Analysis results of the current benchmark data

Randall Sobie University of Victoria
(on behalf of the HEPiX Benchmark Working Group)
May 4 2022

Previous presentations:

Jan 19 2022	Results from the experiment workloads on the CERN tested
March 2 2022	HEPSpec06 (32/64) and SPEC2017 (CPP/INT) results
April 6 2022	Experiment workloads vs HEPSpec06 (32/64) and SPEC2017 (CPP/INT)

May 4 2022 Validation of the Gravity Wave (Ligo ..) benchmark
Study of “workloads vs workloads”

Analysis roadmap:

1. Reconfirm results of HEPSpec06 and SPEC2017

2. Validate workloads — *Other workloads?*
   - 9 Workloads validated on the CERN Testbed

3. Workloads vs HEPSpec06 and SPEC2017 — *No GW data*

4. **<u>Workloads vs Workloads (Today)</u>**
   - All and Top5 CPU architectures
   - Figure of Merits (W vs W)

5. HEPScore Candidates
   - Discussions in the HEPiX WG (selection criteria) — *See summary of this talk*
   - Building infrastructure to study candidates — *Good progress*
   - Need to finalize Workloads — *Critical path that will determine schedule*

https://rjsobie.web.cern.ch/rjsobie/benchmarks.html

Benchmark analysis results

Tables of HS06 and SPEC2017 results for each CPU-System (updated April 8 2022)

- Tables

Histograms of HS06 and SPEC2017 results from every server (updated April 8 2022)

- Thumbnail GIF plots
- HS-32bit PDF
- HS-64bit PDF
- SPEC2017_CPP PDF
- SPEC2017_INT PDF

HS06 and SPEC2017 2D plots per physical core (updated April 21 2022)

- HS06/SPEC 2D plots (per Physical-Core)

Tables of Workloads from every server (updated April 22 2022)

- Server table

Histograms of Workloads from every server (updated April 22 2022)

- Thumbnails of server histograms
- PDF of ATLAS sim_mt histograms
- PDF of ATLAS gen_sherpa histograms
- PDF of Belle2 histograms
- PDF of CMS gen_sim histograms
- PDF of CMS digi histograms
- PDF of CMS reco histograms
- PDF of LHCb gen_sim histograms
- PDF of Juno gen_sim_reco histograms

Fits of Workloads on 3 CERN Testbed servers (updated April 22 2022)

- Fits

Plot 2D (per physical core) Workloads vs HEPSpec06 and SPEC2017 (updated April 22 2022)

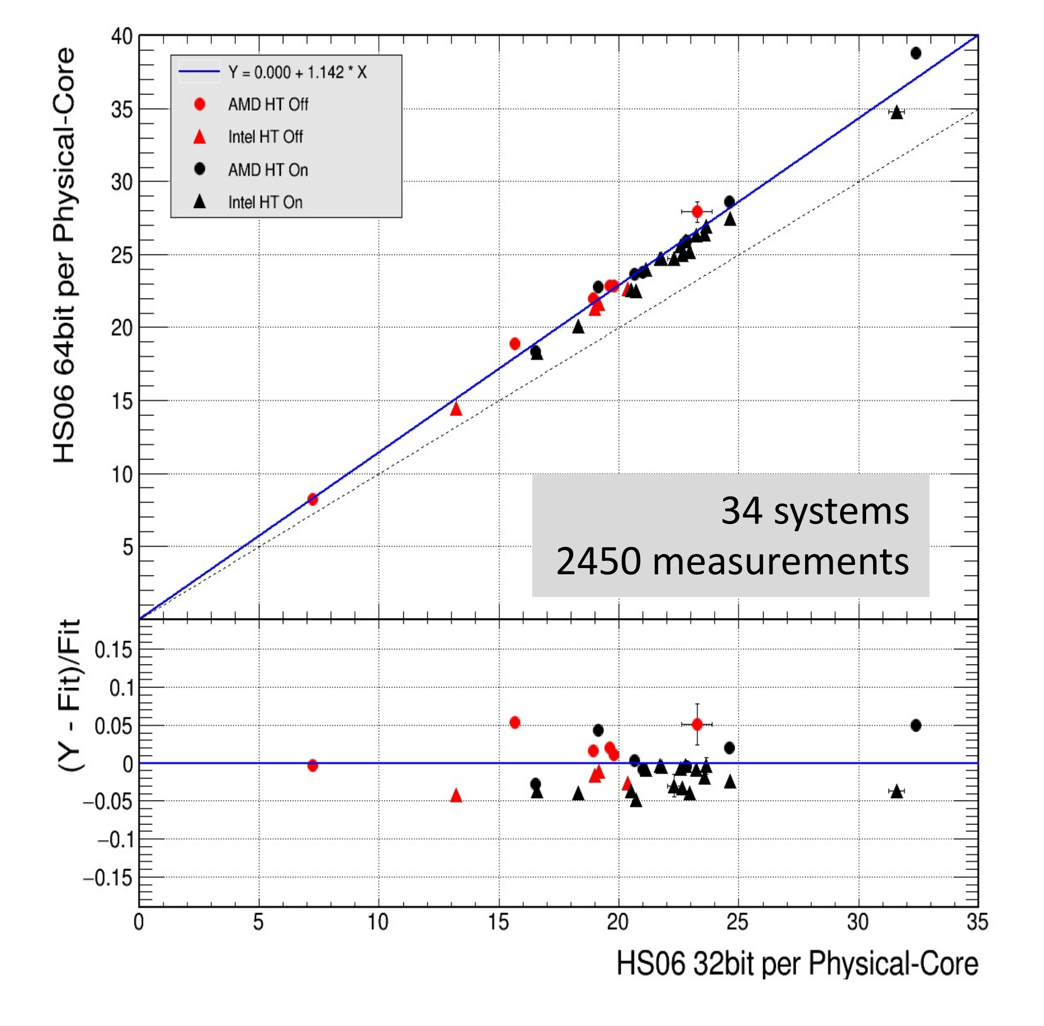
- HEPSPEC06 64bit
- HEPSPEC06 32bit
- SPEC2017 INT
- SPEC2017 CPP

Plot 2D (per physical core) Workloads vs Workloads (updated April 25 2022)

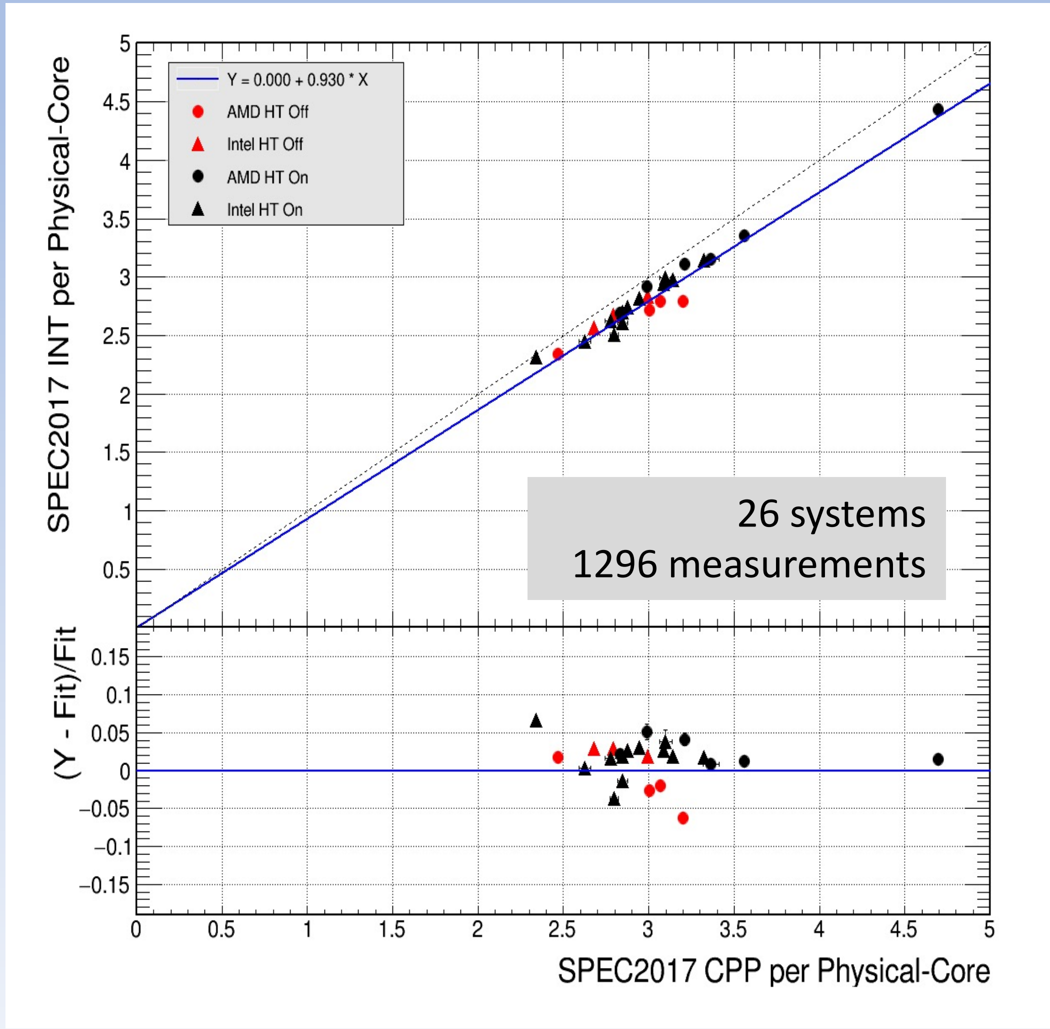
- W vs W plots (all CPU architectures)
- W vs W plots (Top5 CPU architectures)
- FOM (all CPU architectures)

1. Reconfirm results of HEPSpec06 and SPEC2017

HEPSpec06 (HS) 64 vs 32 bit



SPEC2017 (SP) intrate vs cpp

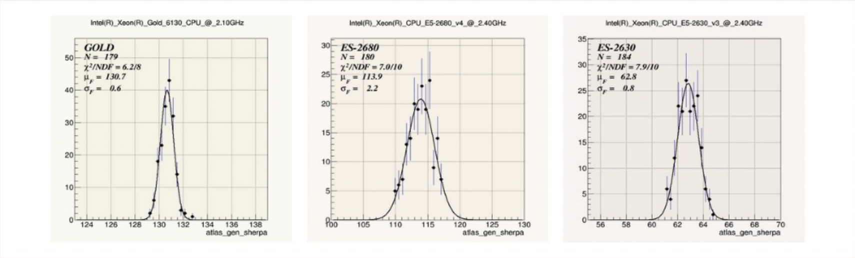


2 Validate workloads

Validated workloads: ATLAS(2), BelleII, CMS(3), LHCb and Juno

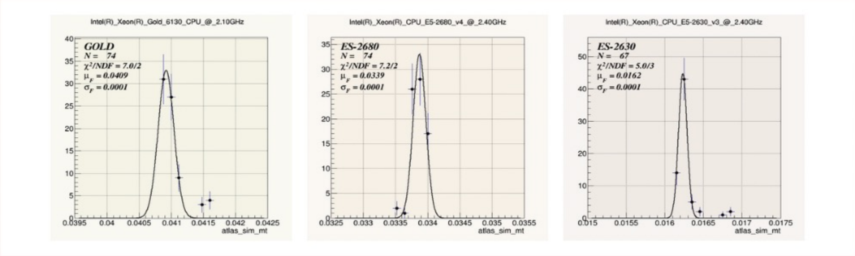
atlas_gen_sherpa

N	HMean	HStd	FSt	Chisq	NDF	FMean	FStd	CPU
179	130.7	0.6	0	6.15	8	130.7	0.6	Intel(R)_Xeon(R)_Gold_6130_CPU_@_2.10GHz-64cores
180	113.8	1.8	0	7.03	10	113.9	2.2	Intel(R)_Xeon(R)_CPU_E5-2680_v4_@_2.40GHz-56cores
184	62.8	0.8	0	7.93	10	62.8	0.8	Intel(R)_Xeon(R)_CPU_E5-2630_v3_@_2.40GHz-32cores



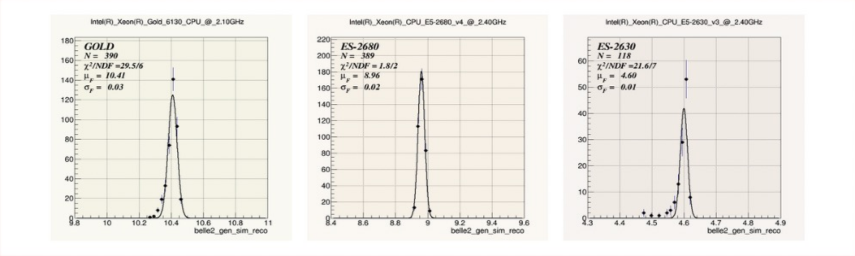
atlas_sim_mt

N	HMean	HStd	FSt	Chisq	NDF	FMean	FStd	CPU
74	0.0410	0.0002	0	7.00	2	0.0409	0.0001	Intel(R)_Xeon(R)_Gold_6130_CPU_@_2.10GHz-64cores
74	0.0339	0.0001	0	7.23	2	0.0339	0.0001	Intel(R)_Xeon(R)_CPU_E5-2680_v4_@_2.40GHz-56cores
67	0.0163	0.0001	0	4.96	3	0.0162	0.0001	Intel(R)_Xeon(R)_CPU_E5-2630_v3_@_2.40GHz-32cores



belle2_gen_sim_reco

N	HMean	HStd	FSt	Chisq	NDF	FMean	FStd	CPU
390	10.40	0.03	0	29.51	6	10.41	0.03	Intel(R)_Xeon(R)_Gold_6130_CPU_@_2.10GHz-64cores
389	8.96	0.02	0	1.75	2	8.96	0.02	Intel(R)_Xeon(R)_CPU_E5-2680_v4_@_2.40GHz-56cores
118	4.59	0.02	0	21.60	7	4.60	0.01	Intel(R)_Xeon(R)_CPU_E5-2630_v3_@_2.40GHz-32cores



Confirm the reliability and reproducibility

Run workload on 3 CERN "testbed" servers

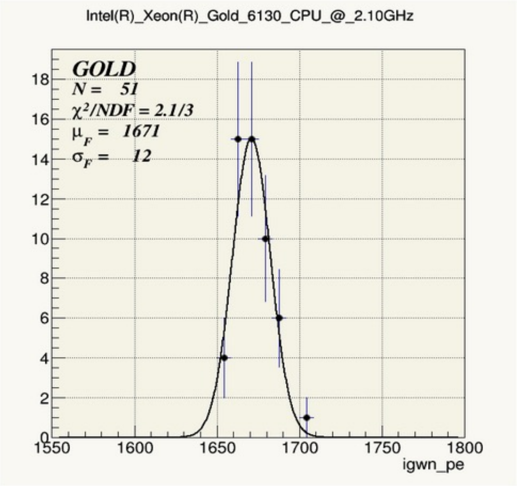
Fit results with Gaussian distributions

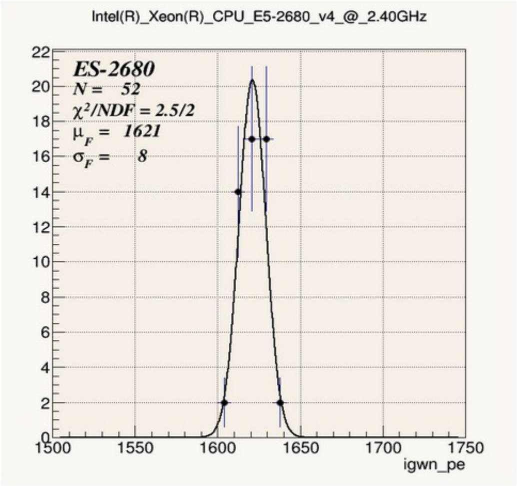
RMS of distributions typically <<1%

2 Validate workloads – new gravity wave (Ligo et al) workload

igwn_pe (Ligo gravity wave)

N	HMean	HStd	FSt	Chisq	NDF	FMean	FStd	CPU
51	1672	10	0	2.11	3	1671	12	Intel(R)_Xeon(R)_Gold_6130_CPU_@_2.10GHz-64cores
52	1621	8	0	2.55	2	1621	8	Intel(R)_Xeon(R)_CPU_E5-2680_v4_@_2.40GHz-56cores

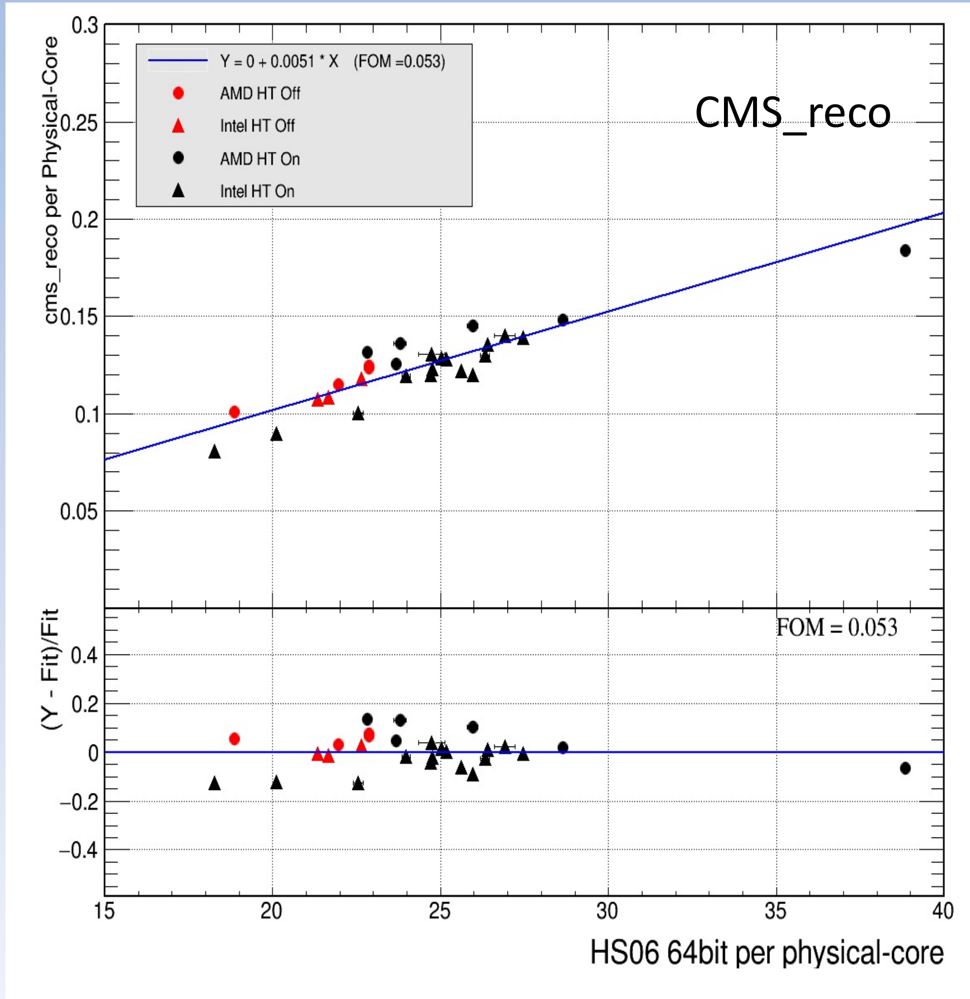


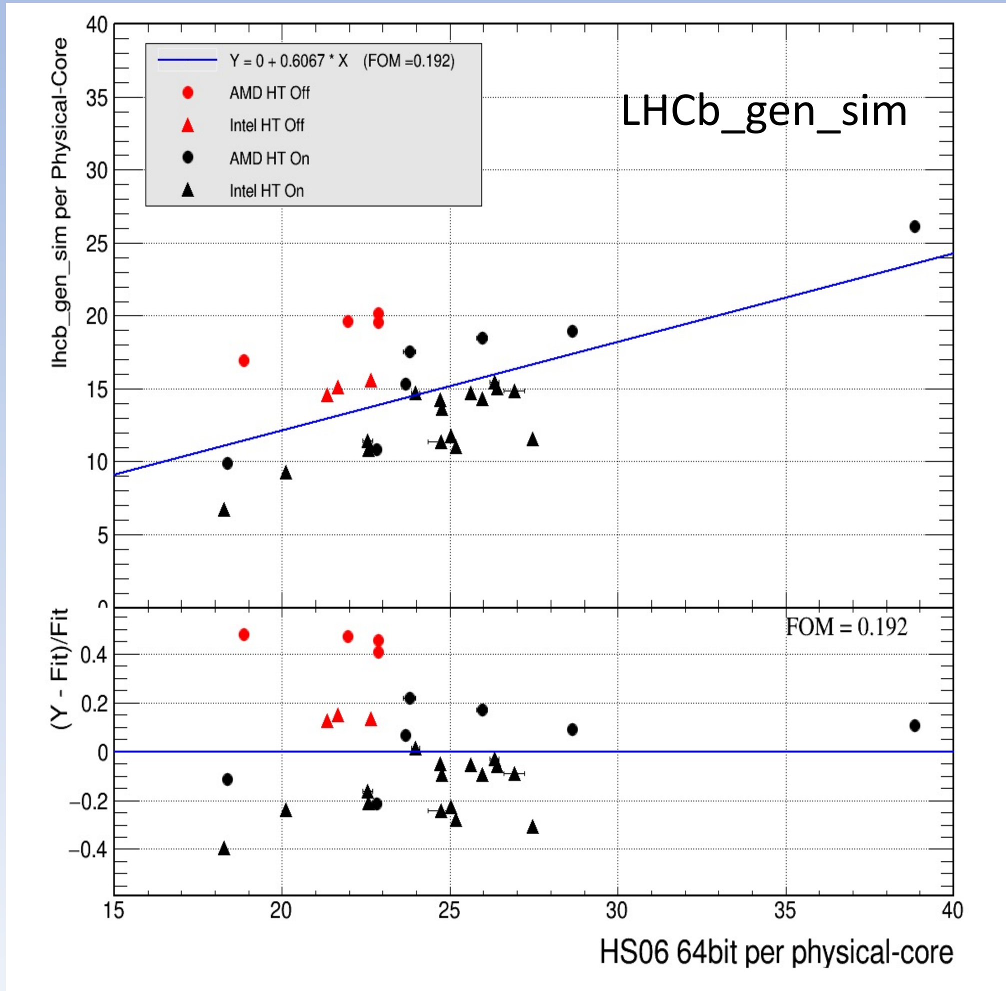


One of the CERN testbed servers is no longer available.

3 Workloads vs HEPSpec06 and SPEC2017

Each workload is compared with the 4 HS/SP benchmarks
Upper plot: events/second VS HS-64 (normalized per physical core)
Lower plot: "residuals" – relative difference of Y-benchmark to the blue fit line



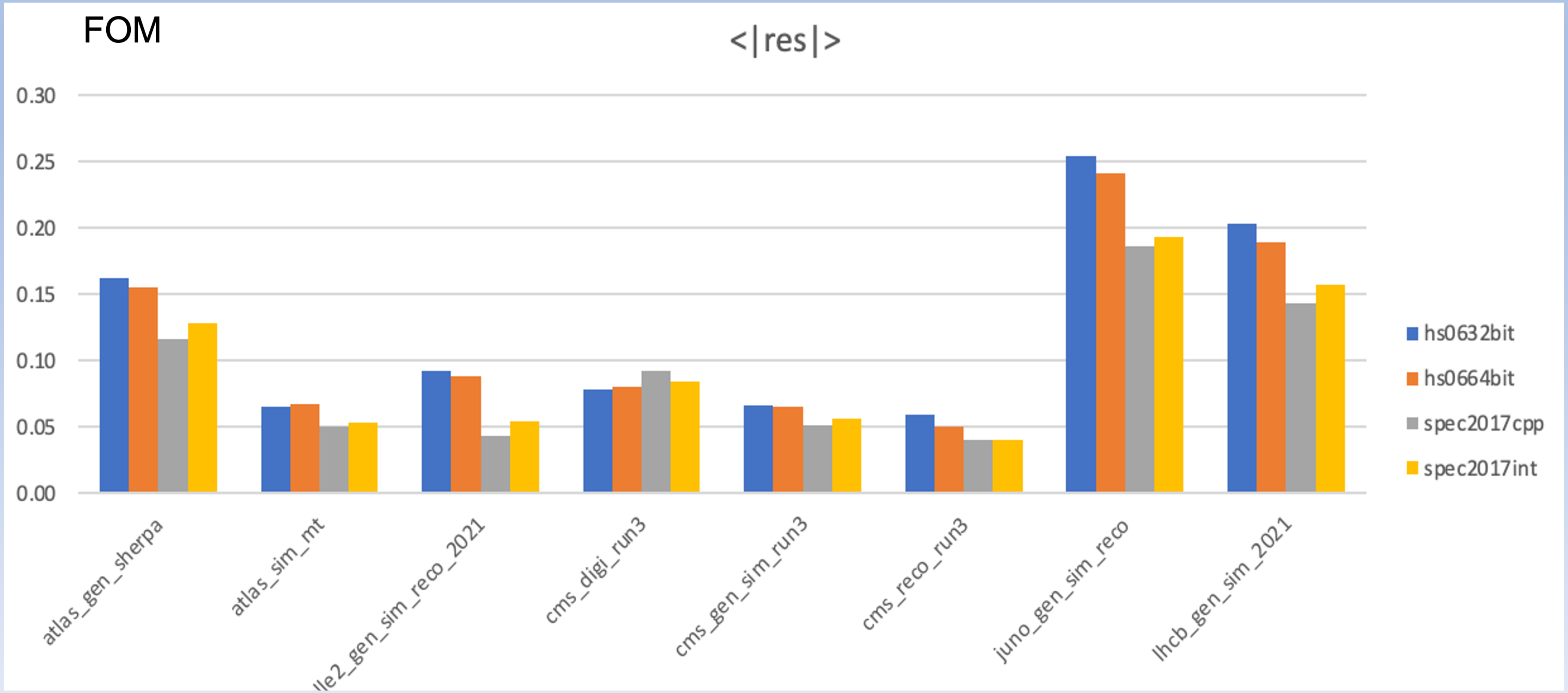


FOM
Figure of Merit
(Y-YFit)/YFit

Mean fractional deviation of the Y-bmk from the blue-fit line

Figure of Merit

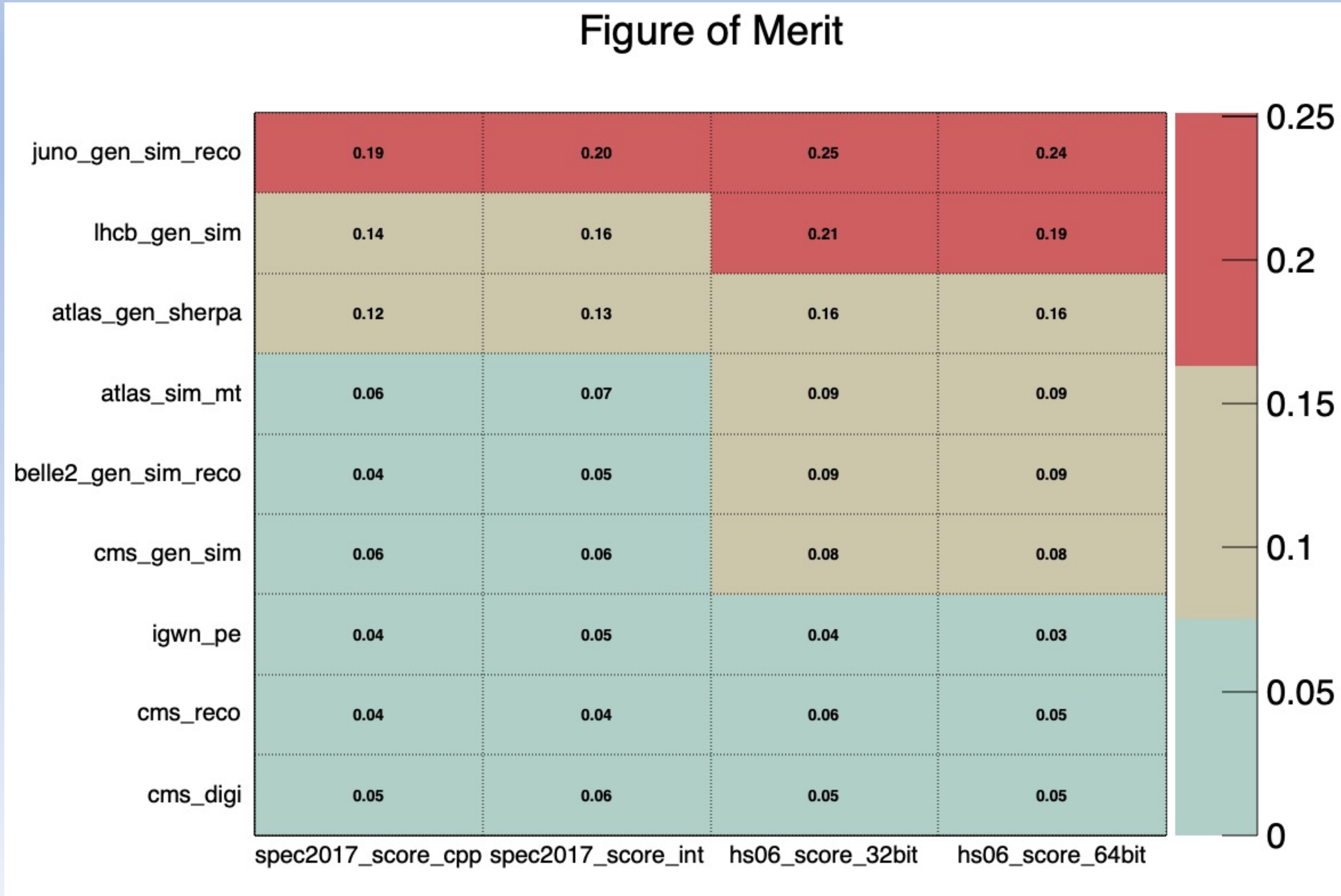
Histogram of FOMs for each Workload and each Benchmark



Value of the FOM reflects the scatter between the Workload and HS/SP benchmark
The scatter observed in the Workload-Workload plots will mirror these results

FOM for HEPSpec06/SPEC2017 vs Workloads

(ignore the gravity wave igwn_pe benchmark – only 2 CPUs so far)



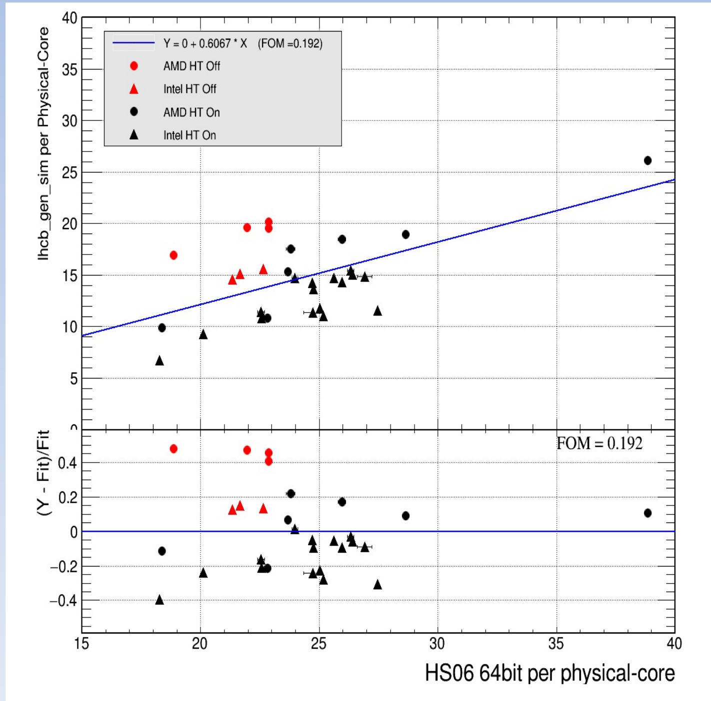
Workloads agree more with SPEC2017

Observation:

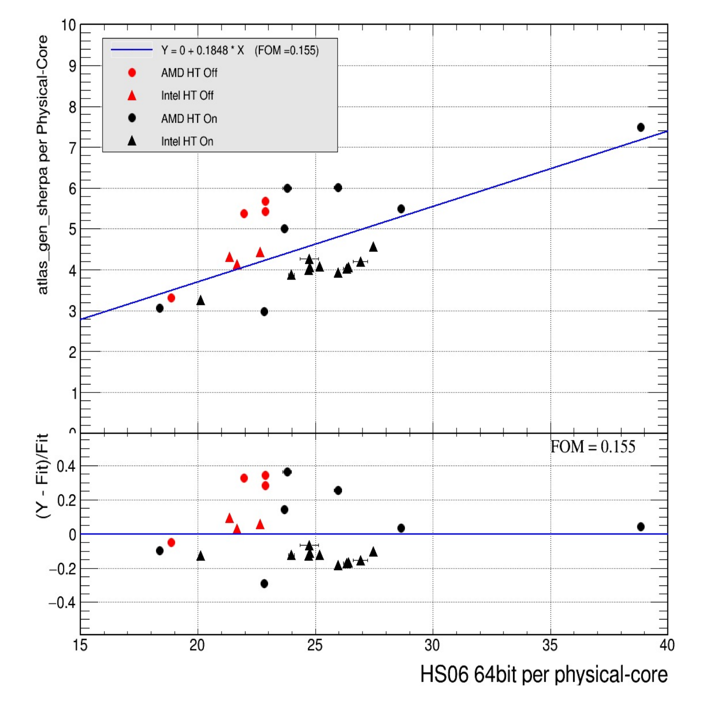
Many workloads have a higher value per physical-core with hyper-threading disabled compared with HS/SP
(More events/second are generated and processed with HT-off)

All plots below are single core applications

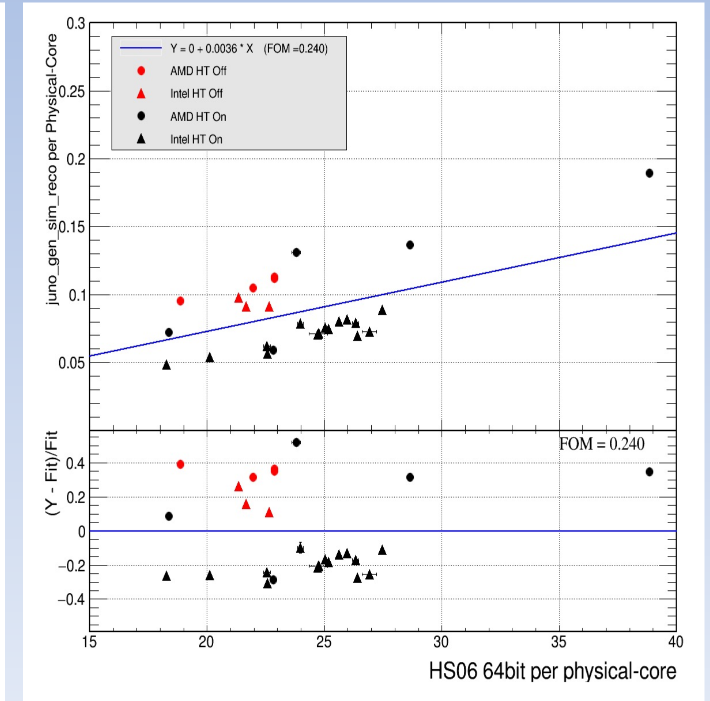
Red= AMD/Intel with HT-off
Blue = AMD/Intel with HT-On



LHCb_gen_sim

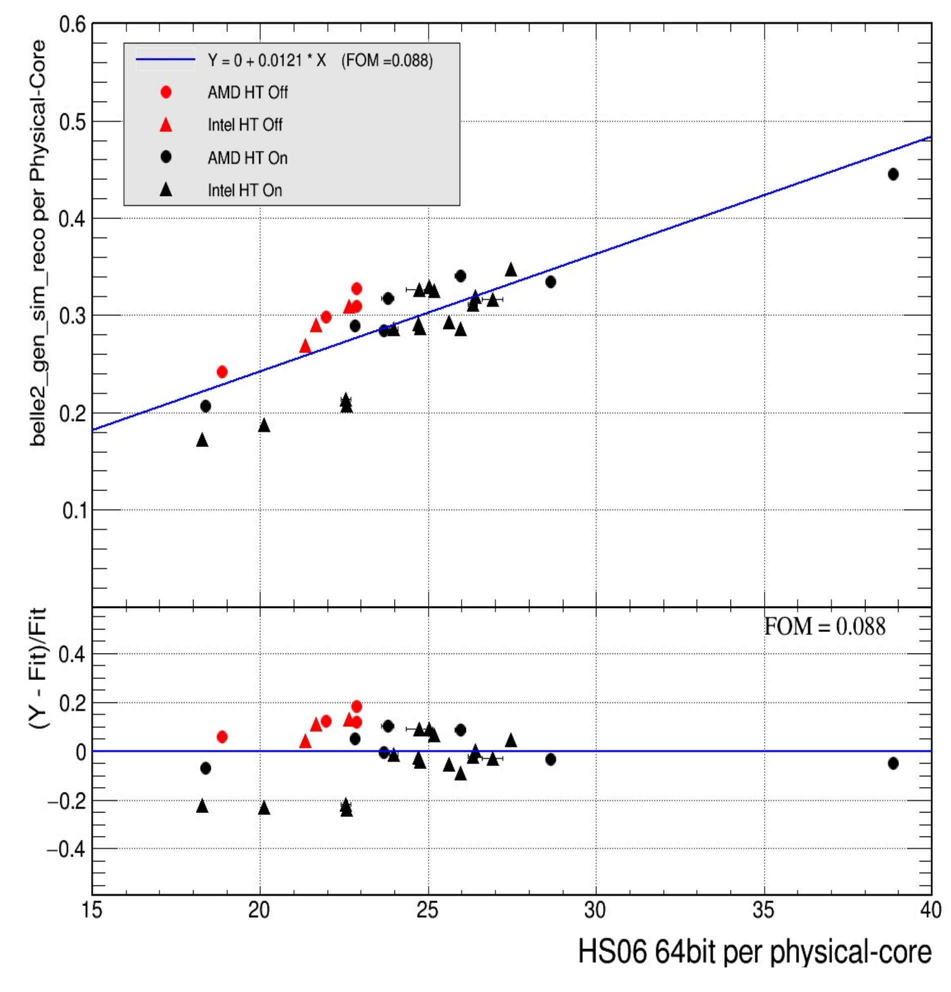


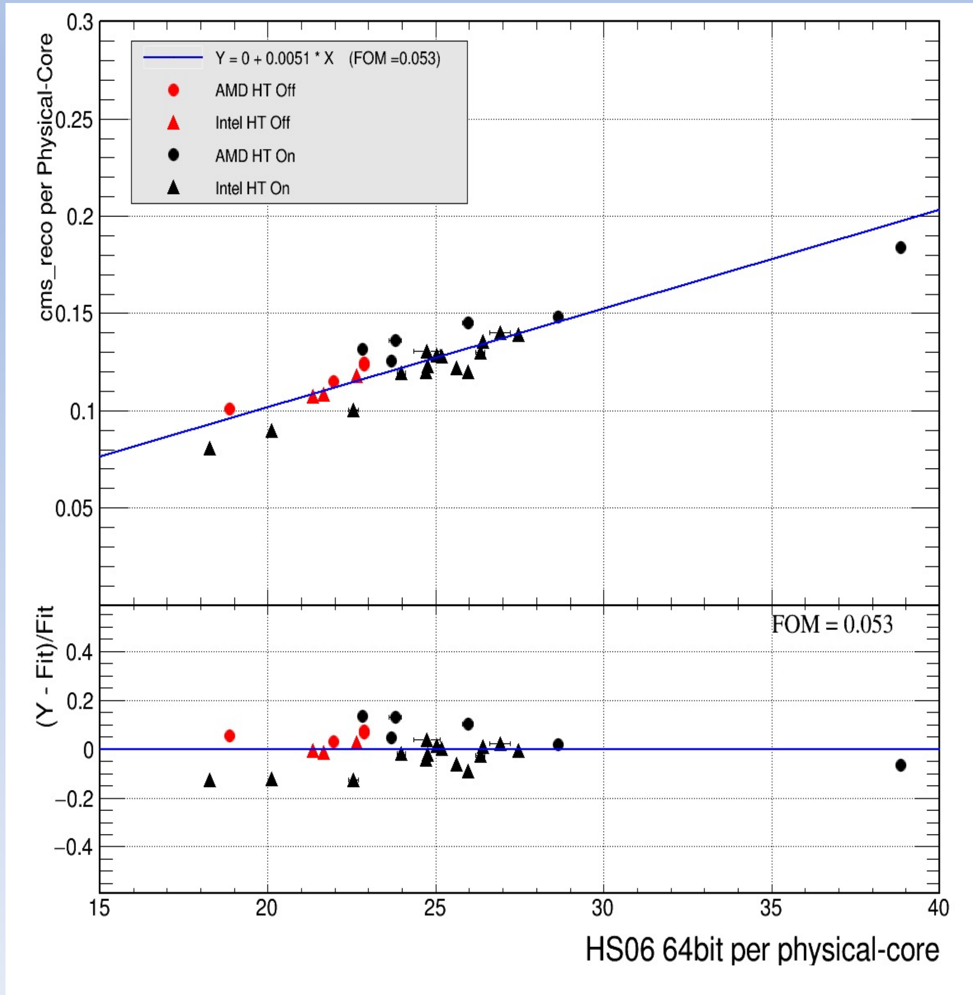
ATLAS_gen_sherpa



Juno_gen_sim_reco

Hyper-threading dependence less for BelleII and none for CMS_reco

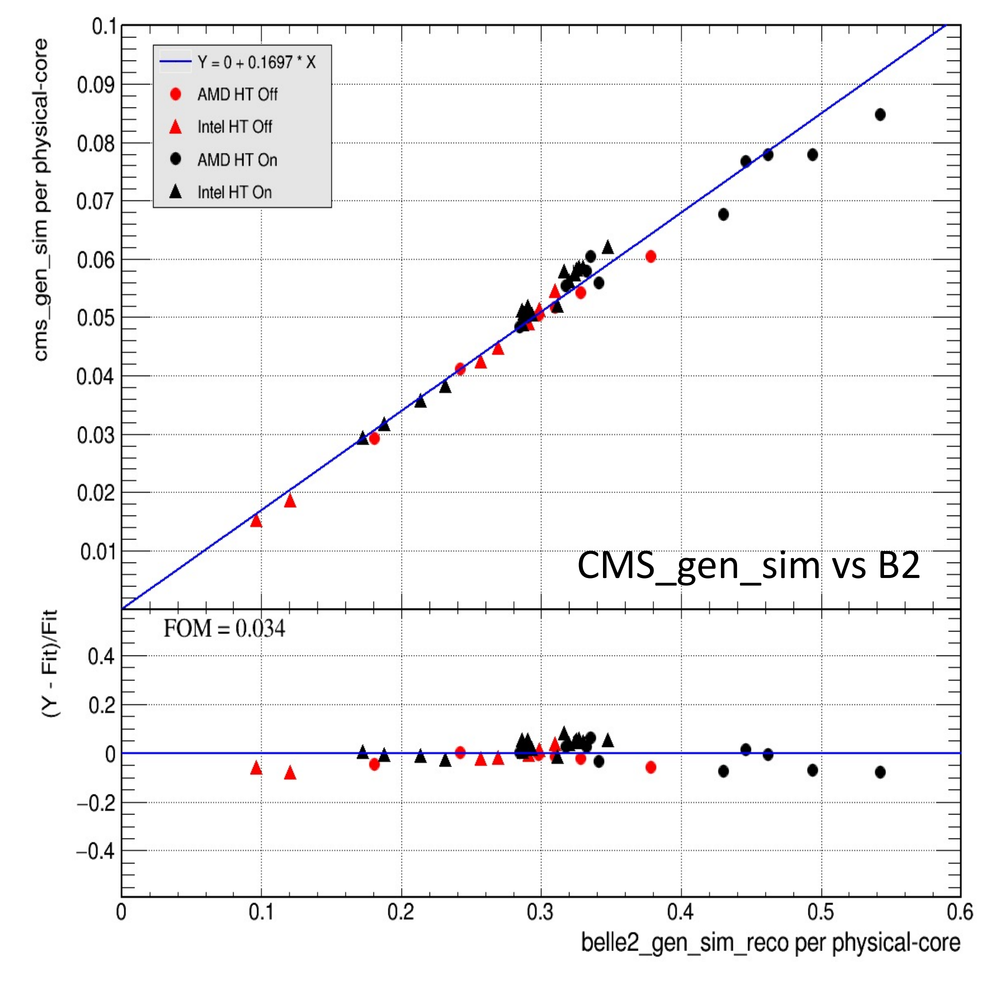


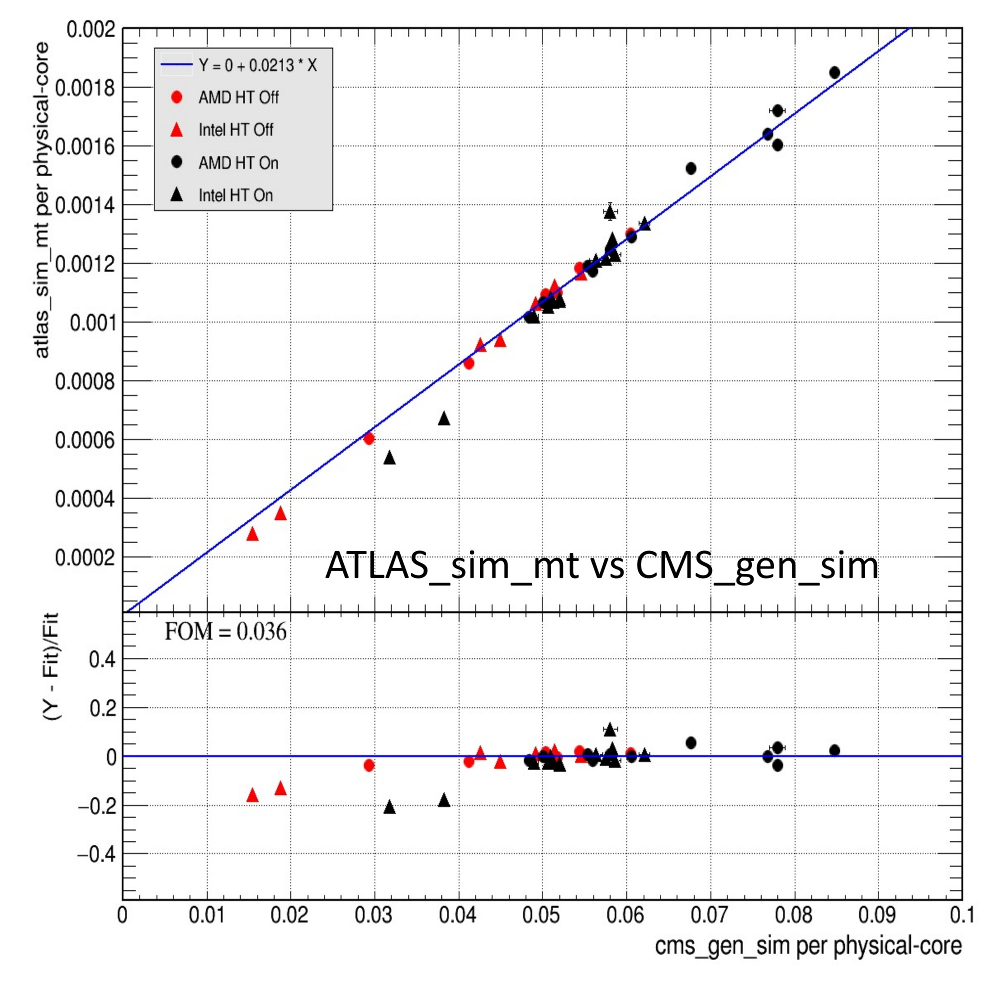


Less difference between HT-on/off for SPEC2017 benchmarks

4. Workload vs Workloads (28 combinations for the 8 workloads)

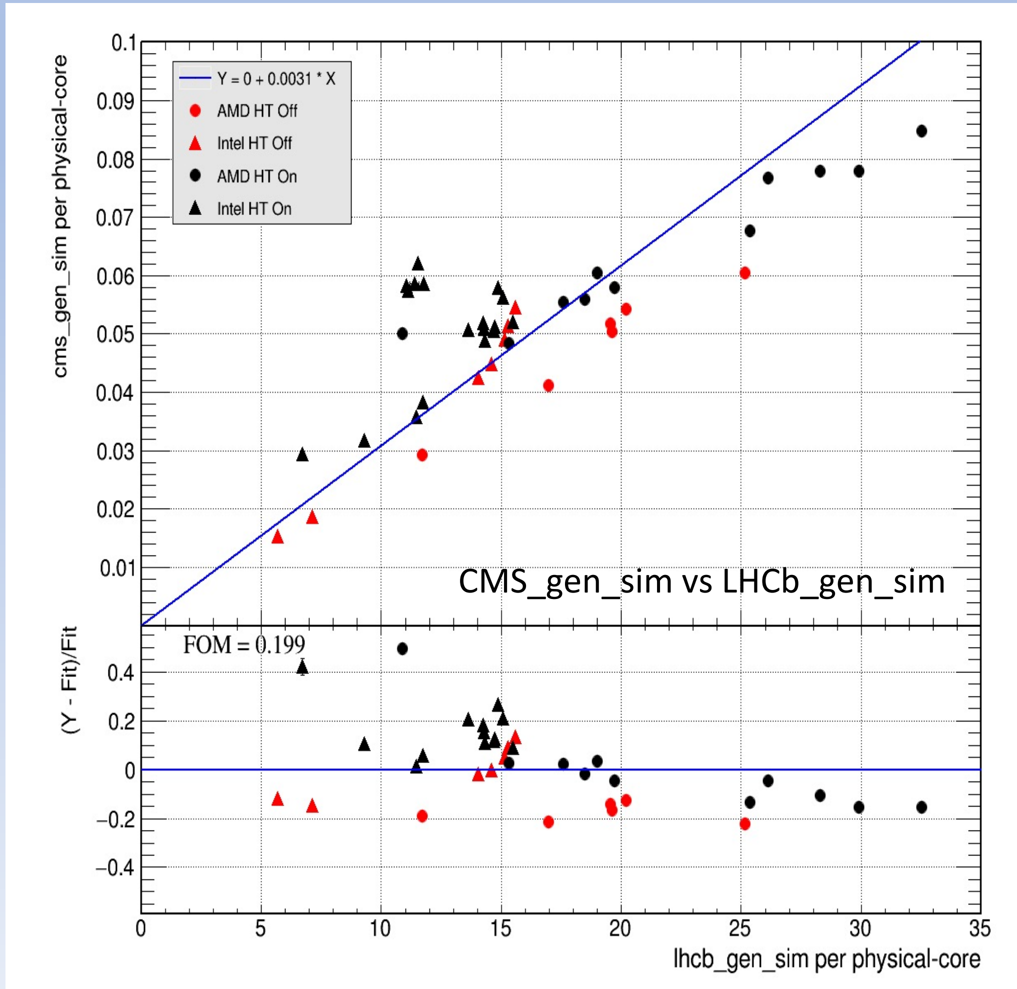
Nice results:

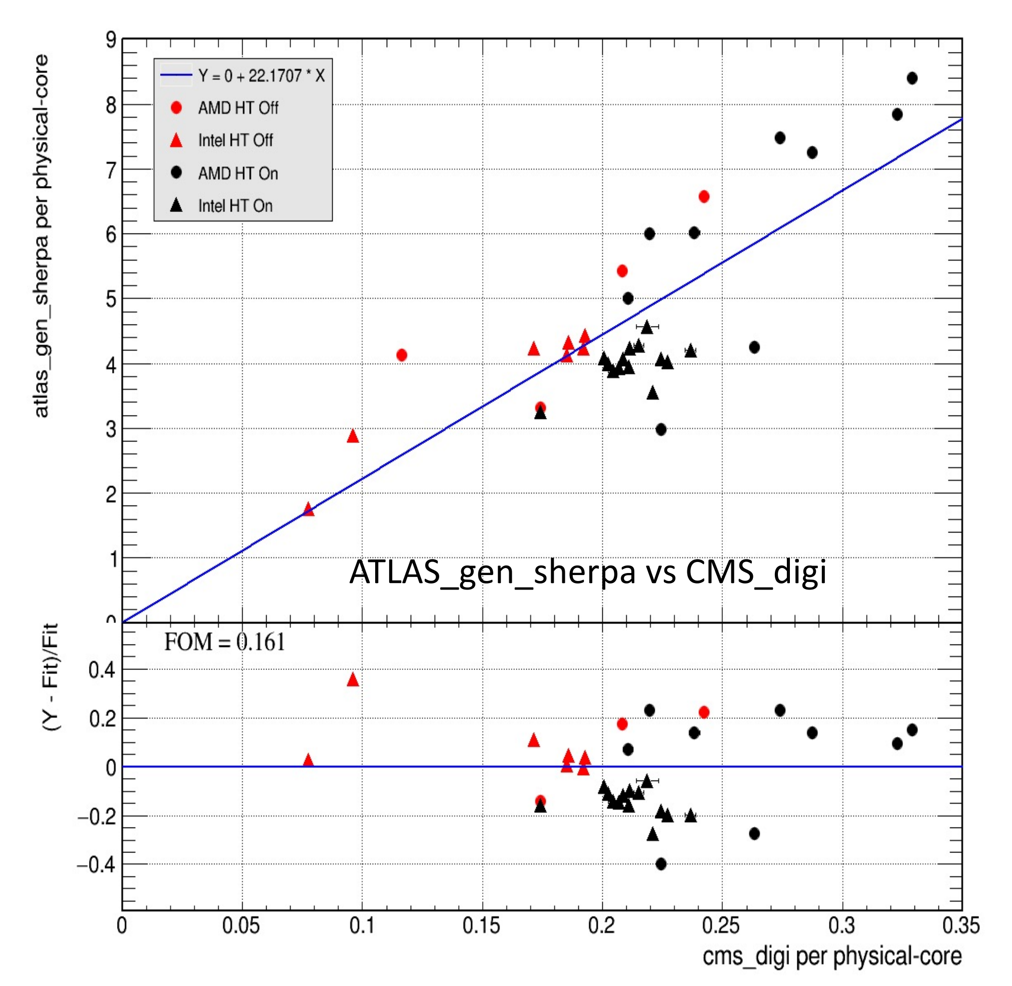




Observation: good agreement for the "simulation" applications (GEANT4?)

Not so nice results: CMS_gen_sim vs LHCb_gen_sim ATLAS_gen_sherpa vs CMS_digi





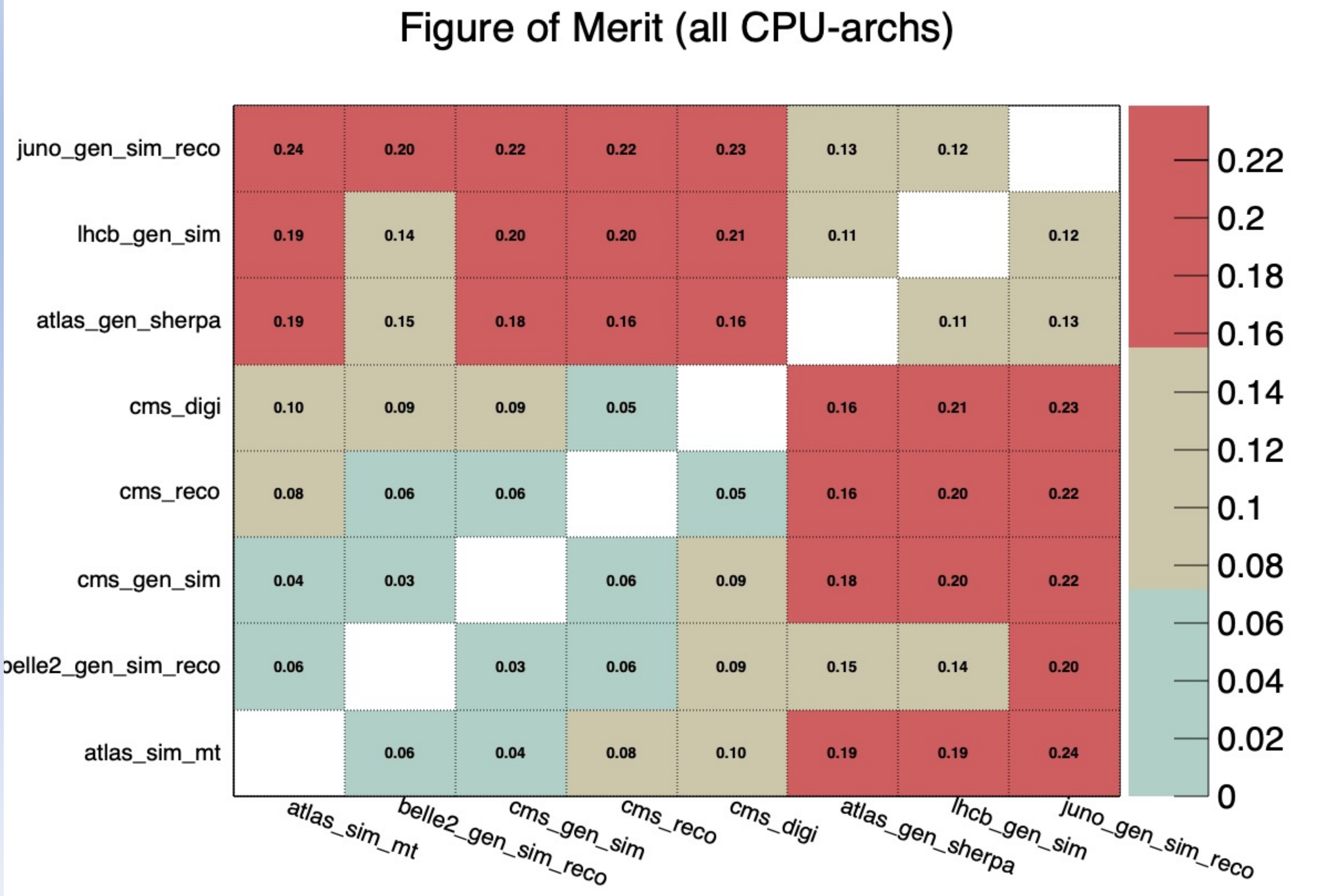


Figure of Merit

Average fractional deviation of the Y-axis benchmark from the fit

Two sets emerge:

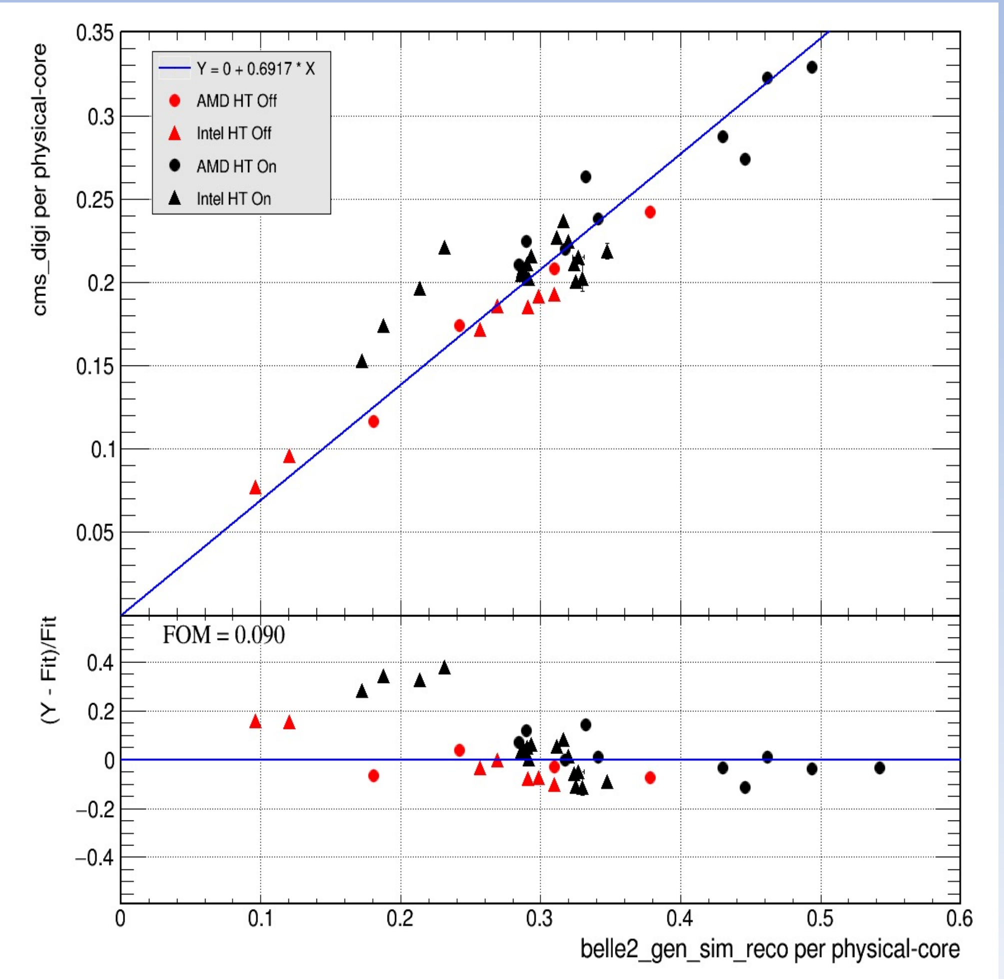
ATLAS_sim_mt
Belle2_gen_sim_reco
CMS_gen_sim
CMS_reco
CMS_digi

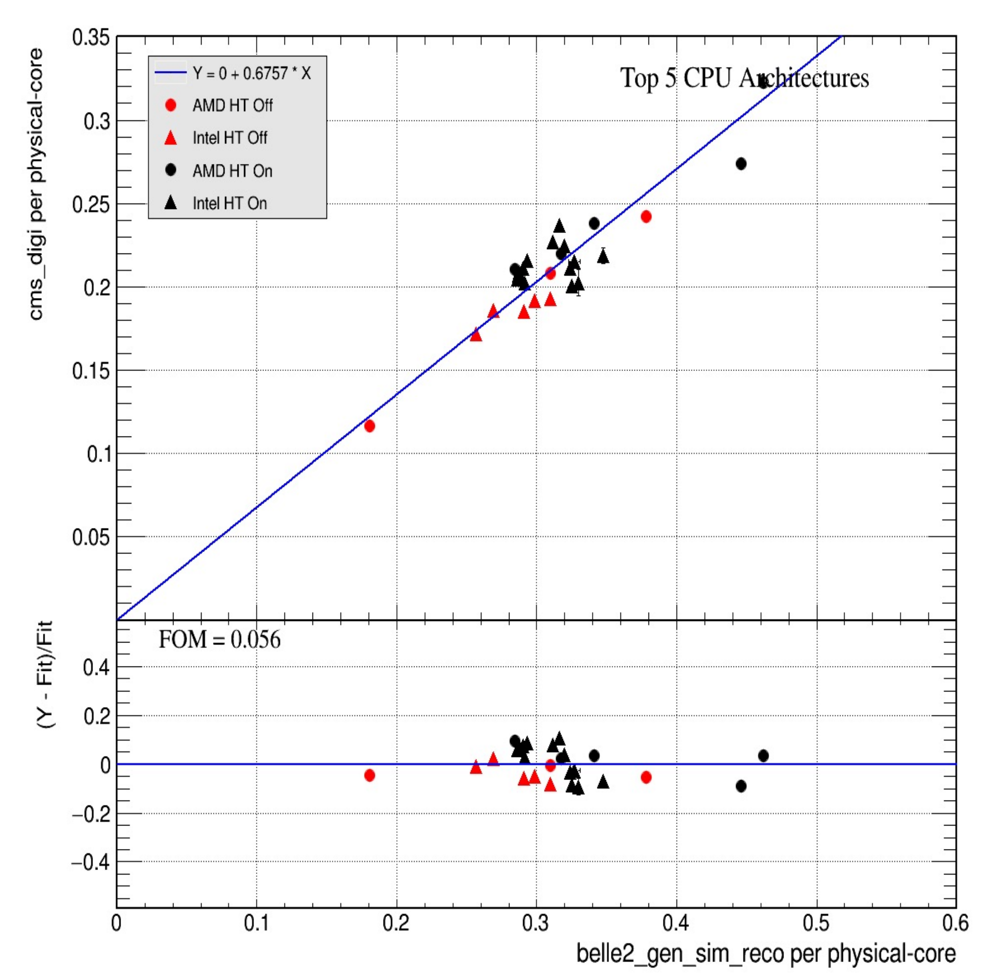
ATLAS_gen_sherpa
LHCb_gen_sim
June_gen_sim_reco

CPU Architecture

Top5 = (Rome, Broadwell, Haswell, Cascade Lake, Skylake)

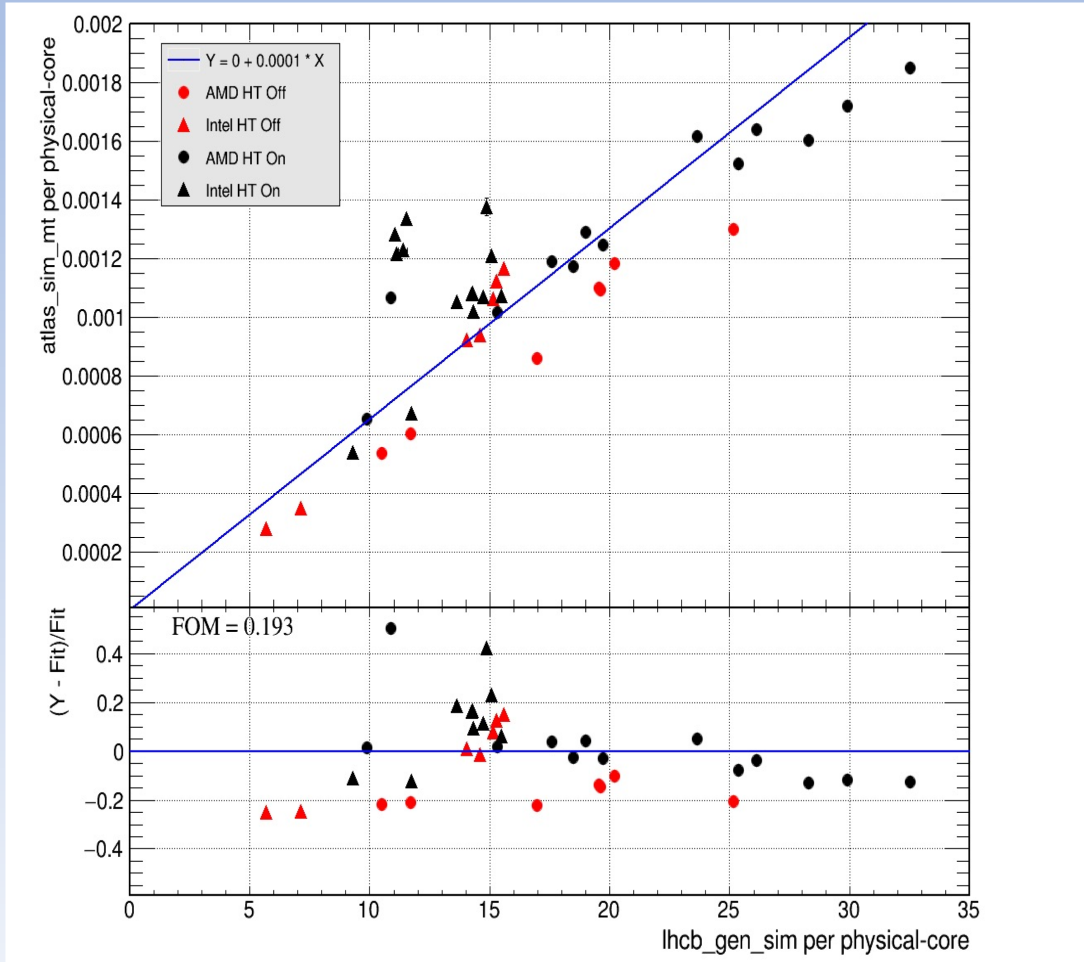
Example: CMS_digi vs Belle2_gen_sim_reco (FOM drops from 0.090 to 0.056)

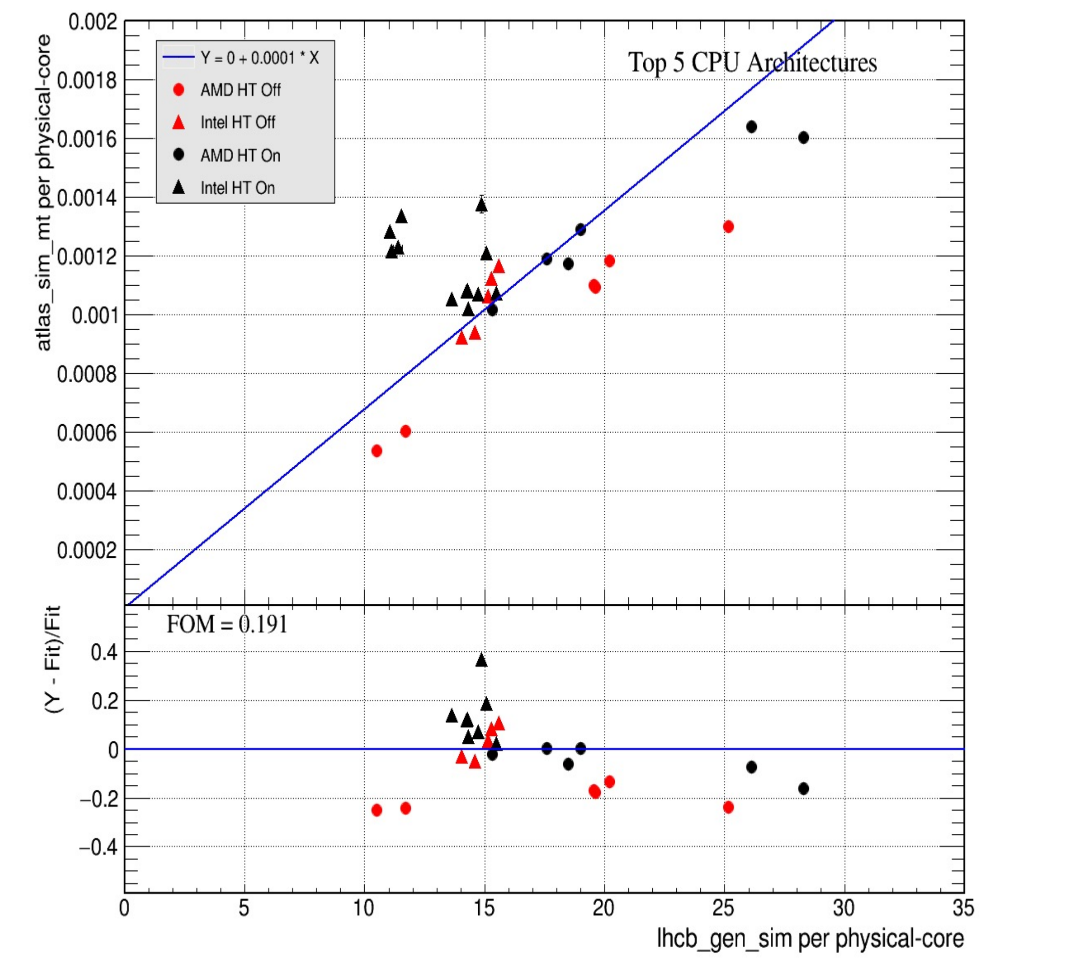




But it does not remove all the scatter …

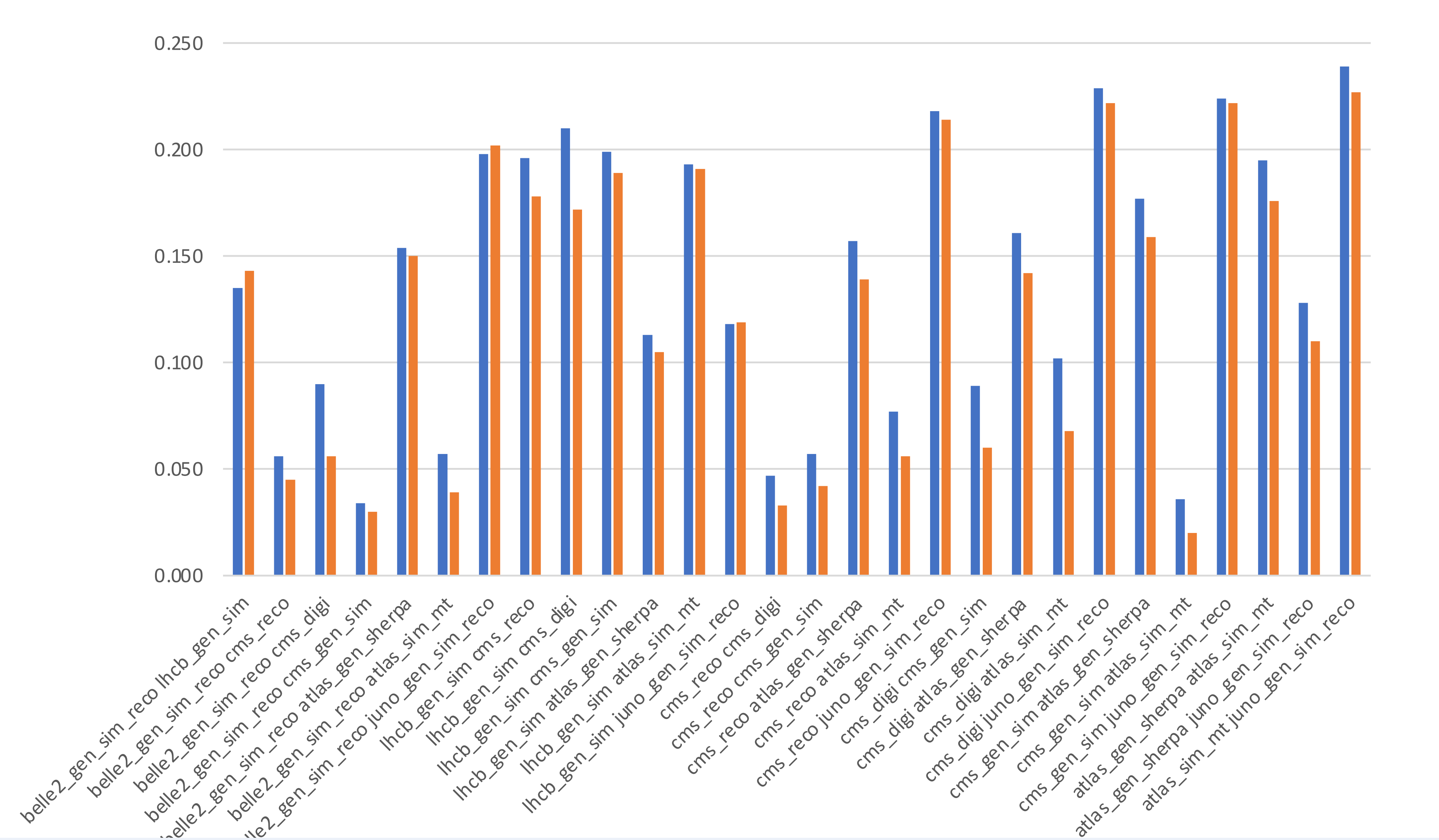
For example, ATLAS_sim_mt vs LHCb_gen_sim_reco (FOM stays constant 0.19)





All CPU-archs (blue) vs Top5 CPU-archs (red)

Reduces the scatter in many cases with FOM dropping by a third but the earlier conclusions are unchanged



5. HEPScore

- Starting to create HEPScore candidates
  - Optimal strategy for creating a long-term benchmark?

Observations

- Simulation is the dominant user of the CPU (and in good agreement)
  - CMS_gen_sim 40% of the 3 CMS benchmarks
  - ATLAS_sim_mt, ATLAS_gen_sherpa are 1000s/event and 0.3s/event, respectively
  - CMS_gen_sim 20s/event and ATLAS_gen_sim 1000s/event

- Some workloads are more performant with hyperthreading off

- Some workloads have large variations for different CPUs

Notes

- We need to finalize the workloads (time is needed to validate, accumulate data and analyze)