Debunking the CUDA Myth

Towards GPU-based AI Systems

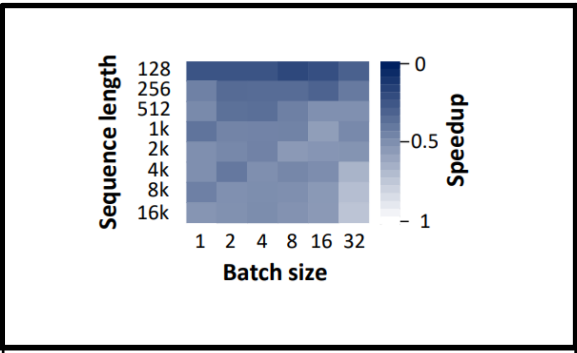
Original Paper by: Yunjae Lee et al. (Prof. Minsoo Rhu's Group, ISCA '25)

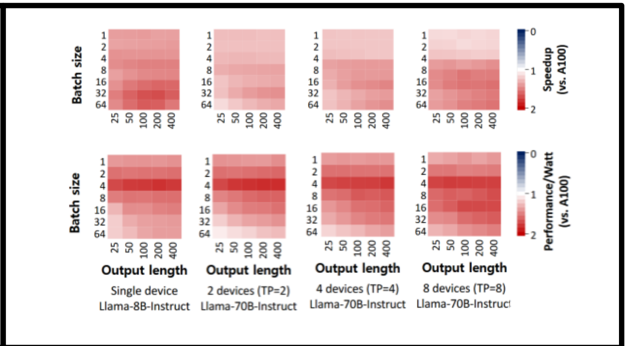
Presented by Jongyun Hur

Conflict Analysis

Conflict Analysis

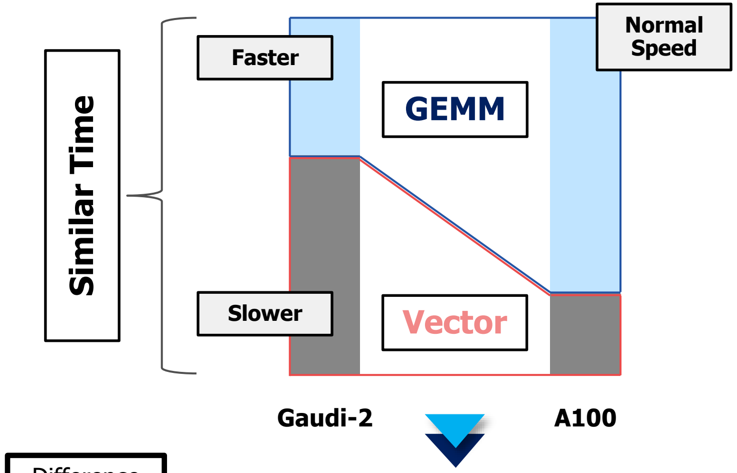
Kernel-level weakness vs. End-to-end strength





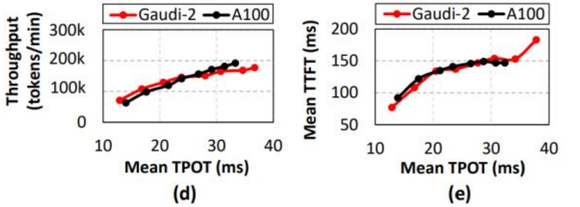
Category	Figure 17(c): PagedAttention (Micro)	Figure 12(a): LLM End-to-End (Macro)
Analysis Scope	**Micro Level** (Kernel-specific)	**Macro Level** (System-wide)
Evaluation Target	Throughput of a **Single PagedAttention Kernel**	**Full Llama-3.1 Pipeline (Synthetic Dataset)**
Performance vs. A100	**~45% of A100** (Inferior)	**Avg. 1.47x of A100** (Superior)
Root Cause	**Vector & Memory-intensive** nature of Attention exposes Gaudi-2's TPC limitations.	**GEMM dominates** the total workload, where Gaudi-2 gains massive acceleration via MME.

Amdahl's Law:
Dominant GEMM performance fully offsets minor Vector latency



Difference

vLLM-based Real Serving (Dynamic Dataset)



Similar End-to-End Performance