Network Issues for Sequoia 2000

Domenico Ferrari

Sequoia 2000 Technical Report 91/6
Computer Science Division
Dept. of Electrical Engin. and Computer Science
University of California, Berkeley
Berkeley, CA 94720

Joseph Pasquale, George C. Polyzos

Computer Systems Laboratory
Dept. of Computer Science and Engin.
University of California, San Diego
La Jolla, CA 92093-0114

Abstract

The goals of the Sequoia 2000 Network are to provide high throughput for the massive observation input data and image output data characterizing Global Change applications, as well as real-time services for animations and collaboration tools such as video conferencing. The first phase of the network will be based on a T3 (45 Mb/s) backbone and FDDI for local distribution. The research issues we are focusing on include protocols that provide deterministic and statistical performance guarantees and take advantage of hierarchical coding of information, and the design of I/O system software that integrates process and device communication software with network protocol software.

Introduction

Past typical wide-area computer networks bandwidths have ranged from 56 Kb/s of the ARPAnet to 1.5 Mb/s (T1) of the NSFNET backbone, which has recently been upgraded to 45 Mb/s (T3). These networks have focused on providing packet-switched services for applications such as electronic mail and file transfer, which are non-real-time and generate messages ranging in size from a few hundred bytes to a few hundreds of thousands of bytes. (Another popular application supported by these networks is remote login, which requires good response, but does not have rigorous real-time constraints and requires minimal bandwidth.)

With the introduction of optical fiber and fast switching technologies into current and future networks, applications that have real-time and high-bandwidth constraints for communication can be supported. The applications of Sequoia 2000 have these demanding communication requirements. Scientific applications for Global Change such as simulation of atmospheric or oceanic processes need processing of massive amounts of data retrieved from a remote database, with the results being visualized at a scientist's workstation as a real-time animation. Furthermore, tools for collaboration between the Sequoia 2000 investigators, such as video conferencing, also require high-bandwidth bounded low-delay communication. We are designing a network for Sequoia 2000 to support these types of applications.

Communication Requirements

Before describing the goals of the Sequoia 2000 Network (SN), it is useful to understand the predominant applications that are to be supported and the type of communications they require. There are two primary types of applications used by Global Change scientists:

1. simulation models of physical processes that attempt to predict what will happen as time progresses;
2. data analyses of observed physical processes, gathered by satellites and remote sensing devices.

The simulation models do generally require a small amount of input, but run for a very long time and generate a large amount of output. The desired output is in the form of a timed sequence of images illustrating the behavior of physical processes as it changes over time. A very simple example is a map of the United States with overlayed colored regions that move over time, where color may indicate pressure or temperature. The data analyses applications have similar characteristics, except that they consume a large amount of input "observation data," which are real measurements of physical processes. Simulation model outputs are often compared with analyses produced from observed data for validation. A timed sequence of differential images, showing the differences between the simulation and the measured system, is very useful in those cases. However, direct frame-by-frame visual comparison is also of interest. The following numbers are typical of Global Change applications. For input, a file containing observation data can easily exceed 1 GB. For output, a single image contains 1 MB, assuming 1K x 1K pixels and 1 byte/pixel for color. Displayed at 20 frames per second, a 1–minute animation generates 1.2 GB (and a rate of 160 Mb/s).

For a concrete example consider the *Analysis of High-Resolution Visible Data for Photosynthetically Available Radiation (PAR).* PAR can be obtained by a method originally developed by Gautier [Gaut80]. This application uses various resolutions of geostationary and AVHRR data, up to half hour and 1 km. The latest available version of the code running on a MicroVAX II under VMS requires a couple of hours to produce results for an analysis period of one day and uses input data up to approximately 1 GB. Interestingly, the data come from many different sources, and, to a great extent, computation can proceed independently on each data set. The main result of the computation is a daily parameter field, i.e., essentially an image. Even though computations are usually run for horizons of many months or years, it is desirable visually to inspect the progress of the computation at various checkpoints to apply consistency checks. The results of this computation are of interest to other researchers and are primary candidates for inclusion in the database. Observation of a sequence of "frames" from this database can reveal interesting dynamic effects. For this reason, the presentation must be smooth and at a speed of tens of frames per second. Furthermore, it is desirable to convert this application to an interactive one through the use of more powerful processors and distributed computing.

Sequoia 2000 is also an experiment in collaboration. The project includes computer scientists of different areas, and Global Change scientists of different areas, distributed over the various campuses of the University of California. For the project to be successful, a high degree of collaboration must exist. This can be enhanced by the availability of tools which promote collaboration, such as video conferencing, distributed window systems, and shared blackboards. These applications have strict delay requirements, and some require high-bandwidth communication. For example, interactive audio and video require end-to-end delays of no more than tens or a few hundreds of milliseconds. For voice quality audio the standard 64 Kb/s rate can easily be provided, given the other requirements of the applications described above. For NTSC quality video, data rates around 30 Mb/s are needed, even after considerable compression. We believe, however, that even more highly compressed video (of "videoconference quality") would be a big asset for projects that involve remote collaboration, such as this one.

Distributed Computing Scenarios

The typical way scientists obtain observed data or output generated by their applications is by sending and receiving magnetic tapes, or by file transfer over the Internet if the files are small. However, a 1 GB file of observed data would take 5 hours or more to obtain, assuming an average throughput of 60 KB/s, which is typical of today’s Internet. Clearly, there is much room for improvement here. A T3 network which provides most of the T3 potential bandwidth (45 Mb/s) would allow delivery of the same file in under 3 minutes.

However, Sequoia will also provide an environment for distributed computing. For instance, computational processes could be placed near the BIGFOOT database [Ston92], which would store the observational data, as well as the output data. Scientists would interact remotely with BIGFOOT using their workstations, which they would use for visualization of output. Output can either be completely downloaded to the workstation, or, more likely, sent in chunks comprising single images (or hierarchically coded sub-images, as described below).

Sending images is more desirable for a number of reasons. First, the complete output file can be extremely large, possibly requiring substantial time (even with a fast network). Second, the scientist will generally not require using the entire file all at once. In fact, a typical scenario is one where the scientist wishes to interactively explore a data set, by fast-forwarding or rewinding an animation, or searching the space defined by an image by zooming and panning. Consequently, what is needed is a close coupling between BIGFOOT and the workstation used for interactive control and visualization. A network which provides high throughput and real-time response is needed.

Global Change researchers are already planning to exploit modes of distributed computation, and have identified the need for high-speed networking. For example, Mechoso *et al.* [Mech91] consider the distribution of a coupled atmosphere-ocean general circulation model across high-speed networks. The basic idea is to use machines with the most suitable architecture for each component of the distributed computation. These specialized machines will probably not be available in the same location; in addition, input and output data might all be distributed in other locations. According to those authors' estimates, the amount of data to be exchanged at each step (i.e., during a simulated hour) is of the order of 40 Mb, while the current execution time for each step is around 10 s of CPU time on one processor of a CRAY Y-MP.

The challenge in designing the SN is to support high-speed transfers of large files, while providing real-time transmission of both images generated by Global Change applications as well as audio and video for real-time collaboration applications.

Sites and Topology

All Sequoia 2000 sites are currently connected to the Internet, mainly through two regional networks, BARRnet in Northern California and CERFnet in Southern California. An NSFNET backbone link between Stanford University and the San Diego Supercomputer Center (SDSC) interconnects the two networks and has recently been upgraded to T3 speed, but most of BARRnet and CERFnet are still based on T1 (1.5 Mb/s) links. The application-to-application throughputs we observed through the Internet in a series of recently performed experiments were limited to 80-90 KB/s under the best circumstances, and more typical values were of the order of 60 KB/s. As described earlier, these throughputs cannot effectively support the applications under consideration.

An important characteristic of the Sequoia 2000 project is the active involvement of the users of the technology under development all the way from conception to actual deployment. This involvement will mainly consist of specifying requirements, providing application descriptions and benchmarks, and evaluating alternative approaches and systems. We would like to take as much advantage of this opportunity as possible. For this reason, we are currently deploying an experimental packet switched network in order to get experience from the applications in question, and to allow experimentation with I/O and network architectures and protocols.

The first phase of the SN will be based on a backbone network of T3 leased communication lines connecting the primary sites. These include the University of California campuses at Berkeley, Los Angeles, Santa Barbara, and San Diego (including the Scripps Institution of Oceanography and the San Diego Supercomputer Center), and the California State Department of Water Resources. Each site will have one or more FDDI rings connecting the various campus labs. The routers connecting the LANs to the backbone will be general purpose DEC machines equipped with T3 communication boards and FDDI interfaces. This will allow us easily to implement and evaluate new network protocols and to experiment with fast and efficient I/O schemes.

We are also investigating the possibility of deploying a gigabit network at a later time. This is of particular interest since two of the Sequoia sites are already involved in gigabit testbed projects (UC Berkeley in Blanca and SDSC in Casa), and the applications under consideration seem really to warrant gigabit/second rates. We believe that the characterization of the application requirements and the identification of scientific applications for which gigabit networks are going to be enabling technologies is important in itself. Therefore, we plan to undertake an extensive application requirements characterization study in collaboration with Global Change researchers.

Research Issues

As described above, a common work scenario for Global Change scientists will be the visualization at their workstation of time-sequenced images accessed from a large object base over a high-bandwidth wide-area network. The data may be produced in real time, or they may not (e.g., because of the computational effort required). The visualization will be interactive with users from remote workstations asking for playback, fast-forward, rotation, and so on; this should be possible without necessarily bringing the entire data set into the workstation at the beginning.

This interactivity and the temporal nature of the data's presentation requires a predictable and guaranteed level of performance from the network. Although image sequences require high-bandwidth and low-delay guarantees, these guarantees often do not have to be deterministic (i.e., absolute), but may be expressed in statistical terms. The protocols to be executed on the host workstations, the gateways, and the switches (or the switch controllers) will have to include provisions for real-time channel establishment/disestablishment [Ferr90a] [Ferr91], so that guarantees about the network's performance (throughput, delay, and delay jitter) can be offered to the users who need them [Ferr90b]. A related issue is the specification of the quality of network service needed by the user. Such a specification must be powerful enough to describe the required guarantees, and yet must be realizable by mechanisms that already exist, or that can be built into the networks of interest.

Mechanisms which support the guaranteed services offered by the network must be integrated with the operating system, particularly the I/O system software, which controls the movement of data between arbitrary I/O devices, such as the network interface, frame buffer, and other real-time devices [Pasq90, Pasq91]. The network software and the I/O system software must work in a coordinated fashion so that bottlenecks, such as those due to memory copying or crossing of protection boundaries, are avoided. The I/O system software, is one of the least understood aspects of operating system design, especially regarding soft real-time I/O. We are exploring the relationship between I/O system software and network protocol software, and how various degrees of design integration affect performance. In particular, we have been investigating the construction of fast in-kernel datapaths between the network and I/O source/sink devices for carrying messages which are to be delivered at a known rate. Since processing modules (e.g., compression/decompression, network protocols) may be composed along these datapaths, a number of problems must be solved, such as how systematically to avoid copying processed messages between modules, or between kernel and user address spaces.

The database server, the network, or even the workstation’s operating system, can take advantage of the statistical nature of guarantees by conveniently dropping packets when necessary to control congestion and smooth network traffic. This is particularly relevant when one is fast-forwarding through a sequence of images; supporting full resolution might not be possible, and users might be willing to accept a lower-resolution picture in return for faster movement. One approach to this problem is hierarchical coding [Karl89], whereby a unit of information such as an image is decomposed into a set of ordered sub-images. A selected subset of these may be re-composed to obtain various levels of resolution of the original image. This gives the receiver the flexibility of making the best use of received sub-images that must be output by some deadline, and gives the network the flexibility of dropping packets containing the least important sub-images when packets must be dropped.

For example, if hierarchical coding and guaranteed performance channels were available, one could specify that the most important (low-resolution) part of an image (or signal) be transported through a guaranteed channel, while the remaining part of the signal could be transported through a lower-quality channel. Such a scheme would make much better use of available resources for the applications considered here than traditional transport mechanisms. One research issue is how to forward hierarchically coded packets in a way that provides the network with the maximum flexibility in congestion control, and how to compose them in time at the receiver so that integrity and continuity of presentation are preserved. In particular, the layers at which multiplexing and demultiplexing will be performed should be carefully designed to take full advantage of hierarchical coding.

Of course, bandwidth (as well as storage space) requirements can be reduced by image compression. Since we are really designing an integrated end-to-end system, the issue of compression is central and complex. In the simplest case, compression will be applied at the periphery of the system, just after the generation of data (e.g., from a simulation), and decompression at the last possible moment, just before presentation. However, this scheme might present difficulties in accessing the data in any way that is different from the original storage scheme. Therefore, intermediate formats might be necessary, particularly if we expect to search the data by content, as some applications suggest. In addition, the compression scheme, or some of its characteristics, will depend on the chosen mode of computation. For example, with a model of local computation and visualization, lossy compression will probably be totally unacceptable. On the other hand, if only the visualization is local, then some loss might be tolerable (and preferable to other forms of system degradation). Therefore, careful design of the overall compression strategy, taking into consideration hierarchical coding and all the components of the system — DBMS, storage hardware, network, and presentation workstation — is required.

Finally, in order to support the group communication required by multi-person collaboration tools such as video conferencing and shared windows, we are investigating multicast routing algorithms which seek to minimize cost (e.g., link bandwidth use), while simultaneously bounding delay between each source and its destinations [Komp92]. This is possible by delivering data using a tree out of the source and avoiding data duplication as much as possible. Finding the minimum cost tree is an intractable problem (known as the Steiner tree problem). However, in the context of interactive audio and video communication the additional end-to-end delay constraints further complicate the problem. We have developed two heuristics, for both source-based routing and distributed implementations, that provide near optimal configurations for typical communication networks. We have also started investigating the conversion of our real-time communication scheme [Ferr90a] to one based on the simplex multicast (instead of unicast) real-time channel as the fundamental abstraction to be implemented by the service.

Current and Future Work

Given the 100 Mb/s data rate of FDDI LANs and their support for "synchronous" traffic with bounded delay, FDDI seems a natural choice for local distribution. However, even though FDDI networks provide the necessary guarantees at the medium access control layer (if they offer "synchronous service"), there are no protocols at the network and transport layers to make effective use of this underlying capability. Moreover, there are no existing protocols that deal with the synthesis of end-to-end guaranteed performance channels from a set of such channels and networks in tandem. A network layer protocol (RTIP) and some ideas that address this problem of providing end-to-end performance guarantees have been described in [Lowe91] following the methodology for real-time virtual circuit establishment provided in [Ferr90a]. An important feature of RTIP is that it is an extension of IP (the Internet network layer protocol), and interoperates with it. We are in the process of implementing RTIP and plan to experimentally evaluate its performance.

We are also implementing prototype versions of two real-time transport protocols, RMTP [Verm91] and CMTP [Wolf91], which allow applications to specify quality of service parameters. While RMTP is message-oriented, and offers a traditional send/receive interface to its clients, CMTP is oriented towards continuous-media applications, and allows senders to continue "producing" information that is continuously "consumed" by the corresponding receivers. Notice that, for real-time traffic, a transport layer relying on RTIP has little functionality to add, and thus can be made very light-weight. Therefore, both RMTP and CMTP are fairly simple protocols. This work is complemented by I/O system software restructuring, specifically to include in-kernel data paths as discussed above. This will allow processing modules such as the network protocols to be uniformly incorporated in the I/O system software.

As was mentioned above, Global Change scientists will occasionally need to visualize more than one sequence of images simultaneously. The system we are building should therefore make it possible to synchronize multiple continuous-media streams. There are several useful types of synchronization besides the obvious one (according to which the streams start being presented at the right times and proceed at identical rates); for instance, it may be desirable to view the streams at different rates, to stop or rewind or step frame-by-frame through some of them while others continue to be presented at their previous rates, to impose precedence constraints on the starting times, to control the presentation of some streams in an automatic event-driven way, and so on. The Sequoia Network should cooperate with the local operating systems in the offering of a synchronization service. We are studying the possibility and convenience of basing such a service on our distributed scheme for delay jitter control, which operates at the network layer [Ferr91].

We also plan to investigate alternative approaches to providing the required network performance based on more subtle forms of resource reservations (e.g., by introducing traffic priorities in conjunction with input rate enforcement mechanisms such as the "leaky bucket"). This technique addresses the question of the delay introduced by the virtual-circuit set-up, but is weaker in terms of expressed performance guarantees. Our plan is to implement PIP, a network layer protocol based on this technique, as an alternative to RTIP and evaluate it against a networking benchmark and RTIP.

In addition, we will develop software to perform hierarchical coding, and will incorporate into the networking code the functionality to deal with the additional information available on the packets. For example, with PIP, this would simply mean different priority classifications for the various coding components; non-essential components could then be dropped if required, exactly as current IP specification allows. This will enable us to demonstrate fast-forwarding capability through continuous image sequences.

In parallel with the effort for real-time protocol development we are investigating the traffic matrix requirements of the Sequoia sites in order to design the topology of the T3 network. This is required at present since we expect that applications will require a substantial portion of the full T3 bandwidth.

Conclusion

The goals of the Sequoia Network are to provide high throughput for the massive observation input data and image output data characterizing Global Change applications, as well as real-time services for animations and collaboration tools such as video conferencing.

REFERENCES

[Ferr90a] D. Ferrari and D. Verma, "A Scheme for Real-Time Channel Establishment in Wide-Area Networks," *IEEE Journal of Selected Areas in Communications,* vol. 8, pp. 368-379 (1990).

[Ferr90b] D. Ferrari, "Client Requirements for Real-Time Communication Services," *IEEE Communications Magazine*, vol. 28, pp. 65-72 (1990).

[Ferr91] D. Ferrari, "Design and Applications of a Delay Jitter Control Scheme for Packet-Switching Internetworks," *Proc. Second International Workshop on Network and Operating System Support for Digital Audio and Video,* Heidelberg, Germany (1991).

[Gaut80] C. Gautier, G. Diak, and S. Masse, ‘‘A Simple Physical Model to Estimate Incident Solar Radiation at the Surface from GOES Satellite Sata,’’ *Journal of Applied Meteorology,* vol. 19, pp. 1005-1012 (1980).

[Karl89] G. Karlson and M. Vetterli, ‘‘Packet Video and its Integration into the Network Architecture,’’ *IEEE Journal on Selected Areas in Communications,* Vol. 8, No. 3, pp. 380-390 (April 1990).

[Komp92] V. Kompella, J. Pasquale, and G. Polyzos, "Multicast Routing for Multimedia Applications," Proc. IEEE INFOCOM ’92, Florence, Italy (1992).

[Lowe91] C. Lowery, "Protocols for Providing Performance Guarantees in a Packet Switching Internet," Technical Report TR-91-002, International Computer Science Institute, Berkeley, CA, 1991.

[Mech91] C. R. Mechoso, C.-C. Ma, J. D. Spahr, and R. W. Moore, "Distribution of a Coupled Atmosphere-Ocean General Circulation Model Across High-Speed Networks," Proc. 4th International Symposium on Computational Fluid Dynamics, 1991.

[Pasq90] J. Pasquale and G. Polyzos, "System Support for Multimedia Applications," Proc. First International Workshop on Digital Audio and Video, International Computer Science Institute, Berkeley, CA, 1990.

[Pasq91] J. Pasquale, G. Polyzos, E. Anderson, K. Fall, J. Kay, V. Kompella, S. McMullan, and D. Ranganathan, "Network and Operating System Support for Multimedia Applications," Technical Report CS 91-186, University of California, San Diego (1991).

[Ston92] M. Stonebraker, "An Overview of the Sequoia 2000 Project" *Proc. CompCon ’92,* San Francisco, CA (1992).

[Verm91] D. C. Verma and H. Zhang, "Design Document for RTIP/RMTP," unpublished report, May 1991.

[Wolf91] B. Wolfinger and M. Moran, "A Continuous Media Data Transport Service for Real-Time Communication in High Speed Networks," *Proc. Second International Workshop on Network and Operating System Support for Digital Audio and Video,* Heidelberg, Germany (1991).