Advanced Communications Systems

Jonathan S. Turner

WUCS-86-21

September 1986

Department of Computer Science
Washington University
Campus Box 1045
One Brookings Drive
Saint Louis, MO 63130-4899

with Mark A. Franklin, Robert E. Morley, Shahid Akhtar, Victor Griswold, Mark Hunter, Shabbir Khakoo, George Robbert, James Sterbenz and Bernard Waxman

ADVANCED COMMUNICATIONS SYSTEMS

Jonathan S. Turner (PI)

The Advanced Communications Systems Project is concerned with new communications technologies that can support a wide range of different communications applications in the context of of large public networks. Communications networks in common use today have been tailored to specific applications and while they perform their assigned functions well, they are difficult to adapt to new uses. There currently are no general purpose networks, rather there are telephone networks, low-speed data networks and cable television networks. As new communications applications proliferate, it becomes clear that in the long term, a more flexible communications infrastructure will be needed. The Integrated Services Digital Network concept provides a first step in that direction. We are concerned with the next generation of systems that will ultimately succeed ISDN.

The main focus of the effort in the ACS project is a particular switching technology we call broadcast packet switching. The key attributes of this technology are (1) the ability to support connections of any data rate from a few bits per second to over 100 Mb/s, (2) the ability to support flexible multi-point connections suitable for entertainment video, LAN interconnection, and voice/video teleconferencing, (3) the ability to efficiently support bursty information sources, (4) the ability to upgrade network performance incrementally as technology improves and (5) the separation of information transport functions from application-dependent functions so as to provide maximum flexibility for future services. The reader is referred to reference [1] for additional background on the project.

The ACS project officially began in January 1986 with support from Bell Communications Research and Italtel SIT. This is the first annual progress report and covers the period though September 1, 1986. In our original research plan, this first year was to be devoted largely to planning and assembling a research group. We have been successful in our major organizational goals and have been able to go beyond our initial plans to achieve significant progress in several research areas. This progress is summarized in the following section of the report. This section is followed by a report on administrative issues, including funding, staffing and space. The bulk of this report is a series of appendices, including all technical reports and publications of the project in the past year.

Research Progress and Plans

The research program of the ACS project can be divided into four major areas: (1) switching system architecture, (2) connection management, (3) network control problems, such as routing and congestion control and (4) design of communications applications in the context of broadcast packet networks. The effort in the last year has concentrated on switching system architecture and related issues, although significant progress has also been made in connection management and network control. Work on application design is being deferred until issues in the other areas are better understood.

Published Papers

"Design of a Broadcast Packet Switching Network," by Jonathan S. Turner, *Proceedings of Infocom 86*, pp. 667–675, 4/86. Also, to appear in *IEEE Transactions on Communications.*

"New Directions in Communications," by Jonathan S. Turner, *Proceedings of the Zurich Seminar on Digital Communications*, 3/86, 25–32. Also, to appear in *IEEE Communications Society Magazine.*

"Design of an Integrated Services *Packet* Network," by Jonathan S. Turner, *Proceedings of the Ninth Data Communications Symposium*, 9/85, 124–133. Also, to appear in *IEEE Journal on Selected Areas in Communications.*

Invited Lectures

ITT Advanced Technology Center, Shelton, CT (9/86)
Digital Equipment Corporation, Littleton, MA (5/86)
Bolt, Beranek and Newman, Cambridge, MA (5/86)
University of California, Berkeley, CA (3/86)
Carnegie-Mellon University, Pittsburgh, PA (2/86)
AT&T Bell Laboratories, Holmdel, NJ (12/85)
IBM Research, Yorktown Heights, NY (12/85)
GTE Laboratories, Waltham, MA (11/85)
MIT, Cambridge, MA (11/85)
AT&T Bell Laboratories, Naperville, IL (11/85)
Bell Communications Research, Red Bank, NJ (10/5)
General Electric Research, Schenectedy, NY (9/85)
Bell Communications Research, Morristown, NJ (5/85)

Figure 1: Publications and Invited Lectures

We have been very active in publishing our earlier results on broadcast packet switching. Papers have been presented at three conferences and revised versions of these papers have been accepted for journal publication. Patent applications have been filed on broadcast packet switching and invited lectures have been given at twelve industrial and academic laboratories during the last year. (See Figure 1 for details.) Our work has generated a great deal of interest throughout the world, and appears to be having an influence on the research programs at several major industrial laboratories. We find this impact of our work particularly gratifying and expect to see it continue as our research program develops.

The following subsections summarize the progress we have made in several specific areas during the past year and outline our plans for the coming year. More detailed information can be found in the appendices.

Switch Architecture and Hardware Design

The most novel aspect of our research program is its focus on networks supporting flexible multipoint communication. Any switching system supporting multi-point communication must be able to connect any subset of its incoming channels to any subset of its outgoing channels. This is in contrast to point-to-point switching systems which need only connect input-output pairs.

Figure 2 shows an architecture of a broadcast packet switching system capable of supporting multi-point connections. The figure also illustrates the handling of point-to-point and multi-point

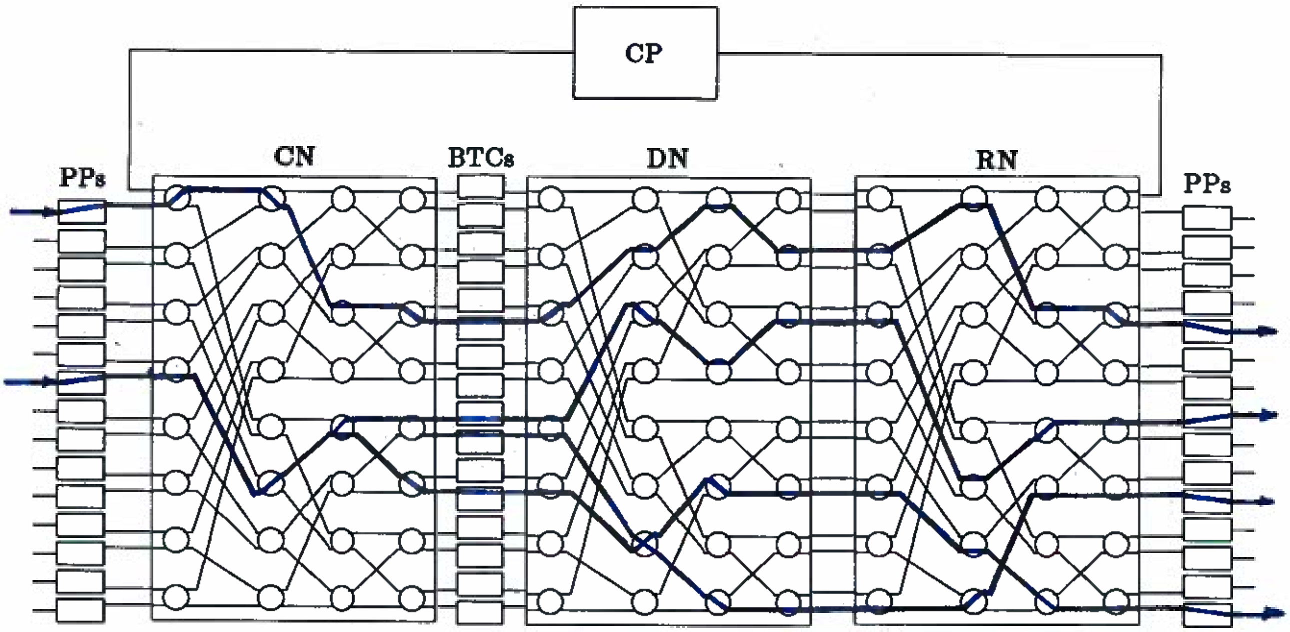

Figure 2: Broadcast Packet Switch Architecture

packets. The system comprises a cascade of buffered packet switching networks. The *Routing Network* (RN) is a conventional binary routing network, routing packets based on successive bits of the destination address. The *Distribution Network* (DN) performs a load distribution function that prevents local congestion in the routing network. The *Copy Network* (CN) performs the packet replication required to support multi-point connections and the *Broadcast Translation Circuits* (BTC) perform a translation that determines how the packets created by the copy network are routed. The *Packet Processors* (PP) perform all per-packet protocol processing, including logical channel translation. The *Connection Processor* (CP) at the top of the figure is a conventional programmable processor that controls setup and modification of connections (it is the call processor in common parlance). The system is designed around a set of custom integrated circuits that we are developing as part of our research program.

In January 1986 we began work on the design of two of the integrated circuits to be used in the laboratory prototype we plan to construct. This work was done in the context of a graduate level VLSI projects course taught in the EE department. We viewed these initial efforts as preliminary, with the primary goal being to get an accurate measure of the scope and complexity of these projects. While the chip designs were not completed during the semester, valuable experience was obtained that will help us to better plan the prototype chips we intend to design and fabricate over the next two years. In addition, this work has given us a head start on these chip designs, which were originally scheduled to begin this fall.

The first of the two chips that is being designed is the switch node that makes up the copy, distribution and routing networks. This chip is a multi-function switch element that can be configured for any of the three networks, with two input and output ports per switch element, each port being four bits wide. The second chip is the broadcast translation circuit which performs the translation for multi-point packets. It contains two random access memories implementing a pair of lookup tables controlling the translation process, plus associated control circuitry. The chips are being designed in a scalable CMOS process with two layers of metal. We expect to fabricate the first chips with a 3 micron feature size, but may scale that down to 1.2 microns at a later time. In the 3 micron implementation, we're expecting each chip to fit comfortably within a standard .5 cm^2 frame in a 64 pin package. The target clock speed for these first chips is 10 Mb/s.

Our plans for the coming year include completing the design, simulation and fabrication of the node and BTC chips. We also plan to design a two chip implementation of the packet processor and a datagram router. These are being tackled in a broader context. We have found that several of the chips we need contain similar parts which are profitably viewed as special cases of a more general *packet transformer*. We are considering the design of a special-purpose silicon compiler that will take as input a specification of a packet transformer and produce a description of a circuit implementing that specification. We expect this approach to greatly reduce the effort required for the design of several of the chips we require. It will also provide a powerful tool for the design of other similar chips.

Performance of Broadcast Packet Switch Fabric

During the past year we have undertaken a series of studies aimed at developing a detailed understanding of the performance of the broadcast packet switch fabric described briefly in the previous section. This work is described in a master's thesis by Richard Bubenik and a technical report by Bubenik and Turner which has been submitted for publication.

The most novel element of this work has been the evaluation of the copy network and the effect on performance of the contention caused by extensive packet replication. The performance evaluation includes both analysis and simulation results, characterizing the dependence of copy network performance on fanout and the location of the active sources. We can place a theoretical bound on the performance degradation possible in the copy network in the worst-case and our simulation results show its behavior in a wide variety of situations both typical and atypical. Summarizing the results briefly, we found that throughput depends strongly on both fanout and the number of active sources with the best performance obtained with fanouts equal to powers of two and many active sources. The worst performance was obtained for fanouts between powers of two and a small number of active sources grouped on adjacent input ports. While the worst-case analysis shows that the factor of two speed advantage provided for the copy network is insufficient to handle the worst-case traffic pattern, our simulation results convince us that it is adequate to handle all but the most contrived of situations.

The performance evaluation work also included a study of architectural alternatives in the design of the routing and distribution networks. The issues examined included, the effect of cut-through operation in the switch nodes on delay and throughput, the effect of the distribution network, the effect of node size and the effect of a modified queueing discipline. Contrary to expectations, our initial results on node sizes showed that networks composed of large nodes performed better than networks composed of small nodes. This turned out to be caused by a subtle effect of the FIFO queueing discipline used by the nodes. The alternative queueing discipline we considered improved performance considerably and gave a further advantage to large nodes over small ones.

Performance evaluation of packet switching fabrics such as the ones used in broadcast packet switches suffers from the lack of a unified theory or even appropriate terminology. Terms like *non-blocking network* which are meaningful in discussions of circuit switching fabrics cannot be applied to packet switching fabrics in any technically precise way. This makes it difficult to compare or classify competing architectures. We feel that our work to date puts us in a good position to begin to correct this situation and we hope to do that in the next year.

Testing of Broadcast Packet Switch Fabrics

The ability to adequately test a switching network is vital to its successful application in a practical communications system. The testing facilities must allow complete coverage of the entire fabric and precise location of failing elements.

While other researchers have developed testing methods for binary routing networks, there has been little work on cascaded networks of the sort used in the broadcast packet switch described above. We have extended earlier work on single networks to cascaded networks, showing how a cascade of m binary routing networks can be tested using $(m + 1)n$ test packets, where n is the number of input ports to the network. Such a test sequence is sufficient to detect and identify any single link or single node fault. We also have shown how to perform the testing when the processor performing the test has access to only one pair of ports on the network.

Connection Management

Connection management refers to the collection of algorithms used to create and maintain multipoint connections in a broadcast packet network. A multi-point connection is intended to be a flexible mechanism that can support a wide variety of different applications. To achieve this flexibility, it must be possible to configure a multi-point connection for different uses. One of the first challenges in creating a useful and practical connection management system is deciding exactly what set of primitive capabilities the network should provide to enable users to configure connections. The subsequent challenge is designing the mechanisms needed to implement these capabilities.

We have identified a method of configuring connections based on the concepts of sub-channels within a connection and permissions. Sub-channels allow a connection to be broken down into several distinct information flows, which can be configured differently but because of their close relationships are controlled by the network in a unified way. Permissions give the user a mechanism for controlling access to sub-channels and assist the network in managing its resources (primarily trunk bandwidth).

Based on these ideas, we have developed a specification of a simple connection management architecture and a series of scenarios showing how it can be used to support a variety of applications including broadcast video and multi-media conferencing. This architecture is described in a draft technical report that appears as an appendix.

The connection management architecture has been designed at several levels of abstraction, with explicit interfaces at each level. The primary abstraction level, from the user's perspective, is the one that defines the interface between the network and a *user agent*. At this level, the network is viewed as a single entity which modifies connections in response to control messages. This level is currently the most well-defined. The next level of abstraction below this defines the interfaces between switching systems in the network and it is at this point that explicit reference must be made to the distributed algorithms and data structures that implement the higher level abstractions. While preliminary work has been done at this level, there is still a great deal that needs to be done. We have also begun looking at a higher level of abstraction corresponding to the interface between user agents. At this level application-dependent issues appear. User agents cooperatively determine how connections should be configured to suit the client applications, and direct the network to configure them via control messages.

Our plans for the coming year include the development of a more detailed specification of the connection management architecture and a preliminary implementation, in the form of a software simulation. We intend to use this simulation to obtain a better understanding of the issues involved in the design and use of a multi-point connection management system. It will also serve as the basis for the prototype system we plan to develop next year.

Routing

The objective of the routing problem is to determine a set of network resources (primarily trunk bandwidth) sufficient to support communication among a specified set of users. In conventional circuit switched networks, all connections require the same amount of bandwidth and (almost all) have exactly two endpoints. Such a network can be described formally as a graph in which each edge has both a capacity and a length. A set of connections for such a network is simply a collection of vertex pairs. A feasible route assignment is an assignment of each connection to a path joining the connection's endpoints that doesn't exceed the capacity of any edge. An optimum routing algorithm is one that can find a feasible assignment whenever one exists.

Of course, this version of the problem is a static one. In a real communications network, the set of connections changes with time and the network must implement a routing policy that manages the changing set of connections in a way that makes it unlikely that a new connection will be blocked. In the interests of efficiency, it is generally assumed that once a connection has been assigned a route, that assignment will remain fixed as long as the connection is present. These considerations lead to a routing policy based on the heuristic strategy of routing connections by the shortest path available at the time the connection is established.

If connections can have an arbitrary bandwidth associated with them, the routing problem becomes a bit more complicated. One must now consider the network to be a graph in which vertices can be joined by multiple edges. To prevent blocking of connections with large bandwidth requirements, new connections should be assigned to the fullest edges with sufficient capacity along the assigned route. This strategy preserves large blocks of bandwidth for use by high speed connections.

In broadcast networks, a connection can involve an arbitrary number of endpoints. A feasible route assignment for a set of connections is an assignment of each connection to a subtree connecting its endpoints, in a way that does not exceed the capacity of any edge. As in the case of point-to-point networks, connections come and go over time, and so the appropriate routing policy is to assign each connection to the subtree with shortest total length available at the time the connection is established. This can be viewed as a generalization of the Steiner tree problem in graphs. This problem is known to be NP-complete, meaning that there is unlikely to be an efficient algorithm that can always find an optimal solution. On the other hand, there is are several efficient algorithms that yield solutions that are close to optimal. The best known one is called the minimum spanning tree heuristic (MST).

Connections in broadcast networks are dynamic in another way. They grow and shrink with time as individual endpoints come and go. The challenge here, is to maintain a good connection topology without doing a great deal of recomputation each time an endpoint is added or dropped. Practical algorithms must be suitable for distributed implementation, with each node making decisions based on local information. The simplest algorithm is a greedy strategy that adds new endpoints by joining them to the connection by the shortest available path and dropping branches of the connection tree when endpoints drop out.

Our research objective is to develop practical and efficient algorithms that can be used in actual multi-point communication networks. To this end, we have been studying the performance of the MST and greedy algorithms from both a worst-case and average case point of view. A prerequisite for our evaluation of the average case performance, has been the development of a simple probability model that can yield data relevant to real networks. We have developed such a model and have begun using it to evaluate the MST and greedy algorithms. We have shown experimentally, that the average case performance of the MST algorithm is excellent, usually within 5% of optimum. While this algorithm is probably impractical for application in a real network, our results show that it can serve as a useful standard of comparison against which other algorithms may be measured. In particular, we have used it to study the performance of the greedy algorithm in dynamically changing connections. Our results show that the solutions produced are generally within 20% of the value obtained for the MST algorithm. The performance deteriorates during long sequences of deletions, because the algorithm simply prunes rather than re-routing during such sequences. This sort of degradation is not unique to the greedy algorithm, but is intrinsic to any algorithm that makes only incremental changes and does not re-route.

Our research plans include continued experimental evaluation of these algorithms and others. We also hope to attack the average case performance of these algorithms analytically, in order to obtain greater insight into the factors limiting their performance. We also plan to design and implement distributed versions of these algorithms.

Congestion Control

A principal advantage of packet switched networks is their ability to dynamically allocate bandwidth to the users who need it at a particular instant. Since networks are subject to rapid statistical variations in demand, care must be taken to ensure acceptable performance under conditions of peak loading. Congestion control refers to the collection of methods used to ensure each user acceptable performance under a variety of load conditions. The high speed and multi-point connection capability of broadcast packet networks place new demands on congestion control methods.

A prerequisite to the development of an effective congestion control method is an understanding of the impact that bursty sources have on queueing in the network. The popular M/M/1 queueing model, while theoretically tractable and widely applicable, is insufficient to model the behavior of a small number of high speed, but very bursty sources. A key part of our work in congestion control has therefore been to obtain an understanding of such sources. We are focussing on a simple model that treats each source as a two state Markov chain. The source is active in one state and idle in the other. Parameters of the model include average holding times in each state and the rate of packet transmission while active. This model can be used for a wide variety of bursty sources, including coded video. Our results to date indicate that such sources can lead to serious performance degradation if not handled carefully.

The basic congestion control mechanism under consideration involves user specification of several parameters defining peak and average bandwidth requirements, plus a measure of burstiness. The Network Interfaces prevent users from exceeding their specified peak bandwidths, by keeping track of the time between successive packet transmissions. The network also measures average bandwidth and burstiness and discards user packets that exceed the specified limits. The mechanism used here can be viewed as a pseudo-buffer for which the user specifies the serving rate and the buffer size. Whenever the user sends a real packet, the network adds a pseudo-packet to the pseudo-buffer. If this does not cause the pseudo-buffer to overflow, the real packet is immediately accepted by the network. Otherwise it is discarded. Note that only pseudo-packets go into the pseudo-buffer. This mechanism is simple enough to be implemented within packet processor chips at the boundary of the network. We hope to demonstrate that it provides sufficient control to handle a wide range of bursty sources.

One initial conclusion is that it may not be possible to allow complete freedom of selection of congestion control parameters. For example, users with a high peak bandwidth and a large ratio of peak to average bandwidth may have to be limited to short burst lengths. Such a policy could allow bandwidth allocation within the network based simply on average bandwidth. A less restrictive policy would probably require the network to allocate bandwidth in a more complicated fashion, taking into account both the bandwidth and burstiness of the channels occupying a given link when making decisions about accepting new connections.

Administrivia

This first year of the project has necessarily involved start-up activities needed to acquire adequate funding and assemble a strong research group and an environment in which they can operate productively. These efforts are summarized below.

Funding

Our efforts to obtain funding have been quite successful. In May we were awarded a major research grant from the National Science Foundation for $750,000 over three years, starting July 1. This provides a good foundation of support upon which to build and we expect it to help us in our efforts to secure additional funding. The Consortium for Research on Advanced Communications Systems was officially established on January 1 after several months of effort. We have been actively trying to expand the Consortium beyond the original two members, but cannot yet report any new members. Two of our prospective sponsors are in the process of internal reviews and in both cases approval is expected shortly, but no commitments have yet been made.

Last year we applied for a grant from the Defense Advanced Research Projects Agency to support our work. While this application has not been rejected, the chances of its being funded appear slim due to budgetary and political disruptions within DARPA. This situation may change in the future, but for the present we cannot plan on funding from this source.

While the project's funding situation is in fairly good shape at the moment, we can see that additional funding will be required next year if we are to achieve our major goals. The most likely source of new funding in the short term is through expansion of the Consortium to 4–6 members. This could provide adequate funding through September 1988. After that time, a substantial new source of funding will be required. We are beginning to explore possible sources of that funding. The options include an NSF Engineering Research Center grant and a Missouri State Technology Center.

Staffing

For most of the first year, the funded staff on the project has been very limited. The awarding of the NSF grant in May allowed us to support one full-time faculty member plus three full-time and three half-time graduate students during the past summer. The research group includes professors Jonathan S. Turner, Mark Franklin and Robert Morley, with professors Franklin and Morley participating on a part-time basis. The graduate student staff includes three doctoral students and four masters students. The graduate students and their research areas are summarized in Figure 3. We have assembled a strong research group and expect the coming year to be a very productive one.

We hope to expand the staffing by one or two graduate students in the coming year and increase the level of participation by professors Franklin and Morley. The project currently has no professional support staff. We hope to add a full-time technician to support the project and help coordinate the prototyping effort. Our ability to take these steps will be determined by budget constraints.

Space and Facilities

For administrative purposes, the ACS Project has been placed within an existing research center directed by Professor Mark Franklin. This center, previously called the Center for Computer System Design, has been renamed the Computer and Communications Research Center, to reflect its expanded scope.

Name	Degree (exp. graduation date)	Research Area
Shahid Akhtar	MS (5/87)	congestion control
Victor Griswold	DSc (1/90)	connection management
Mark Hunter	MS (1/88)	connection management
Shabbir Khakoo	MS (1/88)	switch architecture
George Robbert	MS (1/88)	switch architecture
James Sterbenz	DSc (1/90)	switch architecture
Bernard Waxman	DSc (1/89)	routing

Figure 3: Graduate Student Staff

In connection with this, the office and laboratory space devoted to the center has approximately doubled. The space was renovated in July and August and we have recently moved into the new space, which includes a central suite of offices housing professors Franklin and Turner plus eight graduate students, on the third floor of Bryan Hall, across from our main laboratory facility. This laboratory houses our VAX 750, and a cluster of terminals for graduate student use and also serves as an informal meeting room. We have obtained a second laboratory on the fifth floor of Bryan which will be devoted to our hardware prototyping efforts. We also have several graduate students located in offices on the fifth floor adjacent to the laboratory.

The center's base of equipment includes the VAX 750 mentioned earlier, a collection of terminals in laboratories and offices, plus an AED color graphics terminal supporting our VLSI design work. We have also recently purchased a Tektronix logic analyzer and IC tester. We need to expand our facilities to accommodate the larger group of graduate students working in the center. Our immediate plans include expanding the terminal handling capabilities of our VAX and purchasing a Microvax workstation which will be used for VLSI design work and as a troff/print server. We expect these steps to handle our needs for the next year, but further expansion will be needed later on. We are participating in the preparation of an application for an NSF equipment grant. If this grant is approved, it should satisfy our equipment needs for several years.

Summary

The first year of the Advanced Communications Systems Project has been a productive one. We have successfully managed the transition from what was essentially a one person operation to a research activity involving approximately ten faculty and graduate students. During this transition we have produced some solid research accomplishments in several different areas. Our funding situation is good and there are several good prospects for additional funding in the works. While the coming year presents many challenges, we are in a good position to meet them and look forward to a very productive and exciting period.

References

[1] Turner, Jonathan S. "Research on Advanced Communications Systems," Washington University, Computer Science Department, Research proposal, 8/85.

List of Appendices