

# The Future of AI Inferencing Lies in Open, Flexible Infrastructure Stacks

**Written by:**

**Peter Rutten**

Research Vice President, AI, HPC and Quantum Computing, Core Infrastructure

**Madhumitha Sathish**

Research Manager, AI and HPC, Core Infrastructure

**Published:**

**July 2026**

## IDC opinion

The AI landscape is at a pivotal inflection point, with infrastructure for AI inferencing set to surpass that for AI training. Most organizations are not developing their own AI models but instead deploying prebuilt open-source or commercial models, choosing an infrastructure stack critical to achieving positive ROI. IDC data highlights that involving IT teams early and selecting the right cost/performance infrastructure stack are essential for efficient, ROI-focused AI inferencing. The most consequential decision is whether to adopt an open or proprietary stack.

Historically, proprietary systems have dominated early technology curves but are eventually disrupted by open ecosystems that foster innovation, reduce costs, and enhance flexibility. IDC expects this pattern to repeat in AI inferencing (as well as training) because enterprise datacenters increasingly favor openness and efficiency over vendor lock-in. Open ecosystems enable businesses to quickly adopt new technologies, access broader skill sets, and benefit from community-driven innovation.

The AI inference landscape is defined by four integrated vertical stacks (represented by the ecosystems built around AMD, Intel, NVIDIA, and the UXL Foundation) that translate raw hardware power into AI-infused enterprise applications. These stacks include hardware, orchestration, software stacks (drivers, tools, compilers, and runtimes), AI enterprise suites or solutions, AI models, and enterprise applications, agents, and ISV solutions. One of these four approaches is mostly proprietary, while the other three represent an open strategy. Open platforms support hardware diversity and integration, enabling AI to be embedded in general-purpose workloads rather than siloed in specialized infrastructure.

While AI inferencing factories (dedicated, high-capacity environments) are suitable for extreme-scale workloads, most organizations require AI integration into their existing environments for applications such as email generation, chatbots, or predictive maintenance. This shift toward distributed real-time AI serving underscores the importance of open ecosystems, which facilitate seamless integration across CPUs and GPUs.

IDC believes the future of AI inferencing lies in open, flexible infrastructure stacks that support broad integration and innovation. Organizations should prioritize openness to maximize ROI, efficiency, and adaptability as AI becomes an integral part of enterprise operations.

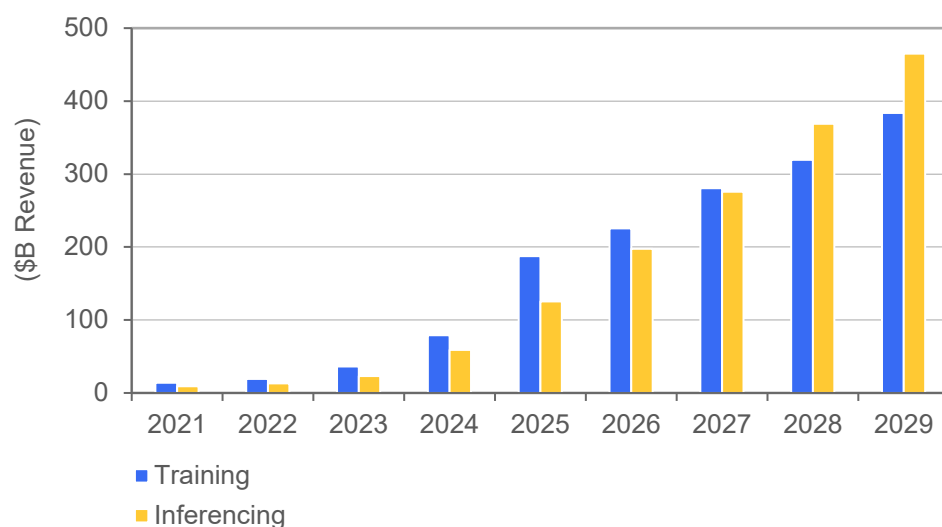
## Situation Overview

### The AI inflection point has arrived

IDC projects that next year, for the first time, annual worldwide investments in infrastructure for AI inferencing will exceed those for AI training. This shift has implications for the nature of the stack built on top of that infrastructure.

**Figure 1**

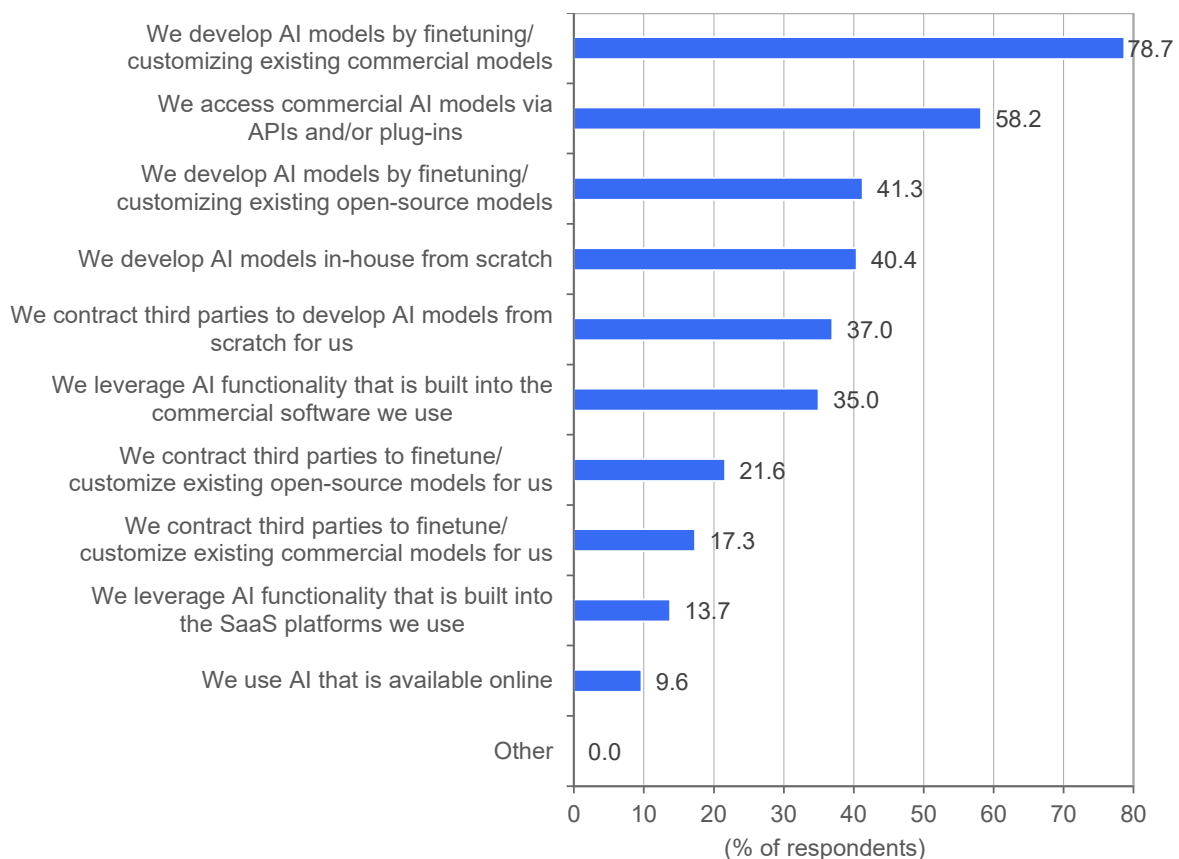
#### Worldwide spending on AI infrastructure: Training versus inferencing (\$M)



Source: IDC's HPC-AI Index, February 2026

For enterprises, this trend reflects an even more pronounced shift. IDC research has found that most enterprises (60%) do not develop their own AI models (Figure 2), but rather rely on commercially available or open-source models that they customize or fine-tune.

**Figure 2**  
**AI model acquisition approaches**



Source: IDC's AI View Survey, 2026 (n = 716)

Businesses that adopt AI functionality with predeveloped open-source or commercial models have largely avoided investments in infrastructure stacks and skill sets for AI training, but they must make technology and skill-set decisions and investments for deploying the AI models. Furthermore, IDC data shows that infrastructure stack investments are the most consequential components in ROI calculations regarding AI initiatives. Worldwide, organizations estimate that 47% of their IT budgets is for AI hardware purchases (servers, storage, networking, excluding cloud), while the remaining 53% is for everything else: cloud, software, third-party services, and staff. Achieving ROI from AI model deployment, therefore, depends heavily on building the right AI inferencing infrastructure stack.

In summary, AI is shifting to the deployment of AI models rather than development, and the decisions organizations make about their infrastructure stacks for AI deployment can significantly influence their anticipated ROI.

## Developing an AI inferencing Strategy for a Positive ROI

Two particularly critical factors are involving the IT team from the very beginning (so it can weigh in on the infrastructure stacks required for deploying the model) and selecting the right infrastructure stack from a cost/performance perspective.

IT teams understand the factors that must be considered to achieve the desired ROI when determining the right AI inferencing infrastructure stack, such as user concurrency, model size, and capabilities (LLM versus VLM versus multimodal); context window size and anticipated daily queries on the model; anticipated average query size; required response times (time to first token, end-to-end latency); required integration with other applications; and for agentic AI, memory layer requirements, tool calling, and multi-turn.

But perhaps the most important factor will be the open versus closed nature of the stack, meaning a stack that is non-proprietary and comprises many contributing vendors and communities (while still providing centralized support), or one that is proprietary and consists primarily of a solution stack from a single vendor.

## Proprietary versus open and what that means for AI Inferencing: A bit of history

We have had this discussion before. Proprietary full-stack systems are very successful at the start of a technology curve (for example, mainframes in the second half of the past century monopolized business processing). They made it easy for enterprises: everything from a single vendor, tightly packaged and optimized, but highly controlled, leading to vendor lock-in.

Because of vendor lock-in, mainframes became increasingly expensive, and there are many other examples of essential proprietary systems for business processing that became increasingly more costly, from operating systems to virtualization layers. At the same time, innovation slowed and responsiveness to changing customer needs became more rigid.

Mainframes, as an example, increasingly faced disruption from more flexible, smaller Unix-based systems that marked the beginning of a more open ecosystem, until Unix vendors made the same mistake and started creating their own flavors of Unix, reverting to proprietary approaches. Then, Unix-based systems were in turn disrupted by truly open Linux-based systems, which now dominate the datacenter landscape. The lesson is that open systems always disrupt proprietary systems, which can't pivot as fast as technologies and business needs change. Indeed, open systems enable open markets that stimulate innovation and provide end users with options. IDC expects this to happen again with AI processing, both for AI inferencing and AI training, which traditionally was more siloed and proprietary

From the outset, AI inference must be fully integrated into enterprises' daily business processes in their datacenters, and enterprise datacenters gravitate toward openness and efficiency, not enterprise software optimized and licensed for only the vendor's own hardware, even if that might serve as a temporary "easy button."

Openness means businesses can quickly and flexibly adopt new technologies and technology stacks rather than having to wait for their single vendor to incorporate them. It

means they don't develop skill sets unique to one vendor's solutions, but can rely on a community from which a variety of evolving skill sets are easier to obtain. In general, there's sufficient evidence to claim that in the long term, open ecosystems benefit innovation, costs, skill sets, resiliency, and security.

## Overview of the current landscape for AI inferencing

To date, several stacks exist or are under development for AI inferencing, from AMD, Intel, NVIDIA, and UXL. Here's an alphabetical overview:

- **AMD ROCm.** This is probably the most advanced open-ecosystem parallel computing platform on the market. It includes AMD Heterogeneous Compute Interface for Portability (HIP), which is a C++ runtime API and kernel language that enables developers to create high-performance, portable applications for both AMD and other GPUs from a single source code. Popular frameworks such as PyTorch and TensorFlow officially support ROCm, and ROCm supports popular models such as vLLM and Hugging Face, as well as a robust set of optimized models in the AIMs catalog.
- **Intel OpenVINO.** Intel's open-source ecosystem is designed to tie together many processor types under a unified strategy. OpenVINO is specifically designed for inferencing optimization, which it does by automatically identifying the most suitable hardware to run a model on with the greatest efficiency. Its oneAPI serves as the core standards-based parallel computing programming model.
- **NVIDIA.** NVIDIA's AI inference stack is a mostly proprietary full-stack platform optimized for deploying high-performance AI models, combining specialized hardware (GPUs such as Blackwell and networking), low-level software (CUDA), and optimized frameworks (TensorRT-LLM, Dynamo, NIM).
- **The UXL Foundation.** The foundation is a broad industry initiative that includes Arm, Intel, Google, and Samsung, and it is developing an open-source parallel computing platform and C++-based programming model called SYCL. Based on oneAPI, the aim is to create a multi-architecture, multivendor software ecosystem that enables AI to run on any processor from any vendor.

## The two extremes on the AI inferencing scale: Integrated versus factories

A compelling reason for adopting an open AI ecosystem is that most general-purpose workloads (web serving, email, enterprise resource management, etc.) already run on open systems. If AI is to be integrated into those workloads, it shouldn't be based on proprietary software. Moreover, it will only make sense to deploy what is sometimes called an AI inferencing factory if the inferencing workload is extreme in scale.

An AI factory is generally defined as dedicated rack-scale infrastructure with very large, very dense GPU capacity, high-speed networking, and performance-intensive storage, used primarily for AI training. AI inferencing factories are similar, but they are designed to deploy and manage AI models in production at scale. They focus even more on advanced networking to keep latency as low as possible, which is important for real-time applications. AI inferencing factories can handle millions of simultaneous requests for stand-alone services such as chatbots, recommendation engines, generative AI services, and agentic AI.

However, most organizations do not need such inferencing scale and prefer not to have a siloed infrastructure stack for inferencing on AI models. They see AI as being integrated into all their other workloads (for example, email generation, document drafting, code generation, customer service chatbots, language translation, content creation, and predictive maintenance). Most organizations are expected to use such AI-powered applications within the next few years, and many report significant efficiency gains. IDC's AI View Survey (2026) found that only 21% consider AI inferencing on an AI model as a stand-alone capability that delivers insights to general-purpose (non-AI) workloads. The remainder (79%) require AI to be integrated into their general-purpose workloads, a move away from centralized, batch-processing models toward more distributed, real-time AI serving. Organizations also overwhelmingly (91%) say that such integration affects the technology on which their general-purpose workloads run. More distributed, real-time AI serving is achieved with what many vendors call AI-ready environments that combine accelerated computing (GPUs or NPUs) with existing CPU-intensive applications via microservices, API layers, and container orchestration (for example, wrapping AI models in containers and deploying them via microservices, then enabling existing apps to make REST calls to the service).

In this approach, inferencing is not just a GPU task. The role of CPUs is vital for functions such as preprocessing, retrieval, and filtering, which is another reason an open ecosystem will be advantageous, as it avoids running CPUs in a closed ecosystem. At the same time, the adjacent GPUs are operated with proprietary software.

## Considering AMD

AMD has an open ecosystem approach to AI inferencing that centers on breaking vendor lock-in by providing a flexible, high-performance, and open source–driven set of solutions. By leveraging the open-source ROCm software stack and collaborating on industry standards such as Ultra Accelerator Link (UALink), AMD enables efficient inference for large language models and agentic workflows on an open infrastructure stack across all deployment scenarios (datacenter, cloud, neocloud, and edge).

Major components of the AMD open ecosystem for AI inferencing include:

- Open software ecosystem (ROCm)
- Open-source software. The ROCm software stack is built as an open-source equivalent to proprietary parallel programming languages, avoiding proprietary silos.
- Portability and compatibility. Through HIP, developers can port code that runs on both AMD and competitor hardware, reducing migration friction.
- ROCm 7 enhancements. The ROCm 7 Platform is designed to provide improved inference capability (over 3.5 times on MI300X in preview), native support for low-precision data types (FP4/FP6), and distributed inference tools (vLLM, SGLang) for massive AI models.
- Open networking and interconnect standards
- UALink Consortium. AMD leads the Ultra Accelerator Link Consortium, which defines an open, high-performance interconnect standard for scale-up systems to avoid vendor-locked interconnect technologies.

- Ethernet support. Through the Ultra Ethernet Consortium, AMD supports open standards to enable AI clusters to use commodity Ethernet for scaling, providing greater flexibility and lower costs than proprietary alternatives.
- Open enterprise AI reference stack
- Turnkey inference services. The AMD enterprise AI reference stack provides prebuilt, optimized containers (AMD Inference Microservices) for fast, efficient deployment on AMD hardware, supporting popular open-source models.
- Standardized workflows. AMD Solution Blueprints offer validated, open-source reference architectures to help enterprises transition from experimental to production-grade AI in specialized areas such as finance and healthcare.
- AMD Developer Cloud. This service provides limited free, remote access to Instinct MI300 series GPUs, enabling developers to test and deploy on AMD hardware using open source tools.
- AMD AI Platforms. Smaller scalable designs with GPUs for AI deployments that are integrated into general-purpose workloads rather than in centralized, batch-processing models
- Partnerships and community adoption
- Major hyperscalers. Meta uses AMD GPUs for Llama 3/4 inference, while Microsoft Azure and Oracle Cloud run production-level inference for large models on AMD hardware.
- Hugging Face collaboration. AMD works closely with Hugging Face to ensure support and optimization for popular models, enabling them to run easily on AMD hardware.
- Edge and client. AMD is extending its open ecosystem to the edge (Versal AI Edge) and on device (Ryzen AI), enabling consistent, open source–driven inference from the datacenter to the PC.
- High-performance, open hardware architecture
- Memory-first strategy. AMD's Instinct MI300X, MI325X, MI350X, and MI355X focus on massive HBM memory capacity (e.g., 288GB HBM3E on MI355X) and bandwidth, enabling large models to fit into fewer GPUs and reducing inference latency.
- Chiplet design and Infinity Fabric. The use of chiplets and Infinity Fabric interconnects provides enhanced memory capacity and bandwidth, crucial for high-throughput, low-latency inference workloads.
- Helios Rackscale Solution. AMD is promoting open-standards rack architecture such as the Helios rackscale solution, which is based on the OCP open rack standards, and integrates AMD EPYC CPUs, Instinct GPUs, and Pensando NICs to provide a cohesive, high-performance platform for generative AI.

## Challenges/Opportunities

AMD is a major participant in the AI infrastructure market with two formidable competitors: on one side, the largest incumbent in CPUs (and a less powerful GPU offering); on the other, the largest incumbent in GPUs (and a fledgling CPU business). As such, AMD is unique in that it has strengths in both processor categories.

While the latter incumbent decidedly dominates the AI training market, the AI inferencing space, which will soon eclipse the training market, is becoming the true battlefield.

The inferencing market does not necessarily benefit from proprietary approaches. Still, the omnipresence of proprietary training systems could lead enterprises to think otherwise. IDC believes enterprises should carefully consider whether they want to run their AI inferencing workloads on proprietary stacks, as they stand to benefit from cost, efficiency, and integration in an open AI inferencing ecosystem. IDC also believes AMD has a comprehensive strategy and solution to deliver these benefits.

## Conclusion

AI inferencing is at a pivotal inflection point, with infrastructure investments set to surpass those for AI training. Most enterprises are deploying prebuilt open-source or commercial models, making the choice of infrastructure stack critical for achieving positive ROI. History has shown that open, flexible infrastructure stacks, supported by broad ecosystems and community-driven innovation, will disrupt proprietary systems, offering greater adaptability, cost efficiency, and integration. The current AI inferencing landscape features four major vertical stacks with three of them favoring open strategies.

AMD's open ecosystem, driven by ROCm, UALink, and partnerships with hyperscalers and Hugging Face, exemplifies this approach, providing high-performance, scalable, and vendor-agnostic solutions. IDC recommends organizations prioritize openness in their AI inferencing strategies to maximize ROI, efficiency, and innovation, as AI becomes integral to enterprise operations.

### *About the IDC analyst(s)*



#### **Peter Rutten**

Research Vice President, AI, HPC and Quantum Computing, Core Infrastructure

Peter Rutten is Research Vice President within IDC's enterprise infrastructure global research domain and part of core infrastructure subdomain. In his capacity as IDC's global research lead for the performance-intensive computing (PIC) practice, Peter leads a team of analysts that deliver quantitative and qualitative insights on High-Performance Computing (HPC), Artificial Intelligence (AI), Big Data and Analytics (BDA) and Quantum Computing (QC) markets and vendors.



**Madhumitha Sathish**

Research Manager, AI and HPC, Core Infrastructure

Madhumitha Sathish is Research Manager within IDC's enterprise infrastructure global research domain and part of core infrastructure subdomain. As part of IDC's coverage of performance intensive computing (PIC), Madhumitha delivers insights on artificial intelligence (AI) and high-performance computing (HPC) infrastructure. In her role, she delivers syndicated and custom qualitative and quantitative insights, including market sizes, vendors shares and forecasts, and custom market models.

---

## Message from the sponsor

AI inferencing is rapidly becoming the dominant AI workload, making infrastructure decisions critical to long-term ROI. As enterprises scale AI deployments, open ecosystems are becoming the preferred approach due to the flexibility, cost efficiency, and vendor independence they enable. History has consistently favored open solutions, and we expect that AI will be no exception. By adopting rightsized, scalable AI infrastructure built on open and interoperable platforms, organizations can seamlessly integrate AI into existing operations while maintaining the agility to evolve with business needs – and put you in control of your AI future. Learn more in the resources below:

- <https://www.amd.com/en/solutions/ai.html>
- <https://www.amd.com/en/solutions/ai/ai-ready-enterprise.html>



IDC Custom Solutions produced this publication. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis that IDC independently conducted and published, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies.

This IDC material is licensed for [external use](#), and in no way does the use or publication of IDC research indicate IDC's endorsement of the sponsor's or licensee's products or strategies.



IDC Research, Inc.

One Beacon Street, Suite 33100, Boston, MA 02108, USA

T +1 508 872 8200

[blogs.idc.com](https://blogs.idc.com) | LinkedIn [@IDC](#) | [www.idc.com](https://www.idc.com)

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries. IDC's analysis and insight help IT professionals, business executives, and the investment community make fact-based technology decisions and achieve their key business objectives.

©2026 IDC. Reproduction is forbidden unless authorized. All rights reserved. [CCPA](#)