Chapter 4
Distributed Storage Systems for Data Intensive Computing

Sudharshan S. Vazhkudai
Oak Ridge National Laboratory, USA

Ali R. Butt
Virginia Polytechnic Institute and State University, USA

Xiaosong Ma
North Carolina State University, USA

ABSTRACT

In this chapter, the authors present an overview of the utility of distributed storage systems in supporting modern applications that are increasingly becoming data intensive. Their coverage of distributed storage systems is based on the requirements imposed by data intensive computing and not a mere summary of storage systems. To this end, they delve into several aspects of supporting data-intensive analysis, such as data staging, offloading, checkpointing, and end-user access to terabytes of data, and illustrate the use of novel techniques and methodologies for realizing distributed storage systems therein. The data deluge from scientific experiments, observations, and simulations is affecting all of the aforementioned day-to-day operations in data-intensive computing. Modern distributed storage systems employ techniques that can help improve application performance, alleviate I/O bandwidth bottleneck, mask failures, and improve data availability. They present key guiding principles involved in the construction of such storage systems, associated tradeoffs, design, and architecture, all with an eye toward addressing challenges of data-intensive scientific applications. They highlight the concepts involved using several case studies of state-of-the-art storage systems that are currently available in the data-intensive computing landscape.

DOI: 10.4018/978-1-61520-971-2.ch004

DATA INTENSIVE COMPUTING CHALLENGES

The advent of extreme-scale computing systems, e.g., Petaflop supercomputers, cyber-infrastructure, e.g., TeraGrid, and experimental facilities such as large-scale particle colliders, are pushing the envelope on dataset sizes. Supercomputing centers routinely generate huge amounts of data, resulting from high-throughput computing jobs. These are often result-datasets or checkpoint snapshots from long-running simulations. For example, the Jaguar petaflop machine (National Center for Computational Sciences [NCCS], 2009) at Oak Ridge National Laboratory, which is No.2 in the Top 500 supercomputers as of this writing, is generating terabytes of user data while supporting a wide-spectrum of science applications in Fusion, Astrophysics, Climate and Combustion. Another example is the TeraGrid, which hosts some of NSF's most powerful supercomputers such as Kraken (National Institute of Computational Sciences [NICS], 2008) at the University of Tennessee, Ranger (Sun constellation linux cluster, 2008) at Texas Advanced Supercomputing Center and Blue Waters at National Center for Supercomputing Applications, and are well on their way to produce large amounts of data. Accessing these national user facilities is a geographically distributed user-base with varied end-user connectivity, resource availability, and application requirements. At the same time, experimentation facilities such as the Large Hadron Collider (LHC) (Conseil Europ'een pour la Recherche Nucl'eaire [CERN], 2007) or the Spallation Neutron Source (SNS) (Spallation Neutron Source [SNS], 2008; Cobb et al., 2007] will generate petabytes of data. These large datasets are processed by a geographically dispersed user-base, often times, on high-end computing systems. Therefore, result output data from High-Performance Computing (HPC) simulations are not the only source that is driving dataset sizes. Input data sizes are growing many folds as well (SNS, 2008; CERN, 2007; Sloan digital sky survey [SDSS], 2005; Laser Interferometer Gravitational-Wave Observatory [LIGO], 2008).

In addition to these high-end systems, commodity clusters are prevalent and the data they can process is growing manifold. Most universities and organizations host mid-sized clusters, comprising of hundreds of nodes. A distributed user base comes to these machines for a variety of data intensive analyses. In some cases, compute intensive operations are performed at supercomputing sites, while post-processing is conducted at local clusters or high-end workstations at end-user locations. Such a distributed user analysis workflow entails intensive I/O. Consequently, these systems will need to support: (i) the staging in of large input data from end-user locations, archives, experimental facilities and other compute centers; (ii) the staging out terabytes of output, intermediate and checkpoint snapshot data to end-user locations or other compute destinations (iii) the ability to checkpoint terabytes of data at periodic intervals for a long-running computation; (iv) the ability to support high-speed reads to support a running application.

In the discussion below, we will high light these key data intensive operations, the state-of-the-art and the challenges and gaps there in to set the stage for how distributed storage systems can help in optimizing them.

Data Staging and Offloading

Large input, output and checkpoint data is required to be staged in and out of these systems. With the exponential growth in application input and output data sizes, it is impractical to store all user data indefinitely at HPC centers. Traditionally, centers have operated under the premise that users come to them with all of their storage and computing needs. The legacy of this approach still weighs heavily when it comes to provisioning a center as significant portions of the operational budget is spent on large data stores and archives. End-