

White Paper

Enabling Advanced AI Computing with IBM Cloud Powered by AMD

Tom Sorensen, and Bob Sorensen
January 2026

Sponsored by IBM and AMD

HYPERION RESEARCH OPINION

The accelerating complexity of artificial intelligence (AI) workloads and high-performance computing (HPC) is reshaping infrastructure requirements across industries. Large language models, multimodal AI, agentic AI, and advanced scientific simulations demand unprecedented compute power, memory bandwidth, and energy. These requirements extend beyond raw performance to include orchestration flexibility, cost optimization, and compliance readiness. Organizations have nearly unanimously adopted AI integration for their traditional workloads, and a recent Hyperion Research study indicates several key elements of continued expansion:

- 74.9% of respondents indicated plans to moderately or significantly expand AI infrastructure to also support HPC workloads.
- Roughly 28% of those characterize this expansion as significant.
- Less than 3% expect to contract their use of AI, none of which would characterize that contraction as significant.

In today's rapidly evolving AI landscape, organizations face a large array of choices when selecting infrastructure for their workloads. Decisions span on-premises deployments and cloud-based solutions, each offering distinct advantages. Likewise, accelerated computing options—particularly GPUs—present multiple paths to performance optimization. Strategic partnerships, such as IBM Cloud with AMD, seek to deliver integrated solutions that combine advanced technical capabilities with flexibility, enabling users to achieve the right balance of performance, functionality, service, and cost efficiency through hybrid solutions while maintaining a forward-looking, agile roadmap.

CURRENT LANDSCAPE

AI is being explored nearly unanimously among user groups across industries and application spaces, even among those users who previously did not adopt AI infrastructure.

While many trends continue to shape the landscape of AI use among large user groups, research suggests three major influences on current circumstances:

Increased investment and allocation: Organizations prioritize hardware optimized for AI training and inference, with growing interest in GPUs supporting lower-precision compute (FP8, BF16). Due to factors such as cost confinement, limited availability of hardware, and continued experimentation, only 16% of users indicate the sole use of on-premises hardware to meet their inferencing compute needs (Fig. 1), highlighting the value of diverse and available cloud resources.

FIGURE 1

Hybrid Characteristics of Compute Resources to Meet Inferencing Needs

Method of Supporting Existing/Planned AI Inferencing for Compute Needs

Method	% Selected
Mostly on-premises hardware, but also supported by cloud resources	35.0%
An even mix of on-premises and cloud resources	20.0%
Mostly cloud resources, but also supported by on-premises hardware	19.0%
Entirely on-premises hardware	16.0%
Entirely cloud resources	8.0%
Don't know/Not sure	2.0%

Source: Hyperion Research, 2025

Growing Environment Complexity: Hybrid cloud, on-prem environments as well as containerization have become a near-unanimous staple of large-scale advanced computing and AI organizations. Hybrid arrangements such as this bring agility in hardware availability and support from the cloud while maintaining the cost effectiveness and reliability of on-premises. However, with this new complexity brings demands such as unique software tools to manage cross-source compute and data management, expanded expertise to use these unique tools, and a more scrutinous attention paid to cost-effectiveness to avoid runaway spending.

Establishment of AI Best Practices: The majority of advanced computing/AI users indicate continued experimentation and exploration of their AI technology and the diverse market offerings to support it.

With the leveraging of AI still a relative newcomer to the continuum of advanced computing, especially among those legacy HPC organizations such as national laboratories, it will take time for standard practices to be truly established. While users take strides to more aptly parameterize and scope their application compute needs, resources like those offered by cloud providers continue to shine.

Users tend to provision their AI inference needs with high-performance GPUs with mainstream GPUs and hyperscalers not far behind in popularity (Fig. 2). This preference to high-performance GPUs is expected to rise as AI adoption and scaling mounts.

FIGURE 2

Hybrid Characteristics of Compute Resources to Meet Inferencing Needs

Primary GPU or AI-Centric Accelerator Used for HPC/AI Inference Needs	
Primary GPU or Accelerator	% Selected
High-performance GPUs (e.g. Blackwell, MI300X, H200)	37.0%
Mainstream GPUs (H100, MI100)	27.2%
Hyperscale AI-centric accelerators (e.g. AWS, Google, Microsoft)	17.4%
AI-centric accelerators from commercial suppliers (e.g. Cerebras, Graphcore, SambaNova)	7.6%
Don't know/Not sure	6.5%
Other	4.4%

Source: Hyperion Research, 2025

END USER PERSPECTIVES ON AI REQUIREMENTS

Recent Hyperion Research studies highlight key user concerns and requirements shaping AI infrastructure decisions:

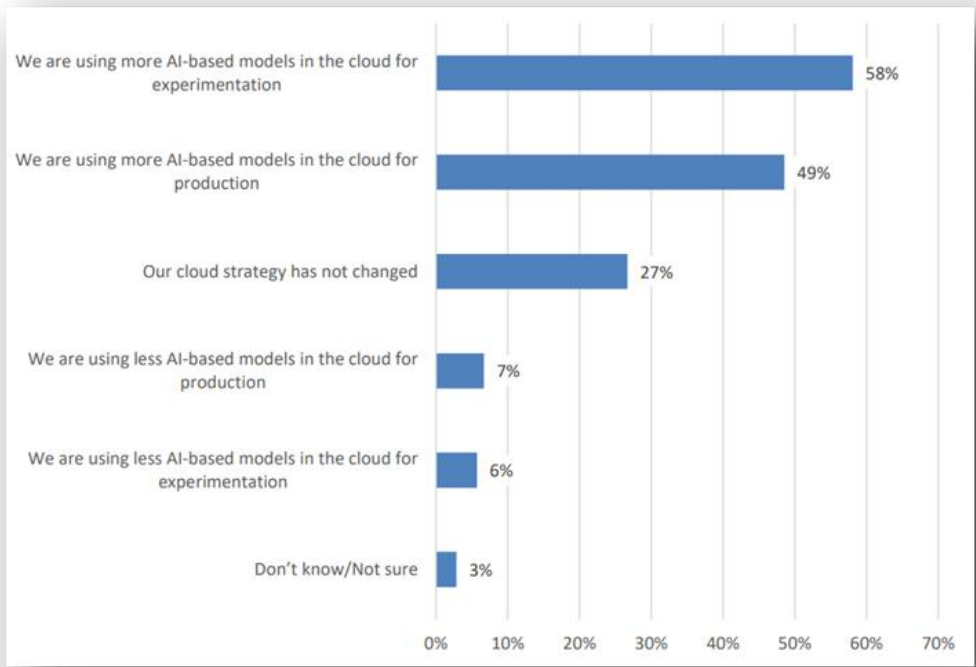
Scale and Complexity: Training trillion-parameter models requires massive parallelism and memory capacity. Furthermore, continued fine-tuning and retraining while inference capabilities are being scaled presents unique challenges. With rapidly evolving software products on the market and growing questions about the future of the availability of new data, training and running a fully scaled advanced AI model in-house can prove to be challenging, costly, and, at times, inefficient.

Hybrid Deployment: Enterprises increasingly operate across on-prem and cloud environments, introducing orchestration and compliance complexity. Managing data locality and provenance, cross-platform software tools, and achieving continuous efficiency and cost-effectiveness requires detailed

and rigorous scheduling and management. Users engaged in experimental or exploratory efforts most often report engaging in these projects in a cloud environment, with production as well not far behind, likely due to its ease-of-access and low commitment. Recent studies indicate a high level of migration among users adopting AI (Fig.3).

FIGURE 3

Advanced Computing Users Respond to AI Needs with Cloud Resources



Source: Hyperion Research, 2025

Affordability and Availability: The hardware and tools required to support leading-edge AI computing are expensive and, at times, wholly unavailable to some users as on-prem resources. Whether it be due to cost constraints, increased pace and expense of chip development, or shifts in trade laws, affordability and availability of required resources to populate data centers are a noted concern of users.

Energy and Sustainability: Rising power and cooling demands make efficiency a procurement priority. Accessing required power, water, space, and other infrastructure is a growing potential obstacle for user organizations, especially those hoping to continue scaling up.

Security and Compliance: Regulated industries require zero-trust architectures, agile responses to new regulations, and resilient security. AI technology and its applications are experiencing a continuous and, at times, confusing barrage of regulatory policies that can prove hard to navigate.

While challenges persist among AI users, there remains a tremendous level of trust in the technology to enhance and reshape current workloads. According to a recent Hyperion Research study, users most often anticipate a return on their AI investments within 3 years while continuing to explore their options and experiment with new model types, hardware configurations, and software tools. These hopes are well founded, with users reporting that their recently integrated AI technology is meeting or exceeding expectations, especially within certain application spaces.

IBM CLOUD WITH AMD SOLUTIONS

[IBM Cloud](#) integrates AMD hardware offerings from their latest generation MI Instinct™ series GPUs to deliver high memory bandwidth for optimized large model inference and fine-tuning, supported by unified memory designed to eliminate bottlenecks between CPU and GPU with optimized support for a diversity of AI data types and sparsity. This hardware, combined with IBM's secure, compliance-ready cloud, enable performance at scale for AI and HPC workloads.

This hardware family, purpose-built to deliver AI and HPC at scale, introduces a lower cost of ownership on the back of efficiency in scaling infrastructure and high memory capacity enabling running larger or a greater number of models on fewer GPUs. Additionally, with a commitment to open standards, users are given freedom from vendor lock-in and flexibility to scale allowing for future growth and agility.

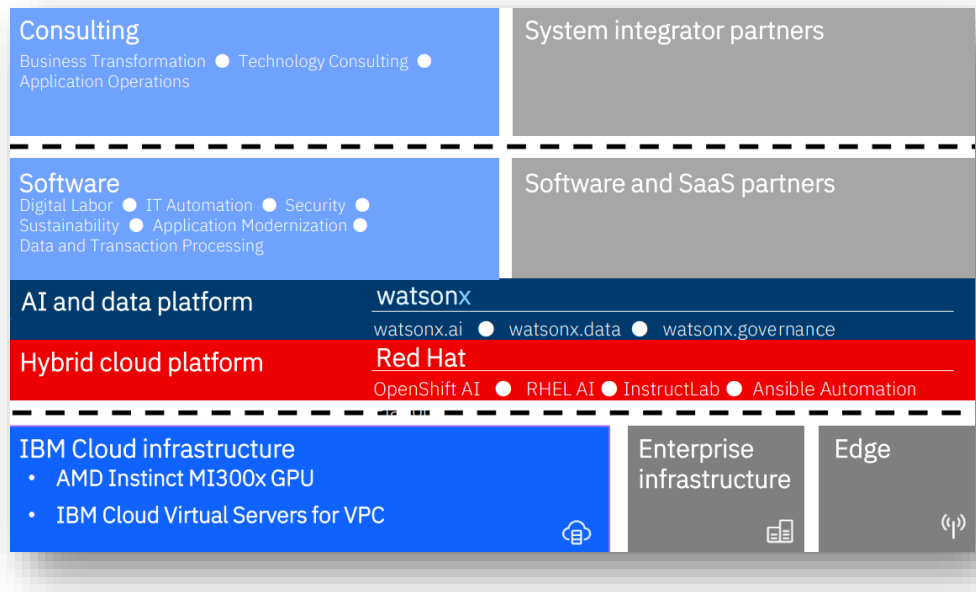
Key Features:

- **Performance & Scalability:** The AMD MI Instinct series supports massive models without GPU fragmentation, reducing training time and cost. Its architecture is optimized for generative AI and HPC workloads requiring large memory capacity. It can manage the largest models with coherent memory sharing across 8 GPUs on a single universal baseboard, and supports FP4, FP6, FP9, and FP16 data types.
- **Hybrid Flexibility:** Red Hat OpenShift on IBM Cloud delivers container portability, both within a cloud environment and on-prem, and Kubernetes orchestration, while OpenShift AI provides the toolbox for building and deploying custom AI solutions. Combined with watsonx.ai as-a-Service on IBM Cloud, enterprises can scale AI workflows seamlessly across on-prem and cloud environments.
- **Security & Compliance:** IBM Cloud offers built-in controls for regulated industries, zero-trust identity management, and post-quantum cryptography readiness—critical for enterprises operating under stringent compliance mandates. The by-default integration methods allow the process to be managed without exposing high-value workloads at runtime across the network or on disk.
- **Quantum Outlook:** IBM and AMD's partnership on quantum-centric supercomputing signals a future where quantum and classical HPC converge for unprecedented problem-solving capabilities.
- **AI Pipeline Coverage:** Applicable across the AI lifecycle—from training and fine-tuning to inference—while supporting traditional HPC applications.

IBM Cloud services offer a full stack approach to advanced computing support with a research-informed approach considerate of common user challenges, behavior patterns, and value creation. In addition to the peripheral support, these cloud services contain a comprehensive Gen AI stack including AI assistants and agent AI, SDKs and APIs for embedding, data platforms and services, and infrastructure tools. The following graphic outlines some of the key user resources offered within the IBM environment:

FIGURE 4

IBM Cloud User Resources



Source: IBM, 2025

AMD hardware on IBM Cloud leverages a suite of cross-industry tools to establish a holistic, research-informed ecosystem to meet users' most pressing needs. These tools, from start to finish, empower modern enterprises with secure, scalable AI. These solutions, which include not only IBM and AMD solutions, integrate open-source solutions with Red Hat OpenShift and Ansible Automation Platform as well as opportunities to modernize with IBM Consulting and Business Partners.

IBM Cloud has identified certain common patterns for AI usage based on user desires and introduced solutions. These needs, their solutions, and some sample use cases are outlined in the figure below.

FIGURE 5

IBM Clouds Needs Meet Solutions

I want	Solution	Sample use cases
Out of the box generative AI solutions	Agents and AI assistants	Customer assist, agentic workflow, document processing
To build customization, have model alignment, and scale AI models	Inferencing and Model-as-a-Service	Financial knowledgebase, policy/governance automation, customer satisfaction analysis
To build my own AI stack, just need AI-ready infrastructure	Build Your Own software stack Deploy on resilient, secure infrastructure	LLM hosting, HPC
Drive scale	Surround cloud services Security tools, observability, application platforms	Logging, monitoring, security and compliance, ISVs, cloud native

Source: IBM, 2025

With user cost expectations often being exceeded and a high expectation of return on AI investments, IBM Cloud partnered with AMD is uniquely positioned to provide users and organizations with the solutions to meet challenges head on and realize their full potential of efficiency, reliability, and return.

AI TRAINING AND SCALE: ZYPHRA CASE STUDY

Zyphra, an open-source AI research and product company based in San Francisco, is building human-aligned superintelligence that empowers individuals and companies to reach their fullest potential. One of the world’s largest open-science superintelligence labs, Zyphra set out to accelerate innovation in multimodal foundation models – spanning language, vision, audio, and brain computer interface (BCI) audio – while advancing breakthroughs in neural network architectures, long-term memory, and continual learning.

Challenges:

To achieve their goals, Zyphra required:

- Massive compute capacity for training cutting-edge generative AI models
- Rapid scalability to keep pace with evolving research demands
- Energy-efficient infrastructure capable of sustaining large-scale workloads
- Open ecosystem flexibility for integrating proprietary and open-source AI stacks

Solutions:

Under a multi-year agreement, IBM and AMD partnered to deliver a dedicated AI training cluster on IBM Cloud featuring:

- AMD Instinct GPUs for extreme model scalability
- AMD Pensando Pollara 400 AI NICs and Elba DPUs for high-speed networking and data movement
- IBM Cloud's secure, compliance-ready environment ensuring reliability and regulatory adherence for enterprise grade AI workloads

This deployment marks the first large-scale integration of AMD's training platform on IBM Cloud, combining compute, networking, and orchestration capabilities into a unified solution. With this deployment, Zephyra has successfully trained ZAYA1-Base, the first large-scale Mixture-of-Experts (MoE) foundation model trained entirely on the AMD platform, from compute to networking to software. During training Zephyra was able to achieve over 750 PFLOPs of real-world training performance on the cluster and the resulting ZAYA1-base model's benchmark performance is extremely competitive with the SoTA Qwen3 series of models of comparable scale and outperforms comparable western open-source models such as SmolLM3, and Phi4.

IBM CLOUD AND HPC WORKLOADS WITH AMD

IBM Cloud has also adopted the AMD EPYC 9005 Turin series designed for simulation and modelling HPC workloads; for example, IBM introduced the top in line frequency model to run HPC/AI workloads which provides unmatched performance. AMD Instinct GPUs and Turin CPUs offer a competitive platform by which traditional HPC workloads can be empowered by ongoing innovations in the hardware space. Advancing capabilities for supporting HPC workloads within the IBM Cloud/AMD environment allows users engaged in the highest levels of scientific and engineering computing to rely on continued support and innovation for their established workloads as well as more easily integrate AI models into workflows, a practice which has become common even among applications leveraging legacy codes.

IBM Cloud HPC is an example of a flexible, cost-effective, and highly-scaling cloud resource, and is uniquely suited for occasional bursting, hybrid computing, or a full cloud native experience. It is powered by an impressive wide-ranging portfolio of IBM IP to choose from, including:

- Software stack: LSF, Symphony, OpenShift, Storage Scale
- Virtual, containerized, and bare metal provisioning
- x86 instances, GPUs and IBM Power
- Security tools

ENTER QUANTUM COMPUTING

Further cementing their relationship, last year IBM and AMD announced in a [press release](#) plans to cooperatively develop the key technologies needed to facilitate the integration of IBM's quantum computers (QC) into the classical HPC ecosystem. As the promise of true quantum computing advantage could be realized in the next few years, IBM, as a leading supplier of quantum systems, is looking to AMD to help develop the classical hardware and related software needed to extract the best

possible performance from an integrated quantum classical computer environment. IBM is one of the first major QC suppliers to establish such a critical link with a leading developer and supplier of key traditional CPUs as well as AI-centric accelerators.

FUTURE OUTLOOK

As users broadly shift toward production-level generative AI applications, their needs are evolving and growing. Infrastructure, support, and raw compute requirements are not only increasing in scale but in complexity. Innovative solutions sensitive to and informed by these requirements are poised to capture favor in this fast-moving ecosystem. **AMD Instinct hardware on IBM Cloud** aims to not only offer leading edge compute power but seeks to aid users in modernization through a complete toolset of solutions for their most common roadblocks. The customization offered by these tools can bring previously unattainable efficiencies, time-to-solution, ease-of-use, and compliance readiness. With continued support, expansion opportunities, and scalability, this compute environment provides forward-looking products that advanced AI users can trust to take into the future.

About Hyperion Research, LLC

Hyperion Research provides data-driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology, and related trend analysis, and both user and vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors, and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue

St. Paul, MN 55102

USA

612.812.5798

www.hpcuserforum.com and www.HyperionResearch.com

Copyright Notice

Copyright 2026 Hyperion Research LLC. Reproduction is forbidden unless authorized. All rights reserved. Visit www.hpcuserforum.com or www.HyperionResearch.com to learn more. Please contact 612.812.5798 and/or email info@hyperionres.com for information on reprints, additional copies, web rights, or quoting permission.