Investigation of the power consumption of the HEPScore23 benchmark as a function of workload and clock frequency on the ATLAS-BFG cluster

Submitted to the
Albert-Ludwigs-Universität Freiburg

by Lucas Ruppert

Supervised by Prof. Dr. Markus Schumacher

Abstract

This thesis studies the reduction of the CO_2-footprint of the ATLAS-BFG cluster operated by the Schumacher-group of the University of Freiburg. The ATLAS-BFG cluster installed new hardware in 2024. By using data measured over in this thesis concerning the performance with the HEPScore23-benchmark, the opportunity to reduce the CO_2-footprint arises. The new composition of the ATLAS-BFG consists of 12 compute nodes with AMD EPYC 7763 64-Core CPUs.The cluster can currently provide up to 42336HS23, whereas before the new hardware the cluster could provide up to 32500 HS23. This would not be able to full-fill the ATLAS-BFG clusters WLCG-pledge of 33300 HS23 for the year 2025.

Since the new composition of the cluster is able to provide 132% of the pledge of 2025, the maximum performance does not have to be provided constantly. By switching to a lower clock-frequency of 2000 MHz compared to the clock-frequency of 2700 MHz when the energy markets share of renewable energy drops below 72% the CO_2-footprint can be reduced by 3.7 t down to 21.2 t per year. This is due to the lower performance of 2568 MHz per compute node for a clock-frequency of 2000 MHz compared to the 3606 MHz per node for 2700 MHz when using 196 vCores.

Declaration

I hereby declare, that I am the sole author and composer of my thesis and that no other sources or learning aids, other than those listed, have been used. Furthermore, I declare that I have acknowledged the work of others by providing detailed references of said work. I also hereby declare that my thesis has not been prepared for another examination or assignment, either in its entirety or excerpts thereof.

Place, Date

Signature

Acknowledgements

I firstly would like to thank Prof. Markus Schumacher for the opportunity to write this thesis. I highly appreciate the time and feedback he provided me with.

Many thanks to my supervisor Dr. Michael Boehler for his great supervision, support and for teaching me a lot during the last three months.

Finally, I want to thank the group as a whole for the welcoming and supportive environment they created. I am thankful that I was able to write my thesis in this group.

Contents

1. Introduction

The CERN (European Organization for Nuclear Research) operates some of the biggest scientific experiments with the goal to expand knowledge about particle physics and "accelerat[e] science"[1]. To achieve this, an enormous amount of computing power is necessary. Because of the complexity and amount of the experiments computations, the needed compute performance has been increasing, which in turn increases the energy consumption and therefore the CO_2-footprint of HEP[1] compute centers. In times of the increasingly urgent topic of climate change, methods to reduce the HEP computations' CO_2-footprint are required. This thesis will explore a possible method to reduce the CO_2-footprint of HEP computations at the example of the WCLG[2]-embedded [2] ATLAS-BFG cluster, operated by the Schumacher-group of the University of Freiburg.

The WLCG is a worldwide computing grid designed to handle and store the data output of big HEP experiments. It is composed of over 170 compute sites in over 42 countries around the world. To keep track of the compute performance each site provides while using different hardware, a benchmark to measure the HEP related compute performance is necessary. This is met with the HEPScore23 benchmark, which is the current WLCG standard to measure the performance of a HTC cluster [3].

All countries taking part in the experiments are required to provide a certain average compute performance every year, measured by the HEPScore23-benchmark. Therefore, every HEP compute cluster has to yearly provide a pledged amount of compute power. The ATLAS-BFG clusters pledge for the year 2025 is 33300 HS23.

To minimize the CO_2-footprint of a HTC[3] cluster, knowledge of the compute nodes performance and power consumption with respect to the nodes configuration is required. The two variables examined in this thesis are the number of vCores used for the computations and the clock-frequency of the used CPUs. Every CPU has a set number of physical cores, which each can run two threads by using hyperthreading. Thus the number of vCores of a CPU is twice the number of physical cores. The clock-frequency governs how fast a task can be finished. Using the HEPScore23-benchmark to measure performance and measuring the nodes power consumption simultaneously for different configurations, the required knowledge of the nodes available at the ATLAS-BFG cluster is obtained.

The ATLAS-BFG cluster currently consists of 12 compute nodes using the AMD EPYC 7763 64-Core CPU spread over three server. However, this hardware was only acquired recently. Before that, the ATLAS-BFG cluster was composed of 84 nodes using the Intel Xeon E5-2630 v4 CPU spread over 11 server. Since the older hardware using the Intel CPUs is still available, this thesis will measure the aforementioned dependences for both the nodes with Intel and AMD CPUs. An important variable to compare different node configurations is the **energy efficiency**, which is the provided performance (HS23) divided by the power consumption (W).

With the acquired knowledge of the nodes with AMD CPUs behaviour for different configurations, a possibility to reduce the ATLAS-BFG clusters CO_2-footprint

[1] **H**igh **E**nergy **P**hysics
[2] **W**orldwide **L**HC **C**omputing **G**rid
[3] **H**igh **T**hroughput **C**omputing

is explored. Since the share of renewable energy at energy production fluctuates and the different configurations are expected to yield different dependences, a method of matching the node configuration of the cluster to the fraction of renewable energy produced at a certain time. By applying this method, the ATLAS-BFG cluster should use a more power-intensive configuration with a high performance only during times when renewable energy has a comparatively high market share. When still aiming to full-fill the ATLAS-BFG clusters pledge, the cluster needs to have a higher maximum performance than the pledge.

To calculate and compare different combinations of the vCore and clock-frequency configuration using this method, the necessary data will be measured first.

2. Experimental setup

This chapter will introduce the experimental setup: the arrangement of the components in a server, the HEPScore23-benchmark [3][4] and the measurement process as well as the data processing.

2.1 Physical server setup

The measurements on both the nodes with the Intel and AMD CPUs available at ATLAS-BFG will be made using the same physical setup, specifically the default operating setup of a server used within the ATLAS-BFG cluster.

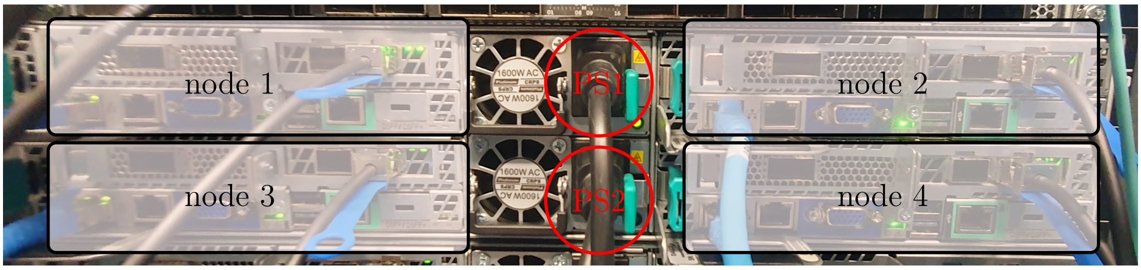


Figure 2.1: Setup of a 4-node server operating in the ATLAS-BFG cluster. Each server has two power supplies and four nodes two CPUs each, in this case the setup with the Intel CPUs is shown. The red circles represent the power supplies (PS).

As shown in fig. 2.1, one server contains four nodes with two CPUs each. While the nodes can be operated separately, they share the same **p**ower **s**upply (PS). In total, there are two PS for this setup, PS1 and PS2. Because the Intel and AMD CPUs differ in build the resulting specifications are different. They first of all differ in the number of physical cores: the Intel CPU has ten cores while the AMD CPU has 64 cores. This creates a significantly higher number of cores per node when using the AMD CPU (table 2.1).

Table 2.1: Specifications of the ATLAS-BFG nodes with Intel or AMD CPUs.

socket type	physical cores	vCores	RAM [GB]	disk storage [GB]
Intel	20	40	128	960
AMD	128	256	512	3500

In table 2.1 the number of vCores per CPU (and also per node) are twice the number of physical cores. This is due to hyper-threading; when operating a CPU in hyper-threading-mode, each physical core can handle two separate processes, or threads, simultaneously.

2.2 HEPScore23

The HEPScore23-benchmark was introduced in 2023 to adapt the metric in charge of measuring the provided compute performance (HEPSpec06 [5]) of a site to the

needs of a HEP computing job [3][4]. The described required system components for the HEPScore23-benchmark to run are 20GB of free disk space and at least 320MB per vCore of free space in the output directory [4].
The HEPScore23-benchmark consists of seven individual benchmarks provided by the four large CERN-experiments ATLAS [6][7], CMS [8][9], LHCb [10][11] and ALICE [12][13] and the in Japan located BELLE-II experiment [14][15] with similar computational needs. The ATLAS and CMS collaborations each provide two benchmarks, one benchmark for the **gen**eration of events and one benchmark for the **reco**nstruction of events. While the other collaborations each provide one, their benchmark combines the functionality of both benchmark-types provided separately by the ATLAS collaboration and the CMS collaboration. Over the course of measuring the HEPScore23-value every benchmark is run three times consecutively. In general, a benchmark of this sort, e.g. the BELLE-II benchmark, will use some HEP workloads [16] and clone those workloads, but not all of those benchmark use single-threading.

Table 2.2: Number of threads used by each benchmark in the HEPScore23-Benchmark [17][16].

benchmark	atlas-gen	atlas-reco	cms-gen	cms-reco	lhcb	belle2	alice
nr. of threads	1	4	4	4	1	1	1

As displayed by table 2.2 three of the benchmarks use multithreading, specifically they utilize four threads. Thus the "default number of copies is the number of [vCores] divided by four" [16]. The benchmark provided by the ALICE collaboration works differently. Because of the high intensity of their usual computing tasks the collaboration parallelised their computing tasks. This also downcast in their provided benchmark. As a consequence, the ALICE benchmark does not follow the specific vCore restrictions the other benchmarks have. One thread generated by the benchmark can be computed by multiple vCores, contrary to the other benchmarks.

As documented in the dedicated GitLab [4] the HEPScore23-benchmark is highly configurable. The CERN also provides a hep-benchmark-suite [18][17], which has a fully functional configuration file as well the option to configure other parameters according to ones needs.

2.2.1 The hep-benchmark-suite

The hep-benchmark-suite is a highly flexible tool to configure and run a number of different HEP related benchmarks, one of which is the HEPScore23-benchmark. Because of its convenience and good documentation it will be used to configure and run the necessary benchmarks for this thesis.

Figure 2.2 displays the functionality of the hep-benchmark-suite. The "Run Logic"-part allows a user to run a HEPScore23-Benchmark with a generated configuration file. Following the documentation by using a provided python environment which has a terminal-command for starting a run, measuring the HEPScore23 with

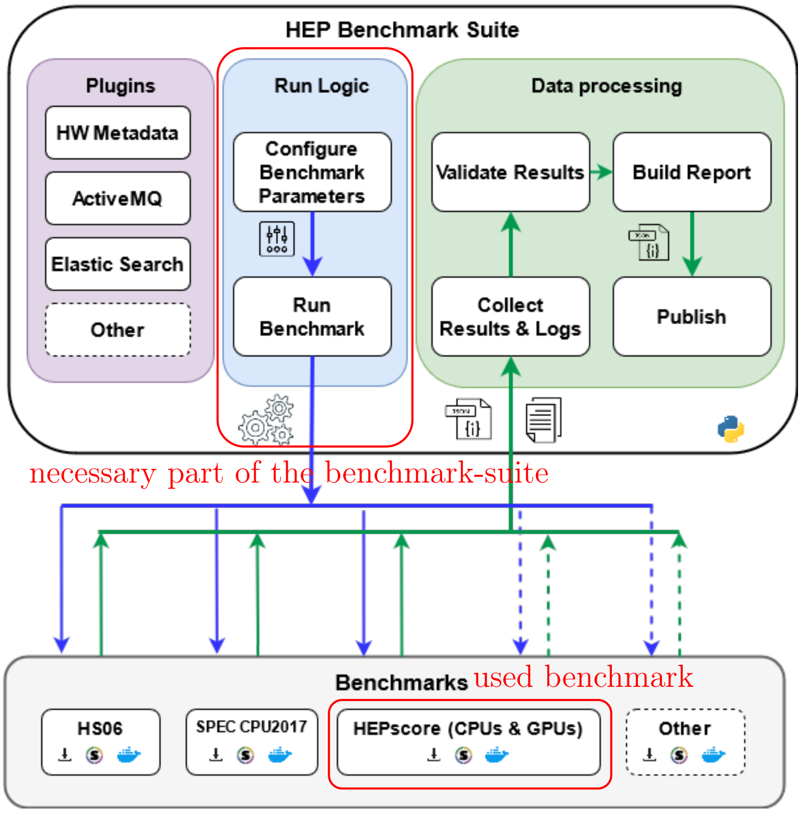


Figure 2.2: Functionality representation of the hep-benchmark-suite. The benchmark HEPScore23 (highlighted in red) will be used with a generated configuration file [18].

a provided configuration or a default configuration. For the purpose of varying the number of vCores one needs to provide a custom configuration because the default configuration is not enabling the user to change the number of vCores used by the benchmark; by default all available vCores are being used. To create a custom configuration for a specific number of vCores, the easiest is to use a provided shell-script from the "examples"-section of the hep-benchmark-suite.

After configuring the benchmark like this the next part is to run a desired benchmark, in this case HEPScore23 as displayed in fig. 2.2.

2.3 Data handling

This section discusses the necessities for measuring the power consumption, load and clock-frequency during a run of the HEPScore23-Benchmark, how the measurement procedure works and gives an overview about how this data is generally handled.

2.3.1 Measurement structure

In section 2.2.1 it is described how a variation in the number of vCores is achieved, but it is clear that the hep-benchmark-suite can not handle the systems configuration to a desired clock-frequency. It is however vital to implement this possibility and it can be achieved with the Linux build-in command "cpupower frequency-set". By adding the options "-f" (default frequency), "-u" (upper frequency limit) or "-d" (lower frequency limit) it is possible to specify an option to configure the clock-frequency settings. By setting all three values to the same it is ensured that the clock-frequency stays at the specified value even under high loads; it sets the scaling governor automatically to "userspace"[1]. However, some kernel restrict the possible values the clock-frequency can be configured to. The clock-frequency of a specific vCore at any time can then be obtained by reading a specific file in the /sys directory [19].

As mentioned in section 2.2.1, the power supply (PS) is not measured by the hep-benchmark-suite, but this can be done by using the open source utility "ipmitool"[20], which requires IMPI-support from the hardware. The command "ipmitool sensor" returns detailed information about the systems power supply. However, the form of this output varies from one system to another. Consequently, the output has to be prepared differently for the two different systems to acquire the desired PS-values (in this case PS1 and PS2). Additionally, one has to consider that this utility returns the power consumption, not the power output of the outlet or the power consumed for cooling the system. It is thus an estimated value of power consumption by the mainboard.

By using the Linux build-in command "uptime" one can obtain the current time and the systems median load over the last minute [21].

In total one has two objectives to accurately acquire all necessary data from one benchmark-run: the benchmark itself has to be started and some mechanism has to take all desired values and store them in a (csv-)file. Since after starting the benchmark the mechanism responsible for it is occupied for as long as the benchmark is running, a separate mechanism has to handle the measurements. Consequently two bash-scripts are necessary. However, they still need to run (and preferably start) simultaneously. Furthermore, because of the servers setup shown in fig. 2.3 regarding the power supply the benchmarks and measurements have to be started simultaneously on all four nodes.

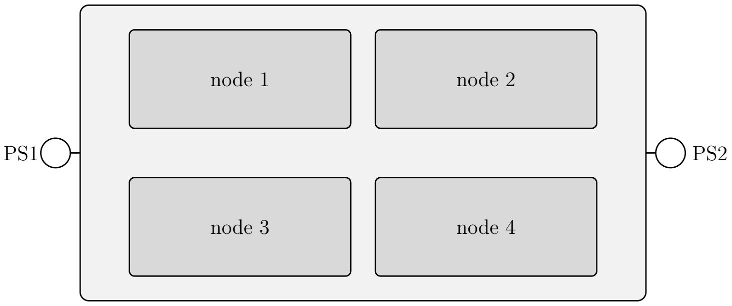


Figure 2.3: Setup of a server operating in the ATLAS-BFG cluster, schematic of fig. 2.1. The power supply (PS1 + PS2) is the same for the four independent nodes, because it is connected to the server. Other values like the load and HEPScore23-values are individual to the nodes.

This is achieved by using the cron-daemon[22] and creating a crontab with the same starting time for starting the benchmark and the measurement on each node.

The script for starting the benchmark follows a simple pattern by just combining all necessary steps described in section 2.2.1 into a bash-script: after starting the

[1] The scaling governor governs the behaviour of the clock-frequency settings [19].

python environment and adjusting the clocking frequency the benchmark is started by referencing a previously created configuration file. When the benchmark is finished the script saves all relevant log-files and stores a stop-signal. The measurement script starts to write the as previously described obtained values into a csv-file every 30 seconds, only stopping upon reading the stored stop-signal by the benchmark script. Therefore a table like table 2.3 and various log-files, e.g. containing the achieved HEPScore23, are received from this measurement process.

Table 2.3: Examplary headers of the raw data of one run on one node (for two Intel CPU-sockets with a total of 40 vCores per node)

time	load	PS1	PS2	freq0	freq1	freq2	...	freq39

2.3.2 Preparing the data

The data analysis and the preparation of the measured data will be handled with the python library pandas [23]. As shown by table 2.3, the csv-file has many columns which need to be combined for the analysis, e.g. the individual power supply of PS1 and PS2 is not of interest, rather their sum PS is. There are also values missing which apply for the benchmark run as a whole, e.g. the number of vCores used and the obtained HEPScore23-value. Therefore a few additions and operations take place producing a pd.DataFrame in the form of table 2.4.

The changes made contain firstly the sum of PS1 and PS2 to PS. Since the clock-frequency is a system configuration it should be fully described by one value per measurement, similar to the number of vCores. Thus it will be taken the mean over all measured frequency-values of all vCores to achieve this. Like the clocking-frequency, the achieved HEPScore23 and the used number of vCores for this value are also important values related to the measurement of a benchmark tun as a whole and shall thus also be added to the pd.DataFrame as column containing a single value.

Table 2.4: Exemplary headers of the prepared data. "freq" contains the mean value of all frequency measurements, "PS" contains the sum of PS1 and PS2, "HS23" contains the produced HEPScore23 and "vcores" contains the number of used vCores. "time" and "load" remain untouched with respect to table 2.3.

time	load	PS	freq	HS23	vcores

The column "freq" in table 2.4 contains the mean of all measurements of the clock-frequency on all vCores as shown in table 2.3. Random samples of a HEPScore23-benchmark runs clock-frequency distribution confirm that the node complies with the configured clock-frequency.

3. Analysis of nodes with Intel CPUs at BFG

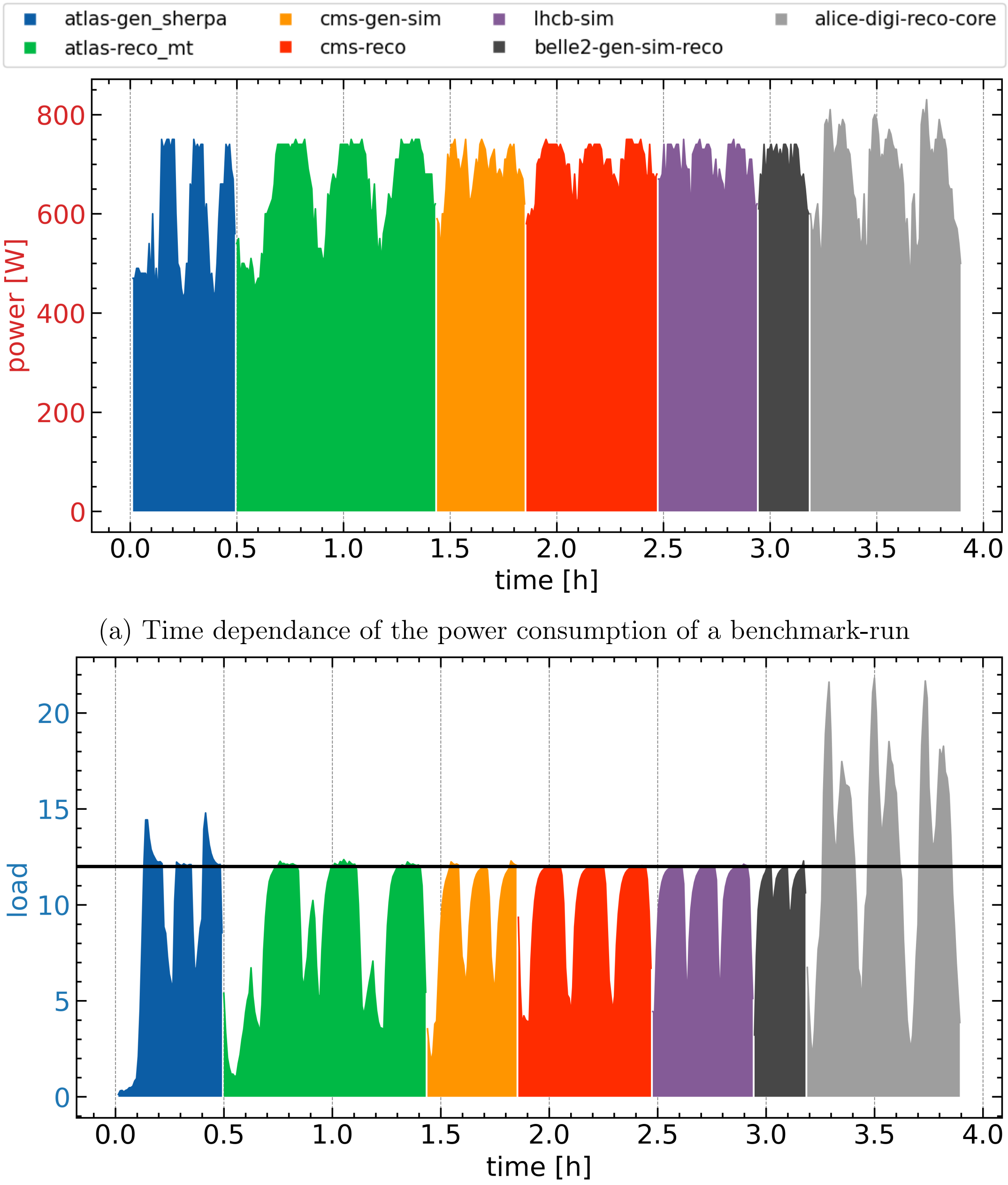
This chapter will analyse the dependence of the HEPScore23-value and the power consumption on the number of vCores and clock-frequency used by a node using the Intel CPUs available at BFG. After discussing the progression of a HEPScore23-benchmark-run of this setup, a metric for power consumption will be developed for both the number of vCores and the clock-frequency variation. After that, the dependence of the benchmarks runtimes on both parameters will be explained.

3.1 Time dependance of a run

The first step is to verify that the benchmark-run follows the configurations. The clock-frequency is validated by taking the mean over all measured values. For the number of used vCores, one examines the behaviour of the load during the run. In general, the load gives information about the CPU utilization where a load of one means one process or thread is being handled (by one vCore) [24]. Which means that in principle the load should equal the number of vCores the benchmark-run is able to use.

The plots in fig. 3.1 are a sample meant to visualize the time dependence of the power consumption and the load during a benchmark-run on one node. Figure 3.1a shows that during times of high load, i.e. when the individual benchmarks are running, the power stays relatively consistent. The load in fig. 3.1b stays at the expected value of 12 during the first six benchmarks but overshoots it during the ALICE-benchmark. This is due to the discussed (section 2.2) difference of the ALICE-benchmark with respect to the other HEP-workloads. Because of the parallelization of the processes the benchmark is not able to stick to the configured number of vCores. While this circumstance is unfortunate it is not easy to fix since it results from the benchmark's design.

However, since the load stays at the expected value for the most part it can be assumed that the benchmark-run sticks to the vCore-configuration.



(a) Time dependance of the power consumption of a benchmark-run

(b) Time dependance of the load of a benchmark-run

Figure 3.1: Time dependence of a benchmark-run on an ATLAS-BFG node with Intel CPUs with a configured clock-frequency of 2400 MHz and vCore configuration of 12. The individual HEP-workloads are colour-coded. (a) shows the power consumption of the node with respect to the time and (b) shows the CPU load of the node with respect to the time. The black line in (b) visualizes a load of 12.

3.2 Analysis of vCore dependence

In this section the dependence of a HEPScore23-benchmark-run and its power consumption on the number of vCores is investigated.

3.2.1 Definition of a power metric

The obtained values of the HEPScore23-benchmark, number of used vCores and clock-frequency differ from the values describing the power consumption. While the other variables describe a property of a benchmark-run on a node, the power consumption is a time dependent function for four nodes describing the power consumption of four nodes. It is thus necessary to find a adequate metric to capture the power consumption of a benchmark-run in one value per node. The simplest choice would be to take the mean of the measured PS-values of a benchmark-run and divide it by four (because there are four separate nodes measuring the same value), there are other choices. When analysing the behaviour of the load and power supply compared to the time (section 3.1) it is noticed that during times when the system operates under high load, the power consumption is also higher. Consequently, the times with lower load and power consumption are not representative of the mean power consumption of the HEP-workloads. The quantile of a certain percentage higher than 0.5 [1] would exclude the unrepresentative part of the PS-measurements. Therefore by taking the mean over the quantiles 0.7 - 0.9, which is called a truncated mean, a somewhat consistent and representative value for all vCore configurations can be obtained [25].

When comparing the histograms in fig. 3.2 for the different vCore configurations, its noticed that they differ in their respective highest value. To be more precise, with a higher number of vCores the distribution relies more on the top-values. This is a first indicator that with a higher number of vCores or processes the power consumption rises. The truncated mean also seems a better metric for those configuration. Nonetheless, even for the fewest number of configurable vCores (4) the truncated mean seems to be a better metric. Therefore it will be used to compare different vCore configurations of the benchmark-runs on the Intel CPUs.

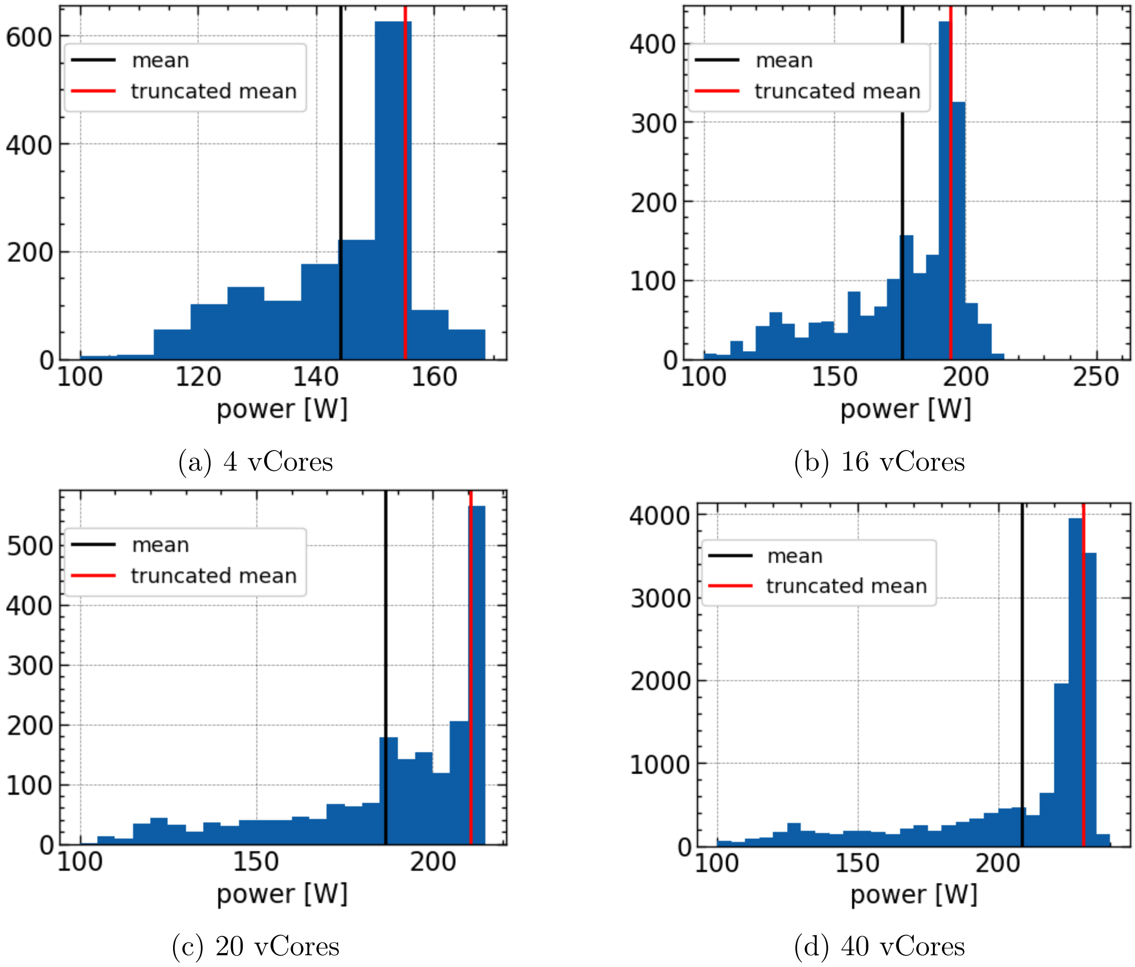
3.2.2 Analysis

With the defined metric which characterizing the power consumption of a benchmark-run and the measured HEPScore23-value characterizing the performance of the benchmark-run the necessary values have been obtained. The different measurements for a specific configuration of clocking-frequency and number of vCores are combined into one value by computing the mean. The error is then given by the standard deviation.

The mean of HEPScore23- and power consumption-values in dependence on the number of configured vCores are shown in fig. 3.3.

Figure 3.3a describes the dependence for the whole node, fig. 3.3b breaks the dependence down to the vCore. The first half of the HEPScore23-graph displays an approximately linear dependence. While the performance for more than 20 vCores also displays a linear dependence, it has a smaller slope. A similar but not as distinct property can be observed for the slope of the power consumption. This saturation effect is generated by the limited number of physical cores. When adding a vCore to the benchmark configuration when below 20 vCores, a whole new physical core is available for the extra process. But after reaching 20 vCores hyper-threading is used. Now two processes need to share the physical resources of a core, thus creating two

[1] A quantile of 0.5 is equivalent to/called the median.



(a) 4 vCores

(b) 16 vCores

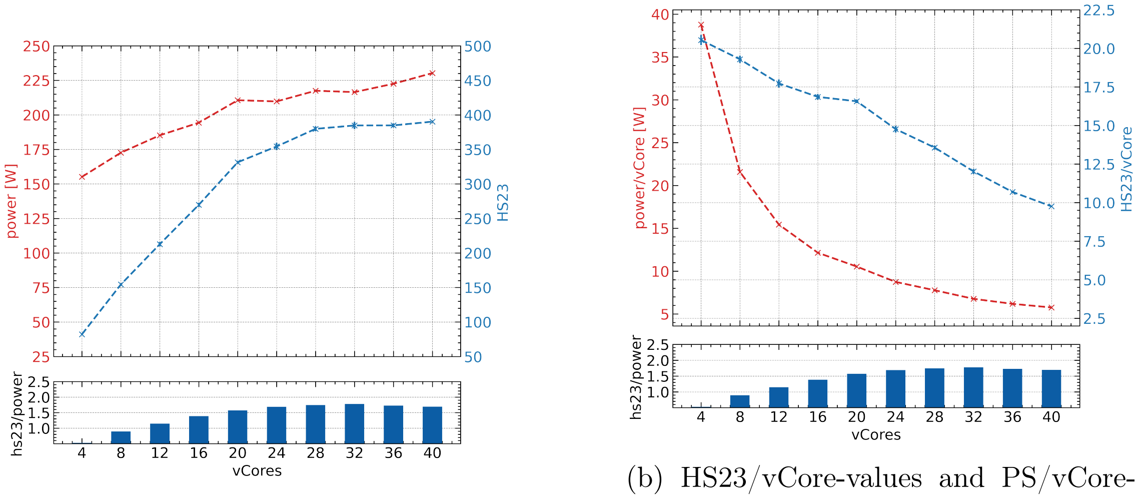
(c) 20 vCores

(d) 40 vCores

Figure 3.2: Histograms of all measured PS-values for a clock-frequency of 2400 MHz and (a) 4 vCores, (b) 16 vCores, (c) 20 vCores and (d) 40 vCores of an ATLAS-BFG node using the Intel CPUs. The black line visualizes the mean while the red line visualizes the truncated mean, specifically the mean of the quantiles 0.7 to 0.9.

vCores on one physical core. This may seem inefficient at first, but when considering the power consumptions curve in fig. 3.3b it becomes clear that the power needed per vCore still becomes significantly smaller by using hyper-threading. It is also important to mention that the additional performance still comes from the same setup by using hyper-threading. However, in case of the HEPScore23-benchmark on the Intel-CPUs this holds only to a certain degree. The bar charts in fig. 3.3 represent the generated HEPScore23 per Watt, thus describing the efficiency of a certain vCore configuration. The bar chart has its peak at 32 vCores, resulting in a efficiency drop when using more vCores. However, the possibility of the total HEPScore23 provided by a node being significantly higher needs to be considered if the goal is to provide as high of a HEPScore23-value as possible. But since the node does not provide a significantly higher total HEPScore23-value in contrast to the power consumption, the optimal working point for most use cases seems to be at 32 vCores.

The data points have error-bars resulting from the standard deviation when calculating them. However, they are hardly visible due to the little deviation of the HEPScore23-values and power consumption of different benchmark-runs from one



(a) HS23- and PS-values with respect to the number of used vCores.

(b) HS23/vCore-values and PS/vCore-values with respect to the number of used vCores.

Figure 3.3: Power consumption and HEPScore23-values in dependence of the number of used vCores of an ATLAS-BFG node with Intel CPUs. For all measurements a clock-frequency of 2400 MHz is configured. (a) shows the vCore-dependence for a node and (b) shows the dependence per vCore.

another.

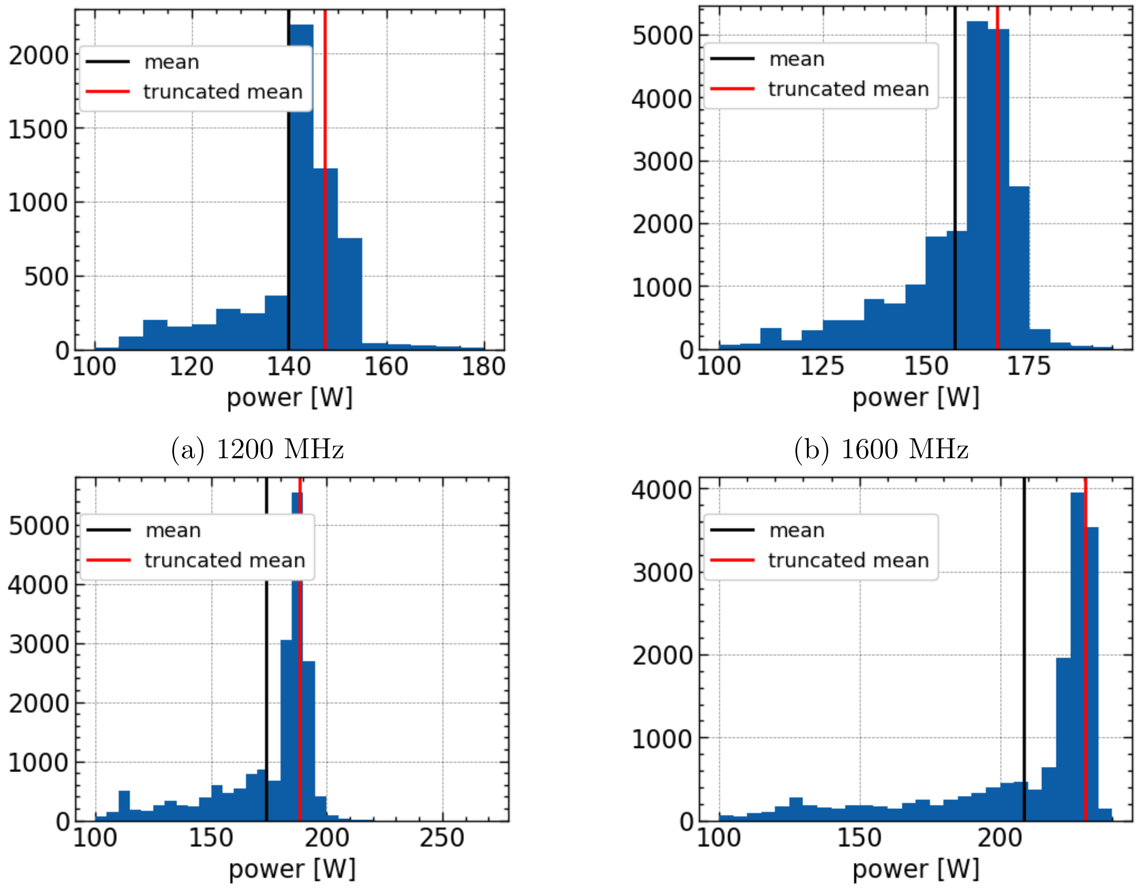
3.3 Analysis of clock-frequency dependence

In this section the dependence of the HEPScore23-benchmark and its power consumption on the clock-frequency is investigated.

3.3.1 Definition of a power metric

As previously conducted and argued for the vCore-analysis in section 3.2.1, a power-metric needs to be defined to compare the power consumption of different clock-frequency configurations. Following the process in section 3.2.1, the black lines in fig. 3.4 represent the mean value of all measurements of the nodes power consumption, while the red lines represent the truncated mean[2]. For a clocking-frequency of 1200 MHz, the mean seems to describe the distribution more accurately than the truncated mean, but for higher clock frequencies the opposite holds. Since there are more measurements with higher clock-frequencies than 1200 MHz (since it is the lowest possible configuration with the Intel CPUs) and those measurements are more likely to yield a HEPScore23-value of interest to the operation of the ATLAS-BFG cluster, the truncated mean will be used as a metric to compare the different clock-frequency configurations. This also provides the opportunity to compare to measurements conducted for the vCore-variation since the same metric is used there.

[2] specifically the mean of the quantiles from 0.7 to 0.9.



(a) 1200 MHz

(b) 1600 MHz

(c) 2000 MHz

(d) 2400 MHz

Figure 3.4: Histograms of all measured PS-values using 40 vCores and a clock-frequency of (a) 1200 MHz, (b) 1600 MHz, (c) 2000 MHz and (d) 2400 MHz using an ATLAS-BFG node with Intel CPUs. The black line visualizes the mean while the red line visualizes the truncated mean, specifically the mean of the quantiles 0.7 to 0.9.

3.3.2 Analysis

With the power metric defined the values of power consumption and performance measured can be handled the same as in section 3.2.2: the mean of the power consumption and HEPScore23-values in dependence of the clock-frequency are depicted in fig. 3.5. The error-bars are again calculated by the standard deviation. Similar to the variation of used vCores, the power consumption and HEPScore23-values do not deviate much for the same configuration, which is why he error-bars are hardly visible in fig. 3.5. In contrast to the vCore dependence of the HEPScore23-benchmark, the dependence on the clock-frequency seems to be entirely linear. However, the dependence of the power consumption on the clock-frequency is only partially linear. Up to a clock-frequency of 2000 MHz, the power consumption is linear with respect to the clock-frequency. For clock-frequencies higher than 2000 MHz, the behavior of the power consumption is not predictable. Considering the bar chart displaying the ratio of provided HEPScore23 and power consumption, and thus the efficiency of the respective clocking-frequency, the efficiency of the node rises until 2000 MHz. While a clock-frequency of 2200 MHz is less energy efficient than a fre-

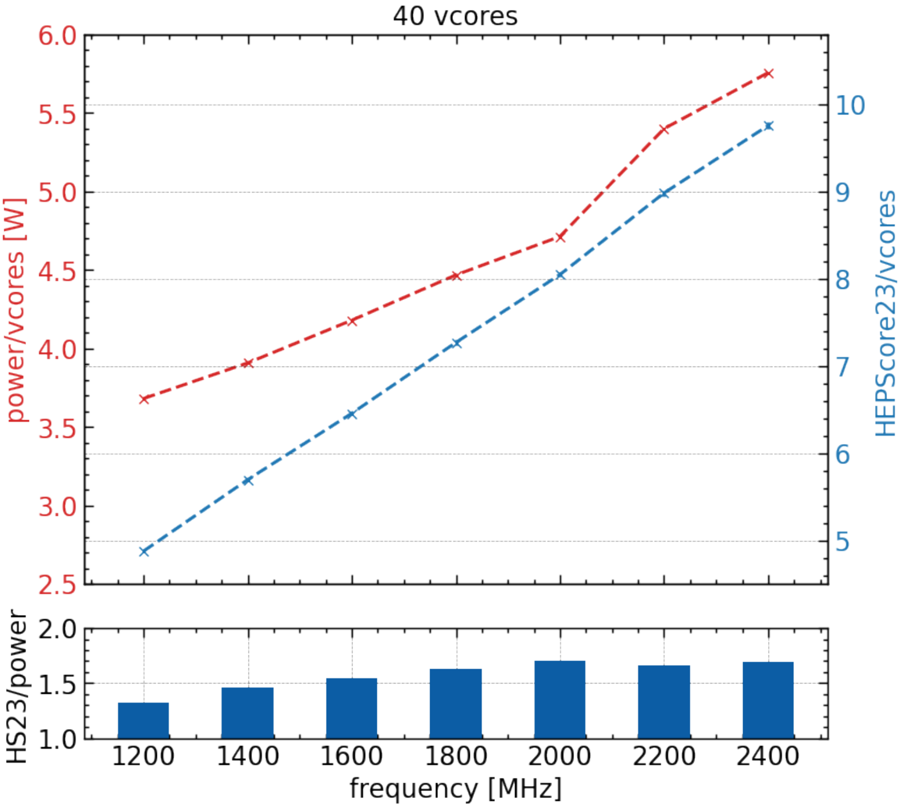


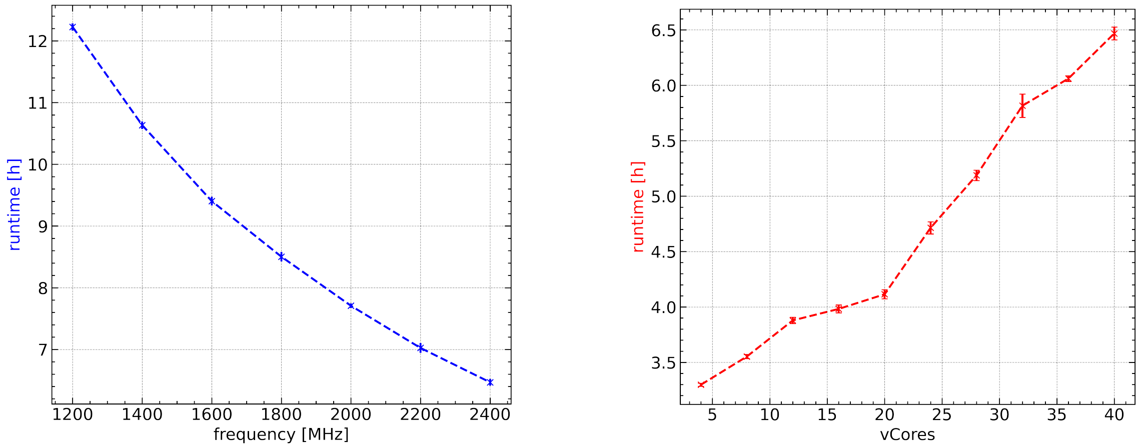
Figure 3.5: HEPScore23-values and power consumption per vCore with respect to the configured clock-frequency of an ATLAS-BFG node with Intel CPUs. 40 vCores were utilized for all measurements. The bar chart on the bottom visualizes the provided performance per Watt (per vCore/node/server), therefore displaying the efficiency of the given frequency with 40 vCores.

quency of 2000 MHz, a clock-frequency of 2400 MHz is approximately as efficient as a frequency of 2000 MHz. As a consequence, two possible operational configurations need to be considered; one at 2000 MHz and the other one at 2400 MHz. Since their efficiency is similar, the two configurations only differ in their total performance (provided HEPScore23-value) and power consumption. Therefore the choice of the operational configuration is purely dependant on how much performance is needed.

3.4 Analysis of benchmark-runtimes

When investigating the power consumption and performance of a benchmark-run in fig. 3.1, it is noticed that the individual benchmarks also have individual runtimes. Furthermore, after analyzing the differences in performance of a node in dependence of vCore and clock-frequency variation the question of their impact on the benchmarks runtimes arises. Answering this question is also relevant for effectively scheduling the benchmarks and their measurements consecutively using the cron daemon (section 2.3.1).

Figure 3.6a shows a decline of the runtime with respect to a higher clock-frequency. This anti-correlation of the benchmark-runtimes is expected, since a higher performance indicates that more processes can be completed faster. The



(a) Runtimes of the HEPScore23-benchmark in dependence of the clock-frequency using 40 vCores.

(b) Runtimes of the HEPScore23-benchmark in dependence of the used vCores using a clocking-frequency of 2400 MHz.

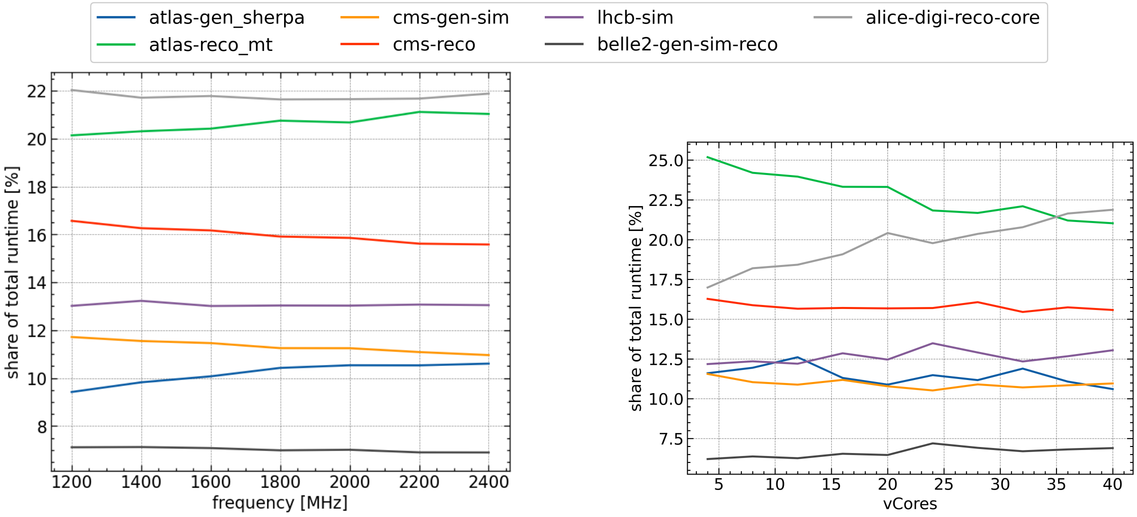
Figure 3.6: Runtimes of the HEPScore23-benchmark running on ATLAS-BFG nodes with Intel CPUs in dependence of (a) the clock-frequency and (b) the number of vCores.

slope however shows a decline with higher clock-frequency, which indicates that the HEPScore23-benchmark runtimes are not solely dependent on the performance (which is linearly dependent on the clock-frequency).

On the other hand, the benchmark-runtimes appear to be positively correlated to the number of used vCores. This seems to be increasing for more than 20 used vCores, displaying two approximately linear dependencies, where the one using only physical cores has a lower slope. Therefore a relation between the benchmark runtime and the number of physical cores used for only one thread each is indicated, similar to the performance (section 3.2.2). However, most HEP-workloads only clone the produced thread(s) by the number of vCores which should be used. Therefore one would expect a dependence on the clock-frequency for their runtimes, but not on the number of vCores, since for those HEP-workloads there is always the same number of threads per used vCore. For example, a benchmark-run with 16 vCores configured will run 16 threads on 16 vCores and a benchmark-run with 40 vCores will 40 vCores use for 40 threads.

The ALICE-benchmark could be responsible for the difference in runtimes, since it uses more resources than configured when running the HEPScore23-benchmark on fewer vCores than the nodes maximum. This hypothesis can be verified by checking the shares of the different HEP-workloads runtimes of the HEPScore23-benchmarks total runtime for vCore and frequency variation, which is visualized in fig. 3.7.

Under variation of the clock-frequency the share of the HEP-workloads' runtimes do not change much, as shown by fig. 3.7a. The same applies to the vCore-variation except for the ALICE-benchmark and the ATLAS-reco-benchmark. While the ATLAS-reco-benchmark tends to decrease when increasing the number of used vCores, the ALICE-benchmarks runtimes share of the total runtime increases. When varying the configured number of vCores, the runtime of the ALICE-benchmark displays a faster increase up to 20 vCores and a slower increase thereafter. This is the



(a) Clock-frequency variation using 40 vCores.

(b) vCore-variation using 2400 MHz.

Figure 3.7: Shares of the runtimes of the individual HEP-workloads of the HEPScore23-benchmark in dependence of the (a) clock-frequency and (b) number of vCores.

expected behaviour from the previously stated hypothesis that the increase of total runtime is primarily due to the ALICE-benchmarks parallelization technology while most of the other HEP-workloads runtime-shares stay comparatively the same.

Since this is a inherent property of the ALICE-benchmark it can not be eliminated by configuring the HEPScore23-benchmark via the hep-benchmark-suite. However, this does not affect the usage the HEPScore23-benchmark as a metric since it is a consistent property and the benchmark is used by the WLCG.

4. Analysis of nodes with AMD CPUs at BFG

This chapter will analyse the dependence of the HEPScore23-value and the power consumption on the number of vCores and clock-frequency used by a node with the AMD CPUs available at the ATLAS-BFG cluster. Similar to the nodes with Intel CPUs, the behaviour of a HEPScore23-benchmark-run with respect to time will be discussed. A metric for power consumption will be specified for both the number of vCores and the clock-frequency analysis.

4.1 Time dependence of a HEPScore23-benchmark run

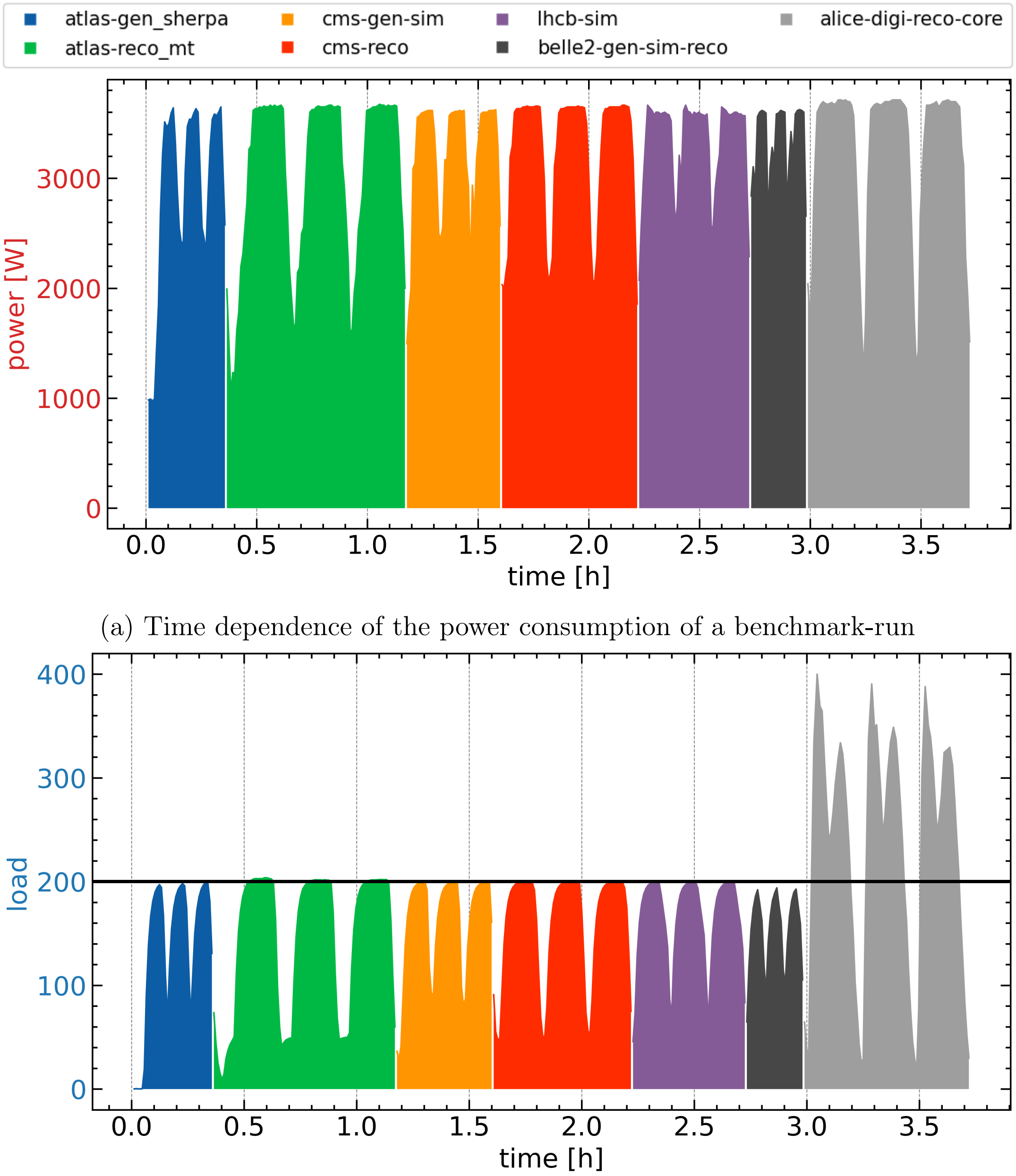
The time-dependence of a HEPScore23-benchmark-run on a node with the newer AMD CPUs available in the ATLAS-BFG cluster shows many similarities with a benchmark-run on the nodes with the Intel CPUs (section 3.1), as expected.

The HEP-workloads, except the one provided by the ALICE-collaboration, follow the configured number of vCores, since the load in fig. 4.1b matches the configured number of vCores. As before with the Intel CPUs, the ALICE benchmark uses more resources than configured. Therefore the same hypothesis regarding the dependence of the benchmark-runtimes with respect to a variation in the number of used vCores is formulated: because the ALICE-benchmark uses more resources if available, its runtime increases if fewer extra resources are available, i.e. a higher number of vCores is configured. This hypothesis is verified by fig. A.1b, whose graph follows the same trend as the one in fig. 3.6b and will therefore not be discussed further.

A difference between the setup of the Intel CPU with respect to the AMD CPUs is the shorter total runtime of the HEPScore23-benchmark, which is evident by the comparison of fig. 3.6 and fig. A.1. This is due to increased performance shown by the newer AMD CPUs.

4.2 Analysis of vCore dependence

This section focuses on the impact different vCore configurations have on the performance and power consumption of a node from the BFG cluster using the available AMD CPUs. While the measurement process as whole and most of the configurations apart from the CPUs are similar to the measurements of the setup with the Intel CPUs, one key difference needs to be mentioned: the scaling governor for all measurements of the setup with Intel CPUs was set to "userspace", thus ensuring that the system follows the set clock-frequency. But for the measurements of the vCore variation with the setup using the AMD CPUs, the scaling governor "performance" was used, which results in a higher clock-frequency for a higher vCore-configuration (table B.3).



(a) Time dependence of the power consumption of a benchmark-run

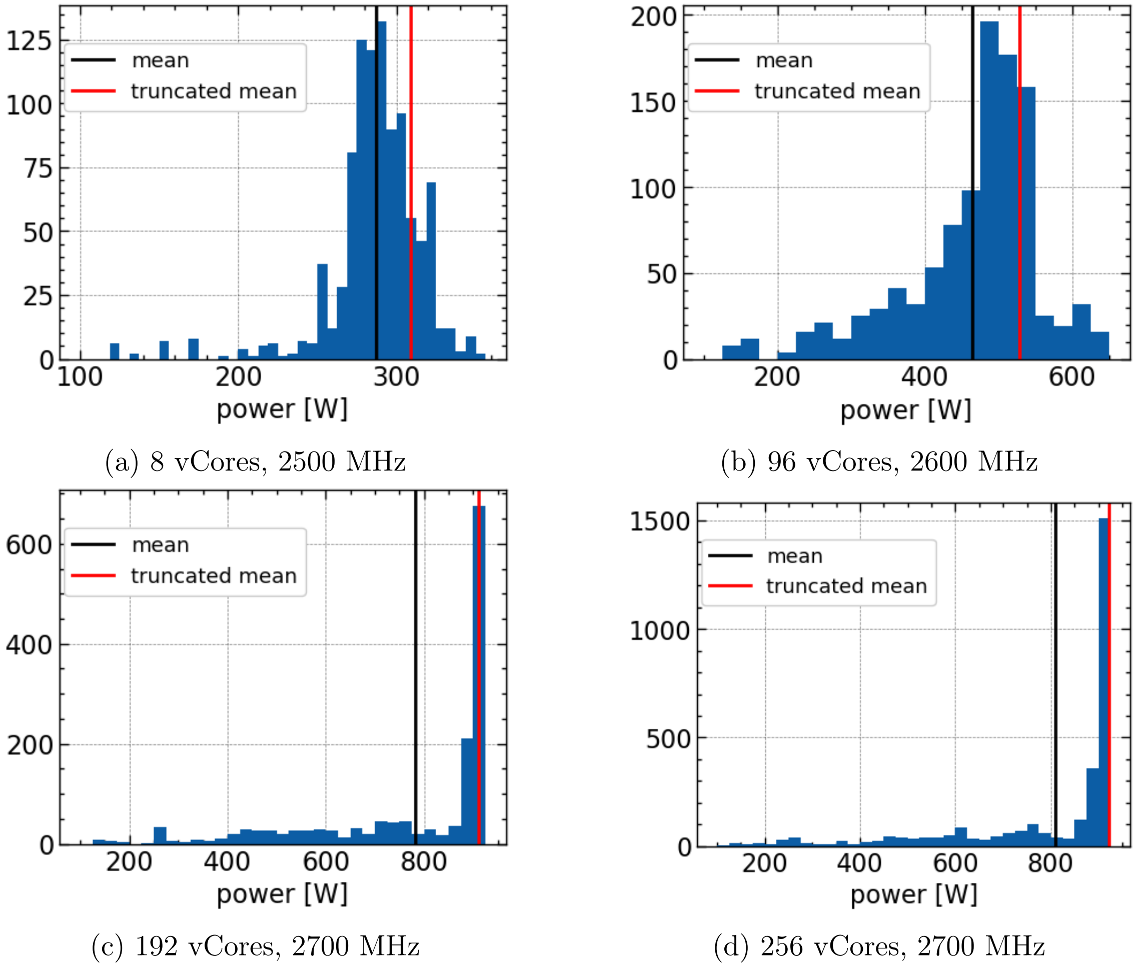
(b) Time dependence of the load of a benchmark-run

Figure 4.1: Time dependence of a HEPScore23-benchmark-run on an ATLAS-BFG node with AMD CPUs of (a) the power consumption and (b) the load. A clock-frequency of 2700 MHz was configured and 200 vCore were used. The individual HEP-workloads are colour-coded. The black line in (b) visualizes a load of 12.

4.2.1 Definition of a power metric

As previously discussed in section 3.2.1 a consistent metric to compare the power consumption of different vCore configurations must be defined.

The histograms in fig. 4.2 compare the quality of the previously presented possible metrics (mean and truncated mean) for power consumption for different vCore configurations spread over the entire set of configurable values. While for a configuration with 8 vCores (second lowest configuration compatible with the HEPScore23-benchmark) the mean seems to be describe the distribution of power supply mea-



(a) 8 vCores, 2500 MHz

(b) 96 vCores, 2600 MHz

(c) 192 vCores, 2700 MHz

(d) 256 vCores, 2700 MHz

Figure 4.2: Histograms of all measured PS-values using (a) 8 vCores and 2500 MHz, (b) 96 vCores and 2600 MHz, (c) 192 vCores and 2700 MHz and (d) 256 vCores and 2700 MHz using an ATLAS-BFG node with Intel CPUs. The scaling governor was set to "performance". The black line visualizes the mean while the red line visualizes the truncated mean, specifically the mean of the quantiles 0.7 to 0.9.

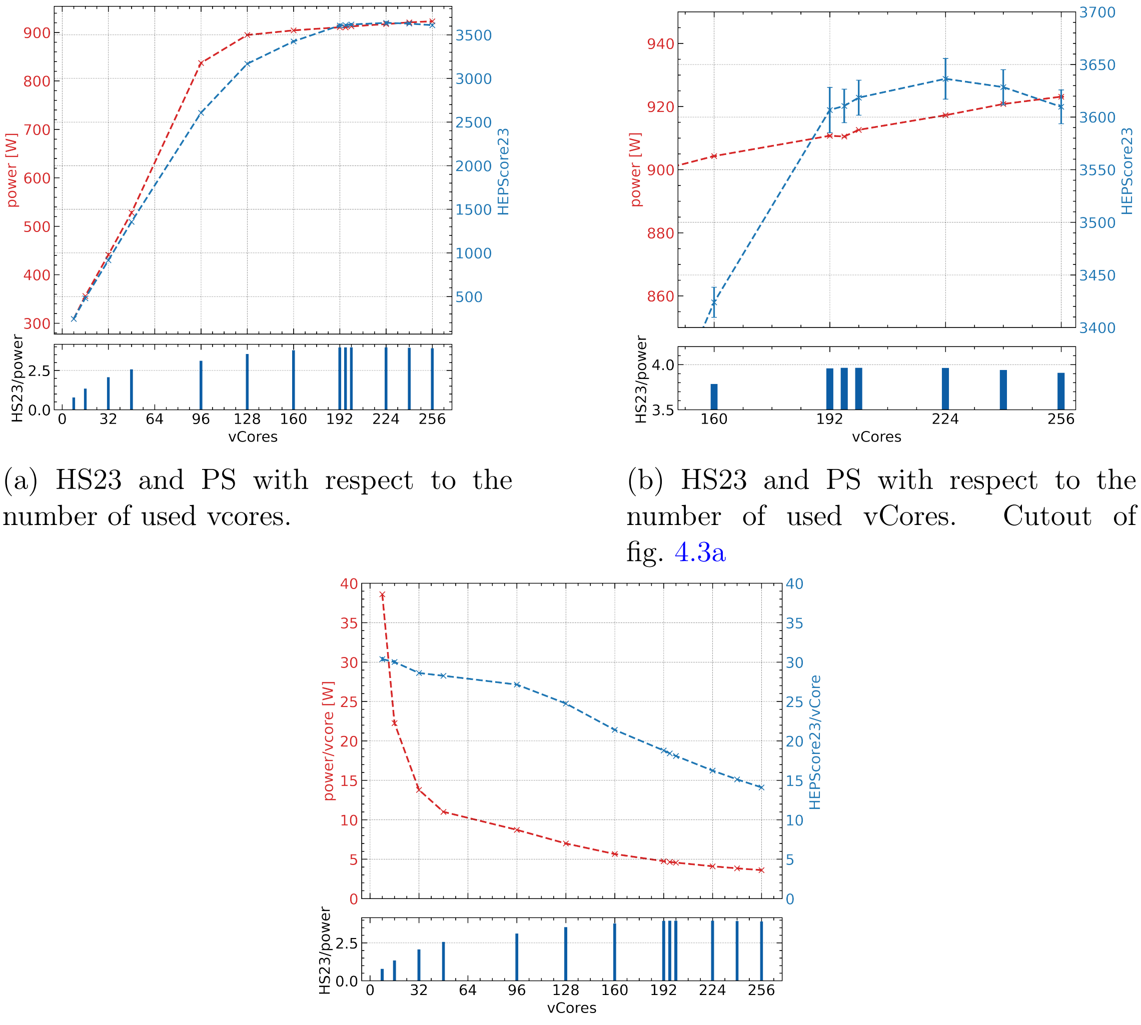
surements more accurately, with a higher configuration of vCore the truncated mean [1] seems to describe the measurements peak more, whereas the mean seems entirely inaccurate. Therefore, following along the argumentation given in section 3.3.1, the truncated mean will be used since a higher vCore configuration is expected to provide a higher HEPScore23-value and thus a more appropriate working point for HEP compute jobs.

4.2.2 Analysis

Using the defined metric of the truncated mean and the general analysis steps described for the vCore variation in section 3.2.2, a study about the dependence of the performance and power consumption of an ATLAS-BFG node using the AMD CPUs on the number of used vCores is now possible.

Figure 4.3a shows a similar pattern compared to fig. 3.3a, but the linearity of the performance whilst only using physical cores (up to 128 vCores) is visualized more

[1] mean of the quantiles from 0.7 to 0.9.



(a) HS23 and PS with respect to the number of used vcores.

(b) HS23 and PS with respect to the number of used vCores. Cutout of fig. 4.3a

(c) HS23/vcore and PS/vcore with respect to the number of used vcores.

Figure 4.3: Comparison of the vCore dependence of the HEPScore23-benchmark and its power consumption using the scaling governor "performance" on a node operating with two AMD CPUs available at BFG. (a) shows the dependence of the whole node on the number of vCores (b) is a cutout of (a) showing the range of vCores considered for operation of the HTC cluster and (c) shows the vCore dependence of performance and power consumption per vCore.

effectively and can be also observed in the power consumption. While the course of the performance with higher vCore configuration seems as linear as with only physical cores with a lower slope, fig. 4.3b provides a more detailed visualization of this part of fig. 4.3a, which is also more interesting to real-world use due to its higher total performance. The dependence of the performance in fig. 4.3b on the number of vCores is contrary to the first impression from fig. 4.3a not linear. For a configuration with vCores higher than 224 the performance drops, while the power consumption still rises linearly. This is also evident when considering the bar chart at the bottom of the figure, which visualizes the efficiency. However, considering the total values of the variation of the performance between 192 vCores and 256 vCores in contrast to the total HEPScore23 provided by the node during those configuration it is evident that the variation is not significant. Though there are other constraints on the use of the nodes for actual HEP-workloads: as described in section 2.2 the system needs to make available a certain amount of RAM and SSD memory per operated vCore. Therefore a working point with around 192 to 200 vCores should be the most suitable configuration.

4.3 Analysis of clock-frequency dependence

This section will provide an analysis of the clock-frequency dependence of an ATLAS-BFG node with AMD CPUs similar to the analysis conducted in section 3.3. However, there is one key difference: while for a node with Intel CPUs it was possible to configure clock-frequencies in steps of 200 MHz starting at 1200 MHz, this is not possible for the AMD CPUs by default. As mentioned in section 2.3.1, for some systems the kernel restricts the configuration of the clock-frequency, which is what occurs for the nodes with AMD CPUs. As a consequence, only a clock-frequencies of 1500, 2000 or 2700 MHz are configurable. In contrast to the measurement of the vCore variation of the nodes with AMD CPUs, the measurements concerning the clock-frequency variation operate with the scaling governor set to "userspace", similar to the measurements conducted on the nodes with Intel CPUs.

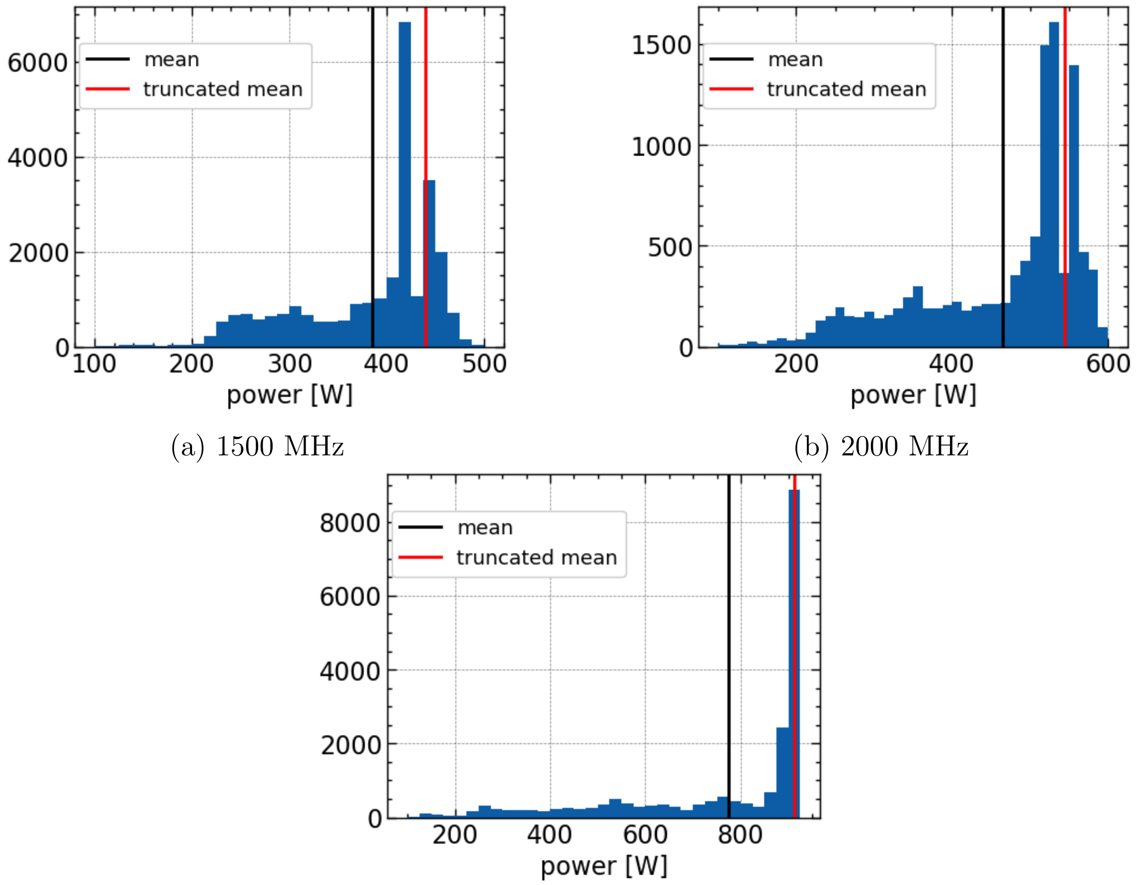
4.3.1 Definition of a power metric

Since there are only three possible working points for the clock-frequency, the figure which visualizes the distribution of the power consumption for a certain frequency configuration now contains all working points and thus all measurements conducted (fig. 4.4). When comparing the compatibility of the two considered metrics, median and truncated mean, for the three frequency configurations, the lowest frequency of 1500 MHz displays a similar compatibility with both metrics. The other two frequency configurations however seem to be more compatible with the truncated mean, especially the configuration of 2700 MHz. Therefore the truncated mean will be used again, which also creates a consistent metric for power consumption throughout this thesis.

4.3.2 Analysis

With the defined metric for power consumption, the dependence of performance and power consumption on the clock-frequency can now be studied for an ATLAS-BFG node with AMD CPUs. The values describing a certain configuration are again the mean values of a all measurements conducted with that configuration (section 3.2.2). The error-bars are the standard deviation of those mean values and are hardly visible in fig. 4.5 because of the small deviation between different measurements with the same configuration of the same node.

As displayed by the HEPScore23-graph in fig. 4.5, the dependence of the performance on the clock-frequency is linear, similar to the node with Intel CPUs. The dependence of the power consumption for a higher clock-frequency is not as predictable: while the rise from 1500 MHz to 2000 MHz is initially lower than the performance rise, the rise from 2000 MHz to 2700 MHz is higher. This results in the highest efficiency at a clock-frequency of 2000 MHz, whereas a clock-frequency



(a) 1500 MHz

(b) 2000 MHz

(c) 2700 MHz

Figure 4.4: Histograms of all measured PS-values using 196 vCores and a clock-frequency of (a) 1500 MHz, (b) 2000 MHz and (c) 2700 MHz using an ATLAS-BFG node with AMD CPUs. The black line visualizes the mean while the red line visualizes the truncated mean, specifically the mean of the quantiles 0.7 to 0.9.

of 2700 MHz is still more efficient than a clock-frequency of 1500 MHz, as described by the bar chart.

This is a contrast to the results for ATLAS-BFG nodes with Intel CPUs: since a clock-frequency of 2000 MHz is similar in efficiency to a clock-frequency of 2400 MHz the working point can be simply determined by the necessary performance. Here however a combination of multiple configurations could be used to meet a certain average performance over a longer period of time with the goal of operating more (energy) efficient.

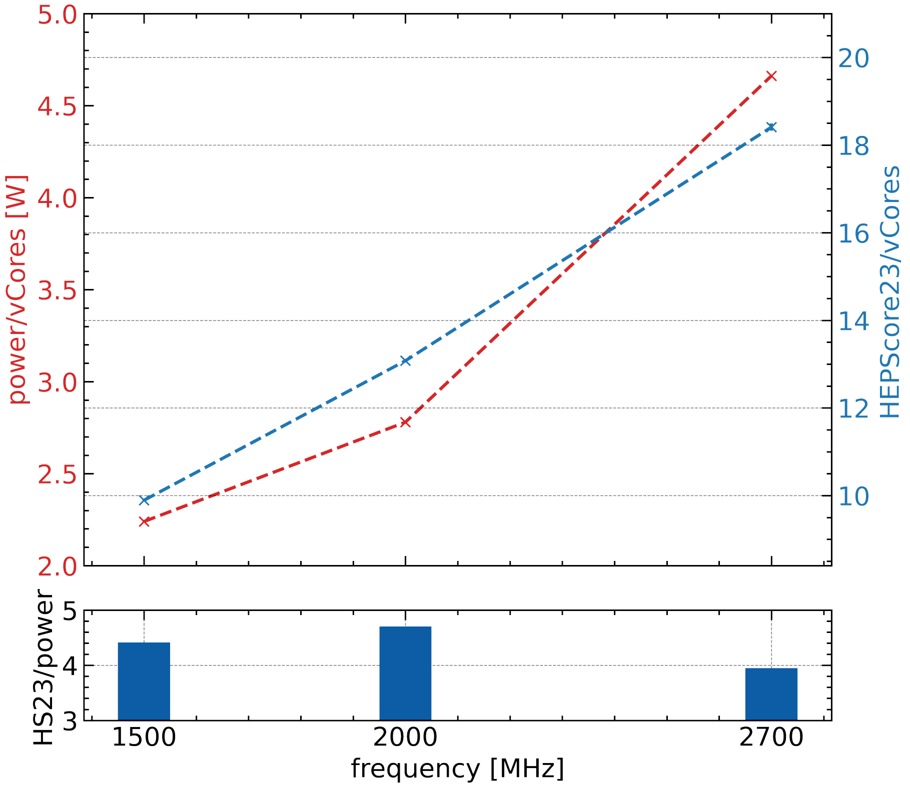


Figure 4.5: Frequency-dependence of the HEPScore23-benchmark compared to the power consumption using an ATLAS-BFG node with AMD CPUs.

5. Dynamic adjustment of CPU clock-frequency according to share of renewable energy at production

Using the obtained data from the analyses conducted for ATLAS-BFG nodes with AMD CPUs, this chapter will attempt to develop methods for the reduction of the CO_2-footprint of the new hardware at the ATLAS-BFG cluster. Afterwards the possible applications of these methods will be discussed.

The computing cluster ATLAS-BFG operated by the Schumacher-group in Freiburg has enhanced its computing power with the hardware changes during 2024. It was previously composed of 84 nodes with Intel CPUs, while the current hardware consists of 12 nodes with AMD CPUs. As per agreement with the ATLAS-collaboration, the computing site has to provide a HEPScore23-value of 33300 as an average value throughout the year 2025. Using the previous site composition of Intel-nodes, the possible maximum provided HEPScore23-average would be $HS_{max} = 33600$, just little more than the agreed upon value. However, with the new hardware the maximum providable HEPScore23 of the BFG-cluster is $HS_{max} = 42336$, which is about 127% of what the site has to provide. This opens the opportunity to use this flexibility to not operate at the highest possible performance at all times to save energy and with that reduce the CO_2-footprint of the ATLAS-BFG cluster.

Turning the cluster off after providing the necessary performance for the year would be the simplest approach to achieve a reduction of total energy consumption, the different frequency configurations of the nodes with AMD-sockets and their different efficiency to split the total necessary performance between two configurations with different clocking-frequency to lower energy consumption. Since the top performance of the nodes at 2700 MHz is needed to even achieve the pledge it will serve as the high-performance configuration. Thus leaving three different frequency combinations (table 5.1) which shall be compared in the following.

Table 5.1: Different frequency configuration for the combinations which energy efficiency is to be discussed in the following. The clock-frequency of 0 represents that the HTC cluster is turned of completey.

high-performance frequency [MHz]	2700	2700	2700
low-performance frequency [MHz]	0	1500	2000

5.1 Energy market

When splitting the total required performance between two configurations another objective could be accomplished apart from saving energy in total: the reduction of the clusters carbon footprint whilst still delivering the same performance. This will be argued with consideration of the German energy market in the following.

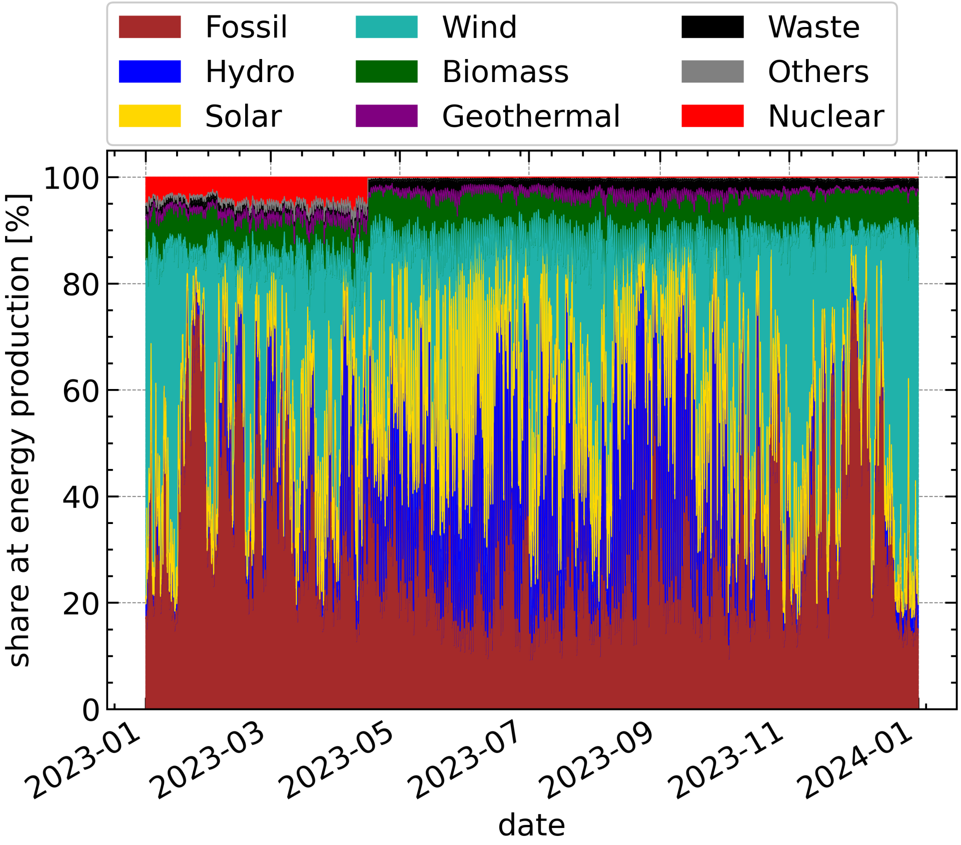


Figure 5.1: Shares of specific energy sources on the German energy market [26] for the calender year 2023 by the Fraunhofer ISE. Some energy sources are split further in the original data but are added here for visualization purposes.

The course of the German energy generation throughout the year 2023 is depicted in fig. 5.1. While the sources in fig. 5.1 are only split between the general energy generators hydro, solar, wind and fossil energy, they consist of several categories themselves, i.e. wind energy can be split further into on- and offshore wind energy.

The development of the energy production over the year displays not a constant but a varying amount, both in total and of the share of renewable energy. It is visible that for example solar energy production varies daily and seasonally, which is due to the suns course throughout the day and year. While a split between two configurations as suggested before has the same total energy consumption independent of exactly when which configuration is used, the carbon emission from that energy consumption dependents on the energy-mix used and therefore also on the time during which a configuration is used. If for example on a day with high solar energy production the more energy-consuming high-frequency configuration is used during the period of more renewable share of generation (during daylight hours) and the low-frequency configuration is used during nighttime hours the carbon footprint of the those computations would be less compared to a combination where the times are inverted.

For a calculation of the carbon emission of a kWh at a certain time a decomposi-

Table 5.2: Carbon emission for different energy sources listed by IPCC [27] and linked to the energy sources listed by Fraunhofer ISE [26].

energy source listed by IPCC	assigned energy sources listed by Fraunhofer ISE	CO_2eq/kWh [g]
Coal	Fossil brown coal/lignite Fossil hard coal	820
Gas	Fossil gas Fossil oil	490
Biomass - dedicated	Biomass	230
Geothermal	Geothermal	38
Hydropower	Hydro Run-of-River Hydro pumped storage Hydro water reservoir	24
Nuclear	Nuclear	12
Concentrated Solar Power	Solar	27
Wind onshore	Wind onshore	11
Wind offshore	Wind offshore	12

tion of the carbon emissions of the separate energy sources in fig. 5.1 is needed. This is provided by data from the IPCC [27] displayed in table 5.2. However, the energy sources listed by the IPCC do not match the sources in fig. 5.1 exactly. Therefore, table 5.2 also displays which sources of the IPCC data will be used to described which source from the data in fig. 5.1.

However, not all of the sources which do not have an equivalent in the IPCC data can be combined as easily as for example the different kinds of hydro energy. In particular, the sources "fossil oil", "waste" and "others" fall in that category. An individual solution is therefore needed: for fossil oil the same value as for fossil gas is used as an estimate and for waste and others the mean CO_2eq/kWh of 380 g is used [28].

5.2 Optimization options on the ATLAS-BFG cluster

To correctly and consistently apply a combination of configurations, a condition for switching from a high clock-frequency to lower (and vice versa) is needed. The objective is to reduce the carbon footprint of the clusters computations, so switching to the high-frequency-configuration only when a certain share of renewable at the energy production is available. But since the pledge of the ATLAS-BFG cluster still needs to be met, this share of renewable energy production at which clock-frequencies are switched needs to be determined[1]. In fig. 5.2, the average HEPScore23-value the HTC cluster provides throughout the year is displayed in dependence of the share of renewable energy production which needs to be exceeded for the configuration to be switched to the higher frequency (to 2700 MHz from 2000 MHz or 1500 MHz). The

[1] To avoid lengthy sentences this will be now referred to as "switching-share".

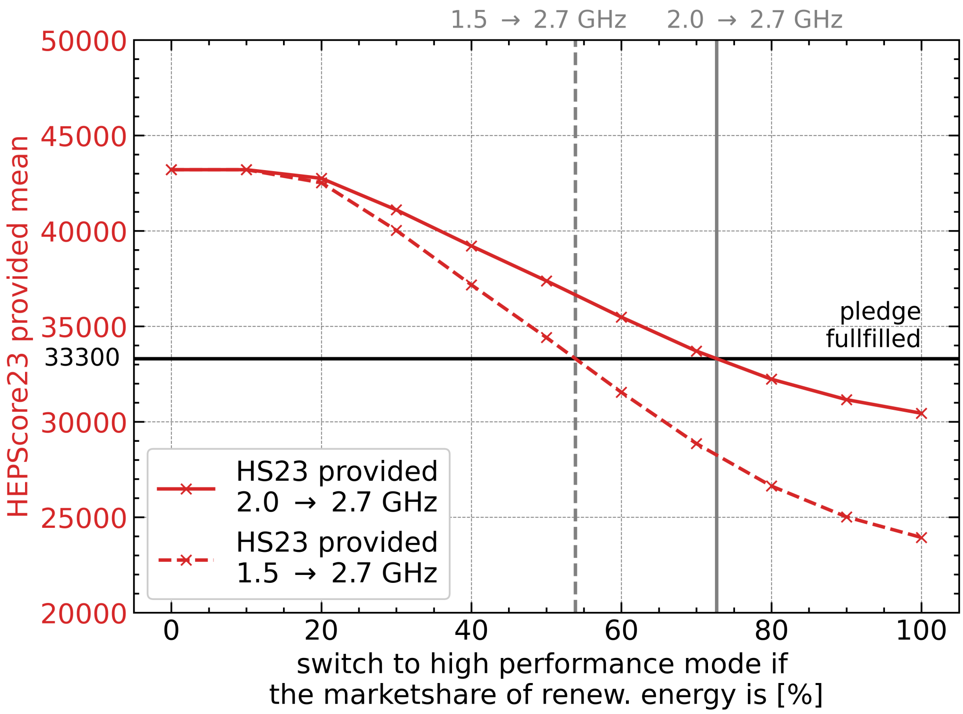


Figure 5.2: Total median provided HEPScore23-value (eq. (5.1)) of the ATLAS-BFG cluster for different configurations. 1.5 → 2.7 GHz represents switching the clock-frequency from 1.5 to 2.7 GHz when the share of renewable energy at production exceeds the specified value on the x-axis. Similar 2.0 → 2.7 GHz, only the lower clock-frequency is 2.0 GHz. The nodes would be set up with the available AMD-sockets (with 196 vCores) and the calculation is run for the year 2023.

provided average HEPScore23-value (HS_a) for the year 2023 is calculated as follows: since the data for the energy production is given in intervals of 15 minutes for each of these intervals a HEPScore23-value is determined based on the switching-share. The values for the different clock-frequencies (1500, 2000 2700 MHz with 196 vCores per node) are displayed in section 4.3.2). This essentially yields a timespan for both the lower and higher clock-frequencies which each have their HEPScore23-value. By taking the weighted mean

$$HS_a = \frac{t_{low} \cdot HS_{low} + t_{high} \cdot HS_{high}}{t_{low} + t_{high}} \tag{5.1}$$

the provided average HEPScore23-value is obtained, where t_{low} is the total time spent in the low-frequency configuration, HS_{low} the HEPScore23-value provided by that configuration.

The black horizontal line visualizes a provided median of 33300 HS23 throughout the year. The gray vertical lines visualize the curves intersection with the pledge-full-filled-line; they indicate the share of renewable energy generation upon the lower frequency needs to switched to 2700 MHz (switching-share) to still be able to full-fill the pledge. These points are

$$s_{1.5\to2.7} = 53.7\%$$

$$s_{2.0\to2.7} = 72.8\%$$

.

The corresponding total energy consumption of a combination of configurations for a certain switching-share is displayed in fig. 5.3 alongside with the switching-shares efficiency.

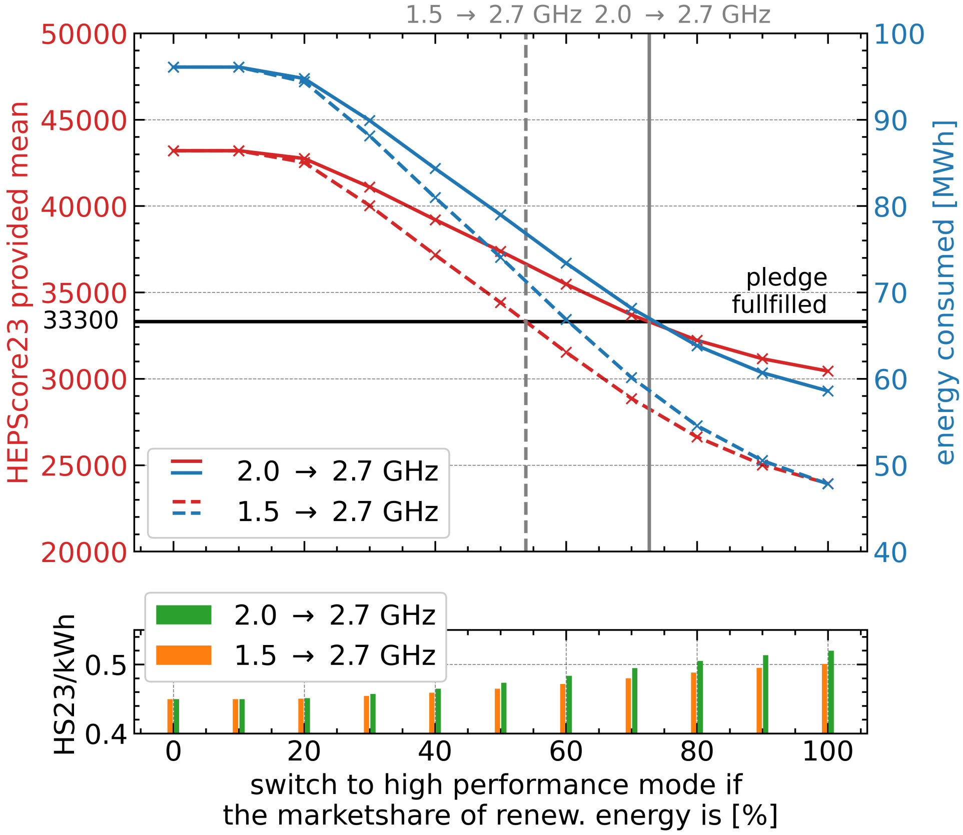


Figure 5.3: Total energy consumption and mean provided HEPScore23-value of the ATLAS-BFG cluster in for different combinations of configurations in dependence of the share of renewable energy production upon to switch those configurations.

The similar values for all curves for the switching-shares up to 30% are due to the minimum share of renewable energy generation. Therefore, especially for a switching-share of up to 20%, the systems always operates on the higher frequency of 2700 MHz.

After the similar efficiency[2] up to 30%, the 2.0 → 2.7 GHz combination becomes more efficient than the 1.5 → 2.7 GHz combination with a higher value of switching-share. This is due to the configuration of 2.0 GHz being more efficient than the configuration of 1.5 GHz, as discussed in section 4.3.2, and the lower frequency-configuration being more used for a higher number of switching-share.

Like before, the total power consumptions of the switching-shares which yield exactly the pledged mean of HEPScore23 can be read to be

$$E_{1.5\to2.7} = 71.4\,\text{MWh}$$

$$E_{2.0\to2.7} = 66.9\,\text{MWh}$$

[2] refers to energy efficiency; HS23/MWh

.

To calculate the total carbon emissions throughout the year, each of the energy sources is broken down to a percentage of the total energy production and then multiplied by a CO_2/kWh-value according to table 5.2. This creates information about the carbon emissions of one kWh energy in a 15 minutes interval. By then multiplying this value with the energy the site consumes during a 15 minute interval with a specific, by the share of renewable energy at energy production determined frequency configuration the total carbon emissions of the site during these 15 minute intervals are obtained. When summed over the whole year, the sites total carbon emissions are obtained, which is what is displayed in fig. 5.4 for both frequency-combinations.

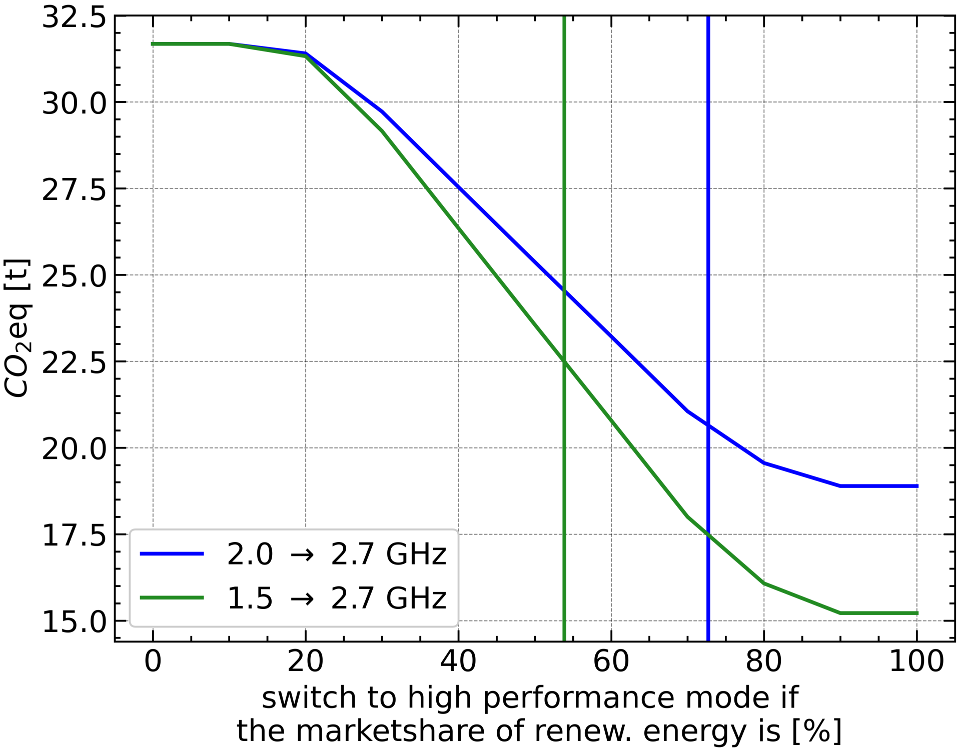


Figure 5.4: Yearly CO_2 emissions of the ATLAS-BFG cluster with different clock-frequency-combinations in dependence of the share of renewable energy at production upon which the configuration is switched to the higher frequency.

The curves in fig. 5.4 behave somewhat as expected: the curve for the 1.5 → 2.7-combination is generally below the one for 2.0 → 2.7-combination, similar to the total energy consumed, which is expected due to a positive correlation of energy consumption and produced carbon. The exact values are

$$C_{1.5\rightarrow2.7} = 23.2\,\mathrm{t}$$

$$C_{2.0\rightarrow2.7} = 21.2\,\mathrm{t}.$$

The combination with the more efficient clock-frequency of 2.0 GHz combination of 2.0 → 2.7 GHz also has a lower total carbon footprint. It is thus advantageous to switch the clock-frequency from 2000 MHz to 2700 MHz compared to switching from 1500 MHz to 2700 MHz.

In conclusion, for the two considered combinations of frequency-configurations, the relevant values when full-filling the pledge of the BFG-cluster, providing a mean performance with a HEPScore23-value of 33300, are listed in table 5.3.

Table 5.3: Values produced by the BFG-cluster when full-filling its pledge.

combination	switching-share	energy consumed [MWh]	CO_2eq produced [t]
1.5 → 2.7GHz	53.7	71.4	23.2
2.0 → 2.7GHz	72.8	66.9	21.2

5.3 Comparison to default configuration

To compare the developed combinations of clock frequency configurations to a combination of using the HTC cluster with the highest performance for as long as it is needed to full-fill the yearly pledge (default configuration) and turning it off after, similar values regarding the power consumption and carbon footprint are necessary.

It is assumed that by turning the ATLAS-BFG cluster off no performance is provided and no power is consumed. For full-filling the pledge eq. (5.2) must hold where t_{op} is the number of days the ATLAS-BFG cluster needs to operate and 43327.2 is the clusters HEPScore23-value in this configuration, namely 12 nodes with AMD-sockets, 196 vCores used and a clock-frequency of 2.7GHz.

$$33300 = \frac{t_{op}}{365} \cdot 43327.2 \tag{5.2}$$

Therefore, the number of days during which the cluster needs to operate is given by

$$t_{op} = \frac{33300}{43327.2} \cdot 365\,\mathrm{d} = 280.5\,\mathrm{d}.$$

The time the cluster needs to compute yields the total consumed energy when multiplied by the power consumption of the clocking-frequency 2700 MHz:

$$E_{default} = 73.8\,\mathrm{MWh}.$$

The carbon footprint of this operation can be calculated the same way as before: by multiplying the consumed energy of a 15 minute interval with the corresponding value of CO_2eq/kWh and then adding all those values the following carbon footprint is obtained:

$$C_{\mathrm{default}} = 24.9\,\mathrm{t}.$$

This is significantly more than the values obtained when using the methods of splitting the operation between two clock-frequency-configurations. In comparison, by using a using clock-frequency of 2000MHz when the share of renewable energy at generation is smaller than 72.8% and a clock-frequency of 2700MHz when the share is higher, the ATLAS-BFG cluster could save 6.9 MWh energy and 3.7t of carbon emissions every year.

5.4 Applicability to other HTC clusters

The results developed in this chapter exemplary with the available hardware at the ATLAS-BFG cluster could be generally applied to other WLCG computing sites. However, there are few differences which need to be considered. First off all, the measurements of the power consumption and performance of a ATLAS-BFG node with the HEPScore23-benchmark would need to be conducted for the hardware used at the specific site. This would allow other HTC computing clusters to also save energy and lower their carbon footprint. The results, specifically the reduction of CO_2-emissions, need to be considered within the boundary conditions they are developed in. While the ATLAS-BFG cluster has an surplus of computing power this isn't generally the case.

Because of an inflation in the necessary computing power of the ATLAS experiment, computing sites need to plan ahead when updating their hardware. As a consequence, the surplus of a WLCG-site decreases over the years as their pledges increase. It is for this reason the ATLAS-BFG clusters composition of nodes with Intel-sockets has a comparatively little surplus of computing power. This also means that the optimal combination of clock-frequency configurations and the share of renewable energy production upon to switch the configurations has to be recalculated using a new pledge yearly.

An aspect which isn't considered within the calculation of the CO_2-emissions of the ATLAS-BFG cluster are the necessary CO_2-emissions in order to manufacture the hardware or rather the necessary surplus of computing power. This means it is unknown if deliberately buying more hardware to then reduce direct CO_2-emissions of the clusters energy consumption would reduce the total CO_2-emissions of the cluster. However, due to the delocalized and international nature of the WLCG computing grid singular sites have an individual surplus of computing power in order to full-fill their pledges, which can be utilized in the demonstrated matter to reduce CO_2-emissions of the WLCG-sites energy consumption.

6. Conclusion

This thesis has studied the power consumption of the compute nodes installed at the ATLAS-BFG cluster. In particular, the dependence of the power consumption on the configured number of vCores and clock-frequency has been analysed. The CO_2-footprint of the HTC cluster for different clock-frequency configurations has been calculated under the condition that the ATLAS-BFG cluster still full-fills its pledge.

The dependencies of power consumption and performance are measured by first setting up the hep-benchmark-suite, a tool to start and configure the HEPScore23-benchmark, on the nodes. The measurement process itself is based on two bash-scripts started simultaneously by a cronjob. One script starts the HEPScore23-benchmark using the hep-benchmark-suite and saves the relevant data while the other script is used to measure the power consumption of the node during the HEPScore23-benchmark-run.

The measured vCore dependencies of the nodes with Intel and AMD CPUs show a similar behaviour. Both display a saturation effect with a linear dependency upon the number of vCores as long as they are less than the number of physical cores (20/128 for nodes with Intel/AMD CPUs). For more vCores than physical cores, the curve starts flattening. The nodes with AMD CPUs display a performance drop when the maximum number of vCores (256) is approached, whereas for the nodes with Intel CPUs the performance stops to rise significantly from 32 to 40 vCores. For both node types, the most energy efficient configuration is around 150-160% of the number of physical cores, at around 32 vCores for the nodes with Intel CPUs and at around 196 vCores for the nodes with AMD CPUs.

For the clock-frequency dependency, a limit on the number of configurable clock-frequencies by the scaling governor of the nodes with AMD-CPUs limits comparability. However, for both node types the HEPScore23-values shows a linear dependence on the clock-frequency. The energy efficiency also rises at first for both node types. While the energy efficiency drops again for the highest measured clock-frequency on nodes with AMD-CPUs (2700 MHz), the nodes with Intel CPUs show a slightly different behaviour. The clock-frequency with the highest energy efficiency is 2000MHz, the clock-frequency of 2200 MHz shows a lower energy efficiency, whereas 2400 MHz (highest measured clock-frequency on nodes with Intel CPUs) has a similar energy efficiency as 2000 MHz.

The power consumption of the ATLAS-BFG cluster with 84 Intel nodes is 19.3MW whilst providing a performance of 32474 HS23[1], whereas the setup with 12 AMD nodes consumes 10.9 MW while providing 43327 HS23[2]. Therefore, by switching the setup from the Intel to the AMD node,s the HTC cluster already consumes 8.4 MW less, even when operating at full capacity. The performance of 43327 HS23 provided by the setup with new AMD nodes corresponds to 132% of the ATLAS-BFG pledges required for 2025. This surplus would enable the HTC cluster to operate at a lower clock-frequency (and therefore less performance) a portion of the year, while still full-filling the pledge. When the lower clock-frequency is used during times with a relatively low production of renewable energy, the CO_2-footprint of the ATLAS-BFG

[1] with each node set to 40 vCores and 2400 MHz

[2] with each node set to 196 vCores and 2700 MHZ

cluster could be reduced.

The two considered splits between clock-frequencies are a high clock-frequency of 2700 MHz and a low clock-frequency of either 1500 MHz or 2000 MHz. The CO_2-footprint of switching from 2000 MHz to 2700 MHz when the share of renewable energy at production exceeds 72.8% equates to 21.2 t, two tons less than when switching from 1500 MHz to 2700 MHz at 53.7%.

The CO_2-footprint of the configuration without switching clock-frequencies but simply turning of the HTC cluster when the pledge is full-filled after 280.5 days is calculated in the same way. This represents a best-case scenario as the overall median CO_2-footprint of the yearly energy market should be more than for the first 280.5 days, because this includes the lower CO_2-footprint of the solar energy produced during the summer. When compared to this configuration of the HTC cluster, the switch from 2000 MHz to 2700 MHz has a 14.8% lower CO_2-footprint and consumes 9.3% less energy (66.9 MWh/21.2 t compared to 73.8 MWh/24.9 t).

Apart from the significantly lower CO_2-footprint when using this switching of clock-frequencies according to the productions share of renewable energy, this should also lower the cost of the energy. As exemplary shown by fig. 6.1, the cost of energy has a high anti-correlation with the share of renewable energy of around -0.8. This could present additional incentive to implement such a method as it yields further benefits.

Because of the reduced CO_2-footprint and energy consumption and possibly lower cost it is recommendable to switch a HTC cluster dynamically to a lower clock-frequency when the share of renewable drops. This presumes the possibility of implementing such a method and surplus of possible performance. To evaluate if purposefully over provisioning the hardware installation to implement this, another study which additionally considers the hardware production and transportation CO_2-footprint needs to be considered.

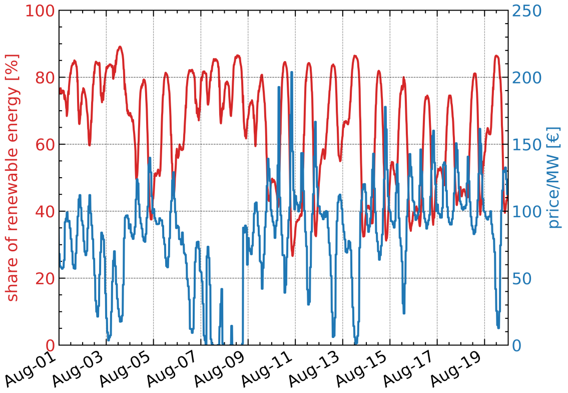
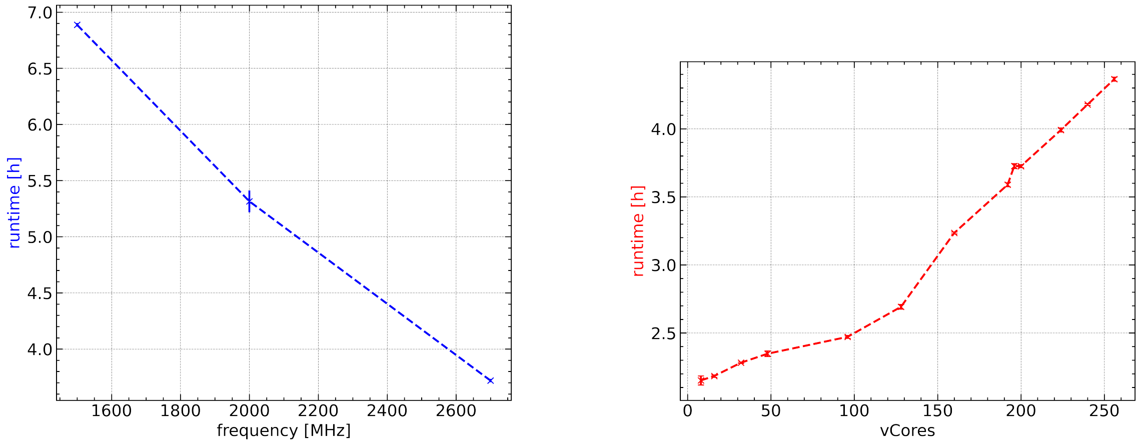


Figure 6.1: Price of energy in comparison to the share of renewable share of energy generation. Exemplary for 20 days in August of 2023.

A. Figures



(a) Runtimes of the HEPScore23-benchmark in dependence of the clock-frequency.

(b) Runtime of the HEPScore23-benchmark in dependence of the used vCores.

Figure A.1: Runtimes of the HEPScore23-benchmark in dependence of (a) the clock-frequency using 196 vCores and (b) the number of vCores with a varying clock-frequency and the scaling governor set to "performance".

B. Tables

Table B.1: Measured and calculated data for the analysis of the vCore dependence of an ATLAS-BFG node with Intel CPUs.

ncores	power/vcore [W]	HS23/vcore	power [W]	HS23	runtime [h]
4.0	38.8	20.5	155.1	82.2	3.3
8.0	21.6	19.2	172.6	153.8	3.6
12.0	15.4	17.7	185.2	212.7	3.9
16.0	12.1	16.9	194.3	269.7	4.0
20.0	10.5	16.6	210.6	331.4	4.1
24.0	8.7	14.8	209.8	354.3	4.7
28.0	7.8	13.6	217.5	379.8	5.2
32.0	6.8	12.0	216.5	384.7	5.8
36.0	6.2	10.7	222.6	384.8	6.1
40.0	5.7	9.7	229.4	386.6	6.5

Table B.2: Measured and calculated data for the analysis of the clock-frequency dependence of an ATLAS-BFG node with Intel CPUs.

frequency [MHz]	power/vcore [W]	HS23/vcore	runtime [h]
1200.0	3.7	4.9	12.2
1400.0	3.9	5.7	10.6
1600.0	4.2	6.5	9.4
1800.0	4.5	7.3	8.5
2000.0	4.7	8.0	7.7
2200.0	5.4	9.0	7.0
2400.0	5.8	9.8	6.5

Table B.3: Measured and calculated data for the analysis of the vCore dependence of an ATLAS-BFG node with AMD CPUs

vCores	power/vCore [W]	HS23/vCore	power [W]	HS23	runtime [h]	frequency
8.0	38.6	30.4	308.8	242.9	2.2	2500.0
16.0	22.3	30.0	356.1	480.1	2.2	2500.0
32.0	13.8	28.6	441.3	915.5	2.3	2500.0
48.0	11.0	28.3	528.1	1356.4	2.3	2600.0
96.0	8.7	27.1	836.9	2605.4	2.5	2700.0
128.0	7.0	24.7	894.8	3167.3	2.7	2700.0
160.0	5.7	21.4	904.3	3424.0	3.2	2700.0
192.0	4.7	18.8	910.7	3606.7	3.6	2700.0
196.0	4.6	18.4	910.5	3610.6	3.7	2700.0
200.0	4.6	18.1	912.6	3618.4	3.7	2700.0
224.0	4.1	16.2	917.2	3636.4	4.0	2700.0
240.0	3.8	15.1	920.8	3628.4	4.2	2700.0
256.0	3.6	14.1	922.8	3609.1	4.4	2700.0

Table B.4: Measured and calculated data for the analysis of the clock-frequency dependence of an ATLAS-BFG node with AMD CPUs.

frequency [MHz]	power/vcore [W]	HS23/vcore	runtime [h]
1500.0	2.2	9.9	6.9
2000.0	2.8	13.1	5.3
2700.0	4.7	18.4	3.7

Table B.5: Calculated data concerning the split of clock-frequencies between 1500 MHz and 2700 MHz. The data is shown in fig. 5.3.

renew. energy share to switch clock-frequencies	median HS23	energy [MWh]	days with 2700 MHz	days with 1500 MHz
0	43200.0	96.1	365.0	0.0
10	43200.0	96.1	365.0	0.0
20	42518.5	94.4	352.5	12.5
30	40021.9	88.1	306.8	58.2
40	37169.8	81.0	254.6	110.4
50	34411.5	74.1	204.1	160.9
60	31540.6	66.9	151.5	213.5
70	28855.1	60.1	102.3	262.7
80	26630.4	54.6	61.6	303.4
90	25016.0	50.5	32.0	333.0
100	23937.0	47.8	12.2	352.8

Table B.6: Calculated data concerning the split of clock-frequencies between 2000 MHz and 2700 MHz. The data is shown in fig. 5.3.

renew. energy share to switch clock-frequencies	median HS23	energy [MWh]	days with 2700 MHz	days with 2000 MHz
0	43200.0	96.1	365.0	0.0
10	43200.0	96.1	365.0	0.0
20	42748.7	94.8	352.5	12.5
30	41095.3	89.9	306.8	58.2
40	39206.5	84.3	254.6	110.4
50	37379.8	79.0	204.1	160.9
60	35478.5	73.4	151.5	213.5
70	33700.1	68.2	102.3	262.7
80	32226.7	63.8	61.6	303.4
90	31157.6	60.7	32.0	333.0
100	30443.0	58.6	12.2	352.8

Bibliography

[1] *CERN*. accessed 17.12.2024. URL: https://home.cern/.

[2] *Worldwide LHC Computing Grid*. accessed 17.12.2024. URL: https://wlcg.web.cern.ch/.

[3] Domenico Giordano et al. "HEPScore: A new CPU benchmark for the WLCG". In: *EPJ Web of Conferences* 295 (2024). Ed. by R. De Vita et al., p. 07024. ISSN: 2100-014X. DOI: 10.1051/epjconf/202429507024. URL: http://dx.doi.org/10.1051/epjconf/202429507024.

[4] Domenico Giordano. *hep-score*. accessed: 12.11.2024. URL: https://gitlab.cern.ch/hep-benchmarks/hep-score.

[5] Michele Michelotto et al. "A comparison of HEP code with SPEC1 benchmarks on multi-core worker nodes". In: *Journal of Physics: Conference Series* 219.5 (Apr. 2010), p. 052009. DOI: 10.1088/1742-6596/219/5/052009. URL: https://dx.doi.org/10.1088/1742-6596/219/5/052009.

[6] The ATLAS Collaboration et al. "The ATLAS Experiment at the CERN Large Hadron Collider". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08003. DOI: 10.1088/1748-0221/3/08/S08003. URL: https://dx.doi.org/10.1088/1748-0221/3/08/S08003.

[7] *ATLAS collaboration*. accessed 17.12.2024. URL: https://atlas.cern/Discover/Collaboration.

[8] The CMS Collaboration et al. "The CMS experiment at the CERN LHC". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004. URL: https://dx.doi.org/10.1088/1748-0221/3/08/S08004.

[9] *CMS collaboration*. accessed 17.12.2024. URL: https://cms.cern/collaboration.

[10] A Augusto Alves Jr et al. "The LHCb detector at the LHC". In: *Journal of instrumentation* 3.08 (2008), S08005.

[11] *LHCb collaboration*. accessed 17.12.2024. URL: https://lhcb.web.cern.ch/.

[12] Kenneth Aamodt et al. "The ALICE experiment at the CERN LHC". In: *Journal of Instrumentation* 3.08 (2008), S08002.

[13] *ALICE collaboration*. accessed 17.12.2024. URL: https://alice-collaboration.web.cern.ch/.

[14] Belle-II Collaboration, II Belle, et al. "Technical design report". In: *arXiv preprint arXiv:1011.0352* (2010).

[15] *Belle II collaboration*. accessed 17.12.2024. URL: https://www.belle2.de/belle-2-experiment/kollaboration.

[16] Domenico Giordano. *HEP workloads*. accessed 17.12.2024. URL: https://gitlab.cern.ch/hep-benchmarks/hep-workloads.

[17] Menéndez Borge, Gonzalo. "HEP Benchmark Suite The centralized future of WLCG benchmarking". In: *EPJ Web of Conf.* 295 (2024), p. 07023. DOI: 10.1051/epjconf/202429507023. URL: https://doi.org/10.1051/epjconf/202429507023.

[18] Domenico Giordano. *hep-benchmark-suite*. accessed: 12.11.2024. URL: https://gitlab.cern.ch/hep-benchmarks/hep-benchmark-suite.

[19] *CPU frequency scaling*. accessed 17.12.2024. URL: https://wiki.archlinux.org/title/CPU_frequency_scaling.

[20] Duncan Laurie. *ipmitool*. accessed 17.12.2024. URL: https://github.com/ipmitool/ipmitool.

[21] *uptime*. accessed 17.12.2024. URL: https://linuxhandbook.com/uptime-command/.

[22] *Cron*. accessed 17.12.2024. URL: https://wiki.archlinux.org/title/Cron.

[23] *pandas*. accessed 17.12.2024. URL: https://pandas.pydata.org/.

[24] *What is Load Average in Linux?* accessed 17.12.2024. URL: https://www.digitalocean.com/community/tutorials/load-average-in-linux.

[25] Dr. Emanuele Simili. *Heterogeneous Tier2 Cluster and Power Efficiency Studies*. accessed 17.12.2024. URL: https://indico.cern.ch/event/1386170/contributions/6151372/attachments/2936329/5157588/ScotGrid_HTCws.pdf.

[26] *energy charts*. accessed 4.12.2024. URL: https://www.energy-charts.info/charts/power/chart.htm?l=en&c=DE&year=2023&source=public&legendItems=3x2vvv6&interval=year.

[27] Steffen Schloemer et al. "Annex III: Technology-specific cost and performance parameters". English. In: *Climate Change 2014: Mitigation of Climate Change*. Ed. by O Edenhofer et al. United Kingdom: Cambridge University Press, 2014, pp. 1329–1356. ISBN: 978-1-107-65481-5.

[28] Dr. Thomas Lauf Petra Icha. *Entwicklung der spezifischen Treibhausgas-Emissionen des deutschen Strommix in den Jahren 1990 - 2023*. accessed 10.12.2024. Umweltbundesamt, 2023. URL: https://www.umweltbundesamt.de/sites/default/files/medien/11850/publikationen/23_2024_cc_strommix_11_2024.pdf.