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

BigData Express: Toward Predictable, Schedulable, and High-performance Data Transfer

Wenji Wu, wenji@fnal.gov

Oct 31, 2018

BigData Express Research Team

- FNAL
  - Wenji Wu (PI)
  - Qiming Lu
  - Liang Zhang
  - Sajith Sasidharan
  - Phil DeMar
  - Amy Jin
- ORNL
  - Nageswara Rao
  - Gary Liu
- KISTI
  - Seo-Young Noh
  - Jin Kim

Note: KISTI and ESnet are unfunded project partners

Why BigData Express?

- **Targeted at optimizing data transfers in high-speed networks**
  - Large-scale data movement of Big Data Science
  - High-speed network environments (40/100GE+)
- **Builds on Multicore-Aware Data Transfer Middleware (MDTM)**
  - mdtmFTP: a high-performance data transfer tool
    - Pipelined I/O-centric design to streamline data transfer
    - MDTM optimizes use of underlying multicore system
    - Extremely efficient in transferring of Lots Of Small Files (LOSF)
  - http://mdtm.fnal.gov
- **Orchestrates system (DTN), storage, & network (SDN) resources**
  - To provide full end-to-end performance optimization

BigData Express

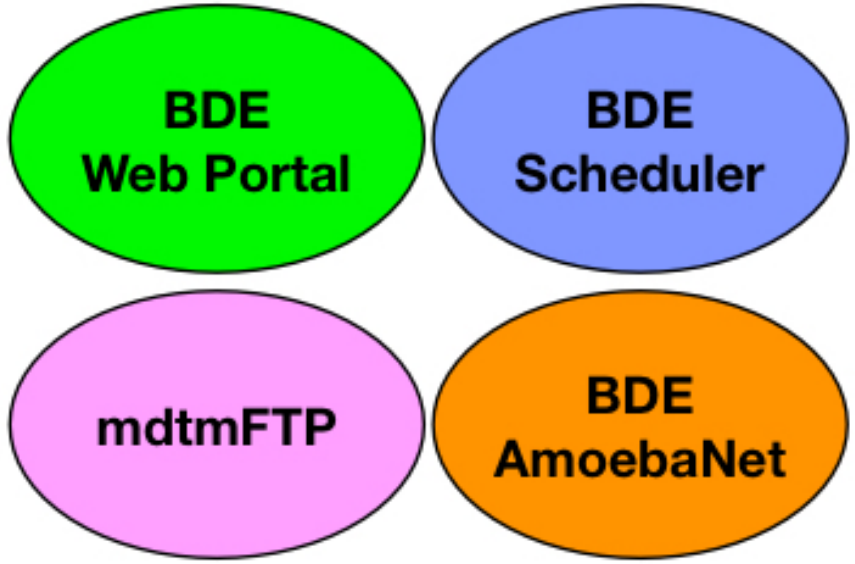
- BigData Express: a schedulable, predictable, and high-performance data transfer service
  - A peer-to-peer, scalable, and extensible data transfer model
  - A visually appealing, easy-to-use web portal
  - A high-performance data transfer engine
  - On-demand provisioning of end-to-end network paths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

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

Fermilab

BigData Express Major Components

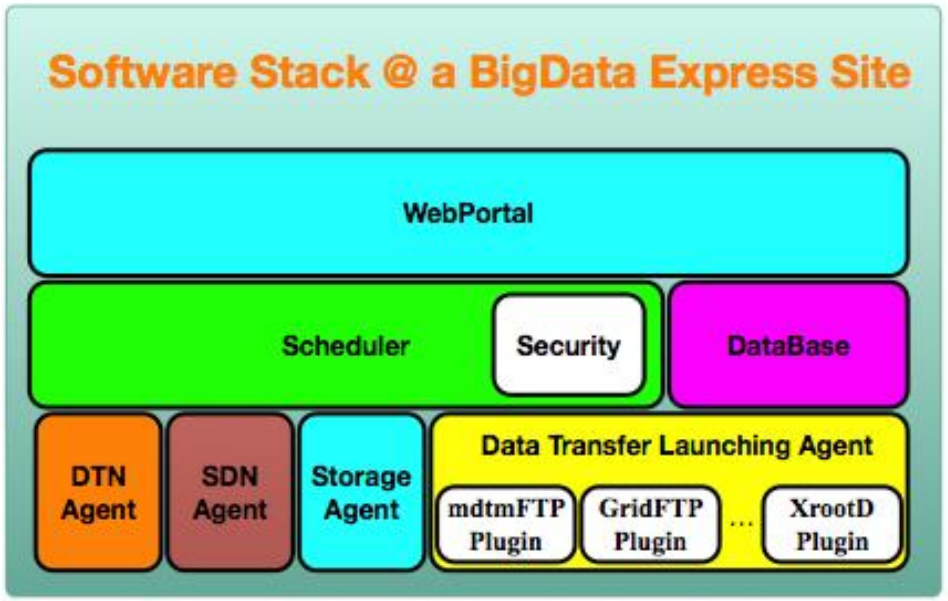
- **BDE Web Portal**
  - Allow users to access BigData Express data transfer services
- **BDE Scheduler**
  - DTN as a service
  - Co-scheduling of DTN, storage, and network
- **BDE AmoebaNet**
  - Network as a service
- **mdtmFTP**
  - a high-performance data transfer engine
  - http://mdtm.fnal.gov

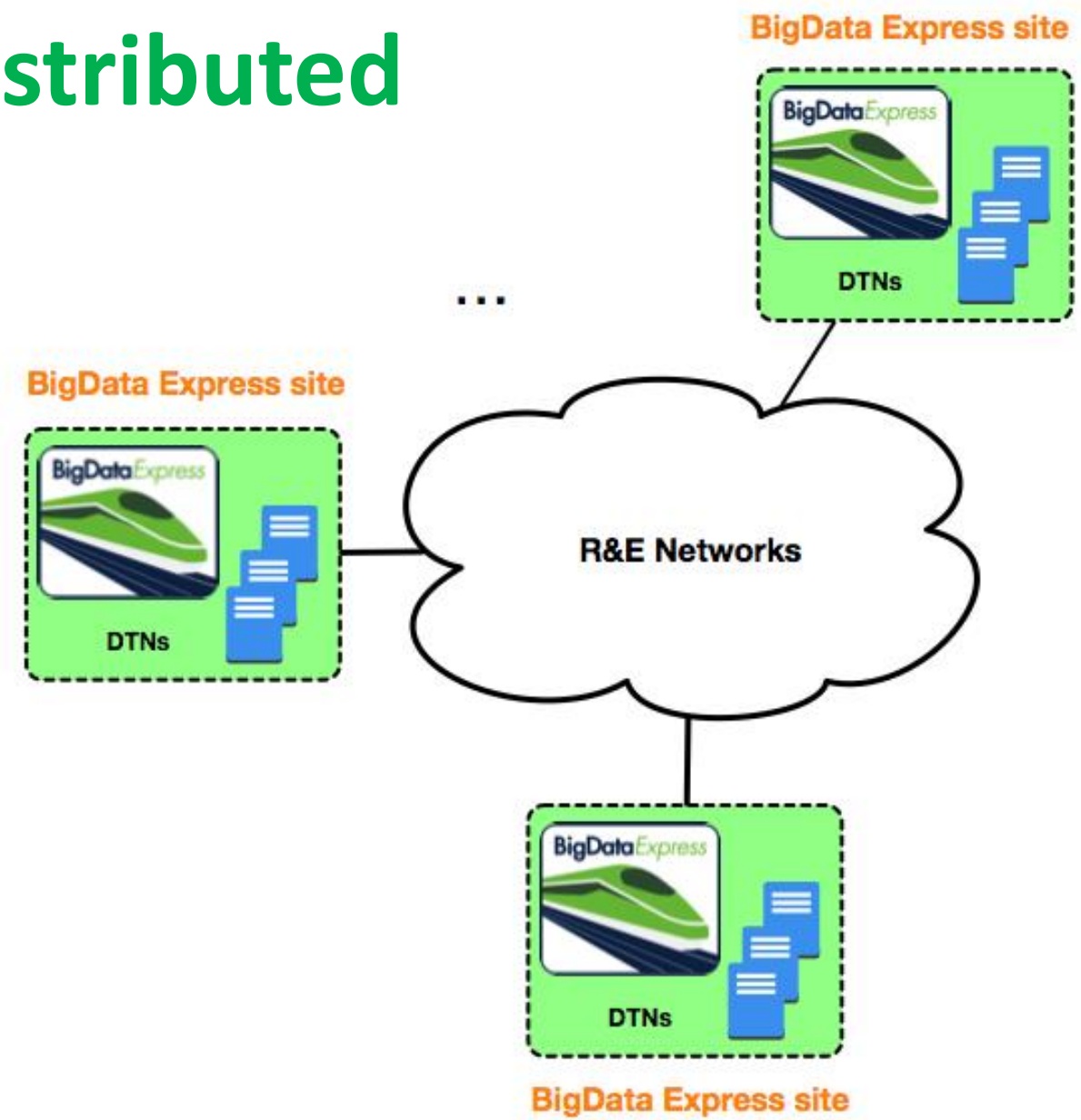


BigData Express Major Components (cont.)

- **DTN Agents**
  - Manage and configure DTNs
  - Collect and report the DTN configuration and status
- **Storage Agents**
  - Manage and configure storage systems
- **Data Transfer Launching Agent**
  - Launch data transfer jobs
  - Support different data transfer protocols

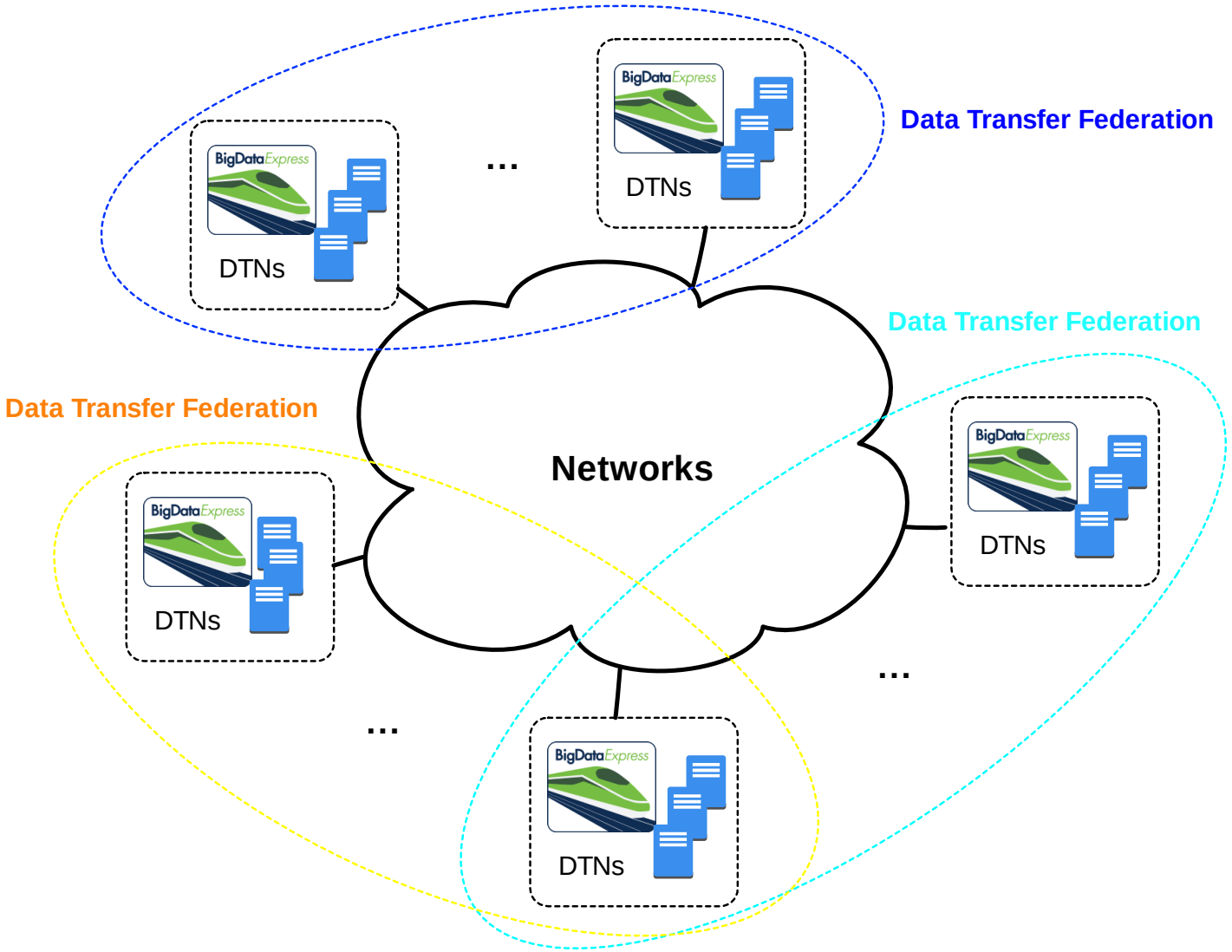
BigData Express -- Distributed





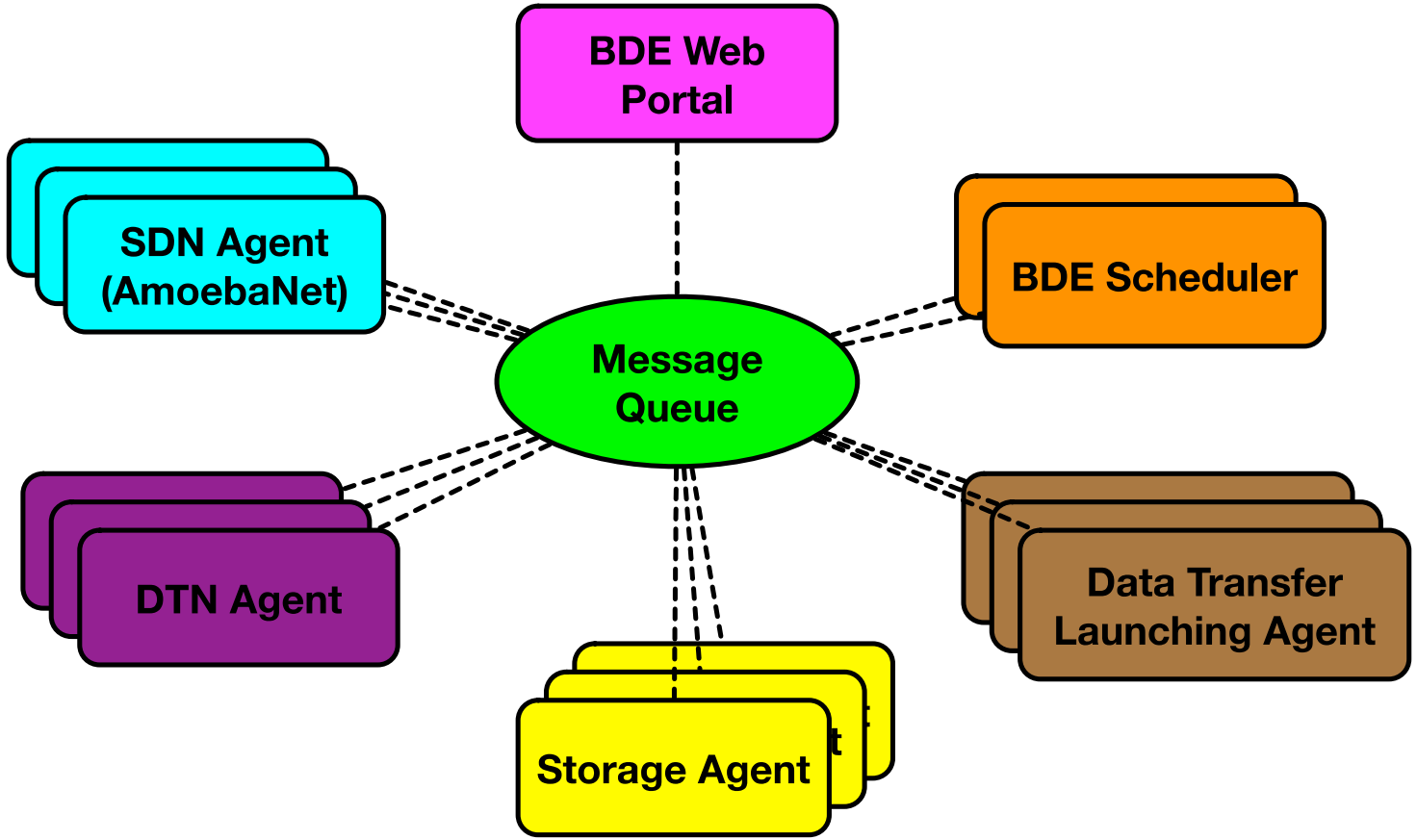
A Peer-to-Peer model

BigData Express -- Flexible



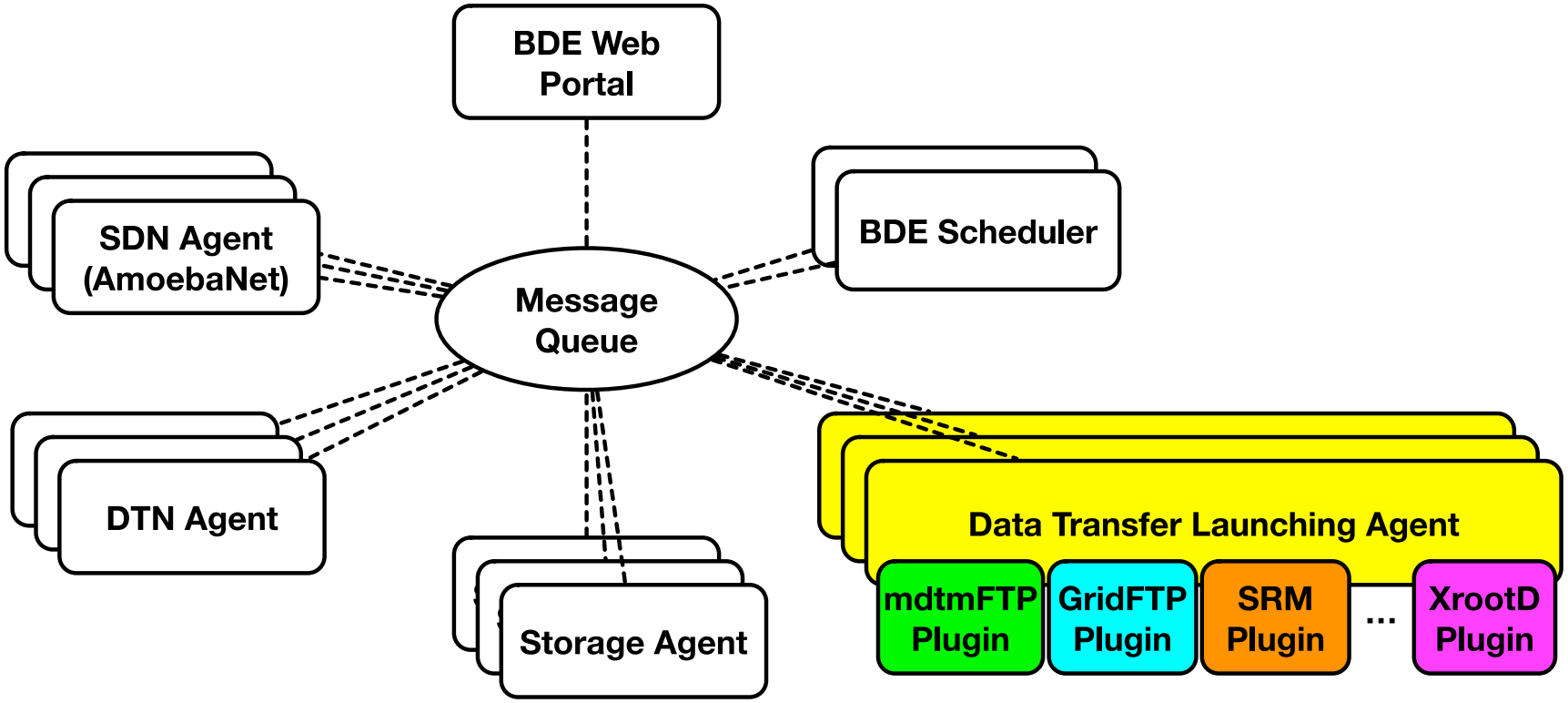
- **Flexible to set up data transfer federations**
- **Providing inherent support for incremental deployment**

BigData Express -- Scalable



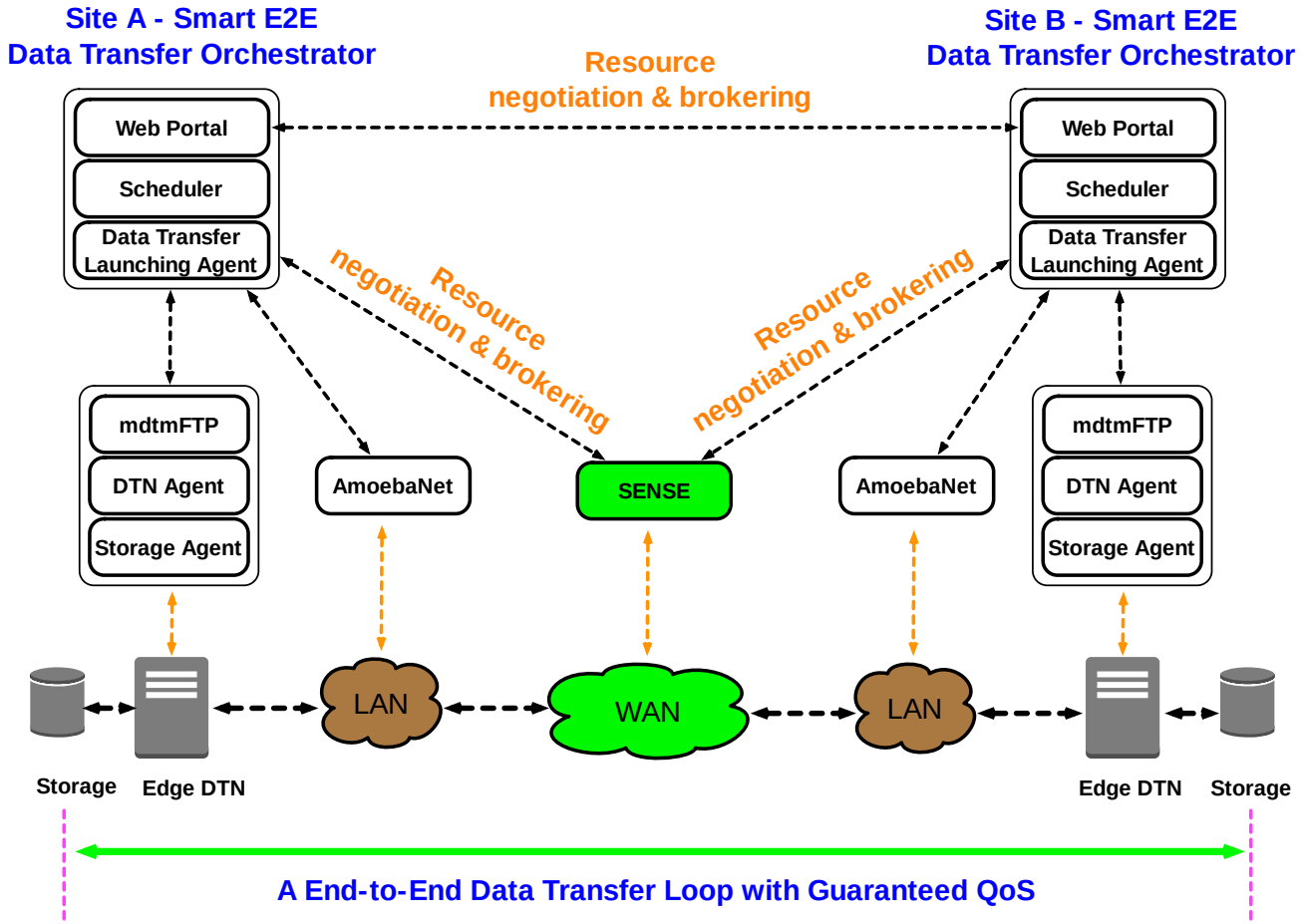
- BigData Express scheduler manages site resources through agents
- Use MQTT as message bus

BigData Express -- Extensible



- Extensible Plugin framework to support various data transfer protocols
  - mdtmFTP, GridFTP, SRM, XrootD, …

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



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

BigData Express – High Performance Data Transfer (I)

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

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

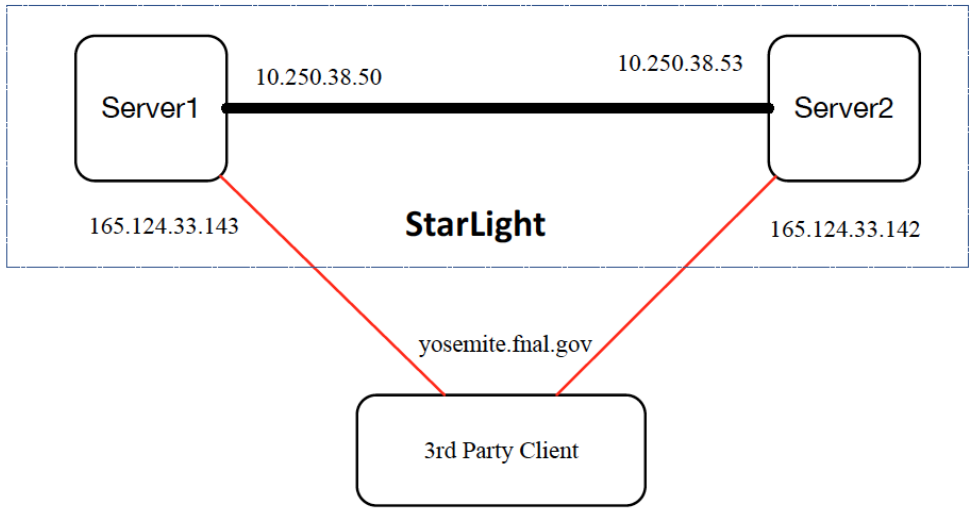
	mdtmFTP	FDT	GridFTP	BBCP
Large file data transfer (1 X 100G)	34.976	N/A	106.84	N/A
Folder data transfer (30 x 10G)	95.61	N/A	-	N/A
Folder data transfer (Linux 3.12.21)	9.68	N/A	-	N/A

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,

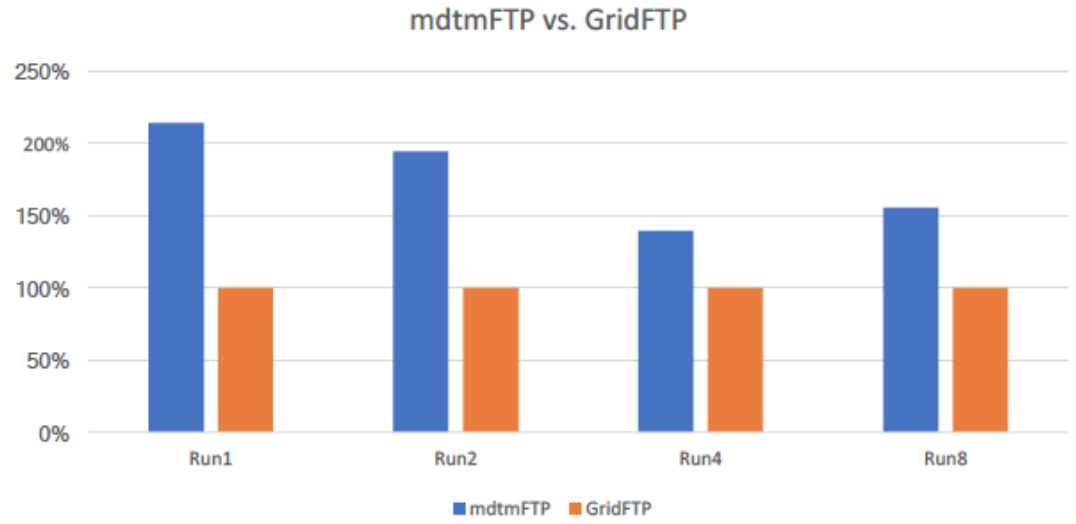
BigData Express – High Performance Data Transfer (II)



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 -- Three Types of Data Transfer

- Real-time data transfer
- Deadline-bound data transfer
- Best-effort data transfer

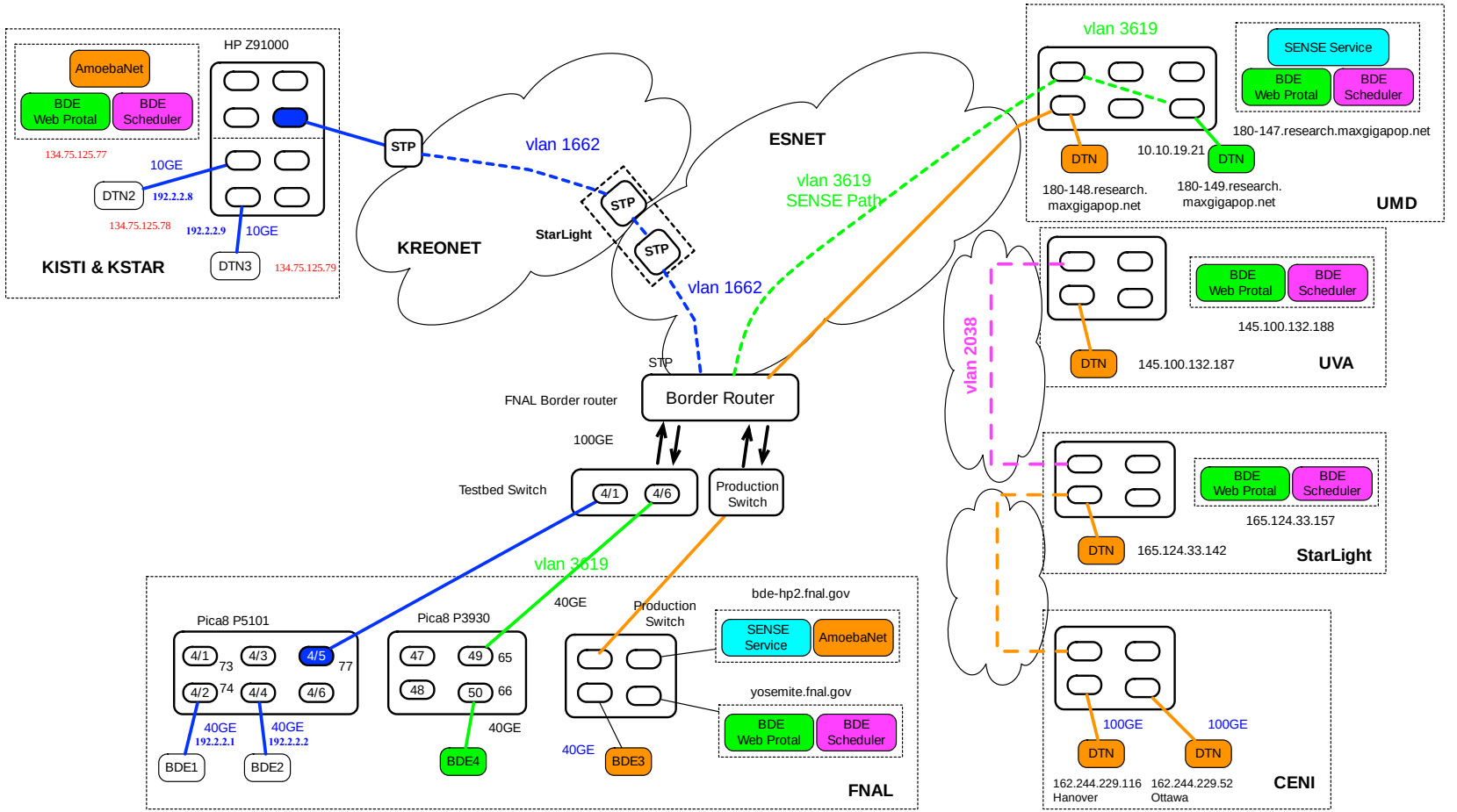
BigData Express vs. Globus Online

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

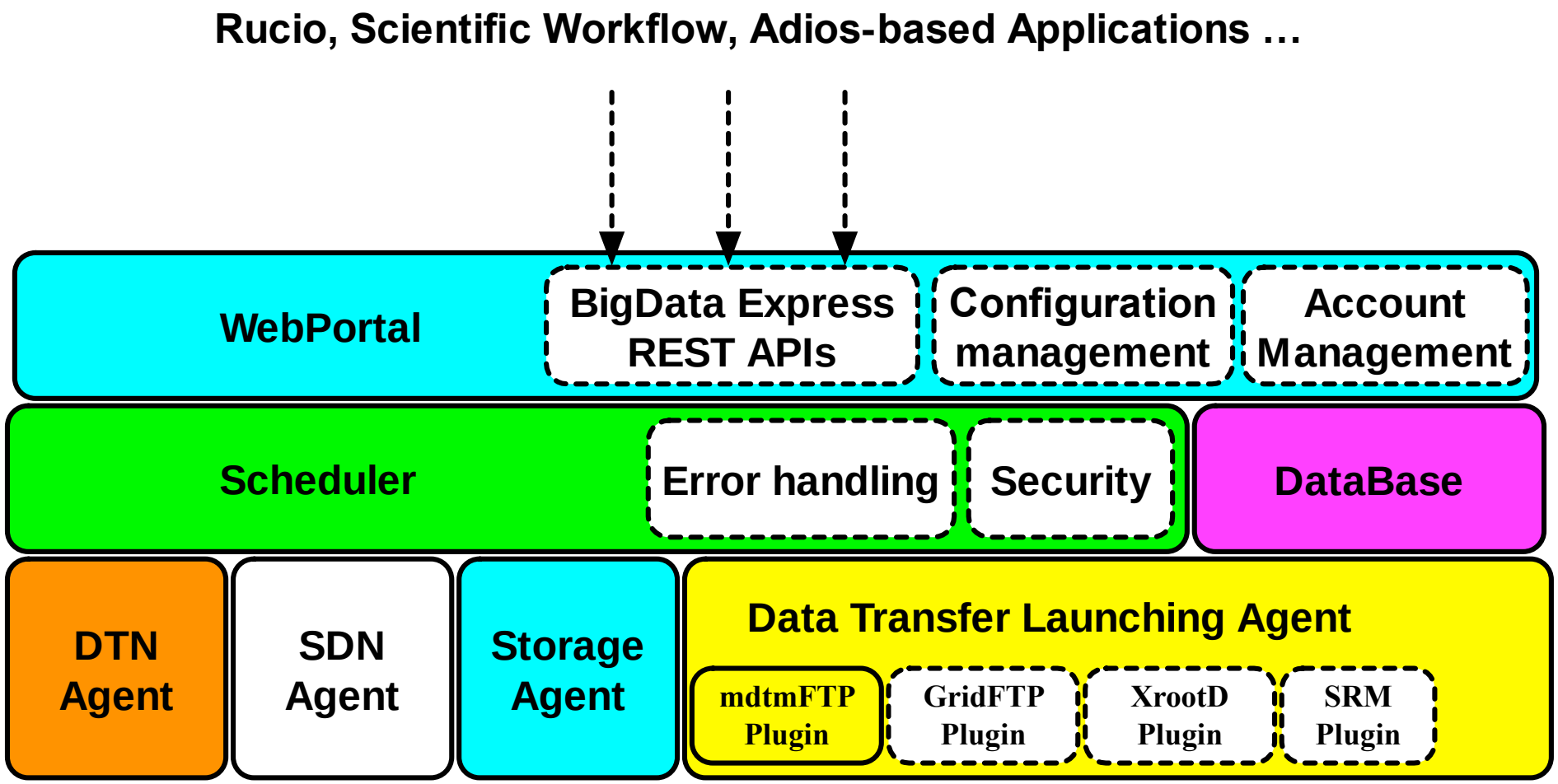
Supported Protocols	• Extensible plugin framework to support multiple protocols:
	  ○ mdtmFTP
	  ○ GridFTP, XrootD, SRM (coming soon)
		• GridFTP

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



Next Stage R&D Plan – Functional Perspective



More information about BigData Express

http://bigdataexpress.fnal.gov

Contact: wenji@fnal.gov