DIRAC
COMMUNITY GRID SOLUTION ™

The LHCb Computing Model and Real Data

Philippe Charpentier

CERN – LHCb

On behalf of the LHCb Computing Group

CHEP
TAIPEI TAIWAN
2010

LHCb Data Characteristics

- Data size and rates are modest compared to other LHC experiments
  - 35 kB RAW event size
  - Trigger rate: 2000 events/s
  - 25 kB RDST (a.k.a. ESD), 85 kB DST (a.k.a. AOD)
  - Typical reconstruction time: 12 HS06.s/event
- Physics research channels are rare
  - b-quark CP violation decay modes (BR ~ 10^{-9} to 10^{-6})
  - Typically a few 10'000s to a million events per year (2 fb^{-1})
    - A needle in a haystack
  - Easier to extract b decay events if only one primary vertex
  - Metrics = average number of visible interactions per beam crossing (μ)
    - For LHC design characteristics μ=0.4 at LHCb
- LHCb is a small experiment
  - Very small Computing Operations Team (< 5FTE)

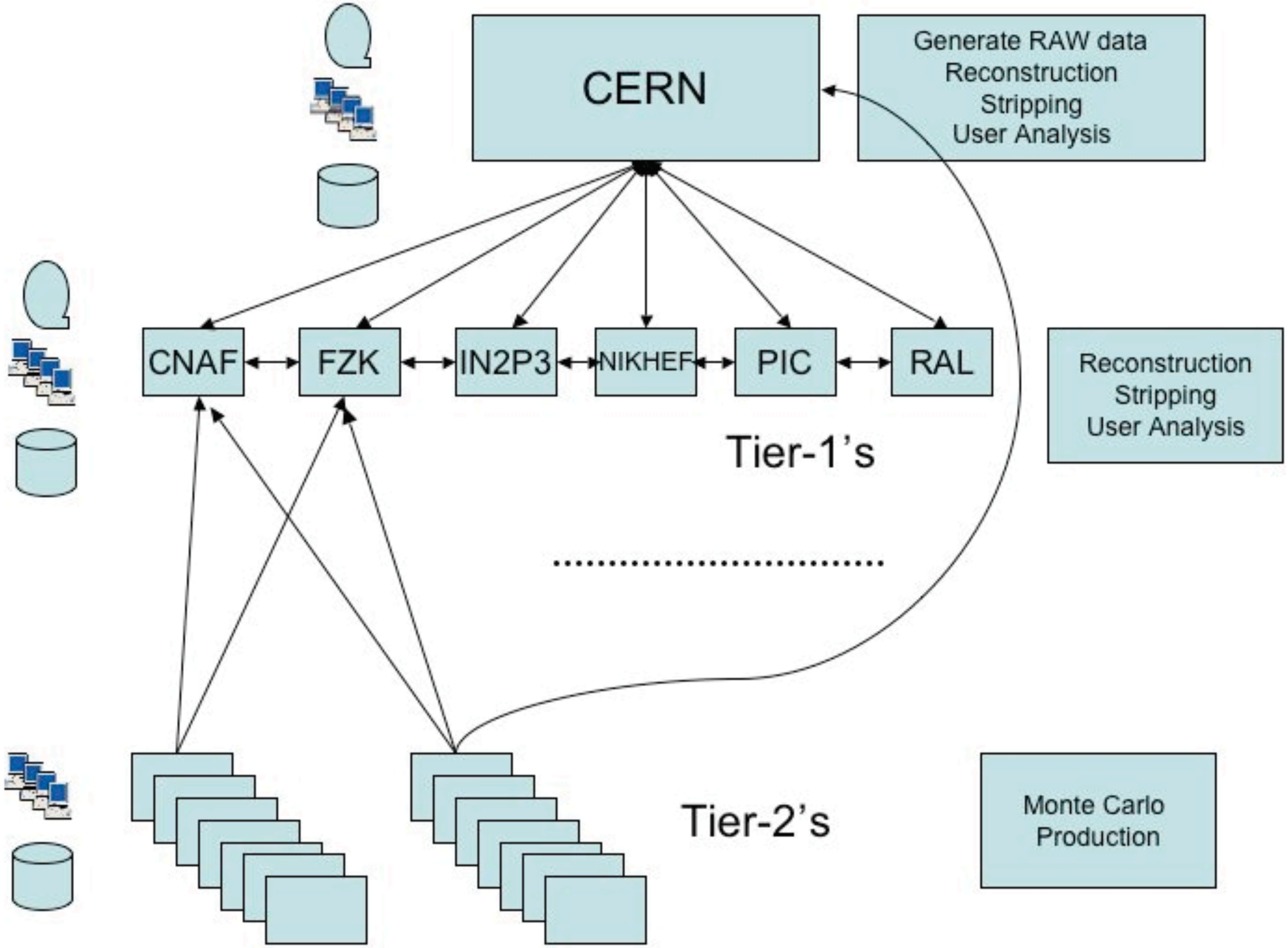
Guidelines for the Computing Model

- Small processing time, but high trigger rate
  - 24 kHS06 required for reconstruction
    - Typically 2000 CPU slots
  - Tier0 could not provide the necessary CPU power
  - Use Tier1s as well for reconstruction (first pass)

- Most problems for analysis jobs are related to Data Management
  - SE accessibility, scalability, reliability…
  - Restrict the number of sites with data access
  - Use Tier1s for analysis

- High requirements on simulated data
  - Background identification, efficiency estimation for signal
  - Typically 360 HS06.s per event
  - Use all possible non-Tier1 resources for simulation

The LHCb Computing Model



LHCb real data in 2010

- LHC started with very low luminosity
  - Very few colliding bunches
  - Not worth for rare b-physics decays
    - Minimum bias trigger for 2 months
    - Introduce tighter triggers when luminosity increases
- LHC change of strategy for higher luminosity
  - Large number of protons per bunch
  - Small squeezing
  - Still low number of bunches (16, 25, 48, increasing since September, up to 400 bunches)
  - Consequence: larger number of collisions per crossing
    - μ=1 to 2.3 !!!
    - Much higher pile-up (1.6 to 2.3 collisions per trigger)
  - Effects on Computing
    - Larger events
    - More complex events to reconstruct
    - Larger pre-selection retention

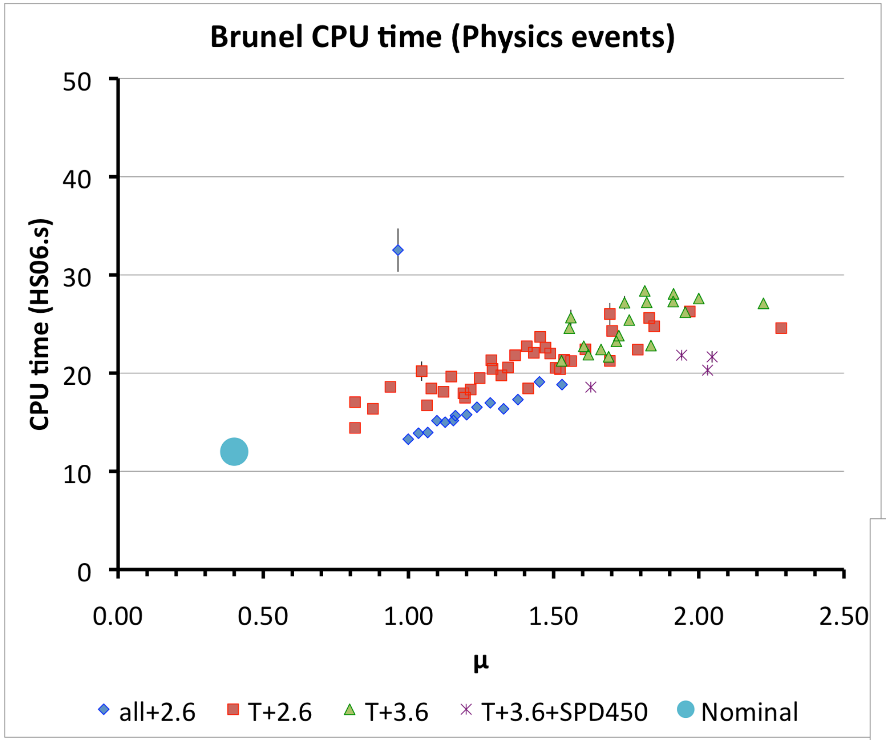
Adaptability of the Computing Model

- Needs to be reactive to continuously changing conditions
- First months: minimum bias data
  - No preselection
    - Reconstruction creating DSTs for all events
- As of July: large μ data
  - Event size
    - 50 to 60 kB
    - Twice more than design
  - Reconstruction time
    - Quadratic with event size
    - 4 times more than design
  - Stripping time and memory
    - Large combinatorics for pre-selection
    - Stripping time exponential with event size
    - Algorithms tuned for μ=0.4 were taking up to 60 HS06.s
      - Twice the reconstruction time!
      - Memory consumption up to 3 GB (nominally 1.5 GB)

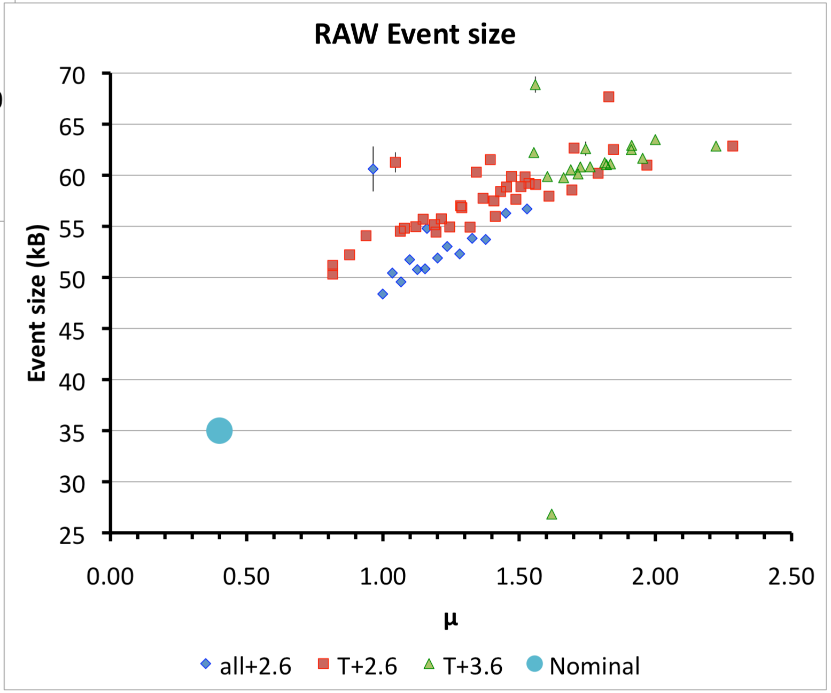
Adaptability (2)

- Reconstruction / stripping jobs
  - Need to fit in Tier1 Grid queues
    - Reduce file size (nominally 3 GB) to 1 GB
  - Extensive work on reducing computing time
    - Reconstruction: factor 2 reduction
    - Stripping: factor 10 reduction in time, large increase in rejection
- Nevertheless, this takes… time!
- For optimisation, data is needed
  - Run with existing applications
    - High failure rate (CPU time limit, max memory exceeded)
  - Use a lot of space for storing too many (too large) events
  - Possible thanks to the available disk (foreseen for more data)
  - Continuous data management operations
    - Remove obsolete processings (keep only 2)
    - Reduce number of replicas (from 7 to 3 or 4)

New Computing Conditions



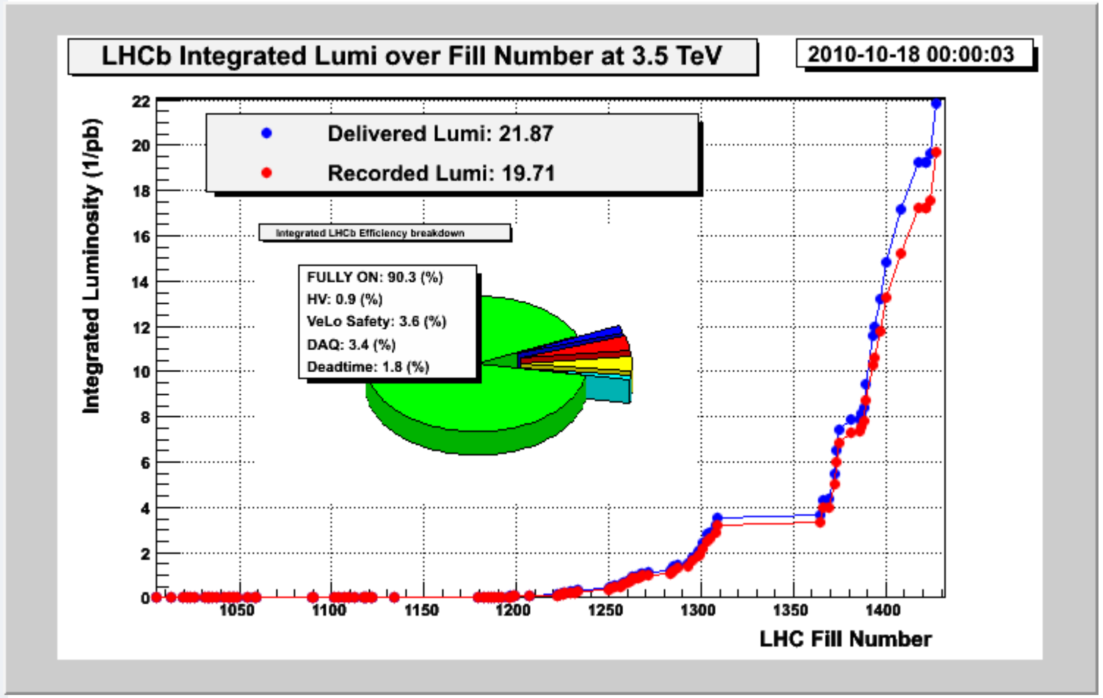
- Both event size and CPU time rise with μ
- Compatible with expectations at μ=0.4



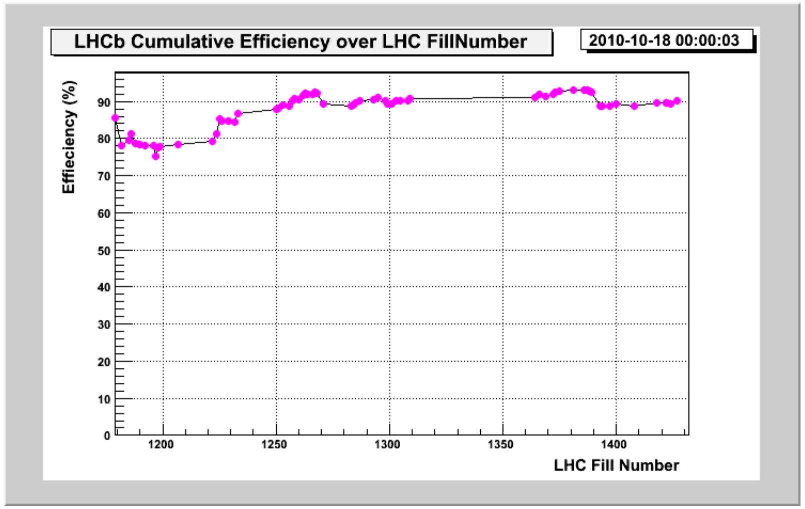
Weaknesses of the Grid

- Workload Management System
  - Mitigated by usage of pilot jobs (DIRAC)
  - Workload optimisation using generic pilot jobs
    - Run multiple payloads (e.g. production + analysis)
- Data Management System
  - Data access by protocol unreliable for long jobs
    - Errors when opening files (servers overloaded)
    - Connections broken when job lasts hours
  - Use as few files as possible, i.e. as large as possible
    - Requires merging of output files (DSTs)
    - Keep runs (1 hour data taking) as granularity of datasets
  - Mitigated by local copy of input data
    - Standard procedure for reconstruction-stripping jobs and merging
    - Not possible for analysis jobs though…

Data collection

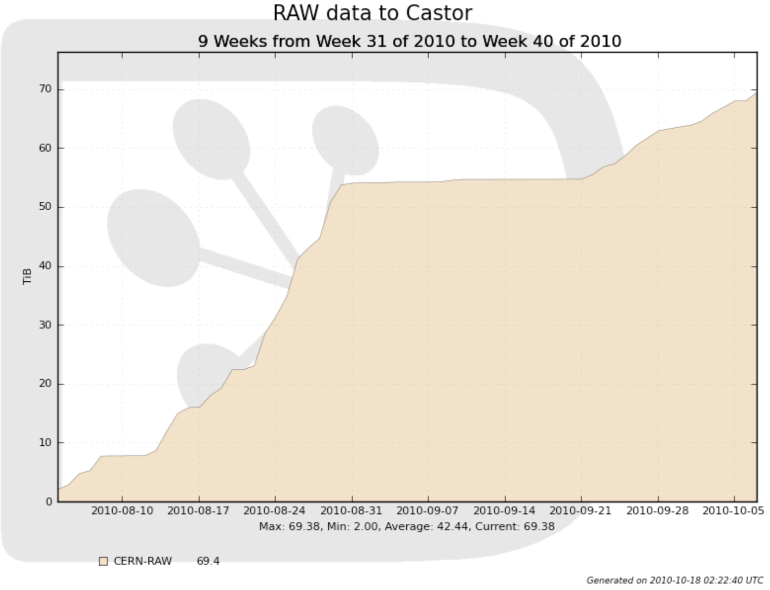


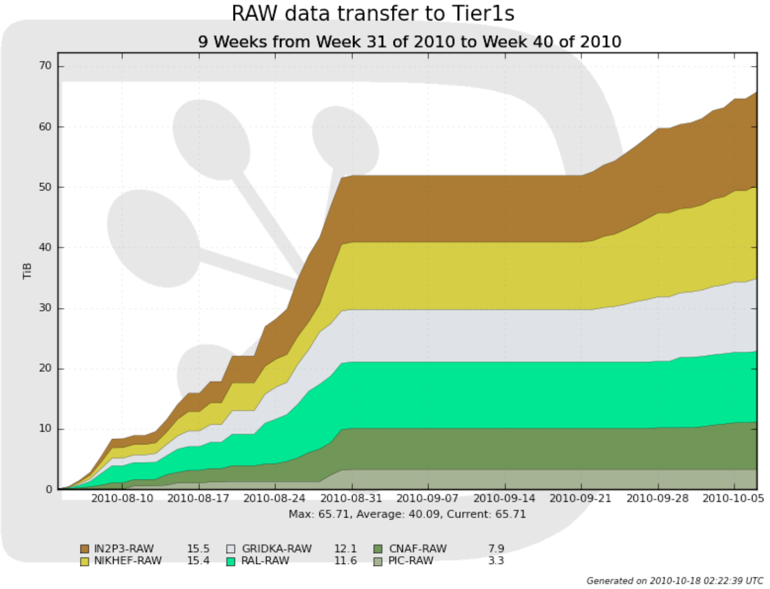
- 21.9 pb^{-1} delivered
- 19.7 pb^{-1} recorded
- 91.2% data taking efficiency
- Most data collected after 15 September



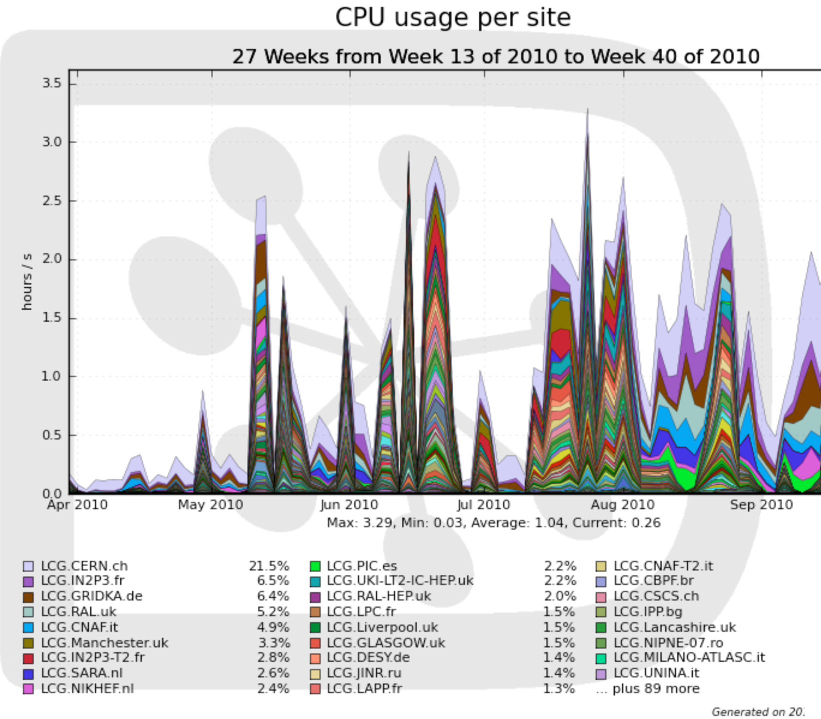
RAW data distribution

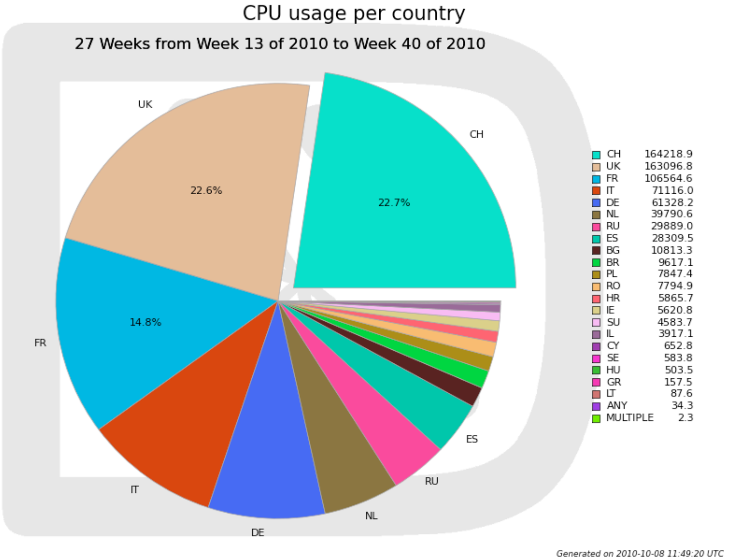
- 65.7 TB of physics RAW data collected
  - Slightly more transferred to Castor
    - Calibration and test data
    - Not distributed to Tier1s
- Distributed immediately to Tier1s
  - A full run (1 hour) goes to a single Tier1
  - RAW data share according to CPU pledges of Tier1s
    - When a Tier1 is unavailable, share temporarily set to 0

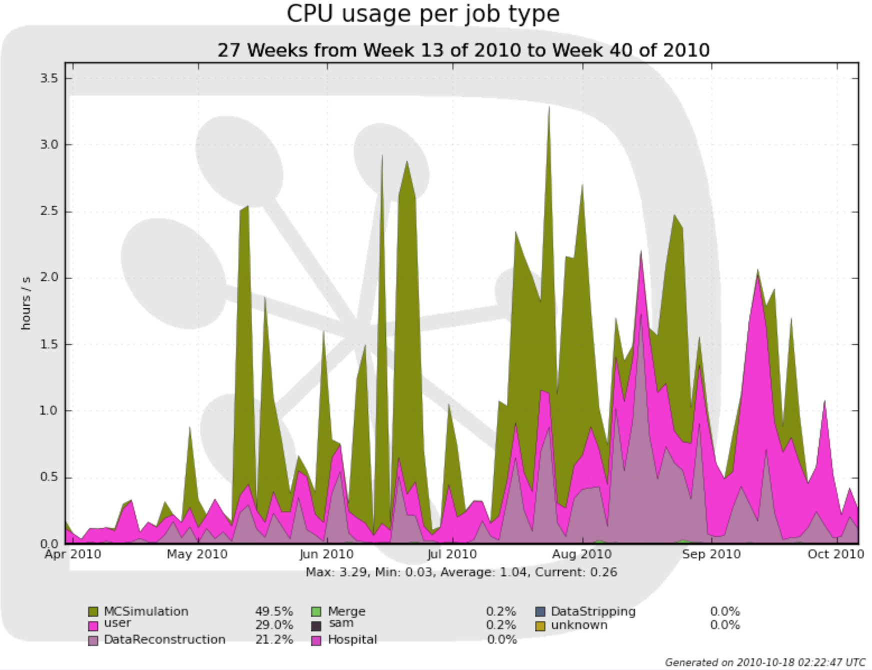




Global Grid usage



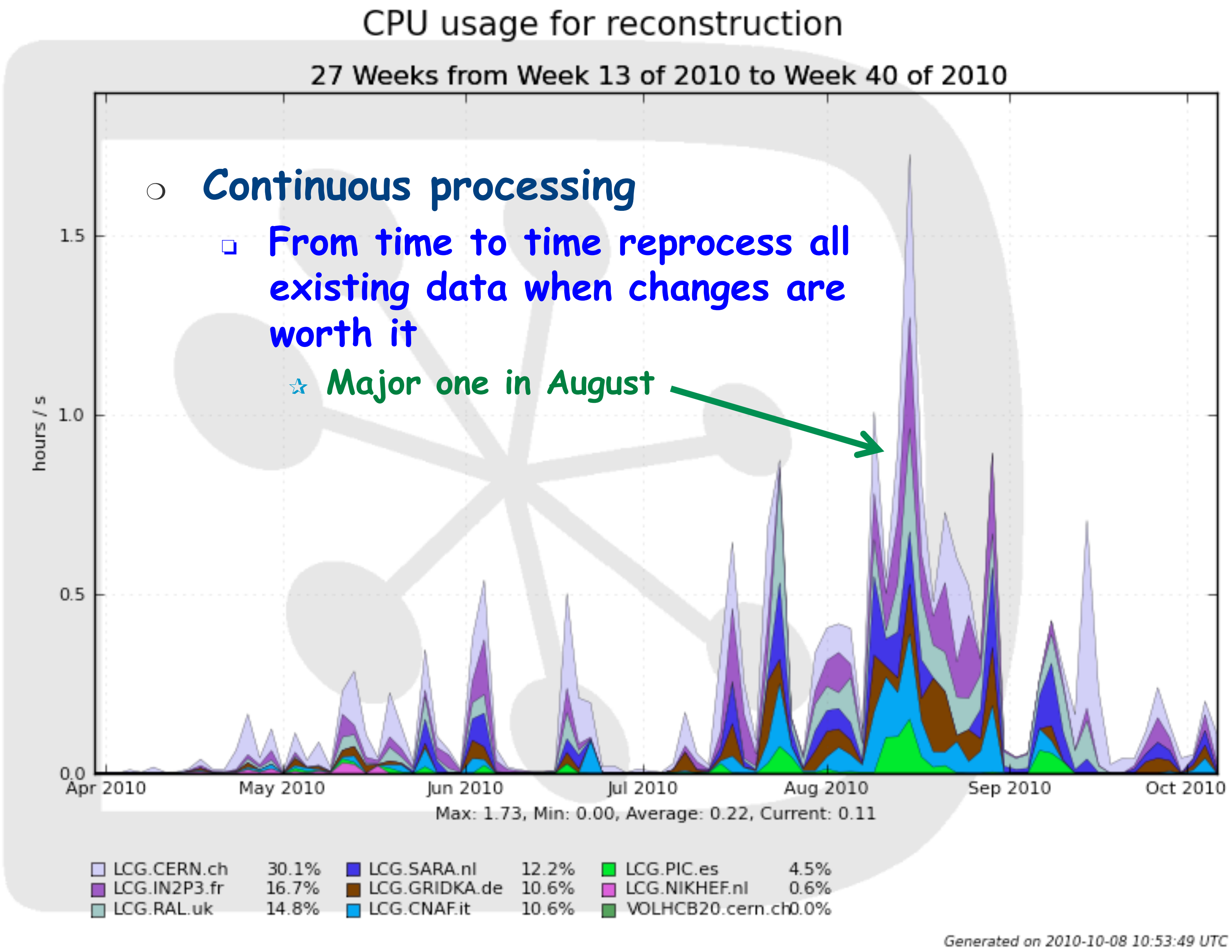




- 115 sites used
  - 21 countries

- Simulation: 50%
- Analysis: 29%
- Reconstruction: 21%

Reconstruction jobs

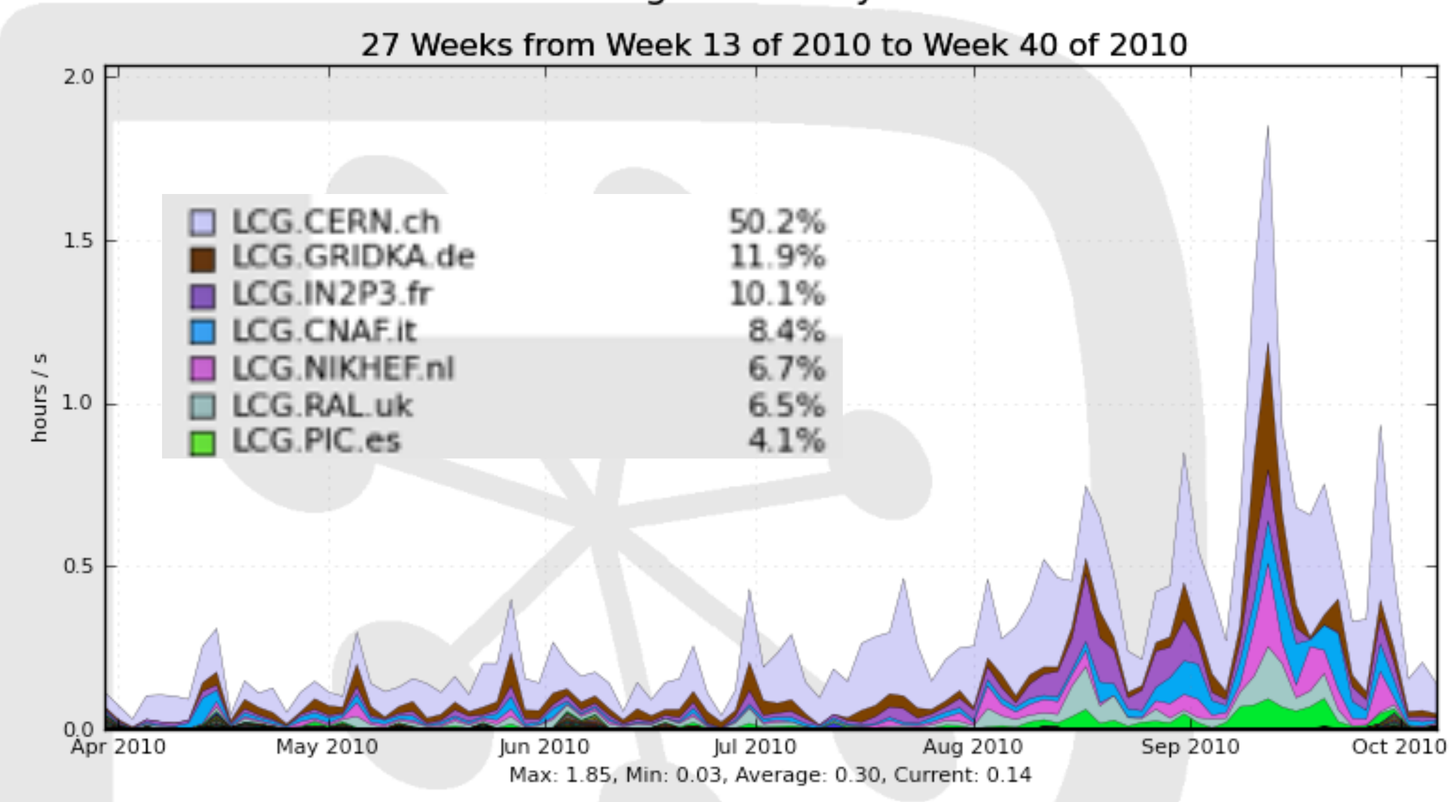


- Continuous processing
  - From time to time reprocess all existing data when changes are worth it
    - Major one in August

Analysis jobs

- Over 250 users used the Grid for analysis
  - Only 2% of analysis at Tier2s (toy MC, private small simulations)
- No a-priori assignment of site: share by availability of resources and data

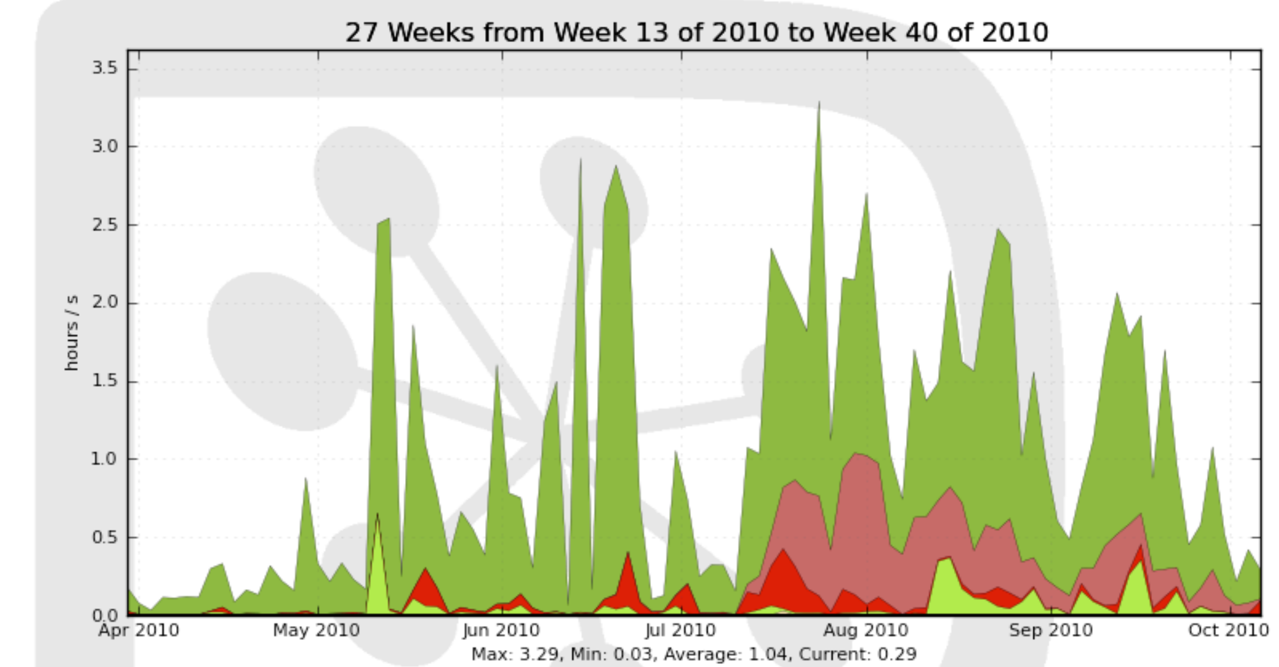
CPU usage for analysis



Job success

- Overall 81% successful jobs
- Main cause of failures (15%): job exceeding CPU time limit
  - Infinite loop in Geant4 on 64-bit
  - Large μ
    - Jobs eventually all completed after several retries!
  - Also few user jobs
- 4% data access problem in application

Job failures



Further adaptations of the Computing Model (1)

- LHCb Analysis Centers
  - Foreseen in Computing TDR: use large Tier2s for Analysis
  - Request from sites/countries to run analysis in Tier2s
  - Conditions
    - Additional CPU and storage resources w.r.t. pledges
    - Local management team (data placement, user support)
    - Open to the whole LHCb VO
      - No “local” or “national” Grid Computing
      - Local analysis done on Tier3s (local job submission, possible Grid storage), desktops, laptops
  - Main caveat
    - Data access is the weakness of the Grid
    - Analysis jobs must use protocol access (rootd, gsidcap, xrootd…)
      - Possibility to include complex local caching in the framework
      - See D.Remenska’s presentation
    - Currently a few sites are under test

Adaptations (2)

- LHCb Reconstruction Centers
  - Recent idea, not yet experimented
  - Keep analysis at Tier1s
    - Mitigate data access problems
  - Move data processing to some Tier2s
    - Anyway using local copy of data
      - Copy from close SE (same site) of not too far SE (close Tier1)
      - Requires good network connectivity from Tier1
        - Avoid CPU inefficiency
    - Use well controlled workflows at Tier2s
      - Simulation
      - Reconstruction / stripping
    - Merging at Tier1
      - Keep entire run at a single Tier1
  - Plan to experiment Reconstruction at Tier2s during winter shutdown

Conclusions

- The LHCb Computing Model looks global sound
- However the new LHC running conditions imply some changes to the offline reconstruction and analysis conditions
- During 2010, several iterations were needed in order to adapt to these conditions
- The full reprocessing of 2010 data will take place starting in November 2010
- Increase in CPU requirements and disk space will have to be watched carefully in order to match the pledges
- Usage or resources beyond Tier1s for reconstruction and analysis are being investigated