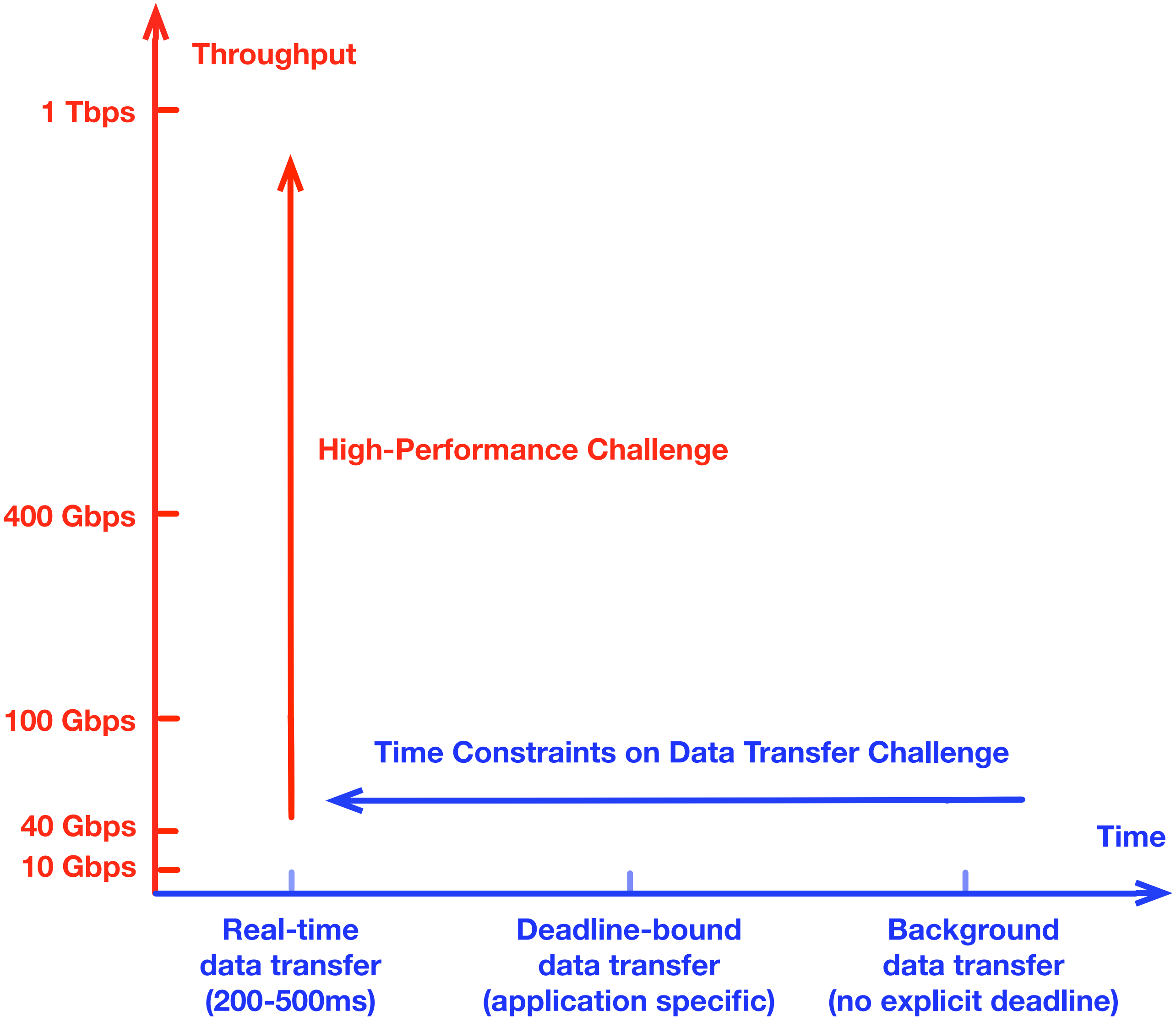
BigData Express

W. Wu, P. Demar, L. Zhang, Q. Lu (FNAL)
G. Liu, N. Podhorszki (ORNL)

DOE/SC/ASCR SDN Projects meeting
Fermilab, Batavia IL
Feb 17, 2016

1. The Challenges

DOE Data Transfer Challenges


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)

Data Transfer – State of the Art

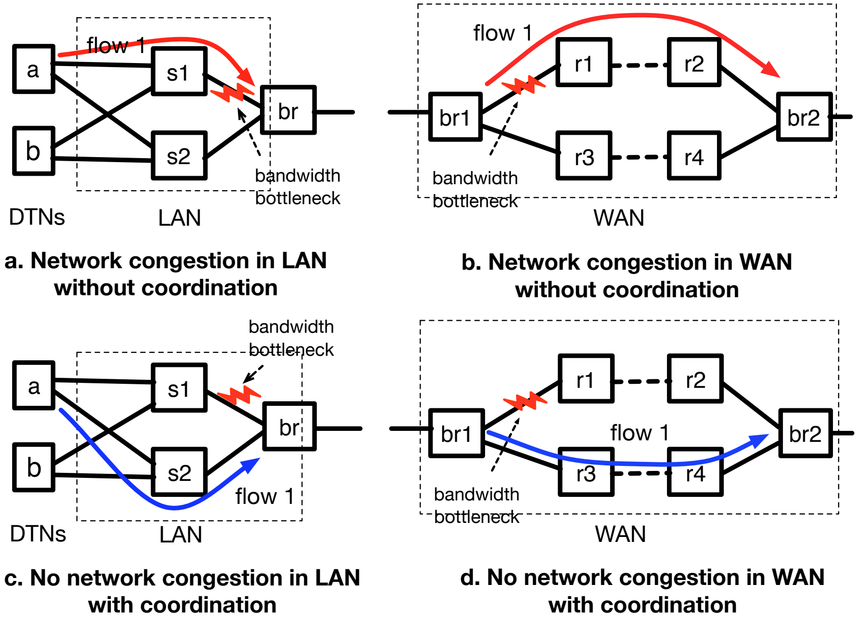
- Advanced data transfer tools and services developed
  - GridFTP, BBCP
  - PhEDEx, LIGO Data Replicator, Globus Online

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

Existing data transfer tools and services will NOT be able to successfully address the challenges of data transfer to support extreme-scale science applications

Problems with existing data transfer tools and services – Problem 1

- Disjoint end-to-end data transfer loop



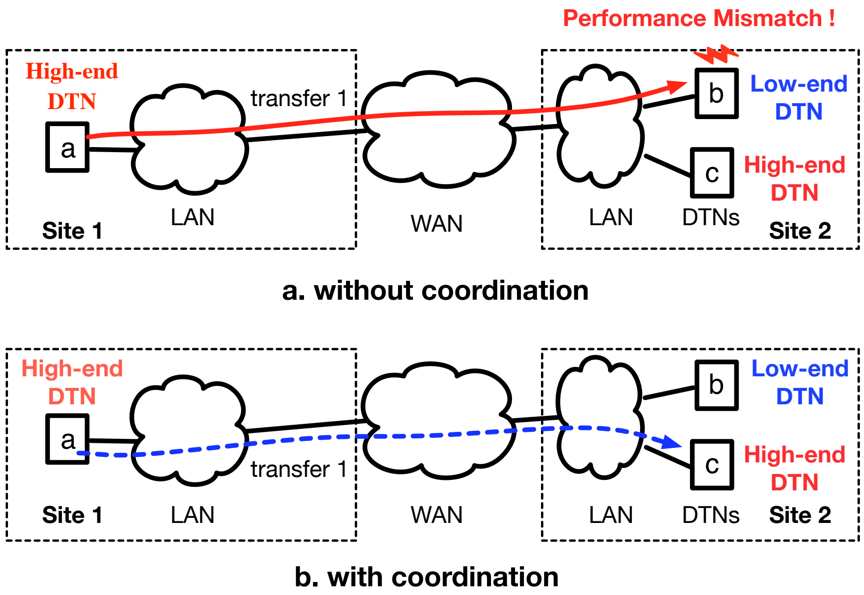
a. Network congestion in LAN without coordination

b. Network congestion in WAN without coordination

c. No network congestion in LAN with coordination

d. No network congestion in WAN with coordination

Network congestion



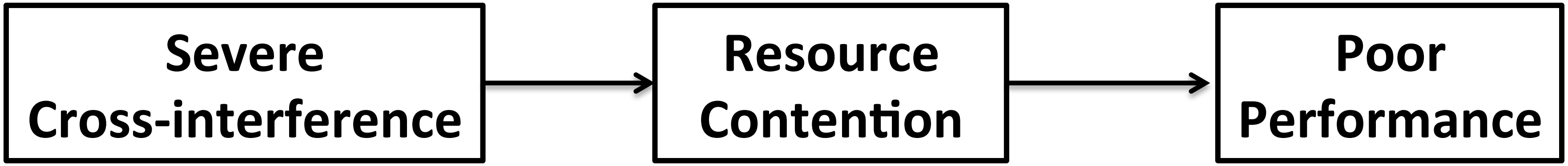
a. without coordination

b. with coordination

DTN performance mismatch

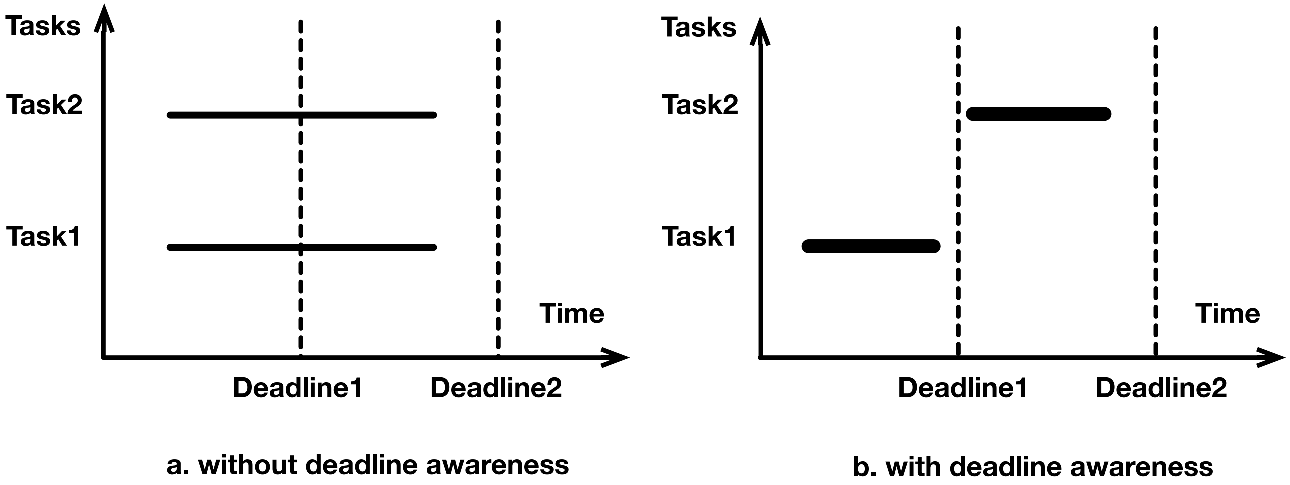
Problems with existing data transfer tools and services – Problem 2

- Cross-interference between data transfers



Problems with existing data transfer tools and services – Problem 3

- Oblivious to user requirements (e.g., deadlines and Qos requirements)



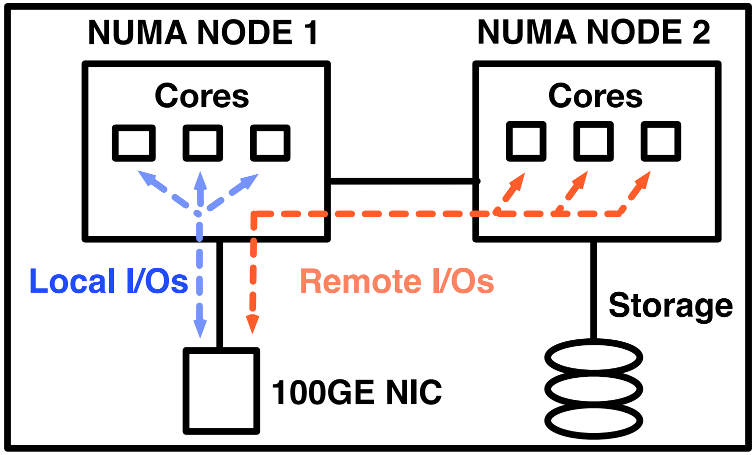
a. without deadline awareness

b. with deadline awareness

Data transfer with and without deadline awareness

Problems with existing data transfer tools and services – Problem 4

- Inefficiencies arise when existing data transfer tools are run on DTNs.



The parallelism vs. I/O locality problem on NUMA systems

Our Solution

- **The BigData Express Project**
  - Collaborative effort by Fermilab and Oakridge National Laboratory
  - Funded by DOE's Office of Advanced Scientific Computing Research (ASCR)
  - A three-year research project
  - http://bigdataexpress.fnal.gov

BigData Express seeks to provide a schedulable, predictable, and high-performance data transfer service for DOE's large-scale science computing facilities (e.g., LCF, US-LHC computing facilities)

2. BigData Express

2.1 Architecture & Design

BigData Express

- A distributed middleware system that will provide a schedulable, predictable, and high-performance data transfer services for the DOE's large-scale science facilities and their collaborators.
- It has two versions
  - A full site version, designed for large data centers
  - A single-node, targeted at small research groups

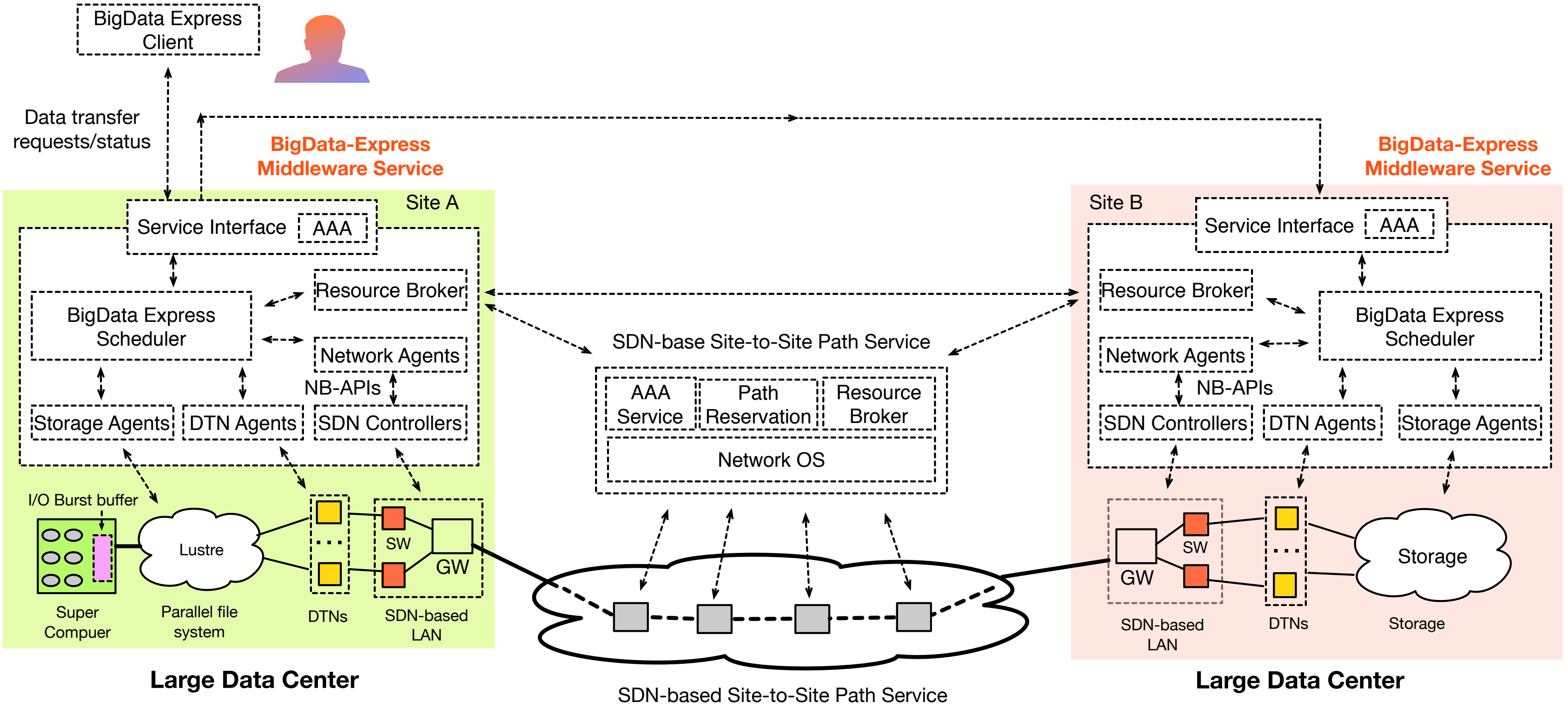
Key Features

- A data-transfer-centric architecture to seamlessly integrate and efficiently coordinate the various resources in an end-to-end loop
  - Directly schedule various local resources within a site
  - a distributed rate-based resources brokering mechanism to coordinate resources across sites
  - A distributed DTN matching mechanism to coordinate and match heterogeneous DTNs at different sites to avoid DTN performance mismatch
- A time-constraint-based scheduler to schedule data transfer tasks

Key Features (cont.)

- An admission control mechanism to provide guaranteed resources for admitted data transfer tasks
- An end-host-based rate control mechanism to improve data transfer schedulability and reduce cross-interference between data transfers
- Extensive use of SDN to improve network I/O performance
- The leveraging of SDS to improve storage I/O performance

BigData Express – Full site version



A large data center typically features

- A dedicated cluster of high-performance DTNs
- An SDN-based BigData Express LAN
- A large-scale storage system

Major entities

- BigData Express scheduler
  - Coordinate all activities at each BigData Express site
  - Manage and schedule local resources (DTNs, storage, and BigData Express LAN through agents (DTN agents, storage agents, and network agents)
  - BigData Express scheduler at different sites will collaborate to execute data transfer tasks.

Major entities

- The service interface
  - – Authenticate, authorize, and audit users and user applications
  - – Allow user to access BigData Express services
  - – For a data transfer task, the following info will be conveyed to BigData Express via the service interface
    - The credentials of the task submitter
    - The paths and filenames of the data SRC/DST
    - The task deadline
    - The Qos requirements

Major entities

- DTN agents
  - Collect and report the DTN configuration and status
  - Assign DTNs to data transfer tasks as requested by the BigData Express scheduler

- Network agents
  - Keep track of the BigData Express LAN topology and traffic status with the aid of SDN controllers
  - Reliably updating SDN-enabled switch rules as requested by the BigData Express scheduler to assign local paths for data transfer

Major entities

- SDN Controller
  - Open-source network operating system (e.g., ONOS)
  - The network agents access the SDN controllers through northbound APIs

- Storage agents
  - Keep track of the usage of local storage systems
  - Provide information regarding storage resources availability to the scheduler
  - Execute storage assignment

Major entities

- Resource broker
  - Implement a distributed rate-based resource brokering mechanism to coordinate resource allocation across autonomous sites

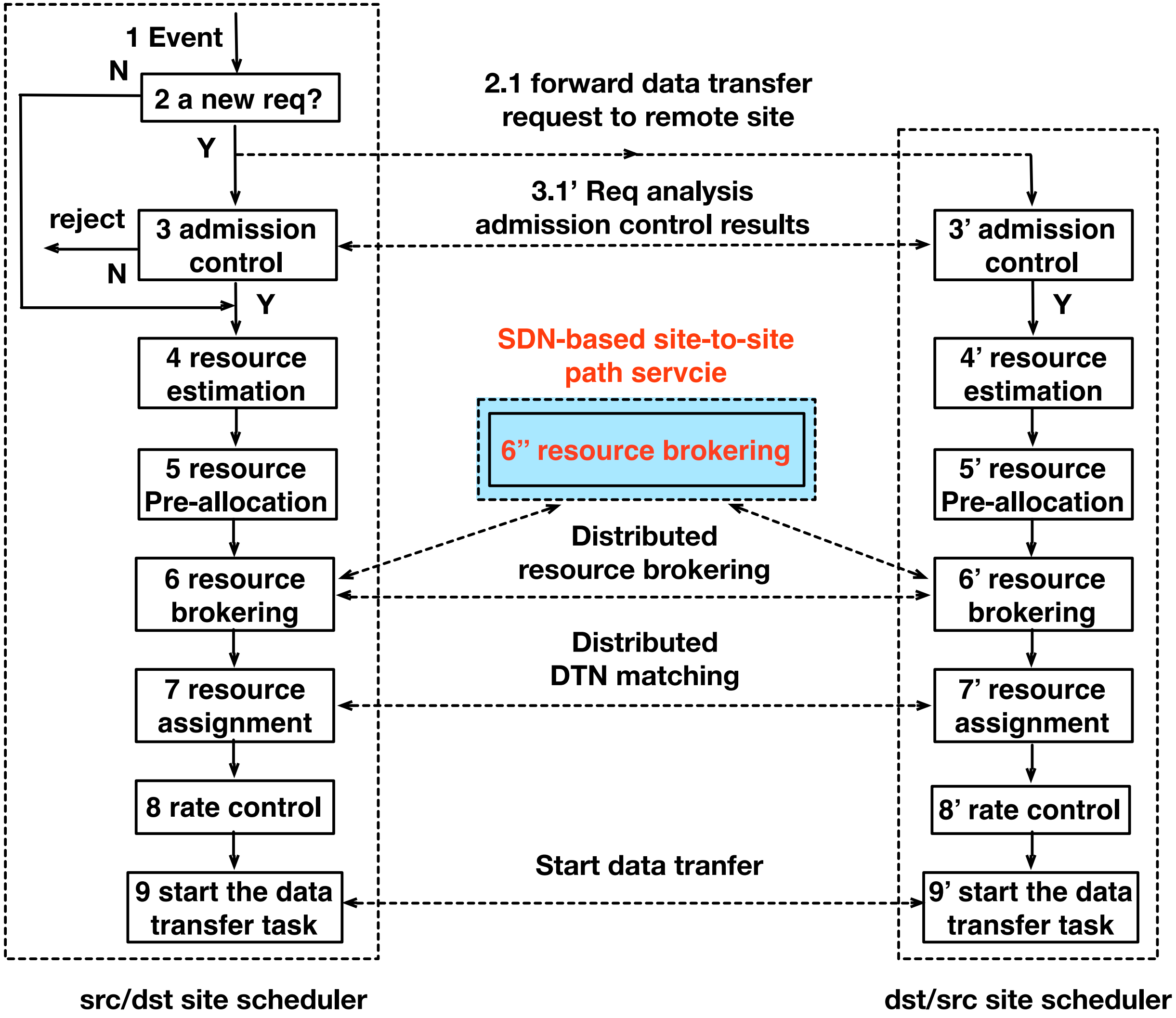
How does BigData Express work? (1)

- The BigData Express scheduler implements a time-constraint-based scheduler to schedule resource for data transfer tasks
- Each resource will be estimated, calculated, and converted into a rate that can be apportioned to data transfer tasks

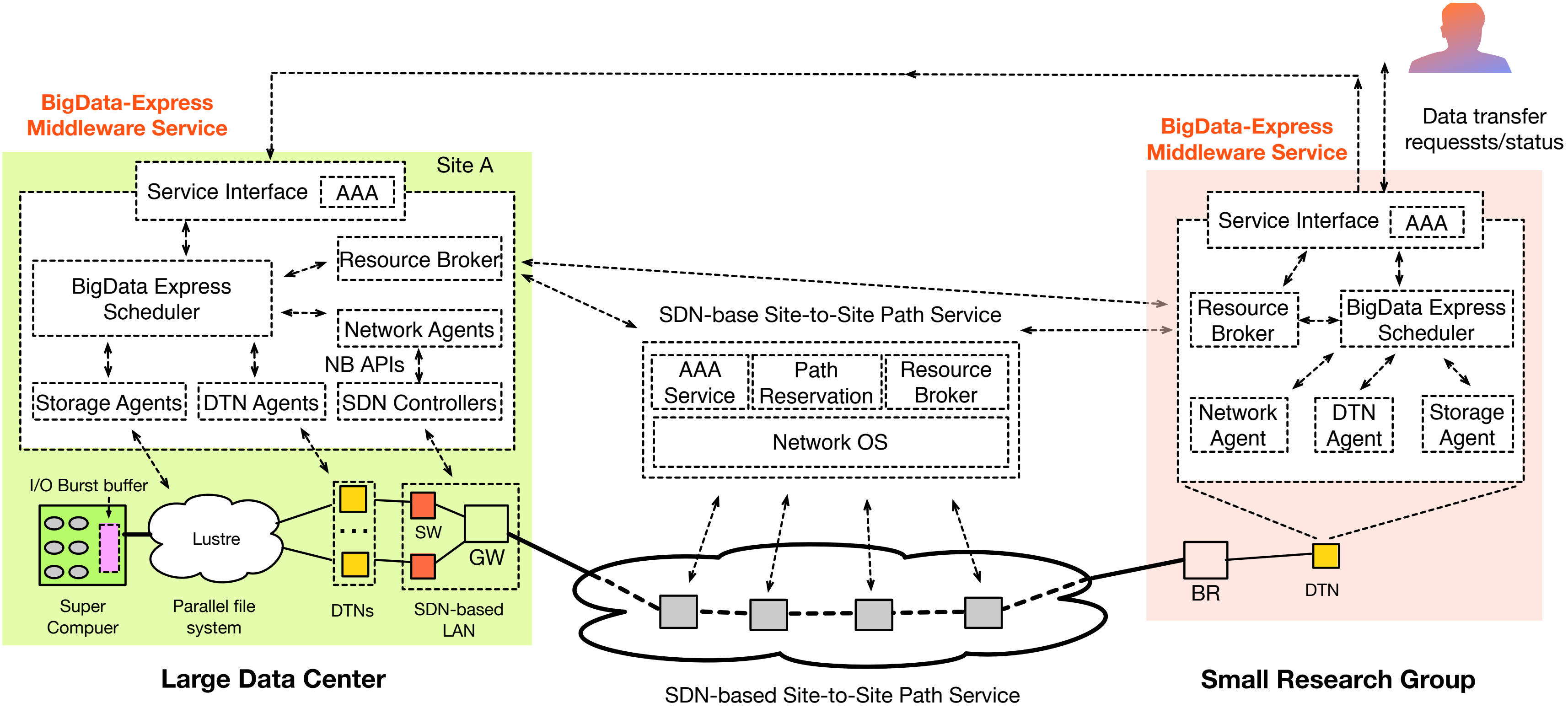
How does BigData Express work? (2)

- On an event-driven or periodic basis, the scheduler will perform the following tasks:
  - Resource estimation and calculation
  - Resource pre-allocation
  - Resource brokering
  - Resource assignment

How does BigData Express work? (3)



BigData Express – Single node version



A small research group typically features

- One or two DTNs, with each DTN having local storage units
- In the Science DMZ architecture, DTNs are directly connected to a border router, No dedicated network slices allocated for DTNs

BigData Express – Single node version

- A single-node version of BigData Express can be deployed at each DTN.
- The single-node version of BigData Express will have functions that are similar to, but less complex than, those of the full site version.

2.2 BigData Express Research Areas

Research Areas (FNAL) – 1

- SDN-enabled network and northbound APIs
  - Transforming network into schedulable resources
  - Calculate optimum network paths and generate suitable OpenFlow rules

Research Areas (FNAL) – 2

- Resource estimation
  - DTN I/O capacity estimation
  - Network throughput estimation and calculation
- Resource pre-allocation
  - Maximizing the number of data transfer tasks whose requirement are satisfied
  - Fully utilizing local site resources

Research Areas (FNAL) – 3

- Resource brokering
  - Algorithm & protocol
- Resource assignment
  - Network assignment
  - DTN assignment and matching
- Admission control
  - A data transfer that cannot satisfy its time constraints without violating others will not be admitted

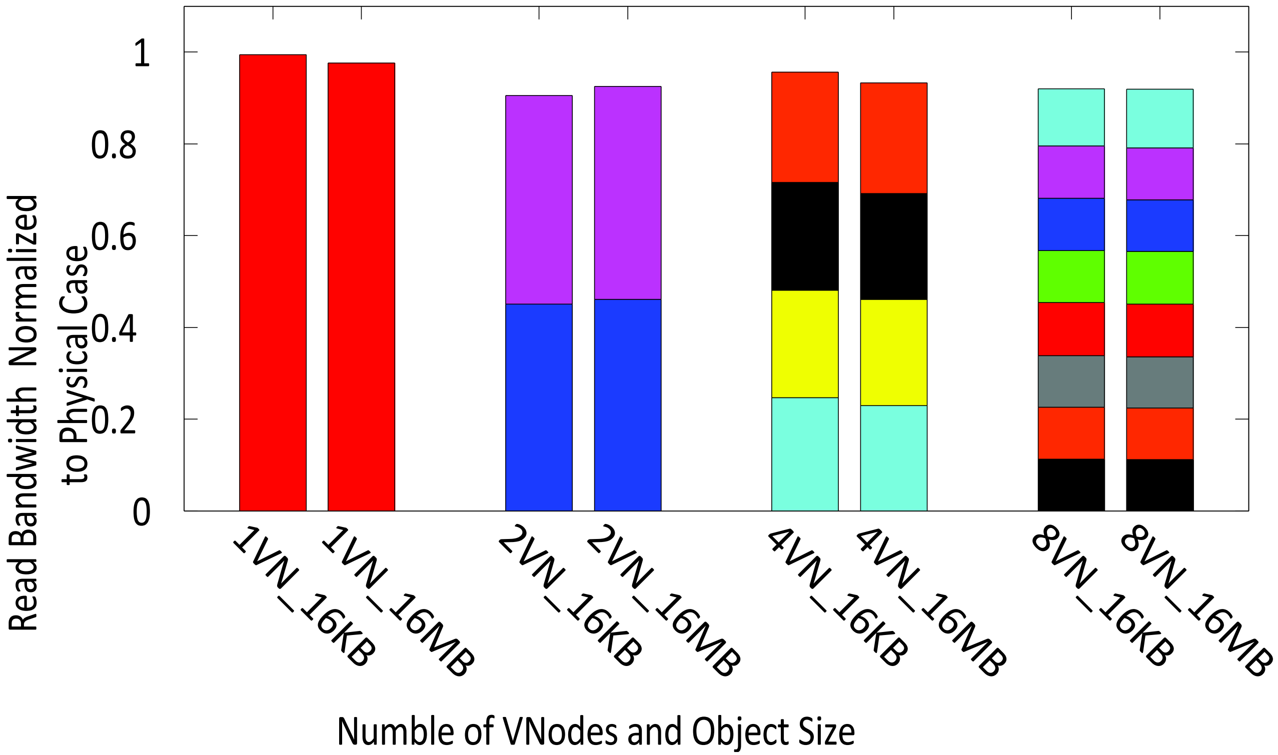
Research Areas (FNAL) – 4

- Employing SDN and multi-NIC parallelism to improve DTN performance
  - Improving I/O locality
  - Maximizing parallelism

Research Areas (ORNL) – 1

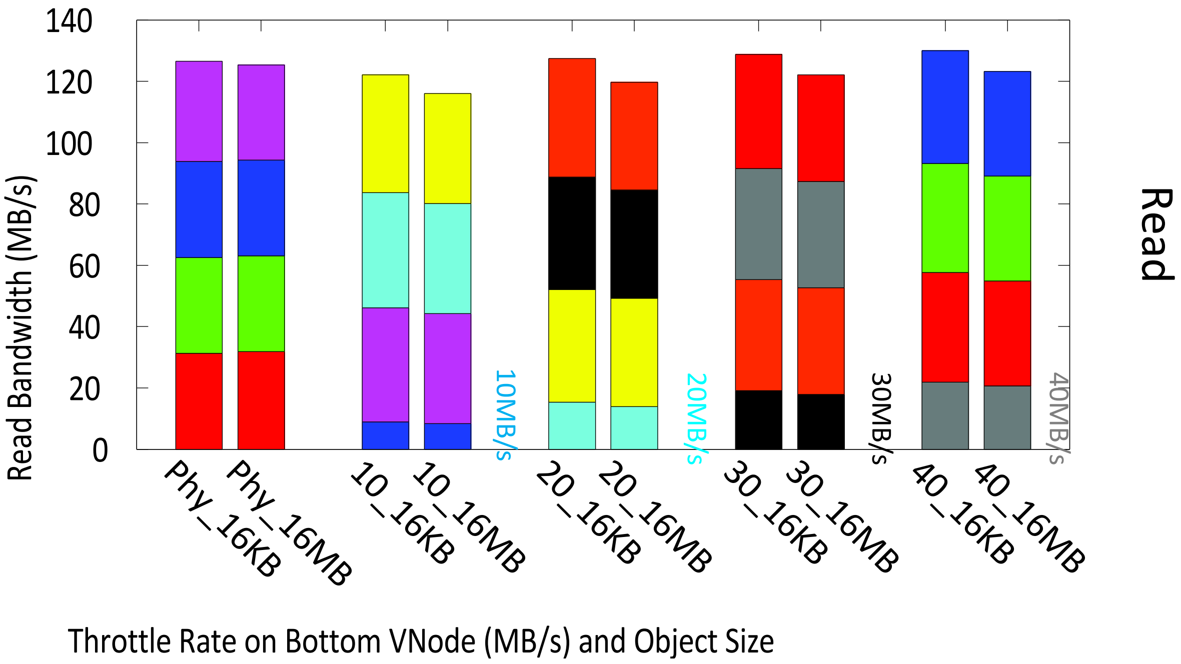
- Managing Block I/O via OS-level Virtualization
- Two vehicles for allocating block I/O in CGroups module.
  - Throttling functionality
    - Set an upper limit to a process group's block I/O
  - Weight functionality
    - Assign shares of block I/O to a group of processes

Read Overhead on Single Node



The worst read overhead is less than10%.

Throttling Read on Single Node



The throttle functionality could guarantee the process's I/O does not exceed the upper limits. But it is largely influenced by other concurrent processes

Research Areas (ORNL) – 2

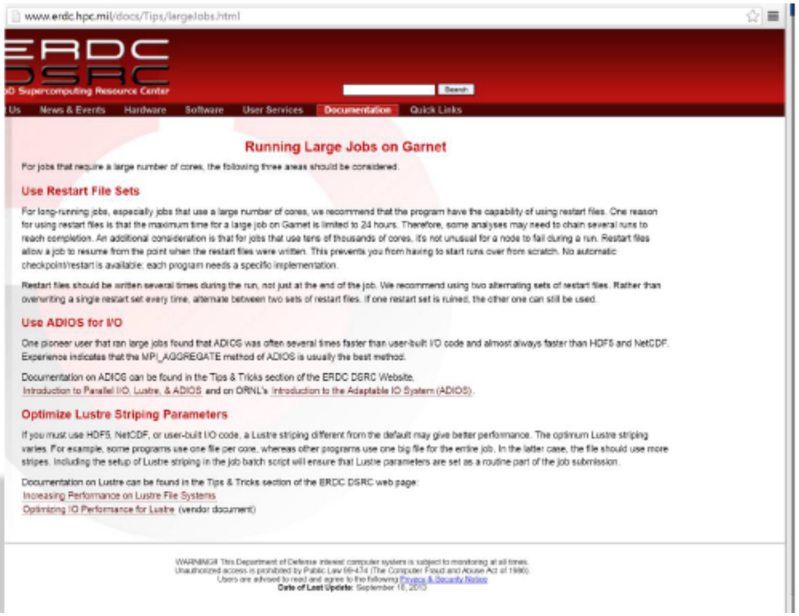
- Storage resource estimation at facilities where Linux containers are disabled
  - Leveraging the bursty nature of HPC I/O
  - Learning techniques

Research Areas (ORNL) – 3

- To support complex data pipeline to achieve data reduction, filtering, and transformation.
  - Leveraging ADIOS to compose efficient data pipeline

Impact at the HPC User facilities

- ALCF
- OLCF
- NERSC
- Tiahne-1A
- Tiahne-2
- Bluelight
- Singapore
- KAIST
- Ostrava
- Dresden
- ERDC
- CSCS
- Blue Waters
- EPFL
- Barcelona Supercomputing Center



Running Large Jobs on Garnet

For jobs that require a large number of cores, the following three areas should be considered.

Use Restart File Sets

For long-running jobs, especially jobs that use a large number of cores, we recommend that the program have the capability of using restart files. One reason for using restart files is that the maximum time for a large job on Garnet is limited to 24 hours. Therefore, some analyses may need to chain several runs to reach completion. An additional consideration is that for jobs that use tens of thousands of cores, it's not unusual for a node to fail during a run. Restart files allow a job to resume from the point when the restart files were written. This prevents you from having to start runs over from scratch. No automatic checkpoint/restart is available; each program needs a specific implementation.

Restart files should be written several times during the run, not just at the end of the job. We recommend using two alternating sets of restart files. Rather than overwriting a single restart set every time, alternate between two sets of restart files. If one restart set is ruined, the other one can still be used.

Use ADIOS for I/O

One pioneer user that ran large jobs found that ADIOS was often several times faster than user-built I/O code and almost always faster than HDF5 and NetCDF. Experience indicates that the MPI_AGGREGATE method of ADIOS is usually the best method.

Documentation on ADIOS can be found in the Tips & Tricks section of the ERDC DSRC Website.
Introduction to Parallel I/O, Lustre, & ADIOS and on ORNL's Introduction to the Adaptable IO System (ADIOS).

Optimize Lustre Striping Parameters

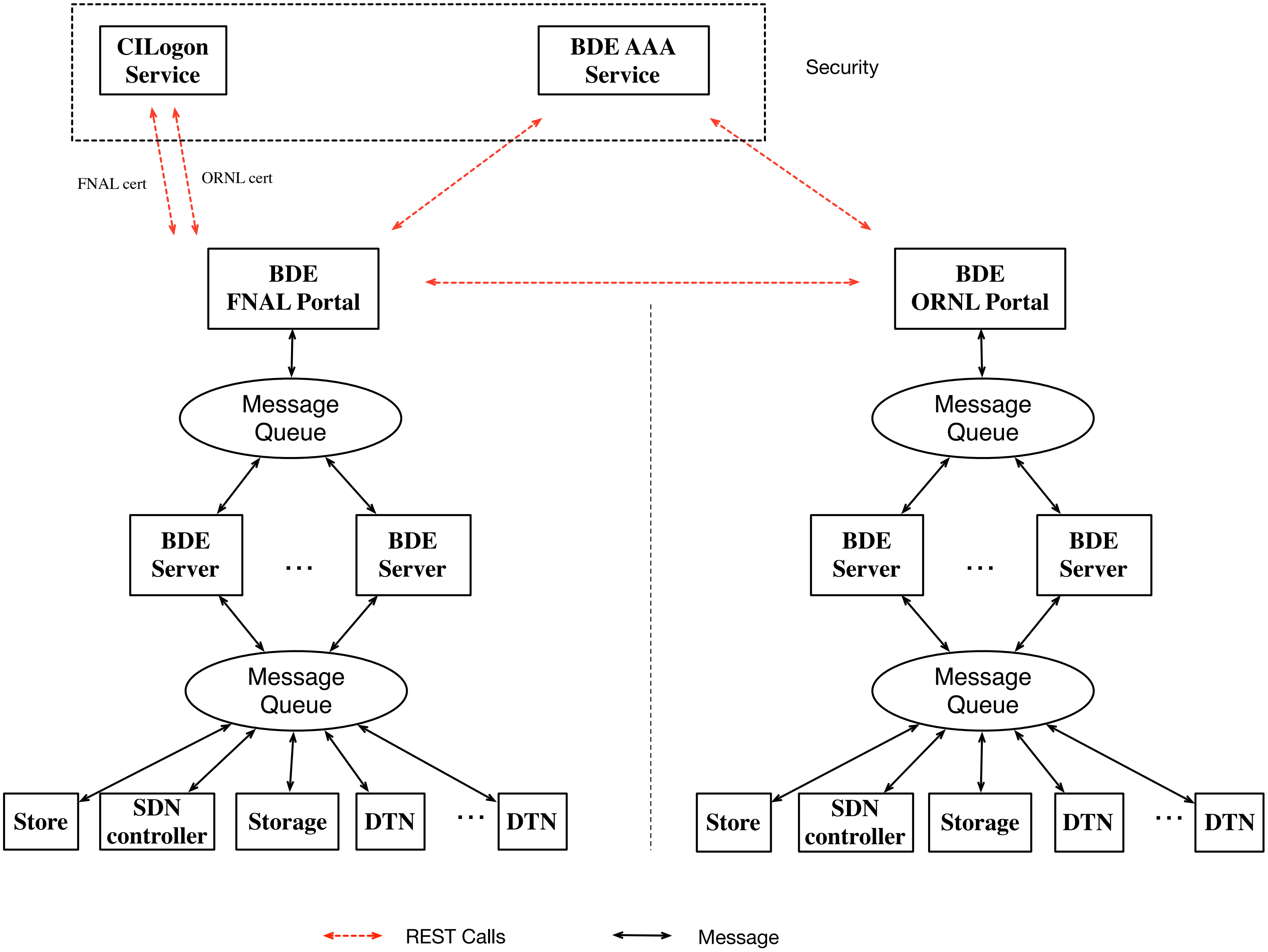
If you must use HDF5, NetCDF, or user-built I/O code, a Lustre striping different from the default may give better performance. The optimum Lustre striping varies. For example, some programs use one file per core, whereas other programs use one big file for the entire job. In the latter case, the file should use more stripes. Including the setup of Lustre striping in the job batch script will ensure that Lustre parameters are set as a routine part of the job submission.

Documentation on Lustre can be found in the Tips & Tricks section of the ERDC DSRC web page:
Increasing Performance on Lustre File Systems
Optimizing IO Performance for Lustre (vendor document)

WARNING!! This Department of Defense interest computer system is subject to monitoring at all times.
Unauthorized access is prohibited by Public Law 99-474 (The Computer Fraud and Abuse Act of 1986).
Users are advised to read and agree to the following Privacy & Security Notice
Date of Last Update: September 16, 2013

2.3 BigData Express Implementation

BDE Architecture



A BigData Express Site

- BigData Express Web Portal
  - Access BigData Express data transfer services
  - Monitor and keep track of data transfer status
  - Monitor DTN status
  - Node.js architecture
  - REST-based web service
- Message Queue (RabbitMQ)
  - Inter-process communication
  - Publish/subscribe
  - Routing
  - RPC

A BigData Express Site (cont.)

- Data Store (Redis)
  - In-memory data store to hold various data
    - DTN resources and status
    - Storage resources and status
    - LAN resource and status

- SDN Controller (ONOS)
  - Open-source network operating system
  - The network agents access the SDN controllers through northbound APIs

A BigData Express Site (cont.)

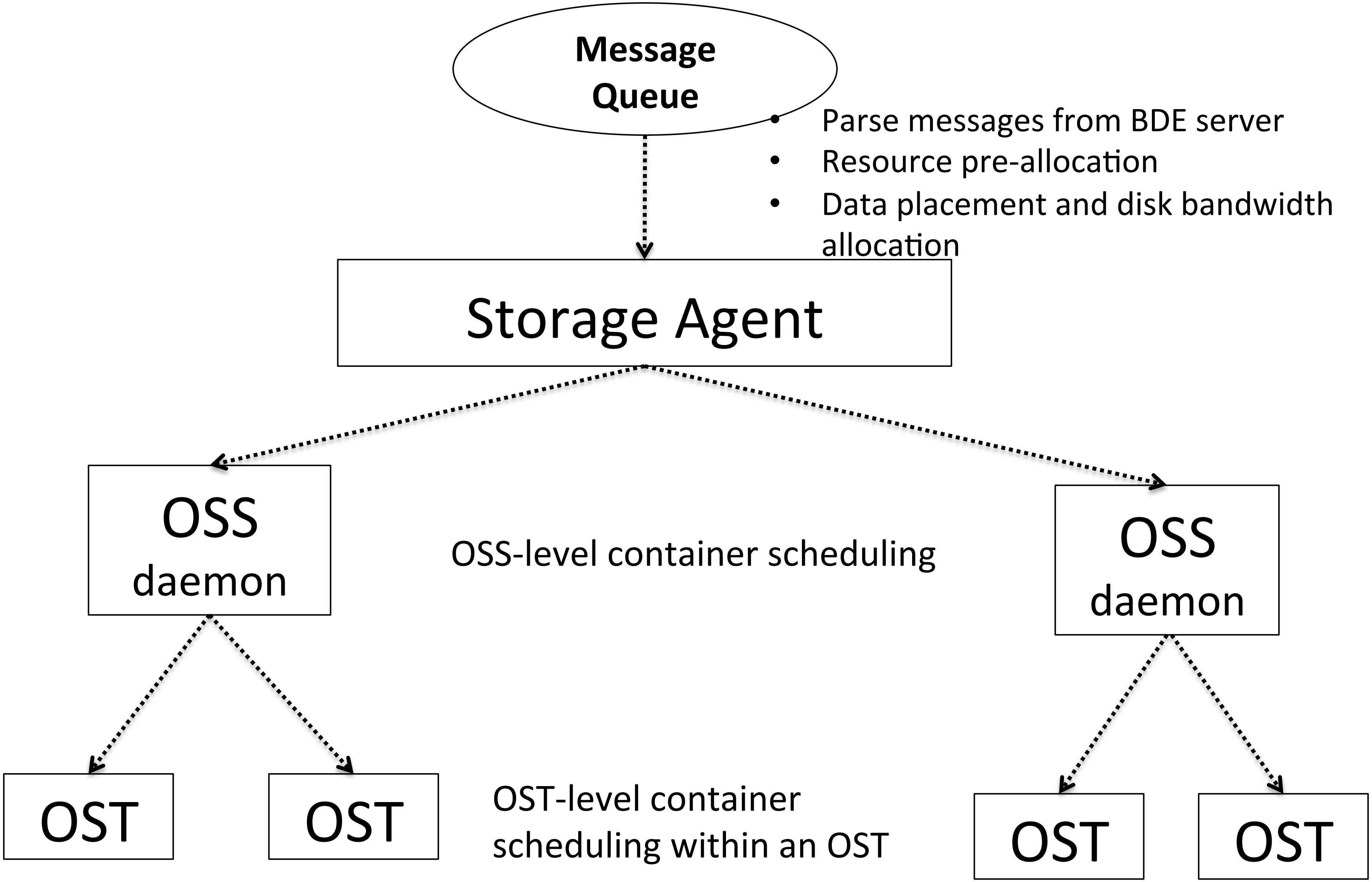
- BigData Express Server
  - Manage local resources
    - DTN
    - Storage
    - LAN
  - Schedule resources for data transfer tasks
    - Resource estimation
    - Resource pre-allocation
    - Resource brokering
    - Resource assignment
    - Admission control
    - Rate control

Security

- BigData Express web service security
  - BDE AAA service
    - Single sign-on
- Local site security
  - Each site has its own security policy.
    - We need to access a site's resources (e.g. DTNs, Storage, LAN, and WAN)
  - CILogon service to obtain certificates for each site
    - Short-lived certificate (max. 1,000,000 seconds)
    - X509

Key BDE Storage Components (ORNL)

- Storage agent
  - Interact with BDE server
  - Storage resource estimation, negotiation and assignment
- OSS daemon
  - Communicate with storage agent and assign weight to each transfer request at each OST



Message Queue

- Parse messages from BDE server
- Resource pre-allocation
- Data placement and disk bandwidth allocation

Storage Agent

OSS daemon

OSS-level container scheduling

OSS daemon

OST

OST

OST-level container scheduling within an OST

OST

OST

Tentative message format between BDE server and storage agent

- Message type (uint8_t)
  - Pre-allocate (along with Pre-allocate ACK message)
  - Reserve (along with Reserve ACK)
  - Release
- Application id
- Various QoS metrics
  - Bandwidth (uint64_t)
  - latency (uint64_t)
  - capacity (uint64_t)

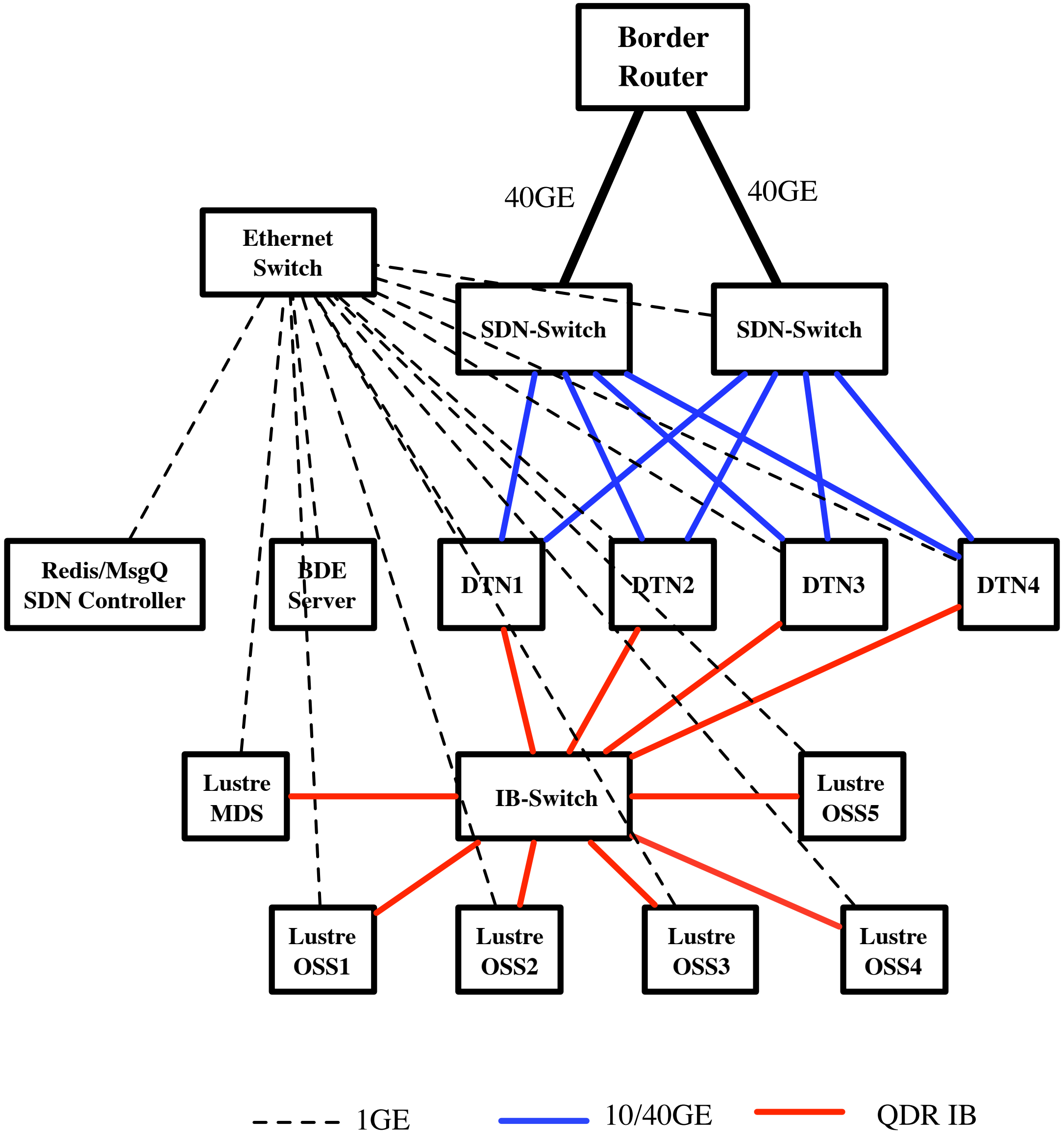
Key data structures

- Resource allocation table at storage agent
    1. <OST id, capacity, latency, bandwidth>
    2. <application id, OST id, weight>

3. BigData Express
Research & Development Plan

3.1 BigData Express Development/Test Environment

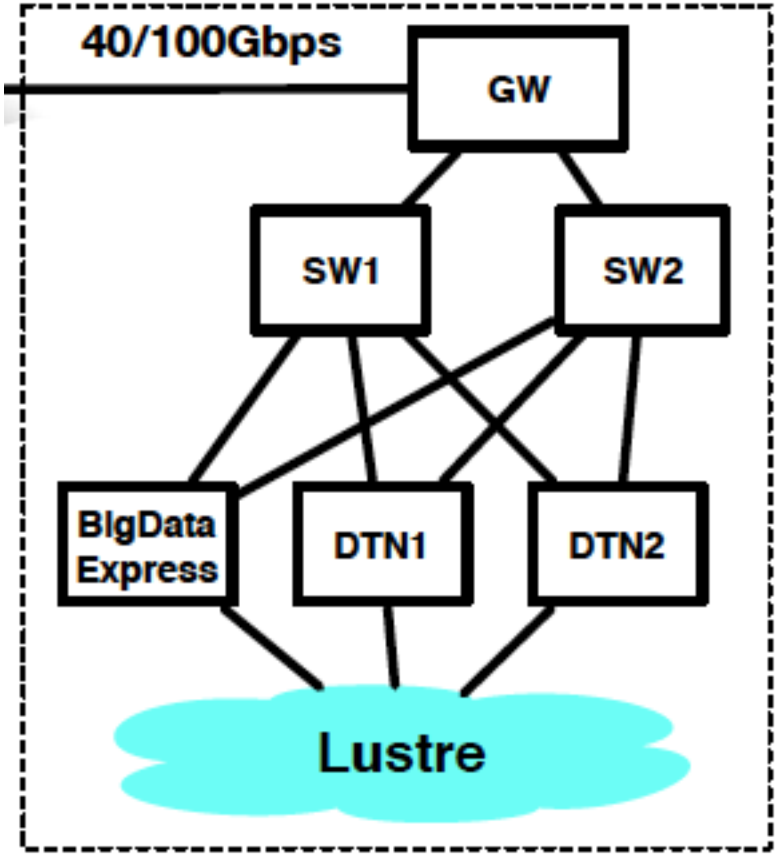
FNAL Development/Test environments



FNAL Development/Test environments

- DTN Cluster
  - 2-4 DTNs
- SDN LAN
  - One SDN controller
  - Two SDN-enabled switches
- Luster cluster
  - 5-8 MDS/MDT + OSS/OST nodes

ORNL Development/Test environments



ORNL BigData Express Development and Testing Environment

- Ongoing discussion with OLCF and ORNL CADES for BDE development and testing

ESNET Test Path between FNAL & ORNL

- End-to-end
- On-demand
- Negotiable
    - Can Bandwidth be brokered?
- Security model
- Need to collaborate with ESnet

BigData Express Project Repository

- We need a software project management toolset
  - Distributed version control
  - Source code management

- Fermilab Redmine
  - https://cdcvs.fnal.gov/redmine

3.2 BigData Express Development Roadmap

FNAL

a. Build up development/test environment

Development/test environment	Yr-1	Yr-2	Yr-3
• DTN Cluster	✔		
• Luster Cluster	✔	✔	
• SDN-based LAN	✔		
• SDN-based WAN	✔	✔	

FNAL

b. Web portal R&D

Web subsystem R&D	Yr-1	Yr-2	Yr-3
• BDE AAA service R&D	✔		
• Interface with CILogon to obtain short-lived certificates	✔		
• File Browsing capability	✔		
• Collect and submit data transfer requests	✔		
• Monitor data transfer status	✔	✔	
• Monitor DTN status	✔	✔	
• Monitor SDN LAN status (optional)			✔
• Monitor Luster file system status (optional)			✔
• Integration and test	✔	✔	✔

FNAL

c. BigData Express site version R&D

Site version R&D	Yr-1	Yr-2	Yr-3
• DTN Agent development	✓		
• SDN Agent development	✓		
• Investigate on how to integrate various Agents & Web portal (Message Queue, Message Format)	✓		
• Resource estimation R&D • DTNs, Storage, and LAN	✓		
• Admission control R&D	✓		
• Resource pre-allocation R&D	✓		
• Resource brokering R&D	✓	✓	✓
• Resource assignment R&D	✓	✓	✓
• Distributed DTN matching R&D	✓	✓	✓
• DTN rate control	✓		
• Integration and test	✓	✓	✓
• Deploy & test with scientific use cases			✓

FNAL

d. BigData Express single-node version R&D

Single-node version R&D	Yr-1	Yr-2	Yr-3
• Resource estimation • DTN	✓		
• Interface with Web portal	✓		
• Admission control R&D	✓		
• Resource pre-allocation R&D	✓	✓	
• Resource brokering R&D	✓	✓	
• Resource assignment R&D	✓	✓	
• Distributed DTN matching R&D	✓	✓	
• DTN rate control R&D	✓		
• Interface with Network	✓	✓	✓
• Integration & Test	✓	✓	✓
• Deploy & test with scientific use cases			✓

ORNL

e. Storage System R&D

Single-node version R&D	Yr-1	Yr-2	Yr-3
• Storage resource estimation using learning techniques	✔	✔	
• Storage resource pre-allocation R&D	✔	✔	
• Storage resource assignment using container R&D	✔	✔	
• Interface with BDE Server	✔	✔	
• Complex data pipeline R&D		✔	✔
• Integration & Test	✔	✔	✔

ORNL

f. Security-related feature development

Single-node & cluster version R&D	Yr-1	Yr-2	Yr-3
• Investigate how grid certificate works	✔		
• Investigate how globus online interacts with file systems	✔		
• Develop similar security features for BigData Express		✔	
• Integration & Test	✔	✔	✔

DEMO and Deployment

- Single node version data transfer demo between FANL and ORNL (SC’16)
- Full-site version data transfer demo between FNAL and ORNL (SC’17)
- Deployment of BigData Express at FNAL and ORNL (Yr-3)
- Deployment of BigData Express at DOE large computing centers