AMD
together we advance_

# END-TO-END AI SOLUTIONS

## ADVANCING AI INFRASTRUCTURE IN THE DATA CENTER

## TOP AI CHALLENGES IN THE DATA CENTER

### Hardware[1]

1. High costs of specialized hardware
2. Incompatabilities with existing hardware
3. Insufficient processing power
4. Inadequate storage capacity
5. Networking limitations
6. Insufficient cooling capabilities

### Technical[1]

1. IT integration
2. Lack of in-house expertise
3. Software issues
4. Finding the right partner
5. Concerns over IP loss
6. Energy demands

## AMD SOLVES KEY AI CHALLENGES

**Modernizes data centers for AI infrastructure**

Up to **2.7x** SPEC® CPU performance[2]

Up to **20%** more AI inference throughput[3]

Up to **87%** fewer servers[4]

Up to **68%** less energy consumption[4]

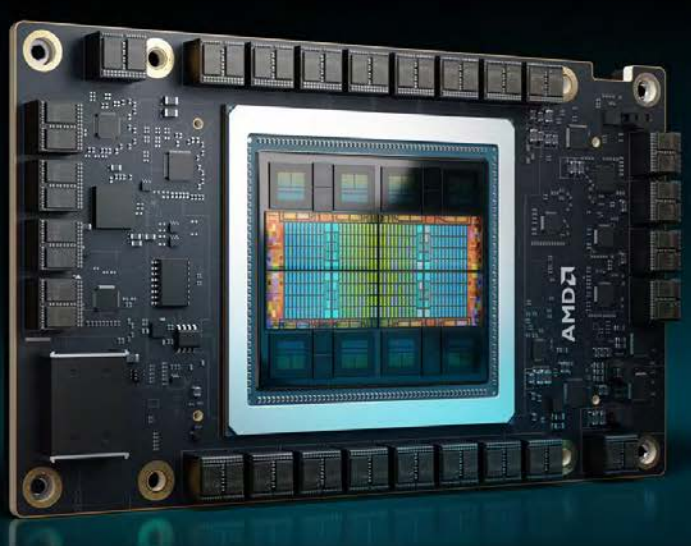
**Addresses data management for AI workloads**

**Optimizes AI innovation**

Up to **$3.4M** energy savings over 5 years[5]

Up to **67%** lower TCO over 5 years[4]

## AI LEADERSHIP FOR NEXT GENERATION DATA CENTERS



**AMD Instinct™ GPUs**
New standards in AI acceleration

**AMD EPYC™ Processors**
Best CPU for AI

**AMD Pensando™**
High performance AI networking

**AMD ROCm™ Software**
Open, proven, ready software stack

**ZT Systems**
Extreme AI systems deployment expertise

**AMD Infinity Guard**
Advanced modern security features built-in at the silicon level

**AMD Ryzen™ AI PRO Processors**
Built-in AI engines for optimum acceleration

**Broadest portfolio of AI | Proven leadership in innovation | Open ecosystem approach | Solutions**

## PARTNER WITH AMD FOR CONFIDENT INNOVATION AND MAXIMUM ROI

AMD powers the world's fastest and most energy-efficient supercomputers – El Capitan (#1), Frontier (#2)

### Accelerate business outcomes

- Prepare for AI faster
- Scale AI faster
- Run models faster

### Enable sustained success

- Minimize operating costs
- Integrate security throughout
- Stay flexible with open standards

AMD is offered by the world's largest infrastructure providers – T1 CSPs, neo-cloud CSPs, top enterprise/ community platforms

AMD is entrusted with the most prolific mission-critical workloads – AWS, Oracle, Netflix, Meta

### Partner with confidence

- Leverage an end-to-end portfolio
- Team up with a trusted partner
- Plan with confidence

1. S&P Global Market Intelligence 451 Research. The State of Datacenter Modernization in an AI-Driven World. Commissioned by AMD. 2025. https://www.amd.com/en/solutions/data-center/insights/data-center-modernization-in-an-ai-driven-world.html.
2. SPECrate®2017_int_base comparison based on published scores from www.spec.org as of 10/10/2024. 2P AMD EPYC 9965 (3100 SPECrate®2017_int_base, 384 Total Cores, 500W TDP, $14,813 CPU$), 6.200 SPECrate®2017_int_base/CPU W, 0.200 SPECrate®2017_int_base/CPU $ (https://www.spec.org/cpu2017/results/res2024q4/cpu2017-20241004-44979.html); 2P AMD EPYC 9755 (2720SPECrate®2017_int_base, 256 Total Cores, 500W TDP, $12,984 CPU $), 5.440 SPECrate®2017_int_base/CPU W, 0.209 SPECrate®2017_in t_base/CPU $ (https://www.spec.org/cpu2017/results/res2024q4/cpu2017-20240923-44824.html); 2P AMD EPYC 9754 (1950 SPECrate®2017_int_base, 256 Total Cores, 360W TDP, $11,900 CPU $), 5.417 SPECrate®2017_int_base/CPU W, 0.164 SPECrate®2017_int_base/CPU $ (https://www.spec.org/cpu2017/results/res2023q2/cpu2017-20230522-36617.html); 2P AMD EPYC 9654 (1810 SPECrate®2017_int_base, 192 Total Cores, 360W TDP, $11,805 CPU $), 5.028SPECrate®2017_int_base/CPU W, 0.153 SPECrate®2017_int_base/CPU $ (https://www.spec.org/cpu2017/results/res2024q1/cpu2017-20240129-40896.html); 2P Intel Xeon Platinum 8592+ (1130SPECrate®2017_int_base, 128 Total Cores, 350W TDP, $11,600 CPU $) 3.229 SPECrate®2017_int_base/CPU W, 0.097 SPECrate®2017_int _base/CPU $ (https://spec.org/cpu2017/results/res2023q4/cpu2017-20231127-40064.html); 2P Intel Xeon 6780E (1410 SPECrate®2017_int_base, 288 Total Cores, 330W TDP, $11,350 CPU $) 4.273 SPECrate® 2017_int_base/CPU W, 0.124 SPECrate®2017_int_base/CPU $ (https://spec.org/cpu2017/results/res2024q3/cpu2017-20240811-44406.html). SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation.See www.spec.org for more information. Intel CPU TDP at https://ark.intel.com/. (9xx5-002E)
3. 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 TotalCores) 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 22.04.3LTS, 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=0mitigations=off, cpupower frequency-set -g performance), BIOS 2.1, (Maximum performance, SR-IOV=On), I/O Tokens Batch Size EMR Turin Relative Difference 128/128 1024 814.678 1101.966 1.353287.288 128/2048 1024 2120.664 2331.776 1.1 211.112 2048/128 96 114.954 146.187 1.272 31.233 2048/2048 64 333.325 354.208 1.063 20.833 For average throughput increase of 1.197x. When scaling to a1000 node cluster (1 node = 2 CPUs and 8 GPUs) comparing the AMD EPYC 9575F system and Intel Xeon 8592+ system: 128/128 achieves 287,288 more tokens/s, 128/2048 achieves 211,112 more tokens/s,2048/128 achieves 31,233 more tokens/s, 2048/2028 achieves 20,833 more tokens/s. Results may vary due to factors including system configurations, software versions and BIOS settings. (9xx5-014A)
4. 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 usedas 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 IntelXeon 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 electricityfactors from the 2024 International Country Specific Electricity Factors 10 –July 2024, and the United States Environmental Protection Agency 'Greenhouse Gas Equivalencies Calculator'. For additionaldetails, see https://www.amd.com/en/legal/claims/epyc.html#q=9xx5TCO-002a. (9xx5TCO-002A)
5. 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 usedas a basis for decision making over actual testing. The AMD Server & Greenhouse Gas Emissions TCO (total cost of ownership) Estimator Tool - version 1.3, compares the selected AMD EPYC™ and Intel®Xeon® CPU based server solutions required to deliver a TOTAL_PERFORMANCE of 391000 units of SPECrate®2017_int_base performance as of November 21, 2024. This estimation compares upgradingfrom a legacy 2P Intel Xeon 28 core Platinum_8280 based server with a score of 391 (https://spec.org/cpu2017/results/res2020q3/cpu2017-20200915-23984.pdf) versus 2P EPYC 9555 (128C) poweredserver with a score of 1630 (https://spec.org/cpu2017/results/res2024q4/cpu2017-20241104-45226.pdf). Environmental impact estimates made leveraging this data, using the Country / Region specificelectricity factors from Country Specific Electricity Factors - 2024, and the United States Environmental Protection Agency Greenhouse Gas Equivalencies Calculator. For additional details, see https://www.amd.com/en/legal/claims/epyc.html#q=9xx5TCO-006. (9xx5TCO-006)



AMD
together we advance_

Connection
we solve IT

**Contact your Connection Account Team for more information.**

| Business Solutions | Enterprise Solutions | Public Sector Solutions |
|---|---|---|
| 1.800.800.0014 | 1.800.369.1047 | 1.800.800.0019 |

www.connection.com/AIOps

3284021-1025