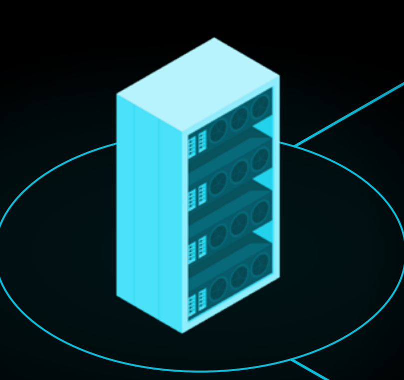
AMD
together we advance_retail

AMD RETAIL AI SOLUTIONS

READY TO RUN
POWERFUL AI FOR ANY STORE

As models shrink and costs fall, cutting-edge AI is beginning to remake mass-market retail applications. With AMD Retail AI Solutions, retailers can deploy advanced AI alongside traditional workloads using affordable, in-store servers and devices.

THE MOMENT TO MOVE ON AI
IS NOW



AI overhead has dramatically dropped

280X
lower inference costs[1]

142X
smaller model sizes[2]

2X
lower hardware costs[1]

AMD EPYC™ Server CPUs have headroom to spare

Today's in-store hardware can deliver AI performance at the right price and power budget.

~58%
higher total AI/ML throughput[3]

~52%
more performance per watt, per dollar[3]

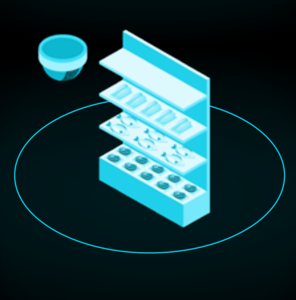
~42%
lower system power[3]

Edge-optimized AMD EPYC 8534PN Server CPUs deliver higher AI throughput than Xeon Platinum 8471N at retail-friendly cost and power budgets. See note 3 for details.

AMD RETAIL AI SOLUTIONS
REINVENT THE IN-STORE EXPERIENCE

Virtualization at the edge
Simplified software image management at the store level helps keep legacy workloads running alongside modern AI.

Smarter POS
Today's AI makes checkout faster and more accurate by recognizing items and spotting shrinkage — no bar codes needed.



Real-time inventory management
Active intelligence helps keep shelves stocked and items fresh by monitoring items and guiding staff.

AI-driven business insights
AI generates real-time feedback on merchandising and operations performance from in-store video feeds.

Proactive AI assistants
Generative AI assistants coach managers, guide employees, and serve customers at the drive-through, in-store, and on their phones.

Behavior-aware loss prevention
New levels of intelligence recognize patterns and behaviors that indicate theft and fraud.

START NOW WITH TESTED AND VALIDATED AI SOLUTIONS FROM RETAIL LEADERS

Retail AI runs on AMD EPYC Server CPUs

- The best CPUs for enterprise AI[4]
- Cost-effective performance for retail AI[3]
- Edge-optimized options for in-store servers
- Turnkey confidential AI is built in

7 OF THE 10 LARGEST FORTUNE 100 RETAILERS RUN ON AMD EPYC SERVER CPUs[5]
IT'S TIME TO JOIN THEM

Transform your operations with affordable AI running on the same high-performance AMD EPYC Server CPUs trusted by the world's largest retailers. Contact AMD Retail AI Solutions to get started.

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 details, see page four in the Artificial Intelligence Index Report 2025, from Stanford University's Human-Centered Artificial Intelligence Institute.
2. By 2024, Microsoft's Phi-3-mini matched the 2022 performance of Google's 540 billion-parameter PaLM model using only 3.8 billion parameters. That's a 142X drop in just two years. For details, see page 99 in the Artificial Intelligence Index Report 2025, from Stanford University's Human-Centered Artificial Intelligence Institute.
3. SP6-006: AI/ML workloads (Neural Magic DeepSparse, oneDNN, OpenCV, OpenVINO, and TensorFlow averaged) performance/system W/system $ comparison based on Phoronix Test Suite paid testing as of 8/18/2023. Configurations: 1P 64C EPYC 8534PN (0.96x relative performance, 247 avg system W, est $8,482 system cost USD) powered server versus 1P 52C Xeon Platinum 8471N (1.07x relative performance, 423 avg system W, est $8,477 system cost USD) powered server for 0.89x the performance, 42% lower system power (1.52x the performance/system W), comparable system cost for 1.52x the overall system performance/W/$. Assuming an 8kW rack deploying servers, 32 ea. EPYC 8534PN vs. 18 ea. Xeon 8471N can fit within the power budget delivering 1.58x the total AI/ML throughput/rack on average. Testing not independently verified by AMD. Scores will vary based on system configuration and determinism mode used (default TDP power determinism mode profile used). 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. Results may vary and are normalized to EPYC 8534P Performance Determinism mode (always 1 in all measurements). Estimated system pricing based on Bare Metal Server GHG TCO v9.52. For details see https://www.amd.com/en/legal/claims/epyc.html#q=SP6-006&sortCriteria=%40title%20ascending
4. 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, leads 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 details see https://www.amd.com/en/legal/claims/epyc.html#q=EPYC-029D&sortCriteria=%40title%20ascending
5. EPYC-059D: 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.

AMD
together we advance_retail