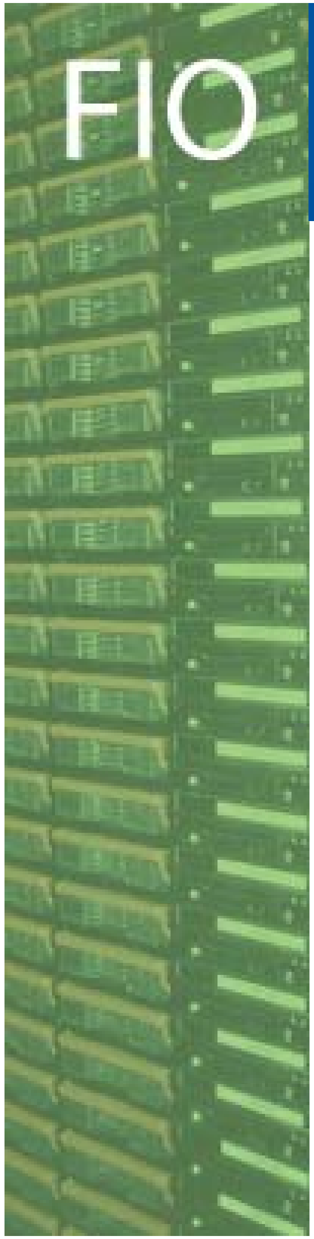
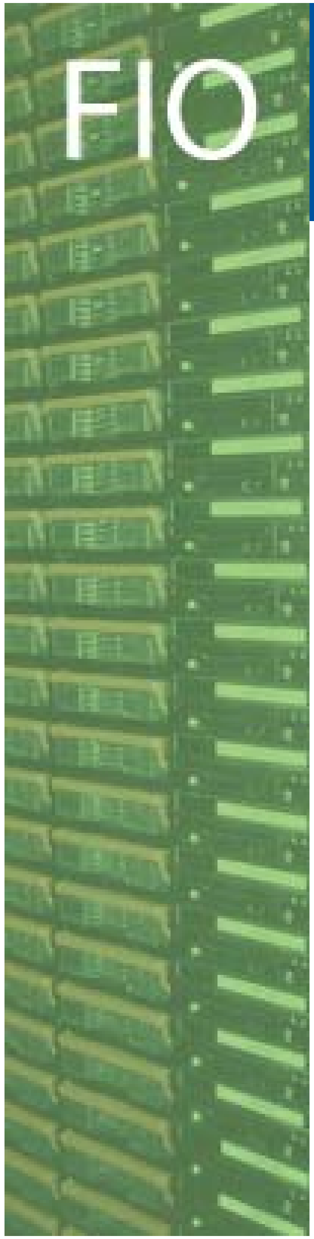


Tier-0
CCRC’08 May
Post-Mortem

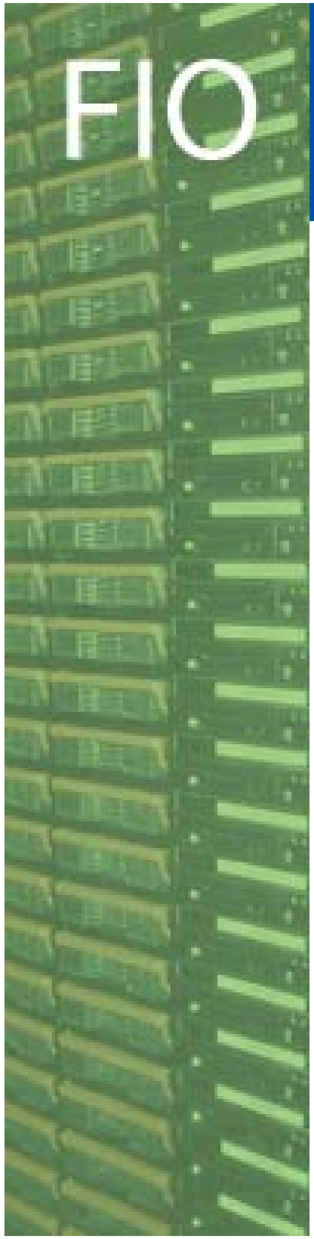
Miguel Santos
Ricardo Silva

IT-FIO-FS



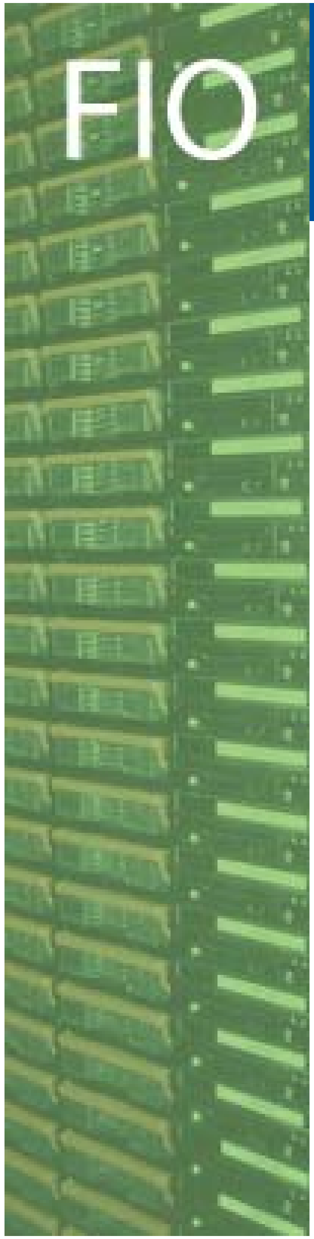
Outlining

- Overall Summary
- Disk Cache
  - CASTOR backend
  - SRM
- Tape
- Castor monitoring
- Batch
- LFC
- WMS



Overall Summary

- Services ran generally smoothly
- The exercise combined activity was handled well
  - Was the activity representative of production load?
- Inter-Experiment interference from the combined exercise was limited
  - Some interference at the SRM layer
  - Reduced interference at the tape layer
- Good experience, some lessons learnt



Disk Cache Overview

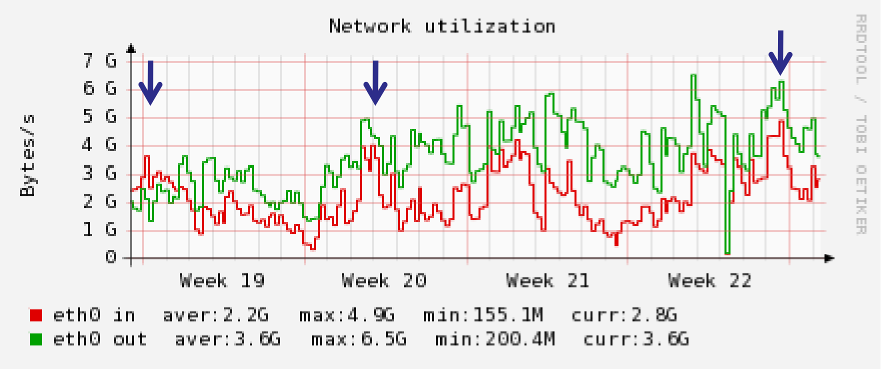
- Normal activity
- In general expecting more load on the system
- Tier-0 storage works well ☺
- A few issues on the CASTOR backend, specially garbage collection on CASTORCMS ☹ ☺
- Alert mailing list used for the first time ☺
- SRM2 ‘teething’ problems ☹
- Monitoring improvements ☺

Glad to have caught the GC problems now rather later!

Good to have exercised and debugged the alert procedure now!

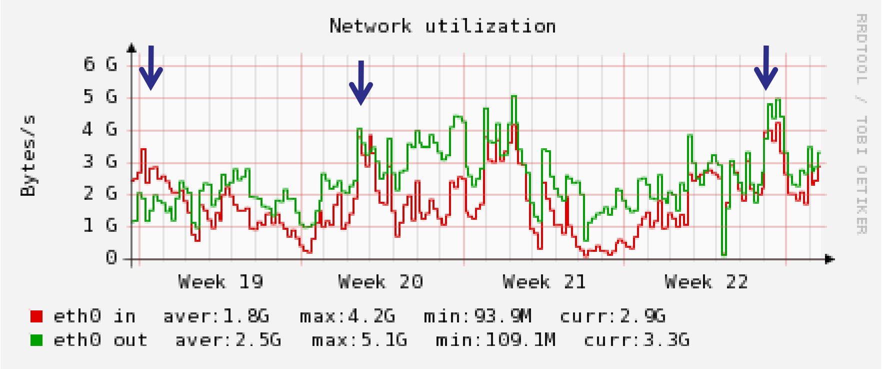
System performance

CASTOR disk cache throughput:



Three CMS GC problems contributed to large peaks of internal traffic ☹

CASTORCMS disk cache throughput:



On the bright side, it is nice to know the capacity is there! ☺

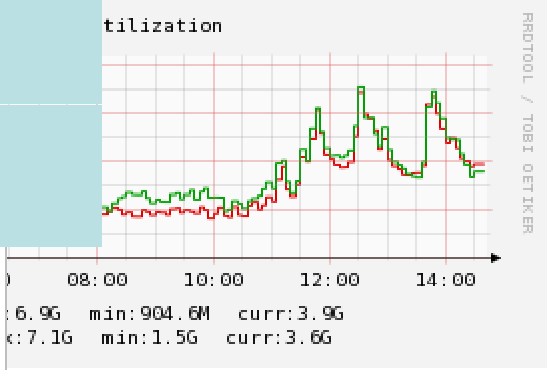
Peaks of 9GBytes/s (IN+OUT)

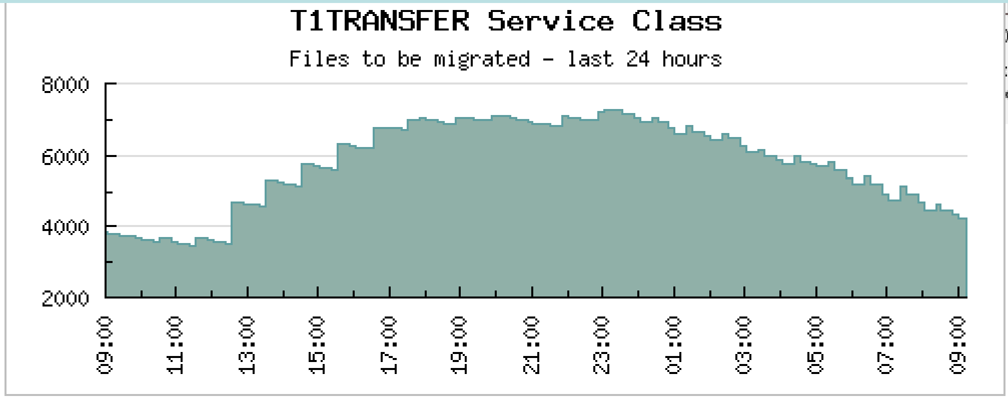
CASTOR incidents

In general the incidents were detected and resolved within short periods of time

Compartmentalized impact
Reduced interference:
- within the instance
- with other instances

g weight +base line





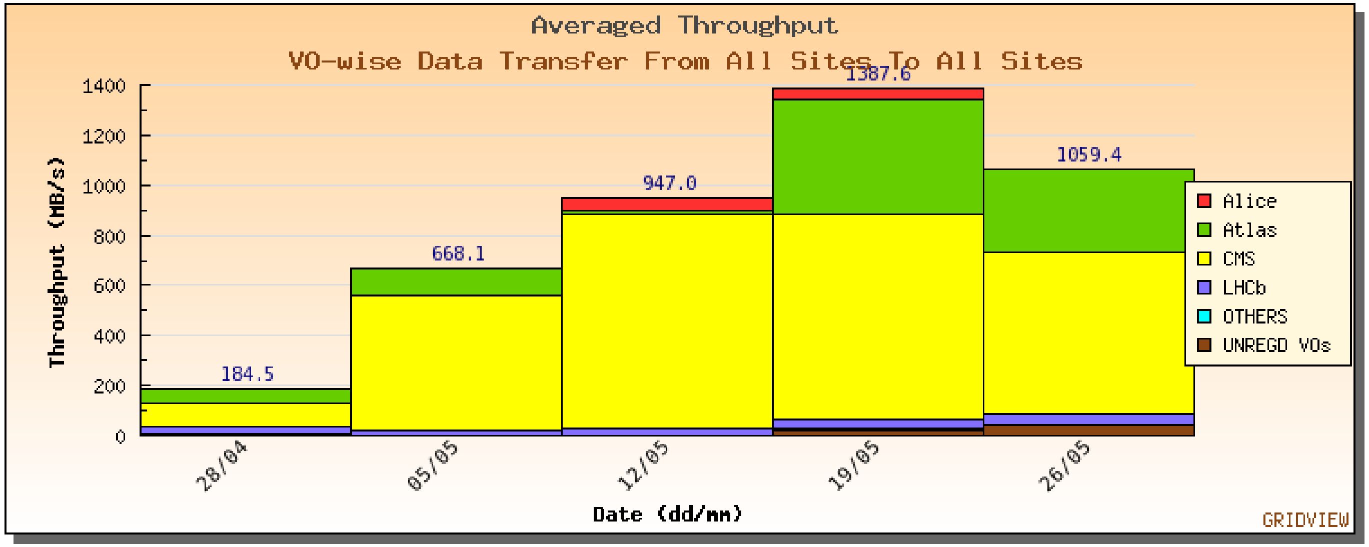
- 20/05: CMS DB slowdown
- 20/05: LHCBFAILOVER has large migration queue ←
- 24/05: PUBLIC slowdown
- 29/05: CMS GC (SRM default weight)←

Alert mailing lists

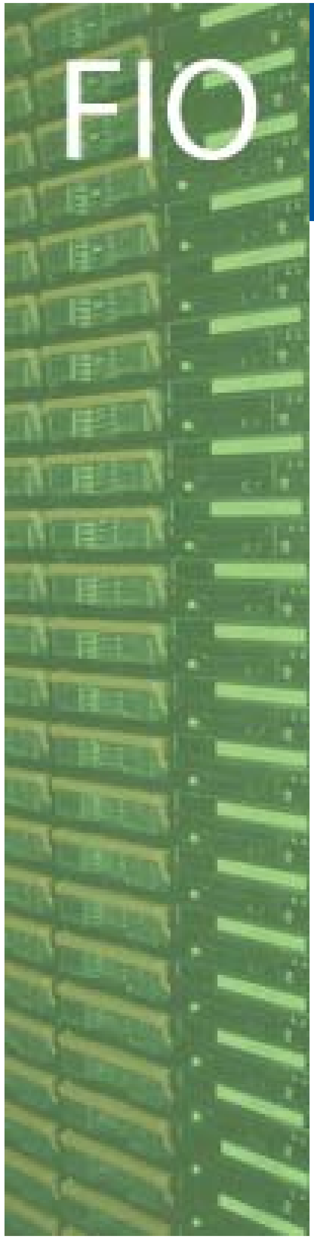
- May 24th PUBLIC slowdown was detected and fixed at 9 a.m. The problem had propagated to the shared SRM and from there to the other SRMs.
- The SRM outage was noticed and reported by ATLAS and a mail was sent to the atlas-operator-alarm list.
- The 24/7 operator had noticed a high load on the SRM db service due to a high number of oracle sessions. The operator rebooted the DB server at ~11:30 after which the SRM endpoints slowly recovered. In parallel the data service standby service manager reported the problem to the SRM service manager.
- The operator procedure has been improved with clearer instructions on how to check the information, verify the service is covered by piquet and how to contact the piquet.

SRM

- When it works it works well
- A large volume of data was transferred
- The average rate was high



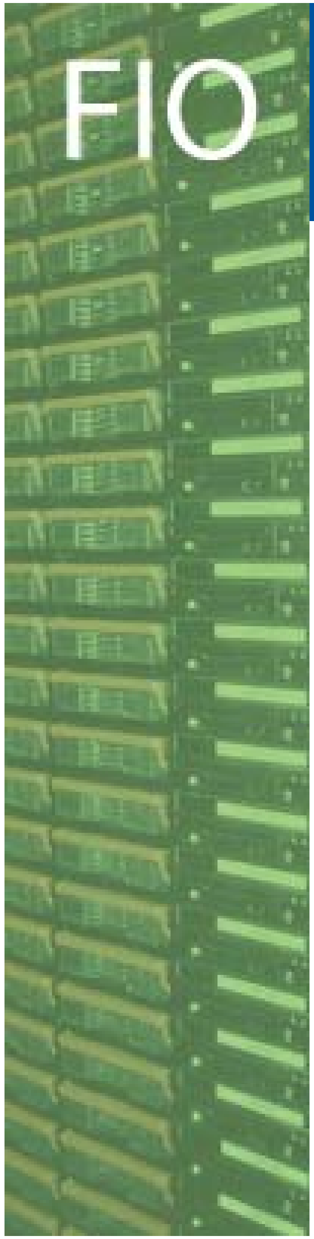
- Reliability is still an issue
- ~10 incidents with impact ranging from service degradation to complete unavailability



SRM Incidents & Conclusions

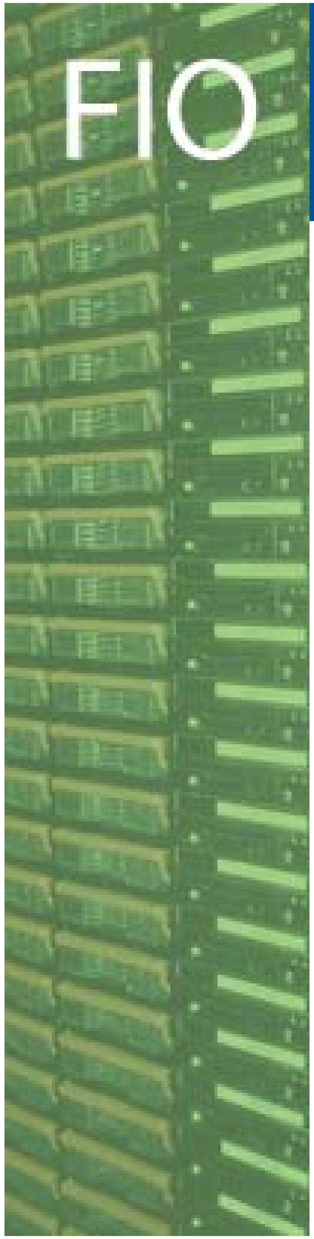
- May 5 – redundant SRM back-ends lock each other in database [ALL VOs]
- May 13th – lack of space on SRM DB [LHCb]
- May 13th – DB “extreme locking” / DB deadlocks [ALL VOs]
- May 9th, May 14th, May 19th – SRM ‘stuck’ / no threads to handle requests [ATLAS]
- May 21st, May 24th – slow stager backend causes SRM stuck / DB overload [All VOs]
- May 30th – get Timeouts due to slowness on Castor backend [ATLAS, LHCb]
- 3 times in May – problematic use of soft pinning caused GC problems [CMS]
- June 6th – patch update crashed backend servers [ATLAS, ALICE, CMS]

- To be improved:
  - Better resiliency to problems
  - More service decoupling
  - Some bugs need to be fixed
  - Better testing needs to be done



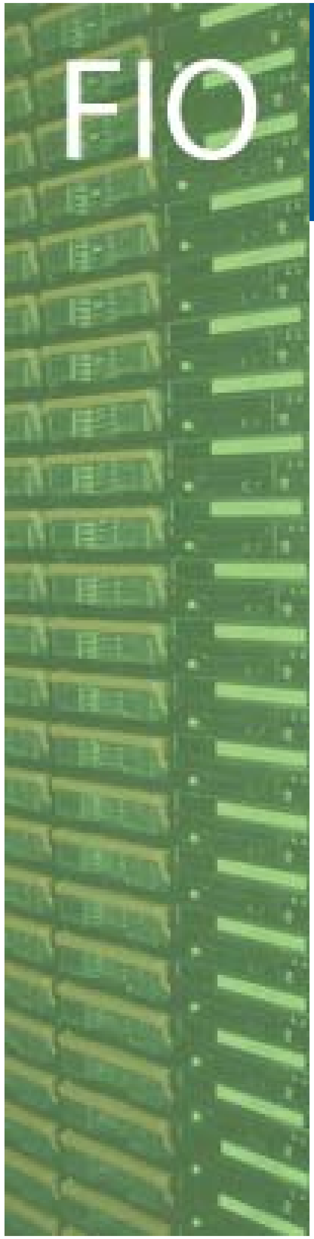
SRM2 Outlook

- Separate out LHC VOs from shared instance
- Migrate all SRM databases to Oracle RAC (done for ATLAS)
- Upgrade to SRM 2.7 and deploy on SLC4
  - Redundant backends
  - Uses CASTOR 2.1.7 API which allows deployment of redundant stager daemons
  - Deploy fixes for identified bugs
- Configure SRM DLF to send logs to appropriate stager DLF
  - Improve our debugging response time
- Continue improving service monitoring



Tape Overview

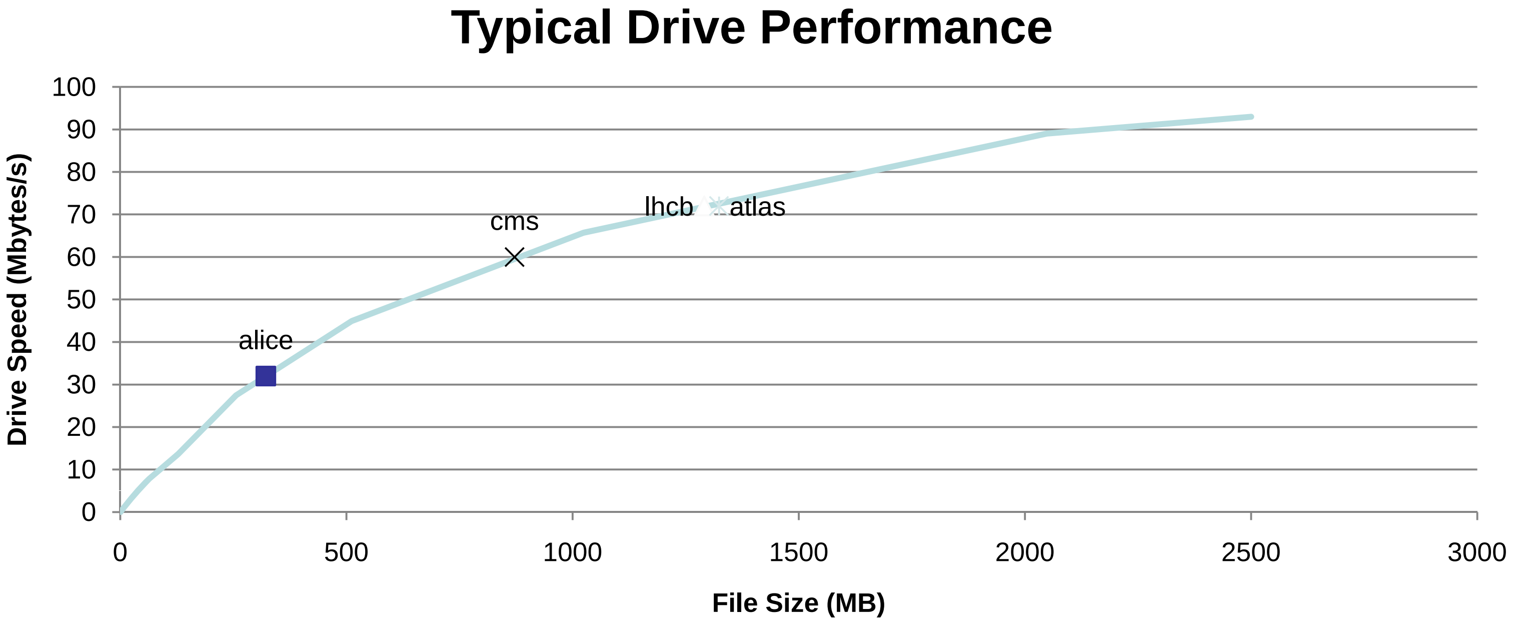
- Writing efficiency continues to improve
  - File sizes, especially Atlas
  - Write policies working well
  - Tiny files still an issue, follow-up will continue
- High read activity continues.
  - Data volume per mount low
  - Non-production access to Tier-0 continues
- Tape service passed CCRC May without issues.
- More load was expected.



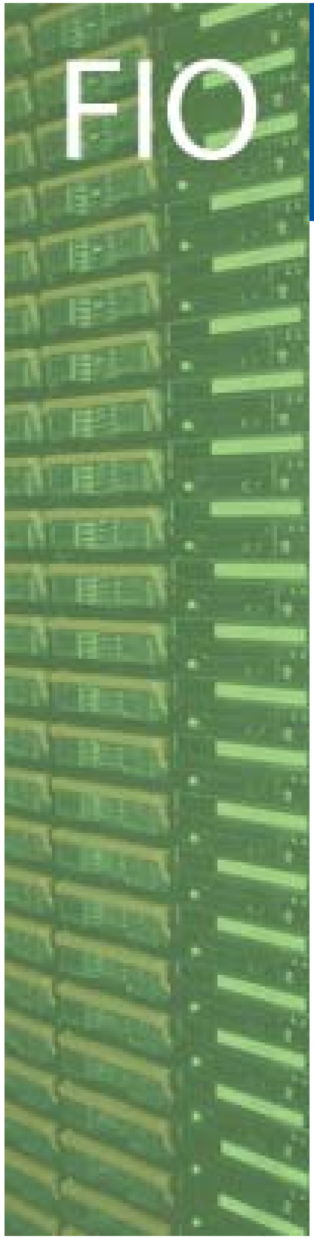
Operational Issues

- Maintenance took place but without significant disruption
  - IBM robot failure lasting 3 hours
  - Sun robot arm failure
  - Firmware upgrade of both IBM and Sun robots
  - Rolling upgrades for Castor tape server code
- Outlook
  - Tape queue prioritisation of production users
  - Tuning bulk transfers
  - Improve automation of metrics

File size and performance



Date	Alice	Atlas	CMS	LHCb
CCRC May '08	322 MB	1291 MB	872 MB	1327 MB
March '08	143 MB	230 MB	1490 MB	865 MB
CCRC Feb '08	340 MB	320 MB	1470 MB	550 MB
Jan '08	200 MB	250 MB	2000 MB	200 MB

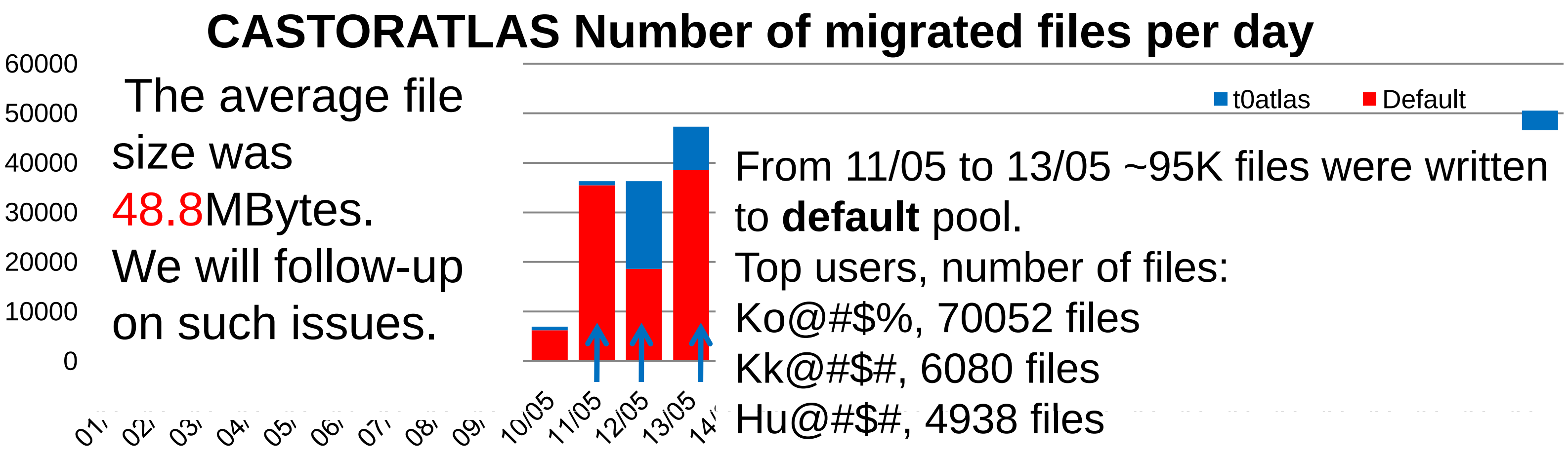


Castor monitoring

- Monitoring improvements continue
- New metrics being proposed, implemented and deployed in a close collaboration between IT-DM and IT-FIO
- Deeper understanding of various activities
- Some examples follow...

Files migrated per day

(CMS & ATLAS)



The average file size was 48.8MBytes. We will follow-up on such issues.

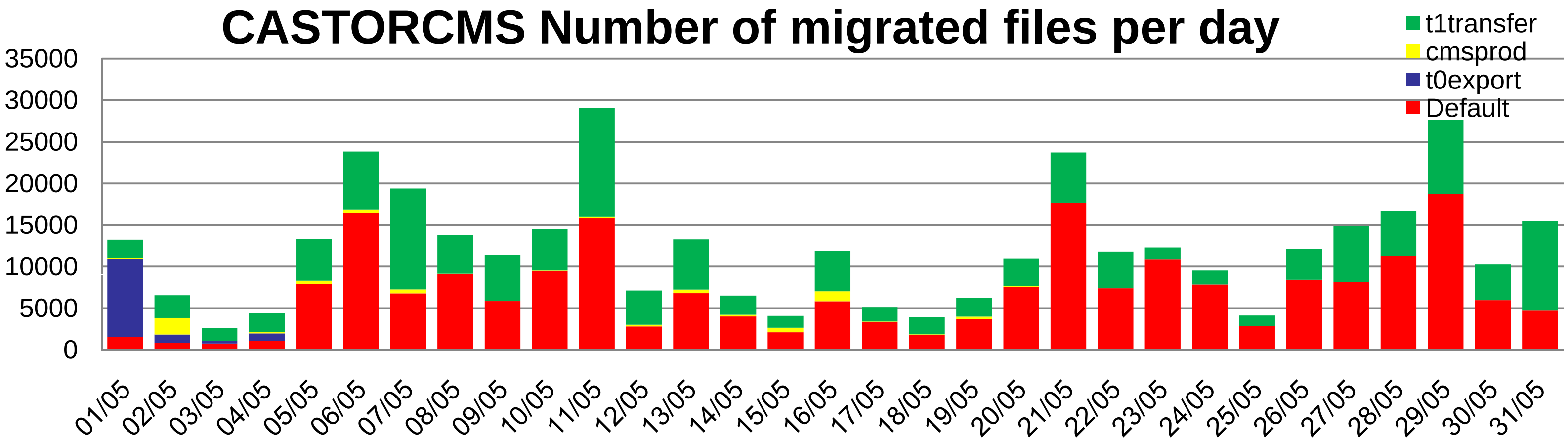
From 11/05 to 13/05 ~95K files were written to **default** pool.

Top users, number of files:

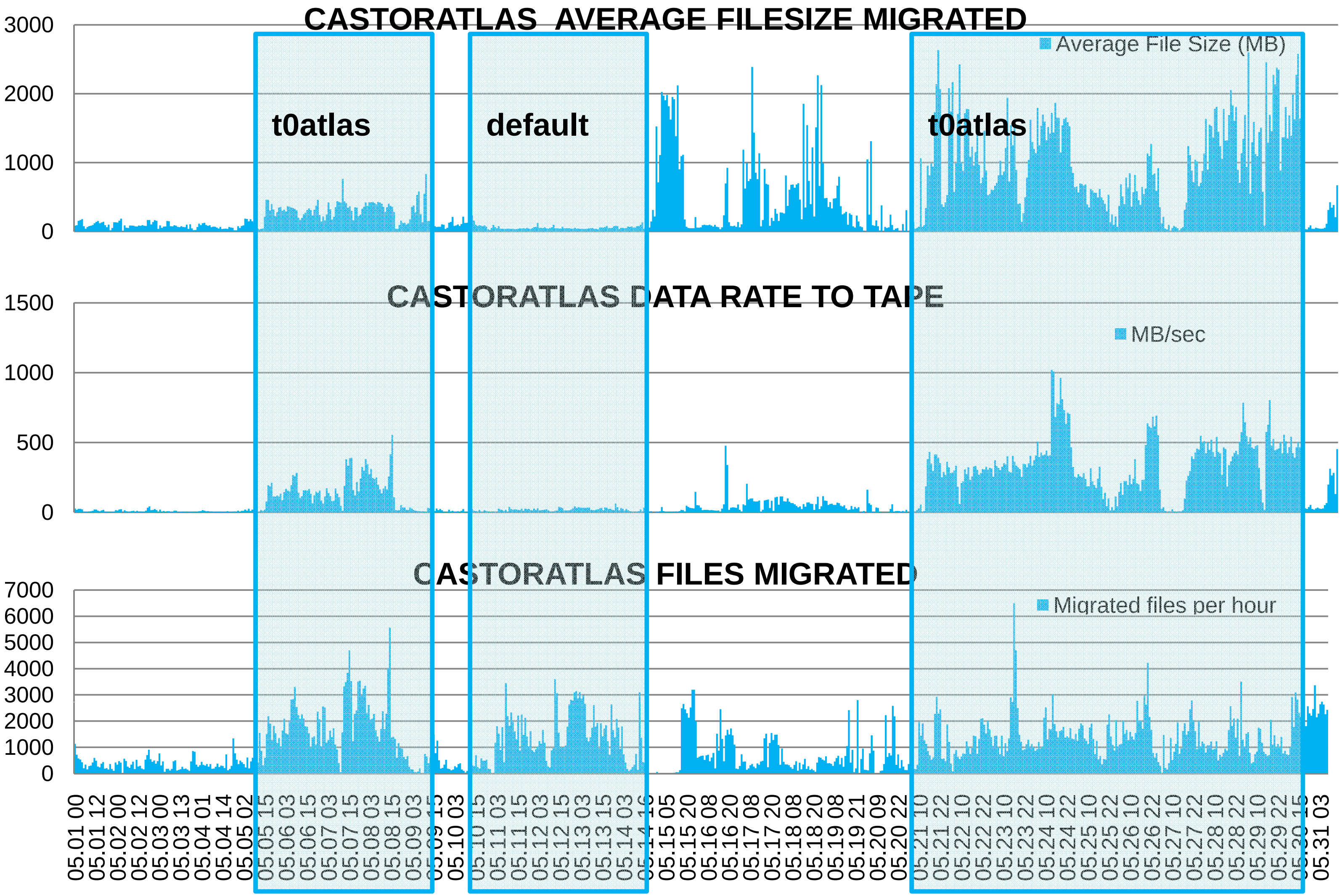
Ko@#$%, 70052 files

Kk@#$#, 6080 files

Hu@#$#, 4938 files



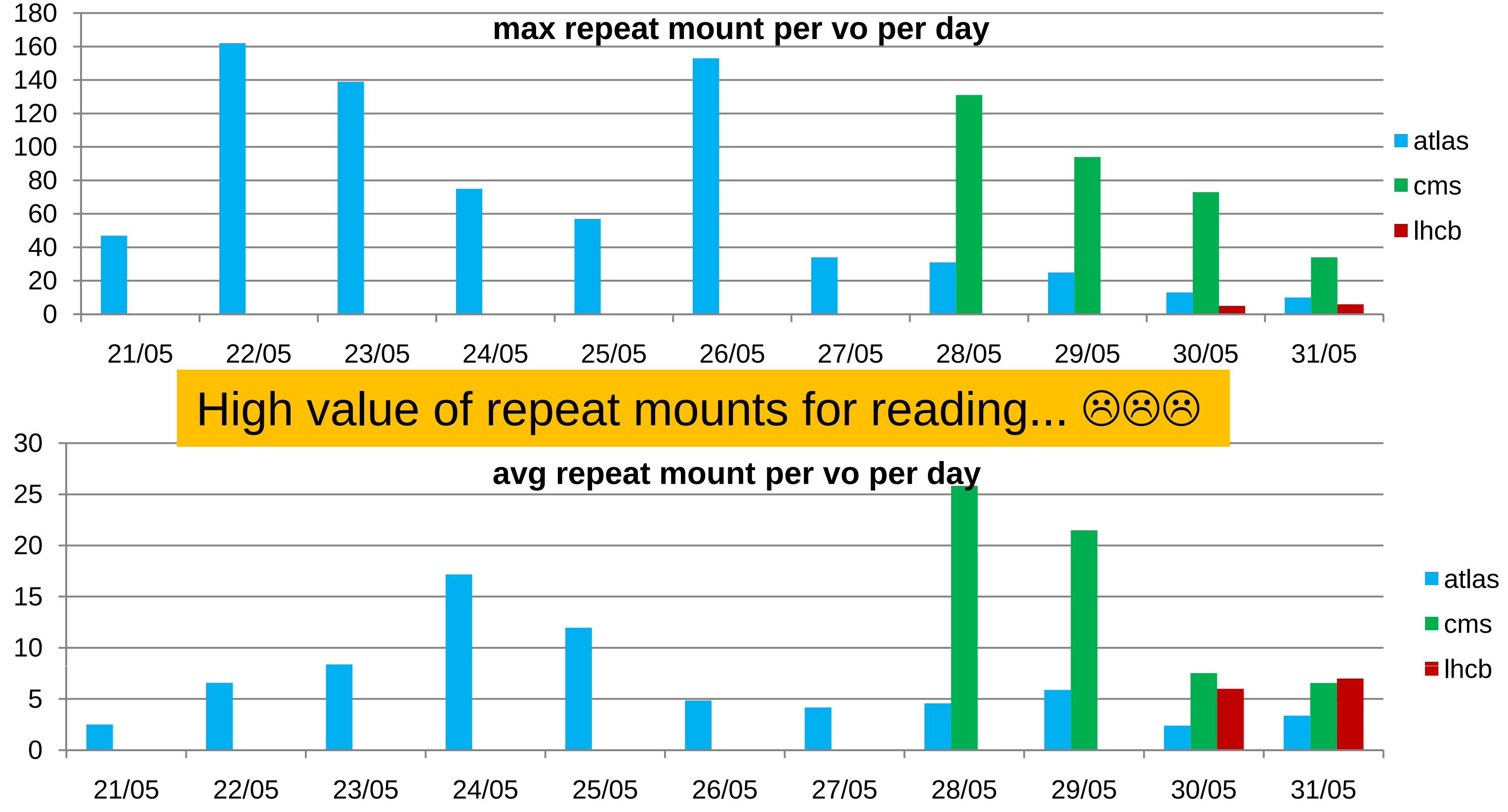
ATLAS migrations



T0ATLAS is working well. Already mentioned issue on default pool...

Repeat tape mounts

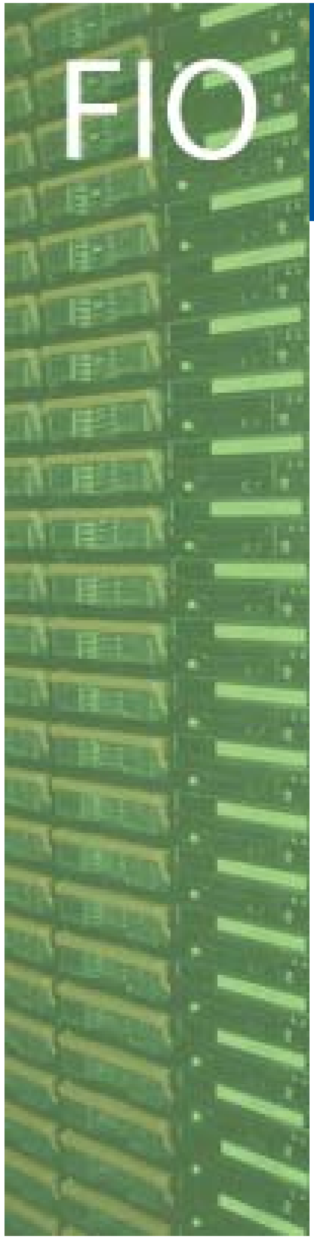
- Repeated (read) tape mounts per VO



High value of repeat mounts for reading... ☹☹☹

Power cut, 30th May

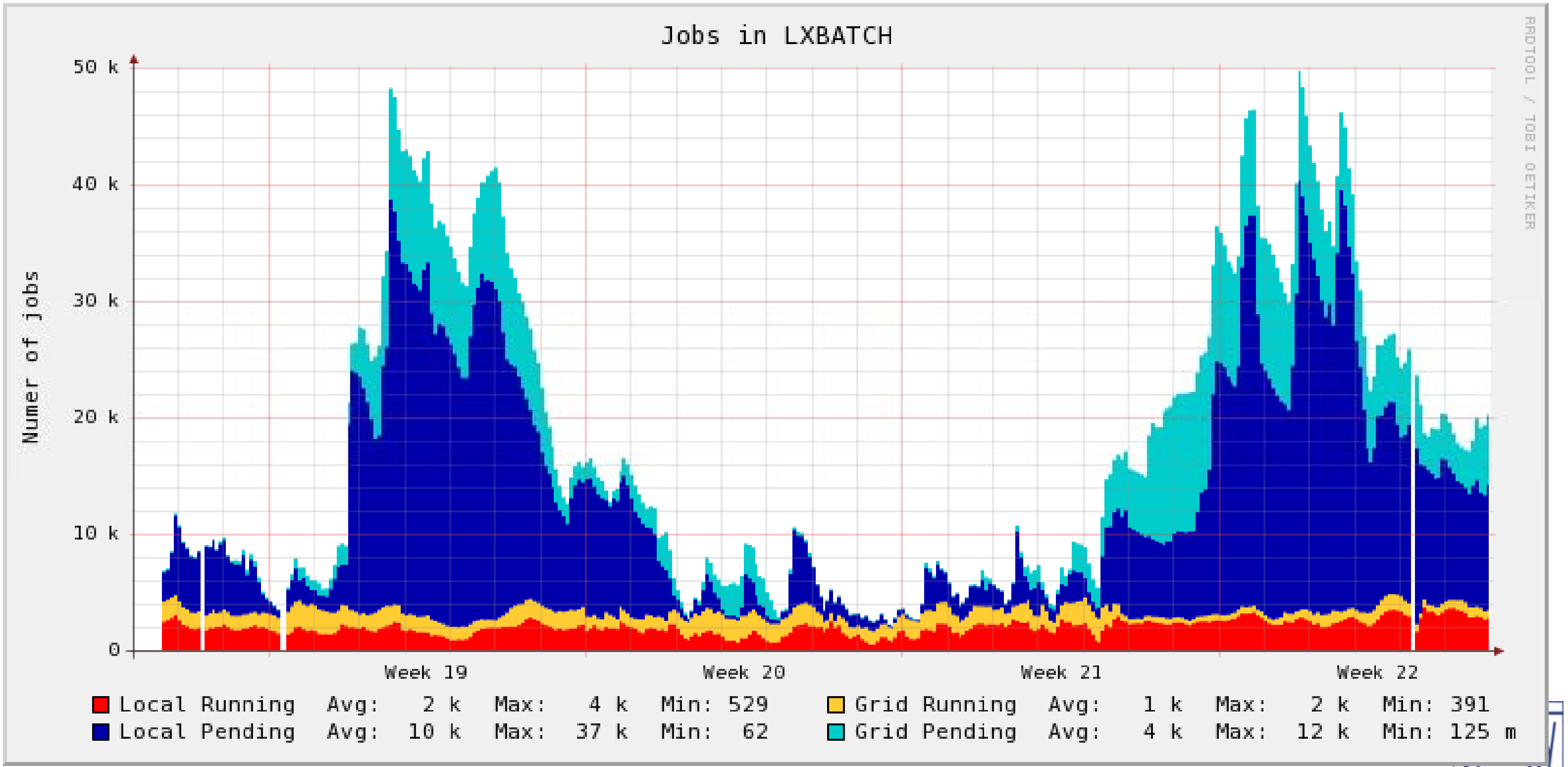
- What happened?
  - Equipment failure internal to CERN caused power to fail on the 18kV loop feeding B513.
- What went right?
  - Resupply of critical zone from dedicated supply.
  - Communications with CERN Control Centre much improved wrt previous incidents.
  - Recovery: Most services fully operational within 4 hours of power being restored.
- What went wrong?
  - Incorrect power connections in critical area, notably for essential network components (DNS, timeservers).
  - Communications between CC Operations and Physics Database Support team: services assumed to be OK due to presence of critical power.
- What next?
  - Re-organisation of power connections where necessary (many already done; network equipment pending).
  - Review recovery procedures for equipment split across physics and critical power (many physics database services).
  - Regular (3-monthly) live tests of physics power failure in the critical zone. First in June if possible.

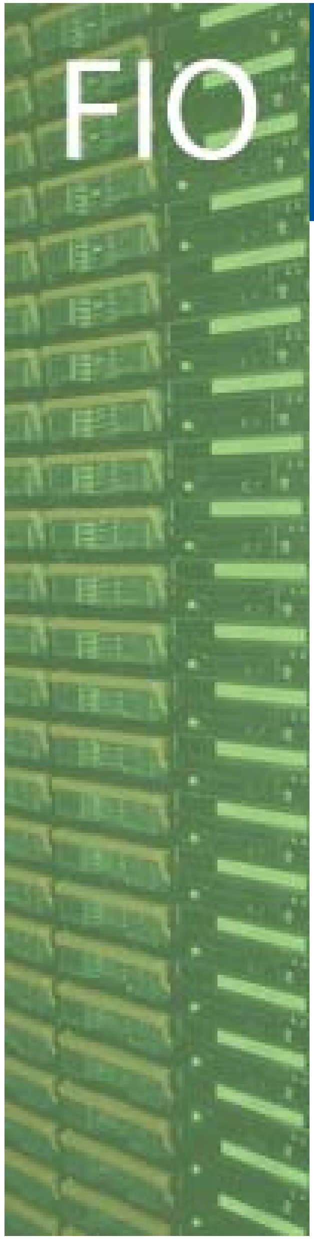


CPU Services

Activity overview

- Overall smooth running
- No visible problems found with the system
- 2.4 M jobs executed including 490 k GRID jobs
- Low average number of pending jobs

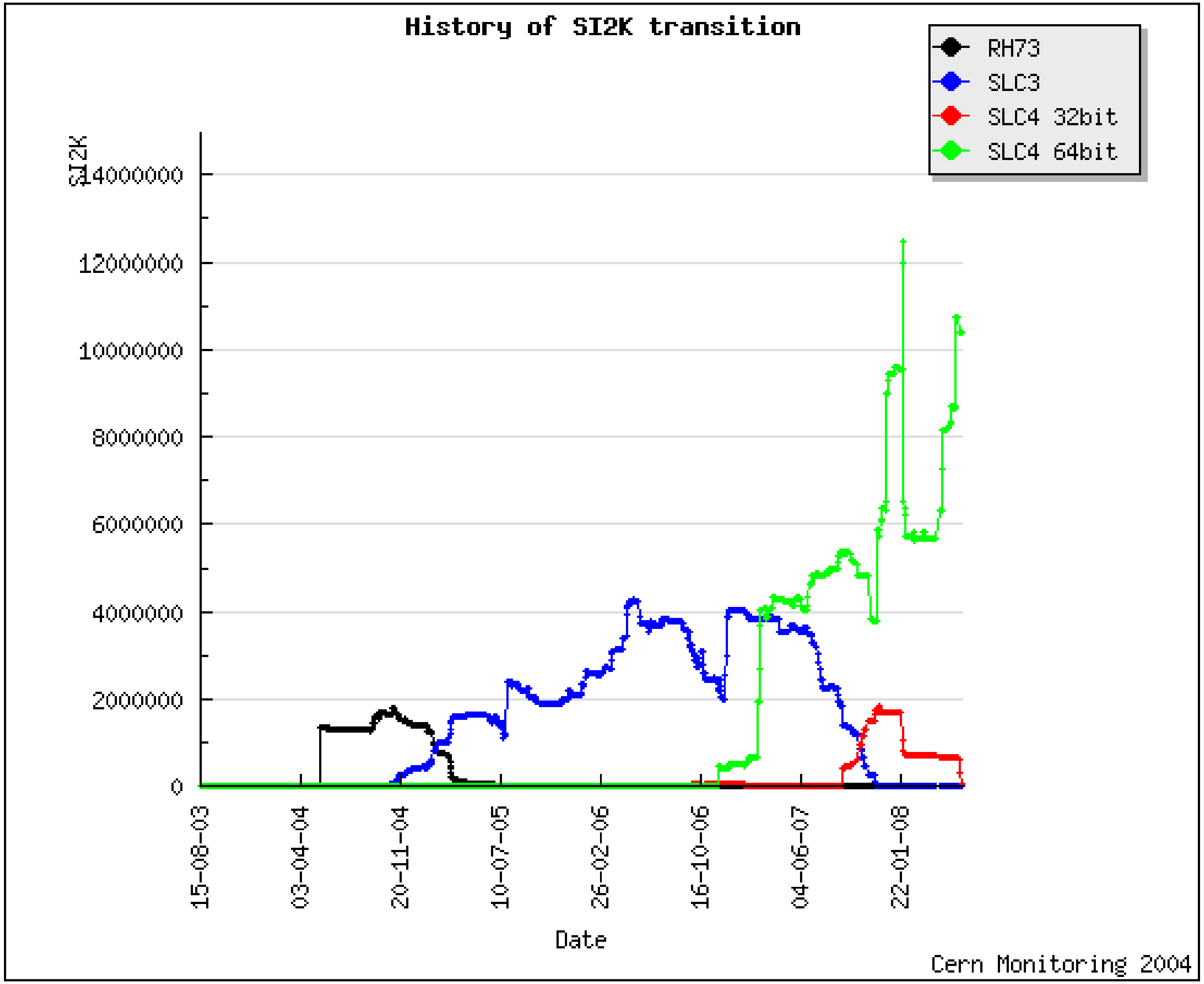


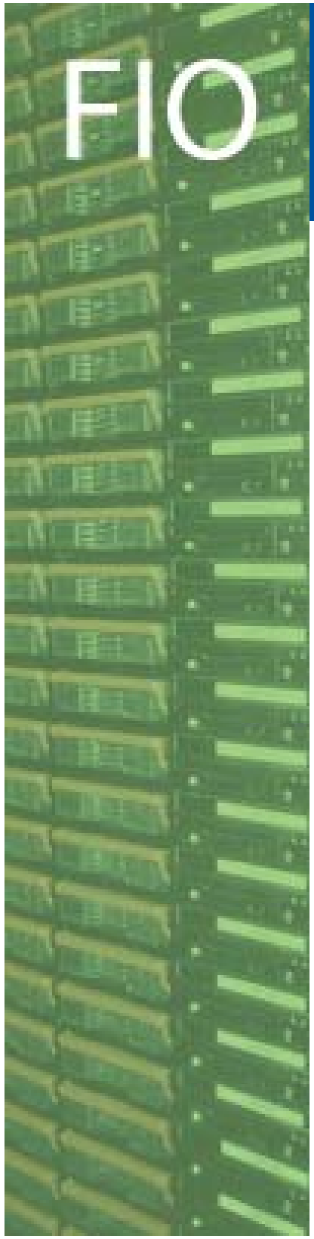


Preparations (1)

- Anticipating heavy usage new resources where added to the dedicated (T0) and public shares
- New version of LSF deployed after problems with double logging appeared in the end of the February run of CCRC'08
  - The load we saw in May now was not enough to properly test the system in production
- WN and CE Software versions were updated
- Old CE hardware replaced by new machines
- As recommended we changed back to publish physical CPUs instead of cores

Preparations (2)





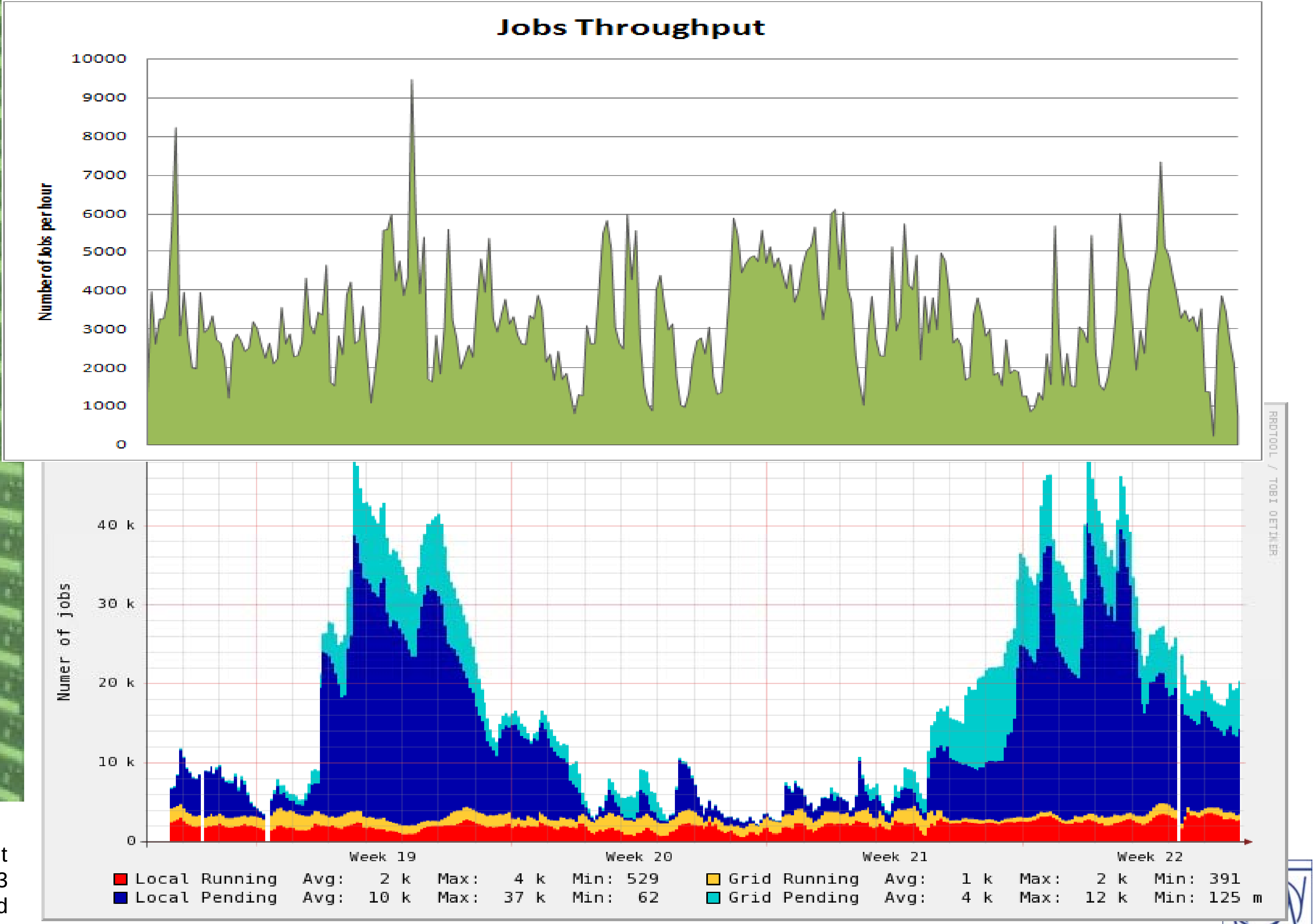
At the start of the exercise (1)

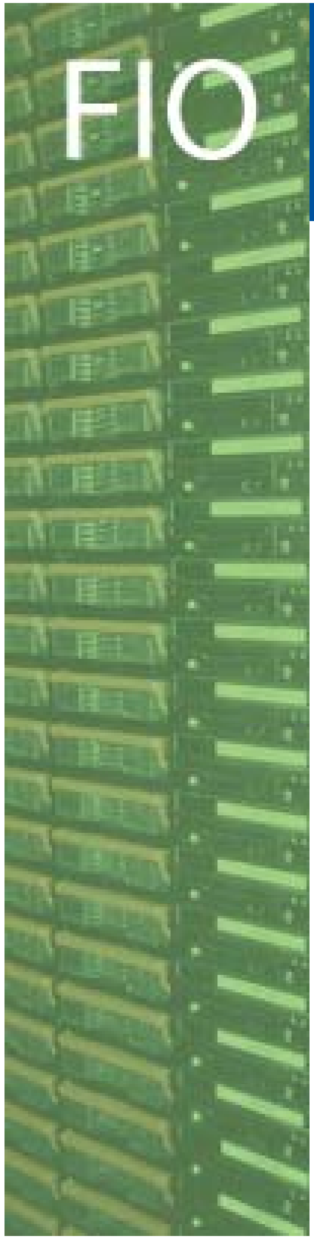
- The number of pending jobs in the system dropped
- No problems found in the system
- Maybe we looked “small” so we got less grid jobs?
  - In GridMap CERN was a ‘small’ site:
    - Not everybody had the same understanding of the recommendations
    - So we did the same, we now publish cores as physical CPUs
    - We saw almost no effect in the number of pending jobs

At the start of the exercise (2)

- Decreased activity overall?
  - 2.4 M jobs executed including 490 k GRID jobs
    - 33% more than in April
  - Average of 80 k jobs per day
- Increased capacity caused faster draining of queues?

Job throughput



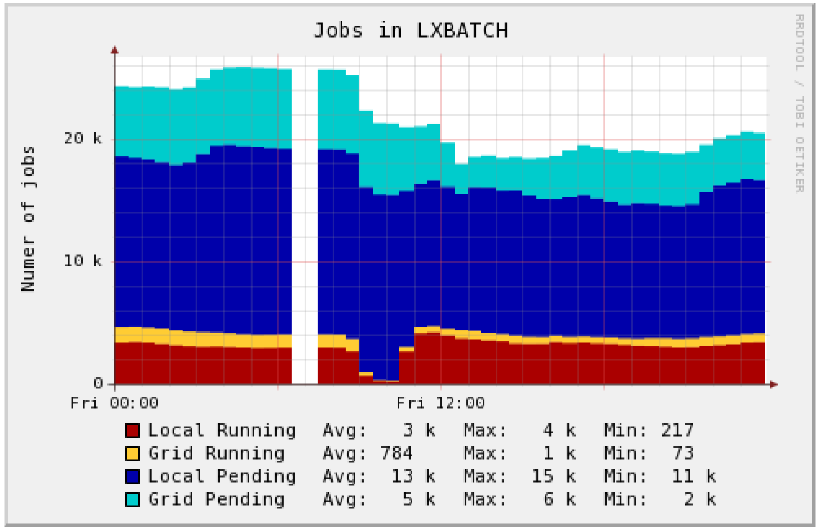


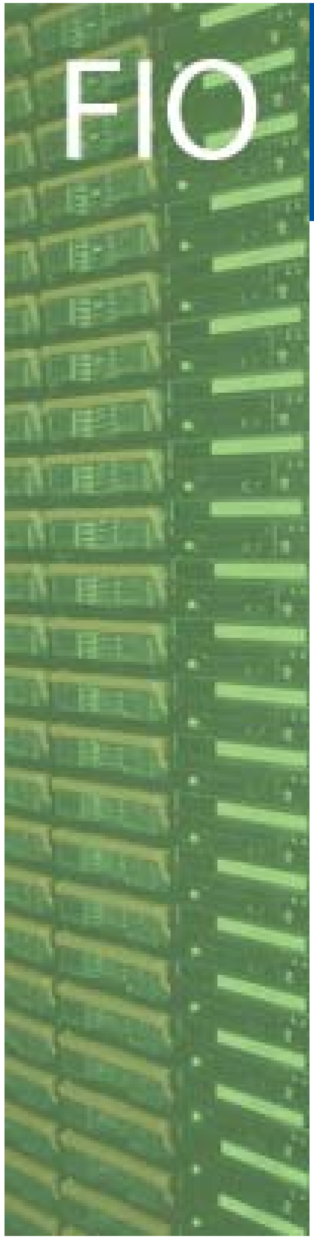
Job throughput

- Is this production running level?
- If yes, we are ready! ;)

No problems until…

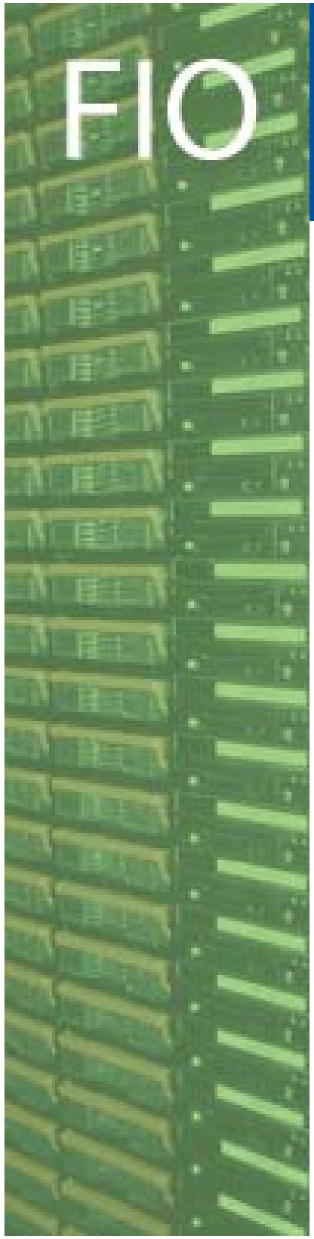
- Powercut at 6:00 am on Friday (30th May)
- lxmaster01 survived (critical power), scheduling was stopped
- Running jobs were not re-queued and were lost (~5k)
  - It would be nice to have at least local jobs marked as re-queable!
- Queues were reopened at 11:00 when CASTOR became available again
- Another 5k jobs terminated a bit to quickly
- Overall ~10k jobs were affected by the event





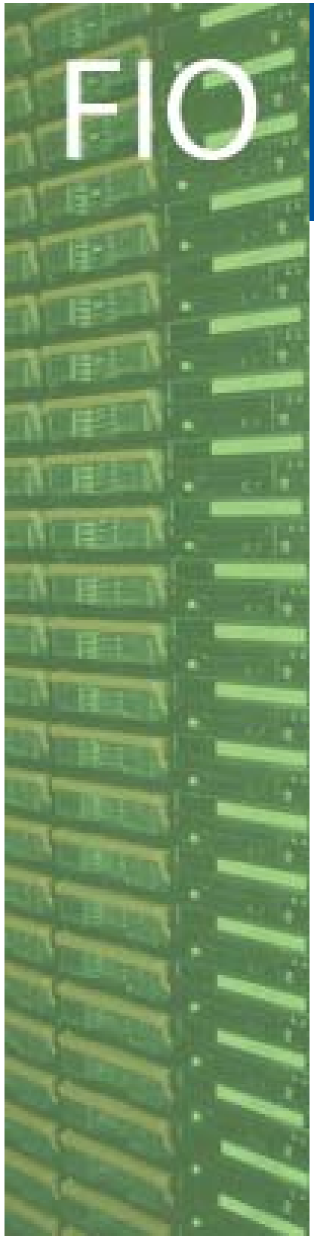
LFC Service

- Smooth operation (except for the power cut)
- Two bugs found in the gLite Middleware
  - Savannah #36550 and #36508
  - Monitoring modified to avoid service impact
- Software version used:
  - 1.6.8-1sec.slc4



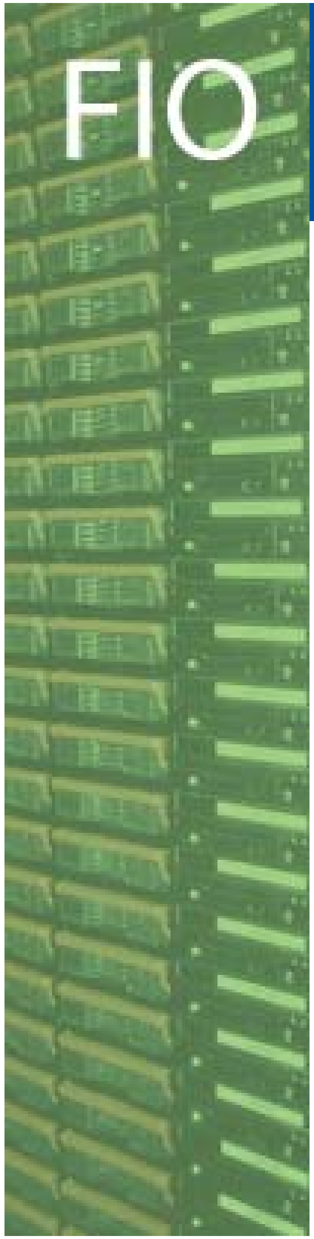
WMS Service

- Smooth operation overall
- A few problems
  - 2nd May:
    - Large fraction of CMS jobs aborted
    - Due to misconfiguration
  - 6th, 14th, 27th May:
    - CMS WMSes overloaded
    - Will hopefully be fixed when moving to SLC4 gLite 3.1 WMS (July?)
  - 30th May:
    - Power cut corrupted active job list on 1 WMS node. Painful recovery, some jobs were lost.
    - gLite 3.1 WMS "jobdir" functionality will avoid such problems, but is not yet configured
- In parallel with CCRC'08:
  - Pilot SLC4 gLite 3.1 WMS installed:
    - Used by CMS
    - Worked OK (but there are known issues)
- Note: LCG-RBs still being used a lot



All services

- Scheduled software upgrade was deployed including a kernel upgrade
  - Reboots needed in all machines
  - No problems seen in most of the services
    - VOBoxes
      - Some internal confusion in CMS lead to “unexpected” reboot of one important machine
    - LXBuild
      - Kernel was downgraded in ATLAS machines because it caused problems with their software. A preproduction machine is missing to test these changes before they go into production



Questions?