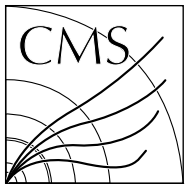
Available on CMS information server

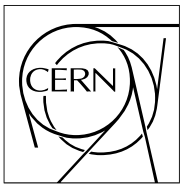
CMS DP -2024/082



The Compact Muon Solenoid Experiment

CMS Performance Note

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland



19 October 2024

2024 HLT timing and throughput results

CMS Collaboration

Abstract

This note presents measurements of the computational performance of the HLT in 2024, in terms of timing (time per event), throughput (events processed per unit time), memory usage and power efficiency. Different configurations are compared, demonstrating the improvements brought by offloading part of the HLT reconstruction to GPUs.

2024 HLT timing and throughput results

October 17th, 2024

The CMS Collaboration
cms-trigger-coordinator@cern.ch

input samples

- The measurements are performed over events
  - collected on July 25th 2024, based only on the Level 1 trigger selection,
  - with an instantaneous luminosity between 2.07×10^{34} and 2.10×10^{34} $cm^{-2}s^{-1}$,
  - with a pileup between 63.0 and 64.0 proton-proton interactions.

- Input data
  - The input data are stored uncompressed, in the same format read by the High Level Trigger (HLT) during data taking, split into files with 100 events each.
  - Before each measurement the input files are read into memory and cached by the operating system.
  - This emulates the conditions used during data taking, where the input data is held on RAM disks and read through a high speed network interface.

High Level Trigger on GPUs

- Part of the High Level Trigger (HLT) reconstruction has been written using the Alpaka *performance portability* library [1][2][3]:
  - the Pixel unpacking, local reconstruction, track reconstruction, and vertex reconstruction [4]
  - the ECAL unpacking and local reconstruction [5]
  - the HCAL local reconstruction [6] and Particle Flow clustering [7]
- The alpaka-based reconstruction can run transparently on CPUs or on GPUs, with almost identical results
  - it has been fully validated on x86-64 CPUs and on NVIDIA GPUs
  - is has undergone a preliminary validation on AMD RDNA3 GPUs
- With the HLT configuration and accelerator and detector conditions used for most of the 2024 data taking, this part covers about **35%** of the online reconstruction time.

references

[1] Alexander Matthes *et al.*, Tuning and optimization for a variety of many-core architectures without changing a single line of implementation code using the Alpaka library, doi:10.1007/978-3-319-67630-2_36

[2] Erik Zenker *et al.*, Alpaka - An Abstraction Library for Parallel Kernel Acceleration, doi:10.1109/IPDPSW.2016.50

[3] Benjamin Worpitz, Investigating performance portability of a highly scalable particle-in-cell simulation code on various multi-core architectures, doi:10.5281/zenodo.49768

[4] Andrea Bocci *et al.*, Heterogeneous Reconstruction of Tracks and Primary Vertices With the CMS Pixel Tracker, doi:10.3389/fdata.2020.601728

[5] Thomas Reis, Developing GPU-compliant algorithms for CMS ECAL local reconstruction during LHC Run 3 and Phase 2, doi:10.1088/1742-6596/2438/1/012027

[6] Martin Kwok, Portable HCAL reconstruction in the CMS detector using the Alpaka library, to be presented at CHEP 2024

[7] Jonathan Samudio, Particle Flow Reconstruction with Alpaka Portability Library, to be presented at CHEP 2024

2022 HLT nodes

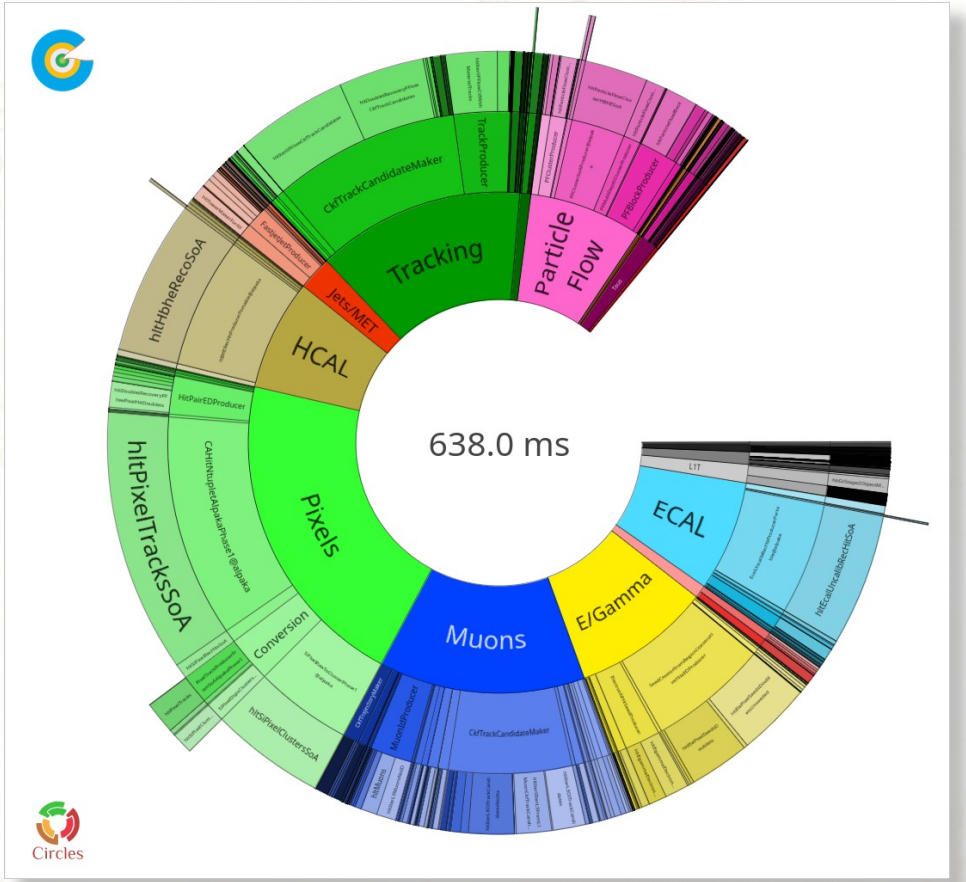
2022 HLT nodes – measurements

- Each of the 200 HLT nodes installed in 2022 is composed of
  - 2 × AMD EPYC “Milan” 7763 processors
    - each with 64 physical cores and 128 hardware threads, partitioned in 4 NUMA nodes
  - 256 GB of RAM
  - 2 × NVIDIA Tesla T4 GPUs
    - each with 16 GB of RAM
- The measurements on these nodes are performed under conditions as close as possible to those used during data taking
  - each measurement consists of 8 jobs running in parallel
    - each jobs uses 32 CPU threads for data processing and processes up to 24 concurrent events
    - each job uses a single NUMA node; there is a single job per NUMA node
  - for the configuration with GPUs:
    - each job uses a single GPU (the one connected directly to the processor and NUMA node)
    - each GPU is shared by 4 jobs; the NVIDIA MPS service is used to share a GPU among multiple jobs more efficiently
  - each measurement is repeated 5 times, and the first measurement is discarded to “warm up” the machine
  - the results are the average ± the standard deviation of the four measurements after the “warm up” one

2022 HLT nodes – timing

- The *timing* measurements on these nodes are performed over 70’000 events with an average pileup of 63.7
  - without GPUs, each 2022 HLT node takes 638.0 ± 5.5 ms per event
  - with GPUs, each 2022 HLT node takes 421.3 ± 2.3 ms per event
  - this corresponds to a speed up of 51.4% ± 1.5%
  - this can be interpreted as 34.0% ± 0.7% of the HLT being offloaded to GPUs

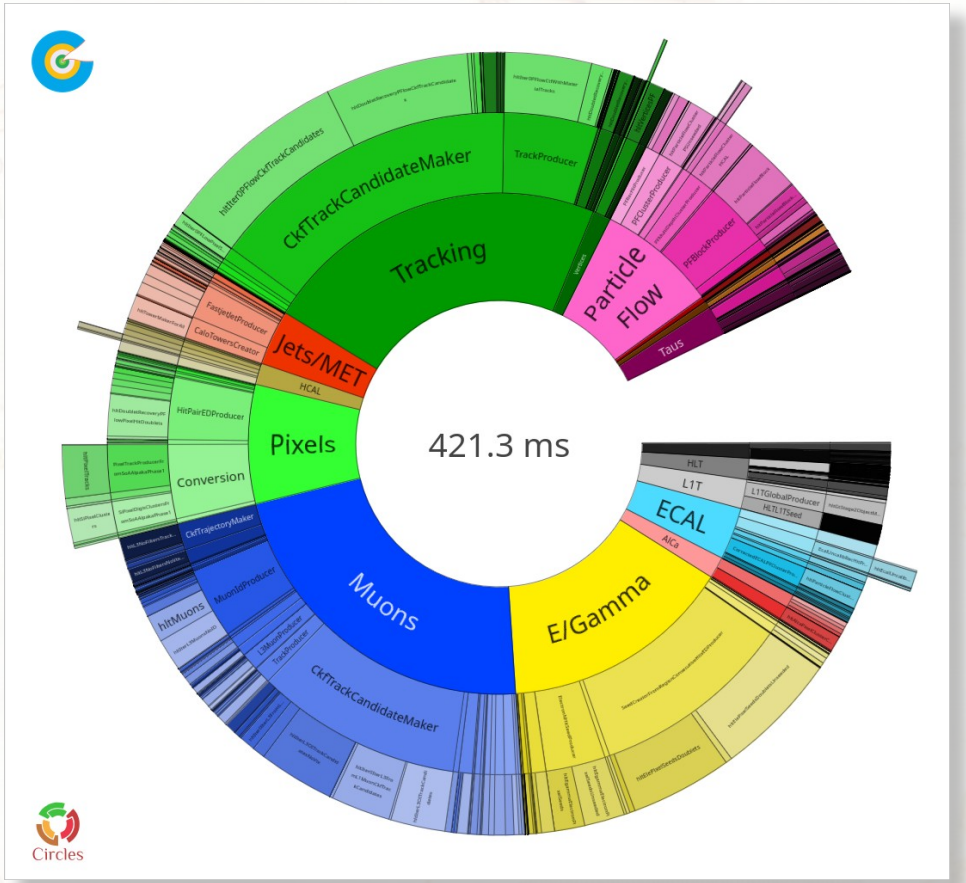
2022 HLT nodes – without GPUs



Element	Time	Fraction
AlCa	5.3 ms	0.8 %
B tagging	1.8 ms	0.3 %
CTPPS	0.0 ms	0.0 %
DQM	1.3 ms	0.2 %
E/Gamma	56.8 ms	8.9 %
ECAL	45.6 ms	7.2 %
Framework	0.0 ms	0.0 %
HCAL	45.7 ms	7.2 %
HLT	5.1 ms	0.8 %
I/O	3.8 ms	0.6 %
Jets/MET	13.9 ms	2.2 %
L1T	7.6 ms	1.2 %
Muons	85.0 ms	13.3 %
Particle Flow	46.6 ms	7.3 %
Pixels	132.8 ms	20.8 %
Taus	7.2 ms	1.1 %
Tracking	85.5 ms	13.4 %
Vertices	4.5 ms	0.7 %
event setup	0.1 ms	0.0 %
idle	0.2 ms	0.0 %
other	89.3 ms	14.0 %
total	*638.0 ms*	*100.0 %*

CPU only

2022 HLT nodes – with GPUs



Element	Time	Fraction
AlCa	5.8 ms	1.4 %
B tagging	1.9 ms	0.5 %
CTPPS	0.0 ms	0.0 %
DQM	1.5 ms	0.3 %
E/Gamma	63.2 ms	15.0 %
ECAL	14.2 ms	3.4 %
Framework	0.0 ms	0.0 %
HCAL	5.8 ms	1.4 %
HLT	5.7 ms	1.4 %
I/O	3.9 ms	0.9 %
Jets/MET	15.0 ms	3.6 %
L1T	7.9 ms	1.9 %
Muons	93.5 ms	22.2 %
Particle Flow	32.5 ms	7.7 %
Pixels	32.4 ms	7.7 %
Taus	8.0 ms	1.9 %
Tracking	94.4 ms	22.4 %
Vertices	4.9 ms	1.2 %
event setup	0.1 ms	0.0 %
idle	0.2 ms	0.0 %
other	30.3 ms	7.2 %
total	*421.3 ms*	*100.0 %*

with GPUs

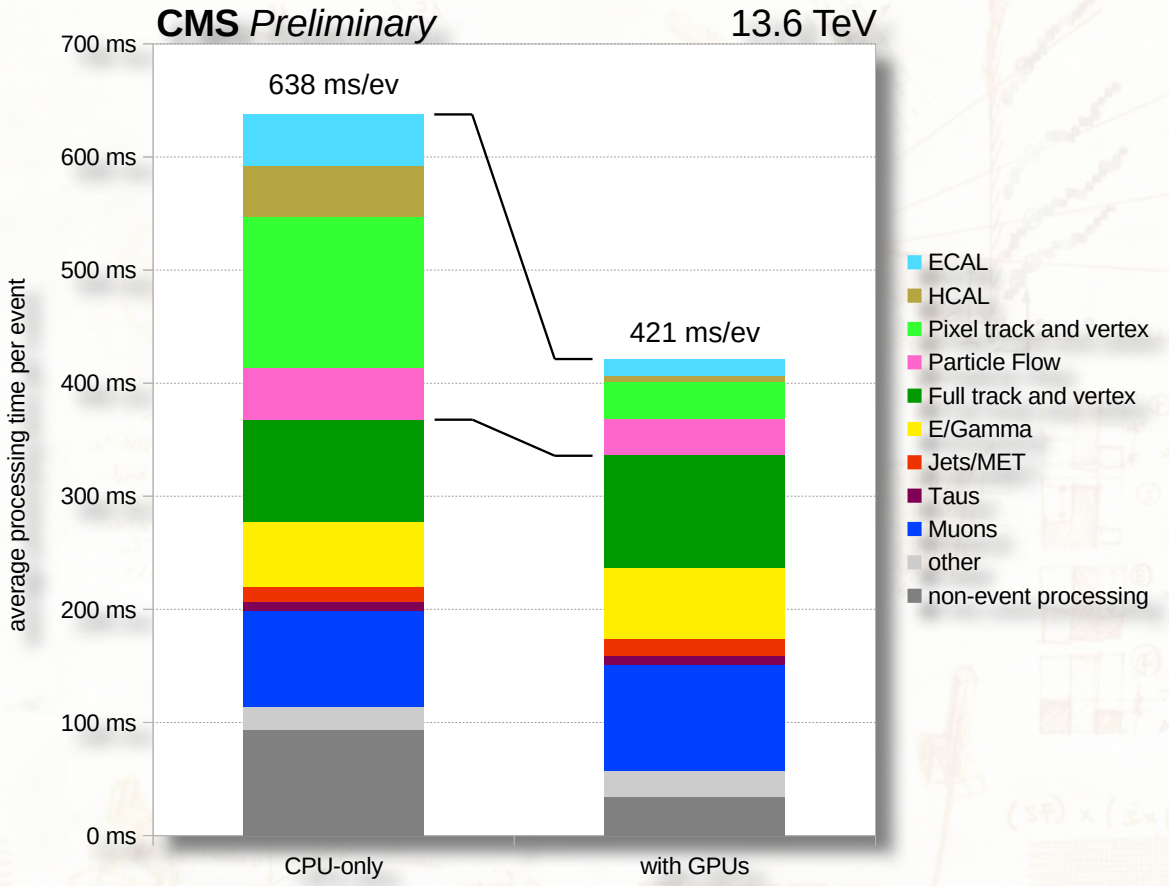
2022 HLT nodes - comparison

Comparison of the average processing time per event, measured on the 2022 HLT nodes

- each nodes is composed of
  - 2 × AMD EPYC “Milan” 7763 processors
  - 2 × NVIDIA Tesla T4 GPUs

The measurements are performed over 70’000 events with an average pileup of 63.7

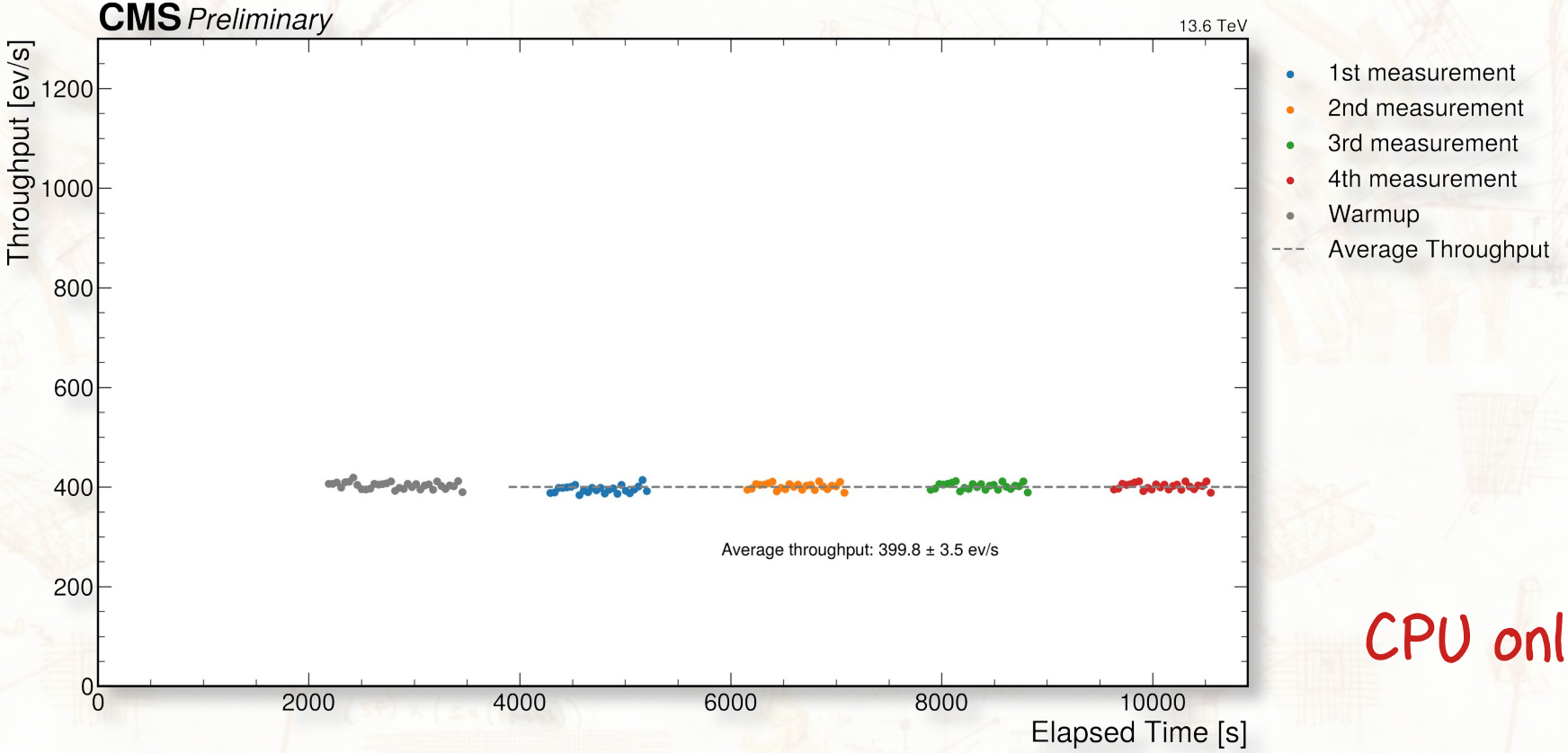
- without GPUs, each 2022 HLT node takes 638.0 ± 5.5 ms per event
- with GPUs, each 2022 HLT node takes 421.3 ± 2.3 ms per event
- this corresponds to a speed up of 51.4% ± 1.5%
- this can be interpreted as 34.0% ± 0.7% of the HLT offloaded to GPUs



2022 HLT nodes – throughput

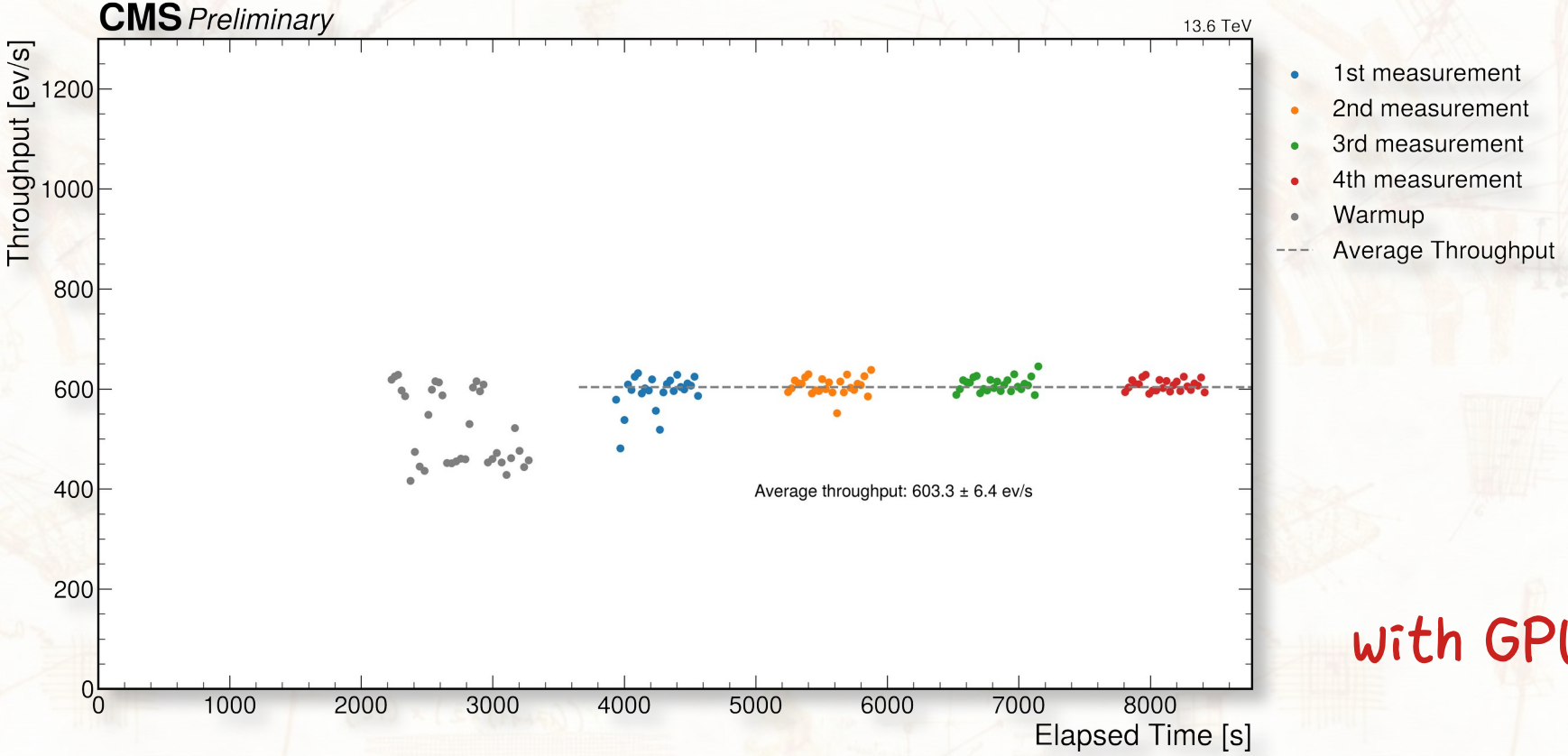
- The *throughput* measurements on these nodes are performed over 50’000 events with an average pileup of 63.9
  - the measurements ignore the processing of the first 20’000 events, and consider only the interval when all jobs are actively processing events
  - without GPUs, each 2022 HLT node can process 399.8 ± 3.5 events per second
  - with GPUs, each 2022 HLT node can process 603.3 ± 6.4 events per second
  - this corresponds to a speed up of 51.3% ± 1.7%
  - this can be interpreted as 33.9% ± 0.7% of the HLT being offloaded to GPUs

2022 HLT nodes – without GPUs



CPU only

2022 HLT nodes – with GPUs

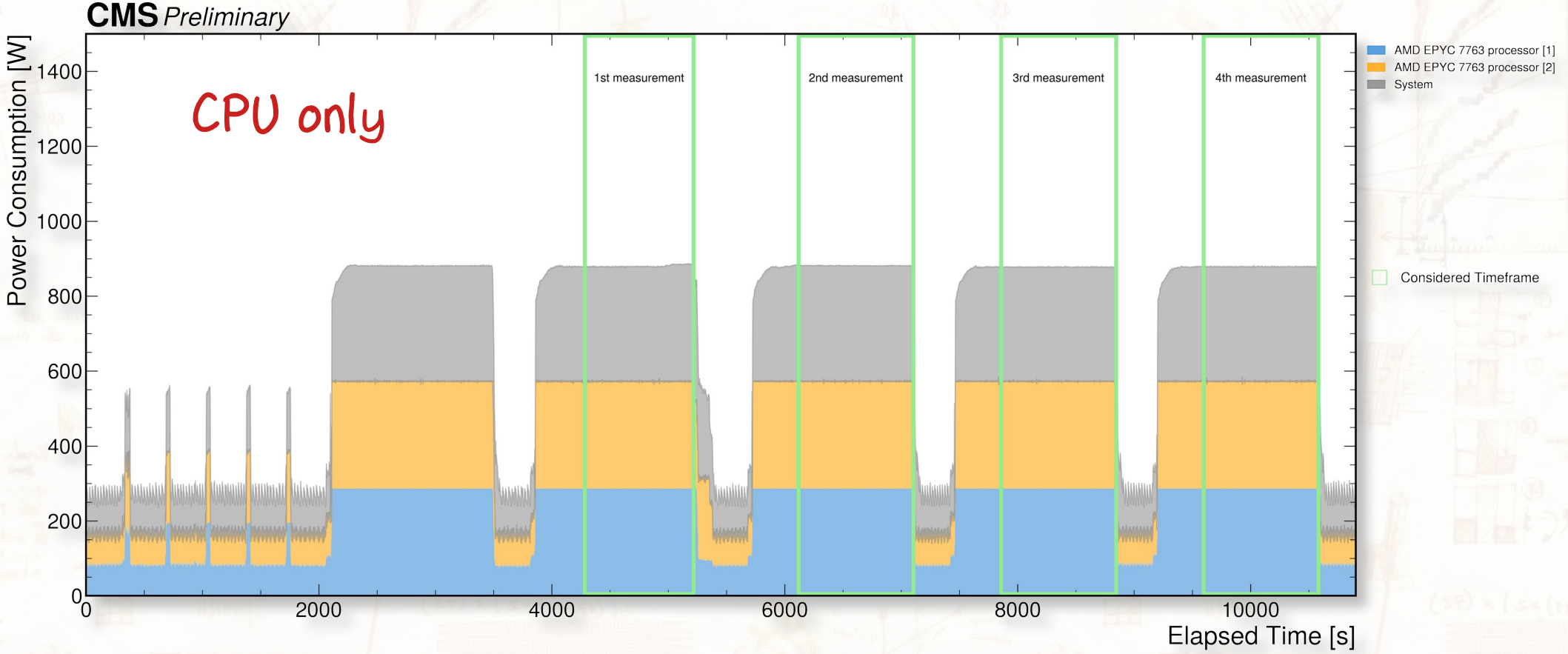


with GPUs

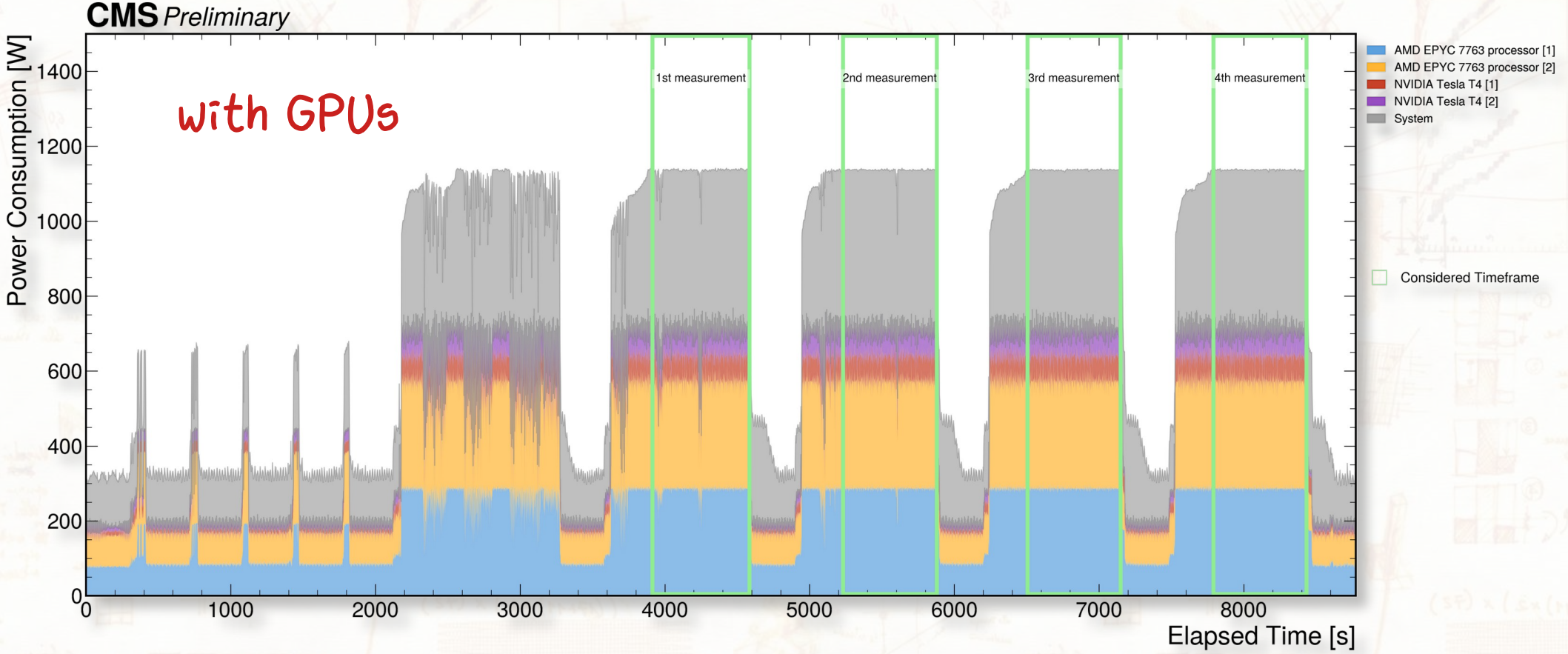
2022 HLT nodes – power consumption

- The *power consumption* measurements on these nodes are performed over 50’000 events with an average pileup of 63.9
  - the measurements ignore the processing of the first 20’000 events, and consider only the interval when all jobs are actively processing events
  - without GPUs
    - each 2022 HLT node can process 399.8 ± 3.5 events per second, consuming 879.8 ± 1.4 W
    - corresponding to 2.20 ± 0.02 J per event
  - with GPUs
    - each 2022 HLT node can process 603.3 ± 6.4 events per second, consuming 1135.5 ± 1.9 W
    - corresponding to 1.88 ± 0.01 J per event
    - this corresponds to an efficiency improvement of 14.5% ± 0.2%

2022 HLT nodes – without GPUs



2022 HLT nodes – with GPUs

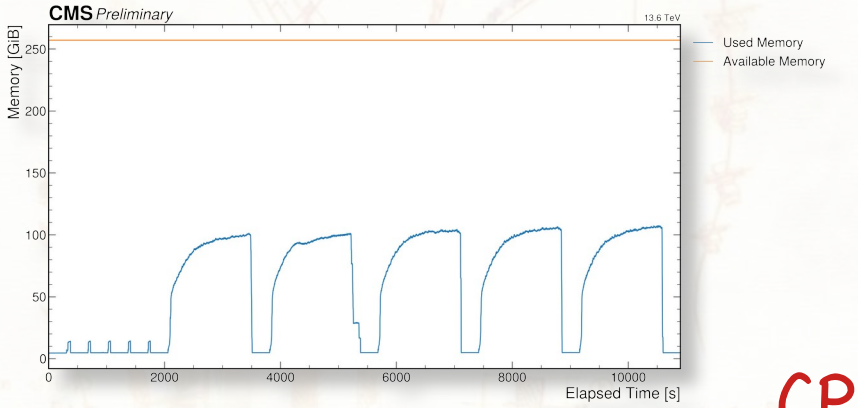


2022 HLT nodes – other details

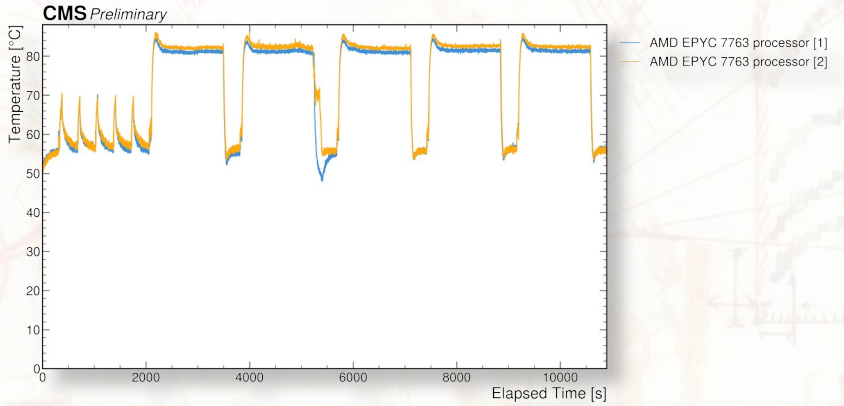
- The following plots show
  - on the *top left*, the memory usage of the whole system
  - on the *top right*, the temperature of each CPU and of each GPU (when used)
  - on the *bottom left*, the memory usage of each GPU (when used)
  - on the *bottom right*, the average utilisation of each GPU (when used)

- These monitoring data was collected while performing the timing, throughput and power consumption measurements.

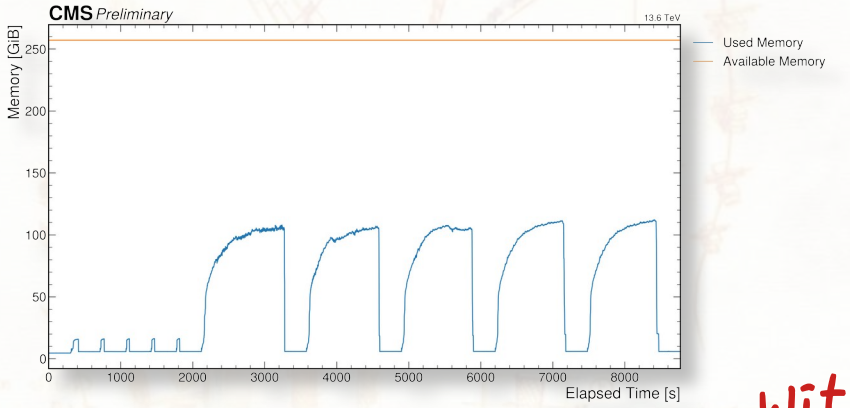
2022 HLT nodes – without GPUs

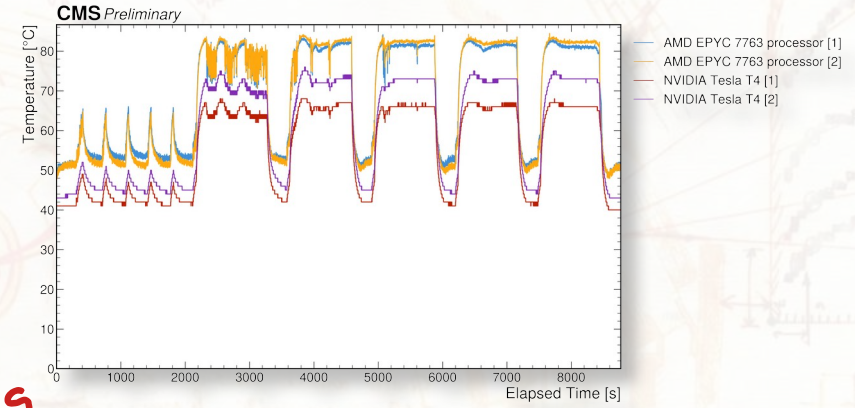


CPU only

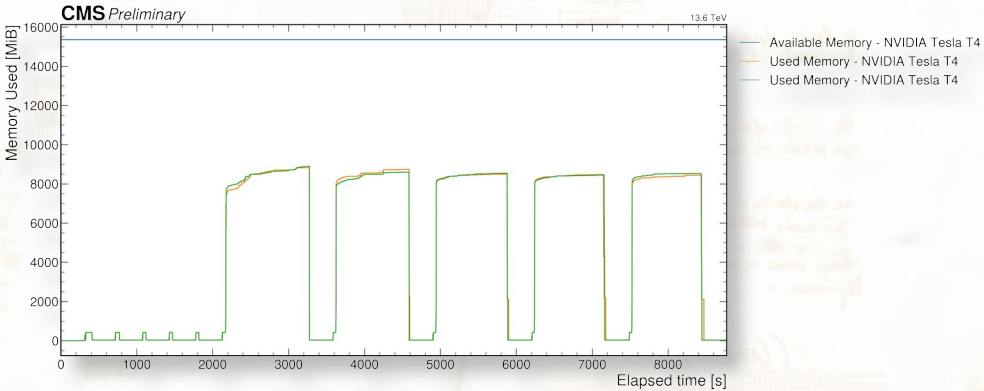


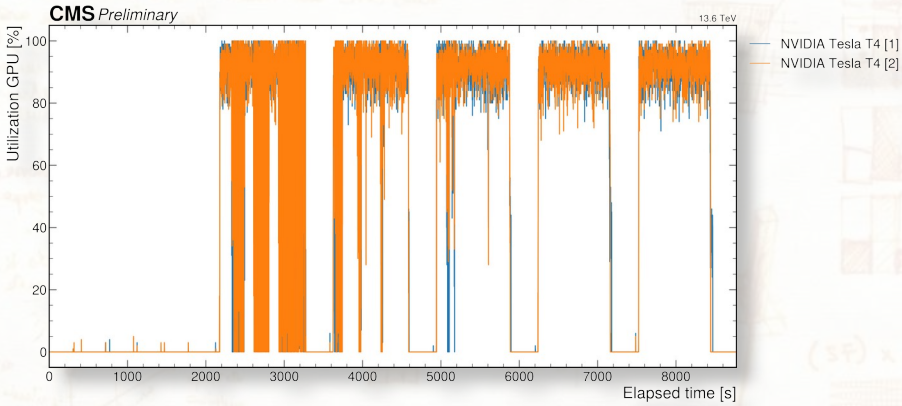
2022 HLT nodes – with GPUs





with GPUs





2024 HLT nodes

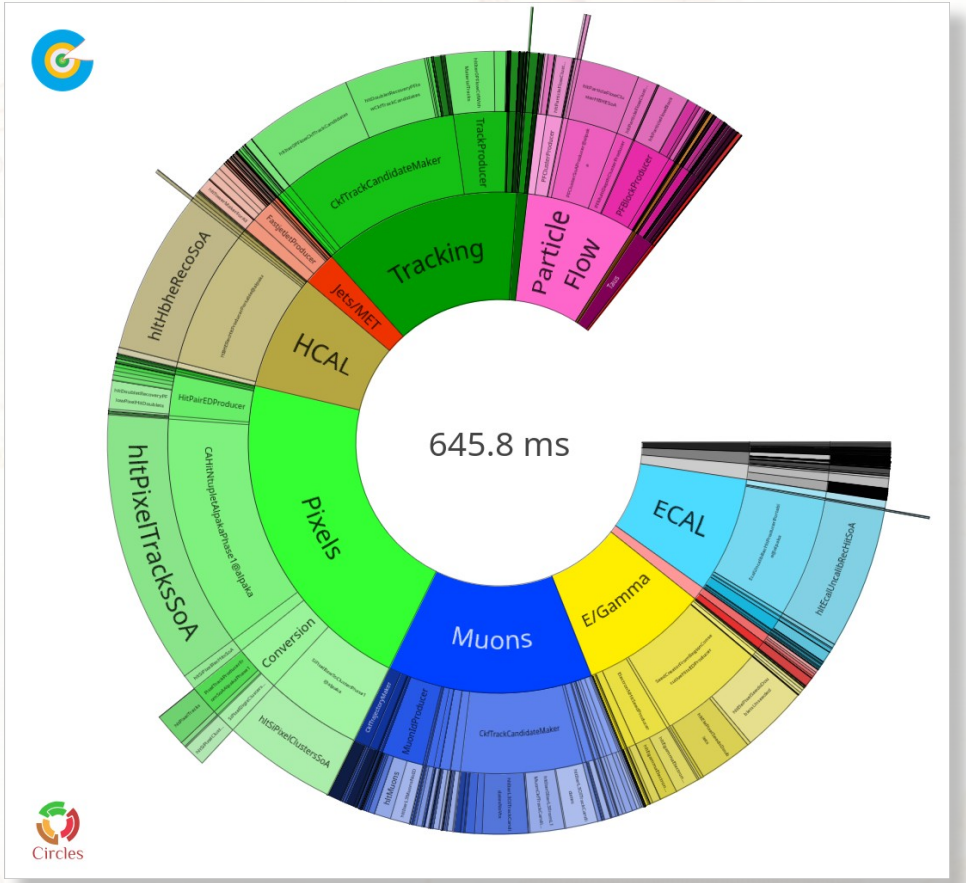
2024 HLT nodes

- Each of the 18 HLT nodes installed in 2024 is composed of
  - 2 × AMD EPYC “Bergamo” 9754 processors
    - each with 128 physical cores and 256 hardware threads, partitioned in 4 NUMA nodes
  - 768 GB of RAM
  - 3 × NVIDIA L4 GPUs
    - each with 24 GB of RAM
- The measurements on these nodes are performed under conditions as close as possible to those used during data taking
  - each measurement consists of 16 jobs running in parallel:
    - each jobs uses 32 CPU threads for data processing and processes up to 24 concurrent events
    - each job uses a single NUMA node; there are two jobs per NUMA node
  - for the configuration with GPUs:
    - each job uses a single GPU, chosen to minimise the latency and optimise the overall GPU utilisation
    - each GPU is shared by 5 or 6 jobs; the NVIDIA MPS service is used to share a GPU among multiple jobs more efficiently
  - each measurement is repeated 5 times, and the first measurement is discarded to “warm up” the machine
  - the results are the average ± the standard deviation of the four measurements after the “warm up” one

2024 HLT nodes – timing

- The *timing* measurements on these nodes are performed over 120’000 events with an average pileup of 63.8
  - without GPUs, each 2024 HLT node takes 645.8 ± 0.5 ms per event
  - with GPUs, each 2024 HLT node takes 415.1 ± 0.4 ms per event
  - this corresponds to a speed up of 55.6% ± 0.2%
  - this can be interpreted as 35.7% ± 0.1% of the HLT offloaded to GPUs

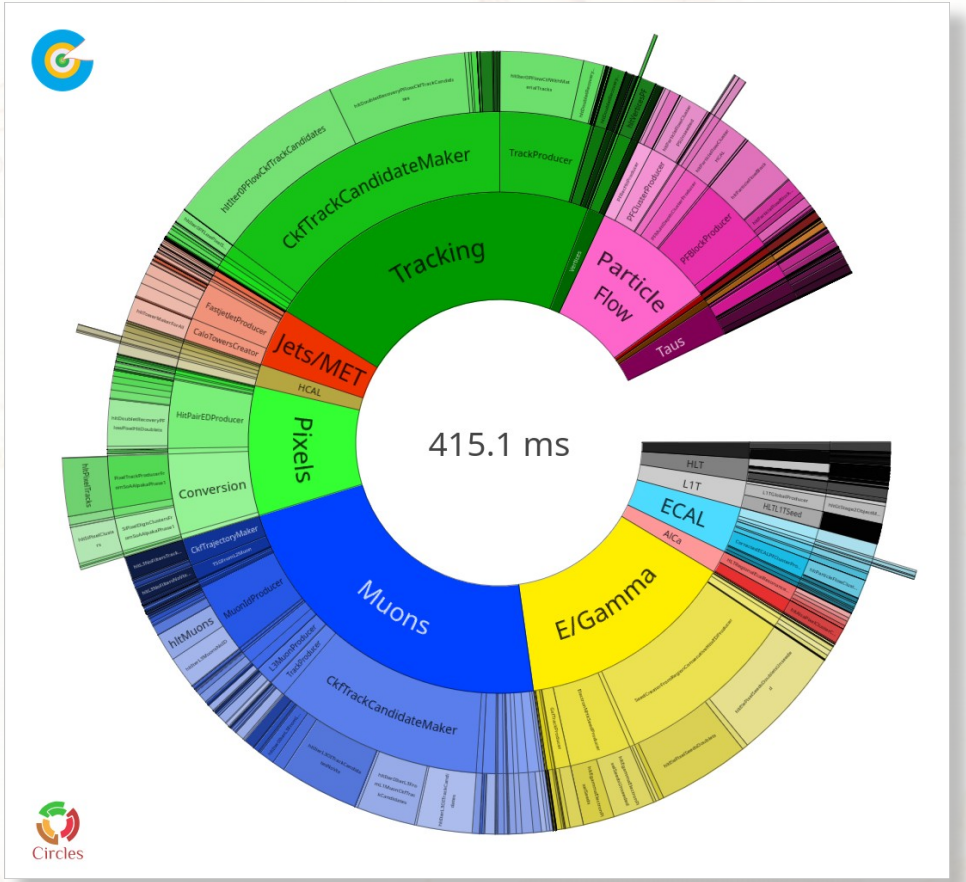
2024 HLT nodes – without GPUs



Element	Time	Fraction
AlCa	5.6 ms	0.9 %
B tagging	1.8 ms	0.3 %
CTPPS	0.0 ms	0.0 %
DQM	1.3 ms	0.2 %
E/Gamma	53.5 ms	8.3 %
ECAL	48.4 ms	7.5 %
Framework	0.0 ms	0.0 %
HCAL	48.3 ms	7.5 %
HLT	5.1 ms	0.8 %
I/O	3.8 ms	0.6 %
Jets/MET	14.2 ms	2.2 %
L1T	6.4 ms	1.0 %
Muons	85.4 ms	13.2 %
Particle Flow	46.1 ms	7.1 %
Pixels	138.5 ms	21.5 %
Taus	7.7 ms	1.2 %
Tracking	82.6 ms	12.8 %
Vertices	4.4 ms	0.7 %
event setup	0.1 ms	0.0 %
idle	0.2 ms	0.0 %
other	92.4 ms	14.3 %
total	*645.8 ms*	*100.0 %*

CPU only

2024 HLT nodes – with GPUs



Element	Time	Fraction
AlCa	6.0 ms	1.4 %
B tagging	1.9 ms	0.5 %
CTPPS	0.0 ms	0.0 %
DQM	1.6 ms	0.4 %
E/Gamma	58.9 ms	14.2 %
ECAL	12.4 ms	3.0 %
Framework	0.0 ms	0.0 %
HCAL	5.8 ms	1.4 %
HLT	6.0 ms	1.4 %
I/O	4.2 ms	1.0 %
Jets/MET	15.6 ms	3.8 %
L1T	7.2 ms	1.7 %
Muons	93.4 ms	22.5 %
Particle Flow	33.2 ms	8.0 %
Pixels	34.7 ms	8.4 %
Taus	8.6 ms	2.1 %
Tracking	91.2 ms	22.0 %
Vertices	4.7 ms	1.1 %
event setup	0.1 ms	0.0 %
idle	0.1 ms	0.0 %
other	29.6 ms	7.1 %
total	*415.1 ms*	*100.0 %*

with GPUs

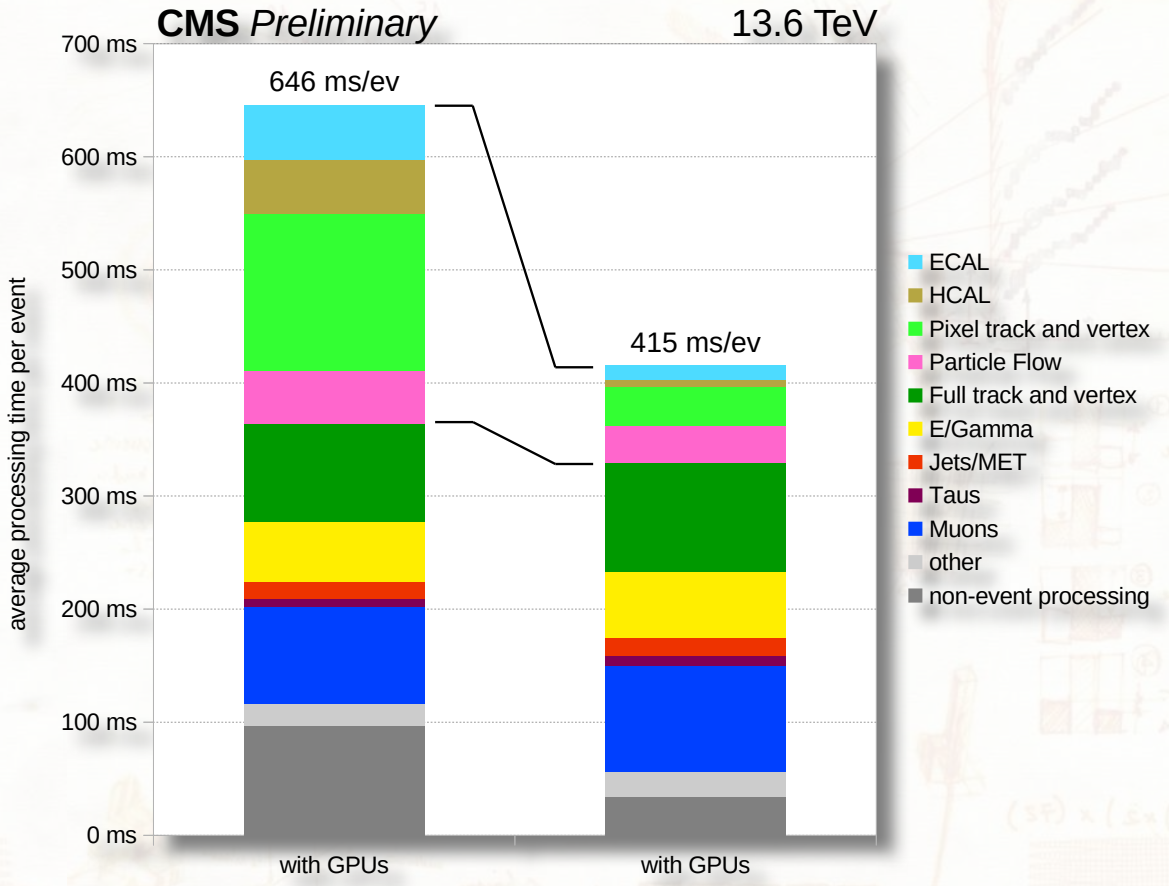
2024 HLT nodes - comparison

Comparison of the average processing time per event, measured on the 2024 HLT nodes

- each nodes is composed of
  - 2 × AMD EPYC “Bergamo” 9754 processors
  - 3 × NVIDIA L4 GPUs

The measurements are performed over 120’000 events with an average pileup of 63.8

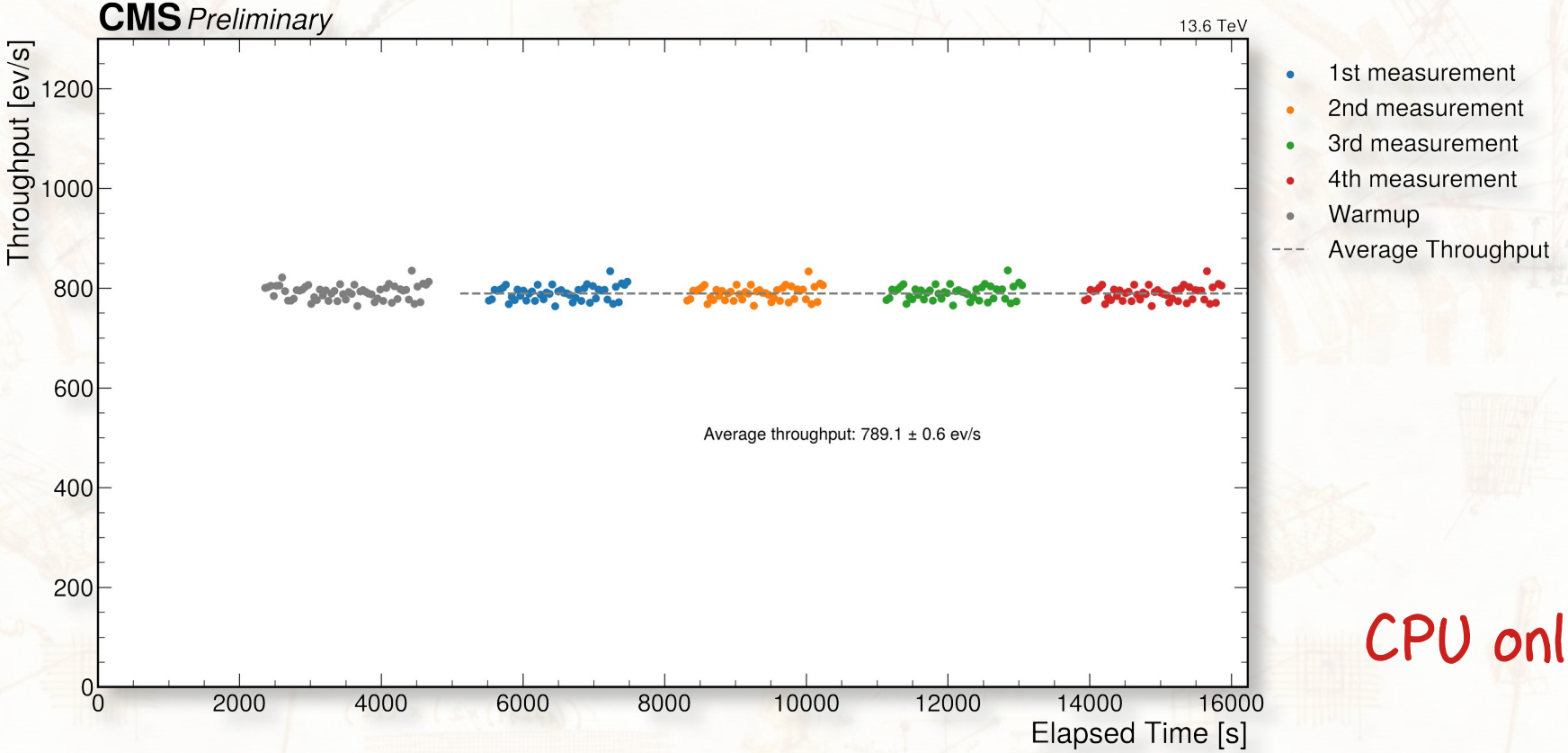
- without GPUs, each 2022 HLT node takes 645.8 ± 0.5 ms per event
- with GPUs, each 2022 HLT node takes 415.1 ± 0.4 ms per event
- this corresponds to a speed up of 55.6% ± 0.2%
- this can be interpreted as 35.7% ± 0.1% of the HLT offloaded to GPUs



2024 HLT nodes – throughput

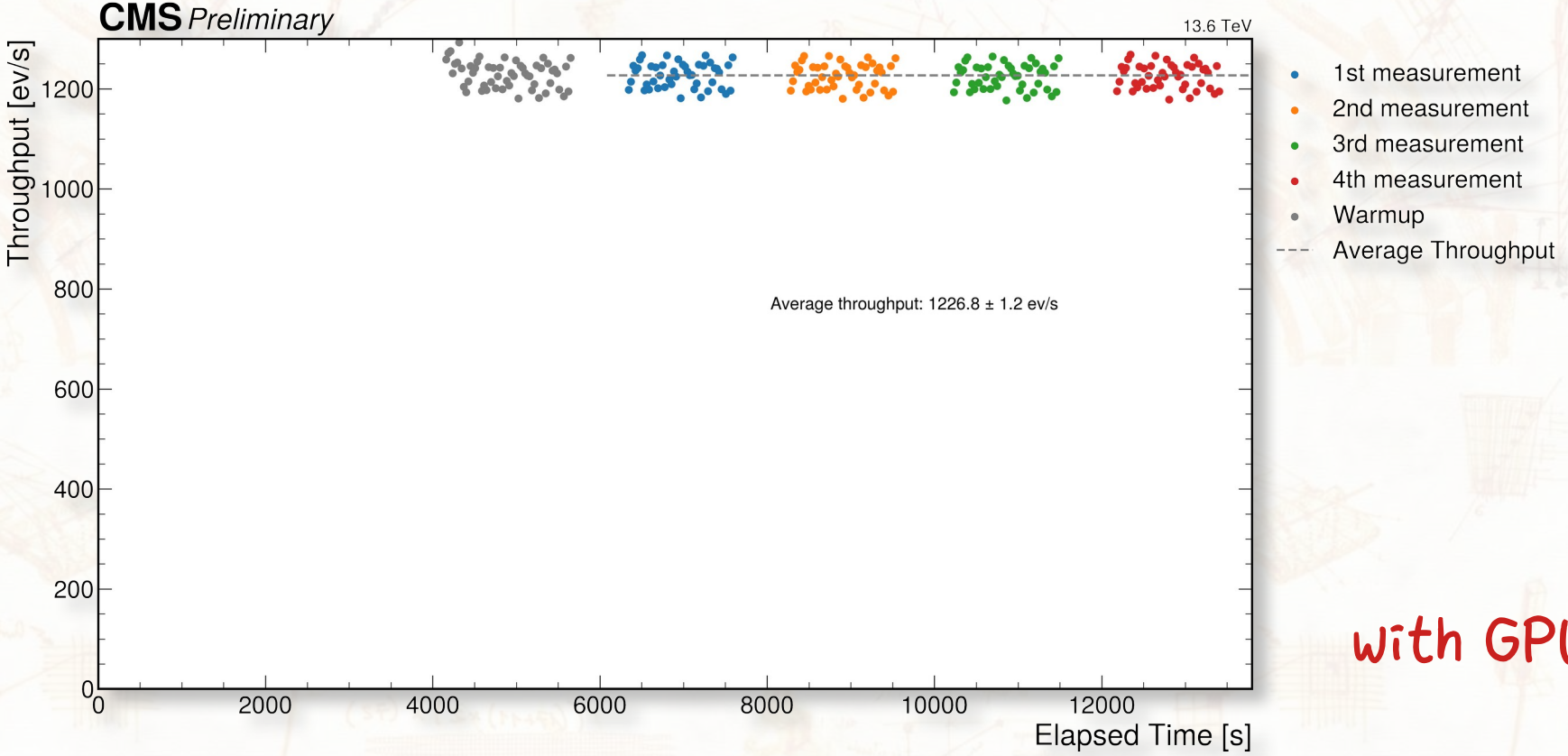
- The *throughput* measurements on these nodes are performed over 100’000 events with an average pileup of 63.9
  - the measurements ignore the processing of the first 20’000 events, and consider only the interval when all jobs are actively processing events
  - without GPUs, each 2024 HLT node can process 789.1 ± 0.6 events per second
  - with GPUs, each 2024 HLT node can process 1226.8 ± 1.2 events per second
  - this corresponds to a speed up of 55.5% ± 0.2%
  - this can be interpreted as 35.7% ± 0.1% of the HLT being offloaded to GPUs

2024 HLT nodes – without GPUs



CPU only

2024 HLT nodes – with GPUs

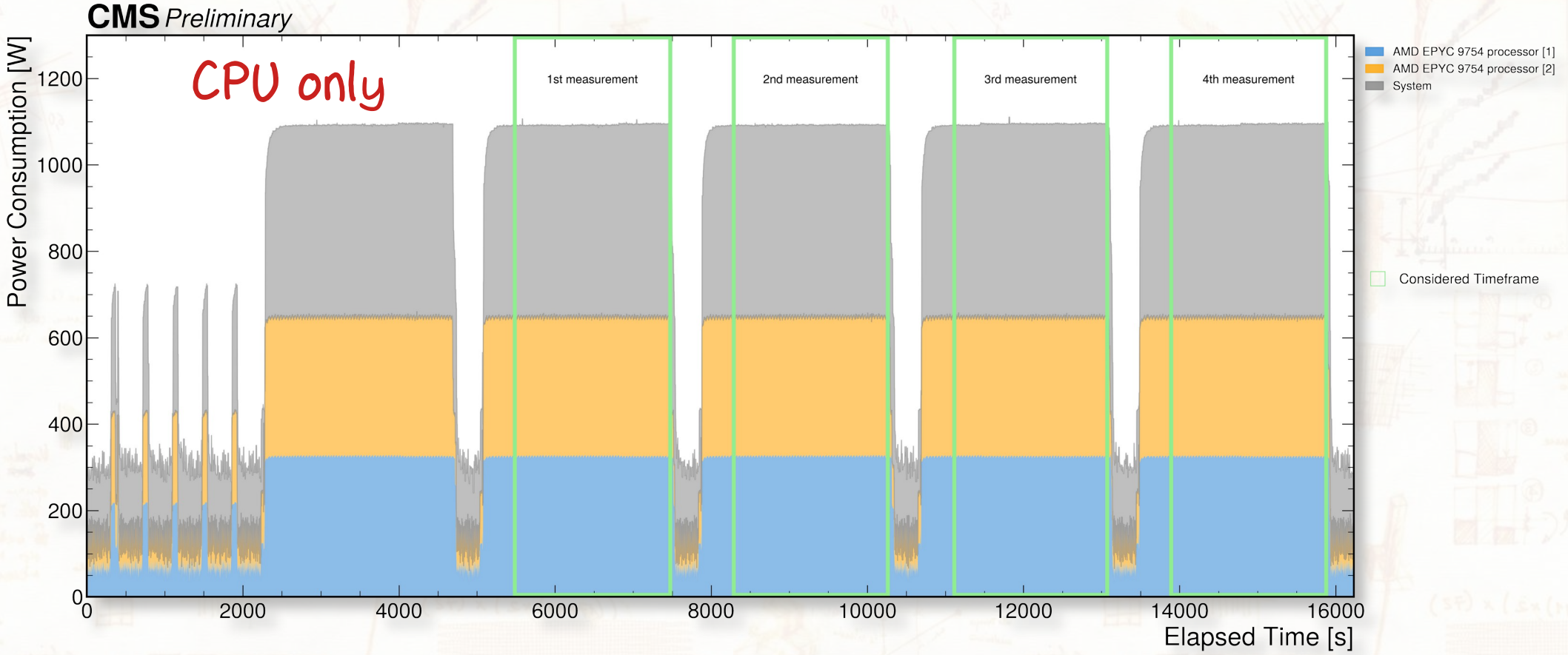


with GPUs

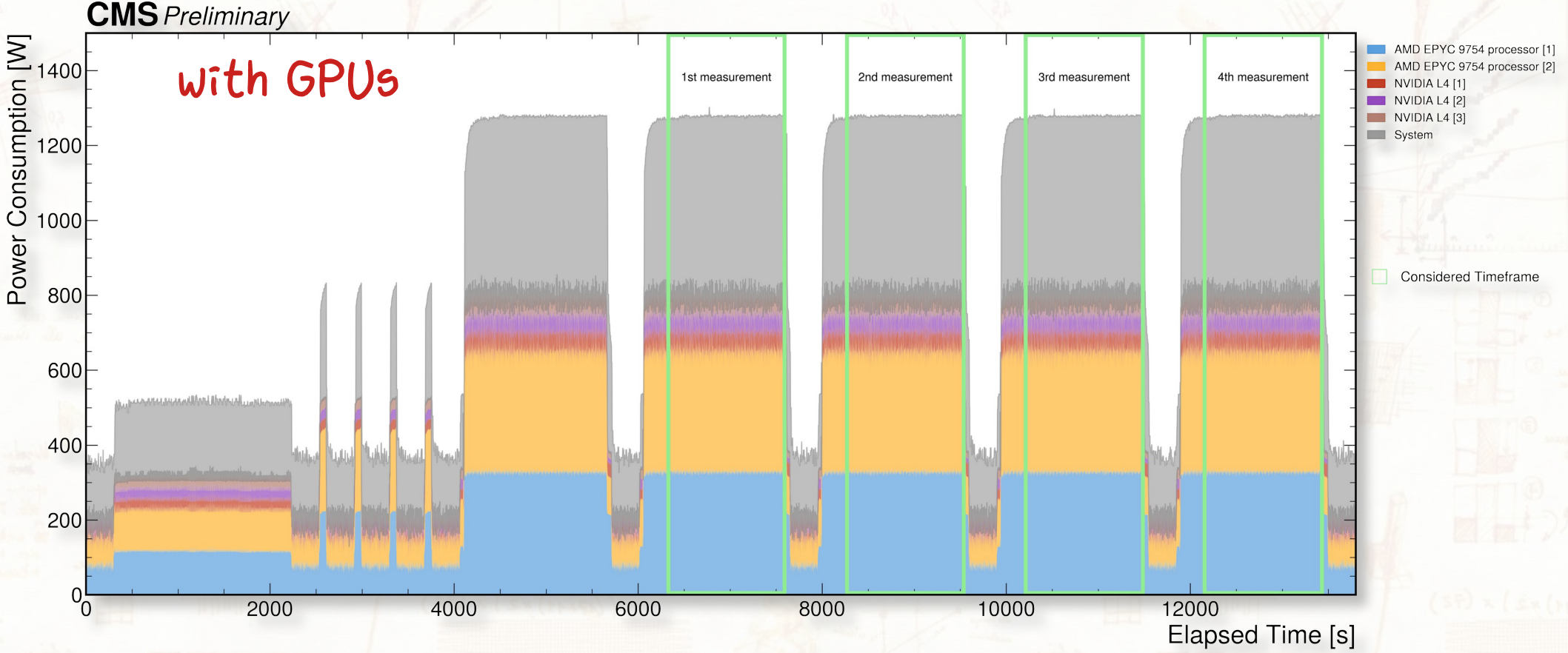
2024 HLT nodes – power consumption

- The *power consumption* measurements on these nodes are performed over 50’000 events with an average pileup of 63.9
  - the measurements ignore the processing of the first 20’000 events, and consider only the interval when all jobs are actively processing events
  - without GPUs
    - each 2022 HLT node can process 789.1 ± 0.6 events per second, consuming 1093.6 ± 0.9 W
    - corresponding to 1.39 ± 0.01 J per event
  - with GPUs
    - each 2022 HLT node can process 1226.8 ± 1.0 events per second, consuming 1278.6 ± 0.4 W
    - corresponding to 1.04 ± 0.01 J per event
    - this corresponds to an efficiency improvement of 25.2% ± 0.2%

2024 HLT nodes – without GPUs



2024 HLT nodes – with GPUs

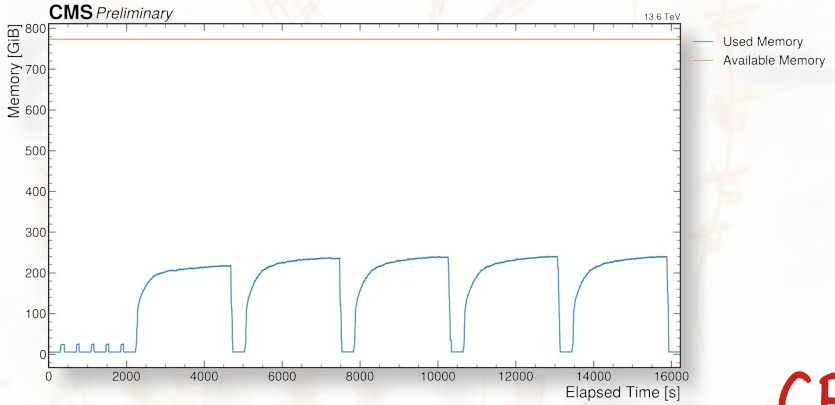


2024 HLT nodes – other details

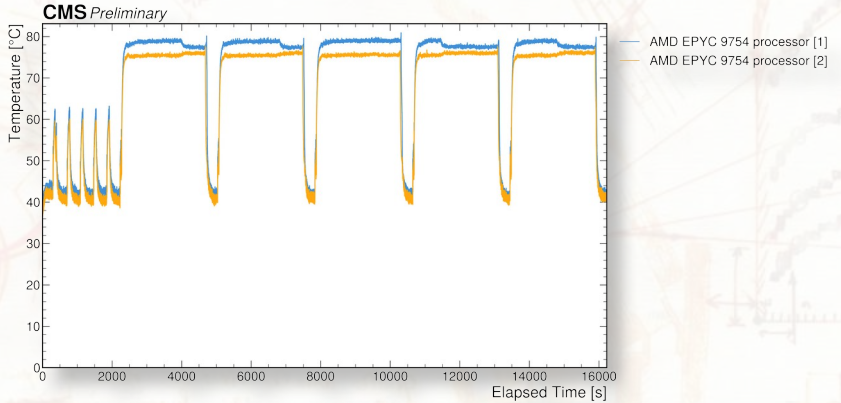
- The following plots show
  - on the *top left*, the memory usage of the whole system
  - on the *top right*, the temperature of each CPU and of each GPU (when used)
  - on the *bottom left*, the memory usage of each GPU (when used)
  - on the *bottom right*, the average utilisation of each GPU (when used)

- These monitoring data was collected while performing the timing, throughput and power consumption measurements.

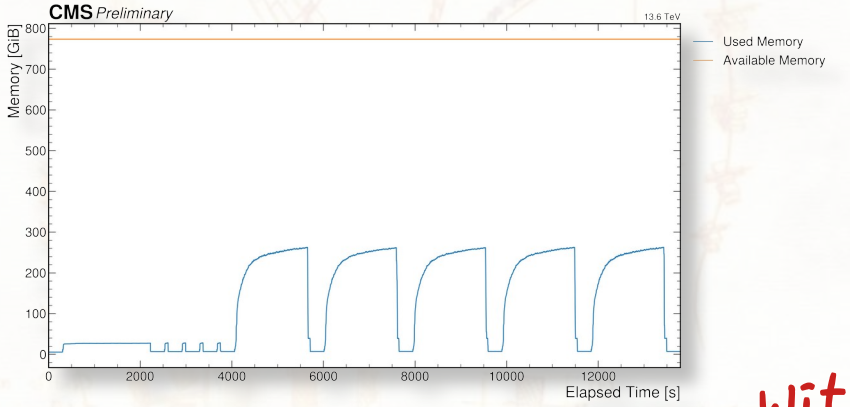
2024 HLT nodes – without GPUs

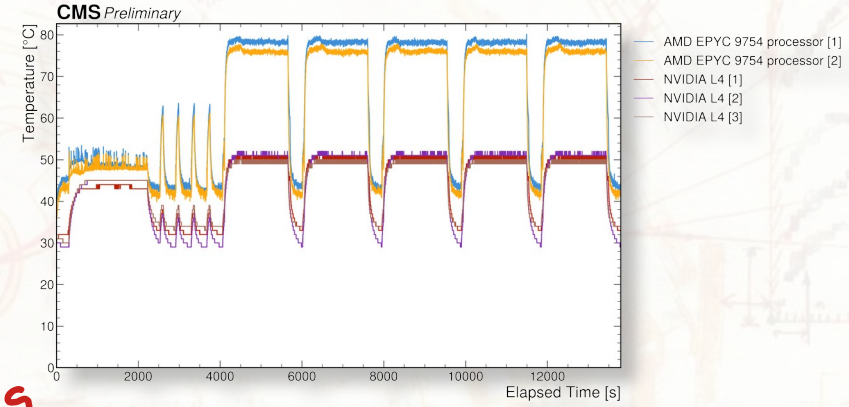


CPU only

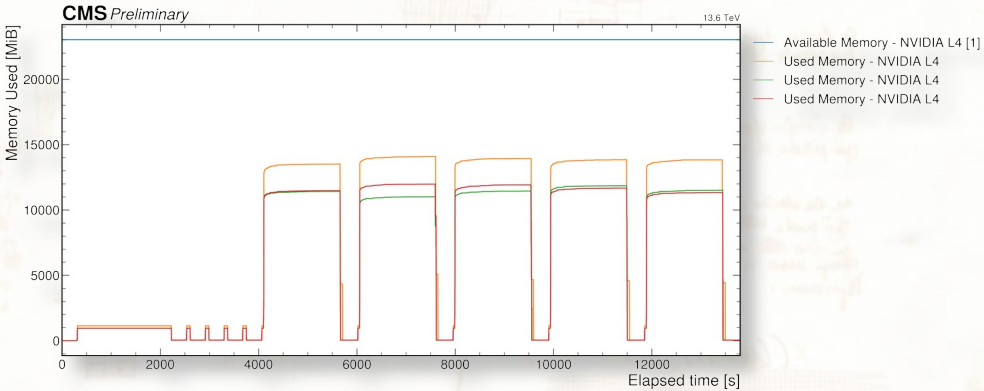


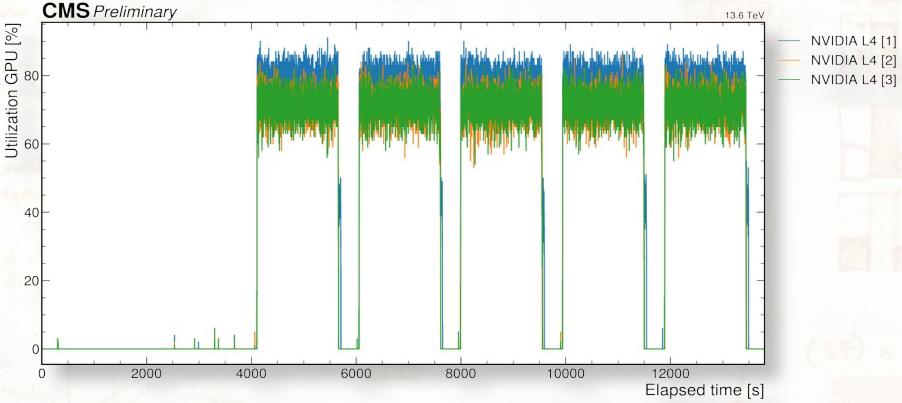
2024 HLT nodes – with GPUs





with GPUs





GPU vs CPU power efficiency

GPU power efficiency

- a “reduced” HLT configuration has been used to estimate the power efficiency of the reconstruction running on GPUs under optimal conditions
  - this configurations includes only the algorithms implemented in alpaka, described on slides 3 and 4
    - the Pixel unpacking, local reconstruction, track reconstruction, and vertex reconstruction [4]
    - the ECAL unpacking and local reconstruction [5]
    - the HCAL local reconstruction [6] and Particle Flow clustering [7]
  - plus a minimal amount of supporting modules necessary to run them
    - the source used to read the input data
    - the HCAL unpacking
- the measurement has been performed on a 2024 HLT node, as described in slide 21
  - with 16 CPU-only jobs running over 100k events
  - varying the number of available GPUs between 0 and 4
  - adding enough jobs running over 200k events on the GPUs to saturate them
- for each point the average throughput and power consumption have been measured
  - taking into account only the interval where all jobs were running

GPU power efficiency - results

	power consumption		throughput	
no GPUs	1050.0	± 0.5 W	1198.2	± 3.8 ev/s
1 GPU	1133.7	± 0.6 W	1690.5	± 8.4 ev/s
2 GPUs	1257.5	± 2.9 W	2229.5	± 6.1 ev/s
3 GPUs	1342.7	± 2.9 W	2764.7	± 2.7 ev/s
4 GPUs	1456.1	± 5.0 W	3222.7	± 8.4 ev/s

- linear fit
  - power draw: $N_{GPUs} \times 102.1 + 1043.8$ W
  - throughput: $N_{GPUs} \times 512.3 + 1196.5$ ev/s
- AMD EPYC "Bergamo" 9754 processor
  - power draw: 525.0 ± 0.2 W
  - throughput: 599.1 ± 1.9 ev/s
  - efficiency: 0.876 ± 0.003 J/ev
- NVIDIA L4 GPU
  - power draw: 102.1 ± 3.4 W
  - throughput: 512.3 ± 8.0 ev/s
  - efficiency: 0.199 ± 0.007 J/ev