



Grid Computing in CMS: Remote Analysis & MC Production

Andre Sznajder
UERJ



Outline

- CMS Detector
- CMS Computing Model
- Data Storage, Transfer & Placement
- Grid Data analysis
- Grid Monte Carlo production

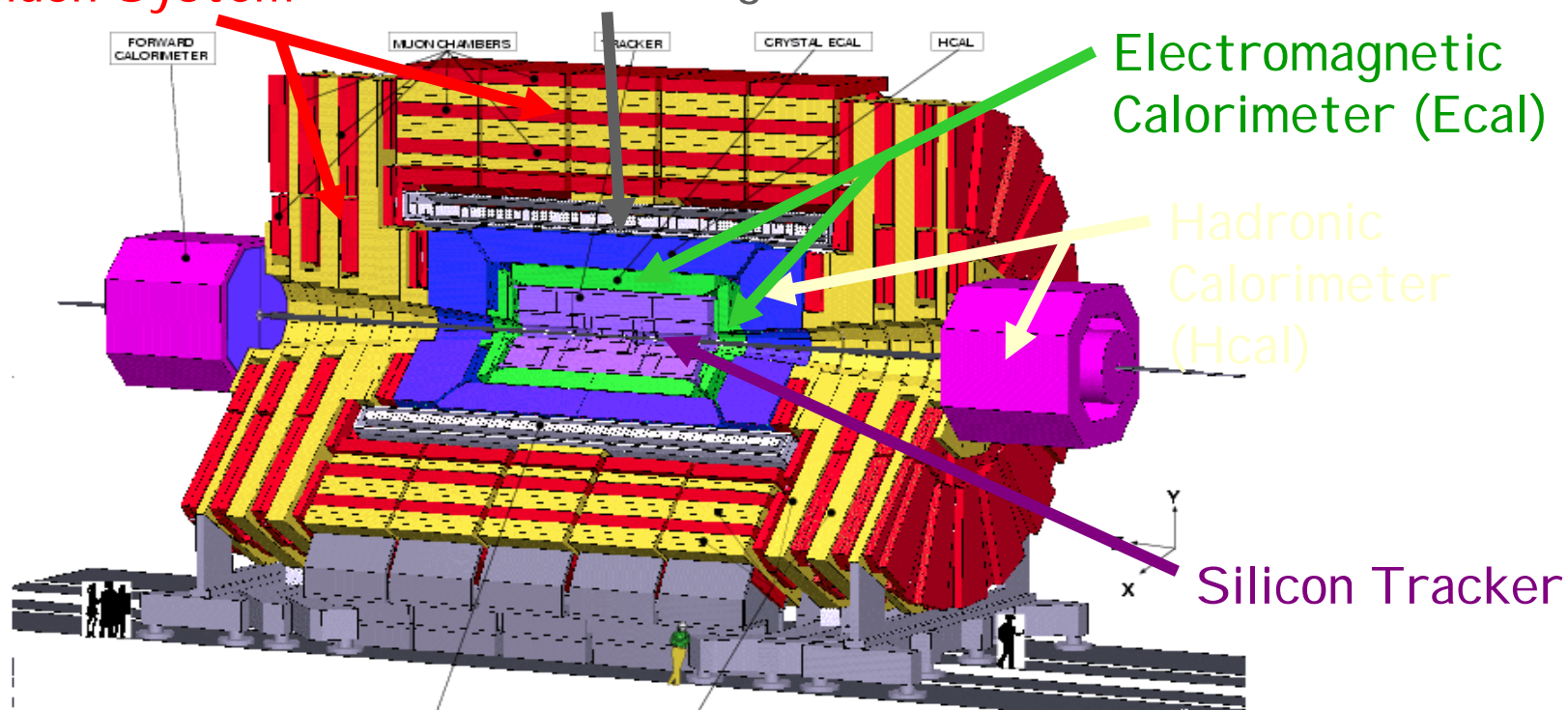


The CMS Detector

SVS modular structure follows the detector sub-systems

Muon System

Solenoid Magnet: 4 Tesla Field



22 m long & 15 m in diameter

More than 1 Million Geometrical volumes



Data Tiers & Volumes(2008)

- RAW
 - Detector data + L1, HLT results after online formatting
 - Includes factors for poor understanding of detector, compression, etc
 - **1.5MB/evt** @ ~150 Hz; ~ 4.5 PB/year (two copies)
 - 1 copy at Tier-0 and one spread over Tier-1's
- RECO
 - Reconstructed objects with their associated hits
 - **250kB/evt**; ~2.1 PB/year (incl. 3 reprocessing versions)
 - 1 copy spread over Tier-1 centers (together with associated raw)
- AOD(Analysis Object Data)
 - The main analysis format; objects + minimal hit info
 - **50kB/evt**; ~2.6PB/year - whole copy at each Tier-1
 - Large fraction at Tier-2 centers
- Monte Carlo in ~ 1:1 ratio with Real Data



CMS Computing Model

- CMS has adopted a distributed computing model which makes use of Grid technologies
- CMS Grid production services are in place:
 - Data transfer and placement system
 - Monte Carlo production
 - Remote Data Analysis
- Steadily increase in scale and complexity
- Basic Grid Infrastructure and Services in place but reliability and stability are the problems



CMS Computing Model

- CMS uses the hierarchy of computing TIERS
- The ensemble of computing GRID resources available to CMS forms the WLCG which is based on different middleware implementations: LCG, OSG, NorduGRID ...
- Details of heterogeneous GRID environments should be invisible for CMS physicists



CMS Computing Model



The CMS offline computing system is arranged in four Tiers and is geographically distributed

Online system

Offline farm

recorded data

Tier 0

CERN Computer center

Tier 1

France Regional Center

Italy Regional Center

Fermilab Regional Center

Tier 2

Tier2 Center

Tier2 Center

Tier2 Center

Tier 3

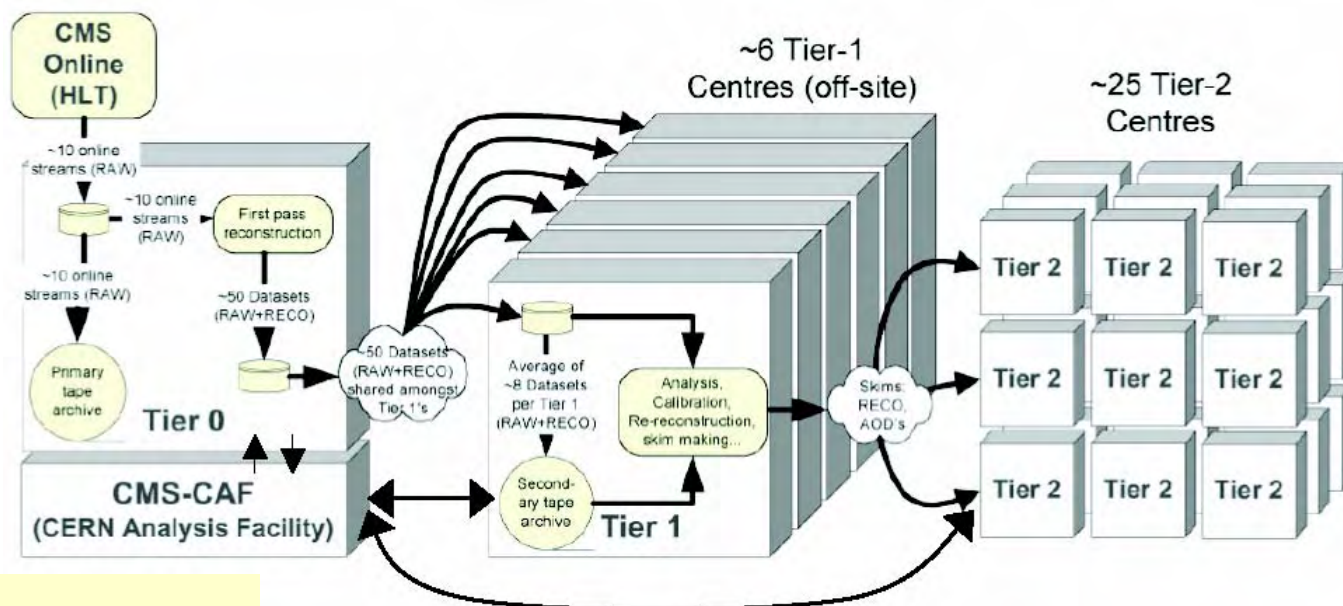
Institute A

Institute B

workstation



Tiered Architecture



Tier-0:

- Accepts data from DAQ
- Prompt reconstruction
- Archives data and distributes to Tier-1's

Tier-1's:

- Real data & MC archiving
- Re-processing, Calibration, skimming and intensive analysis tasks

Tier-2's:

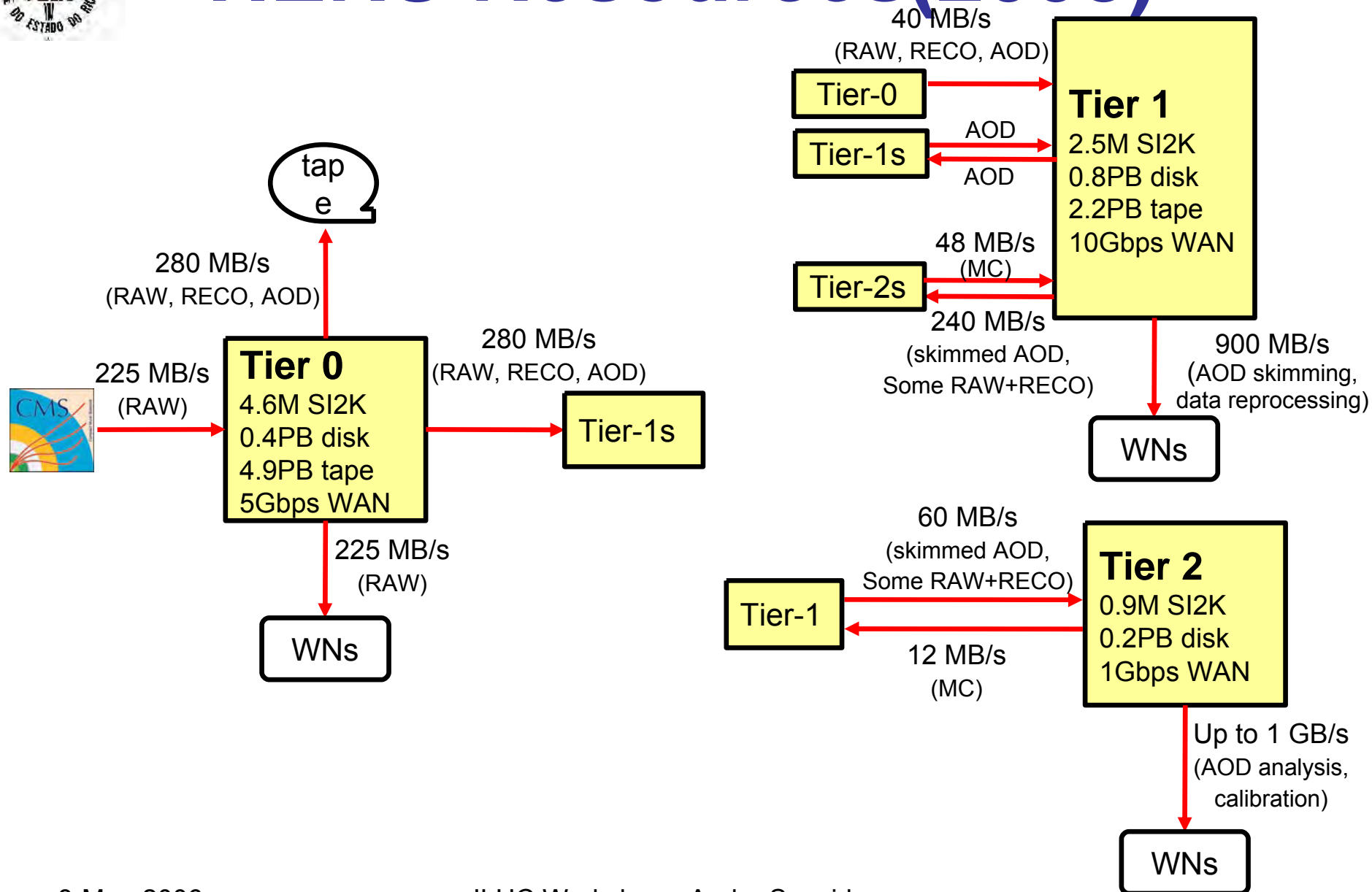
- Data Analysis
- MC simulation
- Import datasets from Tier-1 and export MC data

Tier-3's:

- Opportunistic computing resources
- Not accounted as CMS resources
- Data Analysis & MC simulation



TIERS Resources(2008)



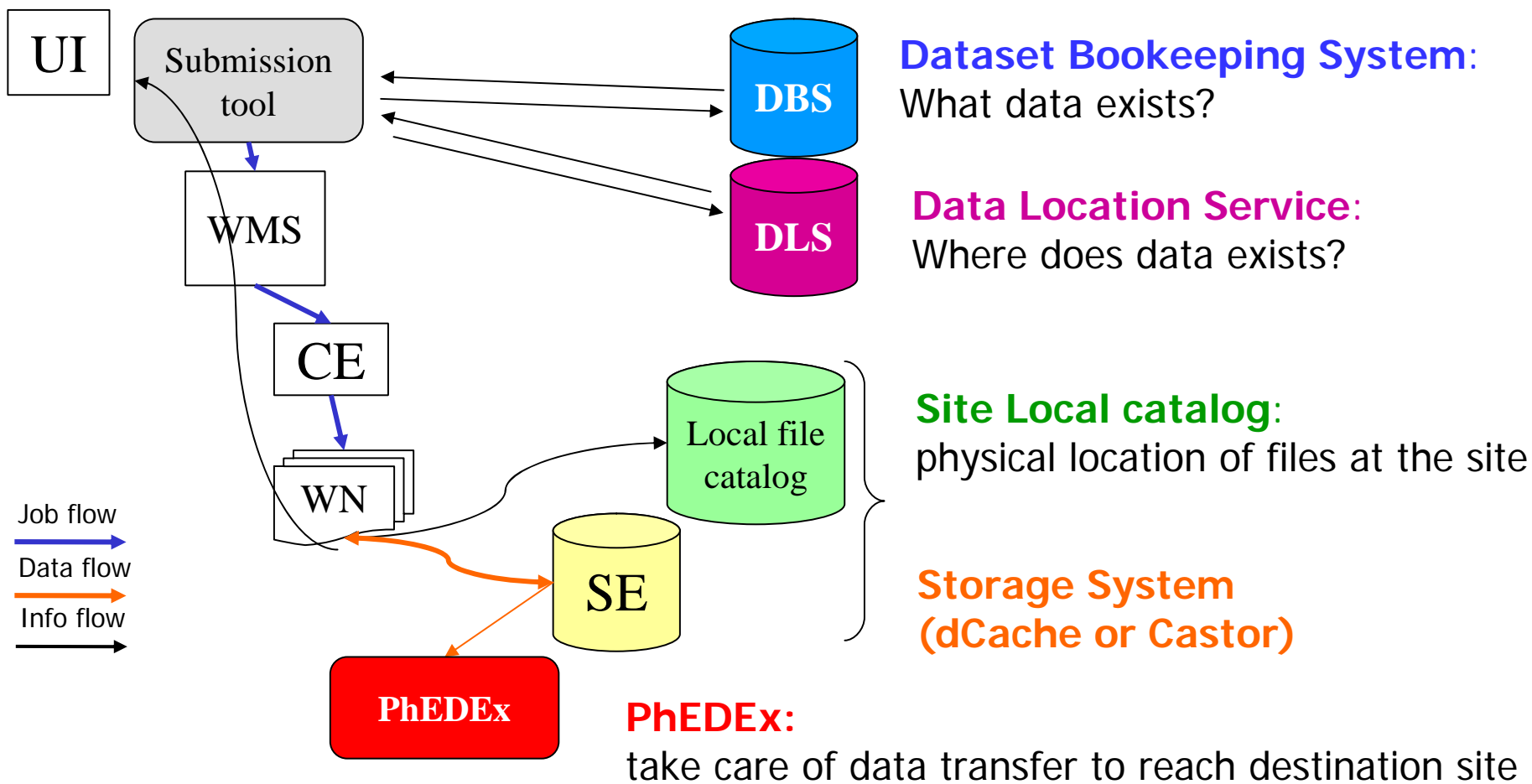
8-May-2006

ILHC Workshop - Andre Sznajder



Data Workflow

Data Management System allow to discover, access and transfer event data in a distributed computing environment





GRID Production Systems

- Data Transfer and Placement System
(**Physics Experiment Data Export – PHEDEX**)
 - In production since almost two years
 - Managing transfers of several TB/day
 - ~150 TB known to PhEDEX, ~350 TB total replicated
 - Running at CERN, 7 Tier-1's, 15 Tier-2's
- Distributed data analysis - **CMS Remote Analysis Builder (CRAB)**
 - Tool for job preparation, submission and monitoring
 - ~ 60K analysis jobs/month
- MC Production - **Monte Carlo Production System(MCPS/McRunjob)**
 - ~ 10M events/month (4x10K jobs), ~ 150M events in total
 - ~ 20% in OSG and 15% in LCG. Rest on local farm mode in big sites although mostly production on the Grid in the last months



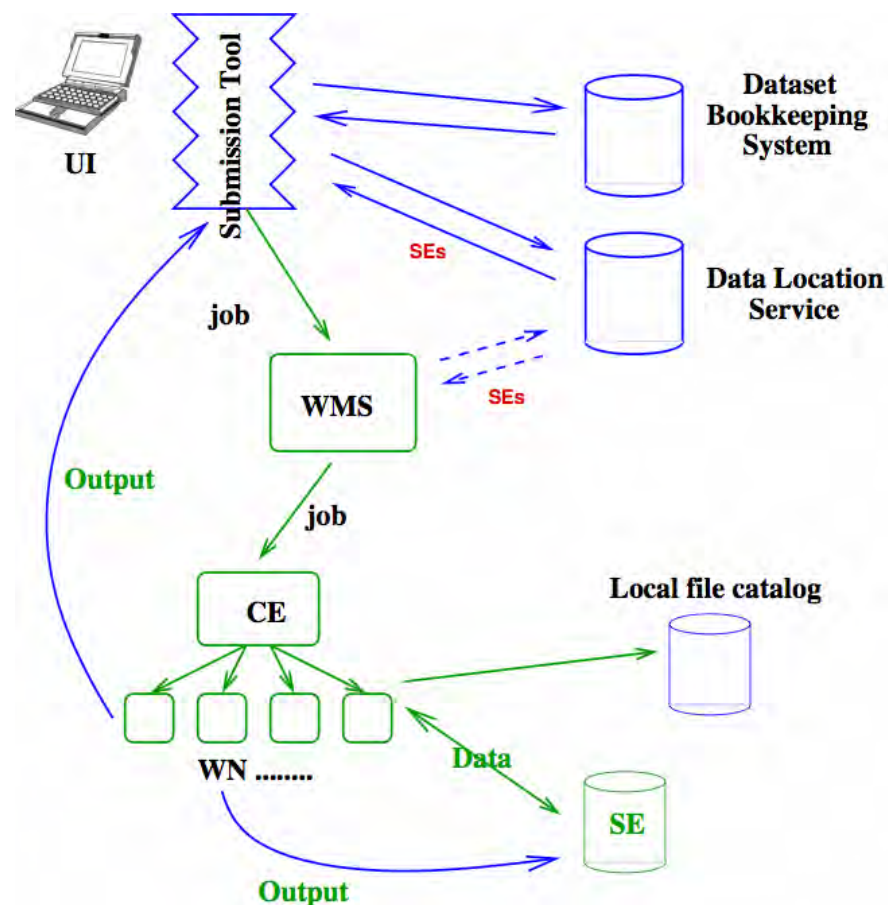
PHEDEX

- **PHEDEX** is responsible for data transfer and placement
- Large scale dataset replica management system
- Manages data flow following a transfer topology:
Tier0 → Tier1 ↔ Tier2
- Routing agents determine the best route
- Enables distribution management at dataset level rather than at file level
- Allows prioritization and scheduling
- In production since more than a year
 - Managing transfers of several TB/day (~200 TB total replicated)
 - Running at CERN, 7 Tier-1's, 10 Tier-2's



Data Analysis on the Grid

- CMS Remote Analysis Builder (**CRAB**) tool for job preparation, submission, execution and basic monitoring
- Data samples distributed in Tier-1 sites
- WLCG(Worldwide LHC Computing Grid) has two main different flavours:
 - LCG/gLite in Europe
 - OSG in the US





CRAB features

- Keeps easy creation of large number of user analysis job
 - Assume all jobs are the same except for some parameters (event number to be accessed, output file name...)
- Allows to access distributed data efficiently
 - Hiding middleware complications. All interactions are transparent for the end user
- Manages job submission, tracking, monitoring and output harvesting
 - User doesn't have to take care about how to interact with sometimes complicated grid commands



CRAB functionalities

- Data discovery
 - Data are distributed so we need to know where data have been sent
- Job creation
 - Both .sh (wrapper script for the real executable) and .jdl (a script which drives the real job towards the “grid”)
 - User parameters passed via config file (executable name, output file names, specific executable parameters...)
- Job submission
 - Scripts are ready to be sent to those sites which host data
 - Boss, the job submitter and tracking tool, takes care of submitting jobs to the Resource Broker



CMS Physics TDR

- CRAB was used to analyze MC data for the CMS Physics TDR (being written now...)



• Most accessed dataset since last July

D.Spiga: CRAB Usage and jobs-flow Monitoring (DDA-252)

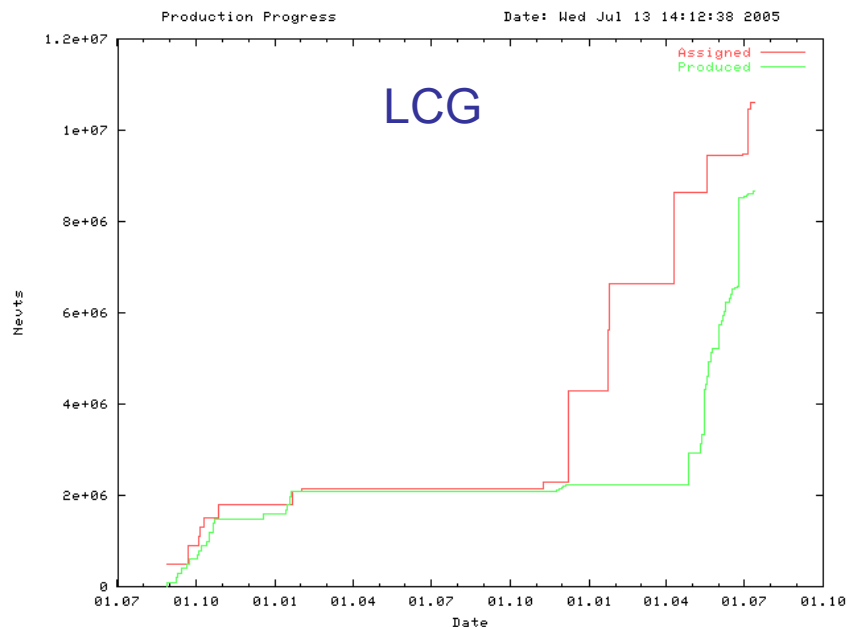
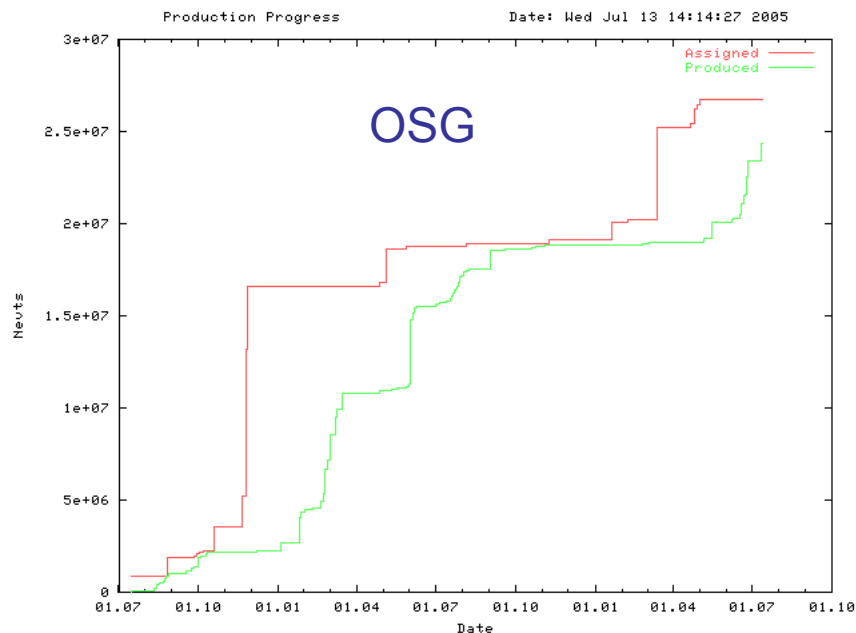


MC Production on the Grid

- Main CMS GRIDS:
 - US Open Science Grid (**OSG**)
 - LHC Computing Grid (**LCG**)
- Production in both Grids(LCG &OSG) since 2003
- **MCPS/MCRunjob** is a tool for running CMS production jobs (preparation, submission, stage-in, execution, stage-out, cleanup)
 - Developed by FNAL with contributions from other CMS people
 - Highly configurable and flexible
 - Interfaced to all Grids and local farm production
- Different production steps (generation, simulation, digitization and reconstruction) currently run separately



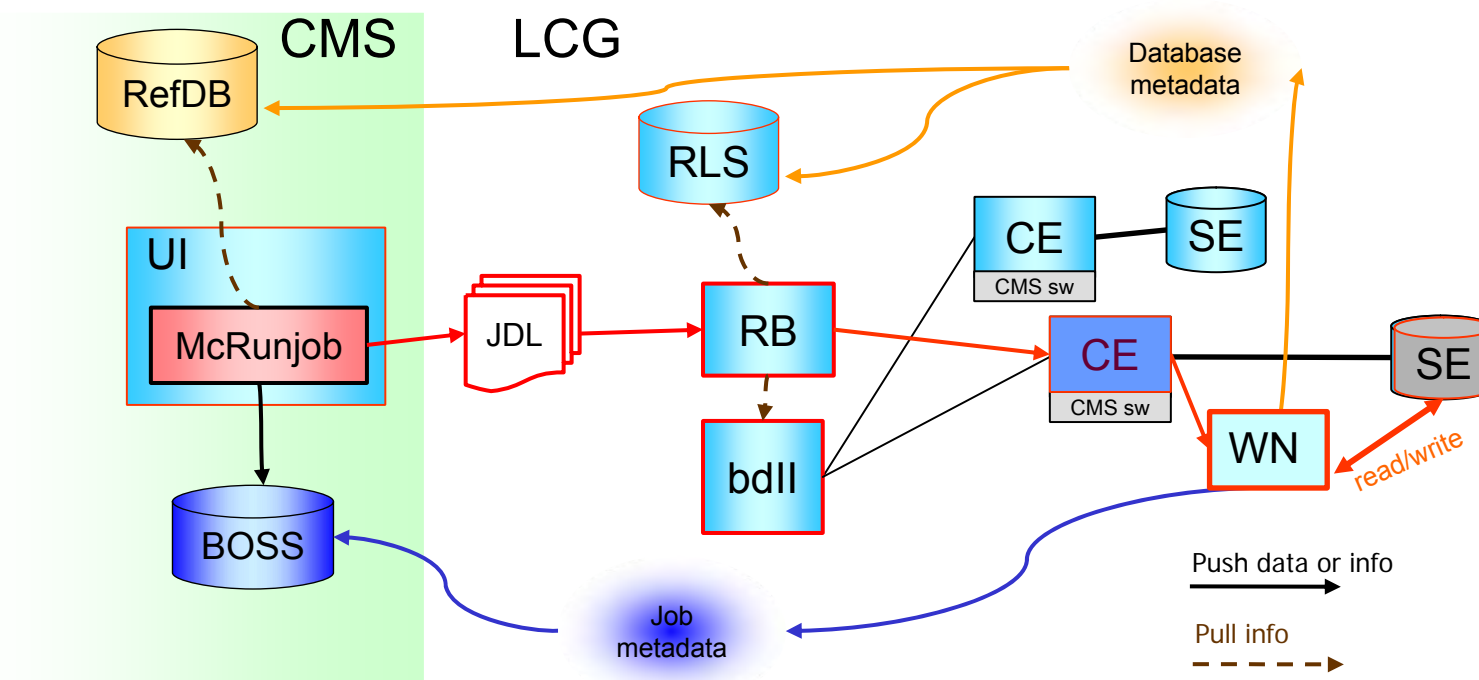
MC Production on the Grid



- Several thousand CPUs available in both Grids
- Few million events per month
- Around 70-90% job efficiency → **The issue is reliability**
 - System issues (hardware failures, NFS access, disk full, site misconfiguration)
 - Software installation problems
 - Grid services, stage-in and stage-out files, LCG catalogue instability



LCG Production Workflow



- Quasi-real-time job monitoring (BOSS)
- CMS software pre-installed



MCPS(OSG)

MCPS started as an idea for how to allow users to make small custom simulation samples.

- CMS official simulation infrastructure is optimized to deliver large numbers of events to many groups, while keeping careful track of provenance information
- MCPS was initially designed to deliver small custom samples: reproducible, reliable events, with a fast turn around.
- The CMS simulation workflow involves 4 steps that are run in serial. MCPS chained the various steps together into a workflow, looking after the relationships between the input and output file.

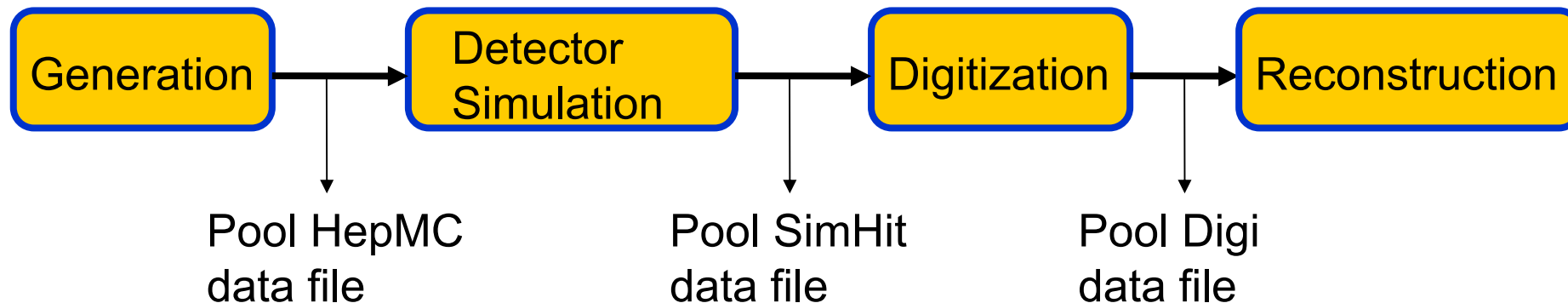


New MC Production System

- New MC production system being developed
- Overcome current inefficiencies, introduce new capabilities and integrate with new Event Data Model and DMS
 - McRunjob was originally designed for local farm and ported to the Grid last year
 - Hopefully less manpower consuming, better handling of grid/sites unreliability, better use of resources, automatic retries, better error report/handling,etc
 - Better coupling with data transfer system
 - Job chaining(generation→simulation→digitization→reconstruction) to overcome bottleneck due to digitization with pile-up (I/O dominated by chaining with simulation (CPU dominated)
 - Data merging, fileblock management, use DBS/DLS



Simulation Software in CMS



Generation – MC truth information from particle gun or physics generator about vertices and particles. Stored in HepMC format.

Detector Simulation – Hit objects with timing, position, energy loss information. Based on the Geant4 tool kit.

Digitization – Constructs Digi objects which include realistic modeling of electronic signal.

Reconstruction – Physics Objects: vertices, photons, e, mu, jets ...



MC Data Tiers Characteristics

- Generation:
 - no input, small output (10 to 50 MB ntuples)
 - pure CPU
- Simulation (hits): GEANT4
 - small input
 - CPU and memory intensive
 - large output: ~500 MB in three files (EVD files), the smallest is ~ 100 KB !
- Digitization:
 - lower CPU/memory requirements
 - I/O intensive: persistent reading through LAN
 - large output: similar to simulation
- Reconstruction:
 - even less CPU
 - smaller output: ~200 MB in two files



THE END

8-May-2006

ILHC Workshop - Andre Sznajder