Inside Technology
edited by Wiebe E. Bijker, W. Bernard Carlson, and Trevor Pinch

Janet Abbate, *Inventing the Internet*

Marc Berg, *Rationalizing Medical Work: Decision-Support Techniques and Medical Practices*

Wiebe E. Bijker, *Of Bicycles, Bakelites, and Bulbs: Toward a Theory of Sociotechnical Change*

Wiebe E. Bijker and John Law, editors, *Shaping Technology/Building Society: Studies in Sociotechnical Change*

Stuart S. Blume, *Insight and Industry: On the Dynamics of Technological Change in Medicine*

Geoffrey C. Bowker, *Science on the Run: Information Management and Industrial Geophysics at Schlumberger, 1920–1940*

Louis L. Bucciarelli, *Designing Engineers*

H. M. Collins, *Artificial Experts: Social Knowledge and Intelligent Machines*

Paul N. Edwards, *The Closed World: Computers and the Politics of Discourse in Cold War America*

Herbert Gottweis, *Governing Molecules: The Discursive Politics of Genetic Engineering in Europe and the United States*

Gabrielle Hecht, *The Radiance of France: Nuclear Power and National Identity after World War II*

Kathryn Henderson, *On Line and On Paper: Visual Representations, Visual Culture, and Computer Graphics in Design Engineering*

Eda Kranakis, *Constructing a Bridge: An Exploration of Engineering Culture, Design, and Research in Nineteenth-Century France and America*

Pamela E. Mack, *Viewing the Earth: The Social Construction of the Landsat Satellite System*

Donald MacKenzie, *Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance*

Donald MacKenzie, *Knowing Machines: Essays on Technical Change*

Susanne K. Schmidt and Raymund Werle, *Coordinating Technology: Studies in the International Standardization of Telecommunications*

Inventing the Internet

Janet Abbate

The MIT Press
Cambridge, Massachusetts
London, England

Wide Web are prominent examples of informally created applications that became popular, not as the result of some central agency's marketing plan, but through the spontaneous decisions of thousands of independent users.

In reconstructing the history of the Internet, I have been struck time and again by the unexpected twists and turns its development has taken. Often a well-laid plan was abandoned after a short time and replaced by a new approach from an unexpected quarter. Rapid advances, such as the introduction of personal computers and the invention of local-area networks, continually threatened to make existing network technologies obsolete. In addition, responsibility for operating the Internet changed hands several times over the course of its first thirty years or so. How, in the face of all this change and uncertainty, did the system survive and even flourish? I believe that the key to the Internet's success was a commitment to flexibility and diversity, both in technical design and in organizational culture. No one could predict the specific changes that would revolutionize the computing and communications industries at the end of the twentieth century. A network architecture designed to accommodate a variety of computing technologies, combined with an informal and inclusive management style, gave the Internet system the ability to adapt to an unpredictable environment.

The Internet's identity as a communication medium was not inherent in the technology; it was constructed through a series of social choices. The ingenuity of the system's builders and the practices of its users have proved just as crucial as computers and telephone circuits in defining the structure and purpose of the Internet. That is what the title of this book, *Inventing the Internet*, is meant to evoke: not an isolated act of invention, but rather the idea that the meaning of the Internet had to be invented—and constantly reinvented—at the same time as the technology itself. I hope that this perspective will prove useful to those of us, experts and users alike, who are even now engaged in reinventing the Internet.

White Heat and Cold War: The Origins and Meanings of Packet Switching

Of all the ARPANET's technical innovations, perhaps the most celebrated was packet switching. Packet switching was an experimental, even controversial method for transmitting data across a network. Its proponents claimed that it would increase the efficiency, reliability, and speed of data communications, but it was also quite complex to implement, and some communications experts argued that the technique would never work. Indeed, one reason the ARPANET became the focus of so much attention within the computer science community was that it represented the first large-scale demonstration of the feasibility of packet switching.[1] The successful use of packet switching in the ARPANET and in other early networks paved the way for the technique's widespread adoption, and at the end of the twentieth century packet switching continued to be the dominant networking practice. It had moved from the margins to the center, from experimental to "normal" technology.[2]

Many computer professionals have seen packet switching as having obvious technical advantages over alternative methods for transmitting data, and they have tended to treat its widespread adoption as a natural result of these advantages. In fact, however, the success of packet switching was not a sure thing, and for many years there was no consensus on what its defining characteristics were, what advantages it offered, or how it should be implemented—in part because computer scientists evaluated it in ideological as well as technical terms. Before packet switching could achieve legitimacy in the eyes of data communications practitioners, its proponents had to prove that it would work by building demonstration networks. The wide disparity in the outcomes of these early experiments with packet switching demonstrates that the concept could be realized in very different ways, and that, far from being a straightforward matter of a superior

technology's winning out, the "success" of packet switching depended greatly on how it was interpreted.

Packet switching was invented independently by two computer researchers working in very different contexts: Paul Baran at the Rand Corporation in the United States and Donald Davies at the National Physical Laboratory in England. Baran was first to explore the idea, around 1960; Davies came up with his own version of packet switching a few years later and subsequently learned of Baran's prior work. Davies was instrumental in passing on the knowledge of packet switching that he and Baran had developed to Lawrence Roberts, who was in charge of creating the ARPANET. This chain of invention and dissemination has become a standard element of origin stories about the Internet; indeed, it is easy to get the impression that packet switching simply took a detour through the United Kingdom before re-emerging, unchanged, in the United States to fulfill its destiny as the underlying technology of the ARPANET.[3]

However, while Baran's and Davies's versions of packet switching had some basic technical similarities, their conceptions of what defined packet switching and of what it was good for were very different. Much of this difference was due to the strong political pressures that were brought to bear on computing research in the United Kingdom and in the United States. Large computer projects in both countries were developed in a context of government funding and control, and national leaders saw computers as a strategic technology for achieving important political goals. But in the very different policy contexts of the United States and the United Kingdom, packet switching took on different meanings for Baran, Davies, and Roberts. Packet switching was never adopted on the basis of purely technical criteria, but always because it fit into a broader socio-technical understanding of how data networks could and should be used.

Networking Dr. Strangelove: The Cold War Roots of Packet Switching in the United States

As the 1960s opened, relations between the United States and the Union of Soviet Socialist Republics were distinctly chilly. The USSR had launched its Sputnik satellite in 1957, setting off alarm in the United States over a "science gap" and prompting a surge of government investment in science and technology. A series of events kept the Cold War in the public consciousness: an American U-2 spy plane was shot down over the USSR in 1960, the Berlin Wall went up in 1961, and 1962 brought the Cuban Missile Crisis. The shadow of nuclear war loomed over popular culture. The novels *On the Beach* (Shute 1957) and *Fail-Safe* (Burdick and Wheeler 1962)—both made into movies in the early 1960s—presented chilling accounts of nuclear war and its aftermath. And in 1964, movie theaters across the United States presented a brilliant black comedy of Cold War paranoia, *Dr. Strangelove* (Kubrick 1963).

Dr. Strangelove, though humorous, highlighted the vulnerability of the United States' communications channels to disruption by a Soviet attack, which might make them unavailable just when they were needed most. In the movie, a psychotic Air Force commander named Jack D. Ripper sets a nuclear holocaust in motion by invoking a strategy of mutual assured destruction called "Plan R." This plan—which allows Ripper to circumvent the president's authority to declare war—is specifically designed to compensate for a wartime failure in command, control, and communications. In the movie, an Air Force general explains:

> Plan R is an emergency war plan in which a lower-echelon commander may order nuclear retaliation after a sneak attack—*if* the normal chain of command has been disrupted. . . . The idea was to discourage the Russkies from any hope that they could knock out Washington . . . as part of a general sneak attack and escape retaliation because of lack of proper command and control.

Plan R allows Ripper to launch a "retaliatory" attack even though no first strike has actually occurred. In reality (as the film's disclaimer states), the US Air Force never had any such strategy. Even before *Dr. Strangelove* opened, the Air Force was exploring a very different solution to the threat of a first strike: building a communications system that would be able to survive an attack and so that "proper command and control" could be maintained. As Edwards (1996, p. 133) has documented, Cold War defense analysts saw robust communications networks as a necessity in any nuclear confrontation: "Flexible-response strategy required that political leaders continue to communicate during an escalating nuclear exchange. . . . Therefore, preserving central command and control—political leadership, but also reconnaissance, data, and communications links—achieved the highest military priority." The need for "survivable communications" was generally recognized by the early 1960s. Among those intent on filling it

was a researcher at the Air Force's premier "think tank," the Rand Corporation.

Founded by the Air Force in 1946 as an outgrowth of operations research efforts initiated during World War II, Rand (originally RAND, derived from "research and development") was a nonprofit corporation dedicated to research on military strategy and technology. Rand was primarily funded by contracts from the Air Force, though it served other government agencies as well. It attracted talented minds though a combination of high salaries, relative autonomy for researchers, and the chance to contribute to policy decisions of the highest importance (Baran 1990, pp. 10, 11). Edwards (1996, p. 116) notes that "Rand was the center of civilian intellectual involvement in defense problems of the 1950s, especially the overarching issue of nuclear politics and strategy." Rand's role was visible enough to be reflected in popular culture—for example, the fictional Dr. Strangelove turns to "the Bland Corporation" when he needs advice on nuclear strategy.[4] Because its approach to systems analysis emphasized quantitative models and simulation, Rand was also active in computer science research (Edwards 1996, pp. 122–124).

In 1959 a young engineer named Paul Baran joined Rand's computer science department. Immersed in a corporate culture focused on the Cold War, Baran soon developed an interest in survivable communications, which he felt would decrease the temptation of military leaders to launch a preemptive first strike:

> Both the US and USSR were building hair-trigger nuclear ballistic missile systems. . . . If the strategic weapons command and control systems could be more survivable, then the country's retaliatory capability could better allow it to withstand an attack and still function; a more stable position. But this was not a wholly feasible concept, because long-distance communications networks at that time were extremely vulnerable and not able to survive attack. That was the issue. Here a most dangerous situation was created by the lack of a survivable communication system. (Baran 1990, p. 11)[5]

Baran was able to explore this idea without an explicit contract from the Air Force (ibid., pp. 12, 16), since Rand had a considerable amount of open-ended funding that researchers could use to pursue projects they deemed relevant to the United States' defense concerns.[6]

Baran began in 1959 with a plan for a minimal communications system that could transmit a simple "Go/No go" message from the president to commanders by means of AM radio. When Baran presented this idea to military officers, they immediately insisted that they needed greater communications capacity. Baran spent the next three years formulating ideas for a new communications system that would combine survivability with high capacity (ibid., pp. 14–15). He envisioned a system would allow military personnel to carry on voice conversations or to use teletype, facsimile, or low-speed computer terminals under wartime conditions. The key to this new system was a technique that Baran (1960, p. 3) called "distributed communications." In a conventional communications system, such as the telephone network, switching is concentrated and hierarchical. Calls go first to a local office, then to a regional or national switching office if a connection beyond the local area is needed. Each user is connected to only one local office, and each local office serves a large number of users. Thus, destroying a single local office would cut off many users from the network. A distributed system would have many switching nodes, and many links attached to each node. The redundancy would make it harder to cut off service to users.

In Baran's proposed system, each of several hundred switching nodes would be connected to other nodes by as many as eight lines (figure 1.1). Several hundred multiplexing stations would provide an interface between the users and the network. Each multiplexing station would be connected to two or three switching nodes and to as many as 1024 users with data terminals or digital telephones. The switching was distributed among all the nodes in the network, so knocking out a few important centers would not disable the whole network. To make the system even more secure, Baran (1964a, volume VIII, section V) planned to locate the nodes far from population centers (which were considered military targets), and he designed the multiplexing stations with a wide margin of excess capacity (on the assumption that attacks would cause some equipment to fail). Baran added such military features as cryptography and a priority system that would allow high-level users to preempt messages from lower-level users.

To move data through the network, Baran adapted a technique known as "message switching" or "store-and-forward switching." A common example of message switching is the postal system. In a message switching system, each message (e.g., a letter) is labeled with its origin and its destination and is then passed from node to node through the network. A message is temporarily stored at each node (e.g., a post office) until it can be forwarded to the next node or the final destination. Each successive node uses the address information

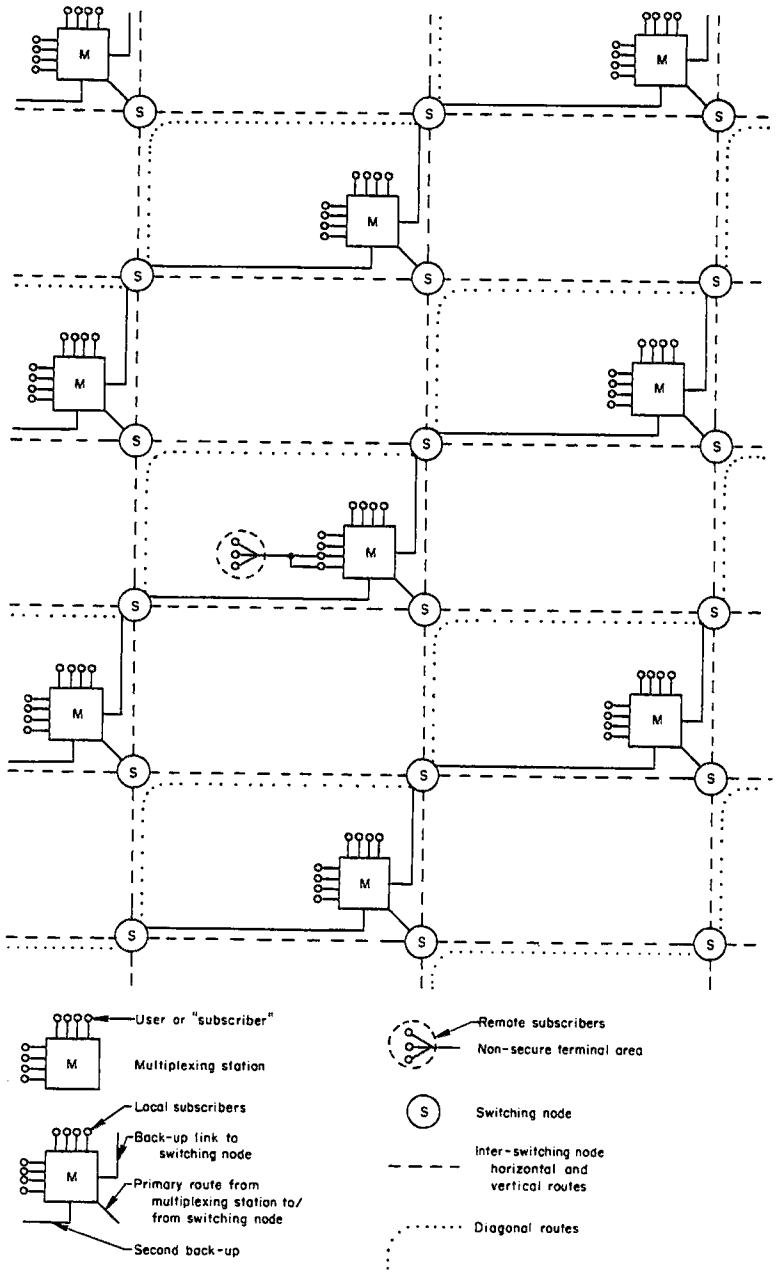

Figure 1.1
Paul Baran's design featuring highly connected switching nodes. Source: Baran 1964a, volume VIII.

to determine the next step of the route. In the 1930s, message switching came into use in telegraphy: a message was stored on paper tape at each intermediate station before being transmitted to the next station. At first, telegraph messages were switched manually by the telegraph operators; however, in the 1960s telegraph offices began to use computers to store and route the messages (Campbell-Kelly 1988, p. 224).

For the postal and telegraph systems, message switching was more efficient than transmitting messages or letters directly from a source to a destination. Letters are stored temporarily at a post office so that a large number can be gathered for each delivery route. In telegraphy, message switching also addressed the uneven flow of traffic on the expensive long-distance lines. In periods of light traffic, excess capacity was wasted; when the lines were overloaded, there was a risk that some messages would be lost. Storing messages at intermediate stations made it possible to even out the flow: if a line was busy, messages could be stored at the switch until the line was free. In this way, message switching increased the efficiency, and hence the economy, of long-distance telegraphy.[7]

Besides appreciating the efficiency offered by message switching, Baran saw it as a way to make his system more survivable. Since the nodes in a message switching system act independently in processing the messages and there are no preset routes between nodes, the nodes can adapt to changing conditions by picking the route that is best at any moment. Baran (1964b, p. 8) described it this way: "There is no central control; only a simple local routing policy is performed at each node, yet the over-all system adapts." This increases the ability of the system to survive an attack, since the nodes can reroute messages around non-functioning parts of the network. Baran realized that survivability depended on more than just having redundant links; the nodes must be able to make use of those extra links. "Survivability," Baran wrote (1964a, volume V, section I), "is a function of switching flexibility." Therefore, his network design was characterized by distributed routing as well as distributed links.

Departures from Other Contemporary Systems

Paul Baran was not the first to propose either message switching or survivable communications to the military. Systems of both types already existed or were in development. A look at the state of the art in these areas makes it easier to see what aspects of Baran's ideas were

really innovative and why he saw opportunities to depart from contemporary practice in certain areas.

Message switching systems were nothing new to the Department of Defense, but the existing systems were cumbersome and inefficient. Baran discovered this when he served as a member of a Department of Defense committee charged with examining several existing or proposed store-and-forward data systems in the early 1960s. These systems had such low capacity that backlogs of messages tended to build up at the switches. Therefore, the switches had to be built with large storage capacity to hold the messages that were waiting to be forwarded, and the switching computers ended up being large and complex. Baran was convinced that a network could and should be built using much higher transmission speeds, eliminating the bottlenecks at the nodes. Besides the obvious benefit of getting messages delivered faster, a high-speed, low-storage system could have switching nodes that were much simpler and cheaper than those used in contemporary store-and-forward data systems. As Baran (1964b, p. 6) pointed out, although the high-speed system would be store-and-forward in its design, in practice messages would spend little time being stored at the nodes; to the user, therefore, a connection would seem to be real-time. Baran's argument (1990, p. 24) that it was possible to build a message switching network with fast end-to-end transmission of messages and small, inexpensive switches was a radical challenge to the existing understanding of such systems.

The concept of "distributed communications" (or "distributed networks") also predated Baran; indeed, his publications cite examples of the idea from mathematics, artificial intelligence, and civilian and military communications (Baran 1964a, volume V, section I). In particular, military planners had already proposed a variety of systems based on a network of decentralized nodes linked by multiple connections (ibid., section IV). Though they shared the idea of distributed communications, however, these other systems differed in essential ways from Baran's proposal. In particular, most of them seem to have entailed the use of simple broadcast techniques, with every message going to every destination, whereas Baran's system would route messages individually through the network.[8]

Most of the distributed systems Baran described were only proposals, not working systems. However, there was one large distributed communications network under actual development in the early 1960s. This was AUTOVON, designed and operated for the Department of Defense by the American Telephone and Telegraph Corporation. In 1961 AT&T had provided the Army with a communications network called the Switched Circuit Automatic Network, and in 1963 the corporation provided a similar network for the Air Force called North American Air Defense Command/Automatic Dial Switching. The Defense Communications Agency, which was charged with coordinating the provision of communications services throughout the armed services, decided to integrate these networks into a new system called the Continental United States Automatic Voice Network (CONUS AUTOVON). AUTOVON was not a message switching system; it was a special military voice network built on top of the existing civilian telephone network. It went into service with ten switching nodes in April of 1964 (Schindler 1982, pp. 266–269).

Describing the AUTOVON system, AT&T's magazine *Long Lines* (1965, p. 3) noted: "The top requirement is that the system can survive disaster." Survivability was sought in part by placing the switching centers in "hardened" sites, often underground, away from major metropolitan targets. The main survivability feature, however, was that the network was arranged in what AT&T called a "polygrid," with each switch connected to several links and with the links distributed evenly throughout the system (rather than having all connections routed through a few central switches). AT&T's publicity stressed that this redundant, decentralized system represented a sharp departure from the hierarchical structure used in the ordinary toll network. AUTOVON had one node for every few hundred lines, whereas in the regular toll system a node typically served a few thousand lines. "The polygrid network," according to the system's architects, "plays a major role in the survivability of AUTOVON. Along with its other virtues of flexibility and economy, polygrid represents the best method that technology can now offer for the rapid and reliable connection of defense communications." (Gorgas 1968, p. 227)

Baran's approach differed from AT&T's in two significant ways. First, although AUTOVON had nodes distributed throughout the system, control of those nodes was concentrated in a single operations center, where operators monitored warning lights, analyzed traffic levels, and controlled system operations. If traffic had to be rerouted, it was done manually: operators at the control center would make the decision and then contact the operators at the switching nodes with instructions to change routes (Gorgas 1968, p. 223; *Long Lines* 1969). In Baran's network, control was fully distributed, as noted above.

Nodes would be individually responsible for determining routes, and would do so automatically without human intervention: "The intelligence required to switch signals to surviving links is at the link nodes and *not* at one or a few centralized switching centers." (Baran 1960, p. 3) Clearly such a system would be more survivable than one dependent on a single operations center—which, Baran noted, "forms a single, very attractive target in the thermonuclear era" (1964a, volume V, section II).

One implication of Baran's design was that the nodes would have to have enough "intelligence" to perform their own routing—they would have to be computers, not just telephone switches. This brings us to Baran's second departure from the AT&T approach: Baran envisioned an all-digital network, with computerized switches and digital transmission across the links. The complexity of routing messages would require computers at the nodes, since the switches would have to be able to determine, on their own, the best path to any destination, and to update that information as network conditions changed. Such computerized switches had never been designed before. "These problems," Baran acknowledged (1964b, p. 6), "place difficult requirements on the switching. However, the development of digital computer technology has advanced so rapidly that it now appears possible to satisfy these requirements by a moderate amount of digital equipment." Preserving the clarity of the signal would require that transmission be digital as well. One consequence of having a distributed network was that a connection between any two endpoints would typically be made up of many short links rather than a few long ones, with messages passing through many nodes on the way to their destinations. Having many links in a route was problematic for the transmission of ordinary analog signals: the signal degenerated slightly whenever it was switched from one link to another, and distortion accumulated with each additional link. Digital signals, on the other hand, could be regenerated at each switch; thus, digital transmission would allow the use of many links without cumulative distortion and errors. Digital transmission was still a novelty at the time; Bell Labs had only begun developing its T1 digital trunk lines in 1955, and they would not be ready for commercial service in the Bell System until 1962 (O'Neill 1985).[9]

Baran's system would push contemporary switching and transmission technology to their limits, so it is understandable that contemporary experts reacted skeptically to his claims. The engineers in AT&T's Long Lines Division, which ran the long-distance telephone service and the AUTOVON system, tended to be familiar only with analog technology, and they doubted Baran's claims that an all-digital system could transcend the well-known limits on the number of links per call (Baran 1990, p. 18).[10] Whereas in AUTOVON there was a maximum of seven links in any route, Baran's simulations of network routing in a small version of his system showed as many as 23 links between endpoints (Gorgas 1968, 223; Baran 1964b, p. 7, figure 11). Evidently, Baran's position outside the community of analog communications practitioners and his awareness of the potential of computer techniques made it easier for him to question the accepted limits. He had no stake in analog telephony, and his training and background in computing made it easier for him to envision an all-digital system as a way of achieving his goal of distributed communication.

And Baran's system departed from traditional telephone company practice in other ways that show the effect of Cold War military considerations on his design assumptions. For instance, AT&T tried to increase the reliability of the phone system as a whole by making each component as reliable as possible, and for an additional fee would provide lines that were specially conditioned to have lower error rates. Baran chose instead to make do with lower-quality communications links and to provide redundant components to compensate for failures. Conditioned lines would be too expensive for a system with so many links, and in any case the reliability of individual components could not be counted on in wartime conditions. "Reliability and raw error rates are secondary," observed Baran (1964b, pp. 4–5). "The network must be built with the expectation of heavy damage anyway."

Packet Switching in Baran's System

Baran's proposed network began as a distributed message switching system. His final innovation was to alter message switching to create a new technique: packet switching. In his system a message could be anything from digitized speech to computer data, but the fact that these messages were all sent in digital form—as a series of binary numbers ("bits")—meant that the information could be manipulated in new ways. Baran proposed that, rather than sending messages of varying sizes across the network, messages should be divided into fixed-size units that he called "message blocks." The multiplexing stations that connected users to the network would be responsible for dividing outgoing messages into uniform blocks of 1024 bits. A short

message could be sent as a single block; longer messages would require multiple message blocks. The multiplexer would add to each block a header specifying the addresses of the sending and receiving parties as well as other control information. The switching nodes would use the header information to determine what route each block should take to its destination; since each block was routed independently, the different blocks that made up a single message might be sent on different routes. When the blocks reached their destination, the local multiplexer would strip the header information from each block and reassemble the blocks to form the complete message. This idea would eventually be widely adopted for use in computer networks; the message blocks would come to be called "packets" and the technique "packet switching."[11]

For all its eventual significance, the decision to transmit data as packets was not the original focus of Baran's work. As the title of his eleven-volume work *On Distributed Communications* indicates, Baran began with the idea of building a distributed network—an idea that had already been identified with survivability by people working in military communications (Baran 1964a, volume V). In describing the system, Baran tended to stress the idea of link redundancy, rather than other elements such as packet switching.[12] But as he developed the details of the system, the use of message blocks emerged as a fundamental element. By the time he wrote the final volume of the series, Baran had changed the name he used to refer to the system to reflect the new emphasis: "While preparing the draft of this concluding number, it became evident that a distinct and specific system was being described, which we have now chosen to call the 'Distributed Adaptive Message Block Network,' in order to distinguish it from the growing set of other distributed networks and systems." (Baran 1964a, volume XI, section I) What, then, was so important about packet switching? What did it mean to Baran and his sponsors?

Transmitting packets rather than complete messages imposed certain costs on the system. The interface computers had to perform the work of dividing users' outgoing messages into packets and of reassembling incoming packets into messages. There was also the overhead of having to include address and control information with each packet (rather than once per message), which increased the amount of data that had to be transmitted over the network. And since packets from a single message could take different routes to their destination, they might arrive out of sequence, which meant that there had to be provisions for reassembling them in the proper order. All this made the system more complex and presented more opportunities for failure. For Baran, these costs were outweighed by his belief that packet switching would support some of the fundamental goals of the system.

Packet switching offered a variety of benefits. Baran was determined to use small, inexpensive computers for his system, rather than the huge ones he had seen in other message switching systems, and he was aware that the switching computers would have to be simple in order to be both fast and inexpensive. The use of fixed-size packets rather than variable-size messages could simplify the design of the switching node. Another advantage for the military was that breaking messages into packets and sending them along different routes to their destination would make it harder for spies to eavesdrop on a conversation. But the biggest potential reward was efficient and flexible transmission of data. "Most importantly," wrote Baran (1964b, p. 6), "standardized data blocks permit many simultaneous users, each with widely different bandwidth requirements[,] to economically share a broad-band network made up of varied data rate links." In other words, packet switching allowed a more efficient form of multiplexing (sharing of a single communication channel by many users).

In the conventional telecommunications systems of the early 1960s, the usual form of multiplexing was by frequency division: each caller would be assigned a particular frequency band for their exclusive use for the duration of their connection. If the caller did not talk or send data continuously, the idle time would be wasted. In an alternative method, called "time division multiplexing," time is divided into short intervals, and each user in turn is given a chance to transmit data for the duration of one interval. Only users with data to transmit are offered time slots, so no slots go idle as long as anyone has data to transmit; this makes time division multiplexing more efficient for usage situations where bursts of information alternate with idle periods. Since computer data tends to have this "bursty" characteristic, Baran (1964b, p. 6) felt that time division was a more "natural" form of multiplexing for data transmission. And since the time slot accommodated a fixed amount of data, Baran believed that the use of fixed-size message blocks was a prerequisite for time division multiplexing. Thus, he associated packet switching with time division multiplexing and its promise of efficient data transmission.[13]

Packet switching would also make it easier to combine links having different data rates in the network. The data rate is the number of bits

per second that can be transmitted on a given link. In the conventional telephone system, each caller is connected at a fixed data rate, and data must flow into and out of a switch at predetermined rates. With packet switching, data flowing into a switch can be divided among the outgoing links in a variety of ways, rather than having to be sent out at a fixed rate. This would make it easier for devices transmitting data at different rates (computers and digital telephones, for instance) to share a link to the network. The system could also take advantage of new media, such as low-cost microwave transmission, that had different data rates than the standard phone company circuits. Though packet switching made the system more complex in some respects, in other ways it made the system simpler and less costly to build.

In sum, packet switching appealed to Baran because it seemed to meet the requirements of a survivable military system. Cheaper nodes and links made it economically feasible to build a highly redundant (and therefore robust) network. Efficient transmission made it possible for commanders to have the higher communications capacity they wanted. Dividing messages into packets increased security by making it harder to intercept intelligible messages. Packet switching, as Baran understood it, made perfect sense in the Cold War context of his proposed system.

The Impact of Baran's Work

For a brief time after its publication in 1964, it seemed that Baran's *On Distributed Communications* might soon become the blueprint for a nationwide distributed packet switching network. In August of 1965, Rand officially recommended that the Air Force proceed with research and development on a "distributed adaptive message-block network." Enthusiastic about the proposal, Air Force representatives sent it for review to the Defense Communications Agency, which oversaw the provision of military communications services (Baran 1990, Attachment 2). The DCA was one of many agencies that had been created in an attempt to bring military operations under the central control of the Department of Defense rather than allowing each of the armed services to build its own systems.[14] In accordance with this centralizing strategy, the DoD administration made it clear during the review process that any new network would be built not by Air Force contractors but by the DCA, which had no expertise in digital technology. Baran and his Air Force sponsors, doubting that the DCA would be able to build the system that Baran had described, reluctantly decided to scrap the proposal rather than risk having it executed badly, which would waste large sums of money and perhaps discredit Baran's ideas (Baran 1990, pp. 33–35).[15]

Though the proposed network was never built, Baran's ideas were widely disseminated among researchers interested in new communications technologies. Following Rand's standard practice, Baran presented his work to various outside experts for comment as he was developing his ideas.[16] Eleven volumes of reports published in 1964 were widely distributed to individuals, government agencies, Rand depository libraries, and other people working in the field. The first volume was also published as an article in the March 1964 issue of *IEEE Transactions on Communications Systems,* and an abstract appeared in the August 1964 issue of *IEEE Spectrum* (a magazine for electrical and computing engineers with an estimated circulation of 160,000).[17] Baran also lectured on his work at various universities (Baran 1990, pp. 32–33, 36). It is not clear how many researchers were immediately influenced by Baran's ideas through these channels. Most academic computer scientists were not concerned with the survivability of communications, and they may not have seen the applicability of Baran's research to their own interests. Several years later, however, his work would begin to receive wide attention as one of the technical foundations of the ARPANET. Curiously enough, the connection between these two American networking efforts would be made via a laboratory in England.

Forging Packet Switching in the White Heat: Networks and Nationalism in the United Kingdom

In the early 1960s, while the United States was caught up in the Cold War, the United Kingdom was experiencing political upheaval of a different type. Just as the Americans were worried about a "science gap" between their country and the USSR, so there were widespread fears in the United Kingdom of a "technology gap" with the United States. Harold Wilson was elected leader of the British Labour Party in 1963, at a time when that party, and much of the general population, felt that the UK was facing an economic crisis. Politicians on all sides warned that the UK was falling behind the other industrial powers in its exploitation of new technologies, that there was a "brain

drain" of British scientists to other countries, and that the country's technological backwardness was at least partly responsible for its economic malaise (Coopey and Clarke 1995; Edgerton 1996, pp. 53, 57).

Wilson addressed the technology issue head on in a speech to the Labour Party's annual conference at Scarborough on 2 October 1963. Calling on labor and management to join in revitalizing British industry, Wilson stressed the importance of keeping up with the ongoing scientific and technological revolution, and he invoked a stirring vision of a new United Kingdom "forged in the white heat of this revolution" (quoted in Edgerton 1996, p. 56). The speech created a sensation in the British media, and Wilson was praised in newspapers across the political spectrum for capturing the concerns of the times and remaking Labour's supposedly anti-progress image.[18] When Labour came to power in the 1964 general election, Wilson was eager to act on his vision by implementing a new economic and technological regime for the United Kingdom.

Wilson's plans included reversing the "brain drain" by training more scientists and giving them the status and the facilities that would persuade them to stay in the United Kingdom, by rationalizing existing industries and creating new high-tech industries, and by shifting resources from unproductive defense and "prestige" areas (such as aerospace and nuclear energy) to commercial applications. To oversee national technological development, Wilson created the Ministry of Technology, a major new department that assumed control of the Atomic Energy Authority, the Ministry of Aviation, the National Research Development Corporation, and a number of national laboratories (Edgerton 1996, pp. 65–70). Mintech, as it came to be called, had two main aims: to transfer the results of scientific research to industrial development, and to intervene in industry so as to make private enterprise more efficient and competitive. Mintech was to have, in Wilson's words (1971, p. 8), "a very direct responsibility for increasing productivity and efficiency, particularly within those industries in urgent need of restructuring or modernisation." These industries included machines tools, aviation, electronics, shipbuilding, and—above all—computing.

Wilson feared that the British computer industry would be destroyed by competition from the United States unless the government intervened quickly. He later recalled: "When, on the evening we took office, I asked Frank Cousins to become the first Minister of Technology, I told him that he had, in my view, about a month to save the British computer industry and that this must be his first priority." (Wilson 1971, p. 9) Cousins responded by increasing funding for National Research Development Corporation, which gave development funds to corporations that wanted to commercialize government research, and by using government contracts to encourage the introduction of new computer products (ibid., p. 63). In addition, Mintech and the Industrial Reorganization Corporation were responsible for pushing British corporate mergers to create large companies, such as International Computers Limited, which would supposedly have the critical mass of resources to compete internationally (Hendry 1990, pp. 155–157; Wilson 1971, p. 63). In 1965 Mintech also took over a government initiative called the Advanced Computer Techniques Project, which had been set up in 1960 to help spin off government-sponsored computing research to industry. Under Wilson, computing research was expected to serve economic aims, and the possibility of government intervention was always present.

One of the British scientists who took the lead in computing research was Donald W. Davies of the National Physical Laboratory in Teddington, a suburb of London. The NPL—established in 1899 to determine values for physical constants, to standardize instruments for physical measurements, and to perform similar activities involving standards and materials testing (Pyatt 1983, pp. 157–158)—had first become involved in computing in 1946, when a team at the laboratory, following a proposal by Alan Turing, built an early stored-program digital computer called the Pilot ACE. Davies had joined the NPL in 1947 and had worked on the Pilot ACE; in 1960 he had become superintendent of the division in charge of computing science, and in 1965 he had been named technical manager of the Advanced Computer Techniques Project (Campbell-Kelly 1988, pp. 222–223). Davies's position kept him in touch with the latest advances in computing technology and with the government's plans to use that technology to aid the British economy.

If the watchword for Baran was survivability, the priority for Davies was interactive computing. Davies was one of many researchers who hoped to improve the user friendliness of computers. Computers of the early 1960s were expensive and in high demand. This meant that their operating systems were designed for maximum efficiency in the use of the computer's central processor. To achieve this, the typical operating system of the early 1960s used batch processing, a technique in which a number of computer programs would be collected and

loaded into the computer together to be executed in succession. Running programs in batches was efficient because it minimized the time the computer spent idle, waiting for data to be loaded or unloaded. The disadvantage of batch processing was that it did not allow users direct interaction with or an immediate response from the computer. As a result, computer users often experienced batch processing as slow, difficult, and tedious.

In the typical programming cycle, the user of a batch processing computer would begin by writing out a program on paper. Then the user or a keypunch operator would punch holes in a set of computer cards to represent the written instructions. The user would bring the deck of punched cards to the computer center, where an operator would feed them into a punched-card reader and transfer the data to magnetic tape. When the computer became available, the operator would load the tape and run its batch of programs, and eventually he or she would return a printout of the results to the various programmers. If a user's program turned out to have errors, the user would have to rewrite it, punch another set of cards, and submit the cards again, perhaps waiting hours for a chance to rerun the program and collect the results. Often users had to repeat this cycle numerous times before a program would work correctly.

Batch processing rationalized the flow of input to the computer, but it was frustrating and inefficient for the programmer. In the late 1950s computer scientists began to talk about a possible alternative, which they called "time sharing." Instead of running a single program from start to finish before going on to the next one, a time sharing operating system would cycle between a number of programs, devoting a fraction of a second of processing time to each one before going on the next (figure 1.2). The wait between cycles would be so short that users would have the impression of continuous interaction with the machine, just as moviegoers have the impression of seeing continuous motion on the screen rather than a rapid succession of still frames.

By sharing the computer's processor among multiple users, time sharing addressed the mismatch between the pace of human action and the much faster processing of the computer. When a computer serves a user at a interactive terminal, it spends most of its time waiting for commands; very little time is spent actually processing data. If a computer can serve many terminals at once, it will spend less time idle and more time doing productive work, which increases the efficiency—and therefore the economically feasibility—of interactive computing.

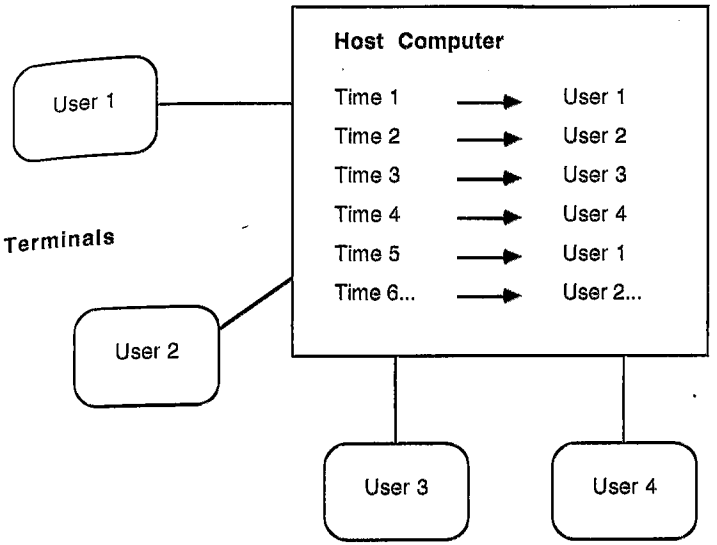

Figure 1.2
Time sharing.

Though time sharing is not necessarily synonymous with interactive computing,[19] the two ideas became closely associated. Time sharing was seen by its proponents as the innovation that would liberate computer users from their punched cards and allow direct and easy interaction with the machine.

The first proposals for time sharing operating systems were presented independently in 1959 by Christopher Strachey of the National Research Development Corporation in the United Kingdom and John McCarthy of the Massachusetts Institute of Technology in the United States. Time sharing caused tremendous excitement in the field of computing, and both academic researchers and industry analysts predicted that it would be the wave of the future. By the mid 1960s, research centers in the United Kingdom and in the United States were using time sharing computers regularly, and computer manufacturers were rushing to bring time sharing products to the market. Businesses began to spring up that offered access to time sharing machines on a commercial basis to customers who would rent or buy a terminal, connect to the service using a modem and a telephone line, and access the service's computers for an hourly rate. Many people thought that time sharing represented the future of interactive computing, since

few if any anticipated the advent of small, inexpensive "personal" computers in the late 1970s.

Davies became interested in time sharing during a 1965 trip to the United States. He had gone there to participate in a computing conference being held at MIT, and he took the opportunity to visit several American computing research sites (Davies 1986, pp. 4–5). Both the conference and his site visits made it clear to him that interest in and knowledge about time sharing were much more widespread in the United States than in the United Kingdom. When Davies returned to the National Physical Laboratory, he decided to organize a seminar on time sharing to disseminate these ideas to the British computing community. The seminar was held in November of 1965, and a number of British and American researchers were invited.

It was during these discussions that Davies became aware of a widely perceived obstacle to interactive computing: inadequate data communications. In early time sharing systems, the terminals had been directly connected to the computer and were located in an adjacent terminal room. As time went on, people began locating terminals at some distance from the computer itself, either for the user's own convenience or, in the case of commercial time sharing services, to offer access to customers over a wide geographic area. Distant terminals could be connected to the computer using dial-up[20] telephone links and modems, but long-distance telephone connections were very expensive, and for data transmission they were also inefficient. Computer messages, as noted earlier, tend to come in bursts with long pauses in between, so computer users paid dearly for telephone connections that were idle much of the time. The high cost of communications put pressure on users to work quickly, sacrificing the user friendliness for which time sharing had been invented.[21] Davies had a long-standing interest in switching techniques. As he thought about the data communications problem, he came up with the idea that a new approach to switching might offer a solution (Davies 1986, pp. 6–7). He knew that message switching was used in the telegraph system to make efficient use of lines, and he believed that by adapting this technique to computer communications he could achieve similar economies. Like Baran, Davies came from a background in computing, rather than communications, so he felt free to suggest a technique that departed from traditional communications techniques but took advantage of advances in computer technology. Davies proposed dividing messages into standard-size "packets" and having a network of computerized switching nodes that would use information carried in packet headers to route the packets to and from time sharing computers. He called this technique "packet switching."

Packet Switching in Davies's System

Like Paul Baran, Donald Davies saw that packet switching would allow many users to share a communication link efficiently. But Davies wanted that efficiency for a different purpose. Packet switching, in his view, would be the communications equivalent of time sharing: it would maximize access to a scarce resource in order to provide affordable interactive computing.[22]

In March of 1966, Davies presented his network ideas publicly for the first time, to an enthusiastic audience of people active in computing, telecommunications, and the military. Afterward, a man from the British Ministry of Defence gave Davies the surprising news that packet switching had already been invented a few years earlier by an American (Baran). The fact that the military man knew about this earlier development when Davies did not underscores the very different contexts in which packet switching evolved in the two countries. Baran's foremost concern had been survivability, which was underlined by his use of terms like "raid," "salvos," "target," "attack level," and "probability of kill" in describing the hostile conditions under which his system was expected to operate (Baran 1964b, p. 2). Davies, on the other hand, did not view packet switching as a way to make the network survivable; after reading Baran, he commented that "the highly connected networks there considered" were "not needed in a civil environment" (Davies 1966b, p. 21).[23] Davies thought the pressing need was for a network that could serve the users of commercial time sharing services. This assumption is evident in his plan to survey businesses' data communications requirements (Davies 1968a). It also shows up in Davies's efforts to make the system easy to use. In his proposal for a national network, he wrote: "A further aim requirement we must keep in mind constantly is to make the use of the system simple for simple jobs. Even though there is a communication system and a computer operating system the user must be able to ignore the complexities." (Davies 1966a, p. 2)

Packet switching served the aim of building a commercial system mainly by bringing down the cost of data communications. However, Davies found further meanings in packet switching that derived from his vision of a commercial system. One of the merits he saw in packet

switching was that it helped achieve fairness in access to the network. In an ordinary message switching system, each message had to be sent in its entirety before the next message could begin. In a packet switching system, time division multiplexing would allow users to take turns transmitting portions of their messages. If a user had a short message, such as a single command for a time sharing system, the whole message could be sent in the first packet, while longer messages would take several time slots to transmit. This way, the user with the short message would not have to wait behind users with long messages (Campbell-Kelly 1988, p. 226). This kind of fairness was appropriate for a system where computers were serving the everyday needs of civilians, rather than transmitting life-or-death messages through a command hierarchy.

Ultimately, Davies thought, packet switching technology could become a commercial product that would contribute directly to Harold Wilson's plan to revitalize the British economy. In a 1965 proposal to have the General Post Office build a prototype for a national packet switching network, Davies (1965, p. 8) wrote:

> Such an experiment at an early stage is needed to develop the knowledge of these systems in the GPO and the British computer and communications industry. . . . It is very important not to find ourselves forced to buy computers and software for these systems from [the] USA. We could, by starting early enough, develop export markets.[24]

Davies (1968c, p. 7) reiterated the need to compete with the United States in 1968, when he compared a proposed Mintech network with the planned ARPANET:

> The proposal resembles the ARPA network being planned. . . . The sponsors of that project believe it will "spearhead" a new kind of data communication system to be developed on a nation-wide scale.
>
> A Mintech network would go beyond the present ARPA plans by providing for a variety of terminals as well as computer to computer communication. To be useful as a "spearhead" project it would need to be started soon and planned with as short a time scale as possible, coming into operation well before a national network.

For Davies, the network was not only a communications tool; it was also a way for British researchers to apply the "white heat" of scientific innovation to counteracting American dominance in the computer market.

Davies's concern with economics and user friendliness underscores the national context in which he conceived the idea of a packet switching network. Davies did not envision a world in which his proposed network would be the only surviving communications system. Rather, he saw a world in which packet switching networks would compete with other communications systems to attract and serve the business user and in which the United Kingdom would need to compete with the United States and other countries to offer innovative computer products.

In December of 1965, Davies proposed the idea of a national packet switching network that would provide inexpensive data communications across the United Kingdom (figure 1.3). He envisioned the network as offering a number of services to business and recreational users, including remote data processing, point-of-sale transactions, database queries, remote control of machines, and even online betting (Davies 1965). In his scheme, a backbone of high-capacity telephone lines would link major cities in the United Kingdom; the proposed network had multiple connections to most nodes, although it was not nearly as redundant as Baran's system.[25] Like Baran, the National Physical Laboratory group designed a network with a dynamic, distributed routing system, each node making routing decisions independently according to current conditions in the network. The nodes would be connected by high-speed telephone lines so as to provide fast response for interactive computing. Users would attach their computers, terminals, and printers to the nodes through dedicated interface computers at local sites.

Davies was convinced that a data communications infrastructure of the sort he was proposing would be necessary to keep the United Kingdom competitive in the information age, and he did not doubt that such a network would someday be built. However, the NPL did not have the resources or the authority to build such a large network on its own. This authority belonged to the General Post Office, which ran the national postal and telephone networks, but managers there had little knowledge of or interest in data communications. Since Davies felt there was no hope of convincing the GPO to collaborate on a national network, he decided that a small in-house experiment would be the only feasible alternative. In the summer of 1966 he made a second, much more modest proposal to build a prototype network at the NPL. This network, named "Mark I," would serve as a demonstration of packet switching, advance the state of knowledge in the field, and support the operational computing needs of the NPL's scientific and administrative personnel.

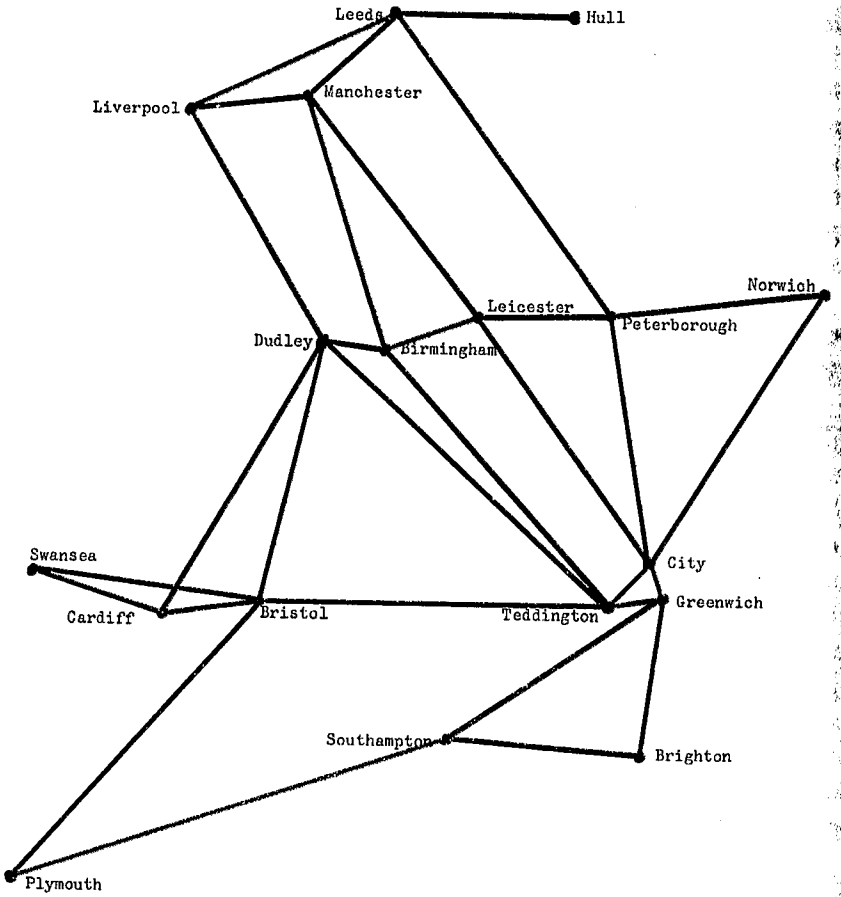

Figure 1.3
Donald Davies's proposed network for the United Kingdom (from archival copy).

The Mark I project started in 1967 with a development team headed by Derek Barber. Roger Scantlebury was the technical leader, Keith Bartlett oversaw hardware development, and Peter Wilkinson was in charge of software development (Campbell-Kelly 1988, pp. 228–229). Though they brought skill and enthusiasm to the project, the members of the NPL team had to struggle against technical and financial constraints. The Mark I was originally designed to have three packet switching nodes, but funding constraints reduced the number to one. With only one node, the NPL team would not get the opportunity to tackle certain issues—such as congestion and routing—that a multinode network would raise. Davies hoped that simulation studies could compensate for the lack of direct experience with a full-scale network (ibid., p. 229). The team had chosen for the node a computer, made by the English manufacturer Plessey, that had been designed specifically for data communications. After the NPL group had spent a year designing the Mark I around this machine, it was abruptly withdrawn from the market, and the team had to switch to a new computer. The chosen computer was the Honeywell 516.

A Honeywell 516 was installed at the NPL in 1969, and over the next two years user services were added to the network. The Mark I had about sixty lines that provided access to a DEC PDP-8 computer and two mainframes. Through the network, NPL researchers could have remote access to computers for writing and running programs, for querying a database, for sharing files, for special services such as a "desk calculator," and for "communication between people" (Davies 1966a, pp. 1–2). The system also included a file server and a "Scrapbook" application that provided document editing and communication tools (Campbell-Kelly 1988, p. 236).

Most of the effort centered on the design of the network interface. This design was shaped by Davies's assumption that businesspeople would turn out to be the main users of networks. Davies (1966b, p. 3) explained: "The emphasis on real-time business systems in this report is due to the belief that they will generate more real-time digital communication traffic than, say, scientific calculations or computer-aided design." Whereas academic researchers might need to transfer large amounts of data from one computer to another, businesspeople would be using terminals to access an interactive computer. The NPL designers therefore focused mainly on providing an easy-to-use terminal interface to the network.

One unusual characteristic of the Mark I that derived from the emphasis on user friendliness was that all terminals, printers, and other peripheral devices were connected directly to the network. The network was actually interposed between a computer and its own peripherals, so that the network became, in a sense, internal to the computer. Davies (1966b, p. 11) commented:

> The overall description of the system shows a major organisational change. Present day multi-access computers each have equipment which assembles messages from keyboards and distributes them to printers. What we are proposing is that this function should be carried out by the network, not the attached computers.

Using the network as a common communication channel for all components would make it possible for any pair of machines to interact. Normally, a terminal user who wanted to print a file would have to log in to a host computer and send a command to a printer attached to that computer. With the Mark I, however, the user could send a command directly from their terminal to the printer, without ever having to go through the printer's host computer. Remote resources would be as easy to use as local ones, since the access procedures were identical. This was a radical concept in user interface design—a concept that would not become a commonplace feature of networked systems for another twenty years.

There was a price to pay for this vision, however. Since all terminals were connected through the network, a failure in the network would mean that terminals would be cut off even from their local host computer.[26] The variety of peripherals attached to the network also made the interface computer more complicated and expensive to build, which delayed the completion of the project.[27] And, in trying to make the terminal interface user friendly, the designers of the Mark I sacrificed flexibility and adaptability. For example, they implemented parts of the user interface in hardware (figure 1.4). A user wishing to set up a connection would punch a button marked TRANSMIT on the front of the terminal, after which a light labeled SEND would light up to indicate that the network was ready to accept data; there were other lights and buttons for different operations. This interface was easy for novices to learn, but it was harder to automate or modify than a procedure implemented in software would have been (anonymous 1967). In the fast-changing world of computing, a system that was not adaptable was in danger of becoming obsolete.

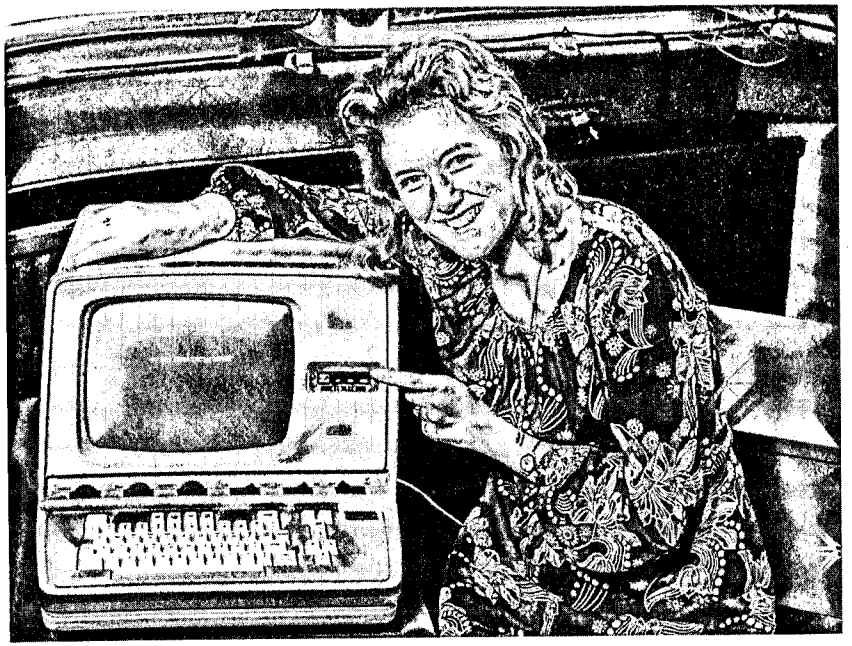

Figure 1.4
A Mark I terminal. The text on the screen reads "NPL Data Communications Network." Source: National Physical Laboratory, Teddington. Reproduced by permission of controller of HMSO.

The Impact of Davies's Work

The Mark I came to be used regularly by researchers at the National Physical Laboratory, and in 1973 Donald Davies's team introduced an upgraded version of the system called "Mark II." The Mark II used most of the same hardware as the Mark I, but software improvements made it two to three times faster. The Mark II remained in service at the NPL until 1986—quite an impressive term of service for an experimental system (Campbell-Kelly 1988, pp. 237–239). Drawing on their experience with the network, members of the NPL team went on to participate in several larger network projects in the United Kingdom and in Europe.

But despite Davies's technical innovations and the local success of the system, the Mark I did not have the kind of influence that the ARPANET would have. Davies was never able to build the national network he had proposed, and the specific techniques used in the Mark I were not transferred outside the NPL. Though Davies had had

a head start on the builders of the ARPANET, it was their work that would come to dominate the field of computer networking.

The politics of the day and the culture of some British institutions hampered Davies's ability to implement his ideas and fulfill his aim of keeping the United Kingdom ahead of the United States in computer networking. In the late 1950s the NPL had been oriented toward pure research, but under the Wilson government there was a marked increase in government oversight and intervention. In the recollection of one NPL scientist (Pyatt 1983, pp. 145–146):

> Schemes for improving the service given to the nation were constantly being hawked from above. . . . Open-ended research was severely cut back and in its place all research projects had to have a 'customer,' who had to be persuaded of the viability and value of each project and agree to make available the funds to carry it out. . . . Meetings [with customers] required regular preparation of cases by Laboratory scientists in time which could ill be spared from practical work.

For Davies and the Mark I team, the emphasis on promoting commercial spinoffs of the network diverted time from actual research and development.

Another source of difficulty for the NPL was Mintech's attempt to "rationalize" the computer industry by forcing manufacturers to reduce the number of different types of computers they offered, on the theory that having a few models with large production runs would create economies of scale. Bowing to this policy, the Plessey Corporation canceled its plan to produce the minicomputer that the NPL team had chosen for its network interface. This delayed the NPL project and forced the NPL designers to make up for the lost functionality of the Plessey computer by increasing the complexity of other parts of the system.

Another major obstacle for Davies was that he needed help from the General Post Office (which had a monopoly on national telecommunications services) to build a large-scale network, and the GPO showed little interest in new computer technology. Davies was not alone in his vexation with the GPO. Early in 1967 a small but influential group of people involved in the British time sharing industry formed the "Real Time Club." This club's main activity was the sharing of information at informal monthly meetings, but it also occasionally lobbied the government to provide more support for data communications (Malik 1989; Foy 1986; Campbell-Kelly 1988, p. 228). Members of the Real Time Club complained about the GPO's reluctance to provide better data communications:

> The entrepreneurs discovered they were all hampered in their time sharing activities by the same thing—what they felt was foot-dragging on the part of the GPO . . . when it came to lines and modems for time sharing services. (Foy 1986, p. 370)[28]

Club members decided that public action was called for, and on 3 July 1968 they held a public event called "Conversational Computing on the South Bank" at London's Royal Festival Hall (Campbell-Kelly 1988, p. 228). Commercial time sharing firms demonstrated their services, leading figures in computing gave talks, and hundreds of computer professionals attended. One of the club's leading members, Stanley Gill, a professor at Imperial College, gave a speech urging that Donald Davies's network design be adopted. The Americans, Gill noted, were already working on plans for the ARPANET. Well attended and widely reported in the press, Conversational Computing on the South Bank generated a public debate on the idea of building a national packet switching network.

Eventually, the activism of computer users forced General Post Office authorities to develop data communications services. The GPO initiated several studies of networking, and with continued pressure from the Real Time Club the government began to give more support to networking research (Campbell-Kelly 1988, pp. 242–243).[29] The NPL's Roger Scantlebury, who had worked on the Mark I, helped supervise the research contracts for the GPO. In 1973 these activities led the GPO to begin work on its Experimental Packet Switching Service (EPSS), which became operational in 1977. However, though Davies's work had helped convince some influential people that a national network could and should be built, the design of EPSS differed significantly from Davies's vision of packet switching (ibid.).[30] Even worse from the NPL's perspective, the Post Office's next-generation Packet Switching Service was based on American rather than British technology; it used a system, developed by the American firm Telenet, that was a spinoff of the ARPANET project.[31] The Wilson government had aimed to encourage the development and exploitation of British computing technology, but its failure to coordinate decision making with the researchers on the front lines of innovation had had—at least in the case of the NPL networking effort—the opposite effect.

Paul Baran and Donald Davies had both envisioned nationwide networks that would use the new technique of packet switching, but neither man had been able to fully realize this goal. Instead, the first large-scale packet switching network would be built by the Advanced Research Projects Agency.[32] The design of this network would draw on the work of both Baran and Davies, but the network's builders had their own vision of what packet switching could achieve.

ARPA was one of many new American science and technology ventures that had been prompted by the Cold War. Founded in 1958 in response to Sputnik, ARPA had as its stated mission keeping the United States ahead of its military rivals by pursuing research projects that promise significant advances in defense-related fields.[33] Throughout its existence ARPA has remained a small agency with no laboratories of its own. ARPA managers initiate and manage projects, but the actual research and development is done by academic and industry contractors. Recognized even by its critics for good management and rapid development of new technologies, ARPA has had some success in transferring its technologies to the armed services and the private sector (Pollack 1989, p. 8).

The director of ARPA reports to the Director of Defense Research and Engineering at the Office of the Secretary of Defense. ARPA has several project offices that fund research in different areas; project offices are created or disbanded in response to the changing priorities of the Department of Defense. Each office has a director and several program managers, all of whom are directly involved in choosing research projects. The first project offices directed research in behavioral sciences, materials sciences, and missile defense. In 1962, with the founding of its Information Processing Techniques Office (IPTO), ARPA became a major funder of computer science in the United States, often outspending universities significantly. Computer science, not yet an established discipline in 1962, developed rapidly once IPTO began funding it. IPTO has been the driving force behind several important areas of computing research in the United States, including graphics, artificial intelligence, time sharing operating systems, and networking (Norberg and O'Neill 1996).[34]

ARPA's funding of basic research was consistent with the philosophy of the administration of President Lyndon Johnson, who, in a September 1965 memo to his cabinet, advocated the use of agency funds to support basic research in universities. In that memo, Johnson, noting that about two-thirds of universities' research spending was funded by various federal agencies, said that this money should be used to establish "creative centers of excellence" throughout the nation (Johnson 1972, p. 335). He urged each government agency engaged in research to take "all practical measures . . . to strengthen the institutions where research now goes on, and to help additional institutions to become more effective centers for teaching and research" (ibid., p. 336). Johnson specifically did not want to limit research at these centers to mission-oriented projects. "Under this policy," he wrote (ibid., p. 335), "more support will be provided under terms which give the university and the investigator wider scope for inquiry, as contrasted with highly specific, narrowly defined projects."

A few months later, the Department of Defense responded to Johnson's call with a plan to create "centers of excellence" in defense-related research. "Each new university program," the DoD suggested, "should present a stimulating challenge to faculty and students and, at the same time, contribute to basic knowledge needed for solving problems in national defense." (Department of Defense 1972, p. 337) IPTO created several computing research centers, giving large grants to MIT, Carnegie Mellon, UCLA, and other universities. By 1970, ARPA had funded a variety of time sharing computers located at universities and other computing research sites across the United States. The purpose of its proposed network—the ARPANET—was to connect these scattered computing sites.

The ARPANET project was managed by Lawrence Roberts, a computer scientist who had conducted networking experiments at MIT's Lincoln Laboratory before joining ARPA in 1966. Roberts had a mandate to build a large, multi-computer network, but he did not initially have a firm idea of how to do this. He considered having pairs of computers establish a connection using ordinary telephone calls whenever they needed to exchange data—a method he had employed in earlier experiments. But the high cost of long-distance telephone connections made this option seem prohibitively expensive. Roberts also worried that ordinary phone service would be unacceptably prone to transmission errors and line failures. Although he was aware of the concept of packet switching, Roberts was not sure how to implement it in a large network.

In October of 1967, with these issues still unresolved, Roberts attended a computing symposium in Gatlinburg, Tennessee, where he

was slated to present ARPA's tentative networking plans. Roger Scantlebury of Britain's National Physical Laboratory also presented a paper at the symposium, where Roberts heard for the first time about Davies's ideas on packet switching and the ongoing work on the Mark I. After this session, a number of conference attendees gathered to discuss network design informally, and Scantlebury and his colleagues advocated packet switching as a solution to Roberts's concerns about line efficiency. The NPL group influenced a number of American computer scientists in favor of the new technique, and they adopted Davies's term "packet switching" to refer to this type of network. Roberts also adopted some specific aspects of the NPL design. For instance, Roberts had planned to use relatively low-speed telephone lines to connect the network nodes. He later recalled that, after the NPL representatives had "spent all night with [him] arguing about the thing back and forth," he had "concluded from those arguments that wider bandwidths would be useful" (Roberts 1989). Roberts decided to increase the bandwidth of the links in his proposed network from 9.6 to 56 kilobits per second. The ARPANET would also use a packet format similar to the NPL Mark I.[35]

After the ARPANET project was underway, the acoustics and computing firm of Bolt, Beranek and Newman, which had the main contract to build the network nodes, continued to interact with the NPL group. According to BBN's Robert Kahn (1990),

> Donald Davies was a very creative guy; he thought a lot about interesting ideas of how networks should be built. He clearly had the concept in his head of what packet networks ought to look like, and he had done it independently in England. I believe Larry Roberts will probably tell you that Donald had a big influence on him.

The NPL's Derek Barber visited the BBN team in 1969; he reported that they "were interested in the possibility of connecting our type of local area [network] directly into" the ARPANET and that they saw the NPL work as "complementary" to the ARPANET project (Barber 1969, p. 15).[36]

Paul Baran, too, became directly involved in the early stages of planning the ARPANET. Roger Scantlebury had referred Lawrence Roberts to Baran's earlier work. Soon after returning to Washington from Gatlinburg, Roberts had read Baran's *On Distributed Communications*. Later he would describe this as a kind of revelation: "Suddenly I learned how to route packets." (Norberg and O'Neill 1996, p. 166) Some of the ARPANET contractors, including Howard Frank and Leonard Kleinrock, were also aware of Baran's work and had used it in their research.[37] In 1967, Roberts recruited Baran to advise the ARPANET planning group on distributed communications and packet switching.

Through these various encounters, Roberts and others members of the ARPANET group were exposed to the ideas and techniques of Baran and Davies, and they became convinced that packet switching and distributed networking would be both feasible and desirable for the ARPANET. Packet switching promised to make more efficient use of the network's long-distance communications links and to enhance the system's ability to recover from equipment failures, which an experimental network would surely encounter. At the same time, however, packet switching was an unproven technique that would be difficult to implement successfully. The decision to employ packet switching on such a large scale reflected ARPA's commitment to high-risk research: if it worked, the payoff would be not only greater efficiency and ruggedness in the ARPANET itself, but also a significant advance in computer scientists' understanding of network properties and techniques. The ARPA managers could afford (indeed, had a mandate) to think extravagantly—to aim for the highest payoff rather than the safest investment.

The Social Construction of Packet Switching

The projects sponsored by Rand, the NPL, and ARPA had much in common in their approach to packet switching, but some crucial differences in ARPA's approach helped the ARPANET play a more enduring and influential role than the other projects. Donald Davies, Paul Baran, and Lawrence Roberts each made technical choices based on specific local concerns, and the extent to which their systems were influential depended in part on whether others shared those concerns. For instance, Baran's system had many elements that were specifically adapted to the Cold War threat, including very high levels of redundancy, location of nodes away from population centers, and integration of cryptographic capabilities and priority/precedence features into the system's design. None of these features were adopted by Davies or Roberts, neither of whom was concerned with survivability.[38] On the other hand, aspects of Baran's system that would be useful in a variety

of situations—such as high-speed transmission, adaptive routing, and efficient packet switching—were adopted for use in later systems.[39]

One thing that Baran, Davies, and Roberts had in common was the insight that the capabilities of a new generation of small but fast computers could be harnessed to transcend the limitations of previous communications systems. Telephone systems of the late 1960s did not use computerized switches, and message switching systems used large, expensive computers that handled messages slowly. When presented with the idea that a network could employ dozens of computers as its switches, people in the communications industry tended to doubt that computers fast and cheap enough to make this idea feasible would be available (Baran 1990, pp. 19–21; Roberts 1978, p. 1307; Roberts 1988, p. 150; Campbell-Kelly 1988, p. 8). Indeed, the first of these small but powerful "minicomputers" did not appear until 1965, when the Digital Equipment Corporation introduced its PDP-8. The fact that packet switching relied on an innovative computer product helps to explain why that technique was consistently explored by computer scientists but not by communications experts, even though it drew on aspects of both fields.

In the 1960s, computing technologies became policy instruments both in the United States and in the United Kingdom. In the United Kingdom, intervention in the computer industry was seen as a symbol of the Labour Party's commitment to modernization and as an engine of economic growth, and the government made efforts to fund research and coordinate industrial production. In the United States, technological prowess was seen as a weapon in the Cold War, and defense-related research was generously funded through organizations such as the Rand Corporation and ARPA. In both countries, individuals and organizations interested in pursuing computer networking often found it necessary to join government-sponsored projects or to present their work as responsive to contemporary political agendas.

Although computer networking had a political role in both countries, there were striking differences in the levels of government funding, in policy makers' interpretation of networks as a military or a civilian technology, and in government's inclination to intervene in private enterprise. These differences are evident in the contrasting outcomes of the attempts by the NPL and ARPA to build large-scale networks. The United States poured much more money into basic computing research than did the United Kingdom, and most of that money was channeled through the Department of Defense. Not only did Roberts have a generous budget for his project; he also was able to call on computer experts from around the country to help build the network. Davies, at the NPL, had a much smaller budget. Faced with a perceived economic crisis and convinced of the need to compete with the United States and other exporters of high technology, the British government tried to rationalize the computing industry and to encourage commercial spinoffs of government research. Eventually much of the research at the NPL and at similar places was directly focused on short-term commercial applications, and the Labour government's industrial policy limited Davies's choice of computers. The US government was less inclined to try to manage the domestic computer industry. Overall, Roberts had much more support and much less interference from his government than Davies had from his.

Davies had been one of the earliest and most articulate advocates of packet switching. He had formulated a detailed plan for a national network at a time when the ARPANET was still just an idea. Yet by the middle of 1968 Davies was already lamenting that his project had been eclipsed by the American effort: "As a force in this discussion NPL is too remote and our own demonstration as planned now is small-scale and likely to be delayed by the reductions in staff and administrative difficulties in purchasing computers." (Davies 1968b, p. 7) Despite their technological vision, neither Baran nor Davies could find the backing to build a national packet switching network. Roberts, in contrast, was able to make the ARPANET an internationally recognized symbol of the feasibility of packet switching only a few years after he learned of the technique.

The fact that packet switching had to be integrated into local practices and concerns led to very different outcomes in the three network projects. Some visions of packet switching were easier to implement, some turned out to be a better match for evolving computer technology, and some were more attractive to organizations in a position to sponsor network projects. Making packet switching work was not just a matter of having the right technical idea; it also required the right environment. Only after the ARPANET presented a highly visible example of a successful packet switching system did it come to be seen as a self-evidently superior technique. The success of the ARPANET may have depended on packet switching, but it could equally well be argued that the success of packet switching depended on the ARPANET.