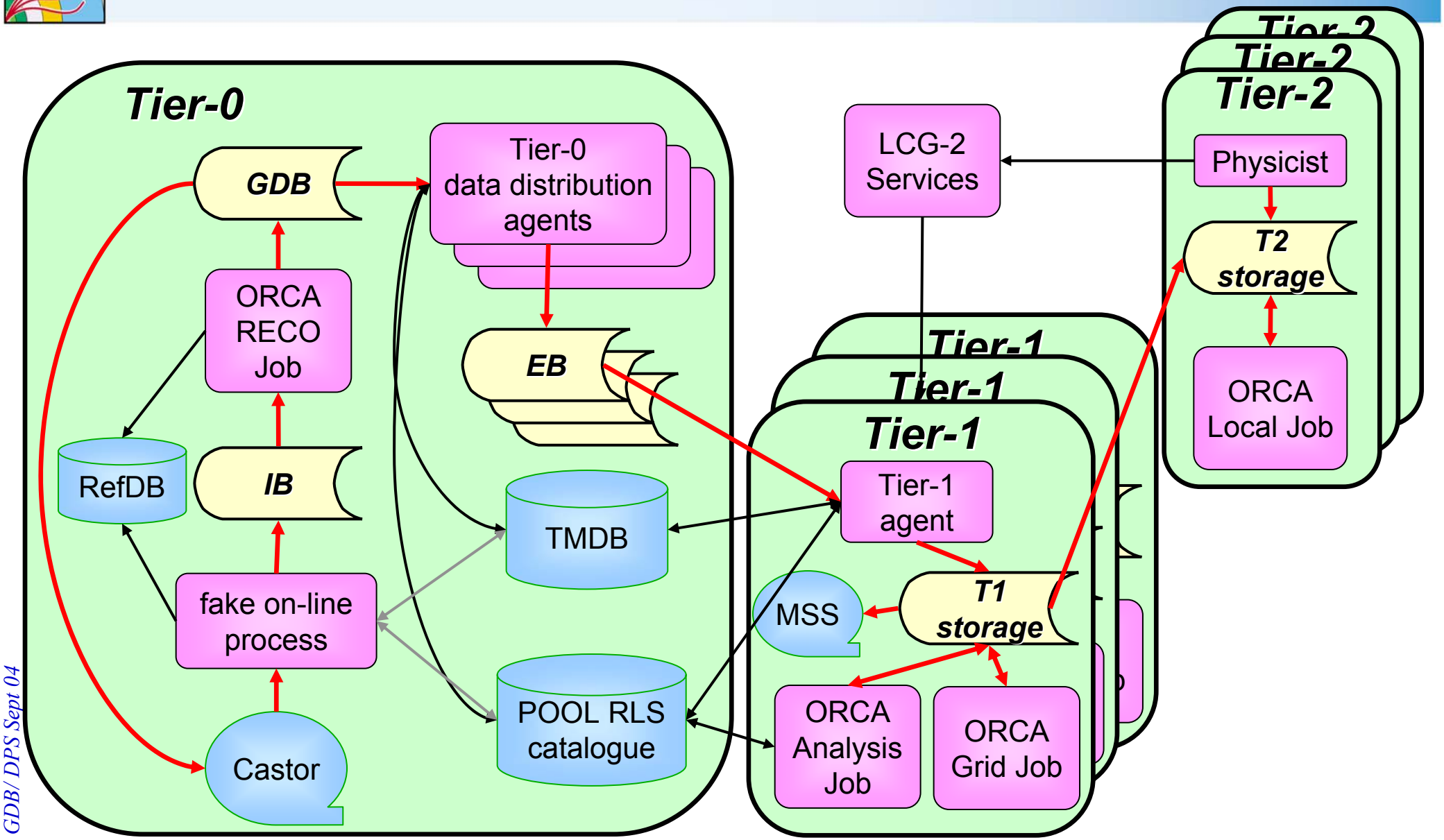
CMS Data Challenges and Operations

Report to the LCG GDB

Sept 2004

David Stickland

DC04 layout


Tier-0
GDB
Tier-0 data distribution agents
ORCA RECO Job
EB
RefDB
IB
TMDB
fake on-line process
POOL RLS catalogue
Castor
LCG-2 Services
Tier-1
Tier-1
Tier-1
Tier-1 agent
T1 storage
MSS
ORCA Analysis Job
ORCA Grid Job
Tier-2
Tier-2
Tier-2
Physicist
T2 storage
ORCA Local Job

DC04 Summary

Achieved 25Hz, but only for short periods. (Only one continuous 24 hour period)

RLS poor performance issues traced and largely corrected.

CMS TMDB over SRB and SRM worked very well. Re-engineered to PheDEx

Solved many internal CMS problems with local catalog publication etc Achieved 20min latency to T1 Analysis jobs

Not well managed in DC04. New PubDB and web tools since implemented

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

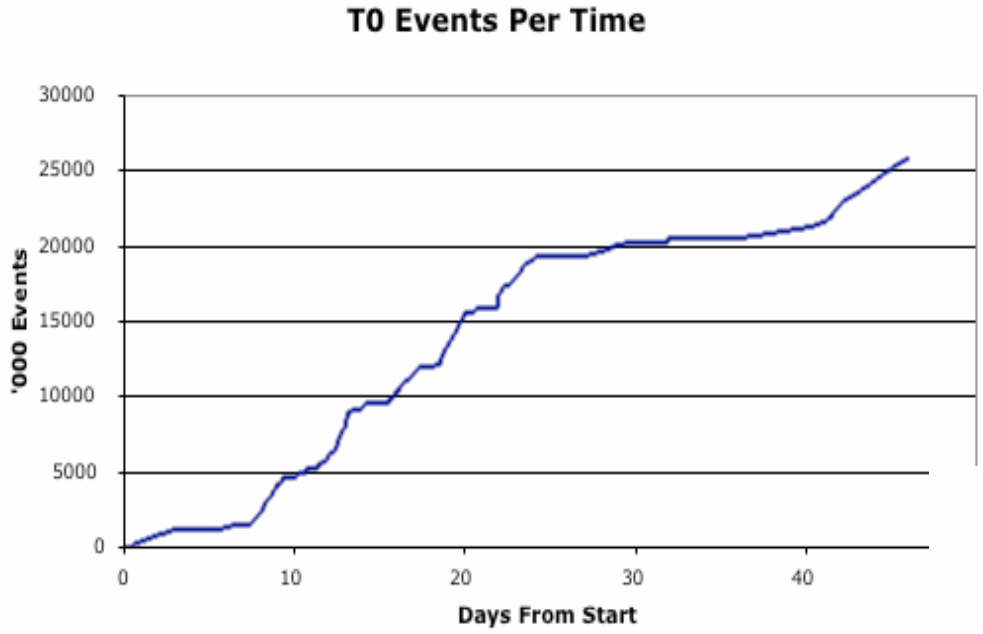
- 70M Monte Carlo events (30M with Geant-4) produced
- Classic and grid (CMS/LCG-0, LCG-1, Grid3) productions

MonaLisa very useful and flexible. GridIce implemented but not heavily used due to non-compute nature of grid challenge

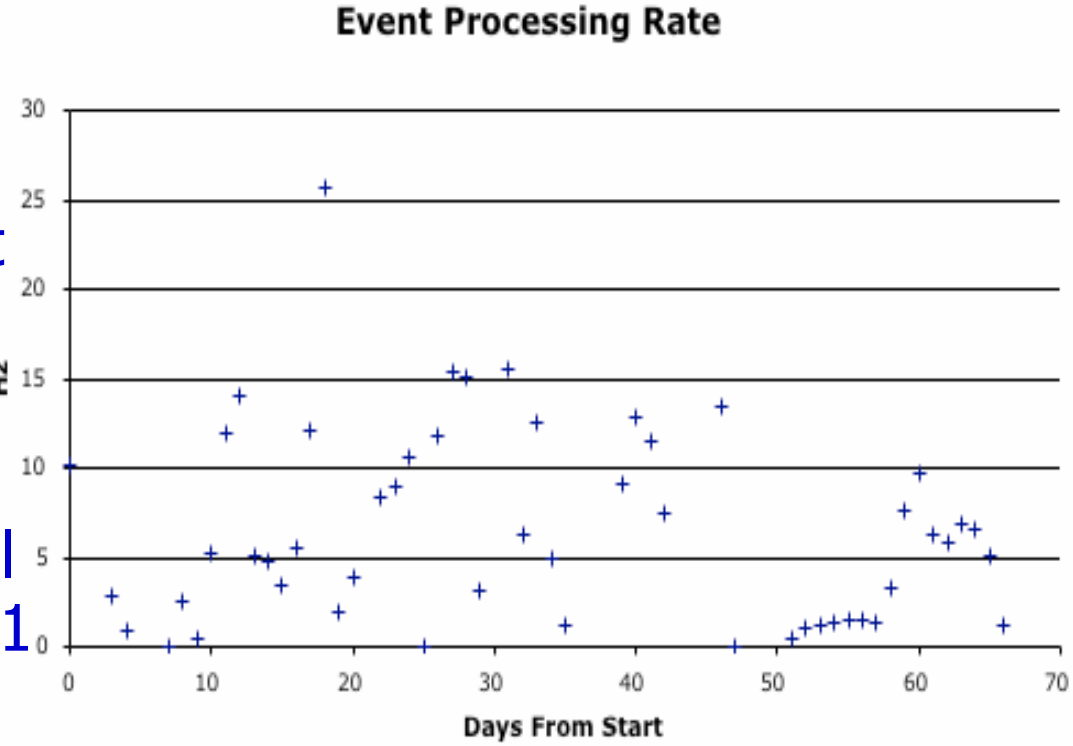
These were the only large scale "computing" grid components of DC04; but they came before LCG2 was functional

Full chain was demonstrated from Tier-0 to T1 and T2 Analysis

DC04 Processing Rate



- Processed about 30M events
  - First version of DST code was not fully useful for Physicists
  - Post-DC04 3rd version ready for production in next weeks
- Generally kept up with SRM based transfers. (FNAL, PIC, CNAF)



- Got above 25Hz on many short occasions
  - But only one full day above 25Hz with full system
- RLS, Castor, overloaded control systems, T1 Storage Elements, T1 MSS, …

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

Testing of Computing Model in DC04

- Concentrated for DC04 on the Organized, Collaboration-Managed aspects of Data Flow and Access
  - Functional DST with streams for Physics and Calibration
    - DST size ok, almost usable by "all" analyses; further development now underway
  - Tier-0 farm reconstruction
    - 500 CPU. Ran at 25Hz. Reconstruction time within estimates.
  - Tier-0 Buffer Management and Distribution to Tier-1's
    - TMDB- CMS built Agent system communicating via a Central Database.
  - Tier-1 Managed Import of Selected Data from Tier-0
    - TMDB system worked.
  - Tier-2 Managed Import of Selected Data from Tier-1
    - Meta-data based selection ok. Local Tier-1 TMDB ok.
  - Real-Time analysis access at Tier-1 and Tier-2
    - Achieved 20 minute latency from Tier 0 reconstruction to job launch at Tier-1 and Tier-2
  - Catalog Services, Replica Management
    - Significant performance problems found and being addressed
- Demonstrated that the system can work for well controlled data flow and analysis, and for a few expert users
  - Next challenge is to make this useable by average physicists and demonstrate that the performance scales acceptably

Post DC04 Actions: CMS-CCC (Computing Coordination Committee)

- Senior physicist + senior computing person from major contributors (initially CERN + T1 countries)
  - Set up by CMS CIB to help bring about needed computing
  - Expected to act - not purely advisory
- Charged to
  - Help define / acquire / secure computing resources
    - "Phase II Computing Estimates for LHC" (with LCG)
    - Work with agencies on LCG MoU's and M&O MoU Addendum
  - Define nature of T1 / T2 centres and how we use them
- Initial task: quantify committed resources and ensure CMS can make effective use of them

APROM - Cross-Project Analysis Coordination

- Mandate
  - Coordinate CMS computing and software activities for analysis from the perspective of end-user physicists
    - considering especially the analysis model and policies, coherence, correctness and ease-of-use
  - Prepare work plans, with the concurrence of PRS and CCS and track identified deliverables
    - in the areas of: analysis tools and user-focussed utilities and interfaces to underlying CCS services.
- Scope
  - Analysis model, policies, overall coherence (meta-data etc.)
  - User Data location and Processing Services
  - Analysis Utilities (COBRA, IGUANA, ROOT),
  - Liaison with LCG / ARDA
- Led by L. Silvestris: already active and generating interest

CCS Re-organisation
New Data and Workload Management Tasks

Data Management (L. Bauerdick)

- Robust engineered solutions
  - data storage, replication, access, movement, mirroring, caching etc.
- CMS Requirements & Technical Assessment Group well-advanced on defining needs and plan
  - Should finish end Sept

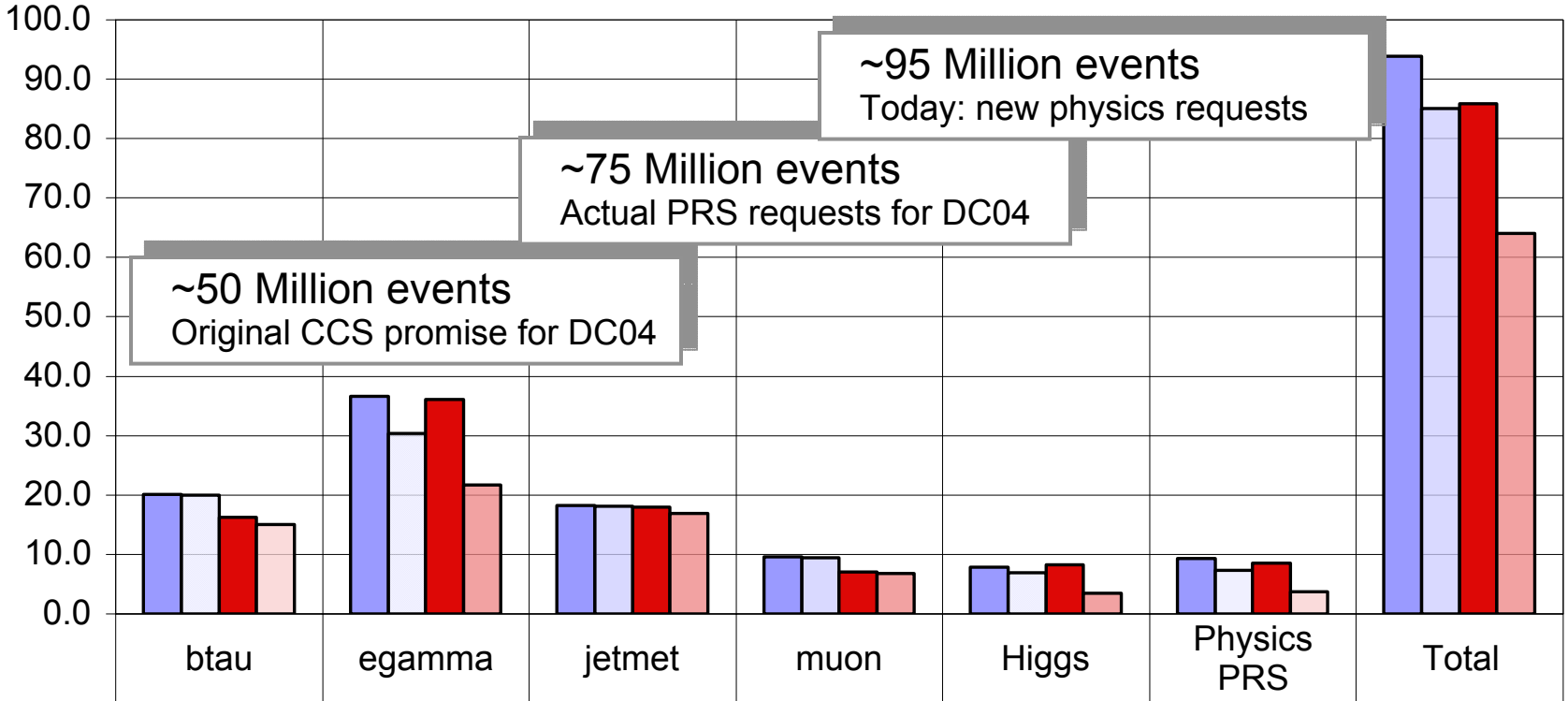
Workload Management (S.Lacaprara)

- Strong end-user (physicist) focus
  - Locate data, prepare and run jobs at regional centers
- Full work plan after DM RTAG and vacations
- Close links to APROM

1.1 Management	1.1.1 Project management 1.1.2 Liaison with non-CMS Projects
1.2 Infrastructure & Services	1.2.1 Central Computing Environment 1.2.2 User support 1.2.3 Software Process Service 1.2.4 Software Process Tools
1.3 Core Applications Software	1.3.1 Application Framework 1.3.2 Basic Toolkits 1.3.3 Development and integration of data / metadata systems 1.3.4 User Interfaces and Graphics
1.4 Production	1.4.1 Production tools 1.4.2 Production Operations
1.5 Data Management	1.5.1 Architecture and basic services 1.5.2 Integration with Production
1.6 Workflow Management	1.6.1 Architecture and basic services 1.6.2 Batch job management tools 1.6.3 User Tools 1.6.5 Validation of Grid for users 1.6.6 Integration with EGEE / ARDA
1.7 Computing TDR	1.7.1 Editorial 1.7.2 Computing Model 1.7.3 Validation of Computing Model

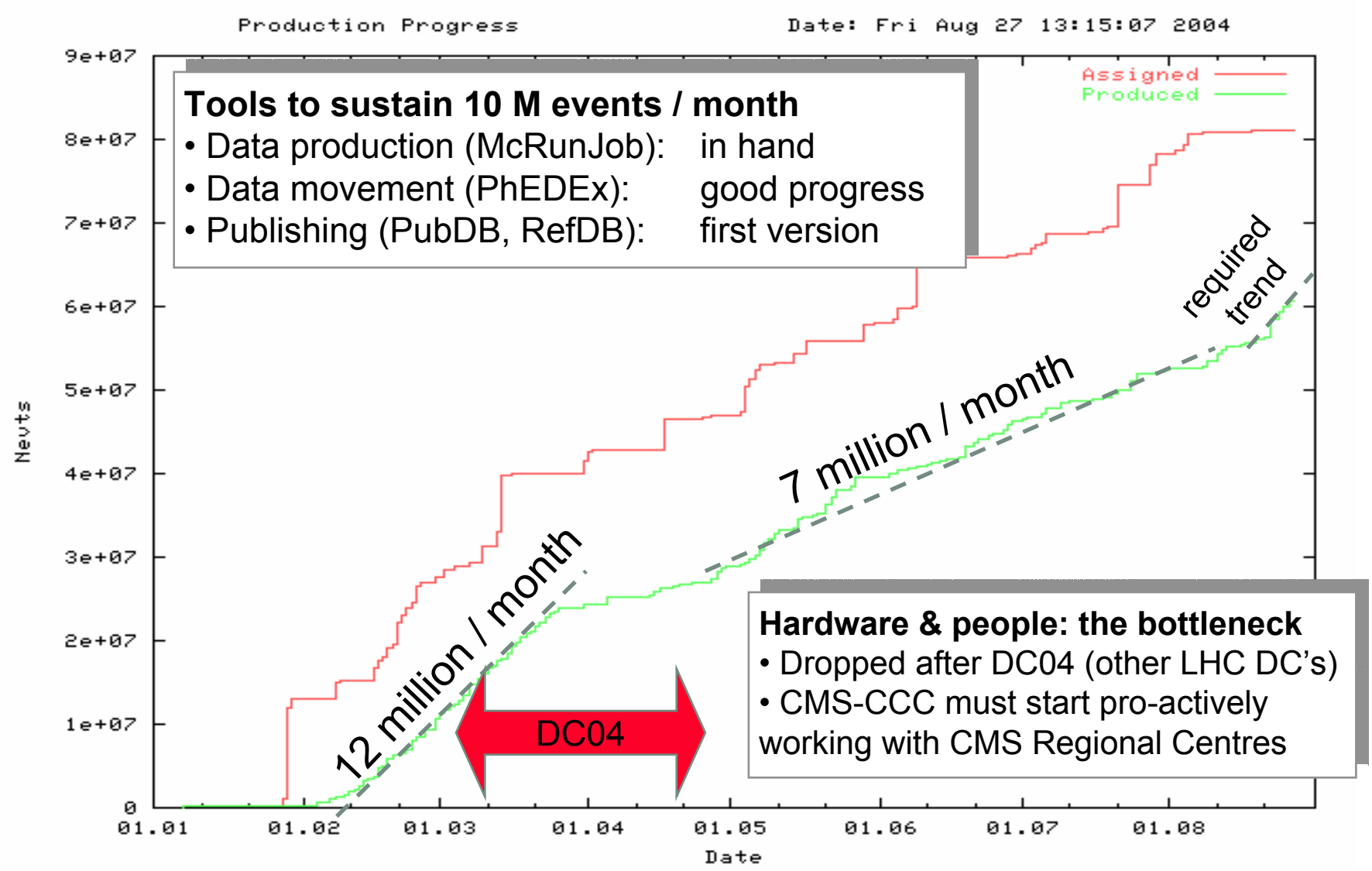
Preliminary

CMS Plan / Status
Simulation / Digitisation Production



	btau	egamma	jetmet	muon	Higgs	Physics PRS	Total
Simulation request	20.1	36.6	18.3	9.5	7.8	9.4	93.9
Simulation done	19.9	30.3	18.1	9.4	7.0	7.3	85.1
Digitisation request	16.3	36.0	18.0	7.0	8.3	8.6	85.9
Digitisation done	15.0	21.7	16.9	6.8	3.5	3.7	64.1

CMS Plan / Status
Digitisation Production

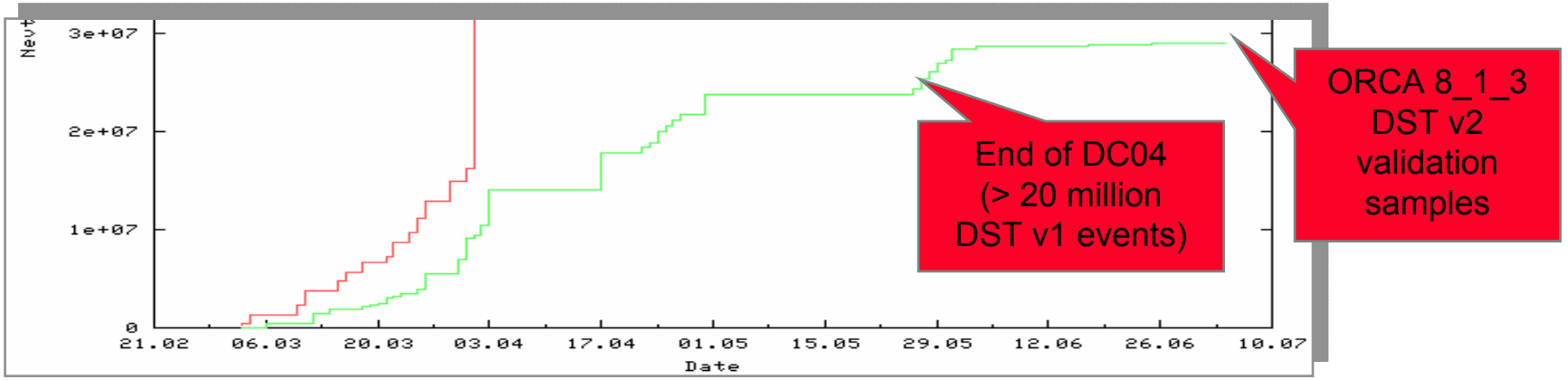


Continuous Operation

- We have been in full time production for a year now
  - Some of this on GRID resources and some on dedicated resources
- We are aggressively working on:
  - Enabling complex tasks like Digitization to run on GRID resources
  - Smoothing the production operations (continuous delivery rather than large blocks at a time;
  - End-to-end delivery;
  - Data movement;
  - Publication;
  - Enabling user jobs
- New CCS Tasks of Data Management, Workload management and Production Operations and cross-project APROM to coordinate and to do the work.
- Continuous user access to data at GRID sites is a big issue.
  - Analysis issues being implemented and tested and made to work

CMS Plan / Status
DST's : software and production

- **DST (v1)** – CCS aspects worked in DC04 but usefullness to PRS limited



- **DST (v2**) - OK for 10 million W→eν calibration sample
- **DST (v3)** - PRS have made physics objects "generally good enough"
  - ORCA 8_[3/4]_0 : have green-light to re-launch W→eν samples
    - modulo higher PRS priorities of digis and SUSY samples
  - ORCA 8_5_0 (~1 more week) with validated DST physics content
    - Then start producing PRS priority samples: O(10M) events
  - Have resolved requirements / design of e.g. configuration tracking
  - Will be used for Reconstruction production this Autumn and first PTDR Analyses

Physics TDR

- Physics TDR scheduled for December 2005
  - Not a "yellow report", but a detailed study of methods for initial data taking and detector issues such as calibration as well as physics reach studies.
- Current Simulated samples more or less adequate for low luminosity
  - About to enter re-reconstruction phase with new DST version
  - Estimate similar sample sizes for high luminosity
- Target 10M events/month throughput
  - Generation, Simulation, Digitization, Reconstruction, Available for analysis
  - New production operations group in CMS to handle this major workload
- In light of DC04 experience, DC05 is cancelled as a formal computing exercise
  - Not possible to serve current physics requirements with data challenge environment
  - However, specialized component challenges are foreseen

Continuous Operation Scale and Scope

- Instead of DC05 we need to progressively bring up a full time operation, including data access
  - Not a test or a challenge.
  - Physicist access to data required.
- Generic GRID resources (for Generation/Simulation)
  - ~750 CPU Continuous
    - CPU means current generation ~2.4+GHz
- CMS T1 resources (Grid with significant Disk and MSS)
  - Needed for data intensive Digitization and Reconstruction steps
  - ~750 CPU Continuous
  - Now 60TB +20TB/month (Spread across T1 centers)
- T1/T2 resources (probably not generic)
  - ~150 CPU Continuous
  - Now 40TB Analysis disk space to grow by about 2-5TB/month (T1+T2)
- We intend to run all this within LCG
  - "LCG" being all those resources available to CMS being steady migration from/between LCG2, GRID3, gLite, …
- We need to reach the 10M event/month level soon (Autumn)
- We need to make the resources available to a growing CMS user base in the same time scale