PNL-SA-22693

COMPUTING & INFORMATION SCIENCES

RECEIVED
JUL 19 1993
OSTI

T. H. Dunning, Jr.
J. R. Lewis

June 1993

DISCLAIMER

Handout

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, 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 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.

Prepared for
the U.S. Department of Energy
Contract DE-AC06-76RLO 1830

Pacific Northwest Laboratory
Richland, Washington 99352

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

Computing & Information Sciences

Thom H. Dunning, Jr.
Acting Associate Director

Janice R. Lewis
Sr. Technical Program Manager

The mission of the Computing and Information Sciences program is to develop the advanced computing infrastructure needed for research in the Environmental and Molecular Sciences Laboratory and to conduct applied research and development in computing and information sciences in support of U.S. Department of Energy missions.

Computing is an integral part of the research being conducted by scientists in the Molecular Sciences Research Center (MSRC) at Pacific Northwest Laboratory (PNL). The advanced computing infrastructure for the Environmental and Molecular Sciences Laboratory (EMSL) will provide a state-of-the-art facility where scientific research can be performed to support U.S. Department of Energy missions in environmental restoration and waste management, energy and environmental quality, and science and technology.

The research and technology development elements of the MSRC Computing and Information Sciences (C&IS) program are aimed at providing technical solutions to problems defined by the theoretical and experimental research programs that support the DOE missions. C&IS activities will include Software development in the areas of the graphical user interface and visualization, development of parallel methods for computational chemistry, and development of data storage systems that provide easy access to pertinent scientific data will occur. There also will be substantial emphasis on laboratory automation, including instrument control and data acquisition technologies and telecomputing through which remote collaborators can participate via the network in the planning, execution, and analysis of experi-

ments. The network infrastructure and distributed computing are areas of technical importance to the efficient operation of the EMSL.

The C&IS program is composed of five research and applied development activities: High Performance Computing, Data Management and Archiving, Graphics and Visualization, Distributed Computing, Network Infrastructure, and Computer Instrumentation and Electronics. These activities are briefly described below.

High Performance Computing. Preliminary investigation and planning for the EMSL has clearly indicated that computational chemistry requires high performance computing services in order to understand the complex systems associated with DOE environmental missions. More precisely, a need has been identified for computer systems capable of executing a single application program at a rate of many gigaflops. Computer systems most likely to provide this level of performance across a wide range of chemistry applications will be advanced architecture parallel machines. Many vendors currently have machines of this general description on the drawing board, and a few large-scale prototypes have been delivered. However, there is as yet no consensus regarding either machine characteristics or a programming model. Some

of the differences are fundamental, such as whether the system efficiently provides the same memory address space to all processors ("shared memory") or requires explicit data copying between processors ("message passing"). Such differences can be important and pervasive. Experience has shown that for some applications it is effective to write software based on the most restrictive assumptions. However, for other applications this approach appears to be prohibitively expensive or ineffective. In some cases, architectures of different machines can even mandate different numerical algorithms. The purpose of the C&IS high performance computing activity is to focus on the computer science and engineering aspects of resolving the questions of what type of computer will be best and how EMSL applications should be structured to exploit it and future machines, so as to provide leverage for efforts in developing chemistry applications. This work involves a number of activities.

- evaluation of programming tools and libraries specific to high performance computing
- evaluation or development of nontraditional numerical methods, algorithms, and code structuring techniques
- evaluation of computer architectures and programming environments, including prototype code development and evaluation on available systems
- theoretical studies based on performance models of applications important to MSRC
- establishment of usable interfaces between the high performance computing system(s) and other MSRC computing systems such as the database computer and visualization facilities.

Data Management and Archiving. The C&IS scientific data management activity involves both the selection and installation of high performance data management and archiving hardware as well as the design and develop-

ment of software to provide data management and archiving tools for the scientists in the MSRC. The EMSL will have about 60 instruments, several simulation and modeling programs, and high performance graphics and visualization analysis stations that will generate streams of scientific data. Some individual data sets from these activities are projected to be in the range of 3 to 30 gigabytes. The total volume of datasets of long term interest is estimated to be several terabytes before the end of this decade. These data sets must be readily available even after having been stored for many years. Rapidly moving such large datasets around the EMSL network of distributed processing systems and effectively storing and managing this large volume of data requires high performance and high capacity storage devices connected to high performance networks. The vision for the database and archiving system is to provide a high-performance, large-capacity, modularly-extendible repository rich in information about data sets of long term interest, and rich in methods to store, manage, and effectively search and browse this information. Data will be stored in a multilevel file system (MLFS) that will provide virtually infinite file and file system size. This capability is made possible by automatically moving or "staging" files up and down a hierarchy of successively faster but lower-capacity storage devices. Tool kits will be developed for scientific users and application developers so that they can exploit the capabilities of the data management and archiving hardware.

Graphics and Visualization. EMSL scientists will be required to interpret large volumes of complex data or to execute complex multistep processes requiring access to diverse information. The computer graphics technology required for display of data sets (visualization) is recognized as a key element of the scientific analysis enterprise. The portion of computer codes that communicate directly with the user for problem definition and setup through exchange of text or graphics information (the user interface) is also recognized as

a crucial productivity tool. The C&IS research in this area will first involve determining what information a scientist needs to extract from a particular data set, and what form of graphical representation will be effective in communicating that information. A technology will be implemented that permits facile display and interactive manipulation of the data. A user interface development environment will be implemented that will result in an increase in application development productivity by simplifying the user interface development process and allowing scientific programmers to concentrate more on their particular science and less on the details of the user interface development. Scientific visualization software tools will be developed to provide independent software modules that are extensions to a commercial core scientific visualization system.

Distributed Computing. The objective of the Distributed Computing task is to establish shared computing services for use by distributed workstations, ranging from low-end to high-performance machines of mixed architecture and manufacturer, located throughout the EMSL in offices, laboratories, and shared spaces. Servers will be used to implement shared computing services. File servers will be used to store user files, application software, and some operating system files. Andrew File System (AFS) software from Transarc, Inc. is currently being used for this service but will be replaced by the Distributed File Services (DFS) software from the Open Software Foundation. A common set of servers for all client architectures will simplify file service management and use. EMSL will implement mail standards set by PNL, but the mail service will be setup to deliver mail to a user's DFS home account through a single machine. Print, name, security, and time services will also be supplied through a central file service capability. System management, configuration, and software license management are a combination of policy decisions and software tools that implement policy. Management of these

elements of the computing environment is important to its stability. Management of EMSL workstations is evolving toward use of software that implements a consistent set of procedures for managing several of the Unix platforms that are strategic to EMSL. A consistent set of administration procedures for different workstations will help reduce support costs and improve service. Configuration management policies will formalize the directory structure, control access to critical files, and establish minimum requirements for documentation and maintainability of application software. Software license management standards will be adopted to improve user access to software, reduce software acquisition and maintenance costs, and provide utilization statistics that are useful for acquisition planning.

Network Infrastructure. The network will be the framework for the EMSL computing environment providing the interconnection fabric for the computational systems within and external to the facility. The goal for the network are to provide a high performance, reliable, and easily reconfigurable network that will support the special experiments and computational intensive processes in the EMSL. This goal will be accomplished by designing and installing multiple, partitioned networks to serve fewer users (10-25) than is currently practiced (hundreds) in classic network design. Users will be grouped logically as well as physically on networks, minimizing contention for bandwidth. Critical or often accessed resources will be directly connected or tightly coupled to users' local networks. This close coupling will be accomplished while maintaining centralized management of network systems. EMSL networks have been broadly categorized into three areas. The end-user subscriber networks (ESNs) will service the users' offices and laboratory workstations. Traffic on these networks are generally smaller packets and require short response time. Reliability, manageability, and reconfigurability are key goals attainable with these networks. The Backbone Information Networks (BINs)

provide the mechanism to interconnect ESNs, and provide the transport for the medium to high bandwidth computers and servers. Performance, reliability, manageability, and robustness are the key attributes for this class of network. Computer-to-Computer Network (C2N) is the highest performance network, and will serve the needs of the relatively few high performance systems in the EMSL. Overall performance for transfers of large blocks of data is the design goal for this type of network.

Computer Instrumentation & Electronics. The goal of this task is to perform research and development on data acquisition hardware and software, instrumentation systems, instrument modeling and simulation, and electronics as applied to instrument development. These capabilities include the Interface Development Laboratory (IDL), which provides the tools and environment to develop instrumentation interfaces, perform instrument modeling and optimization, and research prototype data acquisition and control systems. The IDL serves as a focus for the evaluation of data acquisition and control hardware and software as well as providing an electronics design, testing and assembly area, including CAD/CAE capabilities. The IDL includes electronic circuit design and prototyping for front-end and interface electronics. The IDL also includes data analysis software tools that are closely linked to the data acquisition needs for evaluation and prototyping. The group carries out research in data acquisition instrumentation, including Unix-based, DOS-based and POSIX-compliant systems, neural network and digital signal processing applications, distributed data acquisition, expert systems applications, real-time debuggers, instrument modeling and exploration of approaches to data acquisition software design. The IDL provides facilities for the development of high performance data acquisition systems and is central to the information integration process needed by the EMSL research programs.

END

**DATE
FILMED**

8 / 18 / 93