2026 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS) | 979-8-3315-5103-2/26/$31.00 ©2026 IEEE | DOI: 10.1109/ISPASS69572.2026.00073

Power Ranger: A Comprehensive Framework for GPU Power Characterization

Allison Seigler, Zhixing Jiang and Lizy Kurian John
University of Texas at Austin
aseigler@utexas.edu, zx.jiang@utexas.edu, ljohn@ece.utexas.edu

***Abstract*—As GPUs have grown more popular and become commonplace in datacenters, GPU power consumption has become a more important metric than ever before. Power behavior should be a major consideration in all stages of GPU design, development, and deployment. However, accurate power evaluation is difficult because modern-day GPU workloads are large and complicated, and rapid advances in popular datacenter workloads means that the power profiles of these workloads change often. To solve this problem, we introduce Power Ranger, a GPU power benchmark generator that can replicate power behavior from any arbitrary workload using only a reference device and a trace of power over time. Power Ranger's generated workloads replicate real workload behavior using simple kernel launches selected from a look-up table. Because Power Ranger only requires a power trace and device, and does not require any code from the target workload, it can also create real GPU workloads that match any desired power behavior. This capability enables the generation of custom power benchmarks and stress tests that can be executed on real GPU hardware. We evaluate Power Ranger using several real and synthetic workloads, and we find that Power Ranger is able to reproduce original power traces with a MAPE of 5% or less for all cases.**

***Index Terms*—GPU, Power, Energy, Benchmarking, DVFS**

I. Introduction

GPUs have become crucial for efficient datacenter computation due to the growing popularity of highly parallelizable workloads such as machine learning (ML). However, power budgeting is now a major constraint as datacenters proliferate on existing power grids. In addition, GPU idle and active power levels increase with each new architecture generation [3] [4] [5]. Because GPU architectures and workload characteristics develop rapidly, adaptable and flexible GPU power evaluation methodologies are needed, especially at the device level.

To this end, we introduce Power Ranger, a framework for generating simple GPU workloads that can match the power output of any arbitrary target workload. The generated workloads consist only of a schedule of kernel launches, and each kernel consists of simple assembly-level instructions, eliminating the overhead of complicated runtime frameworks that are commonplace in real GPU workloads. Power Ranger addresses the gap in GPU device-level power benchmarking by providing a framework for custom benchmark generation. Using Power Ranger, GPU hardware engineers can evaluate device power behavior early in the design process. Power Ranger's generated benchmarks are portable and easy to simulate in RTL. Power Ranger also provides infrastructure for creating custom power stress tests, enabling thorough evaluation of device power characteristics at the pre-silicon phase as well as device and system level testing of power mitigation techniques such as DVFS.

II. Background

As cloud computing workloads such as ML evolve, GPU datacenters require more power, and GPU power profile characteristics change often [2] [13] [8]. Several methods of mitigating this power consumption have been proposed, at both the system and device level [9] [16] [12] [10] [15]. To aid in the development of these methods, accurate power characterization and benchmarking are needed. Prior work in GPU benchmarking has focused on the systems level [11] [18], and device-level evaluation is primarily limited to ML models and stressors [1] [7] [14]. To the best of our knowledge, no work for generating individual GPU power benchmarks yet exists. We therefore develop Power Ranger, a flexible method for enhancing device-level power evaluation.

III. The Power Ranger Framework

The Power Ranger framework consists of two main components: a kernel look-up table (LUT) that maps power levels to executable kernels, and a workload generation module that constructs workloads matching target power traces.

A. LUT Generation

The kernel LUT provides the foundation for proxy generation by mapping low-level kernels to their power consumption levels. We generate the LUT by sweeping runtime variables across microbenchmark kernels to cover the GPU's full power range. These variables, including thread count, active warp proportion, and input sparsity, significantly affect power consumption. We retain only kernels with power standard deviation below 3W to ensure reliable, steady power levels.

B. Power Ranger Workload Generation

Power Ranger constructs executable workloads that replicate target power traces by sequentially selecting and validating kernels from the LUT. As illustrated in Figure 1, the framework processes the input trace incrementally, building the replicated workload one sample at a time.

For each timestep, the framework extracts target power and temperature values, then uses binary search to efficiently locate a matching candidate kernel in the LUT. The candidate kernel

979-8-3315-5103-2/26/$31.00 ©2026 IEEE
DOI 10.1109/ISPASS69572.2026.00073

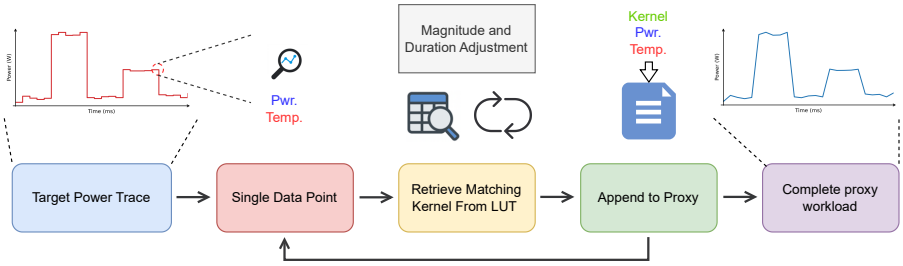
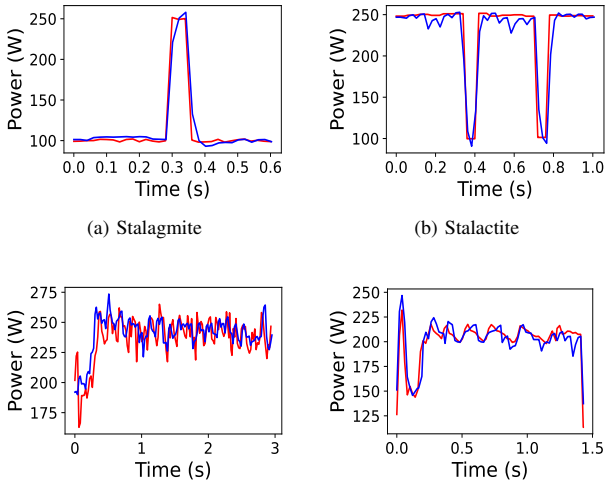


Fig. 1. Framework used for workload generation. For each data point, power and temperature of the input trace are recorded. The power and temperature are fed to the kernel LUT, which outputs a corresponding kernel that will generate this power behavior. The kernel is appended to the new workload.

is executed on the target device, and its actual power output is measured and compared against the target value. If measured power matches the target within acceptable tolerance, the kernel is appended to the workload schedule. Otherwise, the binary search adjusts the kernel selection and repeats the measurement process. This iterative refinement continues until a kernel produces stable, accurate power output matching the target. The final output is a complete program ready for compilation and execution on the target device.

IV. Evaluation



(a) Stalagmite (b) Stalactite

(c) Deepseek Inference (d) Gemma Inference

Fig. 2. Comparison of power traces from our generated replication workloads (blue) with the target trace used as input for various workloads (red).

TABLE I
STATISTIC COMPARISON OF REPLICATED WORKLOADS TO THEIR TARGETS.

Benchmark	MAPE %	DTW	Z-score DTW
Synth_stalagmite	3.76	46.46	1.03
Synth_stalactite	4.19	98.07	1.97
Deepseek_VLLM	4.44	96.42	5.79
Gemma_VLLM	4.91	69.19	3.014

We demonstrate Power Ranger's accuracy by generating replications for several workloads and comparing the power trace generated by the replication workload against the original power trace used as input, which we refer to as the target power trace. Our real target workloads and generated replication workloads are run on an NVIDIA A100 GPU. We use an AMD EPYC 7742 CPU running Ubuntu 22.04.5 LTS as our runtime environment. For our real workloads, we collect power traces of single-batch inferences on Deepseek [6] and Google Gemma [17] models. We also generate several synthetic traces, which are created artificially instead of generated on a real GPU. Power Ranger is able to create a **real** GPU workload that matches a **synthetic** power trace, allowing for custom GPU power stress tests.

Figure 2 presents a visual comparison between the target traces and the traces generated using our replicated workloads. In our synthetic trace replication experiments, our Z-score DTW distance is less than 2, indicating a close match between the target and replication trace. Z-score DTW distance of our real workloads is slightly higher, around 3-5. In all cases, our MAPE is below 5%. The metrics in all cases show close correlation between the target traces and our replication traces.

V. Conclusion

Power Ranger provides a flexible framework for replicating GPU power behavior from arbitrary workloads using only power traces and target hardware. By generating simple kernel-based workloads that accurately match target power profiles, Power Ranger enables comprehensive power evaluation without requiring access to original workload code. This capability supports critical use cases including pre-silicon validation, DVFS testing, and custom stress test generation for evolving datacenter workloads.

VI. Acknowledgement

This research was supported in part by the Semiconductor Research Corporation (SRC) Task 3281.001. Any opinions, findings, conclusions, or recommendations are those of the authors and not of the funding agencies. We would like to thank Tawfik Rahal-Arabi and Mehdi Sadi for their valuable insights and discussions during this research.

REFERENCES

[1] G. Ali, M. Side, S. Bhalachandra, N. J. Wright, and Y. Chen, "Performance-aware energy-efficient gpu frequency selection using dnn-based models," in *Proceedings of the 52nd International Conference on Parallel Processing*, ser. ICPP '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 433–442. [Online]. Available: https://doi.org/10.1145/3605573.3605600

[2] R. A. Bridges, N. Imam, and T. M. Mintz, "Understanding gpu power: A survey of profiling, modeling, and simulation methods," *ACM Comput. Surv.*, vol. 49, no. 3, Sep. 2016. [Online]. Available: https://doi.org/10.1145/2962131

[3] N. Corporation, *NVIDIA A100 Tensor Core GPU*, 2022, device components; Interconnect components; Device ooperation specifications. [Online]. Available: https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-nvidia-us-2188504-web.pdf

[4] ——, *NVIDIA H100 PCIe GPU Product Brief*, 2022, device components; Interconnect components; Device ooperation specifications. [Online]. Available: https://www.nvidia.com/content/dam/en-zz/Solutions/gtcs22/data-center/h100/PB-11133-001_v01.pdf

[5] ——, *NVIDIA GH200 Grace Hopper Superchip Benchmark Step-by-Step Guide*, 2025, benchmarking instructions; Provided benchmarks. [Online]. Available: https://docs.nvidia.com/gh200-superchip-benchmark-guide.pdf

[6] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, and Z. Zhang, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2501.12948

[7] J. Guerreiro, A. Ilic, N. Roma, and P. Tomas, "Gpgpu power modeling for multi-domain voltage-frequency scaling," in *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2018, pp. 789–800.

[8] Q. Hu, Z. Ye, Z. Wang, G. Wang, M. Zhang, Q. Chen, P. Sun, D. Lin, X. Wang, Y. Luo, Y. Wen, and T. Zhang, "Characterization of large language model development in the datacenter," 2024. [Online]. Available: https://arxiv.org/abs/2403.07648

[9] A. K. Kakolyris, D. Masouros, P. Vavaroutsos, S. Xydis, and D. Soudris, "throttll'em: Predictive gpu throttling for energy efficient llm inference serving," in *2025 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2025, pp. 1363–1378.

[10] A. K. Kakolyris, D. Masouros, S. Xydis, and D. Soudris, "Slo-aware gpu dvfs for energy-efficient llm inference serving," *IEEE Computer Architecture Letters*, vol. 23, no. 2, pp. 150–153, 2024.

[11] K.-D. Lange, "Identifying shades of green: The specpower benchmarks," *Computer*, vol. 42, no. 3, pp. 95–97, 2009.

[12] S. Li, X. Wang, X. Zhang, V. Kontorinis, S. Kodakara, D. Lo, and P. Ranganathan, "Thunderbolt: Throughput-Optimized, Quality-of-Service-Aware power capping at scale," in *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)*. USENIX Association, Nov. 2020, pp. 1241–1255. [Online]. Available: https://www.usenix.org/conference/osdi20/presentation/li-shaohong

[13] P. Patel, E. Choukse, C. Zhang, I. n. Goiri, B. Warrier, N. Mahalingam, and R. Bianchini, "Characterizing power management opportunities for llms in the cloud," in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, ser. ASPLOS '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 207–222. [Online]. Available: https://doi.org/10.1145/3620666.3651329

[14] Y. Shan, Y. Yang, X. Qian, and Z. Yu, "Guser: A gpgpu power stressmark generator," in *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2024, pp. 1111–1124.

[15] J. Stojkovic, P. A. Misra, I. n. Goiri, S. Whitlock, E. Choukse, M. Das, C. Bansal, J. Lee, Z. Sun, H. Qiu, R. Zimmermann, S. Samal, B. Warrier, A. Raniwala, and R. Bianchini, "Smartoclock: Workload- and risk-aware overclocking in the cloud," in *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, 2024, pp. 437–451.

[16] J. Stojkovic, C. Zhang, I. n. Goiri, J. Torrellas, and E. Choukse, "Dynamollm: Designing llm inference clusters for performance and energy efficiency," in *2025 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2025, pp. 1348–1362.

[17] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, P. G. Sessa, A. Chowdhery, A. Roberts, A. Barua, A. Botev, A. Castro-Ros, A. Slone, A. Héliou, A. Tacchetti, A. Bulanova, A. Paterson, B. Tsai, B. Shahriari, C. L. Lan, C. A. Choquette-Choo, C. Crepy, D. Cer, D. Ippolito, D. Reid, E. Buchatskaya, E. Ni, E. Noland, G. Yan, G. Tucker, G.-C. Muraru, G. Rozhdestvenskiy, H. Michalewski, I. Tenney, I. Grishchenko, J. Austin, J. Keeling, J. Labanowski, J.-B. Lespiau, J. Stanway, J. Brennan, J. Chen, J. Ferret, J. Chiu, J. Mao-Jones, K. Lee, K. Yu, K. Millican, L. L. Sjoesund, L. Lee, L. Dixon, M. Reid, M. Mikuła, M. Wirth, M. Sharman, N. Chinaev, N. Thain, O. Bachem, O. Chang, O. Wahltinez, P. Bailey, P. Michel, P. Yotov, R. Chaabouni, R. Comanescu, R. Jana, R. Anil, R. McIlroy, R. Liu, R. Mullins, S. L. Smith, S. Borgeaud, S. Girgin, S. Douglas, S. Pandya, S. Shakeri, S. De, T. Klimenko, T. Hennigan, V. Feinberg, W. Stokowiec, Y. hui Chen, Z. Ahmed, Z. Gong, T. Warkentin, L. Peran, M. Giang, C. Farabet, O. Vinyals, J. Dean, K. Kavukcuoglu, D. Hassabis, Z. Ghahramani, D. Eck, J. Barral, F. Pereira, E. Collins, A. Joulin, N. Fiedel, E. Senter, A. Andreev, and K. Kenealy, "Gemma: Open models based on gemini research and technology," 2024. [Online]. Available: https://arxiv.org/abs/2403.08295

[18] A. Tschand, A. T. R. Rajan, S. Idgunji, A. Ghosh, J. Holleman, C. Kiraly, P. Ambalkar, R. Borkar, R. Chukka, T. Cockrell, O. Curtis, G. Fursin, M. Hodak, H. Kassa, A. Lokhmotov, D. Miskovic, Y. Pan, M. P. Manmathan, L. Raymond, T. S. John, A. Suresh, R. Taubitz, S. Zhan, S. Wasson, D. Kanter, and V. J. Reddi, "Mlperf power: Benchmarking the energy efficiency of machine learning systems from watts to mwatts for sustainable ai," in *2025 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2025, pp. 1201–1216.