Operational experience with the new ATLAS HLT framework for LHC Run 3

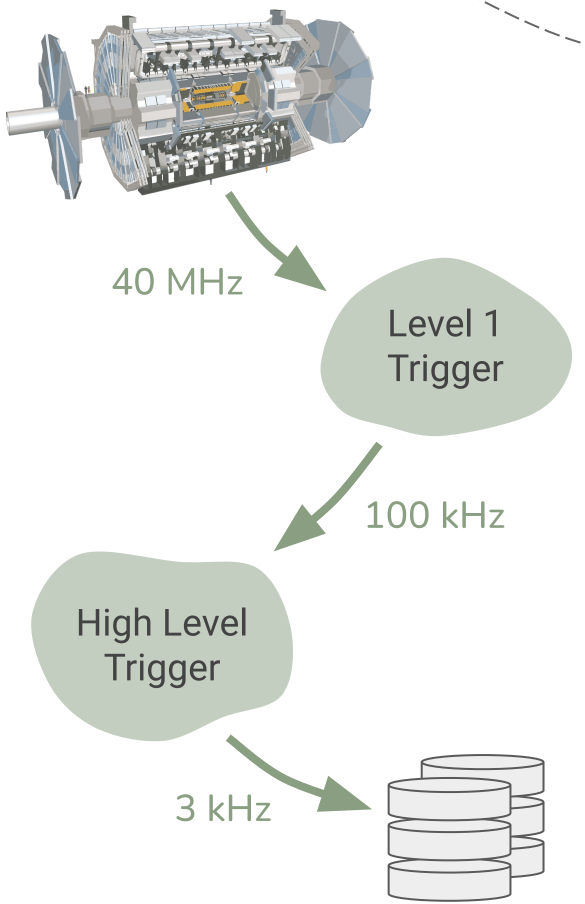
Aleksandra Poręba
(CERN / Ruprecht-Karls-Universität Heidelberg)
on behalf of ATLAS Collaboration
aleksandra.poreba@cern.ch

CHEP 2023, 08.05.2023

UNIVERSITÄT HEIDELBERG
ZUKUNFT
SEIT 1386

Federal Ministry
of Education
and Research

Introduction



40 MHz

Level 1 Trigger

100 kHz

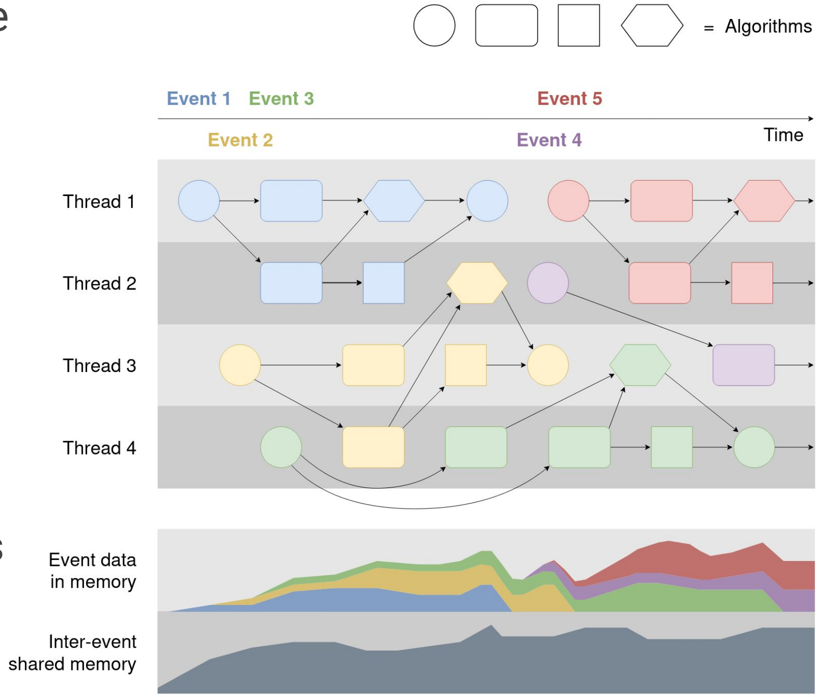
High Level Trigger

3 kHz

- The ATLAS **High Level Trigger** selects events based on
  - Detector readout
  - Level 1 Trigger decision, applying a coarse selection
- Event selection is achieved by a set of selection **chains**
- Early algorithms within a chain reject as many events as possible
  - More CPU-intensive algorithms are executed only on a small subset of events
- The data is processed on a computing farm with ~60,000 real CPU cores (2023)

Run 3 Multi-Threaded HLT

- The HLT was redesigned to share the same code with offline reconstruction
  - Support the **Multi-Threaded** mode
  - Reduce the memory footprint of the code (not an issue for online operation)
- The upgrade benefits:
  - Simplified maintenance of the code,
  - General performance improvements,
  - Integration of computing accelerators for future running periods

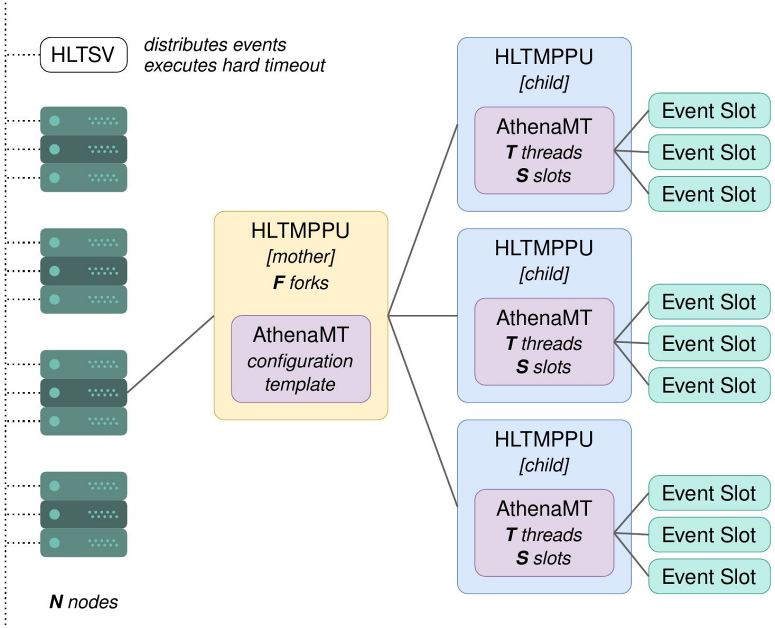


source: ATL-DAQ-SLIDE-2019-738

Online performance

The configuration of the HLT Processing Unit and its CPU resource utilization is defined by three parameters:

- Number of process **forks**
- Number of **threads** within the process
- Number of event **slots** defining how many events can be executed in parallel per node

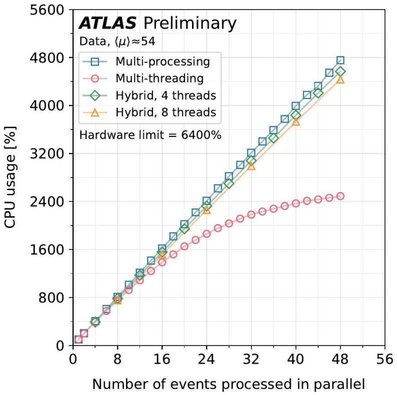


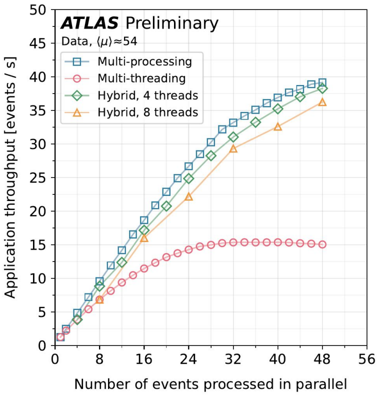
source: TriggerCoreSWPublicResults

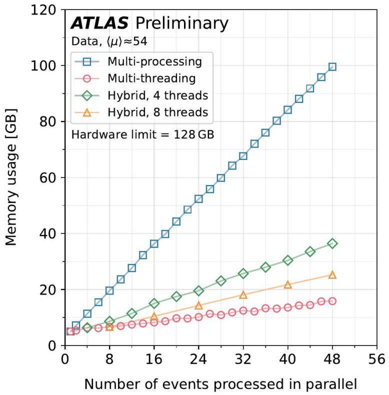
Online performance

- In 2022 the online event processing was maximized with a **pure Multi-Processing** configuration
- A **pure Multi-Threaded** configuration show lower throughput
    - It is still used for MC production, where memory savings are necessary
- **Hybrid configurations** were also considered, giving similar gains in memory usage without throughput penalty

source: TriggerCoreSWPublicResults







Commissioning - CPUs

- HLT farm was upgraded during 2022 to consist of **AMD EPYC 7302** CPUs (sixteen real cores with two hyper-threads per core), improving the total farm performance

year	2018	2022	2023
HS06	1.2M	1.7M[1]	**2.0M**

- The upgrade process was done gradually, therefore different CPU rack configurations had to be used for the old and new types of machines

[1] With 60% of racks being replaced

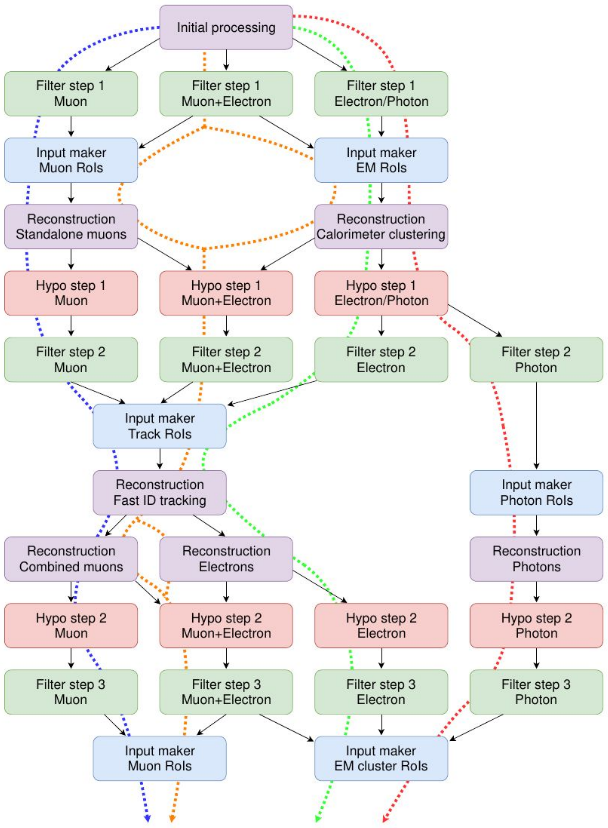
HS06

A rack of ATLAS TDAQ components
https://cds.cern.ch/record/1696907

Run 3 HLT Configuration

- The Run 3 HLT Control Flow is generated based on a list of algorithms organized in **steps**, performing reconstruction and selection
- The steps are combined in **chains** and are organized in a selection **menu**
- The configuration is stored in JSON blob format and can be provided **transparently** to HLT in different ways:
  - from a database,
  - from a file,
  - from a configuration in Python
  - from 'in-file meta-data' (mostly used for offline reconstruction)

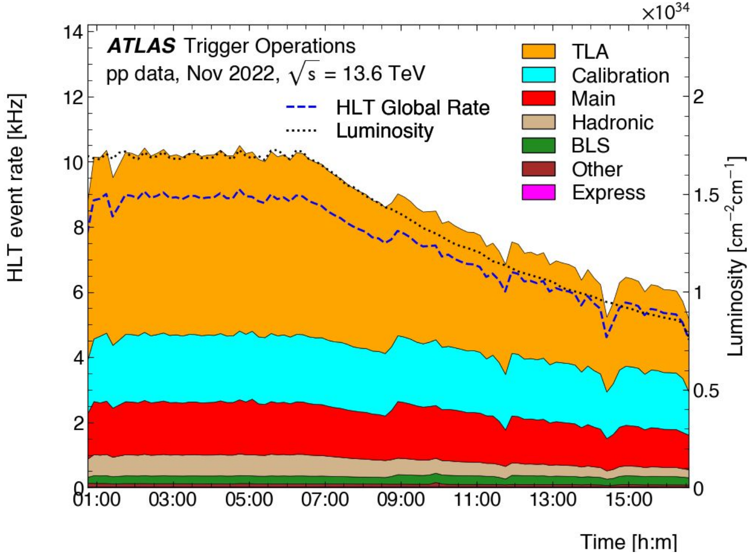
source: ATL-DAQ-PROC-2019-004



Optimizing at end-of-fill luminosity

- As the instantaneous luminosity decreases during a run, more resources (CPU, bandwidth) are available
- The **prescale factors** are adjusted throughout the run
  - Based on preliminary performance studies of selection's cost
  - Some of the chains are enabled only in the end of the run
- The configuration changes are visible in the recorded rates of output streams

source TriggerOperationPublicResults

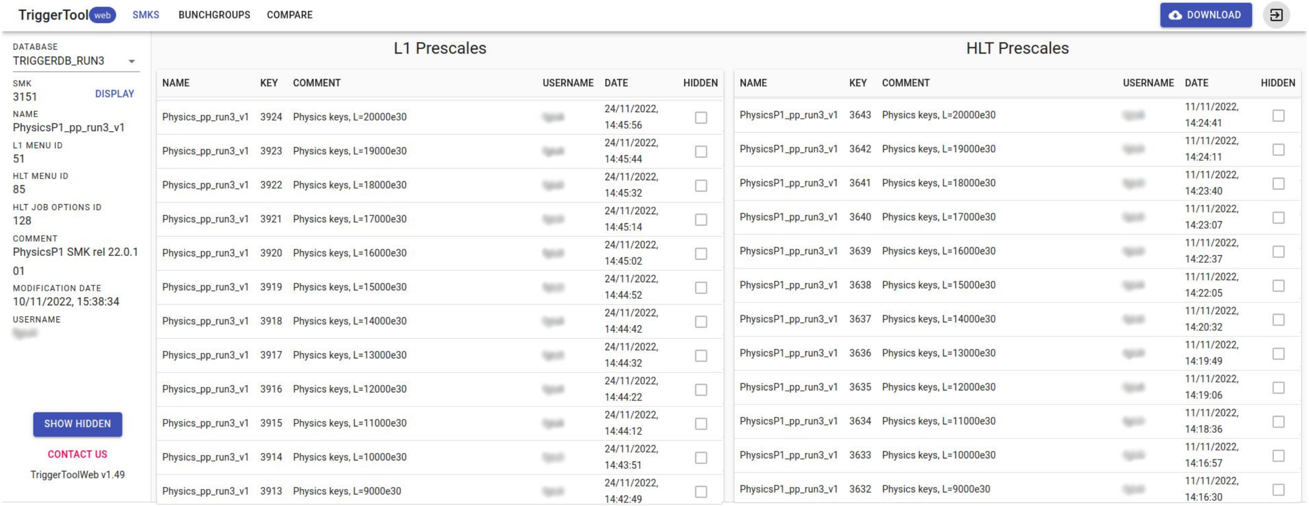


Prescale factor - value associated with HLT chains and L1 items, with prescale of *n* chain will be activated in *one in n* events

Trigger configuration

- The Trigger Configurations are available on **TriggerToolWeb**, a modern web based application, which was developed for Run 3 to replace a Java application
- The application enables display, comparison and modification of the configuration, widely used during the online operation

source: TriggerToolWeb



DATABASE
TRIGGERDB_RUN3

SMK
3151 DISPLAY

NAME
PhysicsP1_pp_run3_v1

L1 MENU ID
51

HLT MENU ID
85

HLT JOB OPTIONS ID
128

COMMENT
PhysicsP1 SMK rel 22.0.1 01

MODIFICATION DATE
10/11/2022, 15:38:34

USERNAME
[redacted]

SHOW HIDDEN

CONTACT US

TriggerToolWeb v1.49

L1 Prescales

NAME	KEY	COMMENT	USERNAME	DATE	HIDDEN
Physics_pp_run3_v1	3924	Physics keys, L=20000e30	[redacted]	24/11/2022, 14:45:56	□
Physics_pp_run3_v1	3923	Physics keys, L=19000e30	[redacted]	24/11/2022, 14:45:44	□
Physics_pp_run3_v1	3922	Physics keys, L=18000e30	[redacted]	24/11/2022, 14:45:32	□
Physics_pp_run3_v1	3921	Physics keys, L=17000e30	[redacted]	24/11/2022, 14:45:14	□
Physics_pp_run3_v1	3920	Physics keys, L=16000e30	[redacted]	24/11/2022, 14:45:02	□
Physics_pp_run3_v1	3919	Physics keys, L=15000e30	[redacted]	24/11/2022, 14:44:52	□
Physics_pp_run3_v1	3918	Physics keys, L=14000e30	[redacted]	24/11/2022, 14:44:42	□
Physics_pp_run3_v1	3917	Physics keys, L=13000e30	[redacted]	24/11/2022, 14:44:32	□
Physics_pp_run3_v1	3916	Physics keys, L=12000e30	[redacted]	24/11/2022, 14:44:22	□
Physics_pp_run3_v1	3915	Physics keys, L=11000e30	[redacted]	24/11/2022, 14:44:12	□
Physics_pp_run3_v1	3914	Physics keys, L=10000e30	[redacted]	24/11/2022, 14:43:51	□
Physics_pp_run3_v1	3913	Physics keys, L=9000e30	[redacted]	24/11/2022, 14:42:49	□

HLT Prescales

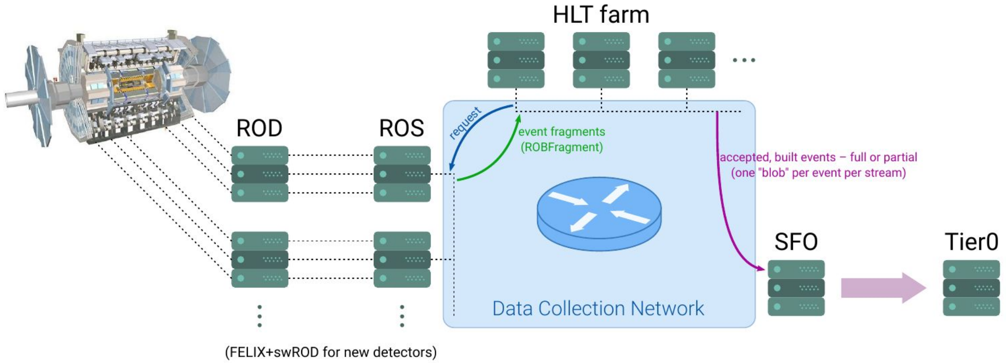
NAME	KEY	COMMENT	USERNAME	DATE	HIDDEN
PhysicsP1_pp_run3_v1	3643	Physics keys, L=20000e30	[redacted]	11/11/2022, 14:24:41	□
PhysicsP1_pp_run3_v1	3642	Physics keys, L=19000e30	[redacted]	11/11/2022, 14:24:11	□
PhysicsP1_pp_run3_v1	3641	Physics keys, L=18000e30	[redacted]	11/11/2022, 14:23:40	□
PhysicsP1_pp_run3_v1	3640	Physics keys, L=17000e30	[redacted]	11/11/2022, 14:23:07	□
PhysicsP1_pp_run3_v1	3639	Physics keys, L=16000e30	[redacted]	11/11/2022, 14:22:37	□
PhysicsP1_pp_run3_v1	3638	Physics keys, L=15000e30	[redacted]	11/11/2022, 14:22:05	□
PhysicsP1_pp_run3_v1	3637	Physics keys, L=14000e30	[redacted]	11/11/2022, 14:20:32	□
PhysicsP1_pp_run3_v1	3636	Physics keys, L=13000e30	[redacted]	11/11/2022, 14:19:49	□
PhysicsP1_pp_run3_v1	3635	Physics keys, L=12000e30	[redacted]	11/11/2022, 14:19:06	□
PhysicsP1_pp_run3_v1	3634	Physics keys, L=11000e30	[redacted]	11/11/2022, 14:18:36	□
PhysicsP1_pp_run3_v1	3633	Physics keys, L=10000e30	[redacted]	11/11/2022, 14:16:57	□
PhysicsP1_pp_run3_v1	3632	Physics keys, L=9000e30	[redacted]	11/11/2022, 14:16:30	□

Commissioning - Data prefetching

- Possible bottlenecks with data requests on the readout machines:
  - Too frequent requests to the readout machine
  - Delays over the network for too big data requests
- Implemented **data prefetching** scheme:
  - At the beginning of a selection step, an 'InputMaker algorithm' requests all necessary detector data in a **RoI** to perform reconstruction and to reduce the frequency of data requests

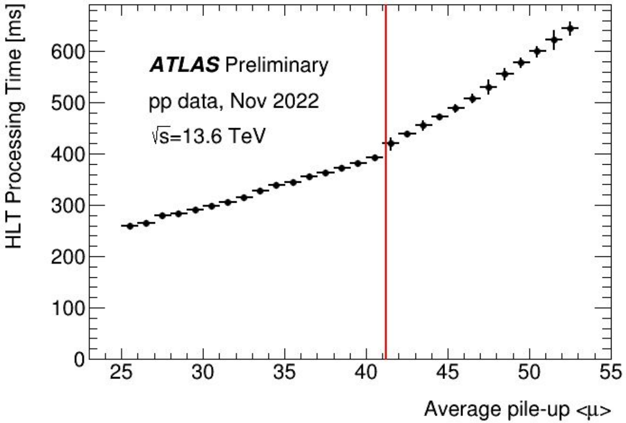
source: TDAQ Public Results

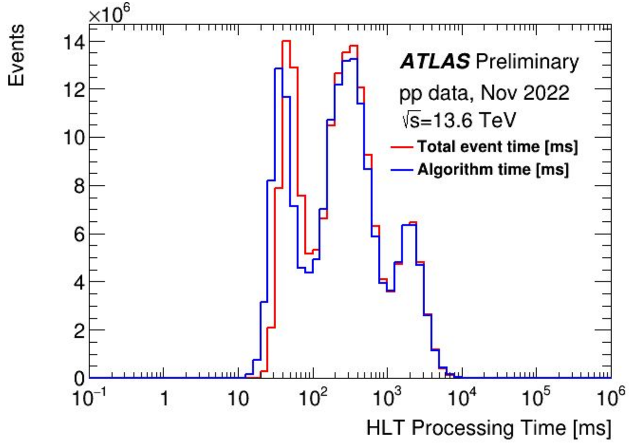
RoI - region in a detector where candidates for particles were identified by L1 Trigger and are passed to the HLT for reconstruction


HLT farm
...
ROD
ROS
request
event fragments
(ROBFragment)
accepted, built events – full or partial
(one "blob" per event per stream)
Data Collection Network
SFO
Tier0
(FELIX+swROD for new detectors)

Online performance monitoring

- **Cost Monitoring** summarizes execution of HLT in form of csv tables and ROOT histograms
- Different monitoring levels are available including performance of:
  - **Algorithms, Selection steps, Chains** e.g. execution time
  - **ReadOut Systems** e.g. readout frequency or data retrieval time
- The Cost Monitoring data is collected in parallel to physics data taking
- After post-processing the results are automatically published on a dedicated website

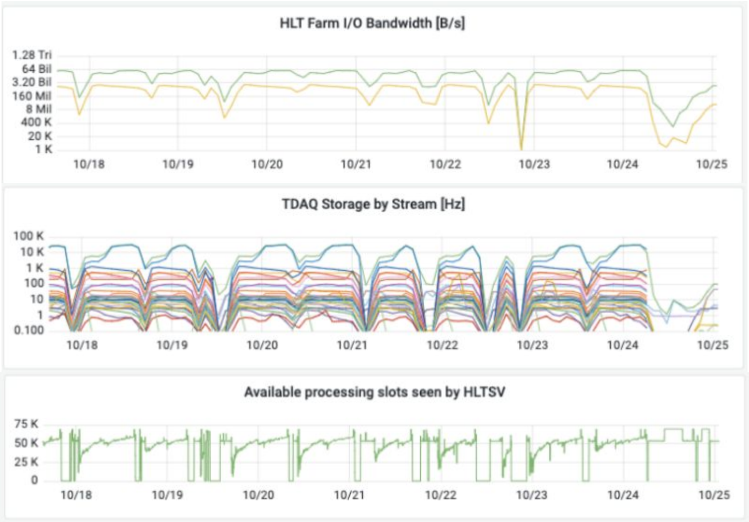




source TriggerOperationPublicResults

Trigger Rate Presenter

- Online monitoring of trigger rates allows to easily notice any performance variations during data taking
  - Occurring e.g. due to detector malfunctioning or change of beam conditions
- **Trigger Rate Presenter** - monitoring of the current trigger rates along diagnostic information including:
  - memory usage
  - performance of readout systems
- Results are displayed on web-based **Grafana** dashboards
- All data is archived with the **P-BEAST** data storage service



HLT Farm I/O Bandwidth [B/s]

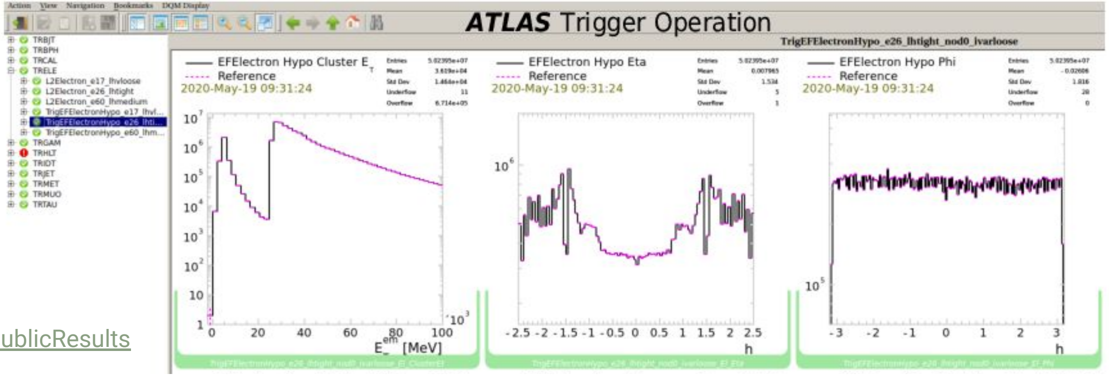
TDAQ Storage by Stream [Hz]

Available processing slots seen by HLTSV

source ATL-DAQ-PROC-2022-008

Data Quality Monitoring

- **Data Quality Monitoring Display** (DQMD) is a display of histograms with signature physics monitoring quantities
- The histograms are compared with references and problems are flagged based on the automatic tests
- Available short descriptions help shifters to take appropriate action
- The list of monitored chains in DQMD is part of the Trigger Configuration



source TriggerOperationPublicResults

Conclusions

- Redesign of HLT to framework to support the **Multi-Threaded** mode and to share reconstruction modules with offline
- HLT farm upgrade, increasing the performance to **2.0M HS06** (start of 2023)
- Read-Out System bottlenecks mitigated by **data prefetching** techniques
  - Will be resolved by the upgrade of the hardware done for 2023
- Available **tools** to assess the HLT online performance, including:
  - physics signature selection,
  - CPU resource needs
  - HLT algorithms optimization

References and Acknowledgements

Performance of the ATLAS Trigger System in 2022 The ATLAS Collaboration (paper in preparation)

Frameworks to monitor and predict rates and resource usage in the ATLAS High Level Trigger Tim Martin and on behalf of the ATLAS Collaboration 2017 J. Phys.: Conf. Ser. 898 032007 DOI 10.1088/1742-6596/898/3/032007

ATLAS Operational Monitoring Data Archival and Visualization Igor Soloviev, Giuseppe Avolio, Andrei Kazymov and Matei Vasile EPJ Web Conf., 245 (2020) 01020 DOI: https://doi.org/10.1051/epjconf/202024501020

Operation of the ATLAS trigger system in Run 2 The ATLAS collaboration 2020 JINST 15 P10004 DOI 10.1088/1748-0221/15/10/P10004

ATLAS Experiment Trigger Operation Public Results
https://twiki.cern.ch/twiki/bin/view/AtlasPublic/TriggerOperationPublicResults

ATLAS Experiment Trigger Core Software Public Results
https://twiki.cern.ch/twiki/bin/view/AtlasPublic/TriggerCoreSWPublicResults

This work has been sponsored by the Wolfgang Gentner Programme of the German Federal Ministry of Education and Research (grant no. 13E18CHA)

Backup slides

Resource utilization estimations

- CPU estimations are performed in advance of deploying a new software release online
- They are scheduled for every release to assess the changes in the CPU resource utilization
- The estimations are based on a special data sample called **Enhanced Bias**, overweighted with high-pT events which are more likely selected by Trigger, compared to events in a zero bias sample
- After applying the prepared weights and reverting the bias, the conditions are comparable to online data taking and the signature rates and CPU cost of running the new software release at the HLT can be predicted

Technical Runs

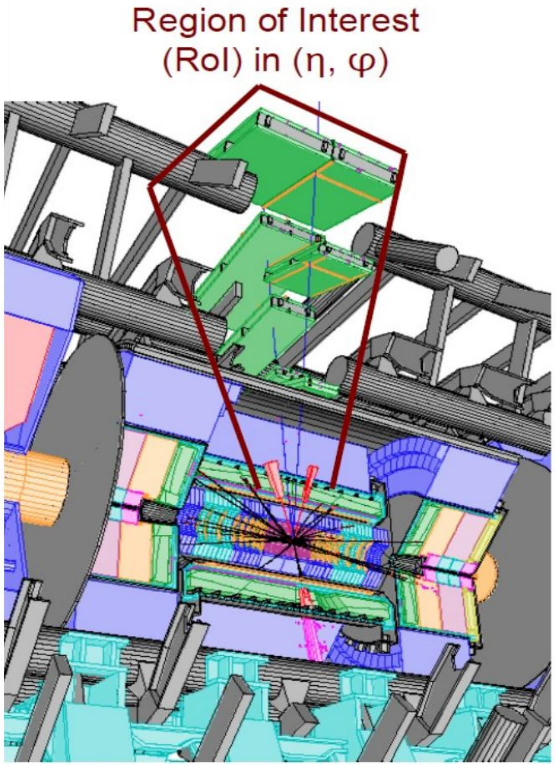
- During the commissioning the HLT software is tested in real-like conditions by executing it with preloaded data
- Tests include overnight runs to catch rare bugs, occurring only in long time running conditions, for example race conditions and to monitor the memory usage and memory leaks

Run 2 Athena MP

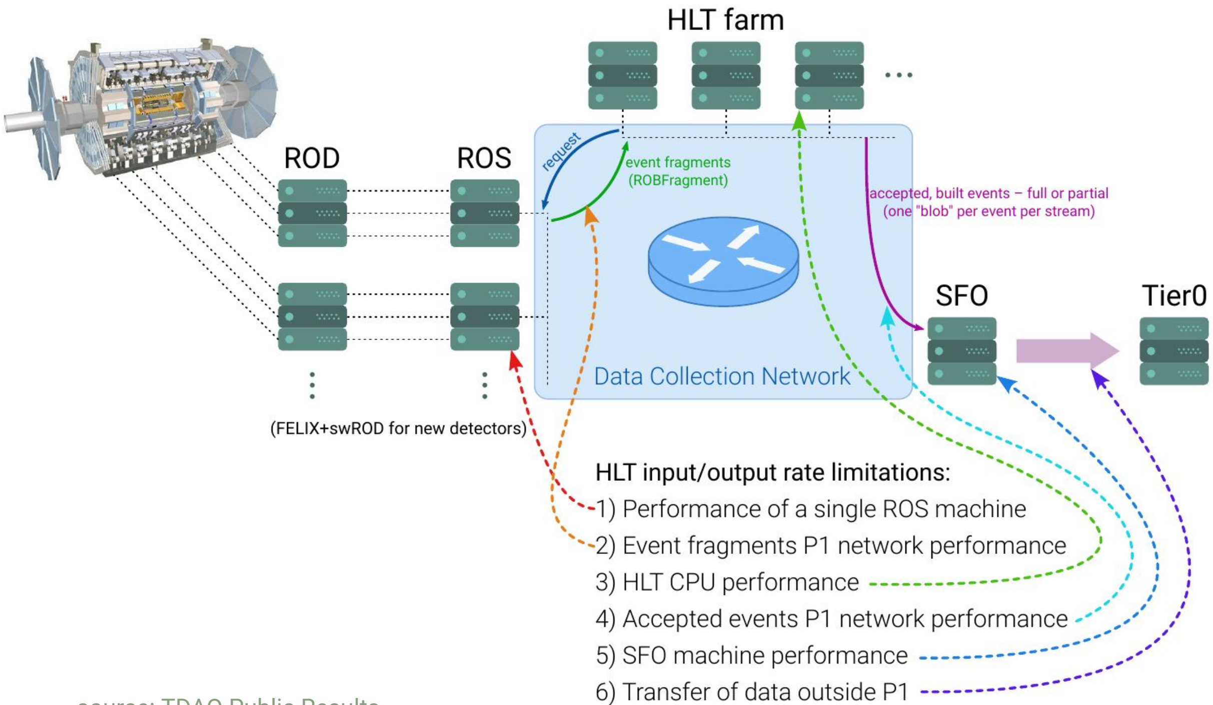
- Athena MP implemented a Multi Processing event model
- The main process is forked after initialisation, into worker processes, according to the number of events that will be processed in parallel
- Each process uses single thread, sharing the read-only memory with other workers
- Reduced memory requirements

Region Of Interest

- During the online data taking in order to save the computing resources HLT algorithms reconstruct the event only in defined Regions-of-Interest (RoI)
- They are defined as regions in the detector where candidates for particles were identified by L1 Trigger and are passed to the HLT for further analysis
- During the data reprocessing (offline) the full detector coverage is used.



Data readout bottlenecks


HLT farm
...
ROD
ROS
request
event fragments
(ROBFragment)
accepted, built events – full or partial
(one "blob" per event per stream)
SFO
Tier0
Data Collection Network
(FELIX+swROD for new detectors)

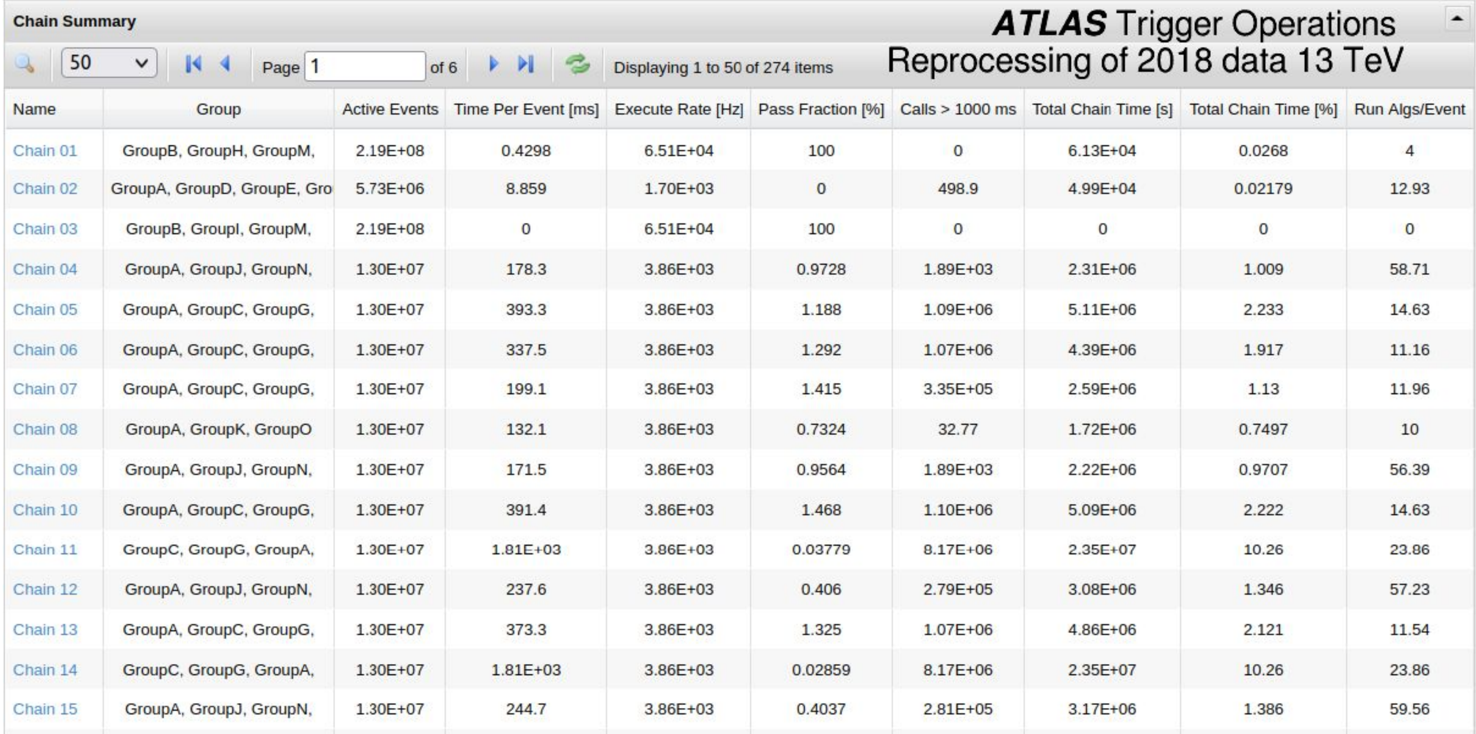
HLT input/output rate limitations:

1) Performance of a single ROS machine
2) Event fragments P1 network performance
3) HLT CPU performance
4) Accepted events P1 network performance
5) SFO machine performance
6) Transfer of data outside P1

source: TDAQ Public Results

Cost Monitoring table result example

source TriggerOperationPublicResults



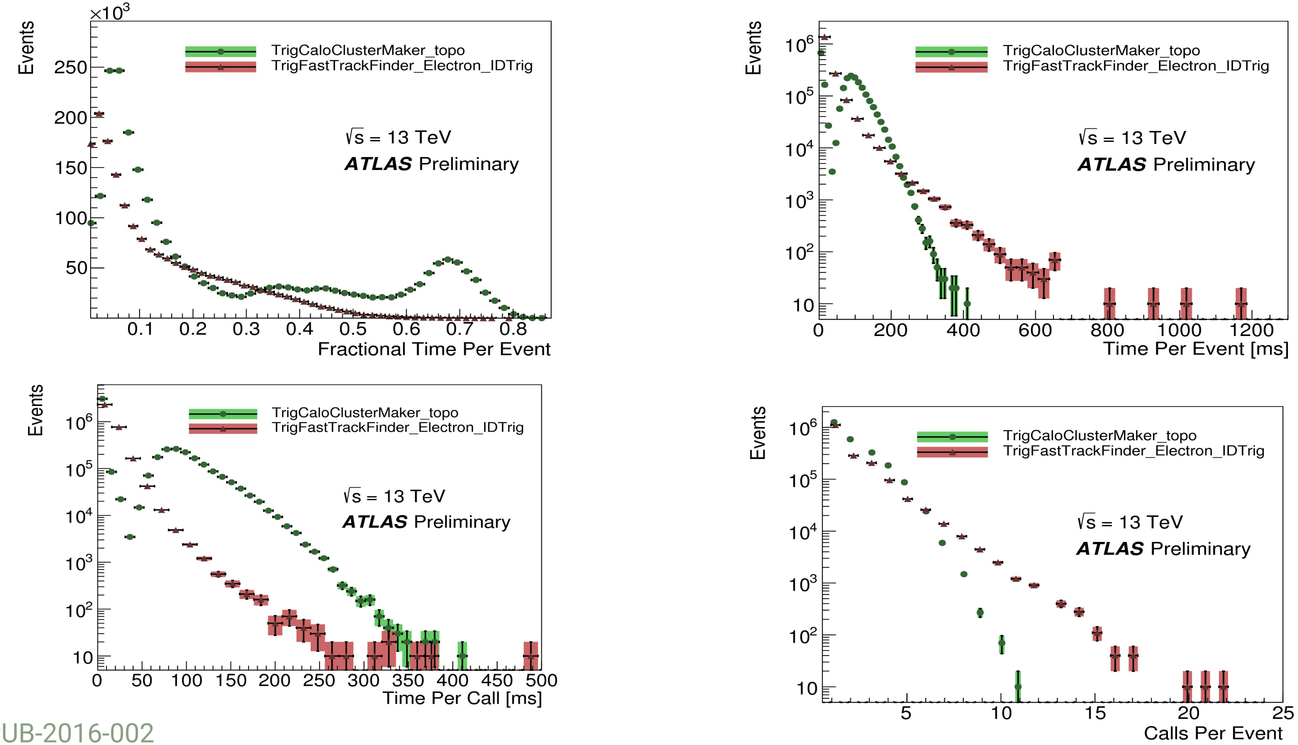
Chain Summary

ATLAS Trigger Operations
Reprocessing of 2018 data 13 TeV

50 | |◀ ◀ | Page 1 of 6 | ▶ ▶| | Displaying 1 to 50 of 274 items

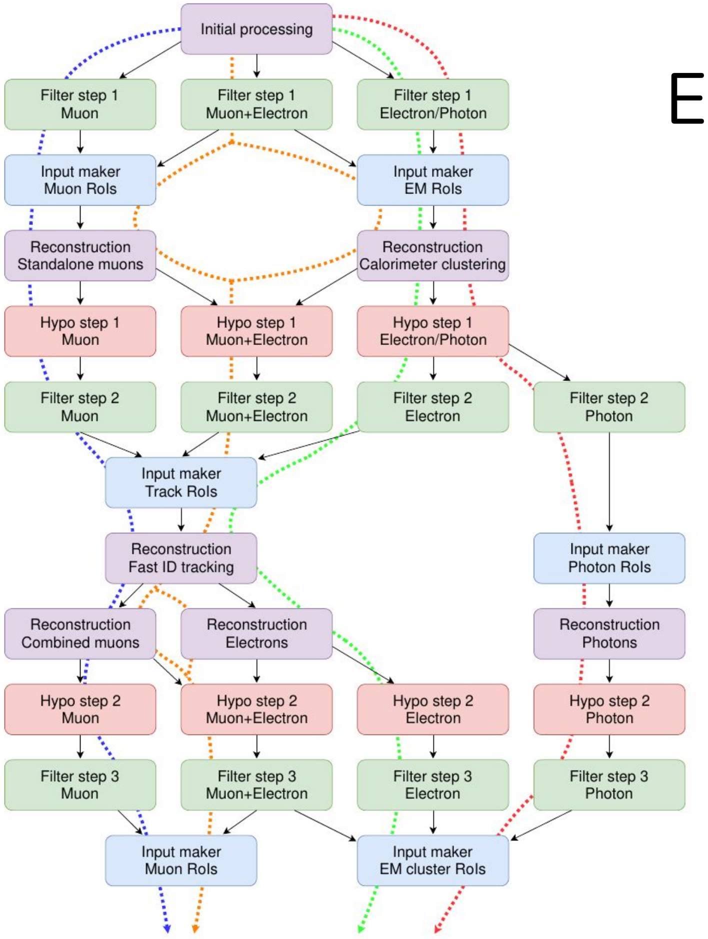
Name	Group	Active Events	Time Per Event [ms]	Execute Rate [Hz]	Pass Fraction [%]	Calls > 1000 ms	Total Chain Time [s]	Total Chain Time [%]	Run Algs/Event
Chain 01	GroupB, GroupH, GroupM,	2.19E+08	0.4298	6.51E+04	100	0	6.13E+04	0.0268	4
Chain 02	GroupA, GroupD, GroupE, Gro	5.73E+06	8.859	1.70E+03	0	498.9	4.99E+04	0.02179	12.93
Chain 03	GroupB, GroupI, GroupM,	2.19E+08	0	6.51E+04	100	0	0	0	0
Chain 04	GroupA, GroupJ, GroupN,	1.30E+07	178.3	3.86E+03	0.9728	1.89E+03	2.31E+06	1.009	58.71
Chain 05	GroupA, GroupC, GroupG,	1.30E+07	393.3	3.86E+03	1.188	1.09E+06	5.11E+06	2.233	14.63
Chain 06	GroupA, GroupC, GroupG,	1.30E+07	337.5	3.86E+03	1.292	1.07E+06	4.39E+06	1.917	11.16
Chain 07	GroupA, GroupC, GroupG,	1.30E+07	199.1	3.86E+03	1.415	3.35E+05	2.59E+06	1.13	11.96
Chain 08	GroupA, GroupK, GroupO	1.30E+07	132.1	3.86E+03	0.7324	32.77	1.72E+06	0.7497	10
Chain 09	GroupA, GroupJ, GroupN,	1.30E+07	171.5	3.86E+03	0.9564	1.89E+03	2.22E+06	0.9707	56.39
Chain 10	GroupA, GroupC, GroupG,	1.30E+07	391.4	3.86E+03	1.468	1.10E+06	5.09E+06	2.222	14.63
Chain 11	GroupC, GroupG, GroupA,	1.30E+07	1.81E+03	3.86E+03	0.03779	8.17E+06	2.35E+07	10.26	23.86
Chain 12	GroupA, GroupJ, GroupN,	1.30E+07	237.6	3.86E+03	0.406	2.79E+05	3.08E+06	1.346	57.23
Chain 13	GroupA, GroupC, GroupG,	1.30E+07	373.3	3.86E+03	1.325	1.07E+06	4.86E+06	2.121	11.54
Chain 14	GroupC, GroupG, GroupA,	1.30E+07	1.81E+03	3.86E+03	0.02859	8.17E+06	2.35E+07	10.26	23.86
Chain 15	GroupA, GroupJ, GroupN,	1.30E+07	244.7	3.86E+03	0.4037	2.81E+05	3.17E+06	1.386	59.56

Cost Monitoring histogram result example (Run 2)



source: ATL-DAQ-PUB-2016-002

Example HLT Control Flow Graph



Initial processing

Filter step 1 Muon

Filter step 1 Muon+Electron

Filter step 1 Electron/Photon

Input maker Muon RoIs

Input maker EM RoIs

Reconstruction Standalone muons

Reconstruction Calorimeter clustering

Hypo step 1 Muon

Hypo step 1 Muon+Electron

Hypo step 1 Electron/Photon

Filter step 2 Muon

Filter step 2 Muon+Electron

Filter step 2 Electron

Filter step 2 Photon

Input maker Track RoIs

Reconstruction Fast ID tracking

Input maker Photon RoIs

Reconstruction Combined muons

Reconstruction Electrons

Reconstruction Photons

Hypo step 2 Muon

Hypo step 2 Muon+Electron

Hypo step 2 Electron

Hypo step 2 Photon

Filter step 3 Muon

Filter step 3 Muon+Electron

Filter step 3 Electron

Filter step 3 Photon

Input maker Muon RoIs

Input maker EM cluster RoIs

→ Data dependencies
define how algorithms are scheduled

⋯ Trigger chains
correspond to different paths through the fixed control flow diagram

🟩 Filter algorithms
run at the start of each step and implement the early rejection

🟦 Input maker algorithms
restrict the following reconstruction to a region of interest

🟪 Reconstruction algorithms
process detector data to extract features

🟥 Hypothesis algorithms
execute hypothesis testing (e.g. $p_T > 10$ GeV) for all active chains

source: ATL-DAQ-PROC-2019-004