Energy-Aware Storage

Yan Li[1], Christina Strong[1], Ignacio Corderi[1], Avani Wildani[1], Aleatha Parker-Wood[1],
Andy Hospodor[1], Thomas M. Kroeger[2], Darrell D. E. Long[1]
[1]*Storage Systems Research Center, University of California, Santa Cruz, CA 95064, USA*
[2]*Sandia National Laboratories, Livermore, CA 94550, USA*

1 Introduction

Energy is swiftly becoming a gating issue in large scale storage systems, from high-performance computing (HPC) to data intensive applications. For example, the Square Kilometre Array (SKA) [3] is a large radio telescope array expected to be finished by 2024. Its dishes will produce about one exabyte (EB) of raw image data per day. However, the power envelope goal for the storage systems of future exascale supercomputers has been set at ≈ 2 megawatts (MW) [2]. Based on this projection, we can predict that, by 2020, under the energy cap, a supercomputer will at most be able to provide 2 to 3 EB storage for the scratch space, which is far from enough for most of the planned applications. Thus, new technologies must be developed to reduce the energy consumption of future storage systems. The amount of power a large-scale computer is allowed to use is not the only concern, however. Computation has become relatively cheap, and new trends [1] suggest that data movement will be a major proportion of the power dissipation in future systems.

2 Approach

We address both concerns by considering energy consumption as a key design constraint in large-scale distributed storage systems. We are developing a new energy profiler for efficient energy consumption measurement. Concurrently, we conduct a workload study of large scale scientific workloads. This will allow us to identify characteristics of scientific workloads and develop optimizations based on the type of workload. In order to evaluate new architectural changes and algorithms, we are developing an energy simulator that can predict the energy consumption of a complex storage system. We are developing data allocation algorithms that take power consumption into account by moving data as little as possible on both a geographical and temporal level.

Profiling Energy Cost A key requirement of designing an energy-efficient system architecture is to accurately measure the energy cost in the runtime phase. Without this information, optimization and trade-offs can only be guesswork. The measurement can be hard because an HPC program normally accesses thousands of pieces of hardware, which can be in many different power-saving modes, not to mention that an HPC system is often shared among many programs. We are researching a new metric called **energy score**, which reflects the energy consumption of operations or the energy needed to generate data objects.

For example, the energy score for compressing a bitmap picture to JPEG depends on (1) the energy needed for loading the source picture from storage to memory, (2) the CPU time used for the compression, and (3) the energy needed for storing the result. (1) and (3) in turn depend on the distance and power-saving mode of the storage, and (2) depends on the CPU's speed.

Energy score is fast to calculate and comparable across applications running on the same HPC system. It can be used both at the runtime phase or at the design phase. It is calculated by an OS-level **profiler** and can work with existing programs with or without actual power meters.

Architectural Changes and Simulating Energy Usage The requirements for reliability and performance vary greatly between applications, and future systems will need to dynamically balance between reliability, performance, and energy cost. To achieve this, we are exploring several architectural changes: (1) near-node storage, (2) using heterogeneous storage devices, like combining SSD and hard drives, (3) splitting the central storage system to several smaller storage systems and dispersing them among the system to reduce the distance between nodes and storage, (4) aggressive power management for storage racks, i.e., powering off a whole storage rack if possible, and (5) extensive use of compression. Thus, we need to find answers to many what-

if questions, like, "How much energy can we save by adding near-compute-node SSD storage?"

However, evaluating the effectiveness of these changes can be difficult and expensive. It is hard to test architectural changes on real hardware because HPC systems can be very expensive to build. Therefore, we are designing an **energy simulator** that can simulate the energy cost of a complex computer storage system, which may include millions of storage devices.

By measuring the energy consumption of individual devices in different running modes, we are building an energy footprint database for devices used in an HPC system. The energy simulator uses virtualization to run the real OS and middleware with virtual devices that calculate the cumulative energy consumption from all operations by using the energy footprint database (Figure 1).

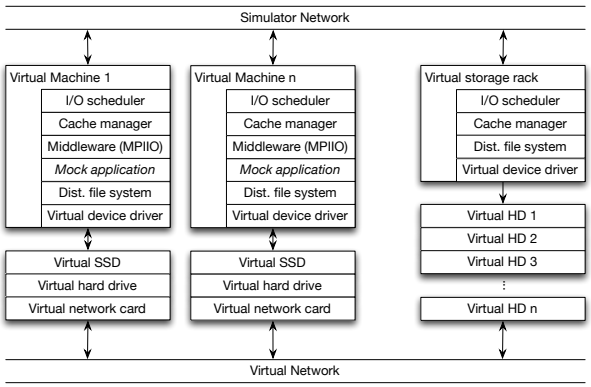

Figure 1: Simulating the storage stacks of multiple connected machines and storage systems. Note that the real computational intensive application is replaced by a "mock application" so that we can simulate hundreds of machines on one computer. Devices will be simulated by virtual devices connected to the virtual nodes. Each virtual device and machine calculates its respective energy consumption.

Energy-Efficient Data Allocation Data allocation places an important role in distributed storage systems, even more so for systems that consist of heterogeneous devices. For non performance critical projects, most of the storage devices should be powered off to save energy, requiring better file grouping strategies. For projects that require high performance I/O, various levels of cache and SSD will be deployed to provide enough bandwidth while at the same time unneeded hard drives will be powered off. For large simulation applications where small errors from a small number of nodes won't affect the final result much, lowering redundancy levels and optimizations, like lazy flushing, show potential for reducing the energy consumption. They all require smarter data allocation algorithms.

In order to reduce the movement of data, we are looking into developing data allocation algorithms that use multi-objective optimizations. Currently we are developing and evaluating an algorithm that explores the redundancy provided by reliability features. Taking the distances between storage devices into consideration, by placing data within a redundancy group diagonally across devices, i.e., different copies of the same data will be stored as far as possible from each other, different computer nodes will use different paths for accessing different copies of the same data, or from neighbor nodes to save the communication with the central storage system. By placing data copies physically far from each other, we can also aggressively power-off unused copies of data to save energy. The drawback of this algorithm is possibly higher metadata management overhead.

Another algorithm we are developing identifies groups of data objects that are often access together and places them near each other for reducing the number of devices needed to run a specific workload. The workload traces we collected help us understand the data access patterns among different workloads, and we are evaluating various machine learning algorithms with these traces for both online and offline data grouping.

3 Current Status

We are currently designing the energy score profiler, simulator, and algorithms for data allocation. To support and evaluate the designs, we are working with our industry partners on acquiring workload traces from various environments. We are in the process of acquiring through our industry partners two separate petabyte storage systems for observing and collecting data: one is for astrophysics computing and the other is for genomic data processing cluster. These two storage systems will give us a starting point, while we continue negotiations with NASA and various national laboratories to get other workloads.

References

[1] S. W. Keckler, W. J. Dally, B. Khailany, M. Garland, and D. Glasco. GPUs and the future of parallel computing. *IEEE Micro*, 31(5):7–17, Sept. 2011.

[2] P. Kogge et al. Exascale computing study: Technology challenges in achieving exascale systems. Technical Report TR-2008-13, Univ. of Notre Dame, 2008.

[3] SKA Organisation. Facts and figures. `http://www.skatelescope.org/media-outreach/fun-stuff/facts-figures/`.