LLNL--9800607G

Development of a Dynamic Time Sharing Scheduled Environment

Final Report
CRADA No. TC-824-94E

Date: Revision:

A. Parties

The project is a relationship between the Lawrence Livermore National Laboratory (LLNL) and Cray Research Inc. (CRI)

University of California
Lawrence Livermore National Laboratory
PO Box 808, L-795
Livermore, CA 94550

Cray Research Inc.
655 Lone Oak Drive
Eagan, MN 55121

B. Project Scope

Massively parallel computers, such as the Cray T3D, have historically supported resource sharing solely with space sharing. In that method, multiple problems are solved by executing them on distinct processors. This project developed a dynamic time- and space-sharing scheduler to achieve greater interactivity and throughput than could be achieved with space-sharing alone. CRI and LLNL worked together on the design, testing, and review aspects of this project. There were separate software deliverables. CRI implemented a general purpose scheduling system as per the design specifications. LLNL ported the local gang scheduler software to the LLNL Cray T3D. In this approach, processors are allocated simultaneously to all components of a parallel program (in a "gang"). Program execution is preempted as needed to provide for interactivity. Programs are also relocated to different processors as needed to efficiently pack the computer's torus of processors.

In phase one, CRI developed an interface specification after discussions with LLNL for system-level software supporting a time- and space-sharing environment on the LLNL T3D. The two parties also discussed interface specifications for external control tools (such as scheduling policy tools, system administration tools) and applications programs. CRI assumed responsibility for the writing and implementation of all the necessary system software in this phase.

In phase two, CRI implemented job-rolling on the Cray T3D, a mechanism for preempting a program, saving its state to disk, and later restoring its state to memory for continued execution. LLNL ported its gang scheduler to the LLNL T3D utilizing the CRI interface implemented in phases one and two.

During phase three, the functionality and effectiveness of the LLNL gang scheduler was assessed to provide input to CRI time- and space-sharing efforts. CRI will utilize this information in the development of general schedulers suitable for other sites and future architectures.

DISCLAIMER

All phases of this project were completed on time and all deliverables were met without significant changes to the original statement of work.

C. Technical

Cray Research was able to provide a highly effective job-rolling mechanism. This mechanism also provided the flexibility of restoring the program's state to different processors than were initially utilized.

LLNL completed the initial installation of the gang scheduler in March of 1996. The next few months were spent perfecting the scheduling algorithms and tuning. Early simulations indicated that interactivity could be improved dramatically, and the computer's saturation point could also be increased. Since LLNL's Cray T3D was originally configured for a high level of interactivity with moderate throughput, we were able to reconfigure the computer for dramatically higher throughput. The gang scheduler was able to provide even greater interactivity while utilization was increased from the 30 percent range to the 90 percent range—a phenomenal level of throughput for a massively parallel computer.

Issues outside of the scope of this CRADA, but of interest for further study include: the comparison of job-paging rather than rolling, gang-scheduling on distributed memory computer architectures, and gang-scheduling on computers architectures in which greater flexibility in processor assignment exists.

D. Partner Contribution

All deliverables have been met. CRI developed a number of documents and a highly effective job-rolling mechanism. The LLNL gang scheduler has proven to be highly effective at improving system throughput and interactivity. In parallel with the development of LLNL's gang scheduler, CRI developed a gang scheduler based upon the Dynamic Job Manager, which has been distributed by CRI. Some U.S. government agencies have installed LLNL's gang scheduler on their Cray T3D computers.

No inventions were created as part of this CRADA.

E. Documents/Reference List

"Improved Utilization and Responsiveness with Gang Scheduling," Dror G. Feitelson and Morris A. Jette, *Job Scheduling Strategies for Parallel Processing Workshop*, (publication pending) April 1997.

"The Gang Scheduler—Timesharing on a Cray T3D," Morris Jette, David Storch, and Emily Yim, Cray User Group Meeting, March 1996.

Substantial documentation available on the internet at "http://www-lc.llnl.gov/dctg/gang."

F. Acknowledgment

Participant's signature of the final report indicates the following:

1) The Participant has reviewed the final report and concurs with the statements made therein.
2) The Participant agrees that any modifications or changes from the initial proposal were discussed and agreed to during the term of the project.
3) The Participant certifies that all reports either completed or in process are listed and all subject inventions and the associated intellectual property protection measures attributable to the project have been disclosed or are included on a list attached to this report.
4) The Participant certifies that if real property was exchanged during the agreement, all has either been returned to the initial custodian or transferred permanently.
5) The Participant certifies that proprietary information has been returned or destroyed by LLNL.

D J Caliga 10 Nov 1997

David Caliga Date
Cray Research Inc.

Morris Jette Date
Lawrence Livermore National Laboratory

Attachment I – Final Abstract
Attachment II – Project Accomplishments Summary
Attachment III – Final Quarterly Report

F. Acknowledgment

Participant's signature of the final report indicates the following:

1) The Participant has reviewed the final report and concurs with the statements made therein.
2) The Participant agrees that any modifications or changes from the initial proposal were discussed and agreed to during the term of the project.
3) The Participant certifies that all reports either completed or in process are listed and all subject inventions and the associated intellectual property protection measures attributable to the project have been disclosed or are included on a list attached to this report.
4) The Participant certifies that if real property was exchanged during the agreement, all has either been returned to the initial custodian or transferred permanently.
5) The Participant certifies that proprietary information has been returned or destroyed by LLNL.

D D Caliga 10 Nov 1997

David Caliga Date
Cray Research Inc.

Morris Jette 1-26-98

Morris Jette Date
Lawrence Livermore National Laboratory

Attachment I – Final Abstract
Attachment II – Project Accomplishments Summary
Attachment III – Final Quarterly Report

Development of a Dynamic Time Sharing Scheduled Environment

Final Abstract
Attachment I
CRADA No. TC-0824-94E

Massively parallel computers, such as the Cray T3D, have historically supported resource sharing solely with space sharing. In that method, multiple problems are solved by executing them on distinct processors. This project developed a dynamic time- and space-sharing scheduler to achieve greater interactivity and throughput than could be achieved with space-sharing alone. CRI and LLNL worked together on the design, testing, and review aspects of this project. There were separate software deliverables. CRI implemented a general purpose scheduling system as per the design specifications. LLNL ported the local gang scheduler software to the LLNL Cray T3D. In this approach, processors are allocated simultaneously to all components of a parallel program (in a "gang"). Program execution is preempted as needed to provide for interactivity. Programs are also relocated to different processors as needed to efficiently pack the computer's torus of processors.

Lawrence Livermore National Laboratory is operated by the University of California for the U.S. Department of Energy
Technical Transfer Initiatives Program, P.O. Box 808 L-795, Livermore, CA 94551