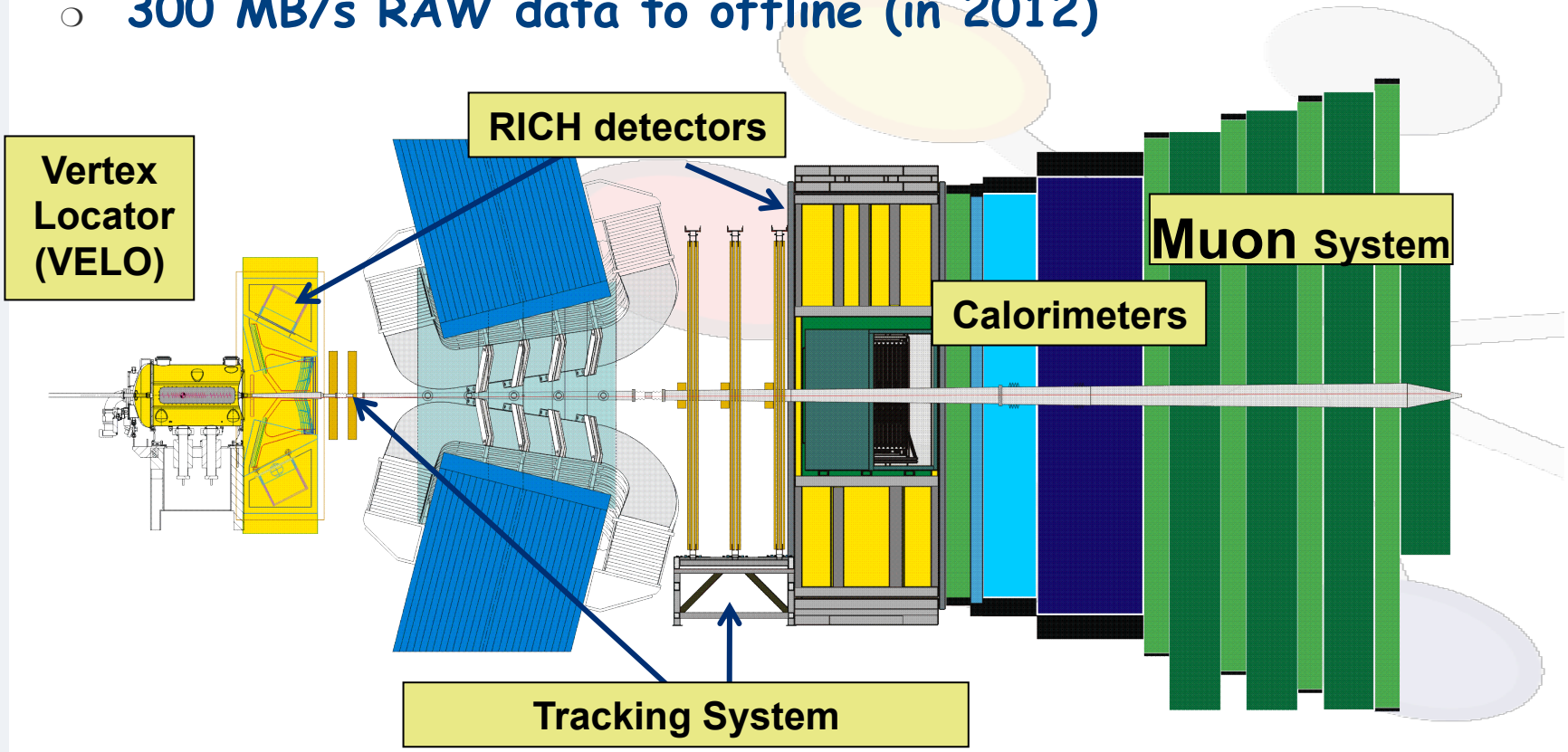
Recent and planned changes to the LHCb computing model

Marco Cattaneo, Philippe Charpentier, Peter Clarke, Stefan Roiser

On behalf of the LHCb Collaboration

LHCb detector

- Forward spectrometer ($1.9 < \eta < 4.9$)
- ~ 2% of solid angle, 27% of heavy quark production cross section inside acceptance
- 300 MB/s RAW data to offline (in 2012)

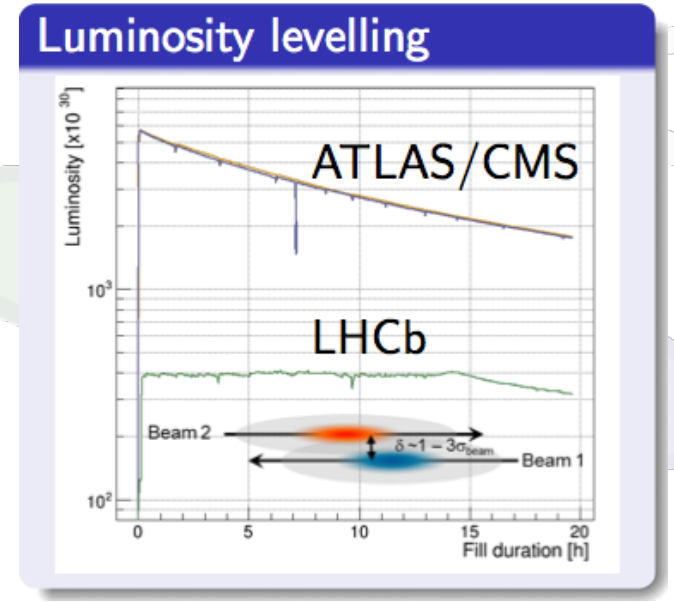


LHCb data-taking conditions

- Exceeding specifications to maximise physics reach

	Design (TDR)	2012 actual	2015 design
Instantaneous luminosity ($cm^{-2}s^{-1}$)	2×10^{32}	4×10^{32}	4×10^{32}
Mean visible p-p interactions/crossing (μ)	0.4	1.7	~1.0 (25ns)
Raw event size (kB)	25	60	60 (25ns)
HLT output rate (kHz)	2	5	12.5

- Luminosity kept constant throughout the run
  - Constant trigger rate
  - Constant event size
- Computing resources scale linearly with trigger rate and length of run

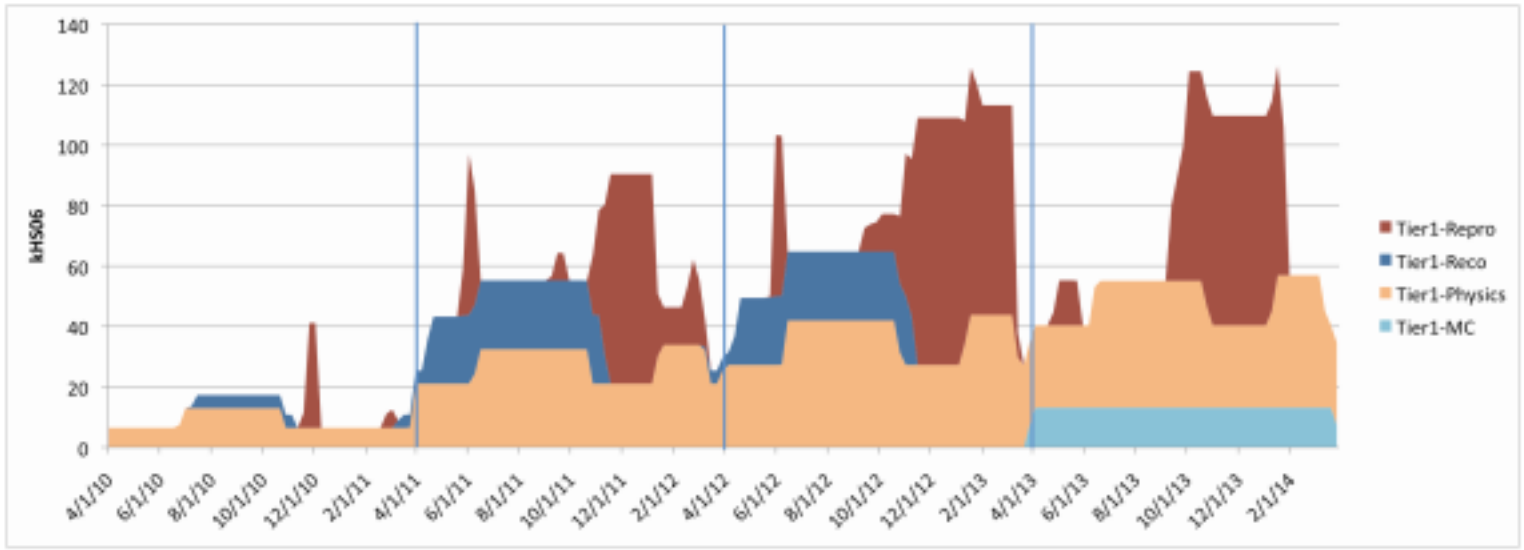


LHCb Computing Model (TDR)

- RAW data: 1 copy at CERN, 1 copy distributed (6 Tier1s)
  - First pass reconstruction runs democratically at CERN+Tier1s
  - End of year reprocessing of complete year's dataset
    - Also at CERN+Tier1s
- Each reconstruction followed by "stripping" pass
    - Event selections by physics groups, several 1000s selections in ~10 streams
    - Further stripping passes scheduled as needed
- Stripped DSTs distributed to CERN and all 6 Tier1s
  - Input to user analysis and further centralised processing by analysis working groups
    - User analysis runs at any Tier1
    - Users do not have access to RAW data or unstripped Reconstruction output
- All Disk located at CERN and Tier1s
  - Tier2s dedicated to simulation
    - And analysis jobs requiring no input data
  - Simulation DSTs copied back to CERN and 3 Tier1s

Problems with TDR model

- Tier1 CPU power sized for end of year reprocessing
  - Large peaks, increasing with accumulated luminosity



- Processing model makes inflexible use of CPU resources
  - Only simulation can run anywhere
- Data management model very demanding on storage space
  - All sites treated equally, regardless of available space

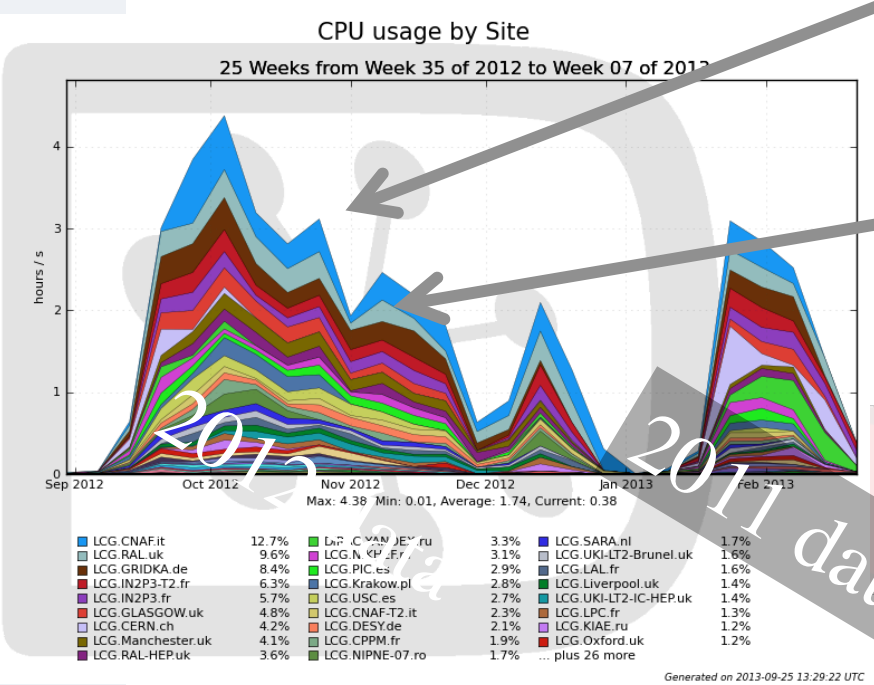
Changes to processing model in 2012

- In 2012, doubling of integrated luminosity c.f. 2011
  - New model required to avoid doubling Tier1 power
- Allow reconstruction jobs to be executed on a selected number of Tier2 sites
  - Download the RAW file (3GB) from a Tier1 storage
  - Run the reconstruction job at the Tier2 site (~ 24 hours)
  - Upload the Reco output file to the same T1 storage
- Rethink first pass reconstruction & reprocessing strategy
  - First pass processing mainly for monitoring and calibration
    - Used also for fast availability of data for 'discovery' physics
  - Reduce first pass to < 30% of RAW data bandwidth
    - Used exclusively to obtain final calibrations within 2-4 weeks
  - Process full bandwidth with 2-4 weeks delay
    - Makes full dataset available for precision physics without need for end of year reprocessing

"Reco14" processing of 2012 and 2011 data

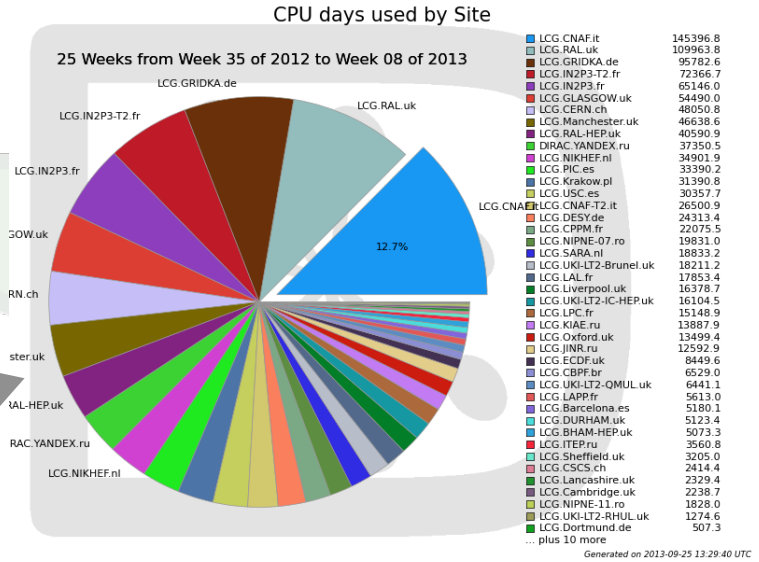
"Stop and Go" for 2012 data as it needed to wait calibration data from the first pass processing.

Power required for continuous processing of 2012 data roughly equivalent to power required for reprocessing of 2011 data at end of year



2012 data

2011 data



45 % of reconstruction CPU time provided by 44 additional Tier2 sites

But also outside WLCG (Yandex)

2015: suppression of reprocessing

- During LS1, major redesign of LHCb HLT system
  - Poster "The LHCb Trigger Architecture beyond LS1"
  - HLT1 (displaced vertices) will run in real time
  - HLT2 (physics selections) deferred by several hours
    - Run continuous calibration in the Online farm to allow use of calibrated PID information in HLT2 selections
    - HLT2 reconstruction becomes very similar to offline
- Automated validation of online calibration for use offline
  - Includes validation of alignment
  - Removes need for "first pass" reconstruction
- Green light from validation triggers 'final' reconstruction
  - Foresee up to two weeks' delay to allow correction of any problems flagged by automatic validation
  - No end of year reprocessing
    - Just restripping
- If insufficient resources, foresee to 'park' a fraction of the data for processing after the run
  - Unlikely to be needed before 2017 but commissioned from the start

Going beyond the Grid paradigm

- Distinction between Tiers for different types of processing activities becoming blurred
  - Currently, production managers manually attach/detach sites to different production activities in DIRAC configuration system
    - In the future sites declare their availability for a given activity and provide the corresponding computing resources

- DIRAC allows easy integration of non WLCG resources
  - In 2013, ~20% of CPU resources from LHCb HLT farm
    - 6.5% from Yandex
  - Vac infrastructure — Talk 119, 11:22 Thursday, Distr. Proc. and Data Handling A
    - Virtual machines created and contextualised for virtual organisations by remote resource providers
  - Clouds — Talk 31, 13:30 Tuesday, Distr. Proc. and Data Handling A
    - Virtual machines running on cloud infrastructures collecting jobs from the LHCb central task queue
  - Volunteer computing
    - Use the BOINC infrastructure to enable payload execution on arbitrary compute resources

Changes to data management model

- Increases in trigger rate and expanded physics programme put strong pressure on storage resources
- Tape shortages mitigated by reduction in archive volume
  - Archives of all derived data exist as single tape copy
    - Forced to accept risk of data loss
- Disk shortages addressed by
  - Introduction of Disk at Tier 2
  - Reduction of event size in derived data formats
  - Changes to data replication and data placement policies
  - Measurement of data popularity to guide decisions on replica removals

Tier2Ds

- In LHCb computing model, user analysis jobs requiring input data are executed at sites holding the data on disk
  - So far Tier1 sites were the only ones to provide storage and computing resources for user analysis jobs
- Tier2Ds are a limited set of Tier2 sites which are allowed to provide disk capacity for LHCb
  - Introduced in 2013 to circumvent shortfall of disk storage
    - To provide disk storage for physics analysis files (MC and data)
    - Run user analysis jobs on the data stored at the sites
  - Status
    - 4 sites currently being put in production
      - Ramping up to a minimum of 300 TB/site over the coming months
    - 1 more site in the pipeline
- Blurs even more functional distinction between Tier1 and Tier2
  - A large Tier2D is a small Tier1 without Tape

Data Formats

- Highly centralised LHCb data processing model allows to optimise data formats for operation efficiency

- Large shortfalls in disk and tape storage (due to larger trigger rates and expanded physics programme) drive efforts to reduce data formats for physics:
  - DST used by most analyses in 2010 (~120kB/event)
    - Contains copy of RAW and full Reco information
  - Strong drive to microDST (~13kB/event)
    - Suitable for most exclusive analyses, but many iterations required to get content correct
  - Transparent switching between several formats through generalised use of analysis software framework

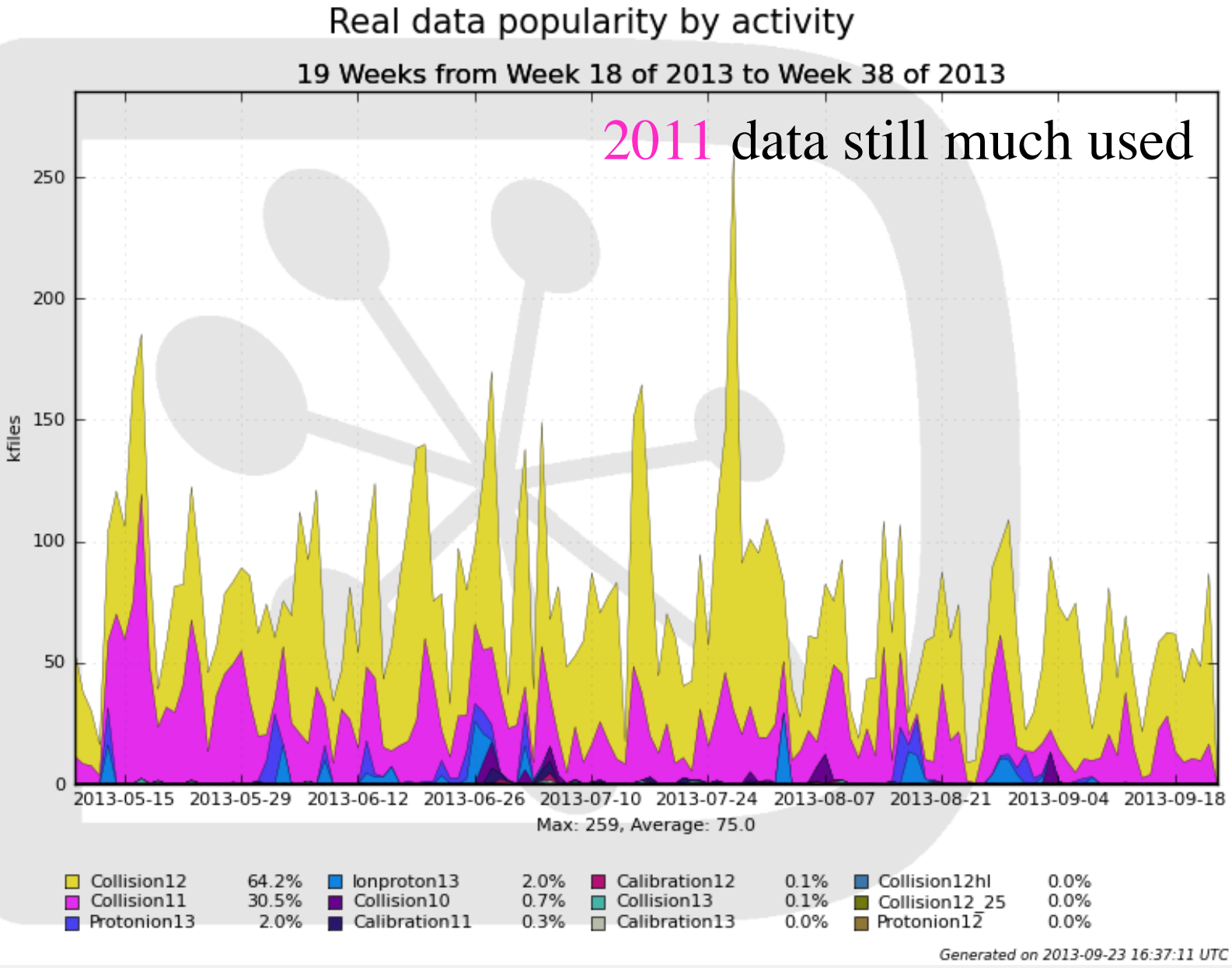
Data placement of DSTs

- Data-driven automatic replication
  - Archive systematically all analysis data (T1D0)
  - Real Data: 4 disk replicas, 1 archive
  - MC: 3 disk replicas, 1 archive
- Selection of disk replication sites:
  - Keep together whole runs (for real data)
    - Random choice per file for MC
  - Chose storage element depending on free space
    - Random choice, weighted by the free space
    - Should allow no disk saturation
      - Exponential fall-off of free space
      - As long as there are enough non-full sites!
- Removal of replicas
  - For processing n-1: reduce to 2 disk replicas (randomly)
    - Possibility to preferentially remove replicas from sites with less free space
  - For processing n-2: only keep archive replicas

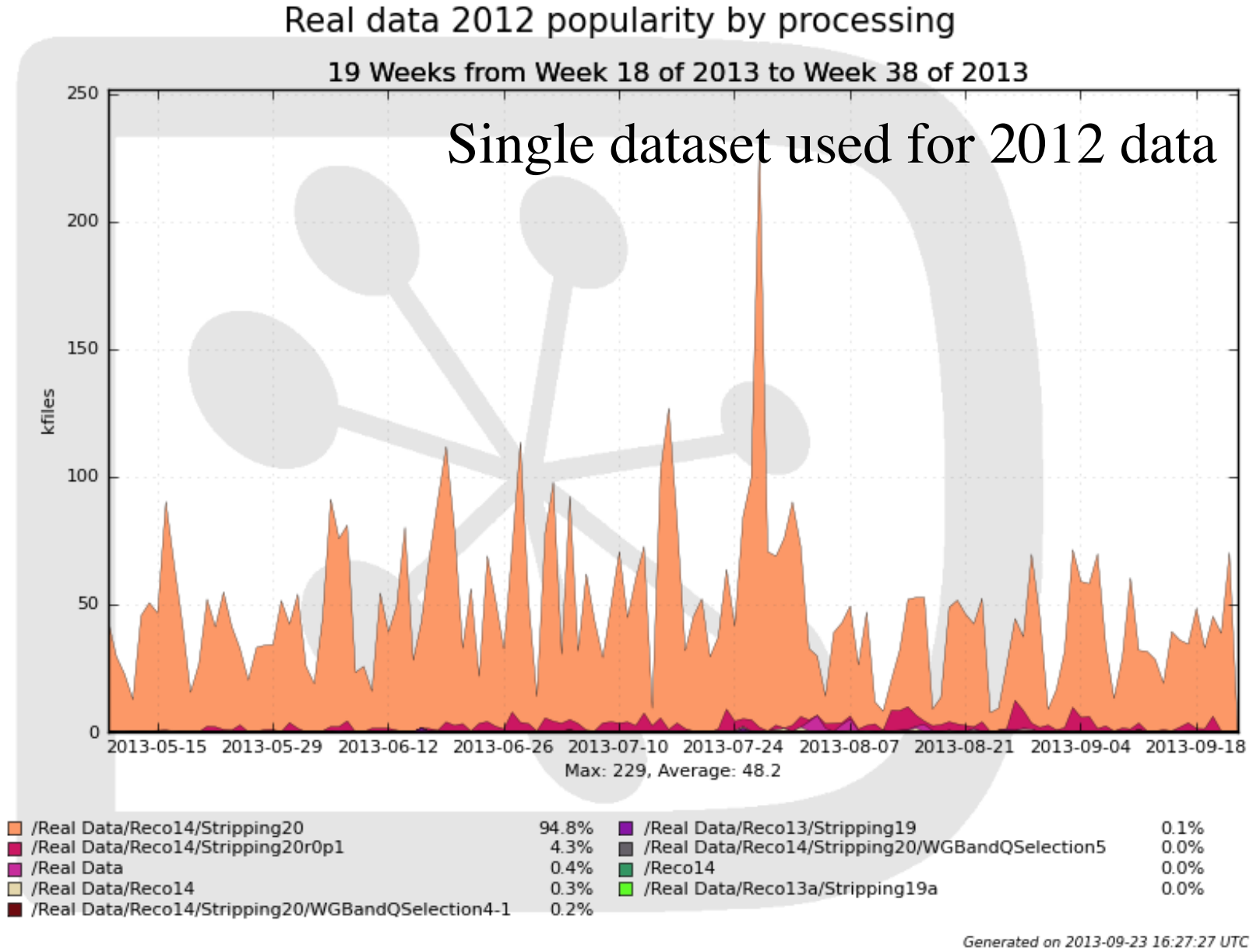
Data popularity

- Enabled recording of information as of May 2012
- Information recorded for each job:
  - Dataset (path)
  - Number of files for each job
  - Storage element used
- Allows currently by visual inspection to identify unused datasets
- Plan:
  - Establish, per dataset:
    - Last access date
    - Number of accesses in last (n) months (1<n<12)
    - Normalise number of dataset accesses to its size
  - Prepare summary tables per dataset
    - Access summary (above)
    - Storage usage at each site
  - Allow to trigger replica removal when space is required

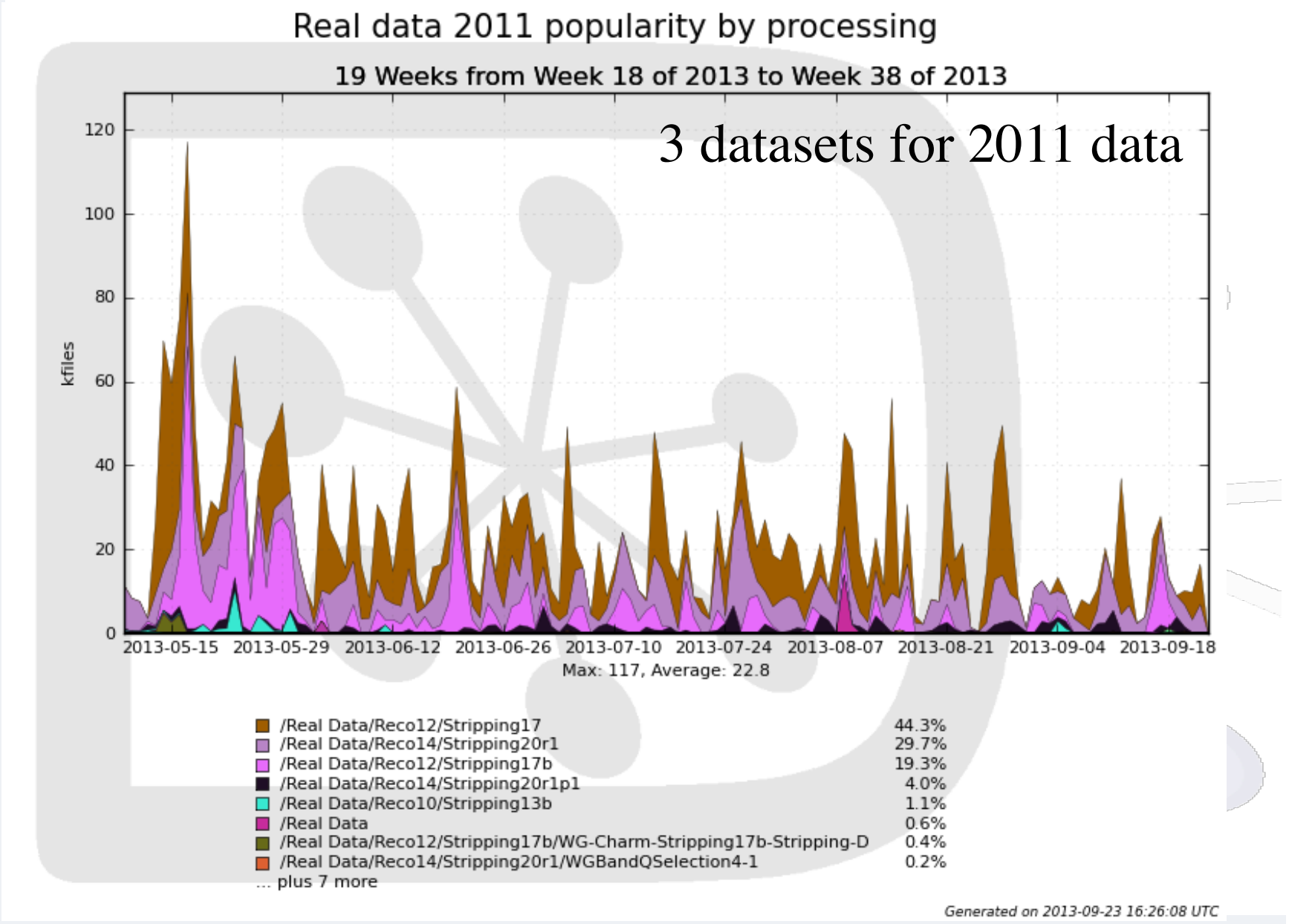
Examples of popularity plots (kFiles/day)



Examples of popularity plots (kFiles/day)



Examples of popularity plots (kFiles/day)



Conclusions

- The LHCb computing model has evolved to accommodate within a constant budget for computing resources the expanding physics programme of the experiment
- The model has evolved from the hierarchical model of the TDR to a model based on the capabilities of different sites
- Further adaptations are planned for 2015. We do not foresee the need for any revolutionary changes to the model or to our frameworks (Gaudi, Dirac) to accommodate the computing requirements of LHCb during Run 2