Evolution of experiments computing: ATLAS

Johannes Elmsheuser

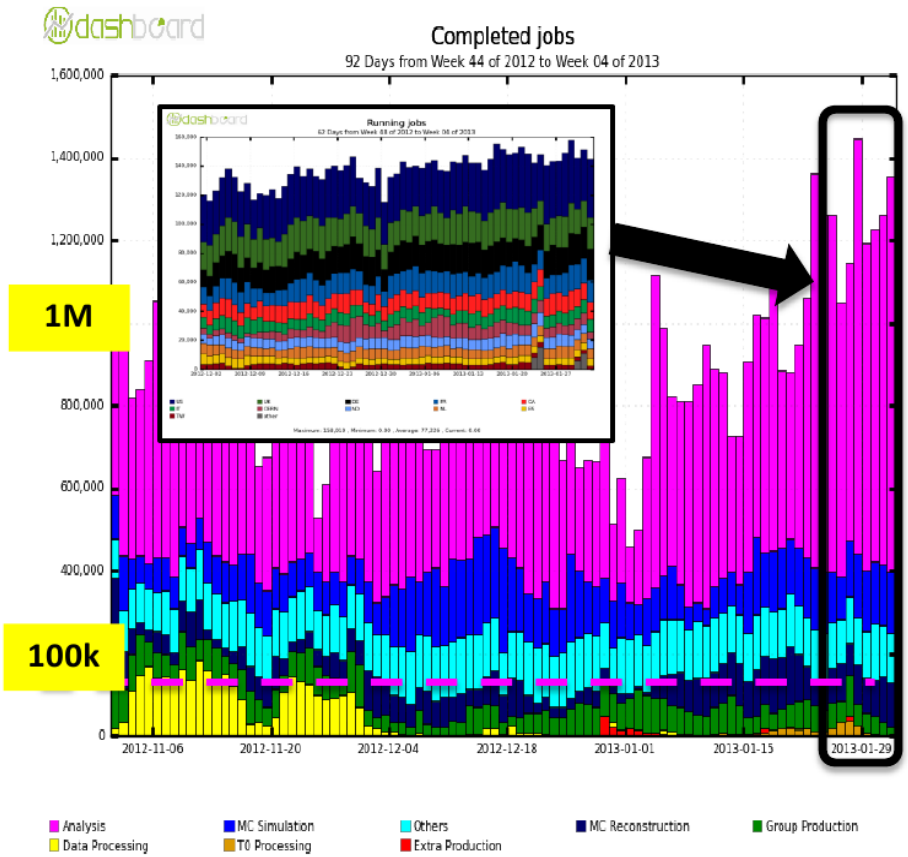
Ludwig-Maximilians-Universität München

3 December 2013/7th Annual Workshop of the Helmholtz Alliance "Physics at the Terascale"

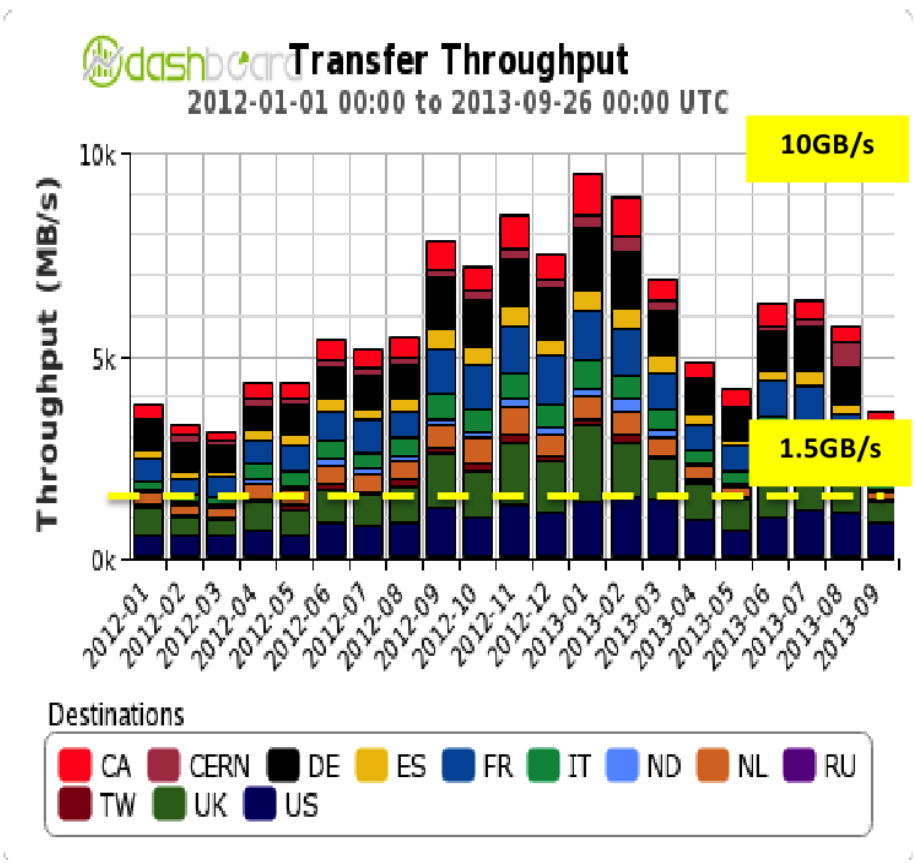
(Some content from Simone Campana's presentation at CHEP'13)

Run 1: Workload and Data Management

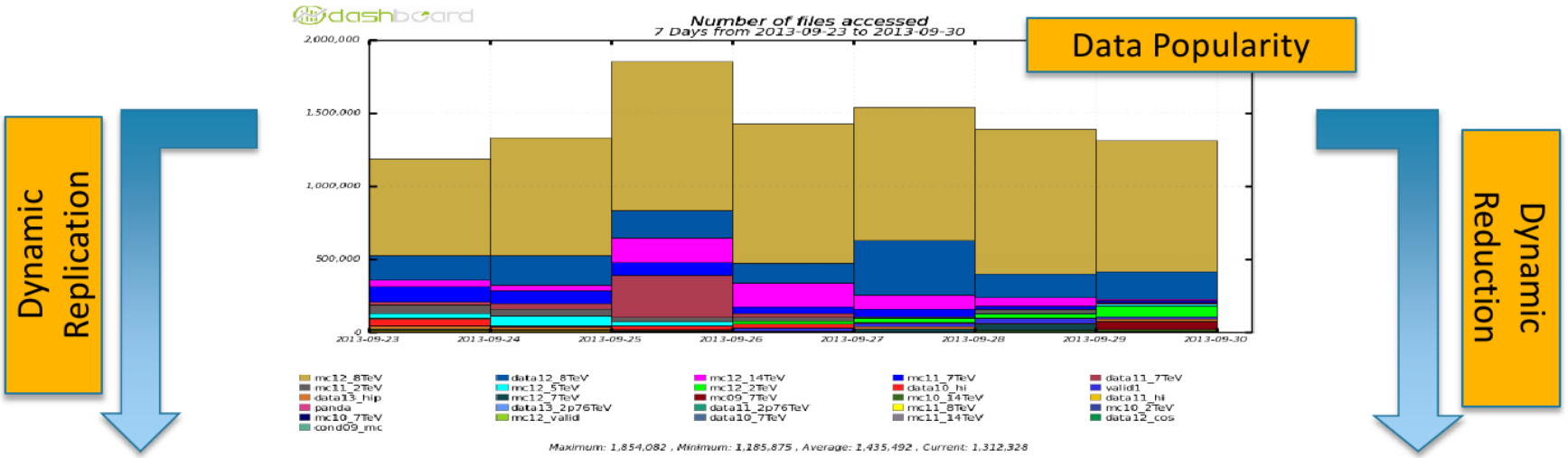
1.4M jobs/day, 150K concurrently running
(2007 gLite WMS acceptance tests: 100K jobs/day)

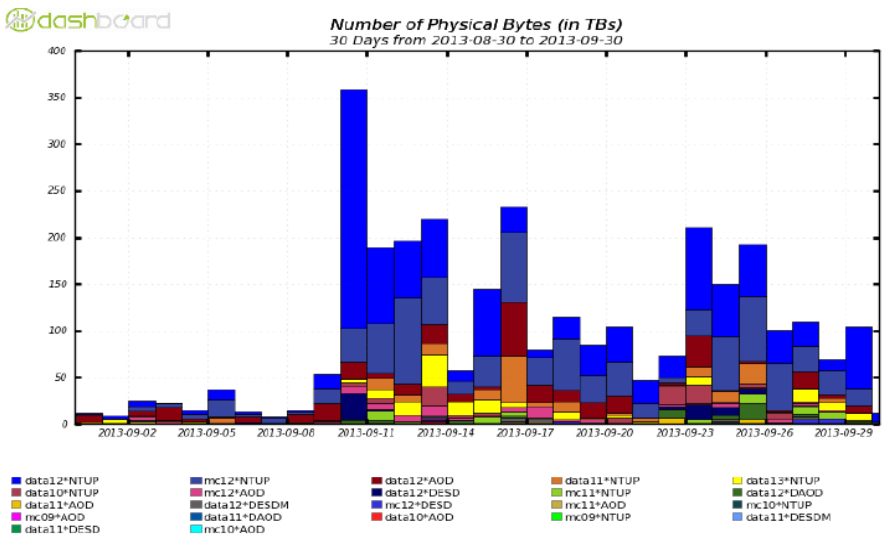


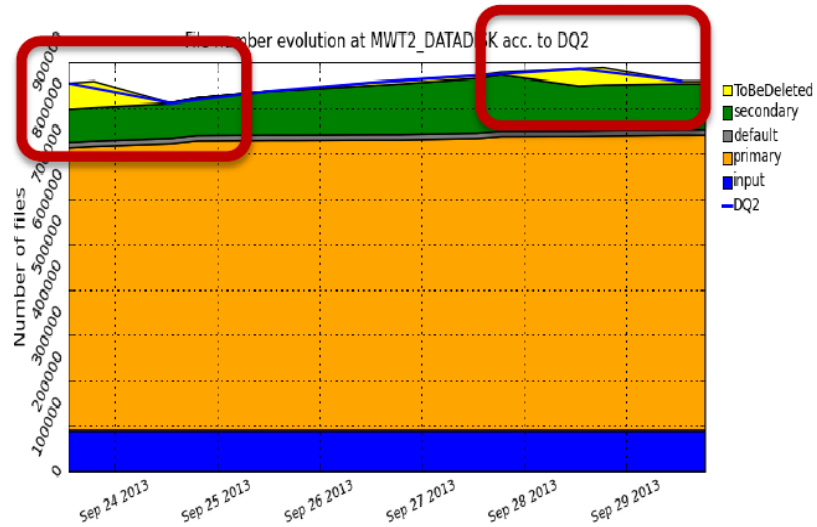
Nearly 10GB/s transfer rate
(STEP09 target: 1.5GB/s)



Run 1: Dynamic Data Replication and Reduction







Challenges of Run 2

Trigger rate: from 550Hz to 1kHz:

- Therefore, more events to record and process

Luminosity increase: event pile-up from 25 to 40:

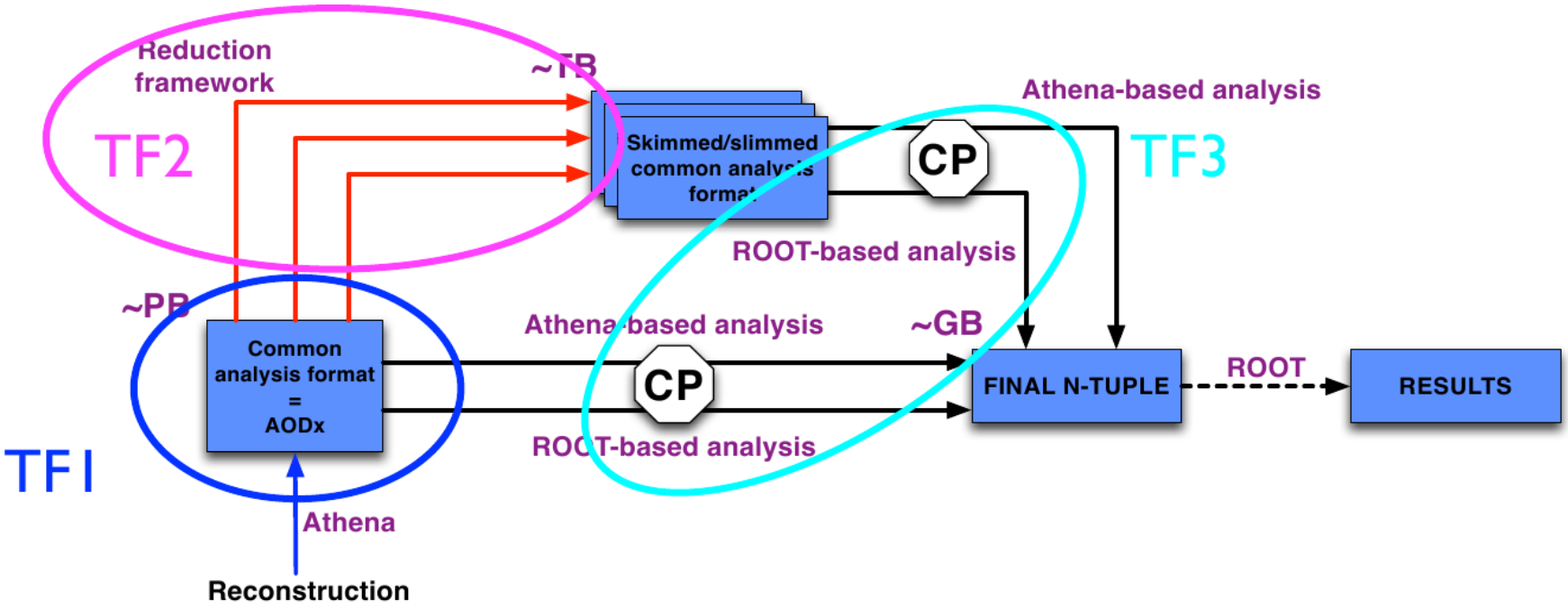
- so more complexity for processing and +20% event size

Flat resource budget:

- For storage, CPUs and network (apart for Moores law)
- For operations manpower
- The LHCC recommends that the hypothesis of flat future resources be removed from the assumptions; instead physics motivated needs should be stated.

The ATLAS Distributed Computing infrastructure needs to evolve in order to face those challenges

Recommendations of Analysis Model Study Group



Analysis Model Study Group recommended change of ATLAS (offline) analysis model - 3 (4) task forces setup

- TF1: design new analysis merged ROOT/Athena EDM ntuple: xAOD
- TF2: data reduction framework
- TF3: analysis framework and tools

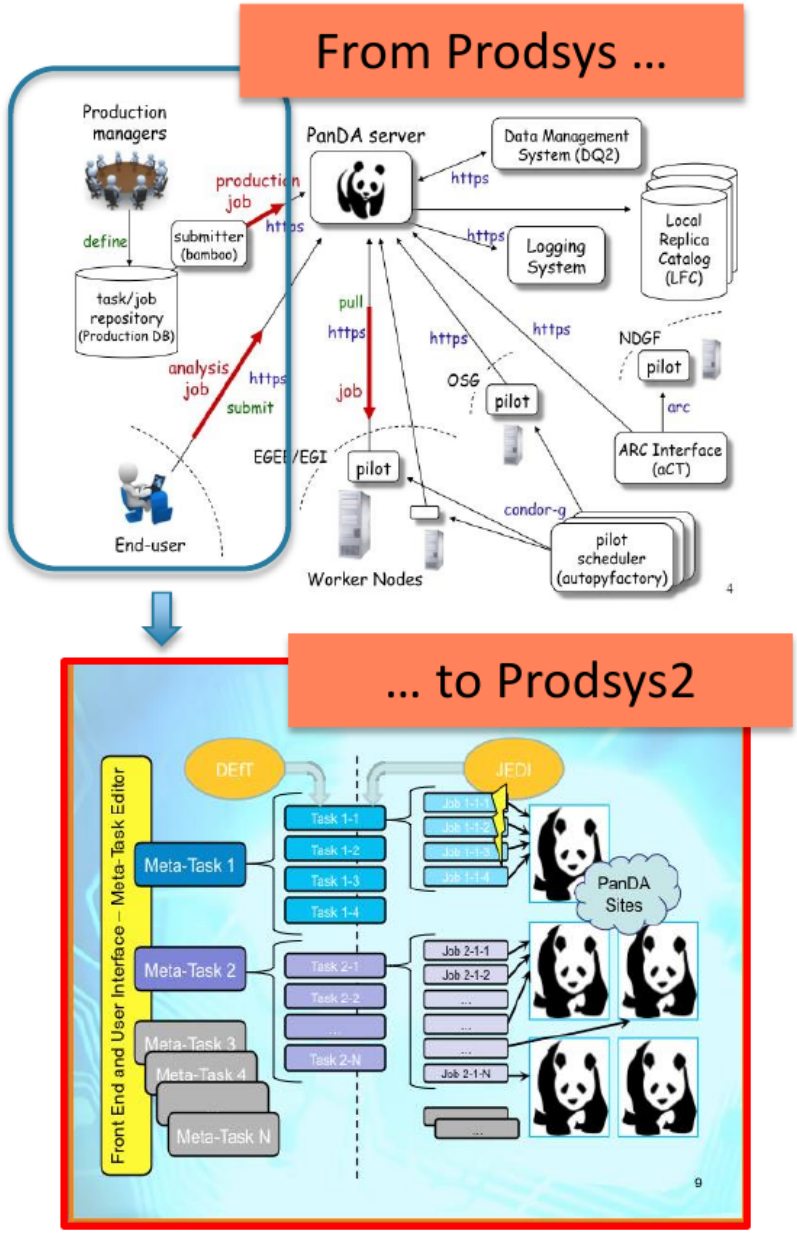
Workload Management in Run 2: Prodsys2

Prodsys2 core components:

- DEFT: translates user requests into task definitions
- JEDI: dynamically generates the job definitions
- PanDA: the job management engine

Features:

- Provide a workflow engine for both production and analysis
- Minimize data traffic (smart merging)
- Optimized job parameters to available resources



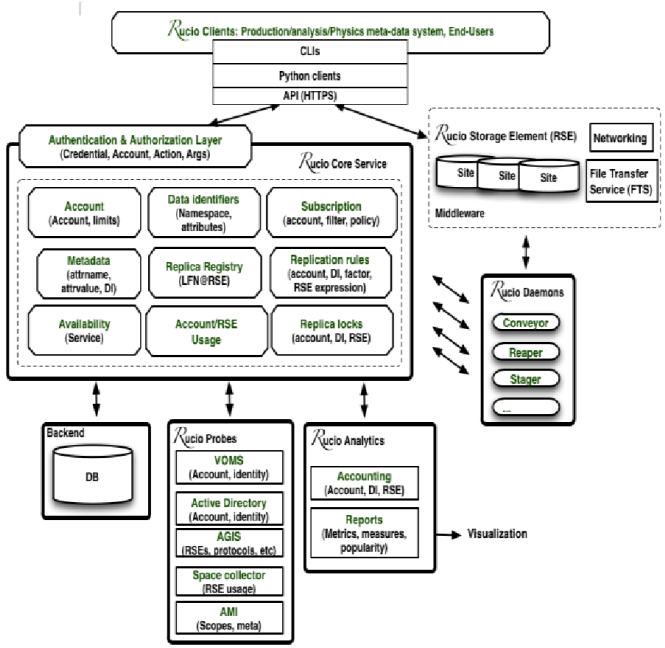
Data Management in Run 2: Rucio

Implements a highly evolved Data Management model

- File (rather than dataset) level granularity
- Multiple file ownership per user/group/activity

Features:

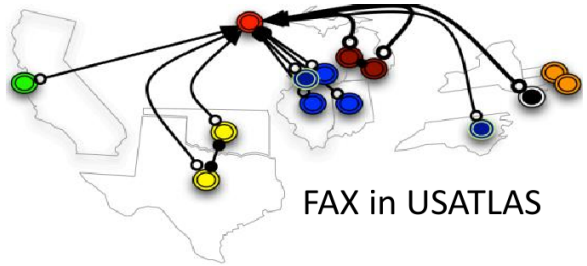
- Unified dataset/file catalogue with support for metadata
- Built-in policy based data replication for space and network optimization
- Redesign leveraging new middleware capabilities (FTS/GFAL-2)
- Plug-in based architecture supporting multiple protocols (SRM/gridFTP/xrootd/HTTP...)

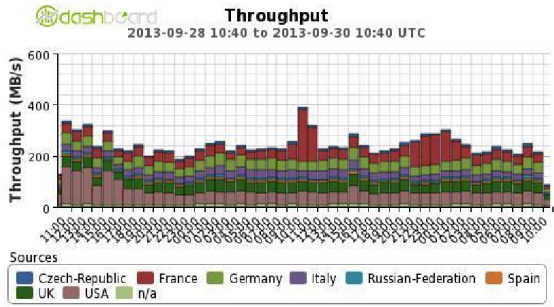


Data Management in Run 2: FAX

ATLAS is deploying a federated storage infrastructure based on xrootd

- Complementary to Rucio and leveraging its new features
- Offers transparent access to nearest available replica
- The protocol enables remote (WAN) direct data access to the storage
- Could utilize different protocols (e.g. HTTP) in future


FAX in USATLAS



Scenarios (increasing complexity):

- Jobs failover to FAX in case of data access failure
  → If the job can not access the file locally, it then tries through FAX
- Loosening the job-to-data locality in brokering
  → From jobs-go-to-data to jobs-go-as-close-as-possible-to-data
- Dynamic data caching based on access
  → File or even event level

Opportunistic Resources: Clouds

A "Cloud" infrastructure allows to demand resources through an established interface

- (If it can) it gives you back a (virtual) machine for you to use
- You become the administrator of your cluster

Free opportunistic cloud resources

- The ATLAS HLT farm is accessible through cloud interface during the Long Shutdown
- Academic facilities offering access to their infrastructure through a cloud interface

Cheap opportunistic cloud resources

- Commercial Infrastructures (Amazon EC2, Google, ...) offering good deals under restrictive conditions

Work done in ATLAS Distributed Computing

- Define a model for accessing and utilizing cloud resources effectively in ATLAS
- Develop necessary components for integration with cloud resources and automation of the workflows

Opportunistic Resources: HPCs

HPC offers important and necessary opportunities for HEP

- Possibility to parasitically utilize empty cycles

Oak Ridge Titan System	
Architecture:	Cray XK7
Cabinets:	200
Total cores:	299,008 Opteron Cores
Memory/core:	2GB
Speed:	20+ PF
Square Footage	4,352 sq feet

Bad news: very wide spectrum of site policies

- No External connectivity
- Small Disk size
- No pre-installed Grid clients
- One solution unlikely to fit all

Good news: from code perspective, anything seriously tried so far did work

- Geant4, ROOT, generators
- Short jobs preferable for backfilling

HPC exploitation is now a coordinated ATLAS activity

Event Service, Monitoring, Grid Information System, Databases

Event service

- Under development: Store small metadata info of every data and MC event in a database and make it accessible for special services

Monitoring

- `http://adc-monitoring.cern.ch/`
- Converged on an "ADC monitoring architecture"
- Rationalization of monitoring system and Porting monitoring to the newly developed components (Prodsys2, Rucio)

AGIS - Grid Information System

- Source repository of information for PanDA and DDM
- More a configuration service than an information system

Databases

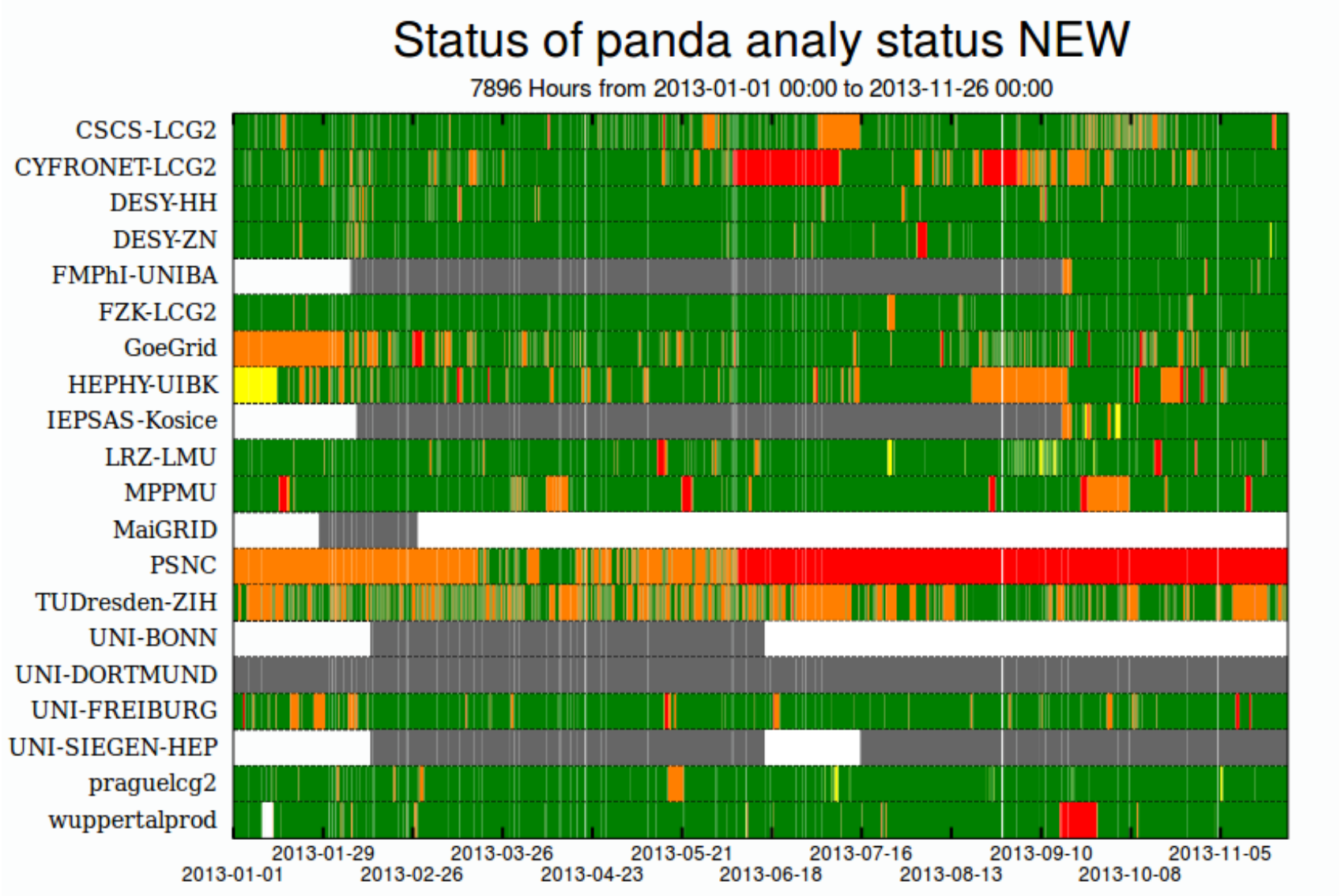
- Relational databases (mostly Oracle) are currently working well
- Many use cases might be more suitable for NoSQL solution (Hadoop: DDM accounting, Event service possible candidate)
- Frontier/Squid fully functional for all remote database access at all sites

Summary and Conclusions

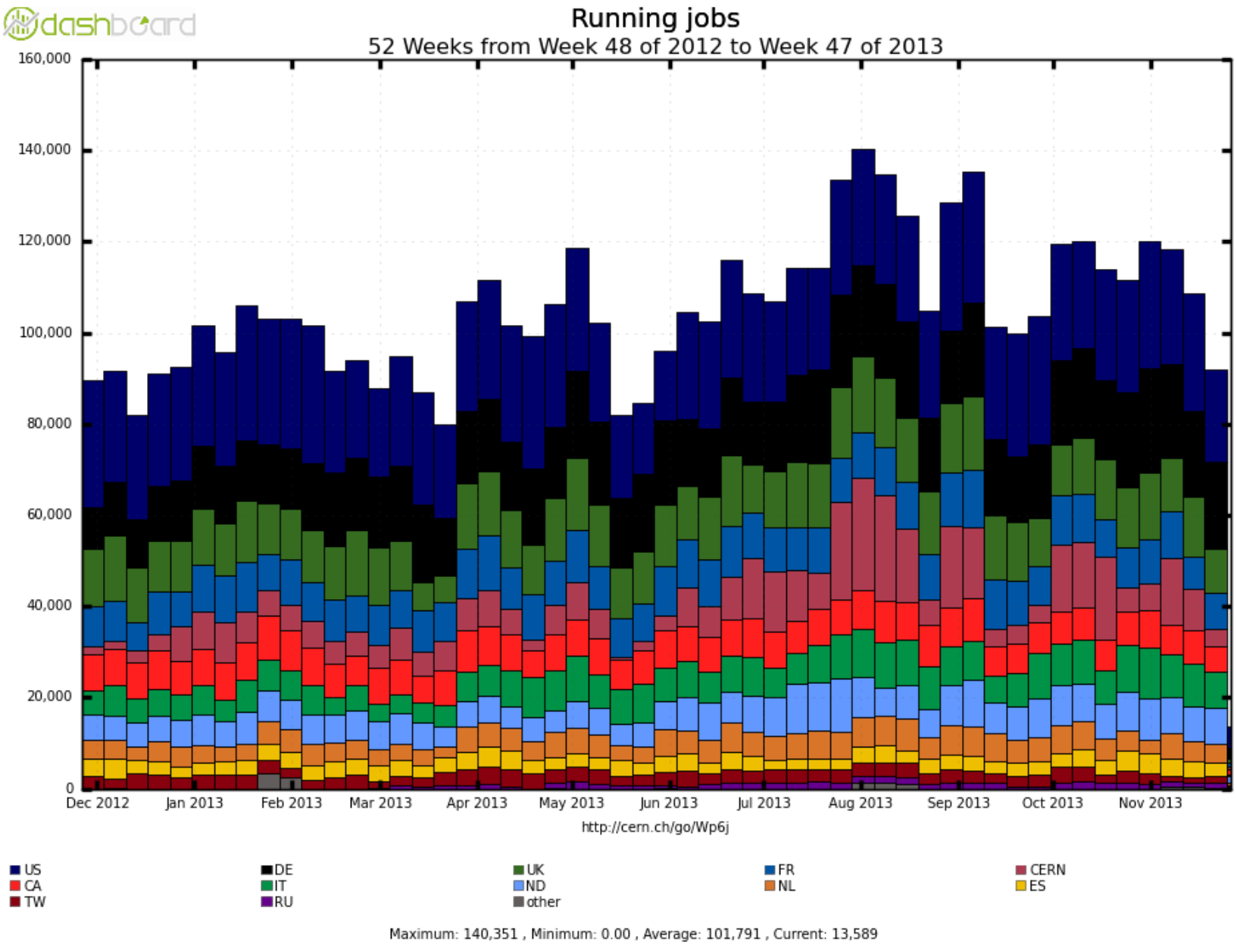
- ATLAS distributed computing development is driven by operations
- Many R&D projects:
  - quickly converge on possible usability in production
  - All R&Ds made it to production (NoSQL, FAX, Cloud Computing)
- Core components (Prodsys2 and Rucio) on schedule
- Model of incremental development steps and commissioning has been a key component for the success of Run1
- New analysis model next big challenge for 2014 before 2015 data taking

BACKUP

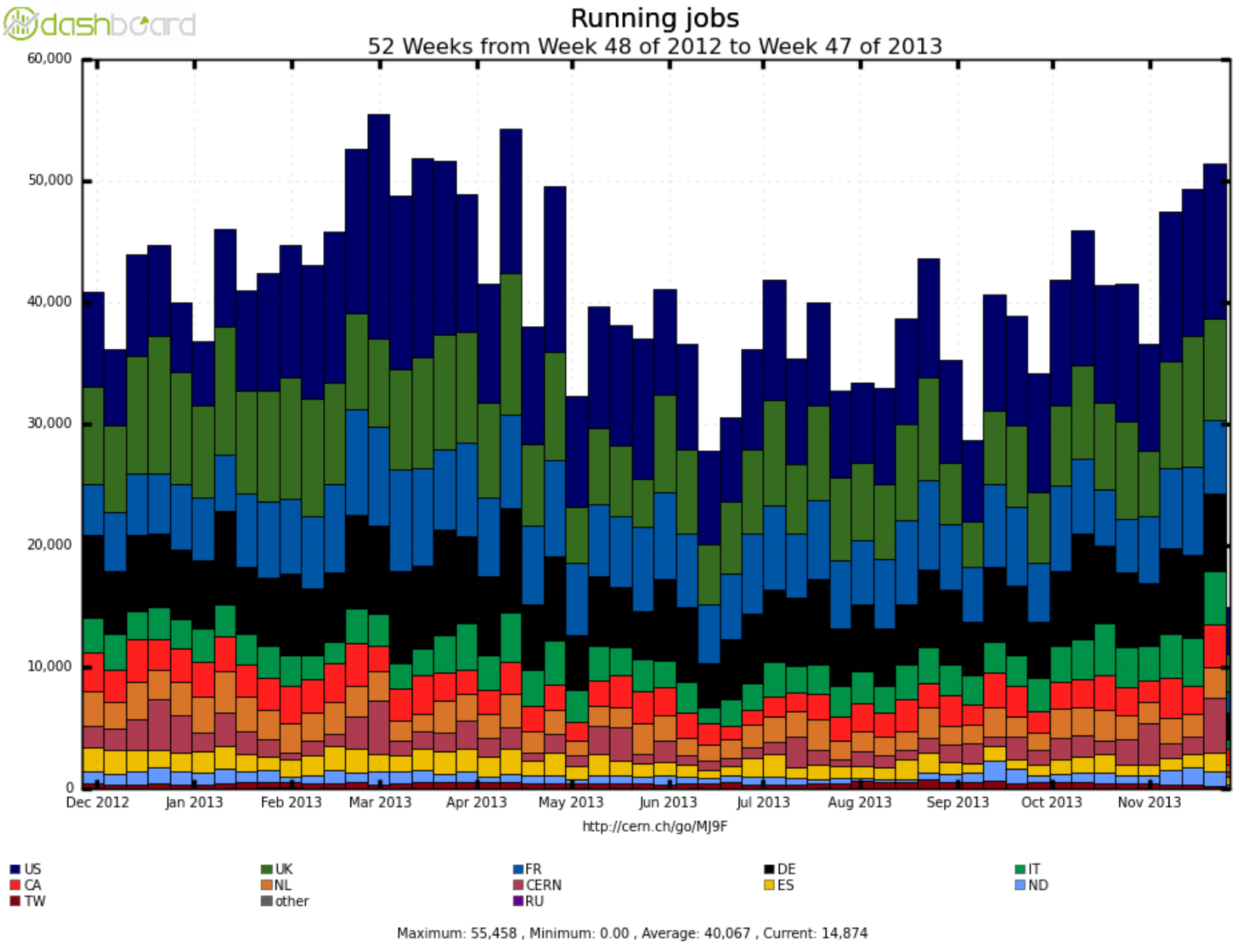
DE Cloud site availability Jan - Nov 13



Running production jobs - last 12 month



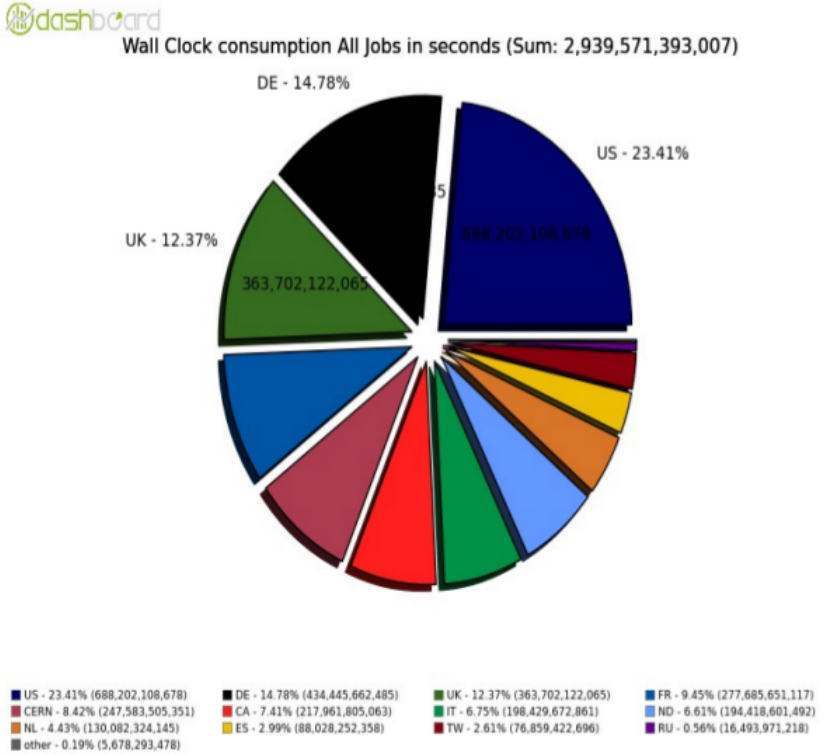
Running analysis jobs - last 12 month



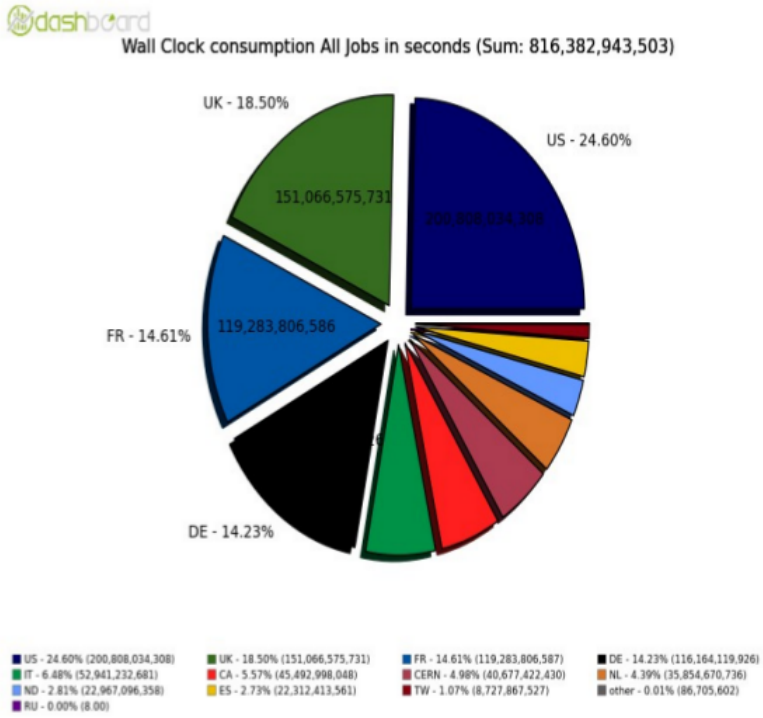
DE-cloud total for ATLAS (Oct12-Sep13)

- DE total: 2nd production (14.8%); 4th for Analysis (14.2%)

(2012 2nd, 14.3%)



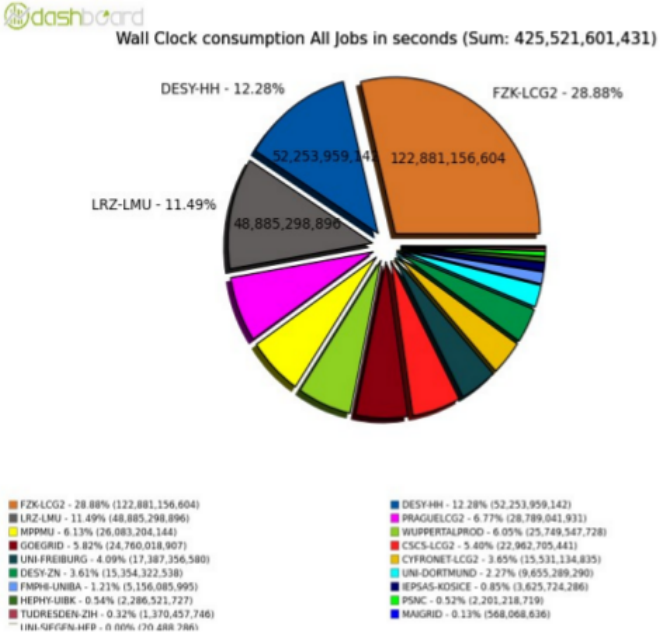
(2012: 3rd, 13.9%)



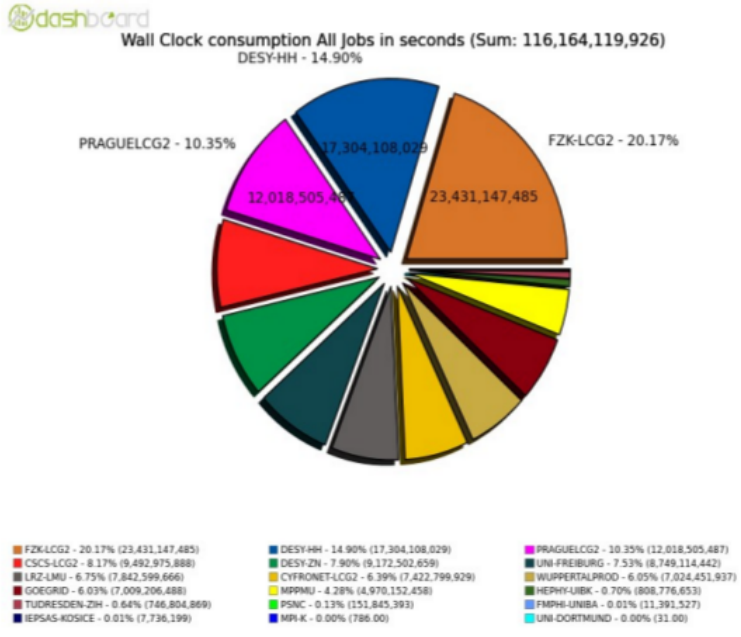
ATLASDE cloud by site Oct12-Sep13

- GridKa: production 29%, analysis 20%
- Desy/MPP : production 22%, analysis 27%
- DE University sites: production 30%, analysis 29%
- non-De sites: production 19%, analysis 24%

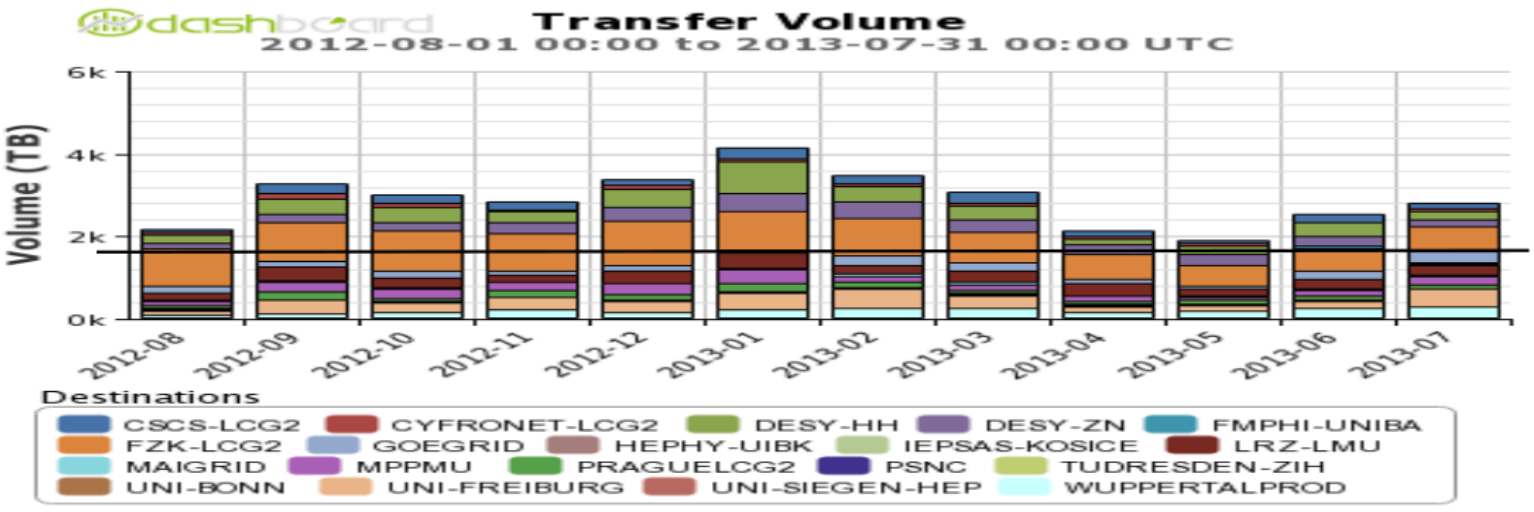
Production



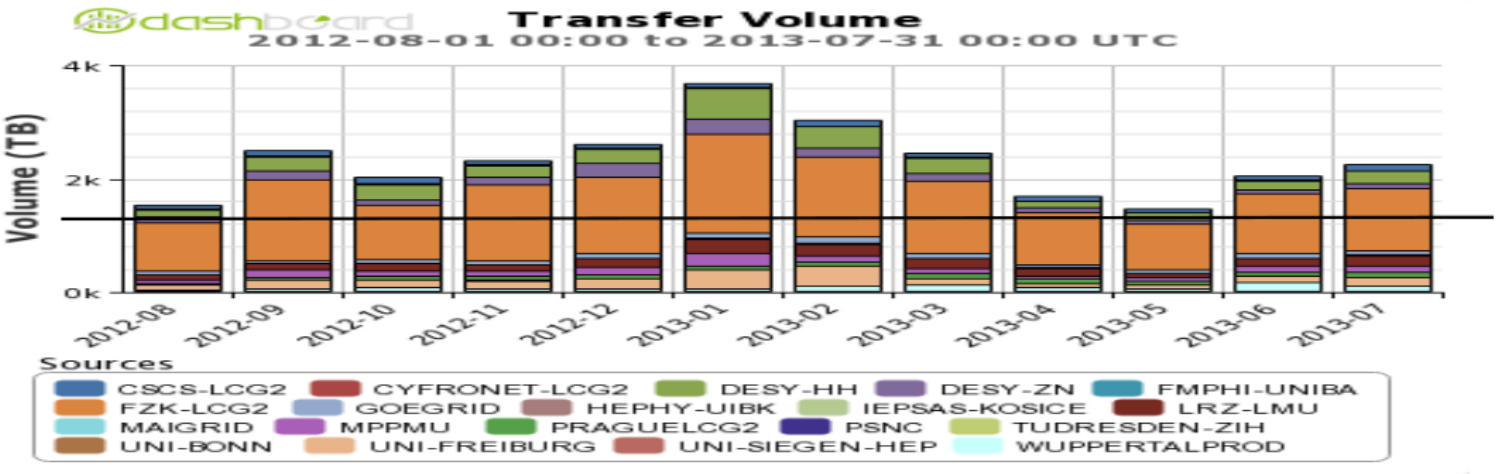
Analysis



ATLAS transfer to/from DE by site



2012



ATLAS transfer to/from DE by cloud