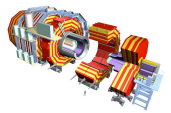
Il Modello di Calcolo di CMS

D. Bonacorsi *(INFN-CNAF Tier-1, Bologna, Italy)*

on behalf of the CMS experiment

Outline

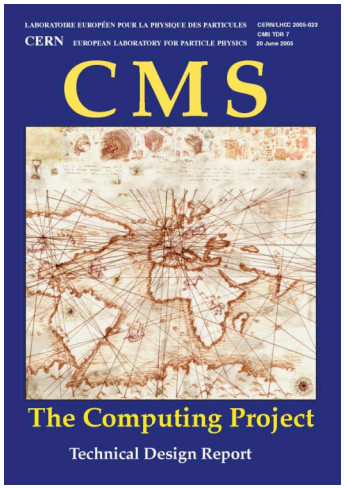


- The CMS distributed computing system
  - from guiding principles to architectural design
- Workflows (and actors) in CMS computing
  - Data Management (DM) and Workload Management (WM)
- The realization of the CMS Computing Model in a Grid-enabled world
  - Implementation of production-level systems on the Grid
    - Data Distribution, MonteCarlo (MC) production, Data Analysis
  - Computing challenges
    - Worldwide LCG challenges, and experiment-specific challenges

CMS Computing Model

- The CMS computing system relies on a *distributed infrastructure* of Grid resources, services and toolkits
  - distributed system to cope with computing requirements for storage, processing and analysis of data provided by LHC experiments
  - building blocks provided by Worldwide LHC Computing Grid [WLCG]
    - CMS builds application layers able to interface with few - at most - different Grid flavors (LCG-2, Grid-3, EGEE, NorduGrid, OSG)

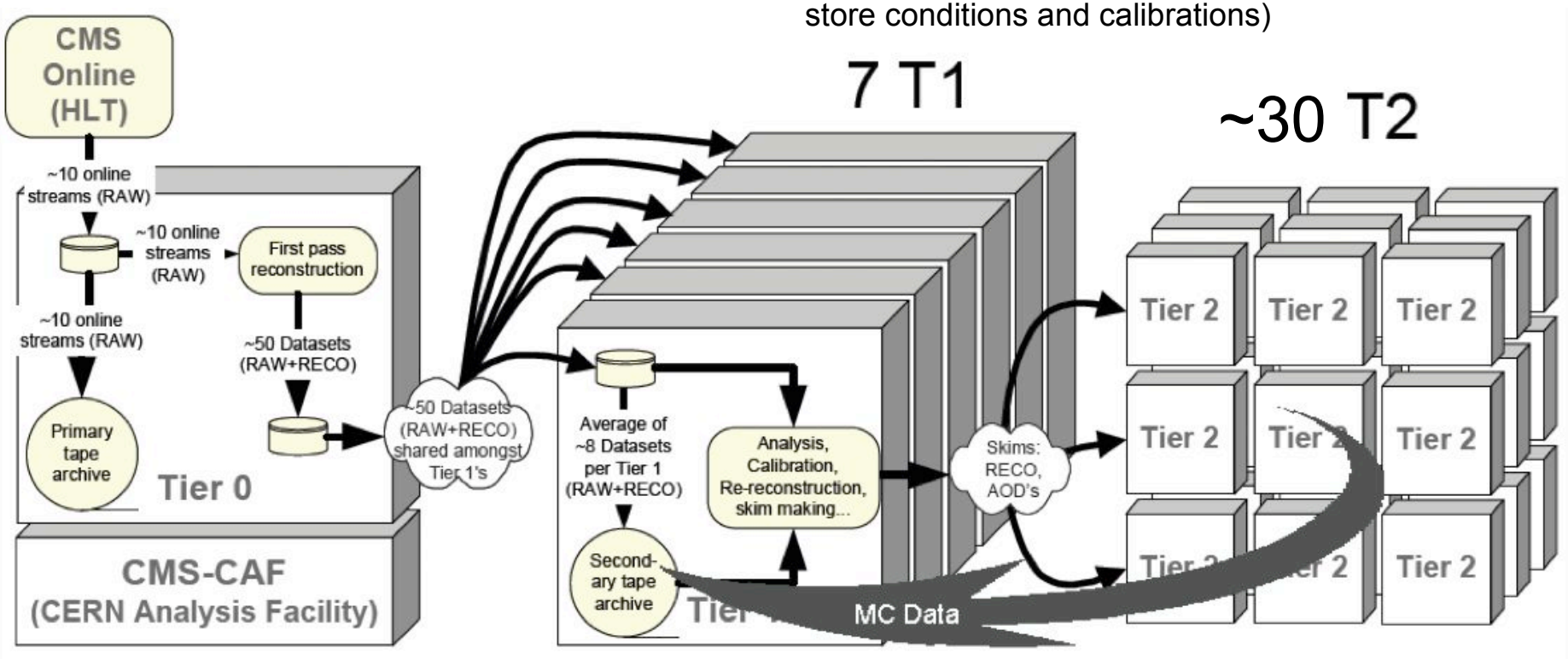
- CMS computing model document (CERN-LHCC-2004-035)
- CMS C-TDR released (CERN-LHCC-2005-023)
  - in preparation for the first year of LHC running (2008)
    - not "blueprint", but "baseline" targets (+ devel. strategies)
  - hierarchy of computing tiers using WLCG tools
    - focus on Tiers role, functionality and responsibility



- Now partially "old" already!

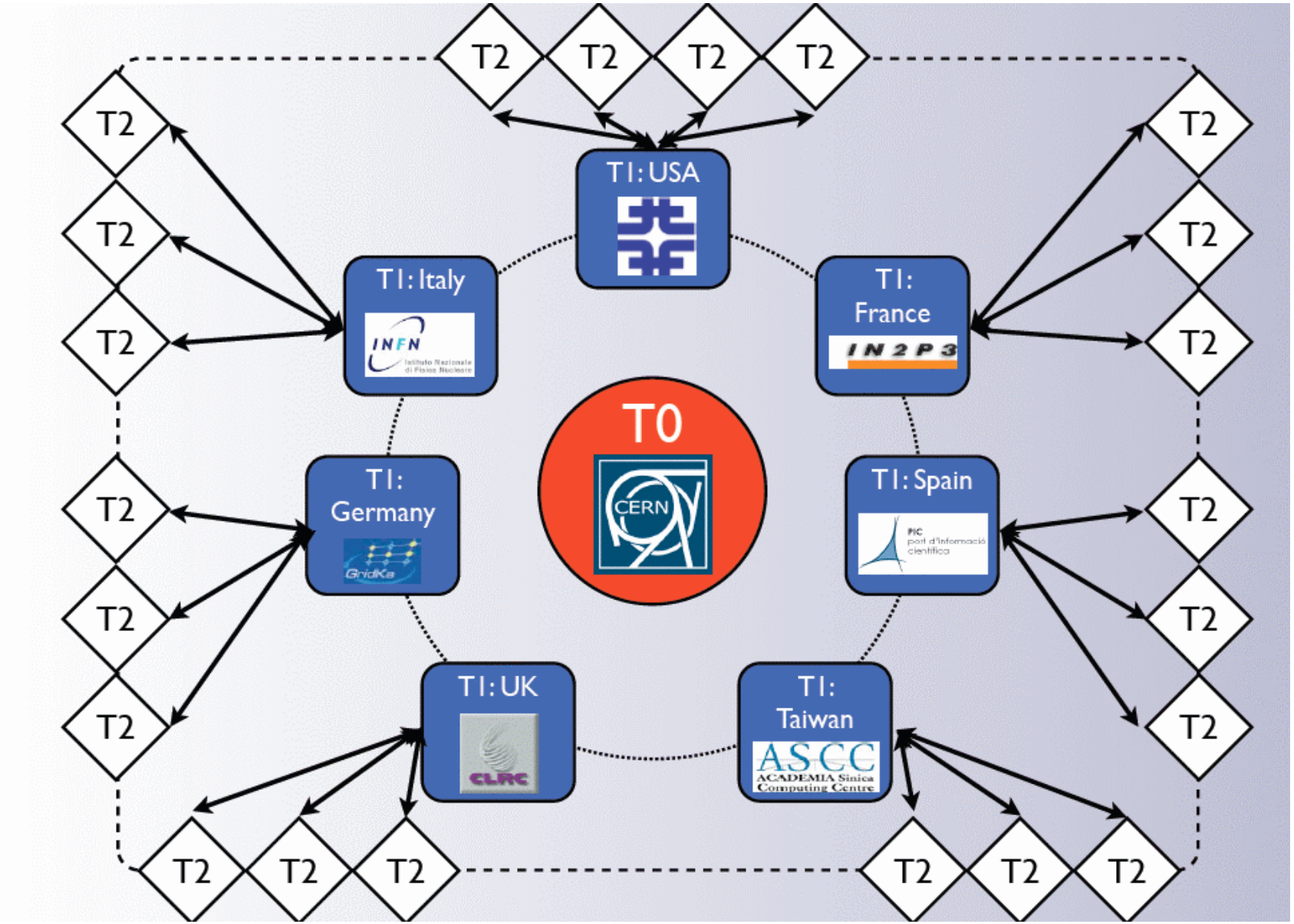
Tiered architecture

- T0:
  - Accepts data from DAQ
  - Prompt reconstruction
  - Data archive and distribution to T1's
- CAF (CERN Analysis Facility for CMS):
  - Access to full raw dataset
  - Focused on latency-critical activities (detector diagnostics, trigger performance services, derivation of Al/Ca constants)
  - Provide some CMS central services (e.g. store conditions and calibrations)



- 7 T1 centers and ~30 T2 centers (see next slide)

Toward a "mesh" model



T1/T2 roles and computing capacities

CMS T1 functions

- Scheduled data-reprocessing and data-intensive analysis tasks:
  - later-pass reco, AOD extraction, skimming, …
- Data archiving (real+MC):
  - custody of raw+reco & subsequently produced data
- Disk storage management:
  - fast cache to MSS, buffer for data transfer, …
- Data distribution:
  - data serving to Tier-2's for analysis
- Analysis:
  - proficient data access via CMS+WLCG services

CMS T2 functions

- User data analysis
- Fast and detailed MC event prod
- Import skimmed datasets from T1s and export MC data
- Data processing for calib/align tasks and detector studies

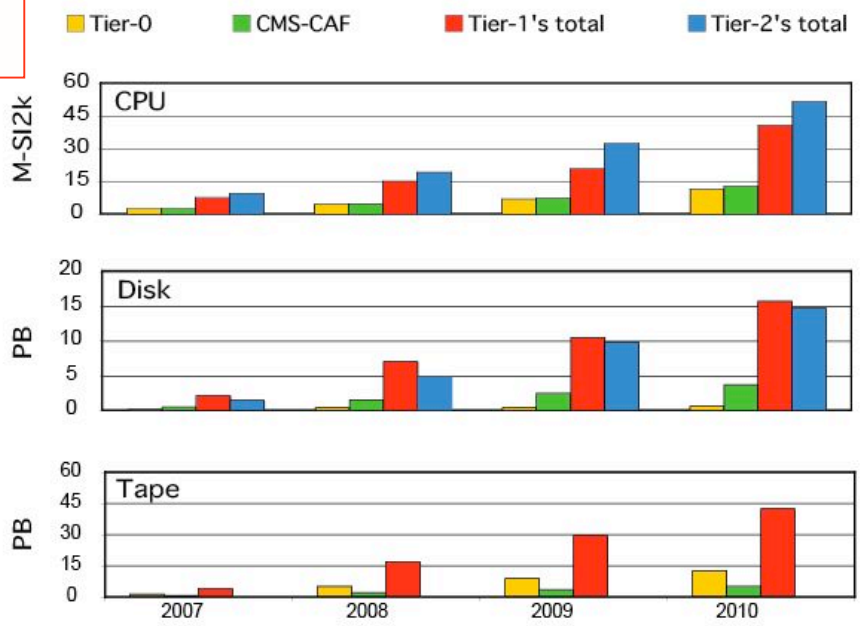
CMS T1 resources (nominal for average T1 in 2008):

NB:1/7

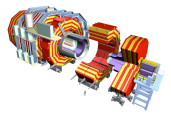
- ✓ WAN: transfer capacity ~10 Gb/s
- ✓ CPU: 2.5 M-SI2k (scheduled reprocessing : analysis = 2 : 1)
- ✓ Disk: 0.8 PB (~85% for analysis data serving)
- ✓ MSS: 2.8 PB (losses ~tens of GB per PB stored)

CMS T2 resources (nominal for average T2 in 2008):

- ✓ WAN: 1 Gb/s (at least)
- ✓ CPU: 900 k-SI2k
- ✓ Disk: 200 TB



Data-driven baseline



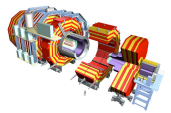
Technical baseline principles

- Baseline system with minimal functionality for first physics
  - ‘Keep it simple!’
  - Use Grid services as much as possible + also CMS-specific services
  - Optimize for the common case
    - for read access (most data is write-once, read-many)
    - for organized bulk processing, but without limiting single user
  - Decouple parts of the system
    - Minimise job dependencies + site-local information remain site-local

- T0-T1s activities driven by **data placement** in the CMS baseline model
  - Data is partitioned by the exp as a whole, do not move around in response to job submission, all data is placed at a site through explicit CMS policy
  - Tier-0 and Tier-1 are resources for the whole experiment
  - Leads to very ‘structured’ usage of Tier-0 and Tier-1
    - activities and functionality are largely predictable since nearly entirely specified
      - i.e. organized mass processing and custodial storage

- ‘unpredictable’ computing essentially restricted to data analysis at T2s
  - T2s are the place where more flexible, user driven activities can occur
  - Very significant computing resources and good data access are needed

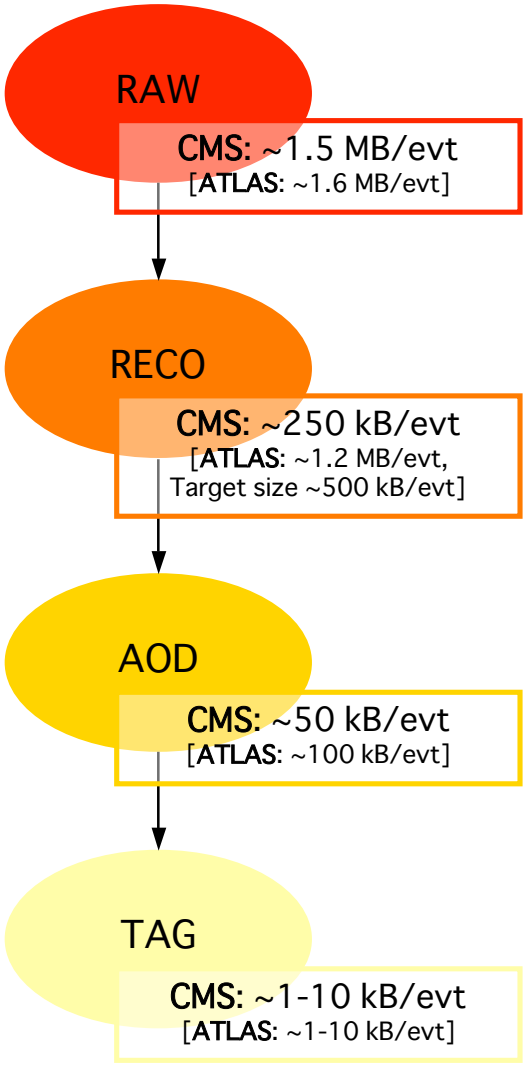
Data organization



- CMS expects to produce large amounts of data (events)
  - O(PB)/year
- Event data are in **files**
  - average file size is kept reasonably large (≥ GB)
    - avoid scaling issues with storage systems and catalogues when dealing with too many small files (+ foresee file merging)
  - O(10^6) files/year
- Files are grouped in **fileblocks**
  - group files in blocks (1-10 TB) for bulk data management reasons
    - exist as a result of either MC production or data movement
  - 10^3 Fileblocks/year
- Fileblocks are grouped in **datasets**
  - Datasets are large (100 TB) or small (0.1 TB)
    - Dataset definition is physics-driven (size as well)

Data types

➢ Data tiers/volumes for 2008 as input parameters for the model*

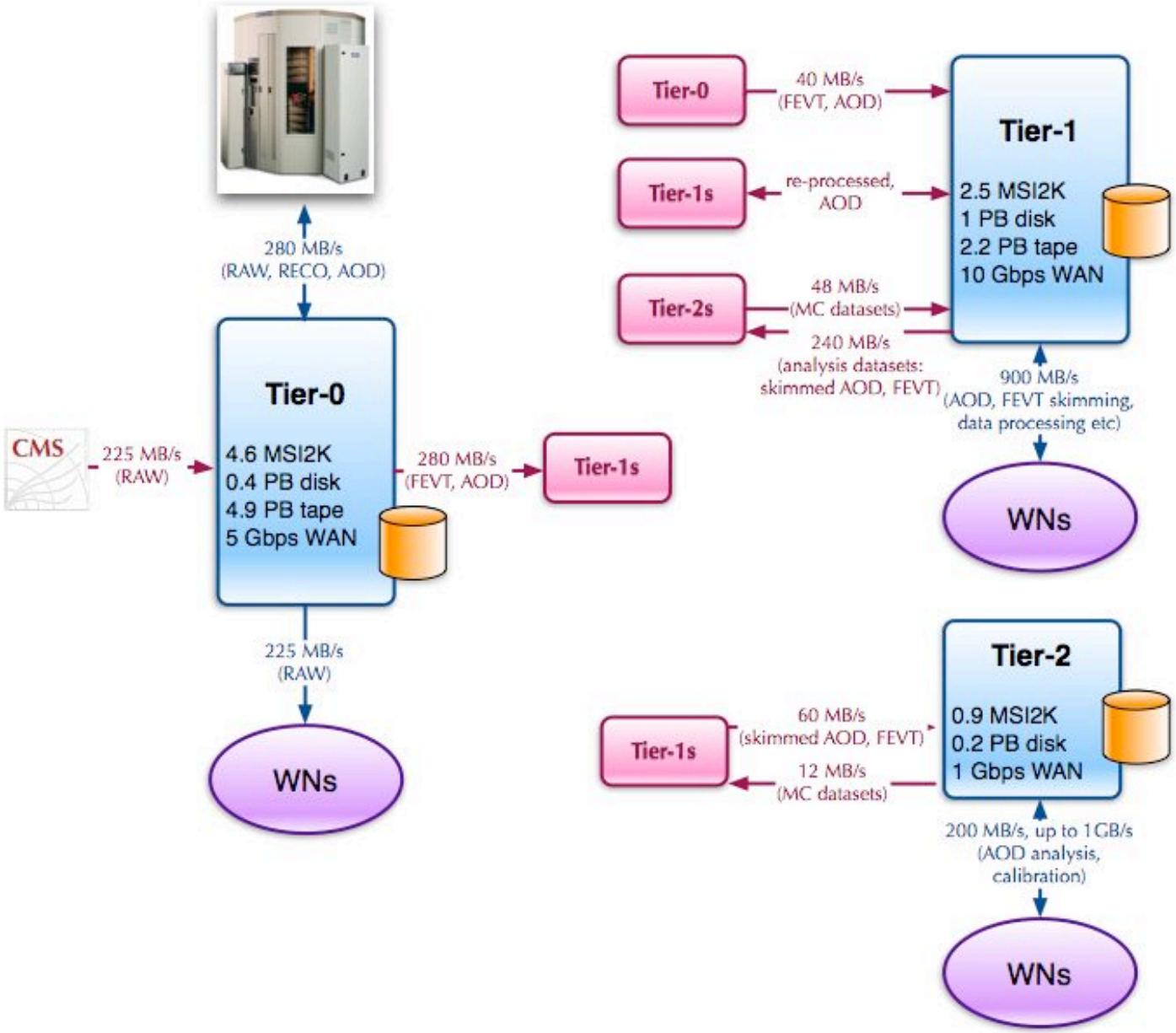


[*] safety factors included (poor understanding of the detector, compression, …)

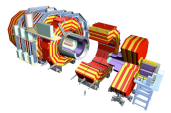
- **RAW**
  - Triggered evts recorded by DAQ
  - ~1.5 MB/evt @ ~150 Hz; ~ 4.5 PB/yr
    - 2 copies: 1 at T0 and 1 spread over T1s
- **RECO**
  - Reconstructed objects with their associated hits
    - Detailed output of the detector reco: track candidates, hits, cells for calib
  - ~250 kB/evt; ~ 2.1 PB/yr (incl. reprocessing)
    - 1 copy spread over T1s (together with associated RAW)
- **AOD** (Analysis Object Data)
  - Main analysis format: objects + minimal hit info
    - Summary of the reco evt for common analyses: particles id, jets, …
  - ~50 kB/evt; ~ 2.6 PB/yr
    - Whole set copied to each T1, large fraction copied to T2
- **TAG**
  - Fast selection info
    - Relevant info for fast evt selection in AOD
  - ~1-10 kB/ev

Plus MC in ~ N:1 ratio with data

CMS data flows



WM and DM



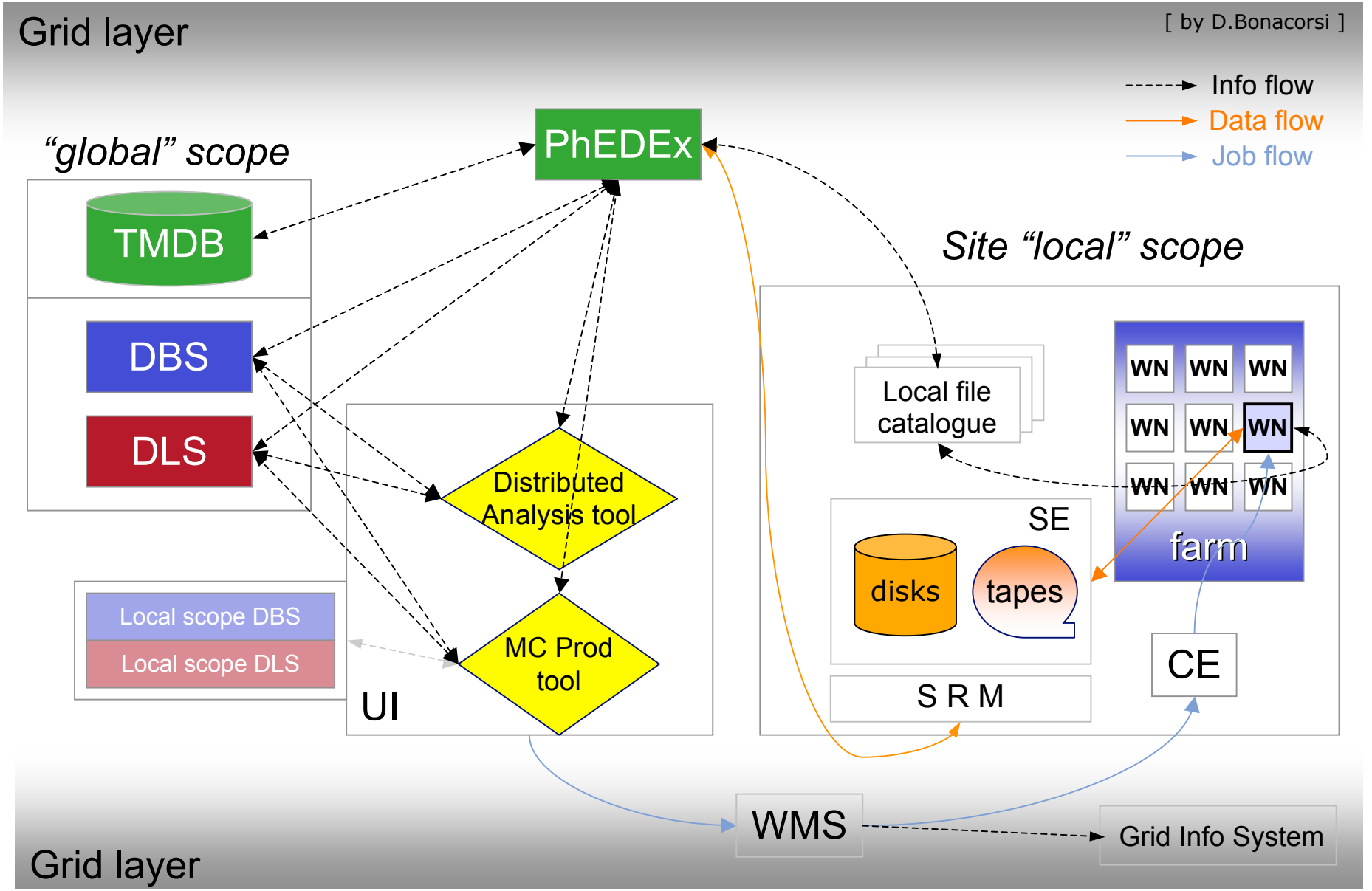
[Note: the migration was not disruptive]

- Recent (2006) migration to **new Data Management**
  - Provide new tools to discover, access and transfer event data in a distributed computing environment
    - Track and replicate data with a granularity of file blocks
    - Reduce load on catalogues

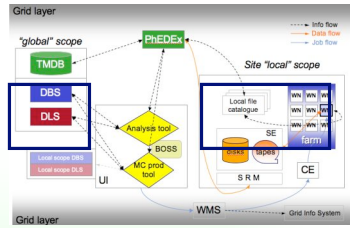
  - ✓ DBS (Dataset Bookkeeping system)
    - DBS provides the means to define, discover and use CMS event data
  - ✓ DLS (Dataset Location Service)
    - DLS provides the means to locate replicas of data in the distributed system
  - ✓ local file catalogue solutions
    - A “trivial” file catalogue as a baseline solution
  - ✓ PhEDEx integration with **most recent gLite services** (see later)

- New DMS is being exercised with **new MC production system**
  - integrate with **new Event Data Model** and **new DMS**

Data processing workflow



DBS / DLS / local file catalogue



Data definition:
- dataset specification
  (content and associated metadata)
- track data provenance

Data discovery:
- which data exist
- dataset organization
  (in term of fileblocks/files)
- site independent information

"What data exists?"

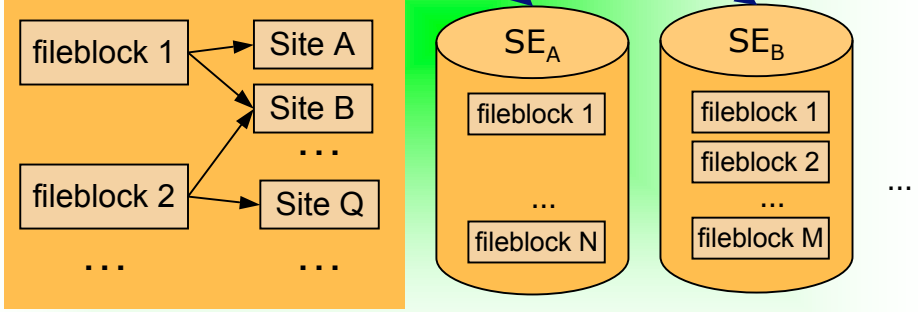
DBS

Interaction with DBS:
- Distributed analysis tool
- MC Production system
- PhEDEx for injection
- User query

"Where is data located?"

DLS

Integration with DLS:
- Insert file-blocks produced at a site
- Insert file-blocks upon data replication
- Query to locate file-blocks
  (e.g. analysis tool)

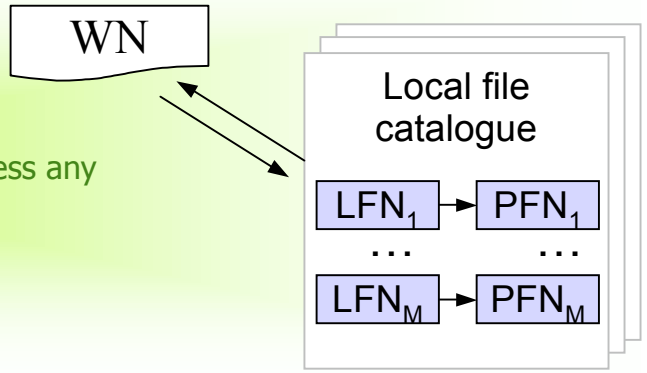


DLS maps fileblocks to SEs where they are located

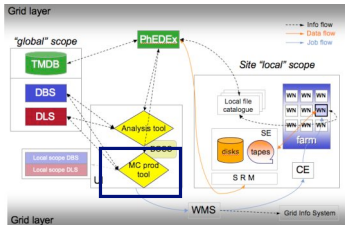
site independent | **site dependent**

site dependent job configuration

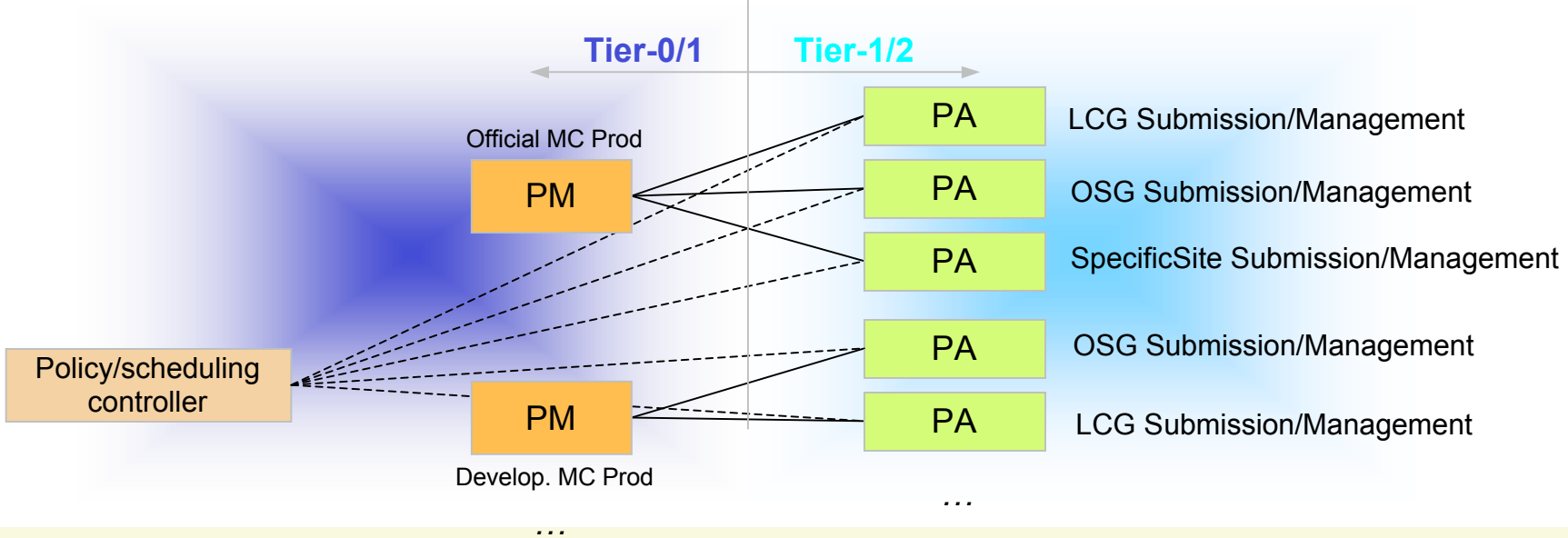
- Need a catalogue + a site local discovery mechanism
  - discover at runtime on WN the site-dependent data organization
  - local file catalogues provide site local information about how to access any given file (aka "**LFN-to-PFN mapping**")
    - CMS baseline solution is to use a trivial file catalogue
    - High-rate large-scale performances required



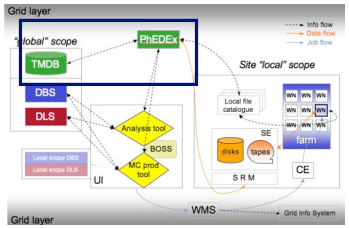
New MC Production system



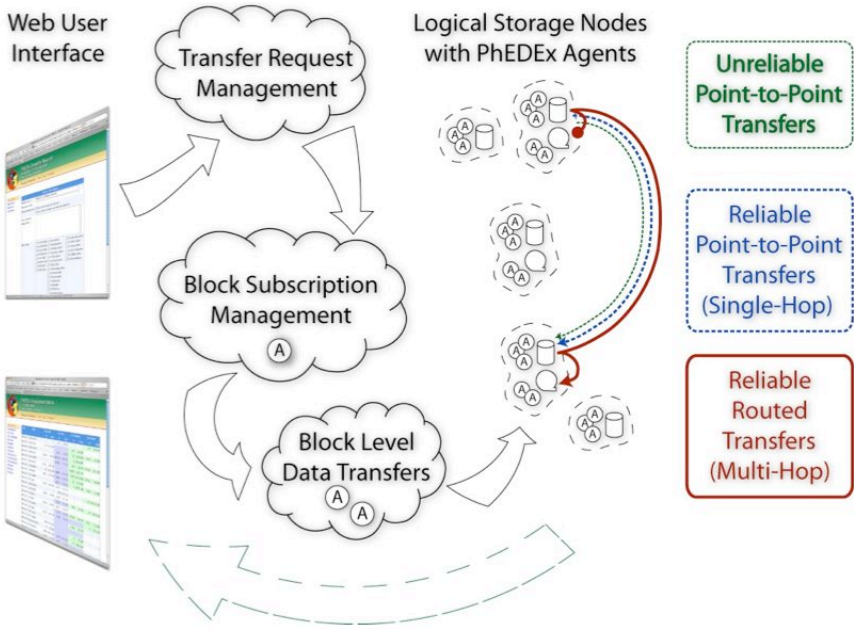
- New MC production system developed in 2006, in production already
  - Overcome current inefficiencies + introduce new capabilities
    - less man-power consuming, better handling of Grid-sites unreliability, better use of resources, automatic retrials, better error report/handling
  - More flexible and automated architecture
    - **ProdManager** (*PM*) (+ the policy piece)
      - manage the assignment of requests to 1+ *ProdAgents* and tracks the global completion of the task
    - **ProdAgent** (*PA*)
      - Job creation, submission and tracking, management of merges, failures, resubmissions, …
        - It works with a set of resources (e.g. a Grid, a Site)
  - Integrate with new Event Data Model and new DMS
    - orchestrate the interactions with local scope DBS/DLS and data placement system

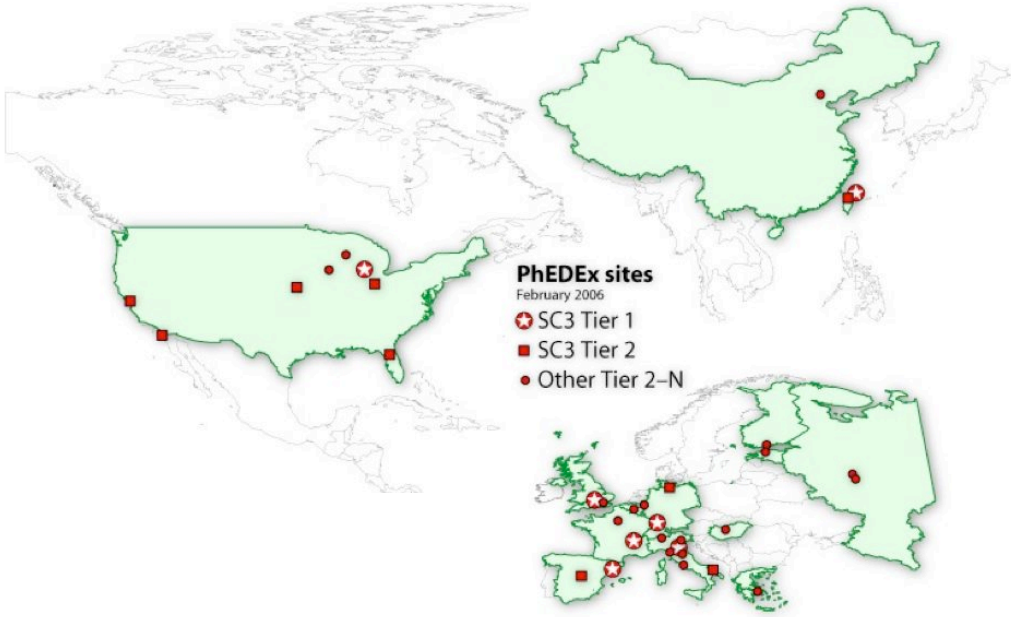


Data placement system

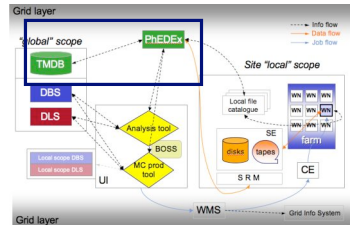


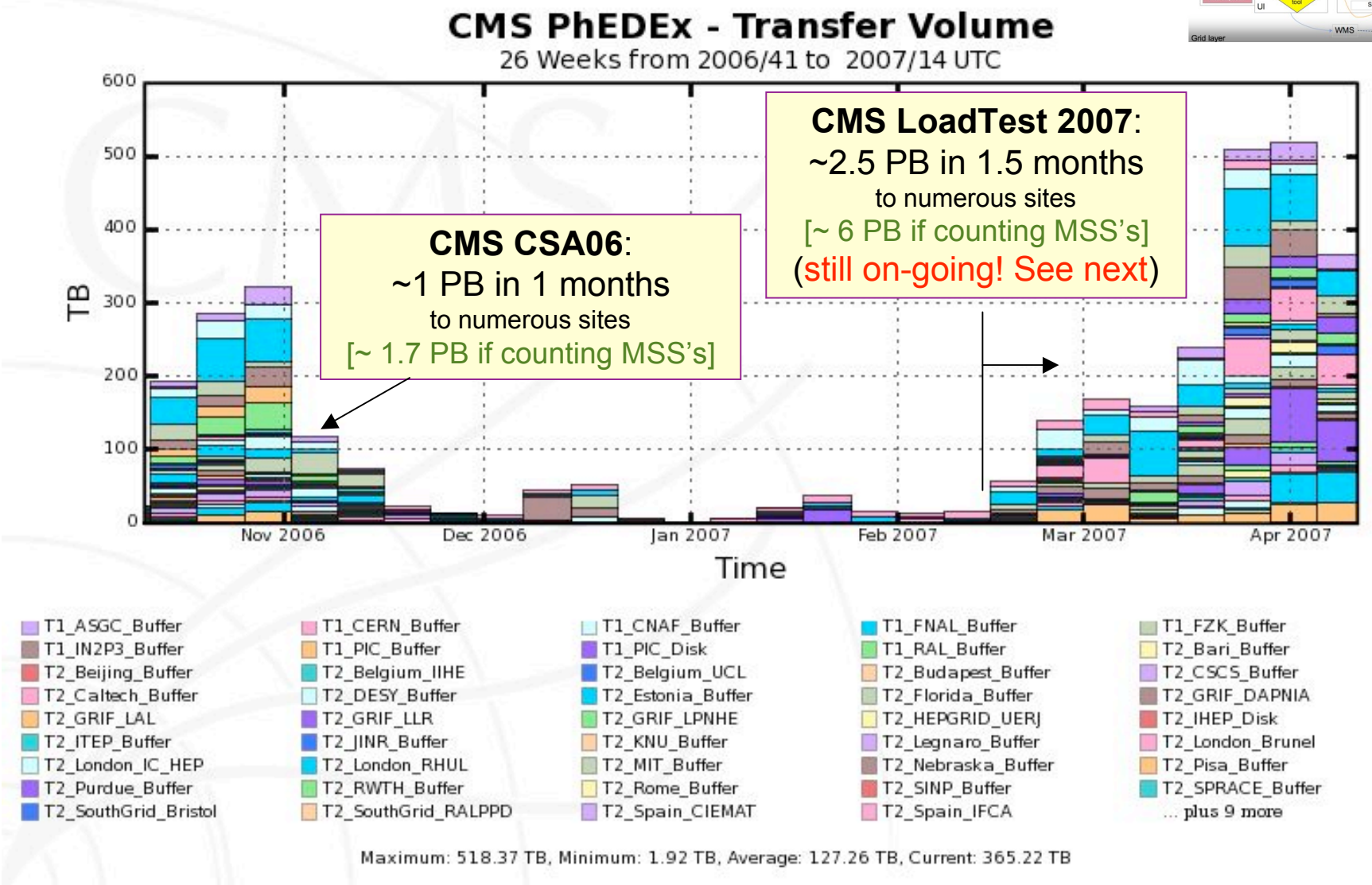
- Physics Experiment Data Export (PhEDEx)
  - large scale reliable dataset/fileblock replication
    - multi-hop routing following a transfer topology (T0 → T1's ↔ T2's), data pre-stage from tape, monitoring, bookkeeping, priorities and policy, etc
  - in production since almost 3 years
    - Managing transfers of several TB/day
    - **See performances in nect slide**
  - PhEDEx integration with gLite services File Transfer Service (FTS)
    - PhEDEx takes care of reliable, scalable CMS dataset replication (and more…)
    - FTS takes care of reliable point-to-point transfers of files





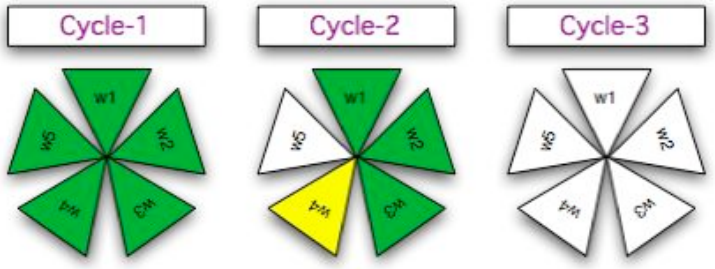
PhEDEx performances in 2006/2007





CMS LoadTest 2007

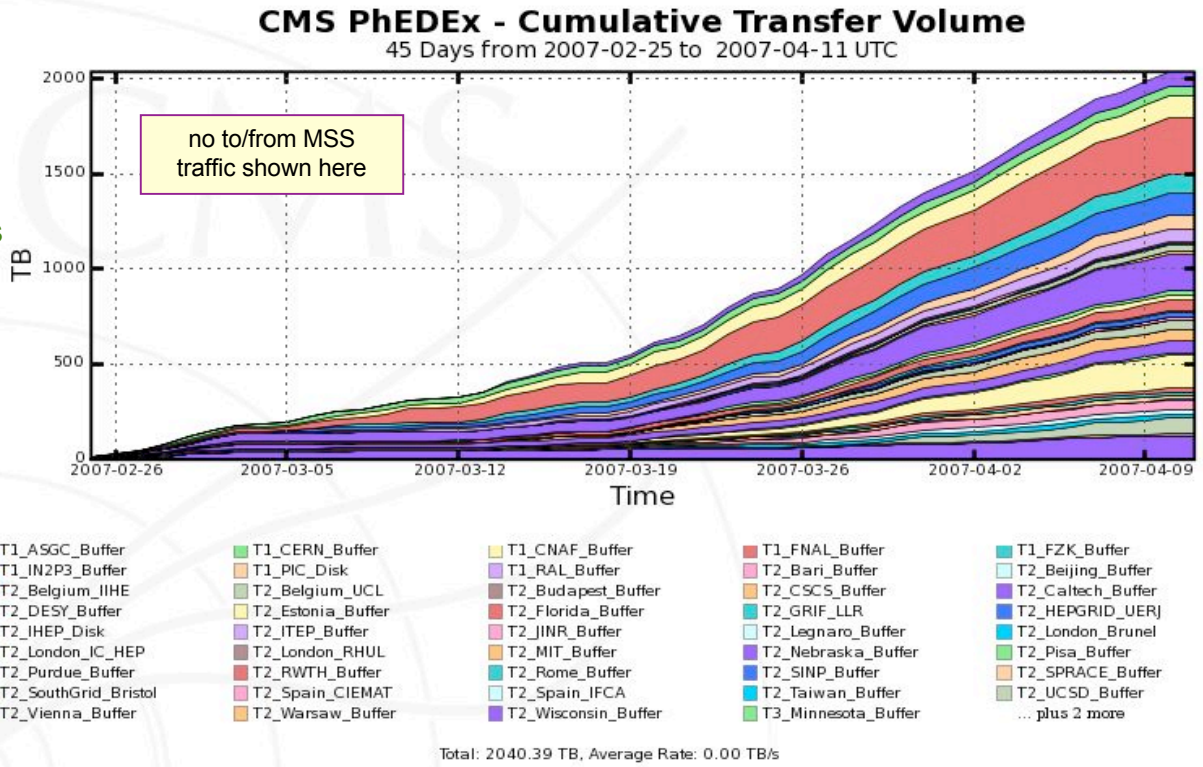
- **Focus**
  - Build and operate a CMS infrastructure for WLCG Tiers to exercise their transfer capabilities, their own storage systems, …



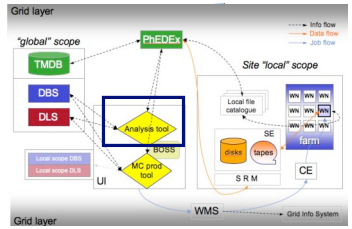
Now running Cycle-2 / week-4

- **Format**
  - 'Cycles' of testing 'weeks'
    - 3 full cycles before CSA07 ramp-up in Jun07

- **Exercises**
  - T0→T1(tape) transfers
  - T1↔T1 transfers
  - T1↔T2 transfers
    - also "non-regional" routes
  - "Variations" of the above

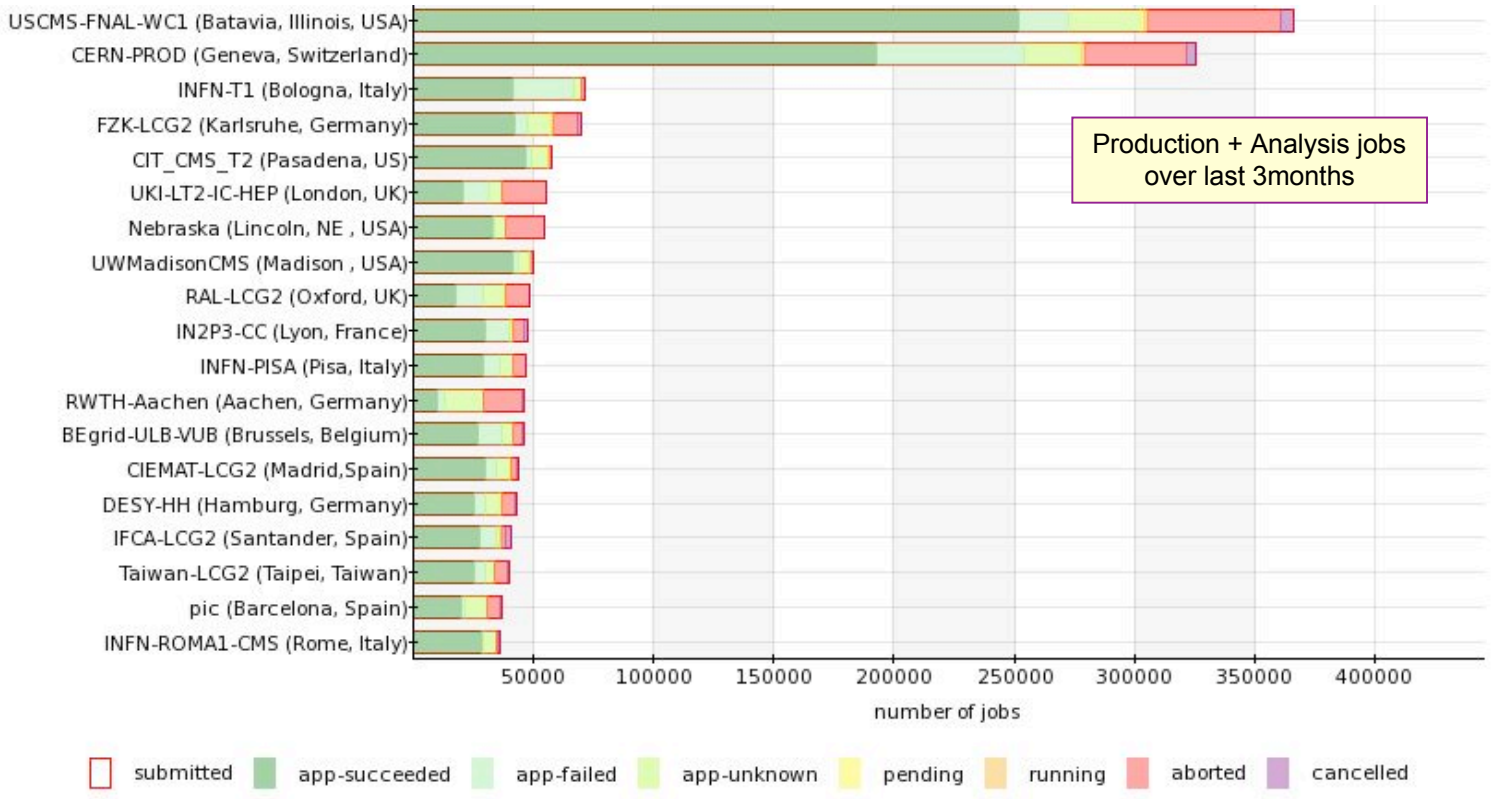


CMS distributed analysis on Grid

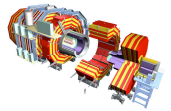


- Production jobs via the ProdAgents
- Analysis jobs via the CMS Remote Analysis Builder (CRAB)
  - Tool for job preparation, submission and monitoring

[→ see also V.Miccio, this conf]



Experience from Computing challenges

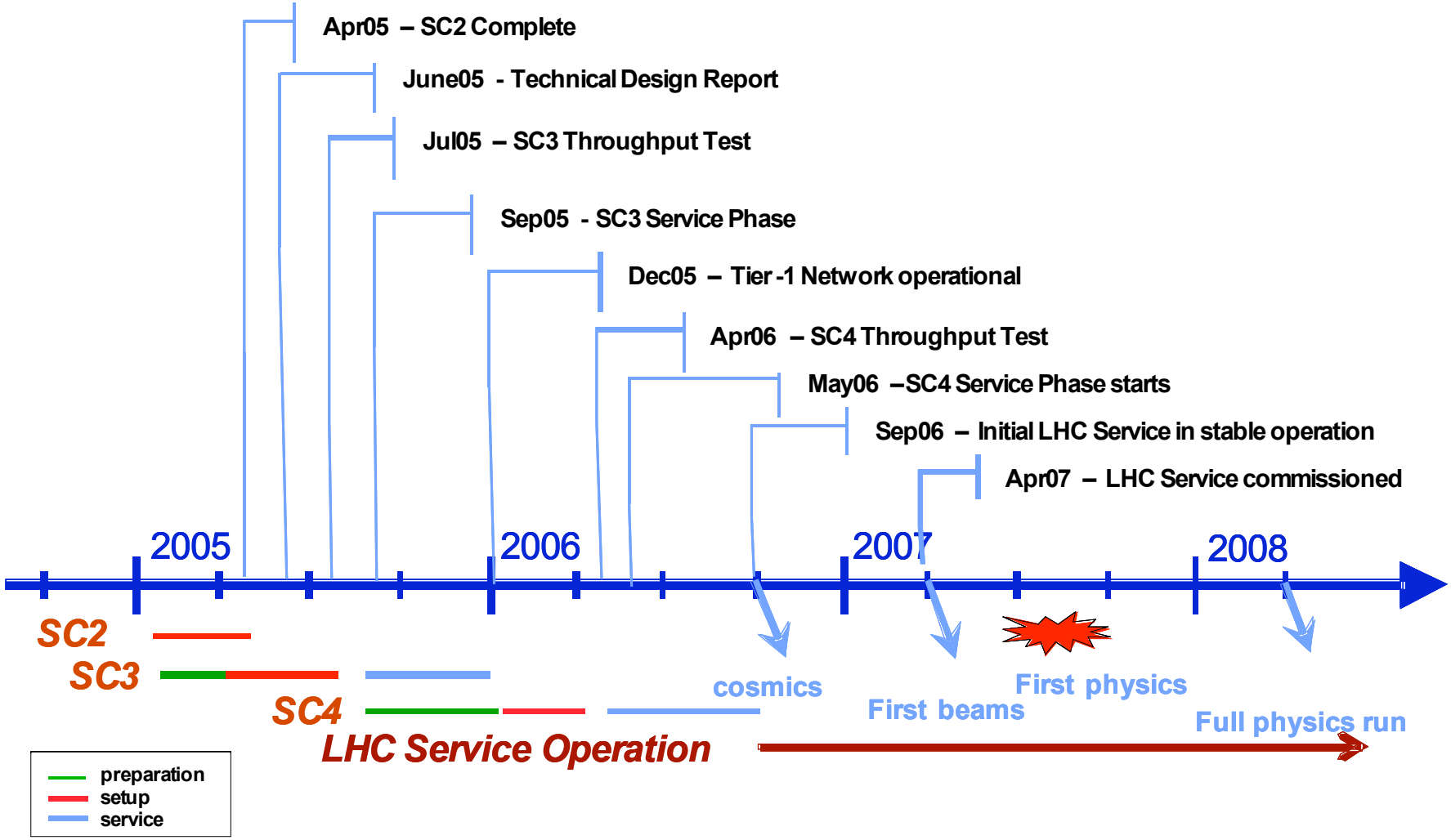


- CMS computing system realization is an iterative process
  - Grid resources/services and CMS solutions for WMS/DMS are tested in scheduled “challenges” of increasing scale and complexity

- Some are WLCG-wide…
  - WLCG Service Challenges
    - a mechanism by which the readiness of the overall LHC computing infrastructure to meet the exps’ requirements is measured and if(/where) necessary corrected
    - understand what it takes to run a **real and wide set of Grid services**

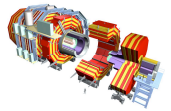
- … some are indeed CMS-specific
  - CMS Data Challenge 2004
  - CMS Computing, Software and Analysis 2006 (CSA06)
  - CMS Computing, Software and Analysis 2007 (CSA07)

WLCG SC's: it was a long path…



[figure: courtesy of J.Shiers]

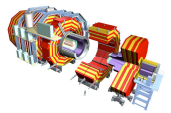
Computing, Software and Analysis Challenges



- Aimed to exercise CMS Computing and Software systems at a defined scale and at a certain level of functionality
  - CSA06 was a 25% activity (wrt 2008), CSA07 will push to 50%

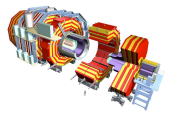
- CSAs include many workflow elements
  - E.g. CSA06:
    - Event reco at T0 center on a mix of samples at ~40Hz for 1 month
    - Data distribution to T1s (for archiving and data serving purposes)
      - T0-T1 rates based on MoU pledges
      - custodial archiving to tape where possible, or disk based archives for 30 days
    - Data skimming (data selection driven by physics groups) at T1s
    - Re-reco at T1s
    - Data serving to T2s and data access at T2s
      - Analysis job submission to T2s

CSA06 metrics (1/2)



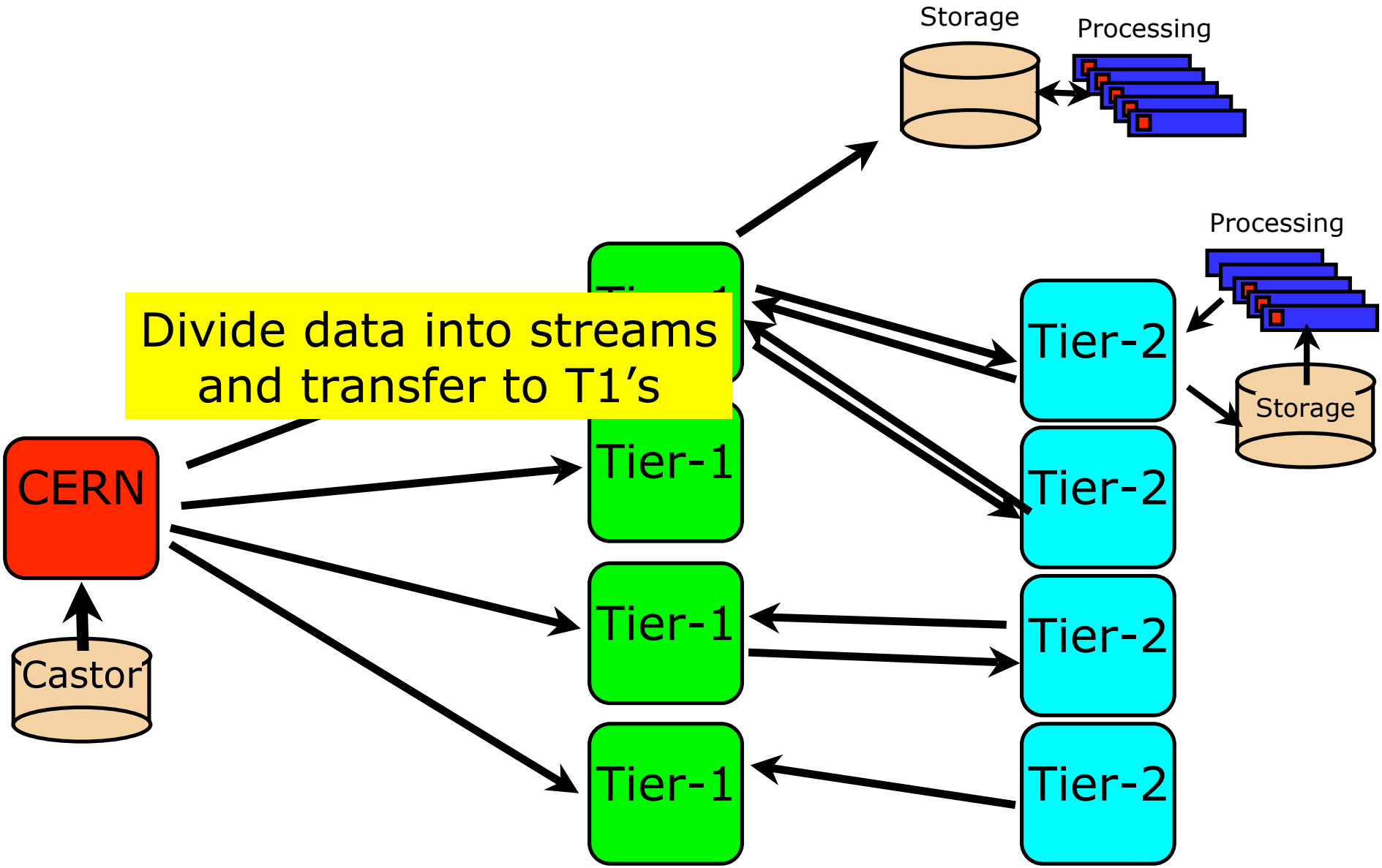
- Binary metrics
  - Automatic FEVT+AOD transfer Tier-0 to Tier-1 via PhEDEx, the data placement tool
  - Automatic transfer of part of FEVT+AOD Tier-1 to Tier-2 via PhEDEx
  - Offline DB accessible via FroNtier/Squid (a caching layer between the reconstruction jobs and the Oracle DB) at participating sites
  - Insertion and use new constants in Offline DB
  - User submission of analysis/calibration/skim jobs via the grid job submission tool CRAB and using the developed Dataset Bookkeeping Service (DBS) and Data Location Service (DLS)
  - Skim job output automatically moved to Tier-2 via PhEDEx
  - Running re-reconstruction-like jobs at Tier-1 that access updated information from the offline DB and per- form a new reconstruction on data distributed from the Tier-0 centre

CSA06 metrics (2/2)



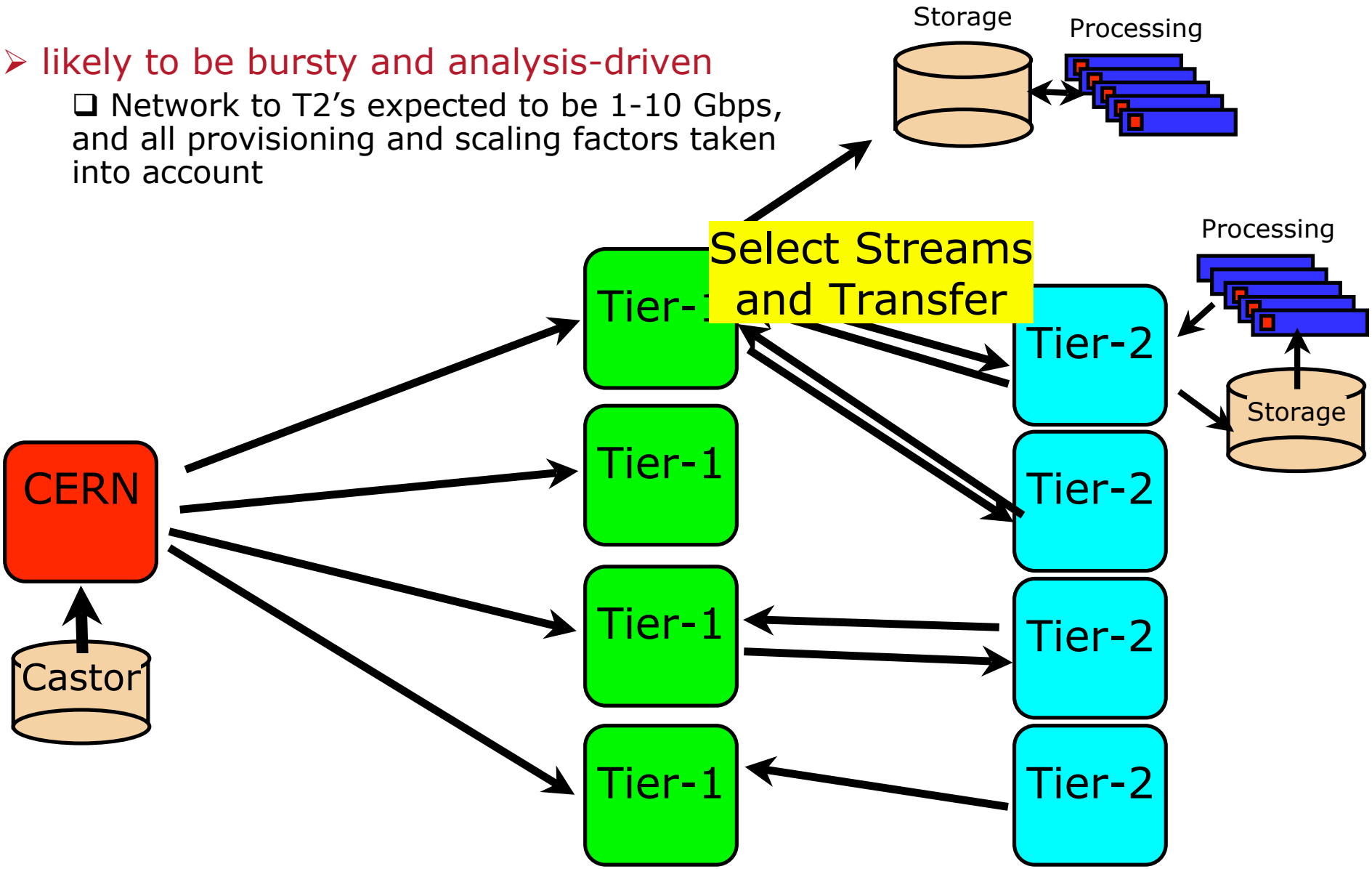
- Quantitative metrics
  - Number of participating Tier-1
    - Goal: 7 Threshold: 5. Passing requires 90% uptime, or < 3 days downtime during challenge
  - Number of participating Tier-2
    - Goal: 20 Threshold: 15
  - Weeks of running at sustained rate
    - Goal: 4 Threshold: 2. This will be the period over which we measure the other metrics
  - Tier-0 Efficiency
    - Goal: 80% Threshold: 30%. Measured as unattended uptime fraction over 2 best weeks of the running period
  - Running grid jobs (Tier-1 + Tier-2) per day (2h jobs typ.)
    - Goal: 50K Threshold: 30K
  - Grid job efficiency
    - Goal: 90% Threshold: 70%
  - Data serving capability at each participating site
    - Goal 1MB/sec/execution slot. Threshold : 400 MB/sec (Tier-1) or 100 MB/sec (Tier-2)
  - Data transfer Tier-0 to Tier-1 to tape
    - Individual goals (threshold at 50% of goal): ASGC: 10MB/s, CNAF: 25 MB/s, FNAL: 50 MB/s, GridKa: 20MB/s, IN2P3: 25MB/s PIC: 10 MB/s, RAL: 10MB/s
  - Data transfer Tier-1 to Tier-2
    - Goal: 20MB/s into each Tier-2. Threshold: 5MB/s
    - Overall success is to have 50% of the participants at or above goal and 90% above the threshold
    - Several Tier-2s have better connectivity and CMS hav higher targets for those
    - Goal for each Tier-2 is to demonstrate 50% utilization of the WAN to the best connected Tier-1

CMS CSA06: T0→T1 flows



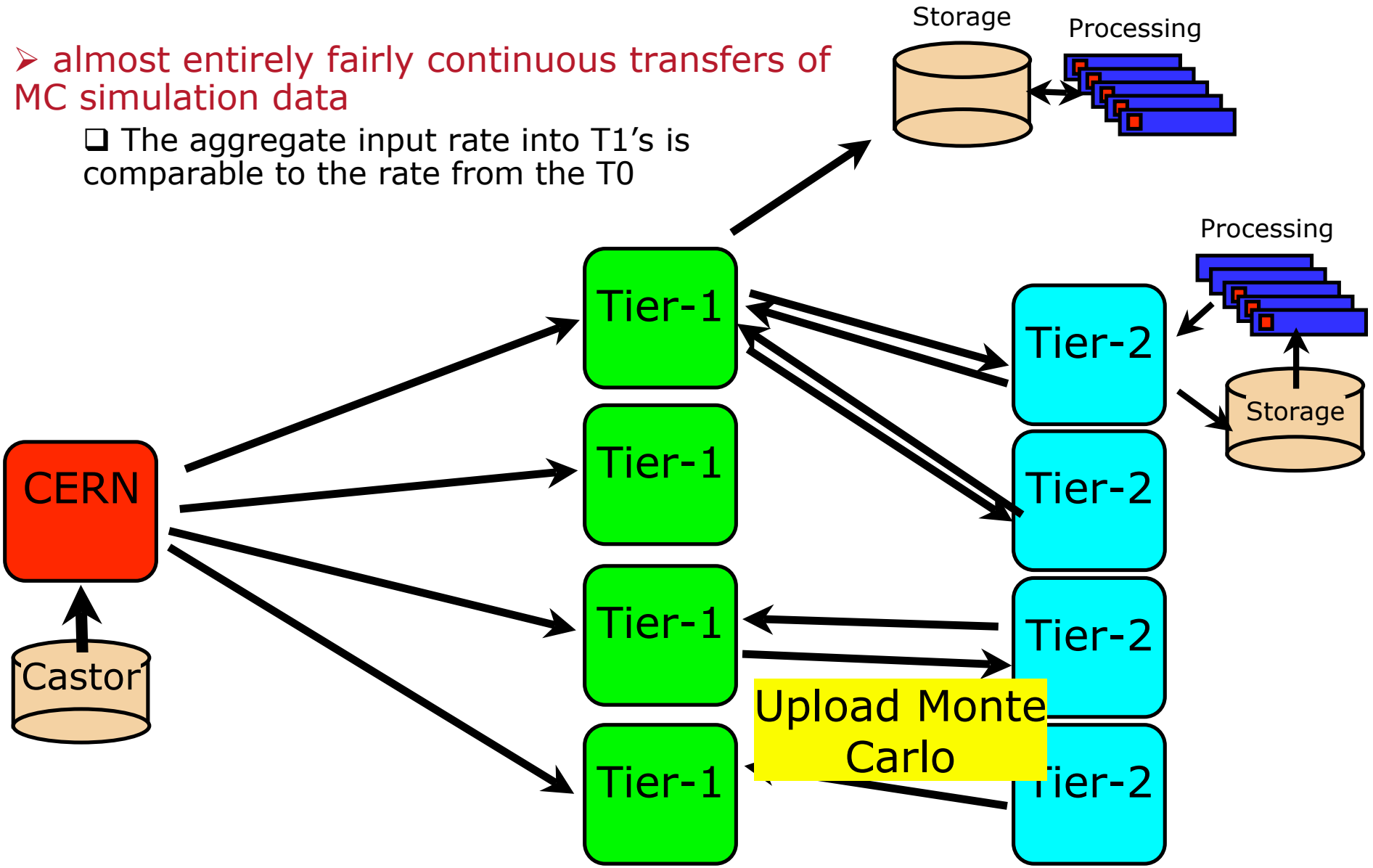
CMS CSA06: T1→T2 flows

- likely to be bursty and analysis-driven
  - Network to T2's expected to be 1-10 Gbps, and all provisioning and scaling factors taken into account



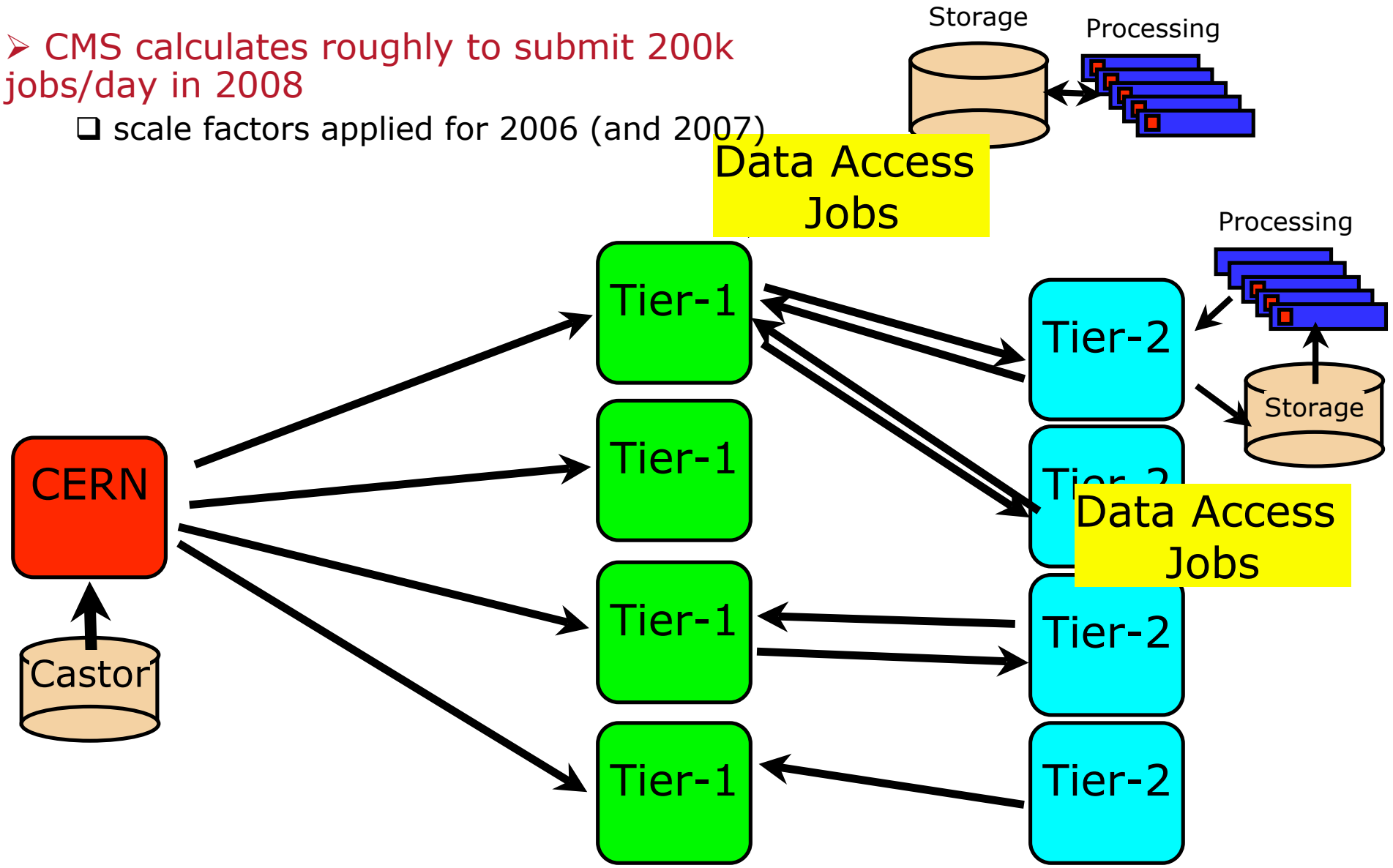
CMS CSA06: T2→T1 flows

- almost entirely fairly continuous transfers of MC simulation data
  - The aggregate input rate into T1's is comparable to the rate from the T0



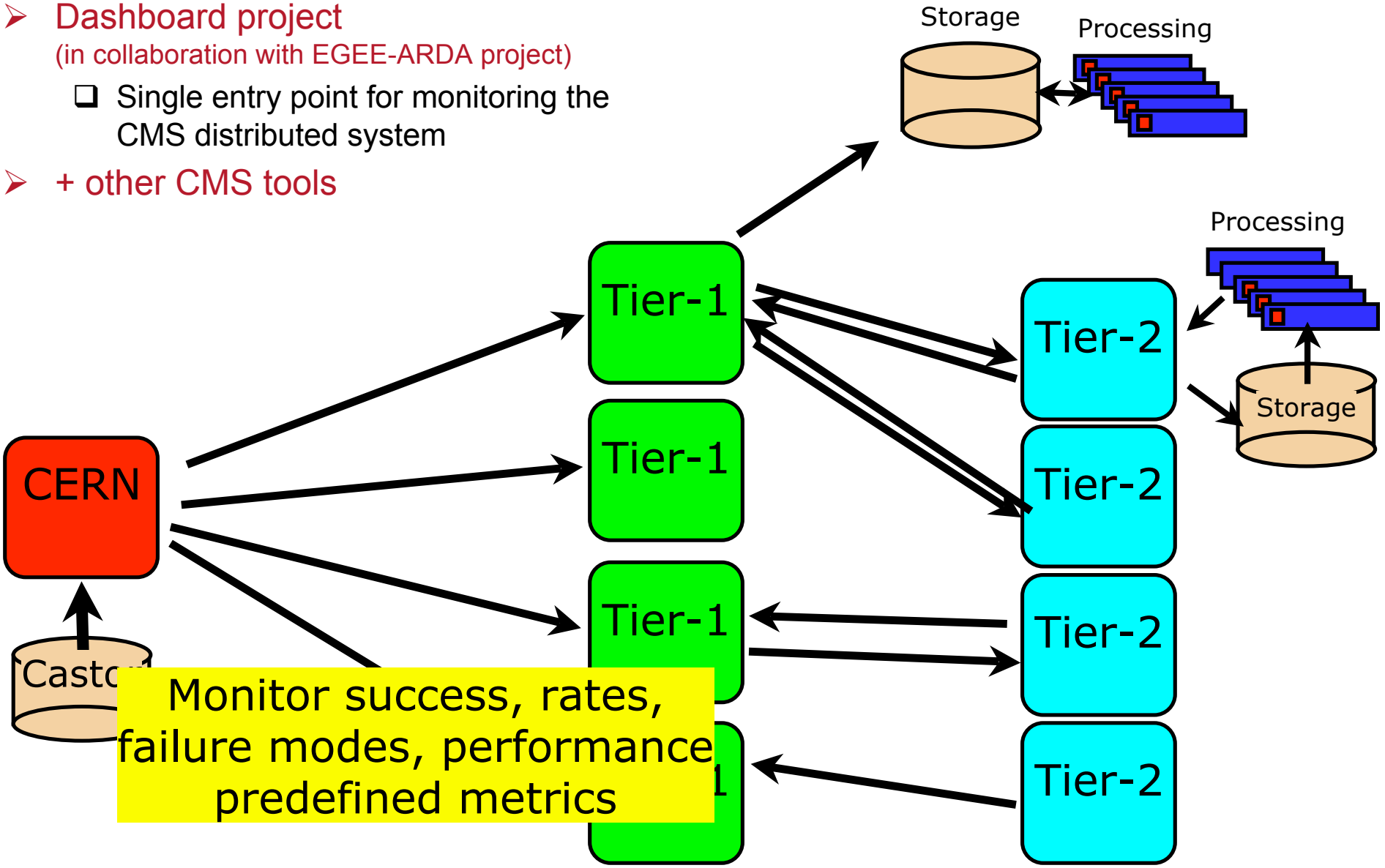
CMS CSA06: data access

- CMS calculates roughly to submit 200k jobs/day in 2008
  - scale factors applied for 2006 (and 2007)

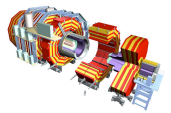


CMS CSA06: monitoring all processes

- Dashboard project
  (in collaboration with EGEE-ARDA project)
  - Single entry point for monitoring the CMS distributed system
- + other CMS tools



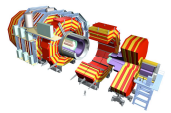
CSA06 summary



- Technical metrics were all met
  - some exceeded by large factors
  - strong engagement of CMS with WLCG community and sites
- Considerable work to do still
  - especially in the integration with data acquisition and on-line computing
  - development work to ease the operations load

- Offline sw
  - Release cycles: OK
  - Sw able to sustain >25% load for the prompt reco at T0
    - Sw error rate, performance, mem footprint well within expectations
      - Margin of further improvements on wider reco workflows
- Production and Grid tools
  - CMS met the very ambitious goals of 25M evts/month of sim event production
    - Improve workflow towards organized processing
- User Analysis workflow
  - Existing LCG and OSG middleware allowed to achieve 50k jobs/day
    - Integration and scale testing continues to be very important
  - CRAB-submitted jobs ran successfully on EGEE and OSG sites
- Data management
  - Tools are fine for the use-cases
  - Transfers: high Tiers participation and uptime, transfer quality and FTS reliability to be improved

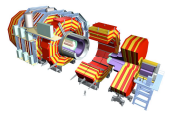
CSAs as scaling tests



- CSA07 as a check of "where we are", scaling to 2008.

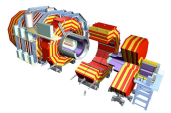
Task or "service"	CSA06 (the reality check)	2007 goal (CSA07 scope)	2008 goal
T0 reco rate	~40 Hz	100 Hz	150-300 Hz
T0→T1 transfers	~140-180 MB/s (continuous)	300 MB/s	600 MB/s
T1→T2 transfers	20-100 MB/s (bursts)	20-200 MB/s	50-500 MB/s
T1→T1 transfers	(not directly tested)	50 MB/s	100 MB/s
Job submission to T1's	(functionality tests only)	25k jobs/day	50k jobs/day
Job submission to T2's	30k-50k jobs/day (intergrated over all T2s)	75k jobs/day	150k jobs/day
MC simulation	25M evts/month	50M evts/month	$1.5\ 10^9$ evts/yr

Preparation activities towards CSA07



- Processing activities
  - Production for HLT and Physics Notes
    - 30 Mevts/month starting now
  - Development on MC Production System
  - CMS-specific tests in Site Availability Monitor (SAM) infrastructure
    - basic CMS analysis job that accesses a known dataset on sites, basic job workflows
  - Job Robot, a job load generator
    - "Updated" robots able to do a scale test and kindly step back (MCprod + analysis)
- Analysis activities
  - Start to estabilish analysis datasets at T2's
    - relies on PhEDEx forthcoming upgrades to give better local control
- Transfer activities
  - PhEDEx improvements
  - LoadTest07, a traffic load generator
    - Hot topics: T0→T1 (tape), T1↔T1 , T1↔T2
- Integration with online, and Global Data Taking (**P5→T0→T1**) tests
  - Includes testbeam data transfers, reconstruction, and access through the complete DM system

Summary



- CMS has adopted a distributed computing model that relies on Grid technologies
- CMS is steadily increasing in quality of tools, and scale and complexity of computing exercises

- Major changes in computing systems done in 2006
  - All tested in CSA06
    - DM, PhEDEx/FTS, processing framework/EDM, MC production system, …
  - Development needed in 2007, but not as wide as in 2006
    - e.g. migration to DBS-2, full Prod System architecture, …

- Challenges ahead?
  - Global Data Taking tests (spring)
  - CSA07 (summer)
  - Magnet Test and Cosmic Challenge 3 (autumn)
  - Engineering run (autumn?)
- And.. operations, operations, operations (CMS keyword for 2007)