ATLAS Software and Computing Effort at BNL

Alexei Klimentov

NPPS group meeting

Jun 12, 2019

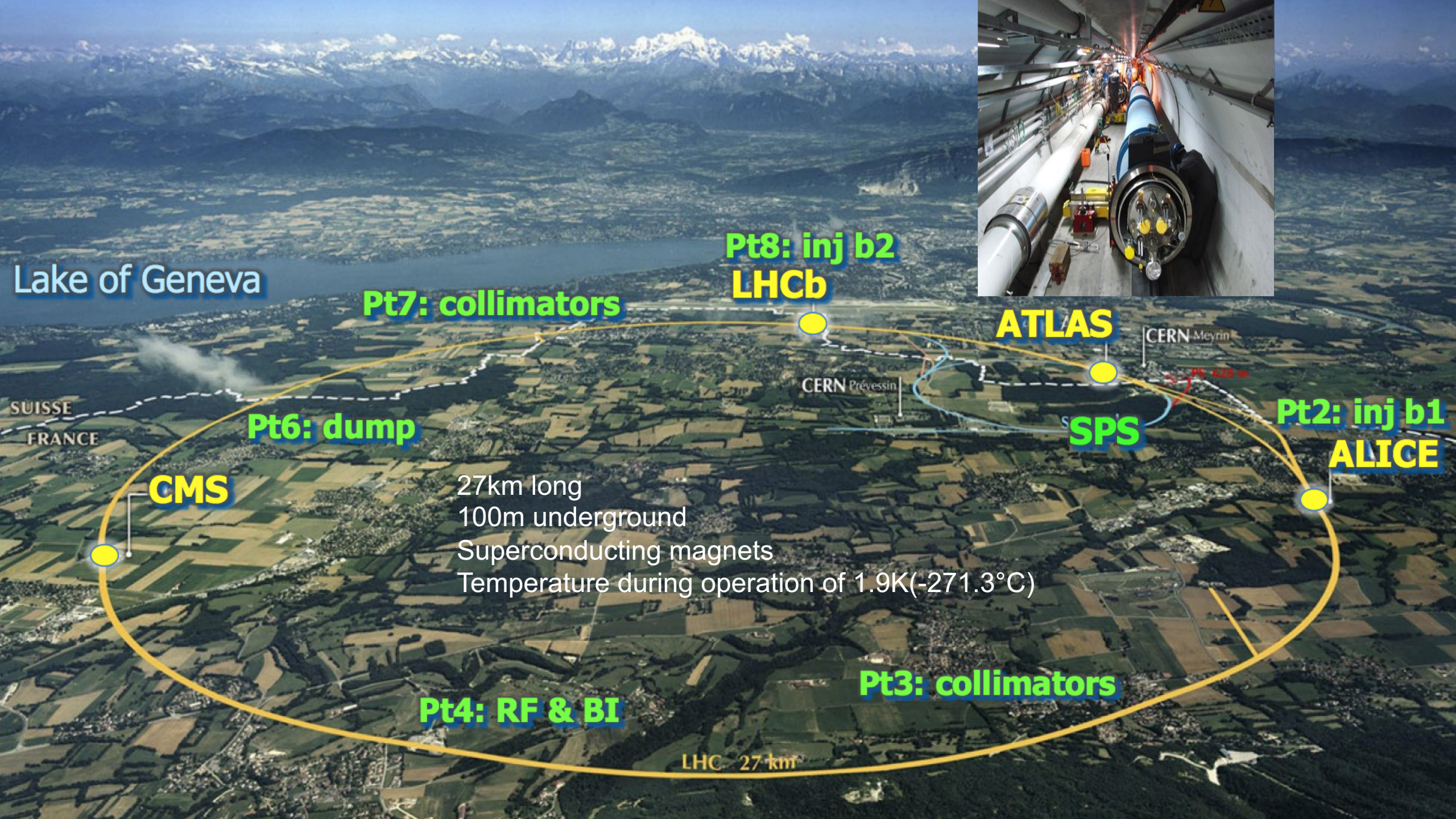
BROOKHAVEN NATIONAL LABORATORY

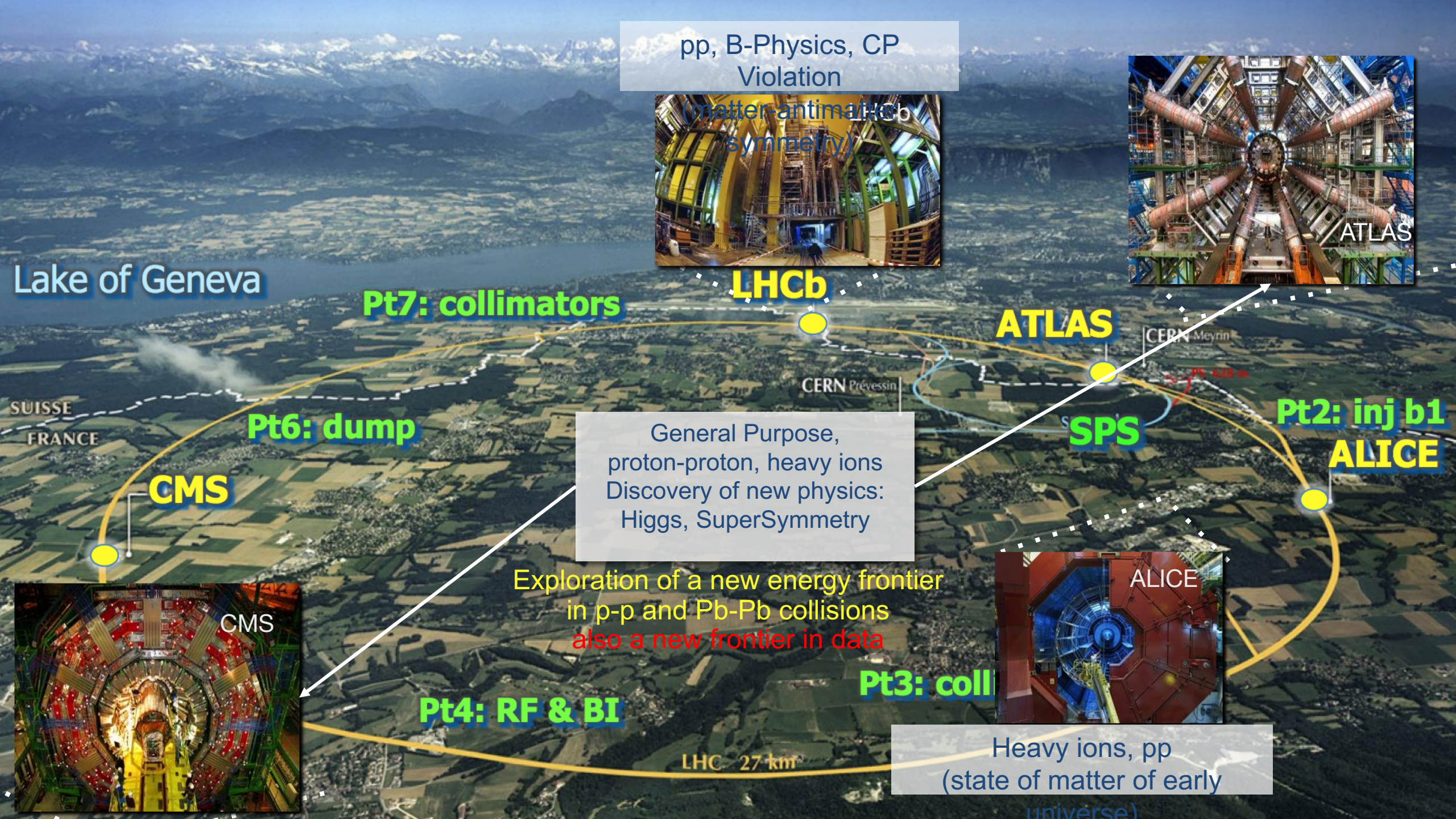
U.S. DEPARTMENT OF ENERGY

BROOKHAVEN SCIENCE ASSOCIATES

- Thanks to many BNL and US ATLAS Colleagues for slides, materials and comments

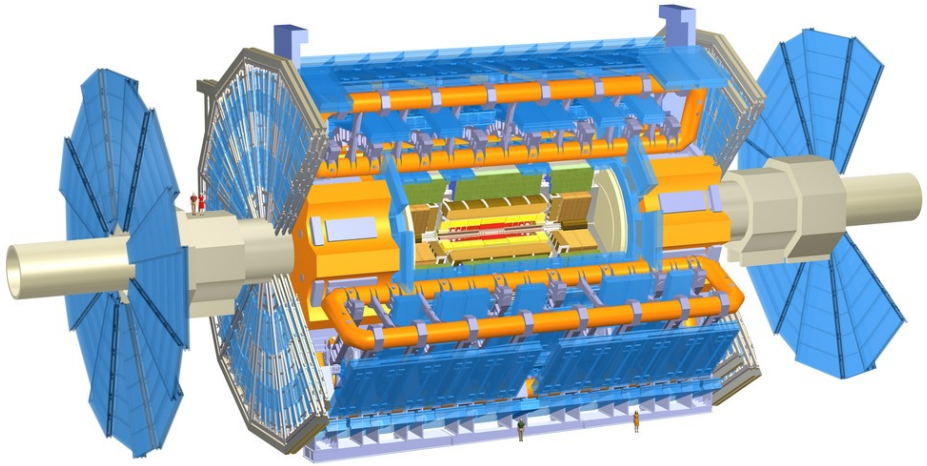
- Caveat.
  - It is an overview talk, more technicalities have been presented and will be presented by group members
  - It is primarily about effort in NPPS. BNL Tier-1 is the biggest ATLAS tier center, there is also a strong SW effort in Omega group

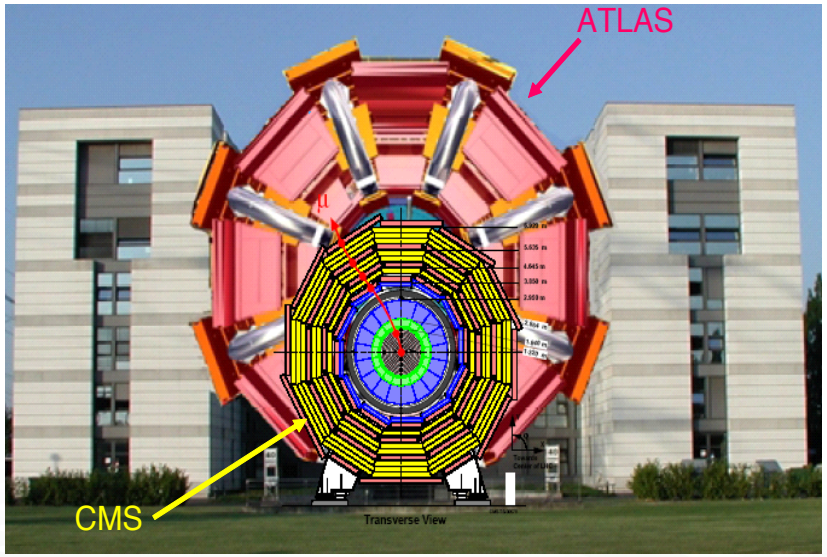

Lake of Geneva
Pt8: inj b2
LHCb
Pt7: collimators
ATLAS
CERN Meyrin
CERN Prévessin
SUISSE
FRANCE
Pt6: dump
SPS
Pt2: inj b1
ALICE
CMS
27km long
100m underground
Superconducting magnets
Temperature during operation of 1.9K(-271.3°C)
Pt3: collimators
Pt4: RF & BI
LHC 27 km


pp, B-Physics, CP Violation
(matter-antimatter symmetry)
LHCb
ATLAS
Lake of Geneva
Pt7: collimators
LHCb
ATLAS
CERN Meyrin
CERN Prévessin
SUISSE
FRANCE
Pt6: dump
SPS
Pt2: inj b1
ALICE
CMS
General Purpose, proton-proton, heavy ions
Discovery of new physics: Higgs, SuperSymmetry
Exploration of a new energy frontier in p-p and Pb-Pb collisions
also a new frontier in data
ALICE
Pt3: coll
Pt4: RF & BI
LHC 27 km
Heavy ions, pp
(state of matter of early universe)

The ATLAS Experiment at the LHC

3000 scientists
174 Universities and
Labs from 38 countries
More than 1200 students





The Nobel Prize in Physics 2013
François Englert, Peter Higgs

The Nobel Prize in Physics 2013

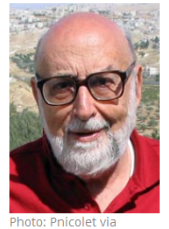


Photo: Pnicolet via Wikimedia Commons
François Englert

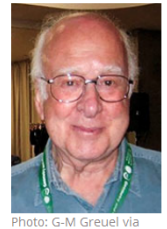


Photo: G-M Greuel via Wikimedia Commons
Peter W. Higgs

The Nobel Prize in Physics 2013 was awarded jointly to François Englert and Peter W. Higgs *"for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at s Large Hadron Collider"*

ATLAS has 44 meters long and 25 meters in diameter, weighs about 7,000 tons. It is about half as big as the Notre Dame Cathedral in Paris and weighs the same as the Eiffel Tower or a hundred 747 jets

U.S. DEPARTMENT OF ENERGY

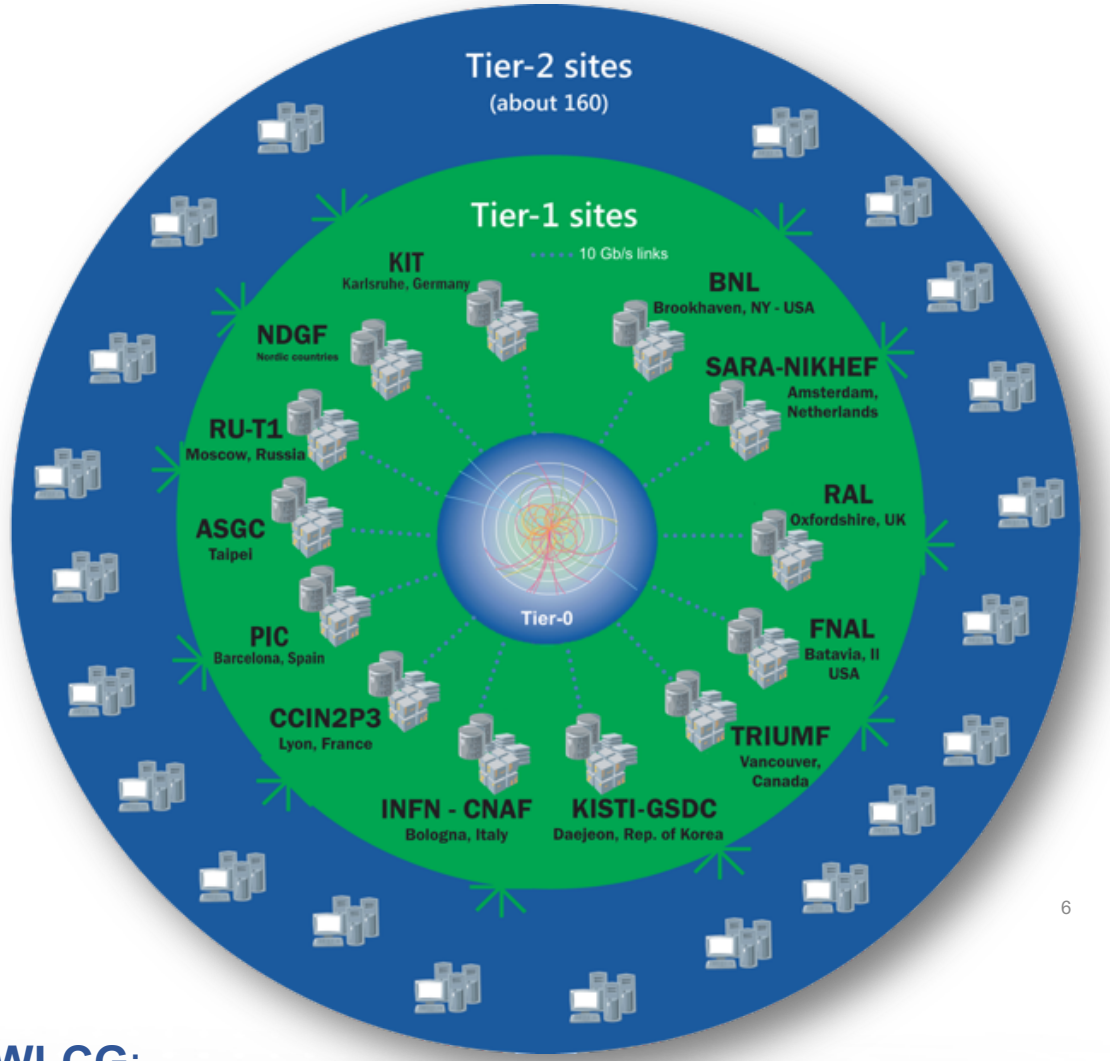
BROOKHAVEN NATIONAL LABORATORY

The Worldwide LHC Computing Grid

Tier-0
(CERN and *Hungary*):
data recording, reconstruction and distribution

Tier-1: permanent storage, re-processing, Analysis
T0 spill-over
HLT
MC Simulation
Derivation production

Tier-2: Simulation, end-user analysis
Re-processing
Derivation production



~170 sites, 42 countries

~750k CPU cores

~1 EB of storage

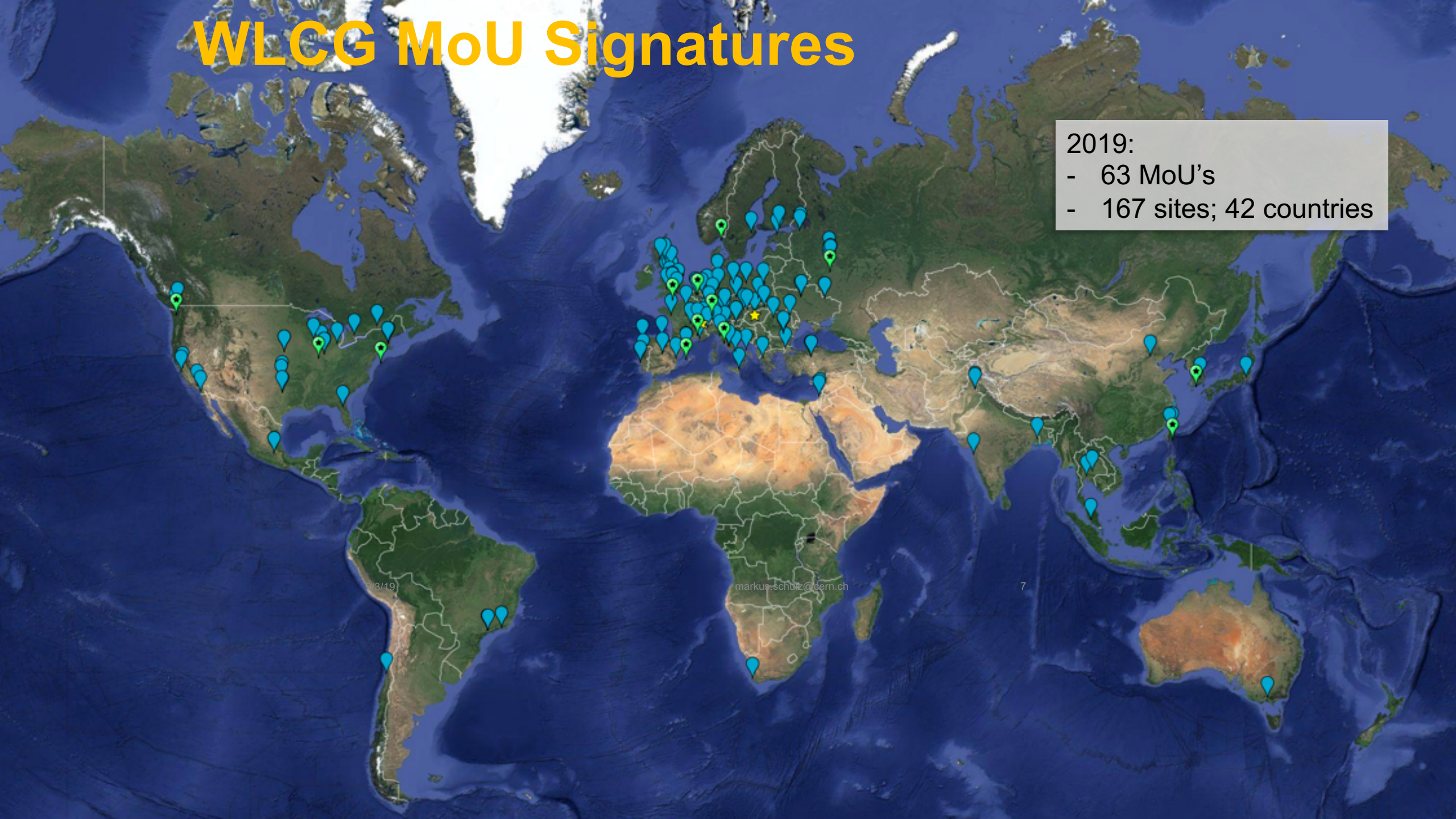
> 2 million jobs/day

10-100 Gb links

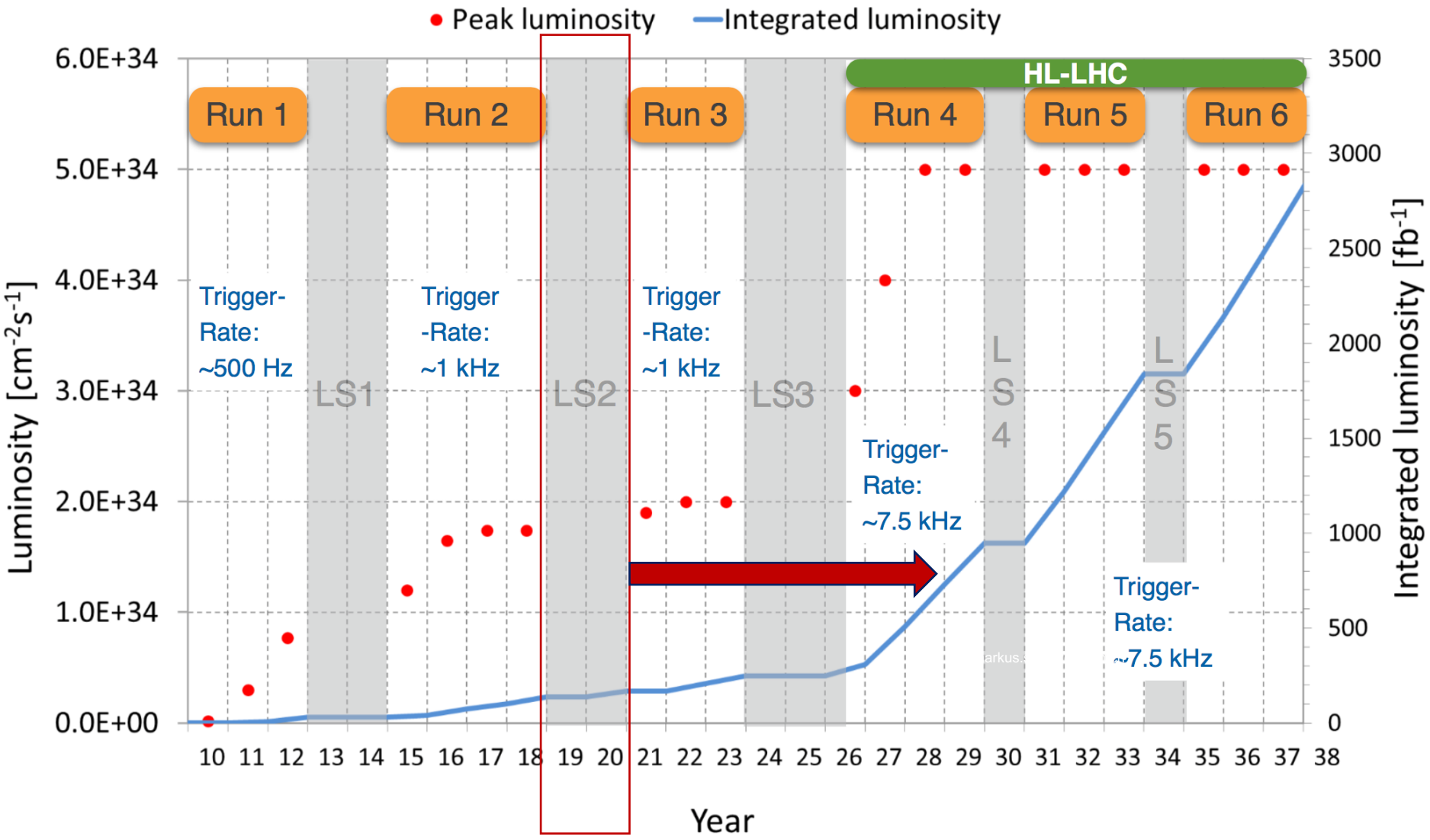
WLCG:
An International collaboration to distribute and analyse LHC data

Integrates computer centres worldwide that provide computing and storage resource into a single infrastructure accessible by all LHC physicists

WLCG MoU Signatures

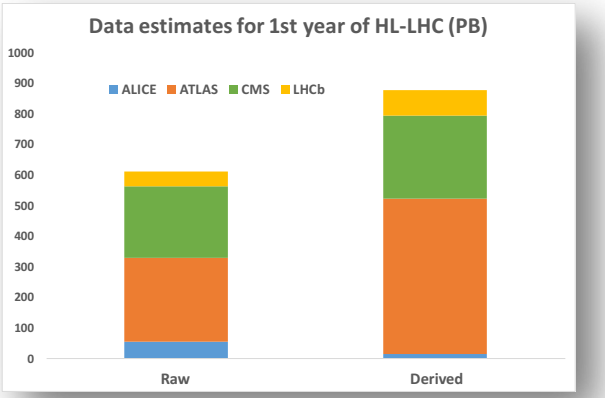


LHC Schedule



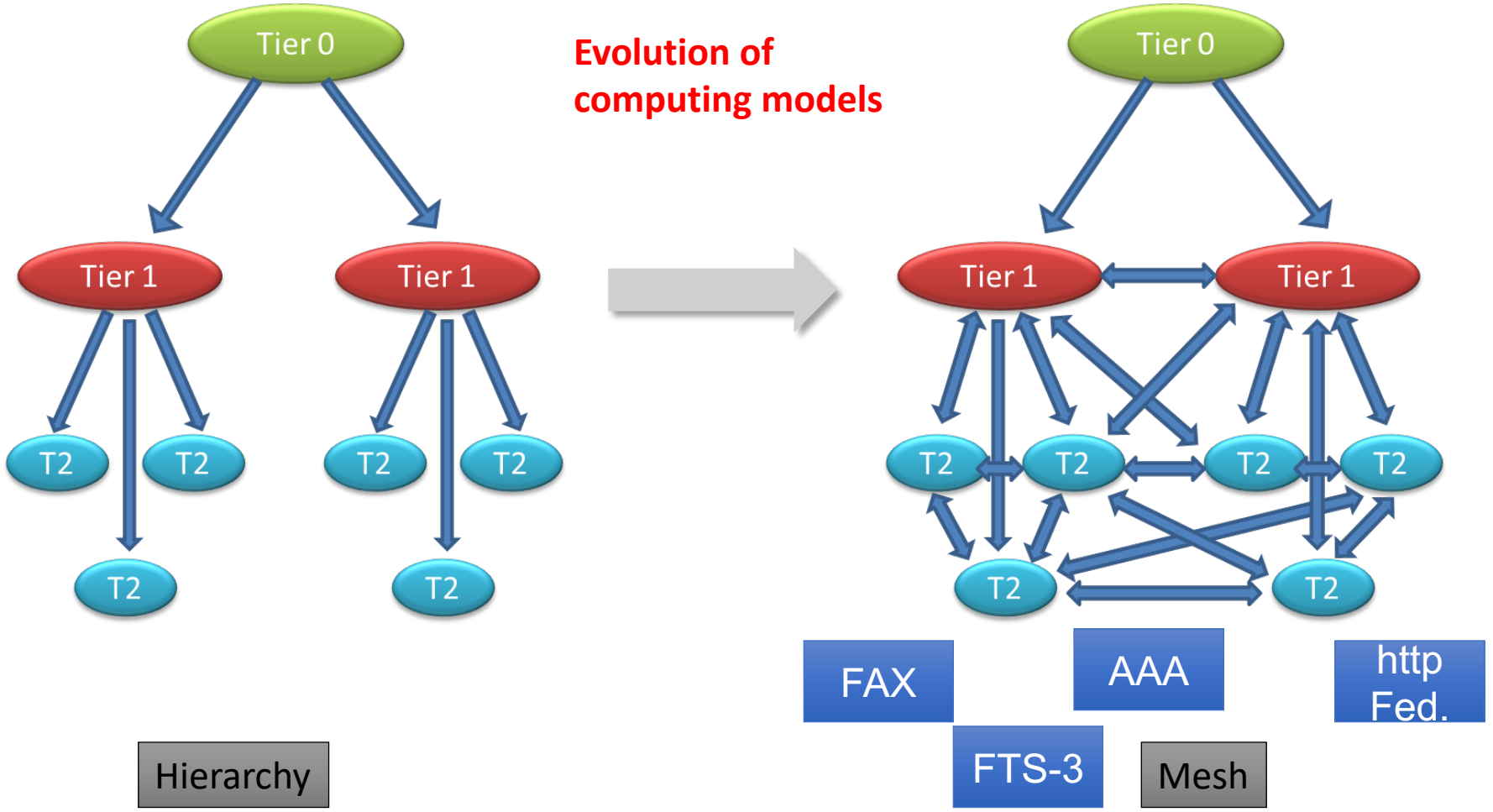
Run 3 ALICE, LHCb upgrades

Run 4 ATLAS, CMS upgrades



Data:
- Raw → 2027: 600 PB
- Derived (1 copy): → 2027: 900 PB

Computing model evolution

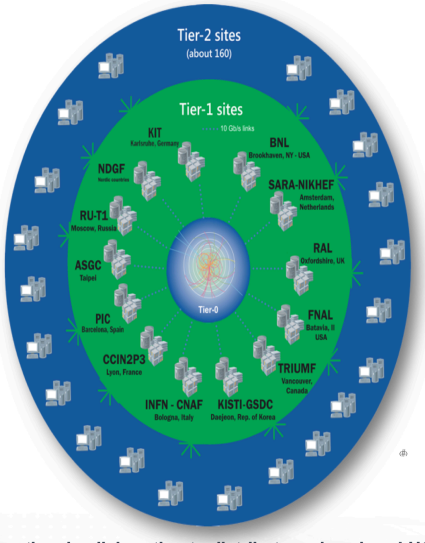


The Worldwide LHC Computing Grid

Tier-0
(CERN and Hungary):
data recording,
reconstruction and
distribution

Tier-1: permanent
storage, re-processing,
analysis

Tier-2: Simulation,
end-user analysis



~170 sites,
42 countries

~750k CPU cores

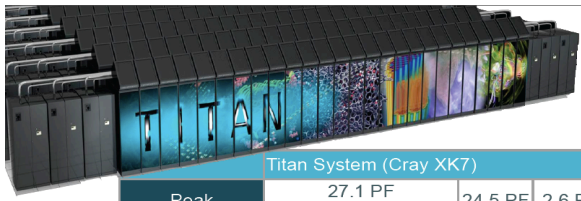
~1 EB of storage

> 2 million jobs/day

10-100 Gb links

WLCG:
An International collaboration to distribute and analyse LHC data

Integrates computer centres worldwide that provide computing and storage resource into a single infrastructure accessible by all LHC physicists



Titan System (Cray XK7)			
Peak Performance	27.1 PF 18,688 compute nodes	24.5 PF GPU	2.6 PF CPU
System memory	710 TB total memory		
Interconnect	Gemini High Speed Interconnect	3D Torus	
Storage	Lustre Filesystem	32 PB	
Archive	High-Performance Storage System (HPSS)	29 PB	
I/O Nodes	512 Service and I/O nodes		

12 OLCF | 20

ATLAS Grid would be around #30 from Top100

nectar

Google Actual Cloud Platform

The ATLAS collaboration have members with access to these machines and to many others…

amazon web services™ | EC2

U.S. DEPARTMENT OF ENERGY

BROOKHAVEN NATIONAL LABORATORY

Paradigm shift in HEP Computing

Old paradigms	New ideas
• Distributed resources are **independent entities**	• Distributed resources are seamlessly **integrated worldwide** through a single submission system • **Hide middleware** while supporting diversity
• Groups of users utilize **specific resources** (whether locally or remotely)	• All users have access to **same resources**
• Fair shares, priorities and policies are managed **locally**, for each resource	• **Global** fair share, priorities and policies allow efficient management of resources
• **Uneven user experience** at different sites, based on local support and experience	• **Automation, error handling, and other features** improve user experience • Central support coordination
• Privileged users have access to **special resources**	• All users have access to **same resources**

Orchestrators

ATLAS central components at CERN

Workflow Management: "translates" physicist requests into production tasks

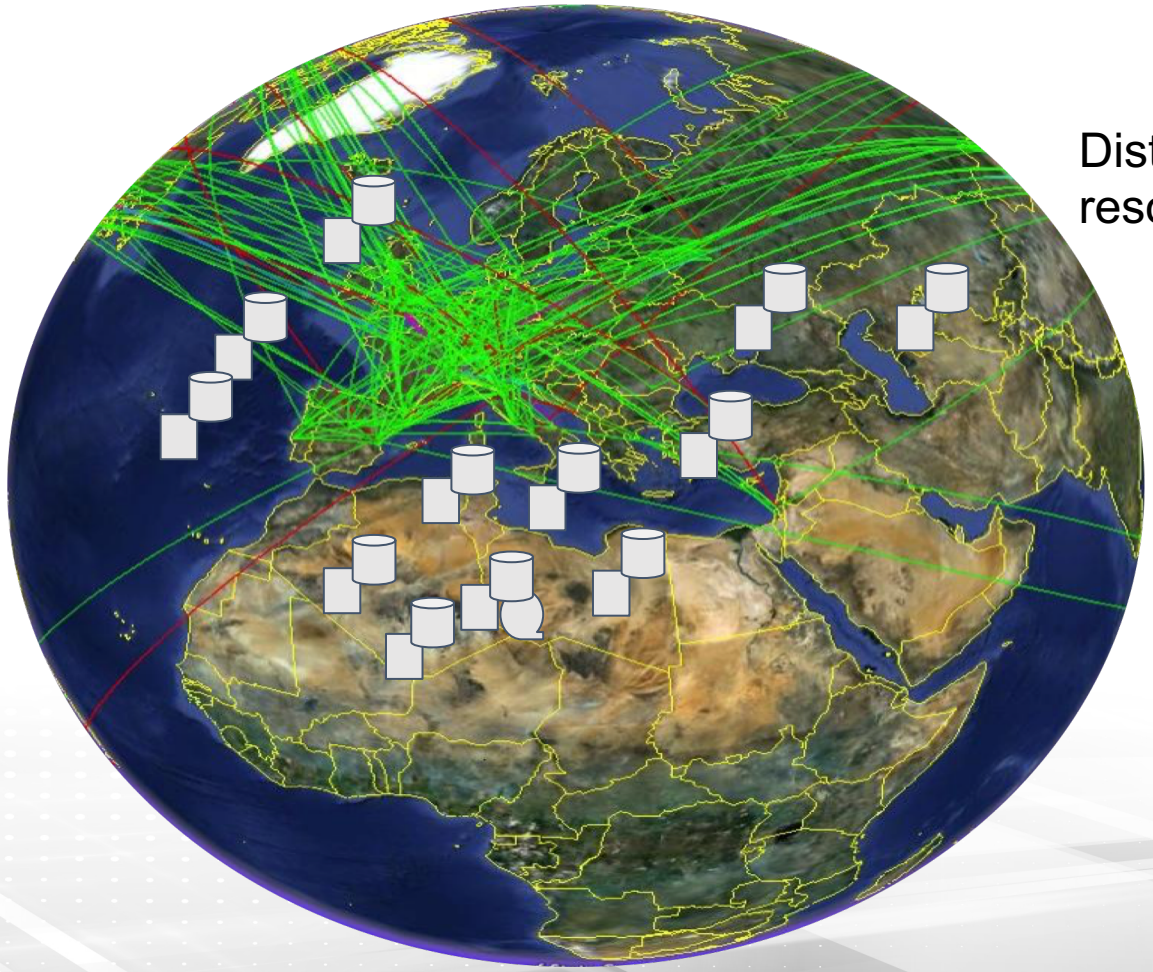
Workload Management: submission and scheduling of jobs & tasks

Information System (AGIS) PanDA queues and resources description

PanDA

SCIENTIFIC DATA MANAGEMENT RUCIO

Data Management: bookkeeping and distribution of files & datasets

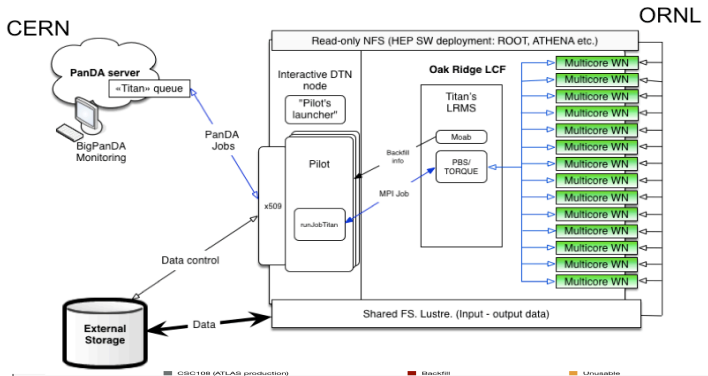


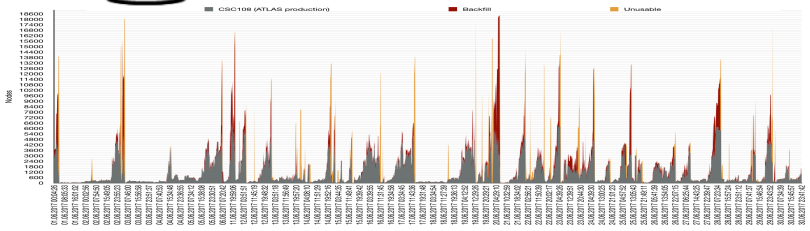
Distributed resources

Out of four principal ATLAS distributed software systems, three came from BNL team

Workload Management. PanDA. Production and Distributed Analysis System

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



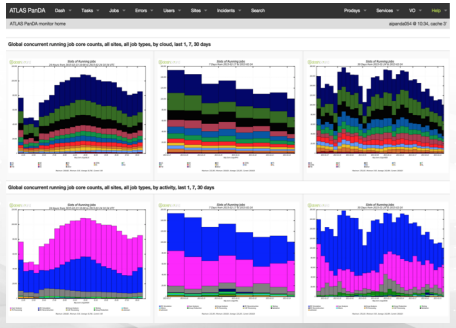


BigPanDA Monitor
http://bigpanda.cern.ch/

Global ATLAS operations

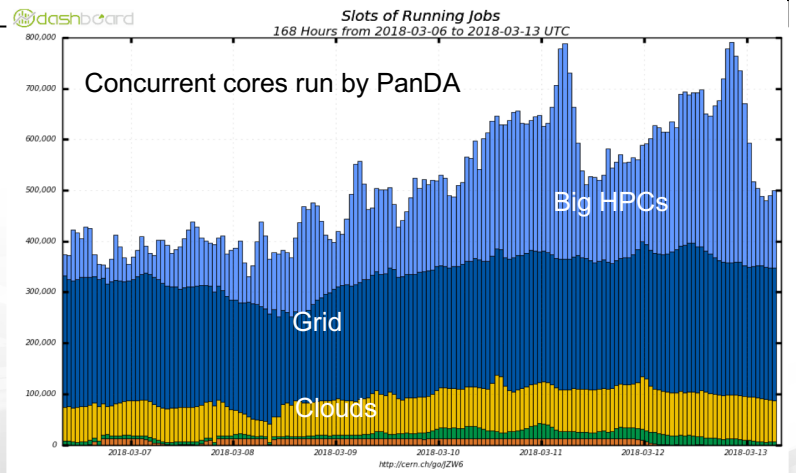
Up to ~800k concurrent jobs
25-30M jobs/month at >250 sites
~1400 ATLAS users

First exascale workload manager in HENP
1.4+ Exabytes processed yearly in 2014 -2018
Exascale scientific data processing today



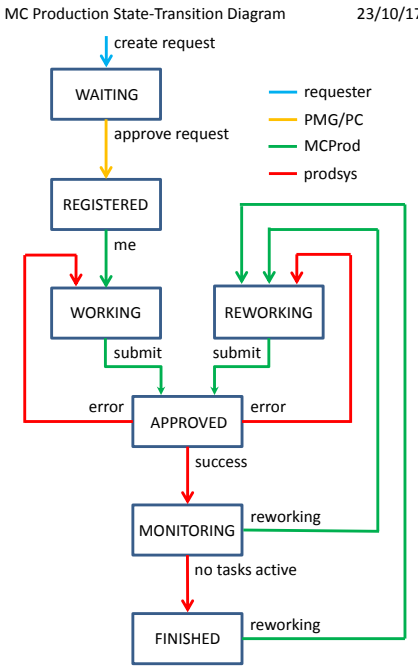
PanDA Brief Story

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
2012-14: 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
2016-19: DOE ASCR BigPanDA@Titan project
2016:PanDA for bioinformatics
2017-2018:COMPASS adopted PanDA , NICA (JINR)
PanDA beyond HEP : BlueBrain, IceCube, LQCD
2018 : Harvester : PanDA edge service



U.S. DEPARTMENT OF ENERGY

ATLAS Workflow and Workload Management

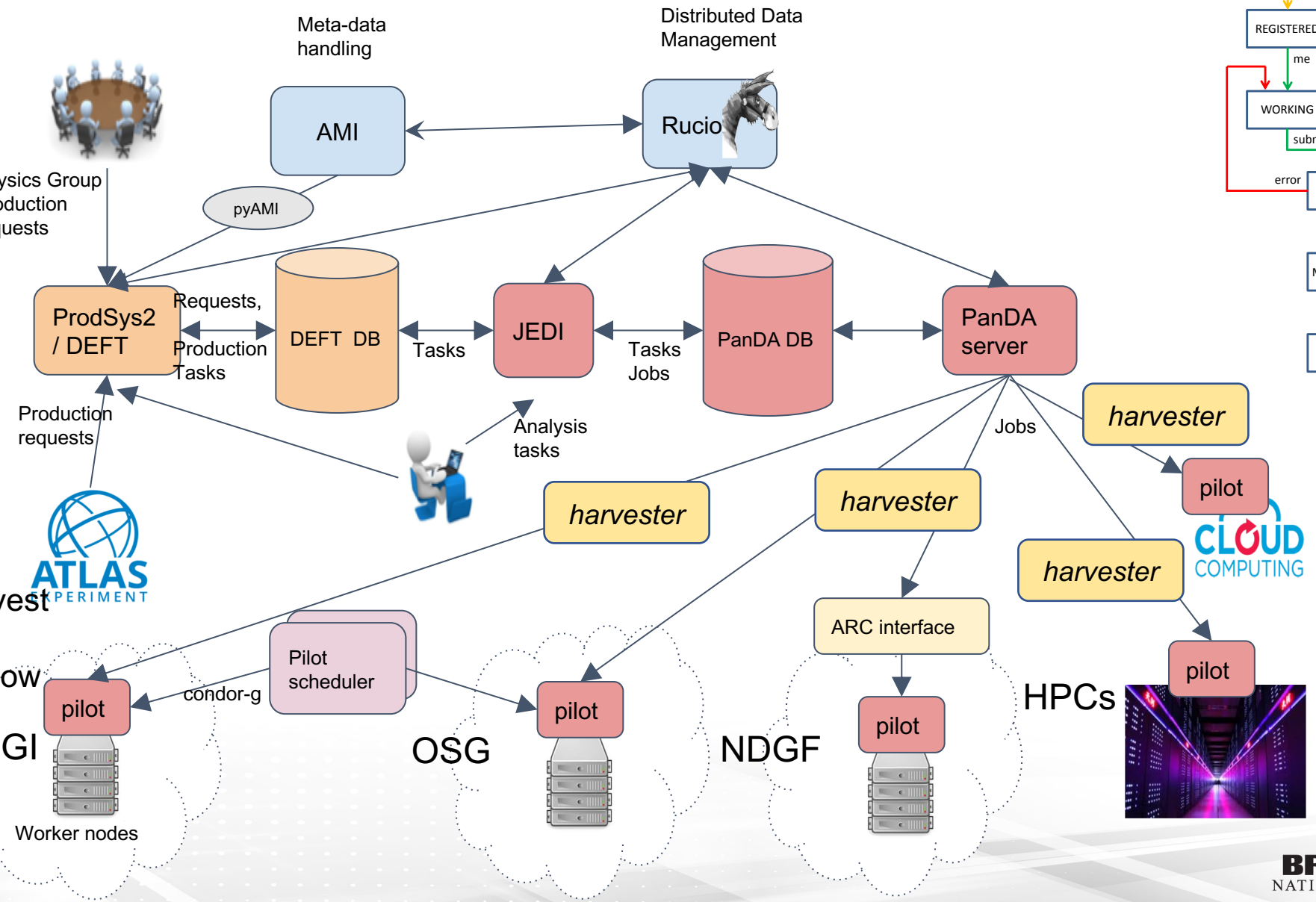


Orchestrate all ATLAS Workflows :

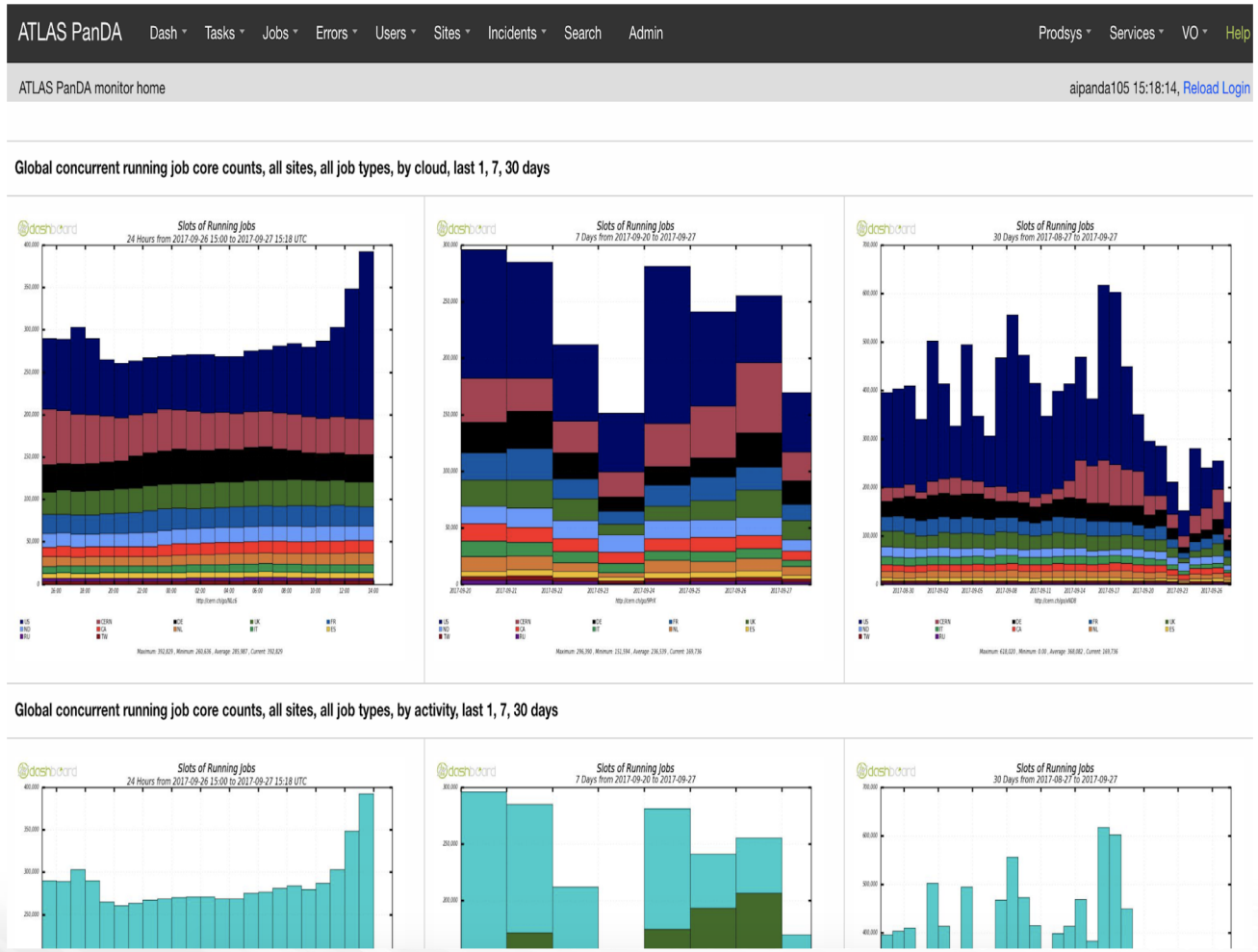
- MC Production
- Physics Groups WF
- Data reprocessing
- T0 spill-over
- HLT processing
- SW validation
- User's analysis

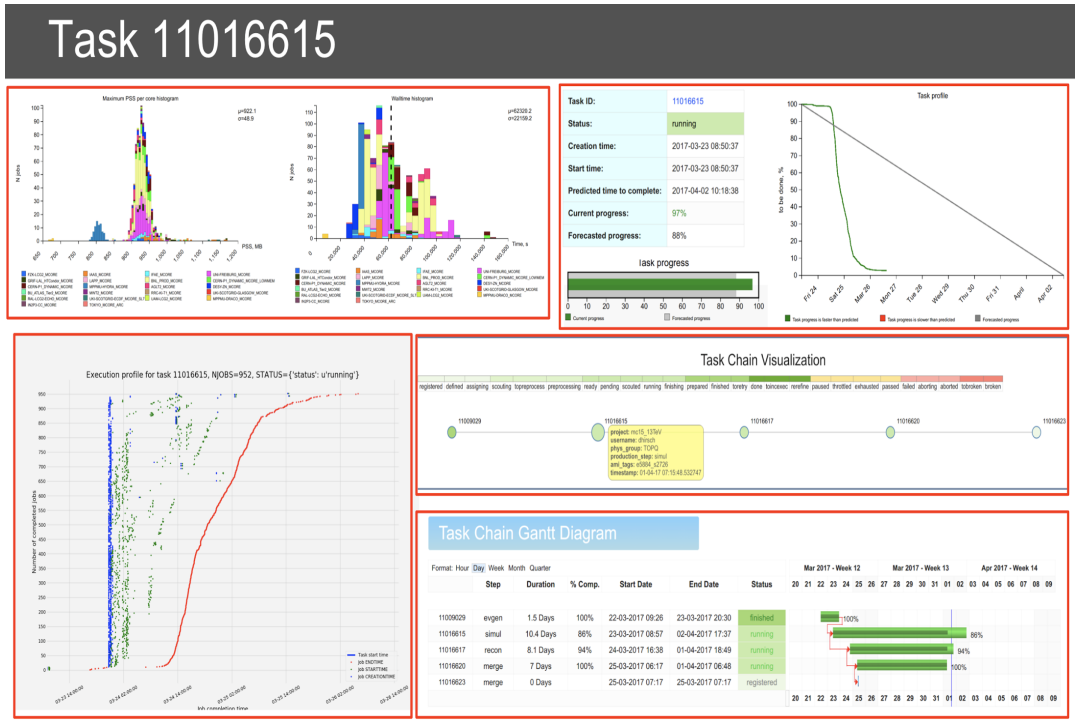
Support ATLAS rich harvest of resources

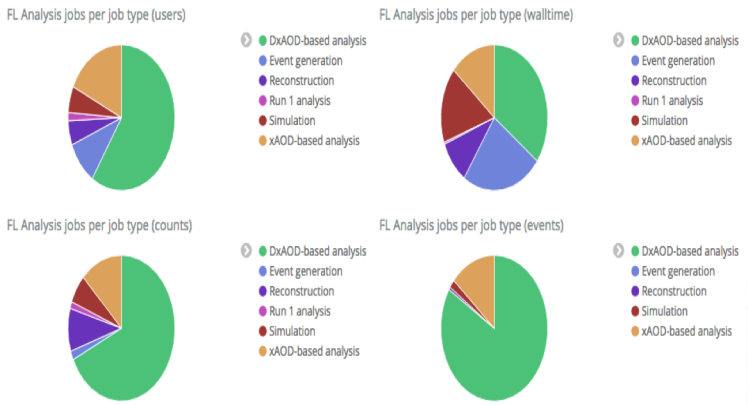
Integrate WF and data flow



Monitoring and Analytics (bigpanda.cern.ch)







WMS Summary and Lessons learned

- ***We designed and implemented a scalable, flexible, automated production that follows physics priorities***
  - Steady state production 24x7x365 with ~300-350k cores across ~140 sites
  - HPC peaks to >1M cores, demonstrating extreme scalability of PanDA
  - PanDA and Prodsys orchestrate ~10 principal workflows and dozens of variants, with automated shares that follow ATLAS physics priorities and allocate work across global resources
  - Also supporting over 1000 analysis users with fair sharing of resources
- ***Integrated workflow and dataflow***
  - Moving >1 PB, >20 GB/s, 1.5-2M files per day
  - 405PB disk+tape, 1+B files in total (and ~540PB in 2019)
  - **PanDA processes over 1.5 Exabytes per year**
- WMS is designed by and serves the physics community
- WMS new features are driven by experiment operational needs. WMS functionality is important as scalability
- Computing model and computing landscape in general has changed

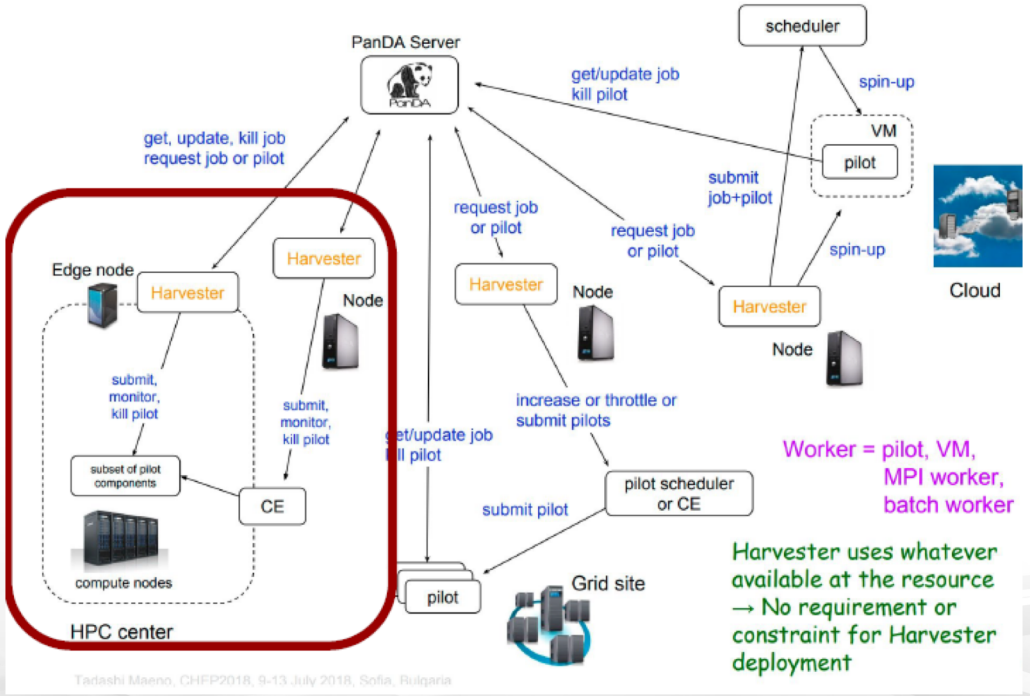
There are several systems with very well defined roles which are integrated for distributed computing : Information system (AGIS), DDM (Rucio), WMS (ProdSys2/PanDA), meta-data (AMI), and middleware (HTCondor, Globus…). We managed to have a good integration of all of them in ATLAS.

Recent Accomplishments. HPCs and Distributed Computing

Bringing HPCs to production has required a distributed computing revolution

Wide range of technologies and policies.
Defined unified resource manager to

- Deliver software and data, retrieve results, report status
- Assign resources to workflows and shape workflows to resources
  - **Harvester** manages > 95+% of ATLAS resources.



Recent Accomplishments. Harvester

Harvester on Titan beyond ATLAS

USQCD

Quantum chromodynamics (QCD) is the component of the Standard Model of elementary particle physics that governs the strong interactions. It describes how quarks and gluons, the fundamental entities of strongly interacting matter, are bound together to form strongly interacting particles, such as protons and neutrons, and it determines how these particles in turn interact to form atomic nuclei.

nEDM

The goal of the nEDM experiment at the Fundamental Neutron Physics Beamline at the Spallation Neutron Source (ORNL) is to further improve the precision measurement of neutron properties by a factor of 100 to search for violations of fundamental symmetries and to make critical tests of the validity of the Standard Model of electroweak interactions

LSST Large Synoptic Survey Telescope

The goal of the Large Synoptic Survey Telescope project is to conduct a 10-year survey of the sky that will address some of the most pressing questions about the structure and evolution of the universe and the objects in it:

- Understanding Dark Matter and Dark Energy
- Hazardous Asteroids and the Remote Solar System
- The Transient Optical Sky
- The Formation and Structure of the Milky Way

PMI

Molecular Dynamics: simulations of enzyme catalysis, conformational change, and ligand binding/release in collaboration with research group from University of Texas at Arllington.

cbi THE CENTER FOR BIOENERGY INNOVATION

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.

IceCube SOUTH POLE NEUTRINO OBSERVATORY

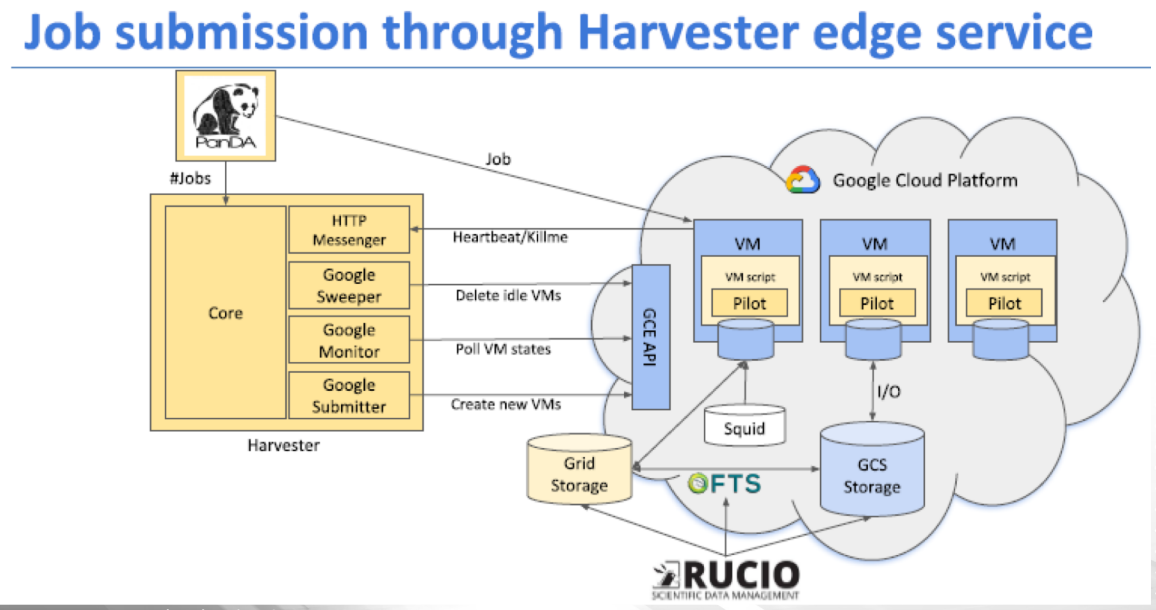
IceCube collaborators address several big questions in physics, like the nature of dark matter and the properties of the neutrino itself. IceCube also observes cosmic rays that interact with the Earth's atmosphere, which have revealed fascinating structures that are not presently understood.

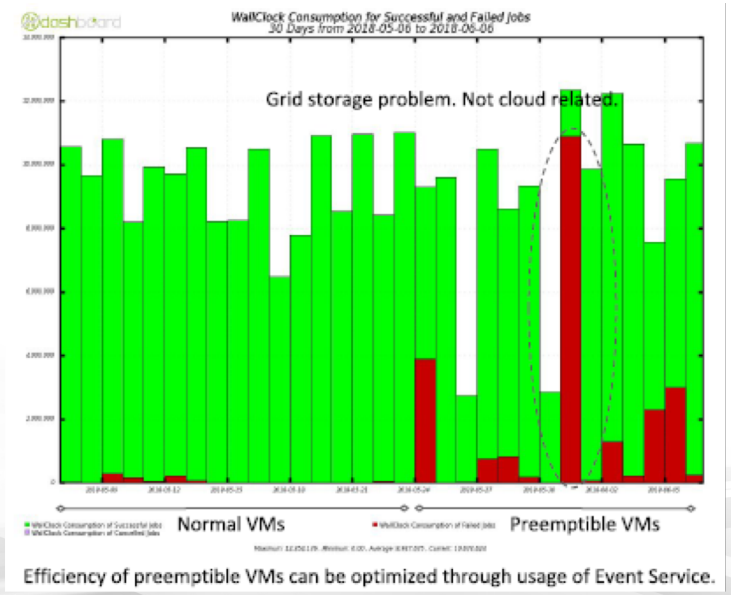
Recent Accomplishments. Data Access and Management Highlights

- BNL-led ATLAS Analysis Model Study Group Run 3 Goals (J.Elmsheuser)
  - 30% less disk storage in Run 3: O(60PB)/year
    - AMSG achieved similar (20%) savings for Run 2.
  - Provide directions for further savings @ HL-LHC
    - Achieved through painstaking analysis of data format utilization (at the variable-by-variable level), of duplication across streams, and impact on physics analysis
      - interacting with analysis groups key part of AMSG work
  - AMSG R3 proposing to introduce **50KB/event DAOD_PHYS**

Recent Accomplishments. Collaboration with Google

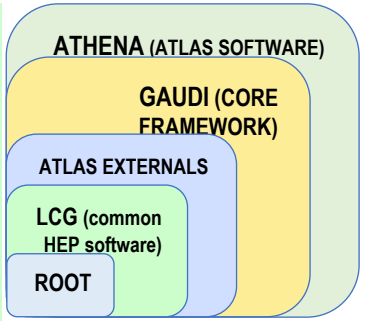
- ATLAS keeps multiple (expensive) copies of data for worldwide distributed analysis - R&D to use Google Storage
- Proof of concept project focused on analysis usage
  - Using Google storage transparently from ATLAS PanDA
  - Tested operating a 120 core Google cluster as PanDA resource
    - Successful with CPU at Google and data at CERN
    - Testing access of Google storage from ATLAS Tier 1 & Tier 2 sites





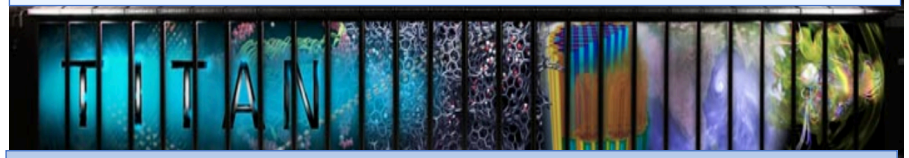
Recent Accomplishments. ATLAS SW installation from source code on HPC

All-inclusive installation from source code, including generators (Geant4, Pythia…), ROOT, LCG stack



- **Full automation feasible:** code upload via **HTTP** (**no CVMFS**)

Friendly Linux, AMD CPUs
(ATLAS kits binaries work)



PowerPC, 10X of Titan
IBM CPUs, GNU Linux
(ATLAS kits binaries do **not** work)

DETAILS

- **5M code lines of ATLAS software release**
- **100 external packages**
- **130 generator packages**
- **Total compilation time: 1 day**
- **Few code adjustments needed (e.g. compiler macro)**

- Major ATLAS production release was installed on Summit LCF. Validation is in progress
- Procedure works on Titan LCF
- Plan to automate it as much as possible

ATLAS SW&Computing effort at BNL

- Total ATLAS NPPS ~8+ FTE
- Core expertise in offline software and databases
  - Athena framework core expertise including its multiprocessing and multithreading variants
  - Deep expertise on the C++ architecture of Athena and C++ itself (S.Snyder, *D.Adams*)
  - BNL develops ROOT I/O for ATLAS and works with the ROOT team on I/O issues (M.Nowak)
- Leading roles in ATLAS distributed software and computing since its inception - . Elmsheuser, A. Klimentov, T. Maeno, P. Nilsson, S. Padolski, T. Wenaus, *R.Mashinistov, S.Panitkin, M.Potekhin*
  - PanDA workload management system manages all ATLAS distributed production and analysis
  - Prodsys production system translates physicist requests into PanDA production
  - Many innovations to grow the resources available to ATLAS (HPCs, clouds, fine grained processing)
- US ATLAS and ATLAS Software infrastructure support – A.Undrus, S.Ye
  - Long term support of ATLAS release build/test tools (~ 20000 Nightlies, CI, stable releases annually). Transitioned to modern open-source tools. Extending to new architectures (e.g. Summit)
- BNL is co-leading US ATLAS HL-LHC SW&Computing effort (T.Wenaus)
- BNL is co-leading ATLAS Distributed Computing effort (J.Elmsheuser)
- BNL is (co)leading many ATLAS distributed computing areas and projects (WFM SW, ProdSys/PanDA, harvester, pilot, HPC…)

Innovating SW&Computing for HL-LHC and R&D Projects

- *Many successful and pioneering R&Ds in the past*
  - Cloud Compuiting, HPC/HTC integration, Event Service, PanDA beyond ATLAS

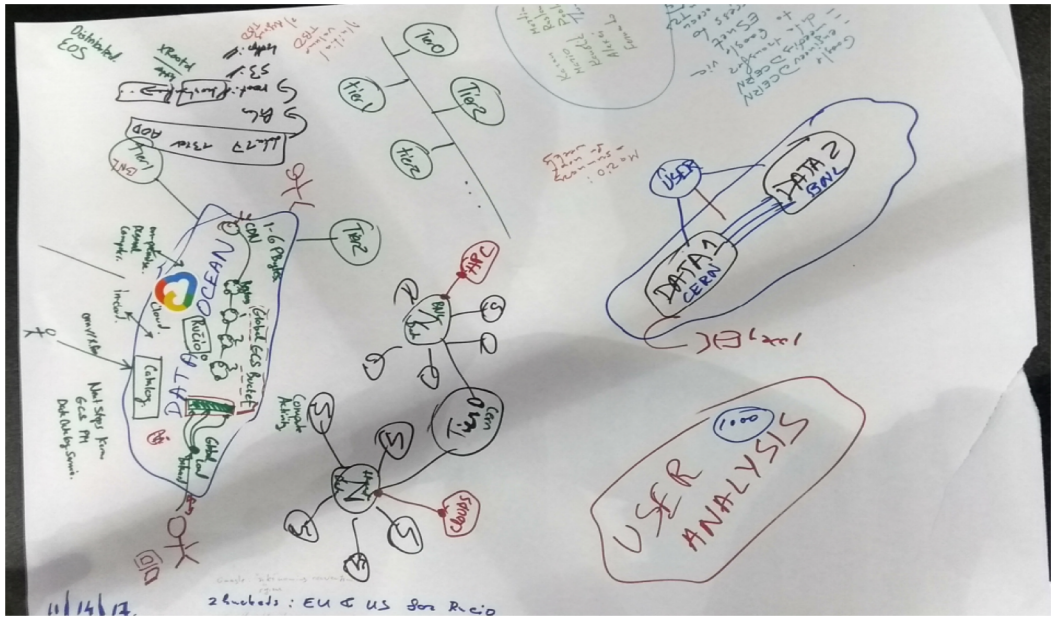
- Google Computing
- Data Carousel
- Community collaboration
  - HEP SW foundation, IRIS-HEP, WLCG DOMA = Data Organization, Management and Access, WLCG Operations Intelligence

- HL-LHC SW&C
  - New architectures and new workflows
  - Data streaming and intelligent data flow and control (ESS/iDDS)
    - Next step in the development of the ATLAS event service
    - Make full use of the network to economize storage
      - Send only the data the consuming client needs
    - Process data with WAN latency hiding to efficiently process data being streamed from far away

Google Computing. Motivation.

- IT landscape has changed dramatically since end of XX century
- US technology sector is recognized as world leaders
  - Amazon, Google, Microsoft, Oracle, … - already play significant role in worldwide scientific computing
- LHC data intensive computing challenges are (and have been) at the cutting edge of technology development
- Foster partnerships with US industries in research and development – and not just as late stage product adopters
- The huge challenges at the HL-LHC have spurred new efforts in US ATLAS to collaborate with technology partners
- Traditionally, US ATLAS Ops program did not support R&D with private sector – we are starting a new front in LHC R&D, with companies willing to invest in open source solutions

(US) ATLAS Google Collaboration.



ATLAS & Google — "Data Ocean" R&D Project, ATLAS note ATL-SOFT-PUB-2017-002 https://cds.cern.ch/record/2299146/, 29 Dec 2017

US ATLAS institutions and the Google Cloud team started collaborating at SuperComputing 17
- at SC17: Google, BNL, UTA, LBNL, & ORNL drafted plans for a demo
  - Rucio and PanDA teams made plugins to access Google cloud
  - Google provided cloud credit for testing prototypes
- Results were presented at NEXT 2018 and CHEP 2018 by ATLAS, and at CERN meetings by Google.
- "Proof of Concept" success has led to expanded work plan
  - Geared towards HL-LHC, leveraging Google expertise
- Expanded technical teams, both within ATLAS and Google
  - Five areas of collaboration identified so far, which are in various stages from planning to active technical work

(US) ATLAS Google Collaboration

- Collaborative research activities are organized along five separate research tracks, each group working independently
- While this collaboration was initiated by US ATLAS, there is now wide international interest and participation
  - CERN IT, CERN OpenLab, WLCG, Tokyo U, UK, EU institutions...
  - Google has now joined CERN OpenLab
- US ATLAS and Google Reps will meet with DOE HEP Management on Jun 14th

Track 1	Data Management across Hot/Cold storage
Track 2	Machine learning and quantum computing
Track 3	Optimized I/O and data formats
Track 4	Worldwide distributed analysis
Track 5	Elastic computing for WLCG facilities

BNL will lead and coordinate it

Data Carousel R&D project

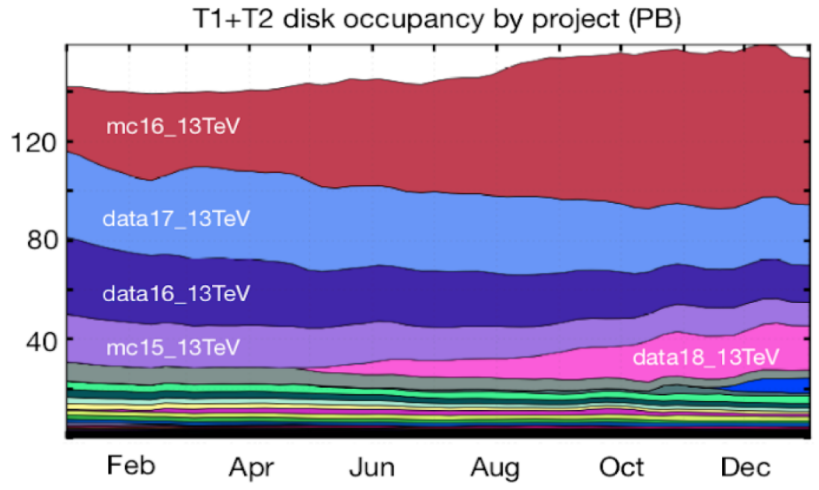
Reduce disk storage by running workflows from tape

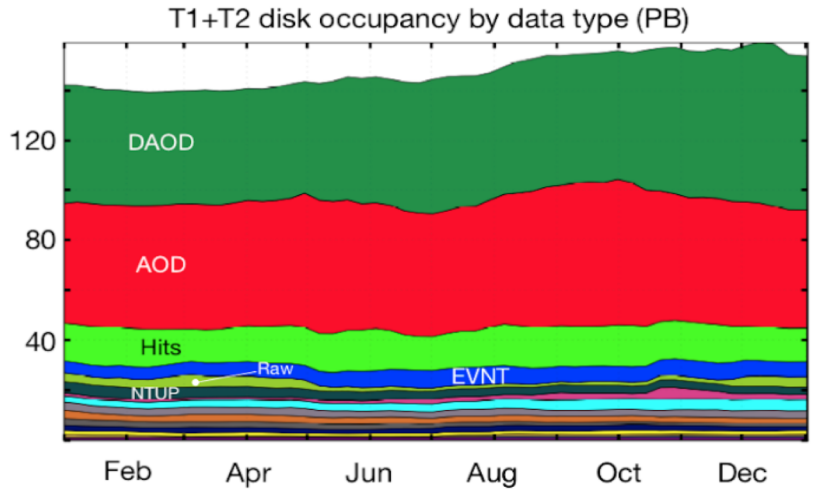
- Stage on disk a sliding window of e.g. 10% of an input dataset which is processed promptly
- Requires tight orchestration between workload management system, data management system, and tape services
  - *Coordinated by X.Zhao and AK*

Current status

- Completed first phase of tape system stress test, on all ATLAS tape sites
- Set up metrics and define ProdSys/DDM protocol
- Completed phase II round2 : run derivation production for realistic data sample, with ProdSys/DDM integration.
  - Now we are in preparation to phase II round 3, which requires a deeper integration of the workload and data management systems. We will also introduce shares and priorities at this stage.

2018





High-Lumi LHC Computing

Strategy : “High-risk High-reward ” R&D in FY19-21 (LS2)

- Goal is to reduce risk that evolutionary approach will impact HL-LHC physics reach (particularly for precision physics).

Task
Implement ATLAS framework support for offloading algorithms/tasks to GPU. Interface ML models to ATLAS framework. Support data science tools
Develop FastChain to run efficiently on LCF machines, including exascale
Implementation of multi-threaded MC Reconstruction workflow on LCF class machines
Develop ML based analytics tools for PanDA monitoring, supporting optimal distributed workload and data magement, including integration with Elastic Search analytics tool.
Get a generator running on next-gen HPCs, e.g. Sherpa or Madgraph
New workflows integrating DDM and WFM like data streaming, intelligent caching and use of hierarchical storage (e.g. data carousel); authentication/authorization
Total new effort : ~7 FTE