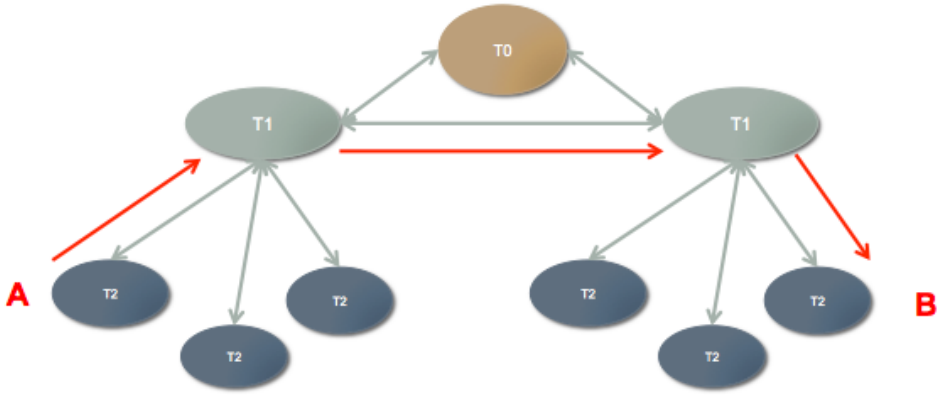
CMS Computing Model Evolution

Claudio Grandi

INFN Bologna

On behalf of the CMS Collaboration

Computing in CMS Run1



Start from MONARC model

Securing of data and transfer to processing sites

PhEDEx

Metadata and conditions

DBS, Frontier

Data processing

(ProdAgent), WMAgent, CRAB

Infrastructure monitoring

Dashboard

Improve data distribution

Full site mesh

Improve job management

Pilot jobs, glidein-WMS

Improve use of storage

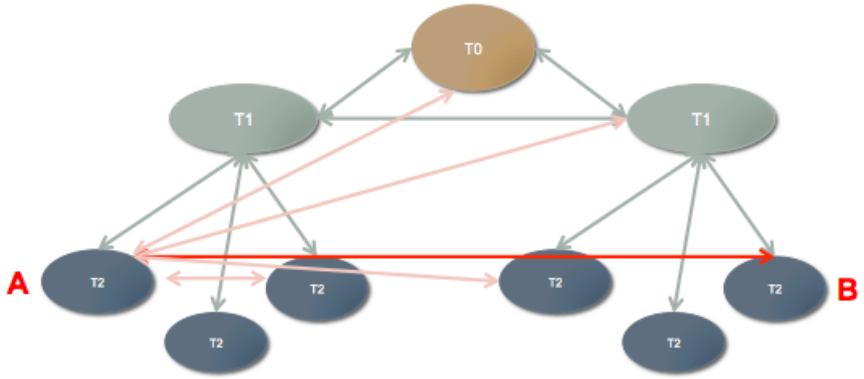
Data Popularity, Victor

Improve data access

Remote access, xrootd

Improve software distribution

CVMFS



Moving to Run2

LHC increased beam energy and intensity

(25-30 pile up events with beam spacing at 25 ns)

➔ { Factor 2.5 in reconstruction time and 30% in AOD size
  { Additional factor 2 due to out-of-time pileup can be avoided

Trigger rate: 0.8-1.2 kHz (to preserve physics potential)

➔ Factor of 2.5 in number of events

A factor of 6 in computing resources would be needed in no changes are applied to the computing system

Rationalization

New platforms

Automation

Opportunistic resources

➔Flexibility➔

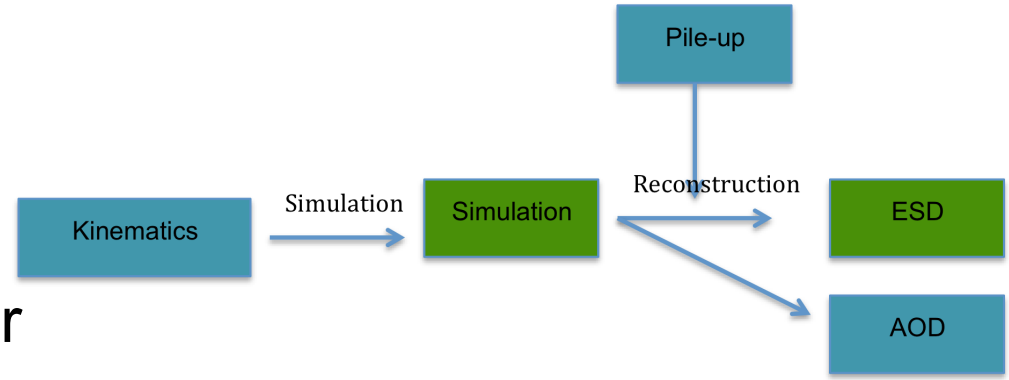
CMS Data Model and workflows

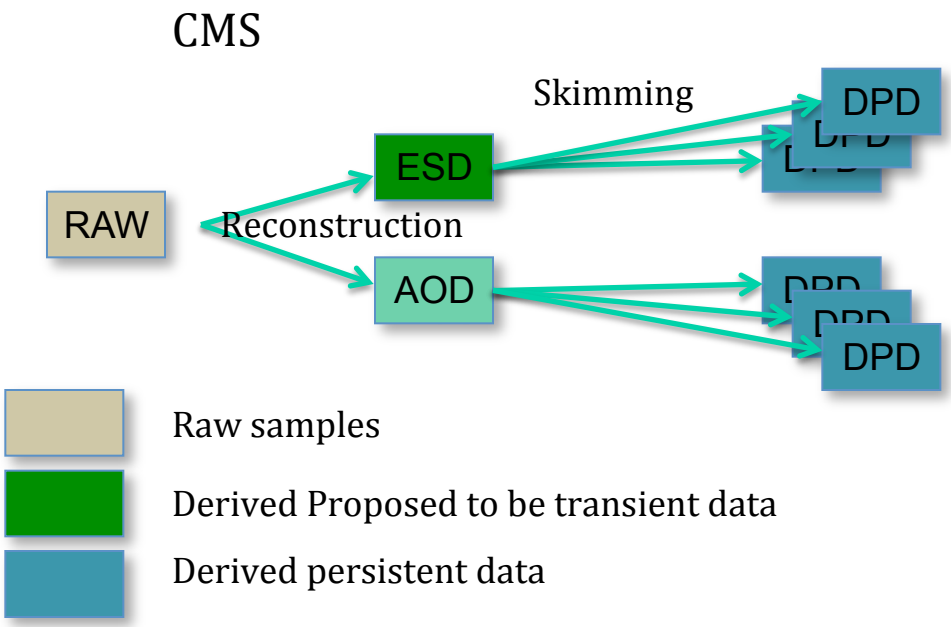
AOD base for analysis

ESD proposed to be transient for most datasets

One re-reconstruction per year

AOD compression





MC/data ratio gradually reduced from 2 to 1

Bunch crossing pre-mixing

Up to 1000 min. bias per event

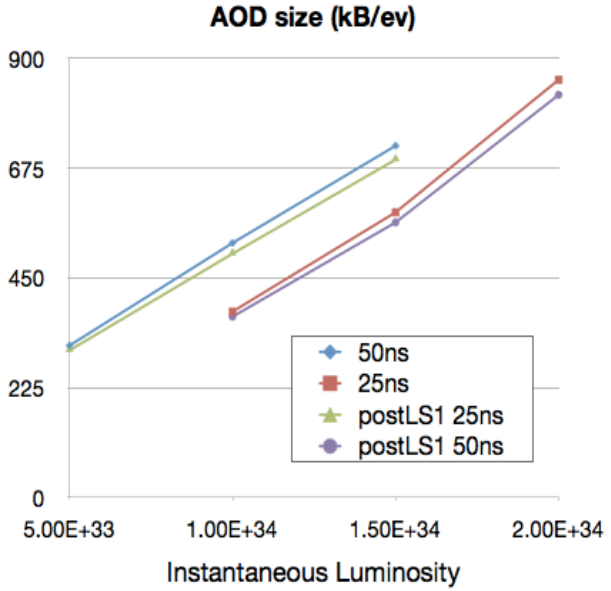
Fast simulation

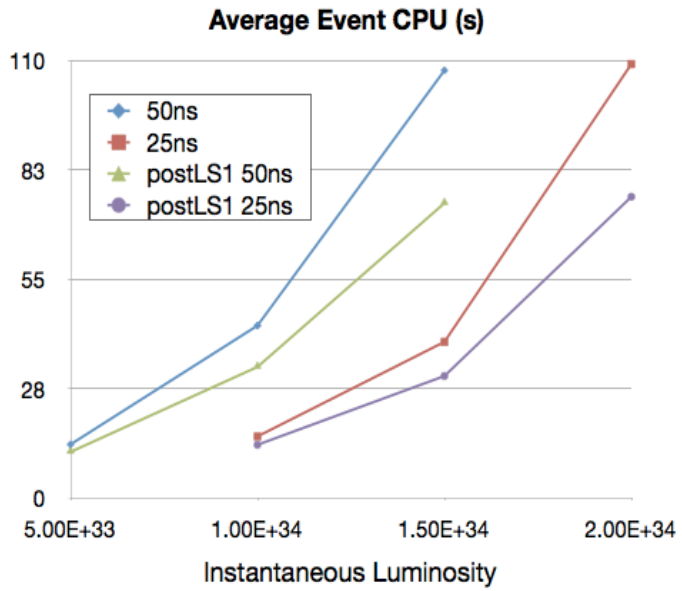
Offline performance

Up to one order of magnitude gained in code speed during Run1 for high luminosity events

Further improvements from better treatment of pile-up

~ 30% reduction in processing time at high-lumi (for the same events)





Multithreaded CMSSW in production in Autumn 2013

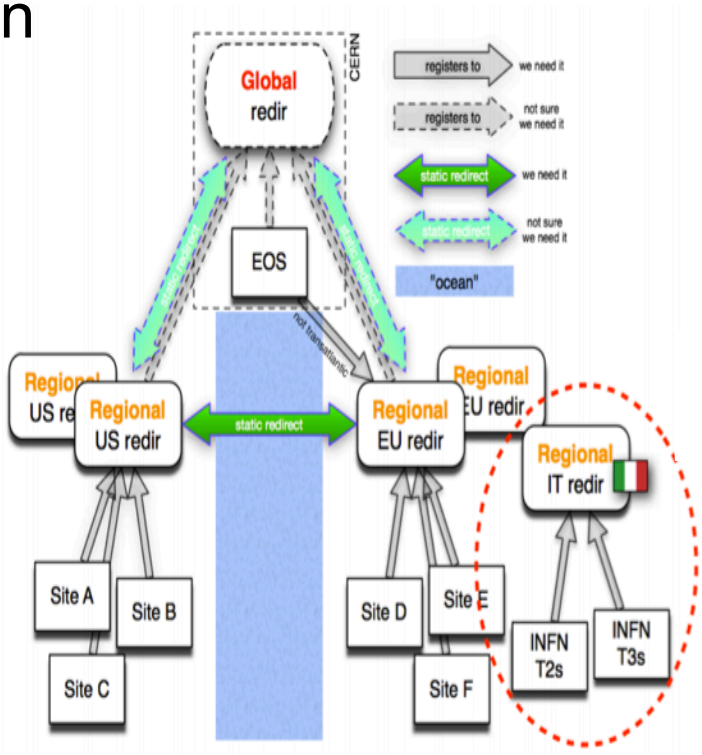
Become compatible with heterogeneous resources

ARM, Xeon Phi, GPGPU, PowerPC farms (Blue Gene)

Compatible with remote I/O

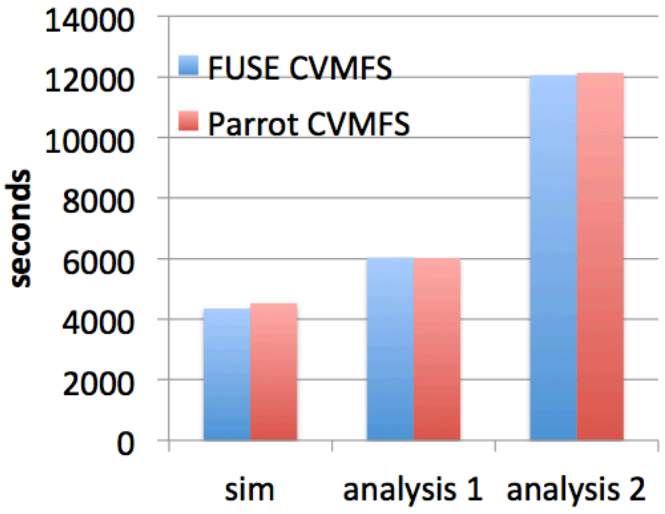
Data Management 1/2

- Pre-placement remains the main method for efficient access to data
  - Addition of tools to automatize data transfer and removal based on a Data Popularity System
- Devote a fraction of the storage to act as a site cache
  - Automatic replication and cancellation
- Storage federation superimposed to current structure
  - Remove the data locality requirement and add flexibility
    - Fall-back in case of missing files
    - Low-rate activities (e.g. visualization)
    - Diskless Tier-3 and opportunistic sites
  - Efficient access provided by a hierarchy of xrootd redirectors



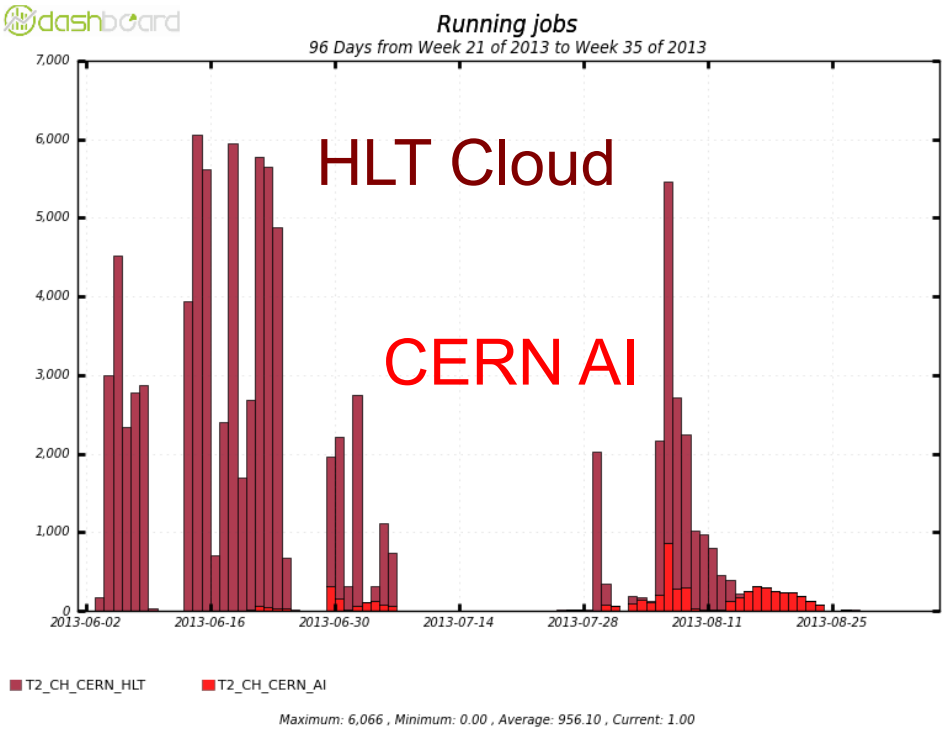
Data Management 2/2

- Opportunistic storage: separate storage management and data transfer in PhEDEx
  - Manage temporary storage at sites without CMS manpower
- Conditions data access will continue to be based on Frontier and squid caches
- Software distribution now based on CVMFS
  - Automatic cache management (also based on squid)
  - Very little pre-configuration required at sites
  - Migration almost terminated
- Possibility to use CVMFS at opportunistic sites via Parrot
  - CVMFS mounted in user space, does not require root privileges
  - Performance penalty is limited



Workload Management

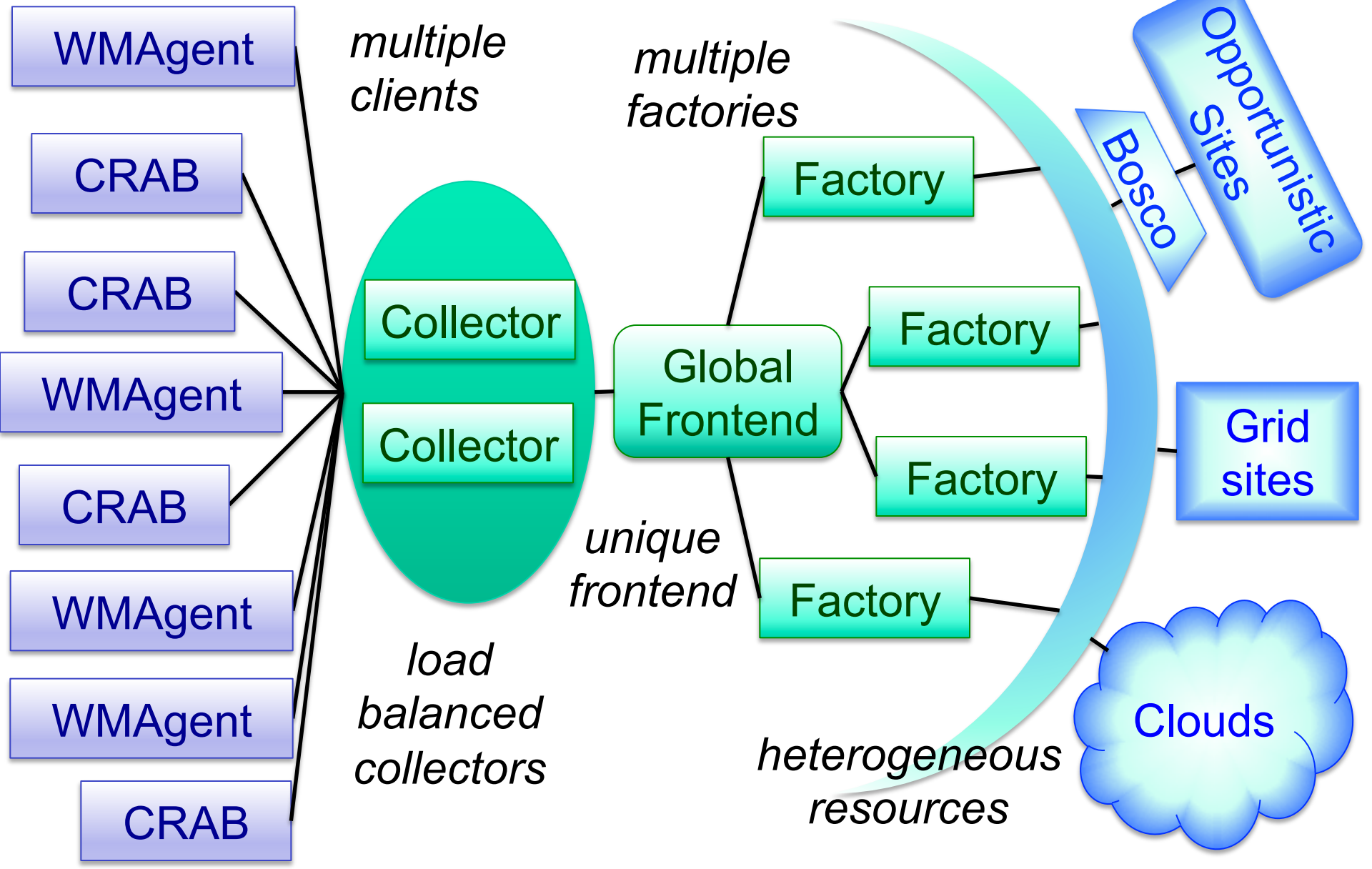
- Separation of resource allocation and job management
  - Via the glidein-WMS
  - Support of Clouds and opportunistic resources in addition to Grids are natural extensions



Use of CVMFS and remote data access are key elements for an easy adaptation of the system to Clouds and opportunistic resources

BOSCO is a thin layer that allows submission of glidein through an ssh gateway to opportunistic resources

Workload Management



Multicore scheduling

CMS will be able to exploit multi-core resources already at the end of 2013

We expect that resource allocation will be "multi-core" early in 2014

The glidein will take care of scheduling single and multi-core jobs on multi-core resources

For efficiency reasons the resource allocation should have a longer duration, to be agreed with sites

Support by sites/middleware is needed to let the system know the characteristics of the allocated resource (including the remaining allocation time)

CMS distributed computing

Resources coming from:

- Tier-0, 7 Tier1s, 49 Tier2s

… but also:

- Tier3s, HLT farm, Clouds, opportunistic resources

The HLT farm corresponds to 40% of the total Tier1 capacity and is available when not taking data

Part of the prompt reconstruction will be done at Tier1s

Disk-tape separation at Tier-1s (ongoing) allows to use them also for user analysis

Opportunistic resources (independently of their access interface) and (Public) Clouds can be used for non IO-intensive tasks, e.g. MC production

➔Towards a flat structure←

Summary

To cope with the increased computing needs for Run2 CMS needs to revisit the Computing Model

Flexibility will be added in order to be able to exploit heterogeneous resources

- New architectures
- New resource allocation interfaces
- Interchangeability of sites

Increased use of automation and of data caches

None of the changes represents a revolution with respect to 2012 since many changes were already applied during Run1

Significant R&D and adaptation of the infrastructure is needed in order to increase efficiency