BigDataX 2015: Distributed Storage System Projects

Ioan Raicu, Illinois Institute of Technology

Extracting knowledge from increasingly large data sets produced, both experimentally and computationally, continues to be a significant challenge for scientific discovery. Experimental fields, such as high-energy physics report that experimental data sets are expected to grow by six orders of magnitude or so in coming years. For computational fields, such as fusion science, in which energy codes run on a million cores (available today), data will be output in bursts of an astounding 2 petabytes/sec with checkpoints every 10 minutes, producing an average of 3.5 terabytes/sec over the entire run of an experiment. At exascales, these burst and average I/O rates would be three orders of magnitude higher.

The requirements to support knowledge discovery at extreme scales, combined with the flop-to-I/O imbalances on today's petascale machines, is increasingly driving the need for more heterogeneous *in situ* computing. With *in situ* computing, a subset of the nodes (or perhaps all the nodes) on the HPC platform are organized differently to act as I/O servers and embedded data analysis engines. To support *in situ* computing at extreme scales, two trends are becoming extremely important:

1) It is critical that storage systems be close to compute nodes to manage large-scale I/O volume and data movement, as well as to optimize I/O performance automatically
2) It is equally important that storage systems be integrated with some capability to capture provenance data so as to provide a valuable means of performing post-hock analysis and perform verification, validation, and retracing after the fact has occurred

Unfortunately in today's petascale platform's not only is storage disconnected from compute, storage systems are not well equipped to capture or store provenance for later analysis. This disconnect has been considered a serious impediment in moving from the petascale to the exascale and to support efficient analysis. This work addresses heterogeneous in-situ computing achieved by a tighter integration of storage systems with provenance data systems, with the goal of delivering a large-scale distributed file system that supports provenance capture transparently. Collocation of the storage systems with compute nodes is bound to improve the I/O efficiency at extreme scales. However, recording the innumerable states and artifacts within a provenance system adds to significant I/O and storage burden. There is a need to make provenance a first class citizen by making storage systems natively compliant with provenance.

To achieve this vision, this work will extend the co-PI's NSF funded project FusionFS [31] by adding distributed provenance data management system [28, 32]. FusionFS is a new distributed file system that will co-exist with current parallel file systems in High-End Computing, optimized for both a subset of HPC and Many-Task Computing workloads. FusionFS is a user-level file system that runs on the compute resource infrastructure, and enables every compute node to actively participate in the metadata and data management. Distributed metadata management is implemented using ZHT [6], a zero-hop distributed hash table. ZHT has been tuned for the specific requirements of high-end computing (e.g. trustworthy/reliable hardware, fast networks, non-existent "churn", low latencies, and scientific computing data-access patterns). The data is partitioned and spread out over many nodes based on the data access patterns. Replication is used to ensure data availability as well as erasure coding algorithms, and cooperative caching delivers high aggregate throughput. Data is indexed, by including descriptive, provenance, and system metadata on each file. Data can be compressed to deliver higher storage efficiencies and in certain cases improved performance [5]. FusionFS supports a variety of data-access semantics, from POSIX-like interfaces for generality, to relaxed semantics for increased scalability.

This work will also leverage existing prior work on the SPADE provenance system, but with a focus on scalability, performance, and overheads minimization. A preliminary analysis has shown that the manifestation of distributed concepts in FusionFS and SPADE are similar and if leveraged appropriately, can provide the necessary provenance support in storage systems. In particular, FusionFS is a user-level file system with a POSIX-like interface that runs on the compute source infrastructure and enables every compute node to actively participate in the metadata data management. SPADE is a decentralized data provenance management system that has modular components for gathering, integrating, filtering, storing, and querying data provenance within the user-level file system. However, SPADE currently cannot scale to millions of nodes, and its implementation does not lend itself well to high-end computing systems. This work will build upon prior work with the SPADE [36] provenance system, the FusionFS [31] distributed file system, and the FusionProv [28, 32] provenance enabled distributed file system, and extend the work towards the design and implementation of a distributed provenance-aware storage systems that could scale to today's largest systems, as well as tomorrow's extreme scale systems to millions of nodes and billions of cores.

Leveraging years of work in distributed file systems [5 - 35] and provenance capture and query, projects surrounding this work would entail making improvements to a variety of critical components such as metadata management, I/O access pattern coalescing, distributed graph databases to help scale the distributed provenance queries, as well as exploring novel interfaces into distributed storage systems. ***These projects are suitable for undergraduate students as they generally target specific modules in a distributed file system, often times limited to only several thousand lines of code. Once modifications are made to the file system layers, benchmarking evaluations are relatively straight forward for a non-expert to complete, and to compare against competing techniques.***

1 Mentor

Dr. **Ioan Raicu** (PI) is an assistant professor in CS at Illinois Institute of Technology (IIT), as well as a guest research faculty in 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 received his PhD in 2009 from University of Chicago. His research work and interests are in the general area of distributed systems, with particular interests in resource management in large scale distributed systems focusing on many-task computing [1, 2], data intensive computing, cloud computing, grid computing, many-core computing, and big data. He teaches both undergraduate and graduate courses, from introduction of parallel and distributed computing, to advanced operating systems, cloud computing, and data-intensive computing. ***He has worked extensively with undergraduate students and published multiple (7 papers [4 - 10]) peer reviewed conference papers at some of the top conferences, such as IEEE/ACM Supercomputing/SC, IEEE IPDPS, ACM HPDC, EuroSys, XSEDE, and ACM HPC. He has also published many short papers (10 papers [11 - 20]) at local workshops. 8 of the undergraduate students have attended international conferences such as IEEE/ACM Supercomputing/SC, ACM HPDC, IEEE CCGrid, and XSEDE with four more attending local workshops such as GCASR.*** Overall, he has co-authored over 100 peer reviewed articles, which received over 4881 citations, with a H-index of 28. His work has been funded by NASA, DOE, and NSF. He is the recipient of numerous awards, such as the IEEE TCSC Young Achievers in Scalable Computing award (2014), IEEE/ACM CCGrid Outstanding Service Award (2014), IIT Junior Faculty Research Award (2013), NSF CAREER Award (2011), NSF/CRA Computation Innovation Fellow award (2009), and the NASA GSRP Fellowship award (2006). He has also founded and chaired several workshops, such as IEEE/ACM MTAGS, IEEE/ACM DataCloud, ACM ScienceCloud, and IEEE CASK. He is on the editorial board of IEEE TCC, Springer JoCCASA, and Springer Cluster. He has been leadership roles in several high profile conferences, such as HPDC, CCGrid, Grid, eScience, Cluster, ICAC, and BDC.

2 References

[1] I. Raicu, I. Foster, Y. Zhao. "Many-Task Computing for Grids and Supercomputers", Invited Paper, IEEE Workshop on Many-Task Computing on Grids and Supercomputers (MTAGS08), 2008

[2] I. Raicu. "Many-Task Computing: Bridging the Gap between High Throughput Computing and High Performance Computing", Doctorate Dissertation, Computer Science Department, University of Chicago, March 2009

[3] Scott J. Krieder, Justin M. Wozniak, Timothy Armstrong, Michael Wilde, Daniel S. Katz, Benjamin Grimmer, Ian T. Foster, Ioan Raicu. "Design and Evaluation of the GeMTC Framework for GPU-enabled Many-Task Computing", ACM HPDC 2014

[4] Kevin Brandstatter, Jason DiBabbo, Daniel Gordon, Ben Walters, Alex Ballmer, Lauren Ribordy, Ioan Raicu. "Delivering 3.5 Double Precision GFlops/Watt and 200Gb/sec Bi-Section Bandwidth with Intel Xeon Phi-based Cisco Servers", Student Cluster Competition (SCC), IEEE/ACM Supercomputing/SC 2014

[5] Dongfang Zhao, Jian Yin, Kan Qiao, Ioan Raicu. "Virtual Chunks: On Supporting Random Accesses to Scientific Data in Compressible Storage Systems", IEEE International Conference on Big Data 2014

[6] 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

[7] Benjamin Grimmer, Scott Krieder, Ioan Raicu. "Enabling Dynamic Memory Management Support for MTC on NVIDIA GPUs", EuroSys 2013

[8] Ke Wang, Kevin Brandstatter, Ioan Raicu. “SimMatrix: Simulator for MAny-Task computing execution fabRIc at eXascales”, ACM HPC 2013

[9] Ke Wang, Anupam Rajendran, Kevin Brandstatter, Zhao Zhang, Ioan Raicu. “Paving the Road to Exascale with Many-Task Computing”, Doctoral Showcase, IEEE/ACM Supercomputing/SC 2012

[10] Scott Krieder, Ben Grimmer, Ioan Raicu. “Early Experiences in running Many-Task Computing workloads on GPGPUs”, XSEDE 2012

[11] Ben Walters, Scott Krieder, Ioan Raicu. "A Survey of State-of-the-Art GPU Profilers", 3rd Greater Chicago Area System Research Workshop (GCASR), 2014

[12] Dustin Shahidehpour, Scott Krieder, Ben Grimmer, Ioan Raicu. "Accelerating Scientific Workflow Applications with GPUs", 2nd Greater Chicago Area System Research Workshop (GCASR), 2013

[13] Tonglin Li, Xiaobing Zhou, Kevin Brandstatter, Ioan Raicu. "Distributed Kev-Value Store on HPC and Cloud Systems", 2nd Greater Chicago Area System Research Workshop (GCASR), 2013

[14] Ben Grimmer, Scott Krieder, Ioan Raicu. "Enabling Dynamic Memory Management Support for MTC on NVIDIA GPUs", 2nd Greater Chicago Area System Research Workshop (GCASR), 2013

[15] Scott J. Krieder, Benjamin Grimmer, Dustin Shahidehpour, Jeffrey Johnson, Justin M. Wozniaky, Michael Wildeyz, Ioan Raicu. "Towards Efficient Many-Task Computing on Accelerators in High-End Computing Systems", 2nd Greater Chicago Area System Research Workshop (GCASR), 2013

[16] Kevin Brandstatter, Tonglin LI, Xiaobing Zhou, Ioan Raicu. "NoVoHT: a Lightweight Dynamic Persistent NoSQL Key/Value Store", 2nd Greater Chicago Area System Research Workshop (GCASR), 2013

[17] Jeff Johnson, Scott Krieder, Benjamin Grimmer, Justin Wozniak, Michael Wilde, Ioan Raicu. "Understanding the Costs of Many-Task Computing Workloads on Intel Xeon Phi Coprocessors", 2nd Greater Chicago Area System Research Workshop (GCASR), 2013

[18] Kevin Brandstatter, Ioan Raicu. “CiteSearcher: A Google Scholar frontend for Mobile Devices", Illinois Institute of Technology Research Day, 2012

[19] Tonglin Li, Antonio Perez De Tejada, Kevin Brandstatter, Zhao Zhang, Ioan Raicu. “ZHT: a Zero-hop DHT for High-End Computing Environment”, 1st Greater Chicago Area System Research Workshop, 2012

[20] Ke Wang, Kevin Brandstatter, Ioan Raicu. “SimMatrix: Simulator for MAny-Task computing execution fabRIc at eXascales”, 1st Greater Chicago Area System Research Workshop, 2012

[21] I. Raicu, P. Beckman, I. Foster. “Making a Case for Distributed File Systems at Exascale”, ACM Workshop on Large-scale System and Application Performance (LSAP), 2011

[22] T. Li, R. Verma, X. Duan, H. Jin, I. Raicu. “Exploring Distributed Hash Tables in High-End Computing”, ACM Performance Evaluation Review (PER), 2011

[23] I. Raicu. “Building Blocks for Scalable Distributed Storage Systems”, NSF CyberBridges Workshop, 2012

[24] T. 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

[25] D. Zhao, Ioan Raicu. “HyCache: A Hybrid User-Level File System with SSD Caching”, 1st Greater Chicago Area System Research Workshop, 2012

[26] D. Zhang, Ioan Raicu. "SimHEC: Simulator for High-End Computing Systems”, 1st Greater Chicago Area System Research Workshop, 2012

[27] Dongfang Zhao, Ioan Raicu. “HyCache: A User-Level Caching Middleware for Distributed File Systems”, IEEE HPDIC 2013

[28] Chen Shou, Dongfang Zhao, Tanu Malik, Ioan Raicu. “Towards a Provenance-Aware a Distributed File System”, USENIX TaPP13

[29] Dongfang Zhao, Da Zhang, Ke Wang, Ioan Raicu. “RXSim: Exploring Reliability of Exascale Systems through Simulations”, ACM HPC 2013

[30] Ke Wang, Abhishek Kulkarni, Michael Lang, Dorian Arnold, Ioan Raicu. “Using Simulation to Explore Distributed Key-Value Stores for Extreme-Scale Systems Services”, IEEE/ACM Supercomputing/SC 2013

[31] Dongfang Zhao, Zhao Zhang, Xiaobing Zhou, Tonglin Li, Ke Wang, Dries Kimpe, Philip Carns, Robert Ross, and Ioan Raicu. "FusionFS: Towards Supporting Data-Intensive Scientific Applications on Extreme-Scale High-Performance Computing Systems", IEEE International Conference on Big Data 2014

[32] Dongfang Zhao, Chen Shou, Tanu Malik, Ioan Raicu. “Distributed Data Provenance for Large-Scale Data-Intensive Computing”, IEEE Cluster 2013

[33] Dongfang Zhao, Kent Burlingame, Corentin Debains, Pedro Alvarez-Tabio, Ioan Raicu. “Towards High-Performance and Cost-Effective Distributed Storage Systems with Information Dispersal Algorithms”, IEEE Cluster 2013

[34] Dongfang Zhao, Kan Qiao, and Ioan Raicu. Hycache+: Towards scalable high-performance caching middleware for parallel file systems. In *IEEE/ACM CCGrid ’14*, 2014.

[35] Ke Wang, Anupam Rajendran, Ioan Raicu. MATRIX: Many-Task Computing Execution Fabric for Extreme Scales, Illinois Institute of Technology, Department of Computer Science, Technical Report, 2013.

[36] Gehani, Ashish, and Dawood Tariq. "SPADE: Support for provenance auditing in distributed environments." Proceedings of the 13th International Middleware Conference. Springer-Verlag New York, Inc., 2012.