LHC Computing in der Hochenergiephysik

Hartmut Stadie
Universität Hamburg

LSDMA 2013 Spring Meeting
11-13 March 2013

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

SPONSORED BY THE
Federal Ministry
of Education
and Research

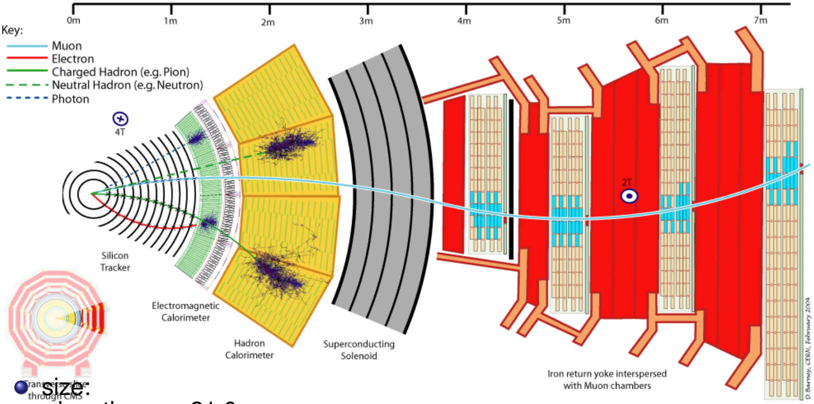
Outline

- Introduction
  - The LHC
  - The data
- Computing model
  - Motivation
  - The CMS computing model
- Current status
  - Tier-0/1
  - Tier-2
- Data management at a site
  - Data management at a site
- Conclusion
  - Conclusion

The Large Hadron Collider (LHC)

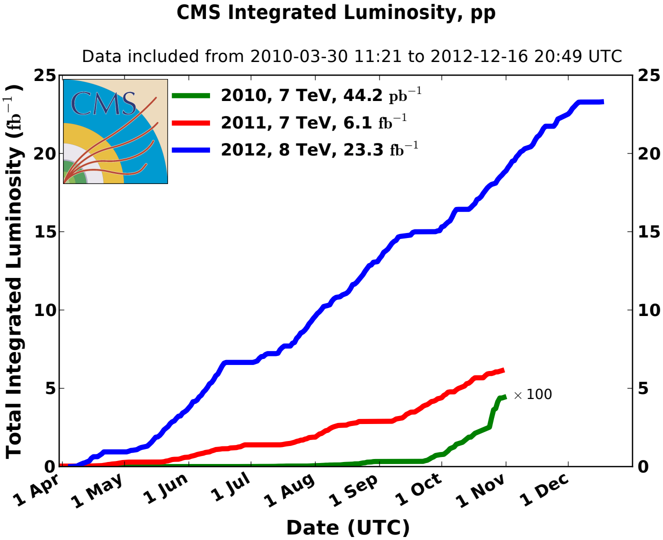
- proton-proton collider
- circumference: 26.66 km
- $\sqrt{s} = 7 - 8$ TeV
- crossing rate: 50 ns rate: 20 MHz
- data per crossing (event): $\sim$ 1 MB
- collider experiments: ALICE, ATLAS, CMS, LHCb,...
- online filter: output rate $\sim$ 500 MB/s

The CMS Experiment


0m
1m
2m
3m
4m
5m
6m
7m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
4T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
Transverse slice through CMS
D.Barney, CERN, February 2004

- size:
  length: 21.6 m
  diameter: 14.6 m
- mass: 12,500 t
- magnetic field:
  solenoid: 3.8 T

Current Status



Data Organization

Physicists analyze the data requiring specific signatures (final states/channels).

Collider data:

- split directly into O(10) primary data sets (PD) based on event signatures
- every PD in 2012 consists of roughly 100 TB

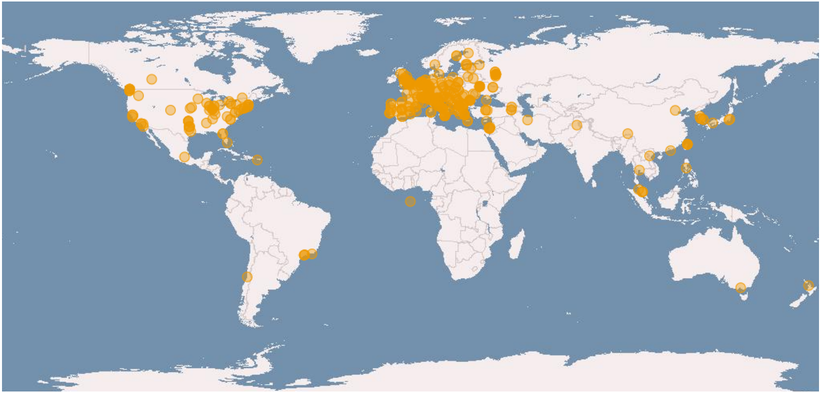
Simulated events:

- comparisons with simulation to find deviations from the standard theory (e.g., new particles like the Higgs boson), measure properties
- overall size (in transfer DB): at least 10 PB

A typical analysis of the 2012 data might run over 200 TB a couple of times.

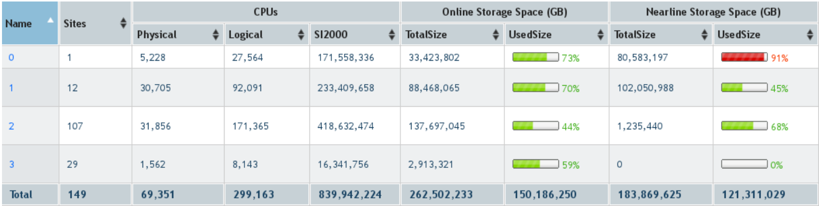
Worldwide LHC Computing Grid (WLCG)

- provides resources and services to LHC experiments
- highly distributed system



Worldwide LHC Computing Grid (WLCG)

current overall resources sorted by the different classes of computing sites (Tiers):

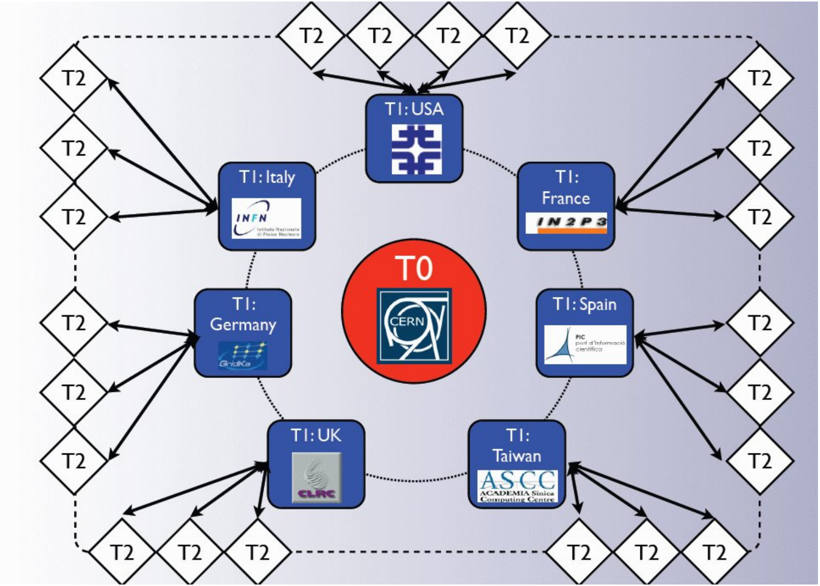


Name	Sites	CPUs			Online Storage Space (GB)		Nearline Storage Space (GB)	
		Physical	Logical	SI2000	TotalSize	UsedSize	TotalSize	UsedSize
0	1	5,228	27,564	171,558,336	33,423,802	73%	80,583,197	91%
1	12	30,705	92,091	233,409,658	88,468,065	70%	102,050,988	45%
2	107	31,856	171,365	418,632,474	137,697,045	44%	1,235,440	68%
3	29	1,562	8,143	16,341,756	2,913,321	59%	0	0%
Total	149	69,351	299,163	839,942,224	262,502,233	150,186,250	183,869,625	121,311,029

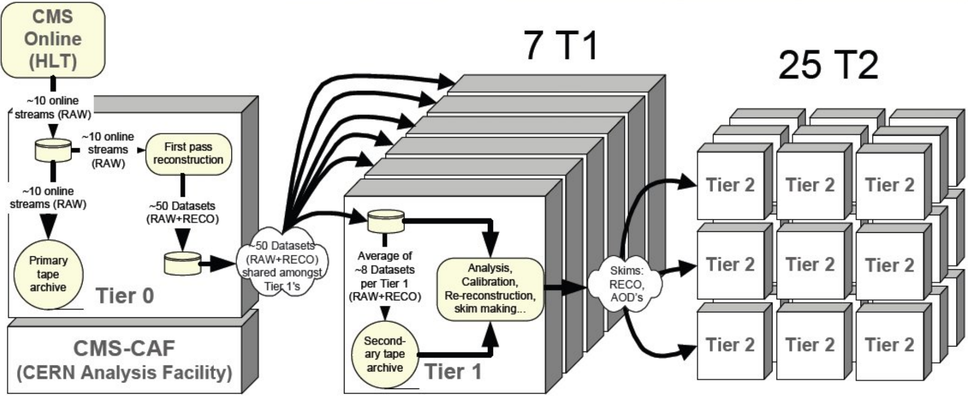
on top of bare resources:

- fabric: batch systems, storage system
- tools for resource sharing
- collective tools

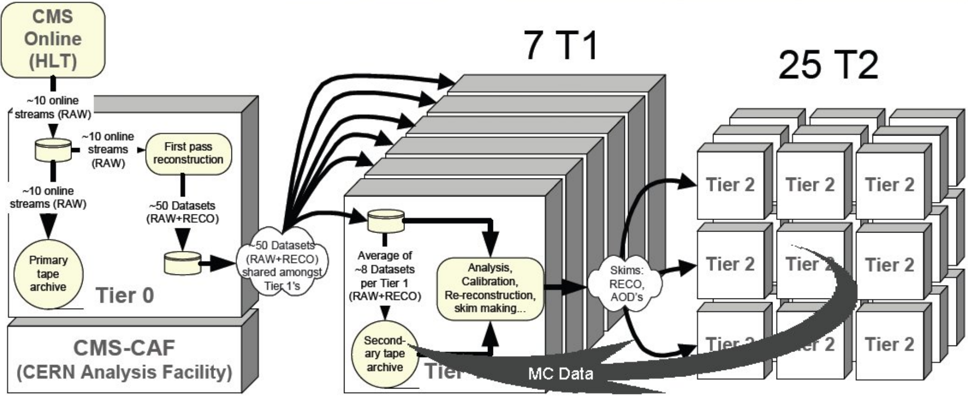
CMS Computing Model



Data Flow


CMS Online (HLT)
~10 online streams (RAW)
~10 online streams (RAW)
First pass reconstruction
~10 online streams (RAW)
~50 Datasets (RAW+RECO)
Primary tape archive
Tier 0
CMS-CAF (CERN Analysis Facility)
~50 Datasets (RAW+RECO) shared amongst Tier 1's
7 T1
Average of ~8 Datasets per Tier 1 (RAW+RECO)
Analysis, Calibration, Re-reconstruction, skim making...
Second-ary tape archive
Tier 1
Skims: RECO, AOD's
25 T2
Tier 2
Tier 2
Tier 2
Tier 2
Tier 2
Tier 2
Tier 2
Tier 2
Tier 2

Data Flow (MC)


CMS Online (HLT)
~10 online streams (RAW)
~10 online streams (RAW)
First pass reconstruction
~10 online streams (RAW)
~50 Datasets (RAW+RECO)
Primary tape archive
Tier 0
CMS-CAF (CERN Analysis Facility)
~50 Datasets (RAW+RECO) shared amongst Tier 1's
7 T1
Average of ~8 Datasets per Tier 1 (RAW+RECO)
Analysis, Calibration, Re-reconstruction, skim making...
Second-ary tape archive
Tier
MC Data
Skims: RECO, AOD's
25 T2
Tier 2
Tier 2
Tier 2
Tier 2
Tier 2
Tier 2
Tier 2
Tier 2
Tier 2

Experiment-Specific Services

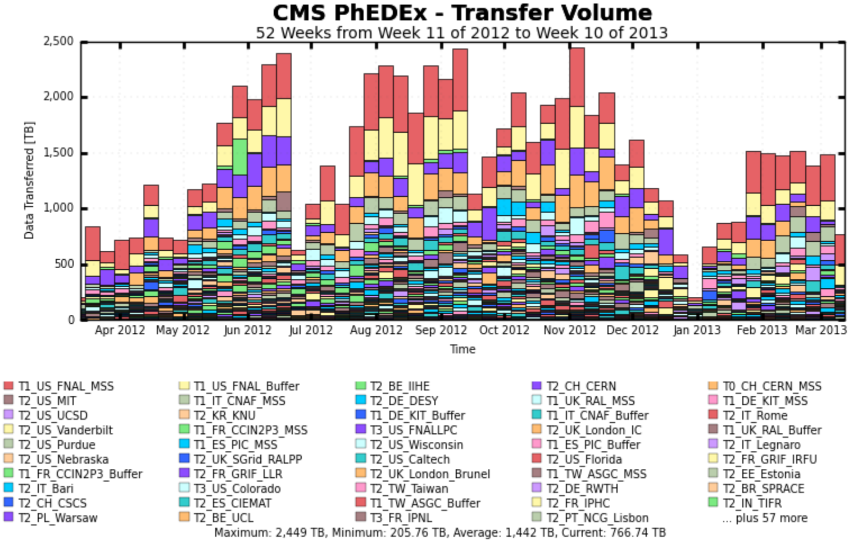
Experiment-Specific Services:

needed for distributed computing:

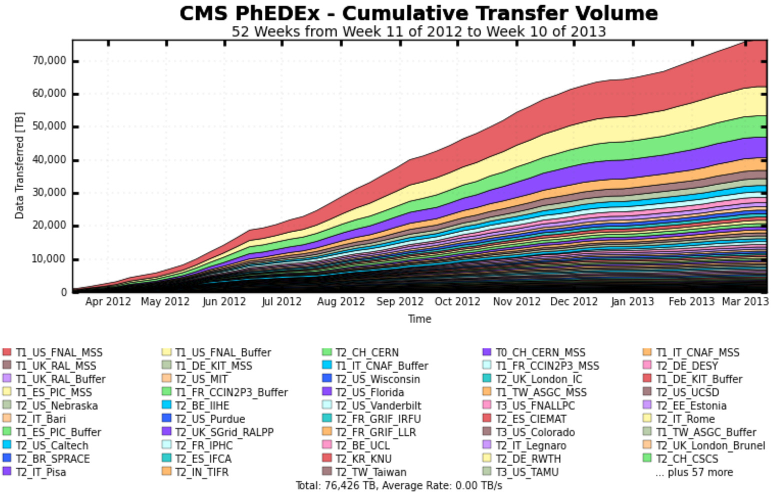
- production agents based on grid tools (WLCG job submission)
- dataset database (DBS) and trivial file catalog at sites
- dataset transfer service (PhEDEx)
    - uses grid tools (FTS, SRM) (as of 5 years ago)
    - special interfaces to Castor, dCache, etc) for file validation (checksums) and integrity tests
    - DB for transfers, dataset locations, commissioned links
    - agents for scheduling transfers, consistency checks, deletion, etc
    - transfers requests need to be approved by data manager of destination site
- (distributed) calibration database (FroNTier)
  squid web cache
- analysis job submission tool (CRAB)
  grid tools (WLCG job submission), bridges to local batch systems

- Introduction
  - The LHC
  - The data
- Computing model
  - Motivation
  - The CMS computing model
- Current status
  - Tier-0/1
  - Tier-2
- Data management at a site
  - Data management at a site
- Conclusion
  - Conclusion

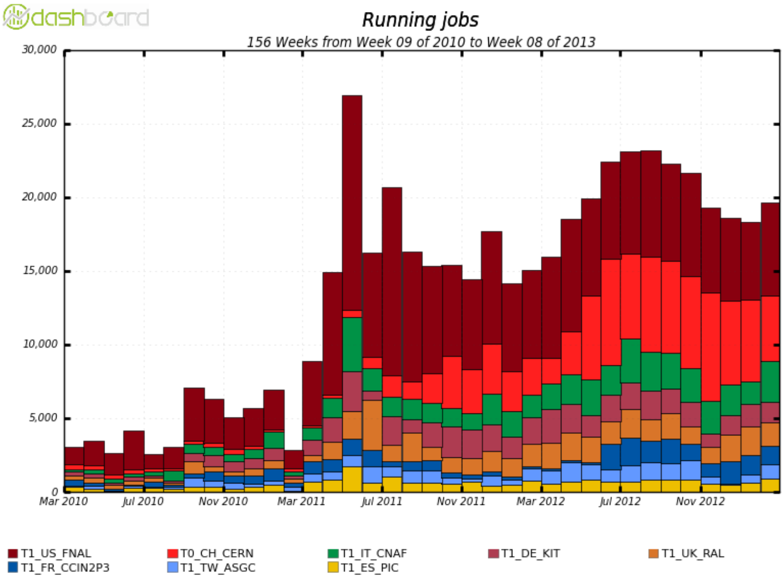
Data transfers by destination (volume per week)



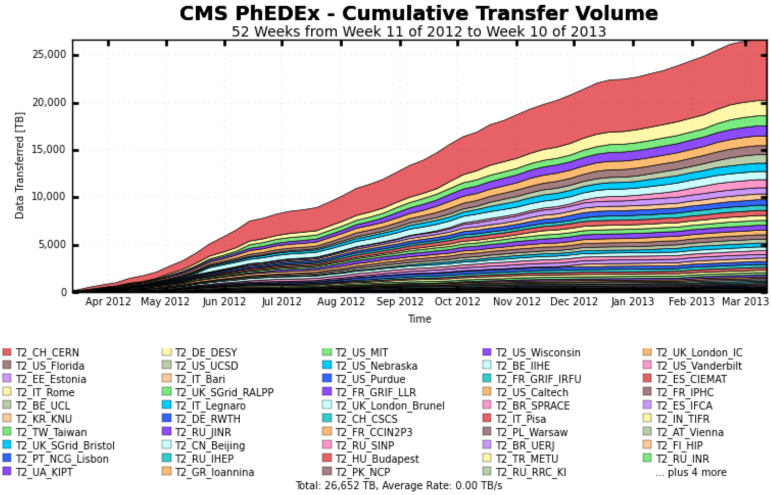
Data transfers by destination (integrated)



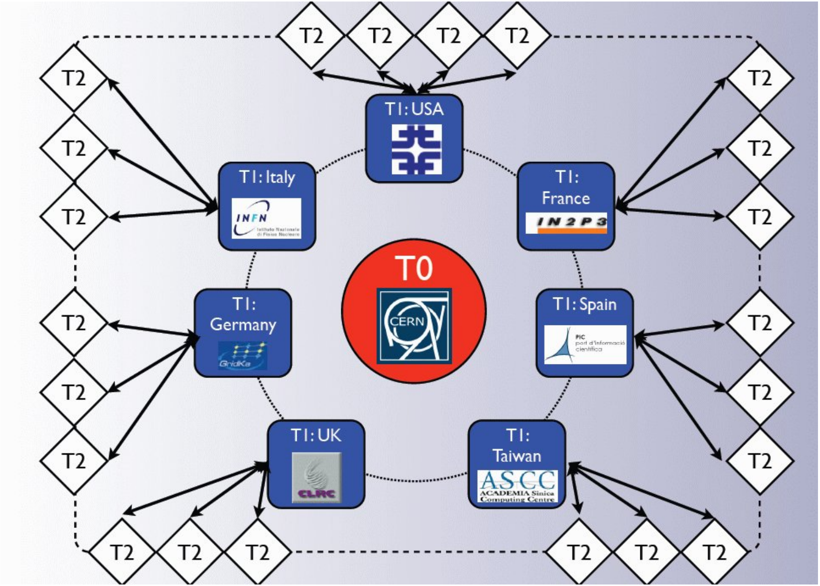
Tier-0/1 data processing obs



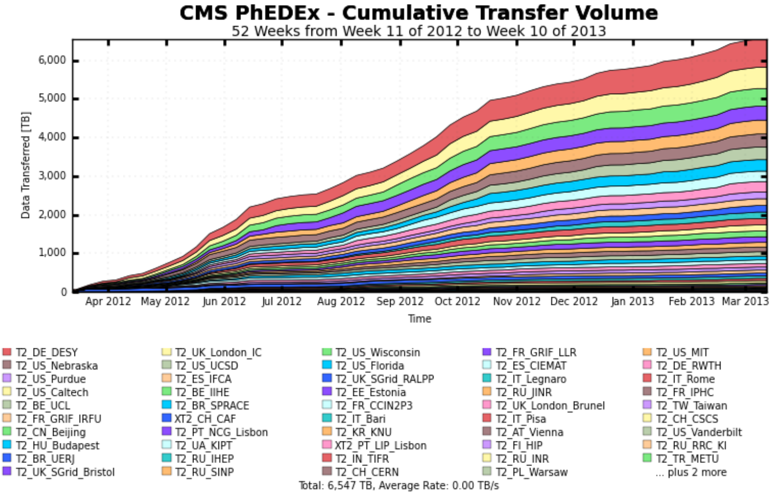
Tier-2 data transfers by destination



CMS Computing Model

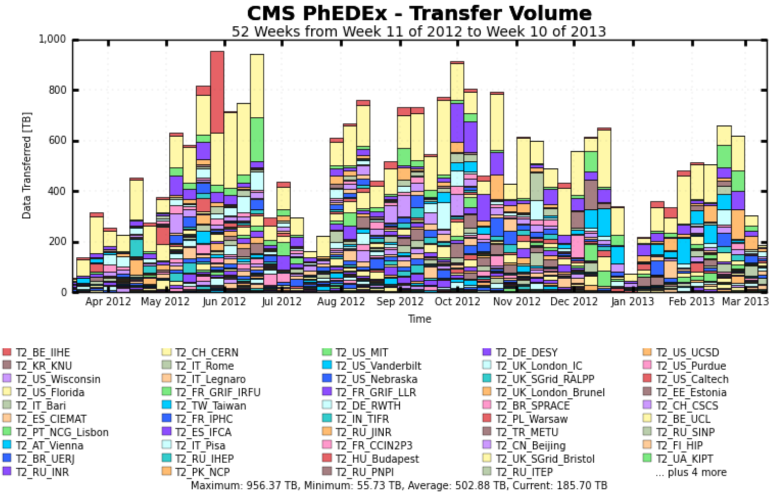

T0
CERN
T1: USA
T1: Italy
INFN
Istituto Nazionale di Fisica Nucleare
T1: France
IN2P3
T1: Germany
GridKa
T1: Spain
PIC port d'informació científica
T1: UK
CLRC
T1: Taiwan
ASCC
ACADEMIA Sinica Computing Centre
T2

Tier-2 data transfers between Tier-2s by source

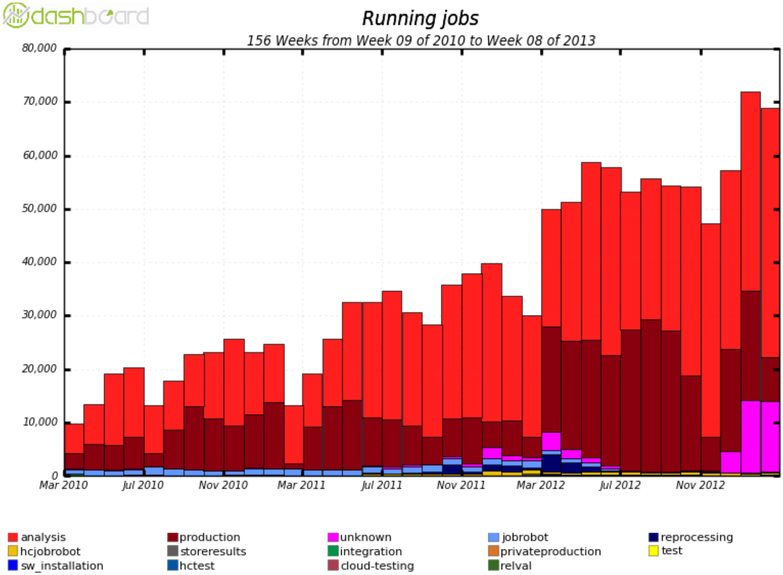


T2-T2 fraction: 25%

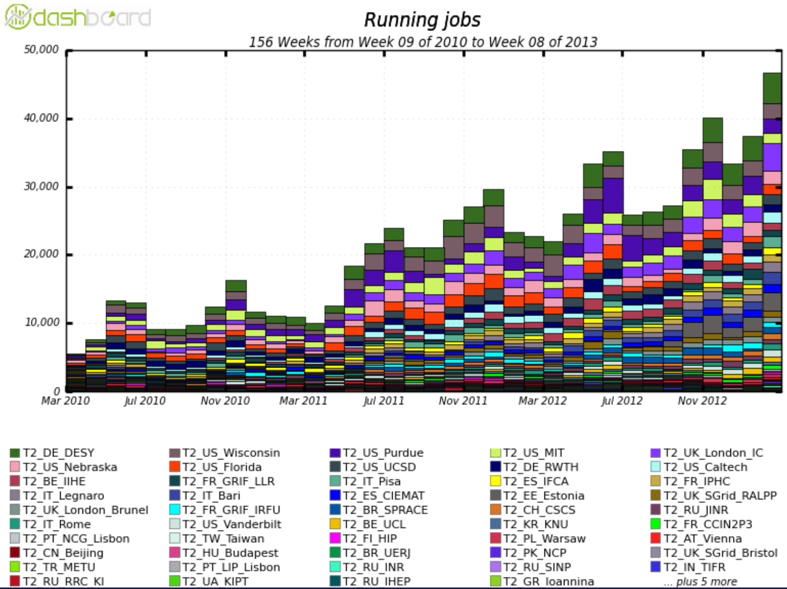
Tier-2 data transfers (volume per week)



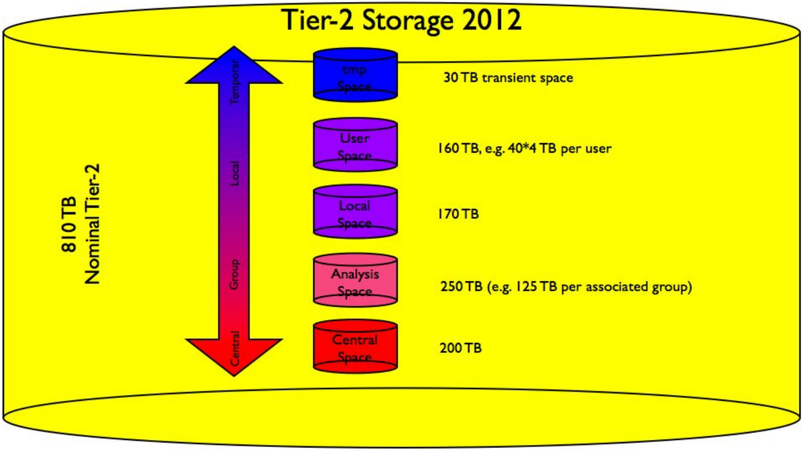
Tier-2 activities



Analysis jobs per site



Data Management at a Tier-2 (e.g. DESY)



DESY houses National Analysis Facility: local space > 1 PB

Management of local data

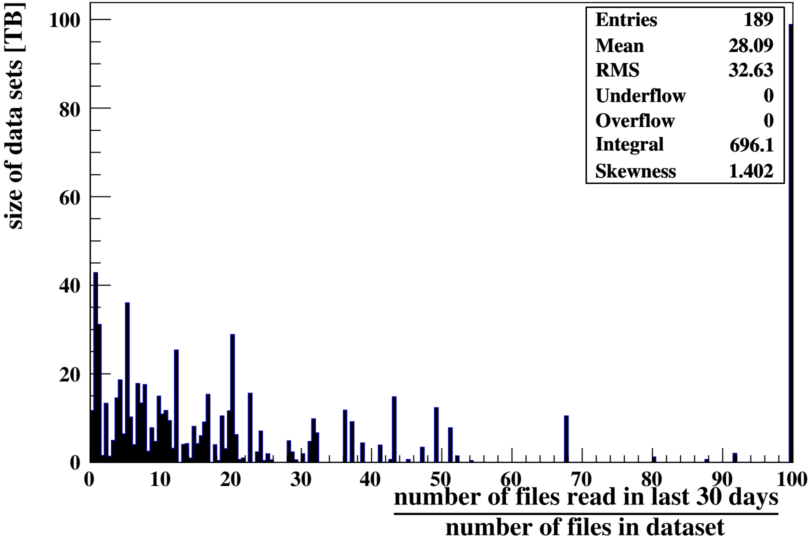
Workflow

- users request datasets
- data manager approves request if justified
- data manager also identifies and requests samples of common interest
- every month popularity is evaluated based on dCache access logs
- unused samples are scheduled for deletion after a grace period

Problems

- PhD thesis lasts three years (requests for outdated data)
- users hardly request the deletion of old data
- no trust in file transfers, do not see Tier-2 storage as a cache
- want to have all data at one place

Example for January



Entries	189
Mean	28.09
RMS	32.63
Underflow	0
Overflow	0
Integral	696.1
Skewness	1.402

Conclusions

Lay person's conclusions

- LHC experiments deal with large data
- use "divide et impera" to break problems down
- resources and middleware from WLCG
- experiment specific tools needed
- good separation and interfaces between sites for scaling needed
- data management very difficult (centralized systems to rigid, by-demand/request system needs resources to scale with data)