DATA CENTER SOLUTIONS | WHITE PAPER

TECHNICAL INFO

AI INFERENCING WITH AMD EPYC™ PROCESSORS

together we advance_AI

June 2024

AI INFERENCING WITH AMD EPYC™ PROCESSORS

CONTENTS

AI IS PERVASIVE

Everywhere you look, artificial intelligence powers our business world. Once an aspirational endeavor, vast leaps in computer power have turned hopes into reality. Information Technology (IT) organizations recognize how ubiquitous AI has become and how easily their businesses can be left behind if they don't begin to use it to innovate in their business. This is why AI is used across areas including commercial and enterprise, cloud data centers, transportation, smart retail, healthcare and life sciences, smart homes, intelligent factories, smart cities, and communication service providers.

AMD PROPELS THE AI LIFECYCLE

AMD EPYC processors accelerate your entire AI journey, making it easier by providing a platform for data center consolidation, a host processor for GPU-accelerated machine learning, and an efficient processor for AI inferencing:

- **MAKE ROOM FOR AI.** AMD EPYC processors help you modernize your data center and make room for servers dedicated to AI model training. For example, to deliver 2000 virtual machines, each with one core and 8 GB of main memory, you can reduce the number of servers you need by 35% when you replace 2-socket servers using 60-core Intel® Xeon® 8490H CPUs with 96-core AMD EPYC 9654 processors. What used to take 17 servers now takes 11, giving you room to expand into AI/ML.[SP5TCO-049]
- **USE AMD EPYC CPU-POWERED SERVERS FOR GPU ACCELERATION.** Training is the most data- and processing-intensive part of the AI lifecycle. You have to sift through mountains of data to normalize it for the model you wish to train, and then let your GPU accelerators ingest all of your training data to move the process along. With up to 160 lanes of I/O bandwidth in 2-socket EPYC 9004 Series processor-powered servers, you have the bandwidth you need to train your model to arrive at the weights of parameters to make it function with the accuracy you need.
- **USE INDUSTRY-STANDARD CPUS FOR INFERENCING.** Once your model is trained, it needs a comparatively small amount of processing power to make real-time inferences based on live data. While models are trained at the beginning of the process and need a large amount of concentrated power, inferencing happens close to the data: in a retail store, in a moving automobile, on factory floors, in radiology departments. The locus of where computing power meets data is the point where efficiency is everything—and with 2-socket servers with 128-core EPYC 9754 processors, you gain ~53% more integer performance per estimated system watt over servers with 64-core Xeon 8592+ CPUs.[SP5-175A]

INFERENCING WITH OFF-THE-SHELF TECHNOLOGY

Off-the-shelf servers with AMD EPYC processors can handle inferencing across a wide range of areas including customer support, retail, automotive, financial services, medical, and manufacturing. The models used in these areas most often fall into the categories of computer vision, natural language processing, and recommendation systems.

These models can be accelerated by a library that is specifically optimized for AMD EPYC processors. The ZenDNN (Zen Deep Neural Network) plug-in enables you to easily move models from training into inferencing on AMD EPYC processor-powered servers. We offer software support throughout your AI lifecycle, inferencing in the core, cloud, and edge, and access to optimized models and software stacks that complement AMD hardware.

BROAD INDUSTRY IMPACT

Three different types of models have had a dramatic impact on business across multiple industries. Computer vision can recognize and classify objects, and also detect anomalies. Natural language processing, generative AI models, and retrieval-augmented generation (RAG) can help recognize text and speech, make meaning out of written words and respond with domain-specific knowledge from research results to help for customers. Recommendation systems can help predict everything from customer needs to anomalies in telemetry data. By focusing on accelerating these three model classes, you can reap the benefits regardless of your industry.

Automotive
Computer vision models help propel self-driving cars and also help recognize signage, pedestrians, and other vehicles to be avoided. Natural-language processing models can help recognize spoken commands to in-car telematics.

Retail
Automate checkout lines by recognizing products, or even create autonomous shopping experiences where the models link customers with the items they choose and put into their bags. Use product recommendation engines to offer alternatives, whether online or in the store.

Medical
Detect anomalies including fractures and tumors with computer vision models. Use the same models in research to assess in vitro cell growth and proliferation.

Manufacturing
Use computer vision models to monitor quality of manufactured products from food items to printed-circuit boards. Feed telemetry data into recommendation engines to suggest proactive maintenance: Are disk drives about to fail? Is the engine using too much oil?

Financial Services
AI-powered anomaly detection helps stop credit-card fraud, while computer vision models watch for suspicious documents including customer checks.

Service Automation
Where IT meets customers, natural-language processing can help take action based on spoken requests, and recommendation engines can help point customers to satisfactory solutions and product alternatives.

INFERENCING WITH AMD EPYC PROCESSORS

AI inferencing generally takes place close to the data, and servers with AMD EPYC processors are often there too, ready to take on the task. In retail environments, video streams can be processed for monitoring inventory with on-site edge servers. In manufacturing, assembly-line images of products can be inspected for defects. In medical imaging applications, hospitals already store and use images on their centralized servers. Financial and consumer services are generally centralized, so the data that needs to be analyzed is already collocated with servers in the data center often powered by AMD EPYC processors.

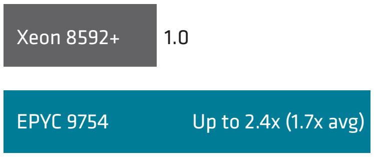
Whether at the core or at the edge, using servers with AMD EPYC processors for inferencing is an easy choice. AMD EPYC 9004 Series processors have made huge gains in areas that accelerate inferencing operations, enabling a hardware boost.

AMD EPYC 9004 SERIES ADVANCEMENTS

The impressive performance and energy efficiency of AMD EPYC processors provides a foundation for data center consolidation and security, making them an Ideal choice for small to midsize AI model inferencing. AMD EPYC 9004 Series processors further improve their aptitude for performing AI inferencing, including the following:

- **'ZEN 4' CORE:** Optimizations in the core architecture have resulted in an overall ~14% increase in instructions per clock cycle compared to our 3rd Gen processors. This improvement is based on the geometric mean of 33 integer, floating-point, and other server workloads.[EPYC-038]
- **POWER EFFICIENCY:** Carefully managing power consumption can help AI inferencing put less of a dent in power budgets. Because of the 'Zen 4' power efficiency, AMD EPYC processors power the most energy efficient servers, delivering exceptional performance and helping reduce energy costs.[EPYC-028D] On 2-socket servers with 128-core EPYC 9754 processors compared to those with 64-core Intel Xeon 8592+ CPUs, Select OpenVINO™ workloads performed inferencing with up to 2.4x maximum (1.7x average) the performance per CPU watt as the Intel processors.[SP5-252]

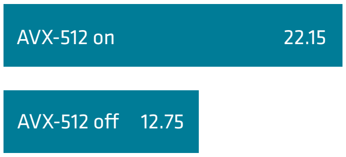
Select OpenVINO 2023.2 segmentation models
Normalized frames per second per CPU watt

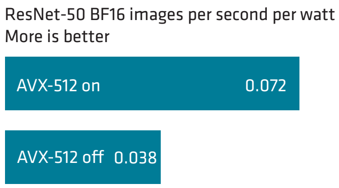


- **MORE CORES:** With a 50% increase in cores from a maximum of 64 in the prior generation to 128 in the current generation, more parallelism is available to power inferencing operations without GPU acceleration.
- **AVX-512 INSTRUCTION EXTENSION SUPPORT:** The 'Zen 4' core supports AVX-512 extensions. The VNNI component enables significant gains in AI inferencing performance across a range of data types. AVX-512 supports BF16 data types for improved throughput without having to risk the quantization challenges of using INT8 data. The AVX-512 implementation uses a two-cycle, efficient 256-bit pipeline, and thus could help maintain frequencies higher than with AVX-2. (Actual achievable frequency will vary depending on hardware, software, workloads, and other conditions.)[EPYC-039]

Phoronix measured the performance impact of the EPYC 9654 processor's AVX-512 instruction by running the ResNet-50 model with TensorFlow 2.10 and the BF16 data type. They ran the benchmark on the same server with AVX-512 turned on vs. off. They measured a 1.73x performance improvement with AVX-512 turned on, a minimal impact on clock frequency, and almost double the images processed per second per watt. This data demonstrates the efficiency of our AVX-512 implementation.

ResNet-50 BF16 Images per second
Higher is better





- **FASTER MEMORY WITH MORE CHANNELS:** The combination of DDR5 memory, plus 50% more memory channels (total of 12) in AMD EPYC 9004 Series CPUs yields a total of 2.25x the memory throughput compared to our prior generation.[EPYC-040]
- **FASTER I/O:** PCIe® Gen 5 I/O, again doubles AMD EPYC processor I/O throughput over the prior generation, enabling faster data ingest for AI inferencing.

SOFTWARE OPTIMIZATIONS

The three most popular frameworks for AI support the most common use cases, including image classification, large-language models, and recommendation engines:

- **TENSORFLOW:** This Google-owned platform focuses on training and inference of deep neural networks.
- **PYTORCH:** Originally developed by Meta, PyTorch was recently welcomed into the Linux® Foundation.
- **ONNX RUNTIME:** The Open Neural Network Exchange (ONNX), a Microsoft-sponsored platform.

We support a range of tools that make it easy to compile, run, and optimize performance of these frameworks:

- **MACHINE-LEARNING GRAPH COMPILER** that can handle graphs for all three frameworks
- **AMD OPTIMIZED CPU LIBRARIES (AOCL),** a set of numerical libraries optimized for the 'Zen' core architecture
- **ZENDNN,** a library of optimized AI primitives
- **ZEN SOFTWARE STUDIO,** including AMD Optimizing C/C++ (AOCC) and FORTRAN compilers
- **RUNTIME SOFTWARE** for both Microsoft Windows® and Linux

ZENDNN PLUG-IN

Our EPYC CPU-specific optimizations are embodied in a set of ZenDNN plug-ins that accelerate deep-learning inference workloads. It replaces basic AI primitives with equivalent operators that are tuned for AMD EPYC processors, plus compilation with optimizing tools such as torch.compile. It delves into comprehensive graph optimizations, including pattern identification, graph reordering, and graph fusion. At the operator level, ZenDNN boasts enhancements with microkernels, mempool optimizations, and efficient multithreading on the large number of 'Zen 4' cores. Microkernel optimizations exploit low-level math libraries that are optimized for AMD EPYC processors, including the AOCL BLIS library.

By optimizing at the primitive level, ZenDNN helps to accelerate a broad range of models across diverse application types. It doesn't matter whether you are using a computer vision model to recognize product defects, a natural language application to respond to customer prompts, or an engine to direct proactive maintenance, ZenDNN helps to accelerate a broad range of inferencing workloads.

While our goal is to upstream all our improvements for the EPYC platforms to the frameworks, we provide the optimizations today through the Plug-ins which you can download either directly from AMD or from PyPI repositories for zentorch and zentf. Additionally, we are working with Hugging Face to enable their models out of the box with ZenDNN.

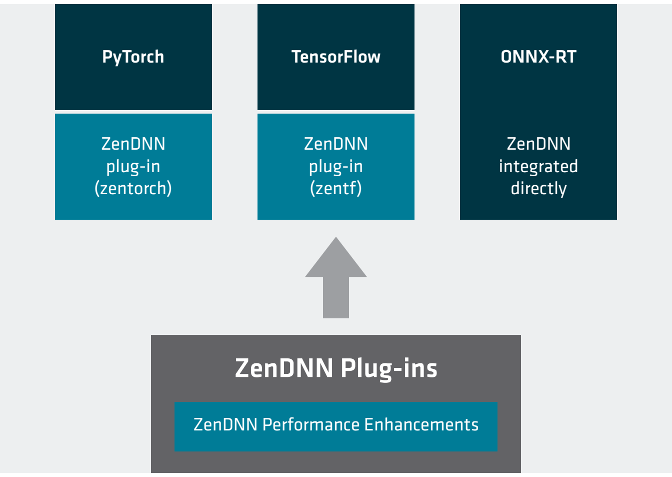
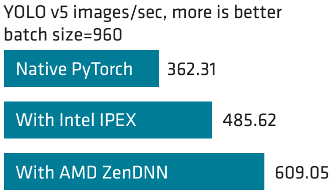

PyTorch
TensorFlow
ONNX-RT
ZenDNN plug-in (zentorch)
ZenDNN plug-in (zentf)
ZenDNN integrated directly
ZenDNN Plug-ins
ZenDNN Performance Enhancements

Figure 1: We provide ZenDNN plug-ins for each of the major frameworks in order to speed the latest optimizations to customers.

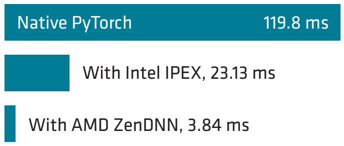
Because AMD EPYC processors support the x86 instruction set architecture, AI platforms can run with off-the-shelf code, but typically not as fast as code optimized for EPYC processors. To demonstrate the dramatic increase in performance, we assembled several PyTorch models as-is, with the Intel® Extension for PyTorch (IPEX), and with the ZenDNN plug-in. We tested inferencing using servers with two 96-core EPYC 9654 processors and observed increases in throughput and reductions in latency with the ZenDNN plug-in:

- **68% HIGHER IMAGE CLASSIFICATION THROUGHPUT**[ZD-052]



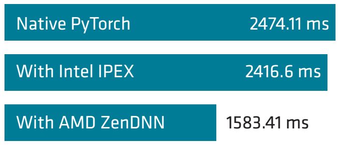
- **96.8% FASTER IMAGE RECOGNITION**[ZD-053]

ResNet50 BF16 latency, lower is better
batch size=1384



- **36% FASTER NATURAL LANGUAGE PROCESSING**[ZD-053]

Llama2 13B FP32, lower is better
batch size=1



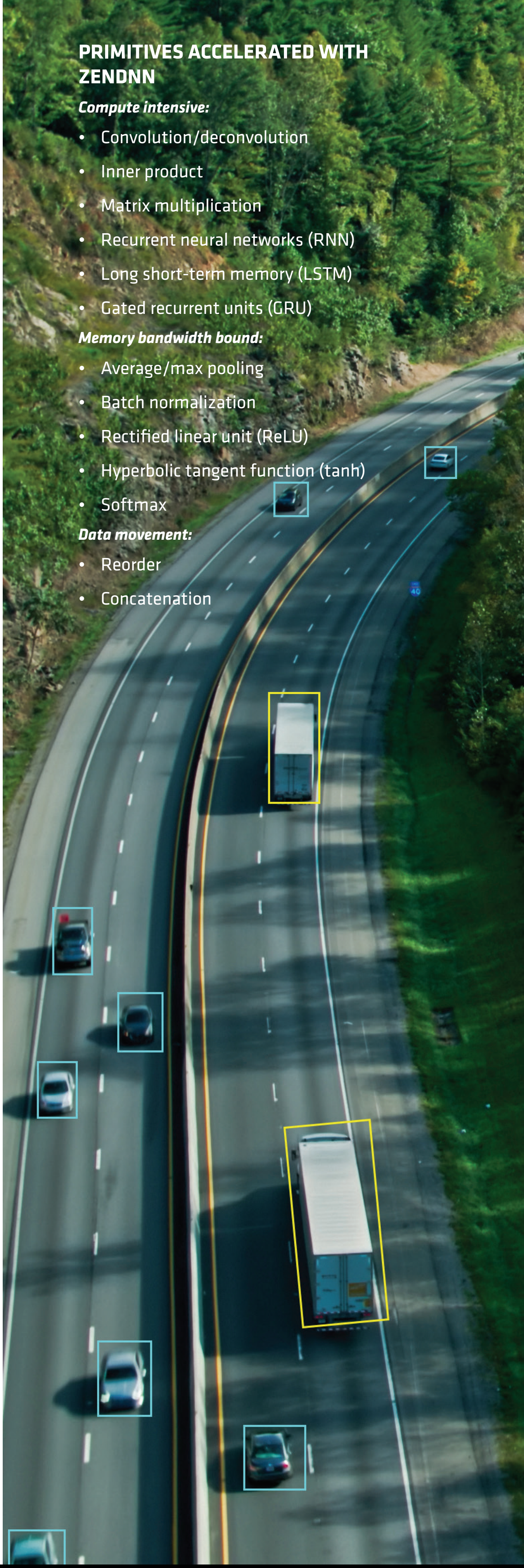
OVERALL PERFORMANCE PROOF

Performance measurements by AMD and third parties confirm the benefits of using AMD EPYC processors for inferencing. AMD EPYC processors deliver high parallelism with high efficiency, and enabling AVX-512 instructions that can speed processing. Compounding these benefits with the optimized ZenDNN library helps deliver even greater performance gains.

Comparing inferencing performance between servers with two 96-core EPYC 9654 processors to those with two 64-core Xeon 8592+ CPUs we find (see next page):

- Classification on random decision forests ~36% faster (SciKit-Learning RandomForest airline_ohe)[SP5-184A]
- Multi-Gate Mixture-of-Experts (MMoE) recommendations help predict customer behavior ~45% quicker (MMoE r1.15.5-deeprec230)[SP5-183A]
- Extreme gradient boosting with the Higgs boson data set runs ~70% faster (XGBoost 2.0.3)[SP5-251]
- Clustering dense vectors runs ~100% faster (FAISS v1.7.4)[SP5-185A]

The bottom line is that when you are considering using server CPUs for AI inferencing, AMD EPYC processors can deliver the performance you need with superb energy efficiency characteristics.



PRIMITIVES ACCELERATED WITH ZENDNN

Compute intensive:

- Convolution/deconvolution
- Inner product
- Matrix multiplication
- Recurrent neural networks (RNN)
- Long short-term memory (LSTM)
- Gated recurrent units (GRU)

Memory bandwidth bound:

- Average/max pooling
- Batch normalization
- Rectified linear unit (ReLU)
- Hyperbolic tangent function (tanh)
- Softmax

Data movement:

- Reorder
- Concatenation

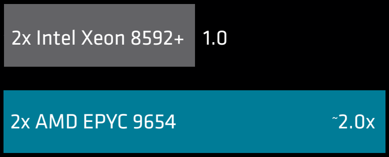
AMD EPYC PROCESSORS WITH ZENDNN PLUG-IN PERFORMANCE PROOF

OUR DATA SHOWS THE DRAMATIC SPEEDUP ON POPULAR WORKLOADS WHEN YOU USE SERVERS WITH AMD EPYC PROCESSORS

UP TO ~100%

MORE CLUSTERING PER SECOND[SP5-185A]

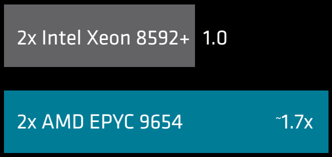
Clustering FAISS
Normalized througput



UP TO ~70%

MORE EXTREME GRADIENT BOOSTING[SP5-251]

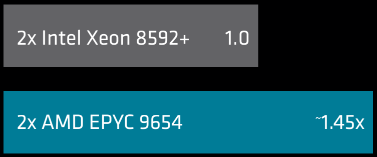
XGBoost with Higgs boson particle explosion
Normalized throughput



UP TO ~45%

FASTER RECOMMENDATIONS[SP5-183A]

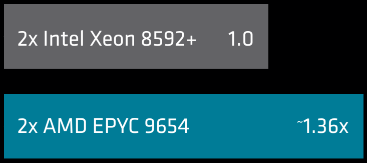
DeepRec multi-gate mixture of experts (MMoE)
Normalized throughput



UP TO ~36%

RANDOM DECISION FOREST SPEEDUP[SP5-184A]

Classification–Random Forest
(airline_ohe workload)
Normalized throughput



WE HAVE YOU COVERED

Your AI models are trained infrequently—but inferencing happens every day, minute by minute, across your business. Inferencing needs to be close to your customers and it must deliver the high performance and energy efficiency that can help AI transform your business.

Servers powered by AMD EPYC processors provide an excellent platform for CPU-based AI inferencing. With performance propelled by an energy-efficient AVX-512 implementation across up to 128 cores of processing power, and an optimized library whose primitives drive the processor to deliver its might to your solutions, it's hard to find a better solution. With ZenDNN performance enhancements to optimize the top inferencing models, you can choose your model and take advantage of the best that AMD has to offer.

END NOTES

For details on the EPYC footnotes used in this document, visit amd.com/en/legal/claims/epyc.html.

EPYC-028D SPECpower_ssj® 2008, SPECrate®2017_int_energy_base, and SPECrate®2017_fp_energy_base based on results published on SPEC's website as of 2/21/24. VMmark® server power-performance / server and storage power-performance (PPKW) based results published at https://www.vmware.com/products/vmmark/results3x.1.html?sort=score. The first 105 ranked SPECpower_ssj®2008 publications with the highest overall efficiency overall ssj_ops/W results were all powered by AMD EPYC processors. For SPECrate®2017 Integer (Energy Base), AMD EPYC CPUs power the first 8 top SPECrate®2017_int_energy_base performance/system W scores. For SPECrate®2017 Floating Point (Energy Base), AMD EPYC CPUs power the first 12 SPECrate®2017_fp_energy_base performance/system W scores. For VMmark® server power-performance (PPKW), have the top 5 results for 2- and 4-socket matched pair results outperforming all other socket results and for VMmark® server and storage power-performance (PPKW), have the top overall score. See https://www.amd.com/en/claims/epyc4#faq-EPYC-028D for the full list. For additional information on AMD sustainability goals see: https://www.amd.com/en/corporate/corporate-responsibility/data-center-sustainability.html. More information about SPEC® is available at http://www.spec.org. SPEC, SPECrate, and SPECpower are registered trademarks of the Standard Performance Evaluation Corporation. VMmark is a registered trademark of VMware in the US or other countries.

EPYC-038 Based on AMD internal testing as of 09/19/2022, geomean performance improvement at the same fixed-frequency on a 4th Gen AMD EPYC™ 9554 CPU compared to a 3rd Gen AMD EPYC™ 7763 CPU using a select set of workloads (33) including est. SPECrate®2017_int_base, est. SPECrate®2017_fp_base, and representative server workloads. SPEC® and SPECrate® are registered trademarks of Standard Performance Evaluation Corporation. Learn more at spec.org.

EPYC-039 The AVX-512 frequency for a given model of the 4th Gen AMD EPYC™ processors are the typical average recorded during AMD internal testing while running an HPC FLOPs workload using AVX-512 binaries in performance determinism mode. AMD's implementation of AVX-512 uses 2-cycle, efficient 256b pipeline and thus could help maintain frequencies higher than AVX2. Actual achievable frequency will vary depending on hardware, software, workloads and other conditions.

EPYC-040 AMD EPYC 9004 CPUs support 12 channels of up to 4800 MHz DDR5 memory which is 460.8 GB/s of maximum memory throughput per socket. Prior generation of EPYC CPUs (7003 series) have a maximum 204.8 GB/s. EPYC 9004 CPUs have 2.25x the memory throughput per CPU. 460.8 ÷ 204.8 = 2.3x (2.25x) the max memory throughput.

SP5-175A SPECrate®2017_int_base comparison based on published scores from www.spec.org as of 01/03/2024. Comparison of published 2P AMD EPYC 9754 (1950 SPECrate®2017_int_base, 720 Total TDP W, 256 Total Cores, $30823 Est system $, 1046.5 est system W, 1.863 SPECrate®2017_int_base/est. System W, 0.063 SPECrate®2017_int_base/est. System $, https://www.spec.org/cpu2017/results/res2023q2/cpu2017-20230522-36617.html . Cost Breakdown: Chassis $2175, Memory $4468, Drives $380, Chipset $0. Watts Breakdown: Chassis 85W, Memory 240W, Drives 1.5W, Chipset 0W is 1.73x the performance of published 2P Intel Xeon Platinum 8592+ (1130 SPECrate®2017_int_base, 700 Total TDP W, 128 Total Cores, $27207 Est system $, 929.5 est system W, 1.216 SPECrate®2017_int_base/est. System W, 0.042 SPECrate®2017_int_base/est. System $, https://www.spec.org/cpu2017/results/res2023q4/cpu2017-20231127-40064.html . Cost Breakdown: Chassis $2175, Memory $1412, Drives $380, Chipset $40. Watts Breakdown: Chassis 85W, Memory 128W, Drives 1.5W, Chipset 15W [at 1.53x the performance/system W] [at 1.52x the performance/system $]. AMD 1Ku pricing and Intel ARK.intel.com specifications and pricing as of 01/03/2024. SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

SP5-183A MMoE r1.15.5-deeprec2306 global_steps/sec workload claim based on AMD internal testing as of 4/19/2024. 2P server configurations: 2P EPYC 9654 (96C/192T), BIOS 1006C (NPS=1, Power Determinism), 1.5TB (24x 64GB DDR5-4800), 2x Samsung MZQL21T9HCJR-00A07 1.7 TB, Ubuntu® 22.04.3 LTS running 12 instances/16 cores/instance scoring 28.86 median global_steps/sec is 1.45x the performance of 2P Xeon Platinum 8592+ (64C/128T), BIOS 1.4.4 (Profile=Maximum Performance), 1TB (16x 64GB DDR5-4800), Intel SSDPF2KE032T1O 3.2TB NVMe, Ubuntu 22.04.3 LTS running 8 instances/16 cores/instance scoring 19.97 avg. global_steps/sec. Results may vary due to factors including system configurations, software versions and BIOS settings.

SP5-184A: SciKit-Learning Random Forest v2023.2 airline_ohe data set throughput workload claim based on AMD internal testing as of 4/19/2024. 2P server configurations: 2P EPYC 9654 (96C/96T) BIOS 1006C (SMT=off, NPS=1, Power Determinism), 1.5TB (24x 64GB DDR5-4800), 2x Samsung MZQL21T9HCJR-00A07 1.7 TB, Ubuntu® 22.04.3 LTS running 12 instances/16 cores/instance scoring 166.8 median throughput is 1.36x the performance of 2P Xeon Platinum 8592+ (64C/64T), BIOS 1.4.4 (HT=off, Profile=Maximum Performance), 1TB (16x 64GB DDR5-4800), Intel SSDPF2KE032T1O 3.2TB NVMe, Ubuntu 22.04.3 LTS running 8 instances/16 cores/instance scoring 123.1 median throughput. Results may vary due to factors including system configurations, software versions and BIOS settings.

SP5-185A FAISS v1.7.4 1000 throughput workload claim based on AMD internal testing as of 4/19/2024. 2P server configurations: 2P EPYC 9654 (96C/96T), BIOS 1006C (SMT=off, NPS=1, Power Determinism), 1.5TB (24x 64GB DDR5-4800), Samsung MZQL21T9HCJR-00A07 1.92 TB, Ubuntu® 22.04.3 LTS running 8 instances/24 cores/instance scoring 39.6 median throughput is 2.04x the performance of 2P Xeon Platinum 8592+ (64C/64T), BIOS 1.4.4 (HT=off, Profile=Maximum Performance), 1TB (16x 64GB DDR5-4800), Intel SSDPF2KE032T1O 3.2TB NVMe, Ubuntu 22.04.3 LTS running 8 instances/16 cores/instance scoring 19.4 median throughput. Results may vary due to factors including system configurations, software versions and BIOS settings.

SP5-251 XGBoost 2.0.3 throughput workload claim based on AMD internal testing as of 4/19/2024. 2P server configurations: 2P EPYC 9654 (96C/192T), BIOS 1006C (SMT=off, NPS=1, Power Determinism), 1.5TB (24x 64GB DDR5-4800), Samsung MZQL21T9HCJR-00A07 1.92 TB, Ubuntu 22.04.3 LTS scoring 203 Airline median throughput (running 12 instances/16 cores/instance) and 2057 Higgs median throughput (running 6 instances/32 cores/instance) for 1.38x and 1.71x the performance, respectively, of 2P Xeon Platinum 8592+ (64C/128T), BIOS 1.4.4 (HT=off, Profile=Maximum Performance), 1TB (16x 64GB DDR5-4800), Intel SSDPF2KE032T1O 3.2TB NVMe, Ubuntu 22.04.3 LTS running 8 instances/16 cores/instance scoring 147 Airline median throughput and 4 instances/32 cores/instance scoring 1200 Higgs median throughput. Results may vary due to factors including system configurations, software versions and BIOS settings.

SP5-252 Third-party testing OpenVINO 2023.2.dev FPS comparison based on Phoronix review https://www.phoronix.com/review/intel-xeon-platinum-8592/9 as of 12/14/2023 of select OpenVINO tests: Vehicle Detection FP16, Person Detection FP16, Person Vehicle Bike Detection FP16, Road Segmentation ADAS FP16 and Face Detection Retail FP16. Testing not independently verified by AMD. Scores will vary based on system configuration and determinism mode used (Power Determinism used). OpenVINO is a trademark of Intel Corporation or its subsidiaries.

SP5TCO-049: As of 05/25/2023 based on AMD Internal analysis using the AMD EPYC™ Server Virtualization & Greenhouse Gas Emission TCO Estimation Tool - version 13.25 estimating the cost and quantity of 2P AMD 96 core EPYC™ 9654 powered server versus 2P Intel® Xeon® 60 core Platinum 8490H based server solutions required to deliver 2000 total virtual machines (VM), requiring 1 core and 8GB of memory per VM for a 3-year period. This includes VMware software license cost of $6,558.32 per socket + one additional software for every 32 CPU core increment in that socket. Environmental impact estimates made leveraging this data, using the Country / Region specific electricity factors from the '2020 Grid Electricity Emissions Factors v1.4 – September 2020', and the United States Environmental Protection Agency 'Greenhouse Gas Equivalencies Calculator. This scenario contains many assumptions and estimates and, while based on AMD internal research and best approximations, should be considered an example for information purposes only, and not used as a basis for decision making over actual testing. For additional details, see https://www.amd.com/en/claims/epyc4#SP5TCO-049"

ZD-052 Testing conducted internally by AMD as of 05/15/2023. 2P AMD EPYC 9654 (192 Total Cores, 1536GB Total Memory w/ 24x64GB DIMMs, 2x960GB SSD RAID 1, HT Off, Ubuntu® 22.04, GCC 11.4.0) powered server running AI benchmarks with ZenDNN Plugin for PyTorch 4.2.0 (zentorch) and IPEX 2.1.2 compared to Native PyTorch Compile 2.1.2 (torch.compile) throughput (Queries/images per second, average of 3 runs)

Test (Data Type, Batch Size, Sequence Length)	zentorch 4.2.0	PyTorch 2.1.2	IPEX 2.1.2	zentorch v PyTorch	zentorch v IPEX
yolov5 (FP32, 960, NA)	335.19	219.87	291.28	1.52	1.1
Bert-large (336M) (FP32, 96, 384)	29.78	22.19	22.2	1.34	1.34
Distilbert (67M) (FP32, 96, 384)	180	135.87	142.78	1.32	1.26
Roberta (125M) (FP32, 96, 384)	104.41	79.04	79.21	1.32	1.32
BART-Large (509M) (FP32, 96, 384)	21.27	16.48	20.26	1.29	1.05
OPT-350M (FP32, 96, 384)	23.15	18.23	18.25	1.27	1.27
GPT2-Medium (380M) (FP32, 96, 384)	25.31	20.33	18.78	1.24	1.35
BART-Base (139M) (FP32, 96, 384)	68.38	55.92	65.1	1.22	1.05
resnet50 (FP32, 960, NA)	978.19	801.19	801.69	1.22	1.22
OPT-1.3B (FP32, 96, 384)	6.65	5.46	5.46	1.22	1.22
Llama7b-chat (FP32, 96, 384)	1.13	0.93	0.94	1.22	1.2
Llama13b-chat (FP32, 96, 384)	0.59	0.49	0.5	1.2	1.18
GPT-J (FP32, 96, 384)	1.62	1.36	1.39	1.19	1.17
yolov5 (BF16, 960, NA)	609.05	362.31	485.62	1.68	1.25
BART-Base (139M) (BF16, 96, 384)	120.09	108.81	106.29	1.1	1.13
Distilbert (67M) (BF16, 96, 384)	154.24	144.18	159.47	1.07	0.97
BART-Large (509M) (BF16, 96, 384)	36.47	34.3	33.14	1.06	1.1
OPT-350M (BF16, 96, 384)	42.04	39.61	35.57	1.06	1.18
Bert-large (336M) (BF16, 96, 384)	47.17	44.51	48.37	1.06	0.98
Roberta (125M) (BF16, 96, 384)	153.4	144.93	159.62	1.06	0.96
GPT2-Medium (380M) (BF16, 96, 384)	40.65	39.37	41.05	1.03	0.99
OPT-1.3B (BF16, 96, 384)	12.53	12.23	11.55	1.02	1.08
GPT-J (BF16, 96, 384)	3.38	3.35	3.36	1.01	1.01

Results may vary based on factors such as software versions and BIOS settings.

ZD-053 ZD-053: Testing conducted internally by AMD as of 05/15/2023. 2P AMD EPYC 9654 (192 Total Cores, 1536GB Total Memory w/ 24x64GB DIMMs, 2x960GB SSD RAID 1, HT Off, Ubuntu® 22.04, GCC 11.4.0) powered server running AI benchmarks with ZenDNN Plugin for PyTorch 4.2.0 (zentorch) and IPEX 2.1.2 compared to Native PyTorch Compile 2.1.2 (torch.compile) has (Latency in ms, average of 3 runs):

Test (Data Type, Batch Size, Sequence length)	zentorch	PyTorch	IPEX
ResNet50 (BF16, 1, NA)	3.84	119.80	23.13
YOLO v5 (BF16, 1, NA)	16.41	65.96	34.13
BERT-Large (FP32, 1, 384)	83.00	119.80	120.59
Llama2 13B (FP32, 1, 384)	1583.41	2474.11	2416.63
GPT-J (FP32, 1, 384)	751.49	1203.59	1214.83