Research Activities at Fermilab for Big Data Movement

Parag Mhashilkar, Wenji Wu, Hyun Woo Kim, Gabriele Garzoglio, Dave Dykstra, Marko Slyz, Phil DeMar
Computing Sector, Fermilab
PO Box 500, Batavia, IL 60510, USA
{parag, wenji, hyunwoo, garzogli, dwd, mslyz, demar}@fnal.gov

Abstract — Adaptation of 100GE Networking Infrastructure is the next step towards management of Big Data. Being the US Tier-1 Center for the Large Hadron Collider’s (LHC) Compact Muon Solenoid (CMS) experiment and the central data center for several other large-scale research collaborations, Fermilab has to constantly deal with the scaling and wide-area distribution challenges of the big data. In this paper, we will describe some of the challenges involved in the movement of big data over 100GE infrastructure and the research activities at Fermilab to address these challenges.

Index Terms — Big Data, 100GE networks, multicore, many core, high-speed networks.

I. INTRODUCTION

Adaptation of 100GE Networking Infrastructure is the next step towards management of Big Data. The ability to efficiently store, retrieve, analyze and redistribute data is fundamental to scientific discoveries. Fermilab is the US Tier-1 Center for the Large Hadron Collider’s (LHC) Compact Muon Solenoid (CMS) experiment and the central data center for several other large-scale research collaborations. To deal with the scaling and wide-area distribution challenges of the data, the Laboratory is now connected to next generation of networking infrastructure, ESnet’s 100G-backbone. This trend towards faster networks has driven the evolution of data management systems. Today, these systems provide access to and manage an increasingly distributed ensemble of resources, storing aggregate data approaching the Exabyte scale. However, this progress is jeopardized by the inability to take into account the disparity in network connectivity and end-host capabilities. To identify and address these shortcomings in the tools and technologies, Fermilab has devised a diverse program of work that spans all layers of computing to ensure full throughput in and across each layer. This paper highlights the multi-prong approach particularly in the Tools and Services Layer, Data Management Layer, Application I/O Layer and the OS Network Driver Layer that is vital to successful adaptation and utilization of the 100GE Networking infrastructure at the Fermilab.

II. HIGH THROUGHPUT DATA PROGRAM (HTDP)

The High Throughput Data Program (HTDP) at the Fermilab focuses on the application I/O layer with the aim of analyzing the performance of the end-to-end analysis systems used by the High Energy Physics (HEP) Experiment by performing large data transfers over the 100GE ESnet networks. This work was fundamental in identifying gaps in the current data movement middleware services when used over the 100GE networks. The experimental results [1, 2] from testing the services such as, GridFTP, Globus Online, SRM, XRootD revealed that over 90% bandwidth utilization can be achieved in average while transferring datasets with medium (8MB – 1GB) and large (2GB – 8GB) file sizes. Doing the same study at a finer granularity of file sizes, however, this higher bandwidth utilization is consistently achieved only when transferring files 32MB or larger. File sizes smaller than 32MB suffer from Lots of Small Files (LOSF) problem, which typically results in significantly lower transfer bandwidth utilization resulting from OS, system and protocol level overhead involved in reading and transferring individual files. As shown in figure 2, pipelining in GridFTP helps reduce the LOSF threshold as subsequent files in a pipeline are transferred without waiting for successful transmission of previous file thus reducing one round trip time (RTT) delay between two consecutive files transferred.

Figure 1 Transfer rates for different file sizes

Figure 2 Pipelining reduces the LOSF threshold

III. CORTEXNET

As networks support higher and higher bandwidth, the progress made on data management may be jeopardized by the...
growing disparity in network connectivity among institutions. While large corporation headquarters, national laboratories, or larger universities will be connected to Terabit-class networks, small institutions may still support previous-generation network connectivity, making available local storage resources through 100 Mb/s to Gb/s links. In this distributed data-centric environment, systems that do not take into account the connectivity and availability of storage resources will become the bottleneck of data processing for all communities. To avoid this situation, it is crucial that the data management systems be aware of the current and historical network conditions to schedule data transfer for staging or just-in-time access accordingly. The CortexNet project at the Fermilab aims at researching new methods and technologies in intelligent selection of storage from the multitude of dispersed resources available. As shown in the figure 2, the intelligent storage selection of CortexNet will take into account historical and current information on the status of the network, sources and sinks of the data and the performance and responsiveness of each component and connection to compute matrices for the cost to move data between pairs of storages. CortexNet will rely on the data gathered by an emerging national network-monitoring infrastructure – collaboration between the ESnet, Internet2 and the Open Science Grid (OSG) to deploy network measurement agents at more than 100 scientific institutions in the United States.

![Figure 3 CortexNet Architecture](image)

**IV. MULTICORE-AWARE DATA TRANSFER MIDDLEWARE**

It is quite clear that, to support Networking at 100GE scales and beyond, new techniques also need to be devised and implemented that supports faster data movement much closer to the hardware and OS level. To date, numerous efforts have been made to exploit multicore parallelism to speed up data transfer performance. The Multicore-aware Data Transfer Middleware (MDTM) project at the Fermilab focuses on the OS/driver layer and implements a user-space resource scheduler. At the OS level, major Operating Systems (e.g., Windows, Solaris, and Linux) have been redesigned and parallelized to better utilize additional cores. Modern network stacks can exploit cores to allow either message- or connection-based parallelism to enhance both performance and processor efficiency. At the hardware level, new multi-queue NIC technologies have been introduced. At data reception, the RSS and Flow Director technologies distribute incoming packets across multiple queues, with each NIC queue handled by a different Interrupt on a dedicated CPU core to achieve scalability. In data transmission, each core in the system is assigned a specific transmit queue. The outgoing traffic generated on a specific core is transmitted via its corresponding transmit queue, and multiple cores can transmit in parallel. The rise in the use of NUMA (non-uniform memory access) systems is due to the scalability advantage of NUMA architecture over UMA (uniform memory access) architecture, high-performance data transfer systems. The user-space resource scheduler implemented, as part of MDTM is NUMA topology-aware and supports core affinity on networking processing and QoS to capitalize on the parallelism of the multicore systems.

![Figure 4 MDTM Architecture](image)

**V. FERMI CLOUD 100GE R&D FACILITY**

To support the R&D activities at different levels, Fermilab is also building FermiCloud 100GE testbed; a state of the art facility connected to the 100GE backbones. This R&D facility is in its final stages of commissioning, will host several performance servers interfacing to different data and storage systems and will support different networking related R&D projects.

Reference:
