STELIA

The Enterprise Edge in an AI Centric World.

An Executive Field Guide for 2025

Paul Morrison, CMO & Petter Olafsen, Market Intelligence Analyst

A note from the authors

When we talk to enterprise leaders about AI, their main three questions are typically “How do I get started with AI? How do I make considered infrastructure decisions? And how do I keep costs under control?” All very sensible trains of thought, which is why we put together this report to help you navigate the exciting, sometimes overwhelming, journey into an AI-centric world.

In our conversations with peers and clients alike, we see that adopting AI eventually becomes a business transformation discussion. This report is a personal guide crafted from our experiences in the field, offering a clear roadmap to address the challenges of getting started, managing infrastructure shifts, and controlling costs.

You’ll discover how evolving concepts like agentic AI and distributed inference are redefining what’s possible. These innovations are not only pushing the boundaries of efficiency but are also reshaping traditional IT models. As we transition from CPU-centric to GPU-driven systems, understanding and harnessing these advanced techniques will become vital for maintaining your competitive edge.

Whether you’re taking your first steps into AI or looking to enhance your current strategy, our goal is to provide practical insights and real-world examples that make these complex ideas approachable. We want you to feel confident in making strategic decisions that not only embrace the transformative potential of AI but also ensure your enterprise remains agile, cost-effective, and ready for the future.

Paul Morrison
CMO

Petter Olafsen
Market Intelligence Analyst

The Enterprise Edge in an AI-Centric World

Contents

Executive Summary

The Enterprise Edge in an AI-Centric World

Executive Summary (1/4)

The emergence of an AI-centric world is radically reshaping the global technological and business landscape in real-time, creating both opportunities and challenges for enterprises across all sectors. IDC projects global AI spending to reach $632 billion by 2028, with a compound annual growth rate (CAGR) of 29.0% from 2024–2028. This report, created by the Stelia in-house market analysis team, provides comprehensive insight into the emerging AI-centric world and its implications for business operations, innovation, and strategy.

13 Key Insights

01. Economic Impact:
AI is experiencing unprecedented growth. **McKinsey** estimates that AI technologies could generate **$17.1 trillion to $25.6 trillion** in economic value globally each year. **PWC** estimates in their 2024 AI report "Sizing the Prize" that the potential impact of AI on the global economy will add an uplift of $15.7 Trillion to the global economy by 2030, effectively adding another India and China to the global economy. Figure 1 below illustrates this.

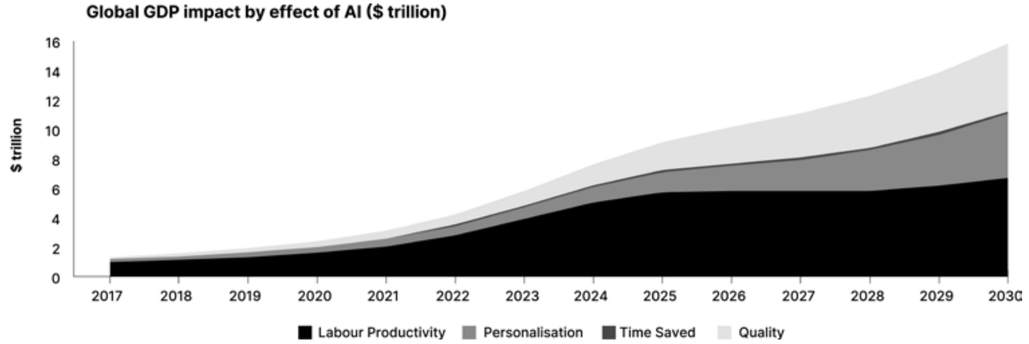


Figure 1 — Source: PwC Sizing the prize

02. Generative AI Growth:
Generative AI is seeing explosive adoption, with **IDC** forecasting its spending to **grow at a CAGR of 59.2%, reaching $202 billion by 2028 and representing 32% of overall AI spending**. This underscores the transformative potential of this technology across various industries.

03. AI Maturity:
The **ServiceNow** 'Enterprise AI Maturity Index 2024' reveals varying levels of AI adoption across enterprises. "Pacesetters" — organisations scoring 50 or more on a 100-point scale — are leading the way, with 33% of them leveraging AI for transformation and innovation, compared to just 14% of other companies.

Executive Summary (2/4)

04. Technological Shift:
A transition from CPU-centric to GPU-centric computing is underway, driven by the demands of AI workloads. This shift necessitates substantial changes in enterprise IT strategies, including adopting high-bandwidth, low-latency network infrastructures to support distributed GPU clusters. Enterprises must reevaluate hardware investments and network designs to accommodate this.

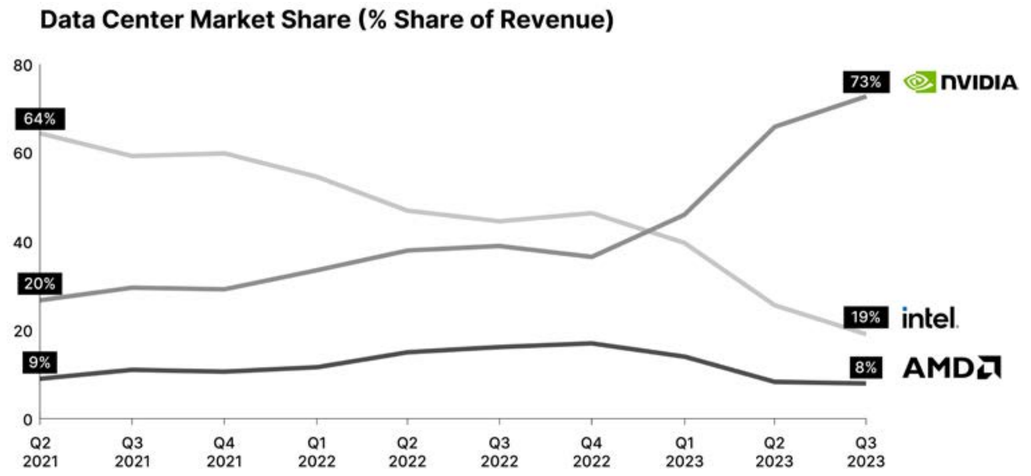


Figure 2 — Source: 10-Qs Filings

05. Data Proliferation:
IDC projects global data volumes to grow at an annual rate of 2.7X until 2027, reaching 291 Zettabytes. AI workloads will increasingly consume storage capacity, driving **data center storage from 10.1 zettabytes in 2023 to 21.0 zettabytes by 2027**. Likewise with network capacity, with estimates suggesting that **by 2030, 75% of all network application traffic will involve AI** content generation, curation, or processing.

06. Infrastructure Challenges:
The growth in AI computation demands is straining existing enterprise IT ecosystems. Issues such as data gravity, integration, scalability, and performance optimisation are becoming critical factors in successful AI implementation. **IDC** reports that **software will represent over half of AI spending**, with hardware as the second-largest category, highlighting the need for robust AI-ready infrastructure.

07. Workforce Transformation:
AI and related technologies have the potential to **automate work activities absorbing 60–70% of employees' time** today. This shift is driving demand for new roles such as AI configurators, data scientists, and machine learning engineers. Organisations are responding with a mix of external hiring and internal upskilling initiatives.

08. Data Sovereignty and Compliance:
Global AI deployments are complicated by data sovereignty requirements. There's a growing need to clarify what types of data fall under sovereignty rules, especially as AI models trained on global datasets and data in motion challenge traditional notions of data sovereignty.

09. Ecosystem Development and Innovative Solutions:

- **Emergence of AI Availability Zones:** Innovative approaches from companies like Stelia include AI Availability Zones or rings being developed to optimise AI wide area network infrastructure. These zones enable GPU-to-GPU workflows within a "metro zone" of a few hundred kilometres, improving data transfer stability and performance.
- **Investments in Network Infrastructure:** Companies such as **Flexential**, **EU Networks**, **Zayo**, and **Light Source Communications** are investing heavily in fibre networks to meet the burgeoning enterprise networking demands of AI. **Lumen's $5 billion deal to connect hyperscalers with AI connectivity** exemplifies the industry's commitment to enhancing private infrastructure.
- **Collaborations and Partnerships:** Enterprises are exploring strategic partnerships with technology vendors, cloud service providers, and network operators to accelerate AI adoption and infrastructure development, recognising that collaboration is key to overcoming complex challenges to accommodate this.

Executive Summary (3/4)

10. ROI and Investment:
On average, **companies are allocating 15% of their technology budgets to AI capabilities**. According to **ServiceNow**, 65% of surveyed organisations are achieving positive ROI from their AI investments, with **23% reporting significant returns (over 15%)**.

11. Cross-Sector Impact:
The Transition to an AI-centric world is driving transformations beyond the tech sector. IDC projects that the **financial services industry will account for over 20% of all AI spending**, while Business and Personal Services will see the fastest AI spending growth at 32.8% CAGR. Banking, high tech, pharmaceutical and healthcare are among the industries that could see the biggest impact as a percentage of their revenues. **McKinsey** illustrates this in their 2024 report 'The economic potential of generative AI', on industry impact.

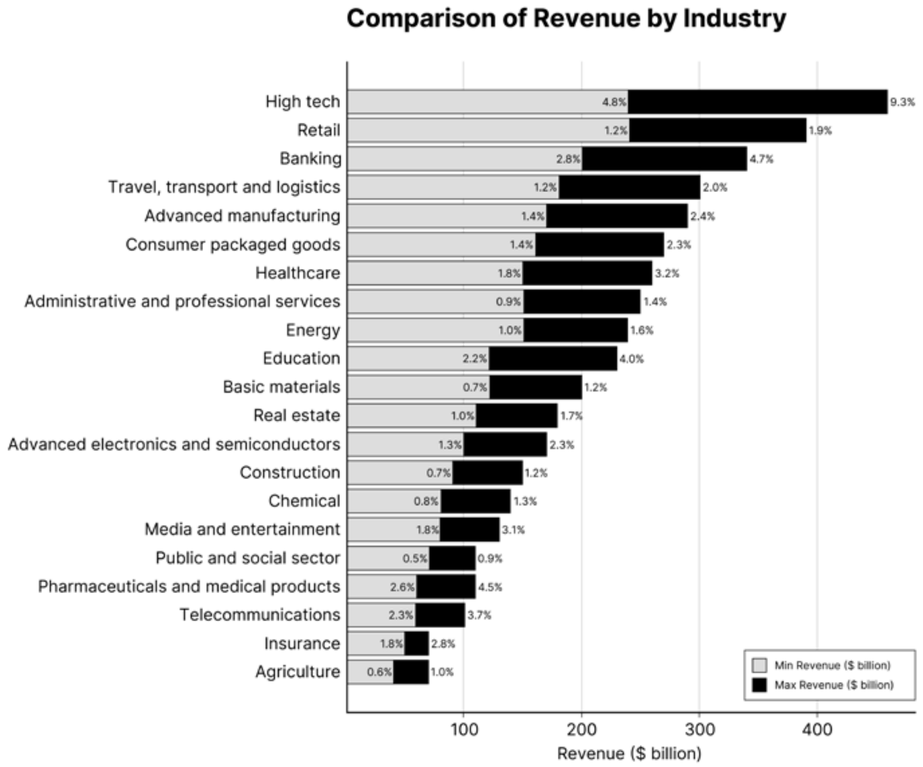


Figure 3— Source: McKinsey & Company, The economic potential of generative AI

12. Geographic Distribution:
AI-driven economic growth is expected to be especially pronounced in China and North America, with **China projected to see a 26% boost to its GDP by 2030 and North America a 14.5% increase**. Together, these regions are anticipated to contribute $10.7 trillion, representing nearly 70% of AI's global economic impact, according to **PwC**. Reflecting this leadership, AI adoption and investment show significant regional disparities. **IDC** forecasts that **AI spending in the United States alone will reach $336 billion by 2028**, making up over half of global AI expenditures and reinforcing the U.S.'s leading role in AI development and integration.

13. Future Trends:
Emerging developments include industry-specific AI chipsets, potential breakthroughs in quantum computing, and new long-haul transport market technologies to address AI networking challenges.

Executive Summary (4/4)

Industry Outlook

As AI continues to develop at a breakneck pace, it will become an integral part of all enterprise operations across all sectors. Through our extensive partnerships with global Cloud Service Providers and technology vendors, we observe that the ability to simply integrate AI capabilities into business processes, breaking down silos between geographies, departments and even between organisations, will be a key differentiator for successful companies.

Our market analysis team, working closely with enterprise clients across sectors, has identified that the future of AI in enterprise is not about adopting or managing “shadow AI” as a quick fix, but about reimagining products, services, business models and operations to fully leverage AI’s potential. This transformation, as evidenced by our cross-industry implementation experience, will require new approaches to data management, skills development, and even organisational structures.

Next Steps for Enterprise Leaders

The transition to an AI-centric world presents both opportunities and challenges. Based on our continuous analysis of successful AI implementations and ongoing dialogue with technology leaders, enterprise leaders should consider the following actions:

01. Assess current AI maturity using frameworks like the Enterprise AI Maturity Index and identify key areas for integration and improvement.

02. Develop strategies for data management and mobility to support AI operations, considering both performance requirements and regulatory compliance.

03. Invest in skills development to build AI capabilities within the organisation, focusing on roles like AI configurators, data scientists, and machine learning engineers.

04. Explore partnerships and ecosystems that can accelerate AI adoption, including specialised infrastructure providers and AI solution vendors.

05. Stay informed about evolving AI technologies, including generative AI and potential quantum computing breakthroughs, and their potential applications in your industry.

06. Establish clear metrics for measuring AI ROI and regularly assess the impact of AI investments on business outcomes.

07. Consider industry-specific AI applications and use cases, particularly in high-growth areas like financial services and business services.

08. Plan for significant infrastructure investments, balancing software and hardware needs to support AI initiatives.

By taking a proactive and strategic approach to AI adoption and integration, enterprises willing to embrace change can position themselves to thrive in the emerging AI-centric business landscape. Our extensive ecosystem of partners, spanning cloud providers, technology vendors, and enterprise clients, consistently demonstrates how this drives innovation, efficiency, and competitive advantage.

The following chapters explore deeper into the primary facets of this AI-centric world, beginning with an exploration of how we arrived at this inflection point in technological evolution. As we examine the drivers, challenges, and opportunities presented by AI, keep in mind the key insights and action items outlined above. They will serve as a framework for understanding the detailed discussions that follow.

Introduction – The Emergent AI-Centric World

Introduction — The Emergent AI-Centric World (1/3)

The AI Tipping Point

Building on the thirteen key insights presented in the Executive Summary, this chapter sets the stage for our comprehensive exploration of the AI-centric world. Through our unique position orchestrating AI operations across multiple industries and our continuous engagement with leading AI technology providers, we've observed first hand the tipping point that has propelled us into this new era of technological transformation.

Our market intelligence team, working in concert with our global network of cloud and technology partners, has documented how AI has evolved from an experimental technology into a foundational driver of business innovation and competition. The rapid acceleration of AI adoption across industries is reshaping the global economy. Our analysis of enterprise implementations reveals that generative AI, in particular, has swiftly moved into practical, revenue-generating applications, signalling the emergence of an AI-centric business environment.

Drawing on real-world observations from our ecosystem of partners and clients, we see AI now permeating daily operations, from personalised customer recommendations to intelligent supply chain optimisation. This widespread integration underscores AI's critical role in enhancing efficiency, fostering innovation, and maintaining competitive advantage.

AI's Transformative Impact Across Industries

AI's influence extends throughout the global economy:

01. Financial Services:
Morgan Stanley is deploying an AI assistant powered by **OpenAI's** GPT-4 to help thousands of wealth managers synthesise vast internal knowledge efficiently, enhancing client advisory services and decision-making processes.

02. Retail:
Stitch Fix utilises generative AI models like **DALL·E** to visualise products based on customer preferences, improving customer experience and driving sales through personalised recommendations.

03. Healthcare:
Companies like **Entos** are pairing generative AI with automated synthetic development tools to design small-molecule therapeutics, potentially revolutionising drug discovery and accelerating time-to-market for new medications.

Just three examples of many demonstrate how AI integration provides significant competitive advantages through improved operational efficiency, accelerated innovation, and enhanced customer experiences.

Drivers of AI Adoption — Why Now?

The surge in AI adoption results from a convergence of factors:

01. Technological Advancements:
The shift to more powerful GPUs, the proliferation of cloud computing, and advancements in machine learning algorithms have significantly enhanced data processing capabilities. GPUs' ability to perform parallel calculations aligns with the requirements of AI workloads, enabling the training of complex models at scale. Market projections underscore this transformation, with GPU deployments expected to surge from 5M units in 2024 to 25M units by 2030, representing a fundamental shift in enterprise computing infrastructure and driving an estimated **$1.2T in total AI infrastructure** spend according to **Coatue**.

02. Economic Incentives:

As noted earlier, substantial economic impact projections motivate enterprises to invest heavily in AI. **McKinsey & Company** estimates that generative AI could add $2.6 trillion to $4.4 trillion annually to the global economy. **IDC** projects global AI spending will reach $632 billion by 2028, with a compound annual growth rate (CAGR) of 29.0% from 2024 to 2028.

03. Data Proliferation:
The exponential growth of data fuels AI systems. **IDC** forecasts global data volumes to reach 291 zettabytes by 2027, growing at an annual rate of 18.5%. This vast amount of data enables enhanced insights, predictions, and outputs from AI models.

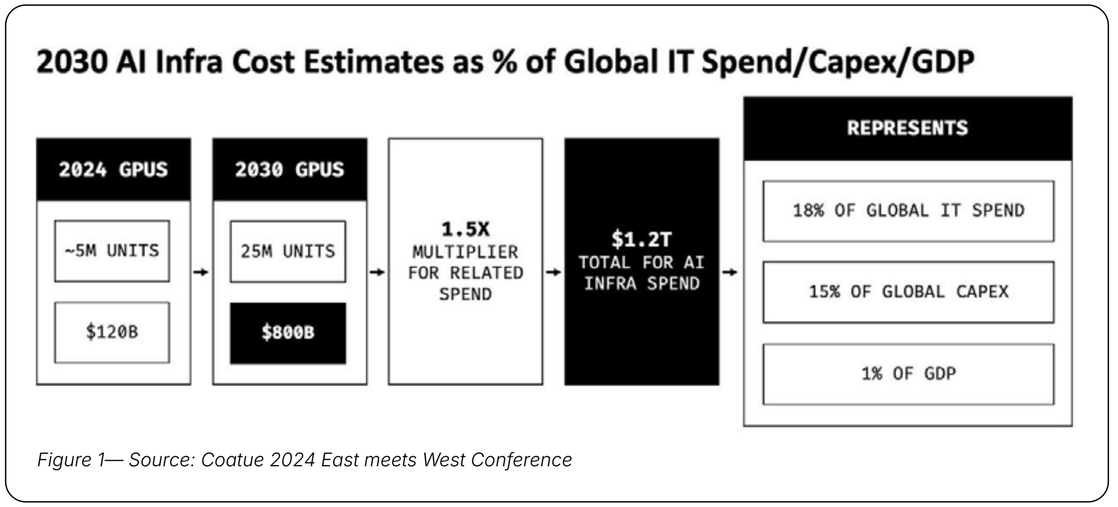


Figure 1— Source: Coatue 2024 East meets West Conference

Introduction — The Emergent AI-Centric World (2/3)

The Infrastructure Imperative

AI's transition from novelty to necessity underscores the importance of robust, AI-ready infrastructure:

01. Computational Demands:
AI workloads require significant computational power. The shift from CPU-centric to GPU-centric computing necessitates substantial investments in new hardware and optimisation of IT strategies. **Nvidia's** H100 GPU, priced at $25,000 per card, illustrates the scale of investment required, with companies like **Meta** in the process of deploying hundreds of thousands.

02. Optimising the IT Stack:
The computational demands of AI require optimisation across the entire computing stack — networking, storage, and software. High-bandwidth, low-latency networks are essential for distributed AI systems, and scalable storage solutions are critical to handle massive data volumes.

03. Scalability and Flexibility:
Enterprises must ensure their infrastructure can scale to accommodate growing AI workloads while maintaining performance and security.

Shifting Business Paradigms and Workforce Dynamics

Implementing AI presents several challenges:

01. Redefining Operations:
AI enables intelligent, adaptive systems that learn and improve over time, redefining productivity metrics and allowing for advanced personalisation in customer interactions.

02. Workforce Transformation:
AI technologies could automate 60–70% of current employee activities. This shift necessitates reskilling and upskilling the workforce. According to the **ServiceNow** "Enterprise AI Maturity Index 2024," organisations are actively acquiring AI-related skills, with plans to hire AI configurators, data scientists, and machine learning engineers.

03. Emergence of AI-Augmented Roles:
New roles that combine human expertise with AI capabilities are emerging, requiring organisations to adapt their talent strategies and foster a culture of continuous learning.

Navigating Implementation Challenges

The surge in AI adoption results from a convergence of factors:

- **Technical Hurdles:** Data quality, model accuracy, and system integration are significant obstacles. Ensuring AI models are trained on high-quality data and integrating them into existing systems requires technical expertise.
- **Ethical Considerations:** AI systems can inadvertently perpetuate biases present in training data. Responsible AI use involves ensuring fairness, transparency, and accountability in decision-making processes.
- **Regulatory Compliance:** Data sovereignty and privacy regulations, such as GDPR and the EU AI Act, create complexity, especially for multinational corporations. Compliance requires robust data governance frameworks.
- **Security Risks:** Protecting AI models and data from malicious activities is essential, as AI systems can be vulnerable to adversarial attacks that compromise integrity and confidentiality.

Industry Examples

Agriculture:
AI is revolutionising agriculture through advanced crop yield prediction. **Xyonix**, an AI consulting firm, combines aerial imagery analysis with custom AI models to assess crop density and environmental impact. This technology aids farmers in decision-making, supports food security policies, and helps seed companies predict plant performance. Such AI applications are crucial for meeting global food demands while minimising environmental impact. Beyond yield prediction, AI in agriculture extends to pest detection, crop optimisation, and automation of farm machinery. As the industry faces challenges like labour shortages and climate change, AI emerges as a key tool for ensuring agricultural sustainability and efficiency.

Energy:
AI is revolutionising energy consumption forecasting in smart grids. **Datategy's papAI** platform demonstrates this transformation. Using historical consumption data from **PJM**, which serves Northeastern US states, the platform employs machine learning to predict energy demand patterns. The process involves data importation, examination, ML pipeline construction, and model evaluation. This AI-driven approach enables energy suppliers to optimise production and distribution, **potentially reducing global energy consumption by 30% by 2040**, according to **IEA** estimates. Such technology enhances grid reliability, enables dynamic demand response, and improves overall energy efficiency.

Introduction — The Emergent AI-Centric World (3/3)

Global Variations in AI Adoption

AI adoption varies globally due to differences in technological infrastructure, regulatory environments, and economic priorities:

- **United States:** Leading in AI investment, **US AI spending is forecasted to reach $336 billion by 2028**, accounting for over half of global AI spending.
- **Europe and the UK:** Emphasis on ethical AI and data protection influences adoption strategies and regulatory compliance requirements.
- **India and Southeast Asia:** Rapid AI adoption, often leapfrogging older technologies, is driving economic growth and addressing societal challenges.

These regional variations have significant implications for global business strategies and competition.

The Future is Here. What Does it Mean?

The AI-centric era demands immediate and strategic action from enterprises. Organisations must adapt to remain competitive in an increasingly AI-driven business landscape. This report will explore:

- **Enterprise Drivers:** from cost efficiency to product innovation, we explore why enterprise is adopting AI
- **The Global AI Landscape:** Analysing AI adoption worldwide, key players, market sizes, and regional differences.
- **Technological Shifts:** Understanding the move from CPU to GPU computing and its implications for infrastructure and operations.
- **Infrastructure Challenges:** Examining obstacles in AI infrastructure development, focusing on networking issues and innovative solutions like AI Availability Zones.
- **Data Growth and AI Workloads:** Assessing the exponential increase in data and optimising systems for AI-specific computational needs.
- **Future Outlook:** Exploring emerging trends, including industry-specific AI chipsets and potential impacts of quantum computing.

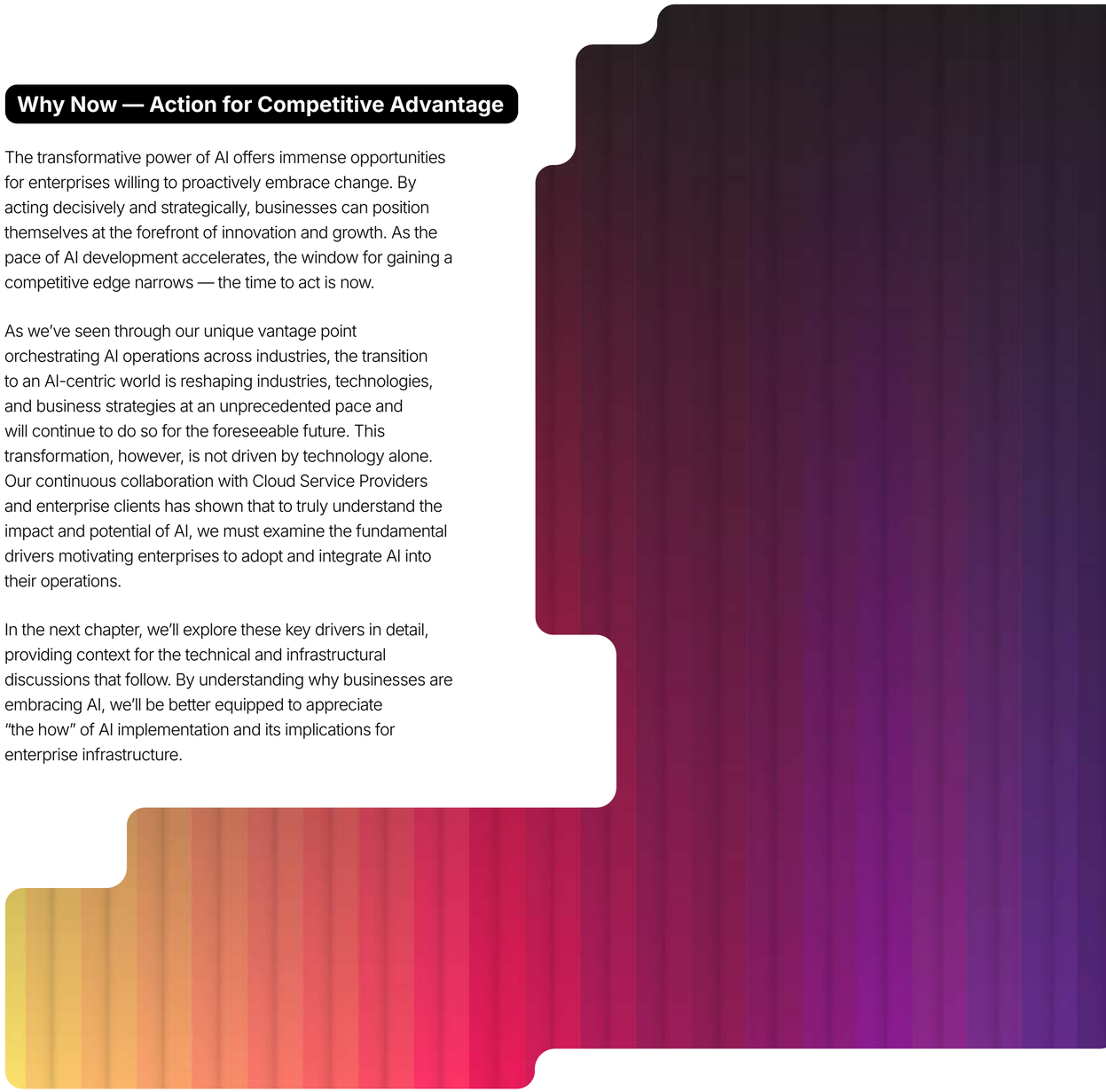
By engaging with these topics, enterprise leaders and AI development teams will be better equipped to navigate the complexities of the Transition to an AI-centric world. Understanding AI infrastructure and the broader AI landscape is crucial for capitalising on opportunities and addressing associated challenges.

Why Now — Action for Competitive Advantage

The transformative power of AI offers immense opportunities for enterprises willing to proactively embrace change. By acting decisively and strategically, businesses can position themselves at the forefront of innovation and growth. As the pace of AI development accelerates, the window for gaining a competitive edge narrows — the time to act is now.

As we've seen through our unique vantage point orchestrating AI operations across industries, the transition to an AI-centric world is reshaping industries, technologies, and business strategies at an unprecedented pace and will continue to do so for the foreseeable future. This transformation, however, is not driven by technology alone. Our continuous collaboration with Cloud Service Providers and enterprise clients has shown that to truly understand the impact and potential of AI, we must examine the fundamental drivers motivating enterprises to adopt and integrate AI into their operations.

In the next chapter, we'll explore these key drivers in detail, providing context for the technical and infrastructural discussions that follow. By understanding why businesses are embracing AI, we'll be better equipped to appreciate "the how" of AI implementation and its implications for enterprise infrastructure.



Chapter 01

Enterprise Drivers For The New AI-Centric World

Chapter 01 — Enterprise Drivers for the New AI-Centric World (1/4)

Introduction

Strategic Decisions Come from Somewhere — the 8 Main Reasons for Enterprise AI

Having established the broad context of the AI-centric world and its transformative impact across industries, we now turn our attention to the specific factors driving AI adoption in enterprises. The urgency of these drivers is underscored by recent market analysis showing that **by 2030, activities accounting for up to 30 percent of hours currently worked across the US economy could be automated** — a trend accelerated by generative AI according to **McKinsey**. Further, McKinsey estimates the top potential revenue impact of generative AI across industry sectors.

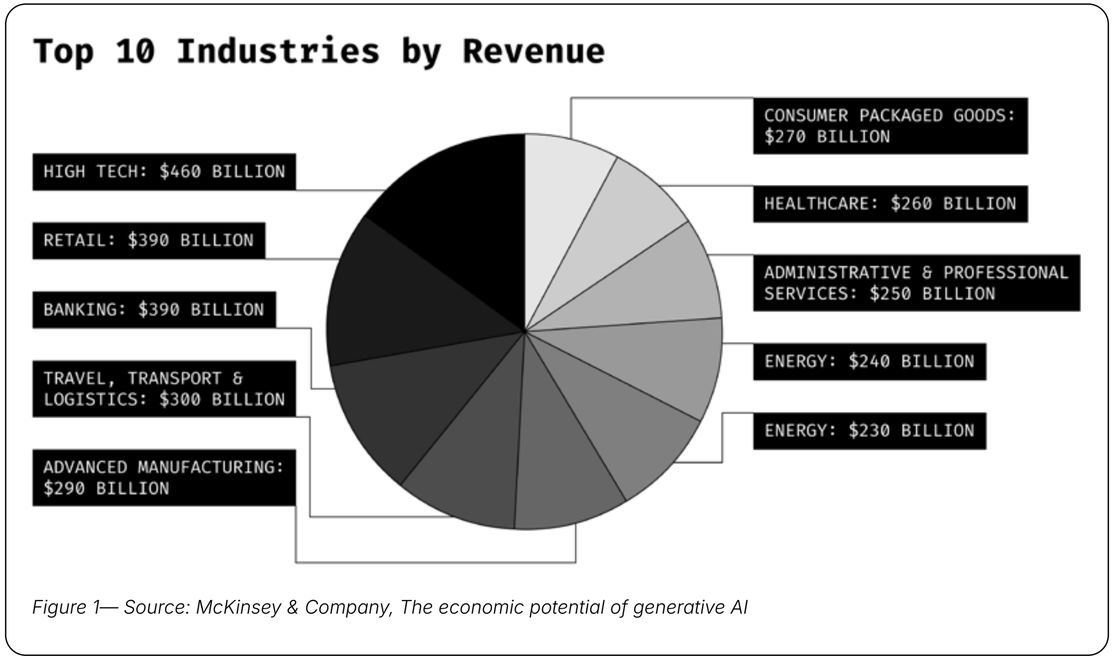


Figure 1— Source: McKinsey & Company, The economic potential of generative AI

These eight main drivers form the foundation of strategic decision-making around AI implementation and directly influence the choices organisations make in building and adapting their AI infrastructure. As we explore each driver, keep in mind how they connect to the broader trends and challenges outlined in the Introduction. Understanding these motivations will provide essential context for the detailed discussions of AI infrastructure, data management, and technological shifts that follow in subsequent chapters.

As artificial intelligence (AI) transitions from an emerging technology to a recognised transformative force, our market analysis team has observed enterprises across industries increasingly recognising its potential to revolutionise their operations and competitive landscape. Working closely with technology vendors and enterprise clients, we've gained unique insight into how organizations are approaching this transformation. This chapter explores the key drivers motivating businesses to adopt AI technologies, providing context for the technical discussions that follow in this report.

Understanding these eight drivers is essential for business leaders and technologists alike. They not only shape the strategic decisions behind AI implementation but also influence the specific infrastructure and operational choices organisations make in their AI journey. As we explore each driver, we'll see how they interconnect and collectively contribute to the AI-centric world.

Chapter 01 — Enterprise Drivers for the New AI-Centric World (2/4)

01. Operational Efficiency and Automation

One of the primary drivers for AI adoption is the promise of enhanced operational efficiency through automation.

01. Process Automation:
AI technologies, particularly Robotic Process Automation (RPA) and intelligent bots, are transforming traditionally labour-intensive tasks. For instance, in customer support, AI-powered chatbots can handle a significant portion of customer inquiries, reducing response times and freeing human agents to focus on more complex issues. **McKinsey** estimates that AI and related technologies have the potential to **automate activities that currently occupy 60–70% of employees' time**.

02. Predictive Maintenance:
In manufacturing and industrial sectors, AI-driven predictive maintenance is revolutionising equipment management. By analysing sensor data and historical performance, AI systems can forecast equipment failures before they occur.
This proactive approach significantly reduces downtime and maintenance costs. For example, a study by Deloitte found that predictive maintenance can **reduce breakdowns by 70% and lower maintenance costs by 25%**.

03. Resource Optimisation:
AI excels at optimising resource allocation across various business functions. In supply chain management, AI algorithms can optimise inventory levels, reducing carrying costs while ensuring product availability. In the retail and hospitality sectors, AI-powered systems optimise staffing schedules based on predicted customer demand, leading to improved service levels and reduced labour costs.

02. Scalability and Flexibility

The ability to scale AI solutions and adapt to changing business needs is a crucial driver for enterprise adoption.

01. Cloud and Hybrid AI Models:
Cloud-based and hybrid infrastructure models offer the scalability and flexibility that modern enterprises require. These models allow organisations to rapidly scale their AI workloads up or down based on demand, without the need for significant upfront investment in hardware. The trade-offs include cloud vendor lock-in and escalating and opaque usage bills.

Enterprises are increasingly adopting hybrid cloud models to manage AI workloads effectively while maintaining flexibility. As noted by Jim Stathopoulos, CIO of **Sun Country Airlines**, the fast-changing landscape of data and AI calls for a hybrid approach to balance experimentation, security, and cost control. Hybrid cloud infrastructures are becoming the preferred choice to ensure scalability while protecting sensitive data and managing resource allocation efficiently.

02. Modular AI Solutions:
Enterprises are increasingly drawn to modular AI solutions that can be customised and expanded over time. This approach allows businesses to start with specific use cases and gradually expand their AI capabilities as needs evolve and expertise grows. Modular solutions also facilitate easier integration with existing systems and processes, reducing disruption during implementation.

In the context of AI, modular hybrid models offer the best of both worlds — flexibility in scaling workloads on the public cloud while securely handling sensitive data in private cloud environments. According to **IDC**, hybrid cloud infrastructure is expected to dominate enterprise decision-making for AI workloads as businesses seek to optimize both scalability and security.

03. Cost Efficiency and Total Cost of Ownership (TCO)

While AI implementation can require significant investment, the long-term cost benefits are a major driver for adoption.

01. Reduction in Labor Costs:
By automating routine tasks, AI helps reduce reliance on manual labour for low-value activities. A report by Accenture suggests that **AI could increase labour productivity by up to 40% by 2035**, enabling people to make more efficient use of their time.

02. TCO Optimisation:
Enterprises are increasingly focused on optimising the total cost of ownership for their AI initiatives. This involves considering not just the initial implementation costs, but also ongoing operational expenses. Solutions that leverage optimised hardware (such as **NVIDIA** GPUs) and strategic partnerships (e.g., with storage solution providers like **VAST**) can help enterprises achieve better TCO. For instance, NVIDIA reports that its DGX systems can reduce AI training time by up to 60%, translating to significant cost savings in GPU usage and energy consumption.

Additionally, as AI workloads are resource-intensive, many companies are rethinking cloud strategies to balance cost predictability with performance. Enterprises are increasingly adopting hybrid cloud models to control escalating costs. According to **IDC, spending on private, dedicated cloud services is expected to reach $20.4 billion during 2024 and more than double by 2027**, driven in part by the need to manage AI-related expenses effectively.

03. AI-as-a-Service Models:
The rise of AI-as-a-Service offerings is attracting enterprises looking to minimise upfront costs and operational burdens. These managed services, such as those offered in **Deloitte's** "Silicon to Services" package, provide AI capabilities with predictable, subscription-based pricing. This model is particularly appealing to organisations that lack the in-house expertise to develop and maintain complex AI systems.

Chapter 01 — Enterprise Drivers for the New AI-Centric World (3/4)

04. Data-Driven Decision-Making and Real-Time Insights

In today's data-rich business environment, the ability to derive actionable insights quickly and accurately is a significant driver for AI adoption.

01. Enhanced Business Intelligence (BI):
AI systems enable real-time data processing and analytics, providing businesses with actionable insights that were previously unattainable. In the financial sector, AI-powered BI tools can analyse market trends, customer behaviour, and risk factors in real-time, enabling more informed investment decisions. **Morgan Stanley's** AI assistant, which helps wealth managers quickly synthesise vast amounts of internal knowledge, exemplifies this trend.

02. Personalisation and Customer Engagement:
AI's ability to process and analyse large volumes of customer data in real-time enables unprecedented levels of personalisation. In the retail sector, companies like **Stitch Fix** use AI to provide personalised product recommendations, significantly enhancing customer engagement and driving sales. The **global AI in retail market is expected to grow from $5 billion in 2021 to $31.2 billion by 2028, according to Grand View Research**.

03. Retrieval-Augmented Generation (RAG):
Enterprises are increasingly seeking real-time RAG capabilities, which enhance traditional AI models by integrating live, context-specific data. This ensures more accurate and up-to-date insights, particularly crucial in dynamic business environments where conditions change rapidly.

However, many enterprises are now adopting private clouds to ensure that real-time AI data is secure. **Somerset Capital Group**, for example, has shifted its critical AI workloads to a private cloud, ensuring sensitive data is not exposed to public cloud infrastructures where it might inadvertently be integrated into external models. This shift in cloud strategy highlights the growing concern over AI data security in real-time applications.

05. Compliance, Security, and Data Sovereignty

As regulatory requirements around data privacy and security become more stringent globally, compliance has become a critical driver for AI adoption strategies.

01. Navigating Regulatory Requirements:
AI solutions that help enterprises adhere to complex regulations such as GDPR in Europe or HIPAA in U.S. healthcare are in high demand. These solutions must ensure data protection while still enabling the powerful analytics capabilities that AI offers.

02. Data Sovereignty and Locality:
With data sovereignty laws becoming more prevalent, enterprises require AI infrastructure that supports hybrid deployment models. These allow organisations to keep data within national borders, addressing regulatory requirements and avoiding compliance issues. According to **Gartner, by 2024, 35% of publicly listed companies will have hybrid and multi-cloud deployments** to address data residency and compliance requirements.

03. AI Governance and Transparency:
There's a growing need for AI models that are not only effective but also explainable and transparent. This is crucial for meeting ethical standards and regulatory compliance requirements, particularly in sectors like finance and healthcare where decision-making processes must be auditable.

06. Competitive Advantage and Market Differentiation

AI is increasingly seen as a strategic tool for creating significant competitive advantage, enabling enterprises to innovate, differentiate their offerings, and respond rapidly to market changes.

01. Product and Service Innovation:
AI enables businesses to develop new products and services faster. In the automotive industry, AI is driving innovation in autonomous vehicles. **Tesla**, for instance, uses AI for its Autopilot system, giving it a significant edge in the race towards fully autonomous driving.

02. Operational Agility:
AI solutions that offer real-time decision-making capabilities allow companies to react swiftly to market changes, customer needs, or operational issues. This agility is particularly valuable in fast-moving sectors like e-commerce and financial trading.

03. Customer Experience Differentiation:
By leveraging AI for personalisation and predictive analytics, companies can provide superior customer experiences. In the banking sector, AI-powered chatbots and virtual assistants are becoming commonplace, offering 24/7 customer support and personalised financial advice.

Chapter 01 — Enterprise Drivers for the New AI-Centric World (4/4)

07. Strategic Partnerships and Ecosystem Integration

The complexity of AI implementation often requires collaboration with technology providers and specialists, driving enterprises to seek strategic partnerships.

01. Technology Collaboration:
Partnerships with technology leaders like **NVIDIA** for hardware or **VAST** for optimised storage solutions or Stelia for data mobility provide enterprises with the technical resources needed to scale AI efficiently. These collaborations can significantly reduce the time and risk associated with AI implementation.

02. Managed Service Offerings:
Collaborations with consulting firms like Deloitte that provide managed AI and infrastructure services allow enterprises to leverage expertise without having to build in-house capabilities. This reduces implementation complexity and accelerates time to value.

03. Ecosystem Integration:
Enterprises increasingly value AI solutions that can connect simply with existing software, cloud platforms, and hardware infrastructure. This integration ensures minimal disruption and smoother deployment, a crucial factor for businesses looking to maintain operational continuity while adopting AI.

08. Innovation and Long-Term Digital Transformation

AI implementation is often a core component of broader digital transformation strategies aimed at positioning enterprises for long-term success.

01. R&D and AI Incubation:
Many enterprises are investing in AI not just for immediate gains but also for long-term innovation. AI incubators and accelerators provide companies with environments to experiment with and develop new AI models and applications safely. For example, **Volkswagen's** AI labs focus on developing AI solutions for various aspects of automotive manufacturing and autonomous driving.

02. Sustainability and Green AI Initiatives:
As organisations focus on sustainability, AI solutions that optimise energy usage or enhance efficiency in supply chains and production processes are becoming increasingly attractive. According to a **PWC** report, **AI could help reduce global greenhouse gas emissions by 4% by 2030**.

03. Future-Proofing Operations:
Beyond immediate needs, enterprises are driven by the need to future-proof their operations. Scalable AI infrastructure that can adapt to future technological advancements or business expansions is crucial for ensuring that today's AI investments remain valuable in the long term.

Challenges and Considerations

While the drivers for AI adoption are compelling, enterprises must also navigate several challenges:

- **Talent Shortage:** The demand for AI skills often outstrips supply. According to a 2021 survey by **O'Reilly, 52% of organisations cited a lack of skilled people as a significant barrier to AI adoption**.
- **Data Quality and Availability:** AI models are only as good as the data they're trained on. Ensuring high-quality, unbiased data sets can be a significant challenge.
- **Integration with Legacy Systems:** Many enterprises struggle to integrate AI solutions with existing legacy IT infrastructure.
- **Ethical Considerations:** As AI becomes more prevalent in decision-making processes, ensuring fairness and avoiding bias become critical considerations.

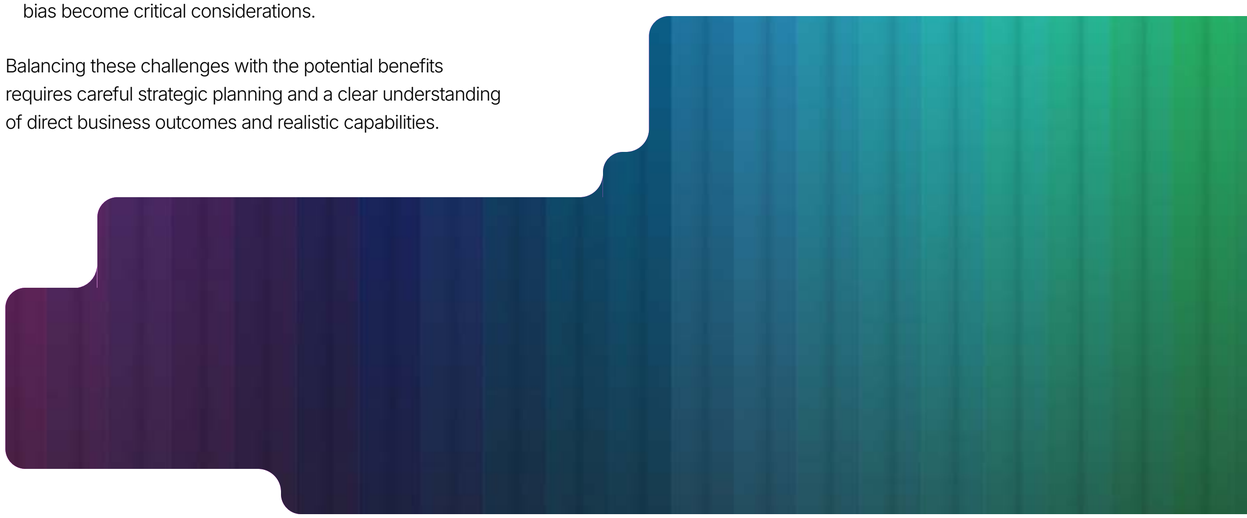
Balancing these challenges with the potential benefits requires careful strategic planning and a clear understanding of direct business outcomes and realistic capabilities.

Linking IT Spend with Business Outcomes

The drivers motivating enterprises to adopt AI are diverse and interconnected, ranging from operational efficiency and cost savings to strategic innovation and competitive differentiation.

The choices made in terms of hardware, software, and networking solutions should ultimately serve broader business objectives. By aligning technical decisions with strategic drivers, enterprises can maximise the value of their AI investments and position themselves for success in an increasingly AI-driven business landscape.

The next chapter will explore the global AI landscape, providing context on how these drivers are shaping AI adoption across different regions and industries.



Chapter 02

The Global AI Landscape

Chapter 02 — The Global AI Landscape (1/6)

Introduction

In the previous chapter, we explored the drivers compelling enterprises to adopt AI as a transformative force in their operations. But while the "why" of AI adoption is increasingly clear, the "where" and "how" remain shaped by regional dynamics, funding models, and infrastructure readiness. AI is not evolving uniformly; it is a mosaic of global investments, regulatory approaches, and technical capabilities.

While **North America** continues to lead the AI race, accounting for **43% of global AI investments in 2024**, its share will decline to **36.5% by 2030** as **Asia-Pacific** — spearheaded by China — takes the lead. China alone is expected to command **two-thirds of Asia-Pacific's AI software revenue**, reflecting its aggressive investments in industrial AI, smart cities, and generative AI. Meanwhile, **Europe** is shaping the conversation around ethical AI, with initiatives like the EU AI Act and projections of a **€1.4 trillion GDP impact by 2030** from generative AI applications.

By 2030, the global AI market is projected to grow from **$214 billion in 2024** to an astounding **$1.34–1.81 trillion**, representing a **CAGR of 36%**. For business leaders, this rapid growth signifies both an opportunity and a mandate: those who harness AI's transformative potential will thrive; those who lag risk irrelevance.

The Building Blocks of the AI Economy

The AI market is broadly segmented into three pillars: **hardware (infrastructure)**, **software**, and **services**, each playing a distinct role in driving enterprise adoption:

01. Infrastructure (Hardware):
Hardware investments — led by GPUs, CPUs, and emerging on-device AI capabilities — will reach **$80 billion by 2024**, with a growing focus on edge AI for real-time applications. Innovations like Qualcomm's Snapdragon processors and Intel's Core Ultra chips are enabling AI to run locally, offering businesses greater efficiency and personalization.

02. Software:
AI software dominates, forecasted to grow from **$98 billion in 2024 to $391 billion by 2030**. Generative AI is the fastest-growing subset, expanding at an unprecedented **CAGR of 49.7%**. It's revolutionizing industries with applications in marketing, retail, healthcare, and finance, delivering both cost savings and new revenue streams.

03. Services:
Services are the connective tissue, facilitating enterprise adoption through consulting, managed services, and integration. These are vital for businesses navigating AI's complexity and addressing talent shortages, ensuring scaled deployments and faster ROI.

Generative AI: The Crown Jewel

Generative AI is at the center of this transformation, forecasted to contribute **$434 billion annually** to enterprise value creation by 2030. Its applications are profound:

- **Retail and e-commerce:** By 2030, generative AI will account for **33% of enterprise use cases**, enabling hyper-personalized shopping experiences, visual search tools, and automated content creation.
- **Finance:** Generative AI is enhancing decision-making through predictive analytics and automating tasks like fraud detection and client interactions, driving **20% of enterprise AI value creation**.
- **Healthcare:** AI-powered solutions are revolutionizing drug discovery, diagnostics, and patient care, making healthcare the fastest-growing vertical in the AI ecosystem.

Economic Impact: Beyond Business as Usual

The economic implications are staggering. Generative AI alone could add **$2.6 to $4.4 trillion annually** to the global economy by 2030. This represents a seismic shift, with AI automating **60–70% of repetitive tasks**, freeing up human talent for higher-order problem-solving. For enterprises, this means enhanced productivity, faster time-to-value, and the ability to scale innovation across functions.

Chapter 02 — The Global AI Landscape (2/6)

What This Means for the C-Suite

The global AI market marks a fundamental shift in how businesses operate, compete, and grow. For leaders, the roadmap is clear:

- **Invest Strategically:** Prioritize AI solutions that align with core business goals, whether it's scaling generative AI for marketing or leveraging predictive AI for operational efficiency.
- **Embrace Regional Dynamics:** Understand where the innovation is happening — North America for cutting-edge startups, China for industrial AI scale, and Europe for ethical governance — and build partnerships accordingly.
- **Prepare for Complexity:** Services will remain essential for navigating AI adoption, addressing talent shortages, and ensuring compliance with emerging regulations.
- **Scale Thoughtfully:** With generative AI poised to disrupt every industry, organizations must focus on scalable, cloud-based solutions while exploring emerging edge AI opportunities for real-time decision-making.

Let's dive into the regional specifics.

Regional Adoption Trends

United States: The Powerhouse of AI Innovation

The United States remains the undisputed leader in global AI innovation, powered by record-breaking private-sector investments, cutting-edge technologies, and a robust ecosystem of startups and research institutions. Over the last five years, **venture capital investments in AI have totalled $290 billion**, fuelling advancements in sectors like autonomous vehicles, healthcare, and IT infrastructure.

Key Initiatives and Investments: Federal programs, such as the **National AI Research Resource (NAIRR) and Department of Defense's $1.5 billion AI budget**, demonstrate the government's commitment to maintaining AI leadership. Meanwhile, tech giants like Microsoft, OpenAI, and Nvidia continue to set benchmarks in generative AI, healthcare AI applications, and GPU-powered advancements.

Economic Potential: Projections suggest AI could contribute between **$1.2 and $3.8 trillion annually** to U.S. GDP over the next decade, underscoring its transformative impact on the national economy.

Sectoral Innovation: The U.S. leads in healthcare AI, with FDA approvals for diagnostic tools like early detection systems for diabetic retinopathy. In autonomous vehicles, initiatives like the **AV START Act** foster investor confidence and rapid progress.

Despite its dominance, the U.S. faces growing scrutiny over data privacy and generative AI's societal impact. Nevertheless, its innovation-driven ecosystem ensures it remains at the forefront of global AI leadership.

China: From Strategic Vision to Global Leadership

China's AI ecosystem is rapidly transforming into a global powerhouse, driven by unparalleled government investments and a thriving domestic market. With **$184 billion invested across 9,600 AI firms** from 2000 to 2023, China's commitment to fostering innovation extends to underdeveloped regions, supported by over **20,000 financial transactions**.

Strategic Government Backing: Central to this growth is China's **"New Generation Artificial Intelligence Development Plan"**, which aims to establish China as the world's leading AI innovation hub by 2030. This plan aligns with the country's broader modernization strategies, such as "Made in China 2025."

Generative AI Excellence: Recent breakthroughs have positioned Chinese generative AI models to rival — and in some cases surpass — those from the United States, with **117 generative AI products** approved by March 2024.

Practical Applications: Unlike the U.S., China focuses on immediate, practical AI uses:

- **Smart Cities:** AI enhances urban planning and infrastructure in megacities like Beijing and Shenzhen.
- **Retail and Healthcare:** AI enables personalized shopping, predictive analytics, and diagnostic tools for underserved rural areas.

Challenges and Opportunities: U.S. sanctions limit access to high-performance semiconductors, but China's focus on AI infrastructure — servers, thermal control, and data technologies — ensures continued progress.

As China narrows the gap with global leaders, its combination of strategic vision, government backing, and consumer-focused innovation positions it to shape the future of AI globally.

Chapter 02 — The Global AI Landscape (3/6)

Europe: Ethical Leadership with Economic Ambitions

Europe's approach to AI prioritizes ethical, human-centric systems, balancing innovation with governance. With the **EU AI Act** poised to take full effect by 2026, Europe intends to set a global benchmark for transparent, trustworthy AI deployment. However, the region's AI ambitions are not limited to governance — projections indicate that generative AI could add between **€1.2 trillion and €1.4 trillion** to the EU's GDP over the next decade, boosting economic output by 8%.

Regulatory Leadership: The **EU AI Act** categorizes AI technologies by risk, establishing guidelines to ensure fairness and accountability. This governance framework fosters trust and confidence in AI solutions across industries.

Public Funding Initiatives:

- **Horizon Europe** has allocated **€2.6 billion for AI research** and plans further investment to support large AI models.
- **Digital Europe Programme** dedicates over **€1 billion** annually to AI deployment and ecosystem growth, with €4 billion set aside for generative AI by 2027.

Sectoral Impact: Europe excels in industries like healthcare and green technology, integrating AI to improve diagnostics, sustainability, and efficiency. However, slower infrastructure development and fragmented markets challenge its competitiveness against global leaders.

Europe's focus on ethical AI and transformative economic growth positions it as a unique counterbalance to the innovation-first strategies of the U.S. and China.

United Kingdom: A Thriving AI Ecosystem

The United Kingdom has established itself as Europe's AI leader, with the largest AI market on the continent, valued at **£72.3 billion ($92 billion)**. This makes the UK **fourth globally**, behind only the U.S., China, and Israel. With record-breaking investments and a thriving innovation ecosystem, the UK is driving transformative advancements across industries.

Private-Sector Momentum: The UK raised **$2.1 billion in AI startup funding in H1 2024**, on track to reach **$4.4 billion by year-end**. Notable investments include:

- **Wayve's $1.05 billion** for autonomous vehicles.
- **CoreWeave's £1 billion London expansion**, advancing data center infrastructure.
- **ServiceNow's £1.15 billion** commitment to UK operations.

Government Support:

Data centers are now classified as **Critical National Infrastructure (CNI)**, ensuring greater stability for AI-driven operations.

The **AI Safety Institute** focuses on responsible AI development.

With globally renowned universities like **Cambridge and Oxford** producing leading AI spinouts, the UK bridges cutting-edge innovation with regulatory oversight. Its thriving AI ecosystem ensures it remains a global leader in healthcare, transportation, and FinTech applications.

Middle East: Infrastructure-Driven AI Growth

The Middle East is rapidly positioning itself as a future AI powerhouse, with AI expected to contribute **$320 billion to GDP by 2030**. Anchored by bold visions like **Saudi Arabia's Vision 2030** and the U**AE's AI Strategy 2031**, the region is transforming its economy through large-scale investments in AI infrastructure and smart technologies.

Saudi Arabia's Investments:

- **Project Transcendence:** Backed by the Public Investment Fund (PIF), this $100 billion initiative is establishing AI hubs and startups.
- **NEOM Smart City:** A $500 million project integrating AI, IoT, and 5G to support urban transformation.
- Data center expansion includes AWS ($5.3 billion), Oracle ($1.5 billion), and **Center3's $1 billion** connecting Asia, Europe, and Africa.

UAE's Leadership:

- The UAE's **AI Strategy 2031** focuses on governance, smart cities, and public-private partnerships.
- Dubai appointed **22 chief AI officers** to implement AI in government services.

Despite challenges like talent shortages, the Middle East's commitment to infrastructure and innovation establishes it as a critical global AI player. -

India and Southeast Asia: Emerging AI Powerhouses

India and Southeast Asia are on the cusp of transformative AI growth, projected to improve Southeast Asia's GDP by **13–18% by 2030**, equivalent to **nearly $1 trillion**. With **$30 billion in AI infrastructure investment** in 2024 alone, the region is scaling rapidly across key industries.

Key Drivers of Growth:

- **Southeast Asia:** Countries like **Malaysia ($15 billion)**, **Singapore ($9 billion)**, and **Thailand ($6 billion)** lead the charge in building AI-ready infrastructure.
- **India:** The **National AI Strategy** supports initiatives like 'Making AI in India,' driving enterprise adoption and workforce upskilling.

High Adoption Rates: Developing economies in the region report **30% higher AI adoption** rates than developed counterparts, driven by a young, tech-savvy population.

Sectoral Focus:

- **Fintech:** AI is transforming fraud detection and personalized financial services.
- **Healthcare:** Predictive analytics and diagnostic tools improve access and efficiency.
- **Retail and E-commerce:** AI enhances customer experiences through personalization.

With proactive government policies, private-sector investments, and a digitally engaged workforce, India and Southeast Asia are poised to become global AI powerhouses.

Chapter 02 — The Global AI Landscape (4/6)

Challenges in Regulating AI: Balancing Innovation and Accountability

Regulating artificial intelligence (AI) presents a host of challenges for sovereign states, driven by the rapid pace of technological advancement and the complexities of governance. Key hurdles include:

01. Pace of Technological Change:
AI evolves faster than legislation can keep up, leaving policymakers struggling to draft regulations that remain relevant and effective. This lag often allows AI technologies to outpace oversight, creating gaps in accountability.

02. Fragmented Regulatory Landscape:
The absence of comprehensive international guidelines has led to a patchwork of laws eg the EU AI Act, complicating compliance for businesses operating across jurisdictions. Companies face increased costs and confusion as they navigate inconsistent requirements.

03. Algorithmic Accountability:
Ensuring transparency and preventing discrimination in AI systems is a significant challenge. While some countries have begun to address algorithmic bias, establishing clear standards for accountability remains a complex task requiring collaboration with industry and civil society.

04. Knowledge Gaps Among Lawmakers:
Many policymakers lack the technical understanding of AI's capabilities and risks, hindering their ability to create informed and effective regulations. Public education on AI remains similarly limited, further complicating stakeholder engagement.

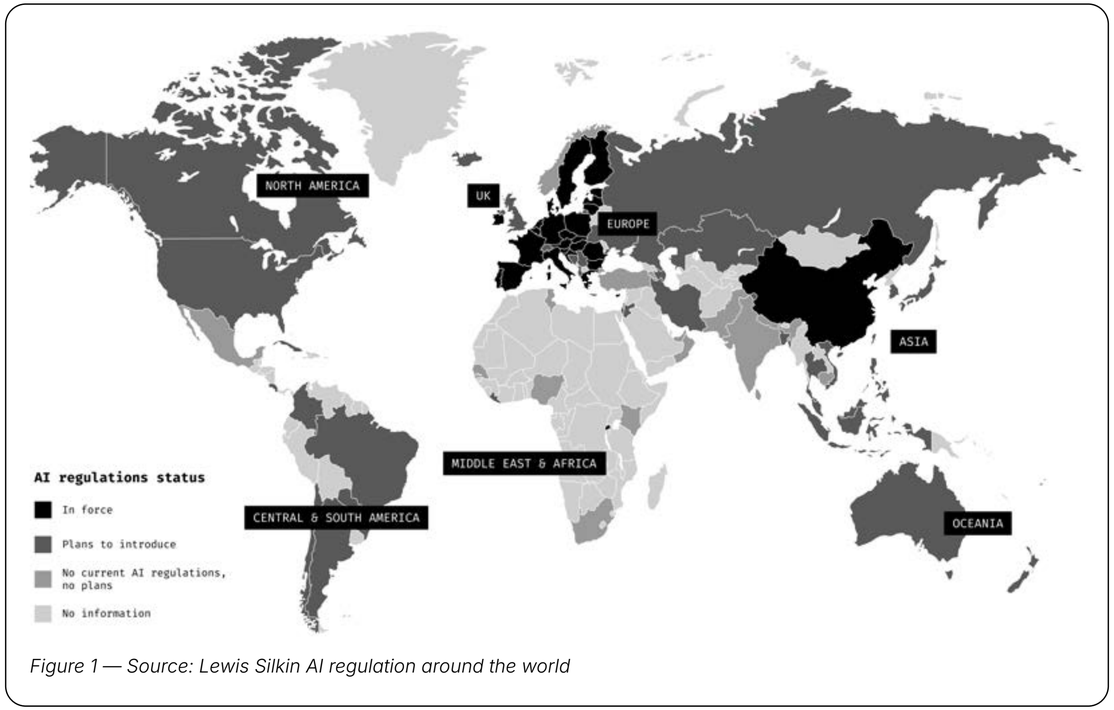


Figure 1 — Source: Lewis Silkin AI regulation around the world

Chapter 02 — The Global AI Landscape (5/6)

Global Cooperation in AI: A Unified Approach to Innovation and Safety

Addressing shared challenges such as ethical AI development, safety, and cross-border innovation has a cross-borders dimension. Recent international initiatives highlight the growing sovereign commitment to collaborative AI governance:

Key Developments

01. The Bletchley Declaration:
Emerging from the **Bletchley Park AI Safety Summit (November 2023)**, this declaration underscores the importance of international collaboration on AI safety. Signatory nations committed to developing common standards for risk mitigation and responsible AI deployment.

02. The AI Seoul Summit:
In **May 2024**, the **Seoul Statement of Intent** was signed by ten countries and the EU, leading to the establishment of an international network of **AI Safety Institutes (AISIs)** across the UK, US, Japan, Singapore, and others. These institutes aim to foster information sharing, align regulatory frameworks, and build AI safety capacity.

03. Bilateral and Multilateral Agreements:

- The UK-US collaboration on AI safety research focuses on the development of interoperable safety protocols.
- The EU-US Trade and Technology Council partnership strengthens cooperation on ethical AI practices and advanced research.

04. Global Forums:
Organizations such as the **G7**, **OECD**, and **United Nations** are central to shaping global AI governance. These forums emphasize multi-stakeholder engagement and the creation of interoperable frameworks to ensure inclusive and ethical AI development.

Currently, international AI agreements lack strong legal teeth, operating more as frameworks for cooperation and shared principles. However, they play a critical role in shaping norms, aligning nations, and fostering trust, which may lead to more binding legal agreements or enforcement mechanisms in the future. For now, their impact depends on the willingness of individual nations and organizations to act on their commitments.

Big Tech's lobbying efforts are heavily influencing the EU AI Act, aiming to minimize regulations on foundational AI models despite public commitments to ethical AI. With 86% of AI-related meetings at the European Commission involving corporate representatives, industry giants like Google and Microsoft argue that strict rules could harm innovation and competitiveness. However, critics warn that this influence undermines critical safeguards for transparency, accountability, and public trust, raising questions about whether Big Tech has become too powerful to regulate effectively.

Case Study: Nordic-Baltic Region: Ethical and Sustainable AI Leadership

The Nordic and Baltic countries are emerging as a unique collaborative hub for artificial intelligence, leveraging their shared values of sustainability, integration, and ethical governance. In **August 2024**, the **Nordic and Baltic Ministers of Digitalisation** met in Copenhagen to unveil the **Nordic AI Vision for 2030**, a roadmap to position the region as a global leader in large-scale AI adoption.

Key Initiatives

01. Nordic Center for Applied AI:
A proposed hub to unify national AI networks, boost cross-border collaboration, and optimize regional investments. This center aims to drive practical AI applications and improve competitiveness while ensuring responsible use.

02. Green AI Infrastructure:
The region is prioritizing sustainable AI solutions, including data centers powered by renewable energy and designed for efficiency. These initiatives align with the goal of making the Nordic-Baltic region the most sustainable in the world by 2030.

03. AI Upskilling Hub:
A new initiative will analyse workforce demands and promote digital upskilling to address labour shortages and ensure the region has a diverse, AI-ready talent pool.

Scaling and Commercialization

To overcome barriers in scaling AI from pilots to production, the region plans to establish a **council for commercialization**, ensuring promising projects reach the market. Additionally, the region is pooling resources to explore **quantum AI**, targeting breakthroughs in life sciences, materials, and energy.

Ethical Leadership

Grounded in Nordic values, the planned **Nordic Center for Responsible AI** will focus on fairness, transparency, and trust, setting global standards for ethical AI governance.

Future Outlook

By integrating ethics, sustainability, and innovation, the Nordic-Baltic region is positioning itself as a global AI leader. With its bold vision and actionable initiatives, it's creating a model for how AI can benefit society while respecting the planet.

Chapter 02 — The Global AI Landscape (6/6)

Emerging Trends and Applications

Emerging technologies like generative AI and agentic AI are transforming industries, but their growing computational demands highlight the need for quantum computing breakthroughs. At the same time, ethical and sustainable AI initiatives are becoming indispensable to ensure these technologies align with societal values.

Stelia, through its deep engagement with enterprise AI adoption, has identified agentic AI as a cornerstone of future operations. Its ecosystem partnerships enable enterprises to scale agentic workflows and streamline operational efficiency across sectors.

01. Generative AI and Multimodal Models
Generative AI is revolutionizing industries, with applications extending beyond text into video, biology, and genomics. Its market is expected to grow at a **CAGR of 59.2%**, reaching **$202 billion** by 2028. Industries such as entertainment, healthcare, and financial services are leading adopters.

02. Agentic AI
Agentic AI is transforming workflows across industries, enabling enterprises to autonomously execute complex tasks while optimizing decision-making. However, its rise places unprecedented demands on networks, necessitating AI-ready architectures to sustain scalability and responsiveness. According to Gartner, by **2026**, over **80% of enterprises** will adopt AI Agents and Agentic Workflows for data management and operational efficiency. By **2025**, agentic AI is expected to become a cornerstone of routine, repetitive, and resource-heavy processes, allowing businesses to refocus human talent on strategic and creative tasks.

03. Quantum Computing
While still nascent, quantum computing holds transformative potential for AI. Early applications focus on optimization and cryptography, with long-term implications for advancing AI capabilities.

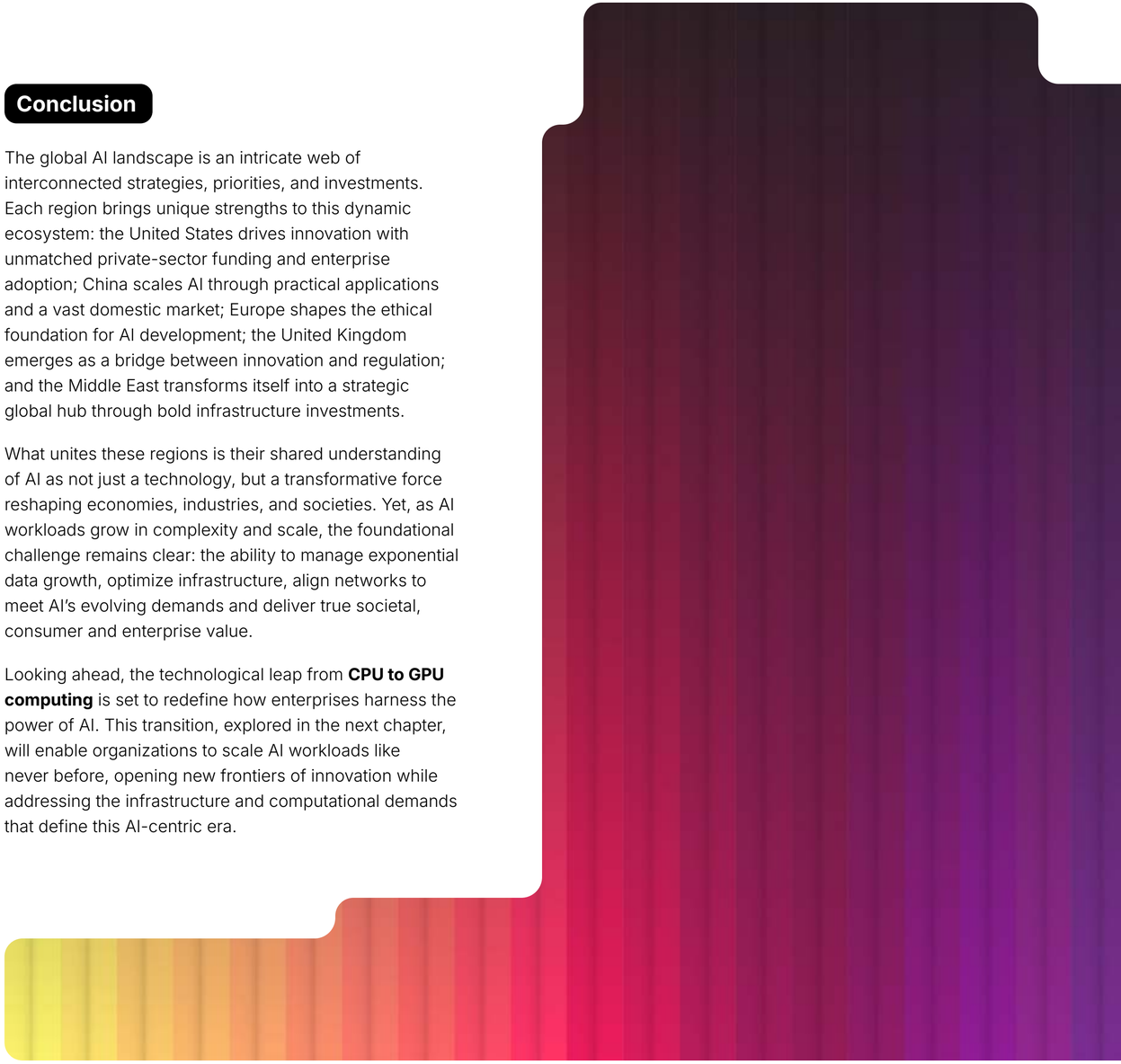
04. Ethical and Sustainable AI
As AI systems become integral to critical infrastructure, ethical considerations are gaining prominence. The **EU AI Act** and similar regulations aim to ensure fairness and transparency and some enterprises are investing in "green AI" to reduce the environmental impact of energy-intensive AI workloads.

Conclusion

The global AI landscape is an intricate web of interconnected strategies, priorities, and investments. Each region brings unique strengths to this dynamic ecosystem: the United States drives innovation with unmatched private-sector funding and enterprise adoption; China scales AI through practical applications and a vast domestic market; Europe shapes the ethical foundation for AI development; the United Kingdom emerges as a bridge between innovation and regulation; and the Middle East transforms itself into a strategic global hub through bold infrastructure investments.

What unites these regions is their shared understanding of AI as not just a technology, but a transformative force reshaping economies, industries, and societies. Yet, as AI workloads grow in complexity and scale, the foundational challenge remains clear: the ability to manage exponential data growth, optimize infrastructure, align networks to meet AI's evolving demands and deliver true societal, consumer and enterprise value.

Looking ahead, the technological leap from **CPU to GPU computing** is set to redefine how enterprises harness the power of AI. This transition, explored in the next chapter, will enable organizations to scale AI workloads like never before, opening new frontiers of innovation while addressing the infrastructure and computational demands that define this AI-centric era.



Chapter 03

Technological Shift: CPU to GPU

Chapter 03 — Technological Shift: CPU to GPU (1/8)

Introduction

The shift from CPU- to GPU-centric computing represents a fundamental reorientation of the global tech landscape, driven by the exponential growth of artificial intelligence (AI) and machine learning (ML). As AI adoption accelerates across industries, the demand for parallel computing has surged, surpassing the limits of traditional CPU architectures.

With global AI spending projected to reach **$632 billion by 2028** (CAGR of 29%) and **McKinsey** estimating AI's economic potential at **$17–25 trillion annually**, the need for powerful, scalable computing infrastructures has never been greater. However, Moore's Law—the doubling of CPU performance every two years—has slowed, prompting enterprises to embrace GPU-based systems. GPUs, with their unparalleled parallel processing capabilities, have become essential to advancing AI training and inference.

This chapter examines the limitations of CPUs, the rise of GPUs, and the implications of this shift for enterprises. Drawing on insights from **Stelia's** GPU Market Tracker, **EpochAI** estimates, and other industry forecasts, we explore how GPU-centric computing is transforming enterprise infrastructure in the AI era.

The Decline of CPU-Centric Computing

Central Processing Units (CPUs) have powered computing for decades, excelling at serial processing. However, Moore's Law, which predicted a doubling of transistor density every two years, is reaching its limits due to challenges like quantum tunnelling, heat dissipation, and rising fabrication costs. As a result, CPU performance gains have slowed, making them less suitable for modern computational demands.

However, CPUs continue to play an indispensable role in modern computing. Rather than being entirely supplanted, CPUs are increasingly paired with GPUs in heterogeneous computing architectures, enabling optimal performance across diverse workloads.

Why CPUs Fall Short for AI

AI and machine learning workloads require immense computational power, particularly for tasks like deep learning and real-time processing. CPUs struggle with these demands due to:

- **Limited Parallelism:** CPUs have fewer cores optimized for serial tasks, making them inefficient for parallel operations central to AI workloads.
- **Time-Intensive Training:** Training AI models on CPUs takes significantly longer, slowing innovation cycles.
- **Energy Inefficiency:** CPUs consume more energy per computation in parallel tasks, driving up costs and environmental impact.

Emerging Trends in CPU-GPU Collaboration

01. Heterogeneous Architectures:
Innovations such as Intel's Xeon CPUs with integrated AI accelerators and ARM's scalable architectures demonstrate how CPUs are adapting to complement GPU-driven workloads. These developments facilitate tighter integration, reducing latency and enhancing overall system performance.

02. Unified Memory Architectures:
The rise of unified memory spaces, where CPUs and GPUs share access to a common memory pool, is simplifying programming models and eliminating bottlenecks in data movement. Technologies like NVIDIA's Unified Memory and AMD's Infinity Architecture exemplify this trend.

03. Specialized CPUs for AI:
Some companies are developing CPUs specifically tailored for AI. For example, Apple's M-series chips integrate both CPU and GPU capabilities with a shared neural engine, enabling efficient on-device AI processing for consumer applications.

Chapter 03 — Technological Shift: CPU to GPU (2/8)

A Symbiotic Future

The relationship between CPUs and GPUs is not one of replacement but of synergy. As AI and machine learning workloads continue to evolve, hybrid architectures leveraging the strengths of both processors are emerging as the gold standard for scalable and efficient computing. Enterprises should focus on building balanced infrastructures that combine the agility of GPUs with the versatility of CPUs to future-proof their systems.

Why GPUs Outperform CPUs

Initially designed for rendering graphics, GPUs have emerged as the backbone of AI infrastructure due to their efficiency in parallel computation. The demand for parallel compute, as offered by GPU-centric computing architecture, is evidenced by the vast increase in machine learning (ML) systems across diverse domains. This growth has been particularly pronounced since the inflection point of widespread AI adoption in recent years.

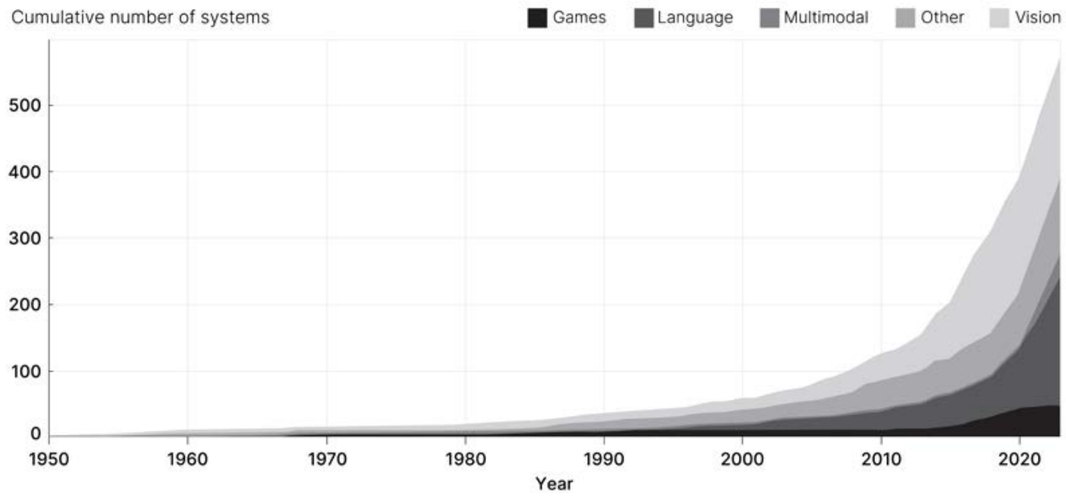


Figure 1 – Source: Jaime Sevilla et al. (2024), "Can AI Scaling Continue Through 2030?". Published online at Epoch AI . Retrieved from: 'Can AI Scaling Continue Through 2030? ' [online resource]

Expanding AI Adoption Across Domains:

- The proliferation of AI applications spans areas such as natural language processing, computer vision, gaming, and multimodal AI systems, all of which thrive on the capabilities of GPU-centric infrastructures.
- This shift represents a paradigmatic transition from the CPU, with its serial processing capabilities that fuelled the technological revolution of the mid-20th century, to GPU architectures optimized for massive parallelism.

Why GPUs Outperform CPUs:

- Massive Parallelism: GPUs feature thousands of smaller cores designed for simultaneous execution of multiple tasks.
- High Throughput: GPUs are optimized for handling large-scale parallel operations, making them ideal for AI and ML workloads.
- Energy Efficiency: GPUs deliver higher FLOPS per watt, reducing energy costs and improving sustainability.

As industries increasingly adopt GPU-centric architectures for inferencing and model training, the transition enables faster, more efficient processing. This, in turn, empowers applications across domains, driving innovation and the adoption of transformative AI capabilities. The widespread integration of AI infrastructure reflects a broader trend toward harnessing GPU power to enable this rapid expansion.

Chapter 03 — Technological Shift: CPU to GPU (3/8)

Market Momentum and Leadership

The growing demand for GPUs is reshaping the semiconductor market:

- **Market Growth:** The global semiconductor market is projected to **grow at a 25% CAGR through 2028**, with GPUs expected to account for nearly half of the total market (**Dell'Oro Group**).
- **Inflection Point:** 2023 marked a turning point, as revenues from accelerators like GPUs surpassed those from CPUs, signalling a fundamental market shift.
- **NVIDIA's Dominance:** NVIDIA has driven this transition with its CUDA platform and a **developer ecosystem exceeding 2 million members**. Over **500 million CUDA-enabled GPUs** are already deployed globally, further cementing its leadership in AI infrastructure.
- Since 2007 NVIDIA's market cap has grown twice as fast as its semiconductors have shrunk in size. Coincidence, causal, or complicated?

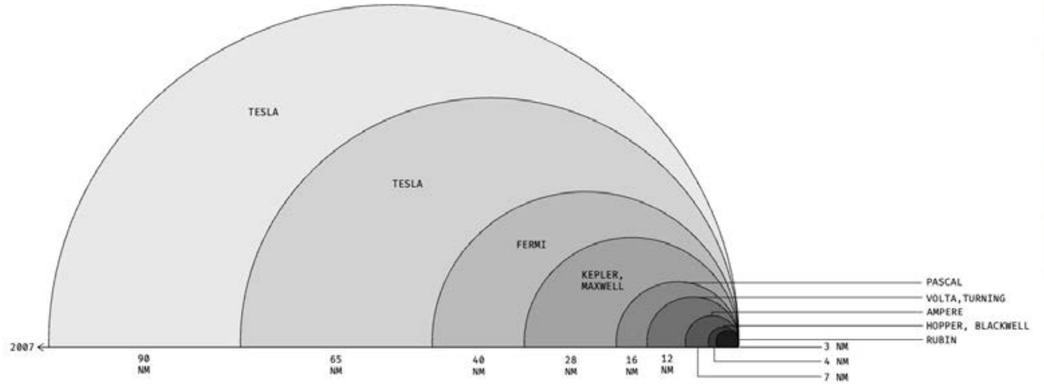


Figure 1A: Scaled visual representation of the progression from Tesla to Rubin GPU architectures, showcasing the advancements in semiconductor technology from 90nm in 2007 to 3nm nodes currently. For reference a human hair is approx. 100,000nm

Drivers of the GPU-Centric Shift

Escalating Demand for AI Compute

The rapid growth of AI and ML applications is fuelling demand for parallel computing:

- **Generative AI Boom:** Generative AI spending is projected to grow at a **59.2% CAGR**, reaching **$202 billion by 2028** and comprising 32% of total AI spending (**IDC**).
- **Rising Training Needs: AI compute scales at 4x per year**, with projections suggesting up to 2e29 FLOP demands at least 20M H100-equivalent GPUs by 2030.
- **Enterprise Adoption:** Leading companies ("pacesetters") are adopting AI at scale, with **33% driving innovation** compared to just 14% of their peers (**ServiceNow**).

Data Explosion and Processing Challenges

Exponential data growth is straining traditional computing systems:

- **Massive Data Volumes:** Global data is expected to reach **291 zettabytes by 2027** (2.7x growth rate), while AI workloads will consume storage at unprecedented scales, which will be discussed in more detail in the forthcoming chapter on Data Growth.
- **AI in Networks:** By 2030, **75% of all network traffic** will involve AI-driven content. The network is already identified as an Enterprise chokepoint with innovative companies such as **Stelia** suggesting a foundational rethink of Internet architecture is essential.
- **Leveraging Untapped Data:** Expanding multimodal data sources could enable a **10,000x increase** in training capacity.

Chapter 03 — Technological Shift: CPU to GPU (4/8)

Market Dynamics and Key Players

Enterprise IT Transformation

To meet AI's demands, enterprises are transforming their infrastructure:

- **Upgraded Hardware:** GPU-centric systems are replacing traditional CPU-based architectures to handle advanced AI workloads. GPU compute is supplanting traditional compute nodes.
- **Enhanced Networks:** Distributed GPU clusters require high-bandwidth, low-latency networks, necessitating significant upgrades at least and architectural changes for optimal futureproofing.
- **Gigawatt Data Centers:** AI is driving the rise of massive data centers, consuming 2x to 10x more power for large-scale training runs. This will be elaborated on in the next chapter – AI Infrastructure Challenges. See also *Constraints and Challenges in Scaling GPU Compute* later in this chapter.

GPU Market Growth

The GPU market is experiencing unprecedented growth, driven by surging demand for AI hardware:

- **NVIDIA's Leadership:** NVIDIA reported **$35.1 billion in Q3 2024** revenue (+94% YoY), with $30.8 billion **(88%) from data center AI workloads**. Its shipments, including 650,000 H100 GPUs in 2023, reflect its market dominance.
- **Scaling GPU Production:** Global production is projected to expand by 30–100% annually, with **estimates ranging from 20 million to 400 million H100-equivalent GPUs by 2030**. This capacity would enable training runs equivalent to between 4651 and 232,558 GPT4 training runs.

Key Players Beyond NVIDIA

While NVIDIA dominates, other manufacturers are shaping the market:

- **AMD:** Focused on AI applications, **AMD projects 70% growth in data center GPUs**.
- **Intel:** Diversifying into GPUs and AI-integrated chips.
- **Apple** and **ARM:** Apple integrates GPUs into custom silicon for optimized performance, while ARM enables GPU solutions for mobile and embedded systems.

Emerging Competitors in AI Hardware

While NVIDIA dominates the AI chip market, innovative startups are developing specialized architectures that challenge traditional GPU designs. These new entrants aim to address specific limitations in GPUs, offering solutions optimized for AI workloads like natural language processing and computer vision.

Silicon Startups:

- **Cerebras Systems:** Known for its Wafer-Scale Engine (WSE), the world's largest AI chip, featuring 4 trillion transistors and 900,000 cores. Cerebras' single-chip design reduces latency, making it ideal for large-scale neural networks. Recent partnerships include building AI supercomputers with **G42**, an Emirati artificial intelligence (AI) development holding company based in Abu Dhabi, founded in 2018.
- **Groq:** Focused on AI inference, Groq's Tensor Streaming Processor (TSP) emphasizes low latency and high throughput. Its streamlined software stack accelerates development, targeting sectors like finance and healthcare.
- **SambaNova Systems:** Offers an integrated hardware-software platform with its Reconfigurable Dataflow Unit (RDU). SambaNova provides turnkey solutions, enabling enterprises to deploy AI quickly and efficiently.
- **Graphcore:** Developer of the Intelligence Processing Unit (IPU), optimized for machine intelligence with fine-grained parallelism and high memory bandwidth. Its Poplar SDK simplifies integration with AI frameworks.
- **d-Matrix:** Focused on AI inference, D-Matrix design chips with digital 'in-memory compute', unlocking efficiency gains for AI workloads. Designed for AI end-user service delivery with high-volume throughput, through applications like chatbots and video generation.

Chapter 03 — Technological Shift: CPU to GPU (5/8)

Market Implications

The rise of specialized AI chips has several implications:

- **Increased Competition:** New players drive innovation and could reduce costs over time.
- **Supply Chain Diversification:** Enterprises benefit from alternative suppliers, reducing dependency on NVIDIA.
- **Performance Advantages:** Specialized chips may outperform GPUs in specific tasks, providing faster training and lower latency.

Considerations for Enterprises

- **Opportunities:** Startups offer customizable, efficient solutions tailored to enterprise needs, potentially reducing costs and deployment timelines.
- **Risks:** Startups may lack the stability or scalability of established providers, posing risks to supply reliability and support.

Constraints and Challenges in Scaling GPU Compute

Scaling GPU compute to meet the explosive demands of AI training presents formidable challenges despite rapid technological advancements. Key bottlenecks include manufacturing capacity for advanced GPUs, the availability of high-bandwidth memory (HBM), and limitations in power infrastructure. As AI systems scale at a historical rate of **4x compute per year**, these constraints are tightening.

Meeting the demands for GPUs and compute by 2030 will require significant expansions in advanced packaging, semiconductor manufacturing, and power infrastructure, with investments potentially exceeding **$100 billion** for large-scale facilities like **Microsoft's** "Stargate."

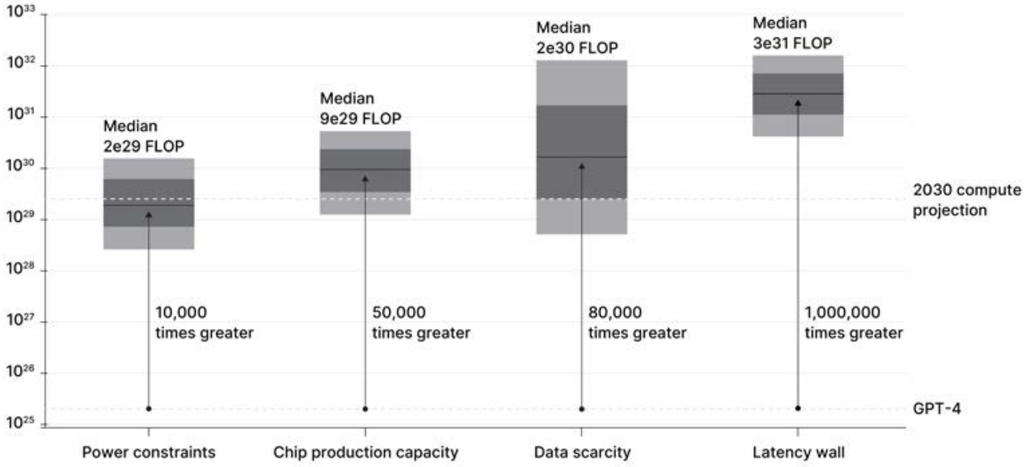


Figure 2 – Source: Jaime Sevilla et al. (2024), "Can AI Scaling Continue Through 2030?". Published online at Epoch AI . Retrieved from: 'Can AI Scaling Continue Through 2030? ' [online resource]

Key Insights

01. Manufacturing Capacity:
The production of GPUs, particularly Nvidia's H100 equivalents, faces challenges due to limited CoWoS packaging and high-bandwidth memory (HBM) availability. Projections estimate **100 Million H100-equivalents will be needed by 2030**, yet current global capacity is significantly lower. Efforts to scale GPU production depend on rapid investment in fabs and high-efficiency packaging technologies.

02. Energy Demands:
Training runs in 2030 could require 6 GW, equivalent to the power needs of a small country. This will be discussed in more details in the following chapter – AI Infrastructure Challenges. With facilities like **Microsoft's "Stargate"** aiming for gigawatt-scale campuses, distributed training networks and co-located power plants are critical to meeting energy requirements. Energy costs could constitute **40% of GPU infrastructure costs by 2030**.

03. Data Scarcity:
Training massive AI models requires enormous datasets, but the availability of high-quality text data may plateau in the next five years. Multimodal data (e.g., image, video, and audio) and synthetic data generation offer promising avenues for scaling datasets.

Chapter 03 — Technological Shift: CPU to GPU (5/8)

04. Latency Wall:
Training larger models encounters a fundamental bottleneck in processing time, as sequential operations grow linearly with model size. **By 2030, multimodal and synthetic datasets could provide up to 20 quadrillion tokens**, supporting training runs as large as 2e32 FLOP, or **80,000x larger than Chat GPT-4**. Without innovations in network topologies and larger batch scaling, the latency wall could limit training runs to 1e32 FLOP unless resolved through hardware improvements and communication protocols.

GPU Production Capacity Projections

- **Scaling Trends:** GPU production is projected to grow between 30% and 100% annually, potentially enabling the manufacture of up to 100 million H100-equivalent GPUs by 2030.
- **Constraints:** Advanced packaging (CoWoS capacity) and HBM production are key bottlenecks, posing risks to sustained production growth.
- **Production Scenarios:**
  - **Low-End Estimate:** 20 million H100-equivalent GPUs, supporting 1e29 FLOPs (5,000x GPT-4's compute scale).
  - **High-End Estimate:** 400 million H100-equivalent GPUs, enabling 5e30 FLOPs (250,000x GPT-4's compute scale).

AI Training Compute Scaling

- **Exponential Growth:** Continuation of the current **4x annual scaling** trend projects training runs up to **2e29 FLOPs by 2030**.
- **Economic Justification:**
  - **Massive Returns:** AI automation could unlock $60 trillion in global economic value annually, justifying investments of $1–2 trillion in AI infrastructure.
- **Industry Milestones:**
  - **Microsoft and OpenAI's 'Stargate':** $100 billion investment in a state-of-the-art data center, launching in 2028.
  - **GPT-5 Projections:** Expected to generate $20 billion in its first year of deployment.
  - **Oracle's Zettascale Initiative:** 131,000 Blackwell GPUs power zettascale AI clusters.
  - **Future of Computing:** Coatue's Philippe Laffont predicts a **$10–20 trillion shift** from CPU- to GPU-based architectures, rebuilding global computing systems.

Despite these constraints, the trajectory for scaling GPU compute remains positive, driven by aggressive investment and the industry's capacity to innovate.

However, packaging and memory bottlenecks, alongside unprecedented energy demands, represent significant risks to the pace of progress. Addressing these challenges will require unprecedented coordination between GPU manufacturers, utility providers, and governments to ensure the necessary infrastructure is in place.

The economic incentives are substantial, as AI capabilities continue to expand into nearly every industry, making these investments essential to sustaining global technological leadership.

EpochAI's predictions highlight rapid advancements in AI, presenting enterprises with significant opportunities and challenges. To capitalize on these developments, organizations must invest in GPU-centric infrastructure to handle advanced AI workloads, ensuring scalability and readiness for future demands. Adopting larger, more capable AI models can drive innovation, enhance products and services, and provide a competitive edge in the market.

Developing AI talent is crucial amid the scarcity of skilled professionals in this field. Enterprises should recruit top talent and upskill existing employees to build internal expertise. Navigating operational challenges—such as increased energy requirements and supply chain risks—requires implementing energy-efficient practices, diversifying suppliers, and investing in robust data strategies.

Maximizing ROI from AI initiatives involves prioritizing projects that offer significant returns through automation and efficiency gains. Establishing AI governance frameworks ensures responsible deployment, while staying informed about evolving regulations maintains compliance and public trust. Integrating AI into long-term strategic planning and proactively adapting to technological trends will enable organizations to remain competitive in an AI-driven landscape.

Chapter 03 — Technological Shift: CPU to GPU (6/8)

Stelia's GPU Market Tracker Insights

Stelia GPU Market Tracker: Comprehensive Insights into Global GPU Landscape

Stelia's GPU Market Tracker combines public data, proprietary insights, and enterprise collaborations to deliver a granular understanding of the global GPU ecosystem. Below, we provide an organized narrative that explains the insights derived from this data, divided into logical sections:

01. Overview of Global GPU Volume by GPU Make and Deployment Location

Estimated volume of GPUs (millions) by manufacturer and location.

	Region				
Make	**North America**	**Europe**	**Asia**	**TOTAL**	**Type**
Google	5.010			**5.010**	TPU
NVIDIA	3.189	0.227	0.860	**3.285**	GPU
Amazon	0.480			**0.480**	GPU
Intel	0.065	0.001	0.064	**0.129**	GPU
Huawei			0.100	**0.100**	GPU
AMD	0.111	0.014	0.001	**0.090**	GPU
Matrix			0.036	**0.036**	GPU
Grand Total	**8.86**	**0.24**	**1.06**	**10.16**	

Estimated GPU Volume (in Millions)

This section details the estimated volume of GPU units across all GPU models by manufacturer (GPU Make). The analysis includes deployment data categorized by key regions: North America, Europe, and Asia.

Methodology:

We ingest public data, signals, and proprietary information through the commercial position of Stelia in the AI networking space to build a picture of GPU deployments. These are tied to known Data Center locations, GPU Make & Model, and other key dimensions to estimate GPU supply.

In cases where dimensions like geolocation are unknown, we revert to the key dimensions associated with the owning entity, until more specific information can be attributed. For example, an AWS GPU deployment will be placed in North America by default until verified information on the specific Data Center location is available.

Future, planned deployments post-Q4 2024 are not reflected in this overview. Future published versions of the Stelia GPU Tracker will showcase deployments from 2025 onward.

Regional Breakdown:

The insights provide clarity on the global distribution of GPUs, essential for strategic AI planning and investment.

Chapter 03 — Technological Shift: CPU to GPU (7/8)

02. Breakdown by Owning Entity Type

Estimated volume of GPUs (millions) by owner type and location.

	Region			
Company Type	**North America**	**Europe**	**Asia**	**TOTAL**
Hyperscaler	6.645	0.000	0.000	**6.645**
Private Cloud	1.442	0.156	0.740	**2.329**
Public Cloud	0.471	0.000	0.105	**0.576**
National HPC	0.272	0.086	0.107	**0.465**
AI-application (Large Scale)	0.000	0.000	0.110	**0.110**
Non-Profit AI Cloud	0.025	0.000	0.000	**0.025**
Grand Total	**8.86**	**0.24**	**1.06**	**10.17**

Estimated GPU Volume (in Millions)

This section examines GPU deployments organized by key owning entity types. The analysis highlights how different organizational categories are utilizing GPU resources, split across key regions: North America, Europe, and Asia.

- **Data Sources:** Leveraging its expertise in AI networking, Stelia gathers and analyzes signals from both public and proprietary channels.
- **Methodology:** GPU volumes are categorized by entity type, with deployments cross-referenced against Data Center locations or owning entity geolocations.
- **Entity Types:**

  01. Hyperscalers: Cloud giants such as Google, Amazon, Microsoft, and ByteDance.

  02. Private Cloud: Includes enterprise private clouds, Venture Capital-backed AI clusters, and Hyperscaler private cloud environments.

  03. Public Cloud: Providers of GPU-as-a-service and Hyperscale public cloud offerings.

  04. National HPC: High-performance computing initiatives, academia, and research projects.

  05. AI-Application (Large Scale): Large-scale enterprise AI deployments (e.g., Tencent, ByteDance).

  06: Non-Profit AI Cloud: GPU clusters owned by non-profit entities.

GPU Distribution Across Entities

01. Dominance of Major Tech Companies:

- **Meta Platforms:** Acquired 25% of NVIDIA's H100 shipments in 2023,
- **Other Leaders:** Microsoft, Google, and Amazon are rapidly expanding their GPU resources for AI development.

02. Key Trends:

- Resource Consolidation: A few enterprises control a significant share of global GPUs, underlining competitive concentration.
- Strategic Investments: Companies are prioritizing GPU infrastructure to maintain leadership in AI innovation.

Chapter 03 — Technological Shift: CPU to GPU (8/8)

Regional GPU Deployment

01. Geographic Breakdown:

- North America:
  - Leadership Position: Home to dominant tech firms and innovative startups.
  - Economic Scale: U.S. AI spending projected to hit $336 billion by 2028, representing over half of global AI investments.
- Asia-Pacific:
  - Growth Hotspots: Accelerated GPU adoption in China and South Korea fuelled by government incentives and emerging AI sectors.
  - GDP Impact: China's AI adoption could increase its GDP by 26% by 2030.
- Europe:
  - Steady Progress: Sustained focus on AI research and infrastructure expansion across the continent.

02. Economic Insights:

- Global Impact: North America and China collectively are set to contribute $10.7 trillion—70% of AI's global economic benefits by 2030.

Conclusion

Summarizing the $1T Shift from CPU to GPU

The transition from CPU-centric to GPU-centric computing is a fundamental shift driven by the imperatives of an AI-centric world. The limitations of traditional CPU architectures, particularly in handling the parallel processing demands of modern AI workloads, have necessitated this change. GPUs, with their superior parallel processing capabilities and energy efficiency, are now very much on the innovation frontlines.

Key Takeaways:

- **Accelerating GPU Adoption:** Enterprises are rapidly adopting GPUs to meet the computational demands of AI and ML applications, with GPU shipments and production capacity scaling rapidly.
- **Market Expansion:** The GPU market is experiencing unprecedented growth, with significant investments in production capacity and R&D from companies like NVIDIA and AMD.
- **Scaling Challenges:** Power constraints, chip production capacity, data scarcity, and the latency wall present challenges that require strategic solutions and innovation.
- **Economic Impact:** Significant investments in GPU infrastructure are economically justified given the potential to capture substantial value from global labour compensation and to drive economic growth.

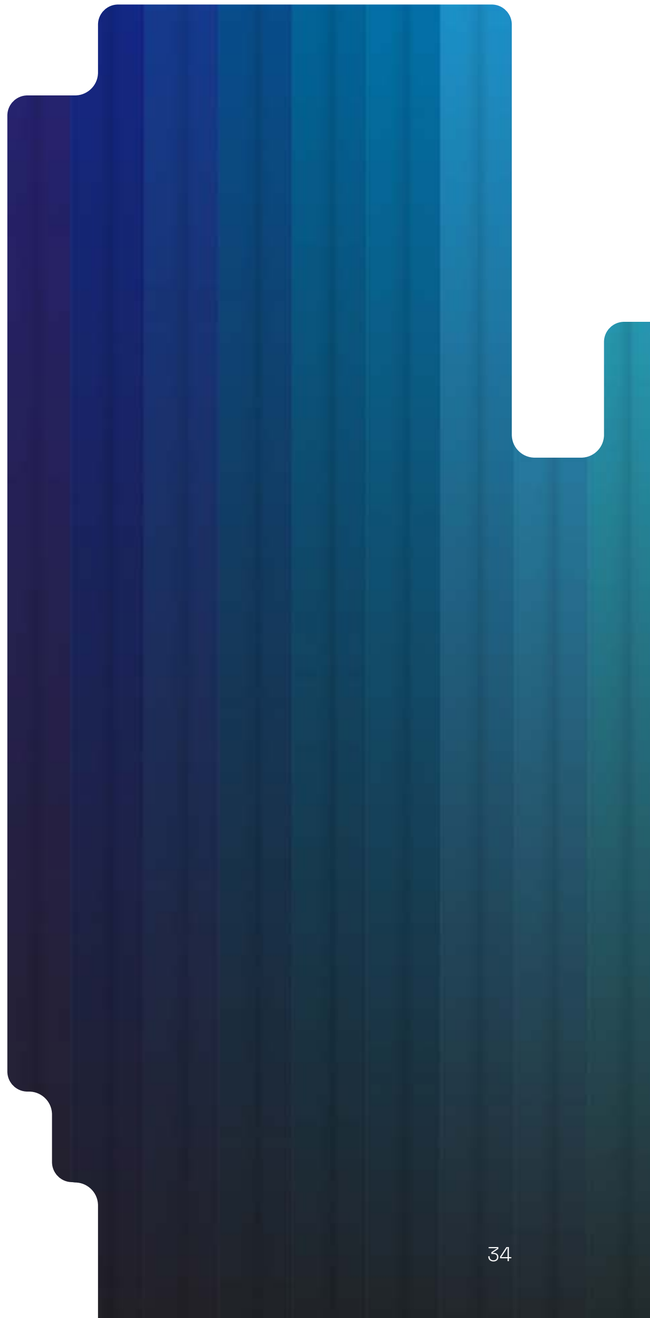
Implications for Enterprises

For enterprise leaders, this shift has profound implications:

- **Strategic Investment:** Organizations must invest in GPU-centric infrastructure to remain competitive in the AI-driven market.
- **Infrastructure Planning:** Upgrading data centers, network infrastructures, and energy management systems is crucial to support large-scale AI workloads.
- **Talent Development:** Building expertise in AI, ML, and GPU programming is essential to leverage the full potential of these technologies. There are approximately **30 million developers globally**, with **300,000 ML engineers** and **30,000 ML researchers**. Enterprises need to attract and develop talent in this competitive landscape.
- **Collaboration and Ecosystem Development:** Partnerships with technology vendors, cloud service providers, and network operators can accelerate AI adoption and infrastructure development.

Looking Ahead

As we move forward, the technological shift to GPU-centric computing sets the stage for addressing AI infrastructure challenges (discussed in the following chapter) and navigating the future landscape of AI and data growth. Enterprises that proactively adapt to this change will be best positioned to harness the transformative power of AI, driving innovation and economic growth in the AI-centric era.



Chapter 04

AI Infrastructure Challenges

Chapter 04 — AI Infrastructure Challenges (1/4)

The Infrastructure Imperative: Scaling AI for the Enterprise

This chapter builds on the preceding sections of the report, which examine **how enterprises can win with AI** by leveraging cutting-edge models, optimising AI-driven decision-making, and navigating the evolving competitive landscape. Earlier chapters have explored the rise of foundation models, AI's role in automation and efficiency, and the economic implications of widespread AI adoption. However, none of these advancements can reach their full potential without a robust AI infrastructure capable of supporting large-scale deployment, inference, and operational resilience. This section specifically focuses on the technological, logistical, regulatory, and energy-related challenges enterprises face when scaling AI infrastructure and provides actionable recommendations for overcoming them.

Artificial intelligence is no longer an experimental capability; it is the strategic foundation upon which modern enterprises are built. AI-driven transformation is permeating every sector, from financial services and healthcare to supply chain logistics and manufacturing. Yet, despite its immense promise, enterprises face significant challenges in deploying scalable, cost-effective, and energy-efficient AI infrastructure. The rapid expansion of AI workloads has outpaced traditional IT architectures, pushing enterprises to rethink compute strategies, networking capabilities, data governance, and—critically—power consumption and grid availability.

As organisations scale AI beyond proof-of-concept deployments, they encounter mounting hurdles: the computational demands of deep learning models, the complexities of distributed AI networking, and the looming constraints of power grids and energy transmission bottlenecks. The declining cost of GPU-hour pricing presents a paradox—while it democratises access to AI computing, it also signals potential risks, such as hardware stagnation, inefficiencies in AI models, and reduced quality of service. Meanwhile, a new reality is emerging: AI is no longer constrained by compute or algorithms but by infrastructure bottlenecks, with energy availability now emerging as the next critical limiting factor.

The explosive demand for AI compute is now colliding with fundamental energy and infrastructure constraints. **McKinsey** estimates that AI-driven power demand could surge by 240GW by 2030—equivalent to adding up to six new UKs worth of power consumption. The **RAND** report projects that AI data centers will require 327GW of power by 2030, a 460% increase over 2022's global data center capacity. The U.S. alone may need 51GW of additional AI data center capacity by 2027, yet permitting delays, transmission constraints, and regulatory uncertainty pose major challenges to scaling infrastructure at the necessary pace. I**f AI data center build-out in the U.S. and other leading markets cannot meet demand, enterprises will be forced to relocate infrastructure abroad, potentially undermining national AI competitiveness and data security.**

Figure 1.1. Estimates of Data Center Power Capacity Required to Host All AI Chips, 2024–2030

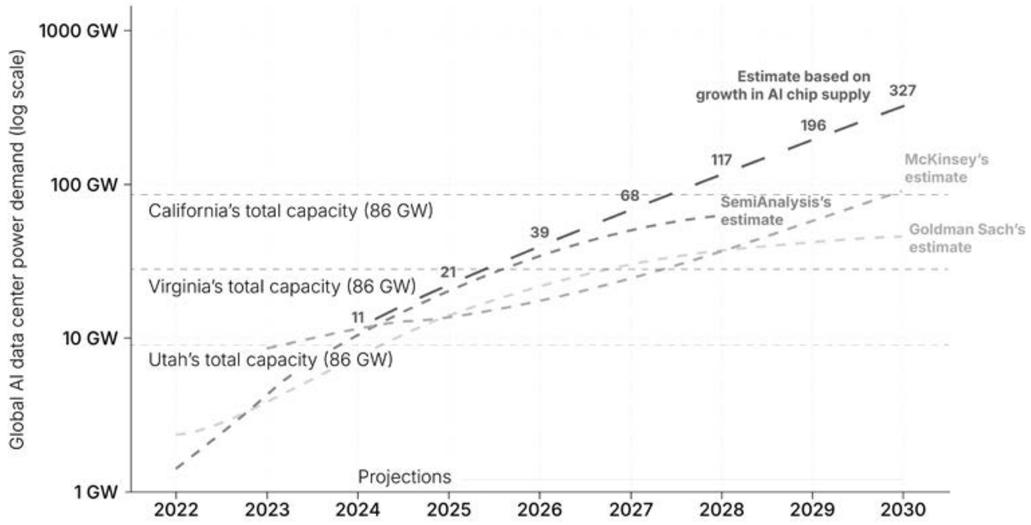


Figure 1.1 Estimates of data center power capacity required to host all AI Chips, 2024-2030.
Source: RAND - 'AI's Power Requirements Under Exponential Growth', accessed January 28th, 2025.

Chapter 04 — AI Infrastructure Challenges (2/4)

The Role of Hardware, Data Center Automation & Energy Strategy in AI Scaling

As AI models grow in complexity and enterprise adoption accelerates, hardware-optimised AI code generation, data center automation, and energy procurement strategies will play a crucial role in sustaining performance and efficiency at scale.

Cloud providers and enterprises running on-premise AI must rethink how they optimise their IT infrastructure for AI-driven workloads. Traditional data center models, built for general-purpose computing, are ill-suited for the extreme demands of modern AI inference and training. This shift raises key strategic questions for AI infrastructure technologists:

- How can enterprises leverage hardware automation to enhance AI performance while reducing operational complexity?
- Will AI-native data centers become the new standard, replacing traditional IT architectures?
- What role will AI play in optimising its own infrastructure, from automated workload allocation to self-repairing data centers?
- How will AI's soaring energy demands reshape corporate energy procurement strategies?
- Will enterprises be forced to prioritise geographic locations based on energy availability rather than traditional tech hub advantages?

Organisations that proactively invest in AI hardware, automation, and efficiency optimisation will gain a competitive advantage, ensuring they remain ahead in the AI race. Companies like Google and Microsoft are already integrating self-optimising AI infrastructure, leveraging machine learning to predict server loads, dynamically adjust power consumption, and mitigate latency in real time. Enterprises that fail to embrace this transformation risk falling behind in AI scalability, cost efficiency, and reliability.

By integrating hardware-accelerated AI processing, automated workload balancing, and predictive infrastructure management, businesses can future-proof their AI investments and maintain sustainable, high-performance AI operations. The next decade will likely see a convergence of AI infrastructure, energy strategy, and automation, with AI not only running on optimised hardware but actively shaping its own infrastructure through intelligent resource allocation and self-tuning architectures.

Defining Key Infrastructure Terms

Before getting further into infrastructure challenges, it's useful to distinguish between commonly used terms in AI infrastructure and deployment.

- **Distributed:** Refers to architectures where compute workloads are spread across multiple nodes or data centers, often working in parallel. Distributed AI enables large-scale model training by breaking tasks into smaller components processed simultaneously across different locations.
- **Decentralised:** A system where decision-making and data processing are spread across multiple entities without a central authority. In AI, decentralised infrastructure reduces dependency on a single provider, increasing resilience and autonomy, particularly for federated learning applications.
- **Disaggregated:** Involves separating AI infrastructure components (compute, memory, storage, and networking) so they can be independently upgraded, scaled, or allocated based on demand. Disaggregated infrastructure optimises efficiency, allowing AI workloads to better match available resources.
- **Inference Efficiency:** The ability to optimise AI model execution for real-time applications while minimising computational overhead and energy consumption.
- **Edge AI:** AI inference and processing performed **closer to the data source**, such as IoT devices, mobile systems, or on-premise servers, to r**educe latency and bandwidth costs**.
- **Federated Learning:** A decentralised AI training approach where models learn across multiple data sources without centralising data, enhancing privacy and security.
- **AI Orchestration:** The management and coordination of AI workloads across hybrid or multi-cloud environments, ensuring efficient resource allocation and scaling.
- **Zero-Trust AI Security:** A security model where every request and transaction is verified before access is granted, ensuring AI workloads remain protected from cyber threats.
- **Self-Healing Infrastructure:** AI-driven infrastructure that can automatically detect, troubleshoot, and recover from failures, reducing downtime and operational costs.
- **Memory-Bound Workloads:** AI tasks that are constrained more by memory bandwidth than computational power, requiring specialised infrastructure optimisations.
- **Heterogeneous Compute:** AI architectures that leverage a mix of GPUs, TPUs, ASICs, FPGAs, and CPUs to optimise for specific workloads.
- **Composable Infrastructure:** A flexible computing model where **compute, storage, and networking resources** are dynamically allocated based on AI workload needs.
- **Model Quantisation:** The process of reducing the precision of AI model parameters (e.g., converting from FP32 to INT8) to improve inference efficiency with minimal accuracy loss.
- **Data Gravity:** The concept that large-scale AI datasets attract computing resources to their location, influencing infrastructure design and data center placement.

Chapter 04 — AI Infrastructure Challenges (3/4)

- **Data Mobility Platform:** High-speed networking technologies that enable efficient AI workload distribution across distributed compute resources.
- **Elastic Scaling:** The ability to dynamically expand or contract AI compute resources in response to workload demands, ensuring efficiency and cost control.
- **Synthetic Data:** AI-generated data used for training machine learning models when real-world datasets are limited, sensitive, or costly to obtain.
- **Latency Budgeting:** The process of allocating acceptable latency thresholds across various AI system components to maintain real-time performance.

Understanding these distinctions helps enterprises align their AI strategies with the right infrastructure choices.

Scalability & Compute Challenges

Chapter 03 of this report covered the fundamental shift from **CPU-centric to GPU-centric architectures**, which has enabled AI breakthroughs, but it has also exposed significant bottlenecks in enterprise IT infrastructure. While training large-scale models like GPT-4 or multimodal AI systems is compute-intensive, the real challenge for enterprises lies in scaling inference workloads efficiently. Deploying AI at scale requires low-latency, high-throughput inference across diverse applications, from real-time customer interactions to edge computing deployments.

Inference Bottlenecks in Enterprise AI

- **Latency Constraints:** AI-driven applications such as fraud detection, conversational AI, and autonomous decision-making demand sub-10ms response times.
- **Throughput Challenges:** Enterprises processing massive AI workloads require optimised inference pipelines that can handle millions of real-time queries per second.
- **Cost of Deployment:** Unlike training, inference is an **always-on** workload, meaning enterprises must balance cost-efficiency with performance scalability.
- **Hybrid AI Compute for Inference:** Many enterprises are shifting to AI inference-optimised architectures, leveraging custom ASICs, FPGAs, and efficient GPU clusters to reduce operational costs.

To navigate these inference challenges, enterprises must align their AI strategies with cost-effective infrastructure planning, adaptive compute scaling, and emerging AI inference accelerators.

The Cloud vs. On-Prem Trade-off

- **Hyperscale AI Inference:** Cloud providers like AWS, Azure, and Google Cloud offer high-density inference-optimised instances, but enterprises face vendor lock-in risks and unpredictable costs.
- **On-Premise AI Inference:** Many enterprises are deploying dedicated inference clusters on-premises to ensure low-latency, high-availability AI services.
- **AI-Specific Inference Chips:** The rise of Google's TPUs, AWS Inferentia, and custom AI accelerators is reshaping how enterprises optimise for real-time inference workloads.
- **The GPU Price War:** The rapid decline in GPU-hour pricing—with costs dropping over 70% in three years—creates an environment of fierce competition and commoditisation. While this benefits AI accessibility, it risks discouraging hardware innovation and reducing service reliability.

The AI Networking Bottleneck

AI inference workloads are latency-sensitive and bandwidth-intensive, creating unprecedented networking challenges. The shift to real-time AI applications across multi-region deployments has pushed existing data center networks to their limits.

More importantly, AI's **real constraint isn't compute, but interconnectivity**. The classic internet was not built for AI's demands, and its limitations are becoming an industry-wide crisis.

- **High-Speed Interconnects:** Traditional Ethernet is insufficient; enterprises are adopting **InfiniBand, NVLink, and AI-specific data mobility platforms**
- **AI Availability Zones:** Companies like **Stelia** and **Microsoft** are redesigning AI data centers with ultra-low-latency metro-scale AI clusters.
- **AI-Native Networking:** Emerging solutions such as **Hyperband by Stelia** are designed to prioritise AI data flows, ensuring real-time processing and mission-critical AI applications are not bottlenecked by legacy internet architectures.
- **Private AI Backbones:** Some enterprises are investing in dedicated AI fiber networks to ensure stable and high-speed AI workloads.

Without a fundamental shift to AI-optimised networking solutions, even the most powerful – read most expensive component in the system - compute infrastructure will remain underutilised due to needless bandwidth constraints.

Chapter 04 — AI Infrastructure Challenges (4/4)

Strategic Recommendations f

For Large Enterprises

01. Invest in AI-Native Data Centers:
Enterprises should explore AI-native data center models that prioritise low-latency, high-bandwidth AI compute while integrating automation for efficiency gains.

02. Develop an Adaptive AI Compute Strategy:
Balancing on-prem, cloud, and edge AI infrastructure can optimise cost, security, and performance.

03. Prioritise Energy Procurement as a Core AI Strategy:
Businesses must engage directly with energy providers to secure reliable, high-availability power contracts, including nuclear, geothermal, and dedicated grid partnerships.

04. Leverage AI-Driven Energy Optimisation:
Implementing AI-powered workload scheduling and intelligent cooling solutions can significantly cut operational costs.

05. Prioritise AI Security & Compliance:
As AI models become critical infrastructure, enterprises must integrate zero-trust AI security models and regulatory compliance frameworks.

For AI Startups

01. Optimise for Cost-Efficient AI Scaling:
Startups should prioritise multi-cloud infrastructure and lightweight AI models to maximise efficiency while minimising costs.

02. Leverage Edge AI for Latency-Sensitive Applications:
Deploying AI at the edge can reduce dependency on centralised cloud compute while improving response times.

03. Secure Energy-Resilient Infrastructure:
Startups should consider colocating AI workloads in regions with reliable, surplus energy capacity to avoid future constraints.

04. Partner with AI Hardware & Power Providers:
Collaborating with chip manufacturers and energy suppliers can ensure access to the latest, cost-effective AI acceleration and power solutions.

For Policymakers & Industry Leaders

01. Accelerate AI-Specific Grid & Power Infrastructure Upgrades:
Governments must fast-track policies that prioritise AI data center energy expansion and streamline permitting processes.

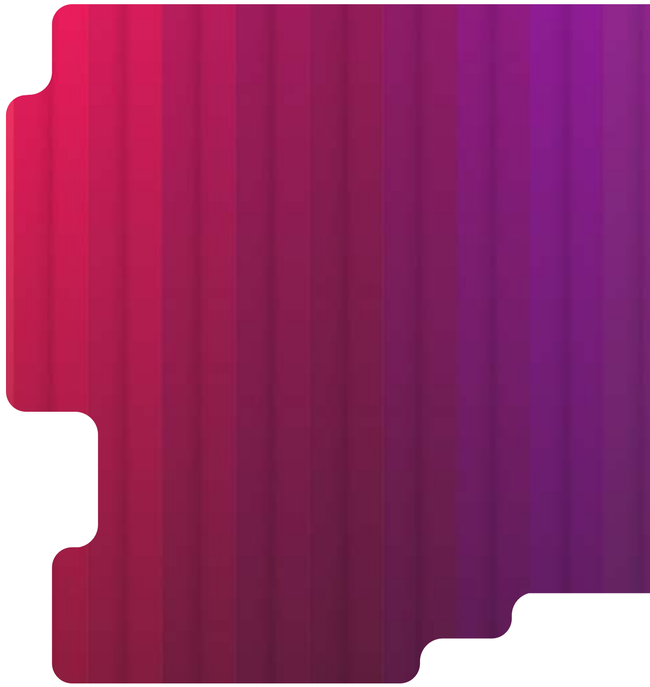
02. Create Incentives for AI Sustainability:
Tax credits and grants should be offered for enterprises investing in energy-efficient AI computing and infrastructure.

03. Ensure AI Infrastructure Security & Sovereignty:
AI infrastructure must be protected from geopolitical risks by strengthening supply chain resilience and cross-border AI governance.

04. Establish Industry Standards for AI-Native Networks & Power Strategies:
The industry must collaborate on setting performance, security, and interoperability standards for AI-specific networking technologies and power grids.

By integrating these strategic recommendations, enterprises, startups, and policymakers can collectively build an AI infrastructure ecosystem that is resilient, scalable, and energy-efficient. AI's next phase will not be determined solely by breakthrough models—but by the ability to deploy, manage, and optimise infrastructure at an unprecedented scale.

As enterprises refine their AI infrastructure strategies, the next major challenge lies in **managing the explosion of data**. AI workloads are not only compute-intensive—they are data-hungry, requiring efficient data pipelines, storage systems, and governance frameworks. In the next chapter, we will explore how enterprises can scale AI workloads while ensuring data integrity, compliance, and accessibility in an era of exponential data growth.



Chapter 05

Data Growth and AI Workloads

Chapter 05 — Data Workloads and the Growth of AI (1/6)

Introduction: Why This Chapter Matters

As enterprises push forward with AI-driven transformation, they face an evolving data landscape that challenges traditional models of compute, storage, and infrastructure. The preceding chapter explored AI infrastructure challenges from hardware and data center capacity to ease of adoption and how enterprise, startups and policy makers can do about them. This chapter extends that discussion by examining the data workloads fueling AI's exponential growth, the infrastructure challenges enterprises must overcome, and the role of AI workloads in reshaping business operations. Understanding these dynamics is critical for organizations seeking to harness AI effectively and stay competitive in the digital economy.

AI is redefining how enterprises generate, store, and process data. The volume of AI-driven workloads is expanding at an unprecedented rate, shifting the center of enterprise computing from traditional data storage to AI inference and execution. By 2030, AI workloads are expected to absorb 50% of enterprise compute capacity, a transformation that demands new approaches to data mobility, storage, and infrastructure scaling.

The challenge for enterprises is not simply managing more data — it is ensuring that AI-driven workloads remain operationally efficient, cost-effective, and scalable. Organizations that fail to build AI-ready infrastructure will face bottlenecks in data movement, compute efficiency, and decision latency, ultimately slowing their ability to compete in an AI-driven economy.

The Exponential Growth of AI Workloads

The volume of global data is accelerating at an unprecedented pace. In 2020, approximately 64 zettabytes of data were created, captured, and consumed worldwide. This number was expected to have reached 160 zettabytes in 2024, and surpass 400 zettabytes by 2028.

What makes this growth even more significant is the shift from human-generated data to machine-generated data.AI models, IoT devices, and autonomous systems now account for a majority of new data production. Unlike traditional enterprise datasets, AI-generated data requires high-speed movement, rapid processing, and scalable storage to deliver real-time insights.

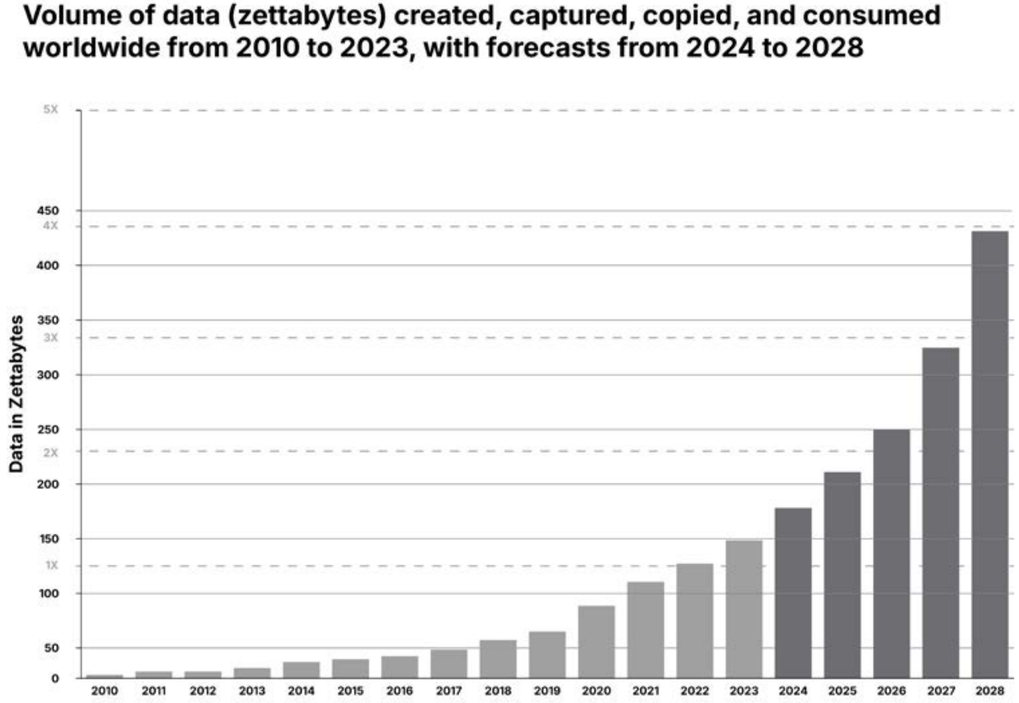


Figure 1: Global Data Volumes (Zettabytes) by Year, from 2010 to 2028. Source: Statista.

Chapter 05 — Data Workloads and the Growth of AI (2/6)

AI Workloads Are Overtaking Traditional Data Workflows

AI is no longer a niche workload—it is becoming the dominant driver of enterprise compute demand.

- **IDC's Workload Evolution Report** estimates that by 2030, AI-driven processes will represent half of all enterprise computing tasks.
  - 2024: IDC expects growth in compute and storage systems spending for cloud workloads (14% CAGR) will grow faster than infrastructure spending on traditional workloads (8.4% CAGR).
  - AI workloads (intensive and less intensive AI) across 'Big 4' professional services firms is expected to exceed 50% share of workload by 2030, increasing from less than 20% in 2019.
- **EpochAI's Compute Scaling Data** shows that the training compute required for state-of-the-art AI models has doubled every six months since 2010.
- **Inference is emerging as the key value driver for AI** —organizations must build infrastructure that optimizes not just training, but real-time decision execution.

This shift has profound implications. Enterprises that previously optimized infrastructure for transactional databases and cloud analytics must now rethink their architecture to support inference-heavy workloads at scale.

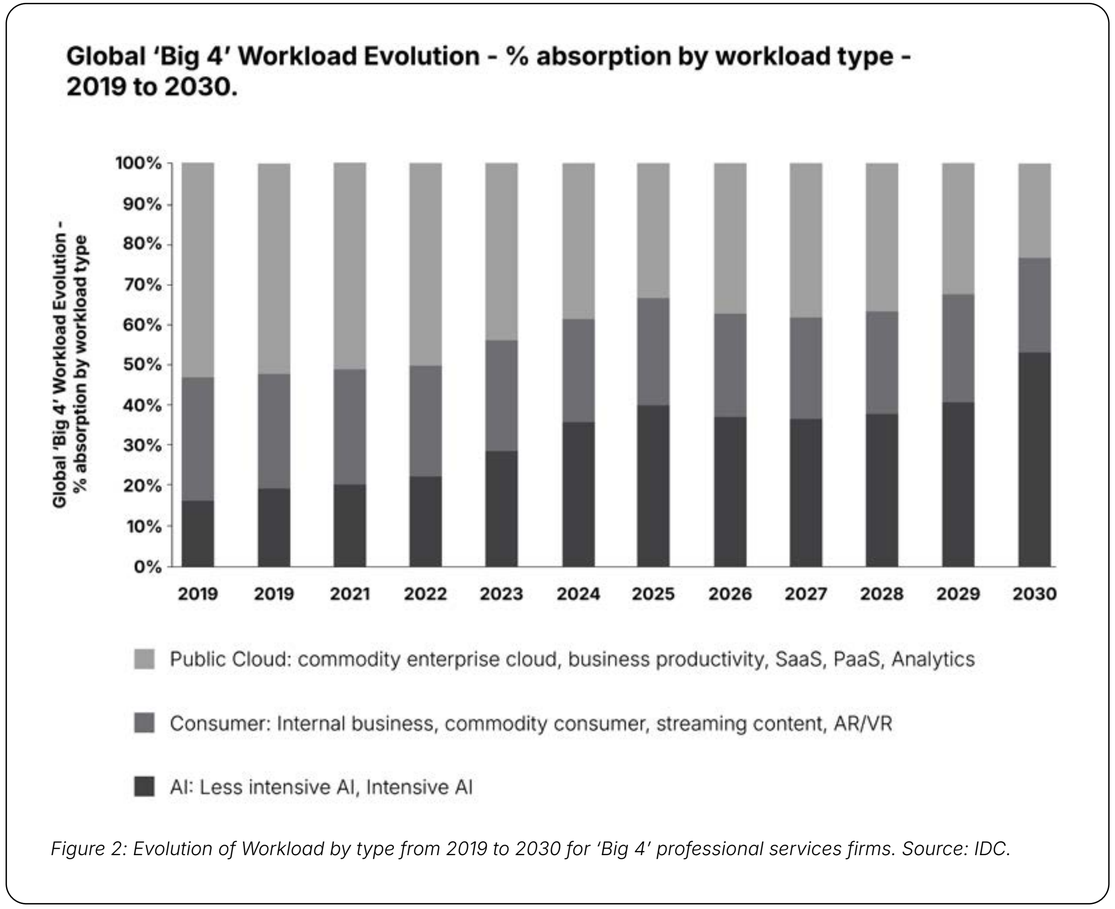


Figure 2: Evolution of Workload by type from 2019 to 2030 for 'Big 4' professional services firms. Source: IDC.

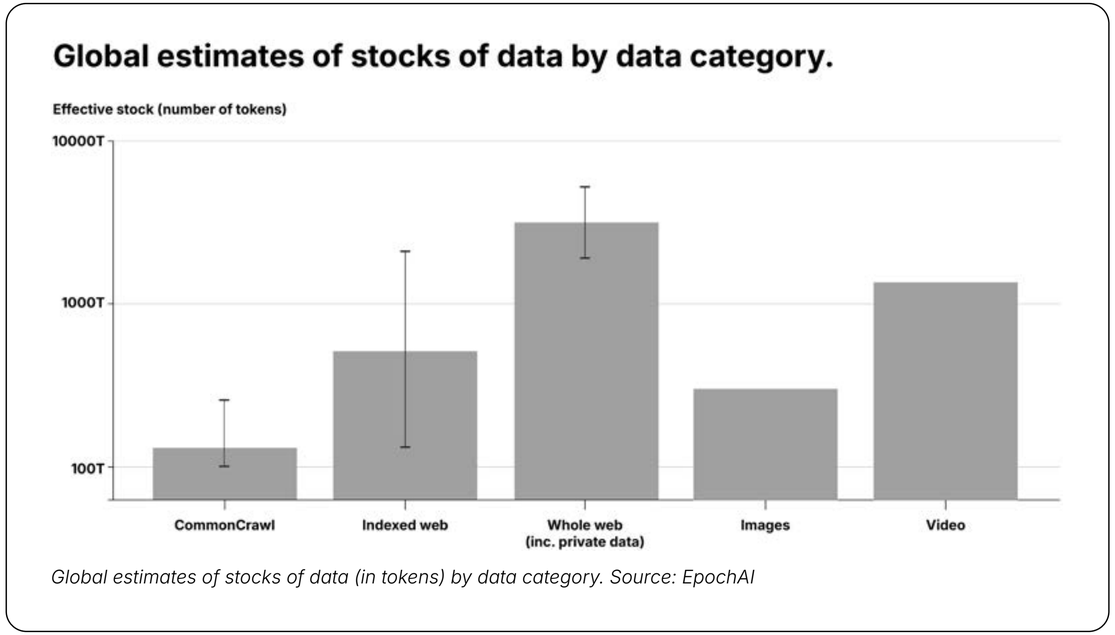
Chapter 05 — Data Workloads and the Growth of AI (3/6)

The Emerging Data Bottlenecks: Storage, Compute, and Energy

As AI workloads scale, enterprises are encountering critical data infrastructure bottlenecks:

- **Storage Constraints** – Traditional storage systems are struggling to keep up with AI-generated data volumes, requiring enterprises to rethink capacity planning and scalability.
- **Compute Inefficiencies** – GPUs, which are critical for AI workloads, are in short supply, leading to compute resource contention.
- **Energy Consumption** – AI data workloads demand 2x to 10x more power than traditional enterprise applications, posing sustainability and cost challenges.

The future of enterprise AI will depend on **purpose-built infrastructure designed for inference at scale** — an approach that prioritizes efficiency in data mobility, GPU utilization, and low-latency execution.



Global estimates of stocks of data (in tokens) by data category. Source: EpochAI

The Looming Data Scarcity Problem

AI models rely on vast amounts of high-quality data. However, industry projections indicate that publicly available high-quality datasets could be exhausted by 2032.

As this happens, enterprises will need to:

- Shift toward **synthetic data generation** to extend AI model training capabilities.
- Leverage **private data ecosystems** to maintain proprietary competitive advantages.
- Build **sector-specific AI models** trained on exclusive enterprise datasets rather than relying on general-purpose public AI models.

This shift marks a **new phase of AI execution**, where enterprises must actively control, curate, and manage their own AI data sources to maintain competitive advantage.

Chapter 05 — Data Workloads and the Growth of AI (4/6)

AI Workloads & The Changing Nature of Enterprise Tasks

The distribution of AI workloads is reshaping workforce dynamics across industries. While **Computer & Mathematical occupations lead AI adoption (37.2%)**, AI-driven automation is expanding into finance, education, and healthcare. Real-world implementations demonstrate how enterprises are adapting to these shifts.

AI's Uneven Distribution Across Job Sectors

High AI Adoption Fields:

- **Software Development & IT** – Companies like Microsoft have integrated AI-driven tools such as GitHub Copilot, which enhances developer efficiency by automating code generation and debugging.
- **Business & Finance** – Bradesco Bank's AI-powered virtual assistant has streamlined customer service, reducing response times from days to hours.
- **Education & Library Sciences** – AI-powered learning management systems like Deel Engage automate course creation and personalize training for employees.

Emerging AI Adoption Fields:

- **Healthcare Practitioners (6.1%)** – IBM Watson Health employs AI-driven diagnostics and NLP-based patient record analysis to enhance clinical decision-making.
- **Manufacturing & Logistics** – Amazon Cobots optimize supply chain operations by reducing manual workloads in fulfillment centers.
- **General Management (6.9%)** – AI-powered strategic planning tools help businesses optimize operations by integrating predictive analytics.

Lagging AI Adoption Fields:

- **Construction (4.1%) and Installation & Maintenance (3.9%)** – AI-based predictive maintenance is improving efficiency, but adoption remains limited compared to other sectors.

As AI workloads become more sector-specific, enterprises must optimize **real-time data access and storage models** to keep up with AI-generated data demands. The ability to move, store, and process AI-driven data at scale will determine how effectively businesses can integrate AI into daily operations.

Data Gravity & the New Enterprise Data Landscape

As enterprises scale AI workloads, the way data moves within an organization becomes a defining factor in performance.

- **The Digital Realty Data Gravity Index™ forecasts** that enterprise data creation will reach **1.2 million exabytes in the next three years**.
- **The challenge:** AI workloads require rapid access to large datasets, but traditional cloud architectures struggle with data movement inefficiencies.
- **The solution:** Enterprises must bring compute closer to where data is generated, whether at the edge, in regional AI hubs, or through optimised hybrid cloud strategies.

Infrastructure optimized for **AI data mobility eliminates latency bottlenecks**, ensuring that AI-driven decision-making happens in real time rather than being delayed by inefficient data transfer processes.

As enterprises adapt to the changing demands of AI in tandem with growing data volumes and changes in data patterns, it is increasingly becoming clear this is an area of focus where enterprise must adapt to remain future proof.

Chapter 05 — Data Workloads and the Growth of AI (5/6)

Enterprise Case Studies: Adapting to the growing demands of data

01. Netflix: Managing Data Growth & Cloud Adoption

Challenge:

- Rapidly growing data storage needs due to global video streaming expansion.
- Need for scalable storage to support millions of users streaming simultaneously.
- Requirement for fast access to massive content libraries while ensuring redundancy.

Solution:

- **Migrated from on-premises data centers to AWS Cloud** for scalable storage.
- **Implemented Amazon S3 and Amazon Glacier** to manage hot and cold storage efficiently.
- **Built a custom Content Delivery Network (CDN)** to optimize data mobility and content distribution.
- Used **machine learning (ML) workloads** for content recommendations, requiring high-performance storage solutions.

Outcome:

- Elastic scalability to accommodate exponential data growth.
- Significant cost savings by optimizing cloud storage usage.
- Improved streaming performance and redundancy globally.

02. Goldman Sachs: AI Workload Evolution & Cloud Adoption

Challenge:

- Need for high-performance storage to handle financial analytics and AI-driven risk modeling.
- Traditional storage infrastructure struggled with AI and ML workloads that required real-time data access.
- Increasing regulatory requirements for secure, scalable, and compliant storage solutions.

Solution:

- **Adopted a hybrid cloud model** using AWS, Google Cloud, and on-prem data centers.
- **Leveraged object storage (AWS S3, Google Cloud Storage)** to store massive datasets used for AI/ML models.
- **Deployed AI-driven analytics platforms** that integrated high-speed NVMe storage for fast processing.

Outcome:

- Enhanced ability to process complex AI workloads with lower latency.
- Improved compliance with financial data storage regulations.
- Greater agility in managing and analyzing data for trading, fraud detection, and risk analysis.

03. Tesla: AI & Edge Data Storage for Autonomous Driving

Challenge:

- Massive amounts of **sensor and telemetry data** generated from autonomous vehicles.
- Need for **real-time AI training** using neural networks that require high-performance storage.
- Requirement for **edge storage** to reduce cloud latency.

Solution:

- **Implemented edge storage on vehicles** with SSDs to store and process sensor data before syncing to the cloud.
- **Built AI data pipelines** using on-prem supercomputers for training Tesla's Full Self-Driving (FSD) algorithms.
- **Leveraged multi-cloud storage solutions** to manage long-term AI training datasets efficiently.

Outcome:

- Improved AI model training efficiency by reducing cloud dependency.
- Faster updates to FSD models due to optimized data pipelines.
- Enhanced vehicle performance through better data mobility between the cloud and edge.

04. Pfizer: Cloud Adoption for Biopharma R&D

Challenge:

- Expanding volume of genomic, clinical trial, and drug development data.
- Need for scalable storage to support AI-driven drug discovery and bioinformatics.
- Requirement for regulatory compliance (HIPAA, GDPR) in data storage.

Solution:

- **Migrated to hybrid cloud (AWS & Azure)** to handle the storage of research data.
- **Used high-performance parallel file storage** to accelerate AI-driven analysis of genomics.
- **Adopted tiered storage strategies** with warm/cold data management for cost efficiency.

Outcome:

- Reduced time to market for drug development.
- Improved collaboration between research teams through cloud-based data mobility.
- Compliance with global regulations while maintaining high-speed access to critical datasets.

Chapter 05 — Data Workloads and the Growth of AI (6/6)

05. Uber: Data Mobility & AI-Powered Analytics

Challenge:

- Need to handle massive amounts of real-time GPS, ride request, and transaction data.
- Storage systems needed to support **machine learning** models for pricing, fraud detection, and demand forecasting.
- Data required high-speed mobility across global data centers and cloud infrastructure.

Solution:

- **Built an in-house data lake (Hudi)** to improve storage efficiency.
- **Implemented multi-region cloud storage** (AWS & Google Cloud) for real-time analytics.
- **Optimized AI storage pipelines** for faster model training and inference.

Outcome:

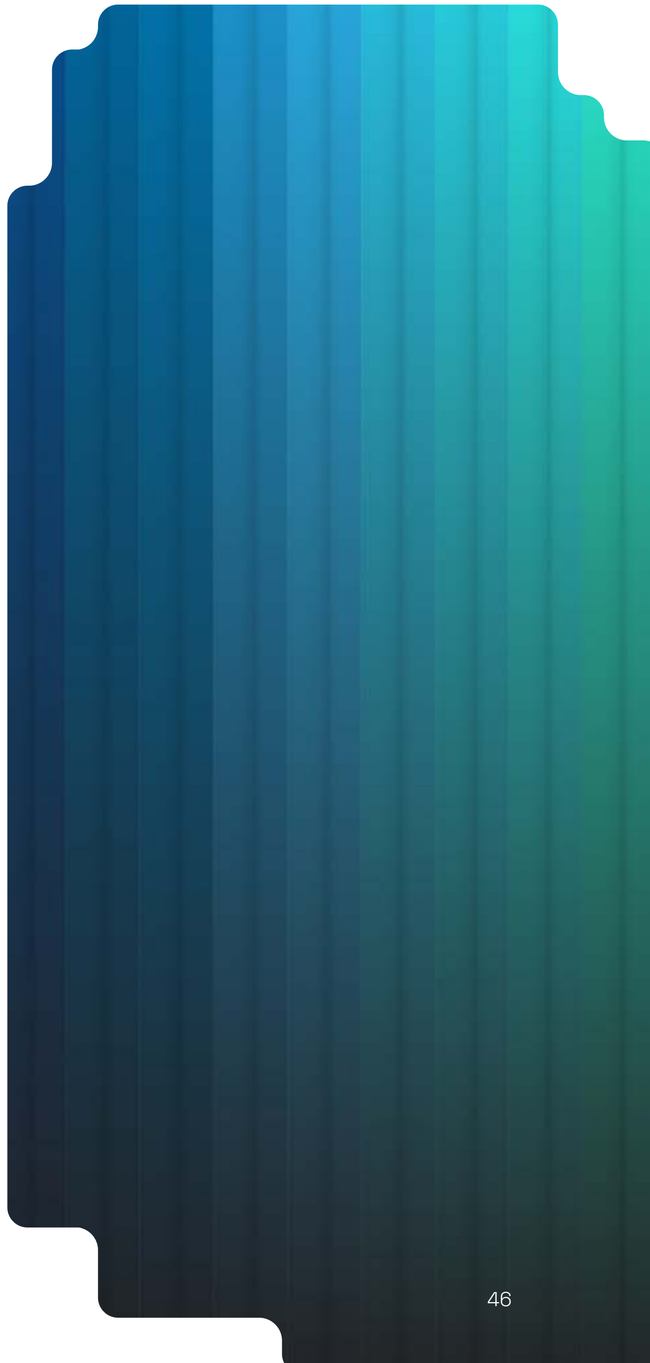
- Scaled to handle petabytes of data with fast retrieval speeds.
- Enhanced AI-driven pricing and routing models with low-latency storage.
- Improved fraud detection through real-time AI analytics.

Key Takeaways from Enterprise Case Studies:

- **Cloud adoption (AWS, Azure, Google Cloud) is a dominant strategy** to handle data growth and AI workloads, although the financial implications may make this unsustainable
- **Hybrid models (on-prem + cloud) are effective for regulated industries** (e.g., finance, healthcare).
- **AI-driven storage optimization** (edge computing, parallel storage, tiered storage) helps in handling AI/ML workload evolution.
- **Data mobility is critical** for industries that require real-time processing (e.g., Uber, Tesla).

The Future Outlook for AI

As AI workloads become increasingly central to enterprise operations, organizations must prioritize scalable infrastructure, efficient data management, and optimized inference execution. The next chapter explores the The Future Outlook for AI: 2025 and Beyond, examining the emerging trends, innovations, and strategic directions that will define AI's role in enterprise success. Businesses that effectively integrate AI-driven workloads into their core operations will position themselves at the forefront of the next digital era.



Chapter 06

Future Outlook

Chapter 06 — The Future Outlook for AI: 2025 and Beyond (1/5)

AI's Transformative Role in Enterprise, 2025 and Beyond

During 2025, AI will evolve from a tool to an active participant in enterprise operations, reshaping industries and unlocking significant value. Key takeaways for leaders:

- **Autonomous AI Agents:** Digital "knowledge workers" will drive efficiency, reduce costs, and enhance decision-making across functions, from supply chain optimization to customer service.
- **Infrastructure Evolution:** Enterprises must adopt scalable, cost-effective AI infrastructures, including edge computing, GPU superclusters, and open standards for seamless integration.
- **Smaller, Specialized AI Models:** Task-optimized AI systems will replace monolithic models, enabling faster, more affordable deployment without compromising performance.

Challenges:

- The "last-mile problem" hampers AI scaling due to data integration and workforce adoption issues.
- Enterprises must invest in holistic systems, upskilling, and ethical governance to bridge the gap between AI's potential and reality.

Introduction

Imagine walking into your office in 2025. Your AI agent has already analyzed overnight market movements, adjusted your supply chain to avoid an emerging disruption in Southeast Asia, and drafted responses to priority customer inquiries for your review. This isn't science fiction — it's the near-future reality of AI in enterprise. As we approach 2025, we're witnessing a fundamental shift from AI as a collection of specialized tools to AI as the nervous system of modern business operations.

The key to understanding this transformation isn't i n speculating about artificial general intelligence (AGI) or distant futures. Instead, it lies in recognizing how today's emerging technologies are rapidly evolving into practical, powerful systems that will reshape organizational operations. The real story of AI in 2025 is about autonomous agents, enhanced infrastructure, and practical applications that deliver measurable business value.

Core Enterprise AI Developments

The Rise of Autonomous AI Agents

During 2025, AI agents will increasingly operate as digital knowledge workers, fundamentally changing how enterprises function. Gartner's prediction that 33% of enterprise software will integrate agentic AI by 2028 actually understates the transformation. Early adopters are already seeing this shift — take Goldman Sachs' AI-driven trading operations, which now handle 35% of trades autonomously, up from 5% in 2019. But trading is just the beginning.

These agents will operate continuously across enterprise functions, learning from every interaction. For example, a global manufacturer implementing autonomous supply chain agents in early 2024 reported a 23% reduction in inventory costs and a 45% decrease in stockout incidents within six months. The agents didn't just execute predefined rules — they actively learned from supplier behavior patterns and market conditions to make increasingly sophisticated decisions.

Advanced Language Models & Multimodal AI

While media attention focuses on consumer applications, the real revolution is happening in enterprise applications of multimodal AI. Consider a pharmaceutical company's recent pilot program where AI systems simultaneously analyze research papers, lab results, and molecular modeling data to identify promising drug candidates. This integration of different types of data and analysis has already reduced early-stage drug discovery timelines by 60%.

By 2025, these systems will routinely process text, audio, visual, and sensor data in real-time, enabling more natural and effective human-AI collaboration. A major aerospace manufacturer is already testing systems that combine visual inspection data, maintenance records, and real-time sensor readings to predict equipment failures with 94% accuracy, weeks before traditional methods would detect issues.

Chapter 06 — The Future Outlook for AI: 2025 and Beyond (2/5)

Infrastructure Evolution

The backbone of this AI transformation requires a fundamental rethinking of computing infrastructure. The emergence of GPU superclusters exceeding 500,000 units isn't just about raw computing power — it's about enabling new types of applications. Microsoft's recent deployment of a 200,000-GPU cluster demonstrated a 40% reduction in model training time while supporting five times more concurrent users than traditional architectures.

The shift from Infiniband to Ethernet for distributed workloads represents a move towards open-standards access to AI capabilities. Early adopters report 30% lower deployment costs and 50% faster scaling of AI applications. Edge computing integration is already showing dramatic results in latency-sensitive applications — a major retailer's edge-based inventory management system reduced response times from seconds to milliseconds, enabling real-time optimization of store operations.

Short-Term Transformations: Laying the Foundations (6–12 Months)

The Pivot to Inference: To date, AI deployments have largely focused on training large language models (LLMs), but the paradigm is shifting toward inference — the execution of these models in real-world applications. This pivot is driving the proliferation of edge locations tailored for inference, particularly in latency-sensitive industries like retail and manufacturing. Enterprises will optimize inference closer to the source, minimizing costs and improving response times.

Consider Scandinavia, where training-focused deployments have struggled to align with regional needs. The next year will see a reconfiguration of resources to address these mismatches, prioritizing inference-ready infrastructure.

Cloud Computing vs. Bare Metal: The AI gold rush spurred the rise of neocloud providers — lean, agile players leasing bare-metal GPU clusters to meet surging demand that hyperscalers couldn't fulfill. While this approach enabled fast scaling, it often left enterprises without the integrated platforms they expect.

CoreWeave's partnership with Microsoft exemplifies how bare-metal solutions bridged gaps temporarily, but the future lies in cloud platforms that deliver fully integrated, elastic GPU compute. Solutions like Lambda's "one-click" deployments signal the beginning of this evolution, making GPU provisioning as seamless as CPU or storage allocation.

Tighter Integration of GPU Compute and AI Workloads: Enterprises are demanding solutions that abstract complexity, enabling effortless deployment of AI workloads. In the coming months, GPU compute providers must enrich their platforms, moving beyond Infrastructure-as-a-Service to deliver AI-native capabilities. This will involve automated workload provisioning, advanced orchestration tools, and intuitive interfaces tailored to enterprise needs.

Smaller, More Specialized Models: The trend toward smaller, task-optimized models marks a departure from the era of monolithic LLMs. These nimble models, championed by NVIDIA as Neural Information Models (NIMs), are designed for efficiency and specific use cases. Enterprises will adopt these models to achieve faster deployment cycles and reduce costs while maintaining performance.

Industry Transformations — 2025 and Beyond

The impact of AI advancement will vary dramatically across industries, each facing unique opportunities and challenges. For model builders — the companies creating AI systems themselves — we'll see a fascinating recursive effect: AI increasingly building AI. Development platforms will shift toward autonomous training and optimization, making sophisticated model development accessible to organizations without massive AI expertise. This democratization will accelerate innovation across all sectors, though it raises important questions about quality control and governance.

In **social media**, the scale of content and interaction demands more sophisticated AI management. By 2025, AI agents will handle content moderation with unprecedented accuracy, moving beyond simple flag-and-review systems to understanding context and nuance in real-time. These systems will simultaneously personalize user experiences while protecting against manipulation, particularly as synthetic media becomes more sophisticated.

Life sciences stands to see perhaps the most profound transformation. The convergence of AI with lab automation and synthetic biology will dramatically accelerate drug discovery and development. Imagine AI systems that not only predict molecular behaviors but actively design and run experiments, learning and adjusting in real-time based on results. This isn't just faster research — it's a fundamental reimagining of how we approach biological discovery.

Chapter 06 — The Future Outlook for AI: 2025 and Beyond (3/5)

Engineering and robotics will see a shift toward truly autonomous systems. Rather than simply executing programmed tasks, robots will adapt to changing conditions and learn from experience. AI will move from assisting design to actively participating in the engineering process, suggesting optimizations and identifying potential issues before they become problems. The key advancement here is the integration of edge computing, allowing for real-time decision-making without relying on cloud connectivity.

Space technologies present unique challenges that AI is uniquely positioned to address. The combination of autonomous systems and advanced predictive capabilities will transform everything from mission planning to satellite operations. AI will help manage the increasing complexity of space operations, from debris avoidance to resource optimization, making space more accessible while improving safety and reliability.

The **research and analysis** sector will experience a fundamental shift in how we approach discovery itself. AI won't just process data faster — it will actively identify patterns and generate hypotheses across disparate fields of study. This capability for cross-domain insight, combined with real-time analysis of massive datasets, will accelerate the pace of discovery across all scientific fields.

Financial services will continue their AI transformation, but with a crucial shift toward more autonomous systems. Beyond current automated trading and fraud detection, we'll see AI systems that can actively manage risk and compliance across entire portfolios. The key advancement will be in combining multiple data streams — market data, news, social sentiment, and more — to make more nuanced decisions in real-time.

Finally, in **defence and surveillance**, the focus will be on enhanced situational awareness and response capabilities. AI systems will move beyond simple threat detection to provide comprehensive battlefield analytics and decision support. The challenge here isn't just technological — it's about developing systems that can make split-second decisions while maintaining appropriate human oversight and ethical constraints.

Across all these sectors, the common thread is a move from AI as a tool to AI as an active participant in decision-making processes. The key to success will be building the right infrastructure and governance frameworks to support these advances while managing the associated risks and ethical considerations.

Implementation Challenges: Bridging Vision and Reality

The Last-Mile Challenge

The gap between AI's potential and practical implementation — what practitioners call the "last 99-mile problem" — remains a critical challenge. A recent McKinsey study found that while 72% of organizations are piloting AI initiatives, only 15% successfully deploy them at scale. The experience of a global retailer illustrates this challenge: despite successful pilots in inventory optimization showing potential 20% cost savings, full deployment took 18 months longer than planned due to data integration issues and employee adoption challenges.

The solution lies in approaching AI deployment holistically. Companies succeeding in AI implementation, like Walmart's successful rollout of autonomous inventory management across 4,700 stores, share common characteristics: they prioritize data infrastructure, invest in employee training, and take an iterative approach to deployment.

Long-Term Transformations: Redefining Infrastructure (12+ Months)

Building for 2025 requires infrastructure that can handle increasingly complex AI workloads while remaining flexible and cost-effective. Meta's recent infrastructure overhaul provides a telling example: their shift to a hybrid architecture combining edge computing with centralized processing reduced AI model latency by 65% while supporting a 3x increase in concurrent AI operations.

Holistic, Integrated Systems: The era of piecemeal IT stacks is fading. Enterprises are rediscovering the value of integrated systems, where compute, storage, and networking are cohesively designed for AI workloads. NVIDIA's Grace Hopper systems exemplify this trend, offering full-rack solutions optimized for AI performance. These systems promise not just speed, but the seamless interconnection necessary for distributed AI workflows across data centers.

The Return of On-Premise Infrastructure: As the cost of cloud-based GPU clusters continues to rise, a subset of enterprises will gravitate back toward on-premise solutions. Companies like Oxide Computer are pioneering "cloud in a rack" systems that blend on-premise control with cloud-like scalability. For organizations with consistent, high-intensity AI workloads, this hybrid approach offers both cost savings and operational flexibility.

Chapter 06 — The Future Outlook for AI: 2025 and Beyond (4/5)

However, many enterprises lack the expertise or capital for full on-prem deployments. These businesses will remain in cloud environments but demand more dynamic, on-demand GPU compute, along with distributed storage solutions that transcend single-location bottlenecks.

Open Standards for AI Infrastructure: The need for universal standards in accelerated computing is becoming urgent. Just as POSIX standardized operating systems in the 1980s, frameworks like NVIDIA's NVLink, Intel's oneAPI, and emerging protocols from the Internet Engineering Task Force are paving the way for interoperability. This will simplify AI deployment, allowing enterprises to leverage diverse hardware without deep technical knowledge.

The key is balance. JPMorgan Chase's approach demonstrates this well — they maintain powerful centralized clusters for complex risk modeling while deploying edge AI for real-time fraud detection, achieving 99.97% accuracy with response times under 50 milliseconds.

Infrastructure Evolution

The foundation for future AI advancement is being laid today. The Internet Engineering Task Force's work on new protocols for AI-optimized networking shows promising early results — test implementations demonstrate 40% improved efficiency in distributed AI workloads. Companies like NVIDIA and AMD are already developing next-generation AI accelerators that promise 5–10x performance improvements by 2025.

The Path to Enhanced AI

While full AGI remains a longer-term prospect, significant advances in AI capabilities will emerge by 2025. DeepMind's latest research shows promising results in transfer learning, with models successfully applying knowledge across different domains with 70% effectiveness — a dramatic improvement from current 30% rates. However, these advances will focus on enhancing specific business capabilities rather than achieving general intelligence.

Tokenomics and AI: Decentralizing Compute for the Future

As AI becomes the nerve center of enterprise operations, tokenomics offers a transformative approach to managing and allocating compute resources. By 2025, decentralized compute marketplaces will emerge, enabling organizations to access a global pool of tokenized computational power. Innovative ecosystem catalysers such as Stelia, powered by HyperBand, will serve as infrastructure enablers, ensuring interoperability and seamless distribution of these tokenized resources.

In this decentralized system, computational power is tokenized, creating a marketplace where unused resources can be traded, perhaps using a stablecoin as currency. Organizations needing compute power for AI workloads can use compute tokens to dynamically access the necessary hardware, while resource providers are incentivized with tokenized rewards.

The Connection Between Compute Tokens and LLM Tokens

- **Compute Tokens:** Represent the hardware power (GPUs, CPUs, etc.) required to run AI systems.
- **LLM Tokens:** Represent the text chunks (words or parts of words) processed by an AI like a large language model (LLM).

How They Interact: Compute tokens enable access to the hardware needed to process LLM tokens. The more LLM tokens your AI needs to analyze or generate, the more compute tokens you may need to provision the required computational resources.

Example:

Imagine a business running an AI chatbot that handles customer questions. The chatbot processes 1 million words of text daily, with each word being part of an LLM token.

01. To process these LLM tokens, the AI needs GPUs to run its computations.

02. The business buys compute tokens from a global marketplace to rent the hardware.

03. If customer inquiries double overnight, the business can buy more compute tokens to scale up and meet demand, without investing in expensive infrastructure.

This interplay ensures flexibility, efficiency, and cost-effectiveness in managing AI workloads.

Chapter 06 — The Future Outlook for AI: 2025 and Beyond (5/5)

Applications of Tokenomics in AI:

- **Autonomous AI Agents:** Dynamically scale AI operations with tokenized compute resources, enabling agents to learn, adapt, and execute in real-time.
- **LLM Workloads:** Manage intensive AI tasks like training and inference for large language models, fueled by tokenized compute.
- **Scientific Research:** Tap into decentralized compute power for simulations and analysis, accelerating discovery across industries.

By enabling a transparent, cost-efficient, and scalable approach to AI, tokenomics democratizes access to computational resources. Stelia and HyperBand power this future by abstracting complexity and unifying distributed resources into a cohesive ecosystem, ensuring organizations can innovate without limits.

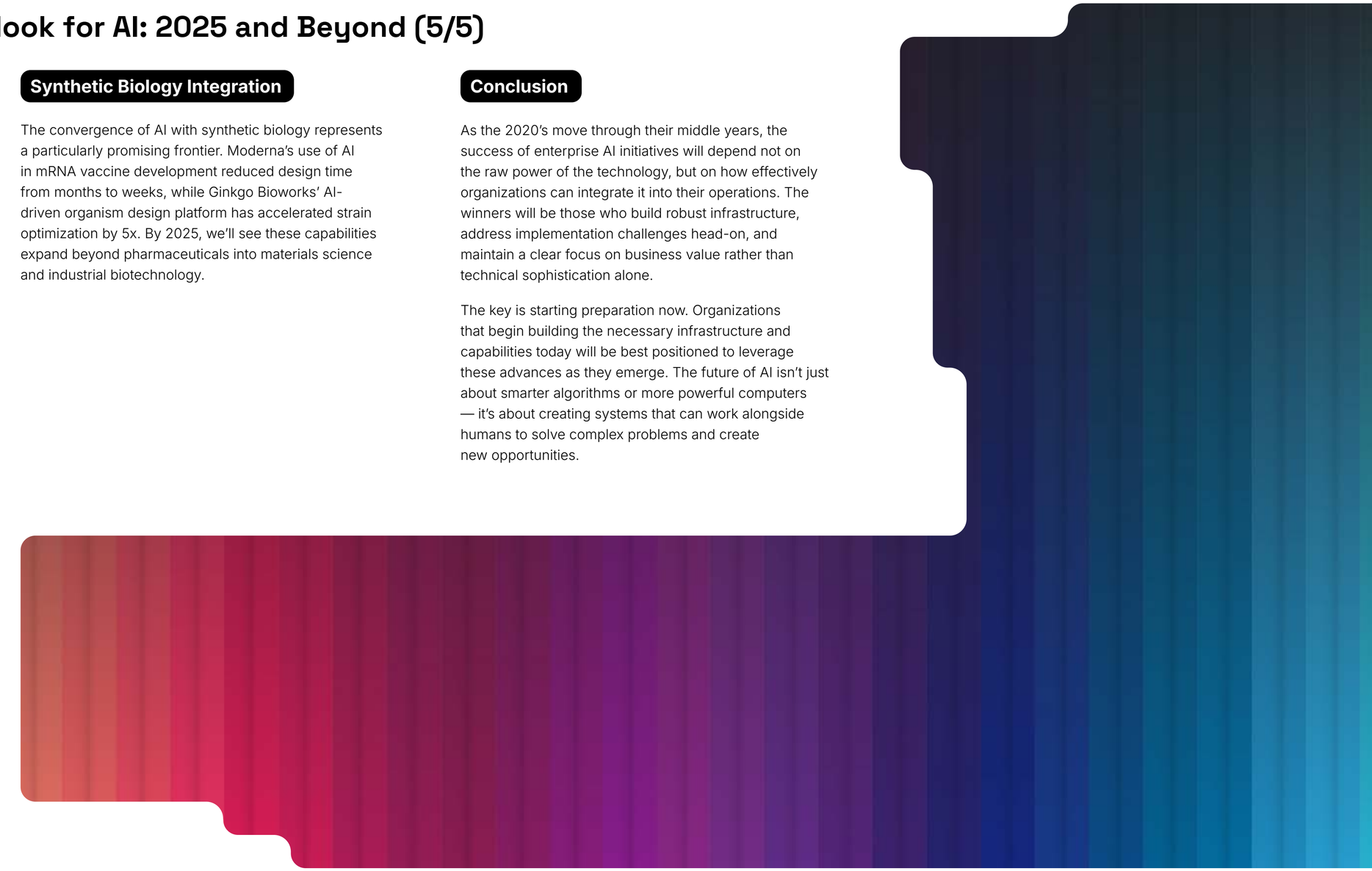
Synthetic Biology Integration

The convergence of AI with synthetic biology represents a particularly promising frontier. Moderna's use of AI in mRNA vaccine development reduced design time from months to weeks, while Ginkgo Bioworks' AI-driven organism design platform has accelerated strain optimization by 5x. By 2025, we'll see these capabilities expand beyond pharmaceuticals into materials science and industrial biotechnology.

Conclusion

As the 2020's move through their middle years, the success of enterprise AI initiatives will depend not on the raw power of the technology, but on how effectively organizations can integrate it into their operations. The winners will be those who build robust infrastructure, address implementation challenges head-on, and maintain a clear focus on business value rather than technical sophistication alone.

The key is starting preparation now. Organizations that begin building the necessary infrastructure and capabilities today will be best positioned to leverage these advances as they emerge. The future of AI isn't just about smarter algorithms or more powerful computers — it's about creating systems that can work alongside humans to solve complex problems and create new opportunities.



Conclusion

Conclusion (1/2)

The AI Infrastructure Race & the Red Queen Effect

01. The AI-Driven Enterprise Transformation: A Continuous Competition

Artificial Intelligence is no longer an emerging technology—it has become the defining force in enterprise strategy, national competitiveness, and economic transformation. This report has explored the structural, technological, and strategic shifts required for enterprises to thrive in an AI-driven world. However, AI is not a one-time innovation—it is a continuously evolving competition, where maintaining a leadership position requires constant optimization, adaptation, and scaling.

This is the Red Queen Effect in action—in the AI race, enterprises and nations must run faster and smarter just to stay in place. The competitive advantage in AI is not simply about adopting the latest models—it is about owning infrastructure, optimizing real-time data pipelines, and securing sustainable compute and energy resources.

The AI economy will be defined by who controls infrastructure, compute, and data mobility—not just who builds the best models. The coming decade will see the AI playing field fragment into infrastructure leaders and infrastructure dependents—and those who fail to build scalable, energy-resilient, and AI-native infrastructure risk being permanently left behind.

02. AI's Next Evolution: From Assistance to Full Automation

- AI has moved beyond human augmentation—it is now automating entire workflows and job functions.
- The AI-first enterprise will be defined by verticalized, industry-specific automation, not just general AI adoption.
- Early AI adopters may gain a temporary lead, but long-term competitiveness depends on real-time AI adaptation and dynamic learning systems. (Red Queen Effect)
- Enterprises must avoid stagnation in AI competitiveness—AI strategy cannot be static, or companies risk becoming obsolete.

In a world where every company is adopting AI, success will not be about AI adoption—it will be about who optimizes AI deployment, inference efficiency, and real-time learning loops the fastest.

03. The AI Infrastructure Imperative: Scaling for AI Marketplaces

- AI infrastructure is no longer an IT consideration—it is a defining factor in business success.
- Enterprises will increasingly consume AI via modular AI marketplaces, where AI services are deployed dynamically rather than being built in-house.
- Multi-cloud and hybrid AI architectures will be required to avoid overdependence on a single AI infrastructure provider. (Strategic Defensive Play)
- The AI economy will be dominated by enterprises that control compute, interconnectivity, and workload optimization at scale.

The AI-native enterprise will not just build AI tools—it will integrate AI into its operational core. Companies must build real-time, dynamic AI infrastructures capable of adapting to changing workloads, shifting regulations, and emerging AI marketplaces.

04. Data Mobility & Distributed AI: The Next Competitive Battleground

- AI is moving from static training models to real-time inference loops, requiring constant data mobility and accessibility.
- Enterprises that fail to optimize data pipelines and interconnectivity will be unable to scale AI effectively. (Data Gravity Effect)
- Compute will move to where data resides, not the other way around. AI-native enterprises must design AI workload placement strategies accordingly.
- The true competitive advantage in AI will be who controls real-time data flows, not just who has the best model architectures.
- The future of AI workloads will be shaped by high-bandwidth AI networking, multi-cloud data sharing, and federated AI learning architectures.

Real-time AI success will be defined by the ability to move, process, and utilize data across distributed environments with near-zero latency. Enterprises that master data mobility, federated learning, and adaptive inference will dominate the AI economy.

Conclusion (2/2)

5. The Energy Bottleneck & AI's Future Growth

- AI is not just a compute race—it is now a power grid competition.
- AI's power consumption is skyrocketing, forcing enterprises to plan energy procurement strategies before infrastructure bottlenecks emerge.
- Nations that control AI-specific energy resources (nuclear, geothermal, renewables) will become the next AI infrastructure hubs. (Geopolitical AI Race)
- Enterprises that fail to lock in long-term power contracts will struggle to scale AI workloads competitively.
- AI energy efficiency will determine long-term AI viability.
- AI-powered grid management will be a key enabler of future AI expansion.
- China, India, and the U.S. are prioritizing AI energy expansion but without a particular emphasis on renewables, relying instead on nuclear, gas, and coal-powered grid scaling to meet AI infrastructure demands. (Geopolitical AI Race)

In an AI-driven world, energy availability will dictate AI scalability—companies must embed power procurement into their AI infrastructure strategy to maintain competitiveness.

6. Strategic Recommendations for Enterprises, Policymakers, and AI Infrastructure Providers

For Enterprises:

- Adopt a Red Queen AI strategy—constant optimization of AI infrastructure is required to avoid falling behind.
- Build AI-native data infrastructures—integrate real-time learning loops and minimize data bottlenecks.
- Secure long-term energy resilience—AI workloads must be optimized for power efficiency and sustainability.
- Invest in AI-native networking—control over data movement is as critical as control over AI models.

For Policymakers:

- Accelerate AI data center permitting—AI expansion is now a national economic and security issue.
- Invest in AI-specific energy infrastructure—grid modernization is essential to support AI's exponential growth.
- Support AI interoperability standards—ensure enterprises can deploy AI across multi-cloud and hybrid environments.

For AI Infrastructure Providers:

- Red Queen in AI networking—AI-native enterprises will demand low-latency, high-bandwidth AI-specific interconnects.
- Build adaptive AI workload allocation—future AI systems must dynamically adjust compute location based on real-time data needs.
- AI marketplaces will define the future of enterprise AI—develop scalable, modular AI platforms that allow seamless deployment and interoperability.

7. The AI-First Enterprise: Preparing for the Next Five Years

- AI success will not be about who builds the best models—it will be about who builds the best AI infrastructure.
- The AI race is accelerating—companies must constantly iterate to maintain their competitive edge. (Red Queen Effect)
- The next five years will separate AI leaders from AI laggards—those that fail to adapt will be left behind.
- Energy, data mobility, and AI-native networking will be the defining factors of AI success.
- AI is no longer about single breakthroughs—it is about continuous evolution, optimization, and scalability.

The AI leaders of the next decade will be those who understand that AI is not a product—it is a continuously evolving system that demands constant optimization. Enterprises that control AI infrastructure, energy, and real-time learning loops will shape the next generation of global AI dominance.

The AI revolution is not slowing down—it is accelerating. The question is no longer **who is adopting AI**, but **who is adapting fast enough to stay ahead**.

STELIA