

Service Challenge

Report From CMS

LATBauerdick/Fermilab
for the CMS Computing Project

Some assessment of where CMS stands in using the WLCG

CMS participation in SC3

CMS goals for 2006 and SC4



CMS Computing at Grid Sites

- ♦ Seven Tier-1 centers are catering to CMS
 - ♦ IN2P3, GridKa, CNAF, PIC, ASGC, RAL, FNAL
- ♦ in total 28 Tier-2 sites have started to work with CMS
 - ♦ most of them listed in the WLCG-MoU
 - ♦ some eleven Tier-2s already working actively in Service Challenge
- ♦ CMS is actively working with sites
 - ♦ participation in CMS program, often through local CMS members
 - ♦ CMS site in service challenge
 - ♦ hosting of CMS data and of CMS analysis
 - ♦ Site Reports in weekly CMS computing integration/operations meeting
- ♦ CMS Computing Integration Program has done site survey:
 - ♦ <http://cmsdoc.cern.ch/cms/cpt/Computing/Integration/documents/T12survey.xls>

Computing Integration

Sites who moved data in SC3-Phasell (18)

- ASGC 5TB
- CNAF 8TB
- FNAL 24 TB
- GridKA 7 TB
- IN2P3 20TB
- PIC 13TB
- RAL 8TB
- Bari 3 TB
- Caltech 1 TB
- Desy 5 TB
- Florida 2 TB
- Legnaro 2 TB
- NCU 2 TB
- Nebraska 4 TB
- Purdue 2 TB
- Spain 2 TB
- SanDiego 6 TB
- Wisconsin 2 TB

Sites who publish data (12)

- From RefDB list of published production datasets
 - excluding SC3 data. More sites will join after SC3

- ASGC
- Bari
- CERN
- CNAF
- FNAL
- GridKa
- IN2P3
- Legnaro
- PIC
- SanDiego
- Florida
- Wisconsin

Sites who host analysis (15)

- Sites who have been targeted with more then 300 CRAB jobs in last month (includes SC3 jobs)

- FNAL
- CERN
- CNAF
- GridKa
- PIC
- Wisconsin
- SanDiego
- Desy
- Nebraska
- IN2P3
- RAL
- Purdue
- Belgium
- NCU
- ASGC



Many Problems reported by Sites

◆ Problems seen by CMS Sites

- ◆ “local” CMS production submission to LCG still too manpower intensive
 - ◆ new MC production system tries to address this
- ◆ fabric often still unreliable, requires to work rather defensively
 - ◆ RLS instabilities, sites downtimes, unreliable I/O, lack of monitoring
 - ◆ data transfers fail for a hosts of reasons, storage interfaces “fragile”
 - ◆ many and diverse problems, but rarely “final” solutions – often just “recipes”
- ◆ CMS application of pile-up generation still too “computing intensive”
 - ◆ “resilient dCache” storage element seems to be winner
- ◆ CMS dataset “publication” too complex and error-prone
 - ◆ new EDM and Data Management system will not have that problem

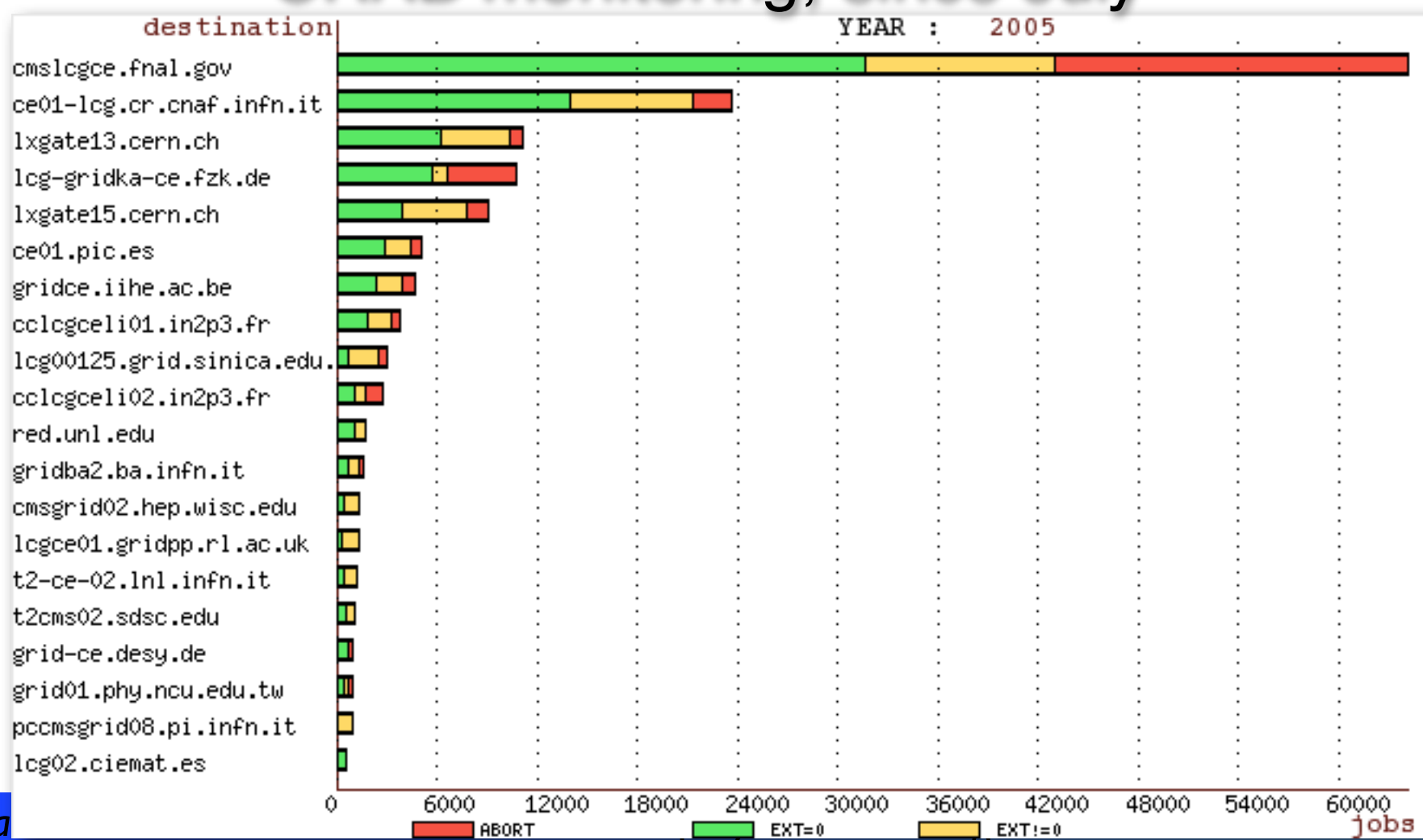
◆ As seen from CMS Computing:

- ◆ most sites are only starting to exercise and demonstrate capability to serve analysis data for user access at required scale and reliability
- ◆ integration program started to stress sites to shake out problems, need more help from sites to follow through on problems

Some Successes

- ◆ **WLCG system is already providing analysis capabilities to CMS!**
 - ◆ many ten thousands of Physics-TDR analysis jobs run successfully
 - ◆ Tier-1 centers gain operational experience (not yet so much Tier-2s)
 - ◆ subset of really reliable and responsive sites used in pTDR production
 - ◆ good fraction of Tier-2s acting fast to become analysis sites for CMS

CRAB monitoring, since July



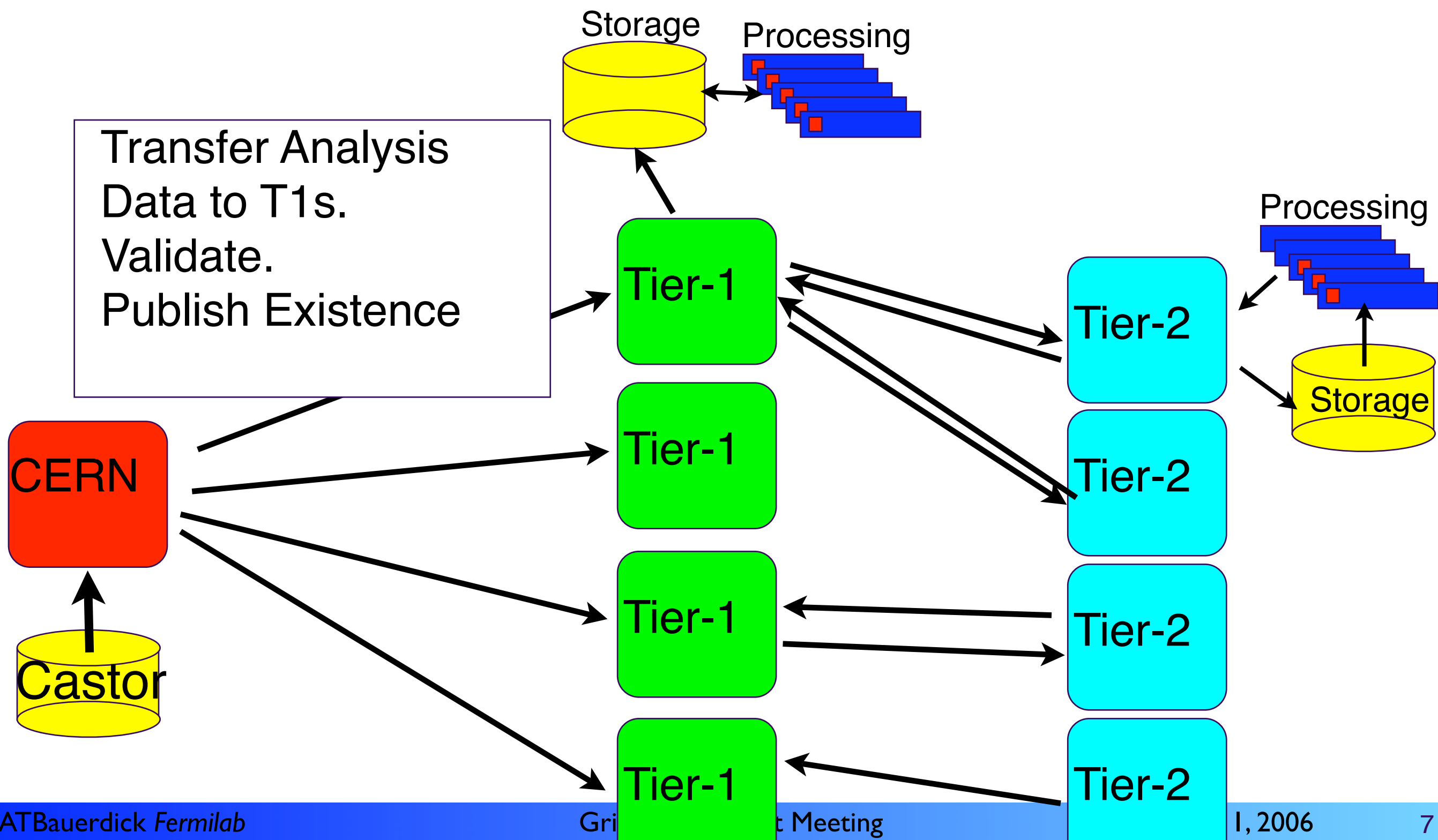


Service Challenge 3

- ◆ SC3 was a major driver of CMS Computing Integration work
 - ◆ “data challenge” driven by WLCG to establish the WLCG “service”
- ◆ in addition we had important SC3 goals for CMS
 - ◆ integration test of next production-level system
 - ◆ establish the major use scenario of data transfer and data serving
 - ◆ test infrastructure for the realistic end-to-end use cases
 - ◆ in particular T0->T1->T2, with tape-tape, tape-in tape-out etc
 - ◆ Including testing the workload management components: the resource broker and computing elements
 - ◆ Bulk data processing mode of operation
 - ◆ come out of SC3 with functional system with room to scale up
 - ◆ major participation from all CMS T1s and CMS T2s
- ◆ Crucial step toward SC4, CMS CSA2006 and LHC running

SC3 CMS Operation

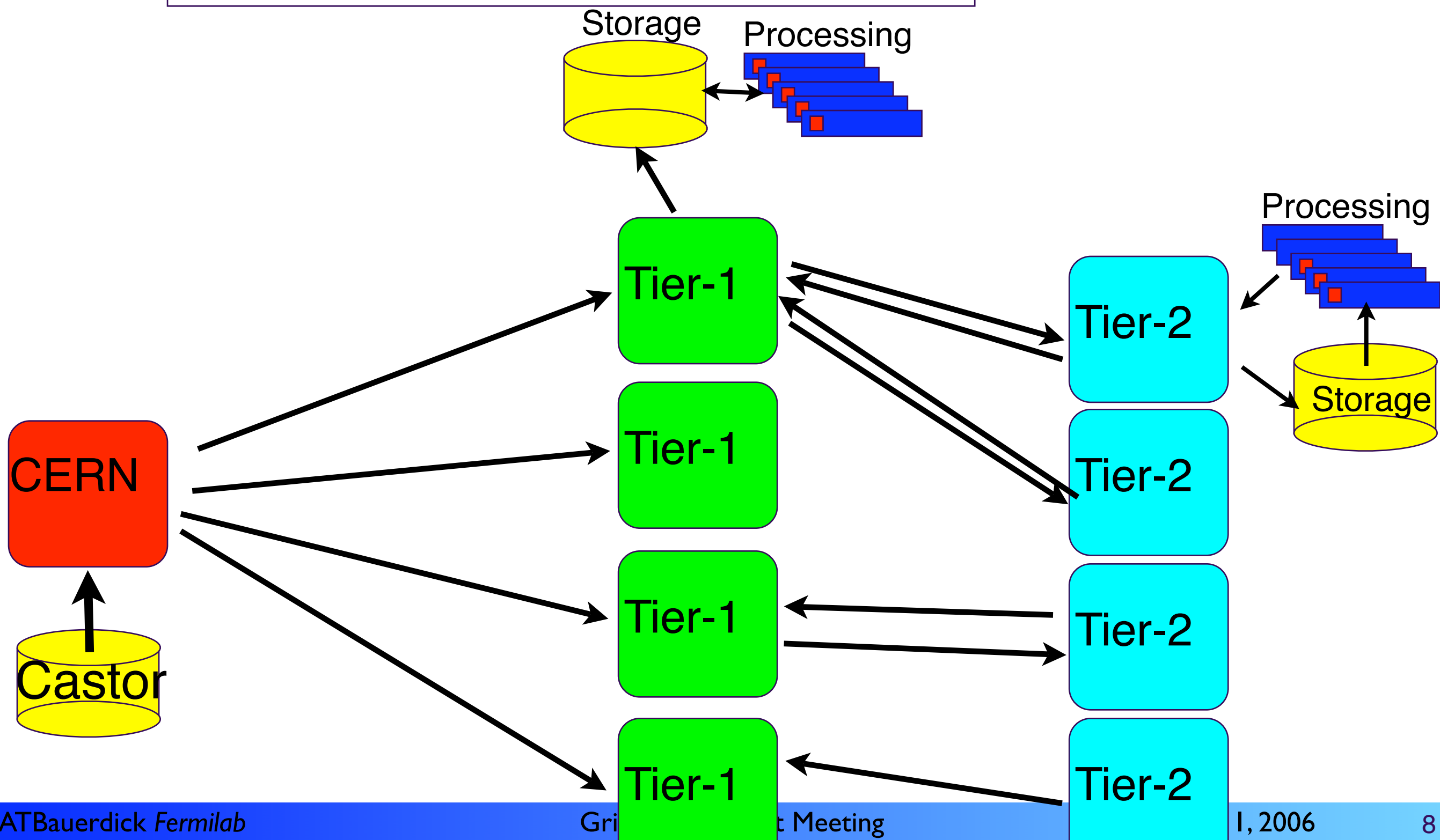
- ◆ Test the basic elements of each Computing Tier simultaneously





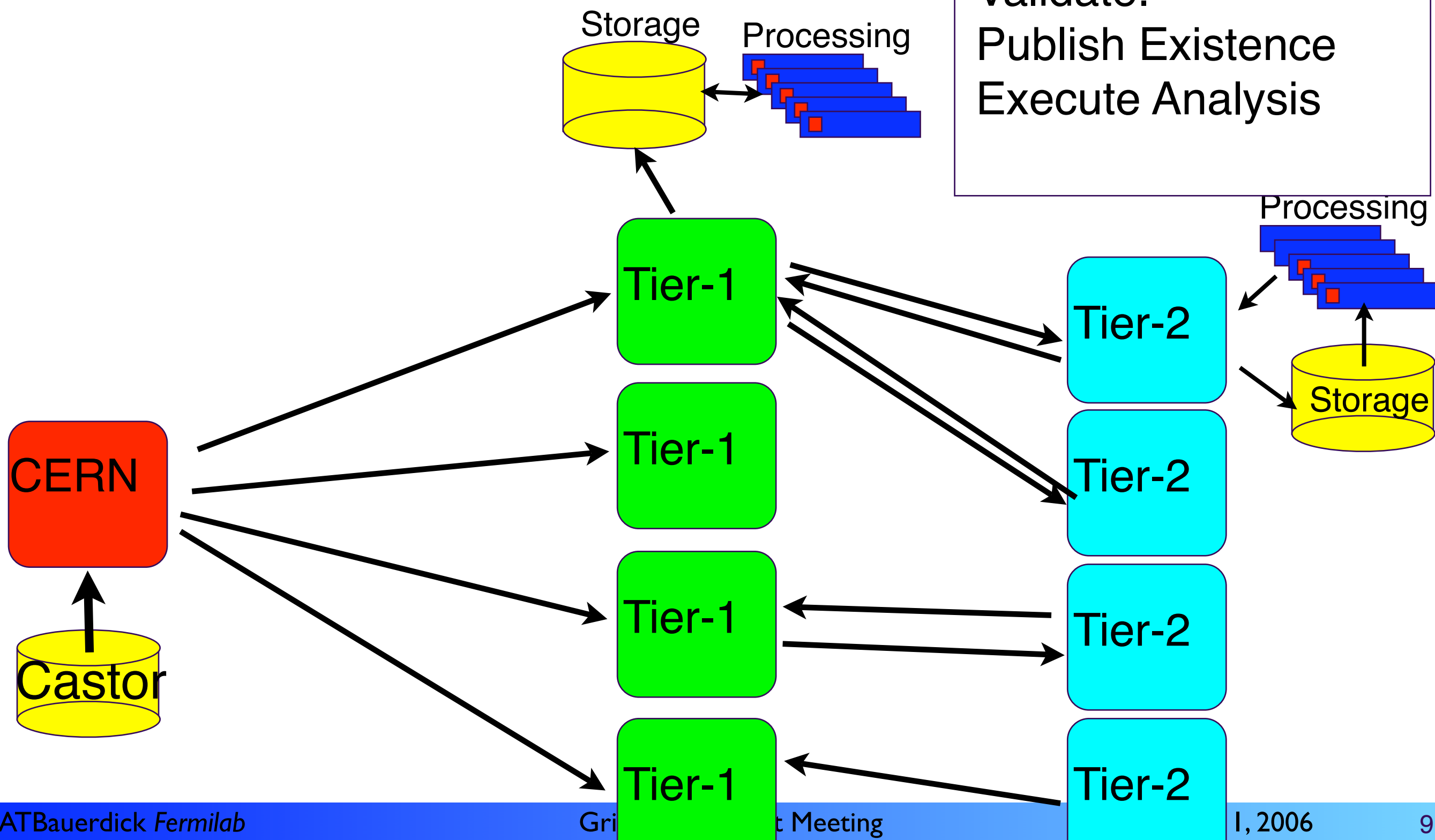
SC3 CMS Operation

Execute Analysis Jobs on Tier-1
Resources. Select Data

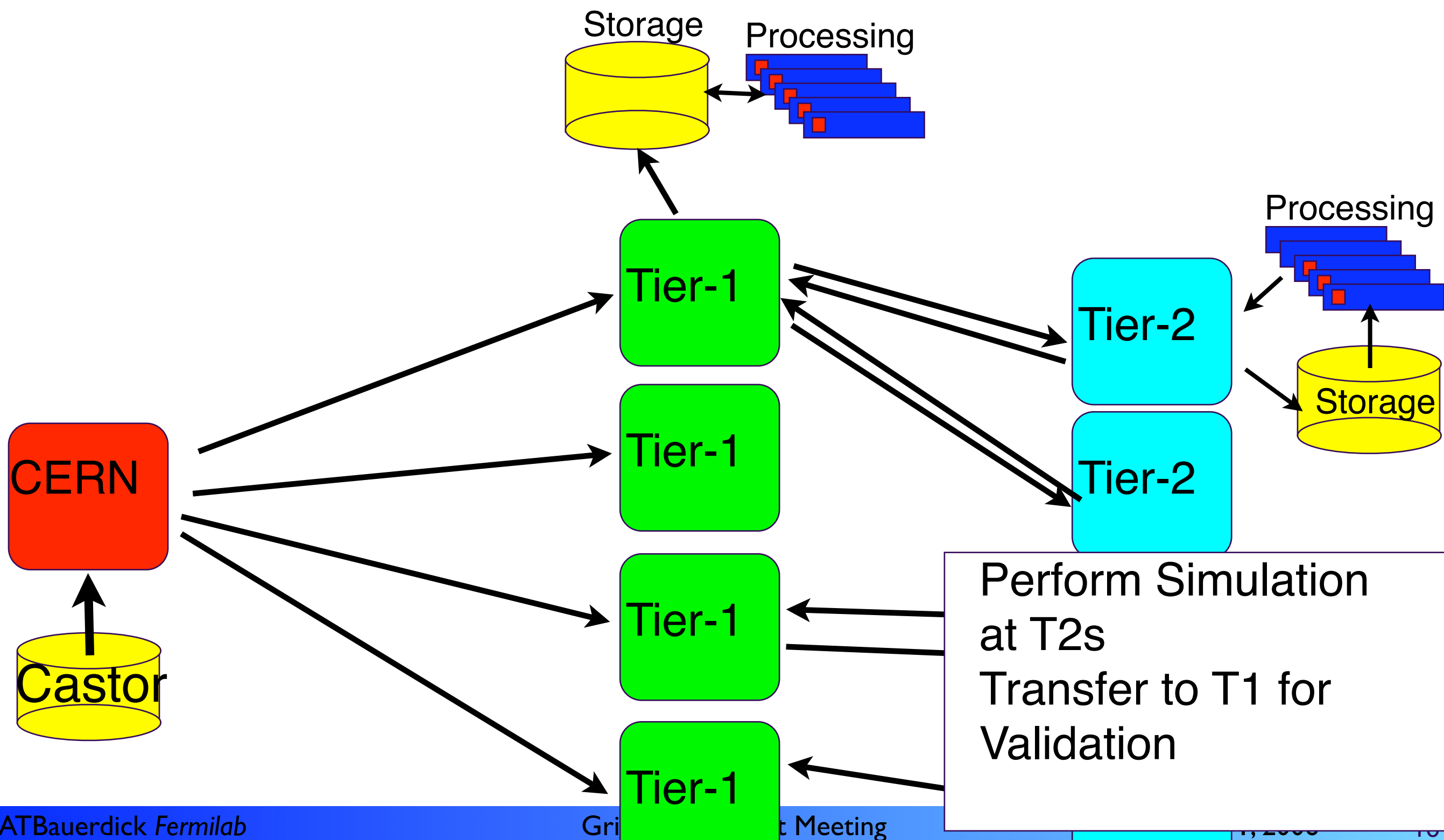


SC3 CMS Operation

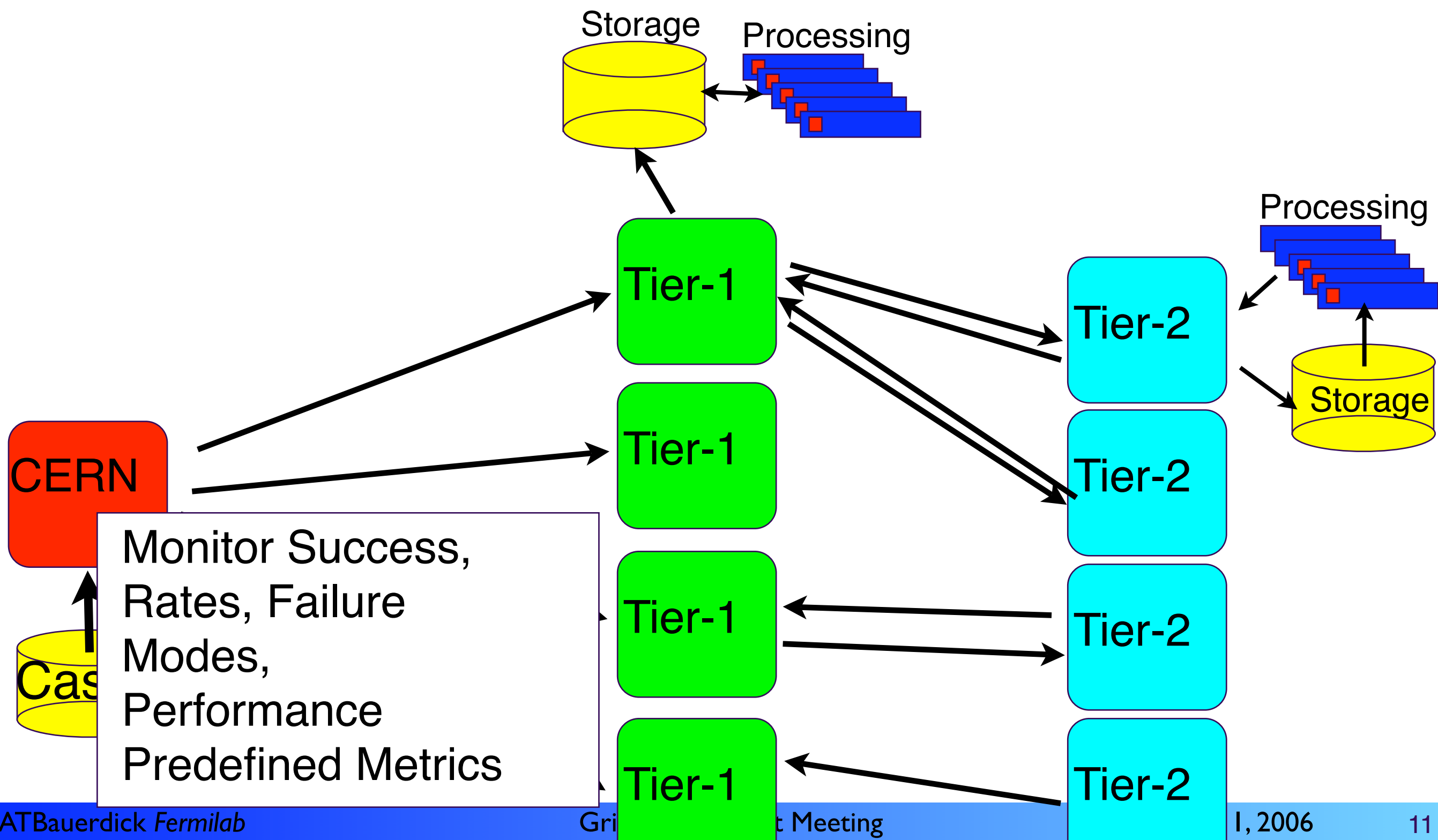
Transfer Analysis
Data to T2s.
Validate.
Publish Existence
Execute Analysis



SC3 CMS Operation



SC3 CMS Operation





SC3 CMS Operations

◆ CMS central operations

- ◆ Dataset placement and transfers, entirely managed with PhEDEx system
 - ◆ Central transfer management database at CERN
 - ◆ Using underlying grid protocols, srmcp, globus-url-copy, eventually gLite FTS
 - ◆ Placing files through SRM on site storage based on Castor, dCache, DPM
- ◆ CMS analysis jobs, submitted by job robot based on CMS CRAB tool
 - ◆ Using 3 LCG resource brokers and OSG Condor-G interfaces
- ◆ Monitoring, info gathered centrally: MonALISA and CMS Dashboard
 - ◆ Data fed from RGMA, MonALISA and site monitoring infrastructure

◆ Per-site responsibilities, done through CMS people at or near site

- ◆ see “White Paper” at <https://uimon.cern.ch/twiki/bin/view/CMS/CmsServicesAtTier1>
- ◆ Ensuring site mass storage and interfaces function, grid interfaces respond, data publishing steps and job data access succeed
 - ◆ Data publishing, discovery: RefDB, PubDB, GLIDE, ValidationTools
 - ◆ Site local file catalogues: POOL MySQL, POOL XML
- ◆ A lot of infrastructure tools provided to the sites, but having the whole chain hang together requires perseverance



SC3 Results for CMS

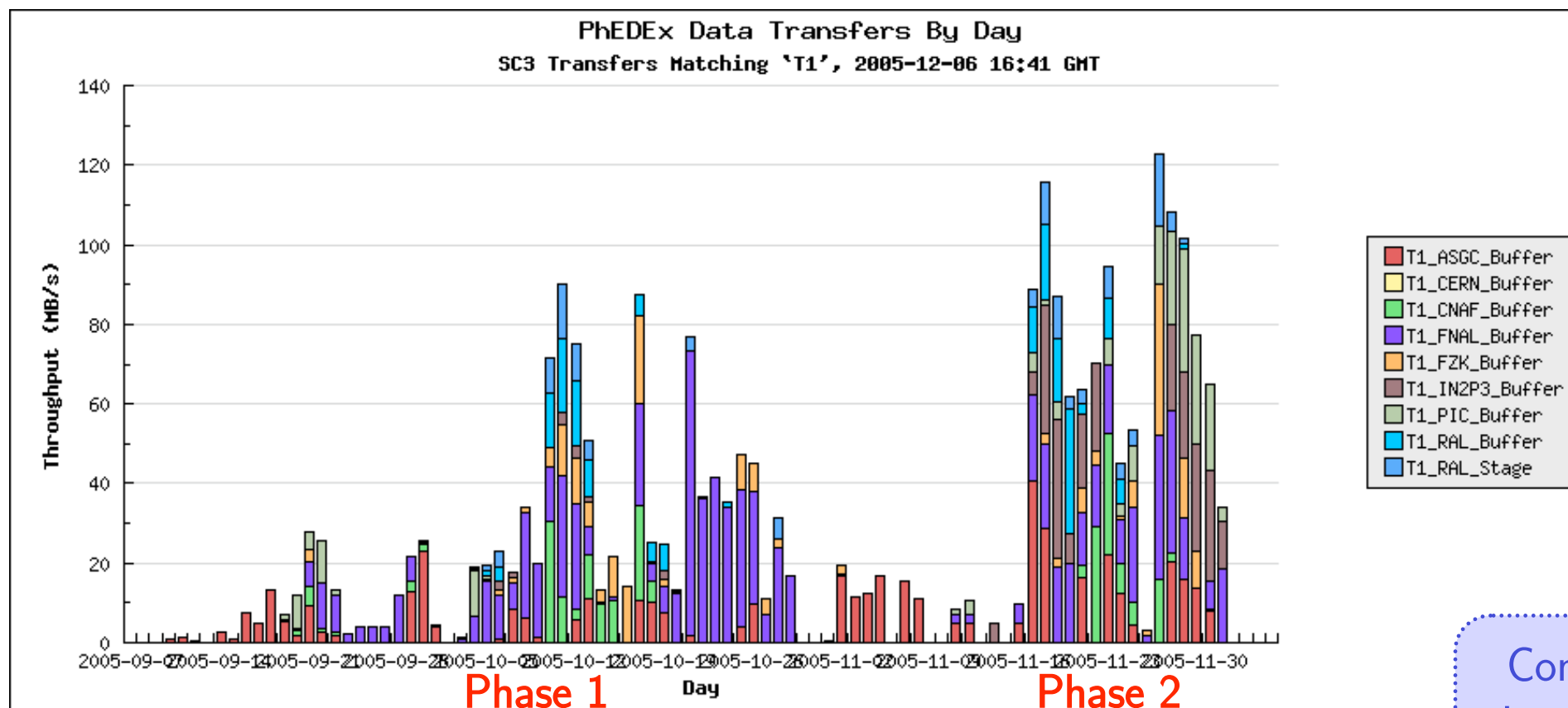
◆ Data Transfers

- ◆ all transfers orchestrated by PhEDEx, interfacing to grid middleware
- ◆ total volume transferred (all SC3) ~0.3 PB
- ◆ in the 2 wks of SC3 Phase2 140 TB, ~ as much as in preceding year!
- ◆ achieved average data rates, end-to-end, of $O(<20\text{MB/sec})$
 - ◆ NB: this is 3 orders of magnitude less than current networking record

◆ Job running

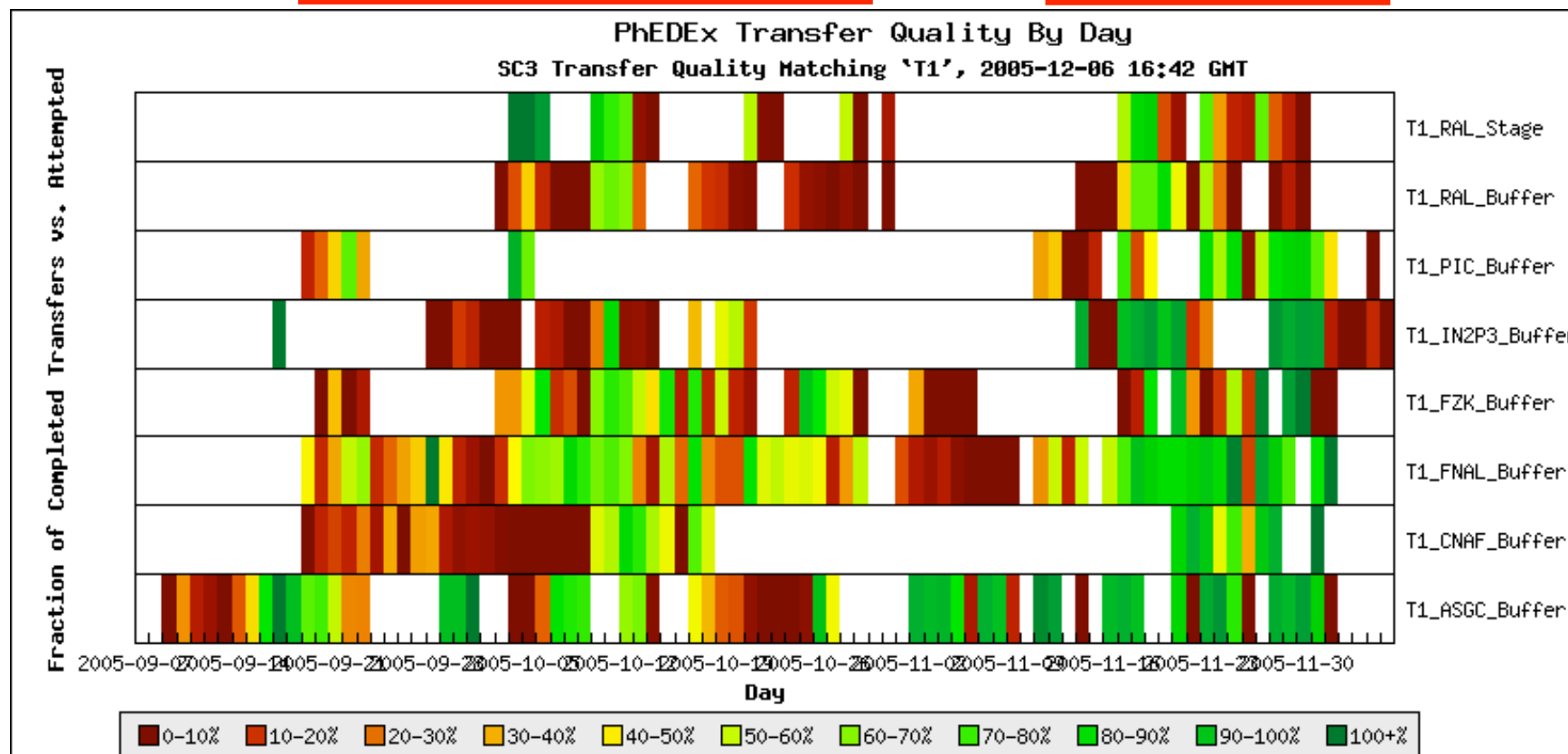
- ◆ orchestrated through job submission “robot”
using CRAB to EGEE/LCG and OSG sites
- ◆ in 2 wks of SC3 Phase2, 32000 jobs running on 38M events
- ◆ not yet comprehensive jobs statistics
estimate ~2/3 of jobs run to completion

SC3 Data Transfers to Tier-1s



Tier-1 WAN Transfers

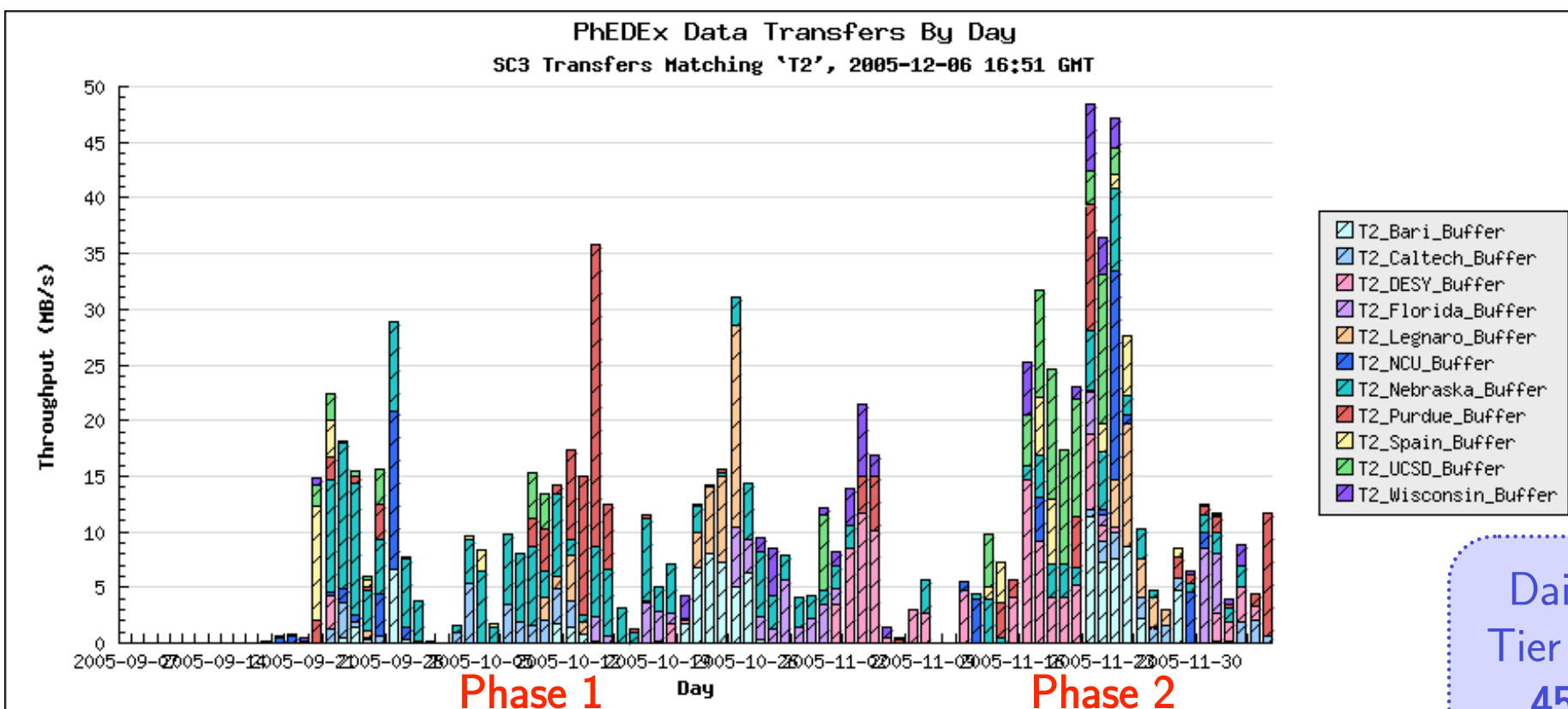
Considerable improvement
in transfer rate and quality
for service phase 2.



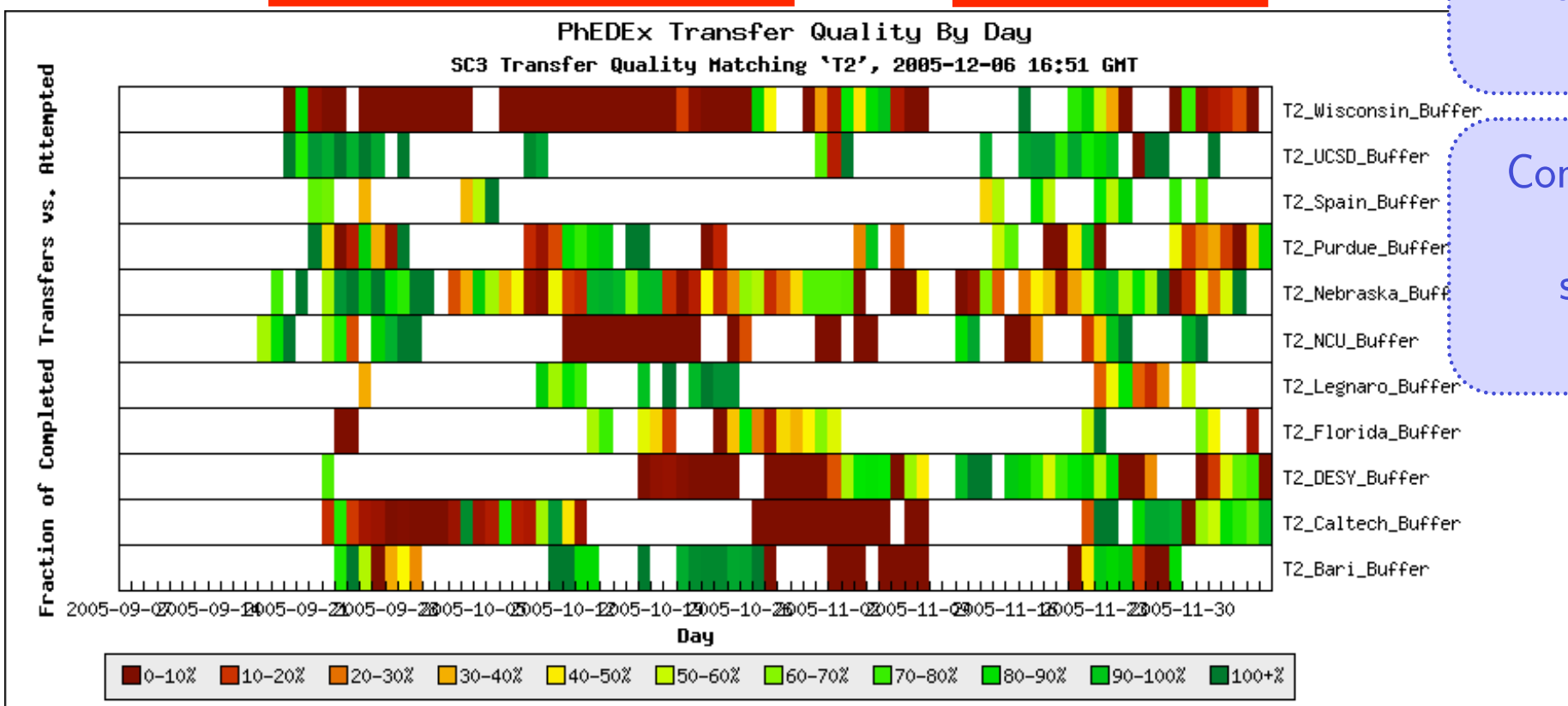
Daily aggregate rate from
Tier 0 to Tier 1s peaked at
120 MB/s.

SC3 Data Transfers to Tier-2s

Tier-2 WAN Transfers



Daily aggregate rate from Tier 1s to Tier 2s peaked at 45 MB/s, typically well below 30 MB/s.



Considerable improvement in quality, rate and system sustainability for service phase 2.

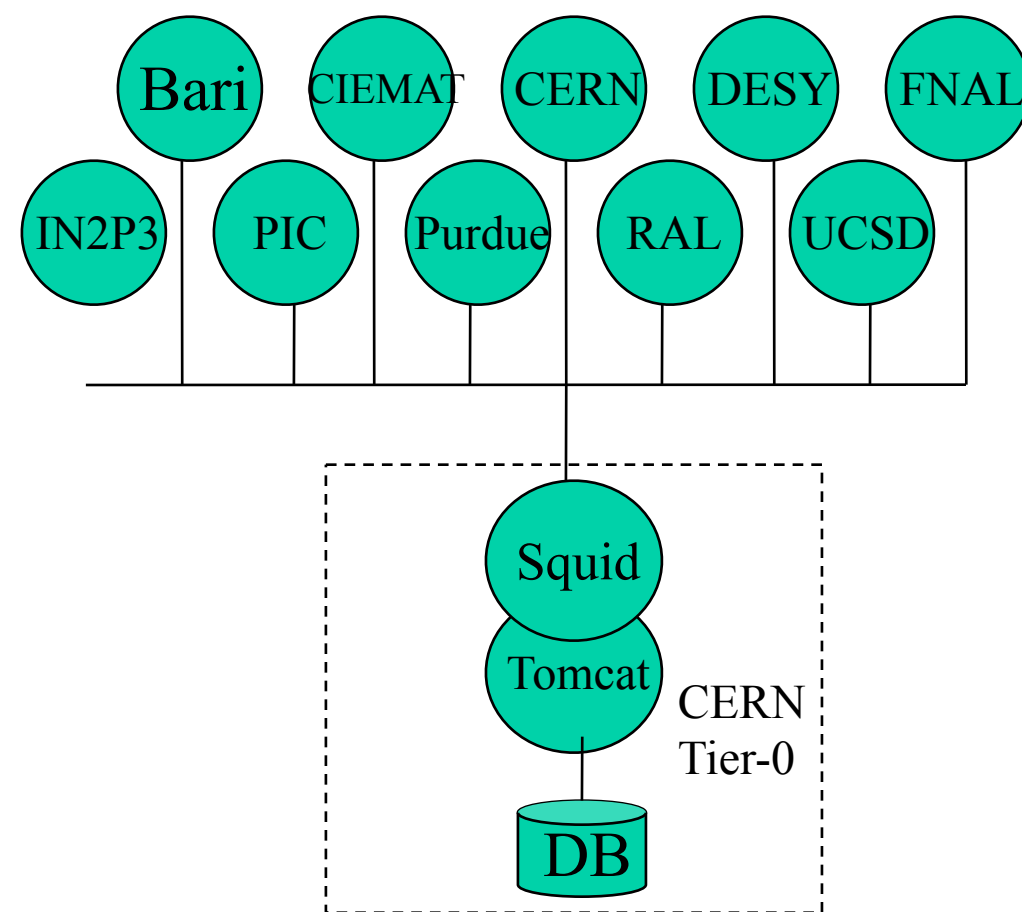
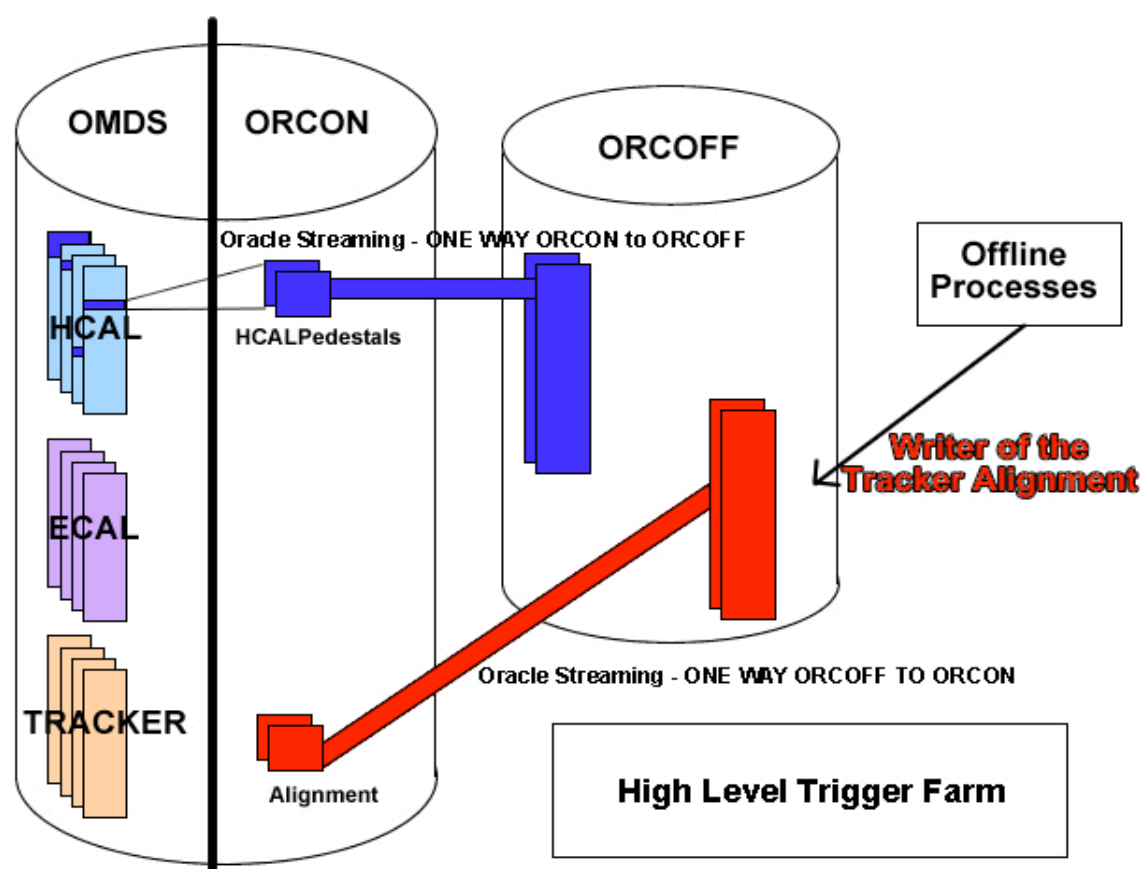


SC3 Lessons Learned

- ◆ **Impressive results, w/ large number of sites ready for real use**
 - ◆ various CMS computing components have done rather well
 - ◆ well-performing large and complex storage systems at many sites
- ◆ **SC3 was mostly about debugging components and systems**
 - ◆ infrastructure did often not work as expected, high error rates
 - ◆ required to de-scope some of the workflows/ data flows
 - ◆ much of (often new) grid software was not tested with CMS before
 - ◆ CMS/LCG integration task force is now addressing this issue
 - ◆ SC3 stretched old CMS data model to the limit
 - ◆ new EDM and data management tools address this, but are not yet operational
 - ◆ extremely costly in terms of manpower
- ◆ **Communicating goals and achievements, successes and failures**
 - ◆ Serious “impedance mismatch” between experiment and WLCG/sites
 - ◆ CMS people at sites are instrumental to success
 - ◆ WLCG as a collaboration gets established, requires to establish excellent relationships between CMS experiment and IT providers

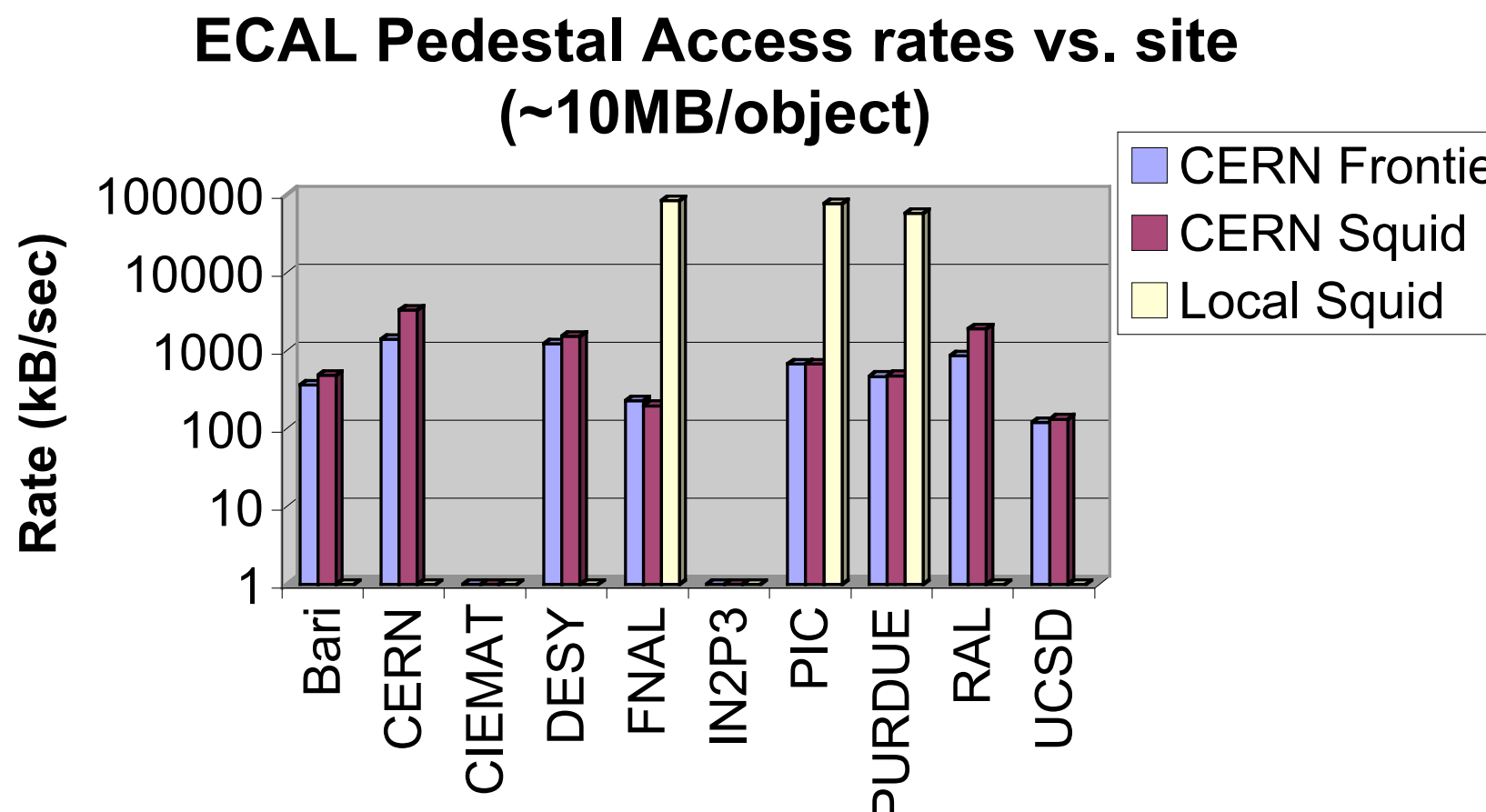
Distributed Conditions DBase Test

- ◆ Online-2-Offline (O2O) transfers of conditions data tested
 - ◆ tables loaded w/ 6-month-equivalent data: HCAL, ECAL, SiTracker
- ◆ Tier-0 to Tier-N (O2N) test w/ Frontier proceeding, initial results
 - ◆ cmsRun (cmssw 0_2_0_condtest release) installed at sites
 - ◆ test jobs access conditions database, through distributed caching
 - ◆ Frontier POOL plugin enables squid-based dBase access
 - ◆ squid-based infrastructure on sites 10 sites



First Results of Frontier Test

- ◆ Demonstrated Functionality with cmssw_0_2_0
- ◆ Instrumented squids to measure access patterns, throughputs
 - ◆ some measurements with cmssw and with Python scripts
- ◆ Results so far
 - ◆ Database configuration seems good, Pool/Frontier Plugin working stably
 - ◆ Installation at sites went smoothly
 - ◆ Local Squids work
 - ◆ High performance access once the cache is loaded
 - ◆ (CERN still on 100Mbps)
- ◆ More stress testing
- ◆ More function. testing





CMS Plans for 2006

- ◆ 2006 is the final year of preparation before start of physics
 - ◆ building on existing components and experience, but also deploying a lot of new and re-factored packages
 - ◆ transitioning to a new Event Data Model and FrameWork
 - ◆ deploy new MC production infrastructure
 - ◆ transition to new CMS dataset management system
 - ◆ work out the CMS-Tier-0 workflows and data flows
- ◆ Three major experiment challenges for CMS Computing
 - ◆ Magnet Test/Cosmic Challenge
 - ◆ commission new software, distribute Detector Conditions Database and event data
 - ◆ CMS participation in WLCG Service Challenge 4
 - ◆ bring data analysis services up to scale
 - ◆ Computing Software and Analysis Challenge 2006
 - ◆ end-to-end system test of CMS workflows



Plans for SC4

- ◆ Details to be discussed in the SC4 workshop(s)
- ◆ main goal:
 - ◆ CMS to be able to use WLCG service for the CSA2006 use case
- ◆ approach should involve stricter site integration and verification
 - ◆ should aim for mass storage validation at 400MB/s by ~ July 2006
 - ◆ aim for 100MB/s at Tier-2 centers
- ◆ demonstration of T1-T1 and T2-any T1 connectivities
 - ◆ CMS computing model has modest T1-T1 and limited permutations of T2-T1 transfers
 - ◆ but need to verify generic T1-T2 transfers for each T2 to “get at any data”
 - ◆ requires discussions on networking architectures and data transfer layer
 - ◆ role of FTS n-n channels, T1 “proxies”, ...?
 - ◆ whole list of site transfer goals 1 ramping up to 4 TB/day
- ◆ scaling of job submission: filling the grid with CMS applications
 - ◆ currently 3000 jobs/day (for user analysis), to scale to 200k/day!



CMS for SC4

- ◆ *SC4 goals for CMS will be very similar to SC3 goals*
 - ◆ run the important CMS use cases: data placement and job running
 - ◆ include workflow at Tier-0
 - ◆ placement of data samples at T1 sites
 - ◆ run skimming and selection and transfer to T2s
 - ◆ include calibration/alignment distributed infrastructure
 - ◆ include “fake” analysis at Tier-2 site
 - ◆ all this will stress the I/O and throughput at each of the site!
 - ◆ by end of challenge try to sustain a submission rate
 - ◆ demonstrate ability for sustained use of WLCG service
 - ◆ goal should be 50% utilization of resources available to CMS
 - ◆ 50% of the required T1-T2 permutations
- ◆ *CMS is very excited about all the sites coming up and forming the WLCG collaboration and infrastructure*
 - ◆ thank you for your close collaboration to prepare for the start of physics!

Conclusions

- ◆ The WLCG Distributed Computing System for CMS is a reality!
 - ◆ it works, has lots of resources, many good people working hard
 - ◆ it also still has many severe problems — users are the “first to know” :-)
 - ◆ there is also a lot of momentum and some convergence
- ◆ Next: Integration into functional Computing Service
 - ◆ SC3 with impressive progress, in particular Tier-2s
 - ◆ at this point the WLCG system is scarily fragile
 - ◆ database test first successful results
- ◆ Many more challenges after SC3
 - ◆ SC4 including “analysis” workflow preparing for CSA2006 system test

