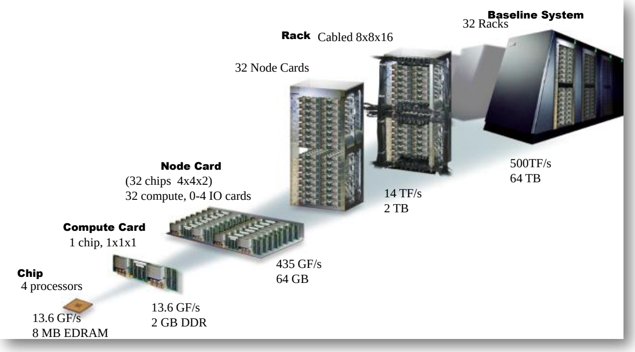
FusionFS:

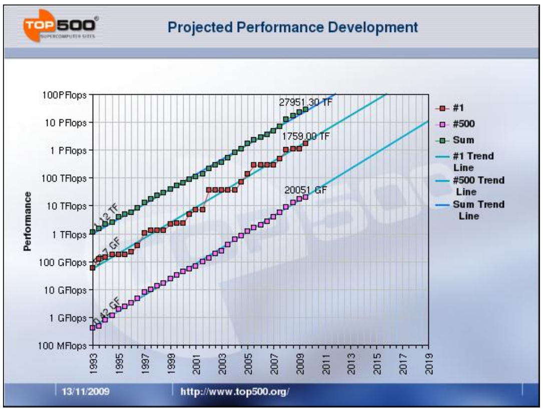
Fusion Distributed File System
http://datasys.cs.iit.edu/projects/FusionFS/

Illinois Institute of Technology
Computer Science Department
Data-Intensive Distributed Systems Laboratory

Motivation



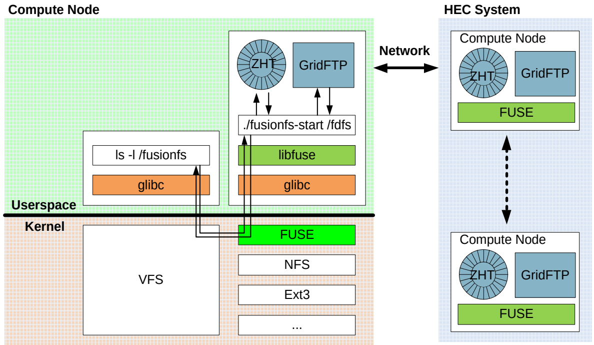
Today's science is generating datasets that are increasing exponentially in both complexity and volume, making their analysis, archival, and sharing one of the grand challenges of the 21st century. Seymour Cray once said – "a supercomputer is a device for turning compute-bound problems into I/O-bound problems" – which drills at the fundamental shift in bottlenecks as supercomputers gain more parallelism at exponential rates, the storage infrastructure performance is increasing at a significantly lower rate. This implies that the data management and data flow between the storage and compute resources is becoming the new bottleneck for large-scale applications. The support for data intensive computing is critical to advancing modern science as storage systems have experienced an increasing gap between their capacity and bandwidth by more than 10-fold over the last decade. There is an emerging need for advanced techniques to manipulate, visualize and interpret large datasets. Many domains share these data management challenges, strengthening the potential impact from generic solutions.



Exascale (i.e. 10^{18} operations/sec) computers will enable the unraveling of significant scientific mysteries, covering many domains (e.g. weather modeling, national security, energy, and drug discovery). Predictions are that exascales will be reached in 2019, with millions of compute-nodes and billions of threads of execution. The current state-of-the-art storage in high-end computing (HEC), in which storage is segregated from compute-nodes and connected by a network (e.g. parallel filesystems), will not scale with the expected exponential growth in concurrency. At exascales, basic functionality (e.g. booting, check-pointing, metadata/data access) at high concurrency levels will suffer poor performance, and combined with system mean-time-to-failure in hours, will lead to a performance collapse.

FusionFS

Professor Ioan Raicu envisions future HEC systems to be designed with non-volatile memory on every compute node, and every node to actively participate in the storage management. He is building a new distributed file system (FusionFS) to address this unprecedented drastic architectural change that requires scalability in the millions of nodes and billions of threads of execution. FusionFS is optimized for a subset of HPC workloads, as well as the emerging Many-Task Computing paradigm, which is more fault tolerant than HPC, diminishing the impact of decreasing mean-time-to-failure. 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, leveraging many-core processors high bisection bandwidth in torus networks. Distributed metadata management is used, implemented in a distributed data-structure, tailored for HEC, supporting constant time operations by emphasizing trustworthy/reliable hardware, fast network interconnects, non-existent node "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, and cooperative caching delivers high aggregate throughput. Data is indexed, by including descriptive, provenance, and system metadata on each file. FusionFS supports a variety of data-access semantics, from POSIX-like interfaces for generality, to relaxed semantics for increased scalability.



Impact

The results of this work has the potential to make exascale computing more tractable, touching virtually all disciplines in HEC. The HEC knowledgebase will extend into commodity systems as the fastest machines generally become mainstream systems in five to seven years. These advancements will impact scientific discovery and economic development at the national level, and they will strengthen a wide range of research activities enabling efficient access, processing, storage, and sharing of valuable scientific data from many disciplines from medicine, astronomy, bioinformatics, chemistry, aeronautics, analytics, economics, to new emerging computational areas in the humanities, arts, and education which are increasingly dealing with ever growing large datasets. The proposed work will revolutionize the storage systems of future HEC systems, and open the door to a much broader class of applications (e.g. Many-Task Computing) that would have not been tractable.

DataSys Data-Intensive Distributed Systems Laboratory
ILLINOIS INSTITUTE OF TECHNOLOGY
Argonne NATIONAL LABORATORY