

Zhe Zhang

Research Staff Member
IBM T. J. Watson Research Center

19 Skyline Drive
Hawthorne, NY 10532-1596
(914) 784-6079

Adjunct Assistant Professor
North Carolina State University

zhe_zhang@ncsu.edu
<https://researcher.ibm.com/researcher/view.php?person=us-zhezhang>

RESEARCH INTERESTS My research mainly focuses on the area of *distributed computing systems*, with specific interests in system fault tolerance and efficient data storage and I/O.

EDUCATION **NC State University**, Raleigh, NC 2009

Ph.D. in Computer Science and Operations Research

Advisor: Dr. Xiaosong Ma

Dissertation Topic:

Adding Coordination to HEC Storage Stacks for Efficiency and Reliability

NC State University, Raleigh, NC 2006

Master of Science in Operations Research

Advisor: Dr. David F. McAllister

Thesis: *Application of Linear Optimization in Stereo Image Rendering*

University of Science and Technology of China, Hefei, China 2003

Bachelor of Engineering in Computer Science

PROFESSIONAL APPOINTMENTS **IBM T.J. Watson Research Center**, Hawthorne, NY Nov. 2010~present
Research Staff Member

NC State University, Raleigh, NC Jul. 2010~present
Adjunct Assistant Professor

Oak Ridge National Laboratory, Oak Ridge, TN Nov. 2009~Nov. 2010
Research Staff Member

Microsoft Research, Cambridge, UK Jul.~Sep. 2009
Research Intern

IBM T.J. Watson Research Center, Hawthorne, NY May~Aug. 2008
Research Intern

Cisco Systems, San Jose, CA May~Aug. 2007
Software Engineering Intern, Service Routing Group

Oak Ridge National Laboratory, Oak Ridge, TN Jun.~Aug. 2006
Research Intern, Distributed Systems Research Group

PUBLICATIONS **Peer-reviewed Conference and Journal Publications**

1. [PDSW '10] "Workload Characterization of a Leadership Class Storage" by Youngjae Kim, Raghul Gunasekaran, Galen Shipman, David Dillow, *Zhe Zhang*, and Bradley Settlemyer. Published in the Proceedings of the The 5th DoE SciDAC Petascale Data Storage Workshop (PDSW), New Orleans, LA, Nov. 2010.

2. [SC '10] “A CCSM-based Terrestrial Ecosystem Model Parameterization Framework” by Dali Wang, Daniel Ricciuto, Zhe Zhang, Haihang You, and Wilfred Post. Appeared in the poster session of *ACM/IEEE Supercomputing (SC)*, New Orleans, LA, Nov. 2010.
3. [CUG '10] “Lessons Learned in Deploying the Worlds Largest Scale Lustre File System” by Galen Shipman, David Dillow, Sarp Oral, Feiyi Wang, Douglas Fuller, Jason Hill and Zhe Zhang. Published in the Proceedings of *The 52nd Cray User Group Conference (CUG)*, 2010.
4. [HPDC '10] “MOON: MapReduce on Opportunistic Environments” by Heshan Lin, Xiaosong Ma, Jeremy Archuleta, Wu-chun Feng, Mark Gardner and Zhe Zhang. Published in the Proceedings of *ACM International Symposium on High Performance Distributed Computing (HPDC)*, Chicago, IL, Jun. 2010 (**25% acceptance rate**).
5. [ICDCS '10] “A Hybrid Approach to High Availability in Stream Processing Systems” by Zhe Zhang, Yu Gu, Fan Ye, Hao Yang, Minkyong Kim, Hui Lei and Zhen Liu. Published in the Proceedings of *IEEE International Conference on Distributed Computing Systems (ICDCS)*, Genoa, Italy, Jun. 2010 (**14% acceptance rate**).
6. [Middleware '09] “An Empirical Study of High Availability in Stream Processing Systems” by Yu Gu, Zhe Zhang, Fan Ye, Hao Yang, Minkyong Kim, Hui Lei and Zhen Liu. Published in the Proceedings of *ACM/IFIP/USENIX International Middleware Conference (Middleware)*, Industrial Track, Urbana Champaign, Illinois, Dec. 2009.
7. [JGC '09] “Improving Data Availability for Better Access Performance: A Study on Caching Scientific Data on Distributed Desktop Workstations” by Xiaosong Ma, Sudharshan Vazhkudai, and Zhe Zhang. Published in the *Journal of Grid Computing (JGC)*, Special Issue on Volunteer Computing and Desktop Grids, 2009.
8. [ISC '09] “Improving the Availability of Supercomputer Job Input Data Using Temporal Replication” by Chao Wang, Zhe Zhang, Sudharshan Vazhkudai, Xiaosong Ma and Frank Mueller. Published in the Proceedings of *International Supercomputing Conference (ISC)*, Hamburg, Germany, Jun. 2009.
9. [EuroSys '09] “Memory Resource Allocation for File System Prefetching – From a Supply Chain Management Perspective” by Zhe Zhang, Amit Kulkarni, Xiaosong Ma and Yuanyuan Zhou. Published in the Proceedings of *European Conference on Computer Systems (EuroSys)*, Nuremberg, Germany, Apr. 2009 (**17% acceptance rate**).
10. [ICPP '08] “On-the-fly Recovery of Job Input Data in Supercomputers” by Chao Wang, Zhe Zhang, Sudharshan Vazhkudai, Xiaosong Ma and Frank Mueller. Published in the Proceedings of *International Conference on Parallel Processing (ICPP)*, Portland, Oregon, Sep. 2008
11. [ICDCS '08] “PFC: Transparent Optimization of Existing Prefetching Strategies for Multi-level Storage Systems” by Zhe Zhang, Kyuhyung Lee, Xiaosong Ma and Yuanyuan Zhou. Published in the Proceedings of *IEEE International Conference on Distributed Computing Systems (ICDCS)*, Beijing, China, Jun. 2008 (**16% acceptance rate**).
12. [SC '07] “Optimizing Center Performance through Coordinated Data Staging, Scheduling and Recovery” by Zhe Zhang, Chao Wang, Sudharshan Vazhkudai, Xiaosong Ma, Gregory G. Pike, John W. Cobb and Frank Mueller. Published in the Proceedings of *ACM/IEEE Supercomputing (SC)*, Reno, NV, Nov. 2007 (**20% acceptance rate**).
13. [EI '06] “A Uniform Metric for Anaglyph Calculation” by Zhe Zhang and David F. McAllister. Published in the Proceedings of *IS&T/SPIE Electronic Imaging (EI)*, San Jose, CA, Jan. 2006.

Other Publications

14. [INFORMS '10] “Supply Chain Models and Heuristics for Data Cache Management” by Zhe Zhang, Xueping Li, Xiaoyan Zhu, Rui Xu, Xiaosong Ma, Galen Shipman. To

appear in the INFORMS (Institute for Operations Research and the Management Sciences) Annual Meeting, Austin, TX, Nov. 2010.

15. [MSR '10] *“Does erasure coding have a role to play in my data center?”* by Zhe Zhang, Amey Deshpande, Xiaosong Ma, Eno Thereska, and Dushyanth Narayanan. Published as Microsoft Research Technical Report MSR-TR-2010-52, 2010.
16. [ORNL '07] *“Improving Data Availability for Better Access Performance: A Study on Caching Scientific Data on Distributed Workstations”* by Xiaosong Ma, Zhe Zhang, and Sudharshan Vazhkudai. Published as ORNL Technical Report 007030, 2007.

RESEARCH PROJECTS

File System Buffer Cache Management: Innovatively observed the similarity between file system prefetching and supply chain management (SCM), and performed mapping of concepts between the two areas. Based on that theoretical analysis, proposed and designed novel mechanisms to measure data access rates and allocate file system memory space among prefetching streams accordingly. The prefetching algorithms are implemented in the Linux 2.6.18 kernel and are able to improve the throughput of mixed workload by up to 33%. Also proposed PreFetching Coordinator(PFC) to address the information distortion in multi-level memory cache management. Rather than being another new prefetching algorithm, PFC acts as a middleman between two adjacent levels of caching/prefetching and optimizes behaviors of existing algorithms. A simulator based study shows that PFC improves different prefetching algorithms including AMP, SARC, and the algorithm used by Linux 2.6 kernel by 14.6% on average.

Related publication : [INFORMS '10, EuroSys '09, ICDCS '08]

Adaptive Monitoring of Large Scale Distributed Systems: Developing adaptive and automatic methods to monitor application I/O behavior on Jaguar, which is currently the world's fastest computer system. Observations from the monitoring framework will improve the performance and scalability of scientific applications as well as mitigate performance issues in center-wide shared file system Spider, which is one of the world's largest scale parallel file systems with 10 Petabytes of storage capacity.

Related publication : [PDSW '10, SC '10, CUG '10]

High Availability in MapReduce Systems Led the effort to develop an erasure coding library for the Hadoop framework which performs MapReduce tasks. The framework is the first implementation of online erasure coding storage for Hadoop, which saves both network bandwidth and storage capacity. We have also provided to the community a comprehensive and empirical evaluation of erasure coding in the data-intensive distributed computing environment. Also collaboratively researched a hybrid architecture and multiple novel techniques to enable Hadoop to run MapReduce applications on resource-scavenging grids with high reliability.

Related publications : [MSR'10, HPDC '10]

High Availability in Overlay Messaging and Streaming Systems: Proposed and designed a hybrid failure tolerance method for providing high availability to overlay messaging and streaming systems, which combines advantages of existing active standby and passive standby methods and allows users to flexibly choose desirable positions in the overhead-delay tradeoff. Also made major contribution in building a software system for evaluating and analyzing different fault tolerance techniques for overlay systems. The system is has been intensively tested on IBM's System S cluster, and has a prospective deployment on larger platforms of IBM's customers.

Related publications : [Middleware '09, ICDCS '10]

Fault Tolerance in Distributed Storage Systems: Developed a prototype data staging/reconstructing manager based on the Moab job scheduler and the Lustre parallel file system. Our prototype allows HPC users to specify data operations in their job scripts, and uses job scheduling information to achieve just-in-time data staging. We also made modifications to the Lustre parallel file system to transparently patch data from remote sources and reconstruct files from partial data loss. With a proposed deployment on ORNL's Jaguar system, the prototype has been tested on multiple clusters at ORNL and NCSU.
Related publications : [ISC '09, ICPP '08, SC '07]

Distributed Service Routing: Improved the fault tolerance and performance of a Distributed Hash Table(DHT) based service routing system by adding backup bootstrap points to accept new nodes joining the network.

Distributed Storage Scavenging: Designed and implemented an on-line cache management algorithm for *FreeLoader*, a novel distributed storage scavenging file system. FreeLoader aggregates underutilized desktop storage space to provide a shared cache/scratch space for large, immutable data sets. Based on the access patterns of this kind of files, the algorithm assigns priorities to data blocks based on both their access recencies and logical offsets.
Related publications : [JGC '09, ORNL '07]

Stereo Image Rendering: Parallelized a computation-intensive anaglyph calculation algorithm. Also applied sensitivity analysis theory of linear optimization and designed a depth-first traversing technique to process areas of similar colors in real world images at low cost(workload reduced by 72% to 89%).
Related publication : [EI '06]

SYNERGISTIC
ACTIVITIES

Invited Reviewer

- Invited as a reviewer for Journal of Parallel and Distributed Computing (JPDC), Aug. 2010.
- Invited as a reviewer for Future Generation Computer Systems (FGCS) – The International Journal of Grid Computing and eScience, Jul. 2010.
- Invited by U.S. Department of Energy as a reviewer for Small Business Innovation Research (DOE SBIR), Feb. 2010.
- Invited as a reviewer for Journal of Grid Computing (JGC), special issue on Volunteer Computing and Desktop Grids, Apr. 2009.

External Reviewer

- The IEEE International Conference on Parallel & Distributed Systems (ICPADS '10)
- The ACM SOSP Workshop on Hot Topics in Storage & File Systems (HotStorage '09)
- The IEEE International Parallel & Distributed Processing Symposium (IPDPS '09)
- The IEEE International Conference on Distributed Computing Systems (ICDCS '08)
- The IACC International Conference on Parallel Processing (ICPP '08)
- The ACM Statistical and Scientific Database Management Conference (SSDBM '08)
- The IEEE International Conference on Networking, Architecture, & Storage (NAS '07)
- The ACM/IEEE Supercomputing Conference (SC '06, '08)

TEACHING

Teaching Assistant, NC State University (Aug. 2004~May 2006)

In the undergraduate courses *Numerical Analysis (CSC 302)*, *Computational Theory (CSC 333)* and *File Organization (CSC 431)*, assisted instructors by holding office hours, grading homework and exams, and giving lectures.

Guest Lecturer, NC State University (Fall 2009)

In the graduate course *Parallel Systems (CSC 548)*, lectured the topic of parallel I/O and file systems.

HONORS &
MEMBER-
SHIPS:

Phi Kappa Phi: Selected as top 10% graduate students, 2004~2007.

ACM: Student member since 2007 and professional member since 2009.

IEEE: Professional member since 2010.

SIAM: Student member since 2004.

Outstanding Student Scholarship: Selected as top 12% students in USTC, 2002.

VISA STATUS

Citizen of China. H-1B Visa.