BigData Express: Toward Schedulable, Predictable, and High-Performance Data Transfer

Wenji Wu, Liang Zhang, Qiming Lu, Phil DeMar
Fermilab
{wenji, liangz, qlu, demar}@fnal.gov

Big data has emerged as a driving force for scientific discoveries. Large scientific instruments (e.g., colliders, light sources, and telescopes) generate exponentially increasing volumes of data. Currently, Large Hadron Collider (LHC) experiments generate hundreds of petabytes of data per year. The aggregated amount of climate science data is expected to exceed 100 exabytes by 2020. To enable scientific discovery, science data must be collected, indexed, archived, shared, and analyzed, typically in a widely distributed, highly collaborative manner. At present, computing facilities for large-scale science, such as ALCF, OLCF, and NERSC, offer the types of computing and storage resources needed to process and analyze science data. The efficient movement of science data from their sources into processing and storage facilities and ultimately on to user analysis is critical to the success of any such endeavor. Data transfer is now an essential function for science discoveries, particularly within big data environments.

Within the U.S. research communities, the emergence of distributed, extreme-scale science applications is generating significant challenges regarding data transfer. We believe that the data transfer challenges of the extreme-scale era are characterized by two relevant dimensions:

- *High-performance challenges*. First, it is becoming critical to transfer data at the highest possible throughputs because the volumes of science data are growing exponentially. Second, the U.S. research communities are working toward deploying extreme-scale supercomputer facilities in support of extreme-scale science applications. To fully utilize these expensive computing facilities, ultra-high-throughput data transfer capabilities will be required to move data in or out of them.
- *Time-constraint challenges*. Scientific applications typically have explicit or implicit time constraints on data transfer. Based on the nature of these time constraints, data transfer tasks can be classified into three broad categories: *(a) Real-time, (b) deadline-bound, and (c) background data transfer*. For *real-time data transfer*, the data transfer task is on the critical path for the end-user experience or a real-time experimental control loop. Scientific applications, such as real-time data analysis, remote visualization, and real-time experimental control, are highly sensitive to data transfer delays. Even small increases in data transfer time can degrade the user experience or result in inaccurate scientific results. For *Deadline-bound data transfer*, the data transfer task is not on the critical path for the end-user experience or a real-time experimental control loop, but it does have an explicit deadline. For example, job startup and scratch storage space purge deadlines in supercomputer centers require deadline-bound data movement. for *Background data transfer*, the data transfer task has a long deadline or no explicit deadline. For example, replicating data from one data center to another data center for long-term storage is a background data transfer task.

To date, several data transfer tools (e.g., GridFTP and BBCP) and services (e.g., the PhEDEx and Ruccio systems, the LIGO Data Replicator, and Globus Online) have been developed to support science data movement. Advanced data transfer features, such as transfer resumption, partial transfer, third-party transfer, and security, have been implemented in these tools and services. There have also been numerous enhancements to speed up data transfer performance, including the following:

- Parallelism at all levels (e.g., multi-stream parallelism, multicore parallelism, and multi-path parallelism) is widely implemented in bulk data movement and offers significant improvement in aggregate data transfer throughput.
- Science DMZ architectures with dedicated high-performance Data Transfer Nodes (DTNs) have been widely deployed. The hardware devices, software, configurations, and policies of Science DMZ are structured and optimized for high-performance data transfer.
- The U. S. research communities are working toward deploying terabit networks in support of distributed extreme-scale data movement. Existing backbone networks are now based on ultra-scalable 100-gigabit

technologies. Advanced virtual path services such as ESnet OSCARS and Internet2 AL2S has been developed.

Although significant improvements have been made in science data transfer capabilities, the currently available data transfer tools and services will not be able to successfully address the high-performance and time-constraint challenges of data transfer to support extreme-scale science applications for the following reasons:

(1) *Problem 1: Disjoint end-to-end data transfer loops*. In current data transfer frameworks, each entity in an end-to-end data transfer loop (e.g., DTN, LAN, WAN, and storage) is scheduled and managed locally, and their policies and mechanisms may act at odds with each other. Without end-to-end integration and coordination, this distributed resource management model may readily lead to resource contention or performance mismatch in the end-to-end loop. As a result, suboptimal (or even poor) performance would occur.

(2) *Problem 2: Cross-interference between data transfers.* A significant amount of cross-interference between data transfers can lead to contention for various resources (e.g., DTN, LAN, WAN, and storage), resulting in degraded performance. This can also lead to high variability in data transfer performance. Existing data transfer tools and services lack effective mechanisms to minimize cross-interference between data transfers.

(3) *Problem 3: Existing data transfer tools and services are oblivious to user (or user application) requirements (e.g., deadlines and QoS requirements).* Without deadline awareness, it is difficult to satisfy the time constraint requirement on data transfer.

(4) *Problem 4: Inefficiencies arise when existing data transfer tools are run on DTNs*. High-end DTNs are typically NUMA systems. However, existing data transfer tools are unable to fully exploit multicore hardware under the default OS support, especially on NUMA systems.

If these problems are not addressed appropriately, they will undermine the ability to support extreme-scale science in the coming years.

Fermilab has been working on the BigData Express project (http://bigdataexpress.fnal.gov) to address these problems. BigData Express seeks to provide a schedulable, predictable, and high-performance data transfer service for big data science. Essentially, *BigData Express* is a middleware data transfer service with the following key features:

- A data-transfer-centric architecture to seamlessly integrate and effectively coordinate computing resources in an end-to-end data transfer loop.
- A distributed peer-to-to model for data transfer services, making it very flexible for the establishment of data transfer federations.
- A scalable software architecture. BigData Express makes use of MQTT as message bus to support communication among its components.
- An extensible plugin framework to support different data transfer protocols, including mdtmFTP, GridFTP, and XrootD.
- An end-to-end data transfer model with fast provisioning of end-to-end network paths for guaranteed QoS. Specifically, the use of an SDN-enabled BigData-Express LANs and SDN-enabled WAN path services to reduce or eliminate network congestion.
- A high-performance data transfer engine. BigData Express adopts mdtmFTP as its default data transfer engine. mdtmFTP is specifically designed for optimization of data transfer performance on multicore systems (DTNs).
- A rich set of REST APIs to support scientific workflows.

The BigData Express software is currently deployed and being evaluated at multiple research institutions, including UMD, StarLight, FNAL, KISTI, KSTAR, SURFnet, and Ciena. The BigData Express research team is collaborating with StarLight to deploy BigData Express on various research platforms, including Pacific Research Platform, National Research Platform, and Global Research Platform. It is envisioned that we are working toward building a high-performance data transfer federation for big data science.