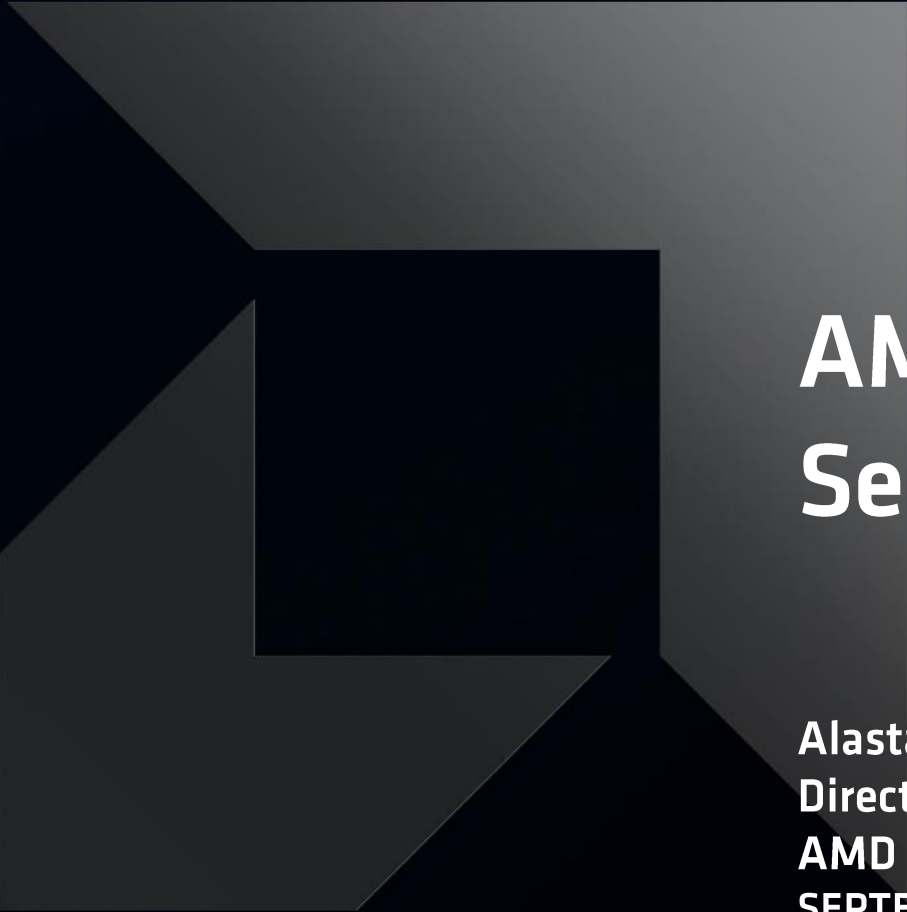




AMD Solutions for Financial Services

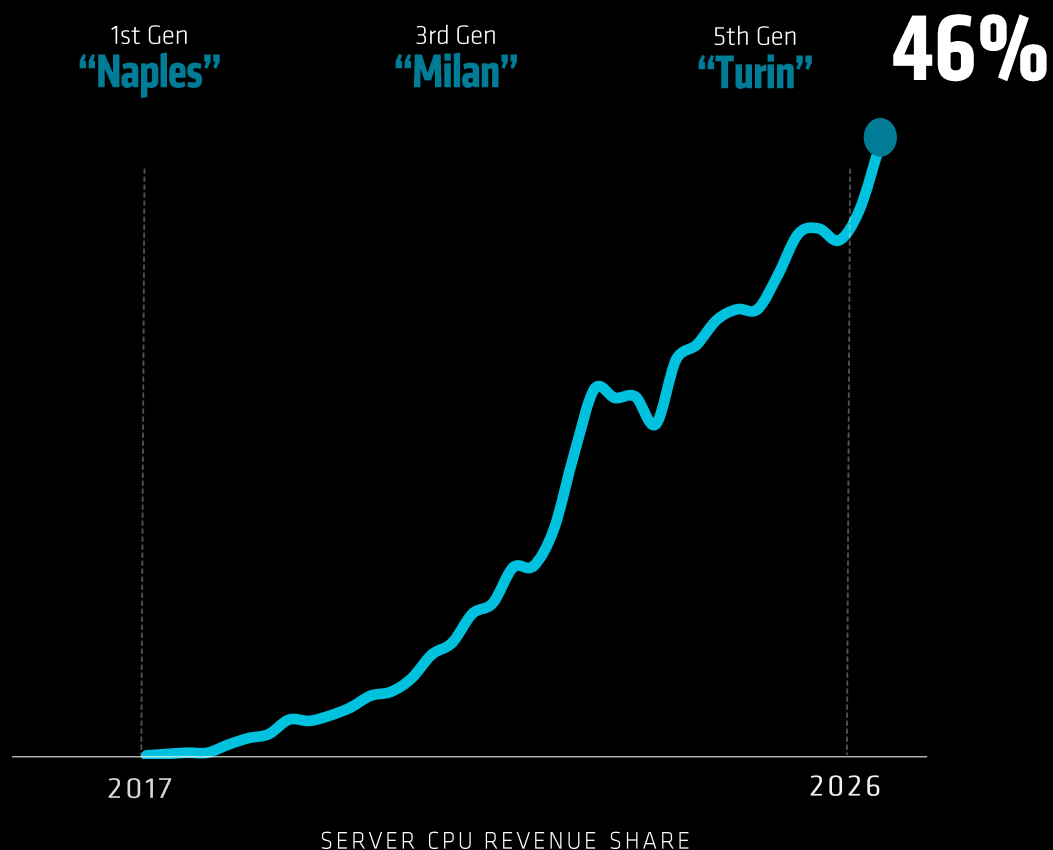
Alastair Richardson
Director of Enterprise Solutions Financial Services
AMD
SEPTEMBER 2026

CAUTIONARY STATEMENT

This presentation contains forward-looking statements concerning Advanced Micro Devices, Inc. (AMD) such as the features, functionality, performance, availability, timing and expected benefits of AMD products including the AMD Instinct™ MI450 Series accelerators; and AMD EPYC™ market share, which are made pursuant to the Safe Harbor provisions of the Private Securities Litigation Reform Act of 1995. Forward-looking statements are commonly identified by words such as "would," "may," "expects," "believes," "plans," "intends," "projects" and other terms with similar meaning. Investors are cautioned that the forward-looking statements in this presentation are based on current beliefs, assumptions and expectations, speak only as of the date of this presentation and involve risks and uncertainties that could cause actual results to differ materially from current expectations. Such statements are subject to certain known and unknown risks and uncertainties, many of which are difficult to predict and generally beyond AMD's control, that could cause actual results and other future events to differ materially from those expressed in, or implied or projected by, the forward-looking information and statements. Investors are urged to review in detail the risks and uncertainties in AMD's Securities and Exchange Commission filings, including but not limited to AMD's most recent reports on Forms 10-K and 10-Q.

AMD does not assume, and hereby disclaims, any obligation to update forward-looking statements made in this presentation, except as may be required by law.

CPU Leadership Has Driven Market Momentum



**Most Fortune 100 Companies
Trust Their Data Centers to
AMD EPYC™ CPUs**

**Every Week, a Forbes Global
2000 Company Switches to AMD**

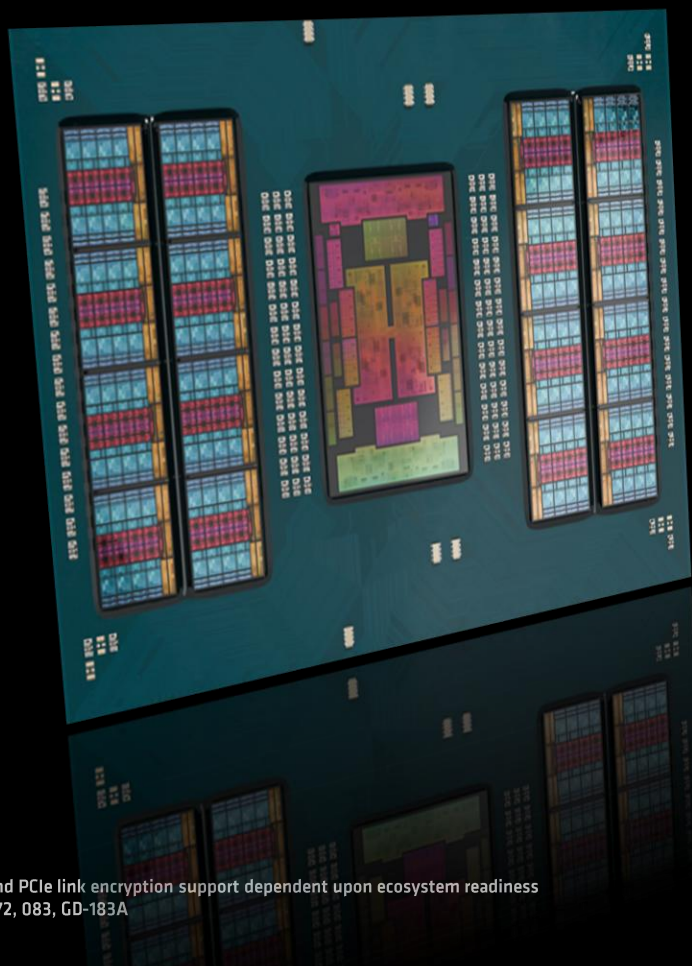
475+ EPYC™ Platforms from OxMs | 1,600+ EPYC™ Powered VMs

See Endnotes: EPYC-055D, 058

From Fortune ©2025 Fortune Media IP Limited. All rights reserved. Used under license.

From Forbes © 2025 Forbes Media LLC. All rights reserved. Used under license.

5th Gen AMD High-Frequency EPYC for Low-Latency workloads



Compute

- “Zen5” **High-Frequency SKUs up to 64 cores**
- **AVX-512 with full 512b data path**
- **SDCI (Smart Data Cache Injection) for low-latency**
- **High-Frequency SKUs up to 5GHz**
- **3/4nm Zen cores**

I/O & Platform

- 2P and 1P Configurations
- **Up to 160 lanes of PCIe® Gen5**
- **PCIe® link encryption**
- **CXL® 2.0¹**

Memory

- 12 ch. DDR5 ECC **up to 6400* MT/s**
- Up to 2 DIMMs/channel capacity delivering up to **6TB/socket**
- **Dynamic Post Package Repair (PPR)** for x4 and x8 ECC RDIMMs

¹ - CXL Type 162 devices and PCIe link encryption support dependent upon ecosystem readiness
See endnotes 9xx5-048, 072, 083, GD-183A

AMD Threadripper PRO 9995WX for Electronic Trading

Up to 96c high-performance cores

- Twice as many cores as the competition

**96 HIGH
PERFORMANCE
CORES**

High frequencies for low-latency response

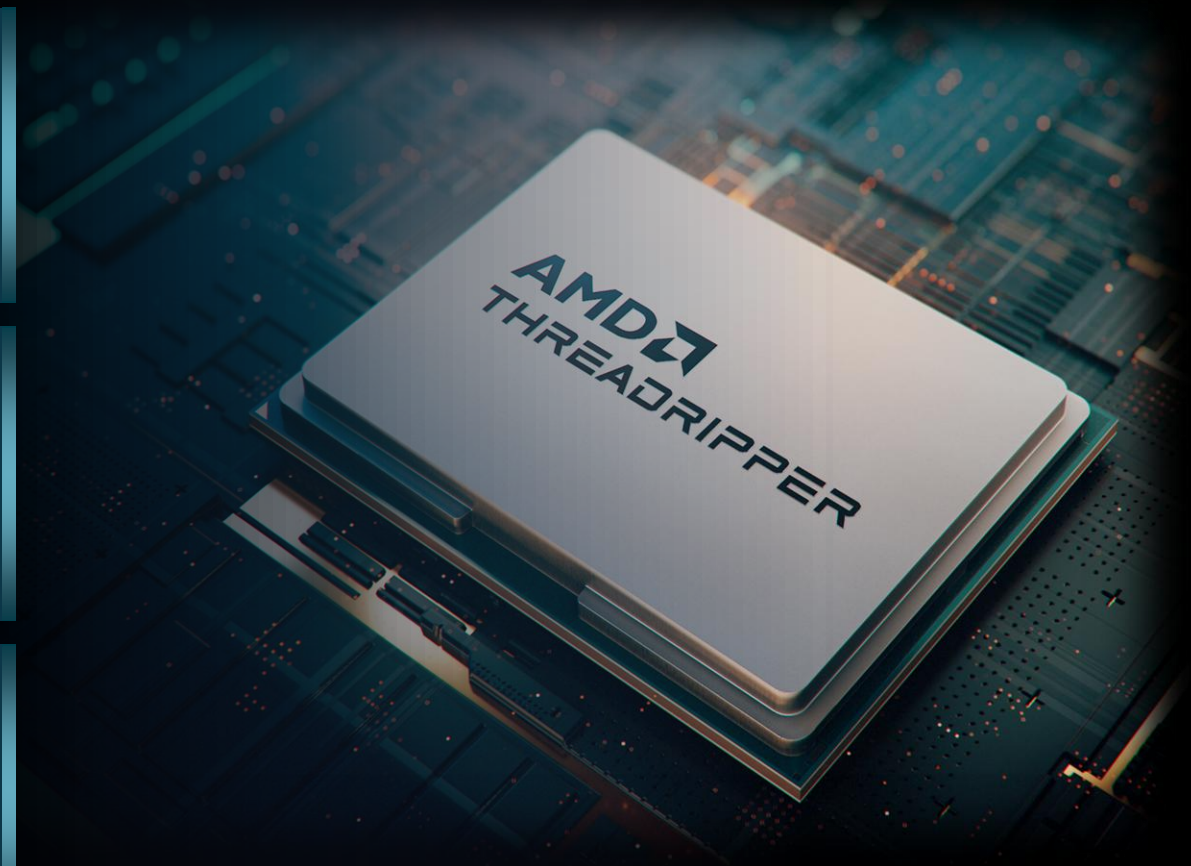
- 96c up to 2.5Ghz (base) / 5.4GHz (Max Boost)

**HIGH CLOCK
SPEED, LOW-
LATENCY**

Support CPU overclocking

- Reducing jitter and ensuring consistent execution of market data

**OVERCLOCKING
ENABLED**



AMD EPYC 4564P & 4565P for Electronic Trading



2X HIGH PERFORMANCE CORES

Up to 16c high-performance cores

- Competition only offers eight performance cores

HIGH CLOCK SPEED, LOW- LATENCY

High frequencies for low latency response

- 16c up to 4.5Ghz (base) / 5.7GHz (boost)

LOW TDP ON OVERCLOCKING CPU

Energy efficient TDP from 65W to 170W

- The competition draws a huge TDP when overclocking CPU for HFT

AMD X4-series Low-Latency Ethernet Adapters Raise the Bar*



Higher Host Throughput
with up to

200%

IMPROVED SYSTEM
PERFORMANCE

Higher Network Throughput
with up to

200%

HIGHER INGESTION
OF MARKET DATA

Lower Power/bit,
Enabling up to

50%

LOWER TOTAL
COST OF OWNERSHIP

The Low-Latency Recipe in Four Steps

1 Step 1 – Platform & BIOS

- F-series EPYC 9005: high clock, large L3
- Cores, memory and NIC in one NUMA domain
- Low-latency BIOS: C-states, ASPM, DF P-states off

2 Step 2 – OS and TuneD

- Minimal RHEL 9 + amd-cpu-partitioning profile
- isolcpus, nohz_full, rcu_nocbs on the hot cores
- IRQs and irqbalance off the clean cores

3 Step 3 – Network and Time

- Kernel bypass: Onload, TCPDirect, ef_vi
- SDCI steers packets straight into L2 cache
- sfptpd for hardware-timestamped PTP

4 Step 4 – Measure and Prove

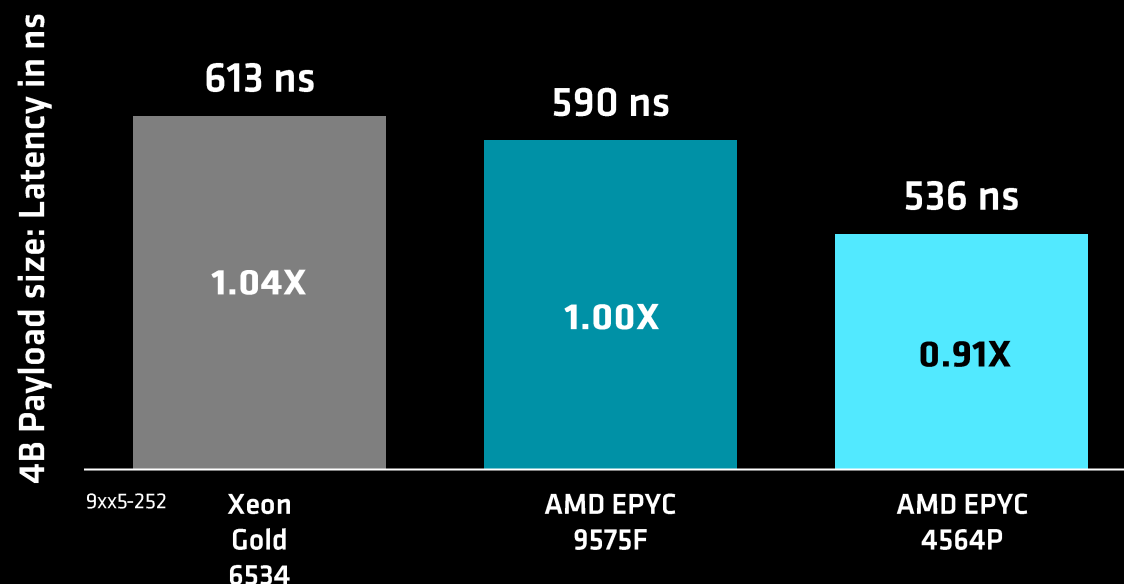
- sysjitter for per-core jitter and outliers
- perf and ftrace osnoise to find the source
- Know when to stop: cost vs the business aim

Find out more on the September 2026 Low latency Tuning Guide on AMD.com
<https://docs.amd.com/v/u/en-US/low-latency-tuning-for-amd-epyc-9005-cpu-powered-servers>

AMD Outstanding Latency with AMD EPYC™ and X4

AMD Combining CPU and NIC Innovation to Offer Leading Performance

10GE Latency for X4 Across Leading Platforms



OUTSTANDING LATENCY

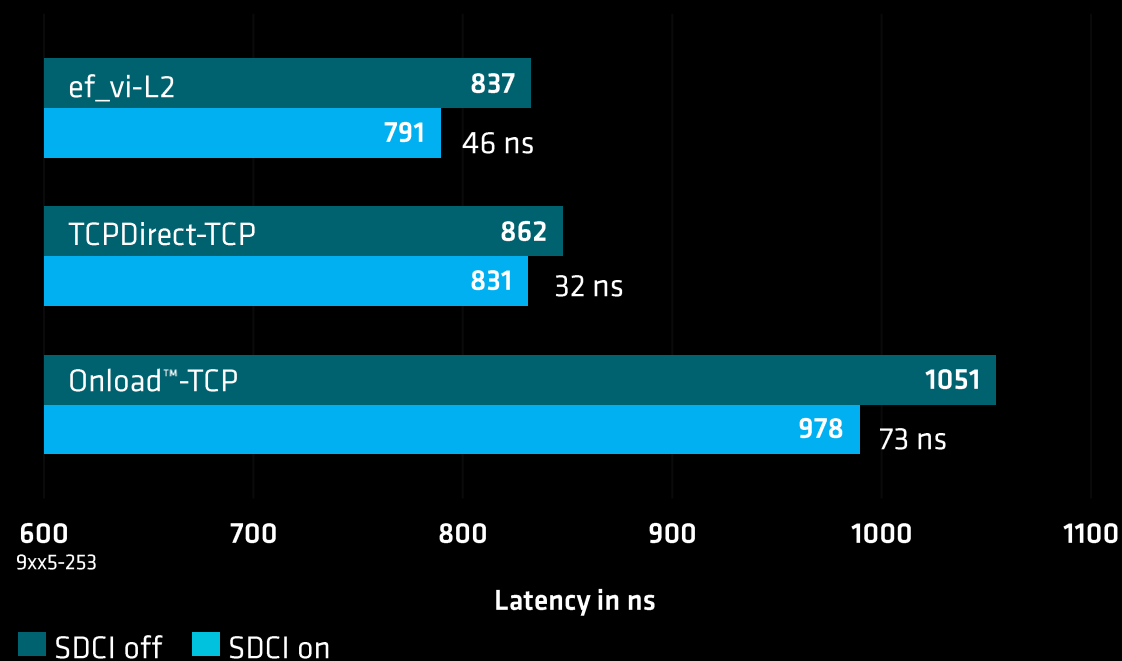
AMD EPYC™ and X4: Co-designed for Lower Latency

- Outstanding small packet performance in Server-class CPUs
- SDCI on AMD EPYC CPUs improves application latency
- AMD EPYC + AMD Solarflare™ provides the best latency

SDCI Application Latency Performance on AMD EPYC™ Zen 5

AMD EPYC Processors with X4 Improve System Performance

X4: 10GE 256B SDCI Improvements



OPTIMIZED TOGETHER

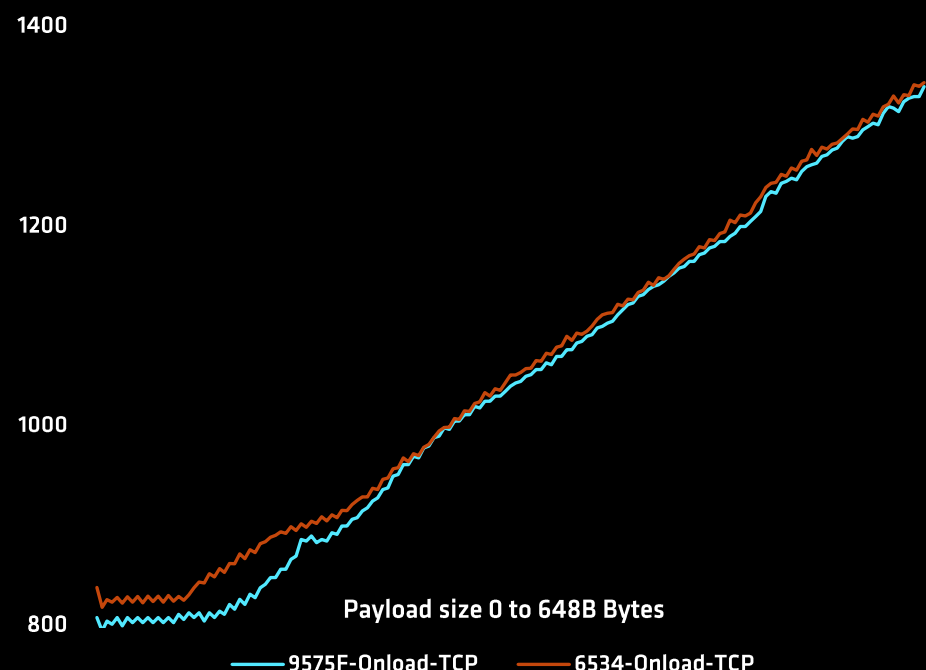
Optimization Using SDCI in CPU Architecture

- Receive data steered to L2 cache of desired core
- Reduced application latency
- Better memory utilization
- Optimized cache management

SDCI Application Latency Performance on AMD EPYC™ Zen 5

AMD EPYC Processors with X4 Improve System Performance

Onload™ TCP with AMD EPYC 9575F vs Xeon 6534



BETTER PERFORMANCE

Measured Using Identical Onload-TCP Applications

- AMD EPYC 9575F achieves greater performance than Intel Xeon 6534 running Onload-TCP Based Applications
- AMD EPYC 9575F is 14ns on average faster than Xeon 6534 with Onload-TCP

Alveo™ V80-LL (Low Latency)

New V80 Card Optimized for Fintech Market

**GTY Low Latency
Networking**

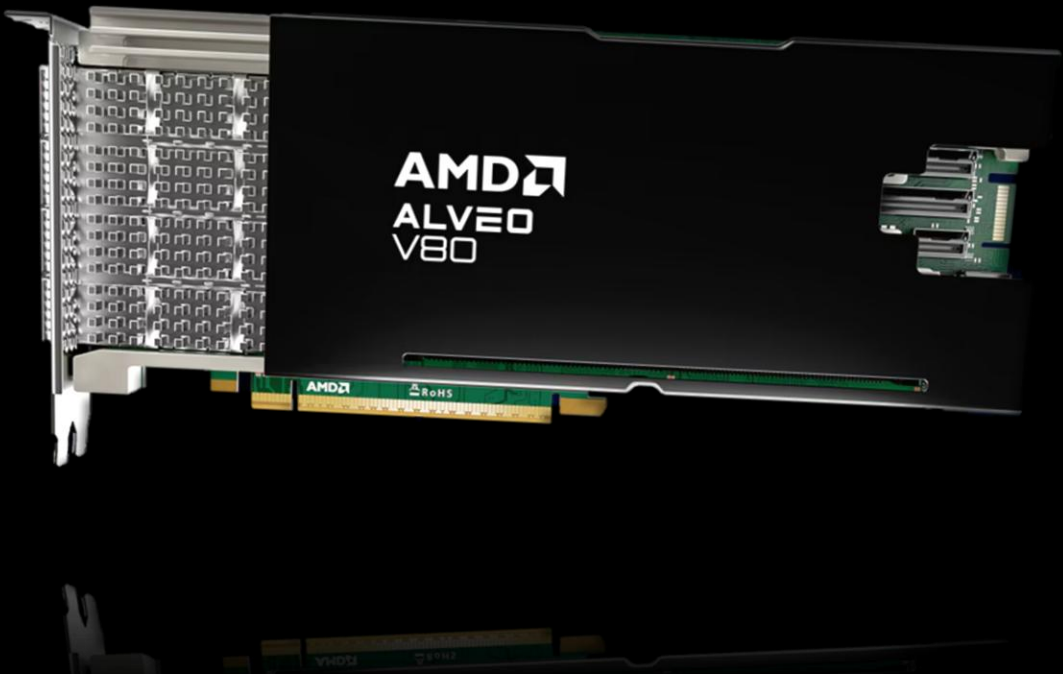
- 4x10/25G
- QSFP56 ports

7nm Versal Architecture

- 2.6M LUTs for flexible compute
- 10.9K DSP slices
- 32GB HBM at 810GB/s

On-Board DDR

- 8GB for Arm processor management
- DIMM Expansion slot



FEATURES

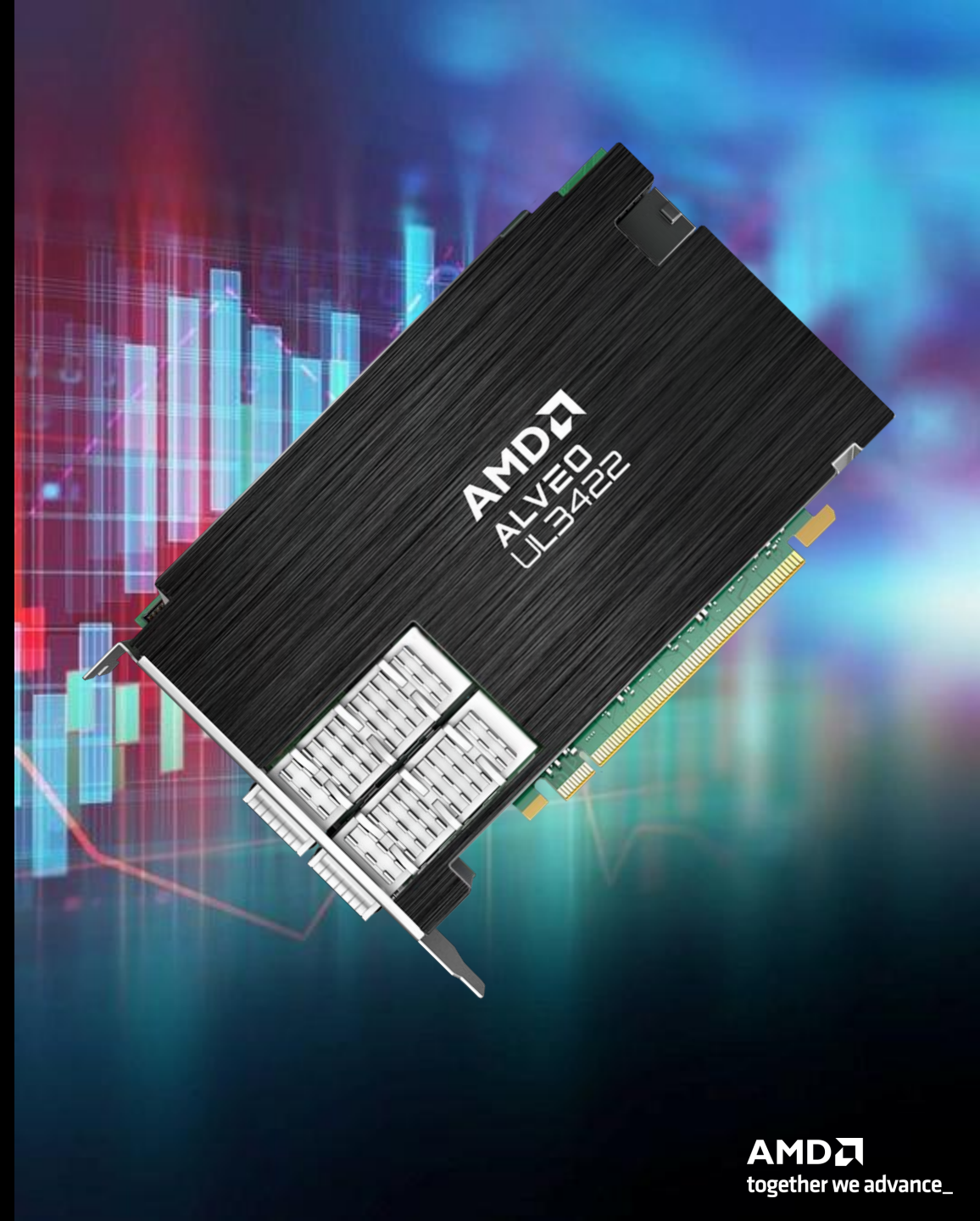
SPECIFICATION

Device Architecture	VH1782 -3 (Versal HBM)
Logic Density	2.6 million LUTs
HBM Capacity	32GB
HBM Bandwidth	820GB/s
DDR4 Capacity	32GB
URAM /BRAM	541Mb/132Mb
DSPs	10,848
Network Interface	• 4x 10/25 • Low Latency GTY's
Expansion	MCIO connectors at PCIe Gen5 rates
Form Factor	Full-height, ¾ Length (FH 3/4L)
PCIe® Interface	PCIe® Gen4 x16, 2x Gen5 x8
Power	300 watts, passively cooled



Alveo UL3422 FPGA Accelerator

FEATURES	SPECIFICATION
On-Board Memory	• 16GB DDR4-2400 MT/s
Network Interface	2x QSFP DD (16x 10/25G Ports)
Expansion Ports	• 2 ARF6 16x10/25G ports • Connects multiple cards • 1 Pico-Clasp connectors for sideband
Form Factor	Full-height, Half Length (FHHL) Single Slot
PCIe® Interface	PCIe® Gen4 x8
1PPS	• 1 PPS In and 1 PPS Out port
Power	• 180W Electrical • 120W TDP • Passive cooling
Product SKU	A-UL3422-P16G-PQ-G



AMD AVX-512 in Electronic Trading

1 Exchange Matching Engine

- Vectorized order-book scans, price-level aggregation and risk checks: one 512-bit op replaces 8 scalar compares, cutting cycles per order. Zen 5's full 512-bit datapath sustains AVX-512 without the frequency penalty seen on legacy x86, so deterministic tail latency is preserved under burst load.

2 Market Data Gateway

- Line-arbitration, gap detection, ITCH/FIX-SBE parsing and checksum/CRC work map directly to AVX-512 byte and mask instructions. VBMI/GFNI accelerate bit-field unpacking and decompression, lifting messages-per-second per core and reducing the core count needed to handle 10M+ msg/s feeds.

3 Trading Applications

- Signal generation, book-building, pricing and pre-trade risk run on AVX-512 FMA and gather/scatter: more Monte-Carlo paths and more strategies per server. Combined with Solarflare X4 kernel-bypass, compute stays off the critical network path.

4 Throughput per Core

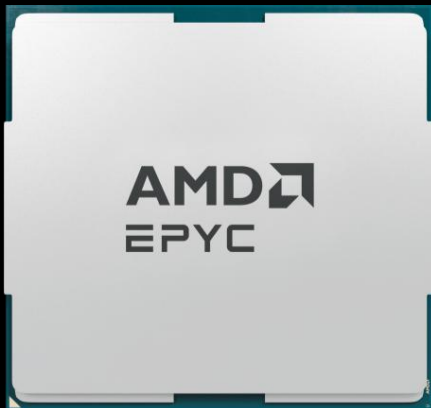
- Up to 2x vector width vs AVX2 with full-rate 512-bit execution on Zen 5 – more work per cycle, fewer cores and sockets for the same feed or strategy load.

5 Latency Determinism

- No AVX-512 clock throttling: sustained clocks mean predictable p99/p99.9 response, the metric matching engines and gateways are actually judged on.

6 Efficiency & Consolidation

- Fewer cores per workload lowers power, rack space and per-core software licensing, while high-frequency EPYC and Threadripper SKUs keep single-thread hot paths fast.



Introducing 6th Gen AMD EPYC “Venice”

World’s best server CPUs for Cloud, Enterprise, and AI

EPYC™ 9006 SP7

“Venice” SP7

PERFORMANCE

Leadership Core Count, Multi
Core Performance
Up to 256 Cores
96 Cores at up to 5.15Ghz

EPYC™ 9006 SP8

“Venice” SP8

ENTERPRISE

Optimized for Performance
and Latency

EPYC™ 9006X SP7

“Venice-X”

HPC

Optimized for High
Performance Computing
and AI Pre-processing

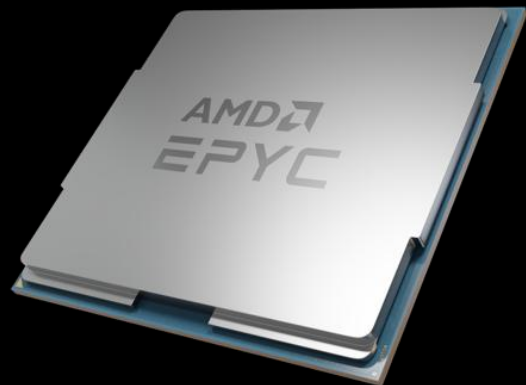
EPYC™ 9006 LP

“Verano”

AI HOST NODE + LPDDR

Optimized for high-
performance AI Host Node

See Endnotes: EPYC-029E



Formerly Codenamed "Venice" SP8

EPYC™ 9006 SP8 Server CPU

Broadest data center coverage

Cover the broadest range of data center workloads while optimizing system economics with balanced compute, memory, and I/O

BROAD ENTERPRISE DEPLOYMENT	BALANCE MEMORY BANDWIDTH & CAPACITY	EXPAND PLATFORM I/O	DEPLOYMENT FLEXIBILITY
8 to 128 Cores	8 Channels + 2DPC	PCIe® Gen 6.0 128 PCIe lanes	1P & 2P, HF & NEBS-Friendly

Note: NEBS-friendly OPN available.

6th Gen AMD EPYC™ 9006 “Venice” Portfolio Availability

“Venice” SP7 in production now!



Platforms from industry leading providers will be available as follows:

9006 SP7

Shipping in Q4 '26

9006 SP8

Shipping in 1H '27

9006X SP7

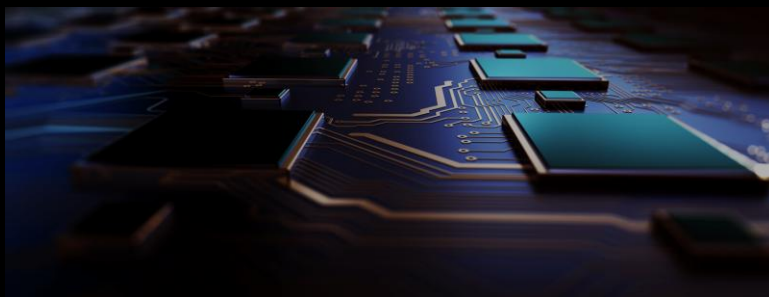
Shipping in 2H '27

9006 LP

Shipping in 2H '27

Subject to change.

AMD EPYC™ Delivers Intelligent Power Management and QoS



UPP

(Unified Power Provisioning)

One power budget for SoC and DIMMs: designed to free rack power for added server density with limited perf impact

Electronic Trading Benefit

Shared SoC/DIMM power budget funds larger in-memory order books and tick stores, so gateways and risk servers scale capacity without a rack-power upgrade.



UBPS

(Utilization Based Power Scaling)

Power that scales with utilization: lower at light loads, rising to full power to help sustain perf under heavier loads

Electronic Trading Benefit

More servers per rack in constrained colo cages – matching engines and market data gateways idle cheap between bursts, then draw full power for the open, close and volatility spikes.



FAST

(Frequency Adaptive Spectrum Technology)

Two CPU personalities in one SoC: prioritize the cores that matter, help maintain perf under power constraints

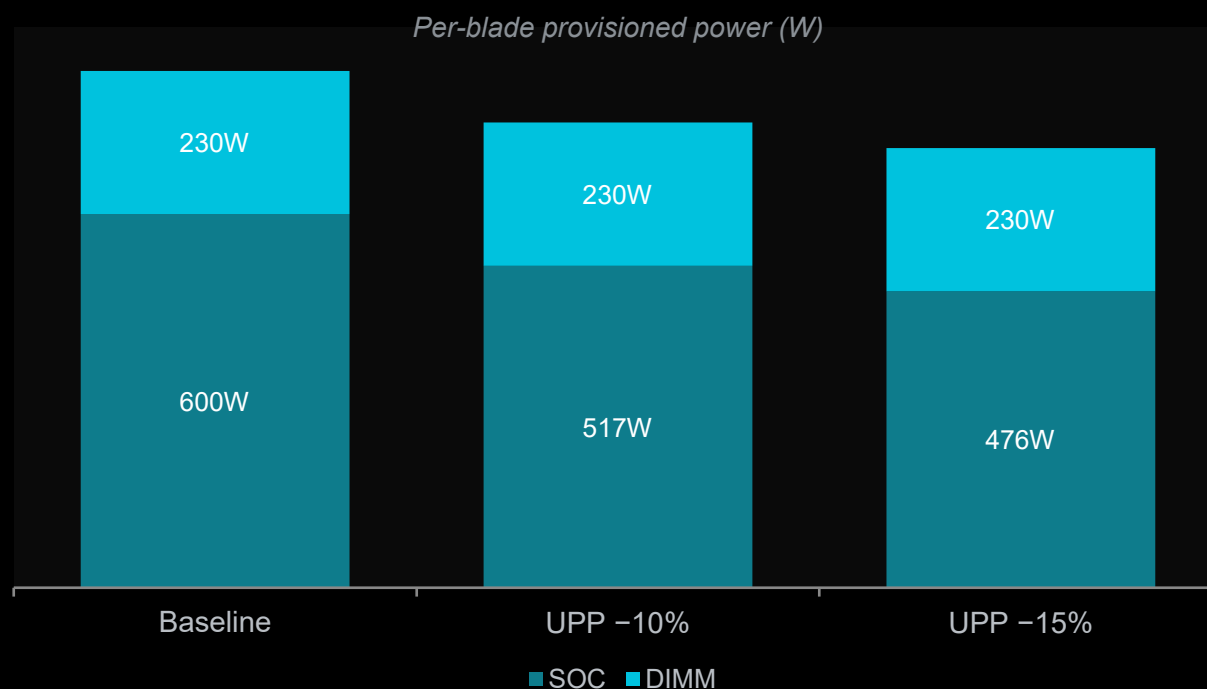
Electronic Trading Benefit

Pin latency-critical threads to boosted cores while background analytics run on the rest – deterministic p99 hot paths and full utilization on one socket.

UPP · THE PAYOFF

More servers per rack, *same power*

Provisioning to typical SOC + DIMM power shrinks per-blade watts and frees rack capacity for more servers — with a fraction of a percent performance impact.



+12%

servers per rack 16 → 18 servers in the same 35 kW rack envelope.

830W → 706W

per-blade power Combined SOC + DIMM envelope at the -15% setting.

~0–1%

avg. performance cost No measurable loss at -10%; ~1% at -15% (SPECint@-class mix).

Representative results; actual gains depend on workload mix and DIMM power profile. Rack assumptions: 35 kW, \$0.12/kWh.

FEATURE UBPS · EFFICIENCY

Let power *follow the load*

UBPS scales socket power with active-core utilization. Set an efficient power ceiling for the low-utilization band you run at most of the time; power ramps to full only as utilization climbs.



Efficient where you live. A lower power ceiling across your typical utilization band.



Full power on burst. Power scales up with utilization, preserving peak SLA under load.



Plan peak, pay typical. Provision for 100% cores but draw power based on real use.

UTILIZATION-BASED POWER SCALING

~25%

Average socket power saved at typical utilization — for a small cost that appears only at low load, not on burst.

BETTER TOGETHER

Three levers, *one AMD EPYC 9006 platform*

Each lever targets one operator pain point — stranded rack power, idle-time bleed, and SKU sprawl — all firmware-managed on the same silicon.

UPP

RECLAIM STRANDED POWER

+12% servers / rack

Provision to typical, not worst-case SOC + DIMM power — 830 W to 706 W per blade.

UBPS

POWER SCALES WITH USE

~25% avg. power saved

Efficient at low utilization; full 600 W on burst, so peak SLA is preserved.

FAST

ONE SOC, MULTIPLE TIERS

1 OPN to premium + batch

Two performance tiers share one physical socket; re-tier in real time.

UPP: Unified Power Provisioning · UBPS: Utilization-Based Power Scaling · FAST: Frequency Adaptive Spectrum Technology

GET STARTED

Pilot in days. Scale *with confidence*

1 Measure

Check provisioned vs. consumed power and real utilization across representative instance classes.

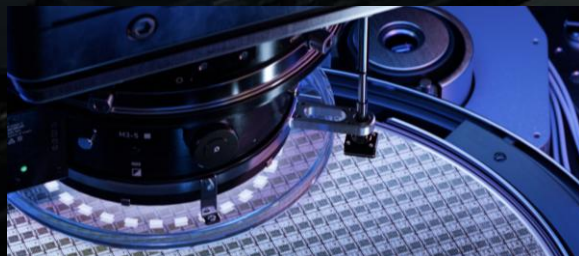
2 Pilot

Enable on a few instance classes, validate with live telemetry, and set guardrails to protect SLAs.

3 Scale

Roll out fleet-wide in measured stages — no rack, PDU, or mechanical changes required.

AMD EPYC™ is Designed to Secure the Modern Datacenter



Device Provenance

Attestable provenance for the complete AMD device manufacturing lifecycle



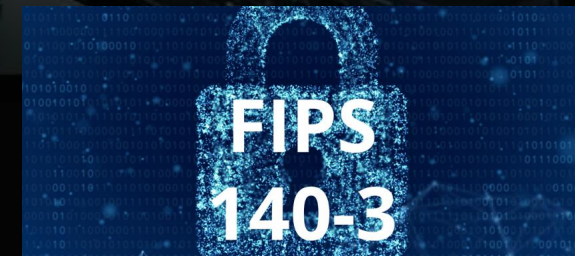
Root of Trust

Hardware-anchored platform integrity from power-on through runtime



Confidential Computing

Isolate and encrypt workloads in hardware-backed secure environments



FIPS 140-3 Certification

Provide assurance for effectiveness of AMD's cryptographic hardware

Note: "Venice" is planned to be among the first AMD products to support AMD Device Provenance Services.

AMD Multi-year Investment Building Confidential Computing

CPU	Transparent Secure Memory Encryption	SEV	Encrypted State	Secure Nested Paging	Trusted I/O
AMD EPYC™ 7002	✓	✓	✓		
AMD EPYC™ 7003	✓	✓	✓	✓	
AMD EPYC™ 8004, 8005, & 9004	✓	✓	✓	✓	
AMD EPYC™ 9005 & 9006	✓	✓	✓	✓	✓

“Venice” + AVX-VNNI-INT8: AI at the Trading Edge

1 Matching Engine – Smarter Risk & Surveillance

- AVX-VNNI-INT8 folds multiply-accumulate into a single instruction, so quantized models for pre-trade risk, self-match prevention and anomaly detection run inline on the CPU. No PCIe hop to an accelerator means inference stays inside the matching-engine budget, in microseconds.

2 Market Data Gateway – Inline Inference on the Feed

- INT8 dot-products classify order flow, detect feed anomalies and score toxicity as messages are parsed. Venice's Zen 6 core count and higher memory bandwidth (12-channel DDR5 / MRDIMM class) keep gateways fed at multi-10M msg/s without adding a separate inference tier.

3 Trading Applications – Alpha & Execution

- Quantized signal models, short-horizon price prediction and smart-order-routing policies run at INT8 with FP32-class accuracy for inference, delivering multiple-x throughput vs FP32 on the same socket – more strategies and more features per server.

4 Throughput Without an Accelerator

- AVX-VNNI-INT8 fuses the INT8 multiply-add sequence into one instruction, raising inference-per-core sharply while keeping models on the same cores as the trading hot path.

5 Consolidation & TCO

- Venice's higher core density, DDR5 bandwidth and PCIe Gen6 I/O collapse compute, feed handling and inference into fewer sockets – less power, rack space and per-core licensing.

6 Ready for the Next Cycle

- Embedded “Venice” SP7/SP8 offer long-life supply and platform continuity for exchange and venue infrastructure, with AVX-512 and AVX-VNNI-INT8 on the same socket.

Consolidate Legacy Infrastructure with Fewer EPYC™ CPU-Powered Servers

1000 Outdated Servers
2P Intel® Xeon® Gold 6258R servers



~12:1 Consolidation

82 AMD EPYC™ Servers
2P AMD EPYC™ 9996 servers



Consolidate with confidence

- Proven deployment path
- Compatible x86 foundation
- Mature ecosystem

Up to **77%**
Less power

Up to **92%**
fewer servers

Up to **40%**
lower 5-year TCO

Servers required to achieve an estimated score of ~397,000 total of units of SPECrate2017_int_base_performance.
See Endnotes: 9xx6TCO-001

6th Gen AMD EPYC™ 9006 “Venice” Portfolio Availability

“Venice” SP7 in production now!



Platforms from industry leading providers will be available as follows:

9006 SP7

Shipping in Q4 '26

9006 SP8

Shipping in 1H '27

9006X SP7

Shipping in 2H '27

9006 LP

Shipping in 2H '27

Subject to change.

Agentic AI Deployment Will Be Distributed to Address These Challenges

FRONTIER



CLOUD



ON PREM



CLIENT



Agentic AI Deployment Will Be Distributed to Address These Challenges

ON PREM



CONSTRAINTS

Power & Cooling

Cost

Ease of Deployment



AMD Instinct MI350P

The Easiest Way to Add AI to Your Data Center

**PERF/\$
LEADERSHIP**

Up to **4.2x**

Tokens/second/\$ vs RTX Pro 6000

**LARGE MODEL
CAPABLE**

Up to **260 Billion**

Model parameters @ FP4

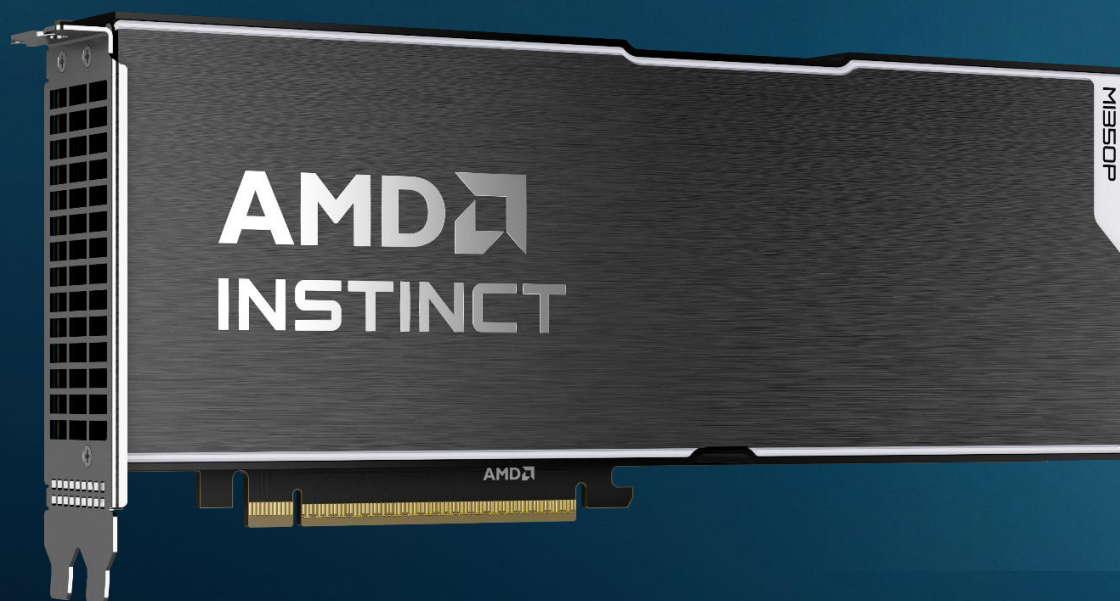
**STANDARD SERVER
COMPATIBLE**

Air Cooled

450-600W TBP

AMD Instinct MI350P

Delivering Higher Productivity



Tokens/Second vs RTX Pro 6000

WORKFLOW AUTOMATION

Llama 3.3 70B Instruct

5.1x

CUSTOMER SUPPORT

Mistral Small 3.2 24B

2.5x

CONTRACT ANALYSIS

GPT-OSS 120B

2.1x

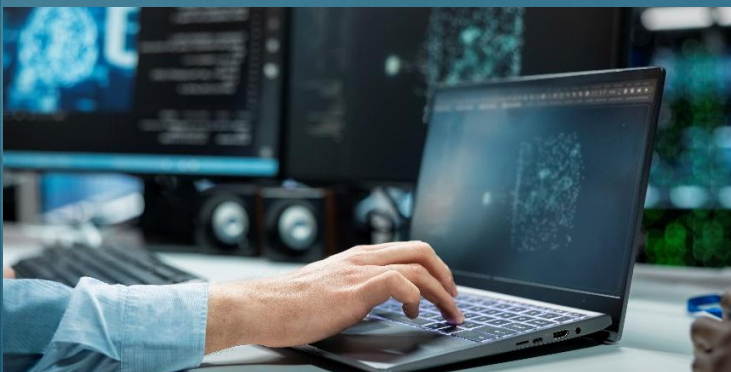
Why Local AI is Taking Off

SECURITY & PRIVACY



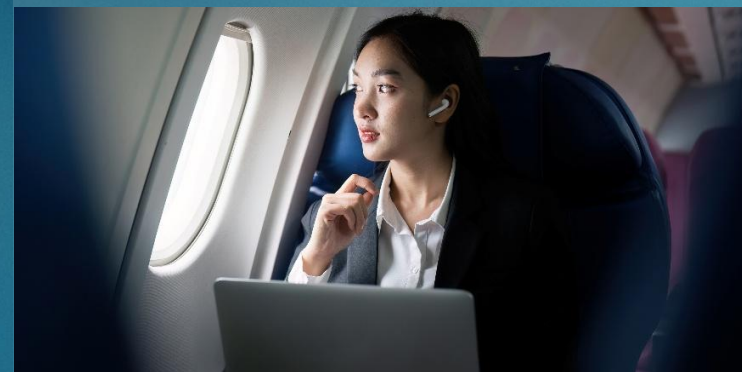
Full control over data and
no third-party handling

“FREE TOKEN” INFERENCE



Optimize recurring
inference costs

RELIABLE ANYWHERE



Continue working
without connectivity

Small Models are Becoming Incredibly Efficient

GPT-OSS



Number of Parameters

120B

GPQA Score*

Graduate-Level Google-Proof Q&A

80.1

August 2025

Qwen 3.5



Number of Parameters

9B

GPQA Score*

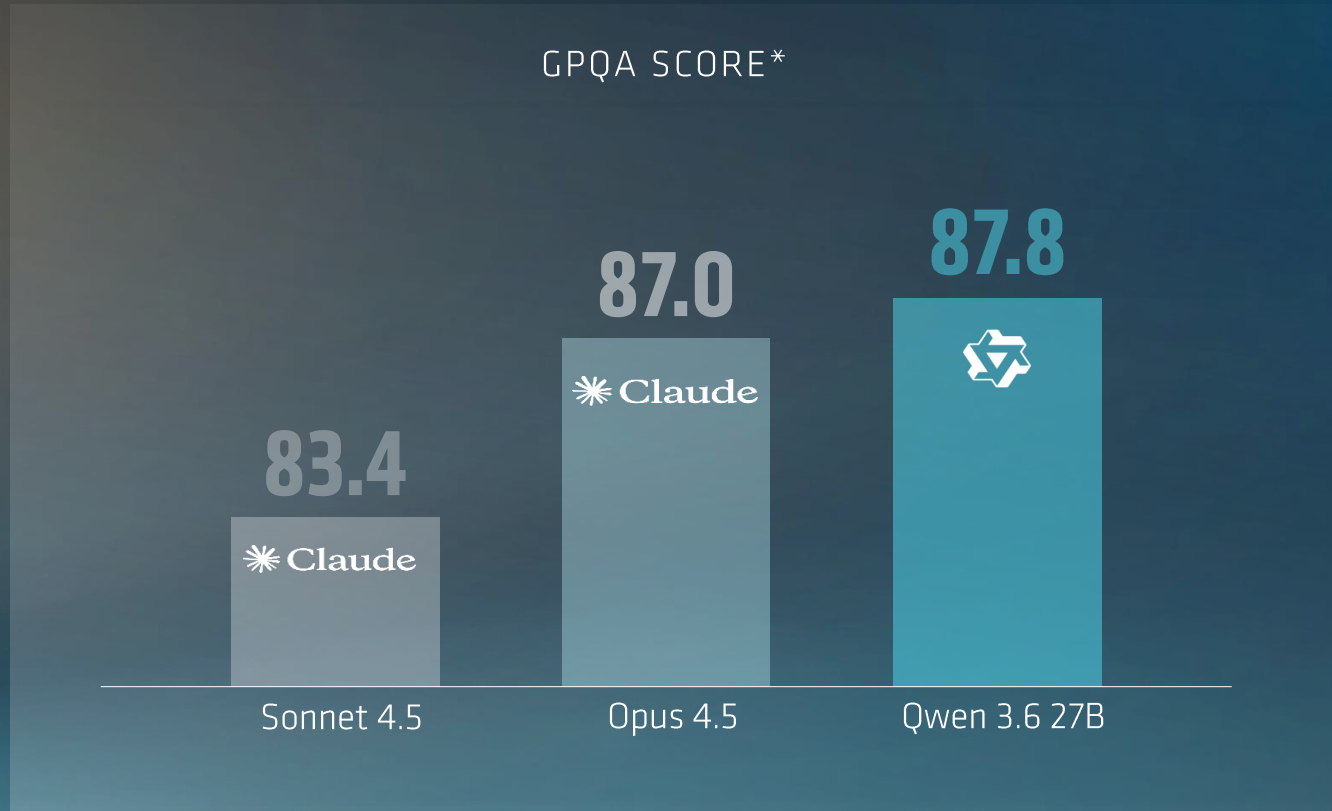
Graduate-Level Google-Proof Q&A

81.7

March 2026

Source: LLM Stats, May 20, 2026 (<https://llm-stats.com/models/compare/qwen3.5-9b-vs-gpt-oss-120b>)

Smaller Models are Closing The Frontier Gap



Smaller open models can now outperform yesterday's frontier models while using a fraction of the cost.

Source: LLM Stats, May 20, 2026 (<https://llm-stats.com/models/compare/claude-opus-4-5-20251101-vs-qwen3.6-27b>)

Efficient Models Unlock Personal AI

Models Fit Across AMD Agentic PCs and Data Center GPUs



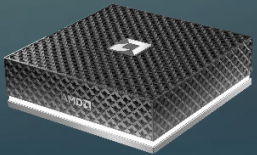
Ryzen™ AI 400

9B

Model Size

24B

Model Size Limit



Ryzen™ AI MAX

27B

Model Size

200B

Model Size Limit



Instinct™ MI350P

260B

Model Size

260B

Model Size Limit



together we advance_

Pushing inference further with “GORGON HALO”

Largest unified memory
from **128GB to 192GB**

Extending AI models
from **200B to 300B**

Run cloud-scale AI
Locally on Your Desktop



Disclaimer

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18u.

© 2026 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, EPYC, EnterpriseOnload, Ryzen, Onload, OpenOnload, Solarflare, Solarflare logo, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective owners. Certain AMD technologies may require third-party enablement or activation. Supported features may vary by operating system. Please confirm with the system manufacturer for specific features. No technology or product can be completely secure.

Endnotes

9xx5-251: Solarflare Generational latency progression: AMD Internal Testing as of 9/15/2025. Workload config: Onload benchmarking tool measuring Latency for payload ranging from 0B to 648B in ns. System Config: 2 Supermicro systems consists of 1 AMD EPYC 9575F (64C, 12x 32GB DDR5-6400, Max speed of 5GHz), Ubuntu 25.04 with kernel 6.14-23, Custom BIOS version 5.35 (pre-release), OS configured with AMDLL profile and BIOS configured for low latency) with AMD Solarflare X4542 (bundle 1.0.3.2 and bootloader 1.0.17.0). AMD Solarflare X2522 (bundle 8.5.2.1000); AMD Alveo X3522 (bundle 1.17.7.30). Release Candidate version of Onload is used (<https://github.com/Xilinx-CNS/onload/commit/0395821bb0685a3019018b3a9b45c4f30915146f>). Driver version 6.6.1.1005. AMD Low Latency Tuning guide available at https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/white-papers/58649_amd-epyc-tg... Average 0 to 648B Latency comparison compared to X4: X3 is 1.27X (27.2%) higher and X2 is 1.43X (42.9%) higher latency 4B latency comparison compared to X4: X3 is 1.35X (34.9%) higher and X2 is 1.55X (55.5%) higher latency.

9xx5-252: X4 latency across platforms: AMD Internal Testing as of 9/15/2025. Workload config: Onload benchmarking tool measuring Latency for payload ranging from 0B to 648B in ns. System Config: 2 Supermicro systems consists of one AMD EPYC 9575F (64C, 12x 32GB DDR5-6400, Max speed of 5GHz), Ubuntu 25.04 with kernel 6.14-23, Custom BIOS version 5.35 (pre-release), OS configured with AMDLL profile and BIOS configured for low latency) with AMD Solarflare X4542 (bundle 1.0.3.2 and bootloader 1.0.17.0). Release candidate version of onload (<https://github.com/Xilinx-CNS/onload/commit/0395821bb0685a3019018b3a9b45c4f30915146f>). Driver version 6.1.1.1005. AMD Solarflare X4542 (bundle 1.0.3.2 and bootloader 1.0.17.0) AMD Low Latency Tuning guide available at https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/white-papers/58649_amd-epyc-tg... System config: 2 Supermicro systems consists of one AMD EPYC 4564P (AS -1015A-MT - 16C, up to 5GHz, 48GB memory). Driver version 6.1.1.1005, Custom BIOS version dated 8/30/24, kernel version 5.14.0-503. OS RHEL 9.4. AMD Solarflare X4542 (bundle 1.0.3.2 and bootloader 1.0.17.0) Driver version 6.1.1.1009 System config: 2 Dell R760 with one Intel Xeon 6534 (8 Core, 4.2GHz, 128GB,), bios version 2.2.4, Ubuntu 25.04 with kernel 6.14.0-23, AMD Solarflare X4542 (bundle 1.0.3.2 and bootloader 1.0.17.0) driver version 6.1.1.1009, 2 NUMA nodes, 4B Payload size AMD Solarflare X4542 adapter installed on 2x 1P Intel Xeon 6534 based Supermicro system produced higher latency by 1.04X (4% higher) (23.7ns) compared to 2x 1P 9575F based Supermicro system AMD Solarflare X4542 adapter installed on 2x 1P EPYC 4564P based Supermicro system produced reduced latency by 0.91X (9% lower) (54ns) compared to 2x 1P 9575F based Supermicro system 0 to 648B Payload size AMD Solarflare X4542 adapter installed on 2x 1P Intel Xeon 6534 based Supermicro system produced lower latency by 1.02X (13ns) (1.6%) compared to 2x 1P 9575F based Supermicro system AMD Solarflare X4542 adapter installed on 2x 1P EPYC 4564P based Supermicro system produced lower latency by 1.07X (52ns) (6.7%) compared to 2x 1P 9575F based Supermicro system

9xx5-253: SDCI improvements: AMD Internal Testing as of 9/15/2025. Workload config: Onload benchmarking tool measuring Latency for payload ranging from 0B to 648B in ns. System Config: 2 Supermicro systems consist of 1 AMD EPYC 9575F (64C, 12x 32GB DDR5-6400, Max speed of 5GHz, Ubuntu 25.04 with kernel 6.14-23, Custom BIOS version 5.35 (pre-release), OS configured with AMDLL profile and BIOS configured for low latency) with AMD Solarflare X4542 (bundle 1.0.3.2 and bootloader 1.0.17.0). Release Candidate version of Onload is used (<https://github.com/Xilinx-CNS/onload/commit/0395821bb0685a3019018b3a9b45c4f30915146f>). Driver version 6.6.1.1005 AMD Low Latency Tuning guide available at https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/white-papers/58649_amd-epyc-tg... 256B latency comparison with and without SDCI: eflatency: 46ns (5.86%) lower latency with SDCI; TCPDirect-TCP 32ns (3.8%) lower latency with SDCI; Onload TCP 73ns (7.5%) lower latency with SDCI . 0 to 648B latency comparison with and without SDCI: eflatency: 55ns (6.5%) lower latency with SDCI; TCPDirect-TCP 44ns (4.89%) lower latency with SDCI; Onload TCP 67ns (6.49%) lower latency with SDCI

9xx5-254: Competitive performance: AMD Internal Testing as of 9/15/2025. Workload config: Onload benchmarking tool measuring Latency for payload ranging from 0B to 648B in ns. System Config: 2 Supermicro systems consist of 1 AMD EPYC 9575F (64C, 12x 32GB DDR5-6400, Max speed of 5GHz, Ubuntu 25.04 with kernel 6.14-23, Custom BIOS version 5.35 (pre-release), OS configured with AMDLL profile and BIOS configured for low latency) with AMD Solarflare X4542 (bundle 1.0.3.2 and bootloader 1.0.17.0). Release candidate version of onload (<https://github.com/Xilinx-CNS/onload/commit/0395821bb0685a3019018b3a9b45c4f30915146f>). Driver version 6.1.1.1005. AMD Solarflare X4542 (bundle 1.0.3.2 and bootloader 1.0.17.0) AMD Low Latency Tuning guide available at https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/white-papers/58649_amd-epyc-tg-low-latency.pdf System config: Two Dell R760 with Intel Xeon 6534 (8 Core, 4.2GHz, 128GB,), bios version 2.2.4, Ubuntu 25.04 with kernel 6.14.0-23, AMD Solarflare X4542 (bundle 1.0.3.2 and bootloader 1.0.17.0) driver version 6.1.1.1009, 2 NUMA nodes, 0 to 648B Payload size Ef_vi (Eflatency): AMD Solarflare X4542 adapter installed on 2x 1P Intel Xeon 6534 based Supermicro system produced lower latency by 4ns (0.5%) compared to 2x 1P 9575F based Supermicro system TCPDirect-TCP (zftcpinping) AMD Solarflare X4542 adapter installed on 2x 1P EPYC 9575F based Supermicro system produced lower latency by 1ns (0.1%) compared to 2x 1P Intel Xeon 6534 based Supermicro system Onload-TCP (sfntpingpong) AMD Solarflare X4542 adapter installed on 2x 1P EPYC 9575F based Supermicro system produced lower latency by 14ns (1.3%) compared to 2x 1P Intel Xeon 6534 based Supermicro system.

Endnotes

EPYC-018: Max boost for AMD EPYC processors is the maximum frequency achievable by any single core on the processor under normal operating conditions for server systems.

EPYC-028C: SPECpower_ssj® 2008, SPECrate®2017_int_energy_base, and SPECrate®2017_fp_energy_base based on results published on SPEC's website as of 11/10/22. 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 80 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 11 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 four 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-028C> for the full list. 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-049: AMD EPYC 9754 is a 128 core dual threaded CPU and in a 2 socket server with 1 thread per vCPU delivers 512 vCPUs per EPYC powered server which is more than any Ampere or 4 socket Intel CPU based server as of 6/13/2023.

GD-83: Use or mention of third-party marks, logos, products, services, or solutions herein is for informational purposes only and no endorsement by AMD is intended or implied.

GD-122: The information contained herein is for informational purposes only, and is subject to change without notice. Timelines, roadmaps, and/or product release dates shown herein are plans only and subject to change. [Insert codenames with quotation marks, i.e. "Zen"; ""Zen 2"""] are codenames for AMD architectures, and are not product names. GD-122

GD-150: Boost Clock Frequency is the maximum frequency achievable on the CPU running a bursty workload. Boost clock achievability, frequency, and sustainability will vary based on several factors, including but not limited to: thermal conditions and variation in applications and workloads. GD-150.

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/technologies/infinity-guard>. GD-183A

9xx5-048: AMD EPYC™ 9005 Series processors require OEM enablement and a BIOS update from your server or motherboard manufacturer if used with a motherboard designed for the SP5 socketed AMD EPYC™ 9004 Series processors. Contact your system manufacturer prior to purchase to determine compatibility.

9XX5-072: CXL Type 1&2 devices and PCIe link encryption support dependent upon ecosystem readiness; type 2 PoC only

Endnotes

9xx5-251: Solarflare Generational latency progression: AMD Internal Testing as of 9/15/2025. Workload config: Onload benchmarking tool measuring Latency for payload ranging from 0B to 648B in ns. System Config: 2 Supermicro systems consists of 1 AMD EPYC 9575F (64C, 12x 32GB DDR5-6400, Max speed of 5GHz), Ubuntu 25.04 with kernel 6.14-23, Custom BIOS version 5.35 (pre-release), OS configured with AMDLL profile and BIOS configured for low latency) with AMD Solarflare X4542 (bundle 1.0.3.2 and bootloader 1.0.17.0). AMD Solarflare X2522 (bundle 8.5.2.1000); AMD Alveo X3522 (bundle 1.17.7.30). Release Candidate version of Onload is used (<https://github.com/Xilinx-CNS/onload/commit/0395821bb0685a3019018b3a9b45c4f30915146f>). Driver version 6.6.1.1005. AMD Low Latency Tuning guide available at https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/white-papers/58649_amd-epyc-tg... Average 0 to 648B Latency comparison compared to X4: X3 is 1.27X (27.2%) higher and X2 is 1.43X (42.9%) higher latency 4B latency comparison compared to X4: X3 is 1.35X (34.9%) higher and X2 is 1.55X (55.5%) higher latency.

9xx5-252: X4 latency across platforms: AMD Internal Testing as of 9/15/2025. Workload config: Onload benchmarking tool measuring Latency for payload ranging from 0B to 648B in ns. System Config: 2 Supermicro systems consists of one AMD EPYC 9575F (64C, 12x 32GB DDR5-6400, Max speed of 5GHz), Ubuntu 25.04 with kernel 6.14-23, Custom BIOS version 5.35 (pre-release), OS configured with AMDLL profile and BIOS configured for low latency) with AMD Solarflare X4542 (bundle 1.0.3.2 and bootloader 1.0.17.0). Release candidate version of onload (<https://github.com/Xilinx-CNS/onload/commit/0395821bb0685a3019018b3a9b45c4f30915146f>). Driver version 6.1.1.1005. AMD Solarflare X4542 (bundle 1.0.3.2 and bootloader 1.0.17.0) AMD Low Latency Tuning guide available at https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/white-papers/58649_amd-epyc-tg... System config: 2 Supermicro systems consists of one AMD EPYC 4564P (AS -1015A-MT - 16C, up to 5GHz, 48GB memory). Driver version 6.1.1.1005, Custom BIOS version dated 8/30/24, kernel version 5.14.0-503. OS RHEL 9.4. AMD Solarflare X4542 (bundle 1.0.3.2 and bootloader 1.0.17.0) Driver version 6.1.1.1009 System config: 2 Dell R760 with one Intel Xeon 6534 (8 Core, 4.2GHz, 128GB,), bios version 2.2.4, Ubuntu 25.04 with kernel 6.14.0-23, AMD Solarflare X4542 (bundle 1.0.3.2 and bootloader 1.0.17.0) driver version 6.1.1.1009, 2 NUMA nodes, 4B Payload size AMD Solarflare X4542 adapter installed on 2x 1P Intel Xeon 6534 based Supermicro system produced higher latency by 1.04X (4% higher) (23.7ns) compared to 2x 1P 9575F based Supermicro system AMD Solarflare X4542 adapter installed on 2x 1P EPYC 4564P based Supermicro system produced reduced latency by 0.91X (9% lower) (54ns) compared to 2x 1P 9575F based Supermicro system 0 to 648B Payload size AMD Solarflare X4542 adapter installed on 2x 1P Intel Xeon 6534 based Supermicro system produced lower latency by 1.02X (13ns) (1.6%) compared to 2x 1P 9575F based Supermicro system AMD Solarflare X4542 adapter installed on 2x 1P EPYC 4564P based Supermicro system produced lower latency by 1.07X (52ns) (6.7%) compared to 2x 1P 9575F based Supermicro system

9xx5-253: SDCL improvements: AMD Internal Testing as of 9/15/2025. Workload config: Onload benchmarking tool measuring Latency for payload ranging from 0B to 648B in ns. System Config: 2 Supermicro systems consist of 1 AMD EPYC 9575F (64C, 12x 32GB DDR5-6400, Max speed of 5GHz, Ubuntu 25.04 with kernel 6.14-23, Custom BIOS version 5.35 (pre-release), OS configured with AMDLL profile and BIOS configured for low latency) with AMD Solarflare X4542 (bundle 1.0.3.2 and bootloader 1.0.17.0). Release Candidate version of Onload is used (<https://github.com/Xilinx-CNS/onload/commit/0395821bb0685a3019018b3a9b45c4f30915146f>). Driver version 6.6.1.1005 AMD Low Latency Tuning guide available at https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/white-papers/58649_amd-epyc-tg... 256B latency comparison with and without SDCL: eflatency: 46ns (5.86%) lower latency with SDCL; TCPDirect-TCP 32ns (3.8%) lower latency with SDCL; Onload TCP 73ns (7.5%) lower latency with SDCL. 0 to 648B latency comparison with and without SDCL: eflatency: 55ns (6.5%) lower latency with SDCL; TCPDirect-TCP 44ns (4.89%) lower latency with SDCL; Onload TCP 67ns (6.49%) lower latency with SDCL

9xx5-254: Competitive performance: AMD Internal Testing as of 9/15/2025. Workload config: Onload benchmarking tool measuring Latency for payload ranging from 0B to 648B in ns. System Config: 2 Supermicro systems consist of 1 AMD EPYC 9575F (64C, 12x 32GB DDR5-6400, Max speed of 5GHz, Ubuntu 25.04 with kernel 6.14-23, Custom BIOS version 5.35 (pre-release), OS configured with AMDLL profile and BIOS configured for low latency) with AMD Solarflare X4542 (bundle 1.0.3.2 and bootloader 1.0.17.0). Release candidate version of onload (<https://github.com/Xilinx-CNS/onload/commit/0395821bb0685a3019018b3a9b45c4f30915146f>). Driver version 6.1.1.1005. AMD Solarflare X4542 (bundle 1.0.3.2 and bootloader 1.0.17.0) AMD Low Latency Tuning guide available at https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/white-papers/58649_amd-epyc-tg-low-latency.pdf System config: Two Dell R760 with Intel Xeon 6534 (8 Core, 4.2GHz, 128GB,), bios version 2.2.4, Ubuntu 25.04 with kernel 6.14.0-23, AMD Solarflare X4542 (bundle 1.0.3.2 and bootloader 1.0.17.0) driver version 6.1.1.1009, 2 NUMA nodes, 0 to 648B Payload size Ef_vi (Eflatency): AMD Solarflare X4542 adapter installed on 2x 1P Intel Xeon 6534 based Supermicro system produced lower latency by 4ns (0.5%) compared to 2x 1P 9575F based Supermicro system TCPDirect-TCP (zftcpinping) AMD Solarflare X4542 adapter installed on 2x 1P EPYC 9575F based Supermicro system produced lower latency by 1ns (0.1%) compared to 2x 1P Intel Xeon 6534 based Supermicro system Onload-TCP (sfntpingpong) AMD Solarflare X4542 adapter installed on 2x 1P EPYC 9575F based Supermicro system produced lower latency by 14ns (1.3%) compared to 2x 1P Intel Xeon 6534 based Supermicro system.

