Nomadic computing (*keynote address*) *

Leonard Kleinrock

Computer Science Department, University of California, Los Angeles, CA 90095-1596, USA

Nomadic computing is a phenomenon in computing and communications that is spreading rapidly. At this stage of its technology, the key problems and the basic understanding of its underlying principles are only beginning to be identified. Analysis and design tools are needed to assist in its development. In this paper, we discuss the nature of the tools and what one might look for in the way of applying some of them. We then pose some of the essential issues of nomadic computing and communications.

1. Introduction

The Internet, with its World Wide Web (WWW) interface has totally revolutionized the way information is accessed in all sectors of the economy: commercial, educational, government, consumer, etc. E-mail, Web browsing, file transfer, access to vast numbers of information sources, people-to-people interaction, and much more are familiar services to tens of millions of people; see figure 1. These global services allow today's users to go almost anywhere they choose and still have access to the Internet. However, the system architecture to support "nomads" as they travel from one location to another is not yet in place. We have "piece parts" available, but no integrated systems support for nomadic computing, and only rudimentary analysis and design tools, both of which we address in this paper.

2. Brief history

Where did all this "Internet" stuff come from? It did not happen overnight. It has been growing exponentially from the time the ARPANET (which later evolved into the Internet) came to life in September, 1969, in the UCLA laboratory headed by the author. And much of its growth has been due to the continued support from the Advanced Research Projects Agency (ARPA). To trace the role that ARPA played, it is useful to divide the history into five "waves" as shown below. In these summaries, we identify the technical and analytical aspects of the waves, rather than the business or social aspects.

* This work was supported by the Advanced Research Projects Agency, DARPA/ITO, under Contract MDA-972–91-J-1011 "Advanced Networking and Distributed Systems" and Contract DBT-63-C-0080 "Transparent Virtual Mobile Environment".

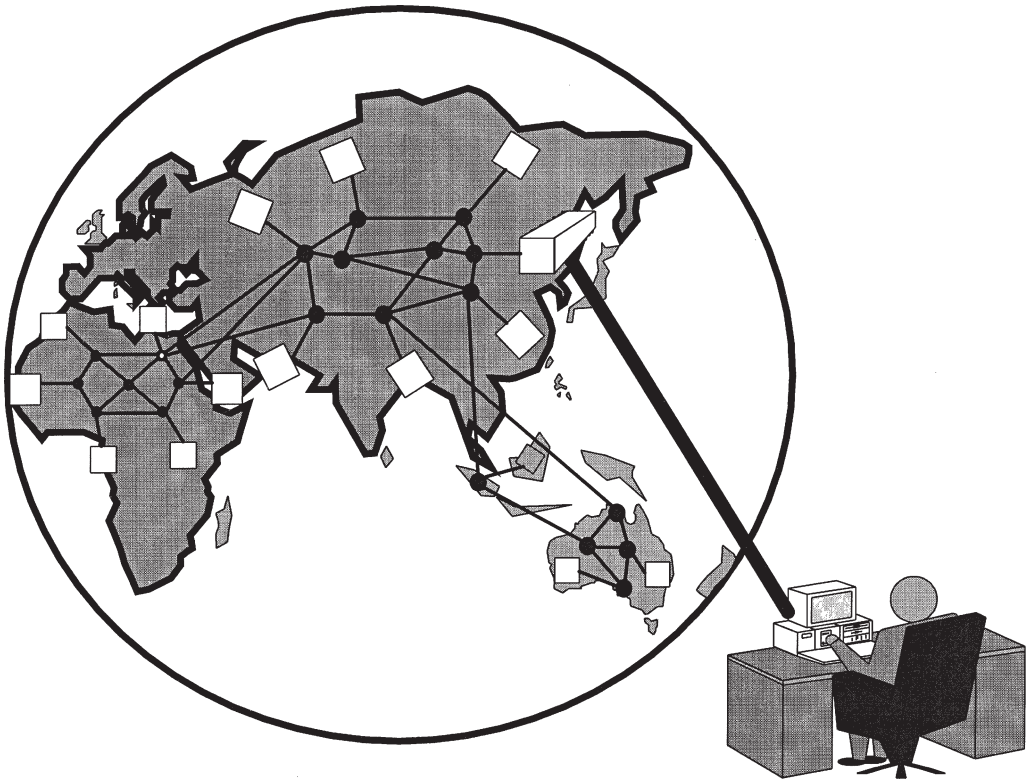

Figure 1. The World Wide Web.

1. The first wave: ARPANET and packet switching [6–8,10–12,16,18,32,33,35,38]. The key ideas, innovations and lessons learned from the invention of packet switching include the following: resource sharing, distributed control, adaptive routing, separation of switch from Host computer, unattended operation, packetized streams, pipelined packets, hop-by-hop transmission, layered protocols. All of these launched the era of effective data networking. Among all of them, the dominant idea was demand multiplexing, which provided access to a resource (bandwidth, cycles, storage, etc.) only when it was needed, and not on a static basis as had been the case before the ARPANET.

2. The second wave: packet satellite [1,4,5,18,24,25,30,34].
In the early 1970's, a satellite link was added to the ARPANET. This link was used in a multiaccess fashion with demand multiplexing, an innovation at the time. The key issues were: propagation delay is critical; the link protocol for a satellite link is different from that of a land-line; the channel is multiaccess broadcast in nature; the distributed nature of the channel is a dominant factor. One of the new phenomena discovered was that the formation of a queue, which is easy and natural in most queueing systems, is a difficult issue in multiaccess broadcast distributed

channels. One cannot find out who is waiting for service without incurring a cost either in collisions, wasted slots, or control channel overhead.

3. The third wave: packet radio [9,14,18,27–29,39].
   In the mid 1970's, ARPA launched an effort in ground radio packet switching. The idea was to create a network of mobile wireless terminals that could create an instant infrastructure with no base stations. A number of access methods for one-hop wireless communications were developed, analyzed and deployed, including: ALOHA, CSMA, URN, collision resolution algorithms, and virtual time CSMA. Similarly, for multi-hop wireless communications the main issues were: hidden terminals, power control, routing, searching, reduced state description, etc. One of the key issues was how to get effective spatial reuse of the channel.
4. The fourth wave: local area networks [26,31,37].
   In the early 1980's, as personal computers began to appear, so did local area networks which were needed to connect these PC's together. The research in the third wave on packet radio had developed the Carrier Sense Multiple Access (CSMA) protocol; once one added Collision Detection to it, to produce CSMA/CD, Ethernet was born and quickly captured the LAN marketplace. However, a number of other access methods were developed for LANs including: Expressnet, Fasnet, Token Ring, FDDI, DQDB, and hub-based LANs.
5. The fifth wave: all the rest.
   In the 1980's and 1990's we have seen continued support by ARPA and other funding agencies in a variety of key areas, most of which have brought us to mobile computing and nomadic computing. The technologies of interest are: parallel architectures and algorithms, teraop machines, fast packet switching, gigabit networks, fiber optics, distributed processing, high-speed interconnect nets, distributed databases, distributed control, networks of workstations, wireless networks, cellular radio, nomadic computing.

In passing through these waves of technology, a number of lessons have been learned. Some of these are:

1. Distributed control works.
2. Demand multiplexing pays.
3. Virtual circuits work well.
4. Topological redundancy is easy.
5. Error control is cheap and necessary.
6. Flow control is essential and dangerous.
7. Nobody wanted networking in the early days.
8. Everybody loves networking today.
9. The economics of the component technologies made networks inevitable.

10. The communication and computer industries, after decades of rivalry, are finally cooperating in integrated products and services.

3. Performance issues

In the spirit of the conference theme for this 3rd INFORMS Telecommunications Conference, namely, the application of the tools and theory of operations research to telecommunications, let us consider the computer and communication systems that we included in the previous section. They all present challenges to the performance analyst. Those systems are difficult to analyze for a number of reasons. They are large, complex, heterogeneous, distributed, dynamic, and stochastic. In order to carry out a performance analysis, we must bring to bear certain tools of the trade. These tools can be grouped into one of six classes as shown in table 1. Following each tool class in the list, we comment on the problems associated with that tool class.

Faced with the problems associated with each of these tool classes, it is apparent that an effective approach to performance analysis is to use a hybrid mix of these tools, using each where it is most effective. This type of approach has been under study at UCLA [3] among other places.

Frequently, it is useful to begin a performance evaluation effort with mathematical modeling. Modeling is often quite effective in explaining the principles underlying the behavior of a system. It can also help to identify the important parameters of system behavior. We devote the rest of this section to specific cases where this author has successfully applied the technique of mathematical modeling and analysis to either uncover basic principles or to identify the core of a difficult problem.

Let us begin by considering a new queueing model (motivated by the study of parallel processing systems) which is extremely difficult to solve (see, for example, [23]). In ordinary queueing systems, a single server is assumed to be available, and that server offers service at the rate of 1 second per second of elapsed time. Moreover, it is usually assumed that the customer can take advantage of service at the rate of 1 sec/sec. However, there are cases (as in parallel processing systems) where the

Table 1

Tool class	Problems
Mathematical analysis	Limited in its ability to solve complex models that include: non-stationary behavior, coupled queues, finite buffers, etc.
Numerical evaluation	The complexity of the evaluation algorithm tends to grow exponentially with the size of the problem.
Iterative solution	It is difficult to predict when the iteration will converge, and at what rate.
Simulation	Gives answers for specific system configurations and parameters. As a result, it is difficult to search a large parameter space.
Emulation	Tends to be expensive in terms of the equipment required and is cumbersome to configure.
Build and measure	Can only be done after deployment, and is essentially guaranteed to bankrupt you.

customer (a processing job) can take advantage of $n(t)$ processors at time t, where $n(t)$ varies. We also may assume that the total service capacity (say, N sec/sec) is available to a queue of jobs, where the job at the head of the queue has preemptive priority for this capacity; however, when this job cannot use all N sec/sec, then the additional capacity is offered to the next job in the queue, etc. No general solution to this queueing problem has been found to date.

As another example, we examine a common feature of the many communication systems described in the five waves above. In most of those systems, it turns out that there are three parameters that interact; these are:

$C =$ capacity of the communication channel (say, in megabits/sec),

$L =$ length of the channel (say, in kilometers),

$b =$ length of the data unit (e.g., a packet) transmitted (say, in bits).

These three can be combined into a single key system variable, the *latency*, which we denote by a, and which is defined as the propagation delay (time for a bit to travel the length of the channel) divided by the time it takes to transmit a packet. It turns out that the system performance in these systems is closely tied to the latency. If we assume that it takes 5 microseconds for energy to travel through one kilometer of the channel, then the latency is simply [20]

$$a = 5LC/b$$

since $5L$ is the propagation delay through the channel and b/C is the time (in microseconds) to transmit a packet. It is interesting to observe the range of values taken on by a for some characteristic systems, and these are shown in table 2. The thing to note from this table is the enormous range over which the key parameter, latency, varies (6 orders of magnitude!).

In evaluating the performance of a system we often compare the mean response-time (T) with the throughput (γ) of a system. This profile often looks like that shown in figure 2.

We note that at low throughput we get good response time, and at high throughput, we get poor response time. So a natural question arises regarding the most effective trade-off between T and γ. We have proposed a single performance measure, power (P), that combines these two. We define power as [19]

$$P = \gamma/T.$$

Table 2

	Bandwidth (Mbps)	Packet length (bits)	Propagation delay (microsec)	Latency (a)
LAN	10.00	1,000	5	0.05
WAN	0.05	1,000	20,000	1.00
Satellite link	0.05	1,000	250,000	12.50
Cross country fiber link	1,000.00	1,000	20,000	20,000.00

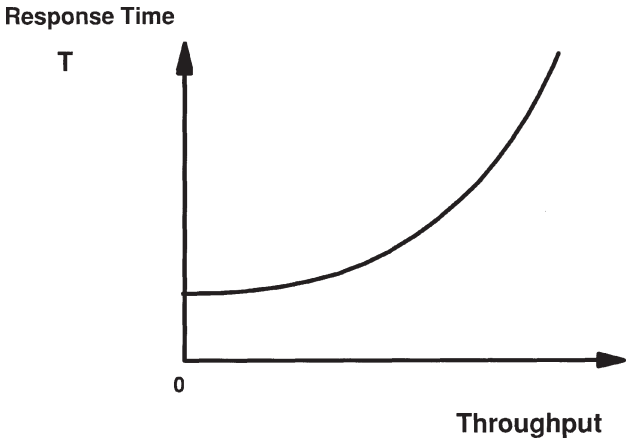

Figure 2. A typical response time – throughput profile.

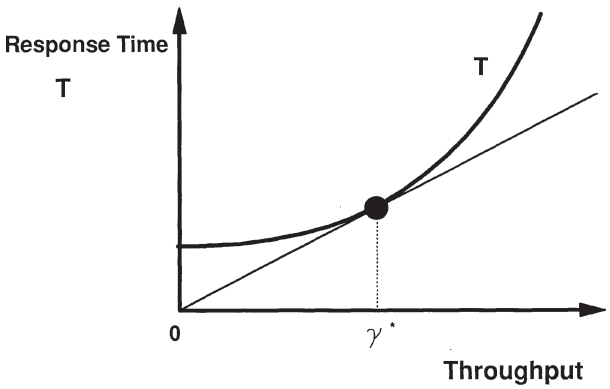

Figure 3. The operating point at maximum power.

It turns out that power is maximized at that point on the response time – throughput profile where a straight line from the origin first becomes tangent to the profile; see figure 3 where we denote the optimum throughput operating point by γ^*.

This result is good for *any* queueing system and *any* flow control system. Interestingly, for all M/G/1 queueing systems [17], this point occurs where $E[N]$, the average number of customers (jobs, messages, packets, etc.) in the system, is *exactly one*! What makes this interesting is that it is intuitively the correct operating point for deterministic systems [19].

In the case of packet radio systems, a totally different consideration leads to exactly the same result we just quoted. Let us consider an ideal multi-hop packet radio system where the power in every radio is adjusted so that each hop covers exactly a radius R. Further assume that the total distance a message must travel is $D \gg R$. Now let $T(R)$ be the average delay (due to interfering traffic from other radios) experienced by a message in traveling one hop. It is clear that if we choose R

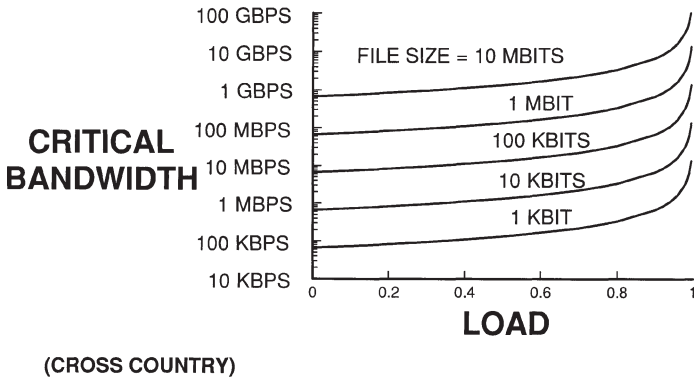

Figure 4. The critical bandwidth for various file sizes.

to be large, then $T(R)$ is large (more interference), but the number of hops is small, and vice versa for a small value of R. The total delay along the path is clearly $T(R)$ D/R. If we now solve for the optimum value of R so as to minimize the total delay, we find that the solution is at exactly the same point as that shown in figure 3. We call this "giant stepping".

As a last example of the power of analytic modeling, let us consider the latency/bandwidth tradeoff in communication [20]. Suppose we choose to transmit a message of b bits through a channel whose capacity is C megabits/sec and whose length is L kilometers (as in our earlier latency discussion). The mean total time T (in microsec) it takes to deliver this message from its source to its destination consists of three components:

$Q =$ the mean time the message spends waiting in the queue for its turn to transmit,

$b/C =$ the time it takes to "pump" the message into the channel,

$5L =$ the propagation time for the last bit to make its way across the channel.

That is,

$$T = Q + b/C + 5L.$$

We note that the time at the source is essentially $Q + b/C$ whereas the propagation time is $5L$. It turns out that when the time at the source is far greater than the propagation time, then the system can be said to *bandwidth limited*; in this regime, more bandwidth can help in reducing T. On the other hand, if the propagation time dominates, the system can be said to be *latency limited*; in this regime, more bandwidth is not important since the overall delay is now dominated by the speed of light. We choose to define a sharp boundary between these two regimes as that point where these two are exactly equal, namely $Q + b/C = 5L$. Note that if we temporarily assume that $Q = 0$ (i.e., no load), then we find that this point is exactly when $a = 1$, where a is our previously defined latency measure. We define C_{crit}, the critical capacity of the system, to be that bandwidth which puts the system exactly at the point where $Q + b/C = 5L$. In figure 4, we show the critical capacity as a function of system

load (i.e., the utilization factor of the channel) for various file sizes, where we have assumed an M/M/1 queueing system [17] for calculating Q.

In this figure, we have assumed that the channel spans the length of the USA. For an arbitrary length channel, the equation for C_{crit} is

$$C_{crit} = b/\big[5L(1-\rho)\big],$$

where ρ is the utilization factor. This condition may be expressed as a condition on a, namely, $a = 1/(1-\rho)$. The boundary which separates these two regimes may also be seen on an interesting graph which can be found in [20].

4. Nomadicity

Let us now return to the issue of nomadic computing. The combination of portable computing with portable communications is changing the way we think about information processing. We now recognize that access to computing and communications is necessary not only from one's "home base", but also while one is in transit and when one reaches one's destination.

But just what is nomadic computing? Whenever you change the computing platform you are using (e.g., start using your laptop when you leave the desktop machine in your office), or change your communications mode, or travel from one location to another, you often face severe compatibility problems. The files on your different computers may not be consistent, your ability to receive video clips disappears, you have no idea how to access a printer in a new location, etc. These dramatic changes in capability brought on by your moving from one place with one set of capabilities to another place with very different capabilities is a source of great frustration today. The field of nomadic computing and communications is emerging to provide solutions and support for your nomadic behavior. The long range goal is to make such discontinuities fundamentally transparent to the nomad through sophisticated systems support. These ideas form the essence of a major shift to *nomadicity* (nomadic computing and communications) [2,15,21,22,36].

We are interested in those capabilities that must be put in place to support nomadicity. The desirable characteristics for nomadicity include independence of location, of motion, of computing platform, of communication device, of communication bandwidth, and with widespread presence of access to remote files, systems and services. The notion of independence here does not refer to the quality of service one sees, but rather to the perception of a computing environment that automatically *adjusts* to the processing, communications and access available at the moment. For example, the bandwidth for moving data between a user and a remote server could easily vary from a few bits per second (in a noisy wireless environment) to hundreds of megabits per second (in a hard-wired ATM environment); or the computing platform available to the user could vary from a low-powered Personal Digital Assistant while in travel to a powerful supercomputer in a science laboratory. Indeed, today's systems treat

radically changing connectivity or bandwidth/latency values as exceptions or failures; in the nomadic environment, these must be treated as the usual case. Moreover, the ability to accept partial or incomplete results is an option that must be made available due to the uncertainties of the informatics infrastructure.

The ability to automatically adjust all aspects of the user's computing, communication and storage functionality in a transparent and integrated fashion is the essence of a nomadic environment.

It is clear that a great many issues regarding nomadicity arise whether or not one has access to wireless communications. However, with such access, a number of interesting considerations arise [13]. Access to wireless communications provides two capabilities to the nomad. First, it allows him to communicate from various (fixed) locations without being connected directly into the wireline network. Second, it allows him to communicate while traveling. Although the bandwidth offered by wireless communication media varies over an enormous range as does the wireline network bandwidth, the nature of the error rate, fading behavior, interference level, mobility issues etc., for wireless are considerably different so that the algorithms and protocols require some new and different forms from that of wireline networks. For example, the network algorithms to support wireless access are far more complex than for the wireline case. Whereas the location of a user or a device is a concern for wireline nets as described above, the details of tracking a user while moving in a wireless environment add to the complexity and require rules for handover, roaming, etc.

There are a number of reasons why nomadicity is of interest. For example, nomadicity is clearly a *newly emerging technology* that already surrounds the user. Indeed, this author judges it to be a *paradigm shift* in the way computing will be done in the future. Information technology trends are *moving in this direction*. Nomadic computing and communications is a *multidisciplinary* and *multiinstitutional effort*. It has a huge *potential for improved capability* and convenience for the user. At the same time, it presents at least as huge *a problem in interoperability* at many levels. The contributions from any investigation of nomadicity will be mainly at the *middleware* level. The products that are beginning to roll out have a *short term focus*; however, there is an enormous level of interest among vendors (from the computer manufacturers, the networking manufacturers, the carriers, etc.) for long range development and product planning, much of which is *now underway*. Whatever work is accomplished now will certainly be of *immediate practical use*.

5. Conclusion

In this paper we have presented nomadicity as a new paradigm in the use of computer and communications technology and have laid down a number of challenging problems. As in all complex systems, the problem of performance evaluation is important and difficult. In such a situation, one must draw upon any tools that are

available to develop the basic understanding of the underlying system behavior. In the case of nomadic systems, it is clear that our existing physical and logical infrastructure must be extended to support nomadicity. The implication is that we must account for nomadicity at this early stage in the development and deployment of the NII (National Information Infrastructure); failure to do so will seriously inhibit the growth of nomadic computing and communications.

References

[1] N. Abramson, Packet switching with satellites, in: *AFIPS Conference Proceedings* 42 (1973) pp. 695–702.
[2] R. Bagrodia, W. Chu, L. Kleinrock and G. Popek, Vision, issues, and architecture for nomadic computing, IEEE Personal Communications (1995) 14–27.
[3] R. Bagrodia and C.C. Shen, MIDAS: integrated design and performance evaluation of distributed systems, IEEE Transactions on Software Engineering (1991) 1049–1058.
[4] R. Binder, N. Abramson, F.F. Kuo, A. Okinaka and D. Wax, ALOHA packet broadcasting – a retrospect, in: *AFIPS Conference Proceedings* 44 (1975) pp. 203–215.
[5] S. Butterfield, R. Rottberg and D. Walden, The satellite IMP for the ARPA network, in: *Proceedings of 7th Hawaii International Conference on System Sciences* (1974) pp. 70–73.
[6] C.S. Carr, S.D. Crocker and V.G. Cerf, Host–Host communication protocol in the ARPA network, SJCC (1970) 589–597.
[7] S.D. Crocker, J.F. Heafner, R.M. Metcalfe and J.B. Postel, Function-oriented protocols for the ARPA computer network, SJCC (1972) 271–280.
[8] H. Frank, I.T. Frisch and W. Chou, Topological considerations in the design of the ARPA computer network, SJCC (1970) 581–587.
[9] H. Frank, I. Gitman and R. van Slyke, Packet radio system – network considerations, in: *AFIPS Conference Proceedings* 44 (1975) pp. 217–231.
[10] H. Frank, R.E. Kahn and L. Kleinrock, Computer communication network design – experience with theory and practice, SJCC (1972) 255–270.
[11] G.L. Fultz and L. Kleinrock, Adaptive routing techniques for store-and-forward computer-communication networks, in: *Proceedings of the IEEE International Conference on Communications* (1971) pp. 39-1–39-8.
[12] F.E. Heart, R.E. Kahn, S.M. Ornstein, W.R. Crowther and D.C. Walden, The interface message processor for the ARPA computer network, SJCC (1970) 551–567.
[13] R. Jain, J. Short, L. Kleinrock, S. Nazareth and J. Villasenor, PC-notebook based mobile networking: algorithms, architectures and implementations, in: *International Conference on Communications* (1995) pp. 771–777.
[14] R.E. Kahn, The organization of computer resources into a packet radio network, in: *AFIPS Conference Proceedings* 44 (1975) pp. 177–186.
[15] R.H. Katz, Adaptation and mobility in wireless information systems, IEEE Personal Communications Magazine 1(1) (1994) 6–17.
[16] L. Kleinrock, Analytic and simulation methods in computer network design, SJCC (1970) 569–579.
[17] L. Kleinrock, *Queueing Systems, Vol. I: Theory* (Wiley, New York, 1975).
[18] L. Kleinrock, *Queueing Systems, Vol. II: Computer Applications* (Wiley, New York, 1976).
[19] L. Kleinrock, Power and deterministic rules of thumb for probabilistic problems in computer communications, *International Conference on Communications* (1979) 43.1.1–43.1.10.
[20] L. Kleinrock, The latency/bandwidth tradeoff in gigabit networks, IEEE Communications Magazine 30(4) (1992) 36–40.

[21] L. Kleinrock, Nomadic computing – an opportunity, Computer Communications Review, ACM SIGCOMM 25(1) (1995) 36–40.
[22] L. Kleinrock, chairman, *Nomadicity: Characteristics, Issues and Applications* (Nomadic Working Team of the Cross Industrial Working Team, 1995).
[23] L. Kleinrock and J.H. Huang, On parallel processing systems: Amdahl's law generalized and some results on optimal design, IEEE Transactions on Software Engineering (special issue on performance evaluation methodology) 18(5) (1992) 434–447.
[24] L. Kleinrock and S.S. Lam, Packet switching in a slotted satellite channel, in: *AFIPS Conference Proceedings* 42 (1973) pp. 703–710.
[25] L. Kleinrock and S.S. Lam, On stability of packet switching in a random multi-access broadcast channel, in: *Proceedings of 7th Hawaii International Conference on System Sciences* (1974).
[26] L. Kleinrock and M. Scholl, Packet switching in radio channels: new conflict-free multiple access schemes, IEEE Transactions on Communications 28 (1980) 1015–1029.
[27] L. Kleinrock and F.A. Tobagi, Carrier sense multiple access for packet switched radio channels, in: *Proceedings of the IEEE International Conference on Communications* (1974) pp. 21B-1–21B-7.
[28] L. Kleinrock and F.A. Tobagi, Random access techniques for data transmission over packet switched radio channels, in: *AFIPS Conference Proceedings* 44 (1975) pp. 187–201.
[29] L. Kleinrock and F.A. Tobagi, Packet switching in radio channels: Part I – carrier sense multiple-access modes and their throughput delay characteristics, IEEE Transactions on Communications 23 (1975) 1400–1416.
[30] S.S. Lam and L. Kleinrock, Dynamic Control Schemes for a Packet Switched Mult-access Broadcast Channel, in: *AFIPS Conference Proceedings* 44 (1975) pp. 143–153.
[31] R.M. Metcalfe and D.R. Boggs, Ethernet: distributed packet switching for local computer networks, Communications of the ACM (1976).
[32] S.M. Ornstein, F.E. Heart, W.R. Crowther, H.K. Rising, S.B. Russell and A. Michel, The terminal IMP for the ARPA computer network, SJCC (1972) 243–254.
[33] L.G. Roberts, Extensions of packet communication technology to a hand held personal terminal, SJCC (1972) 295–303.
[34] L.G. Roberts, Dynamic allocation of satellite capacity through packet reservation, in: *AFIPS Conference Proceedings* 42 (1973) pp. 711–716.
[35] L.G. Roberts and B.D. Wessler, Computer network development to achieve resource sharing, SJCC (1970) 543–549.
[36] J. Short, R. Bagrodia and L. Kleinrock, Mobile Wireless Network System Simulation, in: *ACM Mobile Computing and Networking Conference (MOBICOM '95)* (1995) pp. 195–205.
[37] W. Stallings, *Local Networks – An Introduction* (Macmillan Publishing Company, 2nd ed., 1987).
[38] R.H. Thomas and D.A. Henderson, McROSS – A multi-computer programming system, SJCC (1972) 281–294.
[39] F.A. Tobagi and L. Kleinrock, Packet switching in radio channels: Part II – the hidden terminal problem in carrier sense multiple-access and the busy tone solution, IEEE Transactions on Communications 23 (1975) 1417–1433.

Leonard Kleinrock is Professor of Computer Science at the University of California, Los Angeles. He received his Ph.D. degree from the M.I.T. His research interests focus on nomadic computing, performance evaluation of high speed networks and paralled and distributed systems. He has had over 200 papers published and is the author of six books. He is a member of the National Academy of Engineering, is a Guggenheim Fellow, an IEEE Fellow, and the recipient of the Lanchester Prize, the C.C.N.Y. Townsed Harris Medal, the L.M. Ericsson Prize, the Marconi International Fellowship Award, and the Harry H. Goode Award. He has received numerous best paper and teaching awards, including the ICC Prize Winning Paper Award.