THE PUBLIC SECTOR AI REVOLUTION: ENHANCING SERVICES AND IMPROVING OUTCOMES

TRANSFORMING PUBLIC SERVICES WITH DATA-DRIVEN INSIGHTS AND IMPROVED EFFICIENCY | 2025

TRANSFORMING PUBLIC AGENCIES WITH AI AND MODERN IT: DRIVING INNOVATION AND IMPROVING SERVICES

These are challenging times for government and agency leaders in the public sector, as organizations face budget constraints, workforce crises, and strains on existing infrastructure. Even as these challenges mount, demand for essential public services continues to grow for programs that safeguard public health and safety, support for schools and improved student outcomes, and enhanced citizen safety and emergency response.

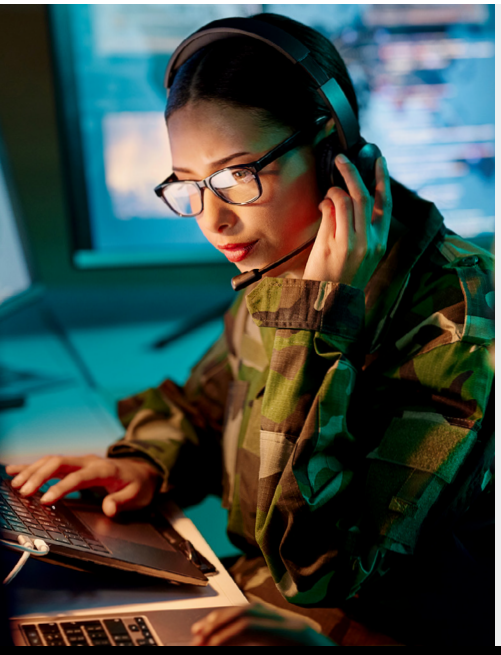
In this white paper, we explore a new generation of innovative solutions to help public sector organizations meet these challenges and introduce new ways to serve the public with improved efficiency, accessibility, and cost management. Artificial intelligence (AI) can play a transformative role in the public sector by delivering real-time insights, improving service delivery, and optimizing resource allocation.

Yet the ability to deliver these critical public services efficiently and equitably is challenged by these realities: most public agencies operate on constrained budgets, limiting their ability to invest in AI infrastructures powered by highly-sought-after, specialized GPU processors, and they lack the skilled IT workforce needed to implement and manage AI-driven systems. For many public sector leaders, it may seem that data-driven AI innovation is simply not in the budget.

To overcome these challenges and realize the potential of AI, public sector organizations need solutions that are cost effective and can modernize agency IT infrastructures to enable the deployment of transformative technologies. This is possible with AMD technology-powered systems, which help improve the performance of your data center, cloud infrastructure, and edge deployments even as you consolidate to save money, space, and energy.

AMD offers a comprehensive end-to-end portfolio of high-performance products and solutions to support AI deployment from the cloud to the data center and to the edge. These include AI-optimized AMD EPYC™ CPUs and AI-accelerated AMD Instinct™ GPUs designed for deep learning training, inference, and large-scale AI models. AMD also offers a suite of adaptive AI solutions for low-power, high-efficiency AI inference at the network edge, ideal for smart city deployments, public safety, and climate monitoring.

AMD technologies can unlock the full potential of AI in your agency to drive innovation, improve outcomes, and transform public services.



AI CAN PLAY A TRANSFORMATIVE ROLE IN THE PUBLIC SECTOR BY DELIVERING REAL-TIME INSIGHTS, IMPROVING SERVICE DELIVERY, AND OPTIMIZING RESOURCE ALLOCATION.

AI HAS THE POTENTIAL TO TRANSFORM PUBLIC SECTOR SERVICES

AI adoption in the public sector is accelerating as governments recognize its potential to enhance efficiency, decision-making, and service delivery. According to a **2024 report from the Hoover Institution**, about 25% of US civil servants are already using generative AI for public-sector work, and 44% of respondents reported they were likely or very likely to be using generative AI in the coming year.

AI's applications in government are manifold: it can help forecast economic trends, spot fraud in procurement, and improve efficiency in data processing for audits and census reports. Public sector agencies manage massive amounts of structured and unstructured data across multiple departments, and AI-driven data analytics help extract insights, detect patterns, and optimize policymaking

Some AI-powered solutions and use cases in the public sector are already in place. Specific examples of how AI is being used to increase efficiency and improving services include:

- **Government-funded healthcare.** AI can predict public health risks, facilitate diagnoses, support pandemic/epidemic modeling, mitigate misinformation, and aid with vaccine delivery. AI-driven chatbots and virtual assistants can provide 24/7 health guidance, while AI-powered remote monitoring devices can track chronic conditions.
- **Defense and national security.** AI enhances real-time surveillance with image recognition and anomaly detection to identify threats and efficiently analyze data to support strategic decisions in real time.
- **Government services.** AI can automate inquiries and document processing with AI-powered virtual assistants to support services for tax filings, license renewals, and unemployment benefits.

EMERGING AI USE CASES TO TRANSFORM THE PUBLIC SECTOR EXPERIENCE

On the horizon are new data-driven and AI-powered use cases that can deliver even more transformative public sector services.

- **Personalized citizen engagement.** AI-tailored interactions with citizens can analyze individual needs and suggest relevant services—such as local job training programs, health screenings, or tax credits. This level of personalization increases service uptake, especially for underutilized programs, and helps ensure that support reaches those who need it most.
- **Emergency response and disaster management.** AI-predictive models can forecast risks such as floods, wildfires, or disease outbreaks by analyzing environmental and social data. During crises, AI can be used for real-time mapping of affected areas, coordinating emergency responses, and sending personalized alerts to citizens based on location.
- **Virtual assistants and chatbots.** AI-driven support can handle routine citizen inquiries by answering frequently asked questions about public services such as passport renewal, social welfare applications, public transportation schedules, or tax filing procedures. Unlike traditional call centers, AI agents operate 24/7 and can handle thousands of interactions simultaneously, reducing wait times and operational costs.

NEW DATA-DRIVEN AND AI-POWERED USE CASES HAVE THE POTENTIAL TO DELIVER EVEN MORE TRANSFORMATIVE PUBLIC SECTOR SERVICES.

THE CHALLENGES OF IMPLEMENTING DATA-DRIVEN AI SOLUTIONS IN THE PUBLIC SECTOR

For many public sector leaders, the promise of AI-powered solutions remains out of reach. Many agencies operate on legacy systems that are not AI-compatible; it can be complex and disruptive to integrate AI into outdated IT infrastructure and fragmented data systems not designed for AI and advanced analytics.

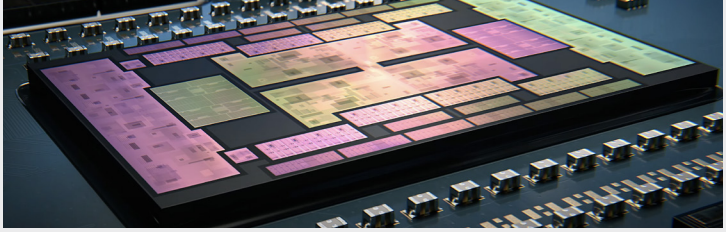
Most AI solutions rely on processing massive amounts of personal data, transaction records, and behavioral information to generate insights. However, data collection, storage, and usage raise regulatory concerns in many jurisdictions, including those with strict data sovereignty laws. For public sector organizations in these regions, data-driven AI adoption must align with local data sovereignty policies to ensure security, privacy, and compliance. This has led to the development of sovereign AI initiatives that comply with local data laws where AI models must be trained and deployed on locally controlled infrastructure within the jurisdiction's borders.

The high price of GPUs and other high-performance computing hardware, driven by supply chain constraints and increasing demand for AI, is another inhibitor to public sector AI implementation. Public sector IT teams are not typically trained in AI and data science, a skills gap that can also slow adoption and limit the impact of AI investments. In addition, AI technologies are evolving rapidly, potentially rendering investments in GPUs and AI frameworks obsolete within a few years.

With these types of challenges, it's critical for public sector organizations to choose the right technology partner, with a large ecosystem of public sector industry partners, to lead AI-driven implementations. AMD develops high performance, energy-efficient processors with advanced security features and IT infrastructure designed for AI workloads. Frontier and El Capitan, the two **most powerful**[1] supercomputers in the world, are based on AMD and being used for groundbreaking work on AI and other applications that are changing the world.

AMD is at the center of a diverse network of industry leaders including OEMs, ISVs, resellers, and developers with deep experience implementing leading-edge AI solutions for the public sector. AMD is in a commanding position to help guide agencies as they harness the potential of AI and data analytics to transform the public sector landscape.

IT'S CRITICAL FOR PUBLIC SECTOR ORGANIZATIONS TO CHOOSE THE RIGHT TECHNOLOGY PARTNER TO LEAD AI-DRIVEN IMPLEMENTATIONS.



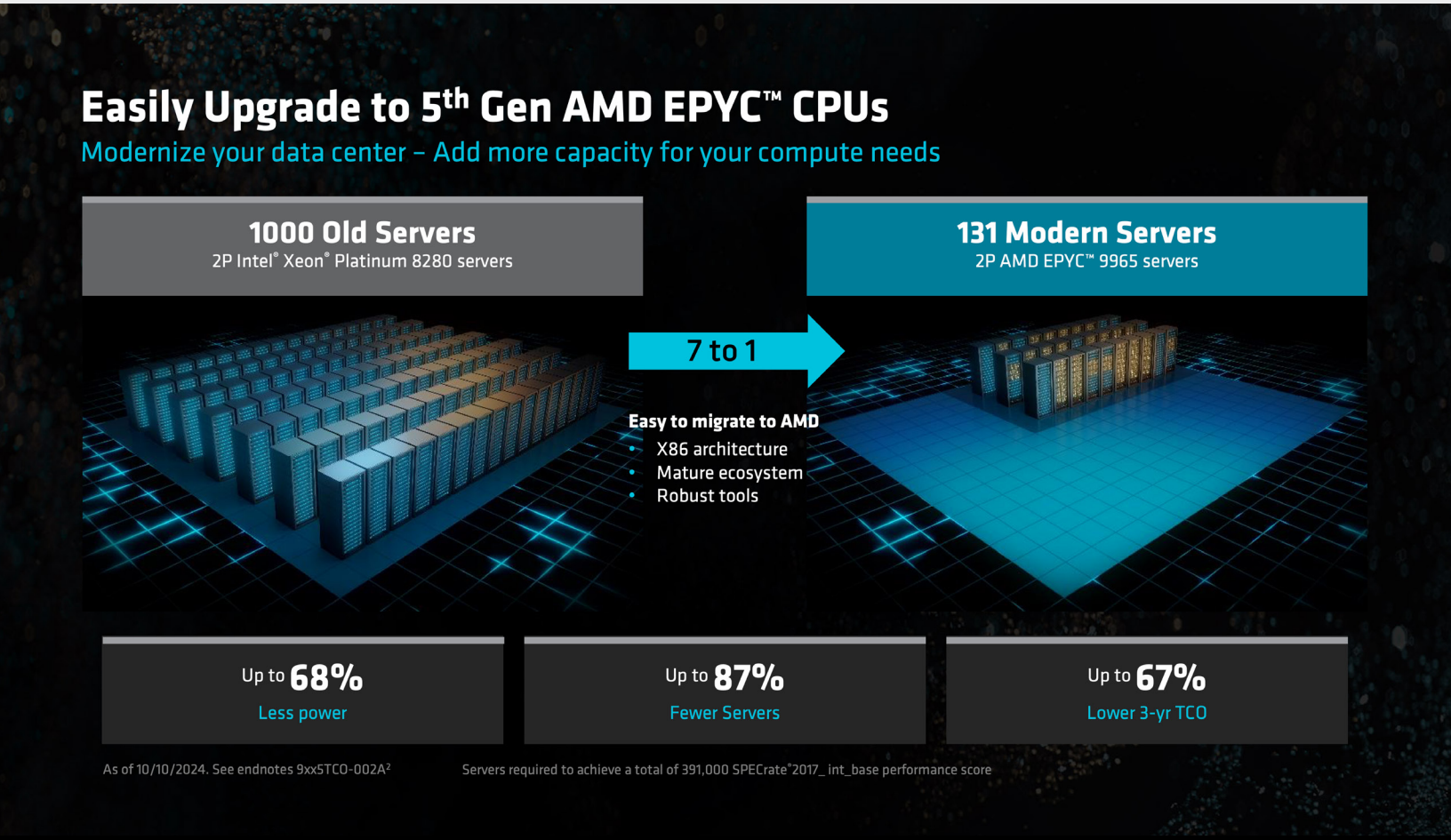
MAKING SMART CHOICES FOR PUBLIC SECTOR AI INFRASTRUCTURE

AI has the potential to transform public sector operations for both agencies and citizens, and early adopters have already harnessed its power. Others are at early or interim stages of integrating AI into their operations and face concerns about the considerable cost of AI investments, particularly related to the affordability of high-performance, high-demand GPUs.

AI workloads extend across a continuum that reflects a range of required "immediacy to insight." GPUs are at one end of the AI-processing spectrum. They are optimized for training large AI models and running large-scale inferencing workloads, and they can deliver insights in milliseconds, which is required for real-time decision-making. However, AI is not just a playfield for GPUs.

The latest generations of CPUs can handle a wide range of AI tasks alongside general-purpose workloads. AMD EPYC™ processors excel for small-to-medium AI models and workloads where proximity to data matters, offering high core counts and exceptional bandwidth. High density CPUs are more than capable of supporting a wide range of AI workloads that don't require massive parallelism, typically with lower upfront cost and power-efficiency. This makes them a cost-effective option for many AI applications including non-real-time generative AI workloads or solutions that determine dynamic pricing and for batch or offline processing applications where latency is not critical. AMD EPYC CPUs can also enable consolidation of IT infrastructure, significantly reducing the number of servers deployed in enterprise data centers and making room for next-generation AI-based infrastructure.

As AI model size, complexity, and volumes increase, GPU clusters can be added to the data center to deliver more performance. GPUs and CPUs can be integrated within the same server to handle different AI workloads most efficiently. AMD provides the end-to-end AI solution with AMD EPYC CPUs and AMD Instinct GPUs.



5TH GENERATION AMD EPYC™ 9005 PROCESSORS ARE A FOUNDATION FOR PUBLIC SECTOR AI INFERENCE

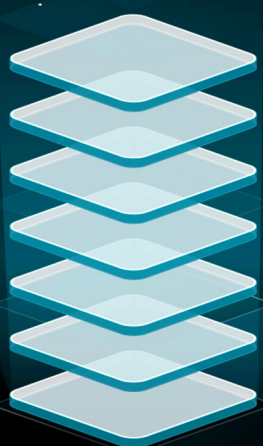
AMD EPYC 9005 processors provide the ideal CPU-based platform to power highly effective AI-driven public sector solutions and deliver operational efficiencies with advanced workload performance and power efficiency, while enabling significantly lower TCO. AMD EPYC 9005 processor-powered solutions drive the innovation that agencies need to power game-changing public sector solutions that use data and AI to improve public services, optimize resource utilization, and enable better outcomes for citizens.

5th generation AMD EPYC 9005 processors offer high core counts, enabling faster processing by distributing AI workloads across multiple cores. High-speed I/O and memory bandwidth address the specific needs of public sector workloads, such as creating large-scale, AI-driven simulations for public health and safety planning. Edge computing is particularly relevant to public sector use cases, as it can support a wide range of business applications, including tax management and citizen analytics, and managing data processing for smart city and environmental monitoring applications.

AMD EPYC 9005 processors and AMD Instinct provide end-to-end AI performance by:

- **Maximizing Per-Server Performance.** AMD EPYC 9005 can match integer performance of legacy hardware with up to 86% fewer servers[3]—dramatically reducing physical footprint, power consumption - freeing up space for new or expanded AI workloads while helping reduce TCO.
- **Delivering Leadership AI Inference Performance.** Many AI workloads—language models with 13 billion parameters and below, image analysis, or recommendation systems—run efficiently on CPU-only servers that feature AMD EPYC 9005 CPUs. Servers running two 5th gen AMD EPYC 9965 CPUs offer up to 2x inference throughput when compared to the previous generation.[4]
- **Maximizing GPU Acceleration.** The AMD EPYC™ 9005 family includes High Frequency EPYC 9005 processors which are excellent host CPUs for GPU-enabled systems to help increase performance on select AI workloads and can improve the ROI of each GPU server. For example, a high frequency AMD EPYC 9575F processor powered server with 8x GPUs delivers up to 20% greater system performance than a server with Intel Xeon 8592+ processors as the host CPU with the same 8x GPUs running Llama3.1-70B.[5]
- **Offering Security at Depth.** AMD EPYC processors include AMD Infinity Guard, a sophisticated suite of system-level security features built-in at the silicon level, with advanced capabilities to help defend against internal and external threats to keep data safe. Infinity Guard provides defense in depth, with security controls at multiple layers, including secure boot and encryption, to help prevent threats. Infinity Guard also seamlessly integrates with most established security, compliance, and operations tools.[6]

AMD INFINITY GUARD TECHNOLOGIES



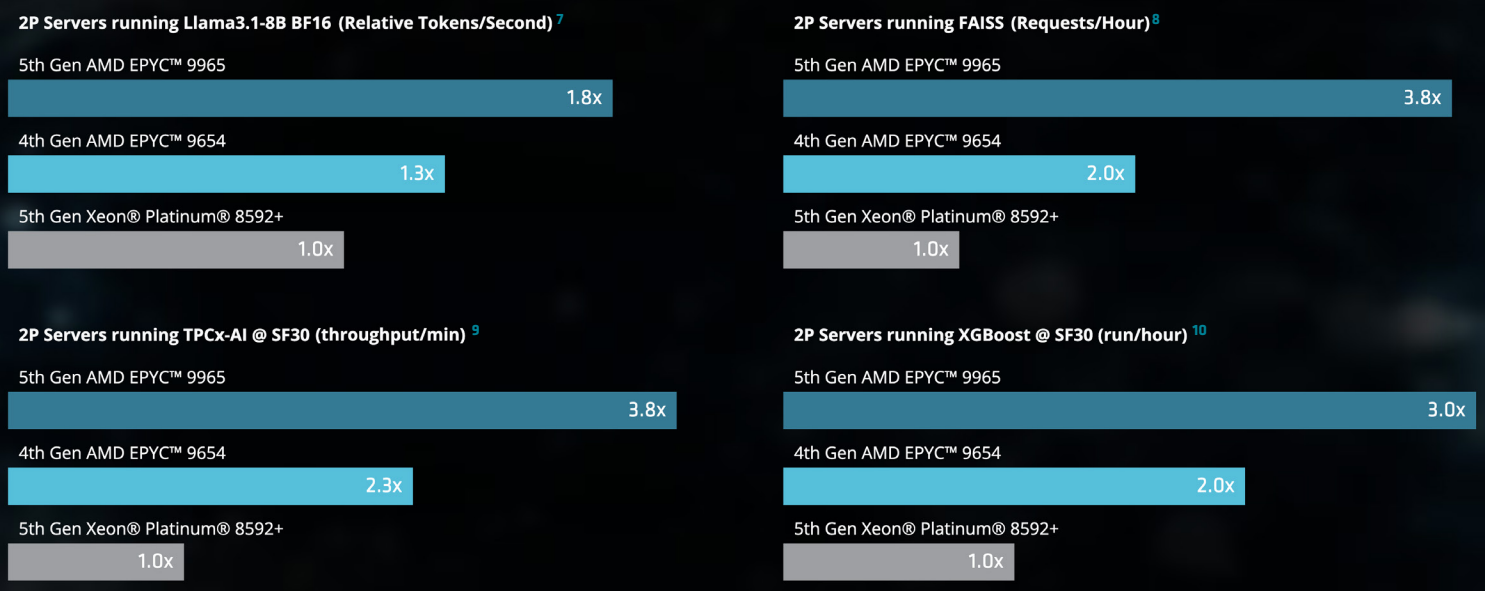
- AMD Secure Boot
- AMD Shadow Stack
- Secure Memory Encryption (SME)
- Secure Encrypted Virtualization (SEV
- Encrypted State (SEV-ES)
- Secure Nested Paging (SEV-SNP)
- Trusted IO (SEV-TIO)

HIGH PERFORMANCE WITHOUT LARGE BUDGET INCREASES

AMD EPYC™ processors-based servers deliver cost-optimized performance across a wide range of public sector workloads and help lay the groundwork for rapid efficiency gains. This can lead to high performance-to-power ratios and low TCO that help users optimize their data centers and physical space and get more performance from the same or less data center power, space, or budget.

AI Workload Models

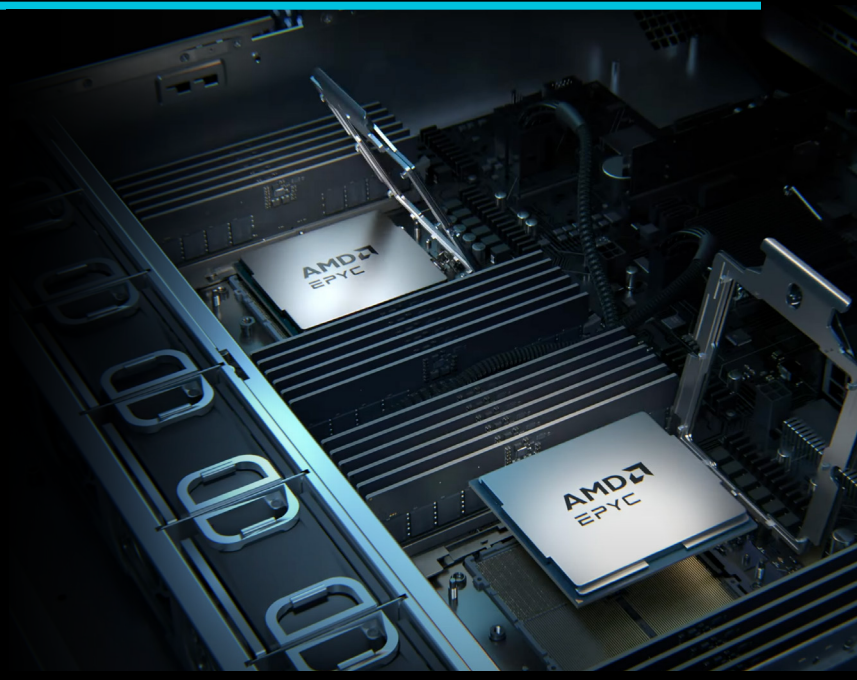
AMD EPYC™ 9005 processor-based servers enable fast, efficient, AI-enabled solutions close to your citizens and data.



LEVERAGE A STRONG ECOSYSTEM

The AMD ecosystem is compatible with major AI and machine learning frameworks. Models built on frameworks like PyTorch or TensorFlow will run on AMD hardware without requiring major code changes, and AI workloads that were previously optimized for competitors' CPUs or GPUs typically operate on AMD EPYC processors without extensive reconfiguration.

In addition, AMD works closely with its OEM partners to help ensure that hardware co-optimization occurs both at the CPU and platform level. These OEMs also address physical security for data center infrastructure for many of its platforms, with features like locking lids and antitamper capabilities to prevent and detect when an unauthorized person attempts to interfere with a server.



CUSTOMER SUCCESS STORIES: BRINGING AI TO PUBLIC SECTOR ORGANIZATIONS

VinAI Empowers Safety Cameras with AMD EPYC™ CPUs

Vietnam-based VinAI wanted to optimize the processing capabilities of the inference servers employed by its GuardPro product, which uses AI to deliver computer vision based on video inputs from IP cameras, used for neighborhood monitoring in smart city scenarios and surveillance for buildings. After testing a range of processors, VinAI deployed AMD EPYC processors for its inference servers, enabling VinAI to double the number of camera streams processed per server from 500 to 1,000, while cutting hardware costs by up to 35%.

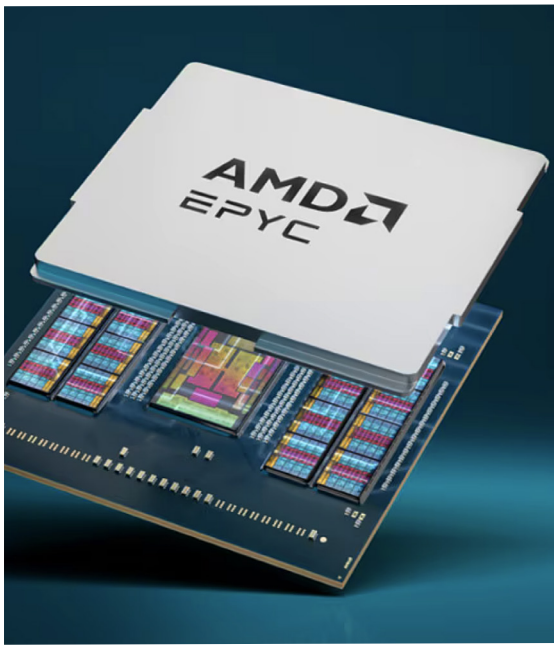
Building Large Language Models with the Power of AMD Instinct™ GPUs and AMD EPYC CPUs.

Researchers at the TurkuNLP Group at the University of Turku in Finland wanted to train a large language model (LLM) in Finnish and other European languages for research and academic purposes. However, most models are proprietary and based on English, and training an LLM requires high-performance computing with massive scalability to ensure sufficiently rapid iterations. The researchers deployed the LUMI supercomputer, powered by AMD EPYC CPUs and AMD Instinct GPUs, to teach its open language model Finnish. This scaled the LUMI infrastructure to 192 nodes, taking two weeks to run training on a 176 billion parameter model for 40 billion tokens and several smaller monolingual Finnish models for 300 billion tokens. These models built for Finnish will serve as the foundation for the next generation of Finnish artificial intelligence technologies, with the goal to create the largest open LLM with comprehensive support for European languages.

AMD EPYC 9005 PROCESSORS CAN DELIVER LEADERSHIP PERFORMANCE AND ENERGY EFFICIENCY WITH OUTSTANDING TCO FOR PUBLIC SECTOR.

AMD, YOUR PARTNER IN PUBLIC SECTOR TRANSFORMATION

Unlock the transformative potential of AI to address pressing public sector challenges and deliver better outcomes for your communities. Partner with AMD to deliver innovative AI solutions that empower agencies to innovate and transform with cutting-edge AI technologies that are scalable, secure, and cost-effective. Together, we can build tailored, future-ready systems that drive efficiency, enhance public services, and improve lives for citizens. Join us in shaping the future of AI in the public sector—let's work together to create impactful solutions that make a difference.



TO LEARN MORE:

- Visit **AMD Public Sector Solutions**
- Contact your AMD representative for more information
- Email the AMD Retail and E-Commerce team with questions

For more information, **visit www.amd.com**

FOOTNOTES

1 Top 500, The List, June 2025

2 9xx5TCO-002A: 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. 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 Intel Xeon 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 electricity factors from the 2024 International Country Specific Electricity Factors 10 – July 2024, and the United States Environmental Protection Agency 'Greenhouse Gas Equivalencies Calculator'. For additional details, see https://www.amd.com/en/claims/epyc4#SP9xxTCO-002A.

3 9xx5TCO-001C: 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. 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 39100 units of SPECrate2017_int_base performance as of October 10, 2024. This scenario compares a legacy 2P Intel Xeon 28 core Platinum_8280 based server with a score of 391 versus 2P EPYC 9965 (192C) powered server with an 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 electricity factors from the 2024 International Country Specific Electricity Factors 10 – July 2024 , and the United States Environmental Protection Agency 'Greenhouse Gas Equivalencies Calculator'. For additional details, see https://www.amd.com/en/claims/epyc5#9xx5TCO-001C

4 9xx5-040A: XGBoost (Runs/Hour) throughput results based on AMD internal testing as of 09/05/2024. XGBoost Configurations: v2.2.1, Higgs Data Set, 32 Core Instances, FP32 2P AMD EPYC 9965 (384 Total Cores), 12 x 32 core instances, 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.4 LTS, 6.8.0-45-generic (tuned-adm profile throughput-performance, ulimit -l 198078840, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C (SMT=off, Determinism=Power, Turbo Boost=Enabled), NPS=1 2P AMD EPYC 9755 (256 Total Cores), 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198094956, ulimit -n 1024, ulimit -s 8192), BIOS RVOT0090F (SMT=off, Determinism=Power, Turbo Boost=Enabled), NPS=1 2P AMD EPYC 9654 (192 Total cores), 1.5TB 24x64GB DDR5-4800, 1DPC, 2 x 1.92 TB Samsung MZQL21T9HCJR-00A07 NVMe®, Ubuntu 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198120988, ulimit -n 1024, ulimit -s 8192), BIOS TTI100BA (SMT=off, Determinism=Power), NPS=1 Versus 2P Xeon Platinum 8592+ (128 Total Cores), AMX On, 1TB 16x64GB DDR5-5600, 1DPC, 1.0 Gbps NetXtreme BCM5719 Gigabit Ethernet PCIe, 3.84 TB KIOXIA KCMYXRUG3T84 NVMe®, Ubuntu 22.04.4 LTS, 6.5.0-35 generic (tuned-adm profile throughput-performance, ulimit -l 132065548, ulimit -n 1024, ulimit -s 8192), BIOS ESE122V (SMT=off, Determinism=Power, Turbo Boost = Enabled) Results: CPU Run 1 Run 2 Run 3 Median Relative Throughput Generational 2P Turin 192C, NPS1 1565.217 1537.367 1553.957 1553.957 3 2.41 2P Turin 128C, NPS1 1103.448 1138.34 1111.969 1111.969 2.147 1.725 2P Genoa 96C, NPS1 662.577 644.776 640.95 644.776 1.245 1 2P EMR 64C 517.986 421.053 553.846 517.986 1 NA Results may vary due to factors including system configurations, software versions and BIOS settings.

5 9xx5-014A: 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 Total Cores) 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 22.04.3 LTS, 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=0 mitigations=off, cpupower frequency-set -g performance), BIOS 2.1, (Maximum performance, SR-IOV=On),
I/O Tokens Batch Size EMR Turin Relative 128/128 1024 814.678 1101.966 1.353 128/2048 1024 2120.664 2331.776 1.1 2048/128 96 114.954 146.187 1.272 2048/2048 64 333.325 354.208 1.063
For average throughput increase of 1.197x. Results may vary due to factors including system configurations, software versions and BIOS settings.

6 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.

FOOTNOTES (CONTINUED)

7 9xx5-009: Llama3.1-8B throughput results based on AMD internal testing as of 09/05/2024. Llama3-8B configurations: IPEX.LLM 2.4.0, NPS=2, BF16, batch size 4, Use Case Input/Output token configurations: [Summary = 1024/128, Chatbot = 128/128, Translate = 1024/1024, Essay = 128/1024, Caption = 16/16]. 2P AMD EPYC 9965 (384 Total Cores), 6 64C instances 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1 DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.3 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198096812, ulimit -n 1024, ulimit -s 8192) , BIOS RVOT1000C, (SMT=off, Determinism=Power, Turbo Boost=Enabled), NPS=22P AMD EPYC 9755 (256 Total Cores), 4 64C instances , 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu 22.04.3 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C (SMT=off, Determinism=Power, Turbo Boost=Enabled), NPS=2 2P AMD EPYC 9654 (192 Total Cores) 4 48C instances , 1.5TB 24x64GB DDR5-4800, 1DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.4 LTS, 5.15.85-051585-generic (tuned-adm profile throughput-performance, ulimit -l 1198117616, ulimit -n 500000, ulimit -s 8192), BIOS RVI1008C (SMT=off, Determinism=Power, Turbo Boost=Enabled), NPS=2Versus 2P Xeon Platinum 8592+ (128 Total Cores), 2 64C instances , AMX On, 1TB 16x64GB DDR5-5600, 1DPC, 1.0 Gbps NetXtreme BCM5719 Gigabit Ethernet PCIe, 3.84 TB KIOXIA KCMYXRUG3T84 NVMe®, Ubuntu 22.04.4 LTS 6.5.0-35-generic (tuned-adm profile throughput-performance, ulimit -l 132065548, ulimit -n 1024, ulimit -s 8192), BIOS ESE122V (SMT=off, Determinism=Power, Turbo Boost = Enabled). Results: CPU 2P EMR 64c 2P Turin 192c 2P Turin 128c 2P Genoa 96c Average Aggregate Median Total Throughput 99.474 193.267 182.595 138.978 Competitive 1 1.943 1.836 1.397 Generational NA 1.391 1.314 1 Results may vary due to factors including system configurations, software versions and BIOS settings.

8 9xx5-011: FAISS (Requests/Hour) throughput results based on AMD internal testing as of 09/05/2024. FAISS Configurations: sift1m Data Set, 16 Core Instances, FP32, MKL 2024.2.1 2P AMD EPYC 9965 (384 Total Cores), 24 16C instances, 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C (SMT=off, Determinism=Power, Turbo Boost=Enabled), NPS=42P AMD EPYC 9654 (192 Total cores) 12 16C instances, 1.5TB 24x64GB DDR5-4800, 1DPC, 2 x 1.92 TB Samsung MZQL21T9HCJR-00A07 NVMe, Ubuntu 22.04.3 LTS, BIOS 1006C (SMT=off, Determinism=Power), NPS=4Versus 2P Xeon Platinum 8592+ (128 Total Cores), 8 16C instances, AMX On, 1TB 16x64GB DDR5-5600, 1DPC, 1.0 Gbps NetXtreme BCM5719 Gigabit Ethernet PCIe, 3.84 TB KIOXIA KCMYXRUG3T84 NVMe, , Ubuntu 22.04.4 LTS, 6.5.0-35 generic (tuned-adm profile throughput-performance, ulimit -l 132065548, ulimit -n 1024, ulimit -s 8192), BIOS ESE122V (SMT=off, Determinism=Power, Turbo Boost = Enabled) Results:CPU Median Relative Throughput Generational 2P Turin 192C 64.2 3.776 1.861 2P Genoa 96C 34.5 2.029 1 2P EMR 64C 17 1 NA Results may vary due to factors including system configurations, software versions and BIOS settings.

9 9xx5-012: TPCxAI @SF30 Multi-Instance 32C Instance Size throughput results based on AMD internal testing as of 09/05/2024 running multiple VM instances. The aggregate end-to-end AI throughput test is derived from the TPCx-AI benchmark and as such is not comparable to published TPCx-AI results, as the end-to-end AI throughput test results do not comply with the TPCx-AI Specification.2P AMD EPYC 9965 (384 Total Cores), 12 32C instances, NPS1, 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C (SMT=off, Determinism=Power, Turbo Boost=Enabled)2P AMD EPYC 9755 (256 Total Cores), 8 32C instances, NPS1, 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT0090F (SMT=off, Determinism=Power, Turbo Boost=Enabled)2P AMD EPYC 9654 (192 Total cores) 6 32C instances, NPS1, 1.5TB 24x64GB DDR5-4800, 1DPC, 2 x 1.92 TB Samsung MZQL21T9HCJR-00A07 NVMe, Ubuntu 22.04.3 LTS, BIOS 1006C (SMT=off, Determinism=Power)Versus 2P Xeon Platinum 8592+ (128 Total Cores), 4 32C instances, AMX On, 1TB 16x64GB DDR5-5600, 1DPC, 1.0 Gbps NetXtreme BCM5719 Gigabit Ethernet PCIe, 3.84 TB KIOXIA KCMYXRUG3T84 NVMe, , Ubuntu 22.04.4 LTS, 6.5.0-35 generic (tuned-adm profile throughput-performance, ulimit -l 132065548, ulimit -n 1024, ulimit -s 8192), BIOS ESE122V (SMT=off, Determinism=Power, Turbo Boost = Enabled) Results: CPU Median Relative Generational Turin 192C, 12 Inst 6067.531 3.775 2.278 Turin 128C, 8 Inst 4091.85 2.546 1.536 Genoa 96C, 6 Inst 2663.14 1.657 1 EMR 64C, 4 Inst 1607.417 1 NA Results may vary due to factors including system configurations, software versions and BIOS settings. TPC, TPC Benchmark and TPC-C are trademarks of the Transaction Processing Performance Council.

10 9xx5-040A: XGBoost (Runs/Hour) throughput results based on AMD internal testing as of 09/05/2024. XGBoost Configurations: v2.2.1, Higgs Data Set, 32 Core Instances, FP32 2P AMD EPYC 9965 (384 Total Cores), 12 x 32 core instances, 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.4 LTS, 6.8.0-45-generic (tuned-adm profile throughput-performance, ulimit -l 198078840, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C (SMT=off, Determinism=Power, Turbo Boost=Enabled), NPS=1 2P AMD EPYC 9755 (256 Total Cores), 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198094956, ulimit -n 1024, ulimit -s 8192), BIOS RVOT0090F (SMT=off, Determinism=Power, Turbo Boost=Enabled), NPS=1 2P AMD EPYC 9654 (192 Total cores), 1.5TB 24x64GB DDR5-4800, 1DPC, 2 x 1.92 TB Samsung MZQL21T9HCJR-00A07 NVMe®, Ubuntu 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198120988, ulimit -n 1024, ulimit -s 8192), BIOS TTI100BA (SMT=off, Determinism=Power), NPS=1 Versus 2P Xeon Platinum 8592+ (128 Total Cores), AMX On, 1TB 16x64GB DDR5-5600, 1DPC, 1.0 Gbps NetXtreme BCM5719 Gigabit Ethernet PCIe, 3.84 TB KIOXIA KCMYXRUG3T84 NVMe®, Ubuntu 22.04.4 LTS, 6.5.0-35 generic (tuned-adm profile throughput-performance, ulimit -l 132065548, ulimit -n 1024, ulimit -s 8192), BIOS ESE122V (SMT=off, Determinism=Power, Turbo Boost = Enabled) Results: CPU Run 1 Run 2 Run 3 Median Relative Throughput Generational 2P Turin 192C, NPS1 1565.217 1537.367 1553.957 1553.957 3 2.41 2P Turin 128C, NPS1 1103.448 1138.34 1111.969 1111.969 2.147 1.725 2P Genoa 96C, NPS1 662.577 644.776 640.95 644.776 1.245 1 2P EMR 64C 517.986 421.053 553.846 517.986 1 NA Results may vary due to factors including system configurations, software versions and BIOS settings.

DISCLAIMERS

COPYRIGHT NOTICE