NHR FAU

FAU Friedrich-Alexander-Universität Erlangen-Nürnberg

Architectural Trade-offs in the Energy-Efficient Era: A Comparative Study of power-capping NVIDIA H100 and H200

Aditya Ujeniya, Jan Eitzinger, Georg Hager, Gerhard Wellein

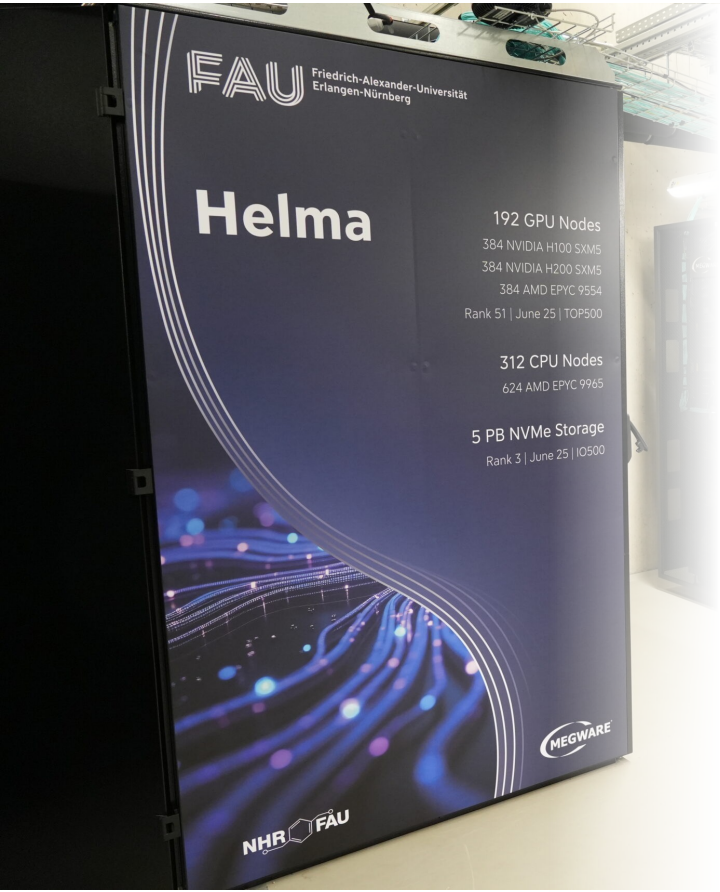
ISC EESP Workshop 2026

26th June 2026

Outline

- Introduce
  - NVIDIA H100 and H200
  - Benchmarks
    - DGEMM
    - TheBandwidthBenchmark (TBB)
- Power-cap analysis
  - Effects on SM frequency throttling
- GPU and memory power breakdown

NVIDIA H100 and H200 characteristics in Helma

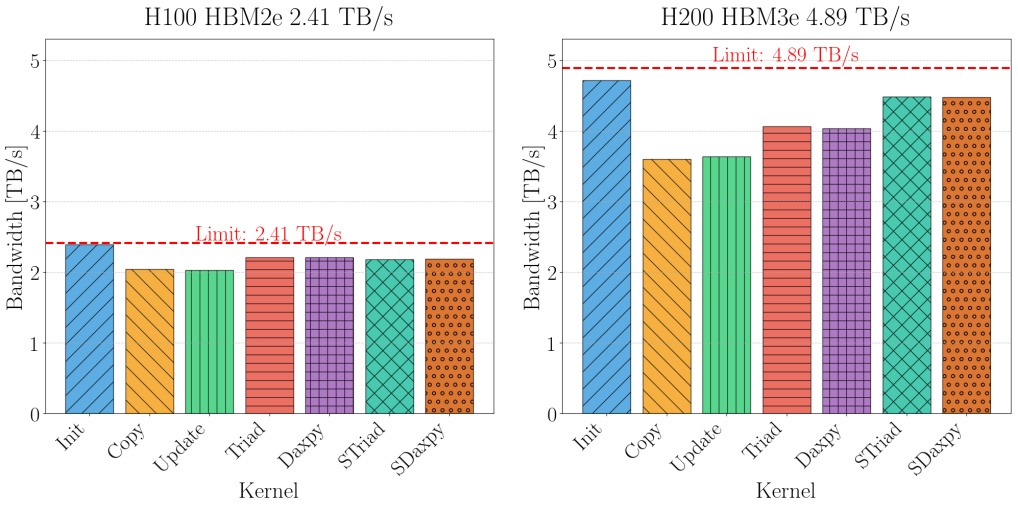


	H100 94 GB HBM2e	H200 144 GB HBM3e
Power Draw [W]	700	700
TF64 Peak [TFlop/s]	67	67
Mem BW [TB/s]	2.41	4.89
Nodes used	4	4
Base SM Freq [MHz]	1665	1665
Boost SM Freq [MHz]	1980	1980

Modules used: *nvhpc/24.11, cuda/12.9.0*

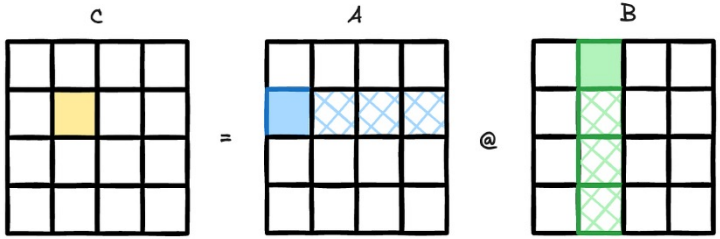
Performance mode: *P0*

Benchmark – TheBandwidthBenchmark (TBB)



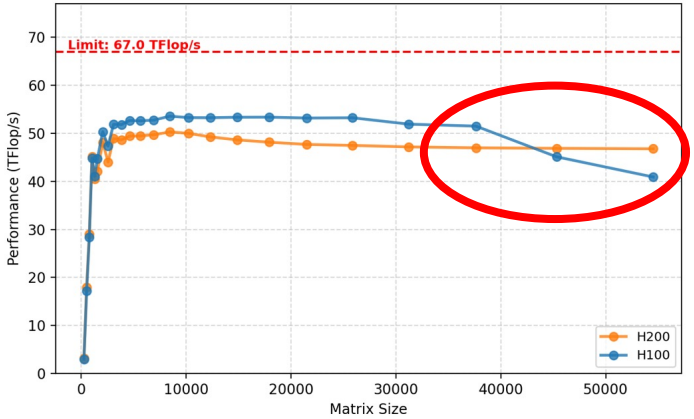
- TBB designed to reveal sustained memory bandwidth on CPU and GPU.
- Different kernels with different number of streams and different data access pattern.
  - **Init**: a[i] = s;
  - **Copy**: a[i] = b[i];
  - **Update**: a[i] = a[i] * s;
  - **Triad**: a[i] = b[i] + c[i] * s;
  - **Daxpy**: a[i] = a[i] + b[i] * s;
  - **STriad**: a[i] = b[i] + c[i] * d[i];
  - **SDaxpy**: a[i] = a[i] + b[i] * c[i];
- We only focus on Schönauer Triad (STriad) kernel going forward.

Benchmark - DGEMM

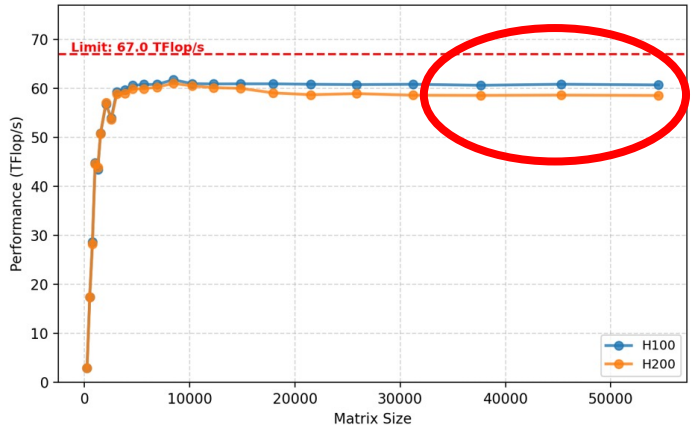


- 2D CuBLAS DGEMM used for benchmarking.
- Matrix size of 32768^2 is used.
- Larger problem sizes tends to be memory-bound as we increase the power-cap.

At 500W

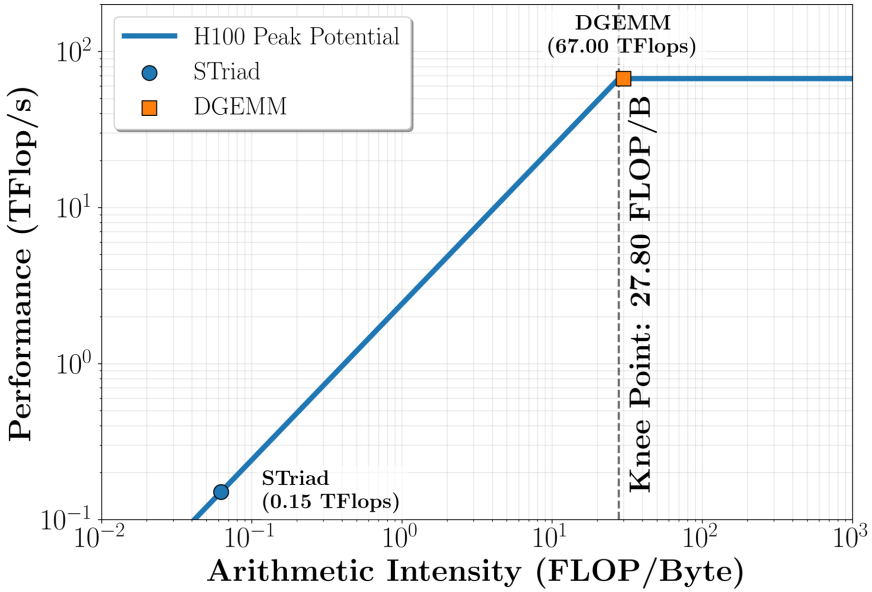


At 700W

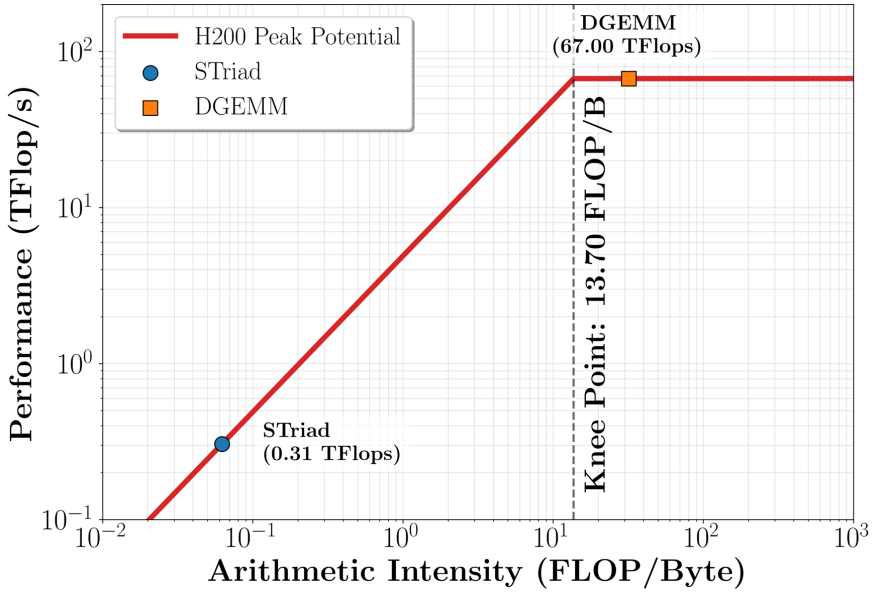


Roofline for the benchmarks

H100 Roofline



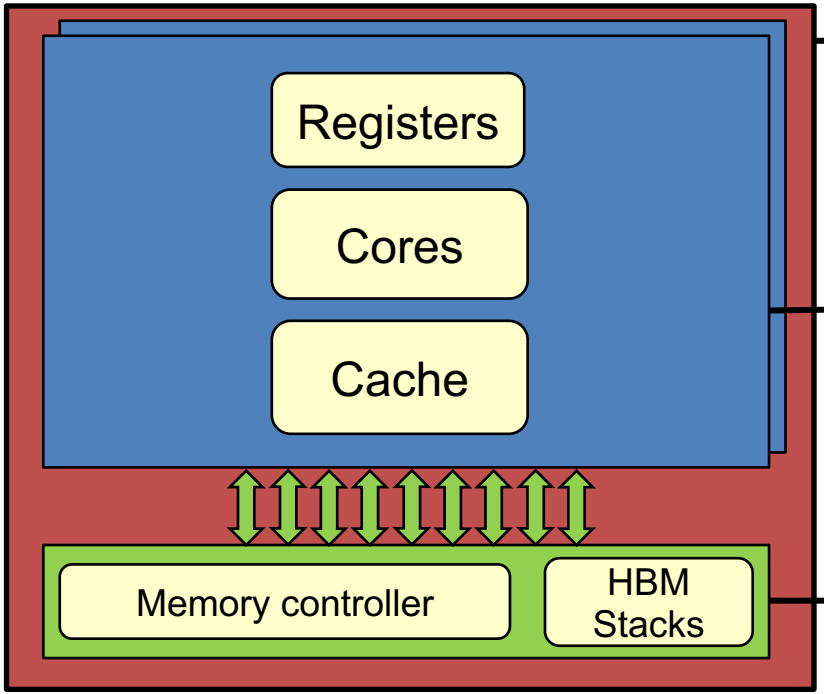
H200 Roofline



Benchmarking methodology

- All benchmarks run for atleast 10 mins at full TDP.
- Power-capping from 200W till 700W with increments of 100W.
- Regression for DGEMM and STriad benchmarks:
  - 4 nodes of NVIDIA H100 and H200 each.
  - Each node has 4 GPUs.
  - Same benchmark binary submitted on each GPU.
  - 50 repetitions per power-cap.
  - 800 data points per GPU per power-cap.
- nvidia-smi used for power measurements.

Power measurements through nvidia-smi



Total Power Draw W_{Total}
(Measured metric)

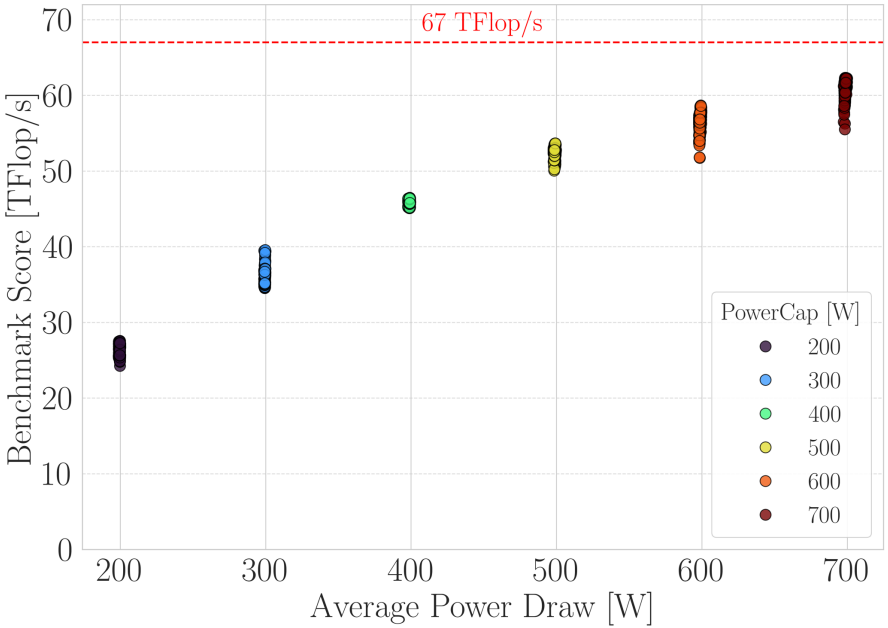
GPU Power Draw W_{GPU}
(Derived metric)

Memory Power Draw W_{Memory}
(Measured metric)

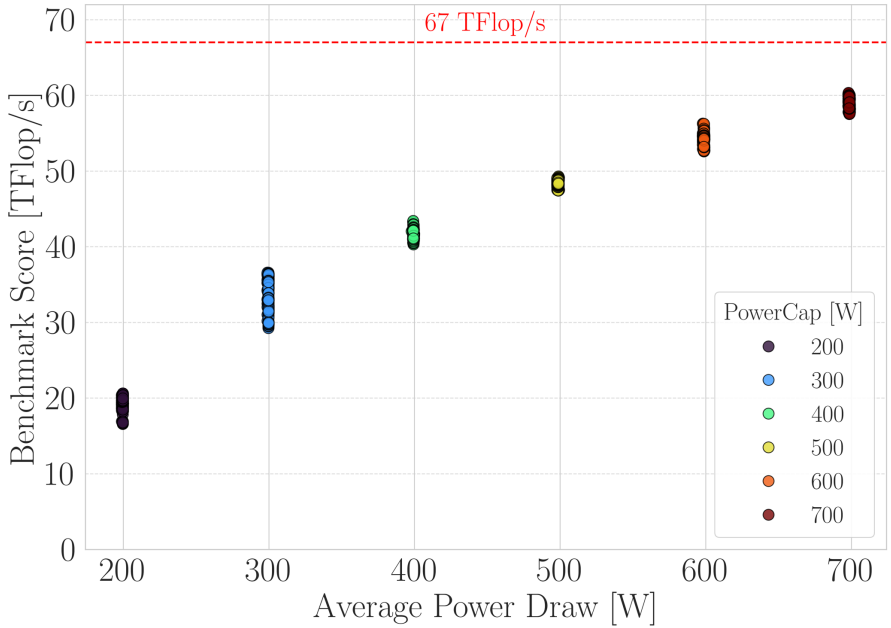
$$W_{Total} = W_{GPU} + W_{Memory}$$

Analysis for DGEMM on NVIDIA H100 and H200

NVIDIA H100

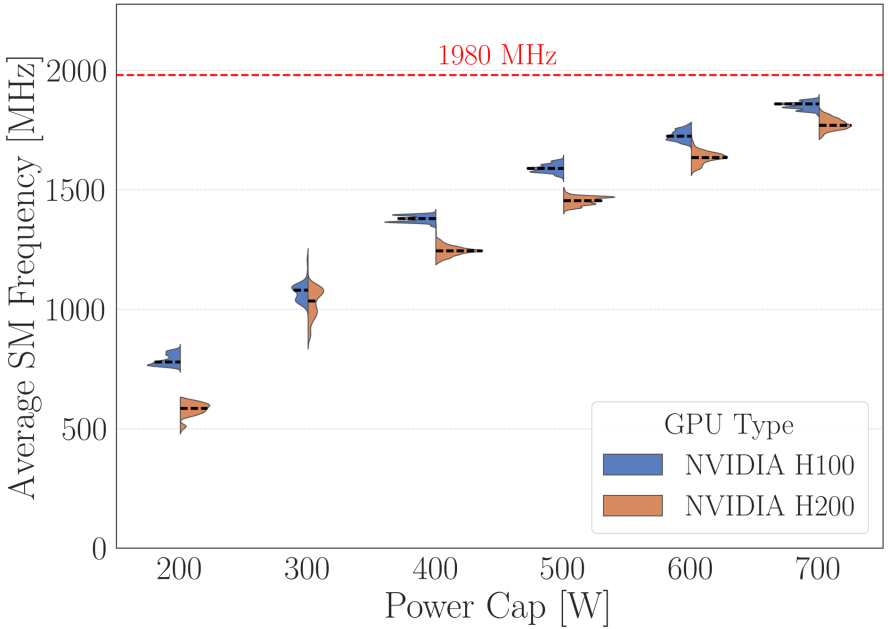


NVIDIA H200

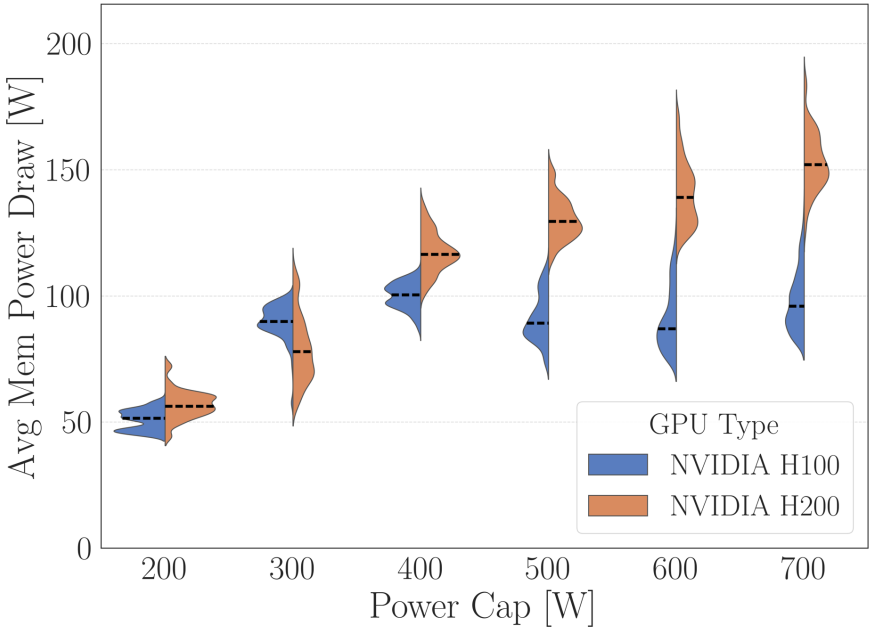


Frequency and memory power analysis for DGEMM

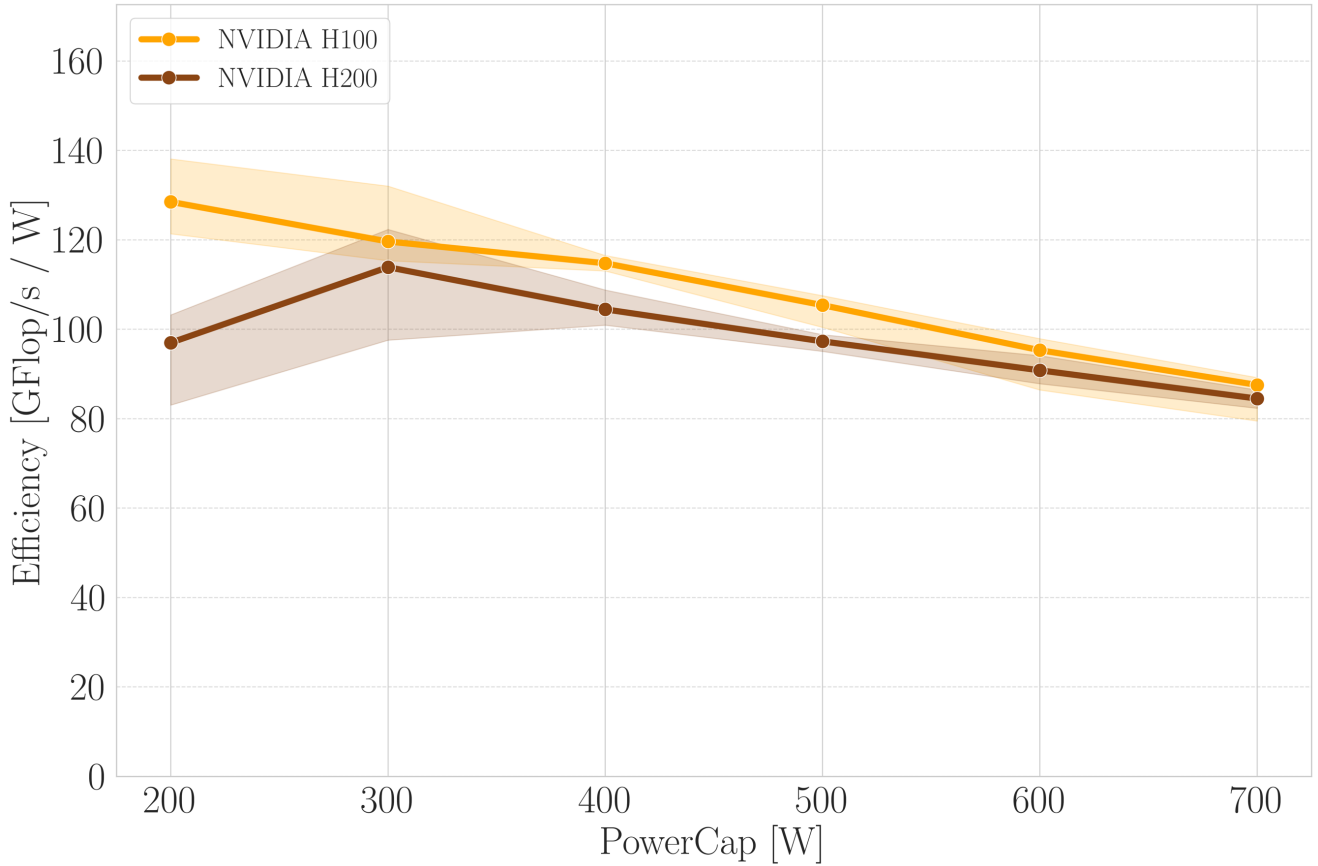
SM Frequency violin plots



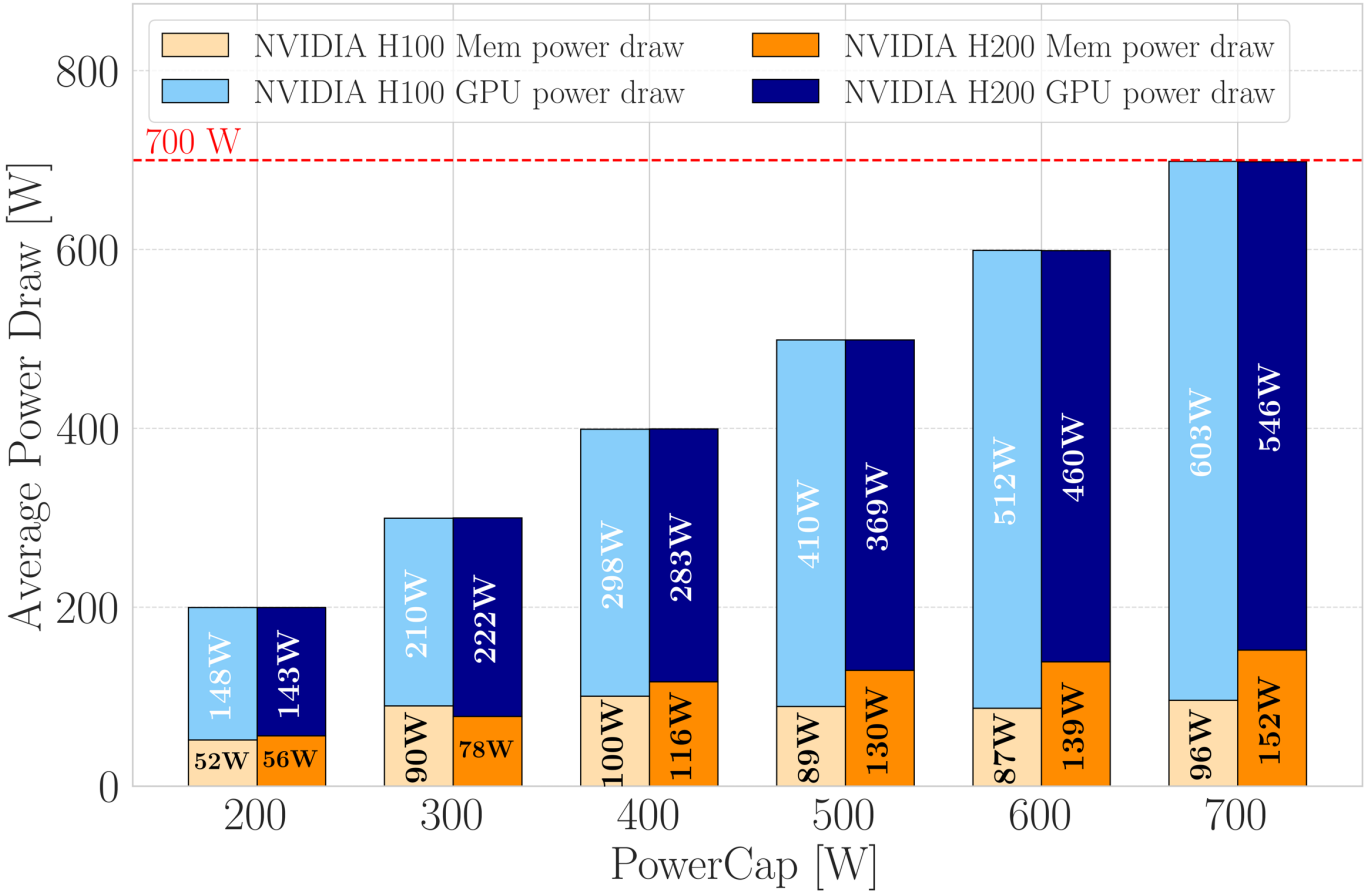
Memory power draw violin plots



Efficiency curve for DGEMM on NVIDIA H100 and H200

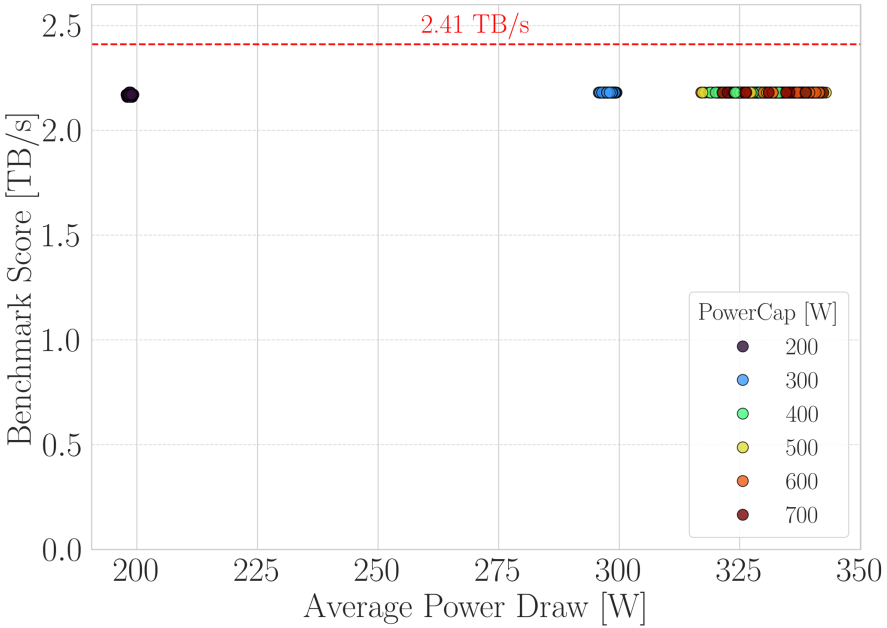


Power breakdown for DGEMM on H100 and H200

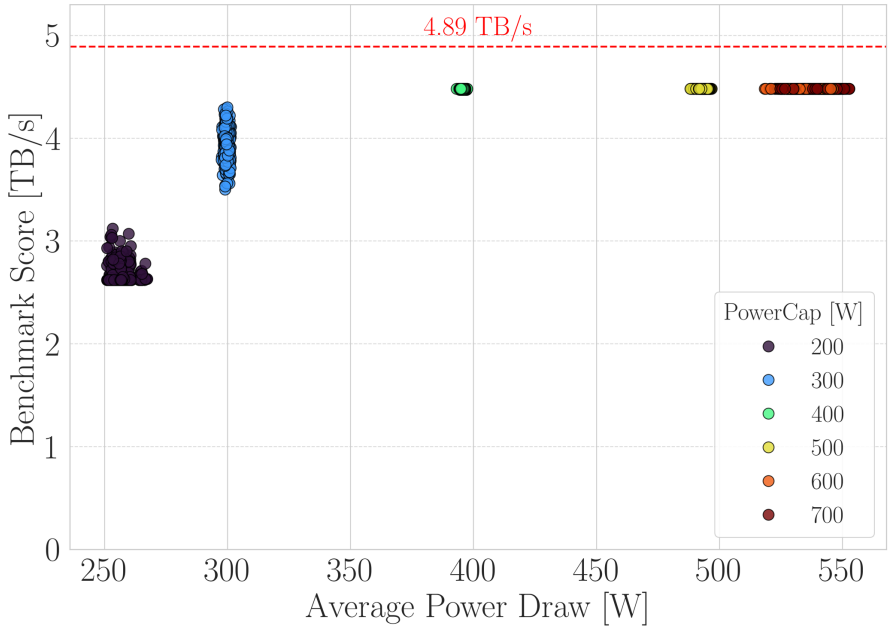


Analysis for STriad on NVIDIA H100 and H200

NVIDIA H100

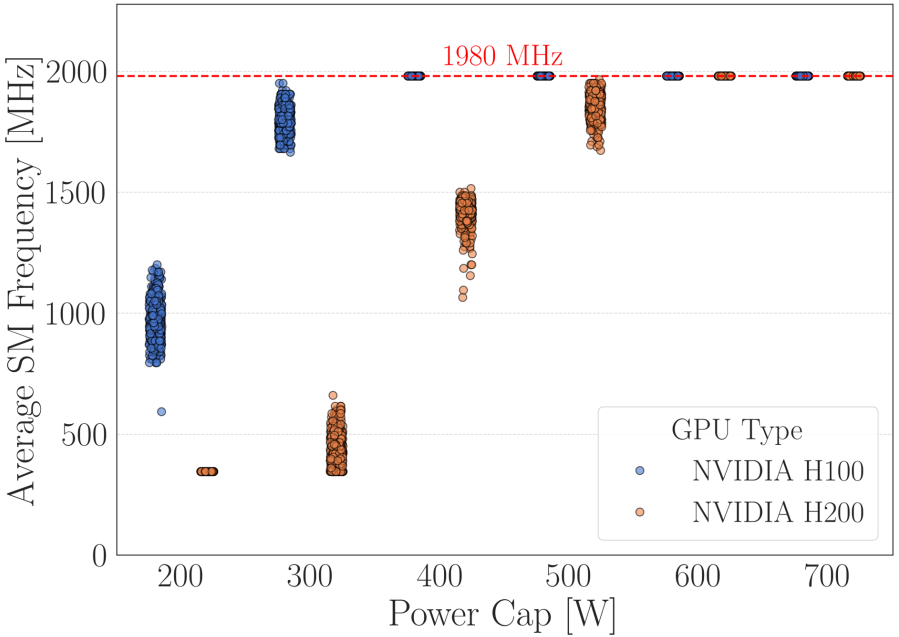


NVIDIA H200

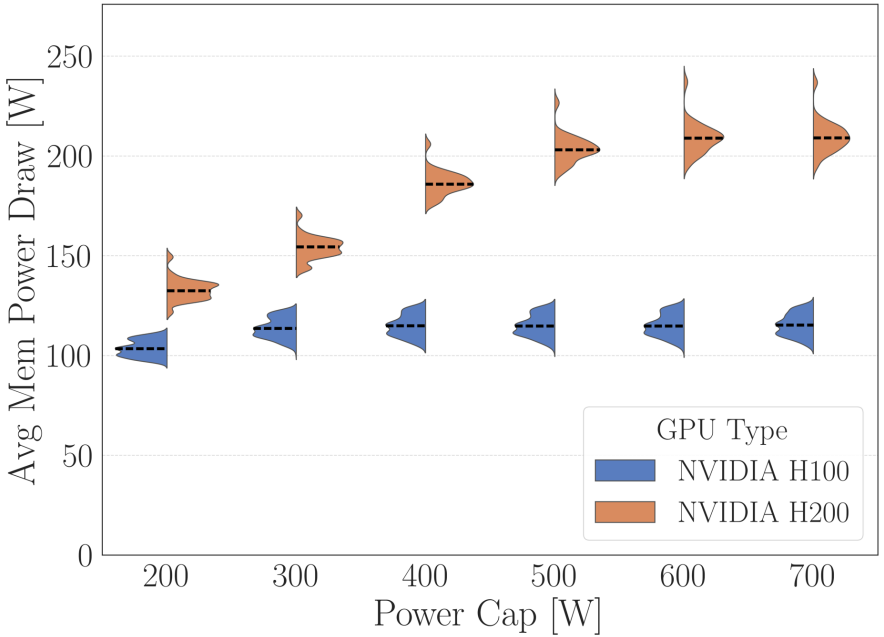


Frequency and memory power analysis for STriad

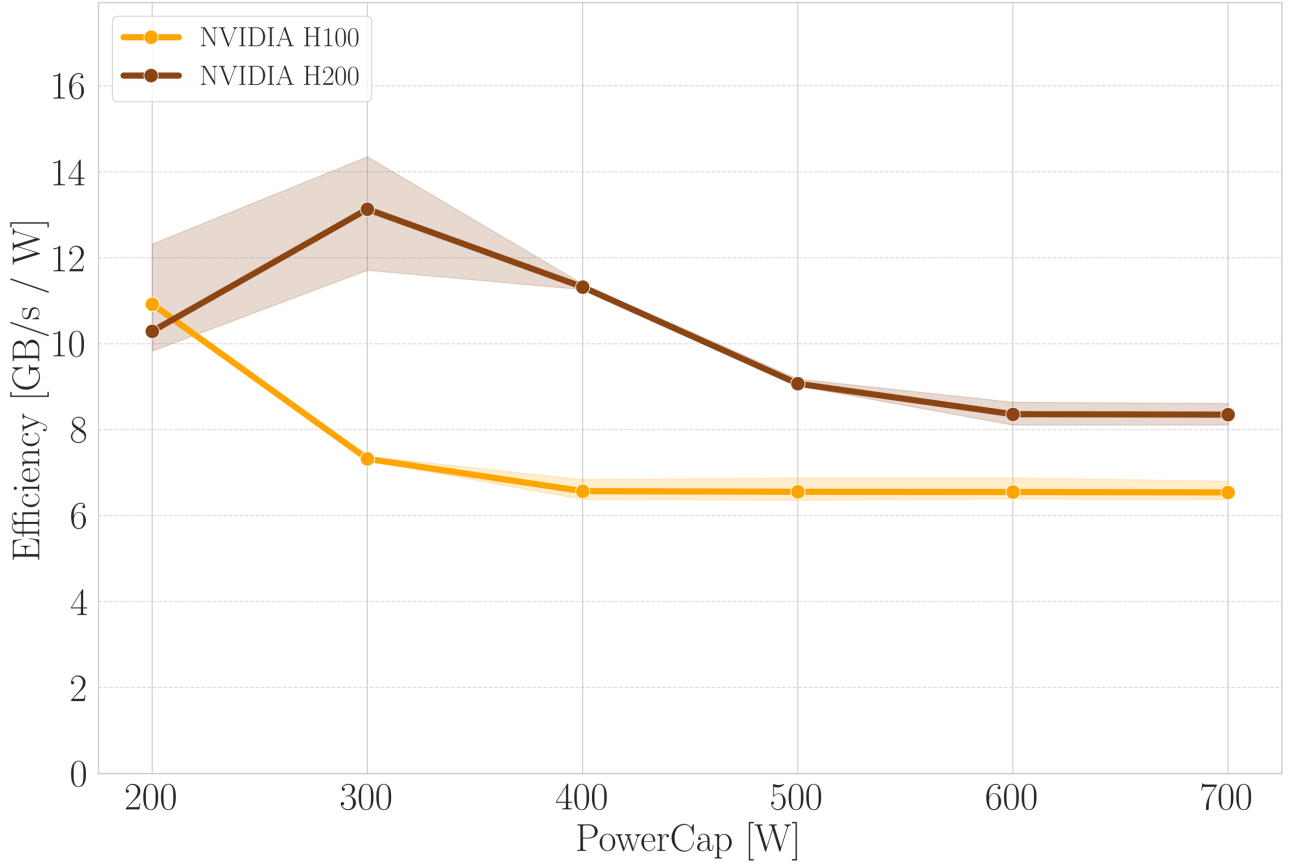
SM Frequency scatter plots



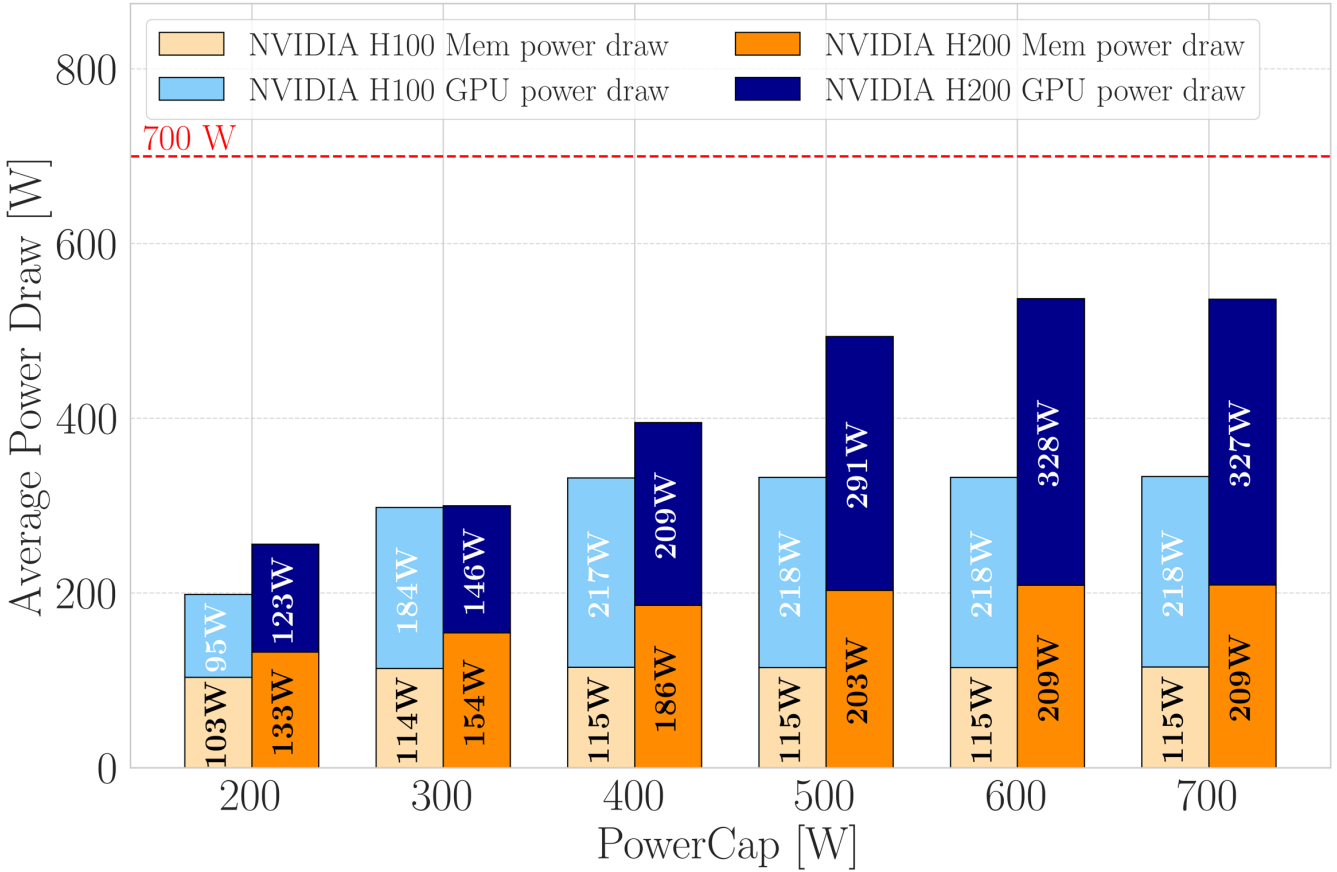
Memory power draw violin plots



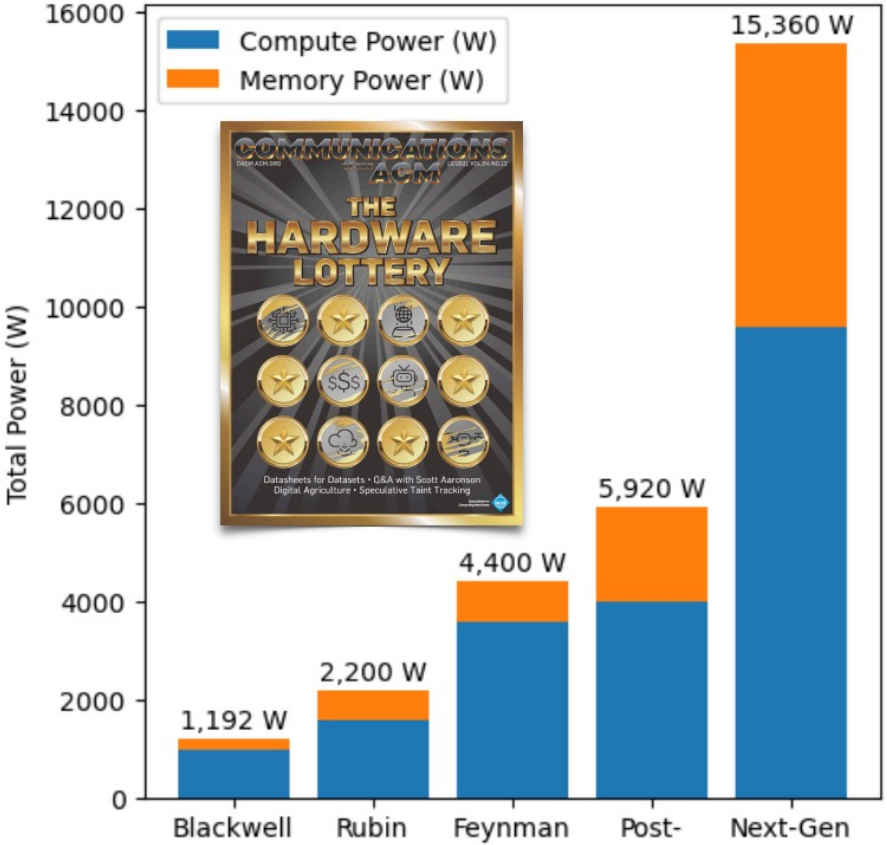
Efficiency curve for STriad on NVIDIA H100 and H200



Power breakdown for STriad on H100 and H200



Why does this study matter ?



Ref: PerfLab Seminar - FROM PICOJOULES TO GIGAWATT-HOURS ENERGY-TO-COMPLETION AND GPU DVFS FOR LLM WORKLOADS - Holger Fröning (https://hpc.fau.de/files/2026/01/2026-01-20_Froening.pdf)

Conclusion

- General points to conclude:
  - The efficient point in the efficiency curve will depend on different datatypes and where the application lies within the roofline.
  - The trade-off in performance comes from difference in GPU power draw share that's left after memory power draw.
  - Memory frequency was 1593 MHz for NVIDIA H100 and 3201 MHz for NVIDIA H200 across all different power-caps.
  - NVIDIA H200 is clearly more energy efficient than H100 for memory bound codes.
- Points to conclude from regression:
  - The memory power draw for H100 is near about 120W and for H200 is near about 220W.
  - Cross specimen statistical variation of performance is more than 5% whereas the statistical variation of performance for individual GPUs remains within 5%.