Final Project Report
April 2009

Runtime Data Management for Data-Intensive Scientific Applications
DOE ECPI Program
Award duration: 08/15/2005 – 08/14/2008
PI: Xiaosong Ma (NC State University / Oak Ridge National Laboratory)

Research Progress and Results

- In the last several months of the project (during the no-cost extension), we have investigated providing parallel data processing services on distributed multi-core workstations. In particular, we explored an innovative hybrid architecture for data-intensive services on top of unreliable environments, by supplementing volatile nodes with a small group of dedicated, highly available nodes. We studied the impact of such an architecture on the data replication requirement, as well as the effectiveness of hybrid-resource-aware data processing task scheduling. Our results, obtained from Virginia Tech cluster using both synthetic node availability traces and a real-world desktop grid trace show promising results. We are submitting a technical paper on this study for conference publication.

- We finished our project on automatic scheduling for transparently parallelized data processing scripts. This work builds on our previous parallel R tool, and enables transparent run-time data processing tasks by building a "personal performance database" through R function and loop cost monitoring and machine learning techniques. With task cost prediction and inter-task dependence analysis, our parallel R tool will be able to use an extended static scheduling algorithm to optimize task partitioning to achieve improved load balance while exploiting data locality. Our results are published in a ISPASS paper in April 2009.

- We have recently achieved impressive scalability using our flexible database fragmentation/replication, as well as hierarchical query scheduling techniques in running our mpiBLAST-pio tool on the new IBM BG/P system. The parallel execution efficiency was measured as 93% in a 32,768-core run. We have reported the results as well as our optimizations in a SC|08 paper. The work is done in collaboration with Virginia Tech, Argonne National Laboratory, and IBM.

- One paper was published at SC|07, generated by joint work between NCSU and ORNL, on augmenting parallel file systems to embed data source information for transparent input data recovery from remote storage.

- We participated in a collaborative effort on scalable parallel BLAST search, which won an award at SC|07. ParaMEDIC, the collaborated project between researchers in Argonne National Laboratory, Virginia Tech, and NCSU PhD student Heshan Lin (supported solely by this ECPI award at NCSU), was chosen

as the ***Winner of the SC|07 Storage Challenge***. ParaMEDIC used 12,000 cores, performed 256 Trillion searches, and generated 1 Petabyte of data. The NCSU work focused on the scalable I/O and output data processing for parallel BLAST.

- We continued our research on massively parallel biological database searches. Our optimized version of the parallel BLAST search tool, due to its performance and reliability, was chosen to be used on the majority of nodes in a cross-continent parallel BLAST endeavor involving 3,000 nodes during the SC|05 StorCloud Challenge event in November 2005. Through this large-scale effort, we addressed run-time query scheduling, database distribution, and fault-tolerance in parallel biological database searches. A conference paper discussing our experience from this SC StorCloud Challenge endeavor appeared in SC|06 and was one of the three papers ***nominated for the best paper award***.

- We obtained initial results in our investigation on efficient data and query scheduling for multiple-database parallel biological sequence search. To obtain an optimized tradeoff between system throughput and individual query's response time, we have developed several scheduling algorithms to make intelligent query and database fragment assignments to processors. We carried out simulation study using a simulator verified against real cluster results. From our experiments, we show that careful scheduling considering parallel search efficiency, system load, and data locality yields an orders-or-magnitude difference in the average query response time of an online parallel BLAST server. The results are published in a conference paper accepted by SSDBM '08.

- We also continued our project on data management in desktop distributed storage of scientific datasets. We furthered our research and studied systematically distributed, striped caching of scientific data on unreliable nodes. We also extended our investigation from the desktop environment to traditional parallel machines, and proposed novel data recovery techniques based on online data patching from parallel job owners' original data sources. Two papers describing the distributed storage architecture and the recovery methodology were published in the HyperI/O Workshop in 2006 and the Operating Systems Review in 2007 respectively, with a related journal paper on caching and a conference paper on data recovery to be submitted soon.

- We have obtained promising results from parallel R project, in collaboration with the pR team at ORNL (a DOE SciDAC project). We have recently developed a research prototype of pR that automatically parallelizes an R script (without requiring any code modification) and efficiently executes it in parallel by performing runtime, incremental code analysis and scheduling. An overall description of this project was published in Journal of Physics in 2006, and a conference paper is currently under submission.

Technology Transfer and Results Dissemination

- Some of our proposed techniques presented in the SC|07 paper on automatic data staging have been incorporated into the production system of the Jaguar machine (currently No. 7 in the top500 list).

- The mpiBLAST-pio tool developed in this research, which was released as a branch of the popular mpiBLAST software (please see our last year's project report), has attracted more attention from industry. IBM researchers have worked with PhD student Heshan Lin (supported by this award) and our collaborator to port the tool to the BlueGene platform. They have reported that the parallel search performance could scale up to 8,000 nodes in a paper recently accepted by the ACM International Conference on Computing Frontiers.

Publications and Presentations

- Chao Wang, Zhe Zhang, Xiaosong Ma, Sudharshan Vazhkudai, and Frank Mueller, "Improving the Availability of Supercomputer Job Input Data Using Temporal Replication", the International Supercomputing Conference, Jun 2009. (Collaboration with ORNL)
- Jiangtian Li, Xiaosong Ma, Karan Singh, Martin Schulz, Bronis R. de Supinski, and Sally A. McKee, "Machine Learning Based Online Performance Prediction for Runtime Parallelization and Task Scheduling", 2009 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS '09), Apr 2009. (Collaboration with LLNL and Cornell University)
- Heshan Lin, Pavan Balaji, Ruth Poole, Carlos Sosa, Xiaosong Ma, Wuchun Feng, "Massively Parallel Genomic Sequence Search on the BlueGene/P Architecture", Supercomputing 2008 (SC|08), Nov 2008. (Collaboration with ANL, IBM, and Virginia Tech)
- Chao Wang, Zhe Zhang, Sudharshan S. Vazhkudai, Xiaosong Ma and Frank Mueller, "On-the-fly Recovery of Job Input Data in Supercomputers", The 37th International Conference on Parallel Processing (ICPP '08), Sept. 2008. (Collaboration with ORNL)
- Heshan Lin, Xiaosong Ma, Jiangtian Li, Ting Yu, and Nagiza Samatova, "Adaptive Request Scheduling for Parallel ScientificWeb Services", 20th International Conference on Scientific and Statistical Database Management (SSDBM '08), Jul 2008. (Collaboration with ORNL)
- Pavan Balaji, Wu-chun Feng, Jeremy S. Archuleta, Heshan Lin, Rajkumar Kettimuthu, Rajeev Thakur, Xiaosong Ma, Semantics-based distributed I/O for mpiBLAST, poster, the 13th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPOPP 08), Feb 2008. (Collaboration with ANL and Virginia Tech)
- Zhe Zhang, Chao Wang, Sudharshan S. Vazhkudai, Xiaosong Ma, Gregory G. Pike, John W. Cobb, and Frank Mueller, “Optimizing Center Performance through Coordinated Data Staging, Scheduling and Recovery”, *Supercomputing 2007*, Nov 2007. (Collaboration with ORNL)
- Sarat Sreepathi, Kumar Mahinthakumar, Emily Zechman, S. Ranji Ranjithan, Downey Brill, Xiaosong Ma, and Gregor von Laszewski: "Cyberinfrastructure for Contamination Source Characterization in Water Distribution Systems", International Conference on Computational Science (ICCS ’07), Beijing, China, May 2007. (Collaboration with ANL)
- Oystein Thorsen, Karl Jian, Amanda Peters, Brian Smith, Heshan Lin, Wu-chun Feng and Carlos P. Sosa, "Parallel Genomic Sequence-Search on a Massively Parallel System", *ACM International Conference on Computing Frontiers*, May 2007. (Collaboration with IBM and Virginia Tech)
- Sudharshan Vazhkudai and Xiaosong Ma, "Recovering Transient Data: Automated On-demand Data Reconstruction and Offloading for Supercomputers", to appear, *Operating Systems Review (OSR)*, Special Issue on File and Storage Systems, Jan 2007. (Collaboration with ORNL)
- Nagiza F Samatova, Marcia Branstetter, Auroop R Ganguly, Robert Hettich, Shiraj Khan, Guruprasad Kora, Jiangtian Li, Xiaosong Ma, Chongle Pan, Arie

Shoshani and Srikanth Yoginath, “High performance statistical computing with parallel R: applications to biology and climate modeling”, *Journal of Physics: Conference Series*, Volume 46, 2006. (Collaboration with ORNL)

- Sudharshan Vazhkudai, Xiaosong Ma, Vincent Freeh, Jonathan Strickland, Nandan Tammineedi and Stephen Scott, "Constructing Collaborative Desktop Storage Caches for Large Scientific Datasets", *ACM Transactions on Storage Systems, 2(3*), August 2006. (Collaboration with ORNL)
- Mark K. Gardner, Jeremy Archuleta, Heshan Lin, Wu-chun Feng, and Xiaosong Ma, "A Continent-Spanning Computational Grid: Architectural Design and Practical Experience", *SC|06*, nominated for the best paper award. (Collaboration with LANL and ORNL)
- Sudharshan Vazhkudai, Douglas Thain, Xiaosong Ma and Vincent Freeh, “Positioning Dynamic Storage Caches for Transient Data”, *the International Workshop on High Performance I/O Techniques and Deployment of Very Large Scale I/O Systems (HyperIO 2006)*. (Collaboration with ORNL)

Collaboration with DOE Laboratories

Almost all the aforementioned projects and publications are through collaborations with DOE research labs, especially ORNL (where the PI is a joint faculty) and LANL. In the last period of the project, our collaboration expanded to LLNL. The PI continues to send students to ORNL for summer internships and all PhD students working on related projects have ORNL researchers on their thesis committees.

Student Progress

The most senior student currently supported by the grant, Heshan Lin, has successfully graduated in April 2009 and joined Virginia Tech as a research scientist. Heshan will continue to work on related projects under the supervision of our collaborator Wu-chun Feng. Another PhD student partially supported by this grant, Jiangtian Li, graduated in January 2009 and has joined Microsoft.