LHCb computing in Run 2 and its evolution towards Run 3

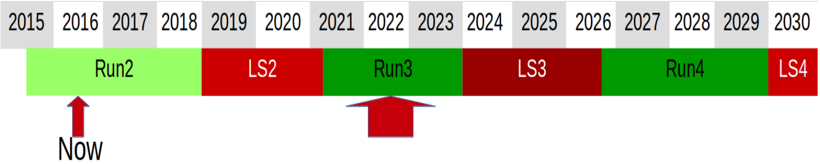
Antonio Falabella - On behalf of the LHCb collaboration

INFN - CNAF (Bologna)

38th *International Conference on High Energy Physics*

3 -10 August - Chicago, Usa

LHCb Upgrade time schedule



Run 2

- Currently running successfully collecting Run 2 data
- Luminosity $4 \cdot 10^{32}$

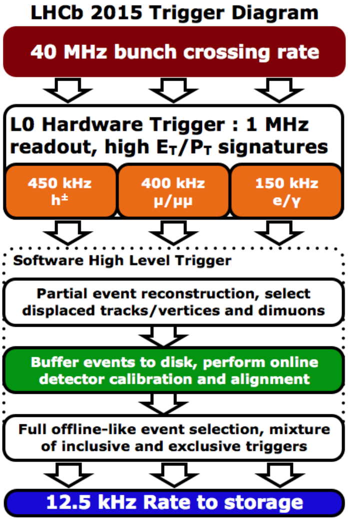
Run 3

- Upgrade foreseen for Run 3 data taking in 2021
- Luminosity $2 \cdot 10^{33}$ 5 times higher
  - $\rightarrow$ pile-up increases
- Trigger rate 5 times higher
- RAW data size 2 times

LHCb Run 1 Computing Model

- RAW data: 1 copy at CERN, 1 copy distributed (7 Tier1s)
    - First pass reconstruction runs democratically at CERN+Tier1s
    - End of year reprocessing of complete years dataset
- Each reconstruction followed by a *stripping* step
    - Event selections by physics groups, several 1000s selections in ~ 10 streams
- Stripped DSTs distributed to CERN and all 7 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
- Disk located at CERN and Tier1s (Tier2s increasingly providing storage)

Run 2 trigger configuration



LHCb 2015 Trigger Diagram

40 MHz bunch crossing rate

L0 Hardware Trigger : 1 MHz readout, high E_T/P_T signatures

450 kHz $h^{\pm}$	400 kHz $\mu/\mu\mu$	150 kHz e/γ

Software High Level Trigger

Partial event reconstruction, select displaced tracks/vertices and dimuons

Buffer events to disk, perform online detector calibration and alignment

Full offline-like event selection, mixture of inclusive and exclusive triggers

12.5 kHz Rate to storage

Run 2 trigger :

- L0 hardware trigger reduces the frequency from 40 MHz to 1.1 MHz
- HLT split in HLT1 and HLT2 reduce further the frequency to 12.5 KHz
- Disk buffer between HLT1 and HLT2
- Calibration and alignment performed from HLT1 output using the first data of each LHC fill in O(min)
- HLT2 output = offline reconstruction

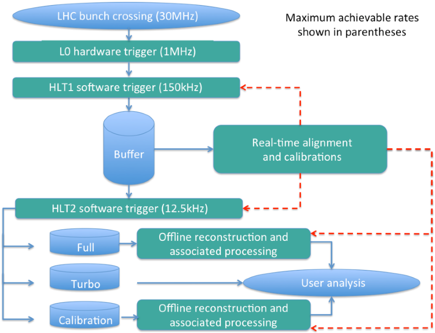
▸ Details on the trigger trigger evolution in Barbara Sciascia's yesterday talk

▸ Detailed description of Real-time Calibration and Alignment in Roel Aaij's yesterday talk

Run 2 data processing evolution

New concept, Turbo stream:

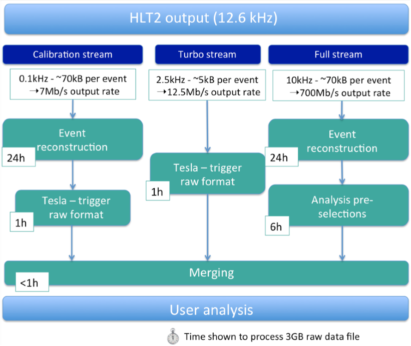
- Selected lines reconstructed and streamed at HLT level
- Special workflow for Offline processing → conversion from Online to Offline data format


LHC bunch crossing (30MHz)
L0 hardware trigger (1MHz)
HLT1 software trigger (150kHz)
Buffer
Real-time alignment and calibrations
HLT2 software trigger (12.5kHz)
Full
Offline reconstruction and associated processing
Turbo
User analysis
Calibration
Offline reconstruction and associated processing

Maximum achievable rates shown in parentheses

- Full Stream: Offline reconstruction
- Turbo Stream: Online reconstruction only. No RAW kept, no reprocessing possible
- Calibration Stream: Data driven efficiency for Online and Offline reconstruction

Run 2 data processing evolution cont'd

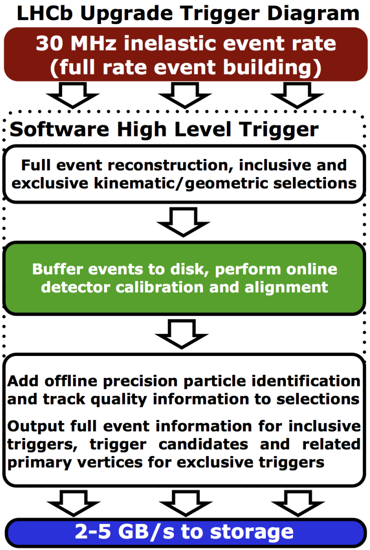

HLT2 output (12.6 kHz)
Calibration stream
Turbo stream
Full stream
0.1kHz - ~70kB per event →7Mb/s output rate
2.5kHz - ~5kB per event →12.5Mb/s output rate
10kHz - ~70kB per event →700Mb/s output rate
Event reconstruction
24h
Tesla – trigger raw format
1h
Tesla – trigger raw format
1h
Event reconstruction
24h
Analysis pre-selections
6h
Merging
<1h
User analysis
Time shown to process 3GB raw data file

- Turbo stream → events for physics analysis
- Event size 10 times smaller
- 20% of the HLT2 output sent to Turbo

Towards the LHCb Upgrade (2020)

- The strategies for the Upgrade must be carefully evaluated → trigger an order of magnitude more data with constant computing budget
- Roadmap towards Run 3
  - Q4/15: Upgrade kickoff meeting
  - Q1/16: Roadmap document for TDR
  - Q1/17: Demonstrate technology changes
  - Q4/17: Software and Computing TDR
  - Q4/18: Definition of Computing Model
- Evolution: Data Processing and Analysis, Externals(Dirac), Simulation
- Revolution: Framework and scheduling, Event Model, Non-event data, Hardware and Dataflow

Run 3 trigger configuration



LHCb Upgrade Trigger Diagram

30 MHz inelastic event rate (full rate event building)

Software High Level Trigger

Full event reconstruction, inclusive and exclusive kinematic/geometric selections

Buffer events to disk, perform online detector calibration and alignment

Add offline precision particle identification and track quality information to selections

Output full event information for inclusive triggers, trigger candidates and related primary vertices for exclusive triggers

2-5 GB/s to storage

Run 3 trigger:

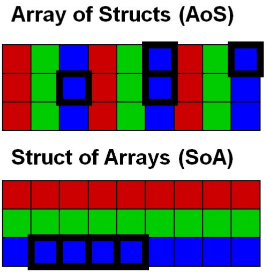
- No hardware trigger (30 MHz is the crossing frequency with non-empty events)
- Software only trigger, only signal events
- Stripping = Streaming
- Turbo will be the default stream
- Little data processing to be done offline
- Much more data selected → proportional growth of MC productions

Framework and scheduling

- LHCb framework based on Gaudi and will continue
- Current only support for single-thread execution
- cache misses become an issue
- must improve SIMD support to profit of modern CPU features
- Task concurrency → prototype exists : GaudiHive
  - Algorithms as reentrant entities that can run in parallel
  - Data immutable → reflects on Event Model

Event Model

- Current implementation based on AoS memory layout make difficult to profit of SIMD features of modern CPUs
- Event model object must be composable
- use single precision where possible


Array of Structs (AoS)

Struct of Arrays (SoA)

- Parallelization on selected algorithms:
  - Forward Track finding, track fitting
  - Investigation of Plain Old Data(POD) FCC event model

Non-Event data

- Detector description (LHCbDD):
    - Current implementation not thread-safe
    - XML persistency format
    - DD4Hep toolkit candidate to replace LHCbDD
- Conditions database (CondDB):
    - DB developed in collaboration with ATLAS
    - Suffer similar problem of LHCbDD
    - Interval of validity in nanoseconds
    - Replace XML with something more compact and fast to read

Hardware and Dataflow

Several new hardware technologies available that require study

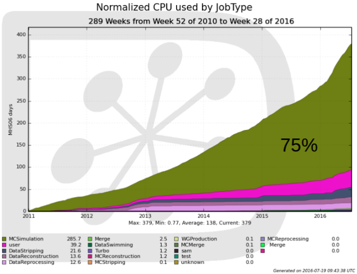
- x86_64
- Knights Landing (KNL)
- GPGPU
- ARM64

Main algorithms under investigation

- Kalman filter
- Forward tracking
- RICH reconstruction

Simulation

- Simulation currently accounts for the $\sim$ 75% of the distributed computing resources
  - CPU resources increase by $\sim$ 20% per year, Run 3 at 5 times luminosity and pile-up



- Collaboration with FCC group on *Gaussino*: an experiment independent version of LHCb Gauss simulation framework
- Simulation requires orders of magnitude of CPU time improvement: fast simulation, code optimization, partial simulation

External constraints (Dirac)

- DIRAC scalability must be considered with respect to:
    - Traffic growth
    - Dataset growth
    - Maintainability
- Introduce message queues to scale with the foreseen increase of traffic
- Backend databases other than MySQL:
    - Support for NoSQL will allow for real time monitoring
    - Block storage support
- Improve functionality testing, add performance testing
- Improve documentation and user support
- → Evolution not revolution

Distributed computing and Analysis

- Run 1 and Run 2 dataflows will not scale to Run 3: RAW storage become expensive, stripping does not scale anymore
- Turbo stream default in Run 3
  - Stripping can be avoided
- New data formats to add between μDST and DST
- Centralised NTuple production
  - Currently produced by users
  - organize them in "trains"
  - WGs would provide the scripts, some experience made during Run 2

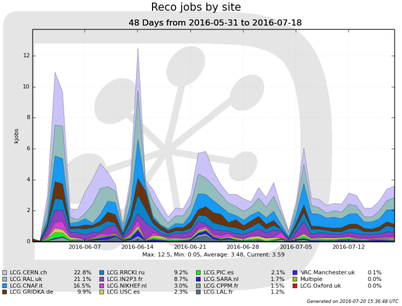
Summary

- Run 2 can be considered as a testbed for Run 3, novel data processing concepts already in place
  - Turbo stream
- Trigger an order of magnitude more data requires a careful evaluation of the suitable technologies
- Some aspects of the computing model will be redesigned, some other will be adapted
- In particular major changes are foreseen for the event model and framework
- Adapt also to the evolution of computing infrastructures
- Tight schedule but clearly defined

Backup slides

Mesh Processing

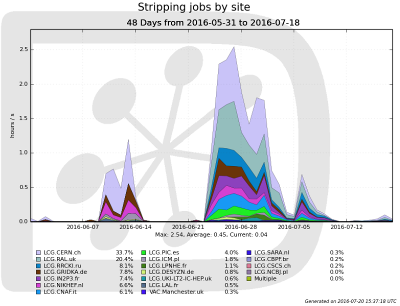
- Reconstruction and Stripping jobs extented to Tier2 sites
  - Tier2 no longer associated with a single Tier1 → Mesh concept



Reco jobs by site

48 Days from 2016-05-31 to 2016-07-18

Generated on 2016-07-20 15:36:48 UTC



Stripping jobs by site

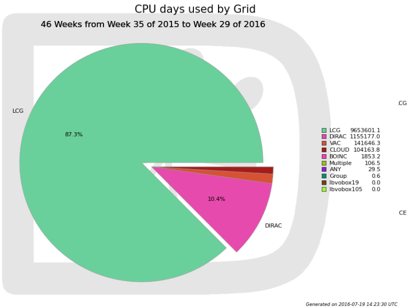
48 Days from 2016-05-31 to 2016-07-18

Generated on 2016-07-20 15:37:18 UTC

- T2 storage:
  - T2 LHCb requirement for 2016 is 2.8PB, 3.8PB for 2017
  - T2-2A + storage element concept: Access to local data or remote if suitable network

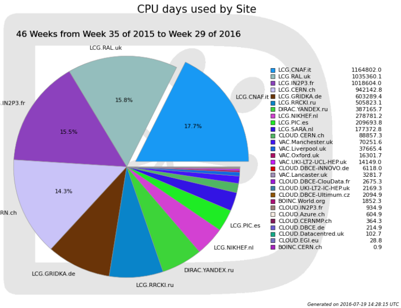
Volunteer Computing and Cloud

- Increasing usage of Non-WLCG resources
- Small fraction of volunteer computing resources: Beauty @LHC project
    - Authentication via trusted machine
    - the VM certifcate can talk only with gateway machine


CPU days used by Grid

46 Weeks from Week 35 of 2015 to Week 29 of 2016

Generated on 2016-07-19 14:23:30 UTC

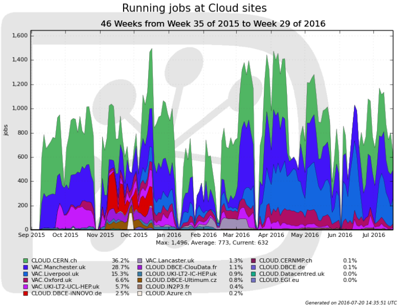

CPU days used by Site

46 Weeks from Week 35 of 2015 to Week 29 of 2016

Generated on 2016-07-19 14:28:15 UTC

Cloud resources

- Cern OpenStack still the largest provider
- Manchester and Liverpool provide considerable fraction
- Only CPU not for storage


Running jobs at Cloud sites
46 Weeks from Week 35 of 2015 to Week 29 of 2016
jobs
Sep 2015
Oct 2015
Nov 2015
Dec 2015
Jan 2016
Feb 2016
Mar 2016
Apr 2016
May 2016
Jun 2016
Jul 2016
Max: 1,496, Average: 773, Current: 632
CLOUD.CERN.ch 36.2%
VAC.Manchester.uk 28.7%
VAC.Liverpool.uk 15.3%
VAC.Oxford.uk 6.6%
VAC.UKI-LT2-UCL-HEP.uk 5.7%
CLOUD.DBCE-iNNOVO.de 2.5%
VAC.Lancaster.uk 1.3%
CLOUD.DBCE-ClouData.fr 1.1%
CLOUD.UKI-LT2-IC-HEP.uk 0.9%
CLOUD.DBCE-Ultimum.cz 0.8%
CLOUD.IN2P3.fr 0.4%
CLOUD.Azure.ch 0.2%
CLOUD.CERNMP.ch 0.1%
CLOUD.DBCE.de 0.1%
CLOUD.Datacentred.uk 0.0%
CLOUD.EGI.eu 0.0%
Generated on 2016-07-20 14:35:51 UTC