Execution Environments for Big Data: Challenges for Storage Architectures and Software

Wolfgang E. Nagel, Ralph Müller-Pfefferkorn, Michael Kluge, Daniel Hackenberg

Over the past years a tremendous increase in research data has been seen. These data originate from a variety of sources - experiments, sensor networks, computer simulations, digitized objects, and even from the public, like the world wide web. Challenges in exploring these data arise not only from their sheer quantity but also from the complexity with which they hide their information. According to [1] data sets up to 100 PB are expected already in 2015.

“Big Data” scientific workflows come in a variety of forms. Scientists analyze vast amounts of small files like pictures or movies from microscopes. Simulations create either a very large output file in a single execution (e.g., in CFD) or medium sized result files in many runs. Others combine the small analysis outputs of large data sets into single samples to be processed further (e.g., gene analysis). Even computer science can generate large amounts of data, for example program event traces.

If an infrastructure wants to meet the productivity and performance goals of such a diversity of “Big Data” approaches it must be able to adapt to their needs.

On the hardware architectural level, an integration of heterogeneous data resources into the processing capabilities is required. A single storage paradigm can't satisfy all needs. Different technologies like phase-change memory [2], flash, hard disks, or optical storage provide disparate performance profiles. A flexible storage infrastructure must adapt to an applications data profile regarding access bandwidth, capacity, or IOPS. In the best-case scenario, the user is unaware of the diverse hardware and only sees a single large file system. The infrastructure switches either automatically using I/O performance data collected online or is guided explicitly by the user. Missing such flexibility would mean that general purpose extreme-scale computing systems (satisfying all user needs regarding data analysis) are no longer possible - a storage system attached to an HPC machine would have to be build for a specific data access pattern to be able to scale with the computing infrastructure.

A “Big Data” I/O infrastructure must also provide quality of service. If an application needs high performance data access, the infrastructure must assure the required service quality for all I/O subsystems and networks. Only a fast and reliable data transfer and access can leverage the exascale compute facilities. To transfer data to the processing units and the large machine memory or to move data between different storage technologies external data movers are necessary, which have access to the same infrastructure as the exascale machine. The data movers have to be external as the exascale resources cannot be wasted for data move operations which might be quite slow.

Appropriate software must be available to support the goals mentioned above. Measuring and recording I/O related performance information from the lowest hardware level up to the application data access patterns will enable a modeling of complex architectures for better understanding, improvement, and optimization. For example, knowing the data flow through the all the hard- and software layers of a file system will help to understand bottlenecks related to specific data access patterns, to realize design flaws of (sub)systems, or to evaluate alternative architectures. Being able to analyze such data even online is the prerequisite for a flexible storage infrastructure that adapts automatically to changing I/O patterns.

To support the users in their daily work software environments specialized on data analysis needs have to be provided. A major requirement is that they must naturally integrate into the working environment of the scientists and they must be easy to use or will otherwise hardly ever be accepted.

The management of data needs to be supported by tools. These tools must include both scalable management and scalable access capabilities. Increasing amounts of data not only need to be managed but also need to be stored in and retrieved from the management system in a scalable way. The same is valid for the metadata, which are needed to describe and reuse the research data. In addition, tools to automatically extract technical, contextual as well as disciplinary metadata from the research data are essential. In future, there is no way to create them manually for the exploding amount of data. Using HPC resources for this expensive step might be necessary and beneficial.

Especially for interdisciplinary work and reuse of data it is necessary to automatically create information and knowledge from the data sources. Techniques and algorithms are needed to read information and to extract context, connections, correlations, interpretations, or ideas. While such technologies are already quite common in business data processing, many science disciplines are still focusing on their original analysis strategies only. Tools to query such information and knowledge in a fast, scalable, and easy way will allow scientists to accelerate their research on data subsets or to observe new findings.

Often a data analysis consists of many single steps and might submit thousands or millions of jobs. Therefore, tools for workflow support need more intelligence and must be scalable as well. For example, they must manage data and computing tasks altogether, organizing and balancing the resources needed for both. For resilience reasons, they must be able to realize exceptions/errors and react to them, which means not only restarting or recreating a workflow − they must react on the source of the exception and find ways to circumvent them automatically. Integrating the data management into the workflow system will ease finding and accessing the data relevant for a task. To support interoperability the experiences of Grid Computing approaches might help in the “Big Data” and exascale arena.

Summary:

- Exascale systems will support an extremely large main memory, which will result in huge input/output data (size and/or number of files).
- Because of size and speed, there will be a need to support different kinds of I/O technologies.
- Depending on dynamic requirements, data will have to be moved between these I/O devices.
- To transfer data between the different storage technologies, external data movers are needed.
- These have to be integrated into a scalable workflow that handles data and computing tasks.
- A single multipurpose file system serving the variety of needs (e.g., IOPS, bandwidth) has to be flexible, which can be achieved by using different storage technologies combined with an intelligent management middleware layer.
- Metadata and information need to be extracted automatically, and tools to explore these will ease data analyses and scientific findings.

[1] Dongarra, J., Beckman, P. et al.,: “The International Exascale Software Roadmap”, Volume 25, Number 1, 2011, International Journal of High Performance Computer Applications, ISSN 1094-3420.

[2] Ning, L., Cope, J., Carns, P., Carothers, C., Ross, R., Grider, G., Crume, A., Maltzahn, C.: “On the role of burst buffers in leadership-class storage systems.” In MSST/SNAPI, April 2012.