Enhancing Functionality and Performance in the PVM Network Computing System

U.S. Department of Energy Grant DE-FG05-91ER25105
Final Project Report

Vaidy Sunderam
Department of Mathematics and Computer Science
Emory University, Atlanta, GA 30322, USA
vss@mathcs.emory.edu

We have no objection from a patent
standpoint to the publication or
dissemination of this material.

1 Overview

The work funded by this grant was part of an ongoing research project in heterogeneous distributed computing with the PVM system, at Emory as well as at Oak Ridge Labs and the University of Tennessee. This grant primarily supported research at Emory on various facets of network computing in general and the PVM system in particular. However, the research efforts also include the PI's ongoing interaction with the other groups in terms of collaborative research as well as software systems development and maintenance. The PVM related phase of our efforts is now complete, although we are continuing work on the next generation heterogeneous computing environments under the *Harness* project. In this report, we summarize our research results and accomplishments with respect to the PVM project, until the end of the funding and project periods, viz. September 14, 1998.

2 Overall Progress

The mainstay of this research project was the PVM system, a software framework to enable heterogeneous network computing. Since its original inception in 1989, the PVM system has been extremely popular and useful, as a software framework, and as a research platform. A large number of copies of the software have been retrieved by numerous institutions worldwide, and are in regular use for production and development applications. The underlying concurrent computing model, and concepts in heterogeneous computing that PVM embodies, have also been the basis for a number of research efforts in distributed processing, and has gained adoption in the teaching of parallel computing. Despite newer models and

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.

interfaces such as MPI, the PVM system continues to be viable and has a sizeable following in the scientific and research community. From the research perspective, the project has made significant contributions in a number of areas that are outlined below.

2.1 Kernel Evolution:

Over the years, a substantial amount of effort was devoted to evolving the PVM kernel. Both in response to user requests as well as subsequent to research investigations, various portions of the core PVM system were enhanced, refined, and optimized. For example, the API was extended to provide hooks to support external schedulers and resource managers, thus enabling third-party subsystems to interface seamlessly with PVM, and to be replaceable without undue programming effort. The runtime system was strengthened considerably, permitting enforcement of security restrictions without requiring system administration privileges on behalf of the user or installer. Buffer management schemes suited to the unique nature of scientific parallel programs (as opposed to transaction or database-oriented systems and multimedia systems) were developed, leading to a more robust and efficient implementation. Recently, we have also developed extensions that allow the PVM system to co-exist with, and complement, MPI implementations.

2.2 Programming Tools:

Work in this area was undertaken along two fronts. The first was HeNCE (Heterogeneous Network Computing Environment), an X-window based software environment designed to assist scientists in developing parallel programs that run on a network of computers. HeNCE provides the programmer with a high level abstraction for specifying parallelism. HeNCE is based on a parallel programming paradigm where an application program can be described by a graph. HeNCE graphs are variants of directed acyclic graphs, or DAGS. Nodes of the graph represent subroutines and the arcs represent data dependencies. Individual nodes are executed under PVM. HeNCE is composed of integrated graphical tools for creating, compiling, executing, and analyzing HeNCE programs. HeNCE proved to be a valuable complement to help parallel programming and monitoring of PVM. The second project concerning programming tools for PVM is the Shared Object system. This framework supports unified address-space programming in PVM, using the object oriented model. The work addresses issues of consistency semantics, language extensions and support, and system level design to support shared object facilities. In the DoPVM system, facilities are provided that can reduce the overheads associated with sharing objects across a distributed environment. Some of these are mechanisms by which the granularity of sharing can be controlled, integral support for scheduling, problem partitioning, and provisions for implementing function and agenda-parallelism.

2.3 Efficiency Aspects:

A major issue in any network computing system is performance. In this project, our efforts in this regard have been aimed at communication protocols that support heterogeneous network computing. Particularly with the imminent advent of gigabit networks, protocols will be a crucial factor in enabling effective network computing. To address this, we have proposed design and implementation strategies for providing general purpose distributed computing primitives on computer networks. This suite of primitives is intended to be a framework within which distributed and concurrent applications may be built in networked environments, in the absence of a distributed operating system. An experimental version of these primitives has been implemented, and performance increases of 20-35% have been attained. On a different front, we also investigated partitioning and scheduling techniques for increasing efficiency in network based concurrent systems. Findings indicated that addressing load imbalance and task allocation strategies have the greatest potential for improved overall performance. Consequently we proposed a new structuring technique, based on domain-specific libraries and load to resource-need matching that is aimed at obtaining optimal performance under PVM.

2.4 Client Server Computing:

To expand the scope and applicability of the PVM system, we have developed an alternate interface to PVM that supports client-server computing. The core of this interface is a remote procedure call (RPC) facility that permits the specification and export of services which may be invoked by clients using the well-established RPC paradigm and mechanics. Previous experiences have established that this model is natural and effective for programming distributed applications, and that RPC can be implemented in message-passing systems like PVM with a minimum of overhead. This model also enables significant additional functionality such as user-transparent load balancing, failure resilience, and adaptive parallelism. A preliminary implementation of the RPC system has been completed and both functionality and performance have been found to be significantly improved and useful in early experiments.

2.5 Input Output Systems:

The PIOUS project is the basis for extending PVM applicability to data bound parallel programs. The PIOUS project has progressed very well, with the development and release of beta level software and libraries, which have attained high levels of scalable I/O performance over networks and clusters of workstations. The PIOUS architecture was developed and the PVM implementation of PIOUS was completed. PIOUS employs data declustering to exploit the combined file I/O and buffer cache capacities of networked computing

resources, and transaction-based concurrency control, to guarantee access consistency without explicit synchronization. PIOUS consists of a set of data servers, a service coordinator, and library routines linked with client processes. An underlying transport mechanism is assumed to carry messages between client processes and components of the PIOUS architecture. A prototype of the PIOUS system has been developed and has been distributed to the PVM community. Further, we have used the PIOUS system to implement the NHT-1 I/O benchmarks over PVM and measure I/O throughput – we have found that upwards of 80% of the transport bandwidth delivered by PVM was available to the I/O subsystem, despite the concurrency control synchronization and the transaction based I/O mechanism used.

2.6 Threads-based computing:

The TPVM (Threads-oriented PVM) system is a subsystem for the PVM distributed system, which supports the use of lightweight processes or “threads” as the basic unit of parallelism and scheduling. TPVM provides a library interface which presents both a traditional, task based, explicit message passing model, as well as a data-driven scheduling model that enables straightforward specification of computation based on data dependencies. The TPVM system comprises three basic modules: a library interface that provides access to thread-based distributed concurrent computing facilities, a portable thread interface module which abstracts the required threads-related services, and a thread server module which performs scheduling and system data management. A prototype implementation has allowed us to perform a number of preliminary experiments. We have also ported TPVM to multiple platforms. Results from its use are very encouraging and numerous external users have retrieved and are using the software. We are also converting the NAS parallel benchmark suite to use TPVM and have had some success in reducing latency and achieving increased communication-computation overlap.

2.7 Monitoring and Visualization

The PVaniM 2.0 system provides online and postmortem visualization support as well as rudimentary I/O for long running, communication-intensive PVM applications. PVaniM 2.0 provides these features while using several techniques to keep system perturbation to a minimum. The online graphical views provided by PVaniM 2.0 provide insight into message communication patterns, the amount of messages and bytes sent, host utilization, memory utilization, and host load information. For online visualization analysis, PVaniM 2.0 utilizes sampling to gather data regarding interesting aspects of the application as well as interesting aspects of the cluster environment. With sampling, necessary statistics are collected and sent to the monitor intermittently. The rate at which the application data is sent to the monitor is a parameter that may be set by the user. With lower sampling rates, the application will experience less perturbation, but the graphical views will not

be updated as frequently. For offline visualization analysis, PVaniM 2.0 utilizes buffered tracing to provide support for fine-grain visualization systems. Currently, PVaniM produces tracefiles for the prototype PVaniM system. A converter is provided that allows the user to also use the popular ParaGraph system using techniques similar to PGPVM. This system has been found to be a very useful complement to XPVM and provides valuable facilities for PVM developers.

2.8 Other Projects

In addition to these major endeavors, a number of smaller research investigations based on PVM were also undertaken. We developed a version of PVM that incorporates fault tolerance that uses the *sender-based logging* technique to maintain consistent state information across PVM processes, and checkpoints processes periodically to enable restart of failed processes in a consistent state in the event of a crash. We also developed a generalized suite of distributed computing primitives and communication architectures for network-based heterogeneous concurrent computing. A benchmark suite consisting of computational fluid dynamics codes was developed and made publicly available. Extensions to PVM were developed to enable performance profiling tools for multicomputers (e.g. ParaGraph) to be usable with PVM programs. Optimizations to the message passing substrate to support heterogeneity without sacrificing efficiency were developed. System add-ons such as those mentioned above contributed to both the research and software system aspects of the software, and such effort are continuing.

3 Research Output

This project has resulted in a number of research publications, professional lectures, workshops and tutorials, and graduate student theses and projects. Research publications and activities related to this project are listed in this section.

3.1 Publications:

- Vaidy Sunderam, Brad Topol, Steve Moyer, Alan Krantz, "Tools and Auxiliary Subsystems in PVM", to appear in *Future Generation Computer Systems*, 1999.
- I. Zoraja, H. Hellwagner, and V. S. Sunderam, "SCIPVM: Parallel Distributed Computing on SCI Workstation Clusters", to appear, *Journal of Concurrency: Practice and Experience*, 1998.
- B. Topol, J. Stasko, and V. S. Sunderam, "Monitoring and Visualization in Cluster Environments", to appear, *Journal of Concurrency: Practice and Experience*, 1998.

- B. Topol, J. Stasko, and V. S. Sunderam, "Dual Timestamping Methodology for Visualizing Distributed Application Behavior", *International Journal of Parallel and Distributed Systems and Networks*, Vol. 1, No. 2, 1998.
- A. Ferrari and V. S. Sunderam, "Multi-Paradigm Distributed Computing with TPVM", *Journal of Concurrency: Practice and Experience*, Vol. 10, No. 3, pp. 199-228, March 1998.
- S. Moyer and V. S. Sunderam, "Characterizing Concurrency Control Performance for the PIOUS Parallel File System", *Journal of Parallel and Distributed Computing*, Vol. 38, No. 1, pp. 81-91, October 1996.
- V.S.Sunderam and S. Moyer, "Parallel I/O for Distributed Systems: Issues and Implementation", *Future Generation Computer Systems*, Vol. 12, No. 1, pp. 25-38, May 1996.
- A. Beguelin, J. Dongarra, A. Geist, R. Manchek, and V. S. Sunderam, "Recent Enhancements to PVM", *International Journal of Supercomputer Applications*, Vol. 9, No. 2, pp. 108-127, summer 1995.
- S. White, A. Alund, and V. S. Sunderam, "Performance of the NAS Parallel Benchmarks on PVM Based Networks", *Journal of Parallel and Distributed Computing*, Vol. 26, No. 1, pp. 61-71, April 1995.
- S. Moyer and V. S. Sunderam, "Parallel I/O as a Parallel Application", *International Journal of Supercomputer Applications*, Vol. 9, No. 2, pp. 95-107, summer 1995.
- V. S. Sunderam, G. A. Geist, J. J. Dongarra, and R. Manchek, "The PVM Concurrent Computing System: Evolution, Experiences, and Trends", *Journal of Parallel Computing*, Vol. 20, No. 4, pp. 531-546, March 1994.
- B. Schmidt and V. S. Sunderam, "Empirical Analysis of Overheads in Cluster Environments", *Journal of Concurrency: Practice and Experience*, Vol. 6, No. 1, pp. 1-32, February 1994.
- A. Beguelin, J. Dongarra, A. Geist, and V. Sunderam "Visualization and Debugging in a Heterogeneous Environment", *IEEE Computer*, Vol. 26, No.6, pp. 88-95, June 1993.
- V. S. Sunderam, "Heterogeneous Environments for Network Concurrent Computing", *Journal of Future Generation Computer Systems*, Vol. 8, Nos. 1-3, pp. 191-203, July 1992.
- G. A. Geist and V. S. Sunderam, "Network Based Concurrent Computing on the PVM System", *Journal of Concurrency: Practice and Experience*, Vol. 4, No. 4, pp. 293-311, June 1992.

- V. S. Sunderam, "PVM : A Framework for Parallel Distributed Computing", *Journal of Concurrency: Practice and Experience*, Vol. 2, No.4, pp. 315-339, December 1990.
- C. Carothers, B. Topol, R. Fujimoto, J. Stasko, V. Sunderam, "Visualizing Parallel Simulations in Network Computing Environments", *Proc. 1997 Winter Simulation Conference*, pp. 110-117, Orlando, December 1997.
- A. T. Krantz and V. S. Sunderam, "Client Server Computing on Message Passing Systems: Experiences with PVM-RPC", *Proc. Third Intl. Euro-Par Parallel Processing Conference*, pp. 110-117, Passau, Germany, August 1997.
- O. Krone, M. Aguilar, B. Hirsbrunner, V. Sunderam, "Integrating Coordination Features in PVM", *Proceedings - First International Conference on Coordination Models and Languages*, Cesena, Italy, Springer-Verlag, pp. 432-435, April 1996.
- V. S. Sunderam and H. Hellwagner, "Parallel Distributed Computing on SCI Workstation Clusters: Early Experiences", *Proceedings - Parallele Algorithmen, Rechnerstrukturen und Systemsoftware Workshop Stuttgart*, Germany, pp. 27-34, October 1995.
- A. Ferrari and V. S. Sunderam, "TPVM: Distributed Concurrent Computing with Lightweight Processes", *Proceedings - 4th High-Performance Distributed Computing Symposium*, Washington, DC, pp. 211-218, August, 1995.
- B. Topol, J. Stasko, and V. Sunderam, "Integrating Visualization Support Into Distributed Computing Systems", *Proceedings - 15th International Conference on Distributed Computing Systems*, Vancouver, BC, pp. 19-26, May 1995.
- S. Moyer and V. S. Sunderam, "PIOUS: A Scalable Parallel I/O System for Distributed Computing Environments", *Proceedings - 1994 Scalable High Performance Computing Conference*, pp. 71-78, Knoxville, TN, May 1994.
- S. Moyer and V. S. Sunderam, "A Parallel I/O System for High-Performance Distributed Computing", *Proceedings - IFIP WG10.3 Working Conference on Programming Environments for Massively Parallel Distributed Systems*, pp. 25.1-25.11, Manno, Switzerland, April 1994.
- S. Y. Cheung and V. S. Sunderam, "Performance of Barrier Synchronization Methods", *Proceedings - International Conference on Computing and Information*, (Eds. D. Krumme et. al.), pp. 175-179, Sudbury, ON, May 1993.
- V. S. Sunderam, "DCL: Protocols and Primitives for Distributed Concurrent Computing", *Proceedings - International Conference on Computing and Information*, (Eds. D. Krumme et. al.), pp. 160-164, Sudbury, ON, May 1993.

- C. Hartley and V. S. Sunderam, "Concurrent Programming with Shared Objects in Networked Environments", *Proceedings - 7th Intl. Parallel Processing Symposium*, pp. 471-478, Los Angeles, CA, April 1993.
- G. A. Geist and V. S. Sunderam, "The Evolution of the PVM Concurrent Computing System", *Proceedings - 26th IEEE Compcon Symposium*, pp. 471-478, San Francisco, CA, February 1993.
- V. S. Sunderam, "Design Issues in Heterogeneous Network Computing", *Proceedings - IEEE Workshop on Heterogeneous Processing*, pp. 101-112, Los Angeles, CA, March 1992.
- A. Beguelin, J. Dongarra, G. Geist, R. Manchek, V. S. Sunderam, "Graphical Development Tools for Network-Based Concurrent Supercomputing", *Proceedings - Fourth IEEE Supercomputing Conference*, pp. 435-444, Albuquerque, NM, November 1991.
- G. A. Geist and V. S. Sunderam, "The PVM System: Supercomputer Level Concurrent Computation on a Heterogeneous Network of Workstations", *Proceedings - Sixth Distributed Memory Computing Conference*, pp. 258-261, Portland, OR, April 1991.
- A. Beguelin, J. Dongarra, G. Geist, R. Manchek, V. S. Sunderam, "Solving Computational Grand Challenges Using a Network of Heterogeneous Supercomputers", *Proceedings - Fifth SIAM Conference on Parallel Processing*, pp. 234-239, Houston, TX, March 1991.

3.2 Talks and Presentations

- "Systems Infrastructures for Distributed Collaborative Computing", invited talk, Parallel and Distributed Processing Techniques and Applications 1998, Las Vegas, NV, July 1998.
- "Metacomputing with the IceT System", invited talk, NSF-INRIA Workshop on Clusters and Computational Grids for Scientific Computing, Townsend, TN, September 1998.
- "Client Server Computing on Message Passing Systems: Experiences with PVM-RPC", contributed talk, Third Intl. Euro-Par Parallel Processing Conference, Passau, August 1997.
- "An RPC Facility for PVM", contributed talk, Cluster Computing Conference, CCC '97, Atlanta, GA, March 1997.
- "The PVM System: Status, Trends, and Directions", invited talk, EuroPVM User's Group Meeting, Munich, October 1996.

- "Performance and Functionality Issues in Heterogeneous Systems", invited talk, University of Alberta, Edmonton, September 1996.
- "PVM Data Transfers on SCI Workstation Clusters: Early Experiences", contributed talk, 4th PVM User's Group Meeting, Santa Fe, NM, February 1996.
- "Performance and Functionality Issues in Heterogeneous Systems", invited talk, Dartmouth College, Hanover, NH, November 1995.
- "Threads Based Concurrent Computing with TPVM, contributed talk, 4th High Performance Distributed Computing Symposium, Washington, DC, August, 1995.
- "Heterogeneous Concurrent Computing with PVM: Recent Developments and Trends", invited talk, Brown University, Providence, RI, December 1994.
- "Heterogeneous Computing with Exportable Services", invited talk, 2nd CNRS-NSF Workshop on Environments and Tools, Townsend, TN, May 1994.
- "PVM: Recent Developments and New Initiatives", invited talk, 2nd PVM User's Group Meeting, Oak Ridge, TN, May 1994.
- "Experimental Systems and Methodologies for Network Computing", invited talk, National Institute of Standards and Technology, Gaithersburg, MD, May 1994.
- "Enhancing Functionality and Performance in the PVM Network Computing System", invited talk, University of Virginia, Charlottesville, VA, March 1994.
- "Heterogeneous Concurrent Computing with PVM: Recent Developments and Trends", invited talk, Washington University, St. Louis, MO, February 1994.
- "Design Issues in Heterogeneous Concurrent Computing", invited talk, George Washington University, Washington, DC, March 1993.
- "Design Issues in Heterogeneous Computing Systems", invited talk, Georgia Institute of Technology, Atlanta, GA, October 1992.
- "Scientific Computing on Heterogeneous Networks", invited talk, NASA Ames Research Center, Moffett Field, CA, May 1992.
- "Heterogeneous Network Computing: The PVM Perspective", invited talk, Sandia National Laboratory, Livermore, CA, May 1992.
- "Towards Standards for Heterogeneous Network Computing", invited talk, *European RAPS Workshop*, Bonn, Germany, May 1992.
- "Heterogeneous Network Computing: The PVM System", invited talk, North Carolina Supercomputer Center, Research Triangle Park, NC, August 1991.

- "Supercomputing on Heterogeneous Networks", invited talk, NCSA, University of Illinois, Urbana-Champaign, IL, June 1991.
- "Supercomputer Level Concurrent Computations on a Network of RS/6000 Powerstations", invited talk, *Large Scale Computer Analysis and Modeling Conference*, Park City, UT, April 1991.

3.3 Graduate Student Theses

- Michael Solomon, M. S. May 1998. "Parallel Database Implementation with PVM".
- Nancy Sedora, M. S. September 1995. "Implementation and Analysis of the NHT-1 I/O Benchmarks on PIOUS for PVM".
- Scott Shipley, M. S. June 1995. "Enhancing the Efficiency of World Wide Web Servers".
- Xiaowu Lu, M. S. April 1995. "Implementation and Performance Issues of NAS Parallel Benchmarks on PVM 3.3"
- Adam Ferrari, M. S. April 1994. "Object-Based Parallelization for the Conch System".
- Benjamin Temko, M. S. August 1993. "Implementing Conch on the Intel Hypercubes".
- Scott M. White, M.S. April 1993. "The NAS Benchmarks on Virtual Parallel Machines".
- Brad B. Topol, M.S. April 1993. "Second Generation Heterogeneous Computing with Conch".
- Daniel Penn, M.S. April 1993. "A Simulation of Attractor Neural Networks in a Distributed Processing Environment".
- Melisa Kelley, M.S. April 1992. "An Interactive Tool for Data Parallel Concurrent Processing".
- Brian Schmidt, M.S. April 1992. "Empirical Analysis of Overheads in Cluster Environments".
- Charles Hartley, M.S. April 1992. "Distributed Object Programming on General Purpose Networks".
- Azita Miahnahri, M.S., April 1992. "Simulation of Attractor Neural Networks".
- Scott Bogartz, M.S., April 1991; B.S. Honors, April 1991. "Design and Implementation of a Distributed Groupware System".