HESHAN LIN

Dept of Computer Science, North Carolina State Univ
EMAIL: hlin2@ncsu.edu
PHONE: (919) 559-5770
WEB: www4.ncsu.edu/~hlin2

EDUCATION

- PhD, Computer Science, North Carolina State University (Expected Mar 2009)
- MS, Computer Science, Temple University, 2004
- BS, Applied Math, South China University of Tech, China, 1998

RESEARCH INTERESTS

My research broadly falls into the general areas of data-intensive parallel and distributed computing. Specifically, I focused on the following areas: 1) Parallel I/O optimizations for petascale bioinformatics applications. 2) Efficient request scheduling for clustered scientific web services. 3) Data-intensive parallel computing on opportunistic desktop grids.

AWARDS AND HONORS

- Distinguished Paper Award, International Supercomputing Conference (ISC), 2008.
- Student Travel Award, TCPP PhD Forum, IEEE International Parallel and Distributed Processing Symposium (IPDPS), 2008.
- HPC Storage Challenge Award (Team Member), IEEE/ACM International Conference for High-Performance Computing, Networking, Storage and Analysis (SC), 2007.
- Best Paper Award Finalist, IEEE/ACM International Conference for High-Performance Computing, Networking, Storage and Analysis (SC), 2006

EXPERIENCES

North Carolina State University
Department of Computer Science
Raleigh, NC

(01/2005-Present) Research Assistant to Dr. Xiaosong Ma

- Proposed, designed and implemented an asynchronous, two-phase parallel I/O technique for petascale bioinformatics applications. The resulting prototype enabled a parallel sequence-matching tool to scale to 32,768 cores with 93% parallel efficiency on the IBM Blue Gene/P supercomputer.
- Researched efficient requests scheduling for data-intensive scientific web services. Work proposed novel adaptive scheduling algorithms that comprehensively exploit parallel computation efficiency and data locality to deliver high-performance scientific processing service on clustered web servers.
- Collaborated with scientists from Argonne National Lab and Virginia Tech to build ParaMEDIC, a semantics-based distributed I/O framework that enables distributed data-intensive supercomputing worldwide.
- Researched building parallel scientific web servers from idle desktop resources. Work explored the challenges and opportunities of aggregating idle CPU resources and unused storage spaces into a dynamic, autonomic web server specialized in scientific data processing.

Virginia Tech
Systems, Networking & Renaissance Grokking Laboratory
Blacksburg, VA

(06/2008-09/2008, 05/2007-08/2007) Summer Intern (Mentor: Dr. Wu-chun Feng)

- Researched providing highly reliable MapReduce services on opportunistic environments. The prototype system extended Hadoop and enabled MapReduce applications to efficiently run on cycle-stealing desktop grids.
- Researched performance optimization for massively parallel genomic sequence search. Work proposed a novel scheduling algorithm that highly improves system throughput by gracefully integrating dynamic load balancing and asynchronous I/O.

Los Alamos National Laboratory
Research & Development in Advanced Network Technology
Los Alamos, NM

(05/2005-08/2005) Summer Intern (Mentor: Dr. Wu-chun Feng)

- Researched, designed and implemented the parallel I/O extension of mpiBLAST, a DOE R&D 100 award-winning bioinformatics application.
- Collaboratively designed and built the GreenGene ad-hoc grid, which aggregated 3000+ processors from distributed supercomputers to discover similarity structures of 6 million genomic sequences (SC|05 StorCloud Demo).

Oak Ridge National Laboratory
Computer Science and Mathematics Division
Oak Ridge, TN

(05/2004-08/2004) Summer Intern (Mentor: Dr. Nagiza F. Samatova)

- Proposed, designed and implemented a novel parallel I/O solution for parallel BLAST (a popular bioinformatics pattern-matching tool). The resulting prototype delivered an order-of-magnitude performance improvement to a state-of-the-art BLAST parallelization.

Parallel Computers Technology Inc.
Philadelphia, PA

(01/2003-12/2003) Software Engineer Intern

- Major designer and developer of ICX-UDS, a middleware product that provides real-time replication for clustered database servers (based on US patent 6,421,688). Designed and implemented core system modules with C/C++ on Linux.
- Proposed, designed and implemented a solution for the multi-server login problem, which allows a database client to login multiple database servers simultaneously without altering either the client or the server APIs. The solution expanded the supporting database systems of ICX-UDS from MS/SQL server to Oracle and MySQL.

PUBLICATIONS

Refereed Conference Paper

- H. Lin, P. Balaji, R. Poole, C. Sosa, X. Ma and W. Feng, "Massively Parallel Genomic Sequence Search on the Blue Gene/P Architecture," *IEEE/ACM International Conference for High-Performance Computing, Networking, Storage and Analysis (SC),* 2008.
- P. Balaji, W. Feng and H. Lin, "Semantics-based Distributed I/O with the ParaMEDIC Framework," *ACM International Symposium on High Performance Distributed Computing (HPDC)*, 2008.

- P. Balaji, W. Feng, H. Lin, J. Archuleta, S. Matsuoka, A. Warren, J. Setubal, E. Lusk, R. Thakur, I. Foster, D. S. Katz, S. Jha, K. Shinpaugh, S. Coghlan and D. Reed, "Distributed I/O with ParaMEDIC: Experiences with a Worldwide Supercomputer," *International Supercomputing Conference (ISC)*, **Distinguished Paper Award**, 2008.
- H. Lin, X. Ma, J. Li, T. Yu and N. Samatova, "Adaptive Request Scheduling for Parallel Scientific Web Services," *International Conference on Scientific and Statistical Database Management (SSDBM),* 2008.
- O. Thorsen, K. Jiang, A. Peters, B. Smith, H. Lin, W. Feng and C. Sosa, "Parallel Genomic Sequence-Search on a Massively Parallel System," *ACM International Conference on Computing Frontiers*, 2007.
- M. Gardner, W. Feng, J. Archuleta, H. Lin and X. Ma, "Parallel Genomic Sequence-Searching on an Ad-Hoc Grid: Experiences, Lessons Learned, and Implications," *IEEE/ACM International Conference for High-Performance Computing, Networking, Storage and Analysis (SC),* **Best Paper Finalist**, 2006.
- A. Ching, W. Feng, H. Lin, X. Ma and A. Choudhary, "Exploring I/O Strategies for Parallel Sequence Database Search Tools with S3aSim," *ACM International Symposium on High Performance Distributed Computing (HPDC), 2006*.
- H. Lin, X. Ma, P. Chandramohan, A. Geist and N. Samatova, "Efficient Data Access for Parallel BLAST," *IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2005.

Other Publications

- P. Balaji, W. Feng, J. Archuleta, H. Lin, R. Kettimuttu, R. Thakur and X. Ma, "Semantics-based Distributed I/O for mpiBLAST (short paper)," *In the ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP)*, 2008.
- H. Lin, X. Ma, J. Li, T. Yu and N. Samatova, "Processor and data scheduling for online parallel sequence database servers." *In Technical Report TR-2007-23. NCSU*, 2007.

In Preparation

- H. Lin, W. Feng, X. Ma and J. Archuleta, "MapReduce on Opportunistic Desktop Environments," *In Preparation.*
- H. Lin, X. Ma, W. Feng and N. Samatova, "An Integrated Approach of Computation and I/O Scheduling for Massively Parallel Genomic Sequence Search," *In Preparation.*

INVITED TALKS

- "Massively Parallel Genomic Sequence Search on Blue Gene/P," IEEE/ACM International Conference for High-Performance Computing, Networking, Storage and Analysis (SC), Austin, TX, 2008
- "Scalable I/O Solutions for High-Performance Genomic Sequence Search," Symposium for Graduate Student Research at NCSU, Raleigh, NC, 2008.
- "Efficient Data Handling in Large-Scale Sequence Database Search," SIAM Conference on Parallel Processing for Scientific Computing, San Francisco, CA, 2006.
- "Internal Design of pioBLAST – A Highly Efficient Parallel BLAST," Los Alamos National Laboratory, Los Alamos, NM, 2005.
- "Efficient Data Access for Parallel BLAST," 19th International Parallel and Distributed Processing Symposium (IPDPS), Denver, CO, 2005.

PROFESSIONAL ACTIVITIES

- Served as the core researcher and developer of mpiBLAST, an open-source parallel pattern-matching application for bioinformatics (http://www.mpiblast.org).
- Served as a reviewer for the First International Workshop on Parallel Programming

Models and Systems Software for High-End Computing (P2S2).

- Co-reviewed papers for a number of conferences: ICDCS'08, SC|08, SC|06, NAS'08, ICPP'08, ICPP'07, ICPP'06, IPDPS'06, PCDN'06.
- Attended the HPC Storage Challenge competition at SC|07 with project "ParaMEDIC: Parallel Metadata Environment for Distributed I/O and Computing."
- Attended the StorCloud demo at SC|05 with project "mpiBLAST on the GreenGene Distributed Supercomputer."

SKILLS

- Experienced in Unix/Linux system programming with C/C++.
- Experienced in parallel programming with MPI, familiar with OpenMP.
- Experienced in designing parallel I/O with MPI-IO and POSIX I/O.
- Knowledgeable in designing and trouble-shooting large-scale parallel/distributed software systems.
- Skillful in software development with various languages (Java, Visual Basic, SQL, RPG).

REFERENCES

Dr. Xiaosong Ma (Ph.D. Advisor) – Email: ma@cs.ncsu.edu, Tel: (919) 513-7577
Assistant Professor, Department of Computer Science, North Carolina State University
Joint Faculty, Computer Science and Mathematics, Oak Ridge National Laboratory

Dr. Wuchun Feng – Email: feng@cs.vt.edu, Tel: (540) 231-1192
Associate Professor, Department of Computer Science, Department of Electrical & Computer Engineering, Virginia Tech

Dr. Nagiza Samatova – Email: samatova@csc.ncsu.edu, Tel: (919) 513-7575
Associate Professor, Department of Computer Science, North Carolina State University
Joint Faculty, Computer Science and Mathematics, Oak Ridge National Laboratory

Dr. Pavan Balaji – Email: balaji@mcs.anl.gov, Tel: (630) 252-3017, Fax: (630) 252-5986
Assistant Scientist, Mathematics and Computer Science Division, Argonne National Laboratory