Service Challenge 3
CMS Recent Experience

Lassi A. Tuura
Northeastern University

Recent Related Experience

- CMS DC04: March-April 2004
  - Tier-0 reconstruction to Tier-1 for analysis
  - Real CMS application stack, files, …
  - No condition data flow
- SC1 (FNAL): December 2004
  - Tier-0 disk to Tier-1 disk (transfers/storage only)
  - Artificial files, no application stack
- SC2: Early 2005
  - Tier-0 disk to Tier-1 disk (transfers/storage only)
  - Artificial files, only FNAL used CMS transfer system (no file catalogue)
- CMS production file transfers: May 2004 - now

Distributed Computing Grid Experiences in CMS Data Challenge

A. Fanfani @ CHEP 2004

Dept of physics and INFN, Bologna
On behalf of CMS collaboration

Real Time Fake Analysis at PIC

José Hernández, CIEMAT

DC04 Post-mortem

Since DC04

- Changes to the project, framework, persistency
  - I will not cover most of these
  - SC3 is an integration test, not joint physics/software/computing
  - We are not under stress to make SC3 a "data challenge", so it's not
- Even more from central to local
  - Keep local what can be local
  - CMS people at site involved, does not depend on site admins
  - Site authorative source for what data it has (PubDB)
  - File PFNs are local, can (re)arrange as necessary
- More use of high-level concepts
  - Data placement, distribution, access in large units (datasets, blocks)
- PhEDEx nee TMDB fared well
  - Sound design evolved, significant development
- Lots of new work on enabling world-wide analysis
  - Allowing jobs to be submitted from anywhere to anywhere

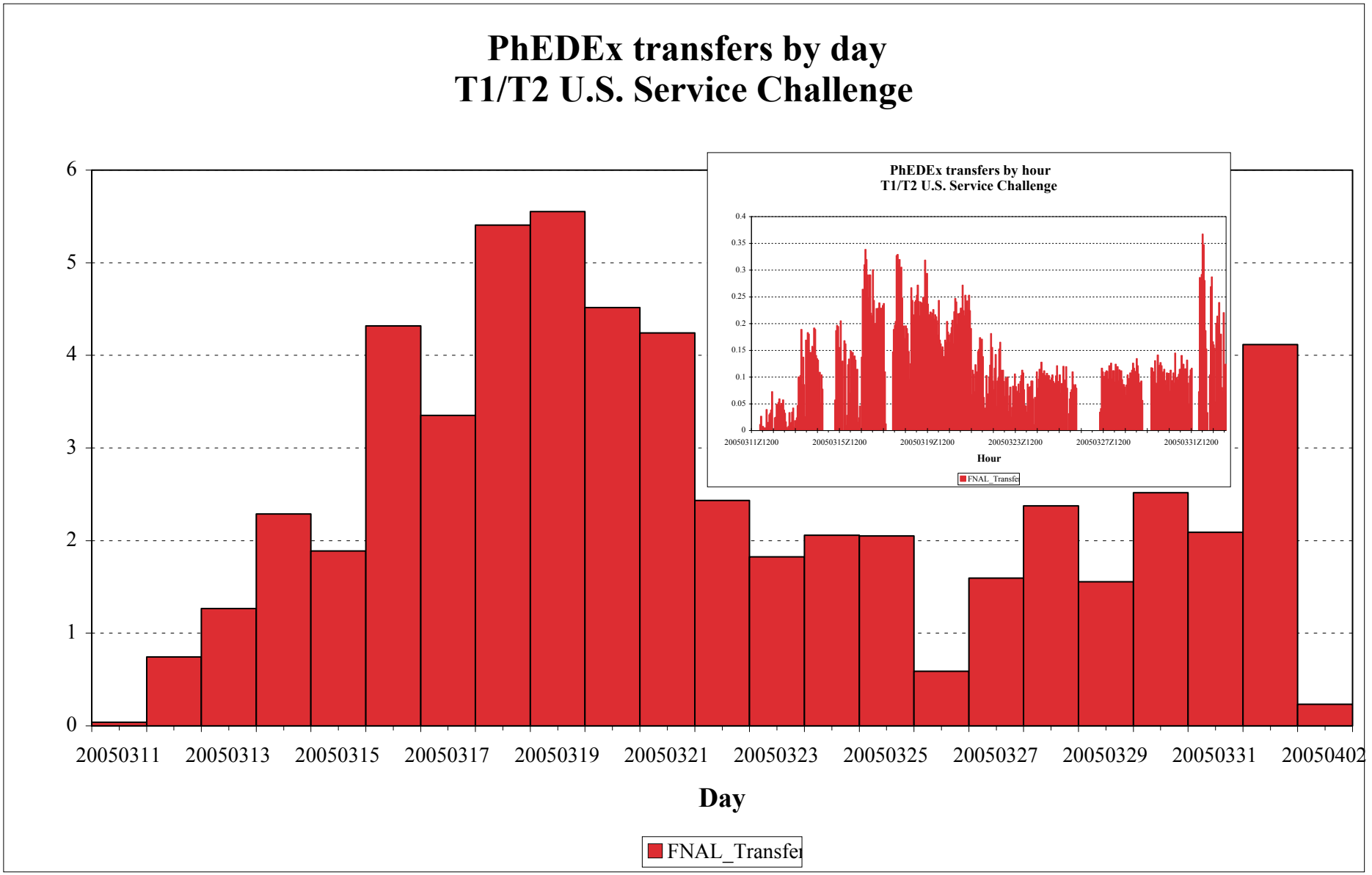
SC1 / FNAL

- Overview
  - Tier-0 disk to Tier-1 disk (transfers/storage only)
  - Artificial files, no application stack, no file catalogues
- FNAL portion of SC1 was carried out by FNAL/CMS storage, transfer experts
  - Using same underlying tools as used by PhEDEx
  - Reached the objectives
- Otherwise CMS as experiment did not participate in SC1
  - No CMS application stack involved
  - FNAL used production storage system
  - To my knowledge most others weren't
  - Files were artificial

SC2

- Overview
  - Tier-0 disk to Tier-1 disk (transfers/storage only)
  - Artificial files, no file catalogues
- Throughput objectives were met: very good
- Participation from CMS: some excellent, some limited
  - FNAL used PhEDEx for transfers and production storage system
  - Several U.S. Tier-2 sites involved in the same manner
    - Transfers were within production system however
  - A few other sites used production-type storage systems: good

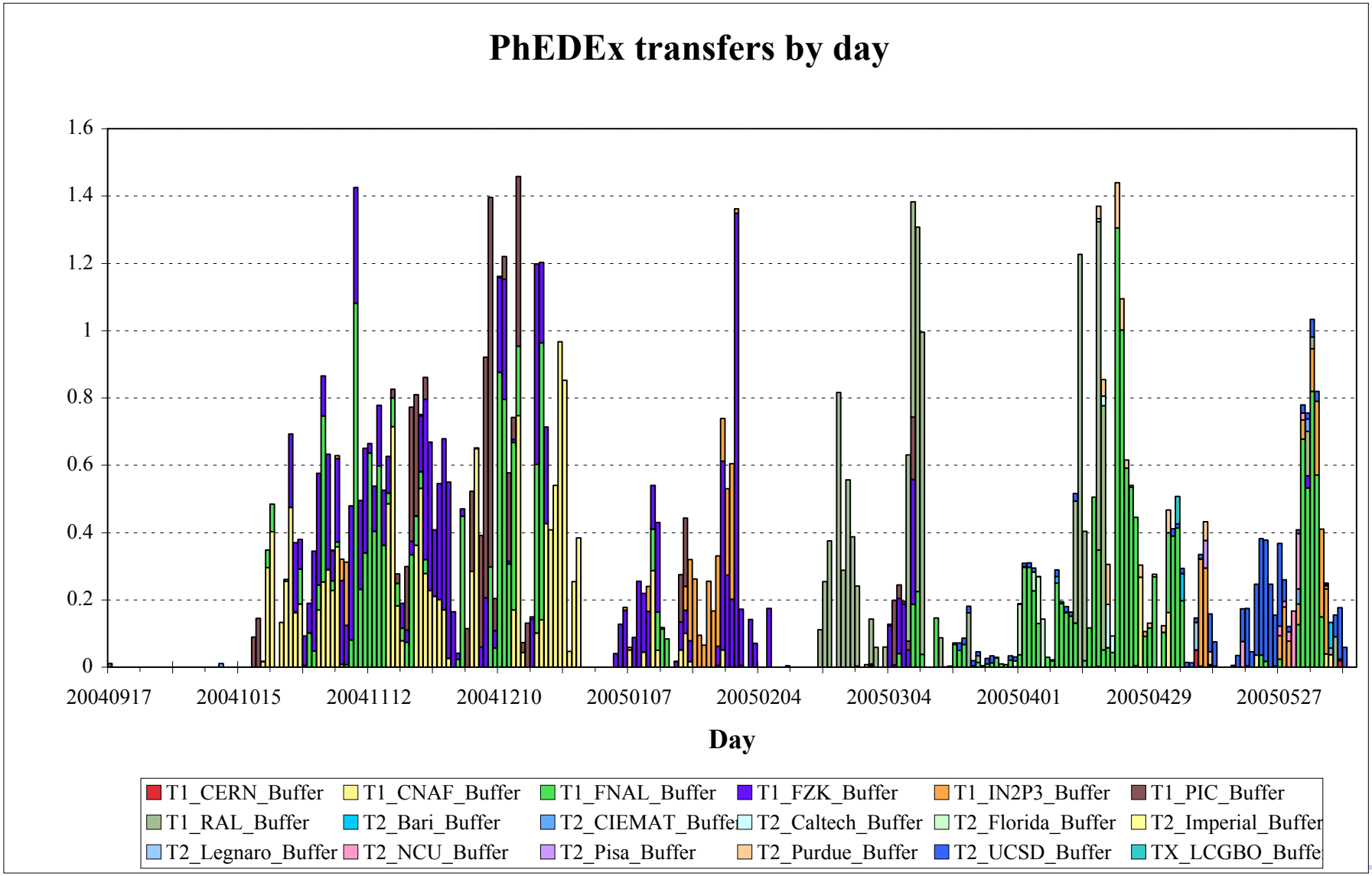
SC2 In Pictures



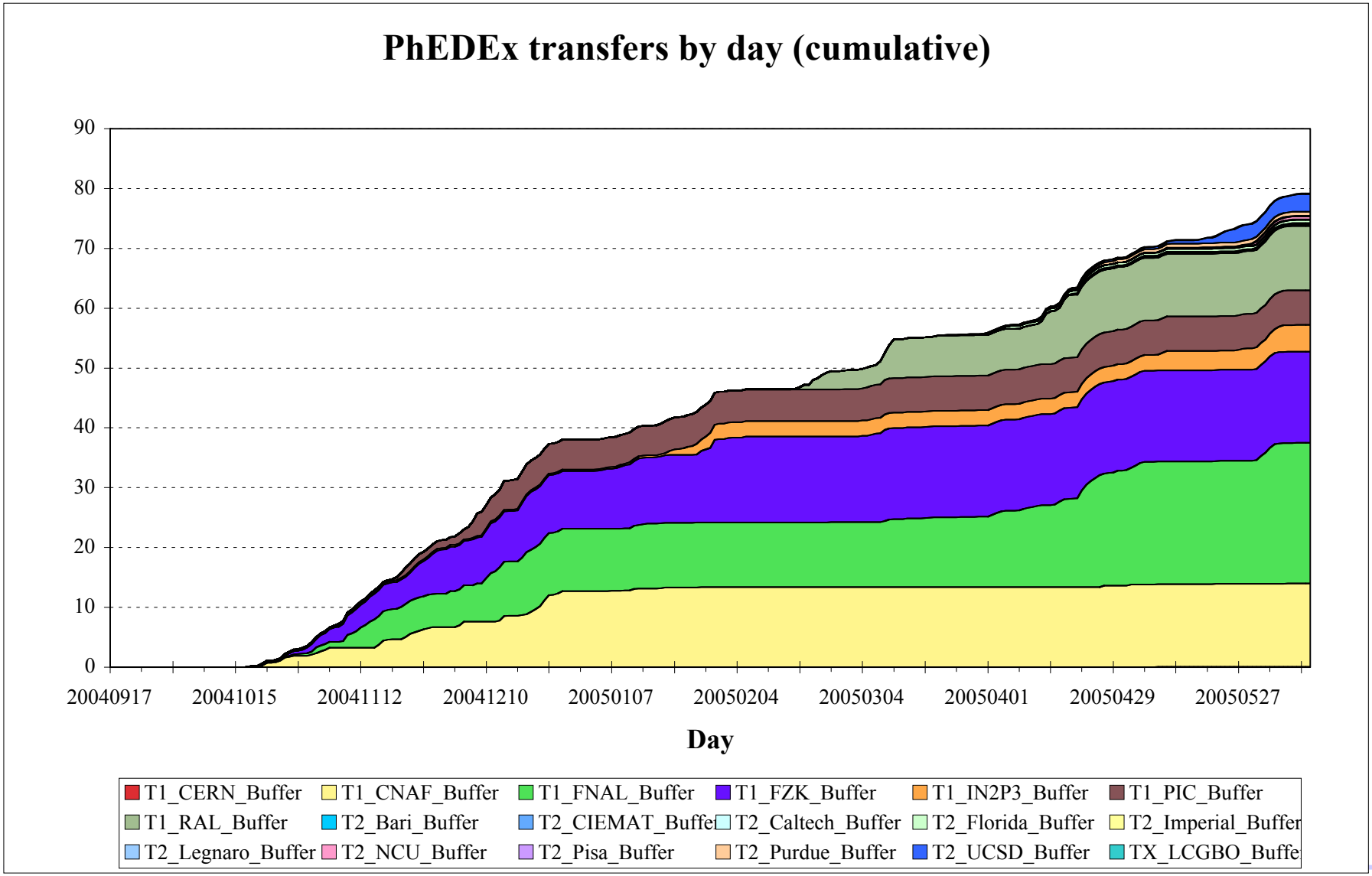
SC2 Observations

- One PhEDEx bug found and fixed, otherwise worked smoothly
  - Tough FNAL expressed strong desire for more and "visual" monitoring
  - Keeping in mind PhEDEx is not network bandwidth monitor
- PhEDEx installation instructions "difficult"
  - Especially installation of Oracle client: switched to instant client
  - Since then deployment significantly simplified: now four tools to run
  - Most of the problems with being "at the end of the stack"
    - Storage, certificates, myproxy, POOL, UI, … daunting list for a new site
- Very good rates reached
  - Very happy with FNAL results, U.S. Tier-2 progress
  - Apart from FNAL what was translated into production systems?
  - Can sites focus both on challenges and production support?
- Not sure what we learnt without experiment applications, files, …
  - Results very different from what we see in production
  - Difficult to say how much can be attributed to file sizes as claimed

CMS Production Transfers



CMS Production Transfers



CMS Production Transfers

- PhEDEx V2.1 overall status
  - ~105 TB known to PhEDEx, ~200 TB total replicated
  - CERN, 7 Tier-1s, 10 Tier-2s, 2 other (13 Tier-2s registered)
  - All Tier-1s operational for inbound transfers, several also export
  - Most Tier-2s doing inbound transfers, several installing
  - V2.2 about to go into production: migration ongoing
- Operational issues
  - Most sites able to keep agents up much of the time unattended
  - Data *cannot be exported from CERN* while production has priority
  - Transfer rate hiccups being looked at, file size implid in some
  - Next big step: getting all sites to export data
    - Plan exists for nearly all sites, but plan very often != solution
    - Exporting data from tape is difficult (in our experience)

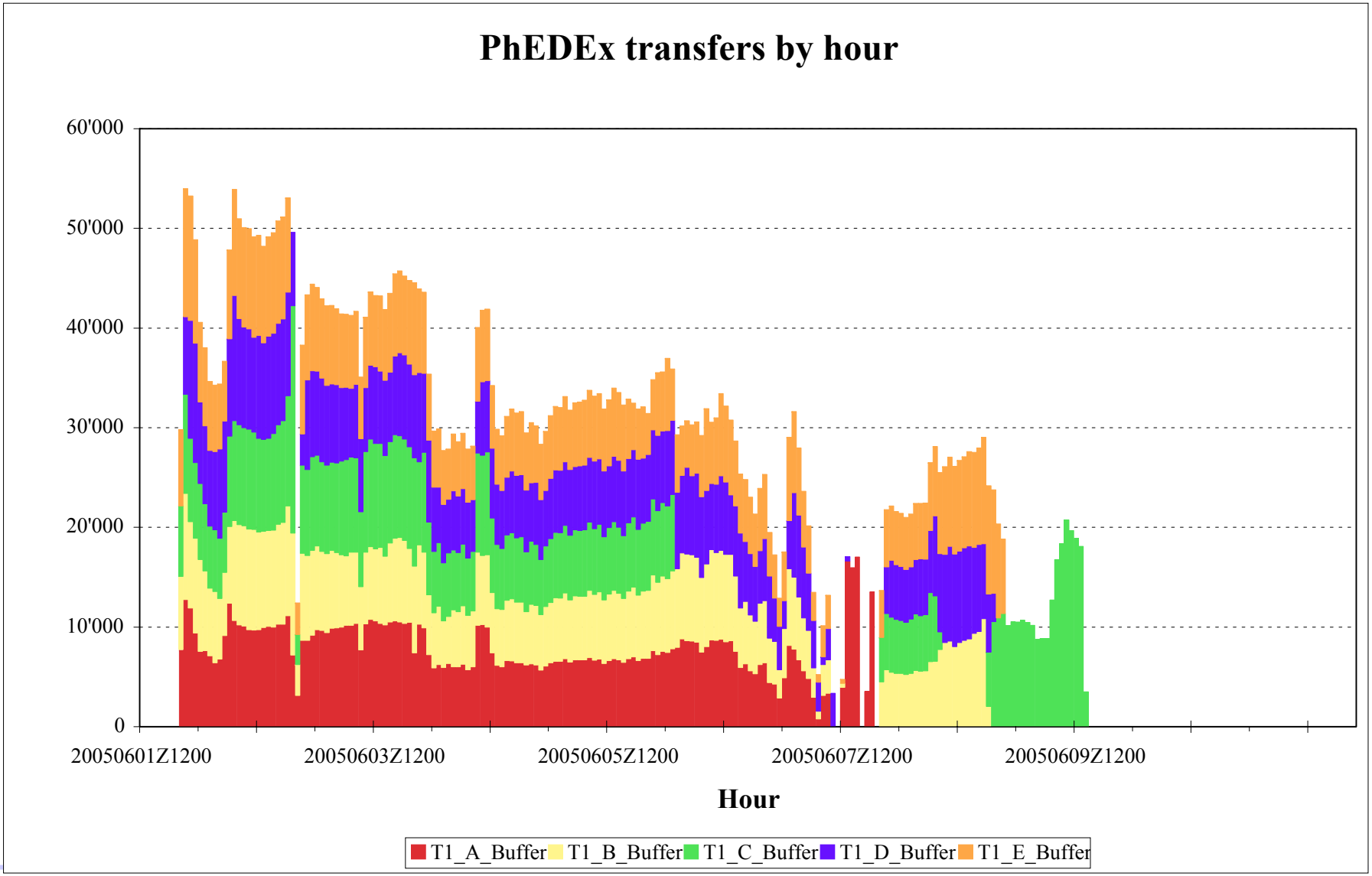
Production Transfer Highlights (I)

- Bandwidth rarely an issue
  - We don't have enough data to transfer to saturate current production networks for extended periods of time
- System stability frequently problematic
  - There are always "good" reasons…
  - We've rarely run much more than 24 hours smoothly
  - On average, about a third of the transfer network is down
    - Doesn't affect other nodes, but descriptive
  - Difficult to find out what's wrong
    - Complicated stacks of software, nobody master all of it
    - Put that monitoring into public, please!
- Popular complaint about CMS file size distribution
  - Implicated in many problems (tape, diretory sizes, stage-in)
  - However not enough being done to look beyond these issues
  - Being addressed: starting to merge files this month

Production Transfer Highlights (II)

- Once system has been configured for data import, transfers easy
  - Provided the underlying infrastructure holds together
  - Agents generally work reliably on the background
  - Generally ample bandwidth available from storage systems
- Exporting data efficiently, especially from tape, is ***difficult***
  - We have consistently had significant issues doing this from most sites
    - Generally takes a month or few to get it to work well
  - Lots of work put into an agent to export efficiently from current Castor
    - Used at CERN, PIC successfully
    - Variant exists for new Castor, partially tested
  - Partly hoping can do large transfer batches with srmcp
    - Let SRM worry about staging in files efficiently
    - Ignore advertising files currently available on disk (pretend all is)
    - For this to work, must be able to put hundreds of files per batch
    - There were issues with this in SC2 that must be addressed
  - Trying to do this as "cold start" in SC3 is scary…

PhEDEx V2.2 Scalability Test



PhEDEx V2.2 Scalability Test

- Can deliver ~30k files/hour to five destinations
  - Translates to up to 200k routing decision per hour
  - If CMS manages to increase files to planned 1.5+ GB and routing scales with number of nodes, can scale up to ~9 TB/hour/destination
- Performance figures
  - Can deal with O(6M) files concurrently in transfer
    - No progress or rate monitoring bottlenecks observed
    - File blocks seem to provide significant server load reduction
  - PhEDEx file transfer overhead well below <1s / file
    - Can be significantly reduced by batch transfers and SRM report
    - Biggest constraint is file catalogue operations
    - Transfers themselves not a limiting factor
  - File routing sets maximum possible rate
    - Much better than V2.1 (~x10), more than sufficient for now
- Further changes in V2.3
  - New dynamic routing in V2.3 opportunity for further improvements

General Observations

- General measurement is events delivered to physicists
  - "Good-quality papers submitted by CMS physicists on time"
  - Everything else is just subservient to that objective
  - Need to get the big stuff right before worrying too much about the smaller things
- Experiment focus has been shifting
  - Components to be used in 2007 pretty much now here
  - Main focus on integration and scaling, not components
    - Getting components to work, not find out what doesn't
  - Lot of complexity yet to be sorted out
- Tests that haven't happened
  - Dedicated RB, etc. tests
  - See comments by others, grumbling "on the field"