AMD EPYC

WHITE PAPER

PRACTICAL STRATEGIES FOR LOW-COST LLM DEPLOYMENTS USING 4TH GEN AMD EPYC™ PROCESSORS

AMD
together we advance_data center computing

First Edition
May 2024

Seong H, Kim
Xiaoqun Wang
Yao Fu

PRACTICAL STRATEGIES FOR LOW-COST LLM DEPLOYMENTS USING 4TH GEN AMD EPYC™ PROCESSORS

CONTENTS

INTRODUCTION - 3

HARDWARE RESOURCES - 3

Calculating Memory Size for LLM Models 3
Memory Performance and Price Considerations 4

TEST METHODOLOGY AND RESULTS - - - - - - - - - - - - - - - - - 4

Models 4
Benchmarking Software 4
System Configurations 4
....................5
Performance Comparison 5
Latency Comparison 6
Power Efficiency 7
Performance Per Unit GB of Memory 7

CONCLUSION - 8

INTRODUCTION

The recent surge in demand for Large Language Models (LLM) has surpassed the industry's expectations, and the growing varieties of use cases that LLM can facilitate changes in the future of the machine learning market. GPT-3 was released in 2020, with more than 50 additional LLMs introduced by 2023. LLM development is continuing, and emerging models are consistently growing in size despite ongoing attempts to reduce model size and improve performance.

Modern LLM deployments tend to rely on Graphics Processing Units (GPUs). However, AI Processors for Cloud and Data Center Forecast Report – 2023 Analysis* published on October 19th, 2023, by Omdia reports that the supply of hardware GPU platforms is struggling to meet this escalating demand. This white paper proposes a CPU-based approach resolving this challenge and presents relevant benchmarking results and corresponding power efficiency across several platforms.

HARDWARE RESOURCES

GPUs represent the fastest-growing method for running LLM training and inferencing workloads. GPUs support large amounts of locally-attached memory that can load and process today's large LLMs.

Today's AMD Instinct™ and NVIDIA Tensor Cores* (A and H) GPUs are popular platforms for running LLM workloads. These GPUs include from 40 GB to 192 GB of High Bandwidth Memory (HBM) which delivers very high performance and very low latency, albeit at substantially higher cost than DDR memory, as you will see on the next page.

CALCULATING MEMORY SIZE FOR LLM MODELS

Training LLM models requires a substantial amount of memory. The Hugging Face Model training anatomy* provides the following baseline memory calculation for floating fp32 (4-byte) model:

- **Model weights:** 4bytes per parameter
- **Optimizer:** 8 bytes per parameter
- **Gradients:** 4 bytes per parameter
- **Activations and temporary memory:** 8 bytes per parameter (high end estimate)

Training this fp32 model requires approximately 20 extra bytes of memory per model parameter beyond the 4 bytes needed for the model parameter itself. The 1B model may therefore require up to 24 GB of memory space if no quantization is used. Using fp16 for the same 1B model requires 12 GB of memory for training. Fortunately, many of the latest models are trained for fp16 or bfloat16 that use much smaller memory footprints than fp32.

Inference does not use optimizers and gradients, meaning that the system does not need additional memory. Thus, the system needs 44 bytes for fp32 plus activation memory. The latest GPT-3 model has 175 billion parameters. Newer models, such as GPT-4, have approximately 1.8 trillion parameters. These large-size models require correspondingly large amounts of memory.

Table 1 shows memory size requirements for different-sized models with different precision options for inference and training.

Model	Inference (float16)	Inference (int8)	Training (fp32)
7 B	14 GB	7 GB	168 GB
13 B	26 GB	13 GB	312 GB
30 B	60 GB	30 GB	720 GB
70 B	140 GB	70 GB	1680 GB

Table 1: Memory requirements for different-sized models

MEMORY PERFORMANCE AND PRICE CONSIDERATIONS

The size and speed of HBM memory varies by platform. Overall, HBM is much faster than DDR, albeit at much higher cost.

- HBM3 performance can reach 819 GB/s to 1075 GB/s
- HBM2e can support 460GB/s
- DDR memory for CPUs can reach several multiples of 10 GB/s per channel. However, using several simultaneous DDR memory channels can provide effective LLM performance.

Price is another important consideration: DRAM Market Monitor*, Q1-2024, Yole Intelligence, reports that DDR4 memory has an Average Sales Price (ASP) of roughly $2.00 per GB as of Q4, 2023. DDR5 memory is expected to cost twice as much as DDR4. HBM has an ASP premium of more than 500% compared to DDR4. Furthermore, the high demand for HBM from GPU and CSP vendors is pushing the price of HBM several times higher than the current market ASP. This trend is expected to continue as LLM expands to support more use cases.

TEST METHODOLOGY AND RESULTS

This section describes the models, systems, and methods AMD used to test the LLM performance of GPU-equipped versus CPU-only systems, how testing was performed, and the results.

MODELS

Table 2 lists the well-known models used for the testing described in this white paper.

Model	Paramter Size (M)	Batch Sizes
opt-350m	350	1/2/4/8/16/32/64/128
opt-1.3b	1,300	
opt-6.7b	6,700	
opt-13b	13,000	
Opt-30b	30,000	

Table 2: Models, parameter sizes, and batch sizes tested

BENCHMARKING SOFTWARE

CPU benchmarking was done using Deep Speed* with OpenVINO™ 2023.3*, and GPU was done using Deep Speed. The results presented below are the average of five runs.

SYSTEM CONFIGURATIONS

This section describes the CPU-only system (Table 3) and GPU (Table 4) used for testing.

Item	Description
CPUs	2 x AMD EPYC™ 9654
Frequency Base | Boost*	2.4 GHz | 3.7 GHz
Cores	96 per socket (192 per node)
Thermal Design Power	400 W
Base Power	137.918 W
L3 Cache	384MB per socket
Memory	768 GB (24 x 32 GiB DDR5 4800)
BIOS	RTI1007D
BIOS Options	default
OS	CentOS Stream 9
Kernel	5.14.0-370.el9.x86_64
OS Options	default
Base Docker image	openvino/DeepSpeed/ ubuntu20_dev:latest
Docker OS Version	Ubuntu® 20.04.6 LTS
Docker Kernel Version	5.14.0-370.el9.x86_64 - default
*EPYC-18: Maximum boost for AMD EPYC processors is the maximum frequency achievable by any single core on the processor under normal operating conditions for server systems.	

Table 3: AMD system configuration

Item	Description
GPU	NVIDIA H100
Cores	16K+ CUDA® cores
Frequency: Base | Boost	1080 MHz | 1785 MHz
Tensor Cores	640
Thermal Design Power	750 W
Base Power Watt	200 W
Memory Type	HBM2e
Memory	80GB
Host OS Version	CentOS Stream 9
Kernel Version	5.14.0-370.el9.x86_64

Table 4: NVIDIA GPU configuration

PERFORMANCE COMPARISON

Figure 1 shows the performance of the opt-350m opt-1.3b parameter models in tokens per second. The results of testing the opt-6.7b and larger models showed similar performance differentials between the GPU and CPU-only systems.

Figure 1: opt-350m and opt-1.3b performance in tokens per second

Figure 1 shows the NVIDIA H100 system clearly outperforms the AMD EPYC CPU-only system in tokens per second. Both systems perform better at large batch sizes. Overall, the AMD EPYC CPU-only systems delivers approximately half of the performance of the NVIDIA H100 system for the same model. This is a compelling result when one compares the high cost of NVIDIA H100 GPU servers against the cost of CPU-only systems. Further, the NVIDIA H100 could not run the opt-1.3b at a batch size of 128 because of the "'CUDA out of memory" issue—another potential benefit of running LLM workloads on CPUs.

LATENCY COMPARISON

Figure 2 shows the single-token latency at different batch sizes. As expected, latencies for large batch sizes are longer than for small ones. Both systems display similar latencies at smaller batch sizes, and the CPU-only system latency shows increasing latency as batch size increases.

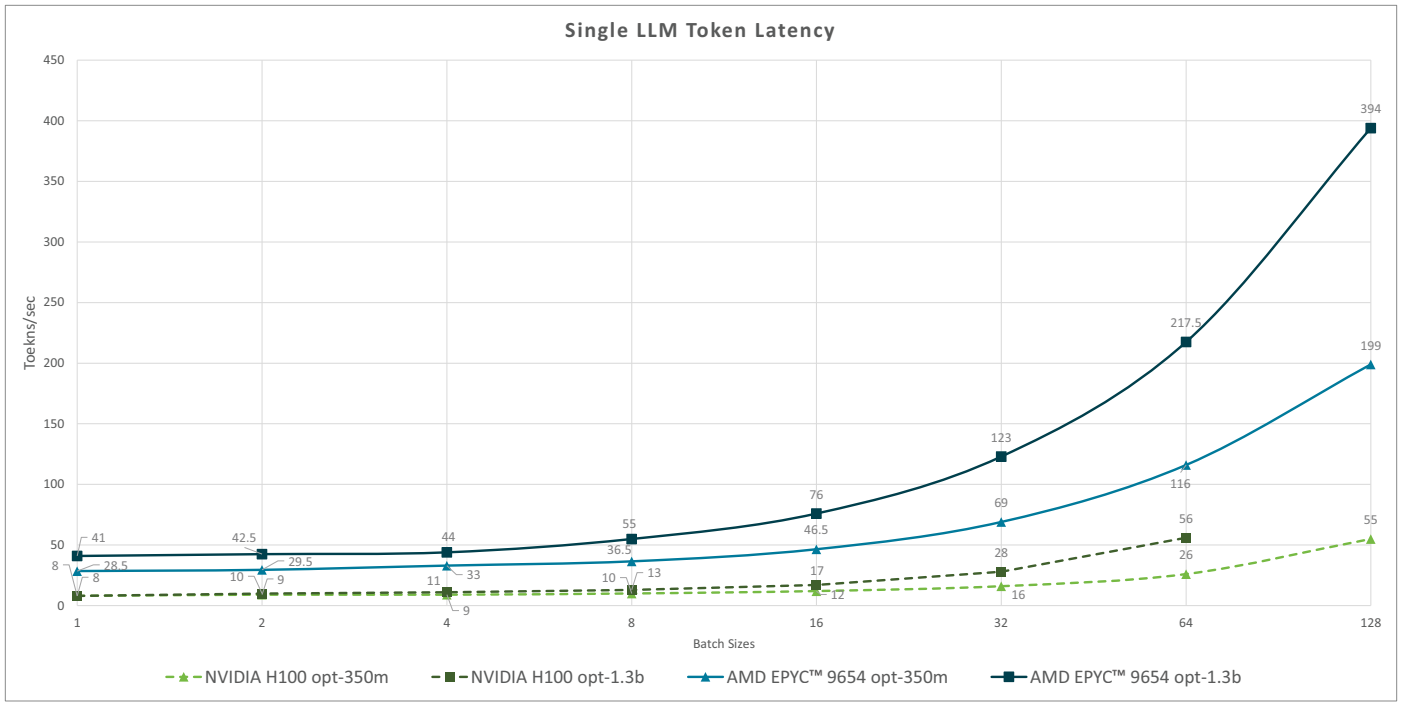

Figure 2: opt-350m and opt-1.3b latency in milliseconds

Latency-insensitive applications (primarily offline applications that perform tasks like summarizing and transcribing) are optimal candidates for CPU processing with performance reasonably close to that of GPUs. Testing also showed that a single NVIDIA H100 GPU can support model sizes up to 13B while the AMD EPYC CPU-only server can support models up to 66B. Adding more memory to a CPU-only system can further increase the supported model size.

POWER EFFICIENCY

In Figure 4, the horizontal (x) axis shows different batch sizes, and the vertical (y) axis shows the server energy consumption required to generate a single token for different batch sizes.

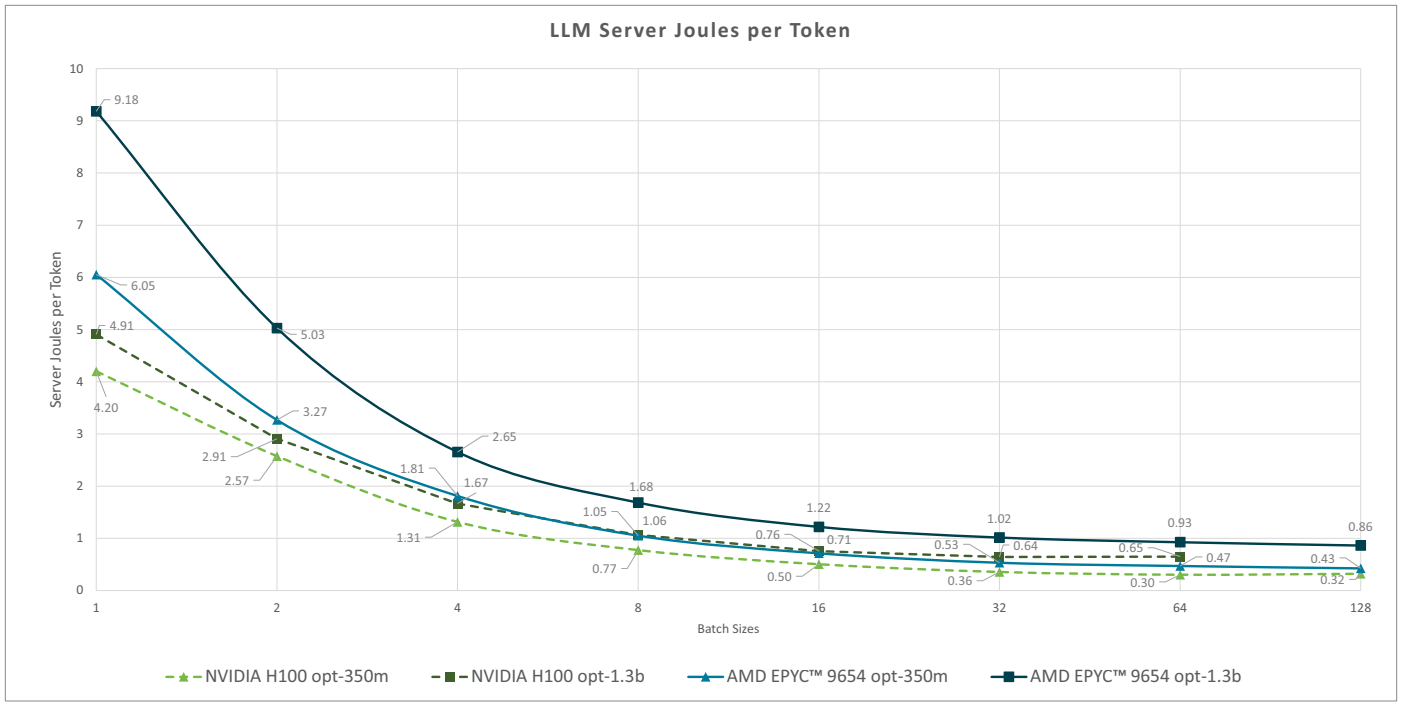

Figure 3: opt-350m and opt-1.3b power efficiency in joules per token

The average socket power in Watts was measured by Scaphandre v1.0.0 (PkgWatt metric) at five-second intervals for the duration of each test. Power consumption was then calculated using the following formula:

Server Joule per token = 1/(tokens/sec) × Total Server W (CPU+GPU)

In Figure 3, lower results are better because they indicate less energy required to generate a token. The power efficiency of the CPU-only system approaches that of the GPU at higher batch sizes.

PERFORMANCE PER UNIT GB OF MEMORY

The systems described in this white paper use different types of system memory:

- The AMD EPYC CPU-only system uses DDR memory.
- The NVIDIA H100 GPU uses HBM memory.

"Memory Performance and Price Considerations" on page 4 described the difference in price between DDR and HBM. Figure 4 shows the throughput achieved per unit memory cost spent. In other words, it shows how much performance one can get for each GB dollar spent on each platform. The substantially lower cost of DDR versus HBM means that the AMD EPYC CPU-only system delivers significantly higher performance per memory price than the NVIDIA H100 GPU.

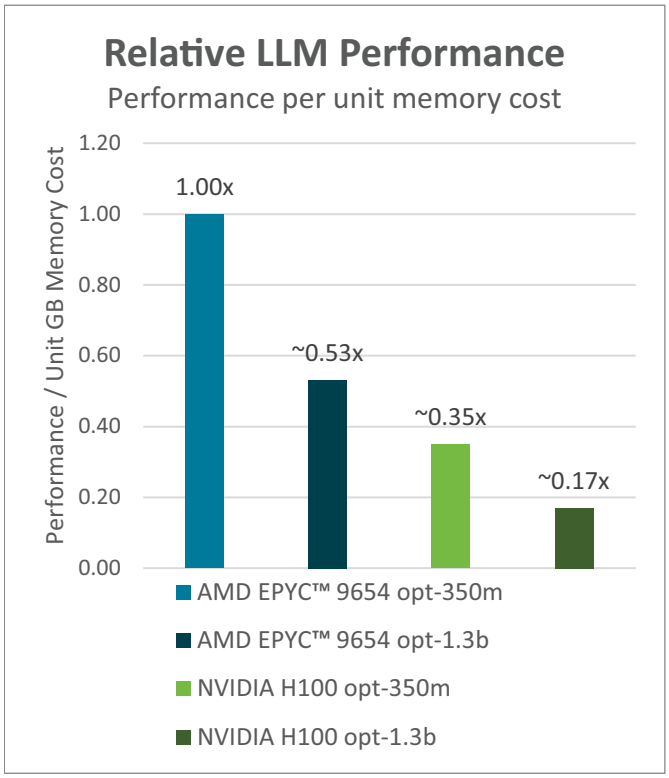

Figure 4: Price-performance per unit memory cost

CONCLUSION

This white paper presented the performance and memory price-performance results of tests carried out on NVIDIA GPU- and AMD EPYC CPU-based inference platforms running various-sized LLMs. The results demonstrate that CPU-based systems can deliver strong LLM performance at a lower cost per memory unit than GPU-based systems.

This is particularly true for workloads that are not sensitive to latency. Moreover, systems with a single GPU cannot complete larger batch operations because of the limited memory available, and adding GPUs further increases costs. By contrast, AMD EPYC CPU-based platforms supported much higher memory capacity and allowed running larger batches. This capability plus the significantly lower per-unit cost of DDR versus HMB make CPU-based servers ideal for LLM inference at optimized cost.

Seong Kim, Ph.D. is a Sr. Director of Cloud Solutions Architecture at AMD.

Xiaoqun Wang, Ph.D, works in Technical Marketing of Strategic AI Solutions at AMD

Yao Fu, Ph.D. is a PMTS, Strategic AI Solutions at AMD

DISCLAIMERS