# CMS Data Distribution Model

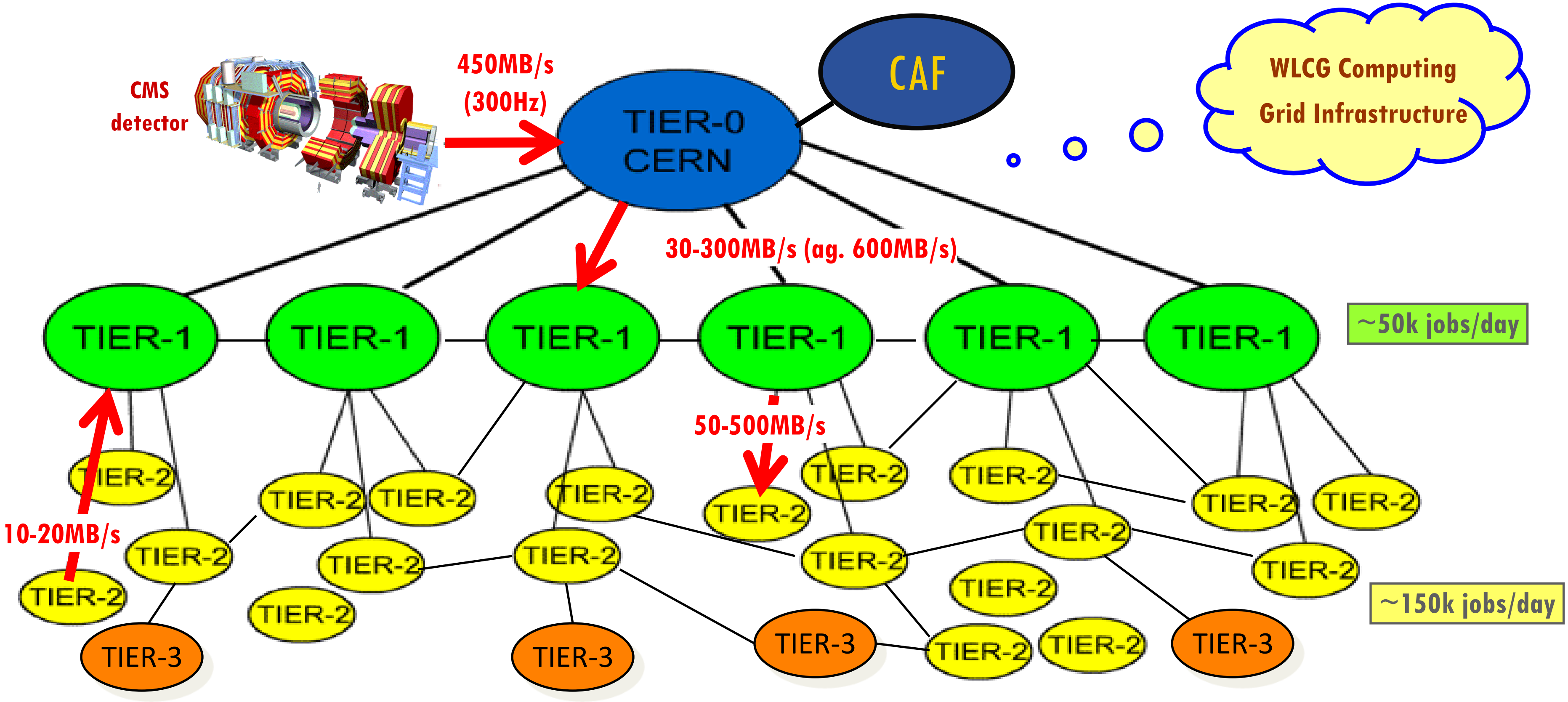
Peter Kreuzer, RWTH Aachen

CMS France Physique meeting
Strassbourg, May 27, 2009

# Outline

- CMS Tiered Computing model
- CMS Data Distribution Model and Performances
  - Specificities of Tier-0, Tier-1, Tier-2
- CMS Analysis Model
- The Role of the CAF(s)
- Distributed Computing Operations
- Challenges/Evolution of the model until physics running

# CMS tiered Computing Model



## Tier-0

Prompt Reconstruction
Archival of Raw
and First RECO data
Calibration Streams (CAF)
Data Distribution → Tier-1

## 7 Tier-1s

Re-Reconstruction
Skimming
Second Archival of RAW
Served Copy of RECO
Archival of Simulation
Data Distribution → Tier-2

## ~50 Tier-2s

Primary Resources for
Physics Analysis and
Detetector Studies by
users
MC Simulation → Tier-1

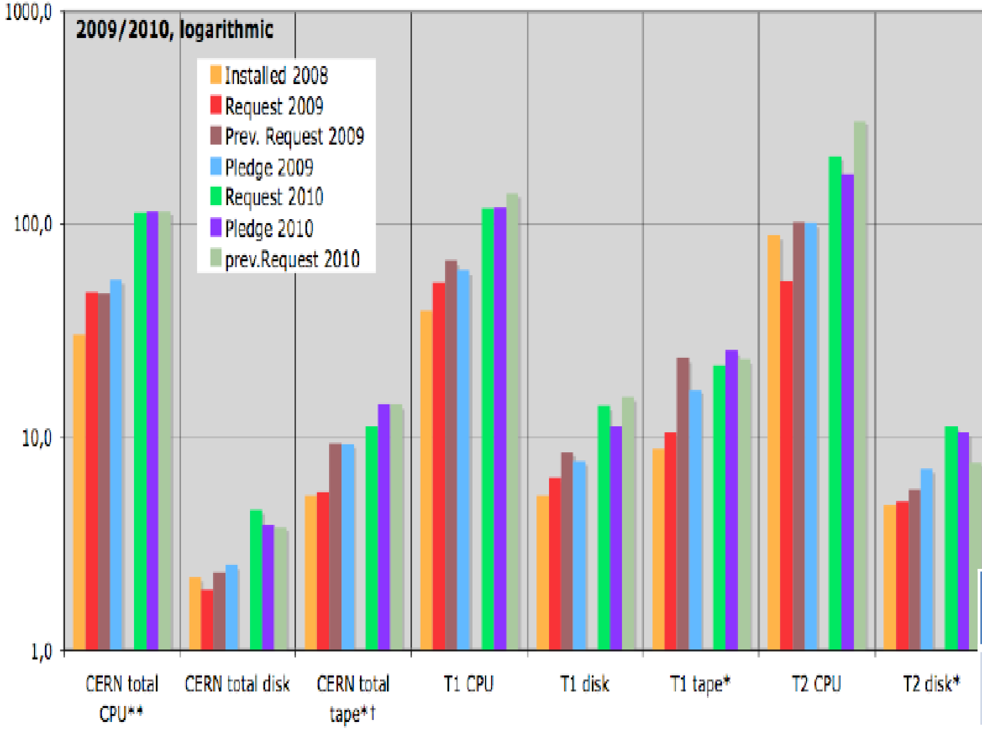
# CMS Data Formats and Comp. Resources

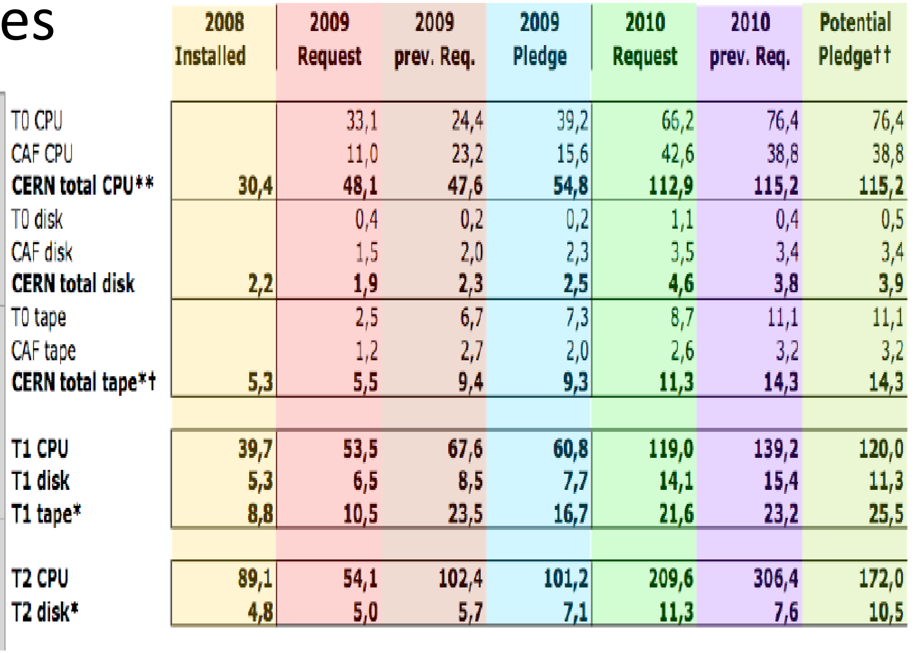
- Data Tiers

(with safety margin)

| Data Tier | RAW | RECO | AOD |
|---|---|---|---|
| <Size> [MB] | 1.5 | 0.5 | 0.1 |

- Data Management Units
  - DataFile : <1.8 GB> , DataBlock : <33 GB>
- CMS WLCG Computing Resources

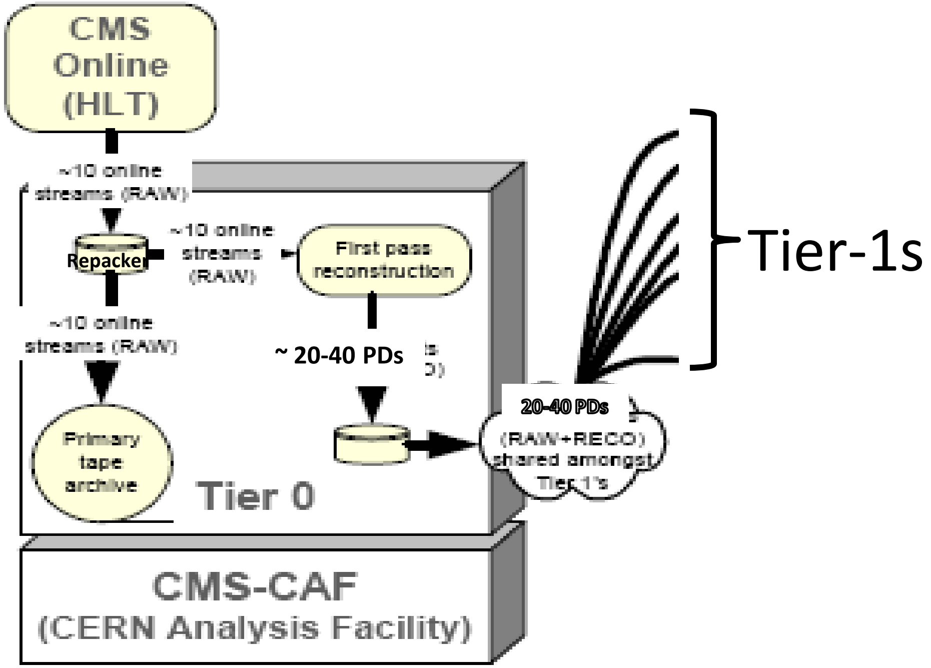




| | 2008 Installed | 2009 Request | 2009 prev. Req. | 2009 Pledge | 2010 Request | 2010 prev. Req. | Potential Pledge†† |
|---|---|---|---|---|---|---|---|
| T0 CPU | | 33,1 | 24,4 | 39,2 | 66,2 | 76,4 | 76,4 |
| CAF CPU | | 11,0 | 23,2 | 15,6 | 42,6 | 38,8 | 38,8 |
| **CERN total CPU**** | **30,4** | **48,1** | **47,6** | **54,8** | **112,9** | **115,2** | **115,2** |
| T0 disk | | 0,4 | 0,2 | 0,2 | 1,1 | 0,4 | 0,5 |
| CAF disk | | 1,5 | 2,0 | 2,3 | 3,5 | 3,4 | 3,4 |
| **CERN total disk** | **2,2** | **1,9** | **2,3** | **2,5** | **4,6** | **3,8** | **3,9** |
| T0 tape | | 2,5 | 6,7 | 7,3 | 8,7 | 11,1 | 11,1 |
| CAF tape | | 1,2 | 2,7 | 2,0 | 2,6 | 3,2 | 3,2 |
| **CERN total tape*†** | **5,3** | **5,5** | **9,4** | **9,3** | **11,3** | **14,3** | **14,3** |
| **T1 CPU** | **39,7** | **53,5** | **67,6** | **60,8** | **119,0** | **139,2** | **120,0** |
| **T1 disk** | **5,3** | **6,5** | **8,5** | **7,7** | **14,1** | **15,4** | **11,3** |
| **T1 tape*** | **8,8** | **10,5** | **23,5** | **16,7** | **21,6** | **23,2** | **25,5** |
| **T2 CPU** | **89,1** | **54,1** | **102,4** | **101,2** | **209,6** | **306,4** | **172,0** |
| **T2 disk*** | **4,8** | **5,0** | **5,7** | **7,1** | **11,3** | **7,6** | **10,5** |

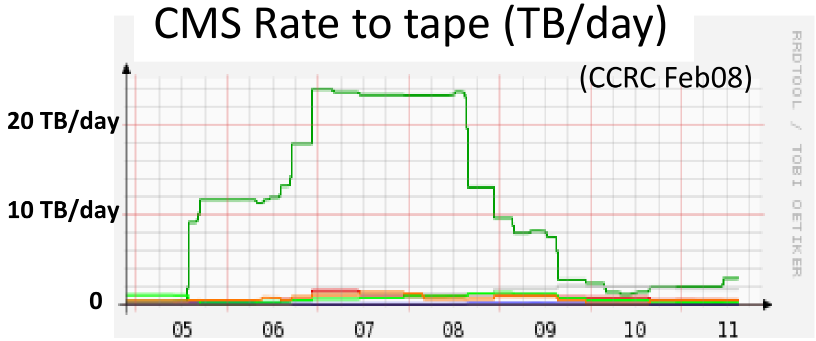
| Resource | $\text{CPU}_{2009}$ | $\text{CPU}_{2010}$ | $\text{Disk}_{2009}$ | $\text{Disk}_{2010}$ |
|---|---|---|---|---|
| CERN/rest | 30% | 25% | 15% | 15% |

# CMS Data Distribution @ Tier-0



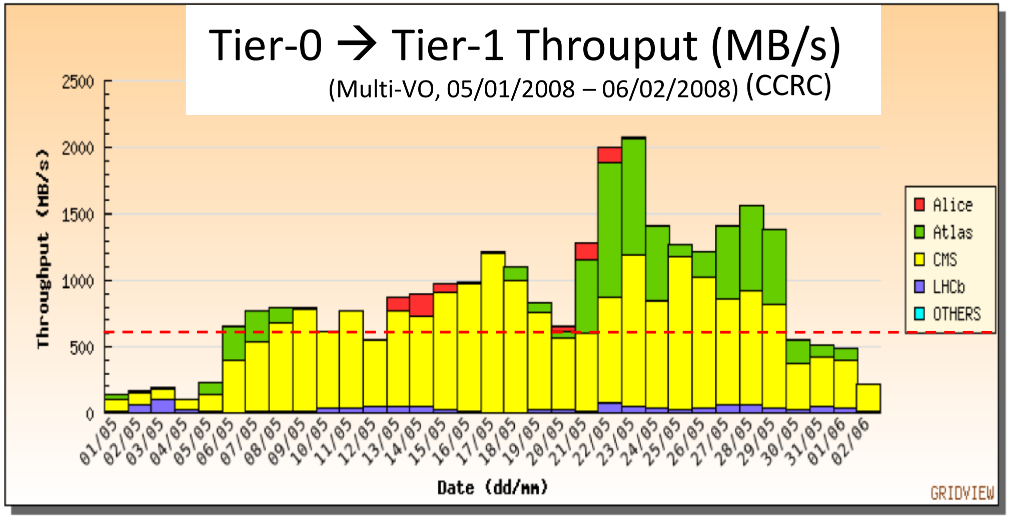
- Online (HLT) → Offline (Tier-0) data flow : Online Streams
- Online Streams contain Physics data or Express (Calibration) streams
- Repacker assembles Primary Datasets (PD) based on Trigger Path
- RAW PDs sent to tape and to Tier-1 for Archival copies + to Tier-0 buffer for the First Reconstruction
- Special Calibration PDs (AlcaReco) are prepared in short (1h) latency and sent to the CAF for Calibration and Alignment, result fed back into Condition DB as input to the First Reconstruction (24h cycle)
- PDs are the basis of the Data Distribution to Tier-1 (depending on the PD size and on each particular Tier-1 capacity)

# Tier-0 Workflow performances


CMS Rate to tape (TB/day)

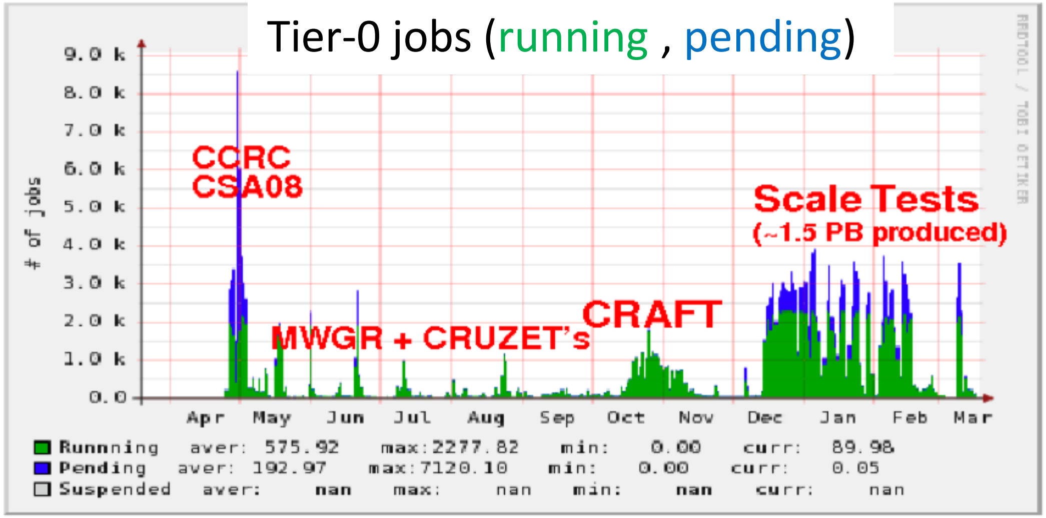
(CCRC Feb08)

- 270 MB/s sustained rates
- 1.5 GB/s peak rates
- Aim for x 2-3 higher (STEP09)


Tier-0 → Tier-1 Throuput (MB/s)

(Multi-VO, 05/01/2008 – 06/02/2008) (CCRC)

- CMS regularly above design rate


Tier-0 jobs (running , pending)

CCRC CSA08

MWGR + CRUZET's

CRAFT

Scale Tests (~1.5 PB produced)

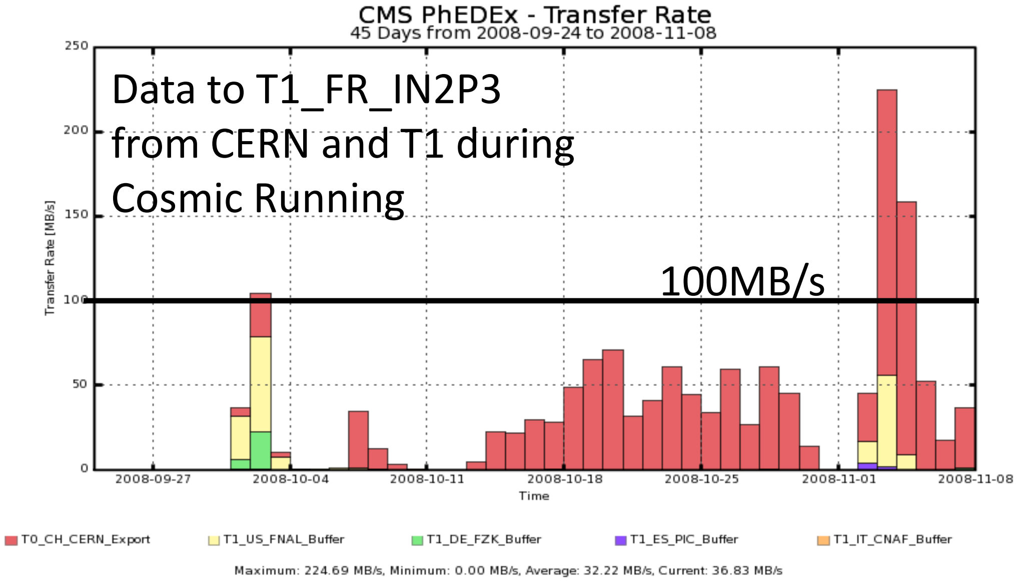
- Tier-0 processing challenges 2008-2009

- Infrastructure used in Cosmic Data Taking, Challenge Activities, Weekly Global Runs and Scale Testing
- Integrated production of RAW PDs, reconstruction, calibration samples, express stream and transfer to Tier-1s

# Data Distribution @ Tier-1s

- CMS Data Distr. Model demands heavy load on Tier-1s
- Tape I/O
  - Write custodial RAW/RECO/AOD copies, Re-RECO output
  - Write/Read MCPROD
  - Read RAW for Re-Reconstruction (6 times in 1st year running)
- Network bandwidth and full mesh link commissioning
  - Tier-0→Tier-1 target rate depending on „size“ of Tier-1
  - #commissioned links to/from other Tier-1s >= 4 , to Tier-2s >= 20
  - Tier-1→Tier-1 and Tier1→Tier-2 target rates dep. on “size”
- Disk storage Capacity
  - custodial RECO/AOD and skimms for Tier-1/2 data serving
  - non-custodial RECO/AOD encouraged
- Other strong requirements
  - CPU Capacity : Re-Reconstruction, Skimming and MC Processing
  - 24/7 coverage and high (98%) availability (WLCG). For CMS: PhEDEx admins, Data Managers, contacts reachable during working hours++

# Tier-1 : Moving and Processing Data



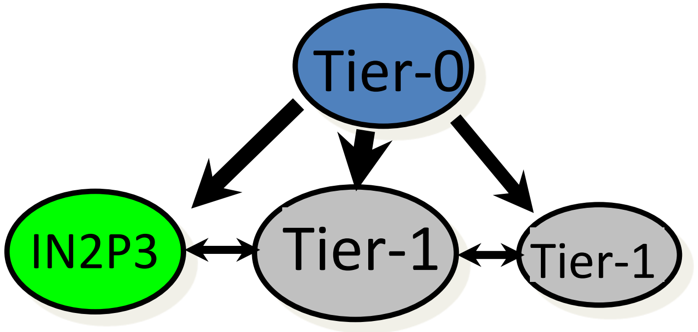
CMS PhEDEx - Transfer Rate

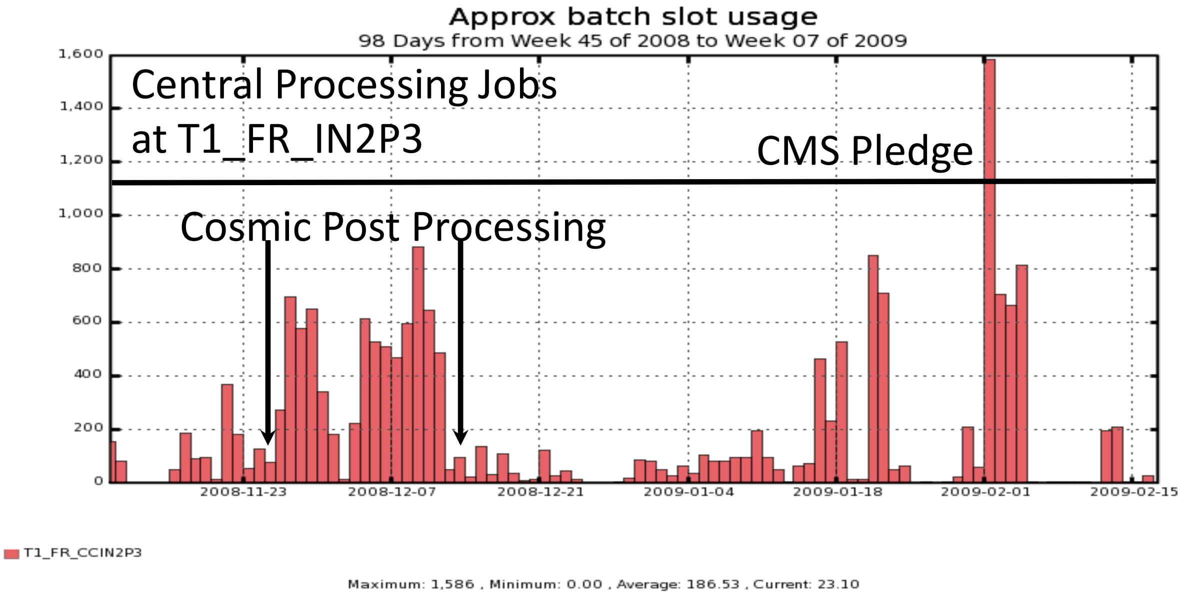
45 Days from 2008-09-24 to 2008-11-08

Data to T1_FR_IN2P3 from CERN and T1 during Cosmic Running

100MB/s

Maximum: 224.69 MB/s, Minimum: 0.00 MB/s, Average: 32.22 MB/s, Current: 36.83 MB/s





Approx batch slot usage

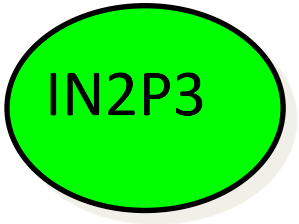
98 Days from Week 45 of 2008 to Week 07 of 2009

Central Processing Jobs at T1_FR_IN2P3

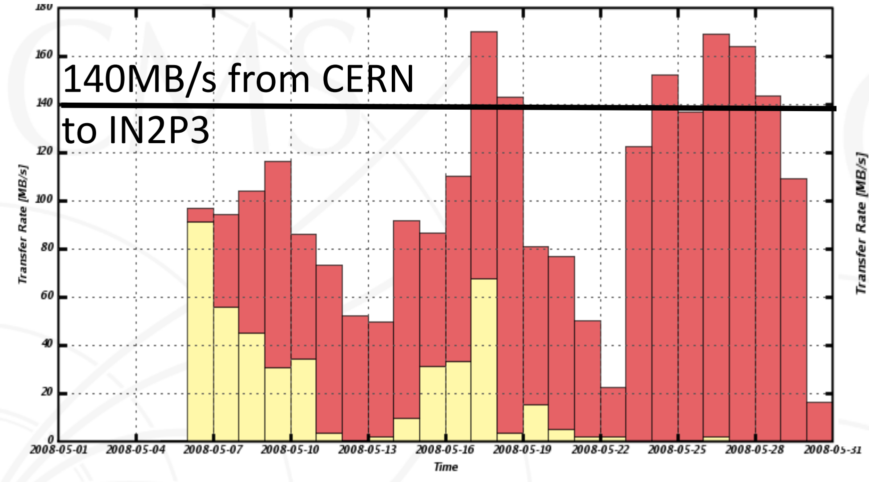
CMS Pledge

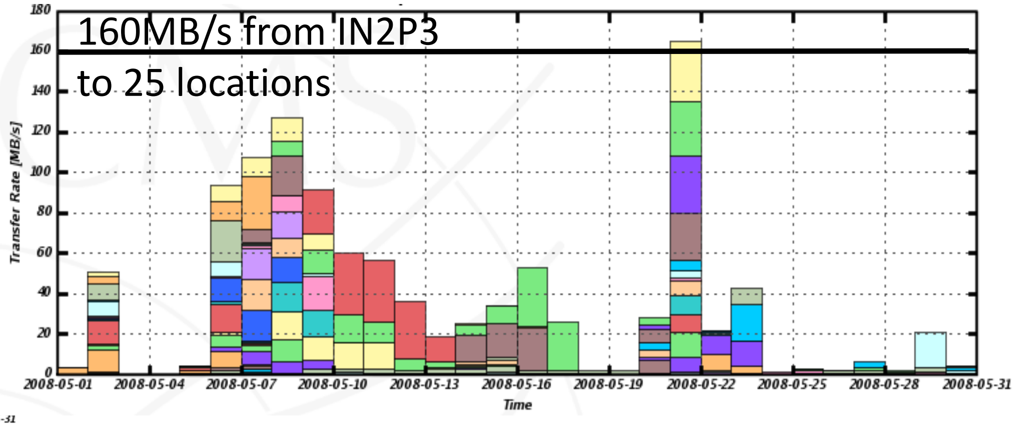
Cosmic Post Processing

Maximum: 1,586 , Minimum: 0.00 , Average: 186.53 , Current: 23.10

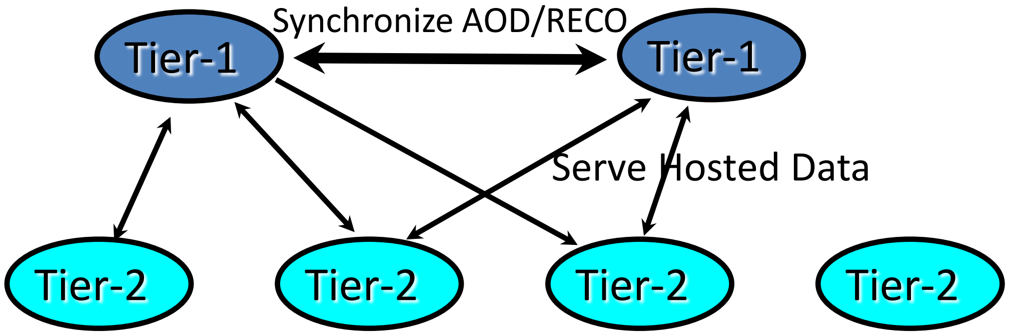


# Tier-1 : Data Serving


140MB/s from CERN to IN2P3


160MB/s from IN2P3 to 25 locations

- In the CMS model the Tier-1s serve the analyzed copy of the data
  - While data is written once, it will be read many times
  - The data serving requirements of the T1s can exceed that of CERN
    - Like CERN the Tier-1s need to ingest and export data simultaneously
- Full mesh of transfers improves data access
  - Also increases commissioning work


Tier-1

Synchronize AOD/RECO

Tier-1

Serve Hosted Data

Tier-2

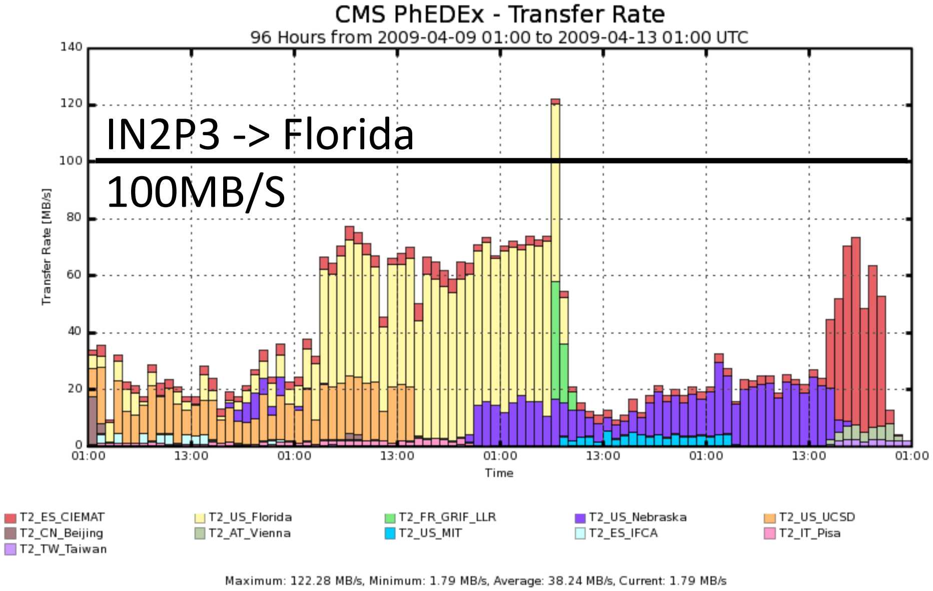
Tier-2

Tier-2

Tier-2

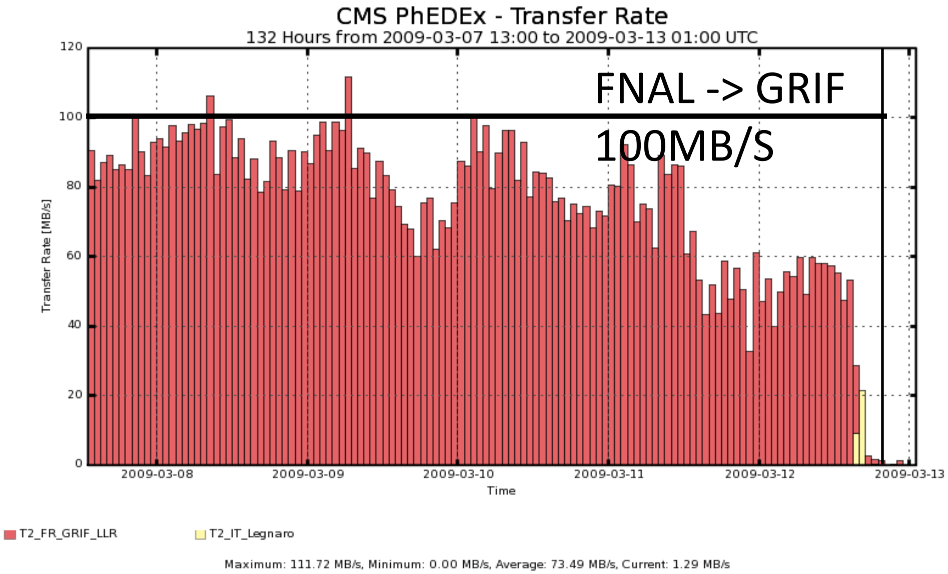
# Tier-1 ← → Tier-2 : Full Mesh Examples

- Link Commissioning in CMS has been a long effort and intensive process
  - Good performance achieved in both directions across the Atlantic
  - Work ongoing



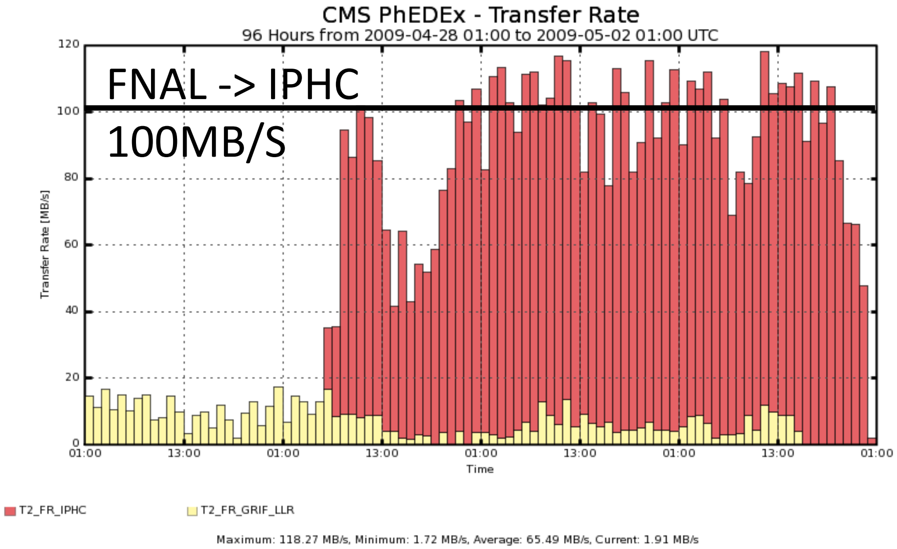
IN2P3 -> Florida

100MB/S



FNAL -> GRIF

100MB/S



FNAL -> IPHC

100MB/S

# Data Distribution @ Tier-2s

- CMS Tier-2 are the Primary resources for user analysis + they host central MC Production
- Efficient Data Distribution from/to Tier-2s relies on

Full Mesh Data Management

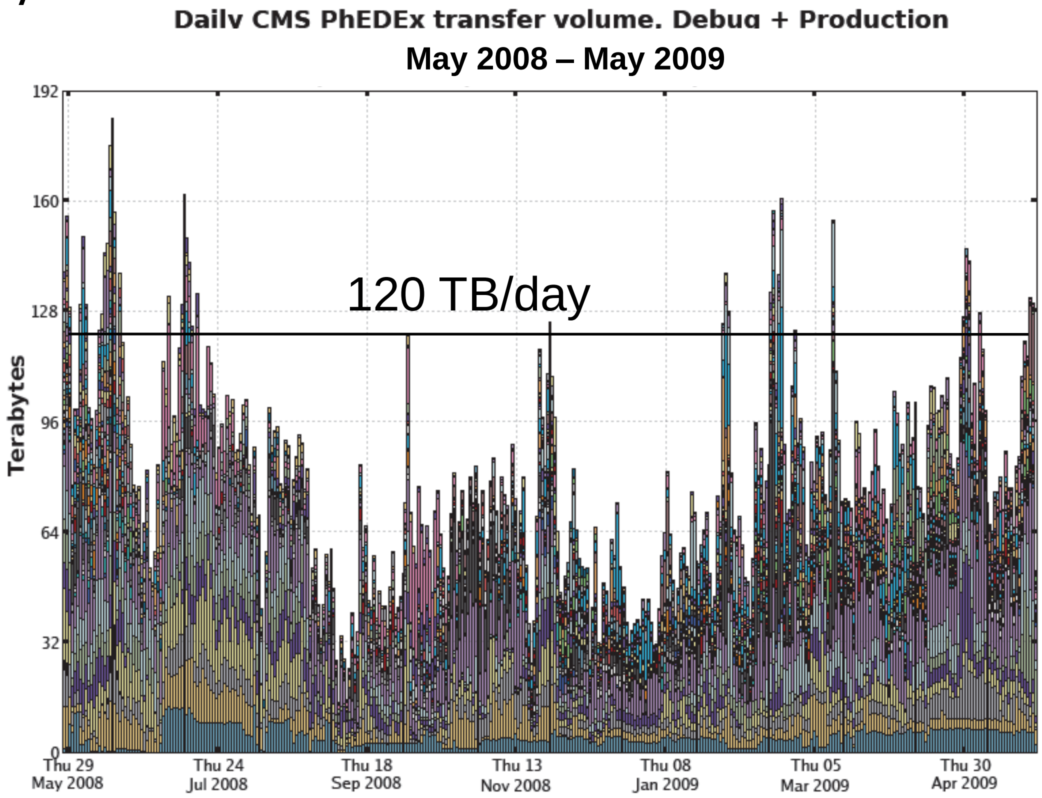
  - >= 2 uplinks and >=4 downlinks required to/from Tier-1s

- 602 $T_i \leftarrow\rightarrow T_j$ links have

been commissioned

- Tier-3 represent another

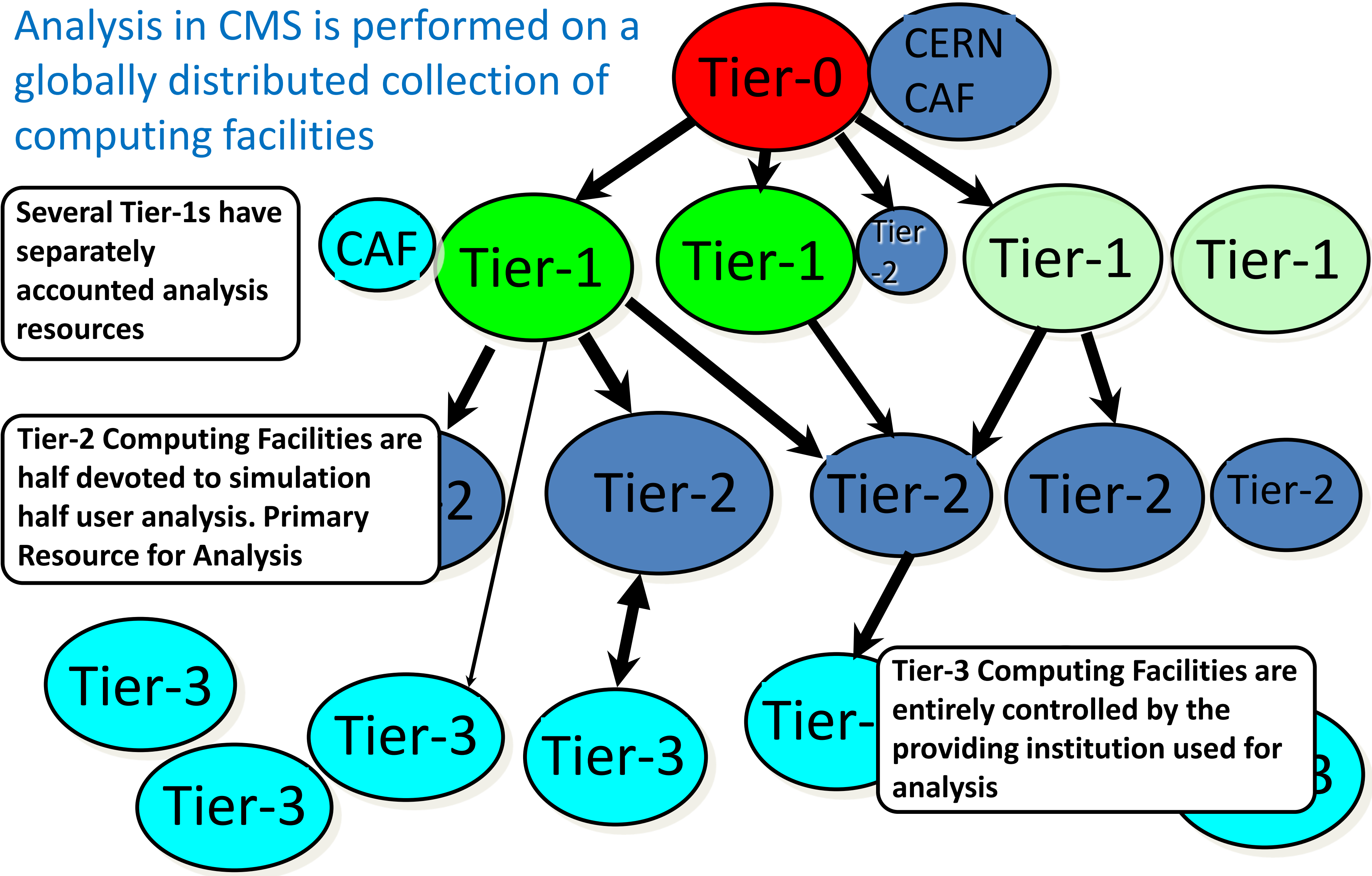
analysis resources controled by the local community. They can receive data from anywhere



- The CMS Analysis Model is build on a Tier-2 to Physics Association Model + Data Management / Worflow Management tools

# Analysis Model in CMS (1/4)

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



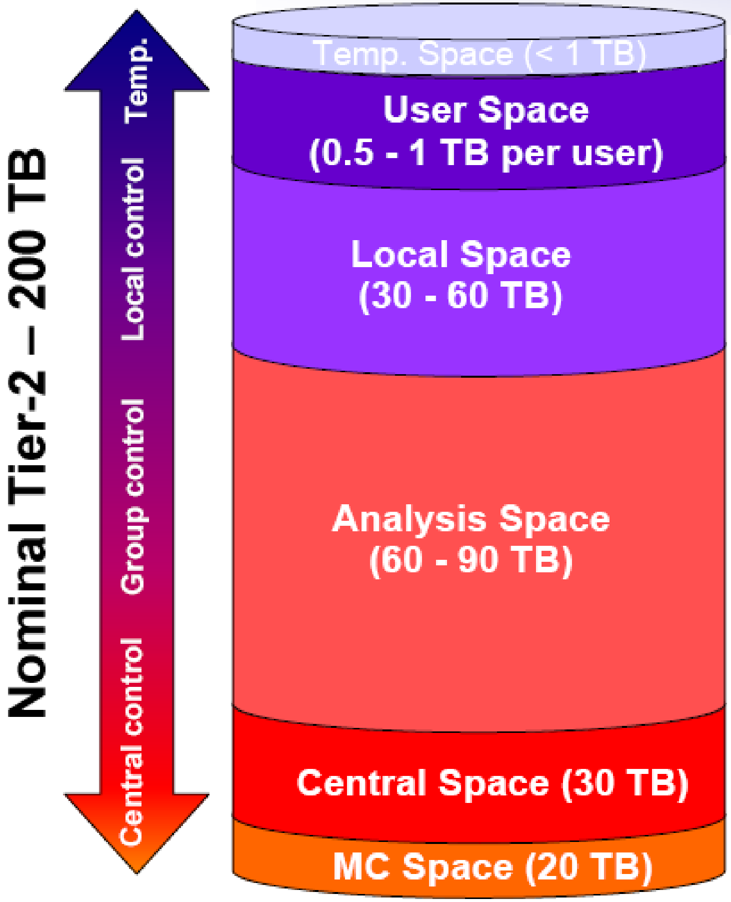
Several Tier-1s have separately accounted analysis resources

Tier-2 Computing Facilities are half devoted to simulation half user analysis. Primary Resource for Analysis

Tier-3 Computing Facilities are entirely controlled by the providing institution used for analysis

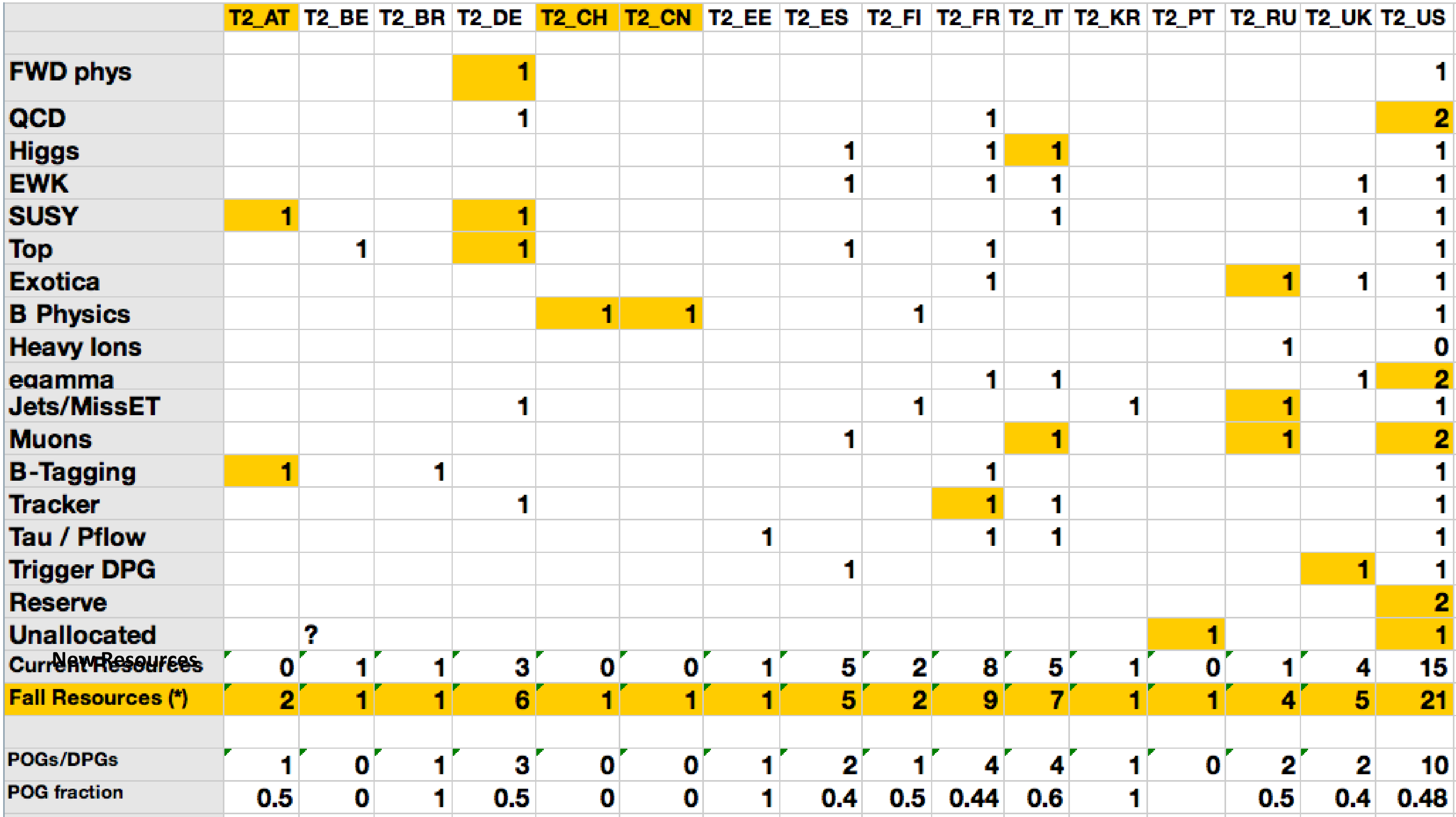
# Analysis Model in CMS (2/4)

- In CMS jobs go to the data
  - The challenging part is making sure the right data is distributed broadly
    - There are 200TB of disk space at a nominal Tier-2.
      - There are 7.7PB pledged total accross the Tier-2s in 2009 (35 sites of nominal Tier-2, currently 5.5PB installed)
    - CMS attempts to share the management of the space across groups
      - Ensures people doing the work have some control
- 20TB is identified for storing local user produced files and making them grid accessible
- 30TB is identified for use by the local group
  - Local community controlled space (PAT, …)
- 30 TB of space at each site is identified for DataOps
  - We expect to be able to host most of the RECO data used in the first year (RECO, AOD)
- 20 TB of space for DataOps for MC staging buffer



# Analysis Model in CMS (3/4)

- Remaining space is divided into chunks and assigned to analysis groups
  - Currently 17 analysis groups in CMS. Balance of physics analysis groups and detector and performance groups



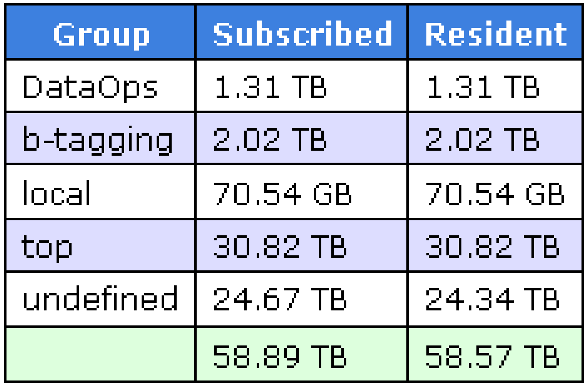
| | T2_AT | T2_BE | T2_BR | T2_DE | T2_CH | T2_CN | T2_EE | T2_ES | T2_FI | T2_FR | T2_IT | T2_KR | T2_PT | T2_RU | T2_UK | T2_US |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FWD phys | | | | 1 | | | | | | | | | | | | 1 |
| QCD | | | | 1 | | | | | | 1 | | | | | | 2 |
| Higgs | | | | | | | | 1 | | 1 | 1 | | | | | 1 |
| EWK | | | | | | | | 1 | | 1 | 1 | | | | 1 | 1 |
| SUSY | 1 | | | 1 | | | | | | | 1 | | | | 1 | 1 |
| Top | | 1 | | 1 | | | | 1 | | 1 | | | | | | 1 |
| Exotica | | | | | | | | | | 1 | | | | 1 | 1 | 1 |
| B Physics | | | | | 1 | 1 | | | 1 | | | | | | | 1 |
| Heavy Ions | | | | | | | | | | | | | | 1 | | 0 |
| egamma | | | | | | | | | | 1 | 1 | | | | 1 | 2 |
| Jets/MissET | | | | 1 | | | | | 1 | | | 1 | | 1 | | 1 |
| Muons | | | | | | | | 1 | | | 1 | | | 1 | | 2 |
| B-Tagging | 1 | | 1 | | | | | | | 1 | | | | | | 1 |
| Tracker | | | | 1 | | | | | | 1 | 1 | | | | | 1 |
| Tau / Pflow | | | | | | | 1 | | | 1 | 1 | | | | | 1 |
| Trigger DPG | | | | | | | | 1 | | | | | | | 1 | 1 |
| Reserve | | | | | | | | | | | | | | | | 2 |
| Unallocated | | ? | | | | | | | | | | | 1 | | | 1 |
| Current Resources<br>New Resources | 0 | 1 | 1 | 3 | 0 | 0 | 1 | 5 | 2 | 8 | 5 | 1 | 0 | 1 | 4 | 15 |
| Fall Resources (*) | 2 | 1 | 1 | 6 | 1 | 1 | 1 | 5 | 2 | 9 | 7 | 1 | 1 | 4 | 5 | 21 |
| | | | | | | | | | | | | | | | | |
| POGs/DPGs | 1 | 0 | 1 | 3 | 0 | 0 | 1 | 2 | 1 | 4 | 4 | 1 | 0 | 2 | 2 | 10 |
| POG fraction | 0.5 | 0 | 1 | 0.5 | 0 | 0 | 1 | 0.4 | 0.5 | 0.44 | 0.6 | 1 | | 0.5 | 0.4 | 0.48 |

# Analysis Model in CMS (4/4)

- In general CMS is concentrating on the groups responsible for commissioning and validation in the first year

- The process of associating analysis groups to sites was challenging
  - Intended to improve communication with sites and put people closer to the work contributing to data management
  - Data Ops uses 400TB at Tier2s for central space. (1.3PB at Accounted at Tier-1s)
  - B-Phys 4TB Jets 22TB Trigger 44TB
  - B-Tagging 35TB Muon 276TB
  - E-gamma 108TB QCD 15TB
  - EWK 280TB SUSY 48TB
  - Exotica 20TB Tau/pflow 11TB
  - Forward 33TB Top 131TB
  - Higgs 44TB Tracker 110TB

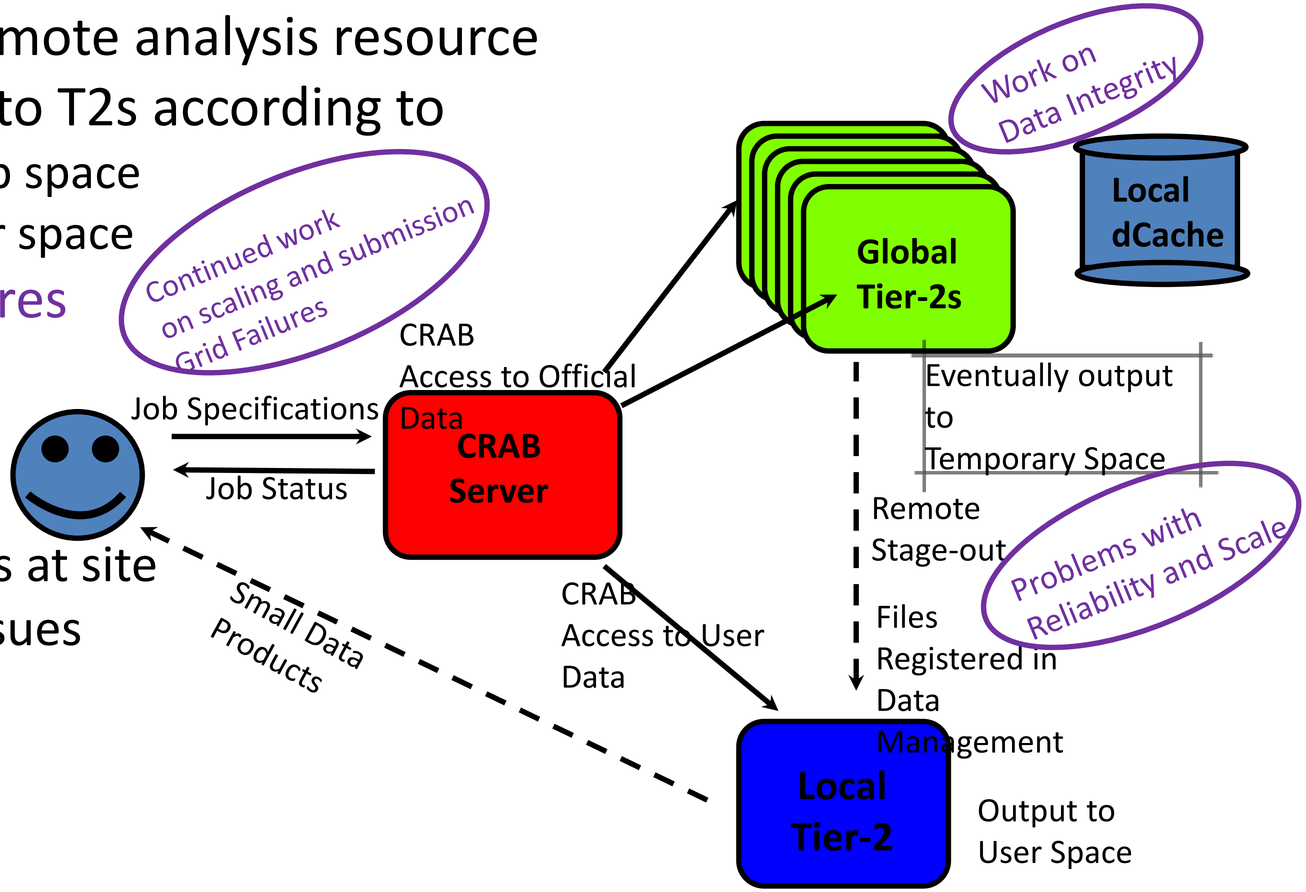
Example: T2_FR_IPHC Group Usage



| Group | Subscribed | Resident |
|---|---|---|
| DataOps | 1.31 TB | 1.31 TB |
| b-tagging | 2.02 TB | 2.02 TB |
| local | 70.54 GB | 70.54 GB |
| top | 30.82 TB | 30.82 TB |
| undefined | 24.67 TB | 24.34 TB |
|  | 58.89 TB | 58.57 TB |

# Accessing the Data for Processing

- How the system will work with 2000 collaborators?
  - CMS Remote Analysis Builder (CRAB) shields the user from the underlying complexity, but a many things have to succeed for analysis to be successful
- Jobs submitted to remote analysis resource
- Output Data moved to T2s according to
  - Central Physics Group space
  - Regional or local user space
- CMS sees ~20% failures on analysis jobs
  - Grid Failures
  - User config. errors
  - Data reading failures at site
  - Remote Stageout issues

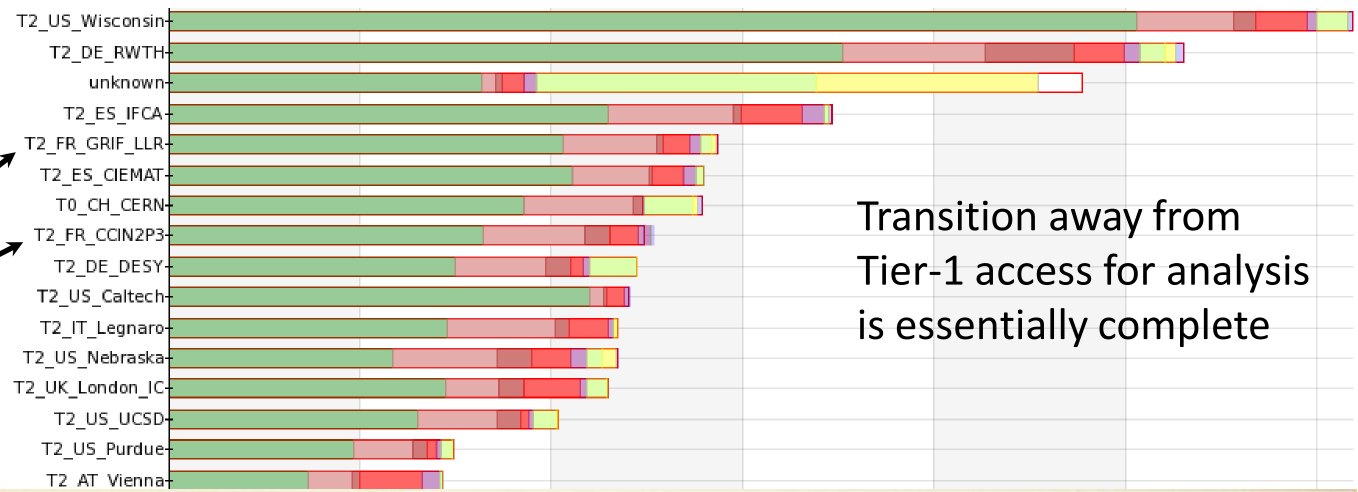

Work on Data Integrity
Local dCache
Global Tier-2s
Continued work on scaling and submission Grid Failures
CRAB Access to Official Data
Job Specifications
Job Status
CRAB Server
Eventually output to Temporary Space
Remote Stage-out
Problems with Reliability and Scale
Small Data Products
CRAB Access to User Data
Files Registered in Data Management
Local Tier-2
Output to User Space

- Adding users and workflows will further stress the system

# Usage of Tier-2s for Analysis

- CMS is currently seeing about 50k jobs per day

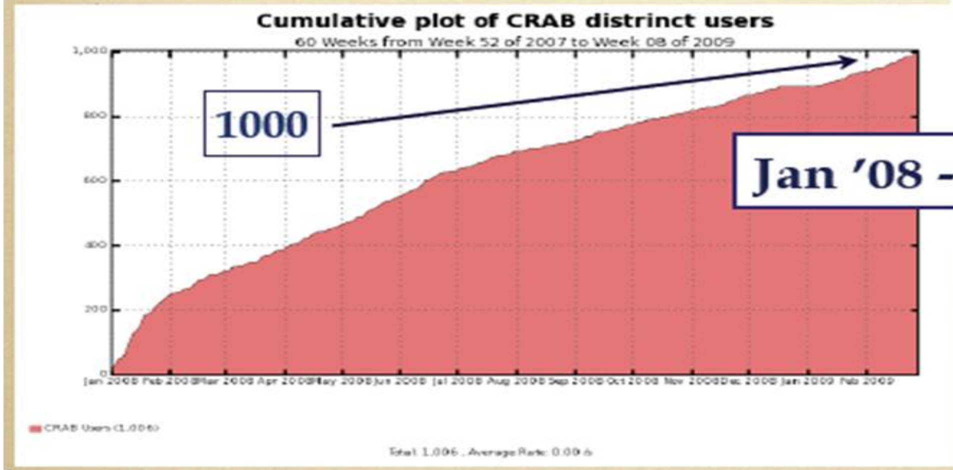
Analysis
Jobs Submitted in the
last Month
Top 25 Sites



French Tier-2 Sites

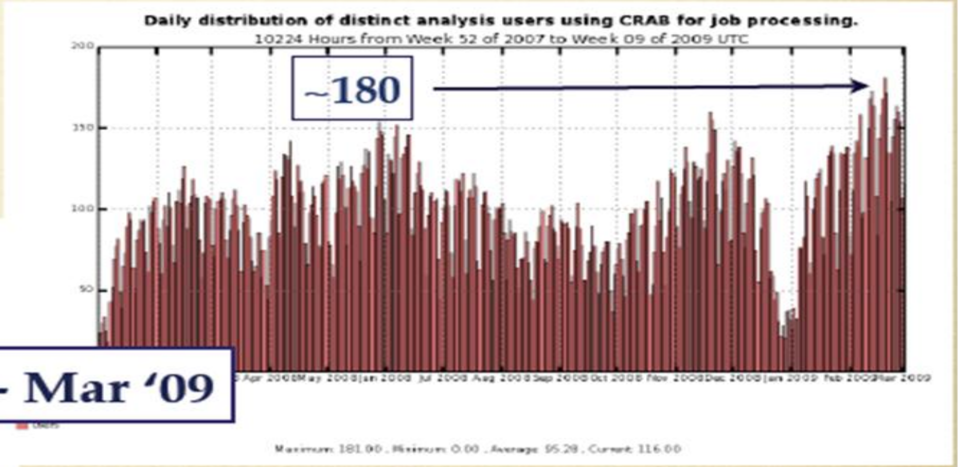
Transition away from Tier-1 access for analysis is essentially complete

More than 40% of the CMS Collaboration make use of the distributed infrastructure



1000

Jan '08 - Mar '09



~180

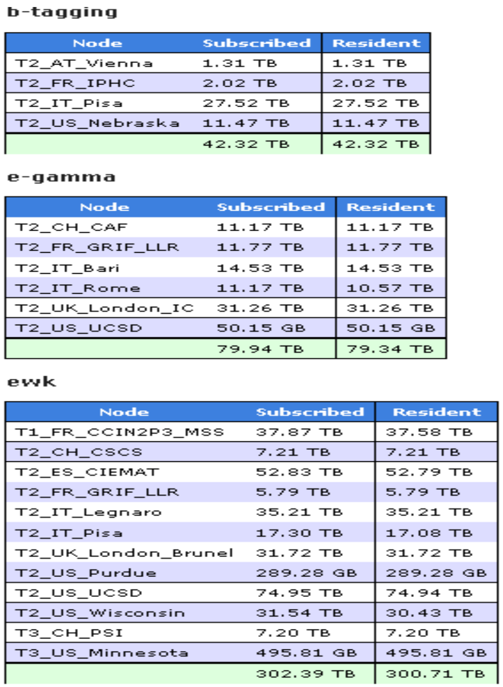
Data from
CMS Dashboard

# Data Management Tools

- Main CMS Components: Data Bookkeeping System (DBS), Local File Catalog and the Data Transfer tool (PhEDEx)
- Any CMS user can place a transfer subscription (Web based)
- Site Data managers or Central Operations approve transfers
- Monitoring of Group Usage for PH group conveners :

https://cmsweb.cern.ch/phedex/prod/Reports::GroupUsage

e usage (for data managers), transfer qualities...



b-tagging

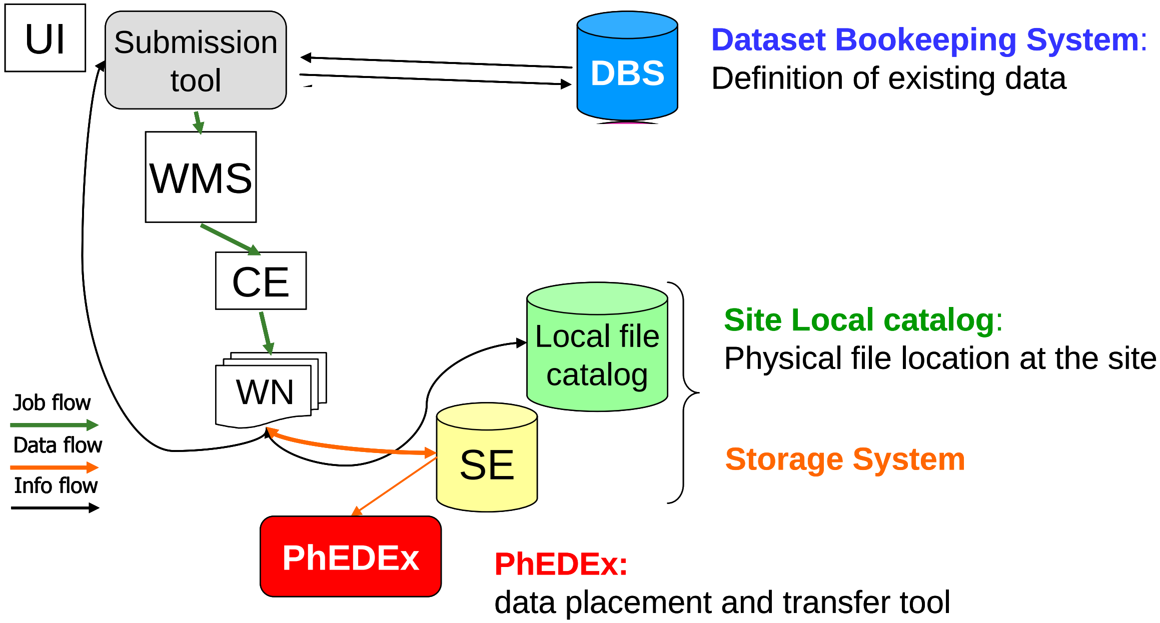
| Node | Subscribed | Resident |
|---|---|---|
| T2_AT_Vienna | 1.31 TB | 1.31 TB |
| T2_FR_IPHC | 2.02 TB | 2.02 TB |
| T2_IT_Pisa | 27.52 TB | 27.52 TB |
| T2_US_Nebraska | 11.47 TB | 11.47 TB |
|  | 42.32 TB | 42.32 TB |

e-gamma

| Node | Subscribed | Resident |
|---|---|---|
| T2_CH_CAF | 11.17 TB | 11.17 TB |
| T2_FR_GRIF_LLR | 11.77 TB | 11.77 TB |
| T2_IT_Bari | 14.53 TB | 14.53 TB |
| T2_IT_Rome | 11.17 TB | 10.57 TB |
| T2_UK_London_IC | 31.26 TB | 31.26 TB |
| T2_US_UCSD | 50.15 GB | 50.15 GB |
|  | 79.94 TB | 79.34 TB |

ewk

| Node | Subscribed | Resident |
|---|---|---|
| T1_FR_CCIN2P3_MSS | 37.87 TB | 37.58 TB |
| T2_CH_CSCS | 7.21 TB | 7.21 TB |
| T2_ES_CIEMAT | 52.83 TB | 52.79 TB |
| T2_FR_GRIF_LLR | 5.79 TB | 5.79 TB |
| T2_IT_Legnaro | 35.21 TB | 35.21 TB |
| T2_IT_Pisa | 17.30 TB | 17.08 TB |
| T2_UK_London_Brunel | 31.72 TB | 31.72 TB |
| T2_US_Purdue | 289.28 GB | 289.28 GB |
| T2_US_UCSD | 74.95 TB | 74.94 TB |
| T2_US_Wisconsin | 31.54 TB | 30.43 TB |
| T3_CH_PSI | 7.20 TB | 7.20 TB |
| T3_US_Minnesota | 495.81 GB | 495.81 GB |
|  | 302.39 TB | 300.71 TB |



**Dataset Bookeeping System**:
Definition of existing data

**Site Local catalog**:
Physical file location at the site

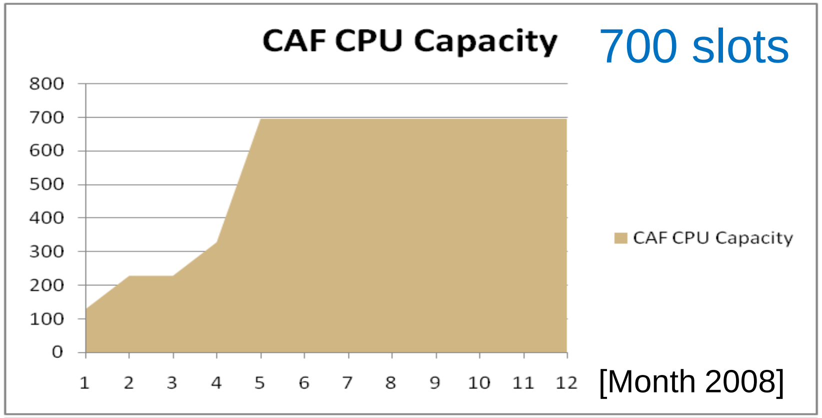
**Storage System**

**PhEDEx:**
data placement and transfer tool

# The Role of the „CAF/T1“ at CERN

- Unique role for latency-critical functions at the source of the analysis chain, with the main goals of :
  - alignment and calibration
  - trigger and detector diagnostics, monitoring and performance analysis
  - physics monitoring, analysis of express stream, fast-turnaround high-priority analysis
- At LHC startup , extended role: Detector Commissioning, Offline DQM, early analysis, requiring large fraction of the RECO (RAW) data on CAF
- The technical requirements on the CAF/T1 are based on the capability to support high-priority and low-latency workflows
  - Large disk-only storage pool
  - Dedicated batch system
  - Interactive facilities
- While enforcing controlled and prioritized access policies to hundreds of users
  - Single entry point for Disk / WN / Interactive access : LSFWEB, controlled by CAF Group leaders

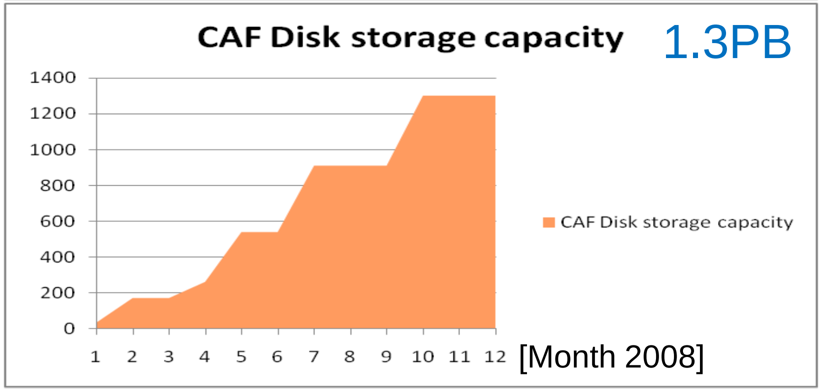
# CAF@CERN Status



CAF CPU Capacity

700 slots

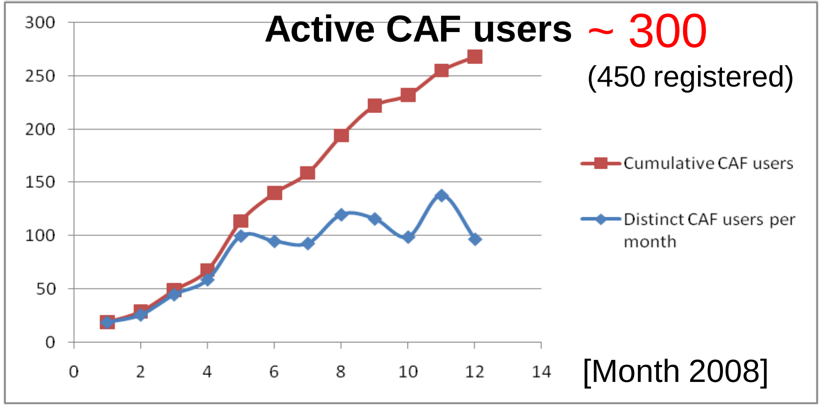
[Month 2008]



CAF Disk storage capacity

1.3PB

[Month 2008]

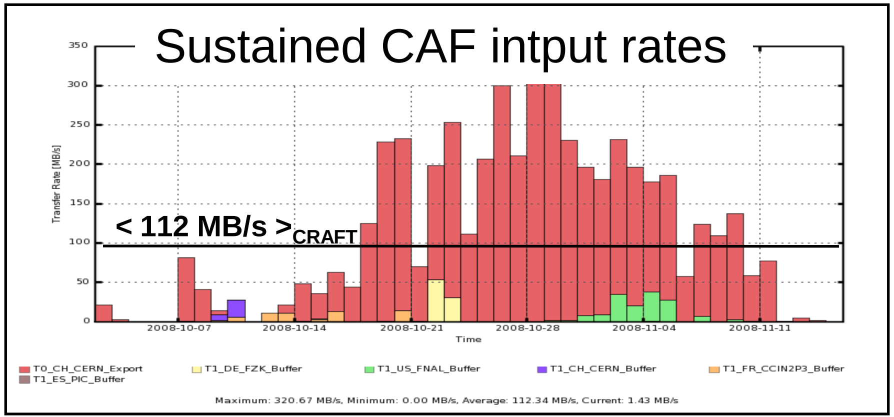


Active CAF users ~ 300

(450 registered)

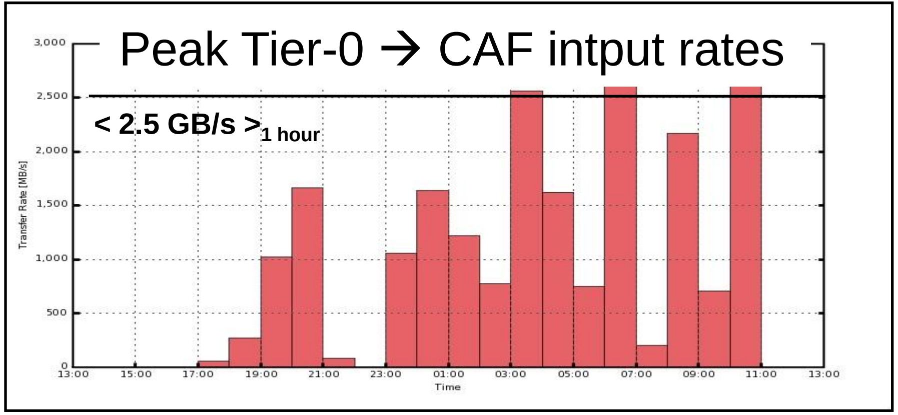
[Month 2008]

# Performance



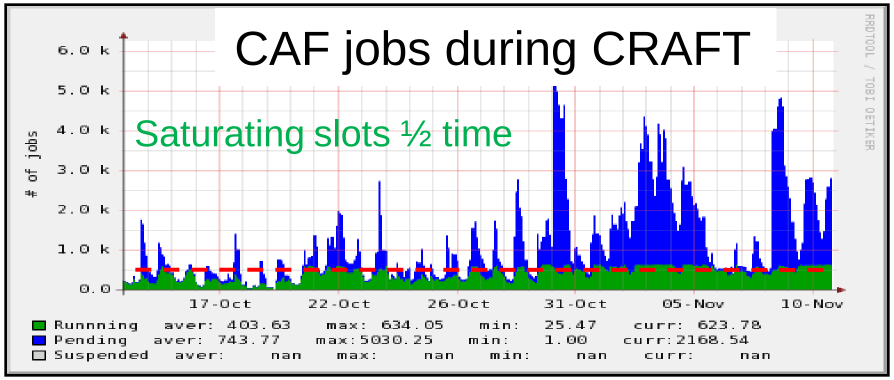
Sustained CAF intput rates

< 112 MB/s >$_{\text{CRAFT}}$



Peak Tier-0 → CAF intput rates

< 2.5 GB/s >$_{\text{1 hour}}$



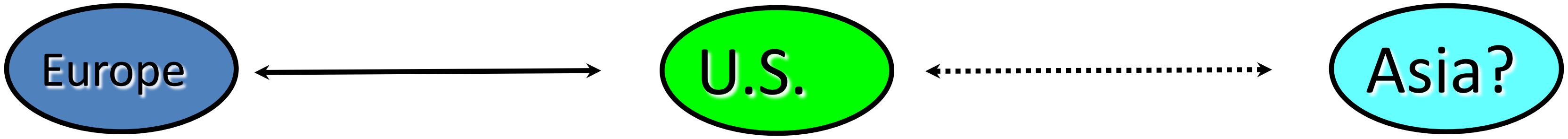
CAF jobs during CRAFT

Saturating slots ½ time

# The role of CAFs in general

- The „CAF/T1“ at CERN has unique role : low latency data access for Calibration&Alignment and Express analysis. To be opposed to general CAFs
- CMS will extend the role of the CAF to serve all CMS users at CERN
  - User Storage : 1-2 TB/user CASTOR, for 200-400 users
  - Requires User Access policies + Space Management (no quotas in CASTOR)
  - Interactive facility : lxplus
  - CAF/T1 will stay „protected“ (but same CASTOR name space)
- „CAF/T2“ also planed at CERN (Grid access, Group space) but not yet well defined
- Other CMS „CAFs“ at Tier-1s :
  - FNAL : inter. access, dedicated batch farm, no GRID connection to CAF resources
  - CCIN2P3-T2 : using physically same SRM endpoint as T1, CPU sperate between T1 and T2 (or at least prioritized with production roles)
- CAFs are capitalizing on the fact that local physics communities have easy access to production data + interactive access. However, the Distributed Analysis Model stays the only viable solution for CMS-wide analysis, given the large amount of data and physicists spread around the world.

# Central/Remote Computing Operations

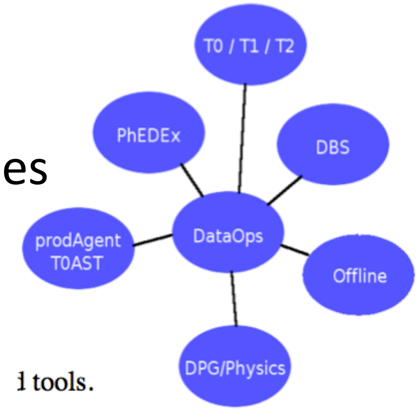


Europe

U.S.

Asia?

- Data Production begins with Data Operations
  - CMS has utilized a two team model for almost 2 years
  - Hand off at the end of the CERN day to FNAL team. Potentially a team for the Asia time zone in the future
- Data Operations Tasks
  - Validate all CMSSW (software) releases
  - Operate T0 for data taking
  - Reprocess Data/MC at T1s
  - Produce MC at the T2s
  - Coordinate central data transfers



- Facilities Operations is more distributed, e.g. Site Monitoring or SW Deployment teams
- Plan to create Analysis Operations team to support users
- The CERN CMS Centre and FNAL Remote Operations Center are "home bases" for Central Operations and for the Offline Computing Shifts, but plan to extend to more Remote Operation Centers

# Computing Operation Centres

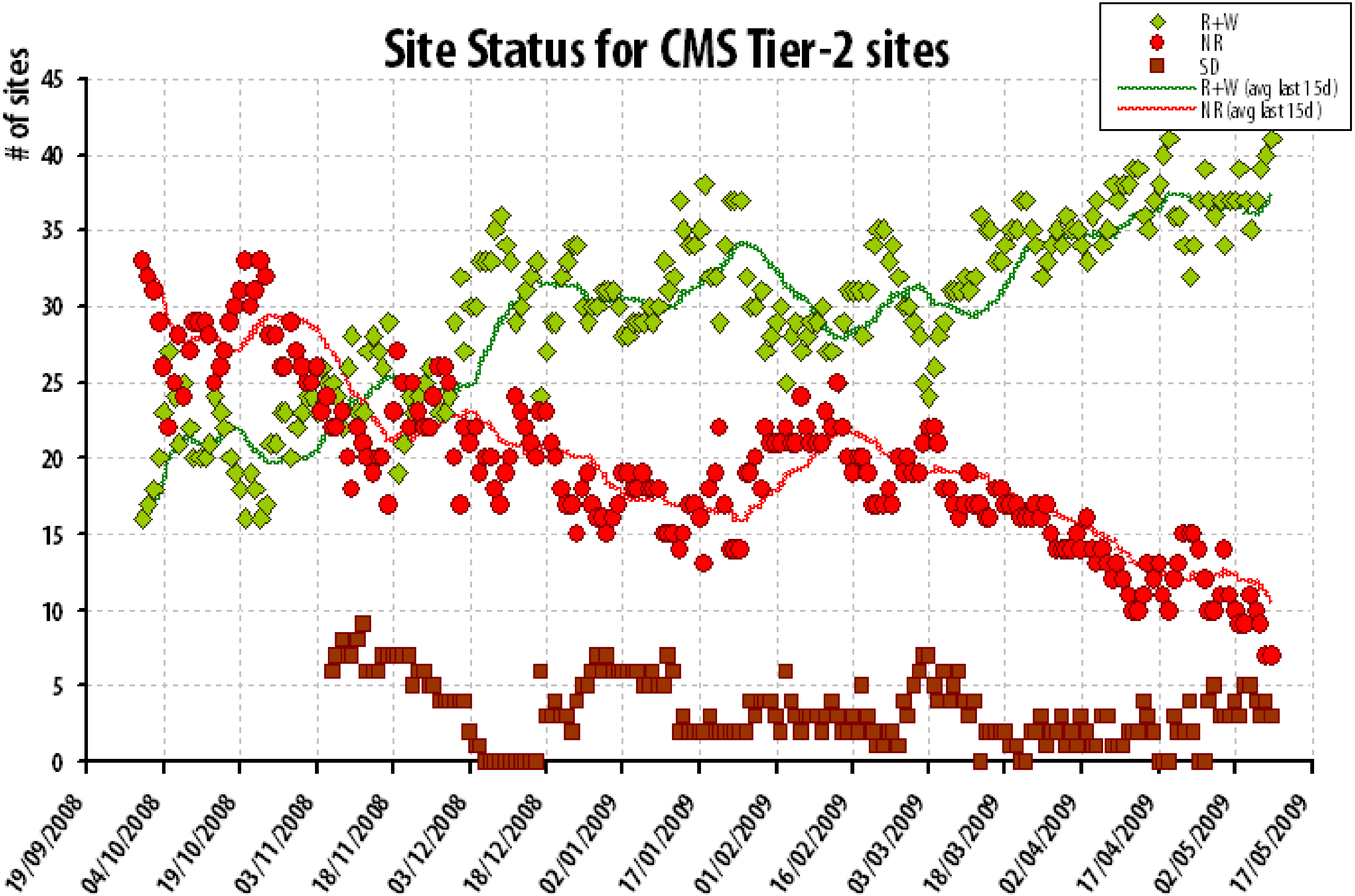
CMS Remote Operations Centre at Fermilab

CMS Experiment Control Room

Permanently-running video links between operations centres

CMS Centre at CERN: monitoring, computing operations, analysis

# Site Availabilty and Readiness



# Challenges/Evolution of the Model until the Physics run

- Step09: Scaling test to be started June 1st
  - Tier-0 : multi-VO tape writing
  - Tier-1 : Processing and Storage I/O
  - Tier-2 : Large Scale Analysis activity
- CRAFT, MWGR, ...

- Evolution of the Data Distribution between today / LHC Startup / High Luminosity running (non-exhaustive list) :

  (i) Today
  - All RECO data at Tier-2s (keep 2 copies) + all RECO at CAF
  - All RECO from Tier-1 reprocessing back to CAF

  ➔ insurrance to make data quickly accessible. Not affordable (and hopefully not needed !) on the long run
  - Occasionally Skimming done at Tier-0
  - Un-equal processing contributions of Tier-1s (CERN and FNAL dominate)

  ➔ Need to make more efficient use all CMS resources

# Challenges/Evolution of the Model until the Physics run (ctnd)

## (ii) 2010 and beyond : Large increase of Data Volume

- Optimize Primary Datasets (10-20) : reasonable and equal size in order to optimize Tier-0 processing and Tier-1 Data Distribution. Also reduce PD ovelap (current estimate 40%)
- CAF : migrate RECO analysis to Tier-2s
- Tier-1 : make better us of resources
- Tier-2 : treat storage more dynamically. Tier-1 → Tier-2 transfer only when needed, then flush Tier-2 space.

➔ Good networking should allow to do that !

- Analysis Model : deploy tool to promote user-produced data to all-CMS : migrate data from /store/user to /store/results, merge small datafiles, inject them into PhEDEx and Global DBS.
- Strengthen central Computing Operations: Asian zone, Analysis Operations support, Computing shifts
- ...

# Conclusions

- The CMS Data Distribution Model is based on Primary Datasets
- As opposed to the ATLAS „clowds“ concept, CMS pulls data from Tier-1s and has broader data-access

➔ The price to pay is the commissioning of the full mesh connectivity between sites

- The CMS Analyis Model tries to empower more users to manage their data (PH group space mgmt at Tier-2s)
- CMS computing workflows and CMS sites have already reached a high level of Readiness in recent years
- To meet our final goals, efforts need to be concentrated on scaling tape I/O, scaling distributed analysis, make better use of resources, go to routine operation mode