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

BigData Express Research Team

November 10, 2018

Many people's hard work

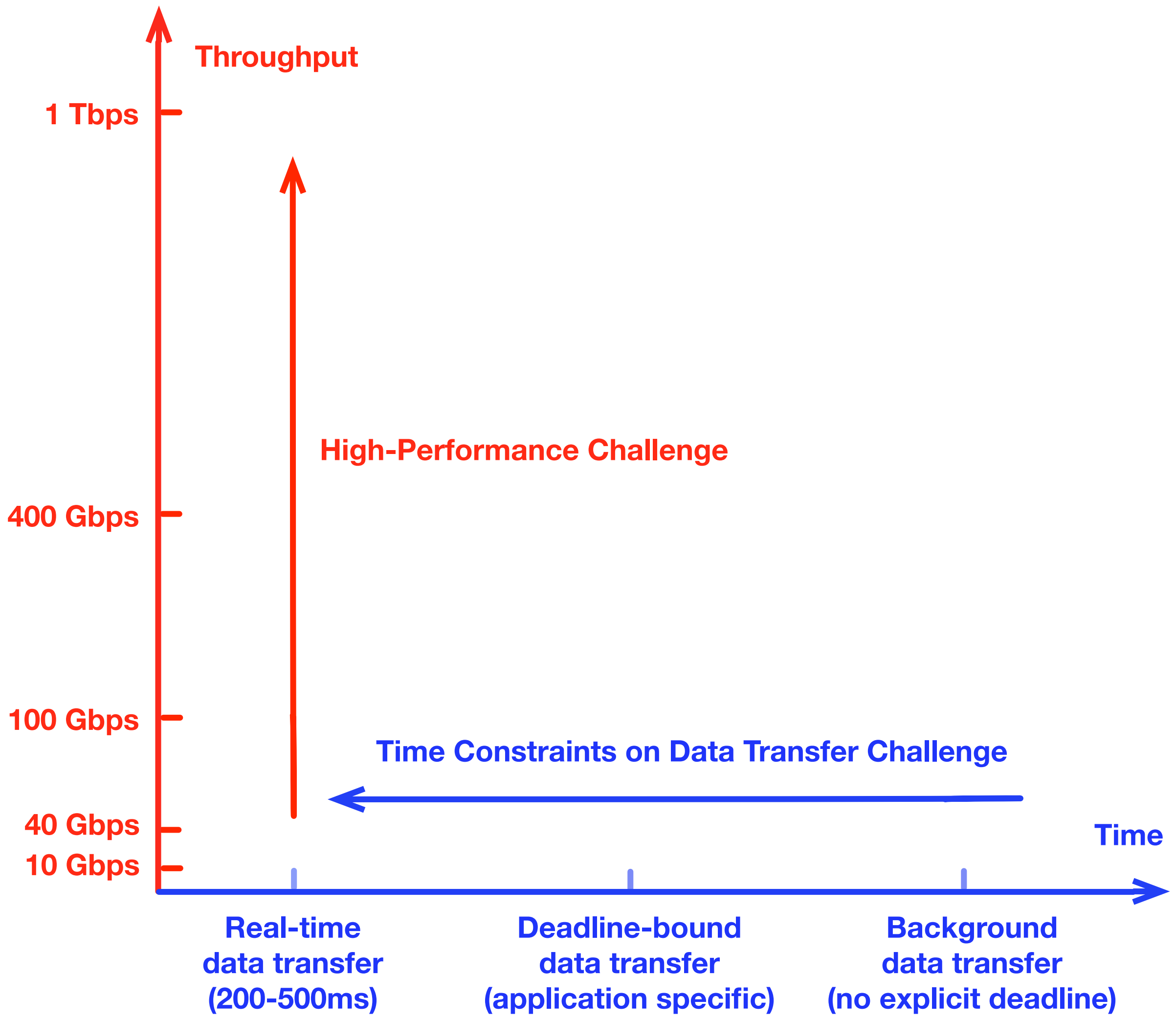
FNAL:	Qiming Lu, Liang Zhang, Sajith Sasidharan, **Wenji Wu**, Phil DeMar
ESnet:	Chin Guok, John Macauley, Inder Monga
iCAIR/StarLight:	Se-young Yu, Jim-Hao Chen, Joe Mambretti
KISTI:	Jin Kim, Seo-Young Noh,
Univ. of Maryland:	Xi Yang, Tom Lehman
ORNL:	Gary Liu

Acknowledgments

This work was supported by the U.S. DOE Office of Science ASCR network research program

U.S. DEPARTMENT OF ENERGY | Office of Science

Data Transfer Challenges in Big Data Era


Throughput
1 Tbps
400 Gbps
100 Gbps
40 Gbps
10 Gbps
High-Performance Challenge
Time Constraints on Data Transfer Challenge
Time
Real-time data transfer (200-500ms)
Deadline-bound data transfer (application specific)
Background data transfer (no explicit deadline)

- 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

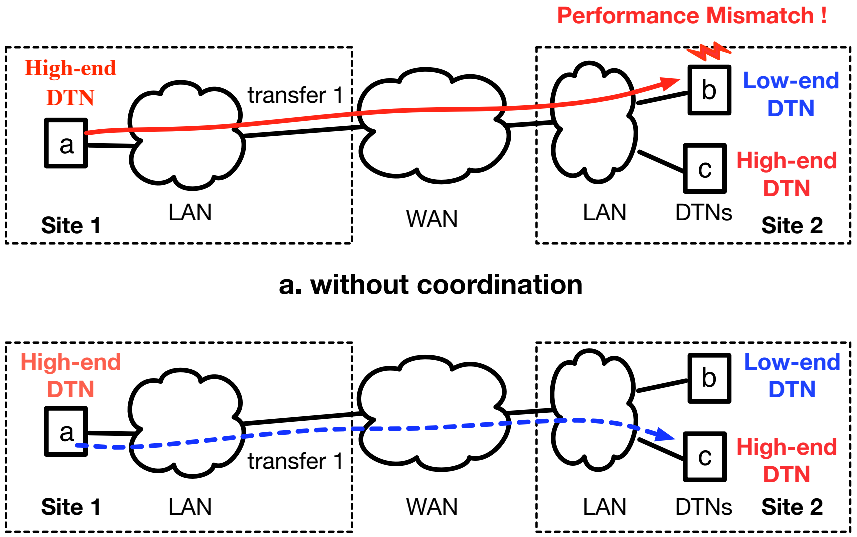
- Numerous enhancements
  - Parallelism at all levels
    - Multi-stream, Multicore, Multi-path parallelism
  - Science DMZ architecture
  - Terabit networks

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 DTNs

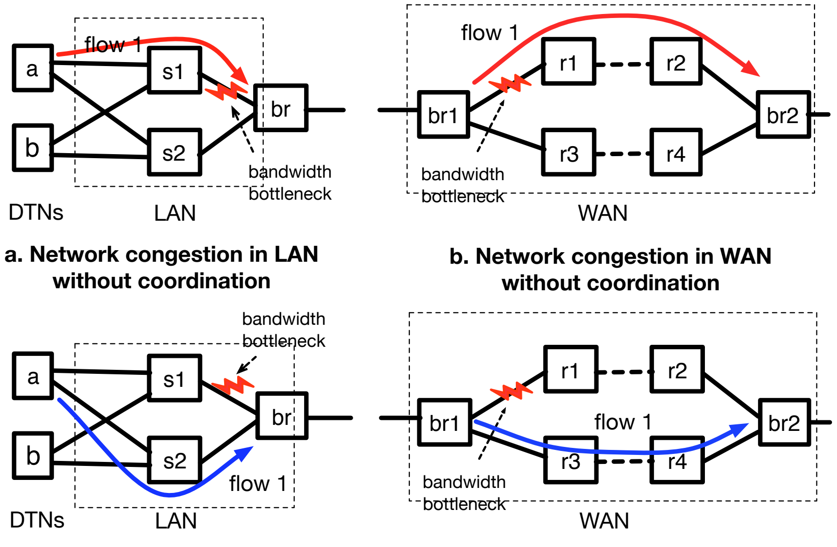
Problem 1 – Disjoint end-to-end data transfer loop

- Distributed resource management model
  - Resource contention
  - Performance mismatch



a. without coordination

b. with coordination



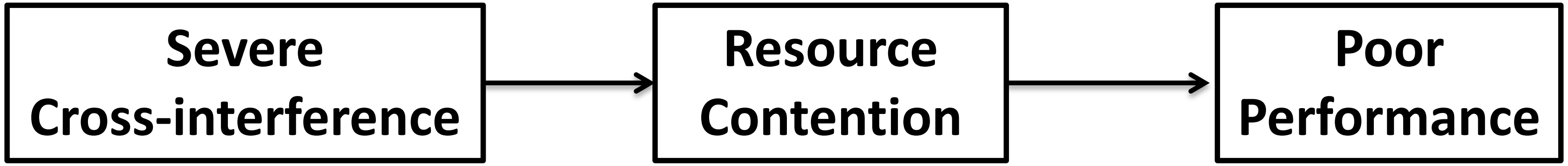
a. Network congestion in LAN without coordination

b. Network congestion in WAN without coordination

c. No network congestion in LAN with coordination

d. No network congestion in WAN with coordination

Problem 2 – Cross-interference between data transfers



Severe Cross-interference

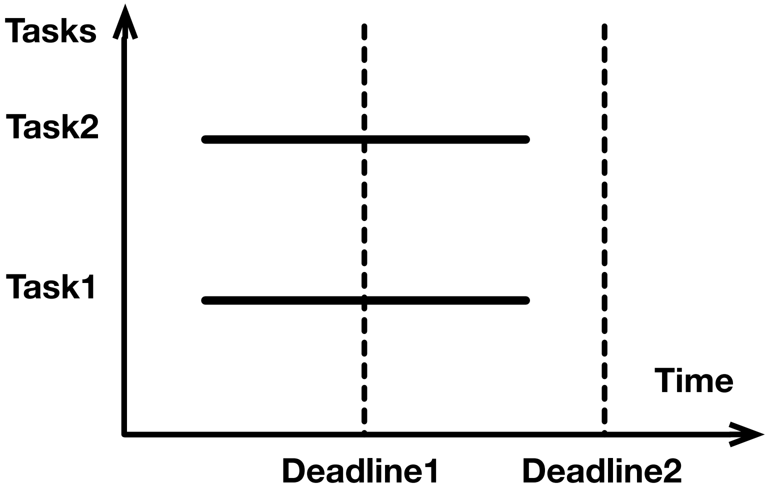
Resource Contention

Poor Performance

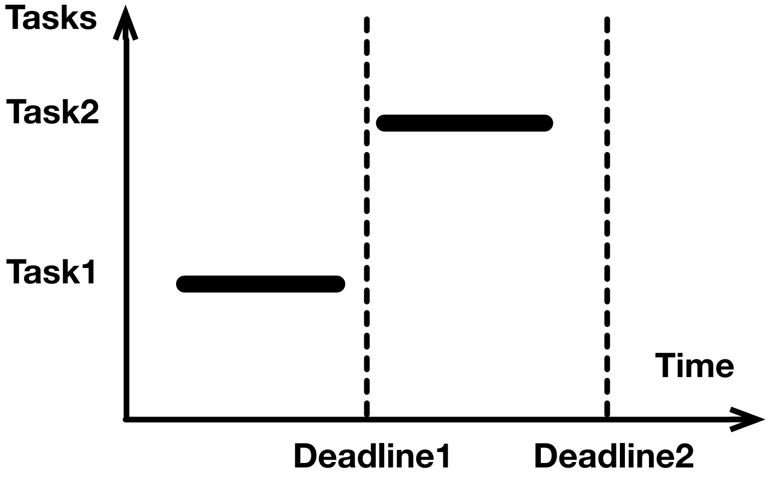
- Degraded performance
- High variability in data transfer performance

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

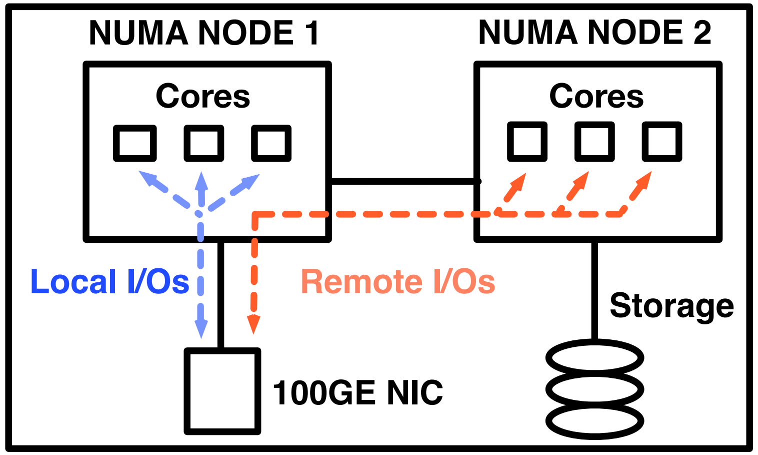


b. with deadline awareness

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

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

....



I/O locality problem on NUMA systems

Need high-performance data transfer tool!

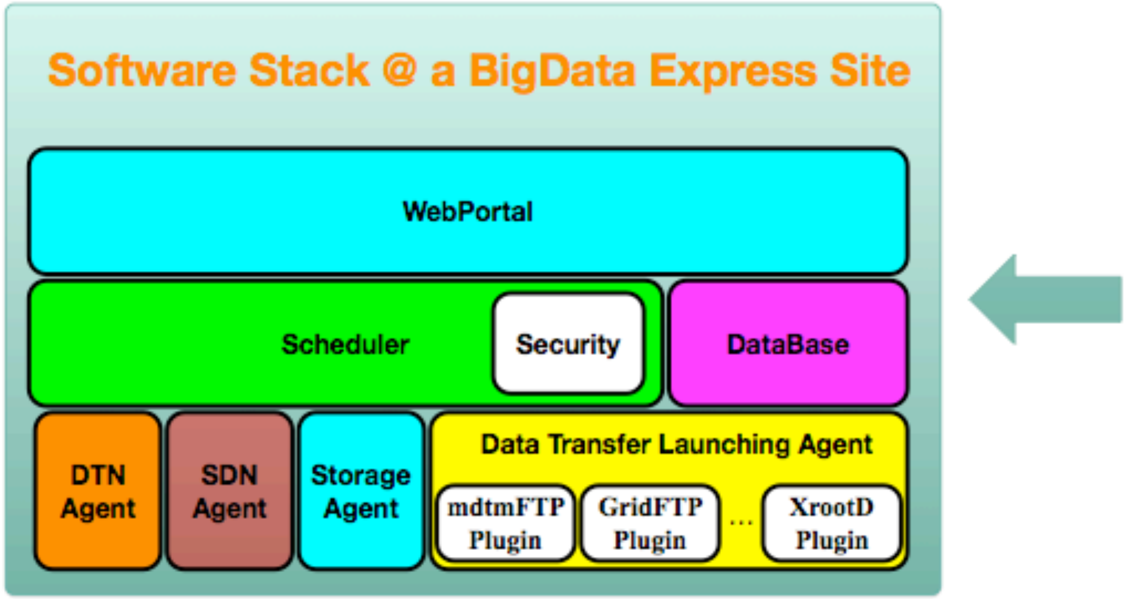
Our Solution - BigData Express

- 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
- **DTN Agent**
  - Manage and configure DTNs
  - Collect & report DTN configuration and status
- **Storage Agent**
  - Manage and configure storage systems
  - I/O estimation
- **Data Transfer Launching Agent**
  - Launch data transfer jobs
  - Support different data transfer protocols

BigData Express -- Distributed



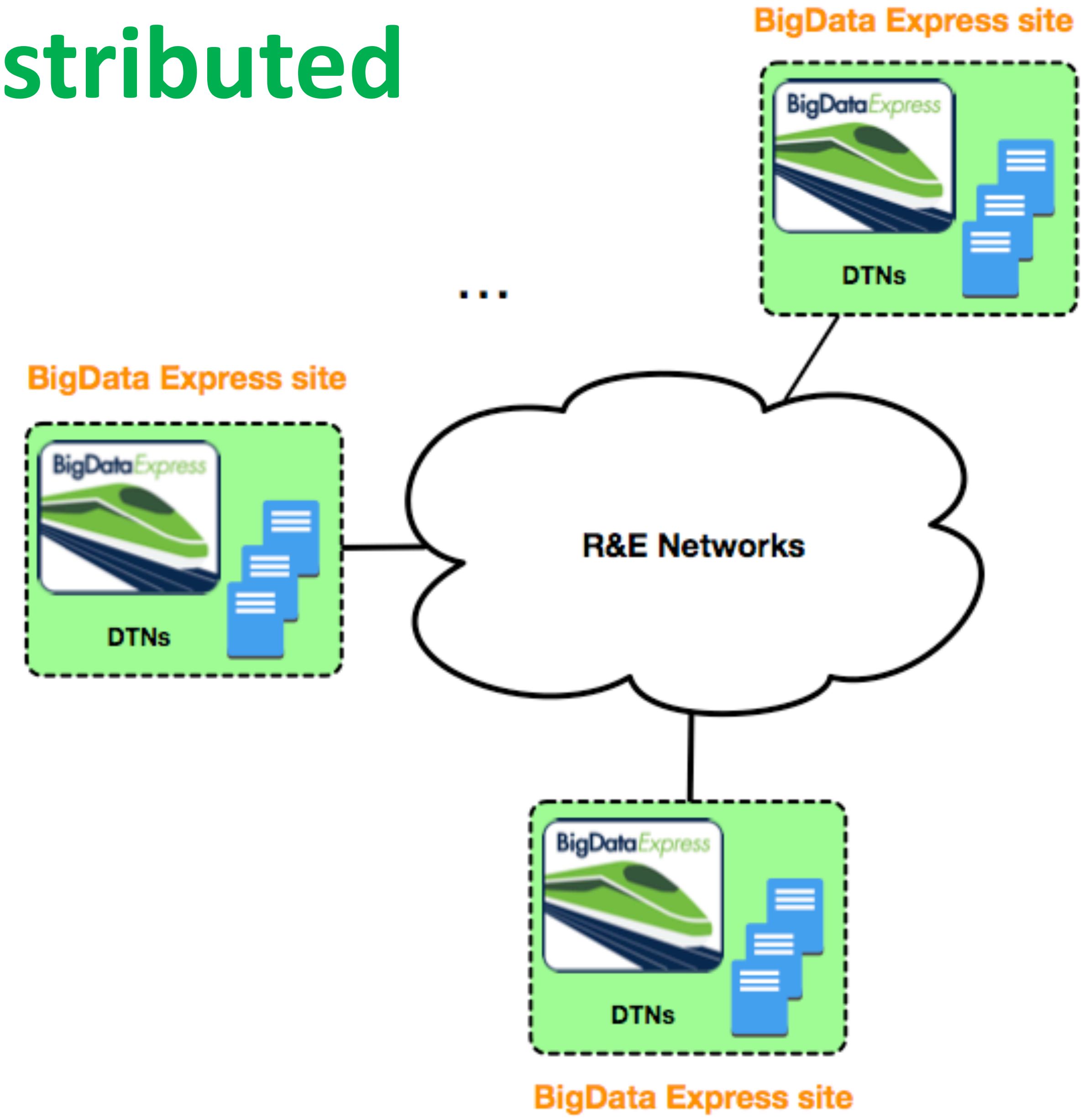
Software Stack @ a BigData Express Site

WebPortal

Scheduler | Security | DataBase

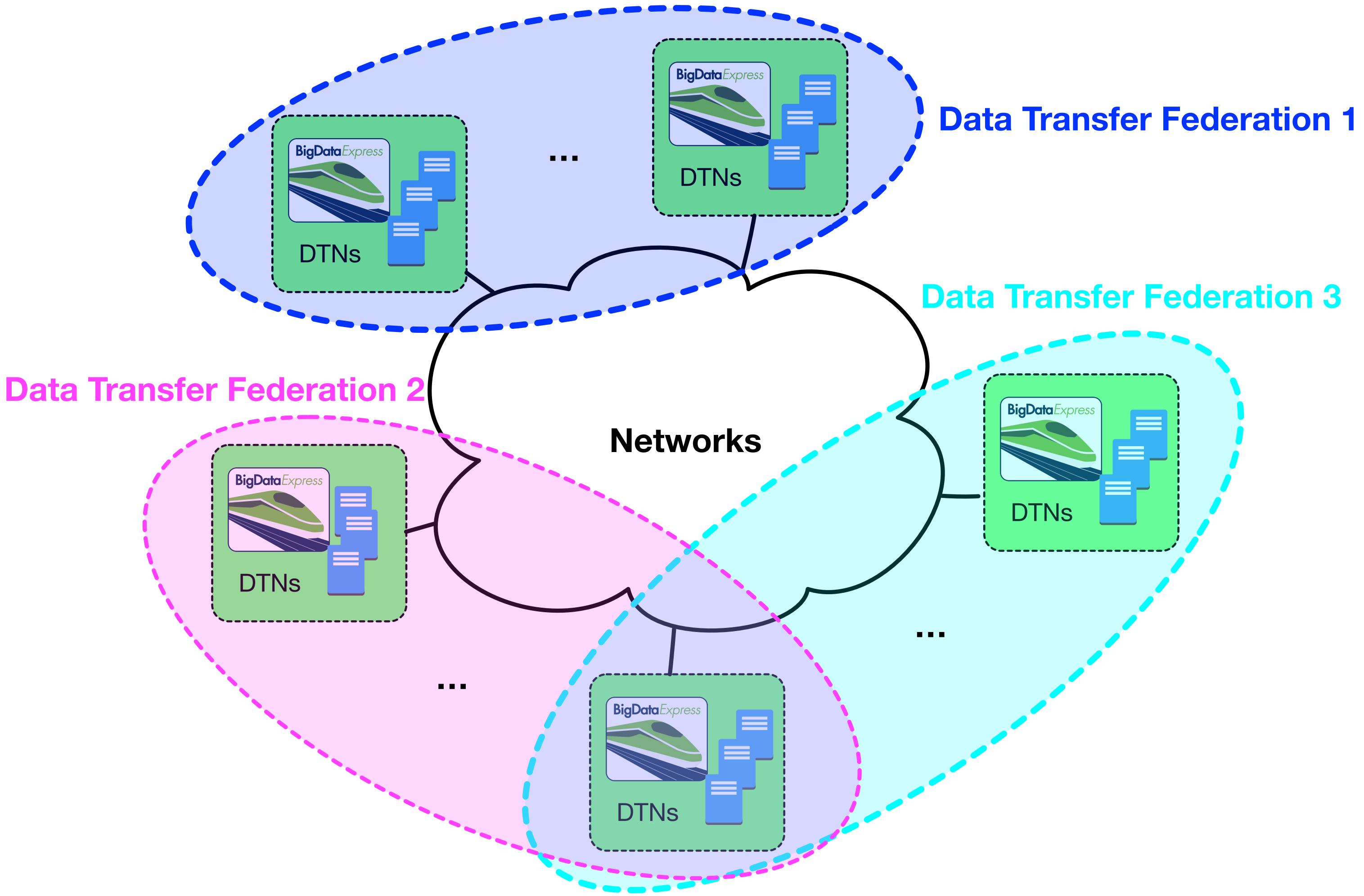
DTN Agent | SDN Agent | Storage Agent | Data Transfer Launching Agent

mdtmFTP Plugin | GridFTP Plugin | … | XrootD Plugin



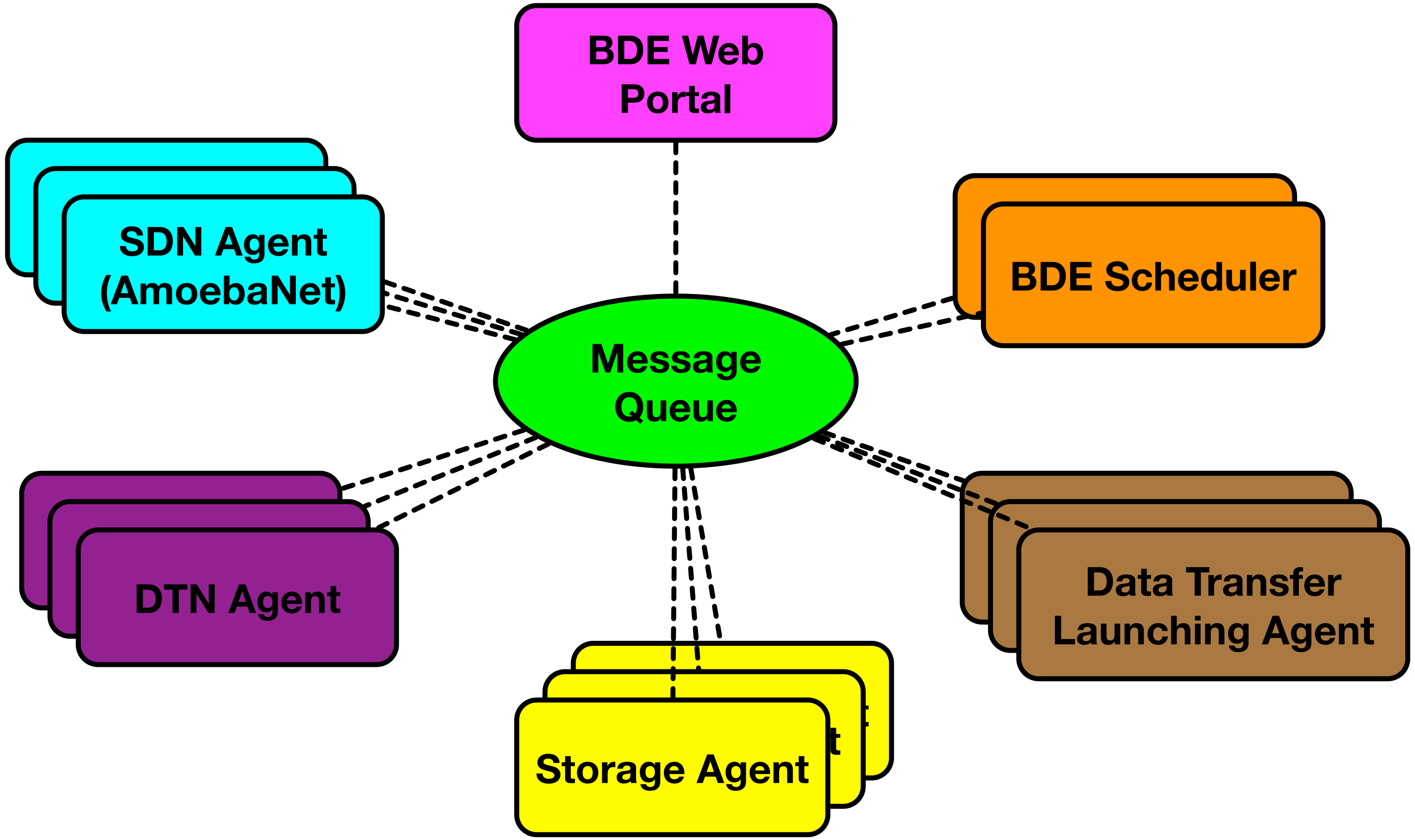
A Peer-to-Peer model

BigData Express -- Flexible


Data Transfer Federation 1
Data Transfer Federation 2
Data Transfer Federation 3
Networks
DTNs
DTNs
DTNs
DTNs
DTNs
BigData Express
...
...
...

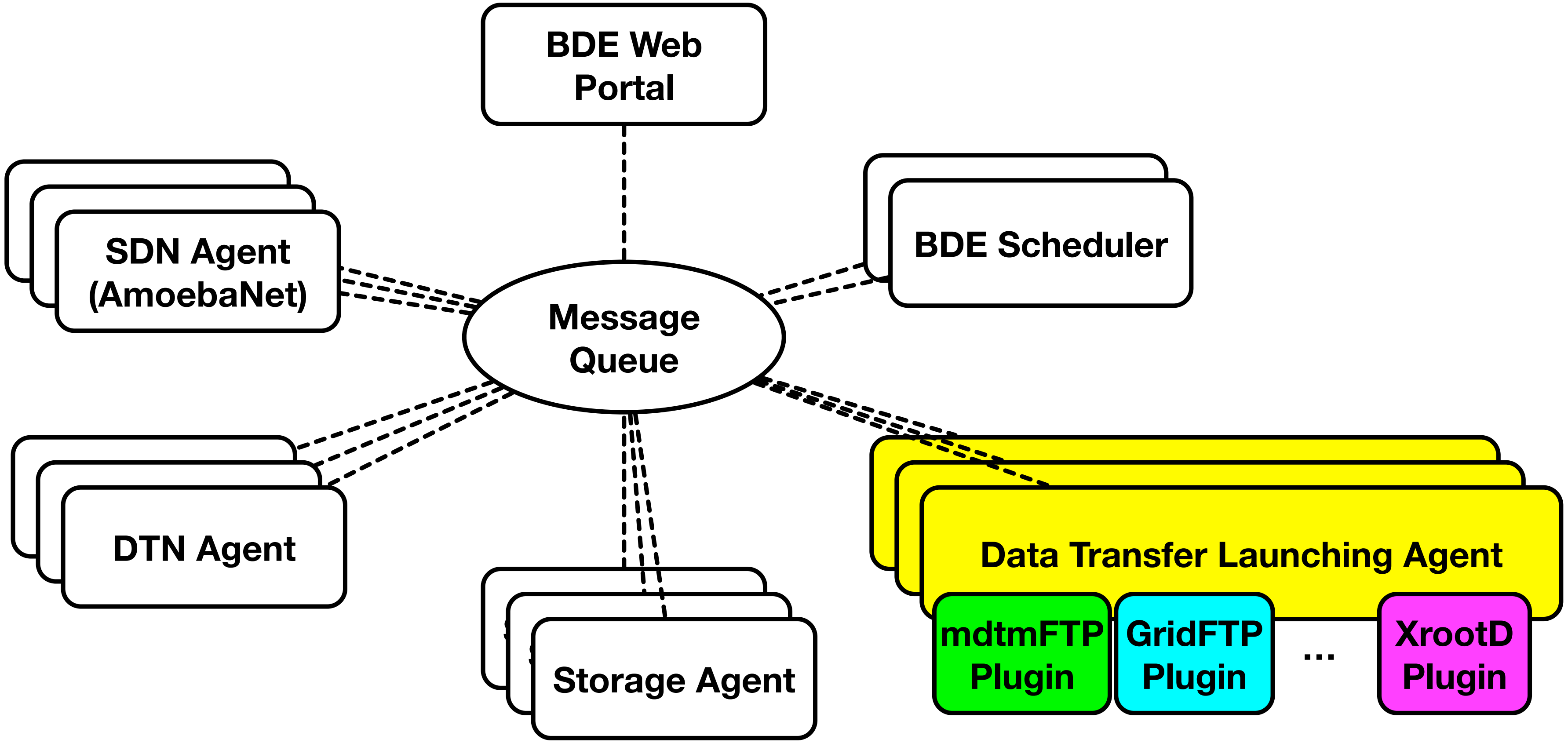
- Flexible to set up data transfer federations
- Providing inherent support for incremental deployment

BigData Express -- Scalable



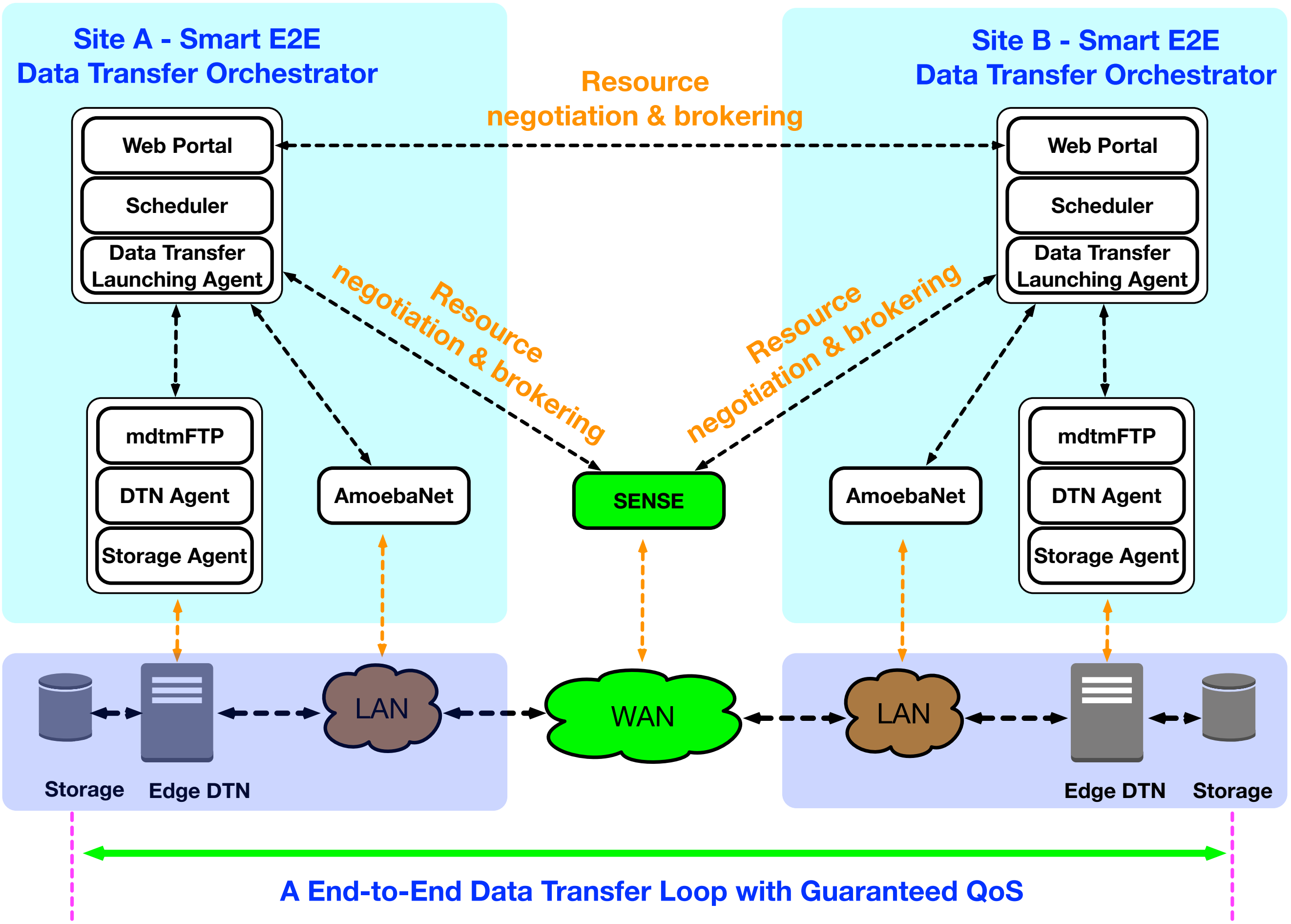
- BigData Express scheduler manages site resources through agents
- Use MQTT as message bus

BigData Express -- Extensible



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

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



Site A - Smart E2E Data Transfer Orchestrator
Resource negotiation & brokering
Site B - Smart E2E Data Transfer Orchestrator
Web Portal
Scheduler
Data Transfer Launching Agent
Web Portal
Scheduler
Data Transfer Launching Agent
Resource negotiation & brokering
Resource negotiation & brokering
mdtmFTP
DTN Agent
Storage Agent
AmoebaNet
SENSE
AmoebaNet
mdtmFTP
DTN Agent
Storage Agent
LAN
WAN
LAN
Storage
Edge DTN
Edge DTN
Storage
A End-to-End Data Transfer Loop with Guaranteed QoS

- 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

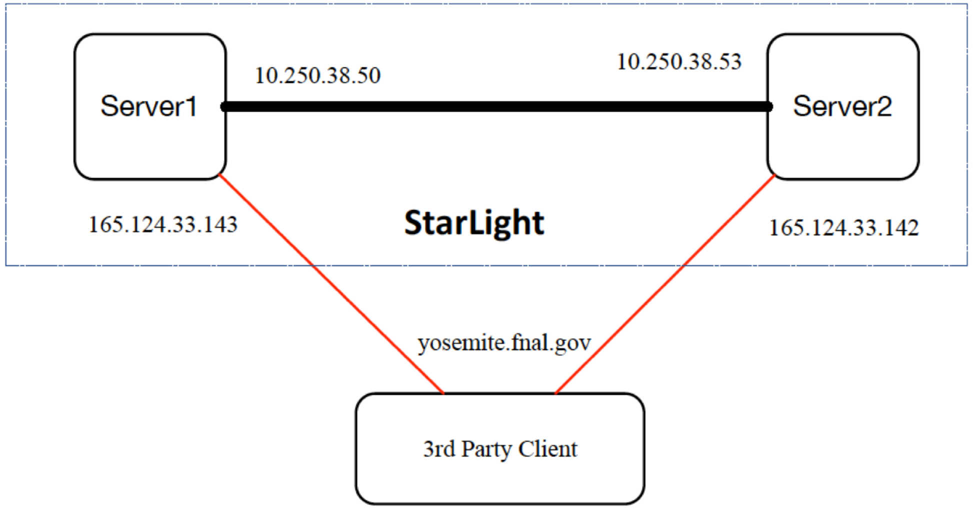
Time-to-completion (Seconds) – 3rd party mode   Lower is better

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)

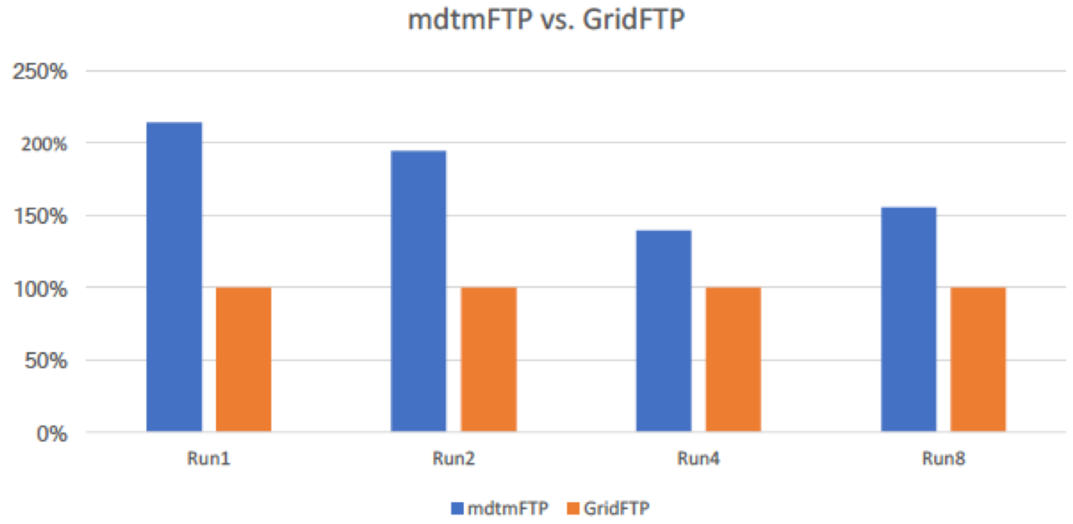
STARLIGHT SDX



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 -- Three Types of Data Transfer

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

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*

BigData Express SC18 DEMO

BigData*Express*

Fermilab

STARLIGHT SDX

OAK RIDGE National Laboratory

ESnet Energy Sciences Network

KREONET

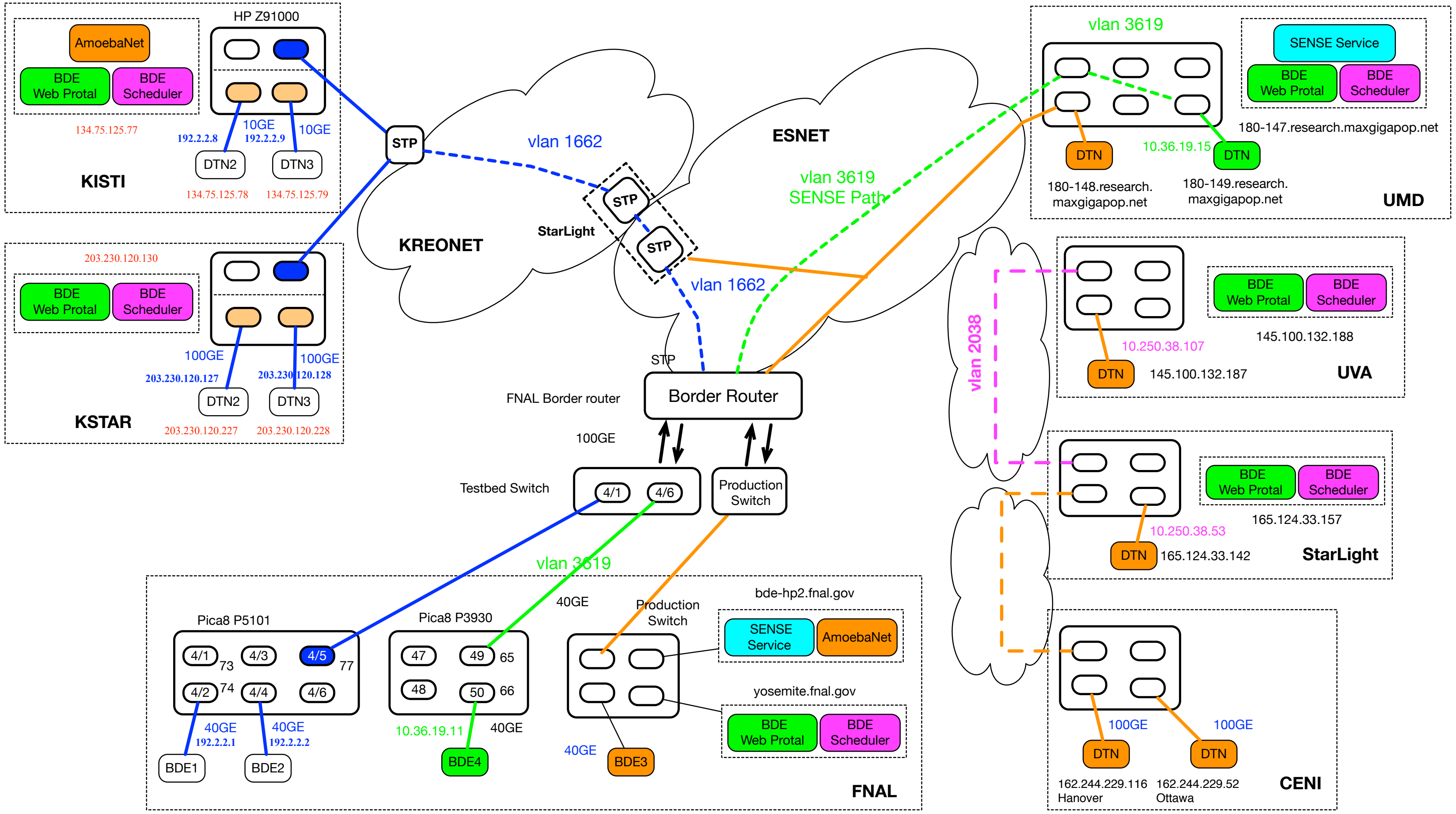
MAX MID-ATLANTIC CROSSROADS

KISTI www.kisti.re.kr

KSTAR

SURF NET

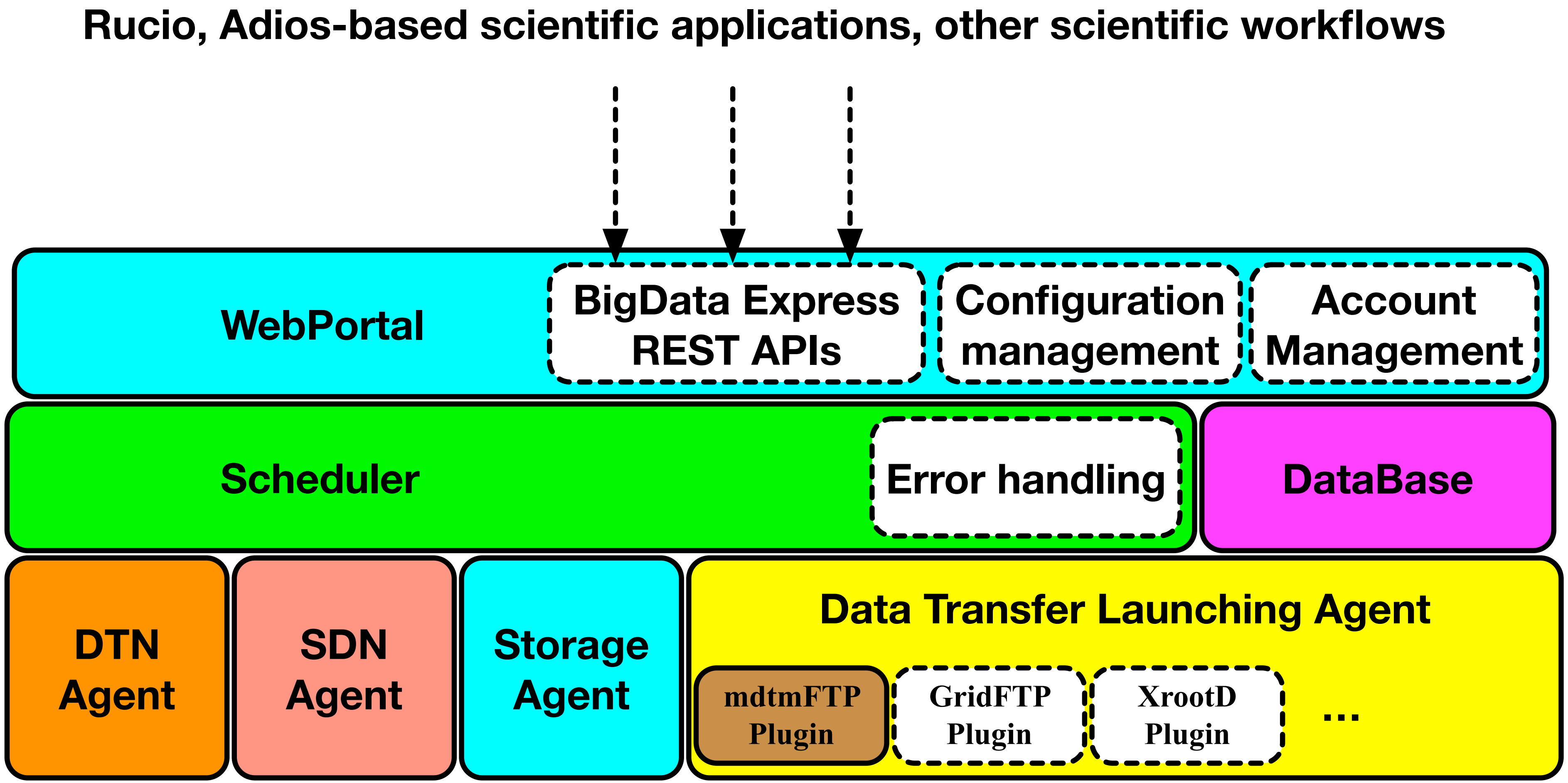
ciena Experience. Outcomes.



BigData Express -- Deployment

- **Asia**
  - KISTI, South Korea
    - https://sc-demo-01.sdfarm.kr:2888/
  - KSTAR, South Korea
    - Https://203.230.120.130:8080
- **Europe**
  - University of Amsterdam, Netherlands
    - https://bde-01.lab.uvalight.net/
- **North America**
  - Fermilab
    - https://Yosemite.fnal.gov:5000
  - StarLight, Northwestern University
    - https://starlight.bigdataexpress.website/
  - UMD/MAX, University of Maryland, College Park
    - https://180-147.research.maxgigapop.net/

Next Stage R&D Plan – Functional Perspective



More information about BigData Express

http://bigdataexpress.fnal.gov

Contact: wenji@fnal.gov

This document was prepared by BigData Express using the resources of the Fermi National Accelerator Laboratory (Fermilab), a U.S. Department of Energy, Office of Science, HEP User Facility. Fermilab is managed by Fermi Research Alliance, LLC (FRA), acting under Contract No. DE-AC02-07CH11359.