AMD

A PRACTICAL GUIDE FOR IT LEADERS

MODERN INFRASTRUCTURE FOR THE AI ERA

Accelerate business outcomes with end-to-end AI infrastructure optimized for scale, openness, and innovation.

TABLE OF CONTENTS

HOW IT LEADERS ARE MODERNIZING TO POWER THE AI ERA

Artificial intelligence (AI) adoption is not letting up; in fact, teams are now racing to scale pilot projects into production models that derive real business value. Notably, 80% of CIOs are evaluating AI additions to their tech stack[1] and 85% are modernizing data centers to support AI[2]. Yet legacy infrastructure is often not designed to meet the compute density, throughput, or energy demands of AI workloads. The result is a bottleneck that threatens to slow innovation, inflate costs, and limit the business outcomes this technology promises to deliver.

For IT leaders, the mission is clear: to deliver measurable business results, they need an infrastructure foundation built for AI performance, scalability, efficiency, and interoperability.

By building an AI-ready infrastructure, organizations can quickly deploy, scale, and operate AI initiatives to drive real business results, enhance operational efficiency, and increase competitive advantage.

AI readiness requires an infrastructure that is:

- **Scalable** to meet rapidly changing AI requirements as workloads move from pilots to organization-wide deployment.
- **Open and Interoperable** to provide freedom of choice across environments.
- **Flexible** to support diverse workloads equipped with a combination of optimized CPUs, GPUs, and DPUs for training, inference, and data processing.
- **Trusted** to maintain security, reliability, and compliance as AI becomes embedded across business-critical operations and sensitive data environments.
- **Energy-Efficient** to deliver high performance per watt with modern power and cooling capabilities to help reduce operational costs and support efficiency goals.

AMD helps IT leaders advance toward AI readiness with a future-ready foundation that accelerates innovation while reducing total cost of ownership (TCO). With a broad portfolio spanning CPUs, GPUs, DPUs, and adaptive SoCs, AMD enables customers to match the right combination of compute engines for any workload while maintaining freedom of choice. Flexible, open, full-stack solutions allow organizations to address specific needs using optimized, integrated platforms that adapt as AI demands evolve.

This eBook provides practical insights for IT leaders on upgrading infrastructure to support AI-ready business transformation.

THE NEW REALITY

ITDM PRIORITIES IN THE AI ERA

AI has evolved from experimentation and is now a mission-critical technology, reshaping how IT leaders define success. They must prioritize architecture that drives measurable business outcomes and ensure operational continuity across all environments.

Five core priorities guide technical decision-makers navigating this transformation:

01

Scaling AI Faster

74% of CEOs say that AI is the technology that will most impact their industry,[3] putting pressure on CIOs, CTOs, and other technical decision-makers to expedite deployments from pilots to production. This requires expanded compute capacity for both training and inference, especially as autonomous and reasoning models emerge.

02

Demonstrating ROI

Showing business value from AI investments, not just experiments, marks the shift from AI adoption to AI transformation. Technical leaders must reduce time-to-value by lowering costs, improving productivity and efficiency, and delivering measurable business outcomes that secure executive and board buy-in.

03

Embedding Resilience and Security

As AI can touch every business function and data system, maintaining data security, privacy, and regulatory compliance is imperative to minimize risk. Adopting zero-trust, compliance-by-design architectures helps protect sensitive information throughout the AI lifecycle.

In essence, IT leaders are moving beyond managing systems to orchestrating AI-driven enterprises, where infrastructure becomes the catalyst for innovation and long-term value creation.

04

Closing the Data-to-Value Gap

Poor data governance and fragmented technology stacks remain big obstacles. Teams are struggling with inconsistent data formats, outdated training data, fragmented access controls, and integration delays across hybrid environments. Leaders need interoperability across platforms and standards while reducing technical debt from legacy systems.

05

Future-Proofing Infrastructure

Organizations are making long-term decisions that can scale with evolving AI workloads, seeking full-stack, easy-to-deploy solutions that optimize performance and reduce TCO. These solutions must both serve today's needs and enable future innovation, requiring decision-makers to balance existing operations with the development of an AI-ready foundation.

THE PERFORMANCE IMPERATIVE

ARCHITECTING AI-READY INFRASTRUCTURE

Building high-performance AI infrastructure requires more than just adding compute. It demands a balance between performance gains, TCO, and operational risk. AI workloads span applications, databases, and unstructured data, requiring optimized and orchestrated compute power. They rely on composable, modular, and agile architectures built to handle fluctuating resource demands.

Three architectural imperatives stand out:

Design for Composability

By disaggregating compute, memory, and storage to enable on-demand provisioning as workloads evolve.

Eliminate Bottlenecks

By increasing throughput and reducing latency between CPUs, GPUs, and memory using next-generation interconnects such as PCIe 5.0 and Compute Express Link (CXL) 2.0.

Optimize Continuously

To forecast capacity, anticipate bottlenecks, and inform orchestration policies with observability tools that provide insights into workload tracing, power efficiency, and accelerator use.

Infrastructure optimization extends beyond compute to the entire data pipeline. Modern memory and storage technologies enable scalable data throughput, while tiered architectures balance performance and control cost. Smart network interfaces and intelligent orchestration keep compute resources fully utilized for AI workloads.

AI-driven orchestration, built on Kubernetes or similar open frameworks, leverages telemetry to balance and scale resources automatically as workloads change. Using machine learning (ML), these systems sustain performance and manage costs as demand spikes. The goal is to achieve the highest performance per watt, per dollar, and per workload with infrastructure that adapts as quickly as AI itself.

High-performance CPUs, GPUs, interconnects, and intelligent orchestration tools can enable organizations to execute on these goals. This is where a partner like AMD comes in, providing open, full-stack solutions to help achieve consistent performance, cost control, and scalability from edge to cloud.

OPEN SOFTWARE

AN ESSENTIAL FOR SCALABLE AI INFRASTRUCTURE

While hardware can aid physical scalability, software helps IT teams effectively orchestrate that performance. AI workloads vary widely, from large language models to real-time analytics, and reconfiguring physical infrastructure for each is impractical. Flexible software stacks can help optimize resource allocation across diverse hardware environments and workload types.

Unlike proprietary stacks that limit growth, open platforms build resilience and long-term scalability through orchestration tools such as Kubernetes, Slurm, and Terraform. These open ecosystems enable automated provisioning, workload portability, and consistent performance across hybrid and multi-cloud infrastructure.

AMD ROCm™ software sets the standard for this open approach, providing transparency, interoperability, and performance tuning across leading AI frameworks. AMD ROCm™ is an open software stack offering a suite of optimizations for AI workloads and supporting the broader AI software ecosystem, including open frameworks, models, and tools.

Modernizing toward AI readiness can introduce challenges such as integration complexity, workforce retraining, and temporary downtime. With an open, full-stack architecture powered by AMD, organizations can mitigate these risks through modular deployment and workload portability across hybrid environments. AMD ROCm™ software integrates seamlessly with popular, rapidly evolving open AI tools, including PyTorch, vLLM, SGLang, and Triton compiler, as well as enterprise platforms and ISVs like Red Hat, VMware, and Cohere. This approach minimizes disruption and enables faster time-to-value.

BALANCING PERFORMANCE AND OPTIMIZING EXISTING POWER CAPACITY

For IT leaders, maximizing existing power capacity is a strategic imperative. Building new data centers is costly, making it essential to optimize the performance and efficiency of the current infrastructure rather than expanding the physical footprint. Energy efficiency becomes both a performance and a cost metric, making it a core part of operational strategy.

Notably, data centers consume as much as 2% of global electricity, and analysts project 160% growth by 2030.[4] AI significantly amplifies this demand as AI compute racks may draw up to seven times the power of typical racks.[5]

More than ever, environmental, social, and governance (ESG) practices are under greater scrutiny from regulators and stakeholders. Many customers' buying criteria now include ESG and regulatory readiness.

Investing in modern infrastructure can help enable high performance and optimize electricity use, supporting both business and efficiency goals.

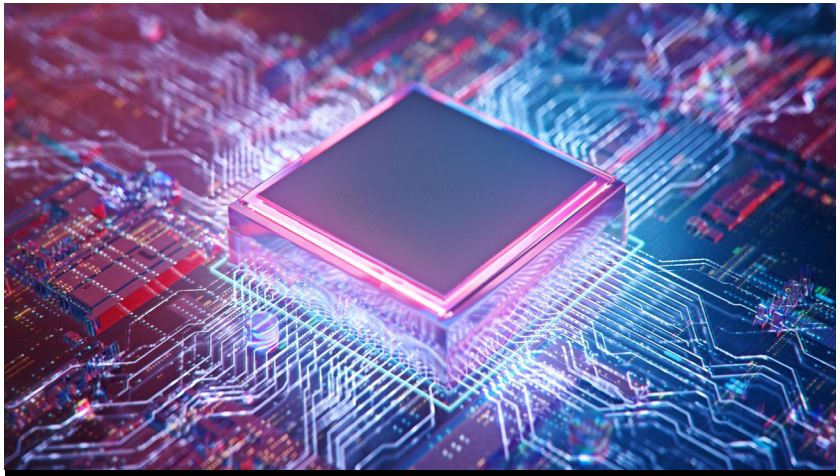
Three main benefits include:

Enhanced AI Performance
Modern CPUs and GPUs accelerate training and inference with efficient parallel processing, reducing runtime and energy use.

Optimized Cooling Costs
Direct-to-chip liquid cooling circulates coolant directly around processors, improving thermal efficiency and supporting sustained performance.

Automated Energy Optimization
Consolidating workloads onto efficient infrastructure frees up space and power budget for additional GPUs and new AI workloads within existing data centers.

Together, these improvements can help enterprises balance innovation demands with efficiency goals while maintaining cost control.



"The price, performance, and memory bandwidth of AMD EPYC™ CPUs are perfect for our needs. They are world-class.[6]

David Baldwin
High Performance Computing Manager, Shell IT

Read the case study

OPTIMIZING CLOUD AND HYBRID STRATEGIES FOR AI

Many organizations are deploying AI in the cloud; however, not all AI workloads belong there. IT leaders need to strategically balance cost, performance, and data sovereignty, and hybrid models can help them achieve this goal.

When deciding how to handle AI workloads, leveraging a structured framework can help maximize performance and minimize costs.

The following framework can help determine optimal placement by evaluating:

To prevent cloud sprawl, implement observability tools that deliver unified visibility across all environments, as well as AI pipelines that scale only where needed. Unified orchestration connects on-prem and cloud resources, enabling a hybrid model that is both agile and cost-efficient. With 71% of IT decision-makers planning increased AI/ML infrastructure spending,[7] this approach helps ensure infrastructure is AI-ready and can scale efficiently as demands grow.

Workload Profile

GPU-accelerated instances are best for compute-intensive tasks such as training large-scale models, while CPU-optimized virtual machines (VMs) are more cost-effective for inference.

AI Lifecycle Phase

Data preprocessing requires CPU instances with high I/O rates; model training needs high-end GPU instances on demand; inference varies by scale and latency requirements.

Cost-Performance Balance

Use independent benchmarking to evaluate performance-per-dollar across VM types and providers, considering underlying hardware specifications that can reduce the number of VMs needed.

MODERNIZING AI INFRASTRUCTURE FOR SCALABLE SUSTAINED GROWTH

BEST PRACTICES FOR IT LEADERS

01

Design Infrastructure to Scale Across AI Workloads

Adopt a flexible foundation by integrating CPUs, GPUs, DPUs, and accelerators to match each task with the right performance profile. By unifying data pipelines for high throughput, applying AI-driven operations for intelligent resource allocation, and embedding sustainability at every layer, IT leaders create adaptive infrastructures that expand seamlessly with business demand and speed up AI innovation efficiently, securely, and at enterprise scale.

Real-World Application

A healthcare organization deploys a shared AI platform that supports both clinical operations and research workloads. EHR and imaging data flow through interoperable analytics pipelines, allowing the same compute resources to power diagnostic models, patient triage automation, and population health studies without requiring separate systems or data silos.

Real-World Application

A telecommunications provider uses unified observability to correlate performance, cost, and efficiency metrics across 5G and edge networks. Predictive analytics automates anomaly detection and resource allocation, enabling teams to resolve issues before they affect service quality and to maintain consistent performance across distributed environments.

02

Establish Unified Visibility Across Your Systems

Connect data from compute, storage, networking, and AI workloads to gain end-to-end insight into performance, security, and cost across hybrid and multi-cloud environments. Through standardized monitoring tools, shared data models, and API-based integrations, IT teams can eliminate silos and ensure consistent visibility from core to edge. Layered with automation and predictive analytics, this end-to-end approach enables proactive optimization, faster troubleshooting, and stronger governance.

03

Embrace Open, Interoperable Ecosystems

Adopt an open environment to gain the freedom to choose which solutions across hardware, software, and cloud platforms to integrate. This interoperability accelerates innovation that enables teams to adopt emerging AI frameworks, data tools, and automation technologies as they evolve, while protecting prior investments. It also enhances scalability and collaboration, allowing workloads to move seamlessly across hybrid and multi-cloud environments. Equally important, open ecosystems strengthen security and compliance by improving visibility and governance across diverse systems.

Real-World Application

An enterprise data center implements AI-driven workload orchestration that dynamically allocates compute to the most energy-efficient nodes. Thermal sensors feed real-time data to power-management algorithms, enabling facilities to automatically adjust cooling and reuse excess heat. These coordinated systems sustain high performance while advancing corporate sustainability goals.

05

Accelerate Time-to-Value for AI Initiatives Across the Enterprise

Speed-to-impact now defines competitive advantage. Faster deployment and integration reduces pilot fatigue, prevents resource drain, and helps teams adapt quickly to market change. By aligning infrastructure, data readiness, and orchestration, organizations can shorten the path from model development to business value, turning AI from a proof of concept into a scalable driver of performance and growth.

Real-World Example

A manufacturing company standardizes its production systems on an open, containerized platform that connects AI-driven quality control, process simulation, and robotics systems. Engineers can integrate new analytics tools without redesigning workflows, enabling faster design iterations, improved detection of defects, and closer coordination between factory and R&D teams.

04

Balance Performance with Energy Efficiency

Shift the focus from raw compute power to performance per-watt, using CPUs, GPUs, and storage optimized for intelligent power management and dynamic scaling. Modern data centers combine liquid cooling, renewable energy, and AI-assisted thermal controls to reduce emissions while maintaining reliability. Aligning workloads with the right resources extends hardware lifecycles and helps avoid unnecessary energy use.

Real-World Application

A financial services firm connects its AI development and production analytics environments through a unified orchestration framework to shorten the handoff between data science and business operations. New models for fraud detection and market analysis are deployed automatically through secure orchestration, enabling risk and trading teams to apply updated insights sooner and adapt more quickly to evolving business needs.

FROM VISION TO VALUE

ADVANCING THE AI ERA WITH AMD SOLUTIONS

AI infrastructure decisions made today will define enterprise capability for years to come. AMD provides the technical foundation and strategic flexibility IT leaders need to deploy AI at scale, with the freedom to integrate across ecosystems and adapt as requirements change.

With a broad portfolio of compute engines, including CPUs, GPUs, DPUs, adaptive SoCs, and AMD Ryzen™ AI processors, AMD empowers enterprises to align compute resources precisely with each workload, from model training to inference at the edge.

With an open ecosystem, anchored by the AMD ROCm™ software stack and deep industry collaborations, organizations can accelerate innovation and experience seamless integration across platforms and frameworks. By unifying compute, networking, and orchestration, AMD provides a cohesive platform designed to scale AI efficiently and sustainably.

From the data center to the edge, the AMD portfolio gives IT leaders the flexibility to scale AI and generate measurable business impact.

DISCOVER HOW YOU CAN DRIVE BUSINESS OUTCOMES WITH AN AI READY INFRASTRUCTURE.

LEARN MORE

TALK TO AN AMD EXPERT FOR YOUR ENTERPRISE NEEDS.

Get started

[1]PwC, Digital Trends in Operations Survey 2025, 2025. [2]FTI Consulting, Global CFO Report, 2025. [3]IDC White paper, IT Modernization Maturity Assessment Prepares Enterprises for AI-Fueled Digital Business Success, doc #US53640525 July 2025. [4]S&P Global Market Intelligence, Data center transformation and AI considerations, 2024. [5]5-Point Likert Scale; Watchtower AI-powered research platform powered by Intercept, Global, n=8,668. [6]5-Point Likert Scale; Watchtower AI-powered research platform powered by Intercept, Global, n=147,984. [7]The Conference Board, CEO and C-Suite ESG Priorities for 2025, 2025. [8]Kearney, Beyond survival: the new operational playbook separating leaders from laggards in, 2025.