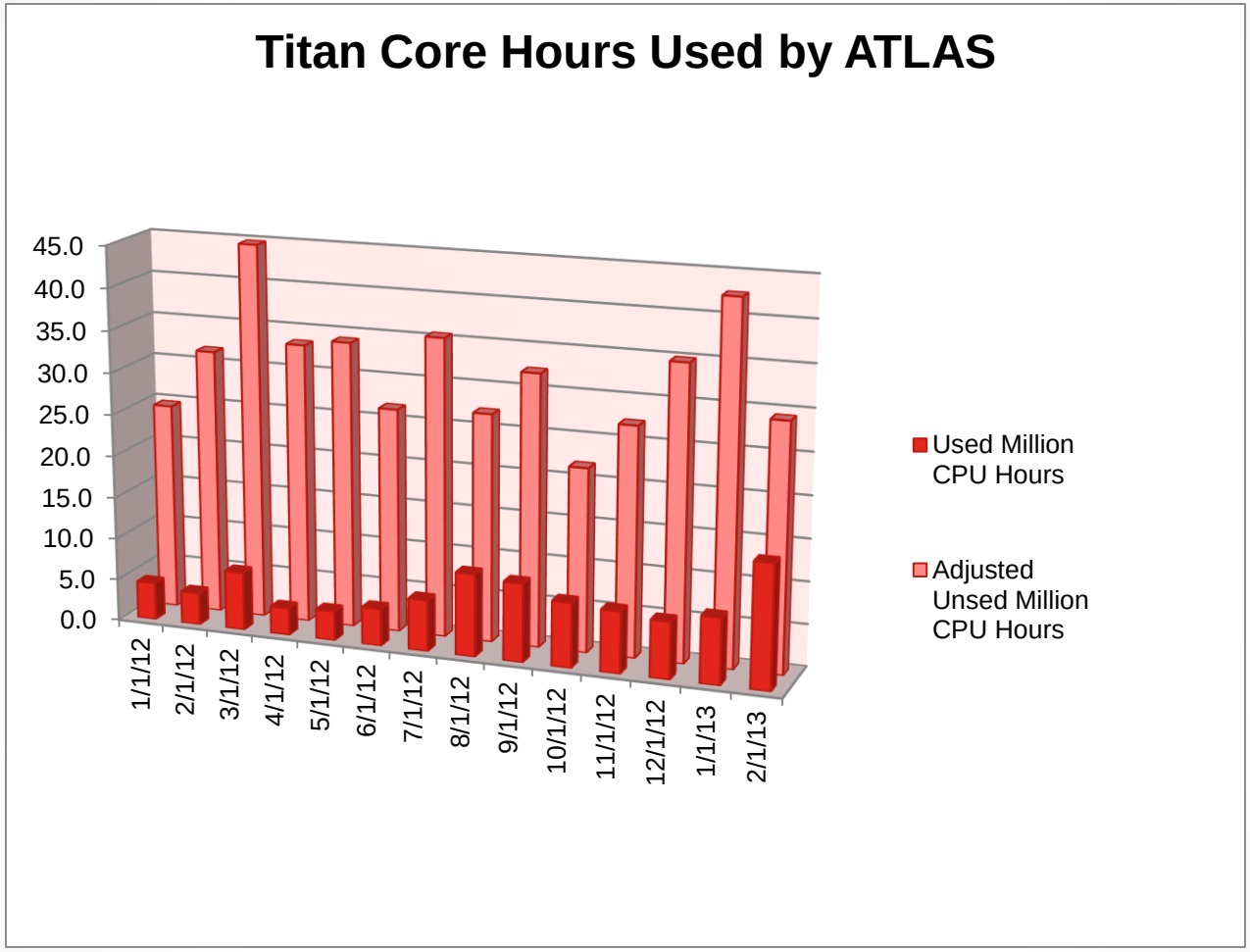
HPC Integration @ DOE sites, Harvester Deployment & Operation

Doug Benjamin
Duke University

Work Flows

- What sort of work flows are currently running at each HPC (NERSC (Cori and Edison), OLCF (Titan), ALCF (Mira, Cooley, Theta))?
  - NERSC - event service simulation and traditional simulation jobs multiple nodes local pilot submission.
  - ALCF (Mira) - Alpgen and Sherpa Event generation
  - OLCF executes regular multicore simulation jobs. In general: set of regular jobs (aligned by number of events) launches in one MPI submission. Size of MPI submission adjusted by number of available nodes (backfill). Specially modified PanDA pilot used to manage of submissions.
- What sort of work flows are you planning on running at your site in the next 6 months, over the next year?
  - 6 Months –
  - NERSC - traditional simulation jobs, traditional simulation jobs running in shifter containers, Harvester event service simulation and Harvester event service merging jobs
  - ALCF (Cooley, Theta) - Harvester event service simulation , Harvester event service merging jobs6
  - 12 Months
  - NERSC - Harvester event service including merging jobs in shifter containers
  - ALCF (Theta) - Harvester event gen.
  - OLCF Atlas Event Service through Harvester and Yoda.

Titan – G4 MC Simulations



Total 74 Million Titan Core Hours used in calendar year 2016

Work Flows (2)

- What is the current status of Event Service jobs running at your site? - What will it take to run the merge jobs at each HPC instead of shipping the files away?
  - NERSC (Edison) - worked through the unit tests. Need to retest using robotic credential. Need secure automatic mechanism for keeping grid proxy with voms extension up to date. Harvest is being tested on login node. Should finish first round of testing within next week or so. except for Rucio-Globus portion
  - ALCF (Cooley) - start deployment and testing in two weeks or less.
  - OLCF Not run yet. Yoda validation on Titan will be performed in nearest future.
  - OLCF provides dedicated facility for IO intensive operations. This facility can be used to perform merging jobs without affecting of Titan computing nodes. Question how CPU intensive is merging? Might it be too CPU intensive for this facility?

Software Installation - Containerization

- How is the ATLAS software installed? How labor intensive is it? How much do you rely on the existing software installed at the site?(ie environmental modules)
  - NERSC - Vakho installs various software releases by hand. We need to rationalize where the software releases and other software needed isinstalled. Taylor is working on a virtual environment for NERSC based on using the existing modules and existing ATLAS software.
  - ALCF - will follow the NERSC plan.
  - OLCF Manually, through packman. ATLAS software should be installed to read-only file system (NFS). Not to much work, but deployment of one release may take few hours.
  - OLCF Proper version of python and some additional python libraries managed through modules. During nearest F2F BigPanDA meeting i am going to discuss possibility to manage LCG middleware (GFAL, VOMS etc.) and Rucio client tools through modules.

Software Installation - Containerization

- Are there plans to use virtualization/containerization at your site? If so what are they?
  - NERSC - has developed shifter - US ATLAS needs to start routinely producing shifter containers for ATLAS use at NERSC
  - ALCF – nothing officially decided on containerization
  - Will be specially discussed during next F2F BigPanDA meeting at the end of march.

Edge Services

- What sort of edge services currently exist at your site?
  - Which ones are provided by the site
  - Which ones are provided by US ATLAS
  - NERSC - Data Transfer Nodes (DTN) - gridftp servers, at NERSC ATLAS runs a cron to turn a gridftp server into a RSE
  - NERSC has setup a gram base Grid compute element
  - Note - all grid credentials used must be registered to a local NERSC account at NERSC
  - ALCF Data Transfer Nodes (DTN) - gridftp servers that only accept ANL CA certificates (with one time password authorization). ATLAS can only use Globus Online for managed transfers to ALCF
  - OLCF –Interactive and batch DTN nodes support several data transfer tools including Globus (with one time password authorization).
- What the plans for new edge service provided by your site? (for example ALCF is going to provide a CondorCE that only accepts ALCF credentials)
  - NERSC should be setting up a HTCondorCE eventually.
  - ALCF is in process of setting up HTCondorCE that will only accept ANL grid credentials

Edge Services

- What is the status of Harvester deployment at your site?
  - Will it be installed inside the HPC firewall or outside?
  - What is the current schedule for deployment and testing?
  - NERSC - ATLAS is starting to deploy Harvester on the login nodes (Edison initially and Cori next)
  - NERSC – have run through all of the unit tests
  - NERSC – currently working through getting Yoda running with a ATLAS Event Service simulation job. Yoda internals changed at NERSC when going from cron based submissions to Harvester submissions with Jumbo jobs
  - Using Athena setup from traditional ATLAS Simulation jobs
  - ALCF ATLAS will run Harvester on edge nodes initially Start deployment this week
  - OLCF - Harvester core components deployed on new DTN nodes.
  - OLCF – unite tests complete. Working through functional tests with simple payloads to check the chain works. Should be finished this week.
  - OLCF – Integration with Pilot2 common components
  - 1.5 weeks – Movers – Rucio Clients, Pilot2 libraries, integration with Harvester through API. Unit tests of data transfers.
  - Note – OLCF is developing its own data motion infrastucture
  - 3 weeks – Proper ATHENA setup , specific handling of setup at OLCF – ie multiple copies of DB Releases to avoid high IO per file, Tests with ATLAS payloads

Wide Area Data Handling

- How do ATLAS data files currently get transferred to and from your site?
  - NERSC - Rucio Storage element using only one gridftp server.
  - ALCF - client tools and globus URL copy to ANL HEP ATLAS group RSE.
  - OLCF Synchronously with jobs, by GFAL mover from/to BNL SE.
- What are your plans for using all the of existing DTN's at your site?
  - NERSC and ALCF - Transfer data from dual use Globus Endpoint - RSE at US ATLAS Tier 1/Tier 2 site and Globus endpoint at HPC
  - With Shared Globus Endpoint at US ATLAS Tier 1/Tier 2 site can have md5 checksums and increased limits on transfers.
  - Need the Globus-Rucio code written and deployed
  - OLCF - For the moment we use 4 interactive DTN. Current state cover our needs for the moment without affecting other users of OLCF. Also OLCF provides batch DTN which can be used if needed

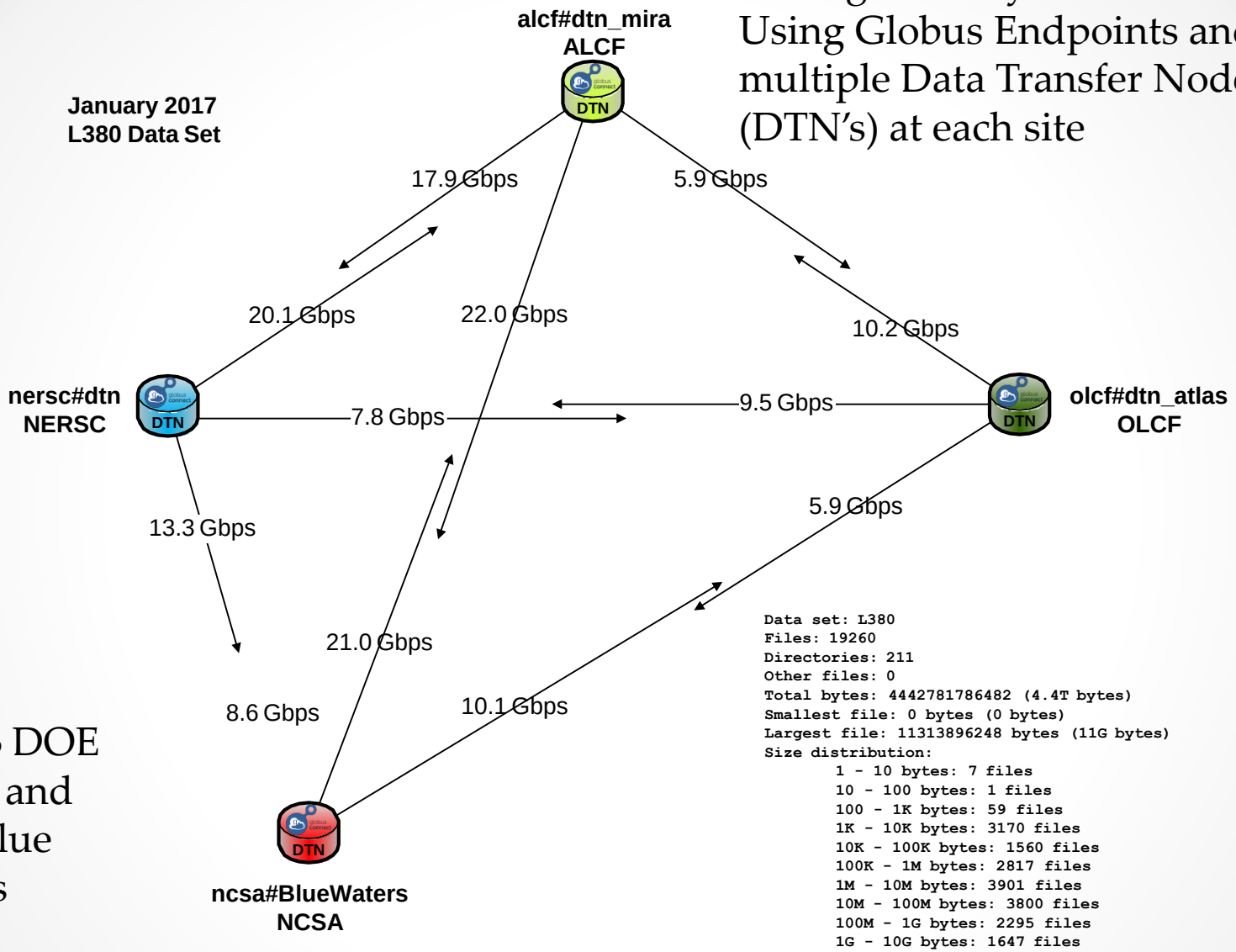
Wide Area Data Handling

- What can be done to aid in the data flows to and from your site?
  - NERSC-ALCF – Dual use Globus/Rucio Storage Element servers and Get Globus - Rucio linkage working.
  - OLCF Data transfers performance looks very good in OLCF. Harvester will allow to decouple payload execution and data transfers, so loading to DTN will be managed, using of Pilot2.0 movers will allow to use different types of data transfer tools.
  - Note – It appears to me that OLCF has no plans for managed 3rd party transfers using Globus and instead will put that functionality into Harvester. Is this really a good idea?

- Why use Globus Endpoints and multiple Data Transfer Nodes (DTN's)?

CCE - data transfer project

Testing Done by Eli Dart – ESNET
Using Globus Endpoints and multiple Data Transfer Nodes (DTN's) at each site



All US DOE HPC's and NSF Blue Waters

```
Data set: L380
Files: 19260
Directories: 211
Other files: 0
Total bytes: 4442781786482 (4.4T bytes)
Smallest file: 0 bytes (0 bytes)
Largest file: 11313896248 bytes (11G bytes)
Size distribution:
        1 - 10 bytes: 7 files
        10 - 100 bytes: 1 files
        100 - 1K bytes: 59 files
        1K - 10K bytes: 3170 files
        10K - 100K bytes: 1560 files
        100K - 1M bytes: 2817 files
        1M - 10M bytes: 3901 files
        10M - 100M bytes: 3800 files
        100M - 1G bytes: 2295 files
        1G - 10G bytes: 1647 files
        10G - 100G bytes: 3 files
```

Future Plans

- Within the next 18 months or so, both OLCF and ALCF will be bringing on new machines. What does ATLAS need to do effectively use the new machines?
  - Ensure ATLAS can run efficiently on KNL machines both at NERSC and ALCF. Work with sites to develop mechanism for getting frontier DB data to sites so we can run more work flows than just simulation
  - OLCF - Support of IBM Power 9 processor and GPU's for Summit or ATLAS' use of OLCF will drop precipitously!
- What can be done to scale up computing done at the HPC sites? How much extra labor will it take?
  - Need a team of US people who have accounts at all DOE HPC sites and can share expertise. Need to breakdown the silos that exist.

Discussion Questions

- How we effectively make use of the Data Handling tools that DOE HPC's have? Is ATLAS really committed to making working with the Globus team?
- Can we come up with a common solution and plan for software installation and standard arraignment at the HPC centers? Much like CVMFS forced standardization – CVMFS is not at DOE HPC centers.
- How would we reduce the labor required to integrate and maintain ATLAS production at the DOE HPC Centers? What about NSF sites?
- How do we prevent having N+1 solutions for the N US HPC centers?
- How we become more of a stakeholder with CCE activities? CCE is the conduit between ASCR and DOE OHEP computing.