Towards Data-Intensive Extreme-Scale Computing

Ioan Raicu[1,2]

[1]Department of Computer Science, Illinois Institute of Technology

[2]Mathematics and Computer Science Division, Argonne National Laboratory

Abstract

State-of-the-art yet decades old architecture of high-performance computing systems has its computation and storage separated. It has shown limits for today's data-intensive applications, because every I/O needs to be transferred via the network between the computation and storage cliques. This work aims design, implement, and evaluate a new distributed storage systems for extreme scale data-intensive computing. We proposed a distributed storage layer local to the compute nodes, which is responsible for most of the I/O operations and saves extreme amount of data movement between compute and storage resources. We have designed and implemented a distributed file system FusionFS for HPC compute nodes to support metadata-intensive and write-intensive operations. It supports a variety of data-access semantics, from POSIX-like interfaces for generality, to relaxed semantics for increased scalability. FusionFS has numerous advanced features to improve performance (e.g. caching and compression), improve reliability (e.g. replication and erasure codes), and improve functionality (e.g. provenance capture and query). FusionFS has been deployed and evaluated on up to 16K compute nodes in an IBM BlueGene/P supercomputer, showing orders of magnitude improvement in metadata and I/O performance. We have compared FusionFS with other leading distributed storage systems such as GPFS, PVFS, HDFS, S3, Casandra, Memcached, and DynamoDB – and FusionFS has always come out ahead in either performance, functionality, or both. We have also done a detailed performance evaluation with various scientific applications. An extensive evaluation of FusionFS was performed through simulations showing near linear scalability up to two million nodes. The long term goals of FusionFS is to scale it to exascale levels with millions of nodes, billions of cores, petabytes per second I/O rates, and billions of operations per second – with real systems, accelerating real data-intensive scientific applications at extreme scales.

Cyber-Infrastructure Used

Current Infrastructure Usage—TeraScale to PetaScale

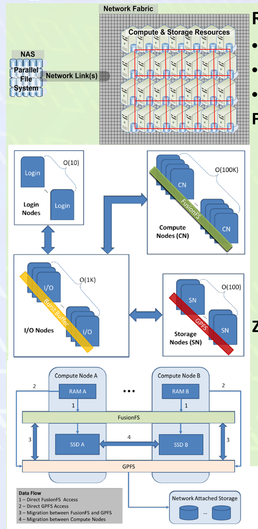
- Dell Linux Cluster @ IIT (512-cores, SSDs/HDD per node)
- SiCortex@ANL (5832-cores SiCortex SC5832)
- Beacon@NICS (54-nodes, 0.2PFLOP/s)
- Kodiak@LANL (1K-nodes)
- Intrepid@ANL (40K-nodes IBM BG/P, 160K-cores, 0.5PFLOP/s)
- Stampede@TACC (~5PFLOP/s Dell w/ Intel MICs)
- BlueWaters@NCSA (~10PFLOP/s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

XSEDE
Extreme Science and Engineering Discovery Environment

CAREER Award Research Area



Research Directions

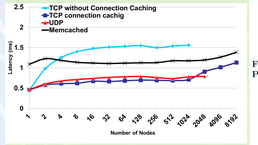
- **Decentralization is critical**
- **Data locality must be maximized, while preserving I/O interfaces**
- **Storage systems will need to become more distributed to scale ==> Critical for resilience and scalability of HPC systems**

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)



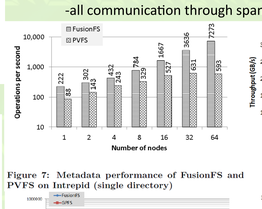
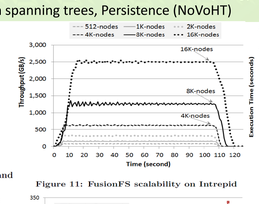
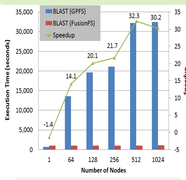
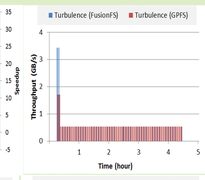


Figure 7: Metadata performance of FusionFS and PVFS on Intrepid (single directory)







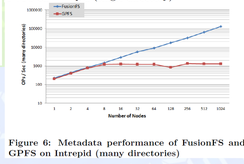
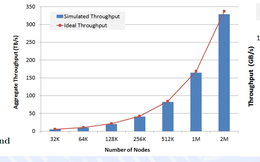
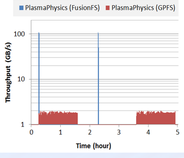


Figure 11: FusionFS scalability on Intrepid





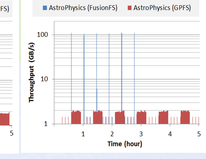
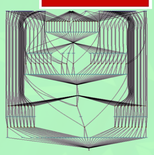


Figure 6: Metadata performance of FusionFS and GPFS on Intrepid (many directories)

Future Research Work

- Scale ZHT and FusionFS to 10PFlops/s systems, such as Mira, Stampede, and Bluewaters
- 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
  - Explore data-aware scheduling to improve real application performance at petascale levels
- Explore extensions to FusionFS through loosely connected projects:
  - Adding provenance support at the filesystem level
  - Improving price/performance ratios through hybrid SSD+HDD caching (HyCache+)
  - Improve storage efficiency through information dispersal algorithms
  - Reduce I/O requirements through novel compression techniques
  - Understand the applicability of FusionFS/ZHT for cloud computing
  - Use simulations (ROSS+CODES) to study the FusionFS architecture at extreme-scales



Field	Description	Characteristics	Status
Astronomy	Creation of montages from many digital images	Many 1-core tasks, much communication, complex dependencies	Experimental
Astronomy	Stacking of cutouts 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	In development
Biochemistry*	Protein structure prediction using iterative fixing algorithm; 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*	Metagenome modeling	Thousands of 1-core integer programming problems	In development
Business economics	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	Ensemble 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 neurosurgical planning research	Execution of MPI application in parallel	In development

Note: Asterisks indicate applications being run on Argonne National Laboratory's Blue Gene/P (Intrepid) and/or the TeraGrid Sun Constellation at the University of Texas at Austin (Ranger).

Missing Cyber-Infrastructure

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

Biography

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 is also the recipient of the IIT Junior Faculty Research Award in 2013. 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 over 100 peer reviewed articles, book chapters, books, theses, and dissertations, which received over 4576 citations, with a H-index of 27. 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 Clouds, 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), and the Springer Cluster Computing Journal (Cluster). He has been leadership roles in several high profile conferences, such as HPDC, CCGrid, Grid, eScience, Cluster, and ICAC. He is a member of the IEEE and ACM. More information can be found at http://www.cs.iit.edu/~iraicu/.

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
- 6 undergraduates
- 7 master students
- 4 PhD students

Introduce new courses:

- Introduction to Parallel & Distributed Computing (CS451)
- Data-Intensive Computing (CS554)
- Cloud Computing (CS553)

Organized Workshops:

- IEEE/ACM MTAGS 2011/2012/2013/2014 at Supercomputing
- ACM ScienceCloud 2011/2013/2014 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:

- Dongfang Zhao, Kan Qiao, Ioan Raicu. "HyCache+: Towards Scalable High-Performance Caching Middleware for Parallel File Systems", IEEE/ACM CCGrid 2014
- Dongfang Zhao, Chen Shou, Tanu Malik, Ioan Raicu. "Distributed Data Provenance for Large-Scale Data-Intensive Computing", 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", IEEE Cluster 2013
- Tonglin Li, Ioan Raicu, Lavanya Ramakrishnan. "Scalable State Management for Scientific Applications in the Cloud", IEEE BigData 2014
- 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 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 a Distributed File System", USENIX TaPP13
- Ke Wang, Zhangjie Ma, Ioan Raicu. "Modeling 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 HPDIC 2013
- Dongfang Zhao, Jian Yin, Ioan Raicu. "Improving the I/O Throughput for Data-Intensive Scientific Applications with Efficient Compression Mechanisms", IEEE/ACM Supercomputing 2013
- Dongfang Zhao, Ioan Raicu. "Distributed File Systems for Exascale Computing", IEEE/ACM Supercomputing/SC 2012
- Yong Zhao, Ioan Raicu, Shiyong Lu, Xubo Fei. "Opportunities and Challenges in Running Scientific Workflows on the Cloud", IEEE CyberC 2011
- Ioan Raicu, Pete Beckman, Ian Foster. "Making a Case for Distributed File Systems at Exascale", ACM LSAP 2011

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