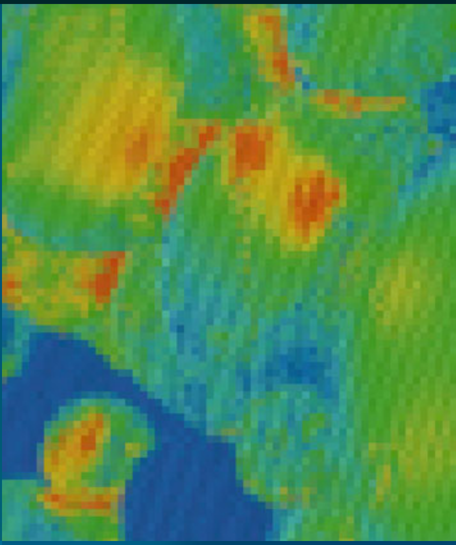
THE FRONTIERS COLLECTION



Bernd Hoefflinger (Ed.)

CHIPS 2020 VOL.2

New Vistas in Nanoelectronics

Springer

THE FRONTIERS COLLECTION

THE FRONTIERS COLLECTION

The books in this collection are devoted to challenging and open problems at the forefront of modern science, including related philosophical debates. In contrast to typical research monographs, however, they strive to present their topics in a manner accessible also to scientifically literate non-specialists wishing to gain insight into the deeper implications and fascinating questions involved. Taken as a whole, the series reflects the need for a fundamental and interdisciplinary approach to modern science. Furthermore, it is intended to encourage active scientists in all areas to ponder over important and perhaps controversial issues beyond their own speciality. Extending from quantum physics and relativity to entropy, consciousness and complex systems—the Frontiers Collection will inspire readers to push back the frontiers of their own knowledge.

More information about this series at http://www.springer.com/series/5342

For a full list of published titles, please see back of book or springer.com/series/5342

Bernd Hoefflinger
Editor

CHIPS 2020 VOL. 2

New Vistas in Nanoelectronics

Springer

Editor
Bernd Hoefflinger
Sindelfingen
Germany

ISSN 1612-3018 ISSN 2197-6619 (electronic)
THE FRONTIERS COLLECTION
ISBN 978-3-319-22092-5 ISBN 978-3-319-22093-2 (eBook)
DOI 10.1007/978-3-319-22093-2

Library of Congress Control Number: 2011943001

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media
(www.springer.com)

Preface

This volume 2 of CHIPS 2020 is a follow-up on CHIPS 2020, published in 2012, which was initiated by the 50th anniversary 2009 of the integrated circuit patent by Robert Noyce. It is indicative of the unique pace of progress of integrated circuits that, from their conception in 1959, it took only six years until 1965 that Noyce's colleague and friend Gordon Moore envisioned the unparalleled potential that these chips could double their complexity and functionality every 18 months. The driving force for this rate of innovation and market growth was that two-dimensional patterning of each new chip generation would allow to double the number of transistors per cm^2, establishing the famous nanometer road map.

As we celebrate the 50th anniversary of **Moore's law**, this nanometer road map has finally reached its limit at about 14 nm, and there are signs of postponing Gigafactory investments in 10-nm facilities. Moore's law on a two-dimensional nanometer scale is dead. But the key message of the 2012 issue of CHIPS 2020 is that Moore-scale exponential growth can continue with quantum steps in the energy efficiency and in the functional efficiency of nanochips.

I am very grateful to the authors of the 2012 chapters that they wrote highly attractive updates on their specific subjects. And we are particularly fortunate that the most distinguished experts on the key innovations for the next decade wrote highly integrated chapters on their fields: Toshiaki Masuhara, recipient of the 2000 IEEE Millennium Medal, acted as the President of the Japanese Project on Low-Power Electronics (LEAP), and he put together the final results on this most holistic research and development program for our book. Zvi Or-Bach, a lifetime creator of highly innovative technology companies and a member of the executive committee of the most promising new IEEE Cooperation S3S, Silicon-on-Insulator, 3D Integration, and Sub-Threshold MOS, wrote the most comprehensive and most realistic overview on monolithic 3D integration. Ulrich Rueckert, with a 30-year record on neural networks and a member of the European "Human Brain Project," wrote a unique assessment of the present worldwide brain-inspired computing activities. The explosion of the video data, occupying 70 % of the mobile Internet, led us to review the inflationary linear video world, caused by the era of charge-coupled devices (CCD). The superb efficiency of recording, coding, and compression of high-performance video, inspired by the human visual system (HVS), is covered by Rafal Mantiuk and Karol Myszkowski, world-renowned for their research. We consider effective HVS-inspired video as the disruptive innovation to save the mobile Internet from its breakdown and to supply reliable machine vision for future intelligent man-machine cooperation, a key perspective for nanoelectronics.

We thank our readers for their great interest in the first CHIPS 2020, which encouraged our publisher to support this second volume. Regarding its direction and strategy, I benefitted from inspiring discussions with the physics editor Claus Ascheron and with Angela Lahee, the coordinator of the Frontiers Collection. Technically, I like to thank Nele Reinders of KU Leuven, Belgium, for the friendly communication on their great work on sub-threshold DTG logic. Stefanie Krug took care again of many illustrations. I thank the Springer team for their careful editing.

As in the first CHIPS 2020 of 2012, our presentation keeps an understandable technical level so that the two books together should be helpful to the broad community concerned about nanoelectronics, from students to graduates, educators and researchers, as well as decision makers like managers, investors, and policy makers.

Sindelfingen Bernd Hoefflinger
July 2015

Contents

Editor and Contributors

About the Editor

Bernd Hoefflinger started his career as an assistant professor in the Department of Electrical Engineering at Cornell University, Ithaca, NY, USA. Returning to Germany, he served as the first MOS product manager at Siemens, Munich. With that background, he became a co-founder of the University of Dortmund, Germany, and later head of the Electrical Engineering Departments at the University of Minnesota and at Purdue University, IN, USA. In 1985, he became the director of the newly established Institute for Microelectronics Stuttgart (IMS CHIPS), a public contract research and manufacturing institute. In 1993, IMS CHIPS became the world's first ISO 9000 certified research institute. He launched rapid prototyping with electron-beam lithography in 1989. He established the institute as a leader in high-dynamic-range CMOS imagers and video cameras from 1993 onward. Among the developments in CMOS photosensors was the chip design and manufacturing for the first sub-retinal implants in humans in Europe in 2005. He retired in 2006.

Contributors

Udo-Martin Gomez Bosch Sensortec GmbH (BST/NE), Reutlingen, Germany

T. Hehn Hahn-Schickard-Gesellschaft für angewandte Forschung e.V., Villingen-Schwenningen, Germany

Bernd Hoefflinger Sindelfingen, Germany

D. Hoffmann Hahn-Schickard-Gesellschaft für angewandte Forschung e.V., Villingen-Schwenningen, Germany

W.J. Jansen Raceplan, DK Assen, The Netherlands

Matthias Keller Department of Microsystems Engineering—IMTEK, University of Freiburg, Freiburg, Germany

M. Kuhl Department of Microsystems Engineering—IMTEK, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany

J. Leicht Department of Microsystems Engineering—IMTEK, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany

Cédric Lichtenau Microprocessor Development, IBM R&D, Boeblingen, Germany

N. Lotze Department of Microsystems Engineering—IMTEK, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany

Yiannos Manoli Hahn-Schickard-Gesellschaft für angewandte Forschung e.V., Villingen-Schwenningen, Germany; Department of Microsystems Engineering—IMTEK, University of Freiburg, Freiburg, Germany

Rafał K. Mantiuk School of Computer Science, Bangor University, Bangor, UK

Jiri Marek Robert Bosch LLC, Research and Technology Center (CR/RTC-NA), Palo Alto, CA, USA

Toshiaki Masuhara LEAP—Low-Power Electronics Association and Project, Tokyo, Japan

C. Moranz Department of Microsystems Engineering—IMTEK, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany

Boris Murmann Stanford University, Stanford, CA, USA

Karol Myszkowski Department 4: Computer Graphics, Max-Planck-Institute for Informatics, Saarbruecken, Germany

Philipp Oehler Microprocessor Development, IBM R&D, Boeblingen, Germany

Zvi Or-Bach MonolithIC 3D™ Inc., San José, CA, USA

Barry Pangrle Starflow Networks, Los Gatos, CA, USA

D. Rossbach Hahn-Schickard-Gesellschaft für angewandte Forschung e.V., Villingen-Schwenningen, Germany

Peter Hans Roth Microprocessor Development, IBM R&D, Boeblingen, Germany

Ulrich Rueckert CITEC, Bielefeld University, Bielefeld, Germany

L. Spaanenburg Department of Electrical and Information Technology, Lund University, Lund, Sweden

K. Ylli Hahn-Schickard-Gesellschaft für angewandte Forschung e.V., Villingen-Schwenningen, Germany

Authors' Biography

Udo-Martin Gómez is chief technical officer of Bosch Sensortech GmbH. He has been director of engineering since 2006 for Advanced Sensor Concepts at Bosch, Automotive Electronics Division, in Reutlingen, Germany, responsible for MEMS sensor predevelopment activities. He also serves as chief expert for MEMS sensors. From 2003 until 2005, he was responsible for the development of a novel automotive inertial sensor cluster platform for active safety applications. Dr. Gómez studied physics at the University of Stuttgart, Germany. In 1997, he received his Ph.D. from the University of Stuttgart for his work on molecular electronics. In January 1998, he joined the California Institute of Technology as a post-doctoral fellow. In 1999, he started his work at Bosch Corporate Research.

Thorsten Hehn received the Dipl.-Ing. degree in microsystems engineering from the University of Freiburg, Germany, in 2006, and the Dr.-Ing. degree from the University of Freiburg in 2014.

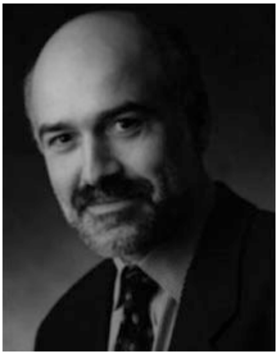
From December 2006 to September 2012, he was research assistant with the Fritz Huettinger Chair of Microelectronics, Department of Microsystems Engineering (IMTEK), University of Freiburg. From December 2006 to December 2009, he was fellow in the graduate school Micro Energy Harvesting, funded by the German Research Foundation (DFG).

In October 2012, he became a research assistant with the group "Energy Autonomous Systems" at Hahn-Schickard, Freiburg, Germany. His research interests include low-power CMOS integrated circuit design, energy harvesting, and power processing circuits for vibration-based energy harvesters.

Walter Jansen studied electrical engineering at Twente University (The Netherlands) and went for his Ph.D. research to Groningen University, where he joined the chair of Technical Computing Science. In 2000, he co-founded Dacolian and was responsible for financial control and HR until the company was merged into Q-Free ASA. Afterward, he consulted for the northern provinces in the Netherlands and subsequently founded the chair for Intelligent Vision at the NHL University of Applied Technology (The Netherlands).

Matthias Keller was born in Saarlouis, Germany, in 1976. He received his diploma degree in electrical engineering from the University of Saarland, Saarbrücken, Germany, in 2003 and a Dr.-Ing. degree (summa cum laude) from the University of Freiburg, Germany, in 2010. From 2003 to 2009, he was a research assistant at the Fritz Huettinger Chair of Microelectronics at the University of Freiburg, Germany, in the field of analog CMOS integrated circuit design with an emphasis on continuous-time delta-sigma A/D converters. In 2009, he was awarded a tenured position as a research associate ("Akademischer Rat"). His research interests are analog integrated circuits based on CMOS technology, in particular delta-sigma A/D converters and patch-clamp readout circuits.

Cédric Lichtenau is a hardware architect. He studied computer science and parallel computing at the University of Saarland, Germany, where he received his Dr. degree, before joining IBM in 2000. He has been working on processor designs for the PowerPC970, cell processor, System P, and System Z architectures. His current focus is on arithmetic and analytics units as well as hardware/software co-design.



Yiannos Manoli was born in Famagusta, Cyprus. As a Fulbright Scholar, he received a BA degree (summa cum laude) in physics and mathematics from Lawrence University in Appleton, WI, in 1978 and a MS degree in electrical engineering and computer science from the University of California, Berkeley, in 1980. He obtained a Dr.-Ing. degree in electrical engineering from the University of Duisburg, Germany, in 1987. From 1980 to 1984, he was a research assistant at the University of Dortmund, Germany. In 1985, he joined the newly founded Fraunhofer Institute of Microelectronic Circuits and Systems, Duisburg, Germany. From 1996 to 2001, he held the Chair of Microelectronics with the Department of Electrical Engineering, University of Saarbrücken, Germany. In 2001, he joined the Department of Microsystems Engineering (IMTEK) of the University of Freiburg, Germany, where he established the Chair of Microelectronics. Since 2005, he has additionally served as one of the three directors at the "Institute of Micromachining and Information Technology" (HSG-IMIT) in Villingen-Schwenningen, Germany. He spent sabbaticals with Motorola (now Freescale) in Phoenix, AZ, and with Intel, Santa Clara, CA. Prof. Manoli has received numerous best paper and teaching awards, and he has served on the committees of ISSCC, ESSCIRC, IEDM, and ICCD. He was Program Chair (2001) and General Chair (2002) of the IEEE International Conference on Computer Design (ICCD). He is on the senior editorial board of the IEEE *Journal on Emerging and Selected Topics in Circuits and Systems* and on the editorial board of the *Journal of Low Power Electronics*. He served as guest editor of *Transactions on VLSI* in 2002 and *Journal of Solid-State Circuits* in 2011.

Rafal Mantiuk is a senior lecturer at Bangor University (UK) and a member of the Research Institute of Visual Computing. Before coming to Bangor, he received his Ph.D. from the Max Planck Institute for Computer Science (2006, Germany) and was a postdoctoral researcher at the University of British Columbia (Canada). He has published numerous journal and conference papers presented at ACM SIGGRAPH, Eurographics, CVPR, and SPIE HVEI conferences. He has been awarded several patents and was recognized by the Heinz Billing Award (2006). Rafal Mantiuk investigates how the knowledge of the human visual system and perception can be incorporated within computer graphics and imaging algorithms. His recent interests focus on designing imaging algorithms that adapt to human visual performance and viewing conditions in order to deliver the best images given limited resources, such as bandwidth, computation time, or display contrast.

Jiri Marek is Senior Vice President Research, Robert Bosch, USA. He has been senior vice president of Engineering Sensors at Bosch, Automotive Electronics Division, since 2003, responsible for the MEMS activities at Bosch. He started his work at Bosch in 1986. From 1990 until 1999, he was responsible for the Sensor Technology Center. In 1999, he became vice president engineering of Restraint Systems and Sensors. Dr. Marek studied electrical engineering at the University of Stuttgart, Germany, and Stanford University, USA. In 1983, he received his Ph.D. from the University of Stuttgart for his work at the Max Planck Institute, Stuttgart, on the analysis of grain boundaries in large-grain polycrystalline solar cells. After a postdoctoral fellowship with IBM Research, San Jose, CA, he was a development engineer with Hewlett-Packard, Optical Communication Division.

Toshiaki Masuhara obtained BS, MS, and Ph.D. degrees in EE from Kyoto University in 1967, 1969, and 1977, respectively. He worked in Hitachi CRL from 1969 to 1991, on depletion-load NMOS, modeling of MOSFET including sub-threshold, and a new high-speed CMOS SRAM. From 1974 to 1975, he was a special student, EECS, University of California, Berkeley. In Hitachi, he was with the Telecom Division (1991–1993), and then became a general manager of the Technology Development Center (1993), and Semiconductor Manufacturing Technology Center (1997) in the Semiconductor and IC Div. In 2001, he was assigned executive director of ASET, and worked for the MIRAI Project. In 2010, he assumed the current position, President of LEAP.

He received the IEEE Solid-State Circuit Award on "NMOS depletion-load circuits and the development of high speed CMOS memories" in 1990, and the IEEE third Millennium Medal in 2000. He was the program chair and general chair in 1993 and 1997 of the VLSI Circuits Symposium. He was an elected Adcom, member of SSCS (1998–2000). He received a Significant Invention Awards, Japan (1994), Tokyo (1984, 1985, 1988, 1992), Yamanashi (1995), and Gumma (1996). He is a member of IEICE, Japan.

Boris Murmann joined Stanford University in 2004, where he currently serves as an associate professor of electrical engineering. He received the Ph.D. degree in electrical engineering from the University of California at Berkeley in 2003. From 1994 to 1997, he was with Neutron Microelectronics, Germany, where he developed low-power and smart-power ASICs in automotive CMOS technology. Dr. Murmann's research interests are in the area of mixed-signal integrated circuit design, with special emphasis on data converters and sensor interfaces. In 2008, he was a co-recipient of the Best Student Paper Award at the VLSI Circuits Symposium in 2008 and a recipient of the Best Invited Paper Award at the IEEE Custom Integrated Circuits Conference (CICC). He received the Agilent Early Career Professor Award in 2009 and the Friedrich Wilhelm Bessel Research Award in 2012. He has served as an associate editor of the IEEE Journal of Solid-State Circuits and as the Data Converter Subcommittee Chair of the IEEE International Solid-State Circuits Conference (ISSCC). He currently serves as the program vice-chair for the ISSCC 2016. He is a Fellow of the IEEE.

Karol Myszkowski is a tenured senior researcher at the Max Planck Institute for Computer Science, Saarbruecken, Germany. In the period from 1993 till 2000, he served as an associate professor (tenured) in the Department of Computer Software at the University of Aizu, Japan. In the period from 1986 till 1992, he worked for Integra, Inc., a Japan-based company, specialized in developing rendering and global illumination software. He received his Ph.D. (1991) and habilitation (2001) degrees in computer science from Warsaw University of Technology (Poland). In 2011, he was awarded with a lifetime professor title by the President of Poland. His research interests include global illumination and rendering, perception issues in graphics, high-dynamic-range imaging, and stereo 3D. Karol published and lectured on these topics widely. He also co-chaired the Rendering Symposium in 2001, the ACM Symposium on Applied Perception in Graphics and Visualization in 2008, the Spring Conference on Computer Graphics 2008, and Graphic on 2012.

Philipp Oehler is a microprocessor development engineering professional, working for IBM R&D Germany. He has seven years of experience in logic design for cryptographic hardware. He holds a degree in electrical engineering (Ph.D.) from the University of Paderborn, Germany, and a degree in mathematics and physics (M.Sc.) from the University Innsbruck, Austria. Philipp's areas of expertise include statistical timing analysis and logic synthesis.

Zvi Or-Bach is the founder of MonolithIC 3D™ Inc., Top Embedded Innovator-Silicon by Embedded Computing Design Magazine and Finalist of the "Best of Semicon West 2011" for its monolithic 3D-IC breakthrough. Or-Bach was also a finalist of the EE Times Innovator of the Year Award in 2011 and 2012 for his pioneering work on the monolithic 3D-IC. Or-Bach has a history of innovative development in fast-turn ASICs for over 20 years. His vision led to the invention of the first structured ASIC architecture, the first single-via programmable array, and the first laser-based system for one-day Gate Array customization. In 2005, Or-Bach won the EE Times Innovator of the Year Award and was selected by EE Times to be part of the "Disruptors"—"The People, Products and Technologies That Are Changing The Way We Live, Work and Play." Prior to MonolithIC 3D, Or-Bach founded eASIC in 1999 and served as the company's CEO for six years. eASIC was funded by leading investors Vinod Khosla and KPCB in three successive rounds. Under Or-Bach's leadership, eASIC won the prestigious EE Times' 2005 ACE Award for ultimate product of the year in the logic and programmable logic category. Earlier, Or-Bach founded Chip Express in 1989 (recently acquired by Gigoptix) and served as the company's President and CEO for almost 10 years, bringing the company to $40M revenue, and to an industry recognition for four consecutive years as a high-tech fast 50 company that served over 1000 ASIC designs, including many one-day prototypes and one-week production delivery. Zvi Or-Bach received his B.Sc. degree (1975) cum laude in electrical engineering from the Technion—Israel Institute of Technology, and M.Sc. (1979) with distinction in computer science, from the Weizmann Institute, Israel. He holds over 150 issued patents, primarily in the field of 3D integrated circuits and semi-custom chip architectures. He is the chairman of the board for Zeno Semiconductors, Bioaxial and VisuMenu.

Barry Pangrle at the time of this writing is working with a number of small startup ventures in low power design and networking. He has a BS in computer engineering and a Ph.D. in computer science, both from the University of Illinois at Urbana-Champaign. He has been a faculty member at the University of California, Santa Barbara, and the Pennsylvania State University, where he taught courses in computer architecture and VLSI design while performing research in high-level design automation. Barry has previously worked at Synopsys as an R&D director for power optimization and analysis and has also worked in power methodology and low power/energy design teams with ArchPro, Atrenta, Mentor Graphics, and NVIDIA. He has published over 25 reviewed works in high-level design automation and low power design and served as a technical program co-chair (2008) and general co-chair (2010) for the ACM/IEEE International Symposium on Low Power Electronics Design (ISLPED). Barry is also a contributing editor writing for Semiconductor Engineering.

Peter H. Roth received his Dipl.-Ing. degree in electrical engineering and his Dr.-Ing. degree from the University of Stuttgart, Germany, in 1979 and 1985, respectively. In 1985, he joined the IBM Germany Research and Development Laboratory in Boeblingen, starting in the department of VLSI logic-chip development. Since 1987, he has been leading the VLSI test and characterization team of the Boeblingen laboratory. Later, Dr. Roth led several development projects in the area of IBM's Mainframe and Power microprocessors. He was also heavily involved in the development of gaming products such as the cell processor for the Sony PlayStation. Dr. Roth is responsible for the hardware strategy for the IBM Germany Research and Development Laboratory. Since 2012, he has an additional focus on hardware/software codesign and Big-data projects like Genomic's and Cognitive Computing. At present, Dr. Roth also is driving OpenPOWER projects in the European Community.

Ulrich Rueckert received the diploma degree (M.Sc.) with honors in computer science, and a Dr.-Ing. degree (Ph.D.) with honors in electrical engineering, both from the University of Dortmund, Germany, in 1984 and 1989, respectively. He joined the Department of Electronic Components, University of Dortmund, in 1985, where he developed the first VLSI implementations of artificial neural networks in Europe. In February 1990, he accepted a position as senior researcher at the Department of Electronic Components, University of Dortmund. From 1993 to 1995, he was associate professor of microelectronics and CAD (computer-aided design) at the Research Centre for Information and Communication Technology of the Technical University of Hamburg-Harburg. From 1995 until 2009, he was full professor of electrical engineering at the University of Paderborn. As a member of the Heinz Nixdorf Institute, he held the chair in "System and Circuit Technology." Since 2009, he has been a full professor at the Cognitive Interaction Technology Centre of Excellence at Bielefeld University heading the research group Cognitronics and Sensor Systems. Since 2001 he is adjunct professor of the Faculty of Information Technology, Queensland University of Technology, Brisbane, Australia. In 2008, he received the first Innovation Award of Northrhine-Westphalia, Germany (together with his colleague Prof. Noé).

His main research interests are bioinspired architectures for nanotechnologies, neural information processing, and cognitive robotics. He has authored or coauthored more than 250 journals and conference publications. His further activities are as follows:

Chairman of the national special interest group "Microelectronics for neural networks" of the ITG (German Information Technology Society)

Founding organizer of the Int. Workshop on Microelectronics for Neural Networks: MicroNeuro (Dortmund 1990, München 1991, Edinburgh 1992, Granada 1995, Lausanne 1996, Dresden 1997).

Founding organizer of the International Workshop on Autonomous Minirobots for Research and Education (AMiRE) which was held in Paderborn, Germany (2001); in Brisbane, Australia (2003); in Fukui, Japan (2005); in Buenos Aires, Argentina (2007); and in Seoul, South Korea (2009).

Reviewer of international conferences (e.g., Int. Conference on Artificial Neural Networks, IEEE Symposium on Circuits and Systems, and Int. Workshop on Advanced Motion Control) and journals (e.g., Neural Networks, IEEE Transactions on Neural Networks, Neurocomputing, IEEE Micro, and IEEE Transactions on Circuits and Systems).

Lambert Spaanenburg received his master's degree in electrical engineering from Delft University and his doctorate in technical sciences from Twente University, both in The Netherlands. He started his academic journey at Twente University in the field of VLSI design, eventually serving as CTO of ICN, the commercial spin-off of the ESPRIT Nelsis project. At IMS in Stuttgart, he co-created the neural control for the Daimler OSCAR 1992 prototype, currently an upcoming standard for car safety measures. He became a full professor at the University of Groningen, The Netherlands. Further research of his group in neural image processing led in 2002 to Dacolian, which held a 60 % market share for license-plate recognition before it merged into Q-Free ASA. Overall, Dr. Spaanenburg has produced more than 200 conference papers, 20 reviewed journal articles, and seven books or chapters. He has served annually on several program committees and journal boards. He has been involved in many technology transfer projects, including eight university spin-offs. Currently, he is a guest at Lund University, researching and publishing on distributed embedded systems and cloud connectivity. There he has co-founded Comoray AB, where m-health products are shaped according to his schemes, starting with an autonomous blood pressure meter implemented on a standard smartphone."

Acronyms

AAEE	American Association for Engineering Education
AC	Alternating current
ACM	Association for computing machinery
ADC	Analog–digital converter
AI	Artificial intelligence
ALE	Atomic layer epitaxy
ALU	Arithmetic logic unit
AMD	Age-related macula degeneration
AMS	Analog mixed-signal
ANC	Analog network core
ANN	Artificial neural network
AP	Action potential
APS	Active-pixel sensor
APSM	Advanced porous silicon membrane
AQL	Acceptable quality level
ARPA	Advanced Research Projects Agency
ASIC	Application-specific integrated circuit
ASIP	Algorithm-specific integrated processor
ASP	Application-specific processor
ASS	Application-specific system
AVG	Available voltage gain
AVS	Adaptive voltage scaling
BAN	Body area network
BAW	Bulk acoustic wave
BEOL	Back-end of line
BI	Backside illumination
BiCMOS	Bipolar CMOS
BiCS	Bit-cost scalable flash memory
BIST	Built-in self-test
BMI	Brain–machine interface

BrainScaleS	Brain-inspired multiscale computation in neuromorphic hybrid systems
BSI	Backside illuminated
BST	Boundary scan test
BW	Bandwidth
C4	Controlled-collapse chip connect
CAD	Computer-aided design
CAGR	Cumulative aggregate growth rate
CAM	Content-addressable memory
CARE	Concerted Action for Robotics in Europe
CAS	Complementary atom switch
CAT	Computer-aided test
CCD	Charge-coupled device
CDMA	Code division multiple access
CE	Certification in Europe
CE	Consumer electronics
CE	Continuing education
CFC	Chip-integrated fuel cell
CIFB	Cascade of integrators in feedback
CIFF	Cascade of integrators in feed-forward
CIS	Charge integration sensor
CISC	Complex instruction set computer
CM	Condition monitoring
CMA	Cool mega array
CMOS	Complementary metal-oxide semiconductor
CNN	Cellular neural network
CNT	Carbon nanotube
COB	Chip-on-board
COO	Cost of ownership
CORDIC	Coordinate rotation digital computer
CPU	Central processing unit
CS	Compressed sensing
CSF	Contrast-sensitivity function
CT	Continuous time
CT-CNN	Continuous-time CNN
DA	Design automation
DAC	Design automation conference
DAC	Digital–analog converter
DARPA	Defense Advanced Research Projects Agency
DBB	Digital baseband
dBm	Power on a log scale relative to 1 mW
DC	Direct current
DCF	Digital cancellation filter
DCT	Discrete cosine transform
DDR2	Dual data rate RAM

DEM	Dynamic element matching
DFT	Design for test
DIBL	Drain-induced barrier lowering
DIBR	Depth-image-based rendering
DIGILOG	Digital logarithmic
DLP	Digital light processing
DMA	Direct memory access
DN	Digital number
DNN	Digital neural network
DOF	Degree of freedom
DPG	Digital pattern generator
DPT	Double-patterning technology liquid-immersion lithography
DRAM	Dynamic random-access memory
DRIE	Deep reactive-ion etching
DSP	Digital signal processor
DT	Discrete-time
DT-CNN	Discrete-time CNN
DTG	Differential transmission-gate
DTL	Diode–transistor logic
D-TLB	Data-cache translation lookaside buffer
DVFS	Dynamic voltage and frequency scaling
DW	Direct-write
EBL	Electron-beam lithography
ECC	Error-correcting code
ECG	Electrocardiogram
ECL	Emitter-coupled logic
EDA	Electronic design automation
EEG	Electroencephalography
EITO	European Information Technology Organization
ELO	Epitaxial lateral overgrowth
ELTRAN	Epitaxial-layer transfer
EMI	Electromagnetic interference
ENOB	Effective number of bits
EOT	Equivalent oxide thickness
ERC	Engineering Research Center
ERD	Emerging research devices
ERM	Emerging research materials
ESD	Electrostatic discharge
ESL	Electronic system level
ESP	Electronic safety package
ESP	Electronic stability program
ESPRIT	European strategic program for research in information technology
EUV	Extreme ultraviolet
EWS	Electrical wafer sort

F^2	Square of minimum feature size
FACETS	Fast analog computing with emergent transient states
FBB	Forward body bias
FC	Fuel cell
FD	Fully depleted
FDSOI	Fully depleted silicon-on-insulator
FED	Future electron devices
FEOL	Front-end of line
FeRAM	Ferroelectric random-access memory
FET	Field-effect transistor
FFT	Fast Fourier transform
FIFO	First-in-first-out
FinFET	Fin field-effect transistor
FIPOS	Full isolation by porous silicon
FIR	Finite impulse response
FIT	Failure in 10^7 h
FIT	Failure in 10^9 h
FLOP	Floating-point operation
FMEA	Failure mode and effect analysis
FOM	Figure-of-merit
FPAA	Field-programmable analog array
FPGA	Field-programmable gate array
fps	frames per second
FR	Floating-point register
FRI	Finite rate of innovation
FSM	Finite-state machine
FSR	Full signal range
GALS	Globally asynchronous, locally synchronous
GAPU	Global analogic programming unit
GBP	Gain–bandwidth product
GFC	Glucose fuel cell
GIPS	Giga instructions per second
GND	Ground
GOPS	Giga operations per second
GP DSP	General-purpose digital signal processor
GPS	Global Positioning System
GPU	Graphics processing unit
GR	General-purpose register
GSM	Global System for Mobile Communication
HAC	Hardware accelerator
HD	High definition *or* high density
HDD	High-definition disk-drive
HDL	Hardware description language
HDR	High dynamic range
HDRC	High-dynamic-range CMOS

HDRV	High-dynamic-range video
HDTV	High-definition television
HEMT	Heterogeneous emitter transistor
HICANN	High input count analog neural network
	HIPERLOGIC High-performance logic
HKMG	High dielectric constant metal gate
HKMG	High-*k* metal gate
HLS	High-level synthesis
HPC	High-performance computing
HRV	Heart-rate variability
HV	High voltage
HVM	High-volume manufacturing
HVS	Human visual system
HVT	High threshold voltage
IBL	Ion-beam lithography
IC	Integrated circuit
ICT	Information and communication technology
IEDM	International electron devices meeting
IEEE	Institute of Electrical and Electronics Engineers
IFFT	Inverse fast Fourier transform
I^2L	Integrated injection logic
IO	Input–output
IoE	Internet of Everything
IOMMU	I/O memory mapping unit
IoT	Internet of Things
IP-core	Intellectual property core
IS	Instruction set
ISO	International Organization for Standardization
ISO	International Standards Organisation
ISSCC	International Solid-State Circuits Conference
IT	Information technology
I-TLB	Instruction-cache translation lookaside buffer
ITRS	International technology road map for semiconductors
ITS	Intelligent transportation systems
JEDEC	Joint Electron Device Engineering Council
JEITA	Japanese Electronics and Information Industry Association
JESSI	Joint European submicron silicon initiative
JIT compiler	Just-in-time compiler
JND	Joust noticeable difference
JPEG	Joint Photography Expert Group
KGD	Known good die
LCA	Life cycle analysis
LDO	Low-voltage dropout regulator
LDR	Low dynamic range

LEAP	Low-power electronics association and project
LOF	Leading-ones-first
LOG	Localized epitaxial overgrowth
LPC	Linear predictive coding
LPDDR	Low-power DDR
LSB	Least significant bit
LSI	Large-scale integration
LTE	Long-term evolution
LUT	Look-up table
LVDL	Low-voltage differential logic
MASH	Multi-stage noise shaping
MBE	Molecular-beam epitaxy
MCC	Manchester carry chain
MCM	Multi-chip module
MCU	Multi-core processing units
ME	Minimum-energy
MEB	Multiple-electron beam
MEBL	Multiple-electron-beam lithography
MEMS	Micro-electro-mechanical system
METI	Ministry of Economy, Trade and Industry (Japan)
MG	Metal gate
MIPS	Mega instructions per second
MITI	Ministry of International Trade and Industry (Japan)
MLC	Multi-level per cell
MMI	Machine–machine interface
MMI	Man–machine interface
MOPS	Million operations per second
MOS	Metal-oxide semiconductor
MOSIS	MOS IC implementation system
MPEG	Motion Picture Expert Group
MPPT	Maximum power point tracking
MPU	Microprocessor unit
MRAM	Magnetic random-access memory
MTJ	Magnetic tunnel junction
MVT	Medium threshold voltage
NA	Numerical aperture
NGL	Next-generation lithography
NHTSA	National Highway Traffic Safety Administration
NIR	Near infrared
NM	Noise margin
NMOS	Negative channel-charge MOS
NoC	Network on chip
NTF	Noise transfer function
NV	Nonvolatile
OCT	Optical-coherence tomography

ODM	Original-device manufacturer
OECD	Organisation for Economic Co-operation and Development
OECF	Optoelectronic conversion function
OFDM	Orthogonal frequency-division multiplexing
OpenCL	Open Compute Language—standard established by the Khronos Group for platform independent description of highly parallel computations
ORTC	Overall road map technology characteristics
OSAT	Out-sourced assembly and test
OSCI	Open SystemC Initiative
OSR	Oversampling ratio
PA	Power amplification
PAE	Power-added efficiency
PC	Personal computer
PCB	Printed-circuit board
PCM	Phase-change memory
PCM	Pulse-code modulation
PCS	Personal Communications Service
PCS	Personal Communication Standard
PD	SOI partially depleted silicon-on-insulator
PDN	Power distribution network
PDN	Power-delivery network
PE	Processing element
PEM	Polymer electrolyte membrane
PLL	Phase-locked loop
PMOS	Positive channel-charge MOS
PMU	Power management unit
POP	Package-on-package
PPA	Power, performance, and area
PPG	Photoplethysmographical
ppm	Parts per million
PROM	Programmable read-only memory
PSCE	Pulsed synchronous charge extraction
PSE	Polymer solid electrolyte
PVDF	Polyvinylidenefluoride
PVT	Power, voltage, temperature
PWM	Pulse-width modulation
PZT	Lead zirconate titanate
QCIF	Quarter common intermediate format
QMS	Quality management system
QXGA	Quad Extended Graphics Array
R&D	Research and development
RAM	Random-access memory
RC	RC time constant
RCAT	Recessed-channel array transistor

REBL	Reflective electron-beam lithography
ReRAM	Resistive Random-access memory
RET	Resolution enhancement technique
RF	Radio-frequency
RFI	Radio-frequency interference (crosstalk)
RFID	Radio-frequency identification
RIE	Reactive-ion etching
RISC	Reduced instruction set computer
ROI	Region of interest
ROM	Read-only memory
ROW	Rest-of-world
RP	Retinitis pigmentosa
RRAM	Resistive random-access memory
RTL	Register-transfer level
Rx	Receive
S3S	Silicon-on-Insulator, 3D Integration, Sub-threshold MOS Cooperation
SAR	Successive approximation register
SBT	Strontium bismuth tantalite
SC	Switch capacitor
SD	Standard definition
SDR	Single data rate
SECE	Synchronous electric charge extraction
SEG	Selective epitaxial growth
SET	Single-electron transistor
SIA	Semiconductor industry association
SIMD	Single instruction multiple data
SIMOX	Silicon isolation by implanting oxygen
SiO_2	Silicon dioxide
SiP	System-in-package
SLD	System-level design
SMASH	Sturdy multistage noise shaping
SMASH	Sturdy multistage noise shaping
SMP	Symmetric multiprocessing
SMT	Simultaneous multithreading
SNDR	Signal-to-noise-and-distortion ratio
SNM	Static noise margin
SNR	Signal-to-noise ratio
SNT	Silicon nanotube
SO	Switch operational amplifier
SOC	System on chip
SOI	Silicon-on-insulator
SOTB	Silicon-on-thin-buried-oxide
SPAD	Single-photon avalanche diode
SPI	Serial peripheral interface
SpiNNaker	Spiking neural network architecture

SQNR	Signal-to-quantization-noise ratio
SRAM	Static random-access memory
SRC	Semiconductor Research Corporation
SSD	Solid-state drive
STF	Signal transfer function
STL	Sub-threshold leakage
STT	Spin–torque transfer
SyNAPSE	Systems of Neuromorphic Adaptive Plastic Scalable Electronics
TAM	Total available market
TFC	Thin film on CMOS
TFET	Tunneling field-effect transistor
TFT	Thin-film (field-effect) transistor
TLM	Transfer-level modeling
TOF	Time-of-flight
TOPS	Tera-operations per second
TQM	Total quality management
TRAM	1T-1R topological switching RAM
TSV	Through-silicon via
TTL	Transistor–transistor logic
Tx	Transmit
UGBW	Unity-gain bandwidth
UL	Underwriters Laboratories
ULP	Ultra-low power
ULSI	Ultra-large-scale integration
ULV	Ultra-low voltage
UM	Universal model (CNN)
UMTS	Universal Mobile Telecommunication Standard
UMTS	Universal Mobile Telecommunications System
VCO	Voltage-controlled oscillator
VDC	Vehicle dynamics control
VDE	Verband Deutscher Elektrotechniker
VDI	Verband Deutscher Ingenieure
VDS	Vehicle dynamics system
VGA	Video graphics array
VHDL	VLSI (originally VHSIC) hardware description language
VHSIC	Very-high-speed integrated circuit
VLSI	Very-large-scale integration
VRD	Virtual retinal display
VSoC	Vision system on chip
VT	Vertical transistor
wph	Wafers per hour
WSI	Wafer-scale integration
WSN	Wireless sensor network
WSTS	World Semiconductor Trade Statistics
XML	Extensible Markup Language

Abstract

CHIPS 2020, published in 2012, predicted the end of the nanometer road map for 2016 at ~ 14 nm, it proposed strategies toward low-energy femto-Joule electronics, and it projected an energy crisis. In 2015, chips have been announced with 1x nm transistors, where x = 4–14, dominated by interconnect and communication energy, which can no longer be reduced in present 2D technologies. The Internet consumed >10 % of the global supply of electric power in 2014. “CHIPS 2020, Vol. 2,” focuses on reducing chip energy and, at the same time, securing the sustained growth of nanoelectronics with increased performance as well as new products and services. A holistic strategy for low-power electronics is presented with the Japanese LEAP project. Monolithic 3D integration offers a strategy to miniaturize systems further with simultaneous gains in speed, energy, materials, and cost. Here, the S3S alliance promises noticeable progress. This is needed desperately, because the data explosion of the mobile Internet, with data rates doubling every 18 months and not supported by corresponding advances in energy efficiency, would lead to the power singularity that the Internet, consuming half of the world’s electric power, could provoke a major blackout by 2020. Since 70 % of the Internet traffic is already due to video, dealing with video from sensing through coding, compression, storing, to display, is a further focus with advice how video data can be reduced by 50–80 % with a simultaneous rise toward the human visual system, a move closer to brain-inspired systems and future human-machine interactions. Energy harvesting has matured significantly with efficient power management. This will allow fully energy-autonomous nanochip systems, if the functional and the energy efficiency of these chips have taken all the quantum steps described in this book. These autonomous nanochip systems will really open a new era for nanoelectronics everywhere and in everything.

Chapter 1
News on Eight Chip Technologies

Bernd Hoefflinger

Abstract The eight chip technologies selected for the 2012 perspective on CHIPS 2020 advanced, with the exception of single-electron IC's. Three of them merged into major cooperative R&D: The S3S initiative stands for Silicon-on-Insulator, 3D integration, and Subthreshold MOS, which we treated then and treat again extensively, particularly in differential logic.

1.1 Overview

Bipolar Technology continues, in the Silicon-Germanium symbiosis, to achieve the highest frequencies, approaching 1 Tera(10^{12})-Hz. PMOS/NMOS Technologies have no individual future but continue in the dominating CMOS technologies.

CMOS-IC Technologies receive by far the largest R&D and equipment investments with over 100 billion $, growing by more than 10 % per year and the threat of diminishing returns at physical transistor lengths below 20 nm. The reason is the fundamental variance of semiconductor transistor properties, analyzed in CHIPS 2020, and now commonly accepted. In order to control this variance at 20 nm and below, two camps are evident, which pursue different kinds of dual-gate transistors: One is the Toblerone-type of transistor, called FINFET or Tri-Gate, which is oriented towards high-transconductance at the price of a large transistor capacitance and footprint. The other is the silicon-on-insulator (SOI) type treated in the following section.

SOI-CMOS IC Technologies are based on the reference nanotransistor in CHIPS 2020. The thin, fully-depleted (FD) transistor channel is formed on a buried oxide, under which any conducting layer has a second-gate, often called buried-gate, effect for specific channel control. This SOTB (silicon-on-thin-buried-oxide) technology

B. Hoefflinger (✉)
5 Leonberger Strasse, 71063 Sindelfingen, Germany
e-mail: bhoefflinger@t-online.de

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_1

allows the highest transistor density, lowest supply voltage V and transistor capacitance C, offering the minimum internal switching energy CV^2.

SOTB is the core technology of the strategic, Sustainable-Low-Power Electronics Project described in Chap. 2.

3D CMOS IC Technologies finally receive wider R&D attention because the 2D shrinking of transistors has reached its fundamental (and practical) limits. Monolithic, crystalline, high-density stacking of transistors at the nano-scale with high crystal and transistor quality is presented in Sect. 3.5 of CHIPS 2020 with selective epitaxy and lateral crystalline overgrowth.

Topography and process complexity now receive rapidly expanding attention, as described in detail in Chap. 3.

Chip Stacks have become rapidly the largest, fast-growing development area besides the continuing development effort on the nanometer roadmap. The stacking of memory chips has advanced faster than our earlier predictions, as well as stacking heterogeneous chips, like processors and memories or MEMS and processors.

Single-Electron IC's with an electron Coulomb-confined in a transistor channel were the research hit at the turn of the millennium. Low operating temperatures and high switching voltages ($\sim$10 V) have put this approach off the list. As we have shown in Chap. 1 of CHIPS 2020, our 10 nm FD-SOI reference transistor is statistically a single-electron transistor operating at normal environment temperatures with a practical operating voltage of 200 mV.

Ultra-Low-Voltage Differential CMOS Logic, in its effective implementation as differential transmission-gate (DTG) CMOS logic, is the most promising direction to lower the operating voltage with sufficient noise margin, to minimize the transistor count and energy while maintaining operating speed. Most recently, this direction with a sophisticated history is picking up speed, since a design library for 40 nm CMOS was published in 2012. However, the force and inertia of less efficient standard-cell, static CMOS circuit libraries with energy improvements of $\sim$10 % per generation proves that any more disruptive innovations with order-of-magnitude improvements have to be accompanied by large-scale infrastructure regarding design, test and technology portability, in order to achieve a broad impact.

The emphasis in Chap. 3 of the 2012 CHIPS 2020 on ultra-low voltage, sub-threshold transistor operation has been echoed broadly in applied nanoelectronics. It is at the core of Chap. 2 in the present book, and it is one of the three foundations of the recent, truly 10× merged R&D Interest Groups, S3S, a merger of the IEEE groups "SOI (**S**ilicon-on-Insulator), 3D System Integration, and **S**ub-threshold MOS". Altogether, these three technology areas have been and continue to be key areas in our attention for the "Future of 8 Chip Technologies" (Fig. 1.1).

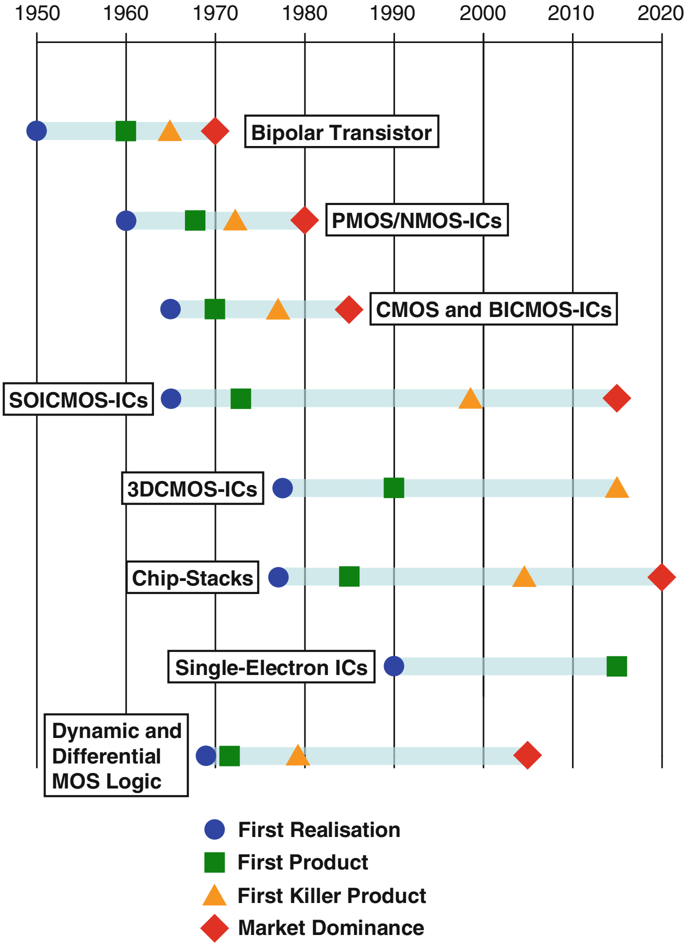
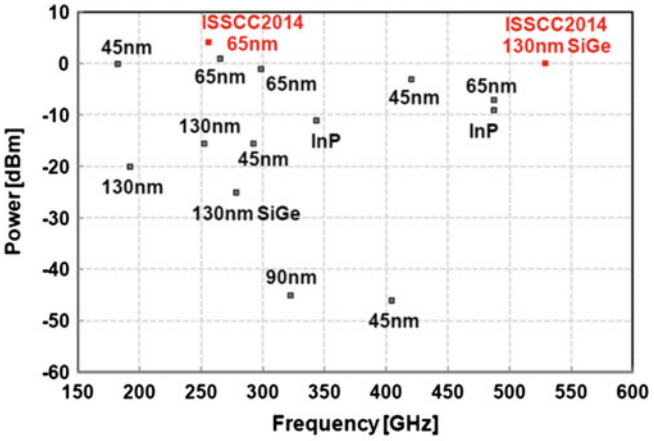


Fig. 1.1 The life-span of chip technologies, 2012 [1]. *Updates:* The first 3D-CMOS killer product arrived 2013 with vertical-gate flash memory. Single-electron-transistor first product will not arrive before 2020. Low-voltage, differential logic has arrived, but will dominate only closer towards 2020

1.2 Bipolar-Transistor Technology

Bipolar transistors were identified in [1] as the transistor type, which continues to offer the highest drive currents, a favorable transconductance and frequency limits of several hundred GHz—at additional manufacturing cost, limited area efficiency and limited compatibility for monolithic, high-density integration with mainstream nano-scale CMOS technologies.

However, their Terahertz capabilities are being pushed further as shown in Fig. 1.2, where 530 GHz have been achieved in 2014 with an output power of



Output power versus frequency for mm-Wave and sub-mm-Wave sources.
Record output power levels are being revealed at ISSCC 2014.

Fig. 1.2 Output power versus frequency for THz sources [2]

1 dBm = 10 mW [2]. The figure shows that Silicon-Germanium offers this performance, outperforming the more expensive Indium-Phosphide transistor technology. Cost-effective heterogeneous 3D integration on large-scale CMOS-chips is the much-needed development for the future of bipolar transistors in nanoelectronics.

$$A_{\mathrm{p}} = \beta_{\mathrm{F}} A_{\mathrm{V}} = \frac{q}{\varepsilon_{\mathrm{Si}}} \frac{D_{\mathrm{nB}}}{D_{\mathrm{pE}}} N_{\mathrm{E}} L_{\mathrm{E}} W_{\mathrm{C}} / V_{\mathrm{t}}.$$

1.3 CMOS Integrated Circuits

Circuits on the basis of complementary MOS transistors make up well over 90 % of all IC's. The reduction in transistor size, to achieve higher densities or functionalities, has come to the limits predicted in [1]. For the key parameter, the length L of the transistor channels:

For Logic: 16 nm,
FOR DRAM: 24 nm
For SRAM: 14 nm
For Flash: 16 nm.

The competition between FinFET and SOI, not considering the cost of manufacturing, can be reduced to the necessary switching energy (C_{Tr} + $C_{Fan\text{-}out}$ + C_{Wire})V^2.

The transistor capacitance for the FinFET in Fig. 1.3 is

$$C_{\mathrm{TrFin}} = C'_{\mathrm{G}}L(kW + H_1 + H_2) + C'_{\mathrm{jB}}LW + C_{\mathrm{jS/D}},$$

where k < 1 accounts for the profile of the fin. In the four cross-section examples, the effective transistor widths ($kW + H_1 + H_2$) are about 105, 82, 116 and 56, respectively. This shows that FinFET technologies are drive-current or charge-oriented with large transistor capacitances. The cross-sections shown relate to technologies with about 24 nm minimum features so that the minimum half-pitch of transistor-rows in the direction of current flow would be between 33 and 40 nm.

The transistor capacitance for the SOI transistor in the upper part of Fig. 1.3 is

$$C_{\mathrm{TrSOI}} = \left(C'_{\mathrm{G}} + C'_{\mathrm{BOX}}\right)LW + C_{\mathrm{FOX/S/D}}.$$

It allows a minimum transistor with W = L and minimum capacitances, albeit with lower currents, leading to different circuit strategies with maximum energy-efficiency as the top priority.

Another persistent argument is the quest for a superior semiconductor material, evaluated with its low-electric-field mobility μ_0. At the nanometer physical distance between source and drain, electric fields along the source-channel-drain space-charge path are mostly >100,000 V/cm, so that the transit-time of the electrons or holes is governed by their maximum velocity, which is the Brownian velocity $v_L = 300{,}000$ cm/s at room temperature, so that the drain current I_D in a nanotransistor with a physical channel-length <40 nm is

$$I_D \sim v_L/L,$$

irrespective of μ_0. The low-field mobility plays a role only in a scattering-limited transport for physical channel lengths >40 nm and electric fields $\sim 10{,}000$ V/cm, where the effect can be modeled by Eq. (3.16) in [1], and we have seen that planar MOS transistors reached their max. currents at 300 µA/µm width with an internal switching energy of 10^5 eV heading towards 190 µA/µm for an 8 nm channel with an internal switching energy of just 50 eV. The on/off current ratio for such a transistor with a max. voltage swing of 400 mV will be 27:1 due to its transfer characteristic of 150 mV per decade of current at room temperature. This can be improved with a dual-gate, tri-gate, fin-gate or surround-gate with larger capacitances and with gate dielectrics, whose dielectric constant is significantly higher than that of Silicon itself. Theoretically, the best achievable would then be 60 mV/decade of current at room temperature.

This limitation has led to increasing research on other types of field-effect transistors with more effective control mechanisms. A fundamentally attractive one is the tunneling effect from the valence band of a source to the conduction band of a drain, controlled by an isolated gate. This leads us to the p-i-n Tunneling Field-Effect Transistor (TFET) with a p-type source, an “intrinsic” channel and an n-type drain. Among the many variants, we show one in Fig. 1.4 with a p-type Ge

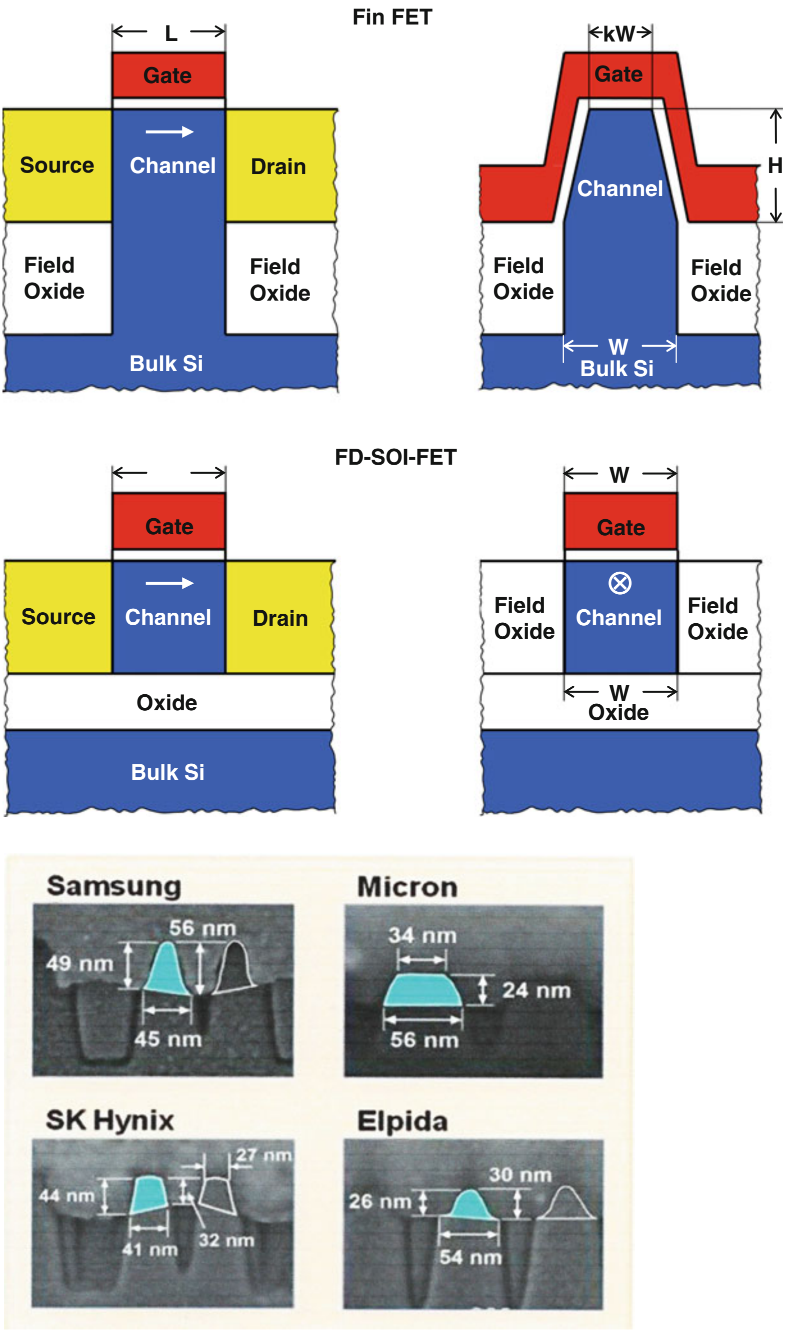


Fig. 1.3 Schematic cross-sections of a FinFET, a fully-depleted (FD) Silicon-on-insulator (SOI) FET and cross-sections of FinFET's for DRAM's

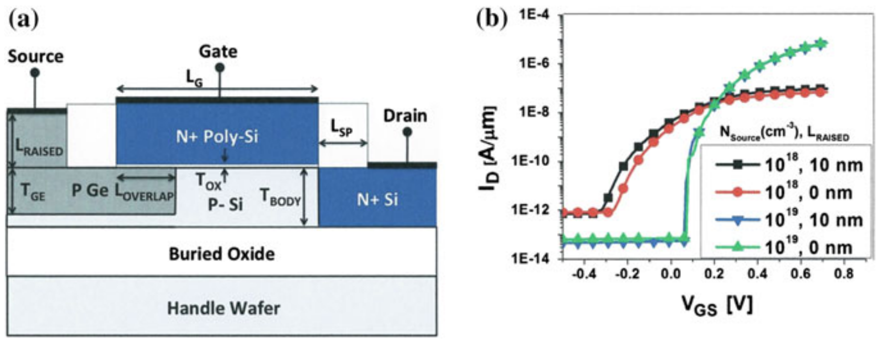


Fig. 1.4 Schematic cross-section of a Ge-Si TFET with a raised source [3]. © eecs.uc-berkeley

source, a Si channel p-doped at 10^{18} to 10^{19} per cm^3, and a highly n-type-doped Si drain [3].

The figure and all publications show that tunneling currents saturate at a maximum of <10 μA/μm channel width (at about 400 mV), about 50-times smaller than comparable Si-MOSFET's. However, their transfer characteristics have slopes between 40 and 60 mV/decade of current so that they achieve very low leakage currents, when the transistors are turned off. And this high transfer slope is independent of temperature in the range of practical operating temperatures. We summarize the properties of TFET's and normal FET's as follows:

Comparison of Normal and Tunneling FET's (Tech. node 14 nm, supply 400 mV)

	FET	TFET
Max. On-current (μA/μm)	200	10
Variance	High	Very high
Transfer slope (mV/decade)	120	50
Temp. dependence	High	Negligible
Off-current	100 nA/μm	<1 pA/μm

Tunneling FET's offer fundamentally much smaller currents, and they will only be used in applications or cases where or if operating speed is not essential like some process of health monitoring. Indeed, for such cases, TFET circuits can operate at such extremely low energy and leakage levels that energy harvesting on-chip or on-system can make these autonomous and independent of batteries or external power-supplies (see Chap. 19).

TFET technology and process integration is still at an early stage. However, it is basically compatible with mainstream nano-CMOS so that there is a high potential for their inclusion in a strategy for energy-efficient nanoelectronics.

We will see in Sect. 1.6 that the direction for energy-efficiency, as detailed in [1] with ultra-low-voltage, sub-threshold operation, has gained wider acceptance in the meantime, and it receives a comprehensive treatment in the following Chap. 2.

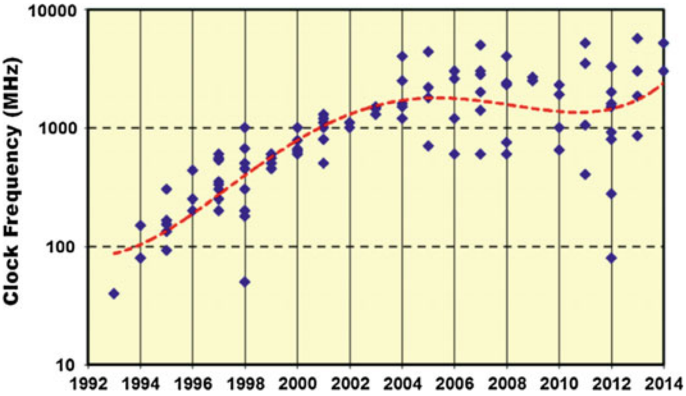


Fig. 1.5 Max. clock frequencies 2014 from [5]. © 2014 IEEE

If drive current is the dominant issue, like in clock drivers and buffers in high-fan-out gates and off-chip drivers, irrespective of energy, advances in max. operating frequencies were hoped for. In a prediction of the millennium year 2000, it had been stated that maximum clock frequencies would settle at 2–5 GHz [4].

Recent data in Fig. 1.5 show that this frequency-limit has become a fact in practice at 2–5 GHz [5].

1.4 Silicon-on-Insulator (SOI) CMOS Technology

The fully-depleted (FD) Silicon-on-insulator (SOI) transistor, (Fig. 1.3), is the reference 10 nm MOS transistor in [1]. The SOI transistor, for a given physical gate-length, is the transistor type with the simplest process flow, the minimum volume, the minimal internal capacitance and minimal source- and drain-parasitic capacitance. The group of SOI proponents and the volume of SOI-based Silicon wafers has increased significantly as the critical dimensions went below ~40 nm. In order to intensify the SOI R&D efforts and to communicate more closely with international (IEEE) working groups on monolithic and heterogeneous three-dimensional (3D) integration, the new joint-interest group

S3S**: Silico-on-Insulator – 3D Integration – Sub-threshold Operation**

has been formed [6], which can really be seen as a **10**× Program, a term introduced in "CHIPS 2020" as an R&D effort, where goals, brain-power and resources were increased by an order-of-magnitude.

The on-going optimization of SOI transistor- and circuit-efficiency and-performance can be seen in the data on a 28 nm fully-depleted SOI signal processor design with body-bias-adjustable threshold voltage (VTBB), with an efficiency of 62 pJ/operation at a frequency of 460 MHz [7]. Although the

Table 1.1 FD-SOI CMOS offers the best low-voltage sub-threshold transistor performance and minimal transistor- and circuit capacitances for optimal energy efficiency

	[7]	[8]	[9]
Technology	28 nm FD-SOI, VTBB	22 nm Tri-Gate	32 nm Bulk
Voltage range V_{DD} (V)	0.39–1.3	0.28–1.1	0.28–1.2
Max. frequency (GHz)	2.6@1.3 V	2.5@1.1 V	0.9@1.2 V
Frequency at min.V_{DD} (MHz)	460@0.4 V	17@0.28 V	3@0.28 V
Total power (mW)	370	227	400
Peak efficiency pJ/operation	62	1.6	170
Frequency/pJ (MHz/pJ) at min. energy	7.6	10.6	0.018

performance comparison of different DSP designs is just qualitative (Table 1.1), this 28 nm SOI design, compared with a 32 nm bulk and a 22 nm FinFET (Trigate) has an efficiency almost 3-times better than bulk CMOS and a frequency at minimum voltage 100-times and 30-times better, respectively. This indicates that

1.5 3D CMOS Technologies

The three-dimensional (3D) integration of transistors tackles the performance, the functional density, the interconnect capacitances and the process complexity of advanced CMOS circuits. Our focus continues to be on monolithic, high-crystalline-quality, 3D arrangements of complementary MOS transistors, possibly with crystalline self-assembly features. One consistent implementation of this strategy was summarized in Sect. 3.5 of “CHIPS 2020” on the basis of reduced-temperature, selective Si epitaxy and lateral overgrowth [10]. Its features are

- Doubling the transistor density,
- Realizing a dual-gate PMOS transistor for equal PMOS and NMOS transistor conductance,
- Reducing the mask count and the processing steps,
- Reducing local interconnect lengths and capacitances.

The key fundamental building block, to merit development efforts, was identified to be the “quad” of two NMOS and two PMOS transistors, connected as two cross-coupled inverters to achieve a differential amplifier, signal regenerator and accelerator, and the heart of a 6T SRAM memory cell. A 3D 6T SRAM cell is shown in Fig. 1.6 with a footprint of 120 F^2.

Alternative monolithic 3D integration of transistor layers with horizontal channels has progressed in recent years [11, 12], and, because of their strategic significance, they receive a detailed treatment in Chap. 3.

Vertical-channel, surround-gate-transistors are the nanotechnology-of-choice for series-connected transistors like in NAND Flash memories, and production-ready

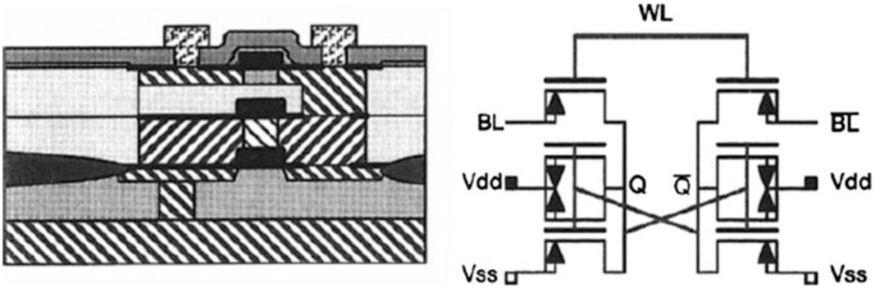


Fig. 1.6 Footprint of a monolithic 3D CMOS 6T SRAM [10]

results have been achieved (Chap. 11). One proposed realization [13] is shown in Fig. 1.7 because of its informative and exemplary process flow.

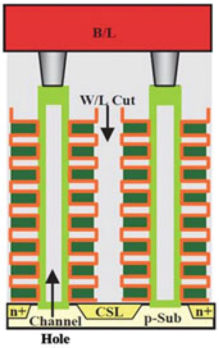
It will be shown in Chap. 11, that this type of non-volatile memory has achieved a density of 10^{11} (100 billion on the American scale) transistors/cm^2 in 2014, extending the validity of Moore's Law in its 50th year, in terms of transistors/cm^2.

1.6 Ultra-Low-Voltage Differential Transmission-Gate CMOS Logic

In the face of the energy- and power-density-crisis in nanoelectronics, the drive towards lower supply voltages and lower voltage swings has been intensified significantly, considering that the energy is basically proportional to CV^2. In the preceding Sects. 1.4 and 1.5, we have described the progress due to Silicon-on-Insulator (SOI) and monolithic 3D integration as a result of reducing the capacitance C. These are two of the three foundations of the S3S cooperation [6]. Low-voltage operation with its quadratic effect on energy takes us into the near- or sub-threshold operation of MOS transistors, where we benefit from a higher transconductance, the ratio of output current over input voltage, as detailed in [1]. CMOS circuits with PMOS pull-up and NMOS pull-down transistors have a relatively wide range of operating voltages. Extensive performance data continues to be published on low-voltage operation. Circuits and results can be classified in two categories:

1. Standard, static CMOS circuits based on widely used libraries of standard cells and
2. Special designs with optimized transmission-gate logic, in particular with fully-differential operation, as advocated in [1].

It is striking to note how persistent the industry and academia work with the standard cells (1), evidently because of the basic design- and test efficiency, another sign for the life-span of our mature CMOS age with incredible investments and



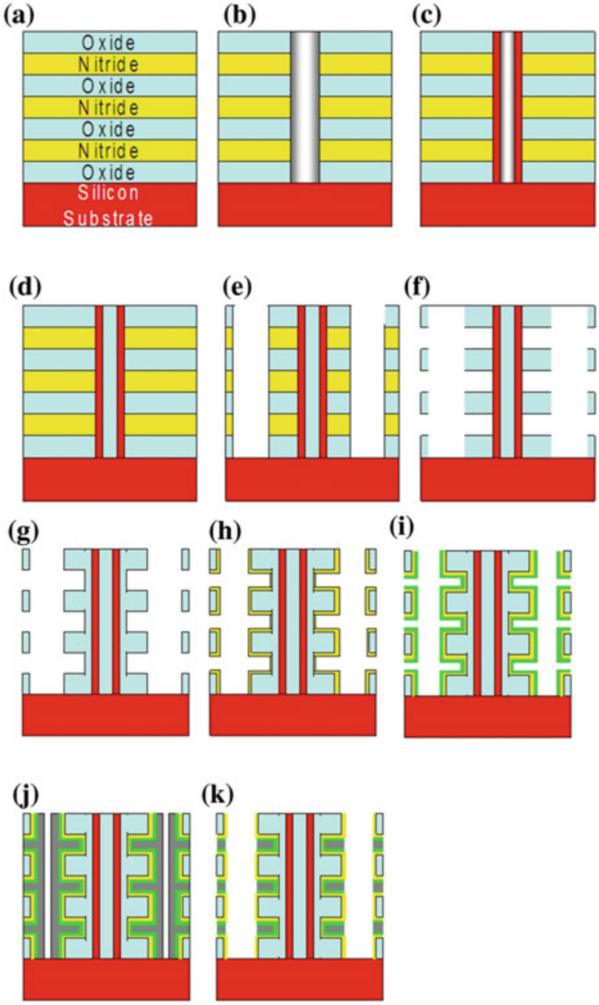


Fig. 1.7 Process-flow for a vertical-channel, 3D-integrated NAND Flash non-volatile memory [13]. **a** Deposit layers, **b** Etch hole, **c** Poly on walls, **d** Fill with oxide, **e** Etch slits, **f** Remove nitride, **g** Tunnel oxide, **h** Nitride trap, **i** Hi-k dielectic, **j** Tantalum fill, **k** Etch Ta & Hi-k. © The Memory Guy

widths of applied developments and products. Representative results are those in [9, 10], as listed in Table 1.1 and shown in Fig. 1.11. While the benefit in energy per operation is quite significant, indeed close to quadratic with the operating voltage, the loss in speed for designs based on CMOS standard-cells, if voltages are reduced, is so serious that, in order to maintain the performance, respectively the throughput, in operations per s, many such low-voltage circuits would have to be operated in parallel. The energy efficiency in operations/s/W or its inverse, the energy/operation in Joule/operation, is the basic figure-of-merit (FOM). However, the real FOM is the throughput-FOM:

- Operations/s/energy/operation.
  We have listed this in the last line of Table 1.1. Evidently, the throughput suffers seriously from the reduction in supply voltage, particularly for standard-cell, static CMOS due to series-connected transistors, which make transitions slow, which have a reduced noise-margin, and which offer little drive capability. The reduction of supply voltage in standard-cell circuits is the way to go only, if throughput does not matter like in a standby-mode.

It is the fundamental message of [1] that

- high-performance, low-voltage CMOS requires a differential operation and differential amplifiers for signal-swing regeneration, speed-up and maximum noise margins.
- With regeneration and drive at the gate outputs, transmission-gate logic is possible with minimum transistor count (all-NMOS with low threshold voltage).

A key example is the carry-generation gate in a differential transmission-gate CMOS adder (Fig. 1.8).

At the heart of this gate is the “Quad”, the cross-coupled pair of CMOS inverters, which is treated in [1] and, with all its applications, in [20]. Based on this low-voltage, sub-threshold, differential transmission-gate logic, a 16 × 16 b multiplier with 6 b accuracy was reported in 2000, called HIPERLOGIC, with 25 MOPS and an energy of only 0.4 pJ/operation in a 800 nm technology. It had a

- Throughput-FOM of 60 MOP/s/pJ in 2000, better than the results in Table 1.1, more than 12 years later. It was projected in [1] that this type of sub-threshold,

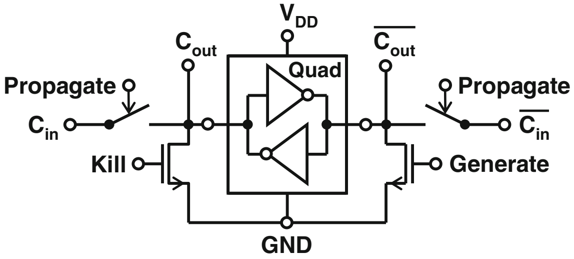


Fig. 1.8 Differential transmission-gate circuit with differential output amplifier [14]

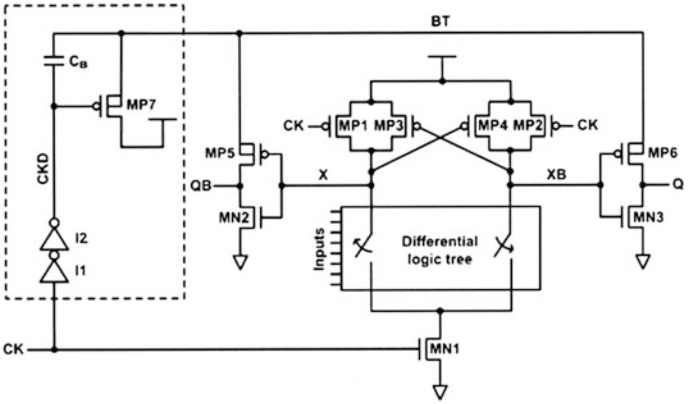


Fig. 1.9 Low-voltage, bootstrapped differential CMOS logic [15], © IEICE Electronics 2008

differential 16 × 16 b multiplier should reach, by 2020, in a 20 nm technology, 600 MOP/s at 1 fJ/operation, resulting in a Throughput FOM of 600 GOP/s/pJ in 2020 in 20 nm at 500 mV.

One noteworthy milestone along this path of sub-threshold, differential logic is a 64 b adder, reported in 2008 [15] for a 180 nm technology. In their extensive simulations, the authors show the significance of differential signal regeneration and differential drivers, enhanced with bootstrapping the output pairs as shown in Fig. 1.9. For a 64 b adder, hey obtained an add-time of 17.3 ns, allowing 58 MOPS, with 175 fJ/operation, resulting in a remarkable

- Throughput FOM of 331 MOP/s/pJ for a 64 b adder in a 180 nm technology at 500 mV in 2008.

In the class of sub-threshold, differential transmission-gate (DTG) logic, the outstanding 2014 result is a JPEG encoder in 40 nm CMOS technology [16]. Its design is based on a library of ultra-low-voltage, DTG, variance-resilient circuits [17], published in 2012. The 2014 publication on the JPEG encoder shows the performance over a voltage range from 210 to 550 mV including the statistics of numerous samples, as shown in Fig. 1.10. The best energy efficiency of 29 pJ/pixel was obtained at 330 mV with a frequency of 41 MHz. The energy rises only linearly to 45 pJ/pixel at 530 mV, while the frequency increases to 240 MHz so that the throughput-FOM would be 3.8-times higher. The macro at the heart of the Discrete-cosine transformation is a 15 b × 15 b multiplier. Thanks to personal communication with the authors, we were able to include the data on this multiplier in Table 1.2 and in Fig. 1.11. The result is a

- Throughput-FOM of 210 MOP/s/pJ at 330 mV, and 880 MOP/s/pJ at 530 mV, in 40 nm CMOS in 2014.

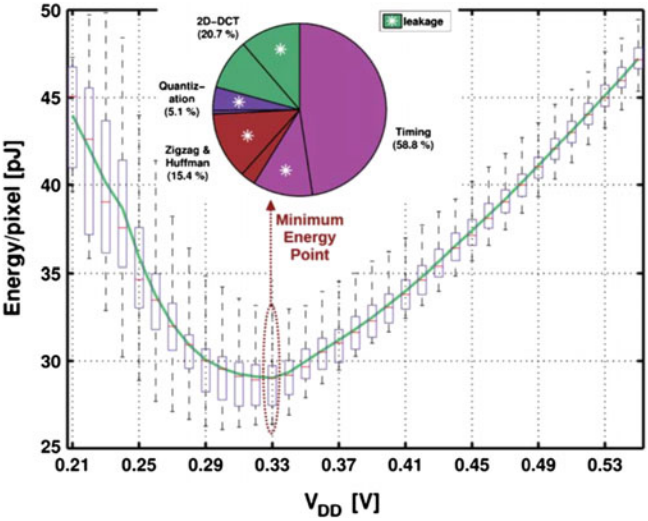


Fig. 1.10 The ultra-low-voltage JPEG encoder in 40 nm transmission-gate logic achieves its min. energy point (MEP) for a supply voltage of 330 mV, where it still performs 41MOPS [16]. The *pie* chart shows the components of energy and leakage at the MEP. © 2014 IEEE-ISSCC

Table 1.2 Energy-efficient multipliers 2000–2020

Ref. Year	Operation	Node (nm)	Voltage (V)	MOP/s	Power	Energy/Op.	FOM: GOP/s/pJ
2000 [14]	16 b × 16 b (6 b)	800/100[b]	0.5	25	10 μW	400 fJ	0.06
2008 [15]	64b adder	180	0.5	58	10 μW	170 fJ	0.34
2012 [10]	24 b × 24 b (6 b)	32	1.05	1400	8.4 mW	6.2 pJ	0.22
			0.325	25	21 μW	0.8 pJ	0.03
2012 [18]	16 b × 16 b	45	1.0	435	7.3 mW	17.4 pJ	0.025
			0.53	85	420 μW	4.8 pJ	0.017
2020 [1]	16 b × 16 b (6 b)	20	0.4	600	0.6 μW	1 fJ	600
2014 [16]	15b × 15b	40	0.53	240	4.3 μW[a]	272 fJ	0.88
			0.33	41	0.5 μW[a]	197 fJ	0.21

[a] 15-cycles pipeline
[b] 800 nm with 100 nm T-gate transistors

As we shall see in the following comparisons, these FOM's on tested multipliers and adders are the highest, reported by the spring of 2014, to our knowledge.

Previously in [1], we have selected multipliers as the key elements for logic operations because their complexity normally rises with n^2, the square of their word lengths n and because they determine the speed, measured in operations/s. Our master chart to assess the state-of-the-art and future projections, is given in Fig. 1.11.

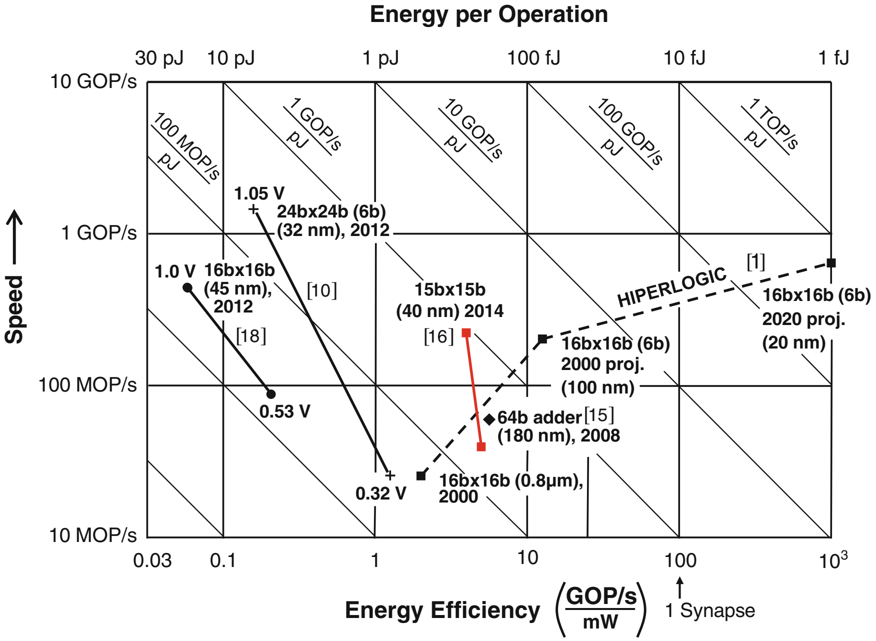


Fig. 1.11 Performance of digital multipliers and adders as the most critical logic circuits (besides memories, Chap. 11). The chart shows the speed in operations/s (OP/s) versus the energy/operation in Joule (J) on the *upper scale*, decreasing from left to right, or the power efficiency in Giga-OP/s per mW on the *lower scale*, increasing from *left* to *right*. The *diagonals* mark the throughput figure-of-merit, measured in speed (GOP/s) over the energy per operation: The highest speeds achieved at minimum energy per operation would be found in the upper right of the chart

The dashed curve [1] with the label HIPERLOGIC summarizes the results reported in [1] and the projections towards 2020, using a DTG logic and a *natural*, leading-one-first multiplier with 6 b accuracy and linear complexity O(n) (see also Chap. 10). The results reported in 2000 were 25MOPS at 70 fJ/operation for an 800 nm-node with 100 nm T-gate transistors [14]. Its potential is 600MOPS at 1 fJ in a 20 nm FD SOI technology with monolithic-3D CMOS quad drivers at 400 mV in 2020, for a

- Throughput-FOM of 600 GOP/s/pJ for a 16 b × 16 b multiplier in 2020, another 600-times better than the top throughput-FOM quoted above for 2014.

Figure 1.11 also shows the top results for standard-cell-based multipliers in 40 and 28 nm technologies [18, 19], which included detailed data on their performance, when the voltages were lowered towards 530 mV and 330 mV, respectively. The data is also listed in Table 1.2.

The standard-cell CMOS multipliers in 2012 needed between 10 and 1 pJ/operation with little progress towards 2014, where they were rated at 0.2 pJ for 8 b × 8 b and 3.1 pJ for 32 × 32 b [19], with the typical quadratic increase of the energy with word-length.

The energy efficiency of standard-cell CMOS multipliers is ~10-times worse than that of differential transmission-gate (DTG) multipliers,

a very clear signal that this circuit technique (DTG) needs to be established as the new logic style for advanced nano-CMOS technologies.

From the 2014 state-of-the art of DTG multipliers at 200 fJ/operation, what will it take to reach 1 fJ in 2020, as Fig. 1.11 suggests?

We see four big steps to get there:

1. **10**×: Leading-Ones-First (LOF) multiplication (Chap. 10) reduces the number of transistors by one order-of-magnitude.
2. **4**×: Monolithic 3D integration of transistors significantly reduces interconnect capacitance and delays = speed-up (Chap. 3),
3. **3**×: Reduced overhead due to few cycles versus 15 cycles. Speed-up.
4. **2**×: Progress of three technology nodes.

The weights may vary, but the opportunities are significant. The multipliers are useful benchmarks. However, there are so many macros, which can benefit from subthreshold DTG logic.

How about scaling beyond 20 nm?

On top of the missing returns,the further scaling of logic poses serious reliability problems [21].

Ultra-low-voltage differential-signal operation also benefits communication and memory, as we shall see in Chaps. 5 and 11. And it is one aspect in the holistic development of low-voltage electronics treated next in Chap. 2.

1.7 Chip Stacks

Chip stacks have become rapidly the largest, fast-growing development area besides the continuing development effort on the nanometer roadmap. The stacking of memory chips has advanced faster than our earlier predictions [1]. Stacks of 32 wafers have been in production for NAND-Flash memories since 2014. The stacking of heterogeneous chips, like processors and memories or MEMS and processors advances faster than expectations (see Chap. 15) (Fig. 1.12).

1.8 Single-Electron-Transistor Technology

Single-Electron IC's with an electron Coulomb-confined in a transistor channel were the research hit at the turn of the millennium. Low operating temperatures and high switching voltages (~ 10 V) have put this approach off the list. As we have shown in this chapter of CHIPS 2020, our 10 nm FD-SOI reference transistor is statistically a single-electron transistor operating at normal environment temperatures with a

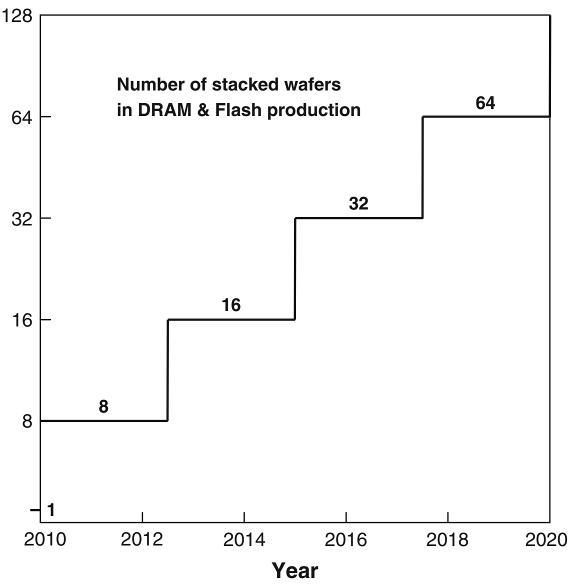


Fig. 1.12 Number of stacked wafers in production for NAND-flash memories. Stacks of 32 wafers were introduced in 2013

practical operating voltage of 200 mV at an internal transistor switching-energy level of just 5 eV. The development of Coulomb-confinement-based circuits and systems is not in the critical path towards energy-efficient nanoelectronics.

1.9 Conclusion

The three years since 2012 have shown more and more the end of the nanometer roadmap at 1x nm, where x > 10 for DRAM and logic and x = 4 for SRAM. Maximum clock speeds have saturated at <5 GHz, mostly because of the variance of transistors and circuits at <40 nm, and because of limitations on the supply-voltage. The optimization of the transistors advanced in two directions: Dual-or triple-gate transistors (FinFET'S) with large effective widths for drive capability and FD-SOI (fully-depleted Silicon-on-insulator) transistors with a primary gate and a second control with a buried gate aimed at min. footprint, capacitance and switching energy. The latter is a part of the S3S special-interest group. Processor performance has shown that FD-SOI beats FinFET at the same clock frequencies in energy-efficiency by significant factors.

The inertia in exploiting the hundreds of billions of $ invested in standard, fully complementary CMOS libraries, has caused little progress in the energy-efficiency of processors, other than the attempts to improve the energy-efficiency by turning down the supply voltage, with dramatic losses in speed, so that the throughput would have to be maintained by operating many of these processors in parallel, losing out on the total energy bill. We introduced a throughput FOM, which checks the ratio of operations/s divided by the energy/operation. This assessment has shown that differential transmission-gate logic beats standard-cell logic, especially for low-voltage, energy-efficient processing, in energy-efficiency by an order-of-magnitude and in the throughput FOM by more than an order-of-magnitude.

A breakthrough in vertical, 3D transistor integration on-chip is the series-connected-poly-Si-transistor towers for ultra-high-density NAND Flash memories, which reached 128 Gb/chip production levels in 2013. 3D chip stacks advanced faster than expected with 32 wafer-layers and TSV's (through-Silicon vias) fused in a production process in 2013.

Ultra-low voltage operation and 3D integration have become the two most important technology drivers, and they receive detailed attention in the two following Chaps. 2 and 3.

References

1. Hoefflinger, B.: The future of 8 chip technologies, chapter 3. In: Hoefflinger, B. (ed.) CHIPS 2020—A Guide to the Future of Nanoelectronics, pp. 37–93. Springer, Berlin (2012). doi:10.1007/978-3-642-23096-7_3
2. Cathelin, A.: RF subcommittee. In ISSCC 2014 Trends, pp. 13–15 (2014), http://isscc.org
3. Kim, S.H.: Germanium Source Tunnel Field Effect Transistors for Ultra-Low-Power Digital Logic. University of California Berkeley Tech. Report No. UCB/EECS-2012-87, May 2012, http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-87.html
4. Hoefflinger, B.: Chips 2020—Ein Ausblick in die Halbleiterwelt von übermorgen, ELEKTRONIK, Heft 1/2000, Seite 10 ff., WEKA Medien, Jan 2000
5. Max. clock frequencies. In: ISSCC 2014 Trends, http://isscc.org
6. IEEE S3S: SOI-3D-Subthreshold Microelectronics Technology Unified Conference, www.ieee.org
7. Wilson, R., et al.: A 460 MHz at 397 mV, 2.6 GHz at 1.3 V, 32 b VLIW DSP, Embedding F_{max} Tracking, 2014 ISSCC Dig. Technical Papers, paper 27.1, pp. 452–453, Feb 2014
8. Hsu, S., et al.: A 280 mV-to-1.1 V 256 b Reconfigurable SIMD Vector Permutation Engine with Two-Dimensional Shuffle in 22 nm CMOS, 2012 ISSCC Dig. Tech. Papers, pp. 178–180 (2012)
9. Jain, S., et al.: A 280 mV-to-1.2 V wide-operating-range IA-32 processor in 32 nm CMOS. In: 2012 ISSCC Digestive Technology Papers, pp. 66–68 (2012)
10. Kaul, H., Anders, M., et al.: A 1.45 GHz 52-to-162 GFLOPS/W variable-precision floating-point fused multiply-add unit with certainty-tracking in 32 nm CMOS. In: 2012 IEEE International Solid-State Circuit Conference (ISSCC), Digestive Technology Papers, pp. 182–183, Feb 2012
11. Sekar, D.C., Or-Bach, Z.: Monolithic 3D-IC's with single crystal silicon layers
12. Courtland, R.: IEEE Spectrum 2014
13. Handy, J.: An alternative kind of vertical 3D NAND string, published Nov. 8, 2013, http://thememoryguy.com/wp-content/uploads/2
14. Grube, R., Dudek, V., Hoefflinger, B., Schau, M.: 0.5 V CMOS logic delivering 25 million 16 × 16 bit multiplications at 400 fJ based on a 100 nm T-Gate SOI technology. Best Paper Award. IEEE Computer Elements Workshop, Mesa, AZ, 2000, 5 p
15. Jung, B.H., Kang, S.C., et al.: Novel bootstrapped CMOS differential logic family for ultra-low voltage SoC's. IEICE Electron. Expr. **5**(18), 711–717 (2008). doi:10.1587/elex.5.711

16. Reynders, N., Dehaene, W.: A 210 mV 5 MHz variation-resilient near-threshold JPEG encoder in 40 nm CMOS, 2014 ISSCC Digestive Digital Papers, paper 27.3, pp. 457–458, Feb 2014 (and private communication)
17. Reynders, N., Dehaene, W.: Variation-resilient building blocks for ultra-low-energy sub-threshold design. IEEE Trans. Circuits Syst. II **59**(2), 898–902 (2012)
18. Pawlowski, R., Krimer, E., et al.: A 530 mV 10-lane SIMD processor with variation-resiliency in 45 nm SOI. In: 2012 IEEE International Solid-State Circuits Conference (ISSCC), Digestive Technology Papers, pp. 492–493, Feb 2012
19. Horowitz, M.: Computing's energy problem (and what we can do about it), ISSCC 2014, paper 1.1, pp. 10–14, Feb 2014
20. Razavi, B.: The Cross-Coupled Pair, Part I, Solid-State Circuits Magazine. Part II, Solid-State Circuits Magazine, Fall 2014, pp. 2–12 (2014)
21. Borkar, S.: Exascale Computing—Fact or Fiction? SSCS Webinar (2014)

Chapter 2
The Future of Low-Power Electronics

Toshiaki Masuhara

Abstract The number of integrated transistors has increased so rapidly that it has become evident that a further increase of the performance is limited by the power dissipation. The future IT and electronics, therefore, require more efficient power-reduction solutions. In the human brain and nerve system, analog signals from sensing orgasms are processed by a network composed of neurons and synapses. This process is slow but very power-efficient because it is done by below-100 mV signal levels.

Although near-threshold or sub-threshold operation of digital CMOS circuits would be possible candidates for the future low-power electronics, the operating voltage was not scaled due to the fact that the sub-threshold characteristics of MOSFETs are not scalable. Various parameter variations of the MOSFETs also deteriorate the noise margin. Most of the existing non-volatile memories and switches have problems in operating at low voltages. It is evident that off-chip interface circuits, power delivery- and control-means, such as back-bias, and the protection devices against electrostatic discharge, also need to be integrated. If the power of the chip is greatly reduced, stacked-chip 3D integration could be an efficient solution in the future IT and electronics.

2.1 Electronics Systems and Power-Efficiency

It is expected by the JEITA Green IT Council that the power consumption of the ten major equipments (PCs, servers, storages, routers, displays, TVs, DVD-players, air-conditioners, refrigerators, and illuminations), and of the data centers increase

T. Masuhara (✉)
LEAP—Low-Power Electronics Association and Project, Niikura-bldg. 8F, 2-2, Kanda
Tsukasa-machi, Chiyoda-ku, Tokyo 101-0048, Japan
e-mail: masuhara@leap.or.jp; toshi-masuhara@bridge.ocn.ne.jp

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_2

rapidly as shown in Fig. 2.1a [1]. In the data centers, the electric power is expected to rise up to 2500 TWh/year in 2050 that is more than 17-times higher than that in 2005. This is due to the processing of big data and the rapid increase of the transactions of big data through the internet and data centers. The fraction of the electric power consumption by servers, storages, routers and switches are also shown in Fig. 2.1b [1]. It is evident that disruptive low-power technologies are needed in the future electronics and IT.

The evolution of the *computations/kWh* of computers is illustrated in Fig. 2.2. The data points for the *computations/kWh* were taken from the paper by Koomey et al. [2]. The vacuum-tube technology was replaced by the bipolar-transistor technology in the 1950s, then by bipolar or NMOS-IC technology in the 1960s. The bulk-CMOS, developed by Wanlass and Sah [3], had a feature that the stand-by power is zero due to the fact that either the pMOS or the nMOS transistor, forming a logic gate, is turned off in the digital circuit. However, since the bulk-CMOS is formed in a highly doped compensated well, the speed was not fast enough to be applied to high-end applications like NMOS. The application was limited to watches and calculators that allow slow speed. The development of the high-speed SRAM [4] in the late 1970s by the author's group was the first demonstration that bulk-CMOS can be applied to high-end integrated circuits. In the 1980s and 1990s, the use of bulk-CMOS has been expanded to most of the major integrated circuits, and it has been used for more than 30 years as the dominant technology. The rapid increase of the number of integrated transistors for more functions and performance, however, caused the rapid increase of the power dissipation. In the paper by Chen [5], it was pointed out that the module heat flux determines the limitation and causes the replacement of the dominant technology solution. He also pointed out

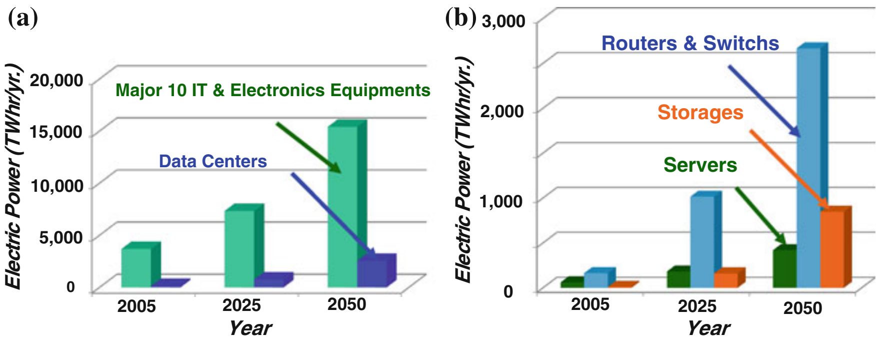


Fig. 2.1 Worldwide electric-power requirements for green IT technology estimated by JEITA (Japan Electronics and Information Technology Industries Association) Green IT Council 2013 [1]. **a** Electric-power requirement for major ten IT and electronics equipments (PCs, servers, storages, routers, displays, TVs, DVD players, air-conditioners, refrigerators and illuminations) and data centers. **b** Electric-power requirement for servers, storages, routers and switches

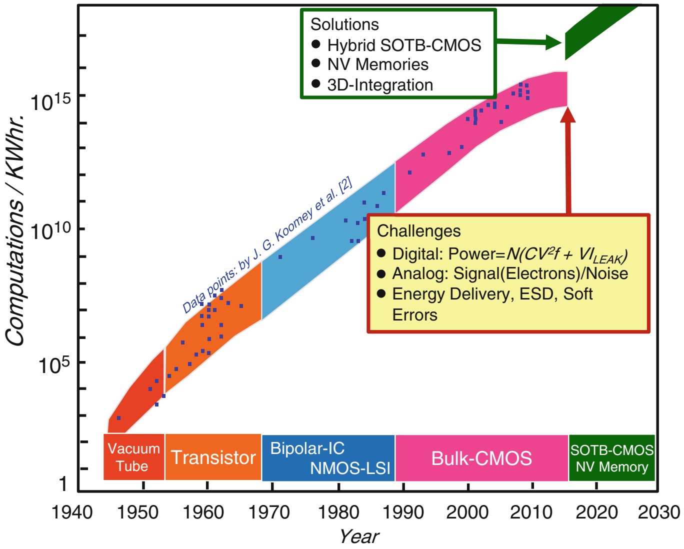


Fig. 2.2 Evolution of the computations/kWh and the major device technologies. Data points are taken from Koomey's paper [2]. Bulk CMOS needs to be exchanged by another technology solution by 2020 due to the power-dissipation limit (© 2009 IEEE)

that, around 2010, ultra-low-voltage and 3D technology would replace the conventional bulk-CMOS. Although many techniques to decrease the power have been developed so far, the saturation of the *computations/kWh* is expected in the future due to the heat-dissipation problem. It is, therefore, expected that the current bulk-CMOS will be replaced by another technology solution by the end of the decade due to the power-dissipation limit.

One of the most effective ways to reduce power in IT systems is the reduction of the power-supply voltage. This is because the power is determined by,

$$Power = N\left(CV^2f + VI_{LEAK}\right).$$

The scaling of the device size and the supply voltage in the past worked when the MOSFET size was above 14 nm. Due to the leakage and sub-threshold slope of MOSFETs that are not scalable, the scaling below 14 nm has not been as efficient as it has been in the past. Challenges associated with the reduction of the power supply need to be solved at the same time. In digital circuits, they must be stable against parameter, voltage and temperature (*PVT*) variation. This means that enough margin needs to be ensured against noises, such as power-supply bounce, EMI noise, 1/*f* and the random telegraph noise of MOSFETs, and the soft errors due to noise by high-energy particles. Secondly, to convert analog signals from the inputs to digital, enough signal-to-noise ratio is needed, which requires that analog circuits operate at a higher supply voltage than that of the digital part. Thirdly, integrated circuits must be operated with on-chip stable and controllable energy-delivery means, and they must be robust against electronic static discharge (ESD).

In the future society, wireless sensor networks (WSN) that utilize autonomous and maintenance-free sensors, will become essential parts of the society as illustrated in Fig. 2.3. The sensor network will be used in many applications including agriculture and food supply, medical services and health care, traffic control, monitoring of the social infra-structure, and ambient sensing such as river, weather, lands, and ocean. A technology solution for both low power and maintenance-free features are required. The sensors require not only digital processing of data, security and ID, but also wireless interfaces with extremely-low-power features. Either the energy harvesting, or perpetuum operation, in other words, the life-long operation with an installed battery becomes essential in such applications. Future mobile equipment, such as robots and smart mobility, also requires an energy-efficient and high *computations/watt* solution.

A project called, "*Ultra-Low Voltage Device Project for Low Carbon Society*", was performed in the *Low-Power Association and Project (LEAP)* since 2010–2015. The purpose of the project was to develop enabling technology solutions that reduce the power of the electronics and IT equipments by one order of magnitude against that of the existing ones by the below-0.4-V operation of the integrated

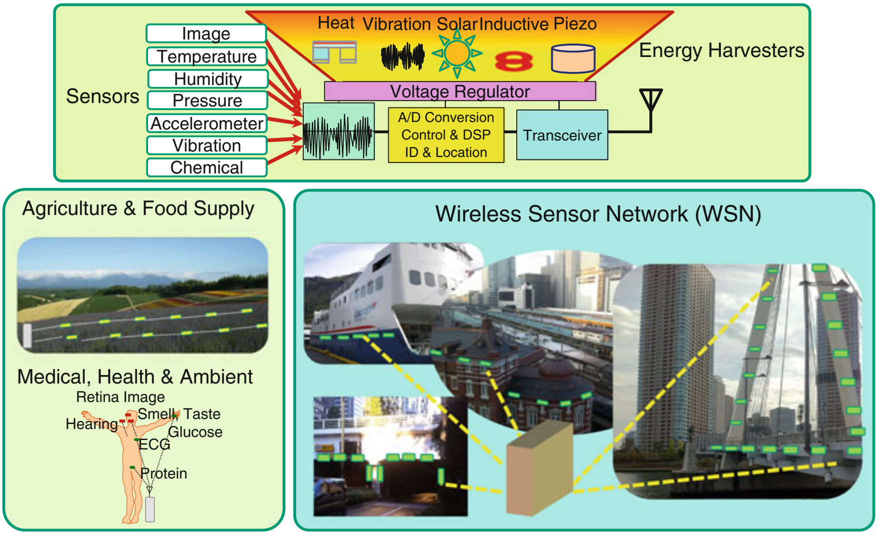


Fig. 2.3 Internet-of-things (IoT) paradigm requiring technology-solutions featuring low power, maintenance-free, and wireless access to internet (© LEAP)

circuits. In the project, three enabling technologies were developed. The first technology is the new CMOS operable at 0.4-V with much less standby current. The second solution is the nano-carbon interconnect for use in 3D flash memories such as BICS [6]. The third solution is the development of low-voltage non-volatile resistive memories. Three types of non-volatile memories and a switch were chosen. These include an embedded MRAM for cache application, a switch for post-fabricated logic configuration, and a new memory called TRAM for tier-0 storage. These were integrated in the backend process to avoid excessive thermal history.

2.2 Low-Power CMOS Technology

2.2.1 Hybrid SOTB CMOS Technology

The most effective way to reduce the power-supply voltage is to reduce the parameter variation of the MOSFETs. The SOTB (Silicon on Thin-buried-Oxide) MOSFETs, a variation of the fully-depleted (FD) SOI MOSFET with thin BOX (buried oxide) layer, was first developed by Tsuchiya et al. [7]. The SOTB-CMOS is a good candidate for low-voltage operation because of the low impurity concentration in the channel region resulting in small parameter variations. It also has a thin BOX layer to control the threshold voltage by applying a back-bias. In the *Ultra-Low Voltage Device Project*, a practical hybrid SOTB CMOS integrated circuit shown in Fig. 2.4, was developed and demonstrated. In the developed SOTB CMOS, the bulk-CMOS portion is fabricated by stripping off the BOX layer used for the SOTB-CMOS, where the SOTB-CMOS are optimized for low-voltage logic and SRAMs. The bulk-CMOS is used for I/O circuits, analog circuits such as A/D, D/A converters, ESD protection devices, and on-chip power converters where a higher voltage capability is required, and the legacy circuits.

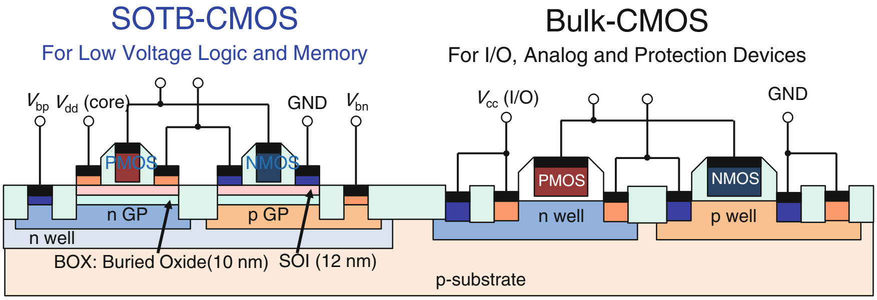


Fig. 2.4 Cross-section of the hybrid SOTB-CMOS. The SOTB-CMOS is optimized for low-voltage logic and SRAMs. The bulk-CMOS is used for I/O, analog, ESD protection and on-chip power converters, requiring higher voltage, and legacy circuits [8] (© 2014 IEEE)

An extensive study of the MOSFET parameter variation for bulk-CMOS was done for threshold-voltage V_{th}, and on-current I_{on}, in the MIRAI project as described in the NEDO project report [9]. The variation of V_{th} and I_{on} is originated by a non-uniform potential barrier in the channel region due to the random dopant fluctuation, and the variation of the gate work-function, the gate edge-roughness, the resistance of the source-drain extension, and the stress and the strain in the MOSFETs. It is proven that the random dopant fluctuation is the major origin of the variation. Figure 2.5 illustrates the simulated *DIBL* and V_{THC} variations for *2000* bulk and SOTB MOSFETs by Prof. Hiramoto, the Univ. Tokyo [10]. It is evident that the variation of the barrier-lowering due to random dopant fluctuations gives rise to the fluctuation of the channel potential, resulting in the variation of the *DIBL* and V_{THC}. It is also shown that both the variation of *DIBL* and V_{THC} is reduced in the FD-SOI as compared to those in the bulk-MOSFET. Better channel potential control by the gate-field in FD-SOI makes it possible to reduce the doping in the channel region, if the adjustment of the V_{TH} by the gate work-function is done properly.

I_d-V_g characteristics of the *high*-V_{th} (*HVT*), *medium*-V_{th} (*MVT*), and *low*-V_{th} (*LVT*) SOTB MOSFETs were optimized for 0.4 V operation as shown in Fig. 2.6 [11]. The threshold voltages are tuned by the Hf and Al in the gate so that the nMOS and pMOSFETs exhibit the symmetrical characteristics at zero gate voltage. The variation of the V_{th} for 1 M MOSFETs is shown in Fig. 2.7 [12]. It is demonstrated that the SOTB MOSFET has a smaller V_{th} variation, in other words, a smaller worst V_{th} value, that makes possible a lower-voltage operation. At the same time, the SOTB MOSFET realizes a smaller I_{ON} variation that results in a larger worst I_{ON}. This is an attractive feature in realizing higher performance at low supply voltages.

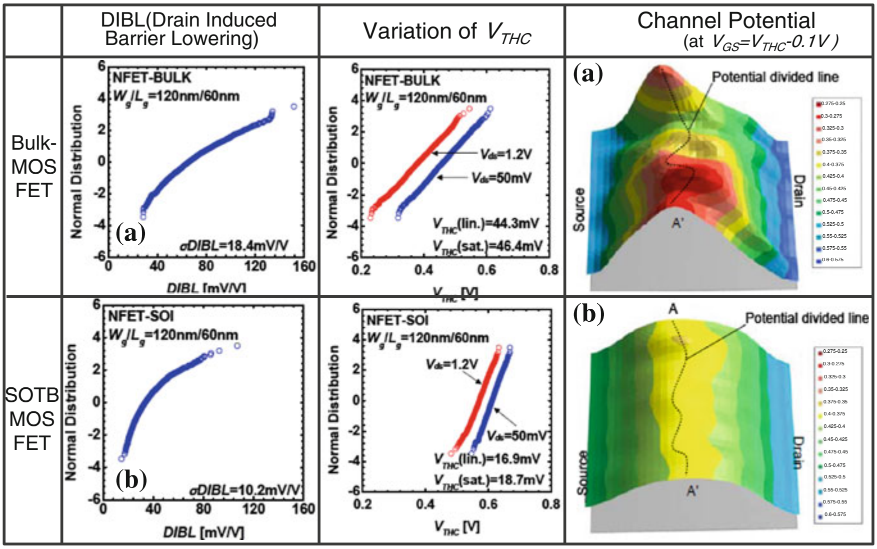


Fig. 2.5 Simulated *DIBL*, and V_{THC} variations for 2000 transistors in the bulk and SOTB MOSFETs. The potential diagrams in the channel are shown to the right by different colors corresponding to 0.025-volt scale [10] (© 2010 IEEE)

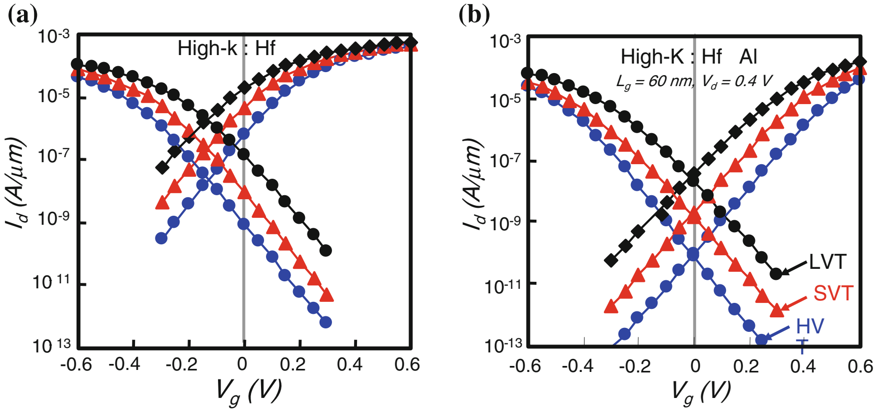


Fig. 2.6 I_d-V_g characteristics of SOTB nMOS and pMOSFETs optimized for 0.4 V operation. The threshold-voltages are tuned by the Hf and Al so that nMOS and pMOSFETs exhibit the symmetrical characteristics at zero-gate voltage [11] (© IEEE). **a** I_d-V_g characteristics with Hf in the gate-material. **b** I_d-V_g characteristics for LVT, SVT, and HVT MOSFETs with Hf and Al in the gate-material

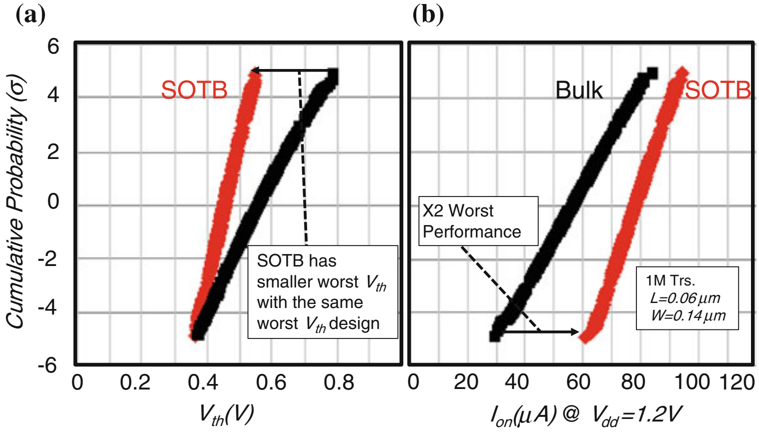


Fig. 2.7 Comparison of the variation (cumulative probability) of the V_{th} and I_{on} for 1 M SOTB and bulk nMOSFETs [12] (© IEEE). **a** Cumulative probability of the V_{th} of 1 M SOTB and bulk nMOSFETs. **b** Cumulative probability of the I_{ON} of 1 M SOTB and bulk nMOSFETs

2.2.2 Low-Voltage SRAM

The low-voltage SRAM was developed using 65 nm SOTB CMOS technology. Figure 2.8 illustrates (a) the cross-section of the SOTB transistor, (b) the SEM photo of the SRAM memory cell, (c) access-time versus supply-voltage V_{dd}, (d) fail-bit distribution versus V_{dd}, and (e) active/standby power of the SRAM when

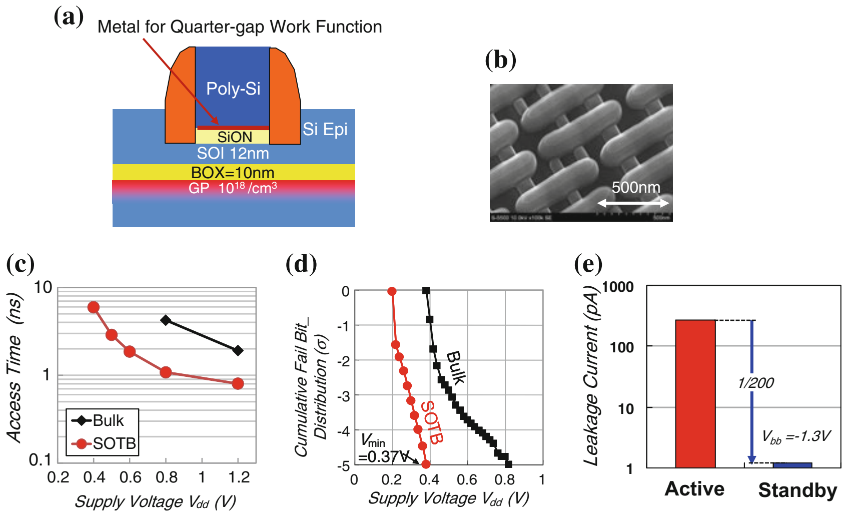


Fig. 2.8 2 Mb, 65 nm SOTB-CMOS SRAM operated at 0.37 V [12] (© 2013 IEEE). **a** Cross-section of the SOTB transistor. **b** SEM Photo of the SRAM cell. **c** Access time of the SRAM. **d** Fail-bit of the SRAM as a function of V_{dd}. **e** Leakage-current of the SRAM cells

a back-bias voltage of −1.3 V is applied [12]. Although several SRAM circuit techniques were developed to operate the SRAM at low voltage, this demonstration was simply done by using the conventional 6-transistor SRAM cell without any voltage-boosting techniques for the word- and bit-lines to compare the technologies. It can be seen that the developed SOTB-CMOS SRAM operates at an access-time of 5.5 ns at 0.4 V, whereas the bulk-CMOS using the same design cannot be operated below 0.8 V. This indicates that the SRAM does not need ECC (error-correction-code) to rescue fail-bits, whereas the tail-portion of the fail-bits in the bulk-CMOS SRAM are usually rescued by the ECC circuits to increase the yield in low-voltage operation. It is also seen in Fig. 2.8e, that the leakage-current in the standby-mode could be efficiently reduced by more than 2 orders by applying a V_{bb} = −1.3 V as compared to the leakage in the active mode. This indicates the effectiveness of the threshold-voltage control in SOTB-CMOS.

2.2.3 *Low-Voltage Microprocessor and Logic Circuits*

A low-voltage microprocessor with 144 KB SRAM was developed using 65 nm SOTB CMOS technology [8]. The photomicrograph of the fabricated chip and the simple block-diagram is shown in Fig. 2.9. Since a bulk-CMOS microprocessor can be integrated on the same chip, a comparison of the performance and of the energy for both designs was done as shown in Fig. 2.10. It is exhibited that the SOTB-CMOS microprocessor could be operated at 0.22 V with a clock frequency

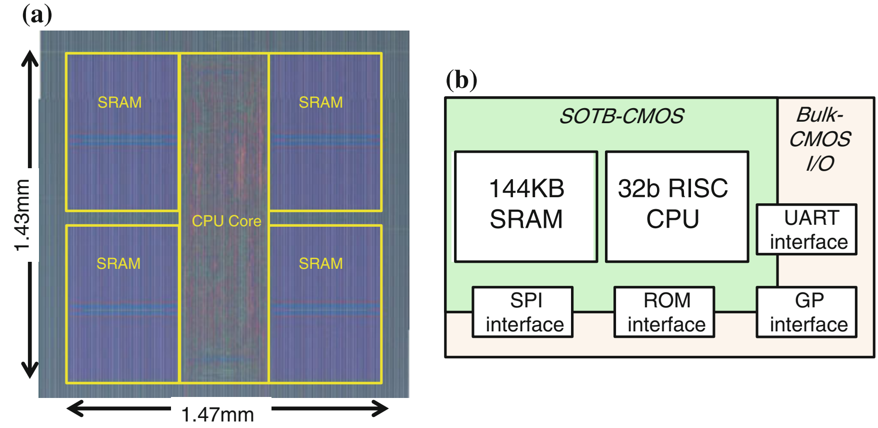


Fig. 2.9 SOTB-CMOS microprocessor with 144 KB SRAM that operates at less than 0.5 V [8] (© 2014 IEEE). **a** Photomicrograph of the chip. **b** Simplified block-diagram of the 32 b-RISC microprocessor

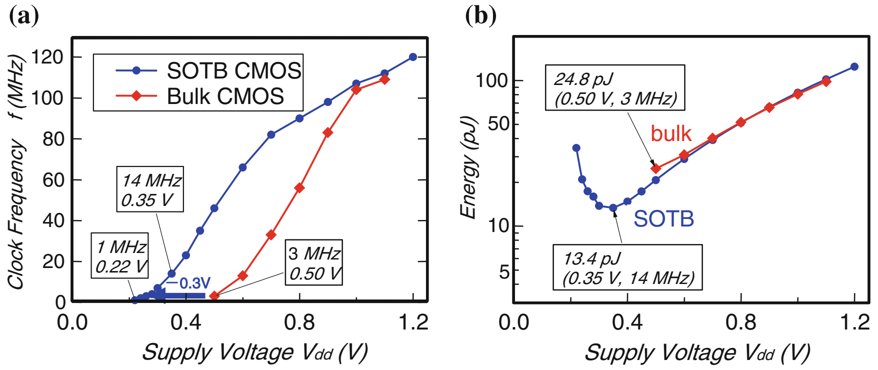


Fig. 2.10 Performance of the SOTB-CMOS and the bulk-CMOS microprocessor [8] (© 2014 IEEE). **a** Clock-frequency as a function of the supply-voltage. **b** Energy (pJ) as a function of the supply-voltage

of 1 MHz whereas the minimum operating voltage of the bulk-CMOS is 0.5 V. It is also seen from (b), that 14 MHz operation was performed at the minimum energy of 13.4 pJ at $V_{dd} = 0.35$ V in SOTB-CMOS. The sleep-current of the microprocessor at 0.35 V is shown in Fig. 2.11 (a) as a function of the supply-voltage V_{dd}, and (b) as a function of the CPU temperature, indicating that the back-bias contributes to effectively reduce the sleep-current by more than two orders-of-magnitude. The comparison of the leakage-current in the bulk and the SOTB nMOSFET is shown in Fig. 2.12. In the bulk nMOSFETs, the substrate-leakage becomes the dominant portion of the total leakage-currents when negative back-bias is applied, and hence, the leakage cannot be reduced below 1/10 of that at zero back-bias. In the

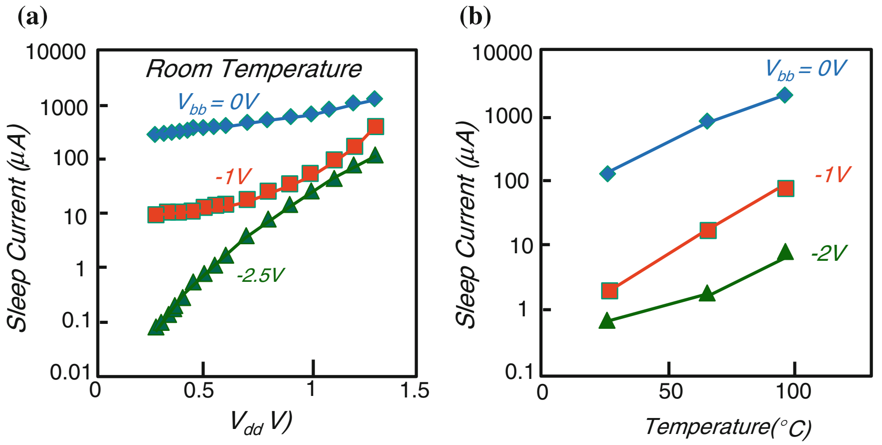


Fig. 2.11 Sleep-current of the SOTB-CMOS microprocessor [8] (© 2014 IEEE). **a** Sleep-current of the CPU as a function of the supply-voltage V_{dd}. **b** Sleep-current of the CPU as a function of the temperature

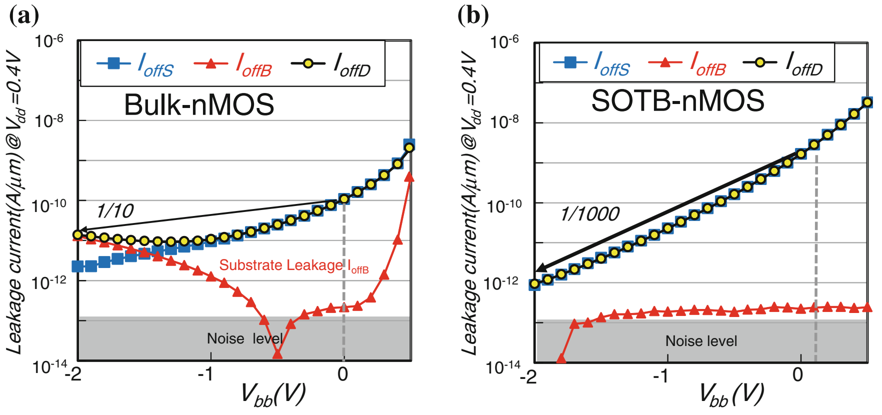


Fig. 2.12 Comparison of the leakage-currents in the bulk- and the SOTB-nMOSFET [12] (© 2013 IEEE). **a** Leakage-currents in the bulk-nMOSFET as a function of back-bias V_{bb}. Substrate leakage increases when a large back-bias is applied. **b** Leakage-currents in the SOTB-nMOSFET as a function of back-bias V_{bb}. Substrate leakage does not increase when a large back-bias is applied

SOTB MOSFET, the leakage-path does not exist because the substrate is insulated by the buried-oxide (BOX) layer. This makes possible a reduction of the leakage-current to 1/1000 of that for zero back-bias voltage as seen in Fig. 2.12b. Effective control of the threshold-voltage by back-bias makes possible to minimize the leakage-current at elevated temperatures, avoiding the abnormal operation at an elevated temperature. It also can compensate for the threshold-voltage increase at low-temperature operation. Such wide controllability by the back-bias makes possible an operation of the processor in a much wider temperature range.

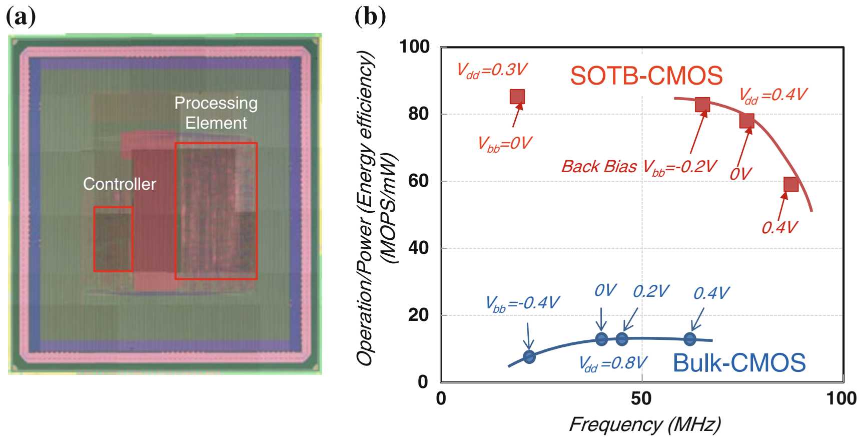


Fig. 2.13 Performance of the SOTB-CMOS accelerator CMA (cool mega array) designed by Keio Univ. (© IEICE). **a** Photomicrograph of the cool-image-array chip designed by using SOTB-CMOS. **b** Energy-efficiency (operation/power) by alpha-blender benchmark versus frequency as a function of supply-voltage and back-bias voltage

To demonstrate the performance of a logic circuit, an SOTB-CMOS Cool-Mega-Array (CMA) consisting of a controller and processing-element was developed by Prof. Amano's group, Keio Univ. [13]. The photomicrograph of the chip is shown in Fig. 2.13a. The comparison of the performance/power (efficiency) of the SOTB-CMOS CMA and the bulk-CMOS CMA was done by using the alpha-blender benchmark. The result is shown in Fig. 2.13b. The comparison shows that the SOTB-CMOS CMA can be operated at V_{dd} = 0.3–0.4 V as compared to 0.8 V in the bulk-CMOS, and it exhibits an energy-efficiency 5-times-higher than that of the bulk-CMOS CMA. This indicates that the low-voltage operation is effective to get higher MOPS/mW in digital signal- or image-processing applications. It has been pointed out that a dedicated-processor solution provides 1000-times higher performance compared to the microprocessor solution. This is due to the CPU-memory bottleneck [14]. This suggests that the combination of the low-voltage CPU and a dedicated-processing element such as CMA or off-loader enables much better performance/power solutions.

2.3 Low-Power Non-volatile Memories and Switches

In present-day electronics and IT equipment, many types of memories are used in each hierarchy. Program-execution is done in the processors with embedded memories, using data from an off-chip DRAM main memory or working buffer. Embedded memories include SRAM/DRAM L1 to L3 cache memories, ROMs or non-volatile PROM for firmware. In storage, HDD has been used as the dominant device. Solid-State Disk (SSD) using NAND-type flash memories are now replacing HDD in the tier 0, 1 class storage device.

In memory-cell circuits, such as DRAM, SRAM, or Flash memories, electronic charge, the product of *CV*, where *C* is the capacity of the memory-node and *V* is the memory voltage, is used as means to store “1’s” and “0’s”. When the capacity *C* or the voltage *V* becomes small, the memory-signal becomes small. As a result, marginal sensing of the small signal becomes a problem. Since these memory-cells are volatile, they lose memory information when the supply-voltage is turned *OFF*. It is, therefore, necessary to keep the power-supply *ON,* or re-load the memory data, when the memory turns *ON* again. In the *ON* condition, the memory-cell leakage due to the sub-threshold current needs to be either compensated for or refreshed to keep the memory charge in the cell as illustrated in Fig. 2.14a. In the non-volatile resistive-change memory cells, illustrated in Fig. 2.14b, once “high” and “low” resistivity-values are written into the memory material, the memory information can be sustained, even if the power-supply is turned *OFF*. The read-operation is done by sensing the current in the memory resistor. The sensing at low voltage, however, requires enough signal-to-noise ratio (S/N) as illustrated. Various resistive-change memories such as MRAMs, ReRAMs, and PCMs, are now in development. In this section, three types of resistive-change memories and switches are described, that have been developed by LEAP for low-voltage operation.

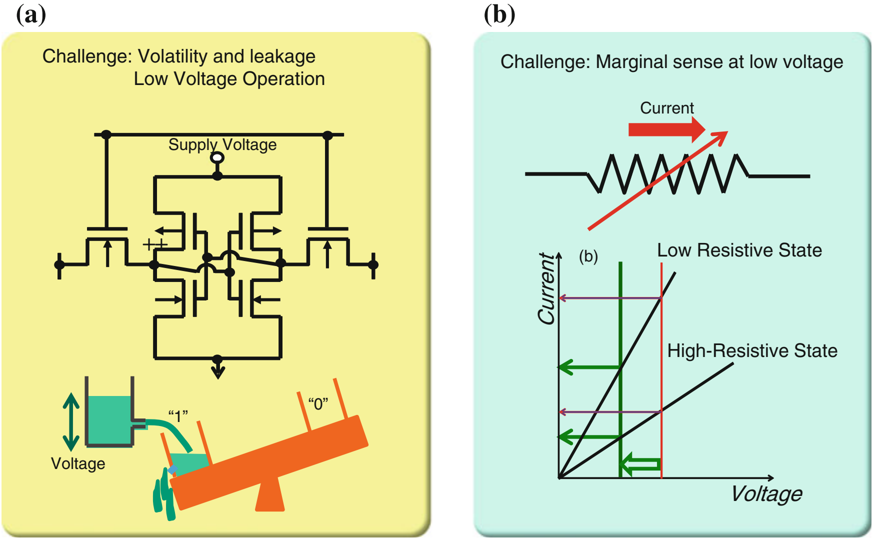


Fig. 2.14 Challenges of the low-voltage memories. In low-voltage operation, the non-volatile resistive memory cell has an advantage over the volatile voltage memory-cell due to leakage compensation and volatility (© LEAP). **a** Voltage memory cell (SRAM or 1T-DRAM). **b** Non-volatile resistive-change-type cell

2.3.1 MRAM for Cache Applications

Figure 2.15 illustrates an MTJ used in STT-MRAM for low-voltage and low-power cache applications developed by LEAP [15]. An essential part of the MRAM consists of a tunnel-insulator sandwiched by pinned and free-ferromagnetic layer. Writing “1’s” and “0’s” is done by the current-flow through the tunnel-insulator and by changing the spin-direction of the free layer. Since enough current-density is required to change the spin direction, a small MTJ is advantageous to realize a low-power MRAM. If the free layer and the pinned layer have the same spin-direction, that state corresponds to the low-resistivity state. On the contrary, if the spin-directions are opposite, the current through the MTJ becomes lower, and that corresponds to the high-resistivity state. The current-ratio depends on the ferromagnetic material of the MTJ tunnel-insulator, and their quality. For marginal operation, the current-ratio of at least over 100 %, including variation, is desirable.

MRAM is supposed to be suitable for the non-volatile data-memory for mobile applications. This does not require a fast switching-speed, but low-voltage and

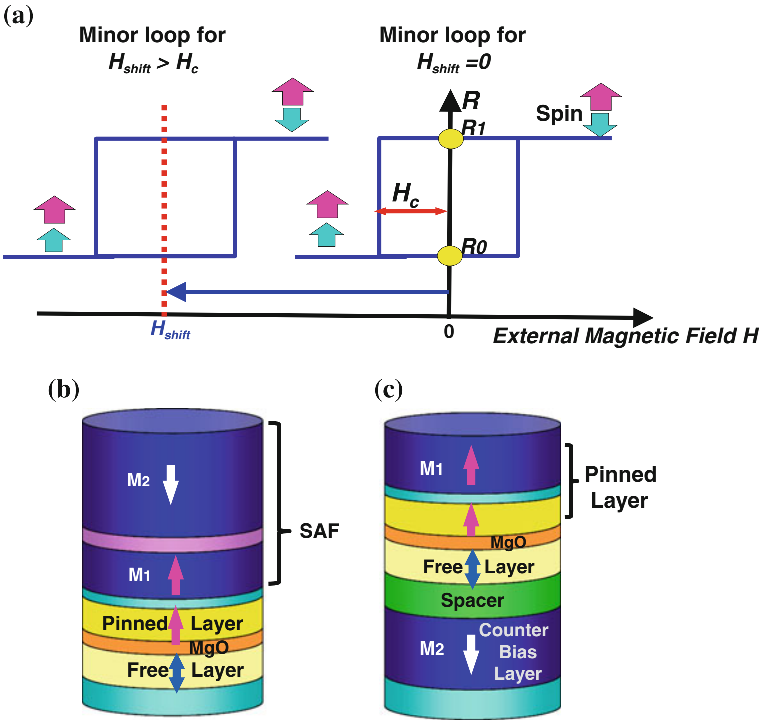


Fig. 2.15 Schematic diagram of the perpendicular magnetic tunnel-junction (MTJ) for suppressing stray-field [15] (© 2013 IEEE). **a** Illustration of the induced minor loop shift in MTJ. **b** Conventional synthetic anti-ferromagnetic (SAF) pinned layer (SP) type MTJ. **c** Counter-bias magnetic field layer (CBF) type MTJ

low-power characteristics are required. Another application of the MRAM is the embedded non-volatile cache-RAM that replaces the currently dominant SRAM or the DRAM cache. This application requires a smaller cell-area compared to that of other cache-memories, and over 10^{15} (infinite) write-erase cycles.

A Counter-Bias magnetic Field layer (CBF) type MTJ shown in Fig. 2.15 was developed for low-power cache applications [15]. In MRAM, the stray magnetic field from the pinned layer causes a shift of the minor loop H_{shift}. It is required that H_{shift} be kept as small as possible so that the symmetrical resistance-change occurs for a positive and negative external magnetic field H as shown in Fig. 2.15a. The stray magnetic field from the pinned layer, therefore, needs to be compensated for by adding an additional layer. Conventional MTJ shown in Fig. 2.15b has Synthetic Anti-Ferromagnetic (SAF) Pinned layer (SP) for this purpose. When the memory cell size is reduced for cache applications, the compensation margin of the SAF type cell may not be enough, and that makes it difficult to assure marginal operation. A Counter-Bias magnetic Field layer (CBF)-type MTJ, shown in Fig. 2.15c, exhibits a larger margin due to the fact that the counter-bias layer and spacer thicknesses could be independently optimized. The CBF MTJ cell, fabricated in the 65 nm CMOS back-end process is shown in Fig. 2.16a. The measured H_c and H_{shift} versus MTJ size in CBF structure cell for suppressing stray field indicates that the measured H_{shift} is maintained below H_c for the cell-size diameter of less than 50 nm. This shows that marginal operation is possible. To achieve read- and write-cycles of over 10^{15}, it is essential to form an MgO layer having good crystal-quality. It was verified that MgO tunnel-oxide with good crystal-quality is formed by inserting a CoFe seed layer on the CoFeB layer and oxidizing Mg deposited on CoFeB [16]. An accelerated experiment of TDDB test indicated that a write-erase cycle of over 10^{16} at 0.65 V was achieved.

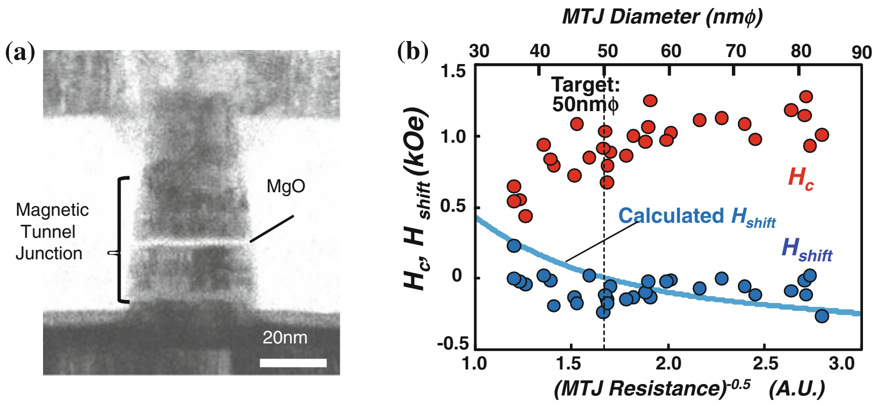


Fig. 2.16 Fabricated MTJ cell in the back-end process and the result of the stray-field suppression [15] (© 2013 IEEE). **a** Cross-sectional TEM photo of the MRAM cells. **b** Measured H_c and H_{shift} versus MTJ size in CBF structure cell for suppressing stray-field

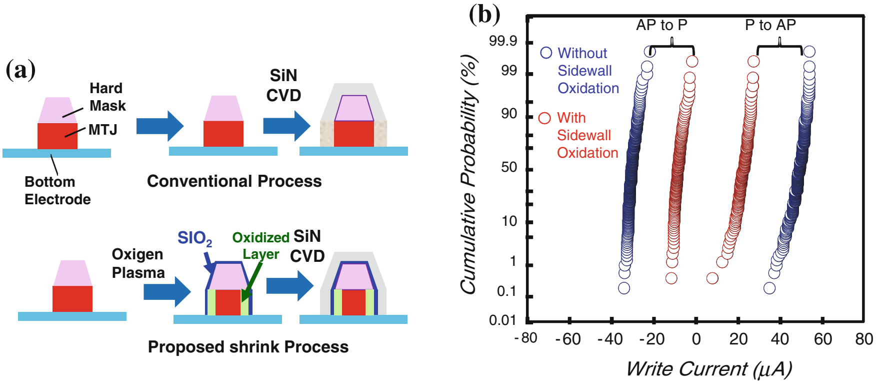


Fig. 2.17 MTJ cell-shrink process using sidewall oxidation, and write-current distribution [17] (© 2014 IEEE). STT-MRAM enables memory cell-size of 25F^2 as compared the 150F^2 SRAM cell where F is the feature size. **a** Sidewall oxidation MTJ cell shrink process. **b** Write-current distribution of the MTJ cell with sidewall oxidation (*red*) and without sidewall oxidation (*blue*)

To reduce the write-current, the MTJ size needs to be minimized. The challenge is to control the variation of the write-current with such a small cell-size. Figure 2.17a illustrates a new cell-size shrink process developed [17]. A sidewall-oxidation technique makes it possible to remove the periphery of the MTJ to reduce the cell-diameter. The write-current of as low as 20 µA was obtained without degrading the variation of the write-current as shown in Fig. 2.17b. Reduction of the write-current, and control of the variation would make possible the MRAM as a promising candidate for non-volatile cache memories in microprocessors applied to electronics and IT equipment.

2.3.2 Complementary Atom-Switch for Programmable Logic After Fabrication

In many data-processing applications, dedicated application-specific ICs (ASICs) have performance advantages over microprocessors. However, due to the rapid increase of the cost and lead (turnaround) time for design, mask, and manufacture of ASICs, the FPGA approach has become desirable in many applications. The FPGA approach is using SRAM cells and switches to configure the logic. Therefore, the SRAM area occupies a large fraction of the switch area, and a non-volatile configuration-data memory is required.

An atom-switch is a non-volatile resistive-change switch for use in programming the logic in the integrated circuits after fabrication. It utilizes the formation and annihilation of a metal-bridge in a polymer-solid-electrolyte (PSE). A schematic illustration of the atom-switch is shown in Fig. 2.18. The switch is called complementary atom-switch (CAS) [18]. As shown in Fig. 2.18a, the switch turns *ON*

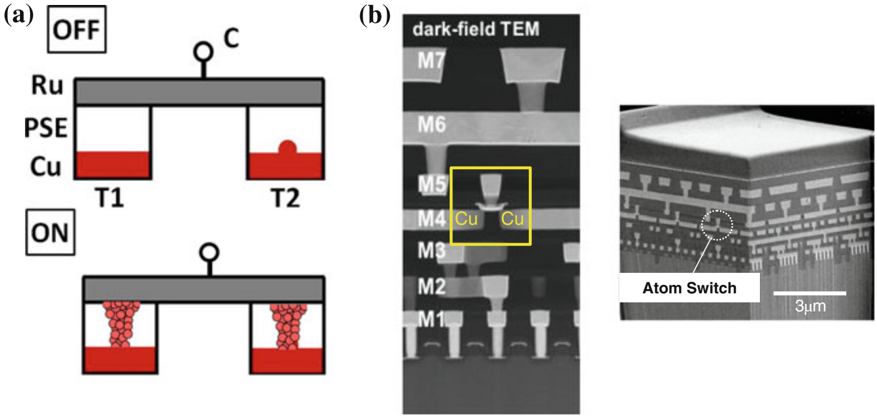


Fig. 2.18 Complementary atom-switch (CAS), a non-volatile switch device for programming after fabrication [18] (© 2012 IEEE). **a** Complementary atom-switch (CAS). **b** Cross-section of the CAS

(set), when a positive bias is applied to the Cu electrode, and it turns *OFF* (reset), when a positive voltage is applied to the Ru electrodes. The bridge is formed by the movement of the Cu ions through a polymer solid electrolyte (PSE). The CAS is formed between M4 and M5 metal layers as shown in the cross-sectional photos in Fig. 2.18b. Figure 2.19 (a) illustrates the current I_{T1C} and I_{T2C} in the set (*OFF* to *ON*)-operation, (b) I_{T1C} and I_{T2C} in the reset (*ON* to *OFF*)-operation, and (c) the ratio of I_{T1T2} (*ON*) and I_{T1T2} (*OFF*) [19]. In the set-mode, the current rapidly

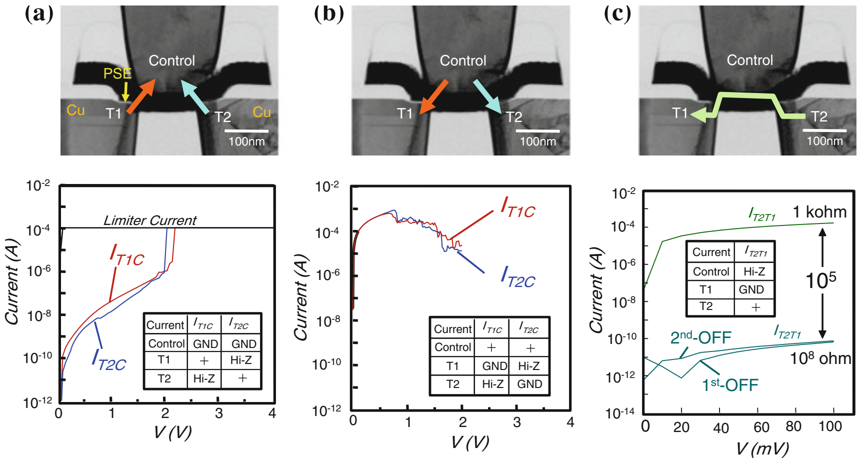


Fig. 2.19 Current versus voltage of the CAS and the ratio of the ON/OFF current I_{T1T2} of the CAS [19] (© 2011 IEEE). **a** I_{T1C} and I_{T2C} in the set (OFF to ON) mode. **b** I_{T1C} and I_{T2C} in the reset (ON to OFF) mode. **c** Ratio of the ON/OFF current of the CAS I_{T1T2}

increases at around 2 V, that is the voltage of the bridge-formation, and in the reset-operation, the current decreases logarithmically around 1–2 V. In the reading operation, the ratio of the current from T2 to T1, I_{T1T2}, is maintained to around 10^5. This makes possible a stable read-operation at a low voltage.

For programming the logic, the cell-architecture shown in Fig. 2.20 was developed [18]. The logic block consisting of 2 × 4-input LUT is programmed by the 386 CAS devices (304 for routing, 64 for LUT, and 18 for condition). The switch-block is laid out above the logic-block as illustrated in the schematic layout shown in (b). Table 2.1 illustrates a comparison of the FPGA, using a SRAM cell and a switch-transistor, and the programmable logic using CAS. The first advantage of the CAS reconfiguration over FPGA is the non-volatility. The second is the small area of the CAS compared to the SRAM cell. The three dimensional feature of the CAS layout also results in a significant area-reduction and corresponding speed and power advantage due to the reduced interconnect resistance and capacitance. Table 2.2 shows a comparison of the ASIC, FPGA, and CAS approaches in configuring logic. Although the ASIC approach has been used for many years, it suffers from design and manufacturing cost and lead-time, particularly in scaled technologies. The FPGA solution has exchanged the ASIC approach due to the zero design-cost and zero lead-time, although it has performance and chip-area penalties.

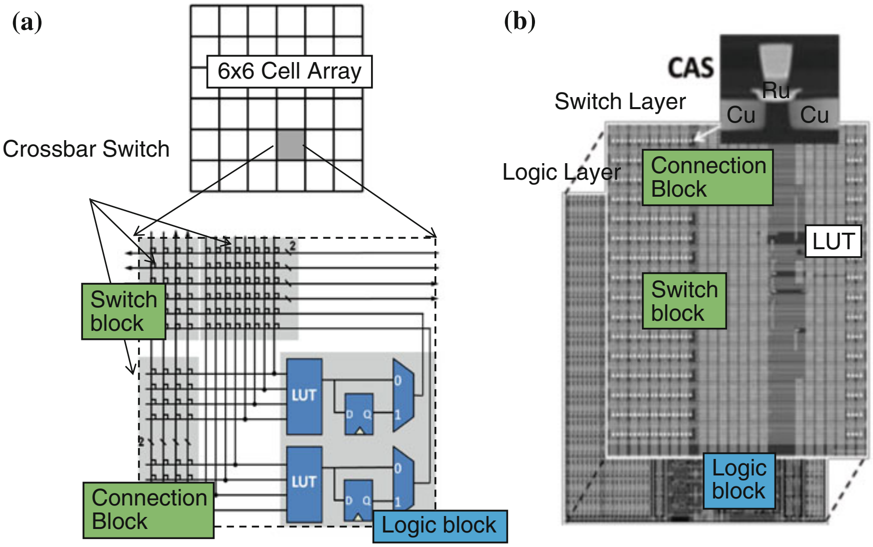


Fig. 2.20 Cell architecture of the programmable logic using CAS [18] (© 2012 IEEE). **a** Cell with 2 × 4-input LUT with 386 CAS (304 for routing, 64 for LUT, and 18 for condition) for 6 by 6 programmable logic cell (*top*). **b** Conceptual 3D-layout of the logic-cell

Table 2.1 Comparison of the FPGA and CAS programmable logic (© LEAP)

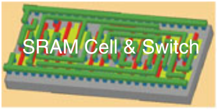
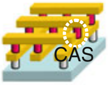
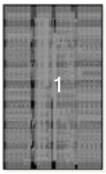
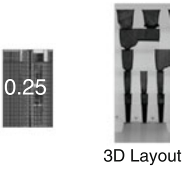
		Conventional FPGA	Programmable Logic Using CAS
Switches	Schematic Figure		
	Area Resistance Capacitance Volatility	1 1 1 Volatile	0.05 0.1 0.1 Non-Volatile
Program-mable Logic	Layout & Area		

Table 2.2 Comparison of the ASIC, FPGA, and programmable logic using CAS (© LEAP)

LSI	ASIC	FPGA	CAS
Power efficiency	10 or more	1	10
Chip area	1/10	1	1/4
Design cost for LSI	High (M$)	0	0
Design turnaround	Months	0	0

Compared to these, the CAS approach could provide cost and power advantages as well as the design and manufacture lead-time advantage.

Since the established Cu bridge in the PSE is very thin, the reliability might be a big concern. To achieve high reliability, the choices and the design of the electrode and PSE material, programming- and erasing-method are essential. The results of the reliability evaluation of the CAS are shown in Fig. 2.21 [19–21]. It illustrates that the proposed CAS switch has proven to be able to provide reliable non-volatile switches for post-fabricated logic programming.

The performance comparison was done between the fabricated programmed logic circuits using SRAM and that programmed by CAS. The comparison was done by using 65 nm bulk-CMOS technology and the 6 × 6 programmable logic-cell arrays. The cell shown in Fig. 2.20 is used. The delay-time and the active power were compared for three configured logics, as illustrated in Fig. 2.22. In 4b multiplexer, the delay and active power advantage of over 60 % is obtained in the CAS configuration over the SRAM configuration [22].

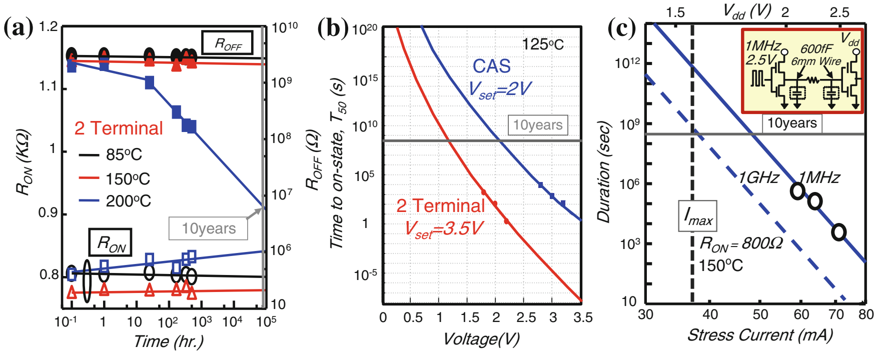


Fig. 2.21 Reliability of the programmed logic using CAS atom-switches [19–21] (© 2012 IEEE). **a** ON/OFF retention of 2-terminal atom-switch at 150 °C, 10 years. **b** OFF retention of CAS and 2-Terminal atom-switch at 125 °C, 10 years. **c** AC ON reliability when *I* = 37 μA, 1 GHz is applied at 150 °C, 10 years

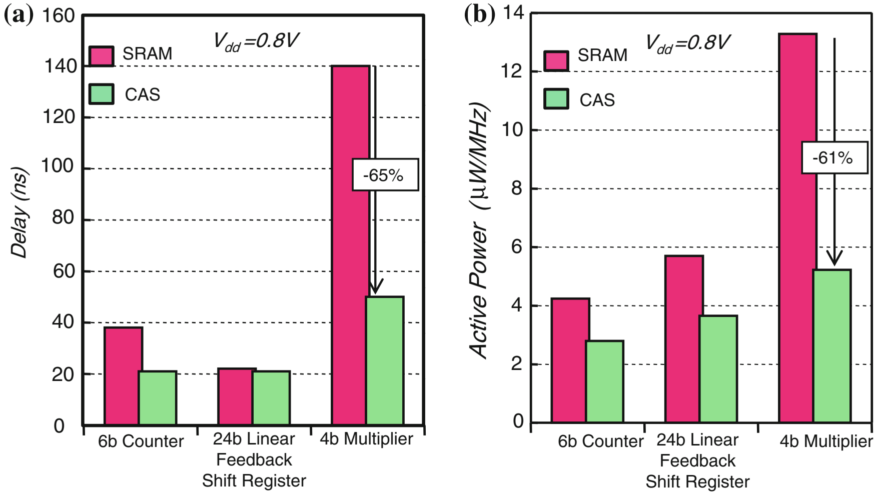


Fig. 2.22 Comparison of the delay and power in conventional FPGA using SRAM and the programmable logic using CAS [22] (© LEAP). **a** Comparison of the delay-time. **b** Comparison of the active power

2.3.3 SOTB-CMOS Microprocessor with Atom-Switch PROM

The atom-switch can be applied to PROM in a microprocessor as a low-voltage firmware memory [23, 24]. A 32 b-RISC chip was developed that utilizes a low-voltage SOTB-CMOS microprocessor and atom-switch PROM as shown in Fig. 2.23a. The microprocessor is the 5-stage pipelined 32 b RISC CPU with 2 blocks of 32 KB SRAM data memory, and 16 KB atom-switch PROM. Since the microprocessor is targeted to operate below 0.5 V, and since it is necessary to apply a relatively high voltage for programming, two series-memory-transistors are used to avoid the breakdown of the PROM-array transistors. The array is also separated by separation transistor during PROM programming. For low voltage read-operation, the PROM circuit, shown in Fig. 2.23b, and the atom-switch PROM-cell, shown in Fig. 2.23c, were designed. At such low voltages, a standard differential sense-amplifier is hard to operate stably. Therefore, a simple

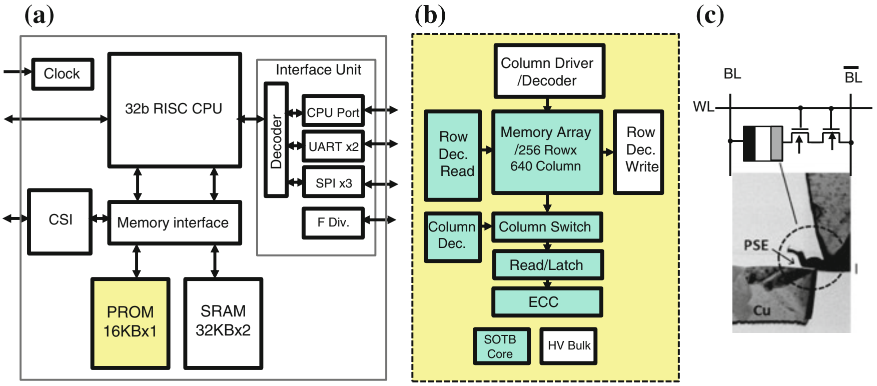


Fig. 2.23 32 b RISC CPU using SOTB-CMOS and PROM using atom-switch cells [23, 24] (© Appl. Phys.). **a** Block diagram of the 32 b RISC. **b** Block-diagram of the PROM. **c** Atom-switch PROM cell

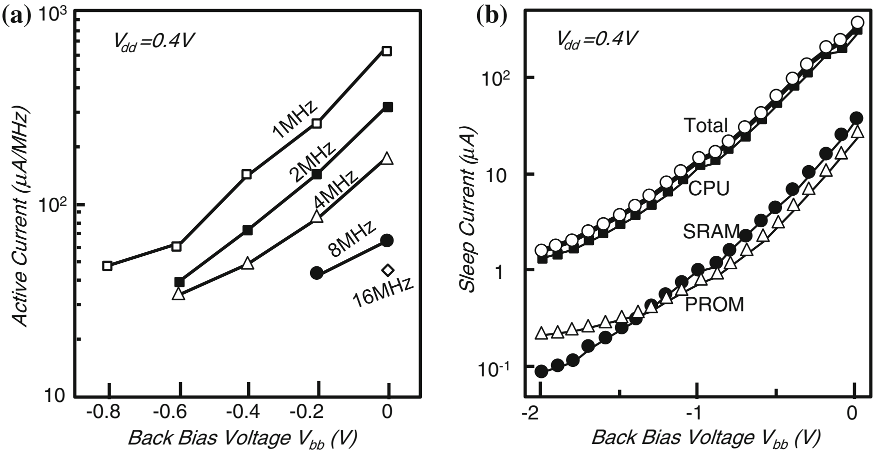


Fig. 2.24 Active current and the sleep-current of the 32 b RISC CPU using SOTB-CMOS and atom-switch PROM [23, 24] (© 2015 IEEE). **a** Active current of 32 b RISC CPU as a function of $V_{BB}(V)$. **b** Sleep-current of 32 b RISC as a function of $V_{BB}(V)$

inverter-type sense-amplifier is employed. The sensing is done by pre-charging the sensing node first, and then, the signal is read by opening the separation-transistor and column-switch. The active current and the sleep current are shown in Fig. 2.24a, b. It is seen that the microprocessor operates with 50 μA/MHz at 0.4 V with less than 2 μA standby current with a back-bias of −2 V. This indicates that the microprocessor can load application-specific firmware after fabrication, and the microprocessor can operate at 0.4 V with very low standby current.

2.3.4 TRAM for Low-Power Storage

In a "big data" era, data-access with high speed and less power has the key importance in the storage systems as shown in Fig. 2.25. In a data center, IT equipment including servers and storage consume 1/3 of the power. The other portion of the power is consumed by the power delivery, UPS and the cooling system. The power of these is correlated to the power of the IT equipment. In a current storage-system hierarchy, SSD has already been used in tier-0 and 1. HDD is dominant in tier 2 and 3. In the future storage systems, SSD would replace the tier 2 and 3, but for tier 0, highly efficient next-generation storage-level memory is needed. This new-generation storage should have a more than 10-times higher data rate and much lower power dissipation as shown in Fig. 2.25c.

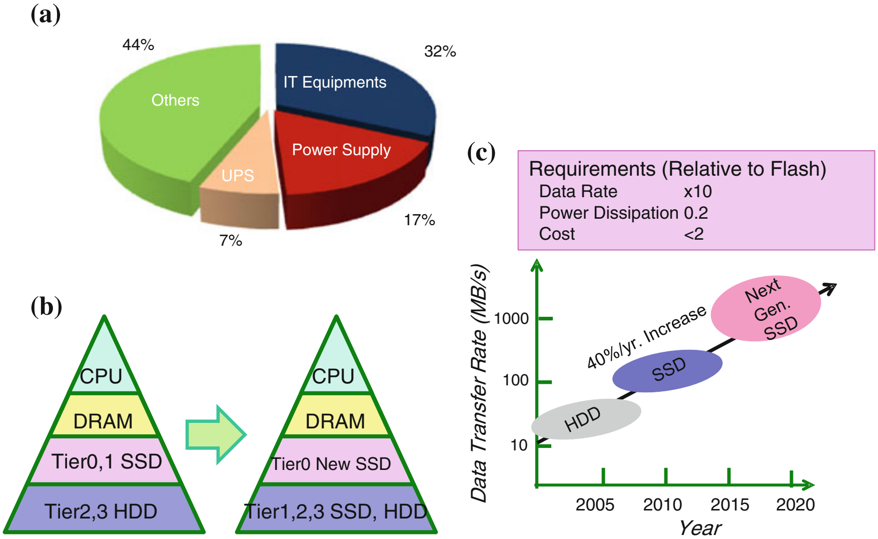
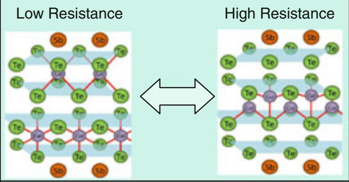
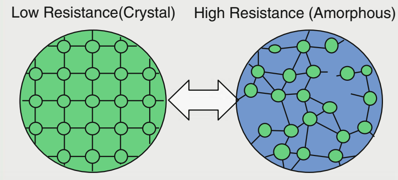


Fig. 2.25 Requirements for the storage-class memory in the era of "big-data" (© LEAP). **a** Power-dissipation in a data center. **b** Shift of hierarchy in storage systems. **c** Requirement for the future storage-class memory

Table 2.3 Comparison of the TRAM (topological switching RAM) and PRAM (© LEAP)

Memory	TRAM	Conventional PRAM
	Topological-switching RAM	Phase Change Memory
Material	GeTe/Sb_2Te_3 Super-lattice	$Ge_2Sb_2Te_5$ Alloy
Memory Mechanism	 Low Resistance ⇔ High Resistance	 Low Resistance(Crystal) ⇔ High Resistance (Amorphous)
State Change	Non-Melting Process by a Short Range Site Change of the Ge Atom	Crystal-Amorphous Phase Change due to Melting Process by Joule Heating

New optical memory, interface phase-change memory using super-lattice material, was first proposed by Tominaga et al. [25]. A new electrical non-volatile memory called TRAM (Topological switching RAM) has been developed by LEAP by modifying the interface phase-change memory [26, 27]. As illustrated in Table 2.3, the TRAM consists of a GeTe/Sb_2Te_3 super-lattice. In the PRAM, set and reset is performed by the phase-change of the $Ge_2Sb_2Te_5$ alloy, in other words, state change between crystal-phase (low resistivity) and amorphous phase (high resistivity) through a melting and re-crystallizing process. In the TRAM, Ge atoms change sites by the electron- and the hole-injection. This state-change is the non-melting process, and requires relatively small energy. Furthermore, the state-change is much faster as compared to the usual melting phase change.

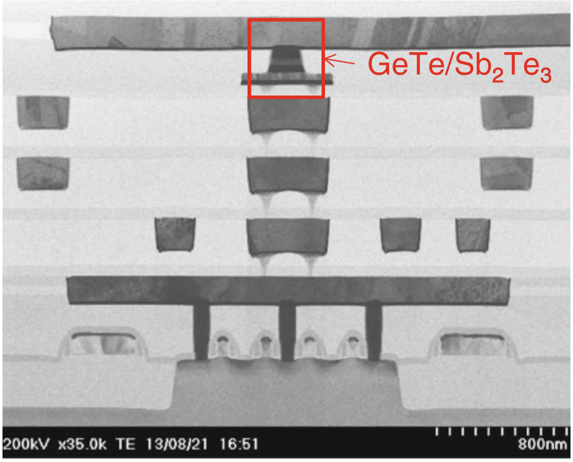


Fig. 2.26 Cross-section of the fabricated 1T-1R GeTe/Sb_2Te_3 Super-lattice TRAM cell [26] (© 2014 IEEE)

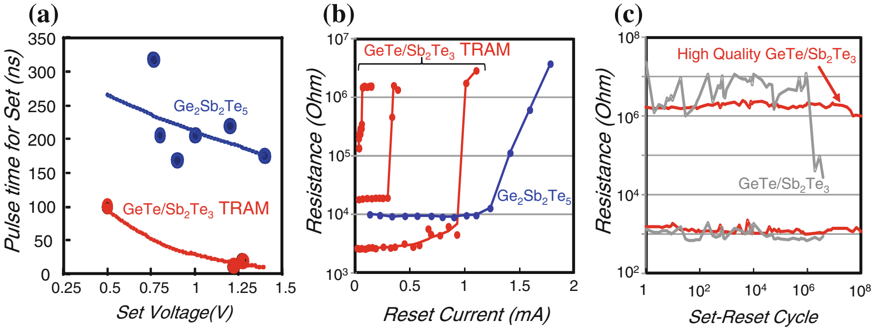


Fig. 2.27 Performance of the $GeTe/Sb_2Te_3$ Super-lattice TRAM [27] (© 2013 IEEE). **a** Pulse time as a function of the set-voltage. **b** Resistance as a function of the reset-current. **c** Set-reset cycle endurance

Figure 2.26 shows the cross-section of the fabricated 1Transistor-1Resistor (1T1R) $GeTe/Sb_2Te_3$ super-lattice TRAM cell. The TRAM cell is formed in the back-end process between metal 4 and 5 by PVD. Some of the performance data of the TRAM cell are compared with those in the PRAM cell in Fig. 2.27. The pulse-time for set is much smaller as compared to that of PRAM as shown in Fig. 2.27a. Resistance-change from low to high (reset) occurs at much lower reset current as shown in Fig. 2.27b. Set- and reset-cycles over 10^8 were observed. It was also demonstrated that the Ge_xTe_{1-x}/Sb_2Te_3 periodic layers, with $x < 0.5$, yield smaller set- and reset-current of 55 μA with less than 1 V set- and reset-voltages [28]. Storage-class memory using the TRAM cell is yet to be developed. These early results suggest that TRAM could be a promising candidate as a storage-class non-volatile memory that coexists with large-capacity flash memories in the storage system.

2.4 3D Integration

3D integration is regarded to be a solution in many applications. These include large-capacity flash memories, DRAM with wide bus interface, and high-performance processors. 3D integration can also realize hetero-integration consisting of power-delivery, analog and sensors, and digital signal-processing. This potential is covered broadly in Chap. 3.

3D flash memories comprising monolithically-integrated 3D memory-cell arrays were developed as shown in Fig. 2.28a [6]. Production of monolithically-integrated 3D flash memory has begun in 2013. This is an effective approach to solve the memory-capacity bottleneck in the flash memories. Since small-pitch interconnects and via with very high aspect-ratio are required in such applications, technology development to apply material other than Cu is now under way in many R&D projects. Metal interconnect such as Cu needs a barrier-metal. Due to the higher

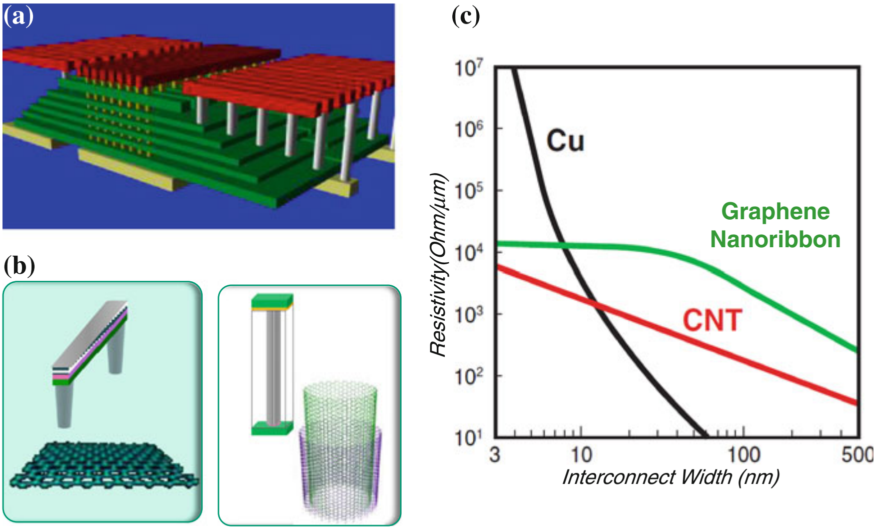


Fig. 2.28 Graphene and carbon-nanotube (CNT) as the candidate material for monolithically-integrated 3D flash memory. **a** Conceptual view of a BICS flash memory [6] (© 2007 IEEE). **b** Horizontal interconnect using graphene (*left*) and carbon-nanotube vertical via (*right*) (© LEAP). **c** Interconnect resistivity challenge in high-density integrated circuits [29] (© 2007 IEEE)

resistivity of the barrier-metal and surface scattering, metal-interconnects with small widths suffer from the increase of the resistivity, whereas graphene nano-ribbon is a promising material for narrow interconnect due to the low resistivity as shown in Fig. 2.28c [29]. Figure 2.28b illustrates a possible technology to solve the interconnect bottleneck, under development in LEAP using nano-carbon interconnect. Figure 2.29 shows the cross-sectional TEM photographs of the prototype

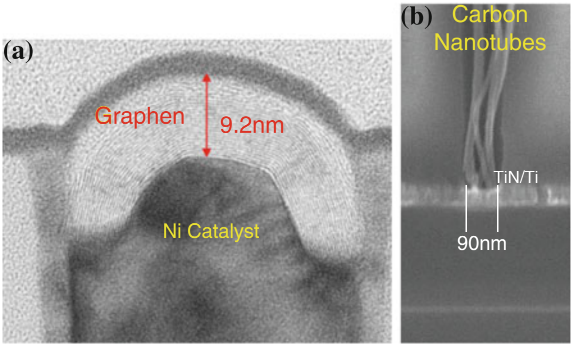
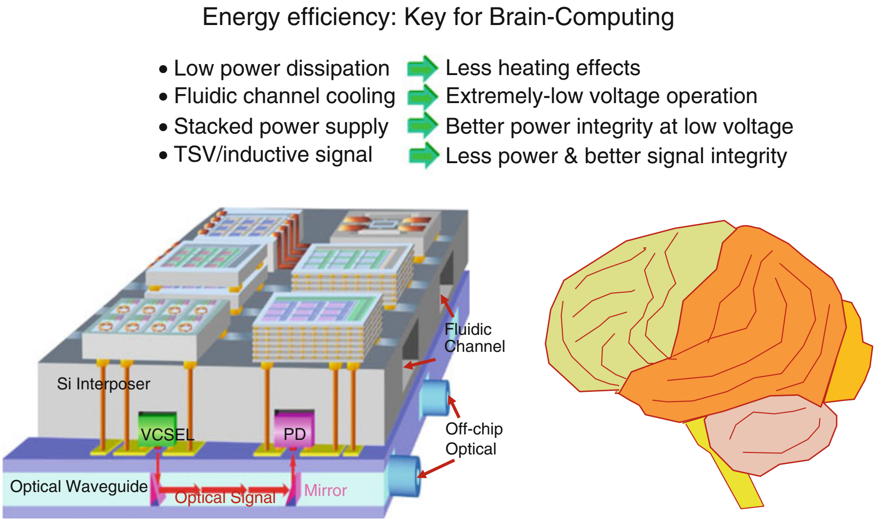


Fig. 2.29 Carbon interconnect technology under development for 3D flash memory [30, 31] (© LEAP). **a** CVD-grown multilayer graphene interconnect on Ni catalyst. **b** Carbon nanotubes grown in a via with aspect-ratio of 19

multi-layer graphene grown by CVD on a Ni catalyst at a temperature below 650 °C, and a carbon-nanotube via for vertical interconnection, both under development in LEAP [30, 31].

In video-data-processing of mobile equipment such as smart-phones or tablets, 3D-stacked DRAM is desirable to realize wide-bus data-transmission requirements between DRAM and processor with small footprint. It has been regarded as a candidate technology to solve several bottlenecks to realizing higher integration. The production of the 3D-stacked DRAM IC was already announced by several manufacturers. To achieve higher performance in high-end processors, 3D-chip stack approaches were also developed. There still remain a number of challenges. In high-end 3D-stacked microprocessors, heat-flux from a chip to the stacked chips causes performance degradation and thermal stress, and that may give rise to reliability problem. Therefore, effective dissipation of power of the 3D-stacked IC's is required. Another challenge is the integration of the chip-design and the chip-supply chain. If a 3D-stacked chip is designed and assembled with chips from different manufacturers, they need to be designed using the same design-rule for 3D-integration. It is also desirable that each chip is the tested known-good-die (KGD). The quality of the chips from different chip-manufacturers also needs to be assured. These challenges need to be solved in the supply-chain of the 3D-stacked chips.

An example of one of the promising chip-stack technologies is the chip-bonding to wafer by a self-assembly technique developed by Prof. M. Koyanagi's group, Tohoku University, as shown in Fig. 2.30. They demonstrated a 38-chip-stack by a



Energy efficiency: Key for Brain-Computing

- Low power dissipation ➡ Less heating effects
- Fluidic channel cooling ➡ Extremely-low voltage operation
- Stacked power supply ➡ Better power integrity at low voltage
- TSV/inductive signal ➡ Less power & better signal integrity

Fig. 2.30 Conceptual 3D stacked-chip integration drawn by Prof. M. Koyanagi. Combination of low-voltage, and energy-efficient chips and 3D stacked-chips could open ways for brain-computing [32] (Courtesy Koyanagi, © 2005 IEEE)

self-alignment technique [32]. Advantages of applying the 3D-integration to low-voltage stacked-IC's is illustrated in Fig. 2.30. If extremely low-power IC's are stacked with wireless power delivery from the stacked power-supply chips, metallic interconnections and wires for power-delivery and bulk heat-dissipation could be eliminated. It also makes possible an ideal low-temperature cooling through micro-fluidic channels to obtain a very steep sub-threshold slope. In the first demonstration of bulk-CMOS operation at 77 and 4.2 K, it was shown that the steep sub-threshold slope at low temperature makes possible low-V_{dd} operation [33]. If the V_{th} variation is minimized, this approach might realize extremely low-power operation. An ideal low-voltage and low-power processing might be possible in the future computing by using such technologies. 3D chip-stack technology is not only effective to be applied to traditional Von-Neumann architecture processors, but also to other types of emerging architectures, such as brain computing (Chap. 18) that require a lot of interconnections. Such an approach would open ways to a future computation paradigm.

2.5 The Future of Low-Power Integrated Circuits

Disruptive low-power IC technologies are required in the era of big data, IoT, and cloud-computing. In this chapter, some of the candidate technologies under development in LEAP were described. These include hybrid SOTB-CMOS, MRAM, CAS switch, TRAM, and nanocarbon-interconnect technology. Due to the explosion of data and transactions via internet, future data-centers require solutions utilizing energy-efficient processors and storages using low-voltage CMOS and non-volatile memories. Mobile terminals and robots also require performance/power efficiency. The future society depends on networked wireless IoT-sensors. They require wireless access and maintainable power-delivery means. These future devices, equipment, and systems are integrated by volume-efficient technologies such as monolithically-integrated 3D (Chap. 3) or 3D-stacked chips. Low-power post-fabricated programmable chips also provide cost-effective solutions for various kinds of applications.

Low-power electronics and IT technologies are also essential to establish a safe and sustainable human society. Past civilizations continuously over-consumed natural resources, and the current civilization is even accelerating the consumption. We expect that the world population would reach 8–10 billion sometime in the near future. In a closed system "earth", a catastrophe may occur to us if we continue our life-style of destroying the nature of earth. Although there have been continuing discussions on the direction of the climate-change, common consensus is that reducing energy-consumption, i.e. the exhaust of CO_2, and preserving forest and oceanic resources are the key issues for the future sustainability. Sensor networks and the advanced climate simulations may help monitoring and predicting the climate change.

The infrastructure of the modern society is formed by steel and steel-reinforced concrete with the lifetime of less than 200 years, much shorter than the Roman concrete used in Pantheon, Rome, built 1900 years ago. Life-cycle management and maintenance of the urban infrastructure have become key safety-issues. Sensor-networks may help monitoring and warning the deterioration and destruction. In such applications, energy-embedded or energy-harvesting (Chap. 19) sensors with wireless interface are the key components in the system.

Establishing a more resilient society against unusual catastrophes such as earthquakes, super-storms, floods, tsunamis, landslides, is also an urgent issue. The 2004 Indian Ocean Tsunami after the *M9.1* Richter-scale earthquake occurred near the west coast of Sumatra, and it claimed 230,000 lives in South-east Asia. In Japan, we experienced the East-Japan earthquake with a magnitude of *M9.0*, on Mar. 11, 2011, and the subsequent tsunami having a height of 14–20 m. This caused the loss of over 18,000 lives and the meltdown of the nuclear reactors in Fukushima. At that time, all the social systems, municipal, transportation, medical, energy and food supply, information, and communication did not function due to the loss of energy. In Philippine, over 6200 lives were lost by the super-typhoon Yolanda in 2013, and in the U.S., over 1800 lives were lost by Hurricane Katrina in 2005.

We have not yet experienced a super-volcanic eruption in the historical era. But geological evidences indicate that our prehistoric ancestors experienced catastrophes in Toba, Sumatra-Indonesia, 70,000 years ago, and in Kikai Island, south of Kyushu-Japan, 7300 years ago. If such super-volcanic activities occur in the future, only a society capable of forecasting the future and making sustainable and quick decisions, can function. Detection and handling of asteroids is also a big concern to the continuation of the human society. A worldwide safety-network is definitely required. Networked sensors, IT's, communication means, and electronics, utilizing low-power

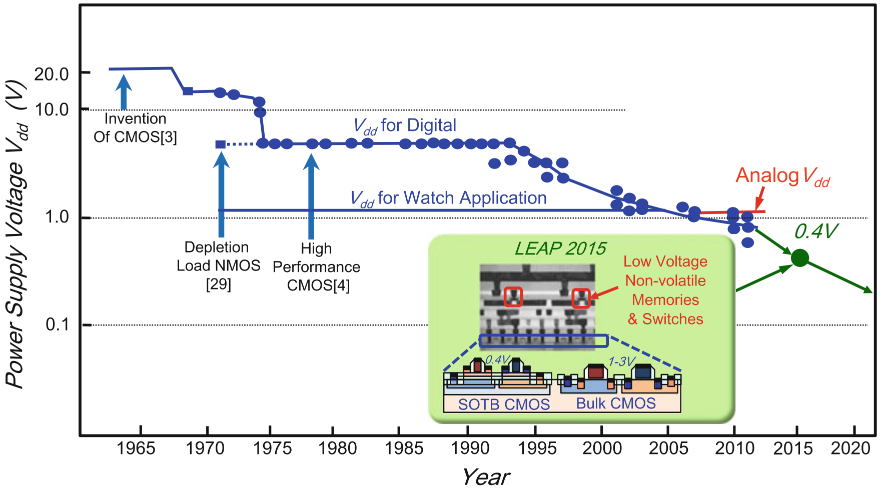


Fig. 2.31 Trend of the power-supply voltage of MOS-IC's in the past 50 years appeared in major circuit conferences (© LEAP)

integrated circuits, are essential elements to gather and analyze data, to perform simulations, to help leaders and to provide information to people in such a catastrophe.

Figure 2.31 illustrates the power-supply voltage in the past 50 years. PMOS-IC's, the major technology during the 1960s, and the first CMOS by Wanlass and Sah in 1963, were operated above 20 V. Then +5 V single-supply, depletion-load NMOS was developed [34], and +5 V became the major supply-voltage due to TTL compatibility and lasted for approximately 20 years. An exception was the low-voltage CMOS for watches and calculators that use 1.0–1.5 V battery. Gradual reduction of the supply-voltage occurred after the late 1990s. In the 2010s, near-threshold or sub-threshold operation has become popular particularly for low-power IC's. The technology developed in LEAP contributed to reduce the power-supply voltage to less-than-0.5 V in digital IC's. In the year 2020 and beyond, integrated circuits may require even lower supply-voltages of less than 100 mV. Transistors with much steeper sub-threshold slope and smaller variation are the key elements, and several candidate technologies, such as tunnel MOSFET (Chap. 1) or nanowire FETs, are under development. It should be noted that the new steep sub-threshold transistors must fulfill the requirement of small variation, controllability of V_{th}, enough I_{ON} (on-current) for digital circuits, good analog device parameters such as f_T, f_{max}, *Noise*, V_{offset}, distortion, and power-handling capability. A possible solution is the hybrid use of some new transistors and bulk-CMOS or SOTB-CMOS as the platform technology.

Most of the figures described in this chapter were provided by, "*Ultra-Low Voltage Device Project for Low Carbon Society*" funded and supported by the Ministry of Economy, Trade and Industry (METI) and the New Energy and Industrial Technology Development Organization (NEDO). The back-end processing was performed by the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan. The author would like to express deep gratitude to Mr. N. Sumihiro, project leader, Dr. S. Kimura, vice project leader, and group leaders Drs. T. Sugii, N. Takaura, H. Hada, T. Sakai and N. Sugii, and all the members of LEAP.

References

1. JEITA Green IT Promotion Council.: Report 2008–2012, Feb 2013
2. Koomey, J.G., Berard, S., Sanchez, M., Wong, H.: Assessing trends in the electrical efficiency of computation over time. In: IEEE Annals of the History of Computing, 17 Aug 2009
3. Wanlass, F.M., Sah, C.T.: Nanowatt logic using field-effect metal-oxide-semiconductor triodes. In: International Solid-State Circuits Conference Digest Technical Papers, pp. 58–59, Feb 1963
4. Masuhara, T., Minato, O., Sasaki, Y., Sakai, Y., Kubo, M., Yasui, T.: A high-speed low-power hi-CMOS 4 K static RAM. In: International Solid-State Circuits Conference Digest Technical Papers, pp. 110–111, Feb 1978
5. Chen, T.C.: Where CMOS is going. IEEE SSCS Newslett. **20**(3), 5–8 (2006)
6. Tanaka, H., Kido, M., Yahashi, K., Oomura, M., Katsumata, R., Kito, M., Fukuzumi, Y., Sato, M., Nagata, Y., Matsuoka, Y., Iwata, Y., Aochi, H., Nitayama, A.: Bit cost scalable

technology with punch and plug process for ultra high density flash memory. In: 2007 VLSI Technical Symposium, Digest Technical Papers, pp. 14–15, June 2007

7. Tsuchiya, R., Horiuchi, M., Kimura, S., Yamaoka, M., Kawahara, T., Maegawa, S., Ipposhi, T., Ohji, Y., Matsuoka, H.: Silicon on thin BOX: a new paradigm of the CMOSFET for low-power high-performance application featuring wide-range back-bias control. In: IEDM Technical Digest, pp. 631–634, Dec 2004
8. Ishibashi, K., Sugii, N., Usami, K., Amano, H., Kobayashi, K., Kha, P.C., Makiyama, H., Yamamoto, Y., Oda, H., Hasegawa, T., Okanishi, S., Yanagita, H., Kamohara, S., Kadoshima, M., Maekawa, K., Yamashita, T., Hung, L.D., Yomogita, T., Kudo, M., Kitamori, K., Kondo, S., Manza, Y.: A perpetuum mobile 32 bit CPU with 13.4 pJ/cycle, 0.14 μA sleep current using reverse body bias assisted 65 nm SOTB CMOS technology. In: Proceedings 2014 IEEE Cool Chips XVII, Apr 2014
9. NEDO Final Evaluation Report, on the Research project.: fundamentals of next generation semiconductor material and process, MIRAI Project-3rd Phase (Japanese). Oct 2011
10. Hiramoto, T., Mizutani, T., Kumar, A., Nishida, A., Tsunomura, T., Inaba, S., Takeuchi, K., Kamohara, S., Mogami, T.: Suppression of DIBL and current onset voltage variability in intrinsic channel FD SOI MOSFETs. In: 2010 International SOI Conference
11. Yamamoto, Y., Makiyama, H., Tsunomura, T., Iwamatsu, T., Oda, H., Sugii, N., Yamaguchi, Y., Mizutani, T., Hiramoto, T.: Poly/high-k/SiON gate and novel profile engineering for low power silicon on thin BOX (SOTB) CMOS operation. In: 2012 VLSI Technical Symposium, Digest Technical Papers, pp. 109–110, June 2012
12. Yamamoto, Y., Makiyama, H., Shinohara, H., Iwamatsu, T., Oda, H., Kamohara, S., Sugii, N., Yamaguchi, Y., Mizutani, T., Hiramoto, T.: Ultralow-voltage operation down to 0.37 V of silicon-on-thin-box (SOTB) 2 Mbit SRAM utilizing adaptive body bias. In: 2013 VLSI Technical Symposium, Digest Technical Papers, pp. T212–T213, June 2013
13. Soo, H.L., Amano, Y.: Evaluation of low power reconfigurable accelerator chip using SOTB transistor (Japanese). In: Technical Digest IEICE, 113, No. 325, RECONF2013-52, Nov 2013
14. Horowitz, M.: Computing's energy problem: (and what we can do about it). In: International Solid-State Circuits Conference Digest Technical Papers, pp. 10–14, Feb 2014
15. Iba, Y., Yoshida, C., Hatada, A., Nakabayashi, M., Takahashi, A., Yamazaki, Y., Noshiro, H., Tsunoda, K., Takenaga, T., Aoki, M., Sugii, T.: Top-pinned perpendicular structure with a counter bias magnetic field layer for suppressing a stray-field in highly scalable STT-MRAM. In: Digest Technical Papers, VLSI Technical Symposium, pp. 11–13, June 2013
16. Yoshida, C., Ochiai, T., Iba, Y., Yamazaki, Y., Tsunoda, K., Takahashi, A., Sugii, T.: Demonstration of non-volatile working memory through interface engineering in STT-MRAM. In: Digest Technical Papers, VLSI Technical Symposium, pp. 12–14, June 2012
17. Iba, Y., Takahashi, A., Hatada, A., Nakabayashi, M., Yoshida, C., Yamazaki, Y., Tsunoda, K., Sugii, T.: A highly scalable STT-MRAM fabricated by a novel technique for shrinking a magnetic tunnel junction with reducing processing damage. In: 2014 VLSI Technical Symposium, pp. 58–59, June 2014
18. Miyamura, M., Tada, M., Sakamoto, T., Banno, N., Okamoto, K., Iguchi, N., Hada, H.: First demonstration of logic mapping on nonvolatile programmable cell using complementary atom switch. In: IEDM Technical Digest, pp. 10.6.1–10.6.4, Dec 2012
19. Tada, M., Sakamoto, T., Banno, N., Okamoto, K., Miyamura, M., Iguchi, N., Nohisa, T., Hada, H.: Highly reliable, complementary atom switch (CAS) with low programming voltage embedded in Cu BEOL for nonvolatile programmable logic. In: IEDM Technical Digest, pp. 30.2.1–30.2.4, Dec 2011
20. Banno, N., Tada, M., Sakamoto, T., Miyamura, M., Okamoto, K., Iguchi, N., Nohisa, T., Hada, H.: A fast and low-voltage Cu complementary-atom-switch 1 Mb array with high-temperature retention. In: 2014 VLSI Technical Symposium, pp. 202–203, June 2014
21. Tada, M., Sakamoto, T., Banno, N., Okamoto, K., Miyamura, M., Iguchi, N., Hada, H.: Improved reliability and switching performance of atom switch by using ternary Cu-alloy and RuTa electrodes. In: IEDM Technical Digest, pp. 29.8.1–29.8.4, Dec 2012

22. Hada, H.: In: Technical Digest, 3rd LEAP FORUM on Ultra Low Voltage Device Project for Low Carbon Society, Tokyo, Japan, Jan 23, 2014
23. Sakamoto, T., Tada, M., Tsuji, Y., Makiyama, H., Hasegawa, T., Yamamoto, Y., Okanishi, S., Banno, N., Miyamura, M., Okamoto, K., Iguchi, N., Ogasawara, Y., Oda, H., Kamohara, S., Yamagata, Y., Sugii, N., Hada, H.: Low-power embedded read-only memory using atom switch and silicon-on-thin-buried-oxide transistor. Appl. Phys. Express **8**, 045201 (2015). doi:10.7567/APEX.8.045201
24. Sakamoto, T., Tsuji, Y., Tada, M., Makiyama, H., Hasegawa, T., Yamamoto, Y., Okanishi, S., Maekawa, K., Banno, N., Miyamura, M., Okamoto, K.. Iguchi, N., Ogasahara, Y., Oda, H., Kamohara, S., Yamagata, Y., Sugii, N., Hada, H.: 0.39-V 18.26-μW/MHz SOTB CMOS microcontroller with embedded atom switch ROM. In: Proceedings of 2015 IEEE Cool Chips XVIII, Apr 2015
25. Simpson, R.E., Fons, P., Kolobov, A.V., Fukaya, T., Krbal, M., Yagi, Y., Tominaga, J.: Interfacial phase change memory. Nat. Nanotechnol. **6**, 501–505 (2011). doi:10.1038/nnano.2011.96
26. Tai, M., Ohyanagi, T., Kinoshita, M., Morikawa, T., Akita, K., Kato, S., Shirakawa, H., Araidai, M., Shiraishi K., Takaura, N.: 1T-1R pillar-type topological-switching random access memory (TRAM) and data retention of GeTe/Sb_2Te_3 super-lattice films. In: Digest Technical Papers, VLSI Technical Symposium, pp. 9–12, June 2014
27. Ohyanagi, T., Takaura, N., Tai, M., Kitamura, M., Kinoshita, M., Akita, K., Morikawa, T., Kato, S., Araidai, M., Kamiya, K., Yamamoto, T., Shiraishi, K.: Charge-injection phase change memory with high-quality GeTe/Sb_2Te_3 superlattice featuring 70-μA RESET, 10-ns SET and 100 M endurance cycles operations. In: IEDM Technical Digest, pp. 30.5.1–30.5.4, Dec 2013
28. Takaura, N., Ohyanagi, T., Tai, M., Kinoshita, M., Akita, K., Morikawa, T., Shirakawa, H., Araidai, M., Shiraishi, K., Saito, Y., Tominaga, J.: 55-μA Ge_xTe_{1-x}/Sb_2Te_3 superlattice topological-switching random-access memory (TRAM) and study of atomic arrangement in GeTe/Sb_2Te_3 structures. In: IEDM Technical Digest, Dec 2014
29. Naeemi, A., Meindl, J.D.: Conductance modeling for graphene nanoribbon (GNR) interconnects. IEEE Electron Device Lett. **28**, 428–431 (2007)
30. Sakai, T.: Technical Digest, 3rd LEAP FORUM on Ultra Low Voltage Device Project for Low Carbon Society, Tokyo, Japan, 23 Jan 2014
31. Sakai, T.: In: Technical Digest, 4th LEAP FORUM on Ultra Low Voltage Device Project for Low Carbon Society, Tokyo, Japan, 6 Mar 2015
32. Fukushima, T., Yamada, Y., Kikuchi, H., Koyanagi, M.: New 3D-integration technology using self-assembly technique. In: IEDM Technical Digest, pp. 359–363, Dec 2005
33. Hanamura, S., Aoki, M., Masuhara, T., Minato, O., Sasaki, Y., Sakai, Y., Hayashida, T.: Operation of bulk CMOS devices at very low temperatures. IEEE J. Solid-State Circuits, **21**(3), 484–490 1986
34. Masuhara, T., Nagata, M., Hashimoto, N., A high performance N-channel MOSLSI using depletion-type load elements. In: International Solid-State Circuits Conference Digest Technical Papers, pp. 12–13, Feb 1971

Chapter 3
Monolithic 3D Integration

Zvi Or-Bach

Abstract As the down-sizing of transistors has arrived at fundamental and practical limits, the technology direction with the largest potential for progress is the integration of transistors in the 3rd dimension on top of each other, maintaining and using the quality of monolithic, crystalline silicon in all successive transistor layers. After decades of exploratory research, monolithic 3D integration is now ready for cost-effective, large-scale implementation of nanoelectronic systems. It offers the largest gains in transistors-per-chip, it solves the on-chip interconnect and communication gridlock and thus the energy, speed and bandwidth problems. It opens a new era of effective industry networks for the sustained growth of the nanoelectronics economy. Monolithic 3D is already being adapted for mass production, in the non-volatile memory segment-3D NAND, and it can be expected that the other segments of the semiconductor industry will follow.

3.1 Why Monolithic 3D

The growth of Integrated Circuits has been driven primarily by the increase of device integration. This technological progress is based on Moore's Law, which is predicated on the notion that the optimum device integration would double the device count every two years. Moore's original prediction accounted for three mechanisms of improvement—decreasing the device (transistor) size, increasing the die size, and improvement of the circuit architecture. Yet, the primary mechanism used over the last 5 decades has been dimensional reduction. Every 2 years, a

Z. Or-Bach (✉)
MonolithIC 3D™ Inc., 3555 Woodford Dr., San José, CA 95124, USA
e-mail: Zvi@MonolithIC3D.com

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_3

new technology node has been developed. Each new node is about 0.7× of the prior node for most critical device dimensions. This dimensional scaling is also known as Dennard scaling.

Dimensional scaling over the prior decades had the added benefits of reduction of cost per function, reduction in power, and increase in speed of device operation. Unfortunately dimensional scaling has reached the point of diminishing returns due to the escalating costs of implementation, as illustrated by Fig. 3.1.

These increasing challenges, which are directly related to dimensional scaling, are due mainly to:

1. Lithography
2. On-chip interconnect
3. Transistor variation.

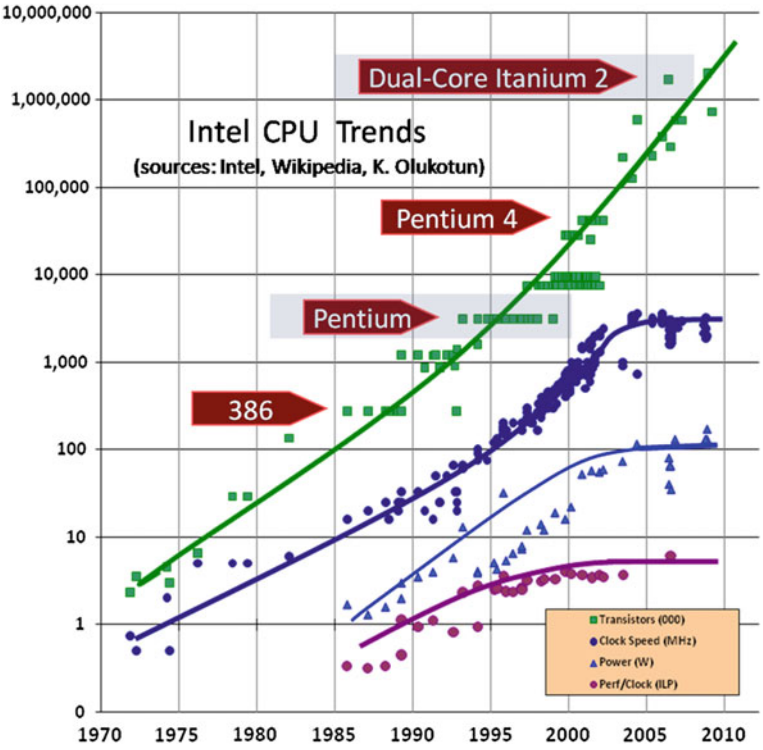


Fig. 3.1 Dimensional scaling is reaching the diminishing-return phase

3.1.1 Lithography

Dimensional scaling has been implemented by a 0.7× reduction of device critical dimensions. Accordingly, the lithography process needs to project smaller features every node, as illustrated in Fig. 3.2.

The industry kept moving to shorter wave lengths utilized in the lithography tool, but reached a technology limit at 193 nm using excimer lasers. The development of an Extreme Ultra Violet (EUV) lithography tool is an on-going major challenge and is not ready for production. Meanwhile, 193 nm immersion lithography is being used in a double and quad processing manner to mitigate the dimensional printing challenge, but at an escalating cost impact on the end device. This is illustrated in Fig. 3.3 as presented at the IEEE IITC 2014 workshop.

3.1.2 On-Chip Interconnect

Dimensional scaling improves transistor switching speed, but it increases the interconnect resistance, wire-to-wire capacitance, and overall interconnect RC. For over a decade, on-chip interconnects have been dominating device performance. First the industry changed interconnections from aluminum to copper, and then the inter-metal dielectric has been changed to low-K materials. Recently, the use of air-bridges was reported in reference to Intel's 14 nm logic process.

At IEDM 2013, Geoffrey Yeap, Qualcomm VP of Technology, stated in his invited talk: "As performance mismatch between transistors and interconnects continue to increase, designs have become interconnect-limited. Monolithic 3D (M3D) is an emerging integration technology poised to reduce the gap significantly between transistors and interconnect delays to extend the semiconductor roadmap way beyond the 2D scaling trajectory predicted by Moore's Law." Yeap provided the following chart—Fig. 3.4—to show the growing gap between transistor delay and interconnect delay.

3.1.3 Transistor Variation

Dimensional scaling has reached the point where some of the critical device dimensions are as small as only a few atomic layers. These and multiple other issues have caused a severe increase of across-the-die transistor variation, thus limiting the industry's ability to reduce the 6-transistor SRAM bit-cell size, as illustrated in the following table (source: imec): Fig. 3.5.

Clearly, below the 28 nm node, SRAM bit-cell scaling is slowing and falls far short of the 2X-per-node needed to maintain Moore's Law.

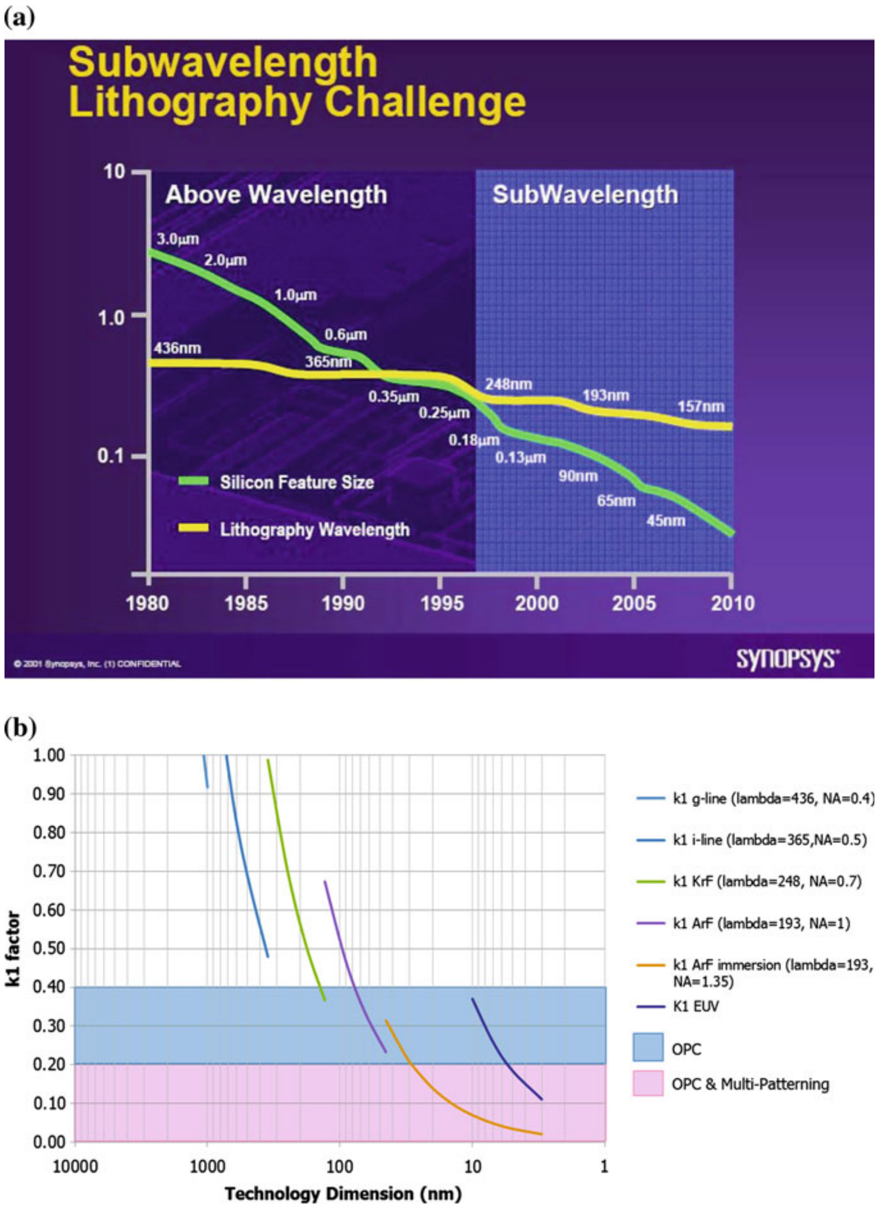


Fig. 3.2 **a** The lithography challenge. **b** Lithographic k1 correction factor by technology node

Monolithic 3D is well positioned to serve as an alternate path for industry's desire to increase device integration. It achieves increases in device integration by effectively having a smaller 2D die size, and, by "folding" the die, the on-chip interconnects are kept relatively short, thus enabling the increase of integration

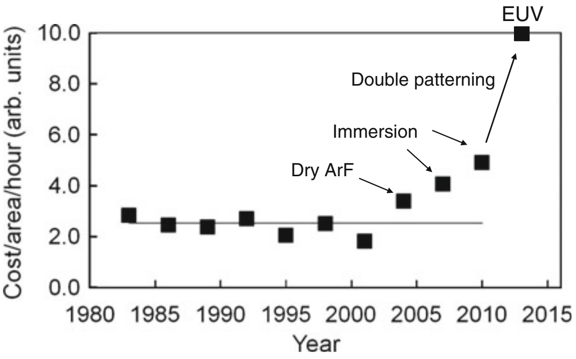


Fig. 3.3 Cost of lithography per wafer area per hour with dimension scaling

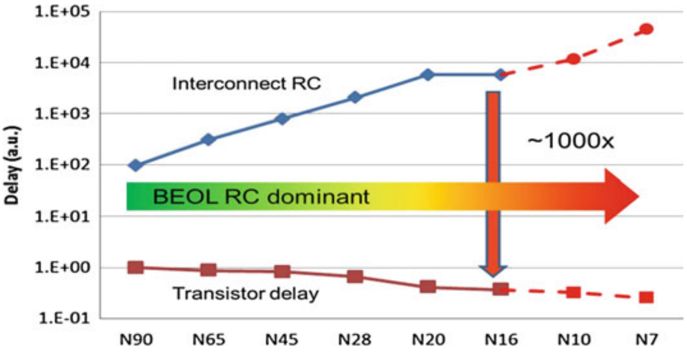
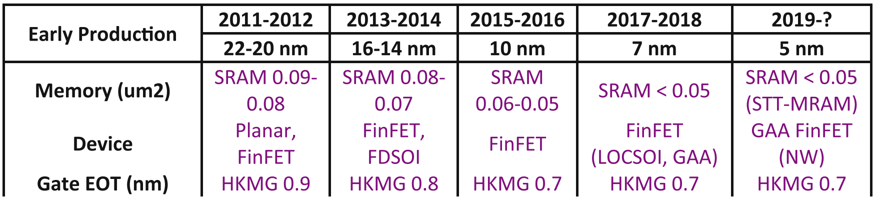


Fig. 3.4 On-chip interconnect delay scaling versus transistor delay



Early Production	2011-2012	2013-2014	2015-2016	2017-2018	2019-?
	22-20 nm	16-14 nm	10 nm	7 nm	5 nm
Memory (um2)	SRAM 0.09-0.08	SRAM 0.08-0.07	SRAM 0.06-0.05	SRAM < 0.05	SRAM < 0.05 (STT-MRAM)
Device	Planar, FinFET	FinFET, FDSOI	FinFET	FinFET (LOCSOI, GAA)	GAA FinFET (NW)
Gate EOT (nm)	HKMG 0.9	HKMG 0.8	HKMG 0.7	HKMG 0.7	HKMG 0.7

Fig. 3.5 SRAM bit-cell scaling

without the escalating costs and interconnect deficiencies associated with dimensional scaling.

3.2 Historical Review of Monolithic 3D Technologies

For many years, monolithic 3D was considered impractical due to the 400 °C temperature limit imposed by the aluminum or copper interconnect. This limitation led to the focus on Through-Silicon-Via (TSV) technology as the only viable path for 3D ICs. It now seems clear that the TSV process flow is intrinsically expensive and, accordingly, is being perpetually pushed to the future. TSVs allow stacking of fully processed devices using wafers that are thinned to about 50 μm.

In monolithic 3D, the upper transistor layers are orders of magnitude thinner, less than about 100 nm. Accordingly, the vertical connectivity is comparable to the horizontal connectivity and is many orders of magnitude better than for TSV technology. Some of the very early work was done in 1989 [1, 2] using selective epitaxial seeding from the bulk to build transistors over transistors without vertical interconnection in-between. In recent years, pioneering efforts were published providing practical paths for monolithic 3D logic devices [1, 3–8]. In the following, we present a brief overview of historical and current works on monolithic 3D technologies.

3.2.1 Thin-Film Polysilicon-Based Monolithic 3D

The simplest approach for monolithic 3D technologies is to use Thin-Film-Transistors—TFT. Most common TFTs use polysilicon devices that could be directly deposited over an existing semiconductor wafer without exceeding the 400 °C temperature limit. TFT performance is inferior to mono-crystalline transistors but could be useful for some memory applications. An early attempt to build a 3D-FPGA using TFTs for the FPGA program memory was made by a start-up named Tier Logic in collaboration with Toshiba [9]. 3D-NAND made with poly TFTs, currently beginning volume production, is considered by most as the future path for non-volatile memories and will be detailed later in the chapter.

3.2.2 Crystalline Overlay

Forming a crystalline overlay can be done by crystallizing a prior deposited poly or amorphous layer. Most common crystallization techniques utilize laser-based melting and various re-crystallization techniques. Some use seeding from the underlying single crystal base to direct the crystal formation [10], others suggest the use of a μ-Czochralski process which is based on pulsed-laser crystallization of a-Si [11]. And some let the layer crystallize as it may [6, 7]. While these techniques may form small regions of crystalized silicon, it seems that layer transfer techniques would be preferable for most 3D logic device applications due to the perfect uniformity of the transferred mono-crystal.

3.2.3 Layer Transfer

Layer transfer techniques became a commonly practiced semiconductor process mostly through the construction of Silicon-on-Insulator (SOI) wafers. The most common layer transfer technique is the ion-cut, also known as Smart-Cut®, invented by CEA Leti and used by Soitec over the last two decades as illustrated in Fig. 3.6.

An alternative layer transfer technique was invented at Canon named ELTRAN —Epitaxial Layer TRANsfer [12]. The ELTRAN layer transfer technique is done at below 400 °C. Variations on the ion-cut techniques were developed by Soitec and SiGen to allow the layer transfer at temperatures below 400 °C by the use of co-implant or mechanical force [13]. Another variation was invented by IBM and is used in the MIT Lincoln Lab [14] where an SOI wafer is processed with transistors, then flipped and bonded to a base wafer with transistors, and then the bulk of the SOI wafer is etched away using the oxide layer as an etch stop. This technique provides a solution that falls in-between TSV and monolithic 3D in terms of the density of vertical (inter-layer) interconnect.

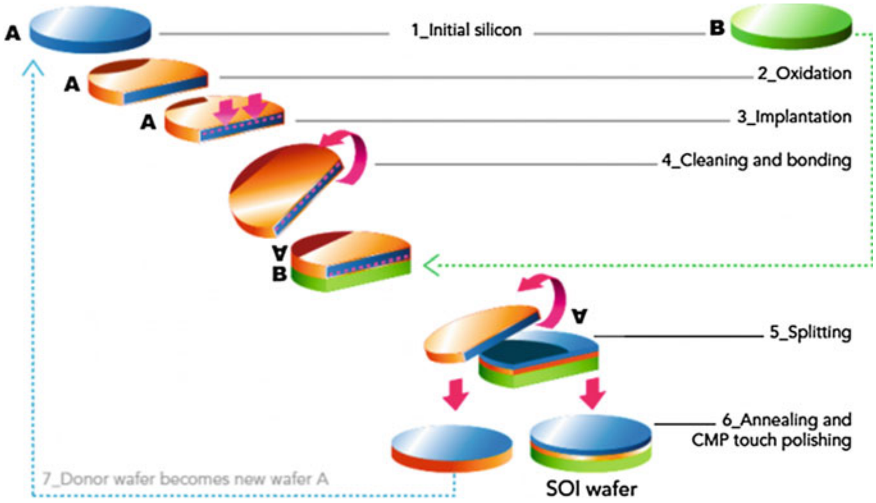


Fig. 3.6 Layer transfer using the smart-cut process

3.2.4 Transistor Activation

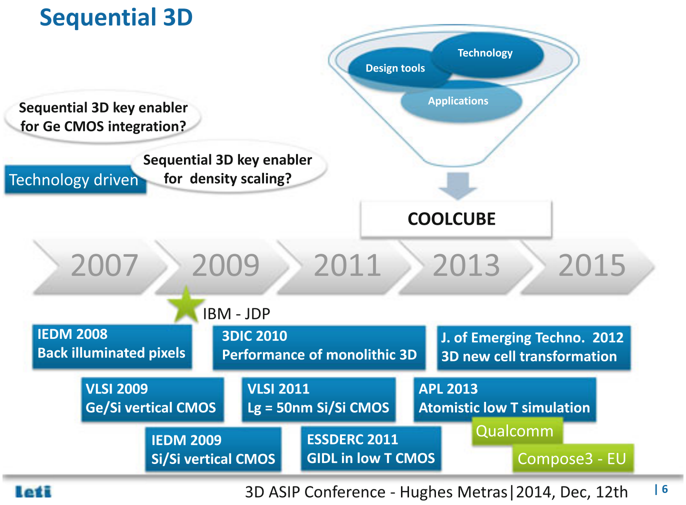
A key step in forming transistors in a mono-crystallized layer is doping activation. Doping activation requires more than 600 °C, typically 800 °C, which needs to be managed in order not to damage the underlying interconnection layers such as copper.

CEA Leti has been one of the more active organizations working on monolithic 3D technology, which they sometimes call Sequential 3D, for logic devices [15]. In its early work CEA Leti developed technologies, which did not use interconnection between the base layer and the upper transistor layer. Recently, CEA Leti has done work with the collaboration and support of IBM, ST Micro and Qualcomm, as illustrated in Fig. 3.7. In some of their recent work, refractory metals were used, such as tungsten, which can withstand high processing temperatures. Some of the upper-layer transistor processes were adapted to be performed at below 600 °C, and the lower transistor structures were modified to survive up to 600 °C. Lately, excimer laser annealing was integrated into the process flow to allow more flexibility for monolithic 3D integrations [16].

Another company, which has been active in the monolithic 3D space, is MonolithIC 3D Inc. The company published several process flows enabling upper-layer transistor activation without damaging the underlying interconnection layers.

3.2.4.1 The RCAT Process

Figure 3.8 describes the "RCAT process" [18], which constructs the RCAT transistors which have been commonly used in DRAM manufacturing since the 90 nm



Sequential 3D

Sequential 3D key enabler for Ge CMOS integration?

Technology driven

Sequential 3D key enabler for density scaling?

Design tools

Technology

Applications

COOLCUBE

2007 2009 2011 2013 2015

IBM - JDP

IEDM 2008
Back illuminated pixels

3DIC 2010
Performance of monolithic 3D

J. of Emerging Techno. 2012
3D new cell transformation

VLSI 2009
Ge/Si vertical CMOS

VLSI 2011
Lg = 50nm Si/Si CMOS

APL 2013
Atomistic low T simulation

IEDM 2009
Si/Si vertical CMOS

ESSDERC 2011
GIDL in low T CMOS

Qualcomm

Compose3 - EU

Leti 3D ASIP Conference - Hughes Metras|2014, Dec, 12th | 6

Fig. 3.7 CEA Leti collaborative road map for monolithic 3D technology [17]

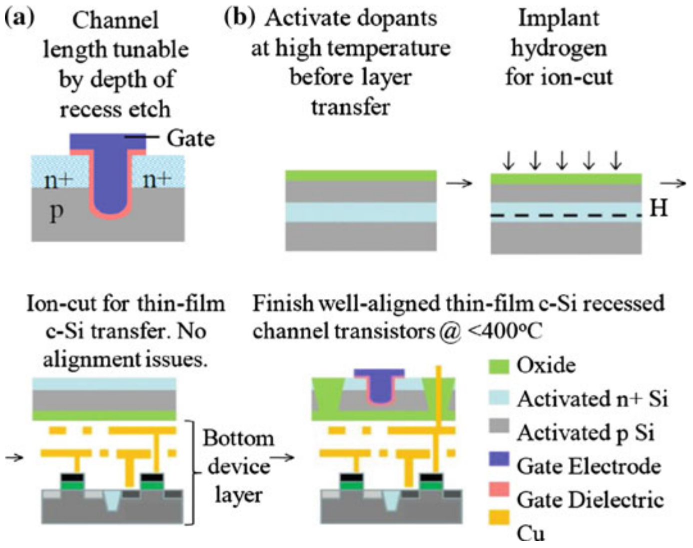

(a)
Channel length tunable by depth of recess etch
Gate
n+
n+
p
(b) Activate dopants at high temperature before layer transfer
Implant hydrogen for ion-cut
H
Ion-cut for thin-film c-Si transfer. No alignment issues.
Finish well-aligned thin-film c-Si recessed channel transistors @ <400ºC
Bottom device layer
Oxide
Activated n+ Si
Activated p Si
Gate Electrode
Gate Dielectric
Cu

Fig. 3.8 **a** A recessed channel transistor. **b** Process flow for Monolithic 3D logic. Bottom device layer with Cu/low k does not see more than 400 °C

node. The RCAT transistor is competitive with standard planar transistors [19] and looks like the inverse of a FinFET. High-temperature dopant-activation steps are done before transferring bilayer n+/p silicon layers atop Cu/low-k using ion-cut. The transferred layers are un-patterned; therefore, no misalignment issues occur while bonding. Following bonding, sub-400 °C etch and deposition steps are used to define the recessed-channel transistor. This is enabled by the unique structure of the device. These transistor-definition steps can use the alignment marks of the bottom Cu/low-k stack since transferred silicon films are thin (usually sub-100 nm) and transparent. Sub-50 nm diameter through-layer connections can be produced due to the excellent alignment.

The key idea for the RCAT process is the activation of the semiconductor-layer doping prior to the layer-transfer step. This completely avoids thermally damaging the underlying-layer interconnect or transistors. Forming the RCAT transistor after the layer transfer uses etch and deposition processes, which do not require high temperatures. This type of flow could be used for other types of transistors such as the junction-less transistor (gated resistor) or for vertical transistors as demonstrated by Besang Inc.

3.2.4.2 The Gate Replacement Process

Recently, the industry has moved to Hi-K metal gates and later fully adopted the “gate last” (gate replacement) approach to avoid exposing the hafnium oxide to high temperatures. This could be used for forming monolithic 3D as illustrated in

Fig. 3.9. First, the dummy gate stack transistors are processed with no temperature restrictions on a donor wafer. Then, using ion-cut and a carrier wafer, a small slice of the donor wafer is transferred to the top of a base wafer. Gate replacement is then performed by removing the H+ damaged gate oxide and replacing it with a HKMG stack using low-temperature etch and deposition processes. This flow has one serious limitation—alignment. As the layer transfer process is now being done on a patterned layer, the transfer misalignment of ∼1 μm would impact the second layer. This misalignment could be reduced in the case of a repeating pattern to the size of the repetition (100 s of nm).

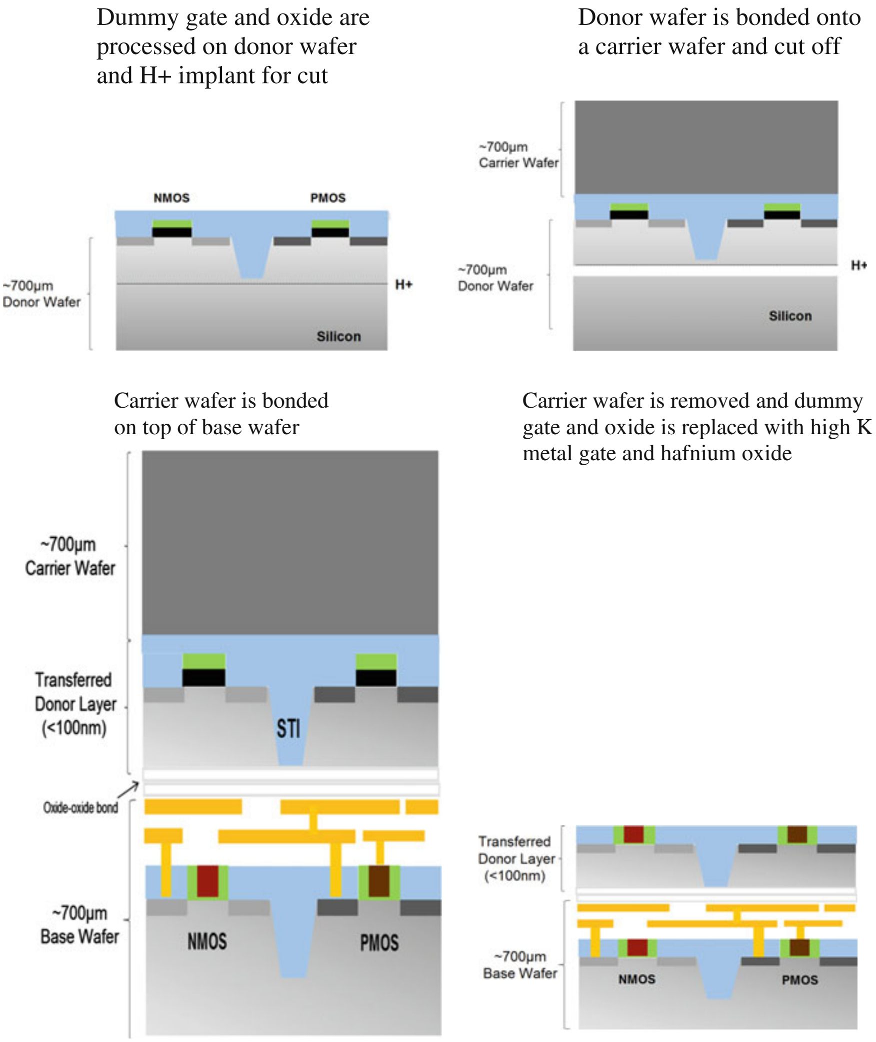

Dummy gate and oxide are processed on donor wafer and H+ implant for cut
NMOS
PMOS
~700µm Donor Wafer
H+
Silicon
Donor wafer is bonded onto a carrier wafer and cut off
~700µm Carrier Wafer
~700µm Donor Wafer
H+
Silicon
Carrier wafer is bonded on top of base wafer
~700µm Carrier Wafer
Transferred Donor Layer (<100nm)
STI
Oxide-oxide bond
~700µm Base Wafer
NMOS
PMOS
Carrier wafer is removed and dummy gate and oxide is replaced with high K metal gate and hafnium oxide
Transferred Donor Layer (<100nm)
~700µm Base Wafer
NMOS
PMOS

Fig. 3.9 Process flow for a gate replacement process

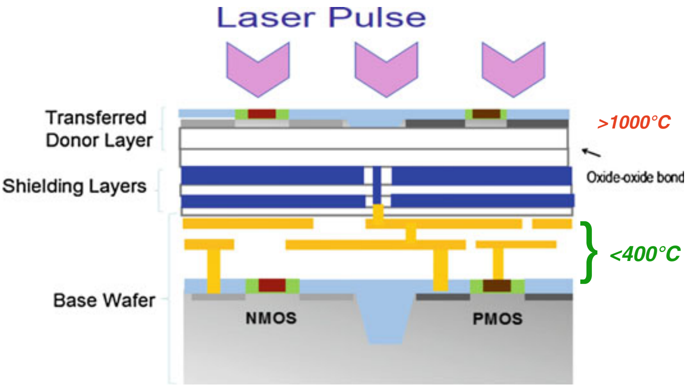


Fig. 3.10 Schematic of an example 3D-IC stack with shielding layers between stacked active layers, with laser annealing to activate the upper active layers

3.2.4.3 Laser-Annealing Process

The process utilizes laser annealing [5] for the 2nd-layer transistor activation while the base layers are protected by an inter-layer shield as illustrated in Fig. 3.10. This process is now becoming viable thanks to the fact that fully-depleted transistors can be integrated atop thin c-Si layers with relatively straightforward modifications of the gate-last CMOS process flow. The transferred donor layer is processed to form transistors with short pulsed-laser exposures providing, for example, annealing of process-induced damage and activation of dopants. It should be noted that the shield/heat sink layers are useful as Vss/Vdd planes and may also serve as a heat spreader and EMI shield.

The design of the shielding layer would be highly dependent on the type of laser used for the annealing process taking into account the thickness of the silicon and oxide layers. It has been shown [2] that, for excimer laser annealing with ~100 ns pulses, there is actually no requirement for a shielding layer as the heat absorbed by the silicon layer dissipates to less than 400 °C before it reaches the underlying interconnect layers.

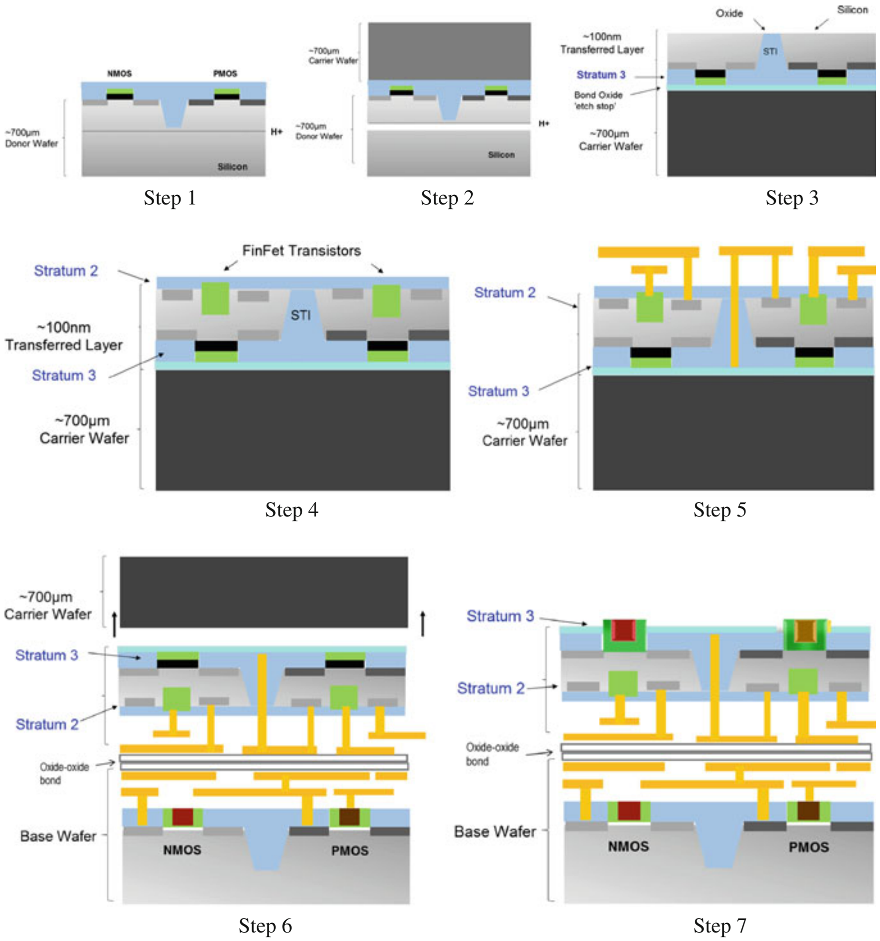
The laser annealing technique could be used as complementary processing to any of the other monolithic 3D process flows. In general, annealing techniques are useful to repair structural damage associated with other processes, and, accordingly, they could be used afterward.

3.3 Precision Bonders—A Game Changer for Monolithic 3D

The monolithic 3D flows presented in the previous section are all practical and would allow monolithic 3D ICs with good performance and competitive cost. Yet they do imply a transistor processing flow that is different from the one already developed and matured in 2D. In fact, all monolithic 3D IC process flows presented in the past required new front-end-of-line process development. Such new development efforts represent an additional barrier for adopting monolithic 3D technology. As it often happens in the semiconductor industry, the improvement in existing processing equipment or the development of a new type of equipment opens the door to new devices or improvements in our ability to process new devices. Past wafer bonders have been limited to about 1 μm wafer-to-wafer misalignment. At the 2014 IEEE S3S conference, two companies presented wafer bonders with about 200 nm wafer-to-wafer misalignment [20, 21]. These types of wafer bonders enable a game change in monolithic 3D manufacturing [22]. For the first time, monolithic 3D technology could be integrated within almost any semiconductor manufacturing facility using their existing transistor processes. In the following section, we provide the description of one such process flow.

3.3.1 Monolithic 3D IC Using Precision Bonders

The flow in Fig. 3.11 is based upon what we call ‘gate-replacement’ [23] processes, and it leverages the precision-bonder alignment accuracy. In step 1, a ‘donor’ wafer will be used to process a transistor layer labeled Stratum 3. The existing front-end process could be used. Alternatively for a gate-last flow, the donor wafer would be held before the gate-replacement phase. Then H+ would be implanted at the desired depth ($\sim$100 nm) in preparation for the layer-transfer step. In step 2, the donor wafer is bonded (oxide to oxide) to a ‘carrier wafer’ and ion-cut off. This bonding step does not require precise alignment. In step 3, the carrier wafer can now be annealed to repair the potential H+ implant damage. In step 4, the donor wafer is now processed to form Stratum 2. The existing front-line process can be used including FinFET or any other available front-line process. The choice of the transistor and the architecture for strata 2 and 3 should consider the need for vertical isolation in-between them. Note that between the transferred layer and the carrier wafer there is an oxide layer, which would be an excellent etch-stop allowing the transfer onto the target layer without the need for ion-cut. A preferred strategy is to use Stratum 2 for the high-performance circuits while Stratum 3 would be used for support of less sensitive circuits. All high-temperature steps should be completed at this point, since in the following step, interconnects are added. In step 5, add contacts and at least one metal layer. In step 6, bond (oxide to oxide or metal to metal) to the target wafer using the precise bonder alignment with less than 200 nm misalignment. Now grind and etch off the carrier wafer. (Not presented here are options to remove the carrier wafer for reuse.) In step 7, the dummy gate and the gate oxide of Stratum 3 can now be replaced, and connections can be made between Stratum 2, Stratum 3 and the underlying target wafer. Alignment and via processing are done just as between conventional BEOL metal layers, as the transferred layer is very thin ($\sim$100 nm).



Step 1 Step 2 Step 3

Step 4 Step 5

Step 6 Step 7

Fig. 3.11 Process flow for gate-replacement process and precise bonding

3.3.2 Smart Alignment

Having a thin transferred layer allows the through-layer via to be as small as a conventional BEOL interconnection via (∼50 nm). Yet the 200 nm bonding alignment window would appear to require a landing pad of 200 nm by 200 nm for each vertical connection. In addition to wafer-to-wafer misalignment, we also need to account for reticle-to-reticle misalignment. This would be highly dependent upon whether or not both wafers come from the same process line using the same lithography tool/stepper. The total misalignment across the wafer between the Stratum 1 and Strata 2/3 would include the reticle-to-reticle misalignment, the wafer bonding misalignment and some other error factors. Assuming that the total across wafer misalignment is less than 300 nm, then it would seem that a bonding pad of

300 nm × 300 nm at the top of Stratum 1 would be needed to allow a safe Strata 2/3 to Stratum 1 via connection. With Smart Alignment the connection is made by two perpendicular 300 nm long strips as seen in Fig. 3.12. The vertical strip is part of the top layer of the target (bottom) wafer.

After bonding, the through-layer via would be aligned to the target wafer in the Y direction and to the transferred layer in the X direction as seen in Fig. 3.13. The top connection strip could then be processed, aligned to the transferred (top) layer. This alignment scheme reduces the vertical connection overhead to a minimum, and it allows for multiple vertical connections per unit area of 300 nm × 300 nm.

3.3.3 Strata 2, 3—Examples

Figure 3.14 illustrates one example for circuit allocation for Stratum 2 and Stratum 3 with an intrinsic vertical isolation. For Stratum 2, the most advanced devices for forming high-speed logic could be used such as FinFET transistors. The SRAM for the high-speed logic circuit could be placed onto the close-by Stratum 3. A compelling option for the SRAM would be the use of Zeno technology [24] where a two-stable-state, single-transistor SRAM is enabled by a deep-implant back-bias. The vertical isolation is achieved by the back-bias. The FinFET transistor by design is also isolated from the substrate. This use of Strata 2 and 3 is compelling as the memory creates no obstructions and offers a very short fetch-path for memory access. Such a dual functional layer (Strata 2 + Strata 3) could be a product offered by itself as an add-on to many designs and single-chip systems.

Such process-flows with a dual functional layer could enable many new innovative devices such as:

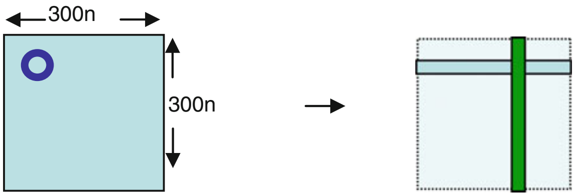


Fig. 3.12 Smart alignment

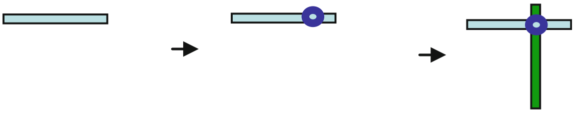


Fig. 3.13 Smart alignment

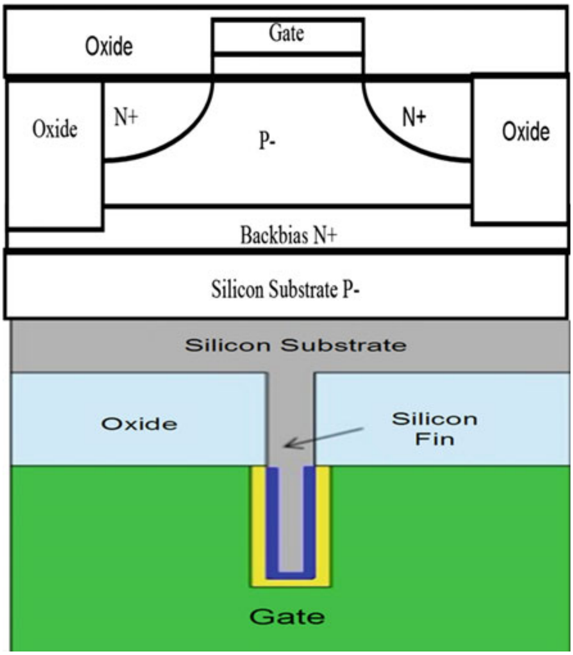


Fig. 3.14 1T SRAM over FinFET

- An image sensor on Stratum 3 with pixel electronics on Stratum 2 could provide an unparalleled dynamic range for cameras.
- A full redundancy layer [25] on Stratum 3 provides redundancy to Stratum 2, allowing almost unlimited logic integration on huge dies, essentially a server farm on a chip.
- A configurable logic fabric as an add-on…

3.3.4 Monolithic 3D Cost Estimates

It is well known that high cost is the number-one issue which slows down the adoption of 3D ICs based on TSV. The proposed monolithic 3D flow has the potential to overcome this barrier as it avoids the use of a thick layer with lengthy etch and deposition processes. In fact, it can provide circuit fabrics for two strata for a cost that is less than one wafer substrate. The donor wafer is reusable, and the cost of the first ion-cut is estimated to be less than $60 [26]. The carrier wafer could be reusable or utilize an inexpensive test wafer costing about $30. The estimated per-wafer cost of precision bonding is less than $20. Other steps involved in layer transfer, cleaning, etch, etc., are estimated at about $30 total. The costs for transistor formation for Strata 2 and 3 and their associated interconnects are no different from any other circuit fabrication costs. Accordingly, we estimate that the cost structure is comparable with the fabrication cost of 2D devices. Yet having the overall design built in a 3-strata fabric provides huge power, performance, and cost benefits.

3.4 EDA for Monolithic 3D

Design tools and algorithms for 3D ICs have been a subject of research work for many years. Some of that work had been summarized in related books [27], and relevant conferences [28]. Academic work, which resulted in many papers and tools, is being done at GeorgiaTech at the GTCAD Laboratory. The challenge is with respect to a commercial tool that would be ready for use by the semiconductor engineering and design community. Here we would have the classical chicken-egg challenge. The commercial EDA industry would wait for the design market to be large enough to justify the attention and the required investment, which is hard to have without the design tools to support such commercial design efforts.

This has been recognized by CEA Leti, which has done research work [29–31] allowing the use of commercial 2D EDA tools for a specific class of 3D designs. In one such case, one stratum has been allocated just for the drive and repeater while the other stratum carries the logic cell with minimum drive as illustrated in Fig. 3.15. The 2D tool would be used to place the 'modified logic cells' for the logic stratum, and the proper drive for each of the logic cells would be placed accordingly in the drive stratum [29].

A more flexible approach, named CELONCEL (Cell-on-cell) stacking [30, 31], allows cells to be placed on top of each other considering the pin-access issues. A physical design tool (CELONCELPD) was proposed that transforms the monolithic 3D placement problem into a virtual 2D problem solved using existing 2D placers. A highly parallelizable zero-one linear program formulation is used for layer assignment followed by a linear-program-based minimum perturbation for a high-quality 3D layout. Figure 3.16 illustrates the general EDA flow. It includes first deflating the cell library by 50 %, then using a commercial EDA placer and then re-inflating the cells after splitting them to two layers.

In the last section of this chapter, multiple monolithic 3D advantages are presented. Many of these leverage the aspect that each stratum would be processed independently—this is often referred to as heterogeneous integration. Accordingly, in many of those, the monolithic 3D design would be a 2D design using 2D EDA with some limited constraints related to the other strata. Such could be:

1. Memory over/under logic. A Memory process is preferably different from a logic process. In a design, which uses some very large memory blocks, those blocks could be either manually placed or assisted by a floor-planner tool. Then a 2D tool could be used to place and route the logic fabric with the memory pins serving as virtual I//O constraints.
2. Image sensor with pixel electronics. These could become mostly a full-custom design where each pixel has the same deep pixel electronics.

Fig. 3.15 Cell-on-Buffer (CoB) concept: **a** conventional 2D cell, **b** its equivalent 3D CoB cell

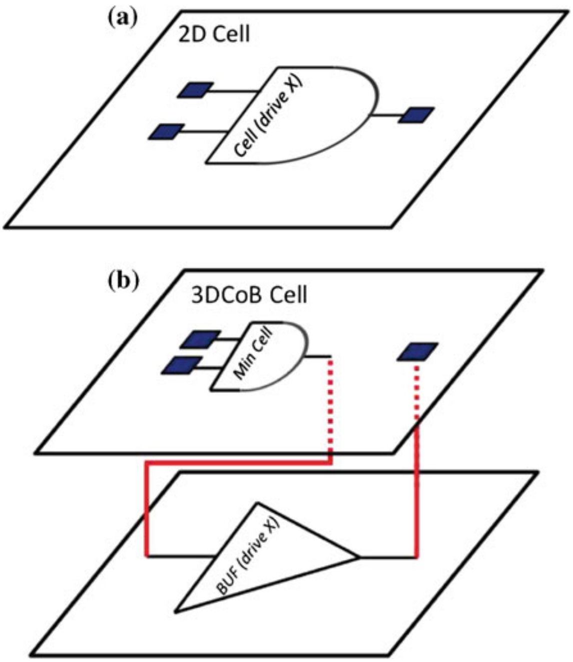
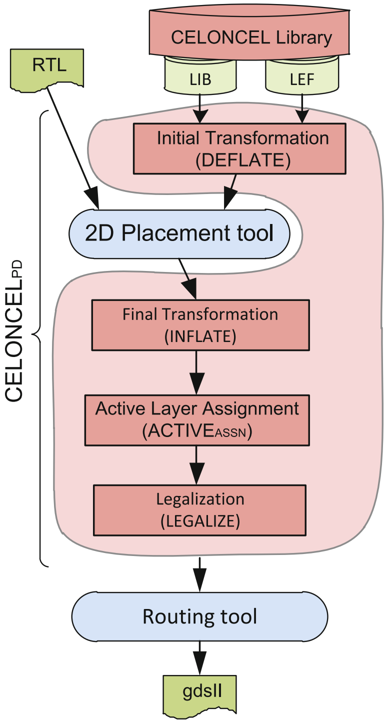


Fig. 3.16 Logic-to-layout using 2D commercial EDA for cell-on-cell


CELONCEL Library
RTL
LIB
LEF
Initial Transformation (DEFLATE)
2D Placement tool
Final Transformation (INFLATE)
Active Layer Assignment (ACTIVEASSN)
Legalization (LEGALIZE)
CELONCELPD
Routing tool
gdsII

3. 3D FPGA. A full-custom-design technique could be used for the regular terrain of the programmable logic array.

3.5 Managing the Heat

Several questions concern the heat-removal aspect for 3D IC's. The first question relates to having more transistors in a smaller space. While more complex circuits present an ever-increasing power challenge, having them built in monolithic 3D is an important part of the solution as it is well documented that 80 % of the power consumption is due to on-chip connectivity [32]. The more interesting question relates to the fact that Stratum 2 transistors are thermally isolated (surrounded by oxide) and without direct access to the silicon bulk for heat removal as is illustrated in Fig. 3.17.

Figure 3.18 illustrates the solution of using the power-delivery network (PDN) for heat removal. This work was reported in IEDM 2012 [33]. Having Stratum 2 only about 1 micron away from the bulk allows a very effective heat removal path through the power-delivery network (Fig. 3.19).

The PDN would enable removing the overall heat, but there could still be specific hot spots and active areas that might not have a thermal connection path to the PDN. The first step would be to add a heat-spreader layer to even out hot spots. In the previously presented Fig. 3.9, a heat spreader is illustrated underneath the transistor layer of the upper strata. Such a heat spreader could be used to shield the interconnect layers from top heat, act as power delivery and provide EMI/RFI isolation. The heat-spreader layer could be constructed from copper with thin isolation from the transistor layer above, or from less common, but even better heat-conducting material such as graphene or CVD diamond. In some cases, the power-delivery path could be too far from an active heat generating source, and some active cells might not have a power connection at all.

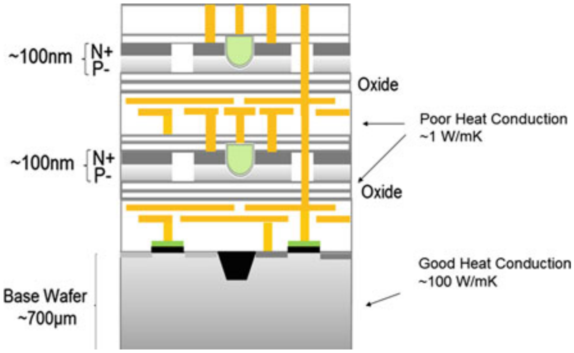
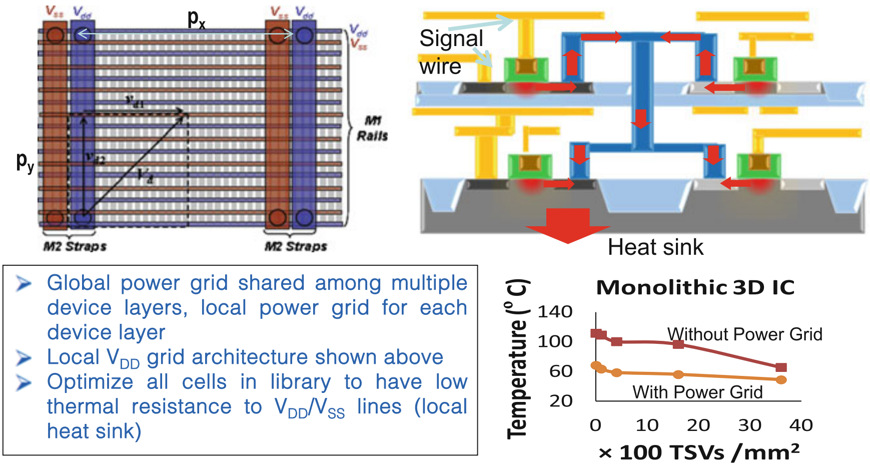


Fig. 3.17 The inner-strata transistor has no natural path for heat removal



- Global power grid shared among multiple device layers, local power grid for each device layer
- Local V_{DD} grid architecture shown above
- Optimize all cells in library to have low thermal resistance to V_{DD}/V_{SS} lines (local heat sink)

Fig. 3.18 Heat removal by the power-delivery network (PDN) [33]. © IEEE 2012

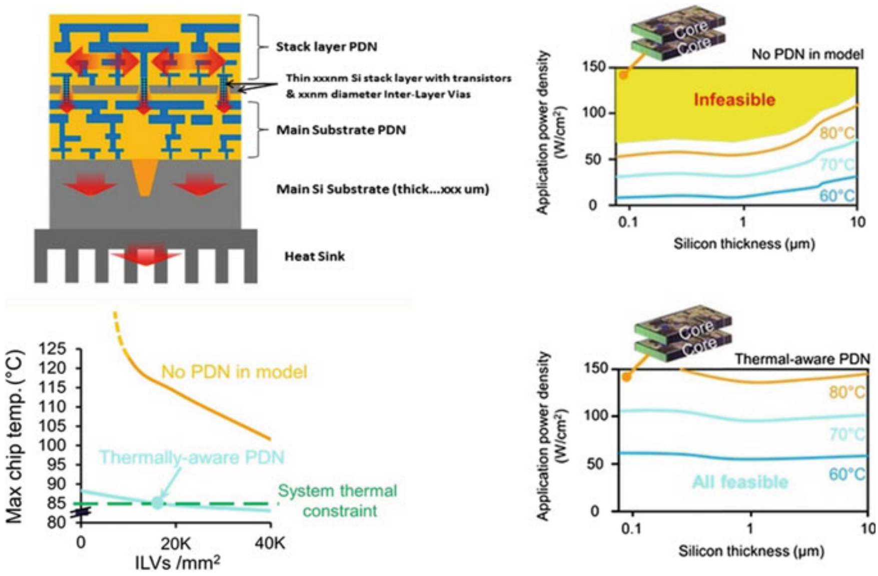


Fig. 3.19 Heat removal by the power-delivery network (PDN) [33]. © IEEE 2012

Figure 3.20 illustrates an important part of the heat removal for monolithic 3D—thermally conducting, electrically isolated contacts [34]. A simple solution could be a reverse-biased contact (illustrated in Fig. 3.20) for a common cell such as a transmission gate, which has no electrical connections to the PDN.

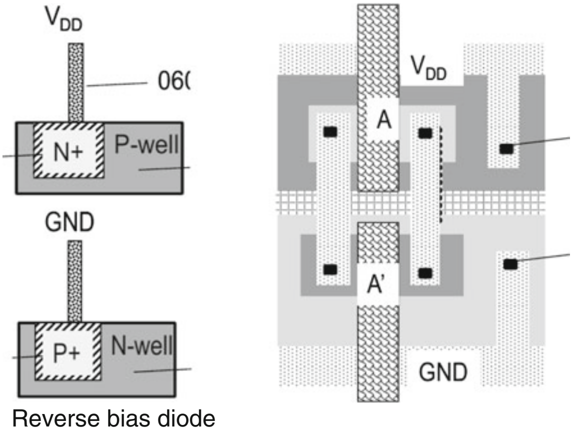


Fig. 3.20 Thermally conducting, electrically isolated contacts

3.6 3D Memories: 3D NAND,…

3.6.1 Introduction to BiCS

In landmark papers [35, 36], Toshiba introduced in 2007 a new approach to memory processing, which they call—BiCS—Bit-Cost Scalable Flash Memory. The key idea of BiCS is to share lithography and processing for multiple layers of processing. By properly architecting a memory-design and -processing flow, many layers could be processed together allowing scaling of the device into the vertical direction while keeping the incremental cost very low as illustrated in Fig. 3.21 [37].

It is now clear that 3D NAND, also called V NAND, is the scaling path for the Non-Volatile (NV) Memory industry. In mid-2013, Samsung announced the mass production of 24-layer 3D NAND, and, in 2014, a second generation with 32 layers has been released to the market. Figure 3.22 is a cartoon by Samsung illustrating the change to monolithic 3D already taking place for memory products.

3.6.2 3D-NAND

Since the first BiCS disclosure, multiple 3D Memories with shared lithography have been made public. It seems that each and every memory vendor has its own architecture claiming it is the best. Most of these architectures use polysilicon transistors, which are good enough for most memory products and easier to process for a multi-layer 3D architecture. The following description presents one [38] of these architectures. The process-flow in Fig. 3.23 shows how shared lithography and processing could be used. Multiple oxide/nitride steps are used with multi-states charge trap dielectric and poly-silicon channel. The polysilicon-channel

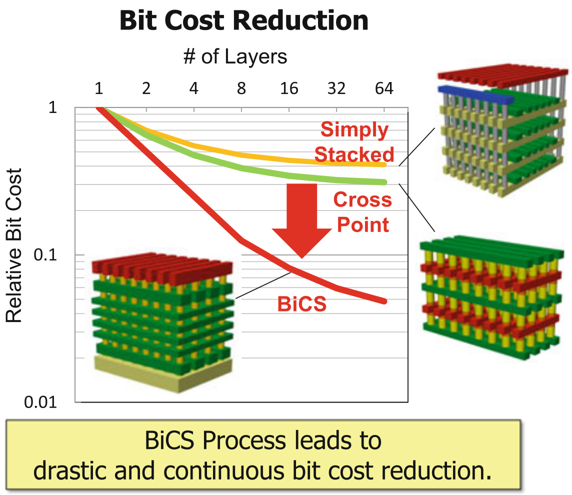


Fig. 3.21 Bit-cost reduction [37]. © IEEE 2014

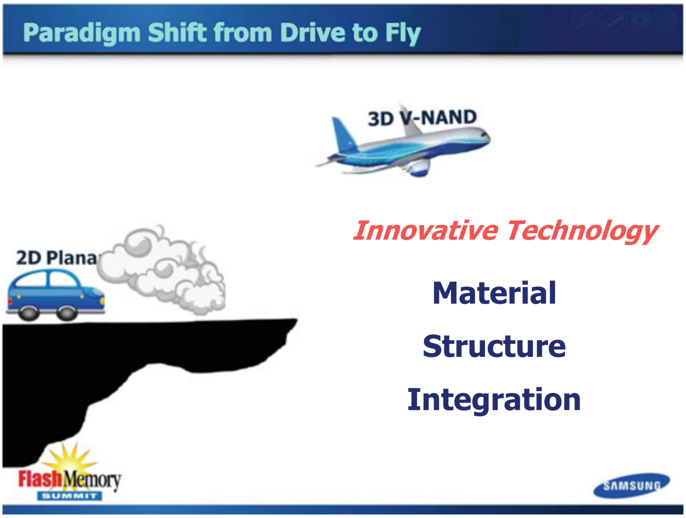


Fig. 3.22 Paradigm shift from 2D scaling to 3D scaling (*Source* SAMSUNG)

performance is acceptable since the grain size is large compared to cell dimensions. The use of an excimer laser to further increase the grain size has been shown to improve the performance. The device uses gate-all-around transistors, which provides excellent electrostatic channel control (Fig. 3.24).

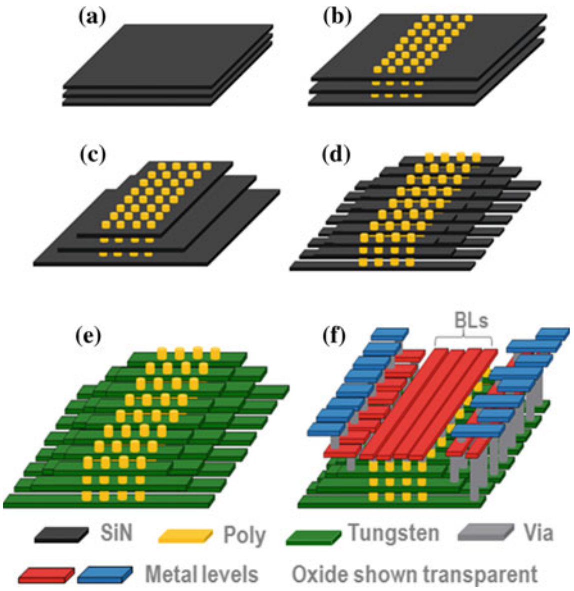


Fig. 3.23 Process-flow for 3D NAND Flash [39]. © IEEE 2009. **a** Deposit multiple Sio$_2$/SiN layers, **b** Etch hole and deposit channel poly (shared litho step), **c** Make staircase pattern for contacts (shared litho step), **d** WL cut etch (shared litho step), **e** Using flow in Fig. 3.24, make charge trap flash dielectrics and gate/WL, **f** BEOL

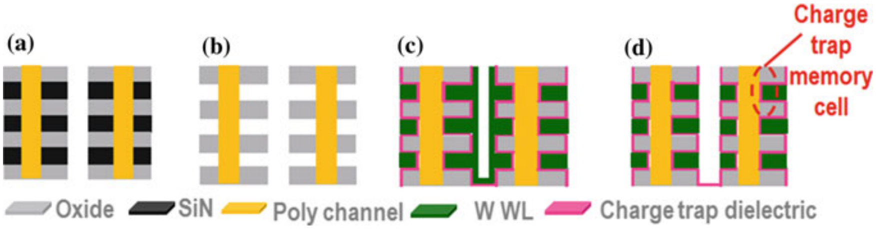
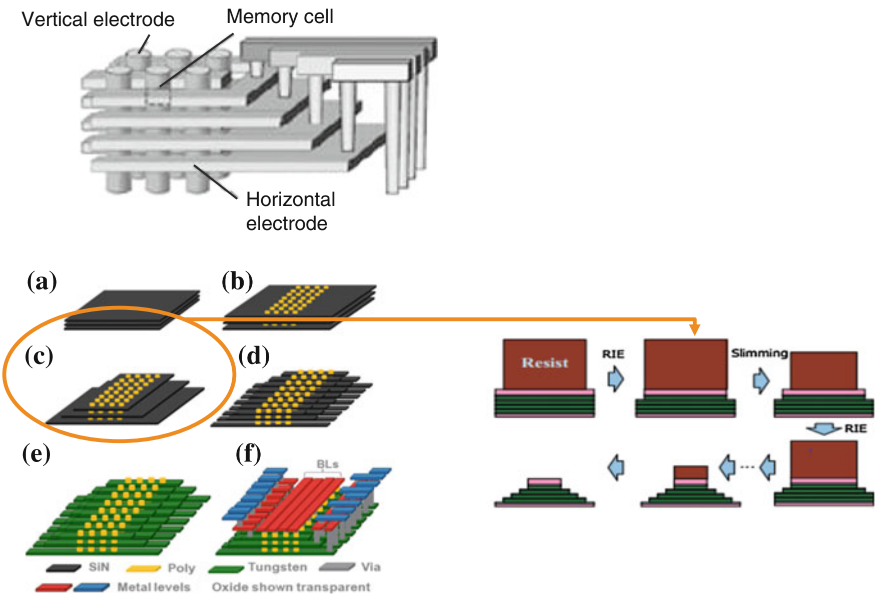


Fig. 3.24 Charge-trap dielectric and electrode definition [39]. © IEEE 2009. **a** After step (d) of Fig. 3.23, **b** Wet etch SiN, **c** Deposit charge trap dielectric and WL, **d** Gate node separation

3.6.3 Making Contact Without Adding Lithography Steps

An important complementary technology allows for making contacts to each of the stacked layers without the need for an individual lithography step. Figure 3.25 illustrates the contact-formation flow using an approach similar to 'spacer technology'. It utilizes isotropic etch/slim followed by anisotropic etch (RIE) to form a staircase structure, which, with one lithography step, can then contact each of the memory stacked planes.



Source: H. Tanaka, et al., VLSI 2007

Fig. 3.25 Staircase patterns for contacts of the individual stacked layers. © IEEE 2007. **a** Deposit multiple Sio_2/SiN layers, **b** Etch hole and deposit channel poly (shared litho step), **c** Make staircase pattern for contacts (shared litho step), **d** WL cut etch (shared litho step), **e** Using flow in Fig. 3.24, make charge trap flash dielectrics and gate/WL, **f** BEOL

3.6.4 3D-NOR Flash

The NAND architecture became the most popular NV Memory architecture by providing the highest memory density and accordingly the lowest cost per bit. In some memory technologies, there is a need for a NOR architecture such as in 3D-DRAM [40] and RRAM [38]. These architectures could use polysilicon or mono-crystallized silicon. Figure 3.26 illustrates such a NOR architecture for RRAM application.

3.7 Advanced Work—Non-silicon Monolithic 3D

3.7.1 III–V Semiconductor 3D Integration

At SMART LEES, Singapore, a manufacturing facility is being put in place [41] to integrate III–V crystal layers with foundry-processed silicon offering truly heterogeneous integration as illustrated in Fig. 3.27.

Here, the heterogeneous integration is in two aspects. First is the ability to mass-produce some base generic function on a CMOS wafer by conventional foundries. Then a far smaller facility processes the monolithic 3D integration of

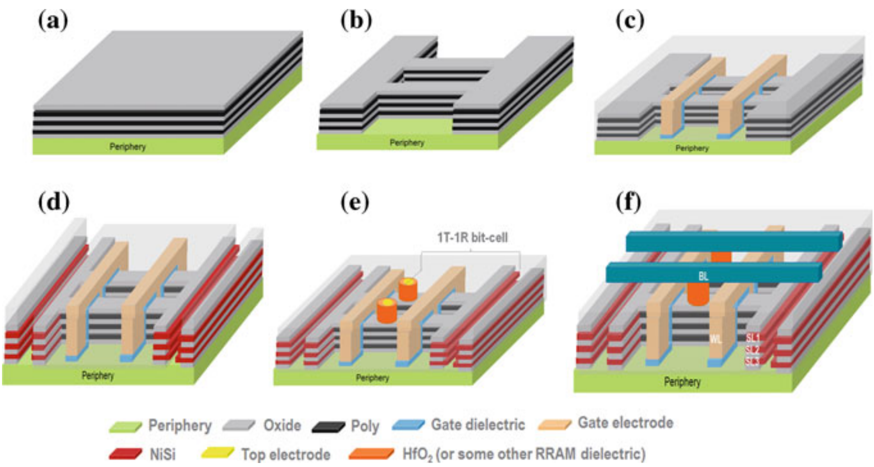
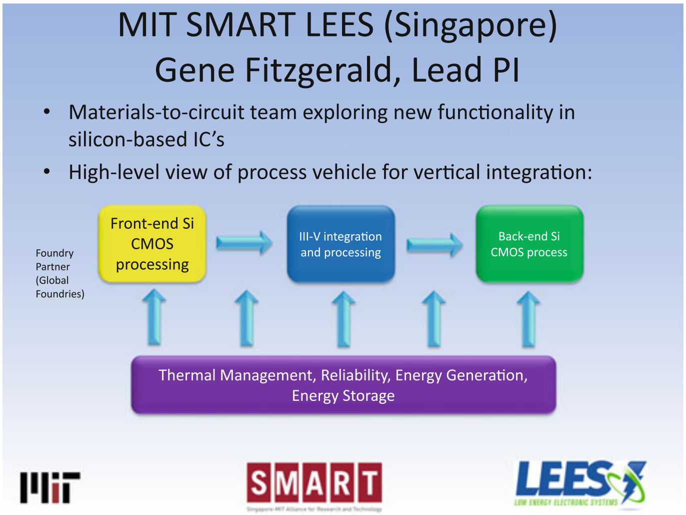


Fig. 3.26 3D RRAM—NOR Architecture [38]. © IEEE 2014. **a** Deposit multiple Sio_2/poly Si layers. Or use ion-cut to make Sio_2/c-Si layers, **b** Pattern (shared litho step), **c** Form gate of select transistors (shared litho step), **d** Pattern SL, then silicide (shared litho step), **e** Form RRAM dielectric and electrode for multi level 1T-1R cells (shared litho step), **f** Form BLs



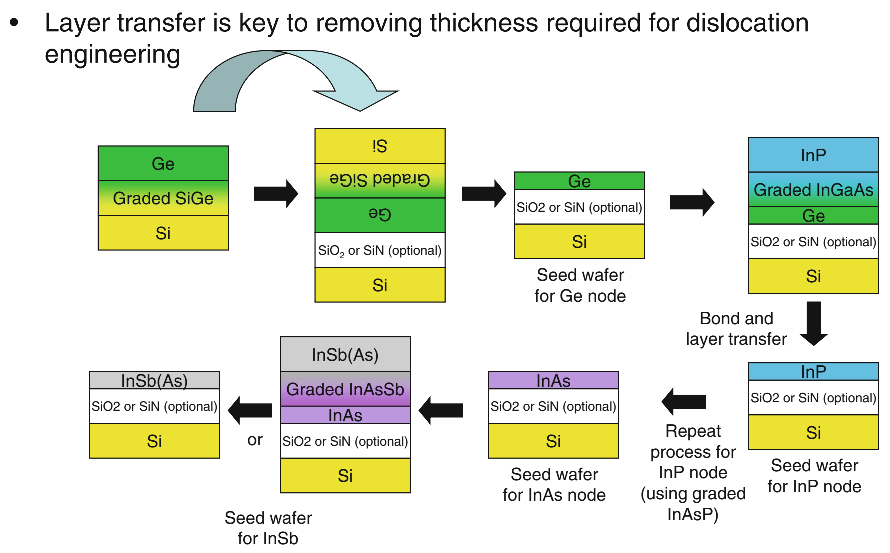
MIT SMART LEES (Singapore)
Gene Fitzgerald, Lead PI

- Materials-to-circuit team exploring new functionality in silicon-based IC's
- High-level view of process vehicle for vertical integration:

Foundry Partner (Global Foundries)
Front-end Si CMOS processing
III-V integration and processing
Back-end Si CMOS process
Thermal Management, Reliability, Energy Generation, Energy Storage
SMART
LEES

Fig. 3.27 Heterogeneous integration of III–V over silicon from foundries [41]. © IEEE 2014

upper strata of a lower-volume structure, customizing the device. The upper strata could be another type of crystal such as III–V materials. The III–V materials could be processed by epitaxial growth over base silicon materials and utilizing ion-cut to transfer over oxide and eventually on top of a foundry-base silicon wafer, like in Fig. 3.28.



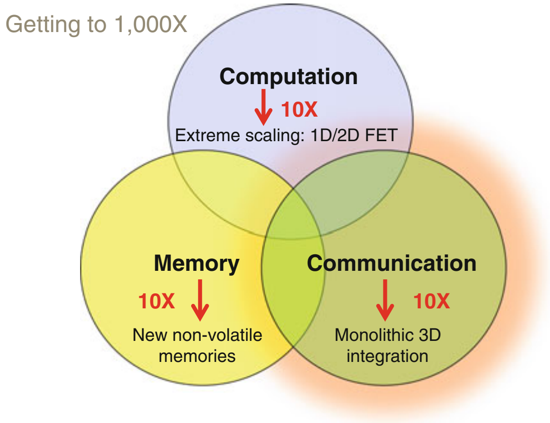
- Layer transfer is key to removing thickness required for dislocation engineering

Fig. 3.28 Using layer-transfer (Ion-cut) for heterogeneous integration [41]. © IEEE 2014

3.7.2 *Monolithic 3D Integration of Semiconductor, Carbon Nano Tube, STT MRAM and RRAM*

A research effort going on at Stanford [42] and other leading US universities is looking to provide 1000× improvements over conventional semiconductor technology by leveraging monolithic 3D integration of semiconductor strata with a CNT STT-MRAM and R-RAM stratum—(Figs. 3.29, 3.30).

Fig. 3.29 1000× improvements [42]



Getting to 1,000X
Computation
10X
Extreme scaling: 1D/2D FET
Memory
10X
New non-volatile memories
Communication
10X
Monolithic 3D integration

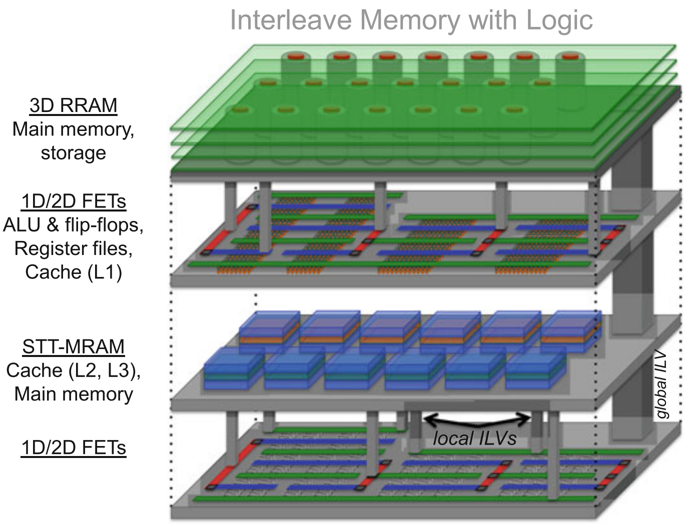


Fig. 3.30 Monolithic 3D integration of Semiconductor, Carbon Nano Tube, STT RAM and RRAM [42]. © IEEE 2014

3.8 The Monolithic 3D Advantages

3.8.1 Introduction

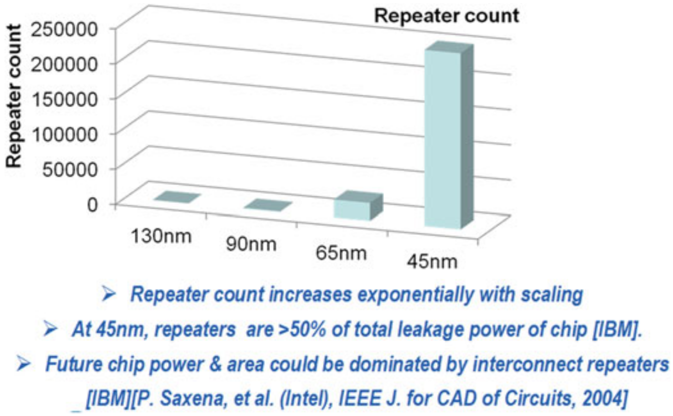
The most important aspect of scaling is the exponential increase of device integration. In this section, we cover some of the other advantages of monolithic 3D integration.

3.8.2 Reduction in Die Size and Power

3.8.2.1 Reduction in Die Size

Dimensional scaling has always been associated with increased wire resistivity and capacitance—see Fig. 3.4. Every node of dimensional scaling is associated with larger output drivers and more buffers and repeaters. Figure 3.31 illustrates the rapid increase of the number of transistors associated with the increased interconnect challenge.

Reduced interconnect length due to 3D stacking provides a reduction of buffers and the average transistor size. MonolithIC 3D Inc. released an open-source high-level simulator IntSim v2.0 to simulate a given design's expected size and

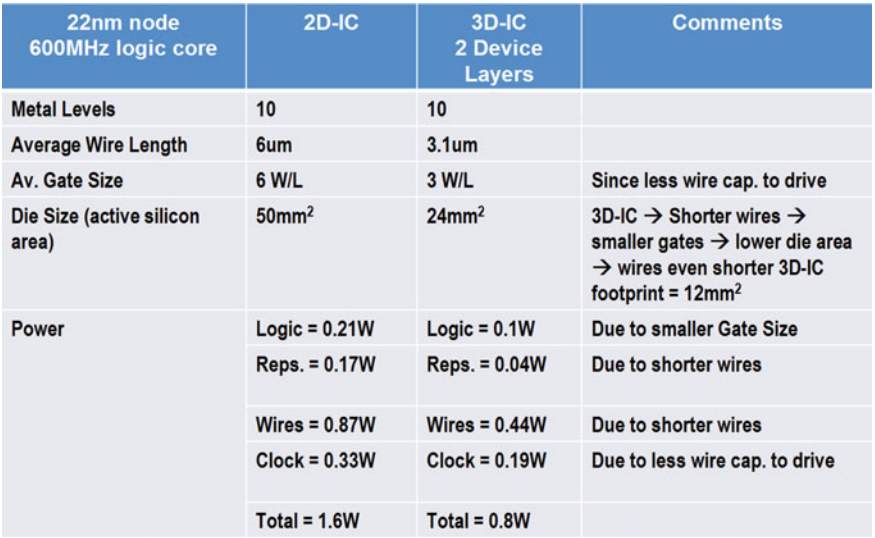


- *Repeater count increases exponentially with scaling*
- *At 45nm, repeaters are >50% of total leakage power of chip [IBM].*
- *Future chip power & area could be dominated by interconnect repeaters _[IBM][P. Saxena, et al. (Intel), IEEE J. for CAD of Circuits, 2004]*

Fig. 3.31 Repeater and buffers consume escalating parts of the end-device power and area. *Source* IBM Power processors R. Puri et al. SRC Interconnect Forum, 2006

power based on process parameters and the number of strata. Using the simulator, we can see in Fig. 3.32 that a 2D design of 50 mm^2 area with an average gate-size of W/L = 6, will only need an average gate-size of W/L = 3 and accordingly only 24 mm^2 of total circuit area, if folded into two strata (the footprint will be therefore just 12 mm^2).

These results are in-line with many other monolithic-3D research results.



22nm node 600MHz logic core	2D-IC	3D-IC 2 Device Layers	Comments
Metal Levels	10	10	
Average Wire Length	6um	3.1um	
Av. Gate Size	6 W/L	3 W/L	Since less wire cap. to drive
Die Size (active silicon area)	50mm^2	24mm^2	3D-IC → Shorter wires → smaller gates → lower die area → wires even shorter 3D-IC footprint = 12mm^2
Power	Logic = 0.21W	Logic = 0.1W	Due to smaller Gate Size
	Reps. = 0.17W	Reps. = 0.04W	Due to shorter wires
	Wires = 0.87W	Wires = 0.44W	Due to shorter wires
	Clock = 0.33W	Clock = 0.19W	Due to less wire cap. to drive
	Total = 1.6W	Total = 0.8W	

Fig. 3.32 Monolithic 3D folding can reduce the required silicon by 50 %

3.8.2.2 Reduction in Power

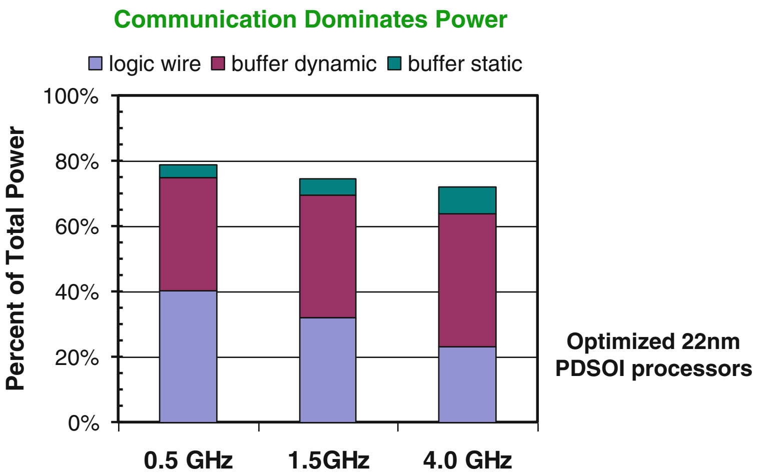
Figure 3.33 illustrates that interconnect is now dominating about 80 % of device power.

Monolithic 3D enables the folding of a circuit, with each stratum only about 1 μ above or below its neighbor, combined with a very rich vertical connectivity between the strata—with the potential to strongly impact the 10 % of wires that consume more than 90 % of the device active power—Fig. 3.34.

3.8.3 Significant Advantages for Using the Same Fab and Design Tools

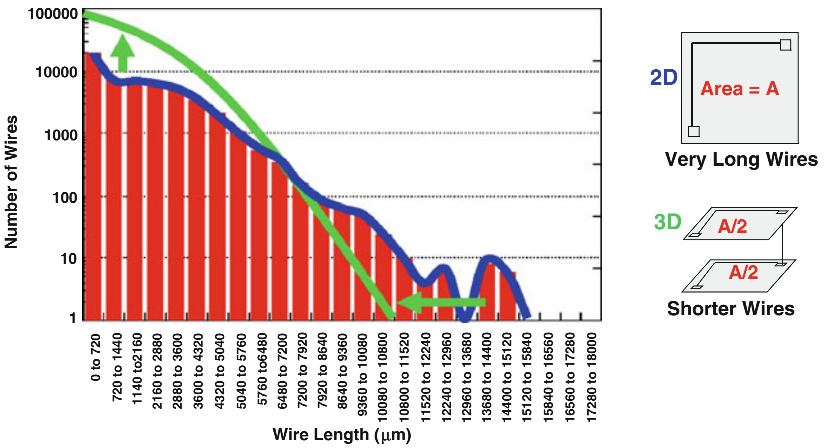
3.8.3.1 Depreciation

With dimensional scaling, every technology/process node requires a significant capital investment for new processing equipment—Fig. 3.35, significant R&D spending for new transistor process and device development—Fig. 3.36, and the building of an ever more complex and costly library and EDA flow.



- 70-80% of total logic power is for communication
  – Need proper consideration of wires!!

Fig. 3.33 Interconnect responsible for about 80 % of device power (IBM, Short-Course IEDM 2012)



- >50% of active power (switching) dissipation is in microprocessor interconnects
- >90% of interconnect power is consumed by only 10% of the wires

Fig. 3.34 Monolithic 3D can reduce the power and cost attributes by the long wires, MIT Lincoln Lab (After K. Guarini, IBM Semi R&D Center HPEC 2006)

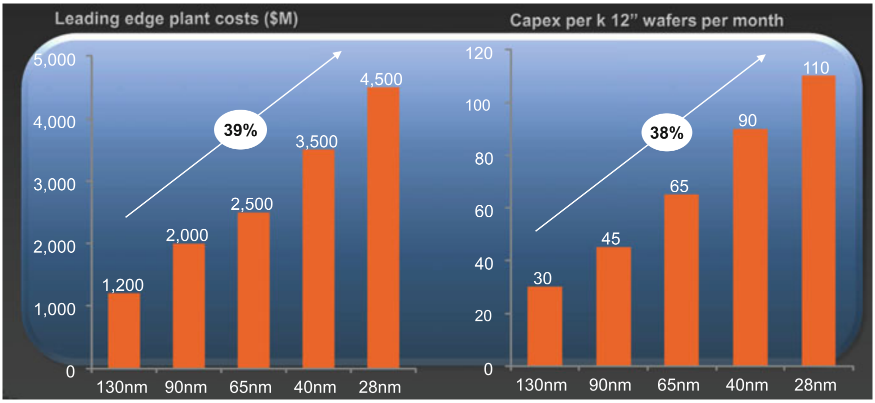


Fig. 3.35 Escalating Capex costs with dimensional scaling (*Source* World Fab Watch)

With monolithic 3D, these costs are not required as dimensions are maintained for multiple generations, and only the number of strata or layers is increased.

If the industry could use the same equipment and the same transistors and libraries for 4 years instead of 2, then all these costs could be depreciated over a longer time, with the resultant significant cost benefit.

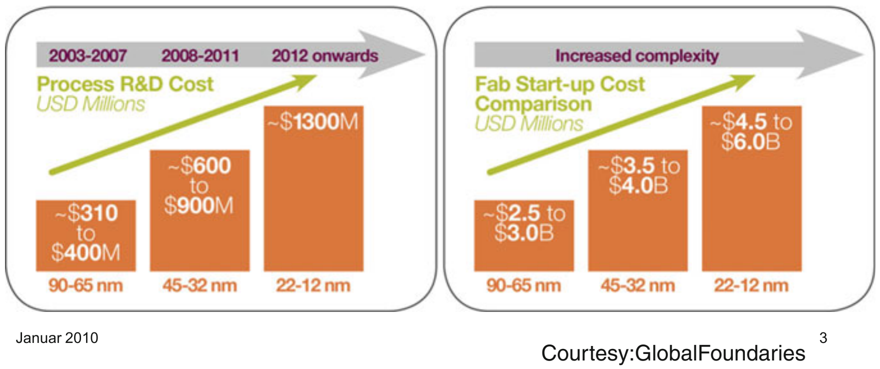


Fig. 3.36 Escalating process R&D costs with dimensional scaling

3.8.3.2 Learning Curve—Yield

Using the same transistor tools and EDA for longer periods has an additional important benefit. Learning curve equals yield improvement. With dimensional scaling, we face the predicament that, by the time we know how to manufacture a process node well, that learning quickly becomes obsolete as we quickly move on to the next node.

With monolithic 3D, the learning of the previous node stacking is directly utilized on the integration development of more strata, rather than on new materials, design-tool issues, etc. Figure 3.37 illustrates the dimensional scaling trend as each node of scaling is taking longer and costing more to get to a mature yield ('ramped-up').

3.8.4 Heterogeneous Integration

3D IC enables far more than an alternative for increased integration. It provides another dimension of design flexibility as shown in Fig. 3.38.

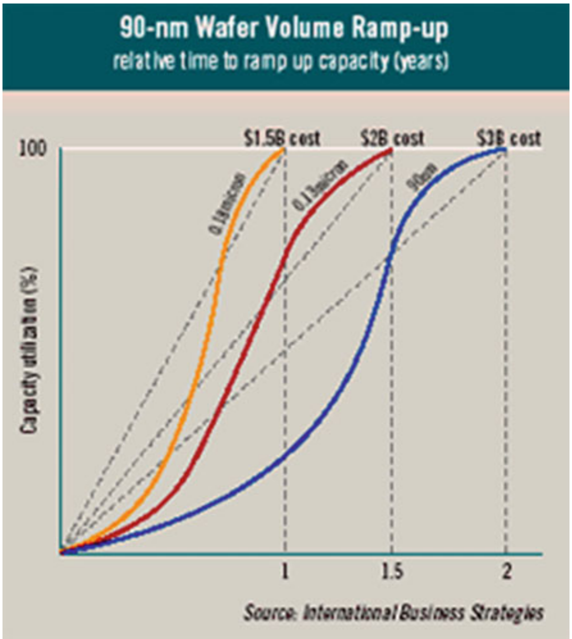
A well-known aspect of this flexibility is the ability to split the design into layers, which can be processed and operated independently, and still be tightly interconnected—especially for monolithic 3D as was presented in Sect. 3.7.

3.8.4.1 Logic, Memory, I/O

Let's start with quoting Mark Bohr, in charge of Intel's process development:

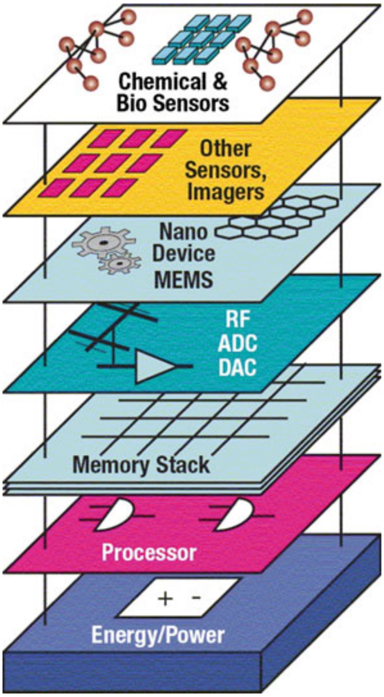
"One important perspective is that chip technology is becoming more heterogeneous. If you go back 10 or 20 years ago, it was homogeneous. There was a

Fig. 3.37 The learning curve with dimension scaling



Source: International Business Strategies

Fig. 3.38 Heterogeneous integration



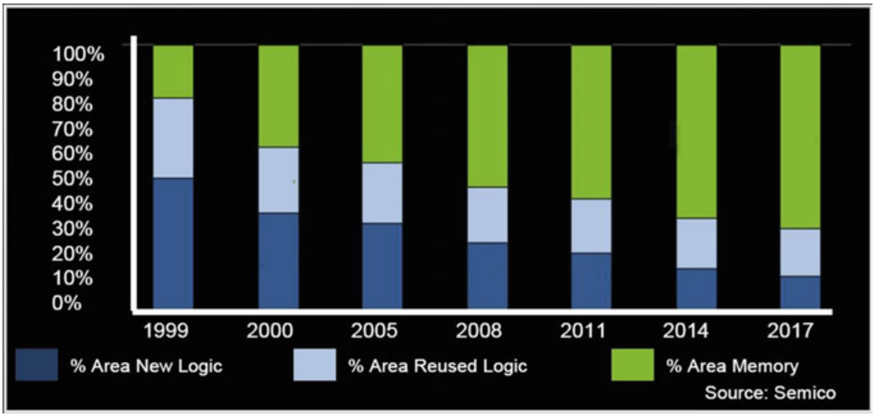
CMOS transistor, it was the same materials for NMOS and PMOS, maybe different dopant atoms, and that basic CMOS transistor fit the needs of both memory and logic. Going forward, we'll see chips and 3D packages that combine more heterogeneous elements, different materials, and maybe transistors with very different structures whether they're for logic or memory or analog. **Combining these very different devices onto one chip or into a 3D stack—that's what we'll see**. It will be heterogeneous integration" (added emphasis).

The most important market for semiconductor products is smart mobility. For this market, the SoC device needs to integrate many functions, such as logic, memory, and analog. In most cases, the pure high-performance logic would be about 25 % of the die area, 50 % of the area would be memory (Fig. 3.39), and the rest would be analog functions such as I/O, RF, and sensors.

In 2D, all the functions need to be processed together and bear the same manufacturing costs. In a monolithic 3D-IC stack using heterogeneous integration, each stratum is processed in an optimized flow, allowing for a significant cost reduction and no loss in optimized performance for each function type.

3.8.4.2 Strata of Logic

The logic itself can be constructed better using heterogeneous integration. In many cases, only a portion of the logic needs to be high performance while other portions could be more cheaply done using an older process node. Other scenarios could include designing different strata with different supply voltages for power savings, different numbers of metal-interconnect layers, or other variations of the design space.



Source: Semico

Fig. 3.39 Embedded memory to occupy about 70 % of SoC die area

3.8.4.3 Strata of Different Substrate Crystals and Fabrication Processes

3D enabled heterogeneous integration can be used as was presented in Sect. 3.7. Some layers can utilize silicon while others may use compound semiconductors. Some layers could be image sensors or other types of electro-optical structures and so forth.

3.8.5 Multiple Layers Processed Simultaneously—BiCS

An extremely powerful, unique advantage of monolithic 3D is the option to process multiple layers in parallel with one lithography step, as was detailed in Sect. 3.6. This option is most natural for regular circuits such as memory, but it is also available for other types of circuits with the right architectures.

The first merchants to recognize this option, and who are moving to monolithic 3D, are the NAND Flash vendors as seen in Fig. 3.40.

One of the clear future trends is the increase of content, often described by terms such as ‘big data’ and ‘abundant data’. This trend certainly illuminates the future

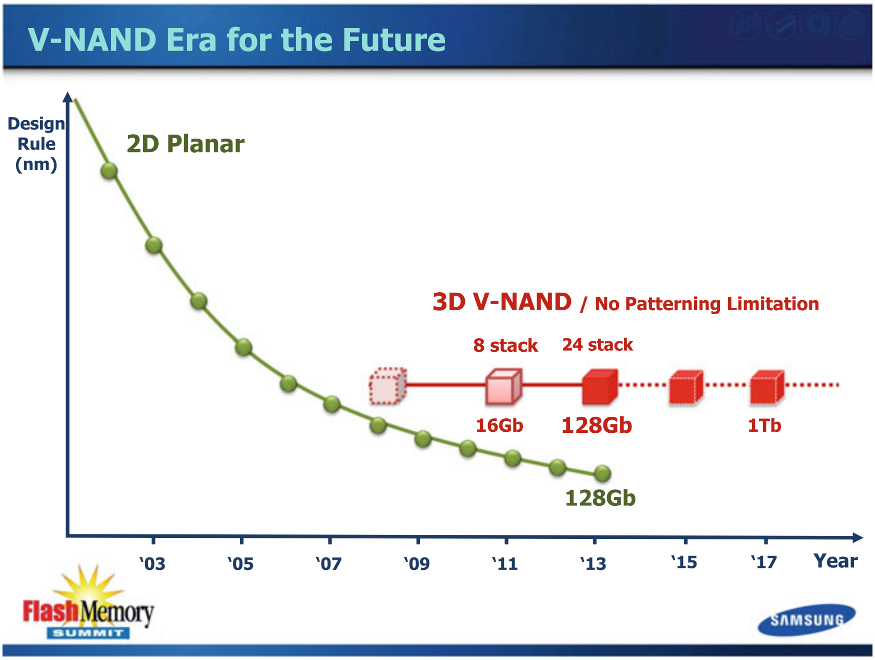


Fig. 3.40 For memory the future scaling is monolithic 3D (*Source* Flash Memory Summit and Samsung)

path for electronic systems. Monolithic 3D is well positioned to provide the best scaling path by integrating large amounts of 3D memory with logic providing a 1000× improvement in cost and power as previously discussed.

3.8.6 Logic Redundancy Allowing 100× Integration with Good Yield

The strongest value of an IC is the integration of many functions in one device. This is and will be the most important driver of Moore's Law, because, by integrating functions into one IC, we achieve orders-of-magnitude benefits in power, speed, and cost. At any given technology node, the limiting factor to integration is yield. As yield relates strongly to device area, most vendors are trying to limit the die size to about 50–100 mm^2. Some product applications require an extremely large die of over 600 mm^2, but those are rare and high value-added cases because the yield goes down exponentially as die size grows.

While memory redundancy is common in the IC industry, logic redundancy is only (and sparingly) used in a few FPGAs—no solution has been found after the failure of Trilogy, where "Triple Modular Redundancy" was employed systematically. Every logic gate and every flip-flop was triplicated with binary two-out-of-three voting at each flip-flop. Quoting Gene Amdahl: "*Wafer scale integration will only work with 99.99 % yield, which won't happen for* ***100 years.***" (*Source: Wikipedia*).

A unique advantage of monolithic 3D is the ability to construct redundancy for circuits including logic, with minimal impact on the design process and while maintaining circuit performance, such as shown in Fig. 3.41.

There are three primary ideas visible here:

- Swap at logic cone granularity.
- Redundant logic cone/block directly above, so no performance penalty.
- Negligible design effort, since the redundant layer is an exact copy.

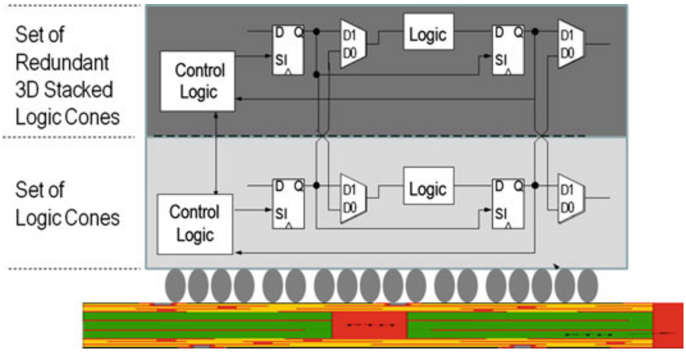


Fig. 3.41 Monolithic 3D enabling logic redundancy and repair

The new concept leverages two important technology breakthroughs:

The first is the scan-chain technology that enables a circuit test where faults are identified at the logic cone level. The second is the monolithic 3D IC, which enables a fine-grained redundancy: replacement of a defective logic cone by an identical logic cone that is only ~1 μm above.

Accordingly, by just building the same circuit twice, one on top of the other, with minimal overhead, every fault could be repaired by the replacement logic cone above. Such repair should have a negligible power penalty and a minimal cost penalty, whenever the base-circuit yield is as low as 50 %. There should be almost no extra design cost, and many additional benefits can be obtained.

This redundancy technique could be used also to repair faults throughout the device life-time, including in the field, which is a powerful advantage for some applications.

In today's designs, we expect more than one million flip-flops (and their logic cones). Consequently, if we expect one defect, then a device with redundancy layer would work unless the same cone is faulty on both layers, which, probability-wise, would be one in a million!

The ultra-integration value could be as much as:

- ~10X Advantage of 3D WSI versus 2D @ Board Level
- ~10X Advantage of 3D WSI versus 2D @ Rack Level
- ~10X Advantage of 3D WSI versus 2D @ Server-Farm Level

Overall, a ~1000× advantage is possible, all due to shorter wires. Instead of placing chips on different packages, boards and racks, we integrate on the same stacked huge chip.

3.8.7 3D-FPGA

Dimensional scaling is associated with escalating mask-set and design costs. Yet designers choose in most case to use old process nodes rather than an FPGA [43]. As a result, the most popular node for design currently is 130 nm, a node that is trailing the leading edge by about 6 process nodes. 3D FPGAs could significantly reduce this huge gap by drastically reducing area- and cost-penalties. Accordingly, a 3D FPGA would have the opportunity to increase innovation and growth in the semiconductor industry at large.

3D-FPGAs could simplify the use of anti-fuse technology for high-density programmable interconnects with the additional benefits of programming the interconnection layers from the top, providing an order of magnitude density improvement.

An additional advantage of a 3D FPGA is in having two compatible products. The prototype device would have extra strata for the interconnect programming, which could be removed for further cost reduction in volume production, illustrated in Fig. 3.42.

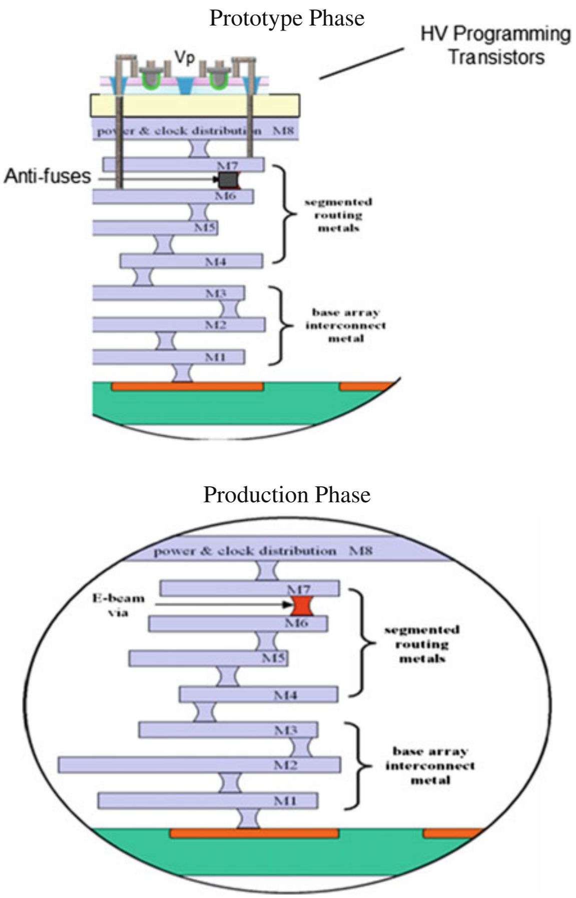

Prototype Phase
HV Programming Transistors
Vp
power & clock distribution M8
Anti-fuses
M7
M6
M5
M4
segmented routing metals
M3
M2
M1
base array interconnect metal
Production Phase
power & clock distribution M8
E-beam via
M7
M6
M5
M4
segmented routing metals
M3
M2
M1
base array interconnect metal

Fig. 3.42 The extra layers associated with interconnect programming could be removed for the volume part

3.8.8 Modular Platform

The 3D monolithic device would be a good fit for platform-based designs, wherein some parts of the device are used by all customers while other parts are tailored to a specific market/customer segment as illustrated in Fig. 3.43.

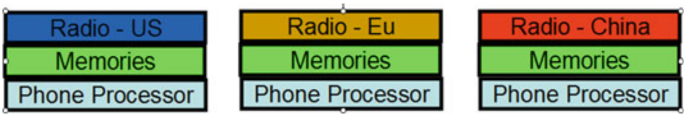


Fig. 3.43 3D-enabled modularity

Such a system architecture could be inexpensively used in many market segments and with multiple variations. One interesting application could be in the FPGA sector where the same platform could come with many flavors of memories and I/O.

3.8.9 Stacked Layers Are Naturally SOI

The upper layers of monolithic 3D devices are naturally Silicon-On-Insulator (SOI). The advantages of SOI are well-established, they increase with scaling, and they include:

- 90 % lower junction capacitance
- Near-ideal sub-threshold swing
- Reduced device cross-talk
- Lower junction-leakage
- Effective back-bias and multi-Vt options
- Multiple-gate operation for superb electrostatic channel control.

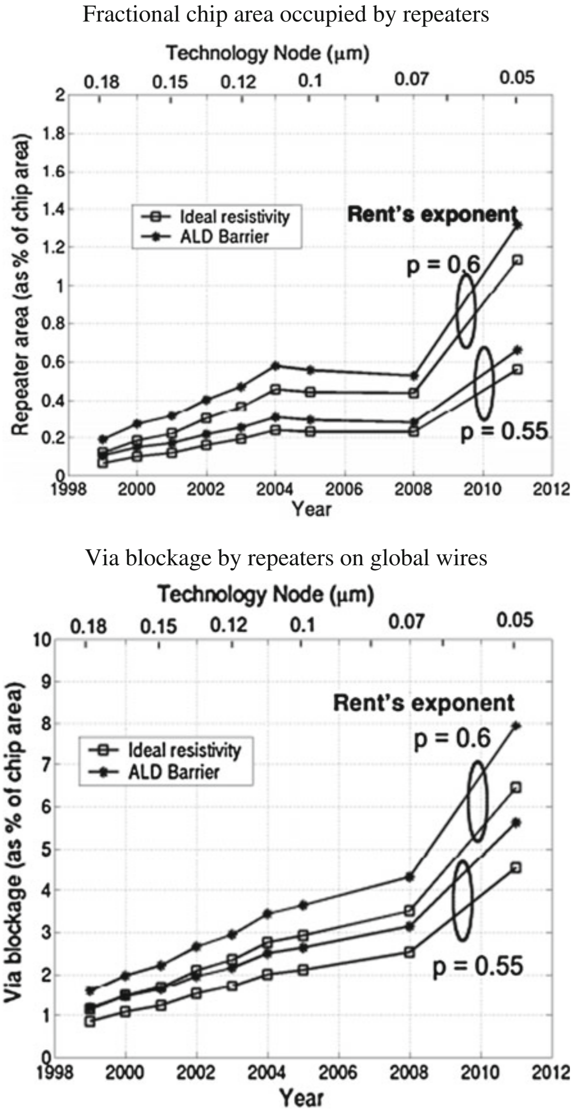
3.8.10 Local Interconnect Above and Below Transistor Layer

Improving on-chip interconnects is critical for enabling the increase in gate count. Simply adding interconnect layers provides limited improvement as each additional layer also adds to blockages/congestion in the intermediate layers created by the need to traverse them up and down the stack. In a monolithic 3D approach, interconnect can be formed and effectively used both above and below the transistor layer, thus doubling interconnect accessibility.

3.8.11 Re-buffering Global Interconnect by Upper Strata

Via blockage resulting from global interconnect buffering is growing exponentially as shown in Fig. 3.44. In addition to the reduction in buffers due to the significant reduction in the average wire-length in a 3D stack, moving those buffers to the upper stratum can effectively address the problem. Using such repeaters on a separate upper stratum does not add to routing congestion on the lower—and congested—metal layers, and it allows the utilization of a greater fraction of the active area.

Fig. 3.44 Escalating area penalty for repeaters (*Source* Prof. Saraswat, Stanford Univ., EE311: Interconnect Scaling, slide 46)


Fractional chip area occupied by repeaters

Via blockage by repeaters on global wires

3.8.12 Other Ideas

3.8.12.1 Image Sensor with Pixel Electronics

The image-sensor industry has moved to back-side illumination to increase the image-sensor area utilization. By adding the option for multiple strata, many additional benefits could be gained such as multi-spectrum day/night with extremely high dynamic range and other options as illustrated in Fig. 3.45.

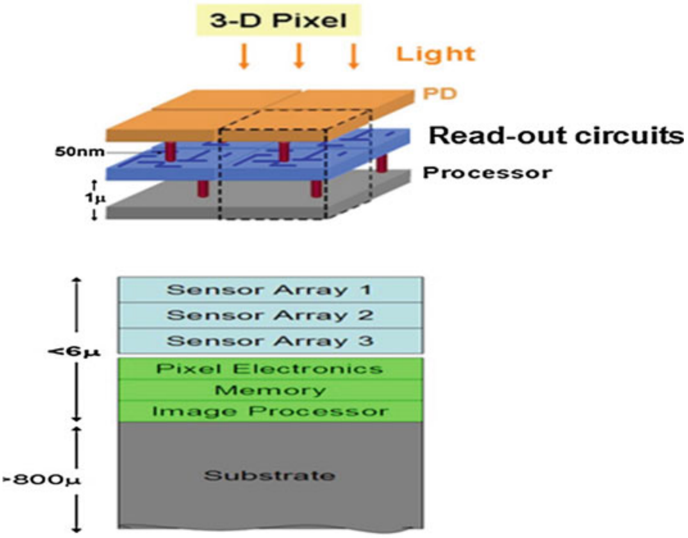
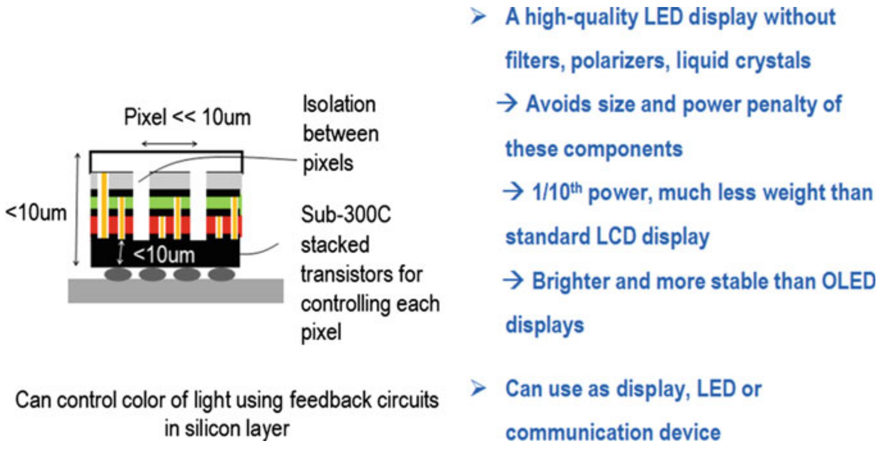


Fig. 3.45 Monolithic 3D image sensor

Pixel electronics behind every pixel could enable a very high dynamic range by counting and resetting individual sensors.

3.8.12.2 Micro-display

The display-market is always looking to reduce power and size while increasing the resolution and brightness. Monolithic 3D could provide ultra-high resolution with



Can control color of light using feedback circuits in silicon layer

- A high-quality LED display without filters, polarizers, liquid crystals
  - → Avoids size and power penalty of these components
  - → 1/10th power, much less weight than standard LCD display
  - → Brighter and more stable than OLED displays
- Can use as display, LED or communication device

Fig. 3.46 Monolithic 3D micro-display

extreme power-efficiency and minimal size, by combining drive electronics with strata of different-color light-emitting diodes as is illustrated in Fig. 3.46.

3.9 Conclusion

Monolithic 3D is a disruptive semiconductor technology. It builds on existing infrastructure and know-how, and it can bring to the hightech industry many more years of continuous progress. While it provides all advantages, once provided by dimensional scaling, monolithic 3D offers many additional options and benefits. Best of all, monolithic 3D can be used in conjunction with dimensional scaling.

This chapter presented various techniques for monolithic 3D processing as well as multiple applications for monolithic 3D devices. It should be noted that the processes and applications could be mixed and matched in various ways to support future market needs.

References

1. Zingg, R.: Stacked CMOS inverter with symmetric device performance. IEDM (1989)
2. Roos, G.: Complex 3D CMOS circuits based on a triple-decker cell. J. Solid-State Circuits **27**, 1067 (1992)
3. Or-Bach, Z.: The monolithic 3D advantage: monolithic 3D is far more than just an alternative to 0.7× scaling. In: IEEE 3DIC Conference (2013)
4. Batude, P. et al.: Direct bonding: a key enabler for 3D monolithic integration. In: Proceedings of the Electro-Chemical Society (ECS) spring meeting, vol. 16, pp. 47 (2008)
5. Rajendran, B.: Pulsed laser annealing: a scalable and practical technology for monolithic 3D IC. In: IEEE 3DIC Conference (2013)
6. Yang, C.-C.:. Record-high 121/62 µA/µm on-currents 3D stacked epi-like Si FETs with and without metal back gate, paper 29.6 IEDM (2013)
7. Shen, C.-H.: Monolithic 3D chip integrated with 500 ns NVM, 3 ps logic circuits and SRAM. IEDM (2013)
8. Lee, S.-Y.: Wafer bonding method, US Patent 7,470,142
9. Natio, T.: World's first monolithic 3D-FPGA with TFT SRAM over 90 nm 9 layer Cu CMOS, VLSI Technology (2010)
10. Wong, S.: Monolithic 3D integrated circuits VLSI-TSA (2007)
11. Ishihara, R.: Monolithic 3D-ICs with single grain Si thin film transistors ULSI 2011
12. Yonehara, T.: Monolithic 3D-ICs with single grain Si thin film transistors. JSAP International No. 4, July 2001
13. Sadaka, M.: Smart stacking™ and smart Cut™ technologies for wafer level 3D integration. ICICDT (2013)
14. Topol, A.W.: Enabling SOI-based assembly technology for three-dimensional (3D) integrated circuits (ICs). IEDM (2005)
15. Vinet, M.: Monolithic 3D integration: a powerful alternative to classical 2D scaling. IEEE S3S (2014)
16. Deleonibus, S.: Future challenges and opportunities for heterogeneous process technology. Towards the thin films, zero intrinsic variability devices, zero power Era, IEDM (2014)

17. ASIP Conf.—Hughes Metras, Dec. 12, 2014
18. Sekar, D.C.: Monolithic 3D-ICs with single crystal silicon layers. IEEE 3DIC Conference (2011)
19. Kim, J.Y.: The breakthrough in data retention time of DRAM using recess-channel-array transistor (RCAT) for 88 nm feature size and beyond. Symposium on VLSI Technology (2003)
20. Uhrmann, T.: Monolithic IC integration key alignment aspects for high process yield. IEEE S3S (2014)
21. Sugaya, I.: New precision alignment methodology for CMOS wafer bonding. IEEE S3S (2014)
22. Or-Bach, Z.: Precision bonders—a game changer for monolithic 3D. IEEE S3S (2014)
23. Or-Bach, Z.: Practical process flows for monolithic 3D. IEEE S3S (2013)
24. Widjaja, Y.: Method of maintaining the state of semiconductor memory having electrically floating body transistor. US Patent 8,514,623
25. http://www.monolithic3d.com/ultra-large-integration—redundancy-and-repair.html
26. http://www.monolithic3d.com/blog/how-much-does-ion-cut-cost1
27. Xie, Y.: BooK: three-dimensional IC Design. Springer, Berlin (2009)
28. Zhou, L.: CASCADE: a standard supercell design methodology with congestion-driven placement for three-dimensional interconnect-heavy very-large-scale integrated circuits. In: IEEE Transactions on CAD of IC and Systems, July 2007
29. Sarhan, H.: 3DCoB: a new design approach for monolithic 3D integrated circuits. ASP-DAC (2014)
30. Bobba, S.: CELONCEL: effective design technique for 3-D monolithic integration targeting high-performance integrated circuits. ASP-DAC (2011)
31. Bobba, S.: Cell transformations and physical design techniques for 3D monolithic integrated circuits. ACM JETCS, Sept. 2013
32. Chang, L.: IBM, technology optimization for high energy efficient computation. Short Course, IEDM (2012)
33. Wei, H.: Cooling three-dimensional integrated circuits using power delivery networks. IEDM (2012)
34. Sekar, D.C.: Semiconductor device and structure. US Patent 8,686,428
35. Tanaka, H.: Bit cost scalable technology with punch and plug process for ultra high density flash memory. VLSI (2007)
36. Fukuzumi, Y.: Optimal integration and characteristics of vertical array devices for ultra-high density, bit-cost scalable flash memory. IEDM 2007
37. Nitayama, A.: 3D NAND flash memories. Tutorials, S3S (2014)
38. Depak, S.: 3D RRAM. IEEE S3S (2014)
39. Jang, J.: Vertical cell array using TCAT (Terabit Cell Array Transistor) technology for ultra-high-density NAND flash memory. VLSI Symposium (2009)
40. Or-Bach, Z.: Semiconductor device and structure. US Patent 8,379,458
41. Fitzerald, E.: Monolithic 3D integration in a CMOS process flow. IEEE S3S (2014)
42. Ebrahimi, M.: Monolithic 3D integration advances and challenges: from technology to system levels. IEEE S3S (2014)
43. Or-Bach, Z.: FPGAs as ASIC alternatives: past & future. EE Times Apr. 21 (2014)

Chapter 4
Analog-Digital Interfaces—Review and Current Trends

Matthias Keller, Boris Murmann and Yiannos Manoli

Abstract By updating the figure-of-merit plots presented in CHIPS 2020 using new survey data collected over the years 2011–2015, this chapter discusses asymptotes and extracts recent improvement rates in the area of low-power, high-performance A/D conversion. Moreover, five years after the writing of CHIPS 2020, the developments in current architectures will be re-iterated, and the emerging concept of analog-to-information conversion will be discussed.

4.1 Introduction

Five years after our survey on analog-to-digital converters (ADCs) in *CHIPS 2020—A Guide to the Future of Nanoelectronics* [1], innovation and progress in data converter design is alive and well. In 2010, we had predicted that the future will bring further improvements in power efficiency, fueled by a combination of technology scaling, minimalistic design and digital assist. The purpose of this chapter is to provide a reality check, quantify recent progress and document the state-of-the-art.

M. Keller (✉) · Y. Manoli
Department of Microsystems Engineering—IMTEK, University of Freiburg,
Georges-Koehler-Allee 102, 79110 Freiburg, Germany
e-mail: mkeller@imtek.de

B. Murmann
Stanford University, 420 Via Palou, Allen 208, Stanford, CA 94305-4070, USA
e-mail: murmann@stanford.edu

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_4

4.2 General ADC Performance Trends

In order to illustrate the progress made over the past five years, we consider the conversion energy and conversion bandwidth plots that were introduced in [1]. As before, the data used for the plots shown in Fig. 4.1 is taken from the online survey data of [2], now extended up to the most recent data set from the 2015 International Solid-State Circuit Conference (ISSCC). From the points added between 2010 and 2015 (marked in gray), one can immediately see that there has been significant progress.

As far as the conversion energy (power divided by Nyquist sampling frequency) in Fig. 4.1a is concerned, we can summarize the key observations as follows. First, there are now 15 designs that reported a Walden figure-of-merit ($\mathrm{FOM_W}$) [3] of less than 10 fJ/conversion-step. In 2010, there was only one such design. Second, while we see improvements across the board, the successive approximation register (SAR) architecture stands out and now clearly dominates the low-energy design space. We will return to this point in Sect. 4.3. Third, and most importantly, we observe that the leading-edge designs for a signal-to-noise and distortion ratio (SNDR) beyond 50 dB align well with a slope of 4× per 6 dB (1 bit). As discussed in [1], this slope corresponds to the "thermal slope," i.e., the trade-off for circuits that are limited by thermal noise. The important conclusion to draw from this is that we have pushed our designs closer to thermal limits, indicating a higher degree of optimization away from technology-imposed limits.

The fact that most leading-edge designs (with SNDR > 50 dB) now follow the thermal slope has led to the widespread adoption of a figure-of-merit that takes this trade-off into account. Recall from [1] that the Walden FOM assumes a slope of 2× per 6 dB, which no longer fits the leading edge (see Fig. 4.1a). The so-called Schreier $\mathrm{FOM_S}$ was first defined in [4] and is based on a 4x per 6 dB slope in the trade-off between energy and dynamic range (DR). For our discussion below, we will utilize a modified version of this FOM that includes distortion [5], i.e., DR is replaced by SNDR:

$$FOM_S = SNDR(\mathrm{dB}) + 10\log\left(\frac{f_{snyq}/2}{P}\right) \tag{4.1}$$

Here, P stands for ADC power consumption and f_{snyq} is the Nyquist output sample rate of the ADC (twice the conversion bandwidth). The bold dashed line in Fig. 4.1a corresponds to $\mathrm{FOM_S}$ = 175 dB, which can be viewed as the state of the art. It is also worth noting that the data we use for SNDR are based on an input frequency near Nyquist to enable a fair comparison (see [2] for a discussion on this subject).

As far as the conversion bandwidth[1] plot in Fig. 4.1b is concerned, we observe that the improvements are significant, but not as pronounced as for conversion

[1] We define the conversion bandwidth as the highest input frequency for which the plotted SNDR was measured. This frequency is typically $f_s/2$ with exceptions noted in the fin_hf column of [2].

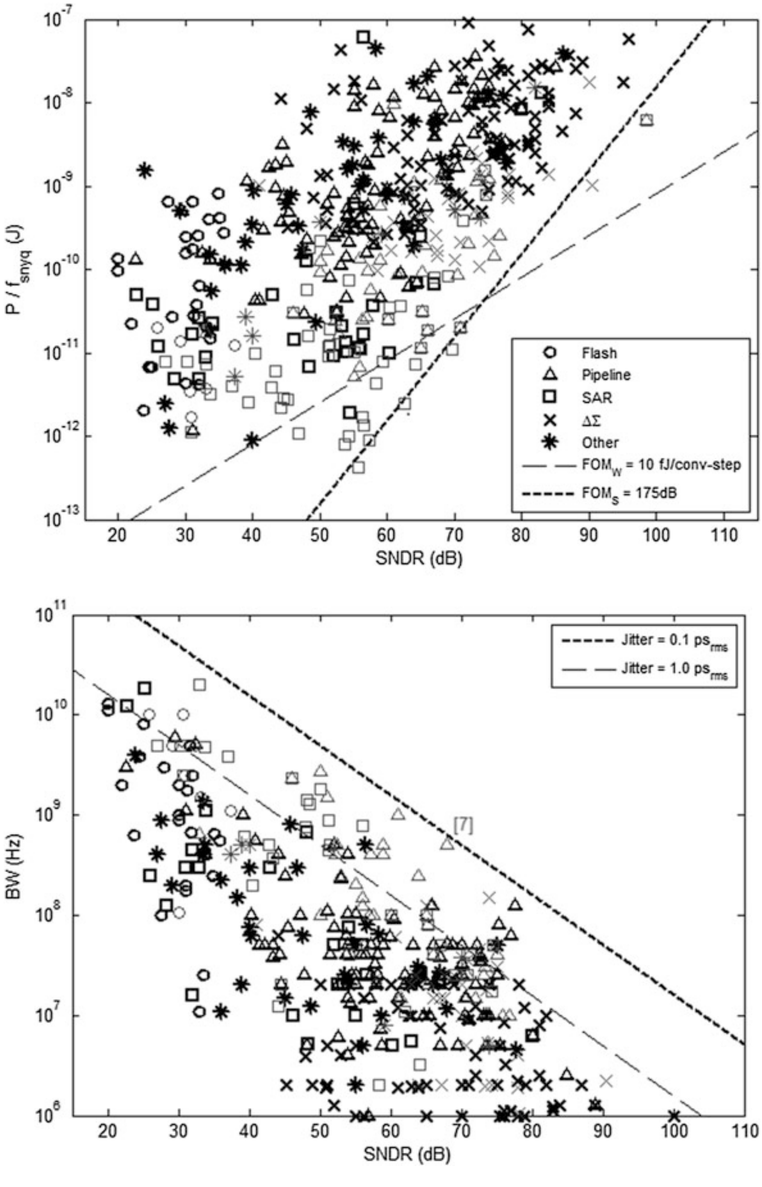


Fig. 4.1 ADC performance data (ISSCC 1997–2015 and VLSI circuit symposium 1997–2014). The *gray markers* indicate data reported after 2010. Conversion energy (**a**) and conversion bandwidth (**b**) versus SNDR

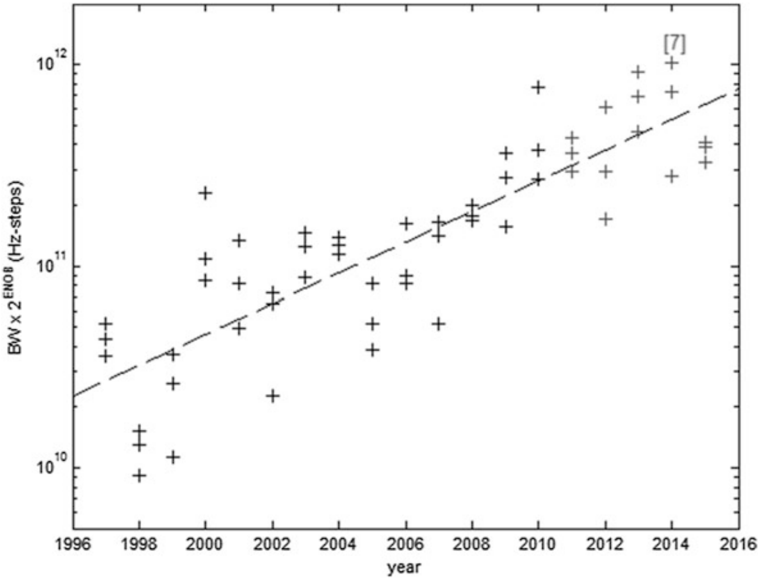


Fig. 4.2 Fit to speed-resolution product of the top 3 designs in each year. The fit line has a slope of 2×/4 years

energy. The reason for this is that the speed-resolution product is limited by our ability to make a low-jitter clock, and it is generally difficult to achieve a standard deviation better than 50 $\mathrm{fs_{rms}}$ [6]. The data point with the best combination of bandwidth and SNDR is [7], located at an equivalent aperture jitter of 127 fs. Note that since this converter suffers from other nonidealities, the actual clock jitter in this design must be significantly better.

To look into the conversion bandwidth trends more closely, we re-plot the speed-resolution chart (Fig. 4.2b) in [1] as shown in Fig. 4.2. From here we see that only two designs reported after 2010 surpass the speed-resolution product of [8] (which is the peak point for 2010). The overall progress slope for the speed-resolution product indicates a doubling every 4.0 years (was 3.6 years until 2010 [1]). Finally, we note that many other SAR-based designs have now managed to pass the line for 1 $\mathrm{ps_{rms}}$. However, pipelined ADCs (like [7]) still dominate the performance in the 60–80 dB range, mostly driven by the needs for wireless base stations [9].

Given that (1) has emerged as a figure-of-merit that not only accounts for the fundamental thermal noise trade-off, but also does a good job at fitting the recent leading edge, it makes sense to use $\mathrm{FOM_S}$ for quantifying conversion-efficiency trends and efficiency-speed tradeoffs. In absence of an acceptable figure-of-merit, our previous analysis [1] used 3D fitting to extract the progress rate in conversion efficiency. Using $\mathrm{FOM_S}$, we can now look at the data in two dimensions, which allows us, among other things, to plot efficiency versus speed.

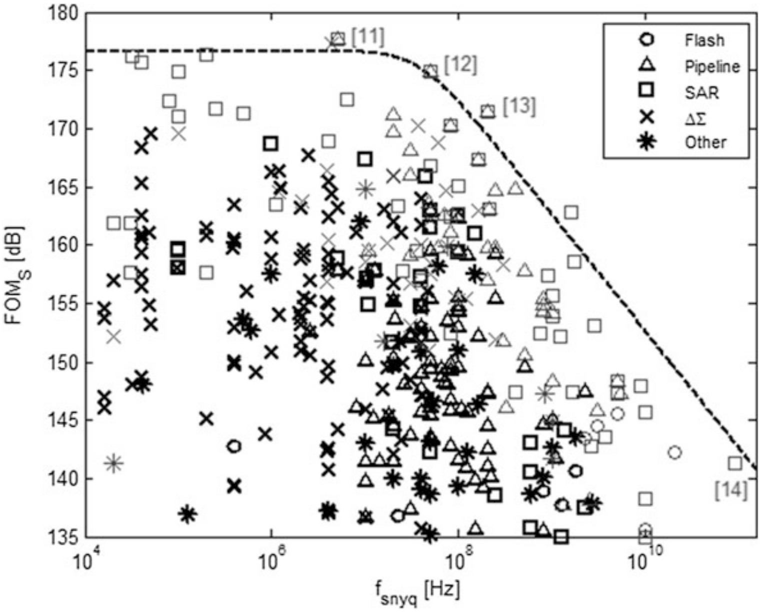


Fig. 4.3 FOM_S versus Nyquist sampling rate. The *gray markers* indicate data reported after 2010

From the plot in Fig. 4.3, we make the following observations. First, note that as expected, the achieved FOM_S is highest for low conversion rates. At frequencies between 10 and 100 MHz, the efficiency begins to deteriorate and rolls off with a slope of approximately −10 dB per decade. As discussed in [10], this indicates that power dissipation scales with the square of speed in this regime. Also notice from the plot that the pipelined SAR designs [11–13] set the peak performance near the corner. The time-interleaved SAR design of [14] marks the rightmost point in this chart and, interestingly, lies almost exactly on the −10 dB/decade roll-off of the drawn envelope. The envelope is constructed by taking the average of the top five data points to define the horizontal asymptote, and the average of the top five designs along a −10 dB roll-off to define the location of that asymptote.

As we can see from Fig. 4.3, the contributions of the past five years have pushed the asymptotes up and to the right. It is interesting to quantify the rate at which this movement occurs. This is done for the location of the low-frequency (LF) asymptote in Fig. 4.4. We observe that the improvements have followed a steady pace with minor variations from year to year (likely due to the finite sample size of the data). Interestingly, the overall progress rate comes out almost exactly to 1 dB per year (or doubling of power efficiency every 3 years). In this context, it is interesting to re-visit the fundamental-limit discussion presented in [1]. There, we noted that a useful bound on conversion energy is given by [15, 16]:

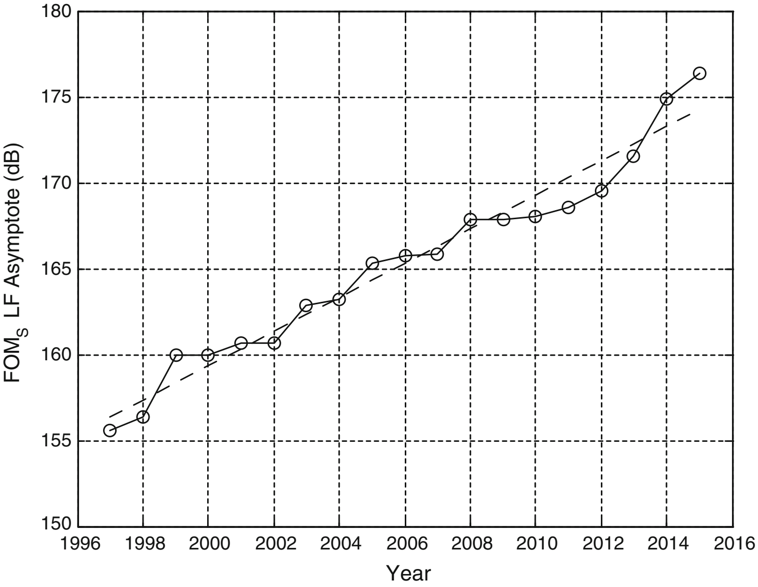


Fig. 4.4 FOM$_S$ trend (low-frequency asymptote)

$$\left(\frac{P}{f_{snyq}}\right)_{min} = 8\ \text{kT} \times SNR \tag{4.2}$$

Approximating SNDR ≅ SNR and inserting into (4.1) gives (assuming room temperature):

$$FOM_{S,max} = SNR(\text{dB}) + 10\ \log\left(\frac{1}{16\ \text{kT} \times SNR}\right) = -10\ \log(16\ \text{kT}) = 192\ \text{dB} \tag{4.3}$$

Since this bound includes only the energy to drive a sampler using an ideal (class-B) amplifier, it is clear that we will likely never reach this number. A more practical limit may be 186 dB, which would be reached in about ten years, assuming that we can maintain the 1 dB per year progress rate.

To extract the rate, at which the high-frequency asymptote of Fig. 4.3 moves to the right, we use FOM$_S$ = 150 dB as an arbitrary reference point and measure (for each year) up to which frequency this level of efficiency is maintained. This yields the plot of Fig. 4.5, from which we observe doubling every 1.8 years, or 10× every 5.9 years (1.7 dB per year). These numbers quantify the rate of power-efficiency improvement for high-speed designs. Since the low- and high-frequency asymptotes shift at different rates, the corner shifts to the right over time. While it was located at about 1.4 MHz in 1997, the corner now occurs at about 42 MHz.

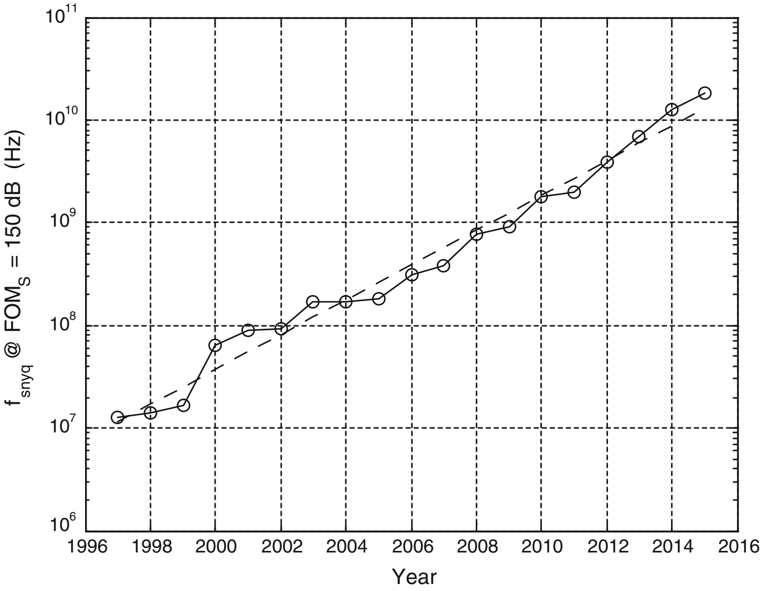


Fig. 4.5 FOM_S trend (high-frequency asymptote)

4.3 Trends in Nyquist A/D Converters

An interesting consequence of the relentless optimization and improvements seen above is the increasing competition among ADC architectures. While it was relatively straightforward to make architectural decisions in the past, today's ADC designer is confronted with an overlapping design space offering multiple solutions that are difficult to differentiate in their suitability. For example, the design space for pipelined ADCs has been encroached by time-interleaved SAR converters. Similarly, wideband delta-sigma converters such as [17] now offer bandwidths that were previously only achievable with Nyquist converters (see also Sect. 4.4).

Figure 4.6 gives an indication on architectural trends. As we had already noted in [1], the SAR architecture continues to be actively researched and conforms with the general trend toward "minimalistic," opamp-less Nyquist ADC architectures. In order to extract high-speed from the SAR topology, time interleaving is typically needed. This explains in part an up-tick in the number of reported designs that use time interleaving, illustrated in Fig. 4.7. More generally, this trend is of course also supported by the increasing integration density available in silicon, which has also enabled multi-core microprocessors.

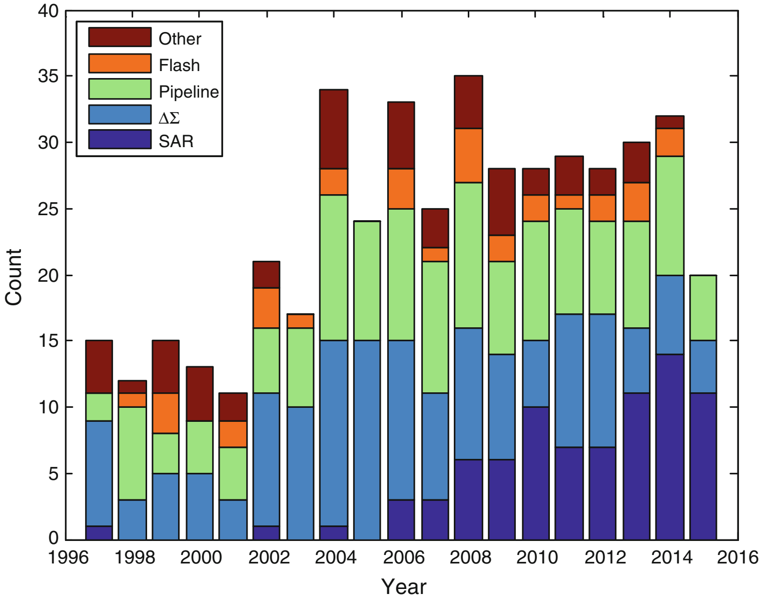


Fig. 4.6 Architectures of published ADCs (ISSCC 1997–2015 and VLSI circuit symposium 1997–2014)

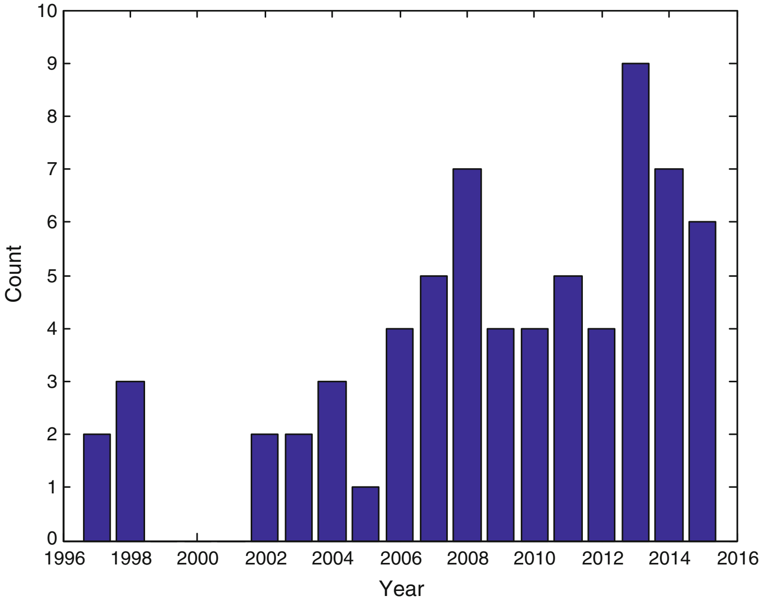


Fig. 4.7 Number of reported time-interleaved ADCs (ISSCC 1997–2015 and VLSI circuit symposium 1997–2014)

4.3.1 SAR ADCs

The SAR ADCs published in recent years show great versatility and range from ultra-low power to ultra-high speed designs (using time interleaving). To see this, contrast the 10-bit 200 kS/s converter of [18] with the 8-bit 90 GS/s part of [14]; both use a very similar circuitry in their converter core. Somewhere in between, we see 10-bit 2.6 GS/s time-interleaved SAR ADCs that can digitize the entire cable TV spectrum [19], as well as highly efficient 100 MS/s, 11-ENOB converters [20] that meet the demands of typical wireless receivers. While much of the progress in SAR converters is enabled by technology scaling, there have been a number of important circuit and architecture innovations as well. These include the combination of SAR conversion with pipelining [21] and the use of dynamic residue amplification in such hybrid topologies [20]. Other recent advancements include the judicious use of redundancy and DAC replica timing [22], majority voting for noise reduction [23], as well as integrated buffering to ease the input drive requirements [24].

4.3.2 Pipelined ADCs

Challenged by the impressive energy efficiency and scaling robustness of SAR converters, the designers of pipelined ADCs have continued their search for "opamp-less" residue amplification techniques. We have seen intriguing innovations in fully-dynamic amplification [25], ring-amplifier-based amplification [26, 27], comparator-based amplification [28], as well as bucket-brigade processing [29]. These and other approaches have helped in keeping the power dissipation of pipelined ADCs competitive for low to moderate sampling rates. Architecturally, the work of [30] reported an intriguing modification to the typical pipeline by splitting the amplifier into a coarse and fine path. This change extends the available settling time in each stage and may prove to be a valuable concept going forward. In the context of high-speed conversion for wireless infrastructure, pipelined ADCs are still the only topology that can meet the stringent application requirements. With proper calibration, we have seen that the pipelined architecture can be pushed to 1 GS/s at 14 bits [7]; a performance level that is hard (if not impossible) to reach with any other topology.

4.3.3 Flash ADCs

Flash ADCs have regained some interest due to the imminent shift from PAM2 to PAM4 signaling in high-speed data links. The time-interleaved flash design of [31] operates at 10.3 GS/s and thereby enables a multi-standard transceiver. As shown in the 32-nm SOI design of [32], the speed can even be extended to 20 GS/s while maintaining outstanding power efficiency. Key to maintaining high efficiency in flash ADCs is to identify a proper offset calibration/mitigation scheme and to minimize the circuit complexity as much as possible. In that vein, the design of [33] introduced a technique that generates extra decision levels using dynamic interpolation at the comparators' regenerative nodes. These and other innovations are strongly linked to the unprecedented speed and integration density that is now at the disposal of the designers. In terms of concept innovation, the approach described in [34] points toward an intriguing new direction. Instead of designing flash ADCs with near perfect thresholds, this work proposes to adaptively control the decision levels to minimize the system's bit error rate, which is the ultimate specification of interest. Broadly speaking, this approach also falls into the categories of digitally assisted and analog-to-information conversion, discussed in more detail in Sect. 4.5.

4.3.4 Digitally Assisted Design

At the front of digitally assisted design, we have seen a variety of ideas applied to all of the above architectures. Perhaps the most complex and sophisticated scheme was implemented in the time-interleaved pipeline ADC of [35], which leverages two million logic gates to reach the unprecedented performance level of 14 bits at 2.5 GS/s. The digital logic is used to correct a variety of analog imperfections including dynamic sampling nonlinearity and signal-dependent self-heating. Similarly, digital equalization concepts are used in [36] to alleviate the residue-amplifier speed requirements and achieve 5.4 GS/s with only two interleaved slices. In the context of background calibration for pipelined ADCs, another noteworthy development was the introduction of algorithms with short convergence times [37]. Another area where digital assist has been pushed to new levels is in the correction of time interleaving artifacts. The time-interleaved SAR converter of [38] uses fully digital compensation of timing skew, which was previously thought to be prohibitively complex. In flash ADC design, digital assist was shown to be effective in reducing the comparator offset trim range by employing a fault-tolerant encoder [31], leading to significant savings in complexity and power. Another area, where digital assisted techniques have found their use, is in emerging topologies, such as VCO-based Nyquist converters [39, 40]. Here, digital calibration is not only an add-on, but needed to make these approaches practical.

4.4 Trends in Delta-Sigma A/D Converters

Delta-Sigma ADCs continued to follow an exploratory focus in both industrial and academic research activities. As a result, numerous findings, answers to previously open questions, and advances over the state-of-the-art were presented at conferences and published in journals. This trend continues to-date.

In the following, we provide an overview and summary of these advancements. Moreover, we reflect on them with regard to the predictions made on the development of Delta-Sigma ADCs back in 2010. Finally, based on recent and actual trends in the design of Delta-Sigma ADCs, we make predictions on future trends. By doing so, the most recent version of the survey provided in [2], extended by Delta-Sigma ADCs published in *IEEE Journal of Solid-State Circuits* from 2010 to 2015, serves as a data base. A detailed survey on Delta-Sigma ADCs covering further conferences and journals can be found in [41].

In accordance with [1], we consider in the following the major sub-blocks of a Delta-Sigma ADC, i.e., the loop filter, the quantizer, and the DAC, in order to present advancements and to discuss trends. Moreover, we stick to the $\mathrm{FOM_W}$ in this subchapter in order to facilitate a comparison with our results presented in [1].

4.4.1 Loop Filter

As outlined in [1], the loop filter of a Delta-Sigma ADC is categorized based on several criteria. Amongst others, the time domain it was designed for, i.e., continuous-time (CT) or discrete-time (DT). In 2010, it was observed that Delta-Sigma ADCs using a CT loop filter seemed to become the vehicle for high-speed implementations, a trend that was predicted to continue. Considering Fig. 4.8, which is an update of Fig. 4.13 in [1] with CT and DT designs published later than 2010, this prediction proved to be true over the past five years: all high-speed implementations with bandwidths larger than 20 MHz were implemented using a CT loop filter. Almost all have achieved SNDRs and FOMs comparable to those of the latest DT implementations while the lowest and thus the best CT FOM for a bandwidth larger than 20 MHz outperforms the lowest DT one by a factor of six. In general, a trend to lower FOMs is clearly visible. However, the minimum CT FOM improved only slightly from 40 to 30 fJ/conversion-step over the past five years.

Considering recent designs for frequency bands up to 20 MHz, CT loop filters were mostly used for the implementations as well. Overall, in comparison with DT Delta-Sigma modulators, nearly twice the number of CT implementations was published from January 2011 to March 2015. An overview of the quantitative distribution of the published architectures is given in Fig. 4.9. As can be seen, a switched-capacitor technique and thus DT circuitry were preferably used only for the implementation of MASH modulators. The better matching between the analog loop filter and the digital cancellation filters may account for this preference. On the other hand, this argumentation implies that, contrary to the prediction made in [1], less research was performed on digitally assisted circuits in order to overcome matching issues in CT multi-stage noise-shaping (MASH) architectures.

Interestingly, only one sturdy multi-stage noise-shaping (SMASH) modulator was among the recent MASH implementations [42]. This fact may be due to the guess ventured in [1], i.e., any SMASH modulator exhibits a feedback loop whose

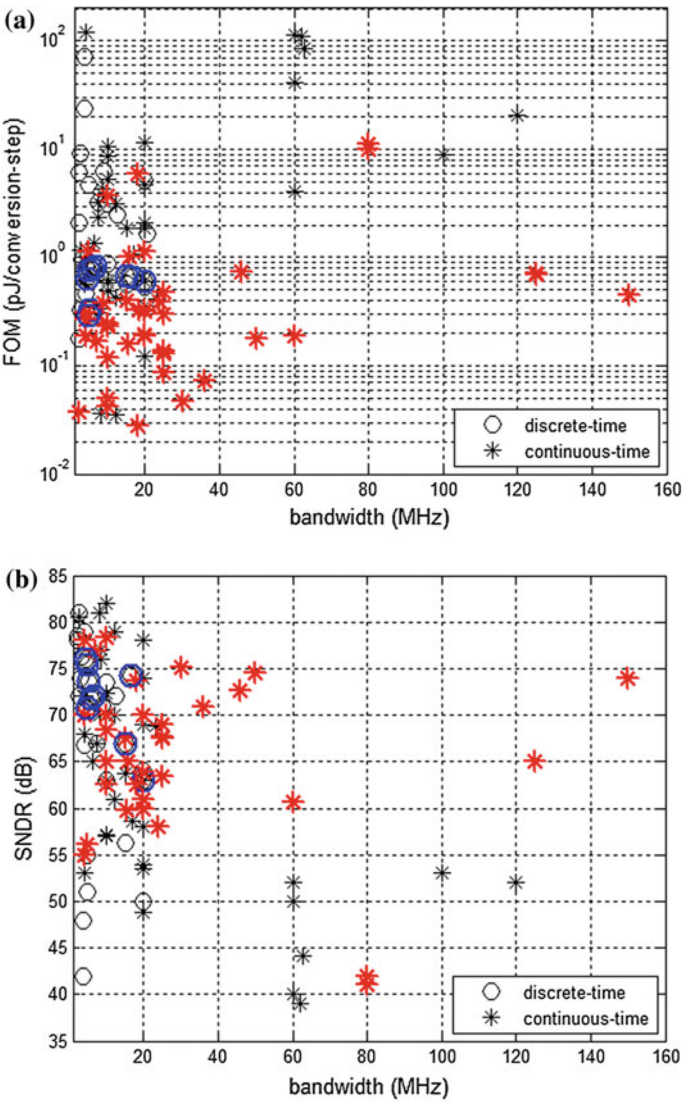


Fig. 4.8 Comparison of FOM and SNDR between DT and CT Delta-Sigma modulators exhibiting a bandwidth larger than 2 MHz. DT and CT designs published later than 2010 are highlighted in *blue* and *red*, respectively. FOM (**a**) and SNDR (**b**) versus bandwidth

order is equal to the sum of the orders of the single-stage modulators used in the cascade. Consequently, they face the stability issues of their equivalent single-stage modulator while requiring at least one additional quantizer for the implementation.

Another criterion for categorizing Delta-Sigma ADCs is based on the architecture of the loop filter, i.e., CIFB, CRFB, CIFF, CRFF, or any mixture of them [4]. In 2010, it was predicted that the CIFF architecture will become the preferred

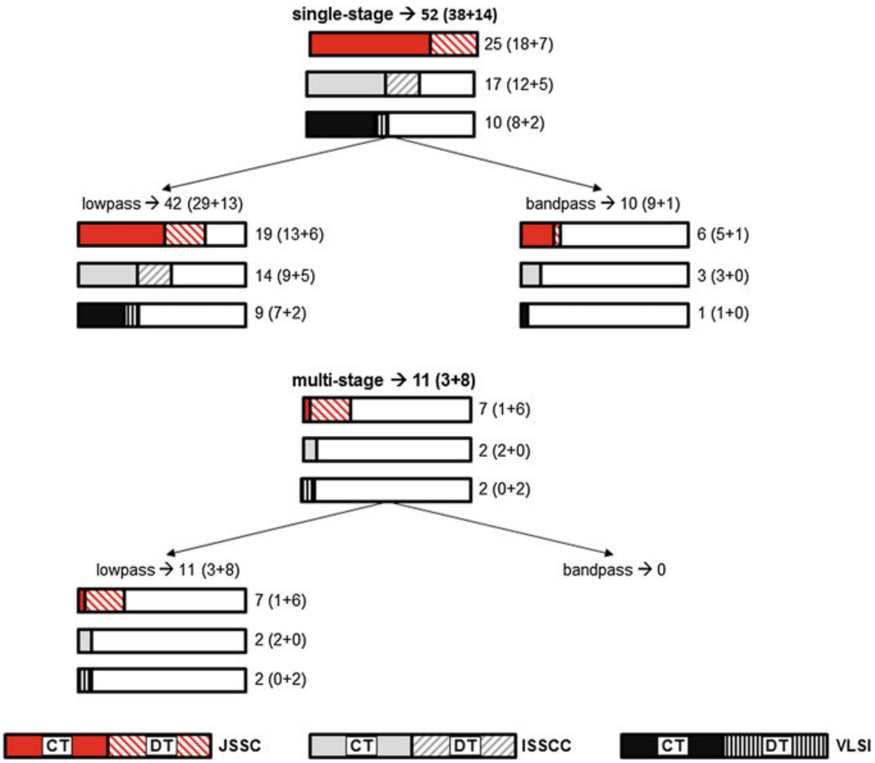


Fig. 4.9 Overview on delta-sigma modulators published between January 2011 and March 2015

architecture for implementations in the latest technology nodes. Despite the peaking of the STF and less anti-aliasing filtering, the signal swings inside a CIFF filter are much lower for input signals close to full scale in comparison to a CIFB filter; a characteristic that makes them very attractive for implementations using low supply voltages. Indeed, all recent designs with supply voltages around 0.5 V used a CIFF loop filter [43–45].

In general, all types of loop filters still enjoy great popularity today, even for supply voltages as low as 1 V. When it comes to the implementation of Delta-Sigma ADCs for telecommunication applications, e.g., wireless receivers, CIFB and CRFB represent the first choice. Not only do they provide better anti-aliasing filtering, but also a flat STF, which is considered a key characteristic with respect to blockers and interferers. However, in 2004, it was proposed to embed a first-order low-pass filter within a fourth-order CIFF loop filter of a Delta-Sigma ADC in order to reduce the peaking of the STF [46]. Both the robustness against interferers and blockers and the anti-aliasing filtering were thus improved. This approach was further pursued recently. In [47], a second-order Butterworth low-pass filter was embedded within a fourth-order CIFF loop filter. In comparison to an implementation with an explicit up-front filter, 25 % less power consumption and 20 % less area were thus achieved. In another approach [48], a second-order CIFB Delta-Sigma ADC was merged with a third-order Chebyshev channel-select filter. As a result, the order of noise-shaping increased from two to five while the low-frequency STF was determined by the Chebyshev channel-select filter. Further research on both approaches may pave the way for CIFF loop filters to become the dominant architecture in the near future, even in telecommunication systems.

Finally, the lower FOMs seen in Fig. 4.9 are in part due to advancements in the implementation of the loop filter, i.e., minimalistic and digitally assisted architectures. Using inverter-based opamps, minimalistic loop filters for Delta-Sigma modulators started to emerge in 2007 [49–52]. Further approaches were presented recently, which allow for implementing second- and third-order loop filters using only a single amplifier [53–58]. These concepts, amongst others, were applied to achieve FOMs of 50 fJ/conversion-step and 41 fJ/conversion-step in a bandwidth of 10 MHz [55, 57].

In [57], another path-breaking technique was presented, which may shape the future of loop filters: integrator loss compensation. Considering an active RC integrator, this technique virtually boosts the DC gain of the amplifier by inserting a negative replica of the resistor R from the input of the amplifier to ground. Implementing high DC-gain amplifiers by means of cascading low DC-gain stages thus becomes obsolete. At present, cascading represents a very popular yet power consuming approach to tackling the reduced intrinsic gain per transistor and the low supply voltages of modern technology nodes that limit the number of stacked transistors and thus the efficiency of cascoding.

4.4.2 *Quantizer*

In [1], two concepts for the implementation of a multi-bit quantizer were considered that comply with the benefits of scaled CMOS technologies for digital applications: the VCO-based quantizer and the time-encoding quantizer. It was predicted that they may become favorite architectures for the implementation of a multi-bit quantizer in the near future, thus replacing power consuming multi-bit flash ADCs. In the following, the developments and advancements of these quantizers are summarized and analyzed.

4.4.2.1 Voltage-Controlled Oscillator-Based Quantizer

Voltage-controlled oscillator-based quantizers highly comply with technology scaling since they consist of digital circuit blocks, e.g., flip-flops and standard logic cells, which provide signal levels equal to the supply voltage. Sampling the phase instead of the frequency, they provide first-order noise shaping to the quantization error by embedding a digital differentiator in order to reconvert the phase to frequency. Without feedback, a VCO-based quantizer thus achieves first-order noise

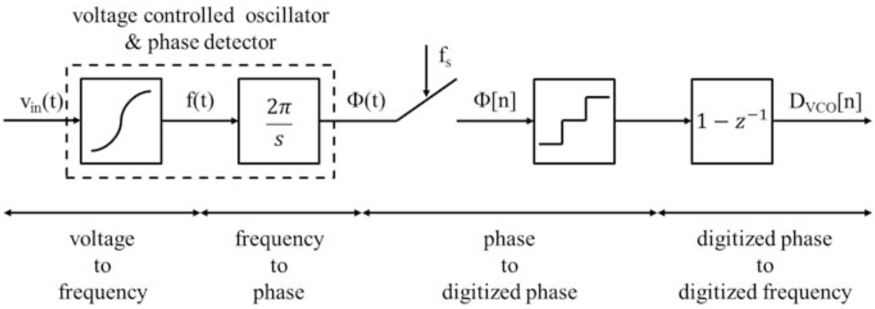


Fig. 4.10 Block diagram of a VCO-based quantizer

shaping by cascading a VCO, a phase detector and a digital differentiator as illustrated in Fig. 4.10. In spite of this fundamental distinction to a Delta-Sigma modulator, where feedback is applied in order to achieve noise shaping, they are often called a first-order Delta-Sigma modulator. Strictly speaking, they belong to the class of noise-shaping ADCs while not representing a Delta-Sigma modulator.

The intrinsic resolution of a VCO-based quantizer is determined by distortion due the non-linear voltage-to-frequency characteristic and phase noise, which are not subject to the intrinsic noise shaping. Thus, a pseudo-differential architecture is used quite frequently in order to reduce even-order harmonics and to improve the SNDR by 3 dB since the signal power quadruples while the noise power doubles. Further means were developed over the past five years in order to reduce the impact of these non-idealities on the resolution of a VCO-based quantizer, e.g., digital background or foreground calibration techniques. While these concepts were developed for improving the performance of stand-alone VCO-based quantizers, other approaches pursued embedding a VCO based quantizer in a Delta-Sigma modulator or using it as a later stage in MASH architectures. Embedding a VCO based quantizer in a Delta-Sigma modulator as performed in [59, 60] or [61] provides two advantages. First, using an N-th order loop filter, an (N + 1)-order noise shaping of the quantization error is achieved since one order is provided by the VCO. Second, distortion and phase noise of the VCO are suppressed by the N-th order loop filter, if referred to the input of the modulator. Using a VCO-based quantizer as a later stage of a MASH architecture as proposed in [62] results in less signal swing at the input of the VCO since later stages in a cascade only have to process the quantization error of the previous stage. Thus, the almost linear range of the non-linear voltage-to-frequency transfer characteristic is used whereby less distortion is induced.

A summary of recent architectures and achieved performances, including [59, 60] which already were considered in [1], is given in Table 4.1, sorted by their publication date. In particular, the advancements achieved in [61] in comparison with [59, 60] should be highlighted: Using a two-step quantizer consisting of a 4-bit flash and a 4-bit VCO, the design achieved an SNR which approximately equals the SNR of an ideal and thus unscaled 2nd-order Delta-Sigma ADC with an 8-bit quantizer. An 8-bit flash ADC, however, is considered to be hardly feasible using a

Table 4.1 Survey on and comparison of recent VCO-based ADCs

	Straayer [59] JSSC 04/2008	Park [60] JSSC 12/2009	Daniels [63] VLSI 2010	Taylor [64] JSSC 12/2010	Asl [62] CICC 2011	Rao [65] VLSI 2011	Reddy [61] JSSC 12/2012	Taylor [66] JSSC 04/2013	Rao [67] JSSC 04/2014
ADC architecture	2nd-order DSMw. VCO quantizer → 3rd-order noise shaping	3rd-order DSMw. VCO quantizer → 4th-order noise shaping	Pseudo differential VCO	Pseudo differential VCO	MASH 1st: lst-order DSM 2nd: VCO ADC → 2nd-order noise shaping	Pseudo differential VCO	lst-order DSMw. 2-step quantizer (4bflash—VCO)	Configurable pseudo differential VCO	pseudo differential VCO
Resolution of VCO (bits)	5	4	4	4	3	4	4	4	4
Calibration	N/A	N/A	Foreground	Background	N/A	Two-level modulation	N/A	Background	Background
Technology (nm)	130	130	65	65	130	90	90	65	90
Supply voltage (V)	1.2	1.5	N/A	2.5 (V/l conv.) 1.2	1.3	N/A	1.4(A) 1.0(D)	0.9–1.2	1.2(A) 1.0 (D)
SNR(dB)	86	81.2	N/A	70	77.3	N/A	83	70–76	75.4
SNDR(dB)	72	78.1	64	69	77	59.1	78.3	69–75	73–74.7
SFDR(dB)	N/A	N/A	79	N/A	91.6	71	78.3	77–83	N/A
Bandwidth (MHz)	10	20	30	18	4	8	10	5.08–37.5	5
Sampling frequency (MHz)	950	900	300	1152	100 (DSM) 1200 (VCO)	640	600	1300–2400	640
Power (mW)	40	87	11.4	17	6.1(A) 7.7(D)	4.3	6.5(A) 9.5(D)	11.5–39	4.1
Area (mm^2)	0.42	0.45	0.02	0.07	0.7	0.1	0.36	0.075	0.16
FOM (fl/cornv.-step)	500	330	150	159	296	366	120	123–305	92–112

supply voltage of 1.4 V since the LSB becomes as small as 5.5 mV. Future research may focus on merging a minimalistic higher-order loop filter and a VCO-based quantizer in order to design a very power and area efficient Delta-Sigma modulator while suppressing distortion of the VCO by the loop filter.

4.4.2.2 Time-Encoding Quantizer

Time-encoding quantizers use a single-bit PWM quantizer whose power of the quantization noise is mostly located at a defined limit cycle outside the signal band. By means of a sinc-decimation filter whose zeros are placed accordingly, this quantization noise is removed [68]. The residual power of the quantization noise within the signal band thus resembles the power of the quantization noise of a multi-bit quantizer although a single-bit quantizer is used. Obviously, the advantage of this approach is its compatibility with scaled CMOS technologies for digital applications, since only a single comparator is needed.

Recently, advancements were achieved in the generation of the limit cycle. Applying describing-function theory, it was shown in [69] that a stable limit cycle with well controlled amplitude and frequency can be generated using a flip-flop instead of an inverter based programmable delay-line in the feedback loop of the time-encoding quantizer. Moreover, an FIR DAC was used in the feedback loop of the Delta-Sigma modulator in order to suppress the limit cycle in the loop, thus reducing the slewing requirement of the first opamp. In [70], the active integrator in the feedback loop of the time-encoding quantizer was replaced by a passive low-pass filter and an amplifying DAC in order to lower the power consumption to 7mW and the area to 0.08 mm^2. Using a 65 nm technology, an SNDR of 61 dB was achieved in a bandwidth of 20 MHz which results in a FOM of 191 fJ/conversion-step.

Another concept for implementing a time-encoding quantizer was presented in [71] which consist of a comparator-based PWM modulator followed by a time-to-digital converter. The publication represents an extended version of [72], which achieved an SNDR of 60 dB in a bandwidth of 20 MHz for a power consumption of 10.5 mW using a 65 nm technology. Similar results were presented almost two and a half year later in [70] while the FOM and the area were improved by 40 and 50 %, respectively; a quite noticeable advancement.

In all considered designs, multi-stage opamp-based loop filters were applied. As for VCO-based Delta-Sigma ADCs, merging the concept of time-encoding quantizer and minimalistic/digitally assisted loop filter may be the next step toward reducing the power consumption and area of these low-voltage compliant ADCs.

4.4.3 DAC

In 2003, it was proposed to use a single-bit quantizer with a multi-tap finite impulse response (FIR) DAC in the feedback path of the Delta-Sigma modulator [73]. Since only a single comparator is needed for the implementation of the quantizer, a power-and area-efficient design thus becomes feasible while the feedback signal resembles the signal of a multi-bit DAC. Consequently, a design using an FIR DAC is nearly as robust against clock jitter as its equivalent multi-bit implementation [74]. Moreover, as in the multi-bit case, the dynamic range of the Delta-Sigma modulator increases since the multi-bit feedback signal contains less power compared to the single-bit case. Finally, since the amplitude of a multi-bit signal better matches the amplitude of the input signal, the signal to be processed by the loop filter, in particular by the first integrator, becomes smaller. As a result, less opamp-related distortions are induced.

The approach has received attention recently in order to replace a multi-bit flash quantizer, since its power consumption determines the overall power consumption of today's Delta-Sigma ADCs to a great extent [69, 75–77]. In [77], it was shown that such a design is more power efficient than an equivalent design using a 4-bit quantizer although a higher sampling frequency must be applied in order to achieve the same resolution. Achieving an SNDR of 70.9 dB in a bandwidth of 36 MHz for a power consumption of 15 mW, the FOM equal to 73 fJ/conversion-step is among the lowest FOMs reported to date for Delta-Sigma ADCs.

4.4.4 Conclusion on Delta-Sigma A/D Converters

Concluding our overview on and summary of advancements of Delta-Sigma modulators: Many innovative cooks all over the world rely on the cooking recipe *Delta-Sigma ADC,* which dictates a loop filter, a quantizer and feedback. By trying ever new ingredients, they are competing with each other for being the first serving the meal with the ultimate taste according to this particular recipe. Naturally, tastes differ, which is why selections of the best recipes are provided as a subchapter in one of the many cookery books entitled *Oversampled Analog-to-Digital Converters* or the like. However, the time may be ripe for focusing on and writing a fundamentally different and thus never before seen cookery book. This book may be entitled *Analog-to-Information Converters*. A glimpse on its first pages and thus what it will be about one day is provided next.

4.5 Analog-to-Information Converters

The term "analog-to-information" was first coined in 2008 [78], and it was discussed in the context of a specific technique called compressed sensing (CS) [79, 80]. From CS theory, it follows that one can recover certain signals (which are "sparse" in some domain), using much fewer samples than required per the Shannon-Nyquist sampling theorem. In recent years, this theory has been translated down to practical realizations, and we are seeing the first few hardware demos

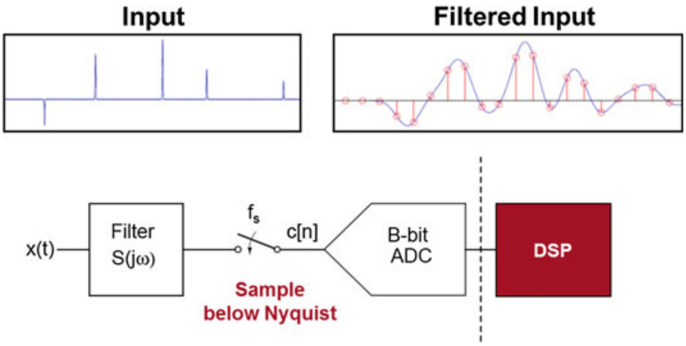


Fig. 4.11 FRI sampling of a pulse train signal

ranging from bio interfaces [81] to CMOS imagers [82] and radios [83]. Due to the requirement of "sparsity," CS is not suitable for arbitrary signals and one must carefully consider the impact of circuit impairments in practical realizations [84].

In the meantime, similar concepts that target sub-Nyquist signal acquisition have emerged. These include Xampling, which uses aliasing in conjunction with CS-like reconstruction in the digital domain [85]. Another approach is finite rate of innovation (FRI) sampling [86, 87], which models a signal in terms of its number of degrees of freedom per unit of time (corresponding to the rate of innovation). This idea is illustrated in Fig. 4.11 using a pulse-train input, which could be viewed as a basic model of an ultrasound or radar signal. The shown waveform has 5 pulses that can be described by 5 arrival times and 5 amplitudes. So, according to FRI theory, there are 10 degrees of freedom and 10 samples should suffice to reconstruct this signal in the digital domain. However, if we take the typical approach of uniform sampling, a very high sampling rate is needed to preserve the information according to Shannon-Nyquist. Namely, the sampling rate must be twice as high as the highest input frequency, which is very large due to the narrow pulses. The solution offered within the FRI framework is to pre-filter the signal before sampling and sample the signal at a rate commensurate with the low-bandwidth content of the smoothed signal ("filtered output" in Fig. 4.11). Xampling and FRI approaches are currently being taken toward practical hardware realizations and promise to offer significant benefits in various applications. In imaging and video, compressed sensing and FRI sampling are particularly powerful directions and receive further treatment in Chaps. 12 and 14.

Another class of analog-to-information interfaces is emerging in the context of feature extraction and classification of patterns that are buried in analog waveforms. Building on similar ideas that have already been explored in the imaging community [88, 89], it is conceivable that new forms of specialized "feature-extraction" A/D converters will emerge. A very recent example is a 6 µW acoustic sensor front-end, featuring analog feature extraction and mixed-signal embedded classification [90]. Such interfaces will undoubtedly gain in popularity as we steer toward the "internet of everything," in which massive amounts of sensor data will force us to feature-extract, classify and interpret signals as close as possible to the sensor itself.

4.6 Conclusions

From Fig. 4.4, it was observed that the overall progress rate of the low-frequency $\mathrm{FOM_S}$ asymptote is approximately 1 dB per year. Based on this rate, the practical limit $\mathrm{FOM_{S,max}} = 186$ dB according to (1) will be reached in ten years, i.e., in 2025. At the same time, the high-frequency asymptote at the (arbitrary) reference level $\mathrm{FOM_S} = 150$ dB in Fig. 4.3 would increase from 20 GHz to 1.28 THz, i.e., by about a factor 2^6 based on the progress rate of ×2 every 1.8 years (see Fig. 4.5). Thus, the intersection point of the two asymptotes will increase from 42 to about 320 MHz.

Accounting for a safety margin as large as three, this trend implies that by 2025, ADCs for Nyquist sampling rates up to 100 MHz cannot be further optimized in terms of $\mathrm{FOM_S}$. Despite some inherent uncertainty in these numbers, this result poses several questions the ADC community must face and deal with in the near future. Will the optimization of conventional ADCs indeed run out of steam, or will new applications emerge that fuel new demand for higher bandwidths, thus keeping the wheels of optimization turning? Will we stick to a general FOM or define application-specific FOMs, which will then provide opportunities for application-specific optimization? Will we overcome the limitation of clock jitter by means of new approaches or architectures that avoid clock-driven sampling? How will the designers of analog-to-information converters respond to the challenges?

In summary, we conclude that the development and progress of ADCs in recent years was mostly driven by the optimization of figures of merit, and remarkable improvements have been recorded over time. However, in the near future, this trend will have to come to a halt, since we are approaching practical limits that are difficult, if not impossible, to surpass. As in the case of the MOS transistor, the broad question on “What’s next?” is looming on the horizon of data converter research.

References

1. Keller, M., Murmann, B., Manoli, Y.: Analog-digital interfaces. In: Hoefflinger, B. (ed.) CHIPS 2020—A Guide to the Future of Nanoelectronics, pp. 95–130. Springer, Berlin (2012)
2. Murmann, B.: ADC performance survey 1997–2015. Available http://web.stanford.edu/∼murmann/adcsurvey.html
3. Walden, R.H.: Analog-to-digital converter survey and analysis. IEEE J. Sel. Areas Commun. **17**(4), 539–550 (1999)
4. Schreier, R., Temes, G.C.: Understanding Delta-Sigma Data Converters. Wiley, New York (2005)
5. Ali, A.M.A., et al.: A 16-bit 250-MS/s IF sampling pipelined ADC with background calibration. IEEE J. Solid-State Circuits **45**(12), 2602–2612 (2010)
6. Ali, A.M.A.: A 14-bit 125 MS/s IF/RF sampling pipelined ADC with 100 dB SFDR and 50 fs Jitter. IEEE J. Solid-State Circuits **41**(8), 1846–1855 (2006)
7. Ali, A.M.A., et al.: A 14 Bit 1 GS/s RF sampling pipelined ADC with background calibration. IEEE J. Solid-State Circuits **49**(12), 2857–2867 (2014)

8. Greshishchev, Y.M., et al.: A 40GS/s 6b ADC in 65 nm CMOS. In: IEEE International Solid-State Circuits Conference—Digest of Technical Papers, pp. 390–391 (2010)
9. Elliott, M., Murmann, B.: High-performance pipelined ADCs for wireless infrastructure systems. In: Manganaro, G., Leenaerts, D.M.W. (eds.) Advances in Analog and RF IC Design for Wireless Communication Systems. Elsevier, Amsterdam (2013)
10. Murmann, B.: Energy limits in A/D converters. In: 2013 IEEE Faible Tension Faible Consommation, pp. 1–4 (2013)
11. Bannon, A., et al.: An 18 b 5 MS/s SAR ADC with 100.2 dB dynamic range. In: IEEE Symposium of VLSI Circuits—Digest Technical Papers, pp. 1–2 (2014)
12. Lim, Y., Flynn, M.P.: A 1 mW 71.5 dB SNDR 50 MS/s 13 b fully differential ring amplifier based SAR-assisted pipeline ADC in 65 nm CMOS. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 458–459 (2015)
13. Verbruggen, B., et al.: A 70 dB SNDR 200MS/s 2.3 mW dynamic pipelined SAR ADC in 28 nm digital CMOS. In: IEEE Symposium on VLSI Circuits—Digest Technical Papers, pp. 1–2 (2014)
14. Kull, L. et al.: A 90 GS/s 8 b 667 mW 64x interleaved SAR ADC in 32 nm Digital SOI CMOS. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 378–379 (2014)
15. Vittoz, E.A.: Future of analog in the VLSI environment. In: Proceedings of IEEE International Symposium on Circuits System, pp. 1372–1375 (1990)
16. Hosticka, B.J.: Performance comparison of analog and digital circuits. Proc. IEEE **73**(1), 25–29 (1985)
17. Bolatkale, M., et al.: A 4 GHz CT ΔΣ ADC with 70 dB DR and −74 dBFS THD in 125 MHz BW. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 470–472 (2011)
18. Tai, H.-Y., et al.: A 0.85 fJ/conversion-step 10 b 200 kS/s subranging SAR ADC in 40 nm CMOS. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 196–197 (2014)
19. Doris, K., et al.: A 480 mW 2.6 GS/s 10 b 65 nm CMOS time-interleaved ADC with 48.5 dB SNDR up to Nyquist. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 180–182 (2011)
20. van der Goes, F., et al.: 11.4 A 1.5 mW 68 dB SNDR 80 MS/s 2× interleaved SAR-assisted pipelined ADC in 28 nm CMOS. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 200–201 (2014)
21. Lee, C.C., Flynn, M.P.: A SAR-assisted two-stage pipeline ADC. IEEE J. Solid-State Circuits **46**(4), 859–869 (2011)
22. Kapusta, R., Decker, S., Ibaragi, E.: A 14 b 80 MS/s SAR ADC with 73.6 dB SNDR in 65 nm CMOS. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 472–473 (2013)
23. Harpe, P., Cantatore, E., van Roermund, A.: A 10 b/12 b 40 kS/s SAR ADC with data-driven noise reduction achieving up to 10.1 b ENOB at 2.2 fJ/conversion-step. IEEE J. Solid-State Circuits **48**(12), 3011–3018 (2013)
24. Kraemer, M., et al.: A 14 b 35 MS/s SAR ADC achieving 75 dB SNDR and 99 dB SFDR with loop-embedded input buffer in 40 nm CMOS. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 284–285 (2015)
25. Verbruggen, B., Iriguchi, M., Craninckx, J.: A 1.7 mW 11 b 250 MS/s 2× interleaved fully dynamic pipelined SAR ADC in 40 nm digital CMOS. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 466–468 (2012)
26. Hershberg, B., et al.: Ring amplifiers for switched-capacitor circuits. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 460–462 (2012)
27. Lim, Y., Flynn, M.P.: 11.5 A 100 MS/s 10.5 b 2.46 mW comparator-less pipeline ADC using self-biased ring amplifiers. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 202–203 (2014)

28. Chang, D.-Y., et al.: 11.6 A 21 mW 15 b 48 MS/s zero-crossing pipeline ADC in 0.13 µm CMOS with 74 dB SNDR. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 204–205 (2014)
29. Dolev, N., Kramer, M., Murmann, B.: A 12-bit, 200-MS/s, 11.5-mW pipeline ADC using a pulsed bucket brigade front-end. In: IEEE Symposium VLSI Circuits—Digest Technical Papers, pp. 98–99 (2013)
30. Chai, Y., Wu, J.-T.: A 5.37 mW 10 b 200 MS/s dual-path pipelined ADC. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 462–464 (2012)
31. Verma, S., et al.: A 10.3 GS/s 6 b flash ADC for 10 G Ethernet applications. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 462–463 (2013)
32. Chen, V.H.-C., Pileggi, L.: 22.2 A 69.5 mW 20 GS/s 6 b time-interleaved ADC with embedded time-to-digital calibration in 32 nm CMOS SOI. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 380–381 (2014)
33. Shu, Y.-S.: A 6 b 3 GS/s 11 mW fully dynamic flash ADC in 40 nm CMOS with reduced number of comparators. In: IEEE Symposium on VLSI Circuits—Digest Technical Papers, pp. 26–27 (2012)
34. Narasimha, R., et al.: BER-optimal analog-to-digital converters for communication links. IEEE Trans. Signal Process. **60**(7), 3683–3691 (2012)
35. Setterberg, B., et al.: A 14 b 2.5 GS/s 8-way-interleaved pipelined ADC with background calibration and digital dynamic linearity correction. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 466–467 (2013)
36. Wu, J., et al.: A 5.4 GS/s 12 b 500 mW pipeline ADC in 28 nm CMOS. In: IEEE Symposium on VLSI Circuits—Digest Technical Papers, pp. 92–93 (2013)
37. Sun, N., Lee, H.-S., Ham, D.: A 2.9-mW 11-b 20-MS/s pipelined ADC with dual-mode-based digital background calibration. In: Proceedings of European Solid-State Circuits Conference, pp. 269–272 (2012)
38. Le Dortz, N., et al.: 22.5 A 1.62 GS/s time-interleaved SAR ADC with digital background mismatch calibration achieving interleaving spurs below 70 dBFS. In: IEEE International Solid-State Circuits Conference—Digital Technical Papers, pp. 386–388 (2014)
39. Rao, S., et al.: A 4.1 mW, 12-bit ENOB, 5 MHz BW, VCO-based ADC with on-chip deterministic digital background calibration in 90 nm CMOS. In: IEEE Symposium on VLSI Circuits—Digest Technical Papers, pp. 68–69 (2013)
40. Taylor, G., Galton, I.: A reconfigurable mostly-digital ΔΣ ADC with a worst-case FOM of 160 dB. In: IEEE Symposium on VLSI Circuits—Digest Technical Papers, pp. 166–167 (2012)
41. de la Rosa, J.M.: CMOS SDMs Survey. Available http://www.imse-cnm.csic.es/~jrosa/CMOS-SDMs-Survey-IMSE-JMdelaRosa.xlsx
42. Yoon, D.-Y., Ho, S., Lee, H.-S.: An 85 dB DR 74.6 dB SNDR 50 MHz BW CT MASH delta-sigma modulator in 28 nm CMOS. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 272–273 (2015)
43. Zhang, J., et al.: A 0.6 V 82 dB 28.6 µW continuous-time audio delta-sigma modulator. IEEE J. Solid-State Circuits **46**(10), 2326–2335 (2011)
44. Michel, F., Steyaert, M.S.: A 250 mV 7.5 µW 61 dB SNDR SC delta-sigma modulator using near-threshold-voltage-biased inverter amplifiers in 130 nm CMOS. IEEE J. Solid-State Circuits **47**(3), 709–721 (2012)
45. Yang, Z., Yao, L., Lian, Y.: A 0.5 V 35 µW 85 dB DR double-sampled delta-sigma modulator for audio applications. IEEE J. Solid-State Circuits **47**(3), 722–735 (2012)
46. Philips, K., et al.: A continuous-time sigma-delta ADC with increased immunity to interferers. IEEE J. Solid-State Circuits **39**(12), 2170–2178 (2004)
47. Rajan, R.S., Pavan, S.: Design techniques for continuous-time delta-sigma modulators with embedded active filtering. IEEE J. Solid-State Circuits **49**(10), 2187–2198 (2014)
48. Andersson, M., et al.: A filtering ΔΣ ADC for LTE and beyond. IEEE J. Solid-State Circuits **49**(7), 1535–1547 (2014)

49. Chae, Y., Han, G.: A low power sigma-delta modulator using class-C inverter. In: IEEE Symposium VLSI Circuits—Digest Technical Papers, pp. 240–241 (2007)
50. Chae, Y., Lee, I., Han, G.: A 0.7 V 36 μW 85 dB-DR audio delta-sigma modulator using Class-C inverter. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 490–491 (2008)
51. van Veldhoven, R.H.M., Rutten, R., Breems, L.J.: An inverter-based hybrid delta-sigma modulator. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 492–493 (2008)
52. Chae, Y., Han, G.: Low voltage, low power, inverter-based switched-capacitor delta-sigma modulator. IEEE J. Solid-State Circuits **44**(2), 458–472 (2009)
53. Perez, A.P., Bonizzoni, E., Maloberti, F.: A 84 dB SNDR 100 kHz bandwidth low-power single op-amp third-order delta-sigma modulator consuming 140 μW. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 478–480 (2011)
54. Chae, H., et al.: A 12 mW low-power continuous-time bandpass delta-sigma modulator with 58 dB SNDR and 24 MHz bandwidth at 200 MHz IF. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 148–150 (2012)
55. Matsukawa, K., et al.: A 10 MHz BW 50 fJ/conv. continuous time delta-sigma modulator with high-order single opamp integrator using optimization-based design method. In: IEEE Symposium on VLSI Circuits—Digital Technical Papers, pp. 160–161 (2012)
56. Christen, T.: A 15bit 140 μW scalable-bandwidth inverter-based Delta-Sigma modulator for a MEMS microphone with digital output. IEEE J. Solid-State Circuits **48**(7), 1605–1614 (2013)
57. Zeller, S., et al.: A 0.039 mm^2 inverter-based 1.82mW 68.6 dB SNDR 10 MHz BW CT Sigma-Delta ADC in 65 nm CMOS using power- and area-efficient design techniques. IEEE J. Solid-State Circuits **49**(7), 1548–1560 (2014)
58. Weng, C.-H., et al.: An 8.5 MHz 67.2 dB SNDR CTDSM with ELD compensation embedded twin-T SAB and circular TDC-based quantizer in 90 nm CMOS. In: IEEE Symposium on VLSI Circuits—Digital Technical Papers, pp. 1–2 (2014)
59. Straayer, M.Z., Perrott, M.H.: A 12-Bit, 10-MHz bandwidth, continuous-time sigma-delta ADC with a 5-Bit, 950-MS/s VCO-based quantizer. IEEE J. Solid-State Circuits **43**(3), 805–814 (2008)
60. Park, M., Perrott, M.H.: A 78 dB SNDR 87 mW 20 MHz bandwidth continuous-time delta-sigma ADC with VCO-based integrator and quantizer implemented in 0.13 μm CMOS. IEEE J. Solid-State Circuits **44**(12), 3344–3358 (2009)
61. Reddy, K., et al.: A 16 mW 78 dB SNDR 10 MHz BW CT delta-sigma ADC using residue-cancelling VCO-based quantizer. IEEE J. Solid-State Circuits **47**(12), 2916–2927 (2012)
62. Asl, S.Z., et al.: A 77 dB SNDR, 4 MHz MASH ΔΣ modulator with a second-stage multi-rate VCO-based quantizer. In: Proceedings of IEEE Custom Integrated Circuits Conference, pp. 1–4 (2011)
63. Daniels, J., Dehaene, W., Steyaert, M.: A 0.02 mm^2 65 nm CMOS 30 MHz BW all-digital differential VCO-based ADC with 64 dB SNDR. In: IEEE Symposium on VLSI Circuits—Digest Technical Papers, pp. 155–156 (2010)
64. Taylor, G., Galton, I.: A mostly-digital variable-rate continuous-time delta-sigma modulator ADC. IEEE J. Solid-State Circuits **45**(12), 2634–2646 (2010)
65. Rao, S., et al.: A 71 dB SFDR open loop VCO-based ADC using 2-level PWM modulation. In: IEEE Symposium on VLSI Circuits—Digest Technical Papers, pp. 270–271 (2011)
66. Taylor, G., Galton, I.: A reconfigurable mostly-digital delta-sigma ADC with a worst-case FOM of 160 dB. IEEE J. Solid-State Circuits **48**(4), 983–995 (2013)
67. Rao, S., et al.: A deterministic digital background calibration technique for VCO-based ADCs. IEEE J. Solid-State Circuits **49**(4), 950–960 (2014)
68. Prefasi, E., et al.: A 0.1 mm^2 wide bandwidth continuous-time Sigma-Delta ADC based on a time encoding quantizer in 0.13 μm CMOS. IEEE J. Solid-State Circuits **44**(10), 2745–2754 (2009)

69. De Vuyst, B., Rombouts, P.: A 5 MHz 11Bit self-oscillating sigma-delta modulator with a delay-based phase shifter in 0.025mm^2. IEEE J. Solid-State Circuits **46**(8), 1919–1927 (2011)
70. Prefasi, E., Paton, S., Hernandez, L.: A 7 mW 20 MHz BW time-encoding oversampling converter implemented in a 0.08 mm^2 65 nm CMOS circuit. IEEE J. Solid-State Circuits **46** (7), 1562–1574 (2011)
71. Dhanasekaran, V., et al.: A continuous-time multi-bit delta-sigma ADC using time domain quantizer and feedback element. IEEE J. Solid-State Circuits **46**(3), 639–650 (2011)
72. Dhanasekaran, V., et al.: A 20 MHz BW 68 dB DR CT delta-sigma ADC based on a multi-bit time-domain quantizer and feedback element. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 174–175, 175a (2009)
73. Oliaei, O.: Sigma-delta modulator with spectrally shaped feedback. IEEE Trans. Circuits Syst. II, Analog Dig. Signal Process. **50**(9), 518–530 (2003)
74. Putter, B.: Sigma-delta ADC with finite impulse response feedback DAC. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 76–76 (2004)
75. Shettigar, P., Pavan, S.: A 15 mW 3.6 GS/s CT delta-sigma ADC with 36 MHz bandwidth and 83 dB DR in 90 nm CMOS. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 156–158 (2012)
76. Srinivasan, V.: A 20 mW 61 dB SNDR (60 MHz BW) 1 b 3rd-order continuous-time delta-sigma modulator clocked at 6 GHz in 45 nm CMOS. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 158–160 (2012)
77. Shettigar, P., Pavan, S.: Design techniques for wideband single-bit continuous-time Delta-Sigma modulators with FIR feedback DACs. IEEE J. Solid-State Circuits **47**(12), 2865–2879 (2012)
78. Healy, D., Brady, D.J.: Compression at the physical interface. IEEE Signal Process. Mag. **25** (2), 67–71 (2008)
79. Candes, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**(2), 489–509 (2006)
80. Candes, E.J., Wakin, M.B.: An introduction to compressive sampling. IEEE Signal Process. Mag. **25**(2), 21–30 (2008)
81. Gangopadhyay, D., et al.: Compressed sensing analog front-end for bio-sensor applications. IEEE J. Solid-State Circuits **49**(2), 426–438 (2014)
82. Oike, Y., El Gamal, A.: CMOS image sensor with per-column ΣΔ ADC and programmable compressed sensing. IEEE J. Solid-State Circuits **48**(1), 318–328 (2013)
83. Yoo, J., et al.: A 100 MHz–2 GHz 12.5x sub-Nyquist rate receiver in 90 nm CMOS. In: Proceedings of RF IC Symposium, pp. 31–34 (2012)
84. Abari, O., et al.: Performance trade-offs and design limitations of analog-to-information converter front-ends. In: Proceedings of International Conference Acoustics, Speech, Signal Processing, pp. 5309–5312 (2012)
85. Mishali, M., Eldar, Y.: Sub-Nyquist sampling. IEEE Signal Process. Mag. **28**(6), 98–124 (2011)
86. Vetterli, M., Marziliano, P., Blu, T.: Sampling signals with finite rate of innovation. IEEE Trans. Signal Process. **50**(6), 1417–1428 (2002)
87. Tur, R., Eldar, Y.C., Friedman, Z.: Innovation rate sampling of pulse streams with application to ultrasound imaging. IEEE Trans. Signal Process. **59**(4), 1827–1842 (2011)
88. Muramatsu, Y., et al.: A signal-processing CMOS image sensor using a simple analog operation. IEEE J. Solid-State Circuits **38**(1), 101–106 (2003)
89. Choi, J., et al.: A 3.4-µW object-adaptive CMOS image sensor with embedded feature extraction algorithm for motion-triggered object-of-interest imaging. IEEE J. Solid-State Circuits **49**(1), 289–300 (2014)
90. Badami, K., et al.: Context-aware hierarchical information-sensing in a 6 µW 90 nm CMOS voice activity detector. In: IEEE International Solid-State Circuits Conference—Digest Technical Papers, pp. 430–431 (2015)

Chapter 5
Interconnects and Communication

Bernd Hoefflinger

Abstract On-chip wiring lengths, delays-per-bit and energy-per-bit no longer decrease with technology nodes <24 nm. Chip-to-chip communication speeds advanced beyond predictions to >30 GB/s/pin in 2014, however, at the price of energy stalled at 1 pJ/b. The expansion in data volume is so fantastic that bandwidth and power present unanswered challenges. Since video takes up >70 % of the Internet traffic, it receives a special treatment in Chaps. 12–14.

5.1 On-Chip and Chip-Chip Communication

The speed and energy of data transport in complex nano-chip systems has become the no. 1 challenge for progress. In [1], we were cautious and yet optimistic about progress. As of 2014, it has been accepted that

- **On-chip wiring lengths, delay/b and energy/b in 2D designs no longer decrease with the next node on the roadmap**.

These critical quantities increase because of increasing chip complexity (see Fig. 3.4 in Chap. 3). This means that on-chip energy is totally dominated by interconnect and would not decrease with a new technology generation, as is evident in Fig. 5.1 [2], where compute energy is the small part, while supply- and interconnect-lines dominate the total on-die IC energy.

For chip-to-chip communication, we assessed the 2010 state-of-the-art as 10 GB/s/pin and 1 pJ/b. We projected 20 GB/s/pin in 2020 and a potentially 100-times improvement in energy/b, of which the transition in supply voltage from

B. Hoefflinger (✉)
5 Leonberger Strasse, 71063 Sindelfingen, Germany
e-mail: bhoefflinger@t-online.de

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_5

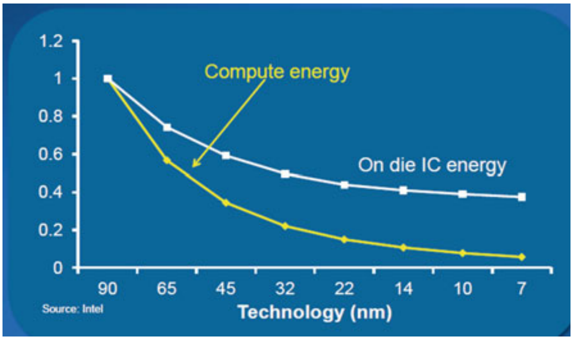


Fig. 5.1 Relative compute energy and total on-die energy for advanced technology nodes [2]

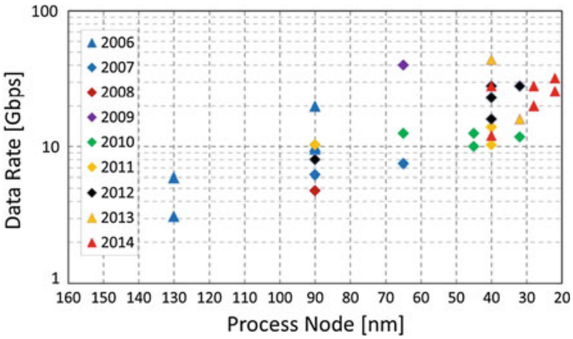


Fig. 5.2 Off-chip data rates per pin versus process node [3] © IEEE 2014

1 to 0.3 V alone would offer a 10-times improvement, requiring fully-differential signaling. While the data rates progressed broadly to 28 GB/s in 2014, as shown in Fig. 5.2 [3] and in the sessions 2 and 26 at ISSCC 2014, the energy is rated at 0.8 pJ (Fig. 5.3), and little progress is anticipated because the supply voltages rest at close to 1 V to maintain speed. A record speed of 60 GB/s/channel was reported in [9] with an efficiency of 0.7 pJ. These energy levels will be needed in scenarios with 20 db channel losses, capacitances close to 1 pF, and supply voltages of 1 V. Since the technology node of 22 nm has already been employed, costly SiGe BiCMOS technology has to be considered to gain speeds beyond 60 GB/channel. At this point, optical communication, Sect. 5.4, has to be considered as well.

5.2 Projections Wireline Communication

Wireline communication bandwidth is expected to double every 4 years, while the required energy/b increases approximately with the square-root of the distance, Fig. 5.3, with a reference value of 5 pJ at 1 m, to be compared with optical interconnect in Sect. 5.4.

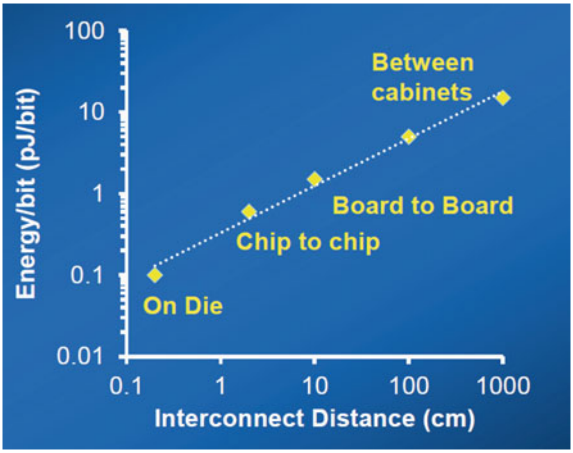


Fig. 5.3 Communication energy versus interconnect distance [2]. *Source* Intel

5.3 Wireless Communication

Wireless communication bandwidth has reached 1 GB/s, Fig. 5.4, 100-times higher than a decade ago. This is an impressive achievement resulting from the sustained progress in transistor bandwidth and in the power-efficiency of transmitters and the sensitivity of receivers.

The progress in power transistors has been particularly remarkable, as illustrated in Fig. 5.5.

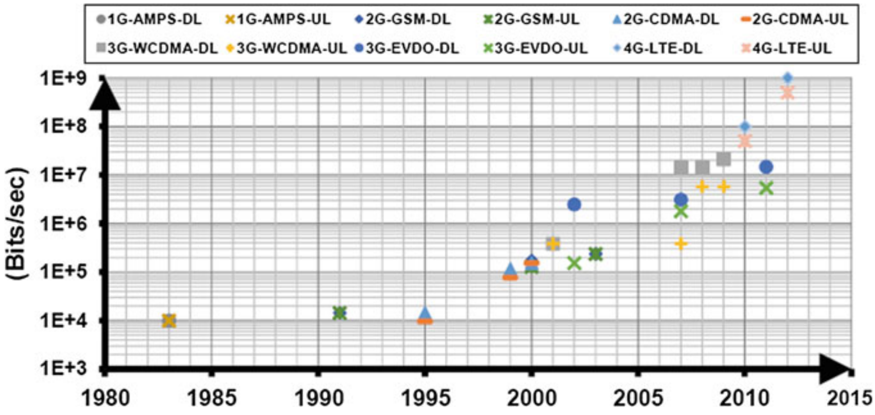


Fig. 5.4 Cellular wireless data rate [4] © IEEE, ISSCC 2014

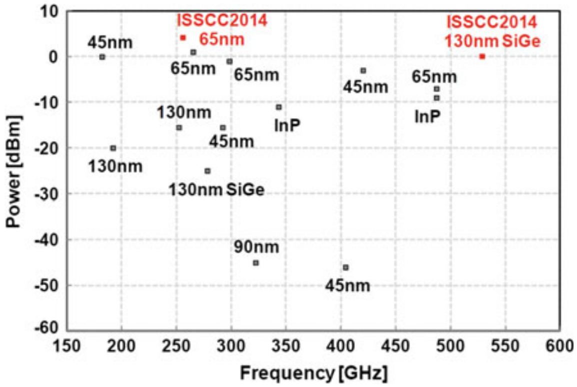


Fig. 5.5 Output power of mm-wave and THz transistors [4]: (0 dBm = 1 mW, 10 dBm = 10 mW, −10 dBm = 0.1 mW) © IEEE ISSCC 2013

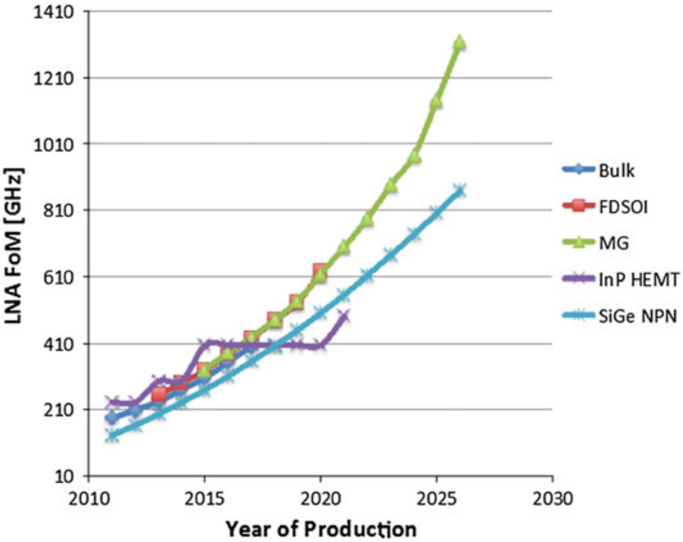


Fig. 5.6 Projected frequency limit of THz transistors [5] © Sematec

The 2014 results demonstrate the advance of Si MOS-FET's and of Silicon-Germanium bipolar transistors, the mainstream technology, which is more compatible with CMOS, against the InP technology.

The future expectations on THz transistors continue to be high, as is evident in the extrapolations in the 2013 ITRS [5], illustrated here with Fig. 5.6. The THz technology is mostly relevant for scanners and imagers.

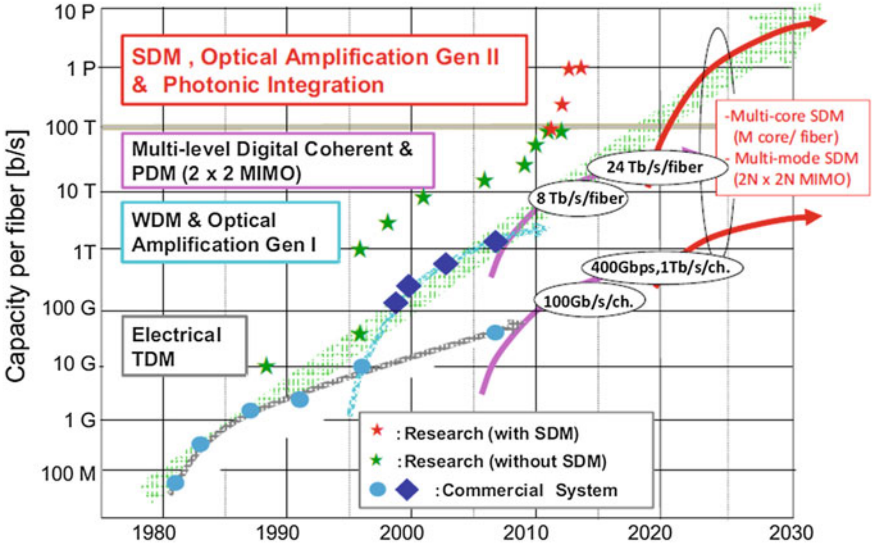


Fig. 5.7 The evolution of fiber capacity for optical communication [8] © IEEE 2014

5.4 Optical Communication

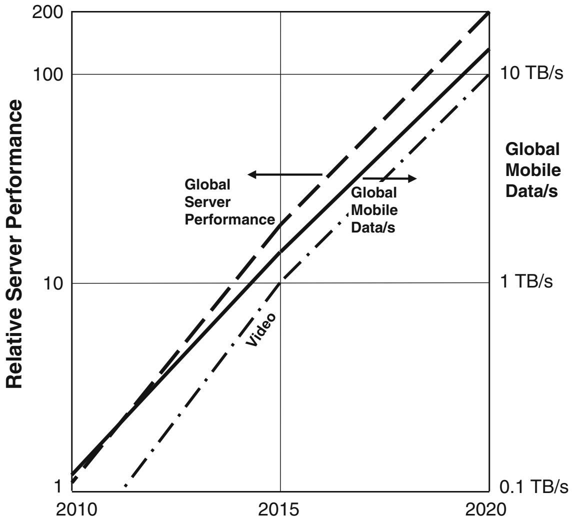
The advent of integrated Laser diodes and integrated low-noise photodiodes produced promising solutions in 2011 (see Table 5.3 in [1]) with up to 24 channels and power efficiencies of 1.5–6 pJ. 2014 can be characterized by 60 channels at 10 GB/s/channel, ∼2 pJ/b, and a pitch of 250 μm in one and 1500 μm in the other direction [6]. Optimized optical receivers have achieved 28 GB/s/channel and a power efficiency of 1 pJ/b [7].

The evolution of fiber capacity and of the related on-chip DSP capacity (for code-conversion) is the big challenge for optical communication. One possible scenario [8] is shown in Fig. 5.7. A fiber reached a capacity of 1 Tb/s in 2010, and a 100 Tb/s/fiber is the speculative goal for 2020. This goal becomes understandable, if we consider the driving global force for data traffic, namely the mobile internet, in the following section.

5.5 Global Mobile Communication

Figure 5.5 in [1], from IEEE Spectrum 2010, had projected the mobile data volume for 2014 at 3.6 Exa-Byte per month (EB/month) and a doubling per year. The CISCO study of 2014 shows 2.6 EB/month, and it corrects the annual growth to 61 %/year [9]. Even so, this means that the traffic will increase another 10-times

Fig. 5.8 Global mobile data rate and the relative performance of servers



from 2015 to 2020, as shown in Fig. 5.8. This growth implies much more bandwidth and—even more seriously:

10-times more data traffic with little progress in the energy-per-bit communicated is not realistic.

The needed paradigm shifts are:

- **Low-power electronics (Chap. 2)**
- **The 3rd dimension (Chap. 3)**
- **Reduce data with intelligent data, with emphasis on video and graphics, inspired by the Human Visual System (HVS) (Chaps. 12–14)**
- **New architectures (Chaps. 16 and 18)**
- **Energy harvesting (Chap. 19).**

5.6 Conclusion

On-chip communication did not improve, neither in speed nor in energy. Speed and bandwidths in off-chip, wireless and optical communication made remarkable progress. However, the energy/b communicated did not improve. At the same time, the data traffic on the Internet doubles every 18 months, projected through 2018, which will necessarily drive us into a major energy crisis. Most chapters of this book are dedicated to avoid this crisis with innovative nanochip-related solutions, which assure future growth and a sustained information society.

References

1. Hoefflinger, B.: Interconnects, transmitters, and receivers. In: Hoefflinger, B. (ed.) Chapter 5 in CHIPS 2020—A Guide to the Future of Nanoelectronics, pp. 37–93, Springer, Berlin (2012). doi:10.1007/978-3-642-23096-7_5
2. Borkar, S.: Exascale Computing—Fact or Fiction? SSCS Webinar, Sept 2014
3. Friedman, D.: Wireline, in "ISSCC 2014 Trends", pp. 106–108 (2014). http://isscc.org
4. Pärssineen, A.: Wireless, in "ISSCC 2014 Trends", pp. 104–105 (2014). http://isscc.org
5. www.itrs.net/reports/2013-edition
6. Morita, H. et al.: A 12 × 5 Two-dimensional optical I/O array for 600 GB/s chip-to-chip interconnect in 65 nm CMOS. In: Solid-State Circuits Conference Digest of Technical Papers 2014, pp. 140–141, Feb 2014
7. Huang T.C. et al.: A 28 GB/s 1 pJ/b shared-inductor optical receiver with 56 % chip-area reduction in 28 nm CMOS. In: Solid-State Circuits Conference Digest of Technical Papers, 2014, pp. 144–145, Feb 2014
8. Miyamoto Y.: High-Capacity Scalable Optical Communication for Future Optical Transport Network, pp. 118–120, ibid., paper 6.2, Feb 2014
9. CISCO Visual Networking Index: Global mobile traffic forecast update (2014). http://www.cisco.com/

Chapter 6
Superprocessors

Systems for a Cloud, Analytics, Mobile and Social Era

Cédric Lichtenau, Philipp Oehler and Peter Hans Roth

Abstract The rise of big data combined with a slow-down in technology advancement has made processor and system design even more challenging. In this paper we describe current processor designs to adapt to a changing landscape, and we discuss future trends towards software-defined computing.

While circuit scaling continues, we have reached the point where transistor performance is saturating and power density becomes an issue. This results in only marginal room to boost single-thread performance by means of a higher voltage and frequency. State-of-the-art processors and systems are now turning towards throughput and specialized hardware to better respond to new emerging workloads working on huge datasets e.g. for analytics or mobile. Wide multi-threading coupled with large caches allows to better use the existing hardware and keep work flowing through the system while waiting for memory or I/Os. Hybrid processors and systems with accelerators closely coupled to the processors and memory are on the rise for an increasing single-thread performance, as not all applications can easily be parallelized and customers expect quick response times.

Big new players like managed service providers (MSPs) are also increasingly looking toward open standards also for processor and system development. They want specific acceleration capacity tailored to their business deep into the system and open innovation inside a wide eco-system instead of just connecting acceleration cards with limited bandwidth and a large software overhead.

Looking into the future, a modern processor in CMOS technology currently integrates over 2 billion transistors. While the transistor performance is saturating, it will still be very challenging for a new technology to match current chip complexity

C. Lichtenau (✉) · P. Oehler · P.H. Roth
Microprocessor Development, IBM R&D, Schoenaicher Str., 220,
71032 Boeblingen, Germany
e-mail: lichtenau@de.ibm.com; cedric.lichtenau@de.ibm.com

P. Oehler
e-mail: oehler@de.ibm.com

P.H. Roth
e-mail: peharo@de.ibm.com

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_6

and performance. We strongly believe that in this constrained space the next innovations will be fueled by hardware-software co-design. This will result in processor and system architectures that are better customized to the workloads running on them.

6.1 Evolving Workloads

In the last decade, we have seen an exponential growth of mobile computing with the wide introduction of the smartphone. Not only does this represents a huge number of new devices that access the backbone infrastructure at the same time, but we also see a new behavior with increased number of requests per device around the clock. The typical example would be a customer checking his account balance, in the past maybe once a week, but now able to do it easily from his mobile device and anytime. This trend is likely to continue, emphasized by the emergence of the Internet of Things, connecting virtually every device to the internet and providing a constant stream of data to analyze and triggering intelligent predictive actions as responses.

This evolution is shown in Fig. 6.1. Previously, data would be collected and processed later in batches to understand what happened. This is no longer sufficient to satisfy the customers' expectation. Across all industry sectors, companies strive

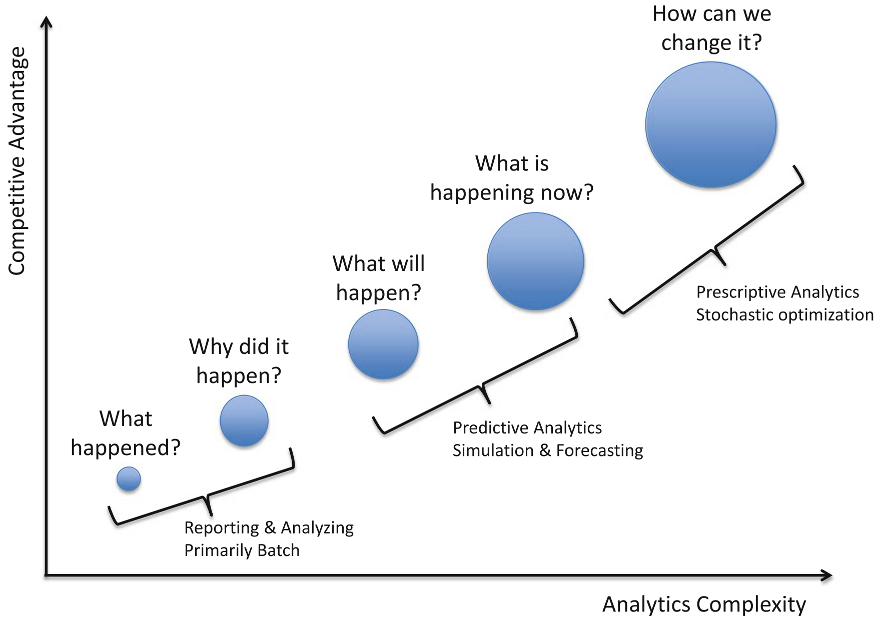


Fig. 6.1 Business analytics evolution

to analyze the stream of data online to take more insightful and relevant actions. Providing additional services or offers through predictive business analytics is the main driver to increase their revenue and margin. This advanced analytics relies on complex mathematic models and is working on large data sets to better understand the desires of a customer.

These new workload characteristics are driving a profound change of the server processor and a system design focused on stronger performance for business analytics (mathematical unit, hardware accelerator), support for a large number of parallel requests (multi-core, multi-threading) and a major additional load on the memory and I/O subsystem (in-memory database, big data).

As classical performance gain through increased frequency and design shrink is saturating, the next big step is workload-specific optimization of the complete system. One example for such a state-of-the-art system is the POWER8 system optimized for big data and analytics to respond to the need to process huge amounts of data in real-time for cloud, analytics, mobile and social workloads.

6.2 POWER8—a Big-Data Processor

In the past, new silicon technology advancements enabled the rate of increase in processor frequency. This has decreased dramatically in recent generations. Many processor designs show very little improvement in single-thread or single-core performance. Instead, a larger number of cores are implemented to compensate for reduced technology advancements.

The POWER8 processor is a balanced multi-core design with significant improvement in single-thread and single-core performance [1, 2]. Additionally, the number of cores has been increased. The RISC (Reduced Instruction Set Computer) processor introduces a twelve-core multi-chip design, large on-chip eDRAM (embedded Dynamic Random-Access Memory) caches, and high-performance eight-way multi-threaded cores. Table 6.1 shows the large amount of cache capacity added to the POWER8 processor. Each core does have a private L2 cache, twice the size compared to the previous generation. The shared L3 cache, implemented on-chip as embedded DRAM, has a size of 96 MB. An additional level—L4 cache —is added to the cache hierarchy: up to eight memory buffer chips (each containing 16 MB) are connected to the processor. Starting with a maximum of two threads on a POWER4 processor, i.e. one thread per core, multi-threading was introduced for POWER5 (two simultaneously active threads running on each core). For POWER7, the number of cores and the number of simultaneously active threads increased, resulting in 32 active threads per socket. The latest generation POWER8 is able to support up to 96 active threads per core (e.g. a 12 core POWER8 processor with 8 active threads on each core).

Enhancements of the micro-architecture of the POWER8 core improved thread and core performance significantly [3]. A list of the most important enhancements contains:

Table 6.1 POWER4 to POWER8 key facts cores, caches, frequency, threads

Processor	Max L1/L2 cache	Max L3/L4 cache	Core frequency (GHz)	Threads per core (Max)
POWER4 (2 cores)	64 kB/core	32 MB off-chip	1.9	1 (2)
	1.41 MB	–		
POWER5 (2 cores)	32 kB/core	36 MB off-chip	2.3	1, 2 (4)
	1.875 MB	–		
POWER6 (2 cores)	64 kB/core	32 MB off-chip	5.0	1, 2 (4)
	4 MB	–		
POWER7 (8 cores)	32 kB	32 MB	4.25	1, 2, 4 (32)
	256 kB/core	–		
POWER8 (12 cores)	64 kB	96 MB	4.25	1, 2, 4, 8 (96)
	512 kB/core	128 MB		

- Advanced eight-way simultaneous multi-threading
- Doubled bandwidth throughout the cache and memory hierarchy
- Extensive out-of-order execution
- Support for fast access to unaligned data and little endian data

Several new differentiating features are supported, e.g. advanced security, enabling dynamic compiler optimization, cryptography acceleration, advanced SIMD (Single Instruction Multiple Data) features, and enabling business analytics optimization.

In addition to traditional workloads, the POWER8 processor is highly optimized for business analytics, big data and system of engagements applications, as well as cloud-based workloads (Table 6.2).

The POWER8 core delivers improved SIMD performance. Through symmetric vector pipelines, and an expanded repertory of SIMD integer instructions. These instructions significantly improve the performance of business analytics applications. Many of these computations offer inherent parallelism, which can be exploited in thread-rich configurations. The POWER8 core doubled the hardware thread parallelism to eight-way multi-threading. The eight-way simultaneous multi-threading enables performance gains on scientific and technical computing, also known as HPC (High Performance Computing).

To deal with a larger memory footprint of typical Big-Data applications, all levels of hierarchy across the cache and memory structure have increased capacity. Compared to the POWER7 core, the POWER8 core has an L1 data cache that is

Table 6.2 Different workload scenarios and contributors to performance improvements

Workload scenario	Performance mainly influenced by
Business analytics	SIMD (single instruction multiple data), multi-threading
Big data	Cache and memory capacity, memory bandwidth
Cloud	Dynamic scripting languages, exploit parallelism across cloud instances

twice as large and has twice as many ports from that data cache for higher read/write throughput. In addition, L2- and L3-Caches in the POWER8 processor are also twice the size compared to the POWER7 processor. A new hierarchy off-chip cache (Level 4) has been introduced. This new L4-cache is included in the new Centaur buffer chips.

In contrast to the parallelism in computations for business analytics, cloud instances often do not have enough simultaneously active application threads to operate all available hardware threads. The POWER8 can exploit the parallelism across cloud instances. The POWER8 core can be put into a special mode “split-core mode”. This mode allows to run four partitions on one core at the same time, with up to two hardware threads per partition. In systems of engagements, i.e. systems, which are decentralized, it enables and incorporates peer interaction, dynamic scripting languages such as Ruby, Python and JavaScript. The POWER8 core improves performance for these workloads through better branch prediction, increased instruction-level parallelism and efficient unaligned data access.

The above mentioned improvements at the micro-architectural level only show a small sub-set of new features added to the POWER8 processor compared to previous generations of the POWER CPUs. The improvements of the POWER8 core have resulted in a significant performance gain. Not only single-thread performance has been improved, but all other modes of multi-threading, too. In addition, a performance comparison at core level shows that running one POWER8 core in SMT2 results in the same performance numbers as running two POWER7 cores in single-thread.

POWER8 Core

Figure 6.2 shows the POWER8 core floorplan. It consists of six units: IFU (Instruction Fetch Unit), ISU (Instruction Scheduling Unit), LSU (Load Store Unit), FXU (Fixed-point Unit), VSU (Vector Scalar Unit), and DFU (Decimal Floating-point Unit). The 32 kB I-Cache (Instruction Cache) is located in the IFU and the 64 kB D-Cache (Data Cache) is located in the LSU. Both, I-Cache and D-Cache, are backed up by a 512 kB unified L2-Cache (Level 2 Cache). The newly designed core can fetch and dispatch up to eight instructions in any given cycle. Up to ten instructions per cycle can be issued to one of the sixteen execution pipelines:

- two fixed-point pipelines (FXU),
- two load/store pipelines (LSU),
- two additional load pipelines (LSU),
- four double-precision floating-point pipelines, which can also act as eight single-precision floating-point pipelines (VSU),
- two fully symmetric vector pipelines (VSU),
- one crypto pipeline (VSU),
- one branch execution pipeline (IFU),
- one condition register logical pipeline (IFU), and
- one decimal floating-point pipeline (DFU).

The POWER8 core has significantly higher load/store bandwidth compared to POWER7. While the predecessor core can perform two load/store operations in a

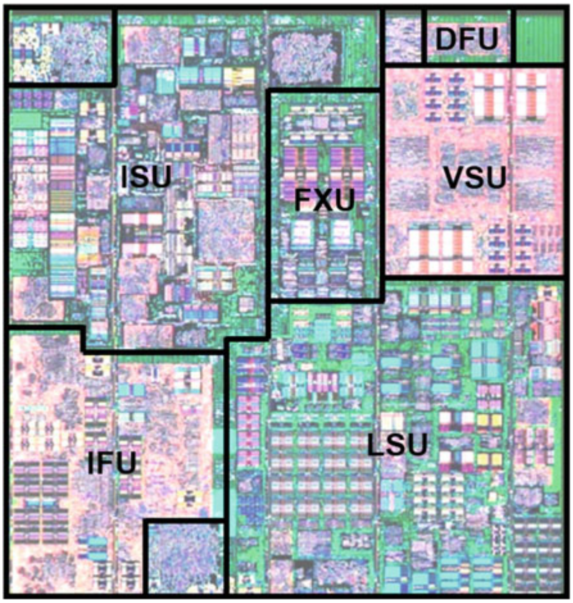


Fig. 6.2 POWER8 core floor plan

given cycle, the POWER8 processor can perform two load operations in the load pipelines, and additional two load or store operations in the load/store pipelines.

Figure 6.3 shows the instruction flow in the POWER8 processor. Instructions are processed from the cache (L2- and L1-cache), through various issue queues and are then sent to the execution units. The Unified Issue Queue, which has two symmetrical halves (UQ0 and UQ1) process most of the instructions (except for branches and condition register logical instructions). The execution pipelines are split into two sets: FX0, FP0, VSX0, VMX0, L0, LS0 associated with UQ0, whereas FX1, FP1, VSX1, VMX1, L1, LS1 associated with UQ1. Not shown in the flow diagram are the two copies of the general-purpose (GPR0 and GPR1) and vector-scalar (VSR0 and VSR1) register files. Which resources (e.g. issue queue, register file, and functional unit) are used by a given instruction, depends on the SMT mode of the processor core.

Workload adaptive performance

New features in the POWER8 processor allow to dynamically adapt the core to specific workload demands [4]. Thread switching and thread balancing at run-time gives tremendous performance improvements compared to previous POWER7 simultaneous multi-threading scenarios.

The operating system maintains the system load, i.e. the number of running or runnable software threads. Software is scheduled to hardware thread T0–T7. If there is only one software thread running, the POWER8 core will switch to ST mode, no matter which hardware thread it is running on. This is an improvement over the

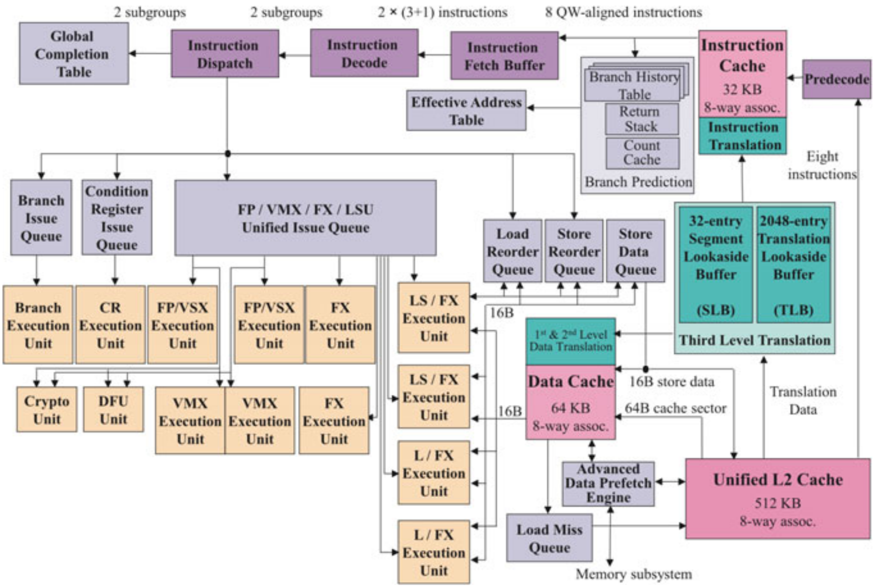


Fig. 6.3 P8 core pipeline flow

POWER7 processor, where expensive software move operation was needed to get the software thread to hardware thread T0 (if it is not already there), before the core can switch into ST mode. If the average load is getting greater then 1, the operating system will begin to schedule work on more hardware threads (T1–T7). The POWER8 core can switch from any SMT mode to any other SMT mode. In SMT mode, the hardware threads are split into two thread-sets: even thread-set (T0, T2, T4, T6) and odd thread-set (T1, T3, T5, T7). Each thread-set uses half of the resources available on the core, i.e. half of the issue queue entries, half of the execution pipelines.

After an SMT mode change or as threads are disabled, there might be an imbalance in the number of active threads per thread-set. The POWER8 processor rebalances the thread distribution across the thread-set. This improves core performance, as core resources are more fairly utilized by the active threads. In addition, the rebalancing of the number of active threads across the thread-set, a better thread mixing, i.e. moving threads across thread-sets depending on the resources which are needed, also improves the core throughput. This feature only takes advantage when the core is running in SMT-4 or SMT-8.

An example of resource allocation by hardware threads T0–T3 are shown in Fig. 6.4. The POWER8 core is running in SMT-4, where one thread-set (T0 and T1) is using a lot of floating-point, VMX, or crypto operations, i.e. VSU-intensive, whereas the other thread-set (T2 and T3) is not, e.g. running operations like fixed-point, load/stores. It is obviously a big advantage to move one of the VSU-intensive threads to the second thread-set where VSU resources are idleing, as shown in Fig. 6.4b.

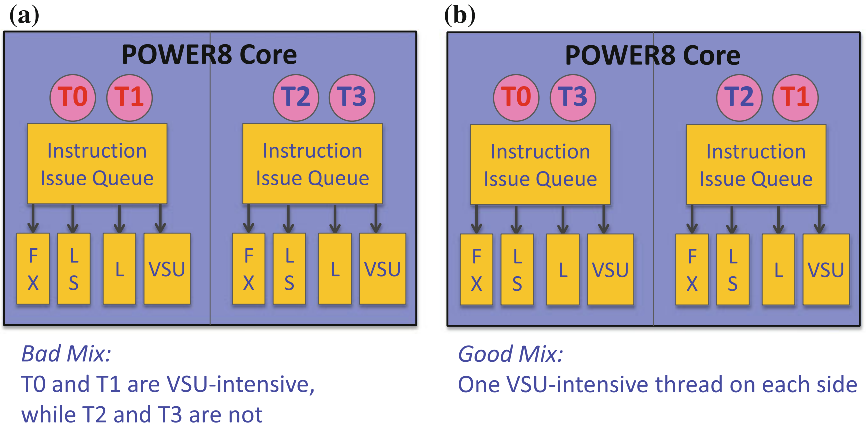


Fig. 6.4 Thread-set mixing on a POWER8 core in SMT-4, **a** shows an example of a bad mix, **b** shows a good mix

6.3 Security

Traditionally, companies kept their data on-premise with their own IT department to manage the data center infrastructure and to guarantee the data security. This has changed, and we expect that the transition will continue from on-premise data centers to hybrid or public clouds, driven mainly by the desire to reduce the cost of IT and to have access to a flexible and elastic infrastructure. A company can grow its IT-footprint as it expands or during a certain time of the year. Typical examples are a tax-service company with country-specific deadlines or holiday season sales for retail. At the same time, there is not a single month going by without another high-profile case of a security break at a large company [5, 6]. Most known cases result in identity theft that has become a big concern to individuals and to the reputation of the affected companies. There is also the risk for a company to loose its competitive advantage, if competitors access internal roadmaps or customer databases. Certain industry branches like healthcare have very strict regulations with data privacy, needing additional certifications. Putting clients'/company's data into a public cloud requires well-considered security concepts and an infrastructure designed to satisfy the confidentiality requirements. Not only does a customer want strong encryption for his data, he is also looking for guarantees that neither the system administrator nor other customers, using the same (virtualized) physical machine, can ever access his data.

To achieve this, security must be built from ground up into the machine. This is especially complex for a multi-node machine with its processors distributed on different motherboards that may be replaced individually while the system keeps running.

A good example for such a system is the POWER8 chip (see Fig. 6.5). It supports secure boot where an on-board micro-controller locks down access to the chip before initializing it based on data contained in an on-board secure memory. It then loads the boot code from flash to the internal cache to prevent any tempering, and it checks the code signature before executing it. Once the chip is initialized and considered secure, the integrity of the other chips in the system is verified before allowing memory-coherent communication between chips. The memory is divided into a trusted area for customer data that can only be accessed by trusted entities in the system and an untrusted area for servicing and diagnostics of the machine. A major effort has been made to prevent unauthorized access to trusted data, by adding tempering detection in the hardware and memory/cache soft destruction on attempts to read out data from the chip via the service interface.

The POWER processors also support cryptographic algorithms in hardware on-chip. Since POWER7+, the following algorithms are supported by the NX (Nest Accelerator) unit:

- AES (Advanced Encryption Standard) in various modes of operation
- SHA (Secure Hash Algorithm)
- AMF (Asymmetric Math Functions)
- RNG (Random Number Generator).

The same NX unit also hosts the memory compression and decompression engines. For POWER8, additionally in-core acceleration is added. The POWER8 in-core cryptographic enhancements are targeting applications using symmetric cryptography (e.g. AES) and security (e.g. Secure Hash Algorithm SHA-2 and Cyclic Redundancy Checking CRC) algorithms. The POWER8 VSU data path has been extended to

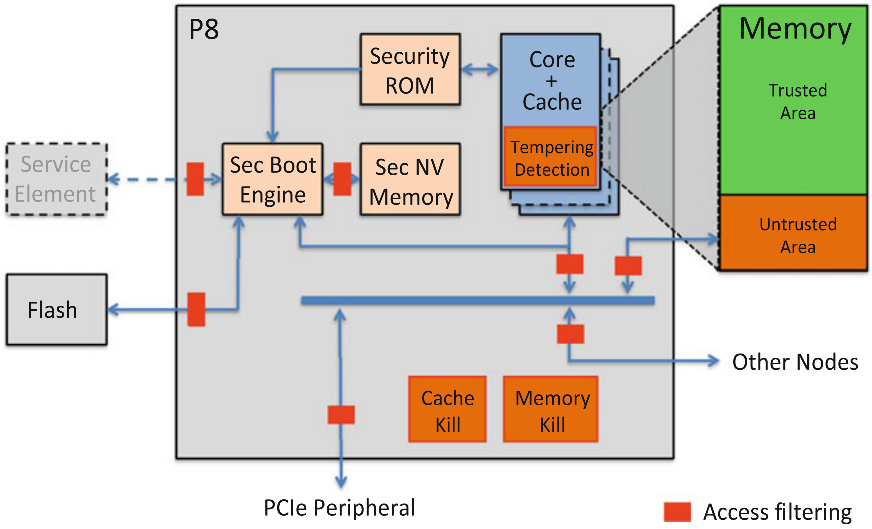


Fig. 6.5 POWER8 security design

include the cryptographic enhancements. The cryptographic sub-unit is fully pipelined. It allows for a fast 6-cycle issue-to-issue latency. AES is a symmetric-key algorithm, which processes data blocks of 128 bits (a block cipher algorithm), and, therefore, naturally fits into the POWER8 Vector and Scalar Unit (VSU). The AES algorithm is completely covered in five new instructions added to the Book 1 PowerPC Architecture, available in Chap. 5 "Vector Facility" (VMX). Encryption performances up to 5.7 GB/s are achieved with the core running at 4 GHz [2].

The AES Galois/Counter Mode (GCM) of operation is designed to provide both confidentiality and integrity. GCM is defined for block ciphers (block sizes of 128, 192, and 256 bits). The key feature is that Galois Field multiplication GF(2^{128}) can be computed in parallel, resulting in a higher throughput than the authentication algorithms that use chaining modes. Four new instructions have been added to the Book 1 PowerPC Architecture Vector Facility to support the GF multiplication on GF(2^{128}); GF(2^{64}), GF(2^{32}), GF(2^{16}), and GF(2^{8}).

The new polynomial multiply instructions can also be used to assist CRC calculation. CRC algorithms are defined by the different generator polynomial used. For example, an n-bit CRC is defined by an n-bit polynomial. Examples for applications using CRC-32 are Ethernet (Open Systems Interconnection (OSI) physical layer), Serial Advance Technology Attachment (Serial ATA), Moving Picture Experts Group (MPEG-2), GNU Project file compression software (Gzip), and Portable Network Graphics (PNG, fixed 32-bit polynomial). In contrast, Internet Small Computer System Interface (iSCSI) and the Stream Control Transmission Protocol (SCTP transport layer protocol) are based on a different, 32-bit polynomial. The enhancements on POWER8 not only focus on a specific application that supports only one single generator polynomial, but they help to accelerate any kind of CRC size, ranging from 8-bit CRC, 16-bit CRC, and 32-bit CRC, to 64-bit CRC.

Another set of instructions have been added to the Book 1 PowerPC Architecture to support Secure Hash Algorithm (SHA-2). SHA-2 was designed by the U.S. National Security Agency (NSA) and published in 2001 by NIST, latest update 2012 [7]. It is a set of four hash functions (SHA-224, SHA-256, SHA-384, and SHA-512) with message digests that are 224, 256, 384, and 512 bits. The SHA-2 functions compute the digest based on 32-bit words (SHA-224 and SHA-256) or 64-bit words (SHA-384 and SHA-512). Different combinations of rotate and xor vector instructions have been identified to be merged into a new instruction. To accelerate the SHA-2 family, two new instructions have been added to the Book 1 PowerPC Architecture Vector Facility. The STD PKCS11 APIs and CLIC software exploit these new instructions on POWER8.

6.4 Optimization Across the Stack

While multi-core and multi-thread keep steadily increasing the throughput performance of modern computer systems, the single-thread performance has not seen a similar advance. Many applications or libraries have been rewritten to better take advantage of multiple cores and threads in a system to increase throughput. This parallelism cannot be extended indefinitely as interlocked business processes still have an inherent need for synchronization between threads and for write access to shared data. Also in a mobile and faster moving world, users are expecting "immediate" answers to their personal request. This requires improved single-thread performance as companies are striving to exploit more and more complex business analytics to better serve its customer needs and improve profit margin through a competitive advantage.

The circuit frequency is topping out due to device performance saturation. At the same time, larger and more complex logic structures to recover performance are affecting signal travel time from unit to unit due to increased wire-resistance. This can result in additional pipeline stages affecting performance. There is a definite need to look into other directions than just technology to reduce the average number of cycles per instruction.

Isolated optimizations in various parts of the hardware and software stack have been and are still on-going but they are quite often looking for a local optimum, that is only modestly contributing to the complete hardware/software stack as depicted in Fig. 6.6. We strongly believe that major innovations, taking the complete stack into consideration, will fuel the future improvements. These innovations fall into two categories. Either do global reaching optimizations of a certain application pattern (e.g. hardware support for a sub-routine call) that will moderately benefit most workloads and that will generally come without a recompile or a code change. Or algorithmic changes coupled to a hardware-specific acceleration (e.g. a string parsing engine) that significantly speeds up a particular class of workload, but is also limited to that class.

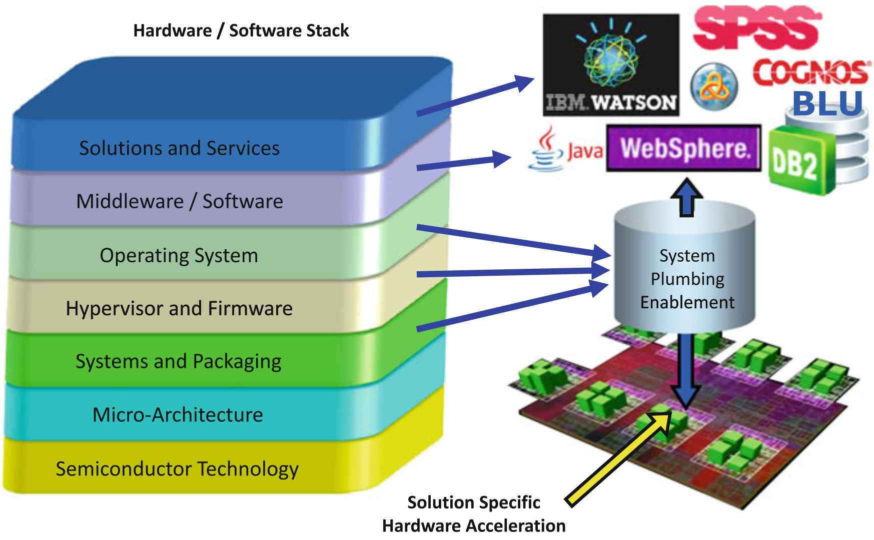


Fig. 6.6 The hardware/software stack (*Source* William Starke)

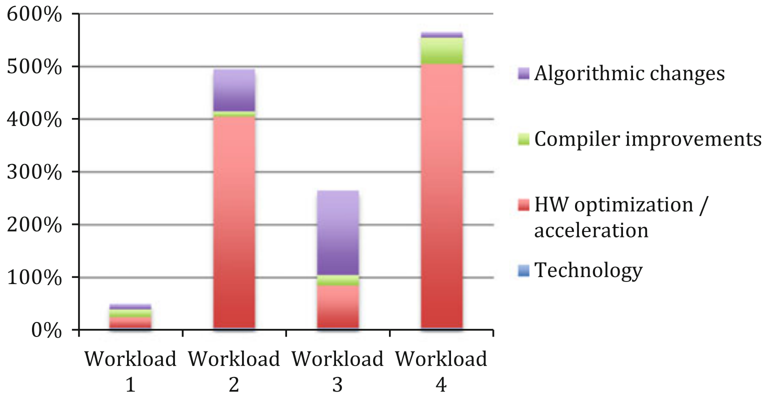


Fig. 6.7 Impact of various optimization methods

Hardware/software optimization relies, in the first step, on a precise measurement to recognize bottlenecks and potential enhancement areas. Collaboration between software and hardware teams, coupled to thinking out-of-the-box, are key to optimize across the hardware/software stack and weigh the performance/risk implications. Some results from optimization done for a POWER7 to a POWER8 system are shown in Fig. 6.7. The effort and runway for optimization is very different: 2–5 years for processor silicon (new instruction), a few months for FPGA/GPU specific acceleration and just a few weeks for algorithm optimization. While architectural changes to the core and system structures generally bring the largest performance gain, it is a bet into a distant future. Software and compiler changes, to better use the underlying hardware or additional accelerators, are much faster to realize, and they can still provide significant performance gains.

Looking at the improvement trend in recent years, we strongly believe that an increase in the application of single-thread performance will mainly rely on hardware/software co-design synergy. There is significant potential lying around. Compared to previous technology-driven enhancements, it will require engineers to have a deep understanding of the complete hardware/software stack and architecture to optimize workloads.

6.5 Accelerators

Some workloads or parts of them are not well suited for a general-purpose processor but can be accelerated significantly by special hardware. It started decades ago with co-processors executing floating-point operations in hardware instead of emulating them with thousands of general-purpose processor instructions [8]. GPUs (Graphic Processor Unit) are now a state-of-the-art accelerator used to speed up graphics representation. Their capacities keep increasing as games' complexity as well as display resolution are steadily increasing. A GPU contains hundreds of simple compute units. Its usage does not need to be limited to graphics. It can be reused to accelerate other workloads, e.g. in the area of technical computing to solve complex and massively parallel problems. For example, the Cell Processor, developed jointly by Sony*, Toshiba* and IBM* for the PlayStation,* was adapted and used as GPU in the Roadrunner supercomputer together with a general-purpose processor and led the Top 500 list some years ago [9]. Another supercomputer—BlueGene—with close to 100,000 GPU chips and no general-purpose processors was developed a few years later.

This concept, while very successful for embarrassingly parallel and mainly independent computations, has limitations when mapped to more traditional workloads. First, workload data need to be transferred from the general-purpose processor's memory to the accelerator(s) and back from after-task completion. This transfer time must be at least offset by the accelerator speed-up gain on the complete workload. Bandwidth and latency of the accelerator's interface to the processor's memory is directly affecting the transfer time.

The second limitation with this architecture is that application code—if not interpreted—must be rewritten to take advantage of the acceleration including synchronization of processors and accelerators during the execution, resulting in a large software bill. GPU vendors have developed ready-to-use libraries and programming language to ease the code development [10]. Nevertheless, the code path for system calls and synchronization overhead between processors and accelerators alone take many hundreds or thousands of instructions, raising the bar high for efficient use of accelerators: large chunks of a workload must be off-loaded in order to achieve any gain.

Figure 6.8a shows the current typical connection of an accelerator via an I/O controller and generally PCI-E as physical interface. This implementation suffers from both limitations mentioned above. Integrating the I/O controller on the processor chip, as well as using the latest generation PCI-E, significantly reduces the

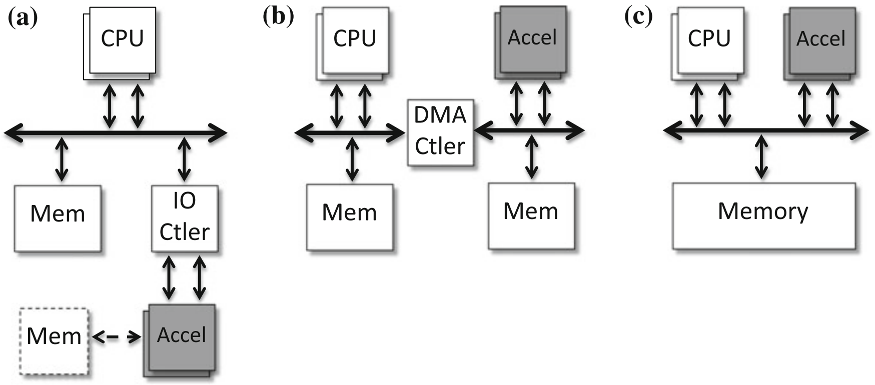


Fig. 6.8 System acceleration architectures

latency and increases the data throughput to the accelerator. This is for example implemented on the POWER8 or Intel Haswell chips to efficiently connect GPUs. Figure 6.8b depicts the use of special hardware DMA engines to do the data transfer while other tasks are executed on the processor and the accelerators. This can hide the data transfer time at the expense of code engineering to carefully schedule the tasks in the system. Finally, Fig. (6.8c) depicts the ideal world where the accelerator (s) is/are coherently connected to a shared memory like the general-purpose processors in the system. With a low-latency, high-bandwidth interface, this will remove all the limitations mentioned above. The POWER8 processor is the first server processor chip to implement such a model by connecting FPGAs for accelerated functions via the Coherent Attached Processor Interface (CAPI). It provides a simple programming model with standard memory semaphores to synchronize the processors and accelerators and to work on common data. The accelerator's interface to memory has only a limited bandwidth and "slow" FPGA-hardware may not be seen at first as a major improvement. But a significant reduction in system footprint (80 %), energy consumption and cost can be achieved with this setup for certain workloads [11, 12]. This is another area of major interest when building large cloud centers, as these parameters are as important as the peek performance.

While GPUs are the most common class of chips used for acceleration, they have a fixed instruction set and I/Os that must fit the workload to be accelerated. FPGA acceleration has been less popular in the past, as it required significant knowledge and effort to modify application code and to design the logic. Successful examples of FPGA acceleration are IBM's Smart Analytics Optimizer* or IBM z System Crypto and Compression.

Another source of acceleration is the Single-Instruction Multiple-Data (SIMD) unit located on modern processor chips like ×86, POWER or zSystem. It executes a particular instruction on n sets of data in parallel (vector operation). State-of-the-art compilers can auto-vectorize code already taking advantage of the SIMD instruction set without any code change. Only certain instructions can be vectorized, and the number of parallel operations is an order-of-magnitude smaller than on a GPU, but it still allows quick access to acceleration without code change.

Scenarios for accelerators in modern workload are not missing. We believe that their usage will significantly increase once integrated in the system to be equal to the general-purpose processors, i.e. connected coherently and with high bandwidth to memory; as well as fully integrated into the software stack with programming and compiler support. Beside the performance enhancement point, also significant reduction in system footprint, power and cost can be achieved. The improvements that can be achieved are huge. We are just at the beginning of this trend.

6.6 Open Computing

Building large scale-out systems in a big-data cloud center puts an increased focus on power efficiency and optimization with speed of innovation giving a particular company an advantage over its competitors. Managed service providers (MSP) are the backbone of our modern life. They build compute centers with typically on the order of 100,000 processors chips, so that every small gain in performance, power or price is making a huge difference. They are looking to get the best-of-class for all components building their systems to maximize infrastructure efficiency and profit.

These big players strongly prefer not to be locked into a single vendor relationship. They are looking for open standards for their whole infrastructure including their hardware components, similar to what exists for software. They want to fuse the innovation of multiple companies around open and clearly specified standards. They want to add custom enhancements, knowing best their specific workload requirements, or to make changes on existing components to get a competitive advantage and best fit their needs.

The Open Compute Project or the Open Power Foundation are working to provide such an open eco-system covering hardware and software to build the infrastructure of tomorrow and to respond to the needs of the industry. The system development moves to a per-customer-centric development (see Fig. 6.9). For example, the complete system design—not just limited to the processor chip—is made available to customers as well as the software stack including an open BIOS and a complete tailored open-source stack. Fully customized solutions with the expertise and with the innovation of multiple companies in all areas of computing

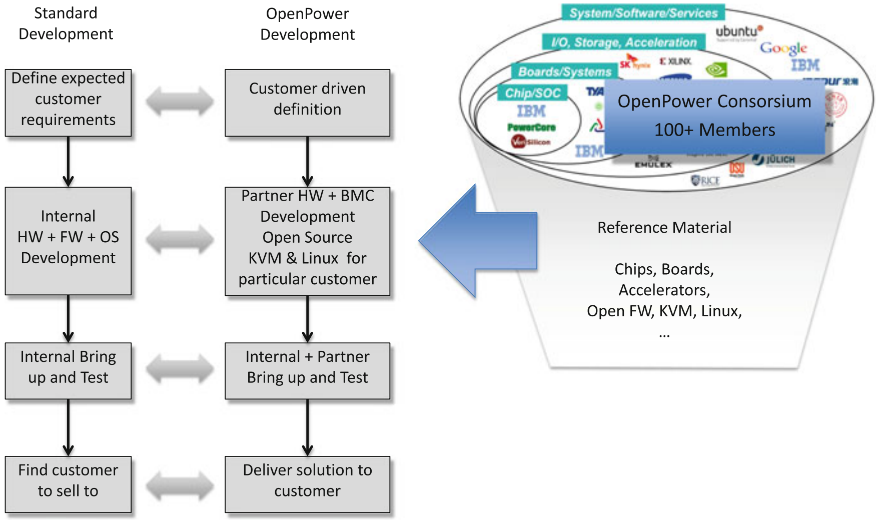

Standard Development
OpenPower Development
Define expected customer requirements
Customer driven definition
Internal HW + FW + OS Development
Partner HW + BMC Development Open Source KVM & Linux for particular customer
Internal Bring up and Test
Internal + Partner Bring up and Test
Find customer to sell to
Deliver solution to customer
System/Software/Services
I/O, Storage, Acceleration
Boards/Systems
Chip/SOC
ubuntu
Google
IBM
OpenPower Consorsium
100+ Members
Reference Material
Chips, Boards, Accelerators, Open FW, KVM, Linux, …

Fig. 6.9 Open power foundation development

can be built. Depending on the expertise of a customer, either a tailored solution can be build, or the customer is picking up parts of the IP and develops his own solution. Both paths allow for a much higher level of system customization while providing a fast and low-risk, agile solution based on an existing toolbox.

Looking at the current server landscape, this trend marks a disruptive change of direction. Today, due to silicon scaling we still see more and more of the peripheral processor functions moving to a monolithic "multi-purpose" processor chip developed by one company and reducing the competition and innovation. We believe that an open eco-system will change this trend and pave the way for companies to take the next step and to make modifications even deeper in the design and to enhance the processor content and architecture towards their needs to better adapt to a rapidly changing world.

Another area, where this could lead to new developments, is the area of technical computing. In the past years, it has been mainly driven by systems having an increasing number of cores running at higher frequency. Previously, new architectures, innovations in core design, system topology and network were fueling the advancement. An open eco-system (open computing) could allow for this to happen again, with smaller companies or universities being able to drive new ideas while reusing most of the existing base infrastructure. Similar developments have been seen in the software area, e.g. with open-source databases tailored to a particular problem like MongoDB or NoSQL databases.

6.7 Outlook

Over the last 40 years, the advancement in technology has been the major driver for improved system performance. While circuit scaling continues, the transistor performance has saturated, and we are increasingly confronted with power-density issues. Moore's law still remains true and will probably for a few more decades as it only addresses the circuit density. The trend towards many cores and many threads is reflecting this and is well suited to respond to the increasing number of requests from a steadily increasing number of mobile devices and the internet of things.

Single-thread performance—or how fast a customer query is answered based on more and more complex analytics—has not seen a similar development. We have even seen in some recent processors a decrease of single-thread performance to accommodate for the multi-core, multi-threading enhancements. Innovation to overcome this will need to come from the remaining part of the hardware/software stack: applications, OS/middleware, system/processor architecture and micro-architecture.

We also see a trend that just a few companies specialize in the CMOS nano-technology manufacturing, as huge wafer volumes and investments in innovations like 3D-chips are needed to offset the investment costs and generate a competitive advantage. The remaining part of the industry is already fab-less or completing its transition and concentrating on the other part of the stack to innovate. This will imply a disruptive change in the focus of the hardware designer to adapt to this new paradigm. His scope will need to embrace the complete hardware/software stack, working closely with software engineers to understand software/application requirements.

With successful developments in hardware/software co-design and optimization across the stack, we expect to keep seeing exponential system-performance improvements for the next decade.

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries

AIX	IBM	POWER	IBM Watson
Cognos	IBM (logo)	zSystem	SPSS
DB2 BLU	Blue Gene	Websphere	CAPI
OpenPower			
IBM's Smart Analytics Optimizer		IBM z System Crypto and Compression	

The following are trademarks or registered trademarks of other companies.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license there from.
Java and all Java-based trademarks are trademarks of Oracle, Inc. in the United States, other countries, or both.
Intel, Intel logo, Intel Inside, Pentium, Haswell are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
UNIX is a registered trademark of The Open Group in the United States and other countries.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
All other products may be trademarks or registered trademarks of their respective companies.

References

1. Stuecheli, J.: Next Generation POWER microprocessor, HC25 A Symposium on High Performance Chips, Palo Alto, CA, 25–27 Aug 2013
2. Mericas, A. et al.: IBM POWER8 performance features and evaluation. IBM J. Res. Dev. **59** (1), 6:1–6:10 (2015)
3. Sinharoy, B., et al.: IBM POWER8 processor core microarchitecture. IBM J. Res. Dev. **59**(1), 2:1–2:21 (2015)
4. Sinharoy, B., et al.: Advanced features in IBM POWER8 systems. IBM J. Res. Dev. **59**(1), 1:1–1:18 (2015)

5. http://www.infoworld.com/article/2607963/security/5-takeaways-from-verizon-s-2014-data-breach-investigations-report.html
6. http://www-935.ibm.com/services/us/en/it-services/security-services/cost-of-data-breach/
7. Federal Register Notice 02-21599, Announcing Approval of Federal Information Processing Standard (FIPS) Publication 180-4, Secure Hash Standard (SHS); a Revision of FIPS 180-3. https://federalregister.gov/a/2012-5400
8. Palmer, J.: The Intel®8087 numeric data processor. In: Proceedings of the 7th Annual Symposium on Computer Architecture (ISCA '80), pp. 174–181. ACM, New York, 1980
9. http://www-03.ibm.com/press/us/en/pressrelease/20210.wss
10. http://en.wikipedia.org/wiki/CUDA
11. http://www-304.ibm.com/webapp/set2/sas/f/capi/CAPI_FlashWhitePaper.pdf
12. IBM Systems Magazine, Power Systems—May 2015, p. 24

Chapter 7
ITRS 2028—International Roadmap of Semiconductors

Bernd Hoefflinger

Abstract In CHIPS 2020, we based our discussion on the 2009 edition of the ITRS, looking forward to 2024. Its 5-year predictions for 2014 have been surpassed by the product introduction of 16 nm Flash (prediction: 20 nm), and the predictions have been reached for processors in 2014 with 24 nm. The 2013 edition has pushed back some of the 2024 data. For comparison, the 2013–2014 reality shows very clearly, that functional chip products are settling at 14 nm for SRAM and Flash, 20 nm for DRAM, and 24 nm for processors. The aggressive data in the ITRS, 5 nm in 2028, are hard to support, with less than one doping atom in the transistor channel and no large-scale lithography in sight for <7 nm.

7.1 General Observations

The International Technology Roadmap of Semiconductors (ITRS) continues its aggressive "one-dimensional" nanometer strategy with

(a) defining nanometer "nodes" out to 1.x nm that no longer have any correlation with either transistor channel-lengths or gate half-pitch, respectively first-metal half-pitch (Table 7.1),
(b) postulating min. metal half-pitch scaling progress (Table 7.1) with math formulas that have no correlation with scientific/technical publications,
(c) using standard lists of challenges in reaching the scaling projections.

The relevance of such a nanometer roadmap becomes increasingly limited. As the 2013 report observed in its introduction, in hindsight, all long-term projections beyond 5 years (and certainly the 15-year time span of the bi-annual editions of the roadmap) had to be reduced again and again, as is particularly evident in the projections for the maximum clock frequency as shown in Fig. 7.1 [3].

B. Hoefflinger (✉)
5 Leonberger Strasse, 71063 Sindelfingen, Germany
e-mail: bhoefflinger@t-online.de

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_7

Table 7.1 Master plan of critical parameters, 2013 edition

Year	2015	2020	2025	2028
Node (nm)	10	4	1.8	?
Logic 1/2 pitch (nm)	32	20	10	7
2D Flash 1/2 pitch (nm)	15	10	8	8
DRAM 1/2 pitch (nm)	24	15.5	10	7.7
FinFET 1/2 pitch (nm)	24	13.5	7.5	5.3
Fin width (nm)	7.2	6.3	5.4	5.0
6T SRAM cell area (nm^2)	6 × 10^4	2 × 10^4	6 × 10^3	3 × 10^3
NAND flash (b/chip)	128/256 Gb	512 Gb/1T	2T/4T	4T/8T
Flash layers	16–32	40–76	96–192	192–384
DRAM (Gb/chip)	8	24	32	32
Wafer diameter (mm)	300	450	450	450
VDD (V)	0.83	0.75	0.68	0.64
CV/I (ps)	0.65	0.5	0.4	0.3

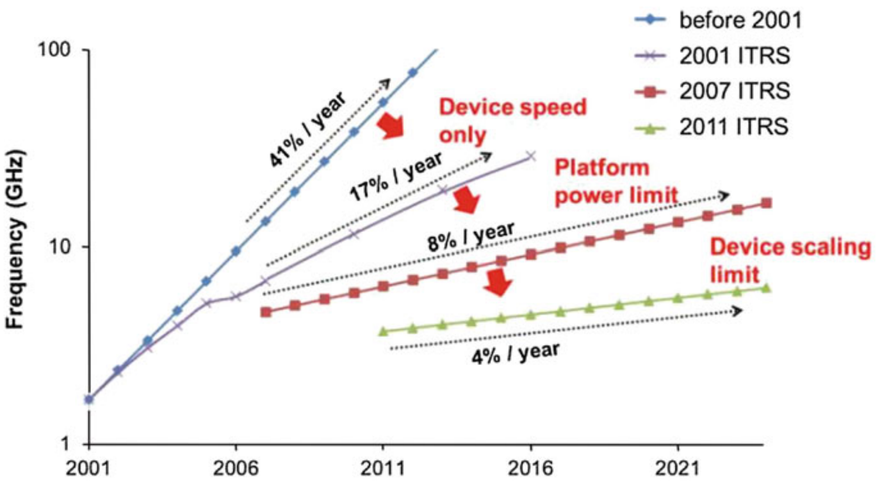


Fig. 7.1 Corrections of the Roadmap for max. clock frequency between 2001 and 2013 from 41 %/year to 4 %/year [3] *Source* Sematech

Projecting clock frequencies >10 GHz in 2000 had limited value, when it was already evident [1] that the limit would be <5 GHz because of limits on operating voltage, max. currents, over-estimated because of simplistic models, and because of RC delays of interconnects. Bandwidths- and density-estimates in the 2009 roadmap received a review in [2], and reality obviously was included in the state-of-the-art starting data of the 2013 edition.

7.2 ORTC—Overall Roadmap Technology Characteristics

The important ORTC forecast table shows new labels and new characteristics, which have become critical in the assessment of progress. Table 7.1 is a condensed adaptation and interpolation regarding the 2020 column. As to the critical items:

The simplistic node-naming has been maintained, although it no longer has any correlation with the minimum features or channel-lengths of transistors on any chip at that node. Therefore, any one of these node names is identified in the four following lines by its 1/2 pitch, (line + space)/2, for M1 in logic, rows of NAND transistors in 2D Flash memory, rows of transistors in DRAM, and minimally spaced fins of FinFET's, respectively. The pace of scaling on paper has been slowed to 70 % in 4 years or 50 % in 8 years, respectively. Final limits are stated in the 2013 report for 2D NAND Flash at 12 nm and for DRAM at 14 nm, presumably, channel length, but contrasting with the values in the table, anyway. To judge the relevance of the data in the table, we should have in mind that

- **In a Si cube of (10 nm)^3, a doping level of 10^{18}/cm^3 means just 1 active p- or n-type atom.**

Since the transistor characteristics are determined by these dopant-atom numbers N within a channel and since their standard deviation is (N)$^{1/2}$, any such numbers N < 10 to 50 make such transistors useless for large-scale integration. This observation is one reason why any of the scaled data, at least beyond 2020, have a limited relevance.

7.3 System Drivers

The System-Drivers Summary in the 2013 ITRS report is governed by the

- Design-Capability Gap:

Although dimensional scaling advanced, at least until 2013, this progress could not be designed into an equivalent progress in transistor density. This statement does not even consider the additional negative effect of scaling on transistor variability.

The design-capability gap is widened further by the handicapped scaling of all Metal pitches due to resistance, granularity, crosstalk and manufacturing problems. 3D integration is mentioned as a relief, however, only in the manner of the vertical poly-Si NAND flash, and not in the sustainable, monolithic 3D strategy, as presented in Chap. 3 in this book.

Admitting that geometry scaling effectively does not offer any density, cost or performance advantages, the report generated the DES = “Design Equivalent Scaling” as the expected performance improvements “per node” by

- Error-correcting codes,
- Lithographic-patterning-related design rules,
- Adaptive voltage and frequency scaling,
- Clock gating,

in other words, “engineering cleverness”, advocated by Gordon Moore as early as 1975 to maintain the Roadmap.

7.4 PIDS—Process Integration, Devices and Structures

This part states the challenges for

- Logic
- DRAM
- Non-volatile Memory.

Its tables of difficult challenges are organized in near-term, 2014–2020, and long-term, 2021–2028.

Immanent scaling limits are quoted everywhere, and the leading hit-words for progress are:

- Multi-gate transistors,
- Gate insulators with a high dielectric constant,
- III–V materials for transistor channels,
- Vertical transistor stacks for NAND Flash NV memory.

The new no. 1 issue is the reliability of devices and circuits, which suffer from variability, ageing, and breakdown related to further reduction of the volume of devices and their interconnects.

7.5 ERD—Emerging Research Devices

The challenges listed in this chapter are the same as in the other chapters like PIDS and SD. No emerging devices are mentioned other than the hit-words in PIDS (see Sect. 7.4). The alternative demand for memories is the replacement of SRAM and Flash by 2018 without any suggestions. No short-term incorporation of III–V channels is envisioned.

7.6 Interconnects

The goal for interconnects on-chip is Tb's per second at the energy level of fJ/b. However, it is stated that no tangible progress has been made between 2009 and 2013 due to basic material limitations, both regarding the metal layers as well as the isolation layers. Therefore, as detailed in Chap. 5, the energy levelled off at ~1 pJ/b. No solutions were found with relative dielectric constants <2. A partial remedy was introduced with air gaps in NAND Flash. The potential of 3D integration is quoted regarding through-silicon vias, but there are no indications of the potential of monolithic 3D integration (see Chap. 3) or of directed self-assembly (DSA) as techniques to fundamentally shorten the interconnects.

7.7 RF-AMS: Radio-Frequency and Analog-Mixed-Signal Technologies

The continuing progress of THz transistors leads to optimistic projections for frequency limits.

Figure 7.2 and high-frequency power-amplification capabilities. The sustained performance level of Silicon-Germanium transistors is proof of the unique significance of this central part of the periodic table (Fig. 7.3).

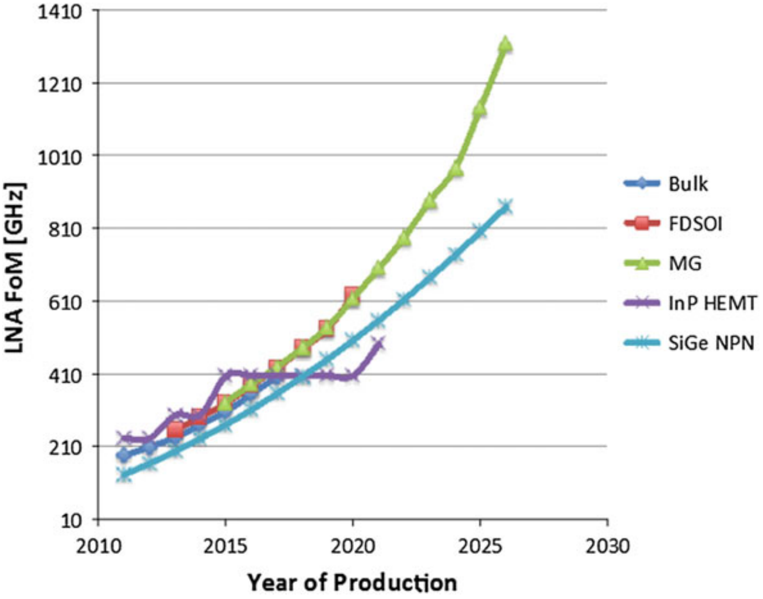


Fig. 7.2 Unity-gain frequency figure-of-merit of THz transistors [3]

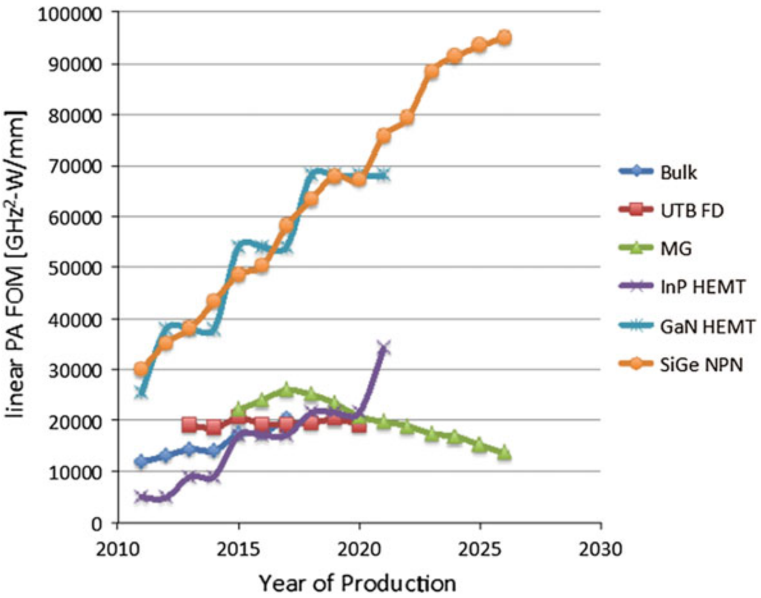


Fig. 7.3 Power-amplification figure-of-merit of RF transistors [3]

7.8 Conclusion

The ITRS has been under pressure at least since 2010 because of its "one-dimensional" exponential-growth philosophy. It had and has no energy- and no monolithic-3D-strategy. The advent of 3D chip stacks for DRAM and Flash memory since 2006 came as a surprise to save Moore's law in the face of the continuous down-ward corrections of progress on the ITRS. Nevertheless, the ITRS has had the unique effect of focusing development resources in the semiconductor industry.

It could continue to play this role, if future editions of the ITRS would concentrate on a holistic strategy for monolithic and heterogeneous 3D integration with energy efficiency of nanoelectronics as milestones.

References

1. Hoefflinger B.: Chips 2020—Ein Ausblick in die Halbleiterwelt von übermorgen, ELEKTRONIK, Heft 1/2000, pp. 10 ff., WEKA Medien, Jan 2000
2. ITRS.: In: Hoefflinger, B. (ed.) Chapter 7 in CHIPS 2020—A Guide to the Future of Nanoelectronics, pp. 37–93, Springer, Berlin (2012). doi:10.1007/978-3-642-23096-7_7
3. www.itrs.net/reports/2013-edition

Chapter 8
Nanolithographies

Bernd Hoefflinger

Abstract Nanolithography, the patterning of hundreds to thousands of trillions of nano structures on a silicon wafer, determines the direction of future nanoelectronics. The technical feasibility and the cost-of-ownership for technologies beyond the mark of 22 nm lines and spaces (half pitch) is evaluated for Extreme-Ultra-Violet (EUV = 13 nm) lithography and Multiple-Electron-Beam (MEB) direct-write-on-wafer technology. The 32 nm lithography-of-choice, Double-Patterning, 193 nm Liquid-Immersion Optical Technology (DPT) can be extended to 22 nm with restrictive design patterns and rules, and it serves as a reference for the future candidates. Burn Lin, who provided the content for this chapter in CHIPS 2020 of 2012, asked the Editor to be the contacting author for this updated chapter.

EUV lithography operates with reflective mirror optics and 4× multi-layer reflective masks. Due to problems with a sufficiently powerful and economical 13 nm source, and with the quality and lifetime of the masks, the introduction of EUV into production has been shifted to 2016.

MEB direct-write lithography has demonstrated 16 nm half-pitch capability, and it is attractive, because it does not require product-specific masks, and electron optics as well as electron metrology inherently offer the highest resolution. However, an MEB system must operate with more than 10,000 beams in parallel to achieve a throughput of ten 300 mm wafers per hour. A data volume of Peta-Bytes per wafer is required at Gigabit rates per second per beam. MEB direct-write would shift the development bottleneck away from mask technology and infrastructure to data processing, which may be easier.

The R&D expenses for EUV have exceeded those for MEB by two orders of magnitude so far. As difficult as a cost model per wafer layer may be, it shows basically that, as compared with DPT immersion optical lithography, EUV would be more expensive by up to a factor of 3, and MEB could cost less than present

B. Hoefflinger (✉)
5 Leonberger Strasse, 71063 Sindelfingen, Germany
e-mail: bhoefflinger@t-online.de

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_8

optical nanolithography. From an operational point-of-view, power is a real concern. For a throughput of 100 wafers per hour, the optical system dissipates 200 kW, the MEB system is estimated at 170–370 kW, while the EUV system could dissipate between 2 and 16 MW.

This outlook shows that the issues of minimum feature size, throughput and cost of nanolithography require an optimisation at the system level from product definition through restrictive design and topology rules to wafer-layer rules in order to control the cost of manufacturing and to achieve the best product.

Optical patterning through masks onto photosensitive resists proceeded from minimum features of 20 mm in 1970 to 22 nm in 2010. The end of optical lithography was initially predicted for 1985 at feature sizes of 1 mm, and hundreds of millions of dollars were spent in the 1980s to prepare X-ray proximity printing as the technology of the future, which has not happened. The ultimate limit of optical lithography was moved in the 1990s to 100 nm, initiating next-generation-lithography (NGL) projects to develop lithography technology for sub-100 nm features and towards 10 nm. This chapter is focused on two of these NGL technologies: *EUV* (extreme-ultraviolet lithography), and *MEB* (multiple-electron-beam lithography). These are compared with the ultimate optical lithography, also capable of delineating features of similar size: *DPT* (double-patterning, liquid-immersion lithography).

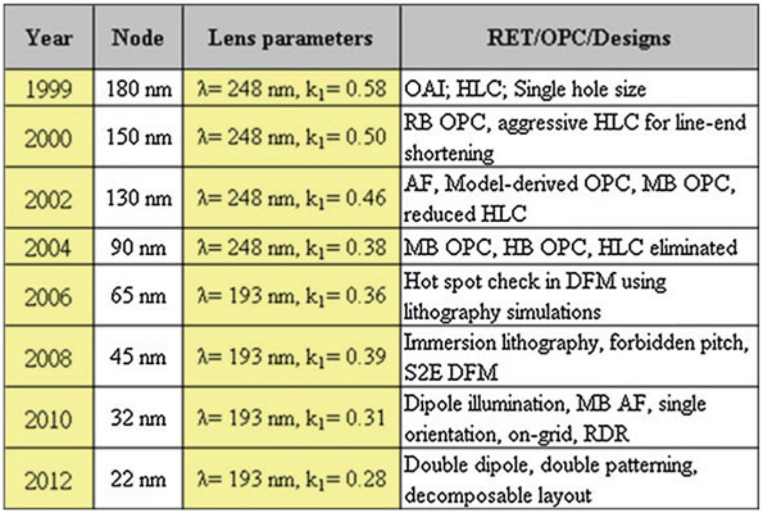
8.1 The Progression of Optical Lithography

Optical lithography has supported Moore's Law and the ITRS (Chap. 7) for the first 50 years of microelectronics. The progression in line-width resolution followed the relationship

$$\text{Resolution} = k_1 \frac{\lambda}{NA}.$$

Shorter-wavelength optical and resist systems, and the continuing increase of the numerical aperture (NA) as well as reduction of the resolution-scaling factor k_1 provided a remarkable rate of progress, as shown by the data on the period 1999–2012 (Figs. 8.1 and 8.2).

The medium for optical patterning was air until 2006, when $k_1 = 0.36$ at the 193 nm line of ArF laser sources was reached. The single biggest step forward was made when the lens–wafer space was immersed in water, and the NA effectively increased by 1.44× according to the refractive index of water at 193 nm wavelength, so that, with the inclusion of restricted design rules and more resolution enhancement techniques (RETs), the 32 nm node could be reached in 2010.



Year	Node	Lens parameters	RET/OPC/Designs
1999	180 nm	λ= 248 nm, k_1= 0.58	OAI; HLC; Single hole size
2000	150 nm	λ= 248 nm, k_1= 0.50	RB OPC, aggressive HLC for line-end shortening
2002	130 nm	λ= 248 nm, k_1= 0.46	AF, Model-derived OPC, MB OPC, reduced HLC
2004	90 nm	λ= 248 nm, k_1= 0.38	MB OPC, HB OPC, HLC eliminated
2006	65 nm	λ= 193 nm, k_1= 0.36	Hot spot check in DFM using lithography simulations
2008	45 nm	λ= 193 nm, k_1= 0.39	Immersion lithography, forbidden pitch, S2E DFM
2010	32 nm	λ= 193 nm, k_1= 0.31	Dipole illumination, MB AF, single orientation, on-grid, RDR
2012	22 nm	λ= 193 nm, k_1= 0.28	Double dipole, double patterning, decomposable layout

Fig. 8.1 Lens and reticle parameters, OPC, and RETs employed for the technology nodes from 1999 to 2012 [1]. OPC: Optical proximity correction, OAI: Off-axis illuminatiom, HLC: Handcrafted layout compensation, RB OPC: Rule-based OPC, MB OPC: Model-based OPC, HB OPC: Hybrid-based OPC, AF: Assist features, DFM: Design for manufacturability, S2E DFM: Shape-to-electric DFM, MB AF: Model-based assist features, RDR: Restricted design rules. © 2009 IEEE

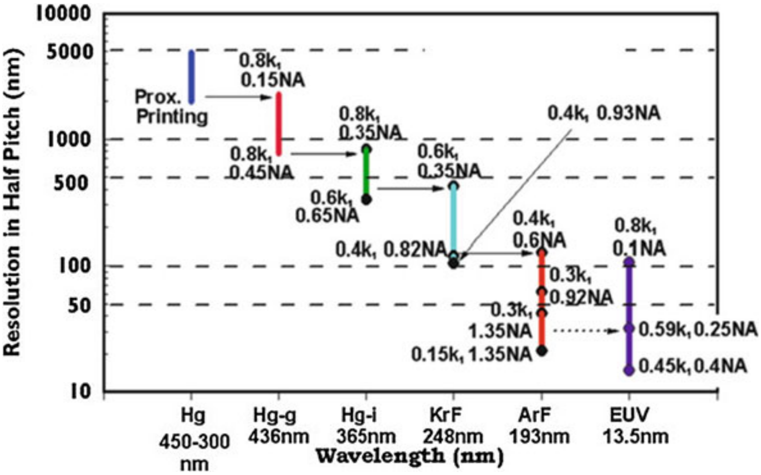


Fig. 8.2 Resolution for half pitch [(linewidth + space)/2] as wavelength reduction progresses [1]. © 2009 IEEE

With the introduction of DPT, the mask cost is doubled, the throughput is halved, and the processing cost more than doubled, making the total wafer production cost high and causing serious economic effects. Even so, multiple patterning technology (MPT) has been introduced to reach the 16 nm node.

A 2010 state-of-the-art optical-lithography system with ArF water-immersion DPT can be characterized by the following parameters:

Minimum half pitch	22 nm at $k_1 = 0.3$, $\lambda = 193$ nm, NA = 1.35
Single-machine overlay	<2 nm
Reticle magnification	4×
Throughput	~ 175 wph (300 mm diameter wafer)
Price	~ €40 M
Related equipment	~ US$15 M
Footprint	~ 18 m^2 including access areas
Wall power	220 kW

This most advanced equipment is needed for the critical layers in a 32 nm process:

- Active
- Gate
- Contact
- Metal 1–*n*
- Via 1–*n*.

These layers form the critical path in the total process, which can involve more than 10 metal layers resulting in a total mask count >40. Liquid-immersion DPT can be pushed to 22 nm with restricted design rules, limited throughput, and a significant increase in cost. It is desirable to use less expensive alternative technologies at and beyond the 22 nm node. The following sections are focused on the two major contenders for production-level nanolithography at 32 nm and beyond: EUV and MEB. Nanoimprinting holds promise for high replicated resolution. However, fabrication difficulties of the 1× templates, limited template durability, expensive template inspection, and the inherent high defect count of the nanoimprint technology prevent it from being seriously considered here [7].

8.2 Extreme-Ultraviolet (EUV) Lithography

The International SEMATECH Lithography Expert Group decided in 2000 that EUV lithography should be the NGL of choice. The EUV wavelength of 13.5 nm means a > 10× advantage in resolution over optical, as far as wavelength is concerned, and sufficient progress overall, even though the k_1 and the NA with the necessary reflective-mirror optics would be back to those of the early days with optical lithography. The system principle is shown in Fig. 8.3.

The 13.5 nm radiation is obtained from the plasma generated by irradiating a Sn droplet with a powerful CO_2 laser. This is realized with an efficiency of 0.2–0.5 % from CO_2 laser irradiation. The numbers in the figure illustrate the power at each component along the path of the beam, required to deliver 1 mJ/cm^2 of EUV light at the wafer at the throughput of 100 wph (wafers per hour). About 0.1 % of the

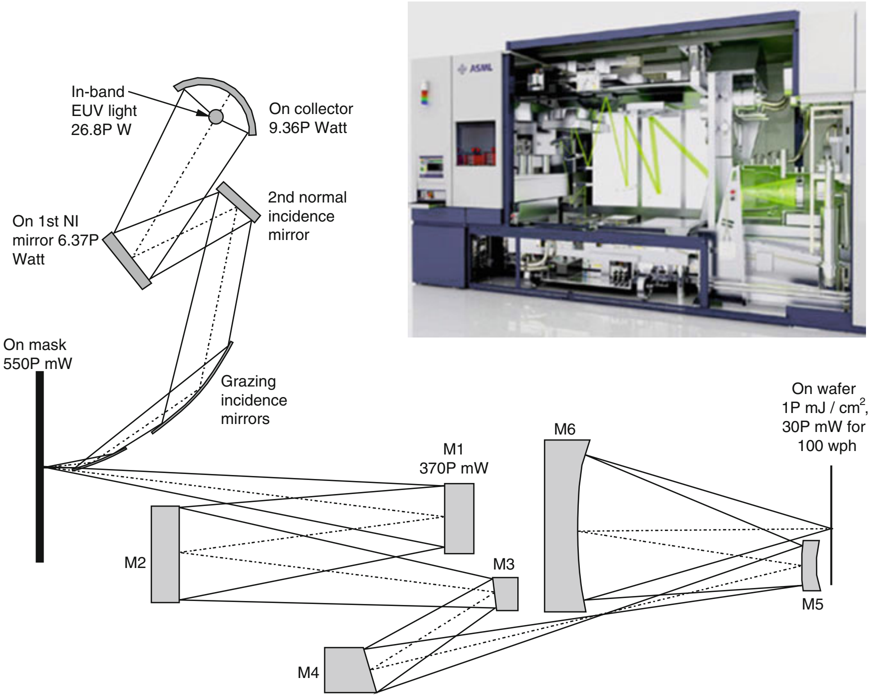


Fig. 8.3 EUV illuminator and imaging lens. It is an all-reflective system [1]. © 2009 IEEE

EUV power arrives on the wafer. Evidently, the overall power efficiency from the laser to the wafer is extremely small. This poses a serious problem: If we consider a resist sensitivity of 30 mJ/cm^2 at 13.5 nm, hundreds of kilowatts of primary laser power are needed.

Other extreme requirements [2] are:

- Roughness of the mirror surfaces at atomic levels of 0.03–0.05 nm, corresponding pictorially to a 1-mm roughness over the diameter of Japan.
- Reflectivity of mirrors established by 50 bilayers of MoSi/Si with 0.01 nm uniformity.
- Reticle cleanliness in the absence of pellicles and reticle life at high power densities.
- Reticle flatness has to be better than 46.5 nm, which is 10× better than the specification for 193 nm reticles.
- New technology, instrumentation, and infrastructure for reticle inspection and repair.

Nevertheless, EUV lithography has received R&D funding easily in the order of US$1 billion by 2010. Two alpha tools have been installed, producing test structures like those shown in Fig. 8.4.

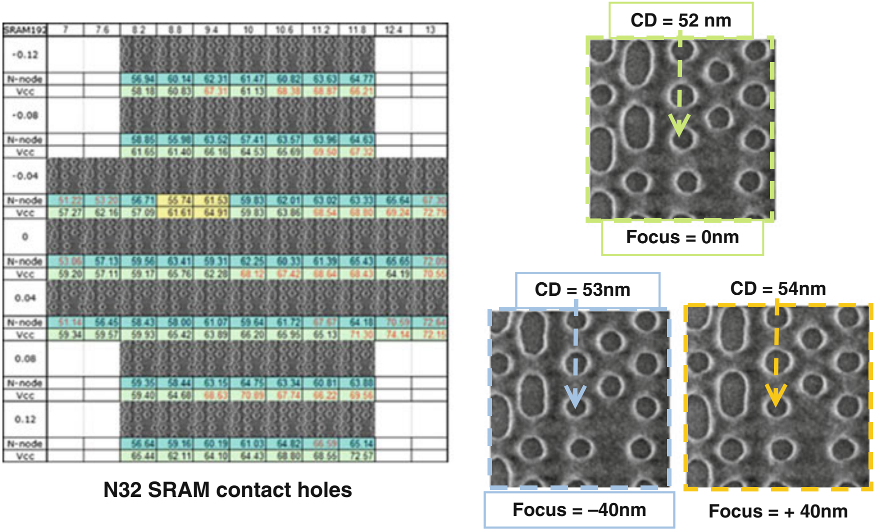

SRAM192
7
7.6
8.2
8.8
9.4
10
10.6
11.2
11.8
12.4
13
-0.12
N-node 56.94 60.14 62.31 61.47 60.82 63.63 64.77
Vcc 58.18 60.83 67.31 61.13 68.38 68.87 66.21
-0.08
N-node 58.85 55.98 63.52 57.41 63.57 63.96 64.63
Vcc 61.65 61.40 66.16 64.53 65.69 69.50 67.32
-0.04
N-node 51.22 53.20 56.71 55.74 61.53 59.83 62.01 63.02 63.33 65.64 67.30
Vcc 57.27 62.16 57.09 61.61 64.91 59.83 63.86 68.54 68.80 69.24 72.79
0
N-node 53.06 57.13 59.56 63.41 59.31 62.25 60.33 61.39 65.43 65.65 72.09
Vcc 59.20 57.11 59.17 65.76 62.28 68.12 67.42 68.64 68.43 64.19 70.55
0.04
N-node 51.14 56.45 58.43 58.00 61.07 59.64 61.72 67.67 64.18 70.59 72.64
Vcc 59.34 59.57 59.93 65.42 63.89 66.20 65.95 65.13 71.30 74.14 72.15
0.08
N-node 59.35 58.44 63.15 64.75 63.34 60.81 63.88
Vcc 59.40 64.68 68.63 70.89 67.74 66.22 69.56
0.12
N-node 56.64 59.16 60.19 61.03 64.82 66.59 65.14
Vcc 65.44 62.11 64.10 64.43 68.80 68.55 72.57
N32 SRAM contact holes
CD = 52 nm
Focus = 0nm
CD = 53nm
Focus = –40nm
CD = 54nm
Focus = + 40nm

Fig. 8.4 Micrographs of SRAM test structures: 32 nm contact holes with a pitch of 52–54 nm [1]. © 2009 IEEE

In any event, EUV lithography puts such strong demands on all kinds of resources that, if successful, it can be used by only a few extremely large chip manufacturers or consortia of similar size and technology status. Attaining lower imaging cost than DPT cannot be taken for granted. A 2012 system can be estimated to have the following characteristics:

Minimum half pitch	21 nm at k_1 = 0.5, λ = 13.5 nm, NA = 0.32
Single-machine overlay	<3.5 nm
Reticle magnification	4×
Throughput	125 wafers (300-mm dia. wafer)
Price	€65 M
Related equipment	US$20 M (excluding mask-making equipment)
Footprint	50 m^2 including access areas
Wall power	750–2000 kW

If we consider that a Gigafactory requires 50 such systems for total throughput, we can appreciate that this technology is accessible to a few players only. Further data on these issues will be discussed in Sect. 8.4.

8.3 Multiple-Electron-Beam (MEB) Lithography

Electron-beam lithography has been at the forefront of resolution since the late 1960s, when focused electron beams became the tools of choice to write high resolution masks for optical lithography or to write submicrometer features directly on semiconductor substrates. As a serial dot-by-dot exposure system, it has the reputation of being slow, although orders-of-magnitude improvements in throughput were achieved with

- Vector scan (about 1975)
- Variable shaped beam (1982)
- Cell projection through typically 25× Si stencil masks (1995)

A micrograph of a 1997 stencil mask with 200 nm slots in 3 μm Si for the gates in a 256 Mb DRAM is shown in Fig. 8.5. Electron beams have many unique advantages:

- Inexpensive electron optics
- Large depth-of-field
- Sub-nanometer resolution
- Highest-speed and high-accuracy steering
- Highest-speed On-Off switching (blanking)
- Secondary-electron detection for position, overlay, and stitching accuracy
- Gigabit/s digital addressing

The systems mentioned so far were single-beam systems, using sophisticated electron optics. A scheme to produce a large array of parallel nanometer-size electron beams is used in the MAPPER system [3], as shown in Fig. 8.6, where 13,000 beams are created by blocking the collimated beam from a single electron source with an aperture plate. The beam-blanker array acts upon the optical signal

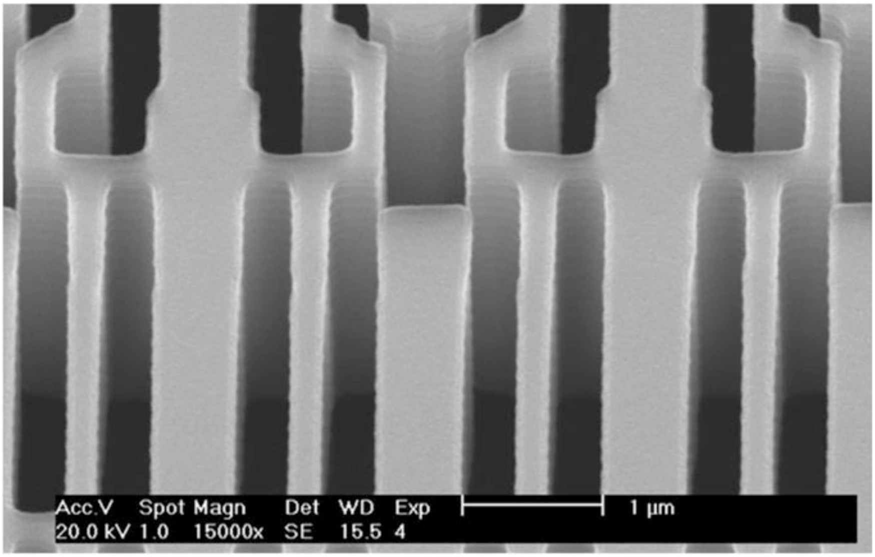


Fig. 8.5 SEM micrograph of a Si stencil mask with 200 nm slots in 3 μm Si for the gates in a 256 Mb DRAM (Courtesy INFINEON and IMS CHIPS)

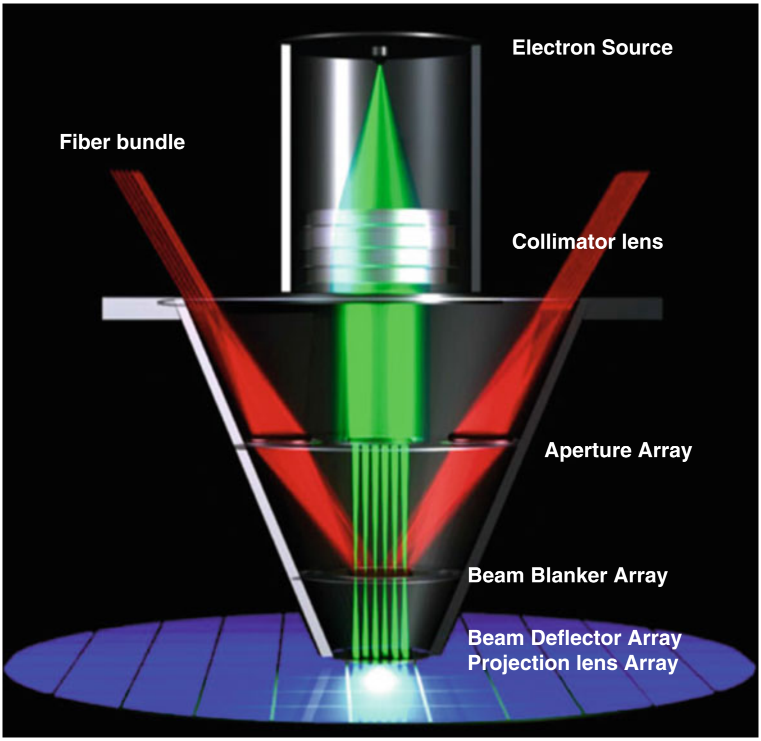


Fig. 8.6 The MAPPER MEB system [1]. © 2009 IEEE

from the fiber bundle to blank off the unwanted beams. The On-beams are Fig. 8.5 subsequently deflected and imaged. All electron optics components are made with the MEMS (micro-electro-mechanical system) technology, making it economically feasible to achieve such a large parallelism [6].

The MAPPER write scheme is shown in Fig. 8.7. For a 45 nm half-pitch system, the 13,000 beams are distributed 150 μm apart in a 26 × 10 mm^2 area. The separation ensures negligible beam-to-beam interaction. Only the beam blur recorded in the resist due to forward scattering calls for proximity correction. The beams are staggered row-by-row. The staggering distance is 2 μm. The deflection range of each beam is slightly larger than 2 μm to allow a common stitching area between beams, so that the 150 μm spaces in each row are filled up with patterns after 75 rows of beam pass through. Each beam is blanked off for unexposed areas [4].

Massive parallelism is not allowed to be costly. Otherwise the cost target to sustain Moore's law still cannot be met. One should forsake the mentality of using the vacuum-tube equivalent of conventional e-beam optics and turn to adopting the integrated-circuit-equivalent highly-parallel electron optics, whose costs are governed by Moore's law. With cost of electron optics drastically reduced, the dominant costs are now the digital electronics, such as CPU (central processing unit), GPU (graphics processing unit), FPGA (field programmable gate array), and

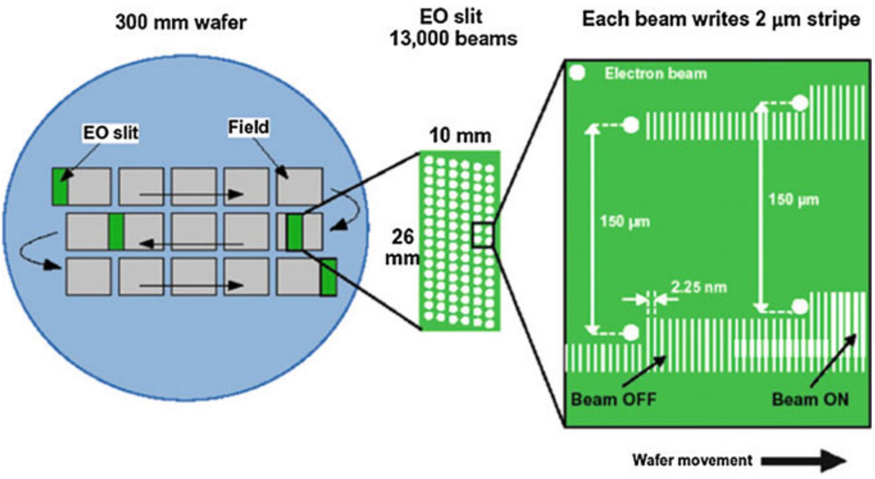


Fig. 8.7 MAPPER writing scheme (EO: Electron optics) [1]. © 2009 IEEE

DRAM, required to run the massively parallel, independent channels. *For the first time, the cost of patterning tools is no longer open-ended but rather a self-reducing loop with each technology-node advance*.

Another multiple-beam system uses a conventional column but with a programmable reflective electronic mask called a digital pattern generator (DPG) as shown in Fig. 8.8.

The DPG is illuminated by the illumination optics through a beam bender. The DPG is a conventional 65 nm CMOS chip with its last metal layer turned towards the illuminating beam. The electrons are decelerated to almost zero potential with the Off-electrons absorbed by the CMOS chip while the On-electrons are reflected back to 50 keV and de-magnified by the main imaging column to expose the wafer. Even though there are between 1 M and 4 M pixels on the DPG, the area covered is too small compared to that of an optical scanner. The acceleration and deceleration of a rectilinear wafer stage to sustain high wafer throughput are impossible to reach. Instead, a rotating stage to expose six wafers together can support high throughput with an attainable speed. Again, all components are inexpensive. The data path is still an expensive part of the system. The cost of the system can be self-reducing [8].

The fascination with these systems, beyond the fundamental advantage that they provide *maskless nano-patterning*, lies in the fact that MEB lithography directly benefits in a closed loop from the

- Advances in MEMS and NEMS (micro- and nano-electro-mechanical systems) technology for the production of the aperture and blanking plates as well as arrays of field-emission tips, and
- Advances in high-speed interfaces to provide terabits/s to steer the parallel beams.

MEB direct-write lithography demonstrated 16 nm resolution (half-pitch) in 2009.

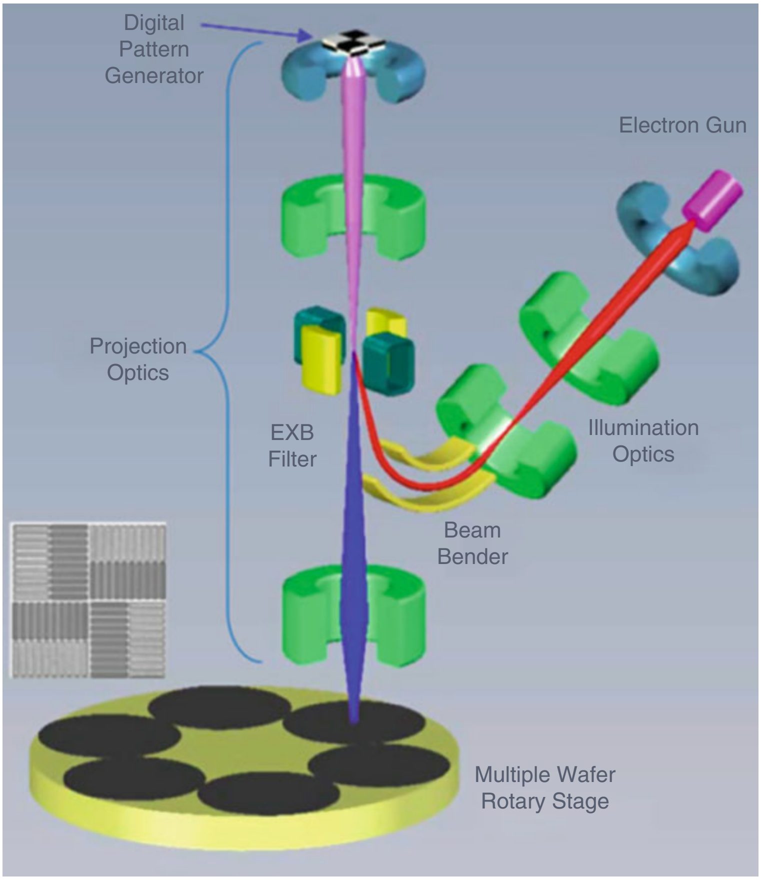


Fig. 8.8 Second generation reflective electron-beam lithography (REBL) system. EXB: charged particle separator $E \times B$

For high-volume manufacturing (HVM), the basic MEB chamber with 13,000 columns and 10 wph has to be clustered into 10 chambers to reach 100 wph, as illustrated in Fig. 8.9.

An MEB system with 10 chambers has about the same footprint as the 193 nm optical immersion scanner. A 2012 assessment would be as follows:

Minimum linewidth	22 nm
Minimum Overlay	7 nm
No reticle	Maskless
Number of e-beams	130,000
Throughput	100 (300 mm) wph
Data rate	520 Tbit/s
Estimated Price	US$50 M
Footprint	18 m^2

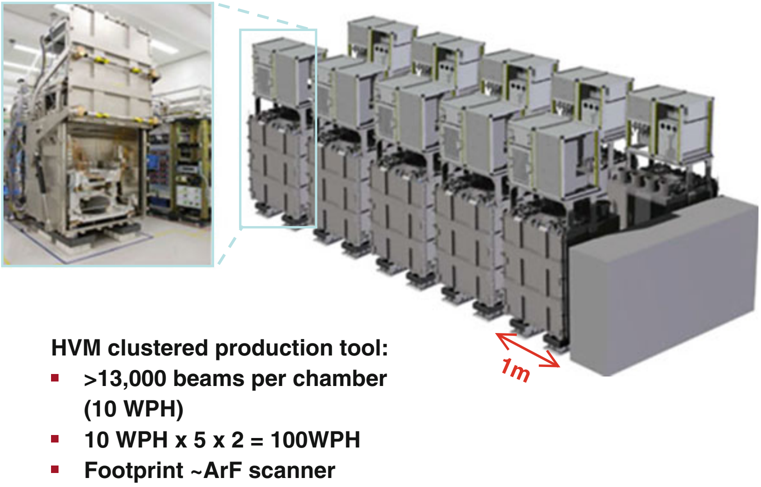

1m
HVM clustered production tool:
- >13,000 beams per chamber (10 WPH)
- 10 WPH x 5 x 2 = 100WPH
- Footprint ~ArF scanner

Fig. 8.9 MEB maskless-lithography system for high-volume manufacturing [1]. © 2009 IEEE

The following calculation of data rates illustrates the challenges and possible system strategies for these maskless systems [5]:

- Total pixels in a field (0 and 1 bitmap)

  = (33 mm × 26 mm/2.25 nm × 2.25 nm) × 1.1 (10 % over scan)
  = 190 Tbits
  = 21.2 TBytes

- A 300-mm wafer has ~90 fields
- 10 wph ⇒ <0.9 s/field
- 13,000 beams ⇒ data rate >3.75 Gbps/beam!
- In addition to data rate, also challenge in data storage and cost
- Use more parallelism to reduce data rate

MEB lithography by 2010 has seen R&D resources at only a few percent compared with those for EUV, and no industry consensus has been established to change that situation. However, this may change in view of the critical comparison in the following section.

8.4 Comparison of Three Nanolithographies

The status in 2010 and the challenges for the three most viable nanolithographies for volume production of nanochips with features <32 nm are discussed with respect to maturity, cost, and factory compatibility. Table 8.1 provides a compact overview.

Table 8.1 Maturity, cost, infrastructure, and required investment

	Wafer cost	Mask cost	Infra-structure	Maturity	Investment
MEB ML2	Potentially low	0	Ready	Massive parallism needs most work	Badly needed
Immersion DPT	More than 2× SPT	Ever increasing	Ready	Mature	No question
EUV	From slightly below to much higher than DPT	Can be higher than DPT	Expensive, to be developed	Not yet	Plenty but never enough

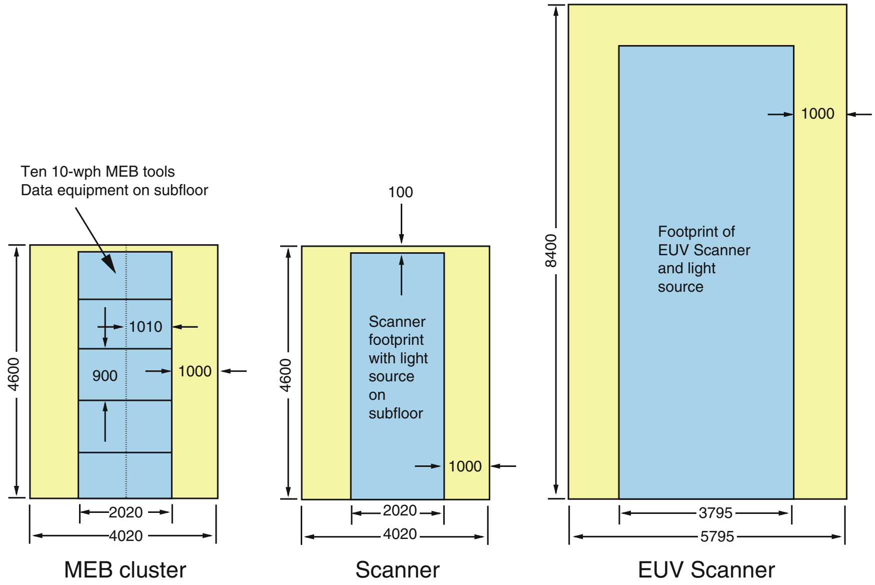

Ten 10-wph MEB tools
Data equipment on subfloor
1010
900
1000
4600
2020
4020
MEB cluster
100
Scanner footprint with light source on subfloor
1000
4600
2020
4020
Scanner
8400
1000
Footprint of EUV Scanner and light source
3795
5795
EUV Scanner

Fig. 8.10 Footprint comparison of MEB cluster, optical scanner, and EUV scanner [1]. © 2009 IEEE

Single-patterning technology optical immersion lithography (SPT) is taken as a reference. Given the 2010 status, neither EUV nor MEB is mature for production. Lithography has become the show-stopper for scaling beyond 22 nm so that serious investment and strategy decisions are asked for. Cleanroom footprint in Fig. 8.10, cost per layer in Table 8.3 and electrical power in Table 8.2 are used to provide guidance for the distribution of resources.

We see that the footprint for the throughput of 100 wph with optical and e-beam technology is about the same, while that for the EUV scanner is more than double.

Cost estimates, difficult as they may be, are important in order to identify fundamental components of cost and their improvement (Table 8.3). The numbers for

Table 8.2 Electrical power (kW) for 32 nm lithography in a 150-MW Gigafactory for 130,000 12 in. wafers/month (HVM: High-volume manufacturing) [1]

kW	Immer scanner	EUV HVM			MEB HVM	
	Supplier estimate	Supplier estimate	30 mJ/cm^2 instead of 10 mJ/cm^2	30 mJ/cm^2 resist + conservative collector and source efficiencies	Ten 10 wph columns	Share datapath
Source	89	580	1740	16,313	120	120
Exposure unit	130	169	190	190		
Datapath					250	53
Total per tool	219	749	1930	16,503	370	173
Total for 59 tools	12,921	44,191	113,870	973,648	21,830	10,222
Fraction of scanner power in fab	8.61 %	29.46 %	75.91 %	649.10 %	14.55 %	6.81 %
	130 k wafers per month 12″ fab, 150,000 kW					

Table 8.3 Cost per lithography layer relative to single-patterning (SP) water-immersion ArF lithography [1]

All costs normalized to SP exposure cost/layer	Imm SP	Imm SP	EUV goal	EUV possibility	MEB goal	MEB possibility
Normalized exposure tool cost	3.63E + 06	3.63E + 06	4.54E + 06	4.54E + 06	4.54E + 06	2.72E + 06
Normalized track cost	5.58E + 05	5.58E + 05	4.88E + 05	2.09E + 05	4.88E + 05	4.88E + 05
Raw thruput (WPH)	180	100	100	20	100	100
Normalized exposure cost/layer	1.00	1.80	2.17	10.37	2.17	1.37
Normalized consumable cost/layer	0.91	2.27	1.43	1.43	1.05	1.05
Normalized expo +consum cost/layer	1.91	4.07	3.61	11.80	3.22	2.42

optical DPT versus SPT are fairly well known. The exposure cost is a strong function of throughput, so it doubles for DPT. However, because of the extra etching step required by DPT, the combined exposure-and-processing cost more than doubles. For EUV lithography, if the development goal of the EUV tool supplier is met, the combined exposure and processing cost is slightly less than that of DPT with ArF water-immersion lithography. If the goal is not met, either due to realistic resist sensitivity or source power, a fivefold smaller throughput is likely.

It drives the total cost per layer to 3× that of optical DPT. With multiple electron beams meeting a similar throughput-to-tool-cost ratio as the EUV goal, a cost/layer of 80 % against DPT is obtained. Further sharing of the datapath within the MEB cluster could bring the tool cost down by 40 % so that overall cost/layer might drop to 60 % of DPT.

The total electrical power required in a Gigafactory just for lithography will no longer be negligible with next-generation-lithography candidates. The optical immersion scanner as a reference requires 219 kW for a 100-wph DPT tool. A total of 59 scanners to support the 130,000 wafers/month factory, would consume 8.6 % of the factory power target of 150 MW.

We saw in Sect. 8.2 that the energy efficiency of the EUV system is extremely small. Even the supplier estimate results in a power of 750 kW/tool or 44 MW for 59 tools, corresponding to 30 % of the total fab power. If the resist-sensitivity target of 10 mJ/cm^2 is missed by a factor of 3 and the actual system efficiency from source to wafer by almost another factor of 10 from the most pessimistic estimate, then the EUV lithography power in the fab rises to unpractical levels of between 141 MW and 974 MW. The cost to power these tools adds to the already high cost estimate shown in Table 8.2. It is also disastrous in terms of carbon footprint.

The overall-power estimate for MEB maskless systems arrives at a total power of 370 kW for the 10-column cluster with 130,000 beams, of which 250 kW is attributed to the datapath. This would result in 22 MW for the fab, corresponding to 16 % of the total fab power, also high compared to that of DPT optical tools. It is plausible to predict that the power for the datapaths could be reduced to 50 kW with shared data, resulting in a total of 173 kW for each cluster and a total lithography power in the plant of 6.8 % or less.

8.5 2015 Perspective on 7 nm Lithography

The crucial role of nanolithography continues to drive elaborate programs on the pro’s and con’s of lithographies for 7 nm [9].

References

1. Lin, B.J.: Limits of optical lithography and status of EUV. In: IEEE International Electron Devices Meeting (IEDM), 2009
2. Lin, B.J.: Sober view on extreme ultraviolet lithography. J. Microlith. Microfab. Microsyst **5**, 033005 (2006)
3. www.mapperlithography.com
4. Lin, B.J.: Optical lithography: here is Why. SPIE, Bellingham (2010)
5. Lin, B.J.: NGL comparable to 193 nm lithography in cost, footprint, and power consumption. Microelectron. Eng. **86**, 442 (2009)

6. Klein, C., Platzgummer, E., Loeschner, H., Gross, G.: PML2: the mask-less multi-beam solution for the 32 nm node and beyond. SEMATECH 2008 Litho Forum. www.sematech.org/meetings/archives/litho/8352/index.htm (2008). Accessed 14 May 2008
7. Lin, B.J.: Litho/mask strategies for 32 nm half-pitch and beyond: using established and adventurous tools/technologies to improve cost and imaging performance. Proc. SPIE **7379**, 1 (2009)
8. Lin, B.J.: Multiple-electron-beam (MEB) direct-write comes of age. In: SPIE Newsroom. 14 Jan 2013. doi:10.1117/2.1201212.004609
9. Lin, B.J.: Optical lithography for the 7 nm node and beyond with or without NGL. In: SPIE Conference, February 2015

Chapter 9
News on Energy-Efficient Large-Scale Computing

Barry Pangrle

Abstract The pinnacle of computing performance is building the fastest computer in the world. Advances in high-performance computing enable scientists and engineers to perform new ground-breaking research with every new generation of supercomputer in important fields like: renewable energy, climate prediction, drug development, genetics and weather prediction. Every field of science and engineering (and all that entails to the welfare of the inhabitants of our planet) benefit from having faster computational resources available. The huge impact that these machines can make in our everyday lives and futures is well known, and countries around the globe compete in this arena and view it as being a vital part of their national (as well as global) interests to continually push the frontier forward on building the world's fastest computers. This chapter looks at the challenges that lie ahead in terms of getting to Exascale (10^18 operations per second) computers and beyond and how critical it is to find solutions that enable more computing speed at ever lower costs in energy per operation.

9.1 History and Background

The first solid-state computers were built by CDC in the early 1960s where Seymour Cray, who would have a large and lasting impact on the supercomputing industry, was then working. The computer industry and supercomputing have ridden and continue to ride the wave of the advancements in semiconductor technology along the way. The exponential growth in the transistor capacity per unit-area of silicon over time (commonly referred to as Moore's Law [1]) has held for over 5 decades. Over the years, this has led to bold predictions of astonishing future capabilities or of the end of scalability and the demise of the industry. So far the naysayers have been held at bay, but it may at least in part be true that it is

B. Pangrle (✉)
Starflow Networks, Los Gatos, CA 95030, USA
e-mail: barry@pangrle.com

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_9

Table 9.1 Engineering prefixes

10^3	*K*ilo	Thousand
10^6	*M*ega	Million
10^9	*G*iga	Billion
10^12	*T*era	Trillion
10^15	*P*eta	Quadrillion
10^18	*E*xa	Quintillion
10^21	*Z*etta	Sextillion
10^24	*Y*otta	Septillion

becoming more challenging to continue at the previous pace of advancement, and in the end it may become more of an economic rather than technical hurdle going forward.

Since this chapter covers a large range of numbers and for the convenience of the reader, Table 9.1 below lists a set of commonly used engineering (and International System of Units) prefixes to indicate the size of an object in the units of reference. For example, a MegaFLOP constitutes one million floating-point operations and would be referred to as an MFLOP and one billion would be one GFLOP, etc.

Another scientifically historical part of the 1960s was the advancements in space exploration culminating with the landing of astronauts on the moon in July 1969 and then safely returning them home to earth. The first pocket calculators were invented by TI and available only 2 years before that Apollo 11 space mission, so the pocket calculators used by the Apollo astronauts, while in space, looked like the one in Fig. 9.1. A low-power unit for sure, but not a very fast computer by today's standards.

This is something to reflect upon, as today mobile chips with 500 GFLOP/s (fp32) [2] of processing power are becoming available, allowing us to carry more computing power around in our hands than was available on the entire planet back in 1969! If we take the CDC 7600, which was released in 1969 at a peak of 36 MFLOP/s and a cost of $5 Million, it would've taken nearly 14,000 CDC 7600 computers at almost $70 Billion to equal the computing power of a chip that will find its way into everyday consumer devices.

Figure 9.2 shows a number of interesting trends in supercomputing starting from back in the 1960s. First, if we draw a line fitting the points starting with the CDC 6600 in the mid-1960s, we notice that the rate of increase in computing

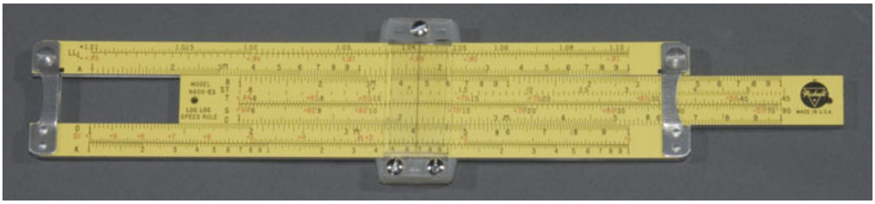
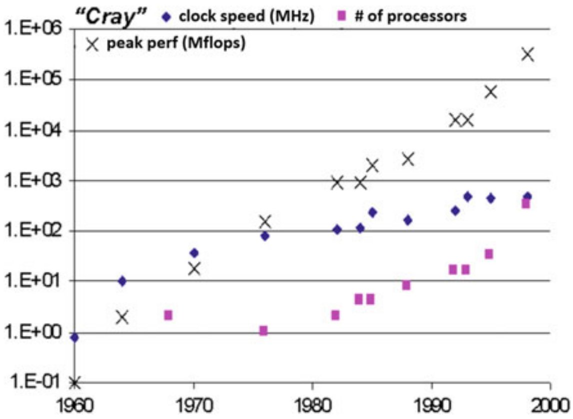


Fig. 9.1 This rule was used by the crew of Apollo 13, in April 1970 [3]

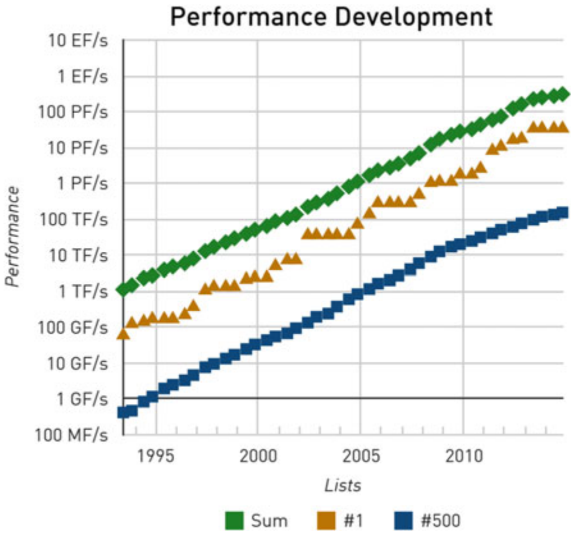
Fig. 9.2 Adapted from slide 13 of "A Seymour Cray Perspective" [4]



performance is about 1000× every 20 years or a little better than 40 % increase in performance per year. If we take 1985 as the year that the GFLOP/s barrier was crossed, then, to get to an EFLOP/s, this plot would project out to reaching that milestone somewhere around 40 years later or ~2025. A second interesting point is that, after 1980, a large portion of the performance increase is being made up by increasing numbers of processors and less reliance on the increase in clock frequencies. It's been widely known that clock frequencies across the computing industry hit an inflection around 2004, which has made the reliance on more processors per computer even more pronounced.

As we can see by the data in Fig. 9.3, the slope of the line for the #1 (fastest machine) still seems to be pretty close to 20 years per 1000x speedup. The point at

Fig. 9.3 Top500 historical supercomputer performance [5] (F = FLOP)



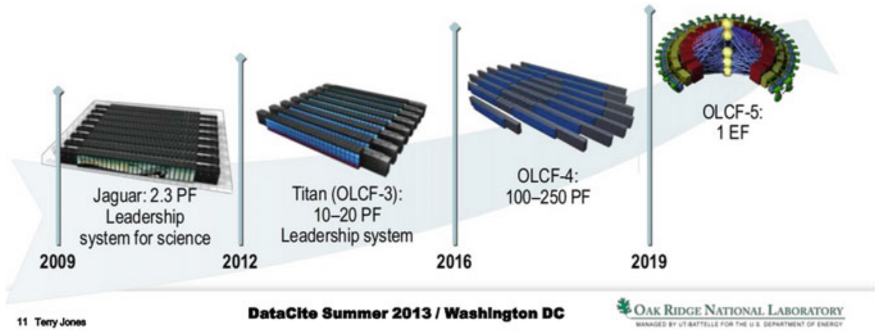


Fig. 9.4 Projected performance 2013 [6]

1 PF/s looks pretty close to 20 years after 1 TF/s, and using that as a baseline would suggest that 1 EF/s would be projected out closer to 2027. The DOE Exascale Challenge from these projections appears to be quite aggressive with an early target of 2020 to reach the 1 EF/s goal. The projections below in Figs. 9.4 and 9.5 show the OLCF-5 supercomputer moving from 2019 to 2022.

Tianhe-2 with an R_{max} of 33.862 PetaFLOP/s is currently sitting at the number 1 position on the Top500 list but has slipped from 49th to 64th on the Green500 list. It's interesting to note that, while Tianhe-2 has a peak of $\sim$55 PetaFLOP/s, the R_{max} is about 60 % of the peak value. It's quite typical for these machines to have peak scores that are significantly higher than their Rmax scores. The Rmax scores are based on the Linpack benchmark, and the Top500 site states, "In particular, the operation count for the algorithm must be 2/3 n^3 + O(n^3) double precision floating point operations." This is supposed to present a more "realistic" measure of the machines' actual performance capabilities on "real" problems. This is mentioned to help put into context the peak claims versus the R_{max} measurements that are used to rank systems on the Top500 and Green500 lists.

It's also interesting to note that the top 23 systems on the Green500 list are all heterogeneous. In the runner-up position on the Green500, from Japan, is Suiren powered by PEZY-SC many-core accelerators paired up with Intel Xeon E5-2660v2 10C 2.2 GHz processors, and in 3rd place, also from Japan, is the former top position holder TSUBAME-KFC using NVIDIA K20x GPU accelerators paired with Intel Xeon E5-2620v2 6C 2.100 GHz processors.

9.2 Energy Efficiency

The Green500 has released its November 2014 list of the top 500 most energy-efficient supercomputers, and L-CSC from the Helmholtz Center is the first supercomputer to surpass the 5 GFLOP/s/Watt barrier. The machine is another heterogeneous system and is based on AMD FirePro™ S9150 GPU accelerators

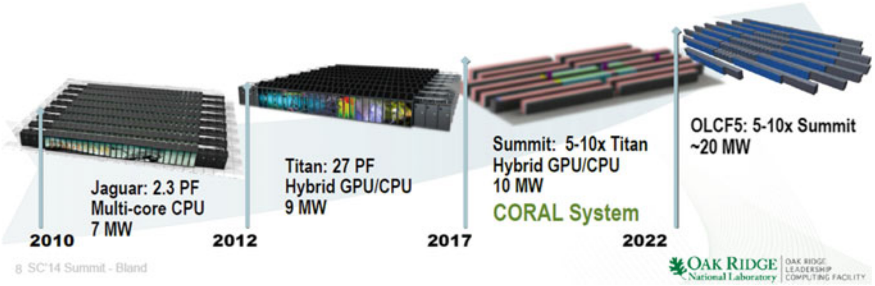
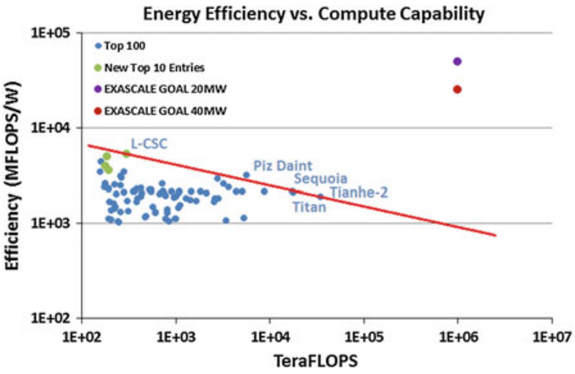


Fig. 9.5 Projected performance 2014 [7]

Fig. 9.6 Energy efficiency versus total compute capability



and Intel Xeon E5-2690v2 10C 3 GHz processors. IBM and NVIDIA have received an award of a significant chunk of a grant from the U.S. DOE for \$425 M to build two new supercomputers targeted to deliver 150–300 peak PFLOP/s. The systems are expected to be installed in 2017. These two new systems will be part of the race towards the Exascale goal (Fig. 9.5).

Looking at energy efficiency versus total compute capability, it would seem to be much harder to build a machine to top the Top500 and still do well on the Green500. Plotting the top 100 of the Green500 list in Fig. 9.6 above shows a "frontier" line in red with L-CSC now the top efficiency system. Even with this raise to the left end of the line, Piz Daint still sits above it indicating that it's perhaps pushing the energy-efficiency frontier further than its competitors. If we look at the slope on this log versus log plot, we get a slope of ~−0.216. This would indicate that, for every factor of 10 increase in system compute capability, we lose about 39 % of our efficiency or, in other words, we lose about an order of magnitude in efficiency for every 4+ orders of magnitude increase in performance. Extrapolating this out to 1 EFLOP/s would indicate that such a system today would operate at theoretically ~915 MFLOP/s/W implying a whopping 1.09 GW of system power. This is about 27x the 40 MW target. If we were to just draw a line through L-CSC and Piz Daint, the slope would indicate about a 33 % loss in efficiency for every factor of 10 increase in performance. Extrapolating from this new line gives us a projected efficiency of ~1300 MFLOP/s/W or about 19x short of the 40 MW efficiency goal at 1 ExaFLOPS. A theoretical factor of 19 could be a challenge to make up in 6 years.

9.3 Conclusions

On the plus-side for improvements going forward, the AMD, NVIDIA and PEXY-SC accelerators are all implemented in 28 nm CMOS technology, so we should be hopeful that moving the accelerators soon to smaller technology nodes will provide a significant boost in efficiencies. It will be interesting to see the efficiency ratings of the new top-end machines in 2017 to see how much farther the technology will need to progress in the final 3 years towards the Exascale goal [8–11].

References

1. Moore, G.E.: Cramming more components onto integrated circuits. Electronics, vol. 38(8), 19 Apr 1965
2. NVIDIA, Whitepaper, NVIDIA® Tegra® X1 NVIDIA'S New Mobile Superchip, January, 2015. http://international.download.nvidia.com/pdf/tegra/Tegra-X1-whitepaper-v1.0.pdf
3. Smithsonian National Air and Space Museum, Home:Collections:Objects:Slide Rule:5-inch: Pickett N600-ES:Apollo 13. http://airandspace.si.edu/collections/artifact.cfm?object=nasm_A19840160000
4. Bell, G.: "A Seymour Cray Perspective", presented at the University of Minnesota on November 10, 1997. http://research.microsoft.com/en-us/um/people/gbell/craytalk/sld001.htm
5. Top500, http://top500.org
6. Jones, T., Vazhkudai, S., Fuller, D.: DOIs and supercomputing, presented at DataCite Summer Meeting, 19–20 September 2013, Washington, DC. http://www.slideshare.net/datacite/2013-datacite-summer-meeting-introduction-adam-farquhar-datacite
7. Bland, B: Present and Future Leadership Computers at OLCF, presented at SC'14, November 17–21, 2014, New Orleans. http://on-demand.gputechconf.com/supercomputing/2014/presentation/SC405-accelerating-ornl-applications-exascale.pdf
8. 43 % of the World's Computing Power Was Manufactured Last Year, June 6, 2014, by Alice and Bob Chronicles. http://www.thedvnc.org/newsblog/2014/6/6/40-of-the-worlds-computing-power-was-manufactured-last-year
9. The World's Technological Capacity to Store, Communicate, and Compute Information, by Martin Hilbert and Priscila López, 1 April 2011, vol. 332, Science. http://www.sciencemag.org/content/332/6025/60.full.pdf
10. Exascale Programming Challenges, Sponsored by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR), Report of the 2011 Workshop on Exascale Programming Challenges, Marina del Rey, July 27–29, 2011. http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/ProgrammingChallengesWorkshopReport.pdf
11. Top Ten Exascale Research Challenges, DOE ASCAC Subcommittee Report, February 10, 2014. http://science.energy.gov/~/media/ascr/ascac/pdf/meetings/20140210/Top10reportFEB14.pdf

Chapter 10
High-Performance Computing (HPC)

Bernd Hoefflinger

Abstract Computing with nano-chip technologies spans highly different applications. The corresponding processor-chips span a range of 1000:1 in performance and in their energy-efficiency. From 2010 to 2014, their performance per product-unit almost doubled every year, while their energy efficiency doubled only every three years, so that their total power consumption quadrupled in three years. With the evidence of this energy crisis, the projections for 2015–2020 have been corrected somewhat to a global performance improvement of 20-times in 5 years, while the energy efficiency shall improve an optimistic 7-times in 5 years, which would raise the total required wall-plug power another 3-times, an unlikely scenario, accompanied by a "Green" movement. This "green" strategy lowers the energy per operation by reducing the operating voltage. However, this reduces the operations per second so much that, wherever the completion of operations in a given time is an issue, more computing units would have to be run in parallel, with a worse balance in required hardware and energy. Therefore, GREEN is not enough. A new throughput figure-of-merit sheds new light on the task of simultaneously improving performance and energy-efficiency, demanding disruptive innovations.

10.1 Highlights on Standard Processors

Standard processors are the highlights of technology scaling and on-chip complexity [1]. The general-purpose, double-precision, high-throughput processor chips of 2014 advanced to 12–16 cores at the "22 nm node" [2]. With master clocks of 5 GHz and 5 Billion transistors on a 5 cm^2 chip, memory/communication had become the no. 1 power issue as shown in Fig. 10.1.

The total share of power at the various memory levels is >40 %, processing operations need ~22 %, power management and clock distribution take up ~18 %.

B. Hoefflinger (✉)
5 Leonberger Strasse, 71063 Sindelfingen, Germany
e-mail: bhoefflinger@t-online.de

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_10

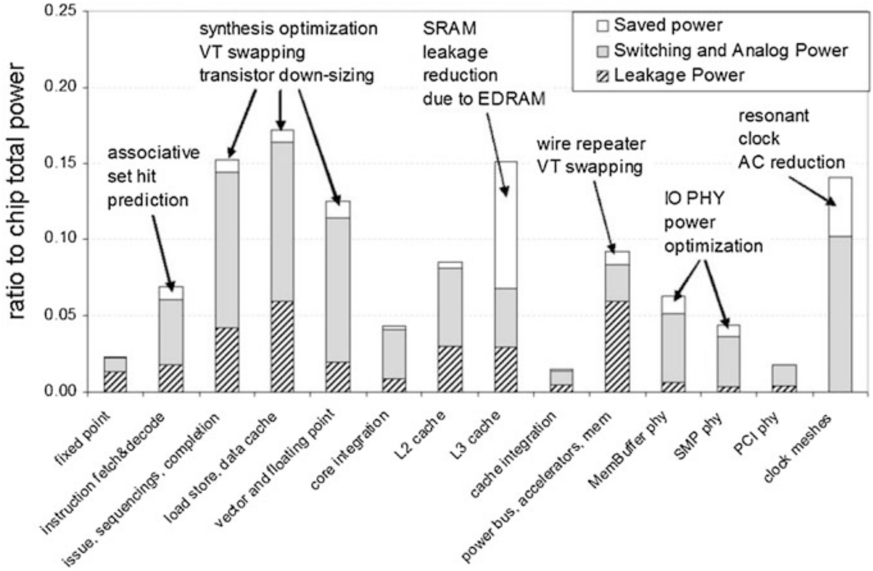


Fig. 10.1 Electric power dissipation on a 12-core CPU in 22 nm SOI technology [2]. © 2014 IEEE

The energy efficiency of these CPU's has doubled in ~3 years to ~4 MOPS/mW or, its inverse, the energy/operation, could be reduced to ~250 pJ/operation in 2014.

High-performance computing (HPC) has been driven by a doubling of the operational throughput every year or a 1000-times improvement every ten years. With the above energy progress, the performance expectations over ten years would need 100-times more power. In order to fill this power gap, multi-processor systems have been made more effective by the inclusion of dedicated, more power-efficient processors, like graphics-processor units (GPU).

10.2 Special-Purpose Processors and Energy Efficiency

DSP's (digital-signal processors), where single instructions initiate the parallel processing of multiple data (SIMD), have had their own remarkable development, and their energy efficiency is well documented by the JPEG coder described in Sect. 1.5, which achieved pJ energy levels per pixel in the coding of images. The DSP's advanced to the class of graphics processors (GPU's), and hybrid systems of CPU's and GPU's achieved energy efficiencies of 10 MOPS/mW in 2014, again, within their category, with a rate of doubling every 3 years. Large contingents of these CPU/GPU systems make up present supercomputers, treated in the following Sect. 10.3.

Figure 10.2 shows the efficiency levels achievable with special-purpose processing.

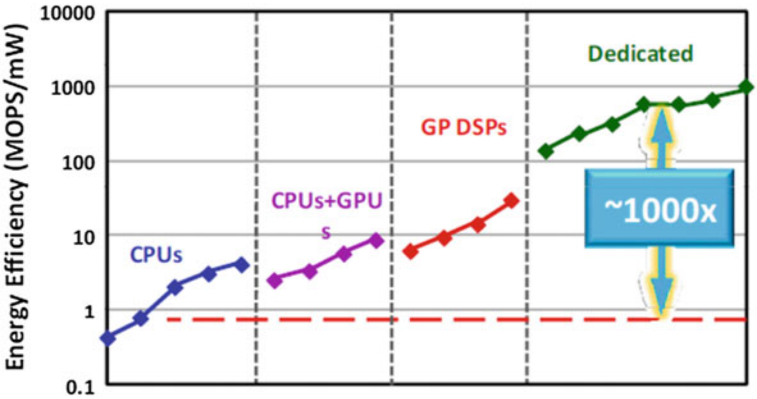
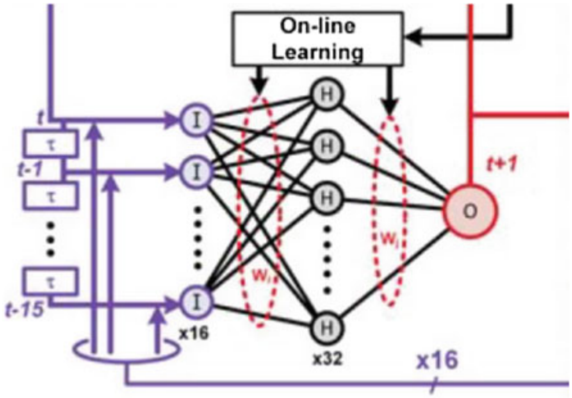


Fig. 10.2 The development of energy efficiency of processors between 2009 and 2013 [11] (2013 data in the upper right of each group). © 2014 IEEE

An order-of-magnitude improvement in energy efficiency beyond the levels of GPU's is evident for the class of dedicated processors. They serve in medical, multimedia and object-recognition applications. The top level of **1 GOPS/mW or 1 TOPS/W** was first reported in 2011, achieved with a neural-network architecture similar to those described in [3], and nicely illustrated (Fig. 10.3) again in a 1.2 TOPS/W publication of 2014 [3]. This was extended to 1.93 TOPS/W in 2015 [4].

Impressive as 1 TOPS/W may be, an always-on mobile companion with heavy video traffic, requiring 10 GOPS processing capacity, would run 1 POP/day (Peta operations/day), consuming 1 kW, not considering communication and display power. We see that even dedicated processors need orders-of-magnitude improvements in energy efficiency. The outlooks in Chaps. 2, 12, 16, 18, and 20 address sustained developments at the performance-energy frontiers, with a focus on video and multimedia.

Fig. 10.3 Three-layer perceptron with input (*I*)-, hidden (*H*)-, and output (*O*)-neuron in [3]. © 2014 IEEE



10.3 Supercomputers

The ultimate performance check on high-performance, general-purpose computing are supercomputers, which, at the 2015 level, should perform Peta-FLOP/s (10^{15} Floating-Point Operations per s). The top 500 are identified continuously [5], and their evolution thru 2020 is projected as shown in Fig. 10.4.

In 2020, the no. 1 is expected to perform 2 Exa-FLOP/s (10^{18}), more than the top-500 together in 2015, and 1000-times more than the no. 1 in 2010. Compared with the projections of 2010 [1], the computing community sticks to the 1000-times/decade target. Obviously, under these expectations, the energy efficiency of these giants has moved into the focus, as shown in Fig. 10.5.

Evidently, the energy efficiency rose 7-times in 5 years. The projection to 2020 makes the optimistic assumption that this improvement rate can be maintained. so that the 20-times improvement of the performance means a ~3-times increase in electric "wall-plug" power over 5 years. The top-10 systems 2014 are listed in Table 10.1.

This data is evaluated with respect to their energy efficiency in GFLOP/s/W or its inverse, the energy/FLOP in nJ = nWs, and it is shown in a plot of PFLOP/s versus the energy efficiency in Fig. 10.6.

The figure also shows the lines of constant throughput figure-of-merit (FOM) in PFLOP/s over the energy/FLOP in nJ. The message of the throughput-FOM is that

"Green" energy efficiency alone is not enough: If the loss in performance = speed = FLOP/s is greater than the gain in energy-efficiency, a certain task of operations, completed in a certain time, would need more computers running in parallel and more total energy than the faster computer with a smaller energy-efficiency.

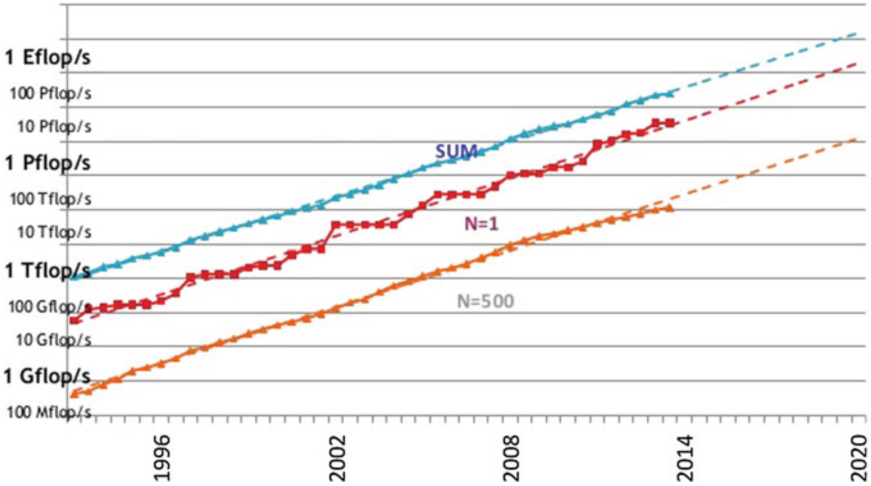


Fig. 10.4 Projected performance development of the top 500 supercomputers, status 2014 [5]: Average growth 80 %/year. © top500

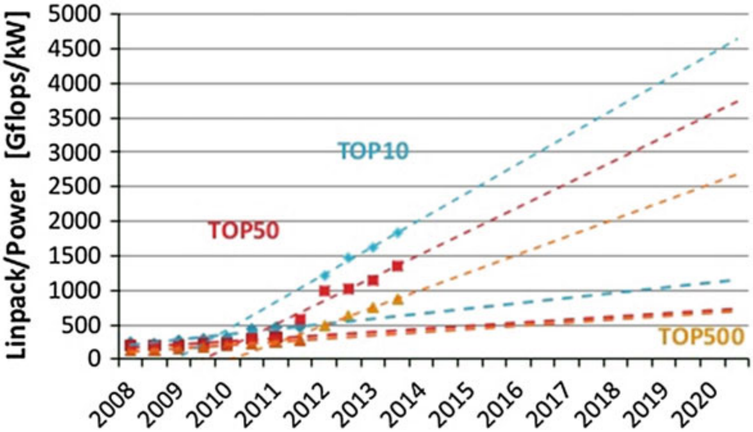


Fig. 10.5 Projected energy efficiency (Gflop/s/kW) of the top 500 supercomputers, status 2014 [5]: average improvement 7× in 5 years. © Top500

Table 10.1 Top-10 supercomputers 2014 [5]

No.	Site	Manuf.	Computer	Country	Cores (thousands)	R_{max} (PFLOP/s)	Power (MW)
1	Univ. Defense Tech.	NUDT	Tianhe-2	China	3120	33.9	17.8
2	Oak Ridge NL	Cray	Titan	USA	561	17.6	8.21
3	Lawrence Liv. NL	IBM	Sequoia	USA	1573	17.2	7.9
4	RIKEN Adv. Inst.	Fujitsu	K Comp.	Japan	795	10.5	12.7
5	Argonne NL	IBM	Mira	USA	786	8.6	3.9
6	CSCS	Cray	Piz Daint	Suisse	116	6.3	2.3
7	Texas ACC	Dell	Stampede	USA	462	5.2	4.5
8	Forsch.zentr. Juelich	IBM	JuQUEEN	Germany	459	5.0	2.3
9	Lawrence Liv. NL	IBM	Vulcan	USA	393	4.3	2.0
10	Government	Cray		USA	226	3.1	?

The strategy in high-performance computing (HPC) has to be to the right ***and*** up in Fig. 10.6. See DARPA [6].

The track record of consumed wall-plug power of the Top 500 is shown in Fig. 10.7. It amounts to a total of 500 MW in 2014, of which 30 % are consumed by the top 10 %. Their optimistic projection for 2020, based on the 7-times

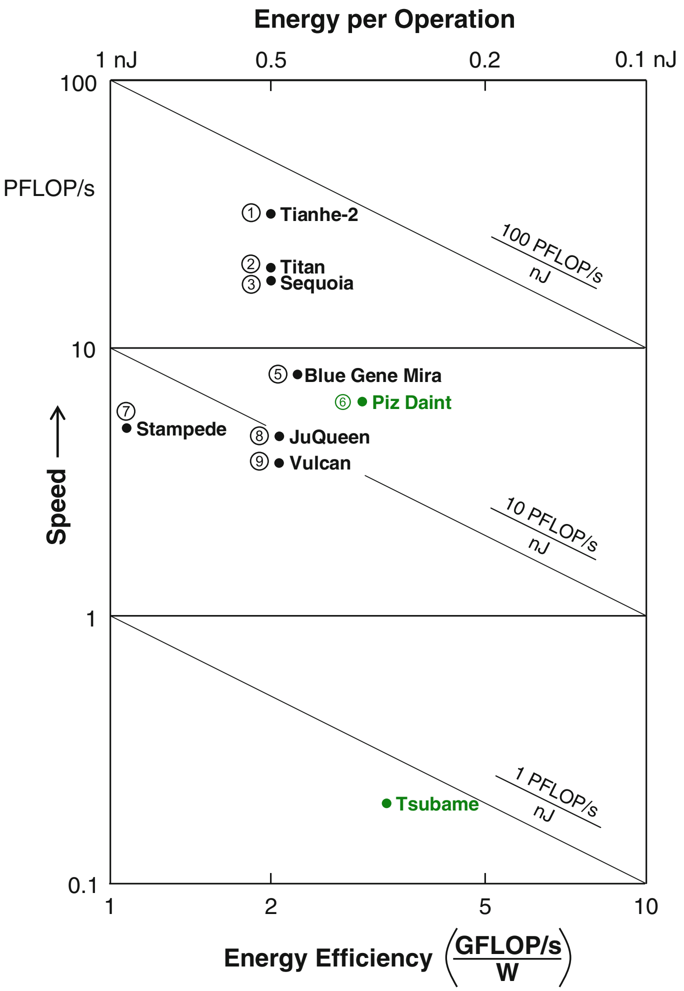


Fig. 10.6 Performance of the top supercomputers 2014 in PFLOP/s (Peta Floating-Point Operations/s) versus their energy efficiency in Giga FLOP/s/W. The top "Green" supercomputer Tsubame KFC is included in this graph with the best energy efficiency of 3.4 GFLOP/s/W [5, 12]. The DARPA UHPC goal 2018 [6] is a computer performing 1PFLOP/s with an energy efficiency of 50GFLOP/s/W, 15-times better than the most efficient system 2014

improvement of the energy efficiency, is that the top 500 will achieve, within 5 years, a 20-times improvement in performance with 3-times more electric power. This would still mean that the no. 1 computer of 2020 would consume 55 MW (the power of a developed city with 40,000 people) and with an energy efficiency 8-times worse than the DARPA UHPC efficiency target [6], extended to 2020. Even with this optimistic projection, parallelism would increase to ~10 Mio. cores, as projected in Fig. 10.8.

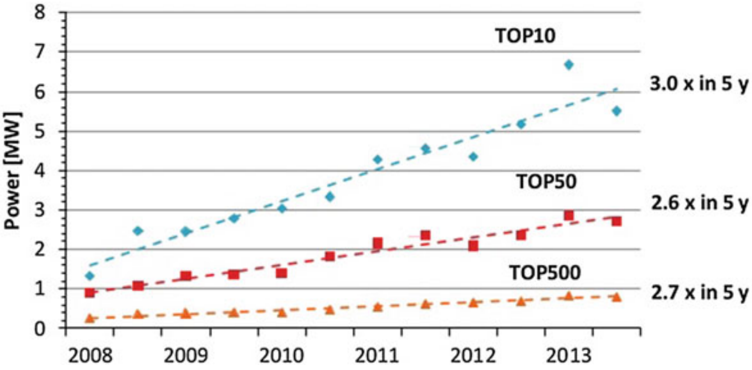


Fig. 10.7 Power consumption of one of the Top 500 supercomputers [5]. © TOP500

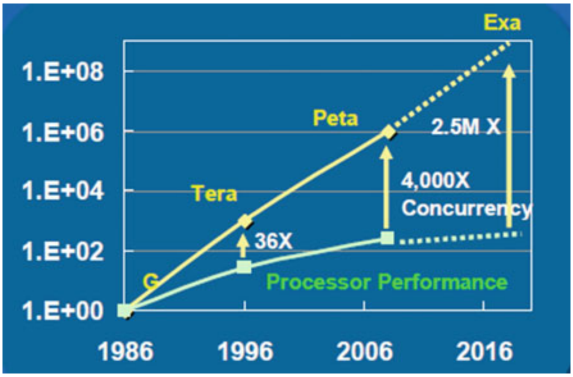


Fig. 10.8 The limited progress in individual processor performance requires the concurrent operation of millions of these to achieve the Exa-FLOP/s supercomputer performance [10]. *Source* INTEL

The Top 500 are a class, exciting to watch and certainly strategically important. However, another group of HPC's is 100-times larger in units- and in power consumption, namely the Internet servers.

10.4 Internet Servers

We estimated the electric-power consumption of the world's servers in 2010 at 36 GW [7]. An in-depth and frequently quoted study [8] arrived at 37–58 GW for internet-related servers, where the authors assumed 50 Mio. cloud servers, using the Internet 90–100 % and 100 Mio. business servers, using the Internet 50–95 %. From 2010 to 2015, the total server performance doubled every year, and this trend is expected to work until 2020, like for supercomputers. With the optimistic assumption that the energy-efficiency continues to improve 7-times in 5 years, the total wall-socket power for servers in 2020 would still be 10-times higher than 2010, requiring 370–580 GW.

The world's total electric power consumption was 2.4 TW in 2012 [9] with an increase of 10 % in 5 years. In this framework, this expansion of server power is not conceivable. Three important remedies for this inflation of computing power are:

- Reduce the data volume by working with effective, intelligent data, particularly for video, instead of inflationary Big Data.
- Improve the energy-efficiency/operation significantly beyond the present pace.
- New architectures for computing have to solve the throughput, the communication and the reliability problems of exponentially expanding, conventionally parallel systems.

We tackle these issues with holistic views in Chaps. 2, 9, 12, 18 and 20.

10.5 Conclusion

Computing is the flagship-domain of nano-chips. The exponential growth of computer performance with an almost 2-times improvement per year could be maintained until 2013, in spite of the nearing end of the nanometer-roadmap, with the massively-parallel operation of multi-core processors at the price of significant increases in wall-socket power. Servers alone consume between 140 and 280 GW in 2015, between 5 and 10 % of the provided global electric power. Even the leading 2014 "Green" supercomputer with its energy-efficiency of 3.4 GFLOP/s/W falls a factor of 4 short of the DARPA roadmap for ubiquitous high-performance computing in spite of its extensive use of special-purpose graphics accelerators. Realizing "Green" efficiency with reduced operating voltage at the price of significant losses in speed forces dramatically more processors/chips operating in parallel to achieve required throughputs, a direction non-"Green" in wasting resources and unreliable and non-resilient [10].

Ultra-low voltage requires new CMOS circuit techniques like differential transmission-gate logic (Sect. 1.6), beating standard CMOS by factors of 10 in energy and >3 in speed. Disruptive moves away from inflationary, dumb numbers to intelligent reality data, particularly in video, are needed as well as their relevance-aware processing (Chaps. 12 and 14), storage and communication, clearly pointing to new architectures (Chap. 18).

References

1. Hoefflinger, B.: Towards terabit memory, Chap. 11. In: Hoefflinger, B. (ed.) CHIPS 2020—A Guide to the Future of Nanoelectronics. Springer, Berlin (2012). doi:10.1007/978-3-64223096-7_11
2. Fluhr, E.J. et al.: POWER8^{TM}: A 12-core server-class processorin 22 nm SOI with 7.6 Tb/s off-chip bandwidth. In: IEEE 2014 International Solid-State Circuits Conference Digest of Technical Papers, pp. 96–97, Feb 2014
3. Kim, G. et al.: A 1.2 TOPS and 1.52 mW/MHz augmented-reality multi-core processor with neural-network NOC for HMD applications, ibid. pp. 182–184
4. Park, S. et al.: A 1.93 TOPS/W scalable deep learning/inference processor with tetra-parallel MIMD architecture for big-data applications, 2015 ISSCC Digest of Technical Papers, paper 4.6, 2015
5. http://TOP500.org/
6. DARPA, Ubiquitous High-Performance Computing (UHPC), DARPA_BAA-10-37_final_ 3-2-10_post_(2).pdf. (2010)
7. Hoefflinger, B.: The energy crisis, Chap. 20. In: Hoefflinger, B. (ed) CHIPS 2020—A Guide to the Future of Nanoelectronics, Chap. 11. Springer, Berlin (2012). doi:10.1007/978-3-64223096-7_20
8. Raghavan, B., Ma, J.: The Energy and Emergy of the Internet, ICSI and UC Berkeley 2011. Copyright 2011 ACM 978-1-4503-1059-8/11/11. http://goo.gl/y4juZ
9. www:eia.gov/installed-electric-power
10. Borkar, S.: Exascale Computing—Fact or Fiction? SSCS Webinar, September 2014
11. Horowitz, M.: Computing's Energy Problem (and what we can do about it), IEEE ISSCC 2014, Dig. Tech. Papers pp. 10–14
12. http://green500.com

Chapter 11
Memory

Bernd Hoefflinger

Abstract News and trends in data storage have advanced to dominate technology and market headlines. They confirm many of the assessments in CHIPS 2020: The low-energy, high-speed static random-access memory (SRAM) with a 6–8-transistor cell is the digital chip function with the lowest supply voltage (<500 mV), the most advanced node on the nanometer roadmap (14 nm), and the potentially most robust nano-circuit due to its fully differential operation. The workhorse DRAM, dynamic RAM, advances only slowly, and, as predicted, is stuck at supply voltages >1 V and technology nodes >20 nm. The overall technology driver is the non-volatile NAND-Flash memory. While its minimum transistor length progresses to 16 nm, the vertical-channel, monolithic 3D transistor integration on-chip and the 3D TSV stacking of >30 chips have produced transistor densities of >10^{11}/cm^2 and memory densities >300 Gb/cm^2, reaching the prediction in Fig. 11.1 of CHIPS 2020. The corollary of this achievement is a powerful technology arsenal for 3D integration, not only of memories, and for the extension of Moore's law in terms of transistors/cm^2. The non-volatile-memory alternatives to NAND Flash continue their competition in the battlefield of programming/writing with thermally induced resistance- or phase-changes. In terms of density as well as write and read speed, the resistive RAM (ReRAM), in the meantime, has made the biggest progress, and it has passed the PCM (phase-change memory). The process compatibility of the ReRAM as a back-end process to CMOS mainstream, as well as its speed and better data retention, make it the technology alternative for non-volatile RAM versus the NAND Flash.

B. Hoefflinger (✉)
5 Leonberger Strasse, 71063 Sindelfingen, Germany
e-mail: bhoefflinger@t-online.de

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_11

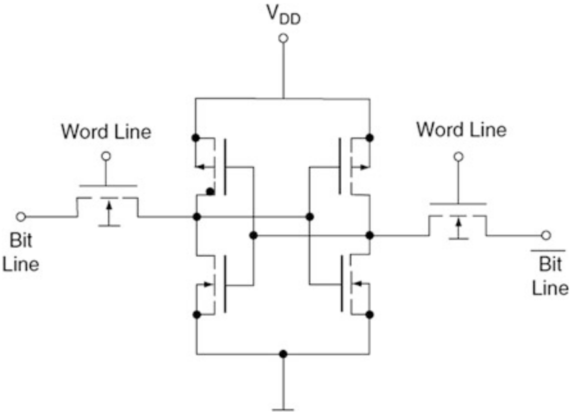


Fig. 11.1 The 6-transistor cell of a CMOS SRAM

11.1 Static Random-Access Memory (SRAM)

The bit-cell (Fig. 11.1), of the static CMOS random-access memory with its 6 transistors, the cross-coupled pair of complementary transistors and two access transistors for differential write and read of the cell content, has become the basic reference circuit for the most advanced technology node, the highest transistor density, the lowest possible supply voltage and the minimum energy per bit.

The 2014 state-of-the-art of planar 2D SRAM cell sizes is shown in Fig. 11.2. The most advanced is a nominal 14 nm SRAM with FinFET transistors [1]. To judge the integration efficiency, we express the cell size of 0.06 µm^2 in terms of fundamental-feature squares F^2, where F = 14 nm. The result is 300 F^2.

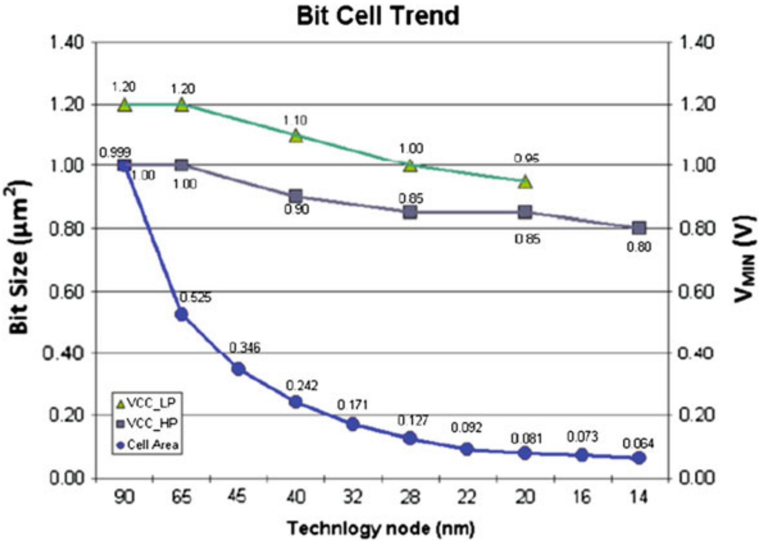


Fig. 11.2 Cell-sizes per bit of CMOS SRAM's [9], ©2014 IEEE

Table 11.1 Comparison of the area efficiency of SRAM's

Year	2010 [2]	2014 [1]	2014 [4]	2014 [3]	2020 [2]
Node F (nm)	90	14	16	20	10
Cell area (μm^2)	1.4	0.06	0.07	0.11	0.012
F^2	180	300	270	270	120
Effective node (nm)	90	18	19.5	24	10

If we compare this with the exemplary cell in [2], we saw 180 F^2 in a 90 nm process, and we can generate Table 11.1 for a comparison.

This table tells us that the overhead for a FinFET transistor topography and for the internal interconnects in the cell lets the gain in area, obtainable from reducing minimum features, fall short of expectations. The number of needed feature squares also gives us the effective cell capacitances, which determine the switching energy and the switching speed. The transistor variance at 14 nm, even for a FinFET topography, reduces the noise margin of the cell [3], and it forces an increase of the SRAM supply voltage, a further increase of the switching energy and of the local thermal-energy density. These negative returns on an expensive scaling effort show the immanent end of the 2D nanometer-roadmap. Figure 11.2 also shows a 2014 cell in a 16 nm silicon-on-insulator technology with a high-dielectric-constant (HK) gate insulator and with metal gates (MG) [4]. At 0.07 μm^2, it is only minimally larger and with 270 F^2 more effective than the FinFET cell. If we take 180 F^2 as a reference for an effective 2D SRAM cell and the 6-transistor cell as our product target, then the 14 nm FinFET cell would effectively be at the 18 nm node and the 16 nm SOI cell at a 19.5 nm node, from the efficiency point-of-view.

Gigabits per cm^2, femto-Joule write- and read-energies, and sub-ns write and read access-times persist as highly critical targets. At the end of the 2D road, evident in Fig. 11.2, there is now no alternative to 3D integration at the transistor-level.

We demonstrated in [2] that the CMOS SRAM is the lead product for the monolithic 3D integration of CMOS circuits and that the result in 10 nm SOI can be a 3D cell area of 0.017 μm^2, corresponding to 120 F^2, a 2.5-times improvement against the 2014 state-of-the-art. This quantum jump can be achieved by growing three transistor-layers on top of each other with reduced-temperature, selective silicon epitaxy and lateral overgrowth, reducing process steps, mask levels, interconnect layers and associated alignment overhead.

11.2 The DRAM (Dynamic Random-Access Memory) at Its Final Stage

The DRAM memory cell with just one transistor and a capacitor per bit is the reference element of practical nanoelectronics and still the fundamental building block of memories for data processing.

However, we learned in [2] that the quality requirements on both the transistor and on the capacitor mean larger feature sizes, typically 45 nm in 2010, and that they will limit the transistor to 22 nm and the capacitor to 20 fF. In fact, at ISSCC 2014, the best reported DRAM's [5, 6] have 26 and 29 nm features with cell sizes of 18 F^2. As predicted, the necessary supply voltages are 1 V or higher. Because of its fundamental switching energy CV^2, the energy efficiency is poor, and it does not scale with the feature size. For 26 nm, the cell size is 0.011 μm^2, offering a density of 9 Gb/cm^2. To keep DRAM on the density roadmap, chip stacks are considered. However, the poor energy efficiency and the power density make that strategy questionable.

We see that the 3D SRAM of 2020 with 0.017 μm^2 will almost reach DRAM density, and its switching energy will be lower by three orders of magnitude. The SRAM continues to replace the DRAM, and the other competitors are those non-volatile memories (NV-RAM's) with adequate write- and read-speeds, which are treated in the following section.

The DRAM at the end of its growth in its energy- and silicon-efficiency presents a serious challenge for the continuing exponential growth of level-1 cache memories for processors. Even if we fill the gap with SRAM and NV-RAM, we finally have to tackle the explosion of data and their processing with the disruptive move to natural data as treated in Chaps. 10, 13 and 14.

11.3 Breakthroughs in Non-volatile Memories (NV-RAM's)

The single transistor with the floating gate for the non-volatile storage of several charge levels (MLC = Multi-Level Cell) has enabled an incredible roadmap of progress for NAND-Flash memories. The 16 nm-technology level, predicted in [2], has been achieved in 2014 [7, 8], and chips with a capacity of 128 Gb have been presented with 2b/cell, reaching storage densities of 70 Gb/cm^2.

A disruptive innovation has been the NAND-Flash memory with series-connected, vertical-channel, surround-gate transistors. This topography is shown in Fig. 11.3, and one possible process-flow is presented in Chap. 1.4. With 24 transistors in series, a 24-layer Si stack, and 2b/cell, a 128 Gb Flash memory has been presented in 2014 [9], and 128 Gb/chip with 3b/cell were reported in 2015 with a highly parallel organization allowing an I/O rate of 1 Gb/s [10].

The density of 96 Gb/cm^2 has been achieved with an effective technology node of ~32 nm so that there is potential for scaling this technology to higher densities. The dramatic progress in memory density, as projected in Fig. 11.1 of [2], is documented in Fig. 11.4 of the present chapter.

This figure shows that the capacity per chip has doubled every two years for 2D MLC NAND Flash with the move to 16 nm where it will stop. Stacks of 32 of these chips reach 5 Tb per chip-stack and 4 TB per package.

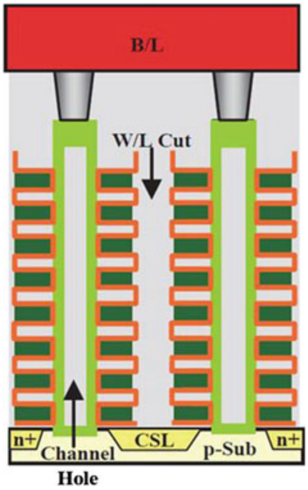


Fig. 11.3 Schematic cross-section of a 3D NAND Flash with vertical transistor series [11]. *Source* MyMemory

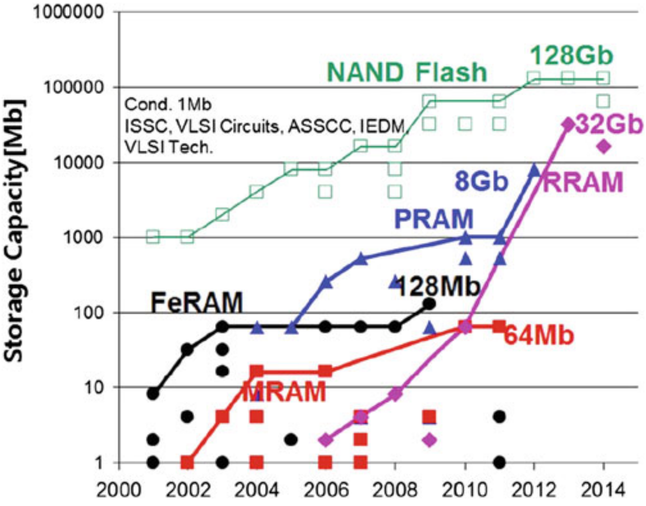


Fig. 11.4 Non-volatile memory capacity in Mb per chip [12]. © IEEE ISSCC Trends 2014

The figure also shows the progress of other non-volatile-memory technologies like

PRAM	Phase-change memories,
FeRAM	Ferroelectric memories,
MRAM	Magnetoelectric memories (spintronics),
RRAM ReRAM	Resistive memories.

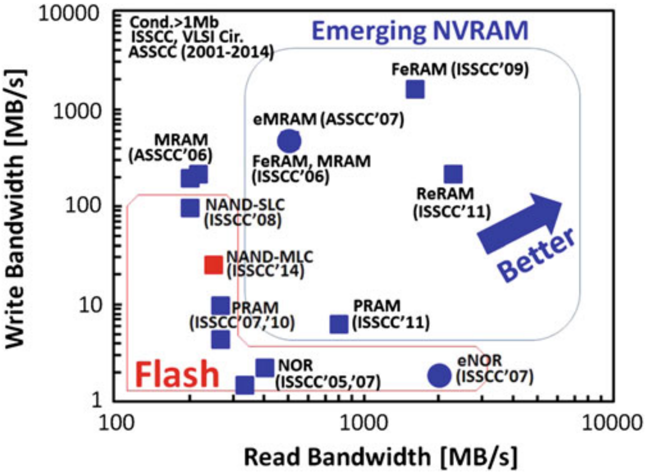


Fig. 11.5 Write and read bandwidth of non-volatile memories [12]. © IEEE ISSCC Trends 2014

We see that the ReRAM, favored in our projections in [2], advanced from 64 Mb in 2010 to 32 Gb per chip in 2014 at a 27 nm technology level. This is a true quantum-step for NV-RAM's, because this type of memory out-performs all others in

- Write- and read-speed,
- CMOS-logic compatibility, since it is added as a BEOL (Back-End-of-Line) feature (CuTe),
- Data retention and re-write capability,
- Low-voltage, low-energy read and possible low-energy write,
- Scalability (1T-1R per cell, requiring only 6 F^2).

The comparison of write- and read-times is shown in Fig. 11.5.

With its write bandwidth of 200 MB/s and its read bandwidth of 1 GB/s, and with its density of 9.5 Gb/cm^2, **the 2014 ReRAM non-volatile memory has reached the performance of the leading DRAM** [13], a sensation and a turning-point for memory strategies.

The ReRAM potential is well documented with a 4T1R non-volatile content-addressable memory with a multi-bit-per-cell capability [14].

11.4 Conclusion

The progress in 2D memory has been slow. The SRAM, proclaimed at 14 nm, in its FinFET area efficiency of 300 F^2, has only the density of a 20 nm node. The 2D DRAM, as predicted, has reached its limit at >20 nm. Their bandwidth and access energy have settled at limits dictated by interconnects ("wires"). This wiring congestion can be removed effectively only with 3D integration, which has finally reached the expected density, compatibility and cost levels (Chap. 3).

Non-volatile NAND Flash memory advanced beyond predictions in density because of both 3D vertical chip stacks as well as 3D vertical transistor stacks, approaching 1 Tb per chip-unit. Their down-sizing has reached limits on retention time, energy and speed. The potential winner in non-volatile memories is the resistive (RE) RAM with its high write- and read-speed, low energy and high retention times.

References

1. Song, T. et al.: A 14 nm FinFET 128 Mb 6T SRAM with V_{min}-enhancement technique. IEEE International Solid-State Circuits Conference Digest of Technical Papers, pp. 232–233 (2014)
2. Hoefflinger, B.: Towards terabit memory, Chap. 11. In: Hoefflinger, B. (ed.) CHIPS 2020—A Guide to the Future of Nanoelectronics. Springer, Berlin (2012). doi:10.1007/978-3-64223096-7_11
3. Yabuuchi, M. et al.: 20 nm high-density single-port and dual-port SRAM's with word-line voltage adjustment system for read/write assists. IEEE International Solid-State Circuits Conference Digest of Technical Papers, pp. 234–235
4. Chen, Y.H. et al.: A 16 nm 128 Mb SRAM in high-k metal-gate FinFET technology with write-assist circuitry for low V_{MIN} applications, ibid. pp. 238–239
5. Oh, T.Y. et al.: A 3.2 Gb/s/pin 8 Gb 1.0 V LPDDR4 SDRAM with integrated ECC Engine for Sub-1 V DRAM core operation, ibid. pp. 430–431
6. Lee, D.U. et al.: A 1.2 V 8 Gb 8-channel 128 Gb/s high-bandwidth memory (HBM) Stacked DRAM with effective microbump I/O test methods using 29 nm process and TSV, ibid. pp. 432–433
7. Helm, M. et al.: A 128 Gb NAND-flash device in 16 nm planar technology, ibid. pp. 326–327
8. Choi, S. et al.: A 93.4 mm^2 64 Gb NAND-flash memory with 16 nm CMOS technology, ibid. pp. 327–328
9. Park, K.T.: Three-dimensional 128 Gb vertical NAND flash memory with 24 WL-stacked layers and 50 MB/s high-speed programming, ibid. pp. 334–335
10. Im J.-W. et al.: A 128 Gb 3b/cell V-NAND Flash Memory with 1 Gb/s I/O Rate 2015 ISSCC Digest of Technical Papers, paper 7.2 (2015)
11. Handy, J.: An alternative kind of vertical 3D NAND string, published Nov. 8, 2013. http://thememoryguy.com/wp-content/uploads/2
12. 2014 ISSCC Trends in Memory
13. Fackenthal, R.: A 16 Gb ReRAM with 200 MB/s/cell write and 1 GB/s/cell read in 27 nm technology, ibid. pp. 338–339
14. Chang, M.-F. et al.: A 3T1R nonvolatile TCAM using MLC ReRAM with sub-1 ns search time. In: 2015 ISSCC Digest of Technical Papers, paper 17.5, 2015

Chapter 12
Intelligent Data Versus Big Data

Bernd Hoefflinger

Abstract The advancement of nanoelectronic functionality and memory capacity has enabled a data explosion, which is promoted as "Big Data", and doubling the mobile Internet traffic every 18 months. In this fashion, the mobile Internet would need over 500 GW of electric power for its operation in 2020 and another 500 GW of embodied power for its manufacturing, maintenance, recycling, and disposal, together one third of the expected total global electric power generation. In the face of this unlikely scenario, a review proposes that this "Big Data" is a result of the artificial, numerical, linear world, inflated by the pocket calculator in the 60s and its successors and now out of control due to linear digital video originated by the CCD imagers since the 70s and mapped onto CMOS imagers. By contrast, we (and our brains) sense our world with a logarithmic response, where minimum response steps record signal increments that are constant ratios of the magnitude of the individual signal. This has been recognized in digital audio, where it led to MP3 as an intelligent data compression. This realism is still missing in digital video, which causes >70 % of the mobile Internet traffic. Transistor-enabled logarithmic recording, invented 1992, and processing of images can compress video data by an order-of-magnitude vs. present mainstream digital video, a significant saving on video traffic and storage. A similar progress from "data" to information can be envisioned towards real data on the "Internet for Everything" like the environment, health, weather, mobility, and robotics.

B. Hoefflinger (✉)
5 Leonberger Strasse, 71063 Sindelfingen, Germany
e-mail: bhoefflinger@t-online.de

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_12

12.1 Progress in Nano-Chips Fosters Data Explosion

The mobile Internet traffic is the best indicator of the data explosion enabled by the performance/cost progress of nano-chips along the expected roadmaps. The CISCO projections of 2014 [1] show a further aggregate growth of 61 %/year through 2018 to 15.9 Exa-Byte (EB) per month (Fig. 12.1), which means ~2 GB/month for each of the ~8 Billion mobile subscriptions to the Internet.

The bandwidth/smartphone will grow 5-times from 2013 to 2018 to 2.7 GB/month, of which more than 2/3 will be video. Overall, video will make up 69 % of the traffic, of which again 69 % will be in the cloud because of mobile-phone storage limitations. Cloud traffic is expected to make up 90 % of all mobile traffic in 2018. The expected 10 Bio. mobile devices in 2018, including tablets, laptops and wearables, will need close to 50 GW of wall-socket power.

For the total Internet energy estimate, we consider the Berkeley study [2] as the most instructive one. It weighed the 1Bio. smartphones of 2010 at 130–450 MW, however, with significant embodied power of 4–14 GW because of their effective use for only two years. The data from [2] are summarized in the 2010 column of Table 12.1. The electric power consumption of 228 GW in 2015 may be compared with a 2013 report of 1500 TWh/year = 175 GW, quoted in [3]. The expectations through 2020 follow Ref. [1] and the progress in computing power as described in Chap. 10.

The result, as illustrated in Fig. 12.2 with the demand of >500 GW, would be 1/6 of the total expected electric power generated globally in 2020, not even considering the embodied power of equal magnitude for manufacturing and life-cycle energy cost, which is not conceivable. A total review of the explosion of digital data is necessary, and we will treat just a few of the needed paradigm shifts.

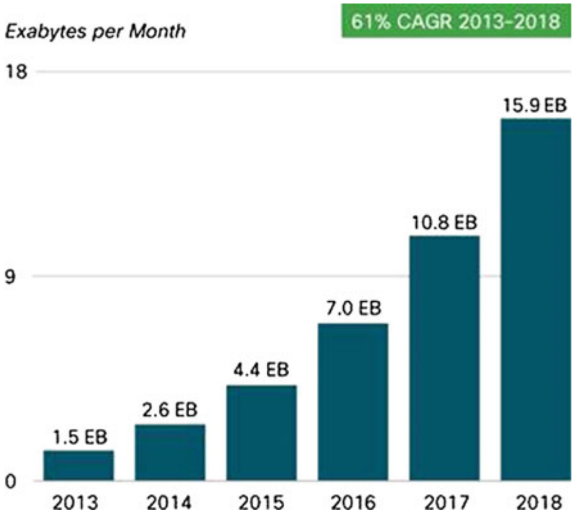


Fig. 12.1 Expansion of the mobile internet traffic 2013–2018 in EB/month [1]. *Source* Cisco VNI Mobile, 2014

Table 12.1 Expected internet performance and its demand for electric power

Year	2010	2015	2020
Mobile traffic (TB/s)	0.11	1.5	13.2
Server performance (rel.)	1	20	200
Energy eff. (rel.)	1	7	25
El. power (GW)	40	120	240
PC's power (GW)	30	30	30
Mobile devices (GW)	0.4	28	56
Infrastructure (GW)	10	50	200
Total Operative El. (GW)	**80.4**	**228**	**526**
Embodied El. power (GW)	80	228	526
Total El. power (GW)	**160.4**	**456**	**1052**
Global El. power (GW) **generation**	**2300**	**2800**	**3300**

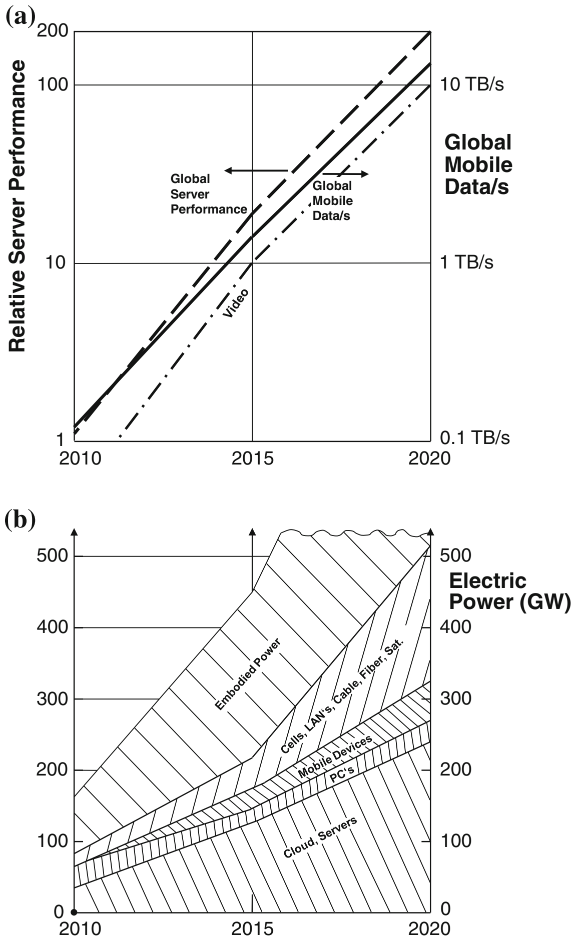


Fig. 12.2 **a** Global mobile internet traffic in Tera-Byte per second (TB/s) and the performance of servers on a relative scale. **b** Mobile internet demand for electric power

12.2 Intelligent Data from and for Our World

The data explosion is a **Digital Data explosion**. Close to being overwhelmed by it, we have to review, how we got into this and how we can manage data. The ITRS 2013 [4] lists, among others:

- A new data representation?
- A new information-processing technology, non-binary, non-Boolean logic?
- A new system architecture?

Aside from our digital artifacts, we have the explosion of natural, physical, chemical, environmental and medical input data acquired by our nano-enabled intelligent systems in the "Internet of Everything".

We have to ask ourselves:

- How do we represent these quantities as data?
- How do we measure or acquire this data?
- How do we process this data?
- How do we store this data?

In Chap. 4 on Analog-to-Digital Converters, these issues are addressed in Sect. 4.5 under the new research direction on Analog-to-Information Converters.

Mostly, we represent the magnitude of a quantity as an analog number N, commonly decimal with a dynamic range from 0 to N_{Max}. The first question is: How efficient is decimal, in other words:

How many elements do we need to describe the range 0 to N_{Max} with a numbering system with a base b? The mathematical answer is that b = e = 2.715… would need the minimum number of elements. This has led to a base-3 or ternary system with levels 0, 1/2 and 1. This is used in some memory systems. However, the next nearest, base-2, the binary system, has the fundamental advantage of near-optimum efficiency and of being the optimum for large logic operations based on simple switches, which are either On or Off, respectively 1 or 0.

How do we cover dynamic range with a binary number system?

We use either integer numbers

$N = a_n 2^n + a_{n-1} 2^{n-1} + \cdots + a_1 2 + a_0$, where $a_i = 0$ or 1, with a dynamic range of 2^{n+1},

or floating-point numbers

$N = (a_m 2^m + a_{m-1} 2^{m-1} + \cdots + a_1 2 + a_0) 2^k$, $k = 0$ to p, $a_i = 0$ or 1, with a dynamic range of 2^{m+p+1}.

The word-lengths of nb for integer or (m + p)b for floating-point immediately provide the basic multiplier for the data explosion, and we should answer the question:

How can we reduce or **compress** word-lengths?

In a classic example, nature or bionics comes to the rescue: We record audio for the purpose of hearing, one of our most important senses. Our sensor, the ear, has a logarithmic response H to sound levels S:

$$H = a\ln(S).$$

For a given sound signal level S, our just noticeable hearing difference dH (JND), dH = a(dS/S), needs larger changes dS, as the signal level S increases. In digital audio coding, this natural response is implemented with a famous, piece-wise linear approximation to the log response, the A-law or μ-law, as shown in Fig. 12.3.

The voice dynamic range of 2.000:1 is represented by 11b plus one signb. The hearing response is approximated by 8 segments (3b) plus signb plus 4b per segment so that a compression from 12b input to 8b output is achieved. Practically, the analog voice signals are converted directly with a pseudo-log A/D converter, built with log-weighted switched capacitors. Similar compression curves are used in high-quality audio. This is not only a reduction in word-length by 30 %, but it also mimics natural hearing with a constant signal-to-noise ratio. This perception-driven coding eventually led to MP3, which revolutionized digital audio.

Another one of our senses, vision, has a similar logarithmic response, however, to a much wider input dynamic range of 7 orders-of-magnitude, with the incredible

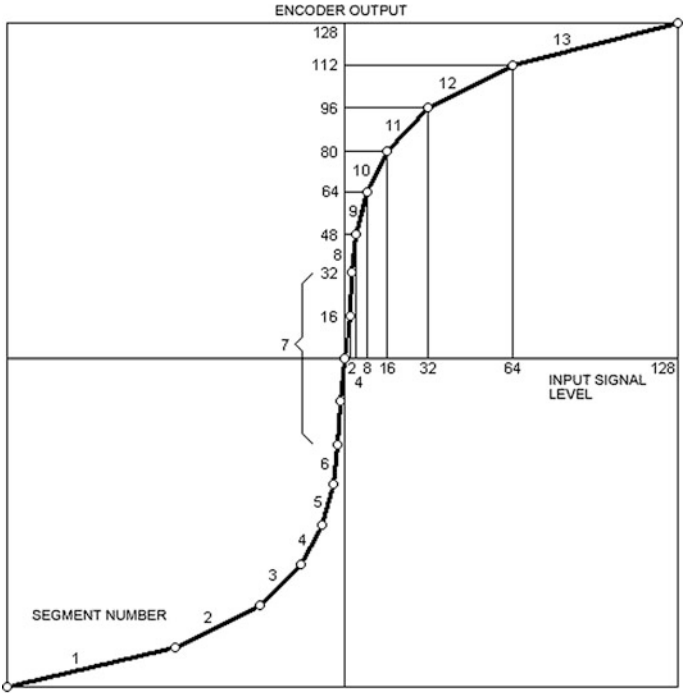


Fig. 12.3 PCM Audio encoding with a piece-wise-linear approximation to the log response characteristic

parallelism of millions of receptors and with a much wider bandwidth. During the epoch of analog chemical film, this logarithmic response was cultivated to an input dynamic range beyond 10.000:1. It is the sad consequence of the microelectronic imaging invention of the charge-coupled device (CCD) with its linear response and a dynamic range of only 1.000:1, that the present digital video drowns the Internet mostly in linear image and video data.

This is even more embarrassing because the implementation of a natural log video response produces better video quality, and because of the miracle, that the HDRC pixel in Fig. 12.4, in the sub-threshold regime of its MOS transistor, offers a log response over a dynamic range of >7 orders-of-magnitude or 28 bits of input dynamic range (see [5, 6] and the following Chap. 13). The HDRC imager characteristic with 10b digital output is represented in Fig. 12.5 together with that of the human eye by their contrast sensitivity, which is the derivative of the sensors' response. Figure 12.6 shows three frames from an HDRC video of a solar eclipse.

The output voltage of the HDRC^{TM} pixel delivers the perfect bionic response to the input luminance over seven orders of magnitude, Fig. 12.5, without any piece-wise linear converter. To mimic the 1 % contrast resolution of the human eye, an output word-length of 10b/pixel is adequate for a dynamic HDR photography and video and because of its high data load on the Internet, the recording, coding and compression of video receive a special treatment in Chaps. 13 and 14. We can state that **the a priori log recording** range of 28 f-stops, while a linear sensor would need seven exposures at different apertures and exposure times leading to 54b/pixel, needing >5-times as many bits per pixel plus latency and post-processing. Further coding and compression advantages, based on logarithmic HDR data, were published in 2007 [6], including two logarithmic codes, LogLuv with 15b for the log of

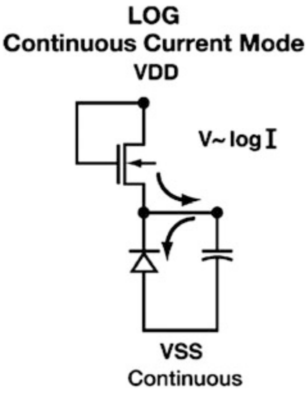


Fig. 12.4 HDRC^{TM} (high-dynamic-range CMOS) pixel with log-response characteristic over seven orders-of-magnitude or 24-f-stops of luminance [4, 5]. The same transistor-photodiode configuration is at the center of each of the 1600 receptors in the retinal implants (Chap. 17) in order to log-compress the charge delivered to the nerve cells of blind patients so that they perceive a natural response and are not irritated by bright light sources

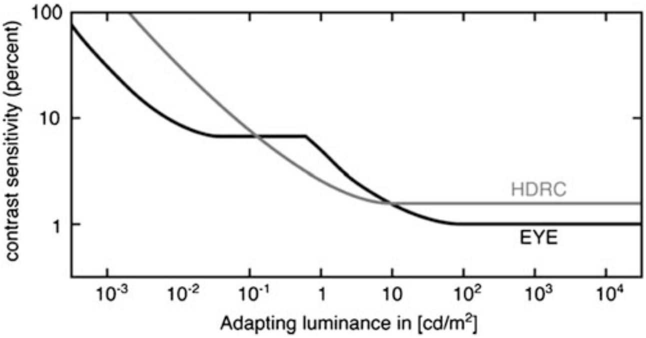


Fig. 12.5 Contrast sensitivity curves: The HDRC sensor shows natural response over seven orders of magnitude

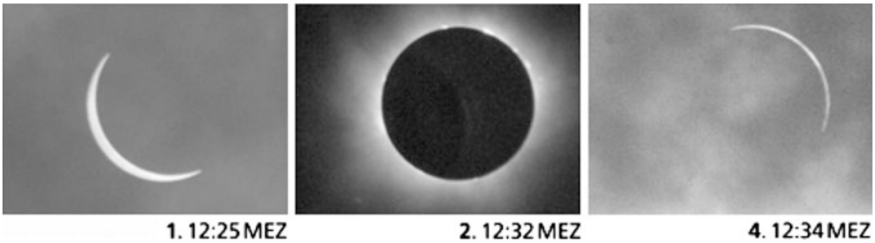


Fig. 12.6 Solar eclipse August 11, 1999, in Germany. Three frames from the video taken with an HDRC camera with 10b output at 30fps., with a 210 mm telephoto lens at f:5.6 [4]. The cloudy sky during the event was well recorded as were the coronas, free from white-saturation, with a constant aperture

luminance and JNDLuv with 12b for log luminance (Fig. 12.7). Because of the great interest in digital HDR photography and video and because of its high data load on the Internet, the recording, coding and compression of video receives a special treatment in Chaps. 13 and 14. We can state that

- **The a priori log recording of images improves the original picture quality together with a natural data compression by more than a factor of 3,**

and that it allows further intelligent coding and compression offering additional factors of improvement.

Eye-like, logarithmic video has to be considered as the biggest alleviation of the video-load on the Internet and, at the same time, as a fundamental improvement of the quality of visual information.

Logarithmic sensing, measuring and acquisition are natural and effective for almost everything in our real world, as well as in virtual and augmented reality. Unfortunately, only with money on its scale of 1, 2, 5, 10, 20, 50 etc. are we used to it.

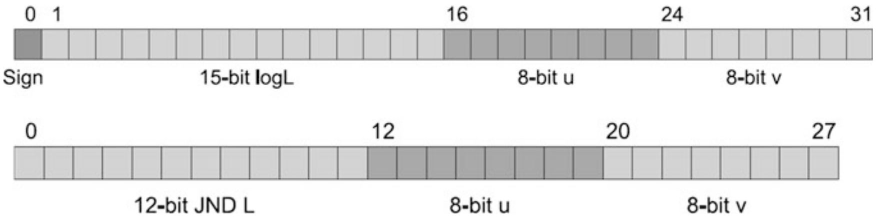


Fig. 12.7 Two logarithmic codes for digital video. *Upper* LogLuv with 15b for the log of luminance and 2 × 8b for color gamut u, v. *Lower* JNDLuv (Just-noticeable difference) Luv with 12b for log luminance, the result of visual-perception research [6]. © Springer 2006

Another example: In the measurement of distance, like for collision-avoidance, we accept that an accuracy of 1 m is ok, if the car in front is 100 m away, that 10 cm would be ok at a distance of 10 m, and that the accuracy should be 1 cm for a distance of 1 m, resulting in a relative accuracy of 1 %. This is the essence of

Weber's Law: We acquire information on a natural quantity N with a constant relative resolution:

$$\mathbf{dN/N = const}.$$

In our example, we would record distances x in a piece-wise linear, pseudo-log fashion:

1.00, 1.01, …, 10.0, 10.1, …, 100, 101 m with a total of 2.000 numbers, instead of the linear 1 cm-resolution requiring 10.000 numbers. If, instead, we record directly logx from 0 to 2 with a resolution of 0.01, we would need just 200 numbers for a distance accuracy of 1 %, a very powerful compression of the "linear-number" world, based on the least-significant value of 1 cm, which we may not even be able to detect at 100 m distance. Examples of the natural distance detection are the 3D TOF (time-of-flight) sensors in the following Chap. 13.

The "number" world is an artifact of numerical calculators, first mechanical, then relay- and tubes-based, finally transistorized since the 1960s with the advent of the pocket calculator. Generally, standard paper-and-pencil multiplication fosters a "number"-world, as we perform a multiplication with the least-significant number first, irrespective of its relevance for the quantities represented by the numbers.

However, many regions in our world developed, over thousands of years, applied-math cultures to start operations with the leading numbers in a quantity, particularly, when the operation is a multiplication of two quantities with a certain relative accuracy. Various types of Abacus are living examples. The most efficient tool, particularly for multiplication, division, powers of 2 and 3 etc. became the slide-rule as shown in Fig. 12.8, (once) widely used in Europe. When the Russians had launched Sputnik into Space in 1957, before the Americans, a shock went through the US that that had happened because the Americans had not taught the slide-rule in school. An introduction in the US was attempted, but effectively undermined by the appearance of the pocket calculator.

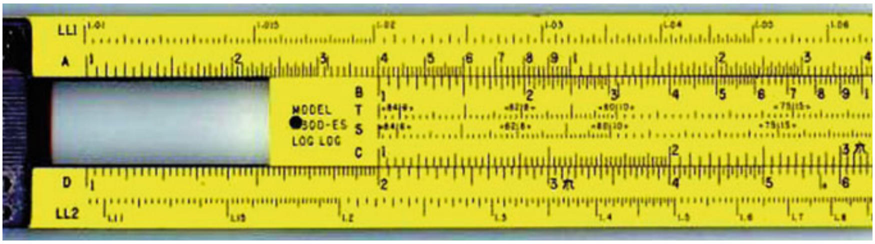


Fig. 12.8 A slide rule. The scales C and D have the same log-scale. The sliding bar is set such that 1 on its scale C is above 2 on scale D, so that the operation 2 × 2 = 4 appears as: log2 on scale D + log2 on scale C appears as log4 on scale D. *Source* Wikipedia

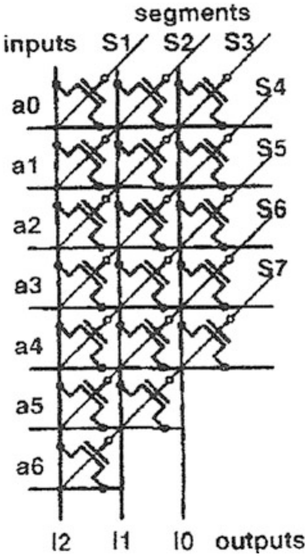
The slide-rule is the genial embodiment of Weber's Law or, in other words, of our real-world concerning accuracy or precision: If the accuracy of the "2" on line C is 10 %, it could be 2.2, and our result would be 4.4, also 10 % higher. If the "2" on line D could also be 10 % higher, we see that the result would be 4.8, 2 × 10 % = 20 % higher:

The relative accuracy of the result of a multiplication is the sum of the relative accuracies of multiplicand and multiplier.

Log conversion is one effective way of representing real-world data, and it reduces multiplication to an addition. Effective transistor arrays for this conversion, often a technically justified log compression like the distance-example above, on a binary base, can be found in [7, 8] and in Fig. 12.9.

Multiplier-intensive operations on real-or virtual-world data in matrix operations or digital neural networks with large numbers of synapses can benefit effectively

Fig. 12.9 Encoding of a 6-bit input into a 3-bit logarithmic output [7]



from log processing, if we imagine that log data are provided and multiplier weights (synaptic weights) are provided on a log scale.

Video processing again is a top candidate for log efficiency gains.

12.3 Digital Multipliers for Reality Data

If log processing is too disruptive or not indicated because of its overheads, digital binary multipliers for reality data can be made much more effective against the present state-of-the-art. Presently, multipliers are area-, transistor- and energy-consuming heavy-weights with critical delays. A number-crunching standard n × n bit multiplier starts with the least-significant bits a_0 and b_0. Eventually, it produces a result with (2n + 1) digits and with a delay of 2n adders. Even with Booth-Wallace coding, its complexity (transistor count) increases with the square of its word-length, $O(n^2/2)$.

If n represented reality data, the accuracy of the n-bit words might be p bits, and the accuracy of the result would be (p − 1)b. In other words:

Only the leading (p − 1)b are relevant. For the rest, just the number of digits is relevant for the order-of-magnitude of the result.

With this sober insight, and in the spirit of Weber's Law, the Abacus and the slide rule, we start multiplication at the significant end, namely with the leading one's. We determine them as $a_j = 1$ and $b_k = 1$, and we proceed as follows in the case of 6b accuracy:

$$A = 2^j + a_{j-1}2^{j-1} + \cdots a_{j-5}2^{j-5},$$
$$B = 2^k + b_{k-1}2^{k-1} + \cdots b_{k-5}2^{k-5},$$
$$P = A \times B = 2^j B + a_{j-1}\left(2^{j+k-1} + b_{k-1}2^{j+k-2} + \cdots + b_{k-4}2^{j+k-5}\right) + \cdots + a_{j-5}2^{j+k-5}.$$

Table 12.2 shows the inputs to the adders, which contribute to the 6-bit result. The complexity is $O(6^2/2)$ with 10 full- and 6 half-adders, independent of the

Table 12.2 Leading-one's-first integer multiplier with 6b accuracy

Integer index	j + k	−1	−2	−3	−4	−5
$a_j = 1, b_k = 1$	**1**	**b_{k-1}**	**b_{k-2}**	**b_{k-3}**	**b_{k-4}**	**b_{k-5}**
If $a_{j-1} = 1$[a]		**(1)**	**(b_{k-1})**	**(b_{k-2})**	**(b_{k-3})**	**(b_{k-4})**
If $a_{j-2} = 1$[a]			**(1)**	**(b_{k-1})**	**(b_{k-2})**	**(b_{k-3})**
If $a_{j-3} = 1$[a]				**(1)**	**(b_{k-1})**	**(b_{k-2})**
If $a_{j-4} = 1$[a]					**(1)**	**(b_{k-1})**
If $a_{j-5} = 1$[a]						**(1)**

The complexity is of the order $O(6^2/2)$ independent of the word length n
[a] Otherwise the inputs from this line are 0

Table 12.3 Transistor counts for standard multipliers and for the accuracy-oriented Leading-Ones-First (LOF) "reality" multipliers

Word length n	8b	16b	24b
Standard Booth-Wallace	1600	6400	14,400
LOF 6b precision (1.6 %) 10b precision (0.1 %)	540	1030 2100	1620 2700

word-length n. A standard 8 × 8 multiplier with Booth-Wallace coding would have a complexity of 32, for 16 × 16 it would be a complexity of 128, 8-times larger than its leading-one's-first version with 6b accuracy. For 10b accuracy, the new multiplier would have a complexity of 50. The related transistor counts are given in Table 12.3.

We see in Table 12.3 that, for word-lengths of 16b or 24b, frequent in signal-processing, the LOF multiplier-type needs 1/3 and 1/5 of the number of transistors, and its delay is 1/5 and 1/8 of standard multipliers, respectively. The LOF 16 × 16 multiplier with 6b accuracy, in ultra-low-voltage, differential transmission-gate (DTG) logic makes up the HIPERLOGIC curve in Fig. 1.11 of Chap. 1, projected towards an energy–efficiency of 1fJ/operation, 200-times better than the best reported result of 2014 and 800-times better than the best results in standard CMOS logic in 2014 (Fig. 1.11).

Besides the multiplication itself, the processing of this reality data with certain accuracies also means shorter, adequate word-lengths with smaller registers and memories, with a leading-ones-based organization, saving memory and, more important, communication bandwidth, the critical show-stopper as detailed in Chap. 5 and in [9].

12.4 Conclusion

The explosion of digital data is not compatible with our resources. It is an artifact of the numerical age generated by the pocket calculator. Now, that visual information fills more than 2/3 of the Internet, and with our expectations on the "Internet of Everything" and on bio-inspired computing with Trillions of sensors, it is time to go back and to start with reality data, with their recording and natural compression, and to process, to communicate and to store them according to their relevance. Section 4.5 also points in this direction with "compressed sensing" and "finite-rate-of-information sampling".

Vision is the no.1 priority, and its digital revolution, like MP3 in audio, is still up for grabs. High-dynamic-range-CMOS (HDRC) imaging offers up to 5-times reduction of luminance-bits per pixel, and it receives further treatment in Chaps. 13 and 14.

References

1. CISCO Visual Networking Index: Global mobile traffic forecast update 2014. www.cisco.com
2. Raghavan, B., Ma, J.: The energy and emergy of the internet, ICSI and UC Berkeley 2011. Copyright 2011 ACM 978-1-4503-1059-8/11/11. http://goo.gl/y4juZ
3. Clark, J.: IT now 10 percent of world's electricity consumption, report finds, The Register, August 16, 2013. http://www.theregister.co.uk/2013/08/16/it_electricity_use_worse-than_you_thought/
4. Hoefflinger, B. (ed.): High-Dynamic-Range (HDR) Vision. Springer, Berlin/Heidelberg (2007). ISBN 13 978-3-540-44432-9
5. Hoefflinger, B.: Vision sensors and cameras, chapter 15. In: Hoefflinger, B. (ed.) CHIPS 2020—A Guide to the Future of Nanoelectronics. Springer, Berlin, Heidelberg (2012). doi:10.1007/978-3-642-23096-7_15
6. Mantiuk, R.: HDR image and video compression, chapter 12. In: Hoefflinger, B. (ed.) High-Dynamic-Range (HDR) Vision. Springer, Berlin/Heidelberg (2007). ISBN 13 978-3-540-44432-9
7. Hoefflinger, B., Selzer, M., Warkowski, F.: Digital logarithmic CMOS multiplier for very-high-speed signal processing. In: IEEE 1991 Custom-Integrated-Circuits Conference (CICC), Digest, pp. 16.7.1–16.7.5 (1991)
8. Hoefflinger, B.: Circuit arrangement for digital multiplication of integers, US Patent 5,956,264, filed 18 Jan 1996, issued 21 Sept 1999
9. Borkar, S.: Exascale computing—fact or fiction? SSCS Webinar 2014. https://courses.edx.org/courses/IEEEx

Chapter 13
HDR- and 3D-Vision Sensors

Bernd Hoefflinger

Abstract Vision chips continue to aggressively follow the complexity of nano-chips. The smartphone vision-chips of 2014 exceeded 16 Mega-pixels, irrespective of the optical-resolution limits of their lenses with focal lengths of 5 mm or less. At 24- to 48-bit per pixel, this volume of video data is about to flood the Internet, where video already exceeds 70 % of the traffic and >90 % of the data stored. On the other hand, high-sensitivity, high-dynamic-range CMOS (HDRC) human-vision-inspired sensors have maintained pixel sizes of 25 μm^2 for ASA 12.800 sensitivity and 1000-by-2000 pixels complexity as well as high-speed global-shutter read-out. HDRC two- and three-chip stereo cameras have become key components for safe and intelligent man-machine cooperation. New compression standards will be needed. More efficient in the long run will be a fundamental review of electronic vision, following human vision with its incredible compression skills. Remarkable benefits from nm technology nodes were achieved for high-resolution time-of-flight (TOF) 3D-vision sensors offering >400-by-600 pixels resolution. TOF-3D and professional stereo cameras enable safe, autonomous mobility and robotics.

13.1 Scaled CMOS Image Sensors

Active-pixel CMOS image sensors have finally taken over from traditional CCD sensors because their development benefitted from the CMOS roadmap together with all its developments of thin-chip and deep-trench processes. As a result, backside-illuminated pixels with high fill-factors and reduced dark-currents have become standard. The 2010 state-of-the-art was described in [1] with a 10 Mpixels sensor in a 140 nm technology with a 1.65 × 1.65 μm^2 backside-illuminated pixel.

B. Hoefflinger (✉)
5 Leonberger Strasse, 71063 Sindelfingen, Germany
e-mail: bhoefflinger@t-online.de

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_13

Table 13.1 Comparison of CMOS Backside-Illuminated sensors 2010–2014

Year	2010 [2]	2014 [3]
No. of pixels	10 M	8 M
Technology (nm)	140 BSI	BSI DTI
Pixel area (μm)^2	1.6 × 1.6	1.12 × 1.12
Saturation (e^-)	9130	6200
Dark current (e^-/s)	3	6
Dyn. range (dB)	71	70

BSI Back-side illuminated
DTI Deep trench isolation, vertical transistor

As the comparison in Table 13.1 shows, the CMOS sensors have reached a mature technology status, which is also due to basic optical correlations of lens apertures, their focal lengths and the spatial resolution achievable (see Table 15.2 in [1]). Therefore, there is no nm-roadmap for sensors other than benefitting from technology advances like high-resolution deep-trench isolation (DTI) and vertical transistors adapted from NAND Flash memories (Chap. 11). These latest enhancements are indicative of the march of CMOS image sensors into 3D monolithic integration (Chap. 3) for image processing and storage, which, in 2014, is still very much a lateral 2D affair, as exemplified in the following section, although such strategies had been demonstrated in 1997 (see Fig. 3.33 in [1]).

13.2 Hi-Speed Feature-Recognition Chips

System integration is well documented with CMOS sensor chips which include pattern-recognition capabilities like recognizing faces. Three chips are represented in Table 13.2 with their technical data. Pixel fields of up to 256 × 256 are recorded

Table 13.2 Comparison of face-recognition chips

Reference	[4] 2014	[12] 2008	[13] 2011
Technology (nm)	180	180	180
Sensor resolution	256 × 256	128 × 128	128 × 128
Clock (MHz)	50	50	100
Power (mW)	630	455	450
Sensitivity (V/luxs)	8.1	NA	NA
Dynamic range (dB)	48	NA	NA
MPU	32b dual-core	32b single-core	32b single-core
Neural network	16 × 16 self-organizing map	No	No
GOPS	12	77	44
Recognitions/s	45,000	1240	160
Frames/s	1340	360	78
Chip area (mm^2)	82	70	13.5

at high frame rates, and the example of 2014 achieves high recognition rates due to the incorporation of a 16 × 16 self-organizing-map neural network, which accelerates the matching of the 16 or 32 characteristics of the reference faces in the library. This feature reduces the processing-time by >98 %.

13.3 High-Dynamic-Range HDR Video Sensors

Standard CMOS sensors like the ones in [2–4] have pixels, which integrate charge generated by the photons so that they have a linear response to the incoming photon density. Their dynamic range is limited, at the low end by the dark current of the photodiode, and at the high end by the capacitance C of the photodiode. Once the charge has reached C × V_{DD}, where V_{DD} is the supply voltage, the pixel is white-saturated. Typical maximum charges are given as saturation in Table 13.2. The result is a dynamic range of 70 dB or 3300:1. However, a natural scene has a dynamic range 10–100-times higher. From a dark night to bright sunshine, the range is 1 Million:1. In order to handle these natural requirements, linear sensors need multiple exposures, serially in time or with several pixels in parallel, inflating data and processing-tasks, and causing motion blur as well as other artifacts (see Chap. 14). The human visual system (HVS) has a natural response characteristic, which can handle an instant dynamic range of close to 1 Million:1 due to its logarithmic response, as discussed in Chap. 12, with many powerful features that form the content of [1, 5, 6]. It is a gift of MOS-transistor characteristics that they replicate this logarithmic optoelectronic conversion function (OCF) in their current-voltage characteristic, if they operate in a pixel as a source-follower in the sub-threshold mode [1, 7]. Figure 13.1 shows the HDRC pixel with sample-and-hold for a global-shutter sensor [8].

The 2014 state-of-the-art of high-speed HDR CMOS sensors with >120 dB intra-scene dynamic range is represented by a 1296 × 1092 pixels sensor with global shutter, 12b digital output up to 80 Mpix/s and a sensitivity of 0.6 mlux, basically not requiring the control of any integration or shutter time.

The building blocks and architecture of the imager, e.g. the column parallel A/D converters, were designed such that up to a full HDTV 1080p resolution

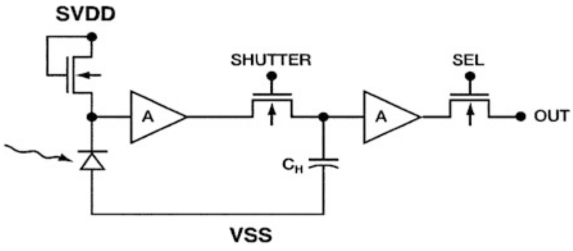


Fig. 13.1 A pixel in the HDRC sensor with global shutter [8]

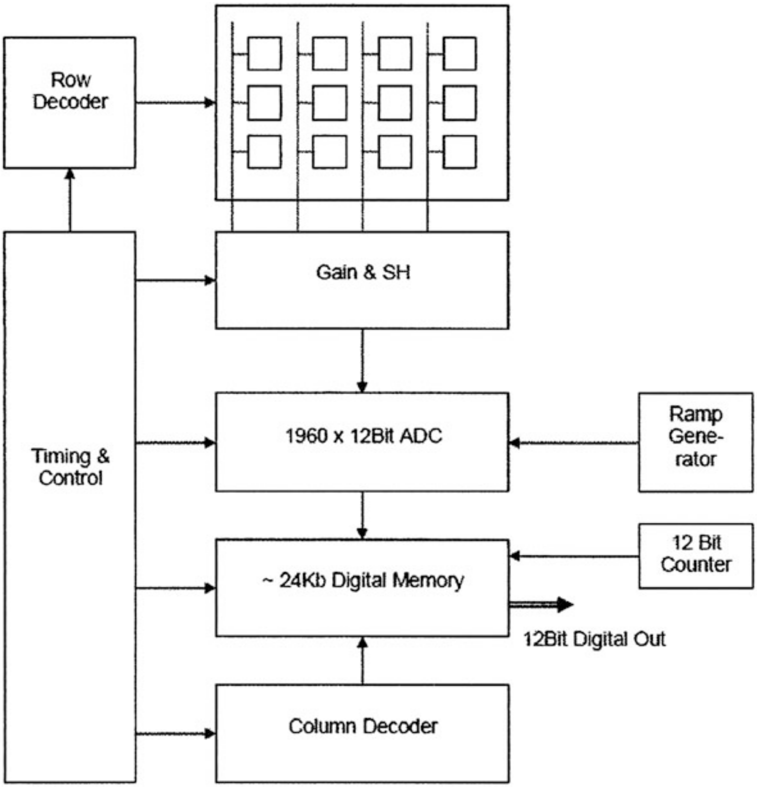


Fig. 13.2 Block diagram of HDR CMOS sensor with global shutter and 1000 fps ROI 12b read-out [8]

(1920 × 1080 pixels) could be easily achieved by abutting additional columns to the pixel array. This is shown in the block diagram in Fig. 13.2.

A high sensitivity <1 mlux at a high speed of 12.5 ns/pixel was achieved with 6.7 × 6.7 μm^2 pixels. This allows 60 frames/s @ SXGA resolution (1280 × 1024 pixels) and up to 1000 fps for 200 × 100 pixels regions-of-interest for professional applications in automation, robotics, surveillance and automotive.

The characteristic in Fig. 13.3 spans an illumination range from 0.1 mlux to 10 klux, 10 orders-of-magnitude, respectively 30 bits or f-stops, with a digital output range of 700 DN (digital numbers) from the 10b outputs. This is a remarkable compression of data, still with a contrast sensitivity, the derivative of Fig. 13.3, of 75DN/decade in the mid-range, corresponding to a contrast resolution or just-noticeable difference (JND) of 1.5 % (see Fig. 13.4).

This log- and sigmoid response has many other desirable features, and it is ideally suited for efficient video coding and processing, as discussed further in the following Chap. 14 (Table 13.3).

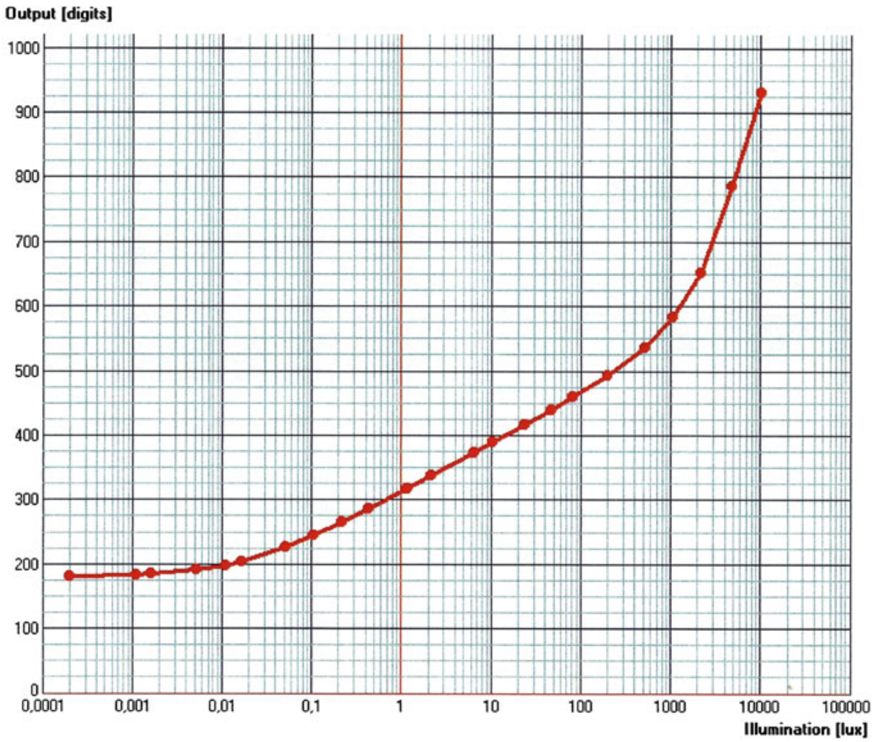


Fig. 13.3 Optoelectronic conversion function (OECF) and digital output of the HDRC video sensor at 33 fps [8]

13.4 HDRC Stereo Cameras

3D scene recording advanced significantly in recent years along the two paths of stereo imaging or time-of-flight (TOF) recording. The most powerful and robust stereo camera is built with three HDRC sensors in the SAFETYEYE camera with unique features [9]:

- x- and y-direction stereo axis
- HDRC 120 dB dynamic range, fixed-aperture photometric sensing
- High-speed global shutter with 12.5 ns/pixel, no integration times.

The SAFETYEYE scene in Fig. 13.5 illustrates the required robustness of video systems supporting the safe cognitive interaction (Chap. 18) of intelligent man-machine cooperation.

The INSERO3D stereo camera with two HDRC sensor chips meets the performance requirements for high-dynamic-range, high-speed photometric vision and, in 2015, became a key enhancement (Fig. 13.6) of the 12 DOF (degrees-of-freedom) bionic handling assistant ROBOTINO [10].

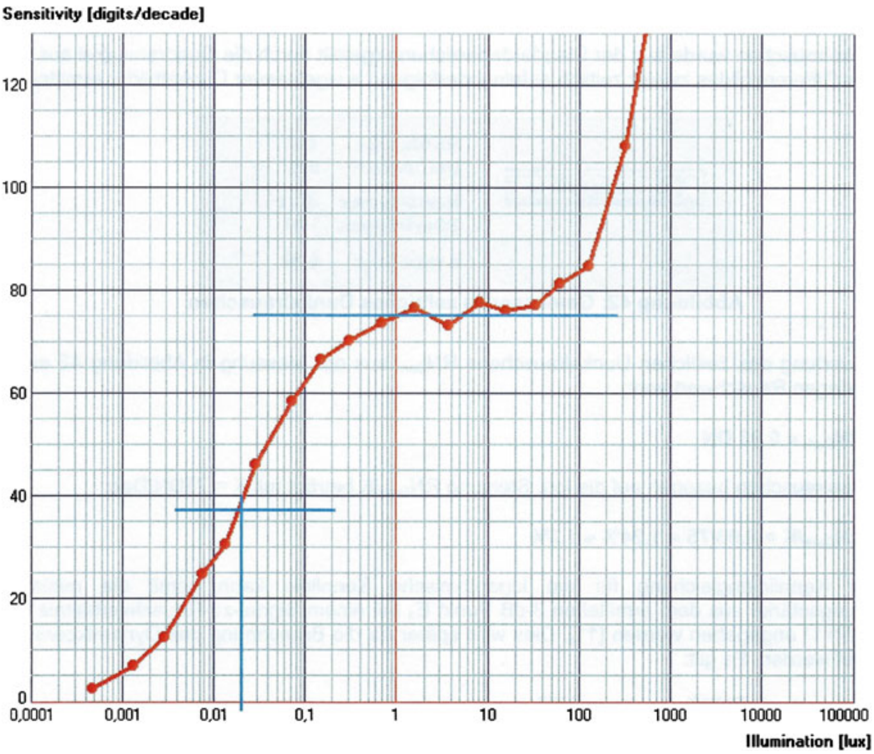


Fig. 13.4 Contrast sensitivity function (CSF) of the HDRC sensor [8]. The dynamic range of 10^7:1 shows a contrast with a sigmoid shape with natural low-light section, a central flat section with 75DN/decade, corresponding to a contrast resolution of 1.5 %, and an upper saturation section towards "white" saturation at ~1000 lux (on the sensor surface)

Table 13.3 Technical Data of the High-Speed HDRC Video Sensor with Global Shutter [8]

Sensor resolution (pixels)	1296 × 1092
Diagonal @SXGA	2/3 in.
Technology	180 nm
Pixel area (μm^2)	6.7 × 6.7
Intra-scene dynamic range	>120 dB
Min. detectable	=0.6 mlux
Sensitivity mid-range	75DN/decade
	=1.5 %
Clock	80 MHz
Pixel rate	80 Mpix/s
Pixel time	12.5 ns
Global-shutter read-out	
A/D converters	1296 × 12b
Regions of interest	Down to 200 × 100 pix
Frame rate @SXGA	60 fps
Max.	1.000 fps

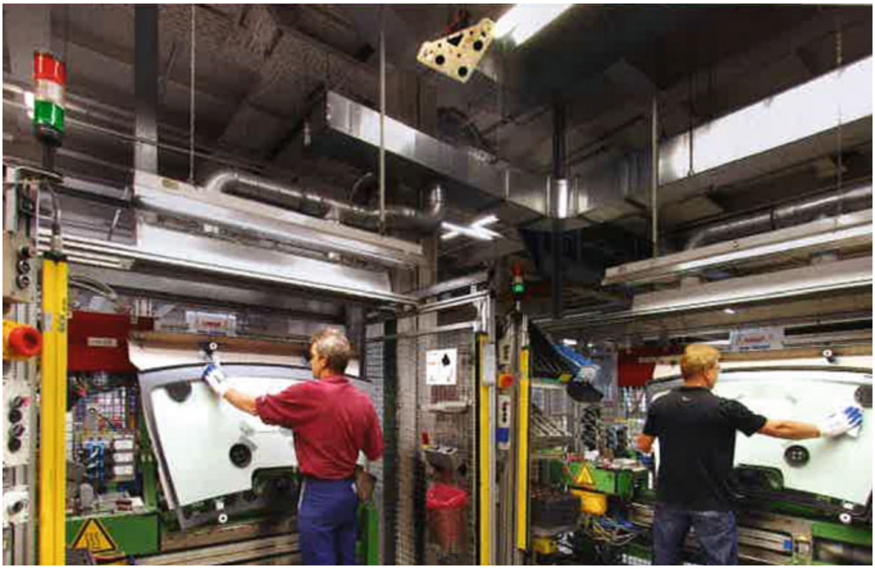


Fig. 13.5 The SAFETYEYE two-axis stereo camera with three HDRC sensors monitoring the safe operations of two operators in the man-machine directed assembly of wind-shields for automobiles [11]

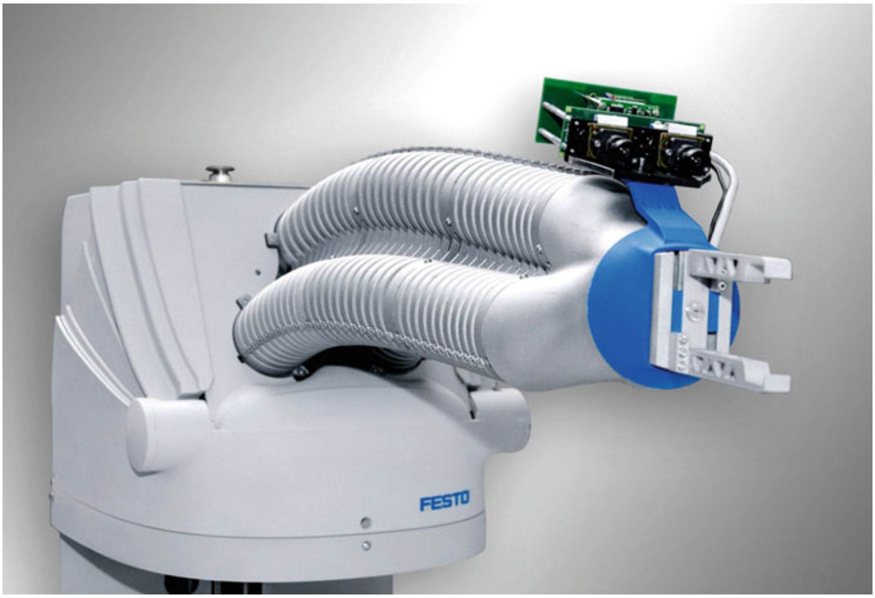


Fig. 13.6 The INSERO3D high-dynamic-range, high-speed, photometric stereo video camera on the bionic handling assistant ROBOTINO (Image courtesy FESTO AG & Co.KG)

Table 13.4 Performance of 3D TOF sensors

Year [Reference]	2004 [11]	2010 [14]	2012 [15]	2012 [16]	2014 [17]
Resolution (pixels^2)	32 × 32	256 × 256	500 × 274	480 × 270	413 × 240
Range (m)	3	0.4	0.75–4.5	1.0–7.0	0.8–1.8
Depth resolution (mm)	1.8	0.5	10–38	5–160	4.0–6.4

13.5 3D Time-of-Flight (TOF) Sensors

A TOF camera of 2004 with 32 × 32 pixels with avalanche photo-diodes, reviewed in [1], had a range of 3 m with a range uncertainty of 1.8 mm. The evolution since 2004 is represented in Table 13.4.

Sophisticated pulsed-illumination and detection techniques led early to remarkable range resolutions, and the sensors since 2012 run programmable modes of standard color imaging and 3D TOF frames simultaneously. This intelligent 3D vision is essential for augmented reality, automation, robotics and autonomous mobility.

13.6 Conclusion

Vision is the most important one of our senses. The sensor resolution in pixels continues to increase with the result that they contribute to the explosion of video data on the Internet. From intelligent acquisition in the pixel through coding, pattern recognition, compression and tone mapping for display, possibly directly in the focal- or near the focal-plane with 3D integration and brain-inspired architectures are essential tasks as described in the following Chaps. 14 and 18.

References

1. Hoefflinger, B.: Vision sensors and cameras, Chap. 15. In: Hoefflinger, B. (ed.) CHIPS 2020—A Guide to the Future of Nanoelectronics, pp. 37–93. Springer, Berlin (2012). doi:10.1007/978-3-642-23096-7_15
2. Wakabayashi, H., et al.: A 1/2.3-inch 10.3 Mpixel 50 frame/s back-illuminated CMOS image sensor. In: IEEE ISSCC (International Solid-State Circuits Conference) 2010, Digest Technical Papers, pp. 410–411
3. Ahn, J.C. et al.: A 1/4-inch 8 Mpixel CMOS image sensor with 3D backside-illuminated 1.12 μm Pixel with front-side deep-trench isolation and vertical transfer gate. In: 2014 International Solid-State Circuits Conference, Digest Technical Papers, pp. 124–125, Feb 2014
4. Shi, C. et al.: A 1000 fps vision chip based on a dynamically reconfigurable hybrid architecture comprising a PE array and self-organizing map neural network. In: 2014 International Solid-State Circuits Conference, Digest Technical Papers, paper 7.3, pp. 128–129, Feb 2014

5. Hoefflinger, B. (ed.): High-Dynamic-Range (HDR) Vision. Springer, Berlin (2007). (ISBN-13 978-3-540-44432-9)
6. Mantiuk, R., Myszkowski, K., Seidel, H.P.: High Dynamic Range Imaging. Wiley Encyclopedia of Electrical and Electronics Engineering, New York (2015)
7. Hoefflinger, B., Seger, U., Landgraf, M.E.: Image cell for an image recorder chip, US patent 5608204, filed 03-23-1993, issued 03-04-1997
8. Strobel, M. et al.: High-Dynamic-Range Low-Noise (HIDRALON) CMOS Imagers, Schlussbericht BMBF-Verbundprojekt, 54 pp., Bundesministerium für Bildung und Forschung, Germany (2014)
9. "Mit dem Dritten sieht man besser", Roboter-Ueberwachung mittels Bildverarbeitung, Daimler-Chrysler Hightech Report, no2. 2006, pp. 34–39. 0874_001 (1). Pdf
10. http://www.festo.com/net/supportportal/File/46266/Brosch-robotino
11. Niclass, C., Rochas, A., Besse, P.A., Charbon, E.: A CMOS 3D camera with millimetric depth resolution. In: Proceedings IEEE Custom Integrated Circuits Conference (CICC), pp. 705–708 (2004)
12. Cheng, C. et al.: iVisual: an intelligent visual sensor SoC with 2790 fps CMOS image sensor and 205GOPS/W vision processor. In: 2008 International Solid-State Circuits Conference, Digest Technical Papers, pp. 306–307, Feb 20008
13. Zhang, W., et al.: A programmable vision chip based on multiple levels of parallel processors. IEEE J. Solid-State Circ. **46**, 2132–2147 (2011)
14. Mandai, S., Ikeda, M., Asada, K.: A 256 256 14 k range maps/s 3-D range-finding image sensor using row-parallel embedded binary search tree and address encoder. In: IEEE ISSCC (International Solid-State Circuits Conference) 2010, Digest Technical Papers, pp. 404–405
15. Han S.-M. et al.: A 413 × 240 Pixel sub-centimeter resolution time-of-flight image sensor. In: 2014 International Solid-State Circuits Conference, Digest Technical Papers, pp. 10–131, Feb 2014
16. Kim S.J. et al.: A 1920 × 1080 3.65 µm-Pixel 2D/3D image sensor with split and binning pixel structure in 0.11 µm standard CMOS
17. Kim W. et al.: A 1.5 Mpixel RGBZ cmos image sensor for simultaneous color and range image capture, ISSCC Digest Technical Papers, pp. 392–393, Feb. 2012

Chapter 14
Perception-Inspired High Dynamic Range Video Coding and Compression

Rafał K. Mantiuk and Karol Myszkowski

Abstract High-Dynamic Range (HDR) images and video can represent much greater color gamut, brightness and contrast than commonly used standard dynamic range images. When HDR video is presented on specialized HDR displays, a substantial increase of realism can be observed, sometimes compared to "looking through a window". Efficient encoding of such video, however, imposes new challenges. In this chapter we argue that adjusting the accuracy of broadcasted HDR video to the capabilities of the human visual system is the key requirement to achieve these goals. Another example of HDR signal is a depth image used in stereoscopic and multi-view imaging systems. When encoding is designed to capitalize from the characteristics and limitations of depth perception, significant compression gains can be achieved.

14.1 Introduction

With increasing computing power, improving storage capacities, and development of wireless and wired network systems, the role of video-based applications increases. Video becomes common in multimedia messaging, video teleconferencing, gaming through remote servers, video streaming and TV signal broadcasting. While in the nearest future a standard video, encoding 8-bit color per-channel, will continue to dominate in all these applications, high-dynamic-range video (HDRV) is likely to become the next step in video evolution, which will enable to handle luminance and chrominance data with much higher precision.

R.K. Mantiuk (✉)
School of Computer Science, Bangor University, Dean St., Bangor LL57 1UT, UK
e-mail: mantiuk@gmail.com; mantiuk@bangor.ac.uk

K. Myszkowski
Department 4: Computer Graphics, Max-Planck-Institute for Informatics, Campus E1 4, 66123 Saarbruecken, Germany

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_14

HDRV employs the so-called *scene-referred* representation of frames, which encodes the approximate photometric characteristic of a scene the video depicts. Since such representation imposes unacceptable storage costs when stored in its native floating-point format, it is important to find more efficient encodings. In this chapter we argue that for efficient scene-referred frame representations the accuracy should not be tailored to any particular imaging technology, which is a common practice today, but it should be tailored to the capabilities and limitations of the human visual system (HVS). First, we present the basic principles behind such HVS-driven representations for HDR pixels (Sect. 14.1), then we demonstrate how modern video compression standards such as MPEG or H.264 can be adapted to accommodate such encodings (Sect. 14.2), including backward-compatible solutions that rely on standard 8-bit frame formats (Sect. 14.3). Before concluding this chapter, we discuss also the compression of dynamic per-pixel depth maps, where the range of depth values makes the resulting video yet another instance of HDR signal. Perception considerations, which are specific for such depth signal, enable more efficient compression (Sect. 14.4).

14.2 HDR Pixel Encoding

The specific choice of pixel encoding used for video compression has a great impact on the compression performance, where the goal is to minimize the number of required bits while providing sufficient accuracy and capability to encode a high dynamic range. If the bit-depth accuracy is too low, banding (quantization) artefacts become visible. If it is too high, invisible noise is introduced into the signal and the efficiency of compression is reduced. In this section, we focus on perceptually uniform encoding, which is based on the idea of aligning the quantization errors with the HVS sensitivity to luminance increments as a function of luminance adaptation levels. For a more complete account of popular HDR pixel encodings, the reader is referred to [1, Sect. 5.1].

Low-dynamic-range (LDR) pixel values, as in the standard JPEG and MPEG formats, have a desirable property that their values are approximately linearly related to perceived brightness of those pixels. Because of that, LDR pixel values are also well suited for image encoding since the distortions caused by image compression have the same visual impact across the whole scale of *signal* values. HDR pixel values lack such a property and, therefore, when the same magnitude of luminance distortion is introduced in low-luminance and high-luminance image regions, the artefacts are more visible in the low-luminance regions. The problem is alleviated, if the logarithm of luminance is encoded instead of luminance [2]. But the logarithmic encoding is not a precise model of the HVS sensitivity to light, which in practice is much lower for the low luminance ranges, where Weber's law does not hold (below 10 cd/m^2). For that reason, several encodings were proposed that employ more accurate models of the eye sensitivity to light changes [3–5].

The derivation of such perceptually uniform encoding is essentially the same as the derivation of the response of the HVS to light by means of the transducer function [6, Sect 4.1]. The transducer function maps physical luminance (in) into the units related to the just-noticeable-differences (JNDs). The first such encoding in the context of HDR compression was proposed in [4], where the threshold vs. intensity function (t.v.i.) was used to determine the smallest noticeable difference in luminance across the luminance range. The paper showed that 11–12 bits are sufficient to encode the range of luminance from 10^{-4} to 10^{8} cd/m^2. The follow-up paper [3] replaced the t.v.i. function with a more modern model of the contrast sensitivity (CSF) [7]. Recently, a similar idea was proposed in the context of encoding HDR images to the Society of Motion Picture and Television Engineers [5] using Barten's CSF [8]. The comparison of those recent encodings is shown in Fig. 14.1. Note that the perceptual encoding curves are located between the logarithmic encoding and the Gamma 2.2 encoding. The latter encoding is commonly used for LDR pixels. From Fig. 14.1, it becomes immediately clear that the logarithmic encoding allocates too much dynamic space for low luminance levels, at the expense of risking quantization errors for high luminance levels. On the other hand, Gamma 2.2 encoding provides excessive precision at such bright regions, while risking artifacts for low luminance levels. A number of such encodings were recently evaluated using psychophysical methods [9]. The study demonstrated that JND-based encodings offers a more uniform distribution of perceivable artifacts

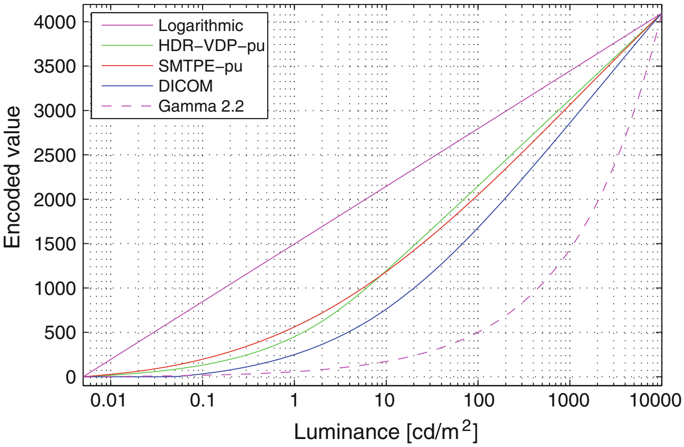


Fig. 14.1 Functions mapping physical luminance into encoded 12-bit luma values. Logarithmic—is the logarithm of luminance; HDR-VDP-pu is the perceptually uniform color space derived from the contrast sensitivity function (CSF) that is employed in HDR-VDP-2 [25]; SMTPE-pu—is the perceptually uniform encoding derived from Barten's CSF; DICOM is the DICOM gray-scale function, also derived from Barten's CSF but using different parameters; "Gamma 2.2" is the typical gamma encoding used for LDR images, but extended to the range 0.005–10,000 cd/m^2. (Reproduced with permission from [10] © Wiley Encyclopedia of Electrical and Electronics Engineering.)

across luminance and require fewer bits to encode, though logarithmic encoding is a viable option if simplicity is a priority. A natural JND-like characteristic is achieved directly in the HDRC pixel (Chap. 13).

14.3 High Bit-Depth Compression

The extension of the existing video compression standards to support HDR is illustrated in Fig. 14.2, and, as can be seen, the scope of required changes is relatively modest [4]. While a standard MPEG4/H.264 encoder takes as input three 8-bit RGB color channels, the HDR encoder must be provided with pixel values in the absolute XYZ color space or linear HDR RGB. Such color spaces can represent the full color gamut and the complete range of luminance the eye can adapt to. Next, pixel values are transformed to the color space that improves the efficiency of such encoding as discussed in Sect. 14.1: HVS-based-encoding in the context of luma (refer to [10, Sect. 4.1] for information on chroma processing). This not only reduces the number of required bits, but it also improves perceptual uniformity of introduced distortions.

Perceptually uniform encoding might require up to 4096 quantization levels (refer to Fig. 14.1). This directly translates into 12-bit encoding, which, in terms of the bit depth, is much higher than commonly used 8 bits. Fortunately, modern compression standards typically have an optional support for higher bit-depths, which allows an easy extension for HDR content. For example, the high-quality content profiles introduced in the MPEG4-AVC/H.264 video coding standard allow to encode up to 14 bits per color channel [11].

Due to quantization of DCT coefficients, noisy artifacts may appear near edges of high-contrast objects. This problem is especially apparent for HDR video, in particular for synthetic sequences, where the contrast tends to be higher than in natural LDR video. This can be alleviated by encoding sharp-contrast edges in each macro-block separately from the rest of the signal. An algorithm for such hybrid encoding can be found in [4].

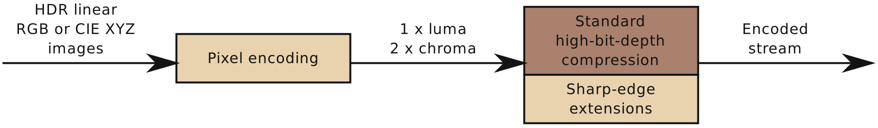


Fig. 14.2 Encoding HDR video content using selected profiles of H.264 for high-bit-depth encoding. The HDR pixels need to be encoded into one luma and two chroma channels to ensure good decorrelation of color channels and perceptual uniformity of the encoded values. The standard compression can be optionally extended to provide better coding for sharp-contrast edges [4]. (Reproduced with permission from [10] © Wiley Encyclopedia of Electrical and Electronics Engineering.)

14.4 Backward-Compatible Compression

Since the standard LDR video formats, such as MPEG and H.264, have become widely adapted and are supported by almost all software and hardware equipment dealing with digital imaging, it cannot be expected that these formats will be immediately replaced with their HDR counterparts. To facilitate the transition from the traditional to HDR imaging, there is a need for backward compatible HDR formats, that would be fully compatible with existing LDR formats and, at the same time, would support enhanced dynamic range and color gamut. Moreover, if such a format is to be successful and adopted, the overhead of HDR information must be low. An example of a backward-compatible HDR encoding is a new JPEG XT standard, which was designed to use 2 standard 8-bit encoders or decoders to store HDR images.

Most backward-compatible compression methods follow similar processing scheme, shown in Fig. 14.3. The following paragraphs discuss this scheme while referring to concrete solutions and possible variants.

Some backward-compatible compression methods expect both HDR and LDR frames to be supplied separately as input [12]. Other methods provide their own tone-mapping operators [10] (see step 1 in the diagram) and expect only HDR frames as input. The latter approach allows to adjust tone-mapped images to improve compression performance. For example, the JPEG-HDR method can introduce a *precorrection* step [13], which compensates for the resolution reduction introduced at the later stages of the encoding. Mai et al. [14] derived a formula for an optimum tone-curve, which minimizes compression distortions and improves compression performance. The drawback of such compression-driven tone-mapping operators is that they introduce changes to the backward-compatible portion of the video, which may not be acceptable in many applications.

The LDR frames are encoded using a standard codec, such as MPEG or H.264 (see step 2 in the diagram) to produce a backward-compatible stream. In order to use this stream for the prediction of the HDR stream, it is necessary to decode those frames (step 3), which can be done efficiently within the codec itself. Then, both LDR and HDR frames need to be transformed into a color space that would bring LDR and HDR color values into the same domain and make them comparable and

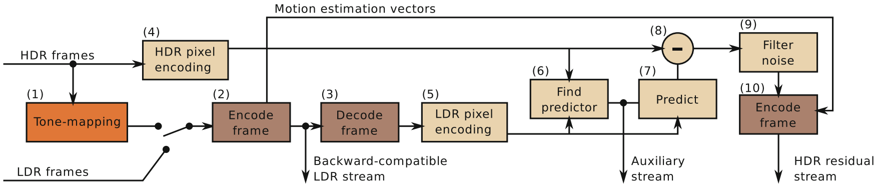


Fig. 14.3 Typical encoding scheme for backward-compatible HDR compression. The *darker brown boxes* indicate standard (usually 8-bit) image of a video codec, such as H.264. (Reproduced with permission from [10] © Wiley Encyclopedia of Electrical and Electronics Engineering.)

easy to decorrelate (steps 4 and 5). Most of the pixel encodings discussed in Sect. 14.1 can be used for that purpose. For example, HDR-MPEG method [12] employs perceptually uniform coding. This step, however, may not be sufficient to eliminate all redundancies between LDR and HDR streams, as they could be related in a complex and non-linear manner. For that purpose, some encoding schemes find an optimal predictor function (step 6), which can be used to predict HDR pixel values based on LDR pixel values (step 7). Such a predictor could be a single tone-curve provided for the entire image [12, 15] or a simple linear function, but computed separately for each macro-block [16]. In the next step (8), the prediction from such a function is subtracted from the HDR frame to compute a residual that needs to be encoded.

The resulting residual image may contain a substantial amount of noise, which is expensive to encode. For that reason, some methods employ a filtering step (9), which could be as simple as low-pass filtering and reducing resolution [13], or as complex as modeling the visibility of the noise in the HVS and removing invisible noise [12].

Finally, the filtered frame is encoded to produce an HDR residual stream (step 10). Although the encoding of the residual stream is mostly independent of the LDR stream, it is possible to reuse the motion vectors stored in the LDR stream and thus reduce both storage overhead and the computing required to find motion vectors for the residual.

14.5 Perceptual Depth Compression for Stereoscopic Applications

Stereoscopic 3D rendering is an important trend towards improving the quality of real-world depiction. Autostereoscopic displays enable glasses-free binocular depth vision at the expense of a larger number of views. Color + depth image formats proved the most versatile in terms of retargeting stereo 3D content for any type of display and viewing conditions (e.g., movie theater vs. cell phone screen), or generating on-the-fly multiple views as required by autostereoscopic displays [17]. Depth-image-based rendering (DIBR) [18] efficiently enables all such adjustments through image warping. Broadcasting and storage of depth data calls for lossy depth map compression, which, similar to the HDRV compression, is the most efficient for depth encodings as fixed precision integers. Another similarity to HDRV encodings relies on the large depth ranges, which make the depth signal high dynamic range.

Per-pixel depth maps have some specific signal characteristics. Depth maps are usually smooth with abrupt discontinuities at object silhouettes [19], and they exhibit significant spatial and temporal redundancies, which are easy to handle using existing video coding standards, such as H.264 after some customization. This involves representing depth data in a perceptually linear domain (similar to luminance encodings in Sect. 14.1), and removing depth information that cannot be perceived by the human observer. For this purpose, instead of a direct depth signal encoding, depth is first converted into the corresponding vergence angle between the two eyes. For stereo 3D applications, where display parameters are usually known in advance, such a representation can be considered a measure of physically (and perceptually) significant depth. For example, depth changes at large absolute depths correspond to small changes in the vergence angle, which better approximates the HVS response to such depth variations.

Pajak et al. [20] observed that color and depth signals are strongly correlated, and in particular color and depth discontinuities often share their spatial locations. On the other hand, if a color edge is missing, the depth difference is often not visible either, as a certain magnitude of color contrast is required for depth perception [21]. In such a case, the accuracy of depth signal encoding can be relaxed. All these observations justify depth downsampling at the compression stage and its subsequent upsampling at the decompression stage, where depth discontinuities are reconstructed from the compressed color signal. This way, color and depth signal are decorrelated. As shown in Fig. 14.4, standard codecs, such as H.264, can be used for efficient compression of such downsampled depth signals.

There is another perception-based motivation, why such a low-resolution depth-map stream is sufficient to reconstruct high-quality depth [20]. For image/video coding, the typical size of 8 × 8 pixels for a DCT block matches well the HVS contrast sensitivity in color vision, which roughly ranges from 1 to 30–50 cpd (cycles per degree). A typical display device shows signals with up to 20 cpd. Therefore, an 8 × 8 pixel DCT block is best at encoding luminance in the range of 2.5 cpd corresponding to the lowest-frequency AC coefficient (half a cycle per block) to 20 cpd (highest frequency AC coefficient). However, the HVS frequency-sensitivity to depth/disparity is much lower and ranges from 0.3 to 5 cpd [22, Fig. 14.1]. (Note that the disparity between two objects located at different depths can be measured as the difference of corresponding vergence angles.) Therefore, an 8-pixel-wide block does not only encode very low frequencies badly, but it also wastes information by encoding frequencies above 5 cpd that do not improve the perceived stereo quality. This suggests that one could downsample the

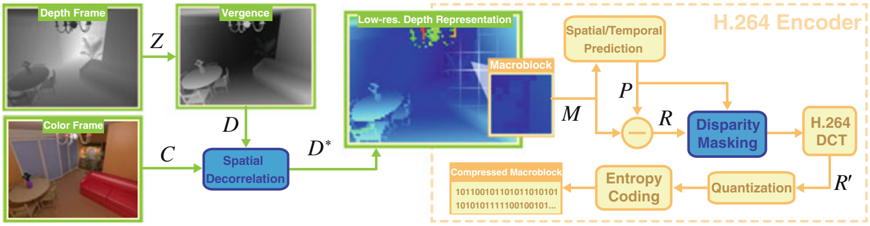


Fig. 14.4 Outline of a perception-based depth compression framework. H.264 compression tools are framed in *orange*, perception-driven enhancements in *blue*. (Image courtesy of Dawid Pajak. Reproduced with permission from [20] © The Eurographics Association 2014.)

depth image by a factor of $k = 4$ before encoding it, which would aim at frequencies in the range from 0.625 to 5 cpd.

To further improve the depth compression performance, another perceptual effect is considered. Disparity masking (refer to Fig. 14.4) reduces the visibility of small depth variations in the presence of a strong depth signal. The residual signal R (refer to Fig. 14.4), which is the difference between the original macroblock signal M and its spatio-temporal prediction P, is filtered based on the model of masking. The magnitude of masking depends on the predicted depth signal P. This means that the precision of the residual R encoding is further reduced in strong depth masking regions.

14.6 Conclusion

It is quite surprising that the well-studied and improved-over-years video compression standards may turn out to be inadequate for new content and displays in the coming years. Although increasing the bit-depth of encoded images seems to be the most apparent solution to this problem, it does not address the major issue: how the encoded code-values should be mapped to the luminance levels produced by a display. The standard color spaces (e.g. rec.709), commonly used for this purpose in low dynamic range images, have been designed for low-dynamic-range displays and reflective print colorimetry, and they are not suitable for high-contrast displays. The problem is even more difficult, if the output device is unknown and may vary from a low-contrast mobile display to a high-end larger-screen display. To fully address this issue, not only the compression algorithms, but the entire imaging pipeline [1], from acquisition to display-adaptive tone-mapping, must be redesigned. Perceptually uniform HDR pixel encodings (Sect. 14.1) offer a general-purpose intermediate storage format, which can represent the colorimetrically calibrated images with no display limitations. Such images could be displayed only on an ideal display, capable of producing all physically feasible colors, which is unlikely to ever exist. Therefore, the HDR video must be adjusted to the actual display capabilities by compressing its dynamic range, clipping excessively bright pixels, choosing the right brightness level, so that all colors fit into the display color gamut. Tone-mapping operators [10, Chap. 5] are intended to achieve these goals, and, while a vast majority of existing operators can handle only static images, some of them are designed specifically for HDR video [23, 24]. Since making radical changes in imaging pipelines in terms of the required bit depth (Sect. 14.2) might render the existing software and hardware obsolete, it is important to ensure backward-compatibility of image and video formats, as discussed in Sect. 14.3. Per-pixel depth maps are yet another example of HDR signal, where specifics of depth perception must be considered to improve the compression efficiency (Sect. 14.4).

References

1. Myszkowski, K., Mantiuk, R.K., Krawczyk, G.: High dynamic range video. In: Synthesis Digital Library of Engineering and Computer Science. Morgan & Claypool Publishers, San Rafael, USA (2008)
2. Mantiuk, R.K., Myszkowski, K., Seidel, H.P.: Lossy compression of high dynamic range images and video. In: Proceedings of Human Vision and Electronic Imaging XI, volume 6057 of Proceedings of SPIE, page 60570 V, San Jose, USA. SPIE, Bellingham (2006)
3. Mantiuk, R., Daly, S.J., Myszkowski, K., Seidel, H.P.: Predicting visible differences in high dynamic range images: model and its calibration. In: Human Vision and Electronic Imaging 204–214 (2005). http://scholar.google.co.uk/scholar?hl=en&as_sdt=2000&q=visual+difference+predictor+high+dynamic+range+mantiuk#0
4. Mantiuk, R.K., Krawczyk, G., Myszkowski, K., Seidel, H.P.: Perception-motivated high dynamic range video encoding. ACM Trans. Gr. (Proceedings of SIGGRAPH), **23**(3), 730–738 (2004)
5. Miller, S., Nezamabadi, M., Daly, S.: Perceptual signal coding for more efficient usage of bit codes. SMPTE Motion Imag. J. **122**(4), 52–59 (2013)
6. Reinhard, E., Ward, G., Debevec, P., Pattanaik, S., Heidrich, W., Myszkowski, K.: High Dynamic Range Imaging, 2nd edn. Morgan Kaufmann Publishers (2010)
7. Daly, S.: The visible differences predictor: an algorithm for the assessment of image fidelity. In: Andrew B. Watson (ed.), Digital Images and Human Vision, pp. 179–206. MIT Press, Cambridge (1993)
8. Barten, P.G.J.: Formula for the contrast sensitivity of the human eye. In: Miyake, Y., Rene Rasmussen, D. (eds.): Proceedings of SPIE 5294, Image Quality and System Performance, pp. 231–238 (2004)
9. Boitard, R., Mantiuk, R.K., Pouli, T.: Evaluation of color encodings for high dynamic range pixels. In: Human Vision and Electronic Imaging, p. 93941K (2015). http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2077715
10. Mantiuk, R.K., Myszkowski, K., Seidel, H.P.: High dynamic range imaging. In: Webster, J.G. (ed.) Wiley Encyclopedia of Electrical and Electronics Engineering. Wiley, New York (2015)
11. Sullivan, G.J., Yu, H., Sekiguchi, S., Sun, H., Wedi, T., Wittmann, S., Lee, Y., Segall, A., Suzuki, T.: New standardized extensions of MPEG4-AVC/H. 264 for professional-quality video applications. In: Proceedings of ICIP'07 (2007)
12. Mantiuk, R.K., Efremov, A., Myszkowski, K., Seidel, H.P.: Backward compatible high dynamic range mpeg video compression. ACM Trans. Gr. (Proceedings of SIGGRAPH), **25** (3) (2006)
13. Ward, G., Simmons, M.: Subband encoding of high dynamic range imagery. In: APGV '04: 1st symposium on applied perception in graphics and visualization, pp. 83–90 (2004)
14. Mai, Z., Mansour, H., Mantiuk, R.K., Nasiopoulos, P., Ward, R., Heidrich, W.: Optimizing a tone curve for backward-compatible high dynamic range image and video compression. IEEE Trans. Image Process. **20**(6), 1558–1571 (2011)
15. Winken, M., Marpe, D., Schwarz, H., Wiegand, T.: Bit-depth scalable video coding. In: 2007 IEEE International Conference on Image Processing, volume 1, pages I-5–I-8. IEEE, New York (2007)
16. Segall, A.: Scalable coding of high dynamic range video. In: 2007 IEEE International Conference on Image Processing, vol. 1, pp. I-1–I-4 (2007)
17. Daly, S.J., Held, R.T., Hoffman, D.M.: Perceptual issues in stereoscopic signal processing. IEEE Trans. Broadcast. **57**(2), 347–361 (2011)
18. Fehn, C.: Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV. In: Stereoscopic Displays and Virtual Reality Systems XI, vol. 5291, pp. 93–104. SPIE, Bellingham (2004)

19. Merkle, P., Morvan, Y., Smolic, A., Farin, D., Müller, K., de With, P.H.N., Wiegand, T.: The effects of multiview depth video compression on multiview rendering. Signal Proc. Image Commun. **24**(1–2) (2009)
20. Pajak, D., Herzog, R., Mantiuk, R., Didyk, P., Eisemann, E., Myszkowski, K., Pulli, K.: Perceptual depth compression for stereo applications. Comput. Gr. Forum (Proceedings of Eurographics 2014), **33**(2), 195–204 (2014)
21. Didyk, P., Ritschel, T., Eisemann, E., Myszkowski, K., Seidel, H.P., Matusik, W.: A luminance-contrast-aware disparity model and applications. ACM Trans. Graph. (Proceedings of SIGGRAPH Asia), **31**(6) Article No. 184 (2012)
22. Bradshaw, M.F., Rogers, B.J.: Sensitivity to horizontal and vertical corrugations defined by binocular disparity. Vision. Res. **39**(18), 3049–3056 (1999)
23. Aydn, T.O., Stefanoski, N., Croci, S., Gross, M., Smolic, A.: Temporally coherent local tone mapping of HDR video. ACM Trans. Graph. (Proc. of SIGGRAPH Asia) **33**(6), 196:1–196:13 (2014)
24. Eilertsen, G., Wanat, R., Mantiuk, R.K., Unger, J.: Evaluation of tone mapping operators for hdr-video. Comput. Gr. Forum **32**(7), 275–284 (2013)
25. Mantiuk, R.K., Kim, K.J., Rempel, A.G., Heidrich, W.: HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. ACM Trans. Gr. (Proceedings of SIGGRAPH), **30**(4), 40:1–40:14 (2011)

Chapter 15
MEMS—Micro-Electromechanical Sensors for the Internet of Everything

Jiri Marek, Bernd Hoefflinger and Udo-Martin Gomez

Abstract Automotive safety applications were the cradle of the MEMS market over many years. Their persistent improvement in cost and size launched the wide application of MEMS in consumer applications as well, starting from year 2009, outpassing in numbers the automotive MEMS volume within only 3 years. The consumer MEMS annual growth rate today is 18 %. Following a steep learning-curve, their form factor was reduced by more than 10-times within 5 years. Concurrently, cost, embodied materials and energy consumption of MEMS devices came down to an extent that basic integrated units contain now a multitude of sensors, with 9 degrees-of-freedom, by 2013. Virtually unlimited new applications which are strong in sensors emerge, enabling the Internet-of-Everything.

15.1 Unique Growth of the MEMS Market

Silicon-based MEMS sensing started with acceleration and pressure sensors. They were a specialty niche in the beginning, but began a unique rise after achieving tough qualifications for the automotive environment, especially in safety-critical systems like airbags and active brakes, as well as in engine management systems for clean and environmentally friendly combustion. Beyond professional standards,

J. Marek (✉)
Robert Bosch LLC, Research and Technology Center (CR/RTC-NA),
4005 Miranda Avenue, Suite 200, Palo Alto, CA 94304, USA
e-mail: jiri.marek@us.bosch.com

B. Hoefflinger
5 Leonberger Strasse, 71063 Sindelfingen, Germany
e-mail: bhoefflinger@t-online.de

U.-M. Gomez
Bosch Sensortec GmbH (BST/NE), Gerhard-Kindler-Straße 9,
72770 Reutlingen, Germany
e-mail: udo-martin.gomez@bosch-sensortec.com

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_15

there were the automotive reliability requirements, which challenged the automotive suppliers to push the new technology by merging micromechanical and microelectronic functions into mass-products of highest reliability at competitively low cost. This development is covered in this chapter of [1]. The description there spans the framework from process technology to the high sophistication of the MEMS gyroscope functionality, sensing rotation and enabling intelligent 4-wheel acceleration/breaking action in the electronic-safety system called ESP, to keep vehicles safely on the road even in critical situations.

As it has happened to all microelectronic inventions since the transistor until today, the eventual results of sustained R&D, driven by tough requirements in the professional sectors like military, telecom, aerospace or automotive, were transformed by creative spirits into consumer electronics innovations. It was a highly productive coincidence that the automotive MEMS sensors were ready for effective mass production when the newly conceived smartphones needed on-board sensors for all sorts of handling or motion detection. MEMS entry into mobile consumer electronics (CE) launched a new dimension of growth in the MEMS market as shown in Fig. 15.1. While automotive MEMS has been achieving a respectable growth rate of 9 %/year, thus almost doubling annual integrated circuits growth, CE MEMS, mostly in mobile equipment, is about to achieve 18 %/year over the period 2011–2016. It can be seen in Fig. 15.2 that the MEMS market started noticeable growth only in 2009, and that CE MEMS fired it up and surpassed automotive MEMS in market-share already in 2012.

The development of the market share of CE MEMS is shown in Fig. 15.3 for the 5-year period 2007–2012, leading to a total of 1 Billion units in 2012.

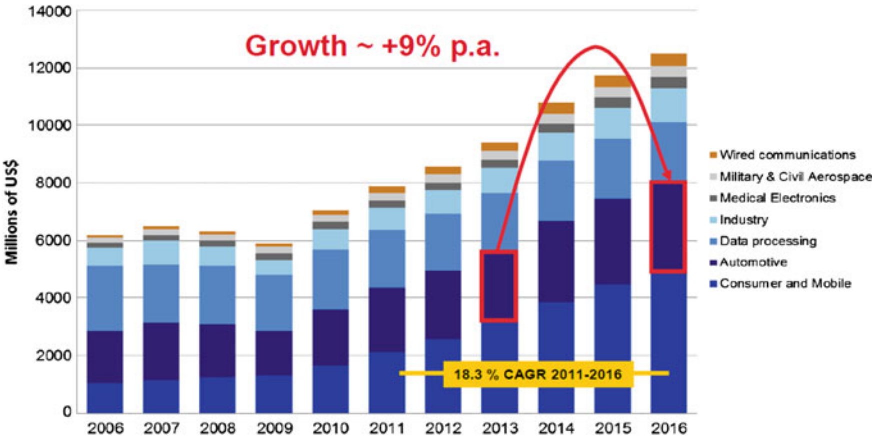


Fig. 15.1 Sensors market in Millions US$ [2]. *Source* IHS iSuppli MEMS Market Tracker—Q3 2012

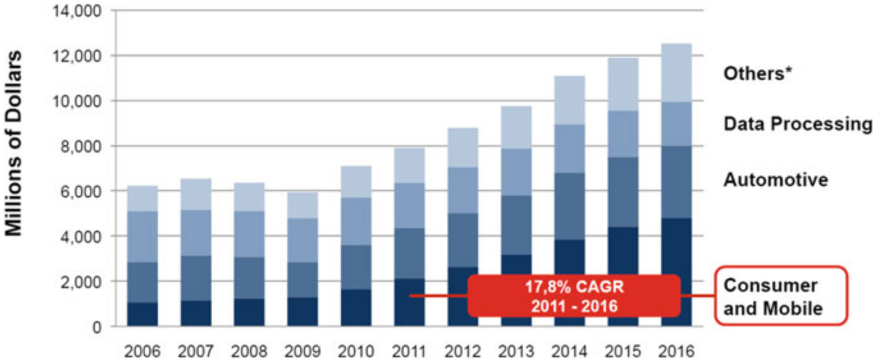


Fig. 15.2 Consumer electronics (CE), the new locomotive of the MEMS market [2]. **Others* Industry, wired communications, medical electronics, military and civil aerospace. *Source* IHS iSuppli

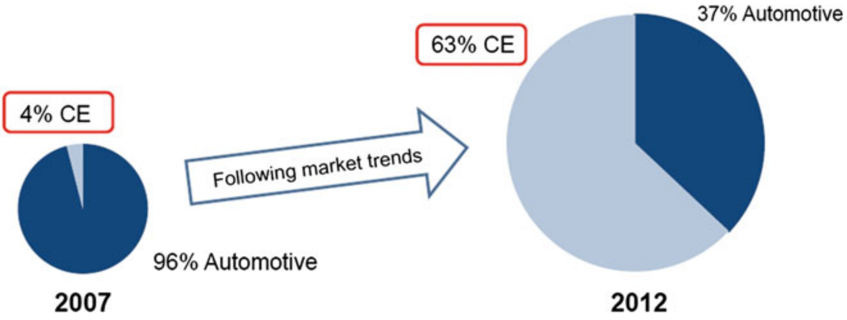


Fig. 15.3 MEMS production units at Bosch

15.2 Automotive MEMS Applications and Scaling

The realization of airbag systems was relying on high-g acceleration sensors for crash detection, which launched intensive R&D on silicon-based sensors to fulfill challenging requirements. Multi-airbag systems added even more challenging specifications for low-g sensors, detecting rather small acceleration, and angular-rate sensors for the detection of rollover. Pressure sensors were implemented for side-crash detection as well. At the same time, clean-engine management systems needed pressure and mass-flow sensors for improved control of a proper and environmentally friendly combustion within the engine.

The control of vehicle dynamics, relying on the data fusion of acceleration and yaw-rate sensor signals, as well as on-wheel rotation speed to initiate wheel-specific braking actions in case of detected driving instabilities, became the MEMS-based prime solution that caused a quantum-leap forward in the large-scale application of intelligent MEMS multisystems. Its broad introduction into the automotive market was due to its

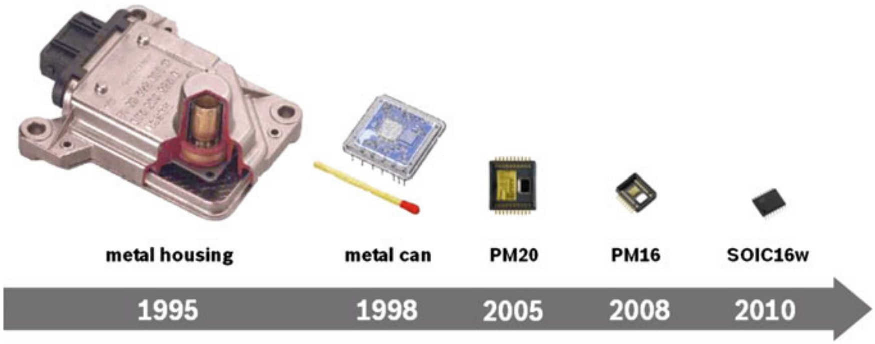


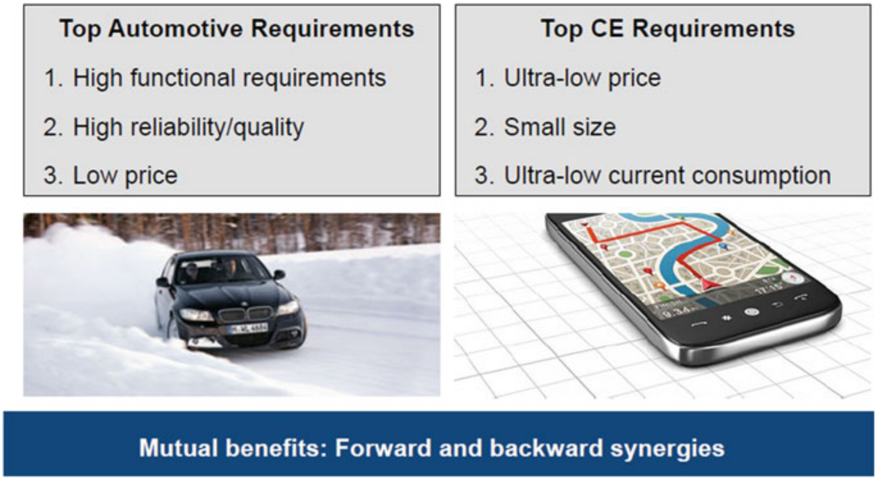
Fig. 15.4 ESP (electronic safety package) gyroscope sensor generations [3]

strong miniaturization potential with respect to volume, size and weight, combined with significant advantages regarding energy consumption, reliability, and cost.

The sequence of generations, since 1995, of the ESP sensors (electronic safety package) is shown in Fig. 15.4. Besides the annual volume of >30 Millions units/year for automotive, this integration- and miniaturization-level initiated the mobile CE market to upgrade its products with intelligent MEMS functions.

15.3 Mobile Consumer Electronics

Automotive MEMS sensors had to advance along a sometimes painful 20-years learning curve, until they could meet the two top priority requirements of consumer electronics: ultra-low prices and extremely small sizes. Priorities for both markets are ranked in Fig. 15.5. These ranking lists explain why the sensor generation of



Top Automotive Requirements

1. High functional requirements
2. High reliability/quality
3. Low price

Top CE Requirements

1. Ultra-low price
2. Small size
3. Ultra-low current consumption

Mutual benefits: Forward and backward synergies

Fig. 15.5 Automotive and CE requirements for MEMS Sensors

2010 from Fig. 15.4 could pass the entry-hurdle for smart MEMS sensors in CE applications. At the same time, the mutual synergies between automotive and CE began to accelerate the progress of MEMS-integrated systems.

15.4 The "Bosch" Process

Another fundamental element of the progress of miniaturized integrated MEMS sensors is the realization of solid-state micro-features with high aspect ratios, as shown in Fig. 15.6. The so-called Bosch DRIE plasma process (DRIE = Deep Reactive Ion Etching) is based on alternating cycles of anisotropic reactive ion-etching (SF_6) and sidewall passivation (C_4F_8) steps, respectively. It enables high etching-rates and aspect-ratios far in excess of 10:1 [4]. This process has become for MEMS, what the planar Si process of 1957 became for integrated circuits. The DRIE process has been refined and exploited as a sustained innovation, for example also with respect to vertical 3D integration (Sect. 1.4 and Chap. 3), so that a very broad development base assures its further progress. As microelectronics triggered off MEMS in 1980, nanoelectronics continues to fertilize MEMS. Beyond this on-going phenomenon, the sophistication of MEMS on one side, and the push forward towards "heterogeneous" 3D integration or "More-than-Moore"approaches (Chap. 3) on the other side, have rallied significant development forces towards intelligent electronic systems which are strong in sensors.

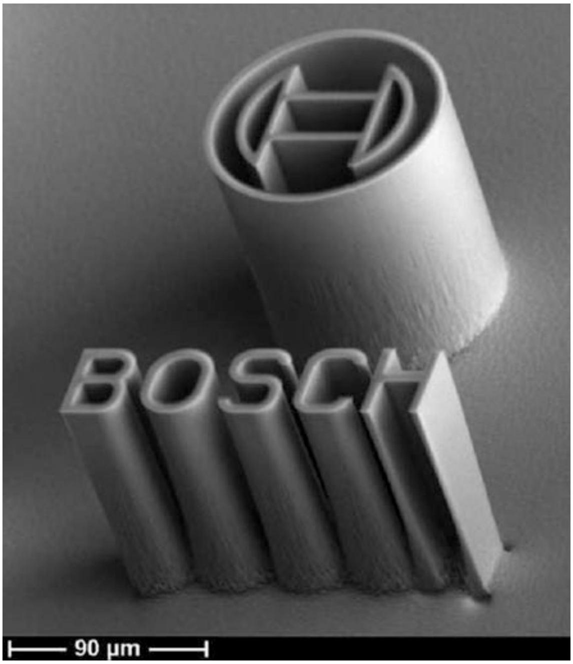


Fig. 15.6 The Bosch process DRIE [4]

15.5 Sensors and Systems-Integration

The feasibility of strong-in-sensors CE mobile devices has led to broad efforts in the device design.

As shown in Fig. 15.7, besides pressure, acceleration and rotation, the MEMS microphone and magnetic sensors can be integrated, challenging heterogeneous technologies and packaging.

The new products immediately began their history of down-scaling in size, cost and power consumption. The evolution of a 3-axis acceleration sensor is shown in Fig. 15.8. Its footprint has been cut in about half every two years over four generations. Some applications are illustrated in Fig. 15.9. Down-sizing reduces power consumption as well, so that, together with new applications in vibration monitoring and wearable intelligence, it emerges as an ideal candidate for energy harvesting and autonomous operation, without a need for battery power any longer (Chap. 19).

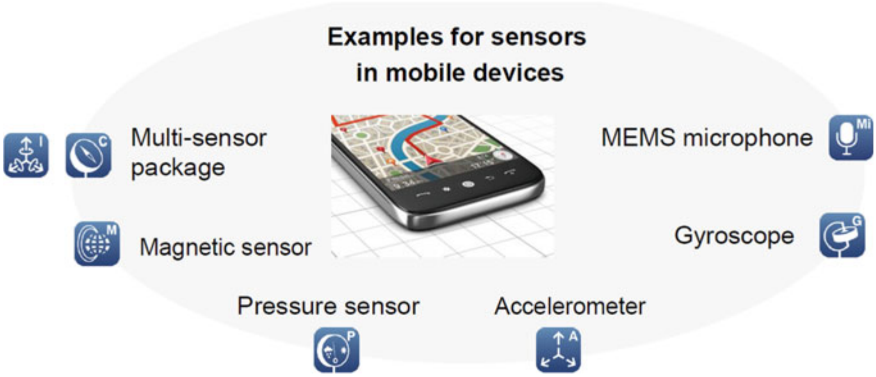


Fig. 15.7 CE MEMS sensors for mobile devices

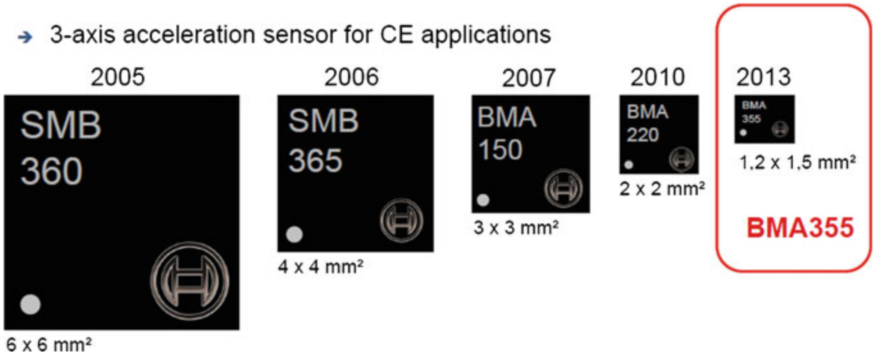


Fig. 15.8 CE MEMS sensors push size reduction. The relative package sizes are to scale [3]

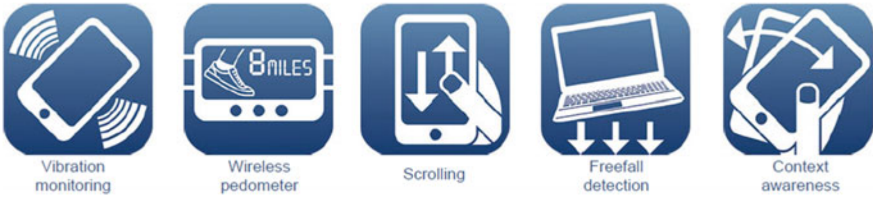
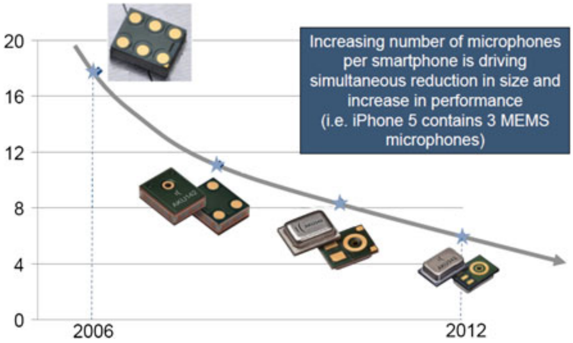


Fig. 15.9 The BMA 355 with a size of 1.2 × 1.5 mm^2 is a 3-axis accelerometer

Fig. 15.10 Size (mm^2) of MEMS microphones [3]



MEMS microphones have a similar history of down-scaling, as illustrated in Fig. 15.10. One consequence is that the number of microphones per product has been increasing from 1 in the beginning to 3 in the iPhone5. This multiplication-effect follows the specific logics of micro- and nano-chips.

15.6 MEMS-Enabled Systems and Their Consistent Development

The rate of progress in MEMS for mobile CE enables the realization of new products in many fields of mobility and communication, from the Human-Machine Interface (HMI) to the Internet-of-Things (IoT). Exemplary functions in the mobile HMI are shown in Fig. 15.11.

The efficient, application-specific design and its sustained introduction of new product generations require a network of qualified sources and tools for hardware/software co-design.

This strategy is illustrated in Fig. 15.12 for the development of sensors ranging over 3-axis (i.e. degrees-of-freedom/DoF) to 9-axis. It involves four technology generations combined with three software generations and use of four SW library generations. The result is a sensor for absolute orientation in five typical applications, as illustrated in Fig. 15.13. It has a high accuracy of 2-to-3° (static), a latency

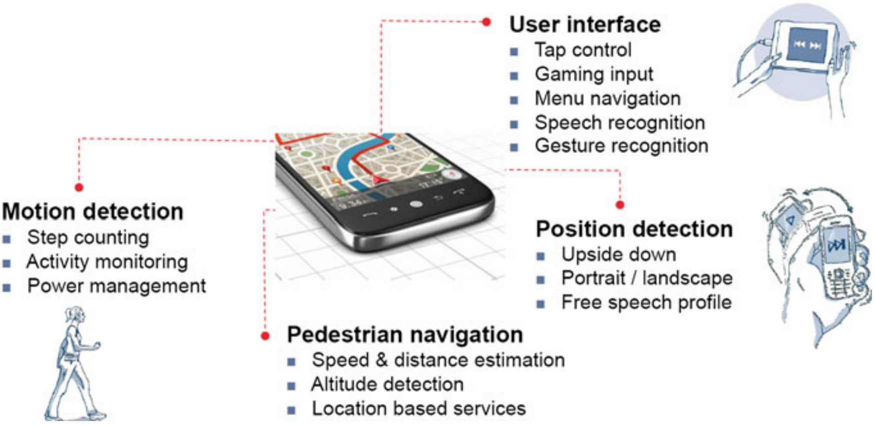


Fig. 15.11 MEMS enable intuitive and realistic implementations of human-machine interfaces

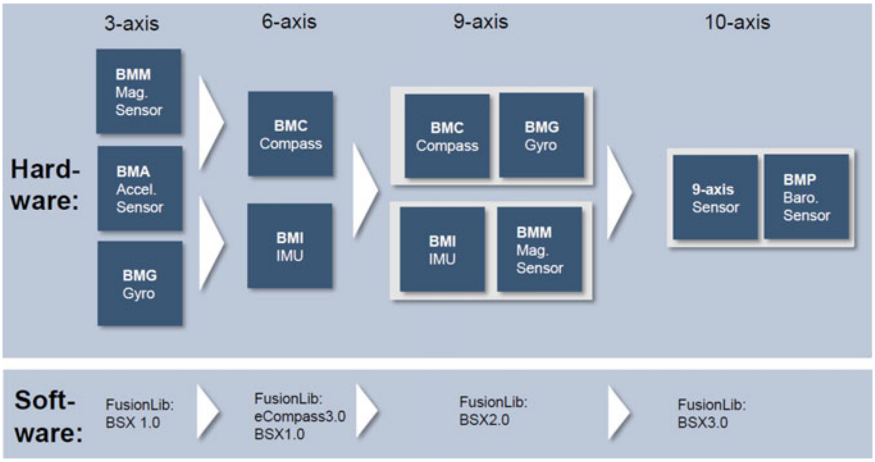
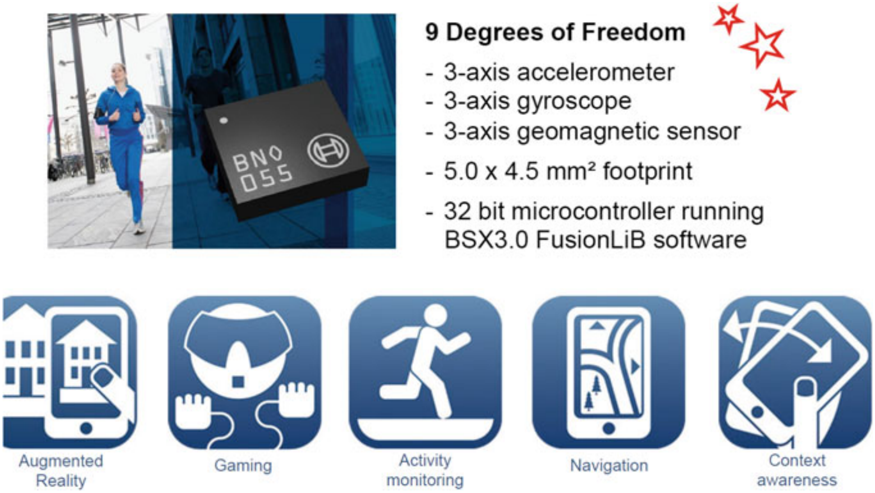


Fig. 15.12 Hardware–software systems integration [3]

of only 20 ms, a short calibration time of 2–3 s, and it is highly resistant to magnetic distortions.

Facilitated by its integrated 32-bit microcontroller running the algorithms for sensor calibration, sensor-data fusion and communication, the device autonomously provides aggregated orientation and motion data (i.e. linear acceleration, rotation, gravity vector, heading, quaternions) for advanced applications in wearable devices and robotics applications.



9 Degrees of Freedom

- 3-axis accelerometer
- 3-axis gyroscope
- 3-axis geomagnetic sensor
- 5.0 x 4.5 mm² footprint
- 32 bit microcontroller running BSX3.0 FusionLiB software

Fig. 15.13 Absolute-orientation sensor, 5.2 × 3.8 mm^2. Integrated system with 9 degrees-of-freedom [3]

15.7 Conclusion

System-level integrated sensors have seen exceptional market growth through consumer electronics applications. With each new generation, their size, performance, cost and energy consumption have improved significantly so that they will swarm into the Internet-of-Everything.

Disclaimer With respect to any examples or hints given herein, any typical values stated herein and/or any information regarding the application of the device, Bosch Sensortec hereby disclaims any and all warranties and liabilities of any kind, including without limitation warranties of non-infringement of intellectual property rights or copyrights of any third party. The information given in this document shall in no event be regarded as a guarantee of conditions or characteristics. They are provided for illustrative purposes only and no evaluation regarding infringement of intellectual property rights or copyrights or regarding functionality, performance or error has been made.

References

1. Marek, J., Gómez, U. M.: MEMS (Micro-Electro-Mechanical Systems) for automotive and consumer electronics, Chap. 14. In Hoefflinger, B. (ed), CHIPS 2020—A Guide to the Future of Nanoelectronics. Springer, Berlin (2012). doi:10.1007/978-3-642-23096-7_14
2. IHS iSupply MEMS Market Tracker—Q3 (2012)
3. Marek, J.: Emergence of sensor swarms for the Internet-of-Things, 23 Oct, 2013
4. Laermer, F.: BOSCH-DRIE shaping MEMS, Invited Keynote to Hiltonhead Workshop, June 2010

Chapter 16
Networked Neural Systems

L. Spaanenburg and W.J. Jansen

Abstract As predicted, intelligent networks with intelligent, wide multi-stage parallel processing have become a favorite in new-media applications like object-recognition and augmented reality with most recent energy efficiencies >1 Tera operations per Watt, 100-times better than conventional von-Neumann processors. Some leading companies in video- and graphics processors and in computer-aided design (CAD) tools are now introducing such products, and the Human-Brain projects enforce this long-term strategy. We see a further outgrowth of neural chips into networks. Firstly, they will appear in health monitoring for healthy living. Non-invasive measurements require time- and space-dependent models that are real-time derived from data. The size of complex neural systems becomes feasible due to further miniaturization by physical innovations like the memristor as well as by novel digital architectures. Together, that will provide effects of self-healing, a level of dependability required for large-scale distributed intelligent systems.

16.1 Introduction

Neural and vision systems have in common that much computing power is required for even the most basic functions. With the increasing abundance of microelectronic computing power, such systems become more and more affordable and gradually move into dedicated devices supporting complex products. The question is usually asked: what kind of new mass markets will open to pay for the required investment?

L. Spaanenburg (✉)
Department of Electrical and Information Technology, Lund University,
P.O. Box 118, 22100 Lund, Sweden
e-mail: lambert.spaanenburg@comoray.com; lambert.spaanenburg@eit.lth.se

W.J. Jansen
Raceplan, Grote Hout 3, 9408 DK Assen, The Netherlands

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_16

This mutual support is posed in [1] to explain historical advances in physics and mathematics.

The vision sensor was added as a feature to the mobile telephone. Soon the feature phone made it to the cheap competitor of the digital camera. Coming of age, the camera phone made the cost of imaging so low that it opens the door for new markets. Over the past years, vision-based Apps and gadgets receive a growing interest at Trade Shows, such as the yearly Mobile World Congress in Barcelona.

Health and Care is a typical upcoming vision market. In the next section, we discuss how health parameters can be extracted through a camera phone. The advance to dedicated health devices is shaped as a wearable. This solution brings a next problem: how to synchronize health devices? As discussed in the next sections, this requires further advances in neural chips. Throughout history, the main dimensions of a microelectronic design cover speed and density. Those dimensions are coupled through parameters like energy consumption and fault sensitivity, morphing the design quality for specific usage. We will discuss how neural networks are made smaller and more reliable.

16.2 Health Monitor

The costs of hospitalisation are going sky-high, while the accessibility of care remains limited. This has caused an increasing interest of hospitals to increase their service to outside the walls of their building. Meanwhile, people are urging to be better informed on their personal status. They want to know about their nutrition, their exercise; basically they want questions answered.

The health meter presented here is based on the capture of visual light reflected on capillary blood covered by skin. Capillary blood can be reached by visual light at various locations on the human body. Literature lists foot, fingertip, hand palm, pulse, tongue, ear lobe and cheek, but there is probably more. A vision sensor can capture the reflected light, and the intensity shows a periodic signal in tune with the blood flow. This presence of the heartbeat reflects the presence of time in the signal, the Photo Plethysmo Graphical or PPG signal (Fig. 16.1).

Hard to do manually over a longer period, a number of opto-electronic devices have been attempted. The major breakthrough is the observation by Aoyagi in 1974 that, by the pulsatile nature of blood flow through the arteries, the contribution of arterial absorbance can be distinguished from tissue absorbance components. From there, pulse oximetry, rapidly spread as a simple, bedside technique to monitor heart rate and arterial oxygen saturation.

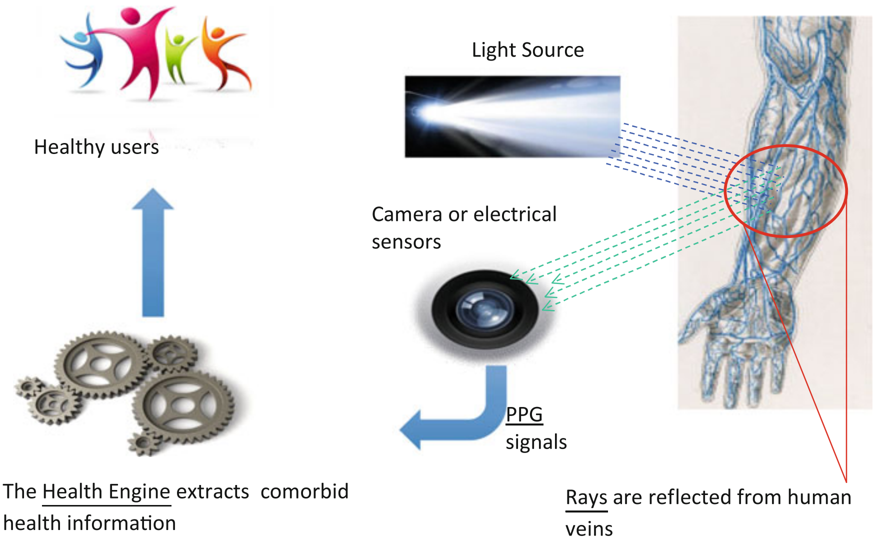

Healthy users
Light Source
Camera or electrical sensors
PPG signals
The Health Engine extracts comorbid health information
Rays are reflected from human veins

Fig. 16.1 Principle of operation (From S. Malki, Comoray Company Presentation)

16.2.1 *PPG (Photo-Plethysmo-Graphical) Analysis*

The PPG technique was originally published by Hertzman in 1938 and became popular in 2000 for perfusion measurements to aid anesthetic monitoring. A good breakdown of the time in-between is given in [2]. The latest developments have confirmed that PPG brings accurate non-invasive acquisition of major physiological parameters. Reflective vision is not the only way to generate a PPG signal, but it is the most popular one. All different sensors have their own noise sources and therefore need their own handlers. A typical example is the use of a standard smartphone platform. The mobile platform has gradually evolved to an easy-to-use camera for family pictures. It tries to take all the diaphragming and lighting away from the user. Unfortunately, this automated process to sharpen the edges and to balance the colors is counter-productive for our effort to get the raw pixels (Fig. 16.2). For this chapter, we simply start from what the various set-ups have in common: the PPG signal.

There are two ways to extract information from the PPG signal: in (a) the time or (b) the frequency domain. The former analyses the signal and its shape over time; the latter looks at the energy/frequency relation after a Fourier transform. Our approach uses the time domain and therefore needs to certify the heart-beat that mechanizes the behavior over time in the presence of abnormalities [3]. In [4], an experiment shows the relative merits of three different algorithms to perform peak detection in such a bio-signal, and it finds that the neural approach gives the best performance allowing as much as 60 % noise.

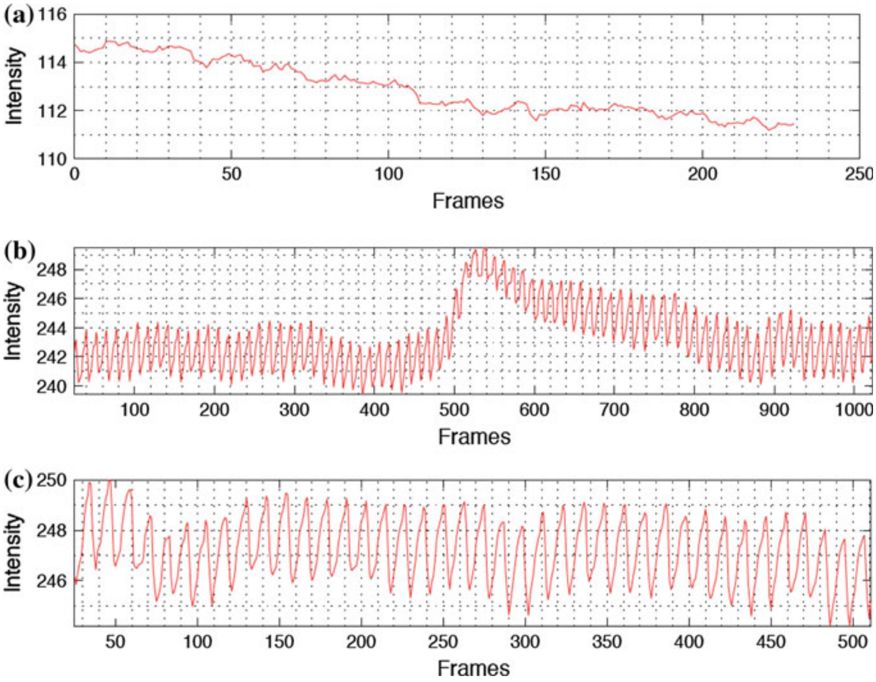


Fig. 16.2 Extracted signals with (**a**) automated color balancing, (**b**) physical motion and (**c**) normal. (Fingertip measurements on i-Phone5)

The length of a signal (or heart) cycle is not a constant. If the heart is exercised at the edge of the performance potential, like during sport exercises or periods of stress, the cycle length can vary. This is called Heart Rate Variability (HRV) [5]. HRV is usually measured from the R-to-R distance, the high peaks in the heart wave. The measurement is complicated by physiological and physical abnormalities. It is important to determine, which of the peaks are caused by the heart, because all other analysis is based on that.

16.2.2 The Need for Modelling

The shape of a PPG signal varies with the measurement location on the body. The further away from the heart, the more the signal has travelled, and therefore the shape will be smoothed. The original heart wave shows a number of clearly discernable peaks, each of which giving meaning to an overall heart diagnosis. The further it travels, the less such peaks are premonitory.

The path the light has to travel between the reflection and the sensor gives another reason for a less clear presence of health information in the PPG signal.

There is air and tissue between the blood and the sensor, and especially the watery tissue will ameliorate the signal strength in dependence of the light frequency (Fig. 16.3). The thicker the tissue, the harder it is to extract a useful signal. With obese persons, it is notably hard to find a pulse and to locate a vein.

A channel model is required to compensate for the above-mentioned negative transmission effects. This is especially important where health analysis gives importance to the details of the signal's shape when inducing the blood pressure value. As the model reflects characteristics of a person's body and/or behavior, it can be trained for personal characteristics. As the model is different for different persons, it gives both health as well as identification data.

Once the main features (the peaks) of the PPG signal are identified and the shape is restored, more health information can be found in further analysis. Oxygenation requires the PPG signal to be extracted in a uniform way through at least two different color channels. Though the green channel is the best to measure with a comfortable dynamic range, the blue and red are needed to discriminate between haemoglobin with and without oxygen. Ideally, the time series in the blue and red channels have to be comparable to come to an accurate oxygen estimation.

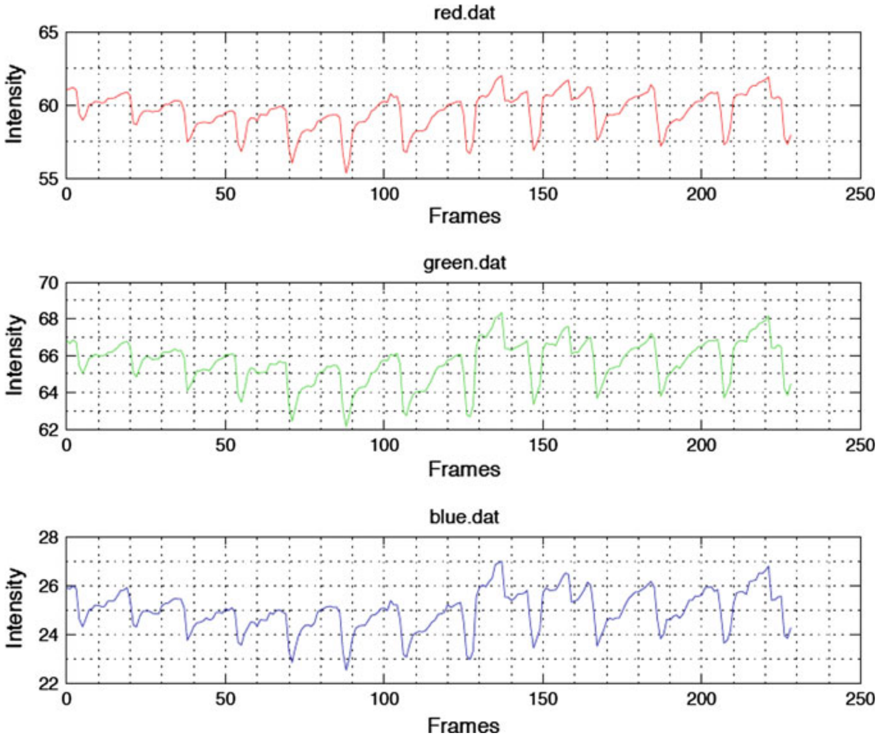


Fig. 16.3 The PPG signal in three color channels (courtesy Tizin) (Color figure online)

The age-old detection of decreased blood oxygenation is by the skin color shift towards blue (also named cyanosis). Together with a weakening pulse pressure, this was known (especially to mountain climbers) as a life threatening condition that requires immediate attention. Blood pressure requires only a single color channel, but the exactness of the shape is very important. A number of experiments have been done to correlate medical blood-pressure measurements and PPG signal features. Such features are shape dependent and therefore will not work without both a proper channel- and a related location-model.

16.3 Integrated Neural Systems

On smaller datasets, vision offers interesting applications because the basic neural equation fits nicely with low-level pixel operations. Analog realizations allow focal-plane processing: a tight integration of light sensors with on-chip low-level operations to reduce the off-chip bandwidth. Parallelism between many small nodes is a given, but programmability is still needed. This led first to analog and then to digital realizations.

With scrutiny of number representations and on-chip network communication, the digital realization has slowly evolved from simple pipelined Central Processing Units (CPU) to tiled Application-Specific Integrated Processors (ASIP). The development can clearly be seen in the history of Cellular Neural Networks, where the intense communication between locally interconnected nodes poses a major hurdle.

The current digital image sensors, such as Xtal, are still based on a co-processor architecture. Often the pixel operations are based on a streaming processor that constantly accesses the memory. Being limited by the memory bandwidth, they do not address the basic need for algorithm acceleration. Raising the stream rate to a higher level brings a next class of ASIPs that takes the special needs for more advanced single-chip applications into consideration. For instance, in biometrics, it is required to align images from different spectral domains for better feature extraction, while for bio-inspired audio processing an elaborate collaboration is needed between multiple timing domains. Such chips will bring human-like understanding into small, low-power smart vision sensors that together will realize Moravec's dream [6].

16.3.1 Synaptic Chips

The early-nineties integrated neural networks were predominantly analog. Digital versions were bound to be small as the required multipliers were simply taking too much area [7]. However, the inherent sensitivity to parameter spread poses a barrier to the potential network size. This has stopped the development for a long time. Larger networks could only be created in software.

Very-large neural networks are of interest for brain modeling. The underlying interest has been real-time bio-mimical pattern matching, trying to understand 'by doing' the working of the human brain. Such sizes cannot even be real-time on supercomputers, and this returned the interest in hardware acceleration.

Early researchers like van der Malsburg have already proposed a bio-plausible circuit for neural behavior [8]. This spike mechanism has been extensively researched, largely in Germany, and Heinrich Klar has produced a clear view of the potential complexity growth [9].

Spurred by large-scale scientific sponsoring by the European community and by DARPA, a number of experiments have been performed [10]. Currently, much attention is given to the IBM TrueNorth chip, offering 1 Mio. programmable neurons and 256 Mio. programmable synapses, based on 4096 neuro-synaptic cores. Each chip is separately usable, containing full support for memory, computation and communication. The inherently massive parallelism helps to bypass the bottleneck in traditional von-Neumann computing that limited the past architectures.

16.3.2 Memristor

When coming to Berkeley in 1971, Leon Chua prepared a course in circuit theory. He noted that an element was missing in the basic element's list, thus far comprising of resistor, capacitor and inductor (Table 16.1). He called this non-linear passive two-terminal electrical component, relating electric charge and magnetic flux linkage, the memristor [11]. Since then, he has trained classes in modeling non-linear systems through analogons using all 4 elements, though one of them was felt to be of theoretical value only.

The memristor has a non-constant electrical resistance. The state depends on the history of current that had previously flown through the device—the so-called non-volatility property [12]. In the powerless situation, the memristor remembers its most recent resistance until it is turned on again [13]. This effect has been noticed before, but generally discarded as an incomplete measurement analysis. For this

Table 16.1 The four circuit elements

Device	Characteristic property (units)	Differential equations
Resistor (R)	Resistance (V/A or Ohm)	$R = dV/dI$
Capacitor (C)	Capacitance (C/V or Farad)	$C = dq/dV$
Inductor (L)	Inductance (Wb/A or Henry)	$L = d\Phi_m/dI$
Memristor (M)	Memristance (Wb/C or Ohm)	$M = d\Phi_m/dq$

reason, Chua has also argued that the memristor is the oldest known circuit element, with its effects predating the resistor, capacitor and inductor.

In 2008, a team at HP Labs made the intellectual link between the non-linear effects observed in the analysis of a thin film of titanium dioxide and the memristor concept; the HP result was published in Nature [13]. Since then, such devices have been intentionally developed for applications in nanoelectronic memories, computer logic and neuromorphic/neuromemristive computer architectures. In March 2012, a team of researchers from HRL Laboratories and the University of Michigan announced the first functioning memristor array built on a CMOS chip. DARPA's SyNAPSE project has funded IBM Research and HP Labs, in collaboration with the Boston University Department of Cognitive and Neural Systems (CNS), to develop neuromorphic architectures, which may be based on memristive systems. An excellent review of the state-of-the-art in memristor-based neural-network design can be found in [14].

16.4 Distributed Neural Networks

The stress meter is another example of how a neural network is needed to split two causes for the same effect. Heart-rate variability and blood pressure come together in signaling physical exertion. However, further information is needed to distinguish between a sporting and/or a stressful life. Fitness can be seen in some personal diary, while psychological signs can also meet stress indications. Therefore, we see the stress meter not as a product by its own means but as the result of merging a health meter into a non-health App.

This is not limited to stress. Nutrition advice, as well, can be supported where bad food and life-style habits are reflected in general health parameters. New areas of health integration, where neural networks accommodate the merging, are in gaming and in air conditioning. Serious gaming is a name for simulation games where decision-makers can be trained in solving real-life management problems. Health is a typical parameter to be considered. Similarly, health sensing can be added to the classical temperature sensing of a central heating system.

16.4.1 Event-Directed Synchronization

A typical example of a distributed neural network can be found in e-fashion. When all the sensors are stuck on a single plaster, the network will still be isochronic. However, sensors can appear all over the body, connected with nano-tube yarns in the textile. Though it is still possible to have the required definition of timing, the realization takes too much power in a battery-operated system.

The problem is not new to technology. With the increase of systems-on-a-chip, it was noted that signal drivers could not maintain signal integrity over large distances. The solution has been to provide asynchronous support for the global communication, while still leaving the local, short-distance control synchronous. This has become popular by the name "GALS". The Globally Asynchronous, Locally Synchronous (GALS) technology is not applicable to e-fashion, as it requires too much wiring.

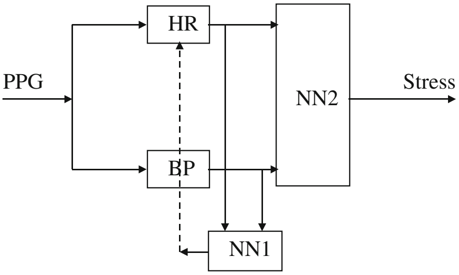
In e-fashion, a number of health sensors are combined. Such readings depend on the placement (for instance blood pressure in the left arm is different from the right arm) and, most important, the time. Typically, parameters change rapidly within bands. Therefore, a coarse reading will be averaged over a time period (say several minutes or even several months) or the placement is restricted (say pulse blood pressure). The latter is caused by signal deformation through path propagation. This makes it hard to find equal points in time. The solution is Context Autonomy, Locally Synchronous (CALS).

An early example of CALS technology appears in the handling of PPG signals. It results from the reflection of visual light within the skin on capillary blood. Therefore, it varies in time in syncopation with the blood flow. Unfortunately, the heart beats are not easily recognizable in the PPG signal. A small error in the discrimination leads to a large error in the determination of the blood pressure value. Neural technology helps to match BP with HR even for low data sampling rates (Fig. 16.4).

16.4.2 Self-healing

Over the 20th century, the Electric Grid expanded unplanned and unattached. The main concern was to establish enough capacity. In the early nineties, the world was shocked by the coming of massive breakdowns with seemingly innocuous causes. A loose wire started to dangle in the wind causing oscillations that blacked-out the Western hemisphere, before the problem was detected centrally. In the years after, the gradually decreasing amount of small disturbances became accompanied by an increasing amount of large accidents.

Fig. 16.4 Stress sensing by neural timing feedback (NN1) and matching (NN2)



This sensitivity of the Electric Grid is primarily caused by the lack of fault containment. Contingent areas are quickly infected. The global structure of such networks with inherent communication delays has become notorious for abnormal behavior. For instance, default distribution of a programming error as part of the maintenance procedure caused the New-Jersey blackout in 1992. Only a couple of years later, the Allston-Keeler (July 1996) and the Galaxy-IV (May 1998) disasters gave rise to a concerted research activity on Self-Healing Networks [15]. A series of three disasters on the Electric Grid in the autumn of 2003 (in America, Sweden and Italy, respectively) suggests that little progress has been made. And these are only the tip of the iceberg [16].

Electronic networks (and especially wireless ones) are not exempt from emergent behavior. Even when the design is perfect, ageing and wear can develop in unknown ways. Moreover, embedded systems are becoming more of the reactive nature that makes abnormal behavior likely to emerge. Where autonomous nodes work together, they tend to pass not only the good but also the bad news. Consequently, special measures are required to make them 'immune' for sick neighbors [17]. Of course, the degree of immunity (or self-healing) must be dependent on the fatality of the sickness. The critical point is clearly how to differentiate between the need to globally adapt and the demand to locally block a fault effect from spreading.

In addition, redundancy allows overruling a malfunctioning part to decrease the fault sensitivity. It is usually eliminated to bring product cost down initially. However, an optimal design should exploit redundancy to the fullest, because it reduces maintenance costs over the lifetime of the system [18]. Partial redundancy is most effective, though this will inevitably raise the need to contain emergent behavior.

16.5 Conclusion

Health values are never stable nor repeatable. Therefore, various means of averaging over arbitrary intervals are used. Together with the location and the function of sensors, this means that trends are more reliable than values under circumstances where continuous operation is performed. For the parts of a health system such time series have to be non-linearly correlated, making the overall system a distributed neural network.

The system has neural parts for the local tactical actions as well as for global, collaborative neural decision-making. This makes a neuron-intensive system where the advances of the Human-Brain projects will be used, though at a much smaller scale. Advances like electronic lenses and hyperscalar sensors are entering the field through surveillance systems and are gradually inherited by the mass-market over

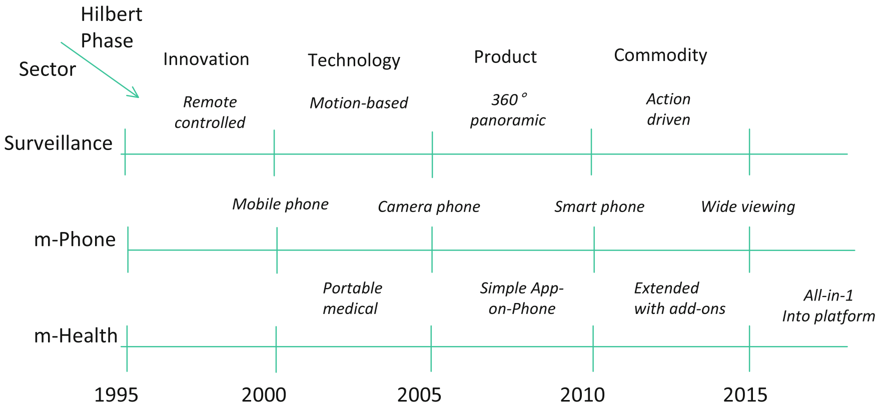


Fig. 16.5 The mobile-health roadmap

mobile telephones. Together with further advances in neural technology, health products will swiftly follow. This sequence in time is illustrated by a roadmap in Fig. 16.5.

References

1. Lederman, L., Teresi, D.: The God Particle. Dell Publishing, New York (2008)
2. Forstner, K.: Pulsoximetrie - Stand und Entwicklung der Technik. Biomed. Tech. **33**(3), 6–9 (1988)
3. Dickson, C.: Heart rate artifact suppression. M.Sc. Thesis, Grand Valley State University, Michigan, USA (2012)
4. Klofutar, A., Höfflinger, B., Neußer, S., Spaanenburg, L.: Robust QRS detection with neuroprocessing. In: Digest 2nd European Conference on Engineering and Medicine, Stuttgart, pp. 64–70 (1993)
5. Jo, J., Lee, Y.K., Shin, H.S.: Real-time analysis of heart rate variability for a mobile human emotion recognition system. In: Recent Advances in Electrical and Computer Engineering, pp. 162–166 (2013)
6. Moravec, H.: Robot–mere machine to transcendent mind. Oxford University Press, Oxford (2000)
7. Nijhuis, J.A.G., Höfflinger, B., Neusser, S., Siggelkow, A., Spaanenburg, L.: A VLSI implementation of a neural car collision avoidance controller. In: Proceedings of IJCNN Seattle WA, vol. 1, pp. 493–499 (1991)
8. Von der Malsburg, C., Schneider, W.: A neural cocktail-party processor. Biol. Cybern. **54**(1), 29–40 (1986)
9. Roth, U., Jahnke, A., Klar, H.: Hardware requirements for spike-processing neural networks. In: Proceedings of IWANN (1995)
10. Hsu, J.: IBM's new brain. IEEE Spectr. **51**(11), 17–19 (2014)
11. Chua, L.O.: Memristor—the missing circuit element. IEEE Trans. Circ. Theor. **18**(5), 507–519 (1971)
12. Chua, L.O.: Resistance switching memories are memristors. Appl. Phys. A **102**(4), 765–783 (2011)

13. Strukov, D.B., Snider, G.S., Stewart, D.R., Williams, S.R.: The missing memristor found. Nature **453**(7191), 80–83 (2008)
14. Thomas, A.: Memristor-based neural networks. J. Phys. D Appl. Phys. **46**(9), 93001 (2013)
15. Amin, M.: Towards self-healing infrastructure systems. IEEE Comput. **8**(8), 44–53 (2000)
16. Amin, M.: North America's electricity infrastructure—Are we ready for more perfect storms? IEEE Secur. Priv. **1**(5), 19–25 (2003)
17. Hofmeyr, S.A., Forrest, S.: Architecture for an artificial immune system. Evol. Comput. **8**(4), 443–473 (2000)
18. Brooks, R.A.: A robust layered control system for a mobile robot. IEEE J. Robot. Autom. **2**(1), 14–23 (1986)

Chapter 17
Insertion of Retinal Implants in Worlwide Prostheses

Bernd Hoefflinger

Abstract Electronic implants for the restoration of vision are particularly challenging. The history of the most advanced one, a subretinal chip with 1600 active pixels, began in 1996 with the invention of a high-dynamic-range recording pixel with adequate stimulation of cells in the retina. Twenty years later, this development has reached world-wide approval in surgical practice and in responses by patients. Its sustained improvement will see similar time-constants.

17.1 HDR Subretinal Implant Inspired by the Human Visual System

Among electronic implants in humans, retinal implants to restore vision for blind people are particularly challenging and close to the human brain. An overview was given in Chap. 17 [1] of CHIPS 2020 of 2012, indicating specifically the on-going clinical trials of the most advanced implant, placed under the retina with 1600 active pixels on a 3 × 3 mm^2 CMOS chip. Its principle is probably the most original electronic implementation of the powerful human visual system (HVS), particularly for two of its most striking powers:

1. Its high-dynamic-range, logarithmic, just-noticeable-difference (JND) sensing (Chaps. 12–14) rods and cones, and
2. Its trained-neural-network compression and pattern-recognition capabilities.

One particular and rapidly proceeding cause of blindness is Retinitis Pigmentosa (RP) where rods and cones fail within time spans of about one year. The top task of an electronic implant is to replace these sensor functions and to provide the proper local activation of the retinal neural network with its natural functionality (2), which had been trained and working until the start of the disease and, generally, survived the disease for some time.

B. Hoefflinger (✉)
5 Leonberger Strasse, 71063 Sindelfingen, Germany
e-mail: bhoefflinger@t-online.de

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_17

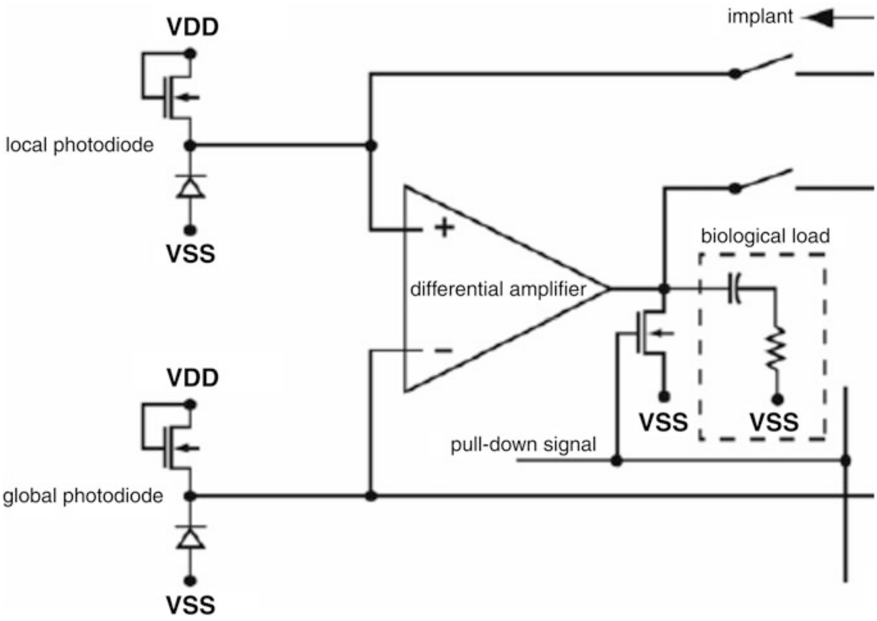


Fig. 17.1 Circuit diagram of one pixel in the subretinal implant alpha-IMS [2]

With the advent of large-scale integration, it had been tried to stimulate the retina with arrays of photodiodes, however, without success, because of a missing opto-electronic conversion function like the one shown in Fig. 12.5 of Chap. 12. It was the incorporation of the log-converting pixel capability in Fig. 12.4, which provided one of the quantum steps [2] in the interdisciplinary research project on the subretinal implant, launched in 1996. A circuit diagram of the eventual pixel is shown in Fig. 17.1. The transistors, local and global, in their subthreshold modes-of-operation, provide input voltages to the differential amplifier, which are the logarithms of the currents from the photodiode. The log response, the consideration of the average luminance and the adjustable amplifier gain are the key elements for a comfortable stimulation of the patient's brain. The alpha-IMS chip with an array of 40 × 40 of these pixels was first manufactured by IMS-CHIPS in Stuttgart in a certified CMOS ASIC process in 1998, and it has been the basis since of all the following processes like Adding the stimulation electrodes, Wiring and packaging, Sealing for the environment inside the eye.

17.2 Chronicle of the Subretinal Implant

This section offers a brief chronicle of the history of a key electronic innovation as an example for the sustained introduction into medical practice:

1996: Beginning of the interdisciplinary research project "Subretinal Implant".
1999: Complete implant admitted for implantation in animal eyes [3].

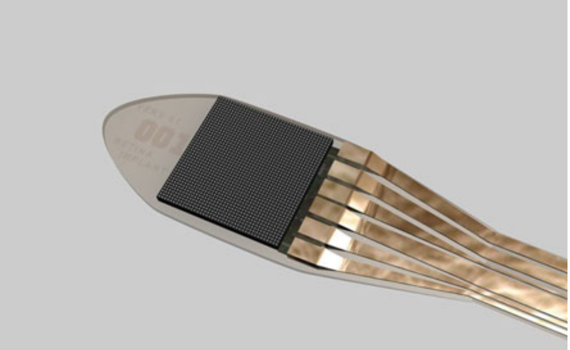


Fig. 17.2 Micrograph of the alpha IMS chip with 1600 pixels on 3 × 3 mm^2 in the assembly with wiring for gain control and electric power

2003: Implant alpha-IMS, Fig. 17.2, qualified for preparation of clinical trials [4].
2004: Foundation of Retina Implant AG, Reutlingen, Germany.
2006: First human implant of alpha-IMS.
2009: Blind patients can read letters [5].
2012: Systematic clinical trial with 8 patients over 1 year [6].
2013: Alpha IMS wins CE certification [7].
2014 October: First report on an international clinical trial [8].

17.3 CE Certification and Results for Blind Patients Worldwide

The tests of the visual outcome for blind patients after the implantation of alpha IMS evolved since 2006 into a standard sequence of perceived patterns, scenarios, and questions. A comprehensive study was performed with 8 patients over a one-year period 2011–2012 [6] (Figs. 17.3 and 17.4).

The following test scores were achieved:

Flat luminance:	8/8
Location of light:	7/8
Motion:	5/8
Grating acuity:	6/8
Landolt C-ring:	2/8.

The best gap recognition of 0.45° corresponded to the pixel pitch on the implant. In the table tests, up to 4-out-of-6 objects were placed on the table and had to be identified as to which, how many and where. With a perfect score of 4, the average for multiple tasks were from 2.5 to 3.8, while the identification of shapes produced scores of 1.5.

The size of the 3 × 3 mm^2 implant covers a visual field of 10° × 10° or 15° diagonal. Letters with sizes of 5°–10° were recognized by the patients. Due to the dynamic range of the implant, patients with the implant were comfortable in any environment indoors and outdoors.

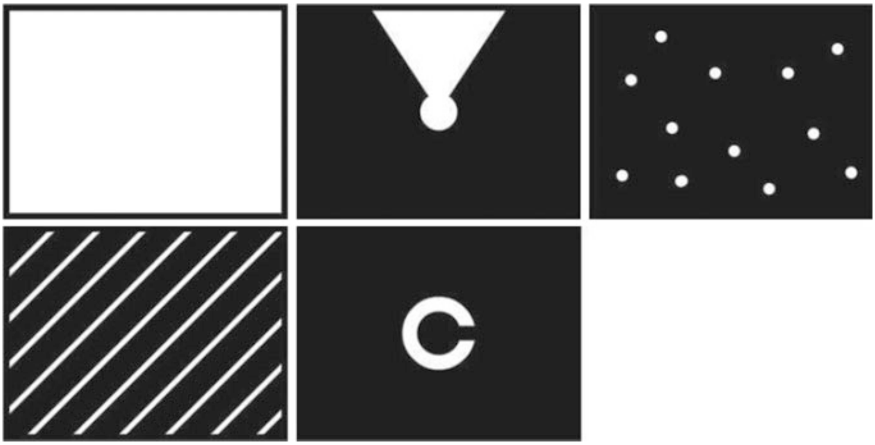


Fig. 17.3 Visual acuity patterns on a 60 cm diagonal screen viewed from a 60 cm distance: Light perception, location, motion with speeds from 3° to 35° per second, grating acuity of 0.1–1.0 line/degree, Landolt ring with gap sizes of 1.6° to 0.45°

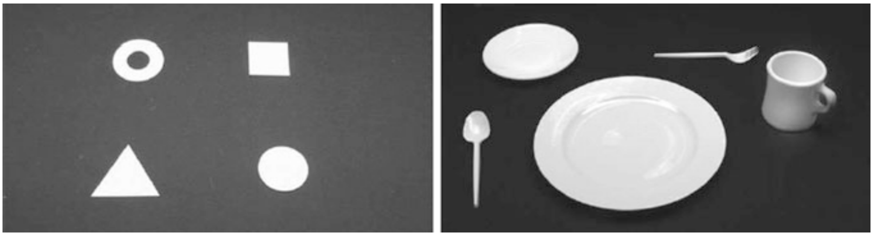


Fig. 17.4 Geometric forms and table settings

> ### Patient reports in the experience of daily life [6]:
>
> In the near-vision range, the most relevant reports included the recognition of facial characteristics, such as mouth shapes (smiles) or the presence/absence of glasses, and differentiation between the contours of people and clothing patterns (striped patterns, black jacket versus white shirt). At home or at work, it was possible to visually localize or distinguish objects, such as telephones, cutlery, parts of the meal (light noodles versus dark beef), red wine versus white wine, and other objects, including door knobs, signs on doors, washbasins or wastebaskets.
>
> In the far-vision range, the most frequently reported perception was finding the line of the horizon and objects along the horizon, such as houses or trees. A river was described as a bright, reflecting stripe. Cars on the street were localized on the basis of bright reflections from their surfaces; the same was true of glass windows in general. One patient reported recognizing stopping and moving cars at night due to their headlights, as well as recognition of the course of the street according to the alignment of the streetlights. Another patient reported seeing the contours of the heads of his colleagues during a work meeting.

By 2013, the records and data on the Alpha IMS including its manufacturing, test, surgical process, as well as the training and experience of patients were accepted by the European Commission on Quality Standards for the CE mark of compliant electronics.

In an international training and cooperation program, seven clinics in five countries worldwide inserted the Alpha IMS chip in blind patients by 2013, and they participated in a joint study, first published in October 2014 [8]:

Germany: Centre for Ophthalmology, University Tuebingen. Staedtisches Klinikum Dresden.
Singapore: Dept. of Ophthalmology, National Univ. Health Systems, Singapore.
United Kingdom: Dept. of Ophthalmology, School of Medicine, King's College, London. Dept. of Clinical Neurosciences, Nuffield Lab of O., Univ. of Oxford.
Hungary, Semmelweis Univ., Budapest.
Hong Kong: Dept. of Ophthalmology, Univ. of Hong Kong.

According to the study, the 12-months visual and safety outcomes on 14 male and 12 female patients, the following results were obtained on the tests described above:

Flat luminance: 22 (85 %),
Localization: 15 (58 %),
Movement detection 6 (23 %),
Grating accuity: 14 (54 %),
Landolt rings: 4 (18 %).

In the visual ability tests regarding objects on a table, the scores on a scale of 0 to 4.0 were

3.12 for detection,
2.94 for localization,
1.06 for identification,

at month 2 after implantation. Regarding their visual experience in daily life, 19 patients (73 %) could localize objects, 12 of them including details.

17.4 Conclusion

Twenty years after the start of the interdisciplinary research project on a sub-retinal implant for blind patients, the restoration of useful vision has been achieved in practice and confirmed by patients worldwide. Future generations of subretinal chips will achieve higher performance, again on respectable time-scales. The electronics of a follow-up chip was described in [1], and we will see its appearance in medical practice in the twenties.

References

1. Rothermel, A.: Retinal implants for blind patients, chapter 17. In: Hoefflinger, B. (ed.) CHIPS 2020—A Guide to the Future of Nanoelectronics, pp. 367–382. Springer, Berlin (2012). doi:10.1007/978-3-642-23096-7_17
2. Graf, H.G., et al.: HDR subretinal implant for the vision impaired, chapter 10. In: Hoefflinger, B. (ed.) High-Dynamic-Range (HDR) Vision, pp. 141–146. Springer, Berlin (2007)
3. Schubert, M., Stelzle, M., Graf, M., Stert, A., Nisch, W., Graf, H., Hammerle, H., Gabel, V., Hofflinger, B., Zrenner, E.: Subretinal implants for the recovery of vision. In: 1999 IEEE International Conference on Systems, Man, and Cybernetics, IEEE SMC '99 Conference Proceedings, Vol. 4, pp. 376–381 (1999)
4. Dollberg, A., Graf, H.G., Hoefflinger, B., Nisch, W., Schulze Spuentrup, J.D., Schumacher, K.: A fully testable retina implant. In: Hamza, M.H. (ed.) Proceedings of the IASTED International Conference on Biomedical Engineering, BioMED 2003, pp. 255–260. ACTA Press, Calgary (2003)
5. Zrenner, E., et al.: Blind retinitis pigmentosa patients can read letters and recognize the direction of fine stripe patterns with subretinal electronic implants. In: 2009 Association for Research in Vision and Ophthalmology (ARVO), Annual Meeting, Fort Lauderdale, FL, D729, p. 4581, May 2009
6. Stingl, K.: Artificial vision with wirelessly powered subretinal electronic implant alpha-IMS. Proc. Royal Soc. B **280**, 20130077 (2013)
7. Retina Implant AG's Alpha IMS Wins CE Mark, http://www.vision-research.eu/index.php?id=868
8. Zrenner, E., et al.: Visual outcome in 26 blind retinitis pigmentosa patients after implantation of subretinal Alpa IMS devices. http://arvo2014.abstractcentral.com/s1agxt/71594D20D

Chapter 18
Brain-Inspired Architectures for Nanoelectronics

Ulrich Rueckert

Abstract Mapping brain-like structures and processes into electronic substrates has recently seen a revival with the availability of deep-submicron CMOS technology. The basic idea is to exploit the massive parallelism of such circuits and to create low-power and fault-tolerant information-processing systems. Aiming at overcoming the big challenges of deep-submicron CMOS technology (power wall, reliability, design complexity), bio-inspiration offers alternative ways to (embedded) artificial intelligence. The challenge is to understand, design, build, and use new architectures for nanoelectronic systems, which unify the best of brain-inspired information processing concepts and of nanotechnology hardware, including both algorithms and architectures. Obviously, the brain could serve as an inspiration at several different levels, when investigating architectures spanning from innovative system-on-chip to biologically neural inspired. This chapter introduces basic properties of biological brains and general approaches to realize them in nanoelectronics. Modern implementations are able to reach the complexity-scale of large functional units of biological brains, and they feature the ability to learn by plasticity mechanisms found in neuroscience. Combined with high-performance programmable logic and elaborate software tools, such systems are currently evolving into user-configurable non-von-Neumann computing systems, which can be used to implement and test novel computational paradigms. Hence, big brain research programs started world-wide. Four projects from the largest programs on brain-like electronic systems in Europe (Human Brain Project) and in the US (SyNAPSE) will be outlined in this chapter.

U. Rueckert (✉)
CITEC, Bielefeld University, Inspiration 1, 33619 Bielefeld, Germany
e-mail: rueckert@cit-ec.uni-bielefeld.de

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_18

18.1 Introduction

Since the beginning of the computer era, the human brain inspires scientists as an alternative approach to artificial intelligence. The computing power of biological neural networks stems to a large extend from a highly parallel, fine-grained, distributed processing and storage of information as well as the capability of learning. Because of their inherent fault tolerance and unrivalled energy efficiency, biological neural networks offer an attractive approach for alternative architectures for ultra-large-scale-integration (ULSI). In nanoelectronics, the growing complexity of ULSI systems turns difficult problems (e.g., power, reliability, testing, connectivity, design complexity, programming…) into great challenges. Models of biological neural networks (so called artificial neural networks, ANNs) aim at system concepts that both exhaust the possibilities of semiconductor technology and solve as far as possible these great challenges. The ULSI neural systems can thus be seen to serve two complementary, but inseparable roles [1]. They help us develop an engineering discipline, by which collective systems can be designed for specific resource-efficient computations. They also lead to hardware components generated from computational neuroscience, components that allow hypotheses concerning neural systems to be tested. Complementary to the traditional (symbolic) artificial intelligence top-down approach for brain-like information processing, ANNs are a bottom-up approach on a biophysical basis of neurons and synapses.

The implementation of ANNs was mainly technology driven in the past. In the 1960s, the transistor replaced the electronic tube, and small discrete electronic components came up on the market. Researchers like Karl Steinbuch (Learning Matrice [2]) in Germany or Bernard Widrow (Adaline [3]) in the United States used these devices in their construction of electronic ANN implementations with a low number of neurons. Computers for simulating ANNs were not widely available at that time; hence building ANNs out of electronic components was a first approach to study functional principles and dynamics of small artificial neuron groups.

Realizing ANNs with discrete electronic devices was tedious, error-prone, and expensive. Furthermore, it was space consuming to scale up the size of the ANN. Therefore, with the increasing availability of computers and especially personal computers (PCs), software simulations of ANNs offer a more comfortable alternative for studying ANNs. The availability of PCs stimulated worldwide the second wave of ANN research in the late seventies. New international conferences came to life (e.g. IJCNN [4], NIPS [5], ICANN [6]). Software simulations offer a high flexibility at reasonable cost but do not exploit the spatio-temporal parallelism that is inherent in biological neural networks. Hence, especially for larger ANNs with hundreds of neurons, the simulation time was quite long in the early days of the computer era. Furthermore, real-time processing in practical applications was not feasible at all.

In the late 1980s, the revolutionary progress of microelectronics had reached feature sizes of one micrometer and became the driving force behind the constant development of new technical products that have markedly improved functionality and higher performance, yet at a lower cost. By this time, Moore′s law gathered momentum, the first independent fabless companies were launched, and the computer-aided design (CAD) automation industry was born. An affordable way to personalized integrated circuit implementations was established even for small design teams from industry as well as from academia. These new and fascinating opportunities motivated intensive efforts to develop ANN chips and neurocomputers for parallel ANN implementation [7]. Systems have been designed based on available circuit techniques, including analog and digital techniques, hybrid approaches, and even optical technology. The European conference on "Microelectronics for Neural Networks" (MicroNeuro [8]) emerged as the only international forum specifically devoted to all aspects of implementing ANNs in hardware. Even hardware products appeared on the market—from both small businesses (e.g. Adaptive Solutions [9]) and large companies (Intel [10], Siemens [11]). All these impressive approaches had a real problem trying to keep up with the effects of Moore's law coming into full swing, as microprocessors, digital signal processors, and field-programmable gate-arrays (FPGAs) all grew faster and faster. The fabless design teams only had access to technologies one generation or two behind the semiconductor companies, who also could afford mass production pricing. As a consequence, ANN hardware had little success at that time.

Since 2005, the performance increase of microprocessors slowed down, and the trend to multi-core and many-core architectures started. Furthermore, GPUs (graphics processing units) became widely available, and the complexity of state-of-the-art FPGAs allowed system-on-chip designs. These off-the-shelf devices offer new perspectives for massively parallel ANN implementation. Thanks to these developments and the above mentioned challenges for nanoelectronics, ANN hardware experiences nowadays a revival manifested in big research programs in the US, in Europe and in Asia. As for massively parallel computer architectures, adequate programming models and software development tools are still missing, ANNs offer an interesting alternative to conventional instruction-set processing because of their inherent massively parallel architecture and learning capability. ANNs will not be programmed, they are trained.

In this chapter, some basic brain facts are summarized, key issues for realizing ANNs are discussed and general approaches for ANN hardware are introduced. Finally, current international large-scale projects of the third generation of ANN hardware will be presented.

18.2 Some Features of the Human Brain

The human brain consists of about 100 billion neurons [12]. The neuron (Fig. 18.1) is the basic working unit of the brain, transmitting information to other nerve cells, muscles, or glia cells. Each neuron is connected via thousands of synapses to other neurons forming an incredibly powerful sensing, information, and action processing system. The structural and functional properties of highly interconnected neurons

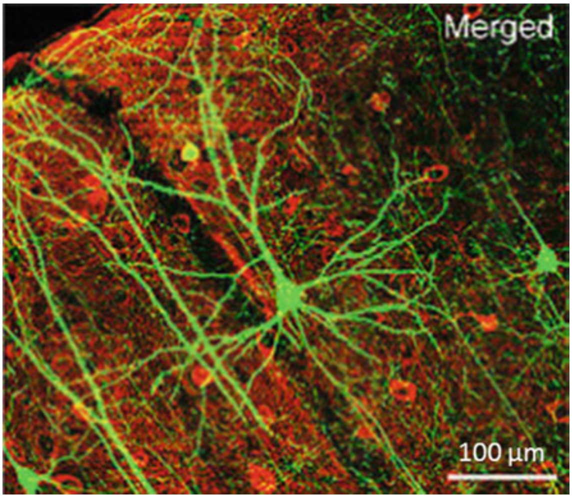


Fig. 18.1 Section of a mouse brain under the microscope (*left*). The *green neuron* in the centre was made visible by a green-fluorescent protein [13]

give rise to the unparalleled intelligence (i.e., learning, comprehension, knowledge, reasoning, planning) and the unique capability of cognition (i.e., the power to reflect upon itself, to have compassion for others, and to develop a universal ethic code) of the human brain.

The neuron consists of a cell body, dendrites, and an axon (Fig. 18.2). The **cell body** receives inputs from its dendrites or directly from axons of other neurons, processes the inputs, and generates the output to other neurons or biological actuators (e.g. muscles). Its function is the firing decision (see the digital solutions in Chap. 16). It has a nonlinear response curve with thresholds, above which an

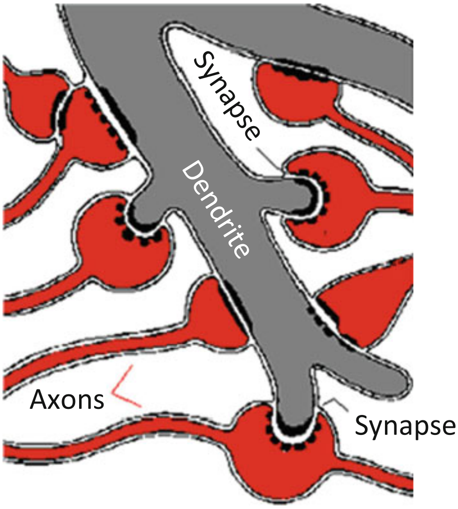


Fig. 18.2 Section of a dendrite with axons and synapses [14]

electrical impulse (spike) is generated (binary output). The key information lies in the action potential, both in its height and in its frequency of occurrence. The action potential is a voltage spike 0–70 mV high and ~2 ms wide.

Neurons signal by transmitting these spikes along their **axons**, which can range in length from millimetre to meter. The electrically excitable axon extends from the cell body as a single output line and branches before ending at nerve terminals, the synapses. The axon carries the output of the neuron through another tree-like structure to couple it to other neurons or physical actuators, incurring a signal-propagation delay that depends on the length of the axon (about hundred miles per hour). The signal-transmission strength is given by the spike density, which resembles a pulse-density-modulation code.

Synapses are the contact points where one neuron communicates with another. They receive as input signal a pulse-density-coded signal with a dynamic range of 0.1–500 Hz with an average of 10 Hz, equivalent to a 10-bit linear code, which could be log-encoded for efficiency and physical meaning. Synapses consist of a presynaptic ending and a postsynaptic ending, separated by a gap of a few nanometers. Neurotransmitters are released across the gap depending on previous history and input signal, for which a simple model is the multiplication of the signal by a characteristic weight. The synapse weight determines to what degree the incoming signal at a synapse is transferred to the postsynaptic output onto a dendrite. The resting potential across the gap of a synapse is 70 mV. When an ionic charge arrives, a specific action potential is generated, reducing the voltage difference. Synapses transmit an output to the dendrite or directly to the cell body, which is weighted by the history of this connection. The coupling process incurs some time delay, but this can generally be added into the axonal delay for modelling purposes. The synapse is the primary location of adaptation in the neural system. The strength of the coupling between two neurons self-adjusts over time in response to factors such as the correlation between the activities of the two neurons that are coupled through the synapse.

Dendrites are the tree-like structures, extend from the neuron cell body and form the input structure of a neuron. They gather the inputs from other neurons or sensory inputs and perform the summation of their associated synapse-outputs. The dendrites and cell body are covered with thousands of synapses formed by the ends of axons from other neurons (Fig. 18.1).

Obviously, this is a very simplified and abstract view on the structure and function of a single neuron. A vast amount of detailed knowledge about the structural and functional properties of single neurons has been gathered during the last decades [15]. For example, neurons are also specialized for certain tasks [16]: e.g. sensor neurons sense inputs (e.g., the retina (Chap. 17)), inter-neurons are the large majority operating inside the brain, motor neurons perform control and actuator functions. The important role of neurotransmitters and neuromodulators on neuron behaviour has been identified [17] and partly analysed. The recent explosion of knowledge about the brain's functions is the result of the advancement of microelectronics and micro-sensing augmented by image processing and micro-physical chemistry. However, there remains a great lack of understanding on the level of micro- and macro-circuits in the brain. On the micro-level, so-called columns have been identified, in which different neuron types are organized as cylinders of dimension 0.5 mm diameter, and about 2 mm in length. Each cylinder (or cortical column) consists of about 10,000 neurons, which are connected in an elaborate, but consistent manner. On the macro level, these columns form certain brain areas with specialized functions (vision, sensory-motor-control, …). Because of the highly interconnected neuronal structure, a clear picture of the organizational structure of the brain is still missing.

Nevertheless, there are interesting 'engineering' aspects of biological neural networks. With structure sizes smaller than 0.1 μ, semiconductor technology starts falling below the level of biological structures forming the brain. However, the brain efficiently uses all three dimensions, whereas nanoelectronics mainly uses the two physical dimensions of the silicon die surface and a restricted number of wiring layers. Nevertheless, taking an area of one square-millimeter—roughly the square dimension of a Purkinje cell (a type of neuron) in the cerebellar cortex, shown in Fig. 18.2 (right), we can use 40-nm CMOS technology to implement a digital artificial neuron (Fig. 18.3, left) with 100,000 32-bit weight synapses and a 32-bit microprocessor as a neural processing unit. Weights are the practical implementation (in hardware, software, theory) of (biological) synapses (contacts between nerve cells).

The scale of the problem of modelling the human brain has been scoped by, among others, Mead [1]. The hundred billion neurons have on the order of 10^{15} connections, each coupling an action potential at a mean rate of not more than a few hertz. This amounts to a total computational rate of about 10^{16} complex operations per second. No computer has yet been built that can deliver this performance in real time, though this gap will be closed in the near future. Current supercomputer developments are aimed at delivering PetaFLOPS (10^{15} floating-point operations

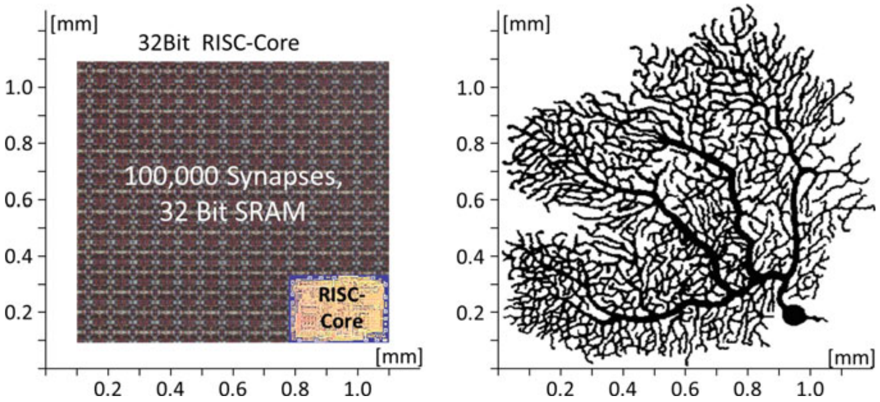


Fig. 18.3 Area comparison of a digital neuron in 40 nm standard CMOS technology (*left*) and a biological neuron (Purkinje cell, *right*)

Table 18.1 Charge and energy model of the human brain [19]

Parameter	Human brain
Number of neurons	10^{11}
Synapses/neuron	10^3–10^5
Ionic charges per neuron firing	6×10^{-11} As
Mean cross section synaptic gap	30 μm^2
Charge/synaptic gap	6×10^{-14} As = 4×10^5 ions
Action potential	70 mV
Energy per synapse operation	4 fJ = 2.5×10^4 eV
Energy per neuron firing	4 pJ
Average frequency of neuron firing	10 Hz
Average brain power	$(2 \times 10^{11}) \times (4 \times 10^{-12}) \times 10 = 8$ W

per second) performance levels, perhaps only one order of magnitude short of the performance required to model the human brain [18].

An even greater challenge is the issue of power efficiency. The power efficiency of neurons (measured as the energy required for a given computation) exceeds that of computer technology, possibly because the neuron itself is a relatively slow component. While computer engineers measure gate speeds in picoseconds, neurons have time constants measured in milliseconds. While computer engineers worry about speed-of-light limitations and the number of clock cycles it takes to get a signal across a chip, neurons communicate at a few meters per second. This very relaxed performance at the technology level is, of course, compensated by the very high levels of parallelism and connectivity of the biological system. Finally, neural systems display levels of fault-tolerance and adaptive learning that artificial systems have yet to approach [18].

Out of the empire of acquired knowledge, some biological data are summarized in Table 18.1 as a performance guide to the human brain, which is interesting to compare with technical data from biomorphic Si brains. The basis of this charge model is a charge density of 2×10^{-7} A s cm^{-2} per neuron firing, fairly common to biological brains [17] and a geometric mean for axon cross sections, which vary from 0.1 to 100 μm^2.

18.3 Brain Simulation Approaches

It seems obvious that the massively parallel computations inherent in artificial neural networks (ANNs) can only be realized efficiently by massively parallel hardware. However, the vast majority of scientists mainly trust on computer simulations, with a clear trend on massively parallel high-performance computers (e.g. IBM Blue Gene [20]), to generate most of their ANN research work. The simulation of ANNs in software on state-of-the-art computers offers a high flexibility and is well suited to test ANN concepts and new algorithms. For networks of limited size, the simulation on standard workstations is practicable, but for many real-world problems the performance that can be achieved with these solutions is not sufficient. Especially, the sequential architecture of today's microprocessors and the memory bandwidth limit the performance that can be achieved for neural network simulation. Due to the parallel structure of neural networks, a considerable speed-up can be achieved by using parallel architectures. Two approaches exist for supporting ANNs in parallel computing architectures: general-purpose neurocomputers for emulating a wide range of neural-network models, and special-purpose ULSI systems dedicated to a specific neural-network model.

General-purpose neurocomputers offer a high degree of observability of the inner workings of neural algorithms as well as their flexibility. Special-purpose ULSI designs offer resource-efficient speed, size, and power consumption. Progress continues in both approaches, and researchers have realized many different architectures in working hardware. There exists a variety of architectures within these two approaches. The performance of neural hardware is commonly reported in connection updates per second (CUPS) during learning of the network and in connections per second (CPS) during recall. Amendments to these measures have been proposed that take into account, e.g., the convergence speed of the algorithm [21] or the precision of the calculations [22]. Furthermore, the resource-efficiency and the robustness of the implementation are of particular interest. Resource-efficiency in this context refers to the appropriate use of the basic physical quantities space, time and energy. However, no one has given the problem of benchmarking a full treatment or found a commonly accepted adequate metric for performance evaluation yet.

Figure 18.4 shows a broad overview of alternative approaches for parallel implementation of ANNs. In spite of the progress in many brain-emulation efforts, a fair comparison of these efforts is very difficult. The main reason is the huge variety of models for the key components of ANNs. Models for the neuron body used to emulate the computation and firing behaviour of biological neurons range from simple threshold gates [23] to more sophisticated complex compartment models [24] requiring spatiotemporal integration. Models for synapses may be as simple as a single bit or represented by floating point values with additional plasticity

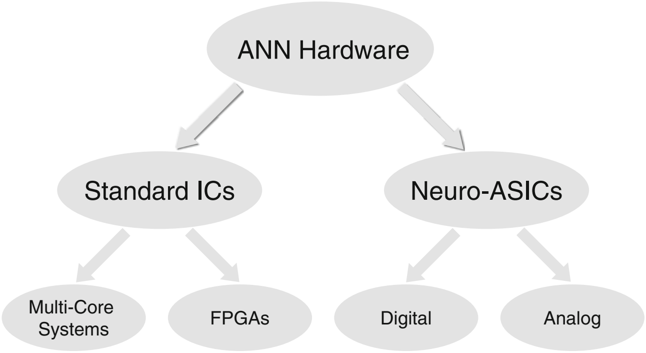


Fig. 18.4 Alternatives for parallel ANN implementations

functionality. Dendrites can be modelled with only one branch or more realistic with many branches with time-varying behaviour. Depending on the level of abstraction, the computational, storage and communication requirements in brain modelling change tremendously.

In the following sections, selected ANN hardware projects will be considered. Compared to the ANN hardware approaches of the second generation in the nineties, most of the projects use the more bio-inspired neuron model of an integrate-and-fire point neuron, summing weighted input from synapses and comparing the resulting sum to a threshold, arriving at a binary decision whether and when to generate an output spike. This model is commonly extended to include a decaying charge, as a “leaky integrate and fire” neuron. The model can also be enhanced in other ways: non-linear summation, time-dependent thresholds, programmable delays in the delivery of spikes, and other variations. The point neuron models are more complex than the abstract neuron models of the second wave of ANNs in the nineties but require only modest computation and hardware, in contrast to biological ion-channel models with spatiotemporal integration [20].

18.4 Neurocomputers Based on Standard ICs

The increasing availability of parallel standard hardware such as Field Programmable Gate Arrays (FPGAs), Graphics Processors (GPUs), and multi-core processors offers new scopes and challenges in respect to resource-efficient implementation and real-time applications of ANNs. Because these devices are inexpensive and available, we can take the first step in implementing neurocomputers with standard devices. ANNs are inherently parallel, and hence it is obvious that many-core processors are an attractive implementation platform for them. The promise of parallelism has fascinated researchers for at least three decades. In the past, parallel computing efforts have shown promise and gathered investment, but in the end, uniprocessor computing always prevailed. Nevertheless, general-purpose computing is taking an irreversible step toward parallel architectures because single-threaded uniprocessor performance is no longer scaling at historic rates. Hence, parallelism is required to increase the performance of demanding applications such as ANN simulation. Since real-world applications are naturally parallel and hardware is naturally parallel, the missing links are programming models and system-software supporting these evolving massively parallel computing architectures. Furthermore, there is no clear consensus about the right balance of computing power, memory capacity, and internal as well as external communication bandwidth of integrated many-core architectures.

Various techniques for simulating large ANNs on parallel supercomputers or computer networks are well known, which can be re-used for mapping ANNs to many-core architectures [25]. Furthermore, many-core processors can be embedded in mobile devices such as robots or smart phones opening up new application vistas for ANNs. Consequently, the number of ANN many-core implementation is increasing [26].

Graphical Processing Units are suited for single-instruction and multiple-data (SIMD) parallel processing. A GPU is a specialized integrated circuit designed to rapidly process floating-point-intensive calculations, related to graphics and rendering at interactive frame rates. The rapid evolution of GPU architectures from a configurable graphics processor to a programmable massively parallel co-processor make them an attractive computing platform for graphics as well as other high performance computing having substantial inherent parallelism such as ANNs. The demand for faster and higher-definition graphics continues to drive the development of increasingly parallel GPUs with more than 1000 processing cores and larger embedded memory at a power consumption of several watts. At the same time, the architecture of GPUs will evolve to further increase the range of other applications. In order to assist the programmers, specialized programming systems for GPUs evolved (e.g., CUDA [27]) enabling the development of highly scalable parallel programs that can run across tens of thousands of concurrent threads and hundreds of processor cores. However, even with these programming systems, the design of efficient parallel algorithms on GPUs for other applications than graphics is not straight-forward. Significant re-structuring of the algorithms is required in order to achieve high performance on GPUs. Furthermore, it is difficult to feed the GPUs fast enough with data to keep them busy. Nevertheless, the increasing number of papers on this topic shows that GPUs are an interesting implementation platform for simulating large ANNs [28].

Field Programmable Gate Arrays have a modular and regular architecture containing mainly programmable logic blocks, embedded memory, and a hierarchy of reconfigurable interconnects for wiring the logic blocks. Furthermore, they may contain digital signal-processing blocks and embedded processor cores. After manufacturing, they can be configured before and during runtime by the customer. Today, system-on-chip designs with a complexity of about a billion logic gates and several Mega-Bytes of internal SRAM memory can be mapped on state-of-the-art FPGAs. Clock rates approach the GHz range boosting the chip-computational power in the order of GOPS (billion operations per second) at a power consumption of several watts. Hence, FPGAs offer an interesting alternative for parallel implementation of ANNs providing a high degree of flexibility and a minimal time to market. The time for the development of FPGA and application specific integrated circuit (ASIC) designs is comparable. A big advantage of FPGAs is that no time for fabrication is needed. A new design can be tested directly after synthesis for which efficient CAD tools are available. A disadvantage of FPGAs is the slower speed, bigger area, and higher power consumption compared to ASICs. Compared to software implementations, FPGAs offer a higher and a more specialized degree of parallelization.

The implementation of ANNs on a reconfigurable hardware makes it possible to realize powerful designs that are optimized for dedicated algorithms [29]. Another great advantage is the feature of reconfigurability that enables the change to a more efficient algorithm whenever possible. If, at the beginning of the training of an ANN, a low data precision is satisfactory, we are able to implement a highly parallel implementation to get a rough order of the network. Using a lower precision allows us to set up an optimized architecture that can be faster, smaller or more

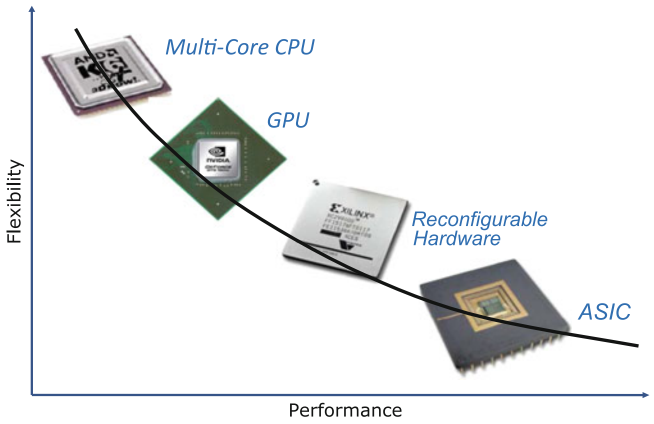


Fig. 18.5 Qualitative performance and flexibility grading of parallel hardware platforms for emulating artificial neural networks

energy-efficient than a high-precision architecture. For a fine-tuning of the ANN, the FPGA can be reconfigured to implement high-precision elements. Additionally, we are able to adapt the implemented algorithms to the network size that is required for a certain problem. Thus, we can always use the most suitable algorithms and architectures. Furthermore, dynamic (or runtime) reconfiguration enables to change the implementation on the FPGA during runtime [30]. Dynamic reconfiguration is used to execute different algorithms on the same resources. Thus, limited hardware resources can be used to implement a wide range of different algorithms. In ANN simulation, we are often interested in providing as much computing power as possible to the simulation of the algorithm. But pre- and post-processing of the input and output data often also requires quite a lot of calculations. In this case, dynamic reconfiguration offers the opportunity to implement special pre-processing algorithms in the beginning, switch to the ANN simulation and in the end reconfigure the system for post-processing. Thus, we do not require the system resources that would be necessary to calculate all algorithms in parallel [31].

In summary, parallel standard hardware like multi-core CPUs, GPUs, or FPGAs are not bio-inspired but cost effective, available, and benefit from market-driven development improvements in the future. They have the highest flexibility (Fig. 18.5) and set the base-line with respect to resource-efficiency and performance for the brain-inspired architectures discussed in the next sections.

18.5 Neurocomputers Based on Neuro-ASICs

The second step in implementing neurocomputers is the realization of the algorithms in silicon (neuro-special-purpose hardware). Each neuron body, synapse and dendrite may take a dedicated piece of hardware either in digital or analog circuit technique. For digital ICs, we can call on efficient software tools for a fast, reliable and even complex design. We can use state-of-the-art process lines to manufacture chips with the highest density in devices.

On the contrary, the design of analog circuits demands much more design-time, good theoretical knowledge about transistor physics, and a heuristic experience of layout design. Only a few process-lines are characterized by analog circuits. In their favor, we point out that, with integrated analog circuits, some neuron functions are quite simple to implement. For example, summation of the dendritic input signals as a current summing is a fairly convenient electronic analog circuit operation and smarter than with common digital accumulators, or a two-quadrant multiplier demands only five transistors. Nonlinearity or parasitic effects of the devices allow us to realize complex functions, as an exponential or a square-root function [1]. Note however, that analog circuits are not as densely integrated as it may seem at first glance. They demand large-area transistors to assure an acceptable precision and to provide good matching of functional transistor pairs, as used in current-mirrors or differential stages.

Various special-purpose hardware implementations of artificial neural networks have been proposed, either dedicated to a wide range of neural networks or optimized for individual algorithms or groups of similar algorithms. Starting in about 1988, there were intensive efforts by many players to develop neural-network chips and boards that would make good on the general concept of sixth-generation hardware [8, 21, 32]. But most of these died out in the following years. As already mentioned, these approaches had a real problem trying to keep up with the effects of Moore’s Law, as CPUs, DSPs and FPGAs all grew faster and faster.

Advances in technology have successively increased our ability to emulate neural networks with speed and accuracy. At the same time, our understanding of neurons in the brain has increased substantially, with imaging and microprobes contributing significantly to our understanding of neural physiology. These advances in both technology and neuroscience stimulated international research projects with the ultimate goal to emulate entire (human) brains. These new approaches are more brain-inspired than the ANN hardware from the nineties. They emulate neural networks on the basis of spiking integrate-and-fire neurons with differences in emphasis. Some approaches aim at a more-detailed and, hence, more computationally-expensive model of neuron behaviour, while others use simpler models of neurons but larger networks. In the following, we will consider projects intended to scale up towards millions of neurons, fabricated and tested with currently available technologies. As a base-line, we summarize the Blue Brain Project using a commercial high-performance supercomputer (IBM Blue Gene) to model neurons and their connections in software. Afterwards, two architectures, based on special-purpose digital integrated circuits emulating neurons in software using many small CPUs networked together, will be presented (SpiNNaker, C2S2). Finally, two neuromorphic systems with special-purpose analog integrated circuits modelling neural circuits on chip (BrainScaleS, Neurogrid) will be considered.

18.6 The Blue Brain Project

The Blue Brain Project [20] at EPFL in Switzerland uses an IBM Blue Gene supercomputer (100 TFLOPS, 10 TB) with currently 8000 CPUs to simulate ANNs (at ion-channel level) in software. The focus of the project is the human cerebral cortex, which takes up 80 % of the brain. The neurons are grouped into functional microcircuits modelling a cortical column. The modelled cortical column is about 0.5 mm in diameter, about 2 mm in height, and consists of about 10,000 neurons. The neurons are not identical and connected in an elaborate but consistent manner. The model does attempt to account for the 3D morphology of the neurons and cortical column, using about 1 billion triangular compartments for the mesh of 10,000 neurons using Hodgkin-Huxley equations [33], resulting in Gigabytes of data for each compartment, and presumably a high level of bio-realism based on floating-point arithmetic. Timing, e.g. propagation delays along the simulated compartments of an axon, are incorporated into the simulation. Synaptic learning algorithms are also introduced, to provide plasticity. A visual representation of parts of the cortical column can be displayed for the simulation, allowing researchers to focus on particular parts or phases of the simulation in more detail.

The Blue Brain project is unusual in its goal to simulate the ion channels and processes of neurons at this fine-grain compartmental level. The time needed to simulate a cortical column is about two orders of magnitude larger than the real biological time. Based on a simpler (point) neuron model, the simulation could have delivered orders of magnitude higher performance. Of course, software emulation of neurons on large computers, including the bio-realistic fine-grain compartmentalized emulation used in Blue Brain, has been used widely in computational neuroscience laboratories [34]. The Blue Brain project is a good example of this approach, because of its combination of large scale and bio-realism. Work on the Blue Brain project is now progressing to a second phase of work merged with the European Human Brain Project [35].

18.7 The SpiNNaker System

The SpiNNaker (Spiking Neural Network Architecture) project at Manchester University [36] aims at a massively parallel multi-core computing system. The basic computing node has one SpiNNaker multi-core chip with 18 low-power ARM 968 processor cores (200 MHz), each with 96 KB of tightly-coupled local on-chip-memory for instructions and data, and a 128 MB SDRAM chip used to store synaptic weights and other information shared by all 18 cores (Fig. 18.6). The SpiNNaker chip was fabricated in a 130 nm CMOS technology. Both chips are integrated as a System-in-Package (SiP) with the SDRAM wire-bonded on top of the SpiNNaker chip (3D packaging). 48 of these nodes are mounted on a PCB, which can be scaled up to 1200 boards for a full SpiNNaker system with more than 1 million of

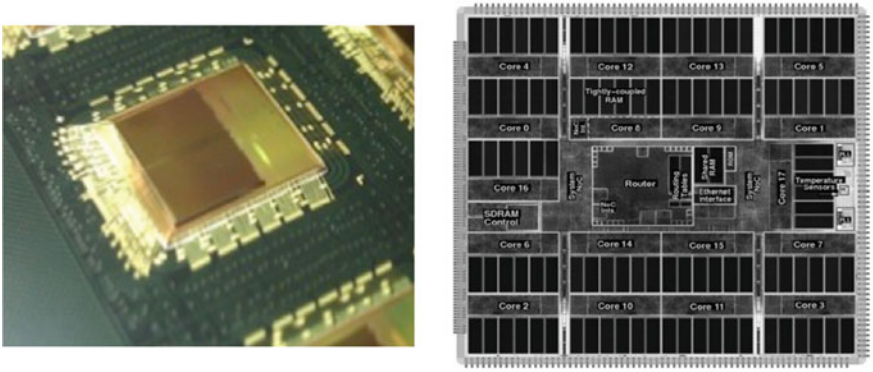


Fig. 18.6 SpiNNaker SiP with SDRAM wire-bonded on the SpiNNaker chip (*left*) and the SpiNNaker die (10.386 × 9.786 = 102 mm^2) with 18 ARM cores [36]. ©IEEE 2013

ARM9 cores and 7.2 TB of distributed RAM. It is expected, that the full system in operation will consume at most 90 kW of electrical power (1 W per node) [36].

The goal is to simulate artificial neural networks with up to a billion neurons in biological real time. Hence, each ARM9 core is designed to simulate about 1000 neurons, communicating spike events to other cores through packets via an on-chip network connecting the 18 cores as well as via an off-chip network connecting SpiNNaker chips. The routing of packets in SpiNNaker is carefully designed to balance complexity and bandwidth. AER (address event representation [37]) packets are used (5 or 9 byte) for asynchronous spike transmission. Although the individual cores execute synchronously, the packets they receive and transmit use an asynchronous transmission protocol (globally asynchronous, locally synchronous mode). The SpiNNaker chips are connected to adjacent SpiNNaker chips in a 2-dimensional mesh network; each chip has 6 network ports, connected to adjacent chips. The speed, at which packets are transmitted over the network, is about 0.2 µs per node hop. The router does not need to know the eventual destination(s) of a packet; it only needs to know, which port(s) to send it to. Routing tables currently remain static during program execution and are built at load-time by the host computer. For structural plasiticty—where a new synaptic connection is established—a separate (progammable) background process will be needed. This has yet to be implemented. Because spike delivery time in SpiNNaker is designed to be faster than in a biological brain (assuming the network and routing delays are adequately controlled), SpiNNaker software directly allows a delay of up to 15 ms to be inserted in delivery of AER packets, in order to simulate longer transmission times. Use of so-called “delay neurons” permits longer delays than 15 ms to be modelled.

SpiNNaker was initially designed with Izhikevich’s point neuron model [24] as the target model. Although their software-based architecture could support a variety of more sophisticated neural models; currently the only other neuron model is a simple leaky Integrate-and-fire point neuron. The neuron algorithm is programmed into the local memory of each of the SpiNNaker cores. Post-synaptic weights for synapses are stored in the SpiNNaker chip's shared memory; the algorithm fetches the corresponding weight into the local core memory whenever a spike arrives at one of its synapses, and it computes neuron action potentials at 1 ms simulation intervals. 16-bit fixed-point arithmetic is used for most of the computation, to avoid the need for a floating-point unit and to reduce energy, computational and space costs. A single SpiNNaker node is able to simulate 16 K neurons with 1000 synapses each within a power budget of 1 W (energy per synaptic event 10^{-8} J) [36].

Future development of SpiNNaker will be done within the European Human Brain Project [35]. The next version of the SpiNNaker chip is planned for a 28 nm CMOS technology, integrating 68 ARM M4 cores (400 MHz, 1 W) with floating-point support, improved power management, energy efficient inter-chip links, and external 2 GByte shared memory.

18.8 The SyNAPSE Program and the IBM TrueNorth Architecture

In 2009, the US DARPA launched the SyNAPSE program: Systems of Neuromorphic Adaptive Plastic Scalable Electronics [38]. It says in its description: "As compared to biological systems…., today's programmable machines are less efficient by a factor of 1 million to 1 billion in complex, real-world environments". And it continues: "The vision … is the enabling of electronic neuromorphic machine technology that is scalable to biological levels." SyNAPSE is a program with explicit specifications. It requires an existing system-simulation background (like the Blue Brain Project [20]) to assess the likelihood of realizing the milestones, which are structured into four phases of $\sim$2 years each:

1. The entry-level specification for synapse performance is:
   - Density scalable to $>10^{10}$ cm^{-2}, <100 nm^2.
   - Energy per synaptic operation <1 pJ, <100× nature.
   - Operating speed >10 Hz (equivalent to nature).
   - Dynamic range of synaptic conductance >10.
2. ($\sim$2010/2012) Specify a chip-fabrication process for $>10^6$ neurons/cm^2, 10^{10} synapses/cm^2. Specify an electronics implementation of the neuromorphic design methodology supporting $>10^{10}$ neurons and $>10^{14}$ synapses, mammalian connectivity, <1 kW, <2 l (the final program goal).
3. ($\sim$2012/2014) Demonstrate chip fabrication $>10^6$ neurons/cm^2, 10^{10} synapses/cm^2. Demonstrate a simulated neural system of $\sim 10^6$ neurons performing at "mouse" level in the virtual environment.
4. ($\sim$2014/2016) Fabricate a single-chip neural system of $\sim 10^6$ neurons and package into a fully functional system. Design and simulate a neural system of $\sim 10^8$ neurons and $\sim 10^{12}$ synapses performing at "cat"-level environment.
5. ($\sim$2016/2018) Fabricate a multi-chip neural system of $\sim 10^8$ neurons and instantiate into a robotic platform performing at "cat" level (hunting a "mouse").

The “Cognitive Computing via Synaptronics and Supercomputing” (C2S2) project is a funded project from DARPA’s SyNAPSE initiative. The six partners (IBM Almaden Research Lab and the five universities Cornell, Columbia, Stanford, Wisconsin Madison, and UC Merced) bringing in expertise in neuroscience, psychology, VLSI, and nanotechnology. They created a massively parallel cortical simulator called C2, which was initially used at the scale of a rat cortex, and more recently at the scale of a cat cortex, running on IBM’s Dawn Blue Gene/P supercomputer, with 147,456 CPUs and 144 TB of main memory. The C2 simulation used a much simpler model of neurons than the Blue Brain, with single-compartment spiking Iszhikevich-type neurons.

The group will turn to digital special-purpose hardware for brain emulation. The prototype chip (45 nm SOI process, 2 mm × 3 mm die size) emulates 256 neurons, using a crossbar connecting 1024 input axons to the 256 neurons with weighted synapses at the junctions. Variations of the chip have been built with 1-bit and 4-bit synapse weights stored in SRAM. Another was built with low leakage transistors to reduce power consumption. Cross-chip spikes are conveyed asynchronously via AER networking, while the chips themselves operate synchronously. Synapses are simulated using the Izhikevich leaky integrate-and-fire model. The results are identical to the same equations simulated in software, but all 256 neurons on the chip update their membrane voltage in parallel, at 1 ms intervals (45 pJ/spike) [39].

The successor of this prototype chip is the IBM True North chip (Fig. 18.7), integrating a two-dimensional on-chip network of 4096 digital application-specific cores (64 × 64) and over 400 Mio. bits of local on-chip memory (~100 Kb SRAM per core) to store synapses and neuron parameters as well as 256 Mio. individually programmable synapses on-chip [40]. One million individually programmable neurons can be simulated time-multiplexed per chip, sixteen-times more than the current largest neuromorphic chip. The chip with about 5.4 billion transistors is fabricated in a 28 nm CMOS process (4.3 cm^2 die size, 240 μm × 390 μm per core). By device count, True North is largest IBM chip ever fabricate and the second largest (CMOS) chip in the world. The routing network extends across chip

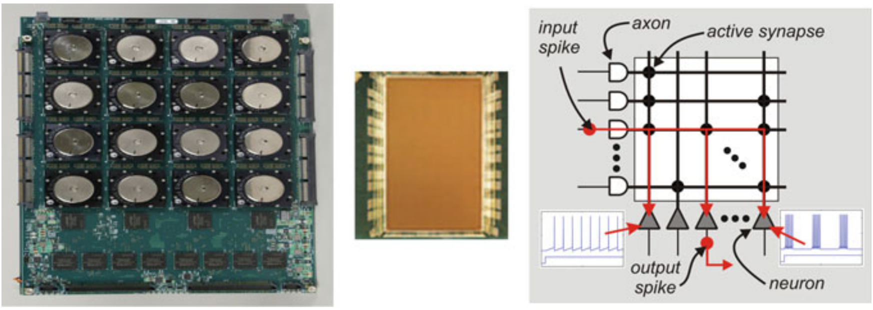


Fig. 18.7 A multi-chip board (*left*) of 16 IBM TrueNorth chips (*middle*) integrating 64 × 64 digital neurosynaptic cores. Each core implements 256 neurons with 1024 spike-inputs (*right*) [41]. ©IEEE 2014

boundaries through peripheral merge- and split-blocks. The total power, while running a typical recurrent network at biological real-time, is about 70 mW resulting in a power density of about 20 mW/cm^2 (about 26 pJ per synaptic event), which is in turn comparable to the cortex but three to four orders-ofmagnitude lower compared to 50–100 W/cm^2 for a conventional CPU.

18.9 The BrainScaleS Wafer-Scale Neuromorphic Hardware System

The European funded research project BrainScaleS (Brain-inspired multiscale computation in neuromorphic hybrid systems) aimed at understanding and emulating functions and interactions of multiple spatial and temporal scales in brain-information processing [42]. It is a collaboration of 19 research groups from 10 European countries and built on the research carried out in the former European funded FACETS project (2005–2010) [43]. Both, numerical simulations on Petaflop supercomputers and fundamentally different non-von-Neumann hardware architectures were employed for this purpose. Within its broad scope of advancing neuromorphic computing, the hardware part is a very-large-scale, mixed-signal implementation of a highly connected, adaptive network of analog neurons. The basic element is the HICANN (High Input Count Analog Neural Network) chip hosting one analog network core (ANC) and necessary support circuitry for communication as well as controlling. The ANC was implemented in a 180 nm CMOS technology and has a total of 112 K synapses and 512 neuron circuits. It could simulate, for example, 8 neurons with 14 K inputs, or 512 neurons with 224 inputs. The goal was to simulate analog neuron waveforms analogous to biological neurons on the same input. The implemented analog neuron model is the exponential integrate-and-fire model (AdExp) [44] and can be configured by 23 individual analog parameters, which are stored in single-poly floating-gate analog memory cells. The area of the analog neuron circuit is 1500 μm^2. The membrane capacitance is implemented as a capacitor on top of the transistor circuits, thus occupying no additional silicon area. Most of the internal currents stay in the range of 100 nA to 1 μA. The synapse weight is stored in a 4-bit SRAM and is represented as a current generated by a 4-bit multiplying DAC. The synapse area is 150 μm^2. Two synapse columns of the ANC can be combined to realize a weight resolution of 8 bit at the expense of bisecting the number of available synapses for the ANC neuron circuits.

A special feature of the BrainScaleS hardware system is wafer-scale integration of the HICANN chips (Fig. 18.8). A total of 384 HICANN chips can be interconnected on an 8-inch silicon wafer, implementing 196,608 neurons and 44 Mio. synapses. The HICANN analog neurons communicate with each other digitally. The backbone of the communication on the wafer is a grid of horizontal and vertical buses enabling the transport of spikes between all analog neurons on the wafer chips. The additional wiring is on top of the manufactured wafer and is fabricated as

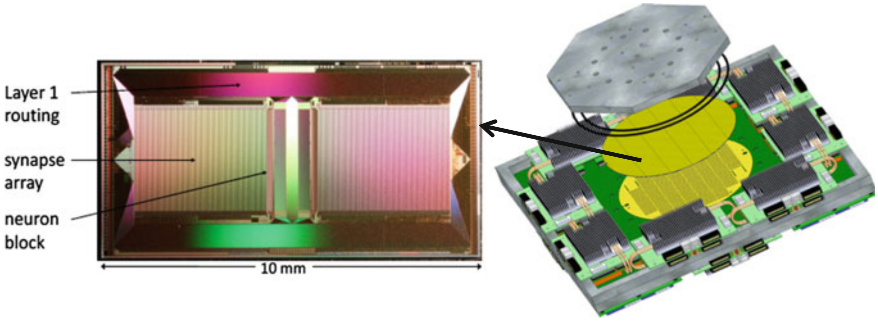


Fig. 18.8 Microphotograph of the HICANN chip (*left*) and view of the BrainScales wafer module (*right*) [46]. ©IEEE 2010

a custom back-end-of-line structure. The manufacturing of this crossbar metal-interconnect network would be a perfect example of future high-speed, maskless electron-beam lithography (Chap. 8). This communication infrastructure is the basis for a hierarchical packet-based routing network with fixed propagation delays. The propagation time for a recurrent connection on a HICANN has been measured as 120 ns. The additional time needed to transmit a pulse across the whole wafer is typically less than 100 ns with a maximum jitter of 4 ns [45].

One key target of the BrainScaleS hardware is a 10^4-fold speed-up of the natural neuron-firing rate of 10 Hz. The wafer is organized into 384 chips with a maximum of 196.608 neurons with 224 inputs (synapses) resulting in about 2×10^9 events/s and 64 Gb/s per wafer. On wafer communication of spikes is digital with 6 bits/event over 64 horizontal and 256 vertical links (1 to 2 Gbit/s) per HICANN chip offering sufficient communication bandwidth. Larger systems can be built by interconnecting several wafer modules. For this purpose, a second communication-protocol layer is implemented within the HICANN chips and on standard chips (e.g. FPGAs) between wafers. Neuron outputs, inputs, circuit parameters, and interconnections can be monitored and controlled by software running on a host computer.

The next version of the neuromorphic wafer-scale system will be developed within the European Human Brain Project. The next HICANN chip is planned for a 65 nm CMOS technology with higher resolution and better precision of the neuron- and synapse-circuits. Furthermore, the communication system will be improved.

18.10 Neurogrid

The Neurogrid project at Stanford University uses programmable analog “neuro-core” chips [47]. Each 12×14 mm^2 CMOS chip (180 nm CMOS) can emulate over 65,000 neurons, and 16 chips are assembled on a circuit board to emulate over a million neurons (Fig. 18.9). The entire 1 M-neuron system consumes about 3.1 W.

The Neurogrid neuron circuit consists of about 300 transistors modelling the components of the cell, with a total of 61 graded and 18 binary programmable parameters. Each programmable neurocore models the ion-channel behaviour and synaptic connectivity of a particular neuron-cell type or cortical layer. Neurogrid uses a two-level simulation model for neurons, in contrast to the point-neuron model used in SpiNNaker, and in contrast to the thousands of compartments used in Blue Brain's simulation. Neurogrid uses this approach as a compromise to provide reasonable accuracy without excessive complexity. A quadratic integrate-and-fire model is used for the neuron body. Dendritic compartments are modelled with up to four Hodgkin-Huxley channels. Back-propagation of spikes from somatic to dendritic compartments is supported.

Neurogrid uses local analog wiring to minimize the need for digitization for on-chip communication. Spikes rather than voltage levels are propagated to destination synapses. To simplify circuitry, a single synapse circuit models a neuron's entire synapse population of a particular type, and each of these circuits must be one of four different types. The synapse circuit computes the net postsynaptic conductance for that entire population from the input spikes received. Although this approach limits the ability to model varying synaptic strength, and it does not model synaptic plasticity, it greatly reduces circuit complexity and size. Synapses can be excitatory, inhibitory, or shunting.

Like the other systems, Neurogrid uses an AER packet network to communicate spikes between chips. Neurogrid's chips are interconnected in a binary tree with links supporting about 80 million spikes/second. Routing information is stored in RAM in each router supporting programmable-weight connection. Like SpiNNaker, the Neurogrid neuron array is designed to run in biological real-time. This means that a single AER link can easily service all of the cross-chip spikes for 65,000 neurons. Furthermore, the on-chip analog connections can easily service their bandwidth, and it seems likely that the binary routing tree, connecting the 16 Neurogrid chips on a circuit board, can easily support a million neurons. The Neurogrid group has demonstrated that their neurons can emulate a wide range of behaviours.

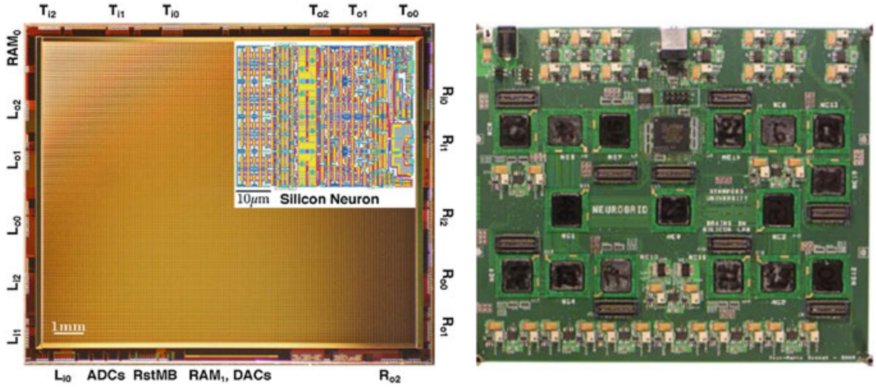


Fig. 18.9 Neurocore with up to 256 × 256 neurons (12 × 14 mm^2) (*left*) assembled on a 16 × 19 cm^2 circuit board (*right*) [48]. ©IEEE 2014

18.11 Comparison

The projects focused on in this chapter use different technological approaches to the implementation of ANNs. The Blue Brain project and the SpiNNaker system simulate ANNs in software on general-purpose processors. Whereas Blue Brain employs high-performance computers (HPC) without bio-inspired architectural hardware adaptations, SpiNNaker relies on embedded low-power processor cores from the mobile world, distributed private memory per core, and a communication network optimized for transmitting "spikes" asynchronously utilizing the address-event-representation (AER). Both approaches make use of the concept of the virtualization that many "neurons" can be simulated time-multiplexed on the same digital core. The TrueNorth implementation is based on a digital application-specific core per neuron, local on-chip-memory for the synapses, and a specialized routing network extending across chip boundaries. BrainScaleS and Neurogrid use a "neuromorphic" approach, with dedicated, adjustable analog circuitry for every neuron in the ANN, adaptive on-chip synapses, and a configurable interconnection network.

The approaches have their specific pros and cons. The neuromorphic ASICs avoid the substantial computational overhead of software simulation and may produce a more biologically-accurate result in less time. On the other hand, for digital implementations, there is no A/D conversion and the cost of the network routing logic is amortized over 1000 emulated neurons per CPU (virtualization). All approaches face the problem of spike networking. Routing AER packets [37] in real-time from tens of billions of neurons is a challenge. The logic circuitry required for decoding and routing may be much larger than the neuron emulation circuit itself. Another issue with AER networking is the timing of spikes. Neurons adapt to early and late signals over time, and some signal-timing tuning is performed by the axons. According to the routing network, the timing of spikes originating from the same neuron varies (jitter) in the proposed network implementations. Proper synchronization can be achieved by inserting delays or reserving communication bandwidth, as proposed in [49].

Synaptic plasticity and learning present the biggest challenges to artificial brain projects. On the one hand, our knowledge about plasticity, learning, and memory is incomplete [50]. On the other hand, our technologies are far less plastic and compact than neural tissue. Experimental evidence for some basic synaptic plasticity mechanisms exist. There is also evidence for neurons growing new dendrites and synapses to create new connections as well as changing the "weight" of existing synapses by increasing or decreasing the number of neurotransmitter vesicles or receptors for the neurotransmitters. Today, the efficient implementation of a writable and non-volatile synapse weight is a hot research topic. Within the discussed projects, the synapses are implemented digitally: 1 bit (TrueNorth), 4–8 bit (HICANN), 13 bit shared (Neurogrid, off-chip), and 16 Bit SpiNakker (off-chip). Whereas TrueNorth and Neurogrid do not support learning, the HICANN chips (hardware-based learning) and SpiNakker (software) include learning. Table 18.2

Table 18.2 Comparison of neuro-ASICs

Neuro-ASICS	Feature size (nm)	Die size (cm^2)	Neurons	Synapses	Bit/synapse	ESE
SpiNakker	130	1.02	1600	128×10^{6a}	8	10^{-8} J
TrueNorth	28	4.3	10^6	256×10^6	1	10^{-11} J
HICANN	180	0.5	8–512	114,688	4–8	10^{-10} J
Neurogrid	180	1.68	65,536	16×10^{6a}	13^b	10^{-10} J

Numbers per chip: [a]off-chip, [b]shared, *ESE* Energy/synaptic event

summarizes chip characteristics of the basic building block (Neuro-ASIC) of the discussed approaches. For the energy demand per synaptic event (ESE) only a rough estimation of the expected magnitude is given. The exact determination of the ESE is difficult and not standardized yet.

Independent of the technological approach, the projects differ in the level of bio-realism and computational sophistication in their emulation of neurons and synapses. SpiNNaker and TrueNorth work with a point-neuron model, as recommended by Izhikevich [24]. A two-level analog model such as Neurogrid's two-compartment circuits, or the BrainScaleS HICANN chip's separate dendritic membrane circuits, allows more sophisticated neural emulations, depending on the complexity of the compartment emulations. The most bio-realistic approach among the projects is the fully compartmentalized model of the neuron of the Blue Brain Project, representing a biological neuron as hundreds of independent compartments, each producing an output based on adjacent ion-channels and regions, and using the computationally expensive Hodgkin-Huxley equations [33] to compute the potential bio-realistically in each compartment.

There is clearly room for scaling for all projects, and it will be interesting to follow the digital-versus-analog strategy, considering the alternative of digital 8b × 8b multipliers with 1 fJ and 100 μm^2 per multiplication (see Sect. 1.5). With respect to system scaling, the power efficiency of chip/wafer-level interconnects is relevant. With the optimum efficiency of 1 mW/(Gb/s) (see Chap. 5), the resulting 74 W could be handled. The more likely level in the project at a less advanced technology node would be 1 kW [19]. Scaling this up to mega-neurons clearly shows that power efficiency is the number 1 concern for these complex systems.

3D integration is the further challenge for any Si brain. The memory Si layer on top of the mixed-signal Si layer can be achieved with a through-silicon-via (TSV) technology (Chap. 3), and it is only one level in the task of building the whole system, because the global programmable digital control could be added on top. In the BrainScaleS and Neurogrid architecture, the digital control is implemented with FPGA (field-programmable logic array) chips on a printed-circuit board.

18.12 Outlook

The building blocks for ICs and for the Brain are the same at nanoscale level: electrons, atoms, and molecules, but their evolutions have been radically different. The fact that reliability, low-power, reconfigurability, as well as asynchronicity have been brought up so many times in recent conferences and articles, makes it compelling that the Brain should be an inspiration (at many different levels), suggesting that future nano-architectures could be neural-inspired. The fascination associated with an electronic replication of the human brain has grown with the persistent exponential progress of chip technology. The present decade 2010–2020 has also made the electronic implementation more feasible, because electronic circuits now perform synaptic operations such as multiplication and signal communication at energy levels of 10 fJ, comparable to biological synapses. Nevertheless, an all-out assembly of 10^{14} synapses will remain a matter of a few exploratory systems for the next two decades because of several challenges.

One challenge lies in mastering the design complexity and achieving economic viability for integrated systems with more than a billion devices per square centimetre. This requires system concepts that both exhaust the possibilities of future technologies and reduce the design-as well as the test-complexity. Due to their highly regular and modular structure, **inherent fault-tolerance**, and learning ability, ANNs offer an attractive alternative for ultra-large-scale integration and the development of **resource-efficient systems** with minimal total energy consumption combined with a small size and fault-tolerant behaviour. These arguments were already a strong motivation for ANN hardware in the 80s. Despite the impressive development of nanoelectronics during the last decades, there is still no clear consensus on how to exploit this technological potential for massively-parallel ANN implementations. The projects discussed in this chapter rely on more brain-inspired ANN models (spiking ANNs) than the forerunners two decades ago. From the point-of-view of basic research, this development is comprehensible, but in respect to applications the benefit of these new architectures is still ambiguous. Another challenge is the adaptation and control of brain-like architectures. Fuzzy-and neuro-control have been practically available for 25 years, and yet their application is still limited to applications that are just feature-oriented and not critical for the technical performance of a product or service.

Hence, it is currently quite difficult to determine the best way to perform ANN calculations for any given application. This is one reason for the huge variety of approaches to ANN hardware implementation known in literature. The problem of benchmarking and an adequate metric for performance evaluation is still open, too. So the discussion is open about the best way to achieve very large neural systems and, in the long term, how to produce so-called artificial brains. We are still a long way from fully comprehending the functional mechanisms of the brain; and the construction of an artificial brain will remain for a very long time, if not forever, a fantasy. This must be taken into account in the new discussion on singularity hypotheses [51] and the emergence of machine superintelligence. We know from

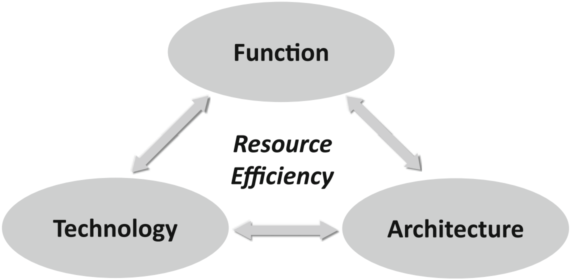


Fig. 18.10 Integrated design view on resource-efficient systems engineering

systems engineering that there is a close interdependence between the main three system views (function, architecture, technology) for the development of resource-efficient technical systems (Fig. 18.10). Biology has taken its own way through evolution based on its own special technology (real wet tissue). Exact brain emulation in software or implementation in dry solid-state circuitry may guide us the wrong way to artificial machine intelligence as we do not adequately account for the influence of the technology on function (behaviour) and system architecture. Probably there are many technological artefacts in brain measurements, which are irrelevant for systems behaviour and hence for system emulation. For example, the impressive brain simulations on supercomputers with neuron numbers comparable to the brains of mouse or cat are still not able to perform some simple task within a natural environment. Consequently, we are far from any accepted description of the principles of information processing in brains and from reconstruction of its capabilities.

Nevertheless, we do have much to learn from brains from the computational standpoint and about the implementation of resource-efficient technical systems. The hardware realization of neural networks should not aim for an exact reproduction of nervous systems, but simply for an efficient use of available technologies for solving practical problems. Remember that creating human life had fascinated researchers for centuries, and every technological epoch had its view on how to do this (e.g. mechanical automata of the eighteenth century). Now that we are close to building an equivalent electronic core, we should also involve the integration of the periphery, namely the innumerable sensors and actuators (motors). In this outside-of-electronics domain, that of MEMS, remarkable progress has been made and continues with high growth (Chaps. 13 through 17). Furthermore, the critical conversion of analog sensor signals to digital signals is advancing rapidly (Chap. 4). And yet it is very early in the evolution of generic neuromorphic peripheries, so that all practical systems will be focused, application-specific solutions, certainly with growing intelligence, but with confined features of neural networks. An outstanding example of an intelligent vision sensor system based on an efficient combination of classical computer vision and brain-inspired hardware architecture was developed by Prof. Ramacher from Infineon Technologies in the course of several research projects funded by the German Federal Ministry of Education and Research (BMBF) [52].

It will take a major, globally consolidated effort involving many disciplines from reliability and ethics to social science to achieve a broad acceptance of brain-inspired hardware for demanding applications. The advancement of care-bots in various world regions will be a good test-ground for this evolution of neuromorphic systems. Particularly attractive is the application of ANNs in those domains where, at present, humans outperform any currently available high-performance computer, e.g. in areas like vision, auditory perception, or sensory motor-control. Neural information processing is expected to have a wide applicability in areas that require a high degree of flexibility and the ability to operate in uncertain environments where information usually is partial, fuzzy, or even contradictory. Even more computational power may be obtained by emerging technologies like quantum computing, molecular electronics, or novel nano-scale devices (memristor, spintronics, nanotubes (CMOL)), but these technologies will not be available on broad basis in the next decade. Today, we are still early in the efficient use of nanoelectronics, and we are keenly awaiting the technology we can use tomorrow.

References

1. Mead, C., Ismail, M. (eds.): Analog VLSI Implementation of Neural Systems. Springer, Berlin (1989). ISBN 978-0-7923-9040-4
2. Steinbuch, K.: Adaptive networks using learning matrices. Kybernetik **2**, 148–152 (1965)
3. Widrow, B.: Pattern recognition and adaptive control. IEEE Trans. Appl. Indus. **83**(74), 269–277 (1964)
4. IJCNN, International Joint Conference on Neural Networks, http://www.ijcnn.org
5. NIPS, Neural Information Processing Systems, http://nips.cc
6. Kohonen, T., et al. (eds.): Artificial neural networks. In: Proceedings of the first ICANN in Espoo, Finland, vol. 1, 2. North-Holland, Amsterdam (1991). ISBN 0 444 89178 1
7. Ramacher, U., Rückert, U. (eds.): VLSI Design of Neural Networks. Kluwer Academic, Boston (1991)
8. MicroNeuro: Conference on "Microelectronics for Neural Networks"; Dortmund, Germany (1990); Munich, Germany (1991); Edinburgh, Scotland (1993); Torino, Italy (1994); Lausanne, Switzerland (1996); Dresden, Germany (1997); Granada, Spain (1999)
9. Hammerstrom, D., Nguyen, N.: System design for a second generation neurocomputer. In: Proceedings of the IJCNN II, pp. 80–83 (1990)
10. Data booklet for Intel 80170NX (ETANN) Electrically Trainable Analog Neural Network. Intel Corp. (1991)
11. Ramacher, U.: SYNAPSE: a neurocomputer that synthesizes neural algorithms on a parallel systolic engine. J. Parallel Distrib. Comput. **14**(3), 306–318 (1992)
12. Brain Facts, Society of Neuroscience, www.snf.org (2008)
13. www.wikipedia.org/Brain (Dec 2015)
14. Chudler, E.H.: Neuroscience for kids: http://faculty.washington.edu/chudler/synapse.html (Dec. 2015)
15. Stufflebeam, R.: Neurons, synapses, action potentials, and neurotransmission. The Mind Project, www.mind.ilstu.edu/curriculum/neurons_intro (2008)

16. Martini, F.H., Nath, J.L.: Neural tissue, chapter 12. In: Fundamentals of Anatomy and Physiology. Prentice-Hall, New Jersey (2008)
17. Sengupta, B. et al.: Action potential energy efficiency varies among neuron types in vertebrates and invertebrates, PLoS Computat. Biol. (2010). doi:10.1371/journal.pcbi.1000840
18. Furber, S., Temple, S.: Neural systems engineering. J. R. Soc. Interface **4**(13), 193–206 (2007)
19. Höfflinger, B.: Chips 2020, chapter 18, vol. 1. Springer, Berlin (2012)
20. Markram, H.: The blue brain project. Nat. Rev. **7**, 153–160 (2006). http://bluebrain.epfl.ch
21. Inne, P.: Digital connectionist hardware: current problems and future challenges, biological and artificial computation: from neuroscience to technology. Lecture Notes in Computer Science, vol. 1240, pp. 688–713. Springer, Berlin (1997)
22. Palm, G., et al.: Neural associative memories. In: Krikelis, A., Weems, C.C. (eds.) Associative Processing and Processors, pp. 307–326. IEEE CS Press, Los Alamitos (1997)
23. Beiu, V., Quintana, J.M., Avedillo, M.J.: VLSI implementations of threshold gates—a comprehensive survey. Spec. Issue Hardware Implementations Neural Netw. IEEE Trans. Neural Netw. **14**(5), 1217–1243 (2003)
24. Izhikevich, E.M.: Which model to use for cortical spiking neurons? IEEE Trans. Neural Netw. **15**, 1063–1070 (2004)
25. Strey, A.: Spezifikation und parallele Simulation neuronaler Netze, Fortschrittbericht, vol. 661. Reihe Informatik/Kommunikationstechnik, VDI-Verlag (2001)
26. Eichner, H. et al.: Neural simulations on multi-core architectures. Front. Neuroinformatics **3** (2009). doi:10.3389/neuro.11.021.2009
27. Gerland, M., et al.: Parallel Computing Experiences with CUDA. IEEE Micro. **28**(4), 13–27 (2008)
28. Oh, K.S., Jung, K.: GPU implementation of neural networks. Patt. Recogn. **37**(6), 1311–1314 (2004)
29. Omondi, A.R., Rajapakse, J.C. (eds.): FPGA Implementations of Neural Networks. Springer, Berlin (2005)
30. Koester, M., et al.: Design optimizations for tiled partially reconfigurable systems. IEEE Trans. Very Large Scale Integr. Syst. **19**(6), 1048–1061 (2011)
31. Porrmann, M., Witkowski, U., Rückert, U.: Implementation of self-organizing feature maps in reconfigurable hardware, in [29], pp. 253–276. Springer, Berlin (2005)
32. Klar, H., Ramacher, U. (eds.): Microelectronics for Neural Networks. VDI Fortschrittberichte, Reihe 21, Nr.42 (1989)
33. Hodgkin, A.L., Huxley, A.F.: A quantitative description of membrane current and its application to conduction and excitation in nerves. J. Physiol. **117**, 500–544 (1952)
34. Djurfeldt, M., et al.: Brain-scale simulation of the neocortex on the IBM Blue Gene/L supercomputer. IBM J. Res. Dev. **52**(1/2), 31–41 (2008)
35. www.humanbrainproject.eu
36. Furber, S., et al.: Overview of the SpiNNaker system architecture. IEEE Trans. Comput. **62**(12), 2454–2467 (2013)
37. Mahowald, M.: VLSI analogs of neural visual processing: A synthesis of form and function, PhD thesis, California Institute of Technology (1992)
38. http://www.darpa.mil/Our_Work/DSO/Programs/Systems_of_Neuromorphic_Adaptive_Plastic_Scalable_Electronics_%28SYNAPSE%29.aspx
39. Merolla, P.A., et al.: A digital neurosynaptic core using embedded crossbar memory with 45 pJ per spike in 45 nm. In: Proceedings of IEEE CICC, pp. 19–21 (2011)
40. Merolla, P.A., et al.: A million spiking-neuron integrated circuit with a scalable communication network and interface. Science **345**, 668–673 (2014)
41. Cassidy, A.S., et al.: Real-time scalable cortical computing at 46 Giga-synaptic OPS/Watt with ~100× speedup in time-to-solution and ~10,000× reduction in energy-to-solution. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 27–38 (2014)
42. http://brainscales.kip.uni-heidelberg.de

43. http://www.facets-project.org; http://facets.kip.uni-heidelberg.de
44. Brette, R., Gerstner, W.: Adaptive exponential integrate-and-fire model as an effective description of neural activity. J. Neurophysiol. **94**, 3637–3642 (2005)
45. Schemmel, J., Fieres, J., Meier, K.: Wafer-scale integration of analog neural networks. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN) (2008)
46. Schemmel, J., et al.: A wafer-scale neuromorphic hardware system for large-scale neuron modeling. In: Proceedings of the IEEE International Symposium on Circuits and Systems (2010)
47. Silver, R., et al.: Neurotech for neuroscience: unifying concepts, organizing principles, and emerging tools. J. Neurosci. **27**, 11807–11819 (2007)
48. Benjamin, B.V., et al.: Neurogrid: a mixed-analog-digital multichip system for large-scale neural simulation. Proc. IEEE **102**(5), 699–716 (2014)
49. Philipp, S. et al.: Interconnecting VLSI spiking neural networks using isochronous connections. In: Proceedings of 99th International Work-Conference on Artificial Neural Networks, Springer LNCS 4507, pp. 471–478 (2007)
50. Ziv, N.: Principles of glutamatergic synapse formation: seeing the forest for the trees. Curr. Opin. Neurobiol. **11**, 536–543 (2001)
51. Eden, A.H., et al. (eds): Singularity Hypotheses, The Frontiers Collection. Springer, Berlin (2012)
52. Ramacher, U., von der Marlsburg, C. (eds): On the Construction of Artificial Brains. Springer, Berlin (2010)

Chapter 19
Energy-Harvesting Applications and Efficient Power Processing

T. Hehn, D. Hoffmann, M. Kuhl, J. Leicht, N. Lotze, C. Moranz, D. Rossbach, K. Ylli and Y. Manoli

Abstract In comparison to the original chapter in CHIPS 2020 Manoli et al. (CHIPS 2020—A Guide to the Future of Nanoelectronics: 329–420, 2012) [1], this chapter presents more application-oriented research with a focus on wearable devices and condition monitoring. It also covers electronic circuit components and systems employed in extracting, processing, and storing the harvested power. In the meantime, many innovative enhancements in terms of efficiency and applicability have been achieved by developing dedicated CMOS integrated circuits.

19.1 Systems and Applications

The first part provides an insight into applications where energy-harvesting systems are used. In this sense, the term "system" means the combination of the generator and power management circuit, which makes it usable in real applications.

19.1.1 Wearable Devices

Improvements in power consumption and the miniaturization of electronic systems are leading to an increasing range of devices designed specifically with mobility in mind. Unobtrusive devices, which support modern users throughout everyday activities, are being developed for integration into textiles in order to reach this

T. Hehn (✉) · D. Hoffmann · D. Rossbach · K. Ylli · Y. Manoli
Hahn-Schickard-Gesellschaft für angewandte Forschung e.V., Wilhelm-Schickard-Straße 10,
78052 Villingen-Schwenningen, Germany
e-mail: Thorsten.Hehn@Hahn-Schickard.de

M. Kuhl · J. Leicht · N. Lotze · C. Moranz · Y. Manoli
Department of Microsystems Engineering—IMTEK, Albert-Ludwigs-Universität Freiburg,
Georges-Köhler-Allee 102, 79110 Freiburg, Germany

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_19

goal. There are a few drawbacks to the body-worn devices however. Battery replacement and maintenance is the key problem, for which the development of energy harvesting devices can provide the solution. Body-worn devices, which can generate electrical energy out of the energy readily available in the surrounding environment, have been presented. Thermoelectric generators exploit the temperature gradient between the body and ambient air [2], woven photovoltaic fibers are being developed to exploit energy radiating from light sources [3, 4], and kinetic harvesters make use of the human body motion [5, 6].

The sole of the shoe offers a confined, but structurally stable space, into which kinetic harvesters can be integrated without major impacts on user comfort and appearance. Consequently, several devices were designed for integration into a shoe, with a strong focus on kinetic-energy harvesting. There are three kinetic excitation sources found in the human gait.

The first source of energy is the force exerted on the shoe sole due to the weight of the user. This reaches dynamic values larger than the body weight [7]. Harvesting devices have been presented, which can directly convert this into electrical energy by means of electroactive materials. Piezoelectric materials (e.g. PZT or PVDF) directly generate a voltage when deformed, while dielectric elastomers (DE, e.g. acrylics, silicones) need to be biased in order to generate a voltage change when deformed [8, 9]. Different approaches translate the vertical motion into a horizontal motion to drive an electromagnetic generator [9] or generate a voltage through contact electrification, e.g. the triboelectric effect [10, 11].

The second source of energy is the impact upon heel strike, which generates acceleration impulses of up to 50*g* when the shoe contacts the ground [12]. An inductive shock-excited harvester as shown in Fig. 19.1 was presented, which uses this impulse to displace an inertial mass attached to a spring [12]. The inertial mass made up of a magnetic circuit then vibrates freely at its Eigen frequency until the motion is damped. During its motion, it moves past coils, which are fixed within the housing. The movement of the magnets thus induces a voltage within the coils to generate average output powers of up to 4 mW across an optimal resistive load [12].

The third energy source is the swing motion of the foot, which generates accelerations of up to 20*g* [12]. Several devices have been presented, in which magnets move linearly through coils to induce a voltage [13, 14]. The magnets act as the inertial mass. Average power outputs of more than 10 mW have been achieved. A flatter device, that can be integrated into the shoe sole more easily and generate average powers of up to 2 mW, was presented in [15]. Figure 19.2 shows an X-ray photography of this device.

19.1.2 Condition Monitoring

Modern technical assets such as production facilities and construction machines become more and more complex. Therefore, the risk of breakdown of a complete system due to the possible failure of crucial components increases. The economic

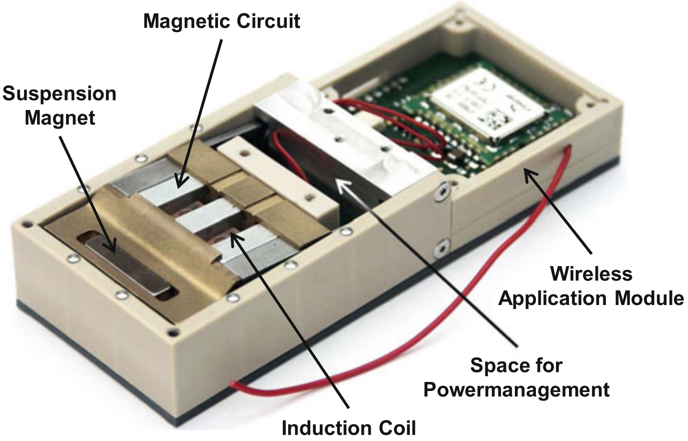
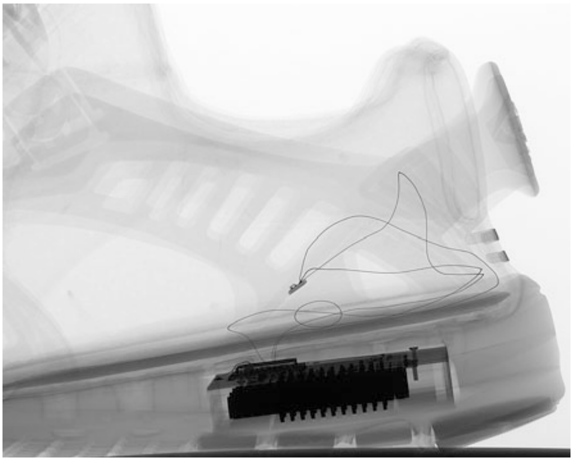


Fig. 19.1 Shock-excited shoe harvester including wireless application module [12]. For the shown configuration, the wireless application module implements the power management functionality. For other applications, a separate power management board can be inserted into the empty space in the middle

Fig. 19.2 X-ray photography of an integrated swing-type harvester [15]. Magnet stacks and coils can be seen (*dark areas*)



efficiency of a complex technical system is determined primarily by its reliability. Systematic preventive maintenance is therefore essential, in particular for technical systems with limited access.

The process of continuous or periodic monitoring of system components provides vital information about the physical condition of the system enabling early detection of developing failures. The knowledge about prospective failures avoids subsequent damage, a total breakdown of the system and allows scheduled maintenance. In this manner expenses caused by maintenance and machine down-time can be minimized. In this respect, condition monitoring (CM) pursues one of two main goals, which is system efficiency. Another important goal of CM is safety, in particular for technical systems, where human life is in danger in case of a breakdown. This applies to wind generators and shipping, for instance. In the last 10 years, CM systems have become an established technology for wind generators.

In the area of shipping, however, CM of large system components such as gearboxes is not state of the art yet. In order to guarantee a safe return to port of a vessel or ship, it is necessary to know about the condition of the machinery and to have some indication of approaching failures. One of the reasons why marine machinery CM has not been established in a greater scope is the investment involved with the installation and operation of a CM-system. In particular, the cabling for power supply appears to be a cost-intensive factor. Therefore, CM becomes more practical and acceptable, if the CM-system is self-sustaining, which means it does not require any cables or maintenance in terms of battery change. In this case, a wireless implementation of the CM-system is achievable.

A key technology for facilitating self-sustaining systems is based on the process of energy harvesting, in which specific devices convert ambient energy into electrical energy. In particular, the development of devices for vibration-energy harvesting is in focus of academic and commercial groups. The conducted research on these devices aims to increase the effectiveness (maximum power output at minimum size) and to broaden the operating bandwidth of conventional vibration transducers. Since variable-speed machinery such as electric drives or power trains in vessels exhibit variable vibration frequencies, self-tunable devices are potentially suitable for effective energy harvesting. Weddell et al. [16] reported on the development of a self-powered sensor system with a tunable vibration-energy harvester for wireless CM of a main engine of a ferry. The tuning mechanism of the tunable energy harvester is based on a magnetic principle, where the total system stiffness is altered by generating and changing an axial magnetic force. This is achieved by adjusting the distance between two magnets by translational motion of the so-called tuning magnet. The second magnet, called coupling magnet, is attached to the transducer structure. The translational motion of the tuning magnet, however, holds two disadvantages: First, the coupling magnet and the tuning magnet can be configured exclusively in the attractive or repulsive mode. Using both coupling modes in combination would enhance the feasible tuning bandwidth. Second, for translational motion, a linear actuator is required, which is more complex than a rotary actuator.

Indeed, a rotary actuator can also be deployed, however, only in combination with a gearing mechanism. Hoffmann et al. [17] presented a concept, which is based on the rotary motion of a cylindrical tuning magnet. Within a rotation angle of only 180°, both coupling modes can be utilized resulting in a broader tuning bandwidth. Moreover, the tuning magnet can be attached directly to a rotary actuator, making a gearing mechanism with its accompanying mechanical losses dispensable

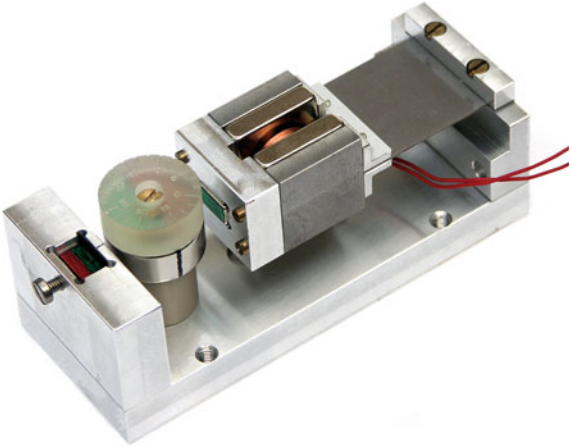


Fig. 19.3 Tuning mechanism using rotary motion of a cylindrical tuning magnet, applied to an inductive vibration based energy harvester

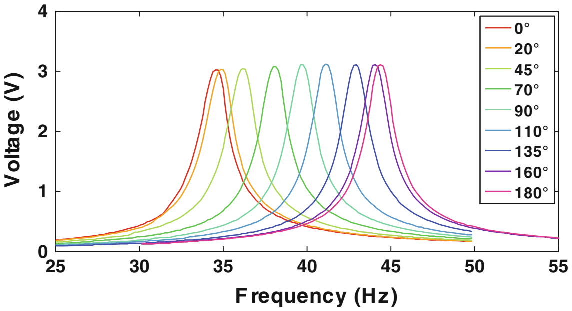


Fig. 19.4 Tuning range of the inductive vibration-based energy harvester

(Fig. 19.3). The tuning range can be designed by careful adjustment of the gap between tuning magnet and coupling magnet (Fig. 19.4). The smaller the gap the larger is the tuning bandwidth. The center frequency of the tuning bandwidth is adjusted by the stiffness of the cantilever in consideration of the mass of the magnetic circuit. The tuning bandwidth and center frequency was tailored to a specific CM application targeting the monitoring of maritime gear boxes (Fig. 19.5).

19.2 Circuit Components for Energy Harvesting Applications

The second part of this chapter describes the electrical circuit components, which are part of the energy harvesting systems. First, circuits suitable for generators with AC output are discussed. Afterwards, a review of the circuits appropriate for

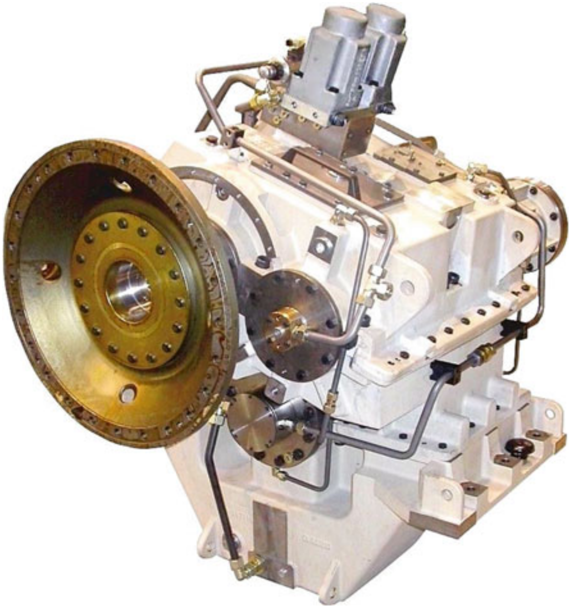


Fig. 19.5 Maritime gearbox whose vibration source is used as the harvesting source for condition monitoring (*Source* Reintjes GmbH)

generators with DC output is given. The last section covers analog and digital circuit techniques enabling ultra-low supply voltages.

19.2.1 AC Sources

19.2.1.1 Wireless Power Transmission Circuits

The possibility to transfer power without a cable connection, as depicted in Fig. 19.6, enables the implementation of isolated, battery-less sensor-nodes that are located behind an unperforated barrier like inside a hermetic housing or the human body.[1] As explained in [18], most wireless supply links incorporate a power transmitter that creates an alternating magnetic (HF) or electromagnetic (UHF) field and a power extraction frontend that is able to derive a regulated dc voltage from the field to supply connected appliances. The physical principle behind it is the magnetic coupling between two inductors that can be separated from each other by a distance of several centimeters and still have an electromagnetic connection. Or, if

[1] The circuits used to extract the power transmitted wirelessly face the same challenges as the circuits used to extract the power generated by energy harvesters, because the power budget available is often very small. Hence, the power transmitted wirelessly is treated as an “AC source” here.

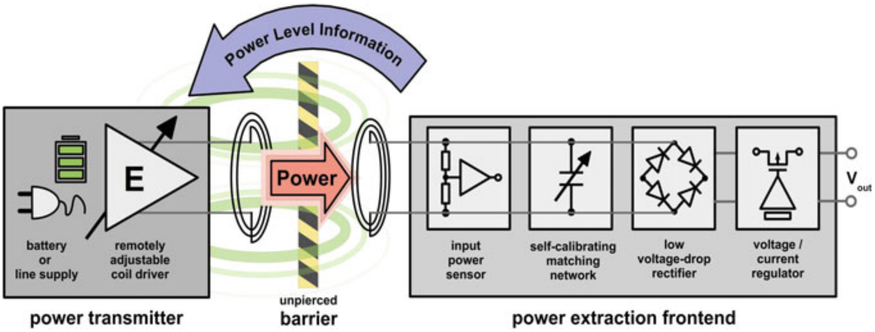


Fig. 19.6 Wireless supply link including a power transmitter and a power extraction frontend connected by two magnetically coupled inductors. The system diagram incorporates the reported efficiency-improving blocks, namely the closed-loop power supply, the self-calibrating matching network and the low voltage-drop rectifier

an even larger distance needs to be bridged, the ability of an electromagnetic wave to transport energy that can be harvested by a dipole antenna as exploited in [19].

When it comes to the design of such wireless supply links for smart sensor nodes, one of the main optimization goals is power efficiency: a strong energetic link between the primary and secondary side in combination with an efficient power management allows reducing the transmitted power dramatically. This promotes a versatile use within a broad field of applications, as it helps to meet legal limitations for radiated radio frequency (RF) power [18], to avoid tissue heating in medical implants [20, 21] or to simply spare the power budget of the transmitter unit, extending the lifetime of battery-driven devices.

One approach for improving the overall link efficiency is an optimized rectifier design. Since the rectifier is one of the few components conveying ALL the transmitted power, the voltage drop across it causes significant heat dissipation and power loss and must therefore be minimized. This can be achieved by using a "cross-coupled MOS" topology as proposed in [22, 23] or by using process-dependent devices like Schottky diodes [24] or even floating-gate diodes [25]. Another elegant solution, explained in [26, 27], is the implementation of a closed-loop wireless supply link, in which an abundance of transmitted power is detected by the power-extraction frontend and reported back to the power transmitter over a data communication channel. In this manner, the transmitter can always adjust the emitted RF power to the current need of the application. In order to achieve optimum transmission characteristics, the power receiver should be tuned to the frequency of the electromagnetic field (comparable to a radio channel) as exactly as possible. Since the relevant parameters are not only difficult to match in a production setup, but are also constantly changing due to variations in the environment, a dynamic self-calibrating tuning algorithm, that periodically adjusts an impedance-matching network at the receiver input to the optimum value is developed in [28, 29].

19.2.1.2 Interfaces for Vibration-Based Kinetic Energy Harvesting

Ambient vibrations are commonly available such as in industrial assembly lines or in means of transportation [40]. Thus, vibration-energy kinetic harvesting is an attractive possibility for enabling autonomous sensor systems. In the following, interface circuits for the two most promising conversion principles, namely the piezoelectric and the electromagnetic conversion principle, are presented.

Interface Circuits for Kinetic Piezoelectric Energy Harvesters

Compared to the capacitive and inductive conversion mechanism, the piezoelectric conversion principle offers the simplest approach for harnessing mechanical vibrations, because there is no need for having complex geometries and numerous extra components [30]. Lead-circonate titanate (PZT) is the piezoelectric material most commonly used at the moment. PZT can be simply deposited using thin-film and thick-film processes, which makes it attractive for MEMS applications. Coated on a carrier beam, PZT directly converts mechanical strain induced by vibrations into electric charge. Usually, two piezoelectric layers are electrically connected in parallel to achieve a higher output current. This configuration is called a bimorph. The electric charge produced when the piezoelectric beam harvester is being deflected has to be processed by an interface circuit. In the following, a review of several interface circuit architectures for piezoelectric generators is given. Two groups of interface circuits can be identified: The first group, which cares about highly efficient rectification including or excluding DC/DC conversion, and the second group, which increases the power output of piezoelectric harvesters by means of active energy extraction.

This paragraph covers the interface circuits belonging to the first group. Peters et al. [31] have presented an active rectifier fabricated in a 0.35 μm process, being able to rectify amplitudes down to 380 mV with an efficiency of up to 90 %, whereas the switch comparator consumes 266 nW. Measurements have shown operability for frequencies up to 1 kHz. A system including an additional DC/DC converter, fabricated in a 0.13 μm process and based on a simple hysteresis control, is presented in [32]. The regulated output voltage is 1.1 V for a power consumption of 67 μW. Rao et al. [33] have proposed an interface circuit composed of a rectifier and a step-up converter, fabricated in a 0.5 μm process. The chip is part of an energy harvester composed of a ball magnet rolling inside a spherical housing. When attached to a human ankle, the ball magnet moves relatively to a coil wound around the housing, generating 300 μW and hence charging a 3.7 V lithium ion battery.

In the following, active energy-extraction interface circuits are highlighted. The 0.35 μm chip presented in [34] implements a pulsed synchronous charge extraction (PSCE) method by means of an inductive boost converter as shown in Fig. 19.7. The energy is extracted from the piezoelectric energy harvester completely during a short time interval when a voltage maximum has been detected. Thus, due to the increased electrical damping effect, more energy can be extracted compared to a

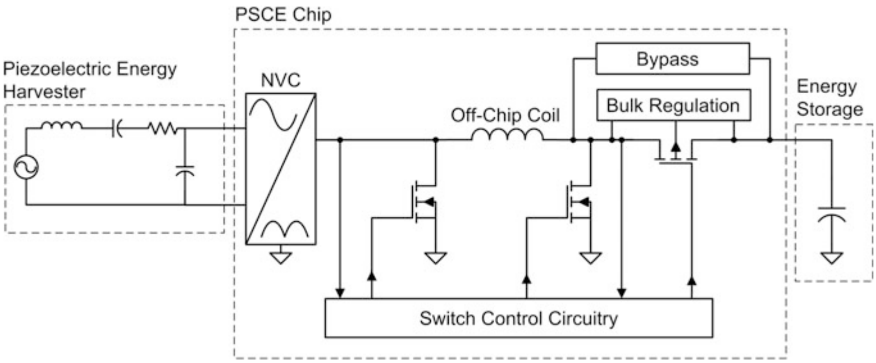
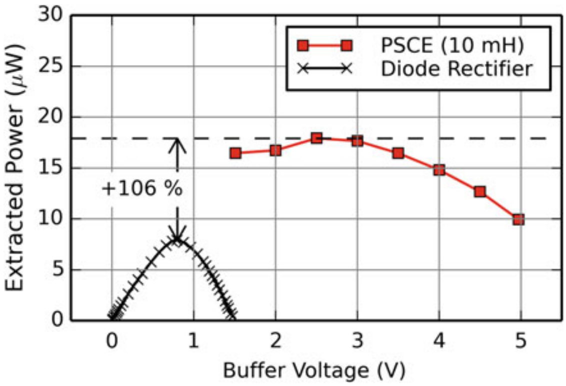


Fig. 19.7 Schematic of the pulsed-synchronous charge extraction (PSCE) circuit

diode rectifier [35]. After the passive cold startup, the chip is powered exclusively by the buffer capacitor and can handle output powers down to 5.7 μW and voltages from 1.3 to 20 V. Compared to a diode rectifier interface, the PSCE chip increases the extracted power to 123 % when the generator is driven at resonance, and to 206 % at off-resonance, as shown in Fig. 19.8. The chip micrograph is shown in Fig. 19.9. A rectifier-free topology performing synchronous charge extraction is proposed in [36]. On one hand, the omission of a rectifier further reduces power losses and saves valuable chip area. But on the other hand, energy is wasted since the energy is extracted only once per vibration period (whereas [34] extracts twice per period). The current consumption of this chip fabricated in a 5 V process is given at 700 nA. Another rectifier-free chip presented in [37] extracts the energy in several steps (multi-shot) when a voltage maximum is detected, hence reducing losses in the parasitic conduction resistances. A fly-back converter using coupled inductors is used instead of a boost converter, whereas the control circuitry, implemented in a 0.35 μm chip, consumes 1 μW. The chip shown in [38] is fabricated in a 0.18 μm process and implements a rectification method which shunts

Fig. 19.8 Comparison of the output power using the PSCE chip and a diode full-bridge rectifier



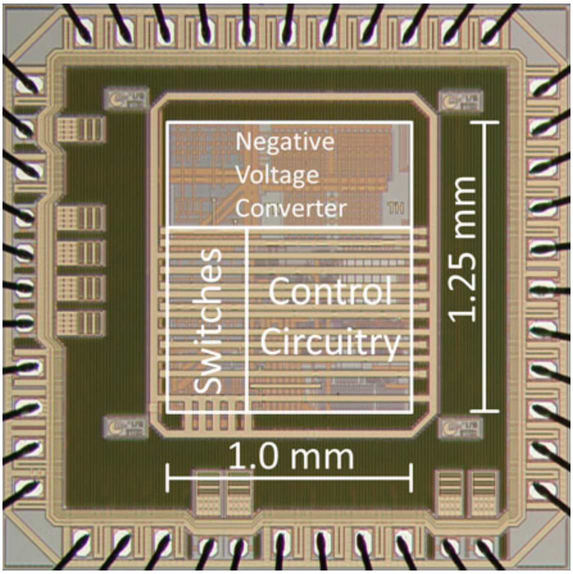


Fig. 19.9 PSCE chip micrograph

the piezoelectric harvester terminals at dedicated instants, increasing the power, which can be extracted. Consuming 1 µW and excited at 1 g, the harvester chip is able to charge a 20 mF super capacitor from 0 to 1.31 V during 8 min. Table 19.1 summarizes all the interface circuits for piezoelectric energy harvesters.

A rather new trend is the development of circuits consisting of organic field-effect transistors, which are flexible and hence adaptive to curved surfaces. Furthermore, circuits based on organic transistors consume less power than their inorganic (i.e. silicon-based) counterpart. In [39], Ishida et al. propose a pedometer built of an organic pseudo-CMOS circuit technique. Therefore, a −2 V charge pump has been implemented, making the circuits, made of p-channel MOSFETs, more robust and reducing their noise floor. The charge pump provides 12 µW with an efficiency of 65 %.

Interface Circuits for Kinetic Electromagnetic Energy Harvesters

Easy system integration by means of smart plug systems can be enabled by small-scaled electromagnetic vibration-energy harvesters, which efficiently convert ambient vibrations into electrical energy occupying only a cubic-centimeter-range volume [41]. Typically, such small transducers deliver electrical power in the micro- to milliwatt range depending on the harvester excitation and the harvester architecture. Interfacing the harvesting device with an energy storage element via a well-designed harvester-specific energy conditioning allows enhanced environmental energy scavenging [40].

In order to charge the storage device, the AC voltage of the harvester has to be converted into a DC voltage with a minimum power demand but nevertheless

Table 19.1 Overview of the interface circuits for piezoelectric energy harvesters

Ref.	Type	Process	Cold start-up capability	Min. power demand rating	Max. efficiency rating
[31]	Active rectifier	0.35 μm	Yes (no pre-charge needed, operation starts at 380 mV buffer voltage)	266 nw (power consumption)	90 % (voltage efficiency)
[32]	Active rectifier + hysteretic output voltage regulator	0.13 μm	Yes (no pre-charge needed)	67 μW (power consumption)	80 % (mechanical to electrical efficiency)
[33]	Active rectifier + inductive boost converter	0.5 μm	Yes (no pre-charge needed)	N/A	40 % (overall)
[34]	Pulsed energy extraction	0.35 μm	Yes (no pre-charge needed, 1.3 V startup voltage)	5.7 μW (power consumption)	73 % (harvesting efficiency, overall)
[36]	Rectifier-free pulsed energy extraction	5 V CMOS	Yes (no pre-charge needed)	700 nA (current consumption)	N/A
[37]	Rectifier-free multishot pulsed energy extraction	0.35 μm	Yes (no pre-charge needed)	1 μW (power consumption)	61 % (chip efficiency)
[38]	Rectifier with shunt-pass technique	0.18 μm	Yes (no pre-charge needed)	1 μW (power consumption)	60 % (harvesting efficiency, overall)

highly efficiently even under varying supply and load conditions. The active rectifier presented in [42] needs less than 140 nW and provides an efficiency of 90–94 % in a wide power range of 1–1 mW. This active rectifier and the one in [43] reduce the forward-voltage drop-down to a minimum of 20 mV.

In addition to the AC-DC conversion, a complete interface needs to optimize the energy extraction from the harvester [40]. Promising techniques for enhancing the electrical energy output of a small-scaled electromagnetic vibration energy harvester are real load matching [43] and maximum power-point tracking (MPPT) driven AC-DC conversion [44].

In the real load matching interface, the energy conditioning circuitry emulates the optimum energy harvester load resistance. In [43], an adaptive charge-pump with minimum power losses of 15 μW is applied. It starts operation at 0.8 V and transfers up to 48 % of the maximum harvester real power into a capacitor. To the knowledge of the authors, this is the only published ASIC applying the load matching algorithm to electromagnetic energy harvesters. However, in order to allow a comparison, the IC presented in [44], originally designed to interface with

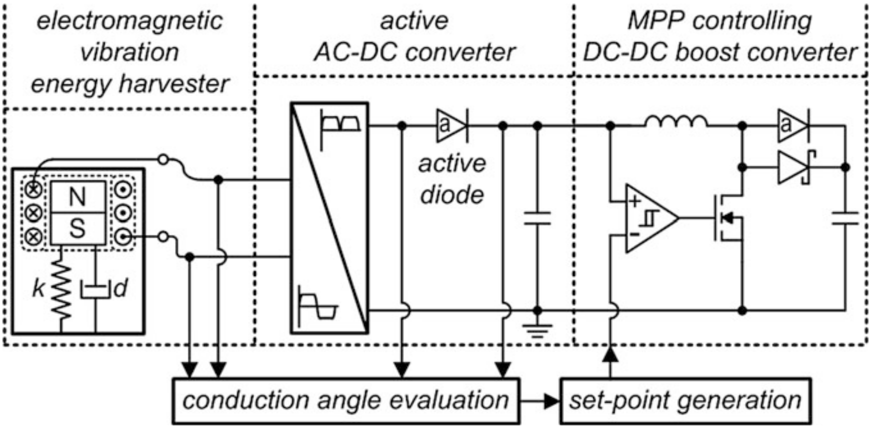

electromagnetic vibration energy harvester
active AC-DC converter
MPP controlling DC-DC boost converter
N
S
k
d
a
active diode
a
conduction angle evaluation
set-point generation

Fig. 19.10 Overview of the conduction angle controlled MPPT

piezoelectric energy harvesters, is described in the following.[2] It employs an inductive duty-cycle-adapted DC-DC converter starting at 300 mV input voltage, and it achieves efficiencies from 46 to 70 % between 4 and 28 μW input power, respectively.

In the MPPT AC-DC conversion, a rectifier with smoothed DC output is applied, and the maximum power-point (MPP) is tracked. The interface IC presented by Shim et al. [45] is designed to interface with a piezoelectric vibration transducer, and it tracks the MPP by applying the fractional open-circuit voltage technique. Periodic disconnection of the energy harvester is required for the set-point generation of the MPPT control. This interface operates self-sufficiently between 33 μW and 10 mW input power over an output voltage range of 1–8 V. The MPP is tracked with up to 99.9 % accuracy, and up to 80 % efficiency is achieved. A self-starting implementation is proposed in [46]. This system tracks the MPP with up to 93 % accuracy applying an extra sensor for set-point generation, and it harnesses autonomously up to 72 % of the maximum harvester real power. The interface IC presented in [47] employs an integrated set-point method avoiding an extra sensor as well as a harvester disconnection (Fig. 19.10). Figure 19.11 shows the micrograph of the fabricated IC. A dedicated circuit measures the conduction angle of the implemented active AC-DC input stage. From the conduction angle, the set-point of an input-controlled boost converter is determined. Consequently, this converter accordingly controls the output voltage of the AC-DC stage. By means of the resulting closed loop control, the optimum conduction angle and hence the MPP are tracked. This interface IC allows cold start-up from a flat energy buffer when the harvester is able to deliver a minimum power of 4.7 μW. Measurements

[2] Under the assumption that any voltage or power specifications are met, the load-matching algorithm employed in the interface ASIC presented in [44] can be used for electromagnetic energy harvesters as well.

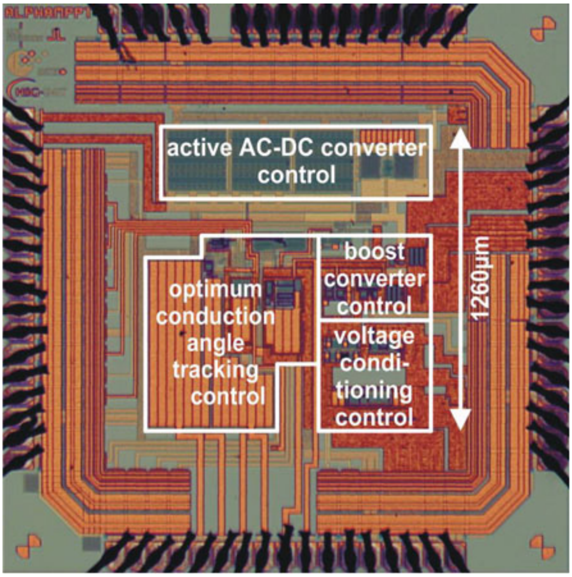

active AC-DC converter control
optimum conduction angle tracking control
boost converter control
voltage condi-tioning control
1260µm

Fig. 19.11 Micrograph of the MPPT IC prototype

demonstrate that over 90 % of the maximum harvester real power can be transferred into an energy storage element via the IC. Additionally, a voltage conditioning for the application is implemented. Table 19.2 summarizes all the interface circuits for electromagnetic energy harvesters.

19.2.2 DC Sources

19.2.2.1 Micro Fuel Cells

The preceding volume of CHIPS 2020 [48] presented the integration of micro fuel-cells (µFCs) in a CMOS process as long-term energy storage for energy-harvesting-powered systems. Recently, advances in integrated CMOS electronics as well as improvements of architecture and output power of the micro fuel-cells lead to an extended intelligent fuel-cell battery, allowing to electrically (re-)charge these fuel cells [49].

The possibility to adjust the interconnection of single µFCs to so-called fuel-cell cascades (FCCs) is presented in [50–52]. A system micrograph is shown in Fig. 19.12. The 42 µFCs can be combined flexibly in series or parallel in accordance to the actual needs of the electronic system being supplied. Three to seven µFCs can be combined in series while up to 14 FCCs are connected in parallel. The integrated voltage regulator powered by the FCCs fixes the output voltage to a digitally selectable value [50–52]. Choosing an FCC configuration that generates an open circuit voltage slightly above the system requirements, can increase the

Table 19.2 Overview of the interface circuits for electromagnetic energy harvesters

Ref.	Type	Process	Cold start-up capability	Minimum power demand rating	Maximum efficiency rating
[43]	Load matching using adaptive charge-pump	0.35 μm	N/A (operation needs 0.8 V buffer voltage)	15–38 μW (power losses)	48 % (harvesting efficiency, sensor not considered)
[44]	Load matching using duty-cycle adapted DC-DC converter	65 nm	Yes (no pre-charge needed, operation starts at 0.3 V buffer voltage)	4 μW (input power)	70 % (harvesting efficiency, overall)
[45]	MPPT driven AC/DC using fractional open-circuit voltage technique	0.35 μm	N/A (output voltage range: 1–8 V)	33 μW (input power)	80 % (just buck-boost converter)
[46]	MPPT driven AC/DC using extra sensor for set-point generation	0.35 μm	Yes (no pre-charge needed)	11 μW (P_{max} of harvester at a_{exp} = 1.6 m/s^2)	72 % (harvesting efficiency, sensor not considered)
[47]	MPPT driven AC/DC using conduction angle control	0.35 μm	Yes (no pre-charge needed)	4.7 μW (P_{max} of harvester at a_{exp} = 0.2 m/s^2)	90 % (harvesting efficiency, overall)

efficiency and thus minimize the hydrogen consumption, resulting in an extended system lifetime.

Not only was the electronic system improved, but also the power density of the fuel-cells increased to 1.68 mW/cm^2 [51]. The intelligent battery was finally used to power a sensor system [52] and a microcontroller [51]. In [49], the extension of such μFCs by a hydrolysis cell is suggested (Fig. 19.13). This hydrolysis cell electrically splits water into hydrogen and oxygen. The resulting hydrogen is stored, while the oxygen is released into the atmosphere. This means, the fuel cell reaction can be reversed to electrically (re-)charge each cell. In combination with an energy harvester with typically fluctuating output power, such rechargeable μFCs can serve as long-term energy storage to even out unpredictable power shortages. This combination will therefore create an uninterrupted power supply to build up reliable, energy self-sufficient sensor and actor systems.

19.2.2.2 Interface Circuits for Thermoelectric Generators

Thermoelectric generators harvest energy from ambient temperature-gradients and deliver DC power. In order to harvest even from low gradients, an energy-harvester

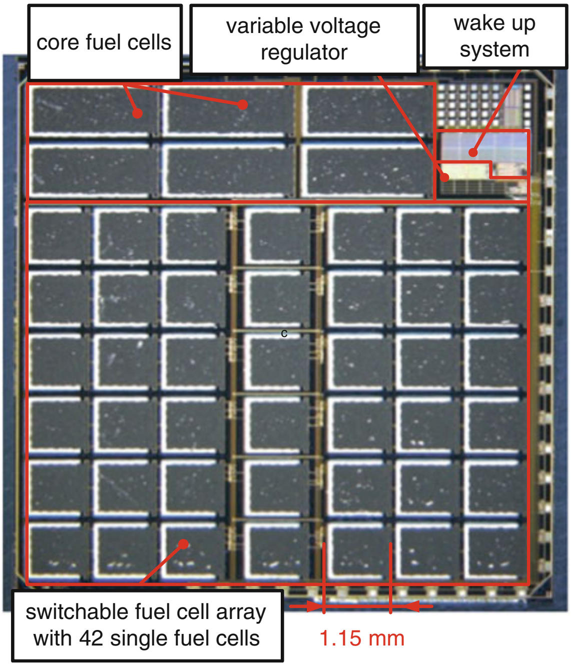

core fuel cells
variable voltage regulator
wake up system
c
switchable fuel cell array with 42 single fuel cells
1.15 mm

Fig. 19.12 Chip-integrated fuel cell devices [51]

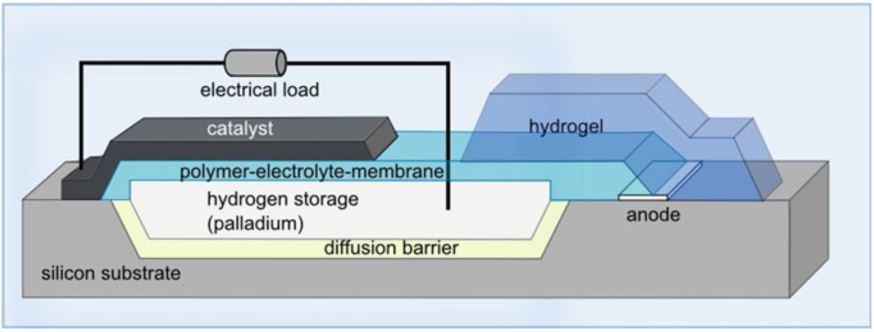


Fig. 19.13 Schematic set-up of fuel cell accumulator [49]. © 2012 SSI/MESAGO

interface has to operate at low input voltages and has to consume only little power. In addition, the interface has to arrange an optimal point of harvester load in order to extract as much power as possible. A capacitive power management with a 1.4 µA controller and a converter peak-efficiency of 82 % has been published by Doms et al. [53]. This thermoelectric generator interface enables to extract 58 % of the available harvester power. A 20 mV input inductive boost converter with

1.1 μW minimum quiescent power is presented in [54]. A battery-less thermoelectric energy harvesting interface circuit with 35 mV startup voltage, maximum power-point tracking (MPPT) and 58 % end-to-end efficiency is given in [55]. In contrast to [55], the IC detailed in [56] needs no support of a vibration-triggered mechanical switch for start-up. Instead, a positive feedback and white noise are applied to allow start-up from 40 mV utilizing a miniaturized transformer. Additionally, 61 % of the available maximum power can be harvested by means of an implemented MPPT. The charge-pump-based interface IC detailed in [57] proposes a fully electrical 50 mV start-up mechanism. The system achieves a peak efficiency of 73 %. In [58], an inductive boost regulator, which allows start-up from as low as 12 mV input while biased by an external 1 V battery, is presented. This design achieves up to 82 % efficiency and can regulate the output from 0.66 to 3.3 V. In standby, 3.5 μA are typically consumed, and the system delivers an output power of 6 μW at an ~2 °C temperature difference. The IC presented in [59] is designed to interface with high resistive power sources such as thin-film thermoelectric generators. Autonomous start-up is enabled at 5.8 μW input power and 21 mV input voltage. At an input voltage of 35 mV, the system can even start-up self-sufficiently at a low input power of 1.3 μW. A complete energy harvesting system is shown in [60, 61]. The system uses an off-the-shelf thermoelectric generator connected to an energy-conditioning IC, which accomplishes MPPT. Moreover, the interface IC autonomously controls the powering of a wireless sensor node via the harvested energy. Table 19.3 summarizes all the interface circuits for thermoelectric generators.

19.2.2.3 Interface Circuits for Solar Cells

Photovoltaic energy harvesting allows transducing ambient light into DC power. A power management with MPPT for solar and thermoelectric energy harvesting is presented in [62]. This IC with battery management enables cold start from 5 μW input power and 330 mV. After cold start, the IC can continue harvesting down to an input voltage of 80 mV, and it consumes only 330 nA quiescent current [63]. In [64], a 5 μW to 10 mW input-power-range inductive boost converter for indoor photovoltaic energy harvesting with integrated MPPT algorithm is presented. This implementation allows the extraction of 70 % of the maximum harvester power. The regulated charge pump with integrated optimum power-point tracking for indoor solar energy harvesting, published in [65], allows harvesting 86 % of the maximum available power. The implemented controller dissipates between 450 and 850 nW. The interface circuit presented in [66] uses a time-based power monitor in order to track the maximum point with an accuracy of up to 96 %. This MPPT implementation dynamically readjusts after solar-cell shading. A 400 nW single-inductor dual-input-tri-output DC-DC buck-boost converter with MPPT for

Table 19.3 Overview of the interface circuits for thermoelectric generators

Ref.	Type	Process	Cold start-up capability	Min. power demand rating	Max. efficiency rating
[53]	Power management based on Dickson charge pump	0.35 µm	No (chip starts operation at 0.76 V input voltage, output buffer pre-charged to 2 V)	1.4 µA * 1.508 V = 2.1 µW (power consumption)	82 % (charge pump) 58 % (overall)
[54]	Inductive boost converter	0.13 µm	No (20–250 mV input voltage, output buffer pre-charged to 1 V)	1.1 µW (power losses)	75 % (overall)
[55]	DC-DC converter with MPPT and mechanically assisted startup	0.35 µm	Yes (no pre-charge needed, chip starts operation at 35 mV input voltage)	N/A	58 % (overall)
[56]	Transformer-based boost converter with MPPT	0.13 µm	Yes (no pre-charge needed, chip starts operation at 40 mV input voltage)	N/A	61 % (overall)
[57]	Inductive boost converter	65 nm	Yes (no pre-charge needed, chip starts operation at 50 mV input voltage)	121 µW (total power losses)	73 % (overall)
[58]	Inductive boost regulator	0.13 µm	Yes (chip self-starts operation at 380 mV input voltage. With 1 V battery attached, chip starts operation at 12 mV)	6.5 µA * 1 V = 6.5 µW (standby power losses)	82 % (no definition given, value is compared to overall efficiencies of other papers)
[59]	Transformer-based boost converter with MPPT	0.13 µm	Yes (no pre-charge needed, chip starts operation at 21 mV)	1.3 µW (at 35 mV input voltage)	74 % (overall)

Table 19.4 Overview of the interface circuits for solar cells

Ref.	Type	Process	Cold start-up capability	Min. power demand rating	Max. efficiency rating
[62]	Power management with MPPT for solar and thermoelectric energy harvesting	N/A	Yes (no pre-charge needed, chip starts operation at 330 mV)	5 μW	80 % (no definition given)
[64]	Inductive boost converter with MPPT	0.25 μm	Yes (no pre-charge needed, chip starts operation at 0.5 V)	2.4 μW (power consumption of control circuit)	70 % (overall)
[65]	Regulated charge pump with MPPT	0.35 μm	No (input voltage range 1–2.7 V)	850 nW (total power consumption of controller)	86 % (overall)
[66]	Boost (solar and thermal) and buck-boost (piezo) converter using shared inductor with MPPT	0.35 μm	No (input voltage ranges 0.15–0.75 V for solar, 20–160 mV for thermal, 1.5–5 V for piezo)	N/A	83 % (solar) 58 % (thermal) 79 % (piezo) (all values are overall efficiencies with inductor sharing)
[67]	Single-inductor dual-input-tri-output buck-boost-converter with MPPT	0.18 μm	No	0.4 μW (power consumption of control circuit)	83 % (overall)
[68]	Self-oscillating switched-capacitor DC-DC voltage doubler	0.18 μm	Yes (no pre-charge needed, chip starts operation at 140 mV)	6 nW	50 % (overall)

indoor photovoltaic energy harvesting is presented in [67]. A peak conversion-efficiency of 83 % is achieved, and the control circuit consumes 400 nW. The charge-pump based interface IC presented in [68] maintains 50 % end-to-end efficiency when harvesting from a 0.84 mm^2 solar cell. The idle power consumption is below 3 nW, and a minimum input power of 6 nW for self-startup is required. The system sustains harvesting down to 1.7 nW input power. In [69], design aspects of on-chip photovoltaic energy conversion, voltage boosting and storage in bulk-CMOS for indoor applications are investigated. Table 19.4 summarizes all the interface circuits for solar cells.

19.2.3 Ultra-Low-Voltage Control Circuits

19.2.3.1 Analog

Power management circuits are needed for optimizing the ambient energy extraction and for the conditioning of this optimal harvested power in order to efficiently charge an energy-storage element or to drive an application such as a wireless sensor-node [40]. Such power management circuits typically need analog implementations that have to operate under variable supply-voltages at low power levels being in the microwatt to milliwatt range [70–78].

Sub-threshold design allows ultra-low power, low-voltage analog circuits [70, 79]. By decreasing the number of stacked transistors, the supply-voltage demand can be further reduced as successfully demonstrated by active rectifiers using stacks with only two transistors [42, 43]. In the power-management circuit presented in [46], this circuit-design technique is applied for a voltage-doubling rectifier, a boost-converter-based maximum power-point tracker, buffer monitoring, and voltage stabilization.

Another possibility to decrease the required operation voltage is forward body-biasing of a MOSFET, which reduces the MOSFET threshold-voltage. The charge pump presented in [71] uses sub-threshold operation as well as body-biasing. The system achieves a pump efficiency of 89 % and operates at 320 mV. In the charge pump shown in [72], the body-biasing is only applied during switching and not in the idle state avoiding reverse currents. This circuit handles a minimum input voltage of 150 mV and reaches a maximum efficiency of 72.5 %. The body terminals of PMOS transistors are used as comparator inputs in the active rectifier presented in [31] avoiding a tail-current source. So, the number of stacked transistors is reduced to two, enabling operation down to 380 mV and voltage-efficiencies over 90 %, whereas the power consumption of the sub-threshold bulk-input comparator is only 266 nW at 500 mV. Instead of applying body-biasing, the threshold-voltage can be lowered by fixed-charge injection into the dielectric of MOSFET gates. In [73], this method minimizes the operating voltage of an oscillator that is used in a charge pump. This charge pump is designed to assist an inductive DC-DC boost converter during start-up allowing operation from an input voltage of 95 mV. In addition to the technique of forward-biasing, floating gates can be applied to reduce the MOSFET threshold-voltages, allowing highly efficient and high frequency rectifiers [74].

Besides straight low-power circuit-design methods, system-level optimization is a major design objective for efficient energy-harvesting system realization. The energy-harvesting charger IC presented in [62] duty-cycles and time-multiplexes several circuit-blocks for power saving, resulting in a low quiescent current of 330 nA. The switched-capacitor voltage-doubler, presented in [68], is self-oscillating, and thus it overcomes the need for an extra clock requiring only 170 pW idle power. Input-controlled DC-DC converters, which are operated in the discontinuous conduction mode, allow efficient voltage conversion even under ultra-low power conditions [75]. This conduction mode allows input-resistance adaptation as demonstrated by the 1.1 nW energy harvesting system presented in [76], which consumes only 544 pW quiescent power. A hysteretic input-voltage controlled DC-DC converter allows maximum power-point tracking, avoids a system clock, is free from fast-scale instability, and its switching frequency scales with the transferred power, which makes it an attractive candidate for ultra-low power applications [46, 80]. Pulsed energy-extraction techniques minimize the number of DC-DC converter switching actions, which allows the reduction of power consumption [30, 34, 35, 77, 78]. The energy-conditioning IC presented in [46] allows harvesting from ambient AC sources as well as from ambient DC sources [60, 61]. Such multi-purpose interface architectures can ease the design of an energy harvesting system because of the given adaptability. The amount of harvested power can be enhanced by extracting energy from several ambient energy-sources by means of a multi-input interface circuit [66].

19.2.3.2 Digital

In numerous energy harvesters, the delivered output voltage is directly correlated to the input excitation (e.g. thermoelectric or solar harvesters), which makes the supply-voltage minimization for the interfacing circuits critical for operation at low excitations. This applies to both digital and analog electronics, but there are various cases when digital blocks are the limiting factor, especially when reliable clock-sources, e.g. for start-up charge pumps as in [73], are required. This section therefore discusses supply-voltage minimization for digital circuit blocks.

Supply-voltage reduction is limited by a corresponding degradation of the on-to-off current ratio of the transistors. The transistor on-currents, which drive the output of a CMOS gate to a high or low level, thus become similar to the leakage currents flowing through the off-transistors. The result is a degradation of output levels until they no longer represent a valid logic level. Transistor variability further aggravates this problem: The drive-strengths of the pull-up and pull-down devices should be equalized for minimum supply-voltage operation, but transistor variability randomly unbalances the drive strengths.

Several approaches exist for minimizing the required supply voltage. First, to reduce the impact of local variability, transistors are sized larger than necessary for optimum speed. Concepts reaching further mostly aim at additionally alleviating the impact of global variability by equalizing PMOS and NMOS drive strengths. This is e.g. possible by the application of body-biasing, as demonstrated in [81] for a system operational to supply voltages down to 85 mV. The applicability of this concept for energy-harvesting applications though is limited by the need for biasing voltages, which are considerably higher, motivating the development of post-silicon programming approaches, where the threshold voltage of the PMOS transistors is altered permanently by an application of large gate-body voltages [73].

An alternative approach, improving both on-to-off current ratio and susceptibility to variability by a reduction of current from the output node, is the use of a

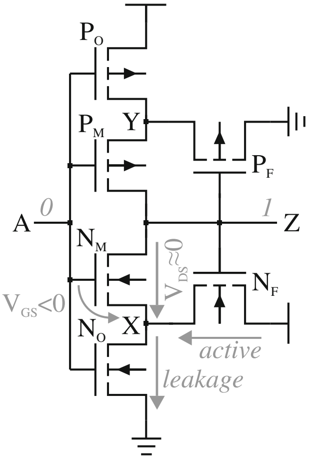


Fig. 19.14 Schmitt-Trigger (ST) inverter with illustration of leakage suppression from critical output node for input 0/output 1 case

Schmitt-Trigger (ST) circuit topology [82]. Figure 19.14 shows an ST inverter (ST topology can be applied to arbitrary gates). Example, for input-low/output-high in Fig. 19.14, the pull-down block is off, and its leakage from node Z is critical for the degradation of the output level. In the ST topology, the feedback transistor N_F is on and pulls the middle node X to a high potential, forcing the middle transistor N_M to a very low drain-source and, more importantly, negative gate-source potential. N_M therefore is switched off very effectively, suppressing leakage from the critical output node. The improved off-behavior furthermore mitigates the effect of drive-strength misbalance, thereby also reducing the impact of global process variability. Full operability of ST circuits has been demonstrated at supply-voltages of 62 mV, the lowest value for CMOS circuits reported to date.

It is interesting to observe that new transistor concepts with multi-gate devices allow for reliability-improving feedback structures (similar to ST) with much simpler circuit topologies, as e.g. reported in [83], making the use of such concepts even interesting for general low-voltage design.

19.3 Conclusion

In this chapter, possible application scenarios making use of energy harvesting have been highlighted. It has been shown that in the area of wearable devices and condition monitoring, significant progress has been achieved in the past, pushing this topic towards practical applications in the industrial and consumer environment. Condition monitoring and smart metering are possible applications where batteries may be omitted in the future, resulting in reduced maintenance effort. Similarly, the power harvested from human walking could supply wearable devices like MP3 players or health tracking equipment.

In research, little attention has been paid to the electronic circuitry interfacing the energy converters (or generators), although this is an essential part of the energy harvesting driven application. An efficient interface circuit might lead to a working application, even if the generator is providing minimal power. The outcome of this review is that interface circuits require careful design in order to fit the special characteristics of each conversion mechanism. Due to their low power consumption and high degree of design flexibility, CMOS integrated circuits are very well suited for this purpose. As shown in Tables 19.1, 19.2, 19.3 and 19.4, most ASIC designs are using the 0.35 μm or 0.18 μm CMOS processes, only few go to smaller processes. Obviously, these two processes offer the best tradeoff between area, speed, power consumption, robustness and flexibility in high and low voltage options. Most designs consume power in the microwatt range, some even in the nanowatt range, resulting in very high efficiency.

Up to now, the majority of research has focused on efficiently interfacing one dedicated conversion mechanism. The vision for the future is to have a robust interface fitting a large variety of different energy-conversion principles, and enabling harvesting from multiple ambient sources (vibration, thermal, solar energy etc.) at the same time. Ideally, the interface circuits are tolerating a wide range of input powers and voltages from ultra-low to high levels, in order to be able to react adequately to alternating conditions. This would be an essential progress for highly reliable power supply of safety-sensitive systems, since this increases the probability that at least one of these ambient energy sources is present.

By means of sophisticated circuit techniques, it is possible to lower the supply voltage down to 320 mV for analog and 62 mV for digital circuits, which is extremely beneficial in environments where only a small power budget is available. Electronics made of such circuits might be driven directly by the output of thermoelectric or solar generators, without requiring additional boost converters or charge pumps, which always reduce efficiency.

References

1. Manoli, Y., Hehn, T., et al.: Energy harvesting and chip autonomy, chapter 19. In: Hoefflinger, B. (ed.) CHIPS 2020—A Guide to the Future of Nanoelectronics, pp. 393–420. Springer, Berlin (2012)
2. Leonov, V.: Thermoelectric energy harvesting of human body heat for wearable sensors. IEEE Sensors J. **13**(6), 2284–2291 (2013)
3. Chen, T., Qiu, L., et al.: Novel solar cells in a wire format. Chem. Soc. Rev. **42**(12), 5031 (2013)
4. Bedeloglu, A.C., Demir, A., et al.: A photovoltaic fiber design for smart textiles. Text. Res. J. **80**(11), 1065–1074 (2010)
5. Zeng, W., Tao, X.-M., et al.: Highly durable all-fiber nanogenerator for mechanical energy harvesting. Energy Environ. Sci. **6**(9), 2631 (2013)
6. Qin, Y., Wang, X., et al.: Microfibre-nanowire hybrid structure for energy scavenging. Nature **451**(7180), 809–813 (2008)

7. Niu, P., Chapman, P., et al.: Evaluation of motions and actuation methods for biomechanical energy harvesting. In: 2004 IEEE 35th Annual Power Electronics Specialists Conference, pp. 2100–2106
8. Kornbluh, R., Pelrine, R., et al.: Electroelastomers: applications of dielectric elastomer transducers for actuation, generation and smart structures. In: SPIE's 9th Annual International Symposium on Smart Structures and Materials, vol. 4698, pp. 254–270 (2002)
9. Kymissis, J., Kendall, C., et al.: Parasitic power harvesting in shoes. In: Digest of Papers of Second International Symposium on Wearable Computers, pp. 132–139 (1998)
10. Bai, P., Zhu, G., et al.: Integrated multilayered triboelectric nanogenerator for harvesting biomechanical energy from human motions. ACS Nano **7**(4), 3713–3719 (2013)
11. Zhu, G., Bai, P., et al.: Power-generating shoe insole based on triboelectric nanogenerators for self-powered consumer electronics. Nano Energy **2**(5), 688–692 (2013)
12. Ylli, K., Hoffmann, D., et al.: Energy harvesting from human motion: exploiting swing and shock excitations. Smart Mater. Struct. **24**(2), 025029 (2015)
13. Ylli, K., Hoffmann, D., et al.: Design, fabrication and characterization of an inductive human motion energy harvester for application in shoes. In: Proceedings of PowerMEMS 2013, London (UK), Journal of Physics: Conference Series, vol. 476, p. 12012
14. Carroll, D., Duffy, M.: Modelling, design, and testing of an electromagnetic power generator optimized for integration into shoes. In: Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, vol. 226, no. 2, pp. 256–270 (2012)
15. Ylli, K., Hoffmann, D., et al.: Human motion energy harvesting for AAL applications. In: Proceedings of PowerMEMS 2014, Awaji (Japan), Journal of Physics: Conference Series, vol. 557, p. 012024 (2014)
16. Weddell, S., Zhu, D., et al.: A practical self-powered sensor system with a tunable vibration energy harvester. In: Proceedings of PowerMEMS 2012, Atlanta, USA, pp. 105–108
17. Hoffmann, D., Willmann, A., et al.: Tunable vibration energy harvester for condition monitoring of maritime gearboxes. J. Phys. Conf. Ser. **557**, 012099 (2014)
18. Finkenzeller, K.: RFID Handbook. Wiley, New York (2003)
19. Karthaus, U., Fischer, M.: Fully integrated passive UHF RFID transponder IC with 16.7-μW minimum RF input power. IEEE J. Solid-State Circ **38**(10), 1602–1608 (2003)
20. Lazzi, G.: Thermal effects of bioimplants. Eng. Med. Biol. Mag. **24**(5), 75–81 (2005)
21. Opie, N.L., Burkitt A.N., et al.: Thermal heating of a retinal prosthesis: thermal model and in-vitro study, Engineering in Medicine and Biology Society (EMBC). In: 2010 Annual International Conference of the IEEE, pp. 1597–1600
22. Peters, C., Kessling, O., et al.: CMOS integrated highly efficient full wave rectifier. In: IEEE International Symposium on Circuits and Systems (ISCAS), pp. 2415–2418 (2007)
23. Ghovanloo, M., Najafi, K.: Fully integrated wideband high-current rectifiers for inductively powered devices. IEEE J. Solid-State Circ. **39**(11), 1976–1984 (2004)
24. Kuhl, M., Gieschke, P., et al.: A wireless stress mapping system for orthodontic brackets using CMOS integrated sensors. IEEE J. Solid-State Circ. **48**(9), 2191–2202 (2013)
25. Peters, C., Henrici, F., et al.: High-bandwidth floating gate CMOS rectifiers with reduced voltage drop. In: IEEE International Symposium on Circuits and Systems (ISCAS), pp. 2598–2601 (2008)
26. Kiani, M., Ghovanloo, M.: An RFID-based closed-loop wireless power transmission system for biomedical applications. IEEE Trans. Circ. Syst. II Express Briefs **57**(4), 260–264 (2010)
27. Silay, K.M., Dehollain, C., et al.: A closed-loop remote powering link for wireless cortical implants. IEEE Sens. J. **13**(9), 3226–3235 (2013)
28. O'Driscol, S.D.: A mm-sized implantable power receiver with adaptive matching. In Proceedings of IEEE Sensors, pp. 83–88 (2010)
29. Kazanc, O., Maloberti, F., et al.: High-Q adaptive matching network for remote powering of UHF RFIDs and wireless sensor systems. In: IEEE Topical Conference on Wireless Sensors and Sensor Networks (WiSNet), pp. 10–12 (2013)

30. Hehn, T., Manoli, Y.: CMOS Circuits for Piezoelectric Energy Harvesters: Efficient Power Extraction, Interface Modeling and Loss Analysis. Springer, Berlin (2014)
31. Peters, C., Handwerker, J., et al.: A Sub-500 mV highly efficient active rectifier for energy harvesting applications. IEEE Trans. Circ. Syst. I: Regul. Pap. **58**(7), 1542–1550 (2011)
32. Colomer-Farrarons, J., Miribel Catala, P., et al.: A 60 μW low-power low-voltage power management unit for a self-powered system based on low-cost piezoelectric powering generators. In: Proceedings of European Solid-State Circuits Conference (ESSCIRC), pp. 280–283 (2009)
33. Rao, Y., Cheng, S., et al.: A fully self-sufficient energy harvesting system for human movements. In: Proceedings of PowerMEMS 2012, Atlanta, GA, USA, pp. 101–104
34. Hehn, T., Hagedorn, F., et al.: A fully autonomous integrated interface circuit for piezoelectric harvesters. IEEE J. Solid State Circ. **47**(9), 2185–2198 (2012)
35. Lefeuvre, E., et al.: Piezoelectric energy harvesting device optimization by synchronous electric charge extraction. J. Intell. Mater. Syst. Struct. **16**(10), 865–876 (2005)
36. Dallago, E., Miatton, D., et al.: Electronic interface for piezoelectric energy scavenging system. In: 34th European Solid-State Circuits Conference (ESSCIRC), pp. 402–405 (2008)
37. Gasnier, P., Willemin, J., et al.: An autonomous piezoelectric energy harvesting IC based on a synchronous multi-shots technique. In: Proceedings of European Solid-State Circuits Conference (ESSCIRC), pp. 399–402 (2013)
38. Aktakka, E.E., Peterson, R.L., et al.: A self-supplied inertial piezoelectric energy harvester with power-management IC. In: IEEE International Solid-State Circuits Conference (ISSCC), Digital Technical Papers, pp. 120–121 (2011)
39. Ishida, K., Tsung, C.H., et al.: Insole pedometer with piezoelectric energy harvester and 2 V organic circuits. IEEE J. Solid State Circ. **48**(1), 255–264 (2013)
40. Manoli, Y.: Energy harvesting—from devices to systems. In: Proceedings of IEEE European Solid-State Circuits Conference (ESSCIRC), pp. 27–36, Sept 2010
41. Spreemann, D., Manoli, Y.: Electromagnetic Vibration Energy Harvesting Devices: Architectures, Design, Modeling and Optimization. Springer, Berlin (2012)
42. van Liempd, C., Stanzione, S., et al.: A 1 μW to 1 mW energy aware interface ic for piezoelectric harvesting with 40 nA quiescent current and zero-bias active rectifiers. In: IEEE International Solid-State Circuits Conference (ISSCC), Digital Technical Papers, pp. 76–77, Feb 2013
43. Maurath, D., Becker, P.F., et al.: Efficient energy harvesting with electromagnetic energy transducers using active low-voltage rectification and maximum power point tracking. IEEE J. Solid State Circ. **47**(6), 1369–1380
44. Gao, Y., Made, D.I., et al.: An energy-autonomous piezoelectric energy harvester interface circuit with 0.3 V startup voltage. In: Proceedings of IEEE Asian Solid-State Circuits Conference (A-SSCC), pp. 445–448, Nov 2013
45. Shim, M., Kim, J., et al.: Self-powered 30 μW-to-10 mW piezoelectric energy-harvesting system with 9.09 ms/V maximum power point tracking time. In: IEEE International Solid-State Circuits Conference (ISSCC), Digital Technical Papers, pp. 406–407, Feb 2014
46. Leicht, J., Maurath, D., et al.: Autonomous and self-starting efficient micro energy harvesting interface with adaptive MPPT, buffer monitoring, and voltage monitoring. In: Proceedings of IEEE European Solid-State Circuits Conference (ESSCIRC), pp. 101–104, Sept 2012
47. Leicht, J., Amayreh, M., et al.: Electromagnetic vibration energy harvester interface IC with conduction-angle-controlled maximum-power-point tracking and harvesting efficiencies of up to 90 %. In: IEEE International Solid-State Circuits Conference (ISSCC), Digital Technical Papers, pp. 368–370, Feb 2015
48. Hoefflinger, B.: The future of eight chip technologies, chapter 3. In: Hoefflinger, B. (ed.) CHIPS 2020—A Guide to the Future of Nanoelectronics, pp. 37–93. Springer, Berlin (2012)
49. Zimmermann, D., Becker, J., et al.: On-chip micro fuel cells as power supply for smart microsystems. In: Proceedings of International Conference and Exhibition on Integration Issues of Miniaturized Systems (SSI) (2012)

50. Moranz, C., Kuhl, M., et al.: A digitally adjusted power supply for systems-on-chip based on CMOS integrated fuel cells. In: Proceedings of PowerMEMS 2011, pp. 86–89 (2011)
51. Moranz, C., Kuhl, M., et al.: CMOS integrierte Spannungsversorgung basierend auf Mikro-Brennstoffzellen. In: Proc. Mikrosystemtechnik-Kongress, pp. 275–278 (2013)
52. Zimmermann, D., Freund, I., et al.: Rechargeable micro fuel cells as power supply for smart microsystems. In: Proceedings of International Conference and Exhibition on Integration Issues of Miniaturized Systems (SSI) (2013)
53. Doms, I., Merken, P., et al.: Capacitive power management circuit for micropower thermoelectric generators with a 1.4 μA controller. IEEE J. Solid-State Circ. **44**(10), 2824–2833 (2009)
54. Carlson, E.J., Strunz, K., et al.: A 20 mV input boost converter with efficient digital control for thermoelectric energy harvesting. IEEE J. Solid-State Circ. **45**(4), 741–750 (2010)
55. Ramadass, Y.K., Chandrakasan, A.P.: A battery-less thermoelectric energy harvesting interface circuit with 35 mV startup voltage. IEEE J. Solid-State Circ. **46**(1), 333–341 (2011)
56. Im, J.-P., Wang, S.-W., et al.: A 40 mV transformer-reuse self-startup boost converter with MPPT control for thermoelectric energy harvesting. IEEE J. Solid-State Circ. **47**(12), 3055–3067 (2012)
57. Weng, P.-S., Tang, H.-Y., et al.: 50 mV-input batteryless boost converter for thermal energy harvesting. IEEE J. Solid-State Circ. **48**(4), 1031–1041 (2013)
58. Ahmed, K.Z., Mukhopadhyay, S.: A wide conversion ratio, extended input 3.5-μA boost regulator with 82 % efficiency for low-voltage energy harvesting. IEEE Trans. Power Electron. **29**(9), 4776–4786 (2014)
59. The, Y.-K., Mok, P.K.T.: Design of transformer-based boost converter for high internal resistance energy harvesting sources with 21 mV self-startup voltage and 74 % power efficiency. IEEE J. Solid-State Circ. **49**(11), 2694–2704 (2014)
60. Leicht, J., Heilmann, P., et al.: Thermoelectric energy harvesting system for demonstrating autonomous operation of a wireless sensor node enabled by a multipurpose interface. J. Phys. Conf. Ser. **467**, 1–5 (2013)
61. Leicht, J., Heilmann, P., et al.: Wireless anti-theft alarm system for automobiles based on thermoelectric energy harvesting powered glass break detection. In: Proceedings of 7th VDE GMM-Workshop 2014, pp. 74–77
62. Kadirvel, K., Ramadass, Y., et al.: A 330 nA energy-harvesting charger with battery management for solar and thermoelectric energy harvesting. In: IEEE International Solid-State Circuits Conference (ISSCC), Digital Technical Papers, pp. 106–107, Feb 2012
63. Texas Instruments: bq25504—Ultra Low Power Boost Converter with Battery Management for Energy Harvester Applications, Rev. A, Oct. 2011, Revised Sept 2012. http://www.ti.com/product/bq25504. Accessed December 2014
64. Qiu, Y., Van Liempd, C., et al.: 5 μW-to-10 mW input power range inductive boost converter for indoor photovoltaic energy harvesting with integrated maximum power point tracking algorithm. In: IEEE International Solid-State Circuits Conference (ISSCC), Digital Technical Papers, pp. 118–119, Feb 2011
65. Kim, J., Kim, C.: A regulated charge pump with a low-power integrated optimum power point tracking algorithm for indoor solar energy harvesting. IEEE Trans. Circuits Syst. II Exp. Briefs **58**(12), 802–806 (2011)
66. Bandyopadhyay, S., Chandrakasan, A.P.: Platform architecture for solar, thermal, and vibration energy combining with MPPT and single inductor. IEEE J. Solid-State Circ. **47**(9), 2199–2215 (2012)
67. Chew, K.W.R., Sun, Z.: A 400 nW single-inductor dual-input-tri-output DC-DC buck-boost converter with maximum power point tracking for indoor photovoltaic energy harvesting. In: IEEE International Solid-State Circuits Conference (ISSCC), Digital Technical Papers, pp. 68–69, Feb 2013
68. Jung, W., Oh, S., et al.: An ultra-low power fully integrated energy harvester based on self-oscillating switched-capacitor voltage doubler. IEEE J. Solid-State Circ. **49**(12), 2800–2811 (2014)

69. Çilingiroğlu, U., Tar, B., et al.: On-chip photovoltaic energy conversion in bulk-CMOS for indoor applications. IEEE Trans. Circuits Syst. I Reg. Pap. **61**(8), 2491–2504 (2014)
70. Maurath, D., Manoli, Y.:, CMOS Circuits for Electromagnetic Vibration Transducers, Springer, Berlin (2014)
71. Peng, H., Tang, N., et al.: CMOS startup charge pump with body bias and backward control for energy harvesting step-up converters. IEEE Trans. Circuits Syst. I Reg. Pap. **61**(6), 1618–1628 (2014)
72. Kim, J., Amayreh, M., et al.: A 0.15 V-input energy-harvesting charge pump with switching body biasing and adaptive dead-time for efficiency improvement. In: IEEE International Solid-State Circuits Conference (ISSCC), Digital Technical Papers, pp. 394–395, Feb 2014
73. Chen, P.-H., Ishida, K., et al.: Startup techniques for 95 mV step-up converter by capacitor pass-on scheme and VTH tuned oscillator with fixed charge programming. IEEE J. Solid-State Circ. **47**(5), 1252–1260 (2012)
74. Peters, C., Henrici, F., et al.: High-bandwith floating gate CMOS rectifiers with reduced voltage drop. In: Proceedings of IEEE International Symposium on Circuits Systems (ISCAS), pp. 2598–2601 (2008)
75. Stanzione, S., van Liempd, C., et al.: A high voltage self-biased integrated DC-DC buck converter with fully analog MPPT algorithm for electrostatic energy harvesters. IEEE J. Solid-State Circ. **48**(12), 3002–3010 (2013)
76. Bandyopadhyay, S., Mercier, P.P., et al.: A 1.1 nW energy-harvesting system with 544 pW quiescent power for next-generation implants. IEEE J. Solid-State Circ. **49**(12), 2812–2824 (2014)
77. Aktakka, E.E., Najafi, K.: A micro inertial energy harvesting platform with self-supplied power management circuit for autonomous wireless sensor nodes. IEEE J. Solid-State Circ. **49** (9), 2185–2198 (2014)
78. Kwon, D., Rincón-Mora, G.A.: A Single-Inductor 0.35 µm CMOS Energy-Investing Piezoelectric Harvester. IEEE J. Solid-State Circ. **49**(10), 2277–2291 (2014)
79. Chandrakasan, A.: Sub-threshold Design for Ultra Low-Power Systems. Springer, Berlin (2006)
80. Redl, R., Sun, J.: Ripple-based control of switching regulators—an overview. IEEE Trans. Power Electron. **24**(12), 2669–2680 (2009)
81. Hwang, M.-E., Roy, K.: ABRM: adaptive beta-ratio modulation for process-tolerant ultradynamic voltage scaling. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **18**(2), 281–290 (2010)
82. Lotze, N., Manoli, Y.: A 62 mV 0.13 um CMOS standard-cell-based design technique using schmitt-trigger logic, solid-state circuits. IEEE J. Solid-State Circ. **47**(1), 47–60 (2012)
83. Hsieh, C.-Y., Fan, M.-L., et al.: Independently-controlled-gate FinFET Schmitt trigger sub-threshold SRAMs. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **20**(7), 1201–1210 (2012)

Chapter 20
2020 and Beyond

Bernd Hoefflinger

Abstract Nanoelectronics is driven by the exponential growth of the Internet. Economists expect an acceleration into the 2020s. However, we find increasing evidence that progress in the energy efficiency of large-scale nanoelectronics is stalling and that the supply of electric power for the Internet would reach one-third of the total global generation by the year 2020 causing an electric-power singularity of major black-outs and a breakdown of the Internet. The Internet service might be charged to regulate this problem. On the other hand, we have pointed out in all chapters of this Volume 2 of CHIPS 2020 that we have great potential to invest into a successful 2020s decade with critical efforts in low-voltage CMOS electronics, monolithic and heterogeneous 3D integration, real-data processing, communication and storage, human-visual-system inspired video, reliable neuro-processing and creative harvesting of enough energy for autonomous nano-chip systems.

20.1 Chip Market Forecasts for 2020

In Chap. 6 of [1], we projected the 2020 chip market to reach between 450 and 600 Billion$. IC Insights [2] puts its 2015 forecast for 2020 at 450 Bio.$, corresponding to the lower aggregate growth of only 4 %/year. Chips along the nano-roadmap have lost their leading role in promoting the growth of the global economy. We saw in Chap. 15 that the "intelligent" MEMS systems, which are often classified as "More-than-Moore", perform significantly better with CAGR's of 17.8 %/year from 2011 to 2016 in their consumer segment, with the potential of reaching 10 % of the electronic chips before 2020. Innovative, heterogeneous "Systems-on-Chips" (SOC's) carry the hopes for more growth.

B. Hoefflinger (✉)
5 Leonberger Strasse, 71063 Sindelfingen, Germany
e-mail: bhoefflinger@t-online.de

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_20

Reference [3] extrapolates the chip market forecast to 2023 with accelerating growth rates to beyond 9 %/year, exceeding 1 Trillion (10^{12}) $. This optimism is driven by speculation about an enormous growth of the Internet-of-Things (IOT), which would mean a continuing exponential expansion of data rates on the Internet. How plausible is the traffic in 2020, for example, of 10 TB/s of video data? With 500 kB per image or frame, it would mean just 20 Mio. images or frames per second worldwide. This data explosion is not unlikely. But how about the supply of the required electric power?

20.2 The Electric-Power Singularity of 2020

Can we manage this data explosion? In Chap. 12, we developed a performance and electric-power model for the Internet, based on this chapter in "CHIPS 2020" of 2012 [4], on the Berkeley study [5], on the CISCO study [6], and on the computational performance in Chap. 10. The outcome for 2020, as shown in Table 12.1 and in Fig. 20.1, would be a total electric power consumption of 1 TW, 30 % of the total global electric power generation estimate. And the 1 TW would be needed just for the electronic operations (the wall-plug power) and the embodied energy in the electronics. It does not consider the secondary needs for cooling, environment and

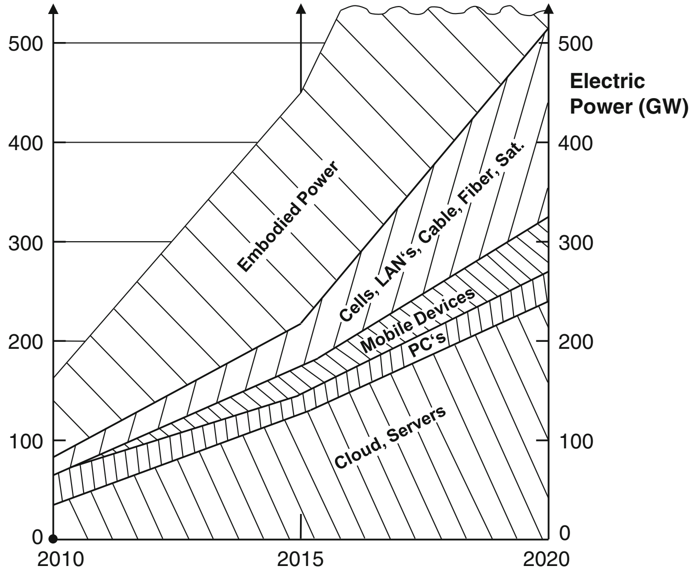


Fig. 20.1 Electric wall-plug power and embodied power for the mobile Internet

buildings. Such electric wall-plug capacity is not likely to be available from savings in other sectors like buildings, traffic, industry and commerce. The construction of new capacity of 1 TW (1000 plants at 1 GW each) within 5 years is not likely either. Therefore, we would face

- **A singularity in electric-energy resources in about 2020.**

A singularity in resources, historically, was introduced in the early 1900s by Adams, see [7], as the event, where the development of a civilization accelerates so much, that it hits a fundamental turning-point because of a lack of essential resources to continue that development, and where totally new strategies will be needed.

The electric-power singularity of the Internet would manifest itself by major black-outs and a breakdown of the Internet caused by global events related to information bubbles like a catastrophe or world-sports events.

This immanent electric-power singularity of the nanoelectronics-enabled Internet can be delayed or possibly avoided.

An obvious delay would be possible by charging Internet users for excessive use of the Internet beyond a certain data volume per month. This might limit, at first sight, the growth of business, but it would generate justified new business income to increase the investment into innovation and into renewable energy resources to handle the data.

A sustained delay would be achieved by leaving the percent/year improvement of the nanometer roadmap and by getting on the Femto-Joule energy road with ultra-low voltage, subthreshold CMOS and monolithic 3D integration.

To avoid the power singularity, 10× steps in nanoelectronics are needed. They are:

- Monolithic and Heterogeneous 3D Integration
- Low-Voltage, new Digital Computing
- New Video
- Reliable Neural Networks
- Fully **Autonomous Nanoelectronics**.

20.3 Monolithic and Heterogeneous 3D Integration

Based on the results in Chaps. 3, 5, 6, 9–12, the communication energy between all electronics functions no longer improves in 2D. The concerted development of 3D integration immediately doubles the energy efficiency, in other words, it halves the needed electric energy, and its sustained development can provide progress for two decades (Fig. 20.2).

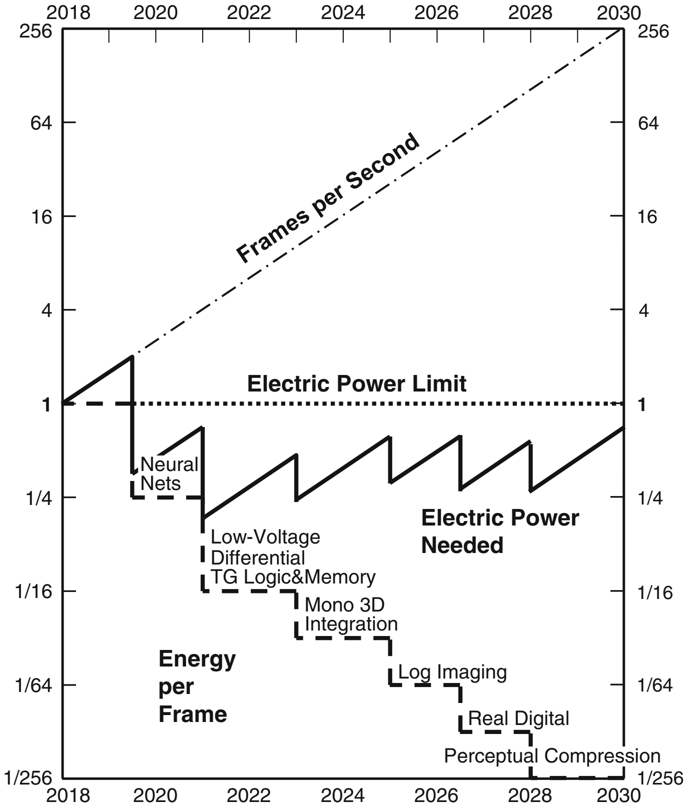


Fig. 20.2 Illustration of an energy roadmap towards 2030: With a fixed total electric power limit (probably 500 GW), six key innovations support the sustained functional growth of the mobile video Internet 200-times in a decade in terms of its video frames/s. This graph is a schematic illustration: There are correlations, but all six technologies have improvement potentials well beyond 2× each

20.4 Low-Voltage, New Digital Computing

The improvement of digital CMOS speed, energy and transistor efficiency with ultra-low voltage, differential transmission-gate (DTG) CMOS logic offers an order-of-magnitude improvement versus present standard full-CMOS logic (see Sect. 1.6). It took standard CMOS logic from 1970 to 1990 to pass NMOS logic. So ultra-low-voltage DTG CMOS has a long sustained career ahead.

On top of this circuit technique, digital processing needs a fundamental review. Binary data on or for our physical world need a relevant representation, guided by significance and accuracy, and math operations on them should be governed by these relevance criteria. It was shown in Chap. 12 that this can reduce the transistor count by an order-of-magnitude and storage by factors of 2–5.

Present binary math started in World War II, and it will extend its rule. Therefore, any new "Physical" Binary Digital may need half a century to take over.

20.5 New Video

The greatest burden on our bill of resources is present digital video, and it will drive us into the electric-power singularity. Yet its reform has the greatest potential for

- **Savings with simultaneous gains in performance and quality.**

Chapters 12–14 have taught us, what we can gain from electronics mimicking the Human Visual System (HVS):

- **JND Log Recording and Coding**
- **Perception-Guided Compression**
- **Display-Oriented Tone Mapping**.

This strategy, in the end, could save an order-of-magnitude in video-data operations, traffic and storage.

20.6 Reliable Intelligent-Learning Nano-Systems

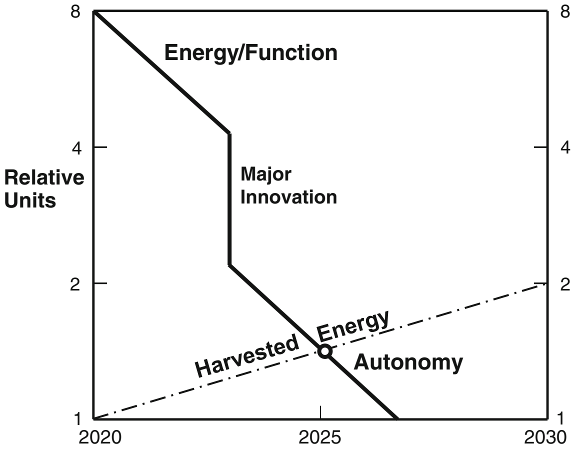
Presently, intelligent- or brain-inspired systems research is governed by mankind's desire to understand the brain and to make a copy of it. If our next generations of products and services should benefit from this research, we have to fulfill the basic requirements of reliability and predictability. These requirements, their methods, rules and verification, in fact, their incorporation into intelligent systems from their conception, are scientific-technical domains that need more attention, respect and rewards in science and business. Otherwise, the present 10× Giga-projects (Chap. 18) will pass by as another wave of AI without benefits for society [8].

FMEA (Failure-Mode-and-Effect Analysis), self-repair, and learning within bounds have to be key aspects of the expected innovations.

20.7 The Era of Energy-Autonomous Nano-Chip Systems

The continuing progress in energy harvesting and its power management (Chap. 19) make the autonomous operation of chips or chip-systems, independent of a battery, more likely. However, only if we reduce the power requirements of present chips more quickly by orders-of-magnitude, will we see a truly new era of nanoelectronics (Fig. 20.3).

Fig. 20.3 The path to energy-autonomous chips



Chapters 1–4, 10, 12–14, 16, 18, and 19 show effective ways to go. As with progress in batteries, the performance of energy harvesters, regarding power per cm^2, cm^3 and gram, will rise slowly. Therefore, the progress in energy-per-function will decide when autonomy will be possible. When this energy autonomy is accomplished on a larger scale and for more complex functions, a totally new era of nanoelectronics-everywhere will start, limited only by creativity. Certainly, the expansion of the Internet-of-Things (IOT) will accelerate, and the local mobile communication of intelligent systems, particularly man-machine systems, will expand beyond imagination.

20.8 Another Singularity?

Imagining nanoelectronics everywhere will lead to the question, if and when nanoelectronics will begin to dominate mankind, the question of the intelligence singularity. The hit-word “Singularity” has been launched into continuing debates since 1999 through Ray Kurzweil’s book “The Age of Spiritual Machines” [7]. A broad assessment of singularity definitions and hypotheses was published in 2012 [8]. Beyond the electric-power singularity in Sect. 20.2, the more speculative question, when machines will take over and govern mankind, receives some answers here from the perspective of the new vistas on nanoelectronics in this book:

1. Between 2010 and 2015, the progress in computing performance (Chaps. 9 and 10) and Internet bandwidth (Chap. 12) had to be reduced from 1000×/decade in [1] to an optimistic 130-to-200×/decade (Table 12.1).
2. Even with this slower growth, we will run into an electric-power singularity by 2020 because of too little progress in the energy efficiency of nanoelectronics.

3. All chapters in this book describe realistic and resources-conscious developments for nanoelectronics, which will deliver another 1000× improvement in the Internet energy per video frame as the most critical data on the Internet and in any communication within and between "intelligent" man-machine systems.
4. The realization of this progress will take into the 2030s due to proven time-constants.
5. An autonomous Mercedes 200T limousine drove with trained steering, automatic keeping of distance and lanes on highways in Germany in 1992 [9]. A wave on autonomous cars has just picked up intensity in 2014, more than 20 years later.
6. A new era is ahead with large-scale energy-autonomous chip systems everywhere for local, particularly man-machine communication and cooperation.

Since the progress in the energy efficiency of nanoelectronics has slowed down between 2010 and 2015 against the forecasts towards 2020, an electric-power singularity of the Internet is immanent by 2020 in the case of an expected efficiency improvement of only <3.5× between 2015 and 2020. This book identifies seven key areas to improve the energy efficiency 1000-times, particularly for video, the prime roadblock for the Internet and the no.1 sense for the cooperation and communication in intelligent man-machine systems. These key areas offer sustained growth for a nanoelectronics-driven global economy for 20 years. Intelligent, fairly autonomous man-machine systems will evolve in this time-frame in the

- **perpetual cooperation and competition between man and his machines**,

with reliability determining the pace [10]. An intelligence singularity will thus continue to be pushed into the future, following a practical understanding of "Artificial Intelligence" as the art of what we were not able to build until yesterday.

References

1. Hoefflinger, B.: Requirements and markets for nanoelectronics, chapter 6. In: Hoefflinger, B. (ed.) CHIPS 2020—A guide to the future of nanoelectronics. Springer, Berlin (2012). doi:10.1007/978-364223096-7_6
2. www.IC Insights.com/2015
3. https://technology.IHS.com/
4. Hoefflinger, B.: The energy crisis, chapter 20. In: Hoefflinger, B. (ed.) CHIPS 2020—A guide to the future of nanoelectronics. Springer, Berlin (2012). doi:10.1007/978-3-64223096-7_20
5. Raghavan, B., Ma, J.: The Energy and Emergy of the Internet, ICSI and UC Berkeley 2011. Copyright 2011 ACM 978-1-4503-1059-8/11/11. http://goo.gl/y4juZ
6. CISCO Visual Networking Index: Global Mobile Traffic Forecast Update 2014. www.cisco.com
7. Kurzweil R.: The Age of Spiritual Machines—When Computers Exceed Human Intelligence. Viking, The Penguin Group, New York (1999)
8. Eden, A.H., et al.: Singularity hypotheses: an overview. In: Eden, A.H., Moor, J.H., Soraker, J. H., Steinhart, E. (eds.) Singularity Hypotheses. Springer, Berlin (2012). doi:10.1007/978-3-642-32560-1_1

9. Neusser, S., Nijhuis, J., Spaanenburg, L., Hoefflinger, B.: Neurocontrol for lateral vehicle guidance. IEEE Micro **13**(1), 57–65 (1993)
10. Gomes, L.: Facebook AI director Yann LeCun on his quest to unleash deep learning and make machines smarter, Feb 18, 2015, http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/facebook-ai-director-yann-lecun-on-deep-learning

Titles in this Series

Quantum Mechanics and Gravity
By Mendel Sachs

Quantum-Classical Correspondence
Dynamical Quantization and the Classical Limit
By Dr. A.O. Bolivar

Knowledge and the World: Challenges Beyond the Science Wars
Ed. by M. Carrier, J. Roggenhofer, G. Küppers and P. Blanchard

Quantum-Classical Analogies
By Daniela Dragoman and Mircea Dragoman

Life—As a Matter of Fat
The Emerging Science of Lipidomics
By Ole G. Mouritsen

Quo Vadis Quantum Mechanics?
Ed. by Avshalom C. Elitzur, Shahar Dolev and Nancy Kolenda

Information and Its Role in Nature
By Juan G. Roederer

Extreme Events in Nature and Society
Ed. by Sergio Albeverio, Volker Jentsch and Holger Kantz

The Thermodynamic Machinery of Life
By Michal Kurzynski

Weak Links
The Universal Key to the Stability of Networks and Complex Systems
By Csermely Peter

The Emerging Physics of Consciousness
Ed. by Jack A. Tuszynski

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2

Quantum Mechanics at the Crossroads
New Perspectives from History, Philosophy and Physics
Ed. by James Evans and Alan S. Thorndike

Mind, Matter and the Implicate Order
By Paavo T.I. Pylkkanen

Particle Metaphysics
A Critical Account of Subatomic Reality
By Brigitte Falkenburg

The Physical Basis of the Direction of Time
By H. Dieter Zeh

Asymmetry: The Foundation of Information
By Scott J. Muller

Decoherence and the Quantum-To-Classical Transition
By Maximilian A. Schlosshauer

The Nonlinear Universe
Chaos, Emergence, Life
By Alwyn C. Scott

Quantum Superposition
Counterintuitive Consequences of Coherence, Entanglement, and Interference
By Mark P. Silverman

Symmetry Rules
How Science and Nature are Founded on Symmetry
By Joseph Rosen

Mind, Matter and Quantum Mechanics
By Henry P. Stapp

Entanglement, Information, and the Interpretation of Quantum Mechanics
By Gregg Jaeger

Relativity and the Nature of Spacetime
By Vesselin Petkov

The Biological Evolution of Religious Mind and Behavior
Ed. by Eckart Voland and Wulf Schiefenhövel

Homo Novus–A Human without Illusions
Ed. by Ulrich J. Frey, Charlotte Störmer and Kai P. Willfiihr

Brain-Computer Interfaces
Revolutionizing Human-Computer Interaction
Ed. by Bernhard Graimann, Brendan Allison and Gert Pfurtscheller

Extreme States of Matter
On Earth and in the Cosmos
By Vladimir E. Fortov

Searching for Extraterrestrial Intelligence
SETI Past, Present, and Future
Ed. by H. Paul Shuch

Essential Building Blocks of Human Nature
Ed. by Ulrich J. Frey, Charlotte Störmer and Kai P. Willführ

Mindful Universe
Quantum Mechanics and the Participating Observer
By Henry P. Stapp

Principles of Evolution
From the Planck Epoch to Complex Multicellular Life
Ed. by Hildegard Meyer-Ortmanns and Stefan Thurner

The Second Law of Economics
Energy, Entropy, and the Origins of Wealth
By Reiner Köummel

States of Consciousness
Experimental Insights into Meditation, Waking, Sleep and Dreams
Ed. by Dean Cvetkovic and Irena Cosic

Elegance and Enigma
The Quantum Interviews
Ed. by Maximilian Schlosshauer

Humans on Earth
From Origins to Possible Futures
By Filipe Duarte Santos

Evolution 2.0
Implications of Darwinism in Philosophy and the Social and Natural Sciences
Ed. by Martin Brinkworth and Friedel Weinert

Probability in Physics
Ed. by Yemima Ben-Menahem and Meir Hemmo

Chips 2020
A Guide to the Future of Nanoelectronics
Ed. by Bernd Hoefflinger

From the Web to the Grid and Beyond
Computing Paradigms Driven by High-Energy Physics
Ed. by Rene Brun, Federico Carminati and Giuliana Galli Carminati

The Language Phenomenon
Human Communication from Milliseconds to Millennia
Ed. by P.-M. Binder and K. Smith

The Dual Nature of Life
By Gennadiy Zhegunov

Natural Fabrications
By William Seager

Ultimate Horizons
By Helmut Satz

Physics, Nature and Society
By Joaquín Marro

Extraterrestrial Altruism
Ed. by Douglas A. Vakoch

The Beginning and the End
By Clément Vidal

A Brief History of String Theory
By Dean Rickles

Singularity Hypotheses
Ed. by Amnon H. Eden, James H. Moor, Johnny H. Søraker and Eric Steinhart

Why More Is Different
Philosophical Issues in Condensed Matter Physics and Complex Systems
Ed. by Brigitte Falkenburg and Margaret Morrison

Questioning the Foundations of Physics
Ed. by Anthony Aguirre, Brendan Foster and Zeeya Merali

It From Bit or Bit From It?
Ed. by Anthony Aguirre, Brendan Foster and Zeeya Merali

Index

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2