The ATLAS Distributed Analysis System

F. Legger (LMU)

on behalf of the ATLAS collaboration

October 17th, 2013
20th International Conference on Computing in High Energy and Nuclear Physics (CHEP), Amsterdam

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Outline

- Distributed Analysis in ATLAS during **LHC run 1**:
  - Computing model: *Job goes to data* paradigm
  - Distributed analysis: performance
    - **Statistics**
    - **Efficiency**
    - **Failures**
    - **Data types**

- Future developments and implications on distributed analysis for **run 2**:
  - new **production system**
  - new **analysis model**
  - **remote file access**

ATLAS Computing Model 2011-2013

ATLAS Grid infrastructure

This talk

Central production & Distributed Data Management (DDM)

T0 → T1, T2

Data processing

MC production

T1-T2

DDM: dq2

Size: PB

AOD: Analysis Object Data

Group production

Size: PB

D3PD: flat ntuple

Distributed Analysis

T1-T2 TB

derived formats

Group Analysis

User Analysis

T1-T2-T3

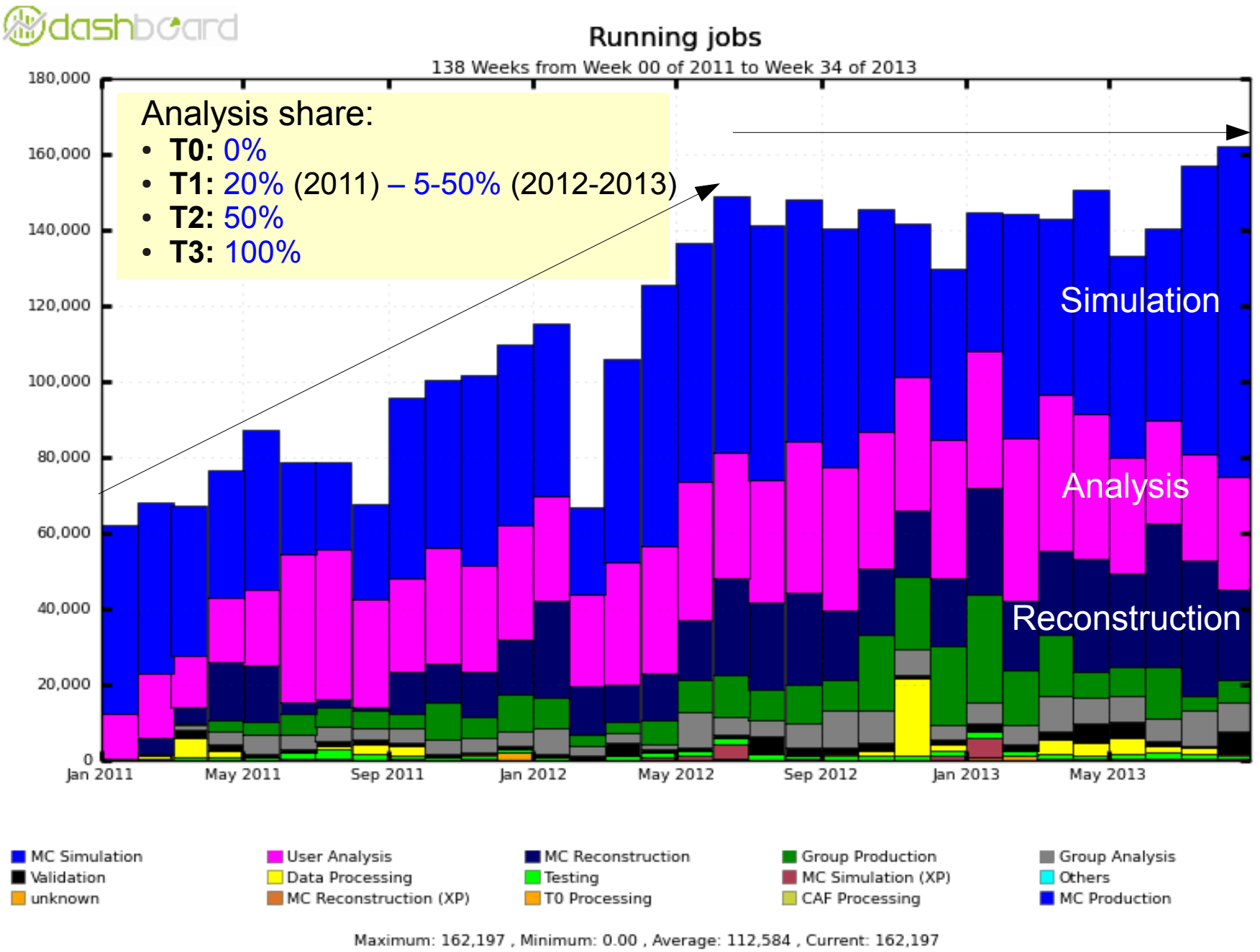
Size: GB

User tools for grid job submission:

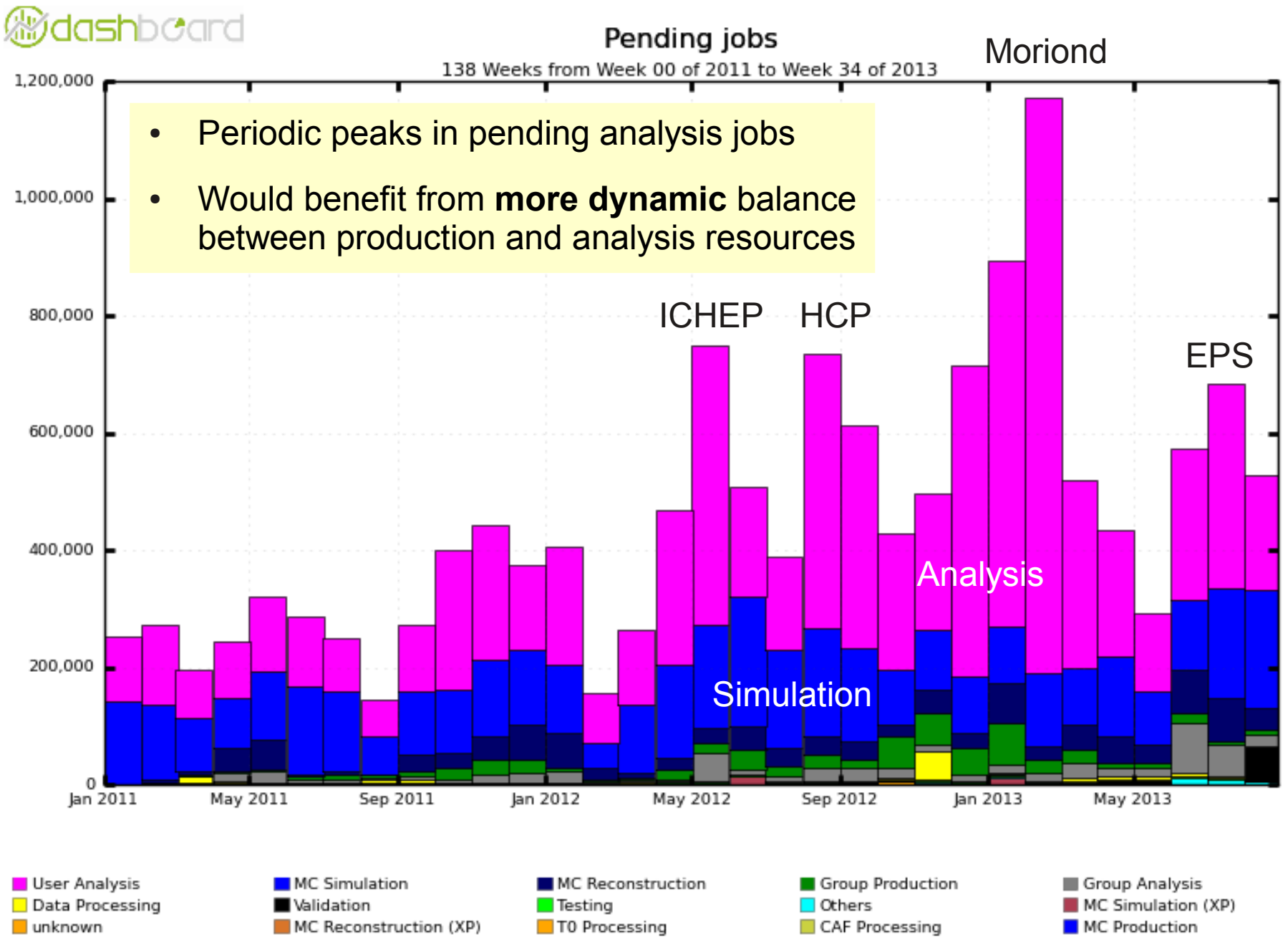
- PanDA
- Ganga

Paradigm: job goes to data

Running Grid jobs 2011-2013



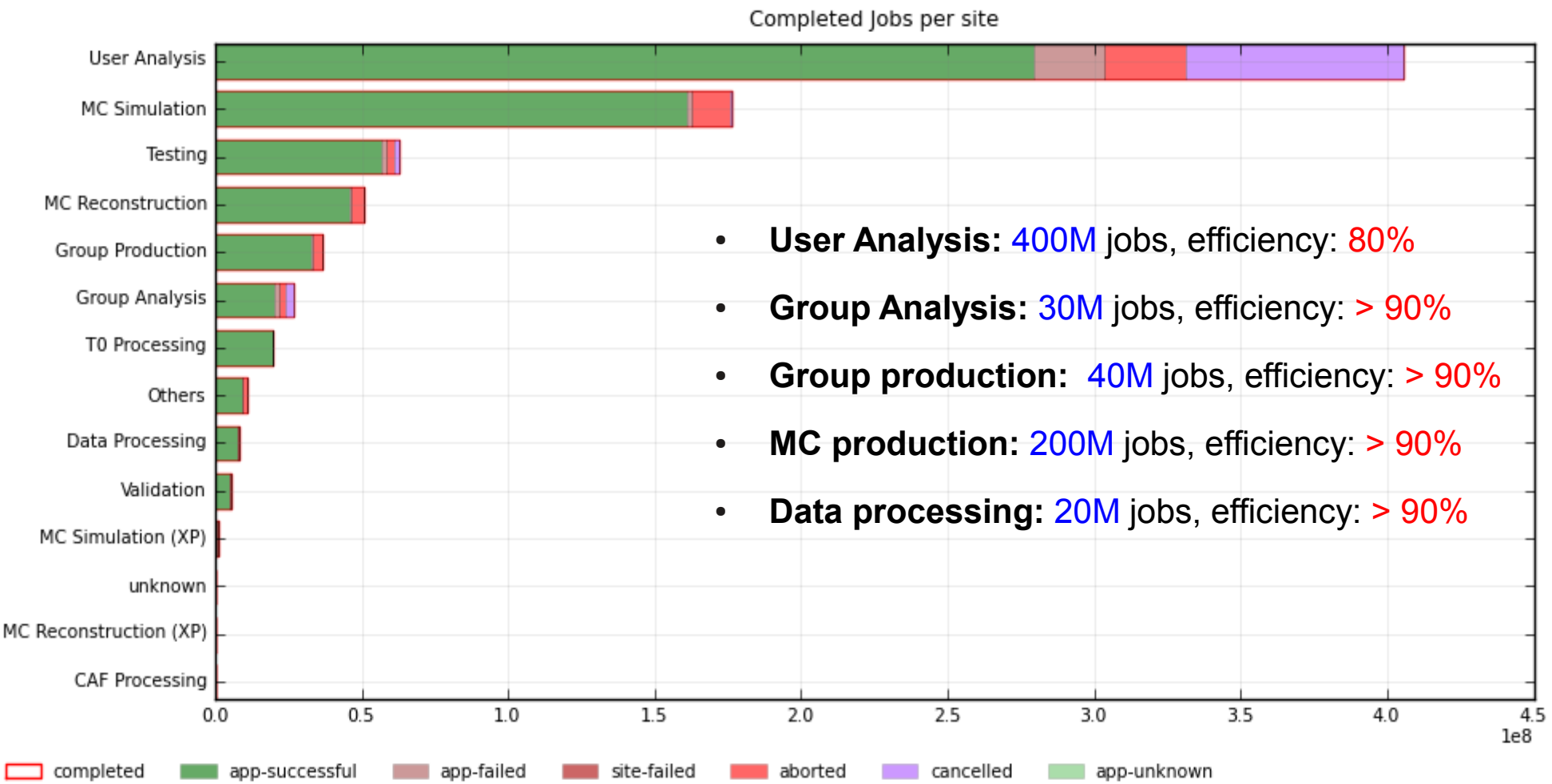
Pending Grid jobs 2011-2013



- Periodic peaks in pending analysis jobs
- Would benefit from **more dynamic** balance between production and analysis resources

ATLAS Grid jobs 2011-2013

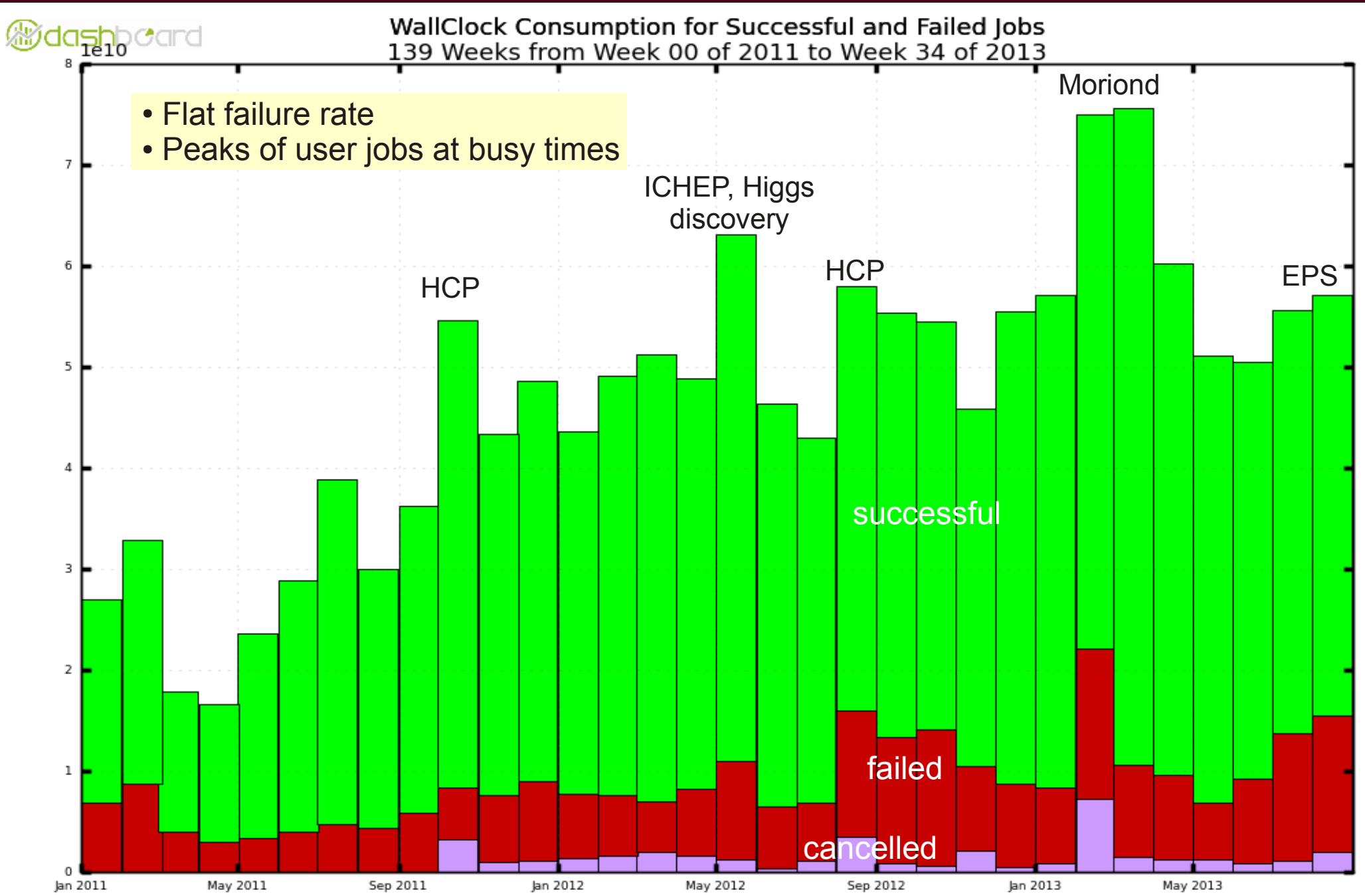
5% grid-related failure rate, mostly due to storage failures



- **User Analysis:** 400M jobs, efficiency: 80%
- **Group Analysis:** 30M jobs, efficiency: > 90%
- **Group production:** 40M jobs, efficiency: > 90%
- **MC production:** 200M jobs, efficiency: > 90%
- **Data processing:** 20M jobs, efficiency: > 90%

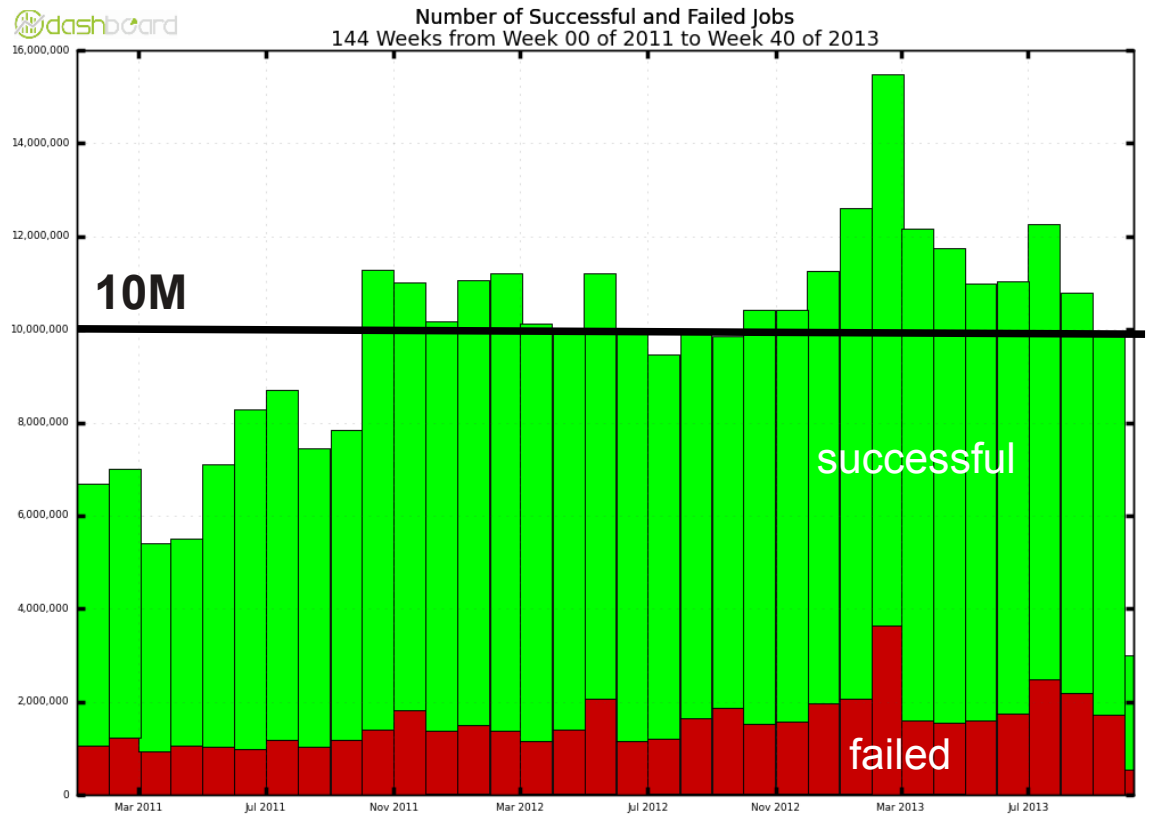
(NB cancelled jobs have walltime close to 0, negligible impact on efficiency)

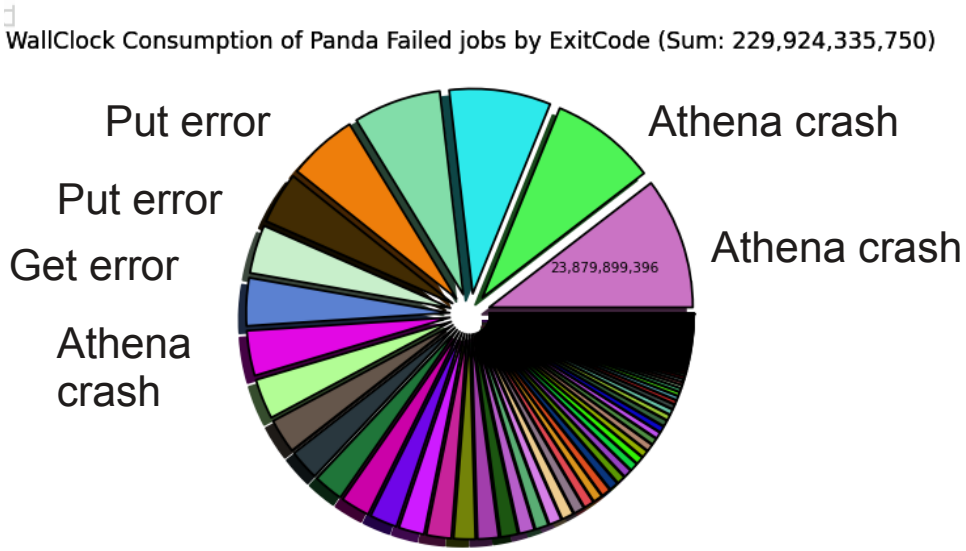
Analysis jobs 2011-2013



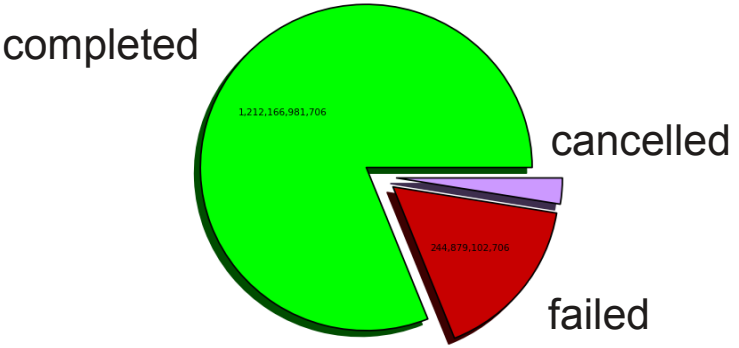
Efficiency analysis jobs: 2011-2013

- Average **400k analysis jobs/day**
- Efficiency : **80%**
- Wall clock consumption of failed jobs: **20%**
- most failures related to user code



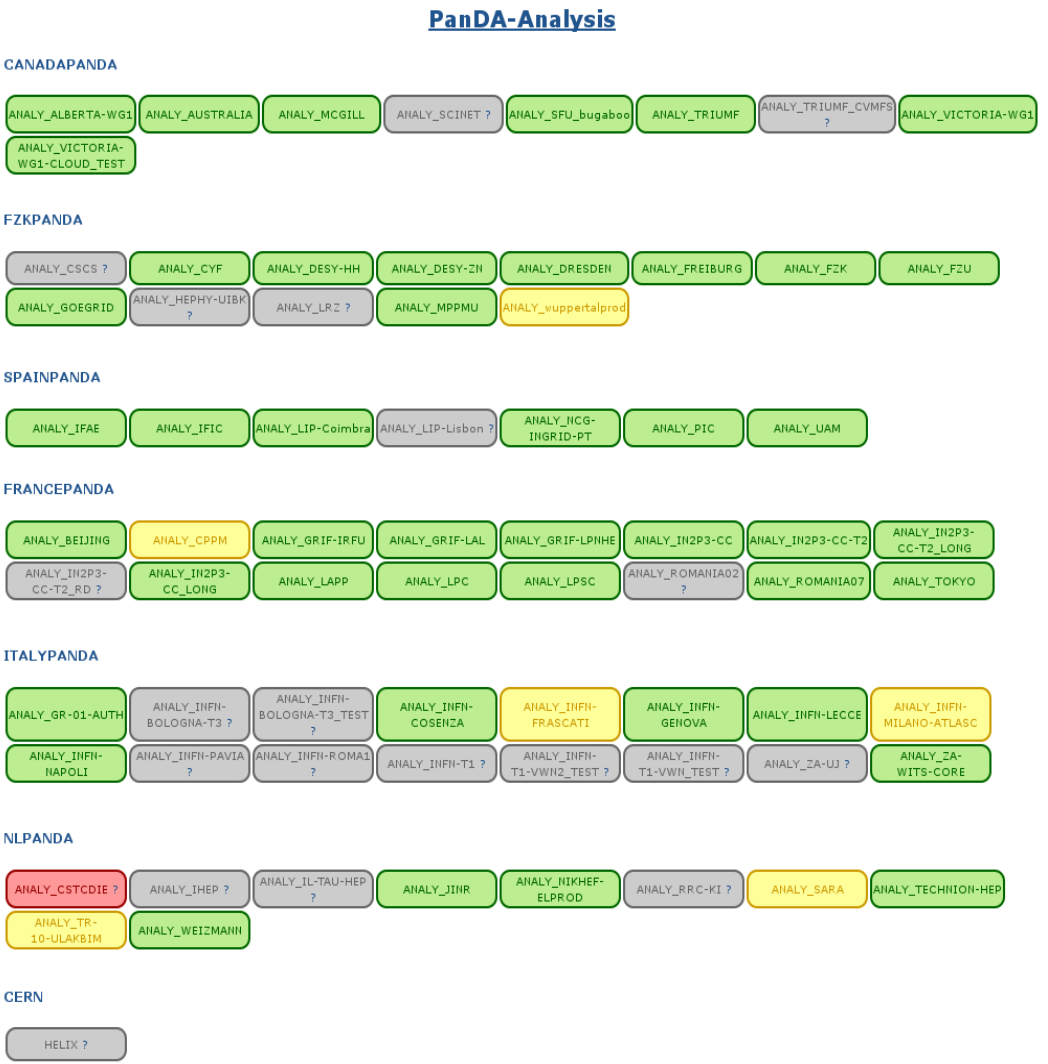


Wall clock consumption



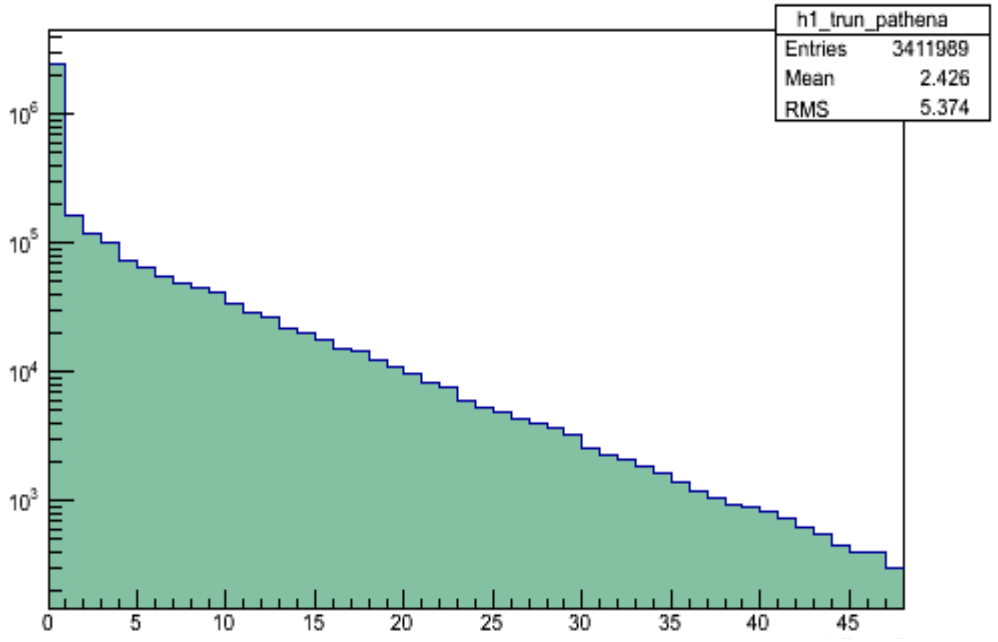
Failed analysis jobs

- What we do to prevent them:
  - Automatic site testing and exclusion from brokerage with **HammerCloud** (see dedicated talk)
  - **DAST** (Distributed Analysis Support Team):
    - User support with dedicated mailing list
    - expert shifters help users 16 hours/day (American and European time-zones)
  - What we plan to do:
    - Better error reporting for automatic retrials/better debugging
    - Automatic kill user jobs in case of massive failures



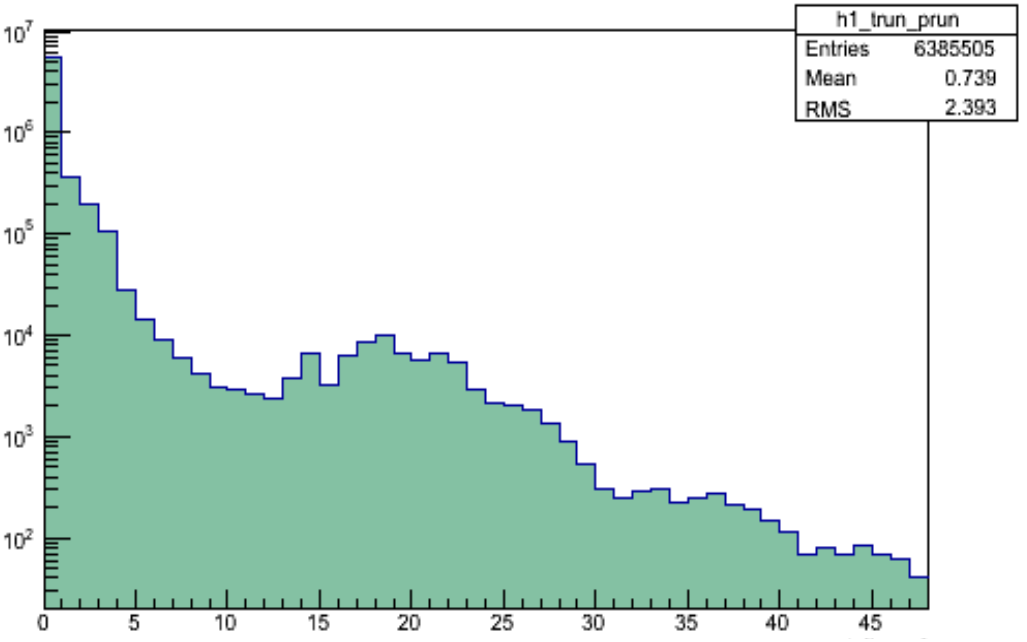
Analysis job length – January 2013

Jobs running on AODs



Run time [hours]

Jobs running on D3PDs



Run time [hours]

Most analysis jobs last 1 hour or less
→ produce many small files
→ also tail of longer jobs

Profit from better optimization of grid resources → need to estimate job parameters

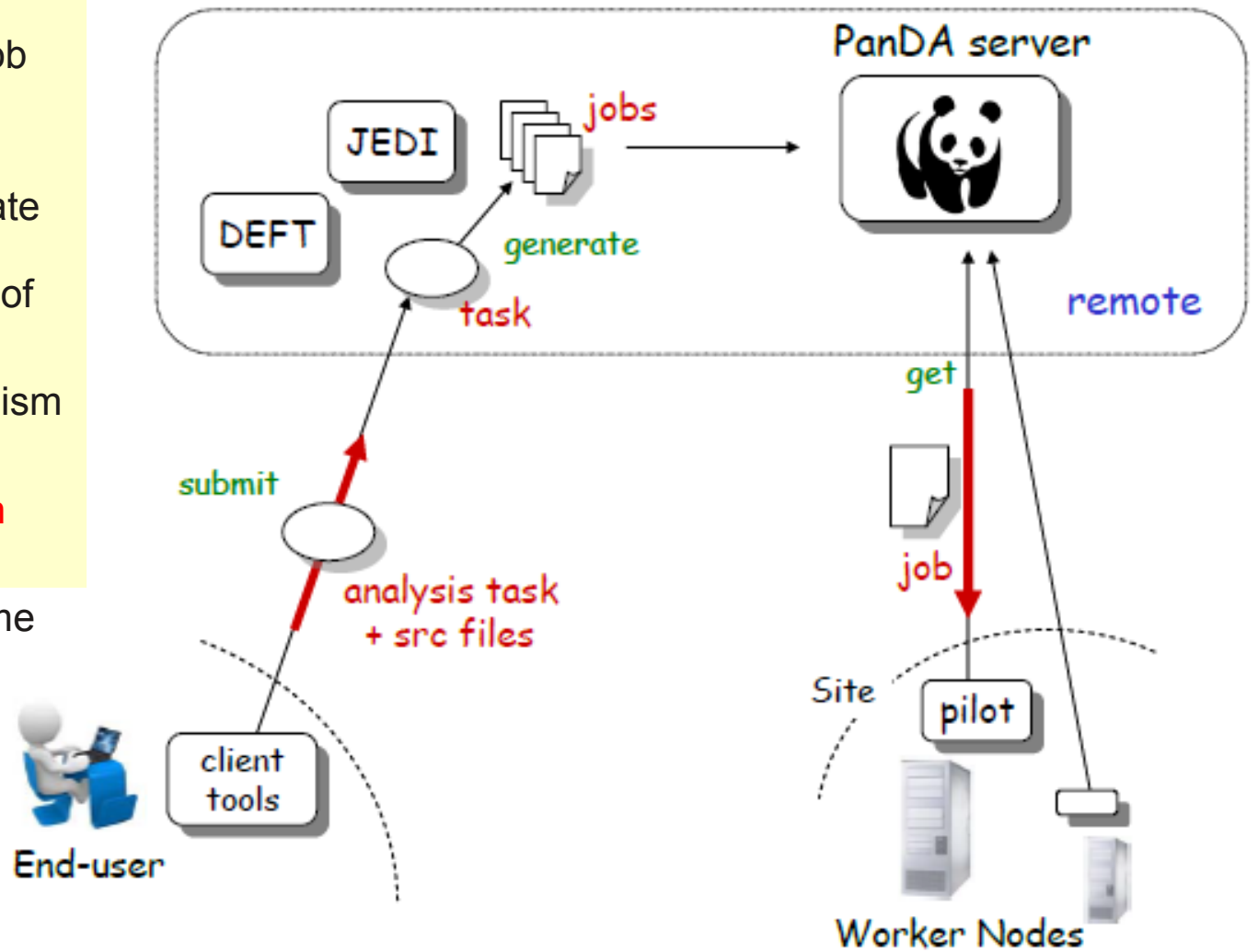
Future Production system

- Based on **tasks** rather than individual **jobs**:
  - More complex work-flows
  - Simplification of client tools:
    - Brokering and task/job management moved **server-side**
    - **Scout jobs** to estimate needed resources → more efficient usage of grid resources
    - Better retrial mechanism for failed jobs
    - faster job submission times!
  - CMS plans to use the same PanDA foundation for analysis

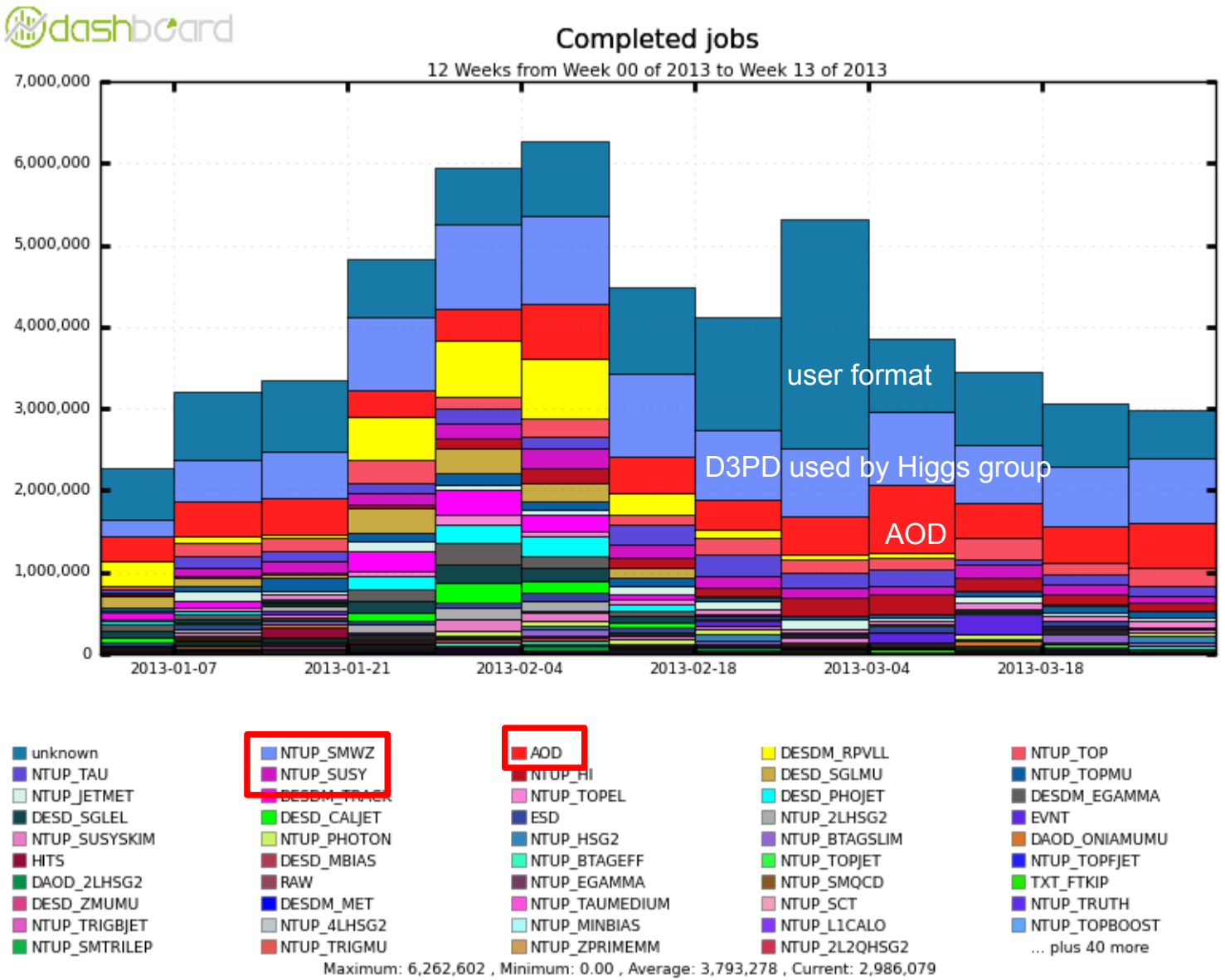
- ProdSys2 (DEFT/JEDI)
- DDM based on Rucio

→ see dedicated talks

New Analysis model



Input data – January-March 2013



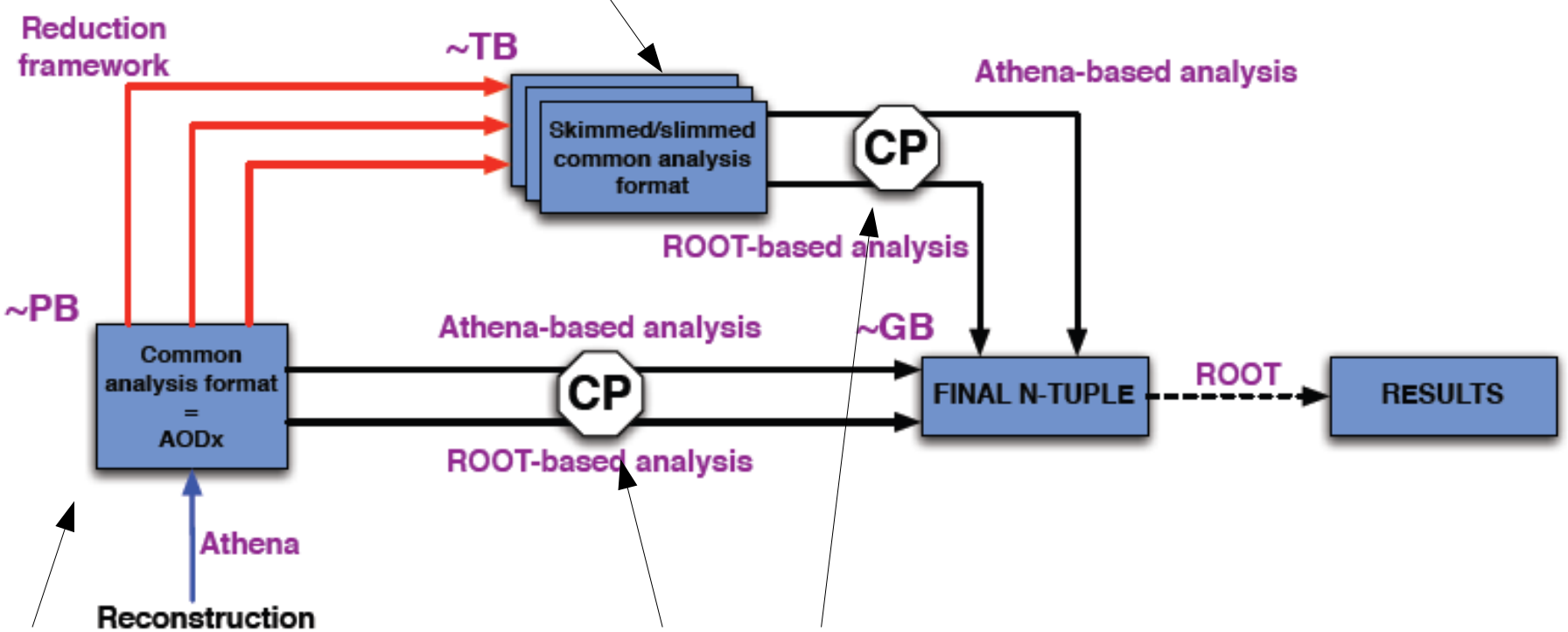
AOD vs D3PD:

- Full reconstruction info *only* available in AOD
- Size D3PD/AOD: up to 50%
- Almost 100 D3PD flavours!
- Total size of a full version of the main D3PDs is 3 times larger than the AODs
- AOD/D3PD users: ¼
- D3PD production: 5-10 s/event, to produce all D3PD flavors, 30 s/event are needed

Future analysis model

Centrally provided
reduction framework
for slimming/skimming

New analysis framework to support both reduced and original AODx analysis, integrated with grid tools



Merged AOD+D3PD format:
AODx, readable with both ROOT and athena

Standardized application of Combined Performance (CP) recommendations

Disk space needed at the end of 2015

Current model: AOD+D3PD New model

	events	MB/evt	AOD (PB)	D3PD (PB)	AODx (PB)
2012, data	$3.9 \cdot 10^9$	0.24	0.9	2.8	1.2
2012, MC	$4.5 \cdot 10^9$	0.40	1.8	5.4	2.2
2015, data	$5 \cdot 10^9$	0.24	2.4	14.4	6
2015, MC	$6 \cdot 10^9$	0.40	4.8	28.8	12

Assumptions:

- AODx / AOD = 1.25, D3PD/AOD = 3
- Total size of run 1 is 2 x size of 2012
- 1 copy of run 1 data
- 2 copies of 2015 data, 2 versions of D3PDs/AODx

	AOD+D3PD (PB)	AODx (PB)
Run 1	22	7
2015	50	18
Total	**72**	**25**

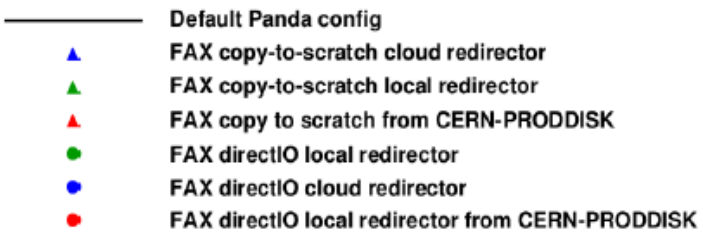
The Federated ATLAS XrootD system (FAX)

- Storage federation aiming to treat Tier1, Tier2 and Tier3 storage space as a *single distributed storage system*

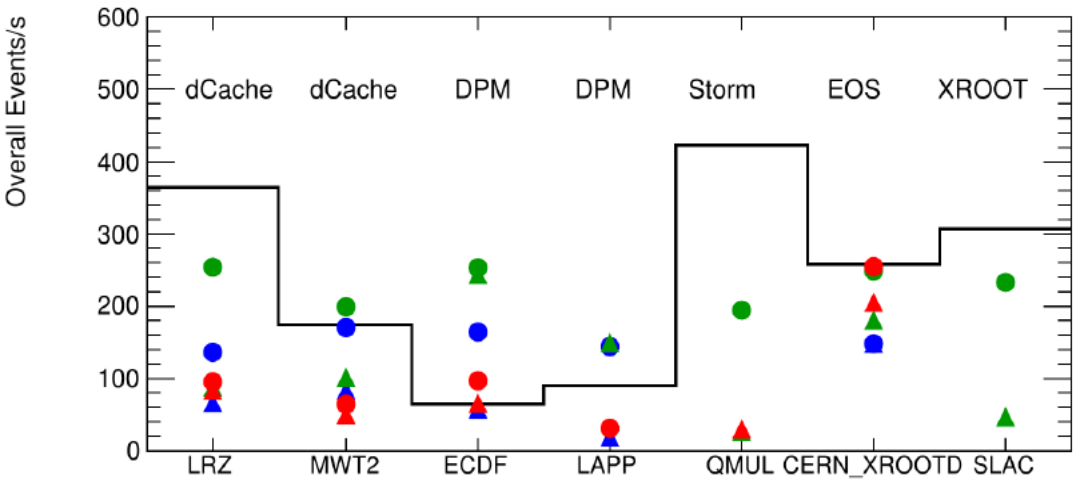
  → relax *data-CPU locality* paradigm

  - **FAX** routes the client to the nearest site with the requested data (scope yet to be defined)
    - data access through a client software like ROOT or xrdcp (other protocols as http under evaluation)
    - requires advanced caching mechanism as TTreeCache
    - transparent for the user
    - currently under evaluation
    - see dedicated talk

Different input access modes:



Result for example sites of different storage systems:



- **Use cases:**
  - Quick scan through large data sample
  - Fetch missing files instead of failed jobs
  - Use of opportunistic resources

Conclusions

- Successful deployment of Distributed Analysis in ATLAS during run 1:
  - 400k/day user jobs: 80% efficiency, 5% grid-related failures (mostly due to storage failure)
  - HammerCloud auto-exclusion to ensure smooth grid operation
  - DAST to provide user support

- Several challenges coming up:
  - new Analysis Model, new Production system, remote file access with FAX

- Future developments:
  - Balance production vs analysis resources, use of opportunistic not ATLAS resources
  - Features to improve user experience:
    - estimate of time-to-completion and improved job monitoring
    - improved automatic job retrial
    - job output to local storage
  - Features to smooth grid operations:
    - Definition of user job limits
    - Monitoring of massive user job failures

Backup

Assumptions disk space calculation

		2012 (8 TeV)	2015 (14 TeV)
Trigger rate	events/s	400 + 150 (delayed)	1000
Data	events	$3 + 1$ (delayed) $\cdot 10^9$	$5 \cdot 10^9$
MC	events	$4.5 \cdot 10^9$	$6 \cdot 10^9$
Raw size	MB/event	0.8	1.1
AOD size (data)	KB/event	250	250
AOD size (MC)	KB/event	400	400
D3PD size (data)	KB/event	50-130	
D3PD size (MC)	KB/event	100-200	