FERMILAB-SLIDES-19-053-SCD

Fermilab U.S. DEPARTMENT OF ENERGY | Office of Science

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
  - 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)

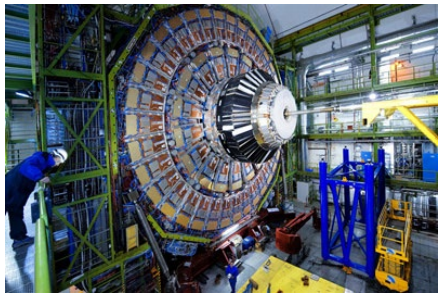
Fermilab is America's particle physics and accelerator laboratory.

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

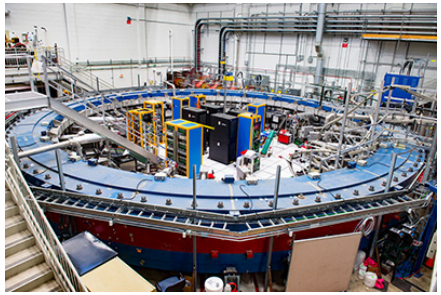
- Over 50 years at the forefront of discovery in high energy physics (top & bottom quarks, tau neutrino,...)
- We build and operate High Power (MW), High reliability Accelerators. Funded by DOE Office of Science.
- Largest U.S. Accelerator complex, 6800 acre site, ~$400M/yr budget, nearly 1800 staff,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

Why Do We Need Advanced network research at Fermilab?

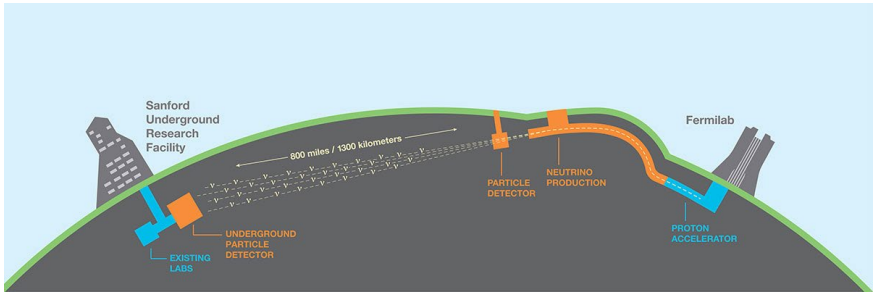
- High-through computing for international particle physics collaborations requires the ability to transport large amount of data quickly around the world



CMS Experiment



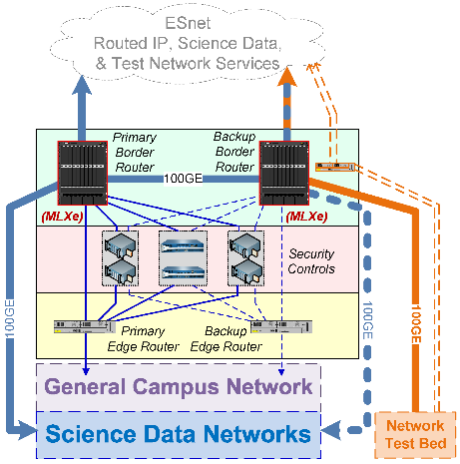
Muons Experiment



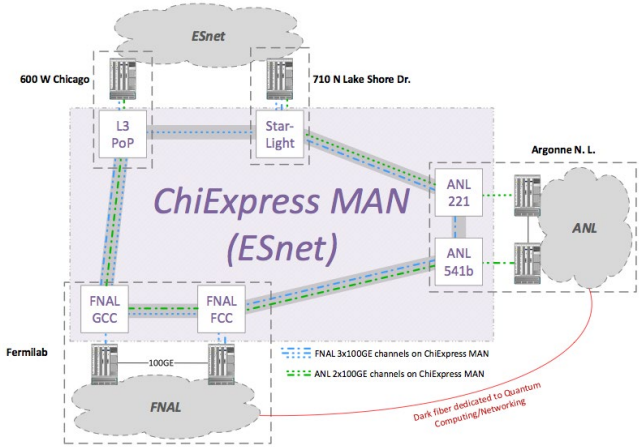
LBNF/DUNE Experiment

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

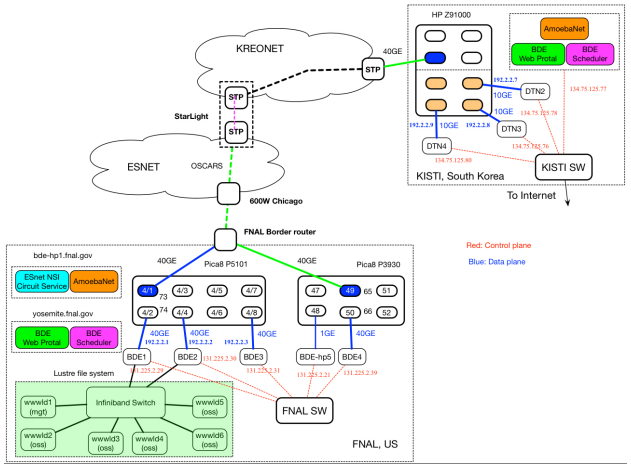
Advanced networking facilities at Fermilab



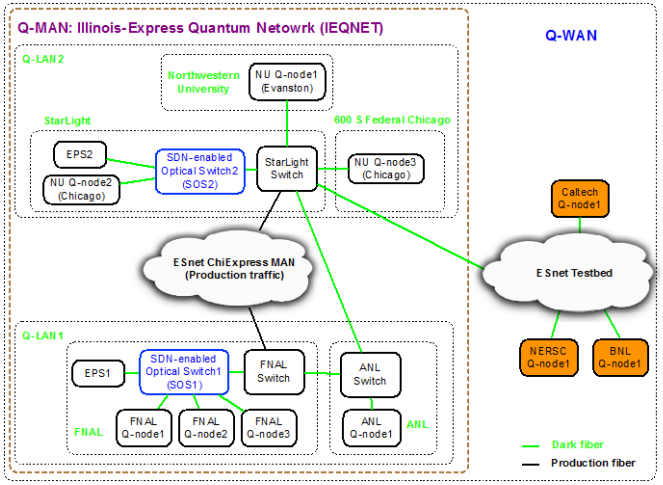
Fermilab Wide Area Networking Facilities



ChiExpress MAN



The Cross-Pacific SDN Testbed



Illinois-Express Quantum Network (IEQNET)

Fermilab

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)

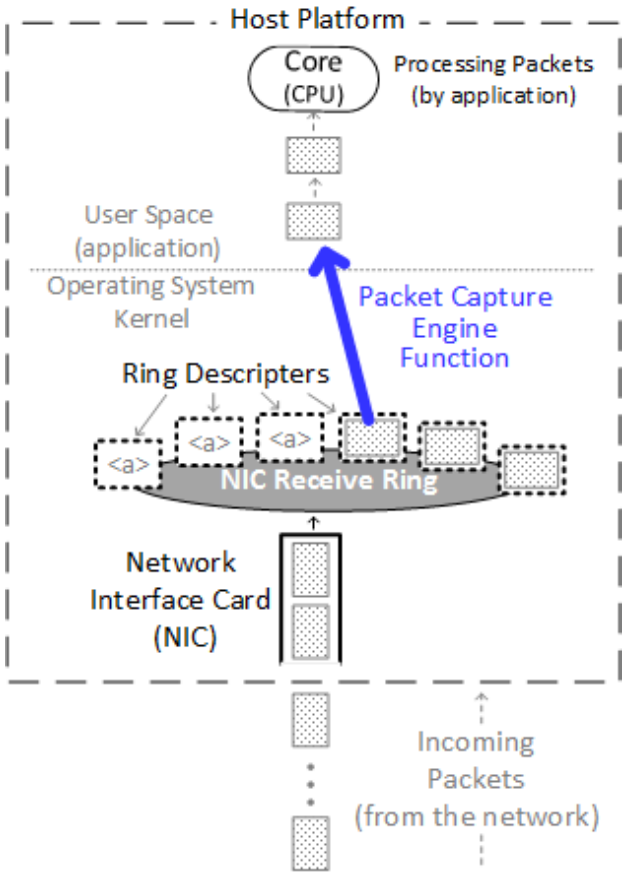
WireCAP: A Novel Packet Capture Engine for Commodity NICs in High-speed Networks

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!!

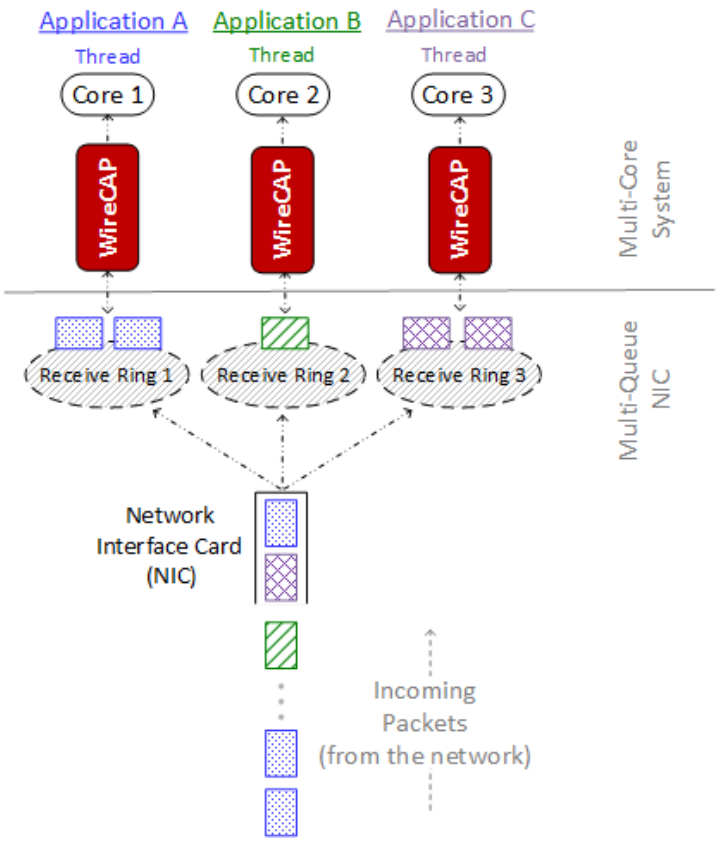


Fermilab

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

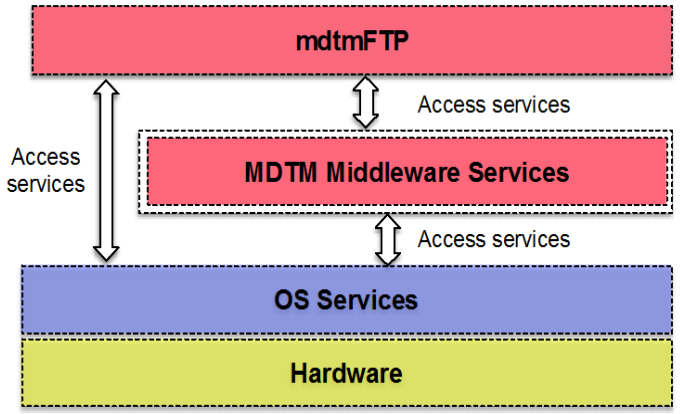
- Project website: http://wirecap.fnal.gov
- Paper
  - Wenji Wu, Phil DeMar, "WireCAP: a Novel Packet Capture Engine for Commodity NICs in High-speed Networks," *IMC'14*, November 05 – 07 2014, Vancouver, BC, Canada.
- Patent
  - U.S. Patent 20160127276A1, filed November 4, 2015, and issued September 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

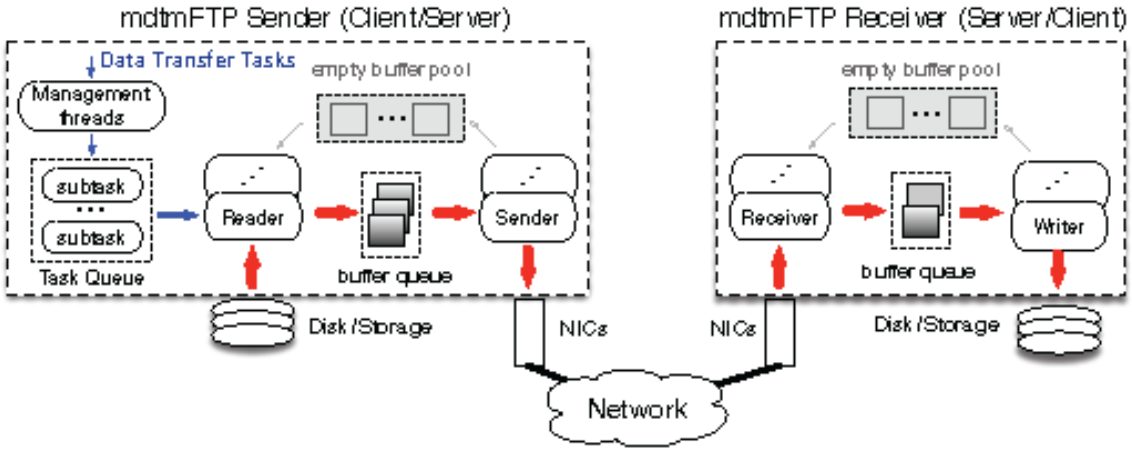
- Pipelined I/O-centric design to streamline data transfer
- Multicore-aware data transfer middleware (MDTM) optimizes use of underlying multicore system
- Extremely efficient in transferring of Lots of Small Files (LOSF)
- Various optimization mechanisms
  - Zero copy
  - Asynchronous I/O
  - Batch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



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

mdtmFTP – A pipelined I/O centric design

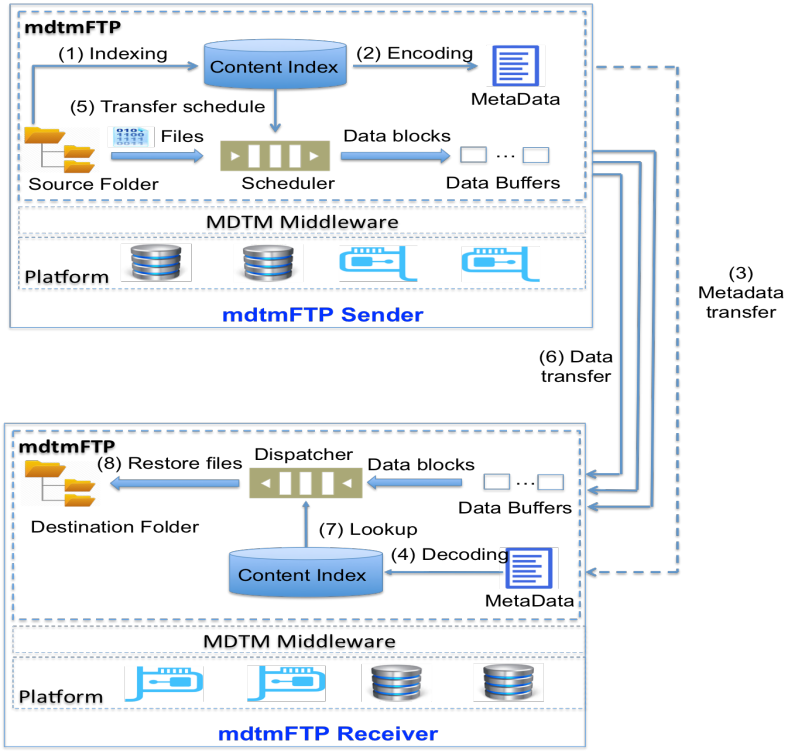
- Dedicated I/O threads to perform network/disk I/Os in parallel
- MDTM middleware to schedule cores for I/O threads
- Advanced data buffer mechanism to improve I/O performance
- Data transfers are executed in a pipelined manner
- A data transfer task is split into multiple subtasks
- Subtasks are executed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



Fermilab

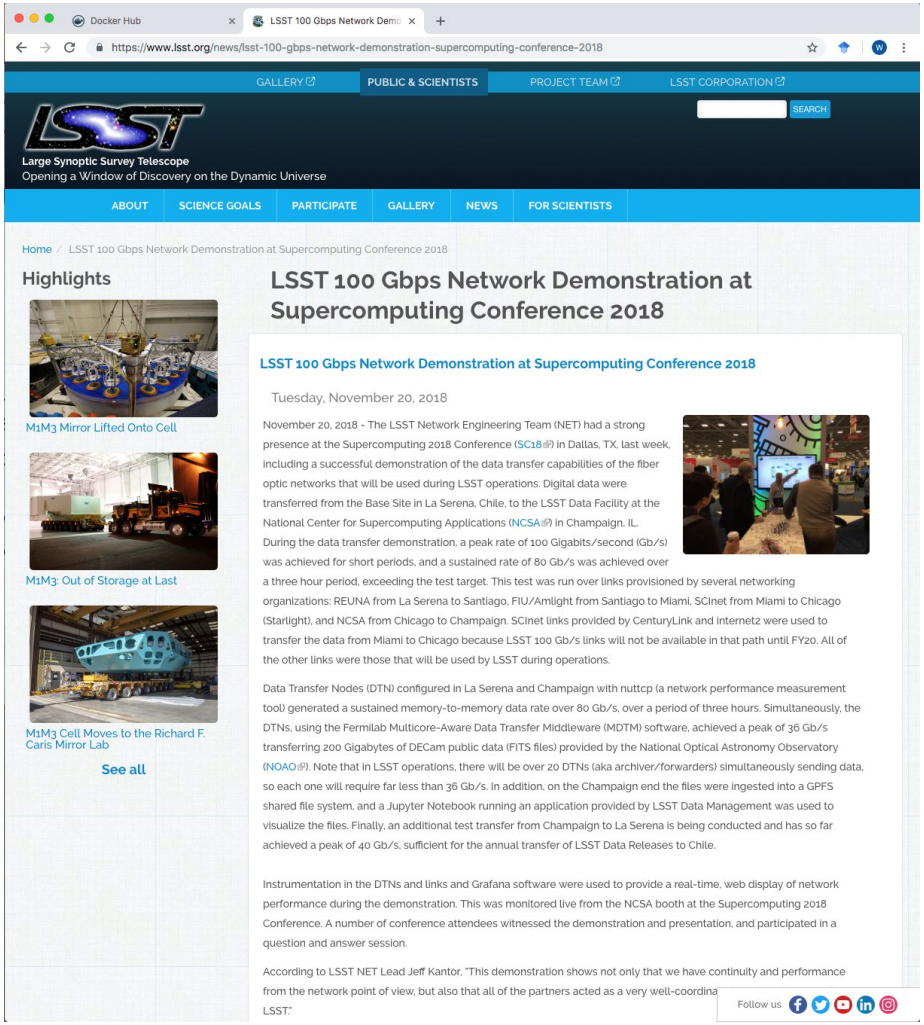
mdtmFTP – A large virtual file transfer mechanism

- 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



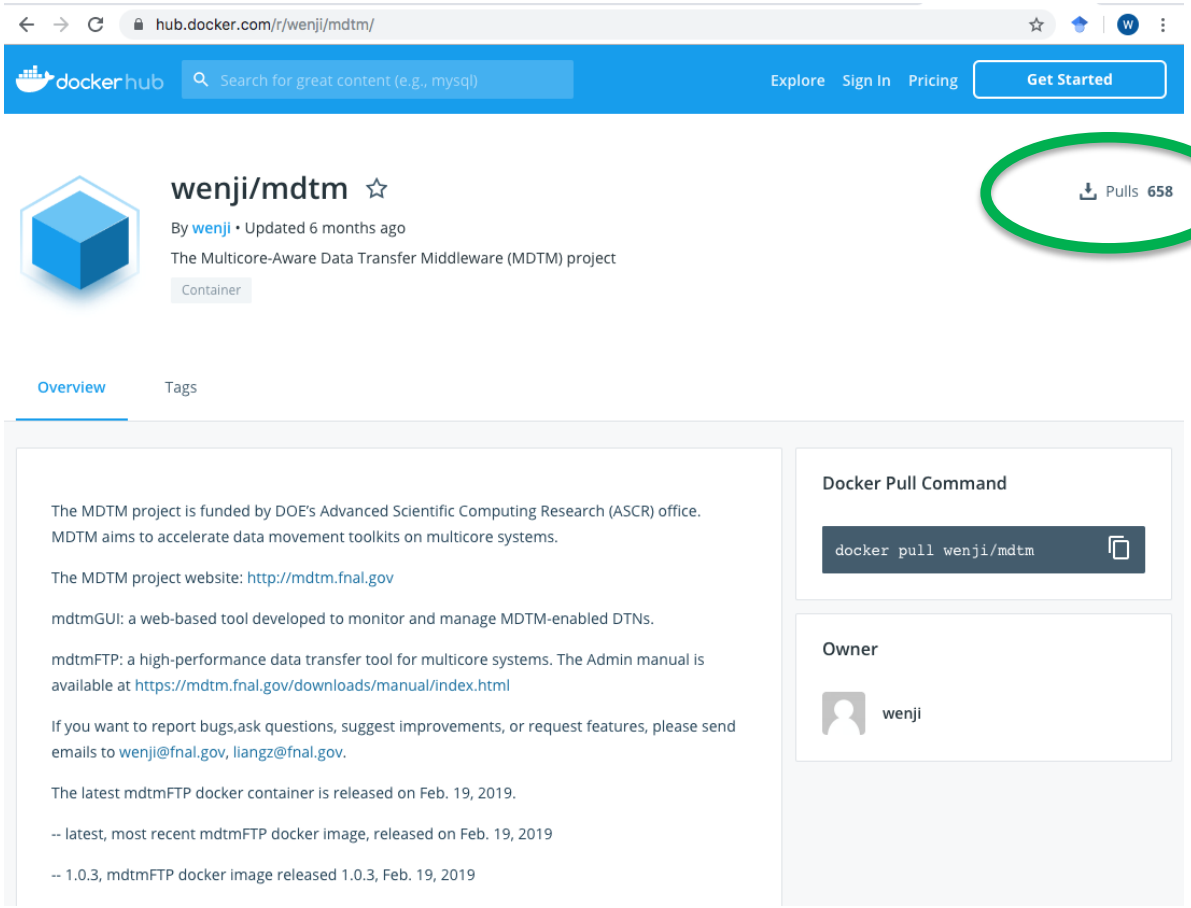
Fermilab

LSST SC'18 News Release



"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



658 downloads

MDTM
MULTICORE DATA TRANSFER MIDDLEWARE

docker

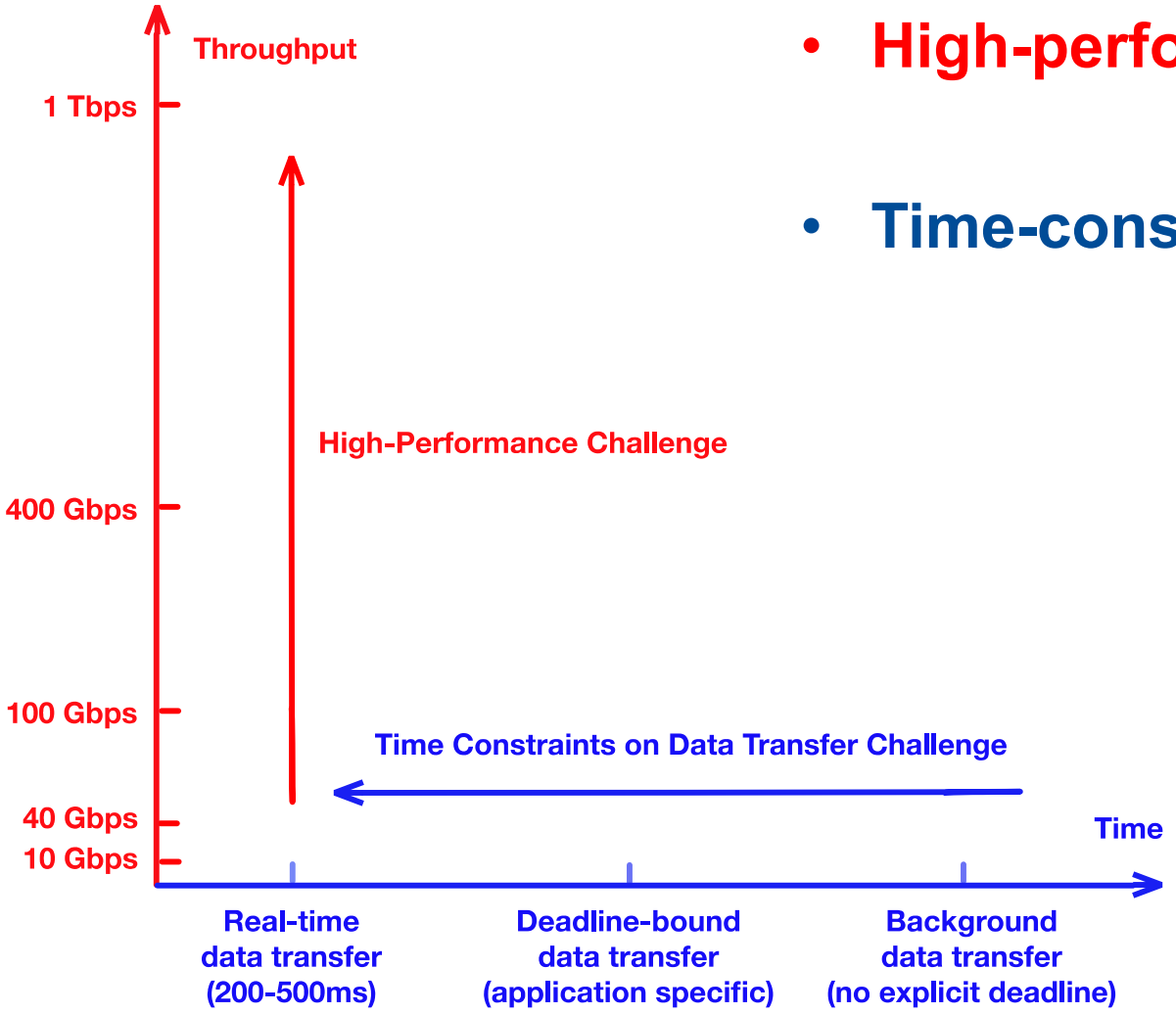
https://hub.docker.com/r/wenji/mdtm/

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**

Fermilab

Data Transfer – State of the Art

- Advanced data transfer tools and services developed
  - GridFTP, BBCP
  - PhEDEx, LIGO Data Replicator, Globus Online

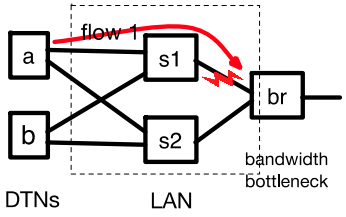
- Numerous enhancements
  - Parallelism at all levels
    - Multi-stream, Multicore, Multi-path parallelism
  - Science DMZ architecture
  - Terabit networks

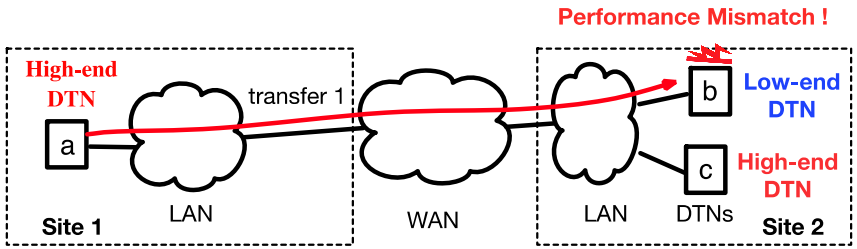
Problems with Existing Data Transfer Tools & Services

- 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

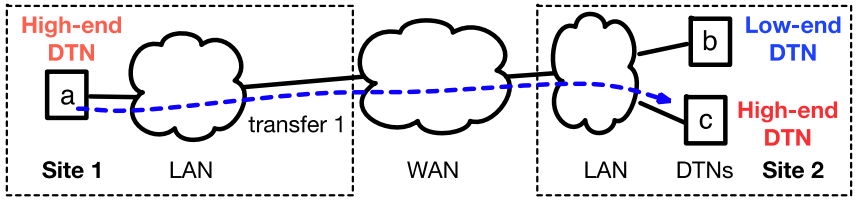
Problem 1 – Disjoint end-to-end data transfer loop

- Distributed resource management model
  - Resource contention
  - Performance mismatch





a. without coordination



b. with coordination

Fermilab

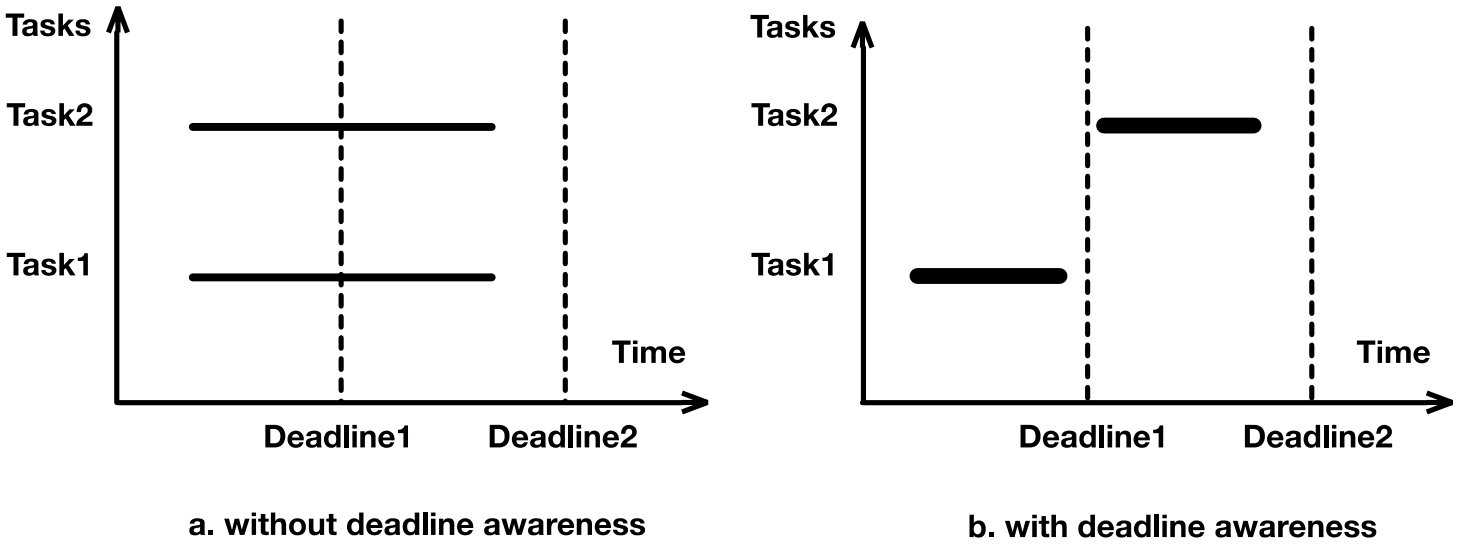
Problem 2 – Cross-interference between data transfers

- Degraded performance
- High variability in data transfer performance

Problem 3 – Oblivious to user requirements

- Data transfer jobs are schedule on a first-come, first serve basis
  - Without deadline awareness

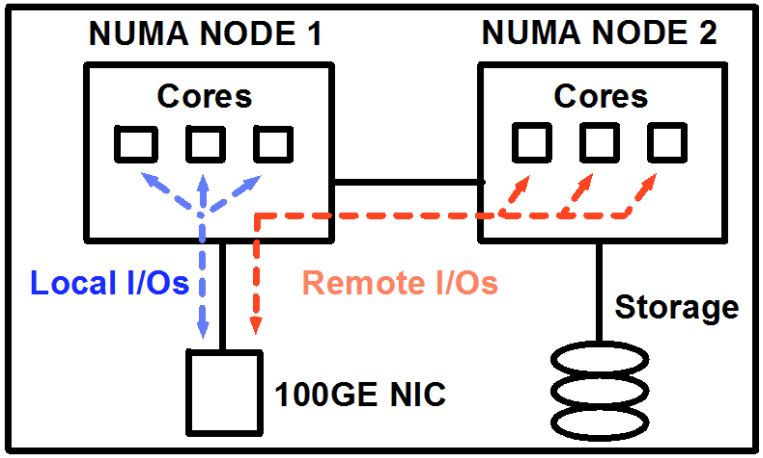
- Resources are shared fairly among data transfer jobs



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!

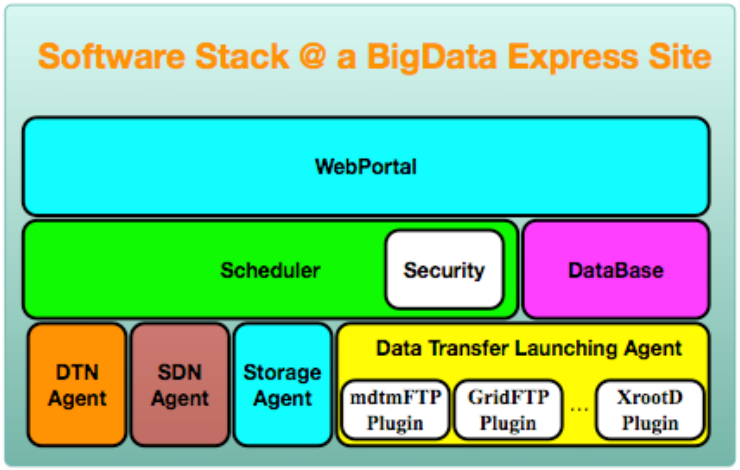
Our Solution – BigData Express

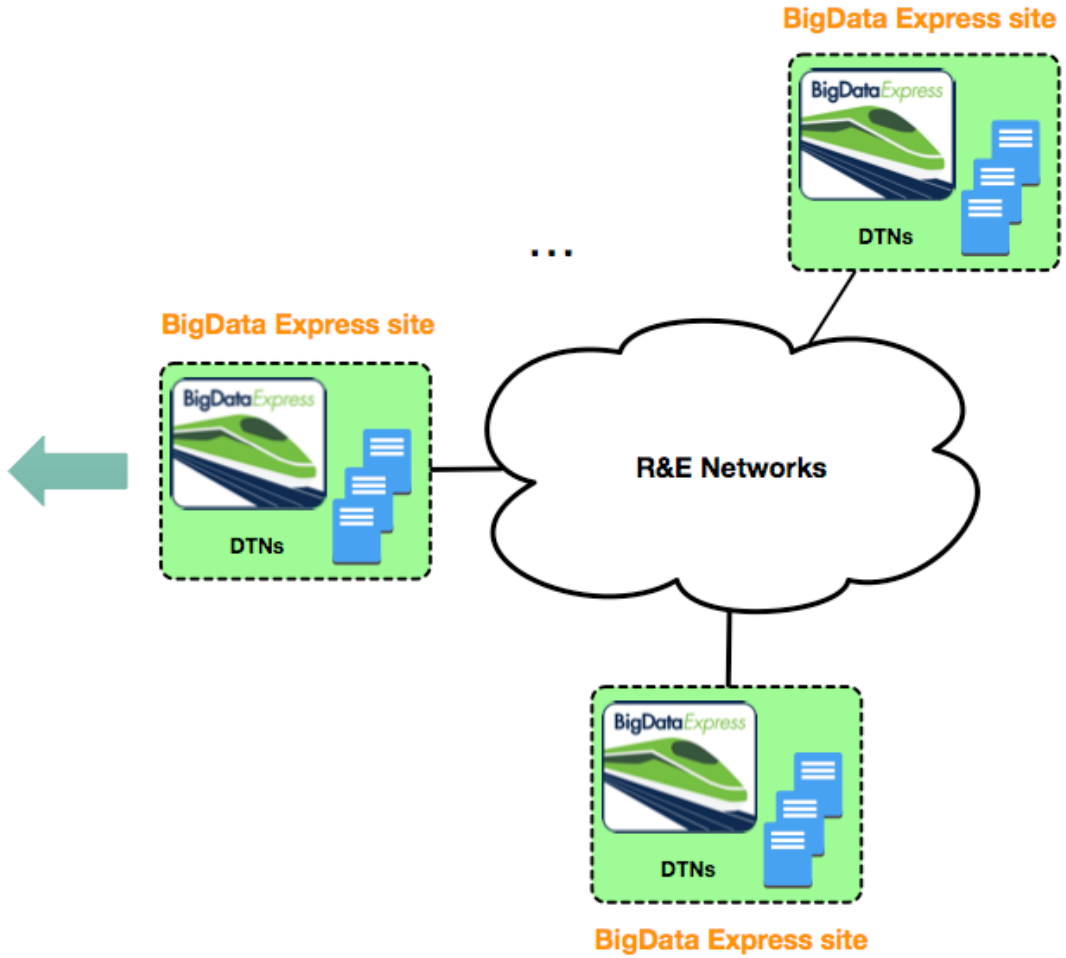
- BigData Express: a schedulable, predictable, and high-performance data transfer service
  - A peer-to-peer, scalable, and extensible data transfer model
  - A visually appealing, easy-to-use web portal
  - A high-performance data transfer engine
  - A time-constraint-based scheduler
  - On-demand provisioning 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

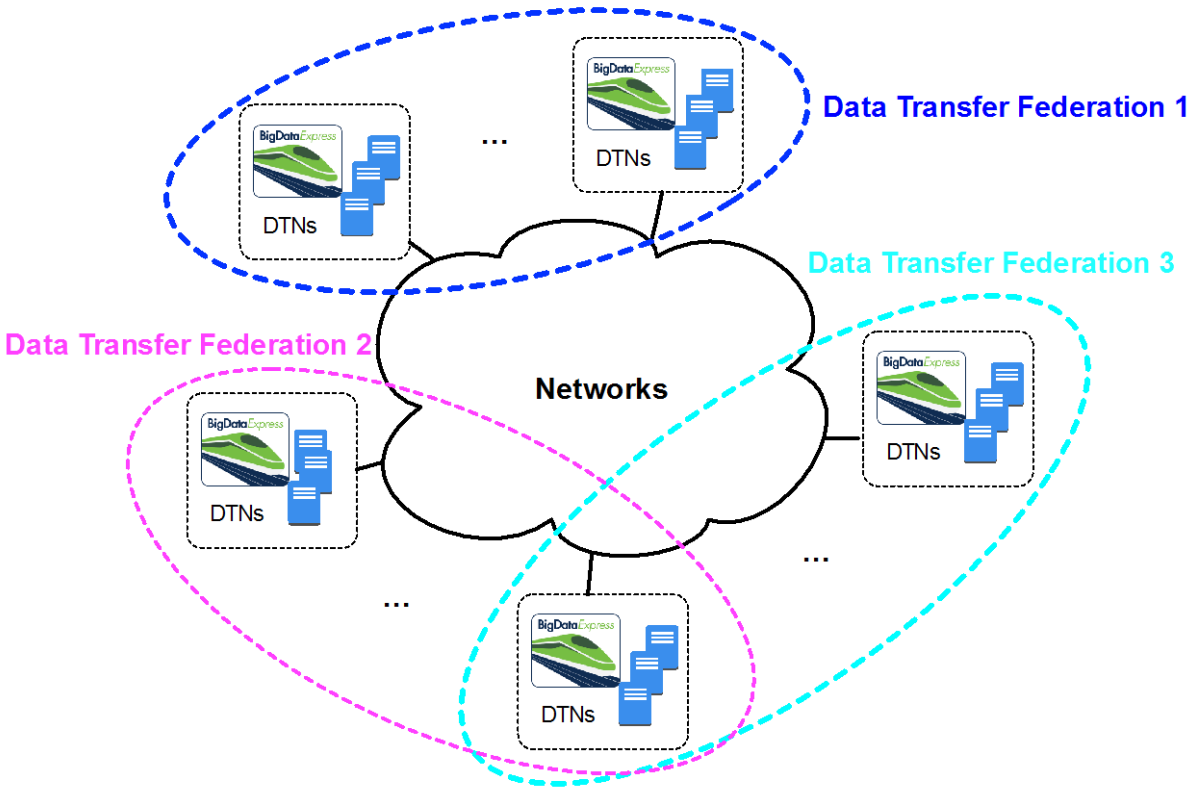
BigData Express -- Distributed





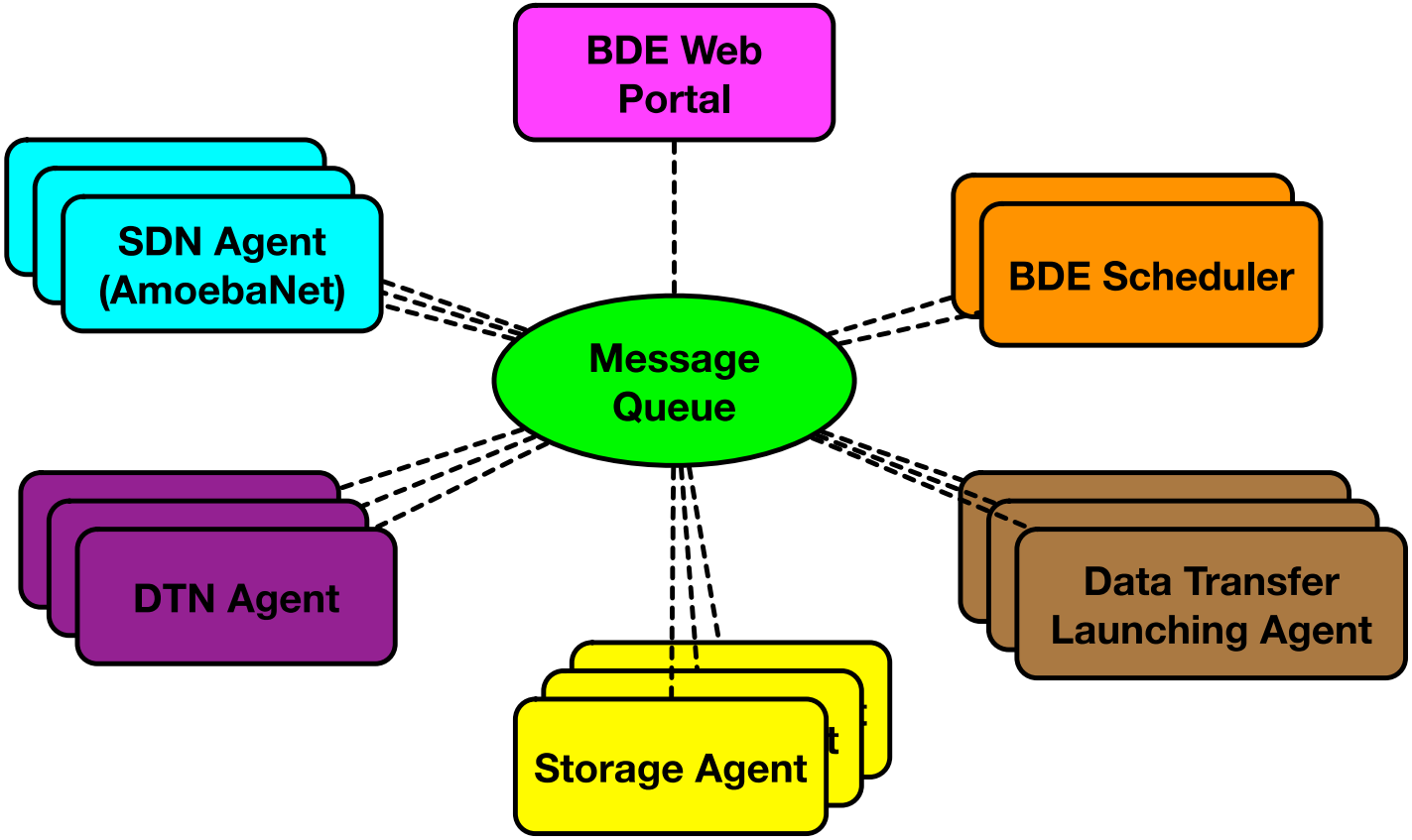
A Peer-to-Peer model

BigData Express -- Flexible



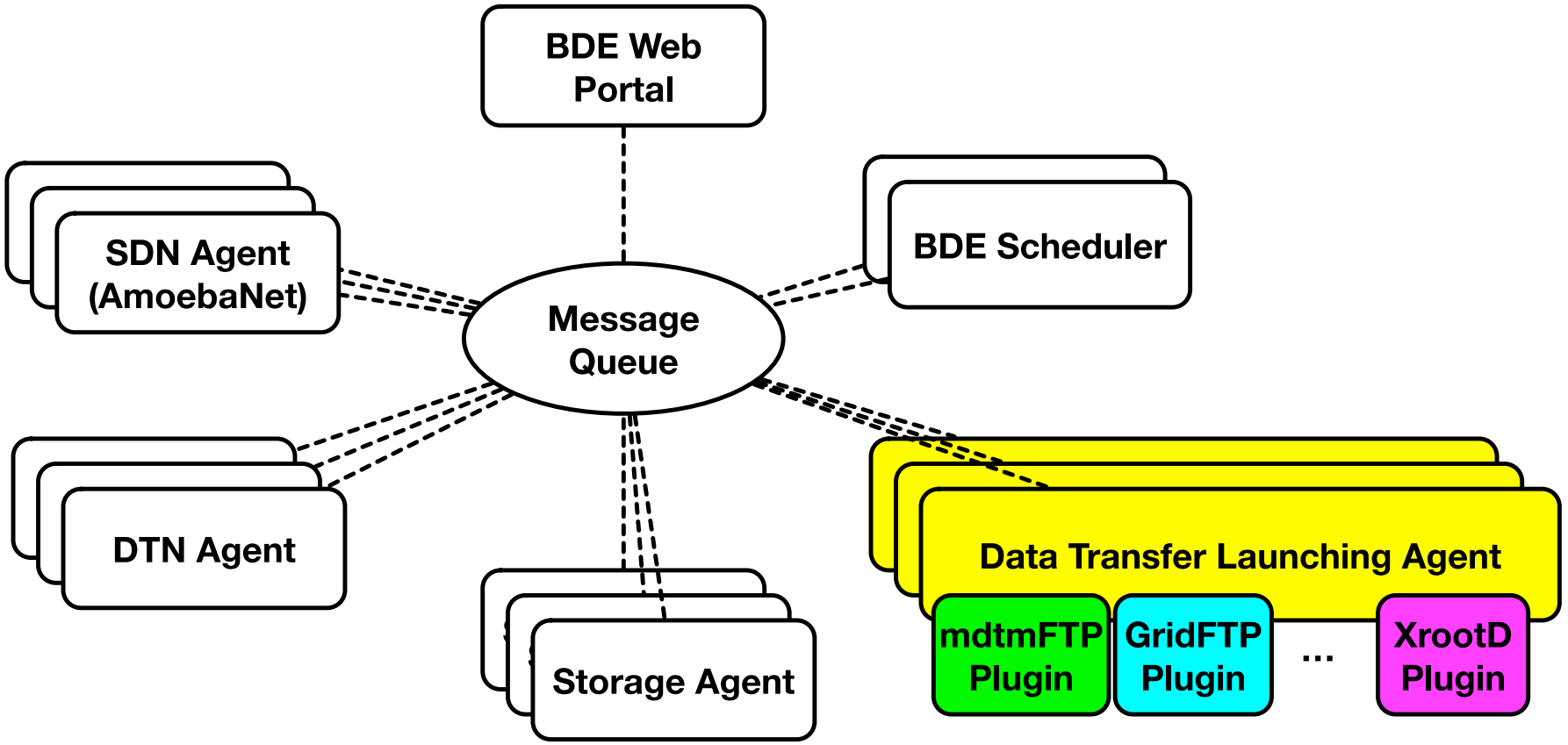
- Flexible to set up data transfer federations
- Providing inherent support for incremental deployment

BigData Express – Scalable



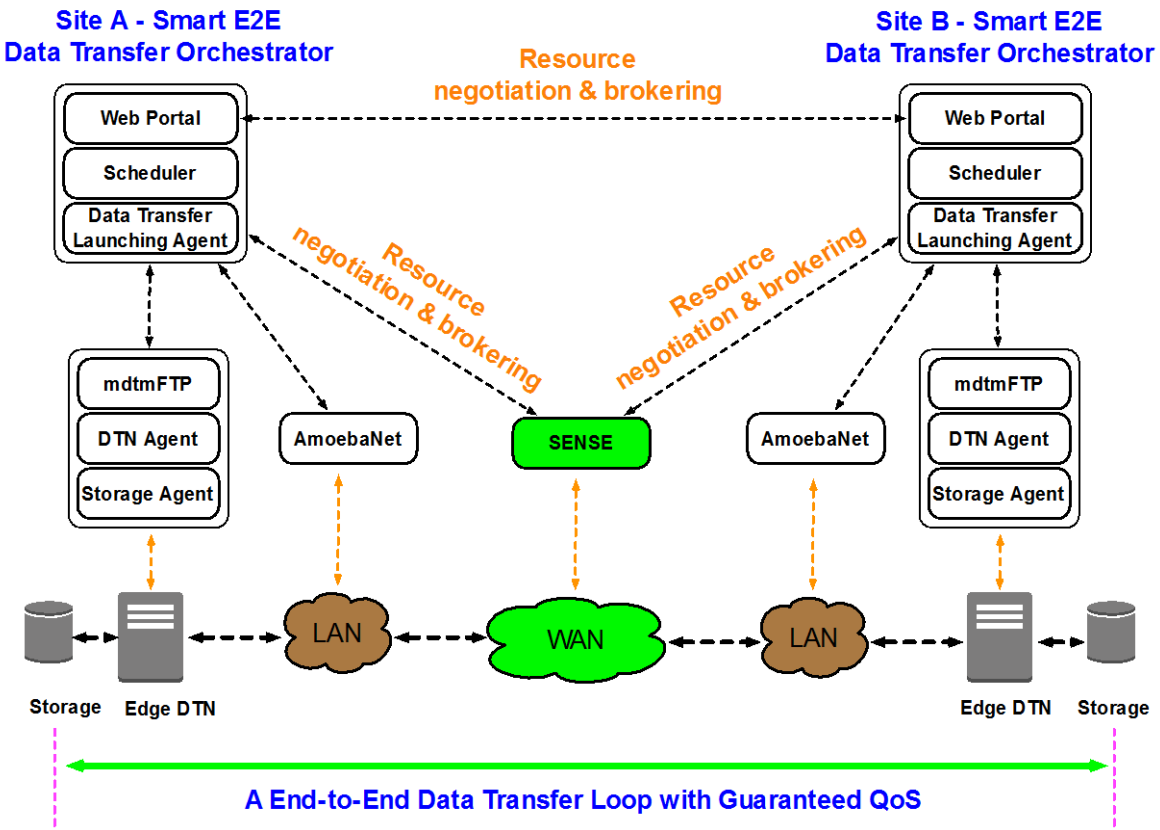
- BigData Express scheduler manages site resources through agents
- Use MQTT as message bus

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**

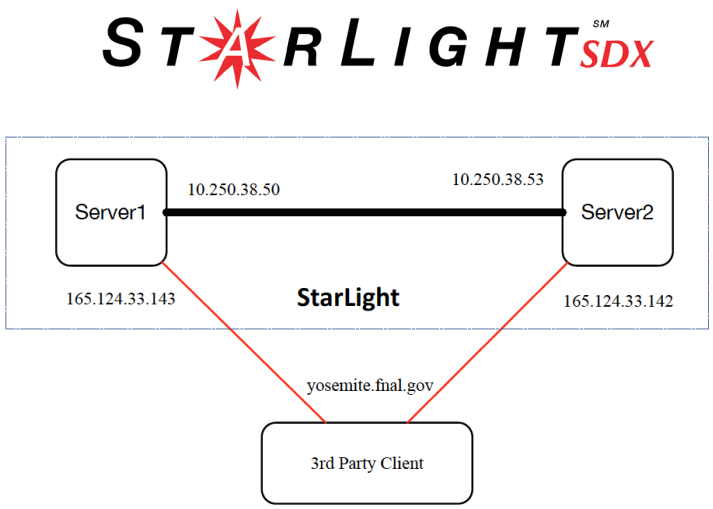
	mdtmFTP	FDT	GridFTP	BBCP
Large file data transfer (1 X 100G)	34.976	N/A	106.84	N/A
Folder data transfer (30 x 10G)	95.61	N/A	-	N/A
Folder data transfer (Linux 3.12.21)	9.68	N/A	-	N/A

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 3rd party data transfer. But BBCP and FDT couldn't run 3rd 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,

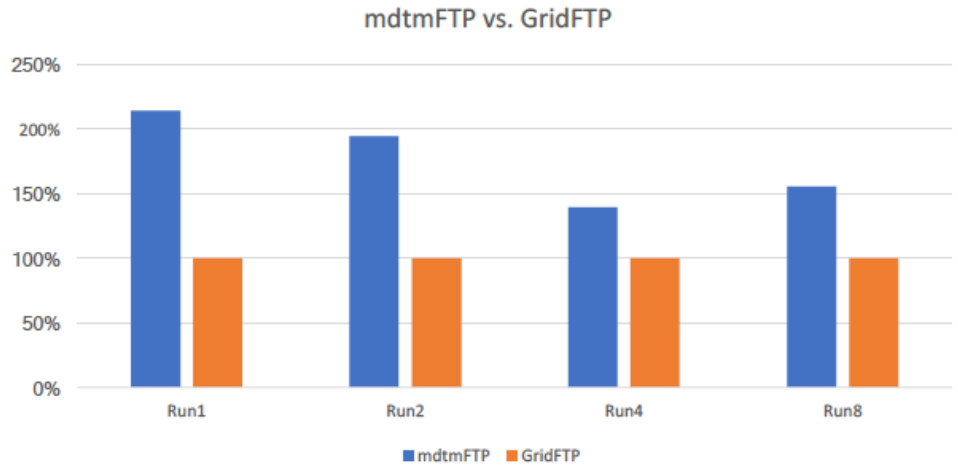
BigData Express – High Performance Data Transfer



StarLight 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
• Disjoint end-to-end data transfer loop	• Distributed resource negotiation & brokering • Co-scheduling of DTN, storage, & networking • On-demand provisioning of end-to-end network path with guaranteed QoS
• Cross-interference between data transfers	• Time-constraint-based scheduler • Admission control • Rate control
• Oblivious to user requirements	• Time-constraint-based scheduler • Three classes of data transfer
• Inefficiencies arises when existing data transfer tools run on DTNs	• mdtmFTP – A high-performance data transfer engine

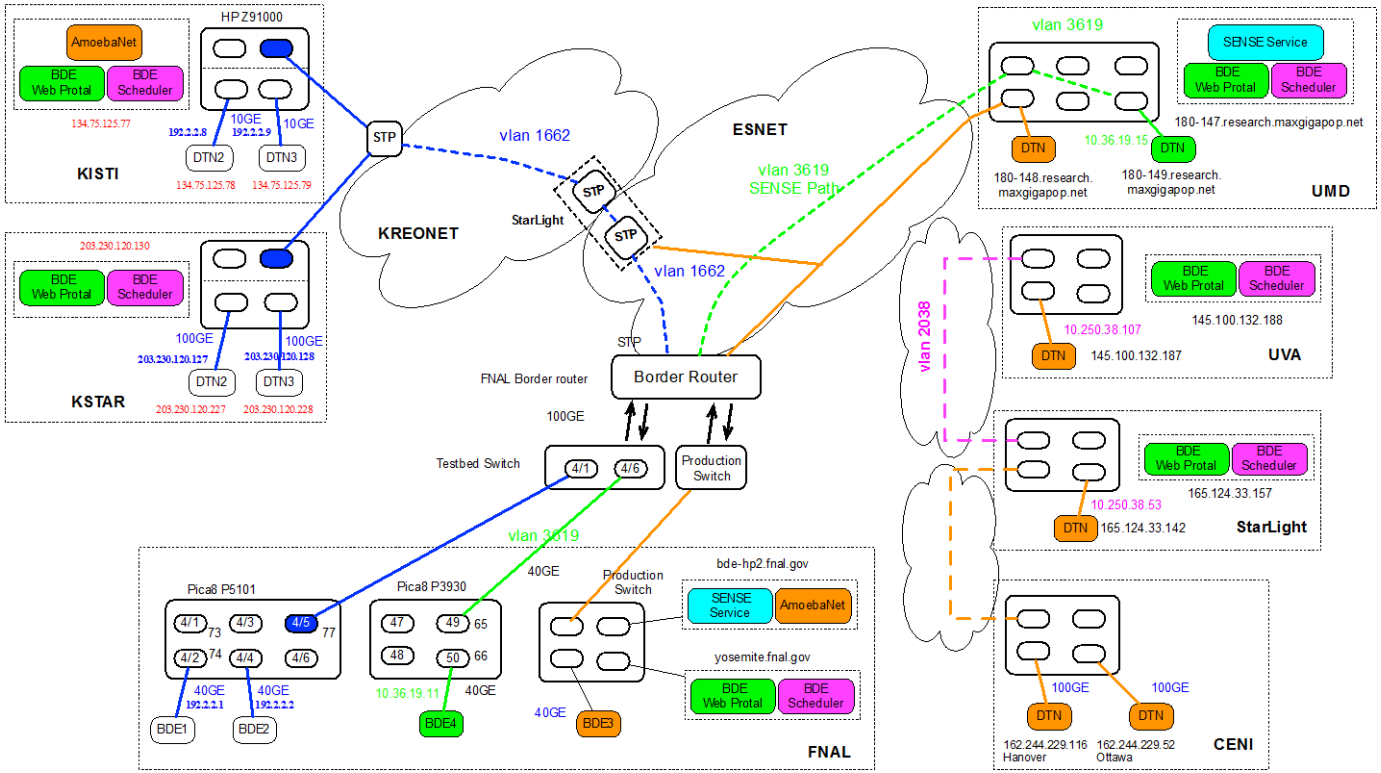
BigData Express vs. Globus Online

Features	BigData Express	Globus Online
Architecture	• Distributed service • Flexible to set up data transfer federations	• Centralized service

Supported Protocols	• Extensible plugin framework to support multiple protocols:
	  ○ mdtmFTP
	  ○ GridFTP, XrootD, SRM (coming soon)
		• GridFTP

SDN Support	• Yes, Network as a service • Fast-provisioning end-to-end network paths with guaranteed QoS	• Not in production
Supported Data Transfers	• Real-time data transfer • Deadline-bound data transfer • Best-effort data transfer	• Best-effort data transfer
Error Handling	• Checksum • Retransmit	• Checksum • Retransmit

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/

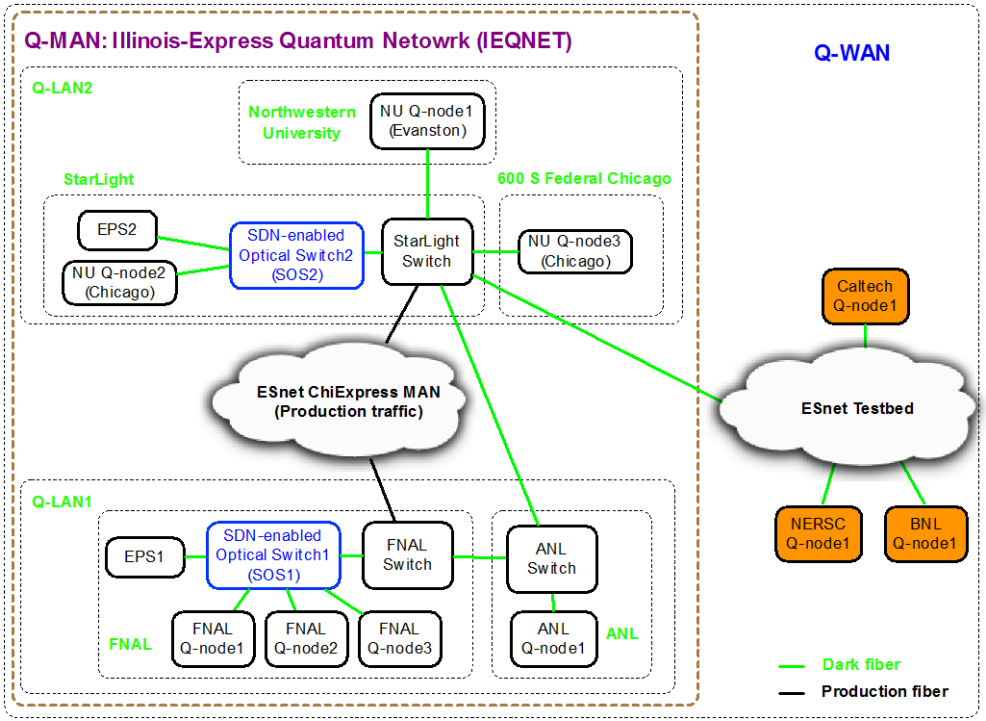
Illinois-Express Quantum Network (IEQNET)

- 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

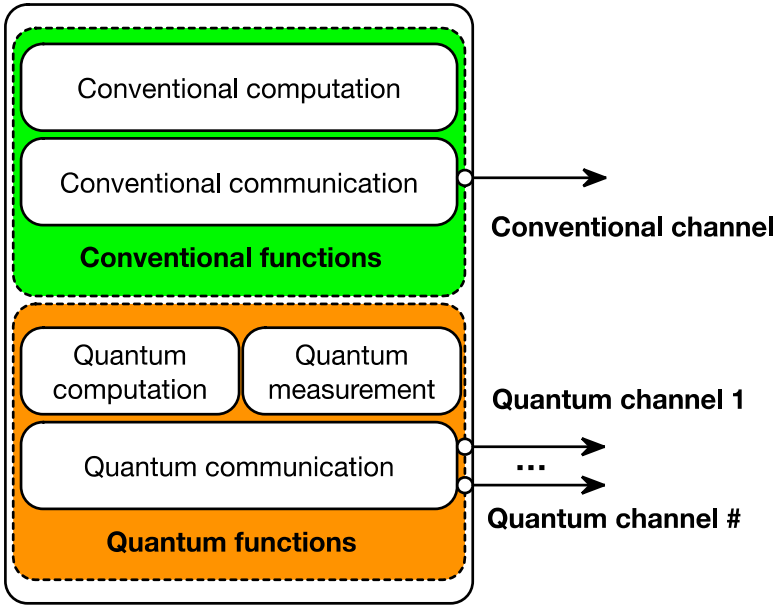


Key features:

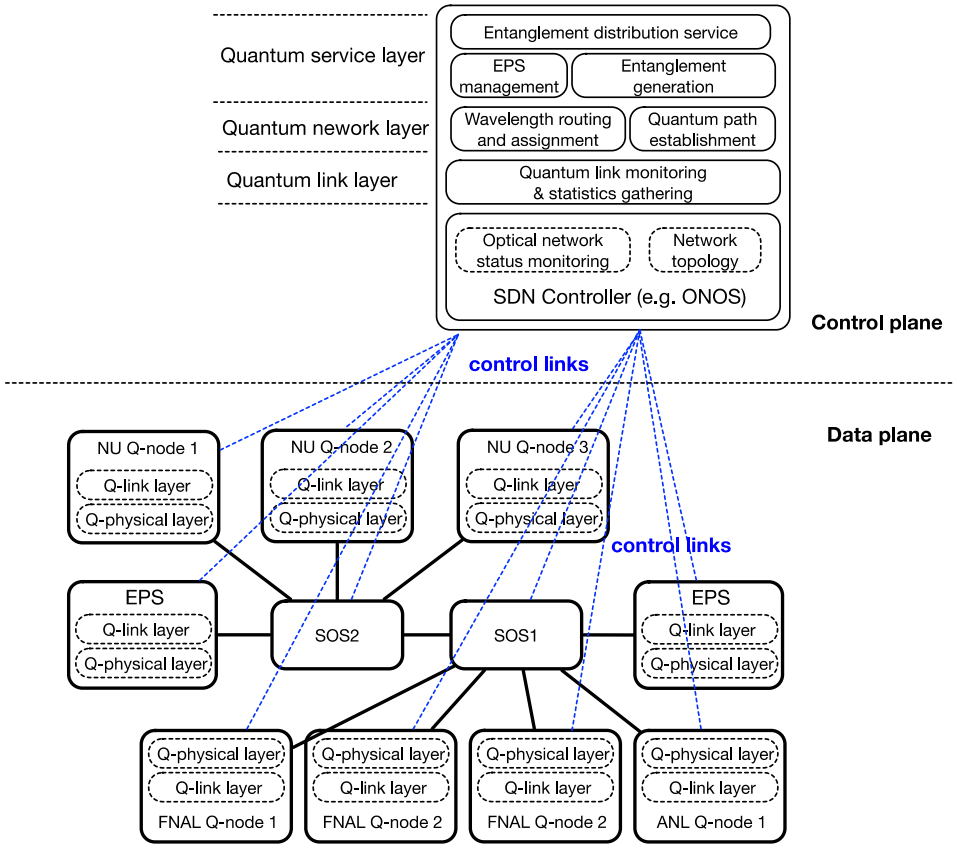
- Support multi-node, flexible, and resilient network configuration
- Support multi-user
- Coexist with traditional networks in the same optical transmission systems
- Adopt a layered architecture and a centralized control

Fermilab

IEQNET – Model and Architecture



Q-node model



Network Architecture

Fermilab

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