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

IEEE HPEC 2024

Ghazanfar Ali [1], Mert Side [1], Sridutt Bhalachandra [2], Tommy Dang [1], Alan Sill [1], Yong Chen [1]

[1]Texas Tech University [2]Lawrence Berkeley National Laboratory

September 25th, 2024

Table of Contents

Introduction

- The exponential performance increase in computing is slowing as Moore's Law reaches its limits. Therefore, future computational capabilities are expected to rely heavily on accelerators like GPUs.

Introduction

- The exponential performance increase in computing is slowing as Moore's Law reaches its limits. Therefore, future computational capabilities are expected to rely heavily on accelerators like GPUs.
- GPU power consumption has significantly increased with each new generation.

Introduction

- The exponential performance increase in computing is slowing as Moore's Law reaches its limits. Therefore, future computational capabilities are expected to rely heavily on accelerators like GPUs.
- GPU power consumption has significantly increased with each new generation.
- The Frontier supercomputer, with AMD MI250X GPUs, uses over 20 MW of power, with GPUs accounting for 80% of node power.

Introduction

- The exponential performance increase in computing is slowing as Moore's Law reaches its limits. Therefore, future computational capabilities are expected to rely heavily on accelerators like GPUs.
- GPU power consumption has significantly increased with each new generation.
- The Frontier supercomputer, with AMD MI250X GPUs, uses over 20 MW of power, with GPUs accounting for 80% of node power.
- Effective GPU power management is crucial for large systems like Frontier and LUMI.

Introduction

- The exponential performance increase in computing is slowing as Moore's Law reaches its limits. Therefore, future computational capabilities are expected to rely heavily on accelerators like GPUs.
- GPU power consumption has significantly increased with each new generation.
- The Frontier supercomputer, with AMD MI250X GPUs, uses over 20 MW of power, with GPUs accounting for 80% of node power.
- Effective GPU power management is crucial for large systems like Frontier and LUMI.
- The effectiveness of GPU power controls, such as power profiles, is not well understood.

Motivation

In this study, we investigate the following questions:

Motivation

In this study, we investigate the following questions:

1

Is TDP rating a reliable metric for estimating the power budget of a node?

Motivation

In this study, we investigate the following questions:

1

Is TDP rating a reliable metric for estimating the power budget of a node?

2

What impact do the GPU power profiles have on GPU and workload parameters?

Contributions

This study provides the following key insights:

Contributions

This study provides the following key insights:

In-depth analysis of GPU power management:

We provide a comprehensive analysis using various workloads to give researchers and architects a foundational understanding of MI100 power management, which is crucial for future energy-efficient GPU designs.

Contributions

This study provides the following key insights:

In-depth analysis of GPU power management:

We provide a comprehensive analysis using various workloads to give researchers and architects a foundational understanding of MI100 power management, which is crucial for future energy-efficient GPU designs.

Evaluation of the supported power profiles:

We evaluated MI100 GPU power profiles and found that altering the profile had little effect on key metrics such as power consumption, performance, and temperature. Additionally, workload-specific insights of these behaviors were provided.

Experimental Setup

- The study was conducted on an AMD MI100 GPU within the ChameleonCloud testbed, running Linux Ubuntu 20.04.

Table 1: Specifications of the AMD Instinct MI100 used in this study.

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

Experimental Setup

- The study was conducted on an AMD MI100 GPU within the ChameleonCloud testbed, running Linux Ubuntu 20.04.
- `rocm-smi` was used for power profile management and metric collection.

Table 1: Specifications of the AMD Instinct MI100 used in this study.

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

Experimental Setup

- The study was conducted on an AMD MI100 GPU within the ChameleonCloud testbed, running Linux Ubuntu 20.04.
- `rocm-smi` was used for power profile management and metric collection.
- Exclusive node access was ensured for data integrity.

Table 1: Specifications of the AMD Instinct MI100 used in this study.

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

Experimental Setup

- The study was conducted on an AMD MI100 GPU within the ChameleonCloud testbed, running Linux Ubuntu 20.04.
- `rocm-smi` was used for power profile management and metric collection.
- Exclusive node access was ensured for data integrity.
- The experimental setup included an AMD EPYC 7763 CPU and an AMD Instinct MI100 GPU.

Table 1: Specifications of the AMD Instinct MI100 used in this study.

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

Experimental Setup *(cont'd)*

Table 2: List of applications used in this study.

Category	Applications
HPC	GROMACS, LAMMPS , NAMD, SPECFEM3D
Machine Learning	BERT, ResNet50, LSTM
Benchmarks	DGEMM, STREAM

- A diverse set of workloads tested the GPU's computational and memory capabilities.

Overview of Methodology

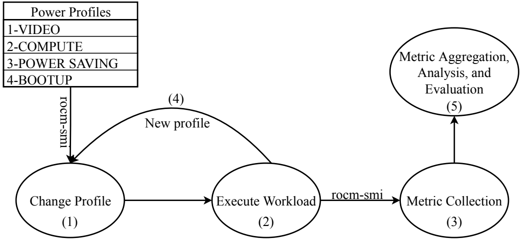

Figure 1: Overview of the methodology to understand the efficacies of the AMD MI100 GPU power profiles.

Overview of Methodology

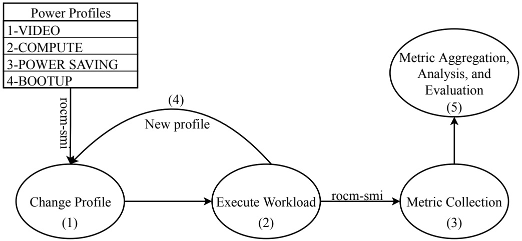

Figure 1: Overview of the methodology to understand the efficacies of the AMD MI100 GPU power profiles.

- Metrics include power usage, voltage, temperatures, clock speeds, GPU, FLOPS, memory usage, and bandwidth.
- Sampled every 250 ms to balance overhead and statistical significance.
- Collected for the default and pre-defined power profiles: `video`, `compute`, `power saving`, and `bootup default`.
- Each profile was tested three times to reduce run-to-run variations.

Performance: Time, GFLOPS/s, and Bandwidth

Table 3: Execution time (seconds) of workloads for each MI100 GPU power 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

- Power profiles had no noticeable impact on key performance metrics, including execution time, GFLOPS/s, and memory bandwidth.
- Execution times for various workloads remained consistent across all power profiles, as shown in Table 3.

Performance: Time, GFLOPS/s, and Bandwidth

Table 3: Execution time (seconds) of workloads for each MI100 GPU power 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

- Power profiles had no noticeable impact on key performance metrics, including execution time, GFLOPS/s, and memory bandwidth.
- Execution times for various workloads remained consistent across all power profiles, as shown in Table 3.

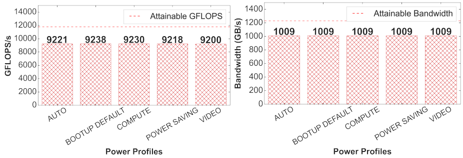

Figure 2: Left illustrates the GFLOPS per second for all the power profiles using DGEMM. Right illustrates the GPU memory bandwidth (GB/s) for all the power profiles using STREAM.

- DGEMM and STREAM achieved over 80% of their peak performance in terms of FLOPS/s and bandwidth, respectively.
- The variation in FLOPS/s across profiles was minimal (38 GFLOPS/s) and close to run-to-run variation, making it insignificant, while STREAM's bandwidth remained unchanged across profiles.

GPU Frequency

- AMD MI100 supports GPU frequencies from 300 MHz to 1502 MHz and a single memory frequency of 1200 MHz, which are controlled internally by power profiles based on workload activity.

GPU Frequency

- AMD MI100 supports GPU frequencies from 300 MHz to 1502 MHz and a single memory frequency of 1200 MHz, which are controlled internally by power profiles based on workload activity.
- Power profiles do not significantly affect GPU frequency variations during workload execution.

GPU Frequency

- AMD MI100 supports GPU frequencies from 300 MHz to 1502 MHz and a single memory frequency of 1200 MHz, which are controlled internally by power profiles based on workload activity.
- Power profiles do not significantly affect GPU frequency variations during workload execution.
- For **hybrid workloads** (GROMACS, LAMMPS, NAMD, SPECFEM3D), GPU frequency fluctuates as the CPU offloads chunks of work to the GPU, leading to frequent switches between low and high frequencies.

GPU Frequency

- AMD MI100 supports GPU frequencies from 300 MHz to 1502 MHz and a single memory frequency of 1200 MHz, which are controlled internally by power profiles based on workload activity.
- Power profiles do not significantly affect GPU frequency variations during workload execution.
- For **hybrid workloads** (GROMACS, LAMMPS, NAMD, SPECFEM3D), GPU frequency fluctuates as the CPU offloads chunks of work to the GPU, leading to frequent switches between low and high frequencies.
- For **machine learning workloads** (ResNet50, BERT, LSTM), the frequency generally remains high during each epoch, with LSTM running at a constant high frequency due to needing only one epoch.

GPU Frequency

- AMD MI100 supports GPU frequencies from 300 MHz to 1502 MHz and a single memory frequency of 1200 MHz, which are controlled internally by power profiles based on workload activity.
- Power profiles do not significantly affect GPU frequency variations during workload execution.
- For **hybrid workloads** (GROMACS, LAMMPS, NAMD, SPECFEM3D), GPU frequency fluctuates as the CPU offloads chunks of work to the GPU, leading to frequent switches between low and high frequencies.
- For **machine learning workloads** (ResNet50, BERT, LSTM), the frequency generally remains high during each epoch, with LSTM running at a constant high frequency due to needing only one epoch.
- **GPU-only workloads** (DGEMM, STREAM) run continuously at higher frequencies after code and data are transferred to GPU memory.

GPU Frequency

- AMD MI100 supports GPU frequencies from 300 MHz to 1502 MHz and a single memory frequency of 1200 MHz, which are controlled internally by power profiles based on workload activity.
- Power profiles do not significantly affect GPU frequency variations during workload execution.
- For **hybrid workloads** (GROMACS, LAMMPS, NAMD, SPECFEM3D), GPU frequency fluctuates as the CPU offloads chunks of work to the GPU, leading to frequent switches between low and high frequencies.
- For **machine learning workloads** (ResNet50, BERT, LSTM), the frequency generally remains high during each epoch, with LSTM running at a constant high frequency due to needing only one epoch.
- **GPU-only workloads** (DGEMM, STREAM) run continuously at higher frequencies after code and data are transferred to GPU memory.
- Peak operating frequencies are inversely related to computational intensity, with more compute-intensive workloads like DGEMM and SPECFEM3D running at lower frequencies.

GPU Junction, HBM, and Edge Temperatures

- Power profiles had a similar impact on GPU junction, HBM (memory), and edge temperatures across all workloads.
- Edge temperatures were consistently lower than junction and memory temperatures.
- Compute-intensive workloads like DGEMM, GROMACS, LAMMPS, and NAMD led to a significant rise in junction temperatures, with DGEMM reaching up to 76°C.
- Memory-intensive workloads, such as STREAM and SPECFEM3D, caused memory temperatures to increase, with STREAM reaching 80°C.

GPU Junction, HBM, and Edge Temperatures

- Power profiles had a similar impact on GPU junction, HBM (memory), and edge temperatures across all workloads.
- Edge temperatures were consistently lower than junction and memory temperatures.
- Compute-intensive workloads like DGEMM, GROMACS, LAMMPS, and NAMD led to a significant rise in junction temperatures, with DGEMM reaching up to 76°C.
- Memory-intensive workloads, such as STREAM and SPECFEM3D, caused memory temperatures to increase, with STREAM reaching 80°C.

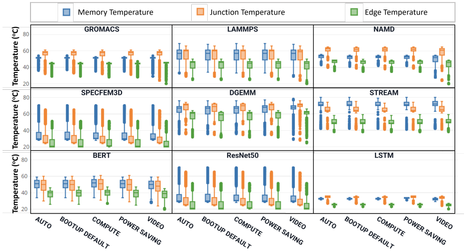

Figure 3: Impact of the power profiles on GPU junction, memory, and edge temperatures ($^\circ C$) for each workload.

- Despite the temperature increases, no thermal throttling was observed during the executions.

Power Consumption

- Power profiles affected power consumption similarly across all workloads, influenced by thermal conditions, computational intensity, frequency, and voltage.

Power Consumption

- Power profiles affected power consumption similarly across all workloads, influenced by thermal conditions, computational intensity, frequency, and voltage.
- No thermal throttling was observed, allowing workloads to use power up to the GPU's thermal budget.

Power Consumption

- Power profiles affected power consumption similarly across all workloads, influenced by thermal conditions, computational intensity, frequency, and voltage.

- No thermal throttling was observed, allowing workloads to use power up to the GPU's thermal budget.

- **Compute-intensive workloads** used lower frequency and voltage to prevent exceeding the TDP, while **hybrid and memory-intensive workloads** operated at higher frequencies and voltages.

Power Consumption

- Power profiles affected power consumption similarly across all workloads, influenced by thermal conditions, computational intensity, frequency, and voltage.
- No thermal throttling was observed, allowing workloads to use power up to the GPU's thermal budget.
- **Compute-intensive workloads** used lower frequency and voltage to prevent exceeding the TDP, while **hybrid and memory-intensive workloads** operated at higher frequencies and voltages.
- These trends highlight that MI100 power management is inversely proportional to workload intensity, with more compute-heavy tasks operating at lower frequencies and voltages to manage power usage.

Power Consumption *(cont'd)*

TDP Violation Magnitude:

Some workloads exceeded the manufacturer's TDP limit, with GROMACS exceeding the TDP by 30%. Low-intensity workloads like STREAM and LSTM stayed within the TDP limit. AUTO profile typically resulted in fewer TDP violations.

Power Consumption *(cont'd)*

TDP Violation Magnitude:

Some workloads exceeded the manufacturer's TDP limit, with GROMACS exceeding the TDP by 30%. Low-intensity workloads like STREAM and LSTM stayed within the TDP limit. AUTO profile typically resulted in fewer TDP violations.

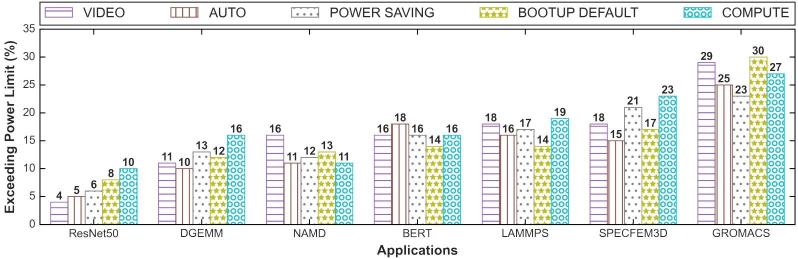

Figure 4: The magnitude of TDP violations for workloads across MI100 power profiles.

Power Consumption *(cont'd)*

TDP Violation Frequency:

HPC workloads frequently exceeded TDP (20%), while ML workloads had fewer violations (10%).

Power Consumption *(cont'd)*

TDP Violation Frequency:

HPC workloads frequently exceeded TDP (20%), while ML workloads had fewer violations (10%).

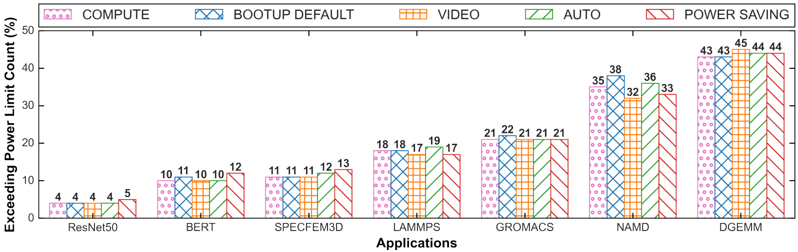

Figure 5: The frequency of TDP violations during the run of workloads across MI100 power profiles.

Summary of Findings

- Our key findings include the following:
  1. Power profiles offer similar power, performance, and utilization metrics across workloads, showing limited adaptability for dynamic power control.

Summary of Findings

- Our key findings include the following:
  1. Power profiles offer similar power, performance, and utilization metrics across workloads, showing limited adaptability for dynamic power control.
  2. TDP breaches were common, with some workloads, like GROMACS, exceeding the TDP by 45% for over 20% of runtime, suggesting a need for overprovisioning in data center designs.

Summary of Findings

- Our key findings include the following:
  1. Power profiles offer similar power, performance, and utilization metrics across workloads, showing limited adaptability for dynamic power control.
  2. TDP breaches were common, with some workloads, like GROMACS, exceeding the TDP by 45% for over 20% of runtime, suggesting a need for overprovisioning in data center designs.
  3. Compute-intensive tasks saw frequency and voltage reductions of up to 50%.

Summary of Findings

- Our key findings include the following:
  1. Power profiles offer similar power, performance, and utilization metrics across workloads, showing limited adaptability for dynamic power control.
  2. TDP breaches were common, with some workloads, like GROMACS, exceeding the TDP by 45% for over 20% of runtime, suggesting a need for overprovisioning in data center designs.
  3. Compute-intensive tasks saw frequency and voltage reductions of up to 50%.
  4. Memory-intensive tasks like STREAM experienced significant temperature increases (80°C), raising concerns about memory reliability.

Conclusion

- This study highlights the importance of GPU power management in addressing rising power demands and improving energy efficiency in HPC environments.

Conclusion

- This study highlights the importance of GPU power management in addressing rising power demands and improving energy efficiency in HPC environments.
- We analyzed the power consumption patterns of the AMD MI100 GPU across various real-world workloads, focusing on how power management adheres to TDP limits and is influenced by computational characteristics.

Conclusion

- This study highlights the importance of GPU power management in addressing rising power demands and improving energy efficiency in HPC environments.
- We analyzed the power consumption patterns of the AMD MI100 GPU across various real-world workloads, focusing on how power management adheres to TDP limits and is influenced by computational characteristics.
- Future work will explore the power, performance, and thermal behaviors of newer AMD GPU architectures to broaden understanding of GPU power management practices across different platforms.

Conclusion

- This study highlights the importance of GPU power management in addressing rising power demands and improving energy efficiency in HPC environments.
- We analyzed the power consumption patterns of the AMD MI100 GPU across various real-world workloads, focusing on how power management adheres to TDP limits and is influenced by computational characteristics.
- Future work will explore the power, performance, and thermal behaviors of newer AMD GPU architectures to broaden understanding of GPU power management practices across different platforms.

Thank you! Let us know if you have any questions?

E-mail: `mert.side@ttu.edu`