Big PanDA.
Next generation Workload Management System for Big Data

Alexandre Vaniachine

(ANL)

Alexei Klimentov, Sergey Panitkin, Dantong Yu, Torre Wenaus

(BNL)

Kaushik De, Gergely Zaruba

(UTA)

Outline

- Introduction
- PanDA in ATLAS
- ASCR project
- Summary

Next Generation “Big PanDA”

- ASCR and HEP funded project “Next Generation Workload Management and Analysis System for Big Data”. Started in September 2012.
- Generalization of PanDA as meta application, providing location transparency of processing and data management, for HEP and other data-intensive sciences, and a wider exascale community.
- Project participants from **ANL**, **BNL**, **UT Arlington**
- **Alexei Klimentov –** Lead PI, **Kaushik De** Co-PI
- **WP1** (Factorizing the core): Factorizing the core components of PanDA to enable adoption by a wide range of exascale scientific communities (UTA, K.De)
- **WP2** (Extending the scope): Evolving PanDA to support extreme scale computing clouds and Leadership Computing Facilities (BNL, S.Panitkin)
- **WP3** (Leveraging intelligent networks): Integrating network services and real-time data access to the PanDA workflow (BNL, D.Yu)
- **WP4** (Usability and monitoring): Real time monitoring and visualization package for PanDA (BNL, T.Wenaus)

PanDA in ATLAS

- The ATLAS experiment at the LHC - Big Data Experiment
  - ATLAS Detector generates about 1PB of raw data per second – most filtered out
  - As of 2013 ATLAS DDM manages ~140 PB of data, distributed world-wide to 130 of WLCG computing centers
  - Expected rate of data influx into ATLAS Grid ~40 PB of data per year
  - Thousands of physicists from ~40 countries analyze the data
- PanDA project was started in Fall 2005. **P**roduction **an**d **D**ata **A**nalysis system
  - Goal: An automated yet flexible workload management system (WMS) which can optimally make distributed resources accessible to all users
  - Originally developed in US for US physicists
- Adopted as the ATLAS wide WMS in 2008 (first LHC data in 2009) for all computing applications
- Now successfully manages O(10E2) sites, O(10E5) cores, O(10E8) jobs per year, O(10E3) users

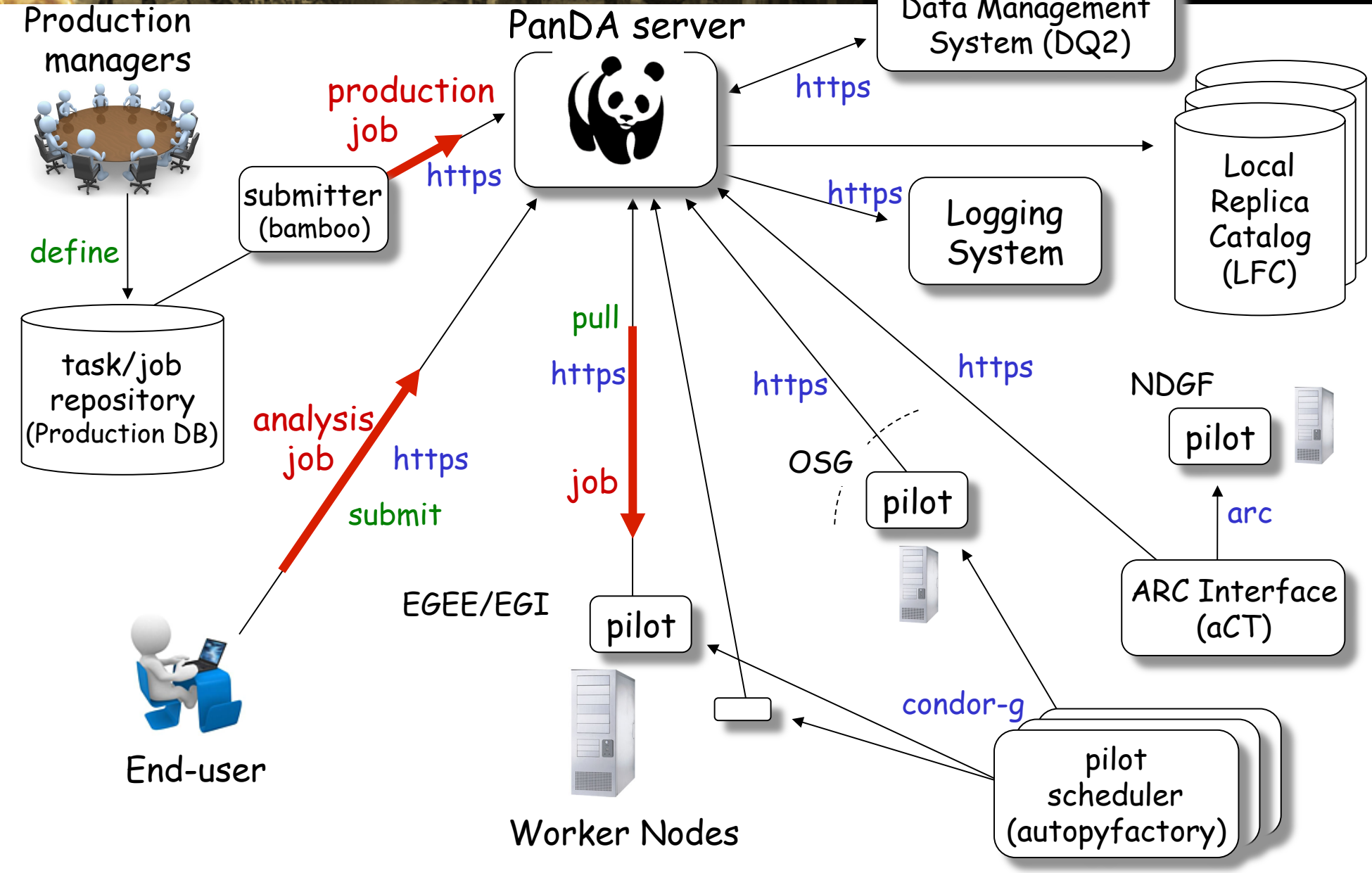
PanDA Philosophy

- PanDA WMS design goals:
  - Achieve high level of automation to reduce operational effort
  - Flexibility in adapting to evolving hardware and network configurations
  - Support diverse and changing middleware
  - Insulate user from hardware, middleware, and all other complexities of the underlying system
  - Unified system for central MC production and user data analysis
  - Incremental and adaptive software development

Key Features of PanDA

- Pilot based job execution system
  - Condor based pilot factory
  - Payload is sent only after execution begins on CE
  - Minimize latency, reduce error rates
- Central job queue
  - Unified treatment of distributed resources
  - SQL DB keeps state - critical component
- Automatic error handling and recovery
- Extensive monitoring
- Modular design
- HTTP/S RESTful communications
- GSI authentication
- Workflow is maximally asynchronous
- Use of Open Source components

Workload Management



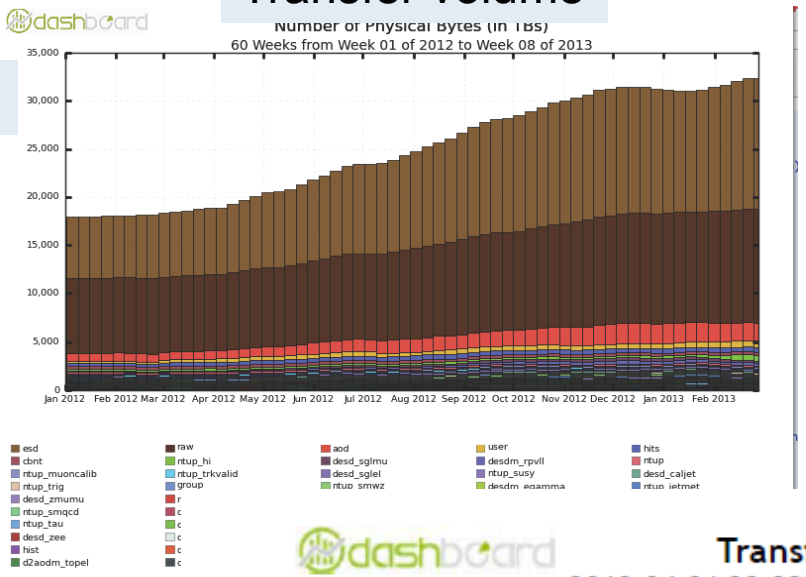
Data Management

- PanDA supports multiple DDM solutions
  - ATLAS Distributed Data Management (DDM) System
  - Pandamover file transfer (using chained Panda jobs)
  - CMS PHEDEX file transfer
  - Federated Xrootd
  - Direct access if requested (by task or site)
  - Customizable lsm (local site mover)
  - Multiple default site movers are available

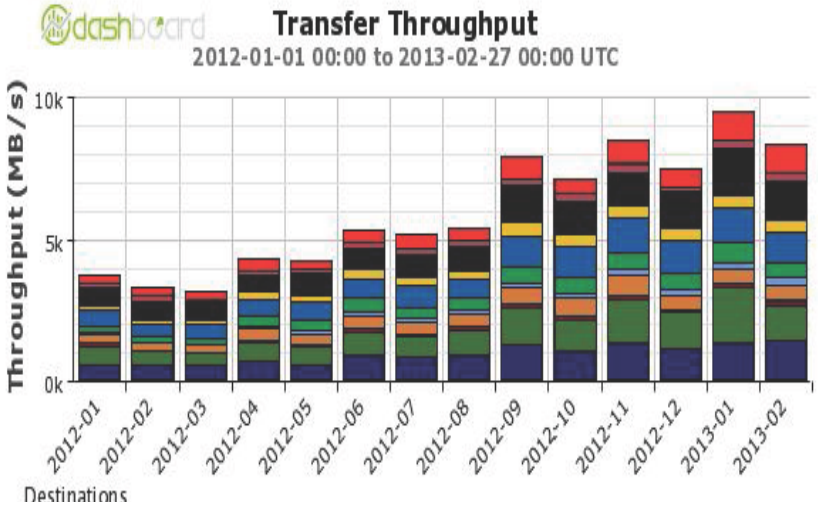
ATLAS Data transfers at a glance

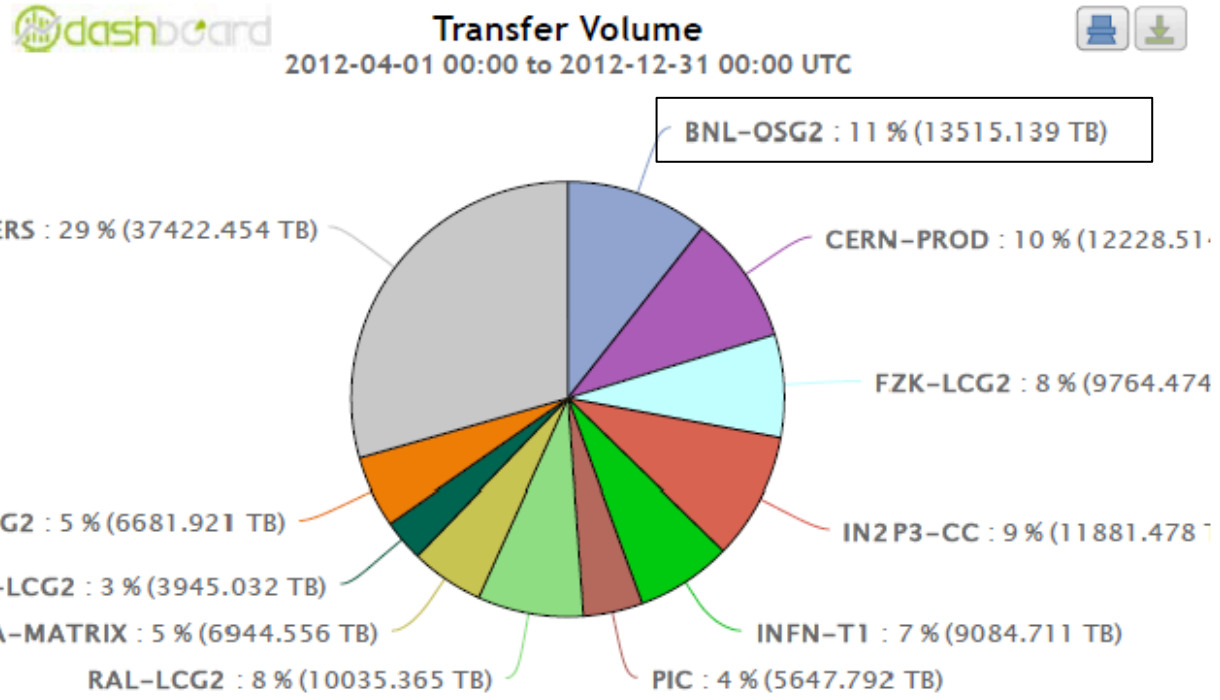
Transfer volume

30 PB ->



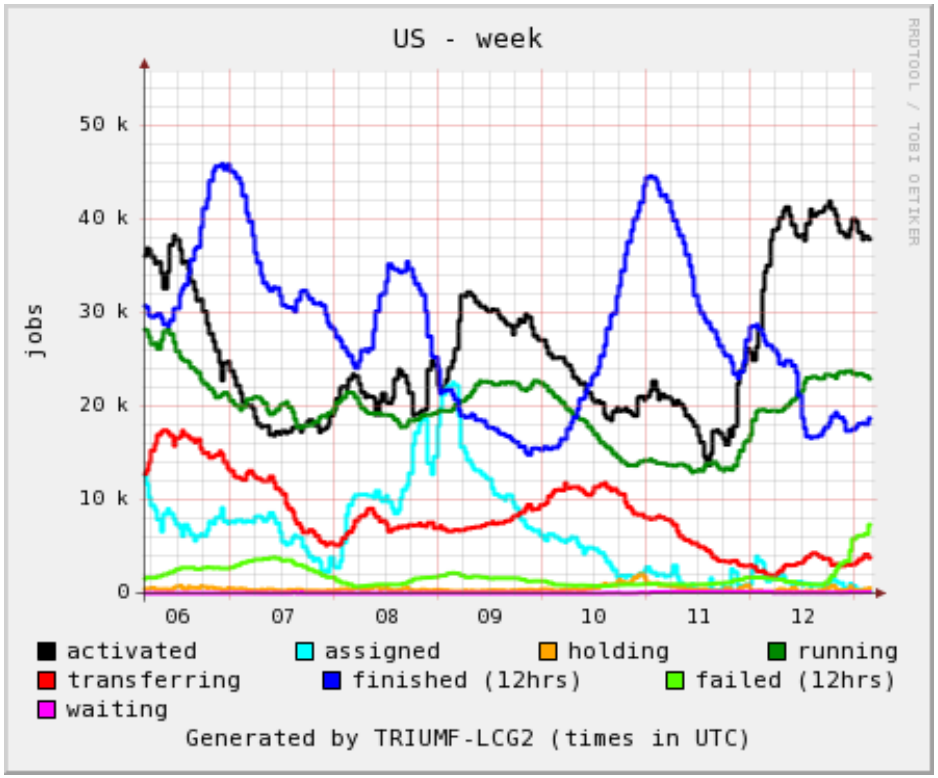
10GB/s->





Job States

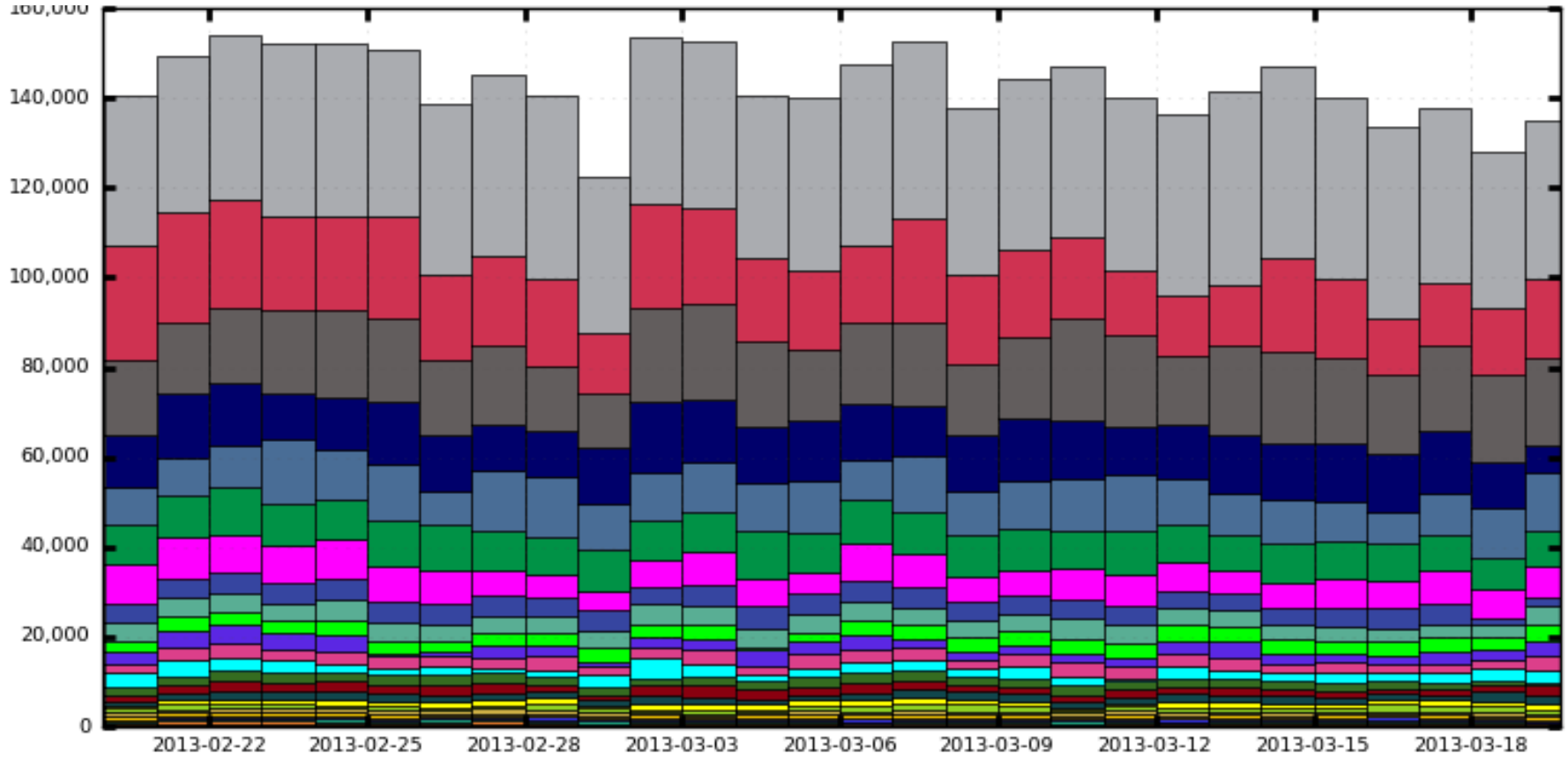
- Panda jobs go through a succession of steps tracked in DB
  - Defined
  - Assigned
  - Activated
  - Running
  - Holding
  - Transferring
  - Finished/failed



PANDA ATLAS running jobs

Average number of concurrently running jobs per day

140k ->



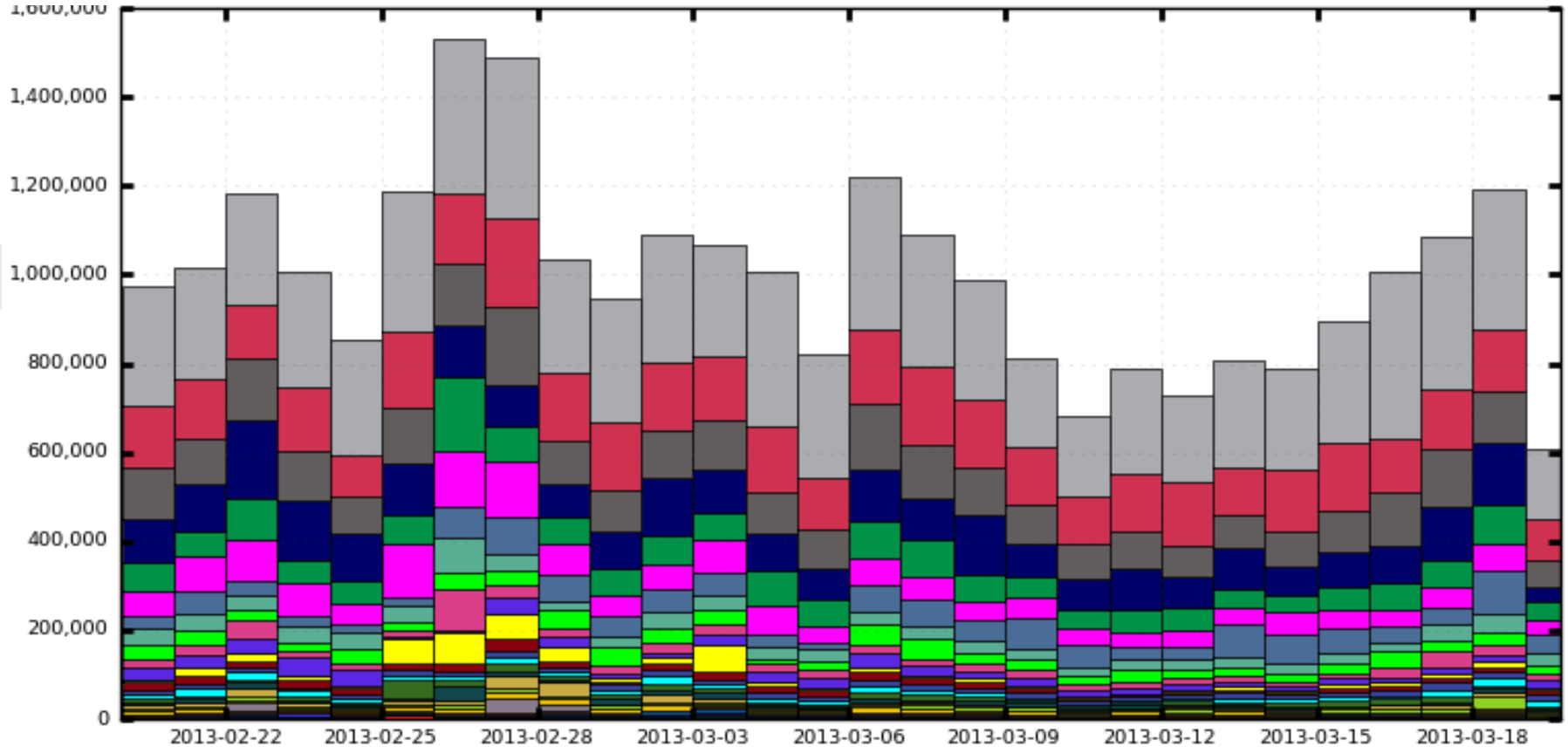
US is a largest resource contributor

USA
FRANCE
ITALY
JAPAN
NETHERLANDS
POLAND
PORTUGAL
SWEDEN
UK
SWITZERLAND
SLOVENIA
TAIWAN
DENMARK, FINLAND, NORWAY, SWEDEN
ISRAEL
AUSTRALIA
CHINA
GERMANY
CANADA
SPAIN
RUSSIA
CZECH REPUBLIC
ROMANIA
SLOVAKIA
CHILE

PANDA ATLAS finished jobs

Finished jobs, per day, for the past month

1M jobs ->



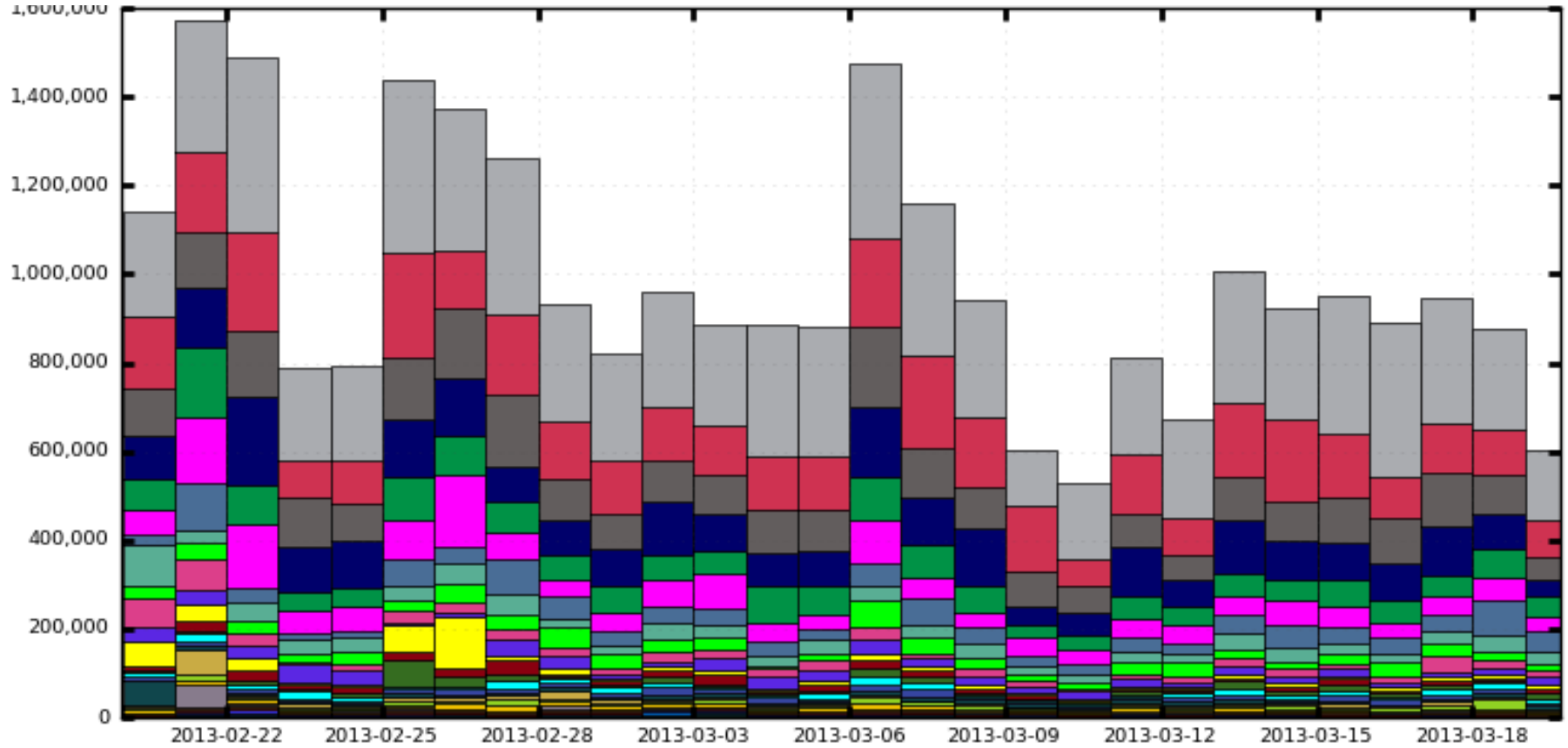
US is a largest resource contributor

USA
FRANCE
SWITZERLAND
RUSSIA
CZECH REPUBLIC
DENMARK, FINLAND, NORWAY, SWEDEN
ROMANIA
CHINA
UK
CANADA
SPAIN
TAIWAN
SLOVENIA
POLAND
AUSTRALIA
AUSTRIA
GERMANY
ITALY
JAPAN
ISRAEL
NETHERLANDS
PORTUGAL
SOUTH AFRICA
SLOVAKIA

PANDA ATLAS submitted jobs

Submitted jobs, per day, for the past month

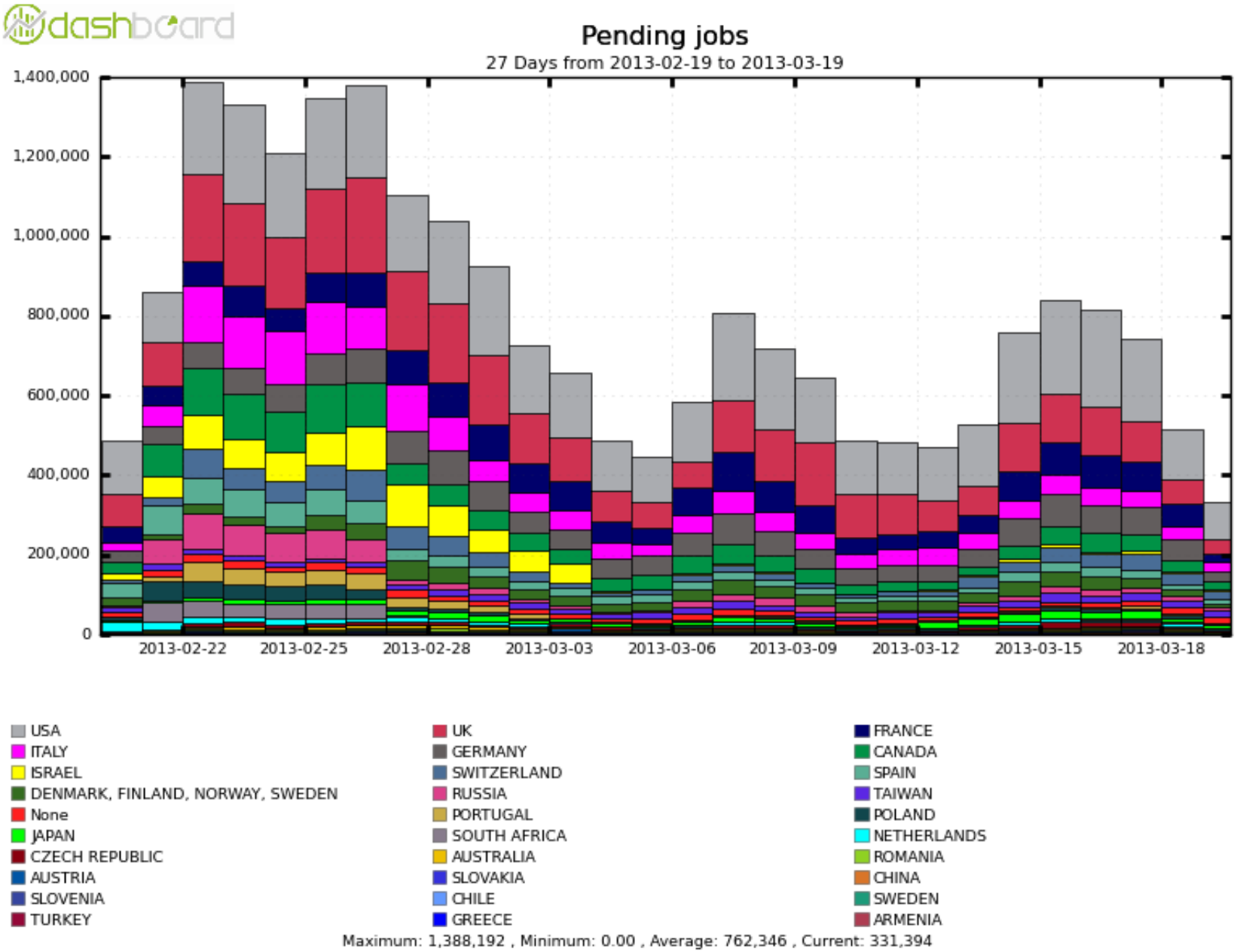
1M jobs



Uneven influx of jobs, spikes in demand, can exceed available resources x10

USA
UK
GERMANY
FRANCE
CANADA
ITALY
SWITZERLAND
SPAIN
JAPAN
RUSSIA
TAIWAN
ISRAEL
CZECH REPUBLIC
DENMARK, FINLAND, NORWAY, SWEDEN
NETHERLANDS
SLOVENIA
POLAND
PORTUGAL
ROMANIA
AUSTRALIA
SOUTH AFRICA
CHINA
None
SLOVAKIA

PANDA ATLAS pending jobs



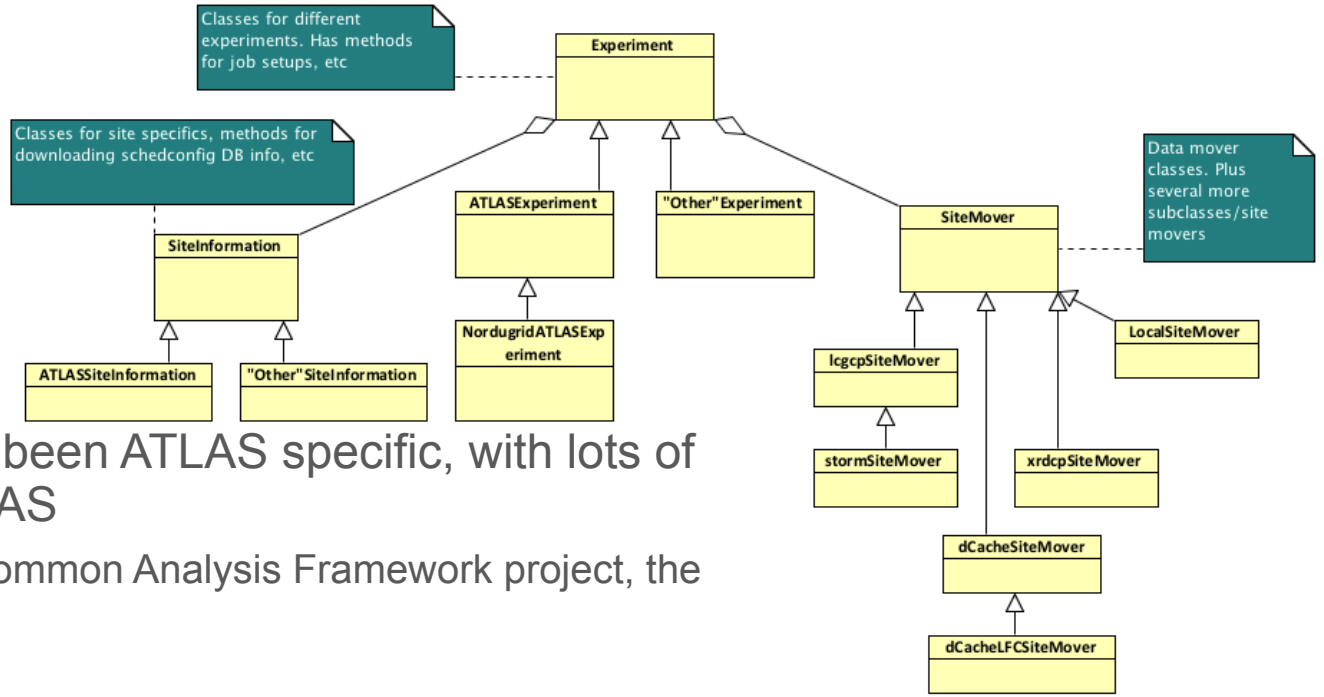
PanDA's Success

- The system was developed by US ATLAS for US ATLAS
- Adopted by ATLAS Worldwide as Production and Analysis system
- PanDA was able to cope with increasing LHC luminosity and ATLAS data taking rate
- Adopted to evolution in ATLAS computing model
- Two leading HEP and astro-particle experiments (CMS and AMS) has chosen PanDA as workload management system for data processing and analysis.
- PanDA was chosen as a core component of Common Analysis Framework by WLCG

Evolving PanDA for advanced scientific computing

- There are tree dimensions to evolution of PanDA
  - Making PanDA available beyond ATLAS and HEP
  - Extending beyond Grid (LCF, Clouds, University clusters)
  - Integration of network as a resource in workload management
- ASCR funded Grant No. DE-FG02-12ER26106
  - 3 new computer professionals were hired by BNL and UTA
  - They will start working on April 1
- In the next few slides we will show activities and progress since September 2012

Extending PanDA beyond HEP. Evolving Panda Pilot



- Until recently the pilot has been ATLAS specific, with lots of code only relevant for ATLAS
  - To meet the needs of the Common Analysis Framework project, the pilot is being refactored
- Experiments as plug-ins
  - Introducing new experiment specific classes, enabling better organization of the code
  - E.g. containing methods for how a job should be setup, metadata and site information handling etc, that is unique to each experiment
  - CMS experiment classes are currently being implemented
- Changes are being introduced gradually, to avoid affecting current production

PanDA for LCF

- Expanding PanDA from Grid to Leadership Class Facilities will require changes
- Each LCF is unique
  - Unique architecture and hardware
  - Specialized OS, “weak” worker nodes, limited memory per WN
  - Code cross-compilation is typically required
  - Unique job submission systems
  - Unique security environment
- Pilot submission to a worker node is typically not feasible
- Pilot/agent per supercomputer or queue model
- Tests on BlueGene/P at BNL. Geant4 port to BG/P
- Got account at NERSC as part of OSG project
- PanDA/Geant4 project at OLCF

PanDA project on OLCF

- Get experience with all relevant aspects of the platform and workload
  - job submission mechanism
  - job output handling
  - local storage system details
  - outside transfers details
  - security environment
  - adjust monitoring model
- Develop appropriate pilot/agent model for Titan
- Geant4 and Project X at OLCF proposal will be initial use case on Titan
  - Collaboration between ANL, BNL, ORNL, SLAC, UTA, UTK
  - Cross-disciplinary project - HEP, NP, HPC

Cloud Computing and PanDA

- ATLAS Distributed Computing set up a few years ago cloud computing project to exploit virtualization and clouds in PanDA
  - Utilize private and public clouds as extra computing resource
  - Mechanism to cope with peak loads on the Grid
- Experience with variety of cloud platforms
  - Amazon EC2
  - Helix Nebula for MC production (CloudSigma, T-Systems and ATOS – all used)
  - Futuregrid (U Chicago), Synnefo cloud (U Vic)
  - RackSpace
  - Private clouds OpenStack, CloudStack, etc…
  - Recent project on Google Compute Engine (GCE)

Running on Google Compute Engine

- US ATLAS and ASCR funded Big Panda project negotiated with Google expansion of the GCE preview project
- Google agreed to allocate additional resources for ATLAS for free
  - ~5M cpu hours, 4000 cores for about 2 month, (original preview allocation 1k cores)
- These are powerful machines with modern CPUs
- Resources are organized as Condor based Panda queue
  - Centos 6 based custom built images, with SL5 compatibility libraries to run ATLAS software
  - Condor head node, proxies are at BNL
  - Output exported to BNL SE
- Work on capturing the GCE setup in Puppet
- We were invited to present results at Google IO 2013

Network as Resource in PanDA

- Work has started on using network information in PanDA with Federated Xrootd Data store
- Continuous stream of PanDA probes is used to evaluate network performance which will be used as a metric of network cost in PanDA
- PerfSonar information will be evaluated as input for job brokering
- Ramp up of these activities in April – May when new hires will come aboard and will start to work under the leadership of Dantong Yu.

Conclusions

- ASCR gave us a great opportunity to evolve PanDA beyond ATLAS and HEP
- Project team was set up
- The work on extending PanDA to LCF has started
  - Submitted proposal to OLCF
- Technical meeting devoted to PanDA on OLCF in summer
- Large scale PanDA deployments on commercial clouds are already producing valuable results
- Strong interest in the project from several experiments and foreign universities and laboratories
  - Opportunity for a common project in the future
  - Workshop in June at BNL

The End

References

- https://twiki.cern.ch/twiki/bin/viewauth/Atlas/PanDA
- http://www.usatlas.bnl.gov/twiki/bin/view/PanDA/WebHome
- http://panda.cern.ch:25880/server/pandamon/query
- Recent Improvements in the ATLAS PanDA Pilot, P. Nilsson, CHEP 2012, United States, May 2012
- PD2P : PanDA Dynamic Data Placement for ATLAS, T. Maeno, CHEP 2012, United States, May 2012
- Evolution of the ATLAS PanDA Production and Distributed Analysis System, T. Maeno, CHEP 2012, United States, May 2012