



# Advanced Network Research at Fermilab

Dr. Wenji Wu

ECE Department Distinguished Seminar

September 18, 2019

# Agenda

- About Fermilab
- Why do we need advanced network research at Fermilab?
- Advanced network facilities at Fermilab
- Major research projects
  - WireCap: a novel packet capture engine for commodity NICs in high-speed networks
    - <http://wirecap.fnal.gov>
  - mdmFTP: a high performance data transfer tool (funded by DOE ASCR, \$1.5M)
    - <http://mdm.fnal.gov>
  - BigData Express (funded by DOE ASCR, \$2.2M)
    - <http://bigdataexpress.fnal.gov>
  - Quantum network research (funded by DOE ASCR, \$3.2M)

# Fermilab is America's particle physics and accelerator laboratory.

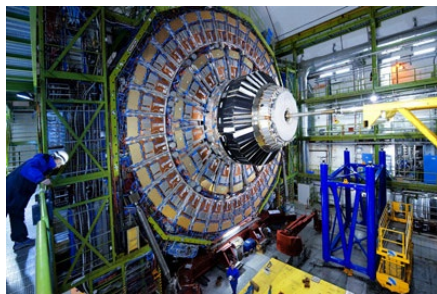
We bring the world together to solve the mysteries of matter, energy, space and time.



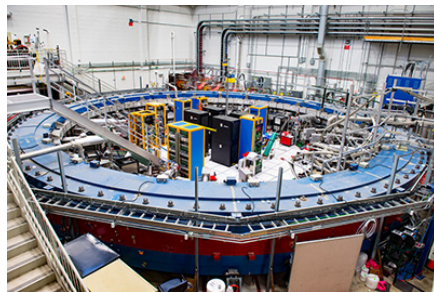
- Over 50 years at the forefront of discovery in high energy physics (top & bottom quarks, tau neutrino,...)
- We build and operate High Power (MW), High reliability Accelerators. Funded by DOE Office of Science.
- Largest U.S. Accelerator complex, 6800 acre site, ~\$400M/yr budget, nearly 1800 staff, > 3200 users.
- Next big project: Long Baseline Neutrino Facility and Deep Underground Neutrino Experiment

# Why Do We Need Advanced network research at Fermilab?

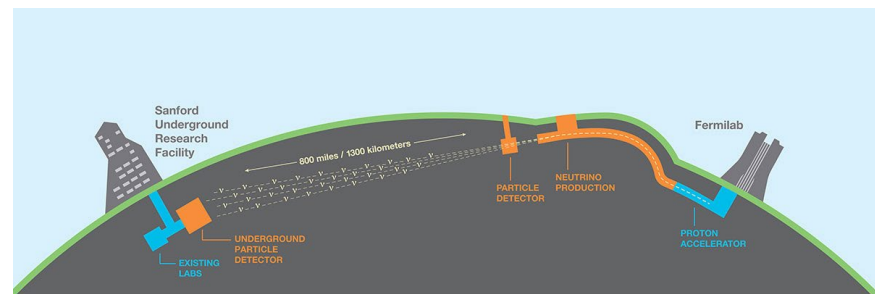
- High-through computing for international particle physics collaborations requires the ability to transport large amount of data quickly around the world



CMS Experiment



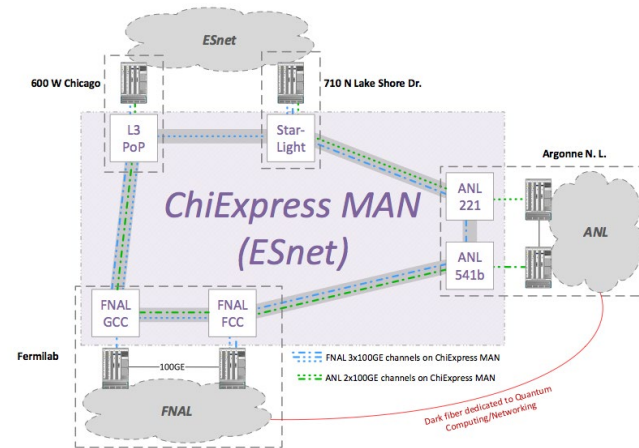
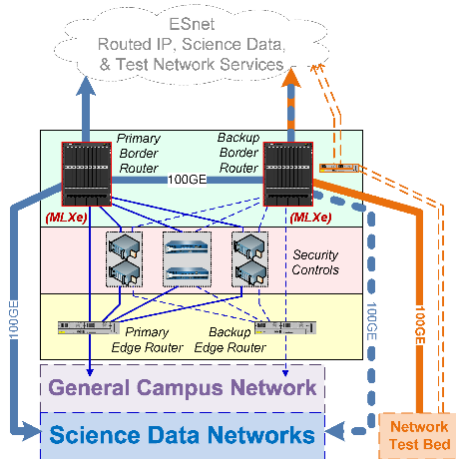
Muons Experiment



LBNF/DUNE Experiment

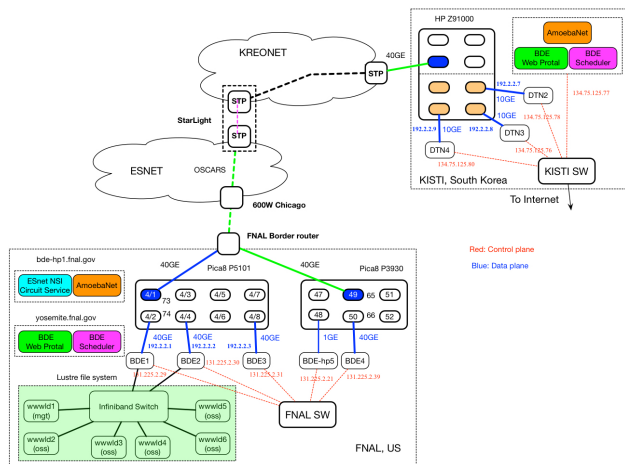
- Fermilab has long engaged in network R&D in support of its mission
  - 100-gigabit connectivity to local, national, and international wide-area networks

# Advanced networking facilities at Fermilab

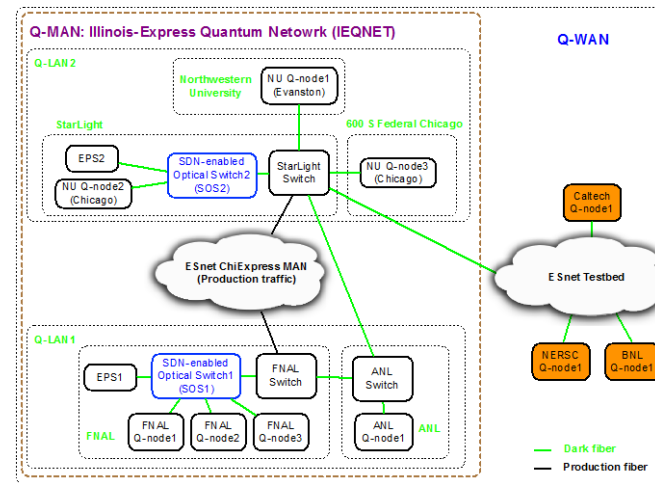


## Fermilab Wide Area Networking Facilities

## ChiExpress MAN



## The Cross-Pacific SDN Testbed



## Illinois-Express Quantum Network (IEQNET)

# Major Network Research Projects that I lead and/or participate in

- WireCap: a novel packet capture engine for commodity NICs in high-speed networks
  - U.S. Patent 20160127276A1
  - <http://wirecap.fnal.gov>
- mdtmFTP: a high performance data transfer tool
  - Funded by DOE ASCR, \$1.5M
  - <http://mdtm.fnal.gov>
- BigData Express
  - Funded by DOE ASCR, \$2.2M
  - <http://bigdataexpress.fnal.gov>
- Quantum network research (funded by DOE ASCR, \$3.2M)

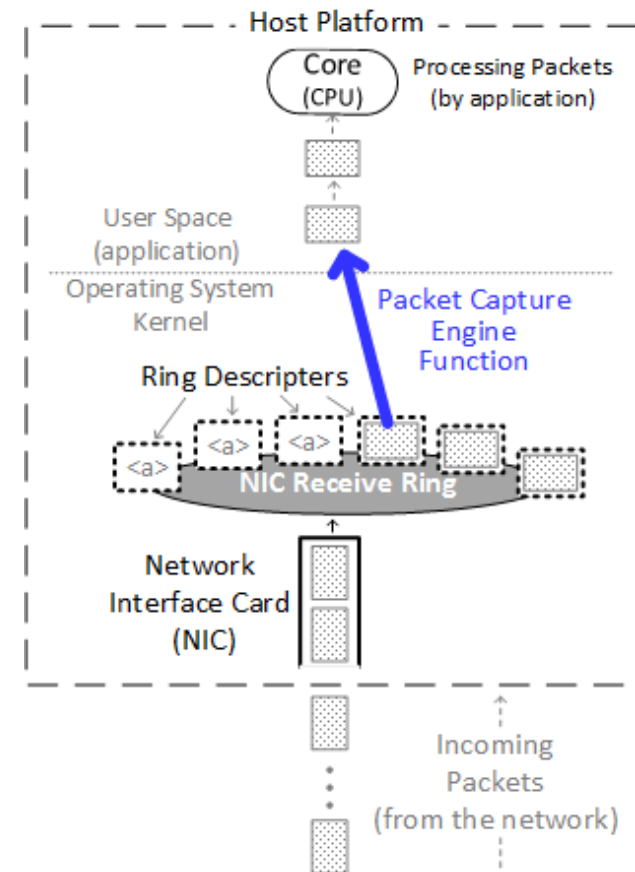
# WireCAP: A Novel Packet Capture Engine for Commodity NICs in High-speed Networks

W. Wu, P. DeMar

# Packet Capture Engine Basics

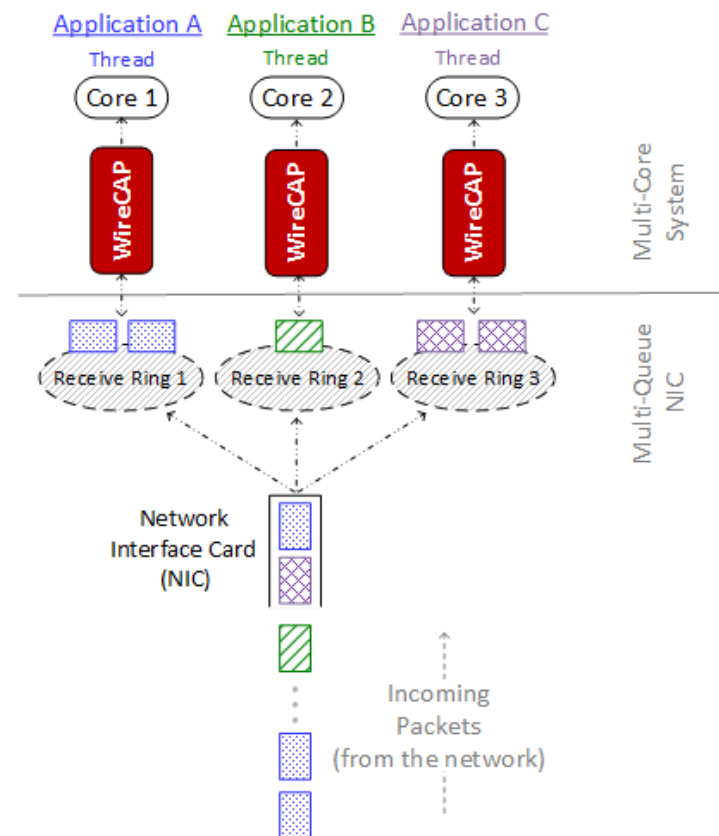
- Capture incoming network data packet for analysis
  - Handle a continuous flow of packets
  - Used by Cyber security and traffic characterization applications
- Implementations found in ASIC, FPGA, and on commodity off-the-shelf (COTS) platforms
- Generic COTS packet capture engine
  - Network interface card (NIC) receives network packets
  - Packets moved to receive ring buffer
  - Packet capture engine provides mechanism(s) to deliver packets from receive ring to application in user space

**No packet drops!!**



# What is WireCAP?

- An innovative packet capture engine:
  - Designed for lossless packet capturing for commodity NICs in high-speed networks
    - Cost-effective
    - Flexible
  - Intended for multicore systems with multi-queue NICs
- Two advanced mechanisms for lossless packet capturing
  - **Ring-buffer-pool mechanism**
    - For short-term traffic burstiness
  - **Buddy-group mechanism**
    - For long-term traffic flooding



# Miscellaneous

- Project website: <http://wirecap.fnal.gov>
- Paper
  - Wenji Wu, Phil DeMar, “[WireCAP: a Novel Packet Capture Engine for Commodity NICs in High-speed Networks](#),” [IMC'14](#), November 05 – 07 2014, Vancouver, BC, Canada.
- Patent
  - U.S. Patent 20160127276A1, filed November 4, 2015, and issued September 18, 2018.
- Issued to multiple agencies, including
  - U.S. Army Research Lab, U.S. Naval Research Lab, U.S. Air Force Research Lab, LANL, ...



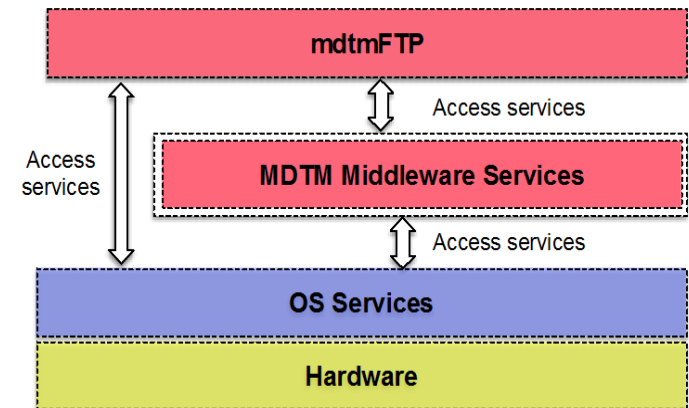
# **mdtmFTP: a high-performance data transfer tool**

**W. Wu, L. Zhang, P. DeMar, L. Carpenter**

# mdtmFTP: a high-performance data transfer tool

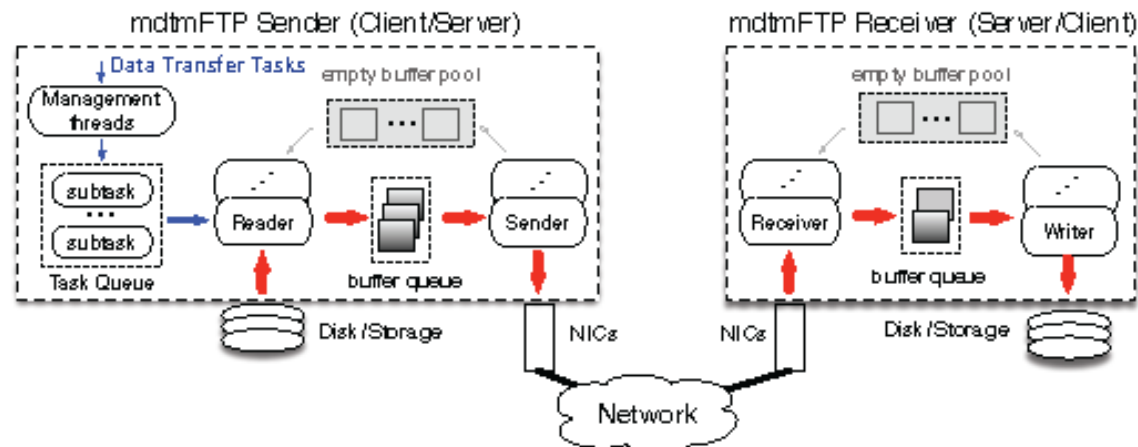
- Pipelined I/O-centric design to streamline data transfer
- Multicore-aware data transfer middleware (MDTM) optimizes use of underlying multicore system
- Extremely efficient in transferring of Lots of Small Files (LOSF)
- Various optimization mechanisms
  - Zero copy
  - Asynchronous I/O
  - Batch processing

A DOE/SC/ASCR-sponsored research project  
Software is available at: <http://mdtm.fnal.gov>



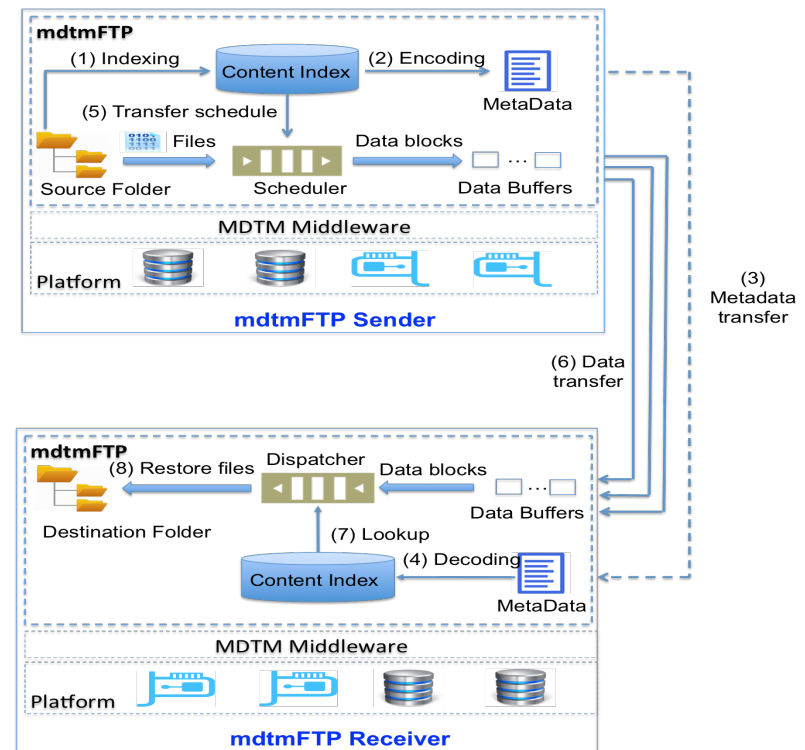
# mdtmFTP – A pipelined I/O centric design

- Dedicated I/O threads to perform network/disk I/Os in parallel
- MDTM middleware to schedule cores for I/O threads
- Advanced data buffer mechanism to improve I/O performance
- Data transfers are executed in a pipelined manner
- A data transfer task is split into multiple subtasks
- Subtasks are executed in parallel



# mdtmFTP – A large virtual file transfer mechanism

- Treat a dataset as a large “virtual file”.
- Each file in the dataset is treated as a file segment in the virtual file, and sequentially “added” to the virtual file.
- The virtual file is logically, instead of physically, created.
  - Different than Tar-based solutions
- The whole data set is transferred continuously and seamlessly as a single virtual file.
  - Different than GridFTP’s per-file-based mechanisms (e.g., pipelining, concurrency)
  - Avoid protocol processing on a per-file basis
  - Allow for batch processing small files in the sender/receiver to optimize I/O performance



# LSST SC'18 News Release

The screenshot shows the LSST website with a news article titled "LSST 100 Gbps Network Demonstration at Supercomputing Conference 2018". The article is dated Tuesday, November 20, 2018. It describes a successful demonstration of data transfer capabilities of fiber optic networks. The article includes three images: "M1M3 Mirror Lifted Onto Cell", "M1M3 Out of Storage at Last", and "M1M3 Cell Moves to the Richard F. Caris Mirror Lab". The article also mentions the use of the Fermilab Multicore-Aware Data Transfer Middleware (MDTM) software and the National Optical Astronomy Observatory (NOAO) public data (FITS files).

Highlights

**LSST 100 Gbps Network Demonstration at Supercomputing Conference 2018**

Tuesday, November 20, 2018

November 20, 2018 - The LSST Network Engineering Team (NET) had a strong presence at the Supercomputing 2018 Conference (SC18) in Dallas, TX, last week, including a successful demonstration of the data transfer capabilities of the fiber optic networks that will be used during LSST operations. Digital data were transferred from the Base Site in La Serena, Chile, to the LSST Data Facility at the National Center for Supercomputing Applications (NCSA) in Champaign, IL. During the data transfer demonstration, a peak rate of 100 Gigabits/second (Gb/s) was achieved for short periods, and a sustained rate of 80 Gb/s was achieved over a three hour period, exceeding the test target. This test was run over links provisioned by several networking organizations: REUNA from La Serena to Santiago, FIU/Amlight from Santiago to Miami, SCinet from Miami to Chicago (Starlight), and NCSA from Chicago to Champaign. SCinet links provided by CenturyLink and internetz were used to transfer the data from Miami to Chicago because LSST 100 Gb/s links will not be available in that path until FY20. All of the other links were those that will be used by LSST during operations.

Data Transfer Nodes (DTN) configured in La Serena and Champaign with nuttcp (a network performance measurement tool) generated a sustained memory-to-memory data rate over 80 Gb/s, over a period of three hours. Simultaneously, the DTNs, using the Fermilab Multicore-Aware Data Transfer Middleware (MDTM) software, achieved a peak of 36 Gb/s transferring 200 Gigabytes of DECam public data (FITS files) provided by the National Optical Astronomy Observatory (NOAO). Note that in LSST operations, there will be over 20 DTNs (aka archiver/forwarders) simultaneously sending data, so each one will require far less than 36 Gb/s. In addition, on the Champaign end the files were ingested into a GPFS shared file system, and a Jupyter Notebook running an application provided by LSST Data Management was used to visualize the files. Finally, an additional test transfer from Champaign to La Serena is being conducted and has so far achieved a peak of 40 Gb/s, sufficient for the annual transfer of LSST Data Releases to Chile.

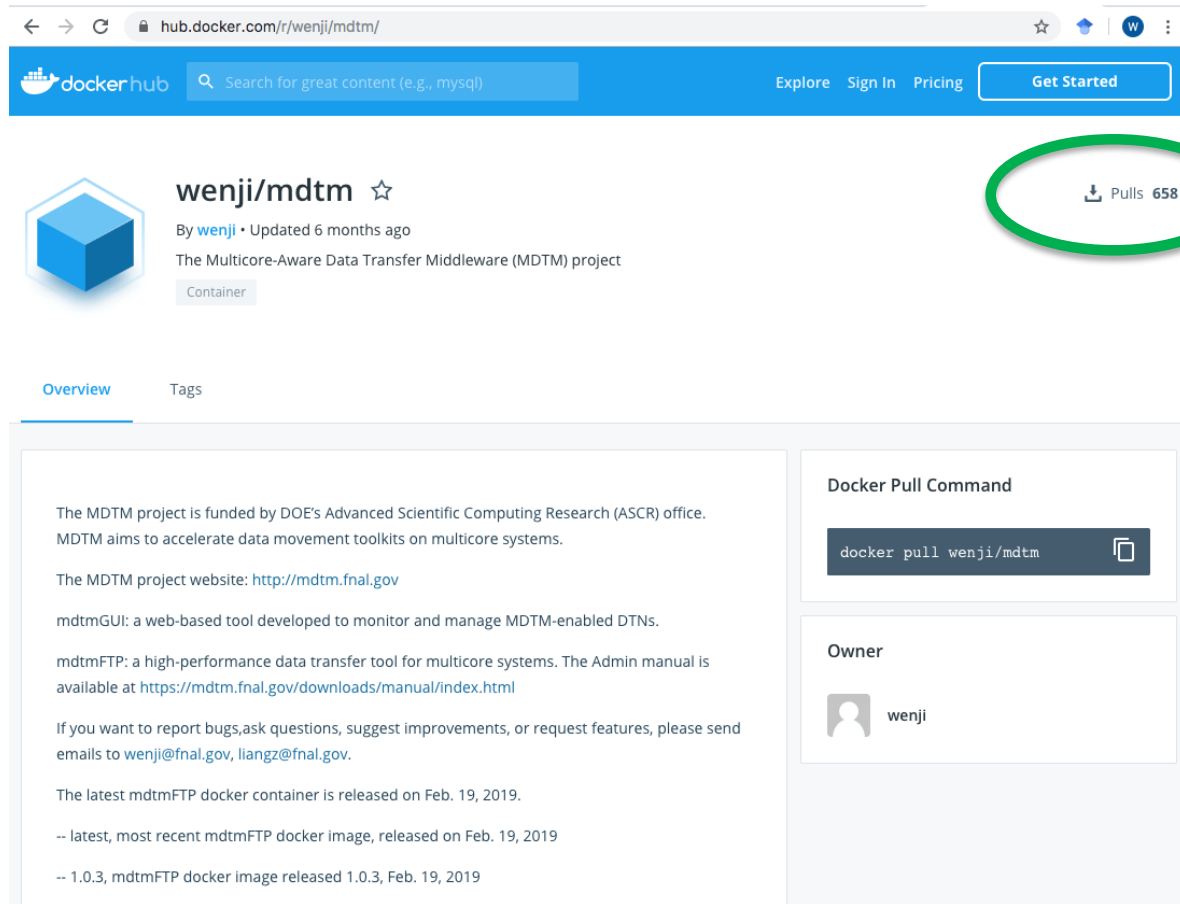
Instrumentation in the DTNs and links and Grafana software were used to provide a real-time, web display of network performance during the demonstration. This was monitored live from the NCSA booth at the Supercomputing 2018 Conference. A number of conference attendees witnessed the demonstration and presentation, and participated in a question and answer session.

According to LSST NET Lead Jeff Kantor, "This demonstration shows not only that we have continuity and performance from the network point of view, but also that all of the partners acted as a very well-coordinated team."

Follow us     

“Simultaneously, the DTNs, using the Fermilab Multicore-Aware Data Transfer Middleware (MDTM) software, achieved a peak of 36 Gb/s transferring 200 Gigabytes of DECam public data (FITS files) provided by the National Optical Astronomy Observatory.”

# mdtmFTP Statistics



The screenshot shows the Docker Hub page for the `wenji/mdtm` container image. The page header includes the Docker Hub logo, a search bar, and navigation links for Explore, Sign In, Pricing, and a Get Started button. The main content area displays the repository name `wenji/mdtm` with a star icon, the creator `wenji`, and the update date "Updated 6 months ago". Below this, a description states: "The Multicore-Aware Data Transfer Middleware (MDTM) project". A green circle highlights the "Pulls 658" statistic. The "Overview" tab is selected, showing a description of the MDTM project, its website (<http://mdtm.fnal.gov>), and information about `mdtmGUI` and `mdtmFTP`. The "Owner" section shows the user `wenji`. The "Docker Pull Command" section displays the command `docker pull wenji/mdtm`.

hub.docker.com/r/wenji/mdtm/

dockerhub Search for great content (e.g., mysql) Explore Sign In Pricing Get Started

wenji/mdtm ☆

By wenji • Updated 6 months ago

The Multicore-Aware Data Transfer Middleware (MDTM) project

Container

Overview Tags

The MDTM project is funded by DOE's Advanced Scientific Computing Research (ASCR) office. MDTM aims to accelerate data movement toolkits on multicore systems.

The MDTM project website: <http://mdtm.fnal.gov>

mdtmGUI: a web-based tool developed to monitor and manage MDTM-enabled DTNs.

mdtmFTP: a high-performance data transfer tool for multicore systems. The Admin manual is available at <https://mdtm.fnal.gov/downloads/manual/index.html>

If you want to report bugs,ask questions, suggest improvements, or request features, please send emails to [wenji@fnal.gov](mailto:wenji@fnal.gov), [liangz@fnal.gov](mailto:liangz@fnal.gov).

The latest mdtmFTP docker container is released on Feb. 19, 2019.

-- latest, most recent mdtmFTP docker image, released on Feb. 19, 2019

-- 1.0.3, mdtmFTP docker image released 1.0.3, Feb. 19, 2019

Docker Pull Command

```
docker pull wenji/mdtm
```

Owner

wenji

658 downloads



<https://hub.docker.com/r/wenji/mdtm/>

# BigData Express

# Many people's hard work

FNAL: Qiming Lu, Liang Zhang, Sajith Sasidharan,  
Wenji Wu, Phil DeMar

ESnet: Chi Guok, John Macauley, Inder Monga

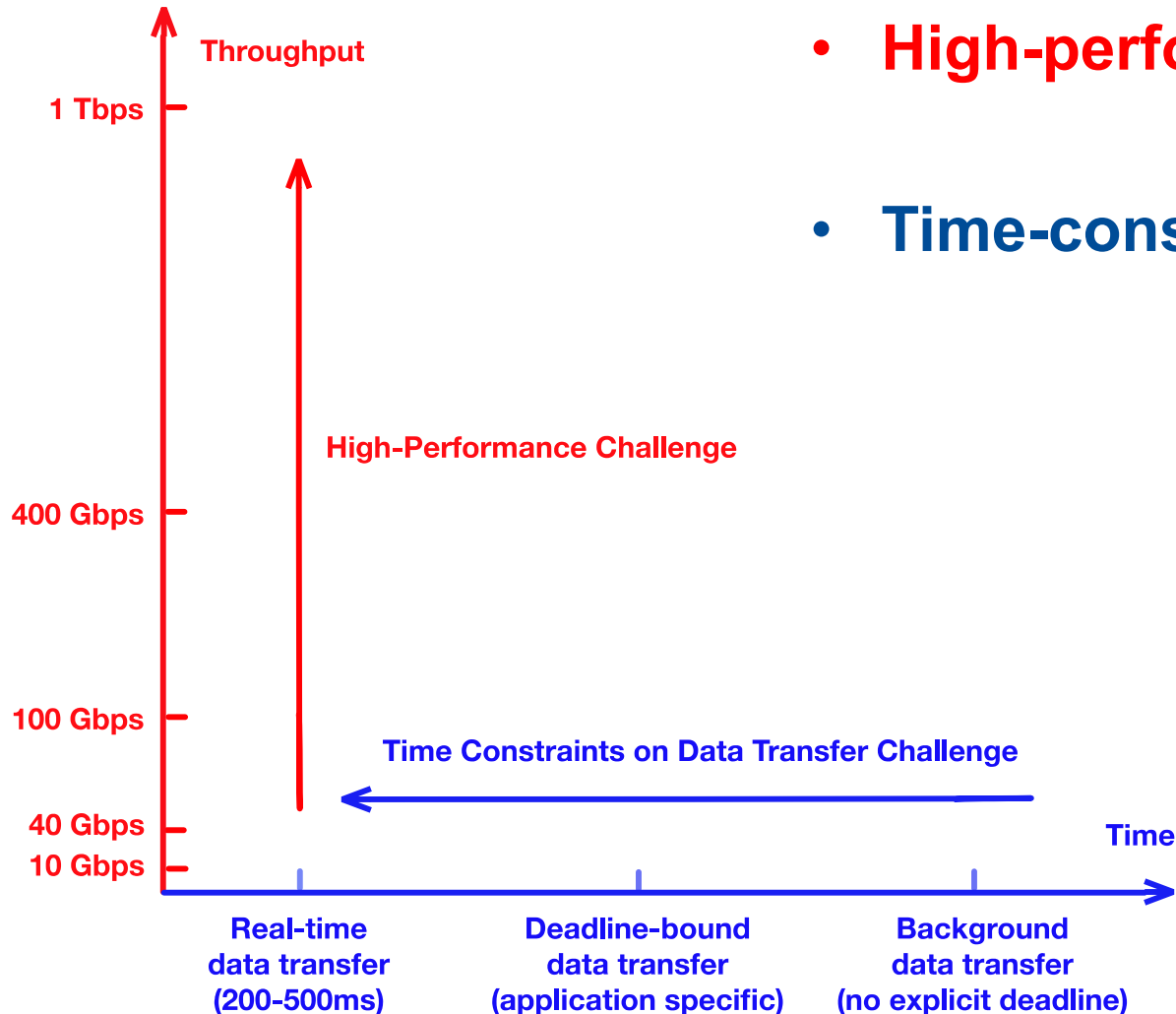
iCAIR/StarLight: Se-young Yu, Jim-Hao Chen, Joe Mambretti

KISTI: Jin Kim, Seo-Young Noh

UMD/MAX: Xi Yang, Tom Lehman

# Data Transfer Challenges in BigData Era

- **High-performance challenges**
- **Time-constraint challenges**



# Data Transfer – State of the Art

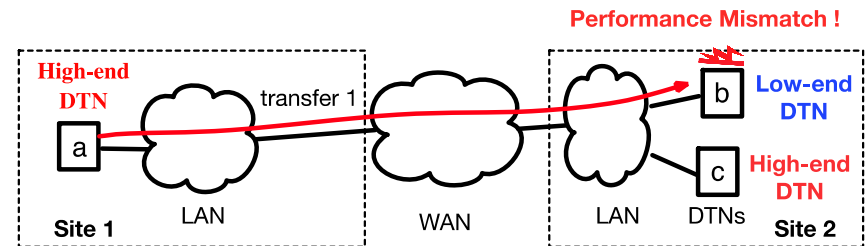
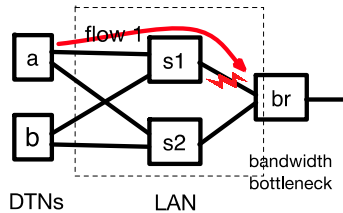
- Advanced data transfer tools and services developed
  - GridFTP, BBCP
  - PhEDEx, LIGO Data Replicator, Globus Online
- Numerous enhancements
  - Parallelism at all levels
    - Multi-stream, Multicore, Multi-path parallelism
  - Science DMZ architecture
  - Terabit networks

# Problems with Existing Data Transfer Tools & Services

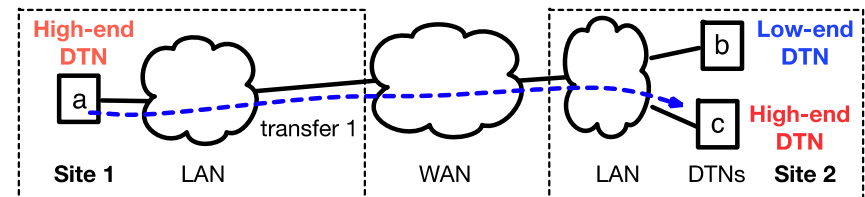
- Disjoint end-to-end data transfer loop
- Cross-interference between data transfers
- Oblivious to user requirements (e.g., deadlines and QoS requirements)
- Inefficiencies arise with existing data transfer tools running on DNTs

# Problem 1 – Disjoint end-to-end data transfer loop

- Distributed resource management model
  - Resource contention
  - Performance mismatch



a. without coordination



b. with coordination

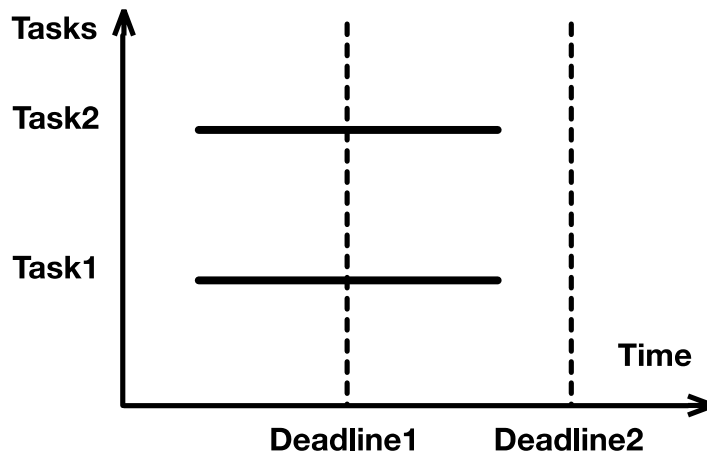
## Problem 2 – Cross-interference between data transfers



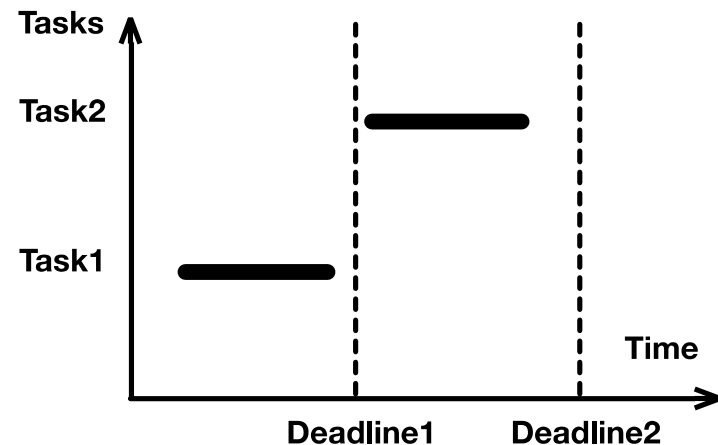
- Degraded performance
- High variability in data transfer performance

## Problem 3 – Oblivious to user requirements

- Data transfer jobs are scheduled on a first-come, first serve basis
  - Without deadline awareness
- Resources are shared fairly among data transfer jobs



a. without deadline awareness

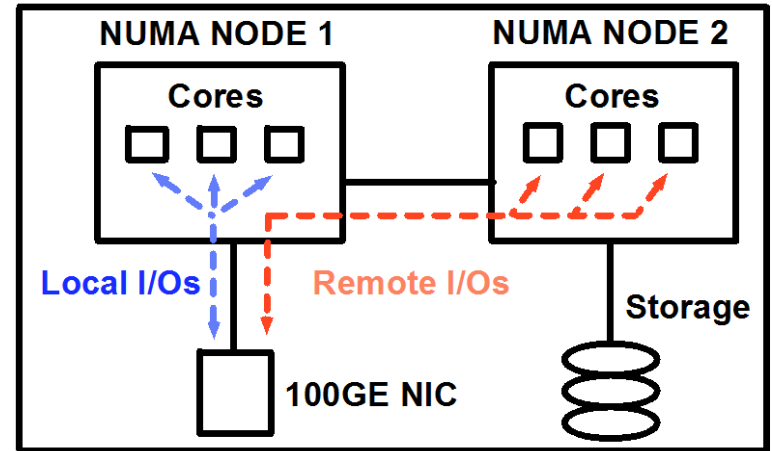


b. with deadline awareness

## Problem 4 – Inefficiencies arises when existing data transfer tools run on DTNs

- I/O locality on NUMA systems
- Cache thrashing
- Scheduling overheads

...



**Need high-performance data transfer tool!**

# Our Solution – **BigData Express**

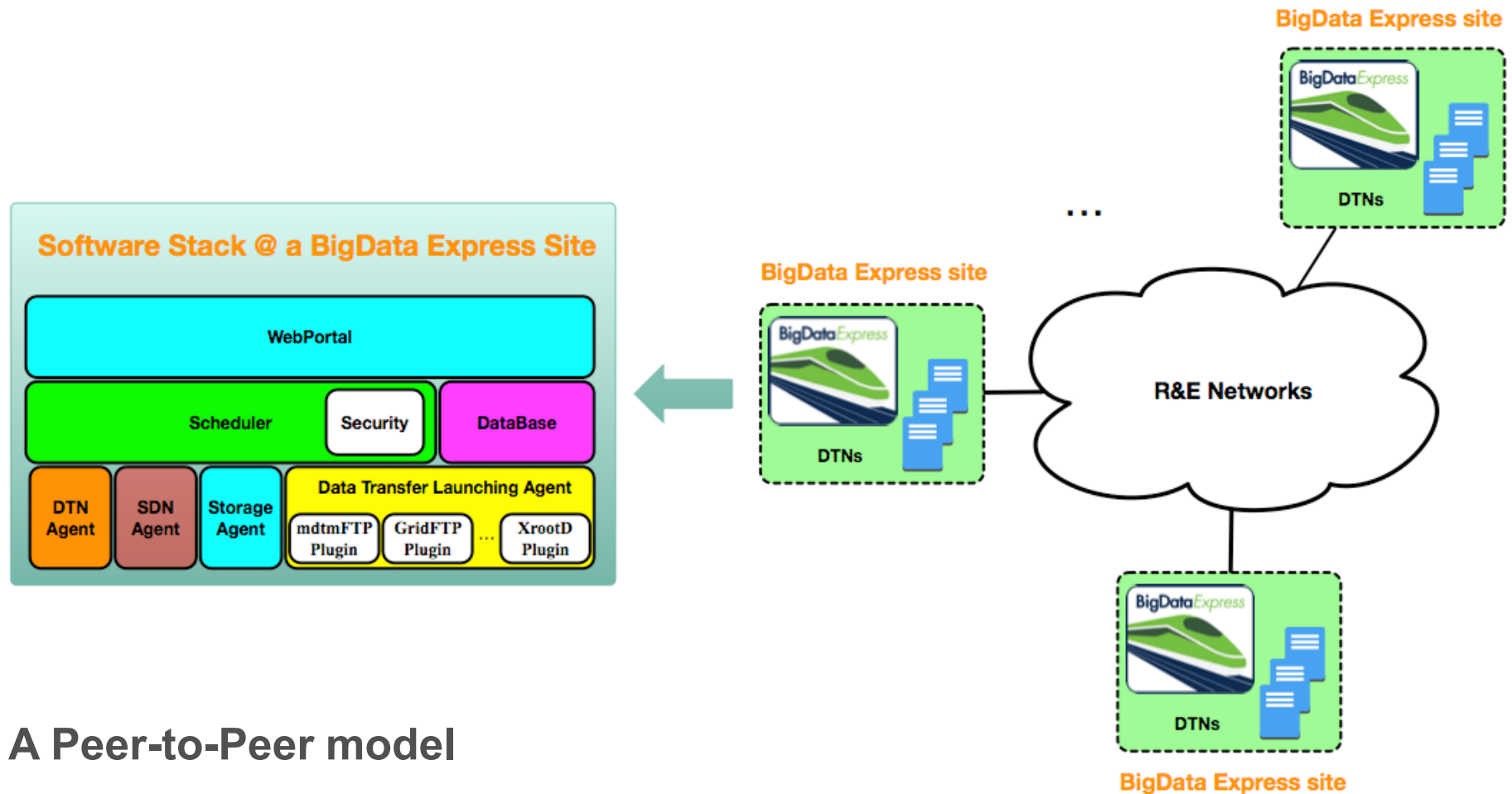


- BigData Express: a schedulable, predictable, and high-performance data transfer service
  - A peer-to-peer, scalable, and extensible data transfer model
  - A visually appealing, easy-to-use web portal
  - A high-performance data transfer engine
  - A time-constraint-based scheduler
  - On-demand provisioning of end-to-end network paths with guaranteed QoS
  - Robust and flexible error handling
  - CILogon-based security

# BigData Express Major Components

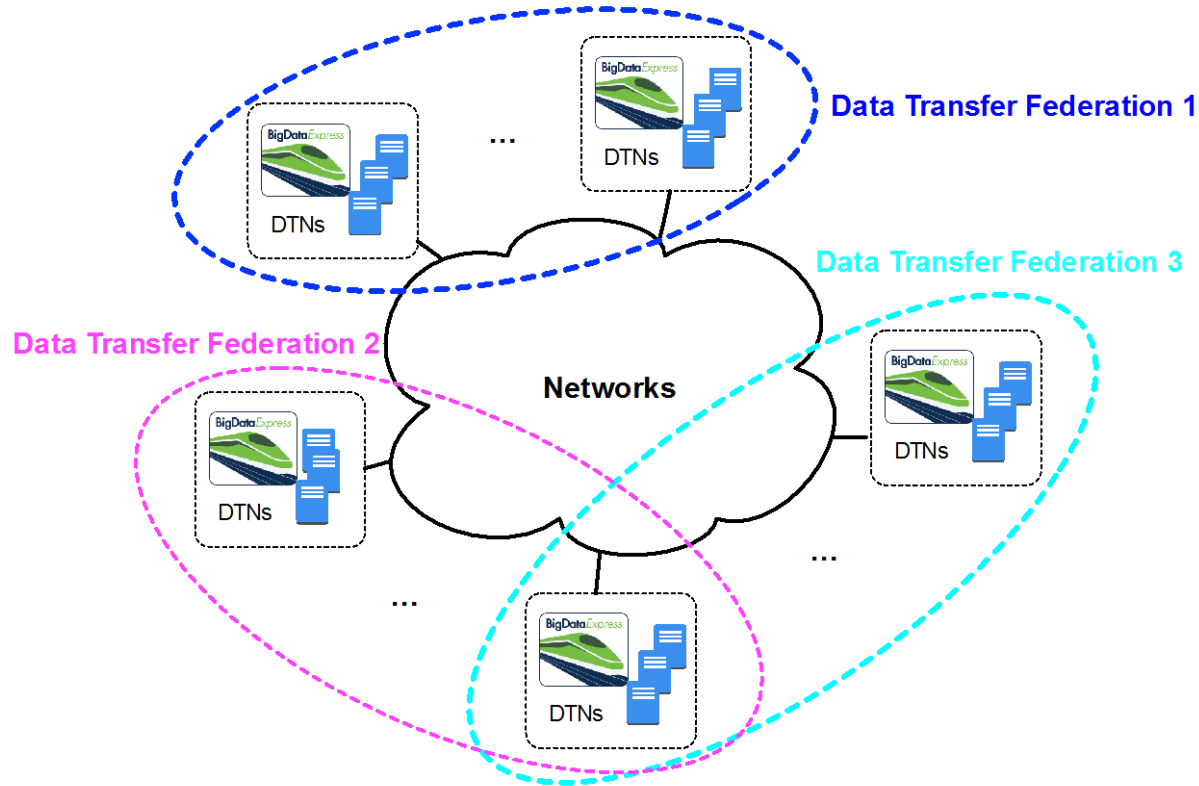
- **BigData Express Web Portal**
  - Access to BigData Express services
- **BigData Express Scheduler**
  - Time-constraint-based scheduler
  - Co-scheduling DTN, storage, & network
- **AmoebaNet**
  - Network as a service
  - Rate control
- **mdtmFTP**
  - High-performance data transfer engine
  - <http://mdtm.fnal.gov>
- **Data Transfer Launching Agent**
  - Launch data transfer jobs
  - Support different data transfer protocols
- **DTN Agent**
  - Manage and configure DTNs
  - Collect & report DTN configuration and status
- **Storage Agent**
  - Manage and configure storage systems
  - I/O estimation

# BigData Express -- Distributed



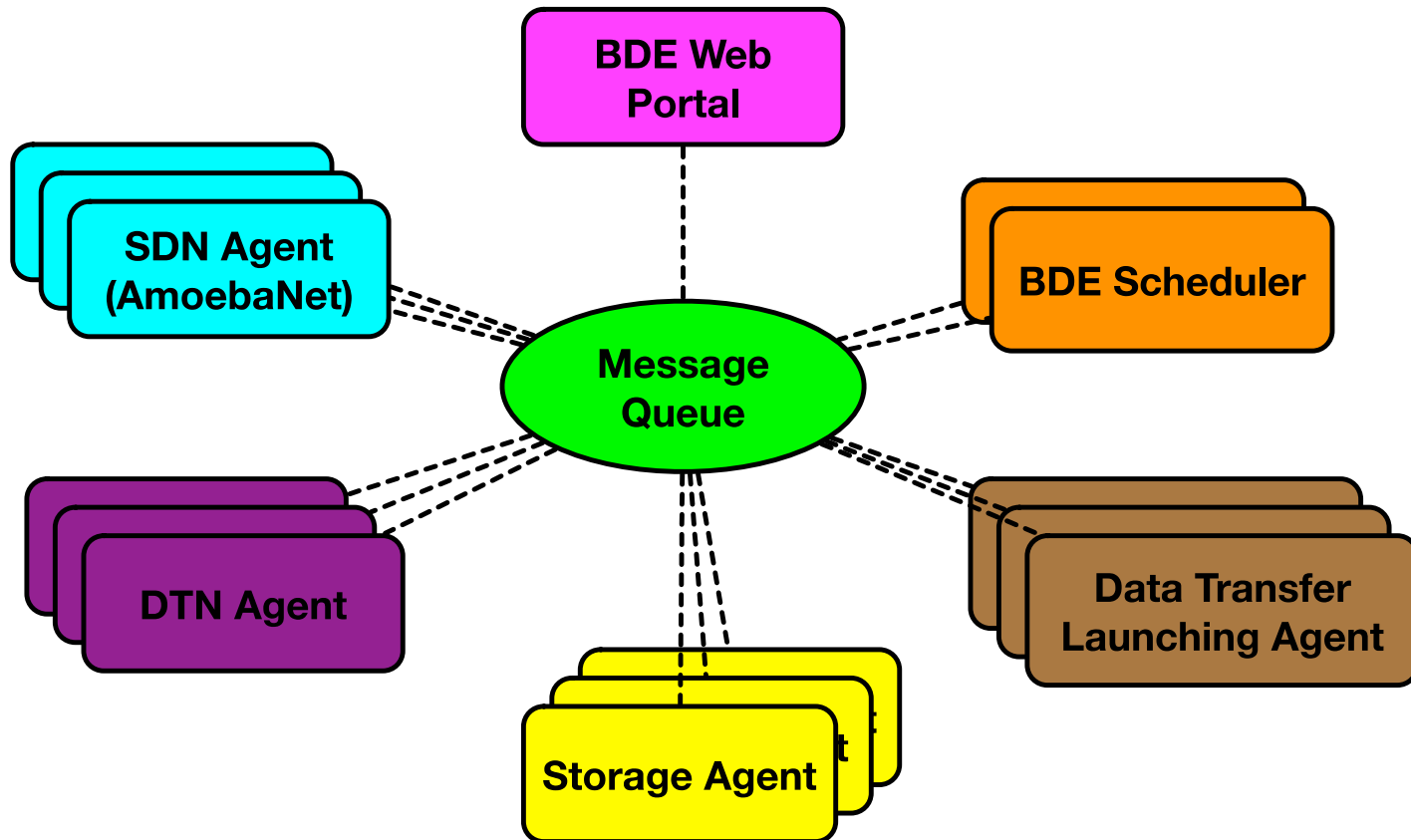
A Peer-to-Peer model

# BigData Express -- Flexible



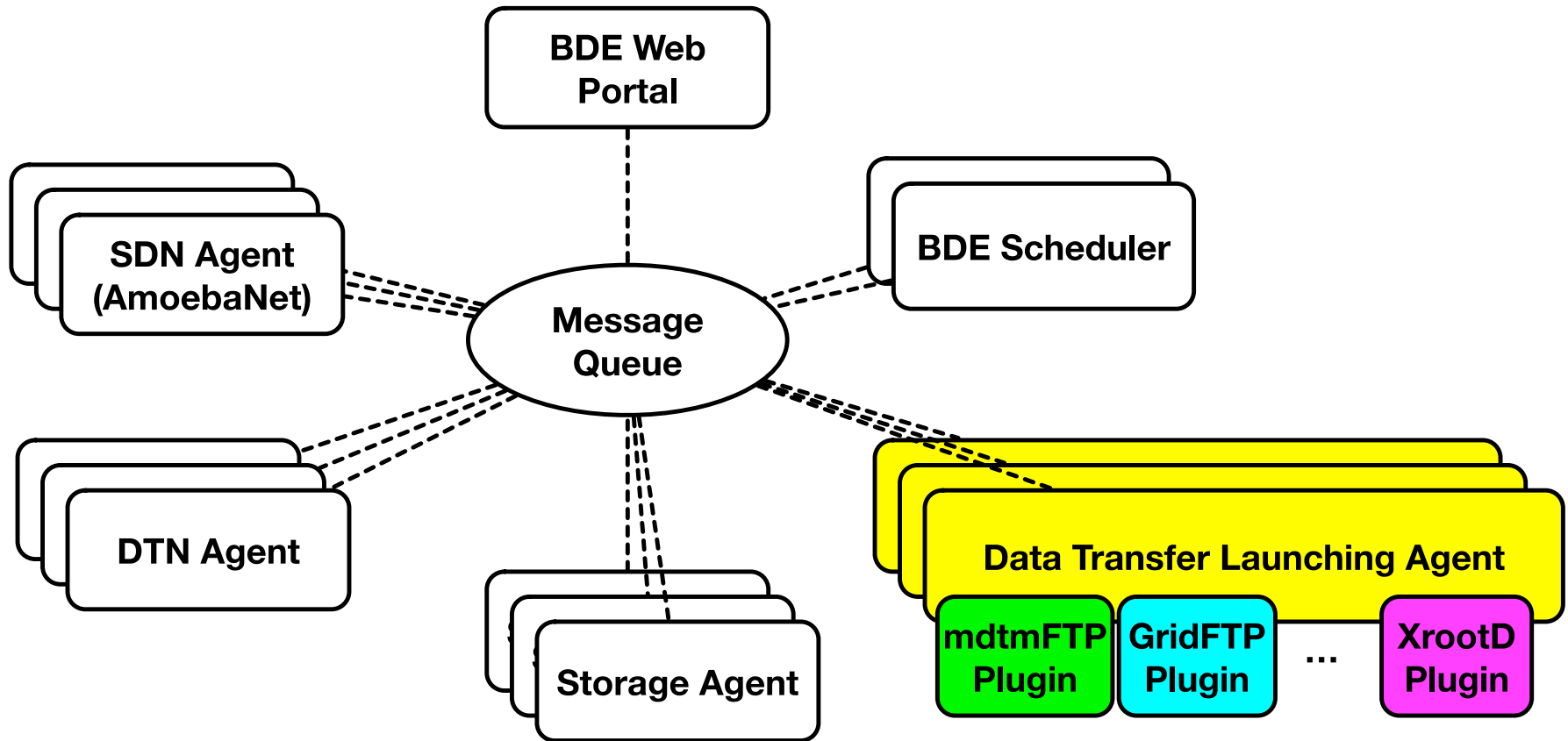
- Flexible to set up data transfer federations
- Providing inherent support for incremental deployment

# BigData Express – Scalable



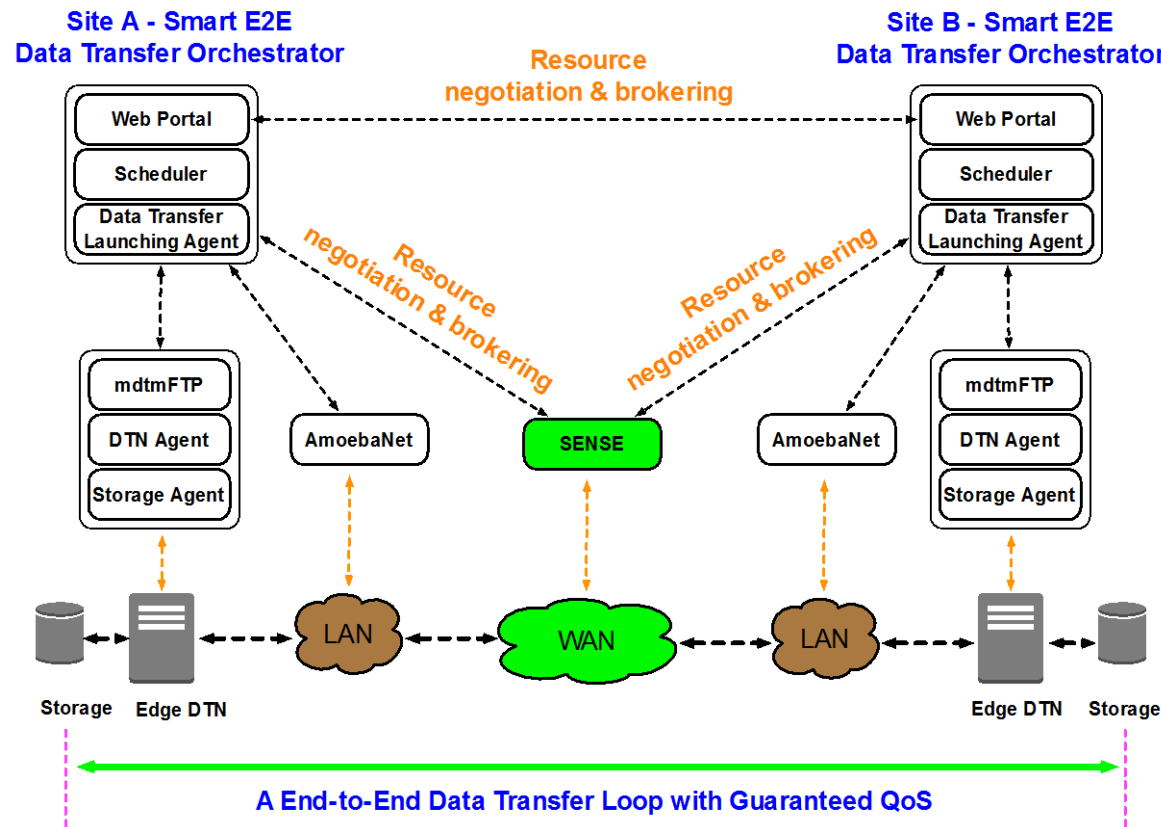
- BigData Express scheduler manages site resources through agents
- Use MQTT as message bus

# BigData Express – Extensible



- Extensible plugin framework to support various data transfer protocols
  - mdtmFTP, GridFTP, XrootD, ...

# BigData Express – End-to-End Data Transfer Model



- Application-aware network service
- Fast-provisioning of end-to-end network paths with guaranteed QoS
- Distributed resource negotiation & brokering

# BigData Express – High Performance Data Transfer

	mdtmFTP	FDT	GridFTP	BBCP
Large file data transfer (1 X 100G)	74.18	79.89	91.18	Poor
Folder data transfer (30 x 10G)	192.19	217	320.17	Poor
Folder data transfer (Linux 3.12.21)	10.51	-	1006.02	Poor

**Time-to-completion (Seconds) – Client/Server mode**      **Lower is better**

	mdtmFTP	FDT	GridFTP	BBCP
Large file data transfer (1 X 100G)	34.976	N/A	106.84	N/A
Folder data transfer (30 x 10G)	95.61	N/A	-	N/A
Folder data transfer (Linux 3.12.21)	9.68	N/A	-	N/A

**Time-to-completion (Seconds) – 3rd party mode**      **Lower is better**

Note 1: “-” indicates inability to get transfer to work

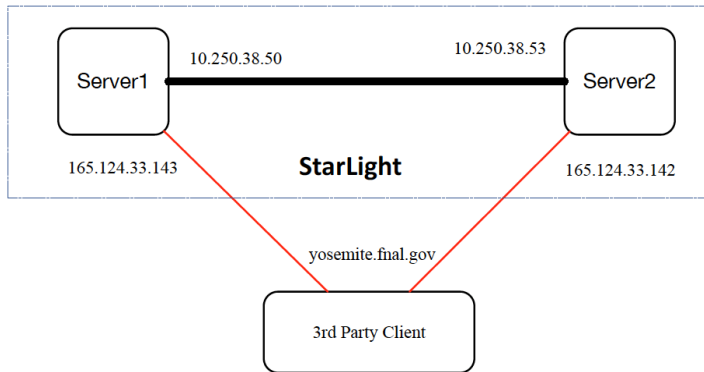
Note 2: BBCP performance is very poor, we do not list its results here

Note 3: BBCP and FDT support 3<sup>rd</sup> party data transfer. But BBCP and FDT couldn't run 3<sup>rd</sup> party data transfer on ESNET testbed due to testbed limitation

**mdtmFTP is faster than existing data transfer tools, ranging from 8% to 9500%!**  
**@ESnet 100GE SDN Testbed,**

# BigData Express – High Performance Data Transfer

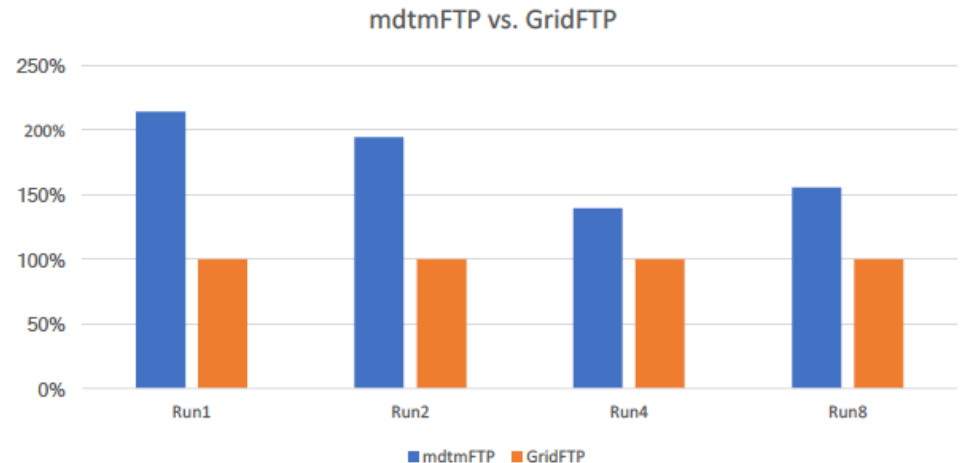
**STARLIGHT<sup>SM</sup>SDX**



**StarLight 100GE Testbed**

Performance – Aggregate throughput

Gb/s	Run1	Run2	Run4	Run8
GridFTP	6.2Gbps	12.24Gbps	20.35Gbps	28.32 Gbps
mdtmFTP	13.27Gbps	23.80Gbps	28.354Gbps	43.94 Gbps



**mdtmFTP is faster than GridFTP, ranging from 40% to 114%!  
@StarLight 100GE Testbed**

# BigData Express – Mechanism Summary

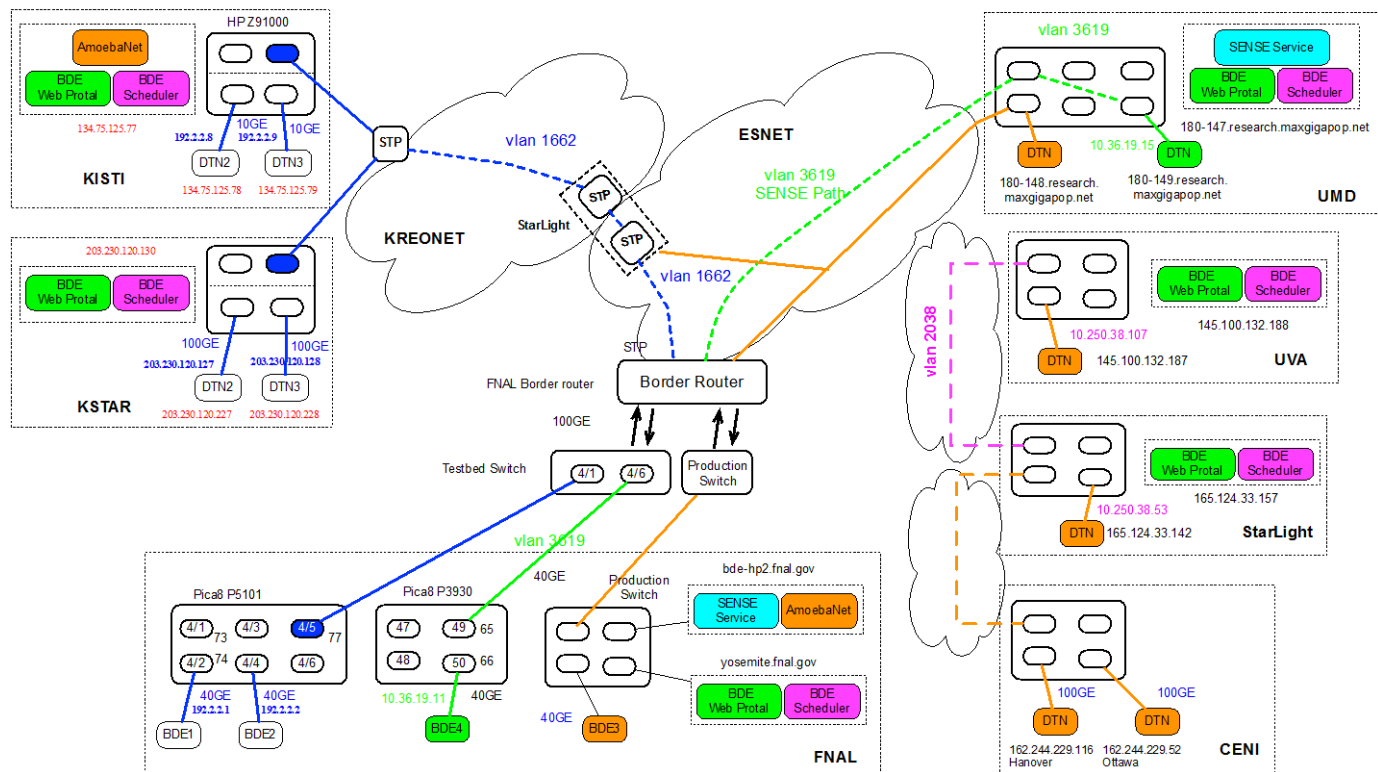
Problems with existing data transfer tools	BigData Express Solutions
<ul style="list-style-type: none"><li>Disjoint end-to-end data transfer loop</li></ul>	<ul style="list-style-type: none"><li>Distributed resource negotiation &amp; brokering</li><li>Co-scheduling of DTN, storage, &amp; networking</li><li>On-demand provisioning of end-to-end network path with guaranteed QoS</li></ul>
<ul style="list-style-type: none"><li>Cross-interference between data transfers</li></ul>	<ul style="list-style-type: none"><li>Time-constraint-based scheduler</li><li>Admission control</li><li>Rate control</li></ul>
<ul style="list-style-type: none"><li>Oblivious to user requirements</li></ul>	<ul style="list-style-type: none"><li>Time-constraint-based scheduler</li><li>Three classes of data transfer</li></ul>
<ul style="list-style-type: none"><li>Inefficiencies arises when existing data transfer tools run on DTNs</li></ul>	<ul style="list-style-type: none"><li>mdtmFTP – A high-performance data transfer engine</li></ul>

# BigData Express vs. Globus Online

Features	BigData Express	Globus Online
Architecture	<ul style="list-style-type: none"><li>• Distributed service</li><li>• Flexible to set up data transfer federations</li></ul>	<ul style="list-style-type: none"><li>• Centralized service</li></ul>
Supported Protocols	<ul style="list-style-type: none"><li>• Extensible plugin framework to support multiple protocols:<ul style="list-style-type: none"><li>○ mdmFTP</li><li>○ GridFTP, XrootD, SRM (coming soon)</li></ul></li></ul>	<ul style="list-style-type: none"><li>• GridFTP</li></ul>
SDN Support	<ul style="list-style-type: none"><li>• Yes, Network as a service</li><li>• Fast-provisioning end-to-end network paths with guaranteed QoS</li></ul>	<ul style="list-style-type: none"><li>• Not in production</li></ul>
Supported Data Transfers	<ul style="list-style-type: none"><li>• Real-time data transfer</li><li>• Deadline-bound data transfer</li><li>• Best-effort data transfer</li></ul>	<ul style="list-style-type: none"><li>• Best-effort data transfer</li></ul>
Error Handling	<ul style="list-style-type: none"><li>• Checksum</li><li>• Retransmit</li></ul>	<ul style="list-style-type: none"><li>• Checksum</li><li>• Retransmit</li></ul>



# BigData Express SC18 DEMO



# BigData Express – Deployment

- **Asia**

- KISTI, South Korea
- KSTAR



- **Europe**

- University of Amsterdam, Netherlands
- CERN (coming soon)



- **North America**

- Fermilab
- StarLight, Northwestern University
- UMD/MAX, University of Maryland, College Park
- Ciena (Canada)
  - US East
  - CA East



- **Australia & Pacific areas**

- National Computational Infrastructure (NCI)
- PAWSEY supercomputing center (coming soon)



# Quantum Network Research

# Quantum Network Research



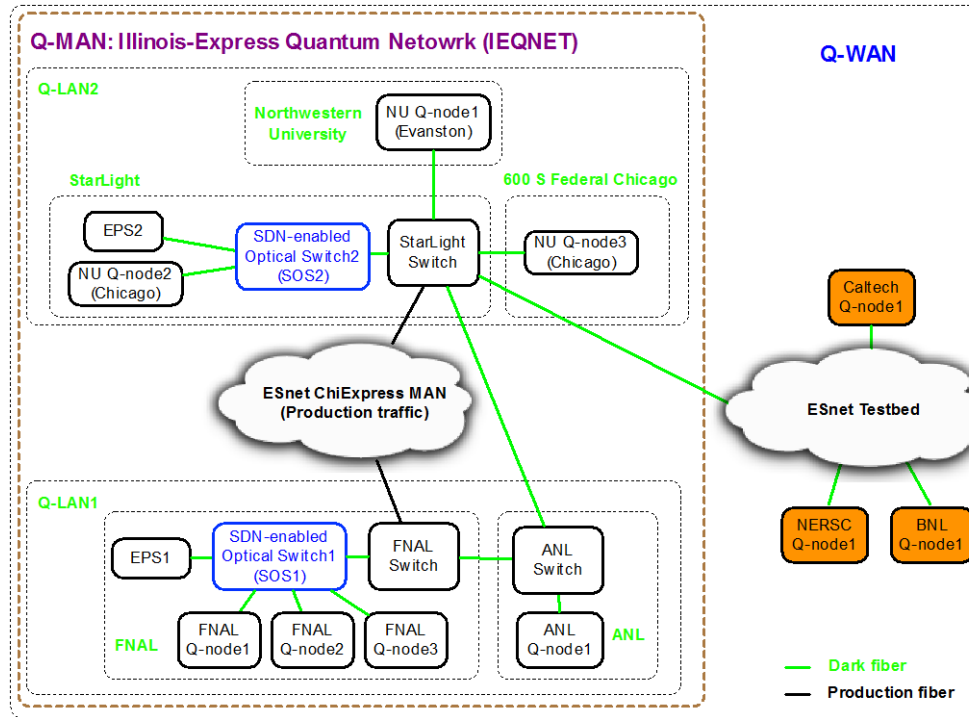
Fermilab has an exciting and promising quantum science program, led by **Panagiotis Spentzouris**

<https://qis.fnal.gov/>

# Illinois-Express Quantum Network (IEQNET)

- Aim to build a metropolitan-scale quantum network testbed that demonstrates important advanced quantum networks capabilities beyond the lab.
- Funded by DOE ASCR, \$3.2M, announced on Sep, 2019.
- Research team
  - Fermilab (lead)
    - P. Spentzouris (PI), C. Pena, W. Wu
  - Caltech
    - M. Spiropulu, N. Lauk
  - Northwestern University
    - P. Kumar, G. Kanter
  - Argonne National Lab
    - J. Chung, R. Kettimuthu

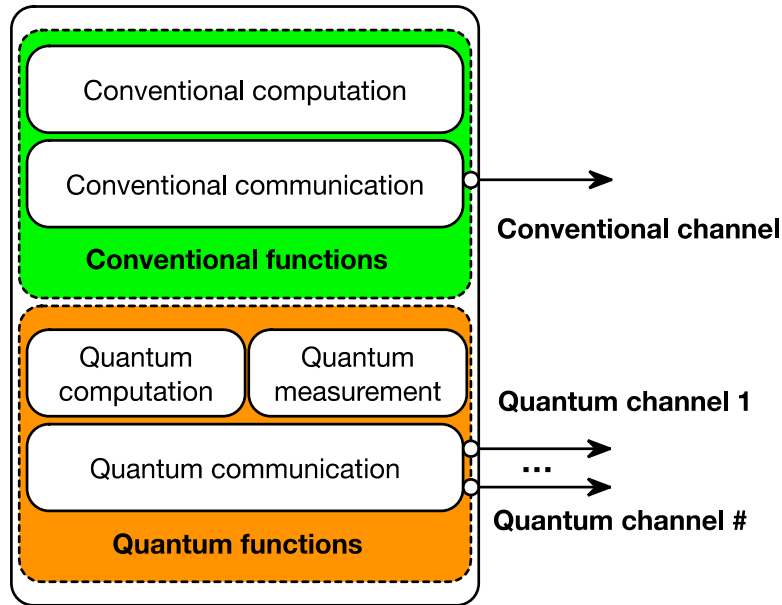
# IEQNET -- Topology



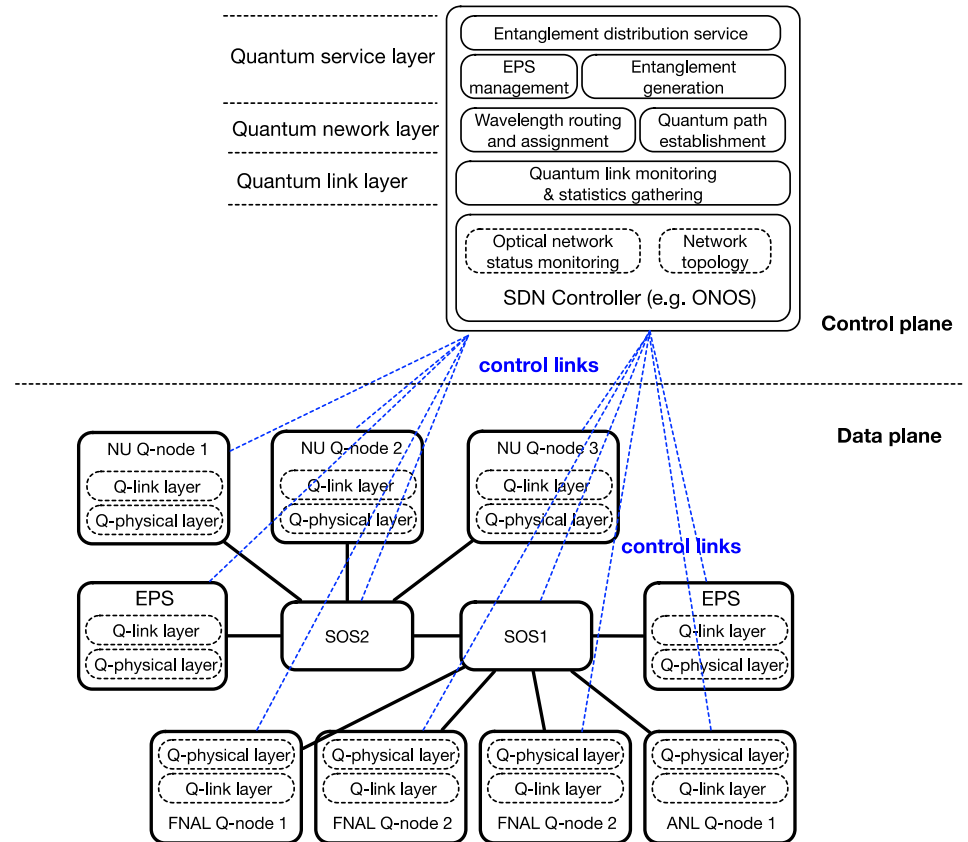
## Key features:

- Support multi-node, flexible, and resilient network configuration
- Support multi-user
- Coexist with traditional networks in the same optical transmission systems
- Adopt a layered architecture and a centralized control

# IEQNET – Model and Architecture



Q-node model



Network Architecture

# IEQNET – Challenges

- **Photon loss**

- Increase exponentially with distance travelled
- No-cloning theorem
  - Quantum state of photons cannot be amplified without any disturbance

- **Phase decoherence**

- Degrade or terminate entanglement

# Questions?

[wenji@fnal.gov](mailto:wenji@fnal.gov)