

Optimizing GPU Frequency for Sustainable HPC: Lessons Learned from a Year of Production on Adastra, an AMD GPU Supercomputer

Gabriel Hautreux

hautreux@cines.fr

CINES

Montpellier, France

Naima Alaoui

alaoui@cines.fr

CINES

Montpellier, France

Etienne Malaboeuf

malaboeuf@cines.fr

CINES

Montpellier, France

ABSTRACT

Power consumption remains a critical concern in today’s GPU-based high-performance computing (HPC) systems. With rising energy costs and growing environmental concerns, optimizing power consumption while maintaining performance has become a priority for HPC centers worldwide. Modern GPUs, such as the AMD MI250X architecture used in the Adastra supercomputer, incorporate features like frequency scaling, which allows dynamic control over power consumption. However, determining the optimal frequency configuration for diverse workloads is a complex challenge due to the varied computational demands of HPC and AI applications. A previous frequency analysis conducted by CINES on a workload equivalent to production, presented at SuperComputing 23, demonstrated that the performance decrease caused by GPU frequency downscaling can significantly improve the energy efficiency of the system. Based on these findings, CINES decided to scale down the default GPU frequency of its machines from 1.7 GHz to 1.5 GHz in July 2024. This new study investigates the impact of reducing the GPU frequency of our AMD MI250X system (Adastra) from 1.7 GHz to 1.5 GHz, hosted at the Centre Informatique National de l’Enseignement Supérieur (CINES) in Montpellier, France. Adastra, ranked multiple times as #3 on the Green500 list for energy-efficient supercomputers, supports hundreds of French researchers across various scientific and engineering domains. This paper focuses on energy savings, application performance, user satisfaction, hardware reliability, job typology, and differences in impact between HPC and AI workloads. Energy consumption analysis over a year—comparing six months before and six months after the frequency adjustment—revealed a 14% reduction in overall energy usage per node. This decrease aligns with the sustainability objectives of HPC centers without requiring hardware modifications. The comparative analysis of HPC and AI workloads underscores the varied effects of frequency scaling, with different outcomes observed depending on workload type. This research provides actionable insights for HPC centers seeking to optimize energy efficiency while maintaining operational performance, without implementing dynamic strategies that could introduce additional overhead in large systems.

KEYWORDS

Energy Efficiency, GPU, Frequency capping, Sustainability

1 INTRODUCTION

In the fast-paced world of High-Performance Computing (HPC), managing power consumption in GPU-accelerated systems has

become crucial. Many top-ranked machines in the Top500 now exceed the 20MW threshold, with this trend accelerating due to AI advancements and government announcements in GW-scale data centers.

As energy costs escalate and environmental concerns mount, HPC centers globally are exploring strategies to optimize power usage without compromising performance. The Centre Informatique National de l’Enseignement Supérieur (CINES) in Montpellier, France, home to the Adastra supercomputer, is at the forefront of this effort. Recognized for its energy efficiency and multiple top rankings on the Green500 list, Adastra supports hundreds of researchers across various scientific and engineering fields. However, CINES continues to seek further energy savings.

This study examines the impact of reducing the GPU frequency from 1.7 GHz to 1.5 GHz on the Adastra system, building on research presented at SuperComputing 23 [2]. The decision to implement this change in July 2024 was driven by findings that demonstrated significant energy efficiency improvements with minimal performance loss at this frequency. By analyzing user data, job performance, hardware replacement, and power draw metrics over a year, this research aims to reveal the potential energy savings and operational efficiencies achievable through GPU frequency scaling.

This study is a pioneering exploration of the real-world implications of GPU frequency scaling in a large-scale HPC environment. The insights gained can guide similar studies by other HPC centers, contributing to a broader understanding of sustainable computing policies. Through comprehensive analysis, we aim to provide valuable insights into sustainable HPC operations, highlighting the varied effects of frequency scaling. The findings offer an actionable strategy for HPC centers seeking to optimize energy efficiency without complex dynamic adjustments, supporting the goal of sustainable and efficient computing at a reasonable cost.

2 RELATED WORK

Power management plays a critical role in the sustainable operation of HPC centers. A power and frequency capping impact study was conducted on Frontier [4] to classify applications and apply, in future works, dynamic settings based on those findings. This paper also shows that frequency capping has a limited impact on MI250X at 1.5GHz.

Another interesting study [3] on this subject was conducted by EPCC at a full CPU system level, applying frequency capping on the CPU. This approach is less deterministic than full system power capping but yields very interesting results for both performance and energy savings.

We hope that our complementary study will promote the effectiveness of frequency capping policies to gather more feedback at the full system level and enable better decision-making for HPC operators.

3 EXPERIMENTAL SETUP

3.1 HPC System Description

The Adastra supercomputer, located at the Centre Informatique National de l'Enseignement Supérieur (CINES) in France, is a state-of-the-art high-performance computing (HPC) system designed to support advanced research and scientific computing. Adastra is equipped with state-of-the-art hardware and software to meet the demanding computational needs of various scientific disciplines. Adastra features an heterogeneous architecture with thousands of compute nodes, each equipped with high-performance processors and accelerators. The system includes 356 accelerated nodes, featuring a 64-core Trento CPU coupled with 4 AMD MI250X GPUs. It is built on a 200GB/s Slingshot interconnect network to facilitate efficient data transfer and communication between nodes. The software environment includes a comprehensive suite of scientific applications, libraries and tools optimized for parallel computing. The operating system (Redhat) and resource management software (SLURM [5]) ensure efficient use of computing resources and support for diverse workloads.

The study presented in this paper is focusing on the 356 MI250X accelerated nodes of the Adastra cluster.

3.2 Frequency Policy Evolution

While the Adastra system has been operating in production since late 2022, the CINES team conducted multiple experiments with power capping and frequency capping of the GPUs [2] and decided to decrease the standard frequency of all the GPUs within the Adastra system, as the performance impact seemed minimal.

The standard frequency was decreased from 1.7GHz to 1.5GHz on the 3rd of July 2024. In this context, we find it useful to share within this paper all the metrics we can between the two semesters:

- **2024-S1:** First semester of 2024, standard frequency of the GPUs at 1.7GHz
- **2024-S2:** Second semester of 2024, standard frequency of the GPUs at 1.5GHz

On top of that, it is important to note that the "power boost" mode is enabled on the Trento CPUs, as the performance gains highlighted are substantial for the tests we performed [2].

4 RESULTS AND ANALYSIS

4.1 User Base Evaluation

4.1.1 Data collection. In order to validate the relevance of this study, we first wanted to ensure that the user base between the two semesters did not change. Using SLURM, we retrieved the data of all the jobs submitted on the system for each semester and counted the users present in both semesters. We did the same for the amount of GPU.hours consumed by the users. All this information is centralized in 1.

4.1.2 Analysis. Based on Table 1, especially the "S1-S2 common" line, we can see that 145 users were active in 2024 on the GPU partition of Adastra. These users represent 48% of the total number of users in 2024. However, these 48% of users consumed 95.2% of the GPU.hours on the system.

In this context, we can validate that both semesters have the same user base, as they are consuming almost all the available hours on the system. If the user base was different between both semesters, we could not really assess the impact of the frequency change at the GPU level.

	Number of users	M of GPU.hours consumed
2024	292	10.693
S1	223	4.794
S2	214	5.899
S1-S2 common	145 (48%)	10.181 (95.2%)

Table 1: Users and GPU.hours consumption for MI250X Partition

4.2 Impact of GPU Frequency Capping on Jobs

4.2.1 Data collection. In order to assess the impact of GPU frequency capping on jobs, we retrieved the information for the two semesters from our SLURM database. For each semester, we retrieved the "job end state" of all the jobs, as well as the elapsed time and number of nodes to compute the GPU.hours volume used by each job. We then grouped the jobs by end state to perform this study.

4.2.2 Analysis. The observations we want to make is the difference in the end state relatively to both the total number of jobs run during these two semesters, as well as the corresponding GPU.hours volume consumed. We plot pie charts showing this distribution in 1.

While the percentage of completed jobs is similar between 2024-S1 and 2024-S2, it is interesting to observe that the number of jobs that went into a Timeout end state did not increase between the two semesters. In fact, it decreased from 2.5

However, the number of jobs may not represent the GPU.hours spent, i.e., the real usage of the system, so we also look at the pie charts for the GPU.hours volume for each end state. In the same figure, we observe that the volume of GPU.hours spent in the machine, ending up in timeout, decreased by 1.8%. However, we notice an increased volume of failed jobs (1.4%) while also observing an 8% increase in completed jobs. The most significant change is the reduction of canceled jobs, which represented more than 11% of the time spent on the machine, compared to only 4% in the second semester.

Globally, we would have expected that some jobs could suffer from longer run times due to GPU frequency capping, but the reduced number of jobs and GPU.hours spent ending in Timeout did not increase during the second semester. On top of that, the GPU.hours volume spent for Completed jobs increased from approximately 71% to 79%.

In this context, we are confident that this GPU frequency reduction did not impact the jobs in a significant way.

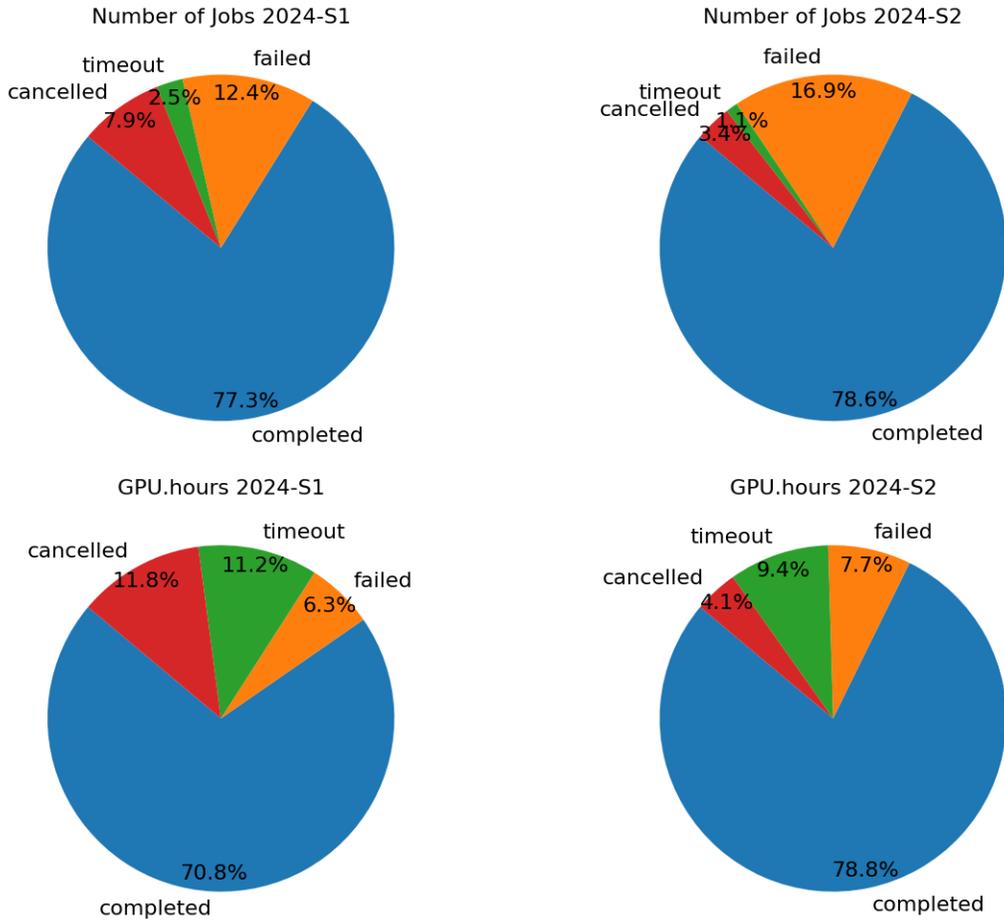


Figure 1: Distribution of job states for S1 and S2 number of jobs

4.3 Impact of GPU Frequency Capping on Jobs Power Draw

4.3.1 *Data collection.* To assess the energy consumption of all our production workload comparatively from the first semester to the second, we retrieved all the job energy information from SLURM, based on the energy performance counters, with a sampling rate of 30 seconds. Unfortunately, we observed that some jobs had no energy information, and short-duration jobs might have incoherent values due to SLURM's 30-second sampling. Once the SLURM data was gathered for both semesters, we removed all the jobs without energy information as well as short-duration jobs (less than 10 minutes) and computed the mean power draw per node for each job.

The mean power draw per node is given by:

$$\text{Mean Power Draw per Node} = \frac{\text{Total Energy Consumption(J)}}{\text{Number of Nodes} \times \text{Time (s)}} \quad (1)$$

4.3.2 *Analysis.* We first plot the distribution of the amount of GPU hours consumed, based on power bins in Figure 2. In this figure, we can observe a clear shift in the mean power draw from the [1500-2000W] range to the [1000-1500W] range from the first semester to the second semester. This may indicate that many applications were impacted by the frequency capping and that this frequency capping had a significant impact on the power draw of those applications.

Due to limited information about the user for each job, we cannot be 100% sure that they did not change their workloads. However, we do not believe that the entire user base would change their workload at the exact same time and thus all shift to a lower power draw per node.

We also observe a small bump around [2200-2400W] that was not expected.

To gain more insights from these values, we took a deeper look at the four scientific fields that use the most GPU hours on the system. These fields include: Astrophysics and Geophysics, Plasma Physics, Molecular Simulation, and AI Research.

The values in terms of power draw compared to the GPU hours used for both semesters are reported in Figure 3. In these plots, we

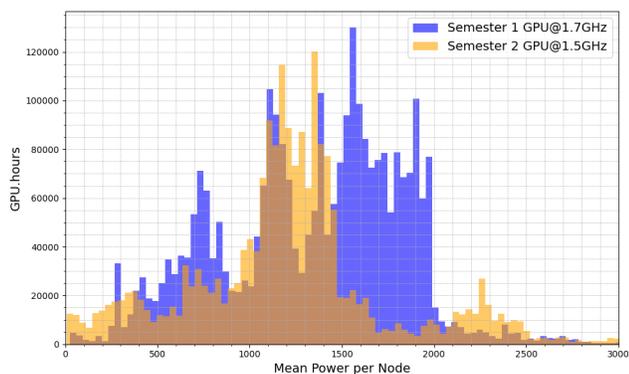


Figure 2: Distribution of power draw on the system for all the jobs of the machine

can observe that the main shifts are coming from Plasma Physics users, as well as Molecular Simulation. The results for Astrophysics and Geophysics are not clear and thus not relevant. However, we do notice that the AI community is using the GPUs more intensively than the others, with a mean power draw in the [2000-2500W] range in the second semester, corresponding to the bump observed in Figure 2. It seems that, unfortunately, the AI community struggled with using the system in the first semester, as the mean power draw was very low during this period (in the 700-800W range). This is not convenient for accurately assessing the impact of the frequency capping for this community. We discuss the possible impact and how to gain better insights into this specific community in the next section.

4.4 Impact of GPU Frequency Capping on AI Workloads

While the impact on AI applications could not be clearly observed in the 2024-S1 and 2024-S2 power draw comparison, as most of the power-hungry AI workloads were only run during the second semester of 2024, we decided to apply a frequency capping comparison on a wide set of AI workloads to assess the impact of this frequency change on a generic basis.

We compared the results on the MilaBench [1] set of AI benchmarks at 1.7GHz and 1.5GHz.

The results are reported in Table 2. Most of the workloads are not significantly impacted by the frequency reduction, and energy gains are generally observed. The mean energy gain is approximately 2%, with a performance loss of around 1%.

Given these results, we validate that this limited GPU frequency reduction has a minimal impact on commonly used AI workloads.

4.5 Impact of GPU Frequency at Full Accelerated Partition Level

4.5.1 Data collection. To present the impact of GPU capping at the full system level, we gathered accountability information retrieved by SLURM on a 10-minute basis. Over the year, we have a data point every 10 minutes with the instantaneous power consumption of the acceleration (MI250X GPUs) partition, as well as its current

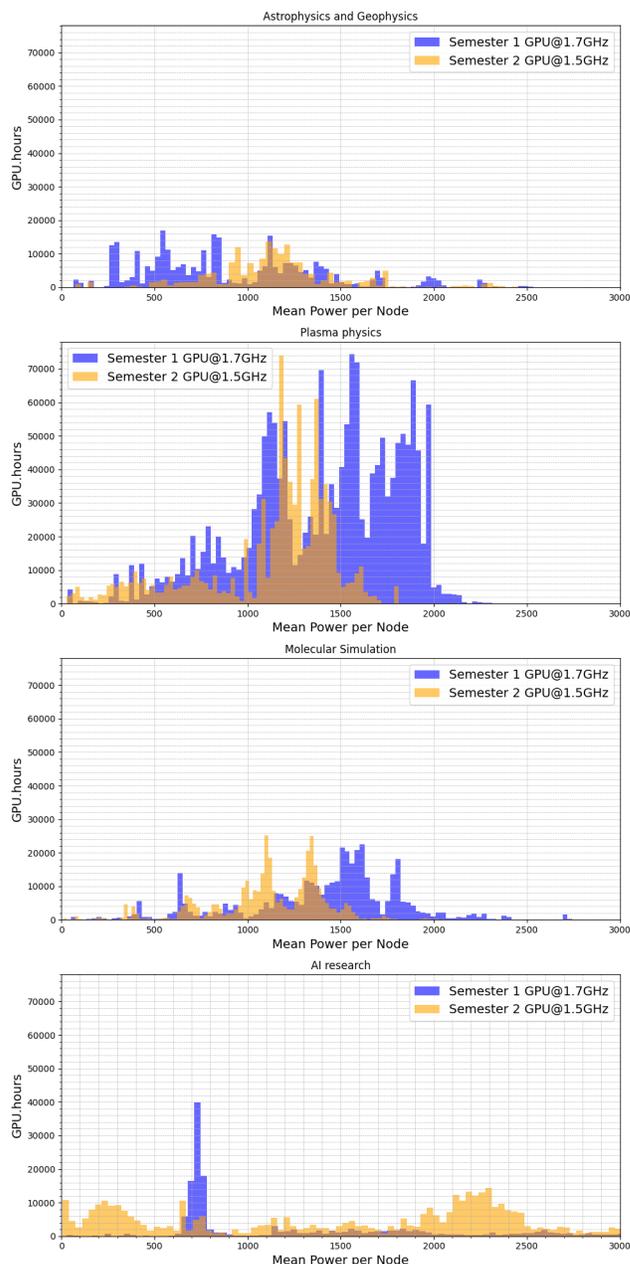


Figure 3: Distribution of power draw on the system for all the jobs of the machine

activity (a job running on the node makes it active). The idle nodes are taken into account for the power consumption of the partition.

Unfortunately, as the Adastra supercomputer is a hybrid machine with both CPU nodes and GPU nodes on the same rack, we cannot validate at the facility level that the power reported by SLURM is exactly correlated to the power used by the CINES infrastructure for those specific nodes. However, we are confident in these numbers based on the performance counters at the node level, but we

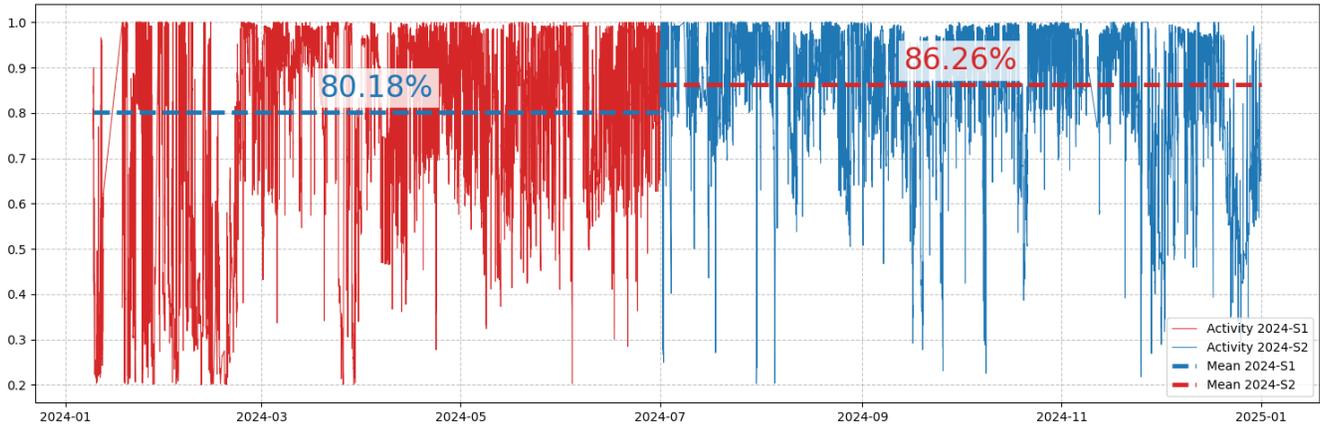


Figure 4: Activity of the MI250X partition in 2024

AI Benchmark	Time to solution	Energy to solution
bert-fp16	+1.13%	+3.27%
bert-fp32	-3.15%	-1.43%
bert-tf32	+1.17%	-1.76%
bert-tf32-fp16	+2.82%	+0.89%
dqn	+1.11%	-0.34%
lightning-gpus	+4.11%	-3.10%
pna	-0.86%	-4.35%
ppo	+4.35%	-15.07%
reformer	+0.28%	-4.57%
resnet152-ddp-gpus	-0.59%	+4.39%
t5	+1.70%	+2.66%
whisper	+4.21%	-5.93%
Average	+1.36%	-2.11%

Table 2: Impact of GPU frequency change from 1.7GHz to 1.5GHz on AI workloads

acknowledge that we miss some data due to our sampling rate of one data point every 10 minutes.

The power data presented here are only related to the nodes and do not take into account the network or file systems.

4.5.2 *Analysis.* This study is designed with three main goals:

- Retrieving the activity on the system in both semesters
- Retrieving the mean power draw of the system in both semesters
- Computing the mean power draw per active node on the system

We can observe in Figure 4 that the activity increased from 80% to 86% between the first and second semesters.

At the same time, the mean power draw of the partition dropped by more than 7%, from 552.75 kW to 513.16 kW, as shown in Figure 5. While the variability of power draw is significant in both semesters, we can clearly observe that the power peaks in the first semester, often higher than 800 kW and reaching up to 1 MW, are limited in

the second semester, with fewer spikes at 800 kW and no values above 900 kW.

In addition, the standard deviation of power draw between 2024-S1 and 2024-S2 decreased, with 22% less deviation in power draw since the 1.5GHz capping was applied at the GPU level. This information is particularly interesting for large infrastructures where power draw swings from large GPU systems can be an issue [6]. GPU frequency capping (or progressive GPU frequency capping) could be an interesting approach to better manage this problem.

Finally, when we compute the mean power draw per active node in the system, we estimate that we save up to 14% of power draw per node. Integrated over a period, this corresponds to 14% less energy consumed by a fully active system.

In our case, as the system was less active in the first semester, the energy gain for the second semester is limited to 7.2%.

The financial impact of such a change for our system is in the order of €100,000 reduction on our power bill.

MI250X Partition (356 nodes)	S1	S2	Gain
Active nodes	285.44	307.09	+7.6%
Mean Power Draw (kW)	552.75	513.16	-7.2%
Power Draw Standard Deviation (kW)	146.75	114.55	-22%
Mean Power per Node (kW)	1.94	1.67	-14%

Table 3: Summary of GPU Frequency Change from 1.7GHz to 1.5GHz impact on Adastra MI250X partition

4.6 Impact of GPU Frequency on Hardware Reliability

4.6.1 *Data collection.* We collaborated with HPE to obtain the hardware failure rates of the Adastra system for 2024. The main information reported by HPE is the number of component replacements, based on their intervention database.

We managed to gather the relevant information, specifically for the Bard Peak nodes (MI250X GPU nodes), and plot it in Figure 6.

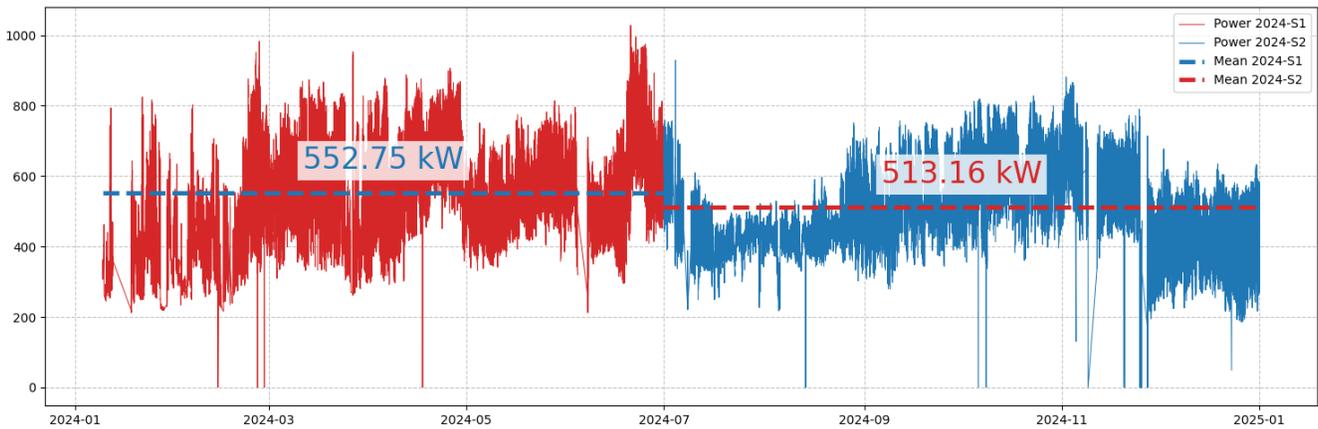


Figure 5: Power Draw of the MI250X Partition in 2024

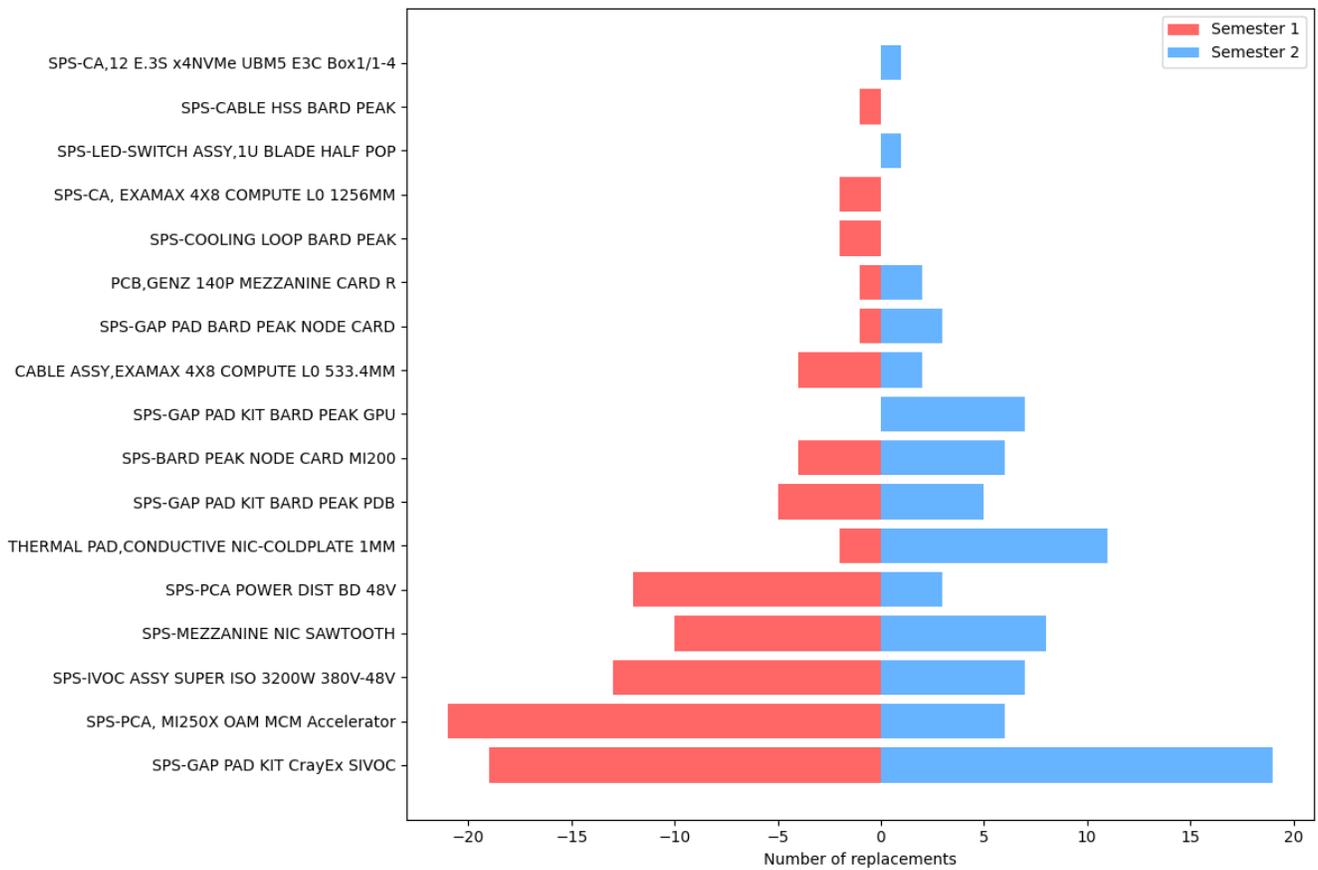


Figure 6: Parts replacement between S1 and S2

4.6.2 *Analysis.* One of the most interesting findings is the number of MI250X OAM MCM Accelerator (i.e., GPU card) replacements. In the first semester of 2024, 21 GPUs were replaced (out of the 1,424 GPUs in the system), while only 6 were replaced in the second

semester. We cannot directly conclude that the frequency reduction led to fewer GPU replacements, but it is still an interesting insight that could be further investigated. If other similar sites manage to gather data from similar systems, we will be able to make longer

comparisons as we continue to operate the system at 1.5GHz and will be able to assess the replacement rate over at least three more years.

4.7 User Satisfaction Impact

One last aspect we wanted to assess during this work was user satisfaction after the frequency policy change. It should be noted that the users were not aware of the frequency change until the end of 2024. In this context, we attempted to identify whether any common users between the first and second semesters of 2024 reported a performance impact on their applications.

Using multiple methods on our user tickets database, including simple classification and more complex LLM inference requests, we could not find any relevant information regarding this specific point.

The main takeaway is that we did not find any relevant information emphasizing performance degradation. However, we cannot definitively say that no one was impacted by this GPU frequency capping.

5 CONCLUSION

The study on the Adastr supercomputer has demonstrated the significant potential of GPU frequency capping in achieving energy efficiency in high-performance computing systems. By reducing the GPU frequency from 1.7 GHz to 1.5 GHz, we observed a notable reduction in power consumption without a substantial impact on computational performance. The analysis of user data and job metrics revealed that the user base remained consistent across the two semesters, validating the relevance of our findings. Furthermore, the reduction in power draw and the associated energy savings highlight the financial and environmental benefits of such optimizations.

However, the study also underscored the need for further investigations into specific areas, such as hardware stability and user satisfaction, to fully understand the nuanced impacts of frequency capping.

In conclusion, this research underscores the importance of dynamic power management techniques in HPC systems. The insights gained from this study can inform future strategies for sustainable and efficient computing, paving the way for more environmentally conscious HPC operations. As we continue to monitor the Adastr system, we aim to gather more data and refine our understanding of the long-term effects of GPU frequency scaling on performance and reliability.

6 DISCUSSIONS AND FUTURE WORK

The findings of this study underscore the potential of GPU frequency capping as an effective strategy for enhancing energy efficiency in HPC environments. Reducing the GPU frequency from 1.7 GHz to 1.5 GHz on the Adastr supercomputer resulted in a significant decrease in power consumption without compromising overall computational performance. This outcome aligns with the broader goals of sustainability and cost-efficiency in HPC operations.

However, several areas require further exploration to fully understand the implications of this approach. One interesting avenue to consider is implementing dynamic frequency adjustments, where

the frequency could be changed on a weekly basis and the impact measured accordingly. This approach could provide a more representative view of the user base and their current workloads, offering deeper insights into the effects of frequency capping on diverse computational needs.

Additionally, while the study observed a reduction in GPU replacements following the frequency adjustment, further investigation is needed to establish a direct causal link. Longitudinal studies across multiple sites and over extended periods could provide more robust insights into the impact of frequency capping on hardware reliability.

Moreover, although no significant user complaints were identified, a survey is scheduled for 2025 to gather feedback from our user base regarding this change. The survey will also address system stability and could leverage findings related to job topology, such as failed or canceled jobs, and hardware replacement across both semesters.

The Adastr supercomputer provides a compelling case study, but replicating these results in diverse HPC environments is crucial for validating the generalization of the approach. Collaborations with other HPC centers could facilitate broader data collection and comparative analyses.

Finally, CINES will evaluate the feasibility of applying similar frequency capping strategies to its new MI300A partition to assess the impact on this technology. By addressing these areas, we can gain a more comprehensive understanding of the benefits and trade-offs associated with GPU frequency capping, ultimately contributing to more sustainable and efficient HPC operations.

ACKNOWLEDGMENTS

We thank the users and administrators of Adastr for their valuable feedback, as well as all the CINES team who helped us gather those data. We also thank Mistral AI for its assistance in correcting the text and generating plotting scripts.

REFERENCES

- [1] Pierre Delaunay, Xavier Bouthillier, Olivier Breuleux, Satya Ortiz-Gagné, Olexa Bilaniuk, Fabrice Normandin, Arnaud Bergeron, Bruno Carrez, Guillaume Alain, Soline Blanc, Frédéric Osterrath, Joseph Viviano, Roger Creus-Castanyer, Darshan Patil, Rabiul Awal, and Le Zhang. 2024. Introducing Milabench: Benchmarking Accelerators for AI. *arXiv preprint arXiv:2411.11940* (2024). <https://arxiv.org/abs/2411.11940v1>
- [2] Gabriel Hautreux and Etienne Malaboeuf. 2023. Reducing HPC energy footprint for large scale GPU accelerated workloads. In *Proceedings of the SC '23 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis*. ACM, 1860–1865. <https://doi.org/10.1145/3624062.3624268>
- [3] Adrian Jackson, Alan Simpson, and Andrew Turner. 2023. Emissions and energy efficiency on large-scale high performance computing facilities: ARCHER2 UK national supercomputing service case study. *arXiv:2309.05440* [cs.DC] <https://arxiv.org/abs/2309.05440>
- [4] Ahmad Maroof Karimi, Matthias Maiterth, Woong Shin, Naw Safrin Sattar, Hao Lu, and Feiyi Wang. 2024. Exploring the Frontiers of Energy Efficiency using Power Management at System Scale. In *SC24*. 1835–1844.
- [5] SchedMD LLC. 2024. Slurm Workload Manager. <https://github.com/SchedMD/slurm> Accessed: 2024-04-03.
- [6] Patki Tapasya, Natalie Bates, Rish Ghatikar, Anders Clausen, Sonja Klingert, Ghaleb Abdulla, and Mehdi Sheikhalishahi. 2016. Supercomputing Centers and Electricity Service Providers: A Geographically Distributed Perspective on Demand Management in Europe and the United States. 243–260. https://doi.org/10.1007/978-3-319-41321-1_13