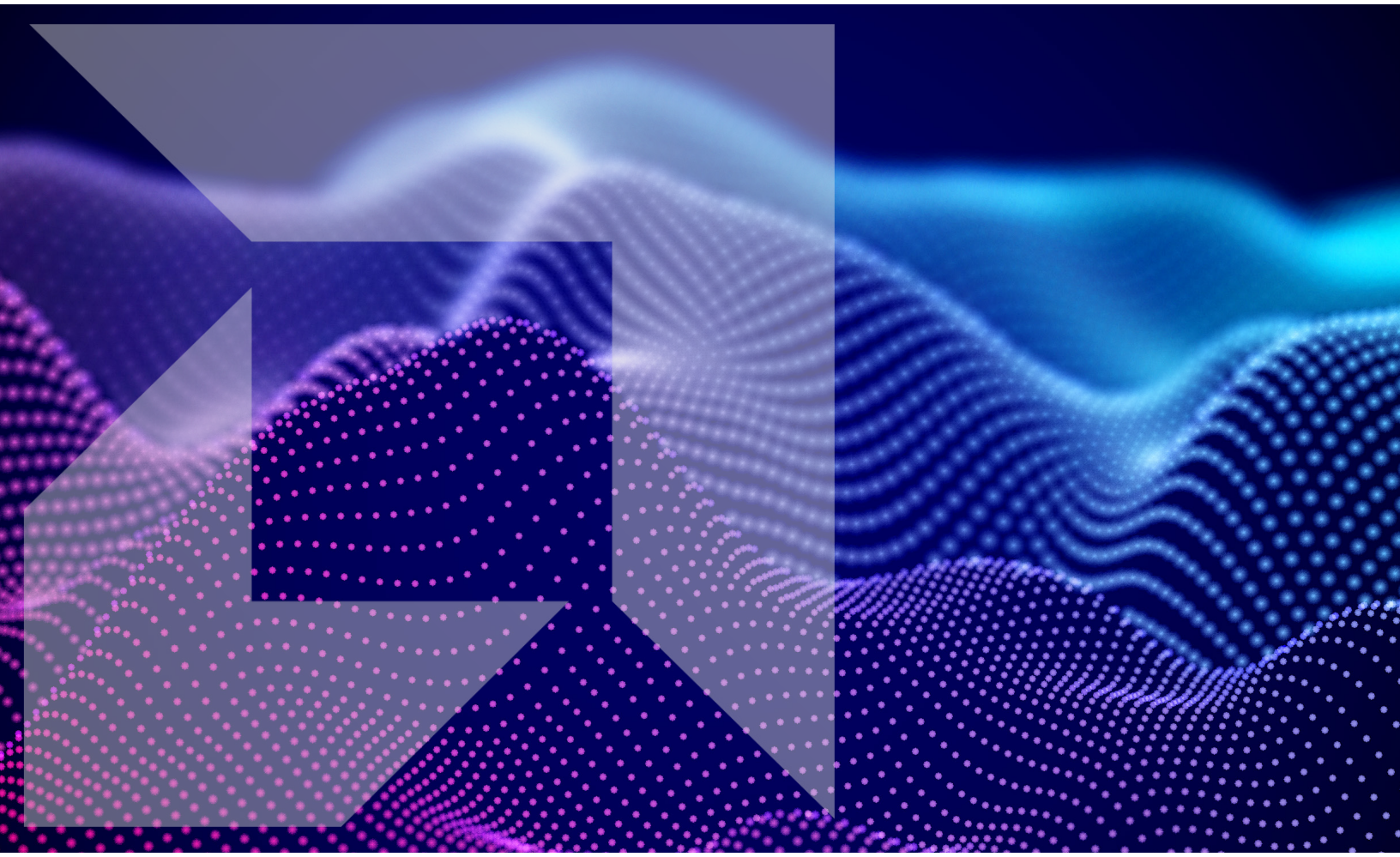


RETAIL'S AI REVOLUTION: ENHANCING CUSTOMER EXPERIENCE AND EFFICIENCY

TRANSFORMING RETAIL WITH IMPROVED CUSTOMER ENGAGEMENT AND DATA-DRIVEN INSIGHTS | 2025

TRANSFORMING RETAIL WITH AI AND MODERN IT: OVERCOMING CHALLENGES AND DRIVING GROWTH

Today's retail industry faces a fast-evolving landscape, with changing consumer behaviors, difficulties hiring and retaining retail employees, and supply chain and logistics issues. Behind the scenes, retailers race to confront the challenging realities of cost control, trend forecasting, and resource management.

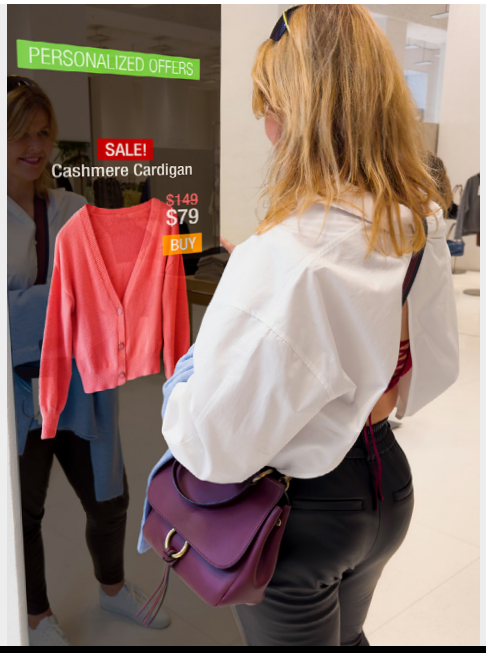
This whitepaper explores a new generation of innovative solutions that can help retailers not only meet and surmount these challenges, but also introduce new ways to delight customers, improve cost management, streamline operations, and increase revenues. Artificial intelligence (AI) and data collecting devices can provide insights into consumer behavior and operational efficiencies, transforming retail with improved customer engagement and personalization and improved protection of financial transactions.

Yet for these solutions to be implemented, some key challenges must be addressed. Most retailers operate on very thin margins, limiting their ability to invest in costly data collection devices and AI infrastructure powered by highly-sought-after, specialized GPU processors. For many retailers, data-driven AI innovation may simply seem not in the budget.

To overcome these challenges and realize the potential of AI, retailers need solutions that are cost effective and can modernize retail IT infrastructure to enable the deployment of transformative technologies. These solutions are possible with AMD, which can help improve the performance of your data center, cloud infrastructure, and edge deployments even as you consolidate to save money, space, and energy.

Read on to discover how AMD offers a comprehensive end-to-end portfolio of high-performance products and solutions to support AI deployments for retail and e-commerce operations. These include AI-optimized AMD EPYC™ CPUs and AI-accelerated AMD Instinct™ GPUs designed for deep learning training, inference, and large-scale AI models. AMD can help modernize the retail datacenter to bring in these new technologies. AMD also offers a suite of adaptive AI solutions for low-power, high-efficiency AI inference at the network edge, ideal for point-of-sale (POS) recommendation engines, autonomous checkout scenarios, and in-store inventory management and loss prevention.

You'll learn that AMD technologies can support your business as you consider how to thrive in a modern retail environment.



TECHNOLOGY ADOPTION IS NO LONGER OPTIONAL BUT ESSENTIAL FOR SURVIVAL IN THE MODERN RETAIL ENVIRONMENT.

TRANSFORMING RETAIL WITH AI

AI adoption in the retail sector is accelerating, driven by the need for increased efficiency, improved quality of customer experience, and enhanced competitive differentiation. According to the report **The Future of AI in Retail** from Everseen, 65% of responding consumers believe that AI makes retail more convenient. Nearly two-thirds (63%) of senior retail managers surveyed think AI deployment is important to maintain or gain a competitive advantage in their industry.

Retail AI adoption is underway, but slowly, as AI models integrated within computer vision, natural language processing, and recommendation systems are making their way into stores and online commerce. Chatbots and virtual assistants are common on e-commerce sites to provide product information and suggest relevant products based on analysis of customer behavior data. AI can support loss prevention through video analytics solutions, such as those offered by the Vadio AI vision platform. AI-powered POS solutions such as RadiusAI's ShopAssist system automate the traditional checkout process with instant product recognition to streamline transactions and provide a frictionless experience for the shopper.

Behind the scenes, AI is also used for automated inventory management to autonomously reorder products based on sales trends and predictive analytics. AI also powers supply chain optimization solutions that facilitate shipping routes and warehouse operations to reduce costs and improve efficiency.

EMERGING AI USE CASES CAN TRANSFORM THE RETAIL EXPERIENCE

On the horizon are many new examples of how AI can personalize the shopping experience for consumers and make it more efficient and effective for retailers. Leading-edge AI retail use cases include:

- **Computer vision to track customer behavior and offer recommendations and discounts.** AI-enhanced cameras and sensors in stores can monitor customer movements, product interactions, and shopping patterns to provide shoppers with real-time recommendations.
- **Hotspot analysis.** Computer vision can track customer movement to identify a store's pinch-points and high-traffic areas. It can also provide data that allows retailers to negotiate higher slotting fees with their suppliers.
- **Automating product detail pages (PDPs) in e-commerce with large language models (LLMs).** LLMs can generate SEO-optimized product descriptions instantaneously for online catalogs, greatly reducing manual efforts and improving sales opportunities in fast-changing retail segments.
- **Creating short-form product videos with generative AI.** Using product images and descriptions, retailers can auto-generate animated explainer videos, 360-degree product spins, or add voice-overs to narrate product features.

INTEGRATING AI, AUTOMATION, AND DATA-DRIVEN INSIGHTS INTO DAILY OPERATIONS GIVES RETAILERS THE OPPORTUNITY TO GATHER INSIGHTS AND OPTIMIZE CONNECTIONS TO THEIR CUSTOMERS.

THE CHALLENGES OF IMPLEMENTING DATA-DRIVEN AI SOLUTIONS IN RETAIL

For many retailers, the promise of AI-powered retail solutions may seem out of reach.

Retail AI solutions rely on processing massive amounts of customer data, transaction records, and behavioral information to generate the personalized insights that make tailored customer experiences possible. However, data collection, storage, and usage raise regulatory concerns in many locales. AI models themselves are also under increasing scrutiny, and retailers must ensure that personal information and data is securely encrypted and stored to prevent unauthorized access. Retailers are responsible, especially in the e-commerce segment, to comply with state, country, and EU level privacy regulations like GDPR and CDPR (in California).

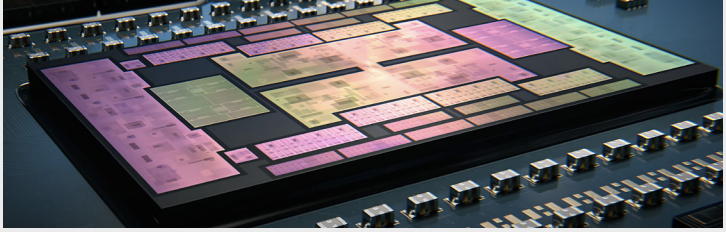
Aging IT infrastructure, or tech debt, is also a challenge for many retailers seeking to embrace new AI solutions. Many retailers run outdated systems, making it difficult to adopt modern AI-driven and cloud-based solutions such as demand forecasting, fraud detection, or customer personalization.

In addition, IT teams at retail businesses are not typically trained in AI and data science, a skills gap that can slow adoption and limit the impact of AI investments. Hiring AI and data science professionals to implement retail AI solutions is not straightforward due to high demand and talent shortages.

This is why it's critical for retailers to choose the right technology partner, with a large ecosystem of retail industry partners, to lead AI-driven implementations. AMD is a leader in the development of high performance and energy efficient processors with robust security features, and software infrastructure designed for AI workloads.

AMD is also at the center of a diverse network of retail industry leaders including OEMs, ISVs, resellers, and developers with deep experience implementing leading-edge AI solutions for the retail and e-commerce industries. AMD is in a commanding position to help guide retailers as they harness the potential of AI and data analytics to transform the retail landscape.

IT IS CRITICAL FOR RETAILERS TO CHOOSE THE RIGHT TECHNOLOGY PARTNER TO LEAD AI-DRIVEN IMPLEMENTATIONS.



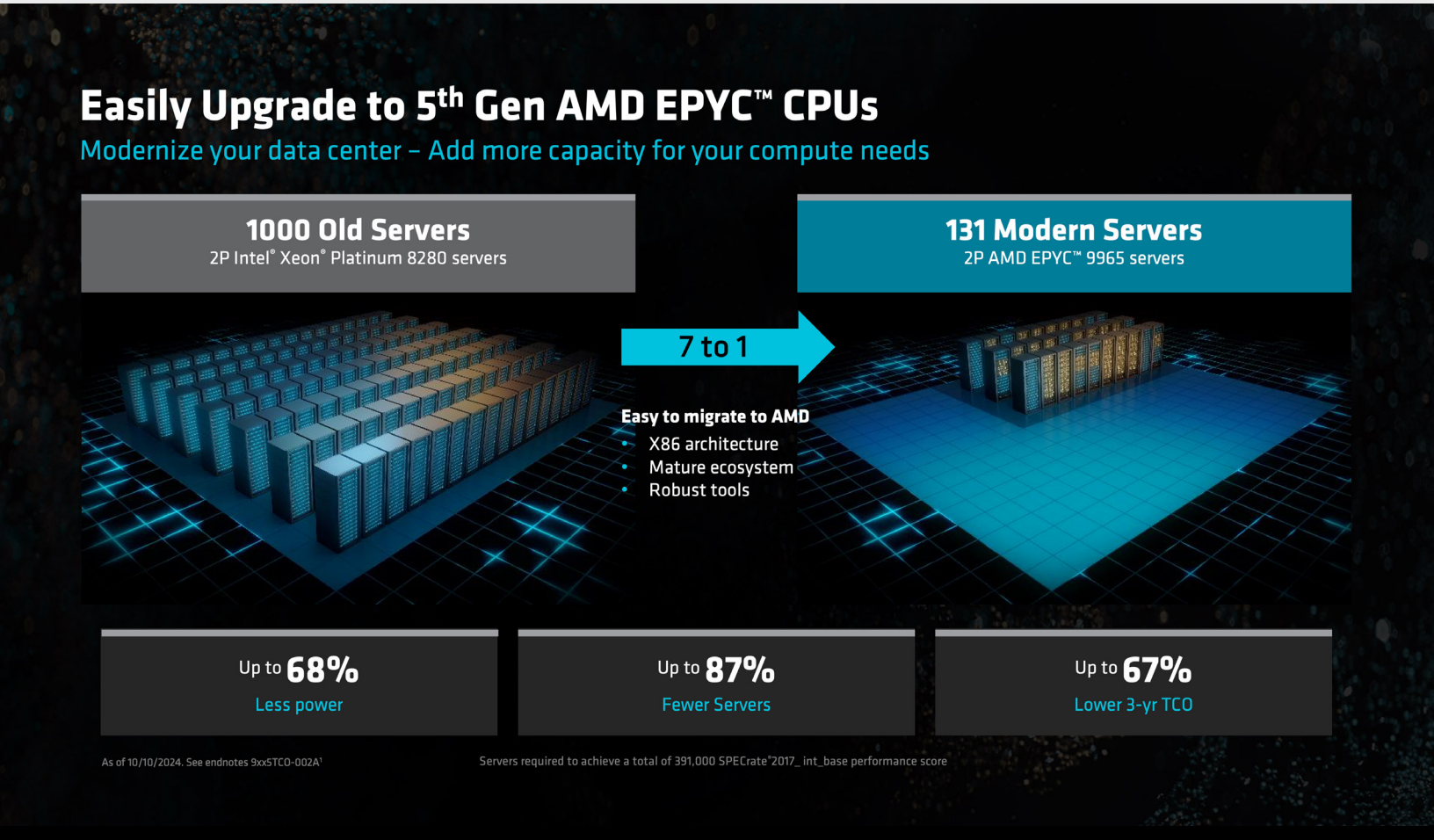
MAKING SMART CHOICES FOR RETAIL AI INFRASTRUCTURE

AI has the potential to transform the retail industry for both consumers and retailers, and early adopters in retail have already harnessed its power. Other retailers are at early or interim stages of integrating AI into their operations and face concerns about the considerable cost of AI investments, particularly related to the affordability of high-performance, high-demand GPUs.

AI workloads extend across a continuum that reflects a range of required "immediacy to insight." GPUs are at one end of the AI-processing spectrum. They are optimized for training large AI models and running large-scale inferencing workloads, and they can deliver insights in milliseconds, which is required for real-time decision-making. However, AI is not just a playfield for GPUs.

The latest generations of CPUs can handle a wide range of AI tasks alongside general-purpose workloads. AMD EPYC™ processors excel for small-to-medium AI models and workloads where proximity to data matters, offering high core counts and exceptional bandwidth. High density CPUs are more than capable of supporting a wide range of AI workloads that don't require massive parallelism, typically with lower upfront cost and power-efficiency. This makes them a cost-effective option for many AI applications including non-real-time generative AI workloads or solutions that determine dynamic pricing and for batch or offline processing applications where latency is not critical. AMD EPYC CPUs can also enable massive consolidation of IT infrastructure, significantly reducing the number of servers deployed in enterprise data centers and making room for next-generation AI-based infrastructure.

As AI model size, complexity, and volumes increase, GPU clusters can be added to the data center to deliver more performance per dollar. GPUs and CPUs can be integrated within the same server to handle different AI workloads most efficiently. AMD provides the end to end AI solution with AMD EPYC cpus and AMD Instinct GPUs.



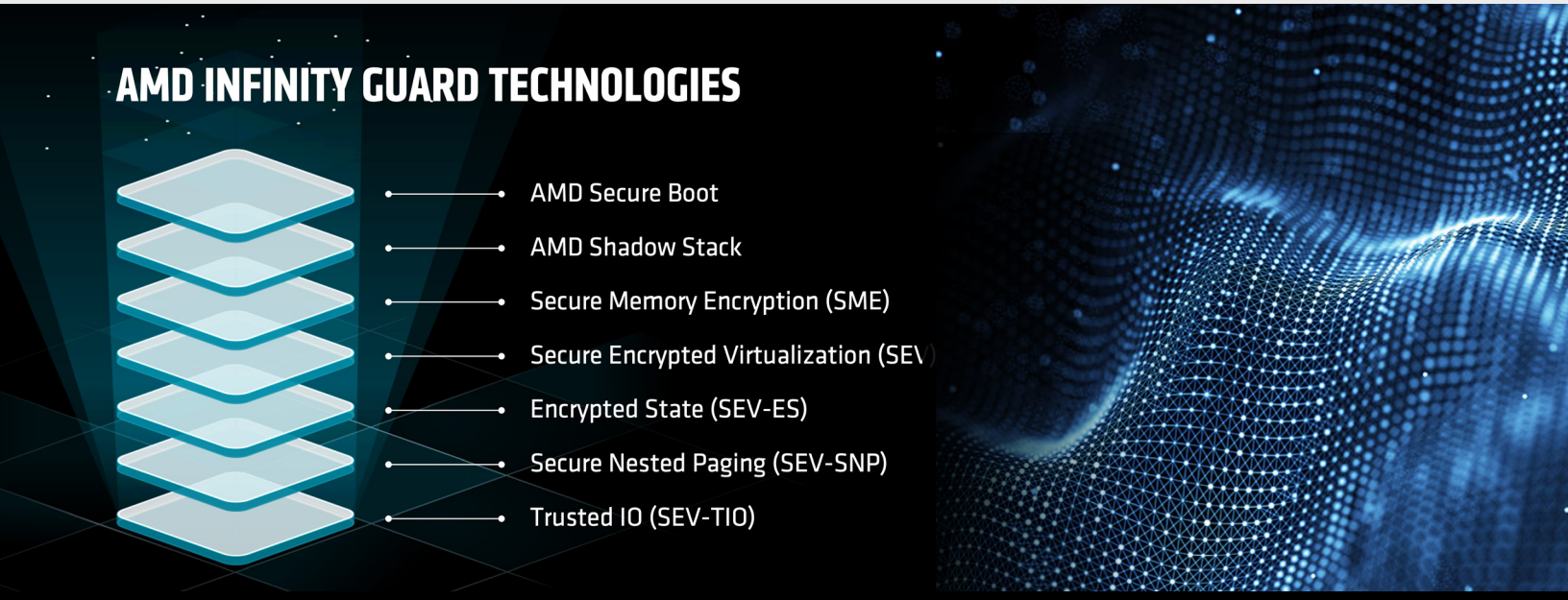
5TH GENERATION AMD EPYC™ 9005 PROCESSORS ARE A FOUNDATION FOR RETAIL AI INFERENCE

AMD EPYC 9005 processors provide an ideal CPU-based platform to power highly effective AI-driven retail operations and deliver operational efficiencies with advanced workload performance, impressive power efficiency, and low TCO. AMD EPYC 9005 processor-powered solutions drive the innovation that retailers need to power game-changing retail solutions that use data and AI to transform logistics/supply chains, distribution, marketing, customer loyalty, in-store interactions, and edge POS systems.

5th generation AMD EPYC 9005 processors offer high core counts, enabling fast processing by distributing AI workloads across multiple cores. High-speed I/O and memory bandwidth address the specific needs of retail workloads, such as creating cloud-enabled content delivery networks (CDNs) and managing data processing in edge systems. Edge computing is particularly relevant to retail use cases, as it can support a wide range of business applications, including POS systems, inventory management, and customer analytics.

AMD EPYC 9005 processors provide end-to-end AI performance by:

- **Maximizing Per-Server Performance.** AMD EPYC 9005 can match integer performance of legacy hardware with up to 86% fewer servers[2]—dramatically reducing physical footprint, power consumption - freeing up space for new or expanded AI workloads while helping reduce TCO.
- **Delivering Leadership AI Inference Performance.** Many AI workloads—language models with 13 billion parameters and below, image analysis, or recommendation systems—run efficiently on CPU-only servers that feature AMD EPY 9005 CPUs. Servers running two 5th gen AMD EPYC 9965 CPUs offer up to 2x inference throughput when compared to previous generation offerings.[3]
- **Maximizing GPU Acceleration.** The AMD EPYC™ 9005 family includes High Frequency EPYC 9005 processors which are excellent host CPUs for GPU-enabled systems to help increase performance on select AI workloads and can improve the ROI of each GPU server. For example, a high frequency AMD EPYC 9575F processor powered server with 8x GPUs delivers up to 20% greater system performance than a server with Intel Xeon 8592+ processors as the host CPU with the same 8x GPUs running Llama3.1-70B.[4]
- **Delivering Security at Depth.** AMD EPYC processors change to include AMD Infinity Guard, a sophisticated suite of system-level security features built-in at the silicon level, with advanced capabilities to help defend against internal and external threats to keep data safe. Infinity Guard provides defense in depth, with security controls at multiple layers, including secure boot and encryption, to help prevent threats. Infinity Guard also seamlessly integrates with most established security, compliance, and operations tools.[5]

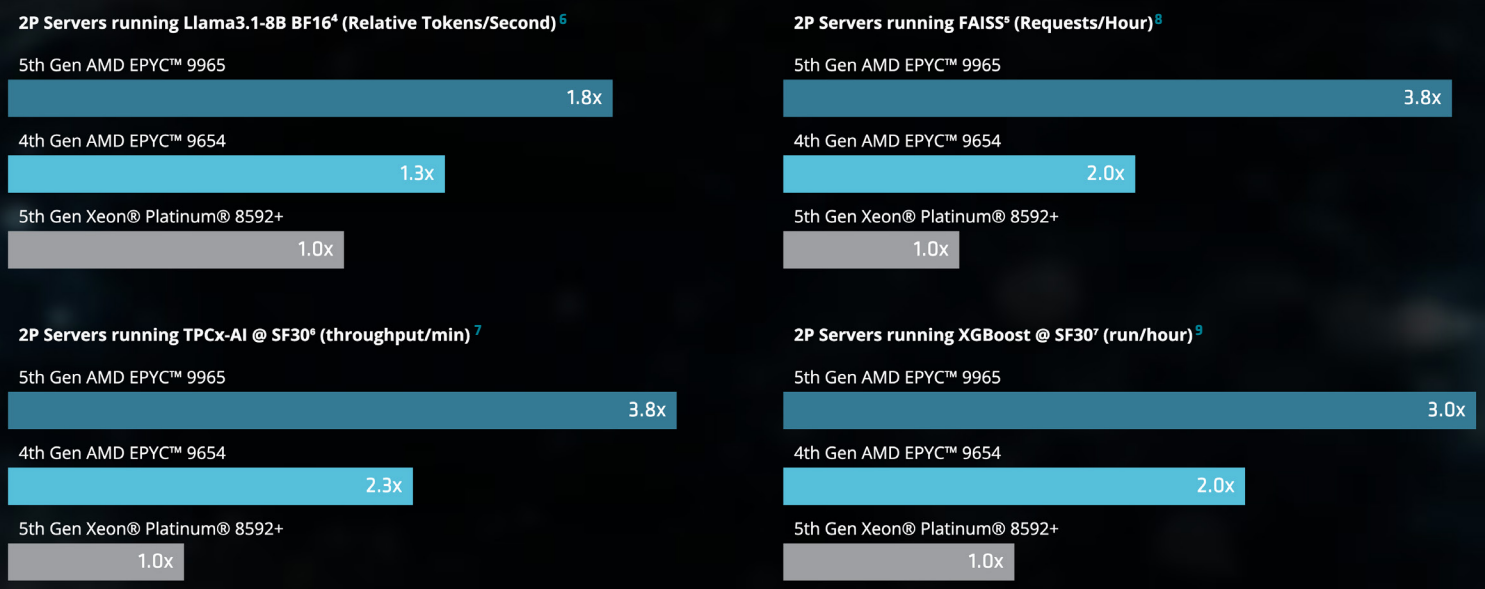


COST-EFFECTIVE HIGH PERFORMANCE WITHOUT LARGE BUDGET INCREASES

AMD EPYC™ processors-based servers deliver cost-optimized performance across a wide range of retail workloads and help lay the groundwork for rapid efficiency gains. This can lead to high performance-to-power ratios and low TCO that help users optimize their data centers and physical space and get more performance from the same or less data center power, space, or budget.

AI Workload Models

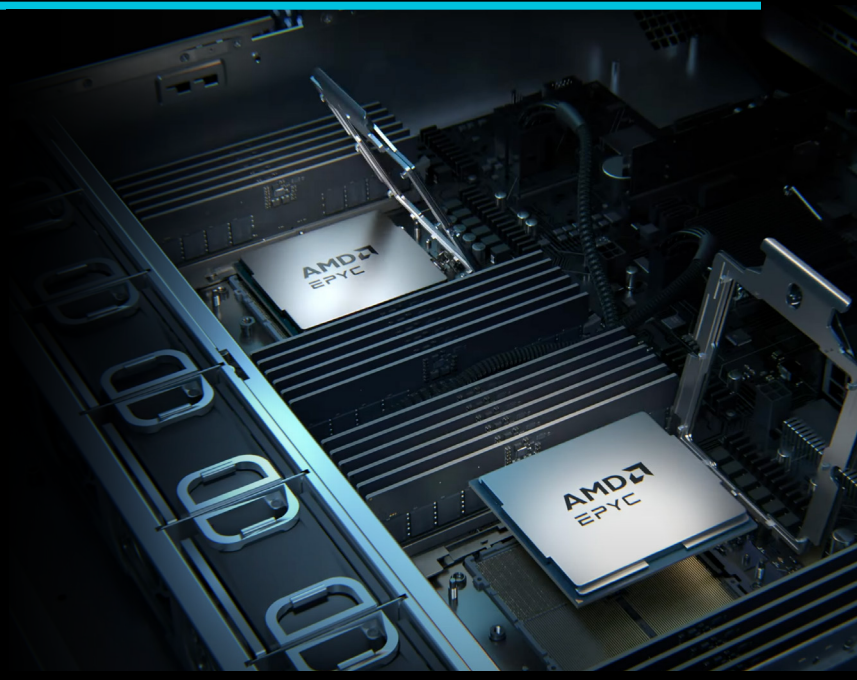
AMD EPYC™ 9005 processor-based servers and cloud instances enable fast, efficient, AI-enabled solutions close to your customers and data.



LEVERAGE A STRONG ECOSYSTEM

The AMD ecosystem is compatible with major AI and machine learning frameworks. Models built on frameworks like PyTorch or TensorFlow will run on AMD hardware without requiring major code changes, and AI workloads that were previously optimized for competitors' CPUs or GPUs typically operate on AMD EPYC processors without extensive reconfiguration.

In addition, AMD works closely with its OEM partners to help ensure that hardware co-optimization occurs both at the CPU and platform level. These OEMs also address physical security for data center infrastructure for many of its platforms, with features like locking lids and antitamper capabilities to prevent and detect when an unauthorized person attempts to interfere with a server.



CUSTOMER SUCCESS STORY: INFRASTRUCTURE REFRESH FOR A BIG BOX DIY RETAILER

A leading home improvement retailer experienced unprecedented demand during the COVID-19 epidemic, putting increased strain on its legacy in-store compute infrastructure. Due to this surge in business, store application response times almost doubled, making it untenable for store associates and consumers alike to have a good store experience. The company decided to refresh its seven-year-old infrastructure with a higher performance infrastructure with the capacity to modernize store applications and support new AI-based solutions—all while spending less than it did on the incumbent infrastructure.

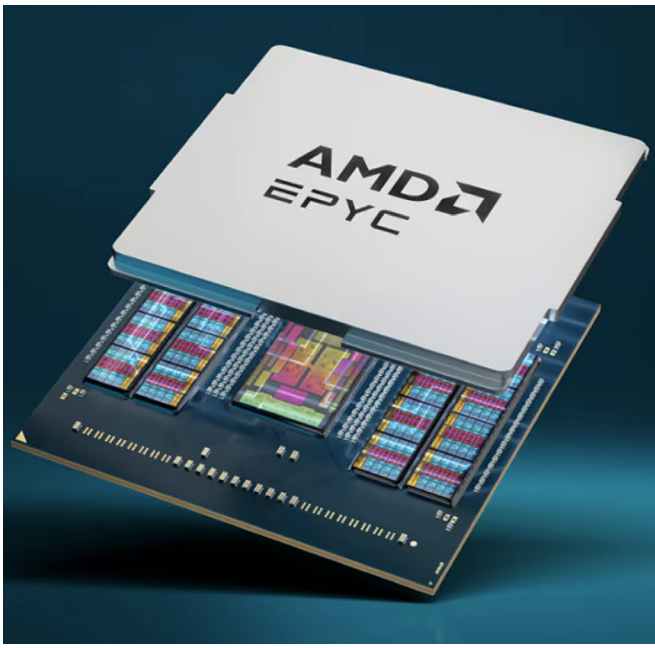
The retailer approached and ultimately selected AMD. They proposed an infrastructure refresh leveraging AMD EPYC CPUs, which provided massive performance advantages compared to existing alternatives, as well as considerable savings through server consolidation, aggressive performance gains, and a superior TCO. AMD's server platform partner also allowed integration of GPUs in a subset of this infrastructure so that the retailer could begin to pilot and roll out AI inferencing capabilities.

This infrastructure modernized the retailer's data centers and edge computing infrastructure in-store, delivering significant cost savings from compute consolidation, and reducing its server infrastructure from 8,800 servers to just 6,600 servers. The new AMD infrastructure also delivered superior TCO, substantial performance improvements, and future-proofing of its infrastructure to support emerging AI workloads.

COST-EFFECTIVE AMD EPYC 9005 PROCESSORS CAN DELIVER HIGH LEVELS OF PERFORMANCE WITH COST-EFFICIENT BUDGET INCREASES.

TRANSFORM YOUR RETAIL OPERATION WITH AMD

AI-driven innovations are transforming the retail industry as consumer expectations evolve, supply chains become more complex, and profit margins tighten. AI-powered retail solutions, including computer vision, dynamic pricing, recommendation engines, and automated inventory management, are increasingly adopted to address these issues. Solutions powered by AMD offer enhanced performance, efficiency, and cost-effectiveness, leveraging high-performance AMD EPYC™ CPUs to enable AI innovations designed for retail's unique needs. Embracing AMD technology enables retailers to boost productivity, make more informed decisions, and leverage AI to transform retail.



TO LEARN MORE:

- Visit **AMD Retail and E-Commerce Solutions**
- Contact your AMD representative for more information
- Email the AMD Retail and E-Commerce team with questions

For more information, **visit www.amd.com**

FOOTNOTES

1 9xx5TCO-002A: 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. The AMD Server & Greenhouse Gas Emissions TCO (total cost of ownership) Estimator Tool - version 1.12, compares the selected AMD EPYC™ and Intel® Xeon® CPU based server solutions required to deliver a TOTAL_PERFORMANCE of 391000 units of SPECrate2017_int_base performance as of October 10, 2024. This estimation compares a legacy 2P Intel Xeon 28 core Platinum_8280 based server with a score of 391 versus 2P EPYC 9965 (192C) powered server with a score of 3000 (https://www.spec.org/cpu2017/results/res2024q4/cpu2017-20240923-44837.pdf) along with a comparison upgrade to a 2P Intel Xeon Platinum 8592+ (64C) based server with a score of 1130 (https://spec.org/cpu2017/results/res2024q3/cpu2017-20240701-43948.pdf). Actual SPECrate®2017_int_base score for 2P EPYC 9965 will vary based on OEM publications. Environmental impact estimates made leveraging this data, using the Country / Region specific electricity factors from the 2024 International Country Specific Electricity Factors 10 – July 2024, and the United States Environmental Protection Agency 'Greenhouse Gas Equivalencies Calculator'. For additional details, see https://www.amd.com/en/claims/epyc4#SP9xxTCO-002A.

2 9xx5TCO-001C: 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. The AMD Server & Greenhouse Gas Emissions TCO (total cost of ownership) Estimator Tool - version 1.12, compares the selected AMD EPYC™ and Intel® Xeon® CPU based server solutions required to deliver a TOTAL_PERFORMANCE of 39100 units of SPECrate2017_int_base performance as of October 10, 2024. This scenario compares a legacy 2P Intel Xeon 28 core Platinum_8280 based server with a score of 391 versus 2P EPYC 9965 (192C) powered server with an score of 3000 (https://www.spec.org/cpu2017/results/res2024q4/cpu2017-20240923-44837.pdf) along with a comparison upgrade to a 2P Intel Xeon Platinum 8592+ (64C) based server with a score of 1130 (https://spec.org/cpu2017/results/res2024q3/cpu2017-20240701-43948.pdf). Actual SPECrate®2017_int_base score for 2P EPYC 9965 will vary based on OEM publications. Environmental impact estimates made leveraging this data, using the Country / Region specific electricity factors from the 2024 International Country Specific Electricity Factors 10 – July 2024 , and the United States Environmental Protection Agency 'Greenhouse Gas Equivalencies Calculator'.

3 9xx5-040A: XGBoost (Runs/Hour) throughput results based on AMD internal testing as of 09/05/2024. XGBoost Configurations: v2.2.1, Higgs Data Set, 32 Core Instances, FP32 2P AMD EPYC 9965 (384 Total Cores), 12 x 32 core instances, 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.4 LTS, 6.8.0-45-generic (tuned-adm profile throughput-performance, ulimit -l 198078840, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C (SMT=off, Determinism=Power, Turbo Boost=Enabled), NPS=1 2P AMD EPYC 9755 (256 Total Cores), 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198094956, ulimit -n 1024, ulimit -s 8192), BIOS RVOT0090F (SMT=off, Determinism=Power, Turbo Boost=Enabled), NPS=1 2P AMD EPYC 9654 (192 Total cores), 1.5TB 24x64GB DDR5-4800, 1DPC, 2 x 1.92 TB Samsung MZQL21T9HCJR-00A07 NVMe®, Ubuntu 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198120988, ulimit -n 1024, ulimit -s 8192), BIOS TTI100BA (SMT=off, Determinism=Power), NPS=1 Versus 2P Xeon Platinum 8592+ (128 Total Cores), AMX On, 1TB 16x64GB DDR5-5600, 1DPC, 1.0 Gbps NetXtreme BCM5719 Gigabit Ethernet PCIe, 3.84 TB KIOXIA KCMYXRUG3T84 NVMe®, Ubuntu 22.04.4 LTS, 6.5.0-35 generic (tuned-adm profile throughput-performance, ulimit -l 132065548, ulimit -n 1024, ulimit -s 8192), BIOS ESE122V (SMT=off, Determinism=Power, Turbo Boost = Enabled) Results: CPU Run 1 Run 2 Run 3 Median Relative Throughput Generational 2P Turin 192C, NPS1 1565.217 1537.367 1553.957 1553.957 3 2.41 2P Turin 128C, NPS1 1103.448 1138.34 1111.969 1111.969 2.147 1.725 2P Genoa 96C, NPS1 662.577 644.776 640.95 644.776 1.245 1 2P EMR 64C 517.986 421.053 553.846 517.986 1 NA Results may vary due to factors including system configurations, software versions and BIOS settings.

4 9xx5-14A: Llama3.1-70B inference throughput results based on AMD internal testing as of 09/01/2024.
Llama3.1-70B configurations: TensorRT-LLM 0.9.0, nvidia/cuda 12.5.0-devel-ubuntu22.04 , FP8, Input/Output token configurations (use cases): [BS=1024 I/O=128/128, BS=1024 I/O=128/2048, BS=96 I/O=2048/128, BS=64 I/O=2048/2048]. Results in tokens/second. 2P AMD EPYC 9575F (128 Total Cores) with 8x NVIDIA H100 80GB HBM3, 1.5TB 24x64GB DDR5-6000, 1.0 Gbps 3TB Micron_9300_MTFDHAL3T8TDP NVMe®, BIOS T20240805173113 (Determinism=Power,SR-IOV=On), Ubuntu 22.04.3 LTS, kernel=5.15.0-117-generic (mitigations=off, cpupower frequency-set -g performance, cpupower idle-set -d 2, echo 3> /proc/syss/vm/drop_caches), 2P Intel Xeon Platinum 8592+ (128 Total Cores) with 8x NVIDIA H100 80GB HBM3, 1TB 16x64GB DDR5-5600, 3.2TB Dell Ent NVMe® PM1735a MU, Ubuntu 22.04.3 LTS, kernel-5.15.0-118-generic, (processor.max_cstate=1, intel_idle.max_cstate=0 mitigations=off, cpupower frequency-set -g performance), BIOS 2.1, (Maximum performance, SR-IOV=On),
I/O Tokens Batch Size EMR Turin Relative 128/128 1024 814.678 1101.966 1.353 128/2048 1024 2120.664 2331.776 1.1 2048/128 96 114.954 146.187 1.272 2048/2048 64 333.325 354.208 1.063
For average throughput increase of 1.197x. Results may vary due to factors including system configurations, software versions and BIOS settings.

5 AMD Infinity Guard features vary by EPYC™ Processor generations and/or series. Infinity Guard security features must be enabled by server OEMs and/or Cloud Service Providers to operate. Check with your OEM or provider to confirm support of these features. Learn more about Infinity Guard at https://www.amd.com/en/technologies/infinity-guard. GD-183A.

DISCLAIMERS