CMS operational issues

[GDB – CERN, 8 Sep 10]

Ian Fisk, Daniele Bonacorsi

on behalf of CMS Computing and all Data Operations and Facilities Operations teams

[special thanks to Josep Flix and Claudio Grandi for their material contributions to the slides]

News

No major changes in CMS Ops teams organization and meetings schedule

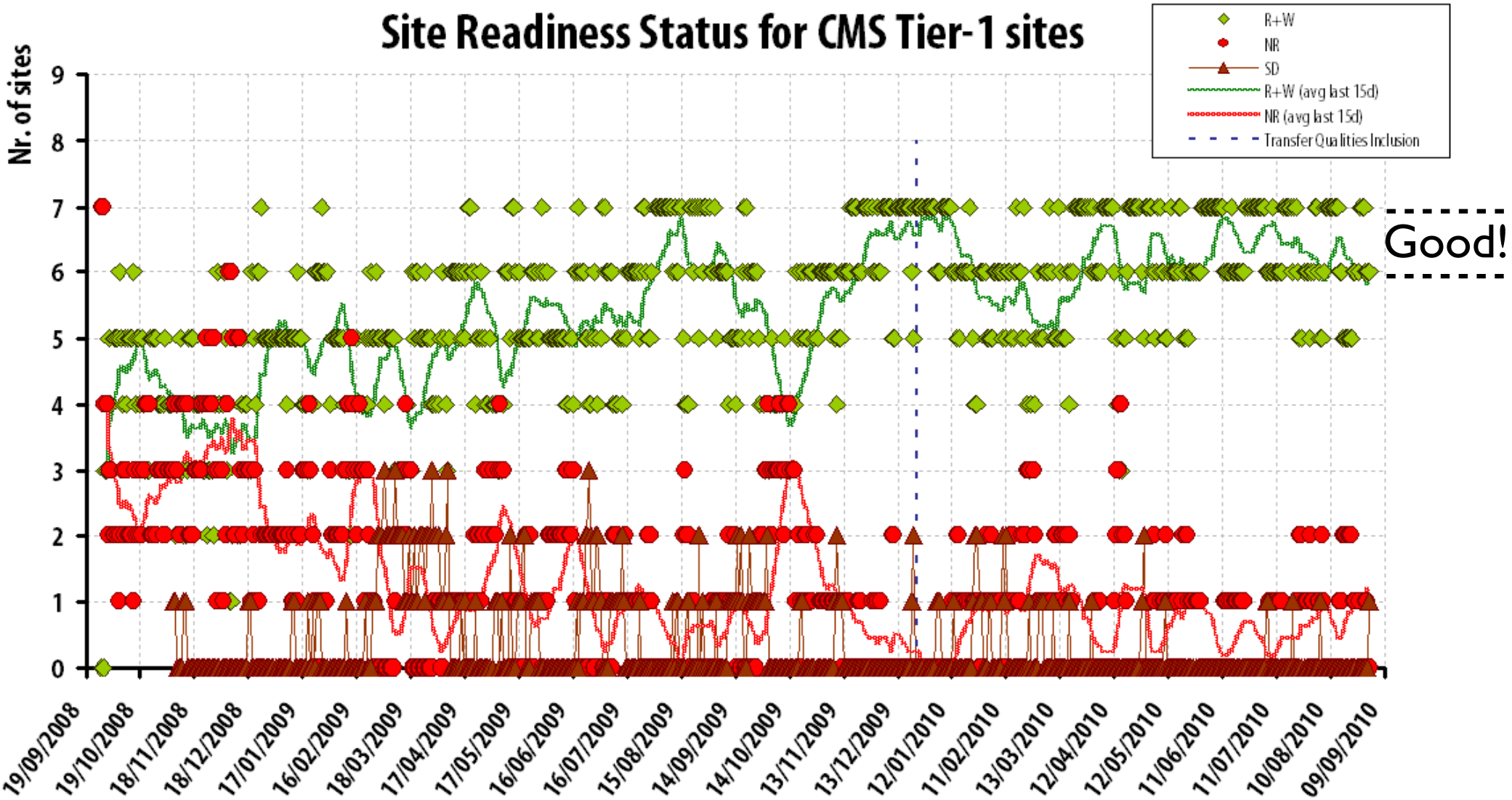
- Detailed info given at May GDB
  - http://bit.ly/d7c0dK (slides 2-3) by J.Flix
- **News**: the Analysis Operations team is fully operational
  - REMINDER: the Analysis Ops team is responsible for central data placement at T2 level, CRAB server operations, validation, and support, and for metrics, monitoring and evaluation of the distributed analysis system

CMS still regularly attends WLCG Ops daily calls

- twiki-based CMS reports since Feb 2009
  - https://twiki.cern.ch/twiki/bin/view/CMS/FacOps_WLCGdailyreports
- **News**: the CMS Computing Run Coordinator (CRC) on-duty regularly attends and reports on behalf of CMS
  - (s)he also reports back WLCG Ops items at Monday's CMS Operations meetings
  - Also attends T1 Services meeting, though normally others from CMS too

CMS Site Readiness: T1s

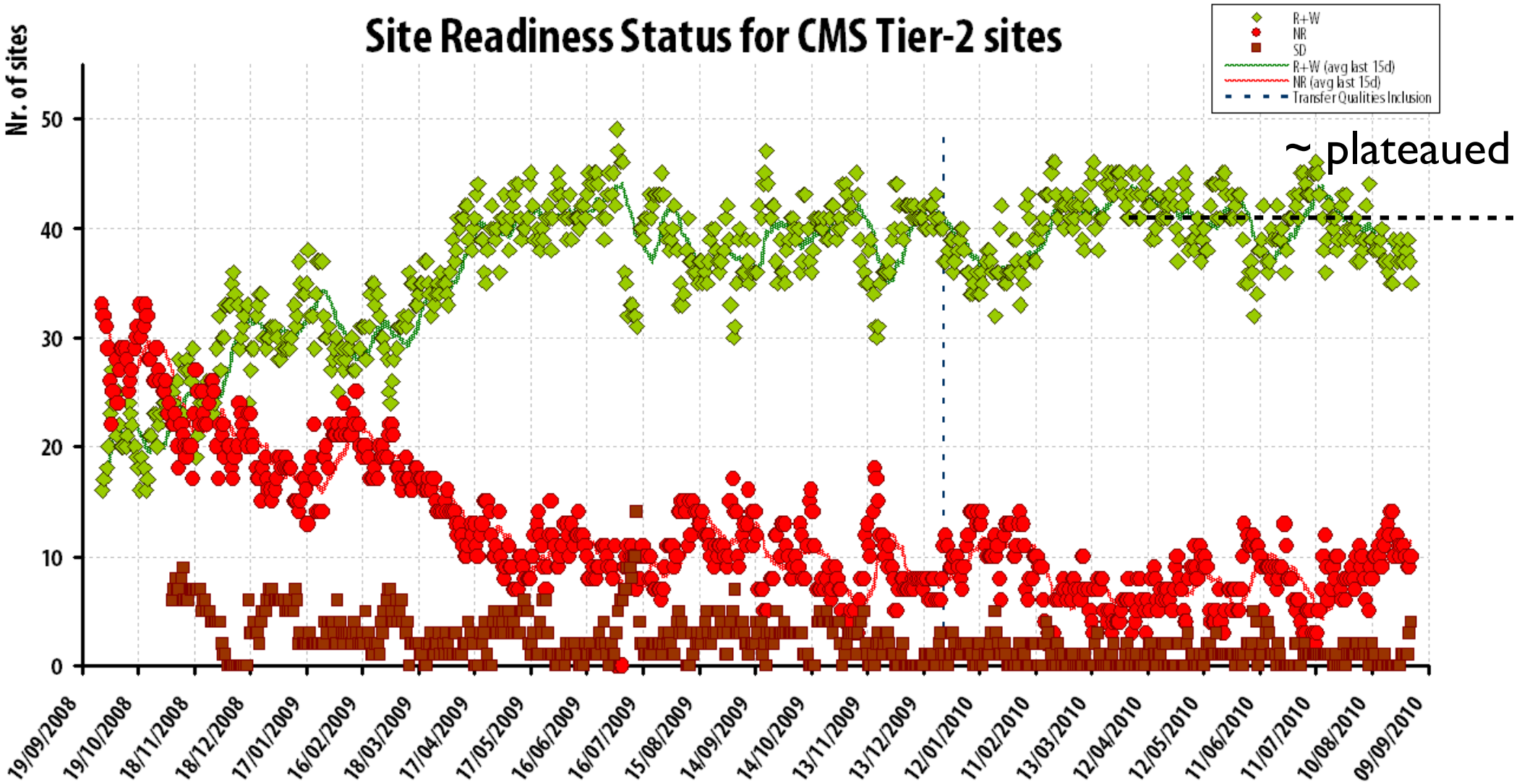
[Credits: J.Flix for CMS FacOps]



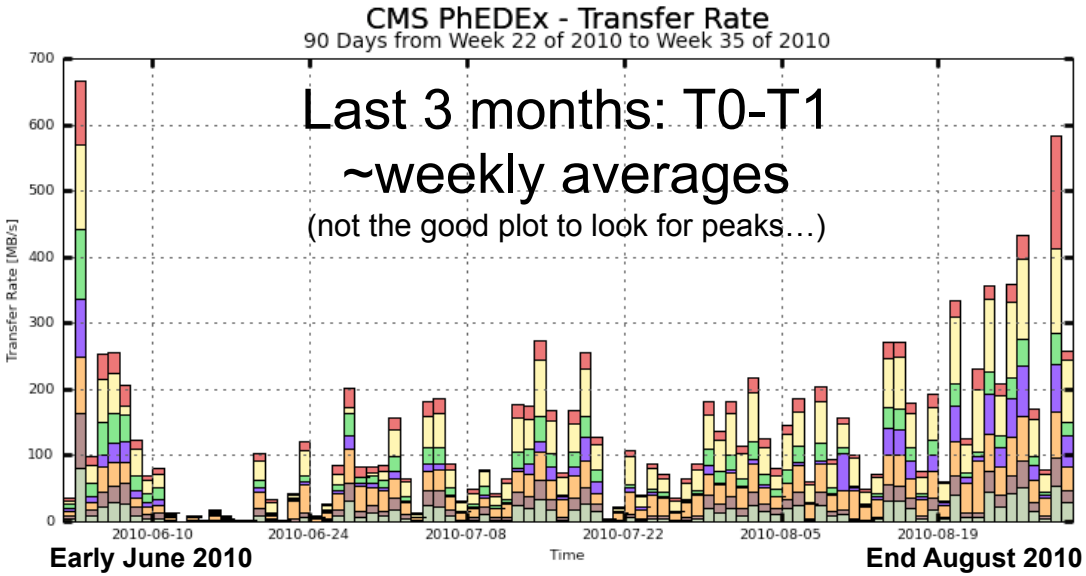
Good!

CMS Site Readiness: T2s

[Credits: J.Flix for CMS FacOps]



CERN outbound transfers

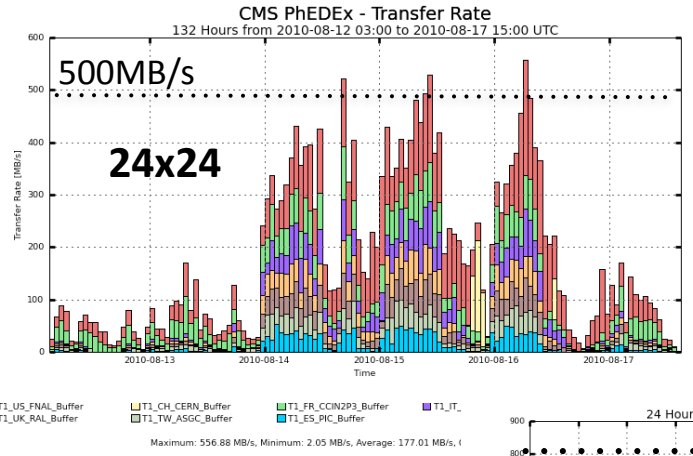


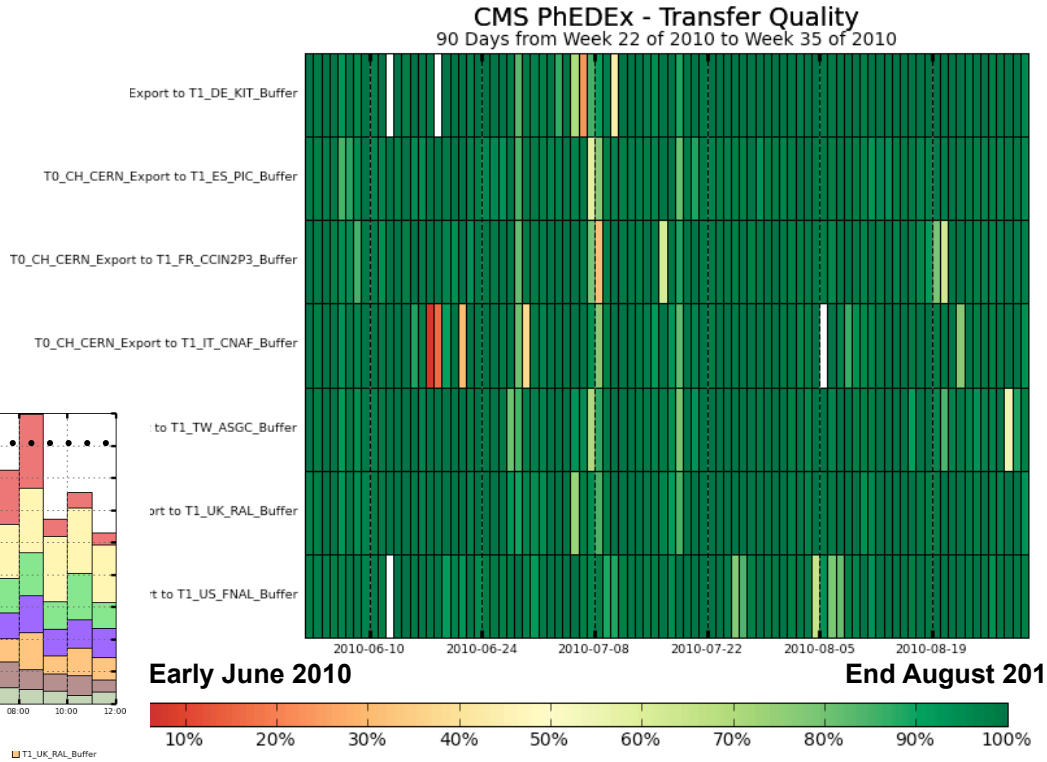
CERN outbound traffic corresponds to LHC fills

- All T1s participating stably
- Good share among T1s
- Peaks at >1 GB/s sustained

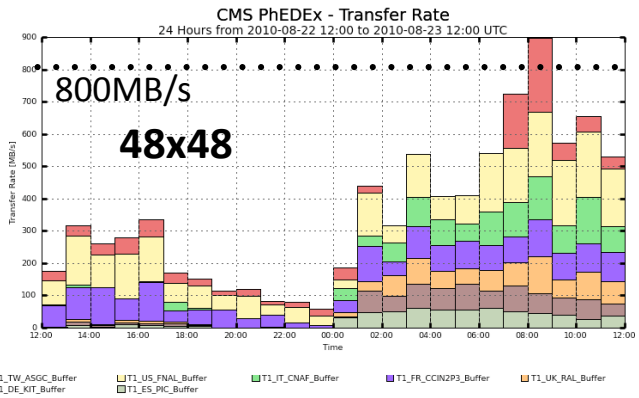
Transfer quality is remarkably high

- No exceptions





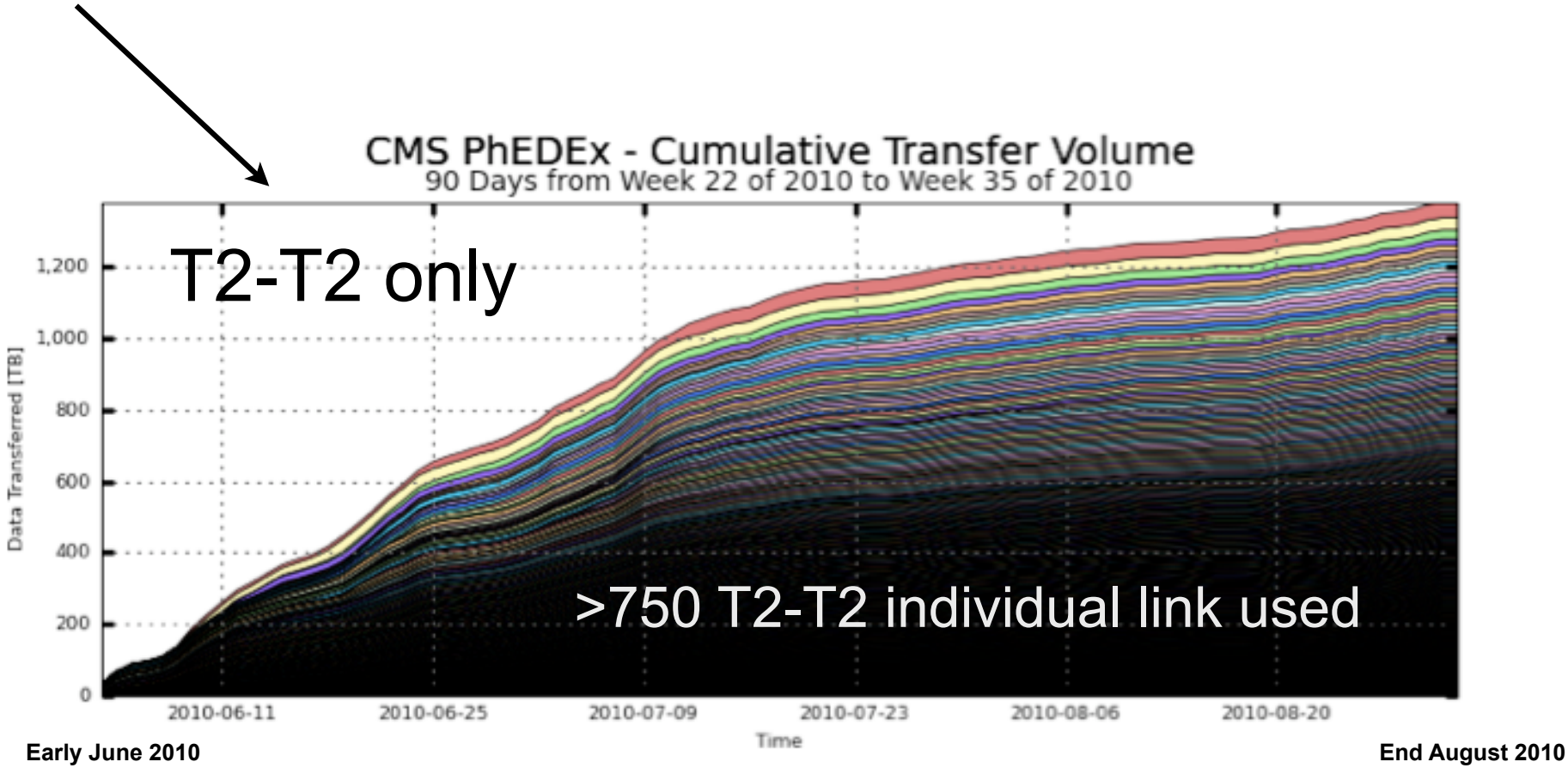
CMS maintained a trigger menu when moving to 48 bunches



Transfers in other routes

Last 3 months:

- T1-T2: peaks at 1.4 GB/s, 3-month aver: ~480 MB/s, ~3.5 PB moved
- T2-T1: peaks at 500 MB/s, 3-month aver: ~95 MB/s, 0.7 PB moved
- T2-T2: peaks at 600 MB/s, 3-month aver: ~190 MB/s, 1.3 PB moved



Major issues at T1s: highlights

[Credits: CMS Facilities Ops team]

ASGC

- SRM instabilities; CRLs issues; power cuts; CASTOR DB issues

CNAF

- BDII instabilities; some small incidents with their storage

FNAL

- some dCache instabilities; unmerge area filling issue (CMS issue, actually)

KIT

- 2 cooling incidents; PBS instabilities; NFS locking issues preventing CMSSW deployments

IN2P3

- Lot of SRM issues; problems with their AFS area; some memory troubles with dCache head node

PIC

- Service degradations (WNs) due to 2 cooling incidents in external module; problems with writing into a pool lately

RAL

- Generally ok

Major issues at T1s: more details [1/4]

[Credits: CMS Facilities Ops team]

ASGC

- Jul. 5: SAM lcg-cp test instabilities. 'Backend DB problem, with a lot of processes occupying large memory. Fixed by killing some processes.
- Jul. 23: CRL problems affecting SAM and jobs running at the site. Some CEs reported MaxWallClockTime incorrectly as well.
- Aug. 8: One squid server down for ~ 2 hours, fixed by restarting.
- Aug. 19-20: SRM server down due to overload. Fixed by restarting. A crond was setup to monitor the SRM server.
- Aug. 21: SRM server down. Local space was full.
- Aug. 23: unscheduled downtime due to power outage in Academia Sinica. Most services were recovered within 2 hours.
- Aug. 24: SRM problem at ASGC finally identified. Misconfig of CMS “t1production” role in SRM server.
- Sep. 1-3: unscheduled downtime due to castor DB issue. CASTOR data in cache not properly sent to disk before scheduled downtime (occurred on 31st Aug). This leaded to some inconsistency inside CASTOR DB when scheduled downtime was finished. It took ~ 3 days to make the recovery.

Major issues at T1s: more details [2/4]

[Credits: CMS Facilities Ops team]

CNAF

- Jun. 10: CMSSW SW tags publishing failure. Fixed promptly.
- Jun. 12/13: CNAF BDII instabilities.
- Jun. 18: declared unexpected downtime due to storage instabilities.
- Jul. 15: bug related to TSM clients preventing migrations to tape to proceed properly. New patch installed and problem fixed promptly.
- Jul. 21: unexpected problem occurred to the LDAP server, preventing some resources to be accessed properly (hardware problem).
- Jul. 22: authentication problem for exports from CNAF. An update process unexpectedly triggered the yaim tool using a wrong configuration file. This led to authentication problems for several cms users. Fixed promptly.

FNAL

- Jun. 8: Backup in Tape recall at FNAL causing transfer errors to CERN.
- Jun. 16: processing stopped completely because all the CMS unmerged space filled up. Data had to be merged first before deleting data. This was a CMS issue not a site issue, actually.
- Aug. 4: all transfers to FNAL failing. FNAL SRM failures. Resolved by restarting some of the dCache services.
- Aug. 5: dcache busy. Problem regarding with memory usage for one of its components. Memory raised and service stabilized.

Major issues at T1s: more details [3/4]

[Credits: CMS Facilities Ops team]

KIT

- Jun. 13: unavailability of cms dcache head node for several hours.
- Jul. 10/11: cooling incident affecting services at the site.
- Ju. 12/13: cooling incident affecting services at the site.
- Jul. 15: CMSSW installation problems due to issues with NFS locking.
- Jul. 23/24/25: unsched downtime to roll back PBS 10.4 to 10.2. CE's PBS misconfiguration, related to security patch.
- Aug. 13: CMSSW installation problems due to issues with NFS locking.
- Aug. 13/14/15: PBS problems affecting jobs. Seem there is pbs bug on server side wh makes the server hang when pbs_mom on remo worker node gets into "problematic state". The problematic state of WNs is not clarified yet in details, people are still investigating root cause incident.
- Aug. 24: CMSSW installation problems due to issues with NFS locking.

IN2P3

- Jun 5: SRM glitch affecting SRM service for 1h.
- Jul. 9: timeout when accesing the CRLs file in the AFS volume onsite. Impacted all import transfers for several hours.
- Jul. 11: SRM down for several hours.Jul. 21: temporary issue on SRM server caused failure of SRM-lcg-* commands.
- Jul. 29: problem installing new CMSSW releases. AFS connection timeouts.
- Aug. 4: Communication problem between WMS and CE.
- Aug. 5: SRM problems. Reboot of the server.
- Aug. 15: SRM failures for several hours.
- Aug. 19/20: Failure on the dcache pool manager. An unscheduled downtime was put in place, and pool manager memory was raised up.
- Aug. 29: SRM failure. Solved by rebooting the system.

Major issues at T1s: more details [4/4]

[Credits: CMS Facilities Ops team]

PIC

- Jun. 3: a DN in glite-data-transfer-veto-mapfile was banning all transfer requests in PIC FTS service. DN added during a test security incident, and few days later it caused a full ban for almost all FTS requests. Developers were asked about this, since it seems a bug or misconfiguration.
- Jul 24/25: cooling failure in PIC external module. Service degradation as some WNs were stopped.
- Aug. 9: cooling incident in PIC external module (where 2/3 of cores are located). Service degradation, as WNs were stopped and ran at 30% levels of CPU for 2 days.
- Aug 26/27+29: SAM SRMv2-lcg-cp failures. One disk pool was kind-of-problematic. Was put in read-only mode and experts started investigating the origin of this writing error on that particular pool.

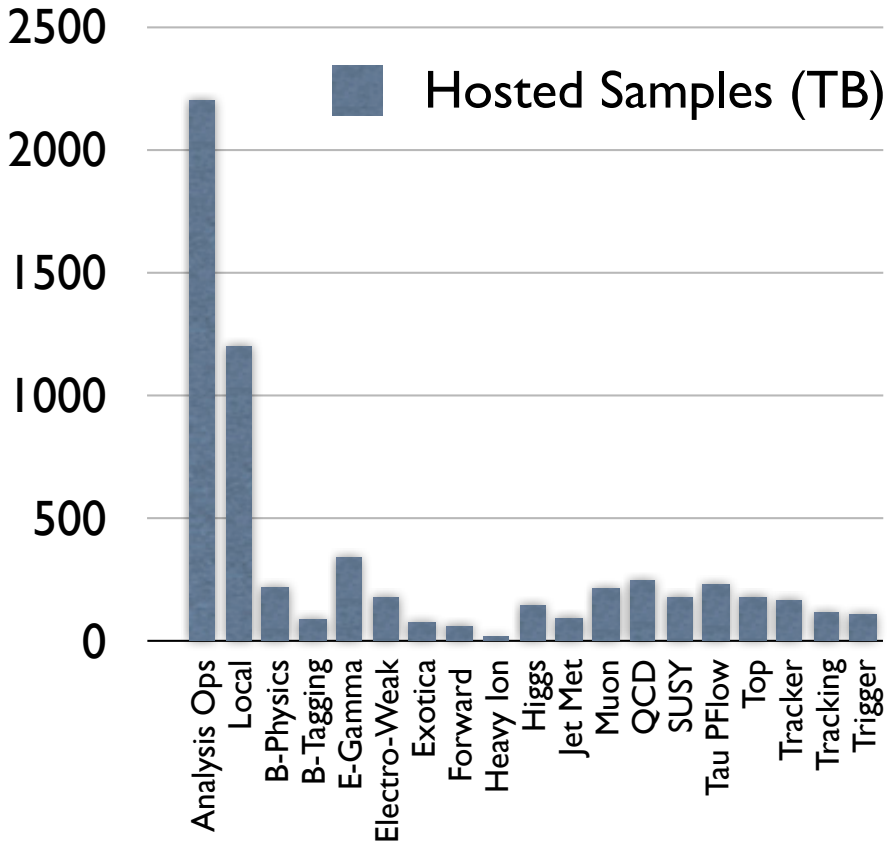
RAL

- Aug. 12: Large amount of idle data at T1_UK_RAL. Some data inconsistent and fixed promptly.

Data Hosted at T2 level

Data hosted by Analysis Operations exceeded 2 PB

- All Physics Groups managing Tier-2 space
  - Some even deleting data



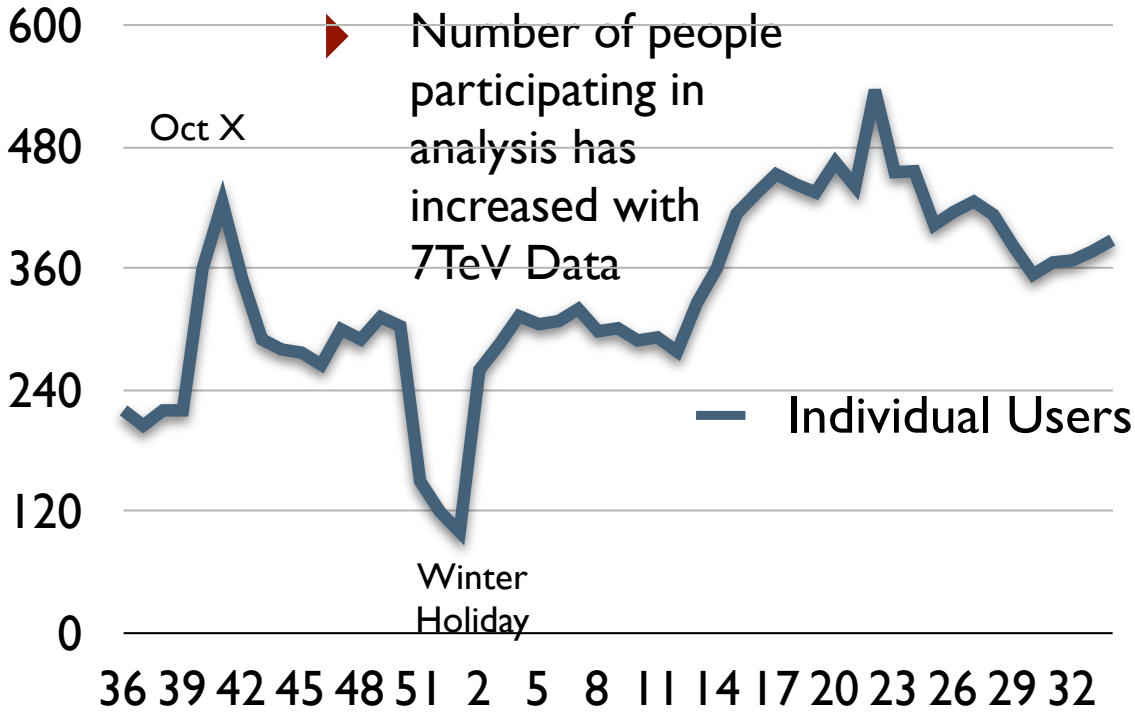
Systematic analysis of the data being accessed currently underway with Analysis Ops

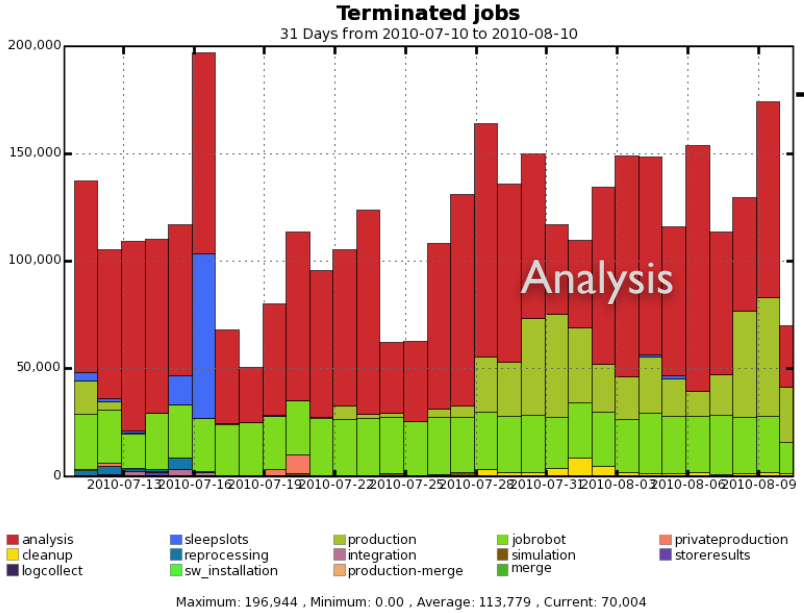
- Aim to more efficiently use the disk space

Analysis

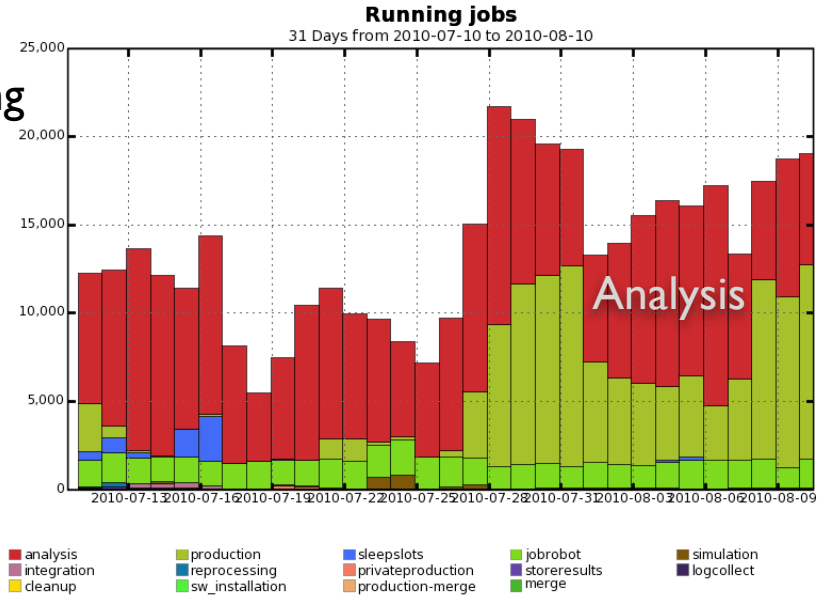
After post-ICHEP decrease and August, analysis users ramping back up

- Analysis and Simulation should be roughly equal. In periods of slow MC production analysis dominates





Terminated and running jobs from mid-July to mid-August



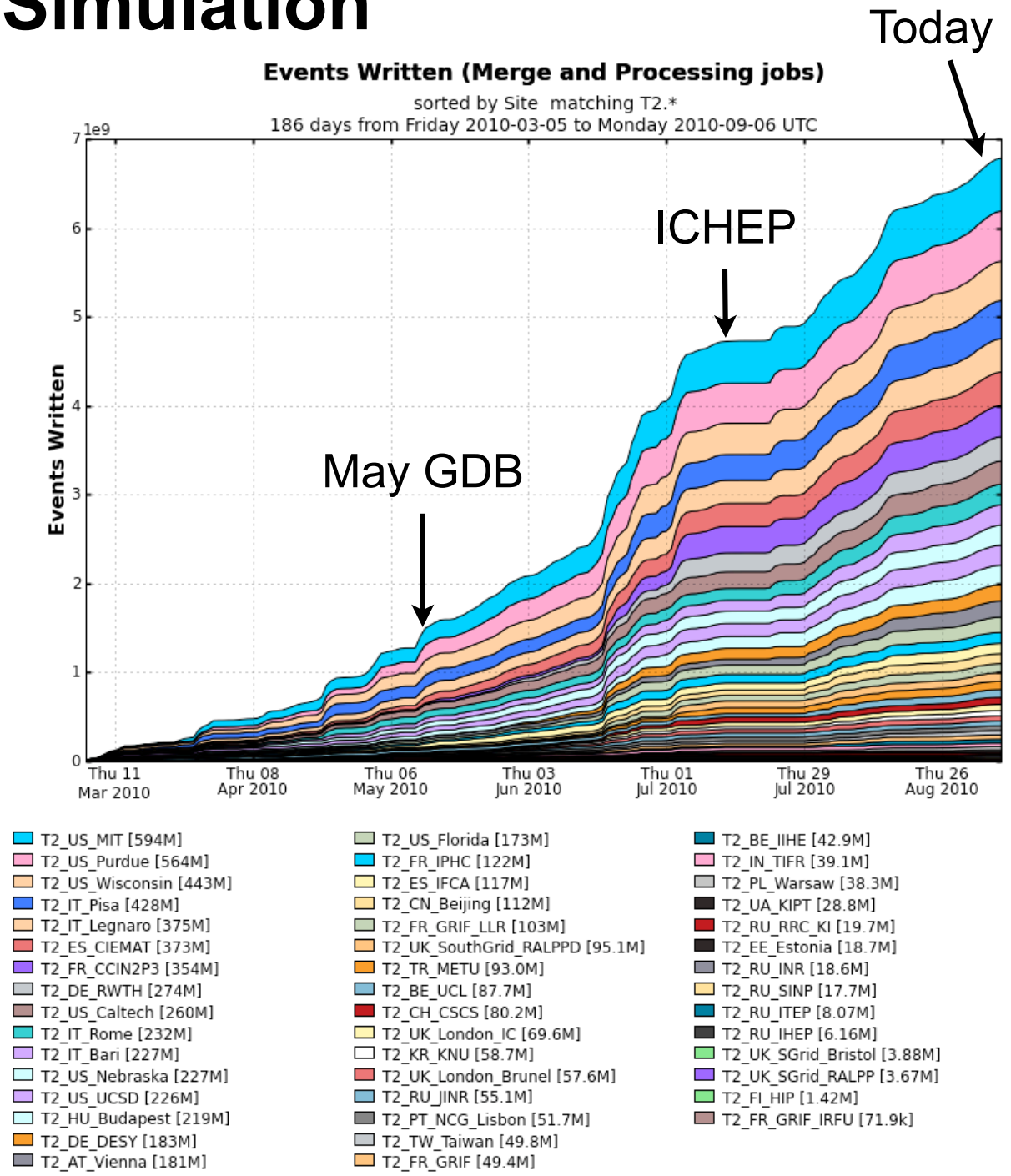
Simulation

Simulation at the Tier-2s going smoothly

- And increasingly Tier-1s
- Peak before ICHEP
- Good balance across Tier-2 sites

CMS just released nearly a 1B event request for new simulation

- Improvements in tuning and beam spot. Many will be pile-up requests.



CMS usage of CREAM-CE

Mid-July 2010:

- all WMSs in use by CMS (INFN + CERN) were updated with patch #3621, that contained all the critical fixes needed by CMS

Summer 2010:

- CREAM CEs were put back in normal production operations
  - No indication of problems for CMS since then
- More: a few more sites are opening CREAM-CEs to CMS so that they are automatically available also to analysis activities
  - from lcf-info: 62 CEs (i.e. queues) open to CMS. The number of machines (one or more queues) open to CMS are 51. Many of them are not T2/T3 of CMS though.
  - Also in this case we didn't get any indication of problems.

No precise information about the use of CREAM via Condor-G by CMS

- Condor->CREAM submission was known to have problems with the current version of Condor
  - problems reported mainly by ATLAS. Fixes should have been made available in Condor 7.5.3 (Development release, released on Jun 29, 2010)
- AFAWK ATLAS is doing tests with that version

Pilot Submission

CMS has a pilot submitter to the glide-in WMS at CERN

- Glide-in factory is still offsite
- Useful for allowing operations team to use both pilot and non-pilot processing

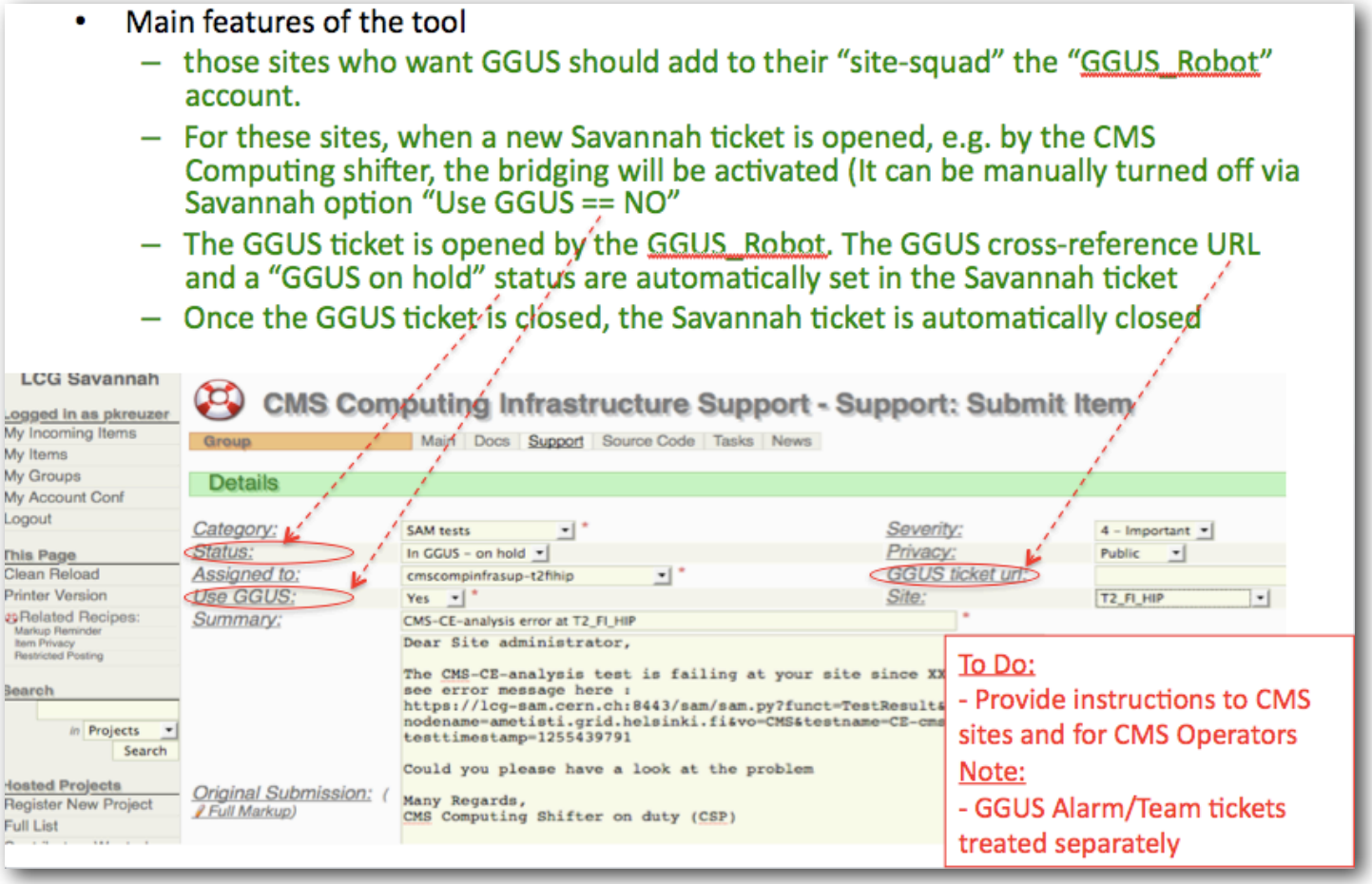
CMS still has roughly half the reconstruction jobs at Tier-1s submitted through the gLite WMS

- Continuing to support both submitters for analysis and organized processing

Bridging of ticketing systems

CMS requested a Savannah-to-GGUS bridging

- Work finalized. Tested. Basically works. Starting to be used in Ops.
- Current preliminary feedback is positive
  - it works and atm CMS is happy
  - we do have several feature requests for the future, need to make sure we know whom to address them



Back-up

CMS Computing shifts

Started in Fall 2008 with 16/7 coverage

- paused during Winter/Spring 2009
- resumed in August 2009, with 24/7 coverage (3 timezones)

Currently CMS Offline/Computing shift crew pool

- 35 people in 3 timezones, mainly non-computing experts

Shift roles

- Computing Shift Personnel (CSP) monitor systems and raise alarms
  - contributes to the standard MoA service work defined by CMS, accounted as the Online Central CMS shifts
- Other roles support the CSP: see box

Shift procedures and checklists

- https://twiki.cern.ch/twiki/bin/view/CMS/ComputingShifts

Computing Run Coordinator (CRC)
- Subscribes to all CSP E-log sub-sections
- Assists CSP in raising alarms/tickets for complex cases
- Calls EOC during off-working hours (see below)

Core Computing Operator (FacOps, DataOps Experts)
- Subscribes to relevant CSP E-Log sub-sections
- Supports CSP during working hours

Computing Expert On Call (EOC)
- Responsible of a particular service
- Alarmed by CSP via Email/IM/Tel during working hours
- Alarmed by CRC if really needed off-working hours

CMS Site Contact Person
- Responds to alarms (e.g. Savannah, GGUS tickets)

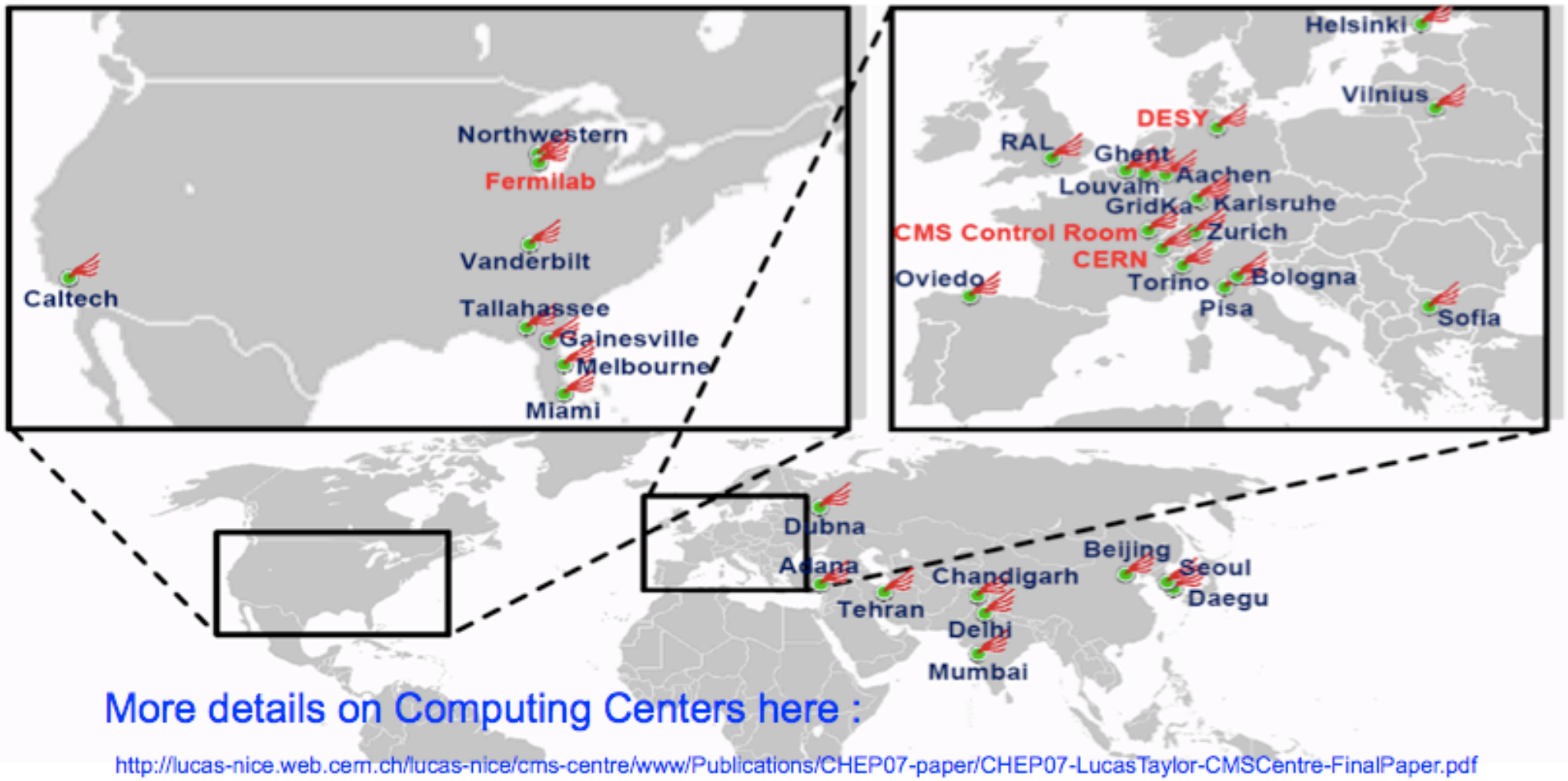
Other shifters (DQM, Online, Detector, ...)
- In temporary absence of CRC, the CSP is the Core Computing contact for any shifter at P5/CMS Center/FNAL ROC

CSP procedure responsible
- Assigns CSP shifts

More CSP tools/info

- Stable EVO in ad-hoc virtual room: http://evo.caltech.edu/
- Shift Sign-Up tool: http://tinyurl.com/r78zsb
- IM under generic “FacOpsShifter” account (instructions on a twiki)
- Computing Plan of the Day: http://cmsdoc.cern.ch/cmscc/shift/today.jsp
- Account in the CSP E-log: https://prod-grid-logger.cern.ch/elog/
- Savannah: https://savannah.cern.ch/projects/cmscompinfrasup/

CMS Computing shifts and CMS Centres



More details on Computing Centers here :

http://lucas-nice.web.cern.ch/lucas-nice/cms-centre/www/Publications/CHEP07-paper/CHEP07-LucasTaylor-CMSCentre-FinalPaper.pdf

- 3 "Primary CMS Centers" : P5 Control Room, CMS Center CERN, FNAL ROC
- Many existing or upcoming "Secondary CMS Centers"
- Primary Centers permanently connected via Tandberg Video system
- Secondary centres using permanent EVO room
- Both primary and secondary centres used for Offline Computing Shifts