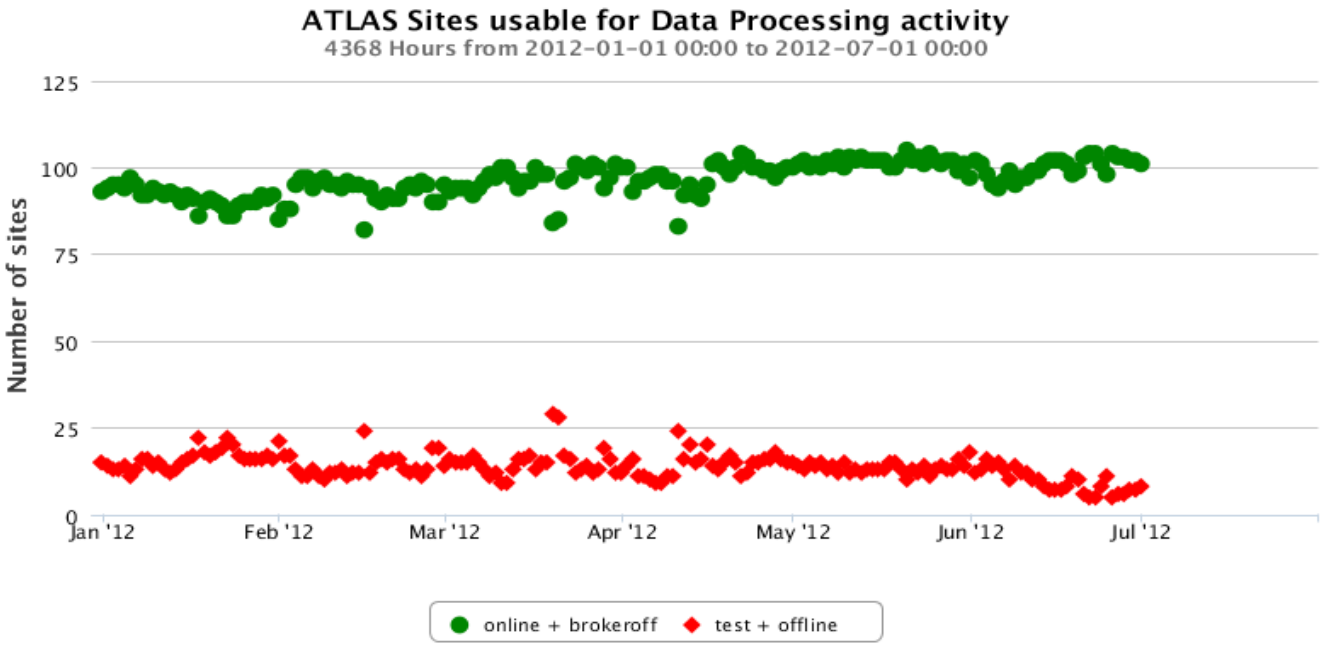
ATLAS Distributed Computing Automation

J. Schovancová[1], F. H. Barreiro Megino[2], C. Borrego[3], S. Campana[2], A. Di Girolamo[2], J. Elmsheuser[4], J. Hejbal[1,5], T. Kouba[1], F. Legger[4], E. Magradze[6], R. Medrano Llamas[2], G. Negri[1], L. Rinaldi[7], G. Sciacca[8], C. Serfon[4], D. Van Der Ster[2]
on behalf of the ATLAS Collaboration

[1] Institute of Physics, AS CR, Prague
[2] CERN, Geneva
[3] Universidad Autonoma de Madrid, Madrid
[4] Ludwig-Maximilians-Universitaet, Muenchen
[5] Czech Technical University, Prague
[6] II. Physikalisches Institut, Georg-August Universitaet, Goettingen
[7] Instituto Nazionale Fisica Nucleare, Bologna
[8] University of Bern, Bern

Grid'2012, Dubna, Russia
16-20 July 2012

ATLAS Computing resources



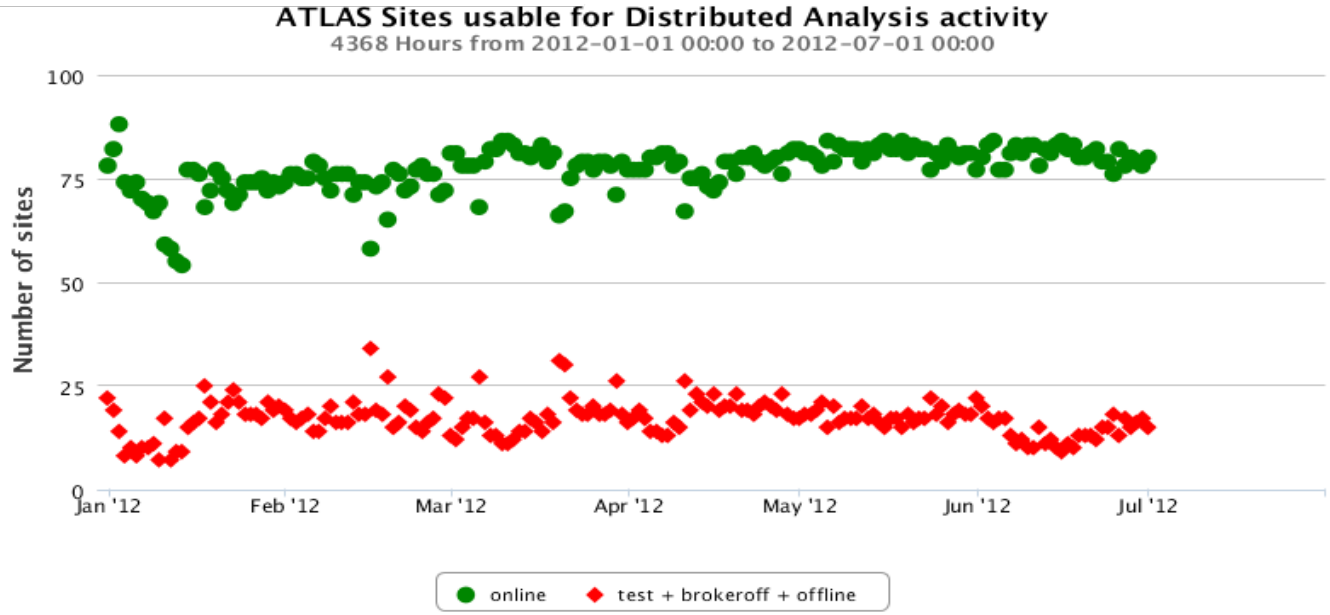
Over 120 grid sites distributed world-wide

Over 100 PB of storage space

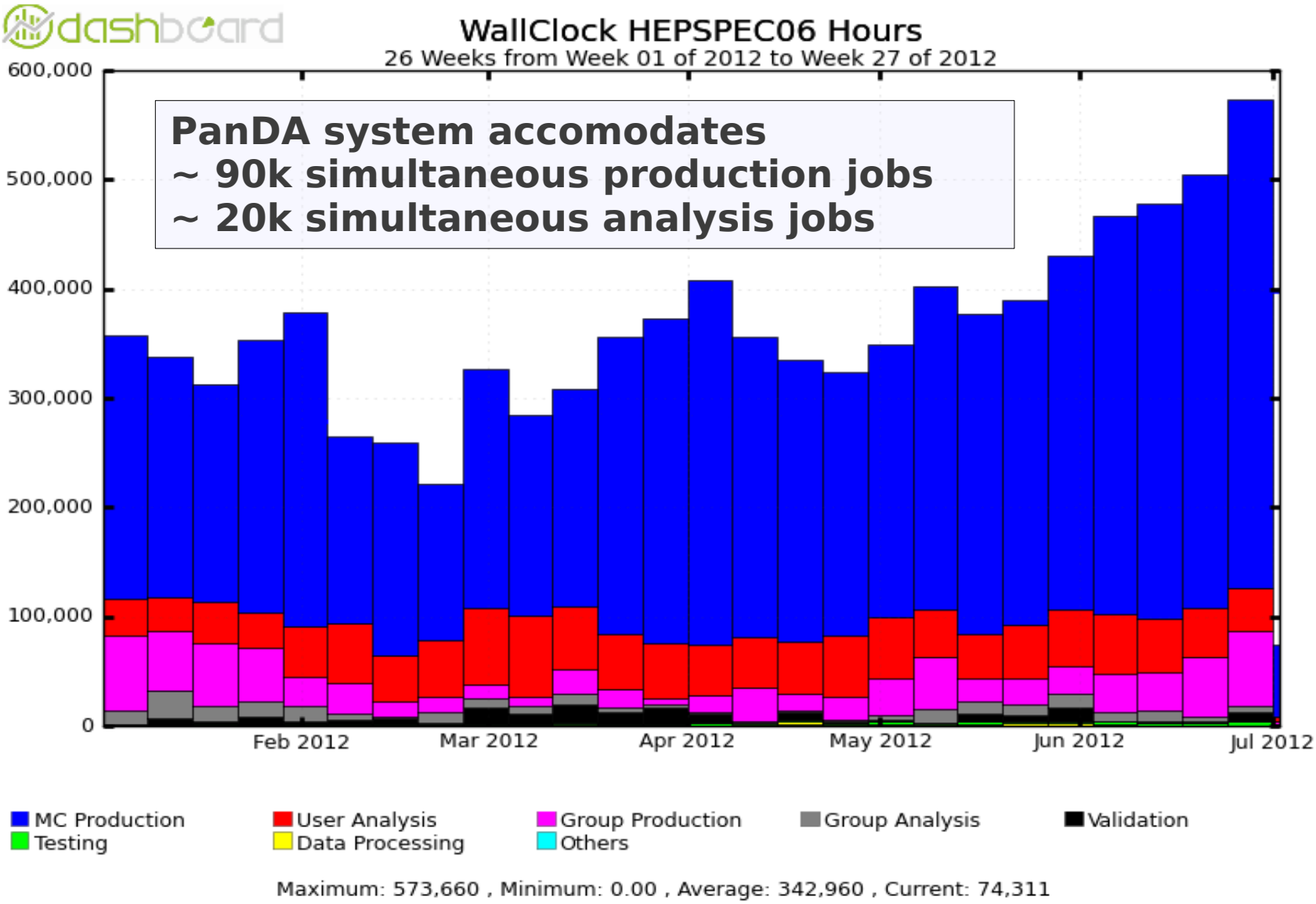
Over 100 k job slots

Computing activities

- Data Transfer
- Data Processing
- Distributed Analysis



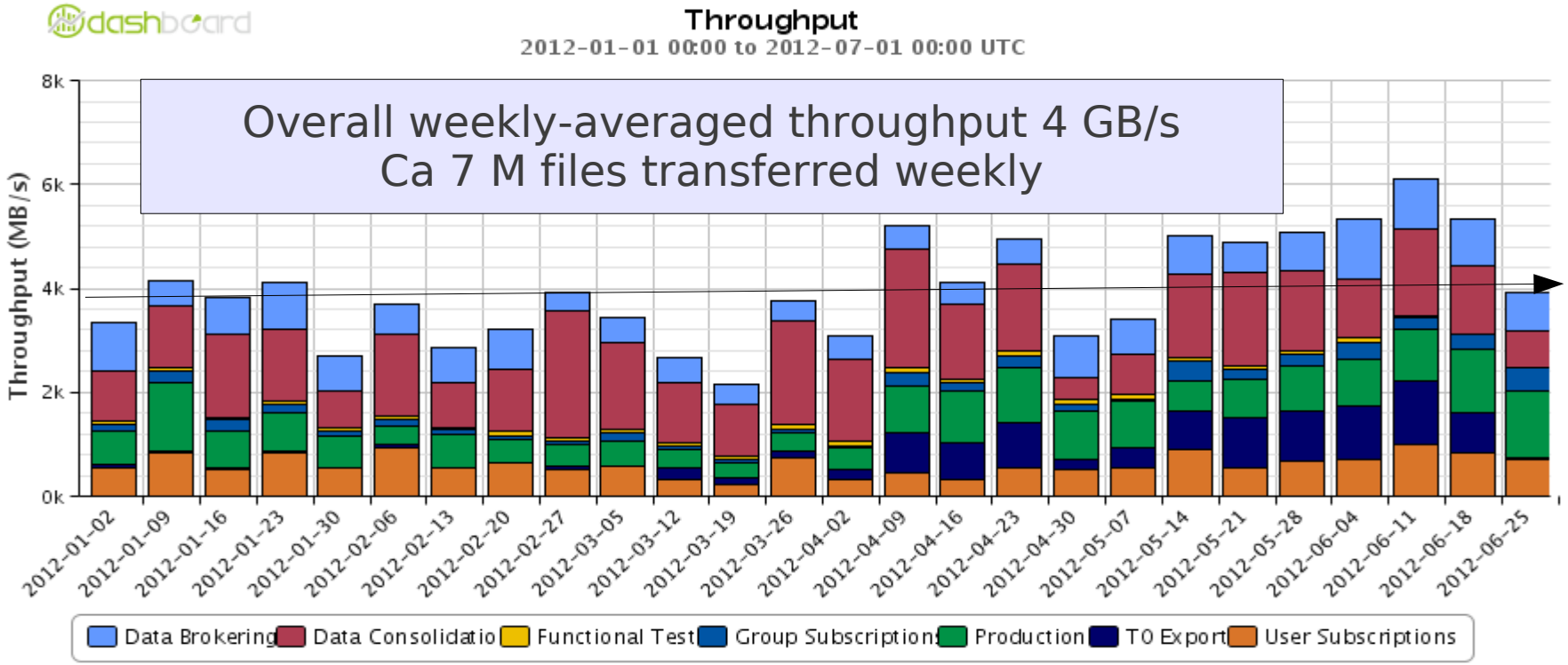
ATLAS Data Processing and Analysis



10 % of jobs (5 % of walltime) fail and are rerun

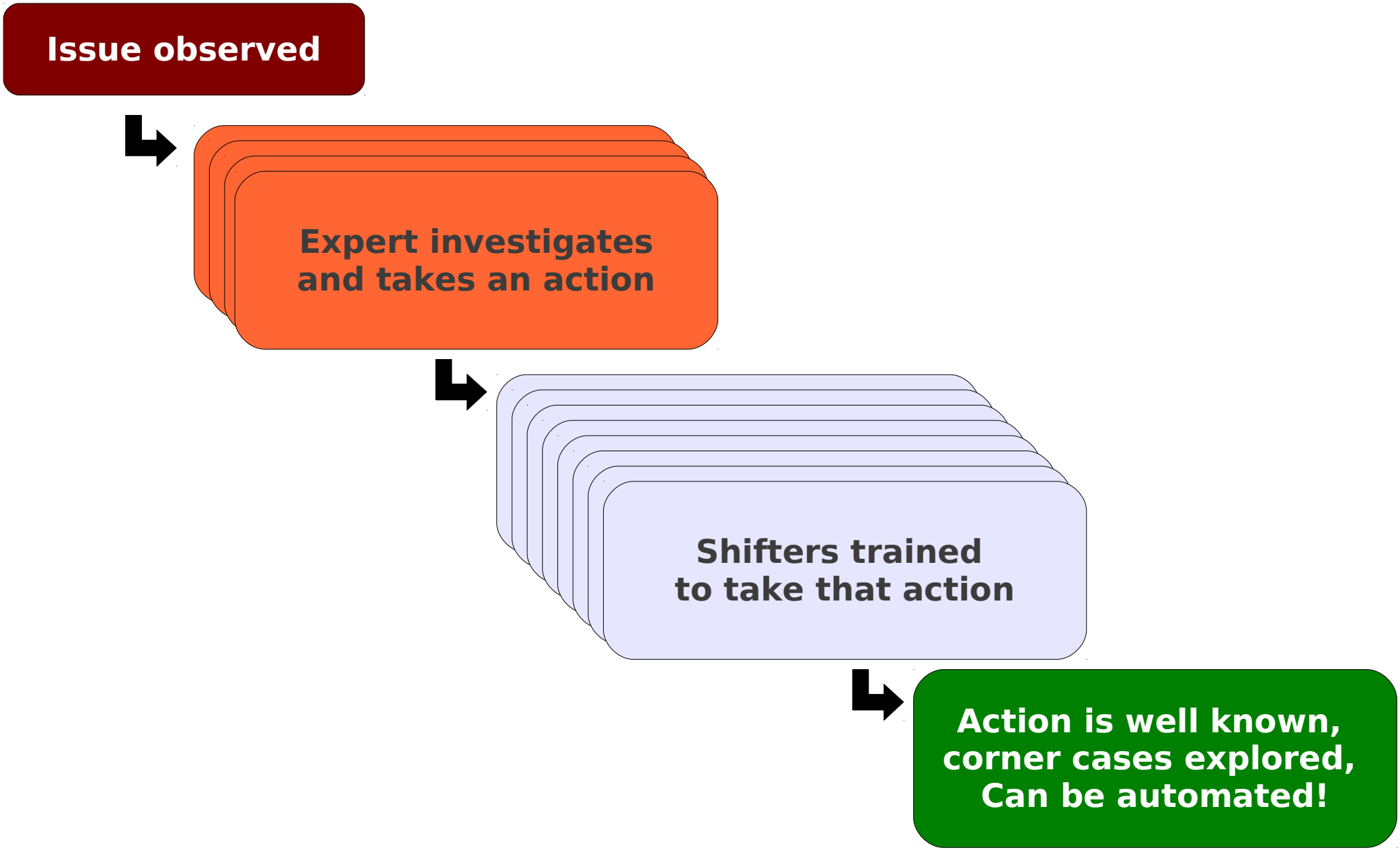
ATLAS Data Transfers

- ATLAS Distributed Data Management (DDM) system
- Data distribution: Preplacement, Dynamic placement, User requests



- Throughput: Peaks over 10 GB/s (10min average), 4 GB/s weekly average
- Average success rate in 2012 Jan-Jun
  - 92 % on the first attempt, 100 % after several retries

Automation workflow



Monitoring the ATLAS Grid Resources

- ADC Operations Team: Monitor ATLAS Grid Resources 24x7
  - 24x7 shift and expert team, site administrators, cloud squads
    - "Human expertise" is essential to smoothly operate, very useful contribution!
- Report issues to sites and cloud squads
  - 6700 GGUS tickets to the sites since 1st Jan 2010
  - Interaction with the cloud squads via e-groups, savannah tickets, ADC meetings
- Functional testing
- Autoexclusion, autorecovery

ADC Monitoring

Data Management | Data Processing | Databases | Point-1 | Sites and Services

ATLAS Grid Information System: AGIS

- List of ATLAS Grid sites
  - Map of services at sites, relations between services
  - Service status and downtime information

Which Activities rely on Services at an ATLAS Site?

Service	Activity		
	Distributed Analysis	Data Processing	Data Transfers
Storage Element(s)	X	X	X
Computing Element(s)	X	X	
File Transfer Service			X
LCG File Catalog	X	X	X
Frontier	X	X	
Squid	X	X	
Network Connectivity			X

Site Exclusion: Service Downtime

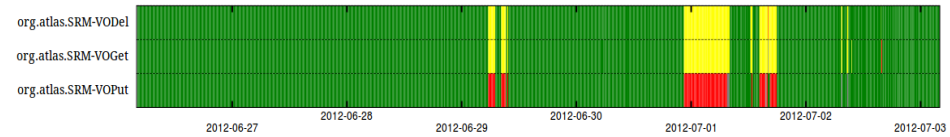
- Downtimes published in GOCDB/OIM, fetched to AGIS
  - downtime for each service
    → downtime for related services
  - **Exclude site from Data Transfer activity with the downtime**
  - **Exclude site from Data Processing and Distributed Analysis activities with downtime of a SE/CE**
    - ***~30 site exclusion/re-inclusion events per week***
- Stop activity when resources not available
  - **Exclude site for write when it's about to get full**
    - ***~6 site exclusion/re-inclusionevents per week***
- Can automatically react on unscheduled downtimes
  - decrease number of manual interventions by the Ops team
  - **Exclude resources from activities to prevent more failures**
    - ***~30 site exclusion/re-inclusion events per week***

Service Functional Testing

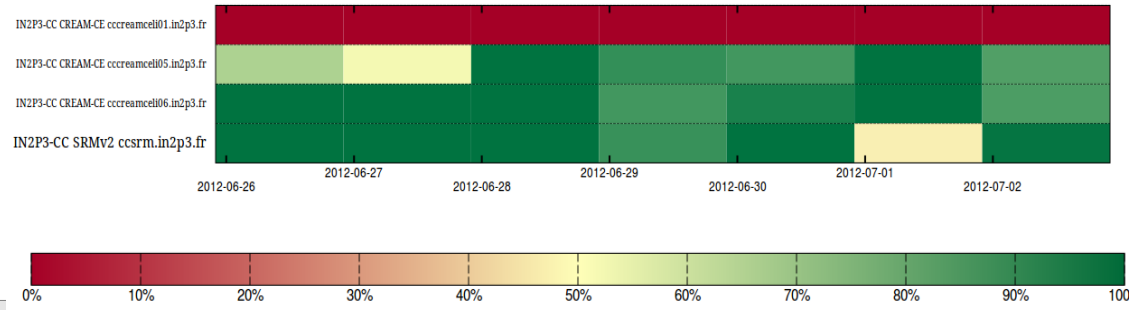
- SAM/Nagios probes to **test availability of services**
  - Critical – used for the WLCG reports
    CE, SE, SW area availability
  - Development – ATLAS internal purposes only
    Frontier, DDM spacetoken, pilot job submission efficiency, FTS, LFC

Plan to exclude services when functional tests are failing.
Need carefully study of *treshold* which distinguishes *failing* service from "*false positives*".

SAM/Nagios Test results for a SRM

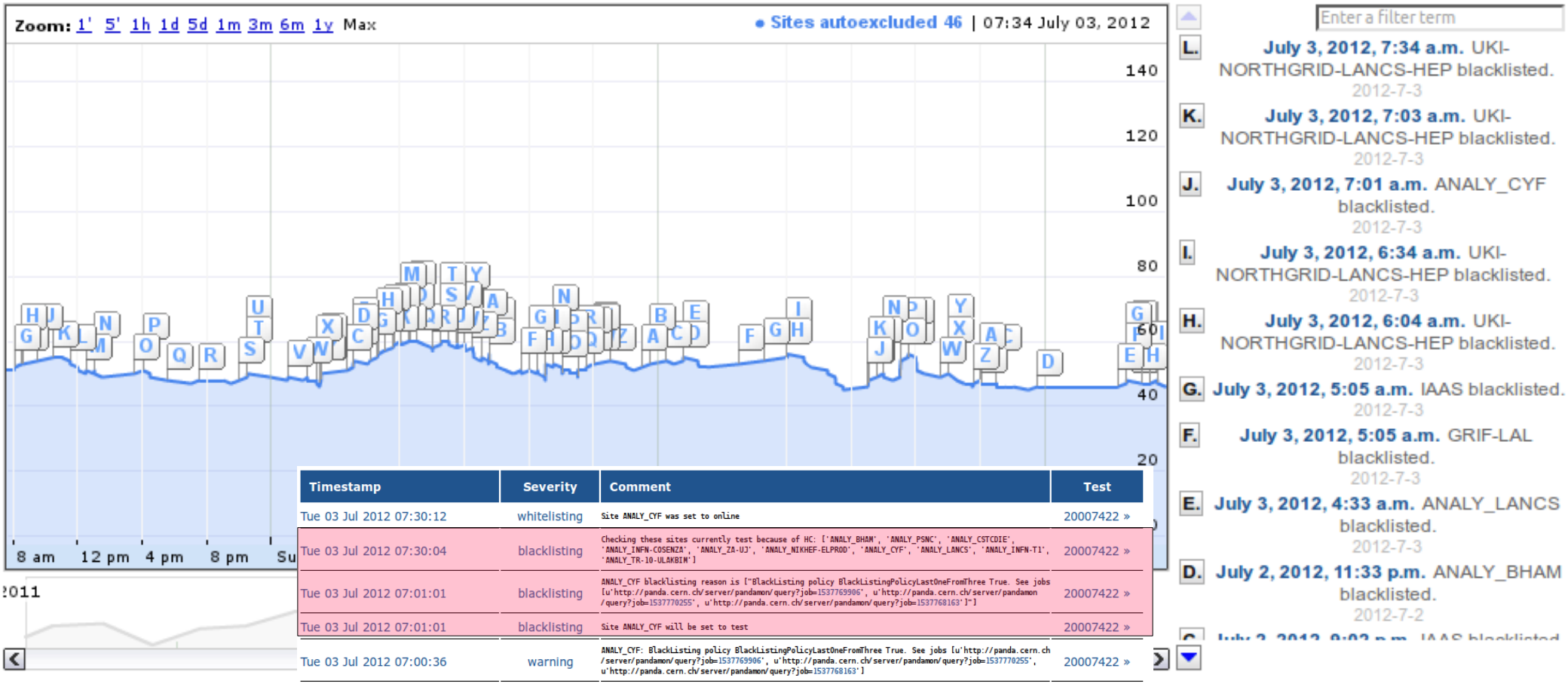


SAM/Nagios Services reliability



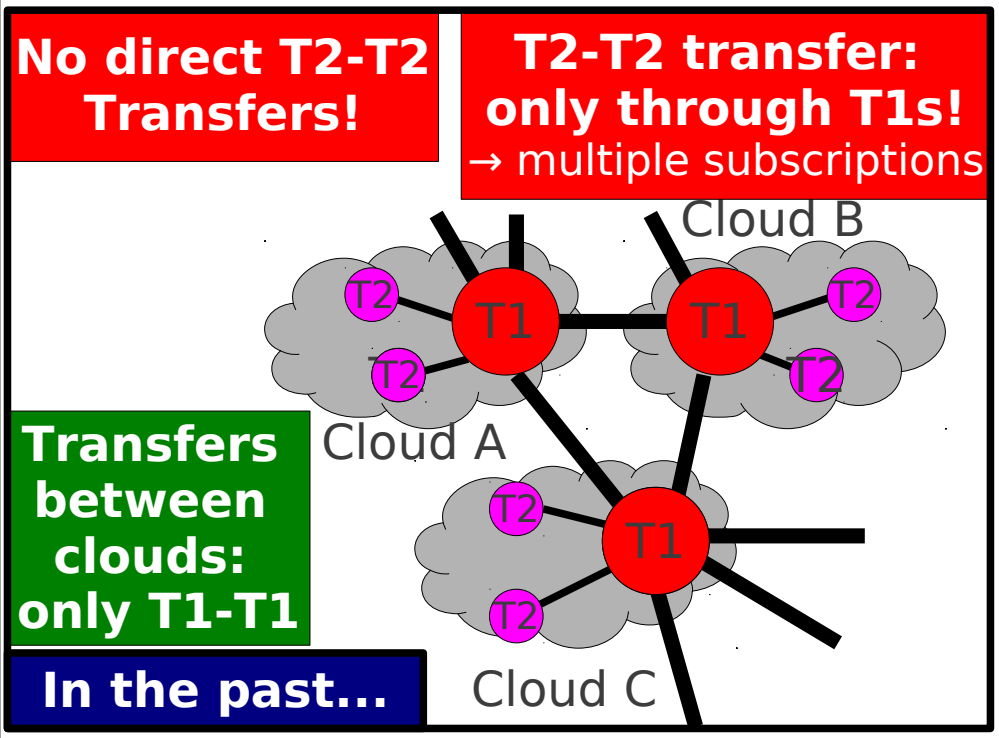
Testing job environment at sites: HammerCloud

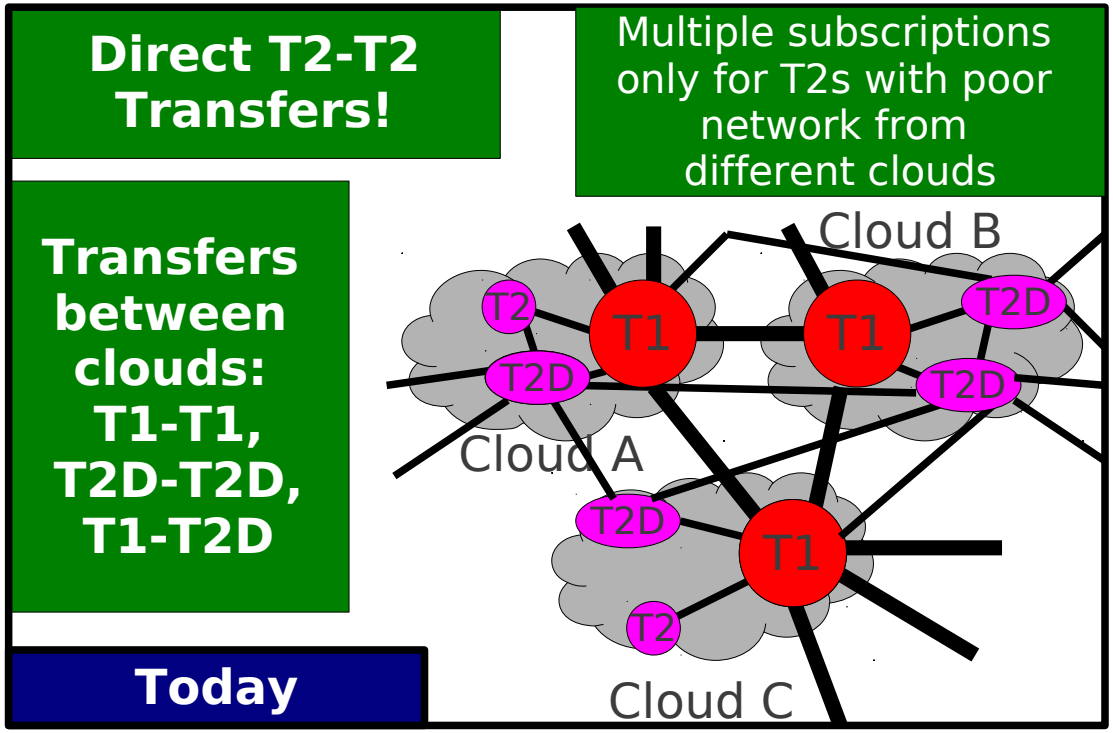
- **Functional tests for Production and Analysis jobs**
  - Same environment and similar workload as the real jobs
  - Continuous flow of jobs, several flavours of jobs within 1 hour
    - ***~240 site exclusion/re-inclusion events per week***



Testing DDM and network throughput

- Sonar tests: periodically measure overall transfer throughput. Goal: **optimise transfer path**
  - Small (~MB), Medium (~ 100s MB), Large (~GB) files
  - Test each pair of DDM endpoints → ~10k pairs!





- PerfSonar tests: measure network characteristics between 2 (storage) endpoints: Only ~15 sites involved, ~ 220 pairs

Service Recovery

- Continuous flow of functional tests
  - DDM: transfers between T1-T2s, transfers between each 2 endpoints
  - Queues: HammerCloud functional tests AFT, PFT
  - Worker Nodes: environment sanity check right before the job computation starts
  - Service availability: SAM/Nagios tests

- Defined Exclusion Policy
  - How many tests in a row can a site fail to stay included
  - After issue fixed what tests to perform, how many tests need to succeed to recover site for an Activity

Summary and Outlook

- ATLAS Distributed Computing successfully operates computing resources (>100k jobslots, >100 PB storage space) at over 100 grid sites world-wide
- Services are excluded/recovered automatically from the Activities: Data Transfer, Distributed Analysis, Data Processing
- Functional tests available for the Activities and for Services
- Rate of available automation allows to spot existing issues
  - Easier to identify the ones requiring action
    - By ATLAS or by the grid site
  - Can “hide” issues from the physics community
  - Concentrate on the remaining issues
    → improve reliability of the system even further