CHEP 2018
23RD INTERNATIONAL CONFERENCE ON
COMPUTING IN HIGH ENERGY AND NUCLEAR PHYSICS
9-13 July 2018
National Palace of Culture
Sofia, Bulgaria

Big PanDA Workflow Management System on OLCF Titan for HENP and extreme scale applications

K.De, J.C.Kincl, A.Klimentov, R.Mashinistov. D.Oleynik, S.Panitkin, P.Svirin and J.C.Wells

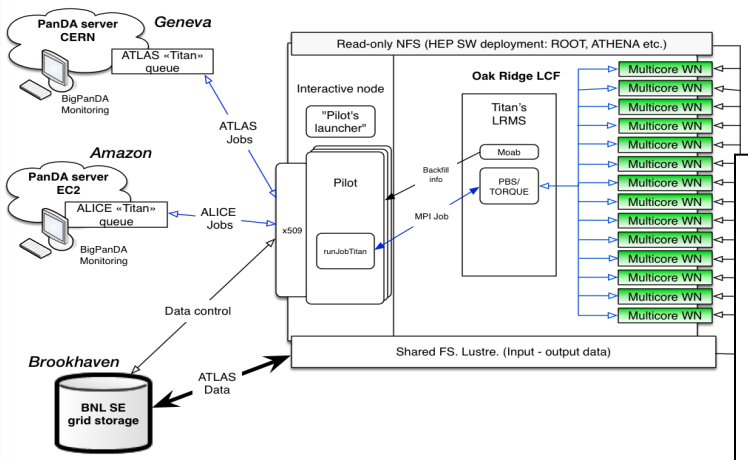
On behalf of the BigPanDA project

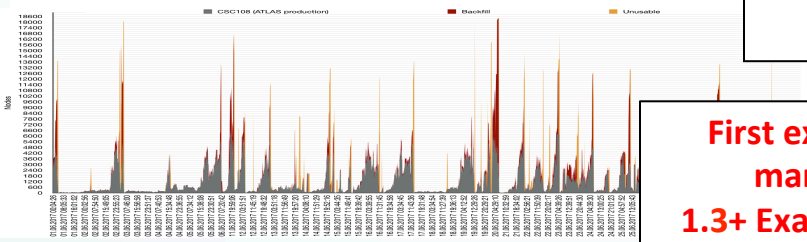
CHEP2018, Sofia, Bulgaria
July 9, 2018

The University of Texas ARLINGTON
BROOKHAVEN NATIONAL LABORATORY
OAK RIDGE National Laboratory
JINR

PanDA. Production and Distributed Analysis Workload Management System

https://twiki.cern.ch/twiki/bin/view/PanDA/PanDA





Global ATLAS operations

Up to ~800k concurrent jobs
25-30M jobs/month
at >250 sites
~1400 ATLAS users

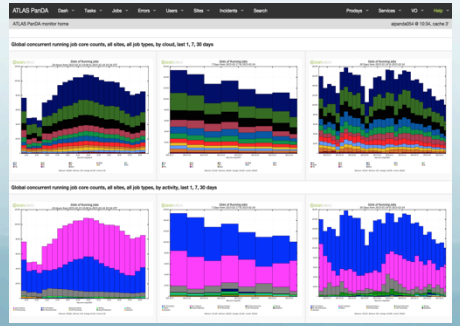
First exascale workload manager in HENP
1.3+ Exabytes processed in 2014 and in 2016-2018
Exascale scientific data processing today

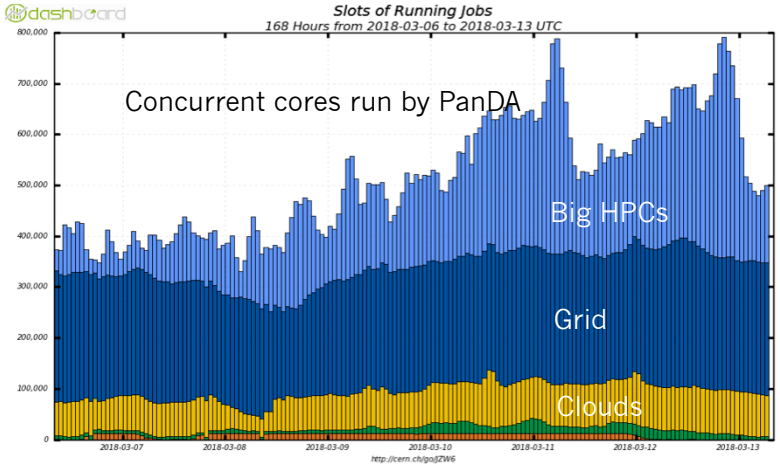
PanDA Brief Story

2005: Initiated for US ATLAS (BNL and UTA)
2006: Support for analysis
2008: Adopted ATLAS-wide
2009: First use beyond ATLAS
2011: Dynamic data caching based on usage and demand
2012: ASCR/HEP BigPanDA project
2014: Network-aware brokerage
2014 : Job Execution and Definition I/F (JEDI) adds complex task management and fine grained dynamic job management
2014: JEDI- based Event Service
2014:megaPanDA project supported by RF Ministry of Science and Education
2015: New ATLAS Production System, based on PanDA/JEDI
2015 :Manage Heterogeneous Computing Resources
2016: DOE ASCR BigPanDA@Titan project
2016:PanDA for bioinformatics
2017:COMPASS adopted PanDA , NICA (JINR)
PanDA beyond HEP : BlueBrain, IceCube, LQCD

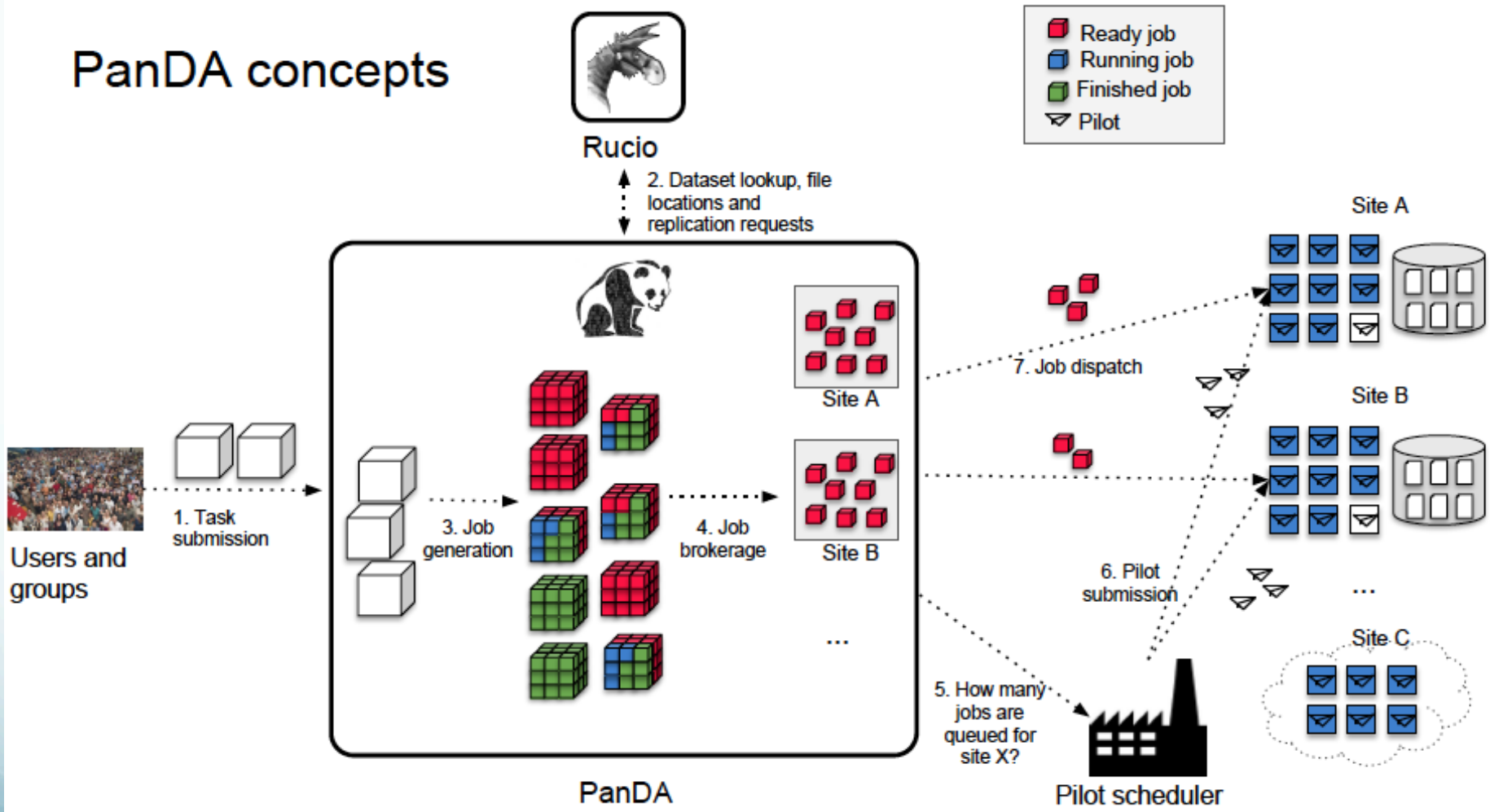
BigPanDA Monitor

http://bigpanda.cern.ch/



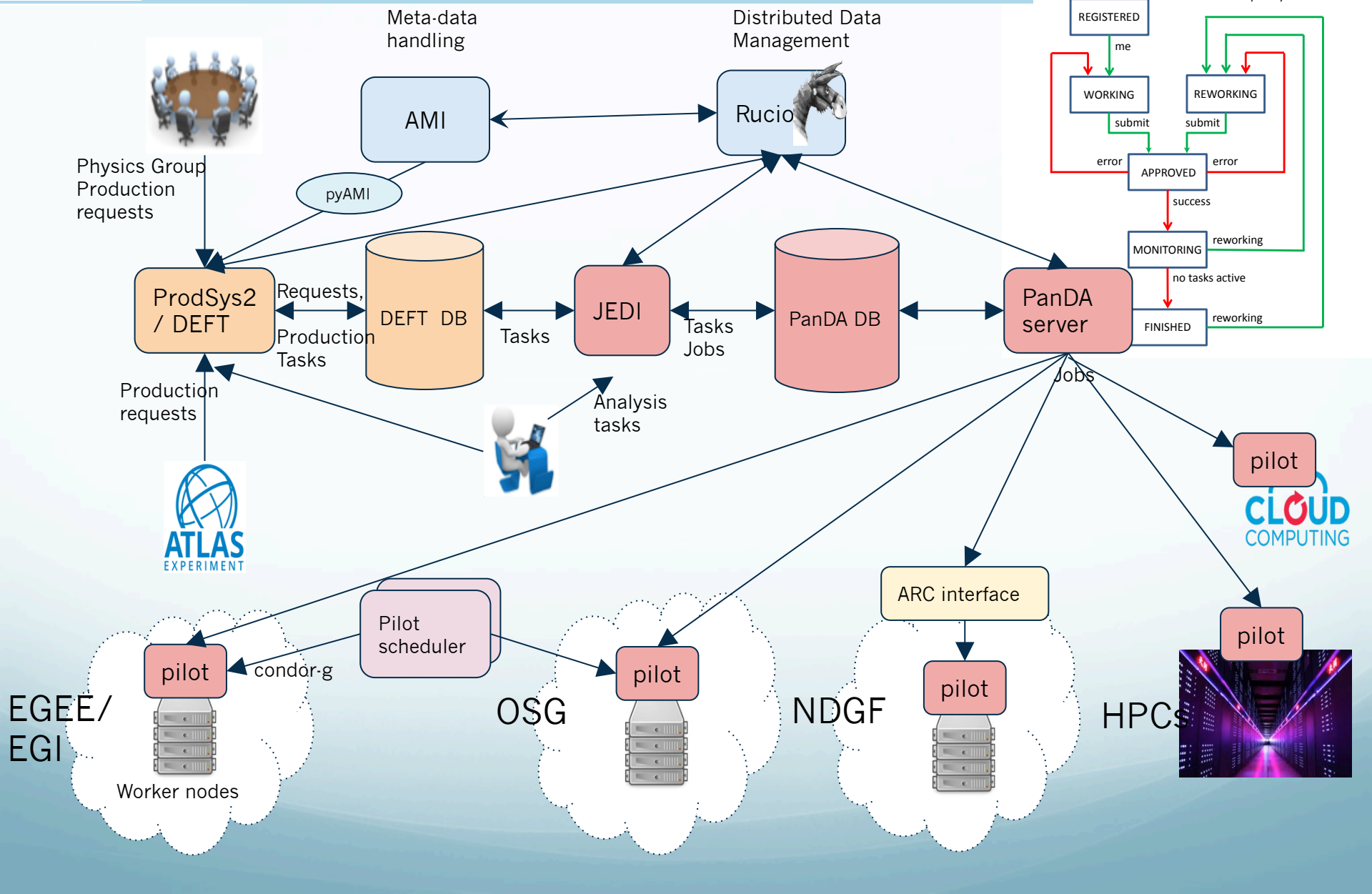


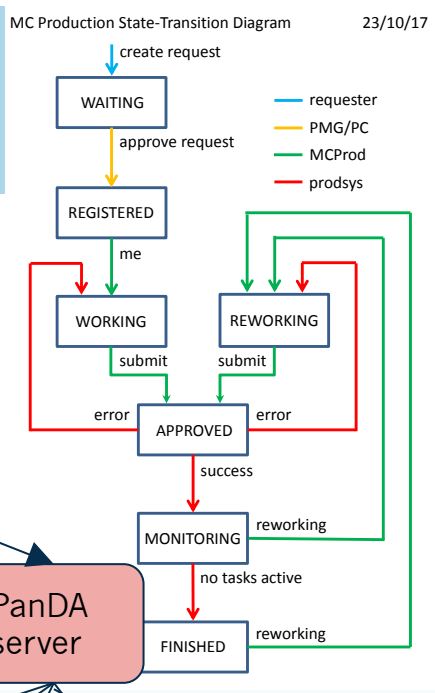
PanDA Workload Management System



CHEP talk about new PanDA developments : "Harvester : an edge service harvesting heterogeneous resources for ATLAS", Track 3 – Distributed computing, 12 Jul 2018, 11:15
https://indico.cern.ch/event/587955/contributions/2937391/

ATLAS Workflow and Workload Management





ATLAS HPC Initiative

- How to get time on supercomputers ?
- How to interface supercomputers to HEP Distributed Computing ?
- How to run HEP code on supercomputers and how to do it efficiently ?

More in CHEP talk :
"The Future of Distributed Computing Systems in ATLAS: Boldly Venturing Beyond Grids",
Track 3 – Distributed computing, 12 Jul 2018, 14:30 https://indico.cern.ch/event/587955/contributions/2937395/

BigPanDA Workflow Management on Titan for High Energy and Nuclear Physics and for Future Extreme Scale Scientific Applications

- BigPanDA project: an extension of PanDA beyond the grid and HEP as well as use of PanDA for projects and experiments beyond ATLAS and HEP
- A DOE ASCR and HEP funded project since 2012; a collaboration between BNL, UTA, ORNL and Rutgers University since 2015 (BigPanDA++)

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

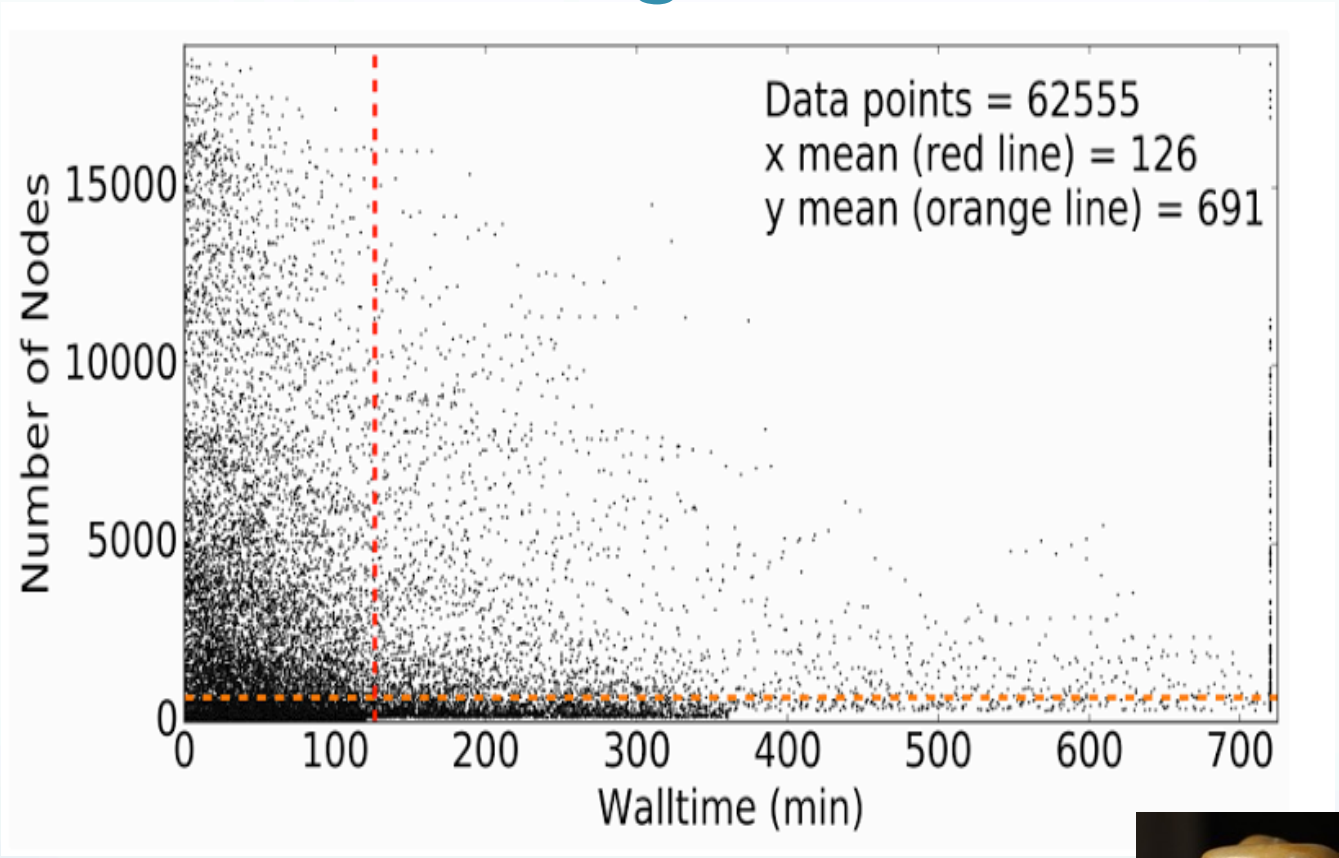
What is a Leadership Computing Facility (LCF)?

- Collaborative USA DOE Office of Science user-facility program at ORNL and ANL
- Mission: Provide the computational and data resources required to solve the most challenging problems.
- 2-centers/2-architectures to address diverse and growing computational needs of the scientific community
- Highly competitive user allocation programs (INCITE, ALCC).
- Projects receive 10x to 100x more resource than at other generally available centers.
- LCF centers partner with users to enable science & engineering breakthroughs (Liaisons, Catalysts).

The Worldwide LHC Computing Grid and a leadership computing facility are of comparable compute capacity.

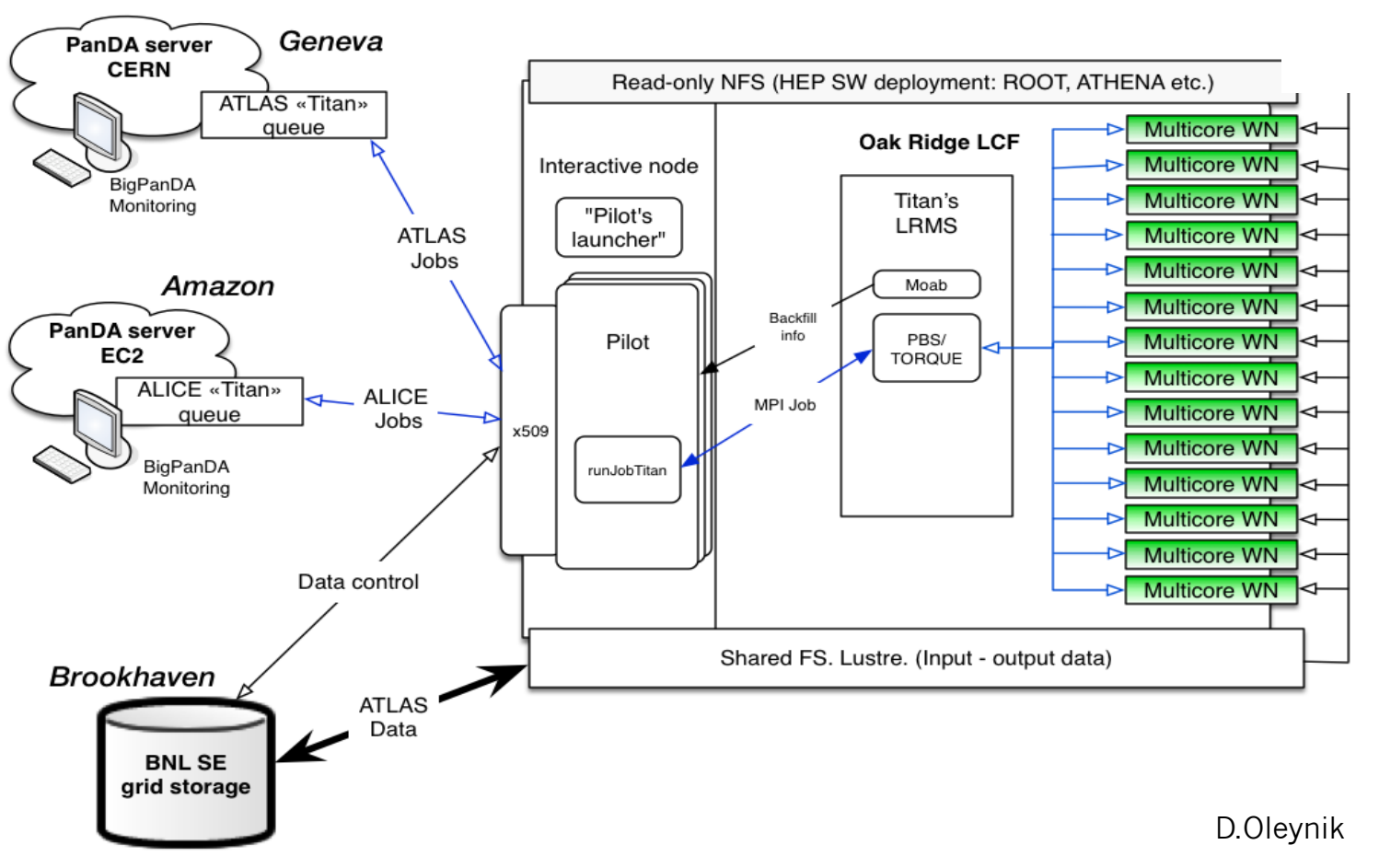
- **WLCG: 750,000 x86 compute cores (ATLAS 300K x86 cores)**
- **Titan: 300,000 x86 compute cores and 18,000 GPUs**

OLCF. Understanding Backfill Slot Availability



- Mean Backfill availability: 691 worker nodes for 126 minutes.
- Up to 15K nodes for 30-100 minutes
- Large margin of optimization

OLCF Titan Integration with PanDA

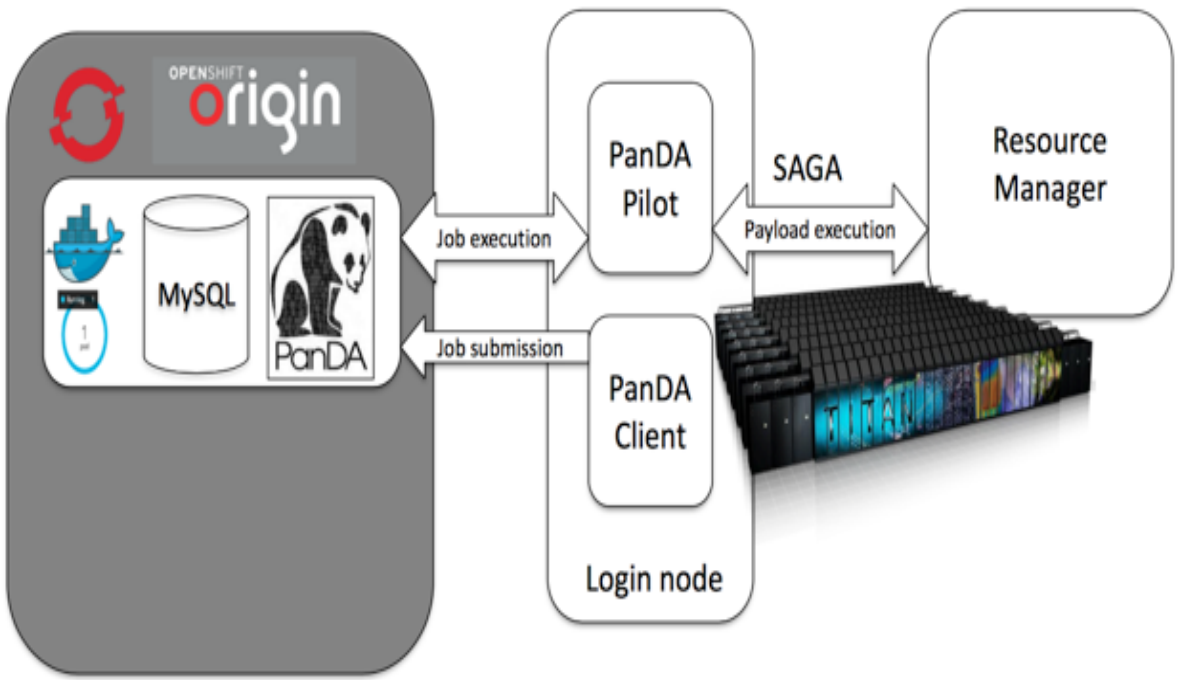


First large scale HPC integrated into ATLAS distributed computing through the BigPanDA project funded by DOE-ASCR
Team leaders: A.Klimentov (BNL), J. Wells (ORNL), S.Jha (Rutgers U), K.De (U of Texas-Arlington)
300 million TITAN core hours in past 12 months, both backfill usage and ALCC allocation

D. Oleynik, S. Panitkin, M. Turilli, A. Angius, S. Oral, K. De, A. Klimentov, J. C. Wells and S. Jha,
"High-Throughput Computing on High-Performance Platforms: A Case Study",
IEEE e-Science (2017) available as: https://arxiv.org/abs/1704.00978

PanDA Server at OLCF: Broad application across domains

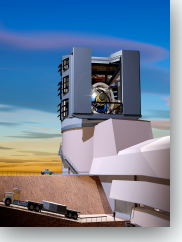
- In March 2017 a new PanDA server instance has been set up at ORNL to serve various experiments. This installation the first at OLCF to demonstrate application of a container cluster management and orchestration system, Red Hat OpenShift Origin.
  - We are looking forward for experimenting with GPU payloads with PanDA provided by ATLAS
- OpenShift, when fully in production, will give OLCF users the ability to deploy and manage infrastructure services
  - https://www.olcf.ornl.gov/2017/06/05/olcf-testing-new-platform-for-scientific-workflows/



Key Contributors:
Jason Kincl (ORNL),
Ruslan Mashinistov (BNL)

PanDA Server at OLCF: PanDA WMS beyond High Energy and Nuclear Physics

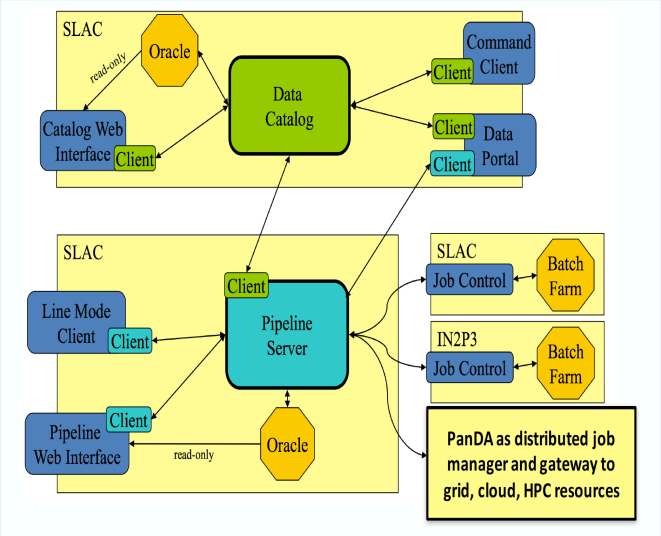
- PanDA designed to support MultiVO
  - Different VO (Experiments) may share same PanDA server instance
  - Server and Pilot plugins allows to tune pre/post-processing VO specific procedures
  - Monitoring is not VO specific
- If VO requires high scalability (hundreds of thousands jobs per day, on wide range of resources) dedicated instance may be deployed
- Beyond HENP
  - Biology / Genomics: Center for Bioenergy Innovation at ORNL
  - Molecular Dynamics: Prof. K. Nam (U. Texas-Arlington)
  - IceCube Experiment
  - Blue Brain Project (BBP), EPFL
  - LSST (Large Synoptic Survey Telescope) project/DESC collaboration
  - LQCD, US LQCD Project
  - nEDM (neutron Electric Dipole Moment Experiment), ORNL

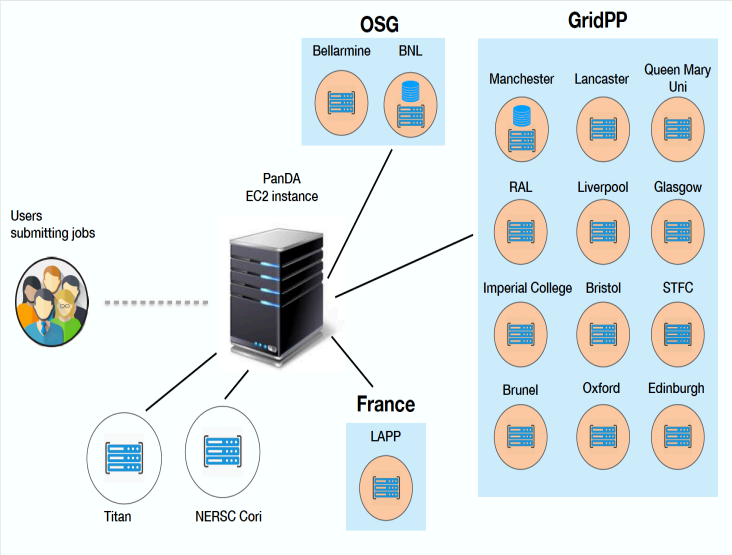


LSST Dark Energy Science Collaboration

- Collaboration with LSST/DESC since 2013 in terms of BigPanDA project
- The LSST Science Pipelines can process data from several telescopes using LSST's algorithms
- Pipeline to PanDA WMS submission has been implemented and tested with some 'Hello word' jobs but also with a standard DESC simulation workflow

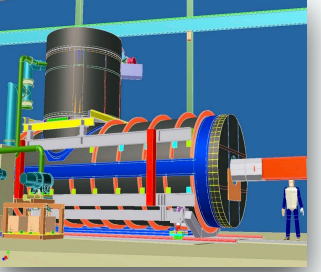
Key contributors : P.Svirin and S.Panitkin (BNL)





- Facilities
  - OSG (BNL and Bellearmine)
  - GridPP
    - 31 endpoints on 12 sites
  - LAPP Annecy (France)
  - LSST storage at 7 EU sites
    - Production results are transferred automatically to NERSC and BNL
- Ongoing process for developing production system (workflow management) above PanDA
- Evaluating backfill mode at OLCF

More information you can find in the presentation: "Using PanDA WMS for LSST Simulations on Distributed Infrastructure"; "Track 3 – Distributed computing"
https://indico.cern.ch/event/587955/abstracts/84566/



nEDM. Neutron Electric Dipole Momentum Experiment at Oak Ridge National Laboratory

- Detailed nEDM detector simulations were executed on Titan via PanDA WMS
  - nEDM SW team is in preparation for a future computational campaign
- nEDM PanDA is migrating from EC2 instance to OLCF instance - since only OLCF resources are going to be used for processing
  - Short time of processing of events makes nEDM good candidate for backfill consumption at OLCF

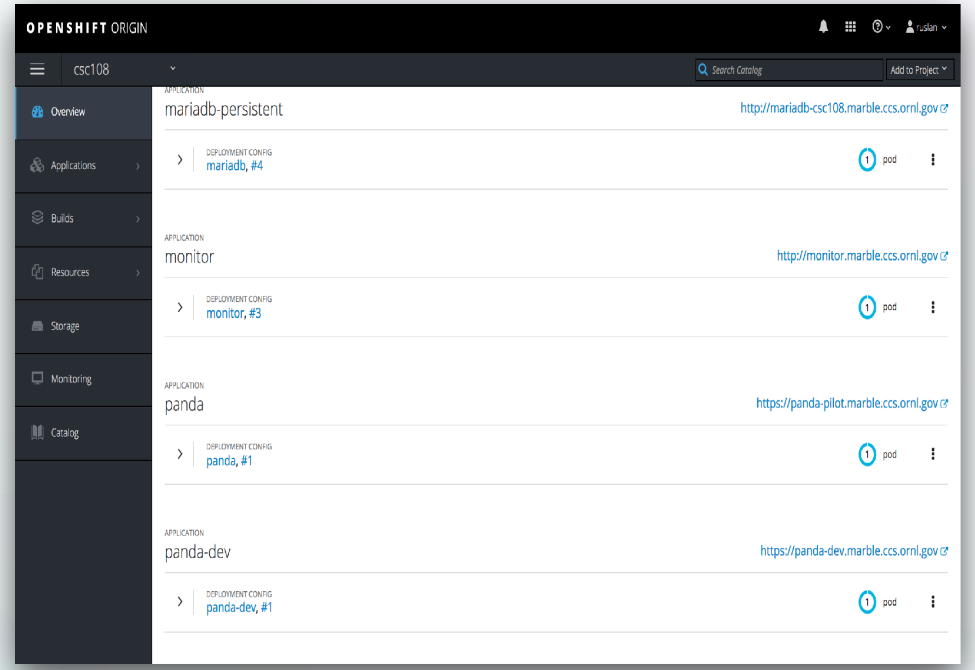
Lattice QCD Computations

- LQCD is a grand challenge subject, with large-scale computations consuming a considerable fraction of publicly available supercomputing resources.
- The computations typically proceed in two phases: in the first phase, one generates thousands of configurations of the strong force fields (gluons), colloquially referred to as gauge fields. This computation is a long-chain Monte Carlo process, requiring the focused power of leadership class computing facilities for extended periods. In the second phase, these configurations are analyzed.
- Until a few years ago, the analysis phase would often account for a relatively small part of the cost of the overall calculation. In recent years, however, focus has turned to more challenging physical observables and new analysis. As a result, the relative costs have shifted to the point where analysis often requires an equal or greater amount of computation than gauge field generation.
- In 2017, as a part of SciDAC-4 funded project, a collaboration was formed between several US LQCD groups and BigPanDA team with the goal to adopt PanDA WMS for the needs of the SciDAC-4 LQCD computational program
- A distributed environment for LQCD computations has been set up using PanDA Server instance deployed at the Amazon Cloud
- Variety of payloads, MPI and non-MPI:
  - HPC (Titan, Cori): GPU-based, multi-node, occupying ~8000 nodes per job, ~20 hours per each job, independent jobs
  - Institutional clusters BNL, TJL: GPU-based, single-node, ~12 hours each, with workflow management
  - New kinds of payloads will be available for Summit in terms of Early Science Program on summit
- LQCD code aggressively uses GPUs and it provides an interesting use-case, when PanDA brokers CPU/GPU applications.

Key contributors : P.Svirin and S.Panitkin (BNL)

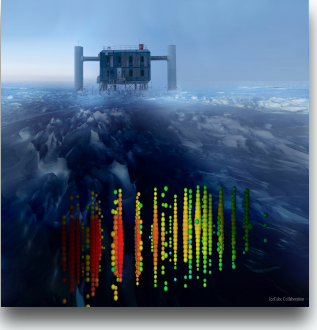
More information you can find on the poster: "PanDA WMS for Lattice QCD Computations" in "Track 3 – Distributed computing"
https://indico.cern.ch/event/587955/abstracts/84565/

PanDA in container



Key contributors : R.Mashinistov(BNL) and S.Wilkinson (UTA)

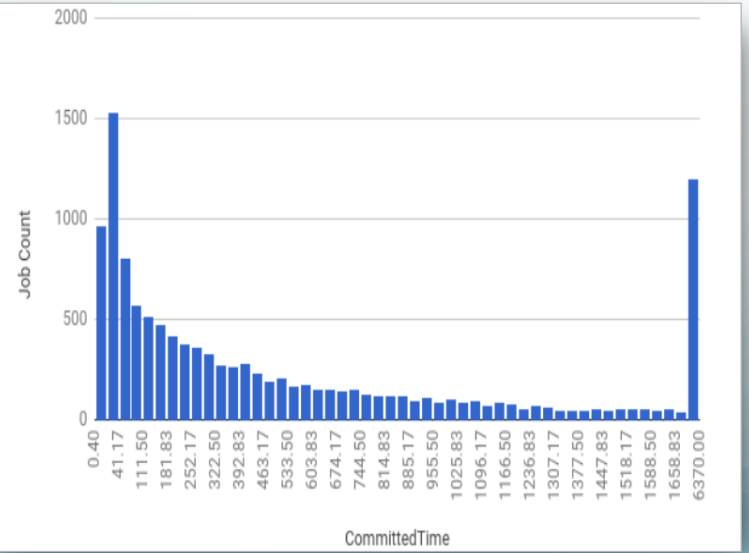
- Persistent MariaDB
- PanDA Monitor (Django)
- Production PanDA server
  - Initial implementation of edge service.
- Developer PanDA server
  - Currently working on support of the Harvester: next generation of PanDA resource facing service



IceCube payloads on Titan

BigPanDA team and IceCube experts joint R&D and demo project

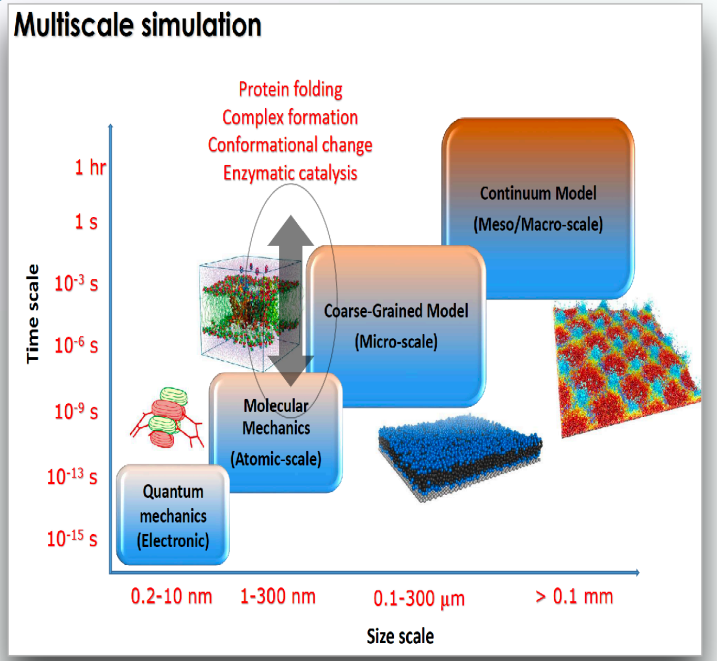
- Application: NuGen package - GPU application for atmospheric neutrinos simulations and analysis
- Application packed in Singularity container.
- Whole node, but not MPI application:
- ~45000 jobs in campaign (5000 input files)
  - Remote stage-in/-out the data from GridFTP storage with GSI authentication
- IceCube payloads were not designed initially for support of MPP, so execution of this kind of payload in the traditional way of HPC will be inefficient (one node per job is not good for LCF's policies)
- PanDA on HPC allows to combine this kind of payload into assemblies and executes them as a simple MPI application (it was required to get efficient execution of ATLAS payloads)
  - To be effective this approach is required to join payloads with similar wall time into assemblies
    - IceCube payloads do not have this type of characterization yet

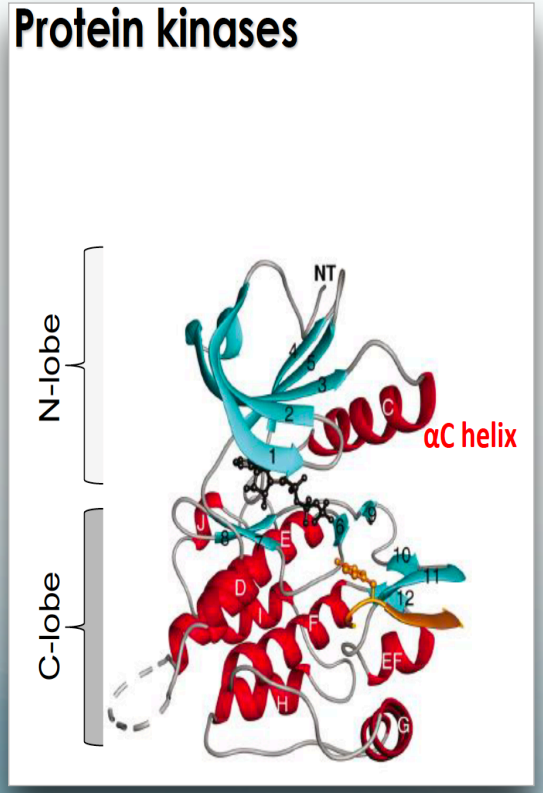


- To optimize quota usage, the processing will be conducted in a few steps.
- On each step, we will run multiple job processes in parallel via MPI wrapper
- Starting with the walltime=20min on the first step, all failed jobs (jobs not completed because of the walltime limit) will be resubmitted on the next step with longer walltime

Molecular Dynamics

- Simulating Enzyme Catalysis, Conformational Change, and Ligand Binding/Release. (Prof. Kwangho Nam (University of Texas at Arlington, USA)

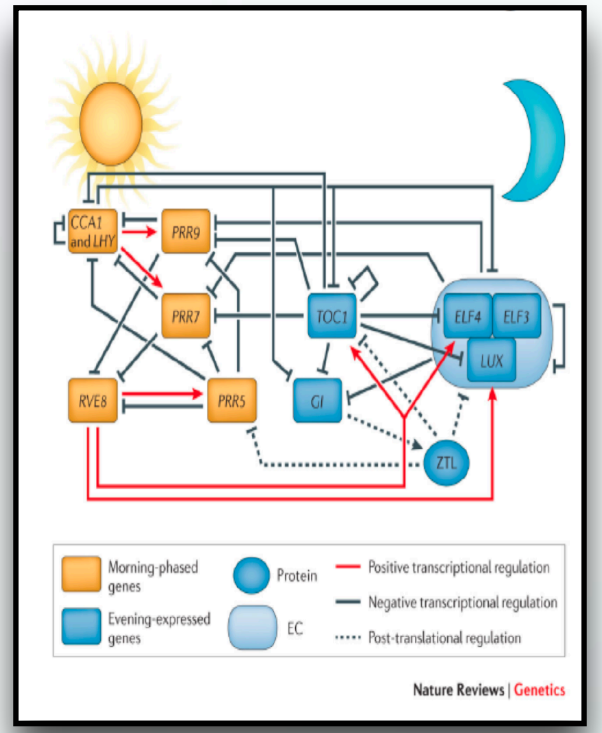




- CHARMM (Chemistry at HARvard Macromolecular Mechanics) payload (hybrid MPI/OpenMP/GPU) example built and executed on Titan
- Depending on the type of projects, payloads can expand beyond 500 nodes on Titan; currently, it uses 60-124 nodes for each project

Biology / Genomics

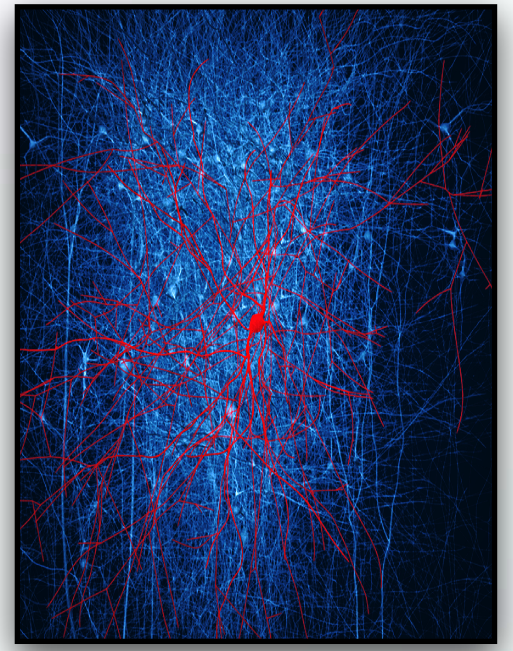
- In collaboration with Center for Bioenergy Innovation at ORNL, the PanDA based workflow for epistasis research was established. Epistasis is the phenomenon where the effect of one gene is dependent on the presence of one or more 'modifier genes', i.e. the genetic background.
- The GBOOST application, a GPU-based tool for detecting gene-gene interactions in genome-wide case control studies, was built and tested on Titan with PanDA as an example
- Input data were located in a set of eight input directories of 152 M each. Every PanDA job was configured to process single input directory in backfill mode on one node and walltime of 30 min. The output data of about 11M per job was located to the corresponding output dir



- Genetic regulatory networks
- Phenotypes, or traits, are determined by genetic regulatory networks
- These genetic networks are composed of genes that are organized to coordinate overt phenotypes
- In quantitative genetic analyses, genes underlying the basis of traits are formalized

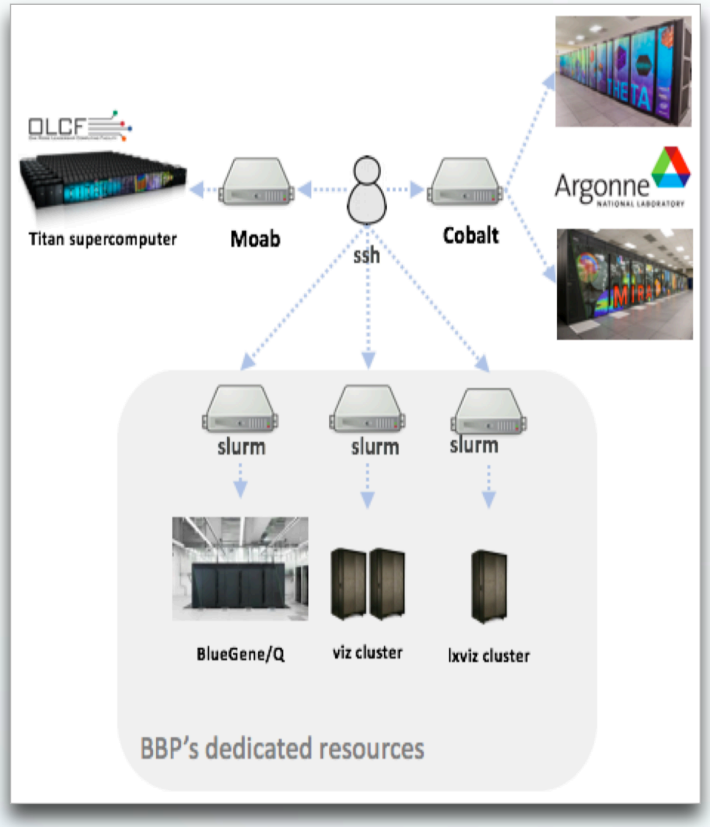
Blue Brain Project

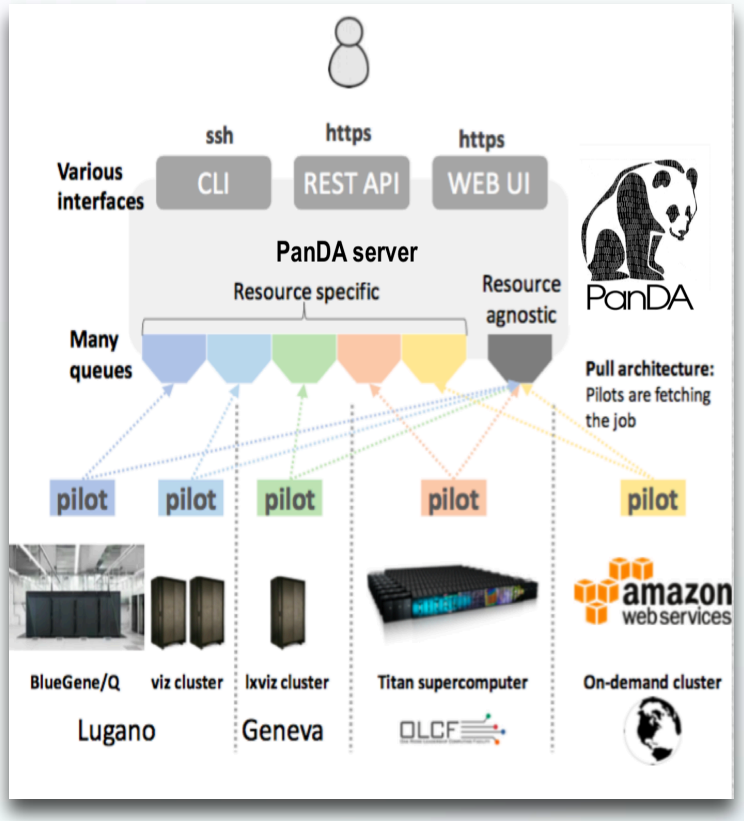
- The goal of the Blue Brain Project is to build biologically detailed digital reconstructions and simulations of the rodent, and ultimately the human, brain
- The supercomputer-based reconstructions and simulations built by the project offer a radically new approach for understanding the multilevel structure and function of the brain
- The novel research strategy allows the project to build computer models of the brain at an unprecedented level of biological detail
- Supercomputer-based simulation turns understanding the brain into a tractable problem, providing a new tool to study the complex interactions within different levels of brain organization and to investigate the cross-level links leading from genes to cognition



The presynaptic neurons of a layer 2/3 nest basket cell (in red) were stained (in blue) in a digital reconstruction of neocortical microcircuitry. Only immediate neighbouring presynaptic neurons are shown.

Blue Brain Project computing model evolution





- Dedicated PanDA instance on VM hosted by BBP
  - Improved PanDA User Interface: from CLI to Web interface
  - Integration with BB authentication system
- "Proof of the concept" phase was successfully finished
  - The test jobs were successfully submitted to the targeted resources via PanDA portal.
  - The project demonstrated that the software tools and methods for processing large volumes of experimental data, which have been developed initially for experiments at the LHC, can be successfully applied to BBP.

Summary

- Over the past 5 years, Titan has delivered the DOE Leadership Computing Mission at ORNL: delivering science and constraining energy consumption.

- ATLAS/PanDA deployment on Titan shows the potential of distributed high-throughput computing to be integrated with high-performance computing infrastructure.
  - Offers significant value for other projects beyond HEP

- BigPanDA project :
  - demonstrates a possibility to improve integration with heterogeneous computing resources
  - extend automation of processing of significant amount of data
  - possibility of integration with workflow management systems for orchestration of complex data processing on supercomputers for broad applications across scientific domains

Acknowledgements

- Thanks to members of the BigPanDA and PanDA projects, and colleagues from the ATLAS experiment at the LHC

- Special thanks to our colleagues from diverse sctientific groups :

  IceCube (G. Merino, V. Brik), LSST/DESC (A. Slosar, E. Sheldon, A. Nomerotski, H. Park), nEDM (L. Broussard, J. Leggett), LQCD (S. Mukherjee, R. Larsen), Genetics (D. Jacobson, M. B. Shah), BBP (A. Beche, F. Delalondre, F. Schuermann), Molecular Dynamics (K. Nam) ...

- This work was funded in part by the U. S. Department of Energy, Office of Science, High Energy Physics and Advanced Scientific Computing under Contracts No. DE-AC02-98CH10886 and DE-AC02-06CH11357.

- The BlueBrain/MegaPanDA pilot project was funded in part by the Russian Ministry of High Education and science under contract No. 14.Z50.31.0024

- This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE- AC05-00OR22725