

Communications and Technology in the Information Age

Robert W. Lucky
Bellcore

Technology has been the enabling and driving force of the information age. It has created the opportunities, challenges, and problems that permeate this book. The purpose of this chapter is to discuss current trends in the technology of computer communications, and the implications of these trends on the world's telecommunications infrastructure.

It is difficult to write about technology in the abstract, as if it were something to be studied under a microscope, isolated, and grown in controlled cultures. More and more we have moved from the conception of technology as an independent thing to that of technology as an unfolding process, inextricably entangled with the social and economic fabrics of life.

The historic conception of technology as an isolated force was exemplified by the motto of the Century of Progress International Exposition held in 1933 in Chicago. This showplace of science and technology had a slogan that today sounds to us like a time capsule from a past civilization:

"Science finds -- industry applies -- man conforms"

How quaint and simplistic this now seems! Certainly it is still true that *science finds*, although that phrase seems to suggest an aimless serendipity that today is less in evidence. Science finds where scientists are looking, and scientists look for the most part where they are paid to look.

It is also true that *industry applies*. However, the implication seems to be that industry has a compulsion to apply whatever science has found. We now realize that industry applications are not always a direct consequence of the findings of science. Industry chooses what it wishes to apply, and puts a spin on its applications -- choices and spins dictated of necessity by their profit motive. Furthermore, industry often decides what it wishes science to find; these are not necessarily independent events.

The biggest disconnect with this 1933 slogan and today's environment is, of course, the last phrase -- *man conforms*. What an antiquated phrase! The very idea today seems like a challenge. We do not conform! Technology offers many possibilities. We *decide*.

However, as quaint as the motto now sounds, there is a germ of underlying truth here. Technology cannot be undone, and there are many concepts of the digital

world that force us inevitably to change our behavior, our laws, and ultimately, ourselves.

Technology Evolution

Billions of dollars of investment ride on being able to anticipate technological change and to understand its implications. Yet there seems to have been very little progress in our ability to make such predictions. Nearly every major technological revolution has been missed by the experts of the previous generation. Partly this is because of the vested interest that technologists have in their own existing areas of expertise, but it is also due to the chaotic nature of the market in which technology operates. Technological progress itself is inextricably interwoven with social and economic factors that often determine which technologies succeed and which fail.

It is often assumed that we cannot predict the future in technology because of the disruptions caused by unexpected, revolutionary inventions. Although these disruptive inventions do happen, the frequency is probably less often than people assume. Almost all the technology in telecommunications and computing today was known a decade ago. It is often said that "technology isn't the problem" or "we have all the technology we need." What is lacking is the ability to project the market acceptance of technologies and any semblance of insight into the future applications of these technologies.

Technology itself often follows a predictable path of progress. In telecommunications the capacity of optical fibers follows a steady exponential expansion, while in computers and electronics the size of memory, the clock speed, and the general capability of integrated circuit chips are entirely predictable.

Moore's Law

The single most important thing to know about the evolution of technology is Moore's Law. Most readers will already be familiar with this "law." However, it is still true today that the best of industry executives, engineers, and scientists fail to account for the enormous implications of this central concept.

Gordon Moore, a founder of Intel Corporation, observed in 1965 that the trend in the fabrication of solid state devices was for the dimensions of transistors to shrink by a factor of two every 18 months. Put simply, electronics doubles its power for a given cost every year and a half.

In the three decades since Moore made his observation the industry has followed his prediction almost exactly. Many learned papers have been written during that period predicting the forthcoming end of this trend, but it continues unabated today. Papers projecting the end are still being written, accompanied with

impressive physical, mathematical, and economic reasons why this rate of progress cannot continue. Yet it does.

Moore's Law is not a "law" of the physical world. It is merely an observation of industry behavior. It says that things in electronics get better, that they get better exponentially, and that this happens very fast. Some, even Gordon Moore himself, have conjectured that this is simply a self-fulfilling prophecy. Since every corporation knows that progress must happen at a certain rate, they maintain that rate for fear of being left behind.

It is also possible that Moore's Law is much broader than it appears. Possibly it applies to all of technology, and has applied for centuries while we were unaware of its consequences or mechanisms. Perhaps it was only possible to be explicit about technological change in 1965 because the size of transistors gave us for the first time a quantitative measure of progress. If this is so, then we are embedded in an expanding universe of technology, where the dimensions of the world about us are forever changing in an exponential fashion.

The notion of exponential change is deceptively hard to understand intuitively. All of us are accustomed to linear projection. We seem to view the world through linear glasses -- if something grows by a certain amount this year, it will grow an equal amount the next year. But according to Moore's Law, electronics that is twice as effective in a year and a half will be sixteen times as effective in 6 years and over a thousand times as effective in 15 years. This implies periodic overthrows of everything we know. An executive in the telecommunications industry recently said that the problem he confronted was that the "mean time between decisions exceeded the mean time between surprises." Moore's Law guarantees the frequency of surprises.

Metcalfe's Law -- Network Externalities

There is another "law" that affects the introduction of new technology -- this time in an inhibiting fashion. Metcalfe's Law, also known to economists generally as the principle of network externalities, applies when the value of a new communications service depends on how many other users have adopted this service. If this is the case, then the early adopters of a given service or product are disincented, since the value they would obtain is very small in the absence of other users. In this situation innovation is often throttled.

Metcalfe's law often applies to communications services. A classic example, of course, is the videotelephone. There is no value in having the first videotelephone, and it only acquires value slowly as the population of users increases. If there are n users at a given time, then there are $n(n-1)$ possible one-way connections. Thus the value grows as the square of the number of users. The value starts slowly, then reaches some point where it begins to rise

rapidly. It seems as if there needs to be a critical mass for takeoff, and that there is no way to achieve that critical mass, given the burden on initial subscribers.

Metcalf's Law has defeated many technological possibilities, left stillborn at the starting gate of market penetration. Nonetheless, there are important examples of breakthroughs. For example, facsimile became a market success, but only after decades of technological viability. Even so, facsimile is a complex story, involving the evolution of standards, the inevitable progress of electronics, the equally-inevitable progress in the efficiency of signal-processing algorithms, and the rise of the business need for messaging services.

Moore's and Metcalfe's laws make an interesting pair. In the communications field Moore's law guarantees the rise of capabilities, while Metcalfe's law inhibits them from happening. Devices that appear to have little intrinsic value without the existence of a large networked community continue to diminish in cost themselves until they reach the point where the value and cost are commensurate. Thus Moore's Law in time can overcome Metcalfe's Law.

The Evolution of the World Wide Web

The most important case study in communications technology is the emergence of the World Wide Web. This revolutionary concept seemed to spring from nothingness into global ubiquity within the span of only two years. Yet its development was completely unforeseen in the industry – an industry that had pursued successive long and fruitless visions of videotelephony, home information systems, and video-on-demand, and had spent decades in the development of ISDN with no apparent application. It now seems incredible that no one had foreseen the emergence of the Web, but except for intimations in William Gibson's science fiction novel *Neuromancer*, there is no mention in either scientific literature or in popular fiction of this idea prior to its meteoric rise to popularity.

There is a popular notion that all technologies take 25 years from ideation to ubiquity. This has been true of radio, television, telephony, and many other technologies prevalent in everyday life. How, then, did the Web achieve such ubiquity in only a few years?

Well, the historians argue, the Web relied on the Internet, which in turn was enabled by the widespread adoption of personal computers. Surely this took 25 years. We might even carry this further. The personal computer would not have been possible without the microprocessor, which depended on the integrated circuit evolution, which itself evolved from the invention of the transistor, and so forth. By such arguments nearly every development, it seems, could be traced back to antiquity.

Although the argument about the origin and length of gestation seems an exercise in futility, the important point is that many revolutions are enabled by a confluence of events. The seed of the revolution may not seem to lie in any individual trend, but in the timely meeting of two or more seemingly-unrelated trends. In the case of the World Wide Web the prevalence of PCs and the growing ubiquity of the Internet formed an explosive mixture ready to ignite. Perhaps no invention was really even required. The world was ready -- it was time for the Web.

While this physical infrastructure was forming in the world's networks and on the desktops of users, there was a parallel evolution of standards for the display and transmission of graphical information. HTML, the hypertext markup language, and HTTP, hypertext transmission protocol, were unknown acronyms to the majority of technical people, let alone the lay public. But the definition of these standards that would enable the computers and networks to exchange rich mixtures of text and pictures was taking shape in Switzerland at the physics laboratory CERN, where Tim Berners-Lee was the principle champion.

The role of standards in today's information environment is critical, but often unpredictable. What is really important is that many users agree on doing something exactly the same way, so that everyone achieves the benefits of interoperability with everyone else. It is exactly the same concept of network externalities that is at work in Metcalfe's law. An international standard can stimulate the market adoption of a particular approach, but it can also be ignored by the market. Unless users adopt a standard it is like the proverbial tree falling in the forest without a sound. Standards are, for the most part, advisory. User coalitions or powerful corporations can force their own standards in a fascinating and ever-changing multi-player game. Moreover, de facto standards often emerge from the marketplace itself.

So in the middle 1980s there was a prevalent physical infrastructure with latent capabilities and an abstract agreement on standards for graphics. One more development and two brilliant marketing ideas were required to jumpstart the Web. The development was that of Mosaic at the National Center for Supercomputing Applications at the University of Illinois. Mosaic was the first browser, a type of program now known throughout the world for providing a simple point-and-click user interface to distributed information. Following the initial versions of Mosaic from NCSA, commercial browsers were popularized by Netscape and Microsoft.

The revolutionary marketing ideas needed for the Web now seem obvious and ordinary. A decade ago, however, they were not at all obvious. One idea was to enable individual *users* to provide the content for the Web. The other idea was to give browsers free to everyone. Between these ideas, Metcalfe's Law was overcome. Even though browsers initially had almost no value, since there were no pages to browse, they could be obtained electronically at no cost. The price

was directly related to the value. Thus browsers spread rapidly, just as their value began to build with the accumulation of web pages.

Allowing the users to provide content was counter to every idea that had been held by industry. The telecommunications and computer industries had tried for a decade to develop and market remote access to information and entertainment held in centralized databases. This was the cornerstone of what were called "home information systems" that were given trials in many cities during the 1970s and 1980s. Later, the vision pursued by the industry was that of video-on-demand -- the dream of providing access to every movie and television show ever made, like a giant video rental store, over a cable or telephone line. Virtually every large telecommunications company had trials and plans for video-on-demand, and the central multi-media servers required for content storage were being developed by Microsoft, Oracle, and others.

The Web exemplifies some powerful current trends -- the empowerment of users, geographically-distributed content, distributed intelligence, and intelligence and control at the periphery of the network. Another principle is that of open, standard interfaces that allow users and third parties to build new applications and capabilities upon a standardized infrastructure.

It is hard to criticize industry for pursuing the centralized approach. Imagine proposing the Web to a corporate board in 1985, and describing how browsers would be given away free, and how industry would depend upon the users to provide whatever content might appear. Even today many corporations wonder and worry about the business model for the Web, and few are making any profits at all.

The Evolution of Digital Telephony

Ironically, in 1975 when Alexander Graham Bell invented the telephone, there was already a nationwide, digital network in place. It was owned by Western Union, and it was the telegraph system.

Bell's invention was essentially the use of analog waves for the transmission of speech, rather than the dots and dashes used in Morse code for the transmission of text. The pressure waves caused by human speech in air were converted by a carbon microphone into an electrical wave, which was then transmitted over copper wires to a distant location.

Over the next century Bell's analog telephone replaced the telegraph as the chosen instrument of human communication. The telephone wires spanned the nation, a giant corporation came together and reigned as a monopoly, and the analog transmission and switching were progressively refined. The transmission medium, which began as copper wires, became microwave carrier, then coaxial cable, then partially satellite, and finally became primarily digital optical fiber.

The switching, which began as manual patchcords, became rotary step-by-step switches, then electronically-controlled relays, and finally centrally-controlled digital switches.

The analog waves so naturally congruent to human speech were replaced gradually by their digital representation. With today's rise of the Internet, another revolution is now taking place in which the medium of exchange is not the bit itself, but a standardized envelope of bits known as a packet. In the succeeding sections we will follow the progress of the telephone infrastructure from waves to bits to packets. Why is the plant digital? Why packets? These are the questions we will seek to answer.

The Conception of Digital Telephony

The first big technological revolution after the telephone had spanned the nation was the digitization of the network. Curiously, the telephone infrastructure began its digitization about 1960, considerably before the concepts of the information age became popularly understood. When the world needed a digital infrastructure in later decades, the telephone system was already there -- not because telephone engineers had foreseen the need for data, but rather because they had wanted a cheaper analog network for voice.

Just as the natural voice waveform is analog, so intrinsically are all transmission media. It takes an artificial discipline to impose a digital signal upon an analog world. As late as the mid-1950s, there seemed no reason to do this. The network was carefully crafted to convey a 3 kHz analog signal as the universal medium of exchange. The bandwidth of 3kHz was calculated to be the smallest range of frequencies that would enable speech to seem perceptually unimpaired. This range of frequencies, roughly from 300 to 3000 Hz, would on a piano be the three octaves beginning at middle-C. Any more bandwidth was wasteful, while any less hurt the intelligibility of the speech. The scientific literature of that day had numerous studies of the "mean opinion score" of speech over a telephone connection, which was designed to be about a "4," where a "5" was equivalent to face-to-face conversation.

Pulse Code Modulation (PCM) was invented by Reeves of IT&T in 1939. In PCM the analog speech signal is converted to a stream of bits by sampling the signal at periodic intervals, and then representing the samples as digital approximations. As we shall see presently, the practice is to convert an analog voice signal into a 64 kilobit per second digital stream. In so doing, the bandwidth required for transmission becomes greatly expanded, and so engineers had little incentive to implement PCM for many years. The only application that seemed to call for PCM was that of encrypted speech during the war years immediately following Reeve's invention.

Analog to Digital Conversion

There is a famous theorem in communications technology, called the sampling theorem (Nyquist), that states that a bandlimited signal may be reconstructed exactly from samples taken at a rate of twice the highest frequency. So that if we assume on the safe side that speech has a 4KHz bandwidth, then a sampling rate of 8000 samples per second could be used to reconstruct exactly the speech signal.

It may seem curious to the uninitiated that a signal can be reconstructed exactly from little snippets taken at regular intervals. Hasn't something been lost? What about the values of the signal in between the sampling instants?

The key here is the assumption that the signal is truly bandlimited. This means that the signal is constrained as to how fast it can change, since there are no "high frequencies" present. This implies a smoothness or predictability that enables the signal to be extrapolated between the sampling instants. In a sense this is a mathematical abstraction, since in theory no signal can be simultaneously time-limited and bandlimited. No matter -- speech can be well reconstructed from samples taken at a rate of 8000 per second. There is no perceptible difference between the reconstructed and original signals.

After the speech signal has been sampled, it has been translated into a stream of numbers. The next step is the approximation of these numbers as a sequence of bits. For example, the first sample might be 1.32956, or some such number. Transmitting the number exactly would require an infinite sequence of bits, but obviously no one is going to hear the difference if we truncate it somewhat. Even in everyday speech, the ambient noise in the room sets a lower threshold on how exactly the speech signal can be perceived.

Engineers determined that 8 bits were sufficient for intelligibility of speech. Thus every sample could ideally be represented as one of 256 possible values. It was found that the representation was better served by having a logarithmic spread of these values, so as to compress the extreme values and expand the smaller values. Again, the engineering of the network infrastructure was based entirely on the perceptual quality of speech transmission. In pulse code modulation the analog voice signal is converted to a stream of 8-bit numbers, occurring 8000 times a second, for a total of 64,000 bits per second. This is the standard digital representation of speech that is used in telephony today throughout the world.

The Philosophy of PCM

Now it is important to understand the implications of this conversion into a digital stream. We started with an analog signal that could be transmitted in a 3kHz bandwidth. How much bandwidth does it now take to transmit the equivalent 64,000 bit per second digital stream? As a rough rule of thumb, it is possible to transmit a digital stream in a bandwidth of about half the bit rate. Thus the digital

equivalent of the voice might require a bandwidth of 32kHz – almost ten times what the original voice required! So why should anyone want to convert speech to digital format?

In the mid-1950s the answer to this question slowly permeated the scientific community. The reason why it was a good idea to convert the speech to bits was because, while an analog signal is fragile, bits are almost indestructible. This is one of the central tenets of the digital world. Since we know that a bit can only be a “1” or a “0”, a “degraded” bit can be restored to its original perfection. In contrast, when an analog wave is degraded, there is no notion of how it can be restored.

Waves come and go, while bits are forever.

It seems miraculous that in the telephone network of the 1950s, voice signals were transmitted in analog form across the continent and around the world. Every source of noise or distortion during transmission would accumulate on this long path. Only the most careful design of equipment was able to convey intelligible speech at the other end of this long and tortuous pipe. If, on the other hand, the speech was converted to bits – even at the cost of greatly increased bandwidth – it was no longer necessary to carry the signal perfectly for long distances. Instead, it could be restored periodically before it had been seriously degraded. Again and again, the bit stream could be regenerated to its perfect original state.

The Digitization of the Telephone Infrastructure

In the first digital carrier system, commercialized in 1961, the digital voice signals were restored about every mile and a half. What a difference! Instead of having to traverse thousands of miles, the signal only had to travel a mile or so. Because it only had to negotiate such a short distance, many more voice channels could be transmitted over the same pair of wires – hence the economic argument for digitization.

The new digital carrier systems were rapidly installed in metropolitan areas, so that in the late 1970s most of the intra-city transmission was digital. However, both the local loop and the long distance systems were still analog. The local loop – the last mile or two to the home – consisted of pairs of copper wires, while intercity transmission was accomplished by modulated analog signals on coaxial cables and microwave carrier systems.

This was the point where one of the great disruptive inventions took place, when Corning announced that it had made an optical fiber sufficiently transparent to carry lightwave signals for about a mile. Very shortly thereafter the telecommunications industry abandoned plans for more advanced microwave transmission systems, and concentrated fully on optical transmission. There

were two important ramifications of this breakthrough. One was the startling promise of capacity that the fibers held – more than anyone at the time thought would ever be needed for any conceivable use of telecommunications. The other ramification of fibers was that they were inherently digital. The quality of analog transmission was so distorted that it seemed impossible to convey any degree of accuracy unless the signal was in the robust digital format. Incredibly, it was only a few short years later that AT&T wrote off its entire analog plant. The long distance network, thanks to fiber, had been suddenly digitized.

Thus the economics of voice telephony resulted in the central portion of the telecommunications network being digitized. This left, however, the local connections to the consumer still in analog format. The analog signals generated by the microphone in the telephone handset were transmitted over a pair of copper wires to the serving central office, usually a distance of one to three miles. At the central office the analog signal would be converted to a 64 kilobit per second digital signal, which would subsequently be interleaved in time with other digital streams for transmission over the long haul facilities.

Although the end portion of the local connection to the home is essentially always analog, there has been a great penetration of a local digital carrier system, called subscriber loop carrier, which carries typically 24 interleaved digital streams into a neighborhood. At that neighborhood point the streams are converted into analog, and the copper wires fan out to serve a number of homes in the area. The effect is as if the central office had been moved closer to the home, but the ultimate connection at the subscriber is still an analog voice channel.

The Packet Revolution

The need for computer communications began to emerge in the late 1950s. At that time the goal was access to time-shared central computers. The telephone network had been designed solely for the transport of analog voice signals, so the only way to transport the computer data was to design apparatus to convert the data into voice-like analog signals. The units that did this conversion became known as modems. In this era only the Bell System was permitted to design and deploy equipment connected to the network, so the first modems were manufactured by Western Electric, and had a transmission speed of up to 300 bits per second. In the early 1960s progressively faster modems using better modulation formats were designed that provided transmission at 1200 and then 2400 bits per second.

Thus in the mid-1960s an increasing number of modems were being used to transmit computer data over the voice telephone network. At the same time, the interior network itself was being changed over to digital format. These two trends continued, so that in 1990 there were millions of modems connected to a network that was essentially entirely digital in its core. In spite of the digital network, the

modems were still necessary to negotiate the two ends of the connection, which remain analog even today.

Consider when a home PC is connected to an Internet Service Provider. The bits generated by the PC for transmission are converted by the modem to an analog signal with voice-like bandwidth. This is done by very sophisticated techniques enabling transmission speeds of up to 33 kilobits per second. The modem's signal then travels about two miles to the serving central office, where it is sampled, quantized, and converted to a 64 kilobit per second digital stream. This stream is interleaved with many other streams, and transmitted as light pulses for long distances. Arriving at the distant central office, it is reconverted into an analog signal that closely resembles the original signal generated by the sending modem at the PC. This analog signal is sent the last mile or so to the Internet Service Provider, where it is demodulated to regain the 33 kilobit per second digital information. It is ironic that the modems have to work so hard to realize only a fraction of the capacity of the digital stream that carries the signal through the long haul network.

The newer technology of 56 kilobit modems is much more efficient in its use of the 64 kilobit network transmission. These modems rely on the Internet Service Provider having a digital connection directly to their computers, so that only the home PC end of the network is analog. In the downstream direction – from the ISP to the PC – a simpler and faster method of transmission is used that couples the high speed data from the ISP directly onto the digital carrier.

While the access ends of an Internet connection are really a kluge of overlays to old technology, the backbone of the Internet is entirely different. It is there that a new paradigm for communications has taken shape – one that threatens to engulf and overwhelm the entire telecommunications infrastructure of the world. The Internet cloud itself – between ISPs and host computers – relies on the relaying of digital packets from one network node to another. The question now being asked everywhere is: Why shouldn't the entire network work this way?

Packet Switching

The conventional telephone network, called the PSTN (Public Switched Telephone Network), is analogous to the child's toy telephone where two tin cans are connected by a string. When a call is originated, the network predetermines a path that can be used to connect the caller A with called party B. Circuit switches are set so as to provide a full time, two-way connection between the parties. Within the network each party has a 64 kilobit per second dedicated path to the other. The bits flow along this path whether or not the parties are actually talking to each other at the moment. It is as if the two parties had indeed installed a very long string between their tin can telephones.

Packet switching, originated by Paul Baran of Rand in 1964, uses a radically different philosophy. The analogy here is much like mailing information on postcards using the ordinary post office mail system. Each postcard has a place for an address, a return address, and a space for information. It is in effect an envelope of bits. If the message to be sent is larger than the space available on a single postcard, then it is broken up and sent in a series of postcards, each of which is numbered so that the recipient can reassemble the entire message. As the postcard (packet) is routed through the network, the address is read at each switching node, and placed in a queue for transmission to another node that is closer to the intended destination.

In packet switching there is nothing similar to the string connecting the tin cans. There is no predetermined, full time path between the sender and receiver. The packets flow autonomously, each in ignorance of the progress of its fellow packets, each trying to wend its way towards its intended destination.

The efficiency and flexibility advantages of packet switching over circuit switching in the case of computer communications are large and obvious. Only as much of the transmission and switching capacity as is needed for a given communication is used. When no information is being sent, there are no packets, and other connections can use the common facilities. When, on the other hand, a great deal of information is required to be sent by a particular connection, then it can surge a great many packets onto the network to meet its sudden demand.

In packet switching the similarities to transportation systems are inescapable. Each packet of bits is like an automobile entering a highway. Network nodes are like interchanges. Rush hours and traffic jams are possible. Queues for entrance to the highway or to a particular exit can develop and build. None of this can happen, of course, with the circuit-switched PSTN. If you have dialtone, then you are assigned a path and guaranteed your own private road. It is, however, an expensive and inflexible road that you own for the duration of your call.

Rules of the Road for Packets -- TCP/IP

When the Department of Defense funded the development of ARPANET in the early 1970s to connect between large computer sites, the academic community (although corporations -- Bolt Beranek and Newman in particular -- also played key roles) designed the network from the beginning using packet switching. The early protocol that defined the rules for traffic flow of the packets was replaced in 1975 with a protocol known as TCP/IP (Transmission Control Protocol/Internet Protocol) written by Robert Kahn and Vinton Cerf. For many years thereafter the name "TCP/IP" was unknown to all but a few hundred academic designers. Today TCP/IP is the center of a multi-billion dollar industry, and dominates networking technology everywhere. The last frontier is whether it takes over from the PSTN for voice telephony.

There is a particular genius in the conception of IP, seen clearly in the retrospective vision from a quarter century after its origin. IP is the protocol for handling packets inside the network, whereas TCP is the end-to-end protocol that resides at the user and host sites at the periphery of the network. Crudely speaking, TCP cleans up any mess left by the inside protocol, IP. (Obviously, it does more than that!) The genius in IP is its utter simplicity: It defines the minimum set of features necessary to connect packets at a network node -- no more, no less. Because of this simplicity, it has been possible through the years to use IP for many more purposes than were ever envisioned when it was designed, while maintaining interconnectivity between disparate networks.

A view that designers have of today's TCP/IP network is that of an hourglass. The wide bottom of the hourglass consists of all the physical media used for electrical and photonic transmission of bits – fibers, copper wires, microwave signals, etc. The wide top of the hourglass consists of all the applications that require communications – speech, video, email, file transfer, etc. The center of the hourglass – the narrow waist – is the restriction to IP inside the network. To pass this through this point, all signals must be formatted in IP packets, no matter what application they represent or what media on which they flow.

There isn't very much to an IP packet. Basically, it is a blank postcard with designated spaces for intended and return addresses, as well as an expandable space for the information bits representing the payload. There are some other bits in the packet which are used for special purposes, such as "time to live" indicators that ensure packets get thrown away before they clog the network like dead letters. But IP packets can get lost, be ruined with errors, get out of sequence, get misdirected -- whatever. IP itself doesn't care. Somebody else will have to worry about this later. This is a job for an end-to-end protocol such as TCP. If such features had been built into IP, as they were with some earlier protocols, it would have been at a cost of efficiency and flexibility.

The addressing scheme used in the Internet also differs conceptually from that used in the PSTN. In telephony, for historical reasons, the telephone number is related to a *place*, rather than a person. In the access network and switching system, the telephone number is something that is physically *wired*. Moreover, the telephone number is confined to something that can be dialed by the old rotary phones – usually a meaningless sequence of digits.

In the Internet, in contrast to the PSTN, the domain name address is alphanumeric, mnemonic, and not at all coupled to a geographical location. Adding a new customer, or changing an address, is a process of registration in a database, not rewiring. When a user sends an email or requests a web page, say www.bellcore.com, it is first necessary to translate this domain name address to an IP address that locates the recipient in the physical hierarchy of systems that comprise the vast Internet. The domain name is forwarded up the hierarchy of

domain name servers, asking successive network servers if they know the translation to the actual address being requested. Ultimately, if lower level name servers are unable to supply the data, the request reaches the root server for the top level domain (.com in this example), which by definition knows all the next-lower translations within its domain. The network address, a sequence of digits, is returned to the user, and is applied to succeeding packets in the email or web page.

The Rise of the Internet

In spite of the genius of IP, in a different world another protocol might have triumphed in the market. Other protocols were formulated and implemented through the years, including even a protocol adopted as a standard by the US Department of Defense, called GOSIP. The success of TCP/IP has some similarity to that of the Web, in that early versions of it were given away free, and were integrated into the UNIX operating systems, which had achieved a great deal of academic popularity. UNIX itself achieved much of that popularity by being given free to universities by AT&T. Perhaps it comes down to the sign hung in one executive's office: *Deployment wins*. Through LANs and the Internet, TCP/IP has won the battle of deployment.

With the meteoric rise of the Internet, the traffic on the telephone network that represents data, rather than speech, has been rising exponentially. Unfortunately, there is no reliable way to estimate the actual data traffic. However, it is believed that data and voice traffic are now about equal. While the voice traffic grows at the historical rate of 3-6%, the data traffic is growing at the alarming rate of about 300% annually. If the current growth rates maintain, the data traffic will overwhelm the speech traffic in the first years of the 21st century. The great majority of this data traffic originates in the TCP/IP protocol.

The Internet today rides as an overlay to the pre-existing voice telephone network. As such, it has been in some ways a parasite, living off the nourishment of its host. However, with the sudden and dramatic rise of Internet users and traffic, a new possibility has arisen. Is it possible that the parasite could eat the host? With this fear in mind, there is much discussion today about whether the telecommunications infrastructure should be entirely converted to the Internet model. This would be a complete reversal of the current paradigm. Instead of data riding on a voice network, we would have voice riding on a network designed for data.

The appeal of a packet-switched, Internet model is best dramatized by the comparison in size and economics between a packet router and a typical end office circuit switch. The router looks like a desktop computer, while the circuit switch looks more like a small building. There is a visceral appeal to the former. Moreover, the idea of having one common format – the data packet – as the only traffic type in the network has both a practical and philosophical attraction. But

the immediate question to resolve in the case of packet switching is its ability to handle voice traffic.

Sending Voice by Packets

Circuit switching seems naturally adapted to voice. The speech signal is continuous and real time, and goes only in general to a single pre-determined destination. In contrast, packet switching seems ill adapted to speech. The packets are asynchronous, unreliable, and suffer variable amounts of transmission delay. They can even arrive out of order at the receiver. Even today many telephone engineers argue that carrying voice by data packets simply doesn't make sense. Why do such an unnatural thing?

Today there is a small amount of voice traffic on the Internet. For the most part this is hobby traffic, like the Morse code transmissions of radio amateurs of a previous era. The quality of transmission is poor, and the ease of use is relatively awkward. Aside from the hobby appeal, there is only one compelling reason for voice on Internet today, and that is the evasion of the existing tariff structure, which imposes a usage-based access fee on all voice traffic except that from the Internet. This is an artificial inducement that cuts the price of communications by about half for domestic traffic and much more for international calls. While this might not be a lasting advantage, it has served to stimulate the technology of voice on Internet.

Even though the reasons for voice over packets today are artificial, there are reasons why in the future packet switching might be a good way to carry voice. First, packet switching offers a considerable multiplexing advantage over circuit switching, possibly by as much as a factor of 30. For example, voice conversations typically are active in only one direction at a time, saving a factor of 2 in bandwidth for the use of as-needed packet switching. Additionally, there are pauses in conversation that further save bandwidth.

The big savings in sending voice by packets, however, is in the efficiency of speech coding itself. Since the voice standard of 64 kilobits per second was set for the telephone network, there have been almost 40 years of progress in speech coding. Today excellent quality speech can be coded at about 8 kilobits per second -- a factor of 8 more efficient than was possible in 1960. Realizing the efficiency of modern speech coding is possible on the Internet because the coding is not fixed within the network, but can be arbitrarily chosen by the users at the periphery of the network -- an important principle in the philosophy of the Internet.

In addition to bandwidth efficiency and less expensive switching, there are potential advantages in the flexibility of packet switching for voice. Since the users can choose their own coding, it is even possible to send high-fidelity, stereo speech. A more compelling advantage, however, is in the integration of

speech with the data environment. Multimedia formats can be integrated with speech, and signaling (network control, such as call setup) can be embedded into the Internet connection carrying the speech. The entire environment can be integrated on the work desktop, and Internet domain name addresses can be used interchangeably with ordinary telephone numbers.

While there are certainly advantages to the integration of the voice and data networks, the original objections to sending voice with packets need to be discussed. It is still true that speech is continuous, while packets are sporadic. It is also true that the reliability of packet transmission is a problem for high quality speech. Probably the worst problem is neither of these, but the inherent delays in the transmission of packets over the Internet. We are very sensitive to delays of more than about one-quarter second in speech transmission, as most of us have experienced with satellite telephony. In typical PC-based systems for IP-telephony today the processing delays are considerably more than this critical value.

In spite of the difficulties of handling speech on a packet network, the problems are not insurmountable. A better quality of service on the Internet would help cure the network delay and lost packet difficulties. The delays inherent in today's PC client software for packet voice will be minimized in future systems. A number of companies are now making systems that work directly from one canonical "black" telephone to another, using Internet packet technology within the network. The advantage of this network-based solution is that it eliminates the PC software delay, and permits intelligent integration of the PSTN and Internet.

If today's trends continue, the network of the near future will be entirely packet switched. We will have accomplished the third of three great revolutions. First was the voice telephone network. Connecting the nation took about 50 years. Next was the digital revolution, which required perhaps 25 years. Most recently we have the packet revolution, which measured from the start of ARPANET has also taken about 25 years.

The Quest for Bandwidth

There is an historical trend in the consumption of bandwidth. Whatever bandwidth is available at a given time seems insufficient. Users demand more bandwidth, while communications designers complain that there is a sufficiency for all foreseeable applications. The problem in the past has always been that the actual future applications were not in the foreseeable category. Nonetheless, the trend has always been evident – while we cannot predict the future uses of bandwidth, we know that more will always be required.

The backbone is always the most amenable portion of the network for upgrading. There is an economy of scale that enables technological solution to bandwidth

enhancement. When a user pays \$20 a month to an ISP for Internet service, something like 50 cents or a dollar of this monthly fee goes to the backbone provider. Thus, for example, the backbone investment could be doubled with almost no impact on the price of service.

The backbone traffic is currently growing at an annual rate estimated at 250-1000%. If such a growth were to continue, it would imply that in about six years we would need a network almost a thousand times larger than that today. In other words, today's network would be almost irrelevant to the network in the near future. To some people this seems marvelous, to others impossible. The latter group argues that this growth rate cannot be maintained, that some Malthusian principle will limit the growth.

What are the components of today's traffic growth? First, we should note that the majority of this growth is from enterprise extranets – businesses implementing their own IP networks from leased digital circuits. New businesses and new users join the fray every day. In the Internet itself, the growth in hosts has maintained at 100 percent annual rate for more than a decade. (Recent measurements indicate that this rate may be slowing somewhat, however.) So there are more users, and those users are staying on longer, and are increasingly using applications that require more bandwidth.

Looking at potential increases, we might assume that the number of users continues to double annually. There are about 100 million users of the Internet today. While these users are spending more time on the Internet, in contrast to other activities such as television, the amount of time spent by an individual obviously cannot increase indefinitely. If the average Internet user spends an hour a day now, how much more is possible? Perhaps as much as a factor of four, but more is hard to envision. Meanwhile, the applications might go from casual web browsing to watching full time video – perhaps a bandwidth increase of a few thousand bits per second to a megabit per second. Putting these factors together we might see bandwidth increasing at a 300% annual rate for at least another half dozen years. Moreover, history has shown us that it is dangerous to predict future uses of bandwidth. We have always been wrong in the past, and there is no reason to believe that pattern will not continue.

Increasing Bandwidth in the Network

Inside the network a new technology, called dense wavelength division multiplexing (DWDM) is emerging at exactly the right time to satisfy today's enormous bandwidth demands. DWDM upgrades existing optical fiber transmission systems by allowing multiple wavelengths to be transmitted simultaneously, each wavelength carrying its own high speed data stream. Wavelengths on the fiber are like colors in the visible spectrum, so that we might think of one data stream being carried by "red" light, another by "green", and so

forth. (Since the actual wavelengths involved are in the infrared spectrum, they are not visible.)

In just the last two or three years the number of wavelengths that can be multiplexed on a single fiber has gone from two to 40. In addition the speed of transmission at each wavelength has increased by a factor of 4, going from 2.5 gigabits per second to 10 gigabits per second. Thus there has been almost a factor of 100 increase in the capacity of fiber systems in the last couple of years.

The new fiber technology has enabled new entrants to the long distance telecommunications market to implement entirely new national and worldwide fiber networks quickly and at lower cost than that of existing carriers. Qwest and Level 3 are new companies that have already begun the rewiring of the world with DWDM. Meanwhile, the traditional long distance carriers are upgrading their capacities, but in some cases are hampered by right-of-way issues or their own earlier investment decisions in the conduit space and number of strands allowed for future growth. The Internet traffic explosion has invalidated their previous network planning assumptions.

Aside from the better utilization of the fiber spectrum, bandwidth is also gained by the use of better compression technology. We have already noted that one of the advantages of packet voice is that it can use more modern speech coding technology – 8 kilobits per second, as opposed to 64 kilobits per second in the standard PSTN encoding. In the transmission of images, the JPEG compression of bit-mapped images can save an order of magnitude in the number of bits required. In video transmission MPEG2 has made digital video possible, giving many more channels to direct broadcast video systems. Improvements continue to be made, and there is a kind of Moore's Law curve in compression technology that buys us continually increased bandwidth relative to our needs.

There is much argument about whether in long distance transmission there is a glut of capacity or a shortage. Are we bandwidth rich or bandwidth poor? On the side of rich, we have the huge increase in capacity offered by DWDM and the extensive new fiber networks being installed. On the side of bandwidth poor, we have the 300-1000% annual traffic growth in data, and the emergence of broadband, multimedia applications. So far it seems a close call with capacity and need fairly well balanced. If the balance would begin to tip one way or the other, it would have serious implications to the business models of companies in the telecommunications field.

Broadband Access

The real problem in the bandwidth demand is the access bottleneck. How do we get multi-megabit streams to the home and the small office? Large offices are not an issue, since wherever there is an aggregated demand, as in a corporation, the economics allow the leasing of high capacity fiber. But in the home and small

office regime, there is no economy of scale, and regardless of the capabilities of technology, issues of expense dominate the arguments.

There are many technologies that can be used to enable broadband data to the home, including traditional copper pairs, fiber, coax, wireless, satellites, airships, and packet radio. The very existence of so many alternatives indicates that none of them is a clear winner. There are enthusiastic proponents and investors behind every one of these possibilities, with little likelihood of any shakeout in the near future.

In a pre-planned world fiber would be a logical choice for consumer access. A new fiber installation is possibly no more expensive than using copper, and it would offer a dedicated and almost limitless bandwidth for the future. The problem is that an extensive rewiring of neighborhoods for fiber access would require a large investment with a considerable associated risk and a long period for recovery. If in the meantime a new solution, such as wireless, should emerge, the investors could be stuck with the sunk cost of an unused fiber network.

Disinclined to run fiber directly to the home itself, the telephone companies instead put fiber into neighborhood nodes, and depend on the traditional copper wire pair for the final connection to the home. Depending on the length of this copper pair, it is possible to send data at rates as high as 52 megabits per second. More typically, however, would be a rate of 1.5 megabits per second over a loop length of 2 miles using a modem pair called ADSL (asymmetric digital subscriber loop). A number of different modem technologies are currently being developed with various acronyms. The family name "DSL" or "xDSL" is usually used to refer to the general technology of sending digits over the local copper loop.

DSL technology is the access method of choice currently for the operating telephone companies. Its advantages are a reasonably high data rate, dedicated per-subscriber capacity, and – most importantly – incremental investment. Moreover, since DSL brings packets directly into the central office, the traffic can bypass the existing circuit switch, and be sent to a router. Thus DSL would enable the evolution of the PSTN towards a packet network – an evolution which is greatly complicated by the preponderance of voiceband modems that depend on analog, circuit-switched connections.

Despite these considerable advantages, there are drawbacks for DSL technology. Most of these disadvantages directly bear on the economics. Most importantly, how expensive is it to provision DSL over the great variety of loops currently in existence? Some fraction of customers will be at the end of a loop that is too long or has poor performance for other reasons. Any custom engineering would require expensive attention, and at present there is not enough experience to be able to quantify this cost.

The DSL modems are still rather expensive, and there is no single standard that permits a mass, competitive market to emerge. However, both of these problems should be solved in the near future.

The best alternative to DSL today is the cable modem. The great majority of homes in the United States are already passed by coaxial cable. Modern cable systems have a gigahertz of bandwidth that is shared among several hundred homes in a neighborhood. That bandwidth can be allocated flexibly for both broadcast and two-way data channels. Cable modems have speeds of as much as 30 megabits per second in the downstream direction. However, this total rate is shared among all users in that branch of the system.

The advantages of cable modems are that it is a relatively low cost add-on to the existing infrastructure. Unlike the telephone companies, the cable companies do not have other data services that might be cannibalized by the cable modem. To some degree the fact that the cable modem uses a shared bandwidth is an advantage. This means that heavier users can take more as needed from the light users at a given time. However, the shared bandwidth brings disadvantages, centered around the possible deterioration of the environment through excess traffic or accumulated noise.

Other proposed broadband connection technologies involve wireless transmission, whether terrestrial, airship, or satellite. For each of these possibilities there are economic tradeoffs involving the total bandwidth and the number of users. Spot beams for the satellites and sectorized antennas in terrestrial systems allow the available spectrum to be reused among sets of users. At present, however, wireless systems have been predicated on voice traffic, so that permitting many data users at megabit speeds may be uneconomic in today's architectures. A number of companies are challenging the assumptions in wireless for data, so it would be premature to believe that broadband data will not be delivered by wireless technology in the future.

Conclusion – Back to Moore's Law

This chapter started with some comments about Moore's Law and the changing price/performance of silicon technology. We need now to recast the packet revolution in this light. While computers have been able to take advantage of Moore's Law, exponentially increasing their performance for a given price, it seems that communications has not followed suit. To highlight this disparity, people have often quoted in jest a similar law that says that communications doubles its cost effectiveness every century. The question is: Can Moore's Law be applied to communications, and is packet switching the means of its application?

Clearly, the silicon chips inside switches and transmission equipment follow Moore's Law, doubling their cost effectiveness at approximately 18 month intervals. However, there are many costly components that are not primarily silicon and do not follow Moore's Law, such as cabinets, power supplies, wiring, etc. Thus the cost effectiveness of traditional circuit switching equipment is doubling its cost effectiveness at about 80-month intervals -- considerably slower than Moore's Law.

Packet switches, on the other hand, are progressing much more rapidly, with doubling periods of 10-20 months. While it is tempting to find some intrinsic technological reason why this is much better than the pace of circuit switching progress, the truth is more likely to be simply that the world is now working on packet switches. If an army of people were put together to rethink circuit switches, they could probably match this pace, but that is not the case -- packet switches are the fashion of today. Just as Moore's Law itself may be a self-fulfilling prophecy, the technologies that are in fashion at any given time show the most dramatic progress, and leave their competitors behind even as they gain more and more adherents.

There is another important consideration in the economics of packet switching, and that is the migration of cost and intelligence to the periphery of the network. One reason that the old circuit switches were so expensive is that they supported the dumbest terminal imaginable -- the common telephone. All the intelligence was centralized in the network so that the peripheral devices could be very inexpensive. Since there were so many telephones on the periphery, this seemed like a good engineering choice.

In an IP network, on the other hand, there needs to be considerable intelligence at the periphery. In fact, we assume that there is a computer or its equivalent at the end of the network in order to execute the complexities of the transport protocols. A packet network may be envisioned as a very cheap network with a \$2000 telephone, while a circuit-switched network is an expensive network connected to a \$20 telephone.

Curiously, the economics of these two contrasting philosophies may be a standoff. The telephone plant in the United States today is said to represent a \$300B investment. If, alternatively, the plant were free and there were 150 million computers connected at the periphery, the total cost would be almost the same. Of course, in the latter case, it is often argued that the computers have other uses, and that only a portion of their cost is attributable to telecommunications.

With packet switches showing exponential cost improvement and optical transmission following an even faster progress (about 12-month doubling), the equipment cost for telecommunications is plummeting. The capital cost per bit transmitted is rapidly approaching zero. At any given time, however, the mix of

equipment in the telecommunications plant ranges from the latest technology to that which may be 30 to 40 years old. The average depreciation period of telecommunications equipment is more than a decade, which may not be appropriate for today's rapidly changing technology.

The dynamics of technology change have given the opportunity for many new entrants into the telecommunications business. At any given time the new technology is much more cost effective than that used in the plants of the established companies. Startups can build entire infrastructures for a fraction of the cost spent by the incumbents. With the technology revolutions of today this causes an instability in the business environment in telecom. Every six months it seems that a new company is bragging about rewiring the world and overthrowing the existing carriers in the process.

Moore's Law, of course, only applies to the technology of communications. Most of the cost in traditional telecommunications is not in the capital equipment, but rather in the operational expense. The real expense in telecom is in maintenance, billing, customer care, provisioning, administration, and so forth. These are people expenses, and the only way that these can be reduced is to steadily decrease the number of people required per access line. In fact, this reduction has occurred throughout this century, but on a pace much slower than that of progress in electronics.

Because of the dominance of operational costs in telecom, some people argue that a packet switched network will ultimately be no less expensive than a circuit switched plant. While this argument is far from resolved, there exist opportunities to lower the operational expense in the new networks. We have already noted, for example, that in the PSTN each new customer requires considerable labor, while in an IP network it is conceivable that the new terminal could be automatically recognized and registered electronically.

The economics of the new networks are also determined by innovative business plans, which may bear more on the cost of providing service than the actual technology. Today's long distance is burdened with high marketing costs. Carriers that sell exclusively to a small set of large industrial companies do not suffer this expense. Moreover, if the new companies do not bill by the minute and mile, they do not need the extensive infrastructure for measurement and processing of data that is required to render such bills.

After decades of relative stagnation, it now appears that telecommunications is indeed following Moore's Law. Bits are getting cheaper, and they are getting cheaper very fast. Going back to the motto of the Chicago Exposition, science has found a better way to communicate, industry has applied this technology, and now it remains for mankind to explore the wonderful new uses of plentiful bandwidth.