Distributed Storage Systems for Extreme-Scale Data-Intensive Computing

Ioan Raicu
Illinois Institute of Technology
Argonne National Laboratory
iraicu@cs.iit.edu

Abstract

This work aims to re-architect distributed storage systems towards extreme scale data-intensive computing by co-locating storage and compute resources, and distributing both metadata and storage operations to all compute nodes at scales up to millions of nodes and billions of cores.

Overview

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. This work has produced several publications [1, 2, 3, 4, 5, 6, 7, 8, 9].

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 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 (MTAGS, DataCloud, ScienceCloud), and has been on the editorial board of several journals (TCC, JoCCASA, TPDS, SPJ, JoGC). He has been leadership roles in several high profile conferences (HPDC, CCGrid, Grid, eScience, ICAC). He is a member of the IEEE and ACM. More information can be found at http://www.cs.iit.edu/~iraicu/.

Acknowledgements

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

References

[1] 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

[2] 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

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

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

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

[6] Ke Wang, Zhangjie Ma, Ioan Raicu. “Modeling Many-Task Computing Workloads on a Petaflop IBM BlueGene/P Supercomputer”, IEEE CloudFlow 2013

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

[8] 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

[9] 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