ALICE in STEP'09 activities highlights
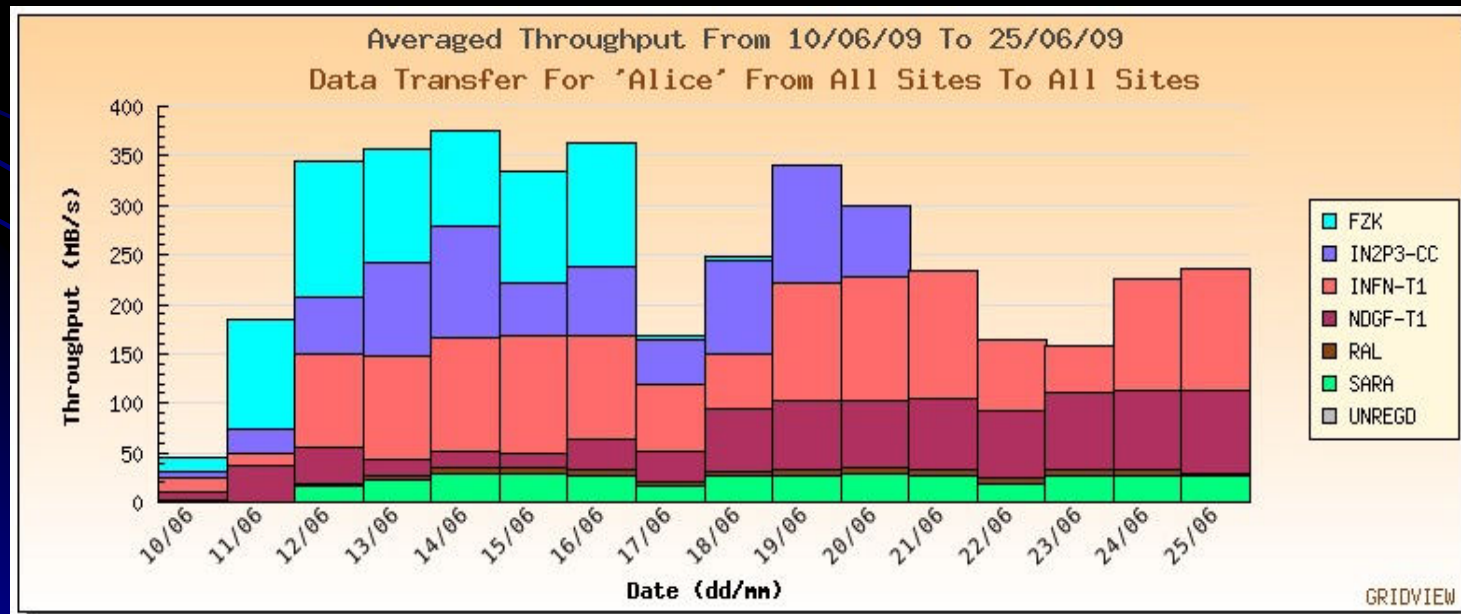
STEP-09 post-mortem
09 July 2009
Latchezar Betev

RAW Replication T0->T1

- Target rate 100MB/sec (p+p)
- Averages during STEP09
  - 350 MB/sec (6 sites, 1 week) – essentially the Pb+Pb rate during accellerator shutdown
  - 200MB/sec (4 sites, 2 weeks)

Averaged Throughput From 10/06/09 To 25/06/09
Data Transfer For 'Alice' From All Sites To All Sites

RAW Replication T0->T1 (2)

- Rates to sites as expected – proportional to ALICE mass storage fraction, all 6 ALICE T1s participating
  - And all at the same time
- Minor tuning of streams to rebalance the relative traffic to each site
  - For example a single stream to a T1 is not optimal
- No incidents – issues were minor and solved quickly by FTS/site experts
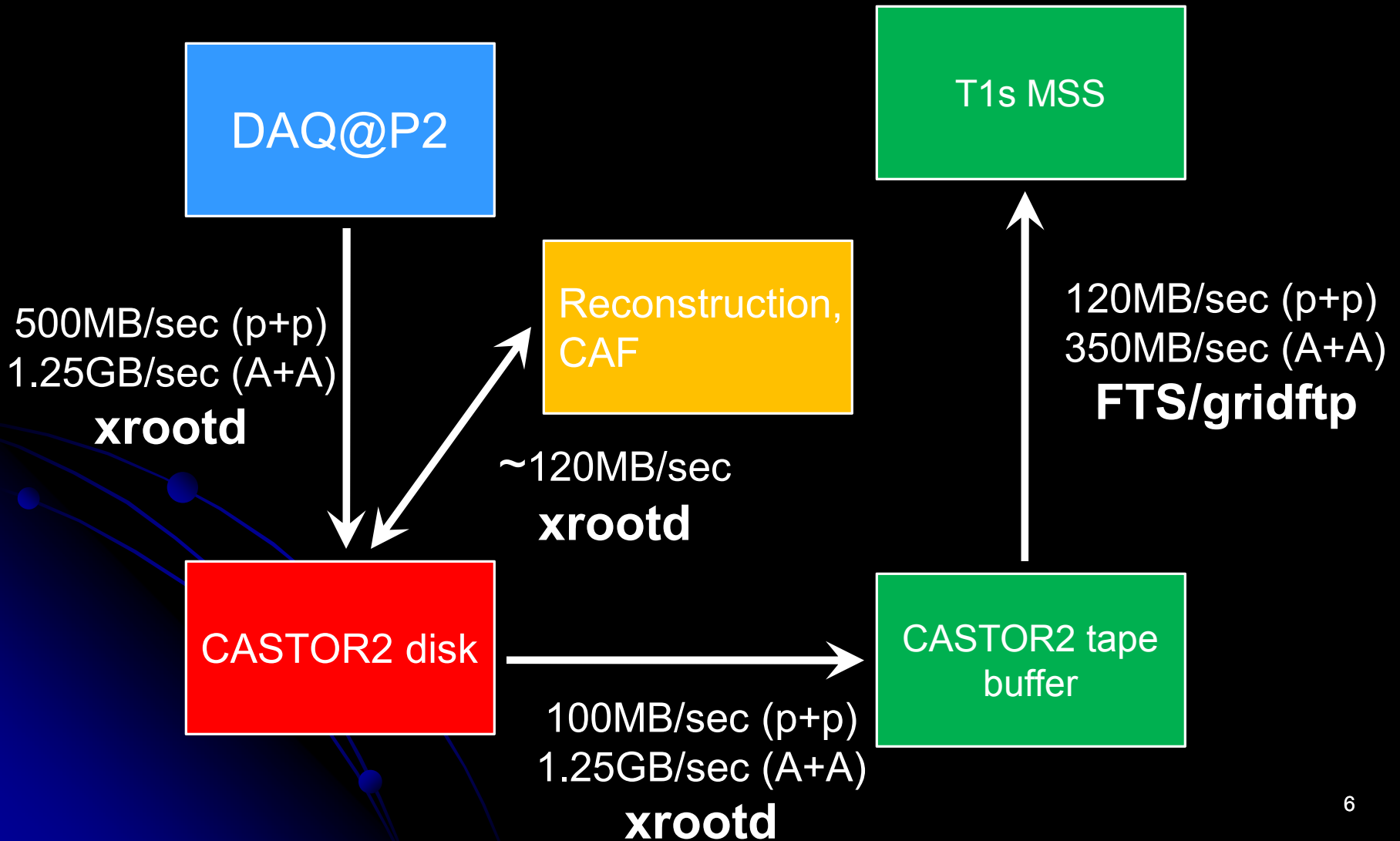- Status of the exercise -> SUCCESS

Reprocessing of RAW with tape recall

- Important for Pb+Pb – first pass reco after full month of data taking (most data only on tape)
- Requires finely tuned pre-staging mechanisms
    - Work on these ongoing
- Essentially postponed until new batch of Cosmics data becomes available (August-September)

Reprocessing of RAW with tape recall (2)

- The operational model is to keep RAW on disk as long as possible
  - This is already in effect for first pass reconstruction in p+p– quasi-online, few hours after data taking
  - The MSS buffers at T1s are as large as possible
- Naturally, tapes are used only as custodial storage for all ESD/AODs – the analysis is fully from disk, users *do not* see or use tape storages
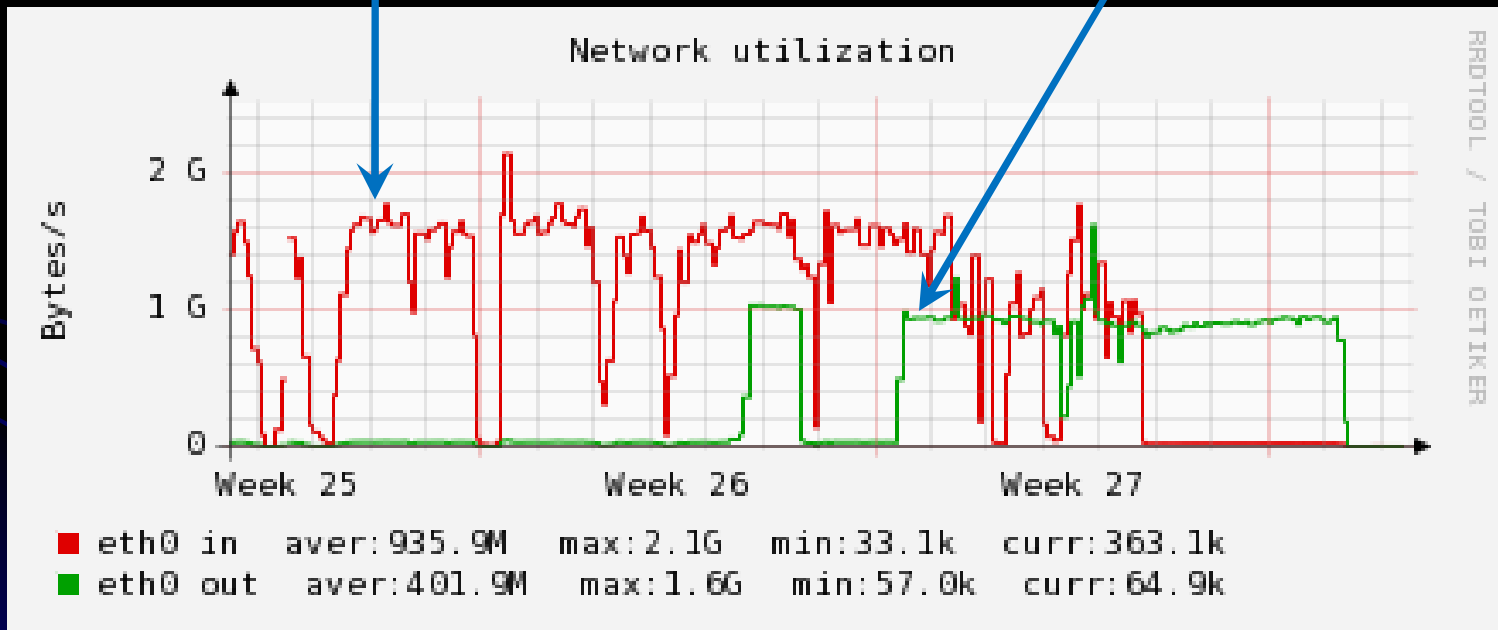- Status of the exercise -> PENDING

CASTOR 2.1.8 with xrootd

DAQ@P2

500MB/sec (p+p)
1.25GB/sec (A+A)
xrootd

Reconstruction, CAF

~120MB/sec
xrootd

T1s MSS

120MB/sec (p+p)
350MB/sec (A+A)
FTS/gridftp

CASTOR2 disk

CASTOR2 tape buffer

100MB/sec (p+p)
1.25GB/sec (A+A)
xrootd

6

DAQ@P2 -> CASTOR2 disk

DAQ stream, average 1.5GB/sec, 2 weeks
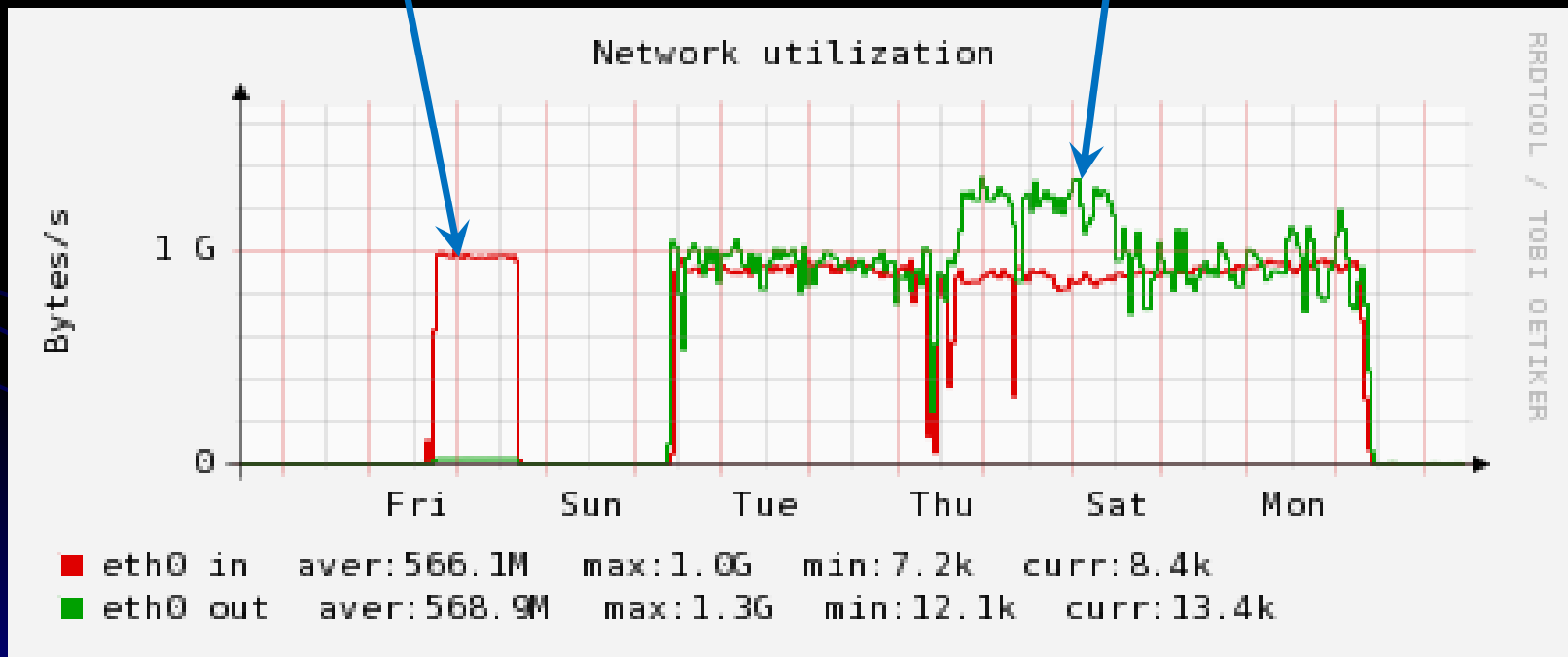
disk->tape buffer, limited 1 GB/sec, 1 week

CASTOR2 disk -> tape

disk->tape buffer, limited 1 GB/sec, 1 week

tape stream, limited 1 GB/sec, 1 week

Network utilization

| | eth0 in | aver:566.1M | max:1.0G | min:7.2k | curr:8.4k |
| | eth0 out | aver:568.9M | max:1.3G | min:12.1k | curr:13.4k |

CASTOR 2.1.8 with xrootd

- Fast development and deployment cycle
- Many thanks to IT/DM and IT/FIO
  - Andreas Joachim Peters
  - Dirk Duellmann
  - Fabrizio Furano
  - Ignacio Reguero
  - Miguel Marques Coelho Dos Santos
  - Olof Baring
  - Vlado Bahyl
  - *For helping move the excercise right along...*
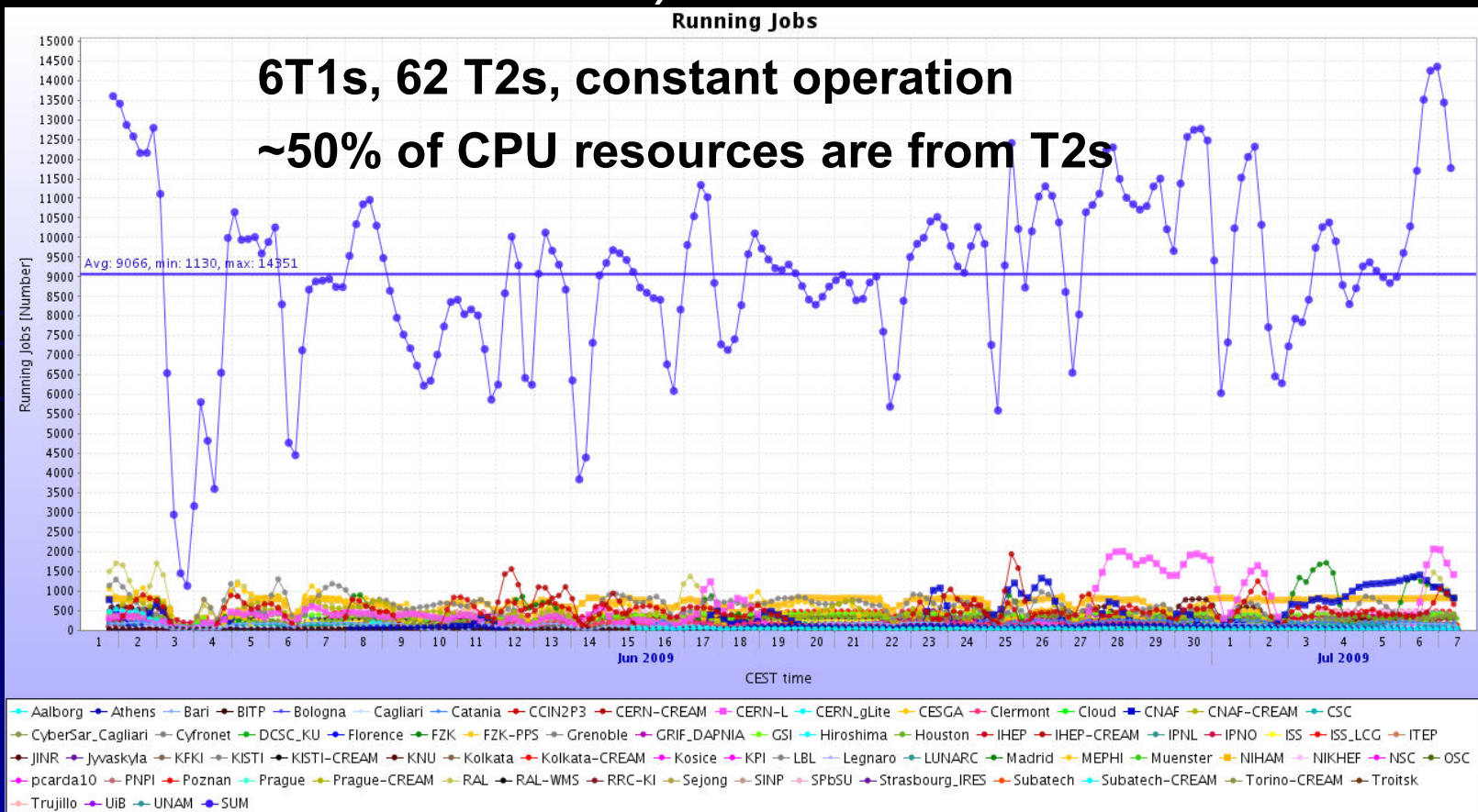
CASTOR 2.1.8 with xrootd

- Data taking rates achieved
  - Data writing/reading – only with *xrootd*
- Additional activities - ongoing
  - Reconstruction
  - Analysis
  - CAF
- CASTOR 2.1.8 in production
- Status of the exercise -> SUCCESS

Production activities

- Running before, during and after STEP'09
- Full scale MC production
  - Large scale (100Mio) p+p minimum bias completed during STEP09
  - Large volume Pb+Pb minimum bias and central
- Regular AliRoot releases, weekly tags with newest analysis code incorporated
  - Build/running on SL4, SL5, Ubuntu (becoming very popular), Itanium (ICC very helpful in weeding out errors), MAC OS

Production activities (2)
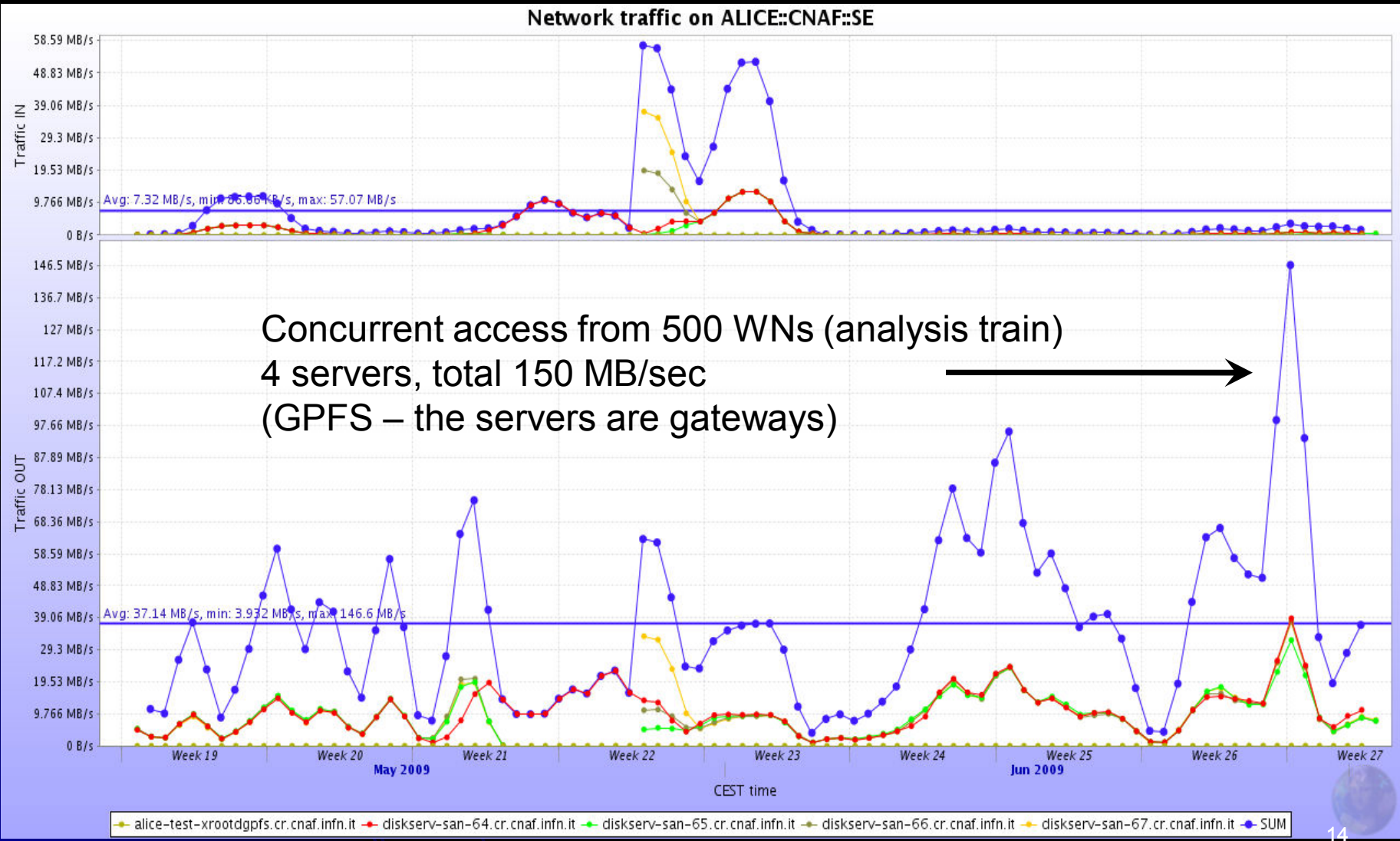
- Stable job profile, WMS under control
- CREAM CE is gaining popularity (still not installed at all sites)

Running Jobs

6T1s, 62 T2s, constant operation

~50% of CPU resources are from T2s
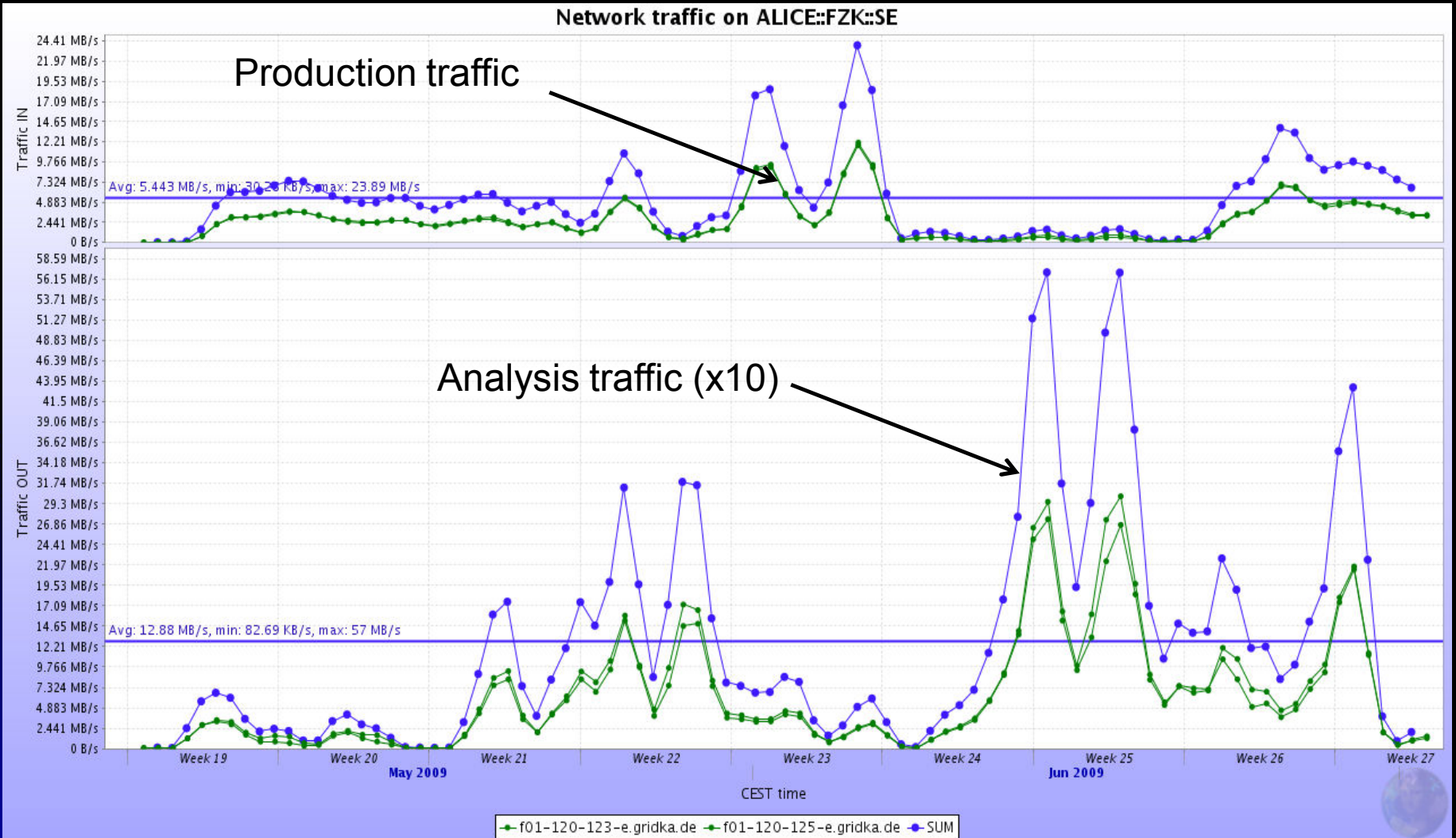
Avg: 9066, min: 1130, max: 14351

Grid analysis activities

- Organized ESD->AOD production (analysis train) running on schedule
  - ALICE strategy presentation by A.Morsch in the analysis session on Friday
- PWG code in AliRoot, new train wagons added as they become available
- ~70 users steady presence on the Grid
  - All end-user analysis is on Grid and PROOF
- ~ 6% of ALICE Grid resources – end-user analysis
  - Does not include the train operation – this is centrally managed activity (+10% resources)

Data access - analysis

Network traffic on ALICE::CNAF::SE

Concurrent access from 500 WNs (analysis train)
4 servers, total 150 MB/sec
(GPFS – the servers are gateways)

Data access - analysis

Network traffic on ALICE::FZK::SE

Production traffic

Analysis traffic (x10)

Avg: 5.443 MB/s, min: 30.23 KB/s, max: 23.89 MB/s

Avg: 12.88 MB/s, min: 82.69 KB/s, max: 57 MB/s

Traffic IN

Traffic OUT

May 2009

Jun 2009

CEST time

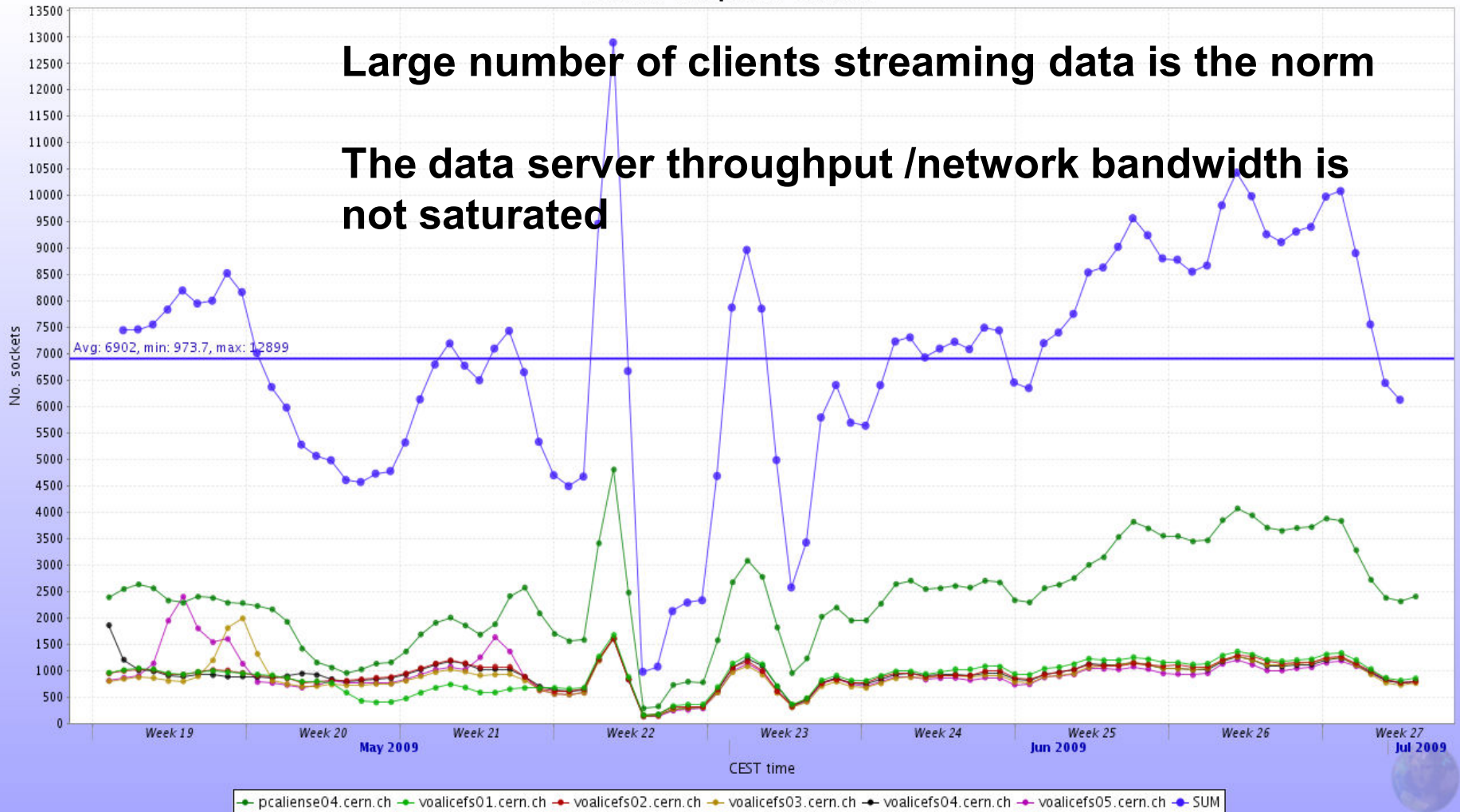f01-120-123-e.gridka.de — f01-120-125-e.gridka.de — SUM

Client connections - analysis

Number of opened sockets

Large number of clients streaming data is the norm

The data server throughput /network bandwidth is not saturated

Avg: 6902, min: 973.7, max: 12899

No. sockets

13500
13000
12500
12000
11500
11000
10500
10000
9500
9000
8500
8000
7500
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0

Week 19   Week 20   Week 21   Week 22   Week 23   Week 24   Week 25   Week 26   Week 27

May 2009   **Jun 2009**   **Jul 2009**

CEST time

pcaliense04.cern.ch   voalicefs01.cern.ch   voalicefs02.cern.ch   voalicefs03.cern.ch   voalicefs04.cern.ch   voalicefs05.cern.ch   SUM
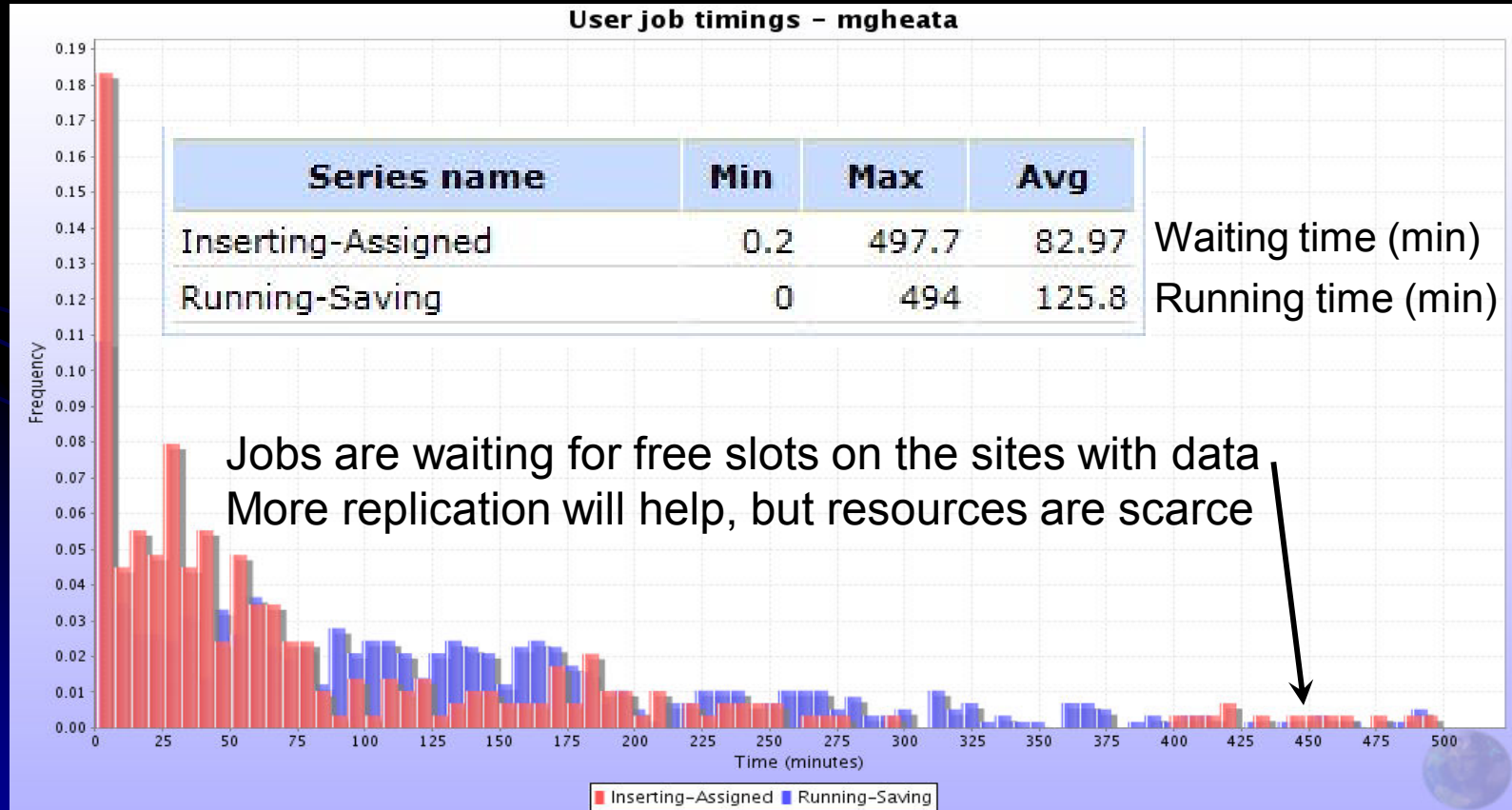
Grid analysis activities (2)

- Single task queue – internal job prioritization is working well
- Biggest concern – availability and stability of disk storage
  - DPM and dCache with updated xrootd

Running jobs per user
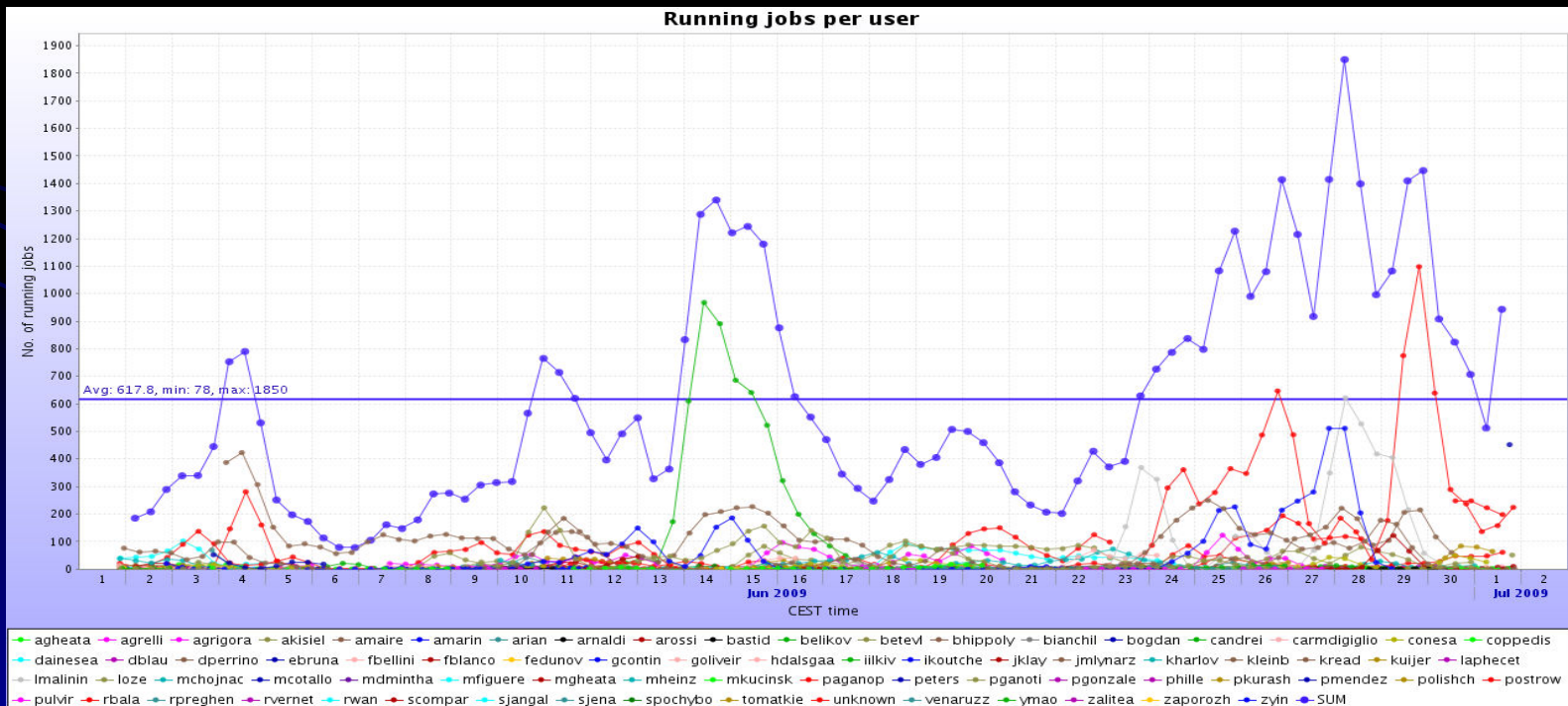
Grid responsiveness

- Minimizing waiting time for users
- Strongly depends on the available data replicas and storage elements

User job timings – mgheata

Series name	Min	Max	Avg	
Inserting-Assigned	0.2	497.7	82.97	Waiting time (min)
Running-Saving	0	494	125.8	Running time (min)

Jobs are waiting for free slots on the sites with data
More replication will help, but resources are scarce

18

Prompt analysis activities

- Detector calibration, interactive analysis
- PROOF-enabled analysis facilities
  - CAF - CERN, GSIAF – GSI Darmstadt, in preparation at CCIN2P3
  - 19 Physics/Detector groups, ~130 unique users
  - Data (ESD/AOD/calibration) from Grid productions and RAW
  - Disk and CPU quotas implemented
  - Direct access to Grid SEs implemented

Prompt analysis activities (2)

- Extremely popular tool, good integration with the Grid at the level of data access
- CAF capacity will be doubled before start of data taking
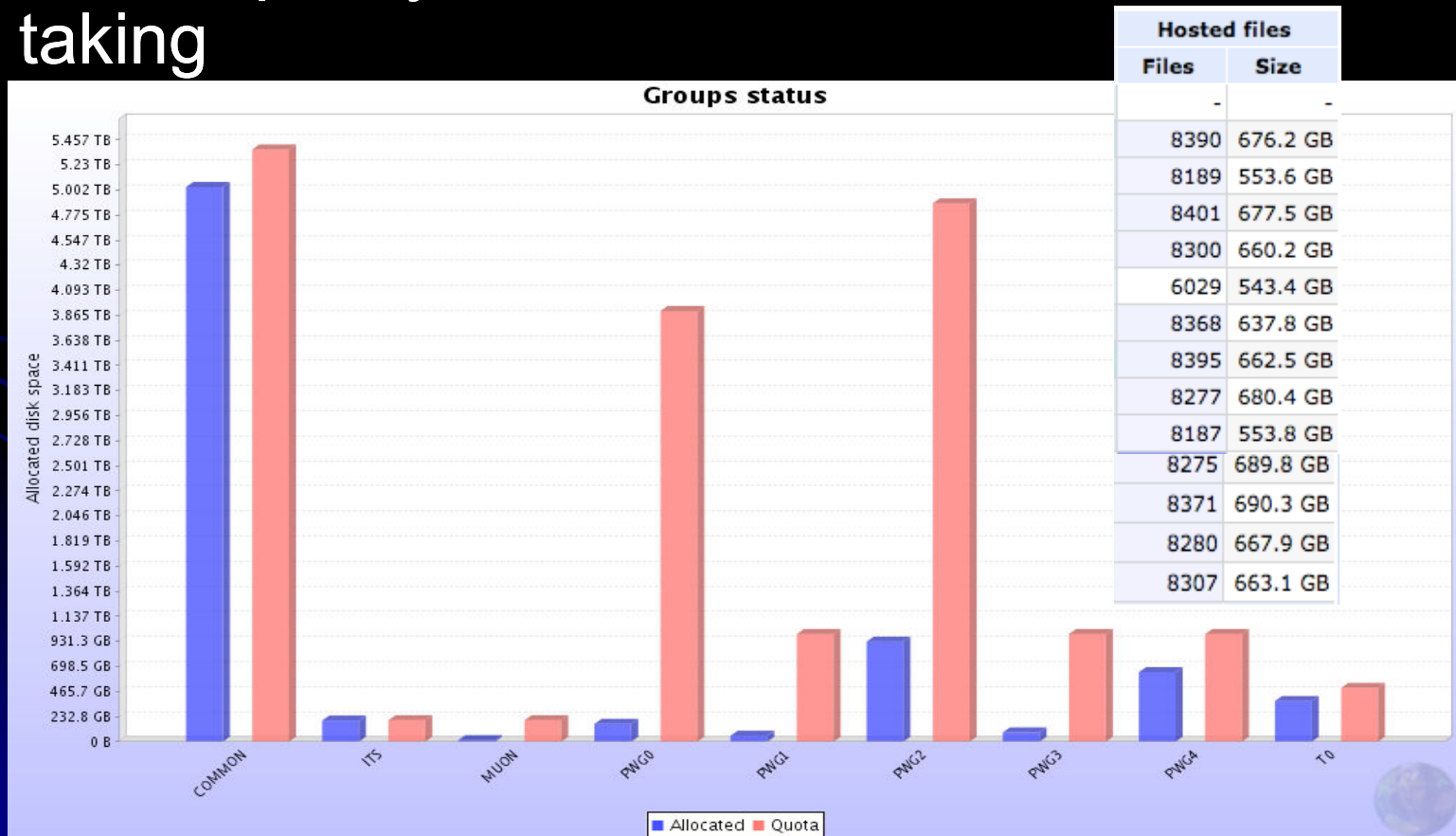
Summary

- All STEP'09 activities successfully completed
  - Re-processing from tape – August/September
- Big success – migration to CASTOR 2.1.8 and validation of RAW data registration at T0
- Grid and interactive user analysis is routine and gaining momentum
- Analysis train strategy is efficient and well adapted to the Grid structure
- Standard activities (MC production) ongoing
- Cosmics data taking will start in August
  - Crucial for further alignment and calibration of ALICE before p+p

Thank you!

- To the experts at the computing centres supporting the ALICE operations
- To the ALICE regional representatives (very few people), who manage the Grid resources operations in entire countries
- CERN IT/GS