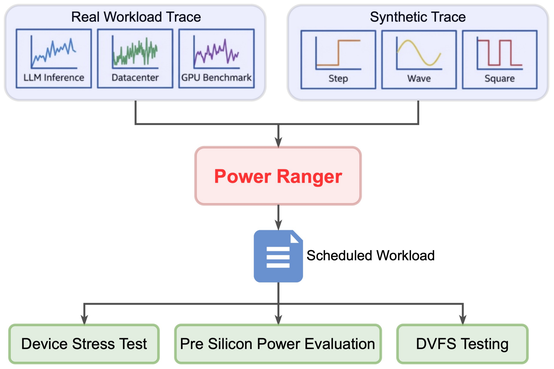
Power Ranger: A Comprehensive Framework for GPU Power Characterization

Allison Seigler, Zhixing Jiang and Lizy K. John
University of Texas at Austin

Introduction

- High demand for GPUs has driven increased interest in power consumption concerns:
  - Unsustainable energy consumption of large workloads
  - High frequency, high magnitude power fluctuations cause soft errors and reduce hardware lifetimes
- Accurate power evaluation for GPUs and GPU clusters is crucial
- **Problem:** design phase power modelling is difficult because large software stacks prevent easy replication of workloads
- **Solution:** Power Ranger, a framework for generating workloads the replicate power behavior



Power Ranger can:

1. Create workloads with a desired power profile given only a target profile and reference device as input
2. Reproduce both real and synthetic power profiles with a high degree of accuracy, enabling the modeling of both current and emerging workloads
3. Create lightweight workloads with minimal software overhead that can be run on RTL models
4. Benchmark GPU power behavior in unusual runtime situations using custom synthetic traces

The Power Ranger Framework

Look-up table (LUT) generated with a variety of power values:

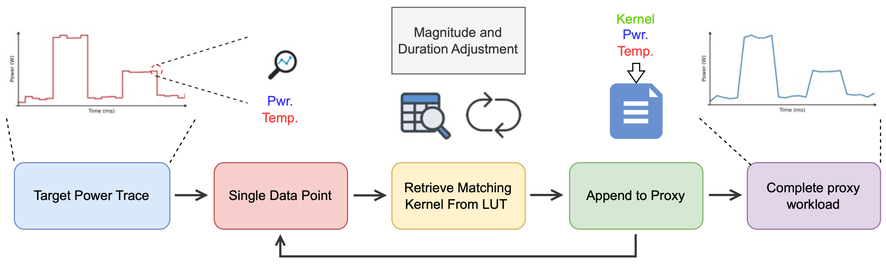
- Several asm-level microbenchmark kernels
- Vary runtime characteristics (grid/block dim, sparsity, etc.)
- Measure steady-state power value for each variation
- Log each kernel and corresponding power value

LUT is generated once per reference device (2-5 hours runtime)

Power Ranger can generate any workload utilizing the generated LUT, *given only the original power trace as input:*

For each point in the input power trace:

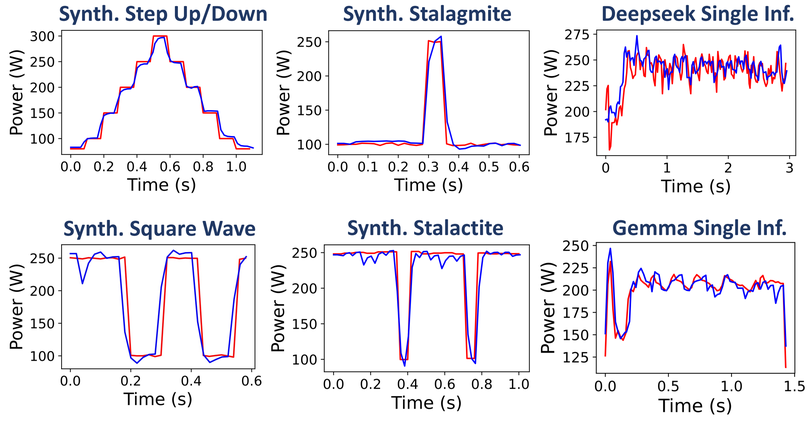
- Select a LUT kernel with matching power value at matching temperature
- Append kernel to replication workload
- Run whole replications workload and log final replication power value
- If power matches, append kernel and move to the next data point
- Else, repeat process with a different LUT kernel, using binary search to locate the kernel



Power Ranger can create real workloads from any arbitrary power trace. This means it can be used to create a *real workload* from a *desired, imaginary* power trace. This makes it a great tool for creating power stressmarks.

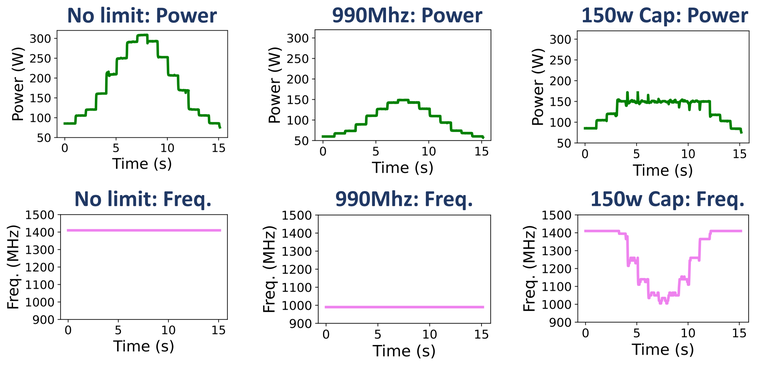
Validation

- We conduct experiments on NVIDIA A100 GPU
- We use Power Ranger to recreate workloads and compare recreations to original target
- **MAPE of 5% or less for all traces**



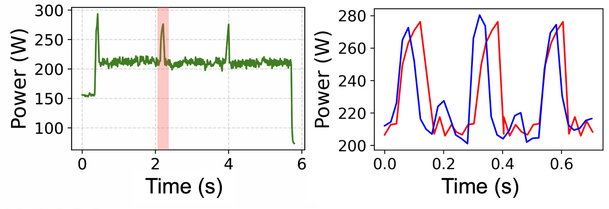
Case Study: DVFS Testing

- We use Power Ranger to create a synthetic workload for evaluating the effect of fixed clock frequency setting and power capping on instantaneous GPU power
- We find that power capping allows for more energy-efficient computation, but it has a much noisier instantaneous power trace
- This noisy power trace is due to the dynamic changes in clock frequency used to enforce the power cap



Discussion: Region of Interest Replication

- What if we are only interested in power behavior at a specific area of a workload?
- We can use Power Ranger to isolate areas of interest and replicate those areas only



References

1. R. Panda and L. K. John, “Proxy benchmarks for emerging big-data workloads,” in 2017 26th International Conference on Parallel Architectures and Compilation Techniques (PACT), 2017, pp. 105–116.
2. Z. Jiang, J. Garrigus, A. Seigler, E. Syed, Y. Huang, M. Sadi, T. Rahal-Arabi, and L. K. John, “Exploration of LLM Workload Reliability based on di/dt Effects and Voltage Droops,” in 2026 IEEE International Symposium on High-Performance Computer Architecture (HPCA), 2026, pp. 1-15.
3. 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.
4. D. Zhao, S. Samsi, J. McDonald, B. Li, D. Bestor, M. Jones, D. Tiwari, and V. Gadepally, “Sustainable supercomputing for ai: Gpu power capping at hpc scale,” in Proceedings of the 2023 ACM Symposium on Cloud Computing, ser. SoCC ’23.

This research was supported in part by the Semiconductor Research Corporation (SRC), as well as by AMD.