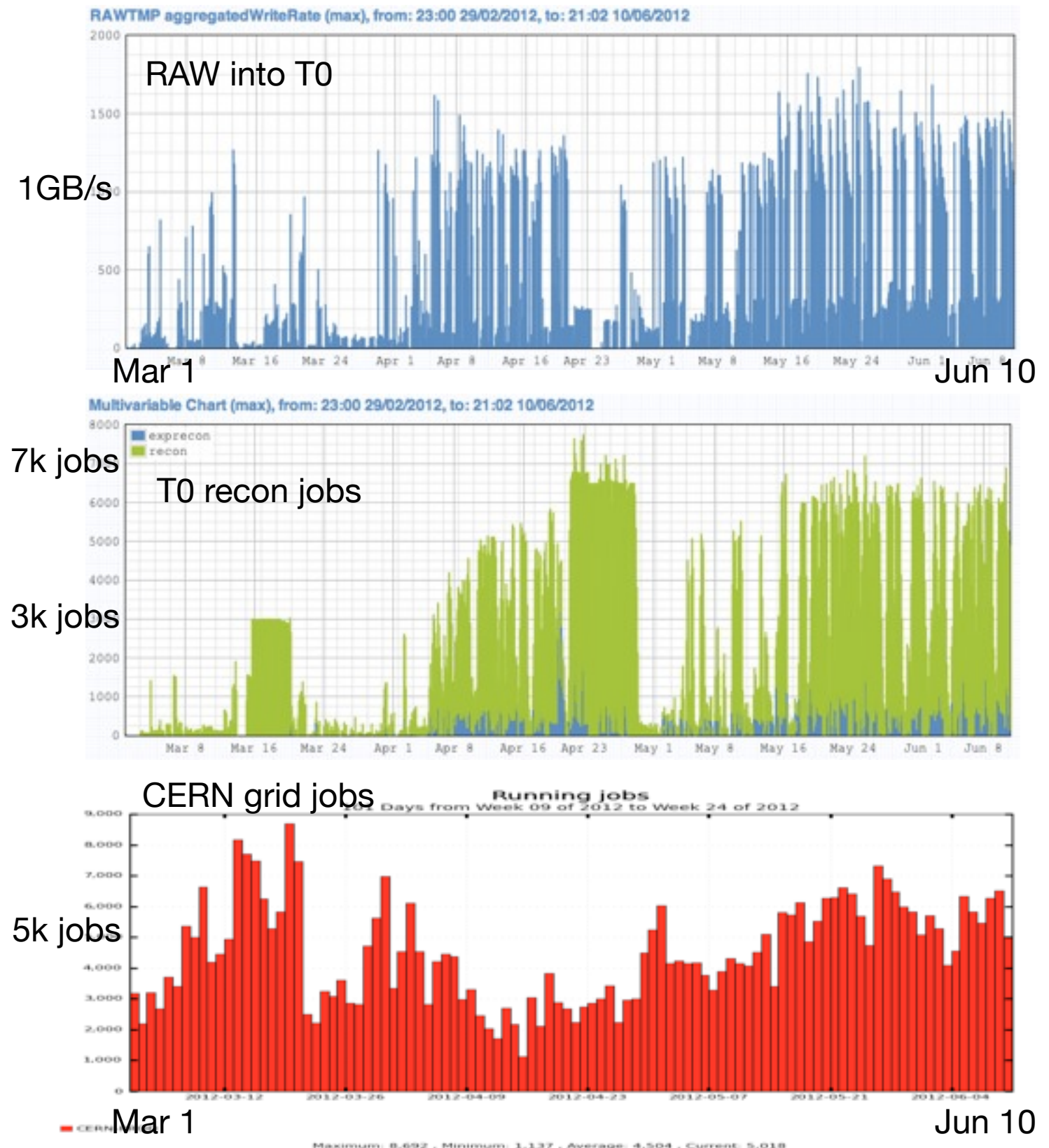# ADC Operations

Ueda I.
2012.06.11. S&C workshop

Tier-0 / CERN Analysis Facility

T0 has been recording RAW data at higher rate than last year (~1GB/s at peaks)

T0 has dedicated resources + spill-over usage onto grid part

- ran up to 6500 jobs during the MD+TS end of April
- GRID jobs filling the rest of available resources

RAWTMP aggregatedWriteRate (max), from: 23:00 29/02/2012, to: 21:02 10/06/2012

RAW into T0

1GB/s

Mar 1 — Jun 10

Multivariable Chart (max), from: 23:00 29/02/2012, to: 21:02 10/06/2012

- exprecon
- recon

7k jobs

T0 recon jobs

3k jobs

CERN grid jobs

Running jobs

Days from Week 09 of 2012 to Week 24 of 2012

5k jobs

Mar 1 — Jun 10

Maximum: 8.692 , Minimum: 1.137 , Average: 4.504 , Current: 5.018
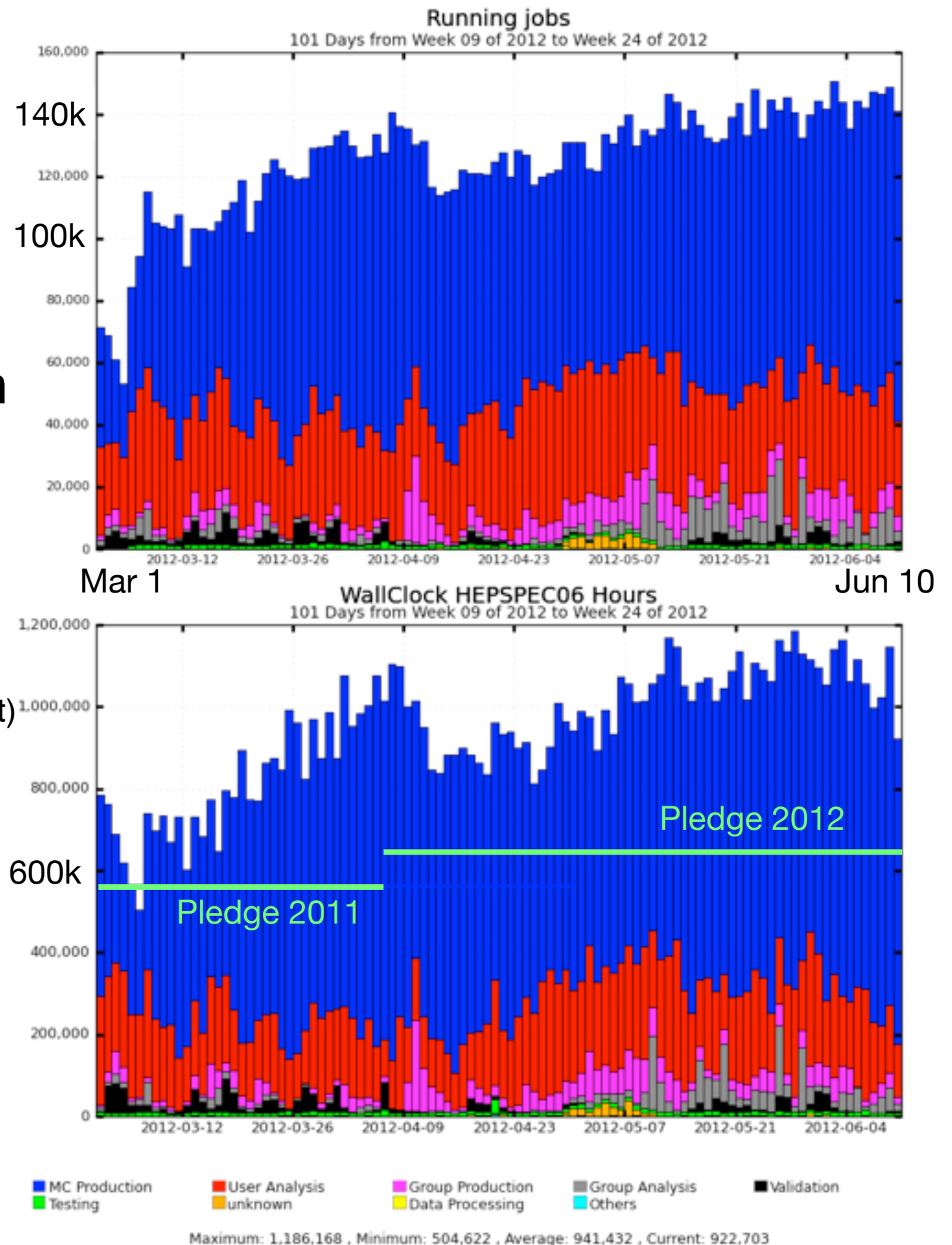
Grid Data Processing

ATLAS Grid processing
120-140 k jobs constantly

Able to use resources much
over promised ('pledged')
by sites

- Request for extra
  resources for May-June
  (ICB May09, ADC weekly May 15 Borut)

Panda parameter tuning

- t1weight being tuned
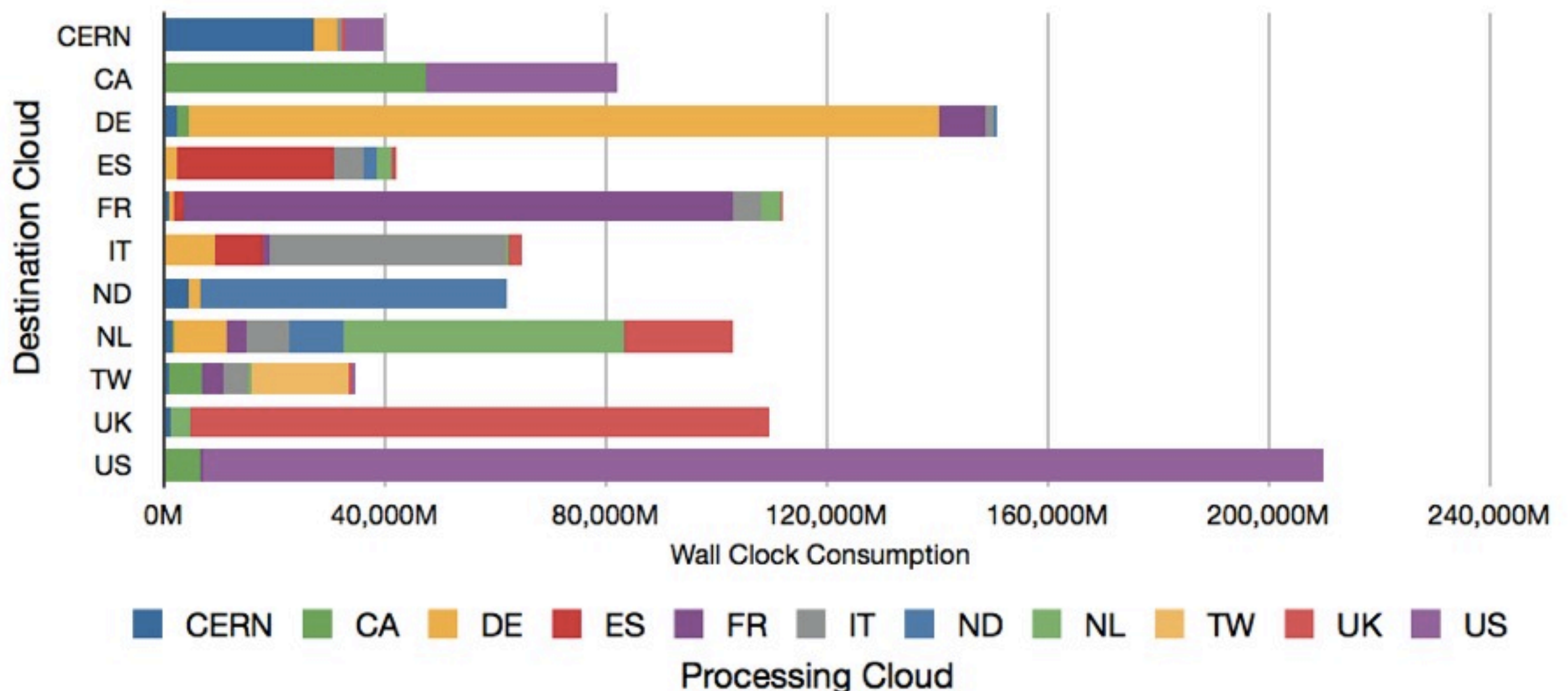- maxinputsize to be
  reviewed

GDP: Multi-cloud production

Wall-clock time consumption (Jan 01 - Jun 10)

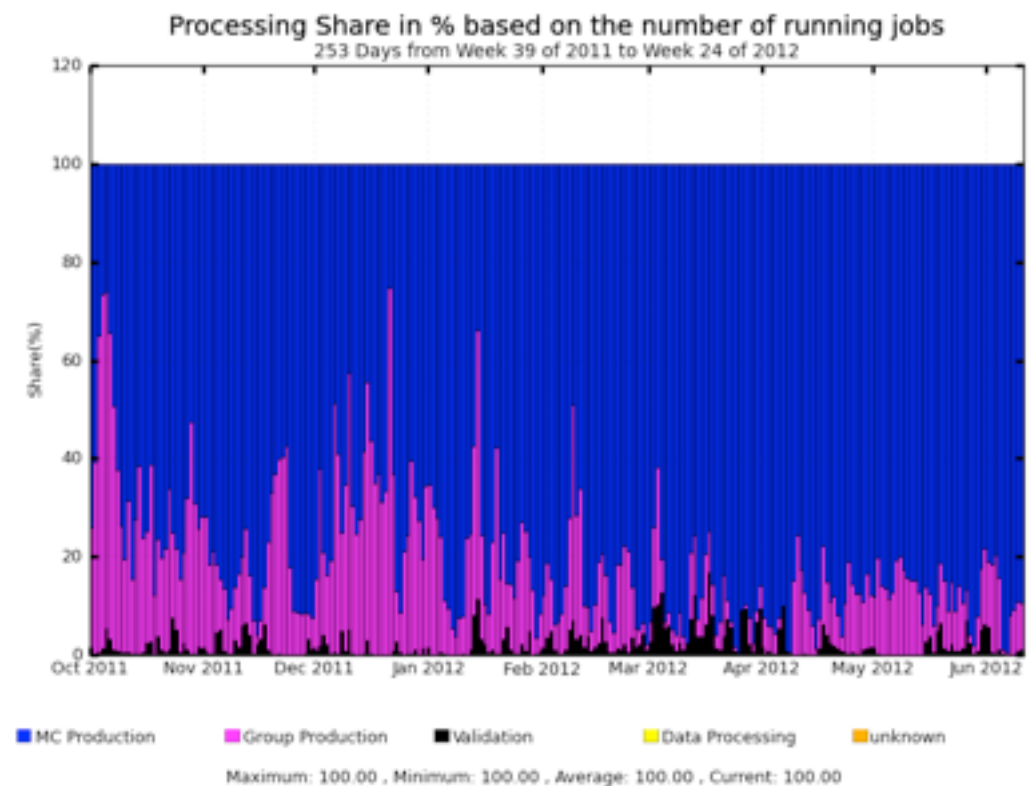|	CERN	CA	DE	ES	FR	IT	ND	NL	TW	UK	US
Processing	3.80%	6.16%	16.17%	3.84%	11.58%	6.74%	6.76%	6.12%	1.74%	12.84%	24.24%
Destination	3.91%	8.12%	14.93%	4.16%	11.09%	6.42%	6.13%	10.20%	3.42%	10.85%	20.78%
2011 share		5	11	5	13	10	5	12	6	11	21
2012 share		9	11	5	11	9	5	12	7	11	20

Multi-cloud production (2012 Jan01 - Jun10)

GDP: New stuff

Fair-share policy for different activities (since Jan 2012)

- T1: 20% for Group Production

- T2/3: 100% for evgensimul at some sites

  ‣ small sites with production only and no analysis

  ‣ small proddisk

  ‣ weak network connection to its T1

Processing Share in % based on the number of running jobs
253 Days from Week 39 of 2011 to Week 24 of 2012

■ MC Production ■ Group Production ■ Validation ■ Data Processing ■ unknown

Maximum: 100.00 , Minimum: 100.00 , Average: 100.00 , Current: 100.00

Task grouping (panda 'working groups')

- provenance + voms group name (eg. AP_phys-higgs)

  ‣ AP (Atlas Production), GP (Group Production), XP (Extra Production)

- implementation underway

Processing types: evgen([103897]) gangarobot-pft([11541]) gangarobot-pft-
Working groups: AP_BPhysics([457]) AP_Btagging([75355]) AP_EGamma(
AP_Susy([5126]) AP_Top([4415]) AP_Validation([5523]) AP_validation([4857])
GP_Higgs([4149]) GP_JetMet([1016]) GP_LAr([8182]) GP_MuDet([7116]) GP_
SM([23579]) Susy([30047]) Top([48]) XP_Higgs([3845]) gr_Higgs([10]) gr_SM(

GDP: Auto-exclusion

Many sites are not very stable

Two auto-exclusion mechanisms running (since Mar 2012)

- PFT -- Production Functional Tests
  - ‣ constant flow of standard test jobs
- Switcher
  - ‣ sets sites to offline before downtime
  - ‣ sets sites to test after downtime

Some small inconsistency between the two systems

- need a global view and adaptation to each other

Status of Panda Production status

GDP: Issues

Prodsys

- scalability in treatment of output dataset (freezing, adding to container, ...)

- chain of tasks

  ‣ one-by-one task request rather than a request for the full chain of tasks

  ‣ child task cannot start until parent task completion when numbers of input-output files are not 1:1

- job definition -- better be dynamic

- task type - group ntup production classified as reprocessing and reco

- Plan for new Job Definition system

  ‣ workshop next week

GDP: Issues

Panda

- shares vs priorities
  - shares can only be defined per site (no global shares) - can be set to small number of sites
  - brokerage not optimized for the fair-share policies yet - priorities may block shares
- schedconfig sometimes fragile -- problem in a site config can affect the whole system
  - integration with AGIS underway
- "extra production" still need some consideration

DDM-related

- multi-cloud production: some issues in transfers -- not really a MCP issue
  - some T2-3s have the problem even within the cloud
- Issues in uploading input for evgen
  - long standing headache for the production managers
  - a bug identified recently (bug #94630)

Others

- software installation system -- expertise is still in a single person

GDP: Operations

May need a review on workflow GDP - AMOD - ADCoS (senior + expert)

- why high priority tasks can stuck for weeks while we have shifters watching the system, tasks and jobs?

Production managers know about priorities and time line

- production management priorities are mapped to prodsys priorities
  - high priority simul may have lower prodsys priority than low priority evgen or reco
- production managers can raise prodsys priorities

GDP has a global view of the overall production

- can tell which prodsys priorities are "high priorities", but may not have enough time to follow everything
- may not know the time line of each task

ADCoS senior shifters can do the first level investigation and ticketing

- currently problem oriented than priorities
- need guidelines for priorities from GDP

Monitoring needs some improvements

- to list important tasks to shifters and see looping jobs (in progress)
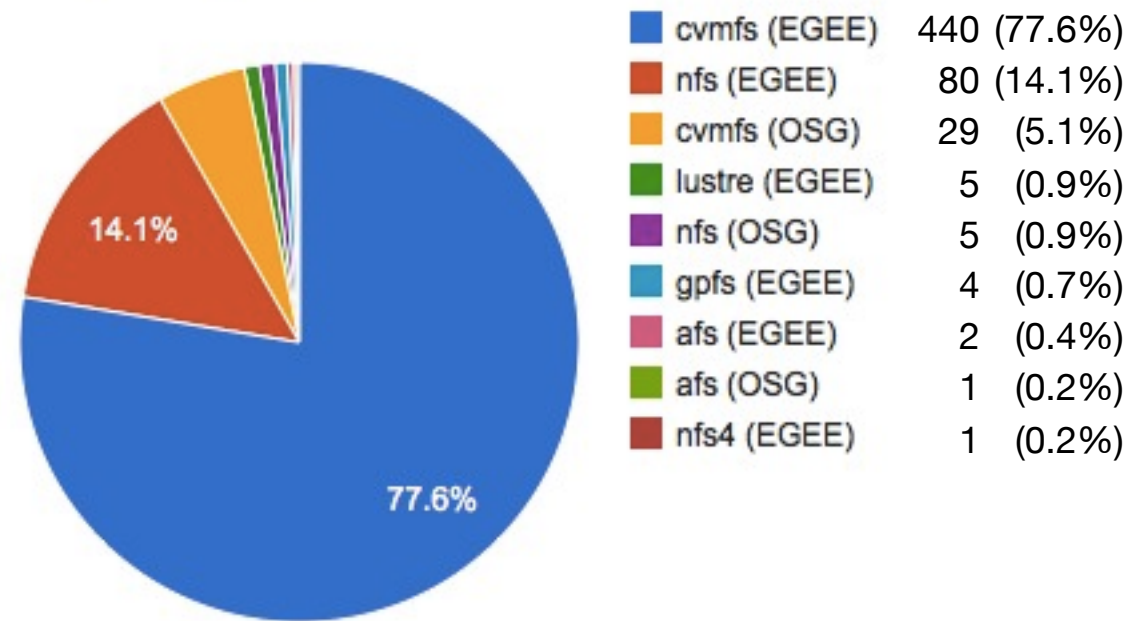
AMOD -- ADC Manager on Duty

- on-call for comp@p1 but also the 'manager' for the other ADC activities

ADCoS expert

- has a global view on ADCoS matters, responsible for some important actions, looks after ADCoS senior, but not 24h

cvmfs

FileSystem Types (active sites)

■ cvmfs (EGEE)	440	(77.6%)
■ nfs (EGEE)	80	(14.1%)
■ cvmfs (OSG)	29	(5.1%)
■ lustre (EGEE)	5	(0.9%)
■ nfs (OSG)	5	(0.9%)
■ gpfs (EGEE)	4	(0.7%)
■ afs (EGEE)	2	(0.4%)
■ afs (OSG)	1	(0.2%)
■ nfs4 (EGEE)	1	(0.2%)

(pie chart: 77.6%, 14.1%)

cvmfs is deployed at ~90 sites (out of ~130) ~ 70%

- 469 CEs (out of 567) ~ 83%
  - ‣ basically deployed at larger sites

cvmfs need to be stable and reliable

- instruction document for AMOD is essential

support level for non-cvmfs sites will decrease

hotdisk

DBReleases in cvmfs

- may go to Frontier/Squid (ADC weekly May 15 Dario)

Frozen Shower libraries in cvmfs (ADC weekly May 29 Misha)

- mc12 FS already in software releases
- mc11 FS will be added soon
- possible need of an old software release would still require hotdisk

Conditions DB in cvmfs (ADC weekly May 29 Misha)

pilot: can utilize DB files on cvmfs -- no need for hotdisk

production: DB files can be delivered as panda dispatch block (as input files)

- need T1 to host them

analysis: panda needs DBReleases in DDM for brokerage (either hotdisk or datadisk)

- Some sites (T3s) have already been working without hotdisk
  - running only AFT and analysis jobs without need for DBR (event generation)

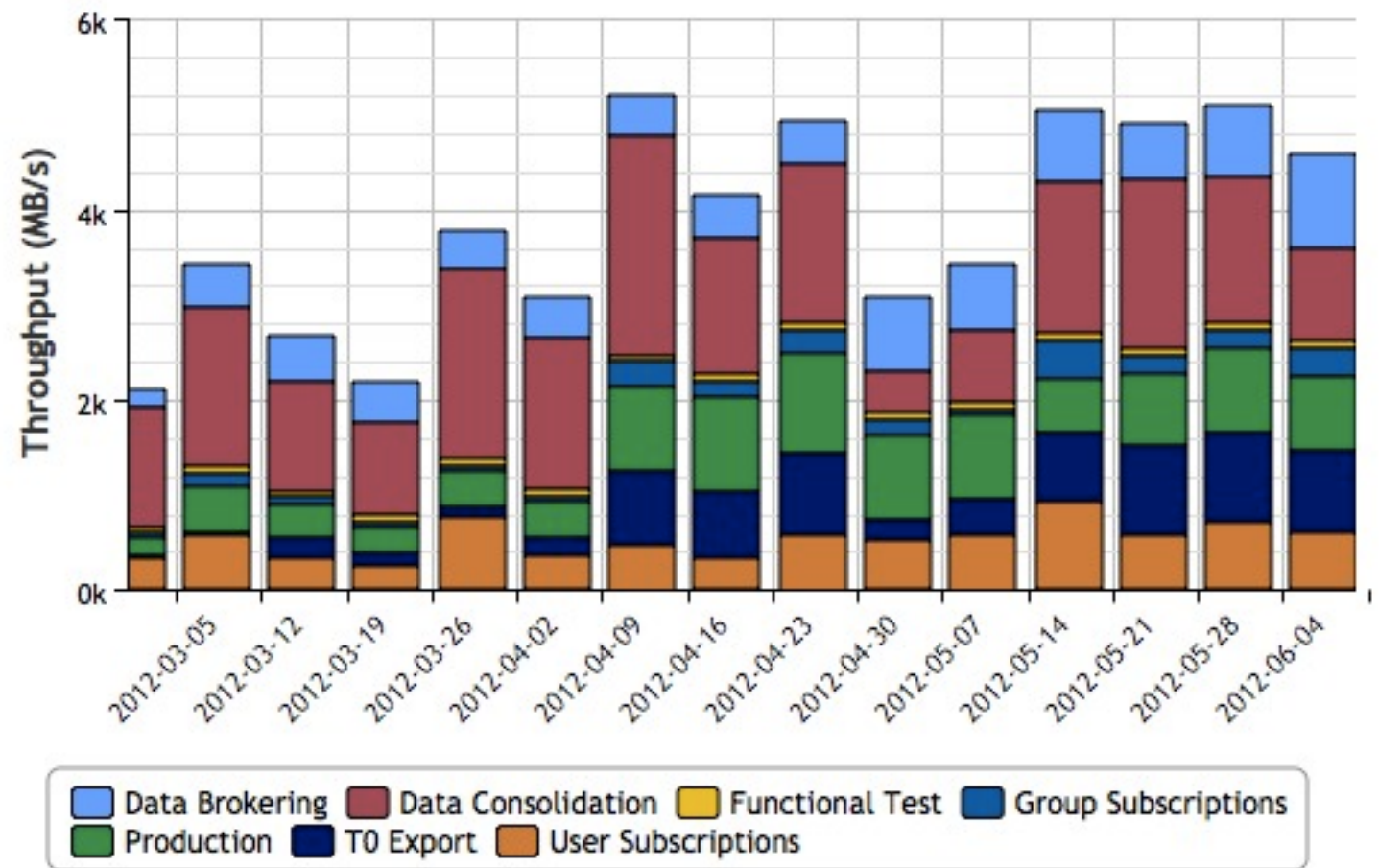work in progress (ADC Development June 4)

- sw-installation : puts cvmfs flag and sw releases into AGIS
- schedconfig : gets sw releases from AGIS instead of BDII

Distributed Data Management
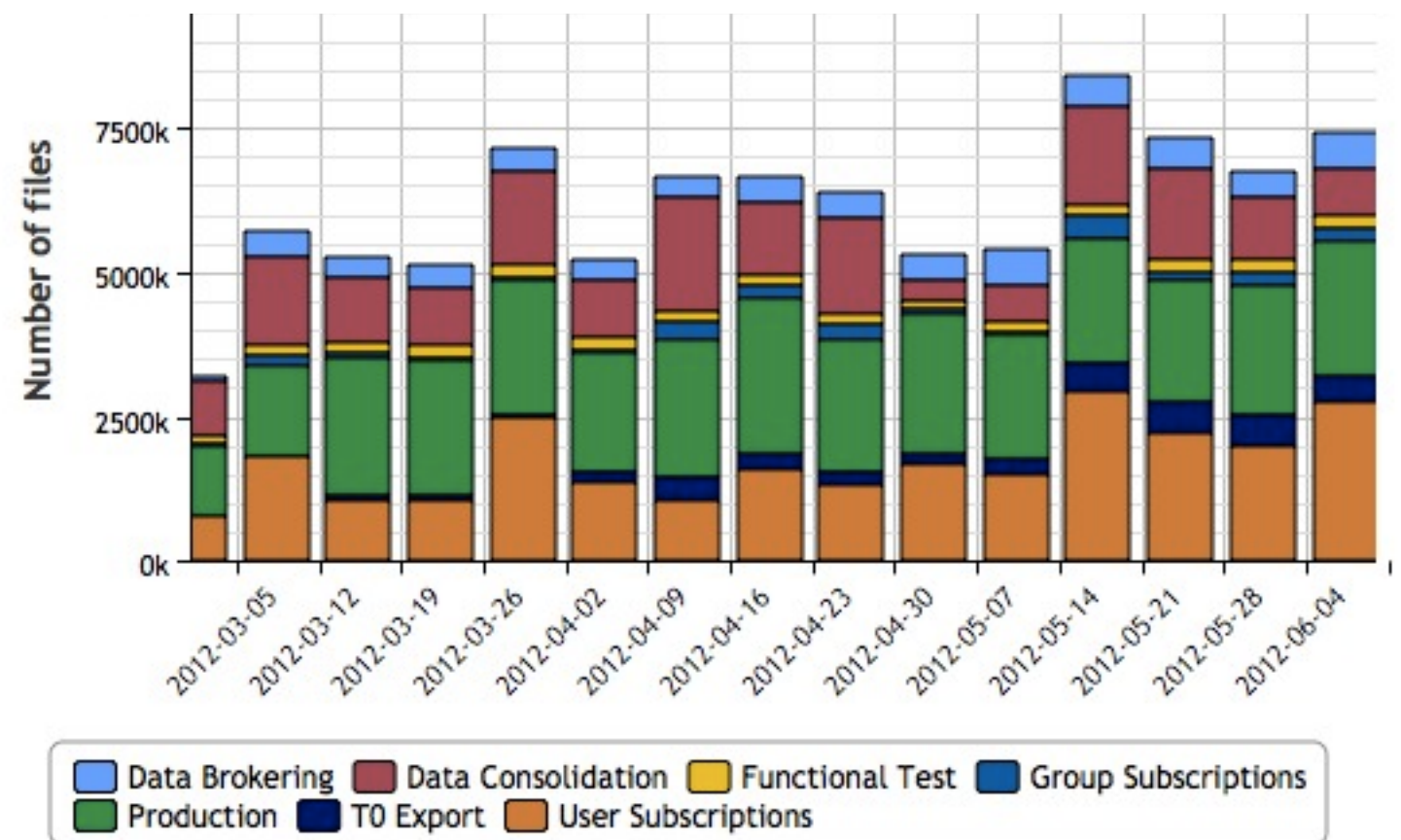
Many of the materials from Cedric

Throughput

- Majority
  - ‣ planned replicas
- The rest
  - ‣ T0, PD2P, GP, User comparable

Number of Files

- Majority
  - ‣ production
  - ‣ user
- == small files

DDM component: LFC

LFC consolidation at CERN :

- done: DE, ES, FR, IT, NL, TW, UK
- Since last SW&C week, one more cloud consolidated (FR cloud).
- Now 230319465 replicas hosted on CERN LFC.

Still to be migrated :

- NDGF (end of the month)
- Triumf (under discussion with CA squad since they have some concerns)
- BNL and US T2s.
  - For US T2s, they will be first consolidated at BNL.

Issues :

- In April/May we suffered from problem with LFC,
  - due to all threads exhausted on frontends.
- 2 new frontends added and now situation under control.
- Need to move more sites to registration on Panda server
  - stopped after seeing slow registration
  - most likely due to huge number of empty directories (HC)
  - cleaning done. should be able to proceed

DDM component: Site Services

News:

- Half of the Site Services have now been runing on VM

Issues:

- not related to VM, since also observed on physical machine
- Agent dying and restarting frequently
    ‣ it doesn't appear in SLS since the threshold (3 restart in last 30 minutes) is too low.
- Fixes applied on the fetcher agent improve the situation.

No news / issues for other DDM components

DDM: matters related to sites

GROUPDISK

- Group Space allocation for 2012 is nearly final (CREM2 May 31)
  - ‣ can be deployed where there is an urgent need
- Merge into DATADISK cannot proceed yet for a technical issue in monitoring

T2D

- get more data transfers
  - ‣ data distribution policy
  - ‣ DDM algorithm in choosing source sites
  - ‣ multi-cloud production
  - ‣ group space
- degradation of transfer throughput / stability causes problems
  - ‣ should be reviewed periodically
  - ‣ criteria being reviewed

DDM: transfer backlog

ADC weekly Jun 5 Stephane
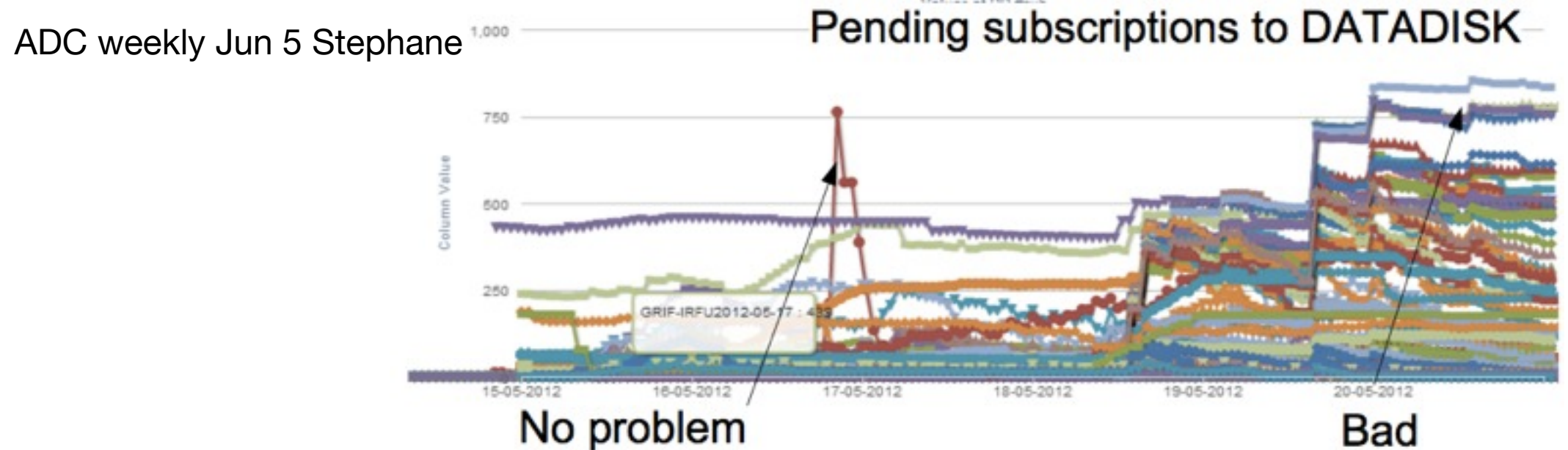
Cloud support teams should check regularly

Sites with huge backlog better be set to 100% for evgensimul

- avoid reco / group production -- shorter jobs with larger files I/O

Subscription agents (AK47, PD2P) should avoid new subscriptions to the sites with large backlog

- should take number of subscriptions per site into account
- Auto-exclusion?
  - ‣ to be worked on with AGIS and SSB

DDM: other issues

DDM do not remove empty directories

- hit the limit at INFN-T1 (bug #94881)

- number of subdirectories on Storm

  ‣ $2^{(16)} -1 = 65535$

data distribution (both predefined & dynamic)

- should take into account

  ‣ the capacity

  ‣ subscription backlog (previous slide)

quite many file losses in April/May (300k)

- incl. RAW (recovered from other replicas)

- 70k irrecoverable

  ‣ HITS, user analysis output, ...

DDM: operations load

Number of savannah tickets vs time still stable :

48 2011-06

60 2011-07

74 2011-08

40 2011-09

45 2011-10

46 2011-11

54 2011-12

46 2012-01

72 2012-02

37 2012-03

31 2012-04

53 2012-05

DDM: DQ2 clients deployment

DQ2 Clients 2.3.0 is ready

- the latest

- and necessary to move towards Rucio

- should not force upgrade before summer conferences

- announcement channels being prepared

Distributed Analysis

A dedicated talk on Tuesday morning

Monitoring

A dedicated talk on Tuesday morning

Network

Proposal to make the Network integral part of WLCG Resource Planning and Operations (LCG MB 5 June 2012)

- We should make the network with its operational and development components part of the global resource planning and operations process
- a Working Group that unites people with expertise in the areas of Applications (experiment computing models, work flows), Workload Management, Distributed Data Management and Networking

LHCONE for ATLAS

- The ATLAS approach is to evaluate LHCONE in a managed way and gain expertese in liaison with Network providers (ADC Weekly 8 November 2011 Michael)
- list of sites for initial evaluation https://twiki.cern.ch/twiki/bin/view/LHCONE/SiteList
  - ‣ with perfSONAR : AGLT2, DESY-HH, LAL, LRZ-LMU, Napoli, Prague, Tokyo, Toronto,, BNL, PIC, TRIUMF, KIT
  - ‣ perfSONAR test not configured (yet) : ASGC, CERN, SARA

Need some person power to monitor the test results

Last words

ATLAS Distributed Computing has been working at large scale

- thanks to great efforts from shifters and experts

- still a room for improvement

    ‣ especially to process urgent tasks smoothly

    ‣ operation workflow to be reviewed

- automation and monitoring are essential

    ‣ SSB as a 'simple' monitoring portal

        - should make it easy to detect problems

    ‣ AGIS as the information provider

        - should become the central place for exclusion information

- network is getting more and more important

    ‣ should check network and transfer monitoring regularly