




BigData and Computing Challenges in High Energy and Nuclear Physics

Alexei Klimentov

CREMLIN WP2 Workshop on BigData Management

Moscow,

Feb 15-16, 2017

Outline

- High Energy Physics and Nuclear Physics (HENP) scale of needs
- BigData Technologies Laboratory at NRC “Kurchatov Institute” (BigData Lab)
  - Russian Ministry of Science and Education mega-grant
- BigData at HENP and computing needs, challenges, evolution and the Laboratory Research Program highlights
  - Workflow and data management
  - Federated data storage
  - Data Knowledge Base
- Summary

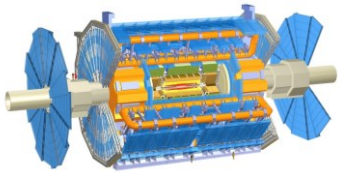
Disclaimer : This talk will have a “slight” bias towards ATLAS experiment @ LHC

The Science Drivers for Particle Physics

Five intertwined science drivers, compelling lines of inquiry that show great promise for discovery :

1. *Use the Higgs boson as a new tool for discovery.*
2. Pursue the physics associated with neutrino mass.
3. *Identify the new physics of dark matter.*
4. Understand cosmic acceleration : dark energy and inflation.
5. *Explore the unknown : new particles, interactions, and physical principles.*

Introduction. The ATLAS detector at the Large Hadron Collider



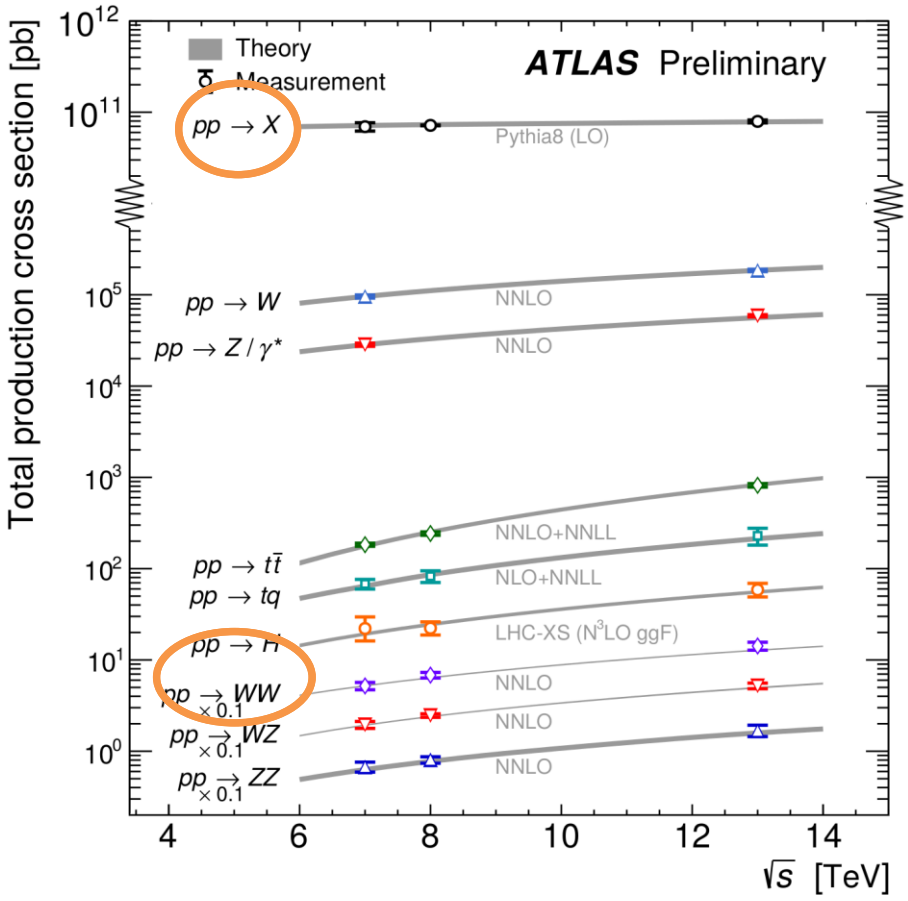
ATLAS at CERN LHC is a flagship experiment in the High Energy Physics with multiple science drivers:

- *Higgs Boson discovery in 2012. Nobel prize 2013 in Physics.*
- *Use the Higgs Boson as a new tool for discovery*
- *Identify the new physics of dark matter*
- *Explore the unknown: new particles, interactions and physical principles*

- Current pace of research and discovery is limited by ability of the ATLAS distributed computing facilities to generate Monte-Carlo events - "Grid luminosity limit" and to process *ALL* LHC data in quasi real-time mode
  - LHC experiments use Grid computing paradigm to organize distributed resources
  - Currently ~300K cores available to ATLAS Experiment worldwide
    - Still not enough CPU power !
      - Many physics simulation requests have to wait for months
  - Supercomputers are rich source of computing power
  - LHC experiments initiated R&D project aimed at integration of LeadershipClassFacilities and HPC resources (in general)

Since 2014 we started a project aimed at integration of ATLAS' PanDA Workload Management System with supercomputers. BigData Lab research is part of this effort – "BigPanDA"

Introduction. How Likely something interesting happen.



- Total Production Cross Section (== probability) vs Energy in pp collisions
- Notice the logarithmic scale on the Y-axis: it spans 11 orders of magnitude
- E.g. you produce 10 Higgs bosons out of 10^{11} billions of collisions
- The probability increases logarithmically with energy
- Theory (lines) agrees very well with measurements (markers)

Introduction. How Likely something interesting happen.

New physics rate ~ 0.00001 Hz

Event Selection :

1 in 10,000,000,000,000

Like looking for a single drop of water from the Geneve Jet d'Eau over 2+ days

RF Ministry of Science and Education mega-grant. BigData Technology Lab at "Kurchatov Institute". Motivation.

- NRC KI is the lead Russian Organization involved in research at the LHC, FAIR, XFEL, RHIC
- NRC-KI plays an important role in WLCG computing
  - *NRC-KI hosts Tier-1 center for 3 LHC experiments*
    - *ALICE, ATLAS,LHCb*
- ATLAS, ALICE are the *BigData* and megaScience Projects
  - megaScience projects are international
- Software technologies development (and software development in general) for such projects is difficult without international collaboration and cooperation
- The Laboratory scientific program is tightly coupled with Russian Fundamental Science priorities, NRC KI program, LHC experiments priorities and address challenges we will meet in 3-5 years
  - *And many challenges are not HENP or LHC specific*

There is a proven use case when software developed for the HENP experiment is expanded beyond HENP and expanded to other areas and disciplines

LHC Computing Needs

Data:
- **RAW 2016: 50 PB → 2027: 600 PB**
- **Derived (1 copy): 2016: 80 PB → 2027: 900 PB**

Run1
2009 - 2013

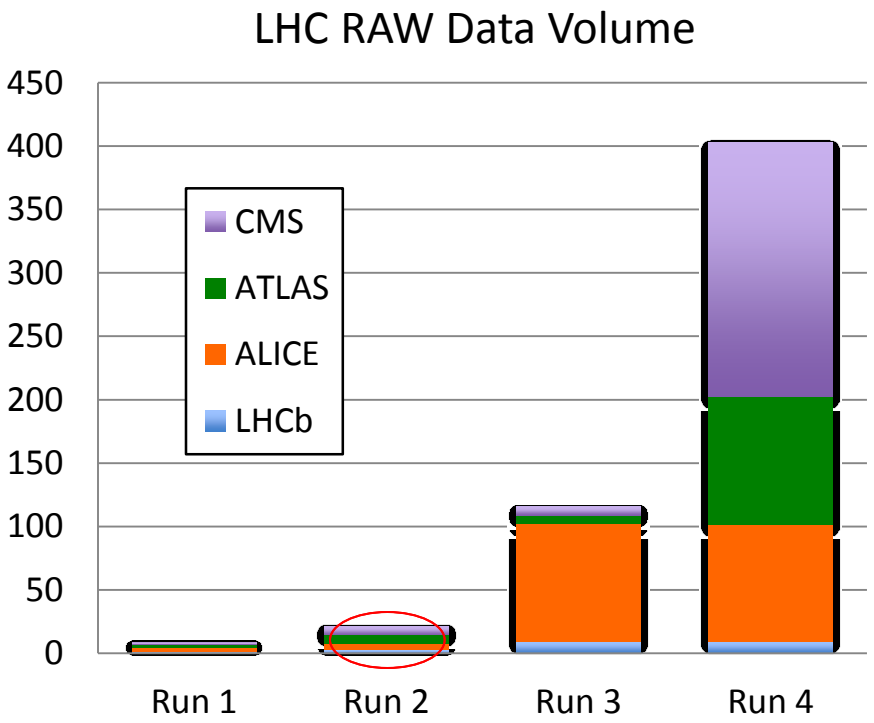
Run2
2015 - 2018

Run3
2020-2022
ALICE + LHCb

Run4
ATLAS + CMS

CPU:
- **x60 from 2016**

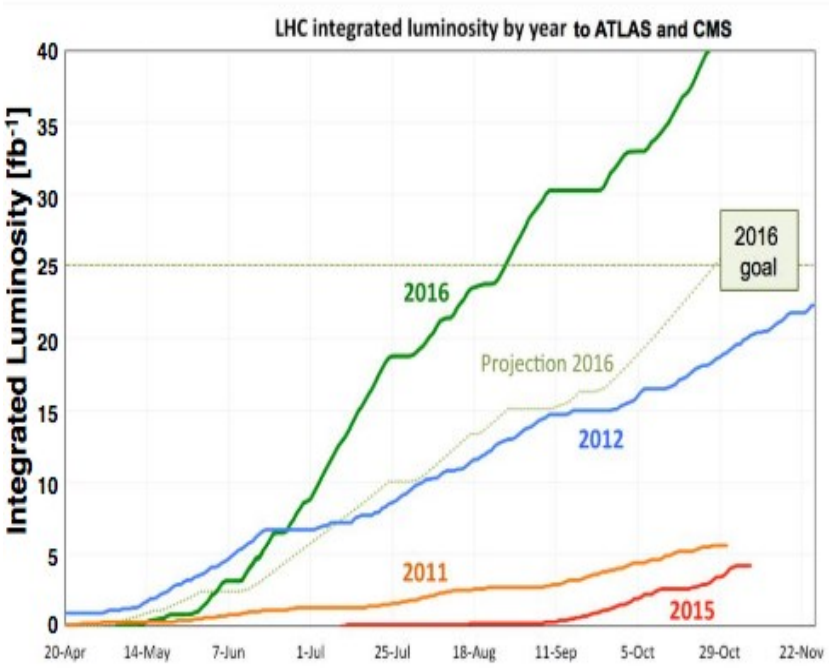
- CPU needs (per event) will grow with track multiplicity (pileup) and energy.
- Storage needs are proportional to accumulated luminosity.
- Grid resources are limited by funding and fully utilized.



Excellent LHC Performance and Immediate Computing Challenges

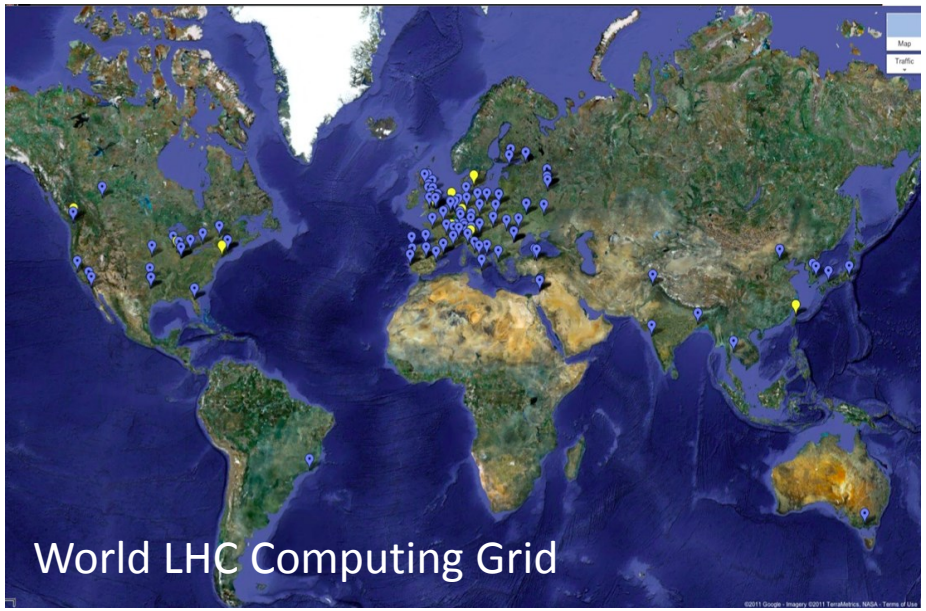
Excellent LHC Performance in 2016 (Run2)

- Unprecedented peak instantaneous luminosity > 40% beyond LHC design
- Data accumulation ~60% beyond 25 fb^{-1} goal for 2016
- High performance of the machine operation and data acquisition



Immediate challenge for the LHC Computing

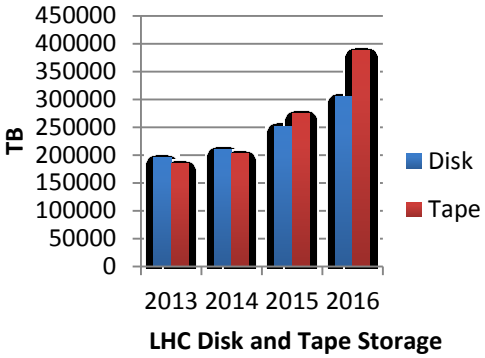
LHC, Amazon & Google Computing Centers. Relative size of things.



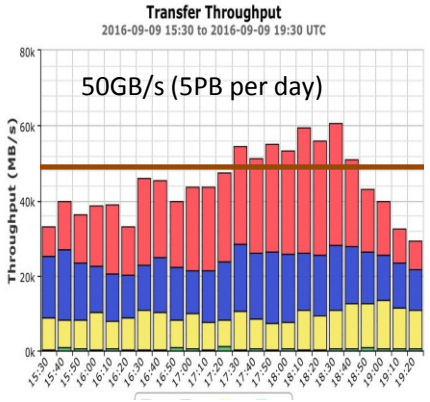
- One Google Data Center is estimated to cost ~$600M
  - An order of magnitude more than the centre at CERN
- Amazon : 9 large sites/zones
  - up to ~2M CPU cores/site, ~4M total
  - 10 x more cores on 1/10 of the sites compared to our Grid
  - 500,000 users
- LHC Computing (WLCG)
  - 167 sites, 42 countries
  - 500+k CPU cores total
  - Disk 350PB, Tape 400+PB
  - ~5000 users

Relative Size of Things. Cont'd

- Storage :
  - Amazon supports millions of queries per second
  - Google has 10-15 exabytes under management
  - Facebook 300PB
  - eBay collected and accessed the same amount of data as LHC Run1
- Processing :
  - Amazon has more than 40 million processor cores in EC2
  - Google has ~1M servers so ~20M cores

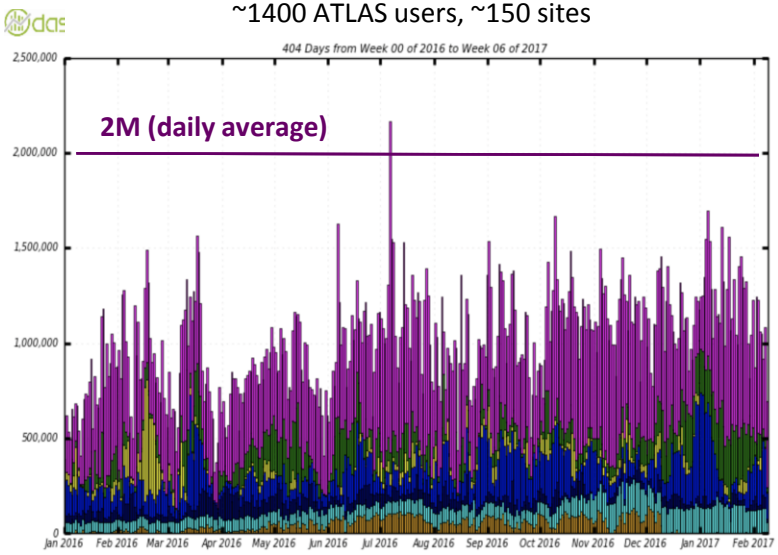


LHC Disk and Tape Storage



ATLAS Production System Performance. Daily Completed Jobs.

Up to ~250k concurrent jobs
25-30M jobs/month at >100 sites
~1400 ATLAS users, ~150 sites



HENP data and processing problems are about 1% the size of the largest industry problems, but we are still distribute more data and lead in the area of data and workflow management, and high-throughput computing in general.

HENP is good in distributed computing :

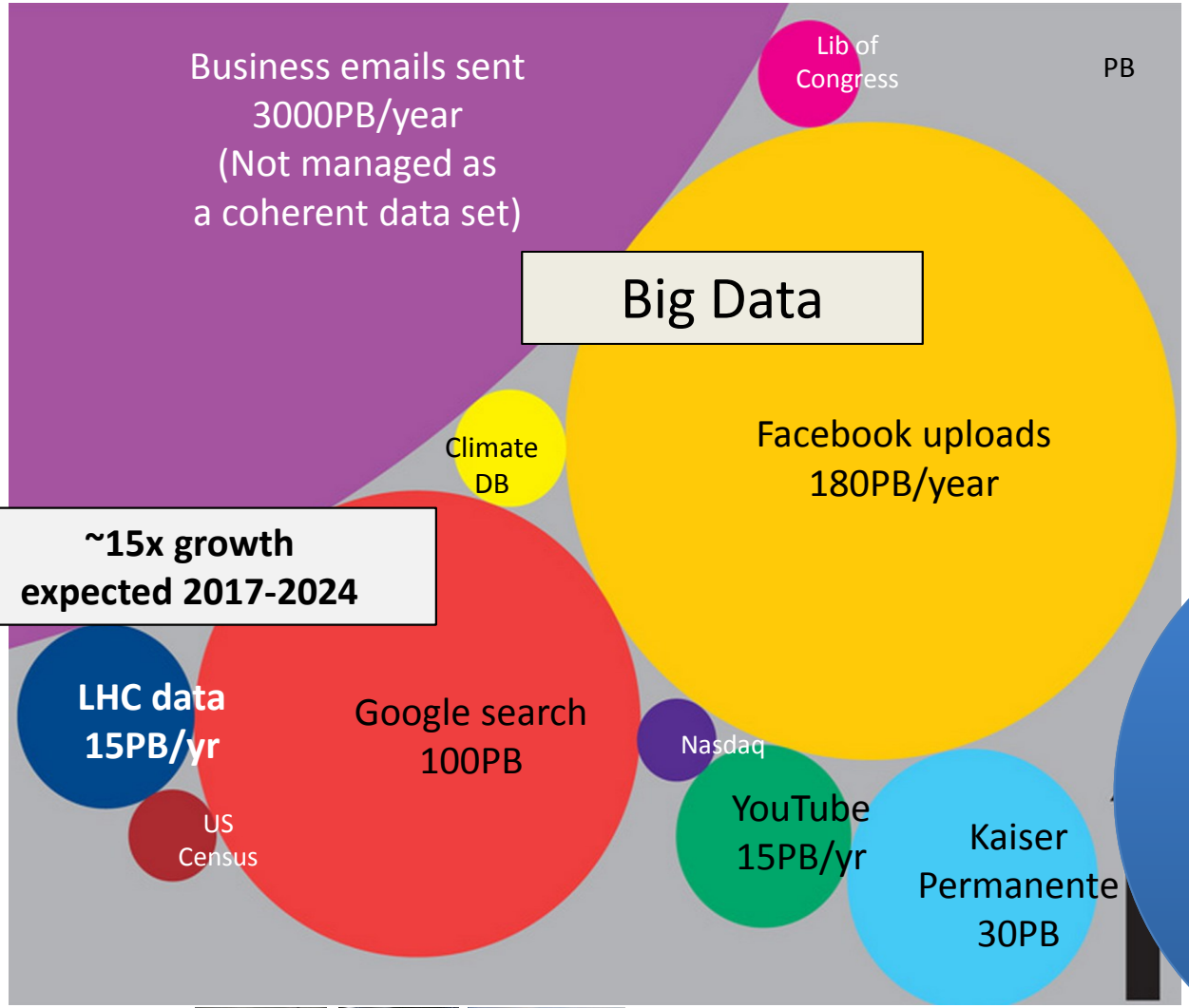
Datasets are large but custodially kept and protected

- We make dynamic use of tape systems
- We move to hundreds of sites
- We make effective use of global network links

We remain leaders in this challenging areas

Big Data: often just a buzz word, but not when it comes to HENP…

BigData?



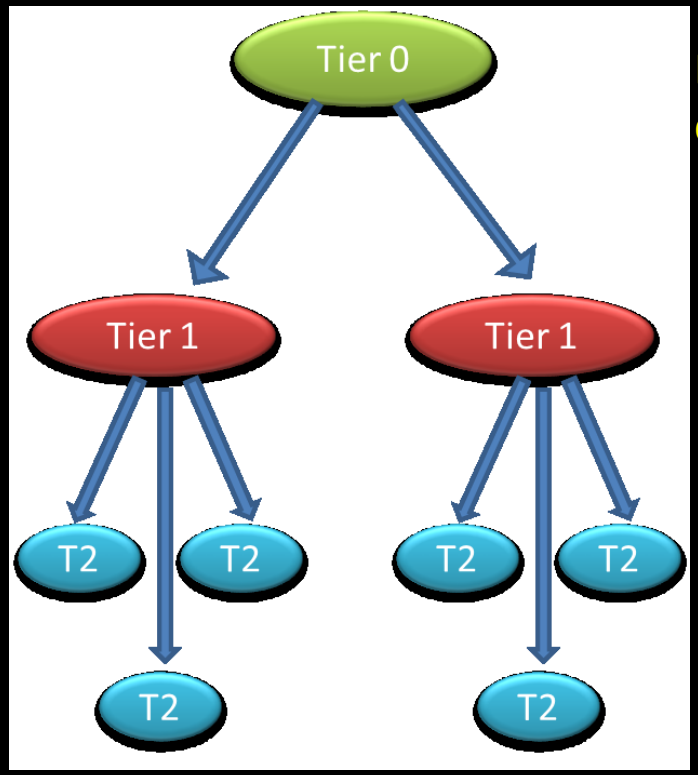
ATLAS data set

250 PB !

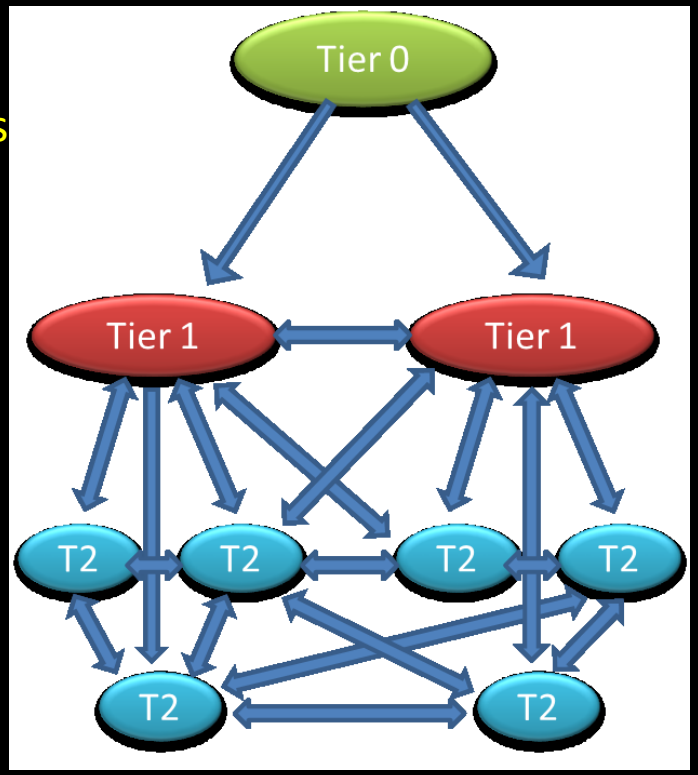
Current ATLAS data set, all data products: 250+ PB

1+M files transferred per day

Computing model evolution. Reducing Complexity



Evolution of computing models



Wide area networks are very stable now

Hierarchy

Network capabilities and data access technologies have significantly improved our ability to use resources independent of location

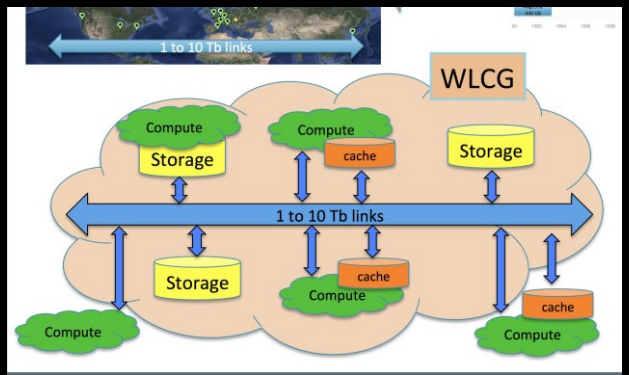
Relaxing hierarchical model : Flat instead of Tiered Grid model

Mesh

Data Management Evolution

- Storage and Compute loosely coupled but connected through fast network
  - Heterogeneous computing facilities in and outside the cloud
  - Different centers with different capabilities, for different use cases
- We want to keep control of data
  - Need to be able to deploy data to a diverse set of resources
    - Clouds, dedicated sites, HPC centers, etc
  - Will need to a combination of real time delivery and advanced data caching

In order to replicate samples of hundreds TB in hours we will need the systems optimized end-to-end and a very high capacity network in between.



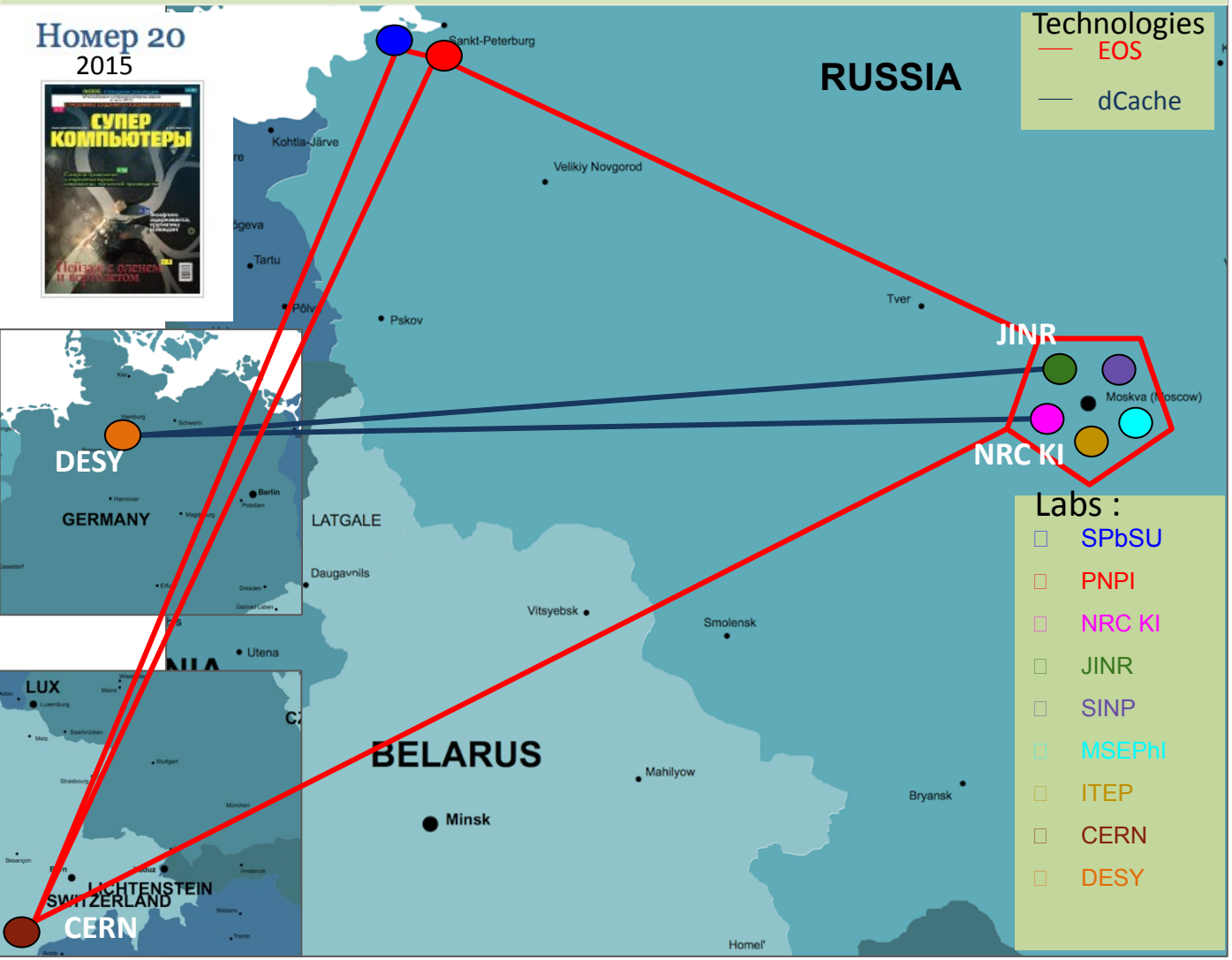
Russian Fund for Basic Research Award. Federated Storage

CERN, DESY, NRC KI, JINR, PNPI, SPbSU, MSEPhI, ….; ATLAS, ALICE (LHC), NICA (JINR)

EOS technology : NRC-KI, JINR, T2 (ATLAS, PNPI, Gatchina), T2 (ALICE, SPbSU, Petergof), CERN

dCache technology : NRC-KI, JINR, DESY

P.Fuhrmann, A.Kirianov, A.Klimentov, D.Krasnopevtsev, A.Kryukov, M.Lamanna, A.Peters, A.Petrosyan, E.Ryabinkin S.Smirnov, A.Zarochentsev, D.Duelmann



R&D Project Motivation

Computing models for the Run3 and HL-LHC era anticipate a growth of storage needs.

The reliable operation of large scale data facilities need a clear economy of scale.

A distributed heterogeneous system of independent storage systems is difficult to be used efficiently by user communities and couples the application level software stacks with the provisioning technology at sites.

- Federating the data centers provides a logical homogeneous and consistent reliable resource for the end users

Small institutions have no enough people to support fully-fledged software stack.

- In our project we try to analyze how to set up distributed storage in one region and how it can be used from Grid sites, from HPC, academic and commercial clouds, etc.

Options for Future Computing & Collaboration

The ultimate question

- How will data be processed and analyzed in 7-10 years and beyond ?

- Buy facilities
  - ✓ Pro : Own it! No impediment to running at full capacity when needed
  - ✓ Con : Must invest for peak utilization, even if not used
- Use services from other providers :
  - ✓ Pro : Others make capital investments
  - ✓ Con : Will usage be available/affordable when needed ?

We worked hard during last years to provide examples of infrastructure not owned by HENP and to integrate HPC with HTC

- Hybrid model
  - ❑ Own baseline resources that will be used at full capacity
  - ❑ Use service providers for peak cycles when needed

Impact on Data and Workflow Management

One of the biggest improvements in joining to a much larger pool of resources is breaking the idea we need to lay out our resources for average load

Workflows could be completed as they are defined and not over months

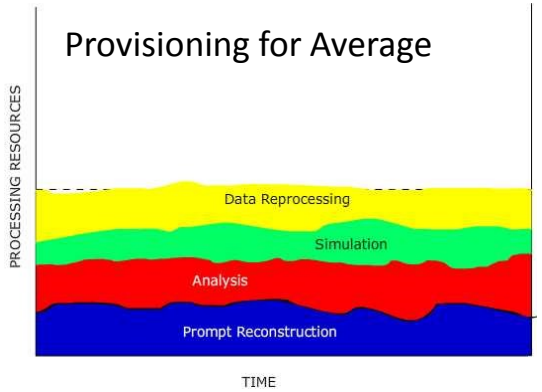
In these processing models the workflow system needs to be able to scale to 5-10 times the average load

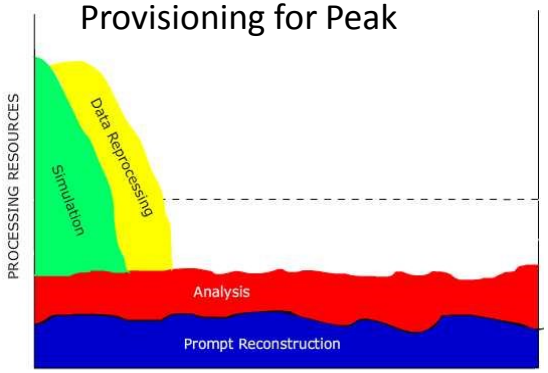
- We want to be able to burst to high values
- The least expense time to be delivered resources might be all at the same

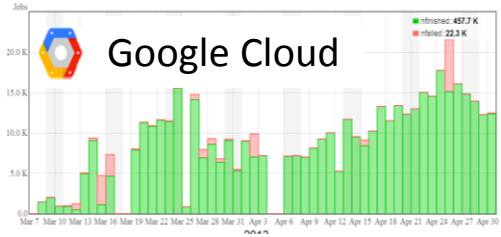
If one is using commercially provided computing faults turn into real money

- Need to focus on potentially wasteful things
  - Infinite loops
  - Giant log output that trigger data export charges
  - CPU efficiency loss

All things we probably should have been worrying about with our dedicated systems, but somehow when you are directly paying for the resources you are a bit more careful

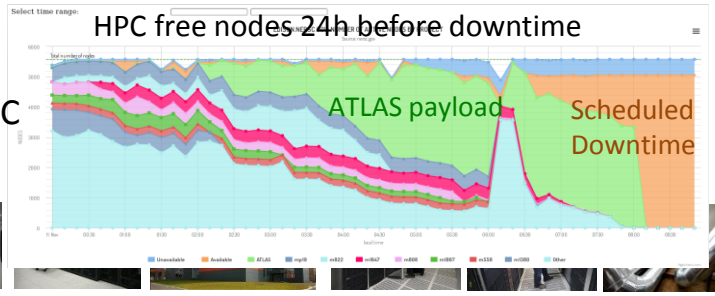






Running ATLAS jobs on GCC

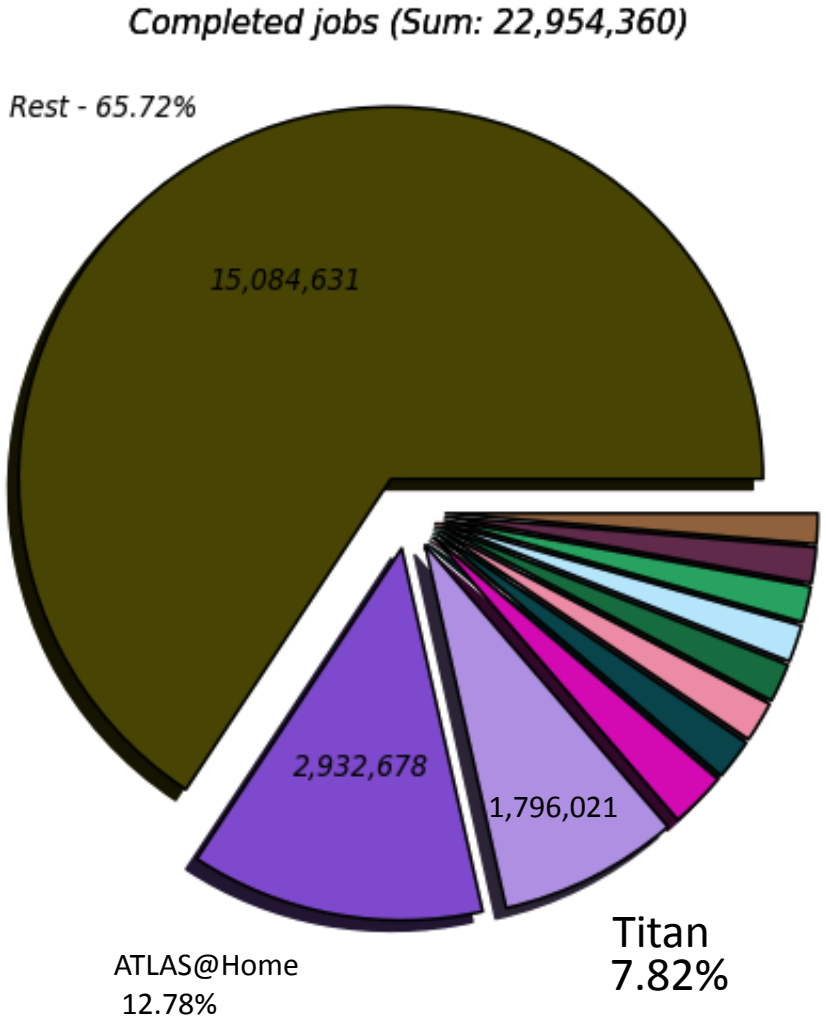
Running ATLAS jobs on HPC
In opportunistic mode



Supercomputers. Titan contribution to ATLAS MC simulations

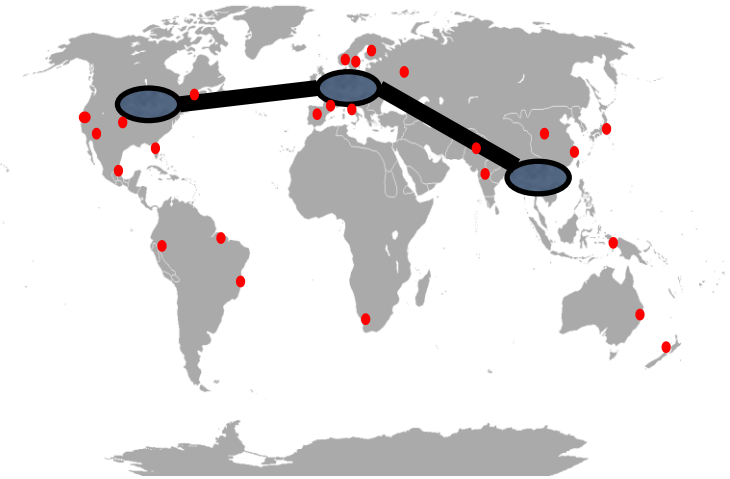
ATLAS simulation jobs completed worldwide: Jan – Oct 2016

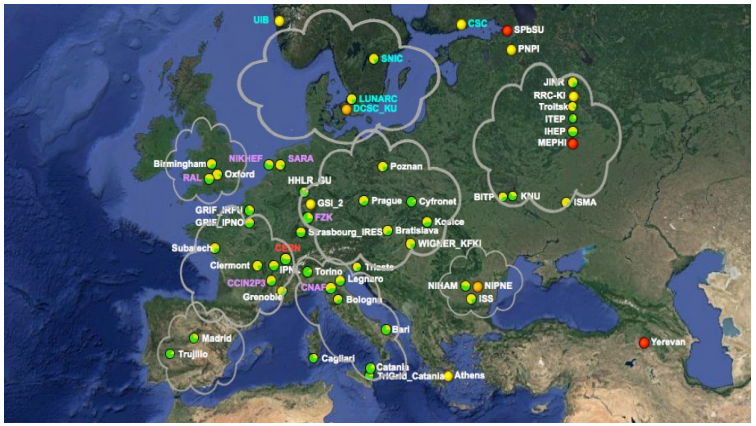
- Titan backfill contributed 7.8% of total simulation jobs CY16 to date.



Workflow and Data Management Evolution

- Big centers for data reduction impacts workflow and data management
- Data selection workflow sits on top of "big data" tools
  - Focusing effort on reproducibility and shared selection criteria
- Data Management involves moving small samples to end sites
- Activity is triggered automatically
  - Needs throttling mechanisms
- The bulk of the data is placed at big sites
  - Reduced samples are moved and replicated
- Still a push to enable the processing on a variety of resources
  - Ability to burst to high capacity becomes even more important when access can trigger processing

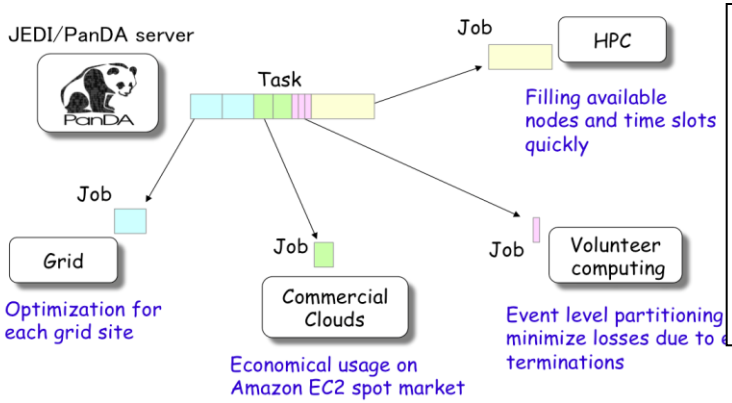




Workflow Management. PanDA. Production and Distributed Analysis System

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

➢ Workload partitioning for traditional and opportunistic resources



Global ATLAS operations

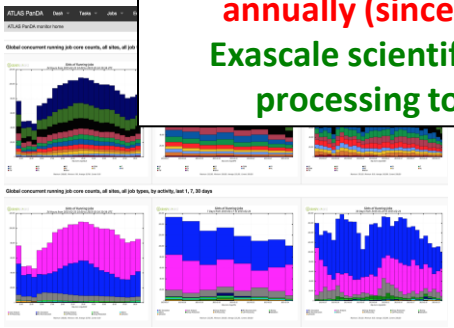
Up to ~250k concurrent jobs
25-30M jobs/month at >150 sites
~1400 ATLAS users

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
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
2016:COMPASS adopted PanDA (JINR, CERN, NRC-KI), NICA (JINR)

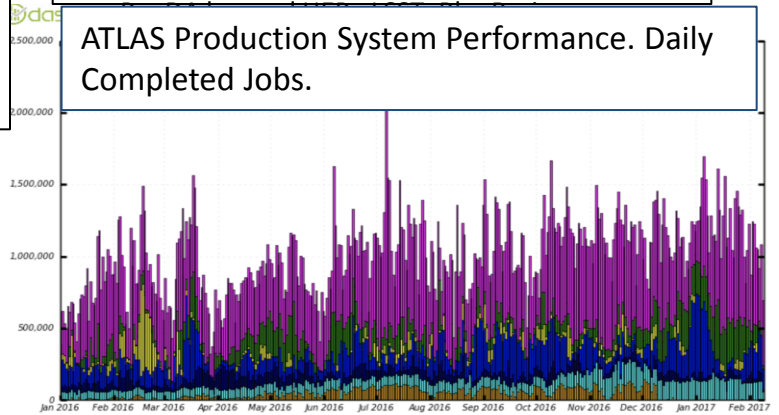
BigPanDA Monitor

http://bigpanda.cern.ch/



First exascale workload manager in HENP
1.3 Exabytes processed annually (since 2014)
Exascale scientific data processing today

ATLAS Production System Performance. Daily Completed Jobs.



BigPanDA. Growing PanDA Ecosystem

COMPASS

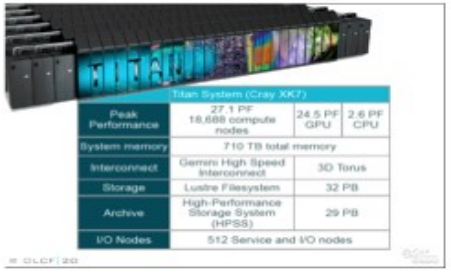
ALICE

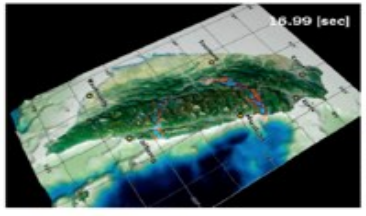
Google Cloud Platform

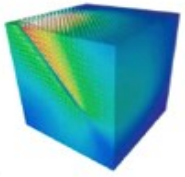
NICA

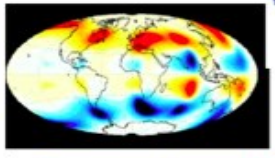
nectar

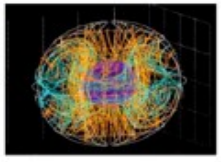
amazon web services™ EC2

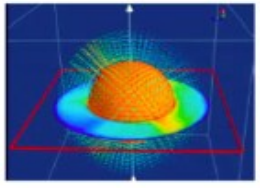


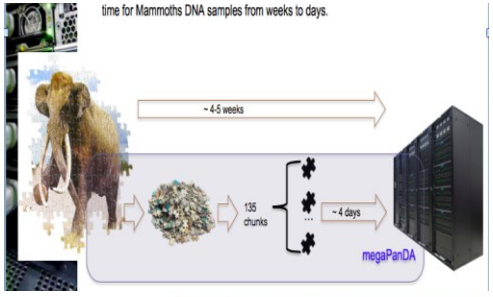






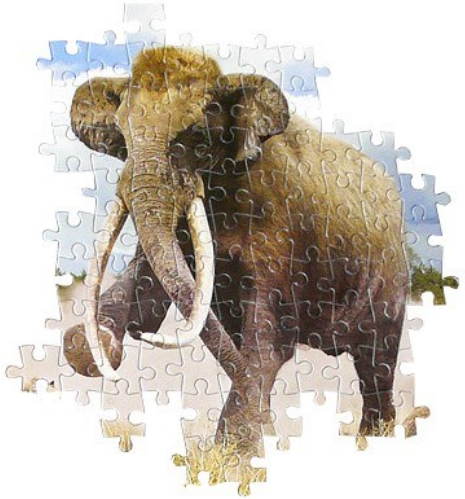


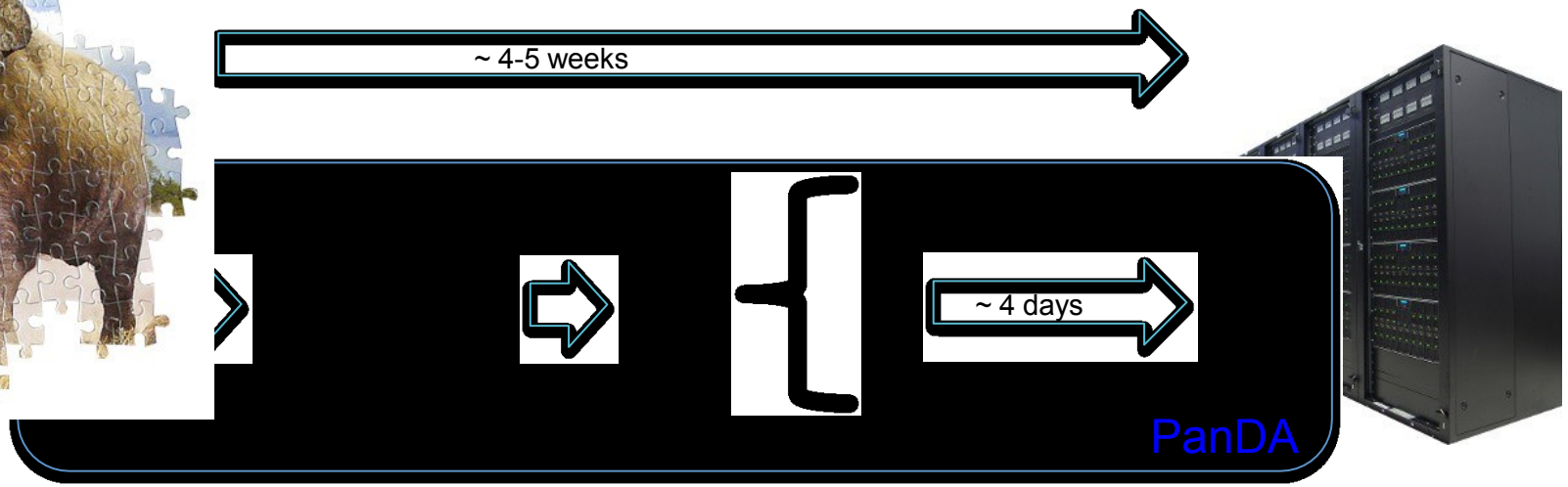




BigPanDA in Genomics

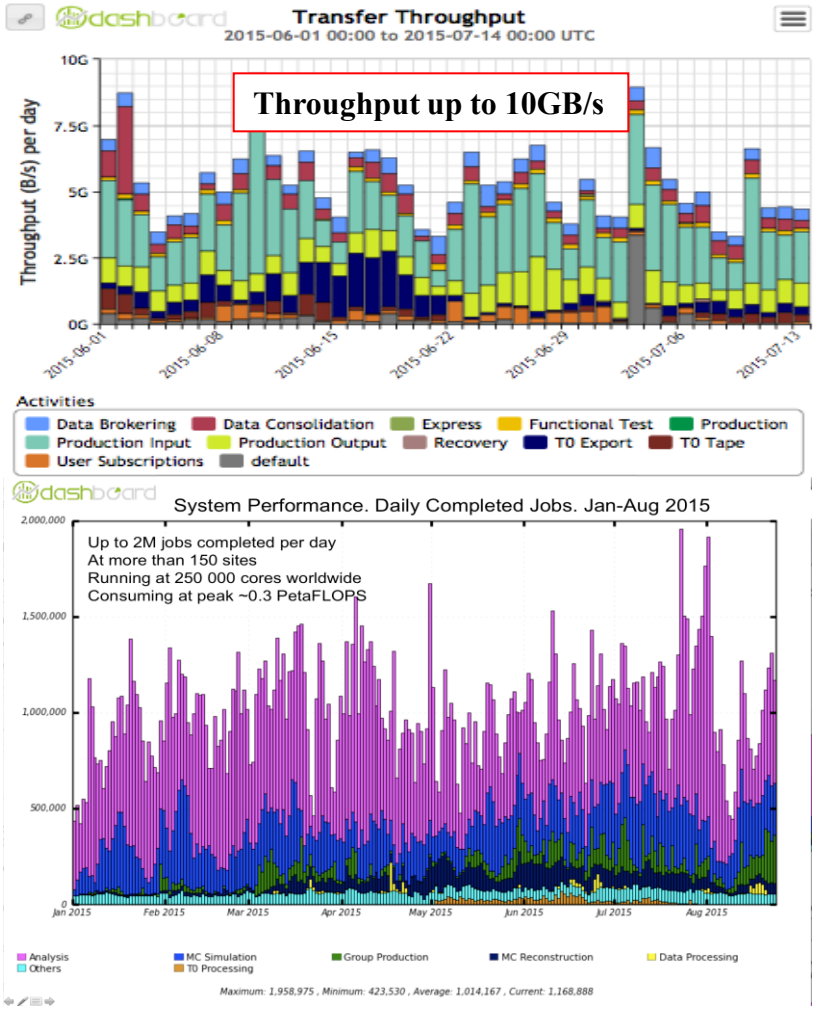
- At NRC KI PALEOMIX pipeline was adapted to run on local supercomputer resources powered by PanDA.
- It was used to map mammoth DNA on the African elephant reference genome
- Using software tools developed initially for HEP and Grid reduced genomics payload execution time for Mammoths DNA sample from weeks to days.

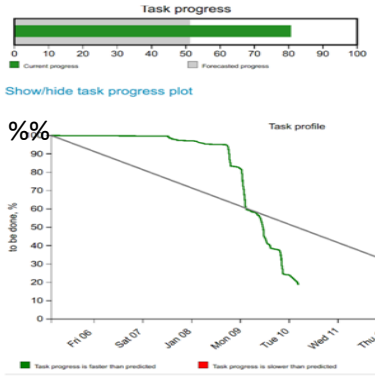




Russian Science Foundation Award. «Machine Learning» algorithms to predict complex system behaviour

- Production System is a large, complicated, distributed system;
  - *Hard to simulate;*
  - *Hard to detect anomalies*
  - *Hard to predict its behavior*
- Very thorough logging;
- Machine learning (ML) algorithms are computationally intensive, using them on raw logs (database rows) is infeasible.
- However, it is possible to use machine learning algorithms if we limit their input to some aggregated metrics of Production System
- The most important metrics are :
  - Time To Complete for tasks/jobs
  - resource utilisation
  - percentage of failed tasks/jobs
  - Running/pending jobs ratio.





Payload Execution profile

days

R&D project (ATLAS, CMS, LHCb)
D.Duelmann, A.Klimentov

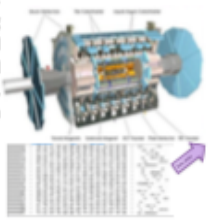
Evaluating non-relational storage technology for HEP meta-data and meta-data catalog

Golosova Marina, Grigoryeva Maria, Klimentov Alexei, Osipova Victoria, Ryabinkin Eygene

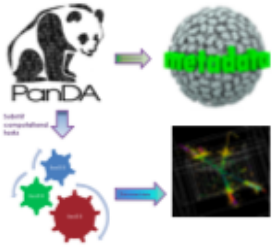
Large-scale scientific experiments produce vast volumes of data. These data are stored, processed and analyzed in a distributed computing environment. The life cycle of experiment is managed by specialized software like Distributed Data Management (DDM) and Workload Management systems (WMS). In order to be interpreted and mined, experimental data must be accompanied by auxiliary metadata, which are recorded at each phase of the data handling.

ATLAS EXPERIMENT
- Petabyte scale
- 3000 scientists
- 180 institutions
- 40 countries

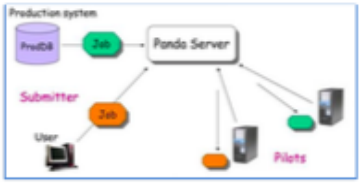
The ATLAS experiment uses **PanDA** *(The Production and Distributed Analysis) Workload Management System* for managing the workflow for all data processing on the WLCG.
PanDA provides location transparency of processing and data access for HEP and other data-intensive research fields.



- Up to 30M jobs completed every month
- At more than 120 sites
- Running at 130 000 cores worldwide



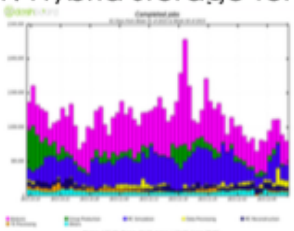
Processing and analysis of huge amounts of auxiliary metadata, which is constantly grows, is no less challenging task than the management and processing of experimental data itself.
Modern LHC collaborations work at petabyte scale production and distributed analysis processing.



- ✓ Pilot-based WMS (pilots – "place holders" for payload)
- ✓ **PanDA Server** is the main component which provides a task queue managing all job information centrally
- ✓ Jobs are submitted to the PANDA server via a simple Python/HTTP client interface
- ✓ Pilots retrieve jobs from the PANDA server to run the jobs as soon as CPU's becomes available

Exploring the metadata storage and processing techniques in ATLAS we have identified potential bottlenecks and proposed methods of improvement, based on non-relational approach:

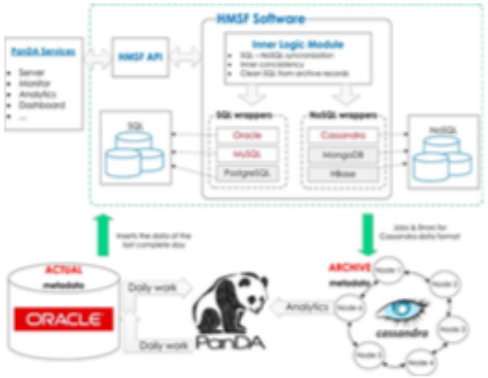
1. Hybrid storage for archived metadata



The number of completed jobs per day for the last million jobs per day

ATLAS PanDA WMS runs *more than 1.5 M jobs per day*. Full jobs archive now hosts information of over *billion of records*. As the metadata volume grows, the underlying software and hardware stack encounters certain limits that negatively affect processing speed and the possibilities of metadata analysis. To improve the scalability and performance of analytical and reporting applications, based on computational jobs metadata, we are developing ***Hybrid Metadata Storage Framework (HMSF)***, providing the metadata segmentation between relational and non-relational database back-ends.

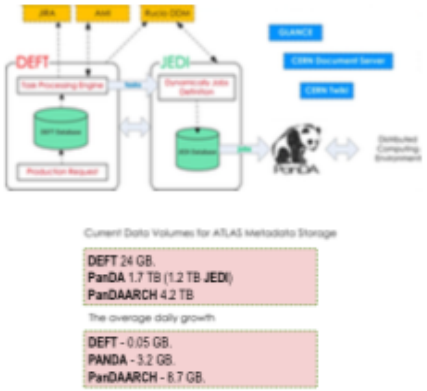
Hybrid Storage Architecture



2. Data Knowledge Catalog (DKC)

LHC Experiments have a set of metadata sources, with data, recorded at each phase of the experiment lifecycle. These metadata sources are *loosely coupled* and potentially may provide to an end-user *inconsistency in requested information*.
To aggregate and synthesize a range of primary metadata sources, and enhance them with flexible schema-less addition of aggregated data, we are developing the ***Data Knowledge Catalog*** serving as the intelligence behind GUIs and APIs.

Data sources:
- AMI (Atlas Metadata Interface)
- GLANCE (search engine for the Atlas Collaboration)
- Rucio (ATLAS Distributed Data Management)
- ProdSys2 (DEFT + JEDI)
- JIRA ITS (Issue Tracking Service)
- Indico (allows you to manage complex conferences, workshops and meetings)
- CERN Document Server
- CERN Twiki (a tool for web page collaborative writing)



DKC Architecture

Data Knowledge Base. ATLAS R&D project

This work was funded in part by the Russian Ministry of Science and Education under Contract N14.Z50.31.0024.

TOMSK POLYTECHNIC UNIVERSITY

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР "КУРЧАТОВСКИЙ ИНСТИТУТ"

BROOKHAVEN NATIONAL LABORATORY

BigData Lab. Major Accomplishments.

- Develop, code and commission Production System on exascale data processing era
- Integrate HighPerformance Computers and SuperComputers for data processing and simulation
- Deploy of BigPanDA workflow management tools on supercomputers
  - Develop, code and implement a new capability in PanDA to collect information about unused worker nodes on Titan, and based on that information; adjust workload parameters to fill free resources.
    - Demonstrated that this new capability can be used at other supercomputing platforms.
      - Deployed on supercomputers at "Kurchatov" SC, NERSC at LBNL and Anselm at IT4I in Ostrava (Czech Rep).
- Generalize PanDA beyond ATLAS and HENP
  - common projects with ALICE, LSST, COMPASS, AMS@ASGC, bioinformatics
- (co)Leading International R&D projects
  - Machine Learning (CERN, NRC KI, RF, EU, US Universities)
  - Federated storage (NRC KI, JINR, CERN and DESY)
  - Data Knowledge Base for megascience experiments and Big Scientific Collaborations (Centers)
- Awarded 8 Russian and International grants for fundamental research for HENP Computing
  - …and beyond

Creation of a new collaboration between NRC-KI, DESY, CERN, JINR, MEPhI, TPU, BNL, GSI, UTA and Rutgers, that led to a project now funded by RF Ministry of Science and Education in 2017-18

Thanks

- This talk drew on presentations, discussions, comments, input from many
- Thanks to all, including those I've missed
  - *I.Bird, P.Buncic, S.Campana, K.De, I.Fisk, M.Grigorieva, M.Gubin, B.Kersevan, A.Kirianov, M.Lassnig, T.Maeno, R.Mashinistov, A.Patwa, A.Poyda, T.Wenaus, A.Zarochentsev …*