CMS Computing Model:

Notes for a discussion with Super-B

Claudio Grandi
[CMS Tier-1 sites coordinator - INFN-Bologna]

Daniele Bonacorsi
[CMS Facilities Ops coordinator - University of Bologna]

Outline

The CMS distributed computing system

- from guiding principles to architectural design

Workflows (and actors) in CMS computing

- Computing Tiers
- A glance to Data Management (DM) and Workload Management (WM) components

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

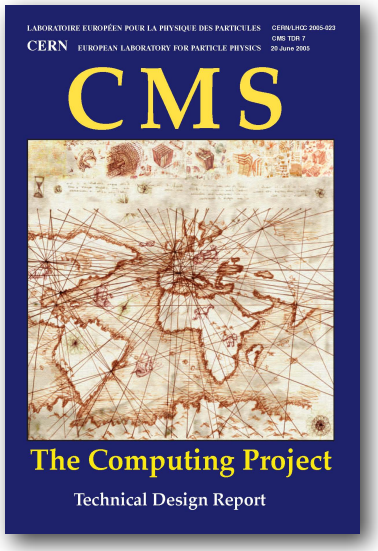
The CMS Computing Model

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-n, EGEE, OSG, NorduGrid, ...)

Several steps:

- 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
    - not “blueprint”, but “baseline” targets (+ development strategies)
  - hierarchy of computing tiers using WLCG services and tools
    - focus on Tiers role, functionality and responsibility
- Now partially “*old*” already?
  - ECoM group
    - To consider Evolution of Computing Model from Startup to Steady State (ECoM)
    - To digest and include the lessons learned



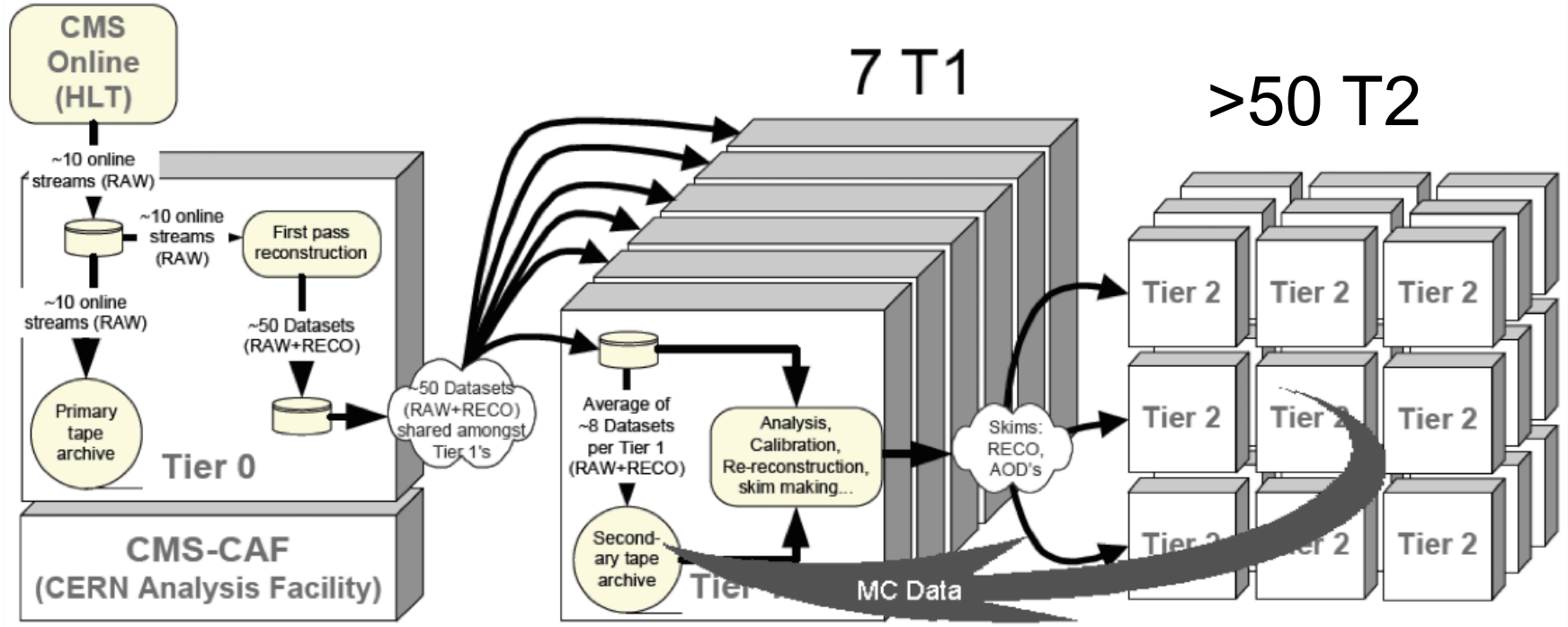
A Tiered architecture

T0 (CERN):

- Accepts data from DAQ
- Prompt reconstruction
- Data archive and distribution to T1's

CAF (CERN Analysis Facility for CMS):

- Access to full RAW datasets
- 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)

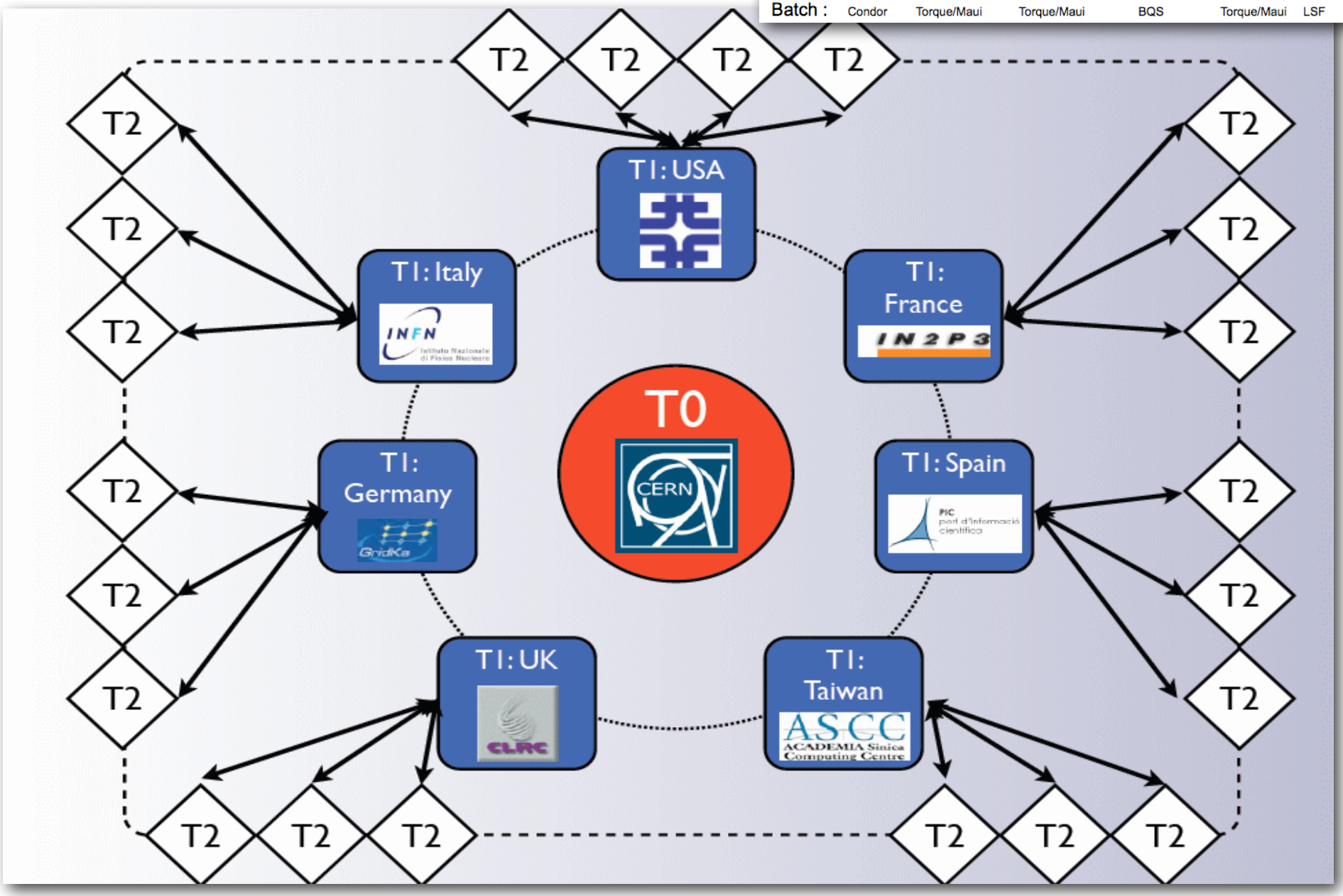


7 T1 centers and >50 T2 centers (and a growing nb of T3's…)

- See next slide

Towards a ‘mesh’ model

T1	FNAL	RAL	CCIN2P3	PIC	ASGC	INFN	FZK
Storage :	dCache/ Endstore	Castor	dCache/HPSS	dCache/ Endstore	Castor	Castor+ Storm	dCache/TSM
Batch :	Condor	Torque/Maui	Torque/Maui	BQS	Torque/Maui	LSF	PBSPro



T1/T2 roles

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
- Data Analysis
  - 5-10% of all processing is RAW data analysis, via special role

CMS T2 functions

- 50% user data analysis
  - Data processing for calib/align tasks and detector studies
  - Proficient data access via CMS+ WLCG services
- 50% MC event prod
  - both fast and detailed
- Import skimmed datasets from T1s
- Export MC data to T1s

A data-driven baseline

Baseline system with minimal functionality for first physics

- ‘Keep it simple!’
- Use Grid services as much as possible + add CMS-specific services if/where needed
- 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
  - Minimize job dependencies + site-local information remain site-local

T0-T1’s activities driven by data placement

- Data is partitioned by the experiment as a whole
- All data is placed at a site through explicit CMS policy
  - do not move around in response to job submission
- Leads to very ‘structured’ usage of Tier-0 and Tier-1
  - T0 and T1 are resources for the whole esperiment
  - 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

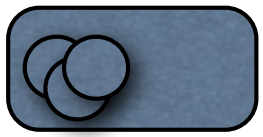
CMS expects to produce large amounts of data (evts)

- O(PB)/yr

Event data are in **files**

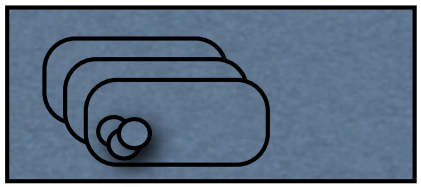
- average file size is kept reasonably large (≥ GB)
  - avoid scaling issues with storage systems and catalogues when dealing with too many small files
  - file merging also implemented and widely used in production activities
- O(10^6) files/year

Files are grouped in **fileblocks**



- group files in blocks (1-10 TB) for bulk data management reasons
  - exists as a result of either MC production or data movement
- 10^3 fileblocks/yr

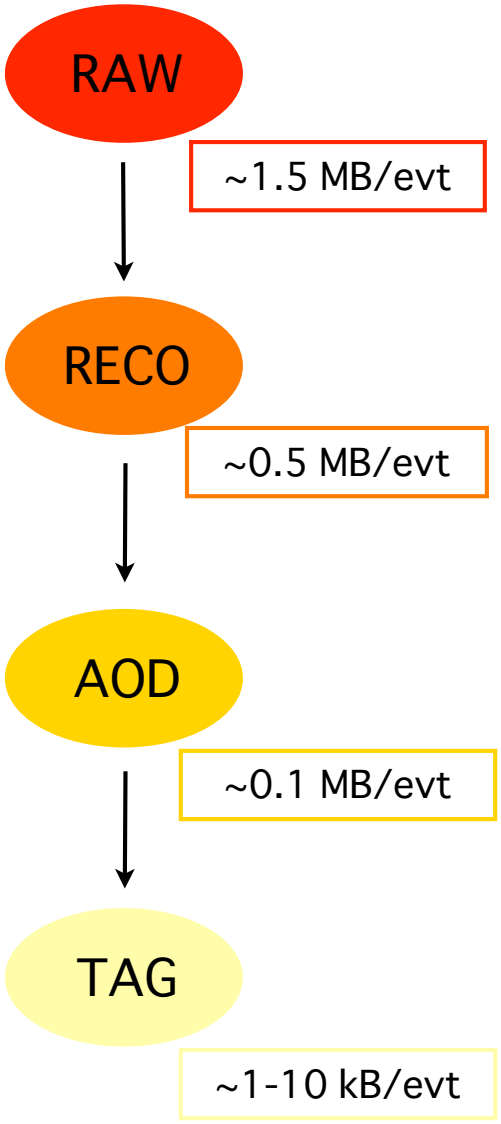
Fileblocks are grouped in **datasets**



- Datasets are large (100 TB) or small (0.1 TB)
  - Dataset definition is physics-driven (size as well)

Data types

Data types/volumes as input parameters for the model



RAW

- Triggered evts recorded by DAQ
  - 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
  - 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, …
  - Whole set copied to each T1, large fraction copied to T2

TAG

- Fast selection info
  - Relevant info for fast evt selection in AOD

Plus MC in ~ N:1 (N>=1) ratio with data

2009/10 Data Taking

- We are all eagerly waiting for the first collisions.
- Estimated the data volume for proton-proton collision for upcoming year:
  - 70 days running in 2009-10,

    assuming 10 month LHC running with 40% availability, 8h fills and 5h turn-around.
  - 300Hz rate in physics stream
  - assume 26% mean overlap
  - $2.3 \cdot 10^9$ events
  - 3.3 PB RAW data (1.5MB/evt)
  - 1.1 PB RECO (0.5MB/evt)
  - 220 TB AOD (0.1MB/evt)
  - We will have multiple copies of the AOD at Tier-1's.
  - We will have multiple re-reco passes.

CMS Data Management

Provide tools to discover, access and transfer event data in a distributed computing environment

- Track and replicate data with a granularity of file blocks
- Minimize the load on catalogues

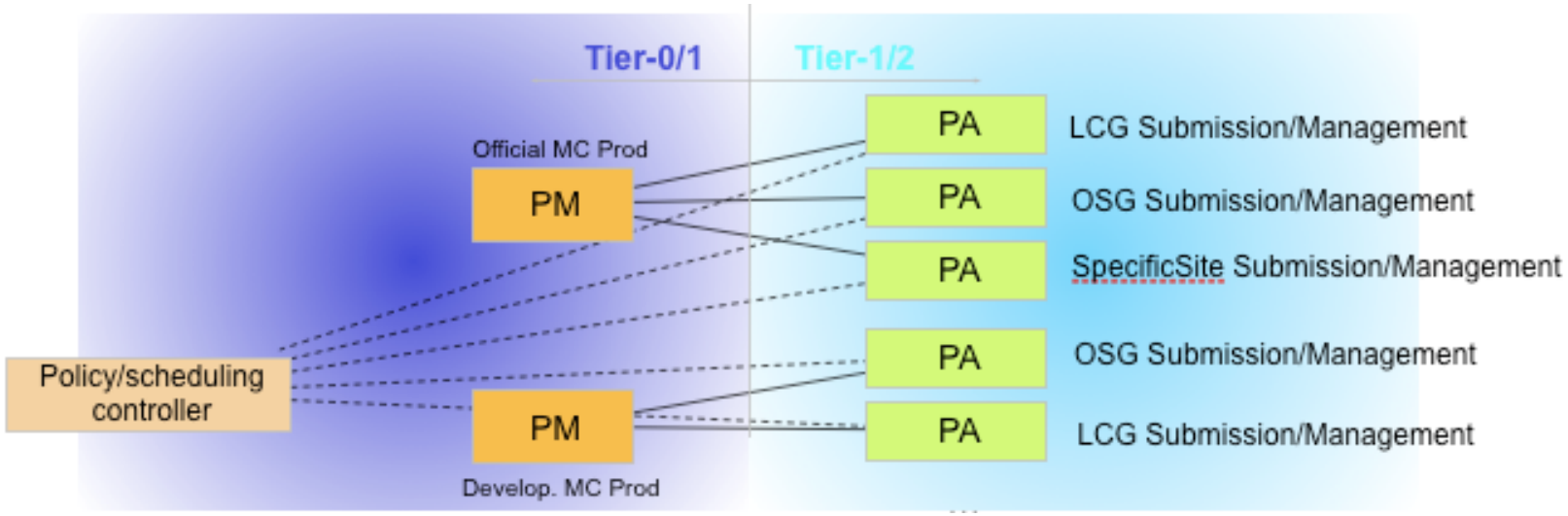
The ‘logical’ components:

- 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 (Physics Experiment Data Export)
  - integration with most recent EGEE transfer services

CMS MC production system

Current MC production system in production since 2006

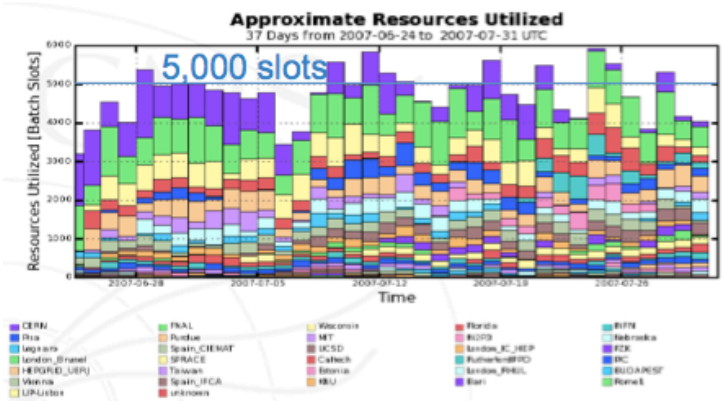
- Overcome previous inefficiencies + introduce new capabilities
  - less man-power consuming, better handling of Grid-sites unreliability, better use of resources, automatic retrials, better error report/handling
- Flexible and automated architecture
  - ProdManager (PM) (+ the policy piece)
    - manage 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)



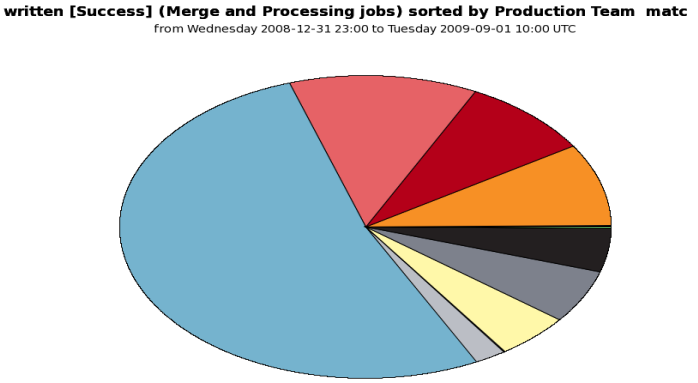
CMS MC production operations

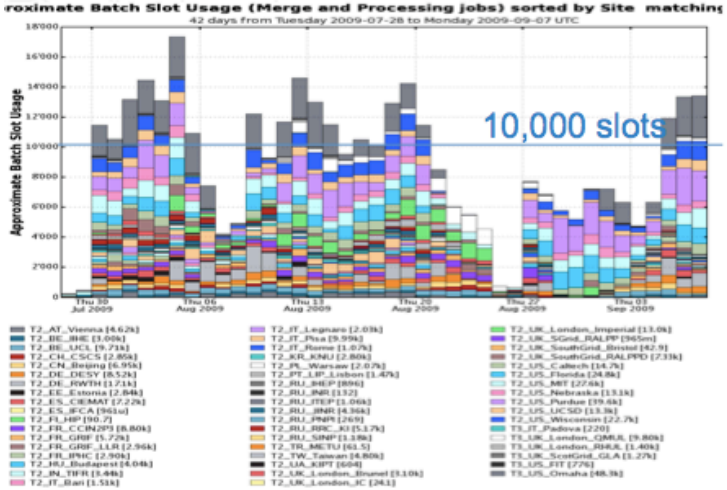
MC prod operations strategy changed:

- 2006-07 : 6 “regional teams” managed by a central manager
- 2008-09 : 1 central team (6 people) managing submissions in defined “T1-regions”



- Summer 2007 :
  - <4700 slots> utilized
  - including 32 T1s and T2s
  - ➔ 75% of total available then





- Summer 2009 :
  - <9-10,000 slots> utilized
  - including 53 T2s and T3s
  - ➔ 60% of total T2 available then
  - ➔ rest of T2 resources used by analysis

CMS Data Placement system

Physics Experiment Data Export (**PhEDEx**)

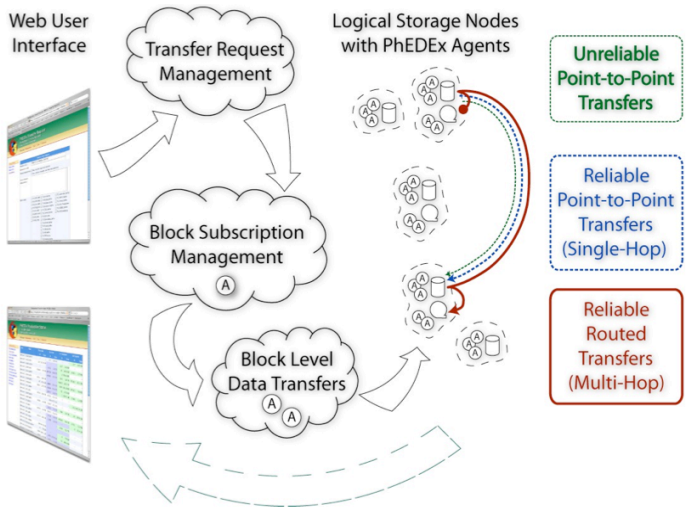
- Large-scale reliable and scalable dataset/fileblock replication
- multi-hop routing following a transfer topology (T0-T1-T2-T3's), data pre-stage from tape, monitoring, bookkeeping, priorities and policy, ...

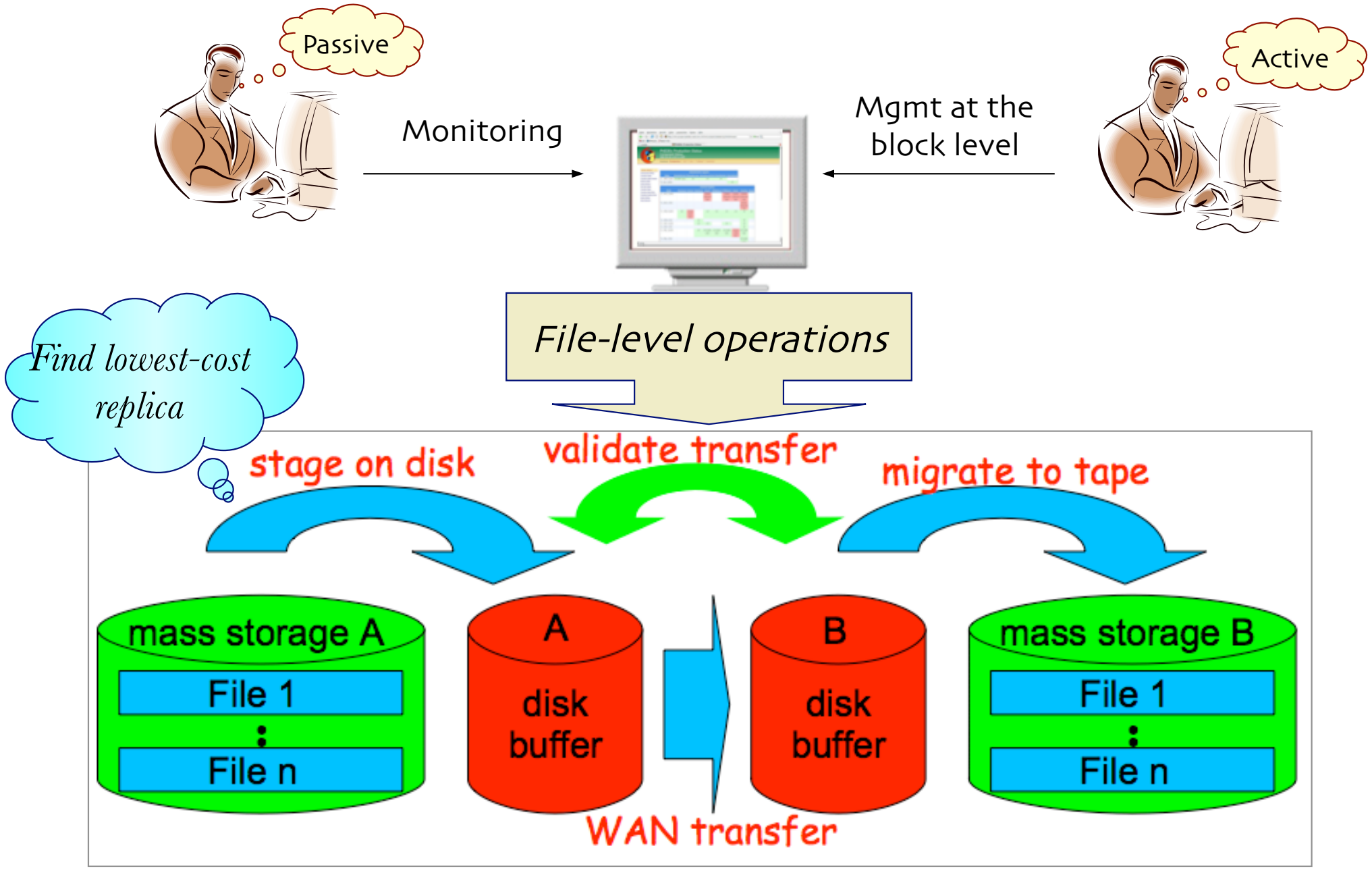
In production since almost 2004

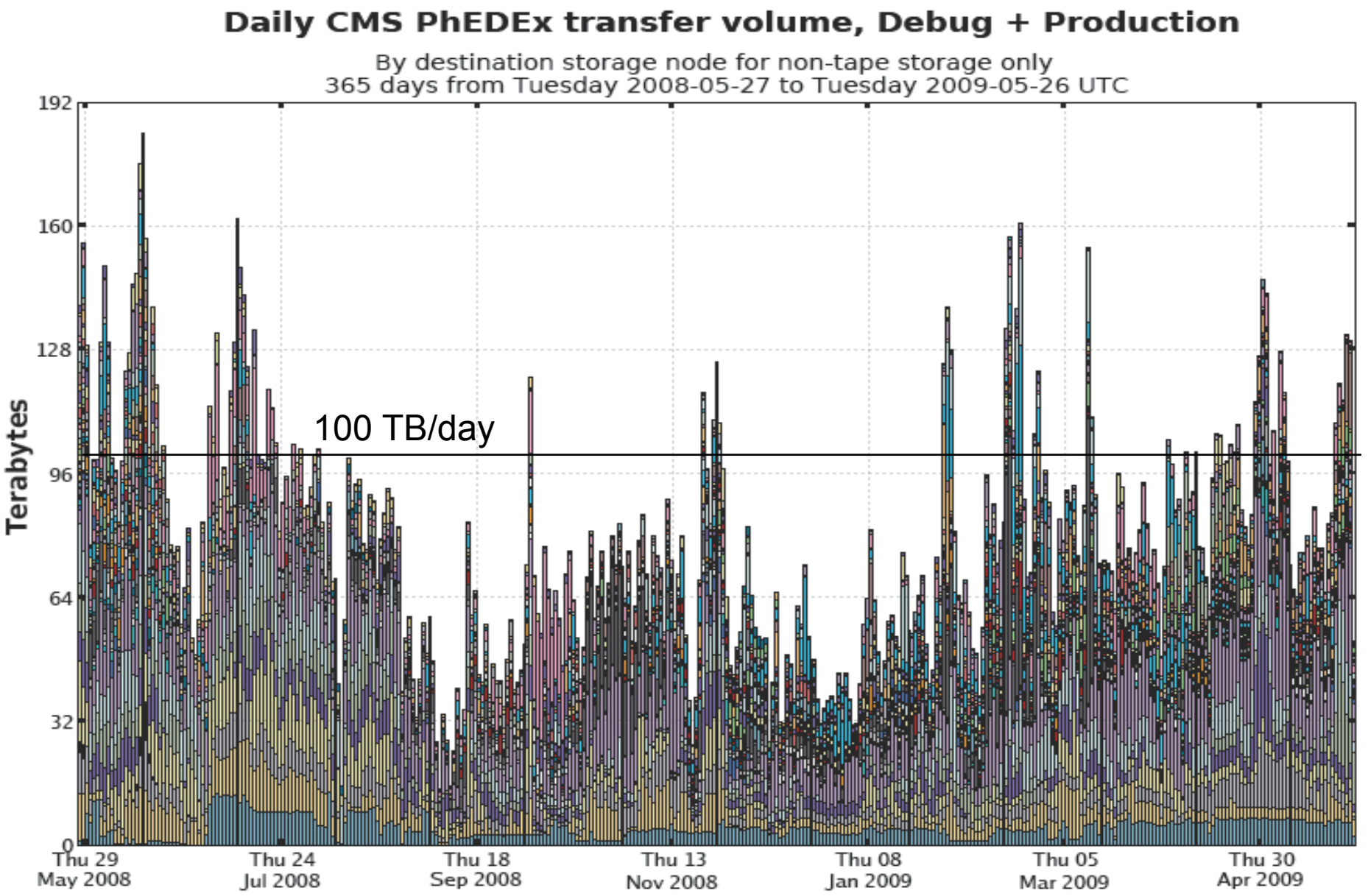
- In the hall-of-fame in terms of mature and high-quality production services for LHC experiments
- Managing transfers of several TB/day

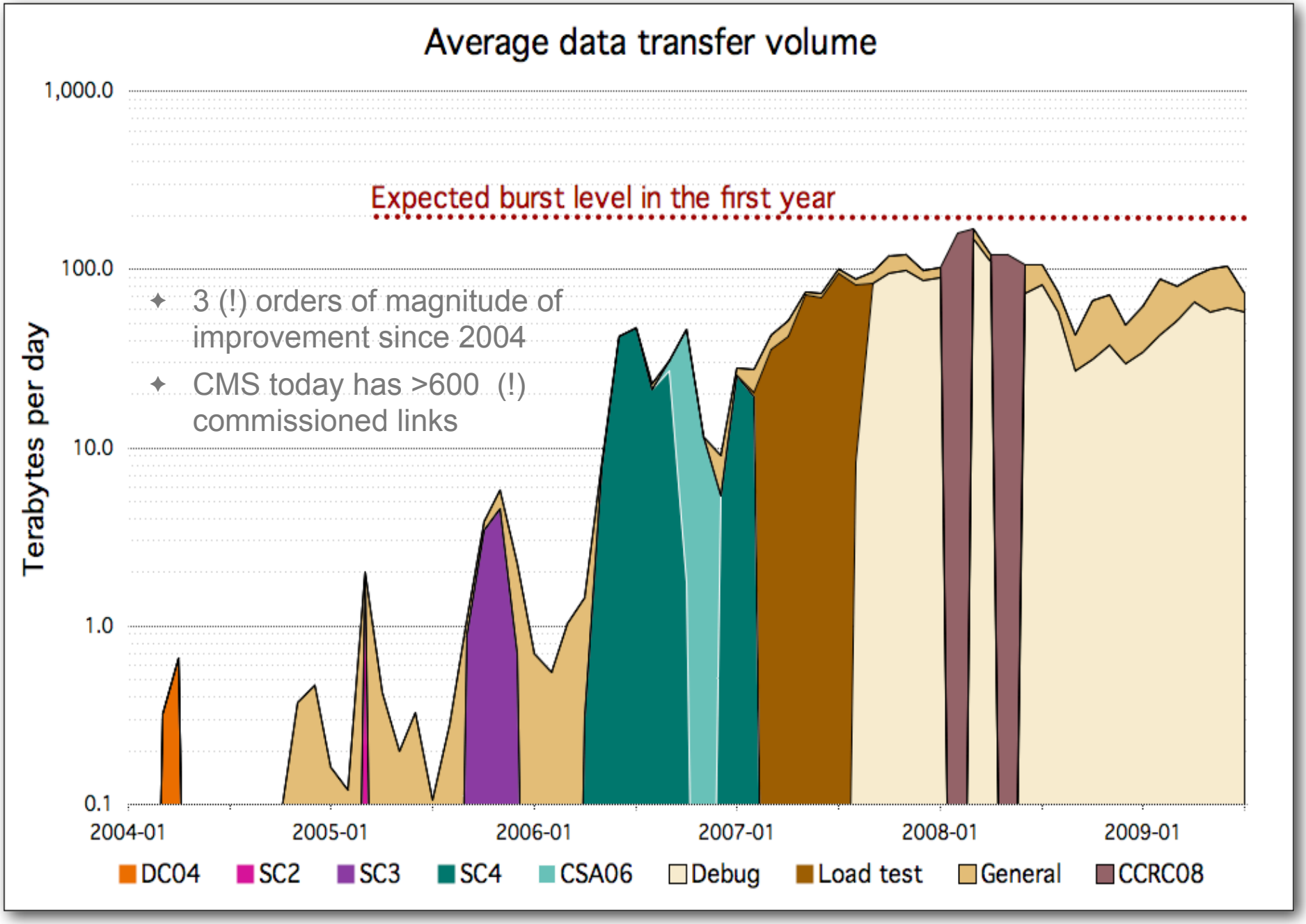
PhEDEx integration with EGEE services

- gLite 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



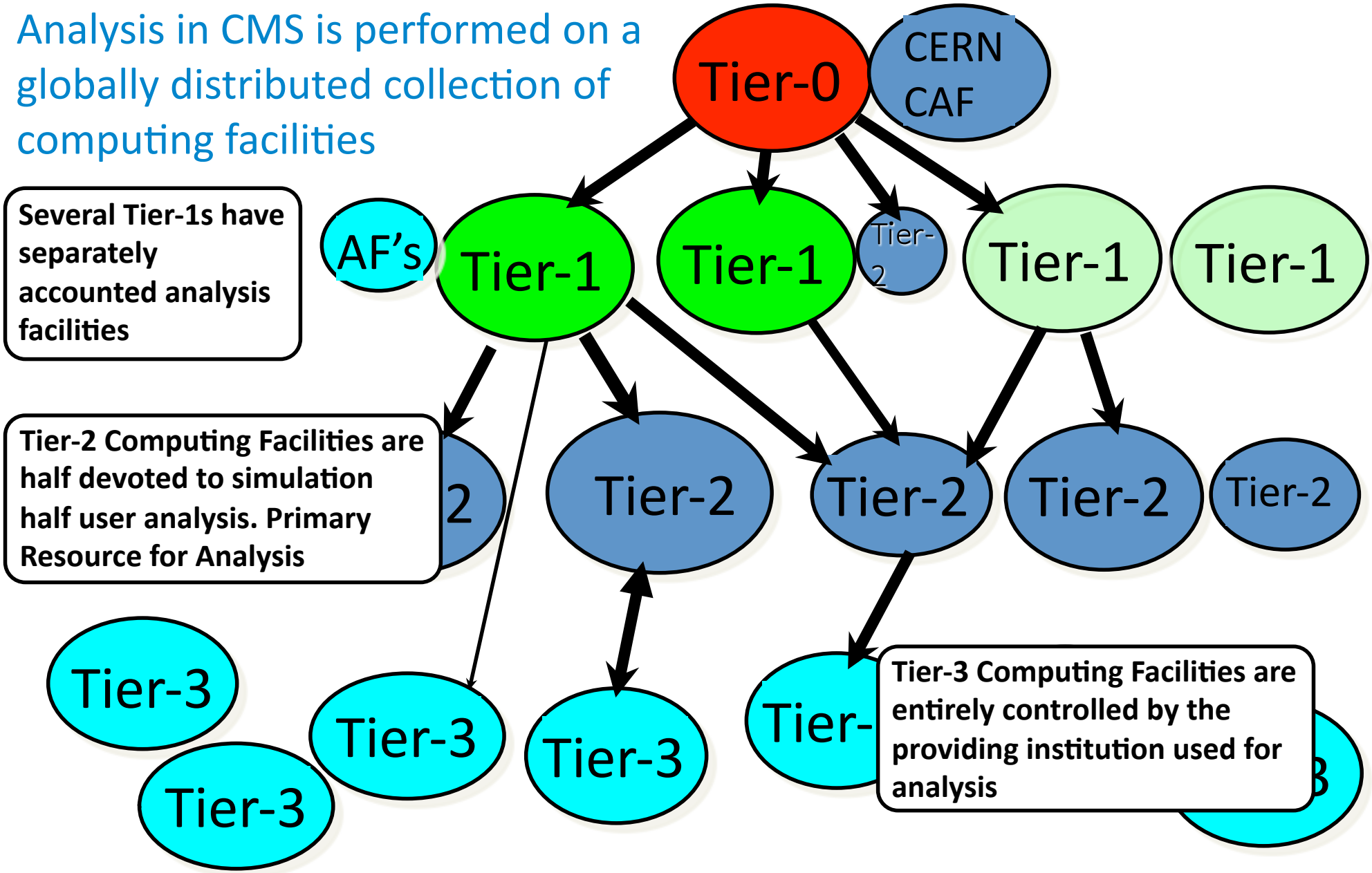

Passive
Monitoring
Mgmt at the block level
Active
File-level operations
Find lowest-cost replica
stage on disk
validate transfer
migrate to tape
mass storage A
File 1
File n
A
disk buffer
B
disk buffer
mass storage B
File 1
File n
WAN transfer


Daily CMS PhEDEx transfer volume, Debug + Production
By destination storage node for non-tape storage only
365 days from Tuesday 2008-05-27 to Tuesday 2009-05-26 UTC
Terabytes
192
160
128
96
64
32
0
100 TB/day
Thu 29 May 2008
Thu 24 Jul 2008
Thu 18 Sep 2008
Thu 13 Nov 2008
Thu 08 Jan 2009
Thu 05 Mar 2009
Thu 30 Apr 2009


Average data transfer volume
1,000.0
100.0
10.0
1.0
0.1
Terabytes per day
Expected burst level in the first year
3 (!) orders of magnitude of improvement since 2004
CMS today has >600 (!) commissioned links
2004-01
2005-01
2006-01
2007-01
2008-01
2009-01
DC04
SC2
SC3
SC4
CSA06
Debug
Load test
General
CCRC08

CMS Analysis resources

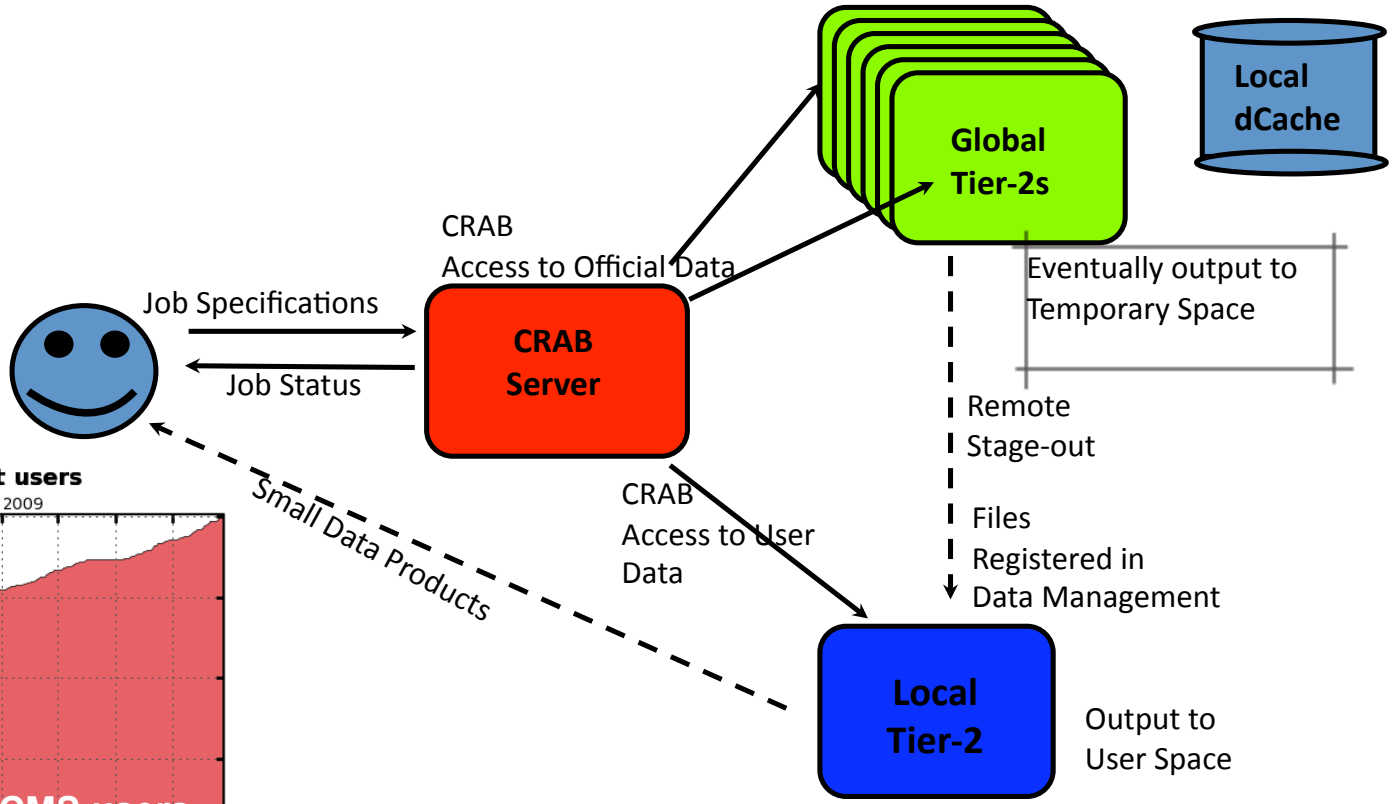
Analysis in CMS is performed on a globally distributed collection of computing facilities

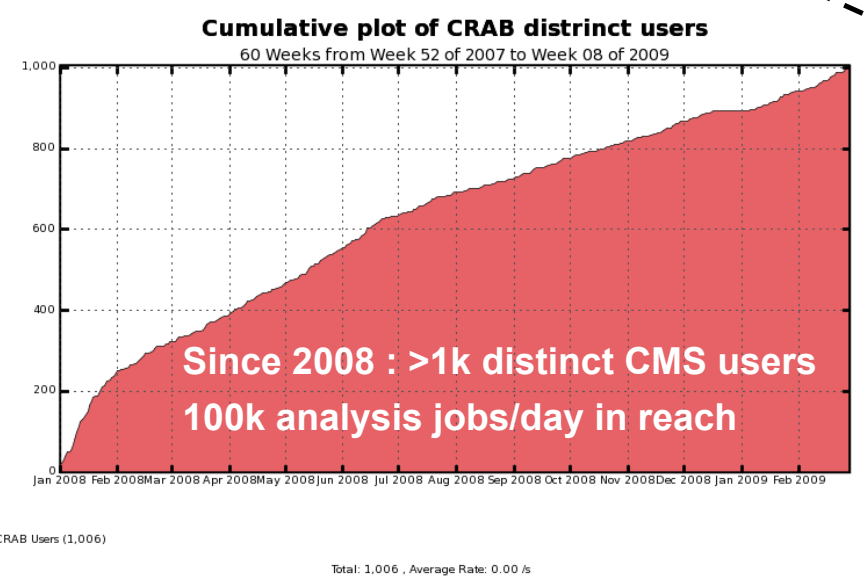


CMS Distributed Analysis on the Grid

CMS Remote Analysis Builder (CRAB)

- Tool for job preparation, submission and monitoring
- Satisfies the needs of CMS users
- Better resource control usage via the CRAB Analysis Server





CMS Tier-2 Disk Space management

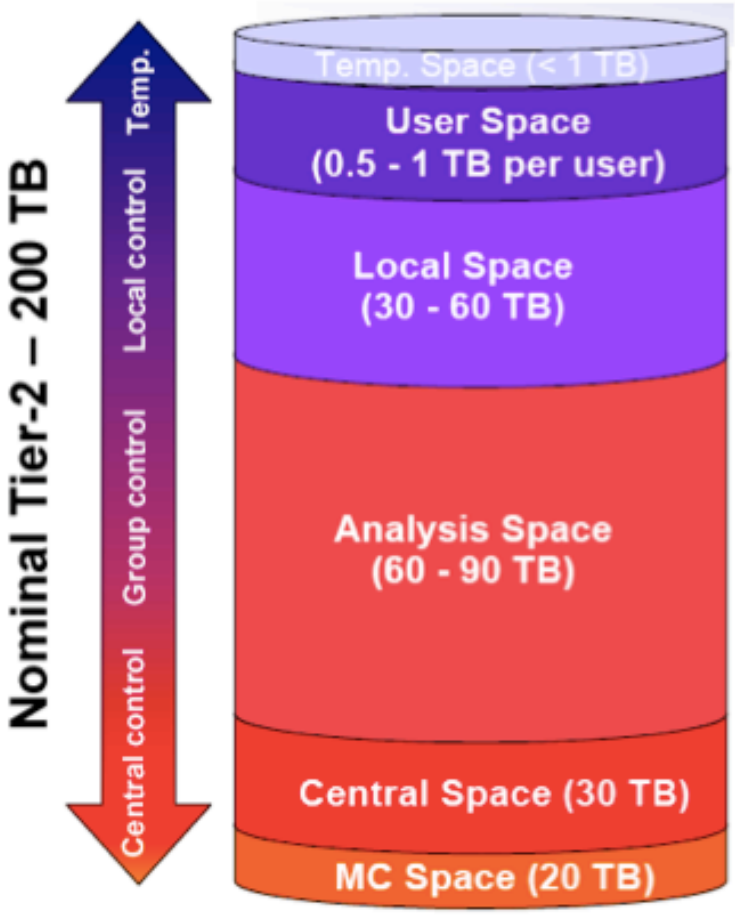
In CMS jobs go to the data : distribute data broadly

CMS attempts to share management of the space across groups

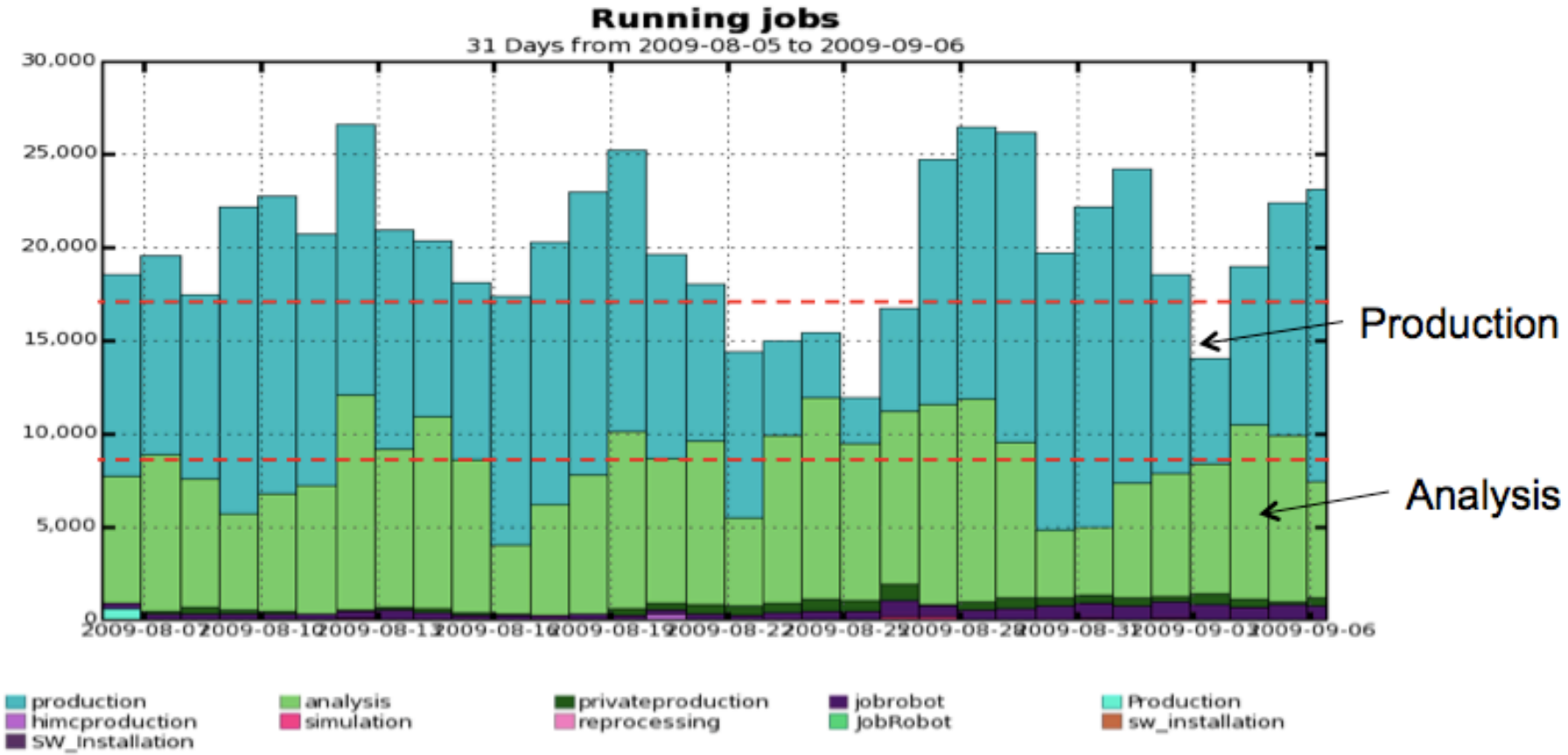
- Ensures people doing the work have some control

200TB of disk space at a nominal Tier-2

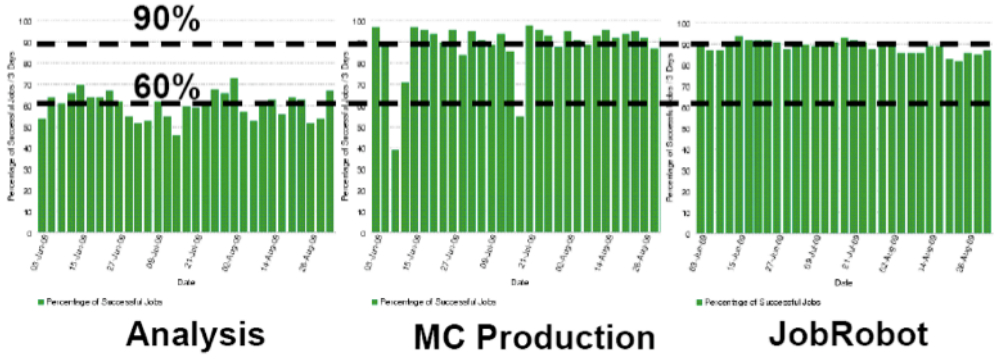
- 20 x 1TB is identified for storing local user produced files and making them grid accessible
- 30TB is identified for use by the local group
- 2-3 x 30 TB reserved to CMS PH Analysis groups
- 30 TB for centrally managed Analysis Operations expect to be able to host most RECO data in 1sr y.
- 20 TB of space for DataOps for MC staging buffer



Job Slot utilization for Analysis



- Current CMS total CPU pledge at T2s : 17k jobs slots
- Analysis pledge : 50%
- Utilization in August was reasonable



ECoM

A group to consider Evolution of Computing Model from Startup to Steady State

- re-examine the CMS Computing Model using various different use-cases that may occur during startup and before "steady state" is reached.
- revisit the utilization of resources, the pattern and distribution of data and exactly what happens where and when.
- several use-cases should be explored in order to understand what flexibility and agility is available ahead of the startup so as to be able to make mid-course corrections as required

A work in progress.

Back-up

CMS Site Commissioning

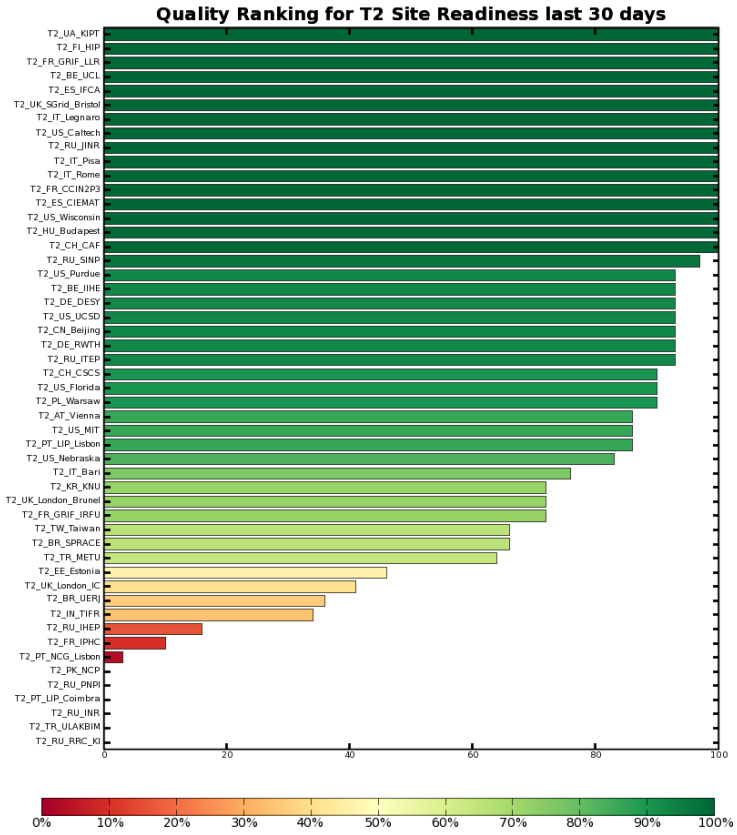
- Objectives
  - Test all functionality required from CMS at each site in a continuous mode
  - Determine if the site is usable and stable
- What is tested?
    - Job submission
    - Local site configuration and CMS software installation
    - Data access and data stage-out from batch node to storage
    - "Fake" analysis jobs
    - Quality of data transfers across sites
- What is measured?
  - Site availability: fraction of time all functional tests in a site are successful
  - Job Robot efficiency: fraction of successful "fake" analysis jobs
  - Link quality: number of data transfer links with an acceptable quality
- What is calculated?
  - A global estimator which expresses how good and stable a site is

Statistics and plots

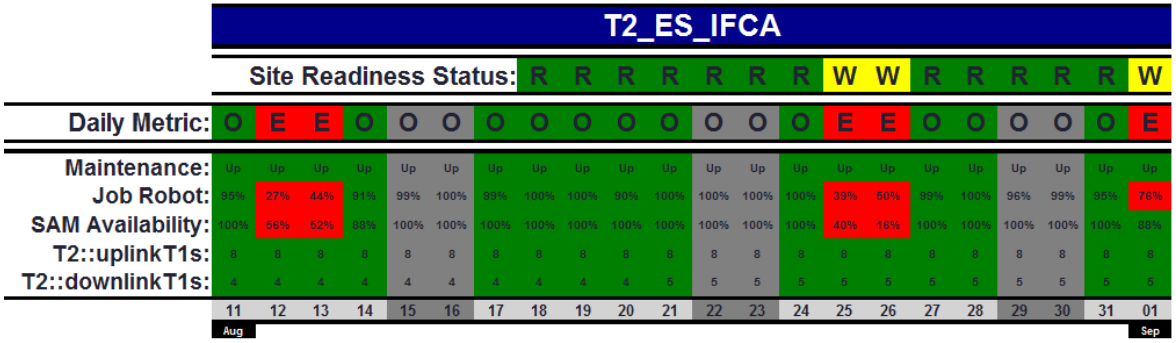
Site summary table

Site Name	SiteComm JR	Commissioned Links (expand this column)	Site availability	SiteReadiness Status	Maintenance in SAM	Good links (expand this column)
T0_CH_CERN	96%(500)	n/a	88%	n/a	n/a	n/a
T1_CH_CERN	n/a	combined	88%	n/a	n/a	combined
T1_DE_FZK	94%(600)	combined	96%	R	n/a	combined
T1_ES_PIC	99%(500)	combined	100%	R	n/a	combined
T1_FR_CCIN2P3	99%(600)	combined	100%	NR	n/a	combined
T1_IT_CNAF	99%(600)	combined	100%	R	n/a	combined
T1_TW_ASGC	98%(601)	combined	100%	R	n/a	combined
T1_UK_RAL	99%(700)	combined	100%	R	n/a	combined
T1_US_FNAL	100%(700)	combined	100%	R	n/a	combined
T2_AT_Vienna	95%(500)	combined	9[illegible]%	R	n/a	combined
T2_BE_IIHE	90%(332)	combined	8[illegible]%	R	n/a	combined
T2_BE_UCL	100%(600)	combined	10[illegible]%	R	n/a	combined
T2_BR_SPRACE	97%(700)	combined	9[illegible]%	R	n/a	combined
T2_BR_UERJ	0%(597)	combined	0%	NR	n/a	combined
T2_CH_CAF	n/a	combined	n/a	R	n/a	n/a
T2_CH_CSCS	84%(600)	combined	0%	W	All services in maint	combined
T2_CN_Beijing	98%(600)	combined	9[illegible]%	R	n/a	combined
T2_DE_DESY	99%(501)	combined	9[illegible]%	R	Some CE in maint.	combined
T2_DE_RWTH	98%(500)	combined	10[illegible]%	R	n/a	combined
T2_EE_Estonia	99%(400)	combined	88%	NR	n/a	combined
T2_ES_CIEMAT	100%(600)	combined	100%	R	n/a	combined
T2_ES_IFCA	76%(592)	combined	88%	W	n/a	combined
T2_FI_HIP	n/a	combined	100%	R	Some CE in maint.	combined
T2_FR_CCIN2P3	n/a	combined	100%	NR	n/a	combined
T2_FR_GRIF_IRFU	48%(284)	combined	0%	NR	n/a	n/a
T2_FR_GRIF_LLR	100%(501)	combined	100%	R	n/a	combined

Site ranking



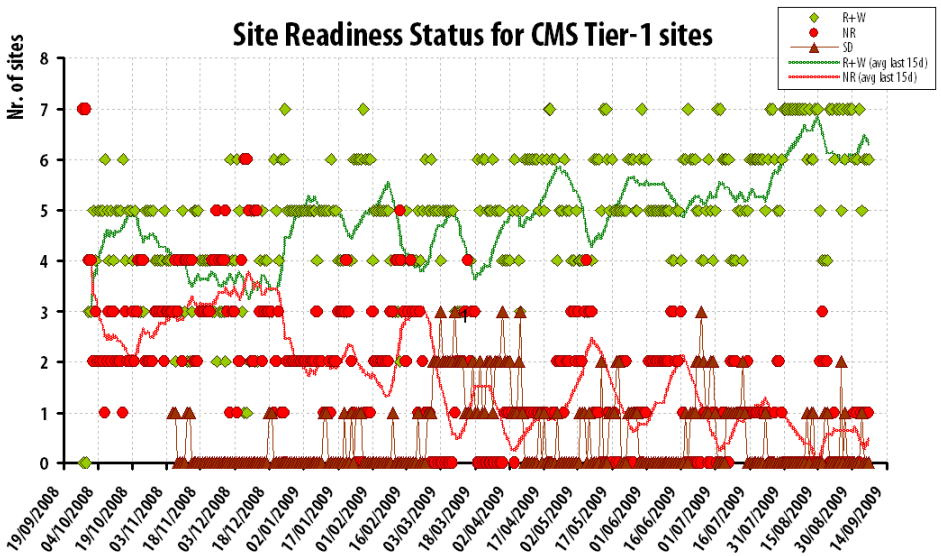
Site history

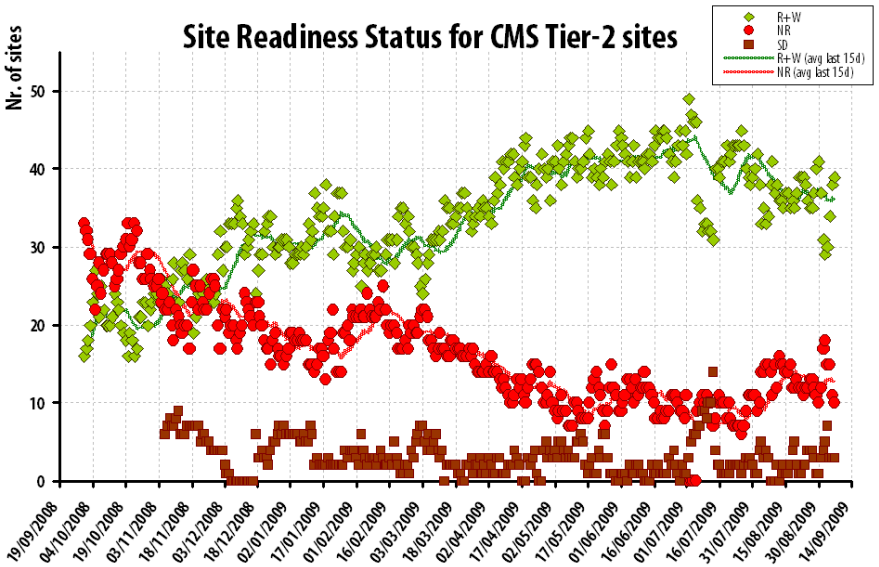


The CMS Site Commissioning team

How can this be used?

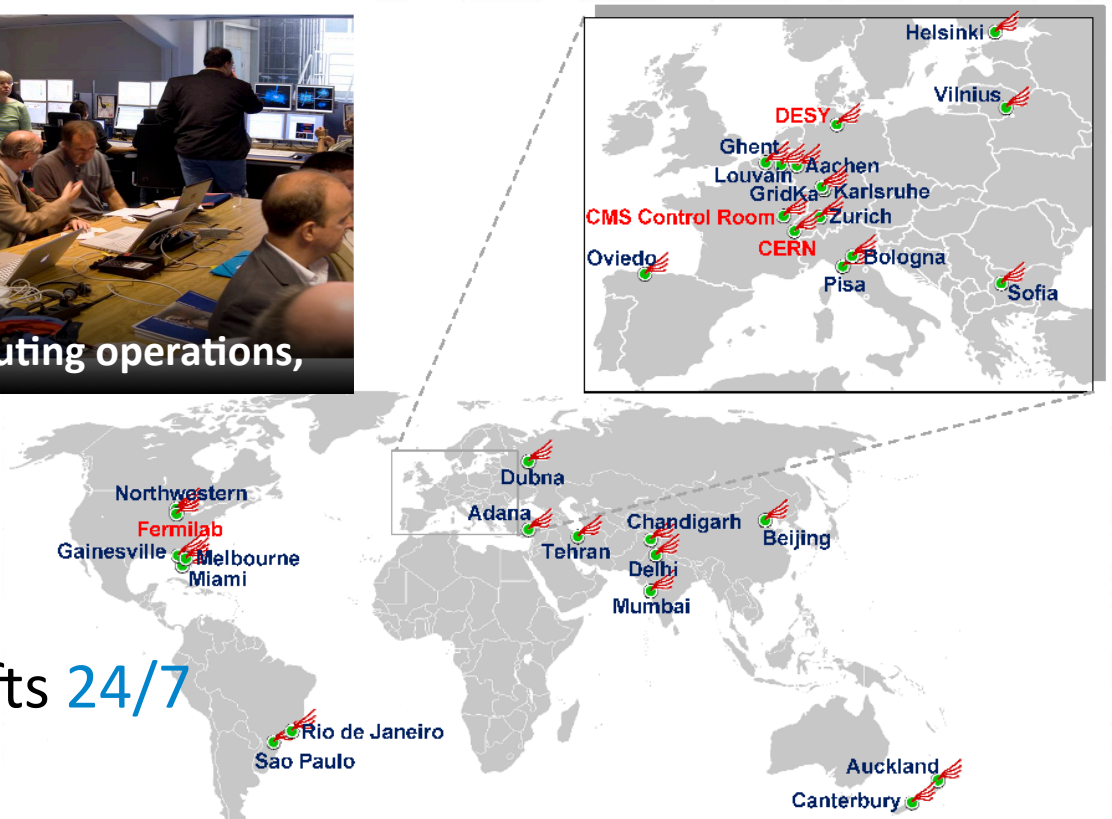
- To measure global trends in the evolution of the reliability of sites
  - Impressive results in the last year
- Weekly reviews of the site readiness
- Production teams can better plan where to run productions
- Automatically map to production and analysis tools ?





The CMS Site Commissioning team

CMS Centers and Computing Shifts



- CMS running Computing shifts 24/7
- Encourage remote shifts
- Main task: monitor and alarm CMS sites & Computing Experts