Atlas Analysis On Grid

Sergey Panitkin

for the ATLAS Collaboration

Outline

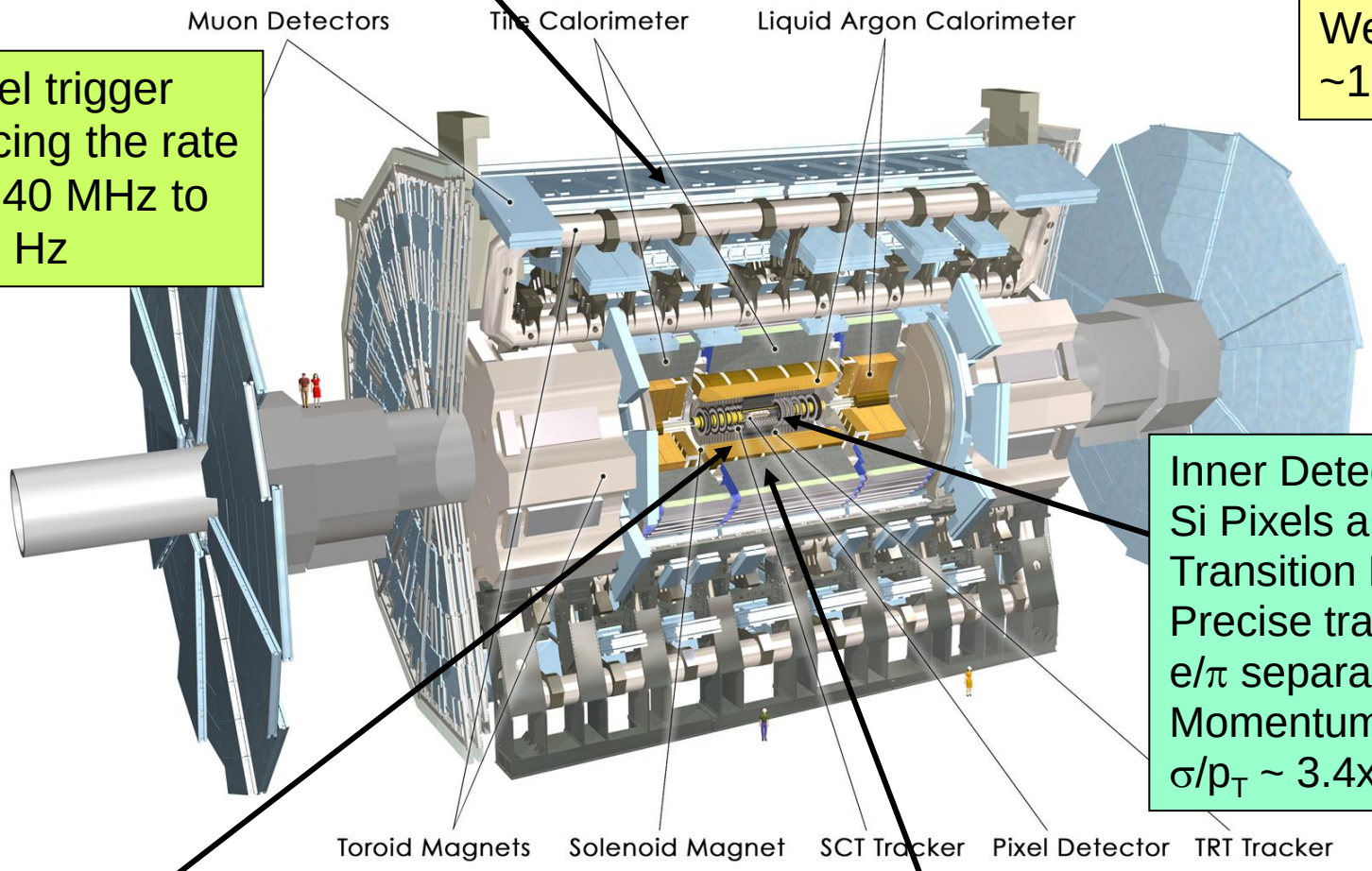
- Brief overview of the (initial) ATLAS Computing Model
  - Some of the key concepts:
    - Tiers of ATLAS, site roles and responsibilities
    - Event data model
    - Grid workload management (Panda)
- Some of the key metrics:
  - Distributed User Analysis
  - Data format popularity
- User Support
- Tier 3 in ATLAS
- Data placement model eveolution

Acknowledgements: J. Shank, A.Klimentov,
D. Rousseau ,D. van der Ster

Muon Spectrometer ($|\eta|<2.7$) : air-core toroids with gas-based chambers
Muon trigger and measurement with momentum resolution < 10% up to E_μ ~ TeV

Length : ~ 46 m
Radius : ~ 12 m
Weight : ~ 7000 tons
~10^8 electronic channels

3-level trigger reducing the rate from 40 MHz to ~200 Hz

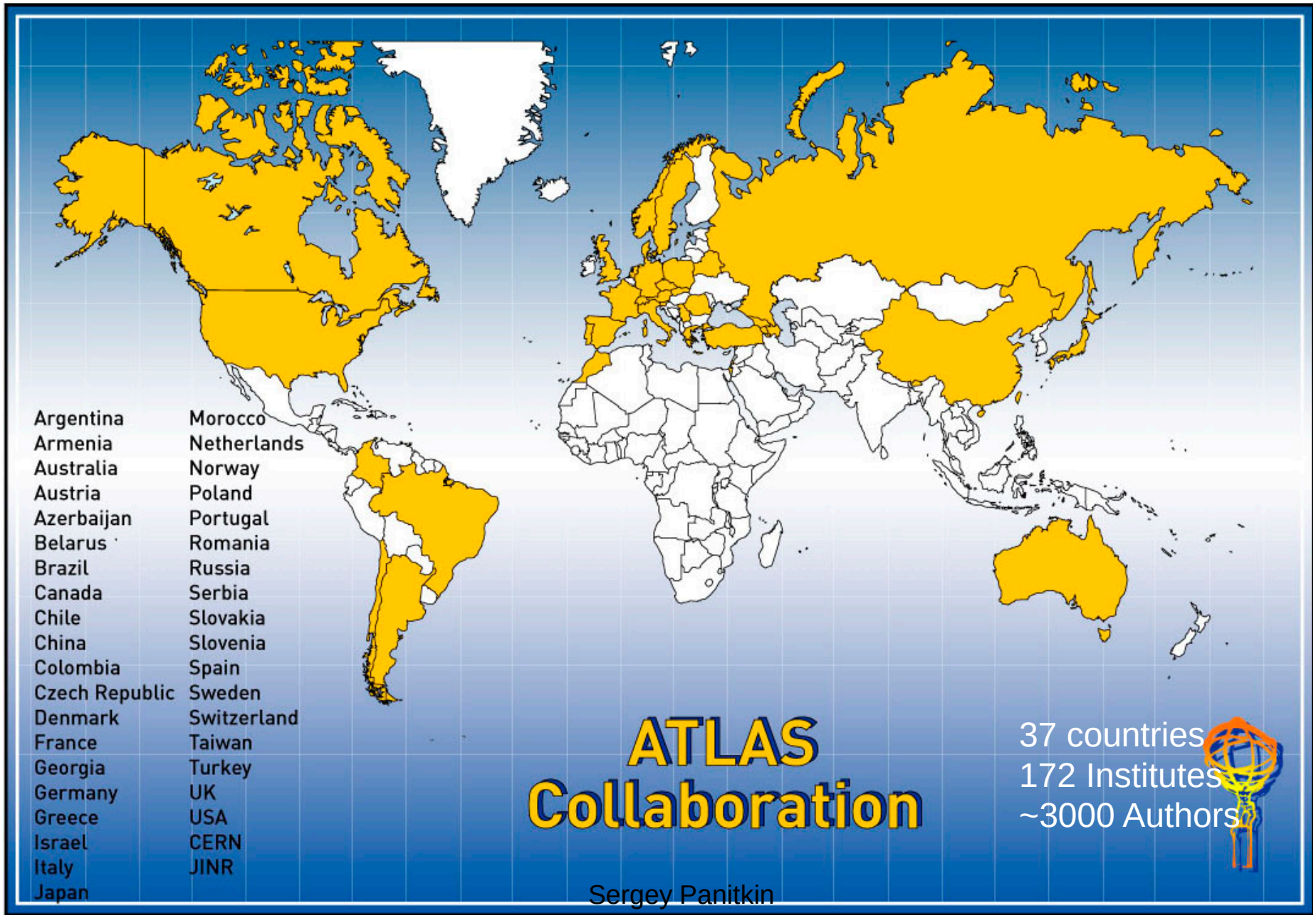


Inner Detector ($|\eta|<2.5$, B=2T):
Si Pixels and strips (SCT) +
Transition Radiation straws
Precise tracking and vertexing,
e/π separation (TRT).
Momentum resolution:
$\sigma/p_T \sim 3.4\times10^{-4}\ p_T\,(\mathrm{GeV}) \oplus 0.015$

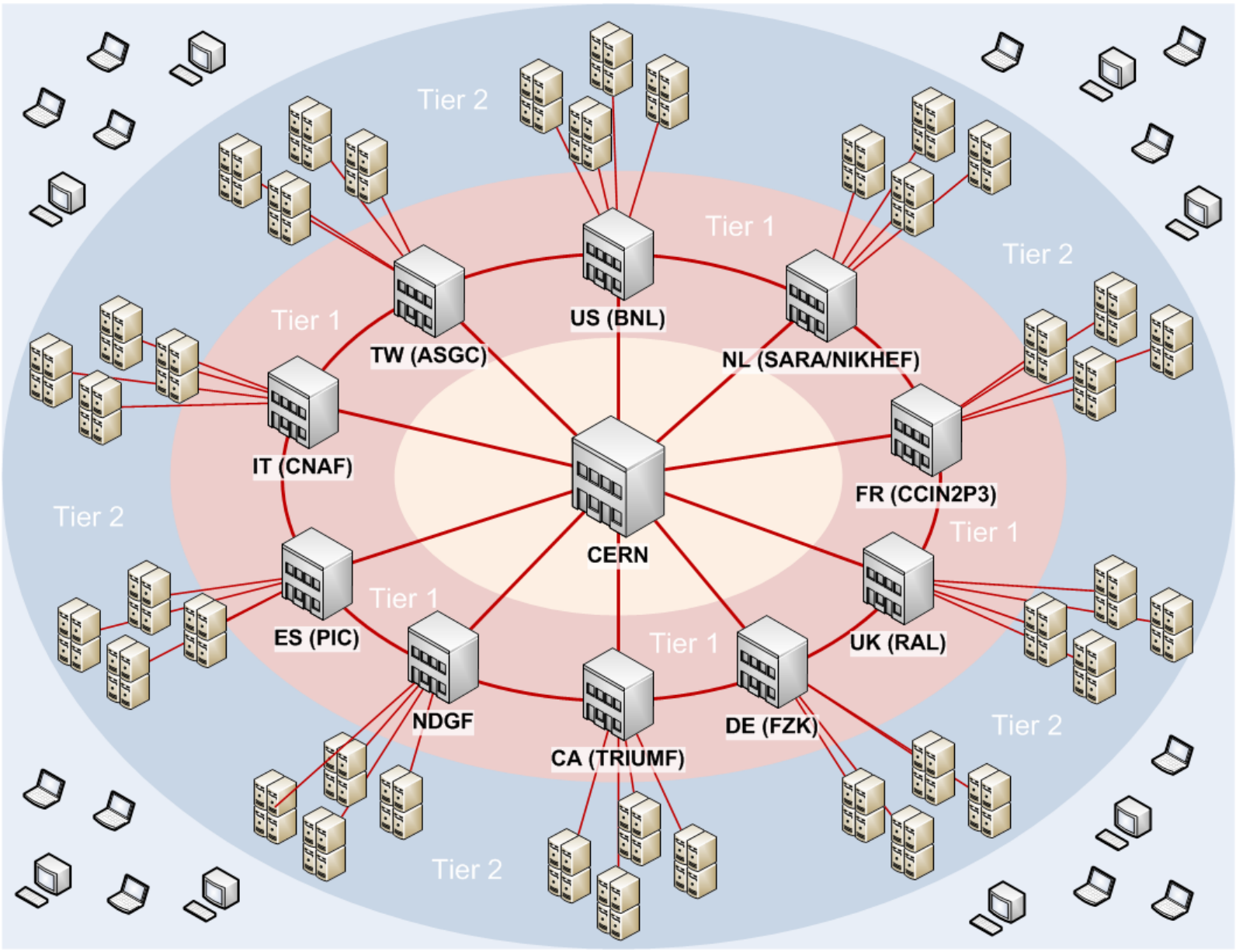
EM calorimeter: Pb-LAr Accordion
e/γ trigger, identification and measurement
E-resolution: ~ 1% at 100 GeV, 0.5% at 1 TeV

HAD calorimetry ($|\eta|<5$): segmentation, hermeticity
Tilecal Fe/scintillator (central), Cu/W-LAr (fwd)
Trigger and measurement of jets and missing E_T
E-resolution: $\sigma/E \sim 50\%/\sqrt{E} \oplus 0.03$

The ATLAS Collaboration



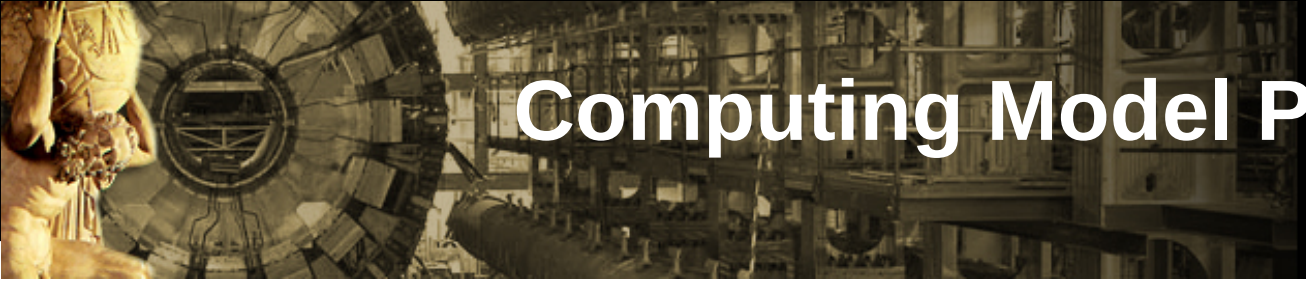
ATLAS Distributed Computing



Tiered, hierarchical model

Computing Tasks per Tier

- **Tier-0 (CERN)**
  - RAW Detector Data Acquisition and archive to tape
  - Calibration and Alignment
  - First processing
  - Data distribution to Tier-1's
- **Tier-1's (10 big Computer Centers)**
  - One Tier-1 at the head of each *cloud*
  - Archive a share of the RAW Detector Data to tape (2^{nd} copy)
  - Re-process those data when needed (new software, new calibration)
  - Archive Simulated data to tape and reconstruct when needed
  - Bulk analysis jobs but also user analysis in some cases
  - Data distribution to Tier-2's
- **Tier-2's (~60 mid size computer centers)**
  - Many attached to a Tier-1 to form a cloud
  - Simulation Production
  - User analysis
- **Tier-3's (100 (?) home institutes, faculty facilities)**
  - End user analysis
  - None pledged resources; Not under ATLAS control

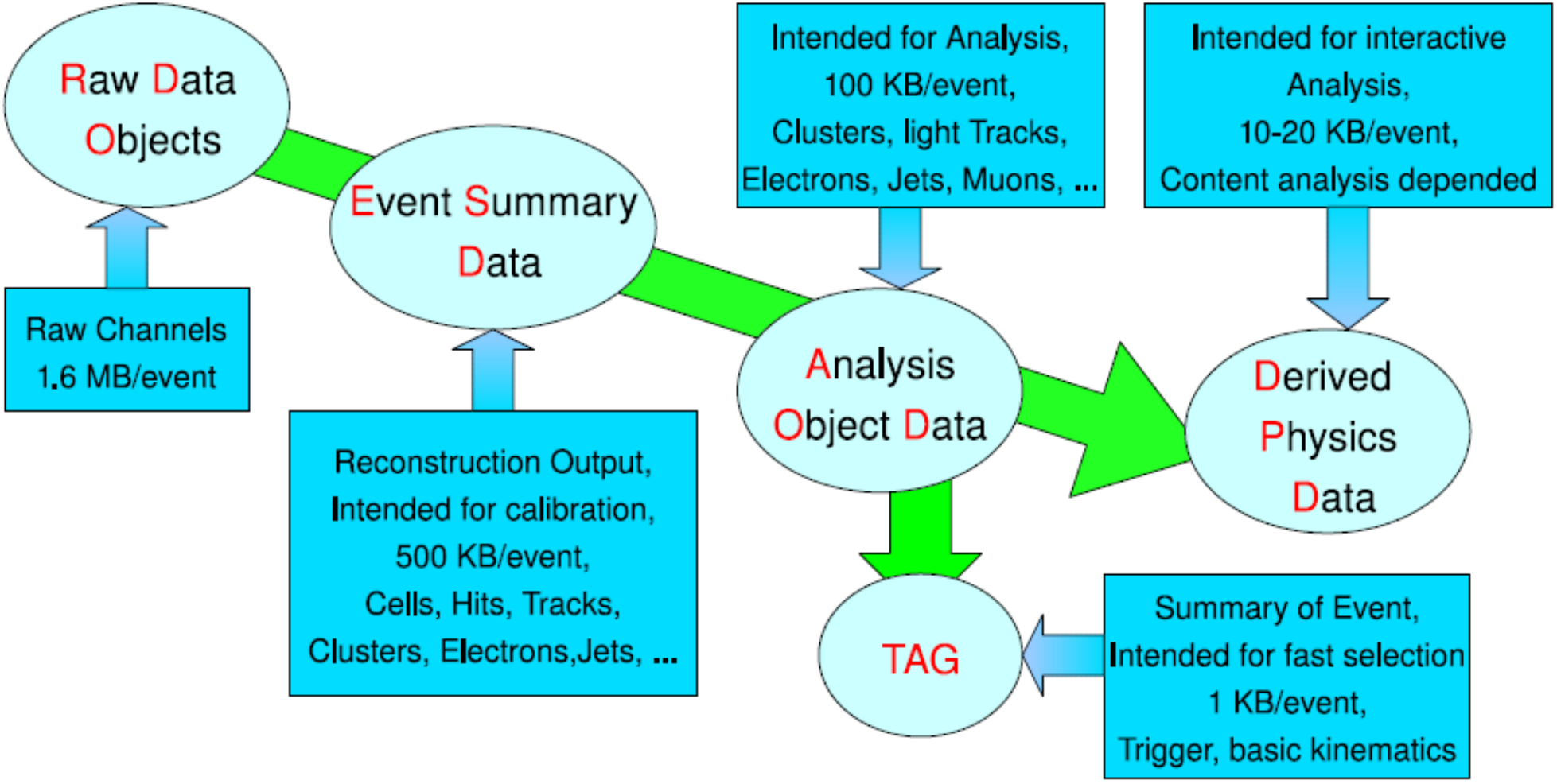


Computing Model Principles

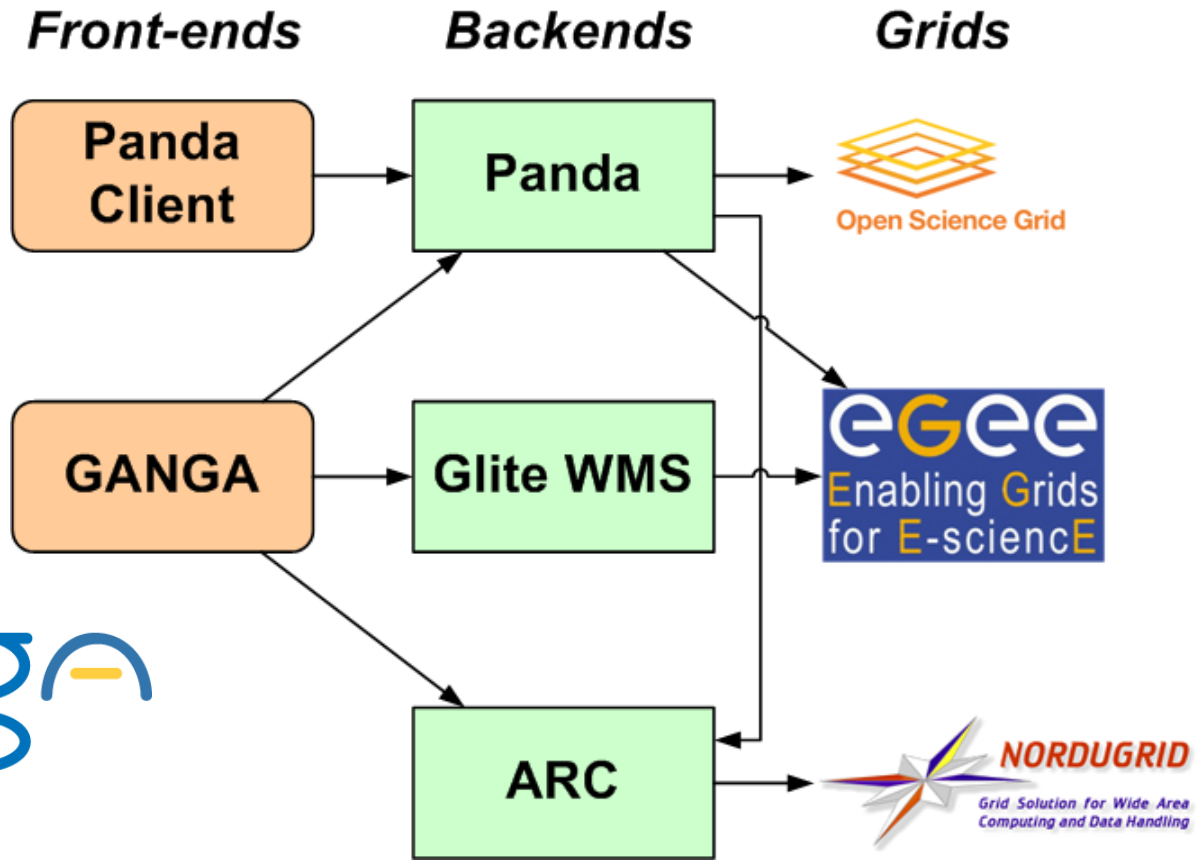
- RAW data master copy stored at CERN
- RAW data distributed over all Tier-1's
  - Tier-1 is responsible for preserving data on tape
  - And recall it for re-processing
- Cloud independence: All derived data available in each cloud
  - Generally, there should be a cloud with free CPU's
  - Generally, data should not have to move between clouds
- All data is pre-placed in each cloud
  - For controlled processing in Tier-1's
  - For user analysis in Tier-2's
- New data produced in a cloud should be archived there
  - Only Tier-1's are required to have tape archives
  - Also true for the Tier-0 (CERN)

ATLAS Event Data Model

Refining the data by: Add higher level info, Skin, Thin, Slim

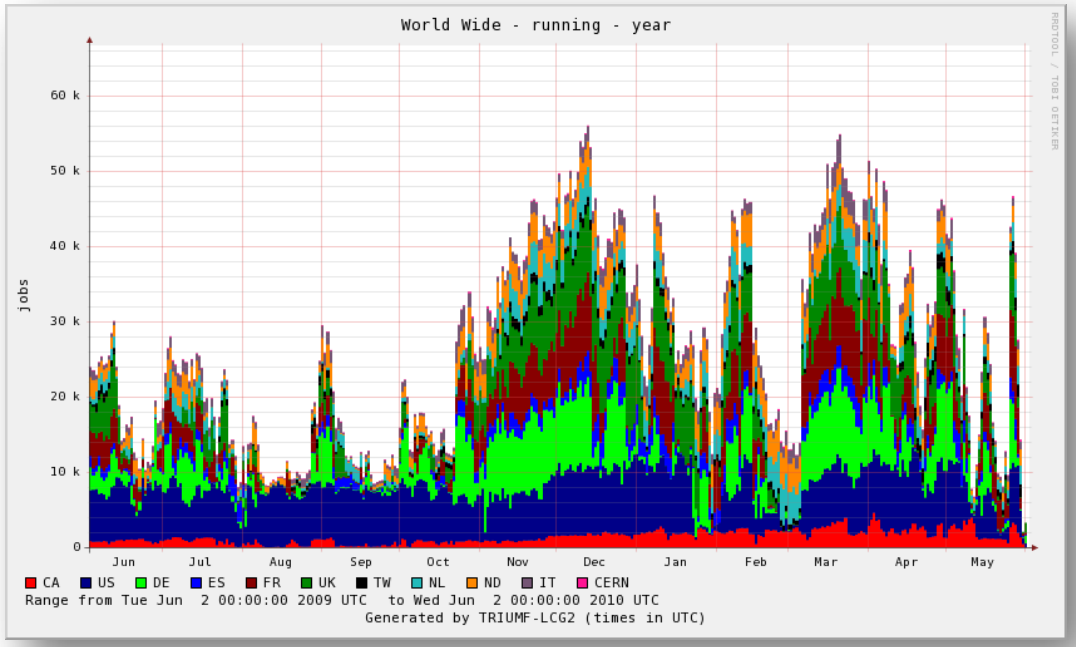


Distributed Analysis



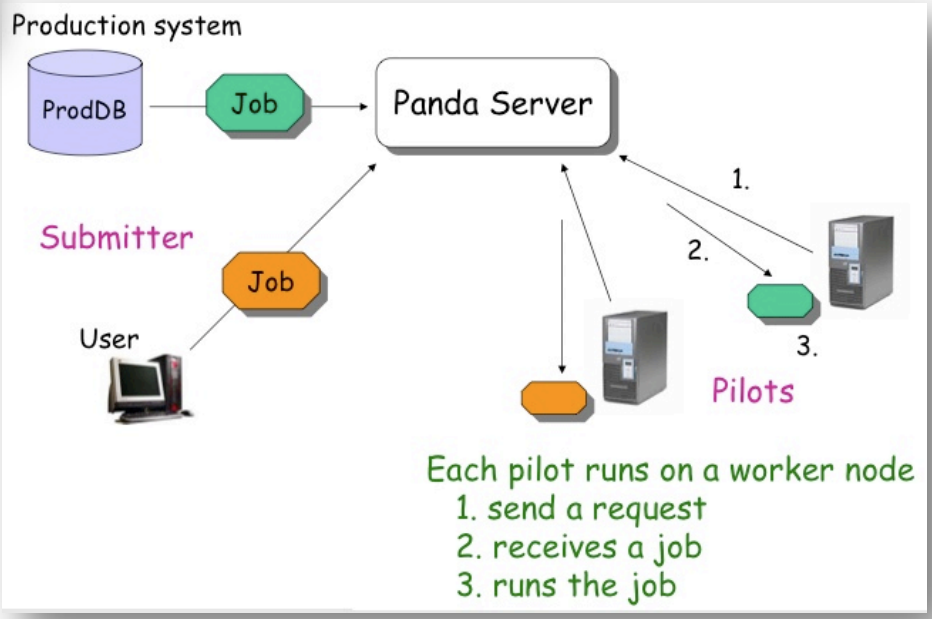
- Basic model: Data is pre-distributed to the sites, jobs are brokered to a site having the data
- Large dataset containers are distributed across clouds, so the front-ends do not restrict jobs to a cloud. i.e. DA jobs run anywhere in the world.

Workload Management: PanDA



- Pilot based Grid workload management system
- Initially used in the US cloud
- PanDA@CERN deployed >2 year ago and is running successfully.
- PanDA is used to run all MC and Reprocessing, and ~90% of the user analysis worldwide

- Panda load depends more on the number of resources (~70 sites), and less so with the amount of data
- Panda provides excellent build-in bookkeeping and monitoring tools – Important for data analysis on the Grid

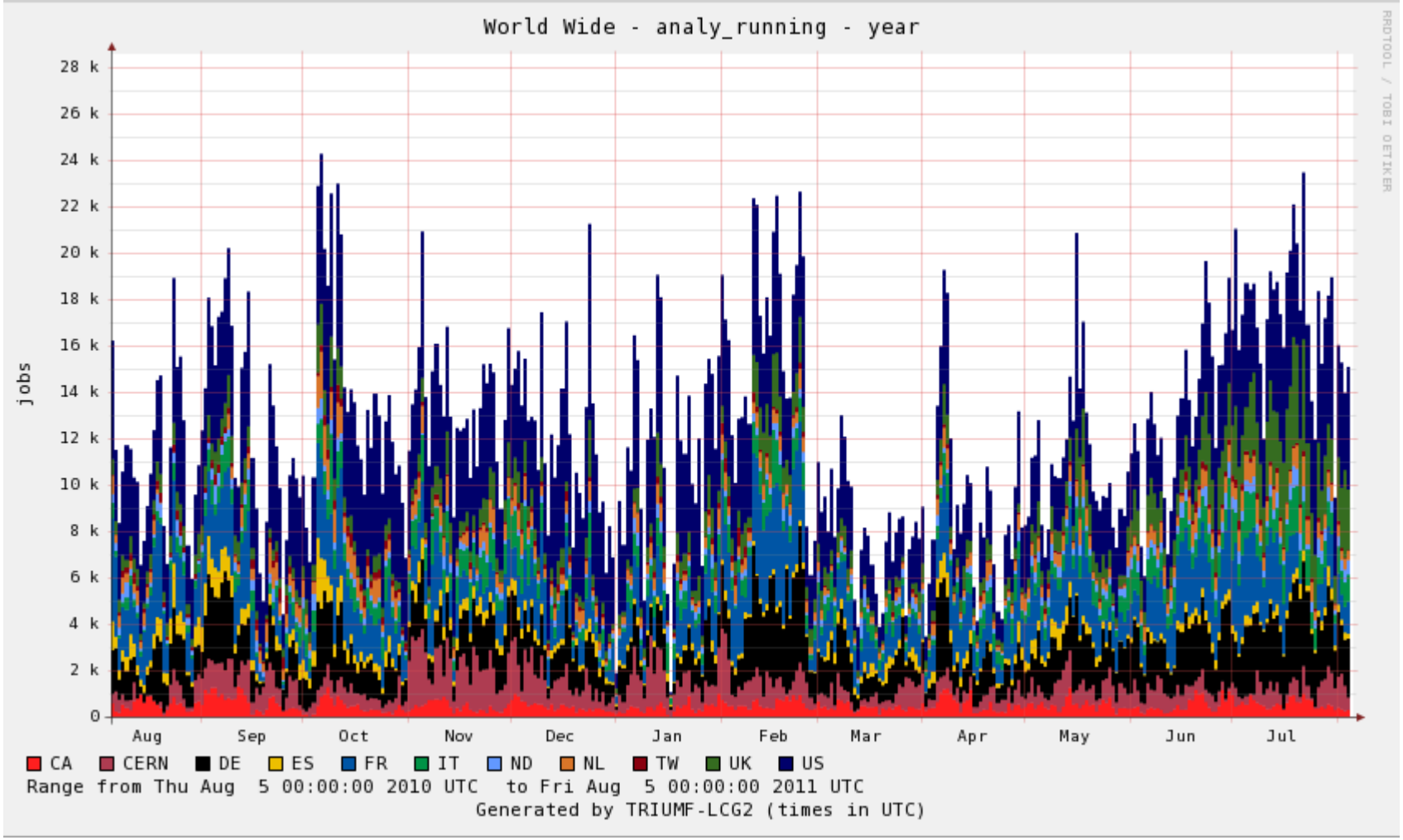


Atlas sw in numbers

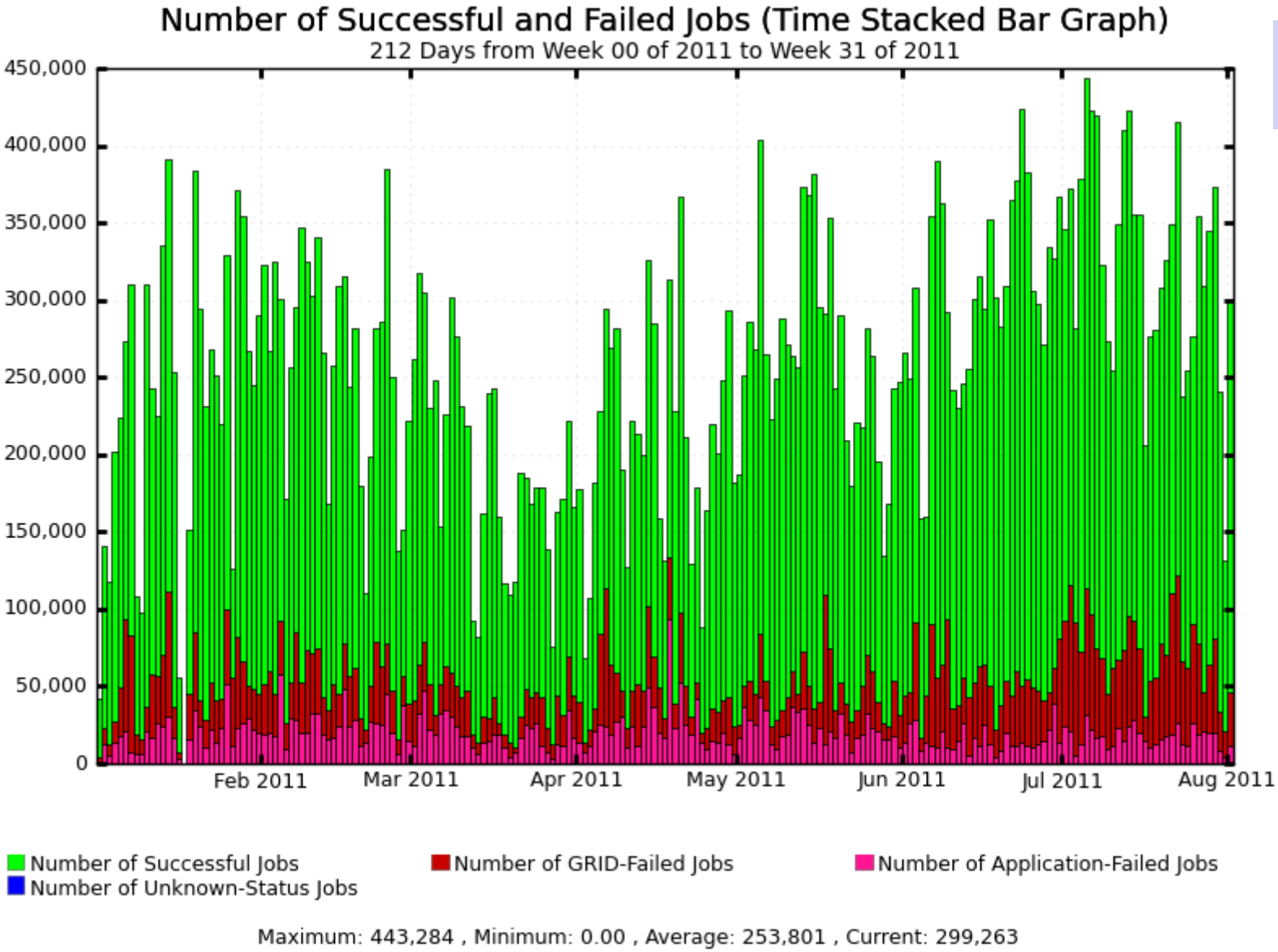
- 2000 packages
- 4 Millions lines C++, 1.4 Millions lines python, 0.1 million F/F90, 0.1 million java,...
- 1000 developers have committed in SVN offline repository for last 3 years
- 300 developers have requested 4000 package changes in first semester 2011 (25 per day)
  - It never stops: data taking, reprocessing, analysis peak for summer winter conference
- 3000 users have grid certificate in atlas vo (able to submit job, retrieve data)

Grid Analysis Activity

- Running analysis jobs since August 2010
- Analysis activity is inherently “spiky” and “chaotic”
- All clouds contribute to analysis
- ~24k jobs at peak load

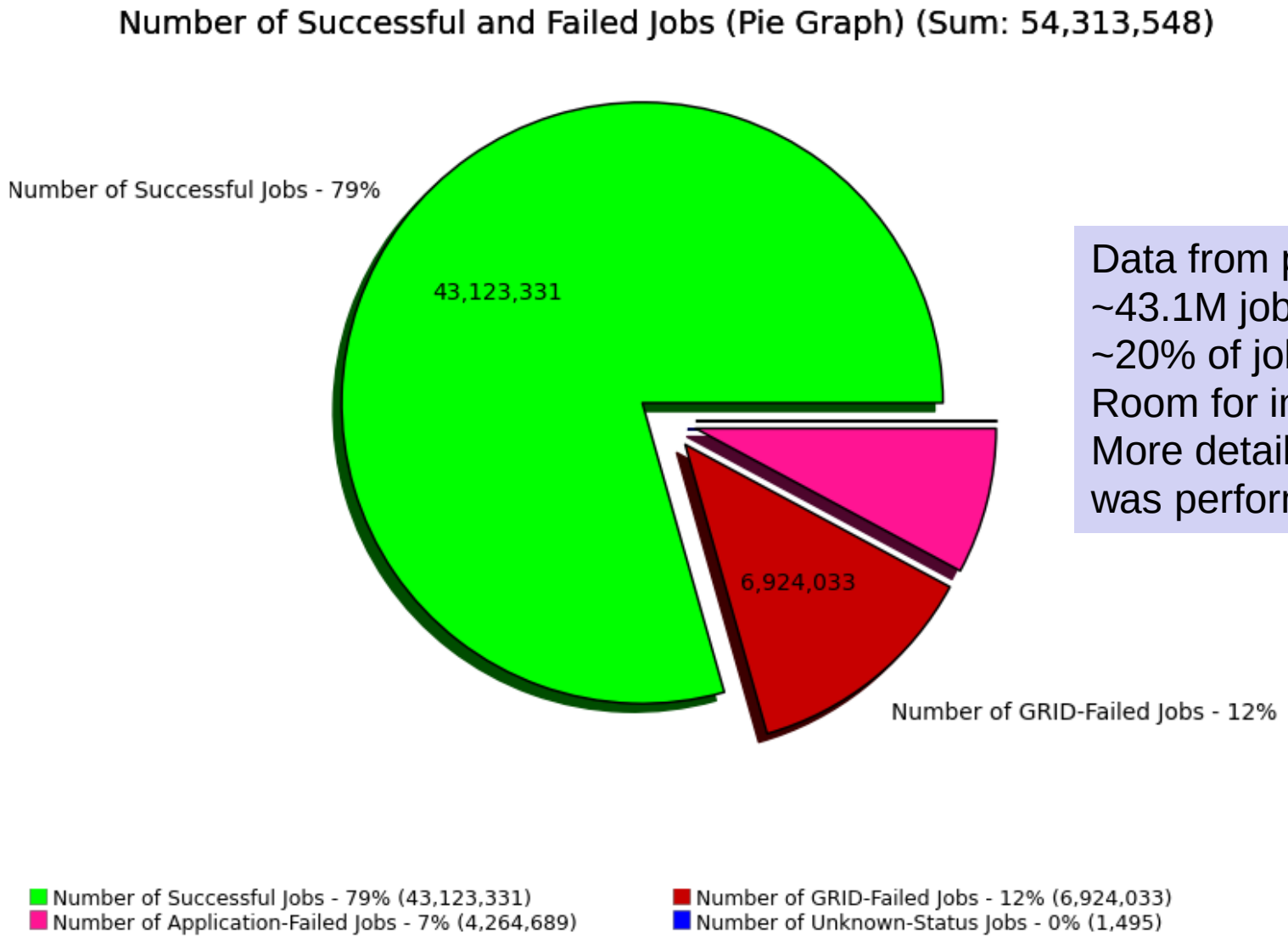


Grid Analysis Activity 2011



ATLAS Dashboard
All analysis sites

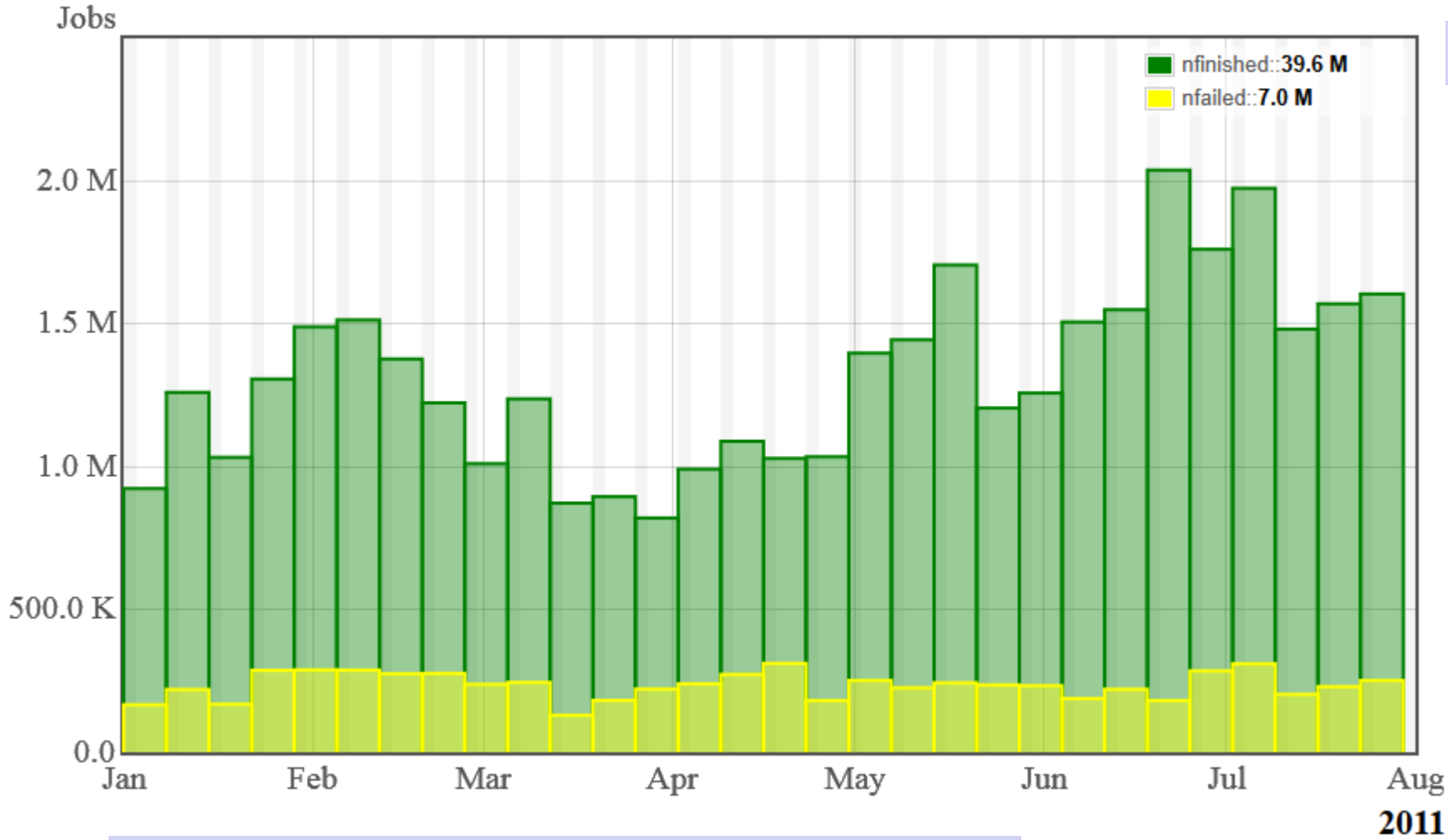
Grid Analysis Activity 2011



ATLAS Dashboard
All analysis sites

Data from previous plot integrated
~43.1M jobs finished in 2011 so far
~20% of jobs failed
Room for improvement
More detailed analysis of job failures was performed

Grid Analysis Activity in 2011



Panda Monitor

Panda driven sites only
Weekly bins
Peaks correspond to major conference seasons?

Number of Analysis Users

Configuration | Production | Clouds | Incidents | DDM | PandaMover | AutoPilot | Sites | Releases | Analysis | Stats | Users | Physics data | ProdDash | StatsDash | DDMDash | SSB

Update

Panda monitor
Times are in UTC

Panda info and help

Recent Panda Analysis Users

Jobs - search
States: running, defined, waiting, assigned, activated, finished, failed
Types: analysis, prod, install, test

Quick search
Panda job ID
Batch ID
Dataset
Task request
Task status
File

Summaries
Blocks: days
Errors: days
Nodes: days
Usage 1, 3 days

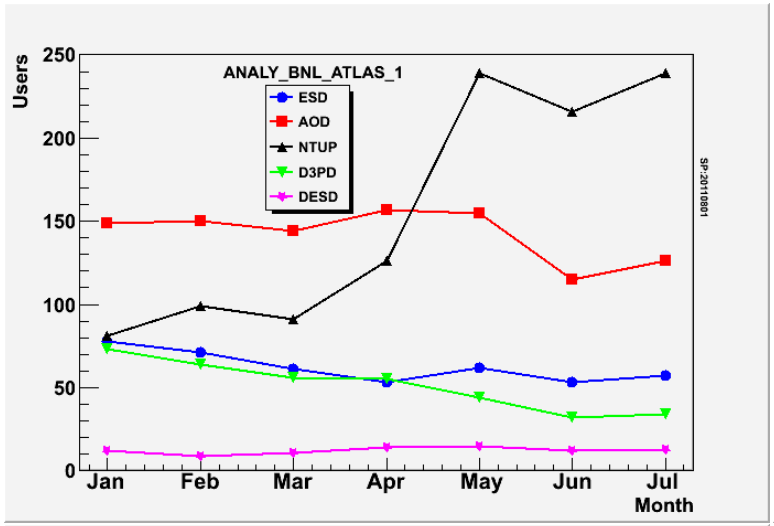
Users in the last 3 days: 459 7: 574 30:926 90:1221 180:1447
Usage in the last 7 days: Job count: 1395422 Users with >1000 jobs: 234 >10k jobs: 22

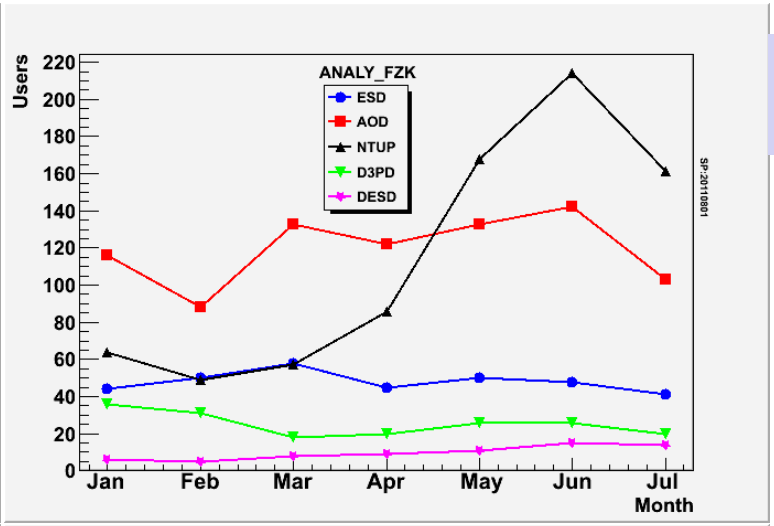
446 users in the last 3 days: (CPU in CPU-hours)

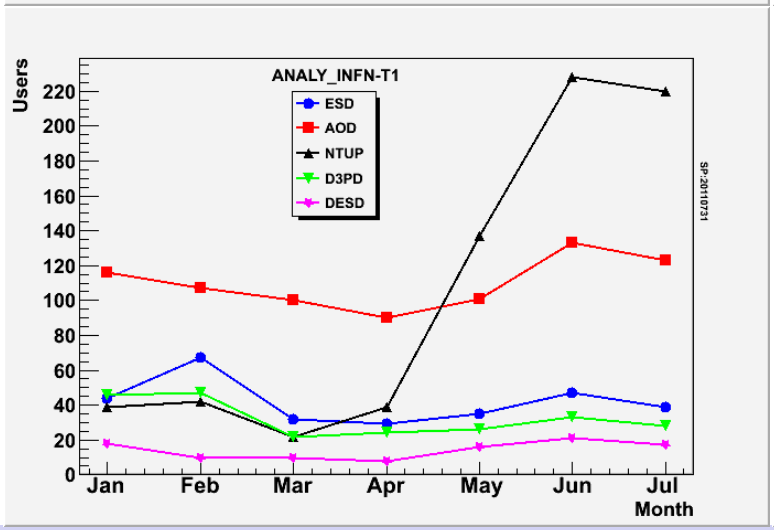
User	NJobs	Latest	Personal CPU 1 day	Personal CPU 7 day	Express CPU 1 day	Express CPU 7 day	Group CPU 1 day	Group CPU 7 day	Groups
Anyes Taffard	5384	08-08 18:03	0	0	0	0	0	0	
Aranzazu Ruiz Martinez	1339	08-08 18:03	0	0	0	0	0	0	
Gregor Kasieczka	10238	08-08 18:03	0	0	0	0	0	0	
James Lacey	8795	08-08 18:03	0	0	0	0	0	0	
Joao Gentil Mendes Saraiva	175	08-08 18:03	0	0	0	0	0	0	
Michael Flowerdew	5085	08-08 18:03	0	0	0	0	0	0	
Shannon Walch	4115	08-08 18:03	0	0	0	0	0	0	
Tomoe Kishimoto	16357	08-08 18:03	0	0	0	0	0	0	

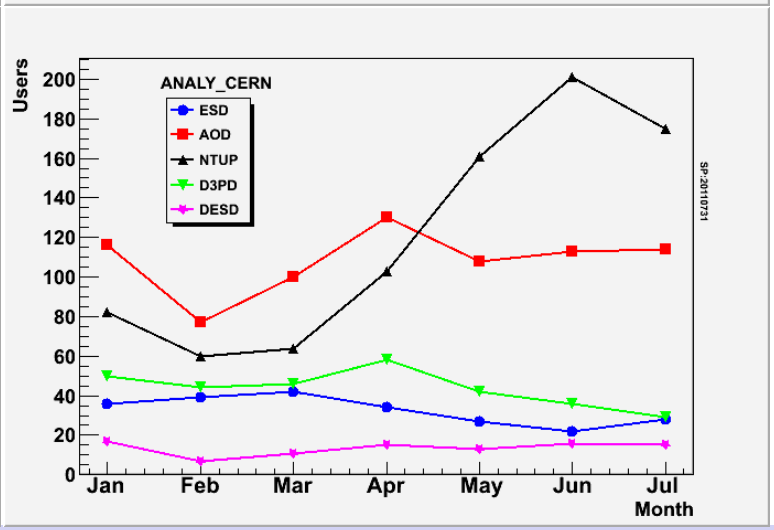
As of Aug 8th , 2011
1447 users were doing analysis on the Grid in the past 6 month
926 users in the past month
459 users in the past 3 days

Analysis users at Tier 1 sites





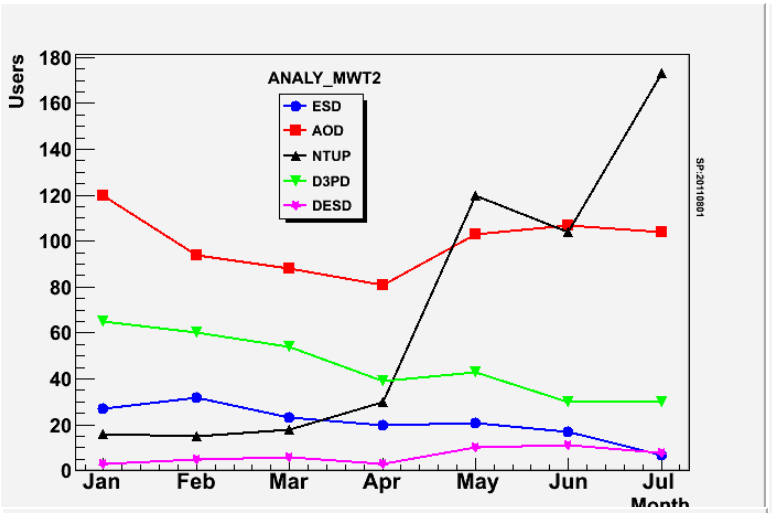


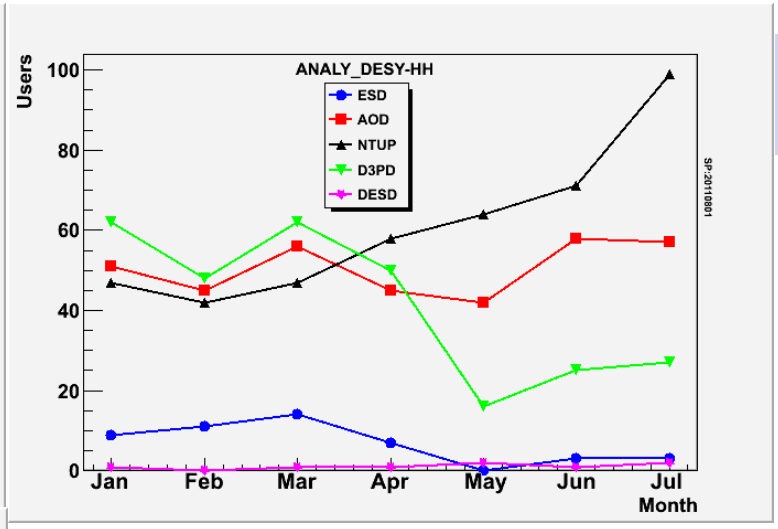


US,DE,IT,CERN sites in 2011

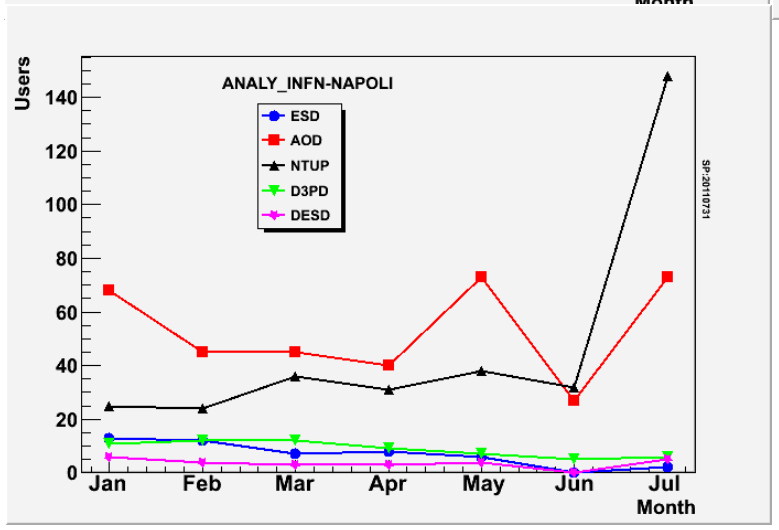
NTUP and AOD are most popular input data formats for analysis in 2011
NTUP popularity grew rapidly in a past few month, AOD popularity is steady
Similar trends at selected T1 sites

Analysis users at Tier 2 sites





Selected US,DE,IT T2 sites in 2011



More varied popularity patterns at T2 sites
AOD based analysis is popular at all T2 sites
Number of people working with NTUP format is growing
ESD popularity is falling

More detailed data popularity analysis on submission and sub-job level was also available

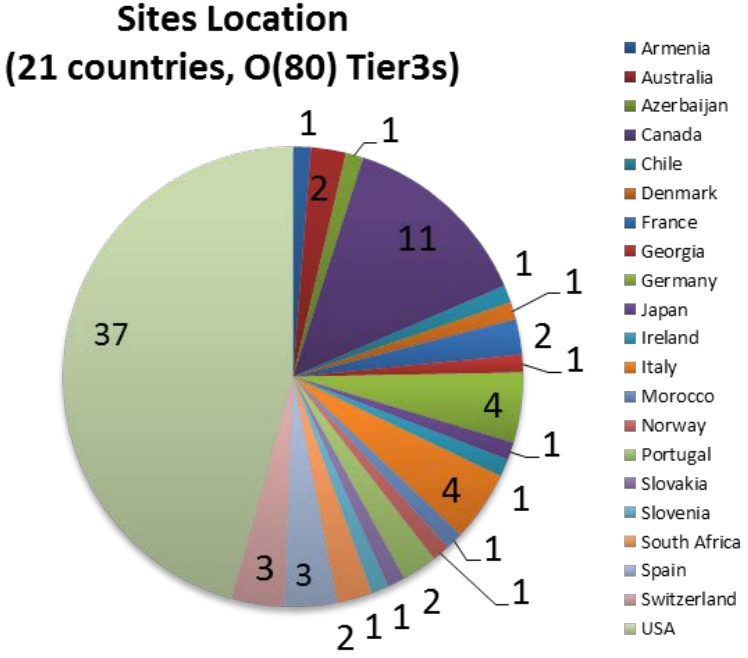
User Support

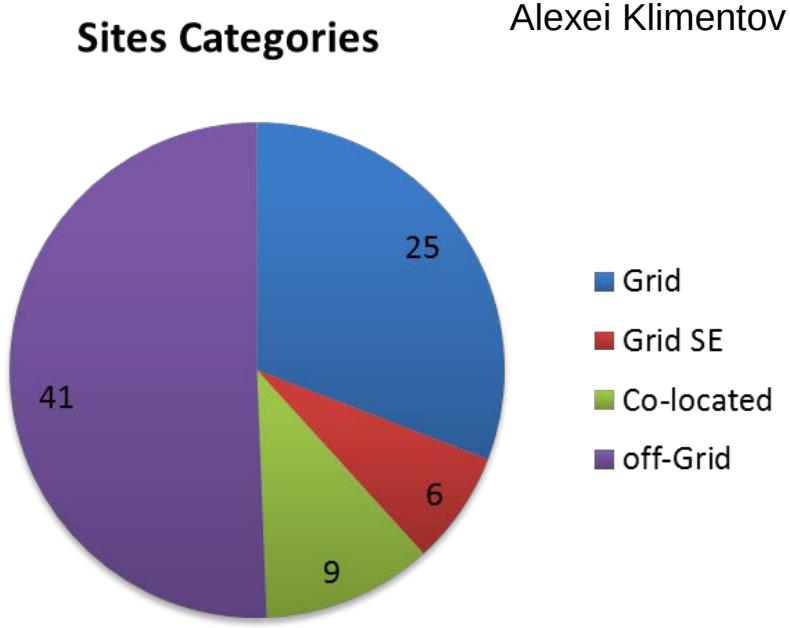
- We have ~1400 active distributed analysis users
  - They should not need to be distributed computing experts – The Grid is a black box that should just work
  - Grid workflows are still being tuned – not everything is 100% naïve user-proof
  - Supporting the users to get real work done is critical (it will stay like this!)
- ATLAS introduced a team of expert user support shifters in fall 2008.

- DAST: Distributed Analysis Support Team
  - Class 2 (off-site) ATLAS shifts; week-long shifts in EU and NA time zones (Asia-Pacific shifters wanted…)
  - 1st and 2nd-level support: better incorporate new shifters and shares the load in times of high demand
  - DAST is a ~15 member team; each takes a shift every 4-8 weeks.
- Users discuss all problems on a single “DA Help” eGroup
  - Discussion about all grid tools, workflows, problems
  - Not just DA – also data management questions

Tier3s in ATLAS

Alexei Klimentov





- O(80) registered Tier3s located in 21 countries. 8 countries have one Grid site
  - *Tier3 size from 16 cores/5 TB to 300 cores/250 TB.*
  - *More than half of sites are off-Grid.*
  - *15 Tier3s participate in DDM functional test , 7 Tier3s participate in Distributed Analysis Functional Test*
- Ongoing effort to provide common solutions and tools for T3 sites
  - DDM tools, workload management
  - Monitoring
  - Support framework

Evolution of Data Placement Model I

2005. Datasets and containers concept

Before 2010 mantra “jobs go to data” served us well

Strictly planned data placement

ATLAS Model First Year (AMFY)

Thermodynamic model

Custodiality (primary/secondary replicas)

Data deletion constraints

2010: “data and jobs move to available CPU resources”

Planned Data Placement + Dynamic Data Placement

(K.De : PD2P – PanDA Dynamic Data Placement) **for more see A. Stradling’s talk on Friday**

2011: Analysis of data format popularity led to

“Life without the ESD”

1 copy of RAW on disk@T1

Evolution of Data Placement Model II

2011

- Move toward caching of data rather than strict planned data placement
  - A dataset is still unit of replication and data placement
- All data replicas are classified as 'primary/secondary'

  Two primary copies of AOD and DESD ATLAS wide (Tier-1s)

  One additional planned copy of 'secondary' data for users (Tier-1s)

  Distribution is done according to MoU shares

  Additional copies :
  - Automatically done by PD2P based on usage pattern
  - Made by users (via Dataset Transfer Request Interface. DaTRI requests)

  Secondary replicas are deleted as soon as disk is 90% full

  No planned data placement at Tier-2s.
- Use data popularity and data access information to regulate the number of Grid replicas.

2012+

- beyond PD2P : file (or even event) level caching, direct reading of remote data
  - Remove cloud boundaries. "Any data, any time" (anywhere)

Summary and Outlook

- ATLAS distributed computing performed well in 2010-2011 and was recognized as a success on several levels
  - “Limitation to release results is not with computing…” F. Gianotti
    - LHCC review
  - ~1440 physicists used Grid to analyze ATLAS data in 2011, ~40M jobs completed
  - Many papers published
- Experience with data has clarified many things
- Still need to optimize usage of computing resources, improve robustness of the system
- Move from strict planned data placement toward caching of data
- Monitor carefully users analysis needs
  - Tune number of data replicas over Grid
    - To decrease jobs waiting time
    - To minimize users manual operations
    - Optimize storage requirements
  - Off Grid analysis monitoring by September
- Will need to adapt new technologies and ideas to fulfill needs of future ATALS data analysis

DA Functional and Stress Testing

- We pre-validate sites for distributed analysis with Functional and Stress tests:
  - GangaRobot is running a continuous stream of short user analysis jobs at all grid sites
    - Carefully selected test jobs, representative of a wide classes of analyses
    - Automatic blacklisting/whitelisting of sites based on a predefined policies. Tunable.
    - Results fed into Panda and are broadcasted to relevant communities
    - Helps to avoid “predictable” job failures. If GR jobs fail - similar type jobs will fail also at a tested site
    - Works well with automatic jobs re-brokerage in Panda
  - HammerCloud is used for on-demand stress tests spanning one or many sites
    - Used to commission new sites, tune the performance at existing sites, and to benchmark sites to make comparisons