SPOTLIGHT
Sponsored by: Intel Corp.

As AI adoption grows, datacenter, cloud infrastructure, and operations teams will need to support new standalone and embedded AI workloads, each with its own computing and data management requirements. Performance, time to market, cost, and security are considerations when buying or expanding cloud environments.

AI Infrastructure in 2025: Balancing Datacenter and Cloud investments

February 2025

Written by: Ashish Nadkarni, Group Vice President and General Manager, Worldwide Infrastructure Research

Introduction

AI is transforming organizations by enhancing efficiency, operations, and intelligent decision-making. Companies are using this technology to codevelop digital products and services, potentially doubling their revenue growth compared with their competitors. Today's AI investments focus on enterprise value creation, which requires net-new spending on IT infrastructure. AI is leading to the induction of newer methodologies like model fine-tuning, retrieval augmented generation (RAG), and the turbocharging of DevOps with AI-infused approaches.

However, businesses are getting smarter about their infrastructure investments. IDC predicts that by 2026, 70% of G2000 C-suite executives will require technology and business outcome alignment paired with return on investment (ROI) and value analysis to justify new AI-enabling infrastructure spending. Unlike past initiatives, most businesses will embrace a hybrid IT operating strategy for their AI workloads. This means hybrid between accelerators and processors, hybrid between public and private cloud environments, and hybrid between standalone and embedded AI workloads.

AT A GLANCE

KEY STATS

According to IDC estimates:

- From 2022 to 2026, enterprise spending on AI-centric systems will grow at 27% annually.
- In 2024, enterprise investments in generative AI solutions surpassed $20 billion. This includes spending on on-premises infrastructure and cloud-based infrastructure-as-a-service solutions.
- By 2028, 75% of enterprise AI workloads will be deployed on hybrid fit-for-purpose infrastructure to turbocharge time to value while optimizing performance, cost, and compliance.

Unlike prior initiatives — which also led to investments in net-new infrastructure — AI-driven investments require planning at all levels and across all teams, especially datacenter and cloud infrastructure teams. AI is poised to have a major impact on the cadence of infrastructure planning and spending as well as on specialized skills development and staffing investments. IDC predicts that by 2028, 75% of enterprise AI workloads will be deployed on fit-for-purpose hybrid infrastructure to turbocharge time to value while optimizing performance, cost, and compliance.

Accordingly, enterprise infrastructure teams (which collectively include datacenter architects, platform engineering, infrastructure and operations [I&O], CloudOps and DevOps teams, and developers) must recommend investments in AI to meet the needs of their business now and in the future. These teams must also consider AI workload–specific requirements when designing and managing an infrastructure stack to meet the requirements for enterprise workloads. The best infrastructure strategy will be an open and flexible hybrid IT strategy, so that DevOps teams will have the ability to quickly experiment and iterate without having to build a special lab environment. Infrastructure team decisions may include:

- Making public cloud infrastructure and platform-as-a-service investments to augment on-premises infrastructure for specific AI model development tasks
- Changing the architecture of on-premises or collocated datacenters to add accelerated or other heterogeneous compute, faster memory, high-bandwidth networking, and flash storage when deploying AI models in production
- Expanding the edge footprint by deploying power- and compute-optimized platforms to handle AI inferencing tasks across multiple remote locations
- Developing a data compliance and infrastructure procurement and operations cost model that ensures the use of appropriate resources for the task in question
- Augmenting existing I&O and developer skills and staff for AI application development and optimized performance-intensive computing (PIC) infrastructure
- Expanding existing DevOps methodologies and processes to include AI application development, AI-centric business workflows, and AI data pipelines

Implementing AI is a multifaceted, businesswide initiative. For infrastructure teams, this means employing strategies to deploy AI workloads fast to realize potential revenue growth while addressing challenges related to cost optimization, data compliance, and security. Ensuring tangible and measurable ROI is a crucial component to any AI initiative.

Investing in AI – The Smarter Way

Enterprises are increasing investments in AI to expand revenue streams in tandem with business differentiation. Whether their focus is to accelerate product and service innovation using newer, never-before-seen approaches; augment human thinking and actions; streamline operations; or reduce costs, AI will be the engine that powers the business. Among the major use cases for AI in business, IDC is seeing momentum in areas like customer engagement (sales, marketing and communications, and customer service), IT optimization, code generation, and augmented threat intelligence. In addition, AI is continuing to deliver value across every major industry, including manufacturing, retail, healthcare, and financial services. IDC finds that from 2022 to 2026, enterprise spending on AI-centric systems will grow at 27% annually. In 2024, for example, enterprise investments in generative AI (GenAI) solutions surpassed $20 billion.

Infrastructure spending accounts for half of more of these investments. It includes spending on on-premises infrastructure and cloud-based infrastructure-as-a-service solutions. However, this spending is being closely scrutinized by the C-suite. Business and IT executives are requiring that technology and business outcomes are aligned and paired with ROI and value analysis to justify new AI-enabling infrastructure spending. This means that now is the time for IT teams to return to the drawing board. IDC predicts that by 2027, after suffering multiple AI project failures, 70% of IT teams will return to basics and focus on AI-ready data infrastructure platforms for data logistics, quality, governance, and trust.

Provisioning the right infrastructure for AI is not simple. IDC's research shows that many AI initiatives either fail or deliver subpar results because of underinvestment in infrastructure. On the other hand, overinvesting in infrastructure can be costly and have a larger-than-expected environmental impact. Making the right set of decisions regarding infrastructure starts with a collaborative approach among various teams. Deployment considerations for AI infrastructure must include:

- **Managing return on investments.** Infrastructure teams must determine the optimal investment strategy, whether purchasing new hardware for deployment in on-premises datacenters or collocation facilities or investing in public cloud services for AI workloads. Each approach requires a particular cost management strategy.
- **Focusing on staffing and skills.** Organizations must plan to address staffing and skills-related issues. IDC predicts that through 2027, finding people with the right skills will remain challenging, reducing the ROI potential for 60% of organizations.
- **Getting to market quickly.** Operationalizing AI is complex, and there is no one-size-fits-all solution. While some AI projects may benefit from new types of hardware and expertise, many can be deployed on existing datacenter and cloud platforms, potentially with upgrades to the latest generation of infrastructure.
- **Maintaining data governance**. Many AI projects require the applications to access sensitive or regulated data globally. Infrastructure teams must take special care to keep data secure and compliant with regulations. Infrastructure teams will need to consider any impact of AI projects on sustainability goals and should be able to address any issues of bias in training data.

Maximize Return on Investments

Improving the ROI in datacenter and cloud infrastructure requires figuring out ways to integrate new AI applications and workflows within currently allocated budgets.

Upgrade General-Purpose Infrastructure

Despite new AI initiatives, budgets will still be under constant pressure. Teams can seek to maximize the value of their infrastructure by being more deliberate with upgrades. The process of replacing old servers and consolidating workloads onto infrastructure that runs the latest generation of processors enables teams to:

- **Add a substantial performance boost for AI workloads.** A refresh of existing CPU infrastructure can enable many AI workloads to be run without specialized hardware. For example, some of the latest CPUs can meet the SLAs to deliver a real-time user experience with less than 100ms token latency on large language models (LLMs) under 20 billion parameters. The latest server processors also deliver higher performance for database, encryption, storage, and other common enterprise workloads.

- **Reduce the overall infrastructure footprint.** Servers running the latest processors have significantly improved energy efficiency. This makes adding capacity to power- and space-constrained datacenters easier while maintaining or reducing energy costs. For example, IDC finds that:
  - Refreshing five-year-old servers with the latest processors can reduce total cost of ownership (TCO) up to 77% by reducing the number of servers needed for the same performance, lowering power costs.
  - When purchasing new servers, selecting the latest processors with built-in AI enhancements over processors that are two generations older can improve performance by a ratio of 5:1 for certain workloads.
- **Reduce software licensing costs.** Many software tools are licensed on a per-core basis. With more powerful cores in newer processors, infrastructure teams can consolidate workloads, reducing the number of software licenses needed.
- **Empower developers.** The most popular open source AI models and frameworks have been optimized to run on x86-based infrastructure (which is the de facto approach in most enterprises). Familiarity with tooling available with this infrastructure enables DevOps teams to deploy AI applications easily without requiring additional highly specialized developer skill sets.

Efficiently Manage Cloud Investments

CloudOps teams can help reduce spending on cloud services by optimizing and automating workload placement, better matching workload needs with the right cloud services. This increases the ROI while providing better service quality.

Invest in Specialized Hardware or Cloud Resources as Needed

After considering how to utilize all existing infrastructure platforms, enterprises should consider adding specialized hardware where and when necessary. Some AI workloads may benefit from GPUs or other discrete accelerators. By strategically procuring only what is essential, enterprises can be more intentional about the type and quantity of additional purchases.

Focus on Staffing and Skills to Extend ROI

IDC predicts that through 2027, the skill sets required to take advantage of net-new AI applications will remain in short supply. This includes the necessary skills to manage hybrid infrastructure environments. Moving from the ideation phase to the successful deployment of AI initiatives will be challenging for organizations that lack the internal skill sets to build and deploy high-performance infrastructure. With applications running on traditional computing platforms, IT departments can generally overcome skills gaps with training. With PIC workloads, IT professionals require sophisticated knowledge to tailor digital infrastructure for specific use cases. Otherwise, they run the risk of skills gaps turning into skills potholes. IDC anticipates that these potholes can and will delay and derail progress to the point that ROI will be achieved only by a significant change in strategy — an achievement that less than half (40%) of organizations will be able to attain. Here are some ways IT organizations can plan now:

- Lean on a technology partner. For those that default to building and deploying their own technology solutions, a mindset shift and a careful evaluation of the most effective partners will be critical to meet distinctive industry needs. Developing the capability to extract the most value from service partner relationships is a strategic imperative.

- Remember that a good technology partner focuses on solutions that provide peak performance and resilient IT services. Such a partner can identify and refine the business unit's needs, often serving as a gatekeeper to protect IP and data while enabling innovation at scale. They will also work closely with line-of-business leaders and service providers to build platforms that drive innovation and differentiation.
- Begin outlining the digital capabilities required by 2028 and restructure line-of-business and IT teams to meet those goals. Identify the industry-specific innovations and types of platforms the organization will need. Invest in industry-specific skill sets and expertise to turn data into insights, transform processes, and accelerate innovation.

Get to Market Quickly to Accelerate Revenue Growth

Getting AI into production quickly can depend on the model type, size of data sets, and choice in hardware.

Except for a few scenarios, most enterprises do not need to develop a model from scratch. Enterprises can usually start their AI journey by fine-tuning or optimizing pretrained models. Pretrained AI models (which can be open source) can often be customized for specific business needs and can help developers avoid training models from scratch. As a result, developers can create AI applications faster and have more time available to innovate with differentiated features. These models don't need to be especially large to be valuable. Smaller "expert" models that are trained on industry-specific or use case–specific data sets may better suit the needs of the enterprise while lowering the compute requirements.

Leverage Existing Datacenter and Cloud Platform Investments

Many enterprises believe they need GPUs to run AI. However, this is not always necessary. Servers with the latest CPUs can deliver excellent performance for inference and even train smaller models (up to about 10 billion–20 billion parameters).

Infrastructure teams must evaluate whether their existing platforms, which are most likely based on x86 architecture, suit some or all AI initiatives. Furthermore:

- If training time is not a critical factor, the existing CPU-based platforms for enterprise workloads can also be used for training, fine-tuning, and inferencing tasks.
- Infrastructure teams with existing public cloud investments must also evaluate general-purpose compute instances for inferencing tasks. This can offer significant advantages in scaling AI applications across multiple geographies.
- Keeping all workloads on existing platforms can streamline workflows (including data preparation, ingest, analytics, and inferencing).
- Infrastructure based on known processor architecture (e.g., x86) helps DevOps teams leverage common toolsets across multiple locations. It also enables greater availability of solutions in the hardware and cloud ecosystem.

Leveraging existing platforms enables enterprises to swiftly transition from proof of concept to production without first procuring expensive, specialized, and scarce infrastructure. Teams can thus scale their infrastructure to keep up with changing business requirements in a rapidly evolving landscape.

Stay Compliant

The stakes with data security, compliance, and adherence to privacy regulations are much higher with AI than with other enterprise workloads. Deploying AI responsibly requires enterprise teams to take an "infrastructure up" approach.

Use Confidential Computing to Deploy AI Globally and Securely

Public cloud services are attractive for many AI deployments, given the global reach of many providers. Unfortunately, this can also expose the organization to data security issues.

Most infrastructure teams are keenly aware of the benefits of securing data at rest (on-disk encryption) and data in flight (network encryption) when designing and implementing infrastructure and application stacks. However, many infrastructure teams do not know that data can still be hijacked with low-level attacks on the system memory. This is where confidential computing comes into play.

Confidential computing (available from select hardware, software, and cloud vendors) lets infrastructure teams move virtual machines, containers, and entire applications into secure computing enclaves. These can be in the public cloud, on premises, or in collocated facilities. Confidential computing complements in-flight and at-rest data encryption with hardware-enabled data-in-use encryption, adding another layer of protection to support a zero-trust security strategy. With confidential computing, teams can:

- Deploy AI applications more securely, at scale, across on-premises infrastructure, collocation facilities, multiple clouds, and edge nodes.
- Set the trust boundary appropriate to their applications to help protect sensitive data and content from advanced attacks, tampering, and theft.
- Employ federated learning (i.e., train AI models from distributed sources without exposing private data). For example, medical researchers can contribute patient data to help train a model that can improve treatment plans.

Gain Transparency into Decision-Making

With increasing data privacy and leakage issues, commercial or prebuilt models are being scrutinized. Infrastructure teams are being asked to come up with ways to examine data sets and methodologies used to train AI models. Using open source AI models, developers can "see" the data used to train models. This can increase transparency, making it easier to identify model bias and understand what is causing it.

Considering Intel

Intel's portfolio of processors, accelerators, software, and networking products is widely used to implement a robust datacenter and cloud infrastructure strategy. Intel's datacenter and AI technologies form the bedrock of solutions that empower companies to transform data into timely and actionable intelligence efficiently and more securely. These solutions cater to a wide range of workload environments, better aligning with budgets.

As predictive AI, GenAI, and high-performance computing (HPC) workloads grow in complexity, their performance and energy-efficiency requirements likewise grow. Intel's approach for achieving an optimal balance of performance and total cost of ownership for these workloads is to design an AI-accelerated system using a host CPU and discrete AI accelerators. For example, Intel Xeon 6 Processors with E-Cores boost datacenter power efficiency with high core density for compute-dense and cloud-scale workloads. Intel Xeon 6 Processors with P-Cores deliver high performance per core

for demanding cloud solutions, offering superior AI performance. Businesses can achieve over 2x better performance than with 5th Gen processors, according to Intel, enabling twice the workload capacity or faster task completion. Businesses can take advantage of this development to improve performance per dollar by running twice as many workloads on the same number of cores or by finishing workloads twice as fast. In addition, Intel's extensive ecosystem of ISV apps, operating system tools, libraries, and frameworks that support these processors means greater ease of use for developers.

Performance

Intel offers an industry-leading portfolio of computing and connectivity hardware for datacenters that deliver the performance and efficiency needed to run a vast range of AI and enterprise workloads at any location or in any deployment and at a low TCO. This includes Intel Xeon processors with built-in AI acceleration for general-purpose computing and inferencing, Intel Gaudi AI accelerators for dedicated AI training and inference, and Intel Ethernet for fast connectivity. Intel's portfolio enables infrastructure teams to implement a cost-effective hybrid AI strategy with seamless core-to-edge-to-cloud coordination. Furthermore:

- Backed by a strong product road map and based on open standards, Intel's portfolio enables smooth upgrades, helping enterprises leverage existing datacenter and cloud platform investments.
- Intel offers processor options that are optimized for specific workload types (e.g., AI training and inferencing and SQL and NoSQL databases) to further improve performance.
- Energy-efficient processors and optimized power settings can lower power consumption while balancing performance.
- Intel works across the full hardware and software stack to optimize performance, including with top enterprise software vendors to leverage the latest hardware performance features.

Choice

Intel's globally distributed ecosystem enables infrastructure teams to build a full stack with industry-leading components from their preferred hardware and software vendors and cloud service providers. In detail:

- Intel hardware and software enable a wide range of production-ready and interoperable AI solutions, cloud services, frameworks, and AI models.
- An open and unified programming model for multivendor, multi-architecture environments enables DevOps teams to streamline development and customize for their needs. Intel offers oneAPI and OpenVINO as part of its open software environment to optimize models for inference across different hardware types.
- Software products and tools, particularly those from the Intel Developer Cloud, make it easier to test, optimize, and place workloads on the best-fit hardware or cloud service.

Trust

Intel's hardware-based security features help secure data in flight, at rest, and in use. Intel's confidential computing ecosystem and trust services enable teams to share data and AI models even when working with sensitive data. They can deploy AI models and other workloads globally while remaining secure and responsible.

Intel's trust goes beyond data security. With decades of engineering expertise, Intel hardware offers a high level of system reliability, availability, and serviceability at scale. Intel's investments in delivering sustainability-focused product features enable enterprises to keep up with workload demands while progressing toward their corporate responsibility goals.

Challenges and Opportunities for Intel

Discrete accelerators have gained considerable momentum as a proxy for AI, but there's more to consider when it comes to the necessary infrastructure stacks for scalable AI. While accelerators such as GPUs provide excellent performance for training and inference, they may be excessive for many AI workloads. Lack of planning during design and implementation can lead to overinvestments. Intel and other vendors will need to help datacenter architects and DevOps teams understand the potential challenges and benefits of investing in the right fit-for-purpose AI stack.

Intel provides a full portfolio of hardware for AI and a strong product road map that lets infrastructure teams predictably upgrade hardware on the same platform, helping to future proof their investments. In detail:

- Intel Xeon processors perform well for inferencing workloads. They are also suitable for some AI training or fine-tuning workloads (e.g., when dealing with smaller AI models).
- Intel Gaudi accelerators can typically achieve better results for popular benchmarks (e.g., MLPerf) at a lower TCO.
- Intel processor and accelerator-based cloud services are widely available and offer a compelling price-performance ratio for AI workloads.
- Intel offers a comprehensive portfolio of confidential computing technologies and services on Intel Xeon processors to meet the diversity of customer needs for AI workloads that depend on sensitive or regulated data.
- Intel has an open ecosystem to enable more choice in AI solutions and confidential computing offerings.

Conclusion

Intel and its ecosystem of partners can enable datacenter architects and DevOps teams to design and implement a long-term AI infrastructure that is resilient, scalable, and responsible. As a result, enterprises can achieve a lower TCO, help ensure data protection, and deploy workloads across virtually any location and deployment type.

Businesses that unlock the power of data will become the leaders of the intelligence era.

About the Analyst

Ashish Nadkarni, *Group Vice President and General Manager, Worldwide Infrastructure Research*

Ashish Nadkarni is group vice president and general manager within IDC's worldwide infrastructure research organization. He specializes in performance-intensive computing infrastructure deployed for AI, HPC, and other engineering workloads.

MESSAGE FROM THE SPONSOR

Intel technologies for bringing AI everywhere

Intel enables the AI continuum in every platform, from client and edge to datacenter and cloud. Intel's portfolio includes a diverse range of processors and accelerators, networking, and software. Together with a world-class ecosystem of hardware, software, and cloud vendors, Intel powers solutions that help enterprises accelerate their adoption of AI and realize value faster.

- For more strategies to maximize the value of your infrastructure investments, register for the IDC and Intel webinar on reducing costs with datacenter and cloud modernization and optimization.
- To learn more about supporting AI workloads to grow revenue, register for the IDC and Intel webinar on innovating with AI.
- To explore how to help protect data while deploying new AI workloads, read the IDC and Intel report on risk mitigation while securing your data.

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
blogs.idc.com
www.idc.com