"CMS Update"

LATBauerdick, July 20, 2005

Overview

- Plans for SC3 sample jobs
  - rather data flow and processing scenario, and SC3 goals
- SC3 resource needs
- Planning updates
- Current problems and issues (SC3 and LCG)

CMS SC3 Organization

- CMS Computing Integration Program working with SC3 team
  - CMS SC3 lead is Lassi Tuura
  - building the integration team helping this effort end-to-end
- CMS contacts at regional centers where CMS hosts datasets
- This worked very well, and people have worked very hard

- Thank you!!

CMS Service Challenge Overall Goals

- An integration test for next production system
  - **Full experiment software stack** – not a middleware test
    - "Stack" = s/w required by transfers, data serving, processing jobs
  - **Checklist on readiness for integration test**
    - Complexity and functionality tests already carried out, no glaring bugs
    - Ready for system test with other systems, throughput objectives
    - (Integration test cycles of ~three months – two during SC3)
  - **Becomes next production service** if/when tests pass
- Demonstrate all CMS data transfers and access the data with analysis applications to stress sites data serving and grid WMS
- Measure and understand efficiencies
- Demonstrate capability to operate at same time as other VO's

Qualitative Goals

Throughput Phase

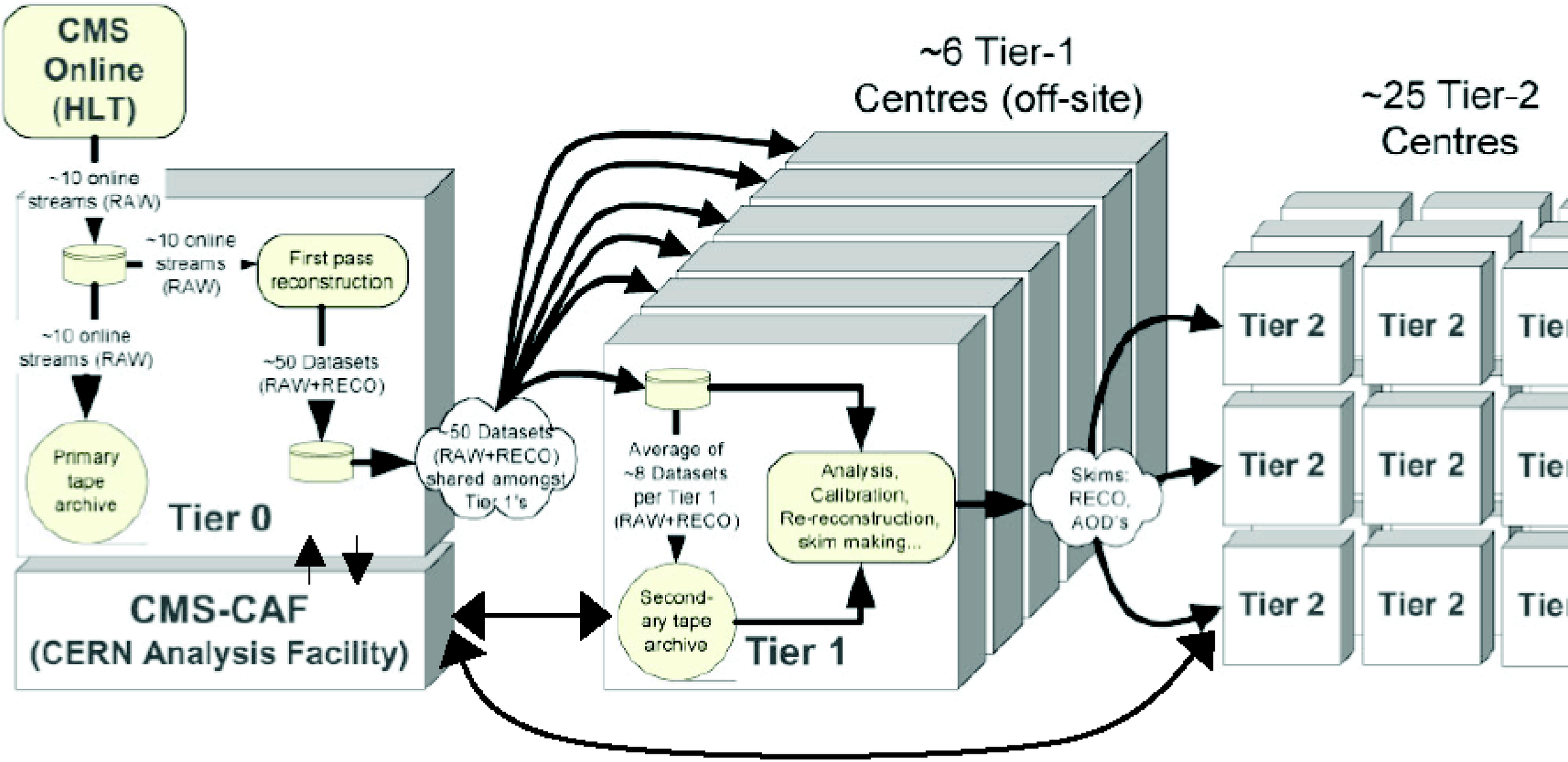
- Overview of throughput exercise
  - Throughput to disk and tape at Tier-1s from CERN Tier-0 disk
  - Fan out transfers to selected Tier-2s, same data but less of it
  - Target: transfer and storage systems work and are tuned
    - Using real CMS files and production systems (or to-be production)
    - Sustained operation at required throughput without significant operational interference / maintenance
- Concretely
  - Part 1: Data from disk buffer at CERN first to Tier-1/2 disks
    - Tier-2s will be subscribed subset of the data going to Tier-1s
    - Data to Tier-2s are routed via Tier-1s
  - Part 2: Same, but data goes to tape at Tier-1s
  - Transfers managed by PhEDEx
  - Files registered to local file catalogue
  - Sufficient monitoring

Quantitative Goals

Throughput Phase

- Rates defined in Jamie's document
  - Tier 0 disk to Tier 1 disk 150 MB/s sustained
  - Tier 0 disk to Tier 1 tape 60 MB/s sustained
  - Tier 1 disk/tape to Tier 2 disk ? MB/s sustained
  - Tier 2 disk to Tier 1 disk (tape?) <1 MB/s (!?) sustained
  - Suggest informally 30 MB/s T1 to T2 if bandwidth is available
- In addition: service quality
  - Transfer failures should have no significant impact on rate
  - Transfer failures <0.1% of files more than 5
  - Catalogue failures after transfer <0.1% of files
  - File migration to tapes (keep up with transfers)

Goal for Service Phase:
Test most of CMS CM Data Flow



Qualitative Goals

Service Phase

- Overview of service exercise
  - Structured data flow executing CMS computing model
  - Simultaneous data import, export and analysis
- Concretely
  - Data produced centrally and distributed to Tier 1 centres (MSS)
  - Strip jobs at Tier 1 produce analysis datasets ("fake" COBRA jobs)
    - Approximately 1/10th of original data, also stored in MSS
  - Analysis datasets shipped to Tier 2 sites, published locally
    - May involve access from MSS at Tier 1
  - Tier 2 sites produce MC data, ship to Tier 1 MSS ("fake" COBRA jobs)
    - May not be the local Tier 1
  - Transfers between Tier 1 sites
    - Analysis datasets, 2nd replica of raw for failover simulation
  - Implied: software installation, job submission, harvesting, monitoring, VO + group roles

Quantitative Goals: Tier 1

Service Phase

- For two periods of at least one week each, sustain
  - Same service quality goals as with throughput phase
  - All transfers and data serving are to/from tape at Tier 1s
  - Data served to worker node jobs: bytes read by instrumented CMS apps (ROOT), not dcap/rfio/… (excludes file transfers!) — 200 MB/s
  - Data stored from worker node jobs — 12 MB/s
  - Transfers from Tier 0 — 3 TB/day (~36 MB/s)
  - Transfers to Tier 2s (all if more than one) — 1.5 TB/day (~18 MB/s)
  - Transfers to Tier 2s (each) — 1 TB/day (~12 MB/s)
  - Transfers to Tier 2s (each, minimum) — >10 MB/s [24+ hours]
  - Transfers to Tier 2s (each, if bandwidth exists) — 30 MB/s [24+ hours]
  - Transfers from Tier 2s (each) — 2.5 MB/s
  - Time from Tier 0 file availability to available for analysis applications at Tier 1 — 10% <15 min; 33% <30 min
  - Skim data to 1/10th and store to tape — (keep up with input)
  - Job success rate — >95%? (to be defined)
  - Job throughput — ?/day (to be defined)

Quantitative Goals: Tier 2

Service Phase

- For two periods of at least one week each, sustain
  - Same service quality goals as with throughput phase
  - Data served to worker node jobs: bytes read by instrumented CMS apps (ROOT), not dcap/rfio/… (excludes file transfers!) — 100 MB/s
  - Data stored from worker node jobs — 2.5 MB/s
  - Transfers from Tier 1 — 1 TB/day (~12 MB/s)
  - Transfers to Tier 1 — 0.2 TB/day (~2.5 MB/s)
  - Time from Tier 1 file availability to available for analysis applications at Tier 2 — 10% <15 min; 33% <30 min
  - Job success rate — >95%? (to be defined)
  - Job throughput — ?/day (to be defined)

Quantitative Goals: Other

Service Phase

- Various constraints
  - Tier 1 strip jobs to keep up with incoming data
  - Tier 1 tape system able to migrate files at incoming rate (T0 + T2s)
  - Tier 1 data export able to keep up with data-producing jobs
  - Tier 2 data export able to keep up with data-producing jobs
- Other components
  - Resource broker able to accept jobs — N secs (to be defined)
  - RB and CEs/WNs able to process jobs — N/day (to be defined)
  - Grid infrastructure-related job failure rate — <5% (to be defined)
- Still undefined (or monitored) quantities
  - Latency from data block request to delivery
  - Number of data requests processed by Tier 1
  - File delay from request to start of transfer for MC and hosted data
  - Time for file to sit in Tier-2 cache
  - Frequency of Tier-2 cache refresh

Checklist Goals

Service Phase

- Automatic installation of CMS software works
- PhEDEx available, all file transfers executed with PhEDEx
- PubDB available, automatically updated from PhEDEx, updates RefDB
- Harvesting of job output files works: injected to PhEDEx, transferred
- File catalogue operational
  - Automatically updated by file transfers, harvesting
  - Functional for all jobs running on worker node
- UI installed with access to CMS software, test data samples accessible
  - Can compile, test, debug and submit CMS jobs to all sites from UI
  - Can receive jobs from all other CMS sites
  - "All sites" = "All CMS sites participating in the challenge"
  - "Submit" = "Submit using CRAB", "Run" = "As submitted fro CRAB"
- Worker nodes have access to CMS environment
  - Software, site configuration scripts, file catalogue, harvest agents, …
- General monitoring sufficient (to be defined)
- Optional: BOSS job monitoring provided (UI, database) and works

Resource Needs: Data Sample Sizes

Service Phase

- Total data capacity
  - 50 TB from CERN to at least two Tier 1 sites
  - ~10 TB from CERN to other Tier 1 sites
  - ~5 TB to each Tier 2
  - 5-10 TB T1/T1 analysis dataset transfers
  - 50 TB T1/T1 2nd raw replica transfers (Tier 1 failover)
- Data can be discarded after a while
  - Data for service phase may need to be kept for a while (month)
- Most likely no need for large CPU capacity
  - Submitting jobs to normal worker nodes,
    expect access to SC storage
  - Reasonable capacity available for two or three periods of a week at a time

SC3 Service Phase CMS Timeline

- Service Phase CMS-1 (Sep/Oct)
  - Move and validate and publish (PubDB) data
    - T0->T1->T2
    - T1->T1
    - T2->T1
  - store data to tape at T1, running CRAB jobs a few days after data had been moved
- Service Phase CMS-2 (Nov?) -- to be refined
  - as above, if possible in "high throughput" by other exp
  - "late" Tier-2s join
  - in addition, "full" data flow use case for Tier-1s: run fake skims with CRAB at T1 and move results to T2

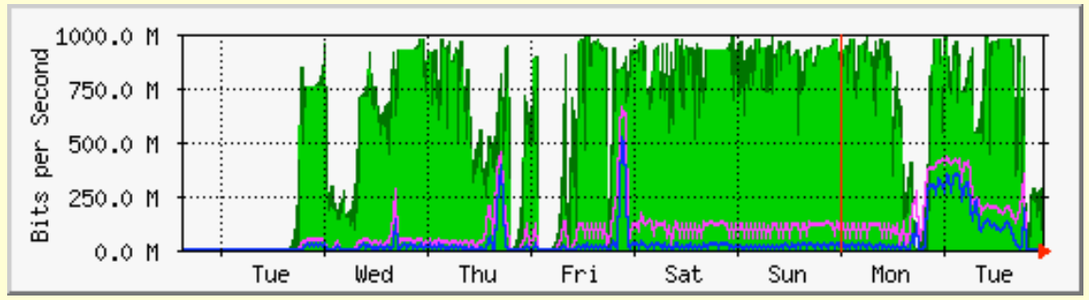
What We Achieved Until Now

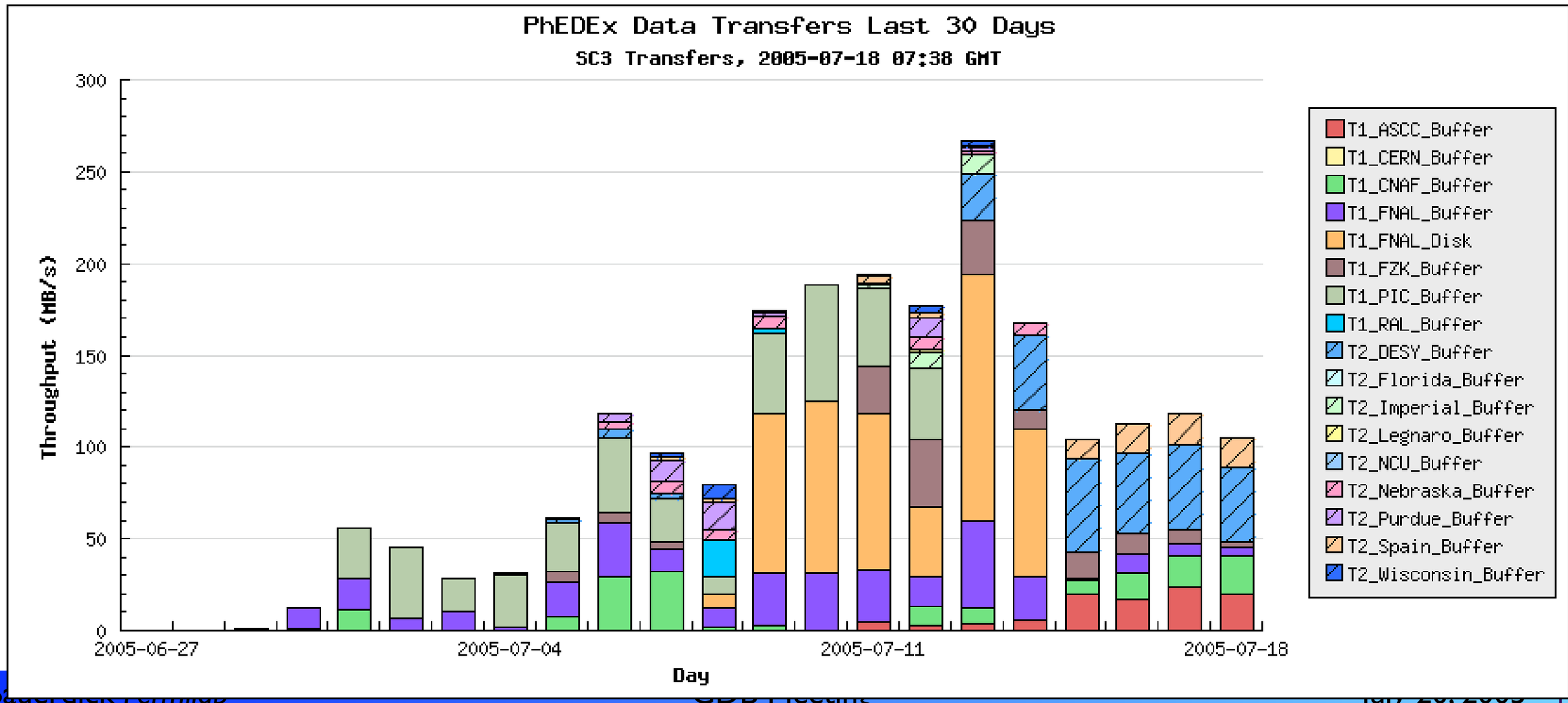
- Still aggregating feedback from sites
  - All this is very preliminary!
- FNAL, PIC reached good sustained rate through PhEDEx
- Some excellent progress on transfers to Tier-2 sites
  - 3 in U.S. (Purdue, Nebraska, Wisconsin)
  - 1 in U.K. (Imperial), Spain federated (CIEMAT / IFCA)
  - How about the rest?
- Minor PhEDEx improvements
  - New PHP-based plotting of transfer rate, pending transfers, transfer quality, based on JpGraph + LCG GridView examples
  - Improving timeout handling of transfer commands
- The full transfer chain (storage, PhEDEx, catalogues) seemed to work generally fine within the limits of what we know
  - Using "big" zipped files were good for everybody....

Achievements

- Very good sustained rate results to PIC, FNAL, DESY
- CNAF cooling crash, SRM, FTS
- RAL, FZK rates vary, timeouts
- ASCC a bit late

PIC network week





Issues

- Main issue: we did not yet really address the CMS goals
  - request to "go back and debug" basic parts of s/w stack
  - need to re-plan to assure CMS goals get addressed in time
  - Except for FNAL, PIC, impossible to conclude anything at this point
- Site installation documentation still lacking
  - Previously sticky issues have been addressed (e.g. PhEDEx deployment)
- most sites started with too short timeouts
  - There's much to improve, but tuning a site in days is not realistic
- downtimes, unavailability of tape services, etc problematic
- Mixed configurations (C1/2 IA64/32) at CERN end caused problems
  - only DESY used Castor-2 pools - learned little about "full load" -> continue
  - Only IA64 boxes ran SRM servers, cross-node RFIO to serve files
- Excessive timeouts (CNAF, RAL all transfers failing at some point!)
- Massive failure rates at CERN end – were these representative?
- Monitoring was unreliable, impossible to gather what was going on

Conclusions

- CMS is fully involved in SC3 and has important goals for the challenge
  - directly relevant for CMS computing integration CMS-LCG/EGEE-OSG
- Continued CMS SC3 test on production environment
  - thank you to sites for agreeing to continue that work in parallel!
- interest in high-throughput test of CMS dataset transfers in presence of other data transfers — "staged" to "service phase"
  - we realize debugging and setup phase needs be extended
  - need to schedule when SC3 is ready to be at the CMS "operation point" of implementing realistic dataflows b/w regional centers
  - require to achieve going beyond the file-level transfers ASAP
- concentrate on SC3 tests for specific configurations, using CMS Computing Integration Program to prepare for these
- require to concentrate on CMS operational point at least during the CMS parts of the challenge
  - still far off from what is a realistic scenario for CMS running
  - focus of service phase needs to be the experiment use case!