CMS/ARDA

Lucia Silvestris

INFN-Bari

Outline

- CMS CPT Schedules (Where we are, what's next..)
- Data Challange DC04
- Next steps After DC04
- Short term plan for CMS Distributed analysis
- Discussion

CMS Computing schedule

2004

- Mar/Apr. DC04 to study T0 Reconstruction, Data Distribution, Real-time analysis 25% of startup scale
- May/Jul. Data available and useable by PRS groups
- Sep. PRS analysis feed-backs
- Sep. Draft CMS Computing Model in CHEP papers
- Nov. ARDA prototypes
- Nov. Milestone on Interoperability
- Dec. Computing TDR in initial draft form.

2005

- July. LCG TDR and CMS Computing TDR
- Post July?... DC05 , 50% of startup scale.
- Dec. Physics TDR

2006

- DC06 Final readiness tests
- Fall. Computing Systems in place for LHC startup
- Continuous testing and preparations for data

Data Challenge 04

Aim of DC04:

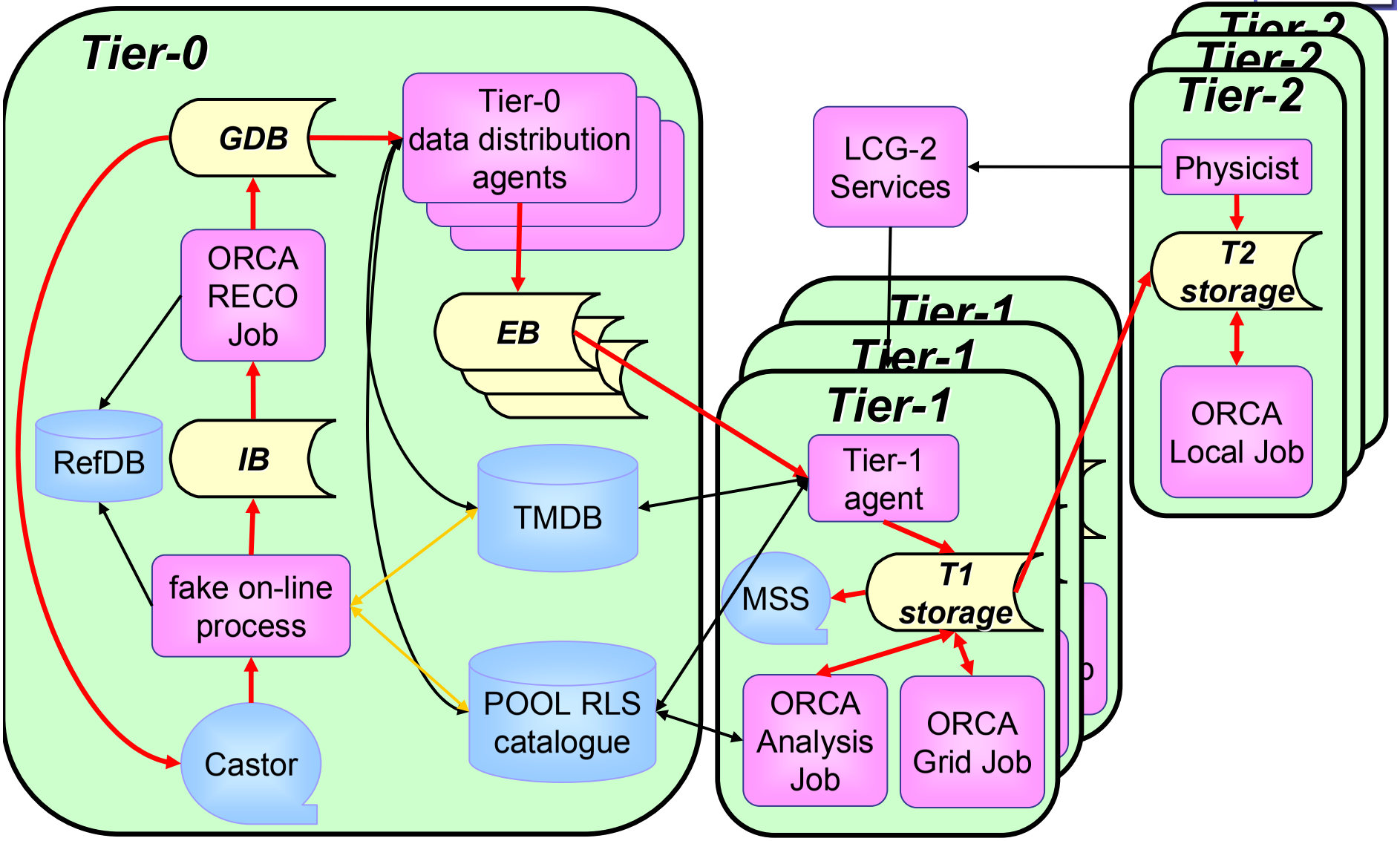
- reach a sustained 25Hz reconstruction rate in the Tier-0 farm (25% of the target conditions for LHC startup)
- register data and metadata to a catalogue
- transfer the reconstructed data to all Tier-1 centers
- analyze the reconstructed data at the Tier-1's as they arrive
- publicize to the community the data produced at Tier-1's
- monitor and archive of performance criteria of the ensemble of activities for debugging and post-mortem analysis

Not a CPU challenge, but a full chain demonstration!

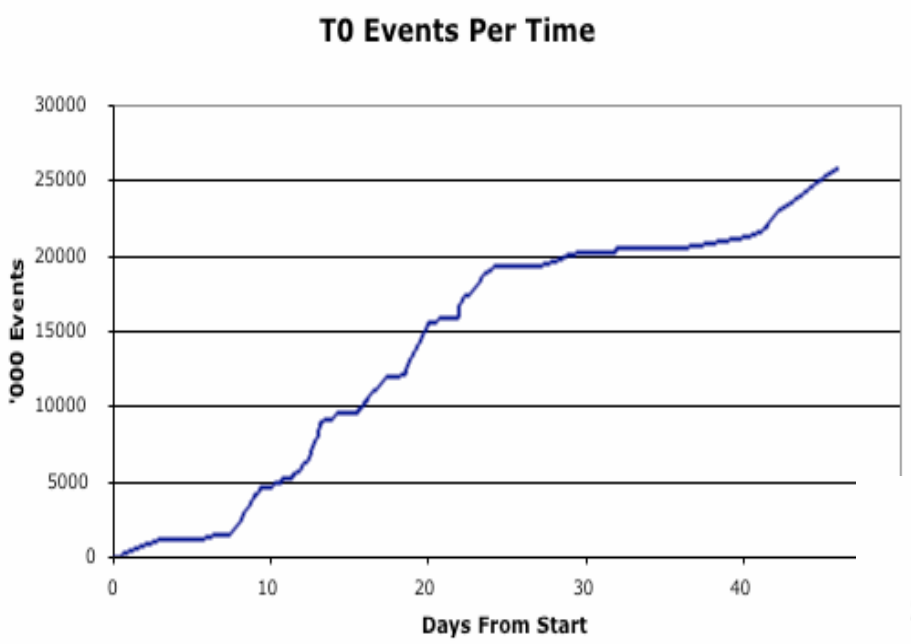
Pre-challenge production in 2003/04

- 70M Monte Carlo events (30M with Geant-4) produced
- Classic and grid (CMS/LCG-0, LCG-1, Grid3) productions

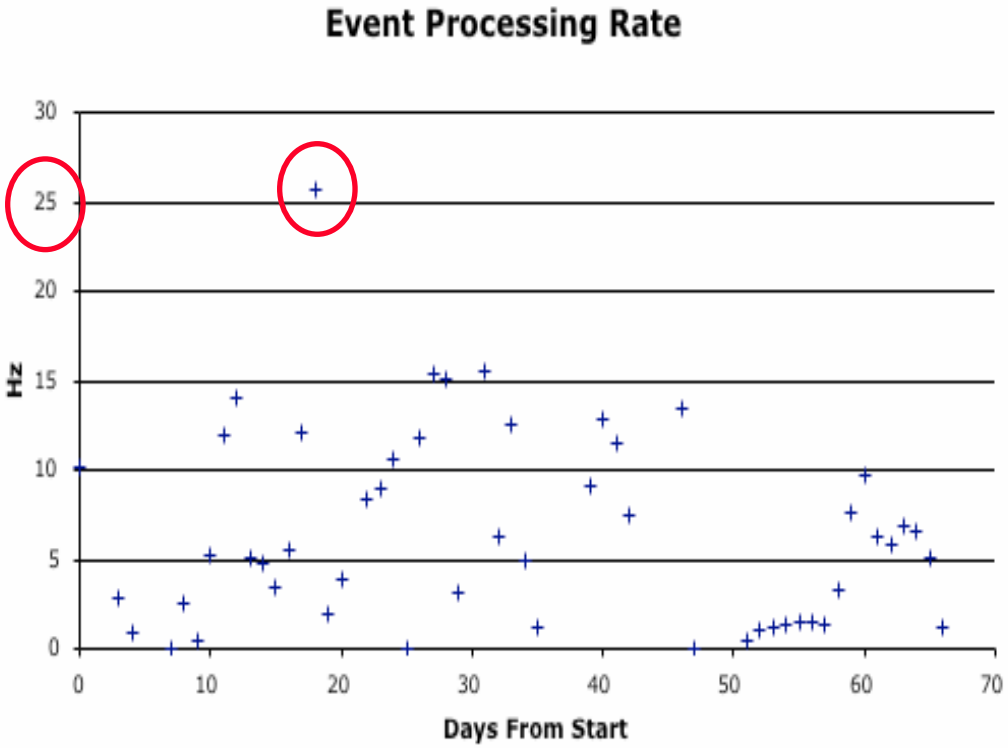
Data Challenge 04: layout



Data Challenge 04 Processing Rate



- Processed about 30M events
  - But DST content not properly tested make this pass not useful for real (PRS) analysis
- Generally kept up at T1's in CNAF, FNAL, PIC



- Got above 25Hz on many short occasions
  - But only one full day above 25Hz with full system
- Working now to document the many different problems

LCG-2 in DC04

Aspects of DC04 involving LCG-2 components

- register all data and metadata to a world-readable catalogue
  - RLS
- transfer the reconstructed data from Tier-0 to Tier-1 centers
  - Data transfer between LCG-2 Storage Elements
- analyze the reconstructed data at the Tier-1's as data arrive
  - Real-Time Analysis with Resource Broker on LCG-2 sites
- publicize to the community the data produced at Tier-1's
  - Not done, but straightforward using the usual Replica Manager tools
- end-user analysis at the Tier-2's (not really a DC04 milestone)
  - first attempts
- monitor and archive resource and process information
  - GridICE

- Full chain (except Tier-0 reconstruction) could be performed in LCG-2

Real-Time (Fake) Analysis

- Goals
  - Demonstrate data can be analyzed in real time at the T1
    - Fast Feedback to reconstruction (e.g. calibration, alignment, check of reconstruction code, etc.)
  - Establish automatic data replication to T2s
    - Make data available for offline analysis
  - Measure time elapsed between reconstruction at T0 and analysis at T1
- Architecture
  - Set of software agents communicating via local mysql DB
    - Replication, data set completeness, job preparation & submission
  - Use LCG to run jobs
    - Private Grid Information System for CMS DC04
    - Private Resource Broker

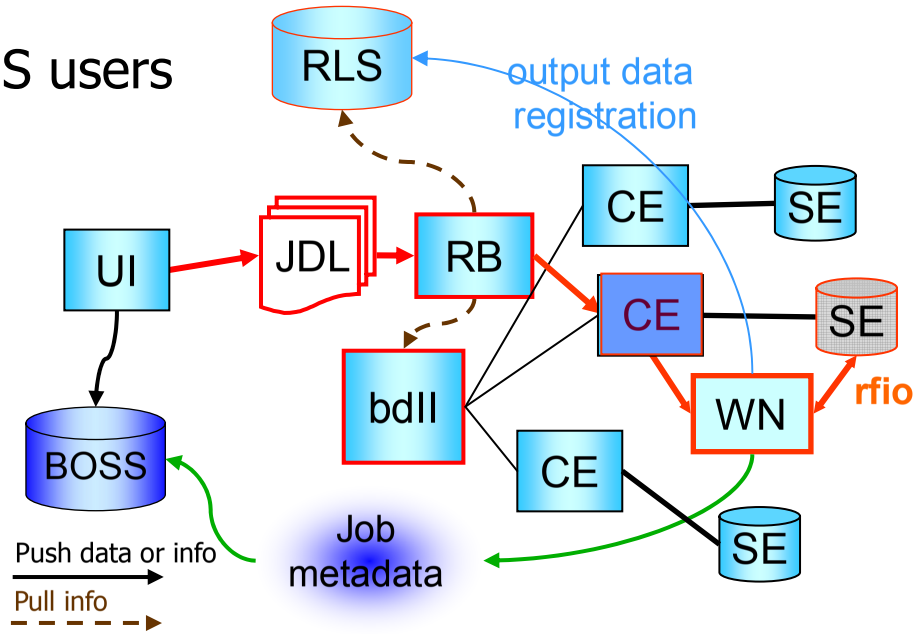
DC04 Real-Time (fake) Analysis

CMS software installation

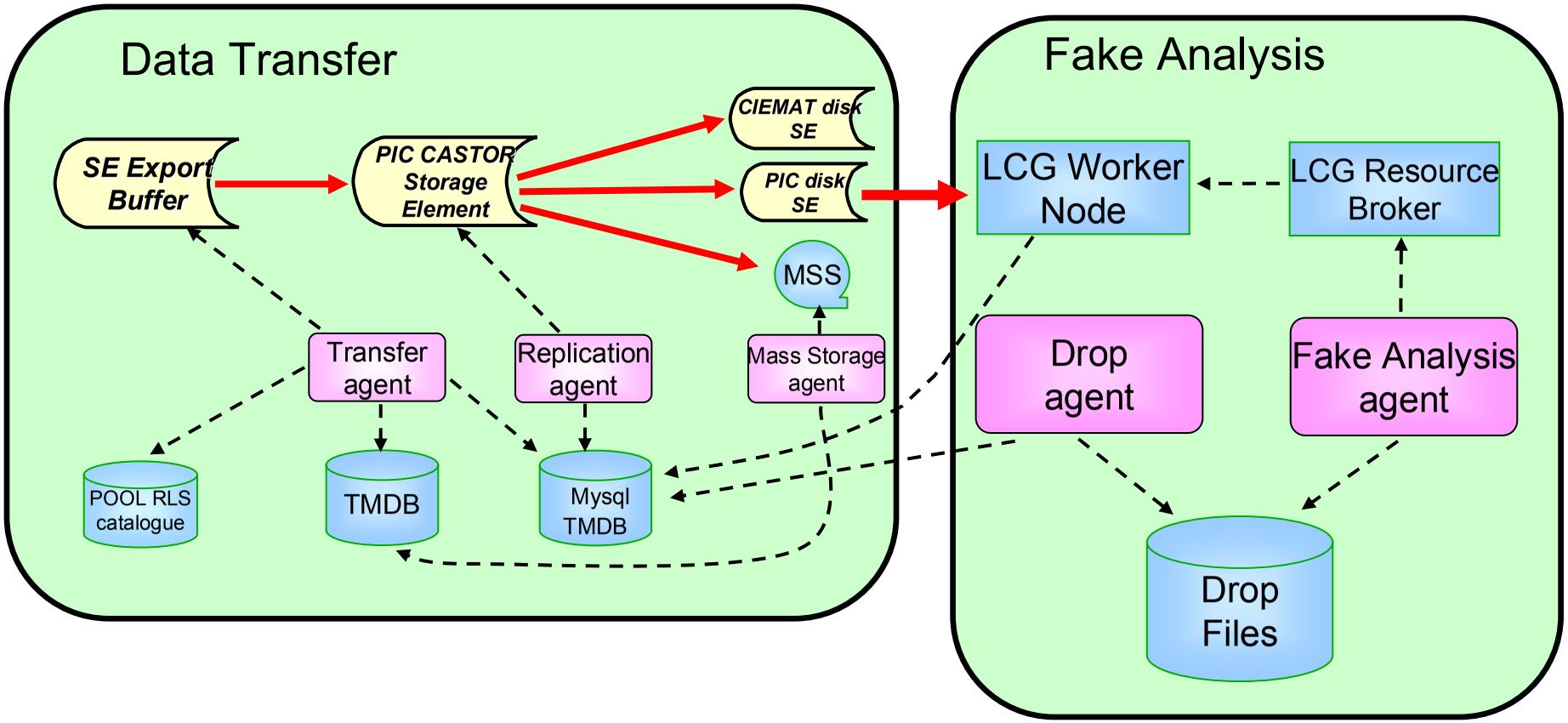
- CMS Software Manager (M. Corvo) installs software via a grid job provided by LCG
  - RPM distribution based on CMSI or DAR distribution
  - Used at CNAF, PIC, Legnaro, Ciemat and Taiwan with RPMs
- Site manager installs RPM's via LCFGng
  - Used at Imperial College
- Still inadequate for general CMS users

Real-time analysis at Tier-1

- Main difficulty is to identify complete file sets (i.e. runs)
  - Information today in TMDB or via findColls
- Job processes single runs at the site close to the data files
  - File access via rfio
- Output data registered in RLS



DC04 Fake Analysis Architecture



- Drop agent triggers job preparation/submission when all files are available
- Fake Analysis agent prepares xml catalog, orcarc, jdl script and submits job
- Jobs record start/end timestamps in mysql DB

Fake Analysis Strategy

- Transfer agent registers in mysql DB files at T1
- Replication agent replicates them to disk at T1/T2s
- Drop agent checks completeness of job data set at T1/T2s
  - Drop file contains GUIDs and PFNs of event data files and ZippedMETA file
- Fake Analysis agent prepares and submits jobs to LCG
  - Get DATASET, OWNER, RUNID, JOBID, GUIDs, PFNs from drop file
  - Prepare local xml catalog, orcarc file, envariables file, jdl script
  - Scripts from INFN fake analysis team adapted to PIC
- Job submission & monitoring
  - BOSS not used due to lack of time, big executables in InputSandbox
  - Instead, job submission, start, end timestamps recorded in mysql DB
- Job running on any data set
- Fake Analysis executable uploaded into SE

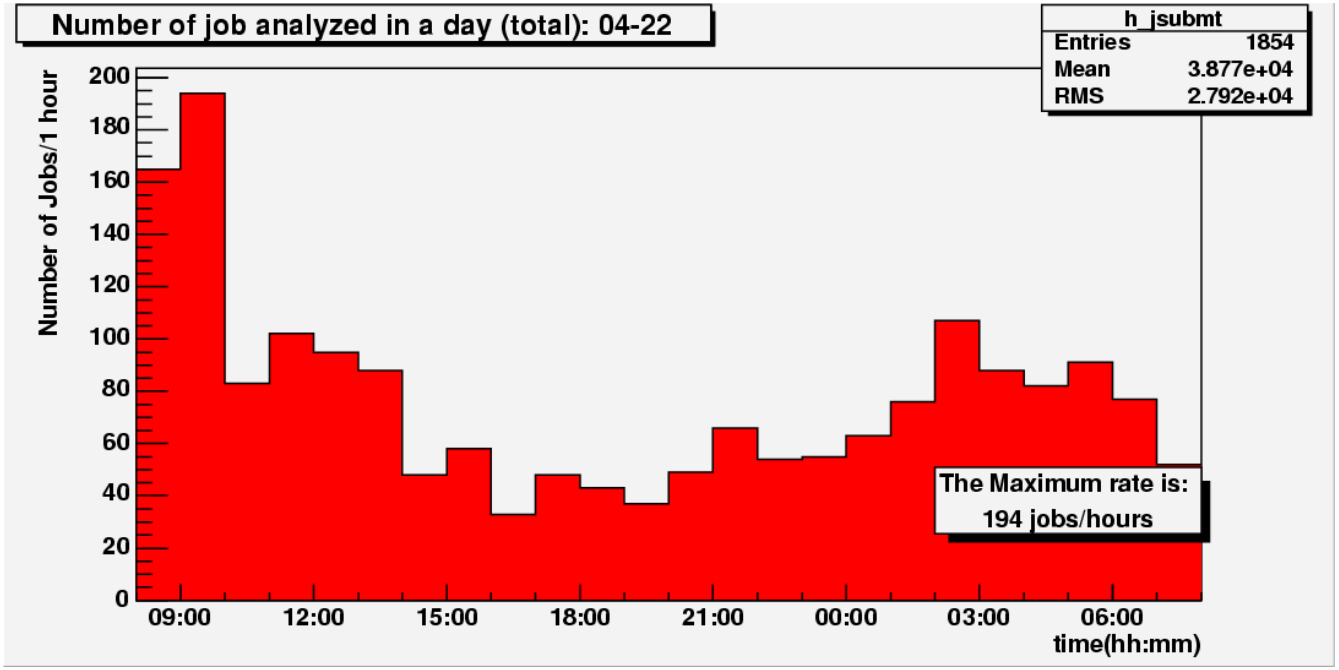
DPS June 04

Slide

DC04 Real-time Analysis

INFN

- Maximum rate of analysis jobs: **194 jobs/hour**
- Maximum rate of analysed events: **26 Hz**
- Total of **~15000** analysis jobs via **Grid** tools in ~2 weeks **(95-99% efficiency)**

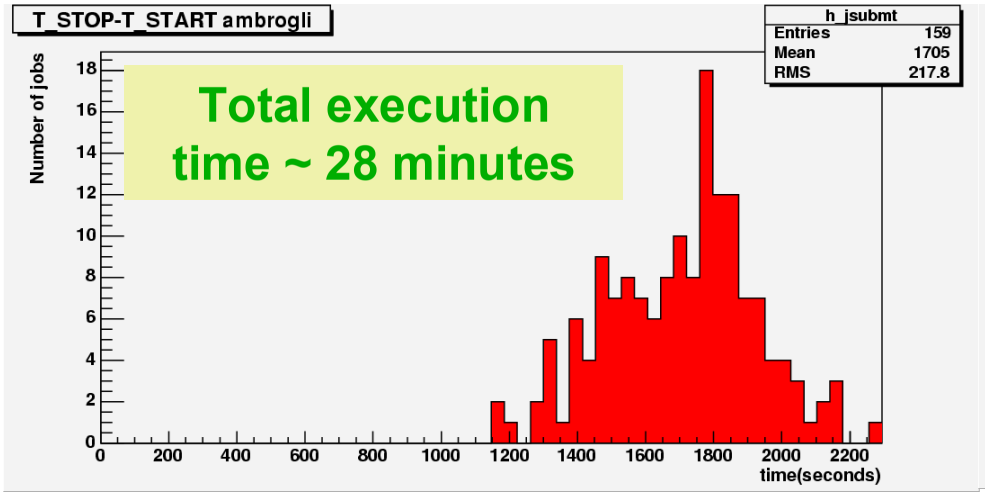


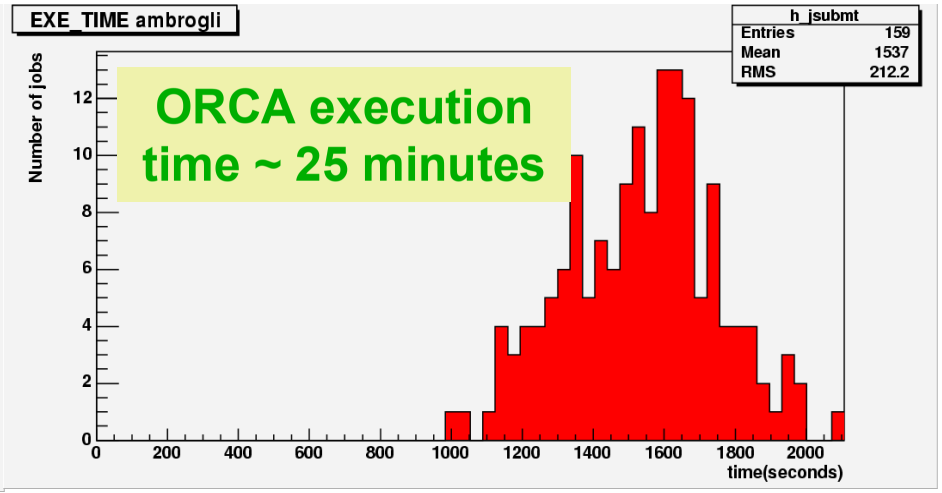
➢Datasets examples:

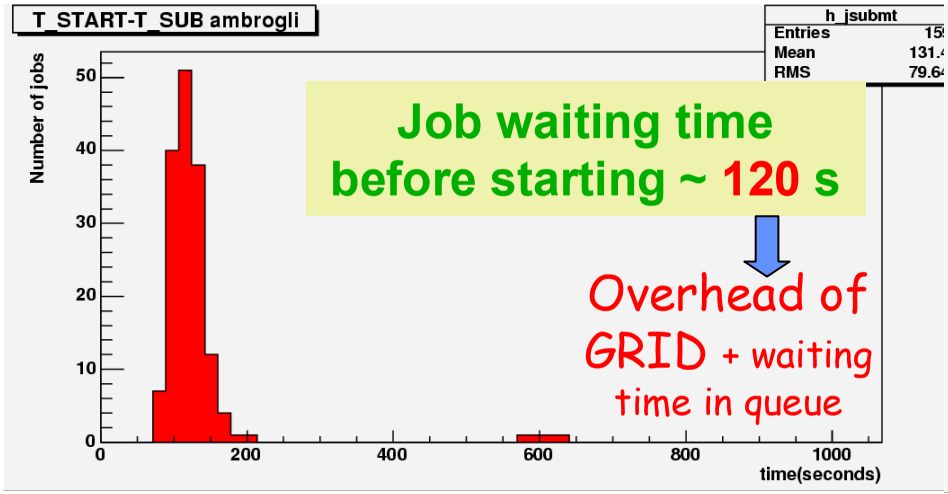
- $B^0_S \rightarrow J/\psi\, \varphi$
  Bkg: mu03_tt2mu, mu03_DY2mu
- **tTH, H → bbbar t→ Wb W → lν T → Wb W → had.**
  Bkg: bt03_ttbb_tth
  Bkg: bt03_qcd170_tth
  Bkg: mu03_W1mu
- **H → WW → 2μ 2ν**
  Bkg: mu03_tt2mu, mu03_DY2mu

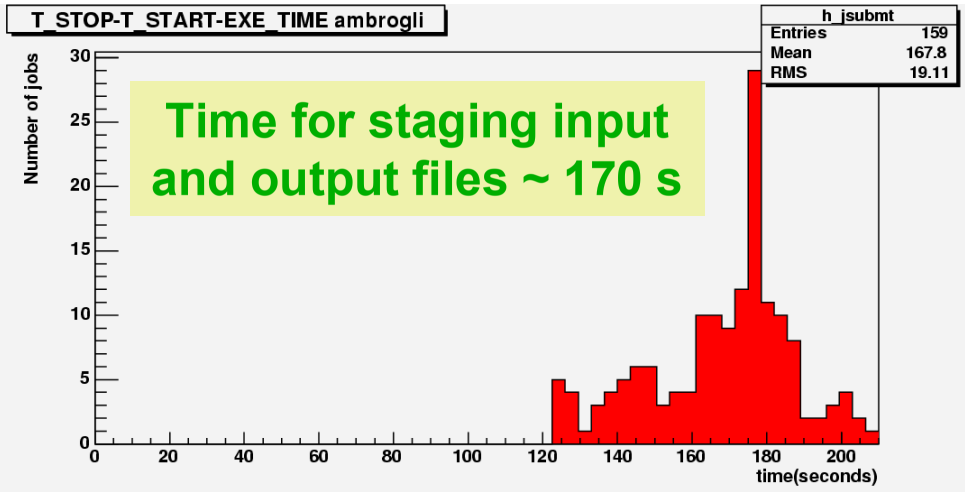
Results: job time statistic

Dataset **bt03_ttbb_ttH** analysed with executable **ttHWmu**









Real time Analysis Summary

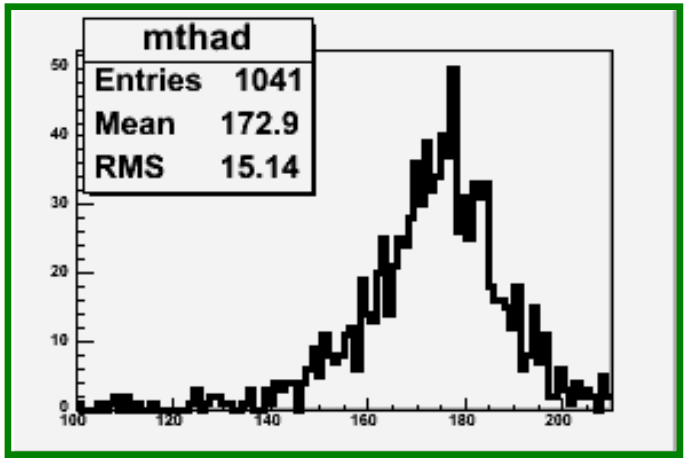
- Real-time analysis: two weeks of quasi-continuous running!
- The total number of analysis jobs submitted ~ **15000**
- Overall Grid efficiency ~ 95-99%

- Problems :
  - RLS query to prepare a POOL xml catalog done using file GUID otherwise much slower
  - Resource Broker disk being full causing the RB unavailability for several hours. This problem was related to large input/output sandbox. Possible solutions:
    - Set quotas on RB space for sandbox
    - Configure to use RB in cascade
  - Network problem at CERN, not allowing connections to the RLS and CERN RB
  - Legnaro CE/SE disappeared in the Information System during one night
  - Failures in updating Boss database due to overload of MySQL server (~30%). The Boss recovery procedure was used

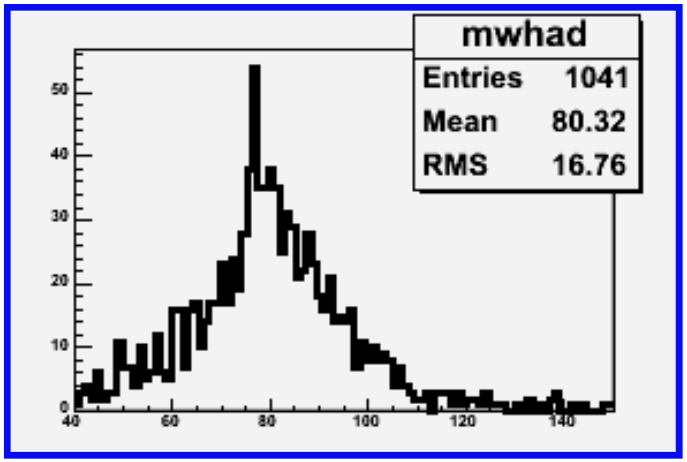
tTH analysis results

Reconstructed Masses:

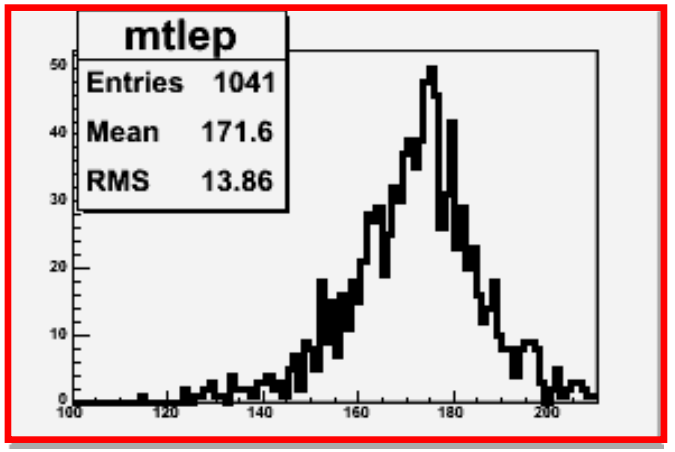
Hadronic Top



Hadronic W



Leptonic Top



Analyzing DC04. Planning Next Steps

- Aim for a rapid series of writeups
  - Particularly for key items such as RLS
- CMS Core-Software worked pretty well
  - G4, Reconstruction, DST, Streaming, POOL... All worked well
  - Usability of CMS Dataset and Collection information needs improvement
  - Ad-Hoc CMS responses to problems worked well. TMDB and Agents!
  - GRID Middleware came very late and mostly with significantly reduced functionality compare to even recent plans
- Do not assume giant leaps can be made
  - Plan for what is available "now" or can be reasonably expected soon
  - Cut the suit to fit the cloth
    - Adapt requirements to reality
    - (Design a small suit of you don't want to go around naked)
- Assume data handling and resource brokering are "orthogonal"
  - Ensure we have the tools to manage the experiment data flows and to implement experimer data flow policy
    - Ensure this will work in "any" GRID scenario and or disaster scenario
    - Leverage work such as SAM/SAMGRID
  - Work with LCG/EGEE/... to test and/or build user tools and applications for remote submission, code installation etc.
  - Both these directions need Experiment effort layered over Common Middleware (old and ne
- Prepare plausible Computing Model scenarios for CHEP

DPS June 04

Possible evolution of CCS tasks (Core Computing and Software)

CCS will Reorganize to match the new requirements and the move from R&D to Implementation for *Physics*

- Meet the PRS Production Requirements (Physics TDR Analysis)
- Build the Data Management and Distributed Analysis infrastructures

❖Production Operations group [NEW]

- Outside of CERN. Must find ways to reduce manpower requirements.
- Using predominantly (only?) GRID resources.

❖Data Management Task [NEW]

- Project to respond to DM RTAG
  - Physicists/ Computing to define CMS Blueprint, relationships with suppliers (LCG/EGEE…), CMS DM task in Computing group
- Expect to make major use of manpower and experience from CDF/D0 Run II

❖Workload Management Task [NEW]

- Make the Grid useable to CMS users
- Make major use of manpower with EDG/LCG/EGEE experience

❖Distributed Analysis Cross Project (DAPROM) [NEW]

- Coordinate and harmonize analysis activities between CCS and PRS
- Work closely with Data and Workload Management tasks and LCG/EGEE

After DC04: CMS-ARDA/DAProm

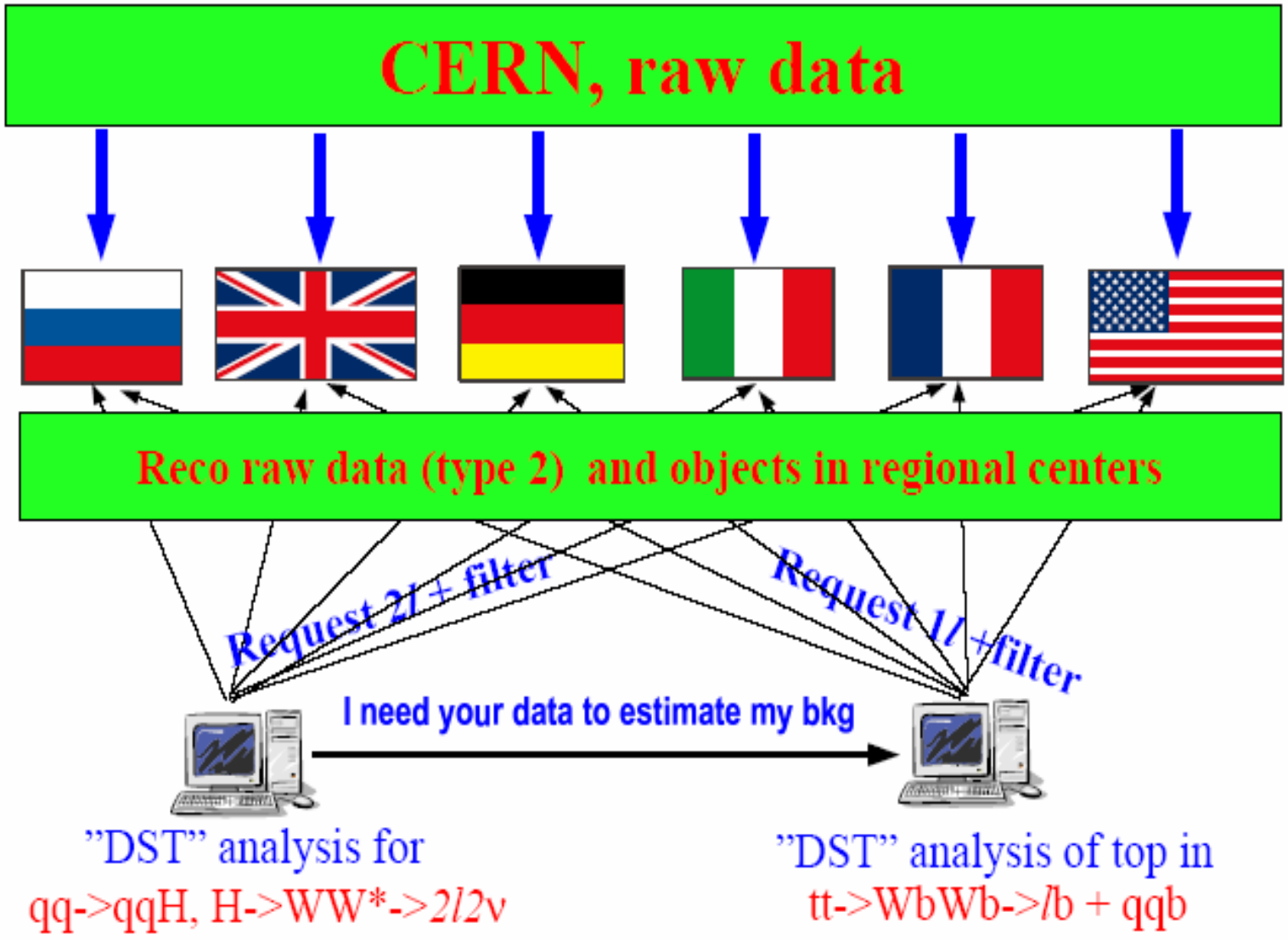
- CMS-ARDA (DAProm) Goal is to deliver an end-to-end (E2E)
- Distributed analysis system.

- The analysis should be possible at all the different event data (SimHits, Digi, DST,AOD, ...) level, using COBRA/ORCA code.

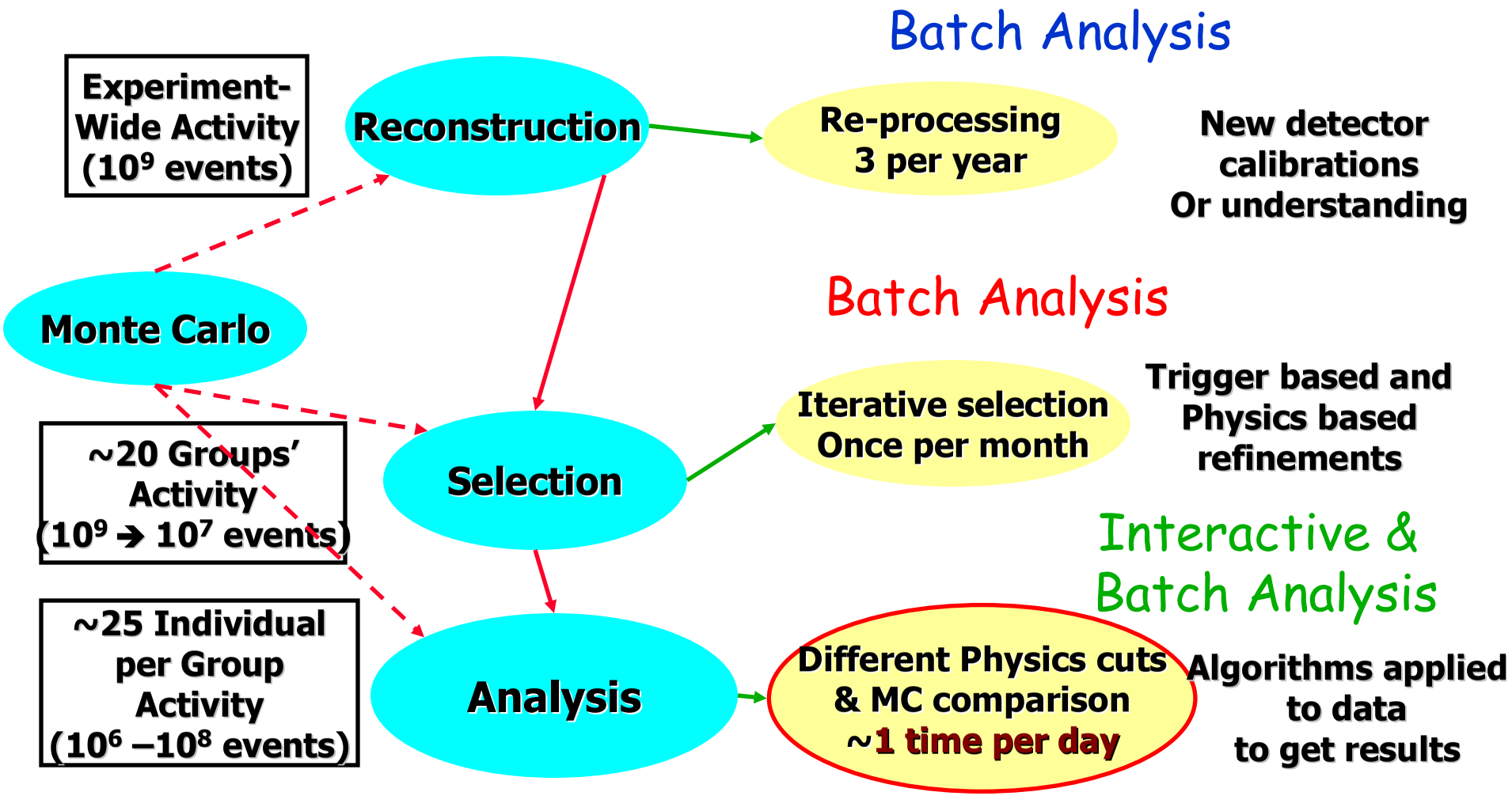
- Distributed analysis extends the analysis to an environment where the users, data and processing are distributed over the grid.

- Cms-Arda Mailing List: cms-arda@cern.ch

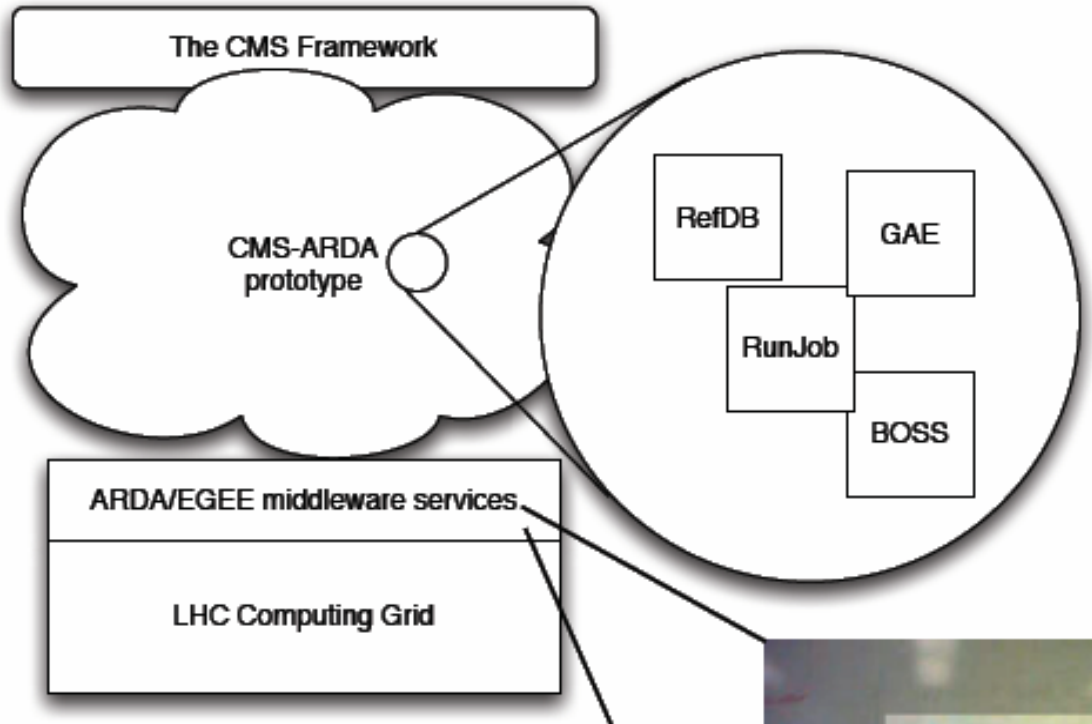
A Vision on CMS Analysis



Hierarchy of Processes (Experiment, Analysis Groups, End-User)



Possible role of CMS-ARDA project



Application-level services interfacing to COBRA and ARDA/ EGGE MW

CMS to diagonalize into engineered end-to-end system

"prototype of generic middleware stack capable of supporting prototypes of distributed analysis environments of experiments, powerful enough to support end-to-end capabilities required"

cms-Arda task Decomposition

- **What are the End-user inputs:**
  - **DataSet and owner (catalog)**
    - ➔ **(high level identification for event samples)**
  - **analysis programs executable + related shared libraries (+ standard cms+external shared libraries)**

- **Possible CMS-ARDA (DAProm) Decomposition**
  - **User Interface to production-service to request DataSet (data-products)**
  - **User Interface to production-service to monitor the status of a request and where data are located (TierN)**

cms-Arda task Decomposition

- **Workflow management tools for analysis-service**
  - DataSet fully attached to a single Catalog for each Tn
  - Grid Tools (services) to submit and monitor analysis job
- **Data Location service**
  - Discovery of remote catalogs and their content
- **Data-transfer service**
  - Like TMDB. Requirements:
  - A local catalog at the production site (one for each dataset) is required and the tool should populate the complete full local catalog at the remote site (Tn)
- **Interface to the storage-service**
  - In order to open/read/write files from an application (srm, nfs, afs, others...)
- **Analysis-task wizard**
  - Requirement: Provide a simple interface and guidance to the Physicists

First Steps for cms-ARDA Project

- **Short term:**

- **Give easy access to PRS for analysis asap to PCP and DC04 data available (July-August):**
  - **This will be done using SW already available (like LCG-RB, Grid-ICE, Monalisa and others.) + TMDB + other CMS specific Data-Management tools + latest COBRA/ORCA versions that already included requirements from end-user, like the possibility to use different level for POOL catalog + other components that could comes from GAE etc + few other new components.**
  - **Clearly we will use LCG-2 /Grid-3 SW that was already used during DC04 (Workload Management, Job monitoring ..)**
  - **People from different institutions (US, INFN, RAL, GRIDKA, CERN..)**

- **This is essential in order to provide data to PRS Groups.**

First Steps for cms-ARDA Project

- **Short term:**
- **Start/continue the evaluation of the EGEE SW (including also LCG-2 components) asap, provide feed-back in order to create hopefully together a system that could work for CMS Analysis offline and online**
  - **People from different institutions (CERN, US, INFN, RAL, GridKA)**
- **Longer term:**
  - **This will be address asap when new organization is in place!**