The CMS Data Grid

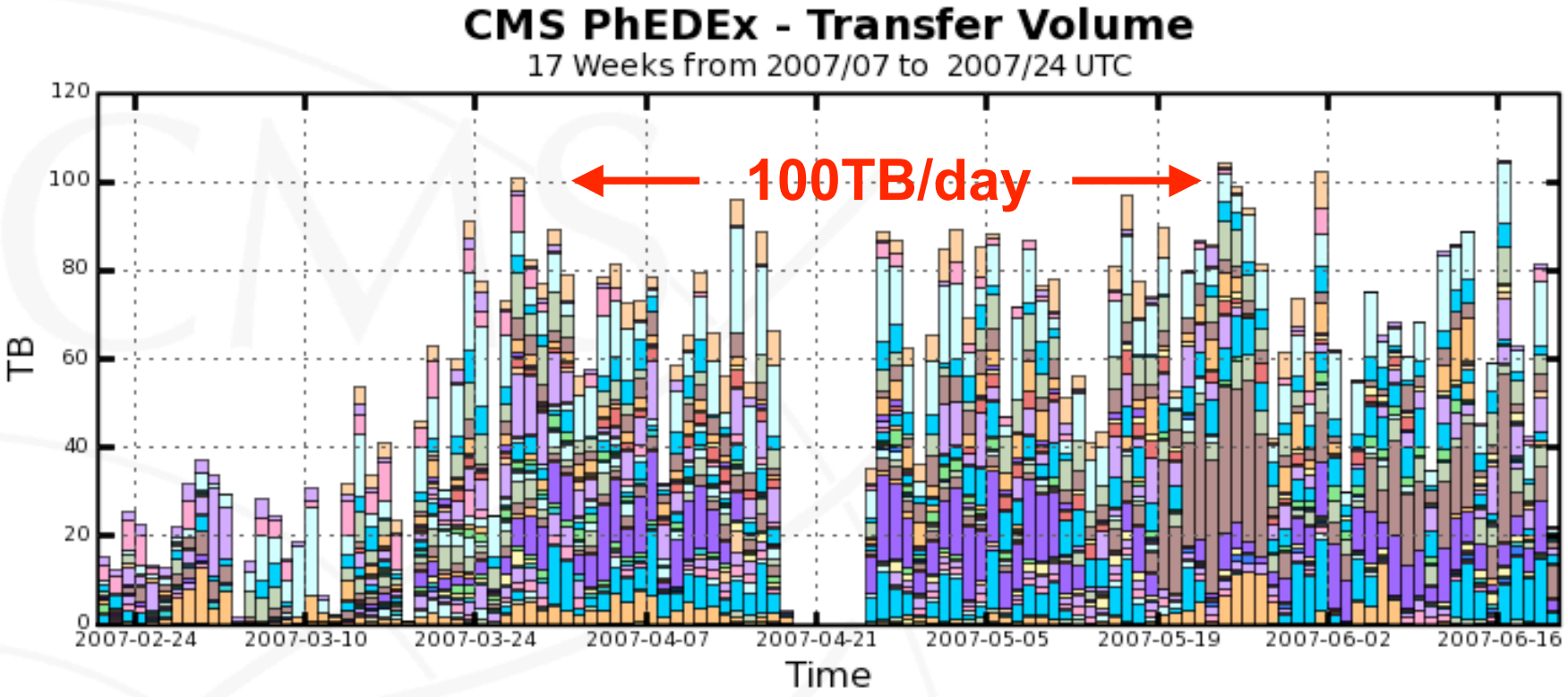
June 25th, 2007

Frank Würthwein (UCSD)

Disclaimer

- The objective of this talk is to provide a high level end-to-end overview.
- In the interest of time, I'll be oversimplifying, omitting, and generally ignoring details.
- My apologies for this in advance!

Setting the Scale



CMS routinely moves up to 100TB of data a day across its Data Grid of more than 50 sites worldwide.

The CMS Experiment

- A particle physics experiment built and operated by ~2000 physicists from 155 institutions in 37 countries.
- Data taking starting in 2008.
- Computing resources in 2008:
  - 34 Million SpecInt2000
  - 11 Petabyte of disk
  - 10 Petabyte of tape
- Distributed across ~25 countries in ~4 continents.

Today 30-50% of 2008 plan deployed and "operational"!

"Computing Model"

- Tier-0: Host of CMS @ CERN, Switzerland
  - Prompt reconstruction & *"back-up"* archive
- Tier-1: in 7 countries across 3 continents
  - Distributed *"life"* archive
  - All (re-)reconstruction & primary filtering for analysis @ Tier-2.
- Tier-2: ~50 clusters in ~25 countries
  - All simulation efforts
  - All physics analysis

CMS Data Grid

Centre	Streams	Associated T2
FZK	5	German T2, Poland, Switzerland
IN2P3	6	French T2, China, Belgium
PIC	2	Spain T2, Portugal
CNAF	7	INFN T2, Hungary
ASGC	5	Taipei, India, Pakistan
RAL	5	UK T2, Estonia, Finland
FNAL	20	US T2, Brazil
CERN		Russia, Ukraine

7 Tier-1 and ~50 Tier-2

All of different sizes and experiences.

“Data Organization”

- “event” ~ 1MByte
  - Atomic unit for purpose of science
- File ~ 1Gigabyte
  - Atomic unit for purpose of data catalogue
- Block of files ~ 1Terabyte
  - *Atomic unit for purpose of data transfer*
- Data volume per year ~ 1-10 Petabytes

A science dataset generally consists of many blocks with same provenance.

A science result generally requires analysis of multiple datasets.

Data Access Model

- Physicists develop custom executable based on CMS software framework.
- Analyze datasets to derive science result
  - Random access within an event
  - Sequential access within files/blocks/datasets
- Dataset Bookkeeping Service
  - Complete list of files->block->dataset, incl. (some) provenance info.
- Dataset Location Service
  - Complete list of location of all blocks.

⇒ **A complete block needs to be moved and registered before it can be analyzed at a T2.**

⇒ **Scientists need not care which T2 has which blocks.**

A word on sociology

- CMS physics community “clusters” according to strong ties for historic, geographic, or physics interest reasons.
- ***Technology shall not restrict sociology, nor policies of CMS.***
  - Some Groups “own” resources.
  - Some Groups are “assigned” resources.
  - Some Groups form (somewhat) dynamically from bottom up.
  - Some Groups are formed top down.
  - Primary data is available to everybody.
  - Derived data can be private for periods of time.
- Overall, there is healthy science competition within CMS, as well as with other experiments.

Layers for data transfer

CMS physics	Transfer request	*Any CMS member*
	Transfer accept	*Dest. Site decides*
CMS IT	PhEDEx.global	*List files, prioritize, decide source*
	PhEDEx.local	*Manage write into local SE*
IT	FTS	*Negotiate IO across VOs*
	SRM	*Storage Resource Management*
	Gftp	*Transfer protocol*
	Disk/tape	

Technological as well as Sociological Layering !!!

Strategies @ Storage Element

- Virtualization:
  - Separate physical and logical namespace
  - Separate request and open

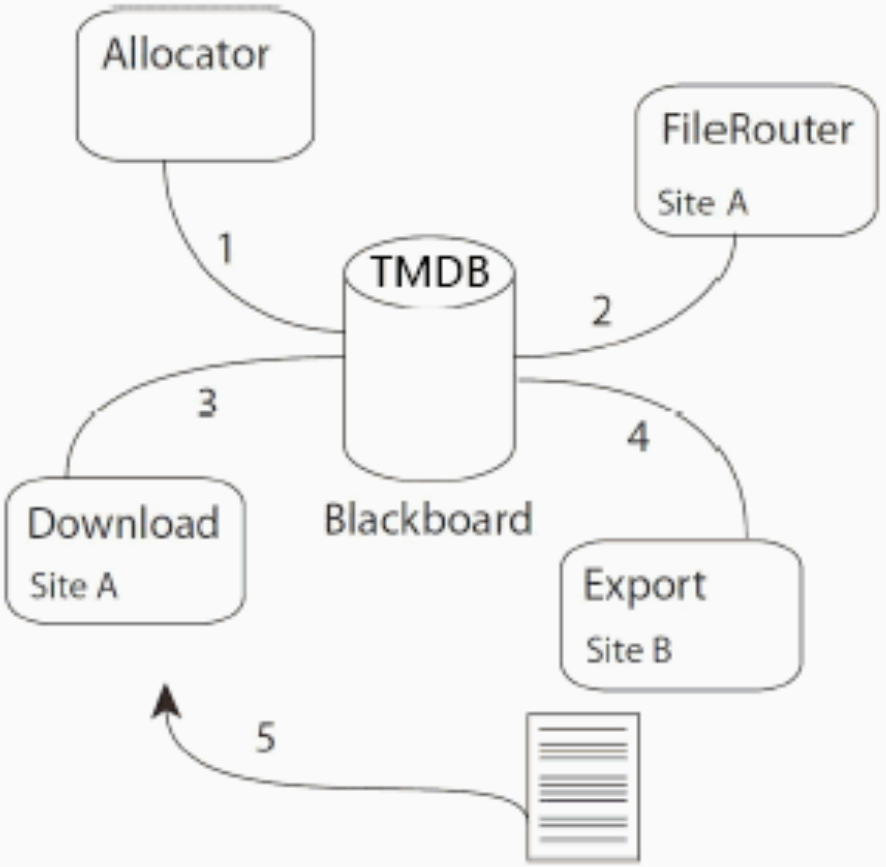
  ⇒ Replication for performance and availability
- Parallelization:
  - Apps. trivially parallel and generally CPU limited

  ⇒ Large (Gbytes/sec) aggregate IO via many (1000s) "slow" reads on LAN and streaming writes (1-10Gbps) from WAN.
- Simplify:

  ⇒ Closed files are imutable

  ⇒ No need for "cache coherence"

PhEDEx



1. Allocator: allocate files to destinations
2. FileRouter: determines closest replica
3. Download: marks files „wanted" from site B
4. Export: initiate staging and provide contact information
5. Download: transfer file

Distributed agents communicating via central "blackboard".

PhEDEx @ T2

- Each T2 runs 4 agents:
  - Download:
    - Manages the actual writes into local SE, including transfer verification and error handling.
  - Deletion
    - Manages the deletion in local SE
    - Deletions are handled via deletion requests, similar to transfers.
  - Registration
    - Watches for completion of blocks, and registers them.
  - Export
    - Controls which files are ready for read. At T1 this may involve file staging from tape.

Implementation of agents not necessarily the same at all sites!

Commissioning the CMS Data Grid

- T0 -> all T1s: 7 links
  - Considered part of the “near online” because files aren’t safe until archived at custodial storage.
- T1 <-> T1: 42 links
  - All 7 sites have a copy each of “AOD”.
    - AOD = “physics summary” ~ 50kB per event
  - Data exchange after T1s reprocesses its archival data.
- T1 -> all T2: ~350 links
  - T2s only cache blocks as needed. They thus all need to go to all T1s to get their data.
- T2 -> regional T1: ~50 links
  - Upload simulated data for archiving.

Roughly 500 links need to be validated and debugged.

Commissioning Challenges & Tools

- Challenges:
  - 500 combinations of sites need to be debugged, and kept functional.
  - All middleware is new and most IT shops have little experience doing operations of this kind at this scale.
- Tools:
  - PhEDEx heartbeat
    - Move small file across each and every link every 30 minutes.
  - PhEDEx loadtest
    - Sustained 24x7 data movement at low priority to measure transfer performance and stability over long periods of time.
  - Lot's of debugging by hand

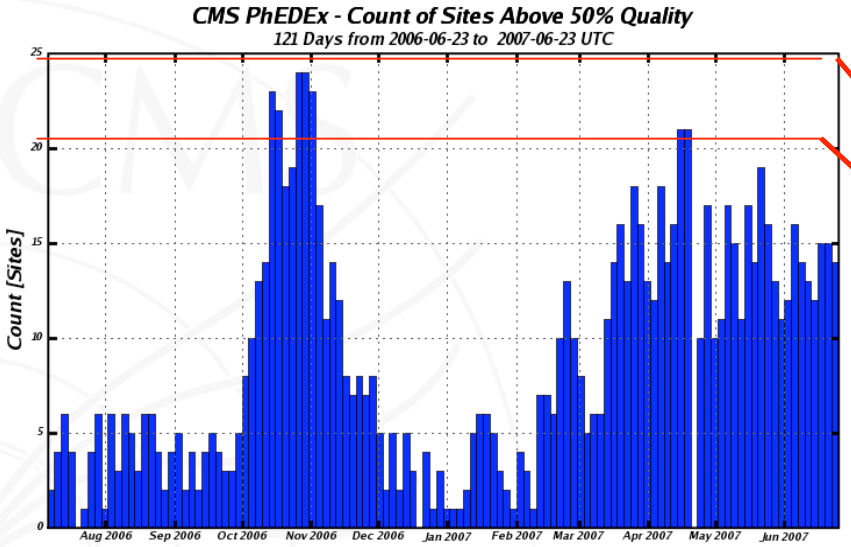
PhEDEx Loadtest

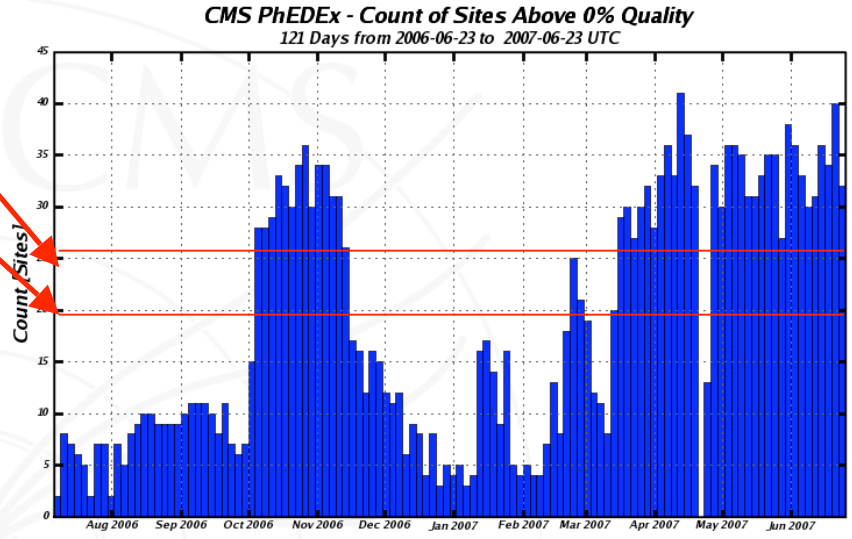
- Go through series of organized exercises.
- Exercises have targets driven by WLCG milestones for Q1-2/2007.

target kind	target (or eventual sub-targets)	period	ASGC	CNAF	FNAL	FZK	IN2P3 ?	PIC	RAL
single target	65% T0 -> T1 **peak** rate	1 week	17.1	23.9	68.3	17.1	20.5	6.8	17.1
simultaneous targets	50% T0 -> T1 **aver** rate	12 hrs	6.6	9.2	26.3	6.6	7.9	2.6	6.6
	50% T1->allT2 sum of aver. rate	12 hrs	38.0	47.0	124.0	31.5	48.0	26.5	42.0
	50% allT2->T1 sum of aver. rate	12 hrs	4.5	5.5	15.0	3.5	7.5	3.0	6.5
simultaneous target	T1 -> each T2 sustain	12 hrs	10	10	10	10	10	10	10
simultaneous target	each T2 -> T1 sustain	12 hrs	5	5	5	5	5	5	5

All targets are in MB/sec to be sustained for some period.

Participation & Success



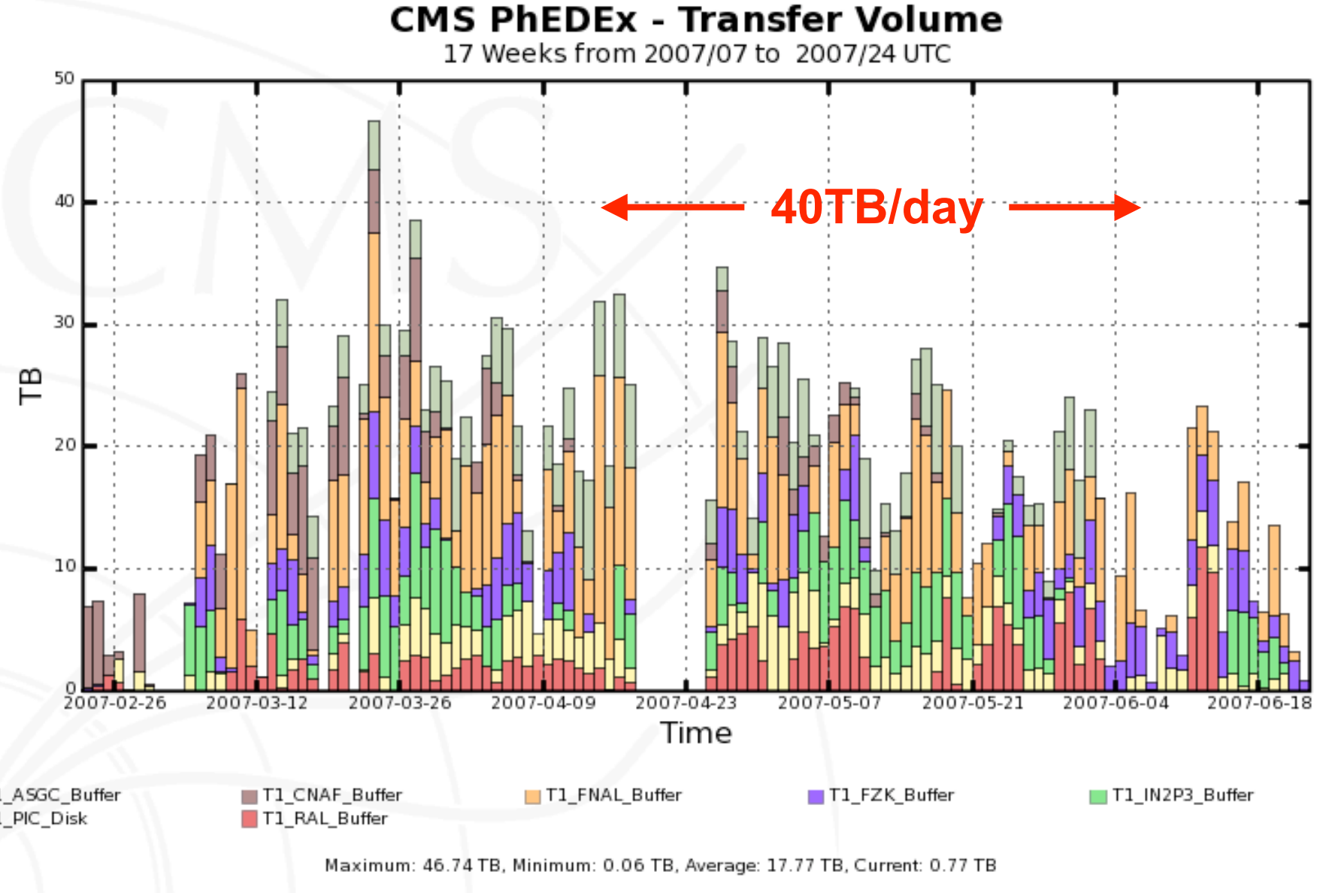


Only sites with more than 50% transfer success rate that day.

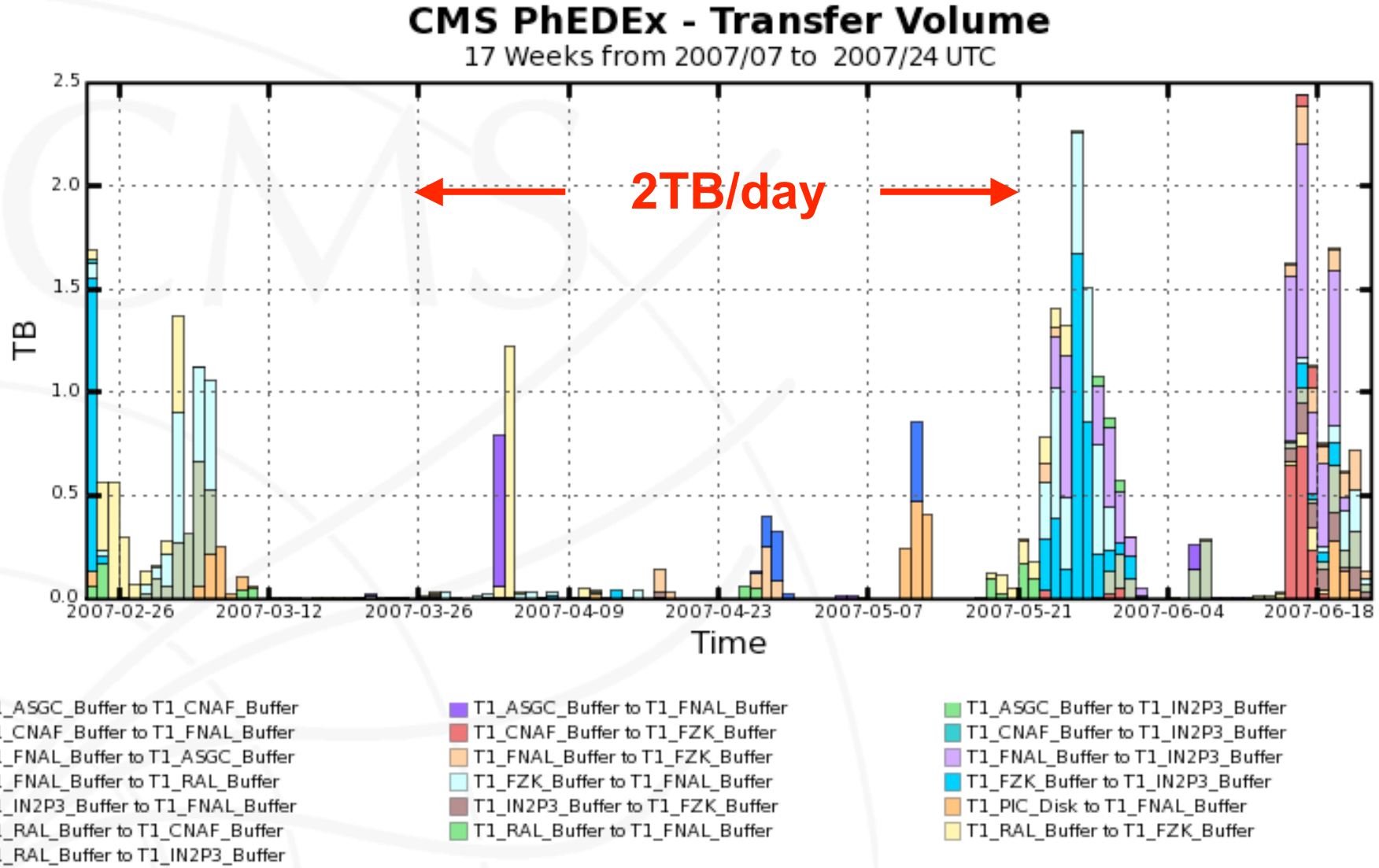
All sites participating in transfers that day.

Not all sites participate every day.
Not all sites are successful when they participate.

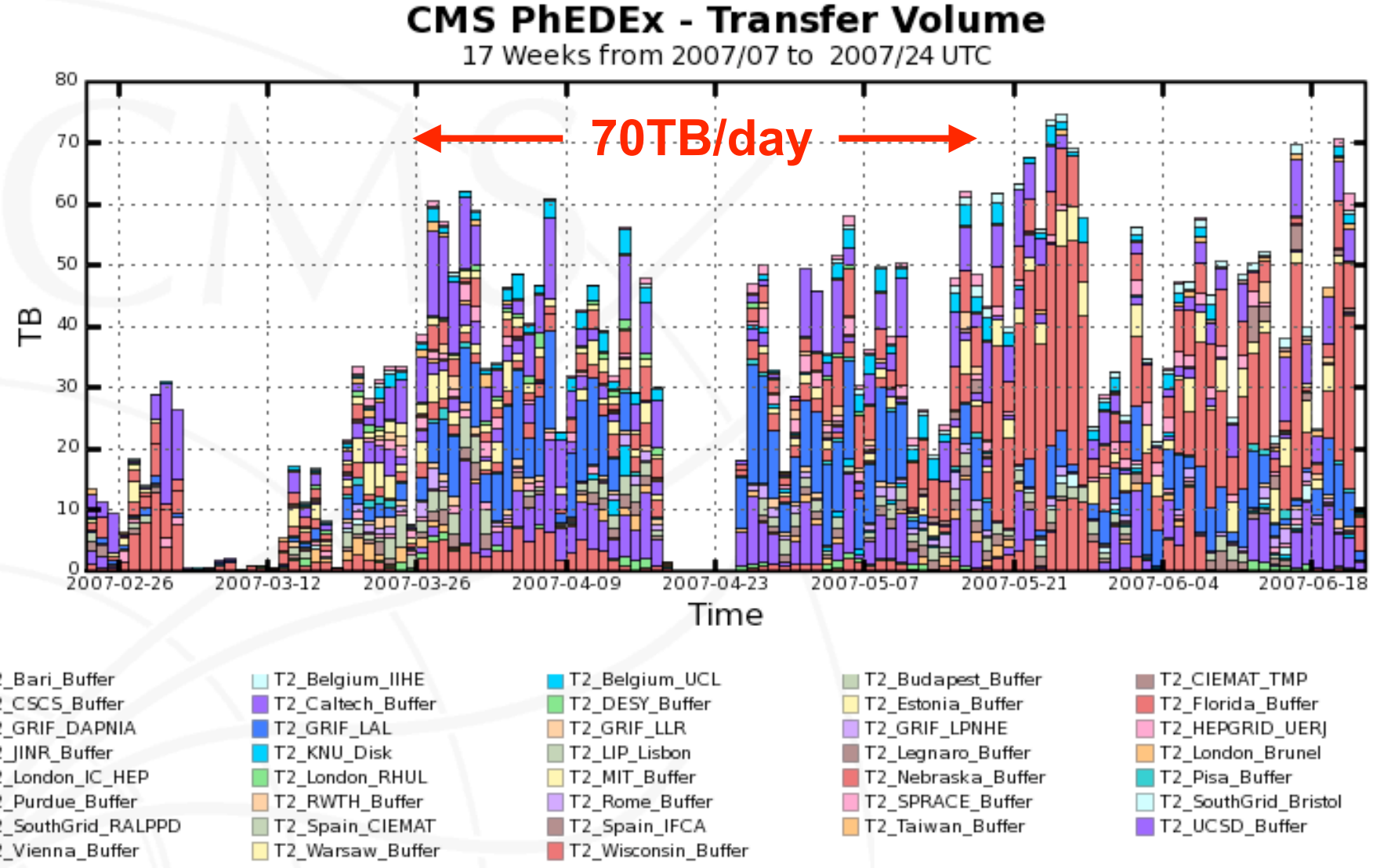
T0 -> T1 Transfers



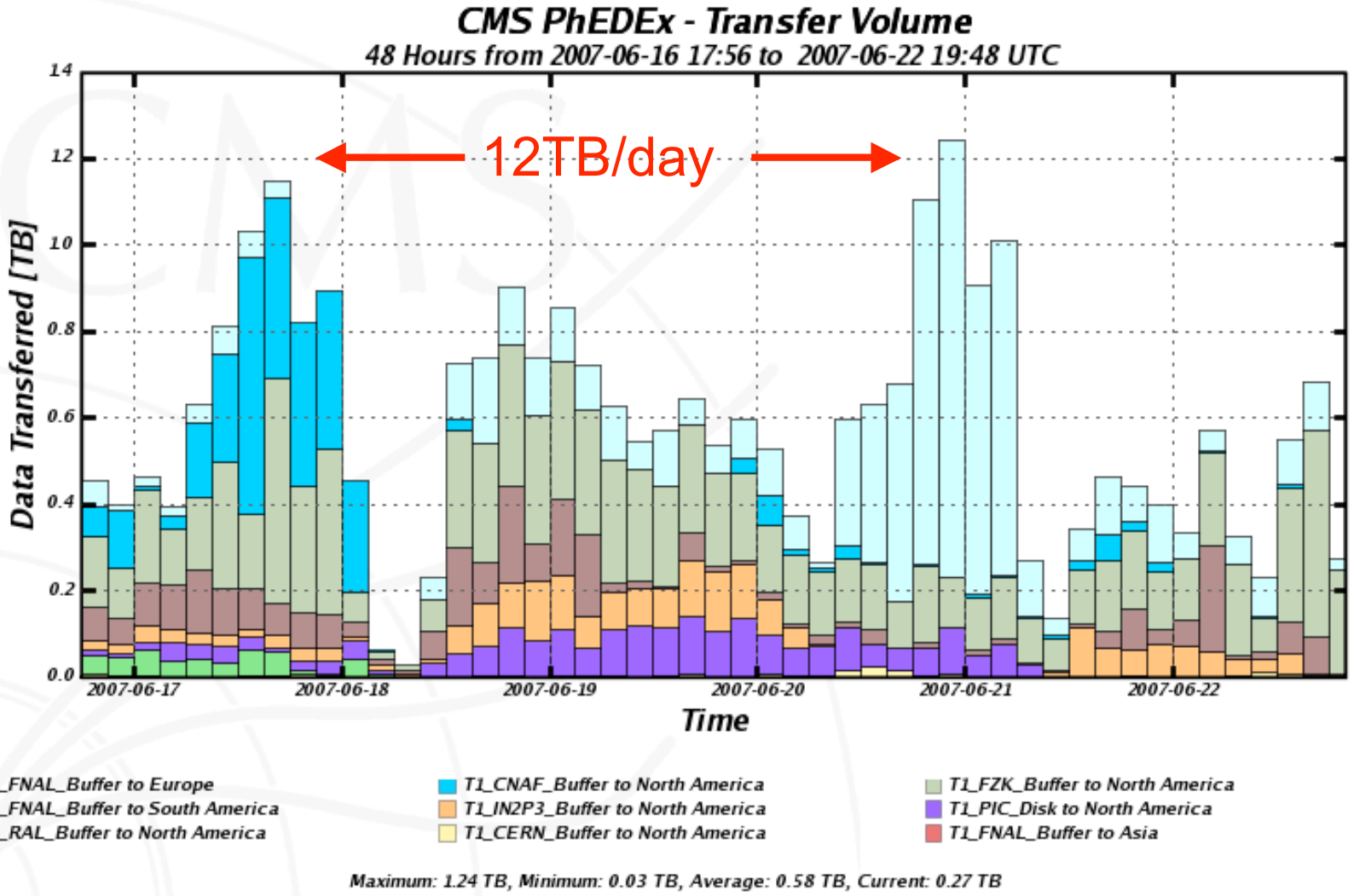
T1 to T1 transfers



T1 to T2 transfers



T1 -> T2 out of region



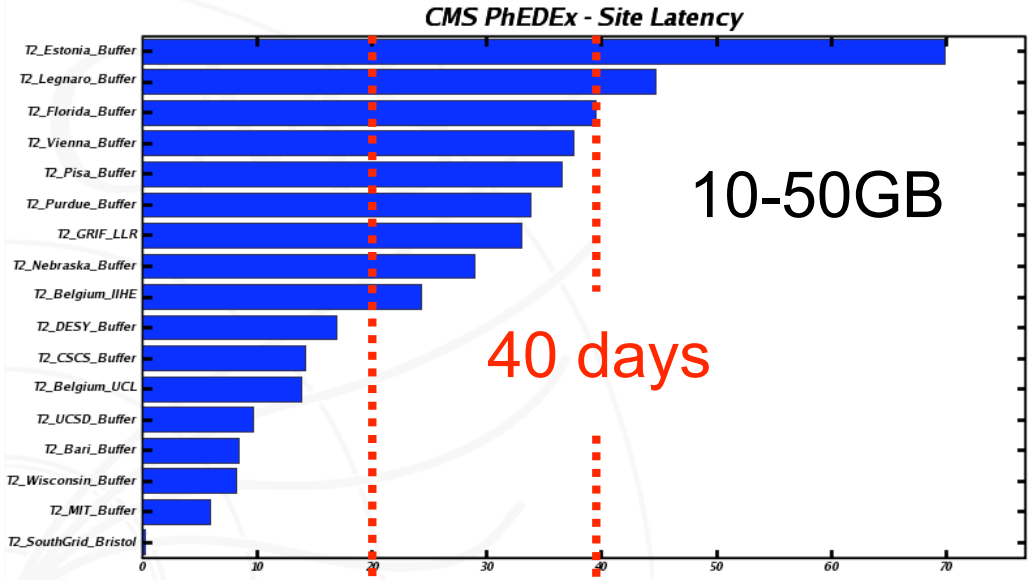
Roughly 15% of total T1 -> T2 transfers at peak.

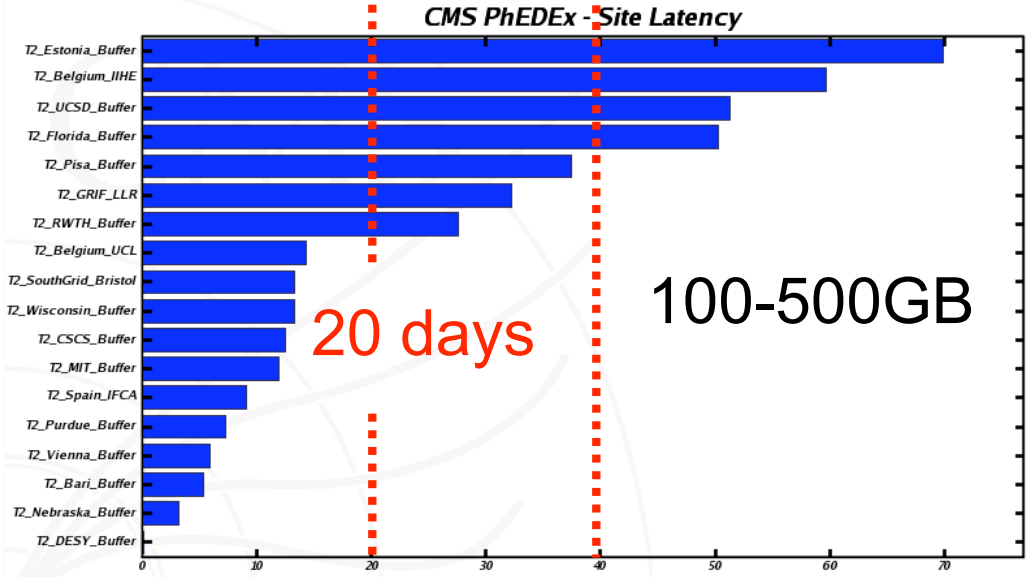
Loadtest Conclusion

- Impressive transfers between some sites, up to few tens of TB per day, often far exceeding reqs.
- *Overall, many of the targets have not been met yet*.
  - Steady performance for T0 -> T1.
  - Very variable for T1 -> T2 within region.
    - Some regions are superb (e.g. US).
    - Some links are excellent (also outside US).
    - Some links are pathetic because sites aren't yet ready.
  - T1 -> T1 exercises barely started.
  - T1 -> out of region T2 started, but lot's more to do.

A lot of work left to do !!!

Transfer latency





Average # of days to completely transfer a block of files, for blocks within a certain range of sizes. Averaged over all blocks transferred within 90 days.

Many of the places that are capable of sinking large rates nevertheless have significant trouble completing blocks, roughly independent of their size! Avg dominated by few blocks with large latency.

We have barely started to look into this issue.

Summary of Challenges

- Rapid deployment and growth of IT infrastructure across more than 50 institutions in 25 countries.
  - *Many people need to learn many new things!*
  - *A lot of strain on operations people.*
- A lot of “bleeding edge” middleware being deployed in a lot of places simultaneously.
  - *Significant stress on developers as we transition from development to operations.*
  - *Try to have developers -> integrators -> operators all be different sets of people to minimize strain.*

Summary & Conclusion

- CMS is operating global data transfer at the 100TB/day scale today.
- While this is a huge success, the details leave a lot to be desired.
- **It's all about deployment, integration, and operations at this point.**
- It is very easy to underestimate the human effort required to transition from where we are to where we need to go.
- ***And there's little time left!***

Afterthought
The next big challenge

1) cms wide data every user can read, but only a a privileged group is allowed to write.
2) Individual users have their own areas where they have write access, and only they have write access.
3) We want one user to read another users area.
4) We want groups (as in voms groups) that can share an area for writing/deleting, but at the same time have an audit or accounting trail to determine who (i.e. which fqan) wrote a file from within that group.
5) Quotas implementable for all spaces described above.