



ATLAS Distributed Computing Automation

J. Schovancová¹, F. H. Barreiro Megino², C. Borrego³, S. Campana², A. Di Girolamo²,
J. Elmsheuser⁴, J. Hejbal^{1,5}, T. Kouba¹, F. Legger⁴, E. Magradze⁶, R. Medrano Llamas²,
G. Negri¹, L. Rinaldi⁷, G. Sciacca⁸, C. Serfon⁴, D. Van Der Ster²
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

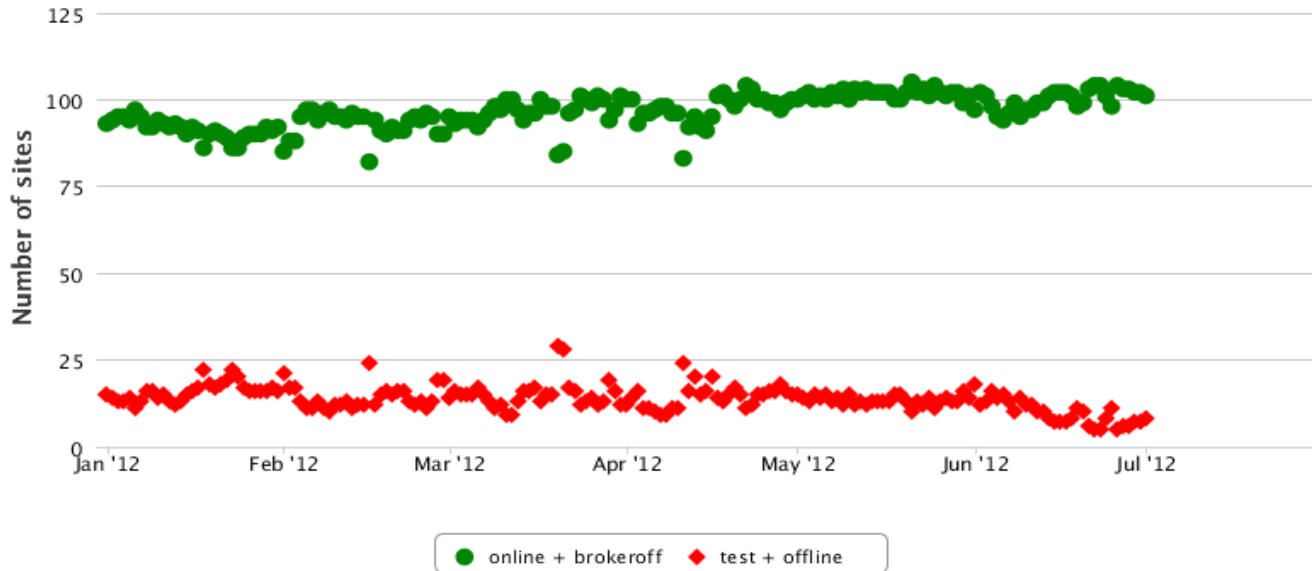
⁷ Istituto Nazionale Fisica Nucleare, Bologna

⁸ University of Bern, Bern

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

ATLAS Computing resources

ATLAS Sites usable for Data Processing activity
4368 Hours from 2012-01-01 00:00 to 2012-07-01 00:00

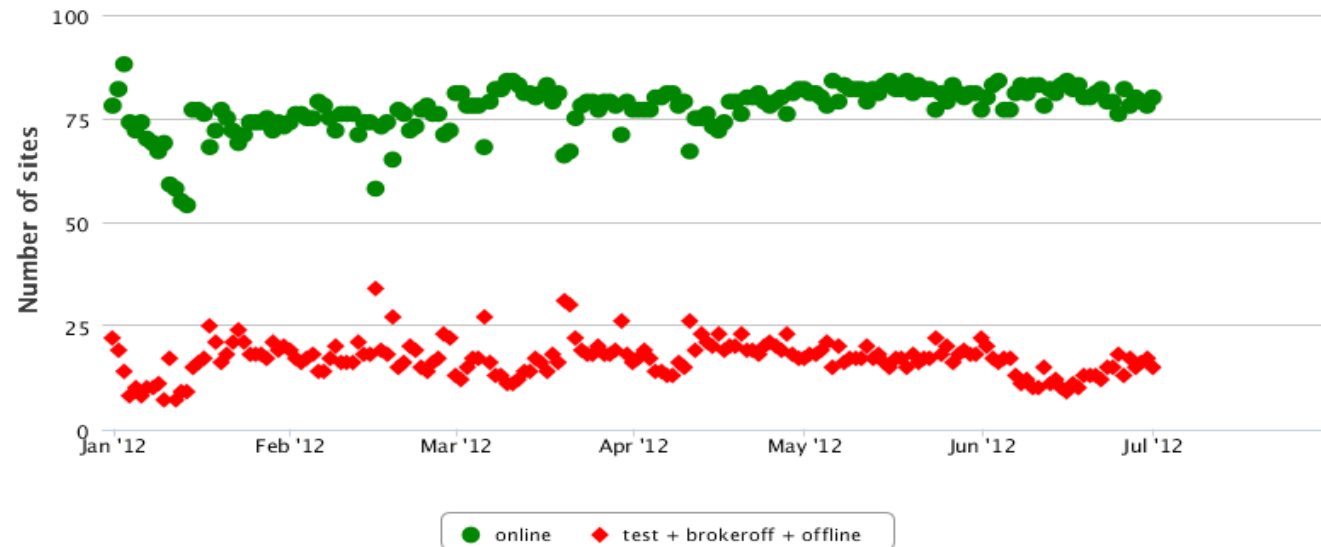


Over 120 grid sites distributed world-wide

Over 100 PB of storage space

Over 100 k job slots

ATLAS Sites usable for Distributed Analysis activity
4368 Hours from 2012-01-01 00:00 to 2012-07-01 00:00



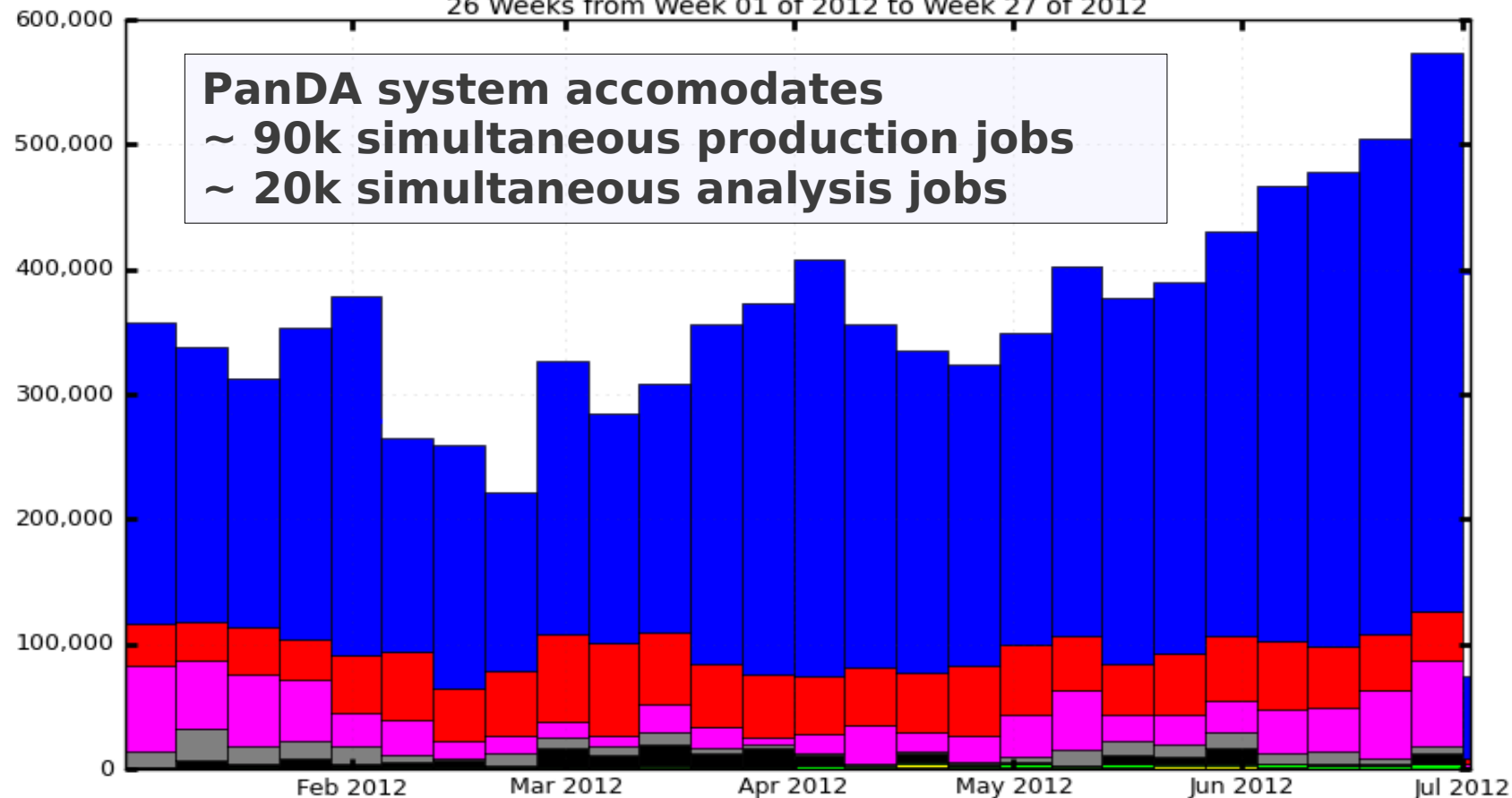
Computing activities

- Data Transfer
- Data Processing
- Distributed Analysis

ATLAS Data Processing and Analysis



WallClock HEPSPROC06 Hours
26 Weeks from Week 01 of 2012 to Week 27 of 2012



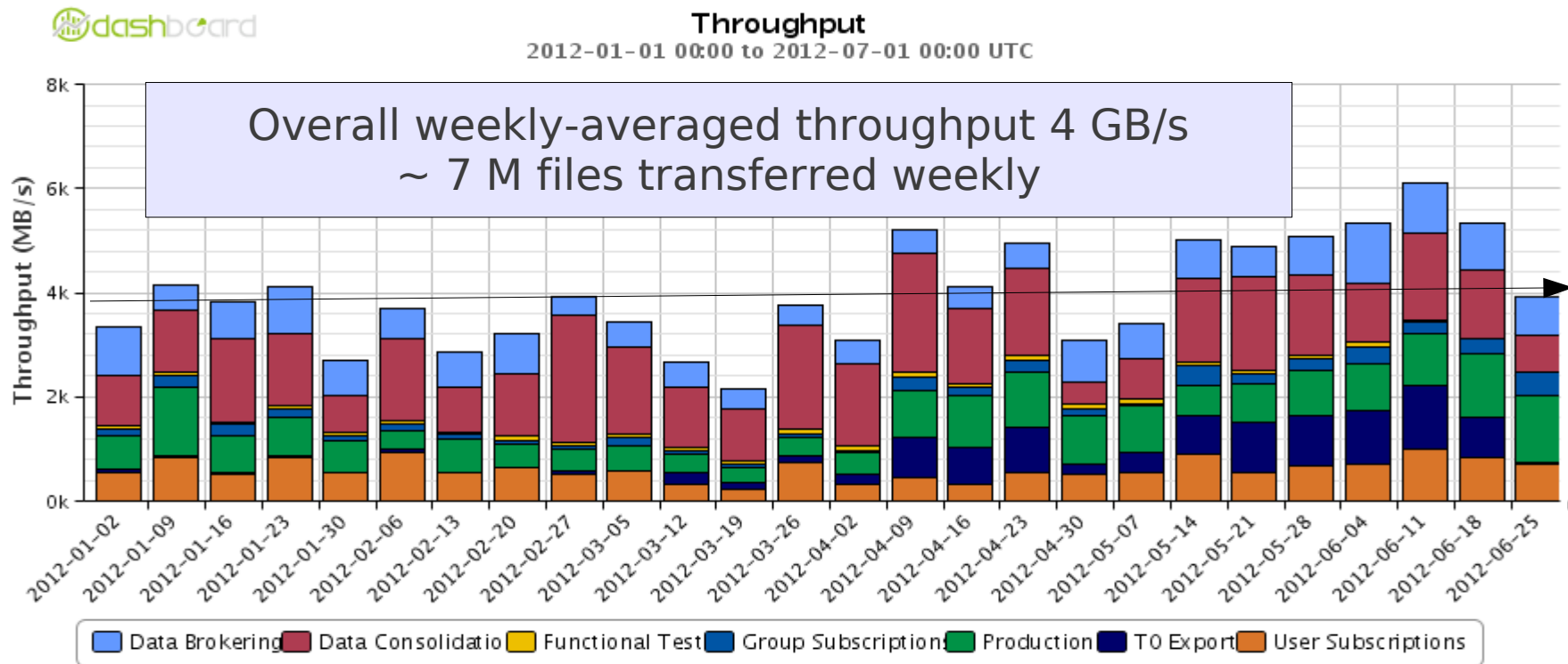
■ MC Production ■ User Analysis ■ Group Production ■ Group Analysis ■ Validation
■ Testing ■ Data Processing ■ Others

Maximum: 573,660 , Minimum: 0.00 , Average: 342,960 , Current: 74,311

~ 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)

Automation timeline

Issue observed



Expert investigates and takes an action



Shifters trained to take that action

New corner case



Action is well known, corner cases explored, Can be automated!

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 fundamental to smoothly operate, their contribution is essential!
- 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

~ 7 GGUS tickets/day

ADC Monitoring

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



- Functional testing
- Autoexclusion, autorecovery

ATLAS Grid Information System: AGIS

- 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

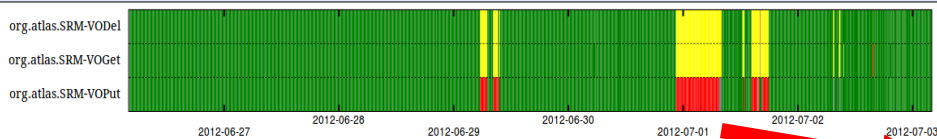
- Downtimes published in GOCDDB/OIM, fetched by AGIS
 - downtime for each service
 - downtime for related services
 - **Exclude site with SE downtime from Data Transfer activity**
 - **Exclude site with downtime of a SE/CE from Data Processing and Distributed Analysis activities**
 - *~30 site exclusion/re-inclusion events per week*
- Stop activity when resources not available
 - **Exclude site for write when available space lower than predefined thresholds**
 - *~6 site exclusion/re-inclusion events 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

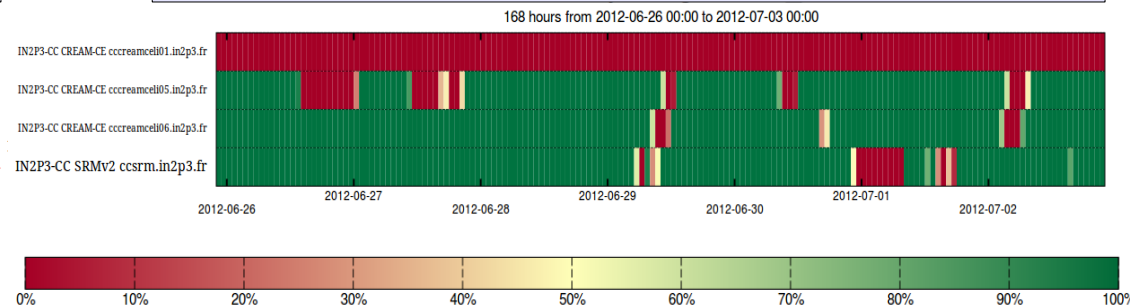
- SAM/Nagios probes to **test availability of services**
 - Critical – used for the WLCG reports
CE, SE, SW area availability
 - Development – ATLAS internal purposes only
Frontier, DDM spacetoken, pilot job submission efficiency, FTS, LFC

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

SAM/Nagios Test results for a SRM

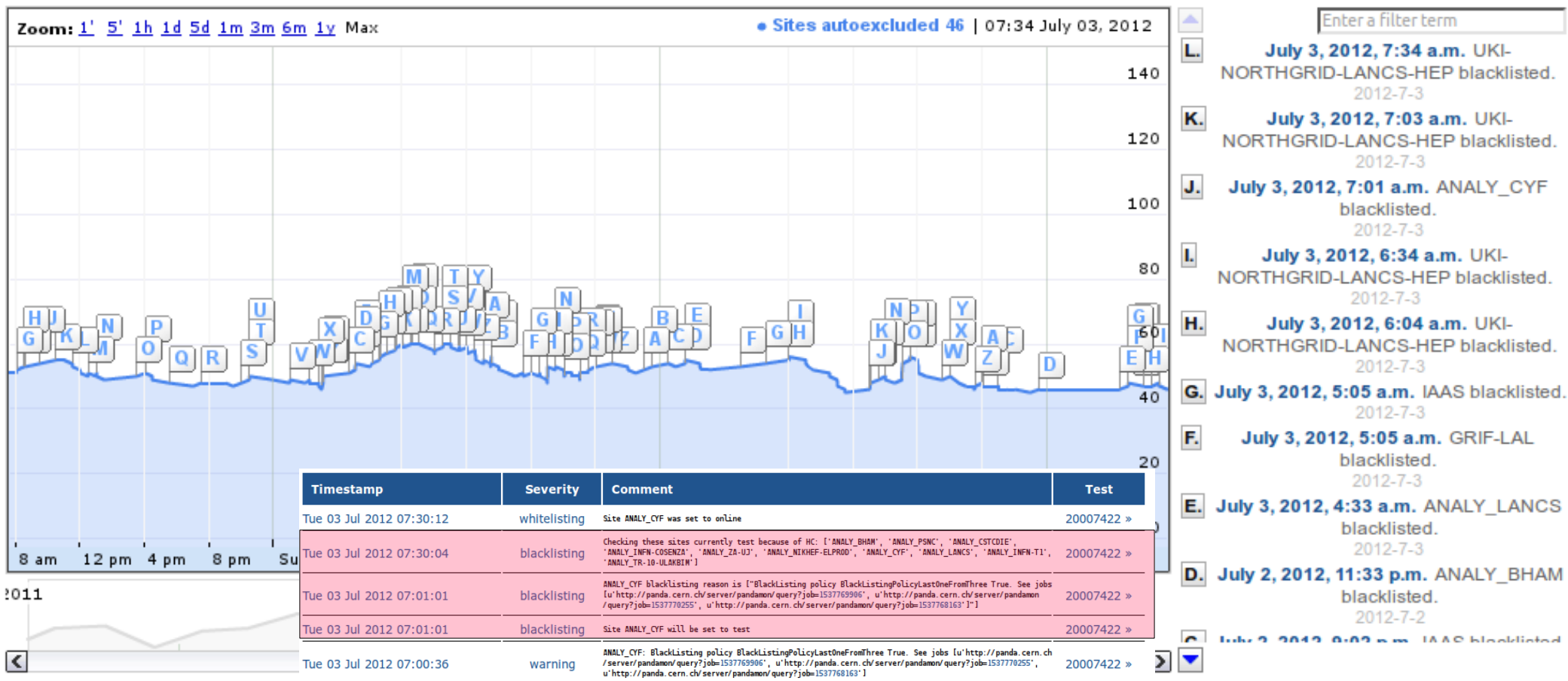


SAM/Nagios Services reliability



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**



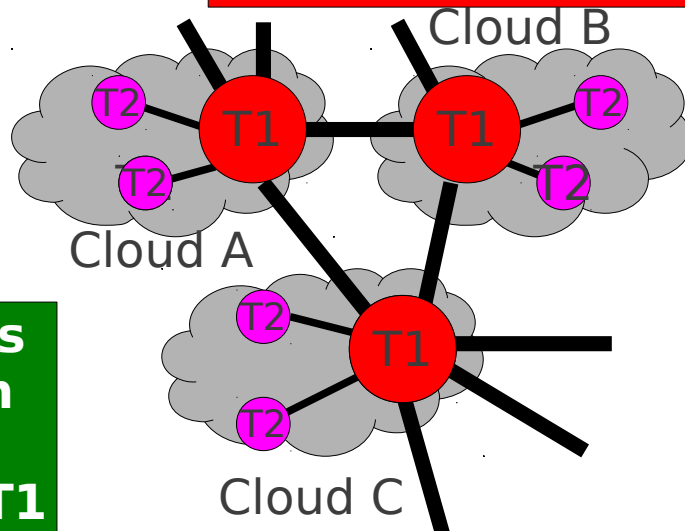
Testing DDM and network throughput

- 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!

In the past...

No direct T2-T2 Transfers!

T2-T2 transfer: only through T1s!
→ multiple subscriptions



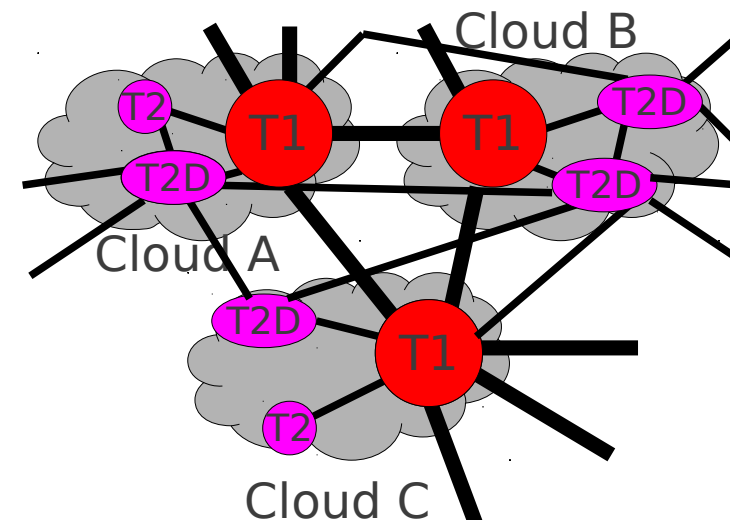
Transfers between clouds: only T1-T1

Today

Direct T2-T2 Transfers!

Multiple subscriptions only for T2s with poor network from different clouds

Transfers between clouds: T1-T1, T2D-T2D, T1-T2D



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

Service Recovery

- Continuous flow of functional tests (even when site excluded from activity)
 - DDM: 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 certain period of time a site may fail without being excluded from an Activity
 - After issue fixed: what tests to perform, how many tests need to succeed to recover site for an Activity

Summary and Outlook

- Functional tests available for the Activities and for Services
- Services are excluded/recovered automatically from the Activities: Data Transfer, Distributed Analysis, Data Processing
- 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**

ATLAS Distributed Computing successfully operates computing resources (>100k jobslots, >100 PB storage space) at over 100 grid sites world-wide