Integration of Titan with PanDA

Sergey Panitkin (BNL) and Danila Oleynik (UTA)

Outline

- I will summarize developments since ATLAS S&C Meeting in June
- Pilot
- Pilot stress tests
- Situation with workloads

#2 TOP500 SUPERCOMPUTER SITES

27 PFlops (Peak)

18,688 compute nodes with GPUs

299,008 CPU cores

AMD Opteron 6200 @2.2 GHz (16 cores)

32 GB RAM per node

Nvidia TESLA K20x GPU per node

32 PB disk storage (Luster)

29 PB HPSS tape archive

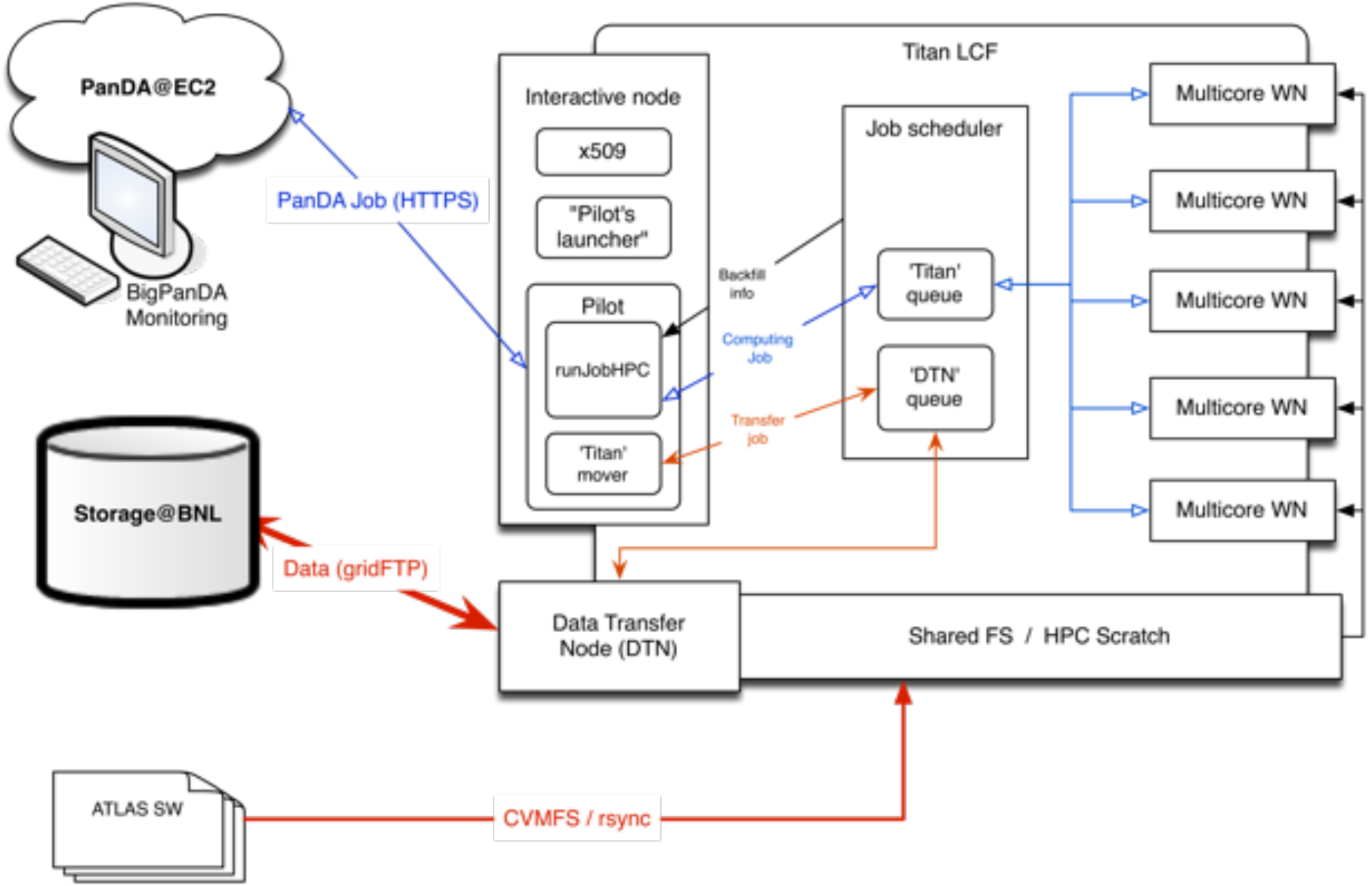
Interfacing PanDA with Titan. Main points

- BigPanDA project on Titan (DD allocation) under ASCR auspices
- 10M hours allocation for 2014-15 on Titan
  - Access to EOS – new Cray XC30 machine at OLCF
  - Also we have access to NERSC via OSG and ATLAS allocations
- Collaboration between ATLAS, ALICE, nEDM experiments
- Project members from BNL, UTA, ORNL/UTK, MSU
  - Strong interest from OLCF, took responsibility for MPI wrapper base and docs
- Technology developed on Titan should be applicable for other HPC centers
  - Already interest from Archer (Edinburgh, UK), IT4Innovation (Ostrava, CZ), Kurchatov Institute (Moscow, RF)
  - Work with ATLAS ANL group on interface to ALCF setup

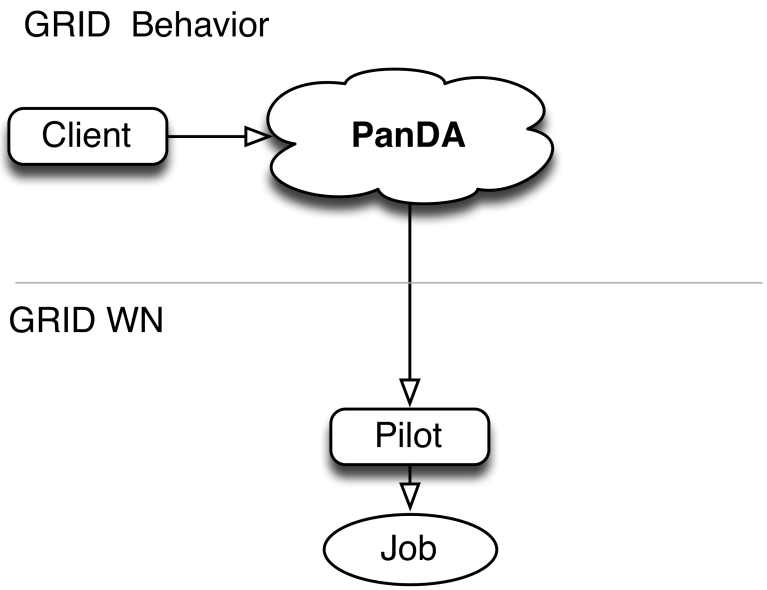
Interfacing PanDA with Titan. Main features

- ATLAS modular pilot augmented with HPC specific classes
  - More details in Danila's and Paul's talk later today
- SAGA (Simple API for Grid Applications) framework as pilot's interface to HPC batch schedulers
  - http://saga-project.github.io/saga-python/
  - http://www.ogf.org/documents/GFD.90.pdf
- MPI wrapper/overlay scripts that allow to run multiple single threaded workload instances in parallel
- "Backfill" functionality in pilot

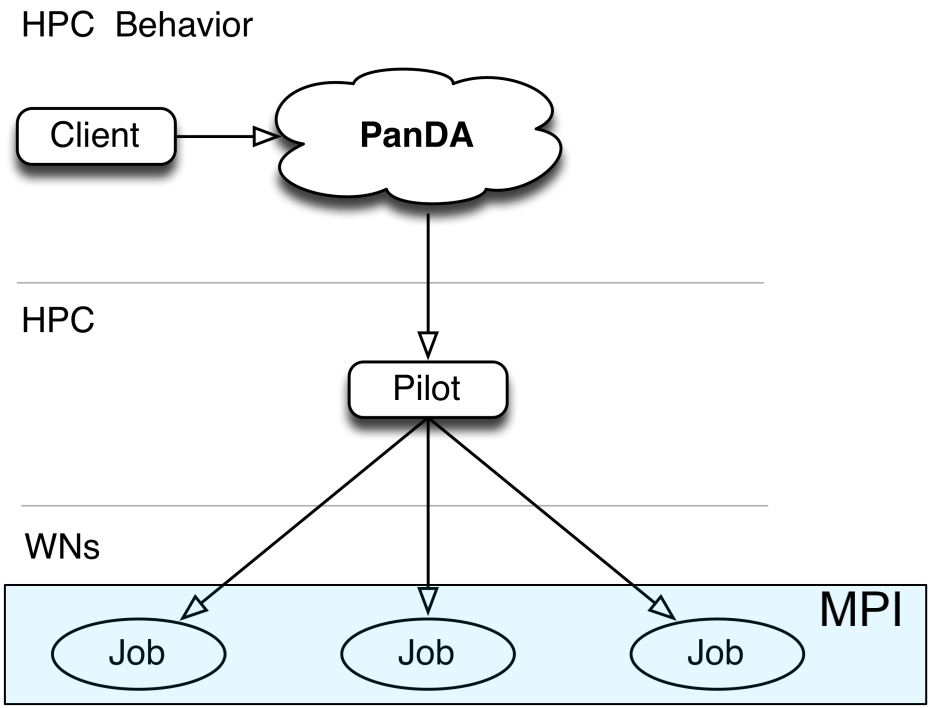
PanDA setup on Titan@OLCF



Pilot on HPC with MPI wrapper



“One to One”

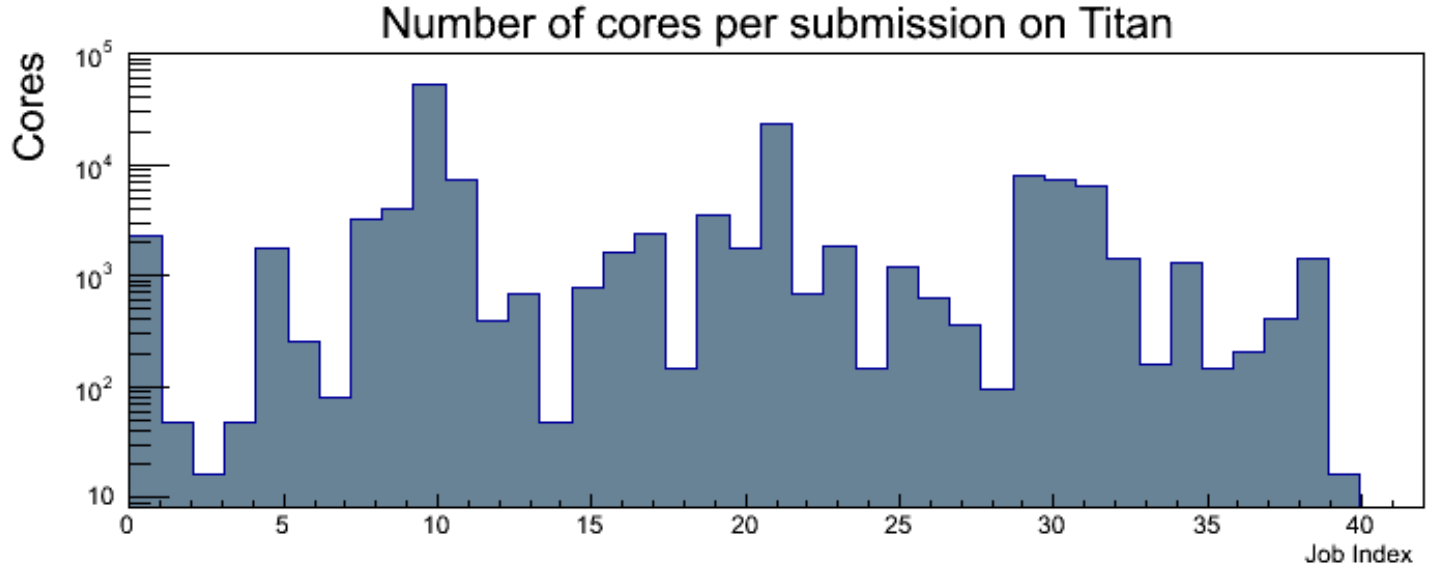


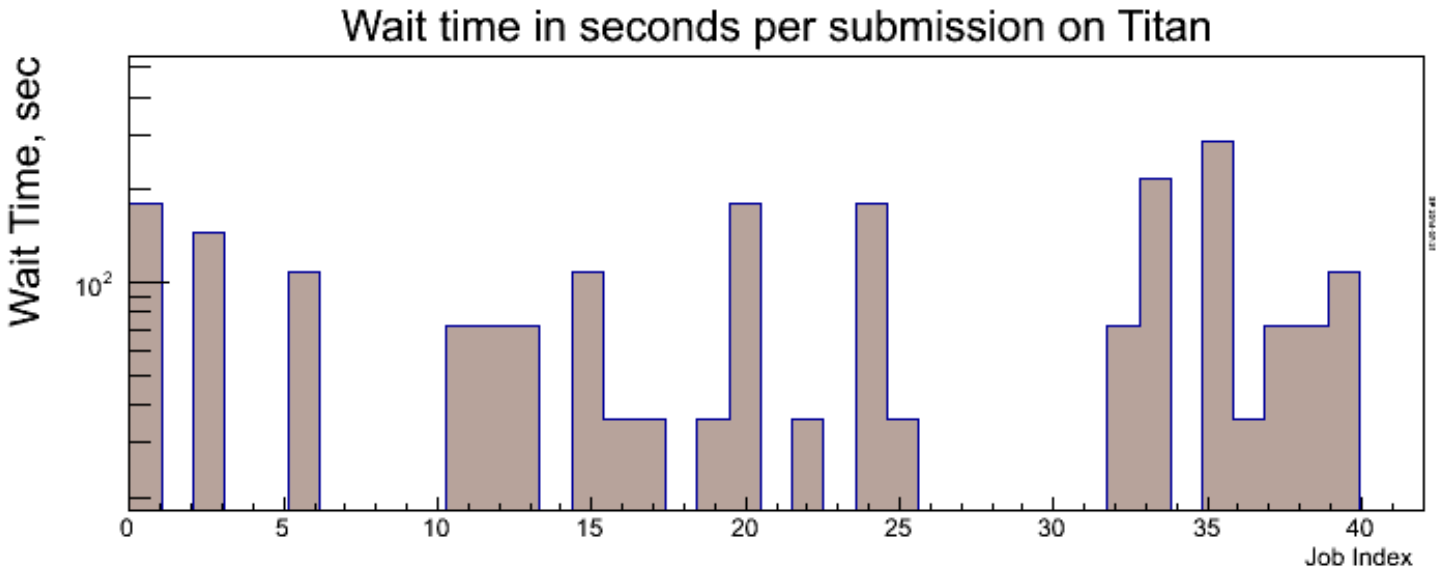
“One to Many”

Pilot stress tests on Titan

- In May 2014 we ran first 24 hour continuous job submission test via PanDA@EC2 with pilot in backfill mode, with MPI wrappers for two workloads from ATLAS and ALICE
  - Stable operations
  - ~22k core hours collected in 24 hours
  - Observed encouragingly short job wait time on Titan ~4 minutes
- Ran second set of tests in July 2014, with pilot modifications that were based on information obtained from the first test
  - Limit on number of nodes removed in pilot
  - Job wait time limit introduced – 5 minutes
  - 145763 core hours collected
  - Average wait time ~70 sec
  - Observed IO related effects that need to be understood better

July pilot tests on Titan



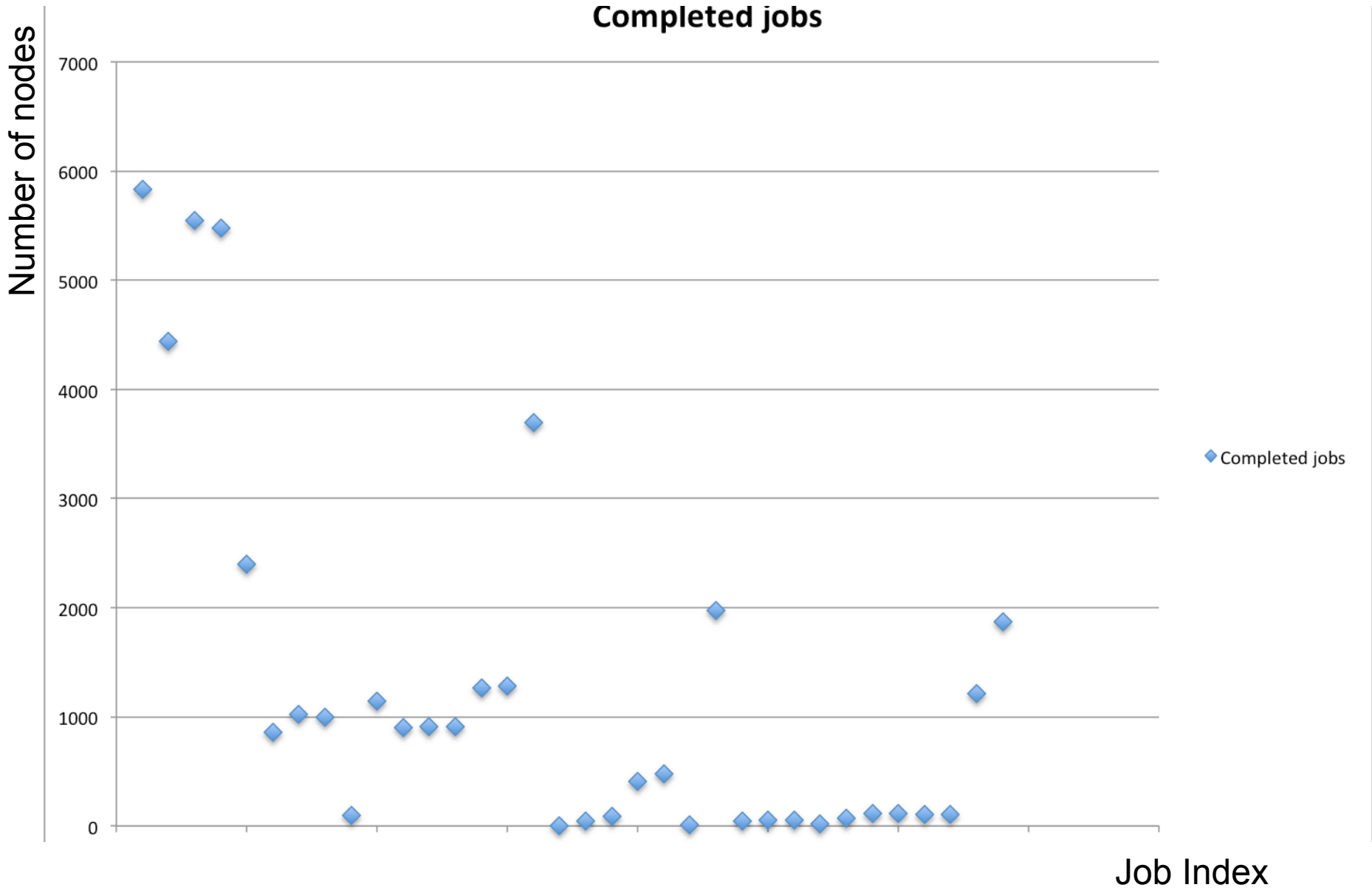


Average wait time 70 seconds !

August pilot tests

- Testing algorithm for internal rescheduling of payload in pilot
  - Pilot gets free resource information from Titan's resource manager
  - Forms job parameters according to free resources and queue policies
  - Submits job to PBS
  - If job exceeds wait time limit, pilot cancels the job and repeats the cycle
- Wait time limit for a job in PBS was set to 2 minutes
- Ran continuously for ~10 hours
- Highly CPU bound payload to avoid IO issues
- Were able to collect ~ 200,000 core hours
- Max number of nodes per job – 5835 (93360 cores)
  - Close to entire ATLAS Grid in size!
- Used ~2.3% of all Titan core hours or ~14.4% of free core hours
- Test data analysis is still in progress

August pilot tests on Titan



Functional pilot tests at NERSC

- PanDA pilot developed for Titan (old version) was tested on Cray machines at NERSC (Hopper, Edison) in July
  - Introduced new corresponding queues in PanDA
  - Minor pilot changes that were needed to comply with NERSC queue policies
  - Most changes were in ‘static’ parameters, like name of the queue and partition, number of cores per node etc.
- Test workloads successfully submitted via PanDA
- Due to different batch policies of NERSC machines against ORNL facilities, usage of backfill mode may be not efficient.
  - Many small jobs
  - Relatively little opportunities for backfill
  - Need to be studied in more details

Workloads on Titan

- Several standalone workloads were ported to Titan
  - Good binary compatibility with Grid, system incomatibility needs to be worked around
- Root,etc
  - Root based ATLAS analysis
  - Limits setting code (aTGC)
- Event generators ported
  - SHERPA (v. 2.0.b2 and v. 1.4.3)
  - MadGraph 5 (v. 1.5.12)
  - ALPGEN v 1.4
  - Simple examples and tutorials for EvGens run
- Geant 4, including multithreaded v4.10
- Full GEANT simulation chains for ALICE@LHC and EIC@RHIC tested
- CVMFS via Parrot works on Titan login nodes. How to expose them to WN?
  - Cvmfs copy to shared file system. Worked nicely for ALICE
  - Didn’t work for full ATLAS repository (copy takes long time, lost sessions, etc)

ATLAS workloads on Titan I

- We started collaboration with group of Prof. Rostislav Konoplich from Manhattan College and NYU
- They are interested in running several large scale simulations to support their ATLAS analysis
- Ported, stand alone, custom MadGraph5 (MG5_aMC_v2.1.2) to Titan
  - Ran several processes on login nodes for validation purposes
    - pp_X2pmin_ZZ_4l_0j1j2j at 14 TeV
    - pp_X2pmin_ZZ_4l_0j1j at 14 TeV and 8 TeV
- Dimitriy Krasnopertsev, a PhD student from MEPhI, started physics validation under Prof. Konoplich guidance this week

ATLAS workloads on Titan II

- CVMFS copy didn't work out for ATLAS
- Needed other mechanism to expose ATLAS software to worker nodes (WN)
- Looked at individual releases installation using pacman
  - Worked very well
  - Releases, 17.2.12, 17.2.11.15 and 18.9.0 installed on Titan shared file system
  - Release 18.9.0 installed at NERSC
  - Many thanks to Grigory Rybkin for help with release installations
  - Many thanks to Vakho Tsulaia for help with missing libraries on HPC!
- Tested Athena "Hello World" on Titan's WNs this week
  - Single threaded Athena on a single WN
  - AthenaMP on a single WN with 16 threads
  - MPI wrapped single and multi-threaded Athena on multiple WN
  - Integration with PanDA pilot started
- Need help from ATLAS with realistic workloads for Titan!

WebFTS on Titan

- Started discussion with WebFTS developers at CERN
- Interest in trying the service on HPC, specifically at OLCF
- Interest from OLCF
- Some indication of interest from NERSC
- Will have a joint meeting between PanDA team, CERN, OLCFin this month

Summary

- Work on integration of OLCF, NERSC machines and PanDA is in progress
- Key PanDA system components ported to Titan@OLCF
- Pilot now uses information about free worker nodes on Titan for job submission. Short job wait times
- MPI wrappers created for several workloads
- Many stand alone workloads ported
- Work on ATLAS workloads is in progress
  - ATLAS release installed on Titan, simple Athena examples tested
  - Help from ATLAS needed with realistic payloads!
- Ran pilot stress tests in backfill mode
  - Stable operations
  - Short wait times
  - Demonstrated significant resource collection capability
- Collaboration with multiple groups and experiments

The End