SPOTLIGHT

Sponsored by: Intel Corp.

Nearly every industry will continue to adopt AI in 2024 to grow revenue and reduce costs. With technology budgets under pressure, the tech C-suite seeks to maximize the return on investment in infrastructure to support AI workloads. They must balance ROI with considerations around time to market, data security, and staffing.

Enterprise AI Strategy in 2024: Maximizing Growth, Return on Investment, and Data Security

March 2024

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

Introduction

AI is a revolution that is accelerating the steady pace of digital transformation (DX) that enterprises have seen in the past decade or so. New AI use cases will turbocharge this transformation. Companies are already using AI to codevelop digital products and services, potentially doubling their revenue growth compared with their competitors. So vast is the scale at which AI — and particularly generative AI (GenAI) — will impact enterprise value creation that in late 2023, IDC's future enterprise research found that 81% of IT leaders expect full-year spending in 2024 to increase or remain stable due to AI initiatives.

AT A GLANCE

KEY STATS

According to IDC estimates:

- Enterprise spending on AI-centric systems will grow at 27% annually from 2022 to 2026.
- In 2023, enterprise investments in GenAI solutions surpassed $19.5 billion. This includes spending for on-premises infrastructure and cloud-based infrastructure-as-a-service solutions.

Unlike traditional digital transformation, which was expected to be a multiyear journey, AI-led digital transformation (AI-DX) is urgent and requires velocity for enterprises to remain competitive and grow. At the same time, it cannot be ad hoc. AI is poised to impact infrastructure strategy and spending at the highest level. Therefore, it requires planning at all levels and across all businesses.

AI has become a board-level conversation. Boards of enterprise companies worldwide are asking their C-suite for a plan on investments in AI to solve business problems. IDC's research on AI adoption found the following sentiments regarding the:

- **State of AI adoption.** 66% of enterprises worldwide said they would be investing in GenAI over the next 18 months.
- **Importance of infrastructure.** Among organizations indicating GenAI will be an increased area of IT spending in 2024, infrastructure will account for 46% of the total spend.

- **Resiliency of investments.** Nearly half of IT organizations expect investments in IT-related security, infrastructure and operations, and AI and automation to be immune from any external economic pressures.

In summary, the tech C-suite must make investments to meet the needs of their business now and in the future. Implementing AI strategies is a multifaceted, businesswide initiative. Tech C-suite executives must therefore look for strategies to get to market fast but also need to address challenges related to cost optimization, data compliance, and security. On the infrastructure side, the tech C-suite must ask their IT teams to develop an infrastructure stack that can be leveraged for any workload, including AI and GenAI. The best infrastructure will be open and flexible, so that DevOps teams will have the ability to quickly experiment and iterate without having to build a special lab environment.

Investing in AI — Business Opportunities

Enterprises are increasing investments in AI to expand revenue streams in tandem with business differentiation. IDC's research finds that enterprise spending on AI-centric systems will grow at 27% annually from 2022 to 2026. In 2023, enterprise investments in GenAI solutions surpassed $19.5 billion. This includes spending on on-premises infrastructure and cloud-based infrastructure-as-a-service solutions. Whether 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.

IDC sees several major use cases for AI in business in customer engagement. For example:

- **Sales.** A constant challenge for sales organizations is finding ways to add new customers and revenue streams. AI can deliver deeper, automated, and targeted insights to sales teams. These insights can lead to better sales margins and operating costs. For instance, sales teams can deploy product or service recommendation systems that deliver more personalized experiences when engaging with potential customers.
- **Marketing and communications (marcom).** With GenAI-enabled tools, marcom teams can speed up and expand their marketing activities, thus gaining the ability to reach new customers and/or newer geographies. For example, automatic language translation, localization, and text and image generation can enable the production of more content while ensuring it is relevant.
- **Customer service.** Customer service teams can leverage GenAI-enabled tools to automate engagement and boost productivity. For instance, they can use chatbots powered by open source large language models to initiate service interactions, expedite the service process, and limit human involvement to escalations and complex issues. AI-enabled fraud detection tools can provide superior customer data protection compared with conventional methods.

AI also has huge benefits on the IT side. With IT budgets constantly strained and IT leaders often asked to do "more with less," AI can augment existing IT staff and processes when it comes to:

- **IT optimization.** AI-enabled tools can improve service quality by optimizing existing infrastructure stacks. Such tools can analyze log files for patterns, detect anomalies, and notify operators of a problem before it disrupts the system. While this process is nothing new, AI can enable deeper data analysis with causality, correlation, and heuristic insights that surpass the competency of traditional tools.

- **Code generation.** Developers can accelerate software development by automating coding-related tasks, freeing themselves up for more strategic initiatives. For example, AI-enabled tools can automate code creation and troubleshoot logic and bugs.
- **Augmented threat intelligence.** With the reputation of a company heavily dependent on its proprietary data and intellectual property, IT security is undoubtedly a priority for CIOs. AI-enabled security tools can significantly enhance the capabilities of IT security teams, leading to better service outcomes. For instance, security operations teams can detect intrusion patterns and halt perpetrators before they access sensitive information. In addition, they can better support the business by integrating fraud detection tools into the IT infrastructure.

Beyond these popular use cases, AI is continuing to deliver value across every major industry, including manufacturing, retail, healthcare, and financial services.

Investing in AI — A Road Map for the Tech C-Suite

Many tech C-suite executives struggle to determine the level of infrastructure investments required to deploy AI. IDC's research finds that skimping on proper investments is a sure-shot way for AI initiatives to fail or deliver subpar results. On the other hand, overinvestments lead to cost overruns and poor returns on investment. The tech C-suite will need to consider balancing their infrastructure investments with the need to get to market quickly, manage costs, and keep data protected:

- **Getting to market quickly.** Operationalizing AI is complex, and there is no one-size-fits-all plan. 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.
- **Managing costs to maximize value.** Tech C-suite executives 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 unique cost management strategy.
- **Maintaining data security and privacy**. Many AI projects require the applications to access sensitive or regulated data globally. Tech C-suite executives must ask their teams to take special care to keep data secure and compliant with regulations.
- **Other considerations**. With the adoption of AI, corporate responsibility has become a significant concern. Tech C-suite executives must consider the impacts of increased infrastructure investments on the sustainability goals set by the organization. They must instruct their development teams to develop a plan for model transparency to avoid unethical bias. Similarly, they need to address staffing and skills-related issues promptly.

There is urgency to invest in AI, but it also requires careful planning. Making the right decisions regarding infrastructure starts with a collaborative approach among various teams.

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 processors can deliver excellent performance for inference and even train smaller models (up to about 10 billion–20 billion parameters).

Tech C-suite executives must, therefore, ask their teams to evaluate whether their existing platforms suit some or all AI initiatives:

- 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.
- Companies 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 — consider that AI often works as a pipeline of many interdependent tasks.
- Infrastructure based on known processor architecture (e.g., x86) helps development operations (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.

Maximize Value and Reduce Costs

Improving the return on investment 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. Tech C-suite executives can seek to maximize the value of their investments by asking their teams to be more deliberate with infrastructure 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 processor hardware can enable many AI workloads to be run without specialized hardware. For example, some of the latest CPUs can deliver a real-time user experience on large language models under 20 billion parameters.
- **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.

- **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

Cloud operations 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 return on investment 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 weigh 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.

Stay Secure and Responsible

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 enterprises to take an "infrastructure up" approach.

Using 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 tech C-suite executives are aware of the benefits of securing data at rest (on-disk encryption) and data in flight (network encryption) when designing and implementing infrastructure and applications. However, many in the tech C-suite 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. With confidential computing, enterprises 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. Tech C-suite executives are asking their teams to come up with ways to examine data sets and methodologies used to train AI models so they can be transparent with decision-making. 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.

In a recent Business Value study on Intel x86–based infrastructure, IDC assessed the impact of running workloads on new Intel-based infrastructure. IDC's analysis showed that the use of 3rd and 4th Gen Intel Xeon processors yielded significant value for study participants. IDC calculated that on average, companies would realize $13.32 million per organization per year ($236,900 per 100 employees using IT services). On the business side, interviewed organizations gain higher net revenue and productivity from getting to market faster, improving the user digital experience, and creating opportunities through new technologies. At the same time, they reported gains from increased development team productivity levels, IT team efficiencies, and IT cost savings from hardware use efficiencies and lower power and facilities costs.

Performance

Intel offers an industry-leading portfolio of computing and connectivity hardware with the performance and efficiency needed to run workloads (including AI) at any location or in any deployment and at a low total cost of ownership . This includes Intel Xeon processors for general-purpose computing and AI inferencing in the datacenter, Intel Gaudi accelerators for dedicated AI training and inferencing, GPUs for AI and general-purpose acceleration, Intel Core processors and Intel Arc Graphics for AI in client and edge devices, and Intel Ethernet for fast connectivity. Intel's portfolio enables the tech C-suite to implement a cost-effective hybrid AI strategy with seamless core-to-edge-to-cloud coordination:

- 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 IT 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 allows 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. Its 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 features in its products 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 the tech C-suite understand the potential challenges and benefits of investing in the right fit-for-purpose AI stack.

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

- 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 demonstrably achieve better results for popular benchmarks (e.g., MLPerf) at a lower TCO compared with GPUs.
- 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 enterprises 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.

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, register for the IDC and Intel webinar 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
Twitter @IDC
idc-insights-community.com
www.idc.com