LBL-32936
UC-405

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

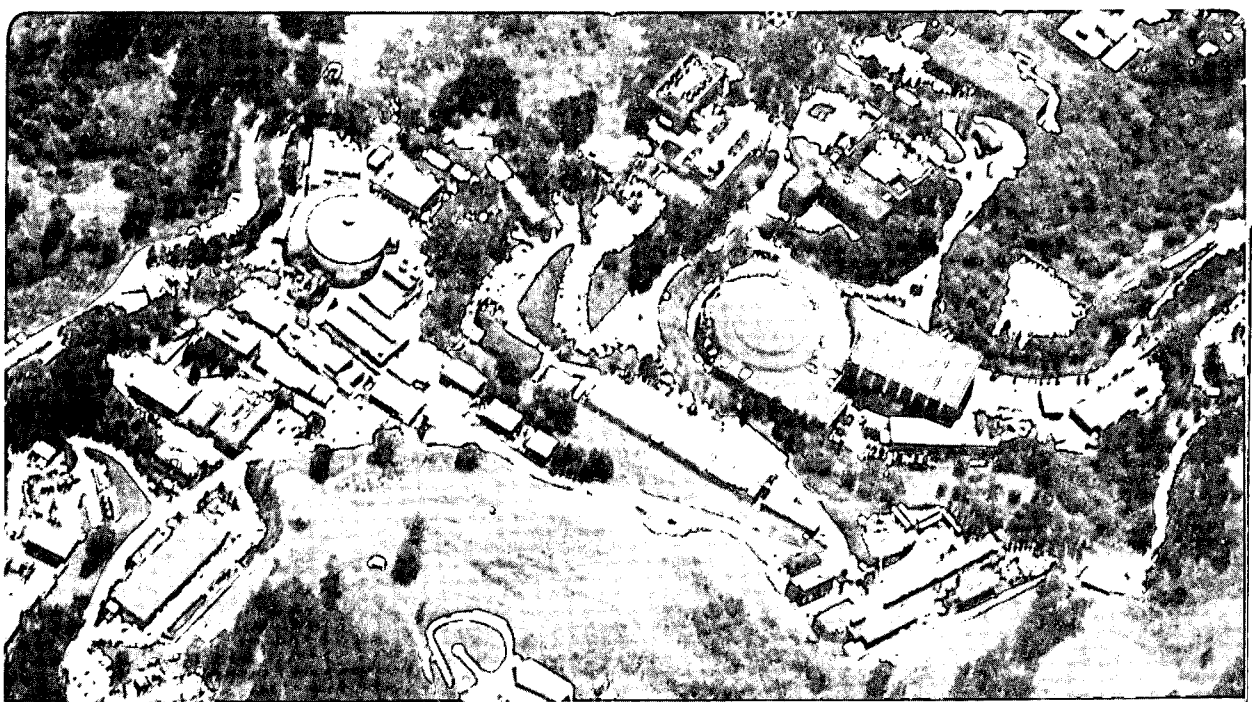
Information and Computing Sciences Division

Presented at the Computing in High Energy Physics 1992
Conference, Annecy, France, September 21–25, 1992,
and to be published in the Proceedings

High-Performance Computing and Distributed Systems

S.C. Loken, W. Greiman, V.L. Jacobson, W.E. Johnston,
D.W. Robertson, and B.L. Tierney

September 1992



LBL-32936
Copy 2
| LOAN COPY |
|Circulates |
|for 4 weeks| Bldg. 50 Library.

Prepared for the U.S. Department of Energy under Contract Number DE-AC03-76SF00098

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. Neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California and shall not be used for advertising or product endorsement purposes.

Lawrence Berkeley Laboratory is an equal opportunity employer.

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

HIGH-PERFORMANCE COMPUTING AND DISTRIBUTED SYSTEMS *

Stewart C. Loken, William Greiman, Van L. Jacobson,
William E. Johnston, David W. Robertson, and Brian L. Tierney
Lawrence Berkeley Laboratory
University of California
Berkeley, CA 94720 USA

Abstract

We present a scenario for a fully distributed computing environment in which computing, storage, and I/O elements are configured on demand into "virtual systems" that are optimal for the solution of a particular problem. We also describe present two pilot projects that illustrate some of the elements and issues of this scenario. The goal of this work is to make the most powerful computing systems those that are logically assembled from network based components, and to make those systems available independent of the geographic location of the constituent elements.

INTRODUCTION

Advances in software paradigms, computing systems, and communications bandwidth over the next few years will help enable an information analysis environment in which scientists have uniform and unimpeded access to computing and data resources regardless of their geographic location. This environment will provide a "just in time" approach to assembling the resources needed to solve specific instances of problems in computational simulation, data acquisition, data analysis, and archiving. It will allow us to design optimal architectures for the solution of specific problems, and then, by using network based resources, to logically assemble and use the required elements only for the time during which they are needed.

*This work is supported by the Director, Office of Energy Research, Office of the Scientific Computing Staff, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

These resources will consist of (1) computing elements (workstations, parallel and vector processors, and specialized processors for encryption, compression, and graphics rendering), (2) data handling elements (large, high speed data "buffers", and distributed mass storage systems), (3) graphics/image display user front end systems, and (4) the software systems to easily interconnect these elements. Not only will this allow powerful capabilities to be brought to bear on large problems, it will also allow access to these capabilities by a much wider community of people than is presently possible. This environment will be enabled through software and hardware architecture advances expected over the next several years, including: an order of magnitude increase in workstation I/O and memory bandwidth; the routine incorporation of coprocessors for special tasks (e.g. video compression); the emerging collaboration between the computing and telecommunications industries for high bandwidth networking; hard-

ware and software improvements permitting multiple heterogeneous computing systems to be easily configured into cooperating elements that form virtual systems; easy access to massive unique data archives enabled through advances in data management and mass storage systems, and; user interface paradigms evolved to allow non-computer specialists to easily assemble the above elements into effective tools to attack scientific problems.

In the following sections, we describe two pilot projects that illustrate some of the elements and issues of this new computing model.

RESEARCH IMAGING

Configurable systems are essential to many scientific endeavors. For example, the research imaging environment is characterized by three elements that are typically geographically dispersed: the imaging device, and its associated control system; computational and data storage elements necessary for processing speed, large memory, high speed data buffers, etc., to capture and interpret the GBytes of data from the imaging device; local workstations for user control of the operation of the imaging device, and the display of the resulting images and visualizations. Advanced scientific imaging will frequently involve separation of these components by their very nature (large, expensive, immobile, etc.).

Another problem in research imaging is that the algorithms for analysis, and therefore the optimal computing environment needed to implement them, may not be well understood. A configurable computing environment allows rapid and economical changes in systems during the learning phase.

Research imaging may seem, at first, rather different from typical problems in High Energy Physics. We believe, however, that the approach described here, with different computing architectures applied to different aspects of a complex calculation, may be useful to a wide range of problems.

A Case Study

The Lawrence Berkeley Laboratory (LBL) Information and Computing Sciences and Research Medicine Divisions have collaborated with the Pittsburgh Supercomputer Center (PSC) to demonstrate the possibilities for such a distributed environment. The prototype application is the interactive visualization of large 3D scalar fields (voxel data sets) by using a combination of heterogeneous supercomputers and low cost workstations for display and control.

This application is designed to test the limits of the recently improved network bandwidth and interprocess communications, and to identify bottlenecks that remain in the way of achieving real-time distributed visualization of large 3D data sets.

The computational part of the application is partitioned into two pieces, one optimal for a massively parallel architecture, and one optimal for a vector processor. The first part is run on a Thinking Machines CM-2, and the second on a Cray Y-MP. These systems are located at PSC, and communicate with each other over a HIPPI, 800 Mbits/sec communications channel, while the remote workstations are connected to PSC via the usual variety of local, regional, and wide area networks (WAN) (e.g. NSFNet and DoE's ESNet).

The Application

The goal is the interactive display of large 3D scalar fields (e.g. a high-resolution MRI data set of the human brain). The data set used for the experiment is 256 x 256 x 128 x 1 byte voxels, or 8.4 MBytes of data. "Interactive" here is taken to mean the ability to generate and display images at a rate of at least 5 frames per second as a result of changing

the isosurface, region of interest, or viewing parameters for the resulting geometry.

Application Architecture

The application consists of three intercommunicating processes. One process runs on a local workstation, and controls the other two processes via an X-window interface. The CM-2 process reads the voxel data and locates isosurfaces using the dividing cubes algorithm [2]. The resulting surface data is sent across a HIPPI channel to a process running on a Cray Y-MP, which does the 3D graphics rendering needed to convert this data into an image. The image is then sent across the network to the local workstation for display as part of the user interface.

Network Issues

There are several network issues that limit end-to-end performance in distributed systems. One of these issues is described here, and several others are discussed in [4]. TCP is used in all of our experiments because it is by far the best developed transport protocol. However, with traditional TCP implementations, there is a problem in that as network speeds increase, throughput becomes limited both by the speed-of-light propagation time between the communicating computers, and by peculiarities in the TCP implementation. The standard version of TCP/IP can send at most 64KB of data per round- trip-time and, in practice, the sustained throughput was at most half this theoretical maximum because of the packet acknowledgement scheme. One of us (VJ) has developed TCP/IP extensions for high performance [1], wide area transport. Working closely with Cray Research and Sun Microsystems these extensions were implemented and used for this experiment. These extensions remove the 64KB per round-trip-time limit and allow TCP/IP to run at the full speed of the underlying network, independent of the end-to-end propagation time. These modifications are essential to achieving fast image display over a wide area network.

Analysis

In this experiment about 500,000 3D points are typically generated to describe one isosurface. This process takes approximately 0.8 seconds using 16,384 processors on the CM-2 (time for surface generation plus time to gather results). It takes 0.06 seconds for the data conversion, 0.1 seconds for the HIPPI transfer, and 0.3 seconds for one Cray Y-MP processor to render the image. Over a 45 Mbits/sec cross-country network, it takes about 0.1 seconds to transfer the resulting 320x320x1Byte (100 KByte) image to the local workstation. Therefore the total time is around 1.3 seconds.

Changing only the viewpoint on the geometric model representing the isosurface in the scalar field is faster because the geometry does not need to be recomputed. The total time in this case is around 0.5 seconds. However, the application can be run in "movie" mode, where images are generated for a set of incremental rotations. In this case the CM-2 and Y-MP are working in parallel, and can generate images of the rotating surface at a rate of about 3 frames per second.

The speed of rotation can be further increased by using a less general hidden-surface removal algorithm (see [4] for more information). Rendering in this case takes 0.03 seconds. Thus in movie mode, network bandwidth becomes the limiting factor, and the maximum speed of rotation is ten frames per second.

We are working on more general methods of distributing the application to ease its usage in a truly heterogeneous "on demand" environment. This work will add workstation

clusters to the collection of computing elements that can be configured into virtual systems. Whether traditional supercomputers or workstation clusters are used will be transparent to the user. Our present work focuses on PVM (Parallel Virtual Machine) [5] as a model of handling interactions among heterogeneous supercomputing resources.

GIGABIT DATA ACQUISITION

The high trigger rate and large event size in future high-energy physics experiments will require new techniques for event readout ("event building"). A high degree of parallelism will be required to achieve this level of performance.

The event builder has become the bottleneck in current data acquisition systems. The performance of current event builders is limited by the interconnection network that is used to multiplex data between sources and destinations. The interconnects used in current event builders, shared buses and multiport memory, can not provide the performance required for SSC and LHC.

During the last few years standards, for networks with several orders of magnitude greater performance than today's Ethernet have been developed. Examples include Asynchronous Transfer MODE (ATM), Fibre Channel, and Synchronous Optical NETwork (SONET). All of these standards specify data links with a bandwidth of at least 1 Gigabit/sec. These standards also specify that the network fabric must be able to provide full bandwidth for simultaneous communication between all possible pairs of nodes.

A common architecture has evolved for the primary flow of data from the detector to processing and recording elements in high performance data acquisition systems. At the highest level of abstraction, almost all proposed high performance data acquisition systems have a classic data flow architecture. A scalable, fully parallel interconnection network is one of the key components required to implement thislternative to the custom interconnects that have been used in high performance data acquisition systems. They have the advantages of supporting standard protocols, being supplied by multiple vendors aut additional engineering cost.

SDC EVENT BUILDER

SDC requires an event builder that accepts data from about 400 readout sources and distributes whole events toors in a level 3 farm. The initial performance requirement is 1000 events per second for one megabyte events. The SDC data must be based on a scalable architecture that will allow a factor of ten increase in performance.

The architecture of a scalable parallel event builder suitable for SDC has three stages, subevent builders, data switch and full event builders. All data paths are assumed to be 1 Gbit/sec Fibre Channel links with a capacity of 100 MB/sec of user data. Ten processors are required in each stage to achieve the one GB/sec required by SDC. Each stage can be scaled by increasing the number of node

The first stage consists of processors that receive data from a subset of readout sources over Fibre Channel 1 subevents. This stage is required to reduce the transaction rate and/or memory requirements in the event builder stage. For smaller detectors such as STAR at RHIC, subevent builders are not needed as readout sources can go directly to the switch.

The second stage is a Fibre Channel switch that distributes subevents to appropriate event builder processors. Current Fibre Channel switches, such as supplied by Ancor Communications are designed to support over 2000 ports at full bandwidth. This exceeds the SDC scalability requirement by a factor of ten.

GIGABIT TESTBED

A prototype test bed has been constructed. This prototype is based on 16 channel Fibre Channel data switch modd processors and communication controllers. This test bed is being used to determine the performance a Fiber Channel based event builder. Performance tests include switch throughput, switch overhead, VME controller performance, error rates and bus usage. The results of tests performance of a full scale event builder.

CONCLUSIONS

We have presented a collection of technologies that, taken together, will provide the possibility for: (1) routinely partitioning problems between heterogeneous supercomputers; (2) doing remote siting of data intensive scientific experiments; (3) providing access to capabilities that could previously only be obtained at a small number of sites due to the size, cost, or experimental nature of the implementation, and; (4) providing new capability enabled by the nature of the networks themselves.

The imaging application demonstrates that wide-area networks are not necessarily the bottleneck to widely distributed imaging applications. Widely deployed Gbits/sec wide-area networks and the associated interconnecting hardware and software promise to alter the way that many large scale problems are approached. These networks will allow the creation of "network" or "virtual" supercomputers– computing systems comprised of geographically distributed components communicating with each other at high speeds, and configured on demand into virtual systems that exist only as long as necessary to solve a particular problem, or until a better combination of elements to solve the problem becomes apparent, at which point the virtual system is reconfigured.

ACKNOWLEDGEMENTS

The authors thank Wendy Huntoon, Jamshid Mahdavi, and Ralph Roskies of PSC for their collaboration with the case study; and Peter Schroder, formerly of Thinking Machines Corporation, Carl Crawford of General Electric Company, and Mark Roos of the LBL Research Medicine Division, for their support and help with this project.

REFERENCES

1. V. Jacobson, R.T. Braden, D.A. Borman, " TCP Extensions for High Performance," Internet Requests for Comment (RFC) 1323, DDN Network Information Center, Menlo Park, CA., May, 1992.
2. H.E. Cline, W.E. Lorensen, S. Ludke, C.R. Crawford, and B.C. Teeter, "Two Algorithms for the Three-Dimensional Reconstruction of Tomograms," Medical Physics, May 1988.
3. M. Schneider, "Pittsburgh's Not-So-Odd Couple," in Supercomputing Review, August, 1991.
4. Johnston, W., V. Jacobson, D. Robertson, B. Tierney, S. Loken, "High Performance Computing, High Speed Networks, and Configurable Computing Environments", LBL Report, LBL-32161, also in "High Performance Computing in Biomedical Research", CRC Press, November, 1992.
5. A. Beguelin, J.J. Dongarra, G.A. Geist, R. Manchek, and V.S. Sunderam, "Graphical Development Tools for Network-Based Concurrent Supercomputing," Supercomputing '91 Conference Proceedings, pp. 435-444.

LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
TECHNICAL INFORMATION DEPARTMENT
BERKELEY, CALIFORNIA 94720