White Paper

Taking a Pragmatic Approach for AI-Ready Infrastructure

Sponsored by: AMD

Ashish Nadkarni

October 2024

IDC OPINION

Artificial intelligence (AI) has become a cornerstone of modern technological advancements, driving innovation across various sectors. With AI as a foundation, organizations can take their efficiency, cost savings, and revenue potential to new heights and usher in a new era of a hyperproductive workforce. However, the successful deployment and operation of AI systems hinge significantly on the underlying infrastructure. Infrastructure is easily the least understood but the most crucial component of an AI "stack."

The reason for organizations to underestimate the role played by infrastructure becomes clear when we examine some of the misperceptions of AI itself. Many businesses think of AI as a brand-new entity in their IT environment. The fact is that AI can be a set of use cases, a set of workloads and, crucially, a set of complex functions that enhance existing use cases and workloads. This landscape of composite AI workloads and functions requires infrastructure and operations (I&O) teams to alter their strategic view of datacenters that house these "AI environments."

There is no "one size fits all" approach to AI infrastructure. On one side, overpivoting to expensive accelerators (e.g., certain types of GPUs) could result in underutilization and low return on investment assuming everything runs well on general-purpose infrastructure. It is imperative for enterprise architecture teams to fully understand the specific AI needs of their business before investing in expensive infrastructure such as GPUs. The specific focus should be on the infrastructure being fit for purpose and providing meaningful return on investment. For example, many AI tasks, especially smaller inferencing workloads, and embedded functions within existing business workloads can be efficiently managed by the primary processors (CPUs), thus avoiding unnecessary costs. Coprocessors or accelerators (e.g., GPUs) can be used for specific functions such as large model training or tuning tasks. It is important for I&O teams to know the competitive landscape and the specific capabilities of various accelerators available in the market today. It is important for these teams to understand how the choice of the primary processor can influence the performance of the accelerator. As such, it is important to see the processor and coprocessor (accelerator) as a combined entity.

Finally, AI is primarily a data problem. I&O teams must consider the implementation of "private AI," which involves running AI workloads on premises to help ensure sensitive data is analyzed securely. IDC sees many enterprises fall short when it comes to protecting data in use in addition to data at rest and data in flight.

IN THIS PAPER

This white paper discusses the strategic role played by infrastructure supporting AI workloads, emphasizing the importance of rightsizing AI solutions based on specific needs. It highlights the potential for using processors (CPUs) for smaller AI tasks and the necessity of coprocessors or accelerators (e.g., GPUs) for larger, more complex models. This paper underscores the need for modernizing datacenters to accommodate AI advancements, the role of private AI for sensitive data, and the importance of a balanced CPU-GPU integration to maximize efficiency and performance.

SITUATION OVERVIEW

Artificial intelligence has become a pivotal technology for businesses across various industries, driving efficiency, innovation, and competitive advantage. Ways in which AI can redefine businesses include:

- **Operational efficiency:** AI enhances operational efficiency by automating routine tasks, optimizing supply chains, and improving asset management. For instance, AI can accurately predict demand, streamline logistics, and help reduce downtime through predictive maintenance.
- **Enhanced decision-making:** AI can provide businesses with advanced analytics and insights, enabling better decision-making. AI-driven processes can analyze vast amounts of data to identify trends, forecast outcomes, and recommend actions, thus supporting strategic planning and operational adjustments.
- **Customer experience:** AI can improve customer service through chatbots, virtual assistants, and personalized recommendations. These technologies often enable efficient handling of customer inquiries, provide tailored suggestions, and enhance overall customer satisfaction.
- **Product and service innovation:** AI enables the development of new products and services by leveraging data insights and advanced algorithms. For example,

AI can assist in drug discovery in healthcare or create personalized marketing content in retail.

- **Risk management and mitigation:** AI systems can detect fraudulent activities and manage risks by analyzing patterns and anomalies in data. This is particularly valuable in sectors like finance and insurance, where AI can help prevent fraud and ensure regulatory compliance.
- **Employee productivity:** AI tools can augment human capabilities, allowing employees to focus on higher-value tasks. AI-driven automation reduces the burden of repetitive tasks, while AI-powered analytics provide employees with actionable insights, enhancing productivity and engagement.

By automating processes and optimizing operations, AI helps businesses reduce costs, minimize errors, and improve resource utilization. This includes savings from reduced labor costs, minimized errors, and improved resource utilization. AI also enables businesses to adapt quickly to changing market conditions. AI systems can scale up or down based on demand, ensuring that businesses remain agile and responsive. AI helps businesses navigate complex regulatory environments by monitoring compliance with data privacy laws and other regulations. AI systems can manage and defend sensitive data, reducing the risk of breaches and noncompliance.

The Many Forms of AI

As mentioned previously, AI is not a singular workload or use case. AI is a collective term that encompasses a variety of use cases, workloads, functions, and algorithms. Moreover, AI is not entirely new. Enterprises have been investing in predictive and interpretive AI systems for a while now. For example:

- **Predictive AI systems:** Enterprises deploy machine learning (ML) systems in use cases such as personalization and pricing optimization engines, to improve fraud and cyberthreat detection, and in process and operations automation and optimization systems.
- **Interpretive AI systems:** Similarly, many enterprises deploy recommendation engines to offer targeted product suggestions based on user patterns, to combine media to optimize consumer interaction and retention, and to enhance customer service, responsiveness, and accuracy.
- **Generative AI systems:** These systems can accelerate research and time to insights, offer frictionless human-machine communication, and provide automated transcription and summarization.

On a relative scale, it is generative AI that is new and has the potential to disrupt the status quo. From a functional perspective, AI takes many forms, including:

- **Training:** It forms the most data processing (read computing)–intensive part of the AI life cycle. It is a set of tasks that seeks to create a model or a system that can understand language, images, and other data types. Variations in training include fine-tuning and optimization of existing models using new data sources. Retrieval-augmented generation (RAG) is a form of model optimization.
- **Inferencing:** Once trained, the AI model requires comparatively less processing power to process incoming data and business records in real time. Inferencing can be a network- or storage-intensive task.
- **Embedded AI:** Traditional workloads and data stores provide the foundational business process computing and many data sources to feed AI innovation and insights. In many cases, the model once trained becomes a function within these traditional workloads. The integration of AI into traditional workloads, referred to as "embedded AI," is becoming more common, enhancing the functionality of existing enterprise applications without significant changes to the core workload.

The Role of Infrastructure for AI

Each of the AI-related tasks mentioned previously has a distinct set of infrastructure requirements. Not all enterprises invest in all three. It is likely that for most enterprises, inferencing and embedded AI are going to form the bulk of their AI investments, with RAG as the only training portion. Enterprises shouldn't invest with the assumption that the infrastructure must be designed for training activities only.

Further, when considering the strategic implications of AI investments on infrastructure, several key factors come into play. These considerations include the choice of compute and storage stacks, the location, deployment model, data security, networking, energy efficiency, and cost optimization:

- **Computing, memory, networking, and storage:** It is true that some AI workloads, especially those involving generative AI and large language models, demand significant computational power, memory, and storage. For such workloads, AI infrastructure should support parallelization and integration with data repositories, enabling high performance and scalability. Similarly, networking is critical for optimizing AI workloads, minimizing training times, and handling network transients effectively. However, smaller AI workloads such as inferencing tasks and embedded AI functions in existing workloads work well on general-purpose infrastructure, such as that supporting virtualized environments. Enterprises need to balance these foundational building blocks to support evolving AI needs.

- **Location and deployment strategy:** For many enterprises, the choice of public cloud versus private infrastructure comes down to the nature and location of the data sets being used in their AI environments. Protecting sensitive data within AI models is crucial and involves robust encryption, access controls, and comprehensive data governance strategies, which influence the deployment location. Depending on where the data generation and placement occurs, it may be important to place infrastructure closer to data sources, enhancing performance and reducing network traffic. Enterprises must also examine the choice of AI platforms and developer tools and the overall budget focus, which comes down to reducing the need for capital investments in private infrastructure.
- **Energy efficiency and sustainability:** Enterprises must acknowledge that as their investments in AI workloads go up, they are also going to impact carbon emissions. Datacenters are already major energy consumers, and AI is only going to make the situation more challenging. It is important for them to consider how to optimize their energy consumption through methods such as advanced cooling systems, the appropriate choice of processors and coprocessors, and renewable energy sources. Deploying microgrids can enhance reliability and sustainability by providing a mix of local and grid-sourced energy. Finally, securing prioritized delivery schedules for critical infrastructure components can help minimize supply chain disruptions.
- **Observability:** A key pillar of AI-ready infrastructure is visibility into the various elements of the computing, storage, and networking stack. Being able to gain insight into infrastructure utilization, security, and service quality is the first step to helping ensure its ability to scale along with the workloads themselves. With proper visibility, organizations can optimize return on investment.
- **Cost optimization:** In addition to the infrastructure considerations, enterprises must examine infrastructure-as-a-service offerings for standalone AI environments, and especially those that require expensive coprocessor workloads can help avoid unnecessary infrastructure investments. Cost optimization must also factor in data movement. Minimizing data movement can reduce costs, especially in public cloud environments where data egress fees are common. Similarly, performing inference at the edge can reduce data movement and associated costs.

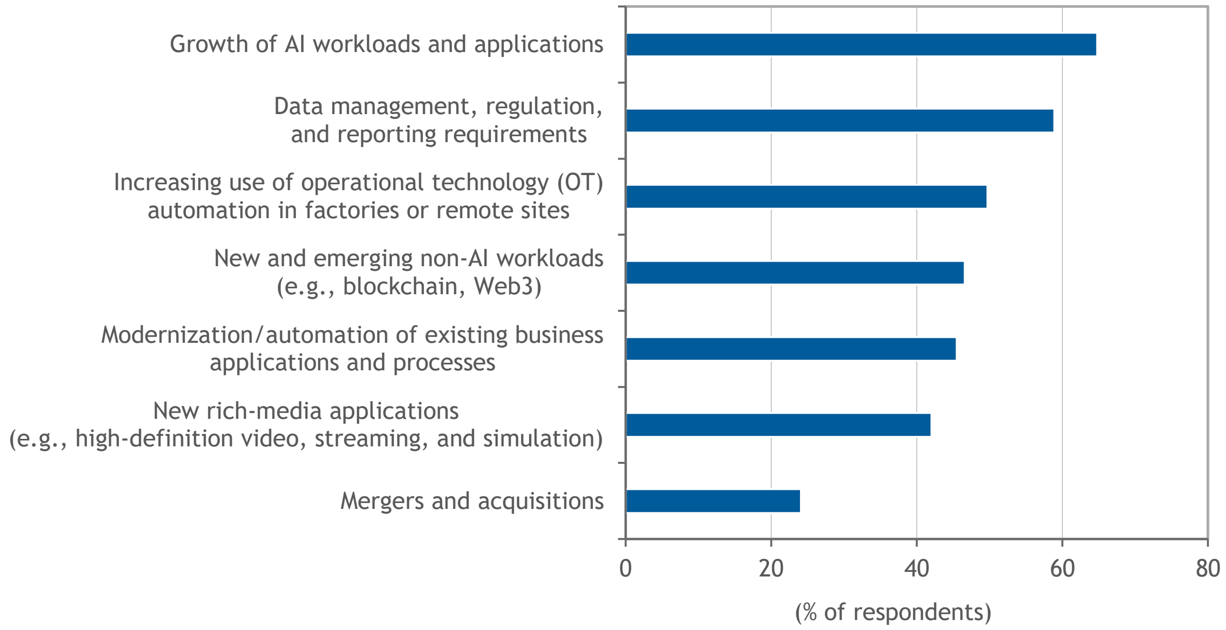
Planning for an AI-Ready Infrastructure

IDC finds that IT buyers are concerned about the drivers of change in infrastructure utilization over the next two years. AI workloads and applications as well as data management, regulation, and reporting requirements are going to be the dominant drivers (see Figure 1). Planning now is imperative so that CIO and IT decision-makers can be prepared for this change as and how it happens.

FIGURE 1

AI and Data Are the Most Important Drivers of Change in Computing and Storage Utilization Over the Next Two Years

Q. *Which of the following trends will drive the most change in your organization's utilization of compute and storage resources over the next two years?*



n = 1,129

Note: Data is weighted by country IT spend.

Source: IDC's *Worldwide Digital Infrastructure Sentiment Survey,* June 2024

Enterprises can take a pragmatic approach to creating an infrastructure environment that fits the evolving needs of their AI workloads. IDC recommends a three-pillar approach:

- **Making room in datacenters by replacing outdated servers:** Repurposing outdated servers can be costly in the long term. Newer servers are typically not just more capable but also generally more efficient.
- **Addressing virtualized, containerized, and bare metal workloads using a private or hybrid cloud strategy:** This helps ensure that each AI workload is mated to an appropriate infrastructure stack.

- **Optimizing coprocessor (accelerator) investments with the right type of processors (CPUs):** This results in accelerator resources that are not underutilized or unevenly utilized.

This strategy aims to enhance datacenter efficiency and performance without the need for extensive new infrastructure.

The Role of a Processor (CPU) in AI Workloads

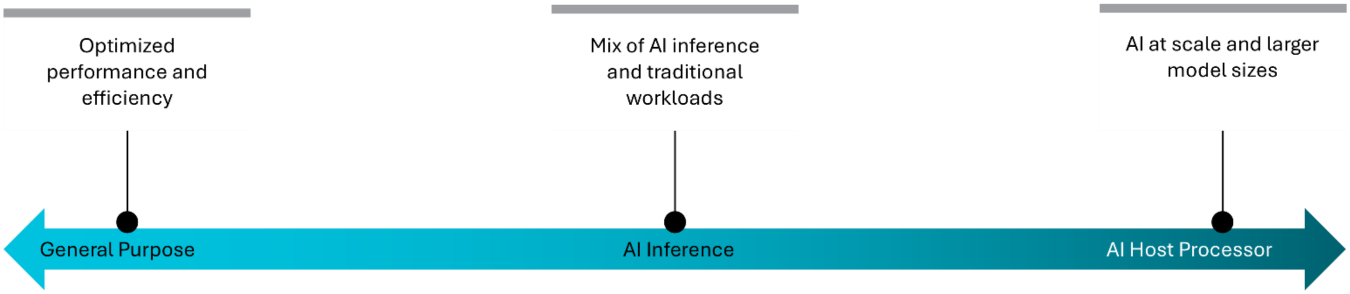
Figure 2 illustrates the three ways in which a host processor (CPU) can enable differentiated outcomes in any infrastructure supporting AI workloads. In detail:

- As a standalone processor (CPU), it can offer optimized performance and efficiency in general-purpose, mixed workload environments.
- As a standalone processor (CPU), it can also offer optimized performance and efficiency in mixed AI inference and traditional workload environments.
- As a host processor (CPU) in accelerated environments, it can enable AI training and inferencing at scale.

In all three cases, traditional and AI workloads can be bare metal, virtualized, and containerized, as well as embedded and standalone in nature.

FIGURE 2

Spanning Traditional Compute, Mixed AI, and AI at Scale with Optimized CPU + GPU Solutions


Optimized performance and efficiency
Mix of AI inference and traditional workloads
AI at scale and larger model sizes
General Purpose
AI Inference
AI Host Processor

Source: IDC, 2024

AMD EPYC PROCESSOR VALUE PROPOSITION

AMD has sought to power the full range of AI workloads in the datacenter and cloud. With its portfolio of processors (CPUs) and accelerators (GPUs), AMD EPYC CPU- and AMD Instinct accelerator–powered servers can be deployed in environments that service embedded AI workloads, inferencing tasks, and training tasks, no matter how large or small.

On their own, the performance and density of AMD processors, combined with their high memory density and I/O characteristics, enables consolidation of mixed traditional workload environments. This is very compelling for embedded AI and AI inferencing use cases wherein the processor (CPU) does the bulk of computations. AMD claims that EPYC processor–based servers and cloud instances enable fast, efficient AI inference close to your enterprise data, driving transformative business performance.

In accelerated server environments, the high CPU frequencies and core counts, large cache, and high memory and I/O bandwidth offered by select AMD processors (CPUs) enable fast time to insights by reducing GPU idle time and improving resource utilization. With more enterprises investing in expensive accelerated server environments (aka "GPU clusters"), investing in the right processors (CPU) can provide significant performance benefits enabling businesses to get the most out of their investments. Here too, according to AMD, its high-frequency 5th Gen EPYC processors are an ideal choice for unlocking the true potential of your large AI workloads by helping maximize GPU accelerator performance and overall AI workload efficiency.

Further, the AMD processor (CPU) and accelerator (GPU) family is designed with datacenter efficiency in mind. The energy-efficient AMD CPU and GPU family optimizes datacenter performance per watt, contributing to effective power and cooling management. Upgrading to the latest AMD EPYC-powered servers and consolidating workloads enable enterprises to save space and energy, reallocating resources for accelerated server environments with high demands. This results in highly performant, high-density compute solutions that support demanding workloads while reducing the need for increased datacenter space and power consumption. By upgrading to the latest AMD EPYC CPU–powered servers and consolidating their workloads, enterprises can free up the space and energy in the datacenter and repurpose these resources for power-hungry, accelerated server environments.

Finally, one of the key pillars of an AI-ready infrastructure is end-to-end data security. AMD EPYC processors are designed with security in mind. Built in at the silicon level, their "Security by Design" approach includes a set of advanced security features and a silicon-embedded security subsystem to help protect an enterprise's most valuable asset: data. AMD encourages businesses to utilize the expanding network of technology partners that implement these features to enable confidential computing, addressing security concerns associated with migrating sensitive applications and data to the cloud.

GUIDANCE FOR IT BUYERS AND DECISION-MAKERS

Key Takeaways for CIOs

CIOs must ask their teams to take a pragmatic approach. It requires acknowledging that AI is not a singular entity. AI workloads and use cases are as diverse as they get: a combination of standalone workloads (both large and small), use cases, and functions within other workloads.

The best way to effectively manage the AI workload spread is to take a fit-for-purpose approach that relies on processors and accelerators, with the choice depending on the specific requirements of the tasks. Processors can handle many AI tasks, especially smaller inferencing workloads, while specialty accelerators (e.g., GPUs) are necessary for larger, more complex training tasks. When specialty accelerators (e.g., GPUs) are necessary, they need a good processor (CPU) paired with them to help them be most productive. In other words, maximizing the potential of your accelerator investment requires a powerful processor (CPU) to drive the system.

Ushering in AI in the environment must be complemented by datacenter modernization initiatives. Replacing outdated servers with newer, more efficient hardware can free up space and improve efficiency, making room for AI workloads without the need for additional infrastructure. Similarly, security and data protection are critical when implementing AI, with a need for robust solutions like AMD Infinity Guard to help ensure data integrity, especially in private AI deployments.

Key Takeaways for I&O Teams and Enterprise Architects

As AI becomes a mainstream component of enterprise IT workloads, it is imperative that the technical and business teams be fully aware of what it means for the infrastructure supporting these workloads. Enterprise architecture teams, for example, can develop clear strategies and guidelines on when and how the workloads can rely on processors (CPUs) only and when they must augment the environment with accelerators like GPUs. With the appropriate and balanced systems architecture that is based on a capable processor platform as the foundation, enterprise architects can create a blueprint for maximizing the value of their organization's investments.

Here are some recommendations to get started:

- Educate yourself and the organization on the diverse types of AI workloads and the appropriate hardware configuration required to meet the desired service-level objectives.

- Start with assessing your current datacenter capacity and identify opportunities for consolidation by replacing outdated servers with current and more efficient hardware platforms.
- Determine your organization's specific AI needs and rightsize the solutions, considering whether small inferencing tasks can be executed in processor-only environments. In many cases, time-to-insight requirements — given the iterative nature of these workloads — could mean a processor-only configuration can get the job done very efficiently.
- Optimize your GPU investment by pairing it with a capable processor (CPU) to enable maximum performance and efficiency. If larger, real-time generative AI tasks require accelerators, determine the type of accelerator best suited for the task. The most expensive accelerator is not always the most appropriate choice.
- Consider the fact that not all accelerators are the same. Even with GPUs, there are products from different vendors that offer better price/performance characteristics and are easily available as part of solutions from leading OEM vendors.
- Consider the role played by the processor in enabling data security. Data security must start at the hardware level and be fully integrated into the higher levels of the stack.

CONCLUSION

AI is transforming businesses by enhancing efficiency, decision-making, customer experience, innovation, risk management, employee productivity, cost reduction, scalability, compliance, and strategic advantage. As AI technologies continue to evolve, their impact on business operations and strategies will only grow, making AI an indispensable tool for modern enterprises. Businesses that effectively implement AI gain a strategic advantage over competitors. AI-driven innovation and efficiency can lead to market leadership and increased profitability. IT organizations must make informed decisions about AI infrastructure, enabling them to choose the right mix of technologies to meet the specific needs of the AI workload.

ABOUT IDC

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, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com

Copyright Notice