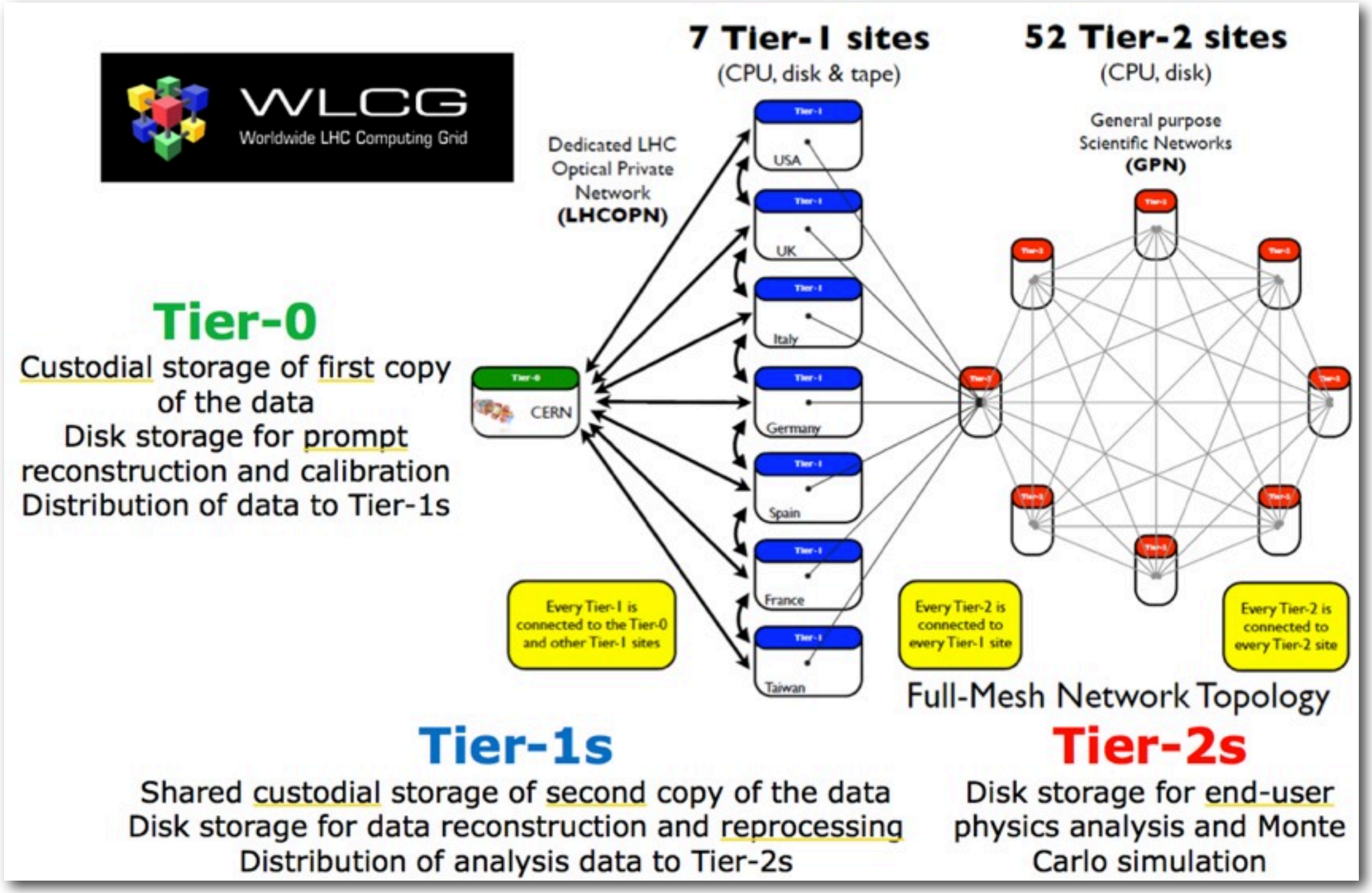
Intelligent data placement models for the CMS experiment

D. Giordano

(CERN IT-SDC)

CERN Openlab Data Analytics Workshop
20 Nov 2013

CMS Computing Infrastructure



[CHEP2013 - The CMS Data Management System]

Motivations (I)

The CMS Experiment makes a considerable usage of the distributed grid resources for the storage and offline analysis of the collected data

- Several hundreds of users submit daily up to 500,000 jobs accessing distributed data
- Data samples for user analysis replicated in several copies and distributed among the 52 CMS Tier-2 sites
  - ~23 PB of data are resident @ Tier-2 sites, **~4 PB** added in the last year
  - **~ 18 PB** transferred among sites in the last year (**~4.5** x resident volume)

Drawback

- the current data management model is manpower intensive and results in inefficient usage of disk space

Motivations (II)

Projections for LHC Run 2 imply a factor of 6 increase in needed computing resources

- Mandatory to optimize the usage of current resources
  - Minimize the job time in accessing/analyzing data
  - Minimize the number of data replicas at sites

Strategy

Optimize disk usage on the Tier-1 and Tier-2 level

- Open Tier-1 resources for analysis
- Simplify the current data management model
  - Tier-2 sites act as centrally managed caches for copies
- Provide more space for user activities

The designed strategy involves

- automatic cache release (i.e. obsolete replica deletion)
- dynamic data placement

Dynamic Data Placement Baselines

Initial pre-placement of new samples

- At least at one Tier-1 on disk, and at one or two different Tier-2 sites

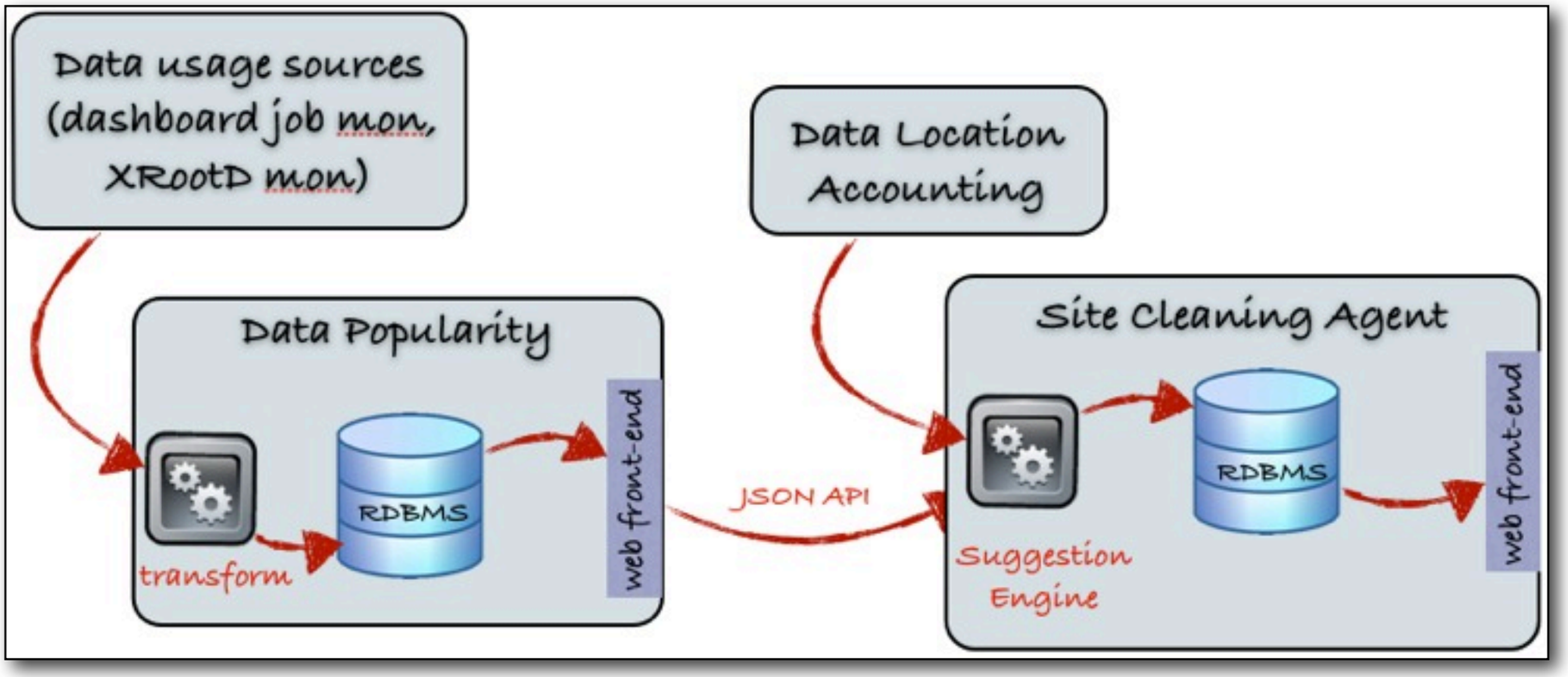
Dynamism

- Automatic replica creation to more sites based on sample usage patterns (a.k.a. data popularity)
  - Sites are selected taking into account current usage of disk and CPU and site readiness
  - Evaluate if a sample is better accessed remotely over the WAN rather than replicated
- Automatic sample cleaner to keep sites from overflowing

Data Popularity Project

CMS has already in production a Data Popularity Service

- monitoring along time the patterns of usage of accessed data samples
- Extended with Site Cleaning agent
  - automatic identification of obsolete replicas per site



Data Analysis inside Oracle DB

Oracle DB is used to

- Store, aggregate, analyze, correlate the monitoring data
- Metrics adopted
  - number of accesses, processing time, number of users/sites
  - Trace time evolution of the popularity metrics
- Multiple aggregations
  - Sites, job success/failure, set of files, analysis activities, time windows

Materialized views are adopted to filter and aggregate incoming data

Data Popularity Service in numbers

Data popularity workflow is steadily collecting data since Jun 2011

Amount of data uploaded:

- ~130k jobs/day , ~500k rows/day ⇔ ~ 400 M records stored

- Size of the Table
  - Raw-Table: ~250 GB, Materialized Views: ~6 GB in ~30 MV tables
- Speed of the procedure
  - Raw-Table upload: ~1 h/day , MV update: ~1.5h/day

Work in progress to include a second monitoring source

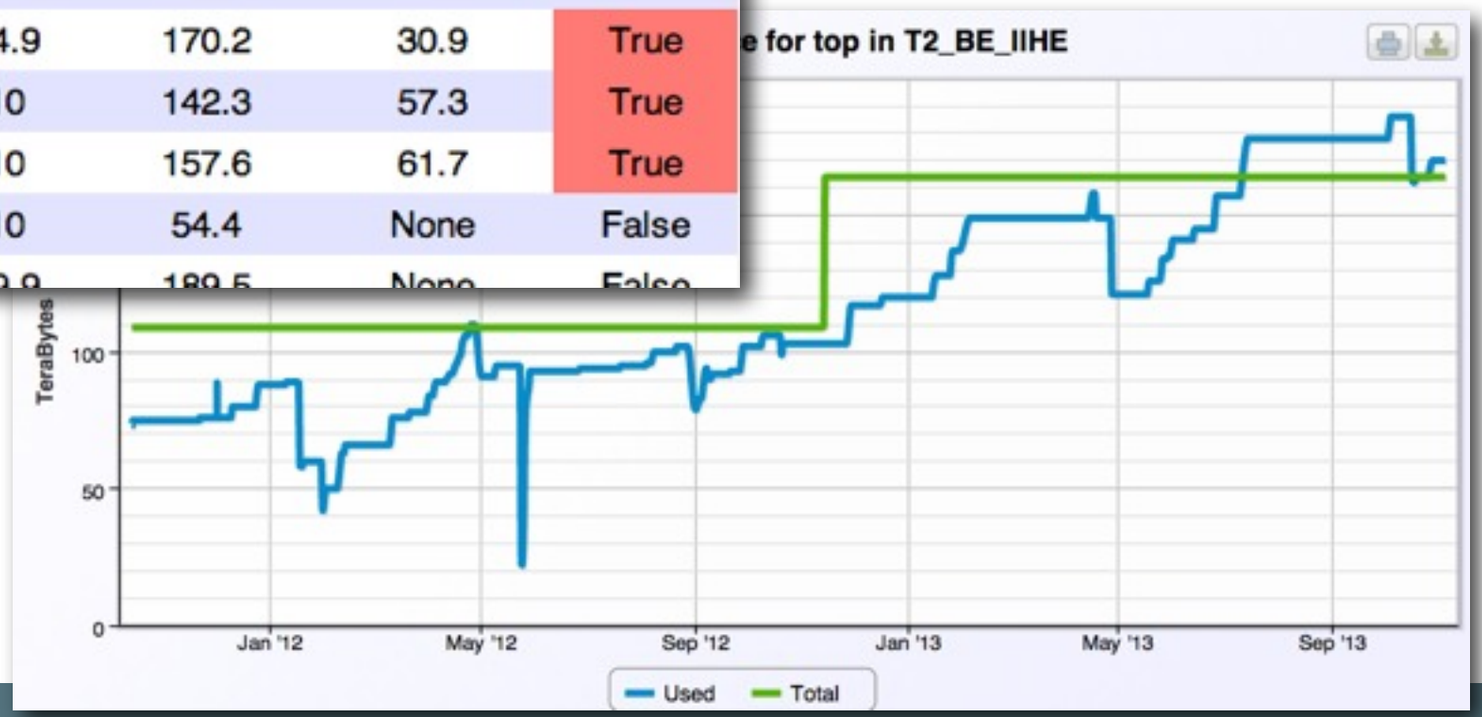
- XRootD data access monitoring
- Will double the rate of data to collect and analyze

Site Cleaning Agent

"Value" extracted from popularity data

- Scans Tier2 sites reaching their space quota and suggests obsolete samples that can be safely deleted

Cloud	Site	Tier	Group	Total [TB]	Used [TB]	Cleaned [TB]	Full
AT	Vienna	T2	AnalysisOps	55	81.6	39.6	True
AT	Vienna	T2	b-tagging	55	55.2	17.7	True
AT	Vienna	T2	susy	55	51.2	12.9	True
BE	IIHE	T2	AnalysisOps	274.9	197	None	False
BE	IIHE	T2	jets-met_hcal	55	45.8	None	False
BE	IIHE	T2	top	164.9	170.2	30.9	True
BE	UCL	T2	AnalysisOps	110	142.3	57.3	True
BE	UCL	T2	exotica	110	157.6	61.7	True
BE	UCL	T2	tracker-dpg	110	54.4	None	False



More to come

Need to extract further knowledge from the monitoring data in order to implement an effective data placement

- Correlate file-access monitoring with site status
  - Readiness, queue length, storage and CPU available
- Classify analysis activities and needed resources
- Making recommendations
- Learn from the past trends and patterns

Possibly adopt data mining tools/techniques common with other projects

- R, Oracle R Enterprise primarily, but also CEP, Hadoop, ElasticSearch, ...

Timescale: Have a new data placement system fully functioning by the begin of LHC run 2 in 2015