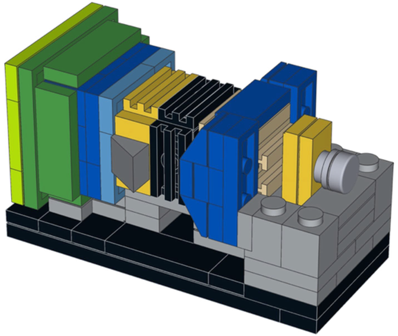
High-Throughput Machine Learning Inference with NVIDIA TensorRT



Maarten van Veghel on behalf of the LHCb RTA project

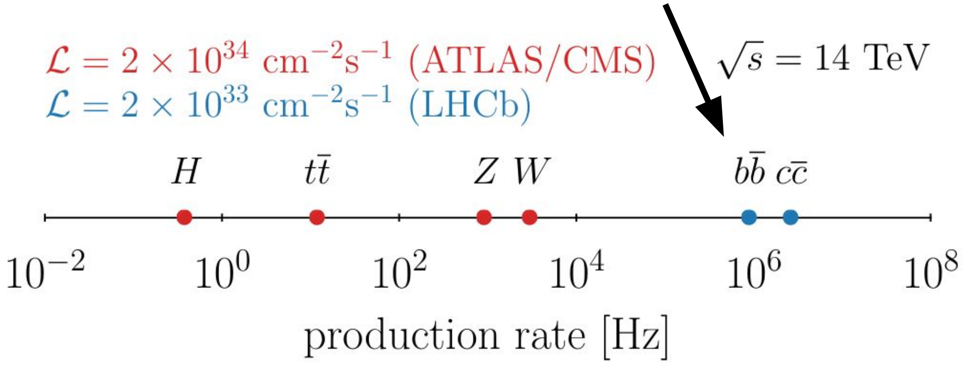
netherlands eScience center

Nikhef

LHCb

High throughput demands of LHCb Run 3

- **LHCb** studies mainly decays of ***beauty*** and ***charm*** **hadrons** with **high signal rates**

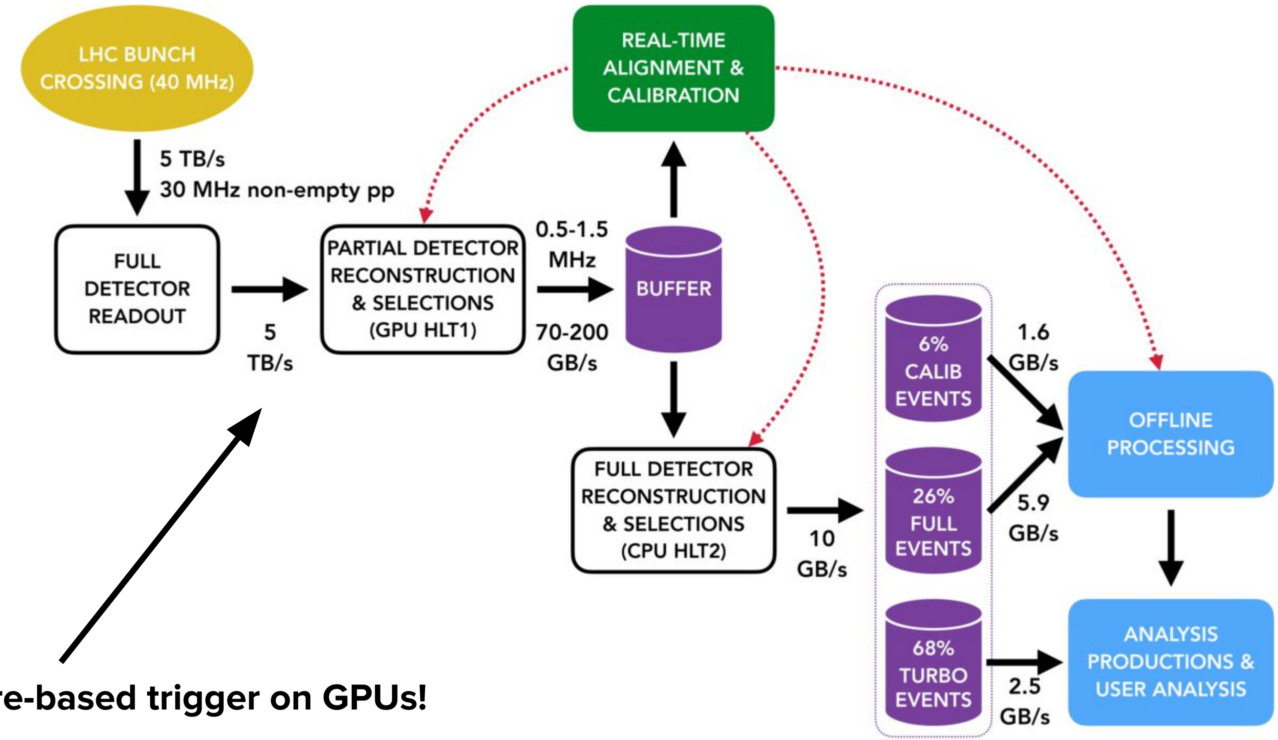


[LHCb-PROC-2022-010]

- **DAQ running at 40 MHz** to cope with **high signal rate**
  - *Reconstruction and selection* with as **many features** as possible, as **early** as possible
  - See also Flavio Pisani's talk on [LHCb's triggerless DAQ]
- Extract information from tracking sub-detectors and subsequently ***reconstruct*** and ***select***
  - Make use of **Machine Learning** (inference) **at earliest selection level as much as possible**

Data flow of the current detector

[LHCB-FIGURE-2020-016]

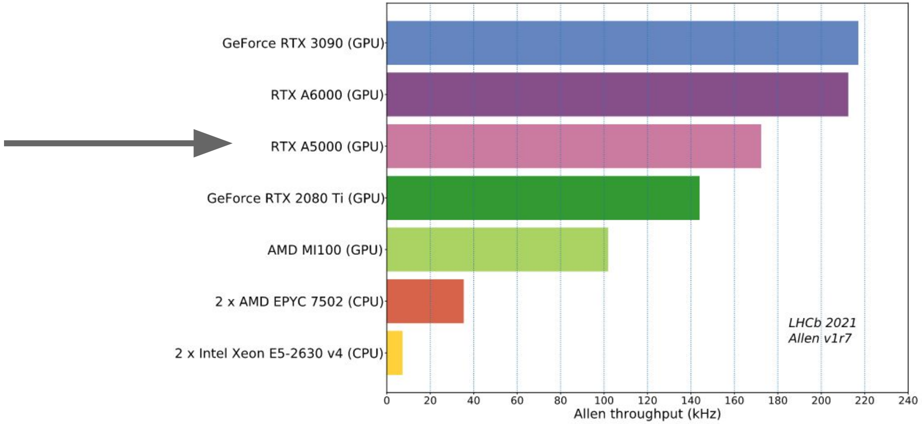


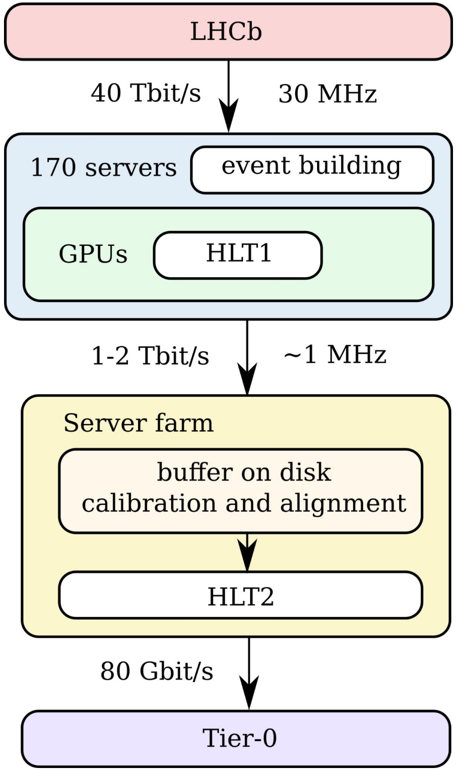
Direct software-based trigger on GPUs!

First level trigger at LHCb HLT1

[Comput Softw Big Sci 4, 7 (2020)](#)

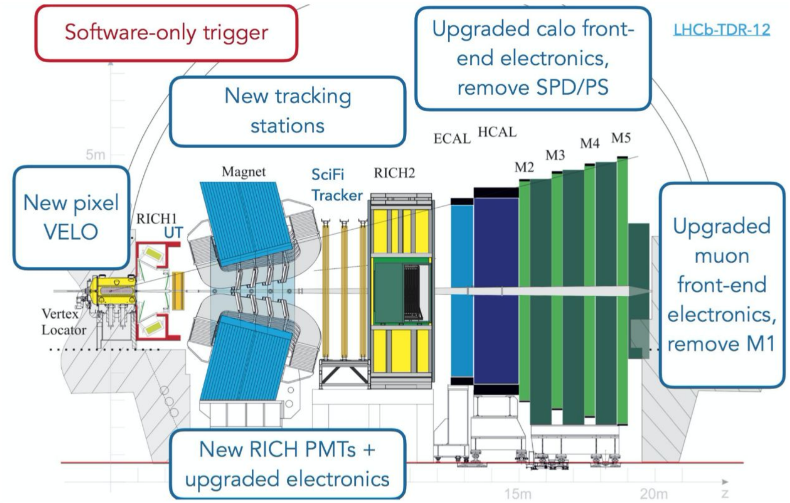
- **326 GPUs** reduce the rate of incoming data from 5 TB/s to approximately 100 GB/s
  - Doubled the number of GPUs this year!
  - About **70 kernels** running, with the [Allen](#) software project
  - See also Conor Fitzpatrick's talk on [HLT1 commissioning](#)
- With 500 GPUs, **minimum** requirement is **60 kHz per GPU** for 30 MHz non-empty bunch crossings
  - **Target achieved!**

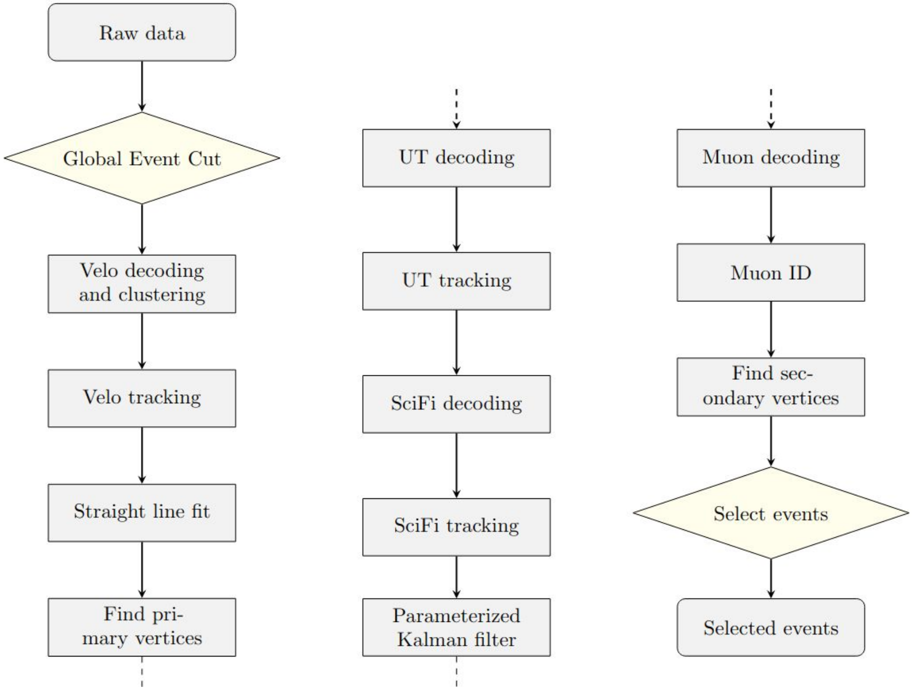




HLT1 reconstruction

- VELO: clustering, tracking, vertexing
- UT, SciFi: tracking
- Track fit and secondary vertex reconstruction
- Muon / Calorimeter reconstruction
  - Muon and Electron PID
  - Neutrals reconstruction
    - See Núria Valls Canudas' talk





LHCB-TDR-021

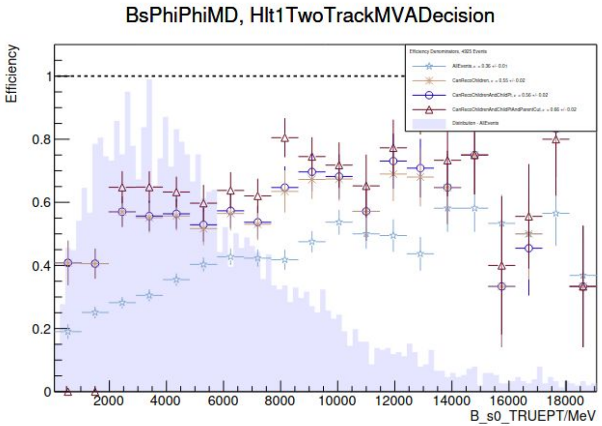
HLT1 selection

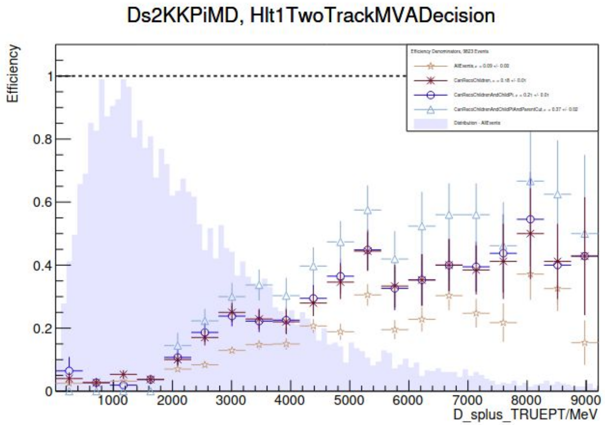
- Selection focused on ***displaced charged track (combinations)***
  - With additional dedicated (displaced) muon and electron lines
- Thresholds tuned to give a combined **output of 1 MHz**

Typical rates

Trigger	Rate [kHz]
1-Track	215 ± 18
2-Track	659 ± 31
High-p_T muon	5 ± 3
Displaced dimuon	74 ± 10
High-mass dimuon	134 ± 14
Total	999 ± 38

[Comput Softw Big Sci 4, 7 (2020)]



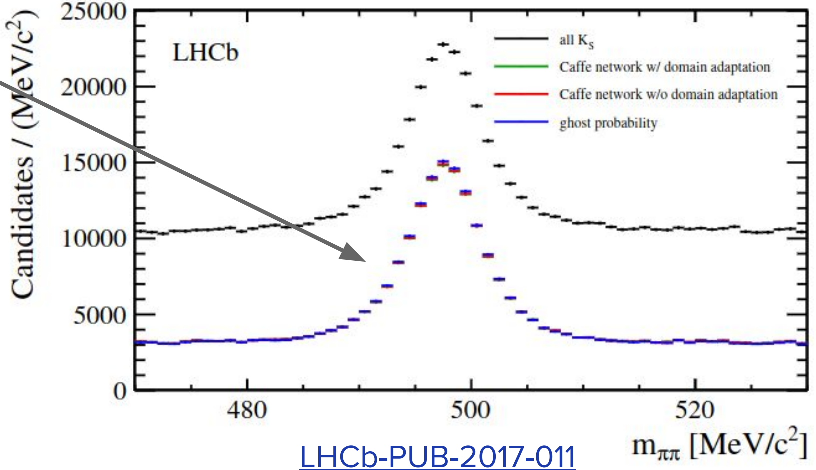


[LHCB-TDR-021]

Applications of ML *in online environment of LHCb*

- **Classification of reconstructed objects** (at all levels)
  - **Reconstruction**
    - Charged tracks
      - Real vs fake (ghost rejection)
    - Type of charged tracks
      - pion / muon / electron / …
  - **Selection** level
    - Higher level objects
      - combination of tracks coming from heavy flavour decays
    - Typically trained / used for **selecting specific signals** with trigger lines
  - **Typical feature counts of 10-20**
- Other tasks like ***pattern recognition*** and ***anomaly detection*** are possible and studied

Ghost rejection MLP from previous LHCb Run 2



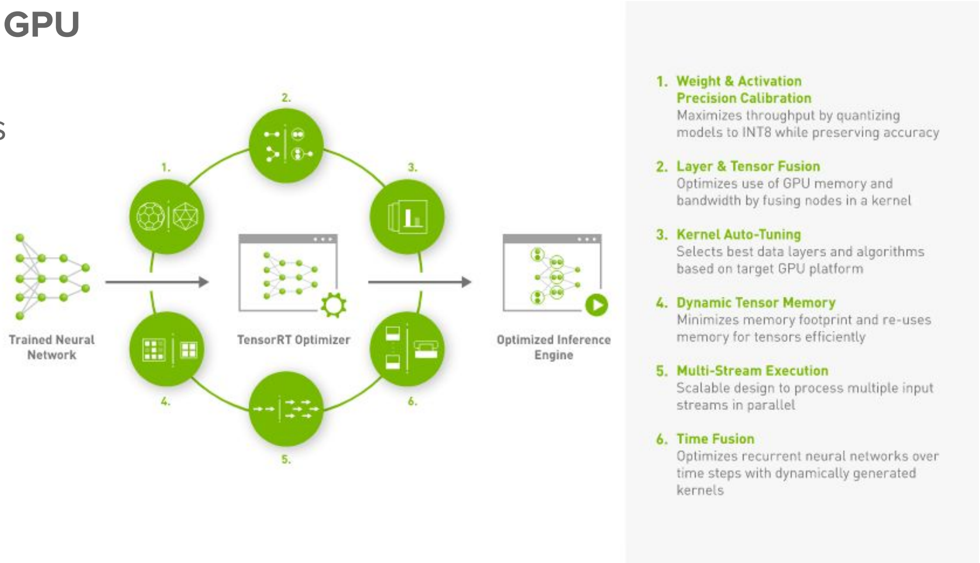
LHCb-PUB-2017-011

Providing a base for ML inference in HLT1 / Allen

- ***Flexibility, maintainability***
  - Hard/hand-coded ML inference is not flexible / not great to maintain
  - Platform to load **standardized ML-model data format**: **ONNX**
    - Supported by many (if not most) training software
    - For LHCb, at CPU (HLT2) level being integrated with *ONNXRuntime*

ONNX

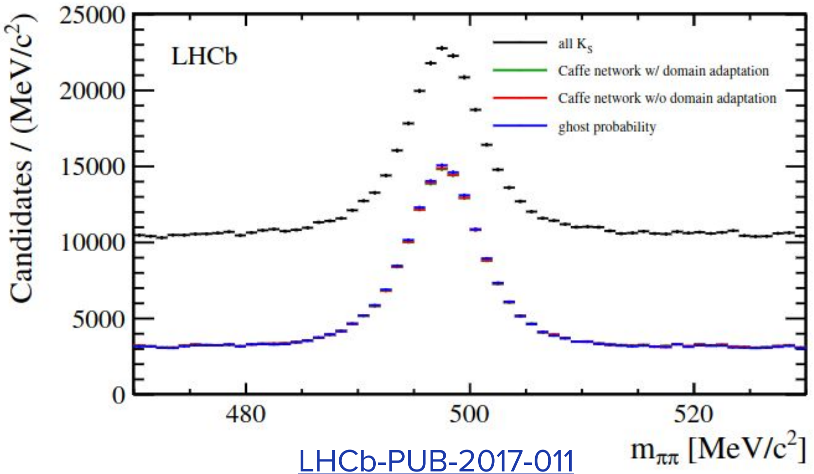
- Providing these features with inference on **GPU**
  - LHCb uses NVIDIA RTX A5000
  - **TensorRT** [link] from NVIDIA provides
    - Fast-inference platform / SDK
    - ONNX files can be read by it
    - Optimization possible within package, like quantization



Testing throughput impact of TensorRT inference

- Testing Machine Learning with **TensorRT** dummy ghost-rejection **MLP**
  - **17 features** (typical size) from tracks / tracking algorithms
  - 2 hidden layers (dim: 25, 20), 1 dimensional classifier output
    - Larger alternative with 6 hidden layers (up to 128 neurons) each tested as well
  - Testing possibility of quantization within TensorRT as well

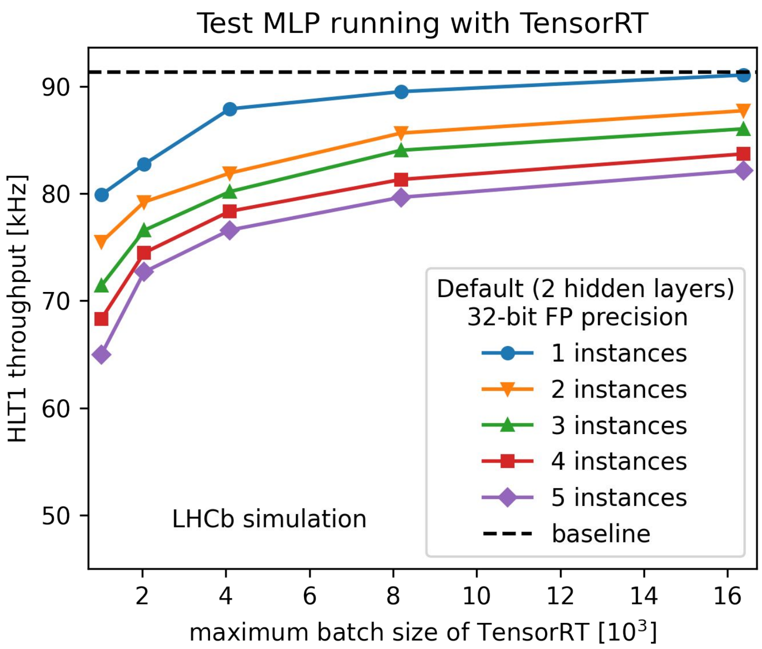
Ghost rejection MLP from previous LHCb Run 2



LHCb-PUB-2017-011

Throughput impact of TensorRT inference

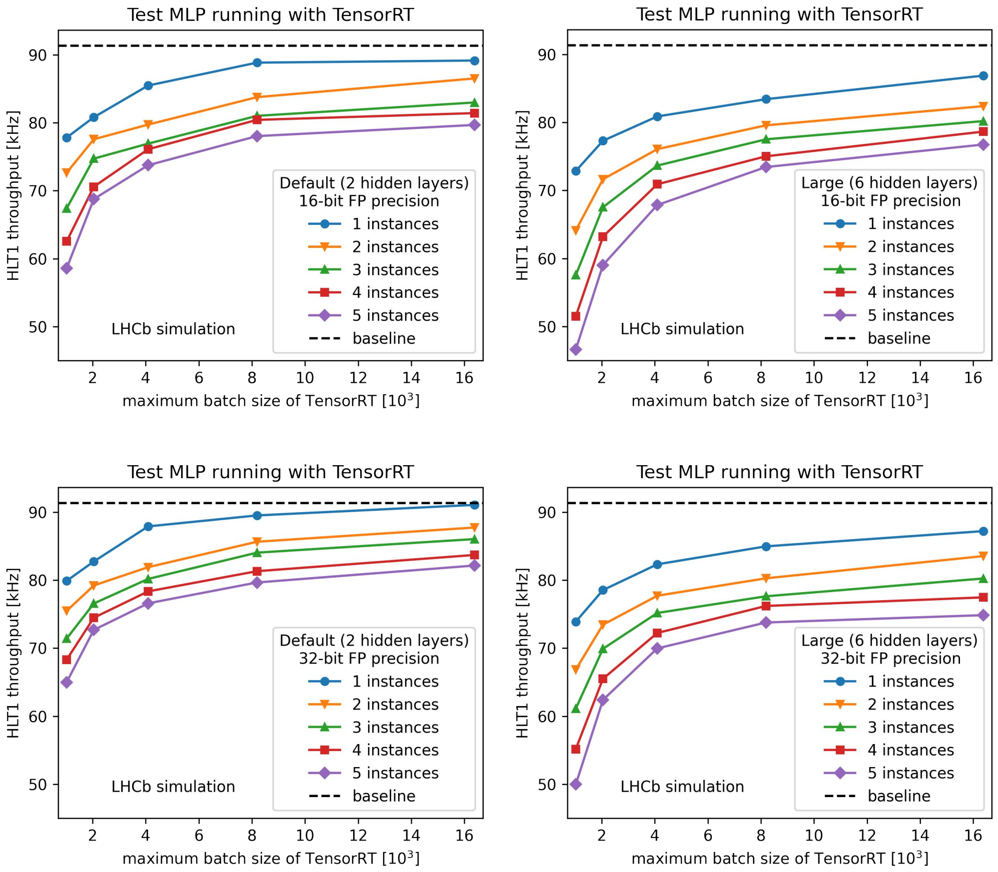
- The **baseline model** tested with respect to TensorRT **batch size**
  - **Kernel overhead is main bottleneck**
    - These MLPs are small
- At high batch size it seems **feasible to run multiple copies** of such neural nets!



Throughput impact of TensorRT inference

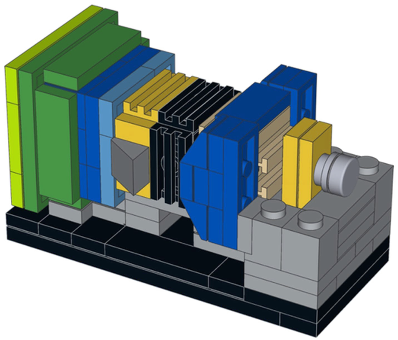
Other variations

- With **larger MLPs**, throughput decrease is stronger, as expected
- **Quantization** differences are minimal
- Most effects seems to be **batch size**
- **No show stoppers so far!**



Conclusions and outlook

- **LHCb** has **high demands of throughput** of reconstruction and selection on GPUs to cope with high signal rates
- **Machine learning ideal to reduce rates while keeping signal efficiencies high**



- Introducing **flexible loading of ML models** at the **first trigger level** (running on GPUs) with **TensorRT**
  - Multiple copies of typical sized MLPs seems to **effect throughput in an acceptable way**
- **Promising avenue of having flexible ML reconstruction and selection at the first trigger level!**