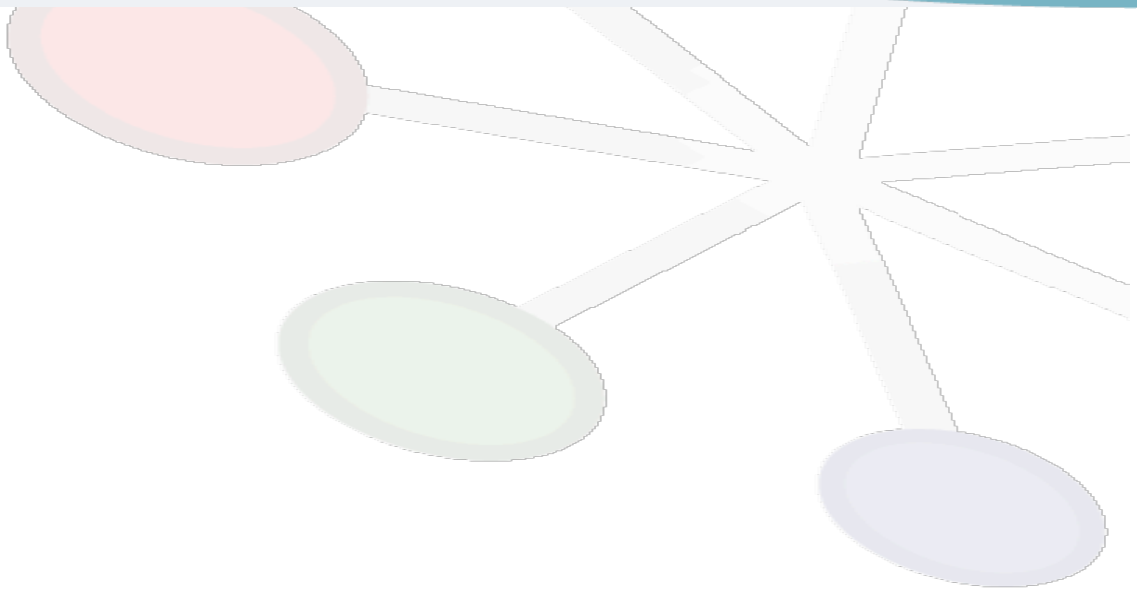
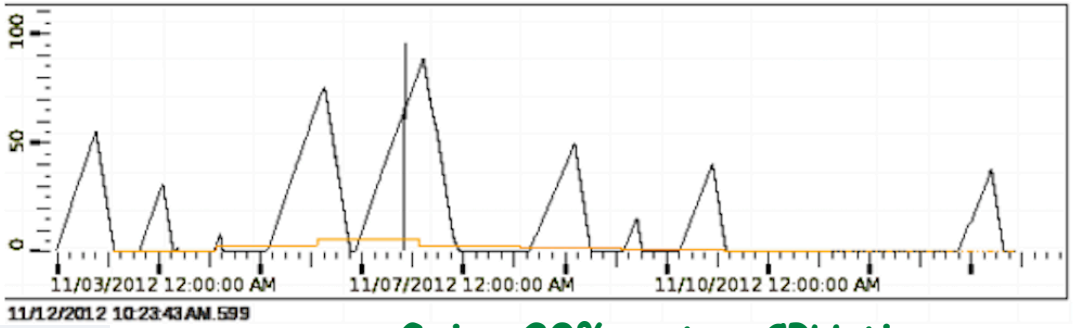
LHCb Computing

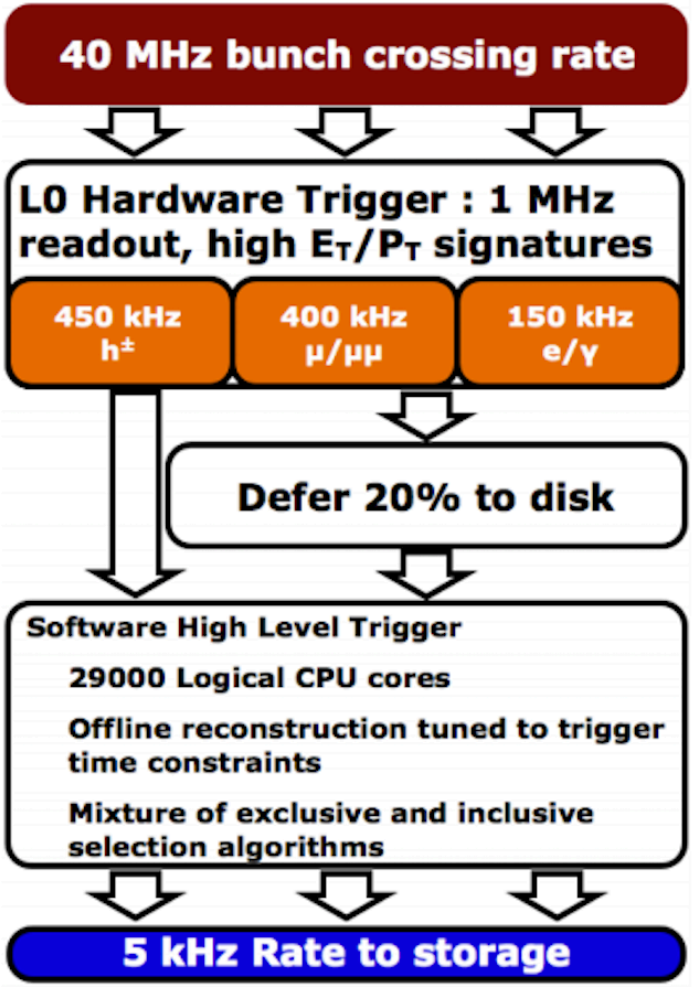
LHCb computing model changes for Run 2 and beyond



Reminder: deferred High Level Trigger in 2012

- LHC "only" delivers collisions 30% of the time
  - Trigger farm idle 70% of the time
- Farm had ~1 PB storage distributed on farm nodes which was not being used
  - Overcommit HLT by 20-30%, buffer the events on the local disks and catch up during the breaks

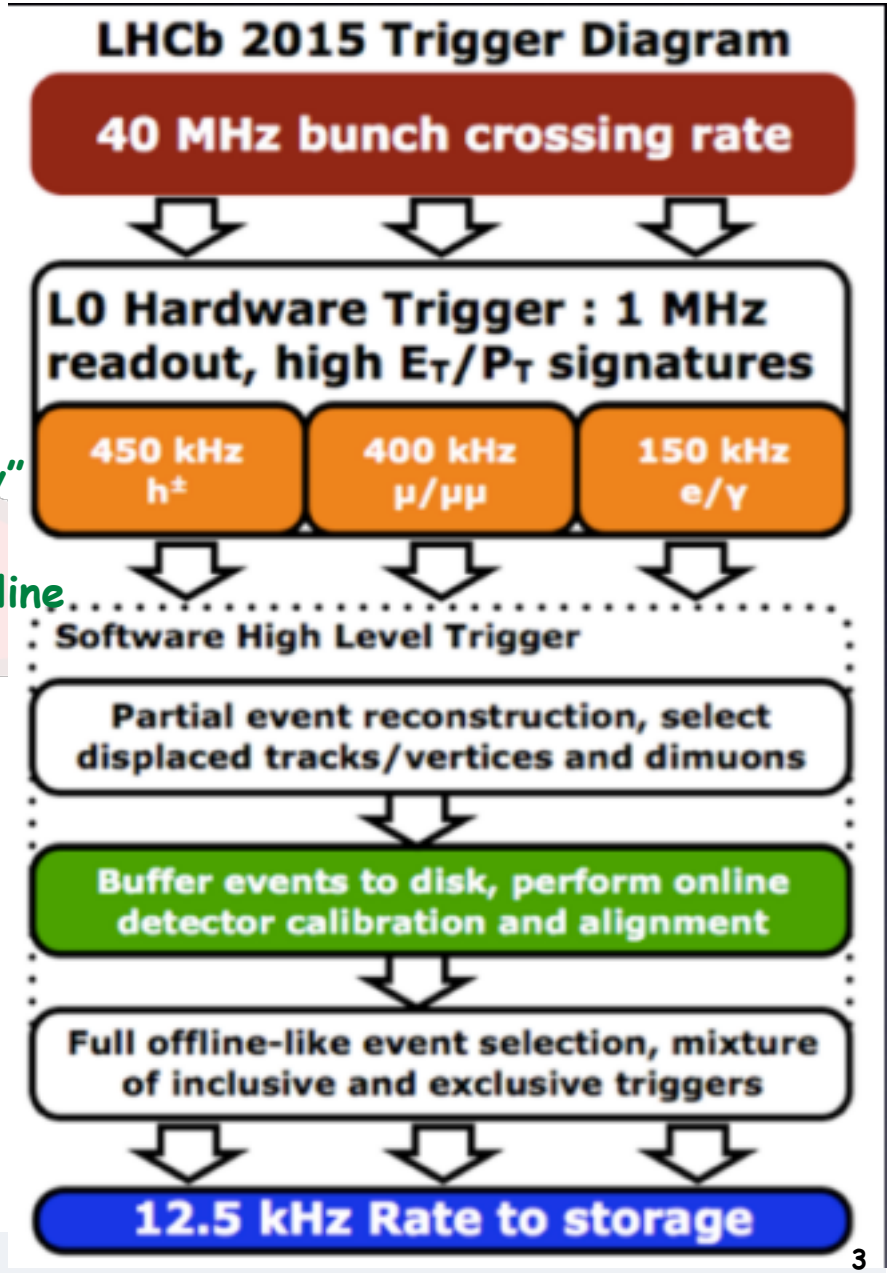




- Gains 20% extra CPU time
  - Lower track reconstruction thresholds
  - Higher trigger efficiency

Run 2: Split HLT

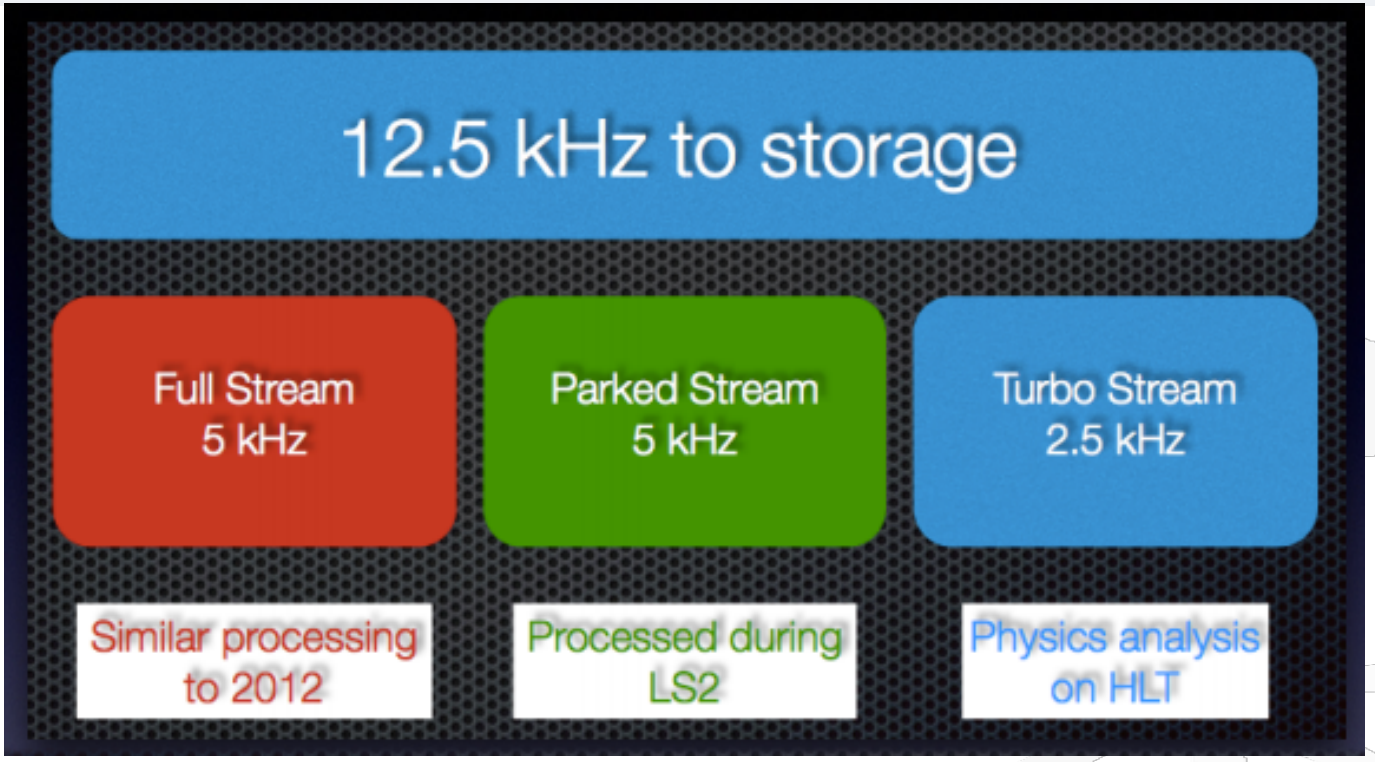
- Logical extension of 2012 deferred trigger:
  - Can we defer processing of ALL events, so we can apply atmospheric pressure dependent Rich refractive index calibration?
    - Allows to use "offline quality" particle ID in the HLT
    - HLT selections closer to offline
      - Higher trigger efficiency
      - More bandwidth for signal
- Split HLT strategy:
  - HLT1 in real time
    - Buffer HLT1 output on disk
    - Wait for Rich calibration
  - HLT2 fully deferred



Run 2: Automated online calibration

- 2012 experience:
  - All calibrations and alignments could be determined in semi-automated fashion before running the offline reconstruction
    - Typical delay 2-3 weeks to run jobs on the grid, verify output, propagate changes to CondDB
- Run 2 strategy:
  - Automate the execution of the alignment and calibration jobs
    - Execute in real time in Online Farm on HLT1 triggered data
  - Automate the validation of the new constants
    - For Rich refractive index:
      - Apply to HLT2 and start processing deferred events
    - For other calibrations:
      - If significant change detected, change run and apply to all HLT
  - Automate the propagation of constants to CondDB
    - Calibration and Alignment available to offline at same time as to HLT
    - Offline reconstruction can run promptly on RAW data
      - No need to buffer RAW data for 2-3 weeks waiting for calibration

Run 2: Reconstruction streams



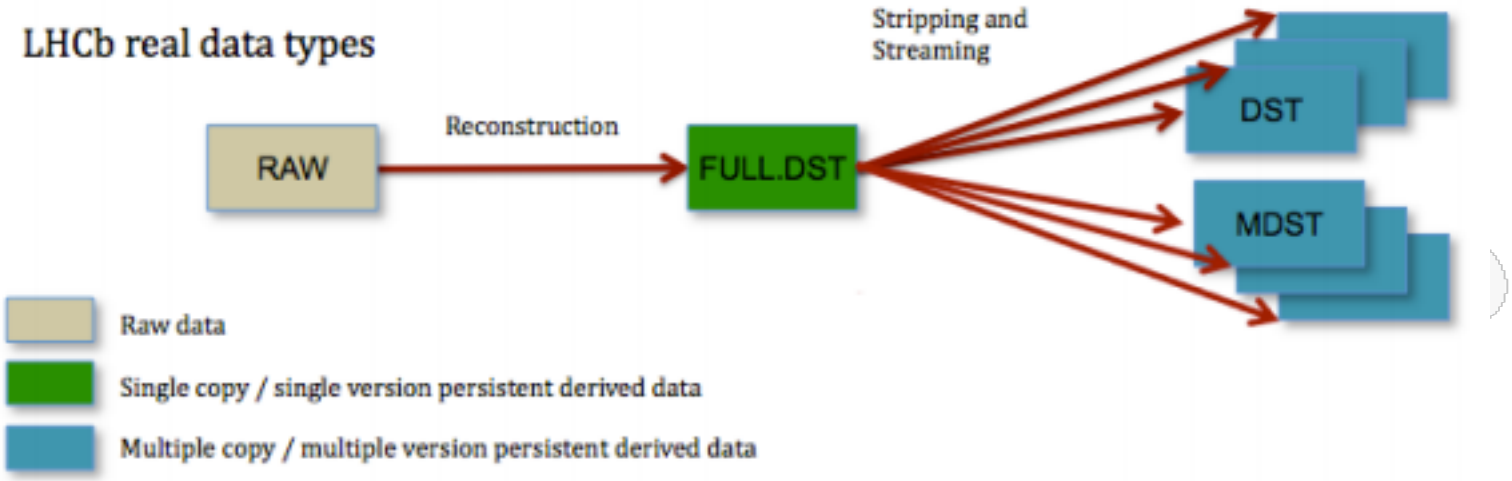
- Full stream: prompt reconstruction as soon as RAW data appears offline
- Parked stream: safety valve, probably not needed in 2015
- Turbo stream: no offline reconstruction, analysis objects produced in HLT
  - Important test for Run3

Run 2: Reconstruction

- In 2015, prompt offline reconstruction will have same quality as most recent 2011-2012 reprocessing
  - Consequence of automatic online calibration and alignment
- Goal: prompt reconstruction will be the ONLY reconstruction
  - Publication quality data within hours of data taking
    - In case of problems with calibration, spotted in online monitoring, can delay reconstruction waiting for fix
  - No reprocessing until LS2
    - Saves factor 2 in CPU per GB of RAW compared to Run 1
    - Smooths out peaks in CPU power requirement
  - No access to RAW tapes until LS2
    - Prompt processing while RAW data still on staging disk
- In practice:
  - Will need some iterations while detector is debugged
    - Partial reprocessings of the early data

Run 2: Stripping

- Workflow unchanged since Run 1:



- Due to PID in trigger and higher CPU budget for HLT2, HLT output is purer in signal
  - Higher stripping retention (>50% in Run 2, ~40% in run 1)
  - Coupled with factor 2 in HLT rate:
    - 2-3 times more events to disk per LHC second
- Need large reduction in event size
  - Goal: >90% of selections on microDST (MDST)
    - MDST: ~8 kB/event, c.f. DST: ~100kB/event
    - In Run 1: ~70% of selections on MDST

2015: Re-stripping and Recalibration

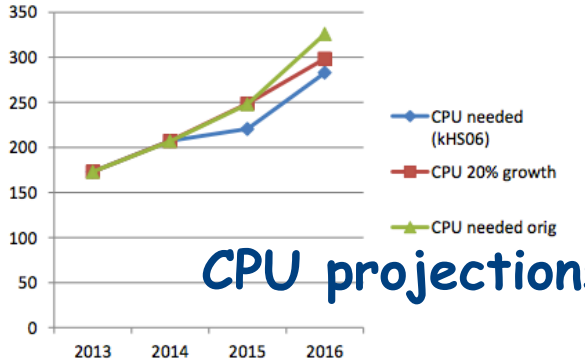
- Re-stripping: re-access FULL.DST to make new selection
  - Heavy tape load (FULL.DST is tape resident)
    - Throughput limited by tape access
      - e.g. 4 PB in 6 weeks for Run 1 re-stripping
  - Limit to ~twice per year (as in Run 1)
- Re-calibration
  - LHCb DST format allows for recalibration of e.g. flavour tagging at analysis stage
      - Also, definition of new variables e.g. for jet isolation cuts
    - In general not possible on MDST
  - MDST.DST
    - New stream containing DST version of all stripping lines that have gone to MDST and are sensitive to calibration
      - Allows to regenerate MDST with new calibration or new variables
      - DST format, but with reduced RAW content, ~50kB/evt
    - Single tape copy (not for user access)
      - Initially also one disk copy
    - Access ~once per month in centralised production
      - Lighter than re-stripping, <25% size of FULL.DST

Run 2: Simulation

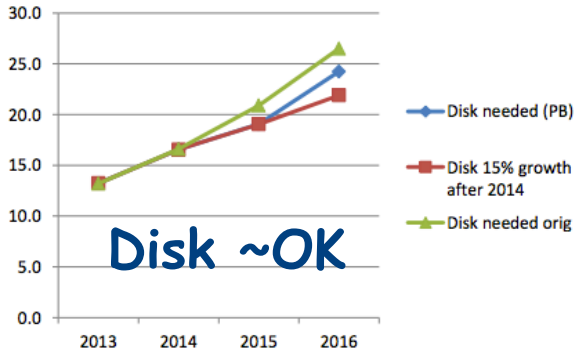
- Reminder: LHCb simulation can run anywhere
  - Take advantage of opportunistic resources
  - For Run 2, factor ~2 increase in simulation CPU needs c.f. Run 1
- Unlike other experiments, cannot use HLT farm during data-taking
  - Farm used between fills for deferred HLT processing
    - Smaller farm than would have been needed for synchronous HLT
- Exploit other resources, going beyond WLCG
  - Yandex
  - "Supercomputers"
  - Volunteer computing (BOINC)

Comparison with "flat budget"

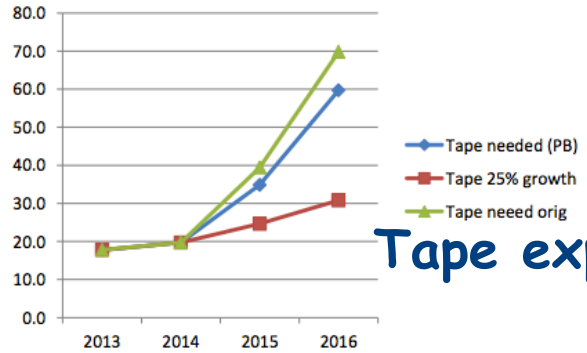
- Definition of flat budget: same money will buy
  - 20% more CPUs
  - 15% more disk
  - 25% more tape



CPU projections ~OK to >2016



Disk ~OK



Tape explodes

- Tape requirement driven by:
  - Two copies of RAW
    - Incompressible, but ~never accessed
  - One copy FULL.DST
    - Contains also RAW. Looking at ways to improve this
  - Two copies analysis datasets (including MC)
    - For Data Preservation. Can we afford it?

Output rate

- We're now drinking from a firehose:

- We can trigger on as much signal as we want within our timing constraints
- But can we keep it? For TDR, 3 output bandwidth scenarios defined:
  - 2 GB/s (20 kHz):
    - Reasonably efficient Topo (10 kHz), approx Run 1 performance
    - Tight exclusive beauty lines
    - Limited exclusive charm selections
  - 5 GB/s (50 kHz):
    - Better exploitation of Topo (25 kHz): SL efficiency $\rightarrow 1.5\times$, hadronic up to $4\times$
    - Comfortable beauty exclusives + LT unbiased
    - Exclusive + some tight inclusive charm
  - 10 GB/s (100 kHz):
    - Fully exploited Topo (50 kHz), even higher hadronic efficiency
    - Beauty and charm exclusives, inclusives
    - LT unbiased charm

Selection	Output rate [kHz]		
Topo	10	25	50
LT unbiased	1	4	5
Excl. b	ϵ	1	3
Incl. di-μ	-	-	2
Charm	9	20	40
Total	20	50	100

- In any scenario we will need to be creative to maximise charm to tape: Trigger-level ntuples, data parking, etc. Final decision planned for 2018.

C. Fitzpatrick

17/06/2014

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Run3: TurboDST?

- Reminder: 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 from the 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 an idea, but would allow a 100kHz HLT output rate without order of magnitude more computing resources.

Conclusions

- Major LHCb change for run 2 is in HLT
  - No major changes in offline
- Main consequence for computing model is simplification of processing model
  - No reprocessing before LS2
  - Minimal access to RAW tape until LS2
    - After initial debugging period
- Evolution of data model towards smaller event sizes for physics analysis
  - Generalised use of MDST
    - With MDST.DST as backup
  - Investigation of Turbo DST for Run 3
- Worry about tape requirement
  - Used mainly for archive and data preservation