High-performance IO

Storage has become a key component in HPC systems, and the challenges for the Exascale era are huge. In this research line we address such problems both for data and metadata.

Summary

Storage is becoming key in HPC systems, and especially when Exascale systems enter the game. The amount of data needed to solve the coming HPC challenges will not fit in memory, thus storage systems need to keep the pace of computing improvements; otherwise Exascale machines will waste energy waiting for the storage system to deliver the needed data. This research line investigates several paths to improve the performance of storage systems at both data and metadata levels.

Objectives

- **Taking advantage of multi-core architectures.** Multicore-architectures are a clear trend in current and future systems. Such increase in computing capacity is a double-edged sword. On the one hand, more cores imply an increase in the concurrent load that can be sent to the storage system. On the other hand, more computing power can be used to optimize the IO path and make IO operations faster. In this research line we investigate how this extra computing power can be used to optimize disk scheduling, or deduplicate the page cache, among others. For example we can dynamically partition...
cores between different tasks guided by a global performance metric, or improve IO Scheduler using machine learning techniques.

- **Parallel File System Optimization.** Using techniques developed to optimize local storage, we are investigating under the JLESC umbrella, the utilization of such techniques to optimize the I/O Scheduler selected in a PFS.

- **Partial stripe avoidance.** When writing data with redundancy, writing full stripes is more efficient than writing partial stripes, because it avoids reading data to rebuild the redundant part of a stripe. When many clients perform parallel write operations to a shared data object, commonly some clients write partial stripes and other clients write the remainder of such partial stripes. In this research line we investigate how to scale Partial Stripe Avoidance (a mechanism where clients notify proxies of write intentions) to extremely large number of clients. Results published in Performance Impacts with Reliable Parallel File Systems at Exascale Level.

- **Million metadata servers.** Along with the increase in storage needs, metadata management is becoming one of the potential bottlenecks for exascale storage systems. Current metadata servers running in a single node will not scale. In this research line we go one step further and research how metadata management could scale to very large numbers of metadata servers (i.e. a million), thus every storage server (or even every client node) could act as a component in the metadata management.

- **Adaptable data placement.** Optimizing the placement of a storage system has been widely investigated: RAID or randomized solutions seem to be the answer in most cases. Unfortunately, when adding more devices to the storage system, the same degree of optimization can only be achieved with a full restriping or moving a large number of data blocks/objects to the new devices. We investigate how to guarantee the performance level of an optimized storage system, without moving large amounts of block/objects every time the storage system grows in capacity.

- **Software Defined Storage.** Working with OpenStack and big data workloads, introduce Software Defined Storage topics to include bandwidth differentiation, isolation and several intelligent services as filters, prefetching and compression. This objective will also explore the idea of moving computation near the data, using IBM Storlets and other mechanisms. An OpenStack modification to support bandwidth differentiation is available at [GITHUB](https://www.bsc.es/ca/research-development/research-areas/big-data/high-performance-io).

- **I/O Hints Research.** Describing or explaining to the storage layer about the application intentions can provide exceptional benefits in performance, energy usage and several other metrics. We investigate and create new I/O hints to improve several aspects and relate them to other objectives (i.e., Adaptable data placement, memory reduction, partial stripe avoidance and advanced prefetching optimizations).

- **Research using new devices.** Devices like NVRAM and Kinetic drives are included in our research. We are using them to create new dynamic and multi-paradigm filesystems. We use them into data schedulers research to improve the I/O Stack. We also built hardware and software to know the energy used on storage devices with different workloads.

---

Barcelona Supercomputing Center - Centro Nacional de Supercomputación