# SC18 Network Research Exhibition: Demonstration Abstract

# MDTMFTP: A High-performance Data Transfer Tool

## Abstract

To address challenges in high performance data movement for large-scale science, the Fermilab network research group has developed mdtmFTP, a high-performance data transfer tool to optimize data transfer on multicore platforms. mdtmFTP has a number of advanced features. First, it adopts a pipelined I/O design. Data transfer tasks are carried out in a pipelined manner across multiple cores. Dedicated threads are spawned to perform network and disk I/O operations in parallel. Second, mdtmFTP uses multicore-aware data transfer middleware (MDTM) to schedule an optimal core for each thread, based on system configuration, in order to optimize throughput across the underlying multicore core platform. Third, mdtmFTP implements a large virtual file mechanism to efficiently handle lots-of-small-files (LOSF) situations. Finally, mdtmFTP unitizes optimization mechanisms such as zero copy, asynchronous I/O, batch processing, and pre-allocated buffer pools, to maximize performance.

At SC'18, we will demonstrate mdtmFTP for bulk data movement over wide area networks. New features and latest optimizations will be showcased. We will compare mdtmFTP with current generation data transfer tools such as GridFTP and BBCP.

## I. Overview

Although Big Data has recently become a hot topic in the media, it has been a driving force in scientific discovery for many years. U.S. Department of Energy (DOE) supercomputing facilities (NERSC, ALCF and OLCF), for example, annually generate hundreds of petabytes of simulation data. Those extreme data volumes are typically stored in geographically dispersed facilities. Scientists often need to transfer large data sets from a local storage system to computation resources at an HPC facility for analysis or visualization purposes. End-to-end data transfer rates become a key performance metric in the efficiency and success of these types of big data workflows.

Today, data transfer tools such as GridFTP and BBCP are commonly deployed to support this type of large-scale bulk data movement. These tools implement many useful data transfer features, including transfer resumption, partial transfer, third-party transfer, and security infrastructure services. There have been numerous state-of-art enhancements in these tools to speed up performance. Parallelisms at many levels are now widely implemented to provide significant improvement in aggregate data transfer throughput. However, these data transfer tools will likely not sufficiently answer the emerging challenges for data movement in an exascale computing environment based on terabit networks.

The end of Moore's Law in chip development is producing a new generation of advanced computer architectures that exploit parallelism to continue Moore's Law-like growth in processing power. Among this new generation of processing technologies are multicore, manycore (e.g., GPU), and hybrid mixtures of the two technologies. These technologies are now commonly found in a wide spectrum of computer hardware, ranging from the large supercomputers in the DOE Office of Science's computing facilities, the data servers in massively parallel data centers, and even mobile devices. Multicore and manycore have become the norm for high-performance computing. These new architectures provide advanced features that can be exploited to design and implement a new generation of high-performance data movement tools.

Fermilab network research group has developed a new high-performance data transfer tool, called mdtmFTP, to maximize data transfer performance on multicore platforms. mdtmFTP has several advanced features. First, mdtmFTP adopts a pipelined I/O design. A data transfer task is carried out in a pipelined manner across multiple cores. Dedicated I/O threads are spawned to perform I/O operations in parallel. Second, mdtmFTP uses a particularly designed multicore-aware data transfer middleware (MDTM) to schedule cores for its threads, which optimize use of underlying multicore system. Third, mdtmFTP implements a large virtual file mechanism to address the lots-of-small-files (LOSF) problem. Finally, mdtmFTP unitizes multiple optimization mechanisms – zero copy, asynchronous I/O, batch processing, and pre-allocated buffer pools – to improve performance mdtmFTP software is available at http://mdtm.fnal.gov.

In this demo, we use mdtmFTP to demonstrate bulk data movement over long-distance wide area networks. New features and latest optimization will be showcased. We will compare mdtmFTP with existing data transfer tools such as GridFTP and BBCP. Our purpose is to show that mdtmFTP perform better than existing data transfer tools.

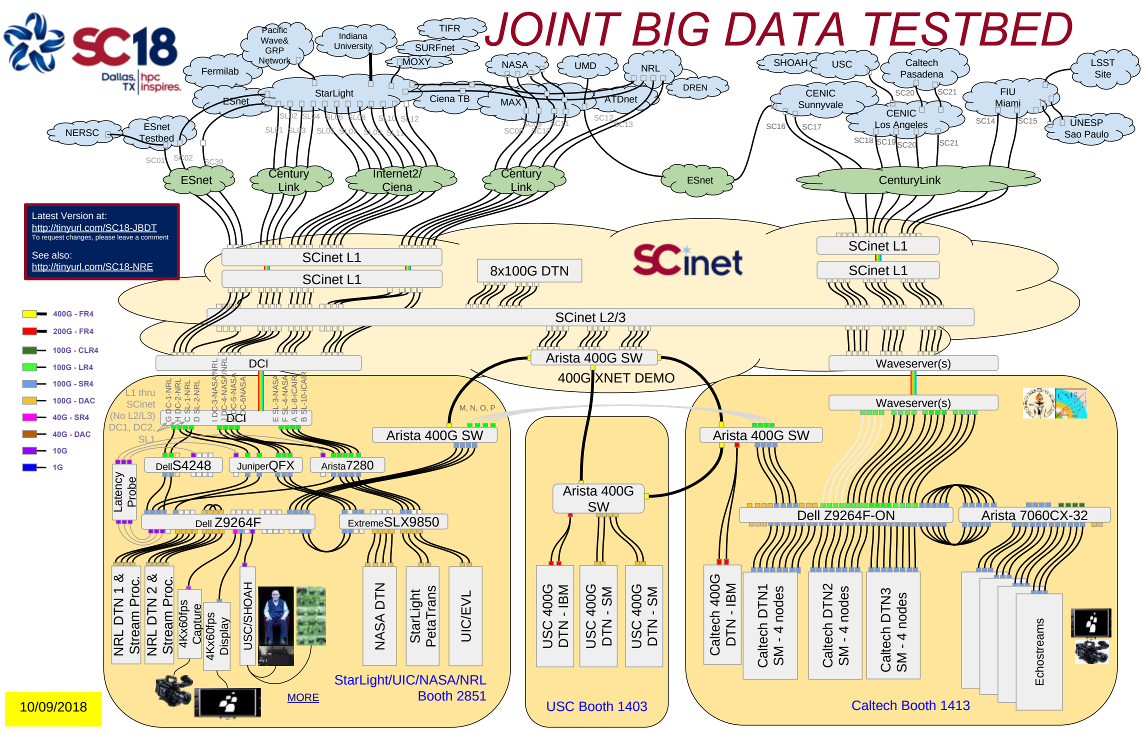

Figure 1 mdtmFTP SC’18 Demo Network Topology

## II. Innovation

mdtmFTP is a high-performance data transfer tool for big data. It has several salient features:

- An I/O-centric architecture to execute data transfer tasks, including dedicated I/O threads for network I/O, disk I/O, and management operations, respectively.
- Optimal scheduling of threads on specific cores to maximize performance and minimize inter-NUMA data movement.
- Reservation of specific cores for an application’s threads to optimize that application’s performance.
- A large virtual file mechanism to concurrently transfer Lots of Small Files (LoSF) extremely efficiently.

## III. HPC and Science Relevance

The next-generation of HPC facilities are expected to support computation for many data-intensive science projects, such as High-Luminosity LHC experiments (HL-LHC). Extremely efficient data movement at BigData volumes will be critical in the success of HPC facilities to support these data-intensive sciences. mdtmFTP and MDTM middleware are designed and developed specifically for this type of extreme-scale data movement.

## IV. SCinet and R&E Requirements

The SCinet WAN team is working with the StarLight facility to implement several 100 Gbps paths from that facility to the SC’18 show floor. One of those 100 Gbps paths will be used to showcase the capabilities of mdtmFTP. At the venue, data will be carried over the terabit network on the show floor.

## V. Network Topology

The proposed demo will utilize high-end NUMA servers located at the StarLight facility and the StarLight/OCC booth on the SC exhibit floor, respectively. The NUMA servers will be back-ended with high-performance storage, such as SSD &RAID, and will have one or multiple 100 GigE NICs. The WAN path will traverse 100GE network infrastructure. Figure 1 illustrates the network topology for mdtmFTP SC’18 Demo.

## VI. Involved Parties

- Liang Zhang, Fermilab, liangz@fnal.gov
- Wenji Wu, Fermilab, wenji@fnal.gov
- Phil DeMar, Fermilab, demar@fnal.gvo
- Se-young Yu, ICAIR, young.yu@northwestern.edu
- Jim Chen, iCAIR, jim-chen@northwestern.edu
- Joe Mambretti, iCAIR, j-mambretti@northwestern.edu
- Fei I Yeh, iCAIR, fyeh@northwestern.edu