Building Blocks for Scalable Distributed Storage Systems

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

Overview

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

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 CIFellow at Northwestern University in 2009 - 2010, and obtained his Ph.D. in CS from University of Chicago under Dr. Ian Foster in 2009. He is a 3-year award winner of the GSRP Fellowship from NASA ARC. 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 (MTAGS, DataCloud, and ScienceCloud). He is on the editorial board of the Springer's JoCCASA, as well as a guest editor for the IEEE TPDS, SPJ, and 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/.

Acknowledgements

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

References

[1] Ioan Raicu, Pete Beckman, Ian Foster. "Making a Case for Distributed File Systems at Exascale", ACM Workshop on Large-scale System and Application Performance (LSAP), 2011

[2] 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

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

[4] Corentin Debains, Pedro Manuel Alvarez-tabio Togores, Ioan Raicu. "Evaluating Information Dispersal Algorithms", 1st Greater Chicago Area System Research Workshop, 2012

[5] Dongfang Zhao, Ioan Raicu. "HyCache: A Hybrid User-Level File System with SSD Caching", 1st Greater Chicago Area System Research Workshop, 2012

[6] Iman Sadooghi, Dongfang Zhao, Tonglin Li, Ioan Raicu. "Understanding the Cost of Cloud Computing and Storage", 1st Greater Chicago Area System Research Workshop, 2012

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

[8] 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