hivenet

Hivenet Benchmark

Consumer-grade GPUs vs data center-grade GPUs

Context

This benchmark evaluates the inference performance of **consumer-grade GPUs** (RTX 4090, RTX 5090) against a **data center-grade GPU** (NVIDIA A100 80GB), focusing on practical inference workloads with medium-sized LLMs and large context windows.

The benchmark is conducted using the official `benchmark_serving.py` script from the VLLM project, using the public **ShareGPT** dataset which contains multi-turn conversational prompts.

Objectives

- Evaluate **latency** and **throughput** across different GPU classes.
- Determine whether **one or multiple consumer-grade GPUs** can surpass or match the A100 for small and medium sized models.
- Provide verifiable results for **infrastructure decision-making** (cost-effective deployment strategies).

Static Configuration

Parameter	Value
Context Length	8192 tokens
Output Length	512 tokens
Model	meta-llama/Meta-Llama-3.1-8B-Instruct
Precision	BF16
Batch Size (auto)	Based on GPU memory
Dataset	ShareGPT (https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered/resolve/main/ShareGPT_V3_unfiltered_cleaned_split.json)
Benchmark Tool	vLLM official `benchmark_serving.py` (vllm/benchmarks at v0.8.3 · vllm-project/vllm)

Test Scenarios

1. Moderate Load (Latency Test)

Attribute	Value
Request Rate	1 req/s
Number of Prompts	100
Goal	Capture **average** latency (TTFT, E2E)

2. Extreme Load (Throughput Test)

Attribute	Value
Request Rate	1100 req/s
Number of Prompts	1500
Goal	Measure **maximum output token throughput (tokens/sec)**

Results and Analysis

Scenario 1 – Latency under Moderate Load (1 req/s)

GPU	Avg ITL(ms)	Avg TPOT(ms)	Avg TTFT(ms)	Avg E2E latency(ms)	Notes
RTX 4090	19	19	349.9	9759.07	
RTX 5090	12.14	12.14	45.41	6058.57	· E2E: 14% faster than A100 · TTFT: 84% faster
A100	13.25	13.25	296.44	7080.9	

Key Insights: All GPUs handle moderate load scenarios effectively. However, the RTX 5090 significantly outperforms all other tested GPUs, including the high-end A100, in all latency categories:

1. **End-to-End Latency (E2E):** The RTX 5090 achieves **14% faster** E2E latency than the A100.
2. **Time-To-First-Token (TTFT):** The RTX 5090 dramatically reduces TTFT by **84%** compared to the A100, a crucial factor for interactive workloads and low-latency applications.

Scenario 2 – Throughput under Extreme Load (1100 req/s)

GPU	Avg Token Throughput (Tokens/sec)	Sustained RPS
RTX 4090	737.65	1.47
RTX 5090	3802.09	7.58
A100	3748.16	7.58

Key Insights:

1. **RTX 5090 surpasses A100 in raw throughput,** delivering ~1.4% more token/sec under load. This is significant given its lower cost and VRAM compared to A100.
2. **Data parallelism further shifts the performance curve:**
   a. **2× RTX 5090 (64 GB combined)** pushes throughput to **~7604.18 tokens/sec,** outperforming A100 by ~103%, essentially **doubling the inference capacity** while using less VRAM and being potentially more cost-efficient than a single datacenter A100.

Conclusion

Across both low-load and high-load inference scenarios with medium-sized model (8B), high-end consumer-grade GPUs demonstrate **comparable or superior performance** to the A100 datacenter-grade GPU.

- Under **moderate load (1 req/s),** the RTX 4090 offers **latencies close to the A100 performances,** and the RTX 5090 delivers superior performances.
- Under **extreme load (1100 req/s),** the RTX 5090 achieves **slightly higher throughput** than the A100, while dual RTX 5090s are expected to deliver **~100% more token throughput,** respectively.

While the A100 remains advantageous for certain workloads requiring larger VRAM, these results show that for medium-sized models, some **consumer-grade GPUs are viable alternatives,** especially when **cost, and scalability** are key considerations.

For most small to meidum-sized LLMs deployment scenarios, **well-configured consumer GPU clusters offer a practical, high-performance option** that challenges the exclusive role of datacenter-grade hardware.