Building Blocks for Scalable Distributed Storage Systems

Ioan Raicu[1,2]

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

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

Abstract

Exascale computers will enable the unraveling of significant scientific mysteries. Predictions are that 2019 will be the year of exascale, with millions of compute nodes and billions of threads of execution. The current architecture of high-end computing systems is decades-old and has persisted as we scaled from gigascales to petascales. In this architecture, storage is completely segregated from the compute resources and are connected via a network interconnect. This approach will not scale several orders of magnitude in terms of concurrency and throughput, and will thus prevent the move from petascale to exascale. At exascale, basic functionality at high concurrency levels will suffer poor performance, and combined with system mean-time-to-failure in hours, will lead to a performance collapse for large-scale heroic applications. Storage has the potential to be the Achilles heel of exascale systems. We propose that future high-end computing systems be designed with non-volatile memory on every compute node, allowing every compute node to actively participate in the metadata and data management and leveraging many-core processors high bisection bandwidth in torus networks. More specifically, this work aims to architect and develop a zero-hop distributed hash table (ZHT), which has been tuned for the requirements of high-end computing systems. ZHT aims to be a building block for future distributed file systems (e.g. FusionFS) to implement distributed metadata management. This work will be evaluated on real workloads on real pre-exascale systems (Cray, IBM, and Sun supercomputers from ANL, NCSA, and ORNL, as well as XSEDE), as well as through simulations at exascales. This work has also been a catalyst in several other storage related projects exploring building blocks for scalable storage systems, such as Hybrid SSD+HHD file systems (HyCache), Persistent Key/Value Stores (NoVoHT), Provenance Enabled Distributed File Systems (PAFS), Increasing Storage Efficiency through Information Dispersal Algorithms (IDA), and understanding reliability through checkpointing (SimHEC). This work will also open doors for further research in programming paradigm shifts (e.g. Many-Task Computing) needed as we approach exascales, but ones that require a significantly more scalable storage infrastructure if it is to be successful at exascales. Work is already underway to better understand the possibility of scaling Many-Task Computing to exascale levels through novel work stealing algorithms (SimMatrix and MATRIX). This revolutionary new distributed storage architecture will make exascale computing more tractable, touching virtually all disciplines in high-end computing and fueling scientific discovery for many years.

Cyber-Infrastructure Used

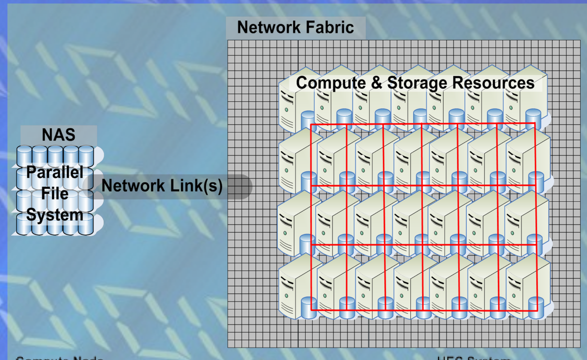
Current Infrastructure Usage—TeraScale to PetaScale

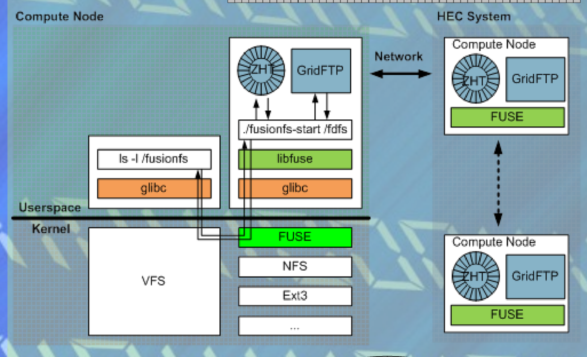
- Dell Linux Cluster at IIT
  - 64-nodes, 512-cores, SSDs and HDD deployed at each node
- SiCortex SC5832 at ANL
  - 972-nodes, 5832-cores
- IBM Blue Gene/P supercomputer at ANL (aka Intrepid)
  - 40K-nodes, 160K-cores, 0.5PFLOP/s

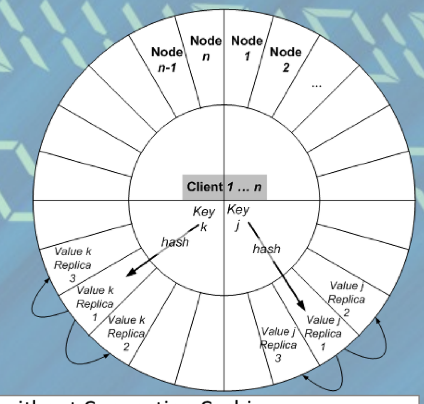
Infrastructures to be used in the Future

- Jaguar (~3PFLOP/s Cray XK6 at ORNL)
- BlueWaters (~10PFLOP/s Cray XE6 at NCSA)
- Mira (~20PFLOP/s IBM BlueGene/Q at ANL)
- XSEDE (formerly TeraGrid, 16 supercomputers in the US)

CAREER Award Research Area







Research Directions

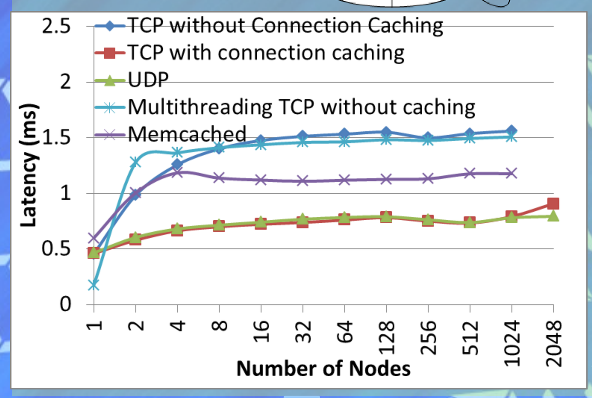
- **Decentralization is critical**
  - Computational resource management (e.g. LRMs)
  - Storage systems (e.g. parallel/distributed file systems)
- **Data locality must be maximized, while preserving I/O interfaces**
  - POSIX I/O on shared/parallel file systems ignore locality
  - Data-aware scheduling 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

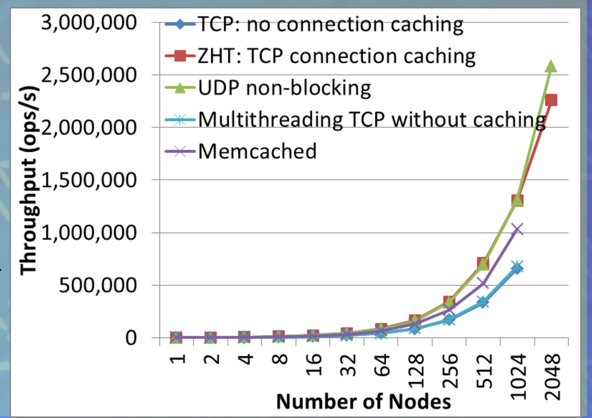
ZHT: Zero-Hop Distributed Hash Table

- 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)



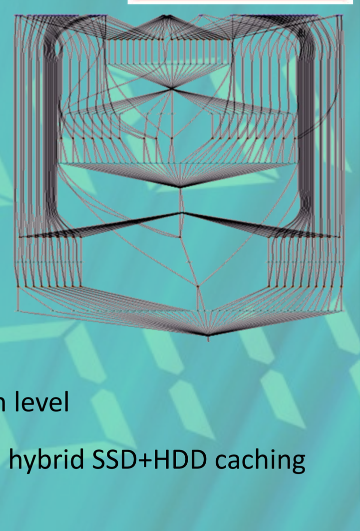
Performance

- More than 2.5M operations/sec aggregated throughput and latencies of less than 1ms with 2K-nodes.
- UDP shows a better scalability; however, TCP can be as fast as UDP when using connection caching.
- The performance differences among three basic operations (insert, lookup and remove) are insignificant.
- ZHT uses a direct 0-hop algorithm (via consistent hashing), with the majority of the overhead coming from network communication.



Future Research Work

- Complete prototype of FusionFS (<1 year)
- Leverage ZHT to other projects: Swift, MosaStore, MATRIX, GlobusOnline (0~3 years)
- Scale ZHT and FusionFS to 10PFlops/s systems, such as Mira BG/Q (1~3 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 extensions to FusionFS through various 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 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 neuro-surgical 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 hard 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 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 50 peer reviewed articles, book chapters, books, theses, and dissertations, which received over 2100 citations. His H-index is 19 , G-Index is 45, and E-Index is 37. 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/DataCloud-SC), and the ACM Workshop on Scientific Cloud Computing (ScienceCloud). He is on the editorial board of 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/.

Collaborations

- **The OCI 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 OCI CAREER recipients at IEEE/ACM Supercomputing conference**
  - This could be used to have both recipients and students funded by these OCI 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 OCI 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
- 3 undergraduates
- 7 master students
- 4 PhD students

Introduce new courses:

- Introduction to Distributed Systems (CS495)
- Data-Intensive Computing (CS554)
- Cloud Computing (CS553)

Organized Workshops:

- ACM MTAGS 2011 at Supercomputing
- IEEE DataCloud 2011 at IEEE IPDPS
- ACM DataCloud-SC 2011 at Supercomputing
- ACM ScienceCloud 2011 at ACM HPDC

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

References

Published:

- 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
- Tonglin Li, Hui Jin, Antonio Perez De Tejada, Kevin Brandstatter, Zhao Zhang, Ioan Raicu. "ZHT: Zero-Hop Distributed Hash Table", 1st Greater Chicago Area System Research Workshop, 2012
- Ke Wang, Kevin Brandstatter, Ioan Raicu. "SimMatrix: Simulator for MAny-Task computing execution fabRIc at eXascales", 1st Greater Chicago Area System Research Workshop, 2012
- Corentin Debains, Pedro Manuel Alvarez-tabio Togores, Ioan Raicu. "Evaluating Information Dispersal Algorithms", 1st Greater Chicago Area System Research Workshop, 2012
- Dongfang Zhao, Ioan Raicu. "HyCache: A Hybrid User-Level File System with SSD Caching", 1st Greater Chicago Area System Research Workshop, 2012
- Iman Sadooghi, Dongfang Zhao, Tonglin Li, Ioan Raicu. "Understanding the Cost of Cloud Computing and Storage", 1st Greater Chicago Area System Research Workshop, 2012
- Da Zhang, Ioan Raicu. "SimHEC: Simulator for High-End Computing Systems", 1st Greater Chicago Area System Research Workshop, 2012
- Tonglin Li, Raman Verma, Xi Duan, Hui Jin, Ioan Raicu. "Exploring Distributed Hash Tables in High-End Computing", SIGMETRICS Performance Evaluation Review-Measurement and Evaluation, 2011

Under Review/Preparation:

- Tonglin Li, Hui Jin, Antonio Perez De Tejada, Kevin Brandstatter, Zhao Zhang, Ioan Raicu. "ZHT: A Zero-Hop Distributed Hash Table for High-End Computing", under review at IEEE/ACM SC 2012
- Ke Wang, Zhao Zhang, Ioan Raicu. "Extreme Scale Distributed Load-Balancing with Adaptive Work Stealing", under review at IEEE/ACM SC 2012
- Dongfang Zhao, Ioan Raicu. "HyCache: A Hybrid User-Level File System with SSD Caching", under review at IEEE/ACM SC12
- Ke Wang, Kevin Brandstatter, Ioan Raicu. "SimMatrix: Simulator for MAny-Task computing execution fabRIc at eXascales", under preparation
- Kevin Brandstatter, Ioan Raicu. "NoVoHT: Non-Volatile Hash Table", under preparation

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