The BigData Express Project (BDE)

BDE Research Team
November 2017
BDE Research Team

• FNAL
  – Wenji Wu (PI)
  – Qiming Lu
  – Liang Zhang
  – Amy Jin
  – Sajith Sasidharan
  – Phil DeMar

• ORNL
  – Nageswara Rao
  – Gary Liu
  – Ryan Prout

• KISTI
  – Syed Asif Shah
  – Seo-Young Noh
  – Jin Kim

Note: KISTI and ESnet are unfunded project partners
BigData Express

- Collaborative effort by Fermilab and Oakridge National Laboratory
- Funded by DOE’s office of Advanced Scientific Computing Research (ASCR)
- A three-year research project
  - Start: Oct 1, 2015
  - End: Sep 30, 2018
- [http://bigdataexpress.fnal.gov](http://bigdataexpress.fnal.gov)
- Code repository
  - [https://cdcvs.fnal.gov/redmine/projects/bigdata-express/repository/](https://cdcvs.fnal.gov/redmine/projects/bigdata-express/repository/)
BigData Express Goal

- A distributed middleware system that will provide a schedulable, predictable, and high-performance data transfer service for the DOE’s large-scale science facilities and their collaborators.
Research problems to solve

• Problem 1: Disjoint end-to-end data transfer loops

• Problem 2: Cross-interference between data transfers

• Problem 3: Existing data transfer tools and services are oblivious to user requirements

• Problem 4: Inefficiencies arise when existing data transfer tools are run on DTNs
BigData Express – Toward Schedulable, Predictable, and High-performance Data Transfer

Data transfer requests/status

BigData Express Client

BigData-Express Middleware Service

SDN-based Site-to-Site Path Service

Resource Broker

SDN Controllers

Storage Agents

DTN Agents

Network Agents

PB-APIs

Site A

Service Interface

AAA

BigData Express Scheduler

Storage Agents

DTN Agents

SDN Controllers

SDN-based Site-to-Site Path Service

SDN-based Site-to-Site Path Service

Site B

Service Interface

AAA

Resource Broker

SDN Controllers

Storage Agents

DTN Agents

Network Agents

PB-APIs

Lustre

Parallel file system

I/O Burst buffer

Super Computer

Large Data Center

Large Data Center
BigData Express Components

• **BDE Web Portal**
  – A web portal that allows users to access BigData Express data transfer services
  – A distributed web application that allows BDE at different sites to coordinate and broker resources

• **BDE Scheduler**
  – DTN as a service
  – Co-scheduling of DTN, storage, and network

• **BDE AmoebaNet**
  – Network as a service
  – Allow applications to program networks at run-time for optimum performance

• **mdtmFTP**
  – a high-performance data transfer engine
Supports three types of data transfer

• Real-time data transfer
  – BDE WebPortal + Scheduler + AmoebaNet + mdtmFTP

• Deadline-bound data transfer
  – BDE WebPortal + Scheduler + AmoebaNet + mdtmFTP

• Best-effort data transfer
  – BDE WebPortal + Scheduler + mdtmFTP
BDE Web Portal

• Providing a visual interface to allow users to submit data transfer requests and obtain data transfer status
• Launching data transfer tasks on users/applications’ behalves
• Providing HTTPS communication channels among BDE sites to support BDE mechanisms
• Allowing system admins to manage and monitor BDE sites
• A common single-point sign-on service - CILogon service – to obtain X.509 certificates for secure access to BDE sites
• [http://bigdataexpress.fnal.gov/Releases.html](http://bigdataexpress.fnal.gov/Releases.html)
BDE Scheduler – Resource Management

BDE scheduler manages site resources through various agents
  • DTN agents
  • SDN agents
  • Storage agents

Use RabbitMQ as message bus to enable communication between BDE scheduler and agents
BDE Scheduler – Resource Orchestration

- Orchestrate resources at local sites
- DTNs
- Storage
- Local network resources

- Negotiate resources with remote sites
AmoebaNet

• A SDN-enabled network service that provide “Application-aware” network
  – Allow application to program network at run-time for optimum performance
  – Network as a service
• Originally designed for BigData Express
• Can be independently deployed to support any applications
• Available at [http://bigdataexpress.fnal.gov/Releases.html](http://bigdataexpress.fnal.gov/Releases.html)
• Docker release [https://hub.docker.com/r/wenji/amoebanet/](https://hub.docker.com/r/wenji/amoebanet/)
AmoebaNet

• Providing AMQP interface to allow application to program network at run-time for optimum performance
  – JSON-RPC style communication
  – Dynamic and flexible interaction with network
• Qos-based routing and path calculation
  – Bandwidth
  – Delay (coming soon)
• Fast provisioning end-to-end network path with guaranteed Qos
  – VLAN-based layer-2 path
  – TCP/IP-based layer-3 path
• REST-based network initialization and configuration
AmoebaNet

- AmoebaNet is implemented using Java upon ONOS
- 3K+ lines of code
- [http://bigdataexpress.fnal.gov/Releases.html](http://bigdataexpress.fnal.gov/Releases.html)
- Docker release
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
• 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
• BigData Express: a schedulable, predictable, and high-performance data transfer service
  – QoS-guaranteed data transfer
  – DTN as a service
  – Network as a service
  – Distributed resource brokering/matching

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