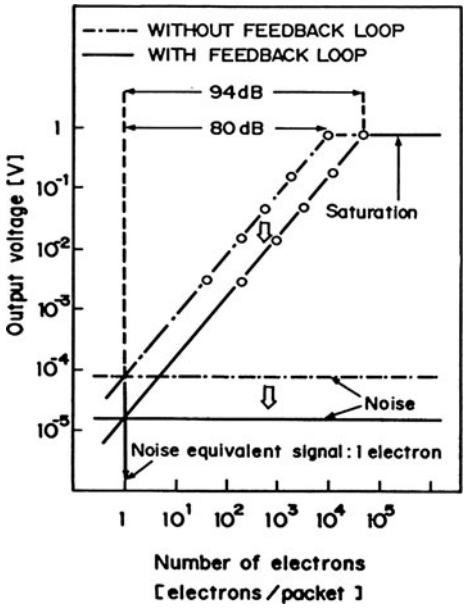

Fig. 10.24 Charge/voltage conversion characteristics of detector

Table 10.1 Characteristics of image sensor

Image area	4.8 (H) × 3.6(V) mm^2 (1/3 in.)
Chip size	7.1 (H) × 5.1(V) mm^2
Number of pixels	398 (H) × 492(V)
Sensitivity	100 mV/L × (2,856 K)
Saturation voltage	1 V
Noise (detector)	20 μ V rms
Dark current (20°C)	2 pa

reset noise, which is the main noise generated by the conventional floating diffusion detector. Figure 10.26 shows the input pulses ϕH1 and RS, the output signal OS, and the waveform SP after the AGI processing. Here, the OS has a constant noise $+N$, as shown in the figure. In the AGI processing, the OS waveform between times t_R and t_S is inverted. The noise $+N$ is inverted into $-N$ in this period. The times t_R and t_S are the intermediates of the reset gate ON period T_R and the OFF period T_W, respectively. The clock cycle T_0 is the sum of T_R and T_W. The noise $-N$ in the feedthrough period cancels the noise $+N$ in the signal output period by mixing the noises $+N$ and $-N$ through a low-pass filter. The transfer function $T(\omega)$ of the AGI processing is shown in Fig. 10.27 It is given by:

$$T(\omega) = Z \cdot \sin(\omega T_0/4) \cdot \frac{\sin(\omega T_0/4)}{\omega T_0/4}, \tag{10.2}$$

where Z is a constant. The second term denotes the transfer function of the well-known correlated double sampling (CDS). In the CDS signal processing,

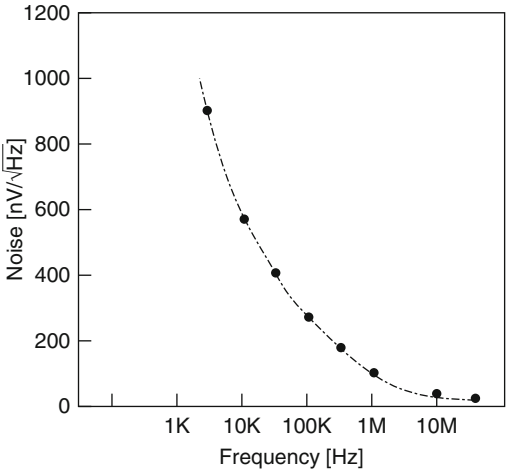

Fig. 10.25 Noise spectrum of detector

the output is obtained by subtracting the noise in the feedthrough period from the CCD output signal. Therefore, the transfer function of the CDS is expressed as a formula of $\sin(Y)$ [8]. The last term corresponds to the Fourier transform of the aperture function of $T_0/2$. The frequency region under π/T_0 is the effective signal band. The noise in the low-frequency region can be suppressed significantly. The high-frequency aliasing noise is much smaller than that of the conventional CDS processing, because no sample-and-hold circuit is used in the AGI circuit [8]. Figure 10.28 shows a schematic diagram of the AGI circuit employed at this time. Emitter followers Q_1Q_2, and Q_3 reduce the impedance to drive the following stages. The voltage gain of the complementary amplifier B is set to 4. The AGI circuit is constructed with the balanced modulator MC1594 (Motorola). Input pulses P_1 and P_1 (bar) control the inverted period in the AGI output waveform. P_1 is the same as the horizontal CCD driving pulse ϕH1. The offset levels of the inverted period and the noninverted period in the AGI output signal are adjusted with the variable resistances R1 and R2. Figure 10.29a,b shows photographs of the noises of the reproduced images before and after the AGI signal processing, respectively. The $1/f$ noise, which is a very conspicuous streak noise in Fig. 10.29a, is reduced to the level shown in Fig. 10.29b. With this processing, the $1/f$ noise decreases 1/8 times at 100 kHz. The reproduced image of a RETMA chart processed with the AGI is shown in Fig. 10.30.

10.4 Charge-Modulation Image Pixel Application

The charge-modulation-type image pixel sensor, which has P-type sensing area and n-type detector transistor, has been proposed by Nakamura et al. proposed with an annular MOS gate transistor [9]. In the MOS transistor, photogenerated charges are stored in the surface depletion layer under the MOS gate, resulting in a change of

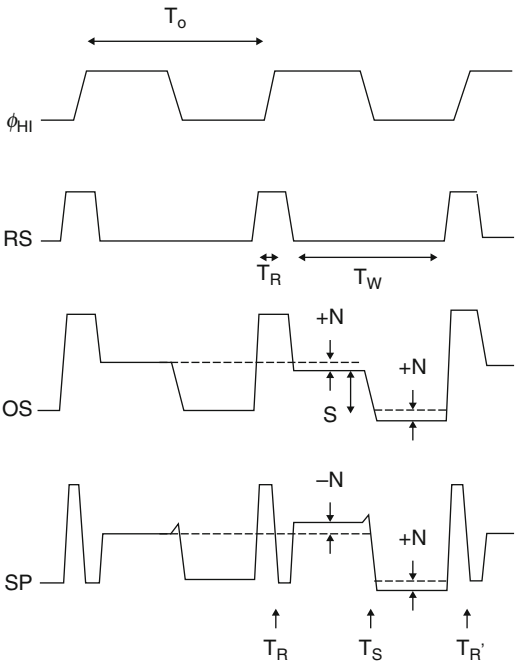

Fig. 10.26 Waveform diagrams in AGI signal processing

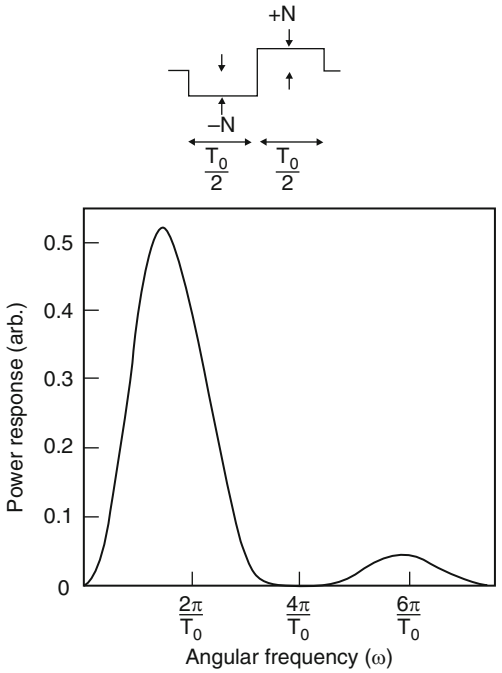

Fig. 10.27 Transfer function of AGI signal processing

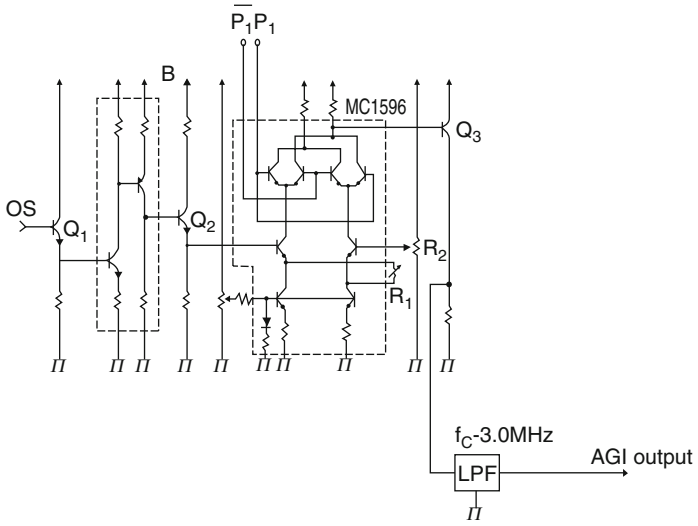

Fig. 10.28 Circuit diagram of AGI

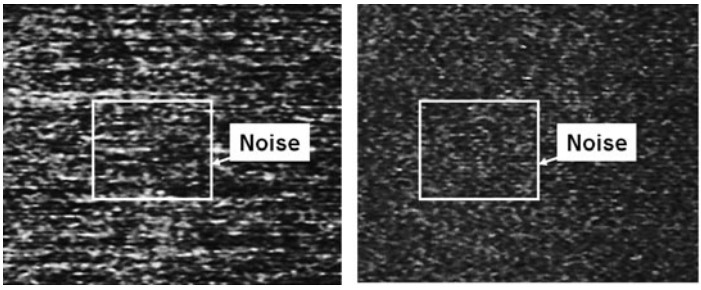

Fig. 10.29 Reproduced imaged of noises of device. (**a**) Before AGI processing. (**b**) After AGI processing

the gate voltage. Drain-to-source current is modulated by this voltage change, and an amplified signal is obtained which is dependent on the light intensity. However, this sensor is not as sensitive as a CCD image sensor. The reason is low optical gain due to large charge sensing area.

As a consequence, a "double-gate floating surface phototransistor" is introduced for high sensitivity image sensor comparable to a CCD image sensor. In the phototransistor cell, the gate area has been divided into two parts, a large accumulation section and a small detection section, to realize a low-input capacitance for a high optical gain. A new readout operation, called "line potential modulation operation," is also introduced.

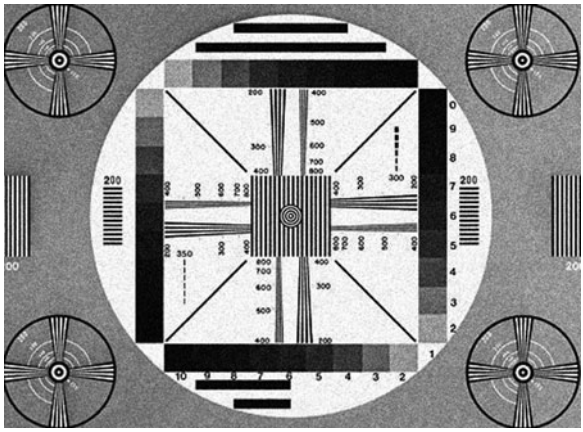

Fig. 10.30 Reproduced image of RETMA chart. (**a**) Before AGI processing. (**b**) After AGI processing

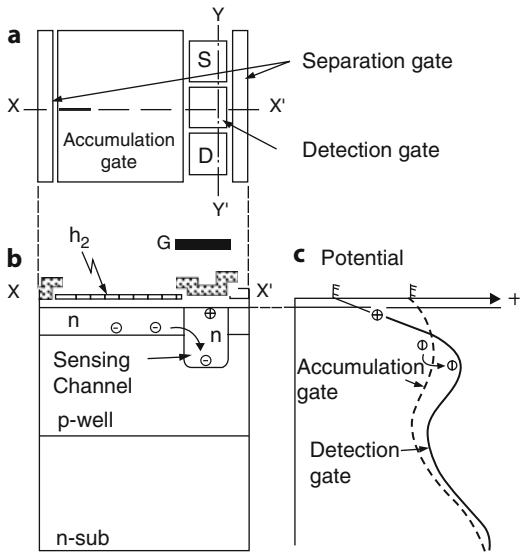

Fig. 10.31 Schematic structure of floating-surface phototransistor. (**a**) Layout. (**b**) Cross-sectional structure. (**c**) Potential profiles under accumulation gate and detection gate

10.4.1 Pixel Construction

Conventional phototransistors have only one annular-type gate [4, 9], whereas the new phototransistor has two gates, an accumulation gate and a detection gate. The accumulation gate is formed in a large area for a large light aperture, and the detection gate is in a small area for small input capacitance. Figure 10.31a,b shows the schematic top and a cross-sectional view of the device. The unit cell consists of two separation gates, an accumulation electrode and a sensing transistor. The unit cell size is $13 \times 13\,\mu m^2$. The fabrication process is similar to that of a CCD image sensor. The device was fabricated using $2\,\mu m$ pattern rule process technology. The

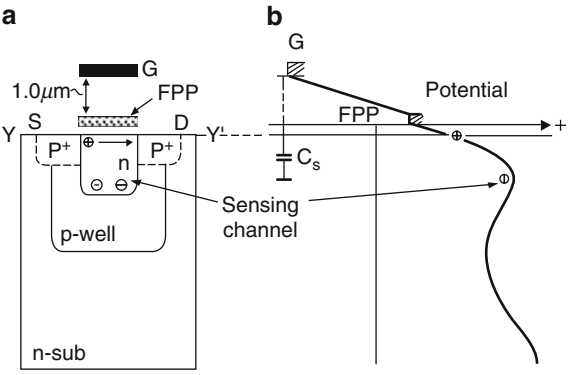

Fig. 10.32 Detection transistor. (**a**) Cross-sectional view of sensing transistor. (**b**) Potential profile in doublegate transistor

accumulation electrode is a thin poly-Si with 60 nm thickness for higher optical efficiency. Figure 10.32a illustrates the cross-sectional view of the sensing transistor adjacent to the accumulation gate. The transistor is formed in a depleted shallow p-well on a 45 Ωcm n-type substrate. The sensing channel of the transistor, which is deeper than the n-type diffusion channel under the accumulation electrode, is made by higher energy phosphor local implantation. As shown in Fig. 10.32b, the small input capacitance Cs has been realized by a double-gate structure. This structure consists of an FPP floating electrically, and a potential control gate (G) over the FPP through a thick gate oxide. The potential control gate area is $4 \times 4\,\mu m^2$ with 1.0 μm thick gate oxide and 120 nm thin gate oxide under the FPP. The input capacitance is estimated to be 0.8 fF. The FPP suppresses the generation of a potential pocket between the accumulation gate and the detection gate, which will be mentioned later. This pocket causes random noise due to the residual charge fluctuation in it. The source and drain regions are made in self-alignment with the accumulation electrode, the separation gate, and the FPP. The shallow p-well under the sensing channel is completely depleted, and its potential is higher than that of the surface potential. Therefore, the source and drain regions are separated from each other.

10.4.2 Operation

The operation of the device is very different from that of a conventional MOS image sensor. In the MOS image sensor, photogenerated carriers stored in the diffused source region are destructively read out by switching the MOSFET on.

In contrast, the incident light that passes through the thin accumulation electrode generates signal electrons in the present pixel. The electrons are collected in the accumulation bulk channel. Subsequently, they are transferred into the sensing channel by applying a negative pulse to the accumulation electrode. The principle of a floating surface sensing transistor is a potential modulation of a floating surface P-type channel formed along a $Si/Si0_2$ interface above an n-type sensing channel.

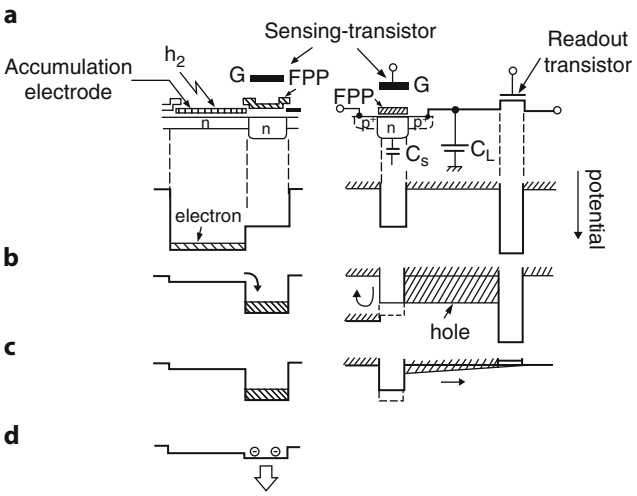

Fig. 10.33 Potential diagrams in LPM operation. (**a**) Photoconversion and accumulation. (**b**) Electron transfer and hole injection. (**c**) Amplified hole readout. (**d**) Signal electron reset

The sensing channel acts as a buried gate in the transistor. That is made possible by an abrupt potential profile of the detection gate in the surface region, shown in Fig. 10.31c. The n-type channel potential of the detection gate is higher than that of the accumulation gate, though the surface potential of the detection gate is lower than that of the accumulation gate. The potential profiles are realized by the deeper sensing channel structure mentioned above.

Figure 10.33 shows the potential diagrams in the line potential modulation (LPM) operation. The sensing transistor source is connected to a P-channel readout transistor and to the capacitance $C_L = 0.9\,\text{pF}$ corresponding to the vertical line capacitance. In Fig. 10.33a, the photogenerated signal electrons are stored under the accumulation electrode. In Fig. 10.33b, the signal electrons are transferred into a sensing channel and modulate the surface potential. Holes are injected into the line capacitance C_L through a modulated surface potential barrier by the fill-and-spill method. In Fig. 10.33c, the amplified holes in the line capacitance are read out by switching on the readout transistor. A significant charge gain Ag is obtained by the following equation:

$$Ag = C_L/C_S. \tag{10.3}$$

In Fig. 10.33d, the signal electrons are expelled into the substrate by applying a negative pulse on the potential control gate G. There is no thermal noise in the sensing transistor, because the sensing channel is completely depleted after the reset operation. In addition, the spatial separation of electrons and holes in the transistor enables us to operate the device in a nondestructive readout mode.

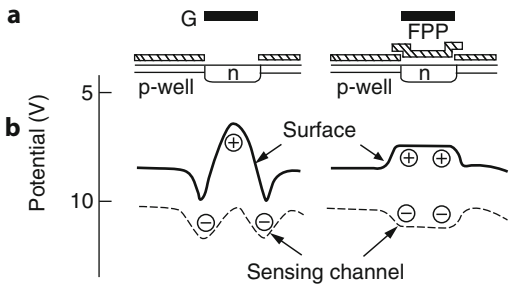

Fig. 10.34 Simulated potential profiles in the sensing transistor. (**a**) Sensing transistor with 1.0 μm-thick gate oxide. (**b**) Sensing transistor with FPP

10.4.3 Simulation

A two-dimensional device simulation was carried out to verify the FPP effect. For higher sensitivity, a lower sensing capacitance has been realized by adopting a 1.0 μm thick gate oxide for the detection transistor. Figure 10.34 shows simulated potential profiles. The solid lines are the surface potentials for holes along the Si/SiO_2 interface, and the dotted lines are those for electrons in the sensing channel. In Fig. 10.34a, potential pockets for electrons are observed at the edges of the accumulation and separation electrodes, where the sensing channel potential is significantly influenced by applied voltages on these electrodes. The potential pockets result in sensitivity degradation, because a small number of signal electrons are stored in these pockets and signal electrons cannot directly modulate the lowest surface potential. On the other hand, with the FPP being inserted under the potential control gate G, the pockets disappear as shown in Fig. 10.34b. The FPP flattens the channel potential in the detecting section through the thin gate oxide under it. Signal electrons directly change the lowest surface potential on which the holes are flowing.

10.4.4 Results

The sensing transistor and the readout circuit were evaluated at 7.16 MHz clock frequency. Figure 10.35 shows the experimental circuit. The signal gain of the sensing transistor was measured not by using the photogenerated electrons but by externally transferring a series of electron packets into a sensing channel from the accumulation region. Figure 10.36 shows applied input pulses and a resultant output signal waveform in the LPM operation. At the start of a reset-read cycle, the positive transistor pulse ($t1$) was applied to the accumulation electrode to receive test electrons in the accumulation channel through the adjacent separation gate. Under this condition, the P-type surface channel of the sensing transistor was pinched off. Subsequently, a negative transfer pulse ($t2$) was applied for transferring the test electrons into the sensing channel of the transistor. By applying a short positive

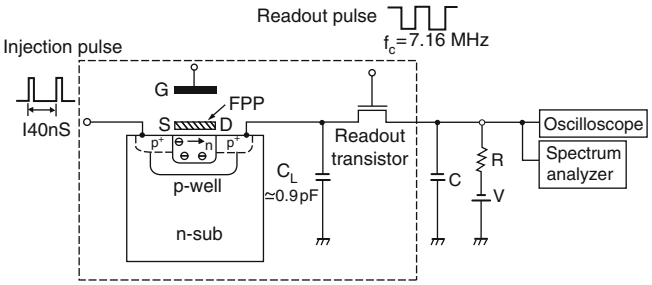

Fig. 10.35 Experimental circuit

injection pulse ($t3$) to the source, holes were injected into the line capacitance C_L. This set the capacitance voltage to the modulated surface potential by skimming the holes in the capacitor. The amplified charge (hole) was read out and removed by switching on the readout transistor with the readout pulse ($t4$). Refreshing the signal electrons was carried out by applying a negative pulse to the potential control gate G. The relation between the input test electrons and the amplified holes was linear over an input charge packet range of 20 to 4,500 electrons/packet with a conversion gain of 1,100, as expected from (10.3). The output noise was measured using a video noise meter. The overall rms noise value within the 10 kHz–3.58 MHz band was 880 holes/packet. Almost all noise is caused by charge fluctuation in the capacitance C_L due to the readout transistor channel thermal noise in the LPM operation. The thermal noise of the external amplifier is negligible. The signal gain of 1,100 and the noise of 880 holes correspond to an equivalent input noise of 0.8 electrons. The electric field in the gate oxide under the FPP is as small as 0.8 MV/cm. Therefore, the leakage current resulting in the FPP charge-up is too small to be measured. No operation point drift due to the FPP charge-up was observed during actual operation.

Figure 10.37 shows the photoresponse of the device, given as a function of illumination with the integration time of 16 ms for a standard NTSC field. The saturation exposure and amplified charge were found to be 1.1 lx and 5×10^{-6} holes/packet, respectively. The noise level exposure is calculated to be 2×10^{-4} lx from the rms noise of 800 holes measured at 7.16 MHz operation frequency. This results in an illumination dynamic range of 75 dB.

10.4.5 Applications of Area Sensor

The described phototransistor has several problems for area imager application. The most serious problem is fixed pattern noise (FPN) in the dark condition. The FPN is caused by dark current and by the detection transistor surface potential variations. The dark current is at the same level as in conventional CCD image sensor. The FPN due to the surface potential variations is suppressed by the clamping and sampling circuit introduced in reference [4]. The photoresponse nonuniformity

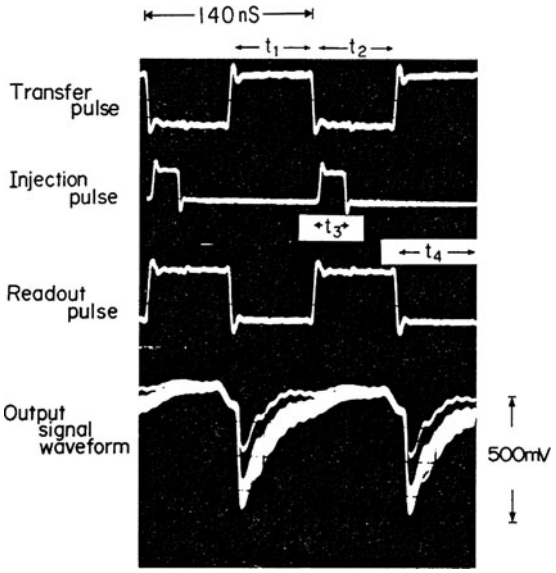

Fig. 10.36 Operating pulses and output signal waveform

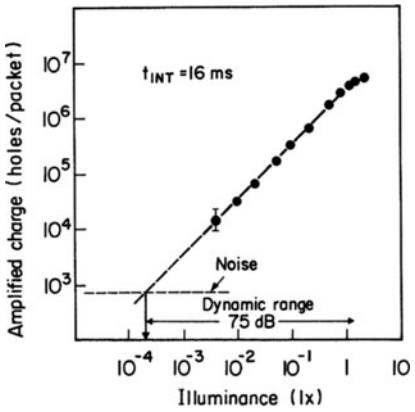

Fig. 10.37 Photoresponse under condition of 16 ms integration

(PRNU) resulting from the detection transistor gain variations will be less than ±5%, because the oxide thickness between the potential control gate G and the FPP, which determines the conversion gain, is flat within 5% variation in a 5-in wafer. Blooming is suppressed by the vertical overflow under the accumulation gate. Arranging this phototransistor in the transversal signal line (TSL) layout can sufficiently reduce the smear. To improve the signal-to-optical-shot-noise ratio, the saturation level can be raised by increasing local phosphorus implantation dose under the FPP.

10.5 Conclusions

Buriedchannel CCDs achieve virtually complete charge transfer with no transfer noise. One of the most effective means for the realization of single photon image sensors is to suppress noises of the CCD output charge detectors under the one-electron level. The double-gate charge detectors have achieved such subelectron noise performance. As a consequence, this structure can be applied to amplified MOS image pixel as chargemodulation configuration.

References

1. J. Hynecek, High-resolution 8-mm CCD image sensor with correlated clamp and hold charge detection circuit, IEEE Trans. Electron Devices **33**(6) 850–862 (1986)
2. R.J. Brewer A Low noise CCD output amplifier, IEDM Tech Dig 610–612 (1979)
3. R.J. Brewer, The low light level potential imaging array, IEEE Trans. Electron Devices **27** 401–405 (1980)
4. J. Hynecek, A new device architecture suitable for highresolution and high-performance image sensors, IEEE Trans. Electron Devices **35**(5) 646 (1988)
5. Y. Matsunaga, S. Oosawa, M. Iesaka, S. Manabe, N. Harada, N. Suzuki, A high sensitivity output amplifier for CCD image sensor, IEDM Tech Dig 116–119 (1987)
6. S. Oosawa et al., An electron counting detector for CCD image sensors, in *SSDM Extend Abstracts*, 1988 pp. 355–358
7. Y. Matsunaga et al., An electron counting detector for CCD image sensors, a highly sensitive on-chip charge detector for CCD area image sensor, IEEE J. SolidState Circuits, **26**(4) 652–655 (1991)
8. Y. Nishida et al, Design concept for a low-noise CCD image sensor based on subjective evaluation, IEEE Trans. Electron Devices **36**(2) 360–366 (1989)
9. K. Matsumoto, T. Nakamura, A. Yusa, S. Nagai, A new MOS image sensor operating in a non-destructive readout mode, Jpn. J. Appl. Phys. **24**(5) L323–L326 (1985)

Chapter 11
Energy-Sensitive Single-Photon X-ray and Particle Imaging

Christian Lotto

Abstract Energy-sensitive detectors perform asynchronous arrival detection of single X-Ray photons and particles. Their ability of measuring the detected particles' energy improves the performance of the particle counting applications and enables spectroscopic applications. In such detectors, either a semiconductor layer for direct conversion or a combination of a scintillator and a semiconductor sensing device for visible photons is used for generation of an electrical charge pulse per absorbed particle. This charge amount, which represents the particle energy, is detected by an asynchronous charge pulse detecting circuit. The noise of such circuits defines the lowest discrimination threshold of counting systems and the energy resolution of spectroscopic applications. Therefore, low noise, low power consumption, and low area are requirements for charge pulse detecting circuits used in segmented energy sensitive particle detectors with a high number of pixels. Choice of a sensing device, definition of the charge pulse detecting circuit's topology, and analysis of interdependences amongst the above performance parameters are covered and a context with employed readout schemes, processing circuits, and target applications is established in this chapter.

11.1 Introduction

Energy-sensitive single-photon or particle imaging systems are able to detect and count single high-energy photons or particles. In addition to detection such systems can also determine the energy of the detected photons or particles. Such imaging systems usually employ semiconductor detectors which sense electrical charge pulses created by the absorption of the photons or particles in an absorbing material. The particles or photons arrive erratically and the imaging system, therefore, needs to perform asynchronous detection of the resultant discrete charge pulses. The detection of these charge pulses gives only part of the required information, however, as the energy of each particle must also be measured. This energy measurement is performed by determining the amplitude of each of the charge

P. Seitz and A.J.P. Theuwissen (eds.), *Single-Photon Imaging*,
Springer Series in Optical Sciences 160, DOI 10.1007/978-3-642-18443-7_11,

pulses, hence the effective energy resolution depends on the accuracy of the charge measurement.

This chapter is organized as follows: The remainder of Sect. 11.1 gives key application examples and introduces the basic topology of energy-sensitive single-photon and particle imaging systems, Sect. 11.2 gives a brief overview of particle sensing devices, Sect. 11.3 describes various architectures of asynchronous charge pulse detecting circuits with an emphasis on noise analysis, and, finally, Sect. 11.4 covers the subjects of voltage pulse processing circuits and information readout schemes.

The quanta of electromagnetic radiation, the photons, can be considered to be particles with rest mass zero. For the sake of readability, the term particle will therefore include the particular case of X-ray photons, unless distinction is made explicitly.

11.1.1 Applications

Existing and emerging applications of energy-sensitive single-photon and particle imaging can be found in a variety of fields including security, material science and inspection, medical imaging, or particle physics research. A nonexhaustive selection of examples is given here.

Absorption X-ray imaging uses radiation sources with a relatively broad spectrum. X-ray photons transmitted through a sample are detected. In conventional detectors, optoelectronic charge generated by a high number of detected photons is integrated during an exposure time in order to produce a transmission intensity image. The integration process implicitly performs a weighting of the detected photons by their energy. As the probability of absorption in the sample generally decreases with increasing photon energy, weighting by energy results in a loss of contrast. Photon-counting detectors avoid weighting by photon energy and can, therefore, inherently achieve better contrast than charge-integrating detectors. Energy-sensitive detectors furthermore allow restriction of the counting to a selected energy window. Spectral ranges having poor absorption can, therefore, be discarded for further contrast enhancement. Generally speaking, energy-sensitive detectors allow selection of a narrow spectral range even if a wide-spectrum source is used. They may, therefore, allow use of X-ray tubes for applications which have previously been reserved for use with synchrotron sources.

The method of *X-ray absorption spectroscopy* determines the X-ray absorption properties of a sample as a function of the photon energy. For this purpose a radiation source of a tunable narrow spectrum is used. Photon counting in a narrow energy window achieves good precision of this spectroscopy method by suppressing detection in nonrelevant energy ranges.

In *X-ray fluorescence* experiments a sample is irradiated by primary X-ray photons of high energy. Subsequent electron excitation and relaxation results in emission of secondary X-ray photons of lower energy. The energy of secondary photons is a characteristic of a material's atomic and crystalline structure [1].

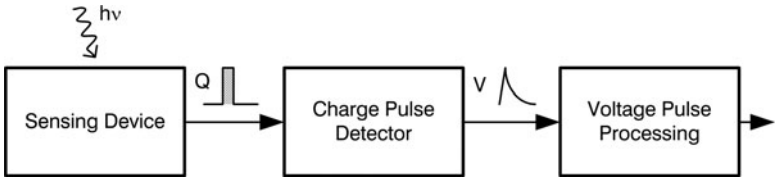

Fig. 11.1 Functional block diagram of a detector pixel for energy-sensitive particle imaging

Suppression of primary photon detection and energy discrimination of secondary photons are requirements that make energy-sensitive single-photon imagers ideal candidates for this application.

11.1.2 Basic Topology

Energy-sensitive X-ray and particle imagers are segmented detectors consisting of one- or two-dimensional arrays of detector channels or pixels. Because of the potentially high number of pixels the electrical power available per pixel is usually restricted. A single pixel circuit, as shown in Fig. 11.1, generally contains the following three building blocks:

Sensing device: A transduction device which absorbs particles in the target range of energies and produces an amount of charge that corresponds to the energy of the absorbed particle.

Electronic charge pulse detecting circuit: An asynchronous electronic circuit able to detect a charge pulse generated by the sensing device and to produce a voltage pulse of an amplitude proportional to the detected amount of charge at any time a charge pulse is created.

Voltage pulse processing circuit: An electronic circuit able to process information of incoming voltage pulses. Depending on the application, required functionalities include: peak amplitude detection, energy-windowed pulse counting or multithreshold binning, pulse width detection, and storage and readout of detected information.

11.2 Particle Sensing Devices

Sensing devices for energy-sensitive particle imaging need to meet a specific set of requirements. First of all, they need to be able to absorb particles of their application to a sufficient degree. In addition, the process of absorption and charge collection needs to take place in a period significantly shorter than the maximum particle hit rate of a detector pixel in order to avoid charge pile-up[1] from consecutive particle

[1]The term "pile-up" denotes accumulation of signal information of consecutive particle detections.

absorptions. Furthermore, as will be explained in Sect. 11.3, meeting the stringent specifications of power consumption and noise in charge pulse detecting circuits require low capacitance of the sensing device.

Sensing devices that fulfill the outlined requirements can be classified according to their principle of operation, i.e., direct conversion vs. scintillator-based, and by the type of their manufacturing and assembly, i.e., monolithic integration vs. hybrid manufacturing.

11.2.1 Direct Conversion Sensing Devices

Direct conversion sensing devices are based on mobile charge carrier generation by absorption of high-energy photons or multiple scattering of massive particles in semiconductor materials. Application of a bias voltage across a p-i-n conversion layer depletes the semiconductor and allows for virtually recombination-free separation of mobile electrons and holes generated by the ionizing events. The degree of absorption for a particle of a given energy and the resulting number of created electron–hole pairs generally depends on the thickness and material of the conversion layer. Silicon is a convenient conversion material, as it enables economic fabrication of sensing devices as well as detecting and processing circuits in the same material. However, the thickness of a silicon conversion layer for high particle energy can be quite high. Semiconductors with higher atomic number, such as cadmium telluride or germanium, can provide more efficient absorption properties. Thicknesses of conversion layers range from tens to several hundreds of micrometers and bias voltages for depletion may be as high as several kilovolts.

11.2.1.1 Monolithic Detectors

In the case of monolithic detectors, a direct conversion layer is fabricated on the same substrate as the detector electronics. The conversion layer may be a part of the substrate used for fabricating the electronics, which is either monocrystalline or epitaxially grown. This approach is limited in practice to silicon and germanium direct conversion layers. The alternative method of growing or depositing direct conversion layers on the detecting circuit substrate gives access to a wider choice of semiconductor materials.

11.2.1.2 Hybrid Detectors

Hybrid detectors consist of a direct conversion layer and an integrated detector circuit that are manufactured on two different substrates. The substrates are assembled using bump bonding or flip-chip technology with one bonding connection per pixel [2,3]. Manufacturing technology for the detector electronics may be chosen without any restrictions concerning the substrate characteristics. Besides a bond pad on the topmost metal layer, the entire area of a pixel may be covered by the electronic circuit.

11.2.2 Scintillators Coupled to Sensing Devices for Visible Light

Scintillator materials are luminescent in response to ionizing radiation such as massive particles or photons of sufficient energy. A scintillator flash, i.e., a pulse of photons of, in most cases, visible wavelength is produced as a result of an ionizing particle interacting with the scintillator. The exact mechanism of this process depends on the type of scintillator material and ionizing particle. The emission of visible-wavelength photons can occur within a fraction of a nanosecond in the case of a fluorescence effect. In the case of phosphorescence or delayed fluorescence, the scintillator flash consists of one or several superimposed decays of longer decay times of up to microsecond length. Depending on the application, suitable solid-state scintillator materials include inorganic and organic crystals, glasses, polymers, and powders.

A particle sensing device is created by optically coupling a scintillator layer to a suitable semiconductor sensing device for visible-wavelength photons emitted by the scintillator. Besides being sensitive to the corresponding wavelength, selection criteria for such a sensing device include short response time and low capacitance at a dimension matching the application. Sensing device dimensions for X-ray and particle imaging applications are typically larger than for conventional visible light imagers, e.g., several tens to a few hundred micrometers for medical X-ray imaging to date. The following subsections give a short overview of the most common visible-light sensing device types that may be used in conjunction with scintillators.

11.2.2.1 p–n Photodiodes

p–n photodiodes consist of a junction between a p- and n-type semiconductor to which a reverse bias voltage is applied. Mobile holes and electrons generated by absorption of a visible-wavelength photon are separated by the resulting electrical field in the depletion zone. The charge collection time of p–n diodes is generally negligible when compared to the width of the visible light peak provided by typical scintillators. Despite their good response time, p–n diodes suffer from relatively high sensing device capacitance as the latter is proportional to the sensing device area.

11.2.2.2 Photogates

Photogates are Metal[2]-Oxide-Semiconductor (MOS) devices in which the application of a bias voltage between the semiconductor and the metal electrode

[2]Note that, the top electrode of MOS devices is made from polycrystalline silicon in most manufacturing processes.

creates a depleted zone with an electrical field in the semiconductor where absorption-generated mobile holes and electrons are vertically separated. A p–n junction, biased at a lower electron potential energy than the semiconductor–oxide interface, is adjacent to the MOS structure. Charge carriers that have drifted vertically to the semiconductor–oxide interface diffuse laterally to this p–n junction called the sense node. The sensing device capacitance of a photogate corresponds to the sense node capacitance and is practically independent of the photogate MOS capacitance. Photogates, thus, have a sensing device capacitance which is area-independent and generally low. The response time of the conventional photogate is limited by the process of lateral diffusion transport and may be excessive for many applications.

Lateral drift field photogates are photogates with a voltage difference applied between two edges of the top electrode. The resulting electrical lateral field at the semiconductor–oxide interface causes drift transport of mobile charge toward the sense node [4]. This method allows overcoming the response time limitations experienced with conventional photogates.

11.2.2.3 Buried Photodiodes

Buried photodiodes[3] are, in many ways, similar to photogates. The fundamental difference being the creation of a depletion zone by the use of a vertical stack of p- and n-type implants in a p-type substrate instead of an MOS structure [5]. The vertical position of minimum electron potential energy, where generated electrons drift to, is buried at a certain distance below the semiconductor surface. An MOS-type transfer gate is available for clocked charge transfer from the buried photodiode to a p–n junction-type sense node. For asynchronous detection of charge pulses, the transfer gate is continuously active. Diffusion transport is responsible for lateral movement of charge carriers onto the sense node. The use of buried photodiodes is, therefore, restricted to applications where long sensing device response times and relatively small pixel areas are acceptable. Besides low sensing device capacitance, an important advantage of buried photodiodes is their availability in modern CMOS manufacturing processes that allow fabrication of high-performance charge pulse detecting circuits.

11.3 Asynchronous Charge Pulse Detecting Circuits

Every pixel of an energy-sensitive particle imager contains an analog circuit able to convert electronic charge pulses into voltage pulses. Destruction of optoelectronic charge after detection, commonly referred to as charge reset, is a further task performed by the detecting circuit. Asynchronous detecting circuits may have

[3] Buried photodiodes are also referred to as pinned photodiodes.

continuous, i.e., asynchronous, or clocked, i.e., synchronous, charge reset functionality. Clocked reset operation generally introduces a dead time, where signal charge pulses are destroyed without being detected. From the following analysis it can be understood that clocked reset is a particular case of continuous reset operation. Therefore, the analysis of this chapter only considers the case of continuous reset.

The requirements for charge-pulse detecting circuits suitable for energy-resolved particle detecting pixel circuits are summarized qualitatively in the following list:

- Low electronic noise
- Low power consumption
- Low semiconductor area
- High charge-to-voltage conversion factor
- High linearity of charge-to-voltage conversion

Sections 11.3.1–11.3.3 discuss three different topologies of charge-pulse detecting circuits used for energy-sensitive particle imaging.

11.3.1 Charge Sensitive Amplifier

The charge sensitive amplifier (CSA) is employed in many particle detectors, in particular for two-dimensional arrays with small pixel circuits [6]. Its invention dates back to the days of discrete circuits.

As illustrated in Fig. 11.2, a CSA consists of an operational transconductance amplifier (OTA), a feedback capacitor, a feedback resistor R_f in parallel with the feedback capacitor for continuous reset, and a load capacitance C_l on the output of the OTA. The sensing device is connected to the virtual ground node of the circuit. The overall capacitance C_{in} of the virtual ground node is approximately equal to the sum of sensing device capacitance and OTA input capacitance.

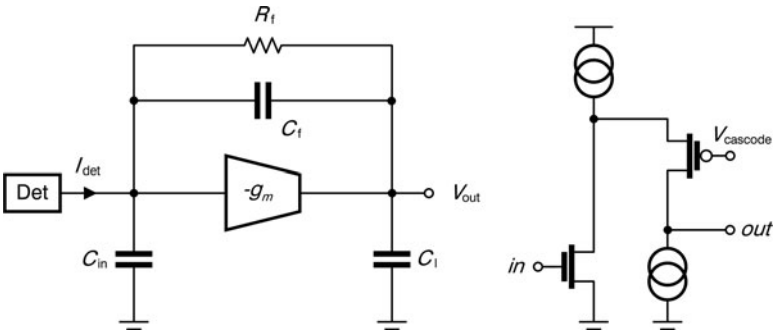

Fig. 11.2 *Left*: Circuit topology of a charge sensitive amplifier (CSA). *Right*: Single-branch folded cascade transconductance amplifier

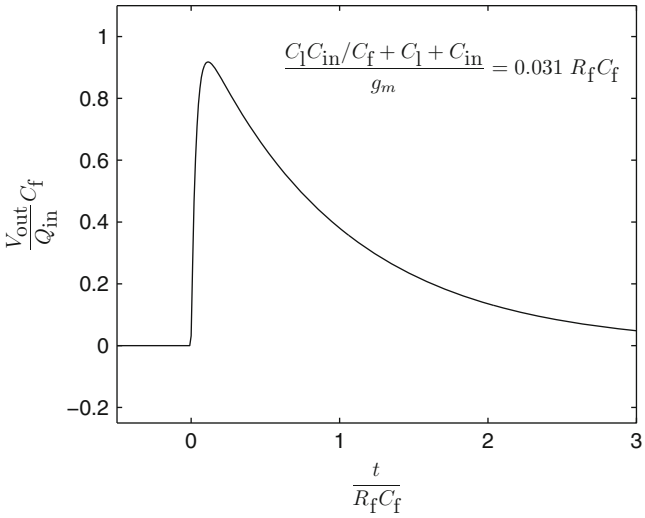

Fig. 11.3 Normalized CSA response to a Dirac input current pulse

Due to feedback operation, a pulse of optoelectronic charge is quickly transferred onto the feedback capacitor and creates a step of the output voltage, as illustrated in Fig. 11.3. The voltage difference across the feedback capacitor is slowly discharged by the feedback resistor. An asymptotic decay, therefore, follows an output voltage peak and brings the detector back to its equilibrium operating point.

The transimpedance function Z_{sig}, which relates the CSA output voltage V_{out} to sensing device current I_{det}, is given by (11.1) and illustrated in Fig. 11.4.

$$Z_{sig} = \frac{R_f\,(1 - sC_f/g_m)}{1 + s\left(R_f C_f + \frac{C_l + C_{in}}{g_m}\right) + s^2\,(C_l C_{in} + C_l C_f + C_{in} C_f)\,\frac{R_f}{g_m}}. \tag{11.1}$$

For sufficiently high transconductance of the OTA, the transimpedance function possesses real dominant and nondominant poles at $1/R_f C_f$ and $g_m/(C_l + C_{in} + C_l C_{in}/C_f)$, respectively. The effect of the high-frequency zero at g_m/C_f is usually negligible. The useful frequency span of a CSA is found between the dominant and the nondominant pole frequencies, where the transimpedance function takes an approximate value of $1/sC_f$. Signal charge pulses shorter than the CSA decay time of $R_f C_f$ are thus converted into output voltage pulses with a conversion factor of approximately the inverse feedback capacitance. The use of small feedback capacitance, therefore, achieves high output voltage for low input charge independently of the input capacitance. The rise time of the output voltage pulse is governed by the nondominant pole.

Low feedback capacitance and large load and input capacitances require high OTA transconductance in order to keep the nondominant pole frequency sufficiently high for the conversion factor to be governed by the feedback capacitance.

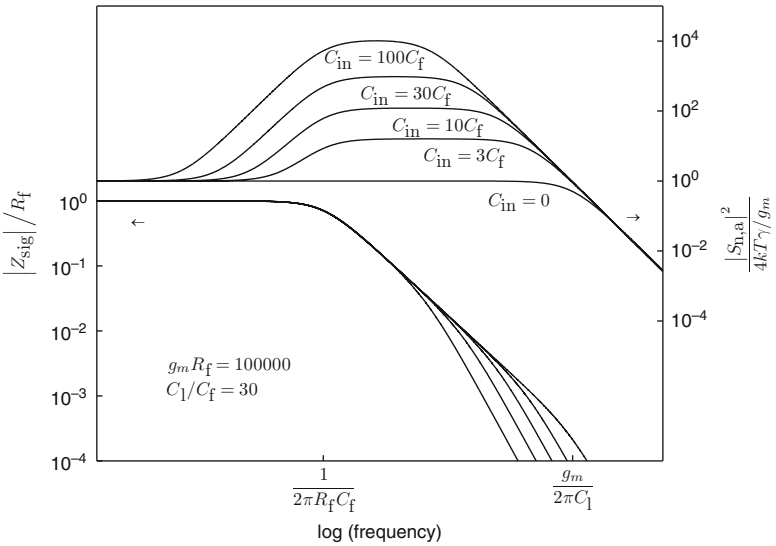

Fig. 11.4 Normalized signal transimpedance function and amplifier noise power spectral density of a CSA for several values of input capacitance

Accordingly, high OTA power consumption is the consequence of high sensing device capacitance.

Furthermore, the combination of high input capacitance and low feedback capacitance results in low values of the CSA's voltage feedback factor. Therefore, loss of conversion factor due to finite amplifier gain is observed if the OTA's open-loop voltage gain is not sufficiently high. This open-loop gain is defined by the product of the OTA transconductance and the OTA's output resistance in parallel with the feedback resistor. In case the feedback resistor is lower than the OTA's output resistance, undesired lowering of the open-loop gain can be avoided by inserting a source-follower type buffer between the OTA output and the feedback resistor [7]. However, very high voltage gain, i.e., high OTA output impedance, is still required in cases of sensing device capacitances several orders of magnitude larger than the feedback capacitance. For this reason, regulated cascode amplifiers have been suggested for such applications [8]. For more moderate ratios of input capacitance over feedback capacitance, single-branch folded cascode OTAs, as shown in Fig. 11.2, are good candidates thanks to their low power supply voltage requirement, decent output signal swing, and low parasitic capacitance in parallel with the feedback capacitor [9–11].

The noise of the detector circuit is a fundamental limit of the energy resolution and, therefore, needs careful consideration in the design of particle detectors. In the following analysis the noise generated in the amplifier[4] and in the feedback resistor

[4] Amplifier noise is also referred to as series noise in literature.

is investigated. Both noise sources are analyzed separately. Furthermore, Poisson noise of the sensing device leakage current is considered.[5]

11.3.1.1 Feedback Resistor Noise of CSAs

The principal noise source in the feedback resistor is thermal noise. The power spectral density (PSD) $S^2_{\mathrm{n,f}}$ of the voltage noise on the CSA output node contributed by the feedback resistor is given in (11.2):

$$S^2_{\mathrm{n,f}} = \frac{4kT\gamma R_{\mathrm{f}} \left|1 + s\frac{C_{\mathrm{in}}}{g_m}\right|^2}{\left|1 + s\left(R_{\mathrm{f}}C_{\mathrm{f}} + \frac{C_{\mathrm{l}}+C_{\mathrm{in}}}{g_m}\right) + s^2\left(C_{\mathrm{l}}C_{\mathrm{in}} + C_{\mathrm{l}}C_{\mathrm{f}} + C_{\mathrm{in}}C_{\mathrm{f}}\right)\frac{R_{\mathrm{f}}}{g_m}\right|^2}. \tag{11.2}$$

The thermal noise excess factor γ needs to be considered if the feedback resistor is actually implemented by use of an MOS transistor. Note that the poles of the feedback resistor noise PSD are equal to the poles of the signal transimpedance function. The frequency of the zero is always higher than the dominant pole frequency in the case of a CSA properly designed for reproducible and high conversion factor. Therefore, the equivalent noise charge (ENC), as given in (11.3), can be approximated by the integral of a first-order noise voltage PSD at a bandwidth defined by the dominant pole, divided by the ideal conversion factor.

$$q_{\mathrm{n,f}} \approx \sqrt{kT\gamma C_{\mathrm{f}}}. \tag{11.3}$$

Tuning of the feedback resistor has no influence on the ENC, because opposite effects are affecting the spectral density and the bandwidth. For a chosen value of feedback capacitance the feedback resistor noise is constant and defines a limit for the CSA noise performance.

11.3.1.2 Amplifier Noise of CSAs

The MOS-transistors of the OTA exhibit thermal noise and flicker noise. The analysis presented here is restricted to thermal noise, which is a valid assumption for a wide range of applications and manufacturing processes. The analysis of flicker noise can be extended from the analysis for thermal noise and is described in [12]. The major component of electronic noise in well-designed OTAs is contributed by the input transistors. The output noise PSD, as given by (11.4) and illustrated in Fig. 11.4, is found from a simplified analysis, neglecting the noise of the remaining transistors.

[5]Feedback resistor noise and sensing device leakage current shot noise are also referred to as parallel noise in literature.

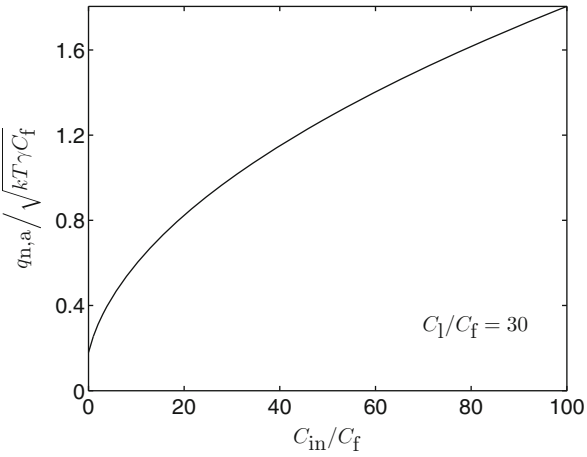

Fig. 11.5 Normalized ENC of the amplifier noise component vs. total input capacitance

$$S_{\mathrm{n,a}}^2 = \frac{(4kT\gamma/g_m)\left|1 + sR_{\mathrm{f}}\left(C_{\mathrm{in}} + C_{\mathrm{f}}\right)\right|^2}{\left|1 + s\left(R_{\mathrm{f}}C_{\mathrm{f}} + \frac{C_{\mathrm{l}}+C_{\mathrm{in}}}{g_m}\right) + s^2\left(C_{\mathrm{l}}C_{\mathrm{in}} + C_{\mathrm{l}}C_{\mathrm{f}} + C_{\mathrm{in}}C_{\mathrm{f}}\right)\frac{R_{\mathrm{f}}}{g_m}\right|^2}. \tag{11.4}$$

Besides two poles equal to those of the signal transimpedance function, a zero at $1/R_{\mathrm{f}}\left(C_{\mathrm{f}} + C_{\mathrm{in}}\right)$ causes peaking of the amplifier noise PSD in the useful frequency range. The effect increases with decreasing voltage feedback factor, i.e., with increasing $C_{\mathrm{in}}/C_{\mathrm{f}}$. At a given OTA transconductance, the noise bandwidth is reduced with increasing total input capacitance. Despite this partial compensation, the amplifier ENC increases with C_{in}, as illustrated for an example design in Fig. 11.5.

A simple approximation of the amplifier ENC is given by (11.5) and suggests a square-root dependence on C_{in}:

$$q_{\mathrm{n,a}} \approx \sqrt{kT\gamma\frac{C_{\mathrm{f}}C_{\mathrm{in}}}{C_{\mathrm{l}}}}. \tag{11.5}$$

This result illustrates that, as opposed to feedback resistor noise, no fundamental limit exists for amplifier noise. By choice of high load capacitance the amplifier noise can always be reduced to a level below feedback resistor noise. High load capacitance and high input capacitance requires, however, very high transconductance for keeping the nondominant pole at a sufficiently high value. Considering the implications on circuit area and power consumption, it is obvious that CSA-based highly segmented particle imagers with good energy resolution require low sensing device capacitance.

11.3.1.3 Sensing Device Leakage Current Poisson Noise in CSAs

Besides signal charge pulses, sensing devices generally contribute leakage current to the detector circuit. In particular, some direct conversion sensing devices suffer from relatively high leakage.

The leakage current's Poisson noise[6] power spectral density, of $2qI_{\text{leak}}$, is multiplied by the square of the signal transimpedance function, $|Z_{\text{sig}}|^2$, in order to obtain the output noise PSD. Besides a difference in high frequency zeros, the Poisson noise displays the same spectral shape as the feedback resistor noise, i.e., the noise power is confined to the low-frequency region below the dominant pole frequency of the CSA. The input referred noise charge due to sensing device leakage Poisson noise is described by the simple approximation of (11.6):

$$q_{\text{n,Poisson}} \approx \sqrt{\frac{1}{2} R_{\text{f}} C_{\text{f}} q I_{\text{leak}}}. \tag{11.6}$$

The impact of sensing device leakage Poisson noise is negligible if its current PSD $2qI_{\text{leak}}$ is significantly below the feedback resistor current PSD of $4kT\gamma/R_{\text{f}}$.

11.3.1.4 Sensing Device DC Current Compensation

DC input current flowing to the detector circuit consists of leakage current of the sensing device and signal current due to constant background light if not avoided in the operating setup. In a conventional CSA, of the type illustrated in Fig. 11.2, the described DC current components flow across the feedback resistor and create a DC shift in the CSA output voltage.[7] This can lead to poor performance of the subsequent voltage pulse processing circuit or, in extreme cases, saturation of the CSA output voltage.

A CSA with DC current compensation, as illustrated in Fig. 11.6, has been shown to be an effective solution to the problem [6, 13].

The functionality of the feedback resistor is implemented by the differential pair, since the small signal drain current of the left-side MOS transistor, injected into the left-hand plate of the feedback capacitor, is proportional to the small-signal output voltage.

A low-frequency high-gain feedback path is provided by the right-side drain-current output of the differential pair, a current source, a compensation capacitor C_{comp}, and a common-source transistor M_{CS}. Thanks to its high transconductance, the resulting two-stage transconductance amplifier can deliver significant DC current to the sensing device at very low changes of its input voltage, i.e., the CSA output voltage. The compensation capacitor is chosen such that the bandwidth

[6] Poisson noise is also referred to as shot noise.

[7] In literature the DC shift of the output signal is also referred to as baseline shift.

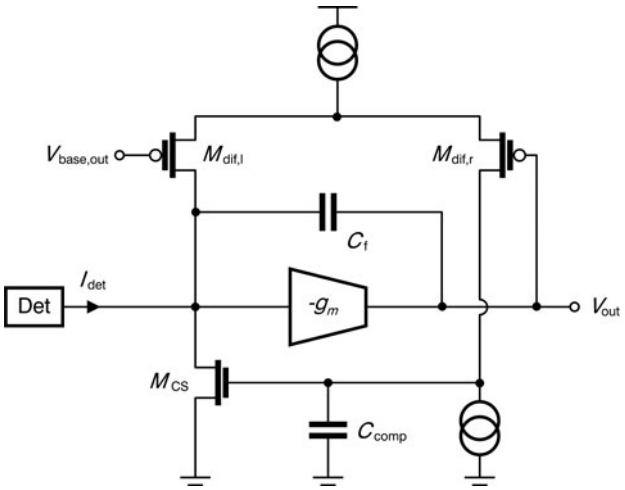

Fig. 11.6 CSA with sensing device DC current compensation

of the high-gain feedback amplifier is orders of magnitude below the dominant pole frequency of the CSA. Therefore, the feedback amplifier does not influence the detection and processing of the charge pulses. There is, however, additional noise associated with the feedback amplifier. Thermal noise of the common-source transistor M_{CS} is the predominant component of this noise, and it is of the same order of magnitude as the noise of the equivalent feedback resistor. Note that the maximum compensated positive and negative DC currents are not equal. The circuit, as shown in Fig. 11.6 is particularly suited for compensating positive current. For negative DC current, the complementary circuit made from p-type transistors is advantageous.

11.3.2 Charge Sensitive Amplifier with Shaper

The noise performance of CSAs is limited by thermal noise of the feedback resistor. The feedback resistor's noise power is confined to the spectral range below the dominant pole at $1/R_fC_f$, as stated by (11.2). The useful frequency span of the circuit and the relevant content of the signal is found above the dominant pole of the CSA. Charge sensitive amplifiers with shapers make use of this spectral separation in order to reduce feedback resistor noise by filtering without compromising the signal [10, 14–17].

Although the basic topology of a CSA used in conjunction with a shaper circuit, as illustrated in Fig. 11.7, is identical to the topology of a nonshaped CSA the design criteria for minimized noise are somewhat different.

The shaper circuit generally is a band-pass filter with a high-pass filtering frequency of f_{hp} and a low-pass filtering frequency of f_{lp}. Shapers are most usually

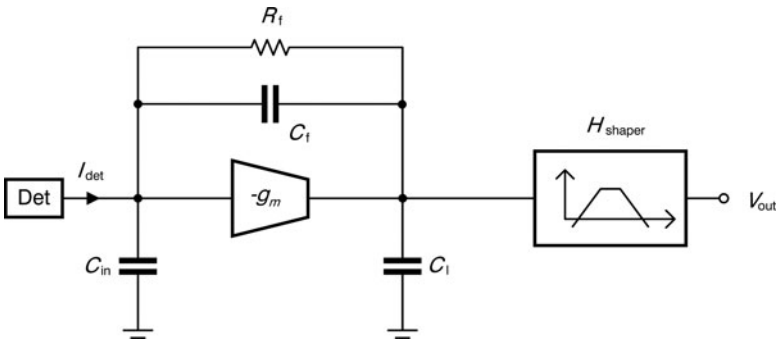

Fig. 11.7 Topology of a CSA-shaper circuit

CR-RC^n filters, i.e., a combination of a first-order high-pass function and an nth order[8] low-pass function. The generic shaper transfer function is thus given by (11.7) and illustrated in Fig. 11.8 for the example of a first-order shaper with unity gain:

$$H_{\text{shaper}} = A_{\text{shaper}} \frac{s \Big/ \left(2\pi f_{\text{hp}}\right)}{\left(1 + s \Big/ \left(2\pi f_{\text{hp}}\right)\right)\left(1 + s \Big/ \left(2\pi f_{\text{lp}}\right)\right)^n}. \tag{11.7}$$

The resulting signal transimpedance function of a CSA-shaper circuit is given by (11.8).

$$Z_{\text{sig}} = \frac{H_{\text{shaper}}\left(R_{\text{f}}\left(1 - sC_{\text{f}}/g_m\right)\right)}{1 + s\left(R_{\text{f}}C_{\text{f}} + \frac{C_{\text{l}}+C_{\text{in}}}{g_m}\right) + s^2\left(C_{\text{l}}C_{\text{in}} + C_{\text{l}}C_{\text{f}} + C_{\text{in}}C_{\text{f}}\right)\frac{R_{\text{f}}}{g_m}}. \tag{11.8}$$

The useful frequency span, where the signal transimpedance function is approximately equal to $1/sC_{\text{f}}$, is now defined by the pass-band of the shaper circuit. The transimpedance function corresponds to the output signal spectrum of the CSA-shaper circuit, stimulated by a Dirac unity current pulse from the sensing device. The shaped pulse at the output of a CSA-shaper circuit will, therefore, display a decay time governed by f_{hp} that will generally be shorter than the decay time of the CSA. Thanks to the use of the shaper, pulses occurring within a short time interval will thus be discriminated more easily by the voltage pulse processing circuit. The rise time of the output pulse is governed by f_{lp}.

The loss in amplitude of the voltage pulse on the shaper output compared to the CSA output[9] is called the ballistic deficit of the detector. It is a function of

[8] n is also referred to as the shaper order.

[9] A ballistic deficit may even be observed on the CSA output if the spacing between the nondominant and dominant poles is poor or if the response time of the sensing device is longer than the reset time constant.

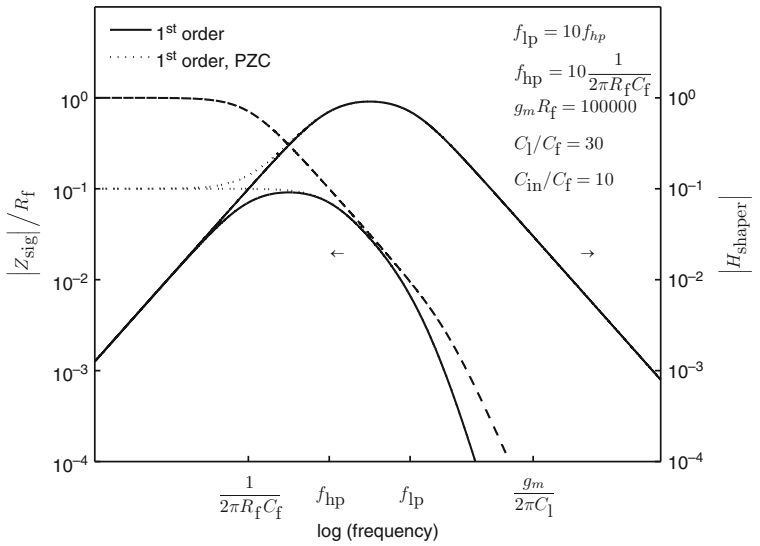

Fig. 11.8 Normalized shaper transfer function and shaped signal transimpedance function in case of first-order shaping with pole-zero cancellation(PZC) (*dotted lines*) and without PZC (*solid lines*). Signal transimpedance function without shaping (*dashed line*) for reference

the deviation of the transimpedance function in its useful range from the ideal capacitance impedance $1/sC_{\mathrm{f}}$. A significant ballistic deficit can be caused by too narrow a bandwidth of the shaper filter, low spectral spacing between the dominant pole frequency and f_{hp}, or low spectral spacing between f_{lp} and the nondominant pole frequency of the CSA.

As the output signal spectrum has a peak at nonzero frequency, the corresponding time-domain response displays some harmonic content, i.e., a signal undershoot, as illustrated in Fig. 11.9. If the signal undershoot would disturb correct operation of the voltage pulse processing circuit, the method of pole-zero cancellation (PZC) can be employed [7, 16, 18–20]. The basic concept of the method is to compensate the dominant pole of the CSA by a matched zero in the shaper transfer function, as given by (11.9). The resulting signal transimpedance function, and hence the output signal spectrum, therefore has a band-limited flat spectrum, as illustrated in Fig. 11.8. The corresponding time-domain response is a well-damped undershoot-free output signal.

$$H_{\mathrm{shaper,PZC}} = \frac{A_{\mathrm{shaper}}\, 2\pi R_{\mathrm{f}} C_{\mathrm{f}} f_{\mathrm{hp}}\ (1 + sR_{\mathrm{f}} C_{\mathrm{f}})}{\left(1 + s \Big/ \left(2\pi f_{\mathrm{hp}}\right)\right)\left(1 + s \Big/ \left(2\pi f_{\mathrm{lp}}\right)\right)^n}. \tag{11.9}$$

On one hand, a shaper circuit reduces feedback resistor noise and, under some conditions, amplifier noise of a CSA, as discussed in the following sections. On the other hand, a shaper circuit also generates additional noise that should be taken

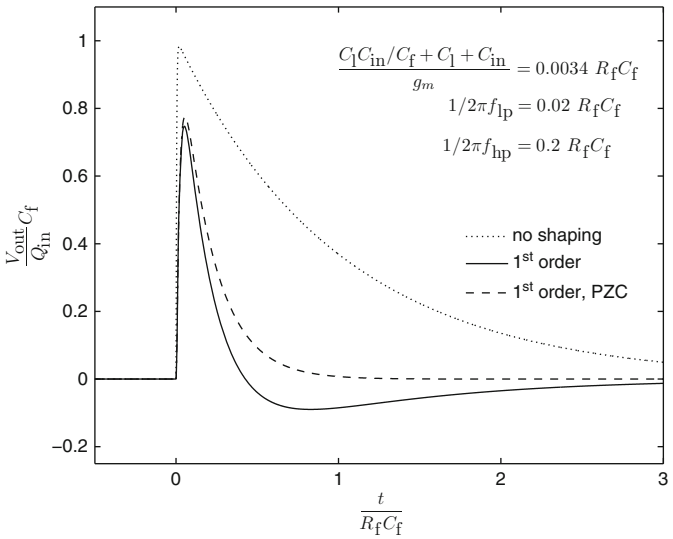

Fig. 11.9 Normalized CSA-shaper output response to a Dirac input current pulse with PZC (*dashed line*) and without PZC (*solid line*). CSA response (*dotted line*) for reference

into account when evaluating the overall performance of the charge pulse detecting circuit. The amplitude of shaper circuit noise depends on many properties and parameters of the shaper. Noise analysis for various shaper architectures can be found in literature, e.g., in [16]. The analysis presented here will be limited to the noise components of the CSA.

11.3.2.1 Feedback Resistor Noise of CSA-Shaper Circuits

A shaper circuit can reduce the input referred noise charge of the feedback resistor below the $\sqrt{kTC_f}$ limit found for CSAs without shapers. The feedback resistor noise power spectral density, as illustrated in Fig. 11.10, corresponds to the product of the noise power spectrum at the CSA output (11.2) and the square of the amplitude of the shaper transfer function (11.7).

For a shaper high-pass frequency significantly higher than the frequency of the CSA dominant pole, shaping has two major effects on the feedback resistor noise spectrum. First of all, the relevant bandwidth of the feedback resistor noise is increased from the dominant pole frequency to the high-pass filtering frequency. Secondly, the relevant noise power spectral density is attenuated by the square ratio of high-pass filter frequency over the dominant pole frequency. The input referred noise charge is, therefore, reduced approximately by the square root of the mentioned frequency ratio, as expressed by (11.10):

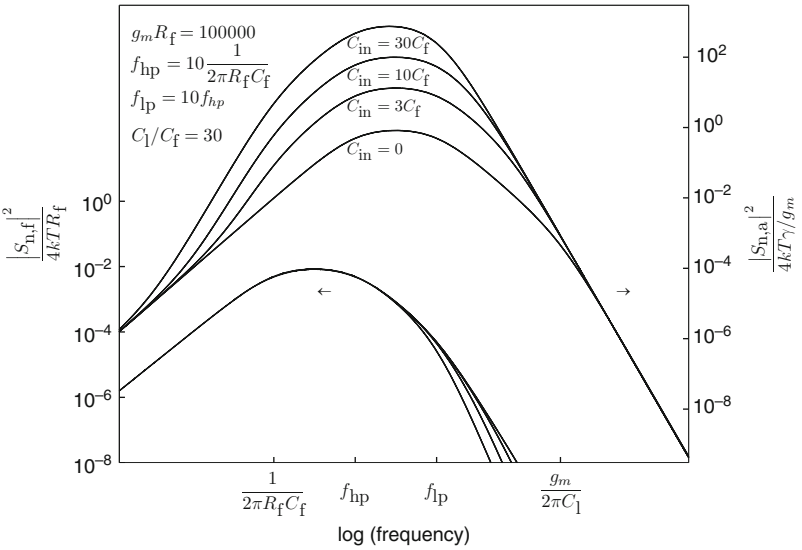

Fig. 11.10 Normalized output noise power spectral densities of the feedback resistor noise and the amplifier noise of a CSA with a first-order shaper

$$q_{\mathrm{n,f,shaped}} \approx \sqrt{kT\gamma C_{\mathrm{f}} \frac{1/\left(2\pi R_{\mathrm{f}} C_{\mathrm{f}}\right)}{f_{\mathrm{hp}}}} = \sqrt{\frac{kT\gamma}{2\pi R_{\mathrm{f}} f_{\mathrm{hp}}}}. \tag{11.10}$$

It becomes apparent that a high frequency spacing between the CSA dominant pole and the high-pass filter frequency is a key factor for low-noise CSA-shaper detectors. It has to be considered that the maximum possible value of the high-pass filter frequency is constrained by the finite width of charge pulses delivered by sensing devices, as well as by power consumption, which increases with required circuit bandwidth. Therefore, the reduction of feedback resistor noise requires reduction of the dominant pole frequency, i.e., increase of $R_{\mathrm{f}}C_{\mathrm{f}}$. Although an increase of the feedback capacitor lowers the output noise voltage, no reduction of the input referred noise charge is observed, due to the corresponding decrease of conversion factor. For these reasons, noise optimization of CSAs used in combination with shaper circuits involves the design of very high feedback resistance, in particular if a high conversion factor is achieved by use of low feedback capacitance.

As outlined in Sect. 11.3.1, the spectrum at the CSA output created by sensing device leakage Poisson noise is very similar to the feedback resistor noise spectrum. The analysis of feedback resistor noise in CSA-shaper circuits can, therefore, easily be extended to sensing device leakage Poisson noise and yields an input referred noise charge equal to the square root of $qI_{\mathrm{leak}}/(4\pi f_{\mathrm{hp}})$.

A pole-zero-canceling shaper offers less noise attenuation in the frequency range below the zero's frequency than a shaper without PZC. The PZC-shaped noise

spectrum is, therefore, flat in the frequency range below the zero, instead of having a first-order roll-off. The noise increase due to PZC, which is already included in the approximation of (11.10), is relatively low in the case of well-designed systems with a high ratio of high-pass frequency over dominant pole frequency.

11.3.2.2 Amplifier Noise of CSA-Shaper Circuits

Besides the significant reduction of feedback resistor noise, the use of a shaper circuit has an impact on amplifier noise as well [21]. The amplifier noise PSD at the shaper output is given by the product of the square amplitude of the shaper transfer function (11.7) and the amplifier noise power spectrum at the CSA output (11.4). The resulting amplifier noise spectrum, as illustrated in Fig. 11.10, shows that shaping does not change the relevant noise power spectral density. Unlike in the case of CSAs without shaper, the noise bandwidth is now limited to a frequency generally below the nondominant pole frequency of the CSA. In the case of a first-order shaper, the amplifier noise component of the ENC can be expressed by the simple approximation of (11.11):

$$q_{\mathrm{n,a,shaped}} \approx (C_{\mathrm{f}} + C_{\mathrm{in}}) \sqrt{kT\gamma \frac{2\pi f_{\mathrm{lp}}}{g_m}}. \tag{11.11}$$

A comparison between the ENC calculated by considering the full spectrum and approximation of the full spectrum (11.11) is given in Fig. 11.11 for a constant OTA transconductance. The amplifier ENC deviates from the simple approximation, as the nondominant pole moves closer to the low-pass filtering frequency for increasing input capacitance.

For CSA-shaper detectors the ENC is reduced for increasing transconductance, because the noise PSD is reduced while the noise bandwidth is kept constant by the shaper. However, increased power consumption and wider input MOS-transistors are required in OTAs with increased transconductance. This results in a higher total input capacitance C_{in} that, in turn, leads to increased amplifier ENC. Increasing the current density of the OTA input transistor is often not a good option, since the thermal noise excess factor γ is increased when going from moderate to strong inversion operation region. An optimum trade-off between increased g_m and increased C_{in} is found as a function of the remainder of the total input capacitance, i.e., the sensing device capacitance [22, 23].

Despite the outlined limitation, the possibility of trading power for amplifier noise without increasing C_{l} is a valuable property of CSA-shaper circuits. Nevertheless, the need for low input capacitance for low-noise, low-power performance, as already found in the case of the CSA, equally applies to the CSA-shaper detector circuit.

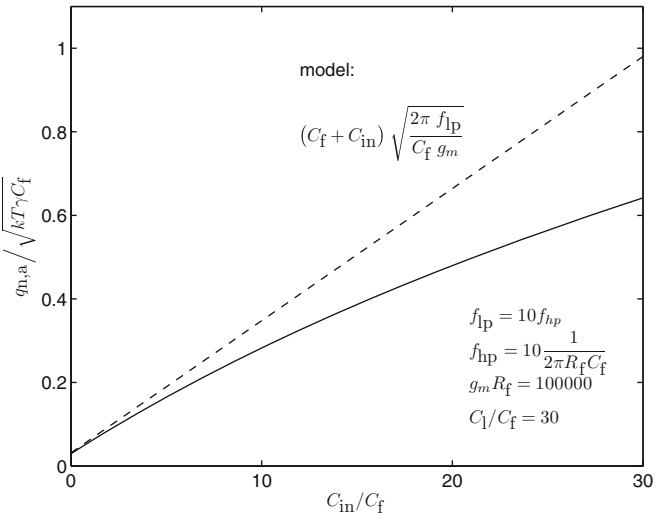

Fig. 11.11 Normalized ENC of the amplifier noise component vs. input capacitance of the CSA with first-order shaping (*solid line*) and simplified model of the amplifier noise ENC (*dashed line*)

11.3.2.3 Feedback Resistor Implementations

The use of very high feedback resistance achieves excellent noise performance in CSA-shaper detectors. In most cases the required value of resistance cannot be practically implemented by the use of linear resistor elements available in CMOS process technologies, such as polysilicon resistors. MOS transistors are, therefore, used for the implementation of feedback resistors in most cases.

The nonlinearity of the resulting resistance causes distortion of the output pulse. If the distortion affects the peak amplitude, a nonlinearity of the detected particle energy is observed.

Feedback resistance decreasing with signal pulse amplitude generally results in an increase of the ballistic deficit, because the dominant pole is temporarily shifted closer to the high-pass filter frequency of the shaper.

If the feedback resistance increases with pulse amplitude, the ballistic deficit is reduced for high pulses. In this case, very low energy nonlinearity can be achieved if the nominal ballistic deficit at low energy is designed to be very low. Conversely, pile-up of signal charge at high detection rates is more likely to occur for such a feedback resistor.

Two examples of feedback resistor implementations using MOS transistors, as illustrated in Fig. 11.12, are introduced and briefly discussed in the following paragraphs.

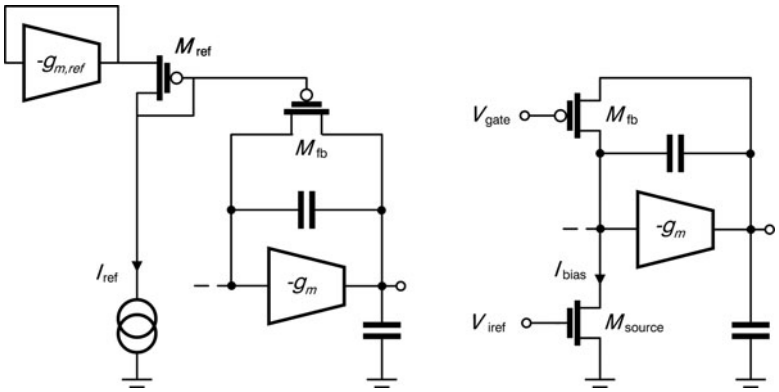

Fig. 11.12 Implementations of feedback resistors. *Left*: Feedback transistor in triode region. *Right*: Common-source transistor in weak inversion region

Feedback Transistor in Triode Region

An MOS transistor connected in parallel with the feedback capacitance, operated in triode region, is used in many applications [9, 11]. When using constant gate bias, threshold voltage variation of the feedback transistor M_{fb} and offset of the OTA lead to significant resistance variations that can be reduced by a tracking gate biasing circuit, as suggested in [7] and shown in Fig. 11.12. Nevertheless, the nominal gate bias excess voltage needs to be chosen with a certain margin in order to avoid high variation of the transistor's on-resistance due to the remaining mismatch between the biasing circuit and the feedback transistor plus the OTA. Practically achievable values of feedback resistance are, therefore, limited despite using a transistor with a long and narrow geometry.

A feedback transistor in triode region shows nonlinear behavior. Consequently increasing pulse height lowers the on-resistance in the case of electron detection and might, therefore, increase the ballistic deficit. If the charge carriers delivered by the sensing device are holes and the feedback transistor is a p-channel device, the on-resistance increases for high signal pulses. In this case the n-well of this transistor can be connected to the CSA output in order to improve linearity. The diode formed between the n-well and the transistor drain connected to the CSA virtual ground node will be increasingly forward biased for high negative output voltage pulses and will compensate for the increase of the MOS transistors on-resistance [19].

Common Gate Stage in Weak Inversion Region

A common gate transistor operated in the weak inversion region, such as illustrated in Fig. 11.12, can achieve very high effective feedback resistance [8]. If the gate bias voltage V_{gate} is chosen sufficiently high compared to the input voltage of the OTA,

the feedback resistance, i.e., the transconductance of the common-gate transistor, is controlled by the bias current source independently of OTA input offset.

In the described configuration the quiescent output level is different from the virtual ground potential and is set by the bias current I_{bias} as well as the gate voltage of the common-gate feedback transistor M_{fb}.

The transconductance of the common gate transistor increases with the output pulse amplitude. Quick return to the baseline and high maximum detection rate are, therefore, achieved at the price of possibly distorted energy detection.

As the additional noise of the bias transistor M_{source} has the same order of magnitude as the common-gate transistor noise, this component should be considered as well for the analysis of the feedback resistor noise.

11.3.3 Voltage Buffer with Shaper

As an alternative to CSA-shaper detector circuits, buffer-shaper circuits, are excellent candidates for highly segmented energy-sensitive particle detectors. The buffer-shaper topology allows very simple implementations with low power consumption and low area expense. Furthermore, excellent noise performance, i.e., very high energy resolution, can be achieved [24, 25].

For this circuit topology the sensing device is connected to the input of an open-loop unity-gain buffer, such as a source-follower circuit. The output of the buffer is filtered by a shaper circuit, very much like in the case of CSA-shaper detectors. As the buffer circuit is a unity gain amplifier without any feedback network, signal charge pulses are converted to voltage steps on the input capacitance C_{in}. A reset resistor R_{r}continuously removes charge from the input capacitance during a slow decay following a voltage step.

Figure 11.13 shows an example of a buffer detector circuit with a first-order shaper without PZC. The shaper implementation is a cascade of a passive CR high-pass-filter and an active low-pass filter consisting of a source-follower with a load capacitor.

The signal transimpedance, as given in (11.12), is found under the assumption of exact unity gain of the source followers as well as neglecting parasitic capacitance of the buffer output and the intermediate node of the shaper circuit. The high-pass filtering frequency f_{hp} of the shown implementation is equal to $1/2\pi R_{\text{hp}}C_{\text{hp}}$, whereas the low-pass filtering frequency f_{lp} corresponds to $g_{m,\text{lp}}/2\pi C_{\text{lp}}$. The signal transimpedance takes an asymptotic value of $1/sC_{\text{in}}$, which illustrates that the conversion factor is governed by the input capacitance.

$$Z_{\text{sig}} \approx \frac{R_{\text{r}}}{1 + sC_{\text{in}}R_{\text{r}}} \frac{s/2\pi f_{\text{hp}}}{\left(1 + s/2\pi f_{\text{hp}}\right)\left(1 + s/2\pi f_{\text{lp}}\right)}. \tag{11.12}$$

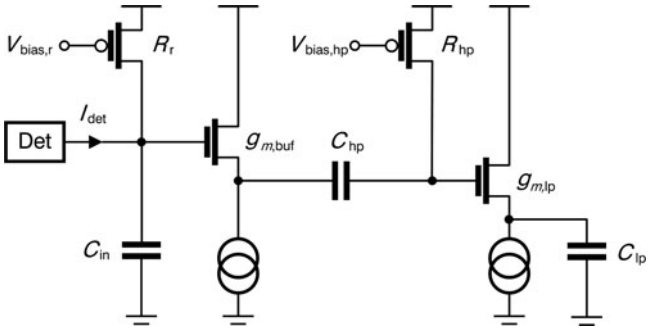

Fig. 11.13 Implementation example of a buffer-shaper detector circuit with first-order shaping

11.3.3.1 Noise Analysis of Buffer-Shaper Circuits

The reset resistor is implemented as an MOS transistor with a constant gate–source bias voltage. As opposed to feedback resistor biasing of CSAs, the biasing of such a reset resistor is not affected by offsets and variations of the buffer circuit. The biasing of the recharge device is, therefore, simplified and the reduced resistance variation allows for choosing of higher nominal values of resistance. In this configuration the recharge device can be operated in triode region, but even weak inversion operation is possible without a bias current flowing across the recharge device.

The expression of the ENC due to feedback resistor thermal noise, as given by (11.13), is analogous to (11.10). The reset resistor noise of buffer-shaper circuits can ultimately be lower than the feedback resistor noise of CSA-shaper circuits because of the fact that higher resistance values can practically be achieved without any need for a bias current source that generates additional noise.

$$q_{\mathrm{n,r,shaped}} \approx \sqrt{kT\frac{1}{2\pi R_{\mathrm{r}} f_{\mathrm{hp}}}}. \tag{11.13}$$

The resistance implementation, as depicted in Fig. 11.13, displays a maximum current slew for large pulses of input charge. On one hand, the corresponding increase of the recharge resistance results in low ballistic deficit for high pulses, i.e., low pulse height distortion. On the other hand, limited detection rate has to be accepted.

The ENC due to the buffer thermal noise is given by (11.14). Note that the expression is similar to the amplifier noise of CSA-shaper circuits (11.11), despite the fundamental difference of the topologies. In the case of buffer-shaper detectors, the input capacitance has no influence on the noise voltage PSD. The important impact of C_{in} on the ENC is, however, explained by the dependence of the conversion factor.

$$q_{\mathrm{n,b,shaped}} \approx C_{\mathrm{in}} \sqrt{kT\gamma \frac{2\pi f_{\mathrm{lp}}}{g_{m,\mathrm{buf}}}}. \tag{11.14}$$

For an identical given value of transconductance, the input capacitance of a buffer circuit is significantly lower than the input capacitance of an OTA. The input capacitance of an OTA mainly consists of the sum of gate–source capacitance and gate–drain capacitance of the input transistor. The input capacitance of a source-follower, however, is not much higher than the gate–drain capacitance of the input transistor, because the impact of gate–source capacitance is greatly reduced by the effect of noninverting near-unity gain. The low ratio of total input capacitance C_{in} over buffer transconductance $g_{m,\mathrm{buf}}$ results in a significant noise advantage of buffer-shaper circuits over CSA-shaper circuits at their respective optimum transconductance and gate dimensions. Furthermore for low total input capacitance C_{in}, of the order of magnitude of the CSA feedback capacitance C_{f} or lower, a buffer-shaper benefits from a systematic advantage over a CSA-shaper, which is illustrated by comparing the capacitance terms in the noise expression of (11.14) and (11.11).

The ENC contribution of a shaper circuit with a given input referred noise voltage is inversely proportional to the conversion factor. If the advantageous buffer noise and reset resistor noise performance is not to be compromised by self-generated noise of the shaper, a high conversion factor is required. As the conversion factor of buffer-shaper detectors is governed by the total input capacitance, low sensing device capacitance is a key parameter. This requirement is, however, in harmony with the general requirements of low-power, low-noise particle detectors. Buffer-shaper circuits with photogate sensing devices and total input capacitance of <10 fF currently reach an equivalent noise charge as low as 10–20 electrons [24].

11.4 Voltage Pulse Processing Circuits

Voltage pulse processing circuits are used in order to extract relevant information from the output signal of the charge pulse detecting circuit. Depending on the application, extracted information comprises pulse arrival time, pulse detection rate, or spectroscopic energy information. Voltage pulse processing circuits are generally part of each pixel in a segmented detector with further processing hardware possibly being shared among several pixels.

Readout circuits selectively transfer information from pixel circuits to shared processing hardware. Given the erratic nature of the particle arrival and the potentially huge amount of information collected by a segmented energy-sensitive detector, pixel-level data storage and data readout are a challenging aspect of systems design.

Voltage processing circuits are generally quite specific for the application of the detector. Classification is possible according to the applied method of energy discrimination, as discussed in Sect. 11.4.1. Different approaches to data readout are discussed in Sect. 11.4.2.

11.4.1 Energy Discrimination Methods

Voltage pulse processing circuits for energy-sensitive particle imaging can be classified into two general groups. Counting circuits with energy criteria, such as single or multiple threshold discriminating counters, retrieve detection rate information. Spectroscopic detectors, such as amplitude detectors or time-over-threshold detectors, generally provide more precise particle energy information, whereas detection rate information is often discarded.

11.4.1.1 Multiple Threshold Discrimination

Multiple threshold discriminators are used to count particle arrivals in one or several energy bins. Several comparators with different thresholds and one or more digital counters are used to implement such a processing circuit [6]. The output of the pixel circuit is, therefore, digital with the binning precision being a function of the charge pulse detector circuit's noise performance.

11.4.1.2 Amplitude Detection

Amplitude detection methods aim at precise particle energy measurement for spectroscopic applications. Analog amplitude detecting or peak detecting circuits are used to detect and hold the peak value of the voltage pulses from the charge pulse detecting circuit. Asynchronous peak detecting [9] or oversampled synchronous peak detecting circuits [17] can be used for this purpose. The energy resolution of the amplitude detection directly depends on the noise of the charge pulse detecting circuit.

11.4.1.3 Time-Over-Threshold Detection

Time-over-threshold detection is an alternative spectroscopic method based on the fact that the absolute pulse width above a constant threshold is proportional to the pulse height, i.e., the input charge for a linear charge pulse detecting circuit. A clock counter can be used in order to generate a digital output that represents the time-over-threshold [26]. This method requires linear charge pulse processing circuits, as nonlinear feedback or reset resistors could heavily distort the time-over-threshold.

11.4.2 Information Readout

Pulse counting applications usually involve counting pulses over a defined exposure interval. Furthermore, digital storage of the count numbers in the pixel circuits is

possible. Synchronous readout is, therefore, an adequate solution of low complexity for counting applications. The asynchronous nature of pulse arrival and the difficulty of storing energy information for many detected particles in a pixel makes readout of spectroscopic information more challenging. Synchronous readout usually involves discarding a considerable number of detected particles. Self-triggered readout is an interesting, but more complex readout method for spectroscopic voltage pulse processing circuits.

11.4.2.1 Synchronous Readout

Synchronous readout involves information storage in each detector pixel as well as sequential readout and reset of information using synchronous timing. Each readout circuit is shared and sequentially used by many detector circuits. For spectroscopic energy discrimination, which generates either digital multi-bit or analog information for each detected particle, the amount of information stored in a pixel is usually restricted to one detection. If a synchronous readout scheme is to be used, the information corresponding to the first or last particle arrival in the observation period is usually read out. Information of remaining detections between two consecutive readout operations is, therefore, discarded.

11.4.2.2 Self-Triggered Readout

In segmented detectors with self-triggered readout, readout circuits are shared row wise or column wise by several detector circuits, as in the case of synchronous readout. The voltage pulse processing circuits are able to deliver a readout trigger signal upon detection of a particle. The trigger signal is used to select the relevant detector circuit and read its information as soon as the corresponding readout circuit is available. Self-triggered readout is not sequential which makes the storage and further processing of the read information more complex. As opposed to synchronous readout, self-triggered readout avoids reading of pixels where no detection has taken place and naturally allocates the readout bandwidth to pixels with nonzero detection rates [27].

References

1. L.S. Birks, E.J. Brooks, H. Friedman, Anal. Chem. **25**(5), 692 (1953)
2. B. Mikulec, M. Campbell, E. Heijne, X. Llopart, L. Tlustos, Nucl. Instr. Meth. A **511**, 282 (2003)
3. M. Chmeissani, C. Frojdh, O. Gal, X. Llopart, J. Ludwig, M. Maiorino, E. Manach, G. Mettivier, M.C. Montesi, C. Ponchut, P. Russo, L. Tlustos, A. Zwerger, IEEE Trans. Nucl. Sci. **51**, 2379 (2004)
4. B. Büttgen, T. Oggier, R. Kaufmann, P. Seitz, N. Blanc, Proc. SPIE **5302**, 9 (2004)

5. J. Ohta, *Smart CMOS Image Sensors and Applications* (CRC Press, Boca Raton, London, New York, 2008), pp. 40–42
6. X. Llopart, M. Campbell, R. Dinapoli, D. San Segundo, E. Pernigotti, IEEE Trans. Nucl. Sci. **49**, 2279 (2002)
7. G. Gramegna, P. O'Connor, P. Rehak, S. Hart, IEEE Trans. Nucl. Sci. **44**, 385 (1997)
8. A. Nardulli, in *IEEE Nuclear Science Symposium Conference Record*, Dresden, Germany, 19–25 October 2008, p. 2731
9. C. Kapnistis, K. Misiakos, N. Haralabidis, in *Proceedings of the 6th IEEE International Conference on Electronics, Circuits and Systems*, vol. 3, 5–8 September 1999, p. 1368
10. T. Noulis, S. Sikos, G. Sarrabayrouse, L. Bary, IEEE Trans. Circuits Syst. I, Reg. Papers **55**, 1854 (2008)
11. E. Beuville, K. Borer, E. Chesi, E.H.M. Heijne, P. Jarron, B. Lisowski, S. Singh, Nucl. Instr. Meth. A **288**, 157 (1990)
12. Z.Y. Chang, W.M.C. Sansen, Nucl. Instr. Meth. A **305**, 553 (1991)
13. F. Krummenacher, Nucl. Instr. Meth. A **305**, 527 (1991)
14. B. Krieger, K. Ewell, B.A. Ludewigt, M.R. Maier, D. Markovic, O. Milgrome, Y.J. Wang, IEEE Trans. Nucl. Sci. **48**, 493 (2001)
15. B. Krieger, I. Kipnis, B.A. Ludewigt, IEEE Trans. Nucl. Sci. **45**, 732 (1998)
16. C. Foirini, M. Porro, IEEE Trans. Nucl. Sci. **51**, 1953 (2004)
17. R. Dlugosz, in *Proceedings of the 15th International Conference on Mixed Design of Integrated Circuits and Systems*, Poznań, Poland, 19–21 June 2008, p. 627
18. C.H. Nowlin, J.L. Blankenship, Rev. Sci. Instrum. **36**, 1830 (1965)
19. P. Gryboś, R. Szczygiel, IEEE Trans. Nucl. Sci. **55**, 583 (2008)
20. R. Szczygiel, P. Gryboś, P. Maj, A. Tsukiyama, K. Matsushita, T. Taguchi, IEEE Trans. Nucl. Sci. **56**, 487 (2009)
21. L.T. Wurtz, W.P. Wheless Jr., IEEE Trans. Instrum. Meas. **42**, 942 (1993)
22. P. Gryboś, M. Idzik, A. Skoczeń, Analog Integr. Circuits Signal Process **49**, 107 (2006)
23. W.M.C. Sansen, Z.Y. Chang, IEEE Trans. Circuits Syst. **37**, 1375 (1990)
24. C. Lotto, P. Seitz, in *Proceecings of 2009 International Image Sensor Workshop*, Bergen, Norway, 26–28 June 2009
25. C. Lotto, P. Seitz, in *Proceedings of EOS Conference on Frontiers in Electronic Imaging*, Munich, Germany, 15–17 June 2009
26. X. Llopart, R. Ballabriga, M. Campbell, L. Tlustos, W. Wong, Nucl. Instr. Meth. A **581**, 485 (2007)
27. R. Bellazzini, G. Spandre, M. Minuti, L. Baldini, A. Brez, F. Cavalca, L. Latronico, N. Omodei, M.M. Massai, C. Sgro, E. Costa, P. Soffitta, F. Krummenacher, R. de Oliveira, Nucl. Instr. Meth. A **566**, 552 (2006)

Chapter 12
Single-Photon Detectors for Time-of-Flight Range Imaging

David Stoppa and Andrea Simoni

Abstract We live in a three-dimensional (3D) world and thanks to the stereoscopic vision provided by our two eyes, in combination with the powerful neural network of the brain we are able to perceive the distance of the objects. Nevertheless, despite the huge market volume of digital cameras, solid-state image sensors can capture only a two-dimensional (2D) projection, of the scene under observation, losing a variable of paramount importance, i.e., the scene depth. On the contrary, 3D vision tools could offer amazing possibilities of improvement in many areas thanks to the increased accuracy and reliability of the models representing the environment. Among the great variety of distance measuring techniques and detection systems available, this chapter will treat only the emerging niche of solid-state, scannerless systems based on the TOF principle and using a detector SPAD-based pixels. The chapter is organized into three main parts. At first, TOF systems and measuring techniques will be described. In the second part, most meaningful sensor architectures for scannerless TOF distance measurements will be analyzed, focusing onto the circuital building blocks required by time-resolved image sensors. Finally, a performance summary is provided and a perspective view for the near future developments of SPAD-TOF sensors is given.

12.1 Introduction

We live in a three-dimensional (3D) world and thanks to the stereoscopic vision provided by our two eyes, in combination with the powerful neural network of the brain we are able to perceive, in real-time, the distance of the objects. Nevertheless, despite the huge market volume of digital cameras that are nowadays ubiquitous in everyday life (e.g., mobile phone, security, and digital still cameras), solid-state image sensors can capture only a two-dimensional projection (2D), or intensity map, of the scene under observation, losing a variable of paramount importance, i.e., the scene depth. This 2D representation of the world is often not sufficient to build the sophisticated models required by automatic systems capable of analyzing

P. Seitz and A.J.P. Theuwissen (eds.), *Single-Photon Imaging*,
Springer Series in Optical Sciences 160, DOI 10.1007/978-3-642-18443-7_12,

and interpreting the environment, emulating the capability of the human brain. On the contrary, a 3D vision tool could offer amazing possibilities of improvement in many areas like security and surveillance, cultural heritage preservation, industrial control, automatic production, cinema industry, virtual reality, gaming, etc., thanks to the increased accuracy and reliability of the models representing the environment. Contactless measuring techniques suitable to estimate the target distance exploit micro-, ultrasonic-, or light-waves [1, 2]. However only the latter technique allows achieving good angular resolution performance, in a compact measuring setup, as required for a 3D-imaging system [3]. Active optical measuring techniques using light-waves can be further classified in three main categories,[1] namely: interferometry, triangulation, and Time-of-Flight (TOF) [4, 6]. The best precision performance can be obtained by means of interferometry, which reaches resolution far below the wavelength of the light source used to actively illuminate the target [4]. Moreover, by using multi-wavelength techniques the distance range can be extended up to several meters [7, 8]. However, interferometric systems require bulky and precise optical setup making them suitable in specific applications where the cost and the complexity of the instrument are not an issue [9, 10]. Active triangulation techniques (e.g., laser triangulation) offer precision performance in the order of tens of micrometers, but the maximum depth range is limited since it is determined by the triangulation baseline length [5]. Many interesting triangulation systems, using a sheet of light to probe the target, have been proposed [11–13]. These systems, thanks to a full custom detector allowing for a fast localization of the light spot onto the pixel array, are capable of high spatial resolution range maps and high frame rates. Nevertheless, laser triangulation-based sensors require scanning mechanisms for the scene illumination, thus making difficult the integration in low-cost, and compact systems. On the contrary, TOF techniques exhibit much better performance in terms of measurable distance, which can range from a few centimeters up to several kilometers, but the precision is poorer in comparison to interferometry and triangulation [5, 6]. However, although scanner TOF systems are available on the market for top-class 3D-measurement apparatus [9, 14, 15], there has been in the last decade an emerging interest to develop scannerless and all-solid-state TOF cameras, where flood illumination, instead of a collimated light beam, is used to avoid light scanning. Many full-custom detectors have been developed for this specific class of systems. They can be mainly classified in three categories:

(a) Intrinsic photon mixing devices. This approach exploits photo-demodulators, in which the photogenerated charge is mixed on two or more collection electrodes thus achieving an intrinsic demodulation effect.

(b) Standard photodiodes coupled to dedicated processing circuitry. This approach exploits the extensive use of electronics (either in the pixel or at the periphery) to recover the distance information from the current photo-generated by the photodiode.

[1] More detailed and broader overviews over optical ranging principles can be found in the references [1–5].

(c) *Avalanche Photodiodes (APDs) or Single-Photon Avalanche Diodes (SPADs) coupled to proper processing circuitry.* With respect to (b), an avalanche diode, operating in Geiger regime to achieve sensitivity to individual photons, is used to collect the light echo and coupled to a readout and processing electronics aimed at extracting the TOF information.

Sensors belonging to (a) represent the most mature solution, and most of the scannerless 3D cameras available on the market are based on this concept [16–18], although different solutions have been adopted to implement the photo-demodulator device. The main advantage of this approach is the read-out channel simplicity, which results in a small pixel size; the problems are the sensitivity to the ambient light due to reduced geometrical fill factor, and the cost of nonstandard technologies (e.g., CCD/CMOS, modified CMOS, high resistivity substrate, etc.), that are often required. Sensors based on (b) are characterized by large pixel pitch due to the extensive use of electronics used to properly process and accumulate the photo-generated charge. Moreover, they typically have lower precision with respect to sensors based on (a) because of the noise contribution introduced by the transistors introduced in the signal path. On the other hand they exhibit an excellent immunity to continuous background light [19–21], which makes this approach the best suited for outdoor operation such as automotive and security. Finally, sensors based on category (c) have been widely used in high-performance single-point scanner systems since the 80s, using avalanche photodiodes (APDs) fabricated within dedicated technologies coupled to discrete-components read out and processing electronics or measurement instruments [22–25]. Only recently, the possibility of implementing good performance APDs/SPADs in CMOS technologies, opened the way to the realization of range image sensors based on this approach. The great advantage of this solution is the extremely high sensitivity of the photodetector, capable of detecting down to single-photon, and the intrinsic low-noise performance that allows operating at the shot-noise limit. Among the great variety of distance measuring techniques and detection systems above mentioned, this chapter will treat only the emerging niche of solid-state, scannerless systems based on the TOF principle and using as, a detector, SPAD-based pixel array. The chapter is organized into three main parts. At first, TOF optical systems and measuring techniques will be described, highlighting the most critical aspects and the main demands of these systems. In the second part, the main application fields of TOF range systems and their relative requirements will be summarized. In the third part, most meaningful sensor architectures devoted to scannerless TOF distance measurements will be described and compared, focusing onto the circuital building blocks required by a time-resolved image sensor suitable for TOF. Finally a performance summary is provided and a perspective view for the near future developments of SPAD-TOF sensors will be given, analyzing the main challenges to be faced and briefly describing on-going, most promising research lines in this field.

12.2 Time-of-Flight Measuring Techniques and Systems

TOF measuring technique relies on the measurement of the round-trip time needed by a light pulse to travel from the measurement system to a target and back again. The first TOF measuring system has been set up by the French physicist Armand Hippolyte Louis Fizeau in the 19th century, to estimate the speed of the light propagating through the air [26]. The ingenious apparatus he built consisted of a light source, a mirror target, and a rotating toothed wheel used as a mechanical light shutter. Knowing the distance between the wheel and the mirror, the number of teeth on the wheel, and the rate of rotation, Fizeau estimated the speed of light with an excellent accuracy for the time, but the most interesting thing is that modern TOF range instruments are based on the same conceptual scheme as explained in the next subsections.

12.2.1 Time-of-flight System

A typical TOF measuring setup is sketched in Fig. 12.1. It consists of five main components:

Modulated light source: LASER or LED typically in the infrared part of the spectrum to make the illumination unobtrusive. LASER diodes allow higher modulation frequencies and typically higher optical power, LEDs on the other hand are cheaper and more suitable to make the system compliant with eye-safety rules.

Optical diffuser: in scanner-less TOF systems the emitted light is spread onto the scene by means of an optical diffuser in order to achieve a flood illumination over a wide field-of-view. This element is not needed if the illumination source has proper angles of divergence along the two axis.

Collection lens: the light echo back-reflected by the target is gathered and focused onto the sensor focal plane by means of an optical objective. An optical band-pass, properly tuned onto the wavelength of the light source, allows improving the background noise rejection.

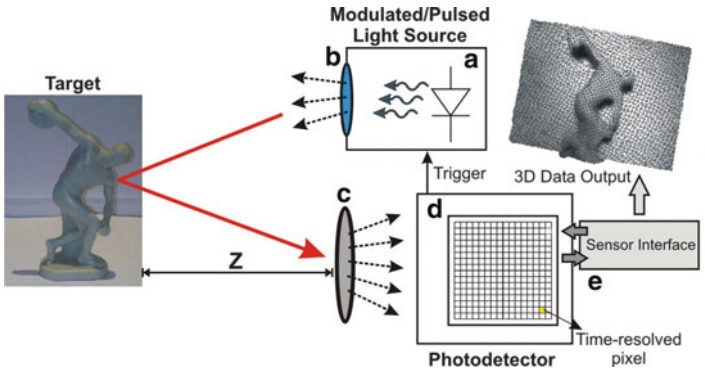

Fig. 12.1 Scannerless TOF measuring system

Solid-state range sensor: constitutes the core of the measuring system. In a scanner-less 3D-imaging system it is composed of an array of photo-detectors (pixels) capable of measuring, in a direct on indirect way, the TOF needed by the light pulse to travel from the light source to the target and back to the sensor.

Sensor interface: providing to the sensor the power supply, required biasing voltage/current signals, digital control phases, and reading out from the sensor the data stream, which typically requires further minor processing to obtain the 3D volume data. Finally, the sensor interface is responsible for the communication with the external world (to a PC, or a processing unit).

12.2.2 Direct and Indirect Time Measuring Techniques

In a classical TOF measurement, referred in the following as Direct-TOF (D-TOF), the detector system starts a highly accurate stopwatch synchronously with the emitter light pulse generation. As the light echo from the target is detected, the stopwatch is stopped and the roundtrip time τ_{TOF} is directly stored. The target distance z can be estimated by means of the simple equation:

$$z = \frac{c}{2} \cdot \tau_{\mathrm{TOF}}, \tag{12.1}$$

where $c = 2.9979 \cdot 10^8\,\mathrm{m/s}$ represents the speed of the light propagating through the air. D-TOF is commonly used for single-point range systems [22–25], but only recently implemented in scannerless TOF systems, because of the difficulties in implementing at a pixel level sub-nanosecond electronic stopwatch. In the following sections the main noise limitation and a description of state-of-the-art systems based on D-TOF will be described. An alternative solution to D-TOF is the so-called Indirect-TOF (I-TOF), where the roundtrip trip time is indirectly extrapolated from a time-gated measurement of the light intensity. In this case, there is no need of precise stopwatch, but of time-gated photons counters, which can be implemented at a pixel level with less efforts and silicon area. The operation principle of D-TOF and an example of a four-gates I-TOF are illustrated in Fig. 12.2 considering both pulsed and modulated light sources. Many other implementations of I-TOF are possible. A particularly interesting one requires a single photon time-gated counter [27], or photo-charge integrator [19, 21], and the light echo is measured through several consecutive time windows (see Fig. 12.3).

The timing of the first window ($W1$) is arranged in a way that the whole duration of the pulses (T_{P}) is collected for objects within the minimum and maximum distance. Each of the following windows ($W2, W3, \ldots$) is shifted in time by T_{SH} with respect to the previous one making a portion of the received pulses fall outside the pixel integration window. Depending on the time of flight of the photons reflected by the scene, pixels looking at close objects get their photons earlier than pixels focused on farther details, so they lose a bigger part of the reflected pulse.

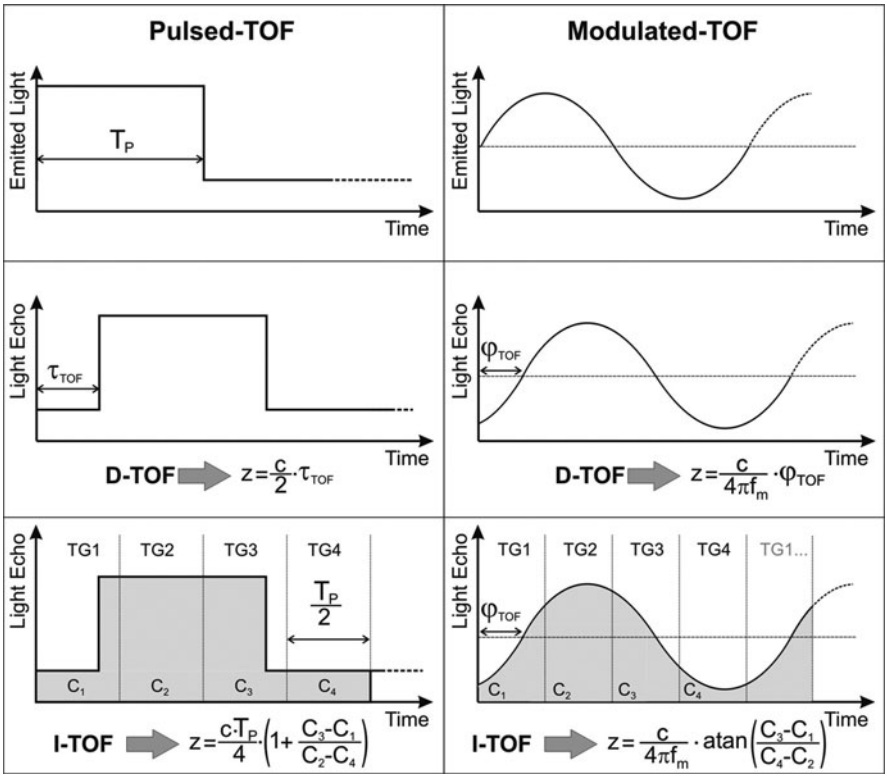

Fig. 12.2 Pulsed, Modulated D- and I-TOF measuring techniques using four time-gated photon counters

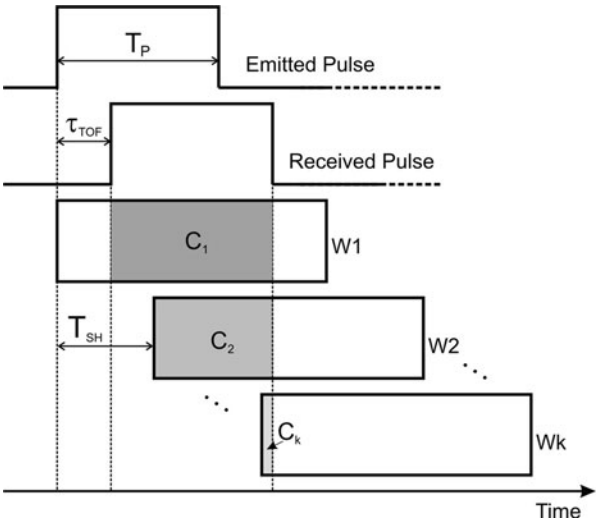

Fig. 12.3 Consecutive time-window I-TOF using a single photon counter

The distance information, achievable from the generic window WK, can be easily extracted by dividing the photons collected within each window by means of (12.2).

$$D_K = \frac{c}{2} \cdot \tau_{\mathrm{TOF}} = \frac{c}{2} \cdot \left[T_{\mathrm{P}} \cdot \left(\frac{C_K}{C_1} - 1 \right) + (K-1) \cdot T_{\mathrm{SH}} \right]. \tag{12.2}$$

Depending on the different distance intervals to be mapped, only one window provides the best measurement accuracy. The main advantage of this technique is that the measurable distance range is not limited by the modulation frequency of the emitter. On the contrary, part of the light echo is wasted being acquired by a single time-gated counter, and more cycles are needed making this approach relatively slow. Although I-TOF can be used with pulsed light as described in the previous example [19, 21], it finds its natural implementation with sinusoidal-modulated emitters [28–32]. In this case, the target distance can be extracted from the phase-shift φ_{TOF}, of the received light echo with respect to emitted light by means of:

$$z = \frac{c}{4\pi f_m} \cdot \arctan \frac{C_3 - C_1}{C_2 - C_4}, \tag{12.3}$$

where $f_m = 1/T_{\mathrm{p}}$ is the modulation frequency of the emitter and C_{1-4} the number of photons detected within each time-gated window. The main advantages of this approach with respect to the pulsed I-TOF are the good linearity of the distance measurement and that it is more suited to generate high frequency light signals using low-cost emitters. The main drawback is that the measurable range is limited by the adopted modulation frequency causing ambiguities in the distance estimation. Finally, pulsed techniques with a duty cycle $\ll 50\%$ offer better immunity to background light, the sensor being exposed to the light only in a short time window, while in the modulated approach the sensor is on average always exposed to continuous light.

12.2.3 Optical Power Budget

To better understand why single-photon detectors are needed in TOF range systems, it is useful to estimate the amount of photons impinging onto each pixel after the travel emitter-target-sensor. Let us consider that the emitter is characterized by a total optical power P_{E}, and suppose this is uniformly spread over an angle (cone of light) θ_{E}. After hitting the target, which can be modelled as a Lambertian reflector [33] (scattered light varies along the observation angle as $\cos\alpha$), the light pulses are then diffused back toward the sensor. The amount of light collected by the lens used to gather light onto the sensor focal plane depends on the target distance and lens characteristics, namely the focal length f and the diameter D, so that its "f number" is $F/\# = f/D$. The power P_{L} collected by the lens is a fraction of the power impinging on the target P_{T}, dependent on the angle α_{L} of light rays intercepted by the lens.

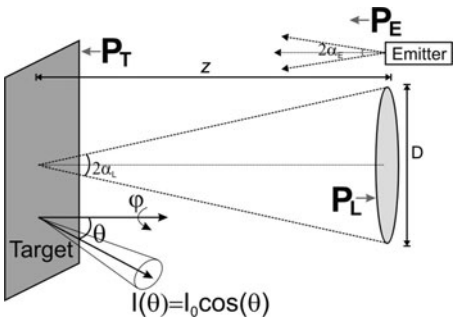

Fig. 12.4 Optical flux intercepted by the collection optics in case of Lambertian target

By considering a rotational symmetry around the φ axis (see Fig. 12.4), the optical flux emitted in a cone having an angle $2\alpha_{\mathrm{L}}$, can be calculated as:

$$\Phi = \iint I(\theta,\phi)\cdot\sin(\theta)\mathrm{d}\theta\mathrm{d}\varphi = 2\pi\int_0^{\alpha_{\mathrm{L}}} I_0\cos(\theta)\sin(\theta)\mathrm{d}\theta = \pi I_0\sin^2(\alpha_{\mathrm{L}}). \tag{12.4}$$

So that the optical power collected by the lens can be extracted by considering the light flux within the solid angle intercepted by the lens over the whole semisphere solid angle as:

$$P_{\mathrm{L}}(z) = P_{\mathrm{T}}\cdot k_{\mathrm{M}}\cdot\sin^2(\alpha_{\mathrm{L}}) = P_{\mathrm{T}}\cdot k_{\mathrm{M}}\cdot\left(\frac{D}{2z}\right)^2 = P_{\mathrm{T}}\cdot k_{\mathrm{M}}\cdot\left(\frac{f}{z}\right)^2\cdot\left(\frac{1}{2F_{\#}}\right)^2, \tag{12.5}$$

where k_{M} is the attenuation due to the propagation through the medium. The photon flux collected by the lens is then focused onto the sensor focal plane array, which typically has a rectangular or square shape. Let us consider a square focal plane array having size L_{s}, and composed by pixels with a pitch p and fill factor FF. In the optimal condition, the illuminator emission angle is matched with the lens field of view, but to avoid dark corner in the image, the focused light spot circle has to include the sensor array (see Fig. 12.5). The sensor field of view, FOV, is a function of the lens focal length and the sensor size, accordingly to (12.6):

$$\mathrm{FOV} = 2\arctan\left(\frac{L_{\mathrm{s}}}{f}\right). \tag{12.6}$$

However, to fully illuminate the pixel array the illuminator divergence should be slightly larger than the FOV (see Fig. 12.5), causing a loss of photons in the wasted zones at the sensor corners. Using (12.4) and (12.5) it is possible to estimate the number of photons $N_{\mathrm{p}}(z)$ impinging onto each pixel at each light pulse as a function of the target distance z:

$$N_{\mathrm{p}}(z) = \frac{1}{2\pi}P_{\mathrm{E}}\cdot\rho\cdot k_{\mathrm{E}}\cdot k_{\mathrm{R}}\cdot k_M^2\cdot\left(\frac{f}{z\cdot F_{\#}}\right)^2\left(\frac{p}{L_{\mathrm{s}}}\right)^2\cdot FF\cdot\frac{\lambda}{hc}\cdot T_{\mathrm{p}}, \tag{12.7}$$

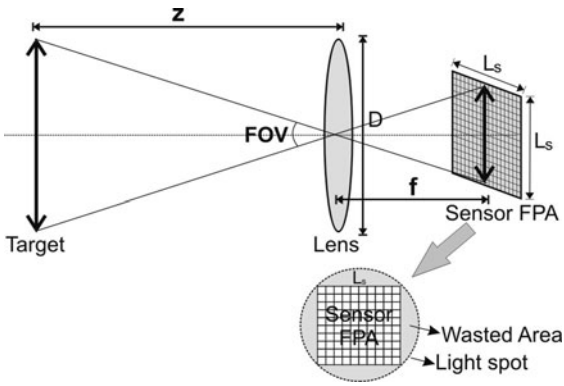

Fig. 12.5 Sensor field of view (FOV) depends on the sensor size and lens characteristics

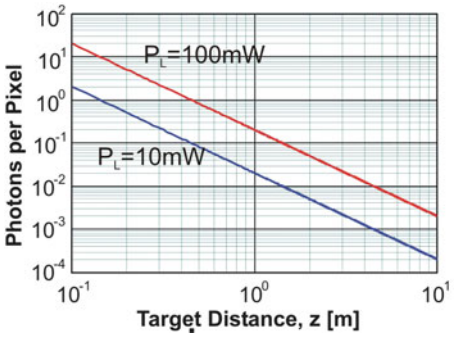

Fig. 12.6 Number of detected photons per pixel for each light pulse of duration $T_p = 10$ ns and for two different optical power levels of the emitter considering the system parameters of Example 1

where P_E is the optical power of the emitter, ρ the target reflectivity, k_E and k_R the optical losses of the emitter and receiver optics, respectively, λ the emitter wavelength, FF the pixel fill factor, T_p the light pulse width and h is Planck's constant.

Example 1: To have a numerical estimation, we can consider as an example the following TOF system parameters:

- *Collection optics*: focal length $f = 6$ mm, F number $F\# = 1.2$, and lens losses $k_R = 90\%$;
- *Sensor*: pixel pitch $p = 10\,\mu$m, fill factor $FF = 20\%$, sensor size $L_S = 1.6$ mm;
- *Target*: target reflectivity $\rho = 70\%$, and propagation in air $k_M = 1$;
- *Emitter*: emitter wavelength $\lambda = 850$ nm, pulse width $T_p = 10$ ns, and diffusive optics $k_E = 1$.

$N_p(z)$ is plotted in Fig. 12.6 considering two values of the emitter light power. It appears clearly that the intensity of the light echo impinging onto the pixel array is only a minimal portion of the emitted light, requiring single-photon detection capability in most of the target distance range. Another consideration, which is

worth nothing, is that the detection system must be capable of a high dynamic range to cope with the large signal variations along the target distance and reflectivity.

12.2.4 D-TOF and I-TOF Noise Considerations

Direct-TOF (D-TOF): The sources of error in this measurement are the statistics of photon arrival, the dark count rate and the time resolution of the SPAD, background light sources, electronic jitter, and laser pulse shape. In the following discussion, we will consider only the contributions of photon arrival statistics and laser pulse shape, in order to investigate the ultimate noise limit. However, typical dark-count rates of SPAD are in the kHz range and can be neglected, while the effect of background light sources can be strongly reduced by optical filtering. As far as jitter noise is concerned, the SPAD time resolution is usually <100 ps, and the jitter noise of the readout electronics has a similar value so that they can be neglected as well. The number of photons incident on the photodetector during the pulse time T_P follows Poisson statistics as long as T_P is much longer than the coherence time τ_c of the light source [34]. If a constant Poissonian light with a photon flux of $n_p(z)$ photons per second impinges on the active area of a SPAD, the detection of a photon after a time t from an arbitrary starting time ($t = 0$) has an exponential probability distribution [17]. The probability density function, $f(t)$, is given by:

$$f(t) = n_p(z) \cdot e^{-n_p(z)t}. \tag{12.8}$$

In case of a light pulse having time width T_P, with $T_P \gg \tau_c$, the light can still be considered Poissonian, but $n_p(z)$ is a function of time so that (12.8) is replaced by:

$$f(t) = n_p(z,t) \cdot \exp\Big[-\int_0^{T_P} n_p(z,\tau)\mathrm{d}\tau\Big]. \tag{12.9}$$

The probability that a photon is detected within each light pulse is given by the integral of $f(t)$, $F(T_P)$:

$$F(T_P) = \int_0^{T_P} f(t)dt = 1 - \exp\Big[-\int_0^{T_P} n_p(z,\tau)\mathrm{d}\tau\Big]. \tag{12.10}$$

We can distinguish two cases: if $F(T_P) \ll 1$, for most of the pulses no photons are detected. For a system working in this condition, a measurement is valid only in the cycles in which a photon is detected, while the other measurements must be discarded. The density function $f(t)$ is approximately equal to $n_p(z)$, and the time resolution is linked to the width of the optical pulse, to $F(T_P)$ and to the number of measurements N. If σ_{tL} is the laser pulse width, the theoretical time resolution of the system σ_t is given by:

$$\sigma_t = \frac{\sigma_{tL}}{\sqrt{N \cdot F(T_P)}}, \tag{12.11}$$

where $N \cdot F(T_P)$ is the average number of photons detected during N measurement cycles. The second limiting case is when $F(T_P) \simeq 1$. In this case a photon is detected on average at every laser pulse and the time resolution of the measurement is given by the standard deviation of $f(t)$. In the case of a rectangular pulse of $n_p(z)$ photons/s·pixel having a duration T_P, the time resolution is given by:

$$\sigma_t = \frac{1}{n_p(z)}. \tag{12.12}$$

The mean of $f(t)$ is also equal to $1/n_p(z)$ and determines an offset in the measurement of the optical pulse arrival time. This offset, which depends on the incident optical power, causes a systematic error that must be removed or at least minimized to perform an accurate measurement. While time resolution can be improved by averaging over a large number of measurements, the illumination-dependent offset is more difficult to remove with a calibration. However, it can be minimized by increasing the optical power and using SPADs with a high quantum efficiency. As an alternative, very short laser pulses (<1 ns) can be used to fall in the case where $F(T_P) \ll 1$, thus obtaining an easy-to-calibrate offset independent on illumination.

Indirect-TOF (I-TOF): To analyze the noise limitation of the I-TOF measuring technique let us consider, for sake of simplicity, the case of two time-gated windows only (see Fig. 12.4). The number of photons C_1 detected within the first time window W1 after N light pulses is given by:

$$C_1 = N \cdot n_p(z) \cdot T_P. \tag{12.13}$$

While the number of photons detected within the second time window W2 is:

$$C_2 = N \cdot n_p(z) \cdot (T_P + \tau_{TOF} - T_{SH}). \tag{12.14}$$

The time resolution σ_t of the measurement is related to the standard deviations σ_{C1} and σ_{C2} of C_1 and C_2. As $n_p(z)$ follows a Poisson statistics, σ_{C1} is expressed by:

$$\sigma_{C1} = \sqrt{N \cdot n_p(z) \cdot T_P}. \tag{12.15}$$

To simplify the notation let us define a parameter α so that $C_2 = \alpha \cdot C_1$, with $0 \leq \alpha \leq 1$. From (12.13) and (12.14) it can be seen that α is a linear function of τ_{TOF} and σ_{C2} can be expressed by the equation:

$$\sigma_{C2} = \sqrt{\alpha \cdot N \cdot n_p(z) \cdot T_P}. \tag{12.16}$$

By means of (12.2) and applying error propagation, the time resolution of the measurement is given by:

$$\sigma_t = T_{\mathrm{P}} \cdot \sqrt{\frac{\alpha \cdot (1+\alpha)}{C_1}} \tag{12.17}$$

σ_{t} reaches its maximum value for $\alpha = 1$, which corresponds to the case $\tau_{\mathrm{TOF}} = T_{\mathrm{SH}}$. In this case the time resolution of the system is inversely proportional to $(n_{\mathrm{p}}(z))^{1/2}$, while in the D-TOF approach (12.12) σ_{t} is inversely proportional to $n_{\mathrm{p}}(z)$. This method, if compared to D-TOF, does not suffer from illumination-dependent systematic errors, and is thus more accurate.

12.3 Single-Photon Sensors for 3D-TOF Imaging

A solid-state detector suitable for scanner-less, single-photon, TOF imaging is generally composed of a two-dimensional array of pixels, each one integrating a SPAD detector, the required bias and readout electronics, a processing circuitry dedicated to time-gated photon counting (I-TOF) or capable of measuring the photon arrival time (D-TOF). The SPAD detector can be either integrated within the read out and processing electronics (full CMOS system) or realized using a dedicated fabrication process and connected in stacked-chip configuration (hybrid system). The main advantage of a full CMOS solution is the low fabrication costs, while hybrid systems allow using SPADs with better performance in terms of quantum efficiency and dark count rate (DCR). Moreover, they can be realized with Class III–V semiconductor material allowing the use of 1,500-nm emitters, making eye-safe operation much easier even at high emission power [35].

12.3.1 Single-photon Detectors

There are many nonsolid-state devices capable of single-photon detection, i.e., photomultiplier tubes (PMTs) and microchannel plates (MCPs). PMTs offer extremely good performance in terms of quantum efficiency and detection speed, however, they are bulky and require high operating voltages preventing their use in two-dimensional arrays [36]. MCPs are arranged in two-dimensional arrays but the detection speed is limited by the response time of the phosphor-screen and CCD exposure and readout time. To overcome these problems, many solid-state alternatives have been developed, such as silicon avalanche photodiodes (APDs) and silicon photomultipliers [37], but only recently CMOS compatible APD device structures have emerged [38–41]. Single-photon detectors can be implemented exploiting the avalanche principles of a p–n junction, biased above the breakdown

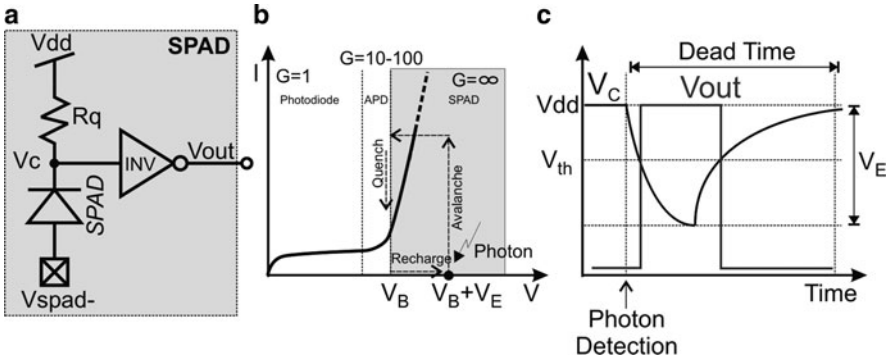

Fig. 12.7 (**a**) Basic SPAD front-end scheme based on passive quenching and A/D conversion by means of an inverter amplifier, (**b**) I/V curve of the reverse biased diode and Geiger operation region, and (**c**) typical detection cycle

voltage V_B, working in a meta-stable regime called Geiger mode, where the absorption of a photon triggers an avalanche current. An APD working in Geiger mode is known as single-photon avalanche diode (SPAD). In this bias condition the gain of the device is virtually infinite, allowing for single-photon detection, and a proper quenching circuitry is needed to stop the avalanche and restore the initial condition of the device (see Figs. 12.7 and 12.8). The simplest quenching circuit is realized by means of a resistor placed in series with the diode, so that as the avalanche current increases the p–n junction bias voltage drops below the breakdown reducing the gain of the device. After that the initial bias condition is restored through the quenching resistor R_q, requiring a minimum dead time (limited by the RC constant given by R_q and the overall capacitances at the SPAD cathode) before the detection of the next photon. To guarantee compatibility with standard CMOS technologies, the SPAD anode is biased at a negative voltage so that the excess bias $V_e = (V_c - V_{SPAD} - |V_{bd}|) < V_{dd}$. The voltage drop at the SPAD cathode can be easily converted into a digital pulse by means of a voltage comparator, which is typically implemented as a simple CMOS inverter. State-of-the-art CMOS SPADs have been successfully implemented in different technologies, ranging from 0.8 μm down to 130 nm [38–47] and exhibit dead times down to about 30 ns [38], photon-detection probability higher than 35% in the visible range [39] and dark count rates of a few Hertz [39, 41]. The great advantage of a CMOS SPAD is the possibility to realize large pixel arrays and include at a pixel-, column-, or array-level sophisticated read out and processing blocks. Many different device structures have been proposed to implement SPADs in standard CMOS technologies, and several circuit solutions have been developed to implement the quenching of the avalanche and the detection of the avalanche event (for a detailed description see Chap. 7 by Charbon, of this book). In the following sections, we will consider the SPAD ensemble as a black box containing the basic scheme of Fig. 12.7 (shadowed square), so that a digital pulse is generated after each photon detection.

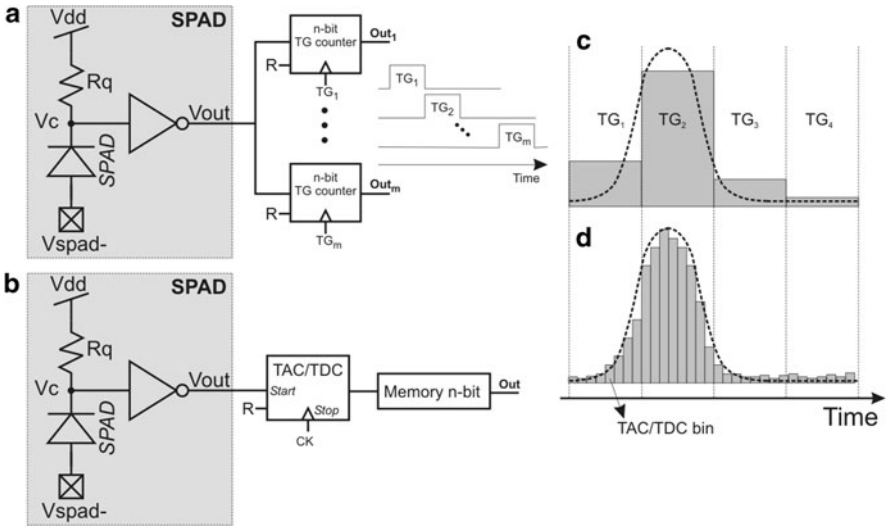

Fig. 12.8 Pixel architectures for TOF imaging using SPADs: (**a**) by using time-gated photon counters or (**b**) by means of TAC/TDC. Slice (**c**) illustrates the detection of the light echo through four time-gates counters, while (**d**) in the case of TAC/TDC

12.3.2 Pixel Architectures for Single-photon TOF Imaging

There are two main architectures suitable to implement a time-resolved SPAD detector: (1) by using one ore more time-gated photon counters as sketched in Fig. 12.8a, or (2) by means of Time-to-Amplitude (TAC) or Time-to-Digital (TDC) converters as shown in Fig. 12.8b. In the first case, when a photon is detected within the i-th time slot, only the relative counter is incremented, so that at the end of the measurement cycle, consisting of several light pulses, it is possible to build a photon-count histogram combining the sensor outputs Out_{1-m} (Fig. 12.8c). If a TAC/TDC is used, a time-stamp information is associated to each detected photon, so that the reconstruction of the detected light echo is more accurate, being the number of bins limited by the resolution of the TAC/TDC (Fig. 12.8d). The main drawbacks of approach (2) are the huge amount of information generated by the sensor in case of large pixel array, which require in-pixel memory and dedicated high-speed readout circuits to properly buffer out the data stream, and the high complexity of TAC/TDC circuits, which can be implemented at a pixel level at the expense of the fill factor. Both (1) and (2) can be realized implementing the processing circuitry at a pixel-, column-, or array-level. Pixel-level implementations are the best suited in case of high photon count rate, being the processing channel dedicated to every detector. The negative aspects are the increase of the pixel complexity, which affects the pixel size, the fill factor, and the overall power consumption. Array-level implementations can hardly be used in case of range imaging, because of the loss

of data due to conflicts in accessing the processing circuit by the pixel array, while column-level solutions offer a reasonably good trade-off.

12.3.3 Circuit Implementations for I-TOF Pixels

The basic scheme of Fig. 12.8a can be easily realized in the digital domain by means of a set of standard n-bit ripple counters. Time-gated functionality can be achieved by using a simple demultiplexer between the SPAD front-end and the counter, or by means of other digital schemes such as the Time-gating (2) scheme shown in Fig. 12.9, where a T-type flip flop and a monostable circuit guarantee the generation of fixed time width pulse that eliminates critical race conditions on the ripple counters. Several SPAD sensors, fabricated within 0.35-μm technologies, based on this approach have been recently proposed: in [32] it is proposed a 60×48-pixel array implementing I-TOF by means of two-counters per pixel based on the Time-gating (1) approach, while Time-gating (2) is used in [42] where a 64-pixel linear array using four in-pixel counters and four memory banks to allow time-interleaved operation, has been developed for lifetime imaging applications, but also validated as a linear TOF detector [43]. A similar architecture, using two counters per pixel and developed within the same context, is proposed in [44, 45], where a 16×4-pixel array is proposed. Lifetime imaging techniques, such as Fluorescence Lifetime Imaging Microscopy (FLIM), are powerful investigation tools emerging in life-sciences, medical research, fluid-dynamics analysis, etc., where fluorescent dyes are stimulated by a fast-pulsed laser and the time-decay of their emission is measured. It is worth noting that a SPAD sensor developed for TOF imaging,

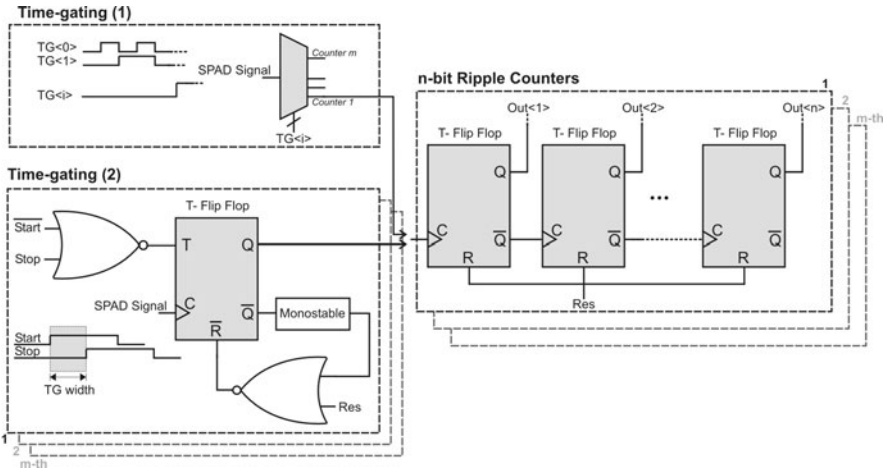

Fig. 12.9 Two alternative schemes for a digital implementation of time-gated photon counters suitable for I-TOF

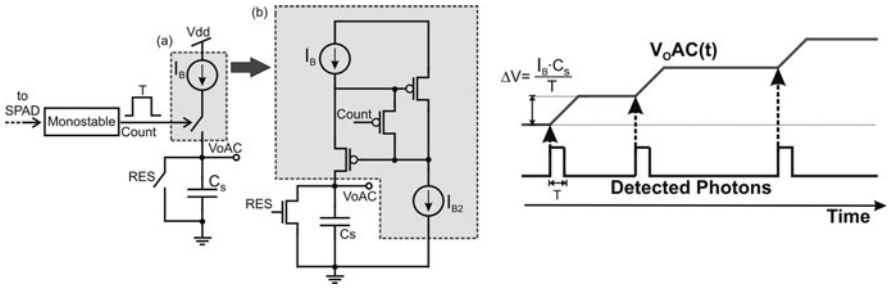

Fig. 12.10 Analog photon counter basic scheme (**a**) and high-speed, low charge injection structure (**b**)

suits very well the requirements of FLIM detectors [46], and therefore they have been mentioned in this chapter. Photon counters can be also realized in the analog domain, requiring much less silicon area with respect to the digital counterpart. The main issues with analog implementations are: to guarantee good uniformity along the pixel array, while maintaining low power consumption and keeping the electrical noise as low as possible, which typically requires quite sophisticated circuit solutions. The basic structure to realize an analog counter is the use of a constant current source I_B which can be selectively connected to a capacitor C_s by means of a switch controlled by the SPAD front-end output (see Fig. 12.10). In order to guarantee a uniform charge-up time T, a multivibrator circuit is needed between the SPAD front-end and the counter. At each detected photon, the voltage V_oAC across capacitor C_s is incremented by $\Delta V = I_B C_S / T$, which represents the analog counter resolution. A switch, controlled by RES signal, is used to restore the initial condition at the end of each acquisition frame. Although this structure is straightforward it has a couple of problems: spurious charge is injected at every count by the switch, and the charging time T must be long enough to effectively close the switch (i.e., create the minority-carriers channel of the transistor used to implement the switch). To cope with these issues the alternative solution shown in Fig. 12.10b can be used, where a feedback loop guarantees very low impedance at node V_oAC reducing charge injection effects and boosting the transient response, at the expense of extra power consumption due to I_{B2}. A 64-pixel linear sensor based on this principle has been proposed in [27] where an analog counter also capable of averaging filtering operation in the analog domain, allows extracting both I-TOF and D-TOF range maps. By means of external calibration, the residual nonuniformity has been roughly reduced down to 2%. To improve the pixel-to-pixel uniformity it is necessary to adopt self-calibration structures, which compensate the unavoidable process and geometrical mismatches. A possible solution has been proposed in [47] where a reference copy of the circuit shown in Fig. 12.10 has been included in each pixel, and used to implement an analog-to-digital conversion intrinsically immune to mismatch effects. The concept is illustrated in Fig. 12.11. The circuit operates in two phases: at the end of the exposure window (phase-1) capacitor C_s contains the

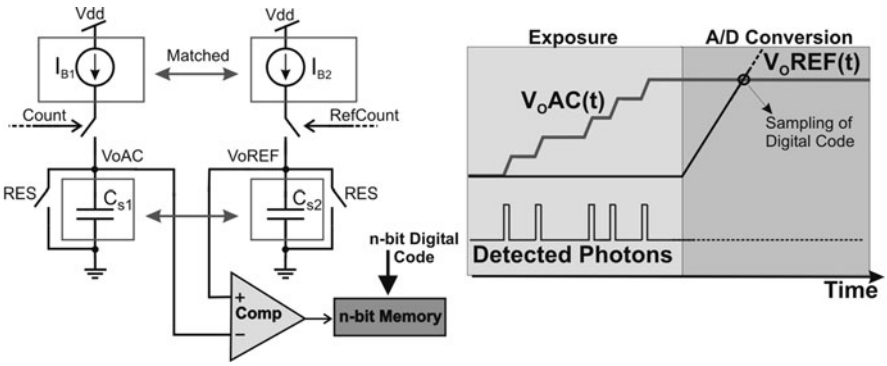

Fig. 12.11 Analog counter structure with A/D conversion intrinsically immune to mismatch effects

analog information proportional to the number of counted photons. At this point, the reference block is externally stimulated by the signal RefCount, globally distributed to the whole pixel matrix, synchronously with n-bit digital code in a single-ramp ADC fashion. The voltage at node V_oREF increases and when it reaches the same voltage previously stored on V_oAC the voltage comparator toggles thus sampling the digital code into the memory. By means of proper interdigitated layout structures it is possible to achieve very good matching between $C_{s1,2}$ and $I_{B1,2}$ making the digitally converted code extremely uniform along the pixel array.

12.3.4 Circuit Implementations for D-TOF Pixels

To implement D-TOF by using the architecture shown in Fig. 12.8b it is necessary to implement on-chip time-measuring circuits such as a TAC or TDC. However, these devices are typically based on complex circuits [48–51], whose area occupation and power consumption make their integration impossible at a pixel level or column level. Many efforts have been recently done in the integration of time-measurement circuits, both TDC and TAC, together with SPAD-array, realizing very compact solutions, which have been implemented in a highly parallel way. In the following, most meaningful architectures will be described. A column-level implementation of 32 TDCs, serving an array of 128×128-pixel array is proposed in [52], where a three-level interpolation scheme has been used to implement a 10-bit TDC characterized by a 100-ns time mapping range. The interpolation principle is schematically shown in Fig. 12.12. A global clock signal CLK is used as a reference signal providing coarse resolution τ_C, and its period is split into 2^k uniformly shifted phases $\varphi[2k : 1]$ by means of a global digital delay line (DLL), in order to obtain medium resolution τ_M. The fine resolution τ_F is realized by using an inverter-based delay line, so that τ_F is limited by the minimum delay of

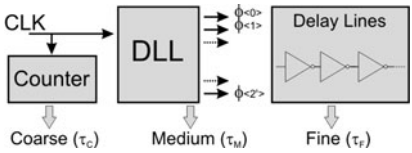

Fig. 12.12 Block diagram of a three-interpolation level digital TDC suitable for array- and column-level implementations

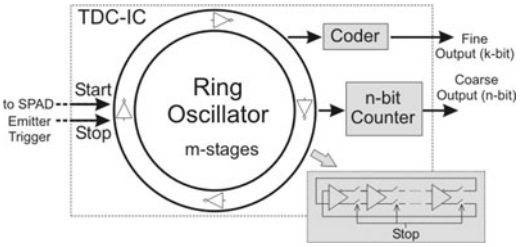

Fig. 12.13 Block diagram of a digital TDC, using internal clock generation, suitable for pixel-level implementations

digital ports, ultimately affected by the adopted technology. Finally, an appropriate interpolation based on a linear combination of $\tau_{C,M,F}$, enables the measurement of a time interval over the whole time range. This solution allows achieving excellent performance in terms of maximum time range, Integral and Differential Non Linearities (INL,DNL), and resolution. The adopted multilevel interpolation allows mapping relatively long time intervals with high resolution without the need for long delay lines, while distributing relatively low-frequency signals only. However, the high circuit complexity requires relatively large silicon area and power consumption, so that its implementation is limited to array- or column-level architectures. An extremely compact, and low power, TDC architecture suitable for pixel-level implementation has been proposed in [53], where a gigahertz ring-oscillator, locally integrated, is started by the SPAD front-end (see Fig. 12.13) when a photon is detected, and a reference signal, synchronous with the emitter trigger, is used to stop the ring, freezing its internal status. An n-bit counter provides a coarse estimation of the photon arrival time (given by the number of oscillation cycles of the ring) while the status of the ring, properly coded, generates the least significant bits. The ring oscillator is realized by means of a chain of current-limited differential inverters connected through pass-transistor switches that are opened by the *Stop* signal in order to sample the ring propagation status. To minimize the impact of dominant supply noise induced jitter effects the cell is constructed from wide swing, differential inverters with bias current regulation, which is included in a feedback loop that guarantees good pixel-to-pixel uniformity. Another TDC architecture having the same level of integration and suitable for pixel-level implementations [54] is sketched in Fig. 12.14. In this structure coarse time measurement is also realized by means of an n-bit counter, and fine resolution is provided by the propagation delay through an inverter chain, but the clock (CK) is externally distributed across the entire pixel array instead of being locally generated. On the first clock edge after the START signal the propagation of the pulse through the delay chain is stopped, thus the number of buffer elements that toggled between

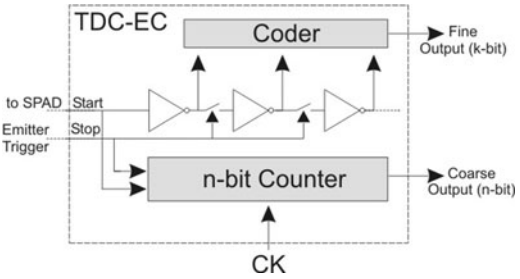

Fig. 12.14 Block diagram of a digital TDC, using external clock generation, suitable for pixel-level implementations

the photon detection and the subsequent clock edge corresponds to the time elapsed between these two events. A coder converts the resulting thermometer code of the fine interpolator output into a k-bit binary number. Finally, the STOP signal interrupts the coarse interpolator (n-bit counter). A calibration loop allows adjusting the tail current of the inverter chain, similarly to the previous implementation. TDC architectures are extremely reliable and offer generally good performance in terms of DNL, INL, and pixel-to-pixel uniformity. However, they require deep-submicron CMOS technologies to be effectively implemented at pixel-level, and to achieve resolutions below 100 ps, while their power consumption is generally high so their use in large pixel array is limited to very low light level applications. A valid alternative is given by TACs that can be realized on the basis of the generic scheme described in Fig. 12.10 for the realization of the analog counter, but substituting the monostable stage with a proper digital circuit, so that the charging of capacitor C_S is started when a photon is detected and stopped by a reference signal synchronous with the emitter trigger. A pixel-level implementation of a TAC implementing A/D conversion using the replica concept illustrated in Fig. 12.11 has been proposed in [47]. The block diagram of the resulting pixel is shown in Fig. 12.15. The digital pulse generated by the SPAD ensemble is first processed by a logic block which generates a pulse whose time-width is proportional to the time occurred between the rising edge of the detected photon signal (Photon) and the falling edge of the clocking signal $STOP$. The basic building block used to generate the voltage ramp is replicated into three layout-matched structures: two of them are used alternatively to measure the photon arrival time, in a time interleaved way, while the third one is used to generate a reference voltage ramp used to implement the embedded ADC scheme. Finally, by using two memory banks, the A/D conversion can be done in parallel with the acquisition of the next frame, avoiding any dead time, and wasting of photons.

12.3.5 State-of-the-art Time-resolved CMOS SPAD Pixel-array

This paragraph will provide an overview of time-resolved sensors implementing SPAD-based pixel-array and on-chip processing on the basis of the structures

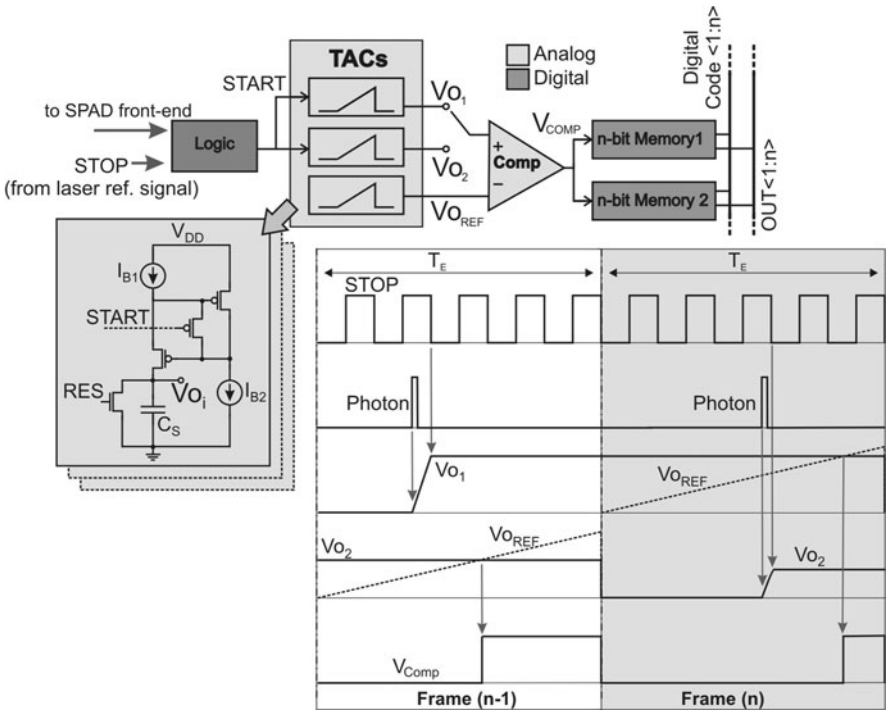

Fig. 12.15 TAC architecture for time-interleaved operation and A/D conversion, with relative timing diagram

previously addressed, summarizing the main technical specifications and achievable performance. Table shown in Fig. 12.16 provides a summary of the architecture choice, the adopted measuring technique, and the main features of the detectors, while table in Fig. 12.17 reports on the characteristics of the adopted TOF setup and the relative measurement performance.

12.4 Challenges and Future Perspectives

The performance of SPAD-based image sensors have been continuously improved in the last decades, mainly thanks to the efforts of a few research teams.[2] The realization of CMOS compatible device structures represented a major breakthrough in this research field, and the maturity of this kind of sensors is continuously

[2]Prof. S. Cova, Prof. F. Zappa (Politecnico di Milano), Prof. E. Charbon (TUDELFT, EPFL), Dr. R. K. Henderson (The University of Edinburgh).

	Ref [32]	Ref [42,43]	Ref [27]	Ref [52]	Ref [54]	Ref [53]
CMOS Technology	0.35 μm	0.35 μm	0.8 μm	0.35 μm	130 nm Imaging	130 nm Imaging
Pixel array:	60x48	64x1	64x1	128x128	32x32	32x32
Architecture	In-pixel digital counters (2x8-bit)	In-pixel digital counters (4x8-bit)	In-pixel TAC and analog counter	32 Column-level 10-bit TDC	In-pixel 10-bit TDC	In-pixel 10-bit TDC
Measuring technique	I-TOF modulated, 2 time-gated window x 2 frames	I-TOF pulsed and modulated, 4 time-gated window	I-TOF and D-TOF	D-TOF pulsed	D-TOF pulsed	-
Pixel pitch	85 μm	26 μm (Linear)	38 μm (Linear)	25 μm	50 μm	50 μm
Fill Factor	0.5%	34% (Linear)	13% (Linear)	6%	~2%	~2%
Power Dissipation	35 mW	600 mW@150kfps	1mW @ 10kHz (core only)	150 mW	90 mW (core only)	38 mW (core only)
Time resolution	-	-	~50 ps	98 ps	119 ps	52 ps
Jitter	-	-	ND	250 ps	185 ps (FWHM)	0.6 LSB
Uniformity	-	-	3% (after calibration)	ND	2 LSB (over the whole array)	8 LSB (over the whole array)

Fig. 12.16 Main technical specifications of SPAD-based sensors for time-resolved imaging. Sensors validated in TOF setup are shadowed. For sensors implementing TAC/TDC time measurement performance is given

increasing. However, there are still several challenges to be faced in order to let CMOS-SPADs enter consumer-applications market. Most relevant ones are listed in the following:

- *Sensitivity enhancement*: Two-dimensional pixel-array reported so far exhibit very low fill factor. This can be improved by means of microlens array [55] coupled on the top of the focal-plane array at the expense of extra fabrication costs. However, if the pixel pitch will be scaled down to a few microns, standard microlenses fabrication processes typically adopted by digital cameras can be used. Moreover, the SPAD quantum efficiency in the near infrared region (750–850 nm) is poor. It can be increased by using deeper junctions, thus requiring some modifications to the CMOS fabrication process. However, dedicated CMOS processes have been developed to fabricate digital cameras currently available on the market, and a similar strategy could be used once CMOS SPADs market will reach the critical production barrier.
- *Dark count rate*: Although DCR of state-of-the-art CMOS SPADs is continuously decreasing [38–41], the uniformity on large pixel array is still a problem. Typically, 10–20% of the devices exhibit DCR values one or two order of magnitudes higher than the average DCR, reducing the overall efficiency of the imager. This aspect is quite critical because mainly related to fabrication defects that can hardly be improved in standard fabrication processes.
- *Time resolution*: Best jitter noise performance of SPAD detectors is about 35 ps [39], while for in-pixel TAC/TDC it is in the order of 50 ps [53]. For many applications this is sufficient, but for submillimeter range detection at high frame

Ref[32]
3D TOF Imaging: demonstrated TOF System Parameters: - Emitter: 850-nm wavelength, modulated at 30MHz, 800mW optical power on a 50° field-of-view - Collection optics: F-number=1.4 - Distance range: 5m - Distance repeatability (σ): 3.8cm@2.4m with a 45ms exposure time.
Ref[42,43]
3D TOF Imaging: demonstrated TOF System Parameters: - Emitter: 650-nm wavelength, modulated at 100MHz or pulsed at 10MHz, 2mW optical power on a 10° field-of-view (light blade) - Collection optics: F-number=1.4 - Distance range: 4m (pulsed), 2.5m (modulated) - Distance repeatability (σ): 10cm@4m with a 10ms exposure time (pulsed@10MHz) - Distance repeatability (σ): <1.5cm@2m with a 10ms exposure time (modulated@100MHz)
Ref[52]
3D TOF Imaging: demonstrated TOF System Parameters: - Emitter: 635-nm wavelength, modulated at 40MH, 250mW peak power, 1mW average power - Collection optics: F-number=1.4 - Distance range: 4m (pulsed), 2.5m (modulated) - Distance repeatability (σ): 10cm@4m with a 10ms exposure time (pulsed@10MHz)
Ref[27]
3D TOF Imaging: demonstrated TOF System Parameters: - Emitter: 905-nm wavelength, pulsed at 100 ns with 0.1% duty cycle, 250mW average optical power on a 10° field-of-view - Distance repeatability (σ): 3cm@2m using D-TOF, 10cm@2m using I-TOF (20k pulses)

Fig. 12.17 Summary of the main performance of TOF range systems based on sensors reported in Fig. 12.16

rates more precision is needed, requiring new SPAD device structures and better performing read out and processing circuits.

- *Pixel dimensions*: The smallest pixel pitch reported so far for a SPAD detector with in-pixel time measuring circuitry is 50 μm [47, 53, 54]. To realize large pixel-array (VGA or million-pixel format) this value should be dramatically reduced. A possible solution is the use of simple analog schemes to implement photon counters or TACs, or adopting circuit-sharing strategies as is done in ultra-small pixels (<2 μm) as used in consumer applications such as cell phones. In ultra low light level applications, where the pixel activity is very low, column level processing implementations represent the best solution.
- *On-chip processing*: When TACs/TDCs are used at pixel level, and in case of large pixel-array implementations, the amount of information generated by the sensor is extremely high. To cope with this issue several solutions are possible,

such as the use of highly parallel read out architectures [46,52,53]. An innovative approach consists in the implementation of on-chip preprocessing of the data. An example of this, applied to the case of lifetime extraction can be found in [56], and in the near future similar implementations could be applied also to TOF measurements. Moreover, the availability of through-silicon via technology allows the realization of stacked multichips sensors, where sophisticated read out and processing interfaces could be realized.

- *Hybrid solutions*: As 3D silicon technologies will become more accessible and widely used, the realization of hybrid sensors, where the SPAD array can be developed using optimized fabrication processes while the read out and processing can be accommodated on a deep submicron CMOS layer, will be possible. By exploiting III–V technologies, the sensitivity range of the SPADs could be possibly extended into the IR region (1,500 nm) where high-power lasers can be used with lesser concern about eye-safety operation.

12.5 Conclusions

Time-resolved image sensors, i.e., imagers capable of detecting arrival times of impinging light signals with nanosecond or sub-nanosecond resolution, are becoming more and more important in many applications. Among them, TOF 3D cameras represent one of the markets having the largest possibilities of expansion thanks to the numerous sectors of exploitation of such a technology. The extra information provided by those sophisticated sensors, with respect to standard 2D cameras, is fundamental to acquire an accurate and reliable model of the scene under measurement, opening the way to new detection paradigms in the field of machine vision. Industrial control, cultural heritage preservation, automatic production, cinema industry, etc. are just a few examples of important sectors using 3D-imaging technology because its beginning, with minor concerns about its cost. However, there are huge markets, like gaming and advanced contactless user-interfaces for any kind of electronic tools used in everyday life, ready to fully exploit 3D-imaging as soon as low-cost and compact systems will be available. This goal can be achieved only by realizing fully-integrated CMOS detectors, which could benefit of the well know mass-production cost rules. This chapter focused on one specific 3D-measuring technique, the TOF, for its numerous advantages with respect to the above mentioned application areas, and provided an overview of state-of-the-art monolithic implementations of time-resolved SPAD-based pixel array. A summary of the most commonly used TOF measuring techniques was given and main noise limitations analyzed. A description of pixel architectures suitable for time-resolved single-photon detection was provided, and a selected choice of circuit implementations among those recently reported by the scientific community was presented and commented. Although SPADs have been well known since the 1980's and exploited in many scientific application areas, their implementation in two-dimensional pixel-arrays has been only recently reported, thanks to the

development of SPADs within deep-submicron CMOS technologies. This allowed also the implementation of smart front-end electronics aimed at the SPAD biasing and read-out, and more and more processing has been included on-chip. This new generation of CMOS-SPAD detectors, capable of single-photon light sensitivity, sub-nanosecond time resolution, processing and filtering of the data-stream and straightforward digital output finds its natural application mainly in niche research domains such as time-resolved fluorescent microscopy but could be, in time, adopted also by consumer application markets.

Acknowledgements The authors are grateful to Prof. G.-F. Dalla Betta and Dr. L. Pancheri for useful discussion and support in reviewing this chapter, and to all the researchers involved in the MEGAFRAME project (www.megaframe.eu) which contributed in developing many of the ideas here contained. The authors express gratitude in particular to Prof. E. Charbon and Dr. R. Henderson for their valuable contributions in the field of CMOS SPADs.

References

1. R. Schwarte, Principles of 3-D imaging technology, in *Handbook of Computer Vision and Applications*, ed. by B. Jähne, H. Haussecker, P. Geißler (Academic Press, New York, 1999)
2. B. Jähne, H. Haussecker, P. Geißler, *Handbook of Computer Vision and Applications*, vol. 1 (Academic Press, New York, 1999), pp. 479–482
3. P. Besl, Active optical range imaging sensors. Machine Vision Appl. **1**, 127–152 (1988)
4. M.-C. Amann, T.B.M. Lescure, R. Myllyla, M. Rioux, Laser ranging: a critical review of usual techniques for distance measurement. Opt. Eng. **40**, 10–19 (2001)
5. F. Blais, Review of 20 years of range sensor development. J. Elect. Imaging **13**(1), 231–243 (2004)
6. B. Hosticka, P. Seitz, A. Simoni, Optical time-of-flight sensors for solid-state 3D-vision. Encyclop. Sensors **7**, 259–289 (2006)
7. R. Dändliker, R. Thalmann, D. Prongué, Two-wavelength laser interferometry using super-heterodyne detection. Opt. Lett. **13**, 339–341 (1988)
8. R. Dändliker, Y. Salvadé, E. Zimmermann, Distance measurement by multiple-wavelength interferometry. J. Opt. **29**(3), 105–114 (1998)
9. Leica-geosystem website: www.leica-geosystems.com
10. Konicaminolta website: www.konicaminolta.com
11. J. Kramer, P. Seitz, H. Baltes, An inexpensive real-time 3-D camera with a smart optical sensor. Sensors Actuators **A31**, 241–244 (1992)
12. Y. Oike, M. Ikeda, K. Asada, Design and implementation of real-time 3-D image sensor with 640 × 480 pixel resolution. IEEE J. Solid-State Circuits **39**(4), 622–628 (2004)
13. Y. Oike, M. Ikeda, K. Asada, A 375 × 365 high-speed 3-D range-finding image sensor using row-parallel search architecture and multisampling technique. IEEE J. Solid-State Circuits **40**(2), 444–453 (2005)
14. Riegl website: www.riegl.com
15. Optech website: www.optech.ca
16. Mesa Imaging website: www.mesa-imaging.ch
17. PMD Technologies website: www.pmdtec.com
18. Optrima NV website: www.optrima.com
19. O. Elkhalili, O.M. Schrey, P. Mengel, M. Petermann, W. Brockherde, B.J. Hosticka, A 4 × 64 pixel CMOS image sensor for 3-D measurements applications. IEEE J. Solid-State Circuits **39**(7), 1208–1212 (2004)

20. G. Zach, M. Davidovic, H. Zimmermann, Extraneous-light resistant multipixel range sensor based on a low-power correlating pixel-circuit, in *Proceedings of the 35th IEEE European Solid-State Circuits Conference ESSCIRC09*, September 2009, pp. 236–239
21. D. Stoppa, L. Viarani, A. Simoni, L. Gonzo, M. Malfatti, G. Pedretti, 16 × 16-pixel range-finding CMOS image sensor, in *Proceedings of the 30th IEEE European Solid-State Circuits Conference ESSCIRC04*, September 2004, pp. 419–422
22. D.V. OConnor, D. Phillips, *Time-correlated Single-Photon Counting* (Academic Press, London, UK, 1984)
23. J.S. Massa, G.S. Buller, A.C. Walker, S. Cova, M. Umasuthan, A.M. Wallace, Time-of-flight optical ranging system based on time-correlated single-photon counting. Appl. Opt. **37**(31) (1998)
24. M.A. Albota et al., Three-dimensional imaging laser radars with Geiger-mode avalanche photodiode arrays. Lincoln Lab. J. **13**(2) (2002)
25. G.S. Buller, A.M. Wallace, Ranging and three-dimensional imaging using time-correlated single-photon counting and point-by-point acquisition. IEEE J. Selected Topics Quant. Electr. **13**(4) (2007)
26. A. Al-Azzawi, *Photonics: Principles and Practices* (CRC Press, Optical Science and Engineering, Boca Raton, 2006)
27. D. Stoppa, L. Pancheri, M. Scandiuzzo, L. Gonzo, G.-F. Dalla Betta, A. Simoni, A CMOS 3-D imager based on single-photon avalanche diode. IEEE Trans. Circuits Syst. I **54**(1), 4–12 (2007)
28. D. Stoppa, N. Massari, L. Pancheri, M. Malfatti, M. Perenzoni, L. Gonzo, An 80 × 60 range image sensor based on 10 μm 50MHz lock-in pixels in 0.18 μm CMOS, in *Proceedings of the International Conference on Solid-State Circuits (ISSCC10)*, February 2010, pp. 406–407
29. T. Spirig, P. Seitz, O. Vietze, F. Heitger, The lock-in CCD-two dimensional synchronous detection of light. IEEE J. Quantum Electron. **31**(9), 1705–1708 (1995)
30. R. Lange, P. Seitz, Solid-state time-of-flight range camera. IEEE J. Quantum Electron. **37**(3), 390–397 (2001)
31. W. van der Tempel, R. Grootjans, D. Van Nieuwenhove, M. Kuijk, A 1k-pixel 3D CMOS sensor, in *Proceedings of IEEE Sensors Conference*, October 2008, pp. 1000–1003
32. C. Niclass, C. Favi, T. Kluter, F. Monnier, E. Charbon, Single-photon synchronous detection. J. Solid-State Circuits **44**(7), 1977–1989 (2009)
33. K. Patorski et al., *Progress in Optics* (Elsevier Science Publisher/North-Holland, NY, 1989)
34. P. Koczyk, P. Wiewior, C. Radzewicz, Photon counting statistics – undergraduate experiment. Am. J. Phys. **64**(3), 240–245 (1996)
35. B.F. Aull, A.H. Loomis, D.J. Young, R.M. Heinrichs, B.J. Felton, P.J. Daniels, D.J. Landers, Geiger-mode avalanche photodiodes for three-dimensional imaging. Lincoln Lab. J. **13**, 335–350 (2002)
36. S. Donati, *Photodetectors: Devices Circuits and Applications* (Prentice Hall, Englewood, Cliffs, NJ, 1999)
37. G. Llosa et al., Novel silicon photomultipliers for PET applications. IEEE Trans. Nucl. Sci. **55**(3), 877–881 (2008)
38. L. Pancheri, D. Stoppa, Low-noise CMOS single-photon avalanche diodes with 32 ns dead time, in *Proceedings of IEEE European Solid-State Device Research Conference (ESSDERC07)*, Monaco, September 2007, pp. 362–365
39. F. Zappa et al., Monolithic CMOS detector module for photon counting and picosecond timing. Proc. ESSDERC, 341–344 (2004)
40. M. Gersbach et al., A low-noise single photon detector implemented in a 130 nm CMOS imaging process, Solid-State Electr. **53**(7), 803–808 (2009)
41. J.A. Richardson et al., Low dark count single-photon avalanche diode structure compatible with standard nanometer scale CMOS technology. IEEE Photonics Technol. Lett. **21**(14), 1020–1022 (2009)

42. L. Pancheri, D. Stoppa, A SPAD-based pixel linear array for high-speed time-gated fluorescence lifetime imaging, in *IEEE European Solid-State Circuits Conference (ESSCIRC'09)*, Athens, September 2009, pp. 428–431
43. L. Pancheri, D. Stoppa, A. Simoni, SPAD-based time-gated active pixel for 3D imaging applications, in *EOS Conference on Frontiers in Electronic Imaging*, Munich, 2009
44. B. Rae, C. Griffin, K. Muir, J. Girkin, E. Gu, D. Renshaw, E. Charbon, M.D. Dawson, R.K. Henderson, A microsystem for time-resolved fluorescence analysis using CMOS single-photon avalanche diodes and micro-LEDs, in *Proceedings of ISSCC*, February 2008, pp. 166–167
45. B. Rae, C. Griffin, J. McKendry, J. Girkin, H.X. Zhang, E. Gu, D. Renshaw, E. Charbon, M.D. Dawson, R. Henderson, CMOS driven micro-pixel LEDs integrated with single photon avalanche diodes for time resolved fluorescence measurements. J. Phys. D: Appl. Phys. **41** (2008)
46. E. Charbon, S. Donati, An ultrafast single-photon image diagnostics sensor with SPAD arrays for industrial and bio-applications, in *7th International Conference on Advanced Laser Technologies (ALT09)*, 26 October 2009
47. D. Stoppa, F. Borghetti, J. Richardson, R. Walker, L. Grant, R. Henderson, M. Gersbach, E. Charbon, A 32 × 32-pixel array with in-pixel photon counting and arrival time measurement in the analog domain, in *IEEE European Solid-State Circuits Conference (ESSCIRC'09)*, September 2009, pp. 204–207
48. J.-P. Jansson, A. Mäntyniemi, J. Kostamovaara, A CMOS time-to-digital converter with better than 10 ps single-shot precision. IEEE J. Solid-State Circuits **41**(6), 1286–1296 (2006)
49. A.S. Yousif, J.W. Haslett, A fine resolution TDC architecture for next generation PET imaging. IEEE Trans. Nucl. Sci. **54**(5), 1574–1582 (2007)
50. M. Lee, A.A. Abidi, A 9b, 1.25ps resolution coarse-fine time-to-digital converter in 90 nm CMOS that amplifies a time residue, in *Digest IEEE Symposium on VLSI Circuits*, June 2007, pp. 168–169
51. P. Chen, C.C. Chen, Y.S. Shen, A low-cost low-power CMOS time-to-digital converter based on pulse stretching. IEEE Trans. Nucl. Sci. **53**(4), 2215–2220 (2006)
52. C. Niclass, C. Favi, T. Kluter, M. Gersbach, E. Charbon, A 128 × 128 single-photon image sensor with column-level 10-bit time-to-digital converter array. IEEE J. Solid-State Circuits **43**(12), 2977–2989 (2008)
53. J. Richardson, R. Walker, L. Grant, E. Charbon, M. Gersbach, D. Stoppa, F. Borghetti, R.K. Henderson, A 32 × 32 50ps resolution 10 bit time to digital converter array in 130 nm CMOS for time correlated imaging, in *Proceedings of IEEE Custom Integrated Circuits Conference,* San Jose, USA, September 2009, pp. 77–80
54. M. Gersbach, Y. Maruyama, E. Labonne, J. Richardson, R. Walker, L. Grant, R. Henderson, F. Borghetti, D. Stoppa, E. Charbon, A parallel 32×32 time-to-digital converter array fabricated in a 130 nm imaging CMOS technology, in *IEEE European Solid-State Circuits Conference (ESSCIRC'09)*, September 2009, pp. 196–199
55. S. Donati, G. Martini, M. Norgia, Microconcentrators to recover fill-factor in image photodetectors with pixel on-board processing circuits. Opt. Exp. **15**, 18066–18074 (2007)
56. D. Li, R. Walker, J. Richardson, B. Rae, A. Nuts, D. Renshaw, R. Henderson, Hardware implementation and calibration of background noise for an integration-based fluorescence lifetime sensing algorithm. JOSA A **26**(4), 804–814 (2009)

Chapter 13
Single-Photon Imaging for Astronomy and Aerospace Applications

Pierre Magnan

Abstract Even with the use of powerful optical systems, astronomy and space imaging applications have often to deal with a very limited amount of photons. Despite major achievements brought to these applications by CCD and hybrid narrow gap semiconductor detectors technologies, and as illustrated by large focal planes used in ground based telescopes, the detection limit of classical solid state imaging devices still remains above the single photon level sensitivity required by several applications. Amongst these, attempts to circumvent the atmosphere turbulence effects limiting image quality have lead to the development of the so-called Lucky Imaging technique and wavefront sensors being part of adaptive optics systems. Both are based on imaging devices requiring the use of internal gain (electron multiplication) to lower the readout noise: EMCCDs and Avalanche Amplifiying pnCCDs using silicon and HgCdTe e-APDs for the infrared domain are considered.

Space based LIDARs, currently becoming more popular for altimetry and atmosphere characterization, also require single photon sensitivity level in order to reduce the emitted laser pulse energy and/or enhance the measurement accuracy. An example of a dedicated imaging device for atmospheric LIDAR is detailed before concluding on the current status of single photon devices for space applications.

13.1 Introduction

Astronomy and space instrument applications have been for a long time precursors in using electronic imaging devices, pushing continuously the corresponding technologies to achieve strong improvements in performance level, reliability, and confidence level in suitability for operational applications. The space telescope Hubble program in the eighties has, for example, led to the development of backside thinning and illumination techniques, applied initially to CCD and that are today commonly used in many scientific imaging or radiometric measurement applications but have also been introduced recently in consumer large volume applications. Although such techniques allow for large improvement in quantum

P. Seitz and A.J.P. Theuwissen (eds.), *Single-Photon Imaging*,
Springer Series in Optical Sciences 160, DOI 10.1007/978-3-642-18443-7_13,

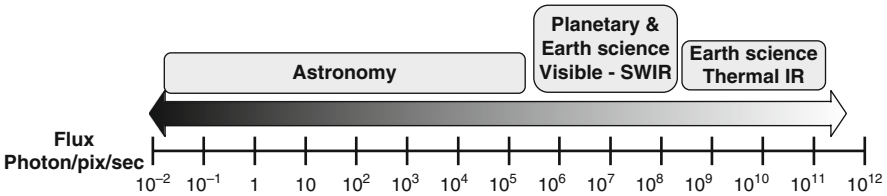

Fig. 13.1 Flux range covered by astronomy and space applications (from [2])

efficiency, one of the main challenges is often to face with the very low level of radiation emitted by the object of interest (e.g., a star), thus requiring dedicated accumulation imaging technique and ideally noiseless detectors. Along the move from the use as sensor coupled to an optical instrument, of initially the human eye, then the photographic plate and finally solid-state imaging devices, the spectral range of interest has broaden from the visible domain to potentially extend from the X-rays (and even the Gamma-rays) to the infrared (and even the microwave). A very comprehensive reference text regarding overall astronomy instrumentation aspects is the book by Mc Lean [1].

The flux range to be covered in astronomy and space instruments applications is very large, as illustrated in Fig. 13.1, explaining the large range of integration time, from tens of microseconds to thousands of seconds, and frame rates from a few millihertz to tens of kilohertz used for this kind of applications in contrast to classical video imaging applications. Astronomy, being either ground- or space-based, is clearly a domain where imaging devices often named "detectors" have often to deal with very few photons available.

The Earth atmosphere composition introduces severe limitations in the observation and measurement capabilities of universe objects using electronic detectors. Its transmission is very low, or even zero, for the largest part of the EM spectrum radiations including Gamma-rays, X-rays, most of the UV and the infrared (except for wavelength less than 25 μm) region, involving mandatory use in this wavelength range of space-based telescopes such as Hubble (HST), Spitzer, Kepler, or the future James Webb space telescope (JWST), which additionally do provide permanent observation capabilities. They also offer a way to avoid Earth atmosphere turbulence effects that degrade the resolution of sky objects image captured from ground and avoid the ground-based telescope to be operated at their theoretical diffraction limit λ/D (λ being the wavelength and D the aperture of the telescope). However, it is much easier and less costly to build large diameter on-ground telescope than equivalent space equipment. As a consequence, the space-based telescopes have reduced diameter (HST: 2.4 m, future JWST: 6.5 m) than their ground counterparts: 8.2 m for the Very Large Telescope (VLT) installed in Chile and 42 m for the future European Extremely Large Telescope (E-ELT), both from European Southern Observatory (ESO). In the common spectral range, the ground-based telescopes can feature a lower diffraction limit at a reduced cost, but only if able to counteract the turbulence effects. Proposal and implementation of techniques allowing for circumventing the effects of turbulence constitute a big challenge for the astronomical

scientific community that results in a very strong motivation for pushing imaging device sensitivity to the single-photon level. Some of the used techniques will be further detailed.

13.2 Scientific Detectors in Astronomy and Space Applications

Detectors in astronomy and space scientific applications are primarily used for imaging the focal plane of the telescope. Their possible functions are the measurement of both the irradiance and magnitude (see Appendix for definitions) of objects (photometry) and also their position in the field (astrometry), for measuring the value of spectral components created by dispersive medium (spectrometry) or interferometry measurements. In addition, they also operate for sensors used either for guiding/pointing the telescope (or the satellite itself) or for providing information about the distortion of the wavefronts for further correction.

Even if some other technologies are still in use, ground- and space-based astronomy imaging and spectroscopic applications are currently completely dominated, respectively:

- By charge coupled device (CCD) imagers for the UV, visible and near-infrared spectral domain (compatible with the Si bandgap energy of 1.12 eV at room temperature)
- By hybrid detectors for the short and mid wavelength infrared domain, made of infrared sensitive material (narrow-gap III–V or II–VI semiconductor) layer structured as pixel photodetectors (often photodiodes) bump bonded on a CMOS readout circuit (ROIC).

Both types of device, that will be shortly reviewed, share the internal photoelectric effect as a photon detection principle and some common requirements, a very large physical size of the devices associated with a good flatness, a very good characteristic uniformity, a low dark current, a high level of quantum efficiency, and a minimum read noise. Scientific ground-based astronomy has especially evolved along time in the direction of larger telescopes and also in the use of larger focal plane arrays (FPAs) being made of precise assembly of elementary devices as mosaics.

Most of the detectors used in ground-based astronomy are custom design in terms of pixel pitch, geometrical format, readout rate, and spectral response optimization to accommodate the specific needs of the telescope mission (what is looked for) and the field of view.

13.2.1 Scientific CCDs

CCDs have been adopted for a long time as first choice detectors in astronomy, for the visible and the near-infrared spectral domain, as the most suited technology for the faint objects of interest involving very long exposure, often ranging from 1 s to

Table 13.1 Characteristics of large mosaics dedicated to survey telescope instruments

	MegaCam (2000) [3]	OmegaCAM (2011) [4]	Pan-STARRS [5]	LSST Project [6]
Telescope	CFHT-Mauna Kea-Hawai (3.6 m)	VST Paranal VLT survey telescope(ESO) (2.6 m)	IFA-Hawai Univ. (Mauna Kea/Haleakala) (1.8 m)	Large synoptic survey telescope Hawai (8.4 m)
FPA size	340 Mpixels	256 Mpixels	1.4 Gpixels	3.2 Gpixels
FPA structure	$36 \times (2K \times 4.5K)$ pixels	$32 \times (2K \times 4K)$ pixels	$256 \times (0.6K \times 0.6K)$ pixels	$189 \times (4k \times 4K)$ pixels
Pixel pitch	13.5 μm	15 μm	10 μm	10 μm
Dimension	$30 \times 30\,cm^2$	26×26 cm	40×40 cm	64 cm diameter
Spectral range	350–1,000 nm	320–1,014 nm	380–1,050 nm	300–1,100 nm
Readout times	30 s (2 readout channels/CCD)	20 s (2 readout channels/CCD	3 s (128 readout channels	2 s (3,024 channels @ 500 kpix/s)
Readout noise	< 10 e- rms	< 5 e- rms	5 e- rms	< 5 e- rms
Operating temp.	−120°C,	−120°C	−75°C	−100°C
Field of view	$1 \times 1(\text{degree})^2$	$1 \times 1\ (\text{degree})^2$	$3 \times 3\ (\text{degree})^2$	$3.5 \times 3.5\ (\text{degree})^2$
Resolution	0.187 arcsec/pix	0.21 arcsec/pix	0.26 arcsec/pix	0.2 arcsec/pix

several thousand seconds. For example, the exposure time of the ground-based Pan-STARRS astronomy camera in the range of 60 s is forecast to allow for reaching a limiting object visible magnitude of 24.

Their use has evolved from the single device (in $2k \times 2k$ or $2k \times 4k$ format until the end of the 1990s) to mosaics being larger and larger size along time.

Table 13.1 summarizes some characteristics for the large mosaics used (MegaCAM and OmegaCAM) or to being developed for the survey telescope instruments, whose role is to map the sky and search in wide fields for various objects (distant clusters of galaxies, asteroids, quasars, structure and content of the Milky Way supernovae, variable stars, microlensing events, extrasolar occultations) performing both astrometry and photometry.

The main features granted to CCD technology, as a result of intensive efforts [7], are a quasiperfect transfer efficiency, a nearly ideal quantum efficiency (around 90% in large spectral range) achieved thanks to backside illumination associated with optimized antireflective coatings, suitable silicon thickness, and deep depleted or fully depleted high resistivity substrate [8], very low dark current reduced by operation at low temperature (around 1 e-/pix/h at −120°C), and low readout noise (2 e- rms can be achieved on the best devices) but at limited pixel rate (a few tens kpix/s). In addition, wafer scale size devices can be fabricated [9]. However, despite significant improvements in output amplifier design (some using JFET rather than MOSFET) and conversion gain ($>5\,\mu V/e$-), this noise figure degrades severely when increasing the pixel rate (reaching more than 10 e- rms for pixel rate of few MHz), leading to a high number of outputs channels (e.g., 16 outputs/CCD planned for LSST).

An architectural innovation [10] has been brought by the orthogonal transfer CCD (OTCCD), which is a CCD device capable of shifting charge in X or Y directions during exposure, then transferring them in traditional direction for readout. It was developed for compensating the image motion across a sensor and avoiding image burring in case of motion of the scene or of the sensor platform during image integration. One of its applications is the "tip-tilt" electronic correction to improve image resolution. It is the base device for the Gigapixels PAN-STARRS camera mentioned in Table 13.1.

In the space instrument domain, the most ambitious CCD focal plane is the very large (935 Mpixels) one developed for the GAIA "all-sky survey mission" of the European Space Agency (ESA), dedicated to the mapping of the 3D position and velocity of more than 1,000 millions of objects down to the 20th magnitude in our Galaxy and beyond [11]. GAIA, located at L2 Lissajous point of the Sun–Earth/Moon system, will perform astrometry, photometry, and spectrometry. It includes a sensitive area of about $0.5\,\mathrm{m}^2$ built with 106 large scientific back-illuminated CCDs, featuring $4{,}500 \times 1{,}966$ pixels of $10\,\mu\mathrm{m} \times 30\,\mu\mathrm{m}$ pixels, operated at 160 K. The angular accuracy is $25\,\mu$as at magnitude 15. The pixels are operated in time delay integration (TDI) mode [12], meaning that the charges are transferred, accumulated, and readout at exactly the same rate as the scene drift by, fixed by the satellite spin rate. The TDI period is $982.8\,\mu$s. Stars are crossing the focal plane at a constant speed and star images are progressively built up as objects cross the CCDs. The maximum level of signal per pixel remains very low, in the worst case 80 e- and 90 e- for, respectively, B3V and G2V, two stars with visual magnitude $m_\mathrm{v} = 20$ [13].

A specific problem of detectors, especially CCDs, for space application is related to their behavior in the space radiative environment. The radiation-induced damage (mainly by protons) can alter transfer efficiency, increase severely the dark current and noise and introduce RTS noise behavior, degrading the overall performances [14]. Mitigation techniques have to be employed such as in GAIA detectors, supplementary buried channels ("notch") and charge injection techniques to counteract the effect of radiation-induced traps [15].

13.2.1.1 Hybrid Detectors for Infrared

Photodiodes realized in narrow-gap III–V (InSb, $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}$) or II–VI semiconductor ($\mathrm{Hg}_{1-x}\mathrm{Cd}_x\mathrm{Te}$) are typically used as photodetectors in the short (SWIR) and mid wavelength infrared (MWIR) domain, Arsenic-doped silicon Si:As (as "impurity band conduction" detectors) being used for long wavelength infrared up to $25\,\mu$m. For some ternary material, the stoichiometry can allow to tune the bandgap energy E_g and the cutoff wavelength λ_C ($\lambda_\mathrm{C} = 1.238/E_\mathrm{g}$ with λ_C in μm and E_g in eV). This is the case for the current key material $\mathrm{Hg}_{1-x}\mathrm{Cd}_x$ Te widely used for infrared detection [16, 17], as illustrated in Fig. 13.2. To reduce the dark current, this material has very often to be operated at cryogenic temperature often around 80 K.

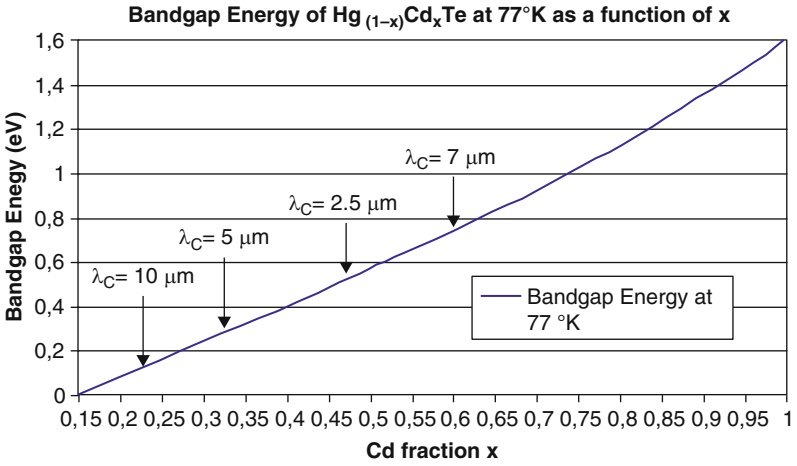

Fig. 13.2 Bandgap energy dependence of $Hg_{1-x}Cd_xTe$ vs stoichiometry (from [17])

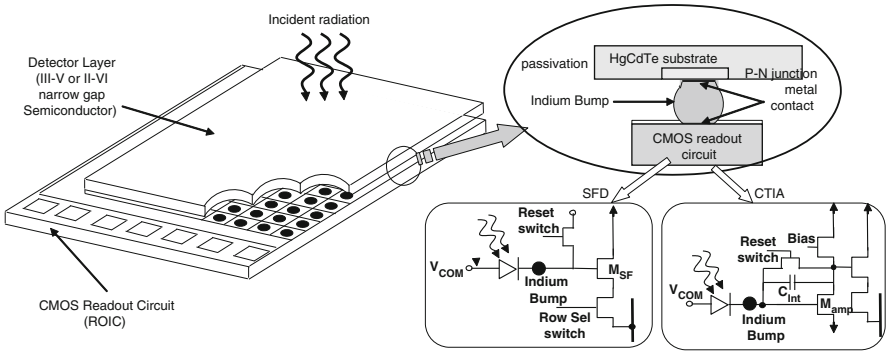

Fig. 13.3 Organization of hybrid infrared detectors and pixel readout circuit

Most of the HgCdTe large devices are currently produced by molecular beam epitaxy (MBE) on CdZnTe substrate, although liquid phase epitaxy (LPE) can also be used. The narrow-gap semiconductor photodiodes are connected to a CMOS readout circuit by indium bump, one per pixel, in a hybrid configuration as described in Fig. 13.3. The ROIC is often called Multiplexer as its main purpose is to allow for sequential pixel signal readout, thanks to a semi-parallel line sampling organization, very similar to the one of visible monolithic CMOS imagers. Photodiode charge is read out either in the voltage domain using a source-follower metal oxide semiconductor (MOS) transistor (SFD: source-follower per detector configuration) or in the charge domain using a capacitive transimpedance amplifier (CTIA) [2].

Infrared focal plane technology has evolved rapidly in the past two decades. The detector format has grown from single pixel detectors to 4k × 4k pixel arrays, and a mosaic of 4 × 4 2k × 2k pixel arrays, 18 μm pixel pitch, 2.5 μm cutoff wavelength

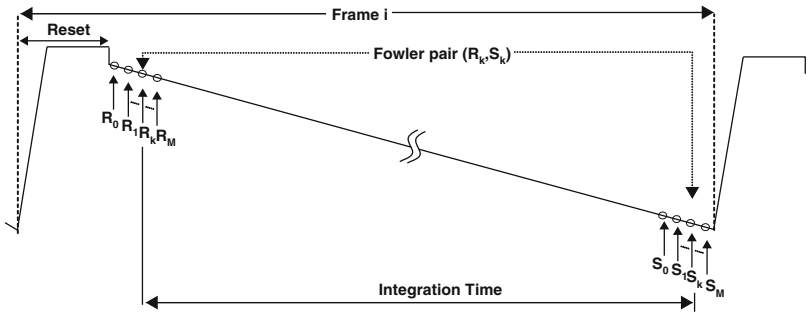

Fig. 13.4 Multiple correlated sampling or "Fowler sampling" for noise reduction in IRFPA readout

are already in operation at ground-based telescopes [18]. The quantum efficiency of these arrays is well above 80% over the whole spectral range [19]. The substrate can be chemically removed, offering extension of the sensitivity down to the visible spectral range and avoiding the fringing phenomenon (resulting from interferences of multiple reflections between the parallel surfaces of the detector substrate), which causes problems in high-resolution spectrographs.

The dark current, dependent on the cutoff wavelength setting, is commonly around 1 e/s/pixel at 80 K but level below 10^{-2} e/s/pixel has been achieved [19]. The read noise with correlated double sampling (CDS) for slow scan is typically in the range of 10–20 e- rms, but recent improvements have brought this level close to 7 e- rms [20].

For relatively long integration times commonly used in astronomical narrow band imaging or spectroscopy, the readout noise can be further efficiently reduced by multiple correlated sampling (MCS) technique, also called "Fowler sampling," proposed by Fowler and Gatley in 1990 [21] and described in Fig. 13.4. Immediately after performing the reset of the pixels, the entire array is repeatedly read out M times nondestructively and the corresponding values for each pixel, stored in a frame memory, are accumulated. At the end of integration, the array is read out again nondestructively M times, the corresponding values for each pixel being accumulated on a pixel-by-pixel basis.

The signal S_i and noise N_i for the frame i are, respectively, given by [22]:

$$S_i = \frac{\sum_{k=1}^{M} R_k - \sum_{k=1}^{M} S_k}{M} \quad \text{and} \quad N_i = \sqrt{\frac{N_{\mathrm{RO}}}{M}}.$$

N_{RO} being the noise associated to single sample readout.

This readout scheme features both the elimination of reset noise and the reduction of the readout noise by a factor $\sqrt{M}$. Each pair $(R_{\mathrm{k}}, S_{\mathrm{k}})$ is named Fowler Pair,

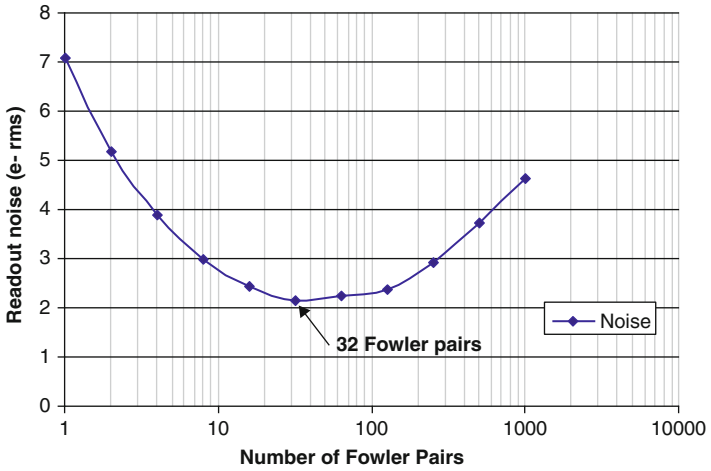

Fig. 13.5 Fowler sampling readout for noise reduction [measurement data from Finger [20]]

the readout scheme being designated for a dedicated value of M by "Fowler-M readout."

Figure 13.5 illustrates the noise reduction obtained from a HgCdTe 2k × 2k pixels array by Fowler sampling [20]. For 32 Fowler pairs ("Fowler-32 readout"), the previous 7 e- rms noise figure translates into 2.4 e- rms.

Two examples of in-operation large Infrared FPAs are illustrated in Fig. 13.6 [18]:

- The Visible and Infrared Survey Telescope for Astronomy (VISTA) infrared camera, a large survey camera built for a dedicated 4 m ESO telescope on Paranal (Chile).[1] The camera, operating in astronomical Z, Y, J, H, and K bands with broad band filters (see Appendix for definitions), has a field of view of 1.65 degrees with a pixel scale of 0.34 arcsec. The focal plane of VISTA has 64 Mpixels and is the largest infrared focal plane in operation at any telescope. It consists of sixteen 2k × 2k HgCdTe LPE arrays with 2.5 μm cutoff. The mean quantum efficiencies of the 16 detectors have been measured as $(J, H, K) = (90, 96, 92)\%$, the mean dark current is 1.2 e/pixel/s and the mean readout noise is 20.9 e- rms with a well capacity between 110,000 and 180,000 electrons.
- The HAWK-I, the High-Acuity Wide-field K-band Imager for ESO's VLT, equipped with a mosaic of four 2K×2K arrays (HgCdTe MBE arrays with 2.5 μm cutoff) covering the wavelength range from 0.9 to 2.4 μm and having a field of view of $7.5' \times 7.5'$ with 0.1 arcsec per pixel[2]. Each array has 32 video channels.

[1] http://www.eso.org/sci/facilities/paranal/instruments/vista/.

[2] http://www.eso.org/sci/facilities/paranal/instruments/hawki/.

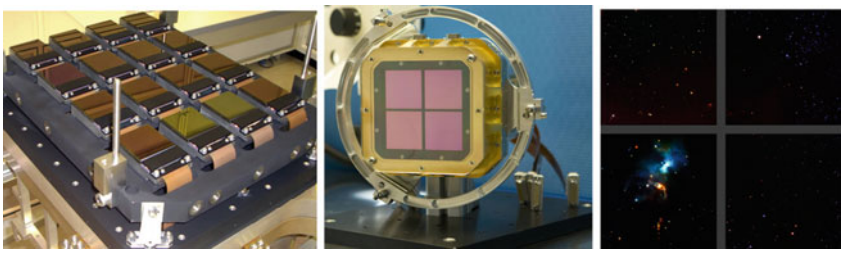

Fig. 13.6 Some infrared detector mosaics used in ESO Pananal Telescopes instruments. *Left*: the VISTA mosaics 4×4 2 K $\times$ 2 K pixels[1]; *center*: the Hawk-I mosaic 2×2 2K $\times$ 2K pixels array[2]; *right*: in bottom left quadrant, the Serpens star-forming region imaged during the commissioning of HAWK-I detectors. For all images, *credit*: ESO

13.3 Imaging Through the Atmosphere

The Earth atmosphere is associated with turbulence effects that distort the wavefronts entering the optical systems, thus degrading the image quality through blurring and twinkling of observed object during exposures even for the top observation locations (Chile, Hawaii or Canaries islands).

Turbulence effects due to atmosphere originate from the perturbations of radiation wavefronts above the telescope (being no more planar) caused by, according to the Kolmogorov theory, the random refractive index variations of stratified air layers (caused by temperature variations) in the turbulent atmosphere. Even if it introduces changes in both wavefront amplitude (leading to scintillation effects) and phase, most of the distortion results from random wavefront phase variations. Short exposure images show some speckle pattern, continuously evolving, as illustrated in Fig. 13.7 by a sequence of images from Gamma Perseus observed with the 4.2 m W. Herschel Telescope in La Palma (Canaria islands)[3]. Even with observatories being installed on high altitude sites chosen for their air stability, the native resolution of images is limited by atmospheric seeing at about 0.5 arcsec [23].

The Strehl ratio, defined as the ratio of the peak intensity of the point spread function (PSF) measured through atmosphere to that expected from a diffraction-limited perfect telescope (no seeing), is commonly used by astronomers to define a quantitative measure of image quality through atmosphere (always less than unity, it can be as low as 0.1 and even 0.01).

The atmosphere coherence time τ_0, timescale on which images change significantly, has been measured in the order of a few milliseconds or tens of milliseconds [24].

The random wavefront phase variations can be characterized by the mean-square phase difference $D_\Phi(\Delta r)$ between two points in the wavefront separated by a distance Δr. $D_\Phi(\Delta r)$ is given, for an observation wavelength λ by [25]:

[3] http://www.photonics.ic.ac.uk/speckle/movies.html.

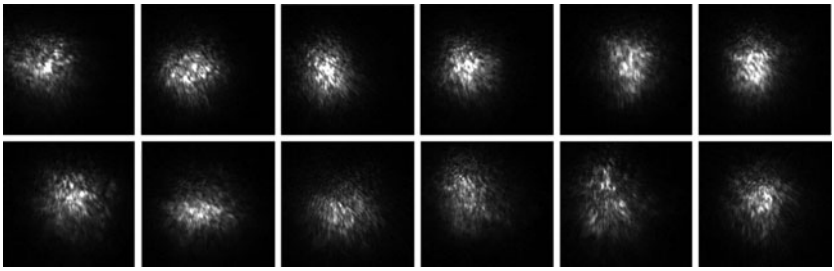

Fig. 13.7 Sequence of Gamma Perseus 25 ms exposure time frames at 570 nm wavelength, observed with W. Herschel 4.2 m Telescope in La Palma (extracted from an original full sequence from Photonics Group at Imperial College[3])

$$D_\Phi(\Delta r) = \langle(\Phi(r) - \Phi(r+\Delta r)^2\rangle = 6.88\,(\Delta r\,/r_0(\lambda))^{5/3} \quad \text{with}$$

$$r_0(\lambda) = \left(\frac{\lambda}{\lambda_0}\right)^{6/5}.r_0(\lambda_0).$$

The quantity $r_0 = r_0(\lambda_0)$ is known as Fried parameter, being the size of an aperture over which the rms phase error is approximately equal to 1 radian, conventionally measured at a wavelength of $\lambda_0 = 550$ nm. r_0 is the diameter of the area where the wavefront perturbations are not significant. It is the aperture defined as having the same resolution as the diffraction-limited aperture without turbulence thus providing a measure of the strength of the seeing distortions. Additionally, the isoplanetic angle θ_0 is defined as the radius of a circle in the sky over which the atmospheric disturbances can be considered as identical. A good approximation of θ_0, being a function of λ is [26]:

$$\theta_0(\lambda) = 0.314\frac{r_0(\lambda)}{H},$$

where H is the average distance of the seeing layer.

To achieve a sufficient signal-to-noise ratio, most of the astronomical images are obtained from long exposures. Long exposure images are the sum of a large number of short exposure images. Because the short-exposure PSF is highly variable, both in structure and in centroid position, the summed long-exposure PSF is highly blurred compared to the diffraction limits, by a factor of 5–15 on a 2.5 m telescope at good sites. For a telescope with aperture D much larger than r_0, the effective resolution without correction, specified by the seeing limited full width at half maximum (FWHM) of the long exposure PSF is given by [25]:

$$\text{FWHM}(\lambda) = 0.98\frac{\lambda}{r_0}.$$

For the best observation sites, the Fried parameter is usually in a range of 10–20 cm, limiting the seeing to 0.5 arcsec range, far from the diffraction limit of the telescope [27] involving mandatory use of turbulence correction methods.

Methods used to allow imaging techniques for counteracting the effects of turbulence can either be based on:

- Performing the measurements at the same timescale as atmospheric coherence involving multiple short exposures. These techniques are categorized as passive methods, which include Lucky Imaging (that will be further detailed), Speckle Interferometry, shift-and-add method [28]. They involve for the main FPA imaging device, a fast readout rate and a very low noise that make the EMCCD very attractive for the wavelength less that 1.1 μm.
- Active correction that will, in real time, eliminate the wavefront distortion before being sent to the main FPA. The most well-known technique is adaptive optics (AO), which generally involves, out the main FPA, a dedicated imaging device called wavefront sensor (WFS) and of a deformable mirror to perform the active correction in a feedback loop configuration.

13.4 Lucky Imaging Technique

The Lucky Imaging technique is based on the record of a sequence of several thousand short exposures and a selection process among them to retain only the few best corresponding to near-perfect seeing moments to provide the final image (a reference star is checked in each frame for image quality). It has been formalized by Fried in 1978 [29] by defining the probability for an exposure to be nearly diffraction limited, the instantaneous wavefront distortion being negligible across the telescope aperture during this exposure, as:

$$P \simeq 5.6 \exp(-0.1557(D/r_0)^2).$$

It typically leads to the selection of 1–10% of the exposure based on their Strehl ratio. The final image is constructed by shifting the selected frames to align them and finally accumulate them.

The exposure times used should be short enough to freeze any atmospheric variations typically, being in the range of few milliseconds or few tens of milliseconds, and the frame rate over 30 fps (up to 100 fps), leading to singlephoton sensitivity level for the fast operating imaging device.

Although Lucky Imaging has been envisaged for a long time and experimented with both conventional CCD and intensified photon-counting detectors, it is only with the achievement of the ultranoise EMCCD that this technique, thanks to the intensive work of Cambridge team (Mackay, Baldwin, Tubb and Law) [30, 31], could demonstrate its real benefits using LuckyCam EMCCD camera in imaging through atmosphere at least in the I spectral band (850 nm) [32] and also in speckle spectrometry [27].

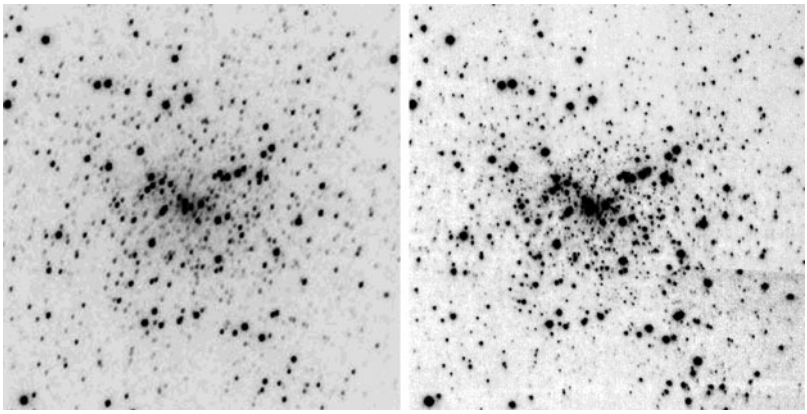

Fig. 13.8 Comparison from StanCam (astronomy 1 Mpixels conventional CCD camera) and LuckyCam images in wide field survey. The *left image* above was taken at La Palma with StanCam, the Nordic Optical Telescope's resident CCD camera, in I-band, a few minutes before the Lucky Imaging mosaic, with 0.6 arcsec seeing, 0.16 arcsec/pix. The *right image* is a mosaic of nine fields taken at 12 Hz, 10% selection giving a 55 × 55 arcsec FOV. Limiting magnitude is about 18.7 in I-band. *Credit*: Mackay and the Lucky Imaging Team, Institute of Astronomy, University of Cambridge[4]

Astronomy cameras using conventional CCD operated at several tens or more frames per second have a readout noise of at least tens electrons. As the process of adding selected frames multiplies this readout noise by the square root of the number of frames, Lucky Imaging technique cannot in that case be used without an unacceptable noise penalty. The main advantage of the EMCCD is its subelectron read noise even at high frame rate (see Chap. 6) When working at low photon rates (less than 1 photon/pixel/frame), it is possible to operate EMCCD devices in photon counting mode (the output signal being thresholded to a single value) thus allowing for the excess noise factor to be reduced to unity and for faint target to be observed [33].

Resolutions in the range 0.1–0.2 arcsec can be obtained routinely under relatively good conditions on a 2.5 m telescope working in I band (850 nm).

The Lucky Imaging website[4] provides actual movies of various observations exposure sequences (Fig. 13.8).

It is important to note that the use of EMCCD for extreme faint flux imaging has required the development of dedicated controllers [34, 35] especially to reduce the clock-induced charge (CIC) described in Sect. 6.5.2 to levels as low as 0.002 e-/pixel/frame.

[4] http://www.ast.cam.ac.uk/~optics/Lucky_Web_Site/.

13.5 Adaptive Optics

13.5.1 Principles

Adaptive optics (AO) for atmospheric seeing compensation proposed in 1953 by Babcock [36] and first used for military application have been widely adopted by the main ground-based observatories. These have setup important R&D programs becoming a driver for new image sensor architectures.

The main components of AO system, illustrated in Fig. 13.9, include a tip/tilt mirror correcting the large global phase slopes (that cannot be corrected with the limited dynamic range of the deformable mirror), a deformable mirror (piezo stacks or bimorph mirrors) that corrects the distortions in real time to reconstruct the wavefront, a wavefront sensor that measures the phase distortions using the light from a reference star (or "guide star" located close to the object in the field of view) and a processor executing the algorithms to generate the commands sent to the

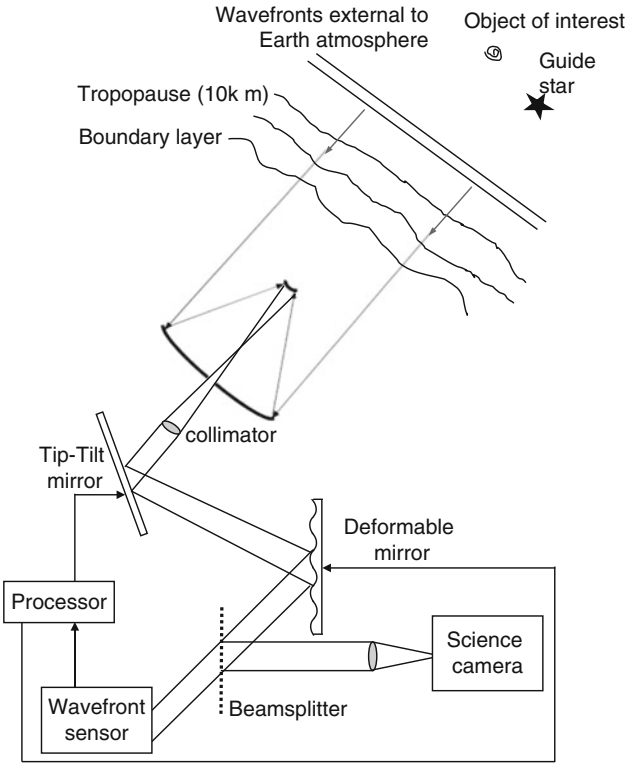

Fig. 13.9 General architecture of an adaptive optics system

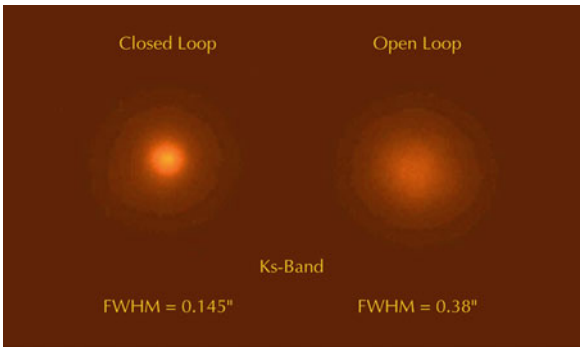

Fig. 13.10 AO-corrected image of a 17.5-magnitude star obtained with SINFONI AO. *Credit*: ESO

deformable mirror to produce a flat wavefront at the WFS. This AO system produces a corrected beam that is imaged by the visible or IR science camera.

Figure 13.10 illustrates the resolution improvement brought by using AO systems in the SINFONI instrument (ESO) on a very faint star ($M_v = 17.5$).

The guide star can be either a natural one (NGS) emitting light in the visible or SWIR domain, or an artificial laser guide star (LGS) created, for example, by excitation by a laser beam of sodium atoms in the so-called sodium layer, located above at around 90 km in altitude, at a wavelength of 589 nm.

Multiple sensing approaches are used to estimate the wavefront distortions, but the most popular include the commonly used Shack–Hartmann, the curvature [37] and pyramid [38] sensors, all requiring an imaging device to map the intensity distribution corresponding to the spatial sampling of phase variations along the wavefront.

The Shack–Hartmann sensor, illustrated in Fig. 13.11, is based on the measurement of the first spatial derivative of the wavefront over a set of subapertures defined by a lenslet array, producing subimages formed on the same imaging detector. The intensity distribution in each subaperture is used to estimate the wavefront's local gradient map by calculating the centroid of the subaperture spots (that would for an ideal plane wavefront, be located on a regular grid defined by the lenslet array geometry).

The number of subapertures determines the spatial sampling of the wavefront and the order of the correction, typically ranging from 4 (2×2) for low order systems used primarily for tip/tilt correction with NGS to up to 1,600 (40×0) subapertures for high order wave front correction systems using LGS or NGS currently developed for ESO instruments [39] and even more for future systems for the ELTs. The current number of pixels per subaperture varies from to basically 2×2 up to 6×6 for the ESO-VLT AOF and SPHERE Shack–Hartmann WFS, but will reach 20×20 for future systems to sample the spot elongation.

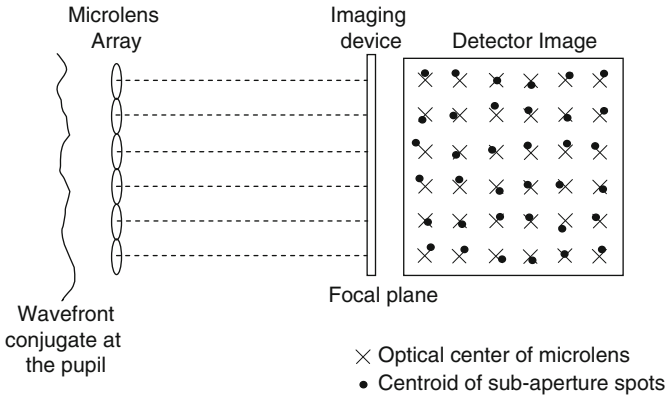

Fig. 13.11 Schematic diagram of Shack–Hartmann wavefront sensor principle

13.5.2 Wavefront Sensor Requirements and Detector Implementations

The performance of the WFS that measures the residual wavefront distortion affects strongly the quality of the overall AO system.

As the timescale of atmospheric distortion changes is in the range of some tens of milliseconds, AO correction rate is typically around 500 Hz (over 1 kHz for planned future systems) [39]. The WFS must measure the wavefront distortion in a very short time and exposure times can be as short as 1 ms. The number of available photons in each frame is severely limited, especially for the distant and faint objects that are of interest with large telescopes. Thus, to reach the photon noise of the WFS signal as being a limit to the performance of an AO system, the readout noise of the WFS has to be minimum, and ideally, completely eliminated. The basic requirements for AO WFS include both high quantum efficiency and a low readout noise.

In addition, a fast readout rate is required to reduce the readout time so as to minimize phase lag in the AO system, the minimum being ideally, the time needed for the WFS exposure. Minimum frame rates of 1 kHz have to be used. The combination of a very low readout noise and high frame rate is very difficult to achieve, even for the relatively reduced format required for WFS detector (compared to science one) typically ranging from 100×100 to less than 250×250 pixels [40]. It makes the use of avalanche photodiodes [APDs, single-photon avalanche diode (SPAD)] attractive. However, using silicon, these devices have been up to now manufactured in very small arrays or dedicated monolithic arrangement of single devices, as demonstrated by the 60 elements circularly arranged SPADA array for AO curvature WFS system used in the MACAO ESO system [41].

These above requirements have mostly led to several custom CCD devices dedicated to WFS (see Table 1 in [39]), based on multiple readout outputs and split frame transfer, that are including an image area split into two, each half being

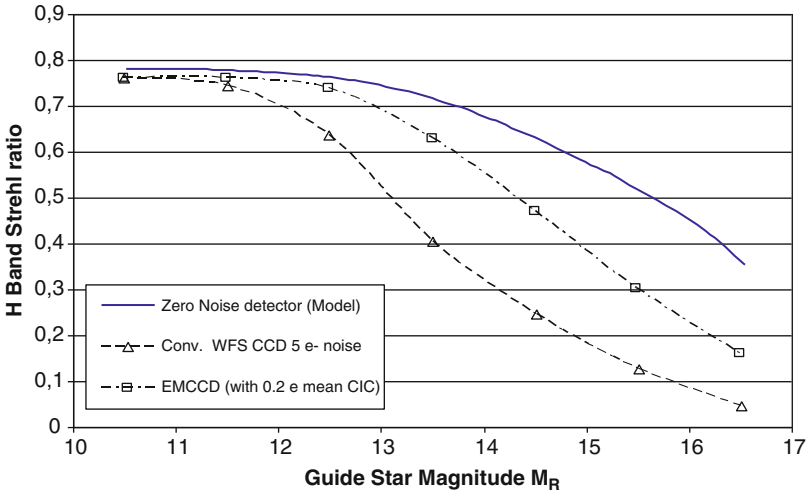

Fig. 13.12 Expected gains from using an EMCDD based WFS for W. Herschel Telescope at La Palma (from [42])

associated with a light shielded storage area). If one considers a 128 × 128 pixels format CCD with 16 output channels, achieving a kHz frame rate requires a pixel rate per channel of 1 MHz. Referring to Fig. 6.2 in Chap. 6, this translates into noise figures between 5 and 10 e- rms, limiting the magnitude of the guide star to be used. This gives a very clear advantage for WFS applications to subelectron noise devices such as EMCCD, as demonstrated by S. Tulloch [42] for the NOAMI AO instrument of Isaac Newton Group of Telescope (La Palma-Canaria) and illustrated in Fig. 13.12. This has been confirmed by Petit et al. for the AO system of the SPHERE Planet Finder instrument for the ESO VLT [43].

As a consequence, EMCCD devices, which can provide a peak QE of 90% as their conventional counterpart do thanks to thinning and backside illumination, have received a lot of interest for AO and WFS application as a way to reach subelectron noise with frame rates over 1 kHz.

An illustrative work is the development of the (EM) CCD 220 device dedicated to WFS in a very significant effort performed by ESO and OPTICON-JRA2 consortium. This device, described in Fig. 13.13, is a split frame transfer EMCCD featuring 240 × 240 back illuminated pixels ($24 \times 24\,\mu m^2$).

The device features eight output channels to achieve the required target frame rate over 1,200 fps. Eight associated electron-multiplying gain registers operating at more than 13 Mpixels/s have enabled to achieve a true sub-electron read noise, the measured mean value for the eight channels being 0.52 e- rms at frame rates of 1,300 fps as reported in [46]. The total transfer time from image to the optically shielded storage area is 18 μs thanks to a 2-phase metal-buttressed parallel clock structures enabling line shifts as fast as 7 Mlines/s. The QE reaches a peak value of 94% obtained with a silicon thickness of 13 μm. A deep depletion version will

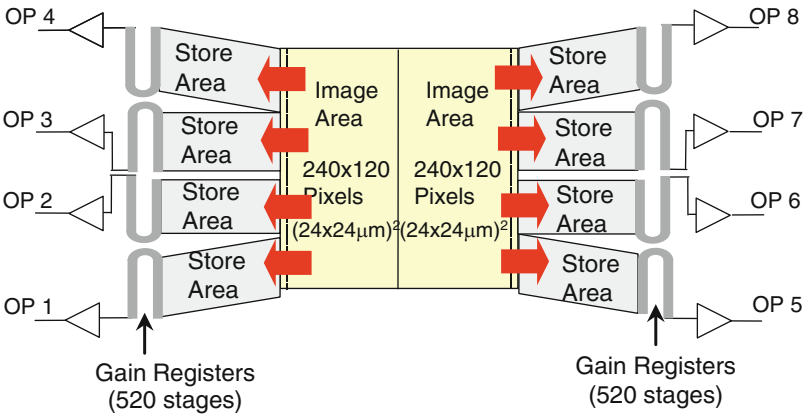

Fig. 13.13 Architecture of an up-to-date EMCCD (CCD220) dedicated to WSF to be used in AOF and SPERE ESO instruments

allow extending this high response level in the red and NIR part of the spectrum. The device is operated at low temperature (−40°C) thanks to a dedicated package integrating a Peltier thermoelectric cooler (sealed and filled with Krypton gas) to achieve a mean dark signal measured as low as 0.009 e-/pix/frame (i.e. 62 e-/frame).

These very impressive results could be obtained thanks to the joint development of a low-noise state of the art controller, called OCam, dedicated to EMCCD driving [44].

Another interesting device for WFS under development is the Avalanche pnCCD (AApnCCD). It is based on pnCCD, a special fully depleted nonconventional CCD device developed initially by the Max Planck Institutes for X-ray detection and successfully used in the XMM/Newton X-ray space telescope (ESA mission). This device does not use MOS gates to perform charge transfer but pn-junctions sequentially and appropriately biased in a three-phase configuration in combination with back-bias, using sideward depletion to move the potential valley [45]. The collection and transfer occur in a buried channel located at a depth of few microns from the surface, the device using a thick (450 μm) high resistivity substrate and being back illuminated. The array is organized in a semi-parallel structure, the charge being transferred along the columns. These are each terminated by a FET transistor buffer for providing column signal to external wire bonded multiplexers. A version of this architecture organized as a split frame transfer allows for a very fast and low-noise readout, and it has been used in a 264 × 264 pixels, 48 μm pitch array allowing for a frame rate of 1,100 Hz to be achieved with a noise of 2.3 e- rms at −55°C.

To further reduce the noise at the sub-electron range, the AApnCCD structure has been proposed, using an avalanche cell (diode) as readout node at the end of the transfer path in each column, operating in Geiger mode with a high value on-chip quenching resistance, before further buffering [46]. A high avalanche field is

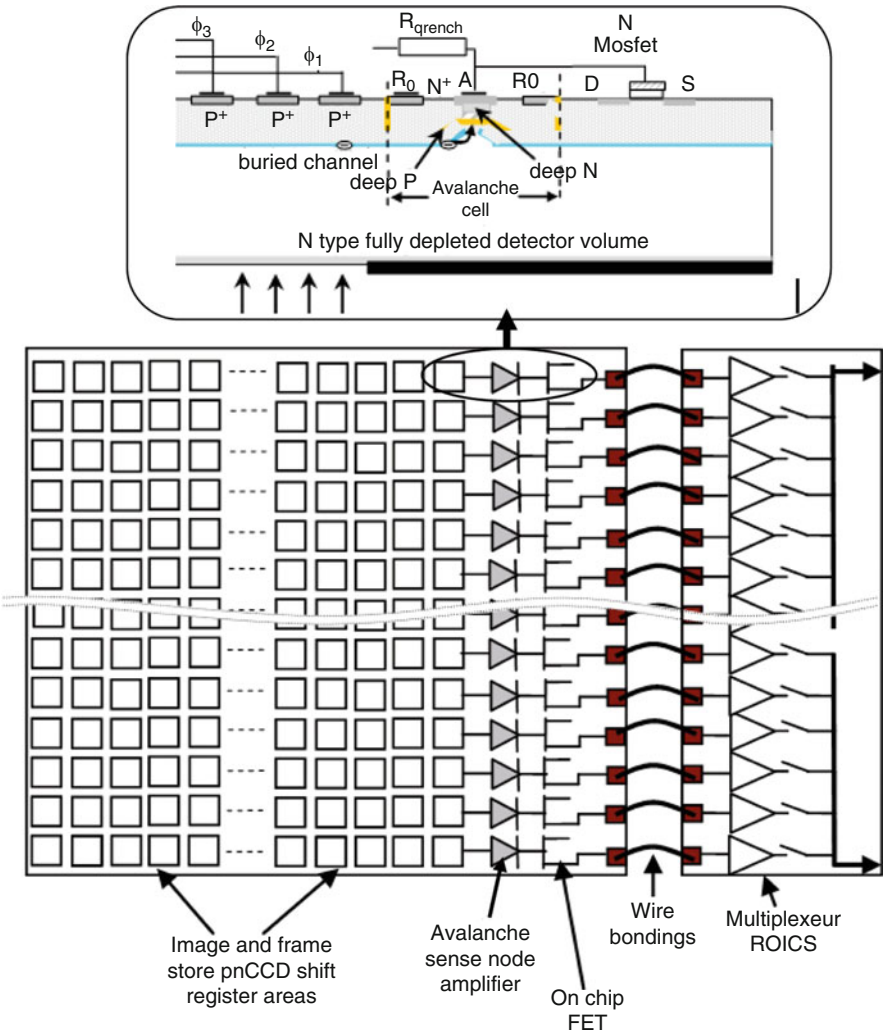

Fig. 13.14 Architecture of avalanche amplifying pnCCD with a detailed view of avalanche sense node area

established between the anode A and a deep P doping controllable by using the bias ring R0 (Fig. 13.14).

The first experimental results, reported in [47], appear promising with a dark count for the avalanche cell measured as 2 dark events/frame at −60°C at 100 Hz readout rate and a good time correlation between transfer and avalanche time.

Finally, a last investigation way for WFS sensing is in the use of high frame rate, low noise (<2 e-rms), high responsivity (around 100 μV/e-), back-illuminated CMOS sensor. Developments promoted by ESO of large size 1,680 × 1,680 pixels, 24 μm pitch, 1 kfps sensor are currently on the way.

13.5.3 Infrared Detectors for Wavefront Sensor

As seen in Sect. 13.3, the isoplanetic angle increases with the wavelength. This provides a motivation to use IR sensors for AO applications. However, up to now, the use of IR detectors has been mostly restricted to low-order tip-tilt sensors, which require small format size and limited readout speeds of 100 fps, essentially because of high read noise (10–20 e-) at fast frame rates. A noticeable exception is the NAOS/CONICA Instrument from ESO VLT, equipped with infrared Shack–Hartmann WFS using a HgCdTe, 1,024 × 1,024 pixels detector (18.5 μm pixel size) working in the 0.8–2.5 μm spectral band [48]. This device has been primarily designed for science applications but, using here some particular readout schemes, frame rate as high as 1,200 Hz could be obtained while keeping a readout noise performance of 10 e- rms.

To reduce the noise [19], one circuit approach consists in using a capacitive transimpedance amplifier (CTIA) for each pixel rather than a source-follower structure, which provides both high gain and bandwidth. In the frame of ESO R&D effort focused on the future Extremely Large Telescope (E-ELT), a 128×128 pixels, 40 μm pitch, detector dedicated to WFS application using a single stage CTIA has been designed featuring high conversion gain (320 μV/e with Cint = 0.5 fF) and readout noise of less than 5 e- rms at a frame rate of 900 Hz (thanks to CDS circuit).

Nevertheless, due to basic limitations of CMOS ROICs, the improvements of the read noise are very difficult to achieve, and a breakthrough is required to reach a lower readout noise level required by WFS (also by fringe trackers) and other challenging goal specifications described in Table 13.2 (extracted of Table 3 in [39]).

The most promising way to overcome the infrared detector ROIC noise barrier is to amplify photoelectron signal inside the infrared detecting material by means of the avalanche gain to increase the signal above the noise floor of ROIC. This motivates active work in HgCdTe APDs array development.

In the last decade and based on initial demonstrations [49, 50], HgCdTe APDs have been actively investigated (see [51, 52] for an extensive bibliography). Avalanche properties of $Hg_{(1-x)}Cd_xTe$ crystal depend on the stoichiometry x and consequently, two regimes are possible depending on the bandgap energy E_g and the related cutoff wavelength λ_c [53]:

- For $E_g < 0.65$ eV ($\lambda_C > 1.9$ μm), the hole ionization coefficient is very low (close to zero). Only electrons can ionize by impact and the gain of the so-

Table 13.2 Challenging requirements of future E-ELT IR AO WFS detector

Parameter	Min spec.	Goal spec.
Full frame rate (fps)	750	1,500
Readout noise at 750 fps	5 e- rms	3 e- rms
Quantum efficiency at Z, J, H, and K ands	>70% s	>80%
Dark current (e-/s/pixel at $T = 80$ K)	<3	<1
Storage capacity (e-)	10k	20k

called e-APD ("electron-initiated APD") increases exponentially with decreasing bandgap energy.
- For $E_g > 0.65\,\text{eV}$ ($\lambda_C < 1.9\,\mu\text{m}$), in the so-called h-APD domain, the holes contribute to the avalanche effect and the APD gain rises to infinity in avalanche breakdown.

This property is related to the direct band gap structure of HgCdTe. Electrons can be accelerated more rapidly than holes because of the large effective mass ratio (25–30) and limited phonon interaction (holes interact more with phonons, due to the larger effective mass and they do not earn enough energy to ionize by impact). This process can be described by a purely ballistic ionization model [54]. In summary, avalanching for $x < 0.53$ results exclusively from electron ionization. As an important consequence, e-APDs do offer an excess noise factor close to unity. This is in contrast to their counterparts realized with III–V material or Si, where photoionization initiates avalanching by both electrons and holes, leading typical excess noise factors often larger than 2 [55], thus excluding noiseless photon counting in linear mode. Exponentially increasing multiplication gains of 100–1,000 for low values of reverse bias (5–10 V) can be obtained [56] as shown in Fig. 13.15.

In addition, consistent with theoretical predictions of constant bandwith for unipolar multiplication, a nearly constant bandwidth has been observed and a value of bandwidth-product (GBW) above 2 THz has been demonstrated in HgCdTe e-APDs [57]. In contrast, the bandwidth in III–V material or Si APDs is known to be inversely proportional to the gain and the GWB product maximum value of 320 GHz has been achieved in III–V APDs [55].

The e-APDs are manufactured using LPE or MBE epitaxy. They use an N+/N/P structure backside illuminated [58], either in a planar structure bump-bonded on ROIC, or in a cylinder geometry (cylindrical "P-around N" photodiode structure [59]) with metal vias for connections to the ROIC.

For wavefront sensing with integration time in the millisecond range, the dark current and its associated shot noise has to be limited. As the dark current scales

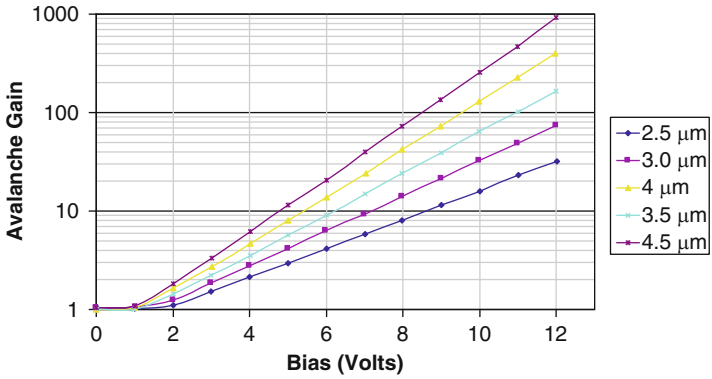

Fig. 13.15 Avalanche gain vs. bias voltage at 77 K for cutoff wavelength (i.e., different setting of material bandgap energy tuning) (from [56])

exponentially with the bandgap energy, this leads to use short cutoff wavelengths λ_C from 2.5 to 3 μm (i.e., high bandgap energy value), with the drawback of a reduced gain.

The dark current has been recently reported as 12 pA/cm^2 (84.5 e/s/pixel) at 80 K [60] for HgCdTe e-APDs with $\lambda_C = 2.64$ μm, operating at a bias voltage of 7.8 V with a gain of 7.6 and hybridized on an existing 320×256 pixels, 24 μm pitch, four outputs CMOS ROIC. This value remains nearly at this level at lower temperature due to the dominant trap-assisted tunneling mechanism. The readout noise for this array was measured as 7.2 e- rms for CDS readout, and 2.8 e- rms for Fowler-8 sampling (eight Fowler pairs). This demonstrates clearly that the HgCdTe e-APD should be matched with a low-noise ROIC offering the suited architecture to fulfill the severe combined low noise and high rate requirements of WFS application of less than 3 e- rms noise at 1 kHz frame rate.

One of the keys of success of HgCdTe APDs detectors will reside in the development of dedicated ROIC and efforts in this direction are pursued actively.

13.6 Space LIDAR Applications

Light detection and ranging (standing for LIDAR) is an optical remote sensing method, often seen as laser radar, which operate by sending laser light to the atmosphere, analyzing light components scattered by the atmosphere as function of time and/or convert the time of flight to a distance for range measurement. It has proved to be extremely useful in atmosphere and environment characterization, analysis of meteorological phenomena, and space exploration. LIDAR is used as ground-based, airborne-based, or space-based systems. An extensive coverage of LIDAR applications can be found in [61]. It has been used for many decades for satellite and lunar ranging using fixed ground-based facilities, such as NASA SLR2000 Satellite Ranging System transmitting 60 μJ of 532 nm energy at 2 kHz frame rate [62], and more recently as altimeter.

The efficiency of a LIDAR is determined by how many useful measurements can be made for a given value of the product of transmitted laser power by the area of the receive telescope aperture (so-called "Power-aperture Product" PAP). This value is typically for satellite-based LIDAR in the range of a few tens of 10^{-11}W to some 10^{-11}W depending on the altitude (several 100 km), pulse energy (several 10 mJ to several 100 mJ), and telescope mirror diameter (typically around 1 m).

It has been demonstrated that photon counting LIDARs, which require the detection of a single photoelectron per measurement, are the most efficient [63]. The GLAS altimeter on NASA ICESat1, using discrete Single Photon Counting Modules, featuring 13×10^6 maximum count rate and less than 500 dark counts/s with a detection efficiency of 65% at 532 nm, provides both atmospheric profiles (at 532 nm) and ice sheets elevation at 1,064 nm using a 40 Hz repetition rate 105 mJ laser (with a minimum detectable signal of approximately 100 electrons). Its successor ATLAS on ICEsat2 (to be launched in 2014), operating in photon counting

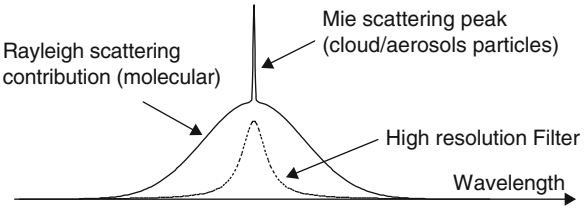

Fig. 13.16 Separation of Rayleigh (molecular) and Mie (cloud and aerosol particles) in scattering contributions (from [67])

mode, will use 2 mJ laser at a repetition rate of 10 KHz [64]. ESA is currently developing a photon counting altimeter for planetary landing assistance based on Silicon SPAD array hybridized on a photon counting readout circuit able to perform ranging with 2 cm accuracy (equivalent to a 150 ps time) over a 5-km distance [65].

Apart from altimetry, the LIDAR principle is more and more used in space for atmosphere characterization (Atmospheric LIDAR). The emitted radiation, back scattered by molecules, aerosol and particles (dust) present in the atmosphere, is sampled and analyzed. As an example, the ATLID instrument from the ESA EARTHCARE mission is an atmospheric LIDAR operating at a single wavelength of 355 nm (3rd harmonic of an Nd-YAG laser) and providing vertical profiles for aerosols and altitude of the highest cloud top [66].

The interaction of light with molecules or aerosols leads to different spectra. ATLID makes use of a high-spectral resolution (HRS) receiver separating, as illustrated in Fig. 13.16, the Rayleigh (molecular) and Mie (cloud and aerosol particles) backscatter return contributions using high-resolution narrow bandwidth Fabry–Pérot etalons.

After the pulsed laser beam has been emitted from the satellite, a fraction of the light backscattered from aerosols and cloud is detected at the focal plane with a time delay (around 3 ms) in both Rayleigh and Mie channels. Along the so-called "atmospheric sequence," the signal is backscattered from the multiple atmospheric layers as the beam is propagating. At the end, a "ground echo" is observed corresponding to the signal reflected from the Earth's surface.

The targeted emitted laser pulse energy (355 nm) is in the range of 19–25 mJ with a pulse repetition frequency in the 100 Hz range. The receiver uses a telescope with a diameter of 0.6 m and the use of narrow bandwidth filter leads to single-photon sensitivity requirements. Although a photomultiplier tube (PMT) could also have been considered, a dedicated EMCCD architecture described in Fig. 13.17 has been developed and evaluated for this class of applications [68], featuring a 6×6 pixels core coupled to a binning register.

The required overall detector area is $120 \times 120\ \mu m^2$, but it is organized as a small array of 6×6 square pixels ($20\ \mu m$ pitch) to allow for a fast charge transfer required by the 667 ns time frame allowed by 100 m targeted spatial resolution, especially taking into account the low signal level ($\sim$1 e- per sample) being handled. The overall charge collected in the pixel array is binned before being transferred into

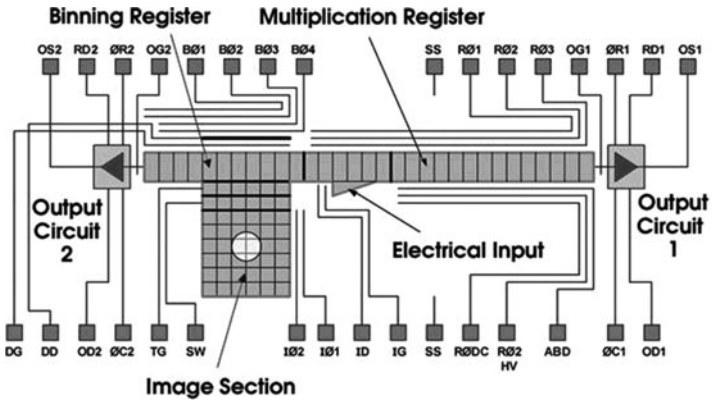

Fig. 13.17 EMCCD architecture dedicated to LIDAR application (*Credit*: e2v)

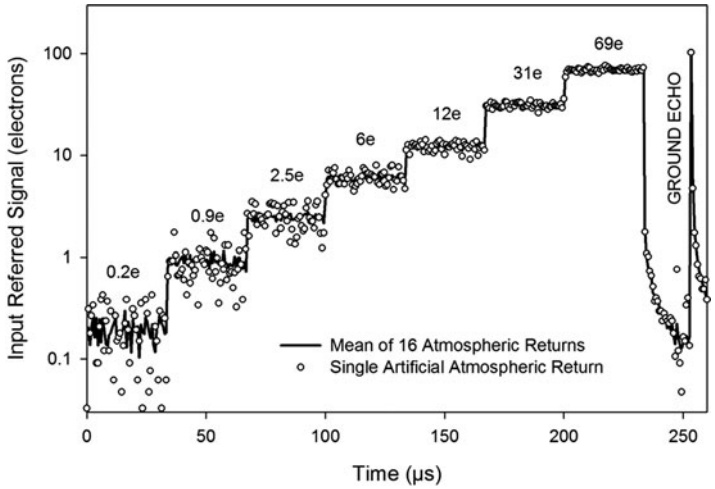

Fig. 13.18 Measured response of an EMCCD lidar sensor to a simulated atmospheric return and ground echo at a data rate of 1.5 MHz and the repetition of the sequence every 10 ms (*Credit*: e2v)

the 520 multiplication-stages register. The dark current has to be less that 0.3 e- per atmosphere sample.

Backside illumination with proper UV optimized AR coating allows for a quantum efficiency of 80% and, thanks to the multiplication process, a total noise in darkness is less than 0.5 e- rms has been achieved.

These features lead to an effective single-photon level sensitivity of this LIDAR dedicated detector as demonstrated by the test simulating the atmospheric returns with injected charge during the acquisition sequence, starting from a mean input signal of 0.2 e- and reaching up to 69 e- per sample at the end of the sequence (Fig. 13.18).

13.7 Concluding Remarks

Even if the science imaging applications from astronomy and space instruments still rely on incremental enhancements of both CCD and HgCdTe arrays, the impressive progress of single-photon imaging devices involving charge multiplication in several form make them very appealing for fast frame rate, very low noise sensing applications. Ground-based astronomy is pushing strongly the development of these technologies as a way to take the best benefit of the large scientific equipments. Space applications, severely concerned by the stability of gain required by absolute radiometric measurements and radiation tolerance, and despite significant on-ground demonstration, for example, EMCCD for the Radial Velocity GAIA instrument [69] or SPAD array for Low Cost Earth Sensor based on Oxygen Airglow [70], are still on demand on additional maturity for deploying these technologies (EMCCD, SPADs, HgCdTe e-APS) in operational missions.

Appendix

Visual Magnitude [71]

The stellar apparent brightness is described by the visual magnitude scale, which is the relative measure of the flux received at the earth. The visual (or apparent) magnitude m_v of a source is defined by

$$m_v = -2.5 \log F_v + m_0,$$

where F_v is the monochromatic flux due to the source and the value of the constant m_0 defines the magnitude zero for a reference star. As the flux evaluation is performed in a spectral band defined by a bandpass filter, F_v has to be replaced in practice by an integral quantity involving the spectral band and the spectral transmission of the filter.

It must be noted that faint objects have larger magnitude than bright ones, a difference of 5 magnitudes corresponding to a brightness ratio of 100.

Astronomical Spectral Bands Denomination [71]

"Johnson" Bands

Band name	U	B	V	R	I
Wavelength (μm)	0.36	0.44	0.55	0.7	0.88
Spectral bandwidth (μm)	0.07	0.1	0.09	0.22	0.24

Infrared Bands

Band name	J	H	K	L	M	N	G
Wavelength (μm)	1.25	1.65	2.2	3.5	4.8	10.6	21
Spectral width (μm)	0.3	0.35	0.4	1.0	0.6	5	11

References

1. I.S. McLean, *Electronic Imaging in Astronomy Detectors and Instrumentation*, 2nd edn. (Springer Praxis, New York, 2008)
2. J.W. Beletic, M. Loose, Scientific imaging sensors, short course presented at the "detectors for astronomy", in Workshop Detectors for Astronomy 2009, ESO Garching, 12–16 October 2009
3. O. Boulade et al., Development of MegaCam, the next generation wide-field imaging camera for the 3.6m Canada-France-Hawaii, Proc. SPIE **4008**, 657–668 (2000)
4. K. Kuijken et al., OmegaCAM: wide-field imaging with fine spatial resolution, Proc. SPIE **5492**, 484 (2004)
5. N. Kaiser, Pan-STARRS: a wide-field optical survey telescope array, Proc. SPIE **5489**, 11 (2004)
6. K. Gilmore, The LSST camera system overview, Proc. SPIE **6269**, 62690C (2006)
7. J.R. Janesick, *Scientific Charge-Coupled Devices* (SPIE Press, USA, 2001)
8. B. Burke, P. Jorden, P. Vu, *Overview Paper – CCD Technology, Scientific Detectors for Astronomy 2005* (Springer, New York, 2006), pp. 225
9. P. Vu et al., Large area detectors and new sensor technologies at Fairchild Imaging, Proc. SPIE **6276**, 627605 (2006)
10. B. Burke, The orthogonal-transfer array: a new CCD architecture for astronomy, Proc. SPIE 5499 (2004)
11. J. de Bruijne et al., Gaia: 1,000 million stars with 100 CCD detectors, *Proceedings SPIE* **7731**, 77311C–77311C–15 (2010)
12. D.F. Barbe, *Time Delay and Integration Image Sensors, Solid State Imaging, NATO Advanced Institute Series* (Noordhoff Int. Publishing, Leyden, 1976), pp. 659–671
13. O. Saint-Pé et al., *Focal Plane Array Design for the GAIA Space Mission, Optical Detectors for Astronomy II: State-of-the-Art at the Turn of the Millenium* (Kluwer Academic, Dordrecht, 2000), pp. 149–163
14. J.C. Pickel et al., Radiation effects on photonic imagers – a historical perspective, IEEE Trans. Nucl. Sci. **50**, 671–688 (2003)
15. A.D. Short, Gaia astrometric CCDs and focal plane, in *Proceedings of the Gaia Symposium*, (ESA SP-576) Paris-Meudon, 4–7 October 2004, p. 343
16. L. Hansen et al., Energy gap versus alloy composition and temperature in $Cd_xHg_{1-x}Te$, J. Appl. Phys. **53**, 7099–7101 (1982)
17. J. Beletic, Teledyne imaging sensors: infrared imaging technologies for astronomy and civil space, SPIE Proc. **7021** (2008)
18. G. Finger et al., Infrared detector systems at ESO, in Workshop Detectors for Astronomy 2009, ESO Garching, 12–16 October 2009
19. G. Finger, "Evolution of IR detection and possible outcomes for astrophysics, Astrophysics", *EAS Publ. Ser.*, **37**, 343–353 (2009)
20. G. Finger et al., "Recent Performance Improvements, Calibration Techniques and Mitigation Strategies for Large-format HgCdTe Arrays", in *Workshop Detectors for Astronomy 2009*, ESO Garching, 12–16 October (2009)
21. A.M Fowler, L. Gatley, Demonstration of an algorithm for read-noise reduction in infrared array, Astrophys. J. Lett. **353**, L33–L34 (1990)

22. A.M. Fowler, I. Gatley, Noise reduction strategy for hybrid IR focal-plane array, Proc. SPIE **1541**, 127–133 (1991)
23. R.N. Tubbs, *Lucky Exposures: Diffraction Limited Astronomical Imaging Through the Atmosphere*, Ph.D. thesis, Cambridge University, 2003
24. J. Vernin, C. Munoz-Tunon, The temporal behaviour of seeing, New Astron. Rev. **42**, 451–454 (1998)
25. D.L. Fried, Statistics of a geometric representation of wavefront distortion, J. Opt. Soc. Am **55**, 1427–1435 (1965)
26. J.M. Beckers, Adaptive optics for astronomy: principles, performance, and applications, Annu. Rev. Astron. Astrophys. **31**, 13–62 (1993)
27. N. Law, *Lucky Imaging: Diffraction Limited Imaging from the Ground in the Visible*, *Ph.D. thesis*, Cambridge University, 2006
28. J.C. Christou, Image quality, tip-tilt correction, and shift-and-add infrared imaging, Pub. Astron. Soc. Pac **103**, 1040–1048 (1991)
29. D.L. Fried, Probability of getting a lucky short-exposure image through turbulence, J. Opt. Soc. Am **68**, 1651–1657 (1978)
30. C.D. Mackay, J.E. Baldwin, R.N. Tubbs, Noise free detectors in the visible and infrared: implications for the design of next-generation AO systems and large telescopes, Proc. SPIE **4840**, 93–116 (2003)
31. C.D. Mackay, J.E. Baldwin, N. Law, P. Warner, High-resolution imaging in the visible from the ground without adaptive optics: new techniques and results, Proc. SPIE **5492**, 128 (2004)
32. R.N. Tubbs, J.E. Baldwin, C.D. Mackay, Diffraction-limited I band imaging with faint reference stars, Proc. SPIE **4839**, 1093 (2003)
33. O. Daigle et al., Extreme Faint flux imaging with an EMCCD, Pub. Astron. Soc. Pac **121**, 882, 866–884 (2009)
34. O. Daigle et al., CCCP: a CCD controller for counting photons, Proc. SPIE **7014**, 70146L–70146L–10 (2008)
35. C.D. Mackay et al., Astronomical imaging with L3CCDs: detector performance and high-speed controller design, Proc. SPIE **5499**, 203–209 (2004)
36. H.W. Babcock, The possibility of compensating astronomical seeing, Pub. Astron. Soc. Pac **65**, 229–236 (1953)
37. F.J. Roddier, F. Rigaut, The UH-CFHT systems, in *Adaptive Optics in Astronomy*, ed. by F.J. Roddier (Cambridge University Press, Cambridge 1999), pp. 205–234
38. R. Ragazzoni, J. Farinato, Sensitivity of a pyramidic wavefront sensor in closed loop adaptive optics, Astron. Astrophys. **350**, L23–L26 (1999)
39. M. Downing et al., Detectors for AO wavefront sensing, Proc. SPIE **7015**, 70151R–70151R–18 (2008)
40. R.J. Dorn, *A CCD Based* Curvature Wavefront Sensor*for Adaptive Optics in Astronomy*, Ph.D. thesis, University of Heidelberg, Germany, 2001
41. F. Zappa et al, Single-photon avalanche diode arrays for fast transients and adaptive optics, IEEE Trans. Instrum Meas **55** 365–374 (2006)
42. S.M. Tulloch, Application of L3 technology to wavefront sensing, Proc. SPIE **5490**, 1167–1176 (2004)
43. C. Petit et al, The SPHERE XAO system: design and performance, Proc. SPIE **7015**, 70151D–70151D–12 (2008)
44. Ph. Feautrier et al, Characterization of OCam and CCD220: the fastest and most sensitive camera to date for AO wavefront sensing, Proc. SPIE 7736 (2010)
45. R. Hartmann et al, Results of a pnCCD detector system for high-speed optical imaging, Proc. SPIE **7015**, 70155C–70155C–11 (2008)
46. P. Holl, A new high-speed, single photon imaging CCD for the optical in *IEEE Nuclear Science Symposium Conference Record*, 2006, pp. 1589–1594
47. I. Ordavo et al, Measurement results from an avalanche amplifying pnCCD for single photon imaging, Nucl Instrum Methods Phys Res A **624**, 498–503 (2010) doi:10.1016/j.nima 2010.04.155

48. E. Gendron et al, NAOS infrared wavefront sensor design and performance, Proc. SPIE **4839**, 195–205 (2003)
49. T.J. de Lyon et al, MBE growth of HgCdTe avalanche photodiode structures for low-noise 1.55 μm photodetection, J. Cryst. Growth **201/202**(3), 980–984 (1999)
50. J. Beck et al., MWIR HgCdTe avalanche photodiodes, Proc. SPIE **4454**, 188–197 (2001)
51. J. Rothman et al., HgCdTe e-APD FPAs for low flux and photon-counting applications, in *Detectors for Astronomy ESO Workshop 2009*, Garching, 12–16 October 2009
52. S. Derell et al, A Monte Carlo study of Hg0.7Cd0.3Te e-APD, IEEE Trans. Electron Devices **56**(4), 569–577 (2009)
53. D. Hall, Electron-avalanche and hole-avalanche HgCdTe photodiode arrays for astronomy, in *Detectors for Astronomy ESO Workshop 2009*, Garching, 12–16 October 2009
54. J. Beck et al., The HgCdTe electron avalanche photodiode, Proc. SPIE **5564**, 44 (2004)
55. J. Campbell et al., Recent advances in avalanche photodiodes, IEEE J. Sel. Top. Quantum Electron. **10**(4), 777–787 (2004)
56. I. Baker, G. Finger, HgCdTe avalanche photodiode arrays for wavefront sensing and interferometry applications, in *Detectors for Astronomy ESO Workshop 2009*, Garching, 12–16 October 2009
57. G. Perrais et al., Study of the transit-time limitations of the impulse response in mid-wave infrared HgCdTe avalanche photodiodes, J. Electron. Mater. **38**, 1790–1799 (2009)
58. J. Rothman et al., HgCdTe e-APD FPAs for low flux and photon-counting applications, in *Detectors for Astronomy ESO Workshop 2009*, Garching, 12–16 October 2009
59. J. Beck, The HgCdTe electron avalanche photodiode, in *2006 Digest of the IEEE LEOS Summer Topical Meetings*, 2006, pp. 36–37
60. G. Finger et al., Development of high speed low noise NIR HgCdTe avalanche photodiode arrays for adaptive optics and interferometry, Proc. SPIE **7742**, 77421K-1–77421K-14 (2010)
61. Cl. Weitkamp (ed.), Lidar, range-resolved optical remote sensing of the atmosphere, Springer Series in Optical Sciences, 102 (Ed.) **460**, 162 (2005). ISBN: 978–0–387–40075–4
62. J.J. Degnan, Present and future space applications of photon counting lidars, Proc. SPIE **7323**, 73230E-1–73230E-12–73230E (2009)
63. J.J. Degnan, Unified approach to photon-counting microlaser rangers, transponders, and altimeters, Surv. Geophys. **22**(5–6), 431–447, Kluwer Academic Publishers (2001)
64. M. Krainak et al., Photon-counting detectors on ICESat1 and candidates for ICESat2, in *Single Photon Workshop 2009*, Boulder, Colorado, USA, 2009
65. B. Dierickx, Hybrid avalanche photodiode ranging and photon-counting altimeter, Proc. SPIE **7826**, 78261Q1–78261Q6 (2010)
66. A. Heliere et al., The EarthCARE mission: mission concept and lidar instrument predevelopment, in *IEEE International Geoscience and Remote Sensing Symposium IGARSS 2007 Proceedings*, 2007, pp. 4975–4978
67. R. Meynart, Lidars in ESA's earth observation missions, in *CLRC Conference*, Snowmass, 09–12 July 2007
68. B. De Monte, R.T. Bell, Development of an emccd for lidar applications, in *International Conference on Space Optics ICSO 2010*, Rhodes, Greece, 4–8 October 2010
69. D.R. Smith, EMCCDs for space applications, Proc. SPIE **6276**, 62760K1–62760K12 (2006)
70. N. Scheidegger et al. *Low Cost Earth Sensor based on Oxygen Airglow*, Final Report ESA Contract #20267/06/NL/JA (2008)
71. P. Léna et al., *Observational Astrophysics*, 2nd ed. (Springer, New York, 1998), ISBN 3–540–63482–7

Chapter 14
Exploiting Molecular Biology by Time-Resolved Fluorescence Imaging

Francis Müller and Christof Fattinger

Abstract Many contemporary biological investigations rely on highly sensitive in vitro assays for the analysis of specific molecules in biological specimens, and the main part of these assays depends on high-sensitivity fluorescence detection techniques for the final readout. The analyzed molecules and molecular interactions in the specimen need to be detected in the presence of other highly abundant biomolecules, while the analyzed molecules themselves are only present at nano-, pico-, or even femtomolar concentration.

A short scientific rationale of fluorescence is presented. It emphasizes the use of fluorescent labels for sensitive assays in life sciences and specifies the main properties of an ideal fluorophore. With fluorescence lifetimes in the microsecond range and fluorescence quantum yield of 0.4 some water soluble complexes of Ruthenium like modified Ru(sulfobathophenanthroline) complexes fulfill these properties. They are outstanding fluorescent labels for ultrasensitive assays as illustrated in two examples, in drug discovery and in point of care testing.

We discuss the fundamentals and the state-of-the-art of the most sensitive time-gated fluorescence assays. We reflect on how the imaging devices currently employed for readout of these assays might evolve in the future.

14.1 Introduction: Time-Resolved Fluorescence as a Uniquely Sensitive Detection Method for the Analysis of Molecular Biology

The proficient design and application of miniaturized in vitro assays have become indispensable for studying and exploiting a large range of molecular interactions and the associated biochemical phenomena in drug discovery in particular and in the life sciences in general. Many contemporary biological investigations rely on highly sensitive assays for the analysis of specific molecules in biological specimens, and the main part of these assays depends on high-sensitivity fluorescence detection techniques for the final readout [1]. The targeted molecules are recognized and their

P. Seitz and A.J.P. Theuwissen (eds.), *Single-Photon Imaging*,
Springer Series in Optical Sciences 160, DOI 10.1007/978-3-642-18443-7_14,

interactions are characterized by these assays. In vitro assays involve the analysis of molecular concentrations, the analysis of biomolecular binding or cleavage processes, or the analysis of conformational changes of the investigated molecular entities. In many assays, the analyzed molecules and molecular interactions in the specimen need to be detected in the presence of other highly abundant biomolecules, while the analyzed molecules themselves are only present at nano-, pico-, or even femtomolar concentration.

This chapter describes the scientific rationale, the design, the sensitivity, and the physical implementation of highly sensitive assays based on fluorescence detection. Many of these assays rely on four major characteristics, that is

1. Labeling of specific molecules by a long-lifetime fluorophore.
2. Integration of the investigated specimens in a planar array.
3. Excitation of multiple specimens in the array by intense light pulses.
4. Imaging of the arrayed specimens on an image sensor conceived for time-gated high-sensitivity readout of the fluorescence signal.

We discuss the fundamentals and the state-of-the-art of the most sensitive time-gated fluorescence assays, we reflect on how the imaging devices currently employed for readout of these assays might evolve in the future, and we conjecture what the impact of such specialized ultra-low-noise-gated image sensors will be in the domain of life sciences.

14.1.1 Labeling of Specific Molecules by a Long-Lifetime Fluorophore

In the last decade, a multitude of miniaturized assays based on fluorescence detection for analysis of biomolecular concentrations and interactions have evolved [1]. Fluorescence assays substitute most of the assays involving radio nucleotide detection, such as radio-ligand binding assays and scintillation proximity assays. Fluorescence assays are broadly applicable; they do not require the specialized labs and safety precautions needed for assays based on radio nucleotides. In contrast to radio nucleotide detection, fluorescence detection as such is not hampered by miniaturization of the assay. Even at femtomolar concentration of the analyte, a minute 1 μl sample still contains 600 analyte molecules. The fluorophores used in the assay can be excited several thousand times before they bleach. In most of the relevant applications, the sensitivity of fluorescence detection is not limited by the size of the specimen, but by the background fluorescence originating from weakly fluorescent but highly abundant molecules in the specimen.

Many of the abundant molecules in biological specimens are intrinsically weakly fluorescent (so-called background fluorescence). In addition, the container of the investigated specimens might also contribute to background fluorescence, in particular when it is made out of glass or a polymeric material. Because the fluorescent light generated in the specimen contains background components, it

is of utmost importance to suppress these unwanted signals from the fluorescence signals indicating the assay result. The method of choice for suppression of the background fluorescence is time-resolved (time-gated) fluorescence detection using long-lifetime fluorophores, exhibiting typical lifetimes of about 100 ns or longer.

Effective suppression of background fluorescence depends on the availability of long-lifetime fluorophores with high quantum yield and efficient excitability and emission in the visible to near-infrared region of the spectrum. Nature only grants scarce access to such fluorophores! The available long-lifetime fluorophores enabling most effective suppression of background fluorescence in fluorescence assays are described in Sect. 14.2.

14.1.2 Integration of the Investigated Specimens in a Planar Array: Homogeneous and Heterogeneous Assays

Carrying out an assay often requires relatively long time spans (some minutes, hours, or even days) for incubation and conditioning of the specimens prior to readout of the assay result. Most assays require comparative measurements at defined time points involving blank, standard and control samples for calibration of the assay result and also for quality control. Arrangement of the specimens in a planar array facilitates "parallel processing and conditioning" of the specimens prior to readout of the assay result [2]. Integration of the specimens to be analyzed in a planar array provides for parallel execution of many process steps in the assay, such as allocation of the samples, sample preparation, incubation and conditioning, separation and washing steps, and – last but not least – parallel readout of the assay results by making use of a high-sensitivity image sensor for time-resolved fluorescence detection.

Time-resolved (time-gated) detection of a long-lifetime fluorophore is widely used in drug discovery for the construction of homogeneous assays that do not require separation steps [1]. In these homogeneous assays, the samples under investigation are processed in arrays of small containers, so-called microwell plates.

Suppression of background fluorescence might also be needed for achieving optimum sensitivity in heterogeneous assays, such as reversed phase protein arrays (RPAs). In the RPA assay, the samples under investigation are immobilized in high-density arrays on solid supports and then detected by fluorescently labeled antibodies. This methodology is most suited to the simultaneous analysis of multiple proteins in small volume samples [3]. The basic principle of time-gated high-sensitivity fluorescence detection in a heterogeneous assay, such as RPA, is still awaiting routine application in the life sciences, because no suitable detection system for time-resolved readout of fluorescence microarrays has been made available at this time. The detection limit of commercially available fluorescence microarray scanners these days is about 0.5 fluorophores/μm^2. This low limit of detection is inaccessible in RPA assays at the present time, because background

fluorescence in the specimen conceals the short lifetime fluorescence labels that are currently used in such assays.

14.1.3 Excitation of Multiple Specimens in the Array by Intense Light Pulses and Imaging of the Arrayed Specimens on an Image Sensor conceived for Time-Gated Readout of the Fluorescence Signal

An example of a miniaturized high-throughput assay used in drug discovery research is the quantification of cyclic adenosine monophosphate (cAMP) in homogenous solution as described in Sect. 14.4.1. This robust, homogeneous "mix-and-measure" assay with optical readout plays a central role in today's drug-discovery process. It is used to explore the biological activity of small, drug-like molecules by high-throughput screening (HTS) of diverse compound collections comprising several hundred thousand to a few million compounds [2, 4].

In high-throughput assays, there is a strong demand for parallel detection of multiple samples at a time. The HTS samples are assembled and handled in micro titer plates containing a planar array of 384 or 1,536 individual microwells. Each of these microwells is individually imaged onto a group of pixels on a high-sensitivity image sensor. Each group of pixels is then employed for the time-resolved observation of the fluorescence response of the sample in each microwell, after all microwells are simultaneously exposed to an excitation light pulse. Since the fluorescence response is a relatively weak signal, the following three measures are necessary to obtain a useful signal-to-noise ratio: (1) The excitation experiment needs to be repeated some hundred to several thousand times, and the generated photocharges must be accumulated at the pixel sites before readout, (2) the samples need to be excited by relatively intense light pulses with a typical pulse duration of the order of nanoseconds and a pulse energy of around 1 mJ for the irradiation of the entire microwell plate. (3) Since accumulation times typically exceed 1 s, very low dark noise densities need to be achieved, either by employing image sensor technologies with dark noise densities of the order of 1 pA/cm^2, or by cooling the image sensor down to appropriately low temperatures.

A typical setup for the simultaneous readout of high-throughput fluorescence assays is depicted in Fig. 14.1. Parallel readout of the analytical results is achieved through imaging of the microwell array on an ultra-low-noise detector array, such as a charge-coupled device (CCD), an intensified charge-coupled device (ICCD), or a high-sensitivity complementary metal oxide semiconductor (CMOS) image sensor.

The 96 fluorescence signals, emitted from 96 volumes of solution under investigation in one 96-well partition of the high-density microwell plate, are condensed by a high-aperture imaging system consisting of a 0.5 numerical aperture (NA) mini-lens array (MLA) and two telescope lenses. The telecentric, high-aperture optical system images 96 volumes of solution onto the detector array. A gated

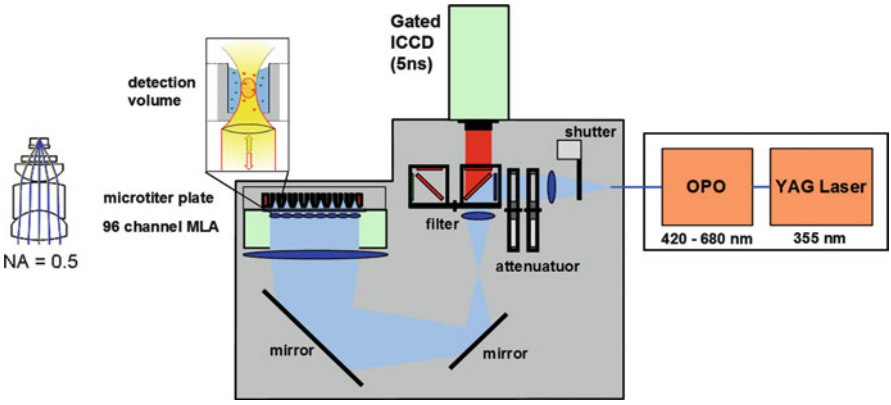

Fig. 14.1 Detection system for time-resolved FRET assay: 96-channel, time-gated detection with nanosecond resolution

ICCD image sensor is used for time-resolved analysis of the fluorescence signals generated by pulsed excitation of the long-lifetime fluorophores used in the assay. In the current setup, a pulsed YAG laser and an optical parametric oscillator (OPO) are used for generation of excitation pulses in the visible to near infrared region of the spectrum.

The time-gated readout of the FRET signal (fluorescence resonant energy transfer, see Sects. 14.3 and 14.4.1) allows for very effective suppression of the background fluorescence in the assay. The background suppression is so efficient that even serum samples exhibiting a very intense background fluorescence can be investigated.

14.1.4 Microarray Assays

The attainment of high sensitivity in microarray assays requires detection of very low surface densities of labeled molecules on the microarray by minimizing background fluorescence.

The use of fluorescence labels that can be discriminated against the background fluorescence permit the detection of 0.01 labeled analyte molecules/mm^2 in heterogeneous assays, suggesting that detection limits approximating 10^{17} mol/L (i.e., $10^3 - 10^4$ molecules/mL) are attainable in miniaturized microarray assays [5].

Microarray assays permit ultrasensitive measurement of multiple analytes in small volume samples, a comprehensive review on the development of this assay format is found in [5].

In this context it is important to note that extremely low concentrations of analyte molecules are not accessible by single molecule detection methods in free solution. Single molecule detection technologies, like fluorescence correlation spectroscopy

(FCS) [6], are hampered at very low concentration of the analyte because of inaccessible long measurement times.

14.2 Properties of the Ideal Fluorophore for Ultra-Sensitive Fluorescence Detection

All molecules of a biological assay can show various types of fluorescence (wavelength, intensity, lifetime, etc.) before, during, and after the assay. To detect an analyte of interest with high sensitivity, a suitable fluorescent label is mandatory.

Labeling as such is warranted for biological specificity, detection specificity, sensitivity, and stability. Most of the fluorescent labels are stable, can be repeatedly excited, and show long shelf life. With their molar mass between 500 and 2,000 Da, they represent, however, relatively bulky molecules which can interact unfavorably – stickiness, steric hindrance – with other assay components. They have also to be chemically coupled to one specific reagent with minimum modification of its reactive properties.

An atom or a molecule only fluoresces if it is first excited by absorption of light through an electronic transition. Relaxation to the ground state proceeds at different timescales through interaction with the neighborhood, and a photon of lower energy than the exciting photon can be emitted. In nature, however, most of the atoms and molecules do not, if at all, fluoresce strongly. This is due to nonexistent or weak absorption and to the quenching of emission through the environment, for example through water and oxygen molecules. Excitation and relaxation is illustrated (Fig. 14.2) by the Jablonski energy level diagram [7]. Absorption occurs on a

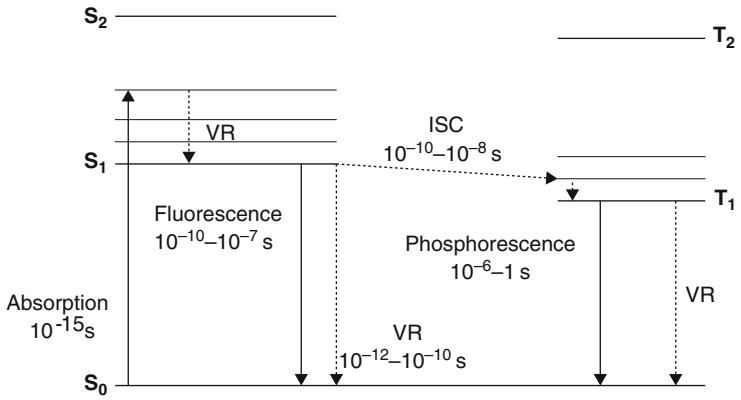

Fig. 14.2 Jablonski energy level diagram. S_0 is the ground state, S_1 andS_2 are the first and second excited singlet states, T_1 and T_2 are the first and second excited triplet states. VR represents vibrational relaxation, while ISC is the inter-system crossing spin conversion from singlet to triplet state. Emission can occur from the lowest excited state

timescale of 10^{-15} s, which is too fast for nuclei displacement. Thus, the electronic transition is vertical (Franck–Condon principle). Then, vibrational relaxation occurs on a timescale of 10^{-12} s. Emission of a photon from the lowest excited singlet state S_1 is called fluorescence and takes place in 10^{-10}–10^{-7} s after completion of vibrational relaxation so that this energy state is already in thermal equilibrium. Thus, the absorption spectrum reflects the vibrational levels of the electronically excited states, while the emission spectrum reflects the vibrational levels of the ground state. Allowed through the spin conversion during the internal system crossing, the lowest triplet state T_1 can be reached. The transition from T_1 to the ground state is called phosphorescence. Because this transition is so-called Laporte-forbidden, the phosphorescence lifetimes can be very long, in the range of microseconds to seconds.

Sensitive time-resolved fluorescence allowing for signal detection against significant background levels requires therefore labels with long lifetimes, that is phosphorescent labels. Independently of the fact that a label does fluoresce or does phosphoresce, the biologists, however, call a luminescent label "fluorophore" and describe the corresponding assays and techniques with the term "fluorescence." Accordingly, the terms fluorescence and fluorophores will be used in the following.

Nature does not offer a lot of compounds with long emission lifetimes of many tens of nanoseconds. Among organic molecules, pyrene shows an extremely long fluorescence decay of ∼100 ns, though being strongly quenched by oxygen present in all biological specimens. Among inorganic elements, the 4f lanthanides ions are well known. In particular, Eu^{3+} and Tb^{3+} exhibit very long lifetimes of several hundreds of microseconds to milliseconds, respectively, with the characteristic fingerprint emission spectrum consisting of very narrow bands, even in solution. Photoabsorption by ions is extremely feeble. A strongly absorbing organic moiety linked with the ion like an antenna is therefore required for efficient excitation.

The ideal fluorophore for ultra-sensitive fluorescence detection in biological specimens must exhibit a high absorption coefficient, a high fluorescence quantum yield and a strong, long-lasting emission, red shifted by more than 20 nm from the excitation wavelength (Stokes shift). It should be noted that a very long lifetime requires a corresponding increase of the fluorescence detection time, thus increasing, unfortunately, also accumulation of detector dark noise. Accordingly, an optimal fluorescence lifetime would lie in the range of 1 μs up to some hundreds of microseconds. Long wavelength excitation, above 400 nm, markedly reduces the background fluorescence because most of the molecules strongly absorb in the UV. Furthermore, for chemical coupling purposes, the fluorophore needs to be chemically stable and water soluble. Since such fluorophores do not exist in nature, they have to be designed according to the properties listed in Table 14.1.

Eu^{3+} and Tb^{3+} tailored with organic moieties building chelates fulfill the first five properties, except property 4. Excitation in the visible range is hampered by the too high energy level of the resonant energy level from which the lanthanide emission occurs. While lanthanide chelates are unstable and dissociate by strong dilution, the Eu-cryptates [8] are stable. However, the Eu-cryptates still suffer from

Table 14.1 Properties of an ideal label for ultra-sensitive fluorescence detection in biological specimens

1	High absorption coefficient	$\varepsilon > 25{,}000\,\mathrm{M}^{-1}\,\mathrm{cm}^{-1}$
2	High fluorescence quantum yield	$\Phi > 0.3$
3	Long fluorescence lifetime	$1\,\mu\mathrm{s} < \tau < 500\,\mu\mathrm{s}$
4	Excitation wavelength	$\lambda_{EX} > 400\,\mathrm{nm}$
5	Emission wavelength	Stokes shift $> 20\,\mathrm{nm}$
6	Chemical stability	
7	Water solubility	

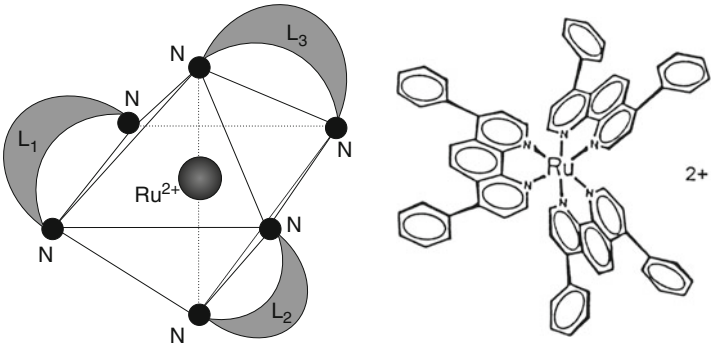

Fig. 14.3 Ruthenium complex in octahedral ligand coordination. N is the nitrogen atom, while L_1–L_3 represents the ligands. On the right side, the structure shows bathophenanthroline as a ligand. The complex can be made water soluble by derivating the phenyl groups of the ligands L_1–L_3 with sulfonyl groups SO_3^-. As an alternative, one phenyl group can be supplied with a spacer and a reactive group such as a carboxyl $COOH^-$ allowing for chemical labeling

efficient water quenching due to O–H vibrations near the Eu^{3+} resonant energy level. Appropriate design of the cryptate molecules, which literally encage the ion, drastically decreases this quenching by keeping water molecules far away from the ion [9].

All seven requirements for an ideal label for ultra-sensitive fluorescence detection are fulfilled by another class of fluorophores, the stable complexes of Ruthenium [10].

14.3 Ruthenium Complexes

Ruthenium is a d^6 transition metal forming very stable complexes with nitrogen donors such as bipyridine and phenathroline derivatives (Fig. 14.3). Water solubility is provided by sulfonyl groups.

In the complex form, the Ru atom is two times positively charged. Its electronic configuration is strongly biased by the field of the octahedral ligand coordination. Figure 14.4 shows the absorption and emission spectra of Ru^{2+}(sulfobathophenanthroline)$_3$ in aqueous solution (PBS, 10 mM phosphate, 150 mM NaCl, pH 7.4).

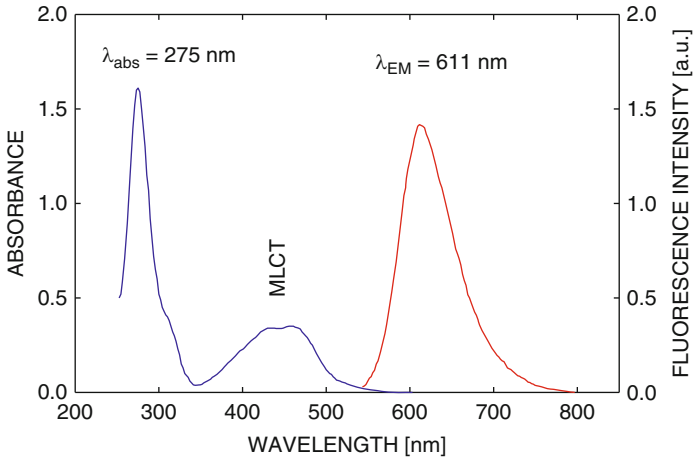

Fig. 14.4 Absorption and emission spectra of the Ruthenium complex Ru^{2+}(sulfobathophenanthroline)$_3$ in PBS (10 mM phosphate, 150 mM NaCl, pH 7.4). The absorption band in the blue range is due to the metal-to-ligand charge transfer (MLCT). The absorption band at 275 nm is a ligand centered transition (LC)

Table 14.2 Spectroscopic properties of Ru^{2+}(sulfobathophenanthroline)$_3$ in PBS

Excitation wavelength λ_{EX}	$\lambda_{EX} = 465\,\text{nm}$
Absorption coefficient at λ_{EX}	$\varepsilon_{453} = 30{,}000\,\text{M}^{-1}\,\text{cm}^{-1}$
Emission wavelength $\lambda_{EM,max}$	$\lambda_{EM,max} = 611\,\text{nm}$
Fluorescence quantum yield	$\Phi = 0.4$
Fluorescence lifetime	$\tau = 4.5\,\mu\text{s}$

Preferentially, the excitation is performed in the blue range of the optical spectrum using the metal-to-ligand charge transfer (MLCT) transition. Thereby, a 4 d metal electron is excited to a ligand excited singlet state. After efficient intersystem crossing, a triplet state is reached from which, due to spin-orbit coupling, a broad emission with lifetime in the µs range occurs [11, 12]. For Ru^{2+}(sulfobathophenanthroline)$_3$, a lifetime of 7.5 µs was obtained through the removal of dissolved oxygen by addition of Na-disulfite to the PBS [13].

All spectroscopic properties of the Ru-complexes, even the emission, are governed by the strength of the ligand field. This is in contrast to the lanthanide chelates where the ion itself determines the emission. Some spectroscopic properties of the Ru-complex with 3 sulfonated bathophenanthroline ligands, Ru^{2+}(sulfobathophen)$_3$, are given in Table 14.2.

The broad emission band facilitates the matching between emission and absorption of the donor/acceptor pair of fluorescence resonant energy transfer (FRET) systems. FRET is a fluorescence technique often applied in biological assays (see Sect. 14.4.1 below). It relies on a nonradiative energy transfer through dipole–dipole interaction between two proximate fluorophores, from an energy

donor to an energy acceptor according to Förster [14]:

$$k_{\mathrm{DA}} = \frac{9 \ln 10 \cdot \kappa^2 \cdot \Phi_{\mathrm{D}}}{128 \pi^5 n^4 N \tau_0 R^6} \int_0^\infty F_{\mathrm{D}}(\lambda) \cdot \varepsilon_{\mathrm{A}} \cdot \lambda^4 \mathrm{d}\lambda, \tag{14.1}$$

where k_{DA} is the transfer rate, κ is the orientation factor which takes values between 0, for perpendicular, and 4 for collinear transition dipoles of both molecules. Φ_{D} is the fluorescence quantum yield of the donor in absence of the acceptor, n is the refractive index of the medium, N is the number of molecules (6.02×10^{20} per millimole), τ_0 is the fluorescence lifetime in s of the donor in absence of the acceptor, R is the distance in cm between donor and acceptor, F_{D} is the donor emission spectrum, and ε_{A} is the absorption coefficient of the acceptor in $\mathrm{M}^{-1}\,\mathrm{cm}^{-1}$. The efficiency of the transfer is directly proportional to the overlap integral between donor emission and acceptor absorption and inversely proportional to the sixth power of the distance. Good transfer is observed between 10 and 50 Å.

14.4 Applications in the Life Sciences

14.4.1 Assay for Drug Discovery

This application concerns the detection of cAMP (see Sect. 14.1.3). Shortly, cAMP is a nucleotide derivative, which serves as signal molecule between cells and also regulates intracellular reactions through phosphorylation of enzymes. Phosphorylation activated enzymes induce metabolic steps, while phosphorylation inactivated enzymes stimulate synthetic pathways. In pharmacological drug screening, the cAMP content is therefore an important parameter to detect the agonist/antagonist activity of a small molecule on an enzyme.

The cAMP–N*ano* – TRF™ Detection Assay [4], developed for use in HTS in drug discovery, directly quantifies cAMP in homogeneous solution (Fig. 14.5). In this assay, cAMP labeled with the Ru-complex competes with endogenous cAMP (unlabeled) for binding to a monoclonal antibody raised against cAMP. This antibody is labeled with the acceptor dye AlexaFluor® 700. Through binding to the antibody, the Ru-complex donor and the AlexaFluor® 700 acceptor are brought together, giving rise to a FRET signal, which is inversely proportional to the cAMP concentration.

Because of the challenging synthesis of $Ru^{2+}(\text{sulfobathophen})_3$, Roche Applied Science uses $Ru^{2+}(\text{sulfobathophen})_2\text{bpy}$ in its cAMP-N*ano*-TRF™ Detection Assay [4]. This Ru-complex (see structure in Fig. 14.6a) has one sulfobathophenanthroline ligand replaced by a modified bipyridine (bpy) with carboxyl group allowing coupling to cAMP. Excitation and emission spectra of the FRET pair are

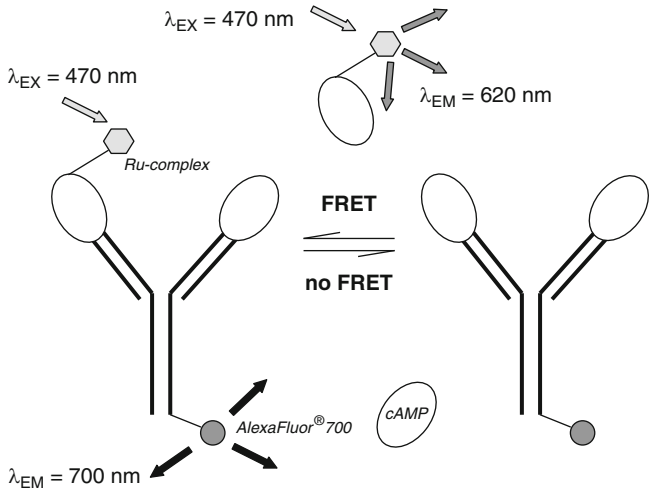

Fig. 14.5 cAMP-N*ano*-TRF detection assay. The cAMP antibody is labeled with the acceptor AlexaFluor700™. Endogeneous cAMP competes with Ru-labeled cAMP for binding to the antibody. Excitation occurs at 470 nm. FRET arises if a labeled cAMP is bound to the antibody. The emission is then detected at 700 nm

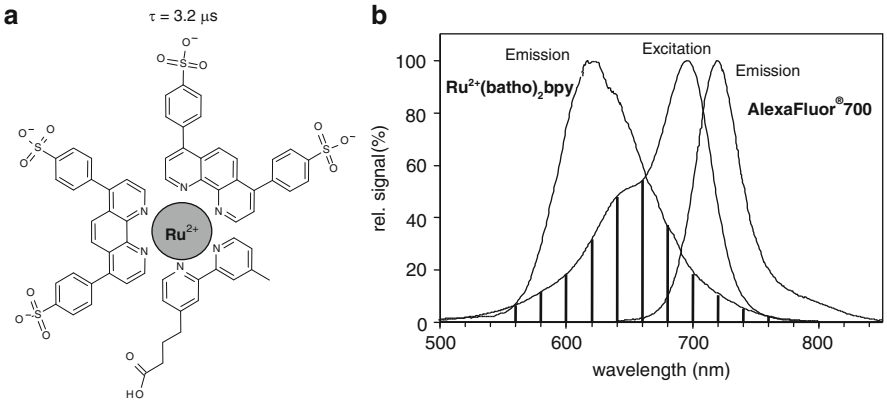

Fig. 14.6 Donor Ru^{2+}(sulfobathophen)$_2$bpy and acceptor AlexaFluor® 700 as FRET pair. (**a**) Structure of the donor: For chemical coupling, a short linker with the carbonyl group $COOH^-$ is attached to the bipyridine moiety. (**b**) Excitation and emission spectra of donor and acceptor. The *dashed area* corresponds to the overlap between donor emission and acceptor excitation

depicted in Fig. 14.6b. The overlap between donor emission and acceptor excitation is indicated by the dashed area.

With a delay of 100 ns after the ending of the excitation pulse, the time-gated detection system opens the window for photon accumulation for the duration of 400 ns (see timing diagram of Fig. 14.7a). The cAMP standard curve shown in

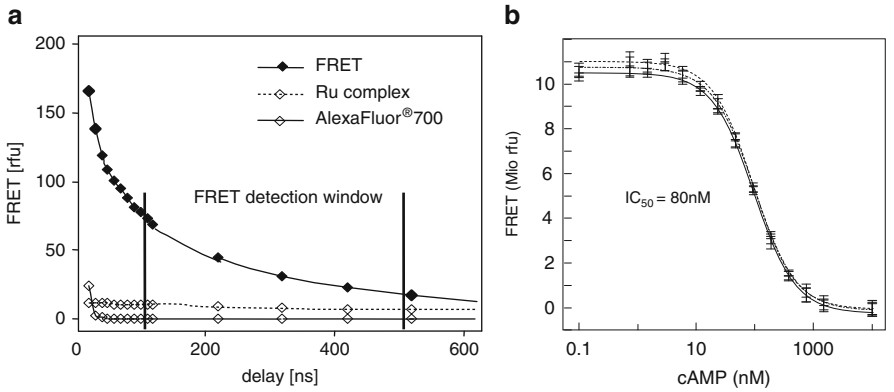

Fig. 14.7 Nano-TR-FRET assay. (**a**) Timing detection diagram: filled diamond: FRET; dotted open diamond: Ru-complex; solid open diamond: AlexaFluor® 700. Detection delay and detection window corresponding to 100 and 400 ns, respectively. (**b**) cAMP standard curve at different incubation times: 30 (solid), 60 (dashed), and 150 min (dotted)

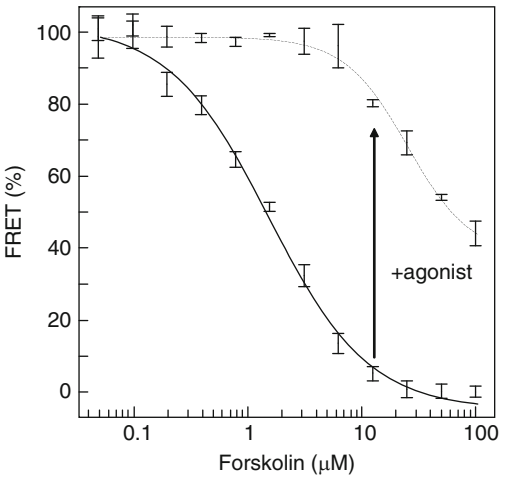

Fig. 14.8 Forskolin dose response in absence and in presence of agonist: *continuous line* without agonist; *dashed line* with agonist

Fig. 14.7b demonstrates an IC_{50} value of 80 nM. This is the cAMP concentration at which the FRET signal is reduced to 50%.

Forskolin is a small molecule commonly used to raise the production of cAMP in cell physiology studies. Figure 14.8 illustrates the forskolin dose response of the cellular assay for a G_i-coupled receptor in absence and in presence of agonist [15]. G_i-coupled receptors activation – due to the presence of agonist – results characteristically in a decrease of cAMP production. The assay is performed in a 384-well plate format with about 10,000 cells/well. After stimulation with forskolin and incubation for 30 min, the detection mix – containing the reagents labeled cAMP, labeled cAMP antibody and detergents for cell lysis – is added. The presence

of the agonist clearly shifts the response curve to higher forskolin concentration referred to the response obtained without agonist.

14.4.2 Assay for Point of Care Testing

Point of care testing requires convenient and robust instrumentation as well as easy-to-use assay procedures. In the following, the principle of a two-step fluorescent CEA assay investigated at Hoffmann–La Roche in the 1980s [10] is described.

Carcinoembryonic antigen (CEA) is a glycoprotein of 200 kDa molecular weight, which appears in higher concentration in the serum of cancer patients. CEA determination is important to monitor the state of those patients as well as to follow the evolution of the applied therapies.

The assay works as follows (see Fig. 14.9). CEA is sandwiched between a first monoclonal anti-CEA antibody (Ab_1), immobilized on a polystyrene bead, and a second polyclonal antibody (Ab_2) directed against another epitope of CEA. Through a linker and a carbonyl group, Ab_2 is labeled with Ru^{2+}(sulfobathophenanthroline)$_3$. After completion of the immunological reaction – resulting from the incubation of the sample with the reagents (bead with immobilized Ab_1 and labeled Ab_2) – the sandwich is removed from the bead with the help of a sulfuric solution wherein the Ru-complex is detected by time-resolved fluorescence photon counting.

Using the MLCT band, Ru^{2+}(sulfobathophenanthroline)$_3$ (see Fig. 14.4) was excited at 453 nm with a Coumarin 460 dye laser pumped by a nitrogen laser at a repetition frequency of 12.5 Hz. The emission, filtered by dielectric edge filters, was detected by a photomultiplier in single photon counting mode. The detection timing was set at a delay of 2 μs with a counting window of 3 μs. The measuring time lasted 3 min, resulting in an effective counting time of only 13.5 ms due to the low repetition frequency of the light source. The excitation light pulse duration was <1 ns with an energy per pulse of 6 μJ.

The sensitivity defined for a signal-to-background ratio of 2 was 60 pg/ml or 3×10^{-13} M. The assay response was linear over 3 orders of magnitude. Under the same excitation and emission settings, the label alone, in a deoxygenated PBS solution, showed a sensitivity of 4×10^{-14} M.

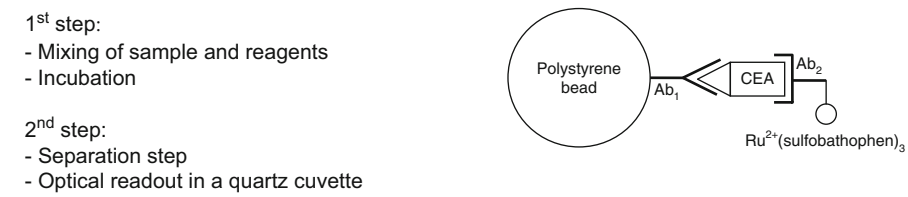

Fig. 14.9 CEA sandwich fluorescent assay. CEA is bound to the antibody Ab_1 immobilized on the polystyrene bead and to the labeled antibody Ab_2 through different epitopes

In this experiment, the excitation light source shows two types of inefficiencies (1) the low repetition rate of the nitrogen laser heightens the accumulated detector dark noise and (2) the energy per pulse of 6 µJ excites at most 3.6% of the fluorophores. An optimal light source would exhibit a pulse duration $T \leq \tau/100$ and a repetition rate of about $f_{\text{rep}} = (10 \cdot \tau)^{-1}$, τ being the fluorophore lifetime. Thus, the ideal light source for the Ru-complex considered here ($\tau = 4.5\,\mu\text{s}$) should have an energy of about 60 µJ per pulse to excite some 30% of the fluorophore molecules, a pulse duration of $T \leq 45$ ns and a repetition rate of 20 kHz. At that time this assay method was unsuitable for diagnostic application for point of care testing because of the cumbersome, bulky and expensive light source and detection system.

Nowadays, this extremely sensitive and highly relevant assay could be carried out on a portable measurement system based on high-sensitivity and ultra-low-noise semiconductor detector technology as described in this book and elsewhere [16].

14.5 Prospective Use of Ultra-Low-Noise CMOS Image Sensors for Time-Resolved Fluorescence Imaging

The healthcare system of the future will rely profoundly on the availability of diagnostic capabilities that are at the same time highly selective, sensitive, reliable, robust, and affordable. In most real-world diagnostic applications, the presence of high numbers of other molecules in essentially unknown composition obstructs the analyte of interest in the chemical reaction employed for detection.

As a consequence, this chapter has clearly shown that the real challenge in medical diagnostics is not the development of innovative photonic devices but rather the availability of a highly refined assay chemistry.

Such assay procedures are available today, and as indicated above, current efforts toward ubiquitous, cost-effective diagnostics microsystems call for affordable miniature light-sources for fluorescence excitation and ultra-low-noise image sensors for high-sensitivity-gated detection of the fluorescence response of the reporter fluorophore.

In summary, the ideal photosensor for diagnostic applications in the life sciences should have the following properties:

- *Imaging capability:* To observe a multitude of samples in a compact assay format, the samples need to be processed, imaged, examined, and evaluated simultaneously. The typical geometrical resolution required for multiwell plates currently in use starts with about 256 × 256 pixels.
- *High detection quantum efficiency:* To achieve high sensitivity of the assay, no photon should be wasted, neither in the imaging nor in the detection process. As a consequence, the quantum efficiency in the interesting wavelength range of 500–700 nm, in particular at 611 nm as indicated in Table 14.2, should be close to 100%.

- *High sensitivity:* For the same reason, each incident photon should be detected, calling for single-photon-sensitivity pixels, as discussed in several chapters of this book.
- *Low dark noise:* The repetition rate of available pulsed light sources, the delivered photon energy in each pulse and the fluorescence light yield are all so low that exposure times of several seconds are necessary in practice to obtain sufficient sensor signals. As a consequence, the dark noise properties of the semiconductor process should be optimally tuned, for example by using special CCD or PPD (pinned photo diode) devices with minimum dark noise density. In addition, it might be necessary to cool the devices, which typically halves the dark noise density for each reduction in temperature by about 8°C.
- *Fast gating:* Because of the predominance of short-lived fluorescence components in the total light output of excited assay samples, gating is of key importance in measuring only the fluorophore part in the fluorescence response of a sample. As explained in Sect. 14.4.1, adequate gating times for Ru-based fluorophores call for a delay time of several 10 ns and an integration time of several 100 ns. Since the detected fluorescence light is so weak, photogenerated charge carriers from several hundred to thousand excitation events need to be accumulated before the photosignal is finally read out. As a consequence, the pixel of our ideal fluorescence image sensor needs to have a gating capability with switching speeds of the order of 10 ns and an extremely high rejection ratio between accumulated photocharges and gated-out photocharges.

As discussed in several chapters of this book, it is clearly possible today to manufacture fluorescence image sensors with the required specifications: A combination of state-of-the-art CMOS or CCD image sensor process with dark current densities of around 10 pA/cm^2 at room temperature, gated pixels using either multi-tap CCDs, pinned photodiodes, or photogate devices, small enough geometries that photocharge diffusion plays a minor role, and antireflection coatings for quantum efficiencies of 80–95% in the wavelength region 500–700 nm are clearly possible. Such pixels can have a gating speed of less than 2 ns [17], and when combined with subelectron noise photocharge detection circuits, as described, for example, in Chap. 8 of this book, the "ideal," affordable fluorescence image sensor for a miniature, high-sensitivity assay-based diagnostic instrument for the life sciences has been realized. Ultra-low-noise CMOS image sensors devised for time-resolved fluorescence assays will become available in the future for numerous applications in the life sciences.

Acknowledgements We thank Peter Seitz for stimulating and clarifying discussions on the physics of low-noise photocharge detection, on the principles and limitations of single-molecule-sensitivity in fluorescence assays, and for proof reading of the manuscript. We are indebted to Doris Roth and Thilo Enderle for clarifying discussions on the cAMP assay.

References

1. E. Trinquet, G. Mathis, in *Exploiting Chemical Diversity for Drug Discovery*, ed. by P.A. Bartlet, M. Entzeroth (The Royal Society of Chemistry, Cambridge, 2006)
2. C. Fattinger, G. Dernick, in *Exploiting Chemical Diversity for Drug Discovery*, ed. by P.A. Bartlet, M. Entzeroth (The Royal Society of Chemistry, Cambridge, 2006)
3. G. Dernick, S. Obermüller, C. Mangold, C. Magg, H. Matile, O. Gutmann, E. von der Mark, C. Handschin, C. Maugeais, E. Niesor, Multidimensional profiling of plasma lipoproteins by size exclusion chromatography followed by reverse phase protein arrays, to be published in J. Lipid Res. (2011)
4. Roche Applied Science, cAMP-*Nano-TRF*™*Detection Assay,* Assay description booklet, Version April 2008, www.roche-applied-science.com
5. R.P. Ekins, Ligand assays: from electrophoresis to miniaturized microarrays, Clin. Chem. **44**, 2015–2030 (1998)
6. K.J. Moore, S. Turconi, S. Ashman, M. Ruediger, U. Haupts, V. Emerick, A.J. Pope, Single molecule detection technologies in miniaturized high throughput screening: fluorescence correlation spectroscopy, J. Biomol. Screen. **4**(6), 335–353 (1999)
7. A. Jablonski, Über den Mechanisnus der Photolumineszenz von Farbstoff-phosphoren, Z. Phys. **94**, 38–46 (1935)
8. G. Mathis, J.-M. Lehn, Macrocyclic rare earths complexes and application as fluorescent tracers, Patent EP 180 492, 1984
9. B. Alpha, J.-M. Lehn, G. Mathis, Energy transfer luminescence of europium(III) and terbium(III) cryptates of macrobicyclic polypyridine ligands, Angew. Chem. Int. Ed. Engl. **26**, 266–267 (1987)
10. F. Müller, D. Schmidt, Metallkomplexe, an die ein immunologisch aktives Material gekoppelt werden kann bzw. ist, deren Herstellung und Verwendung, Patent EP No 178 450, 1986
11. P. Belser, Ruthenium(II)-diimin-Komplexe: Photosensibilisatoren mit einzig-artigen Eigenschaften, Chimia **44**, 226–236 (1990)
12. J.N. Demas, B.A. DeGraff, Design and applications of highly luminescent transition metal complexes, Anal. Chem. **63**, 829A–837A (1991)
13. W. Bannwarth, D. Schmidt, R.L. Stallard, C. Hornung, R. Knorr, F. Müller, Abstracts of Papers Presented at the Sixth International Congress on Rapid Methods and Automation in Microbiology and Immunology, No 285, Espoo, Finland, July 1990
14. T. Förster, *Fluoreszenz Organischer Verbindungen*, (Vandenhoeck & Ruprecht, Göttingen, 1951)
15. D. Roth, H. Matile, J. Nielsen, K. Baumann, H.-P. Josel, K. Stein, L. Pfeiffer, T. Enderle, Tunable nanosecond lightsource for robust and sensitive assays with TRF and lifetime detection, Miptec, Basel, Switzerland, September 2010
16. A. Esposito, T. Oggier, H.C. Gerritsen, F. Lustenberger, F.S. Wouters, All-solid-state lock-in imaging for wide-field fluorescence lifetime sensing, Opt. Express **13**, 9812–9821 (2005)
17. B. Büttgen, F. Lustenberger, P. Seitz, Demodulation pixel based on static drift fields, IEEE Trans. Electron Devices **53**, 2741–2747 (2006)

Index

P. Seitz and A.J.P. Theuwissen (eds.), *Single-Photon Imaging*,
Springer Series in Optical Sciences 160, DOI 10.1007/978-3-642-18443-7,