



Universität Karlsruhe (TH)

Forschungsuniversität • gegründet 1825



CMS Computing Model with Focus on German Tier1 Activities

Armin Scheurer

GridKa School 2009, High Energy Physics Session

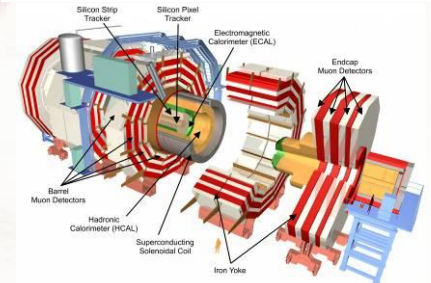
Wednesday, 02.09.2009



Karlsruhe Institute of Technology

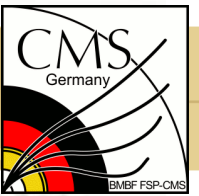


**Bundesministerium
für Bildung
und Forschung**





Overview

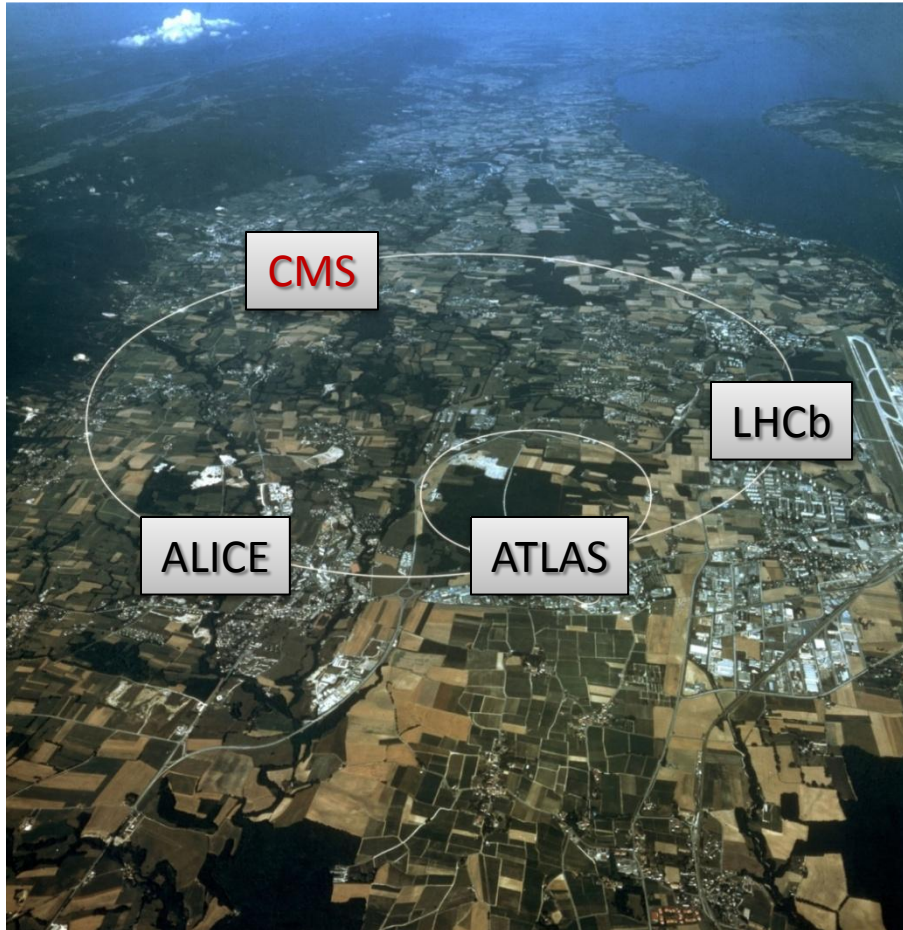


- ❑ The Large Hadron Collider (LHC)
- ❑ The Compact Muon Solenoid (CMS)
- ❑ CMS Computing Model
- ❑ CMS in Germany
- ❑ CMS Workflow & Data Transfers
- ❑ CMS Service Challenges





Large Hadron Collider



Proton-Proton Collider

Circumference:	27 km
Beam Energy:	7 TeV/c ²
Below Surface:	100 m
Temperature:	-271 °C
Energy Use:	1 TWh/a

4 Large Experiments

- CMS (General-Purpose)
- Atlas (General-Purpose)
- LHCb (Physics of b-Quarks)
- Alice (Lead Ion Collisions)

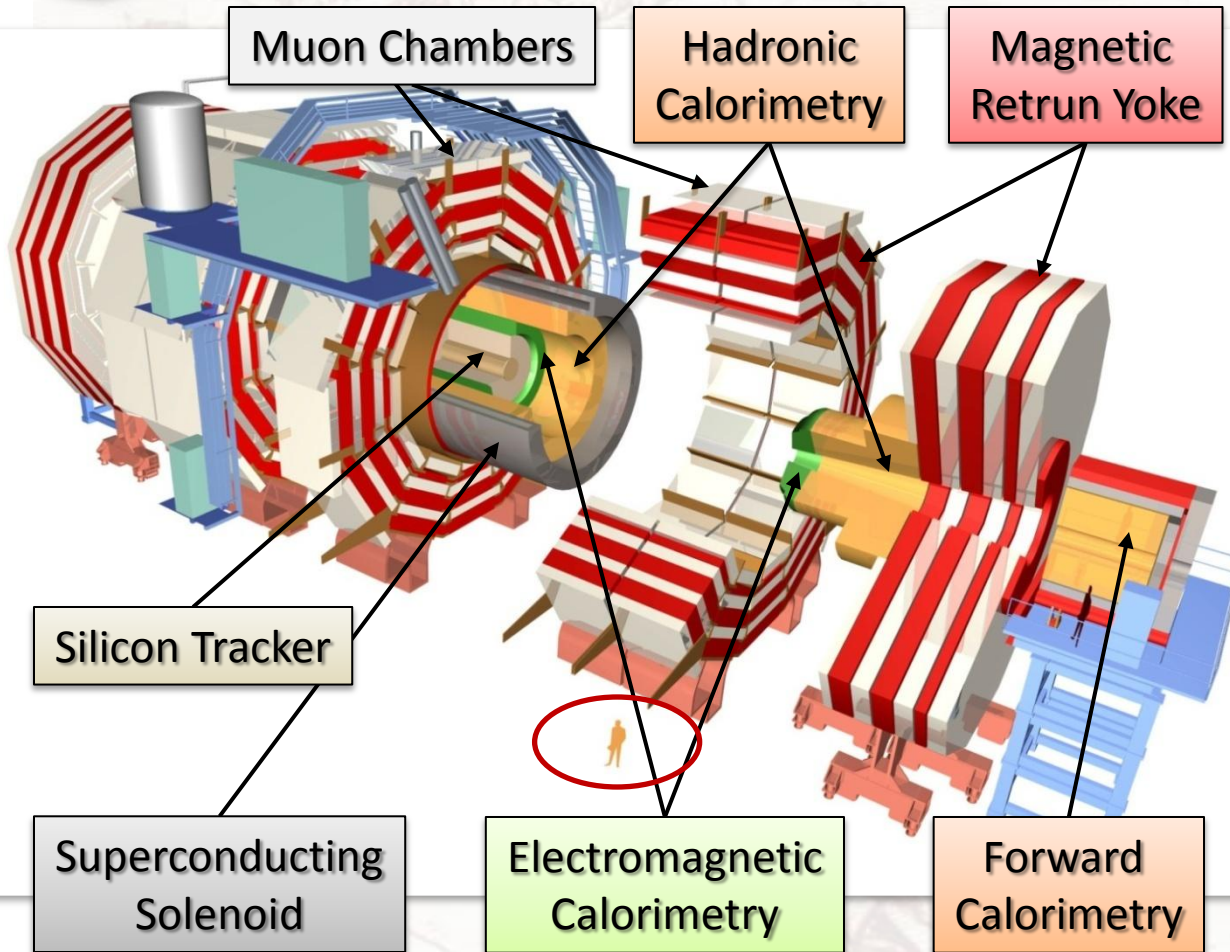
2 Smaller Experiments

- LHCf and Totem






Compact Muon Solenoid



Technical Details:

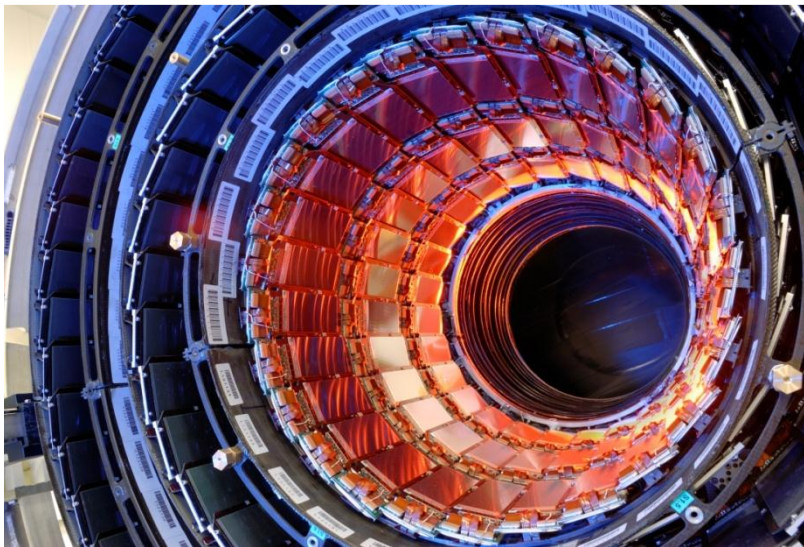
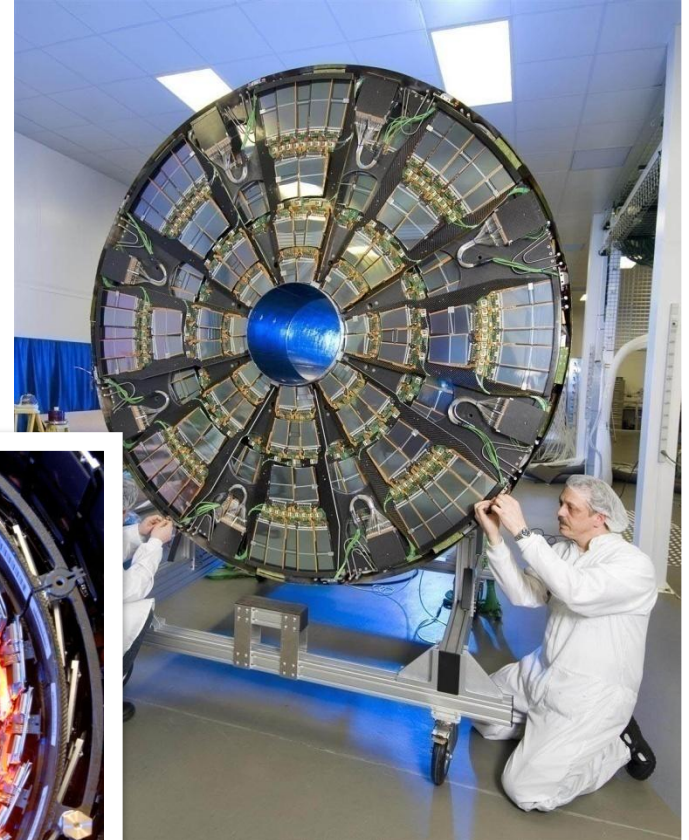
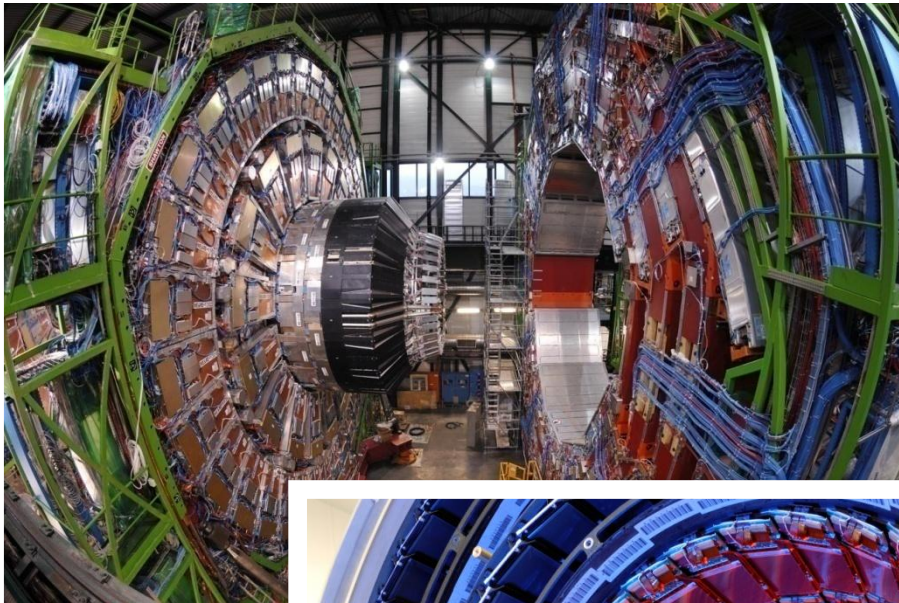
-  Total Weight: 12 500 t
-  Diameter: 15 m
-  Total Length: 21,5 m
-  Magnetic Field
 -  Solenoid: 4 Tesla
 -  Yoke: 2 Tesla
-  Readout Channels, e.g.
 -  Tracker: 10 Mio.
 -  Sum: 100 Mio.
-  Collision Rate: 40 MHz

Data-Rate: Imagine a 100 MPixel-Camera taking 40 Mio. pictures per second!



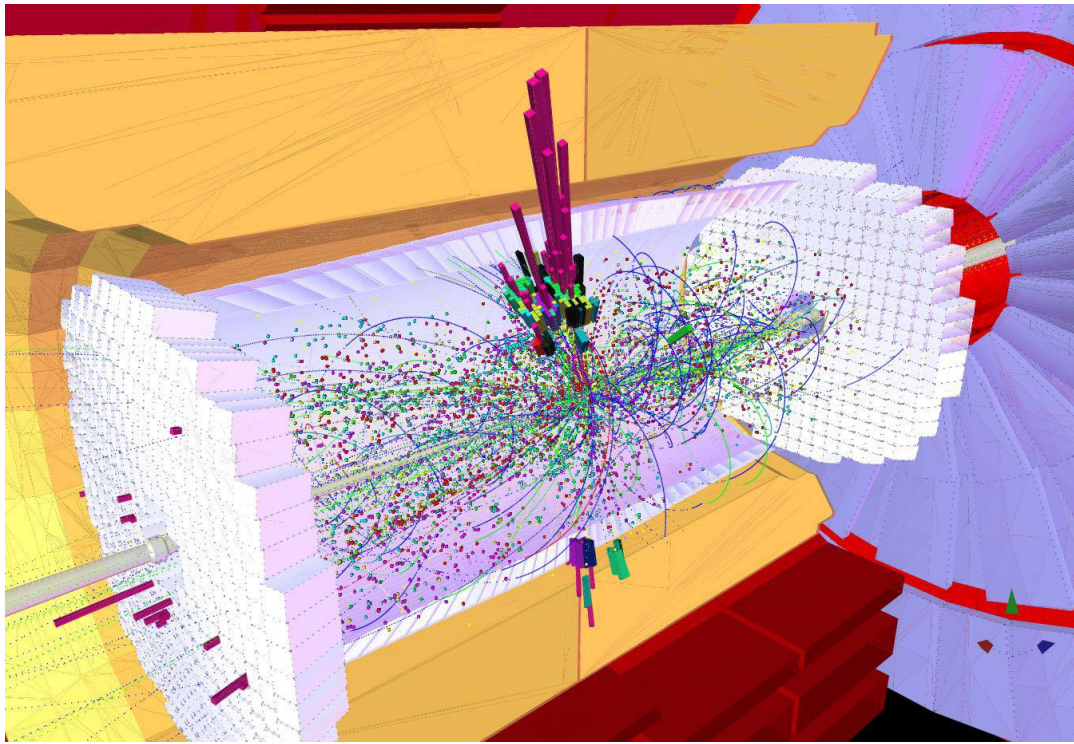


Pictures of CMS





Physics Motivation



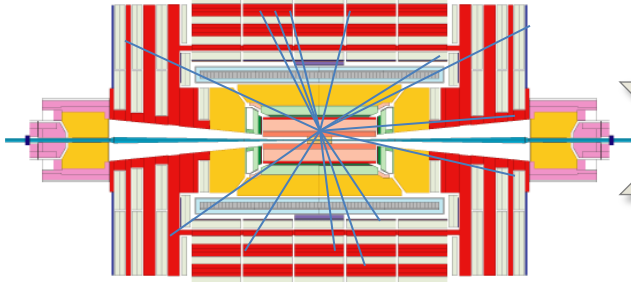
- ❑ In each proton-proton collision, more than **1000 particles** are created.
- ❑ The **decay products** allow to conclude on **underlying physics processes** of the collision.

- ❑ Test of the Standard Model (at the TeV energy scale)
- ❑ Search for the Higgs-Boson
- ❑ Physics beyond the SM (e.g. SUSY, extra dimensions, ...)

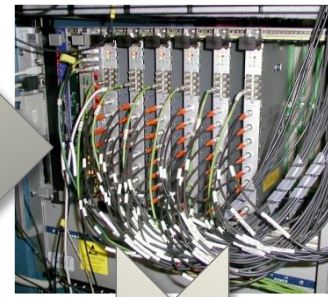




Trigger and Event Rate



60 TB/sec



Level 1 Trigger

Reduction with
ASICs (Hardware)

150 GB/sec

Collision Rate: 40 MHz

Event-Size: 1,5 MB

Tape & HDD
Storage

for Offline-
Analysis



225 MB/sec



High Level Trigger

Software Data
Reduction

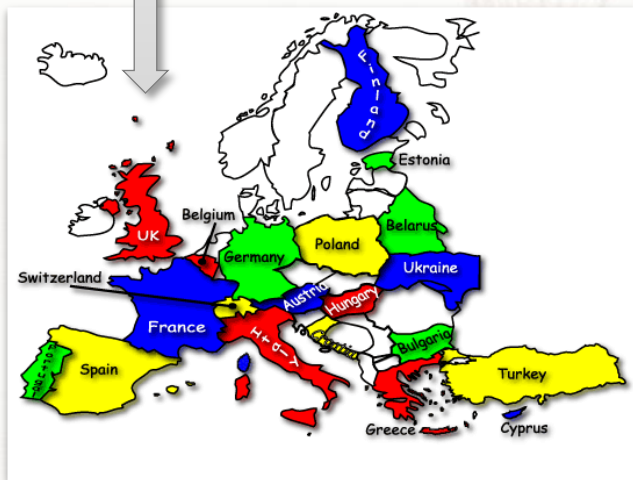
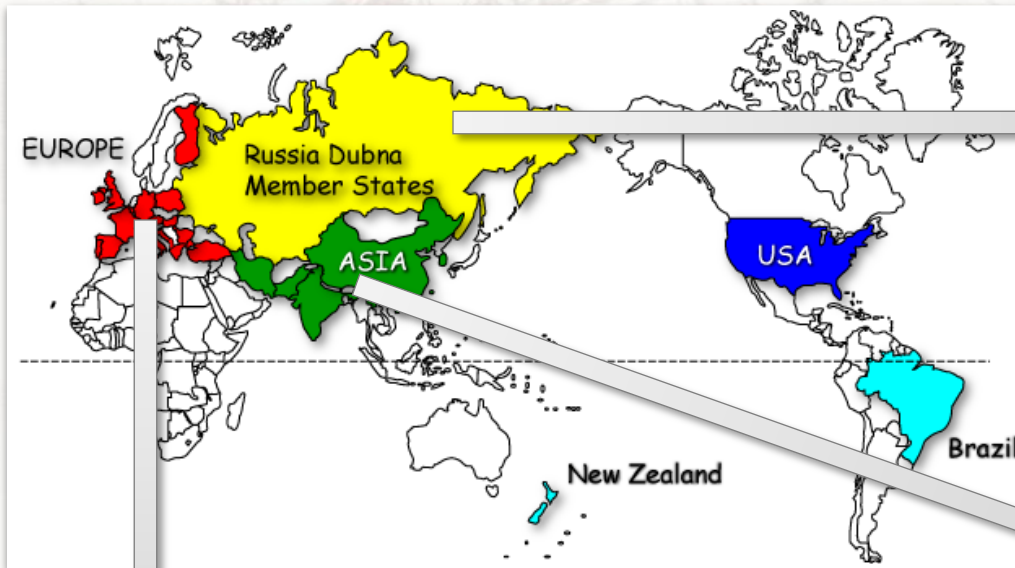
Recorded Events: 150 per second

→ 1.5 PB of data per year

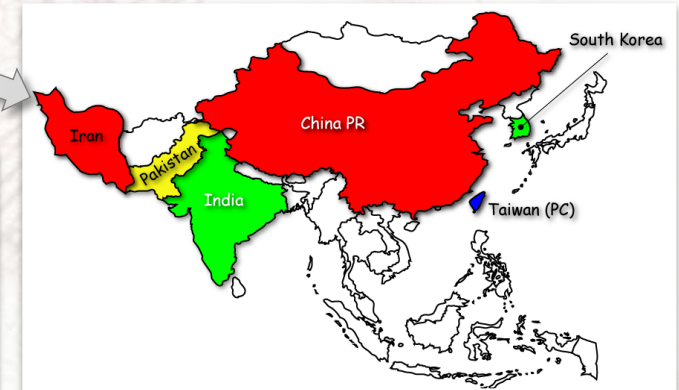




CMS Collaboration



CMS
38 Nations
182 Institutions
> 2000 Scientists & Engineers

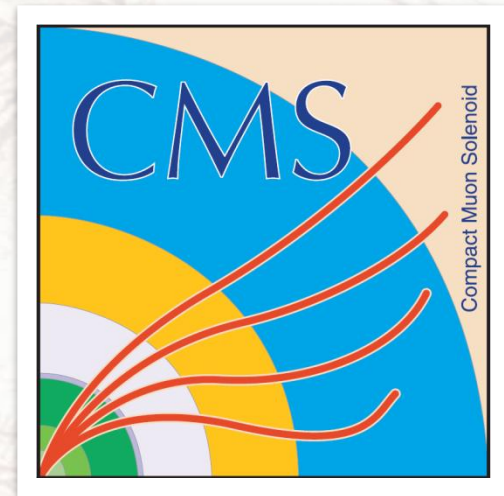




CMS Computing Model

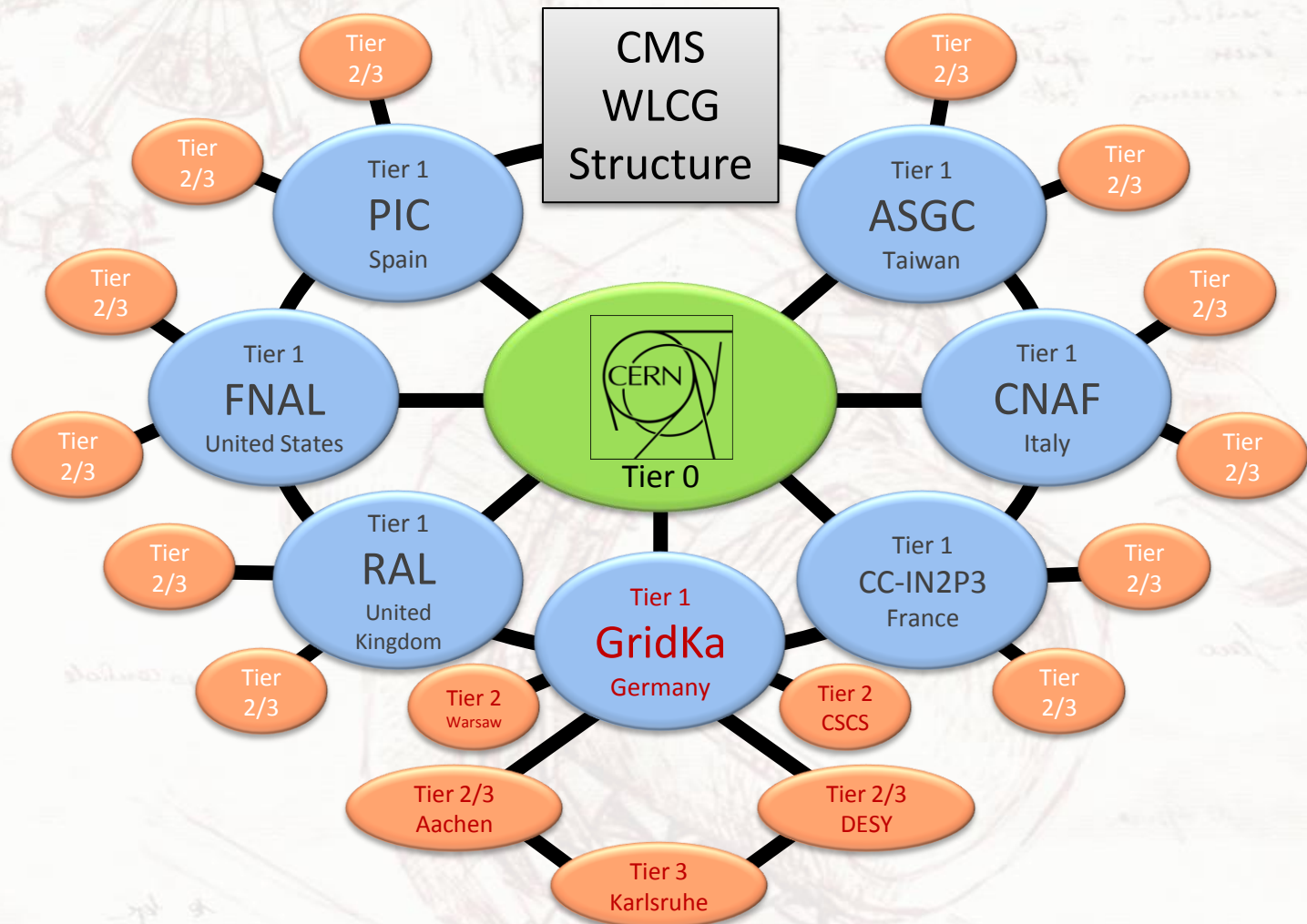


- ❏ LHC experiments have decided to use **distributed** computing and storage **resources**.
- ❏ The Worldwide LHC Computing Grid (WLCG).
- ❏ Grid services based on the **gLite middleware**.
- ❏ Computing centres arranged in a four-tiered **hierarchical structure**.
- ❏ Availability and resources are regulated by MOU (e.g. Downtime per year, response time, etc.).



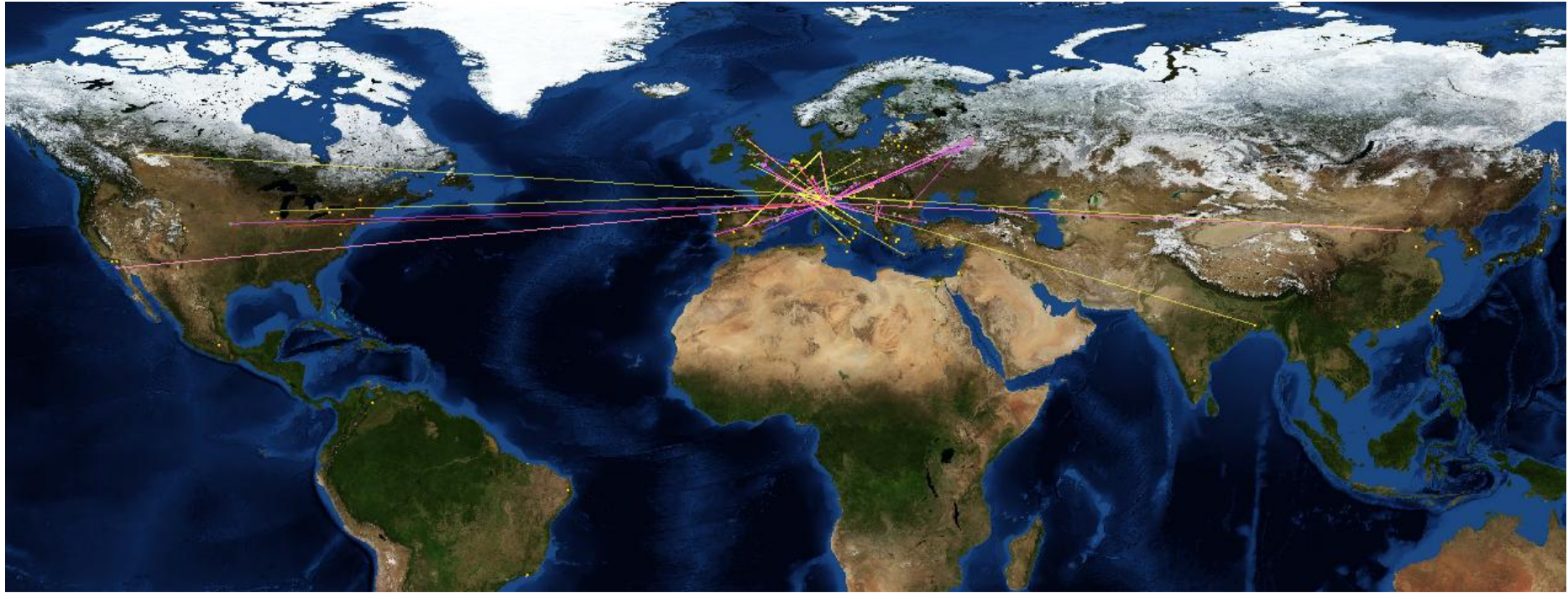
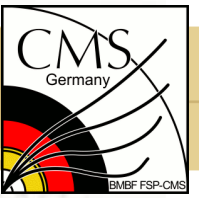


CMS Tier Structure






WLCG Resources



 Taken from the WLCG Real Time Monitor:

 <http://gridportal.hep.ph.ic.ac.uk/rtm/>










Main Tier Responsibilities






Tier0:

-  Storage of RAW detector data.
-  First reconstruction of physics objects after data-taking.



Tier1:

-  Host one dedicated copy of RAW and reconstructed data outside the Tier0.
-  Re-processing of stored data.
-  Skimming (creation of small sub data samples)

Tier2:

-  Monte Carlo production/simulation.
-  Calibration activities.
-  Resources for physics groups analyses.

Tier3:

-  Individual user analyses.
-  Interactive logins (for development & debugging)





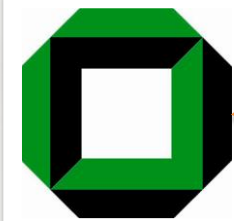
German CMS Members



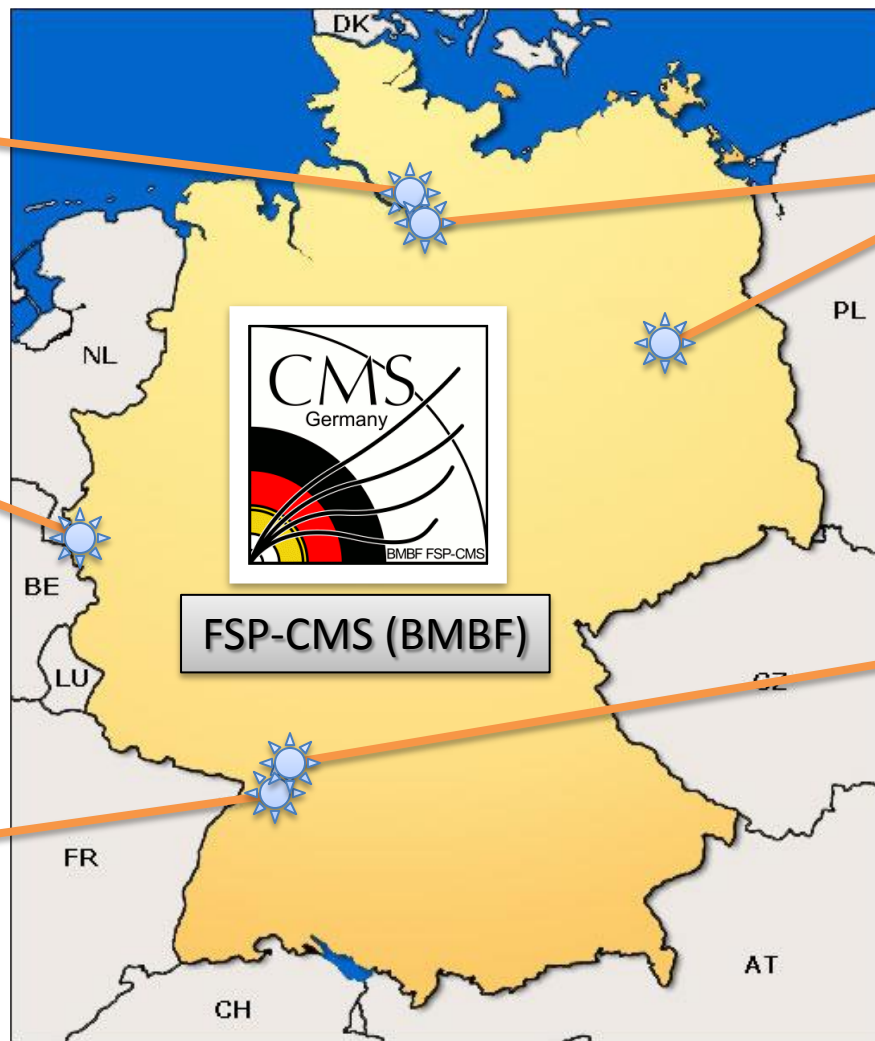
Uni Hamburg



RWTH Aachen



Uni Karlsruhe



DESY Hamburg/
Zeuthen



Forschungszentrum
Karlsruhe





Tier1 GridKa



Located at **Forschungszentrum Karlsruhe** (KIT, Campus Nord)

Part of and operated by the **Steinbuch Centre for Computing** (SCC, KIT)

Multi-VO Tier centre
(supports 8 HEP experiments)

4 LHC experiments: CMS, ATLAS, ALICE, LHCb

4 non-LHC experiments: CDF, D0, BaBar, Compass





GridKa Resources



Network:

- 10 Gbit link CERN – GridKa
- 3 x 10 Gbit link to 3 European Tier1s
- 10 Gbit link to DFN/X-Win (e.g. Tier2 connections)



Storage:

- Based on dCache (developed at DESY and FNAL)
- CMS Disk: ~ 650 TB + D-Grid: ~ 40 TB
- CMS Tape: ~ 900 TB

CPU:

- CMS Resources: ~ 800 cores, 2 GB Ram per core
 - D-Grid Resources: ~ 500 cores, 2 GB Ram per core
- D-Grid resources: partially usable for FSP-CMS

CMS GridKa Resource Pledges

	2010	2011	2012
CPU [MSI2k]	3.5	4.3	5.1
Disk [PB]	1.6	2.0	2.4
Tape [PB]	2.8	3.7	4.6





Tier2/3 and NAF Resources



■ Besides the Tier1 resources, considerable CPU and disk capacities are available in Germany 2008/2009:

■ Tier2 (German CMS Tier2 federation):

■ DESY: 400 cores, 185 TB disk

■ RWTH Aachen: 360 cores, 99 TB disk

■ Tier3:

■ RWTH Aachen: 850 cores, 165 TB disk

■ Uni Karlsruhe: 400 cores, 150 TB disk

■ Uni Hamburg: 15 TB disk

■ NAF (National Analysis Facility):

■ DESY: 245 cores, 32 TB disk

■ D-Grid resources, partially usable for FSP-CMS:

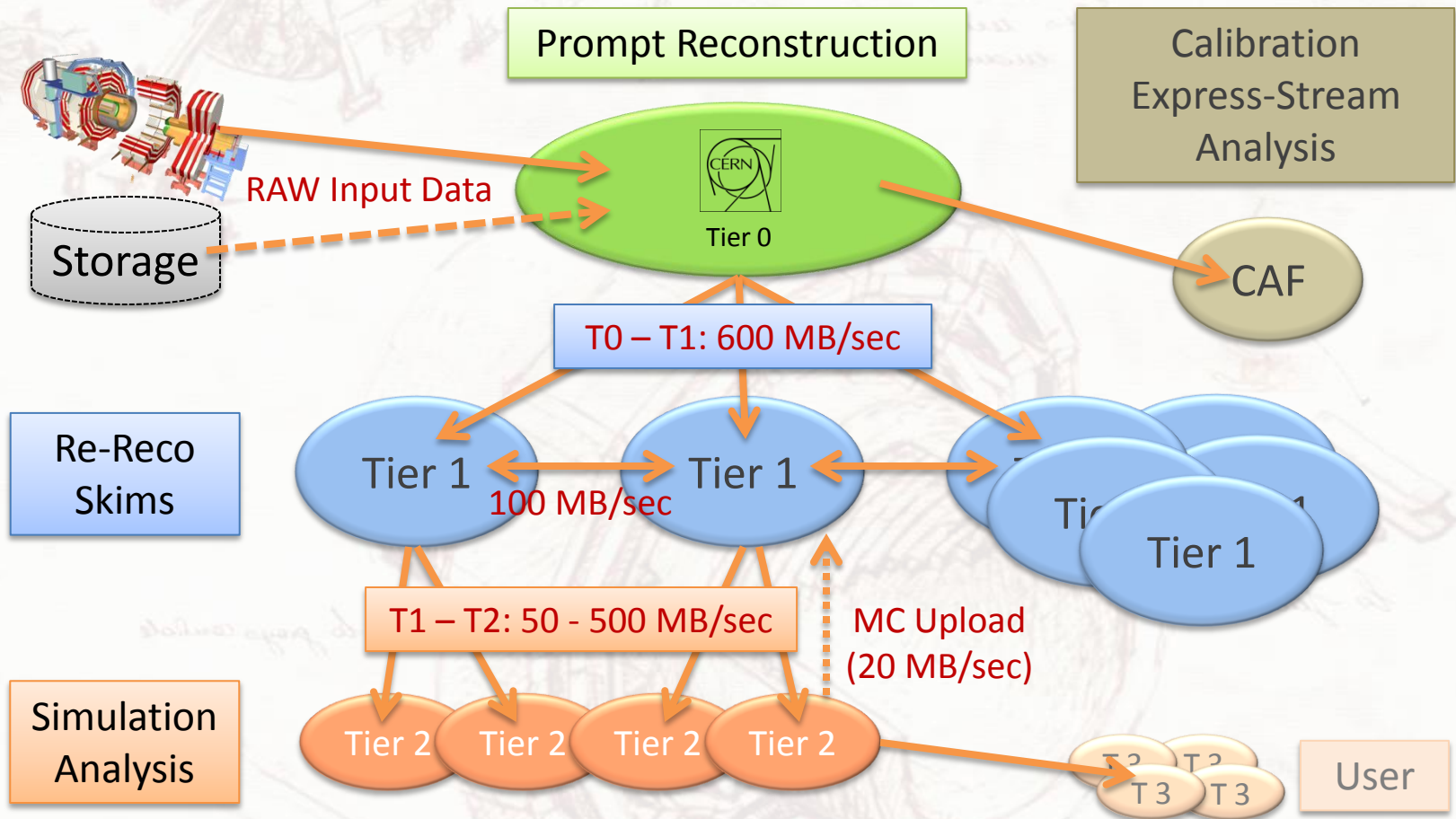
■ RWTH Aachen: 600 cores, 231 TB disk

■ GridKA: 500 cores, 40 TB disk





CMS Workflow











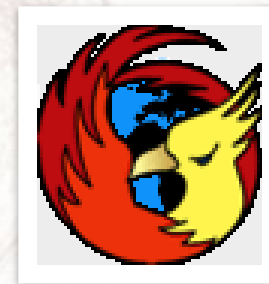


PhEDEx Data Transfers



 **Physics Experiment Data Export** (CMS data placement tool):

-  Various **agents**/daemons are run at the tier sites (depending on the tier level)
 -  upload, download, MSS stage-in/out, DB, ...
-  Provides **automatic** WAN data **transfers**, based on subscriptions
-  Automatic **load-balancing**
-  File transfer **routing** topology (FTS)
-  Automatic **bookkeeping** (database entries, logging)
-  **Consistency** checks (DB vs. filesystem)
-  File **integrity** checks

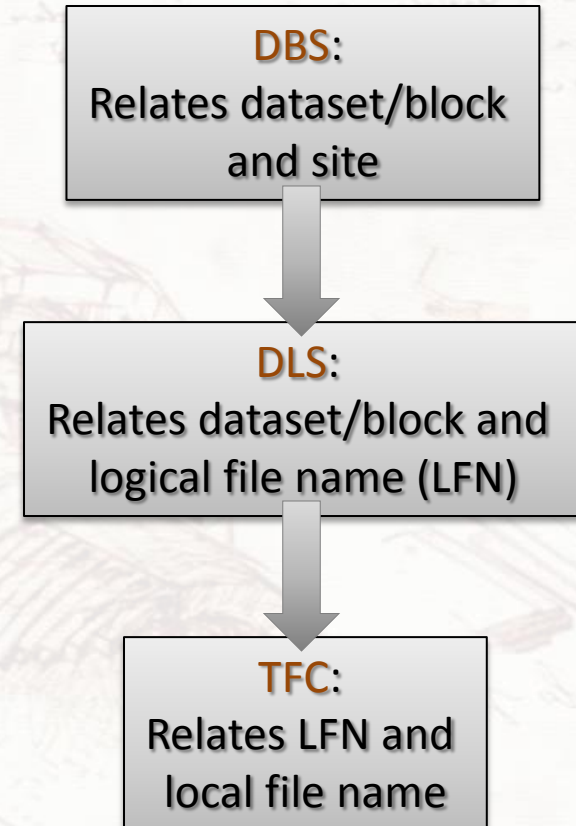




Other Involved Components



- ❏ Monte Carlo Production Agent (ProdAgent)
- ❏ CMS Remote Analysis Builder (CRAB)
- ❏ Dataset Bookkeeping System (DBS)
- ❏ Data Location Service (DLS)
- ❏ Trivial File Catalog (TFC)
- ❏ Grid middleware (gLite)
 - ❏ SE, CE, RB, UI, ...

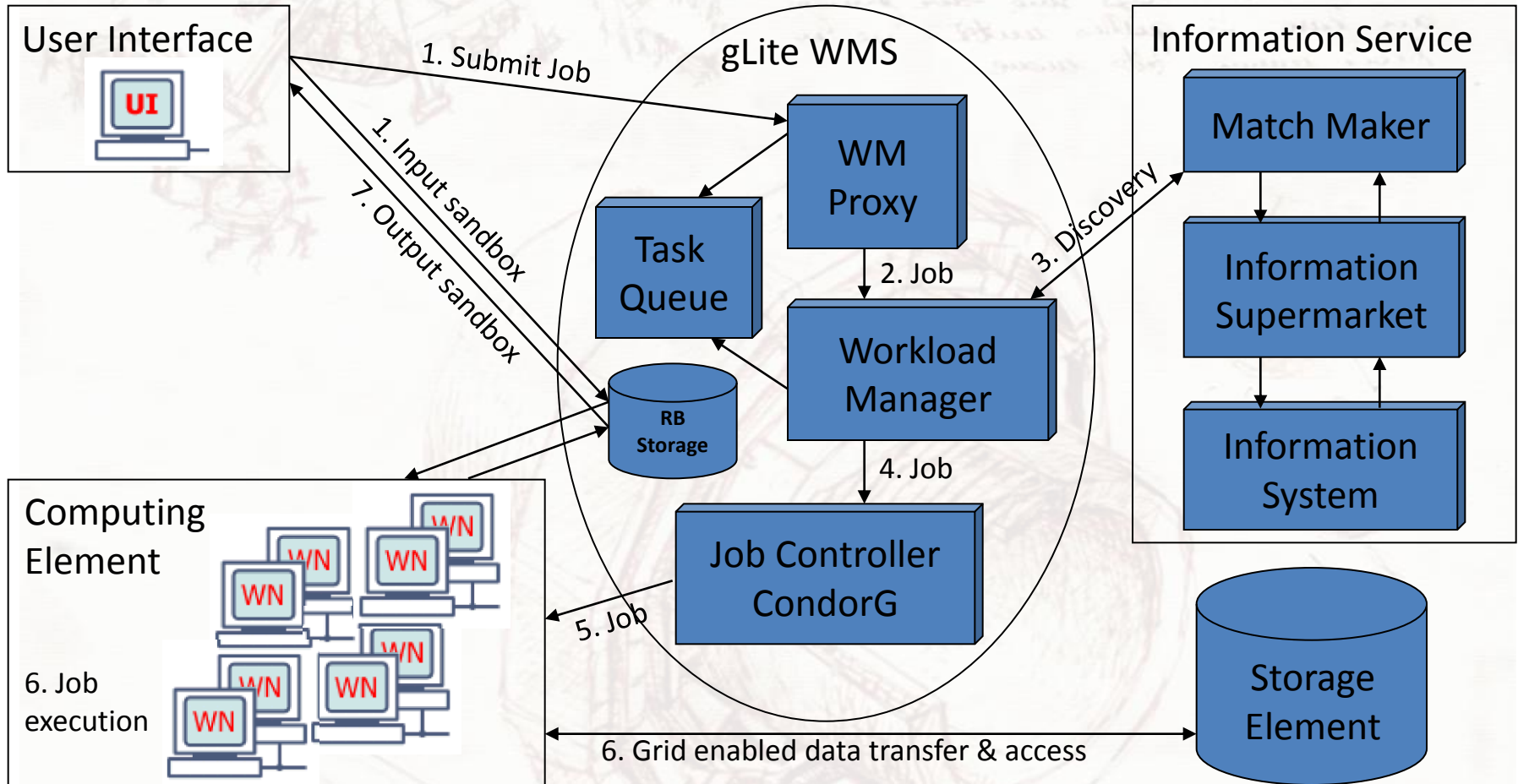


Very complex system → Needs intense testing, debugging and optimisation.





WLCG Job Workflow





CMS Service Challenges - 1



- ❏ Computing, Service and Analysis (CSA) Challenges:
Test the **readiness** for **data-taking**
- ❏ **CSA06**, **CSA07** and **CSA08** were performed to test the computing and network infrastructure and the computing resources at a level of **25%**, **50%** and **100%** required for the LHC start-up.
- ❏ The **whole CMS workflow** was tested:
 - ❏ Production/Simulation
 - ❏ Prompt reconstruction at the Tier0
 - ❏ Re-reconstruction at the Tier1s
 - ❏ Data distribution to Tier1s and Tier2s
 - ❏ Data Analyses

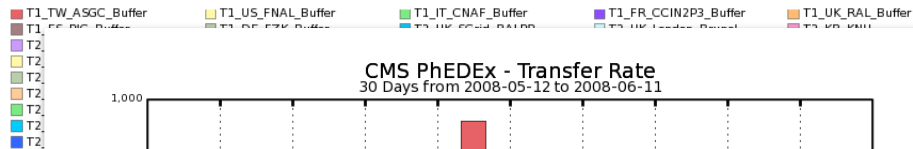
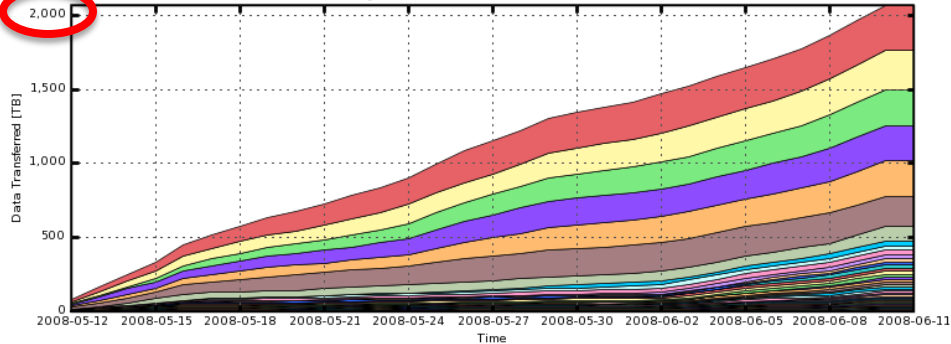




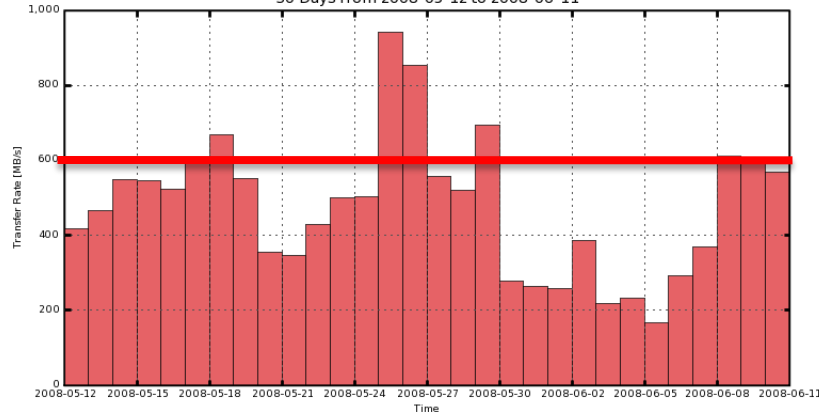
CMS Service Challenges - 2



CMS PhEDEx - Cumulative Transfer Volume
30 Days from 2008-05-12 to 2008-06-11



CMS PhEDEx - Transfer Rate
30 Days from 2008-05-12 to 2008-06-11



TO_CH_CERN_Export

Maximum: 940.68 MB/s, Minimum: 165.78 MB/s, Average: 475.14 MB/s, Current: 568.44 MB/s

Cumulative data volume transferred during CSA08:
> 2 PetaByte in 4 weeks

Transfer rate from the Tier0 to the Tier1s:

< 600 MB/sec

Goal was achieved only partially, problems with:

Storage systems (CASTOR at CERN)

Problems identified and fixed afterwards.





jobs per site

Site	submitted	app-succeeded	app-failed	app-unknown	pending	running	aborted	cancelled
CERN-PROD (Geneva,Switzerland)	0	1,150,000	100,000	100,000	100,000	0	50,000	50,000
unknown	0	0	0	0	650,000	0	100,000	20,000
USCMS-FNAL-WC1-CE2 (Batavia ,USA)	0	280,000	50,000	150,000	0	0	10,000	10,000
GLW-CMS (Madison, USA)	0	250,000	50,000	100,000	0	0	10,000	0
FZK-LCG2 (Karlsruhe, Germany)	0	280,000	50,000	100,000	0	0	10,000	0
USCMS-FNAL-WC1-CE (Batavia ,USA)	0	280,000	50,000	50,000	0	0	10,000	10,000
USCMS-FNAL-WC1-CE4 (Batavia ,USA)	0	280,000	50,000	50,000	0	0	10,000	10,000
Taiwan-LCG2 (Taipei,Taiwan)	0	250,000	50,000	50,000	0	0	10,000	0
Nebraska (Lincoln,NE,USA)	0	200,000	50,000	50,000	0	0	10,000	0
GRIF (Orsay,France)	0	180,000	50,000	50,000	0	0	10,000	0
IN2P3-CC-12 (Lyon,France)	0	180,000	50,000	50,000	0	0	10,000	0
RWTH-Aachen (Aachen, Germany)	0	150,000	50,000	100,000	0	0	10,000	0
INFN-PISA (Pisa,Italy)	0	150,000	50,000	50,000	0	0	10,000	0
DESY-HH (Hamburg,Germany)	0	150,000	50,000	50,000	0	0	10,000	0
INFN-T1 (Bologna,Italy)	0	150,000	50,000	50,000	0	0	10,000	0
INFN-T1 (Bologna,Italy)	0	150,000	50,000	50,000	0	0	10,000	0
INF2P3-CC (Lyon,France)	0	150,000	50,000	50,000	0	0	10,000	0
IFCA-LCG2 (Santander,Spain)	0	150,000	50,000	50,000	0	0	10,000	0
CIEMAT-LCG2 (Madrid,Spain)	0	150,000	50,000	50,000	0	0	10,000	0

number of jobs

submitted app-succeeded app-failed app-unknown pending running aborted cancelled



Results of CMS Challenges



Service	Goal 2008	Status 2008	Goal 2007	Status 2007	Goal 2006	Status 2006
Tier-0 Reco Rate	150-300 Hz	Achieved	100 Hz	Only at bursts	50 Hz	Achieved
Tier-0 → Tier-1 Transfer Rate	600 MB/sec	Achieved partially	300 MB/sec	Only at bursts	150 MB/sec	Achieved
Tier-1 → Tier-2 Transfer Rate	50-500 MB/sec	Achieved	20-200 MB/sec	Achieved partially	10-100 MB/sec	Achieved
Tier-1 → Tier-1 Transfer Rate	100 MB/sec	Achieved	50 MB/sec	Achieved partially	N/A	-
Tier-1 Job Submission	50 000 jobs/day	Achieved	25 000 jobs/day	Achieved	12 000 jobs/day	3 000 jobs/day
Tier-2 Job Submission	150 000 jobs/day	Achieved	75 000 jobs/day	20 000 jobs/day	48 000 jobs/day	Achieved
Monte Carlo Simulation	1.5×10^9 events/year	Achieved	50×10^6 events/month	Achieved	N/A	-





STEP09 Activities @ FZK

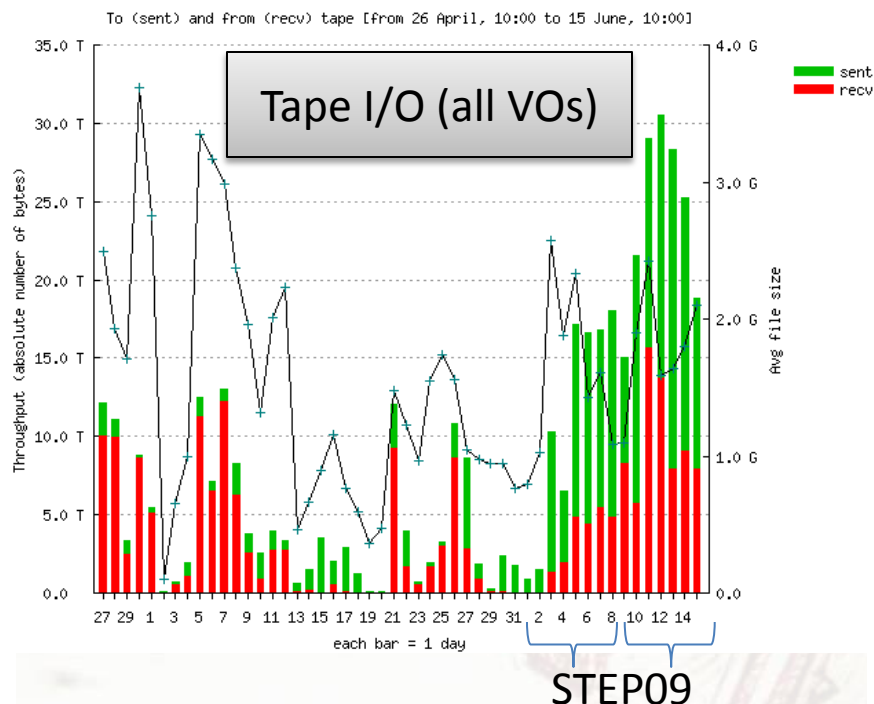


- ❏ STEP09 was the first scheduled large-scale multi-vo test within the WLCG.
- ❏ The following activities have been performed at GridKa during STEP09:
 - ❏ 1st STEP09 week (all files were on disk):
 - ❏ Transfers T0 → T1
 - ❏ Transfers T1 → T1
 - ❏ All AODSIM data were successfully written to tape after transfer.
 - ❏ Transfers T1 → T2
 - ❏ Reprocessing tasks
 - ❏ 2nd STEP09 week (files were flushed from disk):
 - ❏ Same tasks like during the first week





STEP09: Tape Performance



Tape I/O:

1st week: 10 -15 TB/day

mostly writing + reading

2nd week: 25-30 TB/day

reading + writing

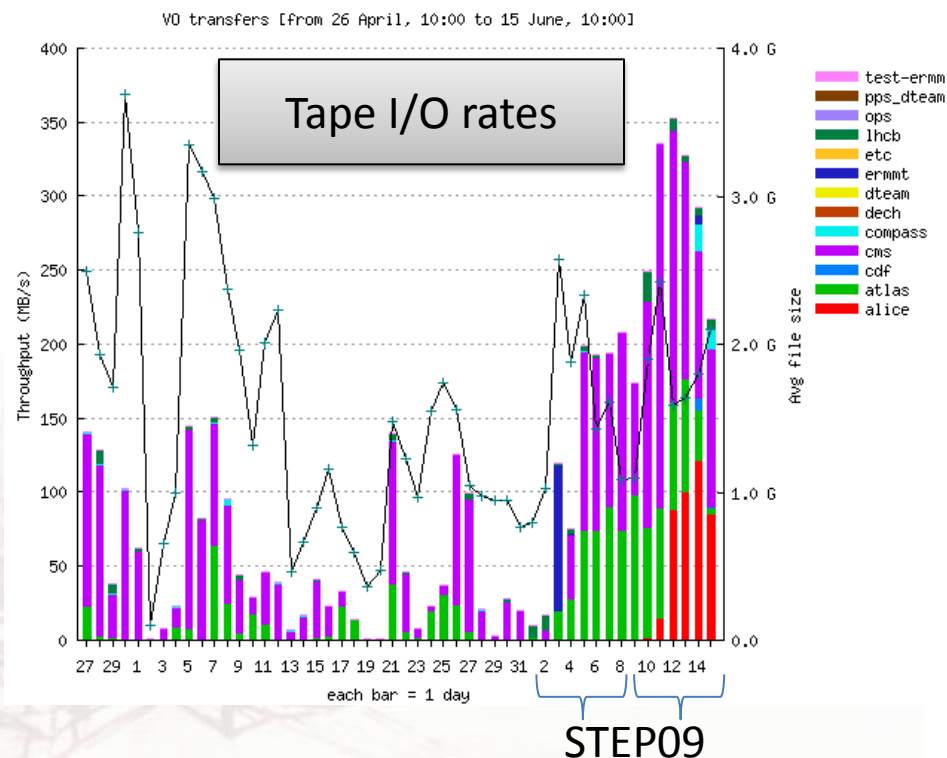
Tape I/O rates:

1st week: 150 MB/sec

mostly writing + reading

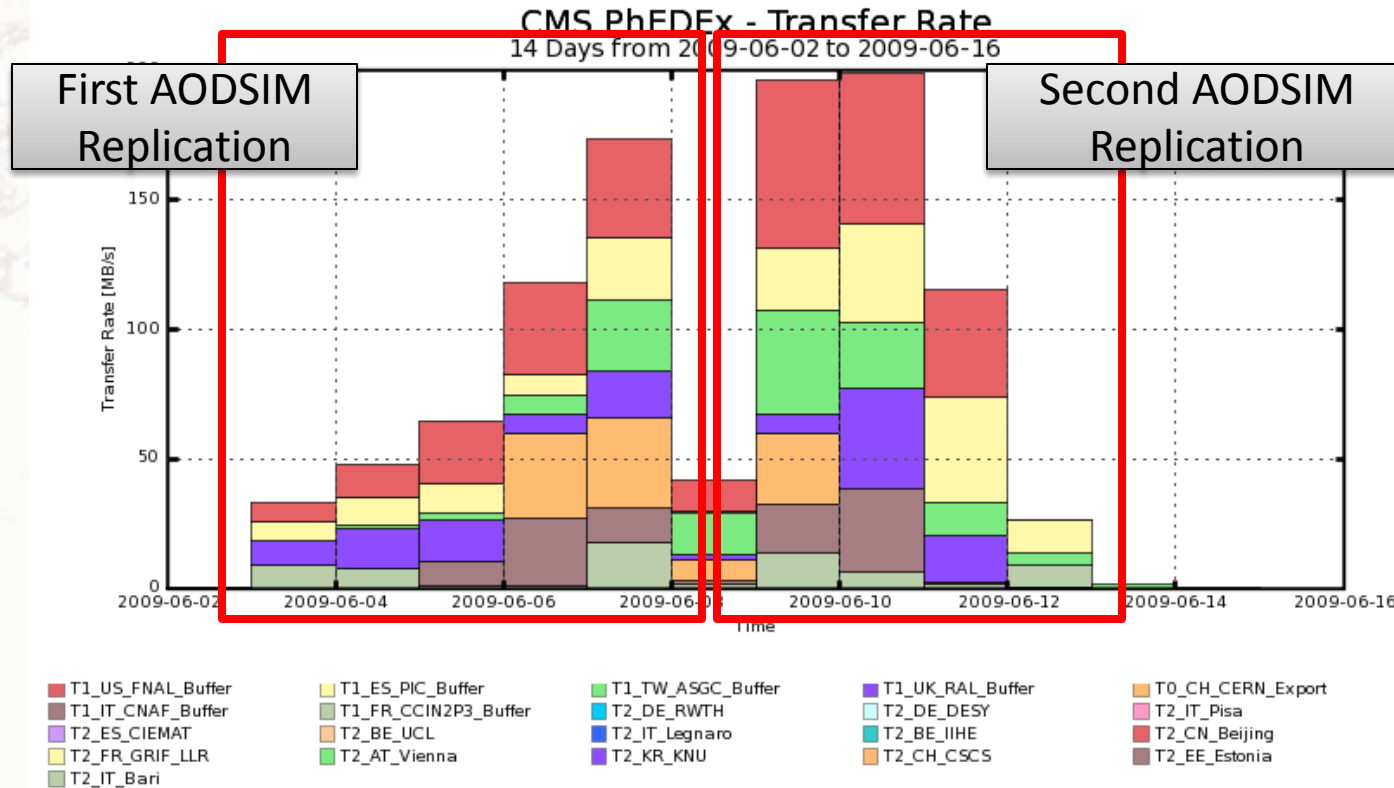
2nd week: 250 MB/sec

reading + writing





STEP09: Transfer Imports



Maximum: 199.03 MB/s, Minimum: 0.09 MB/s, Average: 78.36 MB/s, Current: 0.09 MB/s

T0, T1 and T2 imports to GridKa during STEP09

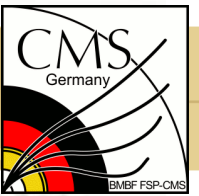
First week: 30 – 170 MB/sec, Second week: 25 – 200 MB/sec

All AODSIM data **successfully written to tape.**





Summary



- ❏ CMS uses Grid technology for data access and processing.
- ❏ Altogether the WLCG/CMS computing model is very complex and needs intense testing.
- ❏ Multiple service challenges have proven the readiness for first data.
- ❏ PhEDEx is used for reliable large-scale data management.
- ❏ Experience has shown, that monitoring is indispensable for the operation of such a heterogeneous system.
- ❏ More information about the requirement of monitoring is presented later in this session.

