Understanding the Efficacy of Power Profiles: A Case Study of AMD Instinct MI100 GPU

Ghazanfar Ali
Texas Tech University
Lubbock, TX, USA
ghazanfar.ali@ttu.edu

Mert Side
Texas Tech University
Lubbock, TX, USA
mert.side@ttu.edu

Sridutt Bhalachandra
Lawrence Berkeley National Laboratory
Berkeley, CA, USA
sriduttb@cs.unc.edu

Tommy Dang
Texas Tech University
Lubbock, TX, USA
tommy.dang@ttu.edu

Alan Sill
Texas Tech University
Lubbock, TX, USA
alan.sill@ttu.edu

Yong Chen
Texas Tech University
Lubbock, TX, USA
yong.chen@ttu.edu

***Abstract*—Graphics Processing Units (GPUs) have become pivotal for modern high-performance computing (HPC) and artificial intelligence workloads due to their substantial computational prowess. However, this computational prowess comes at a cost, as GPUs consume vast amounts of power, presenting a challenge for high-end computing systems, including those aimed at achieving exascale computing capabilities. In response to the power efficiency problem, modern GPUs typically offer the ability to adjust clock frequencies and cap power consumption. However, the AMD Instinct MI100 takes a unique approach by introducing a set of predefined power profiles that internally manipulate clock frequencies to manage power. This study evaluates the effectiveness of these power profiles through a comparative analysis of various power and performance metrics. It indicates that, for most of the selected workloads and during significant portions of their execution, the GPU consumes power exceeding its specified Thermal Design Power (TDP). For instance, the GROMACS workload exceeded its TDP by one-third during almost half of its execution time. Furthermore, the study notes a significant increase in temperature reaching as high as 80$^{\circ}C$. Moreover, DGEMM and STREAM workloads exhibit similar power consumption patterns, suggesting that the underlying power management scheme does not adapt power allocation based on the computational intensity of the workload. Thus, the study demonstrates that changing the power profile does not significantly impact crucial metrics such as performance, clock frequency, voltage, GPU utilization, or temperature. In summary, this research sheds light on the power dynamics of the AMD Instinct MI100 GPU, emphasizing the challenges associated with power, performance, and thermal management in HPC environments. The findings underscore the importance of fine-tuning power management strategies to enhance energy efficiency while maintaining optimal performance in GPUs.**

***Index Terms*—Survey, Power profiles, Energy, Power consumption, GPU, HPC**

I. Introduction

The exponential performance increase at constant cost and power for the computing industry has slowed as we reach the end of Moore's Law [1], [2], [3]. The trajectory of computational progress in the foreseeable future appears to be increasingly reliant on accelerators. While graphics processing units (GPUs) have historically demonstrated impressive computational capabilities, their power consumption has consistently increased with each successive generation. For instance, the Frontier supercomputer, the first exaflop system at the Oak Ridge National Laboratory, uses over 20 MW in power and consists of 37,888 AMD MI250X GPUs [4]. Each node in the Frontier system is configured with one CPU and four GPUs. These GPUs, which have a thermal design power (TDP) of 560 W [5], consume approximately 80% of the total node power. This highlights the necessity for robust power management on GPUs. Furthermore, the fifth fastest system on the Top500 list, LUMI, also uses the MI250X GPUs. Accordingly, we attempt to study the power management of the MI100 GPU, a predecessor to MI250X.

Motivation: We observe that the manufacturer's TDP limits are exceeded over significant durations. To scale power, GPUs provide different power controls, such as dynamic voltage frequency scaling and power profiles. However, the efficacies and impacts of these controls, especially the effect of GPU's power profiles on the workloads, are not well known. In this study, we attempt to investigate the following questions: (1) Is TDP rating a reliable metric to estimate the power budget of a node? (2) What impact do the GPU power profiles have on GPU and workload parameters?

Key insights and contributions: This study makes the following key contributions:

- *In-depth analysis of GPU power management:* We cover a broad range of details with the help of several classes of workloads to provide researchers and architects with a fundamental understanding of the MI100 power management. This is important for designing future solutions that aim to improve the energy efficiency of the GPU.
- *Evaluation of the supported power profiles:* We evaluated the impact of the MI100 GPU power profiles on GPU utilization metrics. We empirically observed that changing the power profile did not noticeably affect power consumption, performance, frequency, voltage, GPU utilization, GPU temperature, TDP exceeding, and magnitude of the TDP exceeding. Furthermore, we provided workload-

specific analyses of these behaviors.

The source code and collected data will be made publicly available. The rest of the paper is organized as follows. Section II describes the experimental setup, including applications and GPU used in this study. Section III evaluates the effectiveness of the GPU power profiles. Section IV describes the related work and Section V highlights the key takeaways and provides the conclusions.

II. Experimental Setup

Our study conducted data collection, analysis, and evaluation on an AMD MI100 GPU within the ChameleonCloud testbed [6], running Linux Ubuntu 20.04, utilizing `ROCm 5.4` for workload deployment and `rocm-smi` for power profile management and metric collection. To ensure data integrity, all tests were conducted with exclusive node access. This diverse set of workloads enabled comprehensive testing of the GPU's computational and memory capabilities. The experimental setup included an AMD EPYC 7763 CPU and an AMD Instinct MI100 GPU. Table I lists the configuration of the AMD MI100 used in this study. Our investigation encompassed

Table I: Specifications of the AMD Instinct MI100 used in this study [7].

Specification	Description
GPU Frequency Range (MHz)	Up to 16 configurations [300:1502]
Memory Frequency	1200 MHz
TDP	290 W
GPU Memory (HBM2)	32 GB
Peak Memory Bandwidth	Up to 1228.8 GB/s

nine GPU-accelerated workloads: (1) GROMACS: Molecular dynamics simulations to study biochemical molecules, specifically a lysozyme solution in water. (2) LAMMPS: Particle simulations for various materials, using a Leanard-Jones 3D melt experiment. (3) NAMD: Simulations of biomolecular systems, using the Apolipoprotein A1 dataset with 92,224 atoms. (4) SPECFEM3D Cartesian: Simulations for wave propagation across various mediums. (5) DGEMM: Compute-intensive matrix multiplication with 25600x25600 matrices. (6) STREAM: Memory-intensive workload, employing a Triad kernel with 655,360,000 elements. (7) BERT: Natural language processing model training on the IMDb dataset. (8) ResNet: Computer vision modeling trained on the CIFAR10 dataset. (9) LSTM: TensorFlow-based sentiment classification on movie reviews, with 25,000 reviews for training and testing.

III. Power Profiles Evaluation

The previous section evaluated the behavior of the default power profile chosen at boot (`auto`) w.r.t. TDP enforcement and energy-proportionality. This section investigates whether the power profiles help reduce TDP violations or improve energy proportionality. Further, we evaluate the impact of the MI100 GPU power profiles on GPU utilization, HBM utilization, execution time, frequency, voltage, and thermal conditions using nine workloads listed in Table II.

Figure 1 illustrates the critical components involved in (1) and (4) changing the profile, (2) executing the workload,

Table II: List of applications used in this study.

Category	Applications
HPC	`GROMACS` [8], `LAMMPS` [9], `NAMD` [10], `SPECFEM3D` [11]
Machine Learning	`BERT` [12], `ResNet50` [13], `LSTM` [14]
Benchmarks	`DGEMM` [15], `STREAM` [16]

(3) collecting the utilization metrics, and (5) evaluating and analyzing the impacts of power profiles. We collected metrics at default, i.e., `auto` and four pre-defined power profiles supported by AMD MI100 GPU – *video, compute, power saving, and bootup default*. The desired power profile was enforced using AMD `rocm-smi` [17]. The workload was executed, as the metrics were collected at a sampling interval of 250 ms. This sampling interval was chosen to keep the overhead of the collection low while getting samples with statistical significance. These steps were repeated three times for a given power profile to mitigate the run-to-run variations. The following metrics `power_usage`, `voltage`, `edge_temperature`, `junction_temperature`, `memory_temperature`, `sclk`, `gpu_usage`, `memory_usage`, `time`, `FLOPS/s`, and `memory_bandwidth` were collected.

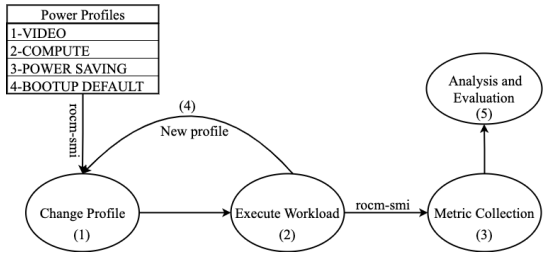

Figure 1: Overview of the methodology to understand the efficacies of the AMD MI100 GPU power profiles.

A. GPU and Memory Usage

Figure 2 shows the impact of power profiles on GPU usage. Each power profile showed approximately 100% GPU usage at peak for each workload. GPU usage merely indicates how busy the GPU is and is agnostic of the *computational intensity* of the kernel activity. For example, DGEMM and STREAM reported the same (i.e., 100%).

Memory usage refers to the amount of used memory in terms of allocations. We observed similar GPU memory usage patterns for each power profile, as shown in Figure 3. Unlike GPU usage, memory usage showed variation for different workloads. As a memory-intensive kernel, STREAM showed the highest memory usage ($\sim 80\%$). LAMMPS and SPECFEM3D also showed higher memory usage ($\sim 75\%$). All other workloads showed $\sim< 50\%$ except LSTM, which used the lowest memory ($\sim< 5\%$). Overall, these observations indicate that power profiles have no significant impact on GPU and memory usage for a given workload.

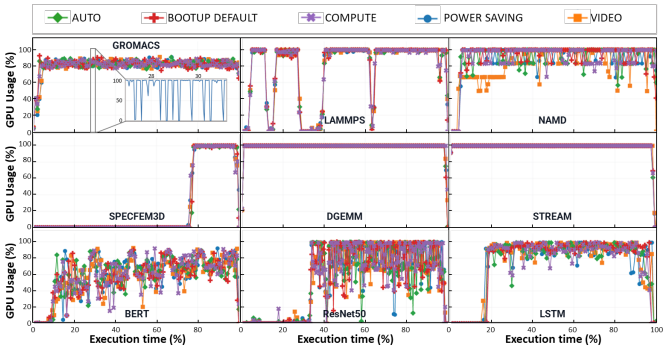

Figure 2: Impact of MI100 power profiles on GPU usage.

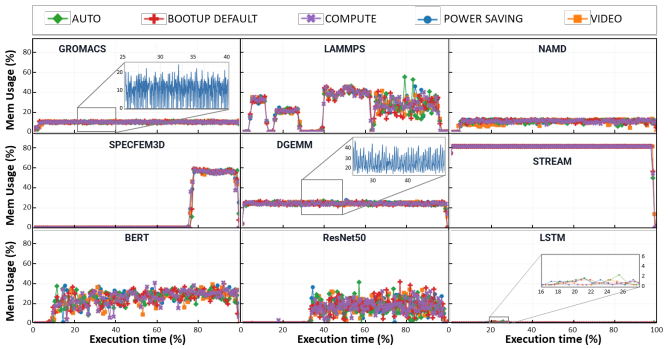

Figure 3: Impact of MI100 power profiles on GPU memory usage.

B. Performance – Time, GFLOPS/s, and Bandwidth

We evaluated the impact of power profiles on key performance metrics – execution time, GFLOPS/s, and memory bandwidth. Table III compares the impact of power profiles on the execution time (in seconds) of different workloads. Similar to GPU and memory usage, none of the profiles impacted the execution time.

Figure 4 shows the GFLOPS/s using DGEMM (left) and GPU memory bandwidth (GB/s) using STREAM (right) for each GPU power profile. DGEMM and STREAM achieved more than 80% of their performance in terms of FLOPS/s and bandwidth. While a maximum variation of 38 GFLOPS/s was observed across the different profiles, the variation is very close to the run-to-run variation and, therefore, insignificant. STREAM's bandwidth across profiles remains unchanged.

C. GPU Frequency

AMD MI100 supports 16 GPU frequencies ranging from 300 MHz to 1502 MHz. This GPU supports a single memory frequency of 1200 MHz. These frequencies are immutable to users; however, power profiles are meant to internally control these frequencies based on workload activity. Figure 5 shows the execution timeline for the workloads with each power profile. Power profiles do not significantly impact GPU frequency variations for a given workload. Generally, each profile likely sets the GPU frequency to high when executing instructions and low when waiting on memory or

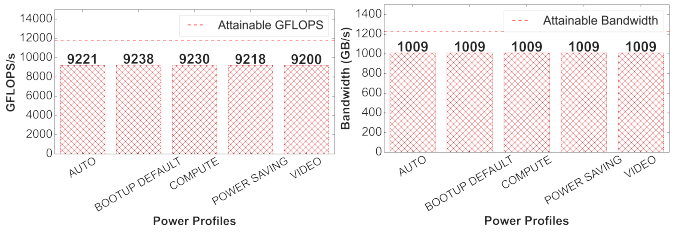

Figure 4: The figure on the left illustrates the GFLOPS per second for all the power profiles using DGEMM. The figure on the right illustrates the GPU memory bandwidth (GB/s) for all the power profiles using STREAM.

idling. More specifically, we observed different patterns for the following workloads: (1) hybrid workloads (GROMACS, LAMMPS, NAMD, SPECFEM3D), (2) machine learning workloads (ResNet50, BERT, LSTM), and (3) GPU-only workloads (DGEMM, STREAM).

For hybrid workloads, the CPU periodically offloads a chunk of work to the GPU, which causes a change in GPU frequency from the lowest frequency to higher frequencies. After execution of the chunk of compute, the GPU clock is scaled to the idle frequency (lowest frequency). This pattern continues during the entire execution of workloads. The size of the offloaded chunk of compute varies from workload to workload. For example, GROMACS offloads comparatively small sizes (∼1 to 2 seconds) of the compute; therefore, there are a lot of back-and-forth switches between the lowest frequency to the higher frequencies. However, the chunk of compute size for LAMMPS and NAMD is comparatively large, so GPU runs on higher frequencies for a longer duration. SPECFEM3D execution consists of three key steps: `mesh generation`, `database creation`, and `solver computation`. The first two steps are performed on the CPU; therefore, the GPU runs idle for a long duration. The `solver computation` is performed on GPU as a single chunk of compute.

For machine learning workloads, models are trained using one or more `epochs`. Each `epoch` is generally executed on higher frequencies. BERT and ResNet50 involve many `epochs`; therefore, frequent switches between the lowest and higher frequencies occur. However, a single `epoch` is sufficient for training the LSTM workload. That is why LSTM is executed at a constant higher frequency.

For GPU-only workloads, the code and data are transferred to GPU memory before execution. After transferring the code and data, these workloads are continuously executed on higher frequency ranges. At the end of executions of these workloads, the results are transferred to CPU memory space.

We observed that peak operating frequencies depend on the workloads' computational intensity. The higher the computational intensity of a workload, the lower the GPU peak operating frequency. For example, DGEMM and SPECFEM3D are comparatively more compute-intensive, and their operating frequencies are lower than other workloads. This behavior is corroborated by Intel CPUs reducing their frequencies

Table III: Execution time (seconds) of workloads for each MI100 GPU power profile.

Application \ Profile	COMPUTE	POWER SAVING	BOOTUP DEFAULT	VIDEO	AUTO
LAMMPS	14	14	14	14	14
NAMD	78.7	78.7	78.9	78.7	78.3
GROMACS	112.7	112.1	112.8	112.5	112.4
SPECFEM3D	180	180	179.9	179.9	180.1
ResNet50	63.9	64.2	63.2	63.8	63.2
LSTM	30	29.2	29.4	30.4	29.4
BERT	277.6	278.1	279.2	277.2	279.2
DGEMM	727.2	728.1	726.5	729.5	727.9
STREAM	467.6	467.6	467.6	467.6	467.6

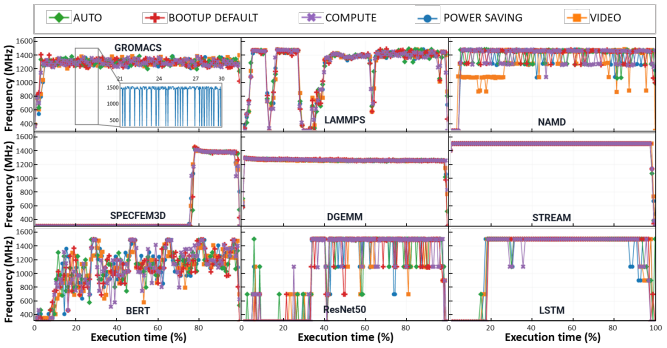

Figure 5: Impact of power profiles on GPU frequency changes during the execution of an application.

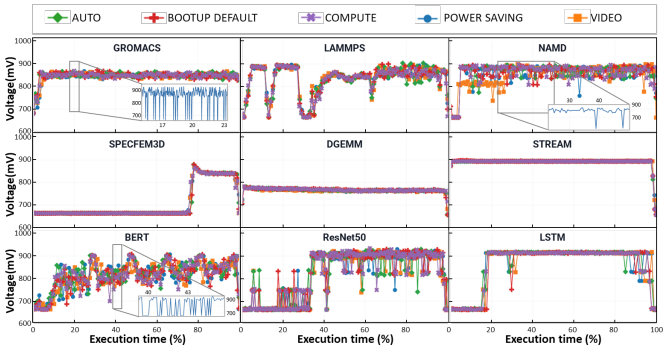

Figure 6: Impact of power profiles on GPU voltage during the entire execution of each workload.

when AVX-heavy instruction threatens to exceed the power limits [18].

D. GPU Voltage

One distinction of AMD GPUs is that they expose voltage. This metric is unavailable in NVIDIA's high-end GPUs such as GP100, GV100, and GA100. We observed that each power profile similarly impacts voltage for a workload. We also observed that voltage follows the footprint of the GPU clock frequency, as discussed above. In other words, a change in frequency causes a change in GPU voltage. Figure 6 shows the impact of power profiles on GPU voltage during the entire execution of each workload. The power profiles generally do not significantly impact GPU voltage variations for a given workload.

Overall, voltage fluctuated in the range of 656 - 956 millivolts (mV) across all workloads. The operating voltage of each workload is generally commensurate with the GPU frequency and the computational intensity of a workload. It is pertinent to mention that the voltage level can be reduced up to 50% for a compute-intensive workload and involves significant memory usage. The higher the computational intensity and memory use, the lower the GPU operating voltage. For example, DGEMM and SPECFEM3D can use 50% and 75% of the maximum available voltage levels, but why? There are two key factors behind restricting voltage for compute-intensive workloads. First, compute-intensive workloads stress all computing resources to the maximum level, which can cause significant power. According to Ohm's Law ($P = \frac{V^2}{R}$), voltage plays a critical role in controlling power usage, given the resistance is constant. Second, GPU compute, and GPU high-bandwidth memory (HBM) blocks share the same voltage configuration; therefore, a change in voltage can impact power consumption by compute and HBM (1200 MHZ) blocks simultaneously. Thus, GPU voltage is adjusted to a comparatively lower voltage for compute-intensive tasks and significant memory usage to keep power consumption under the thermal budget.

E. GPU Junction, HBM, and Edge Temperatures

Figure 7 shows the impact of the power profiles on GPU junction, HBM, and edge temperatures in degrees Celsius ($^{\circ}C$) for each workload. We observed that each power profile similarly impacts thermal conditions for a given workload. The edge temperature is always lower than junction and memory temperatures, regardless of the workload nature. The compute-intensive workloads, such as DGEMM, GROMACS, LAMMPS, and NAMD, cause a significant increase in junction temperature. For DGEMM, we observed an increase in junction temperature up to $\sim 76^{\circ}C$. On the other hand, memory-intensive workloads, such as STREAM and SPECFEM3D, cause a significant increase in memory temperature. For STREAM, we observed an increase in memory temperature up to $\sim 80^{\circ}C$. While DGEMM and STREAM caused a significant increase in junction and memory temperatures, respectively, we did not observe any thermal throttling.

F. Power Consumption

Understanding the impact of power profiles on power consumption behavior is one of our key objectives. In Figure 8,

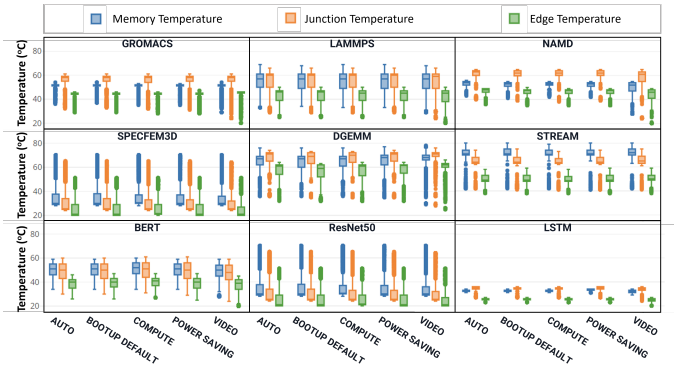

Figure 7: Impact of the power profiles on GPU junction, memory, and edge temperatures ($^{\circ}C$) for each workload.

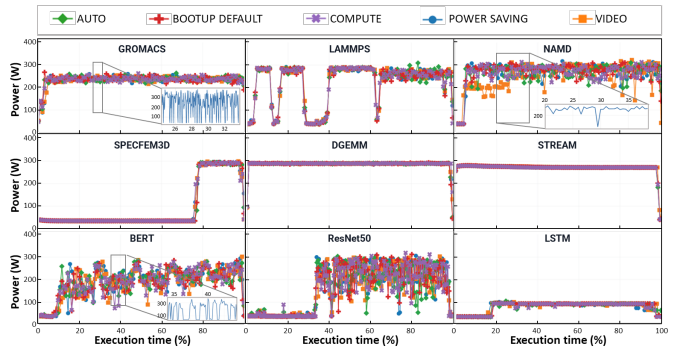

Figure 8: Impact of MI100 power profiles on the power consumption of each workload.

we found that power profiles uniformly affect power for each workload, determined by GPU thermal conditions, computational intensity, operating frequency, and voltage. Notably, GPU thermal conditions play a role in power reduction when junction temperatures exceed thermal throttling thresholds. However, none of our workloads triggered thermal throttling, thus consuming power to its maximum potential within the GPU's thermal budget [19].

We observed distinct trends in power management for different workloads: compute-intensive workloads (e.g., DGEMM, SPECFEM3D) were allocated lower frequency and voltage to minimize TDP breaches. In contrast, hybrid and memory-intensive workloads (e.g., GROMACS, LAMMPS, NAMD, STREAM) operated at higher frequencies and voltages, leading to varied rates of TDP exceedance. Specifically, dense compute workloads were managed with lower frequency and voltage to keep power in check. In contrast, hybrid and memory-intensive workloads did not consistently exceed TDP despite higher settings due to their computational demands. The MI100 GPU's frequency and voltage selection are inversely related to workload computational intensity—lower for compute-heavy tasks and higher for less intensive tasks.

1) TDP Violation Magnitude: Power consumption exceeding the manufacturer's TDP limit refers to power usage beyond the manufacturer's power maximum limit (TDP violation). We calculate this metric using the ratio of the maximum power value of the workload to the TDP of the GPU. Figure 9 shows the power consumption exceeding the manufacturer's TDP limit for seven workloads across MI100 GPU power profiles. We observed that the power consumption of workloads with low computational intensity, such as STREAM and LSTM, did not exceed the TDP limit. Comparing power exceeding the manufacturer's TDP limit across different power profiles shows that AUTO caused lower power exceeding the TDP limit for most workloads. GROMACS's peak power exceeded the TDP limit by 30%. The peak power consumption of the cluster must be designed by considering the workload's peak power consumption. The node utilized in our study was equipped with a single GPU. However, in typical HPC environments featuring more GPUs per node, this could place significant stress on the power supply infrastructure.

2) TDP Violation Frequency: Figure 10 shows the frequency of TDP violations during the entire run of seven workloads across MI100 power profiles. The power profiles showed a similar count of power consumption exceeding the TDP for most workloads. We observed that HPC workloads exceeded the TDP limit by over $\sim 20\%$. The ML workloads exceeded the TDP limit comparatively for a shorter duration, not more than $\sim 10\%$.

IV. Related Work

Numerous studies have delved into the performance analysis of the MI100 GPU. Melesse et al. [20] initially explored the porting of three workloads onto the AMD MI100 within the Summit supercomputer environment, albeit without a comprehensive understanding of the power and energy consumption associated with these workloads on the GPU. In a similar vein, Dufek et al. [21] conducted a study focusing on workload portability across NVIDIA GA100, AMD MI100, and Intel Gen9 GPUs; however, they omitted an in-depth characterization of power consumption.

Most modern GPUs support the DVFS mechanism and pre-designed power profiles to manage power, performance, and energy behaviors [22], [23]. Kang et al. [24] introduced a mechanism leveraging DVFS for energy conservation, but the applicability of their findings primarily extends to scenarios involving data movement over network interfaces. On a different note, Allen et al. [25] extensively examined the impact of DVFS on GPU memory but concentrated their research within the domain of memory-intensive applications. In contrast, the MI100 GPU diverges by adopting power profiles restricting direct GPU frequency adjustments. These profiles autonomously manipulate GPU frequency and voltage, likely in response to workload computational intensity. As demonstrated in this study, these profiles appear to fall short of achieving their intended effects. Jin et al. [26] presented distinct power and performance analyses employing an integer sum reduction kernel. However, while insightful, their work is limited in its coverage of GPU power consumption behaviors, workload diversity, and power profiles' influence.

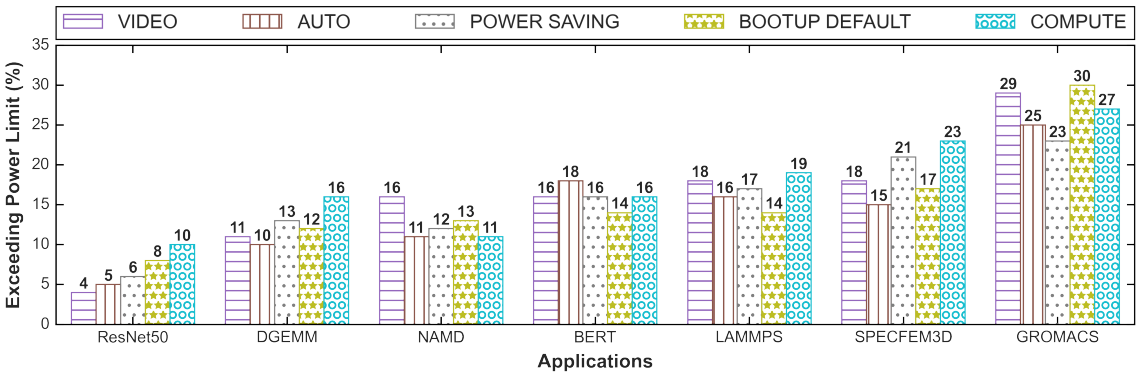

Figure 9: The magnitude of TDP violations for seven workloads across MI100 power profiles.

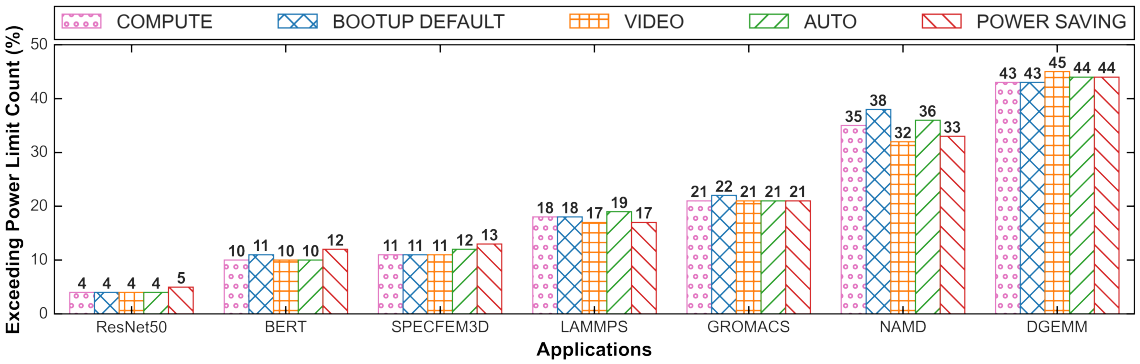

Figure 10: The frequency of TDP violations during the entire run of seven workloads across MI100 power profiles.

To the best of our knowledge, this is the first study that explores the impact of MI100's power profiles on various utilization metrics across a diverse range of workloads encompassing High-Performance Computing, Machine Learning, and microbenchmarks. Furthermore, our study provides workload-specific analyses, delving into topics such as TDP violations, frequency/voltage behaviors, and thermal conditions, collectively contributing to a comprehensive understanding of power management in MI100 GPUs.

V. Conclusions and Future Work

This paper underscores the importance of GPU power management techniques to devise strategies to mitigate escalating power demands and enhance energy efficiency. Through an empirical analysis, we characterize the power consumption patterns of the AMD MI100 GPU across a diverse set of real-world workloads, encompassing four HPC applications, three ML applications, and two benchmarks emphasizing compute and memory intensiveness. Our investigation focuses on the efficacy of the GPU's power management framework in adhering to TDP constraints and examines how workloads' computational characteristics influence power allocation. Additionally, we assess available power profiles to tailor power consumption to user requirements.

Our findings reveal several critical insights: (1) All evaluated power profiles demonstrate comparable power, performance, and utilization metrics across varied workloads, indicating a lack of adaptability in current power profiles for dynamic power control and energy efficiency enhancement. (2) TDP breaches are prevalent under GPU workload conditions, with many workloads surpassing the TDP threshold for significant portions of their execution time. For instance, the GROMACS workload exceeded its TDP limit by approximately 45% for over 20% of its runtime, posing substantial implications for data center design and operational strategies. This necessitates overprovisioning beyond the GPUs' collective TDP to ensure operational safety. (3) Our analysis extends to the frequency and voltage behavior under different workloads, noting a 50% reduction in both parameters for compute-intensive tasks. (4) The STREAM benchmark observed a marked increase in memory temperature, approximately 80°C, which raises concerns regarding the reliability of memory devices and warrants further exploration.

This study aims to furnish researchers and system architects with a foundational comprehension of MI100 power management practices. The insights derived underscore the probable need for power and infrastructure overprovisioning in data centers to manage frequent TDP violations. Future work will extend this inquiry to the power, performance, and thermal dynamics of subsequent AMD GPU architectures to develop a more comprehensive understanding of GPU power management across different architectures.

REFERENCES

[1] C. E. Leiserson, N. C. Thompson, J. S. Emer, B. C. Kuszmaul, B. W. Lampson, D. Sanchez, and T. B. Schardl, "There's plenty of room at the top: What will drive computer performance after moore's law?" *Science*, vol. 368, no. 6495, p. eaam9744, 2020.

[2] J. Shalf, "The future of computing beyond moore's law," *Philosophical Transactions of the Royal Society A*, vol. 378, no. 2166, p. 20190061, 2020.

[3] J. L. Hennessy and D. A. Patterson, *Computer architecture: a quantitative approach*. Elsevier, 2011.

[4] TOP500.org, "Top500, June 2024 Ranking," https://www.top500.org/lists/top500/2024/06/, 2022.

[5] A. M. Devices, "AMDI MI250x," https://www.amd.com/en/products/accelerators/instinct/mi200/mi250x.html, (Accessed on 7/6/2024).

[6] K. Keahey, J. Anderson, Z. Zhen, P. Riteau, P. Ruth, D. Stanzione, M. Cevik, J. Colleran, and Gunawi, "Lessons learned from the chameleon testbed," in *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC '20)*. USENIX Association, July 2020.

[7] "AMD MI100 GPU datasheet," https://www.amd.com/en/products/server-accelerators/instinct-mi100, accessed: 2023-01-19.

[8] B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl, "GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation," *Journal of Chemical Theory and Computation*, vol. 4, no. 3, pp. 435–447, 2008.

[9] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen *et al.*, "Lammps - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales," *Computer Physics Communications*, p. 108171, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010465521002836

[10] J. C. Phillips, D. J. Hardy, J. D. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang *et al.*, "Scalable molecular dynamics on CPU and GPU architectures with NAMD," *The Journal of Chemical Physics*, vol. 153, no. 4, p. 044130, 2020.

[11] D. Komatitsch, J. Vilotte, J. Tromp, J. Ampuero, K. Bai, P. Basini, C. Blitz, E. Bozdag, E. Casarotti, J. Charles *et al.*, "Specfem3d cartesian [software]," 2019.

[12] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, 2019, pp. 4171–4186.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] TensorFlow, "Long Short-Term Memory layer - Hochreiter 1997," https://www.tensorflow.org/api_docs/python/tf/keras/layers/LSTM, 2021.

[15] NVIDIA Corporation. (2013) Cuda samples. [Online]. Available: https://docs.nvidia.com/cuda/cuda-samples/index.html\#matrix-multiplication--cublas-

[16] T. Deakin, J. Price, M. Martineau, and S. McIntosh-Smith, "Gpu-stream v2. 0: Benchmarking the achievable memory bandwidth of many-core processors across diverse parallel programming models," in *International Conference on High Performance Computing*. Springer, 2016, pp. 489–507.

[17] AMD, "Rocm system management interface (rocm smi) library," https://github.com/RadeonOpenCompute/rocm_smi_lib, 2023.

[18] Cloudflare, "On the dangers of intel's frequency scaling," https://blog.cloudflare.com/on-the-dangers-of-intels-frequency-scaling, 2017.

[19] R. Regner, "An analytical approach to quantify the thermal budget in consideration of consecutive thermal process steps," in *10th IEEE International Conference of Advanced Thermal Processing of Semiconductors*, 2002, pp. 15–20.

[20] V. Melesse Vergara, R. Budiardja, and W. Joubert, "Early experiences evaluating the hpe/cray ecosystem for amd gpus," Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), Tech. Rep., 2021.

[21] A. S. Dufek, R. Gayatri, N. Mehta, D. Doerfler, B. Cook, Y. Ghadar, and C. DeTar, "Case study of using kokkos and sycl as performance-portable frameworks for milc-dslash benchmark on nvidia, amd and intel gpus," in *2021 International Workshop on Performance, Portability and Productivity in HPC (P3HPC)*, 2021, pp. 57–67.

[22] G. Ali, M. Side, S. Bhalachandra, N. J. Wright, and Y. Chen, "An automated and portable method for selecting an optimal gpu frequency," *Future Generation Computer Systems*, vol. 149, pp. 71–88, 2023.

[23] ——, "Performance-aware energy-efficient gpu frequency selection using dnn-based models," in *Proceedings of the 52nd International Conference on Parallel Processing*, 2023, pp. 433–442.

[24] K.-D. Kang, G. Park, H. Kim, M. Alian, N. S. Kim, and D. Kim, "Nmap: Power management based on network packet processing mode transition for latency-critical workloads," in *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*, ser. MICRO '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 143–154. [Online]. Available: https://doi.org/10.1145/3466752.3480098

[25] T. Allen and R. Ge, "Characterizing power and performance of gpu memory access," in *2016 4th International Workshop on Energy Efficient Supercomputing (E2SC)*, 2016, pp. 46–53.

[26] Z. Jin, J. Vetter, and J. Vetter, "A study on atomics-based integer sum reduction in hip on amd gpu," in *Workshop Proceedings of the 51st International Conference on Parallel Processing*, 2022, pp. 1–10.