

The Multicore-aware Data Transfer Middleware (MDTM) Project

Wenji Wu (PI)

MDTM Research Team

ACM & IEEE SC'15

Austin, TX November 15-20, 2015



Problem Space

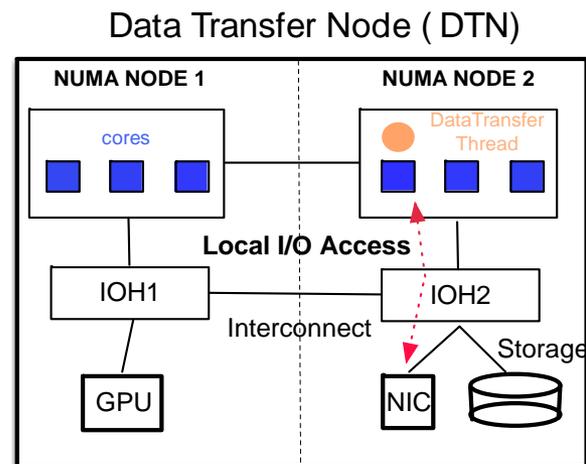
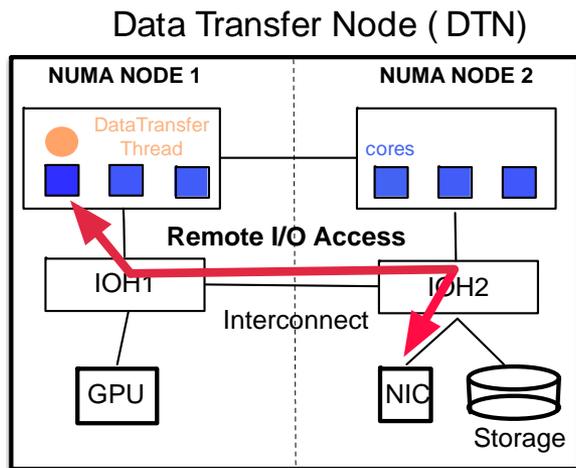
Multicore/manycore has become the norm for high-performance computing.

Existing data movement tools are still limited by major inefficiencies when run on multicore systems

These inefficiencies will ultimately result in performance bottlenecks on end systems. Such bottlenecks also impede the effective use of advanced high-bandwidth networks.



A simple inefficiency case ...



Scheduling without I/O locality

Scheduling with I/O locality

General-purpose OSes have only limited support for I/O locality!

How can we improve?



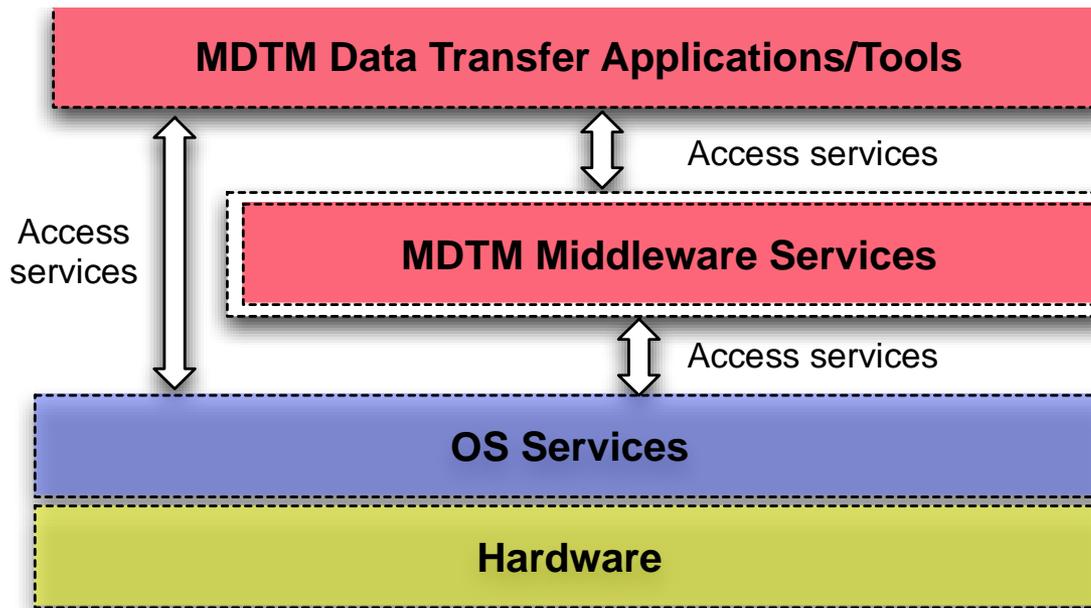
Our solution

- **The Multicore-aware Data Transfer Middleware (MDTM) Project**
 - Collaborative effort by Fermilab and Brookhaven National Laboratory
 - Funded by DOE's Office of Advanced Scientific Computing Research (ASCR)
 - A three-year research project

MDTM aims to accelerate data movement toolkits on multicore systems



MDTM Architecture



MDTM data transfer applications (Client or Server)

- Data transfer applications that use MDTM middleware service

MDTM middleware services

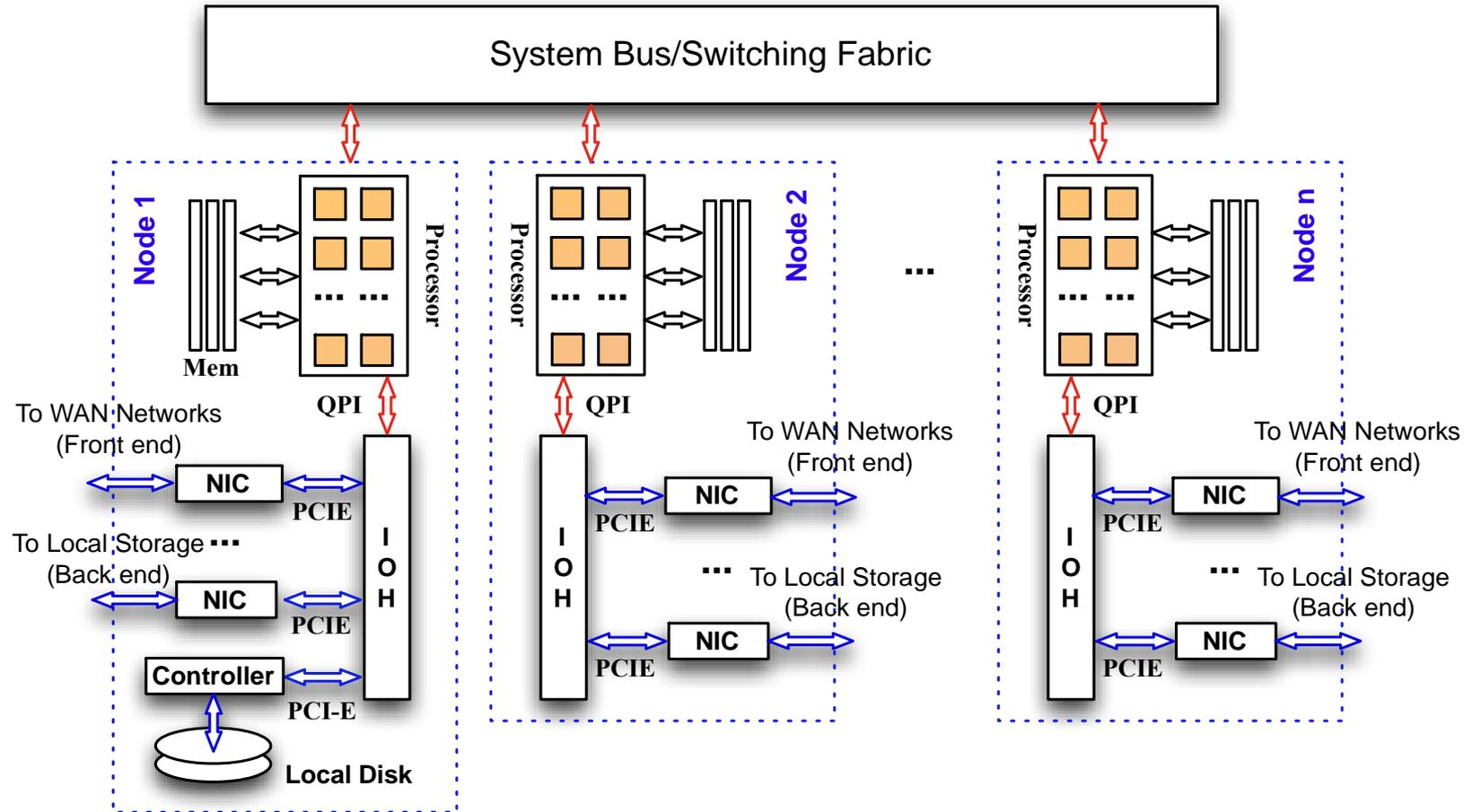
- A user space scheduler that schedule and assigns system resources based on the needs of data transfer applications. It also takes into account other factors, including NUMA topology, I/O locality, and Qos



MDTM Hardware Platform



MDTM is targeted for Data Transfer Node (DTN)

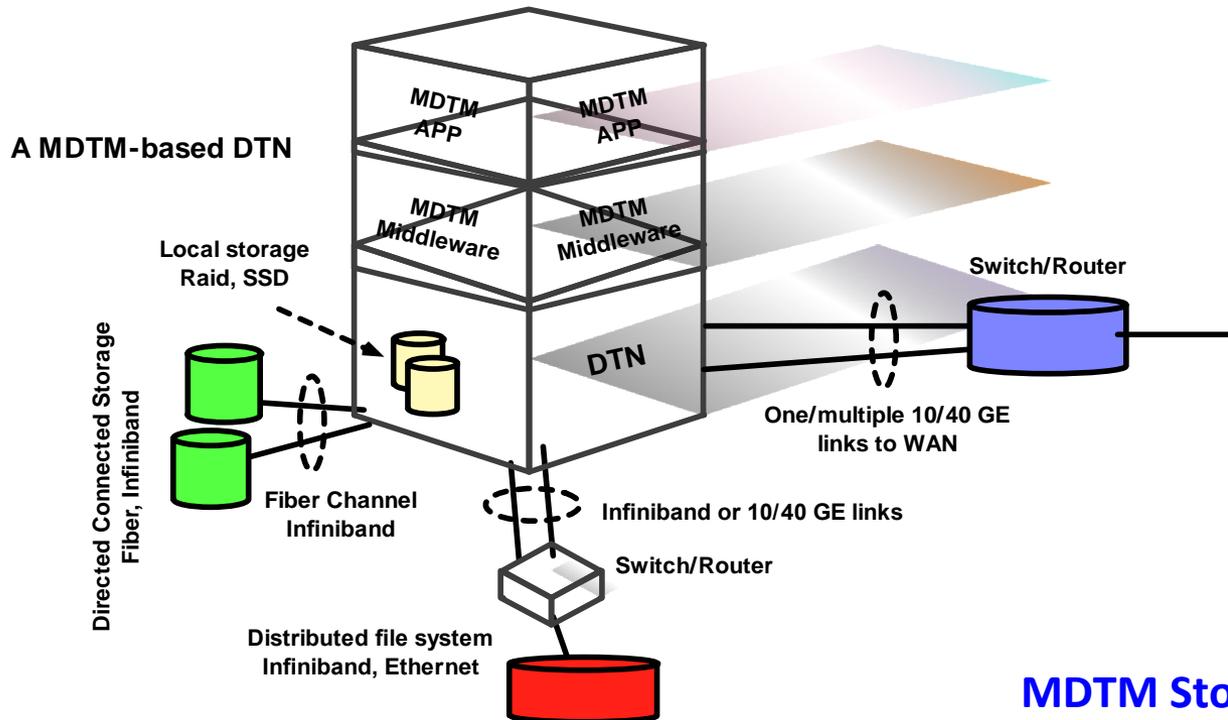


Each DTN features one or multiple NUMA nodes.

Each NUMA node features one or multiple processors that consists of multiple cores.



A MDTM-based DTN Storage and Networking Architecture



MDTM Networking Architecture

- One or multiple WAN links for data transfer
 - Via 10/40 GE NICs
- One or multiple LAN links for storage access
 - Via 10/40 GE NICs, IB adaptors, FC adaptors

MDTM Storage Architecture

- Local storage (Raid, SSD)
- Directed connected storage (FC, IB)
- Distributed file system (IB, 10/40 GE)



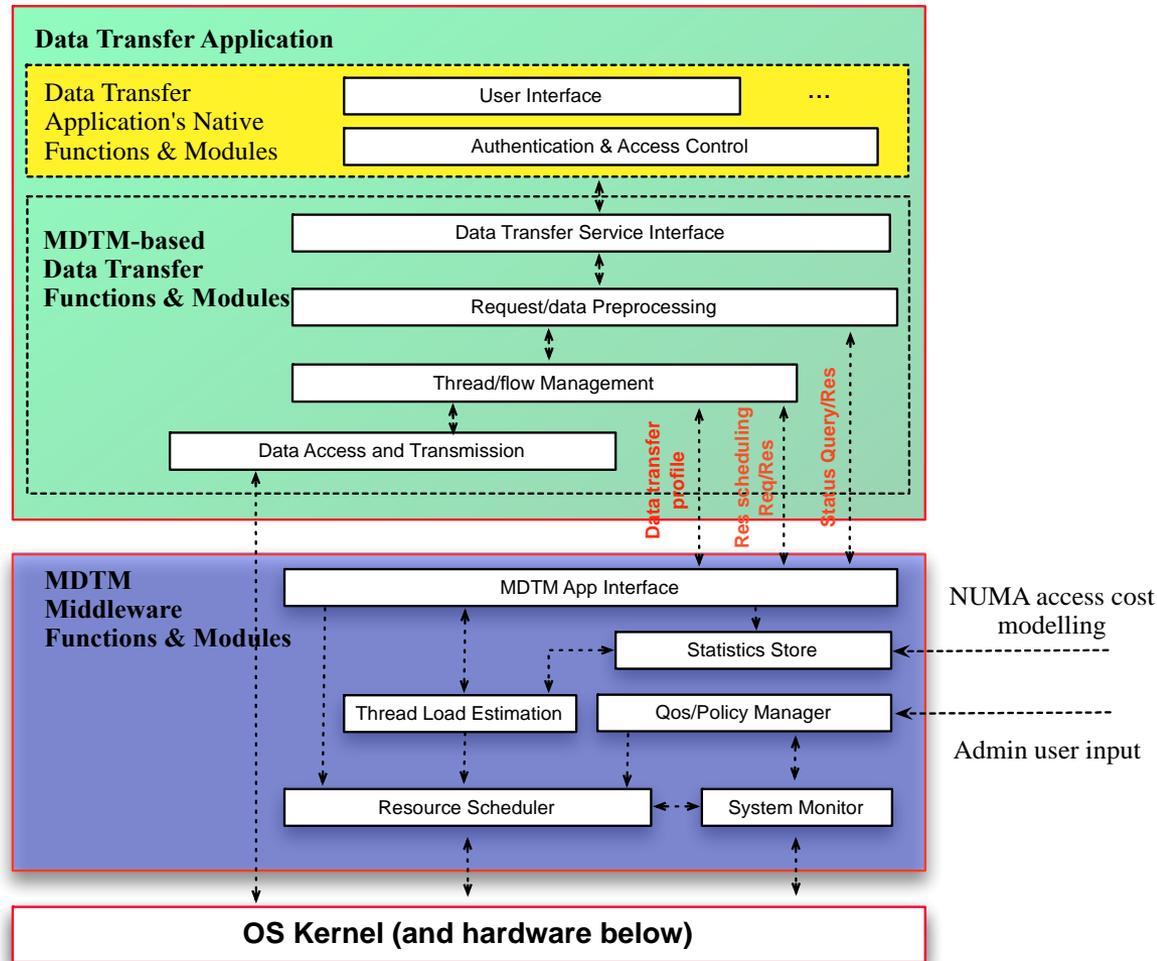
MDTM Software



MDTM Logical Functions and Modules

I/O-Centric architecture
 Parallel data transfer
 Data layout preprocessing

Data flow-centric scheduling
 NUMA-awareness scheduling
 I/O locality optimization
 Maximizing parallelism



How does MDTM works?

A MDTM application spawns three types of threads

- Management threads to handle user requests and management-related functions
- Dedicated disk/storage I/O threads to read/write from/to disks/storages
- Dedicated network I/O threads to send/receive data

A MDTM data transfer application accesses MDTM middleware services explicitly via APIs

In operation, an MDTM middleware daemon will be launched. It will support two types of services

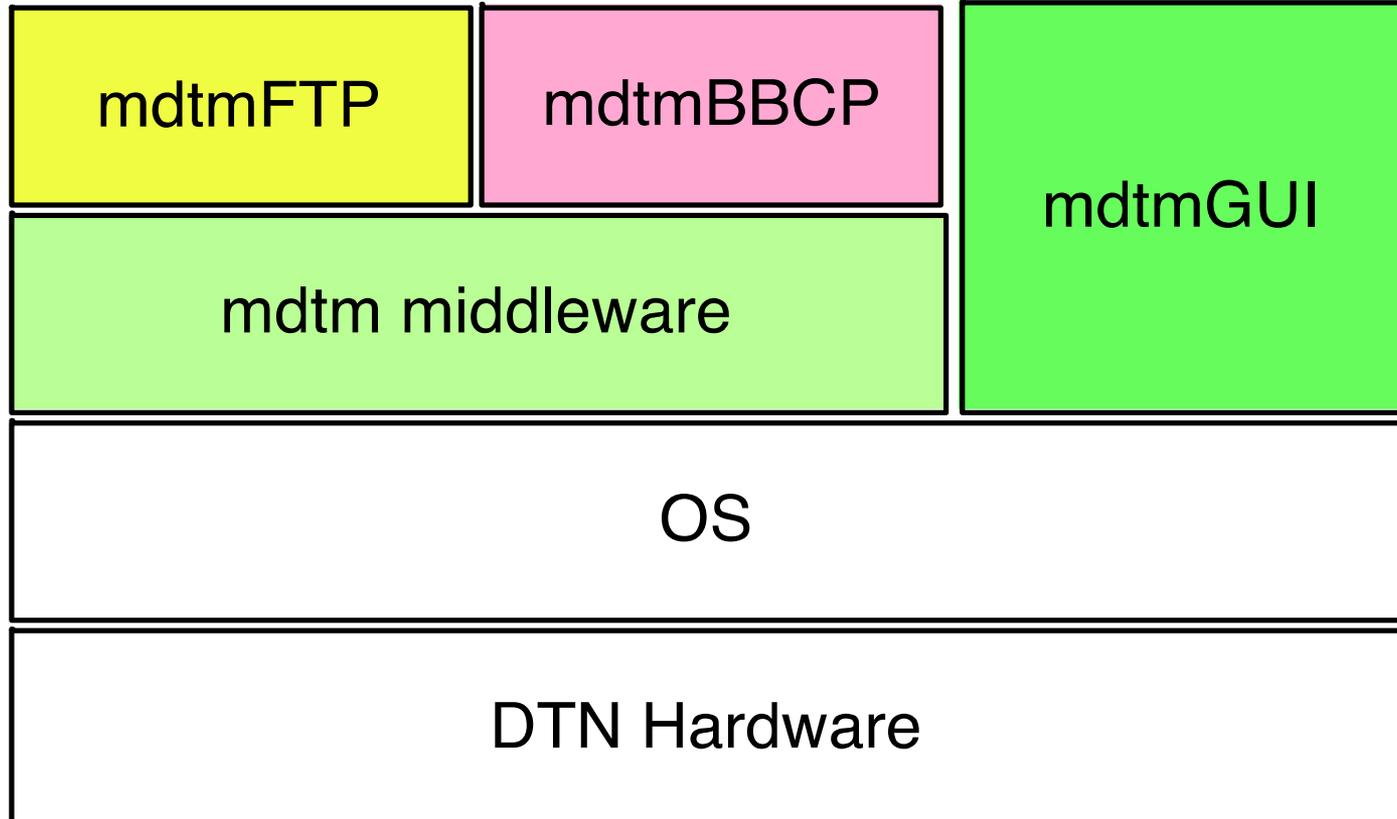
- Query service allow MDTM APP to access system configuration and status
- Scheduling service assigns system resources based on requirements of data transfer applications



MDTM Project Year-2 Status



MDTM Software Suite



Available at: <http://mdtm.fnal.gov>



MDTM Middleware

- Support features
 - Multicore system profiling
 - NUMA system topology layout
 - System status (e.g., CPU/MEM loads)
 - Topology-based resource scheduling
 - Supporting core affinity on network and disk I/O capability
 - Core partitioning on NUMA system
 - Avoid/minimize interference to data transfers



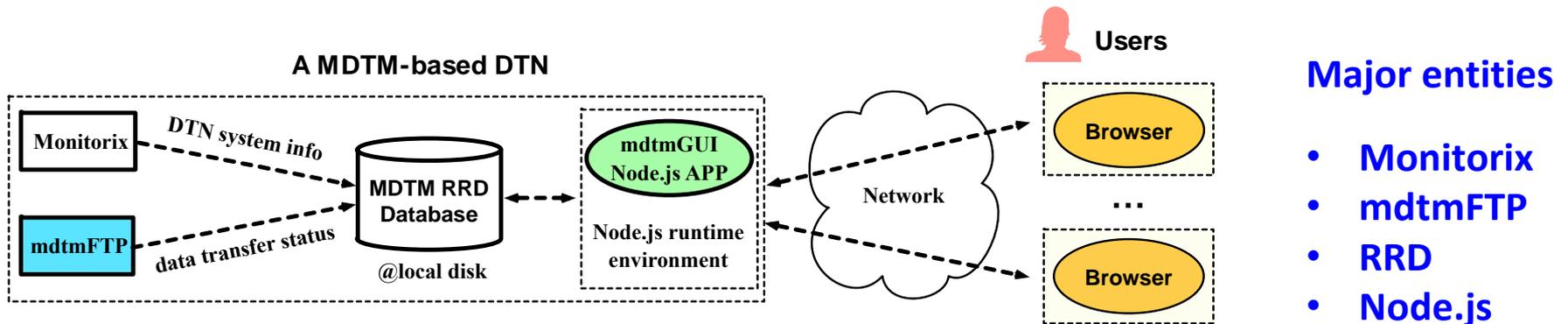
mdtmBBCP and mdtmFTP

- Two MDTM-based data transfer applications
 - mdtmBBCP reuses some BBCP modules for rapid prototyping
 - mdtmFTP reuses some Globus modules for rapid prototyping
- Support features
 - Parallel data transfer with multiple streams
 - Striping data transfer for large files
 - Pipelining data transfer for small files
 - Supporting third-party interaction



mdtmGUI

- Online and real-time monitoring of specific data transfer's status and progress
- Online and real-time monitoring of DTN system status
- <http://mdtm-server.fnal.gov:1337>



mdtmGUI Architecture



MDTM DEMO



MDTM DEMO

- Performance demo
 - Compare MDTM with existing data transfer tools such as GridFTP and BBCP
 - @ESNET 100G Testbed
- mdtmGUI demo
 - Online and real-time monitoring of specific data transfer's status and progress
 - Online and real-time monitoring of DTN system status
 - Between FNAL and BNL



Performance DEMO – 1

Large file transfer

- We use mdtmBBCP, mdtmFTP, GridFTP, and BBCP to transfer a 100GB large file from nersc-tbn-2 to nersc-tbn-1, respectively.
- Performance metric
 - Time-to-Completion



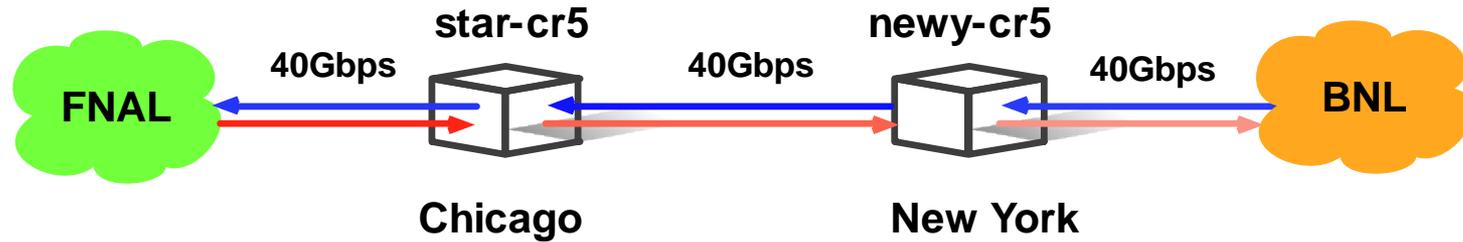
Performance DEMO – 2

Folder file transfer

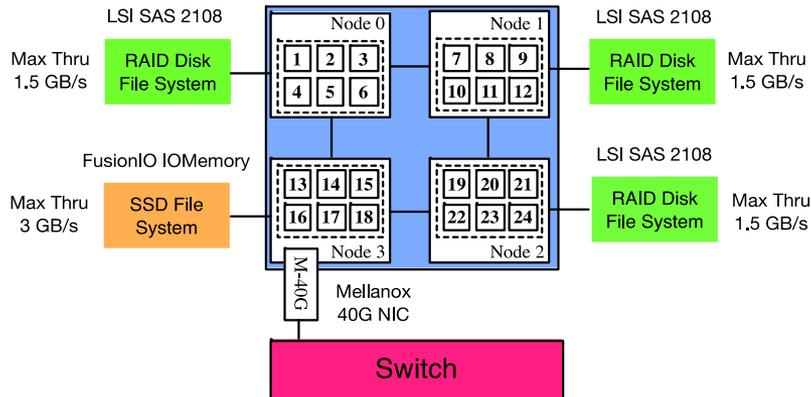
- We use mdtmBBCP, mdtmFTP, GridFTP, and BBCP to transfer a Linux folder from nersc-tbn-2 to nersc-tbn-1, respectively.
- Performance metric
 - Time-to-Completion



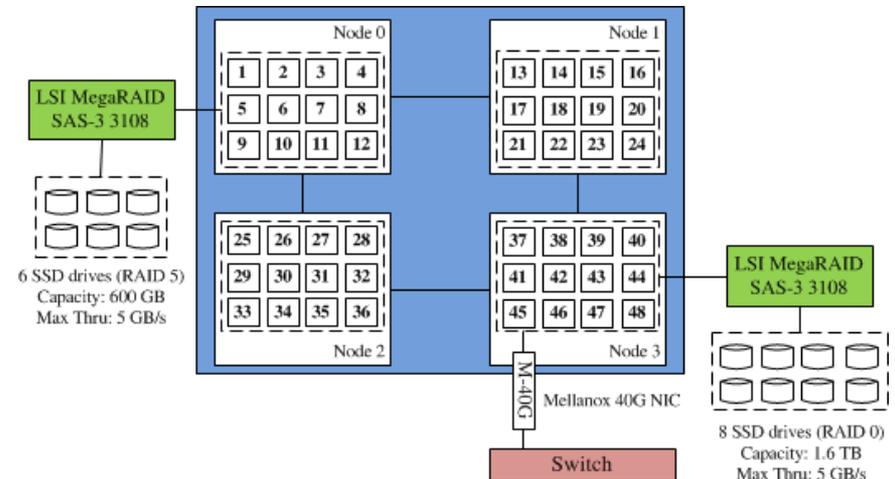
mdtmGUI DEMO - Configurations



Testbed between FNAL and BNL



FNAL DTN



BNL DTN



mdtmGUI DEMO

- We use mdtmBBCP and mdtmFTP to transfer files from FNAL to BNL, respectively.
- We use mdtmGUI
 - to online and real-time monitor specific data transfer's status and progress
 - to online and real-time monitor DTN system status



Questions?

Demo



MDTM Research Team

Wenji Wu (PI)
Liang Zhang
Phil Demar

Dantong Yu (Co-PI)
Dimitrios Katramatos
Shudong Jin
Tan Li