Towards Data-Intensive Extreme-Scale Computing

Ioan Raicu

Illinois Institute of Technology

Argonne National Laboratory

iraicu@cs.iit.edu

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. This work has produced several publications [1-15].

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/.

Acknowledgements

References

[1] Dongfang Zhao, Kan Qiao, Ioan Raicu. "HyCache+: Towards Scalable High-Performance Caching Middleware for Parallel File Systems", IEEE/ACM CCGrid 2014

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

[3] 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

[4] Tonglin Li, Ioan Raicu, Lavanya Ramakrishnan. "Scalable State Management for Scientific Applications in the Cloud", IEEE BigData 2014

[5] 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

[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 IPDPS 2013

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

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

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

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

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

[12] Dongfang Zhao, Jian Yin, Ioan Raicu. "Improving the I/O Throughput for Data-Intensive Scientific Applications with Efficient Compression Mechanisms", IEEE/ACM Supercomputing 2013

[13] Dongfang Zhao, Ioan Raicu. "Distributed File Systems for Exascale Computing", IEEE/ACM Supercomputing/SC 2012

[14] Yong Zhao, Ioan Raicu, Shiyong Lu, Xubo Fei. "Opportunities and Challenges in Running Scientific Workflows on the Cloud", IEEE CyberC 2011

[15] Ioan Raicu, Pete Beckman, Ian Foster. "Making a Case for Distributed File Systems at Exascale", ACM LSAP 2011