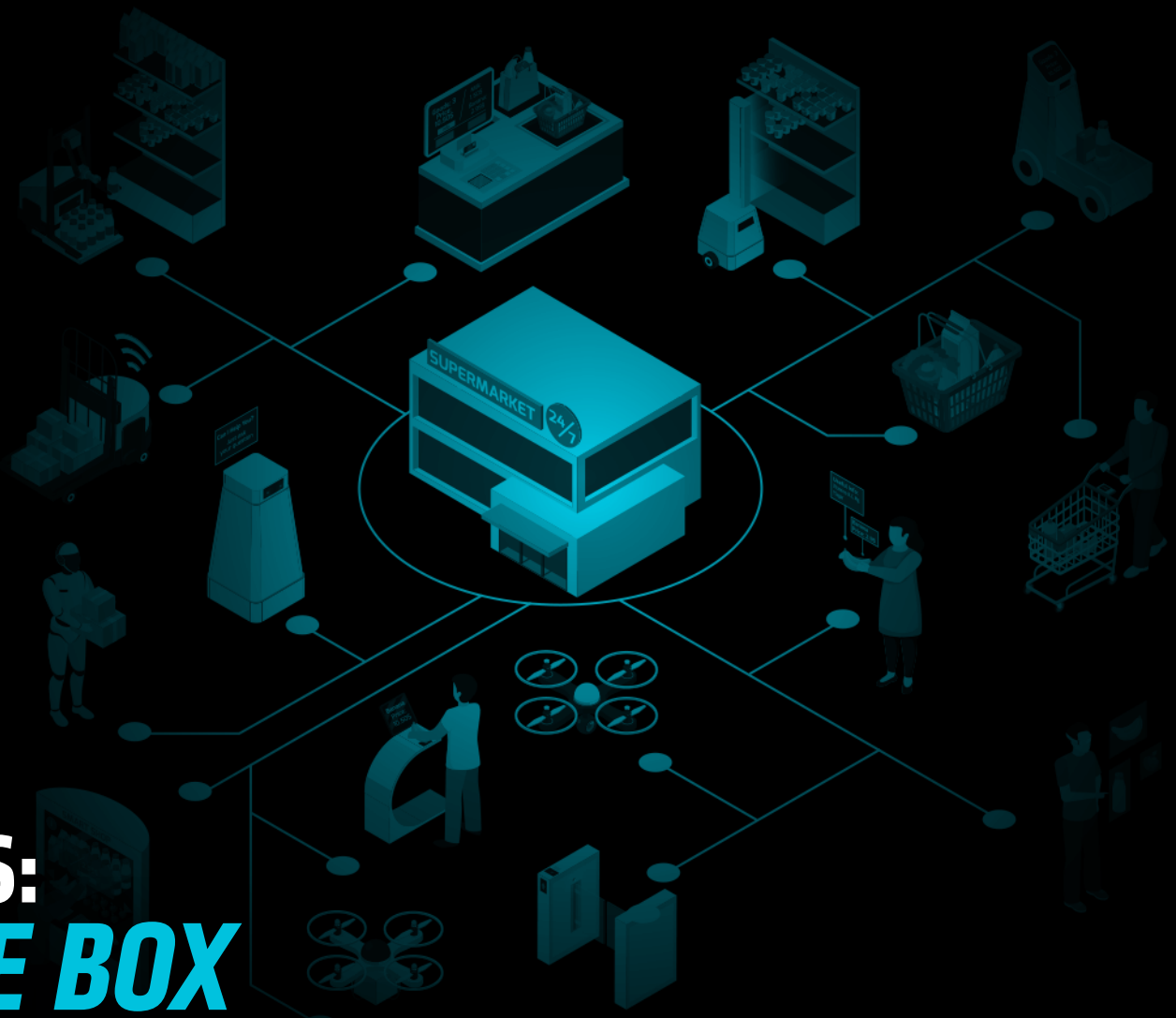
AMD
together we advance_retail



AMD RETAIL AI SOLUTIONS: *OUT OF THE BOX AND INTO THE STORE*

A NEW ALLIANCE IS BRINGING AI TO MASS-MARKET RETAIL

INTRODUCTION

AI IS READY FOR RETAIL

Advanced AI capabilities that were out of reach mere months ago can now run on off-the-shelf servers, unlocking store-wide intelligence and automation that can scale nationwide.

Two things are fueling this change. Drastic drops in AI model sizes and the rise of high-performance, energy-efficient AMD EPYC™ Server CPUs. Together, these forces are driving down computing costs and expanding AI services to retail endpoints like kiosks, surveillance desks, POS systems, and drive-throughs.

Our new venture, AMD Retail AI Solutions, brings integrators, software vendors, and OEMs together in a unified service so that retailers can take advantage of this transformative moment with ease.

THE TIPPING POINT IS HERE

SHRINKING AI OVERHEAD AND RISING CPU PERFORMANCE ARE MAKING RETAIL AI VIABLE AT SCALE

2022-2025

↓**280X**

Lower AI inference costs[1]

2022-2024

↓**142X**

Smaller AI model sizes[2]

AMD EPYC™ Server CPUs

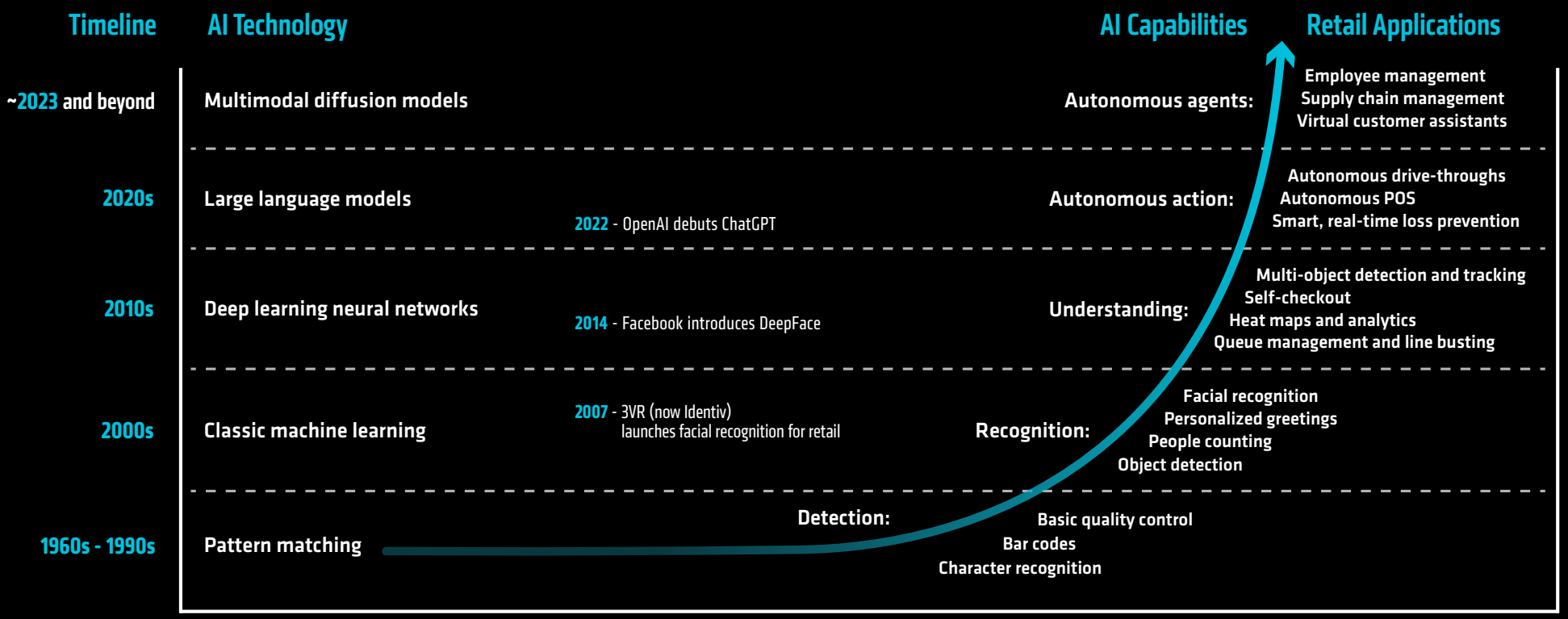
↑**11.3X**

Performance increase 1st – 5th Generation[3]

THE STATE OF RETAIL AI

ADVANCES IN INTELLIGENCE ARE ACCELERATING EXPONENTIALLY

Machine automation has been a key ingredient of retail innovation since the dawn of computers, one that matured gradually up to the dot-com era and has accelerated dramatically since. With the introduction of deep learning and large language models, AI capabilities have increased exponentially, fueling an explosion of smart, autonomous retail applications that are reinventing the industry.



MASS RETAILERS ARE ALREADY USING AI TO GAIN AN EDGE

As LLMs and generative AI transform retail applications, retailers are integrating AI throughout their operations. Recent industry surveys show that AI is rapidly becoming essential to the majority of retailers.

- More than **8 in 10 retailers** have implemented AI to a large or moderate extent.[4]
- **48%** say AI is active in most core functions of their retail operations.[4]

RETAILERS REPORT USING AI THROUGHOUT THEIR OPERATION

- **70%** use AI in marketing[4]
- **62%** use AI in IT/digital[4]
- **58%** use AI in digital commerce[4]
- **54%** use AI in merchandising, strategy, and pricing[4]

THE OPPORTUNITY

GENERATIVE AI HAS GONE FROM CUTTING EDGE TO AFFORDABLE COMMODITY IN THREE SHORT YEARS

Since the debut of ChatGPT in 2022, AI has followed the classic technology progression from high hopes and even higher costs to commodity availability and everyday productivity. The pace, however, has been much faster than previous technology revolutions. Costs have fallen, capabilities have exploded, and the amount of computing power needed to support truly revolutionary AI applications has shrunk from data-center scales to standalone servers.

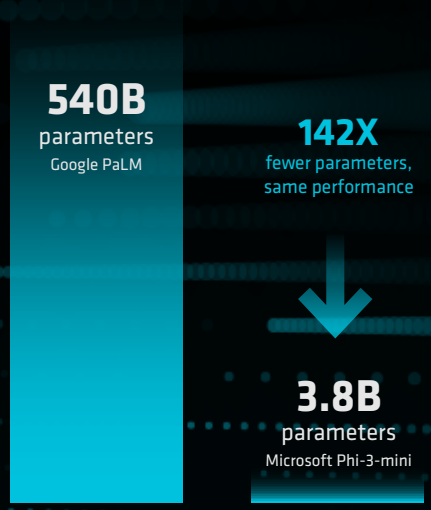
AI costs have dropped dramatically

280X lower inference costs[1]

2X lower hardware costs[1]

2.1X higher energy efficiency[1]

Small models are matching foundation model performance



Graph for illustrative purposes only.
See note 2 for details.

IN-STORE AI AT NATIONWIDE SCALE IS IN REACH FOR RETAILERS

When large language models (LLMs) debuted, only early movers with large-scale data centers and in-house AI expertise could experiment with cutting-edge AI. Training custom models, fielding pilot projects, and procuring expensive AI infrastructure required budgets well beyond what most retailers could afford.

Thanks to falling costs, shrinking models, and vastly more powerful AMD EPYC™ Server CPUs, today's landscape is very different. Now advanced AI that needed large data centers is available off-the-shelf on servers you can deploy on site, in a kiosk, or at the check stand.

	FIRST-GENERATION RETAIL AI	TODAY'S RETAIL AI SOLUTIONS
ADOPTERS	Global enterprises	Large, medium, and small retailers
MODELS	Large, bespoke models tailored to a single retailer	Small, use-case-specific models
HARDWARE	Data centers with GPU arrays	Off-the-shelf servers with general purpose CPUs
POWER CONSUMPTION	High	Low
SKILLS REQUIRED	AI and data science	General IT
TOTAL COST	Very high	Affordable

AMD RETAIL AI SOLUTIONS

A NEW ALLIANCE IS BRINGING AI TO MASS-MARKET RETAIL

Despite the falling prices and a rapid expansion of retail AI technology, many organizations lack the AI and systems expertise it takes to design production-ready solutions and bring them online. That's why we created AMD Retail AI Solutions.

AMD brings together the software developers, hardware manufacturers, and systems integrators. You get a jumpstart on your AI efforts with ready-to-deploy, cost-effective AI solutions.

AMD Retail AI Solutions

AMD AI expertise

Industrial-grade, open-standard AI infrastructure with an open-source software development stack in an ecosystem that includes AI leaders like Meta and OpenAI.

AMD EPYC™ Server CPUs

Full range of server CPUs from powerful, energy-efficient, AI-capable CPUs for in-store AI to high-density, data-center grade processors.

AMD Retail AI Services

Design and fulfillment services that can stand up production solutions at a global scale.

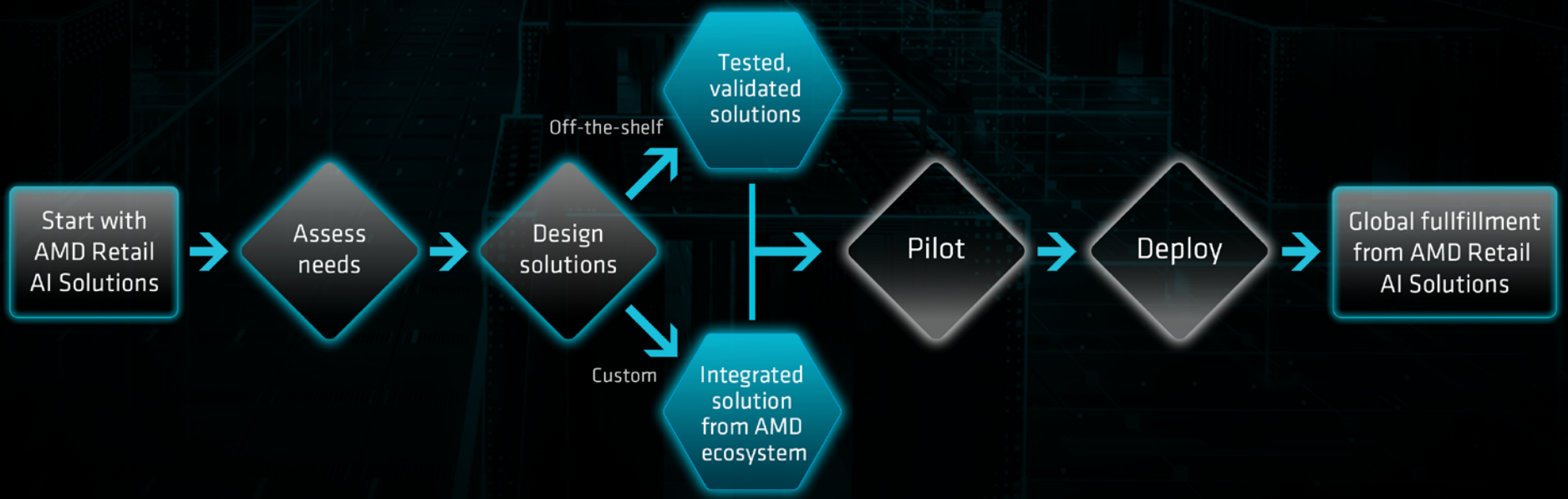
Tested, validated AI solutions

Open ecosystem of retail AI from leading ISVs, pre-qualified on hardware from leading manufacturers.

AMD AND OUR PARTNERS ORCHESTRATE YOUR RETAIL AI TRANSFORMATION

We've designed AMD Retail AI Solutions to support retail IT teams as they explore, design, and implement AI solutions for their organizations. We do an initial assessment and stay at your side from design through final deployment.

Our tested, validated solutions provide off-the-shelf options for key retail applications. If your needs are more complex, we can draw on the entire AMD ecosystem to create a custom solution at practically any scale.



CHOOSE VALIDATED SOLUTIONS OR CUSTOM PLATFORMS FROM LEADING MANUFACTURERS

AMD Retail AI Solutions run on AMD EPYC™ Server CPUs in a range of server configurations that balance cost, performance, and power efficiency. Retailers can choose from tested, validated solutions with ready-to-run software, or work with AMD and our partners to design custom AI solutions.

Validated AMD EPYC Server CPU platforms

HPE

HPE ProLiant DL145, DL325, DL365, and DL385 server platforms

Lenovo

Lenovo ThinkEdge SE455, SR635, SR645, SR655, and SR665 server platforms

SUPERMICRO

Supermicro WIO Systems, Supermicro AS-E300-14GR, AS-1115S-FWTRT and AS-1115-FDWTRT servers

AI SOLUTIONS FOR RETAIL, VALIDATED BY AMD

Our growing portfolio can help deliver AI capabilities to critical use cases throughout retail. Use them off-the-shelf or as building blocks in a custom AI solution. Our open innovation platform is poised for growth with more solutions quickly becoming available.

SOLUTION	USE CASE
ShopAssist™ from RadiusAI	Barcode-free, vision-enabled self-checkout
ShopAssist™ Plus from RadiusAI	Intelligent, real-time inventory management
StorMagic	Hyperconverged infrastructure solution for distributed retail environments
Wobot AI Vaidio® AI Vision Platform	Queue management Customer traffic heatmap Loss prevention Inventory tracking

SHOPASSIST™ FROM RADIUSAI

With ShopAssist, customers don't need to scan barcodes. They simply place items onto a checkout counter. ShopAssist uses computer vision to automatically detect and recognize products, even when items are stacked on top of each other. The solution can even handle age-verified items, like alcohol products.

Retailers can deploy ShopAssist as a standalone solution or integrated into an existing POS terminal.

RadiusAI

USE CASE

Vision-enabled self-checkout

OUTCOMES

Faster checkout

Less demand on store employees

Reduced loss due to theft and errors

VALIDATED HARDWARE

Lenovo

Lenovo ThinkEdge servers
with AMD EPYC™ Server CPUs

SHOPASSIST™ PLUS FROM RADIUSAI

ShopAssist Plus tracks food items in real time, providing unparalleled visibility into food inventory for proactive management of stock levels, freshness, and demand.

ShopAssist Plus uses advanced vision AI to turn existing store security cameras into intelligent inventory systems.

RadiusAI

USE CASE

Intelligent, real-time inventory management

OUTCOMES

Live inventory tracking

Freshness monitoring

Demand forecasting

Automated ordering

VALIDATED HARDWARE

Lenovo

Lenovo ThinkEdge servers
with AMD EPYC™ Server CPUs

STORMAGIC

The StorMagic hyperconverged infrastructure solution, SvHCI, provides lightweight, highly available infrastructure that's ideal for AI workloads within distributed retail environments. Designed for space- and power-constrained stores, it keeps POS, AI-driven analytics, and in-store applications up and running without interruption, thanks to consolidated compute, storage, and virtualization and a resilient two-node architecture.

With a centralized and user-friendly management tool, IT teams can deploy, patch, and standardize applications across hundreds of locations without replacing existing store hardware. Advanced by AMD EPYC™ Server CPUs, SvHCI delivers the performance foundation required for dependable in-store intelligence at scale.

StorMagic

USE CASES

Consolidated compute and storage

Deploy

Patch

Manage

OUTCOMES

Centralized management

Highly available infrastructure

Decoupled hardware / software lifecycles

WOBOT AI: EDGE-BASED VIDEO AI AGENTS FOR SMARTER STORES

Wobot AI delivers intelligent video AI Agents that run on AMD EPYC™ Server CPU platforms, transforming existing camera infrastructure into continuous, in-store intelligence. As part of AMD Retail AI Solutions, Wobot provides pre-validated, edge-ready deployments that autonomously interpret customer movement, staff activity, and vehicle flow in drive-through environments.

These AI Agents generate heat maps, queue metrics, loss-prevention alerts, and AI-powered staff task lists – helping retailers optimize layouts, improve labor efficiency, increase drive-through throughput, and maintain consistent execution across locations. By operating locally at the edge, Wobot enables scalable, cost-effective AI without dependency on GPU-heavy or cloud-bound architectures.

Wobot.ai

USE CASES

Heat maps and customer flow analysis

Queue monitoring

Loss prevention alerts

AI-powered staff task lists

OUTCOMES

Optimized store layouts and merchandising

Improved labor efficiency

Consistent execution across store fleets

VALIDATED HARDWARE

HPE

HPE ProLiant with AMD EPYC™ Server CPUs

SUPERMICRO

Supermicro edge servers with AMD EPYC Server CPUs

VAIDIO® AI VISION PLATFORM

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.

VAIDIO

USE CASES

Loss prevention

Staff optimization

Queue management

Store layout and merchandising

OUTCOMES

Reduced theft

Improved safety

Faster customer service

WHY AMD

WORK WITH AN INDUSTRY LEADER IN AI

From our 6-Gigawatt AI computing partnership with OpenAI[5] to producing the best CPUs for enterprise AI,[6] AMD is at the forefront of the AI revolution. AMD Retail AI Solutions combine our AI expertise with leading hardware manufacturers and software developers, creating an ecosystem of AI solutions for mass-market AI.

RETAIL AI RUNS ON AMD EPYC™ SERVER CPUs

AMD EPYC Server CPUs deliver something competing CPUs simply can't: extraordinary value built on open industry standards powering a vibrant open, ecosystem. You get the performance, core density, and energy profiles you need for data-intensive AI workloads, while keeping initial and ongoing costs in check.

Our highest-core-count CPUs can easily handle up to 20 billion parameter AI models at scale, and our lower-core-count, edge-optimized CPUs can support small to medium AI models and general-purpose workloads with a single processor. They're also ideal for virtualization, allowing retailers to deploy the latest AI apps and keep legacy software up and running while managing nationwide brick-and-mortar stores with cloud-native tools and services.

GET MORE AI FOR YOUR DOLLAR WITH AMD EPYC SERVER CPUs

AMD EPYC Server CPUs deliver more performance per dollar for on-site/edge servers.

UP TO **46%** more performance per rack[7]

UP TO **43%** more throughput per watt, per dollar[7]

Compares SPECrate®2017_int_base performance for single socket on-site/edge servers equipped with 64-core, AMD EPYC™ 8000 Series Server CPUs versus 52-core Intel® Xeon® 8471N CPUs. See note 7 for details.

THE AMD EPYC™ SERVER CPU PORTFOLIO FOR RETAIL AI

AMD EPYC™ 4005 SERIES SERVER CPUs

AMD EPYC™ 4005 Series Server CPUs are an energy-efficient option for POS backends, digital sign CMSs, and payment gateways in small footprint stores with headroom for AI-powered people counting, demographic tracking, and loss prevention at the POS.

Explore AMD EPYC 4005 Series Server CPUs

AMD EPYC™ 8004 SERIES SERVER CPUs

Designed for power-constrained, intelligent edge computing, AMD EPYC™ 8004 Series Server CPUs are ideal for single-socket systems that can support smart, automated drive-throughs and analyze multiple camera feeds to provide advanced AI-assisted heat mapping, in-aisle inventory management, loss prevention, kitchen monitoring, and operations guidance for staff.

Explore AMD EPYC 8004 Series Server CPUs

AMD EPYC™ 9004 SERIES SERVER CPUs

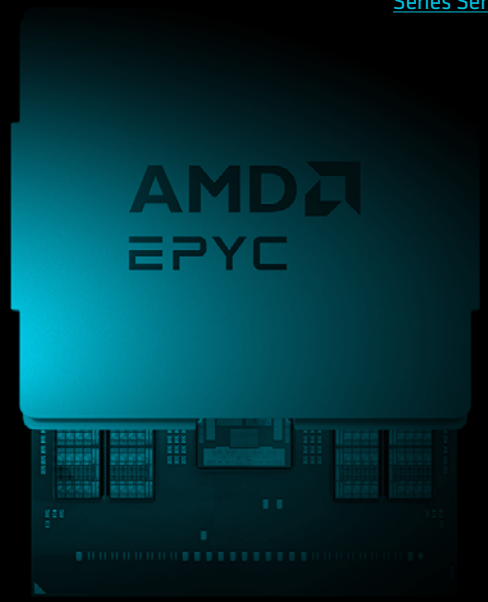
AMD EPYC™ 9004 Series Server CPUs deliver data center-level performance for virtualized deployments of traditional and AI retail applications across high camera counts in larger footprint stores.

Explore AMD EPYC 9004 Series Server CPUs

AMD EPYC™ 9005 SERIES SERVER CPUs

With more optimizations for AI than previous generations, 5th Generation AMD EPYC Server CPUs deliver high core counts and frequencies for multiple AI workloads and new levels of retail intelligence like assembling case files for loss prevention.

Explore AMD EPYC 9005 Series Server CPUs



CONFIDENTIAL AI IS BUILT IN

AMD EPYC 8000 and 9000 Series Server CPUs include AMD Infinity Guard with AMD Secure Encrypted Virtualization (SEV). Infinity Guard provides defense-in-depth to help prevent threats through security controls at multiple layers, including secure boot and encryption. AMD SEV helps protect data in use with virtual machines that can spin up on demand, enabling confidential computing for the entire AI lifecycle.[8]

GET STARTED

JUMPSTART YOUR AI JOURNEY WITH AMD RETAIL AI SOLUTIONS

Get started with ready-to-deploy retail AI along with comprehensive design and fulfillment services. AMD Retail AI Solutions can bring enterprise-caliber offerings and personalized guidance to retailers throughout the sector.

- **Grocery –** Small-format convenience stores to large-scale grocery stores
- **Restaurants –** Quick service restaurants, large-format dine-in chains
- **General retail –** Club warehouses and department stores, discount stores, shopping malls
- **Specialty retailers –** Home improvement, electronics, apparel, jewelry, salons, gas stations
- **Venues –** Stadiums, theaters, museums
- **Transportation –** Airports, railway stations, bus terminals and transit shelters

Learn More

1. The cost of performing inference with a GPT3.5-sized, ~175 billion parameter model dropped an astonishing 280 times since 2022. For more, see Stanford University, Human-Centered Artificial Intelligence, Artificial Intelligence Index Report 2025, page 4.

2. In 2022, the smallest model that scored higher than 60% on the Massive Multitask Language Understanding (MMLU) benchmark was Google's 540-billion-parameter Pathways Language Model (PaLM). By 2024, Microsoft's Phi-3-mini matched it with only 3.8 billion parameters. That's a 142X reduction in just two years. For more, see Stanford University, Human-Centered Artificial Intelligence, Artificial Intelligence Index Report 2025, page 99.

3. 9xx5-069D: SPECrate®2017_int_base comparison based on published scores from www.spec.org as of 07/05/2024. Generational scores are based on highest published scores from www.spec.org in respective launch years. 2P 192C AMD EPYC 9965, 500 W, 3100 SPECrate®2017_int_base, 6.200 SPECrate®2017_int_base/CPU TDP, https://spec.org/cpu2017/results/res2024q4/cpu2017-20241004-44979.html 2P 96C AMD EPYC 9654, 360 W, 1790 SPECrate®2017_int_base, 972 SPECrate®2017_int_base/CPU TDP, https://www.spec.org/cpu2017/results/res2022q4/cpu2017-20221024-32607.html 2P 64C AMD EPYC 7763, 280 W, 861 SPECrate®2017_int_base, 3.075 SPECrate®2017_int_base/CPU TDP, https://www.spec.org/cpu2017/results/res2021q4/cpu2017-20211121-30148.html 2P 64C AMD EPYC 7742. 225 W, 701 SPECrate®2017_int_base, 3.116 SPECrate®2017_int_base/CPU TDP, https://www.spec.org/cpu2017/results/res2019q4/cpu2017-20191125-20001.html 2P 32C AMD EPYC 7601, 180 W, 275 SPECrate®2017_int_base, 1.528 SPECrate®2017_int_base/CPU TDP, https://www.spec.org/cpu2017/results/res2017q4/cpu2017-20171211-01594.html

   SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation.
   See www.spec.org for more information.

4. BRG, "AI in Retail: In Pursuit of Meaningful AI Adoption," November 12, 2025.

5. AMD Newsroom, AMD and OpenAI Announce Strategic Partnership to Deploy 6 Gigawatts of AMD GPUs, October 6, 2025.

6. EPYC-029D: Comparison based on thread density, performance, features, process technology and built-in security features of currently shipping servers as of 10/10/2024. EPYC 9005 series CPUs offer the highest thread density, lead the industry with 500+ performance world records including world record enterprise leadership Java® ops/sec performance, top HPC leadership with floating-point throughput performance, AI end-to-end performance with TPCx-AI performance and highest energy efficiency scores. Compared to 5th Gen Xeon, the 5th Gen EPYC series also has more DDR5 memory channels with more memory bandwidth and supports more PCIe® Gen5 lanes for I/O throughput, and has up to 5x the L3 cache/core for faster data access. The EPYC 9005 series uses advanced 3-4nm technology, and offers Secure Memory Encryption + Secure Encrypted Virtualization (SEV) + SEV Encrypted State + SEV-Secure Nested Paging security features. For additional details, see https://www.amd.com/en/legal/claims/epyc.html#q=epyc5#EPYC-029D

7. SP6-003: General Purpose Application Integer Throughput (SPECrate®2017_int_base) claim based on 1P published results at spec.org as of 9/18/2023. 1P servers: EPYC 8534PN (64-core, est. 329W system power, $8,482 est. system cost, https://spec.org/cpu2017/results/res2023q3/cpu2017-20230828-38760.html) scoring 439 SPECrate2017_int_base vs. 1P Xeon Platinum 8471N (52-core, est. 484W system power, $8,477 est. system cost, https://spec.org/cpu2017/results/res2023q3/cpu2017-20230619-37381.html) scoring 450 SPECrate2017_int_base for 0.97x the performance with 44% better performance/system W (32% lower system W). Assuming an 8kW rack deploying servers, 24 ea. EPYC 8534PN vs. 16 ea. Xeon 8471N can fit within the power budget delivering 1.46x the total integer throughput/rack. Power assumptions based on Phoronix Test Suite maximum system watt measurements. Testing not independently verified by AMD. 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. Estimated system pricing based on Bare Metal Server GHG TCO v9.52, https://www.amd.com/en/legal/claims/epyc.html#q=SP6-003&sortCriteria=%40title%20ascending&numberOfResults=96.

   SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation.
   See www.spec.org for more information.

8. GD-183A: 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/products/processors/server/epyc/infinity-guard.html. https://www.amd.com/en/legal/claims/epyc.html#q=GD-183A&sortCriteria=%40title%20ascending&numberOfResults=96