Network Working Group   C. Partridge
Request for Comments: 1152   BBN Systems and Technologies
                                                               April 1990

Workshop Report
Internet Research Steering Group Workshop on
Very-High-Speed Networks

Status of this Memo

This memo is a report on a workshop sponsored by the Internet Research Steering Group. This memo is for information only. This RFC does not specify an Internet standard. Distribution of this memo is unlimited.

Introduction

The goal of the workshop was to gather together a small number of leading researchers on high-speed networks in an environment conducive to lively thinking. The hope is that by having such a workshop the IRSG has helped to stimulate new or improved research in the area of high-speed networks.

Attendance at the workshop was limited to fifty people, and attendees had to apply to get in. Applications were reviewed by a program committee, which accepted about half of them. A few key individuals were invited directly by the program committee, without application. The workshop was organized by Dave Clark and Craig Partridge.

This workshop report is derived from session writeups by each of the session chairman, which were then reviewed by the workshop participants.

Session 1: Protocol Implementation (David D. Clark, Chair)

This session was concerned with what changes might be required in protocols in order to achieve very high-speed operation.

The session was introduced by David Clark (MIT LCS), who claimed that existing protocols would be sufficient to go at a gigabit per second, if that were the only goal. In fact, proposals for high-speed networks usually include other requirements as well, such as going long distances, supporting many users, supporting new services such as reserved bandwidth, and so on. Only by examining the detailed requirements can one understand and compare various proposals for protocols. A variety of techniques have been proposed to permit protocols to operate at high speeds, ranging from clever implementation to complete relayering of function. Clark asserted that currently even the basic problem to be solved is not clear, let alone the proper approach to the solution.

Mats Bjorkman (Uppsala University) described a project that involved the use of an outboard protocol processor to support high-speed operation. He asserted that his approach would permit accelerated processing of steady-state sequences of packets. Van Jacobson (LBL) reported results that suggest that existing protocols can operate at high speeds without the need for outboard processors. He also argued that resource reservation can be integrated into a connectionless protocol such as IP without losing the essence of the connectionless architecture. This is in contrast to a more commonly held belief that full connection setup will be necessary in order to support resource reservation. Jacobson said that he has an experimental IP gateway that supports resource reservation for specific packet sequences today.

Dave Borman (Cray Research) described high-speed execution of TCP on a Cray, where the overhead is most probably the system and I/O architecture rather than the protocol. He believes that protocols such as TCP would be suitable for high-speed operation if the windows and sequence spaces were large enough. He reported that the current speed of a TCP transfer between the processors of a Cray Y-MP was over 500 Mbps. Jon Crowcroft (University College London) described the current network projects at UCL. He offered a speculation that congestion could be managed in very high-speed networks by returning to the sender any packets for which transmission capacity was not available.

Dave Feldmeier (Bellcore) reported on the Bellcore participation in the Aurora project, a joint experiment of Bellcore, IBM, MIT, and UPenn, which has the goal of installing and evaluating two sorts of switches at gigabit speeds between those four sites. Bellcore is interested in switch and protocol design, and Feldmeier and his group are designing and implementing a 1 Gbps transport protocol and network interface. The protocol processor will have special support for such things as forward error correction to deal with ATM cell loss in VLSI; a new FEC code and chip design have been developed to run at 1 Gbps.

Because of the large number of speakers, there was no general discussion after this session.

Session 2: High-Speed Applications (Keith Lantz, Chair)

This session focused on applications and the requirements they impose on the underlying networks. Keith Lantz (Olivetti Research California) opened by introducing the concept of the portable office - a world where a user is able to take her work with her wherever she goes. In such an office a worker can access the same services and the same people regardless of whether she is in the same building with those services and people, at home, or at a distant site (such as a hotel) - or whether she is equipped with a highly portable, multi-media workstation, which she can literally carry with her wherever she goes. Thus, portable should be interpreted as referring to portability of access to services rather than to portability of hardware. Although not coordinated in advance, each of the presentations in this session can be viewed as a perspective on the portable office.

The bulk of Lantz's talk focused on desktop teleconferencing - the integration of traditional audio/video teleconferencing technologies with workstation-based network computing so as to enable geographically distributed individuals to collaborate, in real time, using multiple media (in particular, text, graphics, facsimile, audio, and video) and all available computer-based tools, from their respective locales (i.e., office, home, or hotel). Such a facility places severe requirements on the underlying network. Specifically, it requires support for several data streams with widely varying bandwidths (from a few Kbps to 1 Gbps) but generally low delay, some with minimal jitter (i.e., isochronous), and all synchronized with each other (i.e., multi-channel or media synchronization). It appears that high-speed network researchers are paying insufficient attention to the last point, in particular. For example, the bulk of the research on ATM has assumed that channels have independent connection request and burst statistics; this is clearly not the case in the context of desktop teleconferencing.

Lantz also stressed the need for adaptive protocols, to accommodate situations where the capacity of the network is exceeded, or where it is necessary to interoperate with low-speed networks, or where human factors suggest that the quality of service should change (e.g., increasing or decreasing the resolution of a video image). Employing adaptive protocols suggests, first, that the interface to the network protocols must be hardware-independent and based only on quality of service. Second, a variety of code conversion services should be available, for example, to convert from one audio encoding scheme to another. Promising examples of adaptive protocols in the video domain include variable-rate constant-quality coding, layered or embedded coding, progressive transmission, and (most recently, at UC-Berkeley) the extension of the concepts of structured graphics to video, such that the component elements of the video image are kept logically separate throughout the production-to-presentation cycle.

Charlie Catlett (National Center for Supercomputing Applications) continued by analyzing a specific scientific application, simulation of a thunderstorm, with respect to its network requirements. The application was analyzed from the standpoint of identifying data flow and the interrelationships between the computational algorithms, the supercomputer CPU throughput, the nature and size of the data set, and the available network services (throughput, delay, etc).

Simulation and the visualization of results typically involves several steps:

1. Simulation

2. Tessellation (transform simulation data into three-dimensional geometric volume descriptions, or polygons)

3. Rendering (transform polygons into raster image)

For the thunderstorm simulation, the simulation and tessellation are currently done using a Cray supercomputer and the resulting polygons are sent to a Silicon Graphics workstation to be rendered and displayed. The simulation creates data at a rate of between 32 and 128 Mbps (depending on the number of Cray-2 processors working on the simulation) and the tessellation output data rate is in typically in the range of 10 to 100 Mbps, varying with the complexity of the visualization techniques. The SGI workstation can display 100,000 polygons/sec which for this example translates to up to 10 frames/sec. Analysis tools such as tracer particles and two-dimensional slices are used interactively at the workstation with pre-calculated polygon sets.

In the next two to three years, supercomputer speeds of 10-30 GFLOPS and workstation speeds of up to 1 GFLOPS and 1 million polygons/second display are projected to be available. Increased supercomputer power will yield a simulation data creation rate of up to several Gbps for this application. The increased workstation power will allow both tessellation and rendering to be done at the workstation. The use of shared window systems will allow multiple researchers on the network to collaborate on a simulation, with the possibility of each scientist using his or her own visualization techniques with the tessellation process running on his or her workstation. Further developments, such as network virtual memory, will allow the tessellation processes on the workstations to access variables directly in supercomputer memory.

Terry Crowley (BBN Systems and Technologies) continued the theme of collaboration, in the context of real-time video and audio, shared multimedia workspaces, multimedia and video mail, distributed file systems, scientific visualization, network access to video and image information, transaction processing systems, and transferring data and computational results between workstations and supercomputers. In general, such applications could help groups collaborate by directly providing communication channels (real-time video, shared multimedia workspaces), by improving and expanding on the kinds of information that can be shared (multimedia and video mail, supercomputer data and results), and by reducing replication and the complexity of sharing (distributed file systems, network access to video and image information).

Actual usage patterns for these applications are hard to predict in advance. For example, real-time video might be used for group conferencing, for video phone calls, for walking down the hall, or for providing a long-term shared viewport between remote locations in order to help establish community ties. Two characteristics of network traffic that we can expect are the need to provide multiple data streams to the end user and the need to synchronize these streams. These data streams will include real-time video, access to stored video, shared multimedia workspaces, and access to other multimedia data. A presentation involving multiple data streams must be synchronized in order to maintain cross-references between them (e.g., pointing actions within the shared multimedia workspace that are combined with a voice request to delete this and save that). While much traffic will be point-to-point, a significant amount of traffic will involve conferences between multiple sites. A protocol providing a multicast capability is critical.

Finally, Greg Watson (HP) presented an overview of ongoing work at the Hewlett-Packard Bristol lab. Their belief is that, while applications for high-speed networks employing supercomputers are the the technology drivers, the economic drivers will be applications requiring moderate bandwidth (say 10 Mbps) that are used by everyone on the network.

They are investigating how multimedia workstations can assist distributed research teams - small teams of people who are geographically dispersed and who need to work closely on some area of research. Each workstation provides multiple video channels, together with some distributed applications running on personal computers. The bandwidth requirements per workstation are about 40 Mbps, assuming a certain degree of compression of the video channels. Currently the video is distributed as an analog signal over CATV equipment. Ideally it would all be carried over a single, unified wide-area network operating in the one-to-several Gbps range.

They have constructed a gigabit network prototype and are currently experimenting with uncompressed video carried over the same network as normal data traffic.

Session 3: Lightwave Technology and its Implications (Ira Richer, Chair)

Bob Kennedy (MIT) opened the session with a talk on network design in an era of excess bandwidth. Kennedy's research is focused on multi-purpose networks in which bandwidth is not a scarce commodity, networks with bandwidths of tens of terahertz. Kennedy points out that a key challenge in such networks is that electronics cannot keep up with fiber speeds. He proposes that we consider all-optical networks (in which all signals are optical) with optoelectronic nodes or gateways capable of recognizing and capturing only traffic destined for them, using time, frequency, or code divisions of the huge bandwidth. The routing algorithms in such networks would be extremely simple to avoid having to convert fiber-optics into slower electronic pathways to do switching.

Rich Gitlin (AT&T Bell Labs) gave a talk on issues and opportunities in broadband telecommunications networks, with emphasis on the role of fiber optic and photonic technology. A three-level architecture for a broadband telecommunications network was presented. The network is B-ISDN/ATM 150 (Mbps) based and consists of: customer premises equipment (PBXs, LANs, multimedia terminals) that access the network via a router/gateway, a Network Node (which is a high performance ATM packet switch) that serves both as a LAN-to-LAN interconnect and as a packet concentrator for traffic destined for CPE attached to other Network Nodes, and a backbone layer that interconnects the NODES via a Digital Cross-Connect System that provide reconfigurable SONET circuits between the NODES (the use of circuits minizes delay and avoids the need for implementation of peak-transmission-rate packet switching). Within this framework, the most likely places for near-term application of photonics, apart from pure transport (ie, 150 Mbps channels in a 2.4 Gbps SONET system), are in the Cross-Connect (a Wavelength Division Multiplexed based structure was described) and in next-generation LANs that provide Gigabit per second throughputs by use of multiple fibers, concurrent transmission, and new access mechanisms (such as store and forward).

A planned interlocation Bell Labs multimedia gigabit/sec research network, LuckyNet, was described that attempts to extend many of the above concepts to achieve its principal goals: provision of a gigabit per second capability to a heterogeneous user community, the stimulation of applications that require Gpbs throughput (initial applications are video conferencing and LAN interconnect), and, to the extent possible, be based on standards so that interconnection with other Gigabit testbeds is possible.

Session 4: High Speed Networks and the Phone System (David Tennenhouse, Chair)

David Tennenhouse (MIT) reported on the ATM workshop he hosted the two days previous to this workshop. His report will appear as part of the proceedings of his workshop.

Wally St. John (LANL) followed with a presentation on the Los Alamos gigabit testbed. This testbed is based on the High Performance Parallel Interface (HPPI) and on crossbar switch technology. LANL has designed its own 16x16 crossbar switch and has also evaluated the Network Systems 8x8 crossbar switch. Future plans for the network include expansion to the CASA gigabit testbed. The remote sites (San Diego Supercomputer Center, Caltech, and JPL) are configured similarly to the LANL testbed. The long-haul interface is from HPPI to/from SONET (using ATM if in time).

Wally also discussed some of the problems related to building a HPPI-SONET gateway:

a) Flow control. The HPPI, by itself, is only readily extensible to 64 km because of the READY-type flow control used in the physical layer. The gateway will need to incorporate larger buffers and independent flow control.

b) Error-rate expectations. SONET is only specified to have a 1E-10 BER on a per hop basis. This is inadequate for long links. Those in the know say that SONET will be much better but the designer is faced with the poor BER in the SONET spec.

c) Frame mapping. There are several interesting issues to be considered in finding a good mapping from the HPPI packet to the SONET frame. Some are what SONET STS levels will be available in what time frame, the availability of concatenated service, and the error rate issue.

Dan Helman (UCSC) talked about work he has been doing with Darrell Long to examine the interconnection of Internet networks via an ATM B-ISDN network. Since network interfaces and packet processing are the expensive parts of high-speed networks, they believe it doesn't make sense to use the ATM backbone only for transmission; it should be used for switching as well. Therefore gateways (either shared by a subnet or integrated with fast hosts) are needed to encapsulate or convert conventional protocols to ATM format. Gateways will be responsible for caching connections to recently accessed destinations. Since many short-lived low-bandwidth connections as foreseen (e.g., for mail and ftp), routing in the ATM network (to set up connections) should not be complicated - a form of static routing should be adequate. Connection performance can be monitored by the gateways. Connections are reestablished if unacceptable. All decision making can be done by gateways and route servers at low packet rates, rather than the high aggregate rate of the ATM network. One complicated issue to be addressed is how to transparently introduce an ATM backbone alongside the existing Internet.

Session 5: Distributed Systems (David Farber, Chair)

Craig Partridge (BBN Systems and Technologies) started this session by arguing that classic RPC does not scale well to gigabit-speed networks. The gist of his argument was that machines are getting faster and faster, while the round-trip delay of networks is staying relatively constant because we cannot send faster than the speed of light. As a result, the effective cost of doing a simple RPC, measured in instruction cycles spent waiting at the sending machine, will become extremely high (millions of instruction cycles spent waiting for the reply to an RPC). Furthermore, the methods currently used to improve RPC performance, such as futures and parallel RPC, do not adequately solve this problem. Future requests will have to be made much much earlier if they are to complete by the time they are needed. Parallel RPC allows multiple threads, but doesn’t solve the fact that each individual sequence of RPCs still takes a very long time.

Craig went on to suggest that there are at least two possible ways out of the problem. One approach is to try to do a lot of caching (to waste bandwidth to keep the CPU fed). A limitation of this approach is that at some point the cache becomes so big that you have to keep in consistent with other systems’ caches, and you suddenly find yourself doing synchronization RPCs to avoid doing normal RPCs (oops!). A more promising approach is to try to consolidate RPCs being sent to the same machine into larger operations which can be sent as a single transaction, run on the remote machine, and the result returned. (Craig noted that he is pursuing this approach in his doctoral dissertation at Harvard).

Ken Schroder (BBN Systems and Technologies) gave a talk on the challenges of combining gigabit networks with wide-area heterogeneous distributed operating systems. Ken feels the key goals of wide area distributed systems will be to support large volume data transfers between users of conferencing and similar applications, and to deliver information to a large number of end users sharing services such as satellite image databases. These distributed systems will be motivated by the natural distribution of users, of information and of expensive special purpose computer resources.

Ken pointed to three of the key problems that must be addressed at the system level in these environments: how to provide high utilization; how to manage consistency and synchronization in the presence of concurrency and non-determinism; and how to construct scalable system and application services. Utilization is key only to high performance applications, where current systems would be limited by the cost of factors such as repeatedly copying messages, converting data representations and switching between application and operating system. Concurrency can be used improve performance, but is also likely to occur in many programs inadvertently because of distribution. Techniques are required both to exploit concurrency when it is needed, and to limit it when non-determinism can lead to incorrect results. Extensive research on ensuring consistency and resolving resource conflicts has been done in the database area, however distributed scheduling and the need for high availability despite partial system failures introduce special problems that require additional research. Service scalability will be required to support customer needs as the size of the user community grow. It will require attention both ensuring that components do not break when they are subdivided across additional processors to support a larger user population, and to ensure that performance does to each user can be affordably maintained as new users are added.

In a bold presentation, Dave Cheriton (Stanford) made a sweeping argument that we are making a false dichotomy between distributed operating systems and networks. In a gigabit world, he argued, the major resource in the system is the network, and in a normal operating system we would expect such a critical resource to be managed by the operating system. Or, put another way, the gigabit network distributed operating system should manage the network. Cheriton went on to say that if a gigabit distributed operating system is managing the network, then it is perfectly reasonable to make the network very dumb (but fast) and put the system intelligence in the operating systems on the hosts that form the distributed system.

In another talk on interprocess communication, Jonathan Smith (UPenn) again raised the problem of network delay limiting RPC performance. In contrast to Partridge's earlier talk, Smith argued that the appropriate approach is anticipation or caching. He justified his argument with a simple cost example. If a system is doing a page fetch between two systems which have a five millisecond round-trip network delay between them, the cost of fetching n pages is:

$$5 \text{ msec} + (n-1) * 32 \text{ usec}$$

Thus the cost of fetching an additional page is only 32 usec, but underfetching and having to make another request to get a page you missed costs 5000 usec. Based on these arguments, Smith suggested that we re-examine work in virtual memory to see if there are comfortable ways to support distributed virtual memory with anticipation.

In the third talk on RPC in the session, Tommy Joseph (Olivetti), for reasons similar to those of Partridge and Smith, argued that we have to get rid of RPC and give programmers alternative programming paradigms. He sketched out ideas for asynchronous paradigms using causal consistency, in which systems ensure that operations happen in the proper order, without synchronizing through a single system.

Session 6: Hosts and Host Interfaces (Gary Delp, Chair)

Gary Delp (IBM Research) discussed several issues involved in the increase in speed of network attachment to hosts of increasing performance. These issues included:

- Media Access - There are aspects of media access that are best handled by dedicated silicon, but there are also aspects that are best left to a general-purpose processor.

- Compression - Some forms of compression/expansion may belong on the network interface; most will be application-specific.

- Forward Error Correction - The predicted major packet loss mode is packet drops due to internal network congestion, rather than bit errors, so forward error correction internal to a packet may not be useful. On the other hand, the latency cost of not being able to recover from bit errors is very high. Some proposals were discussed which suggest that FEC among packet groups, with dedicated hardware support, is the way to go.

- Encryption/Decryption - This is a computationally intensive task. Most agree that if it is done with all traffic, some form of hardware support is helpful. Where does it fit in the protocol stack?

- Application Memory Mapping - How much of the host memory structure should be exposed to the network interface? Virtual memory and paging complicate this issue considerably.

- Communication with Other Channel Controllers - Opinions were expressed that ranged from absolutely passive network interfaces to interfaces that run major portions of the operating system and bus arbitration codes.

- Blocking/Segmentation - The consensus is that B/S should occur wherever the transport layer is processed.

- Routing - This is related to communications with other controllers. A routing-capable interface can reduce the bus requirements by a factor of two.

- Intelligent participation in the host structure as a gateway, router, or bridge.

- Presentation Layer issues - All of the other overheads can be completely overshadowed by this issue if it is not solved well and integrated into the overall host architecture. This points out the need for some standardization of representation (IEEE floating point, etc.)

Eric Cooper (CMU) summarized some initial experience with Nectar, a high-speed fiber-optic LAN that has been built at Carnegie Mellon. Nectar consists of an arbitrary mesh of crossbar switches connected by means of 100 Mbps fiber-optic links. Hosts are connected to crossbar switches via communication processor boards called CABs. The CAB presents a memory-mapped interface to user processes and off-loads all protocol processing from the host.

Preliminary performance figures show that latency is currently limited by the number of VME operations required by the host-to-CAB shared memory interface in the course of sending and receiving a message. The bottleneck in throughput is the speed of the VME interface: although processes running on the CABs can communicate over Nectar at 70 Mbps, processes on the hosts are limited to approximately 25 Mbps.

Jeff Mogul (DEC Western Research Lab) made these observations: Although off-board protocol processors have been a popular means to connect a CPU to a network, they will be less useful in the future. In the hypothetical workstation of the late 1990s, with a 1000-MIPS CPU and a Gbps LAN, an off-board protocol processor will be of no use. The bottleneck will not be the computation required to implement the protocol, but the cost of moving the packet data into the CPU's cache and the cost of notifying the user process that the data is available. It will take far longer (hundreds of instruction cycles) to perform just the first cache miss (required to get the packet into the cache) than to perform all of the instructions necessary to implement IP and TCP (perhaps a hundred instructions).

A high-speed network interface for a reasonably-priced system must be designed with this cost structure in mind; it should also eliminate as many CPU interrupts as possible, since interrupts are also very expensive. It makes more sense to let a user process busy-wait on a network-interface flag register than to suspend it and then take an interrupt; the normal CPU scheduling mechanism is more efficient than interrupts if the network interactions are rapid.

David Greaves (Olivetti Research Ltd.) briefly described the need for a total functionality interface architecture that would allow the complete elimination of communication interrupts. He described the Cambridge high-speed ring as an ATM cell-like interconnect that currently runs at 500-1000 MBaud, and claims that ATM at that speed is a done deal. Dave Tennenhouse also commented that ATM at high speeds with parallel processors is not the difficult thing that several others have been claiming.

Bob Beach (Ultra Technologies) started his talk with the observation that networking could be really fast if only we could just get rid of the hosts. He then supported his argument with illustrations of 80-MByte/second transfers to frame buffers from Crays that drop to half that speed when the transfer is host-to-host. Using null network layers and proprietary MAC layers, the Ultra Net system can communicate application-to-application with ISO TP4 as the transport layer at impressive rates of speed. The key to high-speed host interconnects has been found to be both large packets and large (on the order of one megabyte) channel transfer requests. Direct DMA interfaces exhibit much smaller transfer latencies.

Derek McAuley (University Cambridge Computer Laboratory) described work of the Fairisle project which is producing an ATM network based on fast packet switches. A RISC processor (12 MIPS) is used in the host interface to do segmentation/reassembly/demultiplexing. Line rates of up to 150 Mbps are possible even with this modest processor. Derek has promised that performance and requirement results from this system will be published in the spring.

Bryan Lyles (XEROX PARC) volunteered to give an abbreviated talk in exchange for discussion rights. He reported that Xerox PARC is interested in ATM technology and wants to install an ATM LAN at the earliest possible opportunity. Uses will include such applications as video where guaranteed quality of service (QOS) is required. ATM technology and the desire for guaranteed QOS places a number of new constraints on the host interface. In particular, they believe that they will be forced towards rate-based congestion control. Because of implementation issues and burst control in the ATM switches, the senders will be forced to do rate based control on a cell-by-cell basis.

Don Tolmie (Los Alamos National Laboratory) described the High-Performance Parallel Interface (HPPI) of ANSI task group X3T9.3. The HPPI is a standardized basic building block for implementing, or connecting to, networks at the Gbps speeds, be they ring, hub, cross-bar, or long-haul based. The HPPI physical layer operates at 800 or 1600 Mbps over 25-meter twisted-pair copper cables in a point-to-point configuration. The HPPI physical layer has almost completed the standards process, and a companion HPPI data framing standard is under way, and a Fiber Channel standard at comparable speeds is also being developed. Major companies have completed, or are working on, HPPI interfaces for supercomputers, high-end workstations, fiber-optic extenders, and networking components.

The discussion at the end of the session covered a range of topics. The appropriateness of outboard protocol processing was questioned. Several people agreed that outboarding on a Cray (or similar cost/performance) machines makes economic sense. Van Jacobson contended that for workstations, a simple memory-mapped network interface that provides packets visible to the host processor may well be the ideal solution.

Bryan Lyles reiterated several of his earlier points, asserting that when we talk about host interfaces and how to build them we should remember that we are really talking about process-to-process communication, not CPU-to-CPU communication. Not all processes run on the central CPU, e.g., graphics processors and multimedia. Outboard protocol processing may be a much better choice for these architectures.

This is especially true when we consider that memory/bus bandwidth is often a bottleneck. When our systems run out of bandwidth, we are forced towards a NUMA model and multiple buses to localize memory traffic.

Because of QOS issues, the receiver must be able to tell the sender how fast it can send. Throwing away cells (packets) will not work because unwanted packets will still clog the receiver’s switch interface, host interface, and requires processing to throw away.

Session 7: Congestion Control (Scott Shenker, Chair)

The congestion control session had six talks. The first two talks were rather general, discussing new approaches and old myths. The other four talks discussed specific results on various aspects of packet (or cell) dropping: how to avoid drops, how to mitigate their impact on certain applications, a calculation of the end-to-end throughput in the presence of drops, and how rate-based flow control can reduce buffer usage. Thumbnail sketches of the talks follow.

In the first of the general talks, Scott Shenker (XEROX PARC) discussed how ideas from economics can be applied to congestion control. Using economics, one can articulate questions about the goals of congestion control, the minimal feedback necessary to achieve those goals, and the incentive structure of congestion control. Raj Jain (DEC) then discussed eight myths related to congestion control in high-speed networks. Among other points, Raj argued that (1) congestion problems will not become less important when memory, processors, and links become very fast and cheap, (2) window flow control is required along with rate flow control, and (3) source-based controls are required along with router-based control.

In the first of the more specific talks, Isidro Castineyra (BBN Communications Corporation) presented a back-of-the-envelope calculation on the effect of cell drops on end-to-end throughput. While at extremely low drop rates the retransmission strategies of go-back-n and selective retransmission produced similar end-to-end throughput, at higher drop rates selective retransmission achieved much higher throughput. Next, Tony DeSimone (AT&T) told us why high-speed networks are not just fast low-speed networks. If the buffer/window ratio is fixed, the drop rate decreases as the network speed increases. Also, data was presented which showed that adaptive rate control can greatly decrease buffer utilization. Jamal Golestani (Bellcore) then presented his work on stop-and-go queueing. This is a simple stalling algorithm implemented at the switches which guarantees no dropped packets and greatly reduces delay jitter. The algorithm requires prior bandwidth reservation and some flow control on sources, and is compatible with basic FIFO queues. In the last talk, Victor Frost (University of Kansas) discussed the impact of different dropping policies on the perceived quality of a voice connection. When the source marks the drop priority of cells and the switch drops low priority cells first, the perceived quality of the connection is much higher than when cells are dropped randomly.

Session 8: Switch Architectures (Dave Sincoskie, Chair)

Dave Mills (University of Delaware) presented work on a project now under way at the University of Delaware to study architectures and protocols for a high-speed network and packet switch capable of operation to the gigabit regime over distances spanning the country. It is intended for applications involving very large, very fast, very bursty traffic typical of supercomputing, remote sensing, and visualizing applications. The network is assumed to be composed of fiber trunks, while the switch architecture is based on a VLSI baseband crossbar design which can be configured for speeds from 25 Mbps to 1 Gbps.

Mills' approach involves an externally switched architecture in which the timing and routing of flows between crossbar switches are determined by sequencing tables and counters in high-speed memory local to each crossbar. The switch program is driven by a reservation-TDMA protocol and distributed scheduling algorithm running in a co-located, general-purpose processor. The end-to-end customers are free to use any protocol or data format consistent with the timing of the network. His primary interest in the initial phases of the project is the study of appropriate reservation and scheduling algorithms. He expect these algorithms to have much in common with the PODA algorithm used in the SATNET and WIDEBAND satellite systems and to the algorithms being considered for the Multiple Satellite System (MSS).

John Robinson (JR, BBN Systems and Technologies) gave a talk called Beyond the Butterfly, which described work on a design for an ATM cell switch, known as MONET. The talk described strategies for buffering at the input and output interfaces to a switch fabric (crossbar or butterfly). The main idea was that cells should be introduced to the switch fabric in random sequence and to random fabric entry ports to avoid persistent traffic patterns having high cell loss in the switch fabric, where losses arise due to contention at output ports or within the switch fabric (in the case of a butterfly). Next, the relationship of this work to an earlier design for a large-scale parallel processor, the Monarch, was described. In closing, JR offered the claim that this class of switch is realizable in current technology (barely) for operation over SONET OC-48 2.4 Gbps links.

Dave Sincoskie (Bellcore) reported on two topics: recent switch construction at Bellcore, and high-speed processing of ATM cells carrying VC or DG information. Recent switch design has resulted in a switch architecture named SUNSHINE, a Batcher-banyan switch which uses recirculation and multiple output banyans to resolve contention and increase throughput. A paper on this switch will be published at ISS '90, and is available upon request from the author. One of the interesting traffic results from simulations of SUNSHINE shows that per-port output queues of up to 1,000 cells (packets) may be necessary for bursty traffic patterns. Also, Bill Marcus (at Bellcore) has recently produced Batcher-banyan (32x32) chips which test up to 170Mb/sec per port.

The second point in this talk was that there is little difference in the switching processing of Virtual Circuit (VC) and Datagram (DG) traffic that which has been previously broken into ATM cells at the network edge. The switch needs to do a header translation operation followed by some queueing (not necessarily FIFO). The header translation of the VC and DG cells differs mainly in the memory organization of the address translation tables (dense vs. sparse).

The discussion after the presentations seemed to wander off the topic of switching, back to some of the source-routing vs. network routing issues discussed earlier in the day.

Session 9: Open Mike Night (Craig Partridge, Chair)

As an experiment, the workshop held an open mike session during the evening of the second day. Participants were invited to speak for up to five minutes on any subject of their choice. Minutes of this session are sketchy because the chair found himself pre-occupied by keeping speakers roughly within their time limits.

Charlie Catlett (NSCA) showed a film of the thunderstorm simulations he discussed earlier.

Dave Cheriton (Stanford) made a controversial suggestion that perhaps one could manage congestion in the network simply by using a steep price curve, in which sending large amounts of data cost exponentially more than sending small amounts of data (thus leading people only to ask for large bandwidth when they needed it, and having them pay so much, that we can afford to give it to them).

Guru Parulkar (Washington University, St. Louis) argued that the recent discussion on appropriateness of existing protocol and need for new protocols (protocol architecture) for gigabit networking lacks the right focus. The emphasis of the discussion should be on what is the right functionality for gigabit speeds, which is simpler per packet processing, combination of rate and window based flow control, smart retransmission strategy, appropriate partitioning of work among host cpu+os, off board cpu, and custom hardware, and others. It is not surprising that the existing protocols can be modified to include this functionality. By the same token, it is not surprising that new protocols can be designed which take advantage of lessons of existing protocols and also include other features necessary for gigabit speeds.

Raj Jain (DEC) suggested we look at new ways to measure protocol performance, suggesting our current metrics are insufficiently informative.

Dan Helman (UCSC) asked the group to consider, more carefully, who exactly the users of the network will be. Large consumers? or many small consumers?

Session 10: Miscellaneous Topics (Bob Braden, Chair)

As its title implies, this session covered a variety of different topics relating to high-speed networking.

Jim Kurose (University of Massachussetts) described his studies of scheduling and discard policies for real-time (constrained delay) traffic. He showed that by enforcing local deadlines at switches along the path, it is possible to significantly reduce overall loss for such traffic. Since his results depend upon the traffic model assumptions, he ended with a plea for work on traffic models, stating that Poisson models can sometimes lead to results that are wrong by many orders of magnitude.

Nachum Shacham (SRI International) discussed the importance of error correction schemes that can recover lost cells, and as an example presented a simple scheme based upon longitudinal parity. He also showed a variant, diagonal parity, which allows a single missing cell to be recreated and its position in the stream determined.

Two talks concerned high-speed LANs. Biswanath Muhkerjee (UC Davis) surveyed the various proposals for fair scheduling on unidirectional bus networks, especially those that are distance insensitive, i.e., that can achieve 100% channel utilization independent of the bus length and data rate. He described in particular his own scheme, which he calls p-i persistant.

Howard Salwen (Proteon), speaking in place of Mehdi Massehi of IBM Zurich who was unable to attend, also discussed high-speed LAN technologies. At 100 Mbps, a token ring has a clear advantage, but at 1 Gbps, the speed of light kills 802.6, for example. He briefly described Massehi’s reservation-based scheme, CRMA (Cyclic-Reservation Multiple-Access).

Finally, Yechiam Yemeni (YY, Columbia University) discussed his work on a protocol silicon compiler. In order to exploit the potential parallelism, he is planning to use one processor per connection.

The session closed with a spirited discussion of about the relative merits of building an experimental network versus simulating it. Proponents of simulation pointed out the high cost of building a prototype and limitation on the solution space imposed by a particular hardware realization. Proponents of building suggested that artificial traffic can never explore the state space of a network as well as real traffic can, and that an experimental prototype is important for validating simulations.

Session 11: Protocol Architectures (Vint Cerf, Chair)

Nick Maxemchuk (AT&T Bell Labs) summarized the distinctions between circuit switching, virtual circuits, and datagrams. Circuits are good for (nearly) constant rate sources. Circuit switching dedicates resources for the entire period of service. You have to set up the resource allocation before using it. In a 1.7 Gbps network, a 3000-mile diameter consumes 10**7 bytes during the circuit set-up round-trip time, and potentially the same for circuit teardown. Some service requirements (file transfer, facsimile transmission) are far smaller than the wasted 2*10**7 bytes these circuit management delays impose. (Of course, these costs are not as dramatic if the allocated bandwidth is less than the maximum possible.)

Virtual circuits allow shared use of bandwidth (multiplexing) when the primary source of traffic is idle (as in Voice Time Assigned Speech Interpolation). The user notifies the network of planned usage.

Datagrams (DG) are appropriate when there is no prior knowledge of use statistics or usage is far less than the capacity wasted during circuit or virtual circuit set-up. One can adaptively route traffic among equivalent resources.

In gigabit ATMs, the high service speed and decreased cell size increases the relative burstiness of service requests. All of these characteristics combine to make DG service very attractive.

Maxemchuk then described a deflection routing notion in which traffic would be broken into units of fixed length and allowed into the network when capacity was available and routed out by any available channel, with preference being given to the channel on the better path. This idea is similar to the hot potato routing of Paul Baran’s 1964 packet switching design. With buffering (one buffer), Maxemchuk achieved a theoretical 90% utilization. Large reassembly buffers provide for better throughput.

Maxemchuk did not have an answer to the question: how do you make sure empty "slots" are available where needed? This is rather like the problem encountered by D. Davies at the UK National Physical Laboratory in his isarythmic network design in which a finite number of crates are available for data transport throughout the network.

Guru Parulkar (Washington University, St. Louis) presented a broad view of an Internet architecture in which some portion of the system would operate at gigabit speeds. In his model, internet, transport, and application protocols would operate end to end. The internet functions would be reflected in gateways and in the host/net interface, as they are in the current Internet. However, the internet would support a new type of service called a congram which aims at combining strengths of both soft connection and datagram.

In this architecture, a variable grade of service would be provided. Users could request congrams (UCON) or the system could set them up internally (Picons) to avoid end-to-end setup latency. The various grades of service could be requested, conceptually, by asserting various required (desired) levels of error control, throughput, delay, interarrival jitter, and so on. Gateways based on ATM switches, for example, would use packet processors at entry/exit to do internet specific per packet processing, which may include fragmentation and reassembly of packets (into and out of ATM cells).

At the transport level, Parulkar argued for protocols which can provide application-oriented flow and error control with simple per packet processing. He also mentioned the notion of a generalized RPC (GRPC) in which code, data, and execution might be variously local or remote from the procedure initiator. GRPC can be implemented using network level virtual storage mechanisms.

The basic premise of Raj Yavatkar's presentation (University of Kentucky) was that processes requiring communication service would specify their needs in terms of peak and average data rate as well as defining burst parameters (frequency and size). Bandwidth for a given flow would be allocated at the effective data rate that is computed on the basis of flow parameters. The effective data rate lies somewhere between the peak and average data rate based on the burst parameters. Statistical multiplexing would take up the gap between peak and effective rate when a sudden burst of traffic arrives. Bounds on packet loss rate can be computed for a given set of flow parameters and corresponding effective data rate.

This presentation led to a discussion about deliberate disciplining of inter-process communication demands to match the requested flow (service) profile. This point was made in response to the observation that we often have little information about program behavior and might have trouble estimating the network service requirements of any particular program.

Architectural Discussion

An attempt was made to conduct a high-level discussion on various architectural questions. The discussion yielded a variety of opinions:

1. The Internet would continue to exist in a form similar to its current incarnation, and gateways would be required,

at least to interface the existing facilities to the high speed packet switching environment.

2. Strong interest was expressed by some participants in access to raw (naked ATM) services. This would permit users to construct their own gigabit nets, at the IP level, at any rate. The extreme view of this was taken by David Cheriton who would prefer to have control over routing decisions and other behavior of the ATM network.

3. The speed of light problem (latency, round-trip delay) is not going to go away and will have serious impact on control of the system. The optimistic view was taken, for example, by Craig Partridge and Van Jacobson, who felt that many of the existing network and communications management mechanisms used in the present Internet protocols would suffice, if suitably implemented, at higher speeds. A less rosy view was taken by David Clark who observed (as did others) that many transactions would be serviced in much less than one round-trip time, so that any end-to-end controls would be largely useless.

4. For applications requiring fixed, periodic service, reservation of resource seemed reasonably attractive to many participants, as long as the service period dominated the set-up time (round-trip delay) by an appreciable margin.

5. There was much discussion throughout the workshop of congestion control and flow control. Although these problems were not new, they took on somewhat newer character in the presence of much higher round-trip delays (measured in bits outstanding). One view is that end-to-end flow control is needed, in any case, to moderate sources sending to limited bandwidth receivers. End-to-end flow control may not, however, be sufficient to protect the interior of the network from congestion problems, so additional, intra-network means are needed to cope with congestion hot spots. Eventually such conditions have to be reflected to the periphery of the network to moderate traffic sources.

6. There was disagreement on the build or simulate question. One view was in favor of building network components so as to collect and understand live application data. Another view held that without some careful simulation, one might have little idea what to build (for example, Sincoskie’s large buffer pool requirement was

not apparent until the system was simulated).

Comments from Workshop Evaluation Forms

At the end of the IRSG workshop, we asked attendees to fill out an evaluation form. Of the fifty-one attendees, twenty-nine (56%) turned in a form.

The evaluation form asked attendees to answer two questions:

#1. Do you feel that having attended this workshop will help you in your work on high speed networks during the next year?

#2. What new ideas, questions, or issues, did you feel were brought up in the workshop?

In this section we discuss the answers we got to both questions.

Question #1

There was a satisfying unanimity of opinion on question #1. Twenty-four attendees answered yes, often strongly (e.g., Absolutely and very much so). Of the remaining five respondents, three said they expected it to have some effect on their research and only two said the workshop would have little or no effect.

Some forms had some additional notes about why the workshop helped them. Several people mentioned that there was considerable benefit to simply meeting and talking with people they hadn't met before. A few other people noted that the workshop had broadened their perspective, or improved their understanding of certain issues. A couple of people noted that they'd heard ideas they thought they could use immediately in their research.

Question #2

Almost everyone listed ideas they'd seen presented at the conference which were new to them.

It is clear that which new ideas were important was a matter of perspective - the workshop membership was chosen to represent a broad spectrum of specialties, and people in different specialities were intrigued by different ideas. However, there were some general themes raised in many questionnaires:

(1) Limitations of our traffic models. This particular subject was mentioned, in some form, on many forms. The attendees

generally felt we didn't understand how network traffic would behave on a gigabit network, and were concerned that people might design (or worse, standardize) network protocols for high speed networks that would prove inadequate when used with real traffic. Questions were raised about closed-loop vs. open-loop traffic models and the effects of varying types of service. This concern led several people to encourage the construction of a high-speed testbed, so we can actually see some real traffic.

(2) Congestion control. Despite the limitations of our traffic models, respondents felt that congestion control at both switching elements and network wide was going to be even more important than today, due to the wider mix of traffic that will appear on gigabit networks. Most forms mentioned at least one of the congestion control talks as a containing a new idea. The talks by Victor Frost, Jamal Golestani and Scott Shenker received the most praise. Some attendees were also interested in methods for keeping the lower-layer switching fabric from getting congested and mentioned the talks by Robinson and Maxemchuk as of interest.

(3) Effects of fixed delay. While the reviews were by no means unanimous, many people had come to the conclusion that the most serious problem in gigabit networking was not bandwidth, but delay. The workshop looked at this issue in several guises, and most people listed at least one aspect of fixed delay as a challenging new problem. Questions that people mentioned include:

- How to avoid a one round-trip set up delay, for less than one round-trip time's worth of data?

- How to recover from error without retransmission (and thus additional network delays)? Several people were intrigued by Nachum Shacham's work on error detection and recovery.

- Should we use window flow-control or rate-based flow control when delays were long?

- Can we modify the idea of remote procedure calls to deal with the fact that delays are relatively long?

A couple of attendees noted that while some of these problems looked similar to those of today, the subtle differences caused by operating a network at gigabit speeds led them to believe the actual approaches to solving these problems would have to be very different from those of today.

Security Considerations

Security issues are not discussed in this memo.

Author's Address

Craig Partridge
Bolt Beranek and Newman Inc.
50 Moulton Street
Cambridge, MA 02138

Phone: (617) 873-2459

EMail: craig@BBN.COM