Remote Experimental Site
Concept Development

Thomas Casper
William Meyer
David Butner

January 20, 1995

This is an informal report intended primarily for internal or limited external distribution. The opinions and conclusions stated are those of the author and may or may not be those of the Laboratory.

Work performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, make 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 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. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

LDRD Final Report

January 20, 1995

T. A. Casper

Remote Experimental Site Concept Development

Principal Investigator: Thomas Casper

Co-investigators: William Meyer and David Butner

Tracking Code: 92-SR-027

Core Competency Research in the Programs

Introduction

Scientific research is now often conducted on large and expensive experiments that utilize collaborative efforts on a national or international scale to explore physics and engineering issues. This is particularly true for the current US magnetic fusion energy program where collaboration on existing facilities has increased in importance and will form the basis for future efforts. As fusion energy research approaches reactor conditions, the trend is towards fewer large and expensive experimental facilities, leaving many major institutions without local experiments. Since the expertise of various groups is a valuable resource, it is important to integrate these teams into an overall scientific program. To sustain continued involvement in experiments, scientists are now often required to travel frequently, or to move their families, to the new large facilities. This problem is common to many other different fields of scientific research.

The next-generation tokamaks, such as the Tokamak Physics Experiment (TPX) or the International Thermonuclear Experimental Reactor (ITER), will operate in steady-state or long pulse mode and produce fluxes of fusion reaction products sufficient to activate the surrounding structures. As a direct consequence, remote operation requiring robotics and video monitoring will become necessary, with only brief and limited access to the vessel area allowed. Even the on-site control room, data acquisition facilities, and work areas will be remotely located from the experiment, isolated by large biological barriers, and connected with fiber-optics. Current planning for the ITER experiment includes a network of control room facilities to be located in the countries of the four major international partners; USA, Russian Federation, Japan, and the European Community.

Using LDRD funding at LLNL, we have explored issues associated with development of a Remote Experimental Site (RES) concept which would provide an environment for participation in scientific research from geographically distant sites as if all personnel and resources were at the same location. This concept relies on many of the rapidly developing modern computation and communication technologies and has been made practical by the appearance of high performance workstations in a distributed computing environment. In particular, we explored the use of process-to-process communications over high speed wide area networks (WAN) to provide real-time synchronization and to exchange data among multiple computer networks. An internally developed asynchronous interprocess communication system (IPCS) provided a framework for application of WAN-distributed computations fully synchronized with experimental operations. Information used for data analysis, display, and control system applications is now routinely exchanged between geographically distant sites and the experiment, limited only by network latency times and bandwidth. We also explored the use of the distributed computing environment (DCE) as a means of synthesizing high performance computing using multiple workstations connected over the WAN.

This effort was funded under the "core competency research in the programs" category for a total of \$152K over three years with which we were able to research prototype applications. These applications were demonstrated in the operations environment associated with our ongoing collaboration with the DIII-D tokamak experimental group at General Atomics (GA) in San Diego, CA. The following concepts were explored under this funding:

- An extension of our local area network (LAN) interprocess communication system (IPCS) software to applications over a WAN, ESnet.
- Client-server applications using daemon processes to access experimental data over the WAN.
- Creating a "computer cell" structure for WAN use of IPCS
- Experimentation with the National Storage Laboratory (NSL) for network-accessible, high volume data storage of experimental data.
- Use of the Open Software Foundation (OSF) Distributed Computing Environment (DCE) to demonstrate synthesis of a high performance computer over the WAN.
- Application of the Andrews File System (AFS) for a distributed file system in support of experimental operations.

FY92: Remote Experimental Site Concept Development; funding level \$48K

Our IPCS provides a seamless and transparent method for process-to-process interaction between multiple workstations in a heterogeneous computer network involving workstations from multiple vendors employing both UNIX and VMS operating systems. With the FY92 funding,

we extended this IPCS concept to provide synchronization and communication among processes running on computers at LLNL and those at the DIII-D tokamak fusion experimental facility in San Diego, CA. We installed a UNIX workstation in the DIII-D control room with associated data acquisition hardware modules to provide a real time connection to their experimental operations. Utilizing IPCS and this hardware, we fully synchronized our computer network at LLNL to that at GA, subject only to the round trip network delay time which is typically about 100 ms. We have thereby extended the on-line computational resources available to the experiment in a distributed fashion with interactive use predominantly by researchers located remotely at LLNL. As indicated in Figure 1, any computers connected to the local LLNL network that run IPCS and some additional synchronization and monitoring software can be used for real-time experimental operations; this is the essence of distributed computing model. We have implemented this system using the concept of remote procedure calls that form much of the basis for the now emerging technology of DCE, the distributed computing environment.

By employing "daemon" processes running in the background on computers at LLNL and at GA, we have provided user transparent access to the DIII-D data base from our own analysis software and from commercially available packages such as PV-WAVE (from Precision Visuals, Inc.) and IDL (Interactive Data Language from Research Systems, Inc.). This client/server model for data access provides both interactive and automated intershot analysis and display capabilities that fully utilize the computational power at both institutions even though they are separated by some 500 miles. These newly developed capabilities are now supporting operations involvement on the DIII-D experiment for scientists located at LLNL under our experimental collaboration program with General Atomics.

To explore the concept of network-accessible data storage remote from the data generation location at the experiment, we experimented with techniques to use the National Storage Laboratory (NSL), a prototype terabyte capacity data storage facility located at NERSC. The availability of such large volume disk capacity for network accessible storage of experimental data with automatic migration of data among different storage hierarchies is important for future experiments which are expected to generate gigabytes of data on a daily basis. We used the network file system (NFS) protocols to provide connectivity to this storage facility. Our experience with the NSL indicates that it is a useful concept for our type of data transaction (which is different than the design basis for the NSL) where we access a few megabytes of data "randomly" located within shot data files of a few hundred megabytes in size. Due to the still developmental nature of the NSL facility, we presently do not use it to support our routine daily real-time operations. We would, however, pursue use of a production version of such a facility for future experiments, particularly after a distributed file service is implemented.

Figure 1. Remote Experimental Site Network Connections

FY93: IPCS Cells for Remote Experimental Site Development; funding level \$64K

We originally developed and used our IPCS as a software-based mechanism for automatically communicating among processes (computer programs) running within a local area network here at LLNL. IPCS allows processes to pass messages among themselves provided they are uniquely named within the existing network. This system provides communications within a network of heterogeneous computers consisting of both VAX/VMS and UNIX (e.g. HP, Sun, SGI, AEG, and IBM) operating systems and DECNET and tcp/ip network protocols. IPCS processes within a given computer communicate directly. For processes running on different computers, a special IPCS process called the netmailer is used to pass the information. The netmailer is the only process that has knowledge of other computers running IPCS and of the method required for communication.

When the netmailer receives a message locally, it first looks in a table of IPCS process names and associated computers. If the netmailer finds the process name in its name table, it sends the message to the netmailer running on the associated computer using the appropriate communication method. If the name is not found, it starts a search process by communicating in turn with each connected netmailer to determine on which computer the process is running. Upon receipt of a positive response, the sending netmailer caches the process and computer names in its name table and forwards the message.

Significant problems arise when IPCS is used over a wide-area network in our RES configuration. The name search operation becomes considerably more time consuming as the number of connected nodes increases and unnecessarily uses network bandwidth. It also becomes more difficult to assure unique computer and process names since these are often under the control of the local institution. To alleviate these problems we re-designed and implemented an upgrade to IPCS for WAN applications that utilizes the "computer cell" concept associated with the newly emerging distributed computing environment. As shown in Figure 2, each local network, or cell, has one computer designated to run the cell netmailer which must communicate with all other cell netmailers. Within each cell, computer and IPCS process names must remain unique. Across cells, only the IPCS processes that perform intercell services must possess unique names. Similarly, node names need not be unique except for the cell identification nodes. Thus, we have indicated in Figure 2. multiple references to nodes like "computer-a" and to processes like "proc3" which appear in the different cells.

Processes that perform intercell services register with the cell netmailer for their respective cells, as designated by "reg" in Figure 2. When a name search occurs, only the cell netmailers in other cells are queried, thus greatly reducing the name search time and WAN communications. If a name search

results in the finding of a process in another cell, the netmailer performing the name search caches the process name and the name of the other cell node. We have implemented and tested this cell-based IPCS system. We find that there is no effect on performance in our intracell communication compared with the original IPCS concept but some performance degradation can be experienced in passing intercell messages. Previously, as is still true for intracell communication, a message between nodes was operated on by four processes: the sending process, the sending netmailer, the receiving netmailer, and the receiving process. An intercell message, however, is potentially operated on by six processes: the sending and receiving cell netmailers as indicated in Figure 2, plus the four intracell processes. This slight increase in overhead is offset by the advantage of local autonomy in the naming of processes and the overall reduction in communication times for name searches, especially when dominated by local communications, as is the usual case for our applications to date.

Figure 2. IPCS Cell Structure

This cell-based IPCS system provides us with flexibility for operating experiments in near real time (delayed by the network transmission times) over a wide area, heterogeneous computer network. It is used for synchronization of processes running remotely over a WAN with those running locally at the experimental site. It provides for automatic scheduling of processing tasks and for exchange of experimental data in near real time. It

is a critical link in our cooperative computing application using multiple computer networks in support of experimental operations.

FY94: Distributed File System and Security for the Remote Experimental Site; funding level \$40K

Our main objective of this supporting research was to apply the emerging principles of distributed computing to the development of efficient techniques for remotely conducting research collaborations on large physics experiments. Our main thrust during this final year was to demonstrate the applicability of the Open Software Foundation's Distributed Computing Environment (OSF/DCE) to analysis tasks and to use the Andrews File System (AFS) in an operational environment.

AFS is the current state of the art in network distributed file storage and offers the potential for location transparent data access over a wide area network (WAN). We installed AFS at LLNL and have incorporated its use into all of our UNIX-based systems forming the LLNL RES plus extensions to workstations located at NERSC. Analysis of experimental data typically requires only multiple small records of information (several tens of kilobytes) from large data files (~ 125Mbytes) for interactive data analysis by several scientists. Using our previously developed software, data are moved across the WAN from the experimental facility and stored locally in AFS. We have experienced excellent response times for accessing this data, presumably because of a good match between our data use and the cache characteristics of AFS with a system that is transparent to the users. With this successful demonstration of AFS, we will incorporate the distributed file service (DFS, developed from AFS) when available from OSF.

Our test problem of soft x-ray emission tomography for reconstruction of the plasma shape using data from the DIII-D tokamak experiment was ported to the industry standard distributed computing environment. Using OSF/DCE, we demonstrated its utility for enhanced throughput for analysis of experimental data using multiple CPU's which are widely distributed. These results were demonstrated at the SuperComputing93 conference where this image reconstruction application was run simultaneously on 22 computers from 4 different vendors; HP, IBM, Cray, and DEC at 5 different locations; the Conference floor in Portland, OR, LLNL and Sandia sites in Livermore, GA in San Diego, and Cray Research in Colorado. We were able to calculate and display reconstructed images at up to 10 frames/sec (near video rates) and received the "Best Speedup" award in the "Distributed Computing Challenge". This demonstrates the ability to synthesize high performance, WAN-distributed computing on a typical, real experimental data analysis problem.

Acknowledgment

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract Number W-7405-ENG-48.

Summary

Using this LDRD funding, we have been able to explore several areas associated with the distributed computing environment and file systems as applied to our RES concept for remote interaction on large scale physics experiments. We have positioned ourselves to provide computations applications using the newly developing DCE concepts. We have demonstrated the applicability of DCE to solve real computational analysis tasks in support of experimental operations where high performance computing is synthesized from multiple workstations on a WAN. These networks can be synchronized with the real-time experimental operations environment. We now have available for research at LLNL the computer environment shown in Figure 1. Applications to the DIII-D operations are currently being funded from programmatic resources. Some of our software, in particular the interprocess communications system, has been adopted by the DIII-D group to improve their local computer environment by integrating multiple workstations into the real-time operations and to implement a distributed shot file system for data acquisition from a variety of workstations and the central data acquisition computer. Under collaborative agreements with GA, our RES software will be made available to other collaborators to support the DIII-D national effort.