Available online at www.sciencedirect.com

Computer Communications 28 (2005) 956–966

www.elsevier.com/locate/comcom

# High-speed networks: definition and fundamental attributes

Seshasayi Pillalamarri[a], Sumit Ghosh[b,*]

[a]City of Phoenix Municipal Court, Information Systems and Technology Division, 300 W Washington St., Phoenix, AZ 85003, USA
[b]Secure Network Systems Design Laboratory, Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA

Received 16 September 2003; revised 1 July 2004; accepted 27 July 2004
Available online 8 September 2004

## Abstract

This paper represents an effort to develop a simple yet comprehensive understanding of networking through a novel definition—high-speediness, which encapsulates the high-speed nature of networks and is denoted by the symbol '*s*', two fundamental observations that underlie every network design, and the identification of the fundamental attributes of networks. While the definition is mathematical yet physically intuitive, the two observations are due to the laws of physics and imply fundamental limits on networks, and the attributes fall out from a careful analysis of the primary objectives of networks. In addition to the inherent pedagogical value, high-speediness may be used to classify and compare existing networks. The paper computes the '*s*' values for important past and present networks. The high-speediness factor may also constitute a desirable, target network operating point, which the network provider may choose to sustain during network operation by imposing suitable controls. The fundamental attributes represent a holistic view of network, revealing the most important issues and how they are interconnected to each other. Together with the attributes, the high-speediness promises to serve as a meaningful guide in the design of future high-speed networks, which constitutes the most contribution of this paper.

*Keywords:* Networking; Communications; Fundamental attributes; High-speed networks; ATM networks; IP networks; Multimedia

## 1. Introduction

Observations of the enormous complexity of today's networks, their increasing importance and ubiquitous use in the world, the uncontrollable number of diverse dimensions in which networking is expanding, the apparent ad hoc development and deployment of networking techniques, coupled with the intense debate between the superiority of ATM versus IP, had motivated the authors to seek a systematic, logical, and scientific way of understanding the monster that networks have now come to represent. The most obvious question was, what is a high-speed network? What imparts to a network the label, high-speed? Clearly, a link—optical or twisted pair, that interconnects two ATM nodes at 155.5 Mb/s may not qualify alone as high-speed since one can synthesize it from approximately one hundred T-1 lines, each rated at 1.5 Mb/s and definitely not viewed high-speed. Of great interest is the fact that although the DoD's Arpanet was heralded as a high-speed network back in the early 1980s, today most users are unhappy with its performance. What has caused this change? In an effort to seek answers, it was reasoned that to identify the inherent, indispensable characteristics of high-speed networks, one must first search for a definition of high-speed networks, one that is mathematical yet simple but, most important, physically intuitive. Then, for key past and present networks, their high-speed characteristic must be computed to help us trace the evolution in networking from the perspective of their high-speed nature. Next, utilizing the definition as a reference and recognizing the key functions of a network, one would have to identify the fundamental attributes of networks. The exercise will provide valuable understanding of networks, especially the fundamental limitations, if any. This, in turn, will help us consider the important issues while designing new network architectures and provide us insight into their ultimate potential.

A thorough literature search on the fundamental characteristics of high-speed networks revealed sparse

---

* Corresponding author.

*E-mail addresses:* sesh.pilla@phoenix.gov (S. Pillalamarri), sumit.ghosh@ieee.org (S. Ghosh), sghosh2@stevens.edu (S. Ghosh).

0140-3664/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2004.07.013

results. In characterizing high-speed local area networks (HSLANs), Abeysundara and Kamal [1] compute a quantity, '*a*', labeled normalized propagation delay. It is defined as the ratio of the physical propagation delay between a given pair of source and destination points to the packet transmission time, for a representative packet size for LANs. They report findings that for a packet size of 200 bits, transmission speed of $2\times10^8$ m/s, and a link bandwidth of 100 Mb/s, '*a*' equals 1.25 and 2.5 for LANs spanning 5000 and 10,000 m, respectively. They conclude that as the value of '*a*' approaches or even exceeds unity, the LAN may be viewed as operating at high-speeds. Furthermore, higher the value of '*a*', the more high-speed the LAN. While a pioneering work, the effort incurs three limitations. First, it is limited to LANs. Second and more important, the concept of '*a*' fails to reflect the quality of service (QoS) considerations associated with a message communication. Third, Abeysundara and Kamal fail to link their choice of the definition of '*a*' to any scientific reference. In characterizing network performance, Stallings [2] defines a symbol '*a*', exactly as in Ref. [1], but uses it to explain why congestion control is difficult in networks that exhibit high values of '*a*'. Stallings reports '*a*' values ranging from 0.1 for a 10 Mb/s Ethernet to 21,200 for wide-area ATM network. This definition suffers the same limitations as those in Ref. [1]. In arguing that gigabit networks are radically different from their predecessors, Kleinrock [3], defines a quantity, $a=5\times L\times C/b$, where $L$ represents the length of the network in miles, the factor 5 is approximately the number of microseconds light takes to travel a mile, $C$ is the capacity of the network in Mb/s, and $b$ the number of bits in the data packet. Assuming only a few users sending packets at gigabit speeds, Kleinrock observes that the value of '*a*' ranges from 0.05 for a local area network to 15,000 for a gigabit network across the US, and offers two conclusions. First, at bandwidths in the range of gigabit/s, latency assumes a dominating role. Second, he agrees with [2] in that end-to-end congestion control through feedback is likely to pose an enormous challenge. In contrast to [3], this paper focuses on a definition of practical high-speed networks, as shared by a large number of users with diverse traffic transportation needs.

In explaining techniques to achieve low latency communication, Thekkath et al. [4] observe that the controllers play an important role in determining the overall latency in high-speed networks. For the dynamic, synchronous transfer mode gigabit network with dynamic reallocation of bandwidth and reserved channels, Bohm et al. [5] claim that not only will their network be free from congestion but that well defined real-time services may be provided. Chlamtac and Franta [6] define a network as high-speed, if it succeeds in adequately handling the node transmission needs, both at the individual and aggregate-levels. They argue that whether or not a network is high-speed, is strictly viewpoint dependent. They further state that queuing delays, network algorithms, management, routing, security, and node processing overheads, serve as obstacles to the development of high-speed networks. Maxemchuk [7] illustrates through examples how sharing in network can be effective and that buffering of traffic sources at nodes may be effectively avoided through dispersity routing. Neither of the claims in Refs. [5] and [7] are validated for real-world networks.

Gerla [8] utilizes their operating data rates to classify LANs into high- or low-speed networks. It has been suggested that the widespread proliferation of LANs had motivated the development of new, high-speed interconnecting networks to carry the aggregate traffic. Furthermore, the mismatch resulting from connecting the LANs rated at 10 Mb/s to 1 Gb/s to the WANs rated at Gb/s, gives rise to congestion. Cidon et al. [9] claim that the new high-speed networks will possess higher transmission and switching speeds since they will be hardware-based, compared to the current software-based processing times. They propose new models for distributed algorithms that will differentiate between switching and processing times, thereby isolating the hardware and software functions. The claim is yet to be proven. Bolla and colleagues [10] examine the switching aspect of high-speed networks. Mink et al. [11] confirm that, in high-speed networks, the data copying, checksum generation, and buffer management functions consume a majority of the execution time of the underlying workstations in the form of sending and receiving data through TCP/IP. They hypothesize that large window and packet sizes will very likely enhance the performance of workstation-based communication. Oie et al. [12] conjecture that, since different media and source traffic exist in a switch of an integrated services network such as ATM, at the same time, heavy contention for the switch ports, is likely. Thus, assumptions of uniform traffic for switching purposes are incorrect and superior switching architectures are necessary. In examining the role of network controls, Fendick et al. [13] state that efficient control is a key towards achieving effective bandwidth utilization, error control, fair bandwidth allocation among users, and in limiting queue lengths through effective network memory utilization. Heinaenen [14] defined, in 1991, that for a network to be labeled high-speed, it must support data rates ranging from 2 to 34 Mb/s.

The remainder of the paper is organized as follows. Section 2 first presents two fundamental observations that underlie all networks and are based on physics. Section 3 presents a novel definition of high-speed networks, namely, high-speediness, and computes values for key past and present networks. Section 4 proposes a set of fundamental attributes in networks that are both inherent and indispensable. Finally, Section 5 presents some conclusions.

## 2. Two fundamental observations

The discovery and employment of switching constitutes a significant factor in the development of networking.

Switching represents an economic solution to the problem of interconnecting any user to any other user in the network despite a partially-connected network. Clearly, switching requires knowledge and intelligence. In the initial days of telephone and telegraphy, human operators would manually intercept such requests and establish the links. The links would remain active for the entire duration of the communication and then reset. As the telephony usage increased dramatically, the human operators were replaced first by relays and then, by computers, the key reason being the demand for increased computational intelligence and fast processing.

Thus, the first observation is submitted in that computational intelligence is a fundamental requirement in networking, with the demand for complexity and speed in intelligent processing growing with increasingly sophisticated high-speed networks of the future. Since computational intelligence is tied to energy, from physics, it is inferred that its availability is limited. This, in turn, will exert undeniable influence on the evolution of networking. There is no doubt that, to-date, the availability of the computing engines have been the key instrument in shaping the evolution in networking. This point is illustrated through a true story. On June 15, 1896, an experiment was undertaken [15] to determine the length of time necessary to send a telegraph message around the globe. The three-line message was sent via Chicago, Los Angeles, San Francisco, Vancouver, Winnipeg, Montreal, Cannes, London, Lisbon, Gibraltar, Malta, Alexandria, Suez, Bombay, Madras, Singapore, Shanghai, Nagasaki, and Tokyo. The total transit time was computed at 8 h and 34 min, at a cost of $152. In today's networks, the same message may be transported in as little as a few hundred milliseconds, the key difference being made by nature of today's computational intelligence—fast and powerful. Although the physical transmission time of the electromagnetic (EM) message has not changed since then, in 1896, human operators at every station, not computers, had to first receive and transcribe the telegraph message and then send the message along to the subsequent station by tapping on the telegraph machine. Very broadly, the need for computational intelligence appears in two forms. First, just prior to launching a message, the source entity may utilize computational intelligence and determine the complete route up to the destination and encapsulate this information in either the message itself, in the intermediate switches, or a combination of both. Second, a message is launched into the network without precise knowledge of its route and intelligence is expended at the intermediate switches to help guide the message to its destination. In summary, computational intelligence is fundamentally important in networking, as long as the economics of partially connected networks remain in effect, and network usage is dynamic, i.e. the resource usage cannot be predicted with complete certainty, a priori. Computational intelligence is tied to energy and its limited availability will continue to play a key role in the evolution of networking.

The second fundamental observation is the absence of pure energy computers which, in turn, bears an undeniable impact on scalability in networking. Material transport networks including inventory management or air cargo networks are similar to networks that carry EM messages in virtually every respect, especially from the perspective of the fundamental role that computational intelligence plays in routing items from the source to the destination. However, there is one difference between them, one that is subtle yet profound. To-date, computational intelligence has come to us in material form, whether as human beings, electronic circuitry, or programmable computers. Science has not yet succeeded in designing computing engines that constitute pure energy. As a result, in material transport networks [16], a computer may be carried along with every material item being transported and the computer's computational intelligence may be utilized during the transported item's transit through the network. This, in turn, implies that such networks are scalable [17], from the computational perspective. That is, if the number of items transported is doubled, the number of computing engines associated with them, also doubles and the performance degradation is not appreciable. In contrast, an EM message cannot carry with it a computing engine during its transport through the network and all of its computational intelligence needs must be addressed by the limited number of 'matter-based' switches in the network. Since the number of EM messages transported in a network is, in general, orders of magnitude higher than the number of switches, a EM message transport network is fundamentally, not scalable. The concentration of computational intelligence solely at the switches, coupled with their limited number as dictated by economics, must constitute an important consideration in future network design. Networking will experience a profound evolution if and when science is able to synthesize computing engines in pure energy form.

## 3. A definition of high-speed networks

The term, high-speed, associated with networking, requires qualification. In any two networks that utilize the same media, even if one is labeled high-speed while the other labeled low-speed, the actual speed with which the bits are transmitted, is identical and it equals the speed of propagation of EM transmission in that medium, including electrical, electronic, or optical signals. However, a network may set a limit on the frequency with which a user may insert bits into the network. Thus, contrary to ordinary perception, the term, high-speed, does not refer to the actual speed of a single bit, rather to the time elapsed for a collection of bits, viewed as an atomic unit and referred to as a frame, packet or cell, between its launch at the source point and the eventual arrival at the destination. Consider the following example. Although, at birth in the late 1970s, the Arpanet was viewed as a high-speed network, its present day incarnation, the Internet, is no longer considered high-speed. Introspection reveals that since the 1970s, while the average distance over which users communicate has increased, the number of users and the traffic volume and diversity have increased dramatically. Logically, therefore, the distance and the number and nature of users must play a role in determining the high-speediness of a network.

Intuitively, one expects a network to offer the same fidelity with respect to a communication or control between two geographically separated users A and B, as if they were located right next to each other. Where A and B are human beings, the communication refers to all forms of human interaction. Where A and B are machines, fidelity refers to reproduction of traffic patterns at the receiver exactly as generated at the source, in every respect. This is the ideal case. In reality, however, the fidelity is highly degraded. There are four key factors that prevent the realization of the ideal network, as enumerated below:

- Physical propagation delay that stems from the finite propagation speed of EM waves through any medium.
- The limited bandwidth availability of the links in a network.
- Buffer and/or queueing delays that result from multiple users attempting to utilize the network simultaneously.
- The nonzero processing time at each node associated with routing a packet, subject to different requirements and constraints, introduces a delay.

In proposing a definition, it is recognized that there are two important design parameters in any network. These parameters are set by the network designer and their choices reflect the purpose and expected use of the network. They include:

- The characteristic distance
- The characteristic frame size

The characteristic distance is the typical physical distance over which a representative communication occurs in a given network. For instance, it may refer to the average distance, anywhere in the network, over which most message communication take place. An underlying assumption here is that the characteristic distance is a design parameter and remains in effect during the operational life of the network. The characteristic frame size refers to the 'minimum quanta of information transmitted within the network' that is meaningful to the applications that the network is designed to support. From the perspective of transport, it is viewed as an atomic unit. A frame may include additional bits for header and trailer information where necessary. The notion of characteristic frame size may not apply to classic telephony in which users were allocated fixed bandwidths.

First a quantity is defined, transmission time ($t$), as the elapsed time interval between the first bit of the characteristic frame leaving a reference point and the last bit of the same frame reaching the destination point at a characteristic distance away, minus the physical propagation delay between the source and destination points. Thus, '$t$', measured over a characteristic distance, is given by:

$$\begin{aligned} t &= \text{buffer delay} + \text{processing time} \\ &\quad + (\text{characteristic frame size/link transmission rate}) \end{aligned}$$

In the above equation for $t$, clearly, the transmission time is a function of the available bandwidth of the medium and the characteristic frame size. Specifically, the 'link transmission rate' refers to the typical bandwidth either utilized or made available to representative source traffic. In the equation for $t$, buffer delay and processing time measures refer to the cumulative values across the different nodes that are incurred by a characteristic frame en-route over the characteristic distance. These values reflect representative measures corresponding to normal operating conditions. The buffer delay is a direct consequence of the presence of memory elements in the network, the need for which stems from four causes. First, to process a piece of information, i.e. subjecting it to computing, it must first be stored. Second, traffic is generated asynchronously with respect to the operations of the network elements. Third, the burstiness of traffic sources coupled with the asynchronicity can lead to congestion, i.e. an excess of traffic cells, beyond the design capacity, at specific points in the network and at specific time instants. This requires the use of buffers in lessening the chance of cell loss. Fourth, network resources must be shared among multiple users, simultaneously engaged in generating and inserting traffic into the network. The processing time reflects the first fundamental observation in networking, outlined earlier in Section 2. In general, the characteristic distance will influence the number of nodes the characteristic frame must encounter during its transit and, in turn, affect the processing time and buffer delay values in the equation for $t$. The number and diversity of the users in the network dictates the number and nature of the traffic sources utilizing the links within the characteristic distance and, therefore, influences both the buffer delay and processing time values. QoS controls and other principles underlying the network under consideration also impact the buffer delay and processing time values. The exact units of the quantities in the above equation will depend on the network under consideration. Estimates of '$t$' may be obtained through sophisticated modeling and simulation [18].

Next, another quantity is defined, physical propagation delay ($p$), as the delay associated with the propagation of the packet over a characteristic distance, expressed below. Here too, the exact units of the quantities in the equation will depend on the network under consideration.

$$p = (\text{characteristic distance/propagation speed of EM transmission})$$

Now, the high-speediness of a network (s) is defined as the ratio of $t$ to $p$, expressed as:

$$s = (t/p)$$

In the definition, $p$ is extremely fundamental for, it is tied to the physics of reality. It is independent of the current networking technology and serves as a frame of reference to the high-speediness of a network, enabling the evolution of networking with time, from the perspective of the high-speed nature, to be tracked. In contrast, $t$ is an artifact of the network. It emanates from the network and is determined by its design. The definition of $s$ is general and is applicable to all networks. The inclusion of $p$ into the equation imparts to high-speediness a timeless quality in that the '$s$' values of the networks may be used to compare their high-speed nature as they evolve with time.

The high-speediness value may be viewed from two different perspectives. Under the first perspective, termed 'design perspective', the network is expected to remain operational throughout its life at the '$s$' value for which it was designed. This clearly implies that the network provider intends to use the high-speediness value as the network operating point and must impose whatever controls necessary to ensure the objective. Under the second perspective, termed 'classification perspective', the constituent components of high-speediness are measured in an actual operational network under normal conditions, not exceptional scenarios such as severe congestion, and the '$s$' value is computed to reveal whether the network is high-speed. The measurements reflect average values in the network, not relative to any specific user. Whether steps are taken subsequently to force the network to a desired s value is up to the network provider. While the high-speediness value helps focus on the most important elements of a network, it also underscores the fact that, in a given network, one or more of the factors may play a dominating role. For example, in telephone systems where the link transmission rate is low, the third item, '(characteristic frame size/link transmission rate)' dominates the other two in the numerator. In contrast in ATM networks where the link transmission rates, by design, are high, the third item will exert little influence on the high-speediness value of the network.

This paper proposes that a network be classified high-speed if its high-speediness value is always less than unity. Thus, a high-speed network is one where its link bandwidth, the source traffic controls that guarantee QoS for all traffic sources, and the total processing times within the nodes are such that the transmission time of a characteristic frame's transportation across its characteristic distance, is always less than the physical propagation delay of EM transmission for that distance in that medium, irrespective of where the measurement is obtained in the network.

Intuitively, the meaning of high-speediness $<1$ is as follows. The primary objective of the network is to transport a characteristic frame over a characteristic distance, as quickly as possible. The word 'quickly' is a relative term and must be tied to a meaningful reference. The first bit of the characteristic frame reaches the destination at time $(p+t_0)$, assuming that it is launched from the source at time $t_0$. The last bit of that frame and, therefore, the entire frame itself, reaches the destination at time $(p+t_0+s\times p)$, where high-speediness is less than unity. For increasingly high-speed networks, high-speediness assumes smaller and smaller values and, in the limit, approaches 0 for an ideal network. That is, the entire frame reaches the destination quickly, relative to the time it took to receive the first bit of the frame, the only relevant reference standard for the given network. It is noted that while high-speediness $<1$ is arbitrary, it is a reasonable and logical choice.

A common intuition relative to a slow speed network is that even when the leading bit of a frame has reached the destination, the trailing bit of the frame is yet to be launched from the source node. This intuition is consistent with the definition of high-speediness, explained as follows. Consider, without any loss in generality, that the buffer delay and processing time are both 0. So, for a slow speed network, high-speediness $>1$, implying that,

$$(\text{characteristic frame size/link transmission rate}) > (\text{characteristic distance/propagation speed of EM transmission})$$

While the term on the right hand side of the inequality encapsulates the time it takes for the leading bit of the characteristic frame to propagate from the source to the destination, the left term refers to the width of the characteristic frame in time. When this width exceeds the right term, clearly the trailing bit of the characteristic frame could not have been launched from the source node when the leading bit has been transported to the destination.

From the definition, it follows that an ideal or perfect network design is one where $t$ equals 0. Therefore, $s=0$. For $t=0$, the link transmission rate must be infinite, implying infinite bandwidth. The buffer delay must equal 0, implying that the memory should be infinite in capacity so that loss is nil and that memory accesses must require zero time. Also, the processing delay must equal zero, implying that processors must operate at infinite speeds with computations performed instantly. While, clearly, an ideal network is never realizable, this exercise points to the directions along which networking research should evolve. The definition also yields a new view into the nature of networks. The quantity '$t$' reflects the resistance that the network design under consideration offers to the transport of information packets from sources to the corresponding destinations across the network. The resistance is determined solely by the network design parameters. The higher the resistance, the lower the high-speediness of the network. Clearly, future evolution must focus on reducing the resistance through superior processor and buffer design, efficient algorithms, and higher link bandwidths, subject to desired characteristic frame size and distance measures.

To illustrate the definition, consider two networks that constitute two opposite extremes, relative to their geographical range. The first is an extremely wide area, deep-space network, wherein the characteristic distance is set at $3\times10^8$ m and characteristic frame size at 128 Kb. Assuming an EM transmission speed of $3\times10^8$ m/s in free space, $p$ computes to 1 s. Assume that the processing time and buffer delay are 0 and the link bandwidth is 64 kb/s. Then, $s=(0+0+128/64)/1=2$, which exceeds unity. Thus, the network under consideration is a slow network. For the same network, when the characteristic frame size is reduced to 64 Kb, $s=1$, implying that the network is neither high-speed nor slow. However, when the characteristic frame size is further reduced to 32 Kb, $s=0.5$, implying a high-speed network. Given that the link bandwidth in the network is only at 64 Kb/s, the common perception, ignoring the processing time and buffer delay, is that the network is slow. However, the enormous geographic distance over which packets are transported qualifies this network as high-speed. The inference is logical from the primary viewpoint that networks should be evaluated from the perspective of its purpose and design parameters.

The second scenario consists of an extremely small area network, a VLSI chip. Consider that the characteristic frame size is 32 kb. Assuming a characteristic distance of 1 cm, and given the speed of EM transmission is $2\times10^8$ m/s, $p$ equals $50\times10^{-12}$ s. Assume an available link bandwidth of 64 Mb/s, which is certain to lead to a common perception that the network is high-speed. However, $s$ computes to $10^7$, implying a very slow network. To transform it into a high-speed network with $s=1$, for a packet size of 32 bits and assuming processing time and buffer delay values of 0, the required link bandwidth, $X$, is given by $(0+0+32/X)/(50\times10^{-12})$, which yields $X>1.56$ terabits/s. The inference is as expected. For source and destination points at close proximity, an extremely high transmission rate is indispensable to qualify a network as high-speed.

Thus, the high-speediness value is the first in the literature to reflect the quality of service attribute of networks by encapsulating the notion of delay. It is also unique in that it addresses the computational aspect of networking. Beyond its intrinsic pedagogical value, meaningful usage of the high-speediness may consist in classifying and comparing existing networks. More importantly, it may constitute a desirable, target network operating point which the network provider may choose to sustain during network operation by imposing suitable controls.

The definition reveals how the high-speed Arpanet of the past is presently the slower Internet. Significant increases in the number of users and traffic volume, coupled with the increase in the characteristic distance, have led to enormous increase in node processing times and the buffer delays, pushing the '$s$' value much higher than unity.

### 3.1. Tracing the evolution in high-speed networking

To understand the gradual and continuing evolution in high-speed networks from the perspective of their high-speed nature, this section computes and presents the '$s$' values for key past and present networks. We assume a speed of $2\times10^8$ m/s for propagation of a signal in an optical fiber medium and $3\times10^8$ for free space.

The digital private automatic branch exchange (PABX) network is setup in a local area, an organization, or a campus usually limited to an area of 5 km radius, and operates on the circuit switched principle. The digital PABX provides telephone service and data service through modems. A few trunk lines, relative to the number of sources, are provided to connect to the central office of the public telephone system and exploits the fact that not all sources dial outside the exchange at all times. It employs a switch topology where all the lines are connected to a main distribution frame and the exchange cards provide the required switching function. Addressing a source is achieved through a numbering scheme. Pulse code modulation (PCM) scheme is used for transmission of voice and data in the PABX network. It transmits 193 bit frames (24 channels of 8 bits of voice from different sources plus one frame bit) over a 1.544 Mb/s T1 carrier line. Assuming zero processing time after a call is established, and since network sharing delay in circuit switching is zero, the transmission time for a characteristic frame of 193 bits is 125 μs. The propagation delay being $10^{-5}$ s for a characteristic distance of 2 km, high-speediness equates to 12.5, a slow speed network.

Ethernet LAN provides high degree of interoperability among heterogeneous computers. Even though the throughput has increased with higher data rate transmission standards, the high-speediness of the LAN remains same essentially due to the network sharing delay being high under full load of the network. Assuming a characteristic distance of 3.6 km, its 1522 byte frame yields a high-speediness factor of 297, not a very high-speed network. The voice over IP application on the wireless LAN deployment has high access point congestion and hence high network sharing delay. Its delay variance of 50 ms is a major factor in the '$s$' factor being a high of 7515. Clearly it is a slow network. For the 16 Mb/s token ring LAN architecture, assuming a characteristic distance of 5 km, a representative packet length of 4064 bytes, network sharing delay of 5 ms,and processing time of 50 μs,the high-speediness computes to 212.1. Thus, it is a slow network.

The LAN emulation on an ATM network has a maximum of 1550 byte frame. Transmitting it over a distance of 5 km with an estimated 11 ms of sharing delay yields an ‘*s*’ factor of 449. It is a slow network. For the 155 Mb/s ATM under LAN mode, assume a characteristic distance of 10 km, a fixed frame length of 53 bytes, delay due to traffic at 1 ms, and processing time of 19 μs at the nodes. Under these circumstances, high-speediness computes to 20.38, qualifying the network as higher speed than the ATM LANE network.

For the 100 Mb/s FDDI ring, assuming a characteristic distance of 10 km, a representative frame length of 4096 bytes, delay due to traffic at 500 μs, and processing time of 50 μs, high-speediness computes to 11.82, implying a relatively higher speed network.

For the 800 Mb/s Fiber channel, under the HIPPI protocol, assume a characteristic distance of 5 km, a representative packet length of 2148 bytes, delay due to traffic at 10 μs, and processing time of 100 μs, the ratio high-speediness is obtained as 4.50. Thus Fiber channel under HIPPI protocol qualifies as even higher speed network.

For the 45 Mb/s, distributed queuing dual bus (DQDB) metropolitan area network, identified as the IEEE 802.6 standard, assume a characteristic distance of 100 km, a fixed packet length of 53 bytes, sharing delay due to traffic at 1 ms, and a processing time of 100 μs. The high-speediness value is 2.20, implying the network is even superior.

For the 64 kb/s X.25 wide area network, assume a characteristic distance of 1000 km, a representative frame length of 1053 bytes, delay due to traffic at 100 ms, and a processing time of 500 μs. Then high-speediness computes at 23.39, implying a relatively fast network. For the 2 Mb/s Frame relay network, assuming a characteristic distance of 1000 km, a representative frame size of 1608 bytes, delay due to traffic of 1 ms,and a processing time of 100 μs,high-speediness equates to 0.36. Thus, for longer distances and smaller packet sizes, Frame relay technology constitutes a high-speed network.

For the 155 Mb/s wide area ATM network, assume a characteristic distance of 1000 km, a fixed packet length of 53 bytes, delay due to traffic at 1 ms,and processing time of 110 μs in the nodes. Then, high-speediness computes to 0.223, qualifying ATM as high-speed networking technology.

Table 1 summarizes the results.

## 4. Fundamental attributes of networks

From the definition of high-speediness, it follows that distance, bandwidth, EM transmission medium, intelligent processing, and packet size, are indispensable factors underlying any network. To gain a total understanding of every possible factor that underlies networks, this section analyses networking, starting with its primitive purpose, and yields a comprehensive set of the fundamental attributes. A fundamental attribute is defined, in this paper, as an indispensable, inherent property of a communication network without which the network is inconceivable.

According to the USA Federal standard 1037, ‘a communication network is an interconnection of two or more users or processes which are involved in ‘information transfer’ according to agreed conventions. The meaning assigned to the information must be preserved during these transfers.’

### *4.1. Distance*

Since communication involves at least two users or processes, from the above definition, and as no two entities may occupy the exact same geographical point at the same time, distance inevitably becomes a fundamental attribute Without distance, be it as large as astronomical units or as small as those in microelectronics, no network is conceivable. Every research into new network design must take the issue of distance into consideration.

Distance is a fundamental property of the physical world and, in a network, it behaves as a resistance to the transport of a message. From physics, since the speed of EM waves is finite, distance translates into physical propagation delay, higher number of user traffic sources, greater number of networking resources, higher chance of resource contention among users, greater probability of noise, errors, faults, and security problems. Since network topology depends on the distance, the latter plays a key role in network performance.

### *4.2. Asynchronism*

Clearly, in the most general case, the timing of the communication between two or more users is likely to be asynchronous, i e. irregular in time. Assuming A and B as two users, neither A nor B can know, a priori and with certainty, when the other will initiate communication. Nevertheless, either one of them must be prepared to respond to the other when contacted. Also, when A contacts B, it has no a priori certain knowledge of how quickly B will respond. Although synchronous networks are conceivable and many are in operational use, the synchronous design principle breaks down as networks grow to encompass vast geographical distances, greatly increased number of nodes and users, and very high-speeds. The traffic sources' interactions with the network elements are inherently asynchronous. Therefore, source traffic control techniques to ensure QoS in future networks must take into consideration the asynchronism. While the nodes of a network interact asynchronously among themselves, faults and errors in the interactions between nodes may also occur asynchronously. Hence, the design of the timing and control algorithms as well as security techniques and recovery procedures, in future network evolution must increasingly focus on the asynchronicity.

Table 1
High-speediness for past and present networks

| Network name | Characteristic distance (km) | Characteristic frame size (maximum) | Representative data rate | Typical buffer delay (per frame over the distance) | Typical processing time (per frame over the distance) (µs) | High-speediness (s) |
|---|---|---|---|---|---|---|
| PABX (Base line) | 2 | 193 bits | 1.54 | 0 | 0 | 12.5 |
| 802.11a Wireless LAN | 2[a] | 2346 bytes[b] | 54 Mb/s[a] | 50 ms | 10[c] | 7515 |
| 802.11 b Wireless LAN | 2 | 2346 bytes | 11 Mb/s | 50 ms[d] | 10[c] | 7512 |
| 802.11 g Wireless LAN | 2 | 2346 bytes | 54 Mb/s | 50 ms (SD of delay variance) | 10[c] | 7515 |
| 802.3 Ethernet LAN | 3.6[e] | 1522 bytes[e] | 10 Mb/s | 5 ms | 200[f] | 297.3 |
| Fast Ethernet | 3.6 | 1522 bytes | 100 Mb/s | 5 ms | 200 | 289.7 |
| Gigabit Ethernet | 3.6 | 1522 bytes[g] | 1 Gb/s | 5 ms[h] | 150[i] | 286.1 |
| 802.2 Token Ring LAN | 5 | 4064 bytes[j] | 16 Mb/s | 5 ms | 50 | 212.1 |
| ATM LANE (LAN Emulation) | 5 | 1550 bytes[k] | 155 Mb/s | 11 ms[l] | 219 | 449 |
| ATM LAN | 10 | 53 bytes | 155 Mb/s | 1 ms[l] | 19[m] | 20.38 |
| FDDI Ring (Obsolete) | 10[n] | 4096 bytes[o] | 100 Mb/s | 500 µs | 50 | 11.82 |
| Fiber Channel | 5 | 2148 bytes[p] | 800 Mb/s | 10 µs | 100 | 4.50 |
| DQDB MAN | 100 (typical) | 53 bytes | 45 Mb/s | 1 ms | 100 | 2.20 |
| X.25 (Obsolete) | 1000 | 1053 bytes[q] | 64 Kb/s | 100 ms | 500 | 23.39 |
| Frame Relay | 1000 (typical) | 1608 bytes[r] | 2 Mb/s | 1 ms | 100 | 0.36 |
| PPP over ATM AAL2 | 1000 | 1507 bytes[s] | 155 Mb/s | 1 ms[l] | 110[m] | 0.223 |
| ATM WAN | 1000 (typical) | 53 bytes | 155 Mb/s | 1 ms[l] | 110[m] | 0.222 |

[a] IEEE 802.11a,b,g. This distance depends on the gain of the antenna and peripheral environment. In the case of a point to multipoint configuration the maximum radius is 3 km.

[b] IEEE 802.11a,b,g with 34 addressing and control bytes + 2312 data bytes.

[c] Robert Hoskins, ‘AirFlow Networks First to Deliver Hardware-based Processing for Massive WLAN Scalability’, Broadband Wireless Exchange Magazine

[d] Jan Linden, ‘Latency key in Wireless-net management’, Global IP Sound Inc., Sept 2003.

[e] IEEE 802.3 allowing 4 bytes extra for VLAN tag over 1518 bytes of original standard.

[f] ‘Ethernet switch latency times’, McClellan Consulting, 1999. http://www.mcclellanconsulting.com/analysis/latanal.html.

[g] IEEE 802.3 or IEEE 802.3z.

[h] Donghui Xie, ‘Ring Traffic Convergence with failures—average delay experienced under stress in a server of a switched Ethernet ring with TCP application’, Cisco, IEEE 802 Resilient Packet Ring Study Group, November 6–9, 2000.

[i] ‘Understanding Gigabit Ethernet Performance on Sun Fire Servers’, Sun Micro systems, Aug 2003.

[j] IEEE 802.5, 1985.

[k] ‘LLC multiplexed data frame format (28 + 1522 bytes): LAN Emulation over ATM’, ATM forum—AF-LANE-0084.000, July 1997.

[l] ‘Support of Delay and Jitter Requirements’, Working group of Task Force TF-NFN of National Research Networks of Europe (NREN), May 2000 http://www.cnaf.infn.it/(ferrari/tfng/qosmon/.

[m] ‘(Assuming 6 ATM switches and multiplexers 6×18.41 = 110.46 µs), Frame Delay Through ATM Switches: MIMO Latency‘, ANSI T1A1.3/98-056, November 2–4, 1998.

[n] Maximum segment length is not specified in FDDI specifications. This is an estimated characteristic distance for a 1300 nm wave length, 9 microns core, single mode optical fiber.

[o] ISO 9314.

[p] X3T9.3 Task Group of ANSI, ‘Fiber Channel Physical and Signaling Interface (FC-PH), Rev. 4.2’, October 8, 1993.

[q] CCITT Recommendation X.25, ‘Interface Between Data Terminal Equipment (DTE) and Data Circuit Terminating Equipment (DCE) for Terminals Operating in the Packet Mode on Public Data Networks’, International Telegraph and Telephone Consultative Committee Yellow book, Vol.VIII.2, Geneva, 1981.

[r] ‘Multiprotocol Interconnect over Frame relay‘, RFC 1294.

[s] ‘PPP encapsulated IP packets with header and trailer over ATM AAL2‘, RFC 3336, Dec 2002.

### *4.3. Traffic sources*

Traffic sources are naturally indispensable in networks and estimates of expected traffic in the network must constitute an integral component of any network design In general, a network designer does not possess a precise picture of how the network under design will be utilized after its deployment. As a result, traffic estimation in the real world can pose a significant challenge. Traffic engineering must include both high-level issues including the distribution as a function of time of the (1) number of traffic sources, (2) their bit rates, (3) session durations, and (4) QoS requirements, as well as lower-level issues such as, for each traffic source, (a) the cell arrival distribution, (b) whether the traffic is constant bit rate, bursty, fractal, or self-similar, (c) cell-level traffic mix of video, audio, and data, and (d) active and silence interval times within a frame. As stated in Section 4.2, the arrival of a cell at a network node is, in general, asynchronous.

### 4.4. Faults and errors

The existence of faults and design errors in any human engineered system, including high-speed networks, is only logical. Failures in the processors, switch fabrics, and links are likely. For high-speed network designs of the future, although resistance to failures will be high, the loss of cells and the consequent damage from any failure is likely to be very costly. Wherever possible, automatic and distributed detection of failures, fault tolerance, and self-recovery must be incorporated into the design of future networks. With the asynchronous nature of the faults and errors; discussed earlier in Section 4.2, the task can be especially challenging.

### 4.5. Transmission medium

Although networks exploiting other types of non-EM waves such as sound are clearly conceivable, this paper focuses on EM transmission which invariably requires a medium. Whether it is wires, optical fibers, or free space, the EM transmission medium imposes a finite speed of propagation, which translates into a physical propagation delay. Transmission media also influence the energy required for transport, probability of loss, and the available transmission capacity. Consider, for example, communication between the earth and the moon via two competing media—free space and an optical fiber. The EM propagation speed through free space is $3\times10^8$ m/s while that through the fiber is $2\times10^8$ m/s, thus a single bit by itself is transported faster through free space. However, dispersion and energy dissipation are generally higher in free space than in optical fibers, implying a much higher bandwidth rate accompanied by lower cell loss in the optical link.

### 4.6. Bandwidth/data rate

The ability of a link or medium to transport bits is referred to as bandwidth/data rate and serves as the key to the viability of a network Bandwidth may be viewed as the number of lanes of a wide freeway that permits multiple vehicles to travel along the same route, simultaneously.

### 4.7. Delay

In addition to the physical propagation delay imposed by distance, the limited bandwidth disallows all of the bits of a packet to travel in parallel As a result, some of the bits must travel in sequence, giving rise to transmission delays. There is an additional source of delay which will be discussed subsequently.

### 4.8. Size of the information packet

The information packet is the basic quanta of information transport For the purpose of transport across a link, a packet is viewed as an atomic unit, i.e. all of its bits must be sent as a whole. It cannot be fragmented. Although it is a design parameter, the choice of the packet size always exerts a strong influence on the network performance and, often, on the underlying networking principles. For instance, compared to an IP packet of size up to 1800 bytes, the ATM packet size is set to a paltry 53 bytes to achieve high switching fabric speeds. Since so many more ATM packets would result for a given user message, as opposed to IP packets, it implied one more reason for ATM to introduce a call setup phase which would first establish the route along which all ATM packets would be subsequently transported. The traditional approach to the choice of packet sizes has largely been ad hoc. Although loosely based on the current speed of electronics and available link bandwidths, an evolutionary route to keep up with the advances in electronic switching speeds and higher capacity links, has rarely been explored. Future network designs should undertake theoretical and empirical studies to determine packet sizes for efficient network operation.

### 4.9. Sharing

Although Sections 4.1–4.8 appear to exhaust nearly all of the fundamental attributes of networks, a number of diverse dimensions along which networking has been evolving, such as routing and congestion, appear to defy any connection with the inherent characteristics of networks The missing link is found in the attribute, sharing. Unlike its peers, stated in Sections 4.1–4.8, sharing is not dictated by physics. Rather, it originates in the most primitive objective of a network namely, sharing of the network resources, for the purposes of facilitating interactions between users, maximizing the use of the resources, and economies of scale. Upon introspection, it is obvious that a network dedicated to a single user is a logical contradiction. In any network, sharing assumes the form of exchanging knowledge about the current, dynamic state of the network among the nodes, processing headers or call requests in the processing nodes, processing traffic cells in the switch fabric, and multiplexing traffic cells from different sources along the links. Examples of multiplexing include frequency division multiplexing (FDM), synchronous time division multiplexing (STDM), and asynchronous time division multiplexing (ATDM).

This simple attribute gives birth to a number of other attributes which are, in essence, the missing dimensions.

Under sharing, ‘congestion’ is inevitable since precise knowledge of the state of the geographically distributed network, at any time, is difficult, if not impossible. Also, being asynchronous, users throughout the network may not be controlled precisely by the network to limit the number of injected cells to the maximum traffic carrying capacity of the network. When the combined injected traffic cells exceed the capacity of the network, even temporarily, cells may be lost. Efforts to prevent or minimize cell loss gives rise to the introduction of buffers which, in turn, give rise to cell delays stemming from the temporary storage of the excess cells until the congestion clears. In addition to buffers, network control and management constitute efforts at cell loss control. Thus, congestion is viewed as a fundamental attribute of any shared network.

Sharing causes the geographically-dispersed nodes to execute network related decisions simultaneously. Evidently, 'knowledge of the state of the network' will playa key role in efficient decision-making, elevating it to the level of an important attribute. Ideally, the state of every element of the network should be known to every decision-making element of the network, precisely and instantaneously. This is impossible, given the latency stemming from the finite propagation speed of EM transmission. In general, the state information may be transmitted over the signaling network and, higher the accuracy and quality of the state information, better the network related decisions.

Sharing also necessitates the employment of switching which, in turn, requires unique addresses to be assigned to users. The idea of guiding a traffic cell from source to destination along the network is termed 'routing' and constitutes an important attribute. Clearly, routing will bear significant impact on network performance.

The need to realize efficient sharing of the network resources among the users motivates the introduction of the 'signaling' attribute. Under this property, the network must coordinate through the constituent nodes, the reservation, allocation, and supervision of the geographically distributed resources, to ensure that the users' messages reach their destinations. While signaling is visibly apparent in telephony, ISDN, and ATM networks, in IP networks, signaling assumes a degenerate form where the signal, in the form of the IP header, and the payload are both encapsulated in a single IP packet.

Sharing also implies equitable distribution of opportunities among the users to utilize the network. This important requirement is labeled the 'fairness' attribute. Fairness does not exclude the selective treatment of specific user traffic as dictated by need, priority, tariff base, and other considerations.

With sharing being a primary motivation, as time progresses, the number of users is likely to increase, necessitating an increase in the number of nodes. A chief concern is whether, under these circumstances, the network performance will deteriorate dramatically, rendering it less than useful. A desirable goal is a network design such that its performance remains relatively undiminished with an increase in the number of nodes and users. This property is encapsulated through 'scalability.' The difficulty with scalability in networking had been addressed earlier in Section 2.

The goal of sharing the network resources among a maximal number of users gives rise to the need to transport every individual user's traffic securely. The prevention of any unwanted interference, deliberate or unintentional, of a user's traffic, during its transport through the network, either from other users or the network itself, is encapsulated by the 'security' attribute.

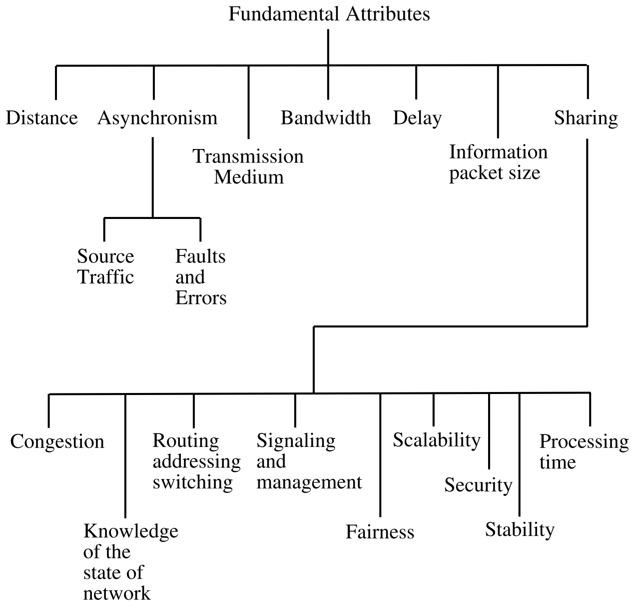

Fig. 1. Fundamental attributes of high-speed networks

The 'stability' attribute refers to the property that the network must resist capitulating into instability when a significant number of users, sharing the network, introduce a sustained excessive number of traffic cells, beyond the design capacity of the network. Instability may also be induced by other factors such as abrupt changes to a network's constituent elements.

The sharing attribute causes the finite computational intelligence resource to be multiplexed among multiple competing users. Thus, 'processing time' is an important, realistic attribute.

Fig. 1 constitutes a graphical presentation of the fundamental attributes of networks.

## 5. Conclusions

This paper has presented a simple yet comprehensive understanding of networking through a novel definition—high-speediness, denoted by the symbol '$s$', which reflects the high-speed nature of networks, two fundamental observations that underlie every network design, and the identification of the fundamental attributes of networks. In addition to the inherent pedagogical value, high-speediness may be used to classify and compare existing networks. The paper has computed the '$s$' values for important past and present networks. The high-speediness value may also constitute a desirable, target network operating point, which the network provider may choose to sustain during network operation by imposing suitable controls. The fundamental attributes represent a holistic view of network, revealing the most important issues and how they are linked to each other.

Together with the attributes, high-speediness promises to serve as a meaningful guide to the design of future high-speed networks. The authors are presently focused on developing practical approaches to determining network operating points.

## References

[1] B.W. Abeysundara, A.E. Kamal, High-speed local area networks and their performance: a survey, ACM Computing Surveys 23 (2) (1991) 221–264.

[2] W. Stallings, ISDN and Broadband ISDN with Frame Relay and ATM, Prentice Hall Publishers, Englewood Cliffs, NJ, 1995.

[3] L. Kleinrock, The latency/bandwidth tradeoff in gigabit networks, IEEE Communications Magazine 30 (4) (1992) 36–40.

[4] C.A. Thekkath, H.M. Levy, Limits to low-latency communication on high-speed networks, ACM Transactions on Computer Systems 11 (2) (1993) 179–203.

[5] C. Bohm, P. Lidgren, L. Ramfelt, P. Sjodin, The DTM Gigabit Network, Journal of High-speed Networks31994; 109–126.

[6] I. Chlamtac, W. Franta, Rationale, directions, and issues surrounding high-speed networks, Proceedings of the IEEE 78 (1) (1990) 94–120.

[7] N.F. Maxemchuk, Dispersity routing in high-speed networks, Computer Networks and ISDN Systems 25 (1993) 645–661.

[8] M. Gerla, High-speed local area networks, IEEE Spectrum 1991; 26–31.

[9] I. Cidon, I. Gopal, S. Kutten, New models and algorithms for future networks, IEEE Transactions on Information Theory 41 (3) (1995) 769–780.

[10] R. Bolla, F. Davoli, A.D. Febbraro, Distributed coordinated deflection routing algorithms for high-speed networks, Journal of High-speed Networks 4 (1995) 287–297.

[11] A. Mink, Y. Fouquet, S. Wakid, Hardware measurement techniques for high-speed networks, Journal of High-speed Networks 3 (1994) 187–207.

[12] Y. Oie, T. Suda, M. Murata, H. Miyahara, Survey of switching techniques in high-speed networks and their performance, International Journal of Satellite Communications 9 (1991) 285–303.

[13] K.W. Fendick, D. Mitra, I. Mitrani, M.A. Rodrigues, J.B. Seery, A. Weiss, An approach to high-performance, high-speed data networks, IEEE Communications Magazine 1991; 74–82.

[14] J. Heinaenen, Review of backbone technologies, Computer Networks and ISDN Systems 21 (1991) 239–245.

[15] Vom Verschwinden der Ferne,. Telekommunikation und Kunst. Zur Ausstellung im Deutschen Postmuseum Frankfurt am Main, ISBN 3-7701-2725-0.

[16] N. Utamaphethai, S. Ghosh, DICAF, a high performance, distributed, scalable architecture for ivhs utilizing a continuous function—congestion measure, IEEE Computer 31 (3) (1998) 78–84.

[17] S. Ghosh, T.S. Lee, Intelligent Transportation Networks: New Principles and Architectures, CRC Press, New York, 2000.

[18] Q. Razouqi, S.-S. Joo, S. Ghosh, Performance analysis of fuzzy thresholding-based buffer management for a large-scale cell-switching network, IEEE Transactions on Fuzzy Systems 8 (4) (2000) 425–441.