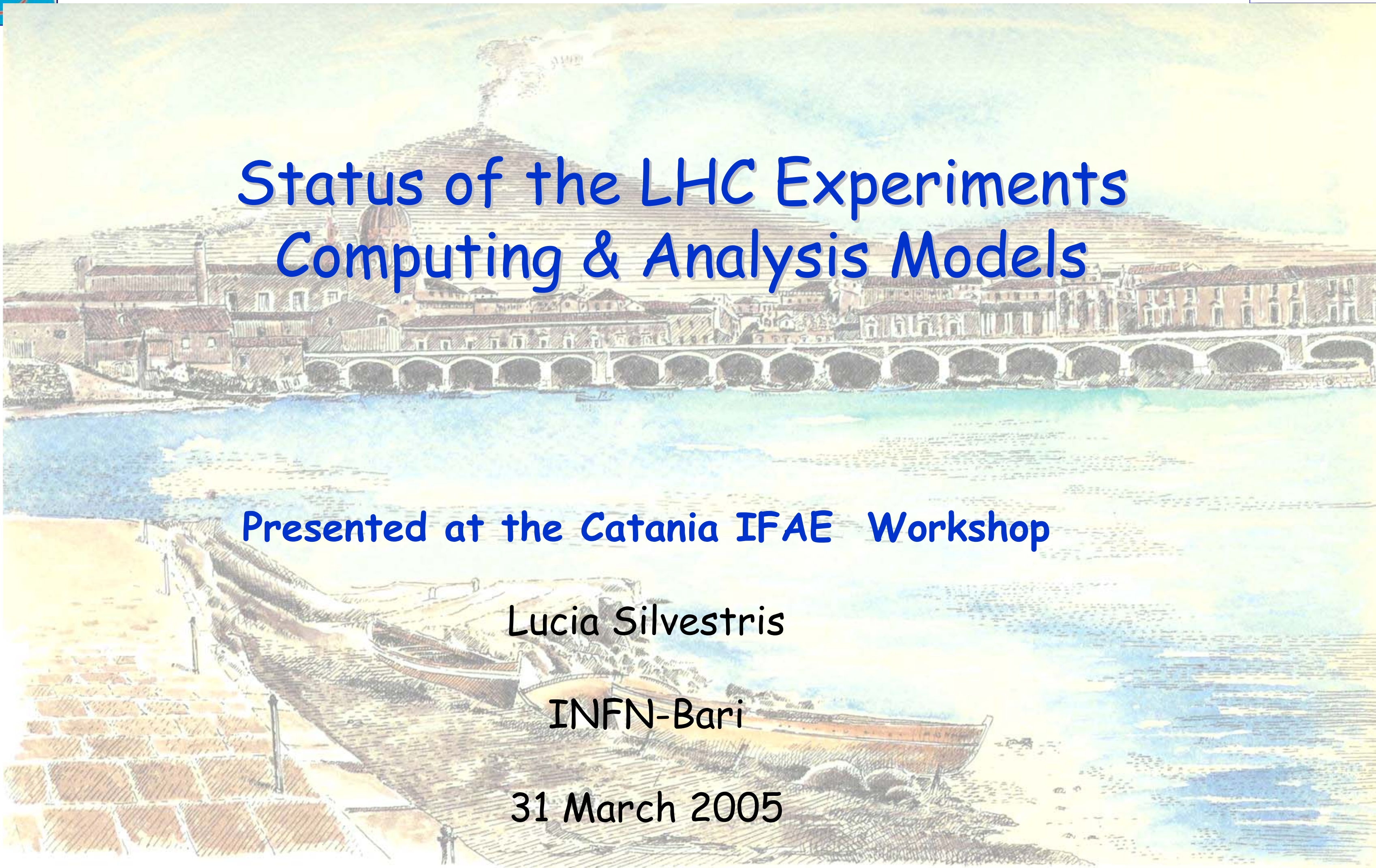


Status of the LHC Experiments Computing & Analysis Models

Presented at the Catania IFAE Workshop

Lucia Silvestris

INFN-Bari

31 March 2005

Computing Model Papers

Requirements from Physics groups and experience at running experiments

- ✓ Based on operational experience in Data Challenges, production activities, and analysis systems.
- ✓ Active participation of experts from CDF, D0, and BaBar
- ✓ DAQ/HLT TDR (ATLAS/CMS/LHCb/Alice) and Physics TDR (ATLAS)

Main focus is first major LHC run (2008)

- 2007 100 days (5×10^6s, 5×10^{32})
- 2008 200 days (10^7s, 2×10^{33}), 20 days(10^6s) Heavy Ions
- 2009 200 days (10^7s, 2×10^{33}), 20 days(10^6s) Heavy Ions
- 2010 200 days (10^7s, 10^{34}), 20 days(10^6s) Heavy Ions

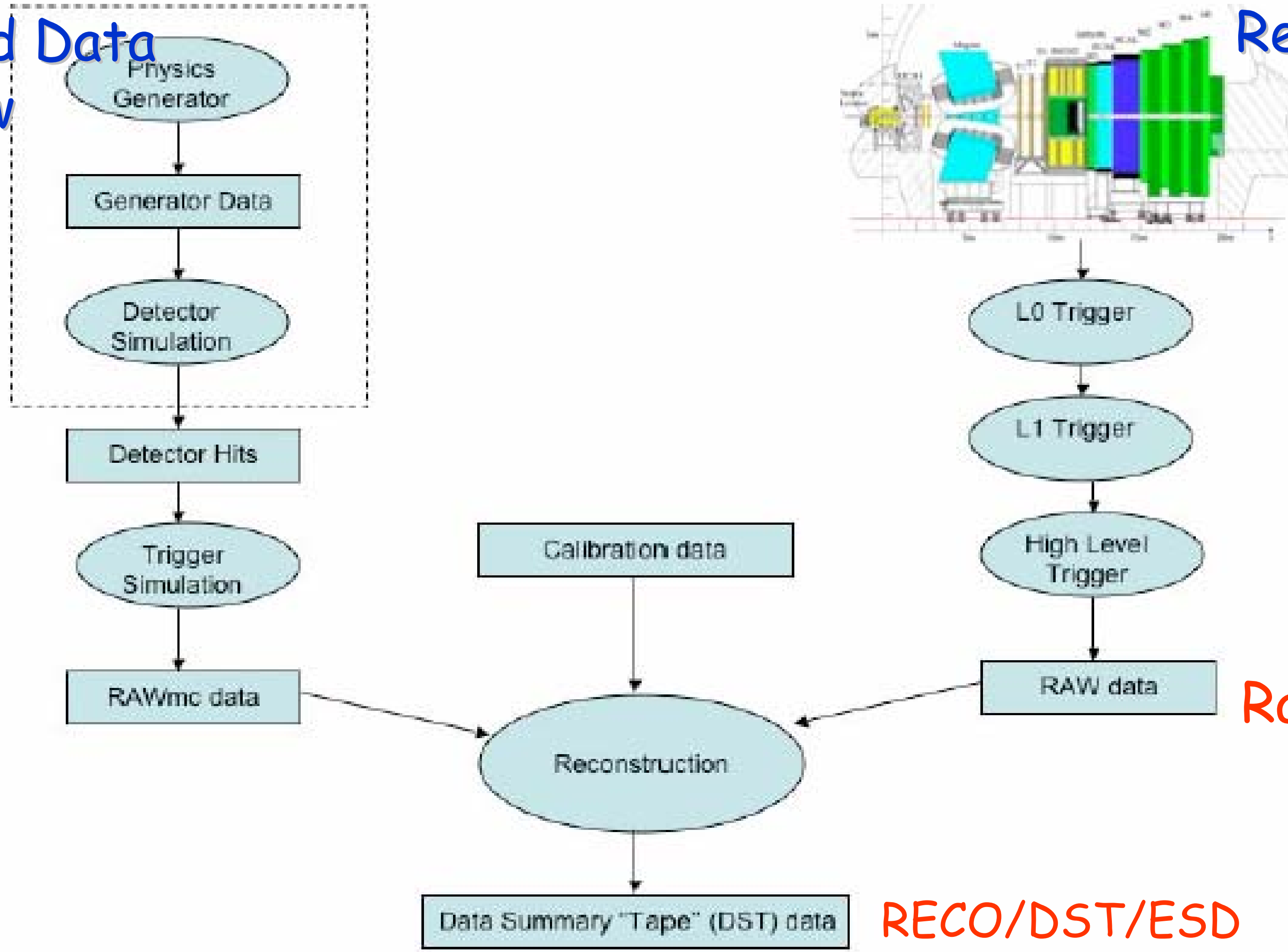
This talk focus on computing and analysis model for pp collision

Numbers from official experiments report to LHCC: Alice: CERN-LHCC- 2004-038/G-086, Atlas: CERN-LHCC-2004-037/G-085, CMS: CERN-LHCC-2004-035/G-083, LHCb: CERN-LHCC-2004-036/G-084

Examples: LHCb Event Data Flow

Simulated Data Flow

Real Data Flow



SIMU

Raw

RECO/DST/ESD

Event Data Model - Data Tiers

RAW
- Event format produced by event filter (byte-stream) or object data
- One copy spread over T1 centers and one at Tier-0
- Used for Detector Understanding, Code optimization, Calibrations,…

SIMU
- Simulated including event generator, geant4 simulation and digitization step
- 1 copy spread over T2, backup at T1 centers

RECO/DST/ESD
- Reconstructed hits, Reconstructed objects (tracks, vertices, jets, electrons, muons, etc.)
- Track Refitting, new MET
- 1 copy spread over T1 centers (together with associated RAW)
  - More copies possible for smaller/hot datasets
- Used by all Early Analysis, and by some detailed Analyses

AOD
- Reconstructed objects (tracks, vertices, jets, electrons, muons, etc.).
- Possible small quantities of very localized hit information.
- All streams at every T1 center, many streams at T2 centers
- Used by most Physics Analysis

TAG
- event-level metadata for fast search and selection in a database
- Self describing data, can be processed without any experiment code
- All at every T1 center, many streams at T2 centers

Inputs to LHC Computing Models

Raw Data size is estimated to be 1.5MB for 2x1033 first full physics run

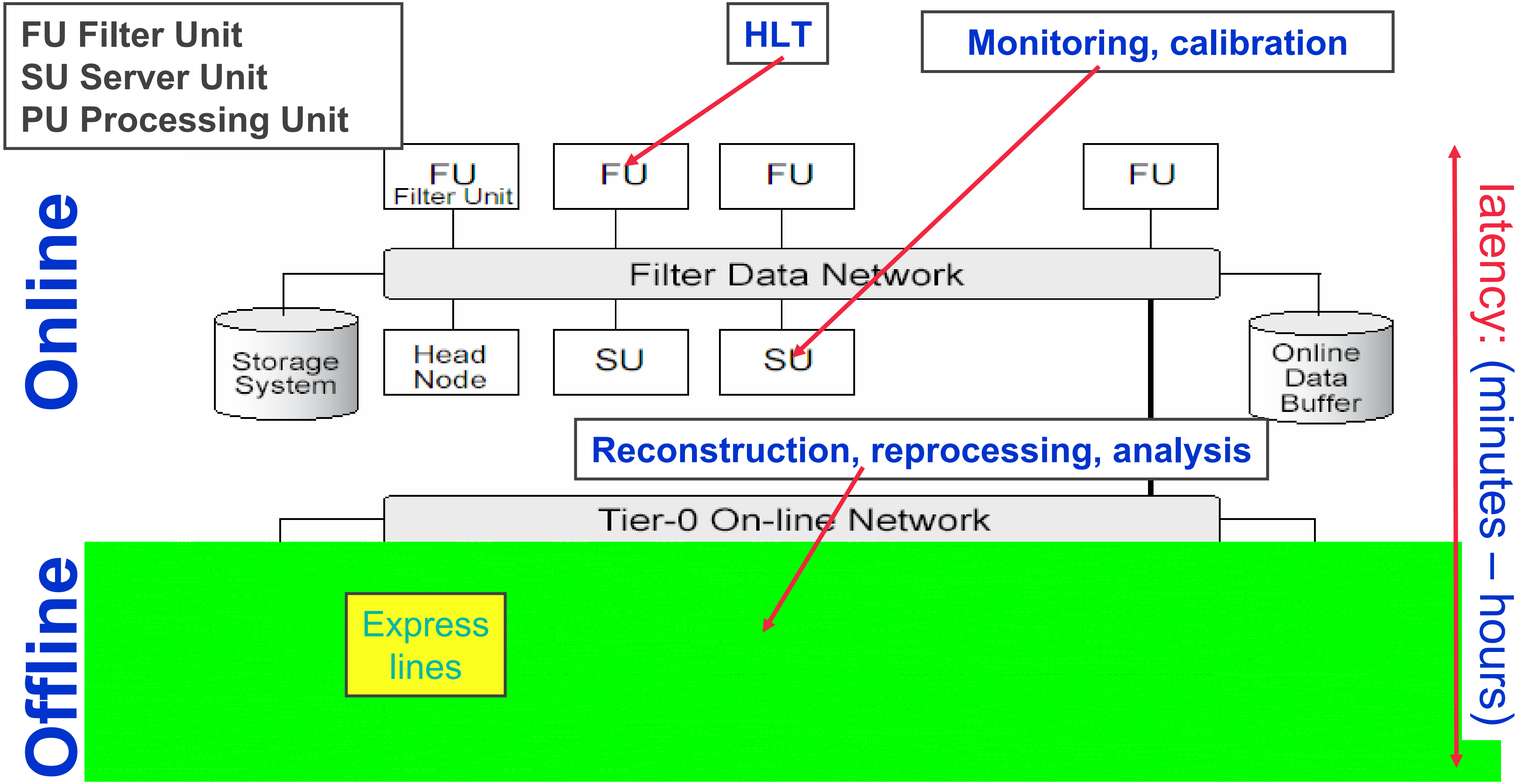
- ~300kB (Estimated from current MC)
- Multiplicative factors drawn from CDF experience
  - -- MC Underestimation factor 1.6
  - -- HLT Inflation of RAW Data, factor 1.25
  - -- Startup, thresholds, zero suppression,…. Factor 2.5
- Real initial event size more like 1.5MB
  - -- Expect to be in the range from 1 to 2 MB
    - Use 1.5 as central value
- Hard to deduce when the event size will fall and how that will be compensated by increasing Luminosity

Event Rate is estimated to be 150Hz for $2x10^{33}$ first full physics run

- Minimum rate for discovery physics and calibration: 105Hz (DAQ TDR)
- +50Hz Standard Model (jets, hadronic, top,…)

CMS Event Data Flow - Event Filter -> Tier 0

FU Filter Unit
SU Server Unit
PU Processing Unit



The result of the reconstruction will be saved along with the raw data in a database (POOL/ROOT)

Event-Filter -> Tier-0

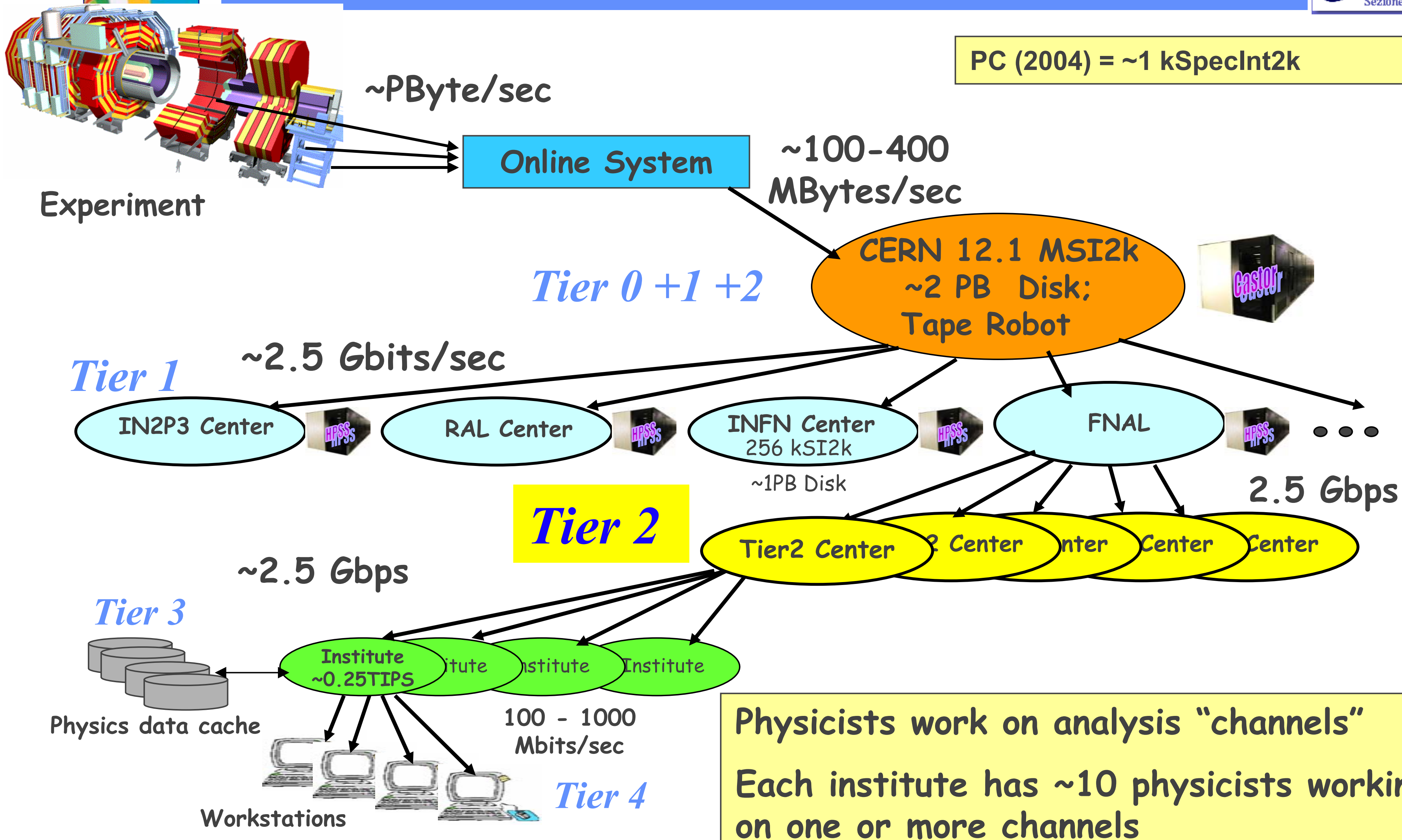
- HLT (Event Filter) is the final stage of the online trigger
- Baseline is several streams coming out of Event Filter
  - Primary physics data streams
  - Rapid turn-around "express line"
  - Rapid turn-around calibration events
  - Debugging or diagnostics stream (e.g. for pathalogical events)
- Main focus here on primary physics data streams
  - Goal of express line and calibration stream is low latency turn-around
  - Calibration stream results used in processing of production stream
  - Express line and calibration stream contribute ~20% to bandwidth
    - Detailed processing model for these is still under investigation

CMS Example: Tier-0 Operations

- Online Streams arrive in a 20 day input buffer
  - They are split into Primary Datasets (50) that are concatenated to form reasonable file sizes
  - Primary Dataset RAW data is:
    - archived to tape at Tier-0
      - Allowing Online buffer space to be released quickly
    - Sent to reconstruction nodes in the Tier-0
- Resultant RECO Data is concatenated (zip) with matching RAW data to form a distributable format FEVT (Full Event)
  - RECO data is archived to tape at Tier-0
  - FEVT are distributed to Tier-1 centers (T1s subscribe to data, actively pushed)
    - Each Custodial Tier-1 receives all the FEVT for a few 5-10 Primary Datasets
    - Initially there is just one offsite copy of the full FEVT
  - First pass processing on express/calibration physics stream
  - 24-48 hours later, process full physics data stream with reasonable calibrations
  - AOD copy is sent to each Tier-1 center

LHC Data Grid Hierarchy

PC (2004) = ~1 kSpecInt2k



Physicists work on analysis “channels”

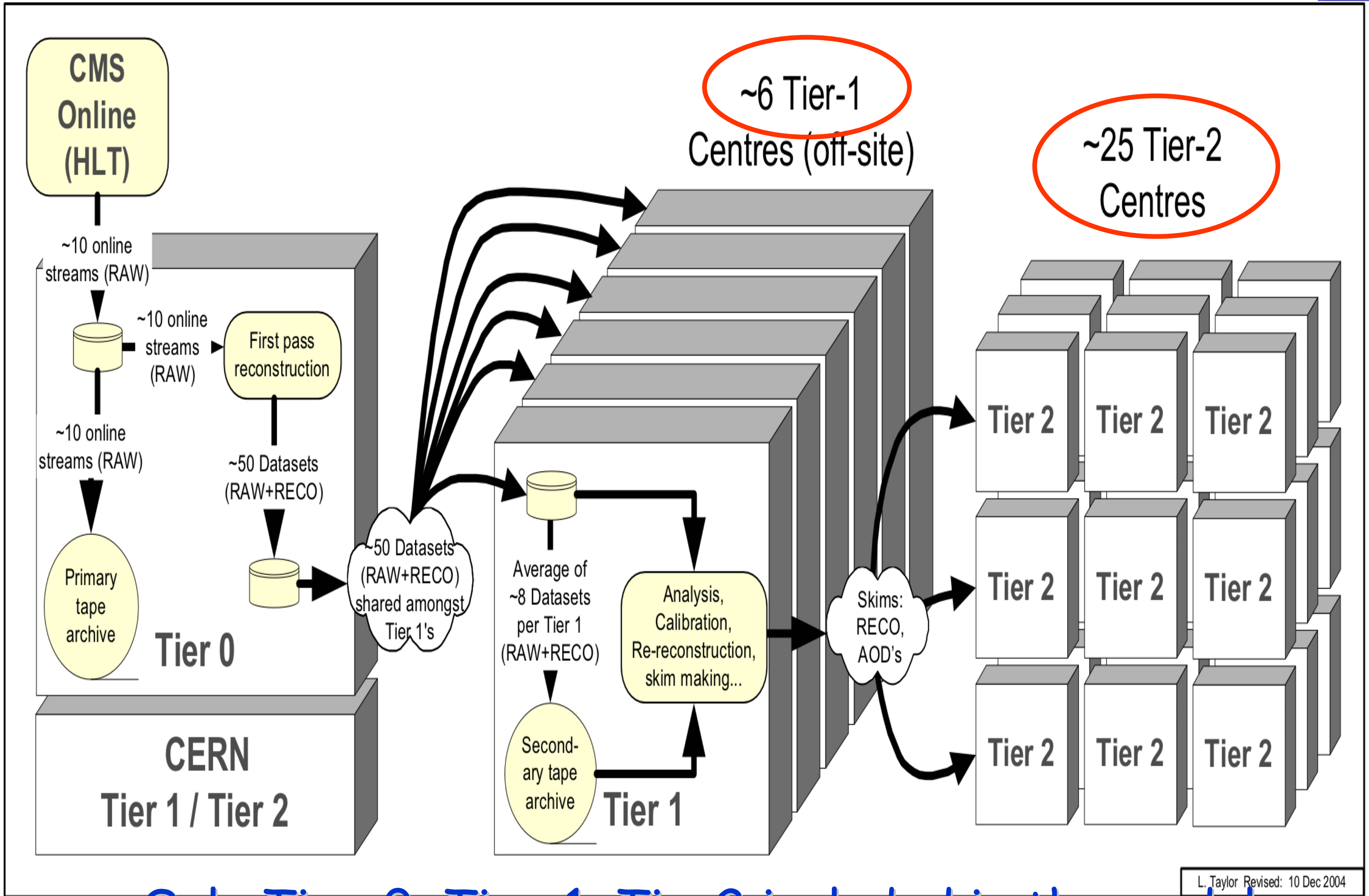
Each institute has ~10 physicists working on one or more channels

Computing Resources

Hierarchy

- Tier-0 has raw+calibration data+first-pass DST (ESD)
- CERN Analysis Facility has AOD, ESD and RAW samples
- Tier-1s curate RAW data and derived samples and 'shadow' the DST (ESD) for another Tier-1
- Tier-1s also house simulated data
- Tier-1s provide reprocessing for their RAW and scheduled access to full DST (ESD) samples
- Tier-2s provide access to AOD and group Derived Physics Datasets and carry the full simulation load

Event Data Flow in the Computing Models



Only Tier-0, Tier-1, Tier2 included in the models

CMS Example Calculation: Tier-0 CPU

Required CPU = 4588 kSI2k =
Scheduled_CPU / EffSchCPU

Scheduled_CPU = 3900 kSI2k =
Reco_CPU + Calib CPU

Reco_CPU = 3750 kSI2k =
(NRawEvts x RecCPU/ev)
/LHCYear

Calib_CPU = 150 kSI2k =
(NRawEvts x CalFrac x
CalCPU/ev)/LHCYear

NRawEvts = 1.5×10^9 =
L2Rate x LHCYear

L2Rate	=150Hz
LHCyear	=10^7 Sec
RecCPU	=25kSI2k/ev
CalCPU	=10kSI2k/ev
CalFrac	=10%
EffSchCPU	=85%

CMS Example Calculation: Tier-0 Tape

Required Tape = 3775 TB =
Annual_Tape / EffTape(100%)

Annual_Tape = 3775 TB =
SUM(RAW+HIRaw+Calib+1stReco+2ndReco +HIReco+1stAOD+2ndAOD)

Raw	=	2250	TB
HIRaw	=	350	TB
Calib	=	225	TB
1stReco	=	375	TB
2ndReco	=	375	TB
HIReco	=	50	TB
1stAOD	=	75	TB
2ndAOD	=	75	TB

Tier-0 Specifications

Efficiency for scheduled CPU	85%
Efficiency for "chaotic" CPU	60-75%
Disk utilization efficiency	70%
Mass Storage utilization efficiency	100%

p-p collision

	Units	ATLAS	CMS	LHCb
Recon. Time/ev	kSI2k sec	15	25	2.4
Simul. Time/ev	kSI2k sec	100	45	50

	Units	ATLAS	CMS	LHCb
Tier 0 CPU	MSI2k	4.1	4.6	
CPU at CERN	MSI2k	6.3	7.5	0.9
Tier 0 Disk	PB	0.35	0.41	
Disk CERN	PB	1.95	1.71	0.8
Tier 0 Tape	PB	4.2	3.8	
Tape CERN	PB	4.6	5.6	1.4

Tier-1 Operations

Receive Custodial data (FEVT (RAW+DST) and AOD)

- Current Dataset "on disk"
- Other bulk data mostly on tape with disk cache for staging
- Good tools needed to optimize this splitting

Receive Reconstructed Simulated events from Tier-2

- Archive them, distribute out AOD for Simu data to all other Tier-1 sites

Serve Data to Analysis groups running selections, skims, re-processing

- Some local analysis possibilities
- Most analysis products sent to Tier-2 for iterative analysis work

Run reconstruction/calibration/alignment passes on local RAW/RECO and SIMU data

- Reprocess 1-2 months after arrival with better calibrations
- Reprocess all resident RAW at year end with improved calibration and software

Operational 24h*7day

CMS Example Calculation: Tier-1 CPU

**Required CPU = 2128 kSI2k =
Scheduled_CPU (1199) / EffSchCPU+
Analysis_CPU (929) /EffAnalCPU**

**Scheduled_CPU = 1019 kSI2k =
ReReco_Data+ReReco_Simu**

**ReReco_Data = 510 kSI2k =
(NRawEvts/NTier1 x RecCPU/ev)
/(SecYear x NReReco/yr x 6/4)**

**ReReco_Simu = 510 kSI2k =
(NSimEvts/NTier1 x RecCPU/ev)
/(SecYear x NReReco/yr x 6/4)**

**Analysis_CPU = 697 kSI2k =
Selection+Calibration**

**Selection = 672 kSI2k =
(NRawEvts+NSimEvts) /
(NTier1-1) x SelCPU/ev) /
TwoDay**

**Calibration = 25 kSI2k =
(NRawEvts / (NTier1-1) x
CalFrac x CalCPU/ev) /
LHCyear**

NRawEvts = 1.5 x 10^9=NSimEvts
LHCyear = 10^7 Sec
RecCPU = 25kSI2k/ev
SelCPU = 0.25 kSI2k/ev

CalCPU = 10kSI2k/ev
CalFrac = 10%
EffSchCPU = 85%
EffAnalCPU= 75%

NReReco/yr = 2
"6/4"- complete rereco
In 4 months, not 6

Example Calculation: Tier-1 Data Serving Rate

Selection = **672** kSI2k = (NRawEvts+NSimEvts) / (NTier1-1) x SelCPU/ev) / TwoDay

Data I/O Rate ≈ 800 MB/s =
Local Sim+Data Reco Sample size / TwoDay

Note, one complete selection pass every two days, is also/only one pass every month for each of 10-15 analysis groups

Tier-1 Specifications

Average for each T1 p-p collision

	Units	ATLAS	CMS	LHCb
Tier 1 CPU	**MSI2k**	1.8	2.1	0.73
Tier 1 Disk	**PB**	1.23	1.11	0.4
Tier 1 Tape	**PB**	0.65	1.85	0.35

ΣT1 Atlas 10
CMS 6
LHCb 6

	Units	ATLAS	CMS	LHCb
Tier 1 CPU	**MSI2k**	18	12.8	4.4
Tier 1 Disk	**PB**	12.3	6.7	2.4
Tier 1 Tape	**PB**	6.5	11.1	2.1

Tier-2 Operations

Run Simulation Production and calibration

- Not requiring local staff, jobs managed by central production via Grid. Generated data is sent to Tier-1 for permanent storage.

Serve "Local" or Physics Analysis groups

- (20-50 users?, 1-3 groups?)
- Local Geographic? Physics interests?
- Import their datasets (production, or skimmed, or reprocessed)
- CPU available for iterative analysis activities
- Calibration studies (and calibration processing?)
- Studies for Reconstruction Improvements
- Maintain on disk a copy of AODs and locally required TAGs.

Some Tier-2 centres will have large parallel analysis clusters (suitable for PROOF or similar systems).

- It is expected that clusters of Tier-2 centres ("mini grids") will be configured for use by specific physics groups.

T2 Specifications

CMS Example: Average T2 center

p-p collision

			Eff Factors
CPU scheduled	250	kSI2K	*85.00%*
CPU analysis	579	kSI2K	*75.00%*
Disk	218	Tbytes	*70.00%*

ΣT2 Atlas ~30
CMS ~25
LHCb ~14

	Units	ATLAS	CMS	LHCb
Tier 2 CPU	**MSI2k**	**16.2**	**19.9**	**7.6**
Tier 2 Disk	**PB**	**6.9**	**5.3**	**0.02**
Tier 2 Tape	**PB**	**0**	**0**	**0**

Networks

CMS more than ATLAS and LHCb is pushing available networks to their limits in the Tier-1/Tier-2 connections

- Tier -0 needs ~2x10Gb/s links for CMS
- Each Tier-1 needs ~10Gb/s links
- Each Tier-2 needs 1Gb/s for its incoming traffic

- There will be extreme upward pressure on these numbers as the distributed computing becomes more and more useable and effective

Service Challenges with LCG, CMS Tier-1 centers and CMS Data Management team/components planned for 2005 and 2006

- Ensure that we are on path to achieve these performances.

Main Uncertainties on the Computing & Analysis models

- Chaotic user analysis of augmented AOD streams, tuples (skims), new selections etc and individual user simulation and CPU-bound tasks matching the official MC production

- Calibration and conditions data.

Calibration & Conditions data

Conditions data: all non-event data required for subsequent data processing

1. Detector control system data (DCS) – 'slow controls' logging
2. Data quality/monitoring information – summary diagnostics and histograms
3. Detector and DAQ configuration information
   - Used for setting up and controlling runs, but also needed offline
4. 'Traditional' calibration and alignment information

- Calibration procedures determine (4) and some of (3), others have different sources
   - Also need for bookkeeping 'meta-data', but not considered part of conditions data

Possible strategy for conditions data (ATLAS Example):

- All stored in one 'conditions database' (condDB) - at least at conceptual level
- Offline reconstruction and analysis only accesses condDB for non-event data
- CondDB is partitioned, replicated and distributed as necessary
   - Major clients: online system, subdetector diagnostics, offline reconstruction & analysis
   - Will require different subsets of data, and different access patterns
   - Master condDB held at CERN (probably in computer centre)

Calibration processing strategies

Different options for calibration/monitoring processing – all will be used

- Processing in the sub-detector readout systems
  - In physics or dedicated calibration runs, only partial event fragments, no correlations
  - Only send out limited summary information (except for debugging purposes)
- Processing in the HLT system
  - Using special triggers invoking 'calibration' algorithms, at end of standard processing for accepted (or rejected) events – need dedicated online resources to avoid loading HLT?
  - Correlations and full event processing possible, need to gather statistics from many processing nodes (e.g. merging of monitoring histograms)
- Processing in a dedicated calibration step before prompt reconstruction
  - Consume the event filter output – physics or dedicated calibration streams
  - Only bytestream RAW data would be available, results of EF processing largely lost
  - A place to merge in results of asynchronous calibration (e.g. optical alignment systems)
  - Potentially very resource hungry – ship some calibration data to remote institutions?
- Processing after prompt reconstruction
  - To improve calibrations ready for subsequent reconstruction passes
  - Need for access to DST (ESD) and raw data for some tasks – careful resource management

Event streaming for prompt calibration

Data streams from the event filter

1. Bulk physics data stream (~300 MB/sec)
2. Express physics stream (duplicating events in bulk stream)
3. Dedicated calibration streams
4. Diagnostic and debugging stream (problem events)

Motivation and role of calibration streams

- Read out of calibration triggers not useful for physics
  - May be processed differently
- Partial detector readout (selected subdetectors only, regions of interest through whole detector around lepton candidates)
  - Implications for TDAQ system being studied
- Separate out events useful for calibration and subdetector diagnostics from bulk physics sample
  - Easier and more efficient access to selected data, especially during start up phase
  - Implies some duplication of data in bulk physics and/or express stream
- Calibration + express stream should consume ~20% of bandwidth

Prompt reconstruction latency

- Calibration streams provide input to determine calibration/alignment for first-pass reconstruction
  - Calibration data arrives at Tier-0 buffer disk with minimal latency
  - Processing can start soon after end of fill, or even during fill itself
- Typical tasks during calibration step
  - Process calibration stream data for fill or subset (may need event reconstruction)
  - Derive updated calibration constants and upload to conditions database
    - Also incorporate results of 'asynchronous' calibration processes (e.g. optical alignment)
  - Verify correctness of constants
    - Re-reconstruct control samples of events (part of calibration stream, or express?)
    - Manual human checking may be required, at least initially
  - Initial target to be ready for bulk physics reconstruction 24 hours after end of fill
    - Time to process, derive constants, re-reconstruct and check on ~10% of full data sample – needs O(10%) Tier 0 reconstruction resources in steady state
    - Anticipate need to devote greater resources during startup, process over and over
    - Obvious place to use remote resources – ideas, but no concrete plans as yet
- Process is not fast enough for express stream – use constants from last fill?

Offline calibration and alignment

Processing after pass 1 reconstruction

- To improve calibration constants ready for subsequent reconstruction passes
- 'Analysis' type processing – individual groups working independently to understand all details of subdetector performance and calibration
- But requires access to ESD and sometimes RAW data – resource-hungry
  - Passes over large samples of RAW (and ESD) data will have to be centrally scheduled and coordinated

Subdetector calibration groups starting to consider these issues

- First definition of DST (ESD) now available
  - What calibration tasks can be done with what datatype?
  - What changes could be made to improve usability of samples
  - – e.g. on ESD add hits not associated but close to a track to allow iterating ID pattern recognition after alignment, without going back to RAW data
- Calibration issues starting to receive higher priority after combined testbeam
  - Detailed definition of calibration streams and samples, going beyond what was presented today
  - Discussions with Tridas (TDAQ) on feasibility of various calibration strategies and run types

Getting ready for April `07

LHC experiments are engaged in an aggressive program of "data challenges" of increasing complexity.

Each is focus on a given aspect, all encompass the whole data analysis process:

- Simulation, reconstruction, statistical analysis
- Organized production, end-user batch job, interactive work

Past: Data Challenge `02 & Data Challenge "04

Near Future: Cosmic Challenge end '05-begin "06

Future: Data Challenge `06 or Software & Computing Commissioning Test.

Examples : CMS HLT Production 2002

Focused on High Level Trigger studies

- 6 M events = 150 Physics channels
- 19000 files = 500 Event Collections = 20 TB

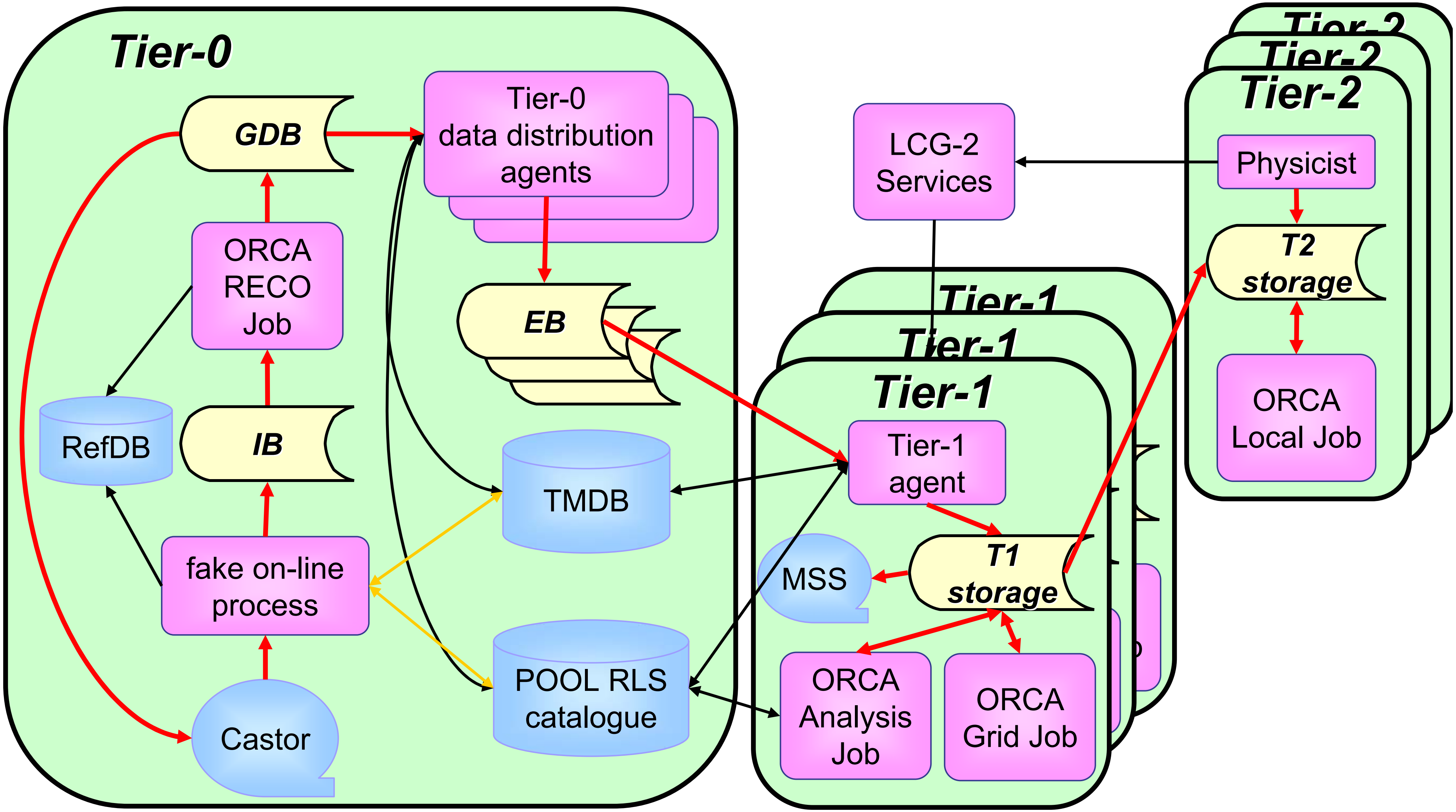
  NoPU: 2.5M, $2x10^{33}$PU:4.4M, 10^{34}PU: 3.8M, filter: 2.9M
- 100 000 jobs, 45 years CPU (wall-clock)
- 11 Regional Centers
  - > 20 sites in USA, Europe, Russia
  - ~ 1000 CPUs
- More than 10 TB traveled on the WAN
- More than 100 physics involved in the final analysis

GEANT3, Objectivity, Paw, Root
CMS Object Reconstruction & Analysis Framework COBRA and applications ORCA

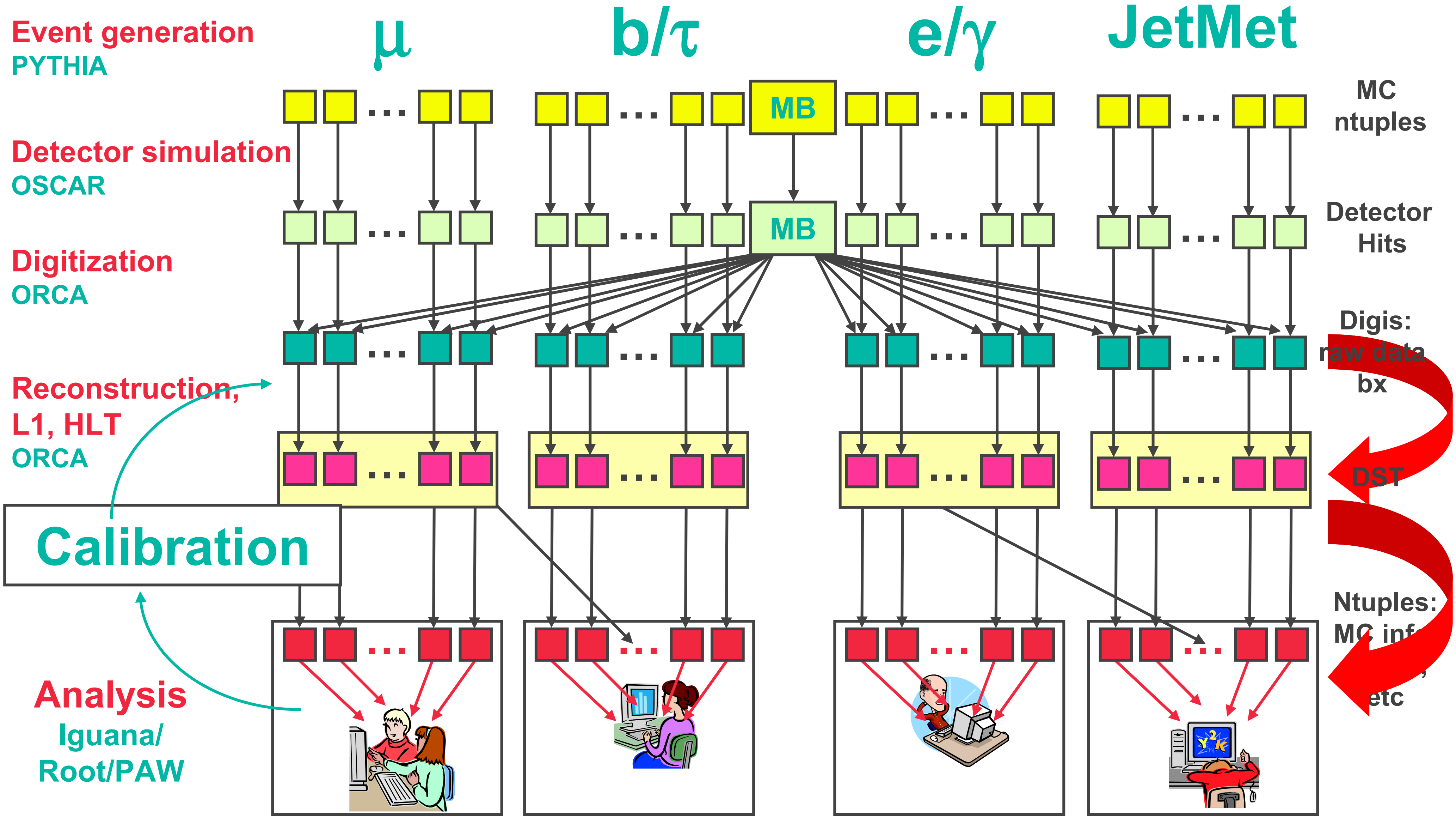
Successful validation of CMS High Level Trigger Algorithms

Rejection factors, computing performance, reconstruction-framework
Results published in DAQ/HLT TDR December 2002

Examples : CMS Data Challenge 04



Examples : Data Challenge 2004

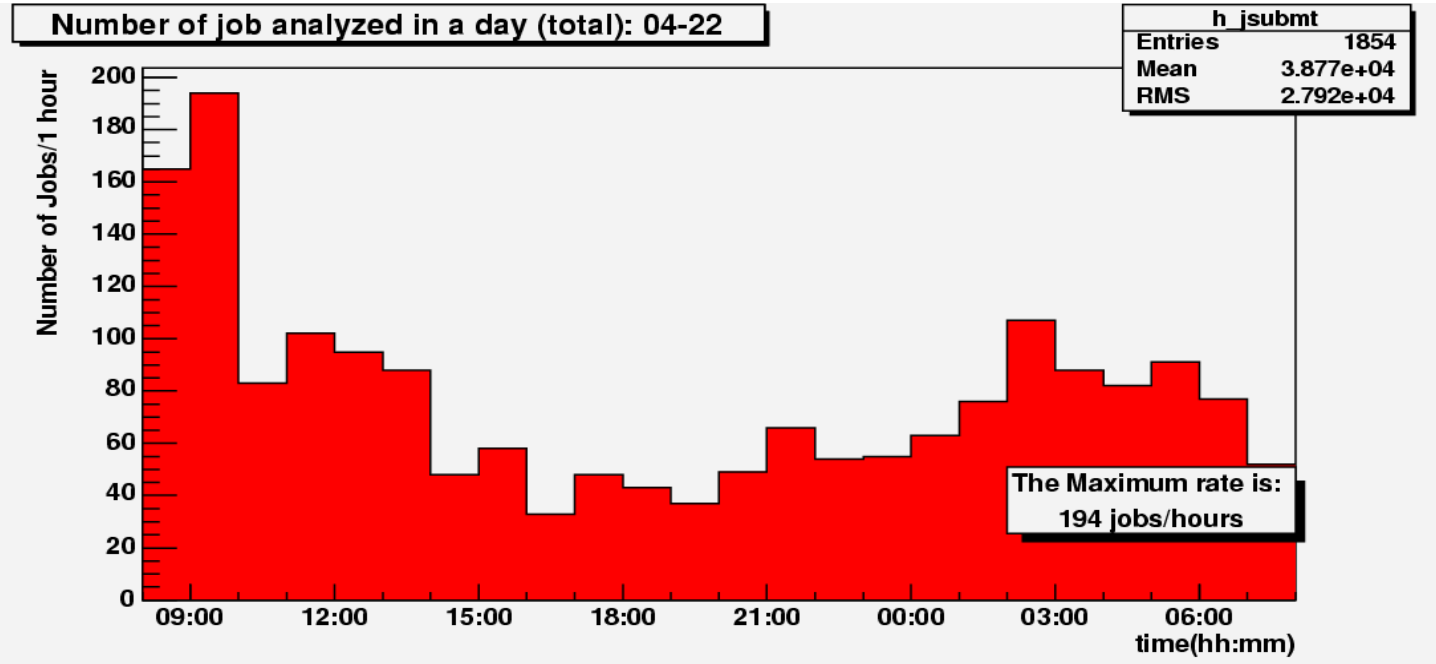


DC04 Real-time Analysis

- Maximum rate of analysis jobs: **194 jobs/hour**

INFN

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

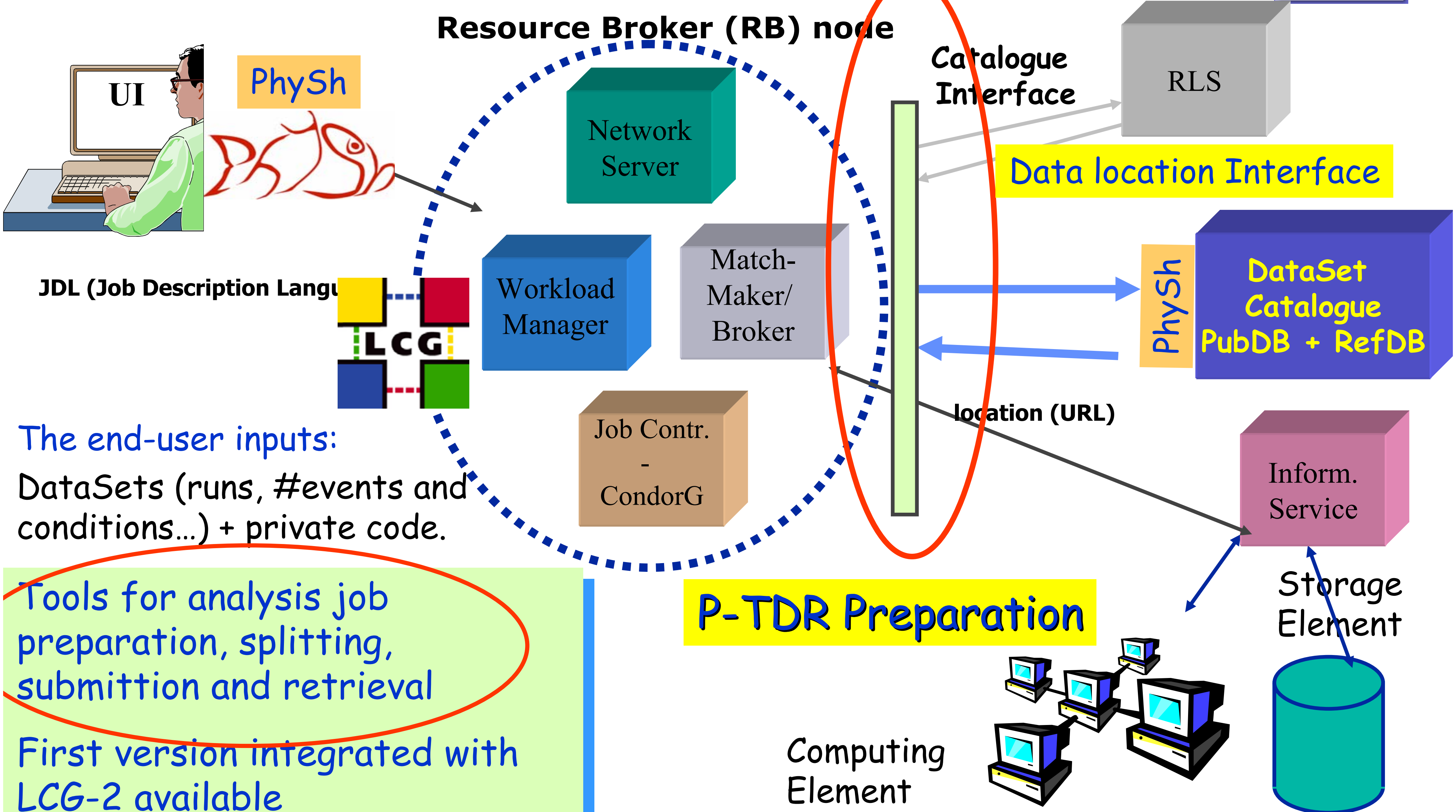


➢ Datasets examples:

- $B^0_s \rightarrow J/\psi\ \varphi$
  Bkg: mu03_tt2mu, mu03_DY2mu
- tTH, $H \rightarrow bbbar$ $t \rightarrow Wb$ $W \rightarrow l\nu$ $T \rightarrow Wb$ $W \rightarrow had.$
  Bkg: bt03_ttbb_tth
  Bkg: bt03_qcd170_tth
  Bkg: mu03_W1mu
- $H \rightarrow WW \rightarrow 2\mu\ 2\nu$
  Bkg: mu03_tt2mu, mu03_DY2mu

Using LCG-2

CMS Examples: Analysis Submission using Grid



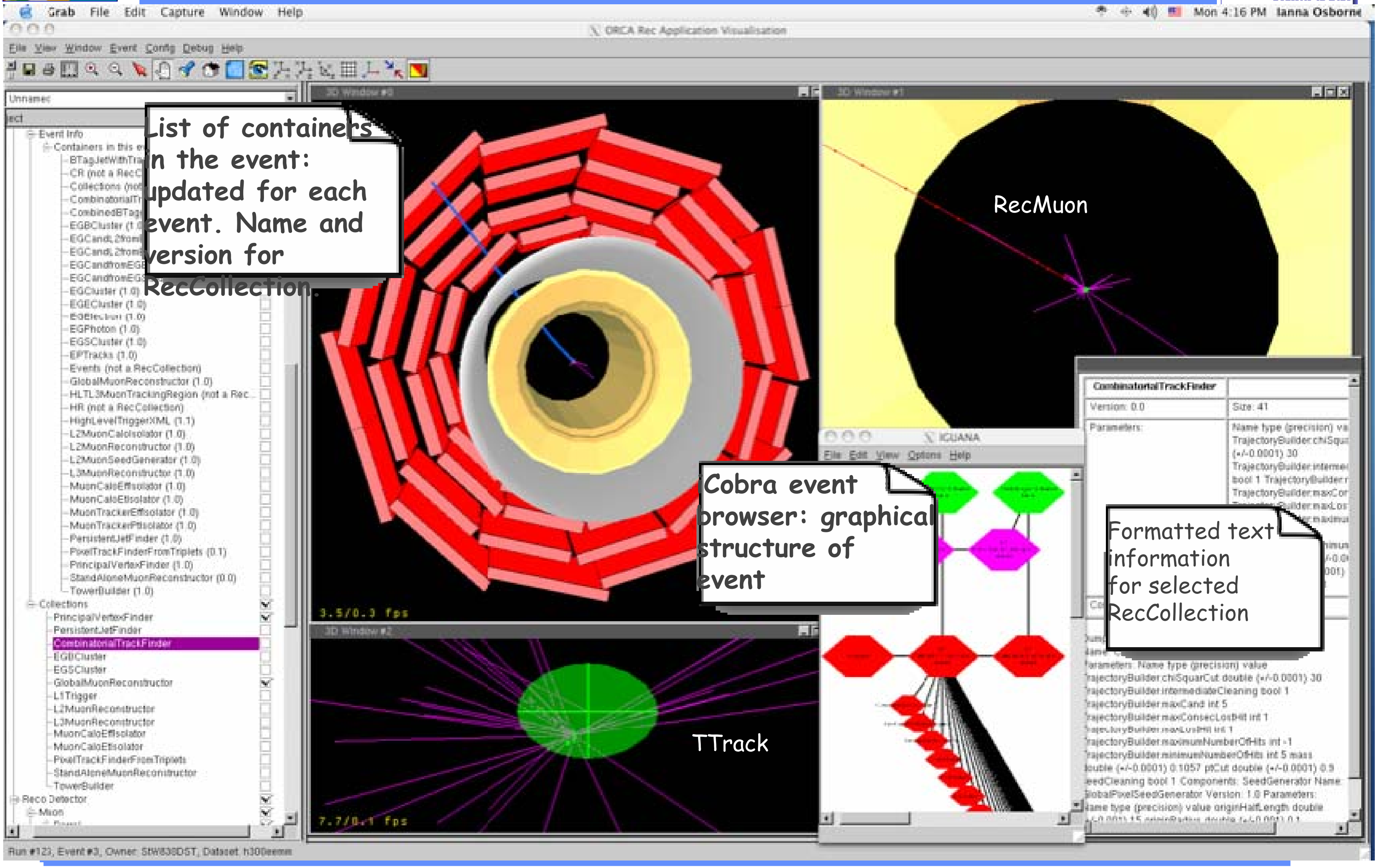
The end-user inputs:

DataSets (runs, #events and conditions…) + private code.

Tools for analysis job preparation, splitting, submittion and retrieval

First version integrated with LCG-2 available

Interactive Analysis/Inspection/Debugging: First version for DST Visualization



Computing & Software Commissioning Goals

Data challenge "DC06" should be consider as a Software & Computing Commissioning with a continuous operation rather than a stand-alone challenge.

Main aim of Software & Computing Commissioning will be to test the software and computing infrastructure that we will need at the beginning of 2007:

- Calibration and alignment procedures and conditions DB
- Full trigger chain
- Tier-0 reconstruction and data distribution
- Distributed access to the data for analysis

At the end (autumn 2006) we will have a working and operational system, ready to take data with cosmic rays at increasing rates.

Computing & Software Commissioning

- Sub-system (component) tests with well-defined goals, preconditions, clients and quantifiable acceptance tests
  - Full Software Chain
    - Generators through to physics analysis
  - DB/ Calibration & Alignment
  - Event Filter & Data Quality Monitoring
  - Physics Analysis
  - Integrated TDAQ/Offline
  - Tier-0 Scaling
  - Distributed Data Management
  - Distributed Production (Simulation & Re-processing)
- Each sub-system is decomposed into components
  - E.g. Generators, Reconstruction (DST creation)
- Goal is to minimize coupling between sub-systems and components and to perform focused and quantifiable tests

Computing & Software System Commissioning

Several different tests

- Physics Performance - e.g.
  - Mass resolutions, residuals, etc.
- Functionality - e.g.
  - Digitization functional both standalone and on Grid
- Technical Performance - e.g.
  - Reconstruction CPU time better than 400%, 200%, 125%, 100% of target (target need to be defined)
  - Reconstruction error in $1/10^5$, $1/10^6$, etc. events
  - Tier-0 job success rate better than 90%, etc.

Summary

Computing & Analysis Models

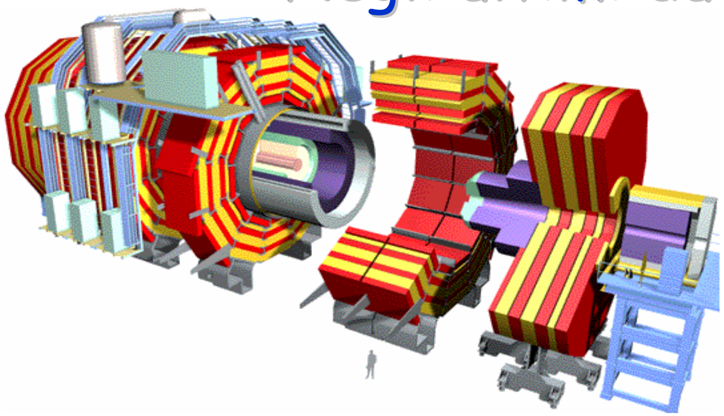
- Maintains flexibility wherever possible

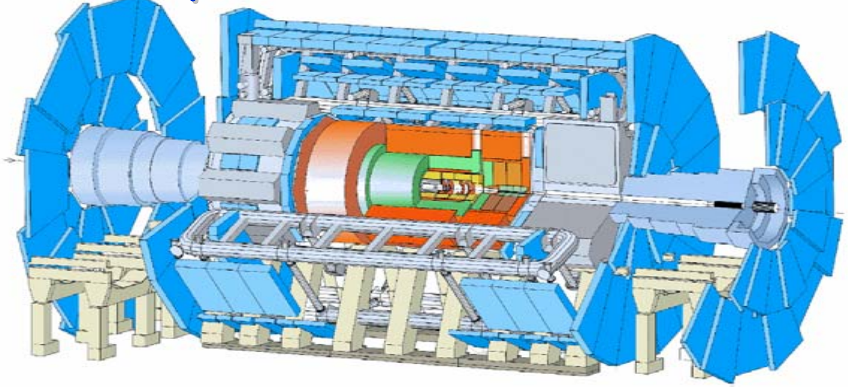
There are (and will remain for some time) many unknowns

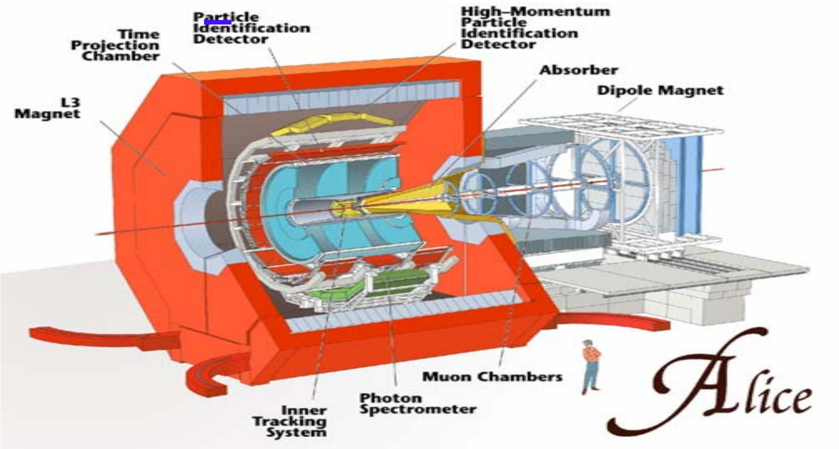
- Calibration and alignment strategy is still evolving (DC2 Atlas) & Cosmic Data Challenge (CMS)
- Physics data access patterns start to be exercised this Spring (Atlas) or P-TDR preparation (CMS)
  - Unlikely to know the real patterns until 2007/2008!
- Still uncertainties on
  - the event sizes
  - # of simulated events
  - on software performances (time needed for reconstruction, calibration (alignment), analysis …)
  - ……….

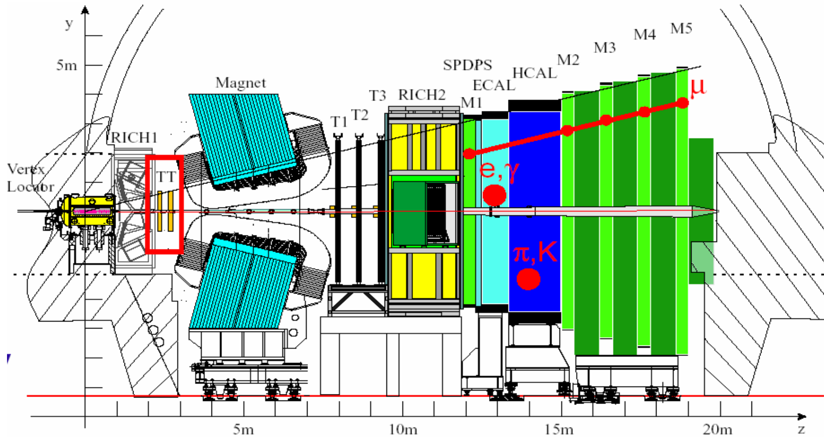
Summary

Negli ultimi due anni prima della presa dati…









Sara' fondamentale un corretto commissioning dei Rivelatori, del Event-Filter e del sistema di software e computing ….
Questo consentira' di esplorare …

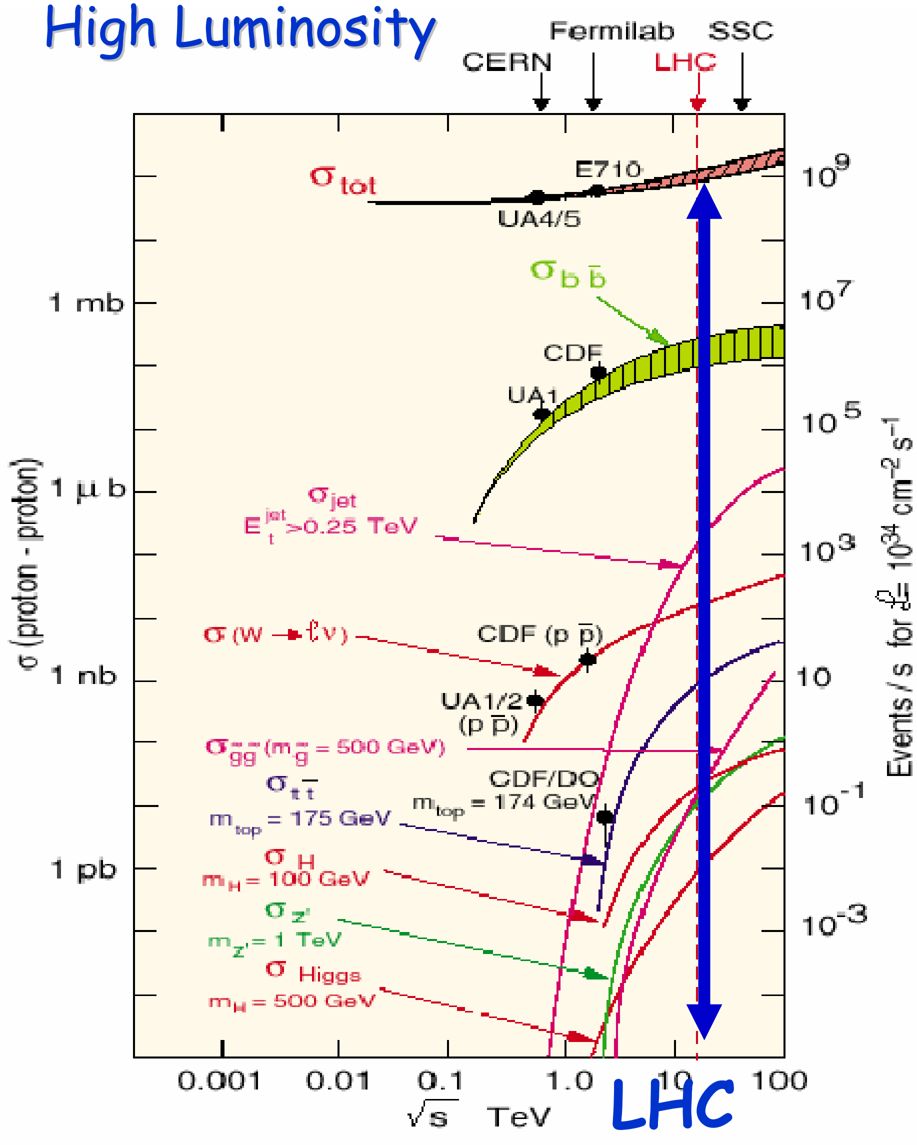
Summary

This physic program..

Cross-sections of physics processes vary over many orders of magnitude:

- inelastic: 10^9 Hz
- $b\bar{b}$ production: 10^6-10^7 Hz
- $W \to \ell\nu$: 10^2 Hz
- $t\bar{t}$ production: 10 Hz
- Higgs (100 GeV/c^2): 0.1 Hz
- Higgs (600 GeV/c^2): 10^{-2} Hz
- SuSy and BSM....

High Luminosity

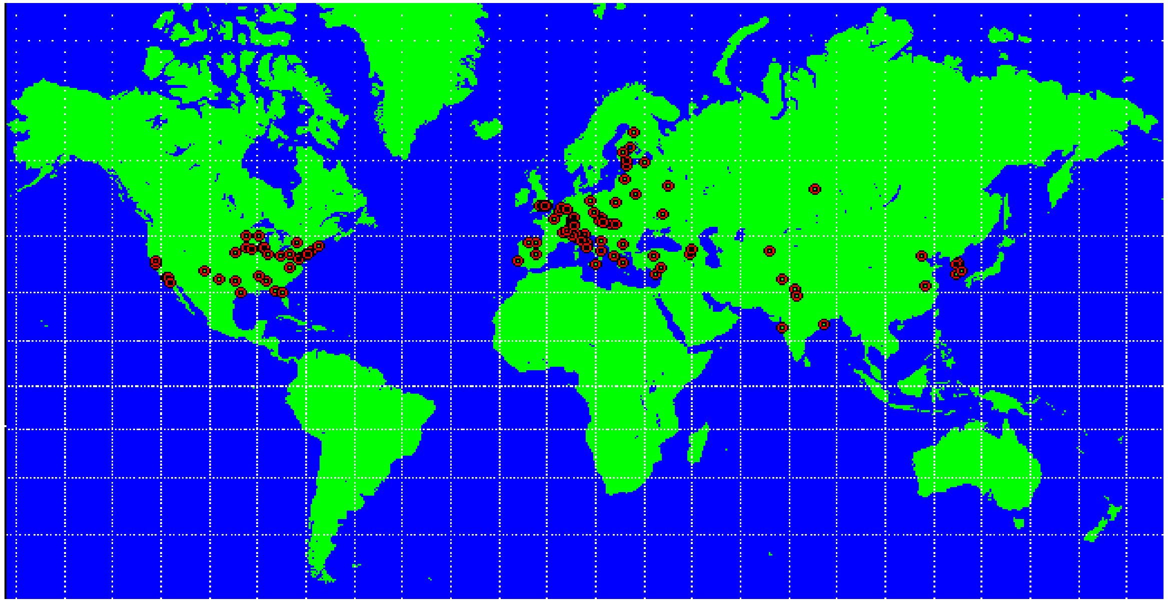


Back-up Slides

LHC Challenges: Geographical Spread

Example in CMS

~1700 **Physicists**
~150 **Institutes**
~ 32 **Countries**
(and growing)



CERN Member state **55 %**
Non Member state **45 %**

Major challenges associated with:

Communication and collaboration at a distance
Distributed computing resources
Remote software development and physics analysis

Challenges: Physics

Example: b physics in CMS

A large distributed effort already today

- ~150 physicists in CMS Heavy-flavor group
- > 20 institutions involved

Requires precise and specialized algorithms for vertex-reconstruction and particle identification

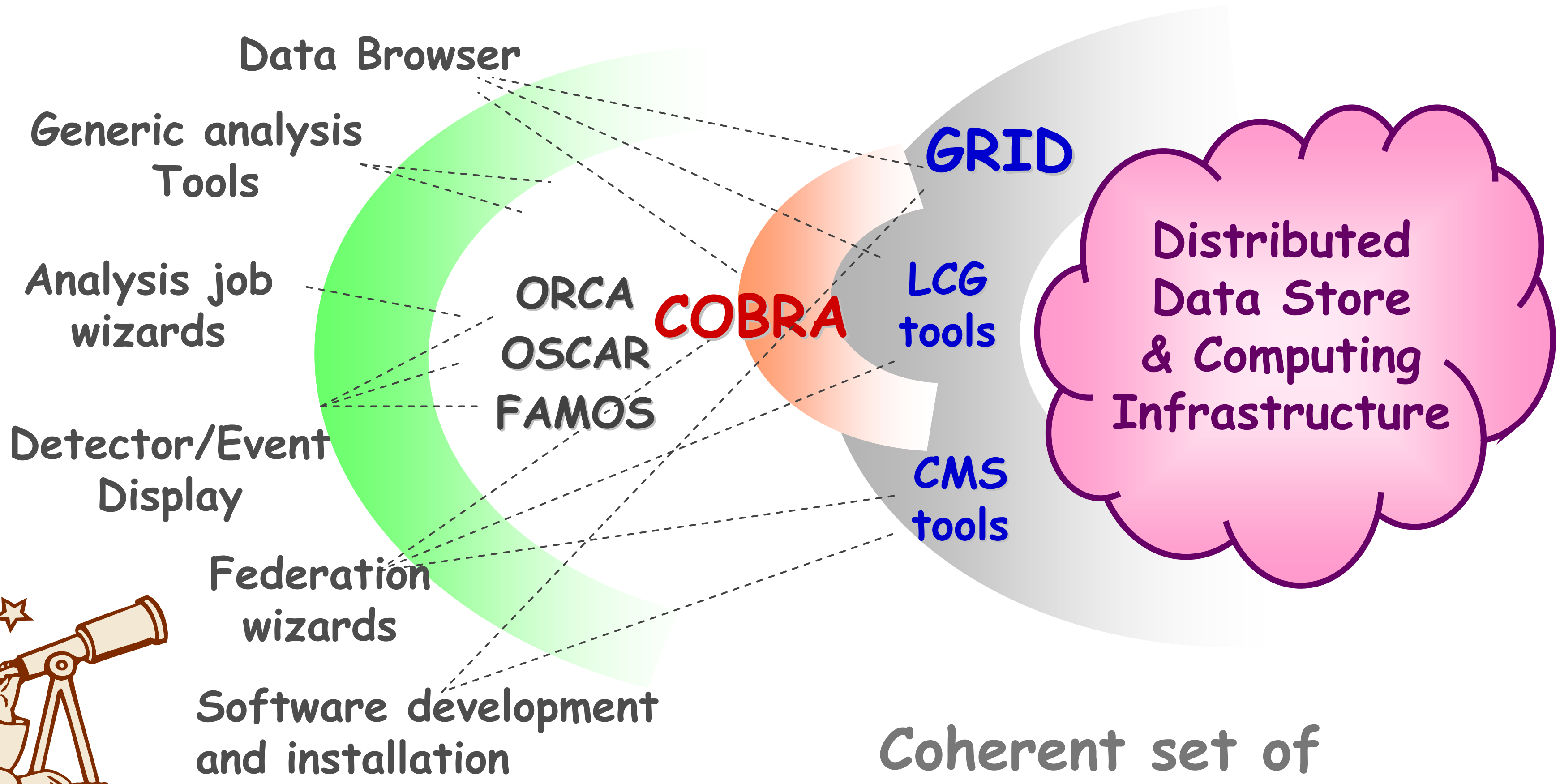
Most of CMS triggered events include B particles

- High level software triggers select exclusive channels in events triggered in hardware using inclusive conditions

Challenges:

- ***Allow remote physicists to access detailed event-information***
- ***Migrate effectively reconstruction and selection algorithms to High Level Trigger***

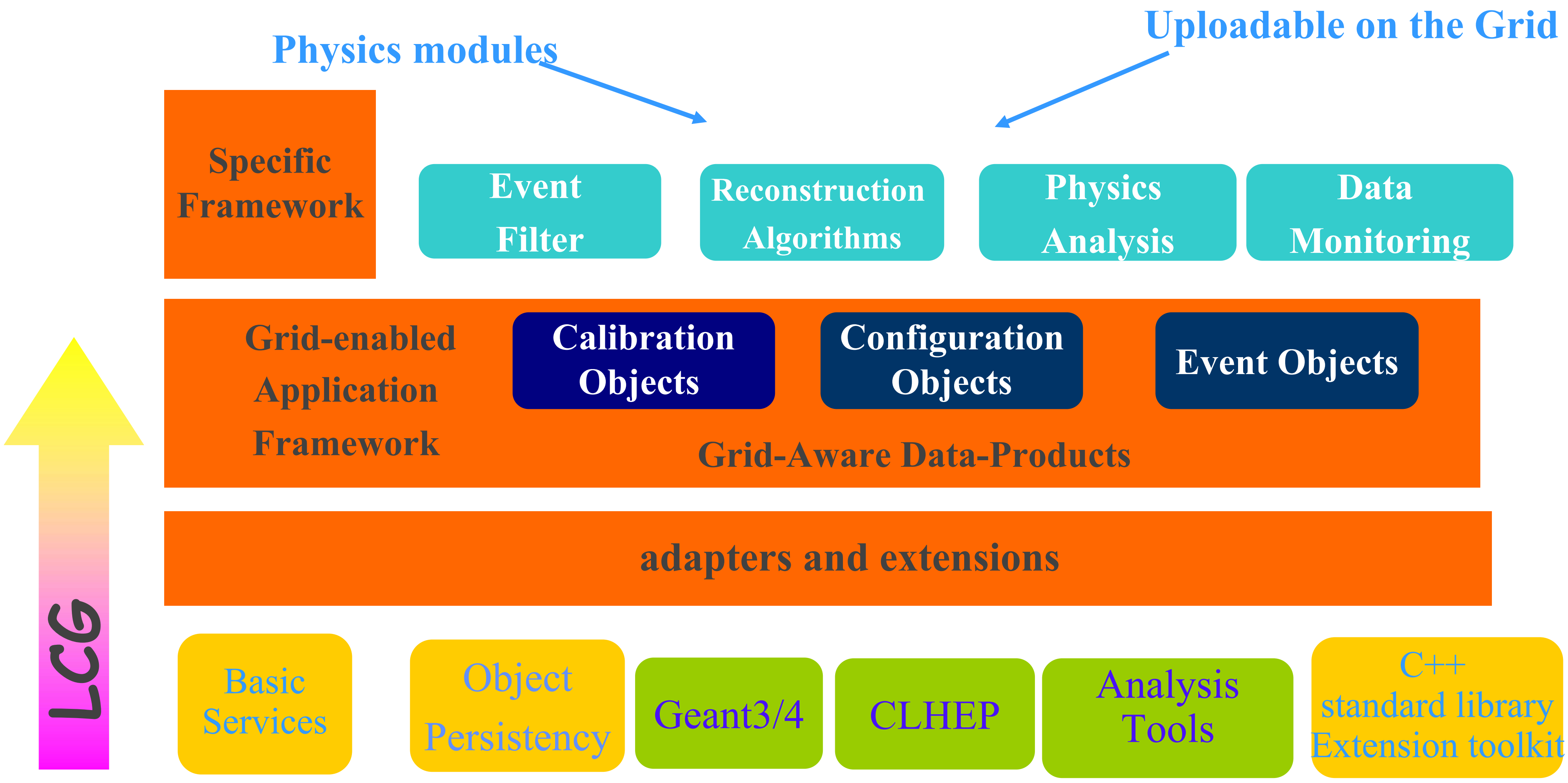
Architecture Overview



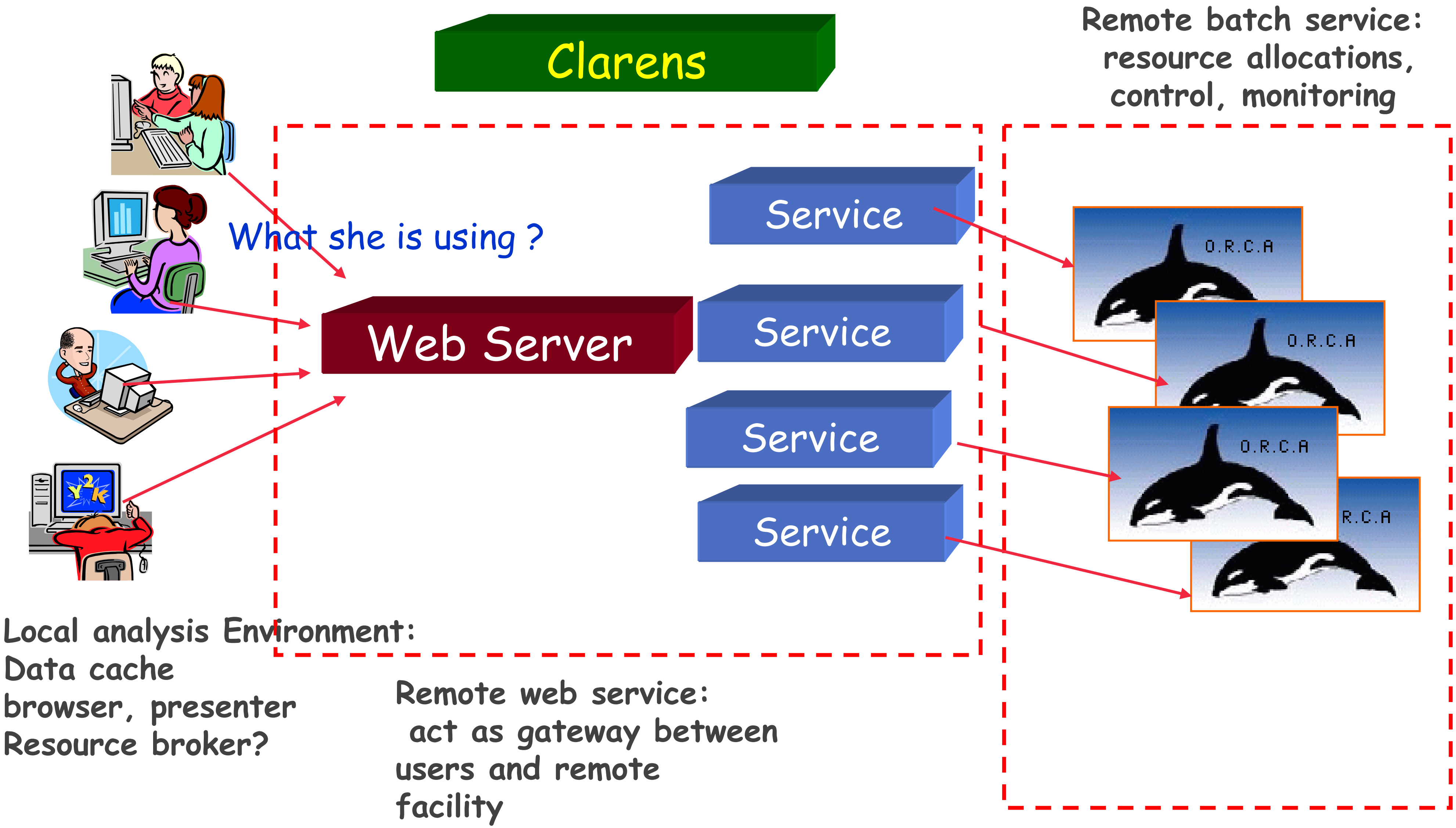
Consistent User Interface

Coherent set of basic tools and mechanisms

Simulation, Reconstruction & Analysis Software System



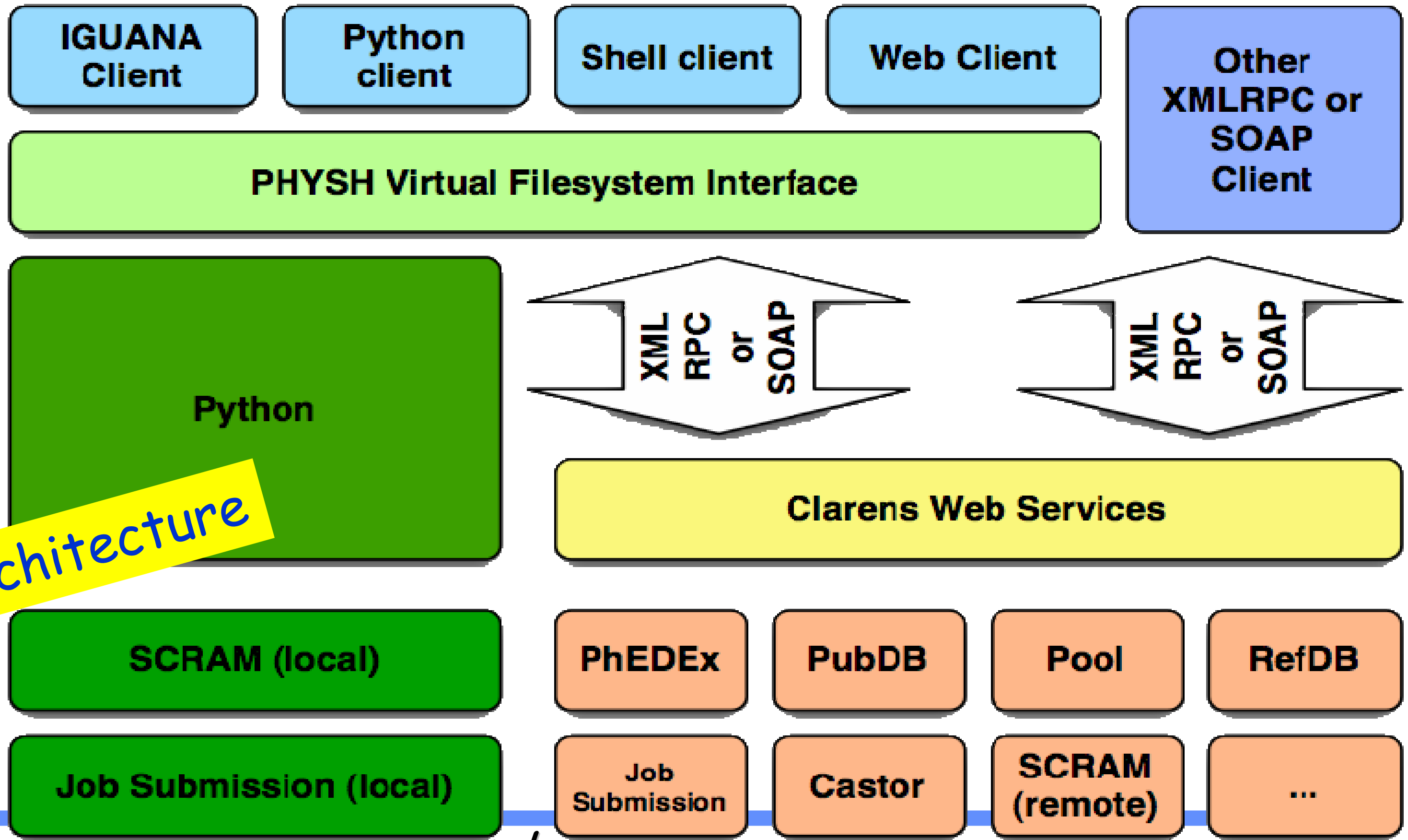
Analysis on a distributed Environment



PhySh is thought to be the end user shell for physicists.

- It is an extendible glue interface among different services (already present or to be coded).

- The user's interface is modeled as a virtual file system interface.



WebService based architecture

Interactive Analysis/Inspection/Debugging

Visualization applications for ORCA, OSCAR, test-beams (DAQ application);

Visualization of reconstructed and simulated objects: tracks, hits, digis, vertices, etc.;

Full DDD detector visualisation;

Magnetic field visualisation;

Interactive modification of configurables at run-time;

Custom tracker selection;

Event browser;

IGUANACMS Today