LHCb Computing activities

Marco Cattaneo

CERN – LHCb

On behalf of the LHCb Computing Group

(update of talks by Philippe Charpentier in March 2011)

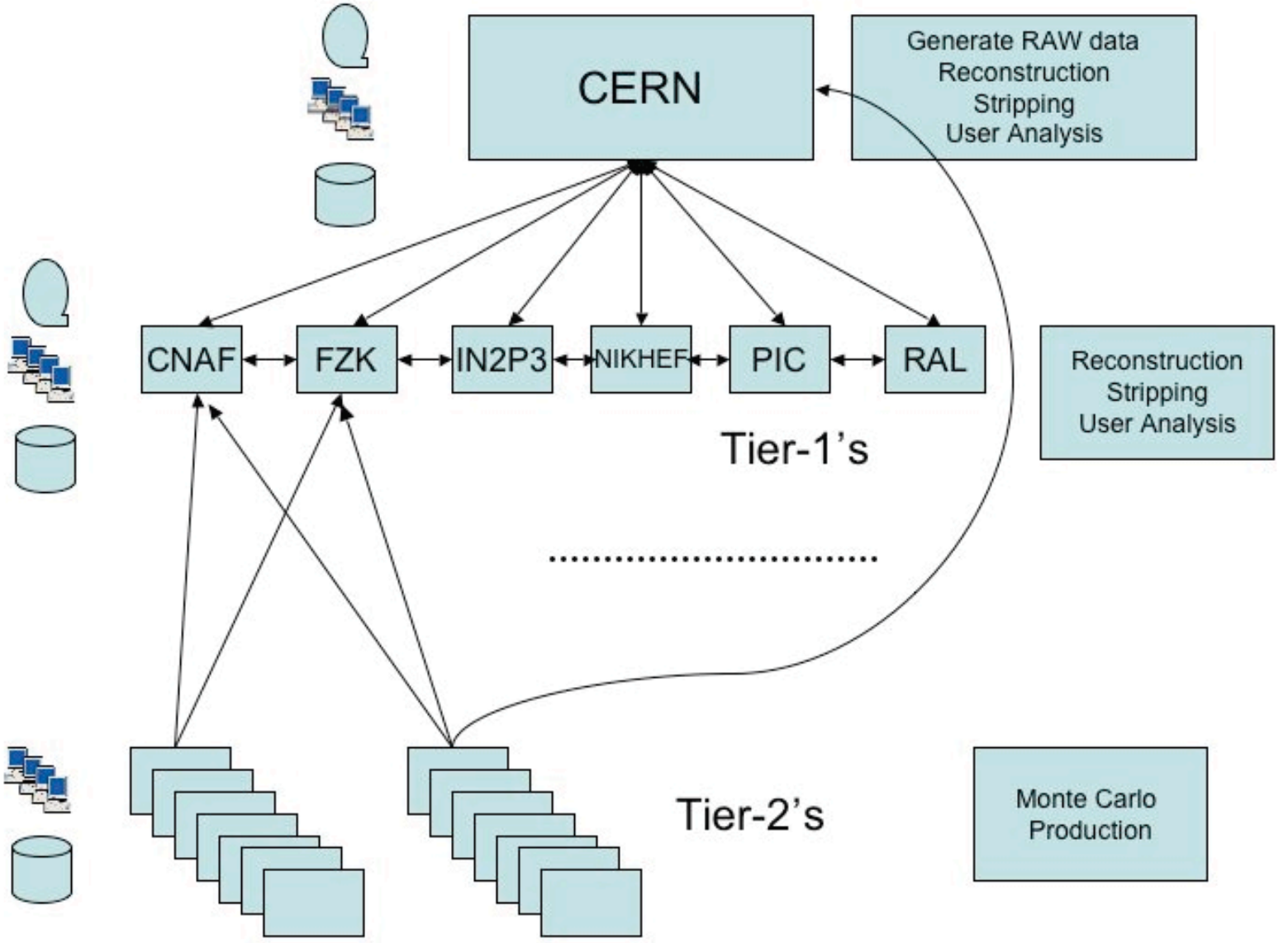
Summary of Computing activities

- Simulation
  - Mainly used for identifying background and evaluating acceptances and efficiencies
  - Simulates an ideal detector, however with realistic geometry
  - Event generation and detector response tuned to real data
    - Iterative process, depends on the data taking year
- Real data handling and processing
  - Distribution to Tier1s (RAW)
  - Reconstruction (SDST)
  - Stripping and streaming (DST+μDST)
  - Group-level productions (DST+μDST)
- User analysis
  - MC and real data processing
  - Detector and efficiency calibration
  - End-user analysis (usually off-Grid: Tier3 or desktop)

Guidelines for the Computing Model

- Small processing time, but high trigger rate
  - 30 kHS06 required for reconstruction
    - Typically 2500 CPU slots
  - Tier0 could not provide the necessary CPU power
  - Use Tier1s as well for reconstruction (first pass)
    - Has been extended to Tier2 for reprocessing
- 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 1700 HS06.s per event
  - Use all possible resources for simulation
    - Priority given to Tier2, but used also to smooth usage at Tier1

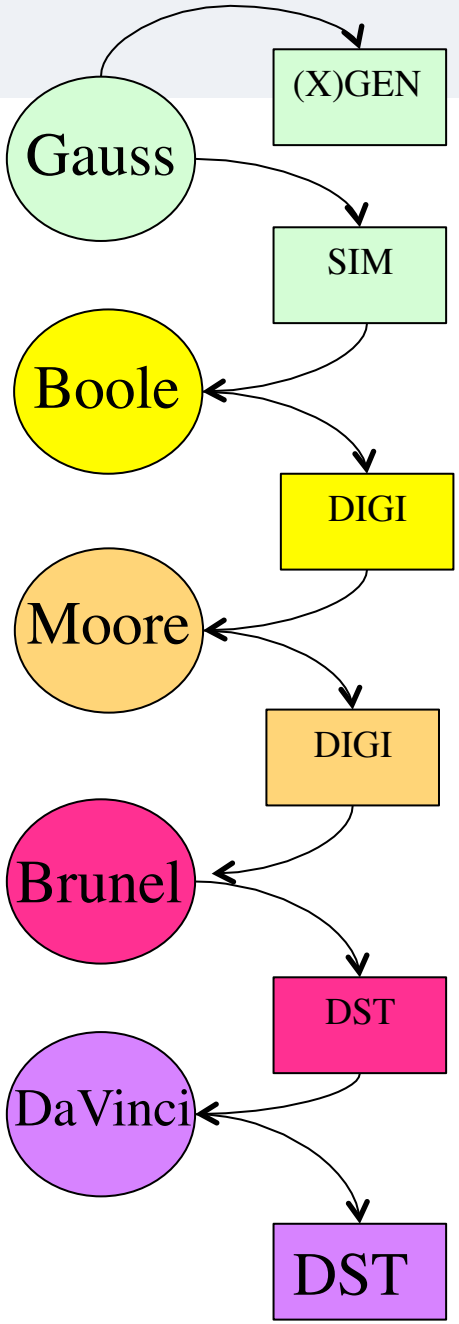
The (original) LHCb Computing Model



Software distribution

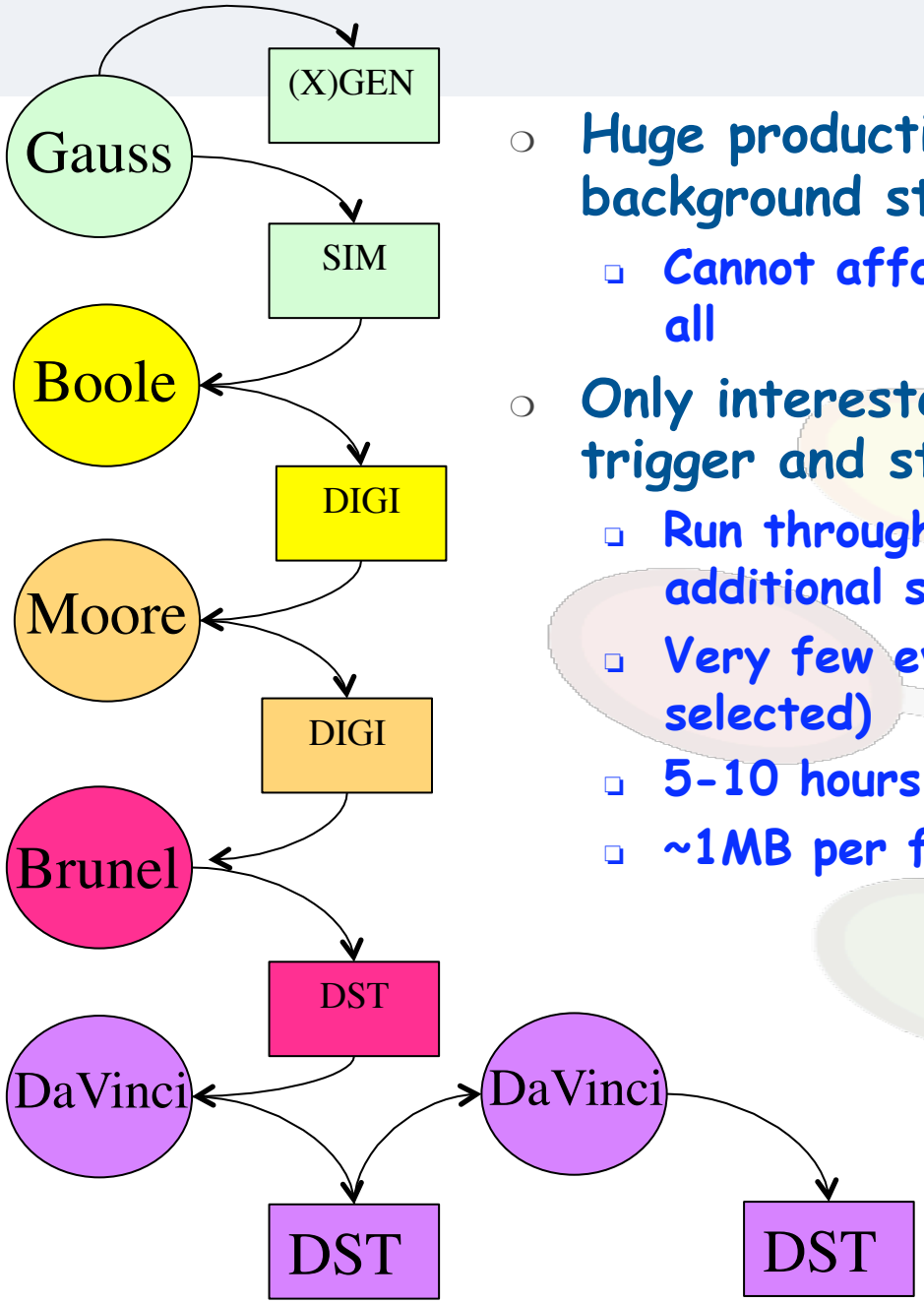
- Applications software distribution:
  - CVMFS (new)
    - Software installed (and eventually removed!) in cvmfs as part of software release procedure
    - LHCb was early adopter, extremely happy with it
      - stability, scalability, availability
  - Installation of tar balls in shared areas
    - Automated by SAM jobs
      - Exception: CCIN2P3 required manual installation in AFS
    - Availability of shared areas is one main cause of job failures
    - Software removal not obvious
  - LHCb would like CVMFS everywhere
    - But tar ball possibility will remain
- Conditions database:
  - Real time replication of new conditions via Oracle streams
    - Needed only when running on new data
    - Looking at FronTier
  - SQLite database regularly distributed as a software package
    - No Oracle access needed when running on older data, e.g. reprocessing, or simulation

Simulation jobs



- 5 steps jobs
- Gauss: simulation, based on Geant4
- Boole: digitisation
- Moore: trigger
- Brunel: reconstruction
- DaVinci: stripping (single stream)
- Any file may be saved, usually only the final DST (uploaded to CERN or "nearest" Tier1)
- 100 to 200 events per job
- 5 to 10 hours duration
- 40 to 80 MB per file
- Merging required (see later)

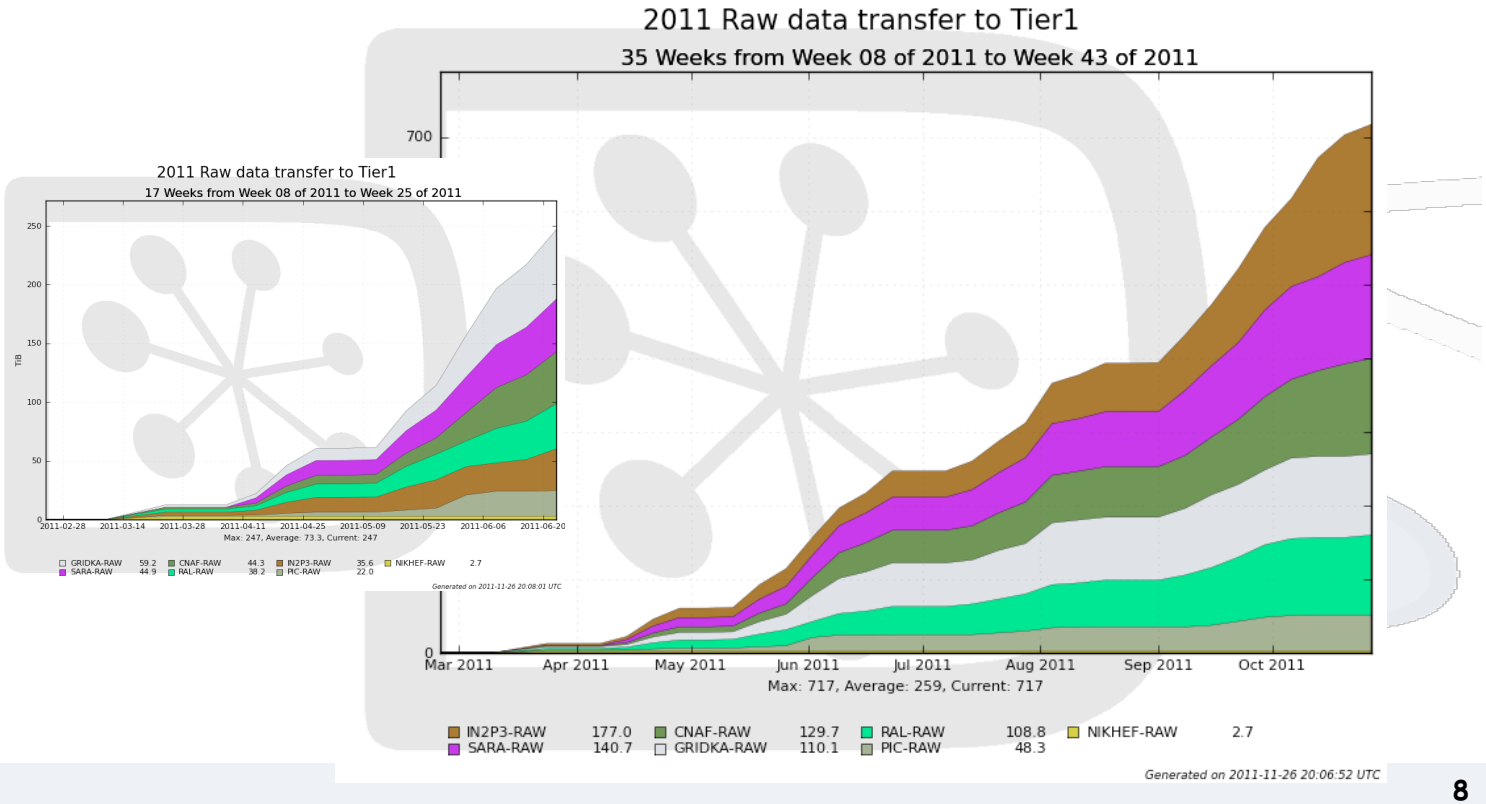
Filltered simulation



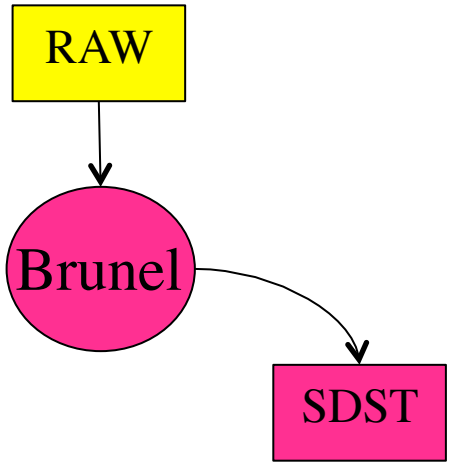
- Huge productions (~100M events) for background studies
  - Cannot afford disk space to store them all
- Only interested in events that pass trigger and stripping
  - Run through full chain but with additional selection step
  - Very few events (1-2 events per job selected)
  - 5-10 hours per job
  - ~1MB per file

RAW data distribution

- ~700 TB of physics RAW data collected in 2011
- 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
    - But share must be recovered later, affects future CPU share, when reprocessing

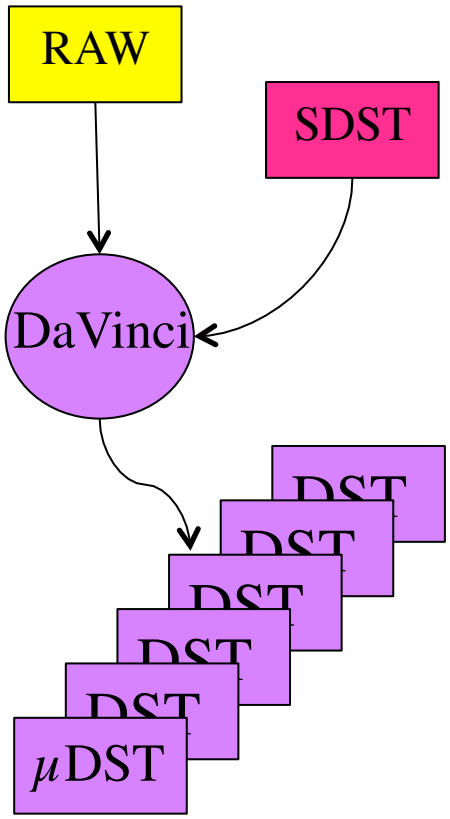


Reconstruction jobs



- 1 step jobs
- One input file: copy to local disk
- Brunel: reconstruction
  - Prompt reconstruction: requires access to Oracle CondDB for latest conditions
  - Reprocessing: SQLite files in software shared area
- SDST saved (local Tier1)
  - local T1D0 (LHCb-Tape)
  - single copy, constrains where further processing can run
- RAW files up to 3 GB (45,000 events)
- 2s per event: 1 day jobs
- SDST: 80% of RAW size

Stripping jobs



- One step jobs
- Multiple input files (e.g. 4 SDSTs)
  - RAW files required as well
  - All files must be present on disk cache (LHCb-Tape, possibly staged)
    - Job throughput limited by cache size and/or number of disk spindles
  - Access by protocol (xroot, dcap…)
  - Memory limited
- DaVinci: stripping and streaming
  - Around 10 to 13 streams
- (μ)DSTs saved (locally)
  - Temporary T0D1 (LHCb-Disk)
- Sum of DSTs: ~20% of RAW size
  - Individual files small (10-100's MB)
  - Merging required

Memory limitations of stripping jobs

Memory

- Big improvement by moving to ROOT persistency (as seen in Stripping16 validation), but memory usage still 20% too much.

Stripping13, POOL persistency

Stripping16, ROOT persistency

Merging jobs

- Automatically generated from job output
  - Simulation
  - Stripping
- Typically 5 GB files
  - For real data, only merge data from the same run
  - Run duration can easily be adjusted online
    - was 1 hour in 2011 (10 Mevts, ~250 RAW files of 3 GB)
- Merged files uploaded locally to TOD1
  - LHCb-Disk

Data replication

- Performed by a data driven transformation
  - Same mechanism as productions
    - See Federico Stagni's talk at ACAT 2011
- Distribution policy implements the LHCb Computing Model
  - RAW files: one Tier1 from the Tier0 replica
  - MC: 3 Tier1 (or CERN) T0D1 replicas, 2 T1D0 archive copies
    - Sites selected randomly
    - Foresee to implement space driven policy
  - Real data: 4 T0D1 replicas (CERN + Tier1), 2 T1D0 archives
    - Differs from CM (should be one replica per Tier1)
    - Adaptation following larger event sizes
    - Each run is distributed to the same sites
    - Replicas reduced to two copies for previous processing
- Replication using FTS

Group level productions

- Two types of central group productions are foreseen:
- Group selections
  - Similar to stripping, but with stripped DST as input
  - Low CPU use, I/O intensive
    - Input data resides on lhcb-disk
    - 20-100 input files per job
- "Swimming" selection
  - Special type of analysis job, with very large CPU requirement (several seconds per event)
    - Similar to reconstruction but with stripped DST as input
    - One file per job

Recent changes

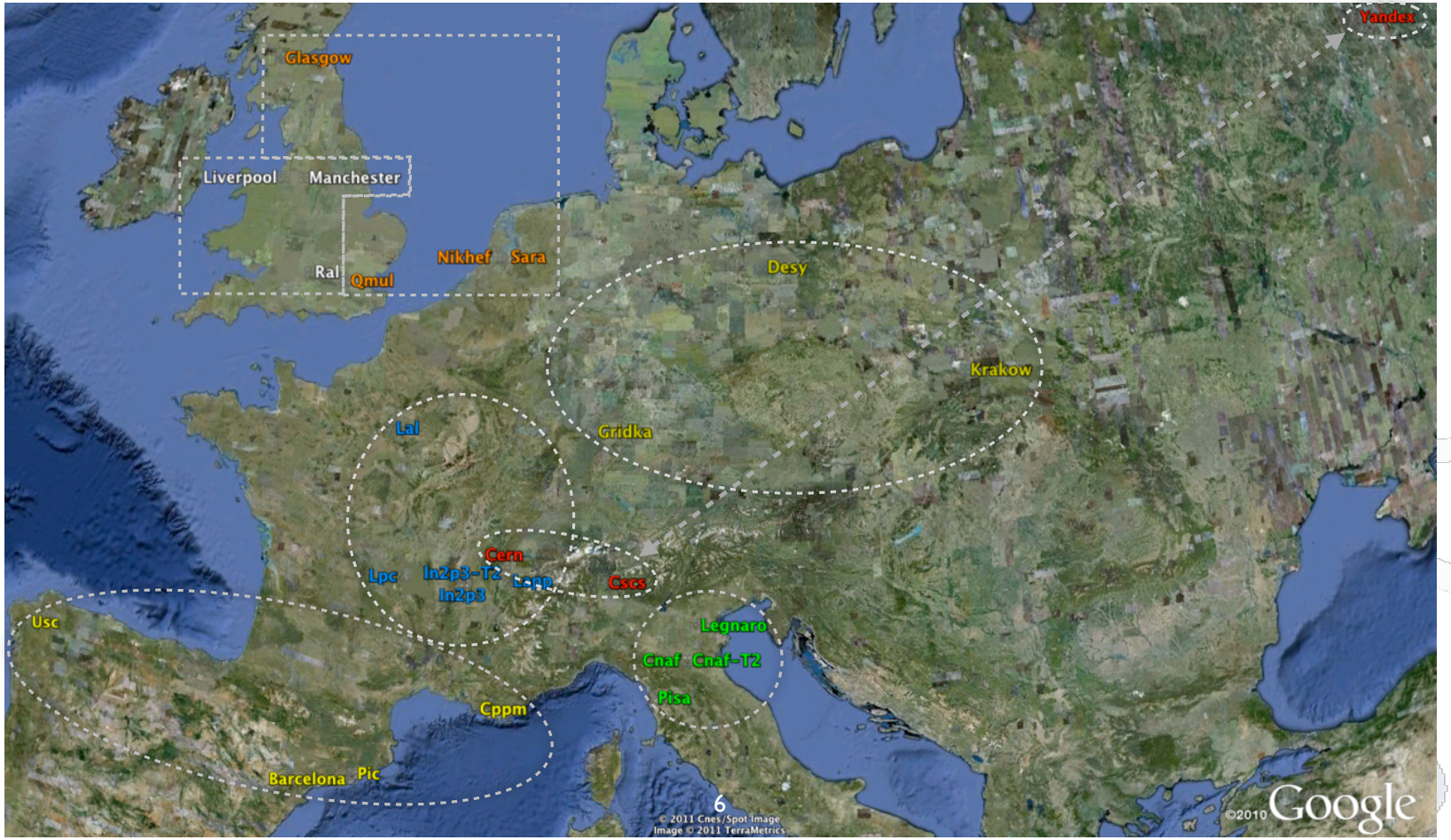
- Reconstruction at Tier2:
  - Using files download from a well connected Tier1
  - Selected number of Tier2s
  - Better control than Analysis
    - All handled by the Production team, better organised, less chaotic
  - Used for end-of-year reprocessing
- Group productions at Tier2s:
  - Possible option for CPU intensive "swimming" productions
    - But do we need it?
  - Keep things simple: use Tier2 only if Tier1 are saturated
    - Tier1 turn-around very good (all user jobs finished overnight)

2011 end of year reprocessing

- Goal: reprocess all 2011 data as fast as possible, to have full dataset available for winter conference analyses
  - Had to start before end of data-taking
    - Had to run in parallel with first pass reconstruction
  - Insufficient CPU resources at CERN+Tier1s

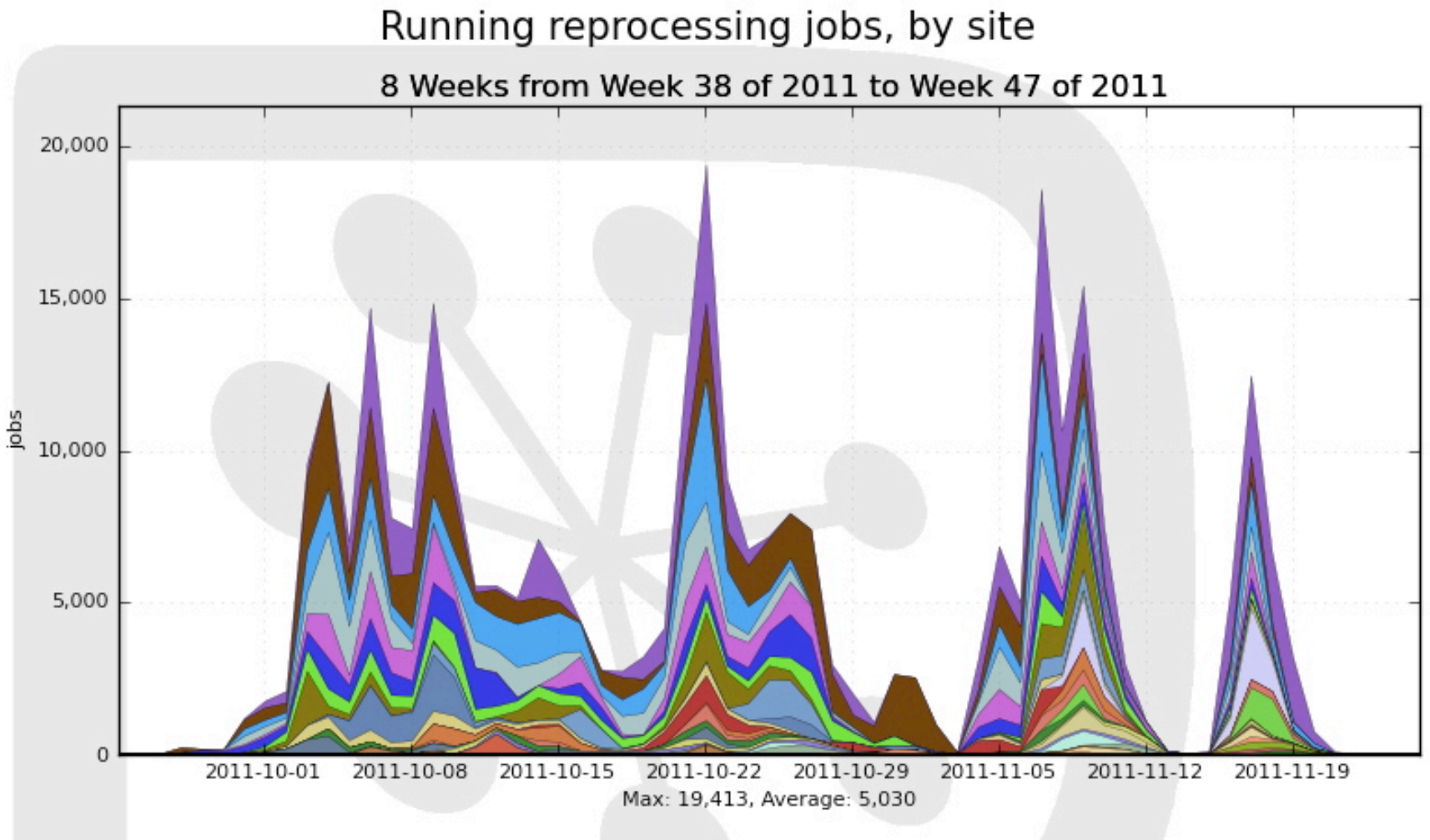
- New model:
  - Do all first pass reconstruction at CERN
    - Not a permanent change, only during October 2011 while still taking data
  - Dedicate the Tier1's to reprocessing
  - Do some of the reprocessing at selected Tier2s
    - Associate Tier2s to a "nearby" Tier1 site
      - Input data (RAW+SDST) downloaded to Tier2
      - Output uploaded to associated Tier1

2011 reprocessing: Tier1-Tier2 associations



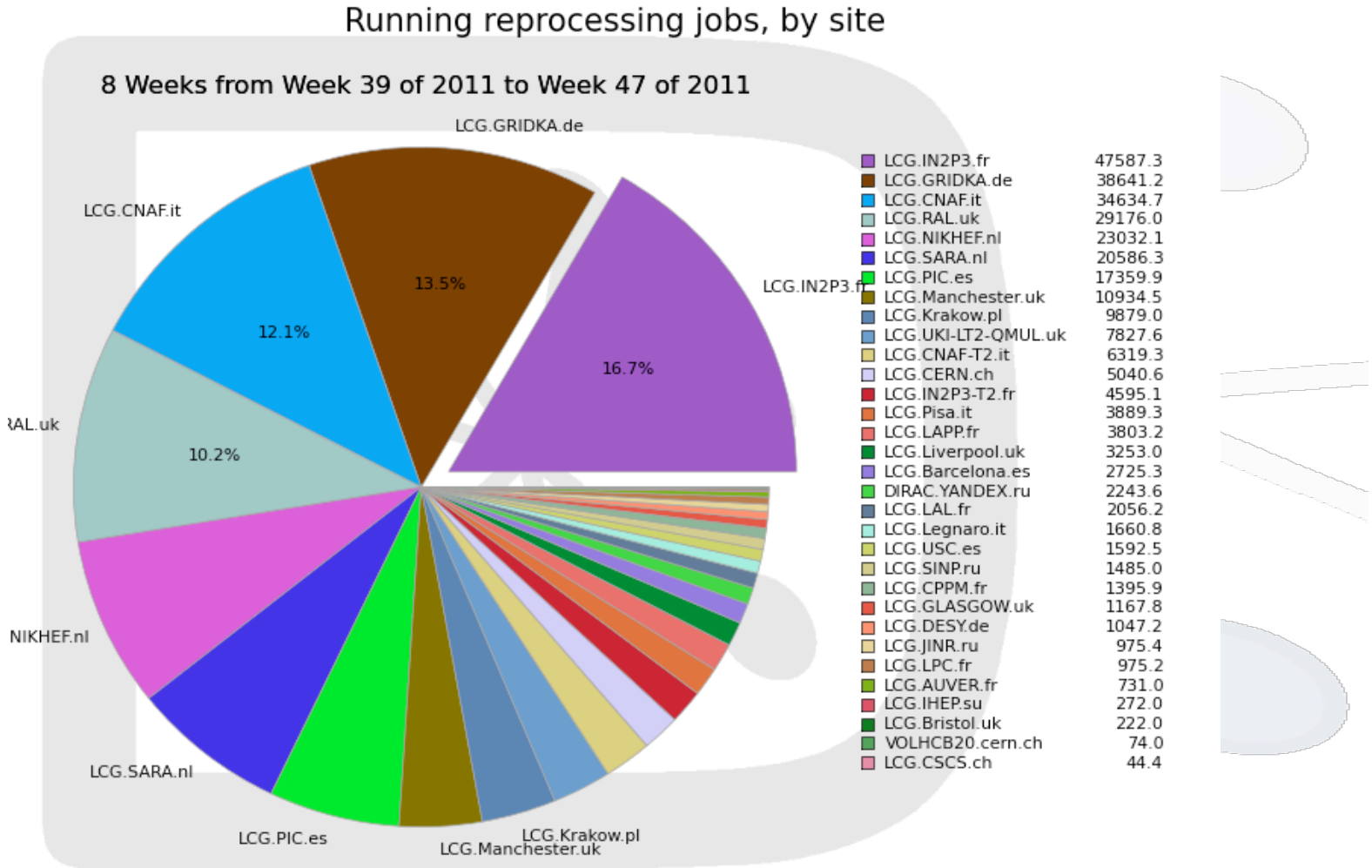
Reprocessing 2011 progress

- Reprocessing started end September, completed 20th November
  - One month faster than forecast



2011 reprocessing, by site

- Biggest single contribution from CCIN2P3
- Total Tier2 contribution > 25%
  - But relatively small impact of French T2 sites



Forthcoming production plans

- MC11 simulation
  - Huge production over several months, starting now
    - 2 billion events, ~one minute per event
    - Need to go as fast as possible for winter conference analysis
- "Swimming" production on 2011 data
  - Several seconds per event, low I/O. Starting soon
- Re-stripping of 2011 data
  - Requires restaging of all 2011 RAW+SDST. In February
    - Will run at Tier1's hosting the SDST
      - Requires replication to CERN of its share of SDST
      - Replication from Tape not tried before
- Prompt reconstruction of 2012 data
  - Starting March, higher data rate than in 2011
    - Larger events (higher pileup), maybe higher HLT rate
- Reprocessing of all 2012 (+2011) data
  - Starting in September