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

# BigData Express

Wenji Wu

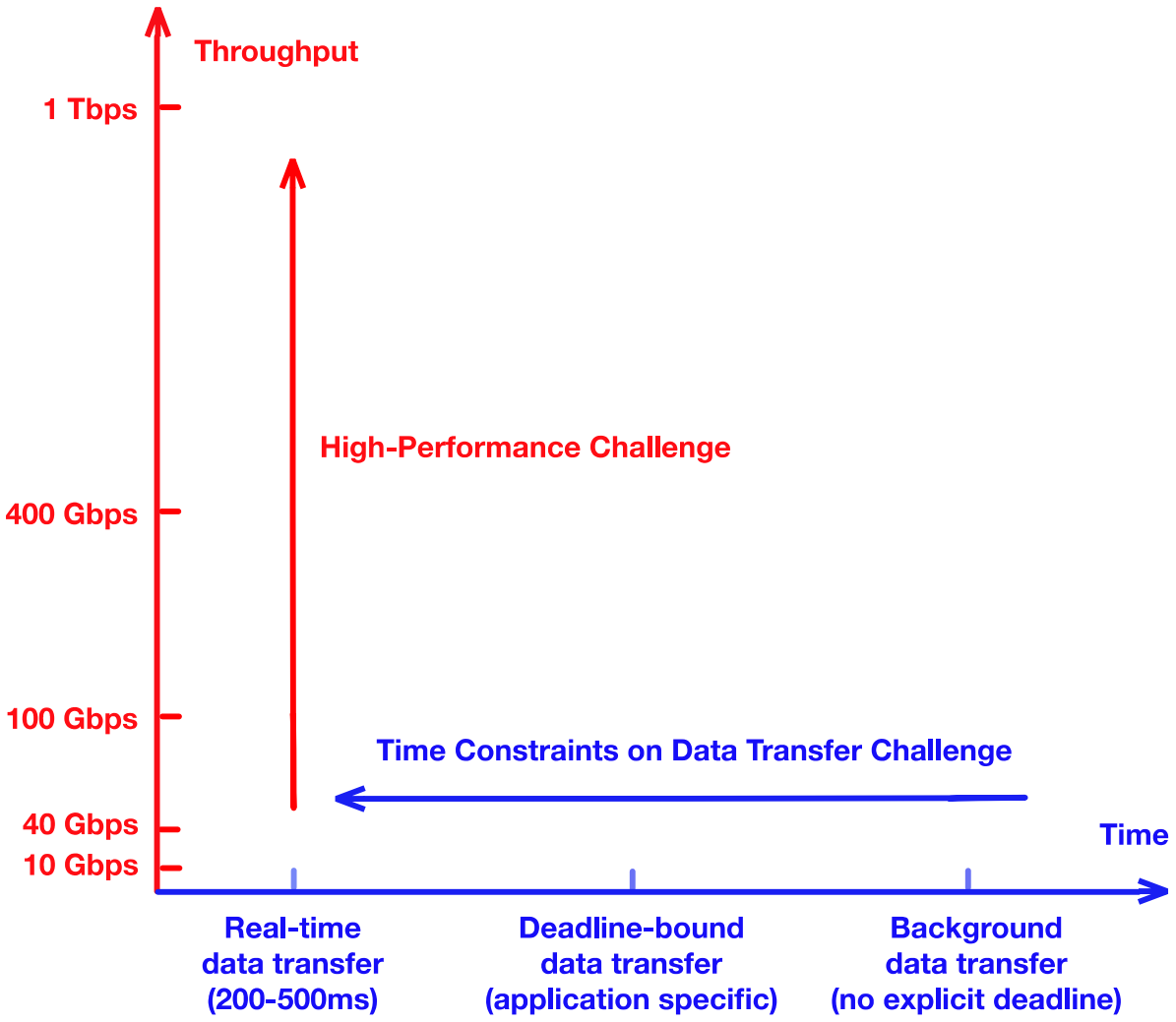
Workshop on Science of Security through Software-Defined Networking

16 June 2016

# Content

- DOE Data Transfer Challenges
  - High-performance challenges
  - Time-constraint challenges
- Problems with existing data transfer tools and services
- The BigData Express Project
  - Architecture and Design
  - How does BigData Express work?
  - The use of SDN and SDS in BigData Express
- Conclusion

# DOE Data Transfer Challenges






Fermilab

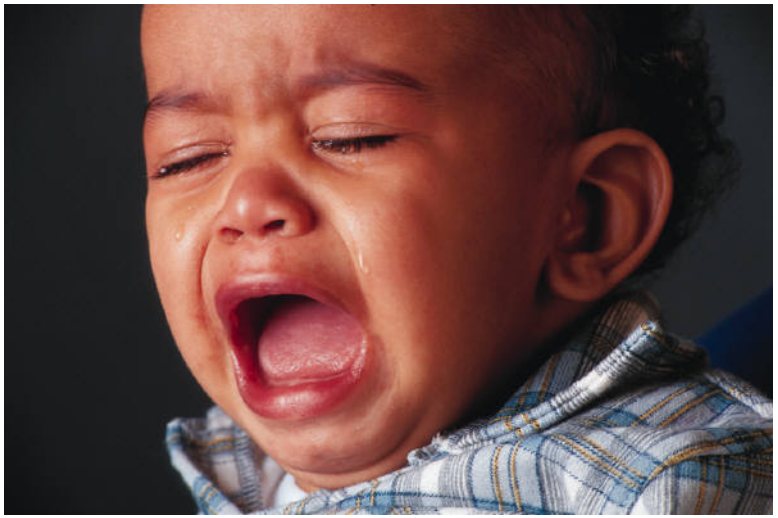
# Data Transfer – State of the Art

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

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

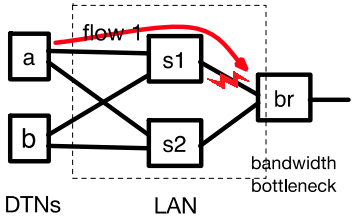
# Can Today's data transfer tools and services support extreme-scale science applications well?

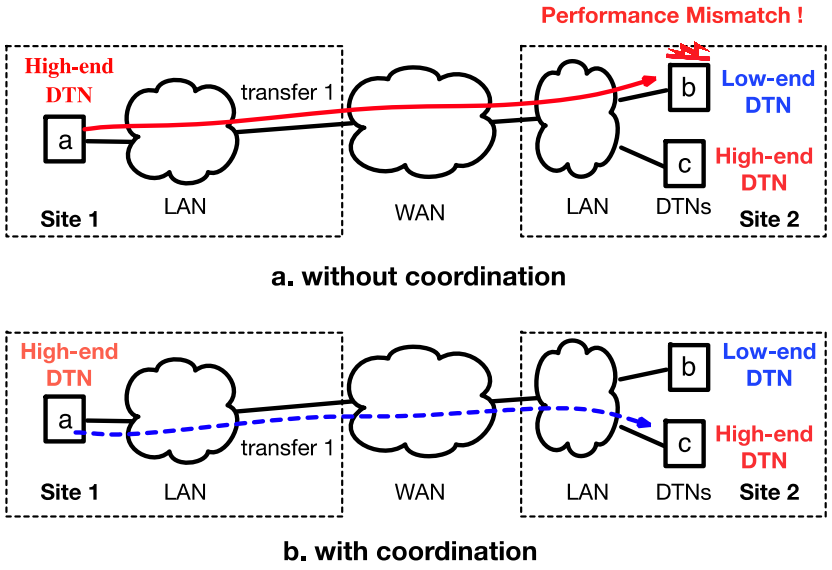
No!



# Problems with existing data transfer tools and services – Problem 1

- Disjoint end-to-end data transfer loop




Network Congestion

DTN Performance Mismatch



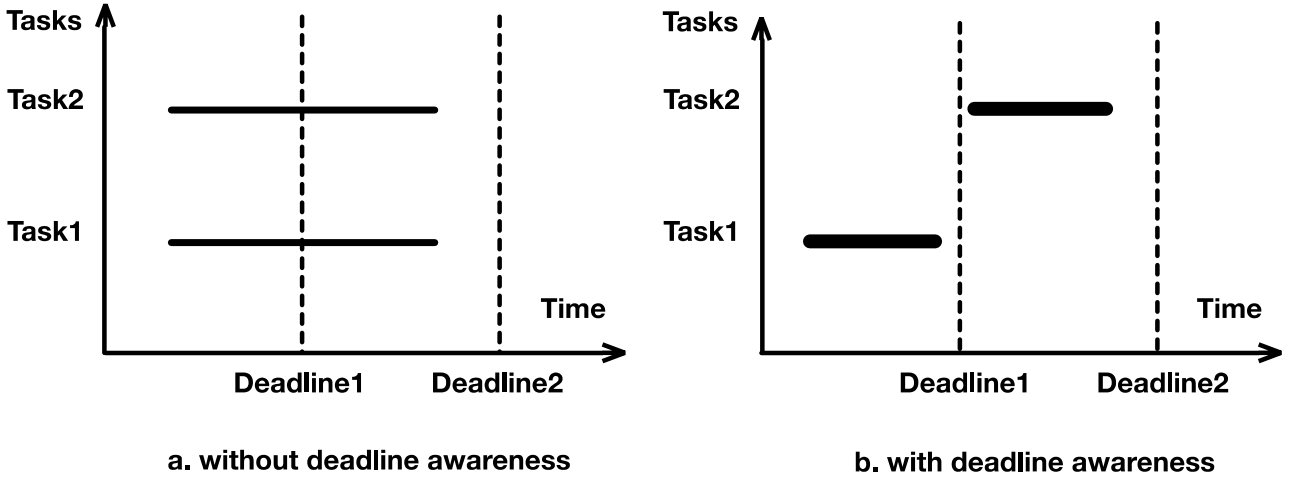
Fermilab

# 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

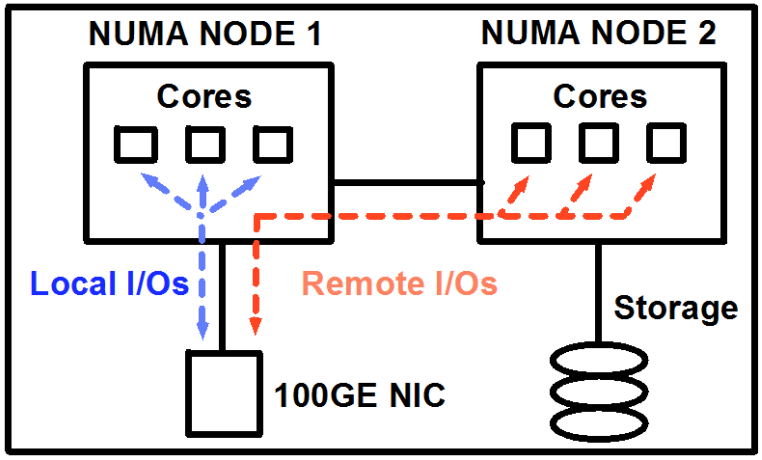
- Oblivious to user requirements (e.g., deadlines and Qos requirements)




**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 (data transfer nodes)



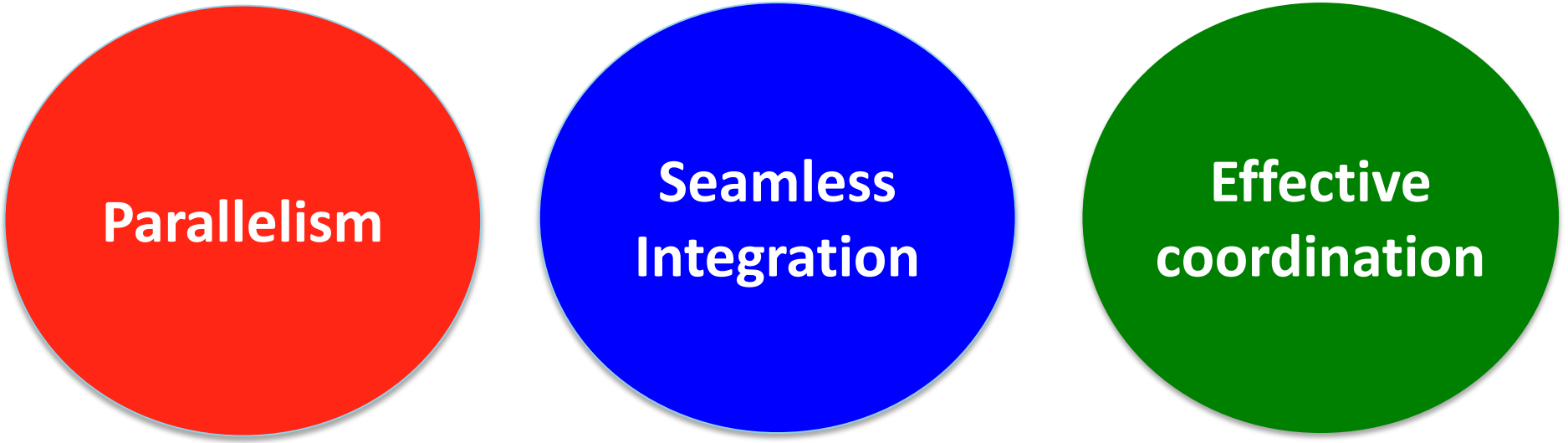

**The parallelism vs. I/O locality problem on NUMA systems**

# 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)
  - http://bigdataexpress.fnal.gov
  - Capitalize on the MDTM project
    - http://mdtm.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)

# BigData Express Design Principles



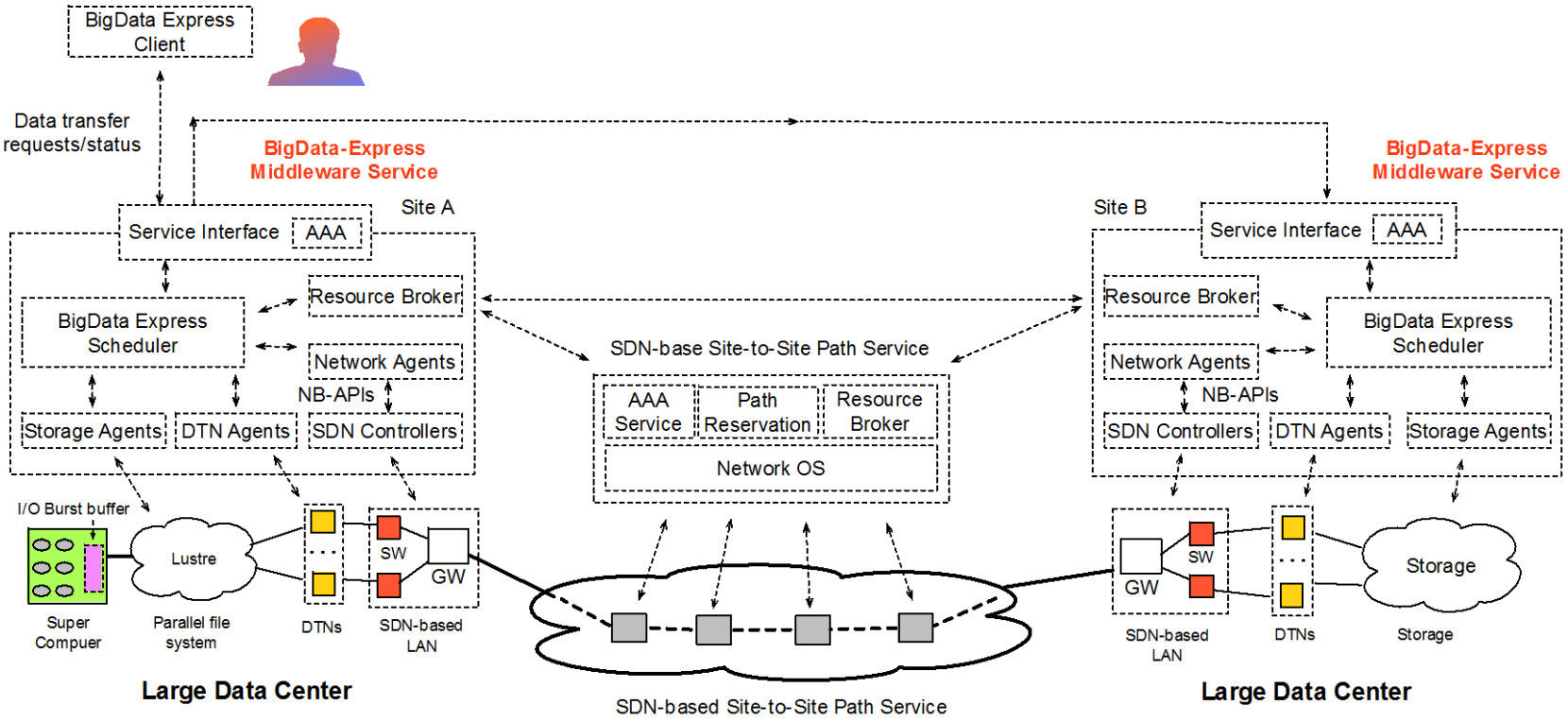
# BigData Express - Key Features (1)

- 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

# BigData Express - Key Features (2)

- 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 - Architecture




A large data center typically features

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



Fermilab

# BigData Express - Major Entities (1)

- 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.

- The service interface
  - Authenticate, authorize, and audit users and user applications
  - Allow user to access BigData Express services

# BigData Express - Major Entities (2)

- 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

# BigData  Express - Major Entities (3)

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

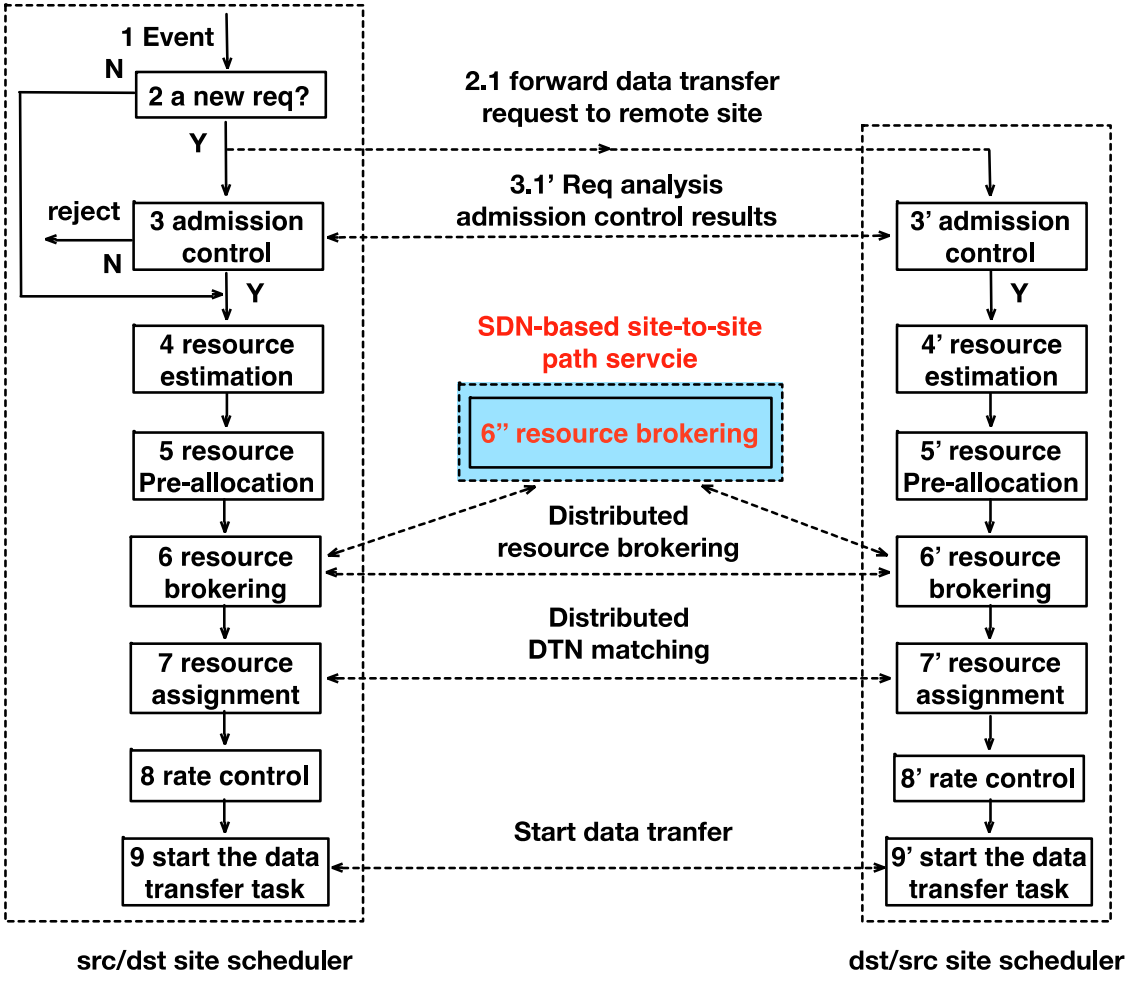
# How does BigData Express work? (1)

- 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

- 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? (2)





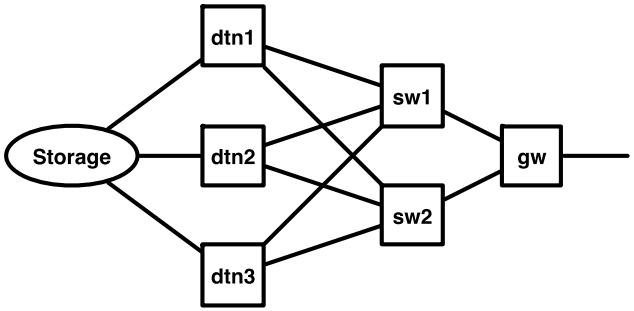

Fermilab

# The use of SDN in BigData Express (1)

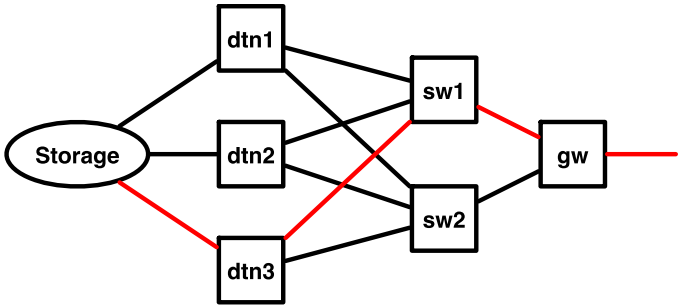
- To transform networks into schedulable resources to enable a data-transfer centric architecture
- To improve network I/O performance
  - Reduce/eliminate network congestion
- To improve DTN performance
  - Eliminate remote network I/Os in DTN

# The use of SDN in BigData Express (2)

- Deploy ONOS controller to control and manage networks
- Use REST APIs to manage SDN-enabled networks
  - Obtain network information
    - Topology, Devices, Links, Hosts
  - Install/delete open flow rules in switches to setup/tear down network paths




Obtain network topology



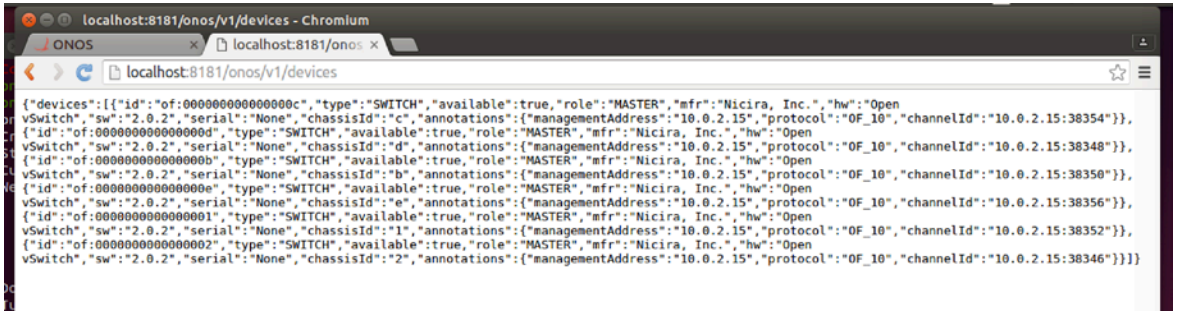

Set up network path



Fermilab

# The use of SDN in BigData Express (3)

- Obtain SDN network links with REST APIs




- Install OpenFlow Rules with REST APIs
  - curl —user karaf:karaf -d @post-intent.json -H “Content-Type: application/json” -X POST http://localhost:8181/onos/v1/intents

- Delete OpenFlow Rules with REST APIs
  - curl —user karaf:karaf -X DELETE http://localhost:8181/onos/v1/intents/org.onosproject.gui/0x1a

# The use of SDS in BigData Express

- Aim to provide guaranteed, high-performance storage I/O
- The idea is to manage block I/Os via lightweight Linux-container-based virtualization
- Two vehicles for allocating block I/Os in a Linux container
  - Throttling functionality
    - Set an upper limit to a process group's block I/O
  - Weight function
    - Assign shares of block I/O to a group of processes

# BigData Express 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

# Conclusion

BigData Express is a middleware data transfer service that provides a **schedulable**, **predictable**, and **high-performance** data transfer service for DOE's large-scale science computing facilities (e.g., LCF, US-LHC computing facilities)

# Questions?