LHCb Computing

LHCb computing model in Run1 & Run2

Concezio Bozzi

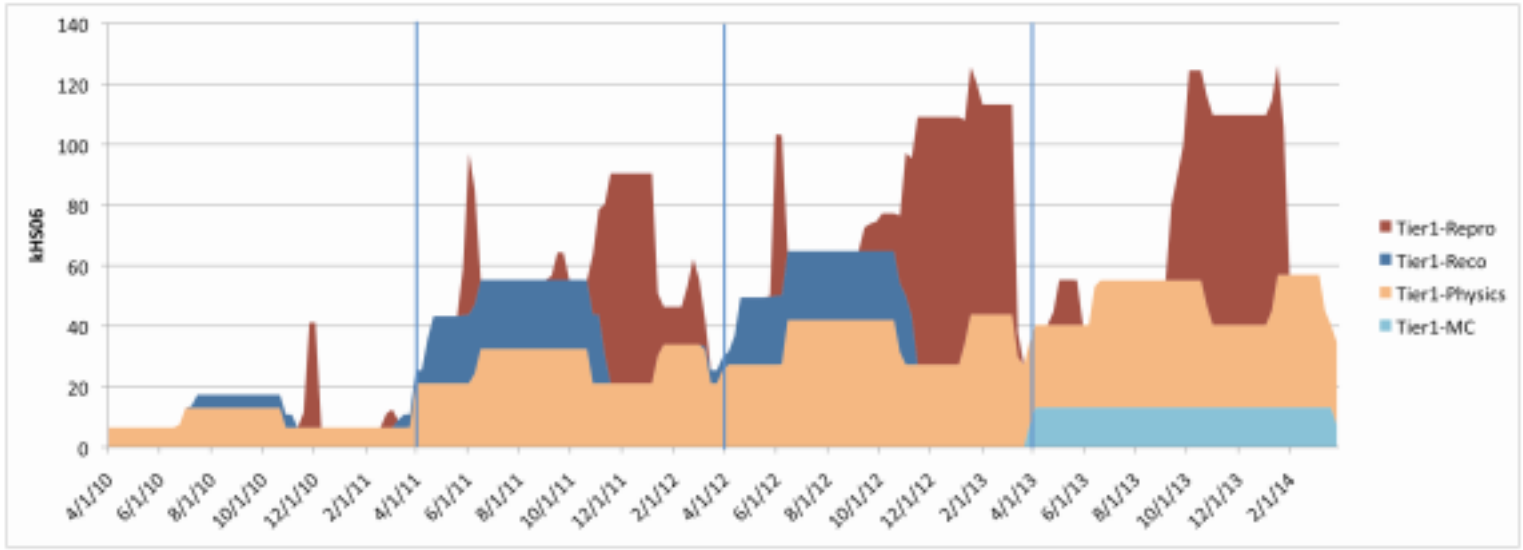
Bologna, Feb 19th 2015

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

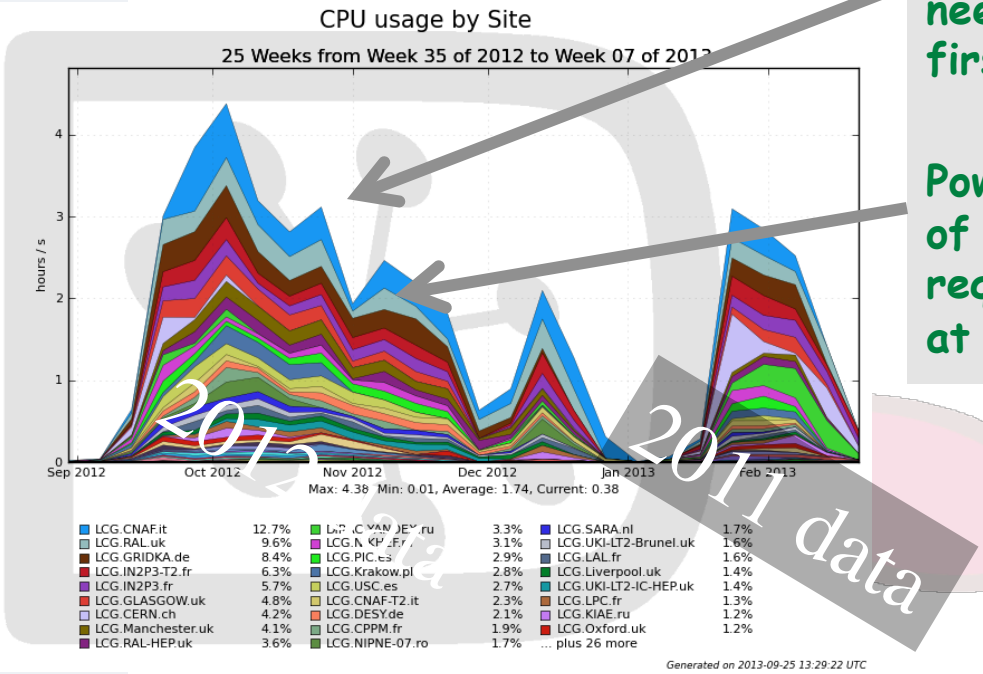
Changes to processing model in 2012

- 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

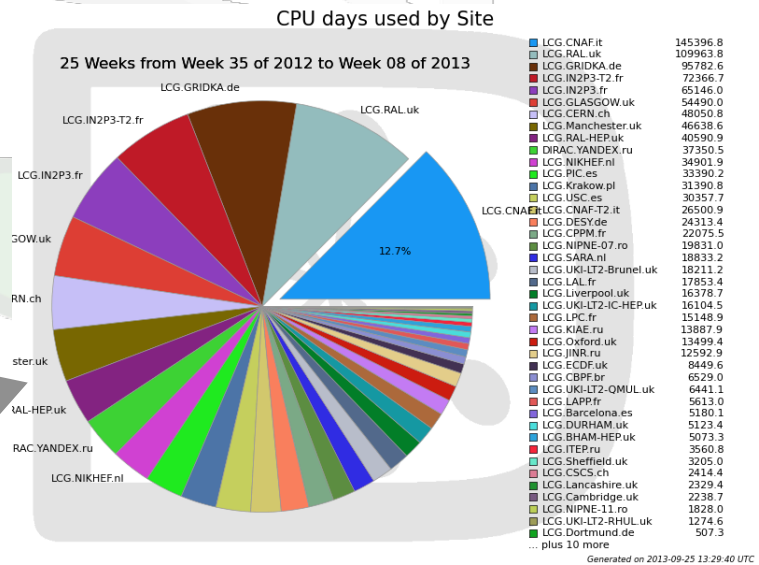
"Stop and Go" for 2012 data as it needed to wait calibration data from the first pass processing.

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



2012

2011 data



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

But also outside WLCG (Yandex)

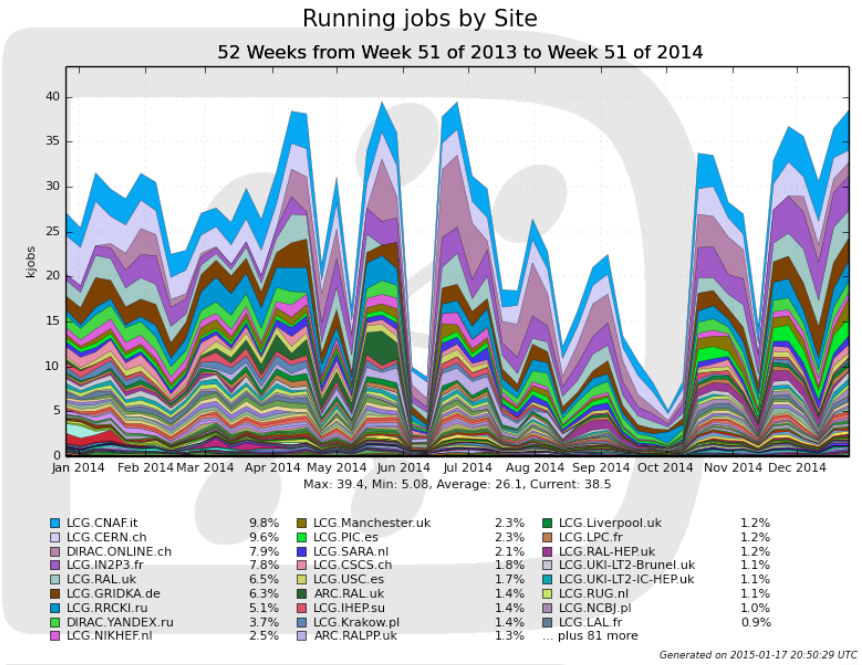
2015: suppression of reprocessing

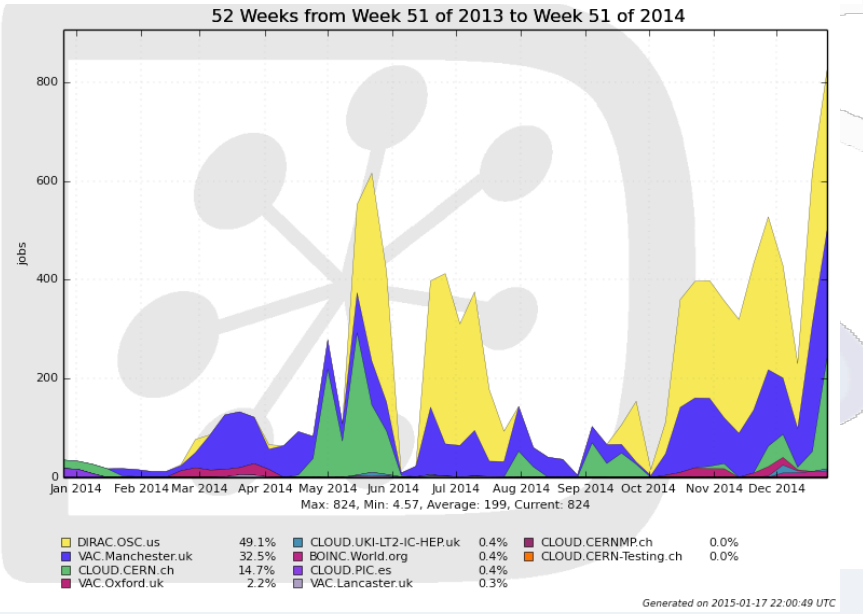
- During LS1, major redesign of LHCb HLT system
  - HLT1 (displaced vertices) will run in real time
  - HLT2 (physics selections) deferred by several hours
    - Run continuous calibration in the Online farm to allow use of calibrated PID information in HLT2 selections
    - HLT2 reconstruction becomes 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

DIRAC allows easy integration of non WLCG resources

- In 2014, ~10% of CPU resources from LHCb HLT and Yandex farms
- Vac infrastructure
  - Virtual machines created and contextualised for virtual organisations by remote resource providers
- Clouds
  - 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



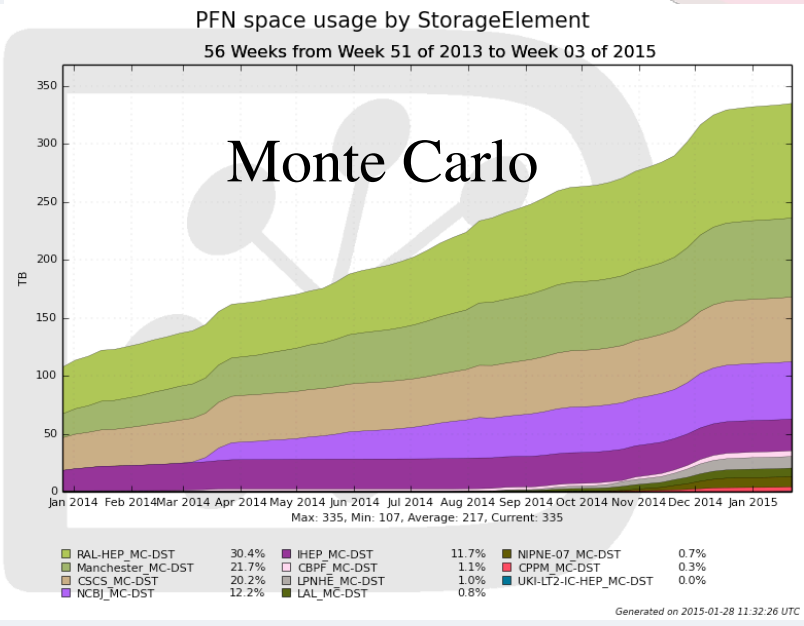


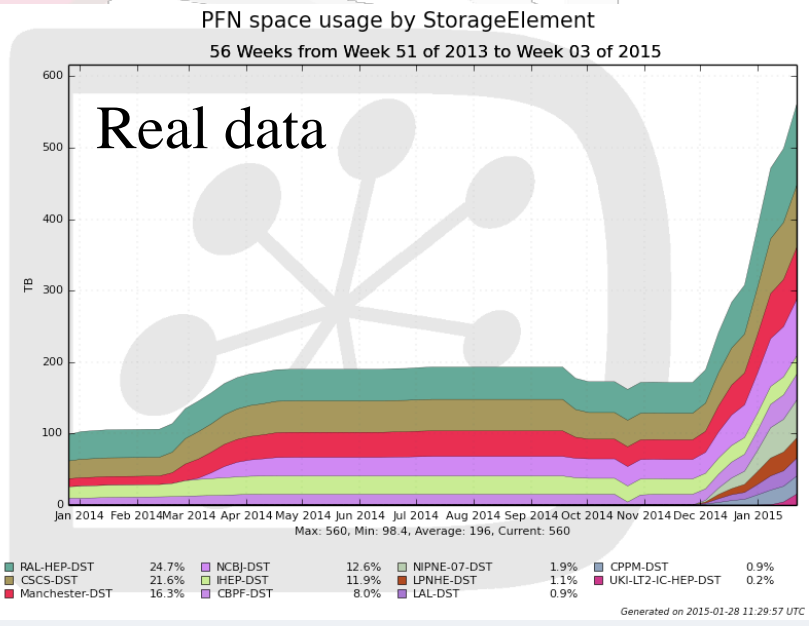
Changes to data management model

- 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
  - Re-introduce a second tape copy in Run2, to cope with data preservation "obligations"
    - Re-generation in case of data loss is an operational nightmare and an overload of computing resources
- 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

- 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
- Blurs even more functional distinction between Tier1 and Tier2
  - A large Tier2D is a small Tier1 without Tape
- Status (Jan 18th 2015): 2.4 PB available, 0.83 PB used





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 μDST (~13kB/event)
    - Save information for signal only
    - Suitable for most exclusive analyses, but many iterations required to get content correct
    - User-defined data can be added on demand (tagging, isolation,…)
  - "Legacy" stripping campaign of Run1 data just completed
    - Will allow to test μDST
    - MDST.DST == FULL.DST of all events passing a μDST stream. Temporary format (2015-2016) to allow regeneration of μDST in case of missing information without running the stripping again

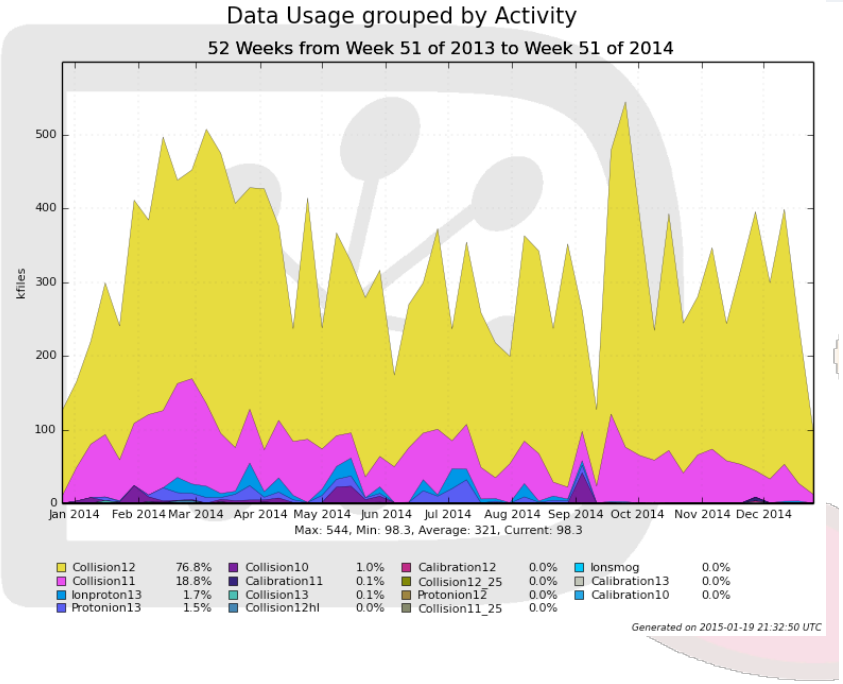
Data placement of (μ)DSTs

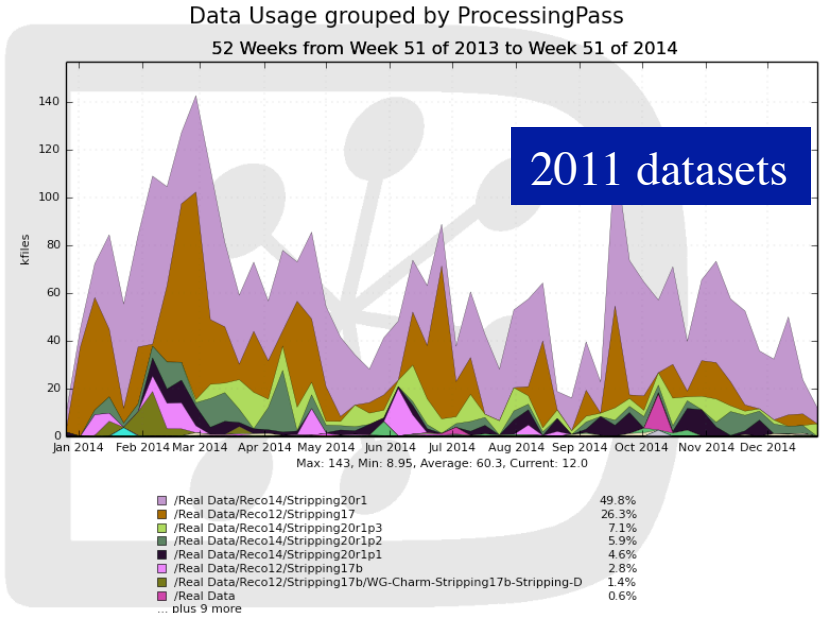
- Data-driven automatic replication
  - Archive systematically all analysis data (T1D0)
  - Real Data: 4 disk replicas, 1(→2) archives
  - MC: 3 disk replicas, 1(→2) archives
- Selection of disk replication sites:
  - Keep together whole runs (for real data)
    - Random choice per file for MC
  - Chose 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

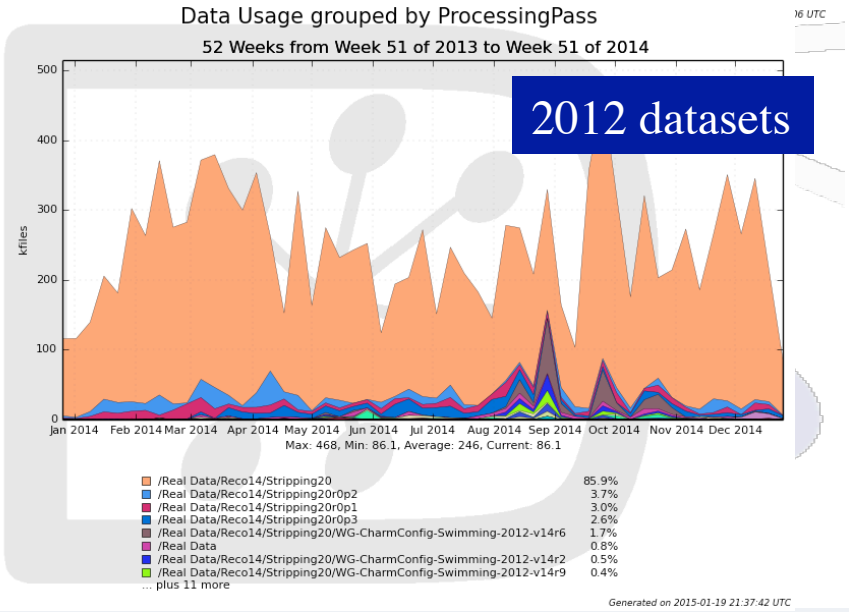
Data popularity

- 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

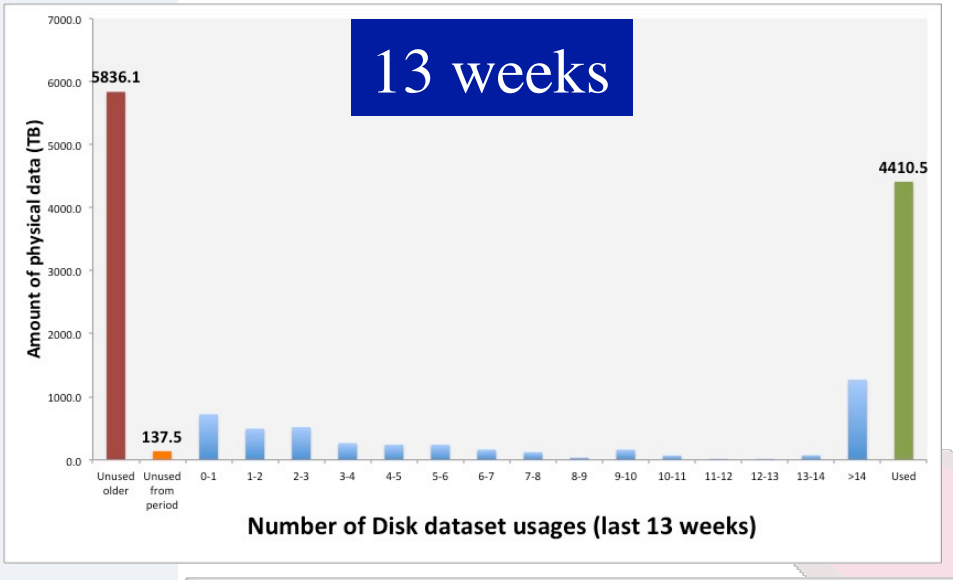


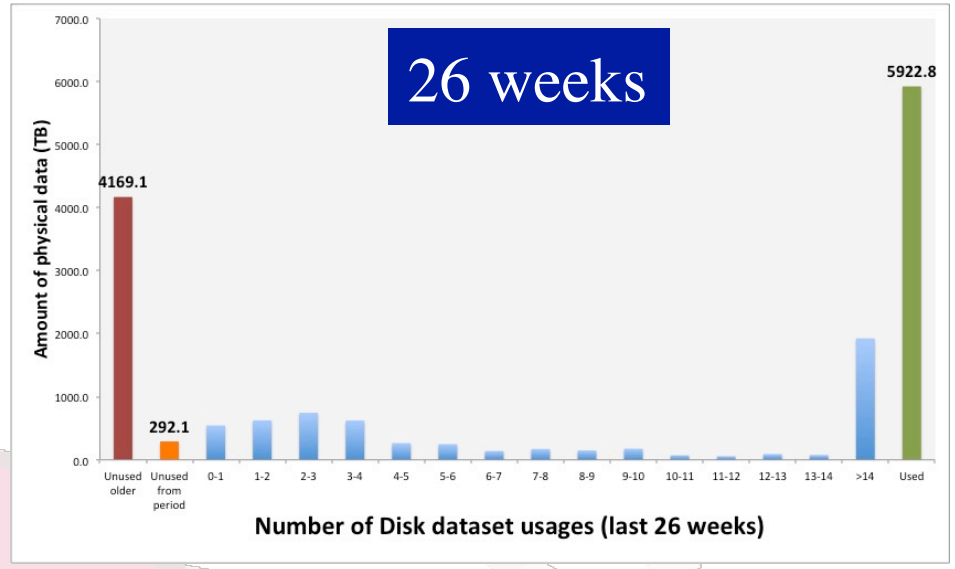


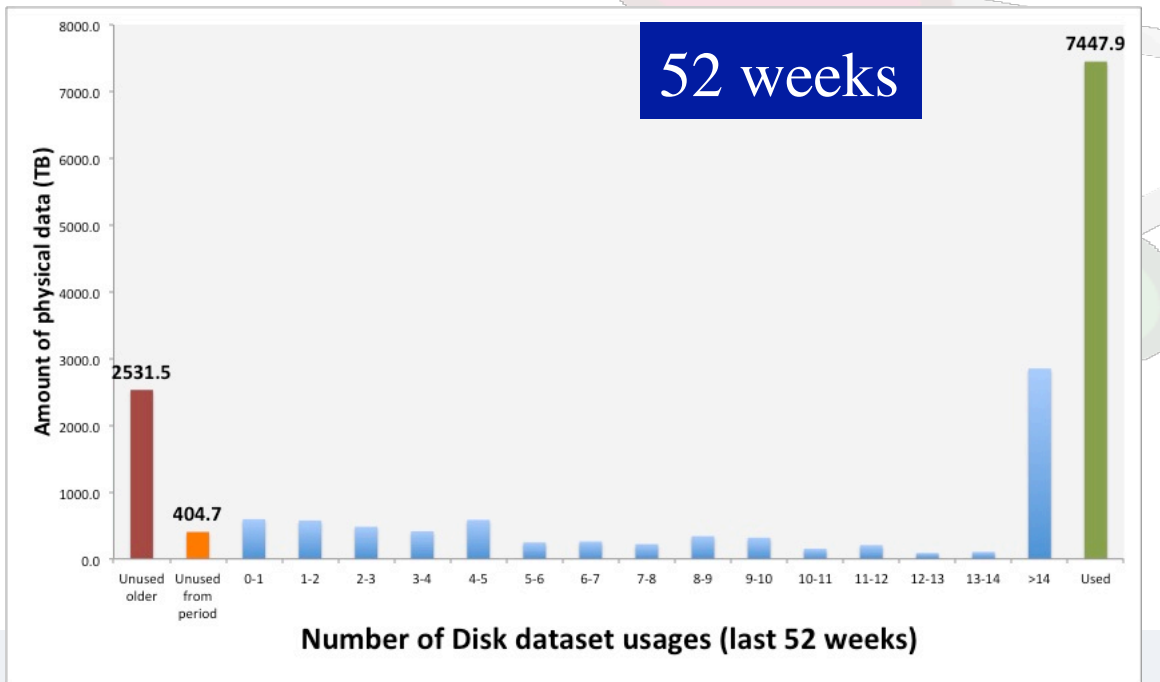


- We are also working on a classifier that, based on all metadata and popularity history, allows to classify datasets into those that are likely to be used within the next *n* weeks and those that are not.

Usage of datasets







~1PB disk space recovered by purging old data

Summary of 2015-2017 requests

Running assumptions

		LHC schedule		
Proton physics	LHC start date	01/05/2015	01/04/2016	01/04/2017
	LHC end date	31/10/2015	31/10/2016	15/12/2017
	LHC run days	183	213	258
	Fraction of days for physics	0.60	0.70	0.80
	LHC efficiency	0.32	0.39	0.39
	Approx. running seconds	$3.0\ 10^6$	$5.0\ 10^6$	$7.0\ 10^6$
Heavy Ion physics	Approx. running seconds	-	$0.7\ 10^6$	$0.7\ 10^6$

CPU

Power (kHS06)	Request 2015	Request 2016	Request 2017
Tier 0	36	51	62
Tier 1	118	156	191
Tier 2	66	88	107
Total WLCG	**220**	**295**	**360**
HLT farm	10	10	10
Yandex	10	10	10
Total non-WLCG	**20**	**20**	**20**
Grand total	**240**	**315**	**380**

Disk

Disk (PB)	2015 Request	2016 Request	2017 Request
Tier0	5.5	7.6	9.1
Tier1	11.7	13.5	15.0
Tier2	1.9	4.0	5.5
Total	**19.1**	**25.2**	**29.6**

Breakdown of CPU requests

pp Running

CPU Work in WLCG year (kHS06.years)	2015	2016	2017
Prompt Reconstruction	19	31	26
First pass Stripping	8	13	11
Full Restripping	8	20	11
Incremental stripping	0	4	10
Simulation	134	153	207
VoBoxes and other services		4	4
User Analysis	17	20	24
Total Work (kHS06.years)	**186**	**246**	**293**
Efficiency corrected average power (kHS06)	**220**	**291**	**348**

HI Running

Resources for heavy ion running	2015 Request	2016 Request	2017 Request
CPU (kHS06)	0	24	32

Breakdown of DISK requests

pp Running

Disk storage usage forecast (PB)	2015	2016	2017
Stripped Real Data	7.3	13.1	15.3
Simulated Data	8.2	6.9	10.4
User Data	0.9	1.0	1.1
MDST.DST	1.5	1.9	0.0
RAW and other buffers	1.0	1.2	0.9
Other	0.2	0.2	0.2
Total	**19.1**	**24.3**	**27.9**

HI Running

Resources for heavy ion running	2015 Request	2016 Request	2017 Request
Disk (PB)	0	0.9	1.7

Tape

Tape (PB)	2015 Request	2016 Request	2017 Request
Tier0	11.2	20.6	30.9
Tier1	23.7	42.1	62.2
Total	**34.9**	**62.7**	**93.1**

pp Running

Tape storage usage forecast (PB)	2015	2016	2017
Raw Data	12.7	21.7	34.5
FULL.DST	8.7	15.2	20.7
MDST.DST	1.8	5.2	7.9
Archive – Operations	8.6	11.6	15.0
Archive – Data preservation	3.1	6.0	9.2
Total	**34.9**	**59.7**	**87.3**

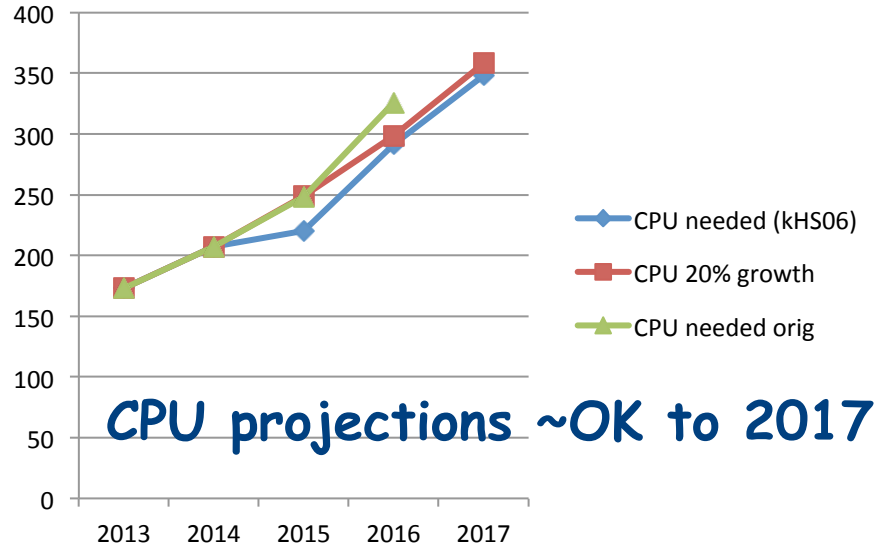
HI Running

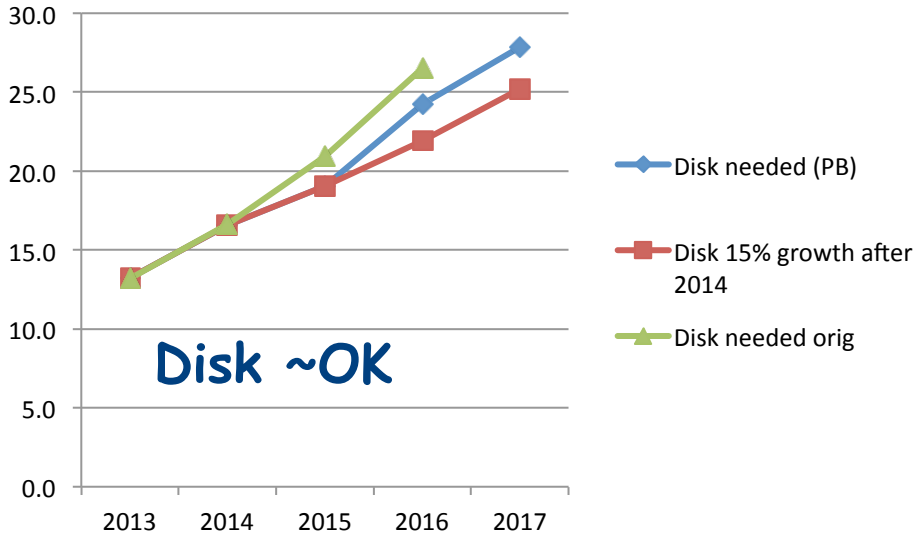
Tape (PB)	0	3.0	5.7

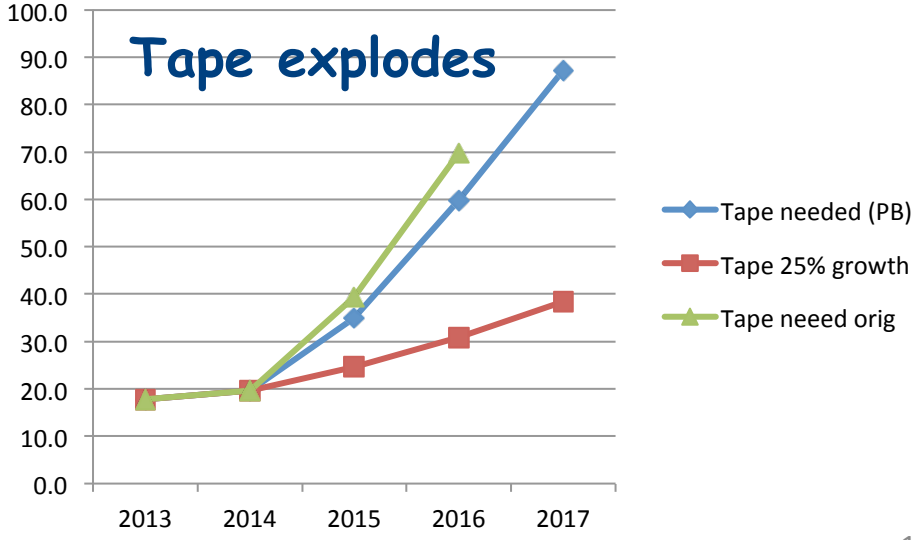
Please note:
WLCG estimates of tape costs include a 10% cache disk.
This is too large for our purposes.

Comparison with “flat budget”

- Definition of flat budget: same money will buy
  - 20% more CPUs
  - 15% more disk
  - 25% more tape





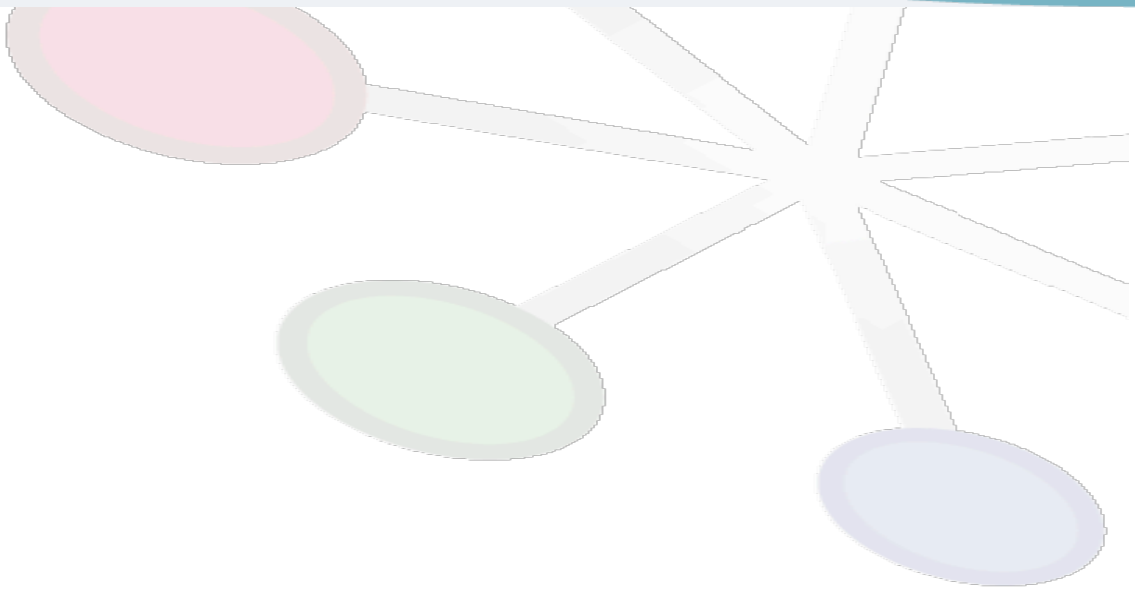


Mitigation strategies

- Increase in tape request is beyond flat-budget expectation
  - Ask resource providers for advance purchases in order to ease ramp-up
  - Trade some other resources for tape
    - But lever arm is short!
- Remove second tape copy of derived dataset?
  - Regeneration of even a small portion of data implies massive tape recalls and computing load, which might jeopardize other production activities
- Continue developing data popularity algorithms and data placement strategies
  - Potential significant savings on disk space
- Continue using available CPU resources "parasitically"

LHCb Computing

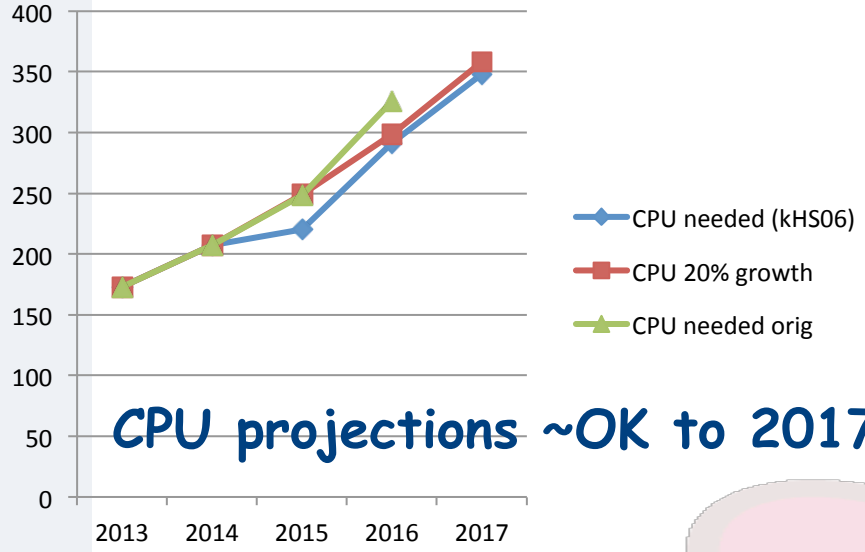
LHCb computing plans for Run 3



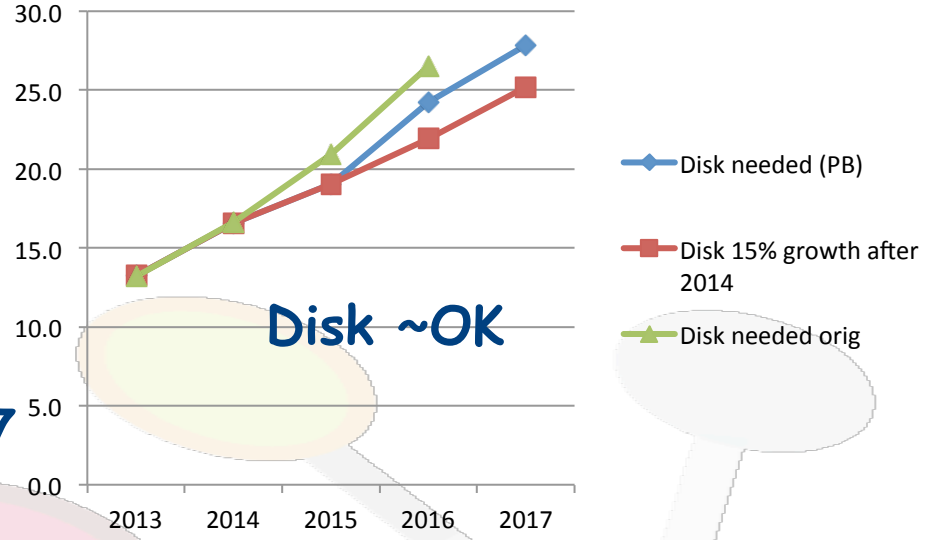
Towards the LHCb Upgrade (Run 3, 2020)

- We do not plan a revolution for LHCb Upgrade computing
- Rather an evolution to fit in the following boundary conditions:
  - Luminosity levelling at 2×10^{33}
    - Factor 5 c.f. Run 2
  - 100kHz HLT output rate for full physics programme
    - Factor 8-10 more than in Run 2
  - Flat funding for offline computing resources
- Computing milestones for the LHCb upgrade (shift proposed to gain experience from 2015-2016 data taking):
  - TDR: 2017Q1 → shifting to Q3
  - Computing model: 2018Q3 → shifting to 2019 Q1
- Therefore only brainstorming at this stage, to devise model that keeps within boundary conditions

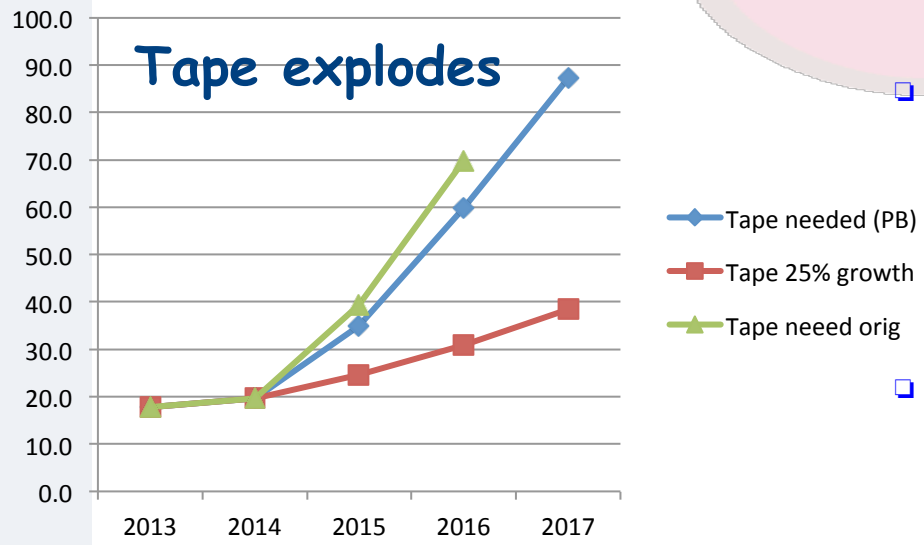
Run 2: computing resources



CPU projections ~OK to 2017



Disk ~OK



Tape explodes

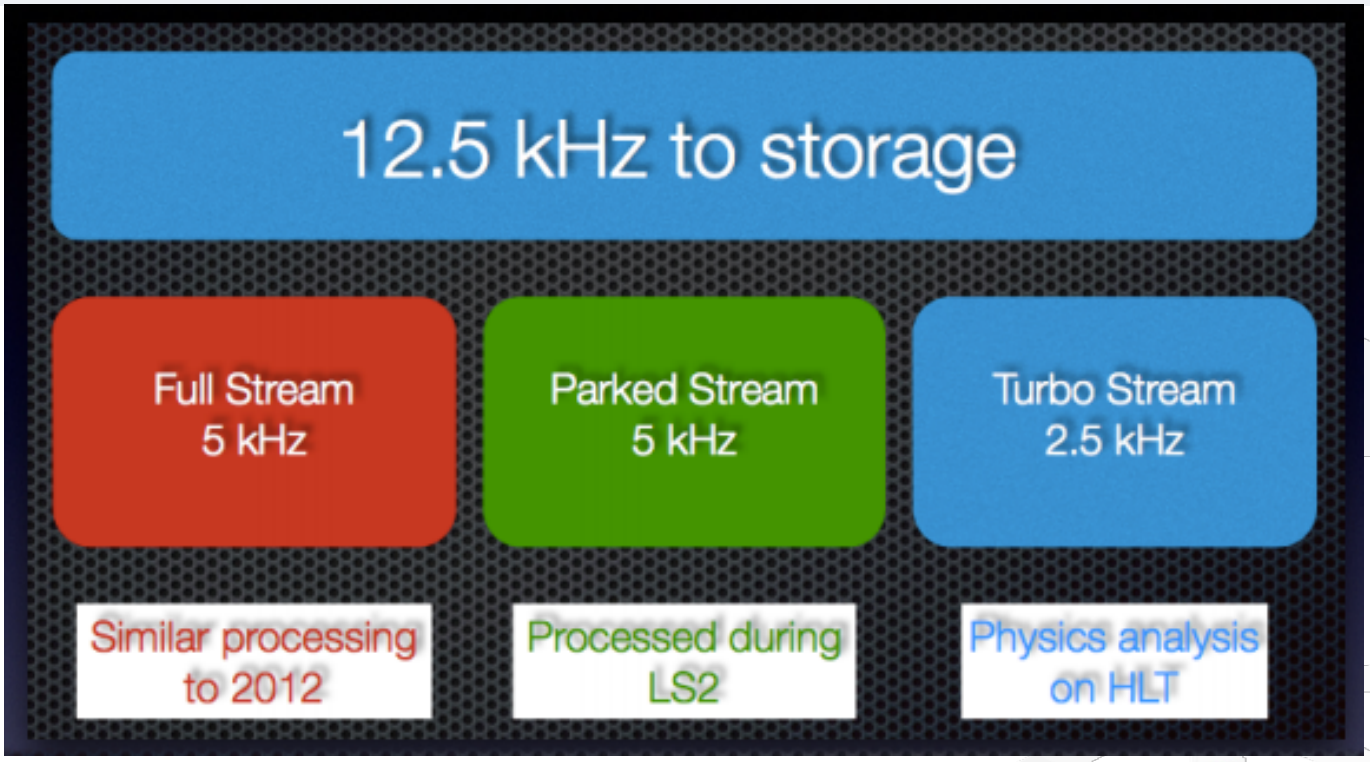
- Tape requirement driven by:
  - Two copies of RAW
    - Incompressible, but ~never accessed
  - One copy Reconstruction output (FULL.DST)
- Flat funding cannot accommodate order of magnitude more data expected in Run 3 - need new ideas

flat budget: same money will buy 20%, 15%, 25% more CPU, disk, tape

Evolution of LHCb data processing model

- Run 1:
  - Loose selection in HLT (no PID)
  - First pass offline reconstruction
  - Stripping
    - selects ~50% of HLT output rate for physics analysis
  - Offline calibration
  - Reprocessing and Restripping
- Run 2:
  - Online calibration
  - Deferred HLT2 (with PID)
  - Single pass offline reconstruction
    - Same calibration as HLT
    - No reprocessing before LS2
  - Stripping and Restripping
    - Selects ~90% of HLT output rate for Physics analysis
- Given sufficient resources in HLT farm, online reconstruction could be made ~identical to offline

Run 2: Reconstruction streams



- Full stream: prompt reconstruction as soon as RAW data appears offline
- Parked stream: safety valve, probably not needed until 2017
- Turbo stream: no offline reconstruction, analysis objects produced in HLT
  - Important test for Run3

TurboDST: brainstorming for Run 3

- In Run 2, Online (HLT) reconstruction will be very similar to offline (same code, same calibration, fewer tracks)
  - If it can be made identical, why then write RAW data out of HLT, rather than Reconstruction output?

- In Run 2 LHCb will record 2.5 kHz of "TurboDST"
  - RAW data plus result of HLT reconstruction and HLT selection
  - Equivalent to a microDST (MDST) from the offline stripping
  - Proof of concept: can a complete physics analysis be done based on a MDST produced in the HLT?
    - i.e. no offline reconstruction
      - no offline realignment, reduced opportunity for PID recalibration
    - RAW data remains available as a safety net
  - If successful, can we drop the RAW data?
    - HLT writes out ONLY the MDST ???

- Currently just ideas, but would allow a 100kHz HLT output rate without an order of magnitude more computing resources.

Simulation

- LHCb offline CPU usage is dominated by simulation
  - Already true in Run 2: simulation >60% of CPU needs in 2016
    - Many measurements start to be limited by simulation statistics
- Simulation suited for execution on heterogeneous resources
  - Pursue efforts to interface Dirac framework to multiple computing platforms
    - Allow opportunistic and scheduled use of new facilities
  - Extend use of HLT farm during LHC stops
- Several approaches to reduce CPU time per event
  - Code optimisation, vectorisation etc.
    - Contribute to and benefit from community wide activities, e.g. for faster transport
  - Fast simulations
    - Not appropriate for many detailed studies for LHCb precision measurements
    - Nevertheless many generator level studies are possible
  - Hybrid approach
    - Full simulation for signal candidates only
    - Fast techniques for the rest
      - e.g. skip calorimeter simulation for out of time pileup
- To avoid being limited by disk space
  - Deploy MDST format also for simulated data

Conclusions

- LHCb event output rate will be an order of magnitude larger in Run 3 (2020)
- Currently brainstorming on ideas for reducing data rate without reducing physics reach
  - Run 2 as a test bed
- Computing efforts concentrated on
  - Code optimisation, e.g.
  - vectorization and GPU usage in HLT
  - Opportunistic use of diverse resources