From Microbenchmarks to LLM Inference – An End-to-End Analysis on the Energy Efficiency of the Grace Hopper Superchip

Markus Velten[1], Christian von Elm[1], Lena Jurkschat[2], Gülçin Gedik[1], Sebastian Döbel[1], Daniel Hackenberg[1]

[1]ZIH, CIDS, TU Dresden [2]ScaDS.AI, CIDS, TU Dresden

1. Introduction & Background

Motivation:

- GH200 leads Green500 (5 of top 10, Nov. 2025); KAIROS: 73.3 GFlop/J (#1 in Green500), JUPITER: 63.3 GFlop/J (#14 in Green500)
- Traditional GPU systems limited by PCIe bandwidth ($\leq$64 GB/s); GH200 solves this with 900 GB/s NVLink
- Shared CPU–GPU power budget (200–1000 W) with adjustable GPU limit (1–900 W)
- LLMs are the largest single consumer of GPU resources – energy efficiency analysis can yield significant savings

Our Analysis: How do GH200's shared CPU-GPU power controls translate into energy, performance, and cost outcomes across compute-, memory-, and mixed-phase workloads?

Contribution: A measurement-validated, cross-workload characterization linking sensor fidelity, power-budget enforcement, workload bottlenecks, and TCO.

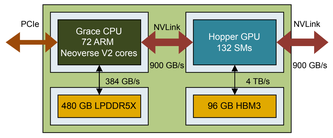


Fig. 1: Grace Hopper Superchip architecture (based on Nvidia whitepaper). Values for evaluated system.

- **CPU:** 72 ARM Neoverse V2 (SVE2), 480 GB LPDDR5X (384 GB/s)
- **GPU:** Hopper H100, 132 SMs, 96 GB HBM3 (4 TB/s)
- **NVLink:** 450 GB/s unidirectional, 900 GB/s bidirectional, cache-coherent (ATS)
- **Power:** Module 200–1000 W; GPU 1–900 W
- **System:** Air-cooled

2. Power Sensor Precision

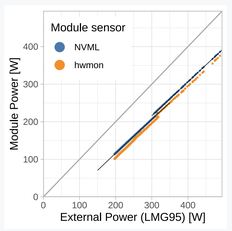


Fig. 2: roco2-based sensor precision analysis. Black lines: quadratic model on NVML data.

Method: roco2 microbenchmarks; 8–72 CPUs $\times$ 30 frequencies; external power meter ZES ZIMMER LMG95 ($\pm$0.015%·P+0.5 W).

Findings:

- NVML: $R^2_{\text{lower}} = 0.9998$, $R^2_{\text{upper}} = 1.0$ – exceptional precision, unlike contemporary x86 (AMD EPYC, Intel)
- hwmon: $R^2 = 0.9991/0.9997$ with 9–13 W offset vs. NVML
- Precision enables energy-based billing in shared HPC/cloud environments
- CPU power no longer leaks into NVML GPU metric [4]
- hwmon no longer exhibits >100 W noise fluctuations

3. LLM Energy Efficiency (Llama-3.1-8B-Instruct via vLLM)

Time-to-First-Token (TTFT) Scaling

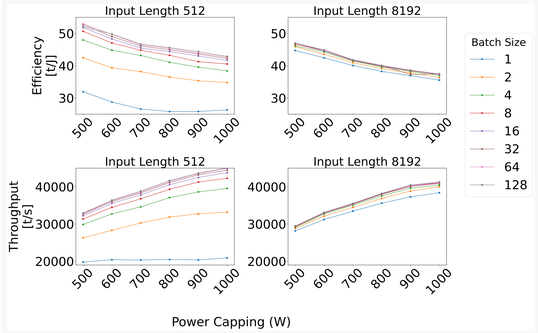


Fig. 3: TTFT energy efficiency & throughput across batch size, input length, and module power limits for Llama-3.1-8B-Instruct (prefill-dominated). Evaluated with BALI benchmark.

Findings:

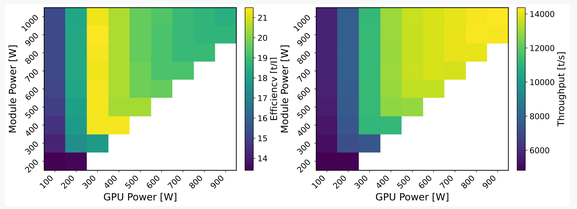
- Non-compute-bound (small batch/short input): up to 21% efficiency gain at only $\approx$5% throughput loss
- Compute-bound (large batch): 26.4% efficiency gain but 28.7% throughput loss – near-linear trade-off
- Throughput saturates at batch 32, 512–1024 input tokens (GPU fully utilized)
- Increasing input length slightly reduces efficiency ($-$6 tokens/J) due to sequence overheads

Output Length Scaling

Findings (1000 W vs 500 W module, batch 32, 1024 input tokens):

- **Short outputs** ($\leq$128 t): power limits unfavorable – throughput loss ($-$14–28%) exceeds efficiency gain ($+$14–23%)
- **Long outputs** ($\geq$512 t): memory-bound decoding dominates – efficiency gains up to $+$11% at only $-$3–6% throughput cost; power limits beneficial

GPU & Module Power Limits Impact



(a) Energy Efficiency [tokens/J] *(b) Throughput [tokens/s]*

Fig. 4: Module & GPU power limit impact with 512 input tokens, 128 output tokens, batch size 32. Module limit barely affects efficiency; GPU limit dominates.

Findings:

- Module power limit has negligible effect unless equal to GPU limit; GPU limit dominates
- **Best efficiency at 300 W GPU** ($+$15% efficiency, $-$22% throughput); balanced trade-off at 500 W ($\approx\pm$6%); below 100 W: $\sim$60% efficiency loss – minimum limit required

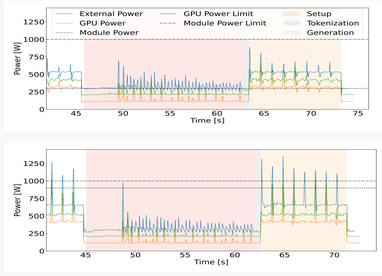


Fig. 5: Power consumption over time during inference at 300 W (top) and 900 W (bottom) GPU power limits (128 input, 128 output tokens, batch 32).

Total Cost of Ownership (TCO)

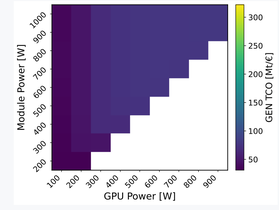
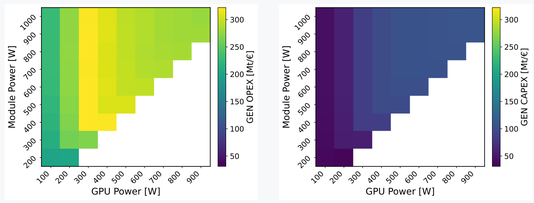


*Fig. 6: Tokens/€ (TCO = OPEX + CAPEX). 0.20 €/kWh, €20 833/node, 5-year amortization. **Best TCO: 78 M tokens/€ @ 900 W GPU.***

Findings:

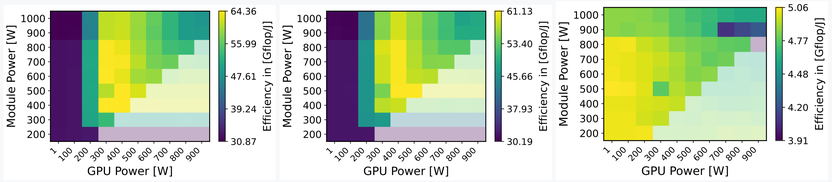
- **OPEX**: efficiency-driven – best at 300–500 W GPU ($\approx$320 M tokens/€, i.e. $\approx$€0.003/1M tokens)
- **CAPEX**: throughput-driven – reducing power limits *hurts* amortization
- **TCO**: follows CAPEX; **best at 900 W GPU** (78 M tokens/€ $\approx$ €0.013/1M tokens); comparable to commercial API: $\approx$52 M tokens/€ ($\approx$€0.019/1M tokens, OpenRouter)



(a) OPEX *(b) CAPEX*

Fig. 7: OPEX and CAPEX breakdown showing different optimal operating points for energy costs vs. hardware amortization.

4. HPL Efficiency



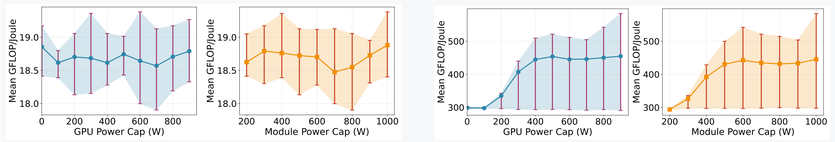
(a) Hopper GPU, GPU memory only *(b) Hopper GPU, GPU + CPU memory* *(c) Grace CPU only*

*Fig. 8: HPL energy efficiency [GFlop/J] across module and GPU power limit combinations. De-saturated cells: module power < GPU power. **Peak: 64.36 GFlop/J @ 400 W (GPU-mem only)**.*

Findings:

- Peak: 51 TFlop/s at default limits; 38.1 TFlop/s at optimal efficiency (400 W)
- Out-of-core (GPU+CPU mem): peak 61.1 GFlop/J at 500 W module / 400 W GPU (NVLink + LPDDR5X overhead)
- CPU has power priority over GPU; GPU allocation reduced when combined power exceeds module limit
- Module power can exceed configured limits (225 W drawn at 200 W module limit; CPU alone: $\approx$380 W measured at 200 W limit)
- Thermal throttling at 1000 W default after $\approx$8 s (air-cooled); idle GPU: $\approx$55 W despite 1 W limit

5. HPL Mixed-Precision Efficiency



(a) FP32 factorization & FP64 iterative refinement entirely on Grace CPU *(b) FP16 factorization on Hopper GPU, FP64 iterative refinement on Grace CPU*

*Fig. 9: Marginal GFlop/J response ($\pm$1 std) to module and GPU power limit changes. FP8 (not shown) similar to FP16 but higher efficiency. **FP8 peak: 816 GFlop/J – 12.7$\times$ vs FP64 HPL.***

Findings:

- FP8 default (1000/900 W): 488 TFlop/s, 648 GFlop/J; optimal (700/600 W): 446 TFlop/s, **816 GFlop/J** ($-$9% perf, $+$26% efficiency)
- FP16: default settings achieve *both* peak throughput (378 TFlop/s) and best efficiency (584 GFlop/J) simultaneously
- CPU-only (FP32): <1% efficiency variation across all power configs – power limits ineffective
- Over-provisioning beyond saturation is ineffective; GPU needs module headroom $\geq$600 W
- Memory-bound iterative refinement on CPU limits phase-specific tuning due to CPU-priority power allocation
- GPU power limit exceeding module limit is configurable but undocumented; effect matches module$-$100 W

6. Conclusions & Future Work

- ✓ **On-board sensors exceptionally precise** ($R^2 > 0.999$) – enables energy-based billing in shared HPC/cloud; sensor behavior can change with software updates
- ✓ **Power limits not always enforced** – module power can exceed configured limits (380 W at 200 W limit); GPU idle: 55 W at 1 W limit; GPU limit exceeded during compute-bound LLM prefill (558 W observed at 300 W limit)
- ✓ **Mixed-precision:** FP8 HPL-MxP is 12.7$\times$ more efficient than FP64 HPL; FP16 uniquely achieves peak throughput and efficiency simultaneously
- ✓ **LLM prefill** benefits from power limits only with latency bound configurations (e.g. small batch sizes and sequence length) (up to 21% efficiency at 5% throughput cost); **decoding** heavy workloads (e.g. $\geq$512 tokens) favor power limits.
- ✓ **TCO:** OPEX efficiency-driven (best at 300–500 W GPU, $\approx$320 M tokens/€ $\approx$ €0.003/1M tokens); CAPEX & TCO throughput-driven – comparable to commercial APIs ($\approx$52 M tokens/€ $\approx$ €0.019/1M tokens); **best TCO: 78 M tokens/€ @ 900 W GPU** ($\approx$€0.013/1M tokens)
- → **Future work:** Evaluate Grace Blackwell Superchip; cross-architecture power sensor quality review; extend LLM analysis to larger models, multi-GPU configurations, and diverse inference workloads

Acknowledgments

Funded by Bundesministerium für Bildung und Forschung via EECliPs project (grant 16ME0602) and NHR Graduate School.

References

[1] Velten et al. *From Microbenchmarks to LLM Inference – An End-to-End Analysis on the Energy Efficiency of the Grace Hopper Superchip*. ISPDC'26, pp. 121–129.
[2] Top500. *Green500 List – Nov. 2025*. top500.org, 2025.
[3] NVIDIA Corp. *GH200 Grace Hopper Superchip Architecture Whitepaper*. 2024.
[4] Yang et al. *Accurate and Convenient Energy Measurements for GPUs*. SC'24.
[5] Hackenberg et al. *Power Measurement Techniques on Standard Compute Nodes*. ISPASS'13.
[6] Dongarra & Luszczek. *HPL-MxP Benchmark*. IJHPCA, 2025.
[7] Jurkschat et al. *BALI – A Benchmark for Accelerated Language Model Inference*. IEEE Access, 2025.
[8] Hernandez et al. *Fine-Grained Power Measurements on GH200*. HPCAsia Workshops, 2025.