A SIMPLIFIED ON-RAMP TO AI-POWERED SMART RETAIL

WHITEPAPER | 2026

Overcoming The Challenges That Slow Down Reinvention

Retailers are accelerating investment in AI, moving from pilot projects to strategic initiatives in an industry push to modernize store operations, embrace omnichannel, improve customer experience, and strengthen margins. Joint AI solutions from AMD and software partners streamline the path to value.

AI has become a core operational driver across the retail industry, with 77% of eCommerce professionals using AI daily and retailers that leverage AI reporting up to 15% annual revenue growth and 30% reduction in operating costs in 2025.[1] At the same time, retail lags other industries in adoption, despite prioritization of speed, automation, and real-time insight.[2]

This gap between ambition and execution reflects persistent challenges. Many retailers lack in-house expertise to overcome deployment and operational complexity, and they balk at the cost and risk of integrating multiple software and hardware components. Overcoming those challenges is critical to advance retailers toward AI's full business potential. AMD offers a streamlined path for retailers to use retail AI, powering tested and certified solutions designed to eliminate those expertise gaps, reduce complexity, and push down cost and risk.

A growing set of pre-integrated packages combines proven software and infrastructure, validated to work together smoothly and be rolled out consistently. The solutions are built to integrate with the AMD full portfolio of server-class CPUs and GPUs, for AI from data center to cloud to edge. They are built with edge servers based on AMD EPYC™ Server CPUs, which simplify retail AI by running efficient CPU-only inference that avoids complex, multi-component integrations. Server consolidation and power efficiency help lower TCO while validated solution stacks cut deployment friction and risk, helping retailers adopt AI quickly and confidently.

Certified retail AI solutions accelerate deployment and can reduce cost and risk.

A New Chapter In Retail Technology

AI has the potential to improve every major aspect of retail operations and strategy. It multiplies efficiency with services on the sales floor including shelf replenishment, loss prevention, visual self-checkout, and smart planograms as well as in non-customer-facing operations such as video analytics and inventory management.

In particular, AI computer vision is the foundation for visibility and insight into customer flow, staff activity, product interaction, and store conditions. These capabilities reduce friction for both customers and employees, while improving consistency across locations. Data-based planning and decision making are now more powerful than ever, with potential that is growing so fast that it can be hard to stay ahead of the curve.

Emerging technologies such as Generative AI and AI agents are further expanding what retail AI can do, by broadening the autonomy of solutions to reason, interpret, and respond. New models for in-store intelligence are giving retailers a deeper understanding of the business, stronger control over execution, and the ability to respond to issues before they escalate.

Growing data volumes are an inevitable requirement for advances in retail AI. Unlocking their value requires robust computing resources, with their potential maximized by pushing them to the edge, placing AI servers directly in stores and other distributed locations.

Processing data close to the sensors, cameras, and systems that generate it avoids the bandwidth costs and loss of responsiveness from transmitting back and forth to a data center or public cloud. It also reduces security exposure for payment information and other sensitive data by keeping it in-house instead of traversing public networks.

In this moment, edge-based AI is no longer optional to remain competitive. It is fast becoming the connective tissue of modern retail and a primary way to pursue competitive advantage.

Certified retail AI solutions accelerate deployment and can reduce cost and risk.

Greater Intelligence Without GPUs

GPUs excel at model training, large model inference with real-time responsiveness, and large scale workloads, all of which are critical to retailers managing fleets and doing backend analytics and planning across multiple locations. For retail environments where modest, sustained workloads and cost sensitivity prevail, GPUs may not quite fit. CPUs offer a more practical, scalable, cost-efficient foundation for retail AI that aligns with the operational realities of running in distributed retail locations.

CPU-based computing hardware fits into existing store infrastructure, runs multiple workloads concurrently, and makes distributed AI feasible at scale. These operations benefit from the predictable throughput, low power consumption, and simple deployment that CPUs offer. Deploying retail AI solutions with CPUs can reduce hardware investment, simplify rollout, reduce maintenance, and integrate seamlessly into the environments retailers already manage. AI-powered video analytics based on CPU processing are being adopted to track customer traffic patterns, POS exceptions, and transaction irregularities, analyze queue times, improve loss prevention, and more.

The advantages of CPUs over GPUs compound at scale with balanced performance, efficiency, and operational simplicity:

- **Comparatively lower power consumption and thermal requirements** can reduce operating expense and avoid costly electrical and HVAC upgrades, which is critical when deploying AI across hundreds or thousands of stores.
- **Ability to run multiple workloads concurrently** can consolidate multiple AI and non-AI workloads on fewer servers to simplify infrastructure and lower capital expense even further.
- **Compatibility with existing store infrastructure**, for easy deployment and reduced requirements in terms of specialized hardware or IT retraining.
- **Predictable performance for continuous inference**, helping improve reliability for mission-critical operations such as loss prevention and out-of-stock alerts.

The cost, power, and integration advantages of CPUs make them the optimal choice for retail AI. Beyond immediate requirements, the full x86 ecosystem of software and expertise provides the broadest scope of solutions and capabilities for future-ready success.

The CPU Advantage For Distributed Edge Retail AI

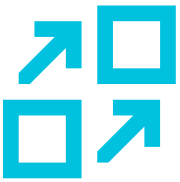
TABLE 1: CPUs VS. GPUs IN DISTRIBUTED EDGE RETAIL AI

FACTOR	CPUs (EDGE OPTIMIZED)	GPUs (COMPUTE AT SCALE)
POWER AND THERMAL FIT	Lower power requirements; runs with existing store facilities	Higher power requirements; may require colling or electrical upgrades
WORKLOAD MIX	Built for multiple, simultaneous AI and non-AI workloads	Optimized for single, heavy parallel workloads
DEPLOYMENT AT SCALE	Easy to roll out across thousands of stores	Higher cost and complexity prohibit distributed deployment at scale
OPERATIONAL OVERHEAD	Simple management; minimal specialized support	More complex to maintain; specialized skills required

CPUs are the AI inference hardware of choice for next-generation endpoint retail.

The Right Fit For Inference At The Edge

AMD EPYC Server CPUs are engineered to address the demanding mix of real-time inference, rapid data movement, and strict compliance requirements that drive retail AI forward. They make it possible for retailers to run reliable, high-performance, energy-efficient AI at the edge with outstanding TCO.



Throughput When It Counts

AMD EPYC Server CPUs drive **performance and scalability** with high compute density, including up to 128 cores in the AMD EPYC 9005 Server CPU series and 64 cores in the AMD EPYC 8004 Server CPU series. Run simultaneous inference across computer vision, GenAI, and Agentic AI workloads, avoiding bottlenecks. Optimize real-time analytics, self-checkout, supply-chain management, loss-prevention, and more to run smarter and drive innovative customer experiences at the edge, even in space-constrained store environments.

Lightning Fast Data Movement

Memory and I/O innovations **accelerate data handling** for real-time streams from cameras, sensors, and POS systems. Multi-channel DDR5 supports consistent high throughput, avoiding stalls when models, sensor streams, and in-store applications all compete for access. Flexible I/O lets retailers attach accelerators, high-speed storage, and dense networking seamlessly, creating a balanced system that scales cleanly as AI deployments grow.

Reduce Power Bills

High-efficiency CPUs can **lower operating costs** for retail AI deployments. High performance per watt makes it possible to reduce power draw and cooling needs across stores and distribution centers, even while expanding the scope of intensive AI inference workloads. The low-power silicon of AMD EPYC Server CPUs is tuned for sustained throughput in broad edge-AI deployment without expanding electrical or thermal footprints.

Security and Compliance in Depth

AMD EPYC Server CPUs **enhance security and trust** with AMD Infinity Guard, a suite of hardware protections that includes secure boot, memory encryption, and advanced isolation technologies. These features safeguard customer data, prevent tampering, and reduce exposure to certain firmware or supply-chain attacks. Aligning with major privacy and compliance frameworks enables retailers to deploy AI confidently while enabling strong trust and regulatory compliance.

Out-of-the-Box Compatibility

AI models and applications **migrate frictionlessly with minimal code changes** and broad compatibility across frameworks such as PyTorch, TensorFlow, and ONNX. Streamlined rollout of new services and pipelines can lower costs and integration risk. Combined with a mature software ecosystem and developer tooling, the stable, future-ready platform accelerates retail AI adoption with low, predictable operational overhead.

---

7 of the 10 Largest Fortune 100 Retailers Run on AMD EPYC Server CPUs.*

Driving Exceptional TCO Into Retail AI

The economics of data-processing infrastructure are critical for retail AI success. AMD EPYC Server CPUs are engineered for the realities of distributed deployment at any scale. The AMD EPYC 8004 Server CPU is purpose-built for power-efficient, always-on inference. The AMD EPYC 9005 Server CPU scales up AI throughput for heavy workloads such as processing large numbers of high-resolution video streams. The CPUs deliver TCO benefits wherever reliable, efficient AI inference is needed, including locations with constrained power, limited space, and no local IT presence.

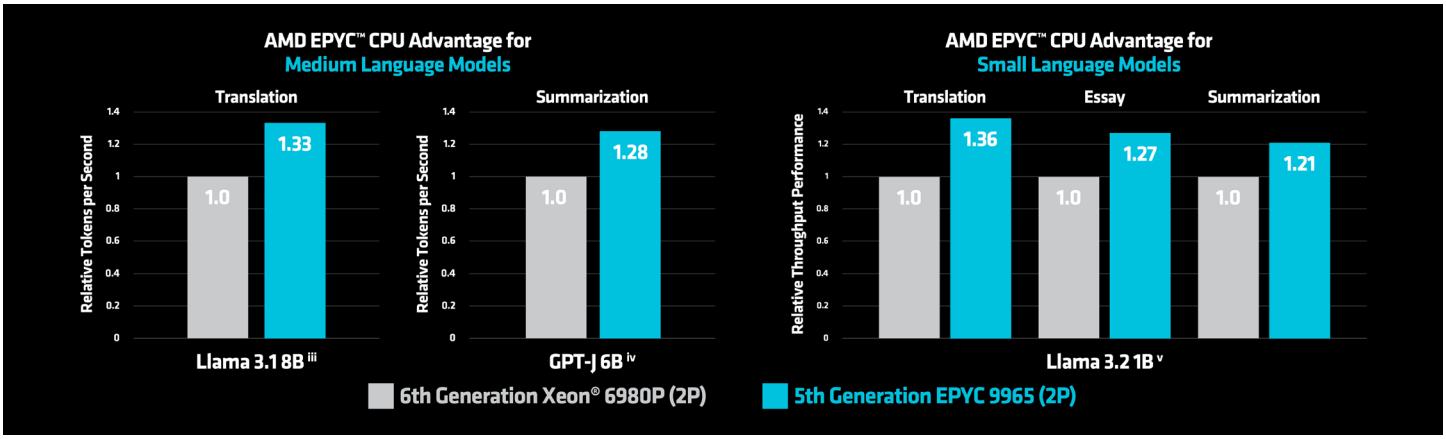
To optimize multi-pipeline AI, the AMD EPYC Server CPU's cores are optimized for high concurrency and throughput without the need for dedicated hardware accelerators. High core counts extend the ability to converge workloads onto fewer servers, and initial equipment cost savings can compound with operating costs over the life of the solution.

The compelling TCO story of AMD EPYC Server CPUs for retail AI extends to simplifying infrastructure management and integrating cleanly with existing IT processes. They can reduce the need for specialized skills, and hardware-based virtualization provides for smooth coexistence of AI, POS, back-office, and operational applications on shared hardware.

The processor's efficiency gains generate meaningful TCO outcomes that manifest in multiple advantages, for high performance per dollar that can include:

- **Fewer nodes per store** with server consolidation driven by high inference throughput.
- **Lower energy consumption** enabled by efficient CPU core microarchitecture.
- **Reduced cooling and infrastructure requirements** in existing constrained facilities.
- **Lower operational cost** through tested and certified deployment and single-socket, low-power designs.

AMD EPYC Server CPUs deliver the performance needed for real-time retail AI while keeping capital and operating costs under control in deployments across large retail networks.



Plug-And-Play Deployment Reduces Time, Cost, And Risk

Deploying AI across retail environments is notoriously complex, with each store carrying its own mix of hardware, software, and integration constraints. The growing portfolio of tested and certified AI solutions from AMD and solution partners eases the burden of multi-vendor engineering and speeds up time-to-value, while mitigating risk and TCO.

Preconfigured systems can be drop-shipped to stores, where they arrive ready to run. CPU-optimized stacks of hardware and software provide rapid onboarding and predictable performance. They can eliminate trial-and-error, reduce overhead, and provide retailers with a proven foundation to operationalize AI at scale. Here are many examples of common retail workloads where AMD has partnered with a solution provider to offer these preconfigured, ready to run systems from our solutions partners.

Acumera Reliant Platform: Cloud-Managed Edge

The Acumera Reliant Platform is a cloud-managed edge foundation for AI-driven store operations. It replaces fragmented hardware with a virtualized layer capable of running AI workloads alongside core networking and security services. Retailers deploy, update, and monitor applications from a single cloud console without on-site IT.

The platform maintains continuous availability across distributed stores, accelerates application rollouts, and streamlines maintenance, helping lower TCO for emerging retail AI usages. Built-in security and scalable controls help ensure reliable, protected deployments and give retailers a flexible foundation for introducing new digital experiences without re-architecting store infrastructure. Learn More >>*

Tested and Certified solutions optimize time to value and reduce adoption risk.

Radius ShopAssist™: Visual Self-Checkout

Radius ShopAssist™ enhances transaction speed and consistency in the checkout lane, replacing manual scanning with automatic item recognition. ShopAssist is built for real store conditions, where space is tight, labor varies by shift, and customers expect rapid service. It integrates with existing POS systems, allowing retailers to modernize lanes without redesigning counters or adding specialized equipment.

The platform continuously learns new SKUs and strengthens loss prevention by catching missed scans and reducing opportunities for error. ShopAssist fits into existing store operations to shorten queues, improve customer satisfaction, and increase the effective productivity of front-end labor. Learn more >>*

Radius ShopAssist™ Plus: Inventory Management

Radius ShopAssist™ Plus monitors fresh, prepared, and perishable inventory in real time, for continuous visibility into stock levels, freshness, and shelf life. By surfacing aging items, forecasting demand, and signaling replenishment needs, ShopAssist Plus helps reduce waste and ensure customer access to products.

The system increases the accuracy of production planning for bakery, deli, and grab-and-go programs. By grounding decisions in live data, retailers can right-size prep volumes and align labor with demand. Because ShopAssist Plus uses existing cameras and edge compute, it scales easily across large networks of stores. Learn more >>*

StorMagic SvHCI™: Full-Stack Hyperconverged Infrastructure

StorMagic SvHCI™ is a full-stack hyperconverged infrastructure solution optimized for distributed retail environments. Its mirrored two-node architecture delivers enterprise-grade high availability for AI workloads, POS systems, and in-store applications without the cost or complexity of traditional data center infrastructure.

Designed to run on standard server hardware and powered by AMD EPYC™ processors, SvHCI enables retailers to deploy edge intelligence quickly, standardize operations across locations, and maintain centralized control through a single management interface.

AI built for retail with results from day one.

Vaidio: Video Analytics

Vaidio transforms retail camera systems into real-time, actionable intelligence. By applying advanced Vision AI to existing infrastructure, retailers can reduce shrink, improve safety, optimize store operations, and enhance customer experience – all on a single, scalable platform. Unlock the power of your cameras to drive smarter retail performance. Learn more >>*

Wobot AI: Intelligent Computer Vision with Wobot AI: Edge-Based Video AI Agents for Retail Operations

Wobot AI deploys intelligent video AI Agents that run on AMD EPYC™ Server CPUs at the retail edge. These agents continuously observe live camera feeds, interpret customer and staff activity, and convert visual signals into structured operational intelligence. The platform analyzes customer flow, queue length, drive-through vehicle throughput, and in-store conditions to generate real-time alerts and actionable insights.

By running directly on AMD EPYC-powered edge servers, Wobot enables autonomous, bandwidth-efficient AI that operates inside each store. Retailers gain consistent, scalable execution across distributed fleets through use cases such as queue management, layout optimization, loss prevention signals, AI-powered staff task lists, and drive-through performance monitoring.

Activate AI with confidence and clarity.

Building Retail's Smarter Future

Even as AI adoption accelerates, retailers are limited by headwinds of cost sensitivity, limited expertise, deployment complexity, and integration risk. Ready-to-deploy, validated AI solutions based on AMD EPYC Server CPU-based servers and proven software offer a compelling solution.

As the optimal engine for AI workloads at the retail edge, AMD EPYC Server CPUs deliver outstanding TCO. AMD EPYC Server CPUs optimize that potential with high core density and performance, outstanding energy efficiency, and advanced hardware security features. Integrating proven AI software for retail with AMD EPYC Server CPU-based servers creates proven solution stacks that mainstream retailers can adopt today.

The tested and certified AI solutions arrive fully integrated and ready to run across every store, with rapid onboarding and predictable results. A new competitive edge is now available to the global retail industry, with the tools they need to thrive in the age of AI.

Explore future-ready retail solutions with an AMD expert

To Learn More

Visit the AMD Retail & E-Commerce Solutions page
Read the Retail AI Revolution whitepaper

***EPYC-059:** Top 10 U.S. retail companies by revenue according to 2025 Fortune 500 list as of June 2, 2025. https://fortune.com/ranking/fortune500. https://www.50pros.com/fortune500. 'Fortune 100' refers to the top 20% ranked companies in the 2025 Fortune 500 list, published in June 2025. From Fortune Magazine. ©2025 Fortune Media IP Limited. All rights reserved. Used under license. Fortune and Fortune Media IP Limited are not affiliated with, and do not endorse products or services of Advanced Micro Devices, Inc.

*Links to third party sites are provided for convenience and unless explicitly stated, AMD is not responsible for the contents of such linked sites and no endorsement is implied. GD-97.

ENDNOTES

[i] **Allaboutai, December 11, 2025.** "AI in Retail Statistics 2026: The $14.49B Market Transforming Global Commerce." https://www.allaboutai.com/resources/ai-statistics/ai-in-retail/.

[ii] **Retail Dive, December 23, 2025.** "A Look at Retail's Year of AI News." https://www.retaildive.com/news/a-look-at-retails-year-of-ai-news/807946/.

[iii] **9xx5-156:** Llama3.1-8B throughput results based on AMD internal testing as of 04/08/2025. Llama3.1-8B configurations: BF16, batch size 32, 32C Instances, Use Case Input/Output token configurations: [Summary = 1024/128, Chatbot = 128/128, Translate = 1024/1024, Essay = 128/1024]. 2P AMD EPYC 9965 (384 Total Cores), 1.5TB 24x64GB DDR5-6400, 1.0 Gbps NIC, 3.84 TB Samsung MZWLO3T8HCLS-00A07, Ubuntu® 22.04.5 LTS, Linux 6.9.0-060900-generic, BIOS RVOT1004A, (SMT=off, mitigations=on, Determinism=Power), NPS=1, ZenDNN 5.0.1 2P AMD EPYC 9755 (256 Total Cores), 1.5TB 24x64GB DDR5-6400, 1.0 Gbps NIC, 3.84 TB Samsung MZWLO3T8HCLS-00A07, Ubuntu® 22.04.4 LTS, Linux 6.8.0-52-generic, BIOS RVOT1004A, (SMT=off, mitigations=on, Determinism=Power), NPS=1, ZenDNN 5.0.1 2P Xeon 6980P (256 Total Cores), AMX On, 1.5TB 24x64GB DDR5-8800 MRDIMM, 1.0 Gbps Ethernet Controller X710 for 10GBASE-T, Micron_7450_MTFDKBG1T9TFR 2TB, Ubuntu 22.04.1 LTS Linux 6.8.0-52-generic, BIOS 1.0 (SMT=off, mitigations=on Performance Bias), IPEX 2.6.0 Results: CPU 6980P 9755 9965 Summary 1 n/a1.093 Translate 1 1.062 1.334 Essay 1 n/a 1.14 Results may vary due to factors including system configurations, software versions, and BIOS settings.

[iv] **9xx5-158:** GPT-J-6B throughput results based on AMD internal testing as of 04/08/2025. GPT-J-6B configurations: BF16, batch size 32, 32C Instances, Use Case Input/Output token configurations: [Summary = 1024/128, Chatbot = 128/128, Translate = 1024/1024, Essay = 128/1024]. 2P AMD EPYC 9965 (384 Total Cores), 1.5TB 24x64GB DDR5-6400, 1.0 Gbps NIC, 3.84 TB Samsung MZWLO3T8HCLS-00A07, Ubuntu® 22.04.5 LTS, Linux 6.9.0-060900-generic, BIOS RVOT1004A, (SMT=off, mitigations=on, Determinism=Power), NPS=1, ZenDNN 5.0.1, Python 3.10.12 2P AMD EPYC 9755 (256 Total Cores), 1.5TB 24x64GB DDR5-6400, 1.0 Gbps NIC, 3.84 TB Samsung MZWLO3T8HCLS-00A07, Ubuntu® 22.04.4 LTS, Linux 6.8.0-52-generic, BIOS RVOT1004A, (SMT=off, mitigations=on, Determinism=Power), NPS=1, ZenDNN 5.0.1, Python 3.10.12 2P Xeon 6980P (256 Total Cores), AMX On, 1.5TB 24x64GB DDR5-8800 MRDIMM, 1.0 Gbps Ethernet Controller X710 for 10GBASE-T, Micron_7450_MTFDKBG1T9TFR 2TB, Ubuntu 22.04.1 LTS Linux 6.8.0-52-generic, BIOS 1.0 (SMT=off, mitigations=on, Performance Bias), IPEX 2.6.0, Python 3.12.3 Results: CPU 6980P 9755 9965 Summary 1 1.034 1.279 Chatbot 1 0.975 1.163 Translate 1 1.021 0.93 Essay 1 0.978 1.108 Caption 1 0.913 1.12 Overall 1 0.983 1.114 Results may vary due to factors including system configurations, software versions, and BIOS settings.

[v] **9xx5-166:** Llama3.2-1B throughput results based on AMD internal testing as of 04/08/2025. Llama3.3-1B configurations: BF16, batch size 32, 32C Instances, Use Case Input/Output token configurations: [Summary = 1024/128, Chatbot = 128/128, Translate = 1024/1024, Essay = 128/1024]. 2P AMD EPYC 9965 (384 Total Cores), 1.5TB 24x64GB DDR5-6400, 1.0 Gbps NIC, 3.84 TB Samsung MZWLO3T8HCLS-00A07, Ubuntu® 22.04.5 LTS, Linux 6.9.0-060900-generic, BIOS RVOT1004A, (SMT=off, mitigations=on, Determinism=Power), NPS=1, ZenDNN 5.0.1, Python 3.10.2 2P Xeon 6980P (256 Total Cores), AMX On, 1.5TB 24x64GB DDR5-8800 MRDIMM, 1.0 Gbps Ethernet Controller X710 for 10GBASE-T, Micron_7450_MTFDKBG1T9TFR 2TB, Ubuntu 22.04.1 LTS Linux 6.8.0-52-generic, BIOS 1.0 (SMT=off, mitigations=on, Performance Bias), IPEX 2.6.0, Python 3.12.3 Results: CPU 6980P 9965 Summary 1 1.213 Translation 1 1.364 Essay 1 1.271 Results may vary due to factors including system configurations, software versions, and BIOS settings.