Characterizing GPU Performance in AI Inference with Power Metrics

Sean Tai
Queensborough Community College
New York, USA

Mohammad Atif*, Sankha Dutta, Meifeng Lin
Brookhaven National Laboratory
New York, USA
*fmohammad@bnl.gov

***Index Terms*—LLM, GPU, Power Consumption, HPC**

Introduction: Large language model (LLM) inference performance depends on compute capability, memory bandwidth, runtime behavior, and GPU power efficiency. This study compares inference performance and energy use across an edge system-on-module and two datacenter GPUs. Experiments are performed on NVIDIA Jetson AGX Thor, A100-SXM4-80GB, and H200 using vLLM and Hugging Face Transformers with three open-weight models: Llama-3.1-8B-Instruct, Gemma-4-E4B-it, and Qwen2.5-7B-Instruct. For each hardware, framework, and model combination, we measure throughput, GPU power, utilization, KV-cache usage, generation time, and memory behavior. Pareto analysis is used to evaluate throughput-power trade-offs. STREAM and BabelStream-style benchmarks measure sustained memory bandwidth, while phase-resolved roofline analysis separates prefill and decode to identify compute- and memory-bound regimes.

Methods: We use two benchmark modes. In whole-generation runs, we measure end-to-end throughput and power. In phase-resolved runs, we isolate prefill and decode so that we can analyze their different hardware demands. We measure hardware ceilings directly rather than using vendor peak specifications. We obtain the BF16 compute roof from the best sustained performance in a GEMM sweep. We measure the memory roof with GPU copy and triad kernels using STREAM/BabelStream methodology. We measure sustained ceilings of 137.9 TFLOP/s and 239.7 GB/s for Jetson Thor, 291.7 TFLOP/s and 1794.5 GB/s for A100, and 717.2 TFLOP/s and 4231.9 GB/s for H200. These values produce ridge points of 575, 163, and 169 FLOP/byte, respectively.

For each hardware, framework, and model combination, we execute matched prompts and collect performance and power measurements with PyNVML. We restrict monitoring to the GPU under test and measure idle power before inference. We subtract idle power from measured power and use the resulting dynamic power to calculate energy per token and power-per-throughput. We calculate arithmetic intensity and achieved performance for prefill and decode and place both phases on device-specific rooflines. We also record clock state, framework configuration, phase boundaries, and resident weight precision to keep comparisons consistent.

Results: The roofline results show a clear separation between prefill and decode. Decode clusters near 4 FLOP/byte on all three systems and remains memory-bound. Prefill has substantially higher arithmetic intensity, approximately 200–800 FLOP/byte. On A100 and H200, prefill reaches or crosses the ridge point and becomes compute-bound in many cases. Jetson Thor has a much higher ridge point, so prefill often remains at or below the memory-bound side of the roofline. We observe that the H200 achieves the highest absolute throughput because of its high HBM3e bandwidth and compute capability, but decode remains memory-bound. Additional compute alone does not remove the decode bottleneck; performance improves mainly because the memory roof is higher. Jetson Thor provides lower throughput but operates at much lower power. We find that throughput and energy efficiency do not scale proportionally. For Llama with vLLM during decode, Jetson Thor reaches about 42 tokens/s at 0.29 dynamic J/token, A100 reaches 305 tokens/s at 0.69 J/token, and H200 reaches 639 tokens/s at about 0.32 J/token. A100 therefore increases throughput relative to Thor but at higher energy per token, whereas H200 combines the highest throughput with substantially better energy efficiency than A100. In the power-throughput Pareto analysis, Thor defines the low-power end of the frontier and H200 defines the high-throughput end. Framework choice also affects performance. vLLM generally provides the higher-throughput results, particularly during decode, and achieves the best whole-generation performance in all 9 comparisons. Transformers is more sensitive to runtime overhead, kernel selection, and resident weight precision.

Discussion and Conclusions: The results show that LLM inference efficiency is controlled by memory movement and runtime behavior as well as compute capability. Decode operates at low arithmetic intensity and remains memory-bound on all three GPUs, so peak FLOP/s is a poor predictor of decode efficiency. Prefill has higher arithmetic intensity and benefits more from datacenter GPU compute, particularly on A100 and H200. Jetson Thor remains more bandwidth-constrained because of its high ridge point. We also show that framework implementation matters. vLLM outperforms Transformers in most whole-generation comparisons, which indicates that scheduling, batching, kernel selection, and memory management can materially affect both throughput and energy efficiency. To make fair comparisons, we subtract idle power, monitor only the target GPU, separate prefill from decode, and record resident weight precision. We find GPU utilization useful for identifying underuse, but once utilization

saturates, it provides less information about performance. Dynamic power and energy per token provide more useful measures of inference efficiency. The three platforms occupy different operating regions. We identify Jetson Thor as the low-power option and H200 as the high-throughput option. In the reported Llama vLLM decode case, H200 approaches Thor's dynamic energy per token while delivering a much higher token rate. A100 provides intermediate throughput but higher energy per token in the same comparison. We therefore select hardware based on deployment objectives such as power, latency, throughput, and energy efficiency.

In future work, we will extend the tested sequence-length and batch-size ranges, validate datacenter measurements with kernel-level counters using tools such as Nsight Compute, and perform precision-matched comparisons for models that use mixed or compressed weight formats.

Acknowledgements: We are grateful to LDRD program at Brookhaven National Laboratory, which is sponsored by the US Department of Energy, Office of Science, under Contract Number DE-SC0012704. Sean was supported by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Community College Internships (CCI) Program.