

Distributed Storage Systems for Extreme-Scale Data-Intensive Computing

Ioan Raicu^{1,2}

¹Department of Computer Science, Illinois Institute of Technology

²Mathematics and Computer Science Division, Argonne National Laboratory

Abstract

State-of-the-art yet decades-old architecture of HPC storage systems has segregated compute and storage resources, bringing unprecedented inefficiencies and bottlenecks at petascale levels and beyond. This work presents FusionFS and ZHT, two new distributed storage systems designed from the ground up for high scalability (8K-nodes) while achieving significantly higher I/O performance (1TB/sec) and operations per second (18M/sec). FusionFS and ZHT will achieve these levels of scalability and performance through complete decentralization, and the co-location of storage and compute resources. FusionFS currently supports POSIX-like interfaces important for ease of adoption and backwards compatibility with legacy applications. ZHT has a simple yet functional NoSQL key/value datastore interface to remove unneeded overheads and limitations inherited from POSIX. Both systems are made reliable through data replication with strong and weak consistency semantics, while FusionFS also supports information dispersal algorithms. FusionFS supports scalable data provenance capture and querying, a much needed feature in large scale scientific computing systems towards achieving reproducible and verifiable experiments. Both systems have been deployed on a variety of testbeds, ranging from a 32-node (256-cores) Linux cluster, to a 96-VM virtual cluster on the Amazon EC2 cloud, to a 8K-node (32K-cores) IBM BlueGene/P supercomputer with promising results, when compared to other leading distributed storage systems such as GPFS, PVFS, HDFS, S3, Casandra, Memcached, and DynamoDB. The long term goals of FusionFS and ZHT are to scale them to exascale levels with millions of nodes, billions of cores, petabytes per second I/O rates, and billions of operations per second.


Cyber-Infrastructure Used

Current Infrastructure Usage—TeraScale to PetaScale

- Dell Linux Cluster @ IIT (512-cores, SSDs/HDD per node)
- SiCortex@ANL (5832-cores SiCortex SC5832)
- Beacon@NICS (54-nodes, 0.2PFLOP/s)
- Kodiak@LANL (1K-nodes)
- Intrepid@ANL (40K-nodes IBM BG/P, 160K-cores, 0.5PFLOP/s)
- Stampede@TACC (~5PFLOP/s Dell w/ Intel MICs)
- BlueWaters@NCSA (~10PFLOP/s Cray XE6)

Infrastructures to be used in the Future

- Jaguar@ORNL (~3PFLOP/s Cray XK6)
- Mira@ANL (~9PFLOP/s IBM BlueGene/Q)
- Titan@ORNL (~18PFLOP/s Cray XK7)



CAREER Award Research Area

Research Directions

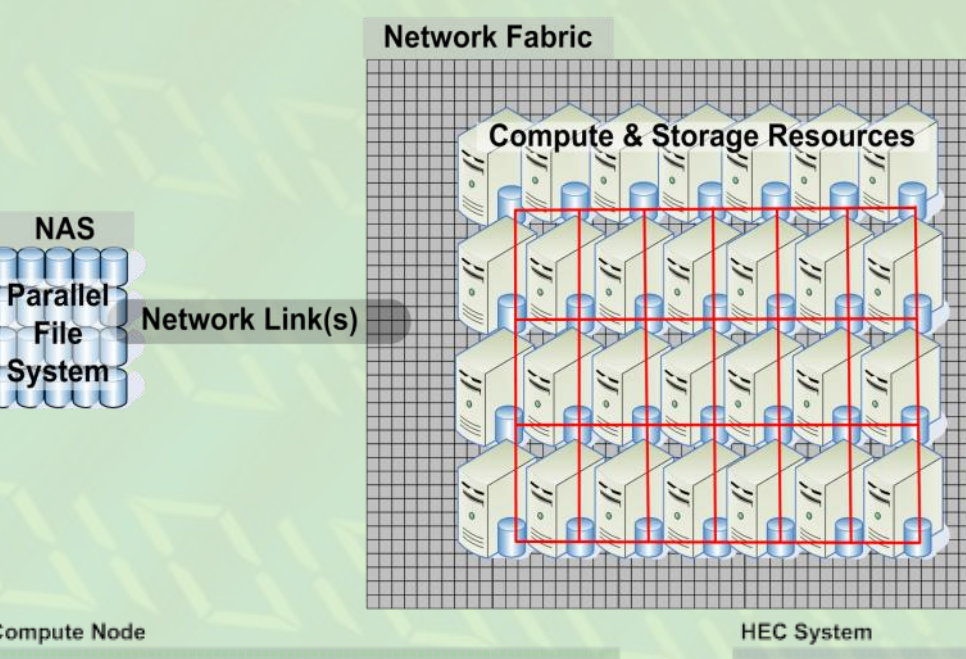
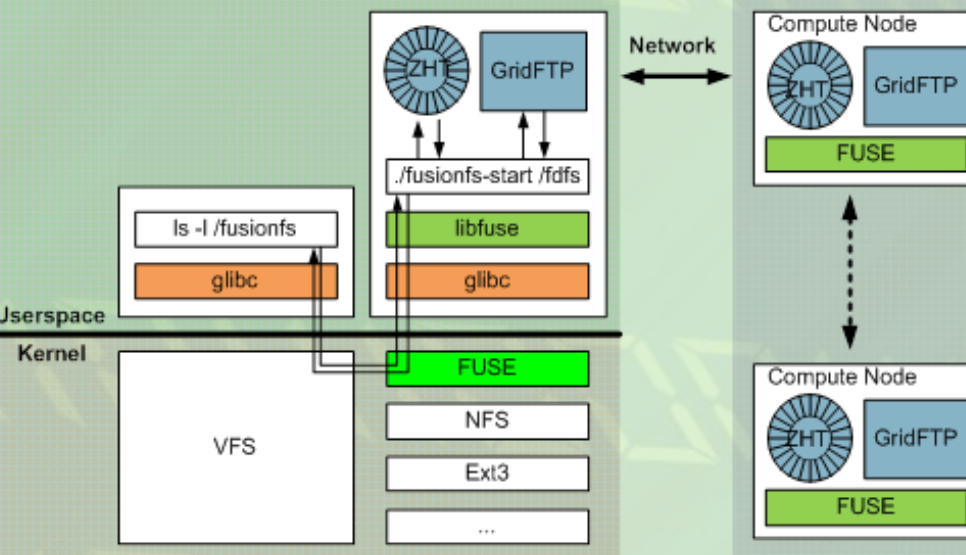
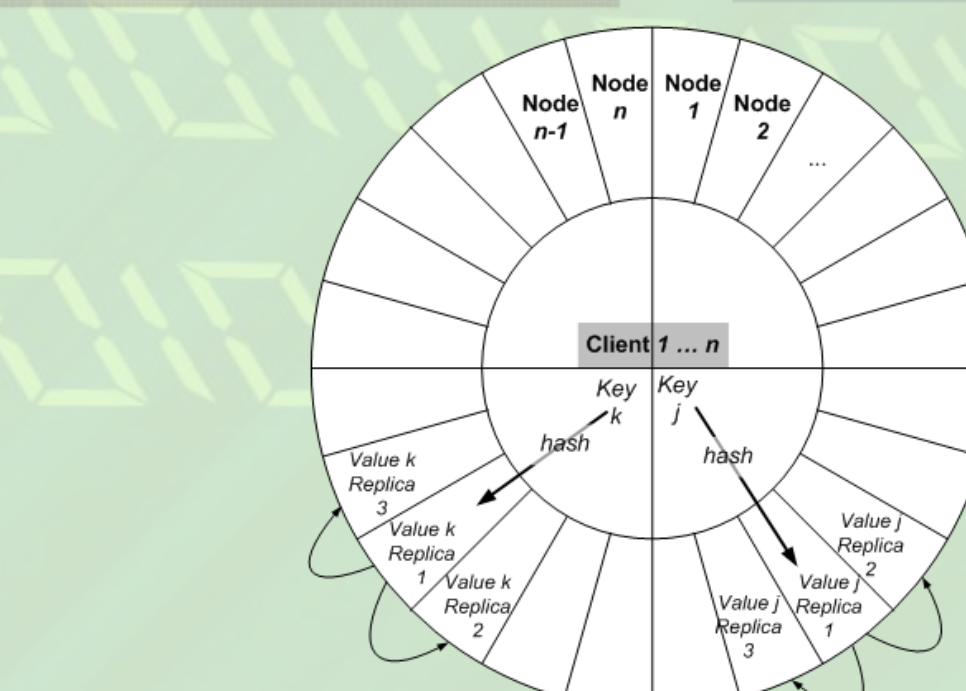
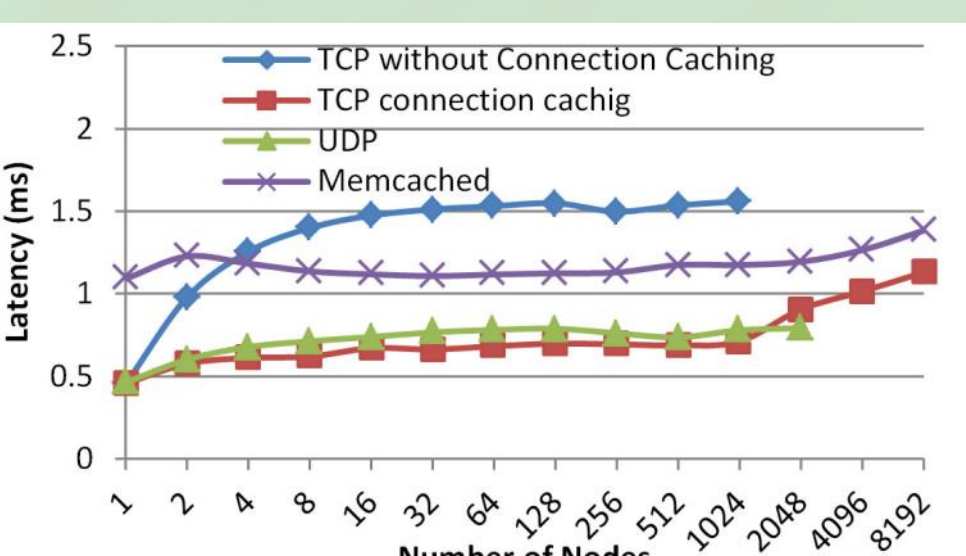
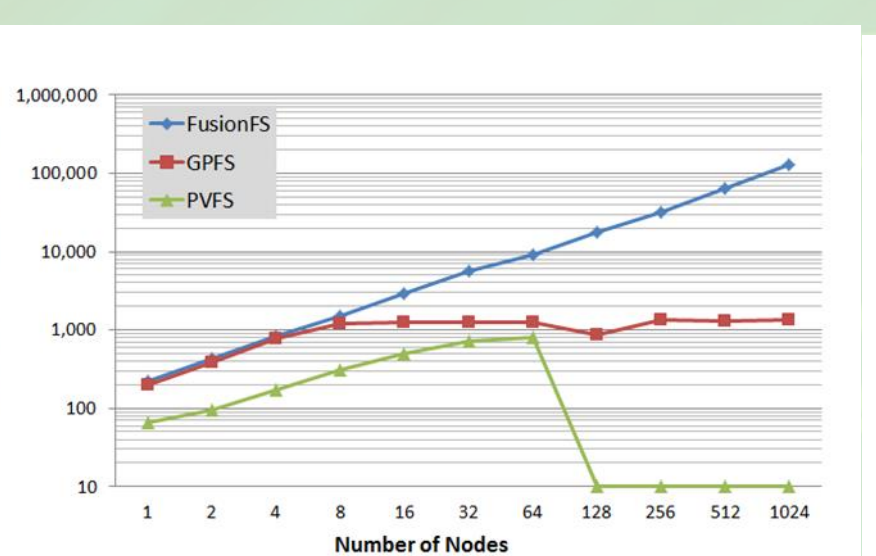
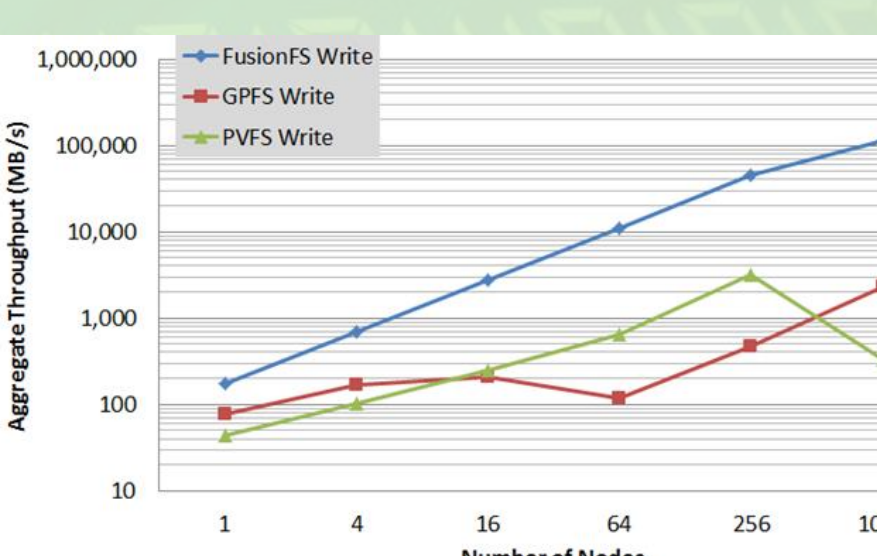
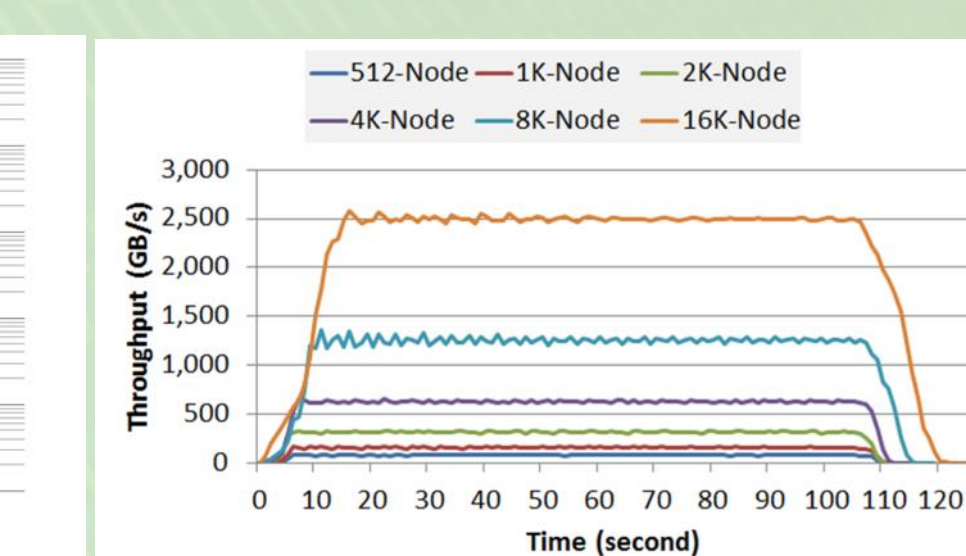
- Decentralization is critical**
 - Computational resource management (e.g. LRMs)
 - Storage systems (e.g. parallel/distributed file systems)
- Data locality must be maximized, while preserving I/O interfaces**
 - POSIX I/O on shared/parallel file systems ignore locality
 - Data-aware scheduling coupled with distributed file systems that expose locality is the key to scalability over the next decade

FusionFS: Fusion Distributed File System

- Distributed Metadata and Management
- Data Indexing
- Relaxed Semantics
- Data Locality
- Overlapping I/O with Computations
- POSIX
- Provenance Support
- Reliable & Efficient through Information Dispersal Algorithms

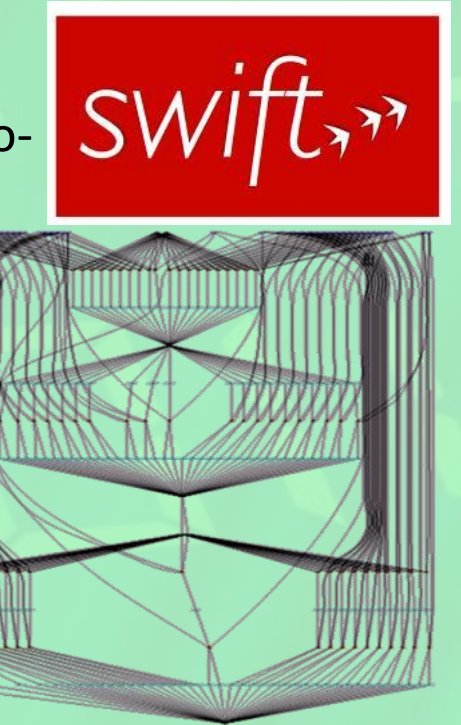
ZHT: Zero-Hop Distributed Hash Table

- Simplified distributed hash table tuned for the specific requirements of HEC
 - Emphasized key features of HEC are:** Trustworthy/reliable hardware, fast network interconnects, non-existent node "churn", low latencies requirements, and scientific computing data-access patterns
- Primary goals:** Excellent availability and fault tolerance, with low latencies
- ZHT details:** Static/Dynamic membership function, Network topology aware node ID space, Replication and Caching, Efficient 1-to-all communication through spanning trees, Persistence (NoVoHT)

Future Research Work

- Complete prototype of FusionFS (<1 year)
- Leverage ZHT to other projects: Swift, MosaStore, MATRIX, GlobusOnline (0~2 years)
- Scale ZHT and FusionFS to 10PFlops/s systems, such as Mira, Stampede, and Bluewaters (1~2 years)
- Work closely with the Swift parallel programming system to evaluate the impact of FusionFS and ZHT for a wide array of Many-Task Computing applications at petascale levels (0~3 years)
 - Explore data-aware scheduling to improve real application performance at petascale levels
- Explore extensions to FusionFS through loosely connected projects (0~5 years):
 - Adding provenance support at the filesystem level
 - Improving price/performance ratios through hybrid SSD+HDD caching (HyCache)
 - Improve storage efficiency through information dispersal algorithms
 - Understand the applicability of FusionFS/ZHT for cloud computing



Field	Description	Characteristics	Status
Astronomy	Creation of mosaics from many digital images	Many 1-core tasks, much communication, complex dependencies	Experimental
Astronomy	Stacking of catalogs from digital sky surveys	Many 1-core tasks, much communication	Experimental
Biochemistry	Analysis of mass-spectrometer data for post-translational protein modifications	10,000-100 million jobs for proteomic searches using custom serial codes	Operational
Biochemistry	Protein structure prediction using iterative fitting algorithms exploring other biomolecular interactions	Hundreds to thousands of 1- to 1,000-core simulations and data analysis	Operational
Biochemistry	Identification of drug targets via computational docking/screening	Up to 1 million 1-core docking operations	Operational
Bioinformatics	Management modeling	Thousands of 1-core integer programming problems	In development
Business sciences	Mining of large text corpora to study media bias	Analysis and comparison of over 70 million text files of news articles	In development
Climate science	Essential climate model runs and analysis of output data	Tens to hundreds of 100- to 1,000-core simulations	Experimental
Economics	Generation of response surfaces for various economic models	1,000 to 1 million 1-core runs (10,000 typical), then data analysis	Operational
Neuroscience	Analysis of functional MRI datasets	Comparison of images, connectivity analysis with structural equation modeling, 100,000+ tasks	Operational
Radiology	Training of computer-aided diagnosis algorithms	Comparison of images, many tasks, much communication	In development
Radiology	Image processing and brain mapping for neuro-surgical planning research	Execution of MPI application in parallel	In development

Missing Cyber-Infrastructure Biography

Formal proposal process to gain access to NSF funded cyberinfrastructure

- Getting significant time on large supercomputers is non-trivial for systems research
- DOE has the INCITE awards, but they primarily fund applications research
- Discretionary allocations on large systems are generally small and limited, and require close collaborations with researchers at the respective laboratory





Dr. Ioan Raicu is an assistant professor in the Department of Computer Science (CS) at Illinois Institute of Technology (IIT), as well as a guest research faculty in the Math and Computer Science Division (MCS) at Argonne National Laboratory (ANL). He is also the founder (2011) and director of the Data-Intensive Distributed Systems Laboratory (DataSys) at IIT. He has received the prestigious NSF CAREER award (2011 - 2015) for his innovative work on distributed file systems for exascale computing. He was a NSF/CRA Computation Innovation Fellow at Northwestern University in 2009 - 2010, and obtained his Ph.D. in Computer Science from University of Chicago under the guidance of Dr. Ian Foster in March 2009. He is a 3-year award winner of the GSRP Fellowship from NASA Ames Research Center. His research work and interests are in the general area of distributed systems. His work focuses on a relatively new paradigm of Many-Task Computing (MTC), which aims to bridge the gap between two predominant paradigms from distributed systems, High-Throughput Computing (HTC) and High-Performance Computing (HPC). His work has focused on defining and exploring both the theory and practical aspects of realizing MTC across a wide range of large-scale distributed systems. He is particularly interested in resource management in large scale distributed systems with a focus on many-task computing, data intensive computing, cloud computing, grid computing, and many-core computing. Over the past decade, he has co-authored 86 peer reviewed articles, book chapters, books, theses, and dissertations, which received over 3250 citations, with a H-index of 22. His work has been funded by the NASA Ames Research Center, DOE Office of Advanced Scientific Computing Research, the NSF/CRA CIFellows program, and the NSF CAREER program. He has also founded and chaired several workshops, such as ACM Workshop on Many-Task Computing on Grids and Supercomputers (MTAGS), the IEEE Int. Workshop on Data-Intensive Computing in the Clouds (DataCloud), and the ACM Workshop on Scientific Cloud Computing (ScienceCloud). He is on the editorial board of the IEEE Transaction on Cloud Computing (TCC), the Springer Journal of Cloud Computing Advances, Systems and Applications (JoCCASA), as well as a guest editor for the IEEE Transactions on Parallel and Distributed Systems (TPDS), the Scientific Programming Journal (SPJ), and the Journal of Grid Computing (JoGC). He has been leadership roles in several high profile conferences, such as HPDC, CCGrid, Grid, eScience, and ICAC. He is a member of the IEEE and ACM. More information can be found at <http://www.cs.iit.edu/~iraicu/>, <http://datasys.cs.iit.edu/>, and <http://www.linkedin.com/in/ioanraicu>.

Collaborations

- The ACI CAREER Workshop is a great start**
 - Running this annually will greatly enhance this program
 - It should drive awareness of our research work and spark collaborations
- Running a BoF, workshop, or meeting for ACI CAREER recipients at IEEE/ACM Supercomputing conference**
 - This could be used to have both recipients and students funded by these ACI CAREER awards to present their latest results
 - NSF Program Officers could also attend to get more interaction with the recipients, their work, and their results
- Mentoring system where senior ACI CAREER recipients work with junior recipients**
- This work deals with large-scale storage systems, helping make compute-intensive systems also suitable for data-intensive systems (covering both traditional POSIX based file systems and NOSQL storage systems)**
 - Interested in collaborations with people looking to scaling up their data-intensive applications

Educational Activities

- Mentored students:**
- 3 highschool girls
 - 5 undergraduates
 - 7 master students
 - 4 PhD students
- Introduce new courses:**
- Introduction to Distributed Computing (CS495)
 - Data-Intensive Computing (CS554)
 - Cloud Computing (CS553)
- Organized Workshops:**
- IEEE/ACM MTAGS 2011/2012/2013 at Supercomputing
 - ACM ScienceCloud 2011/2013 at ACM HPDC
 - IEEE/ACM DataCloud 2011/2012 at IPDPS/Supercomputing
- Editor of Journal Special Issues**
- Journal of Grid Computing, SI on Data Intensive Computing in the Clouds, 2011
 - Scientific Programming Journal, SI on Science-driven Cloud Computing, 2011
 - IEEE Transactions on Parallel and Distributed Systems, SI on Many-Task Computing, 2011
 - IEEE Transactions on Cloud Computing, SI on Scientific Cloud Computing, 2014
- 
- 

References

Major Publications:

- Ke Wang, Abhishek Kulkarni, Dorian Arnold, Michael Lang, Ioan Raicu. "Using Simulation to Explore Distributed Key-Value Stores for Exascale Systems Services", IEEE/ACM Supercomputing/SC 2013
- Tonglin Li, Xiaobing Zhou, Kevin Brandstatter, Dongfang Zhao, Ke Wang, Anupam Rajendran, Zhao Zhang, Ioan Raicu. "ZHT: A Light-weight Reliable Persistent Dynamic Scalable Zero-hop Distributed Hash Table", IEEE International Parallel & Distributed Processing Symposium (IPDPS) 2013
- Ke Wang, Kevin Brandstatter, Ioan Raicu. "SimMatrix: Simulator for Many-Task computing execution fabRIC at exascales", ACM HPC 2013
- Dongfang Zhao, Da Zhang, Ke Wang, Ioan Raicu. "Exploring Reliability of Exascale Systems through Simulations", ACM HPC 2013
- Chen Shou, Dongfang Zhao, Tanu Malik, Ioan Raicu. "Towards a Provenance-Aware Distributed File System", USENIX TaPP13
- Ke Wang, Zhangjie Ma, Ioan Raicu. "Modelling Many-Task Computing Workloads on a Petaflop IBM BlueGene/P Supercomputer", IEEE CloudFlow 2013
- Dongfang Zhao, Ioan Raicu. "HyCache: A User-Level Caching Middleware for Distributed File Systems", IEEE HPDC 2013
- Yong Zhao, Ioan Raicu, Shiyong Lu, Xubo Fei. "Opportunities and Challenges in Running Scientific Workflows on the Cloud", IEEE International Conference on Network-based Distributed Computing and Knowledge Discovery (CyberC) 2011
- Ioan Raicu, Pete Beckman, Ian Foster. "Making a Case for Distributed File Systems at Exascale", Invited Paper, ACM Workshop on Large-scale System and Application Performance (LSAP), 2011

Publications Under Review:

- Chen Shou, Dongfang Zhao, Tanu Malik, Ioan Raicu. "Distributed Data Provenance for Large-Scale Data-Intensive Computing", under review at IEEE Cluster 2013
- Dongfang Zhao, Corentin Debains, Pedro Alvarez-Tabio, Kent Burlingame, Ioan Raicu. "Towards High-Performance and Cost-Effective Distributed Storage Systems with Information Dispersal Algorithms", under review at IEEE Cluster 2013

This work is supported in part by the National Science Foundation grant NSF-1054974.