A comparison of HEPSPEC benchmark performance on ATLAS Grid-Sites versus ideal conditions

Michael Boehler on behalf of the ATLAS Collaboration
Physikalisches Institut, Hermann-Herder-Str.3, D-79104 Freiburg
E-mail: michael.boehler@cern.ch

Abstract. The goal of this study is to understand the observed differences in ATLAS software performance, when comparing results measured under ideal laboratory conditions with those from ATLAS computing resources on the Worldwide LHC Computing Grid (WLCG). The laboratory results are based on the full simulation of ttbar events and use dedicated, local hardware. In order to have a common and reproducible base to which to compare, thousands of identical ttbar full simulation benchmark jobs were submitted to hundreds of Grid sites using the HammerCloud infrastructure. The impact of the heterogeneous hardware of the Grid sites and the performance difference of different hardware generations is analysed in detail, and a direct, in-depth comparison of jobs performed on identical CPU types is also done. The choice of the physics sample used in the benchmark is validated by comparing the performance on each Grid site measured with HammerCloud, weighted by its contribution to the total ATLAS full simulation production output.

1. Introduction

The ATLAS [1] experiment uses a worldwide network of distributed data centres for data processing, linked together in the Worldwide LHC Computing Grid (WLCG) [2]. These centres each comprise a variety of CPU architectures and generations. To properly compare the work done on the different sites, the time spent performing the work is normalised based on CPU benchmark results. The metric used to measure the CPU performance is the HEP-SPEC06 benchmark. It is based on the cpp benchmark subset of the SPEC CPU2006 [3] benchmark suite. For each CPU type on each site the HEP-SPEC06 score has to be determined with this benchmark. The score divided by the number of cores used in the benchmark, yields the so-called *corepower* value. Since many sites operate different CPU types and the corepower is one unique value per queue, a weighted average corepower value considering the fraction of the different CPU types and their corepower value needs to be derived. The metric $hs06$, with unit HEP-SPEC06 seconds per event, is then calculated as follows:

$$hs06 = \frac{t_{walltime} \times n_{cores} \times corepower}{n_{events}} \tag{1}$$

where $t_{walltime}$ is the wall-time (time from start to end of a job), n_{cores} is the number of cores per job, and n_{events} is the number of events per job. The hs06 value is stored in the database of the ATLAS workload management system (PanDA [4]), for each ATLAS job performed on the WLCG. The ATLAS Software Performance Optimization Team (SPOT) continually measures

Table 1. The table lists the number of jobs (nJobs), the number of queues (nQueues) and the number of different CPU types (nCPU) at each selection step.

	Cut	nJobs	nQueues	nCPU
0	total	102066	96	154
1	no TEST queue	98196	89	154
2	nJobs per CPU & queue >= 25	96805	86	125
3	total nJobs per queue >= 50	96757	85	125

the CPU time per event for different workloads. One of these workloads is the full simulation (full sim) of top-anti-top-quark pairs ("ttbar"), which is a "standard-candle" sample, representing a complex process with activity in each sub-detector. A large discrepancy has been found when comparing the hs06 value averaged over all successful full sim production jobs in 2020 (4.7 kHS06 sec/evt) with results from the SPOT performance tests (3.0 kHS06 sec/evt). The aim of this study is to identify the root cause(s) of this large discrepancy.

2. Analysis

The global hs06 value averaged over all ATLAS full sim production jobs, comprises a variety of physics processes, which may vary in event processing time. To form a better basis for comparison, a dedicated test template, containing the same setup as the ttbar full sim jobs evaluated by SPOT, has been added to HammerCloud (HC) [5], a test and benchmarking infrastructure used widely by ATLAS and other experiments. This allows many identical jobs to be sent to all Grid sites worldwide, as is required. In PanDA, jobs are assigned to "queues", which typically represent a single physical computing resource in a specific WLCG site. The meta-data of ttbar simulation jobs submitted through HC was analysed in order to quantify the performance differences between the Grid queues and the results from SPOT. The HC test jobs were successfully submitted to a total of 96 different queues. To ensure that these jobs have sufficient memory available and finish in a reasonable time, all jobs were submitted as multicore (8-core) jobs with 2 GB RAM per core. In total 102k successful ttbar jobs were executed. In order to guarantee sufficient statistics, only CPU types with at least 25 successful jobs on a given queue and only queues with at least 50 successful jobs are considered. This queue selection is described in Table 1, 96k HC test jobs and 85 queues have met the criteria. The data was enriched by the release date of the individual CPU types [6, 7, 8].

3. Results

Since each ATLAS Grid job publishes its corresponding hs06 value, the performance of an individual queue can be retrieved by sorting the benchmark jobs by queue and calculating the average hs06 value per queue. The mean value of all HC benchmark jobs is 4.3 kHS06 sec/evt. Compared to the global average of all full sim events generated in 2020 from the ATLAS production system (4.7 kHS06 sec/evt), it shows a smaller discrepancy compared to the SPOT measurements (3.0 kHS06 sec/evt), but still differs by 1.3 kHS06 sec/evt. Figure 1 shows the average hs06 per queue (dots) measured by the benchmark jobs with the standard deviation as error bar.

3.1. Detailed study of queues with multiple CPU types

Since the number of cores per job and the number of processed events are constant for the benchmark jobs, the hs06 value is proportional to the wall-time multiplied by the corepower value. One single corepower value per queue cannot properly account for differences in the wall-time distribution due to different hardware generations combined into one queue as shown

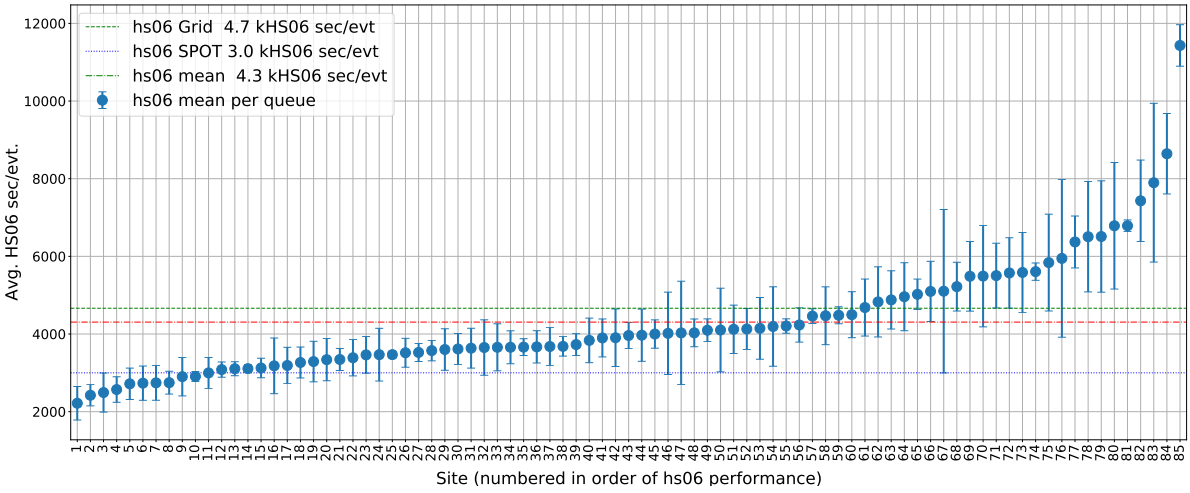

Figure 1. Mean hs06 value per queue. Horizontal lines indicate mean values of full sim 2020 ATLAS production (dashed), SPOT results (dotted), and HC benchmarks (dash-dotted).

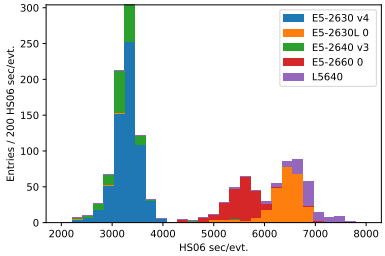

Figure 2. hs06 value of queue with 5 CPU types.

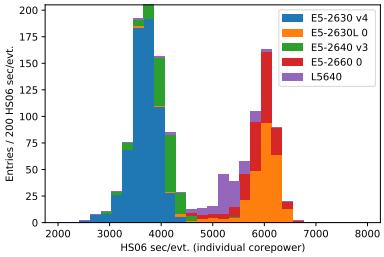

Figure 3. Recalculated hs06 with individual corepower.

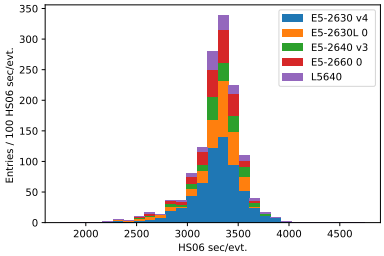

Figure 4. hs06 value scaled to latest CPU type.

in Figure 2. Different corepower values per CPU type should correct for differences in the performance and yield one peak in the hs06 distribution. Figure 3 shows the recalculated hs06 according to Equation 1, replacing the corepower values per queue by the individual corepower values per CPU type. The re-computed hs06 distribution still does not show a single peak, meaning that the corepower values determined for the different CPU generations do not properly reflect the processing performance. In order to quantify the deviation of the benchmark results from the test job performance, the target hs06 value was fixed (here the hs06 value of the latest CPU type of the queue) and the corepower value was adjusted accordingly. The results of this method are shown in Figure 4. All relevant numbers: the release date per CPU type, the original corepower, the recalculated corepower, the corepower values scaled to the latest CPU, and the decrease in percent are listed in Table 2. Scaling the hs06 value based on the benchmark performance to the latest CPU type would reduce the hs06 values of old hardware up to 50%, as indicated in the last column of Table 2. To conclude, the HEPSPEC06 benchmark is not a sufficient metric to compare HEP workload performances of CPU architectures spanning several decades. A more modern benchmark suite, which considers representative HEP-benchmarks is necessary and is already in development - the HEPSCORE benchmark framework [9].

3.2. Closure test

In order to rule out "Grid effects", one can compare Grid jobs executed on the identical CPU type as used in the SPOT measurements. This CPU type is the *Intel(R) Xeon(R) CPU E5-2630 v3 @*

Table 2. Comparison of the release year of the CPUs, the hs06 values, the corepower and the recalculated hs06 values, the corepower values re-scaled to latest hardware generation, and the decrease of the corepower in percent.

CPU type	CPU rel. year	HS06 sec/evt	corepower	HS06 sec/ evt recalculated	corepower re-scaled	decrease [%]
E5-2630 v4	2016	3293	10.0	3663	10.0	0
E5-2640 v3	2014	3177	11.4	4028	11.8	-3
E5-2630L 0	2012	6350	8.4	5933	4.3	48
E5-2660 0	2012	5505	9.6	5878	5.8	40
L5640	2010	6785	7.0	5283	3.4	51

Table 3. hs06 values for full sim production jobs in 2020, HC results, and the relative deviation.

Resources	job state				HC benchmarks			
	all		finished					
	hs06	CPU eff.	hs06	CPU eff.	tot. frac	hs06	hs06 w	rel dev [%]
all	**4664**	0.783	**4127**	0.787	80.3	3585	3906	5.4
Grid & Cloud	4215	0.946	**3987**	0.946	90.5	3618	3921	1.6
Grid	3462	0.955	**3244**	0.953	83.3	3595	3263	-0.6

2.40GHz CPU and it is installed on 11 different Grid queues, which permits a direct comparison. Figure 5 shows a histogram of the hs06 value of Grid jobs performed on this particular CPU type. This test shows an acceptable closure, the agreement is within 4%, comparing the mean value of the HC Grid jobs of 3127 HS06 sec/evt with the SPOT result of 3000 HS06 sec/evt (blue dashed line). To compare the numbers retrieved from the 2020 full sim production jobs and the

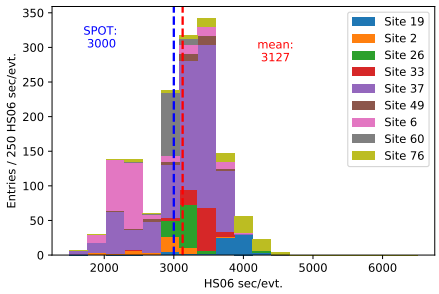

Figure 5. Comparison of the hs06 value of HC jobs performed on several queues with the SPOT results.

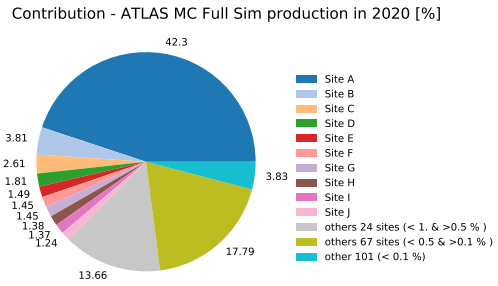

Figure 6. Contributions of all queues to 2020 full Sim production with job state finished.

numbers measured with the HC benchmark, one needs to understand the reference values in more detail. Resources with label *all* in Table 3 include jobs processed on queues at Grid sites, Cloud resources, and HPC centres. The number quoted before (4664 HS06 sec/evt) includes jobs executed on all types of resources with any final state (first column). Since failed jobs or jobs with other final states might bias the results, further comparison concentrates on finished jobs only. In order to compare the results of the HC test jobs running on the different queues with the number extracted from the ATLAS production system, it is necessary to consider the contribution of each queue to the MC production, and calculate a weighted average. The fractional contribution to the total ATLAS production of each queue can be retrieved from ATLAS' monitoring and accounting infrastructure. Figure 6 shows the contributions to the MC production in 2020 for all resources, with the job final state finished. Considering only the

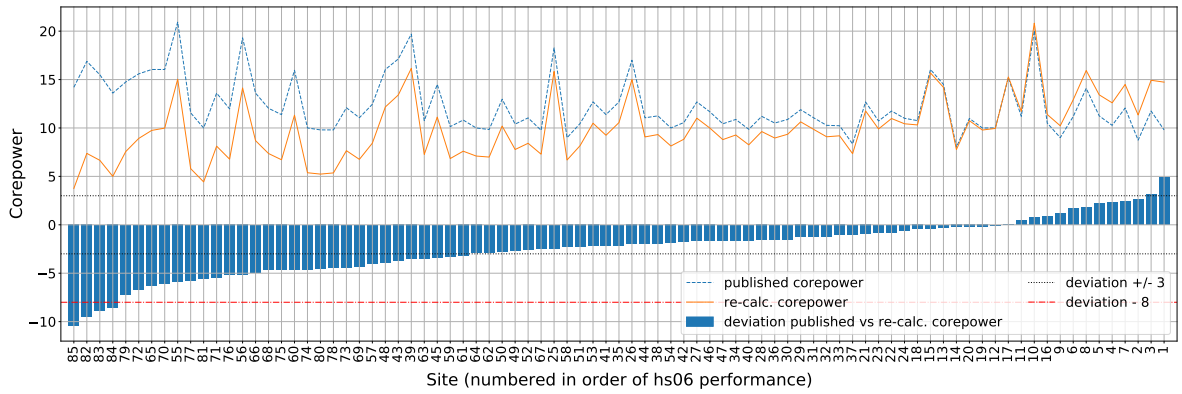

Figure 7. The blue bars show the performance deviation in corepower units, comparing the published corepower value per queue (blue line) with the re-calculated corepower value (orange line) with a reference value of 3 kHS06 sec/evt.

successful jobs and weighting the benchmark jobs properly according to the individual queue contributions to the ATLAS 2020 production, Grid (and Cloud) resources are in agreement within 2% comparing the ATLAS full sim production, see in Table 3. In conclusion, the ttbar sample is an appropriate benchmark process, after correcting for the inefficiencies due to the small number of events in the benchmark.

3.3. Detecting queue performance deviations

The HC measurements with the full sim ttbar process can be used in order to extract the deviation of the published corepower value with respect to an ideal corepower value. The ideal corepower value can be determined from Equation 1 solved for the corepower, inserting the wall-time, the number of cores, number of events from the benchmark jobs, and the SPOT hs06 value of 3.0 kHS06 sec/evt as ideal conditions. Figure 7 shows the published corepower value as a blue line, the "ideal" corepower value as an orange line, and the deviation from published to recalculated as blue bars. Queues with a negative deviation overestimate their compute performance and may need to downsize the corepower value, whereas queues with a positive deviation may be underestimating the queue performance and could increase their corepower values. A total of 94 queues have been tested: 58 queues (62%) show corepower deviations smaller then 3. Only 6 queues (6.4%) have deviations larger than 8, which may require further investigation.

4. Conclusion

Running identical test jobs with HC on the Grid allows a first direct comparison of the performance measurements from the SPOT team with the performance of the Grid queues. The closure test with the benchmark jobs weighted by the individual contributions of the queues to the MC full sim production in 2020 confirms that the ttbar sample is a reasonable choice. The selection of finished jobs performed on Grid (Grid and Cloud) resources reduces the hs06 value of 2020 full sim production from 4.7 kHS06 sec/evt to 4.0 kHS06 sec/evt (3.2 kHS06 sec/evt) which is in good agreement with the weighted benchmark results. The analysis of several Grid queues shows that hs06 is not a sufficient metric to compare old and new hardware with each other and one single corepower value cannot account for heterogeneous hardware within one queue. The HC benchmark test can be used to detect deviations from the quoted corepower value for a given queue, and the site administrators can be notified if necessary.

References

[1] The ATLAS Collaboration et al 2008 *JINST* **3** S08003
[2] Bird I 2011 Computing for the Large Hadron Collider *Annual Review Of Nuclear And Particle Science.* **61**, 99-118
[3] Henning J 2006 SPEC CPU2006 benchmark descriptions *SIGARCH Comput. Archit. News.* **34**
[4] Maeno T et al 2014 Evolution of the ATLAS PanDA Workload Management System for Exascale Computational Science *J. Phys.: Conf. Ser.* **513** 032062
[5] Elmsheuser J et al 2014 Grid site testing for ATLAS with HammerCloud *J. Phys.: Conf. Ser.* **513** 032030
[6] Wikipedia contributors 2021 Wikipedia List of Xeon microprocessors *Wikipedia The Free Encyclopedia* `https://simple.wikipedia.org/w/index.php?title=List_of_Xeon_microprocessors&oldid=7600890` [Online; accessed 24-June-2021]
[7] Wikipedia contributors 2021 List of AMD Opteron microprocessors *Wikipedia, The Free Encyclopedia* `https://en.wikipedia.org/w/index.php?title=List_of_AMD_Opteron_processors&oldid=1026572974` [Online; accessed 24-June-2021]
[8] Wikipedia contributors 2021 Epyc *Wikipedia The Free Encyclopedia* `https://en.wikipedia.org/w/index.php?title=Epyc&oldid=1025383410` [Online; accessed 24-June-2021]
[9] Giordano D, Alef M, Atzori L et al. 2021 HEPiX Benchmarking Solution for WLCG Computing Resources *Comput Softw Big Sci* **5**, 28