

# **PanDA : распределенная система обработки Больших Данных в неоднородной компьютерной среде**

**Пятый Московский Суперкомпьютерный Форум**

**Алексей Климентов**

**Владимир Кореньков**

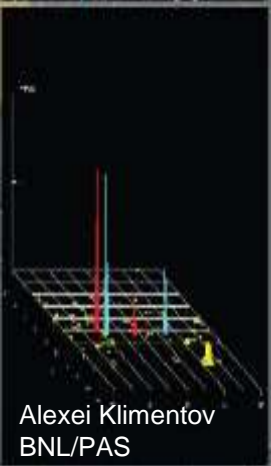
**Василий Велихов**

**21 Октября 2014**



# Main topics

- **CERN and Large Hadron Collider (LHC)**
  - LHC data processing challenges
  - Evolution of LHC computing model
  - Workload Management System core ideas.
- **Workload Management System - PanDA**
- **LHC Computing needs for Run2+ (2015-2018)**
- **megaPanDA project :**
  - Expanding PanDA beyond the Grid and High Energy Physics (HEP)
  - PanDA at Super-Computing Facilities
- **Summary**





# Tools: LHC and Detectors

pp, B-Physics, CP Violation  
(matter-antimatter symmetry)



LHCb



ATLAS



CMS

General Purpose,  
proton-proton, heavy ions  
Discovery of new physics:  
Higgs, SuperSymmetry

Exploration of new energy frontier  
in p-p and Pb-Pb collisions

ATLAS

CERN Meyrin

SPS 7 km

ALICE



ALICE

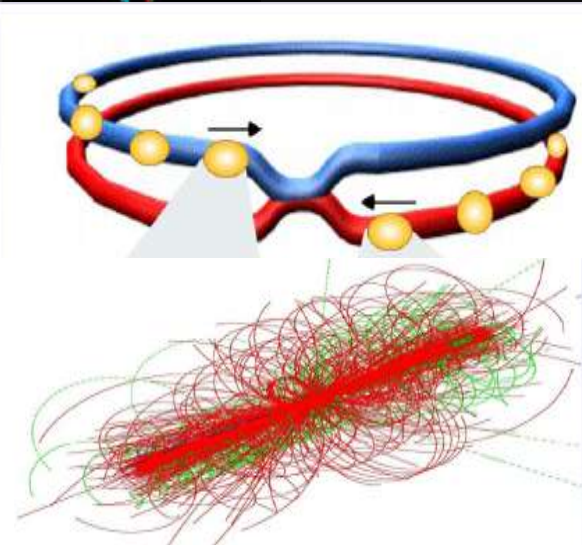


LHC ring:  
27 km circumference

Heavy ions, pp  
(state of matter of early universe)



# Proton-Proton Collisions at the LHC



LHC delivered billions of collision events to the experiments from proton-proton and proton-lead collisions in the Run1 period (2009-2013)

- collisions every 50 ns  
= 20 MHz crossing rate
- $1.6 \times 10^{11}$  protons per bunch  
at  $L_{pk} \sim 0.8 \times 10^{34} / \text{cm}^2 / \text{s}$   
 $\approx 35$  pp interactions per crossing – pile-up
- $\approx 10^9$  pp interactions per second !!!
- in each collision  
 $\approx 1600$  charged particles produced

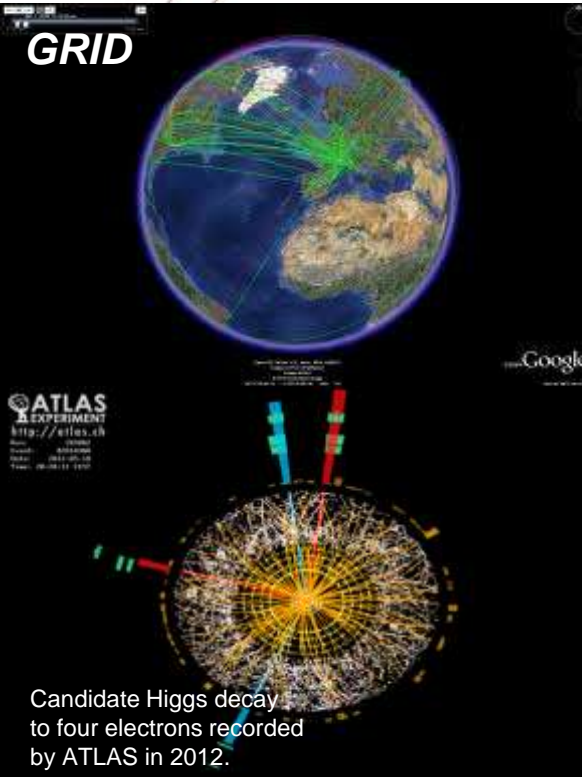
*enormous challenge for the detectors and for data collection/storage/analysis*

**Raw data rate from LHC detector : 1PB/s**

This translates to Petabytes of data recorded world-wide (Grid)

**Up to 6 GB/s to be stored and analysed after filtering**

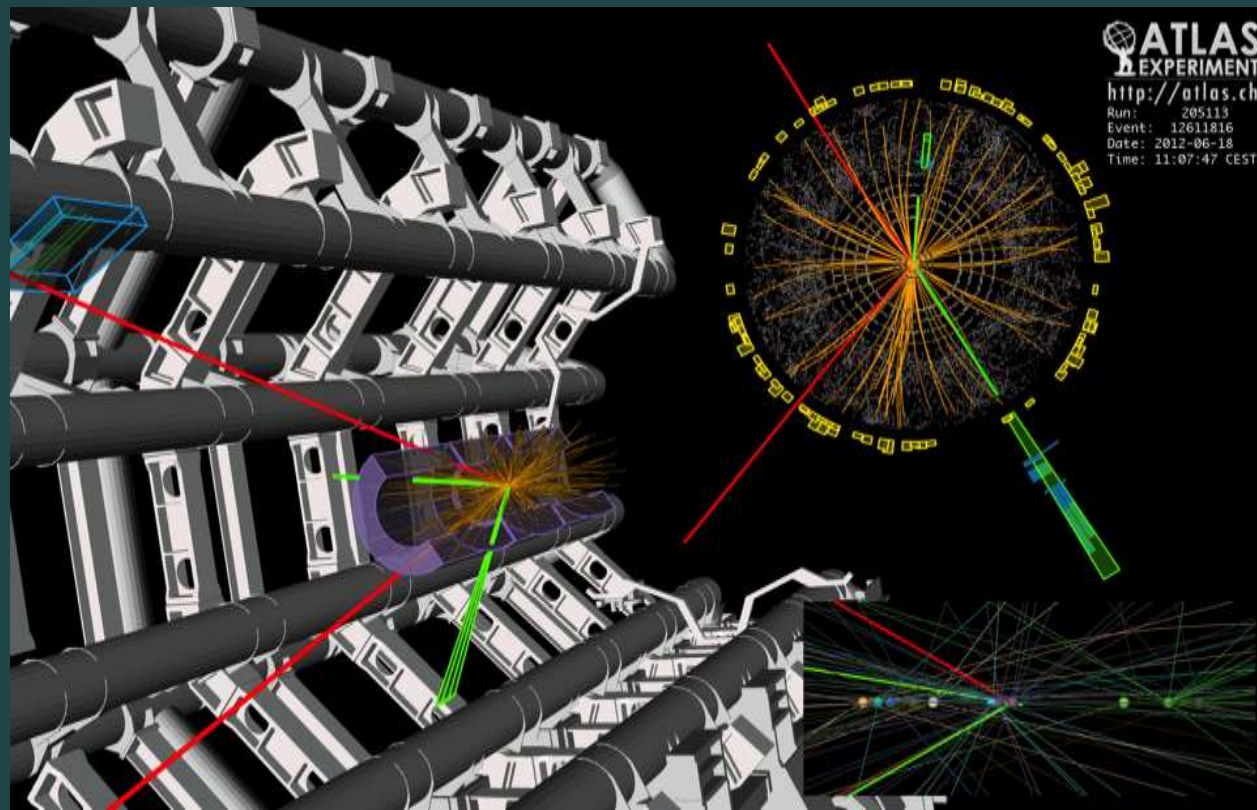
*The challenge how to process and analyze the data and produce timely physics results was substantial, but at the end resulted in a great success*



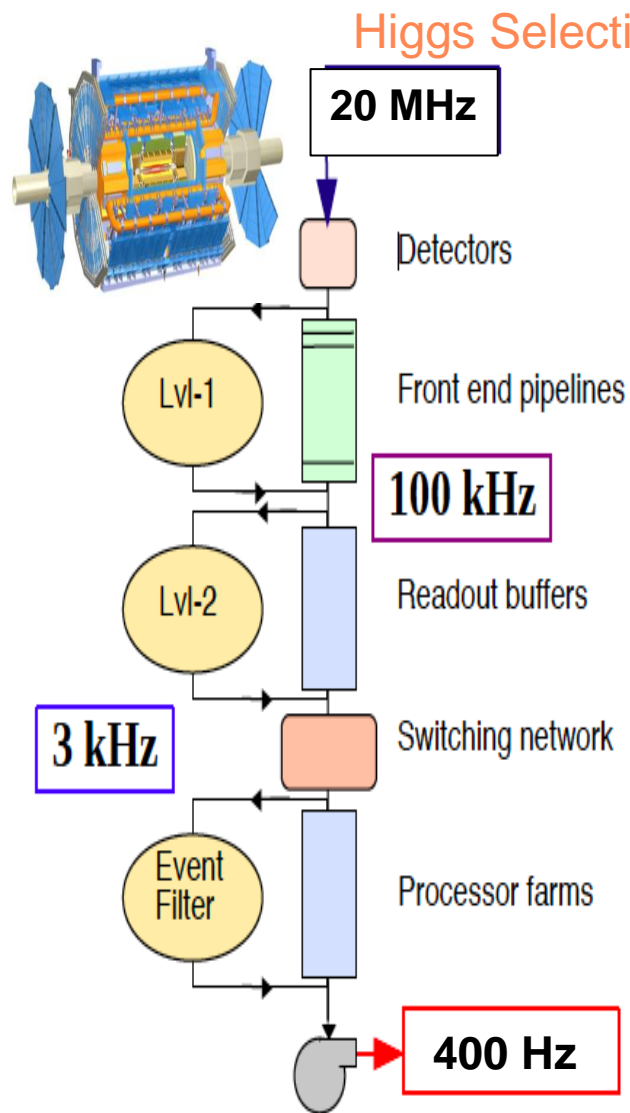
Candidate Higgs decay to four electrons recorded by ATLAS in 2012.

# What is this data?

- **Raw data:**
  - Was a detector element hit?
  - How much energy?
  - What time?
- 150 Million sensors deliver data ... ~ 40 Million times per second
- Up to 6 GB/s to be stored and analysed after filtering
- **Reconstructed data:**
  - Momentum of tracks (4-vectors)
  - Origin
  - Energy in clusters (jets)
  - Particle type
  - Calibration information
  - ...

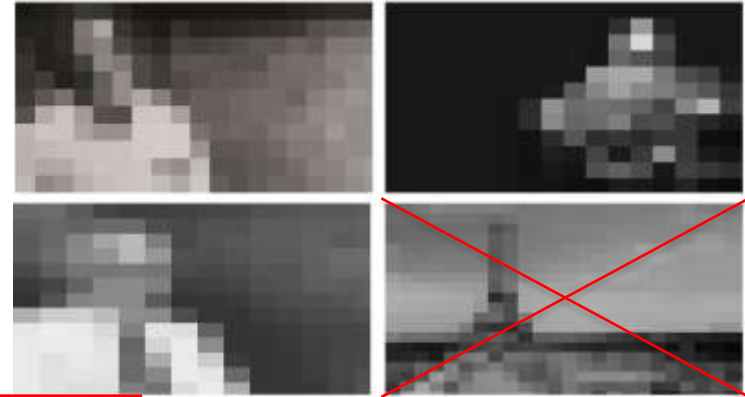


# HEP Online. Reduce the data volume in stages.

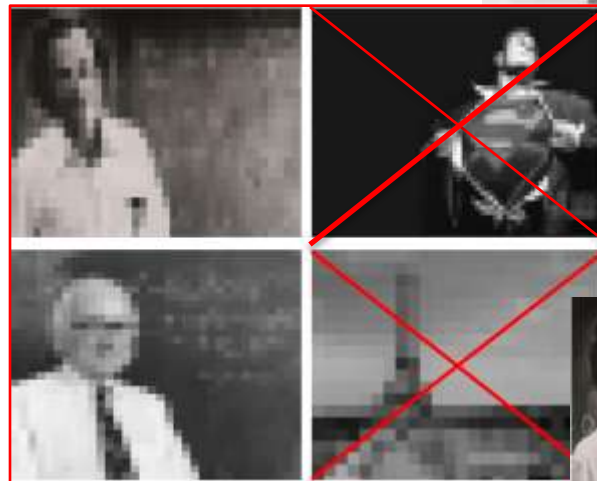


## Higgs Selection using the Trigger

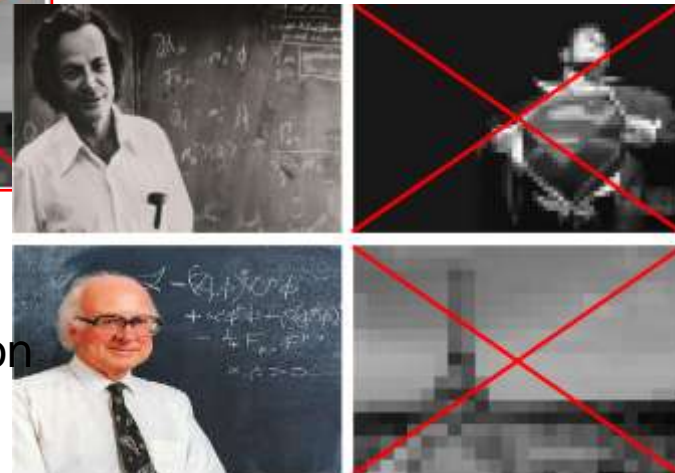
**Level 1:**  
Not all information available, coarse granularity



**Level 2:**  
Reconstruct events  
Improved ability to reject events



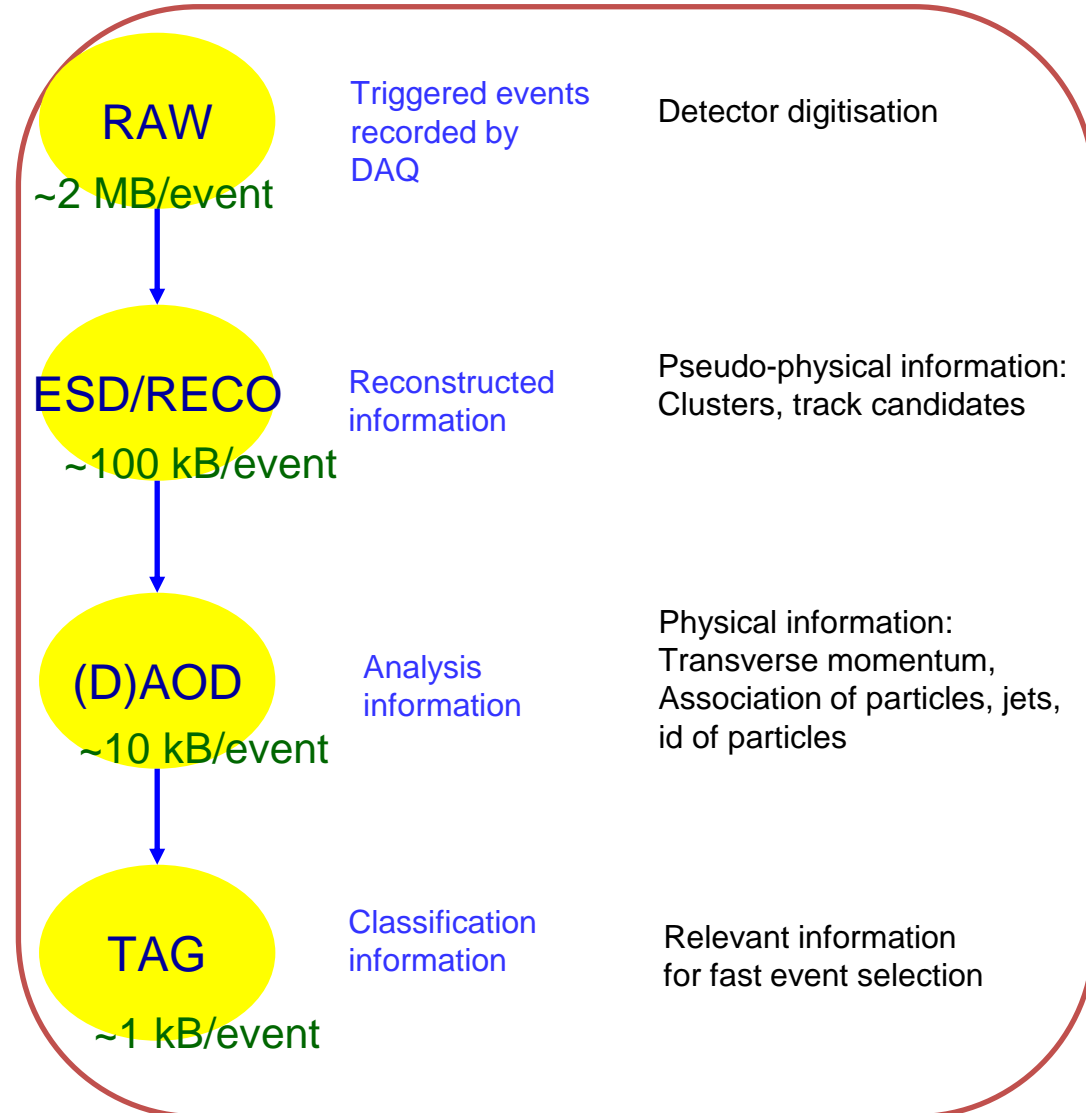
**Level 3:**  
High quality reconstruction algorithms, using information from all detectors



**400 Hz. Run1 ATLAS RAW data rate to tape**


# HEP Offline. Data and Algorithms.

- HEP data are organized as *Events* (particle collisions)
- Simulation, Reconstruction and Analysis programs process “one event at a time”
  - Events are fairly independent → Trivial parallel processing
- Event processing programs are composed of a number of Algorithms selecting and transforming “raw” event data into “processed” (reconstructed) event data and statistics
- **~4 million lines of code (reconstruction and simulation)**
- **~1000 software developers on ATLAS**





# Some history of the scale...



Date	HEP Collaboration sizes	Data volume, archive technology
Late 1950's	2-3	Kilobits, notebooks
1960's	10-15	kB, punchcards
1970's	~35	MB, tape
1980's	~100	GB, tape, disk
1990's	700-800	TB, tape, disk : LEP
2010's	~3000	PB, tape, disk : LHC

## For comparison:

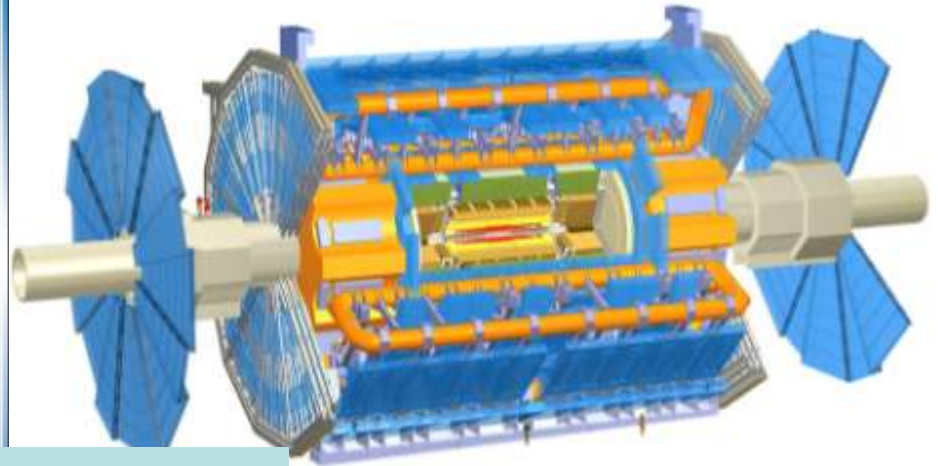
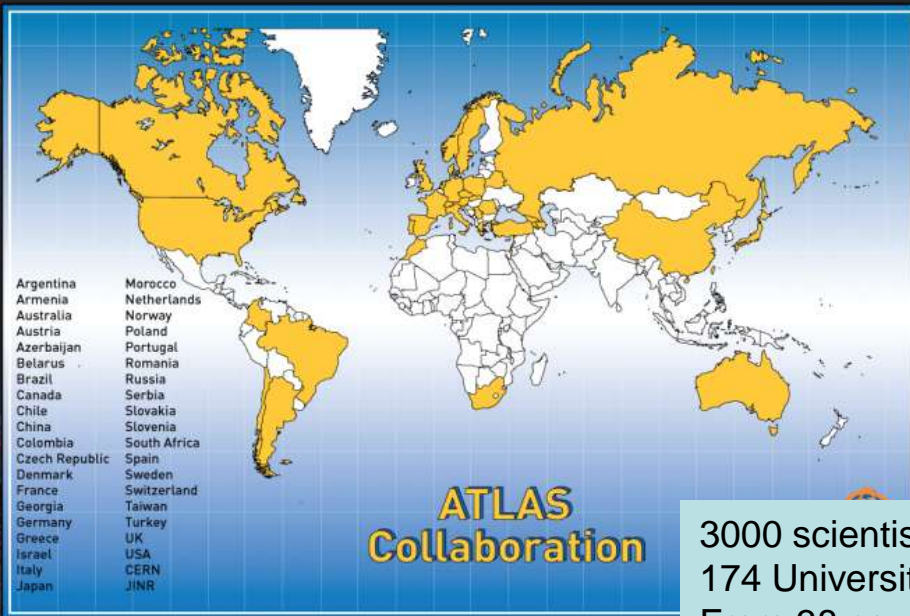
1990's: Total LEP data set ~few TB  
Would fit on 1 tape today

Today: 1 year of LHC data ~25 PB

*LEP – Large Electron-Positron Collider.  $e^+e^-$  collider at CERN in 1989-2000*



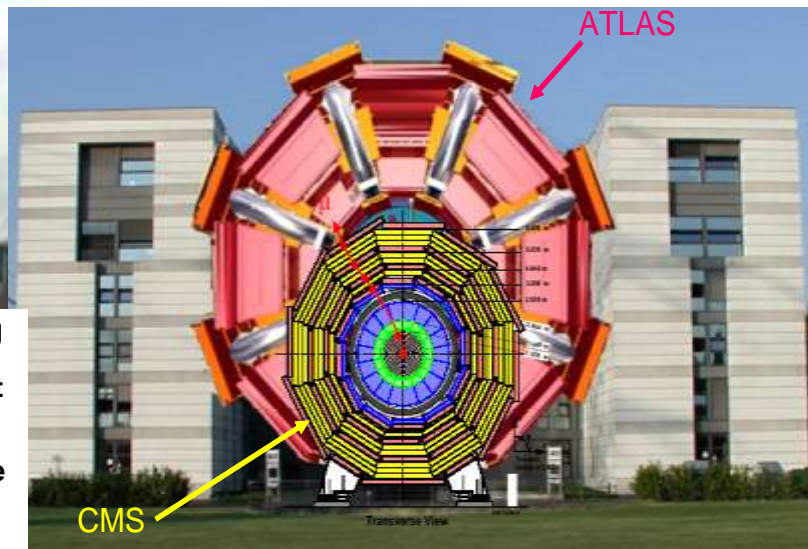
# The ATLAS Experiment at the LHC



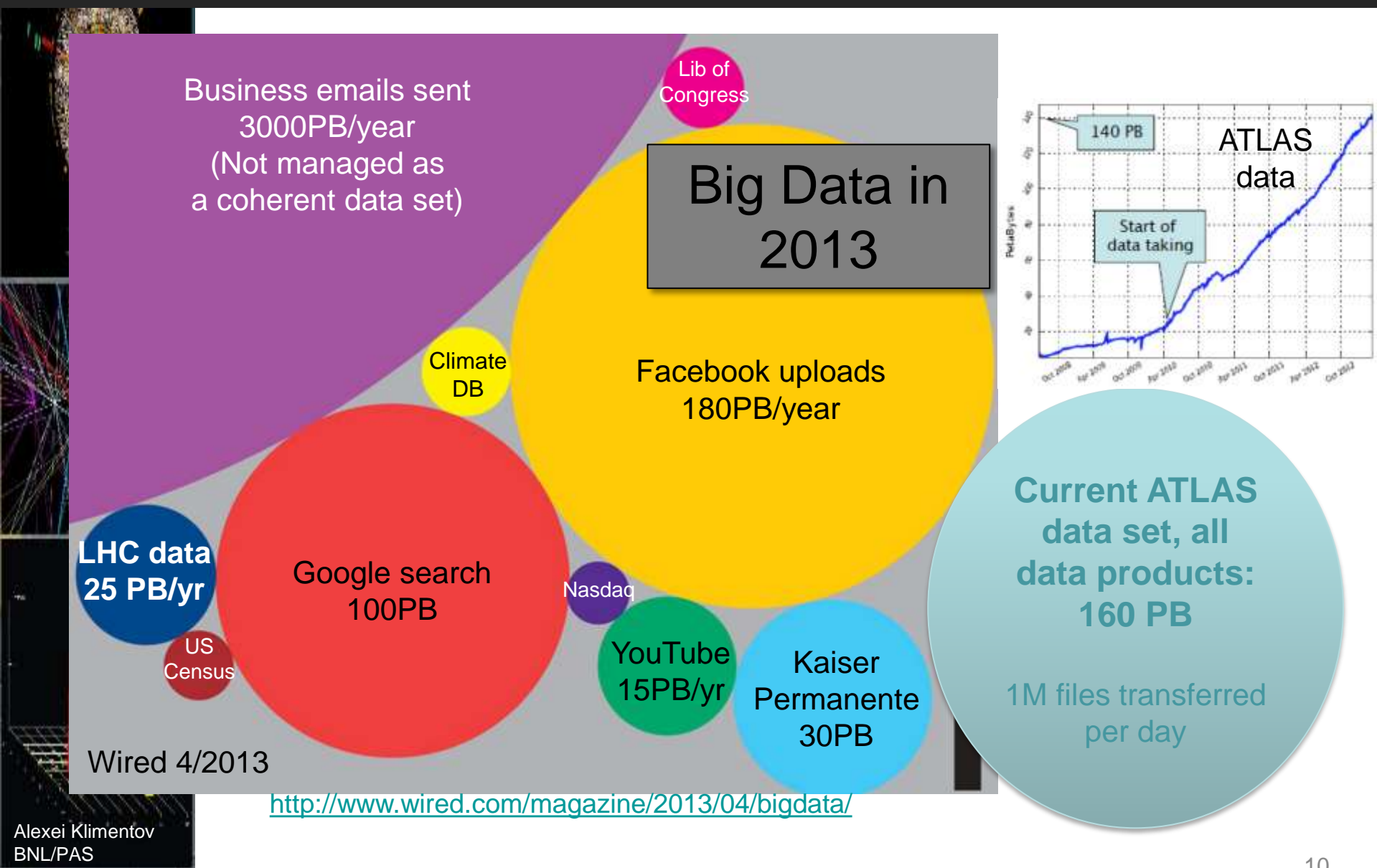
3000 scientists  
174 Universities and Labs  
From 38 countries  
More than 1200 students



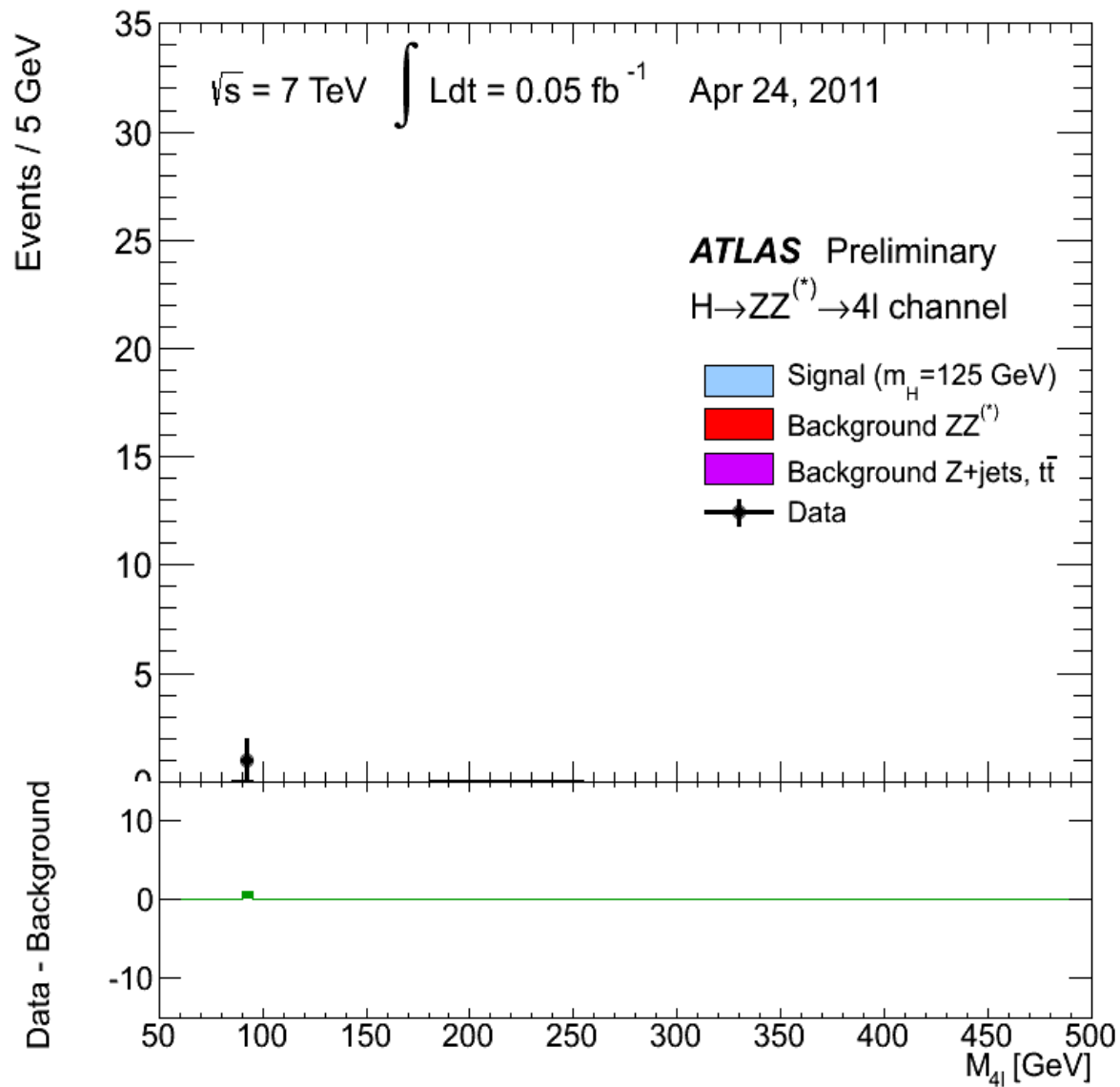
- ATLAS has 44 meters long and 25 meters in diameter, weighs about 7,000 tons. It is about half as big as the Notre Dame Cathedral in Paris and weighs the same as the Eiffel Tower or a hundred 747 jets



# Big Data: often just a buzz word, but not when it comes to ATLAS...



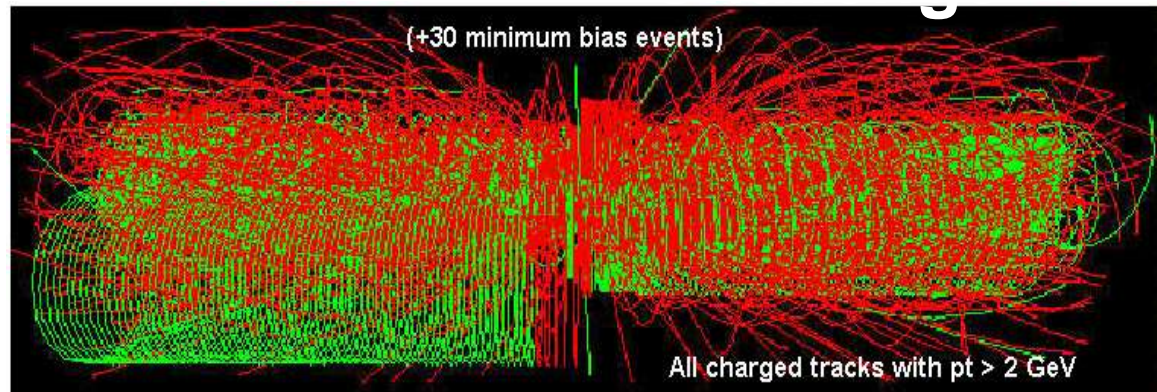
# Higgs Boson Discovery





# ATLAS Data Challenge

- 800,000,000 proton-proton interactions per second
- 0.0002 Higgs per second
- ~150,000,000 electronic channels

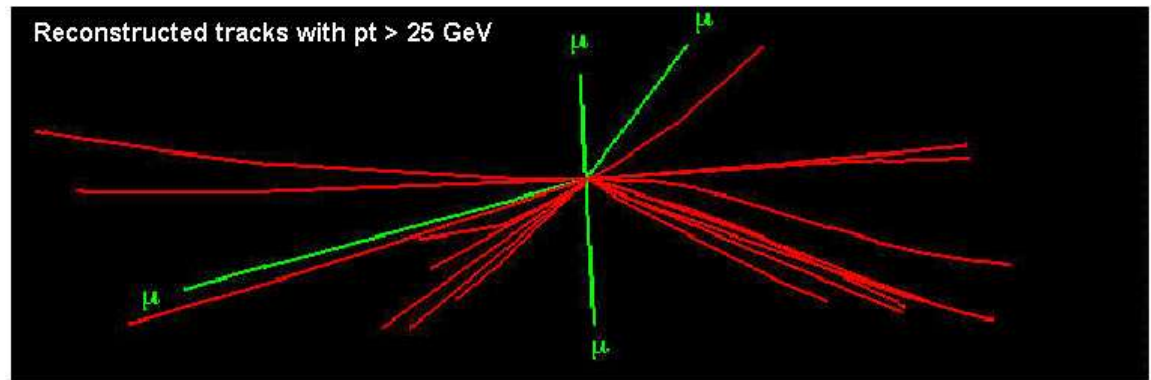


Selectivity: 1 in  $10^{13}$

Like looking for 1 person in a thousand world populations

Or for a needle in 20 million haystacks!

We are looking for this “signature”



PanDA



Drop of water: Roughly 0.1 mL

New physics rate  $\sim 0.00001$  Hz

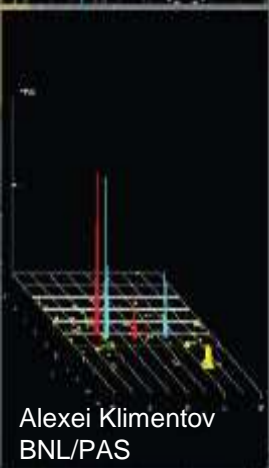
Event Selection :

1 in 10,000,000,000,000

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

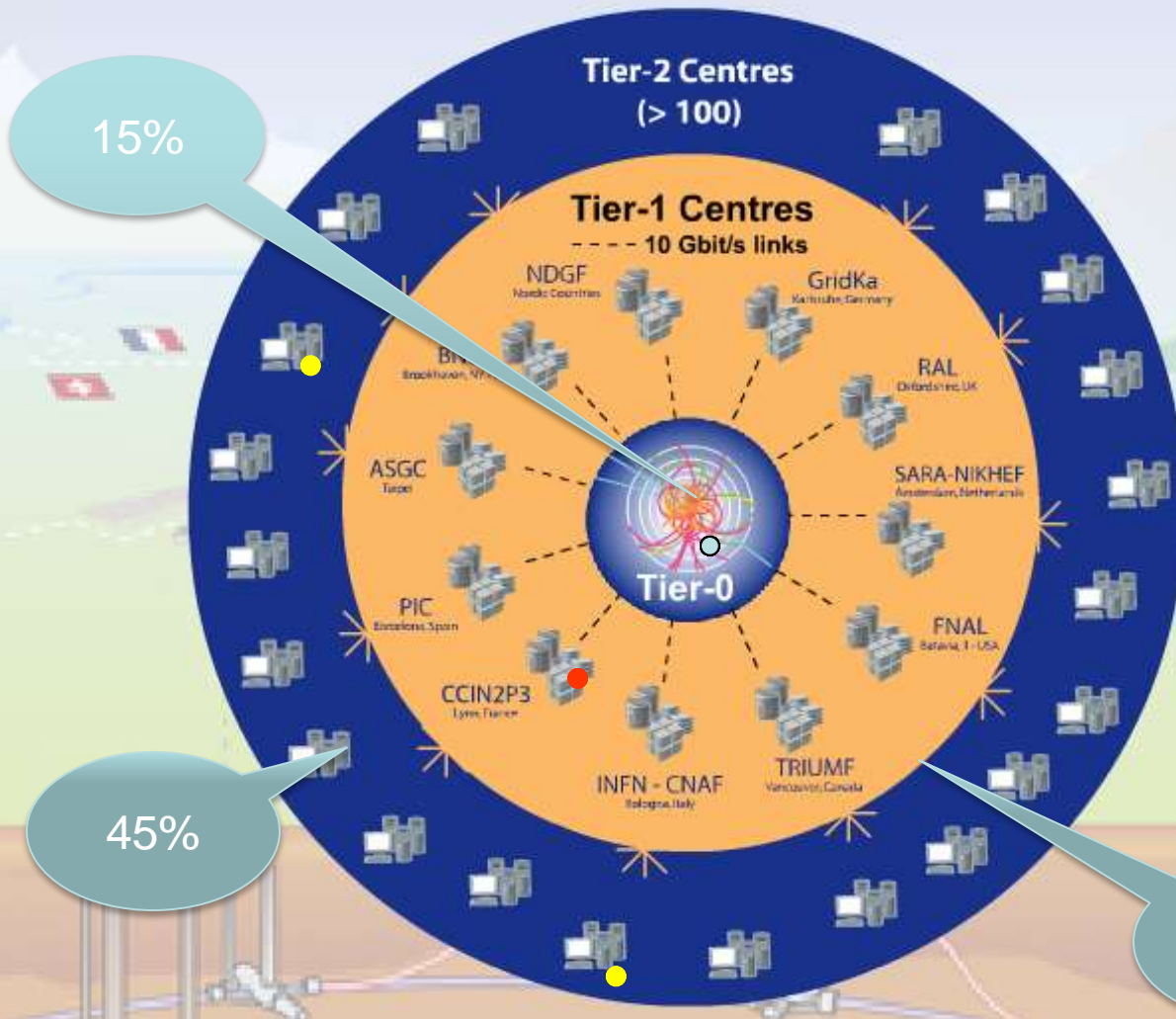


- **Offline computing in HEP**
  - Has changed and evolved dramatically over the past decades
  - Especially for the biggest experiments – at the LHC
- **The situation ~15 years ago**
  - Data processing was performed in large computing centers using local batch systems with dedicated shares
  - A few satellite centers did simulations, occasionally data reprocessing
  - Users were mostly located near large computing centers, usually at the laboratory where the experiment was located, and used a combination of desktops and batch systems for analysis
  - Final Data Summary Files versions were physically shipped for remote analysis





# World-wide LHC Computing Grid



## Tier-0 (CERN): (15%)

- Data recording
- Initial data reconstruction
- Data distribution

## Tier-1 (11 centres; +2 in Russia in 2015 – NRC-KI, JINR): (40%)

- Permanent storage
- Re-processing
- Analysis
- Connected by direct Gb/s network links

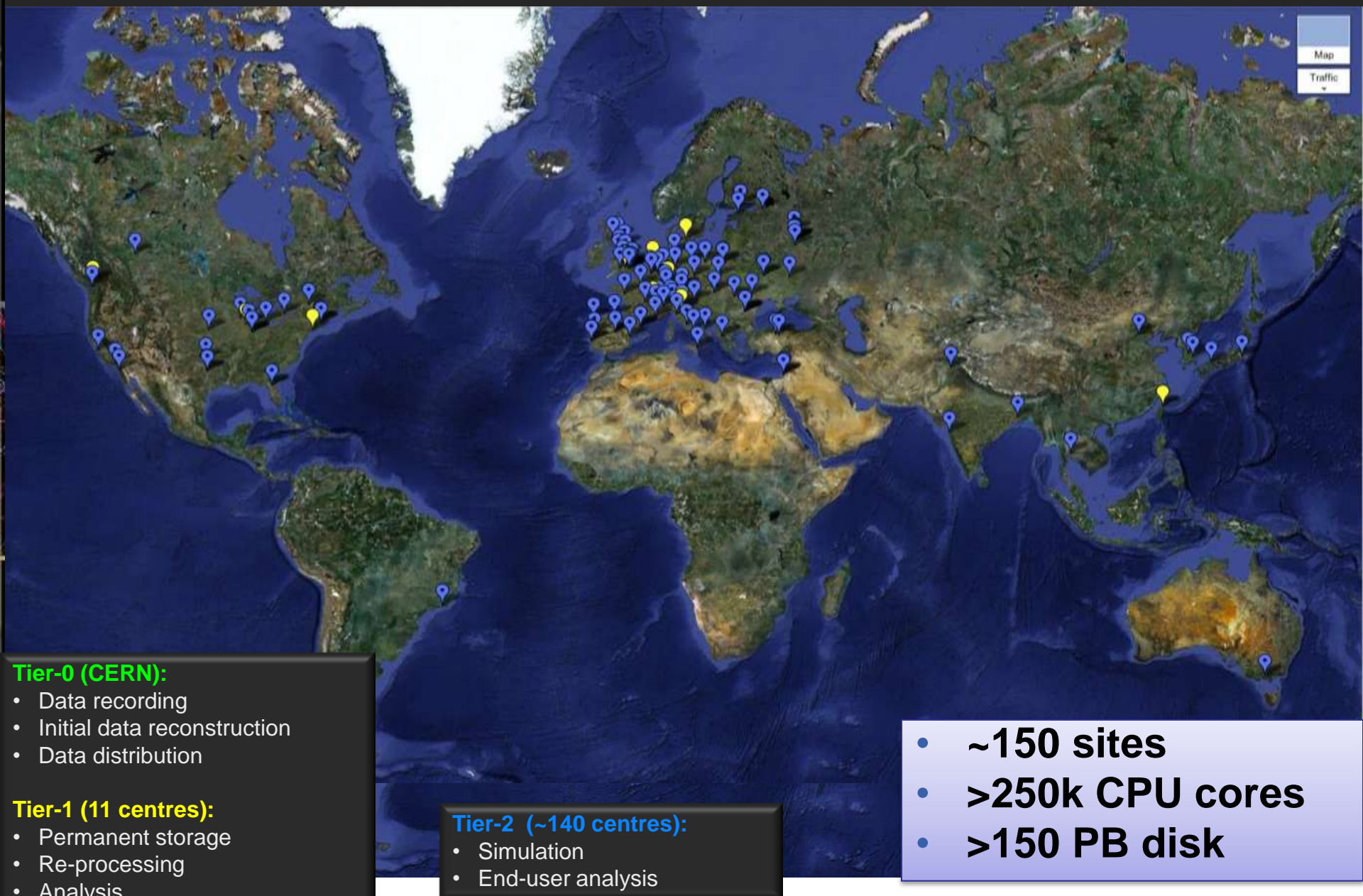
## Tier-2 (~140 centres): (45%)

- Simulation
- End-user analysis



WLCG  
Worldwide LHC Computing Grid

# LHC Computing Grid: A global collaboration...

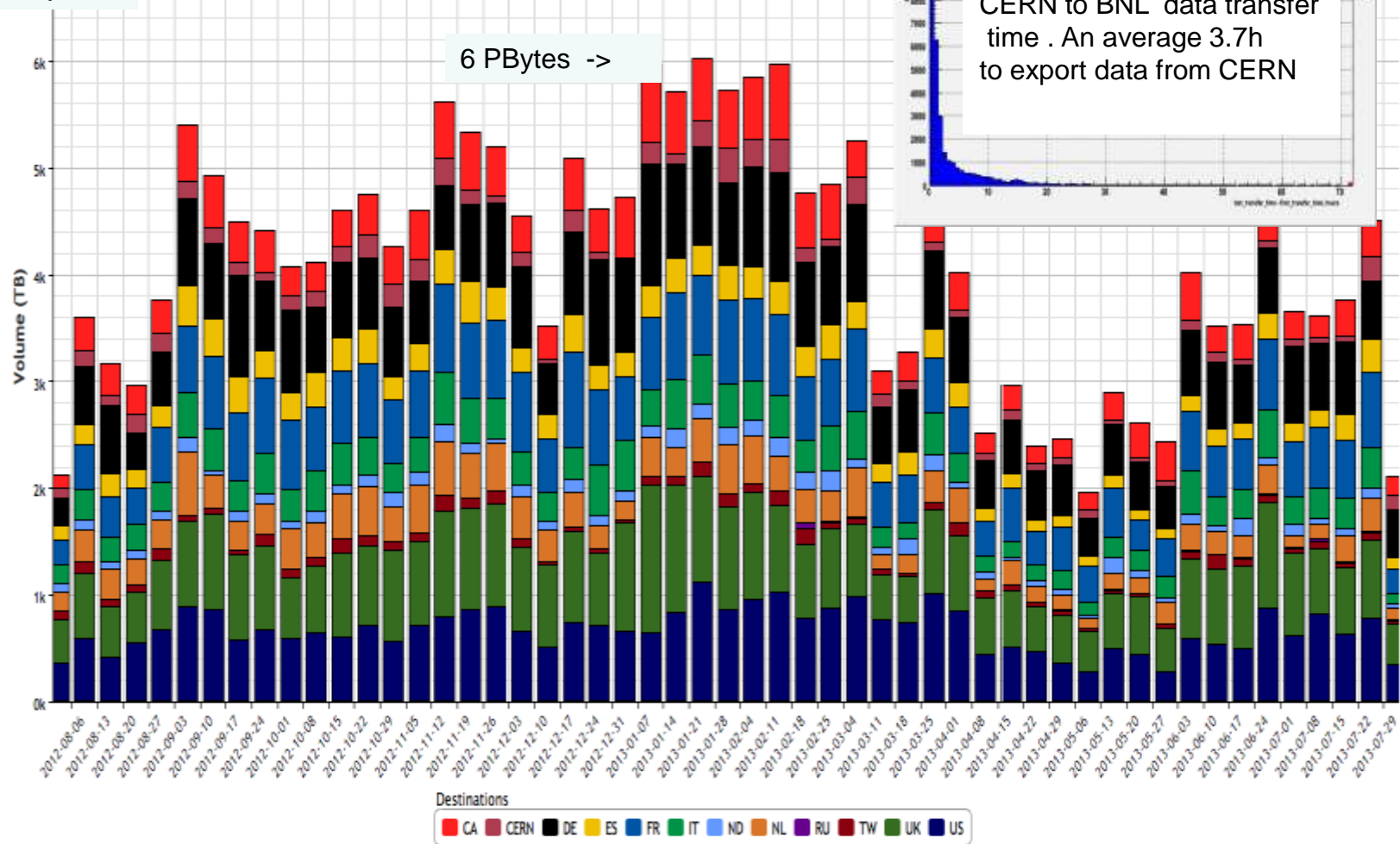




# ATLAS Distributed Computing. Distributed Data Transfer

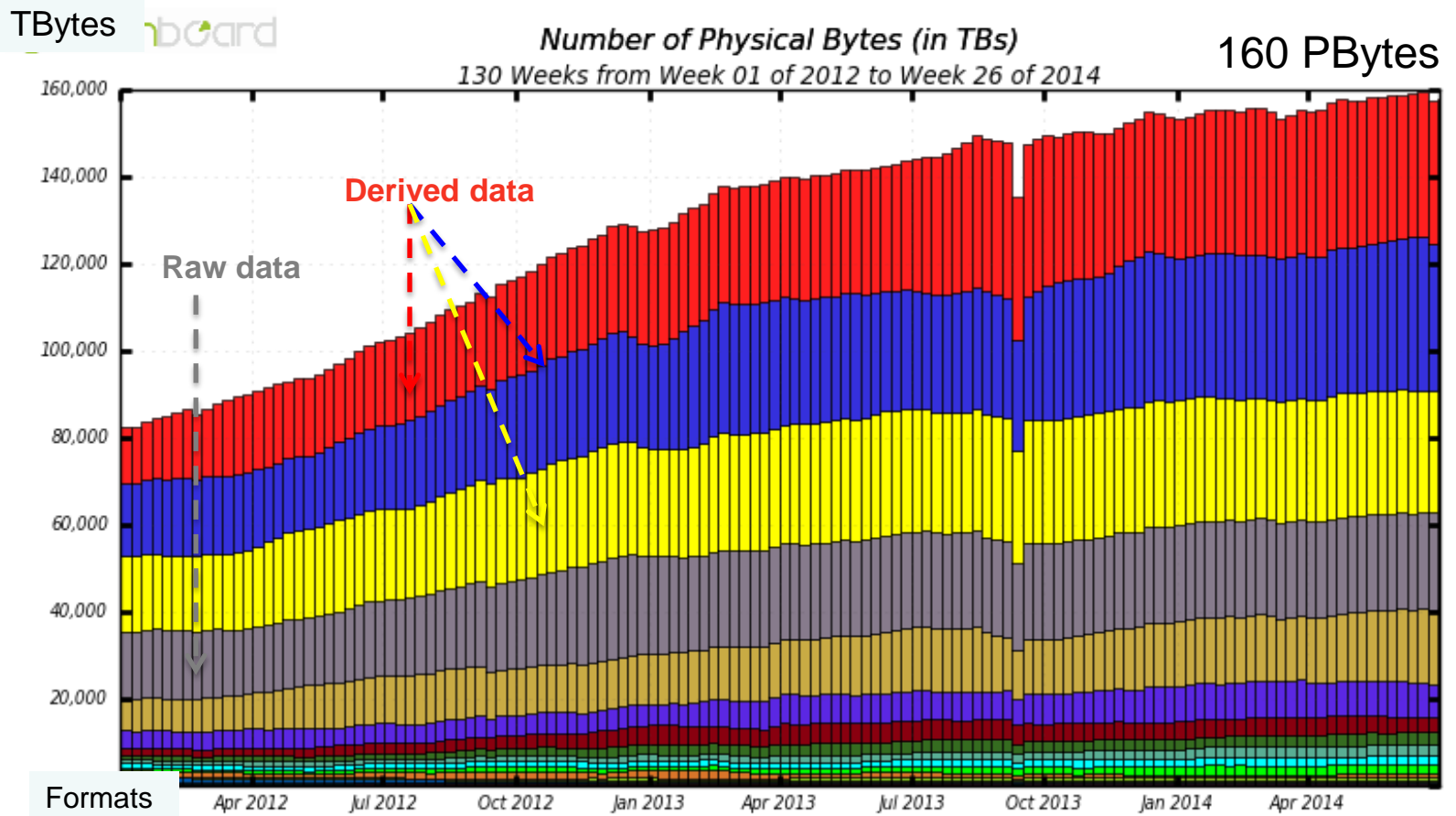
Data Transfer Volume in TB (weekly average). Aug 2012 - Aug 2013

TBytes





# ATLAS Data Volume on the Grid Sites



RAW data volume ~3 PB/year,  
Managed Data Volume on Grid sites 160 PB

# Processing the Experiment Big Data

- **The simplest solution in processing LHC data is using data affinity for the jobs**
  - Data is staged to the site where the compute resources are located and data access by analysis code from local, site-resident storage
  - However
    - In distributed computing environment we don't have enough disk space to host all our data at every Grid site
      - Thus we distribute (pre-place) our data across our sites
    - The popularity of data sets is difficult predict in advance
      - Thus computing capacity at a site might not match the demand for certain data sets
  - Different approaches are being implemented
    - Dynamic or/and on demand data replication
      - Dynamic : if certain data is popular over make additional copies on other Grid sites
      - On-demand : User can request local or additional data copy
    - Remote access : data can be accessed remotely
    - Both approaches have the underlying scenario that puts the WAN between the data and the executing analysis code
- **Intelligent Workload Management System is needed to manage resources and to automate data processing, analysis and simulation**



# Workload Management System. Core Ideas

- **Make hundreds of distributed sites appear as local**
  - Provide central queue for users – similar to local batch systems
- **Reduce site related errors and latency**
  - Build a pilot job system – late transfer of user payloads
  - Crucial for distributed infrastructure maintained by local experts
- **Hide middleware while supporting diversity and evolution**
  - WMS interacts with middleware – users see high level workflow
- **Hide variations in infrastructure**
  - WMS presents uniform ‘job’ slots to user
  - Easy to integrate grid sites, clouds, recently HPC sites
- **Use the same system for Monte-Carlo simulation, data processing and users analysis**

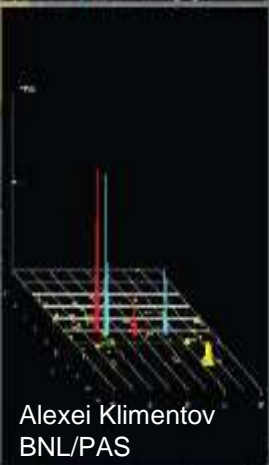




# PanDA. Production and Data Analysis System

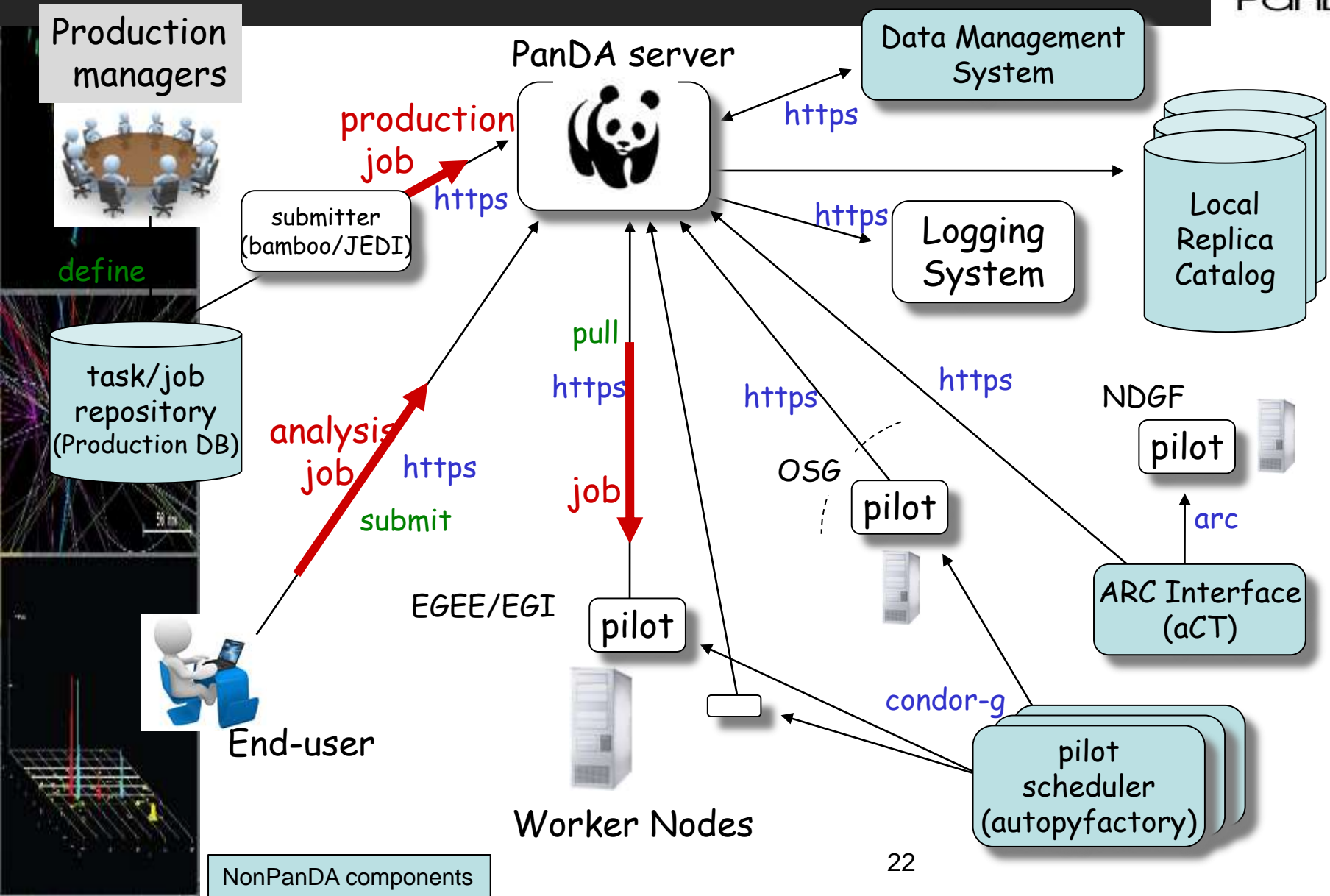


- **ATLAS computational resources are managed by PanDA Workload Management System (WMS)**
- **PanDA project was started in fall of 2005 by BNL and University Texas at Arlington (UTA) groups**
  - An **automated** yet **flexible** workload management system which can **optimally** make **distributed resources** accessible to **all users**
  - Standards based implementation
    - REST framework – HTTP/S
    - Oracle or MySQL backends
    - About a dozen Python packages available from SVN and GitHub
    - Command-line and GUI/Web interfaces
- **Through PanDA, physicists see a single computing facility that is used to run all data processing for the experiment, even though data centers are physically scattered all over the world.**
- **Now successfully manages  $O(10^2)$  sites,  $O(10^5)$  cores,  $O(10^8)$  jobs per year,  $O(10^3)$  users**



Alexei Klimentov  
BNL/PAS

# Workload Management System

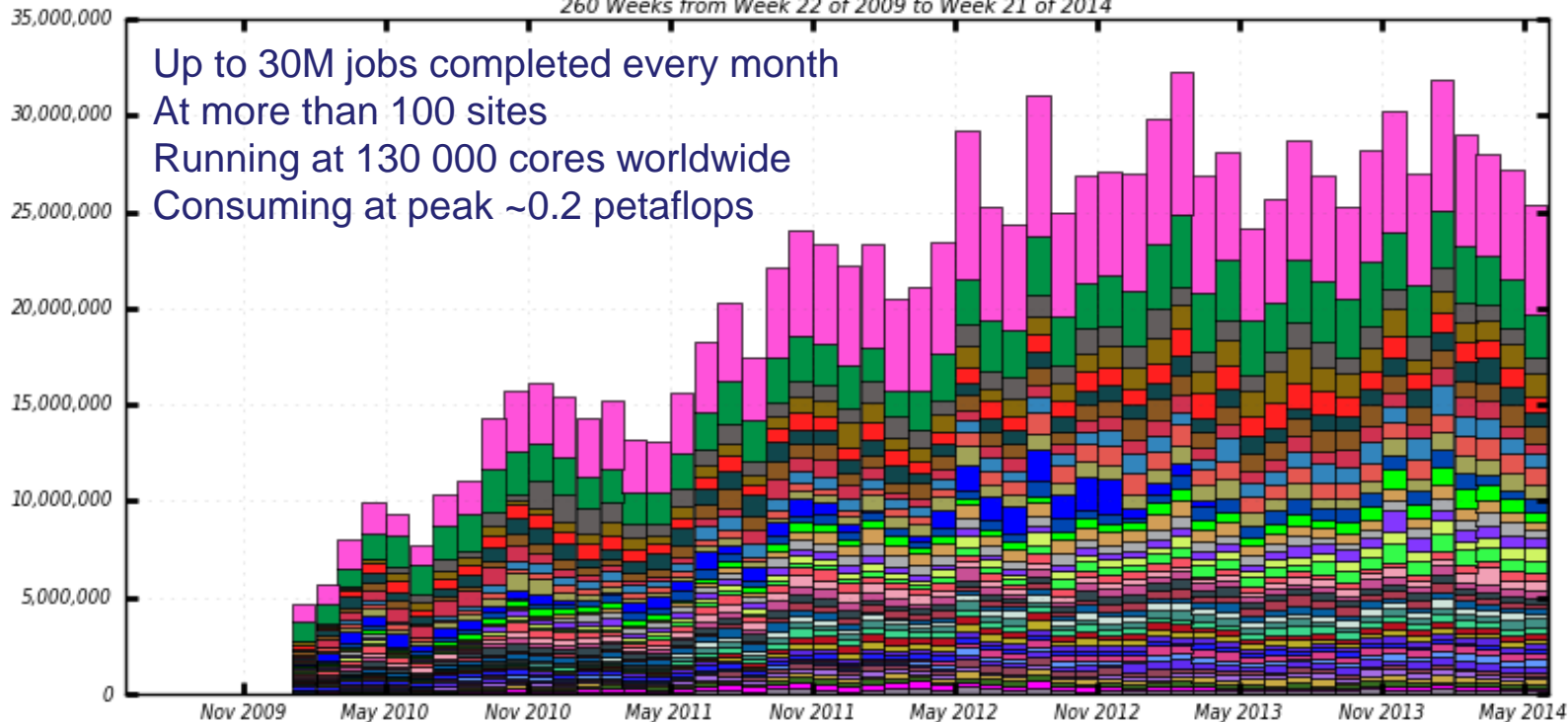


# PanDA Performance

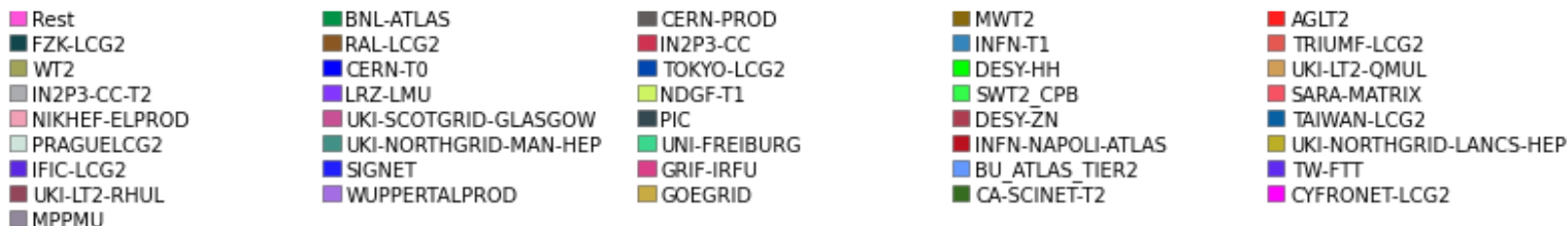


Completed jobs

260 Weeks from Week 22 of 2009 to Week 21 of 2014



*PanDA is exascale now : 1.2 Exabytes of data processed by PanDA in 2013*



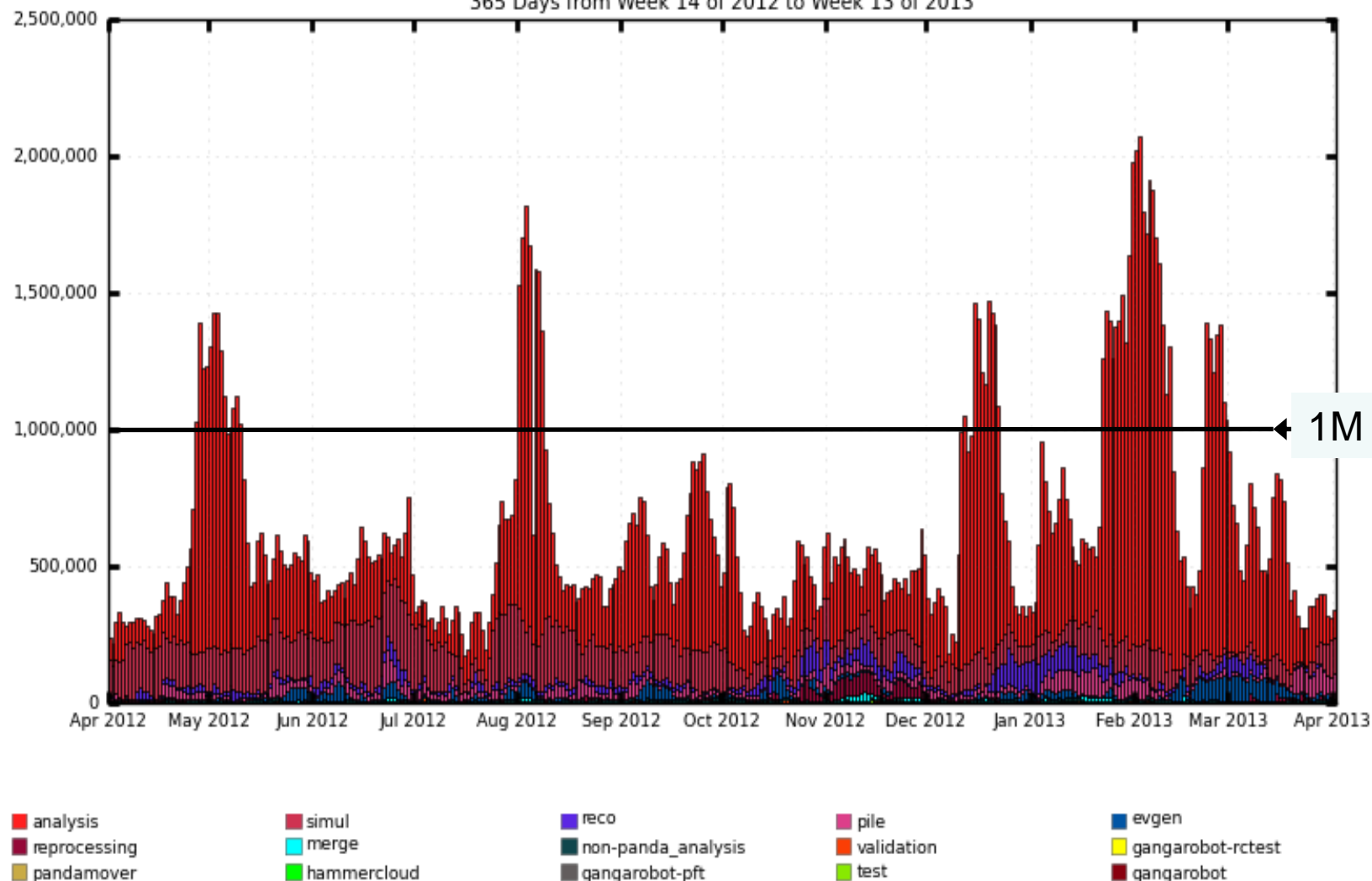
Maximum: 32,306,592 , Minimum: 0.00 , Average: 20,857,355 , Current: 25,377,087

**All available resources are fully used**



# Pending jobs

365 Days from Week 14 of 2012 to Week 13 of 2013



Spikes in demand for computational resources  
Can significantly exceed available ATLAS Grid resources  
Lack of resources slows down pace of discovery

# LHC Upgrade. Computing Needs

CPU needs per event

Run1 :  
2009 - 2013

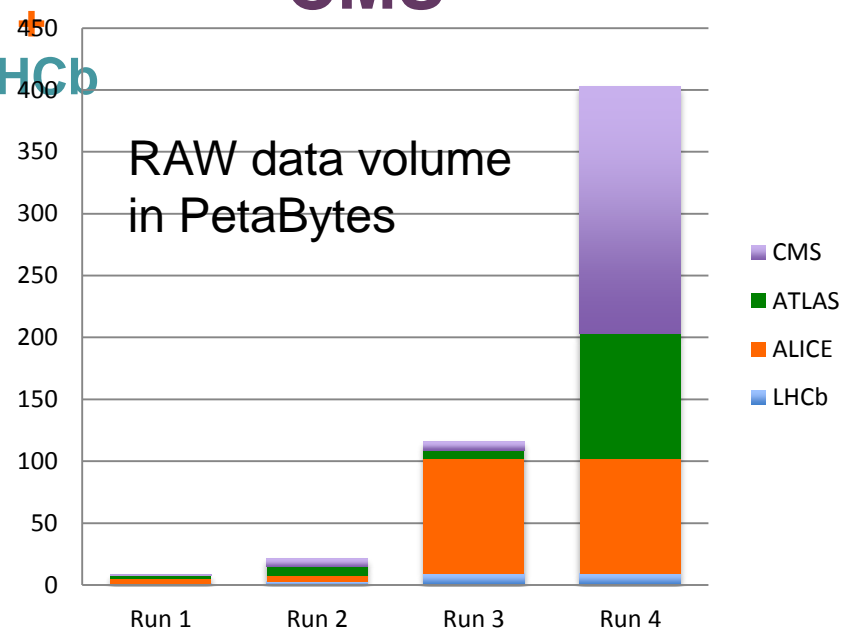
Run2 :  
2015 - 2017

Run3  
2019-2021

Run4  
**ATLAS**  
+  
**CMS**

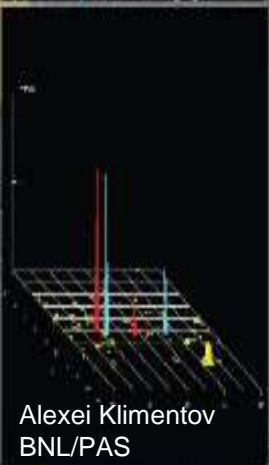
**LHCb**

- 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



# ATLAS Scale of Needs

- The ATLAS experiment uses a geographically distributed grid of approximately 130,000 cores continuously, to simulate, and analyse its data.
- The need for simulation and analysis would overwhelm the expected capacity of LHC Grid computing facilities unless the range and precision of physics studies were to be curtailed.
  - Physics requires to increase rate
    - Run1 data-taking rate 400 Hz
    - Run2 data-taking rate 1kHz
- Leadership Computing Facilities contributions of the order of 10 Million or more core hours per year become important and valuable.
- ATLAS computing can also be a close to ideal “crack-filling” application. PanDA WMS is being upgraded to make it aware of dynamically changing resources, and thus able to exploit groups of processors that become available for relatively short times.
- Extending PanDA beyond the Grid will further expand the potential user community and the resources available to them.







- **There are five dimensions to evolution of PanDA**
  - Making PanDA available beyond ATLAS and High Energy Physics
  - Extending beyond Grid (Leadership Computing Facilities, Clouds, University clusters)
  - Integration of network as a resource in workload management
  - Integration of data management and workload management
  - Using modern technologies such as non-relational databases

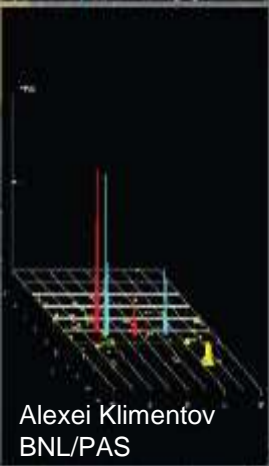


# BigPanDA. Extending the scope. Cloud Computing.



**Google** Compute Engine (GCE) preview project : *Google allocated additional resources for ATLAS for free : ~5M CPU hours, 4000 cores for about 2 months*

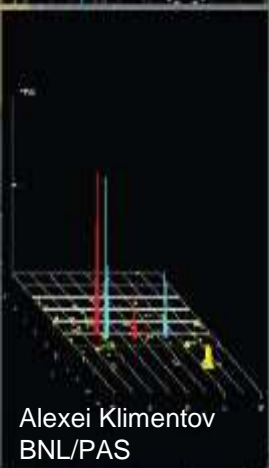
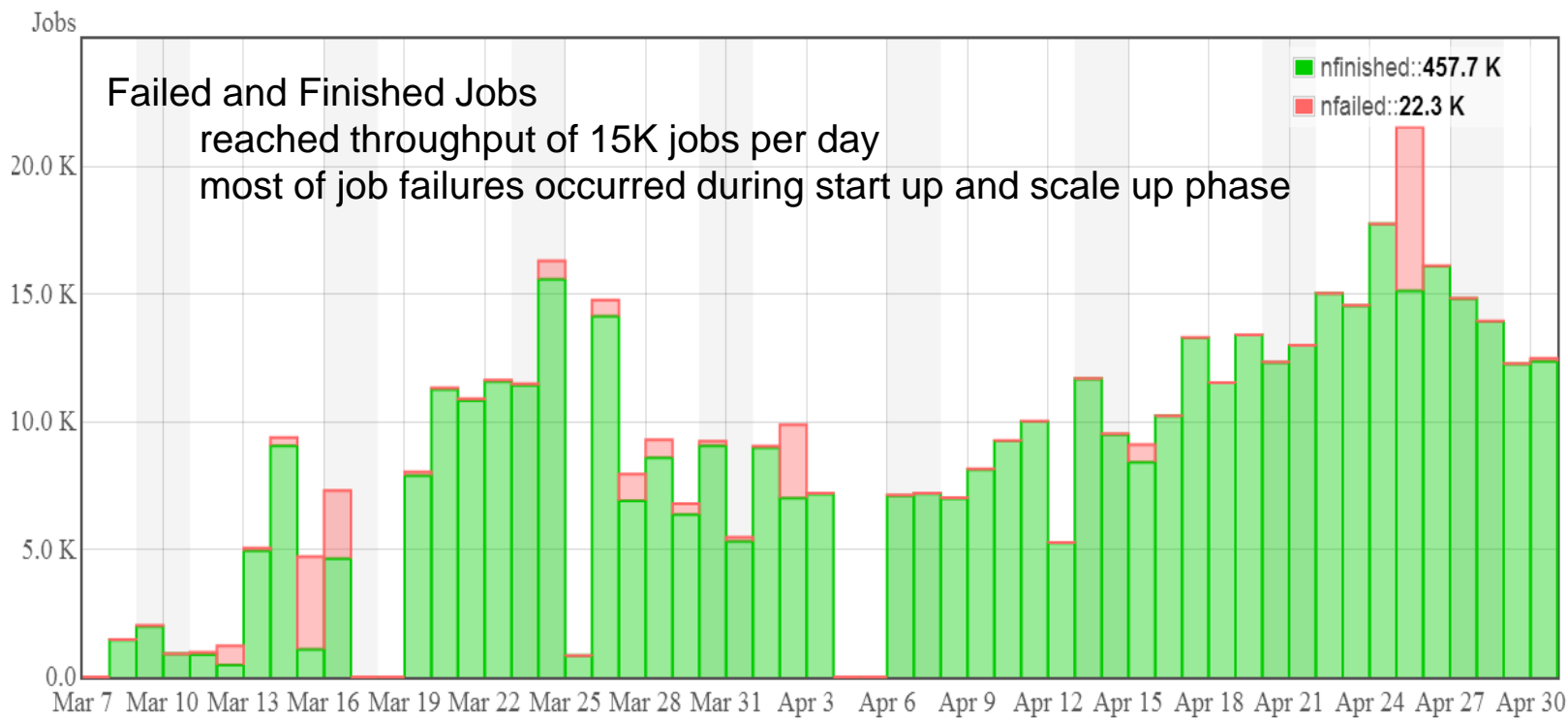
- ❑ Resources are organized as PanDA queues
- ❑ Output exported to BNL
- ❑ Transparent inclusion of cloud resources into ATLAS Distributed Computing
- ❑ The idea was to test long term stability while running a cloud cluster similar in size to Tier 2 center in US ATLAS
- ❑ Intended for CPU intensive Monte-Carlo simulation workloads, planned as a production type of run. Delivered to ATLAS as a resource and not as an R&D platform.



# Running PanDA on Google Compute Engine



- We ran for about 8 weeks (2 weeks were planned for scaling up)
- Very stable running on the Cloud side. GCE was rock solid.
- Most problems that we had were on the ATLAS side.
- We ran computationally intensive jobs
  - Physics event generators, Fast detector simulation, Full detector simulation
- Completed 458,000 jobs, generated and processed about 214 M events
- Invited BNL PAS talk at Google IO conference





# BigPanDA for Leadership Computing Facilities

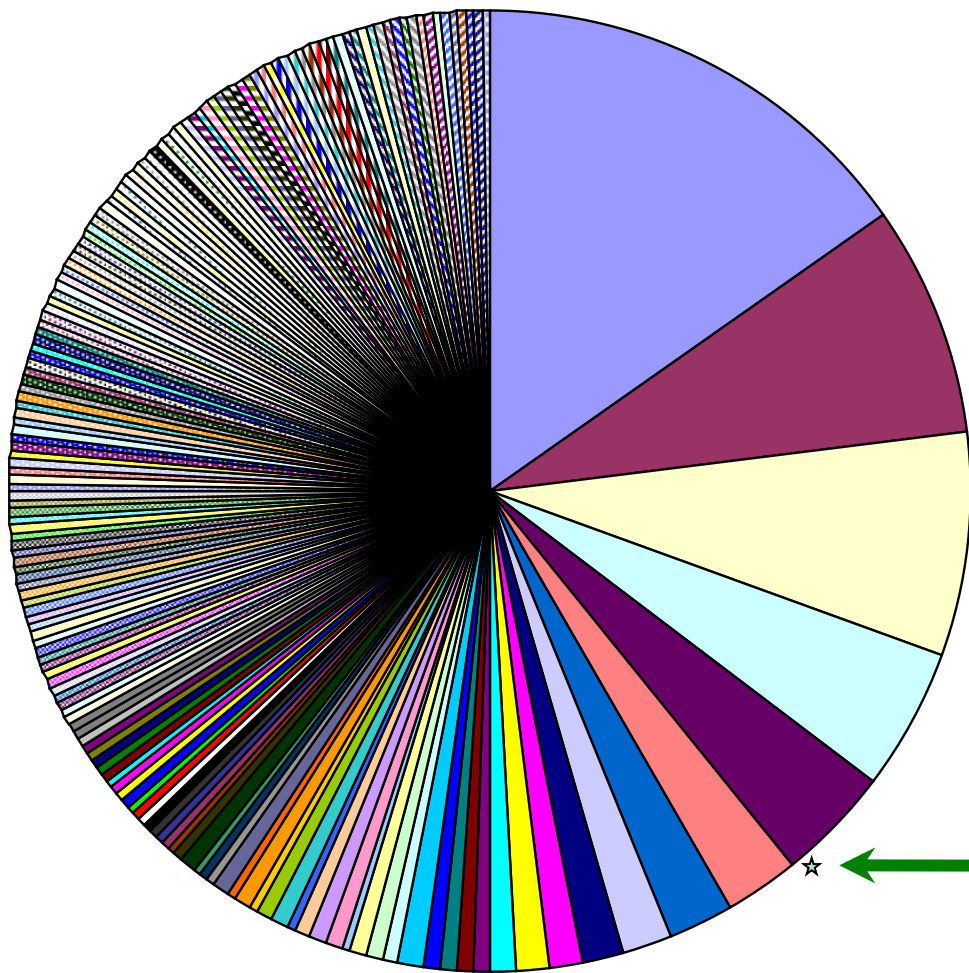


- **Expanding PanDA from Grid to Leadership Class Facilities required significant changes in our system**
  - Each LCF is unique
    - Unique architecture and hardware
    - Specialized Operating System, “weak” worker nodes, limited memory per worker node
    - Code cross-compilation is typically required
    - Unique job submission systems
    - Unique security environment
- **HEP applications (such as Geant or ROOT) can effectively use a single core**
- **PanDA has potential to generate 300M hours per year**
- **PanDA submission algorithm has been adopted to use backfill information and to submit jobs in backfill mode.**
- **Final tests have been conducted in August together with OLCF team**
  - Were able to collect ~ 200,000 core hours
  - Max number of nodes per job – 5835 (93360 cores)
    - Close to 75% ATLAS Grid in size!

***Used ~2.3% of all Titan core hours or ~14.4% of free core hours***



# The Top 500 Computers



- **Most of the computational power is concentrated in a small number of machines**
  - Half the total power is in the top dozen computers
- **Equivalent ATLAS Grid use is about the size of the sector (little green star)**
- **Equivalent ATLAS Grid would be around #27 on this chart**

# High Performance Computing (Top 10, Nov 2013)

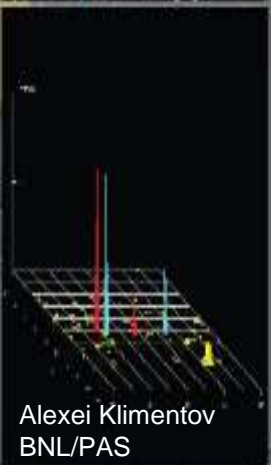
			cores	Rmax	Rpeak	Power	
1	National Super Computer Center in Guangzhou China	<b>Tianhe-2 (MilkyWay-2)</b> - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3120000	33862.7	54902.4	17808	
2	DOE/SC/Oak Ridge National Laboratory United States	<b>Titan</b> - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560640	17590.0	27112.5	8209	★
3	DOE/NNSA/LLNL United States	<b>Sequoia</b> - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1572864	17173.2	20132.7	7890	
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705024	10510.0	11280.4	12660	
5	DOE/SC/Argonne National Laboratory United States	<b>Mira</b> - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786432	8586.6	10066.3	3945	★
6	Swiss National Supercomputing Centre (CSCS) Switzerland	<b>Piz Daint</b> - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Cray Inc.	115984	6271.0	7788.9	2325	★
7	Texas Advanced Computing Center/Univ. of Texas United States	<b>Stampede</b> - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi SE10P Dell	462462	5168.1	8520.1	4510	★
8	Forschungszentrum Juelich (FZJ) Germany	<b>JUQUEEN</b> - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	458752	5008.9	5872.0	2301	
9	DOE/NNSA/LLNL United States	<b>Vulcan</b> - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	393216	4293.3	5033.2	1972	
10	Leibniz Rechenzentrum Germany	<b>SuperMUC</b> - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR IBM	147456	2897.0	3185.1	3423	★

★ Collaborations have members with access to these machines and to ARCHER, NERSC, ... Some HPCs are already successfully used as part of Nordu Grid (Abisko, Abel, Triolith, C2PAP, Hydra)



# HPC Scheduling

- **HEP applications (such as Geant or ROOT) can effectively use a single core**
- **HPC is full, means that the system have allocated all the cycles it is able to deliver**
  - It is probably not all cycles it has
  - Just as there is room for sand in the jar of rocks, there's room for HEP jobs on even a “full” HPC



# HPC Scheduling. Cont'd



- This is not a typical workday view of HPC machine
- At this moment machine is 85% full, the largest open partition has 1024 nodes
- But the shortest job in the queue required 4096 nodes
- The scheduler will be happy to run a short job in R12
- 24h backfill tests have been conducted on Titan together with OLCF team

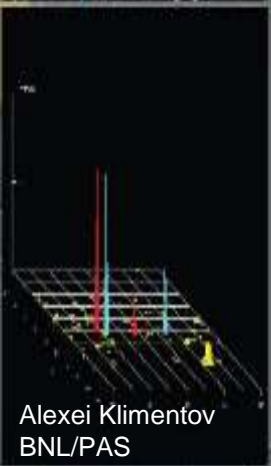


Free nodes



**#2**  **TOP 500<sup>®</sup>**  
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

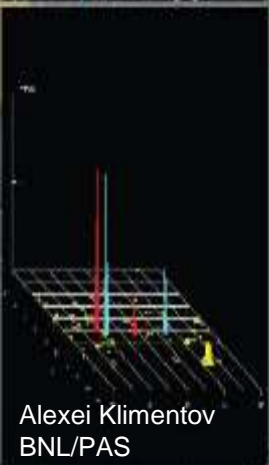




# Interfacing PanDA with Titan



- **BigPanDA project on Titan 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 CERN, BNL, UTA, ORNL, UTK, LBNL, ...**
  - Strong interest from OLCF, took responsibility for MPI wrapper base and docs
- **PanDA has potential to generate 300M hours per year**
- **Technology developed on Titan should be applicable for other HPC centers**
  - Functionality tests have been conducted on NERSC
  - Interest from ASGC, NRC-KI, JINR, Ostrava

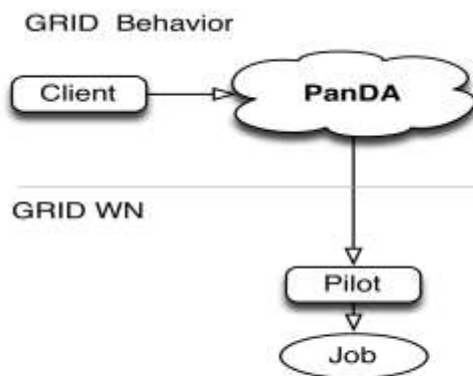


# Interfacing PanDA with Titan. Cont'd

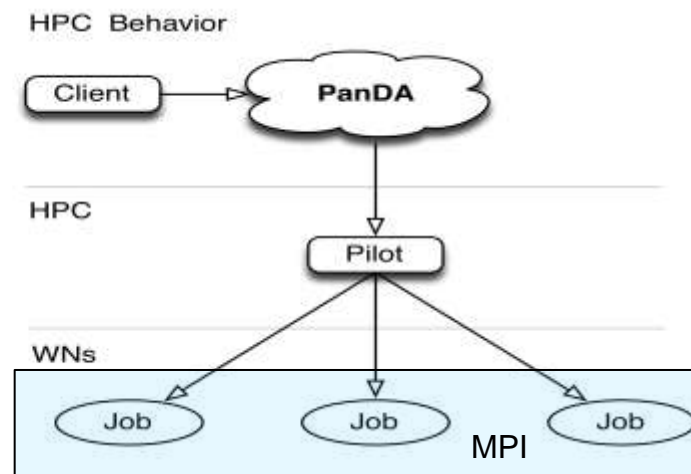


- PanDA modular pilot augmented with HPC specific classes
- SAGA (Simple API for Grid Applications) framework as pilot's interface to HPC batch schedulers
  - <http://saga-project.github.io/saga-python/>
  - <http://www.ogf.org/documents/GFD.90.pdf>
- MPI wrapper/overlay scripts that allow to run multiple single threaded workload instances in parallel
- “Backfill” functionality in pilot

## Pilot on HPC with MPI wrapper

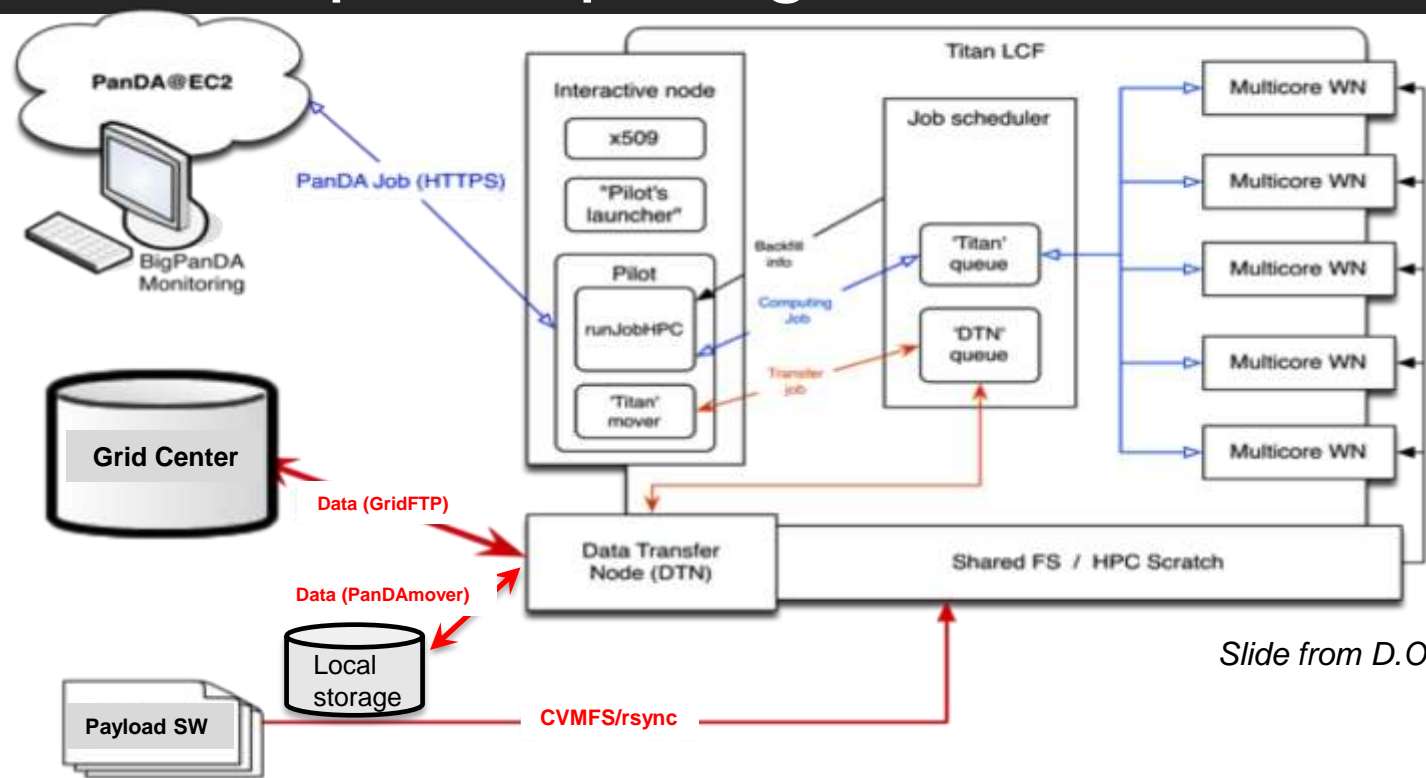


“One to One”



“One to Many”

# Extending PanDA to Oak Ridge Leadership Computing Facilities

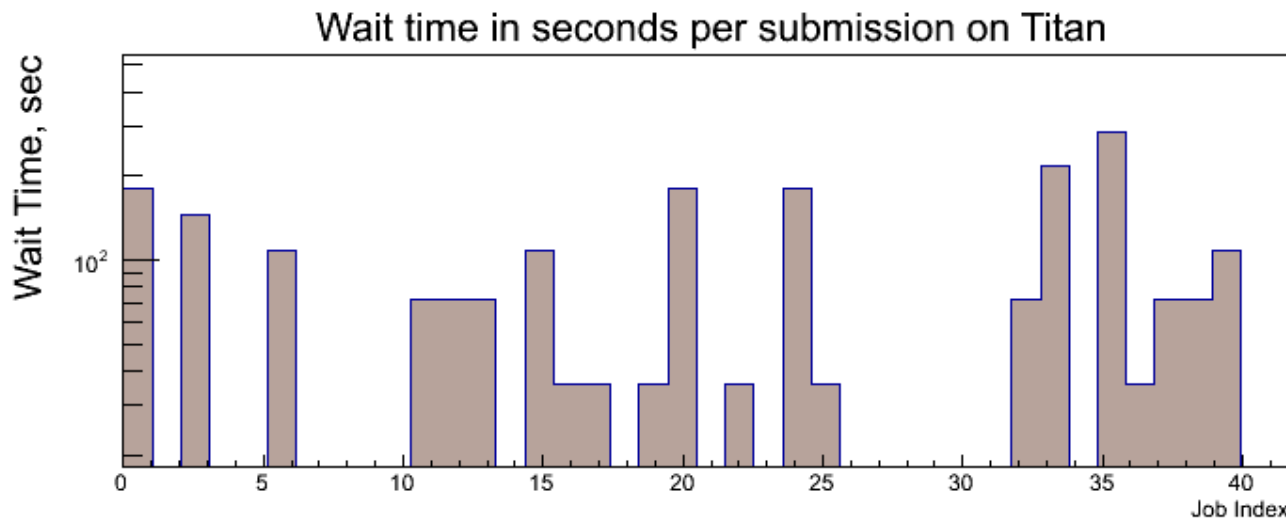
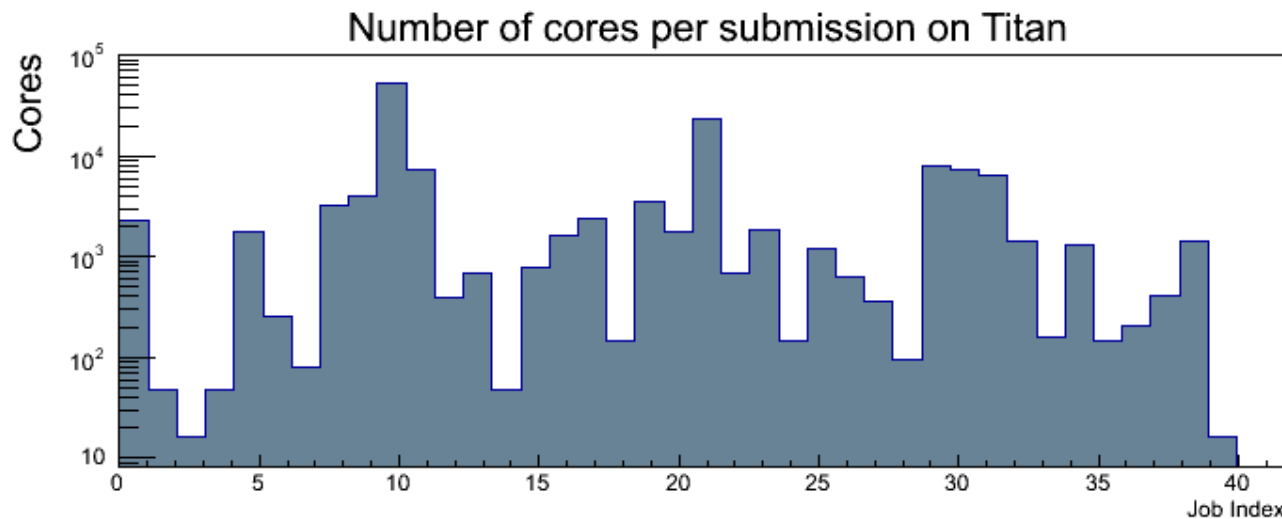


Slide from D.Oleynik

- **ATLAS (BNL, UTA), OLCF, ALICE (CERN,LBNL,UTK) :**
  - adapt PanDA for OLCF (Titan)
  - reuse existing PanDA components and workflow as much as possible.
  - PanDA connection layer runs on front-end nodes in user space. There is a predefined host to communicate with CERN from OLCF, connections are initiated from the front-end nodes
  - SAGA (a Simple API for Grid Applications) framework as a local batch interface.
  - Pilot (payload submission) is running on HPC interactive node and communicating with local batch scheduler to manage jobs on Titan.
  - Outputs are transferred to BNL T1 or to local storage
- **The same architecture has been tested on other super-computers**



# PanDA pilot tests on Titan




Average wait time 70 seconds !

# Resources Accessible via PanDA



Many  
Others



OLCF

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		

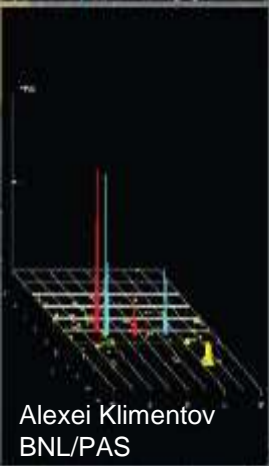
© OLCF/2010

OLCF



# Summary

- **2000s – The decade of the Grid**
- **First years of LHC data – Distributed Computing has helped deliver physics rapidly**
- **Entering a phase of computing evolution**
- **Challenges for computing – scale & complexity – will continue to increase dramatically**
- **The distributed computing model allows us to incorporate clouds and LCF/HPC centers and to use them efficiently for LHC Run 2 and beyond**
- **Access to the LCF coupled with collaborative help in the transformation of HEP code would be a major scientific contribution to the physics discoveries of the next ten years**
- **The work on extending PanDA to Leadership Computing Facilities has started. PanDA has been successfully ported on OLCF Titan and NERSC, and it is underway to the supercomputer at NRC-KI**



# Summary. PanDA's Success



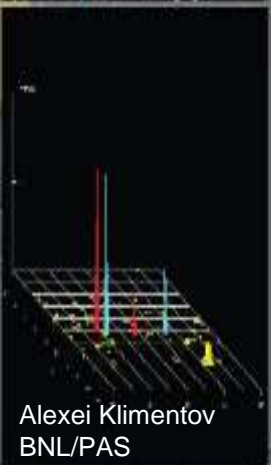
- PanDA was able to cope with increasing LHC luminosity and ATLAS data taking rate
- Adopted to evolution in ATLAS computing model
- Several leading HENP and astro-particle experiments (ALICE, AMS, LSST, COMPASS) has chosen PanDA as workload management system for data processing and analysis or evaluating it.
- It is also used for Biomedical applications
- Cost effectiveness. PanDA enables BigData processing at a fraction of the cost at Google/Amazon data centers

***PanDA is exascale now : 1.2 Exabytes of data processed by PanDA in 2013***





# Backup slides



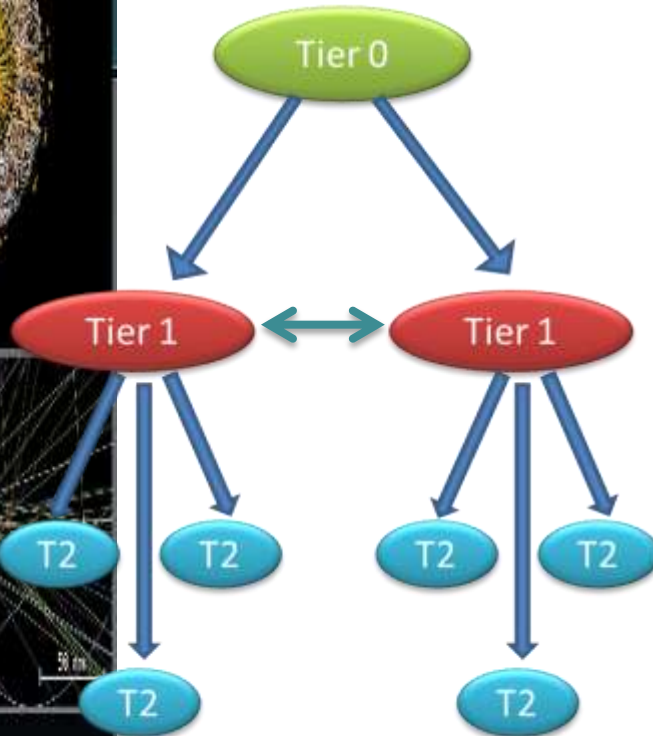
Alexei Klimentov  
BNL/PAS

9/2/2014

ACAT 2014

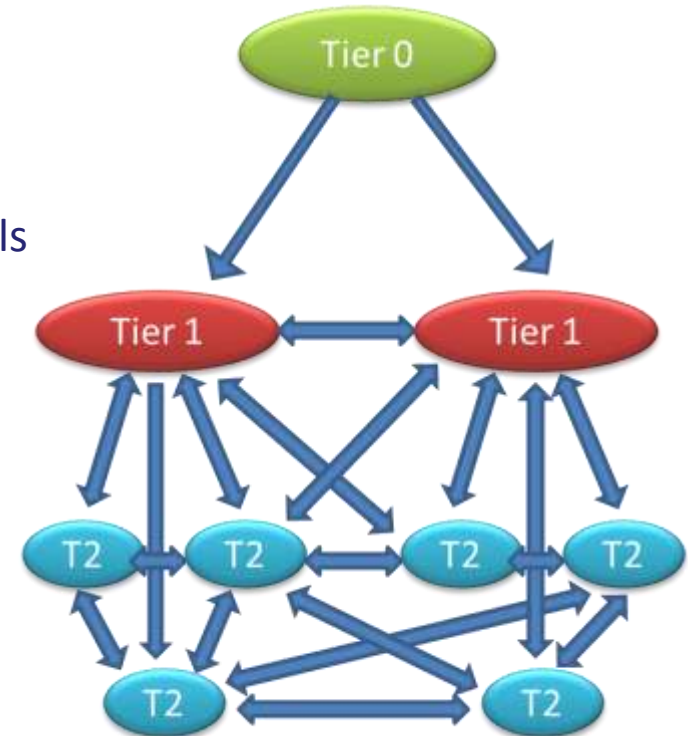
43

# LHC Computing Model Evolution. Tiers Hierarchy



Hierarchy

Evolution of  
computing models

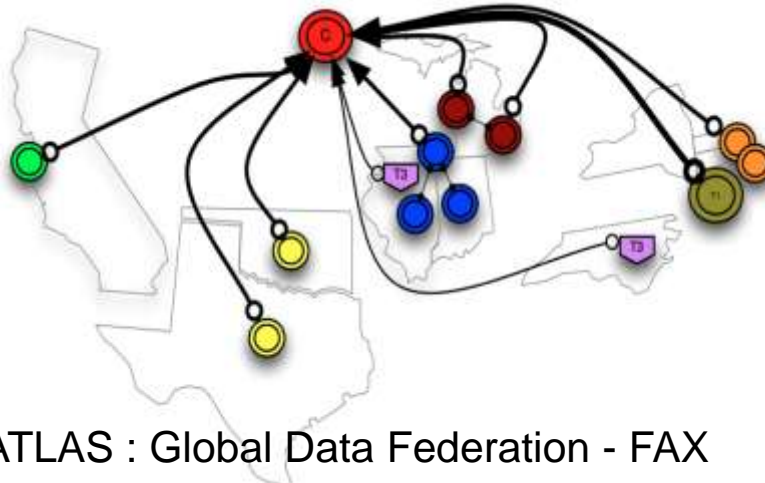


Mesh

- Network capabilities and data access technologies have significantly improved our ability to use resources independent of location
- Now we are relaxing hierarchical model : Flat instead of Tiered Grid model

# LHC Computing Model Evolution.

## Global Data Federations... (Reducing Complexity)



ATLAS : Global Data Federation - FAX

### ALICE :

- Virtually joining together the sites based on proximity (latency) and network capacity into Regional Data Clouds
- Each cloud/region provides reliable data management and sufficient processing capability
- Dealing with handful of clouds/regions instead of the individual sites

CERN-IT : site or cloud is represented by Virtual Storage Cloud Node

