vespertec

Embracing Enterprise AI: Navigating Budget Increases, Hardware Choices, and Market Forces

A Vespertec Whitepaper

Table of Contents

Executive Summary

AI, machine learning, and advanced data analytics have transformed enterprises in recent years—from sub-millisecond fraud detection systems in financial services to edge-based predictive maintenance in manufacturing, and medical imaging in healthcare.

But a new generation of AI, generative AI, is rapidly transforming enterprise and promising a revolution. For those organisations that get it right, it will fundamentally change business models, and offer a step change in capabilities versus competitors.

There is significant investment required in hardware to maximise this opportunity. It requires a new generation of custom, AI-optimised hardware. The cost of running this hardware in the cloud is pushing many to look at cloud repatriation.

Standard, off-the-shelf hardware is no longer sufficient for the scale and complexity of these tasks. Instead, enterprises are adopting custom, AI-optimised architectures—systems designed to maximise computational efficiency while minimising power consumption and thermal constraints.

For AI infrastructure, offloading complex tasks to GPUs like NVIDIA's H100 is now standard practice. The H100's up to 9x faster AI training and 30x faster AI inference speedups than the A100 allows for more intensive model training, crucial for enterprises processing massive datasets in real-time.

Simply buying cutting-edge hardware isn't enough—aligning it with existing systems, handling integration challenges, and ensuring scalability for future AI workloads are where enterprises often struggle.

Real-world deployment demands precision in infrastructure decisions: advanced liquid cooling systems, now deployed by Microsoft in some of its Azure data centres, starting with its facility in Quincy, Washington, can reduce server power consumption by 5% to 15%. Microsoft is also developing custom racks with integrated liquid cooling for its new AI chips.

'Ultimately, enterprise infrastructure decisions around AI have progressed beyond immediate performance gains to focus on designing systems that can adapt to the increasing complexity and scale of AI workloads.'

Allan Kaye, Co-Founder and Managing Director at Vespertec

Key Findings

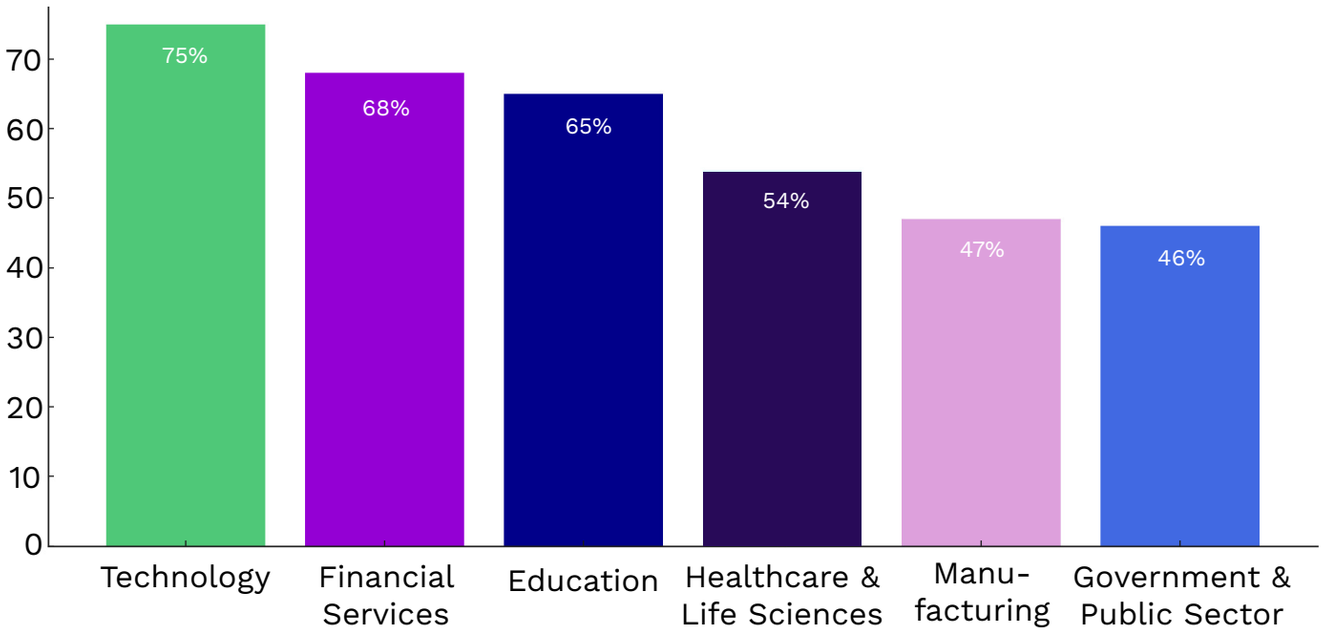
83% of organisations across multiple sectors have already deployed AI solutions, with an additional 15% actively planning implementation in the near future.

Remarkably, only 2% of respondents remain without plans to adopt AI, showing that this technology is now a central component of modern enterprise infrastructure.

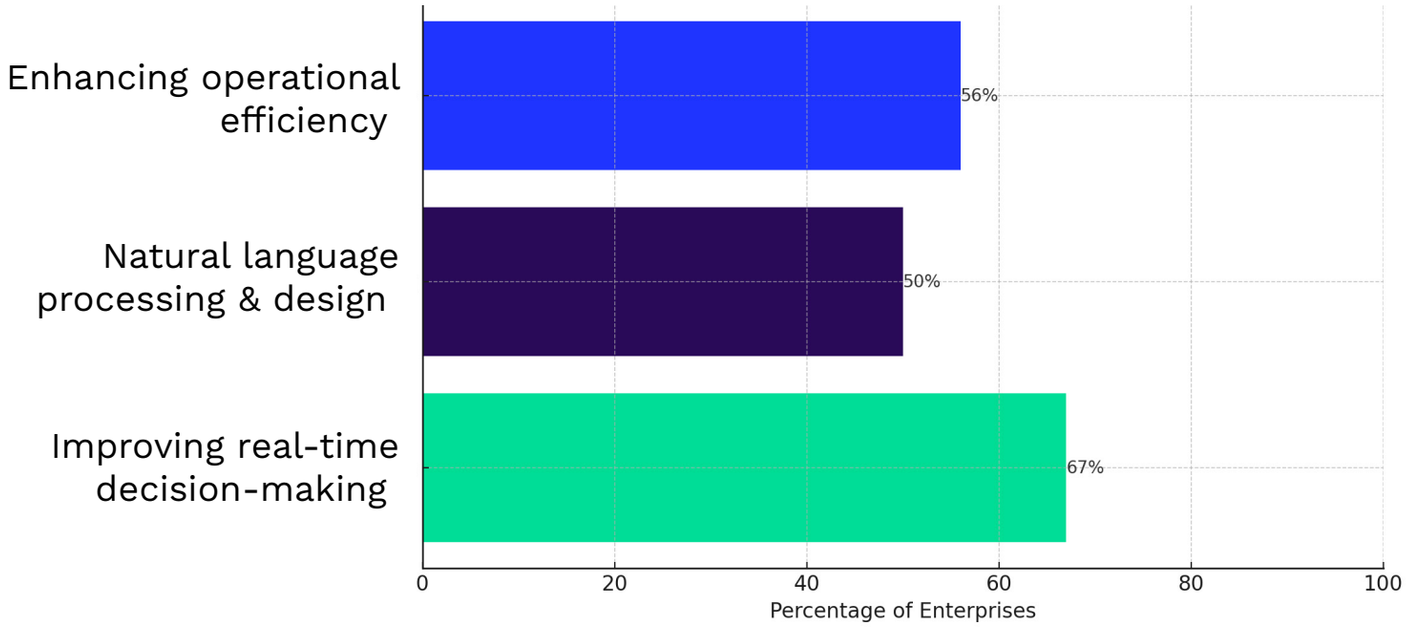
60% of those surveyed went further still, stating that AI is a major priority for them going forwards. Asking the same question by sector revealed some interesting insights:

Most organisations are prioritising AI as a concrete line item in their budget—and even more are at least incorporating it into their day-to-day operations in some way.

Is AI a major priority at your organisation?



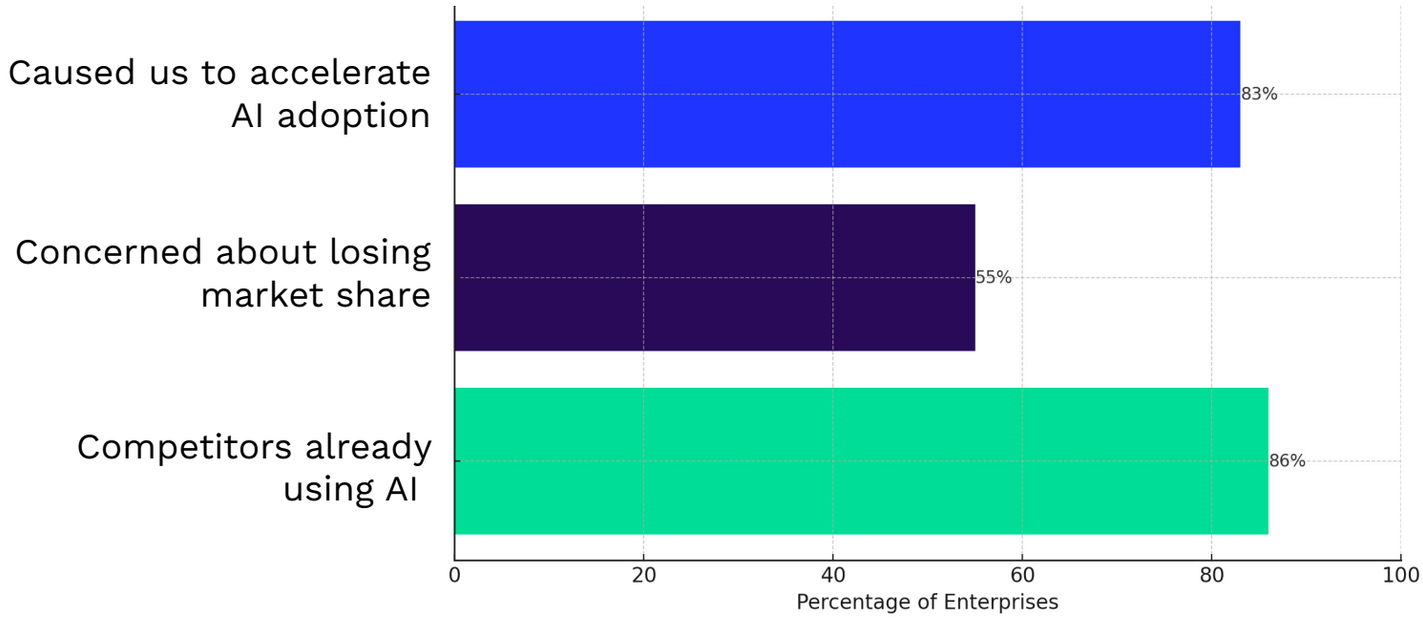
How do you plan to use AI?



- Two-thirds (67%) of AI-enabled organisations use Gen AI for tasks like natural language processing and creative design

Strategic decisions primarily focused on:
- Improving real-time decision-making (50% of enterprises)
- Enhancing operational efficiency (56% of enterprises)

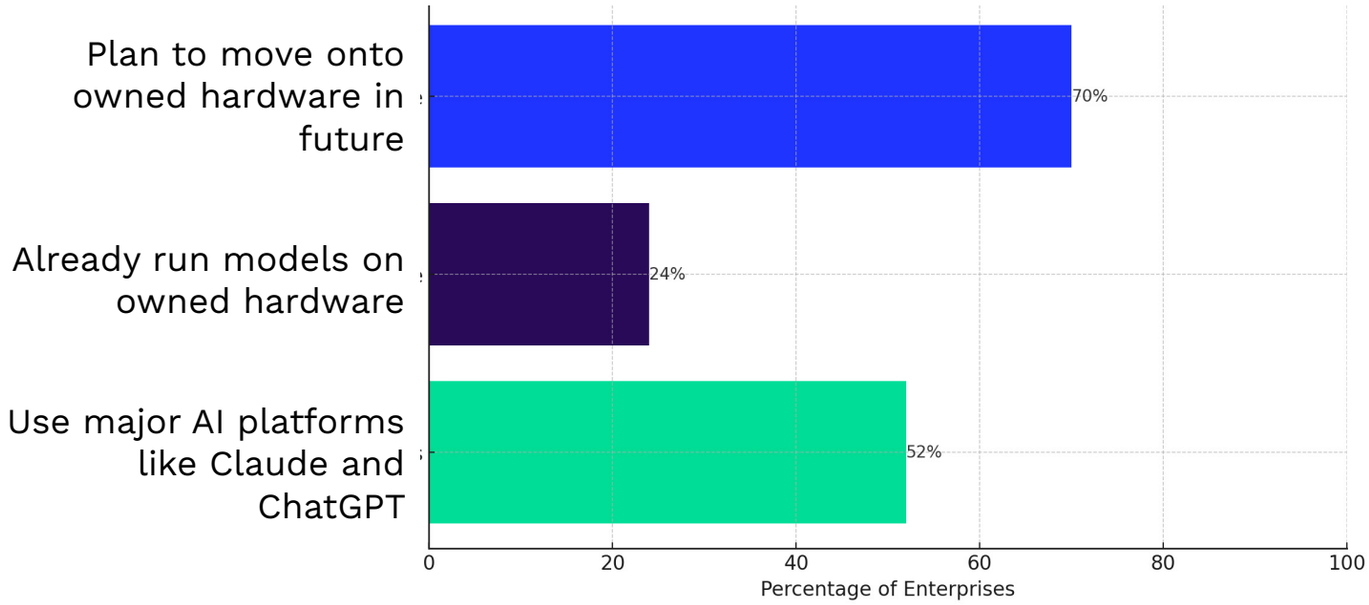
Competitive Concerns



- 86% of enterprises believe their competitors are already deploying AI
- 55% are worried about losing market share due to AI-driven competition
- 83% have accelerated AI adoption due to competitive concerns

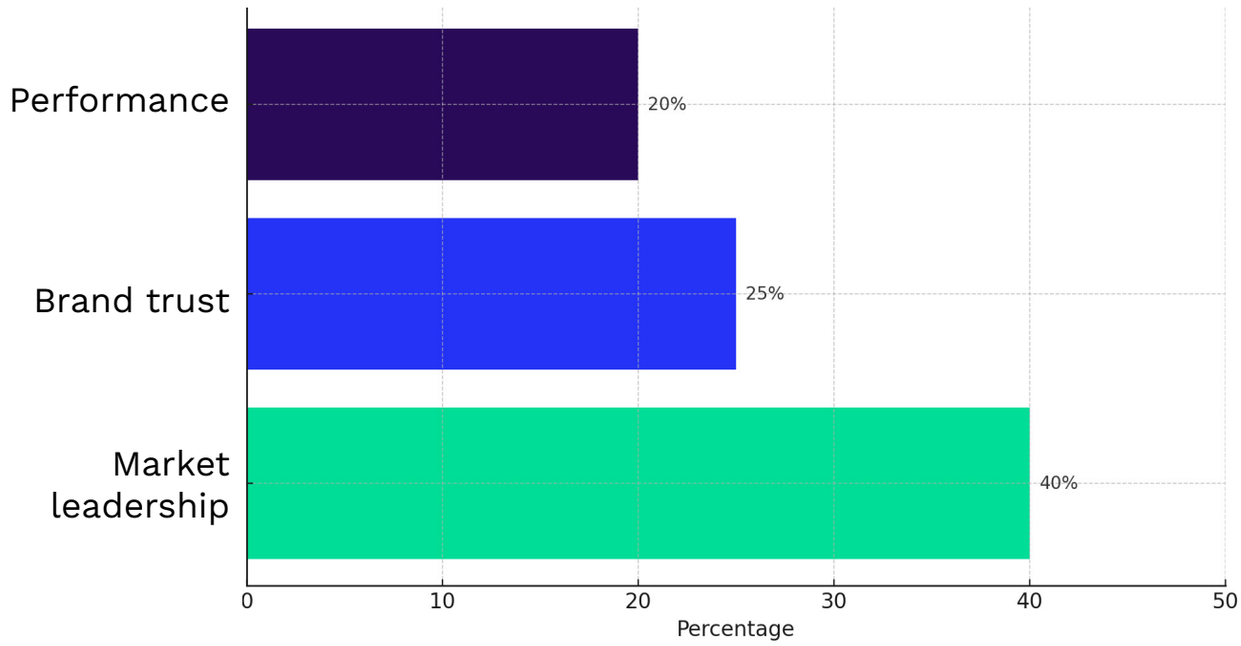
vespertec

How are you implementing AI?



- 52% currently use platforms like OpenAI, Claude, Microsoft, and Google
- 24% use models on their own hardware
- 70% of enterprises plan to move AI models to their own hardware in the near future, driving interest in private AI

Why would you prefer NVIDIA as your hardware provider?



NVIDIA remains the preferred AI hardware provider due to its:

- Market leadership (40%)
- Brand trust (25%)
- Performance (20%)

vespertec

Chapter 1: AI Budgets are Rising, with NVIDIA Leading the Hardware Market

Introduction

Our data shows 54% of enterprises reporting a 10% increase in hardware budgets. While 66% of this funding comes from reallocations, 32% comes from net new budget allocation.

Enterprises are moving beyond basic public cloud services, opting to build private, high-performance AI systems. As 70% of organisations aim to run bespoke AI models on-premises, maximising budget efficiency has become many firms' top priority.

Spending wisely demands a deep understanding of hardware choices. Enterprises must align their AI investments with use cases that demand not only computational power but also scalability and energy efficiency. That's why the question isn't just about increasing budgets—it's about which hardware brands enterprises trust and why.

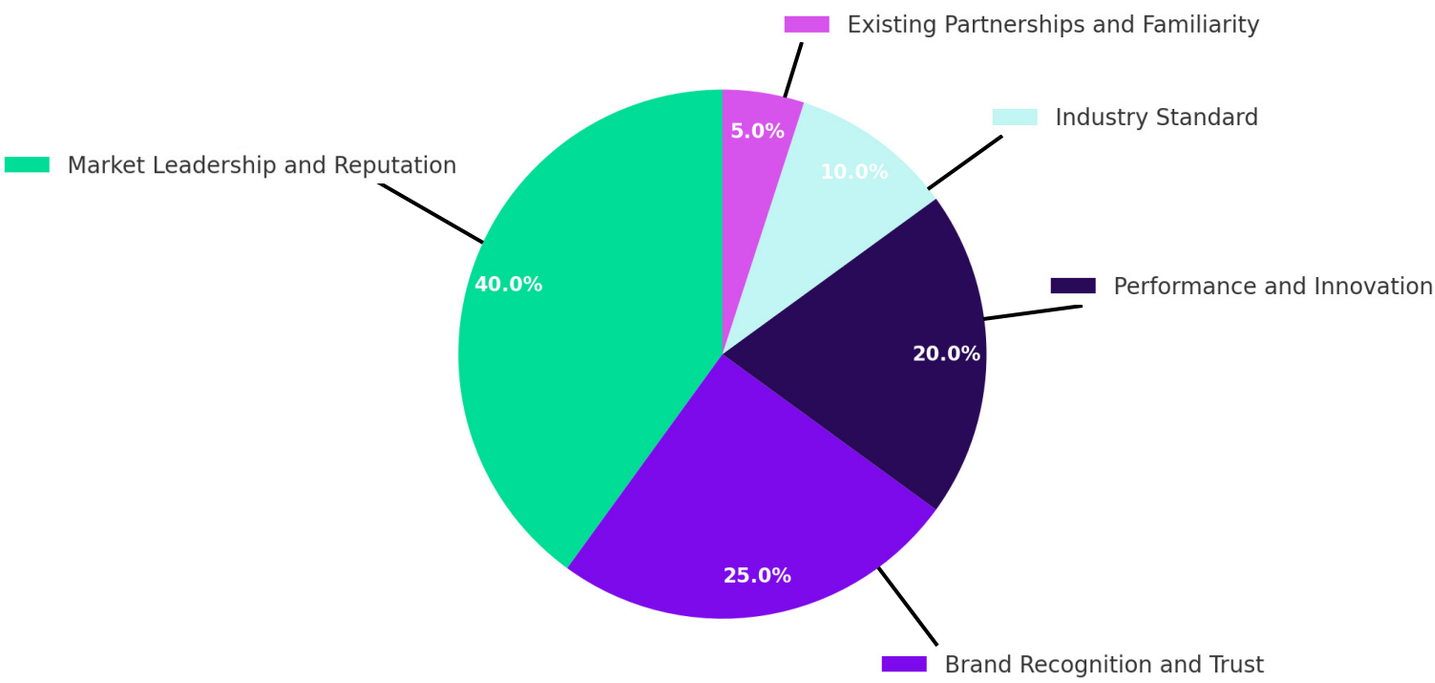
NVIDIA'S Market Leadership

NVIDIA dominates the AI hardware market, powering more than twice (40%) as many enterprise AI deployments as its nearest competitor.

A key driver of this leadership is its unmatched performance in training and inference, particularly with its H100 Tensor Core GPUs and DGX systems.

The H100, featuring 80 billion transistors, 3.35 TB/s memory bandwidth, and 7.9 PFLOPS of FP8 precision, offers up to 4X faster training for large-scale models like GPT-3 compared to its predecessor. The integration of NVLink provides 900 GB/s of GPU-to-GPU interconnect, enabling scalable performance for exascale workloads and next-gen AI applications like trillion-parameter language models.

Reasons enterprises choose NVIDIA



Implementation & Use Cases

NVIDIA's supremacy is evident in real-world deployments like Meta's AI Research SuperCluster (RSC), which:

- Integrates 16,000 NVIDIA GPUs.
- Delivers up to 5 exaflops of AI performance.

The RSC and systems like it are built to handle large language models (LLMs) with trillions of parameters, underscoring the kind of infrastructure needed for enterprises developing complex AI models. In fact Meta has since announced even larger clusters with 24,576 NVIDIA H100 GPUs each, which are being used for training newer models like Llama 3.

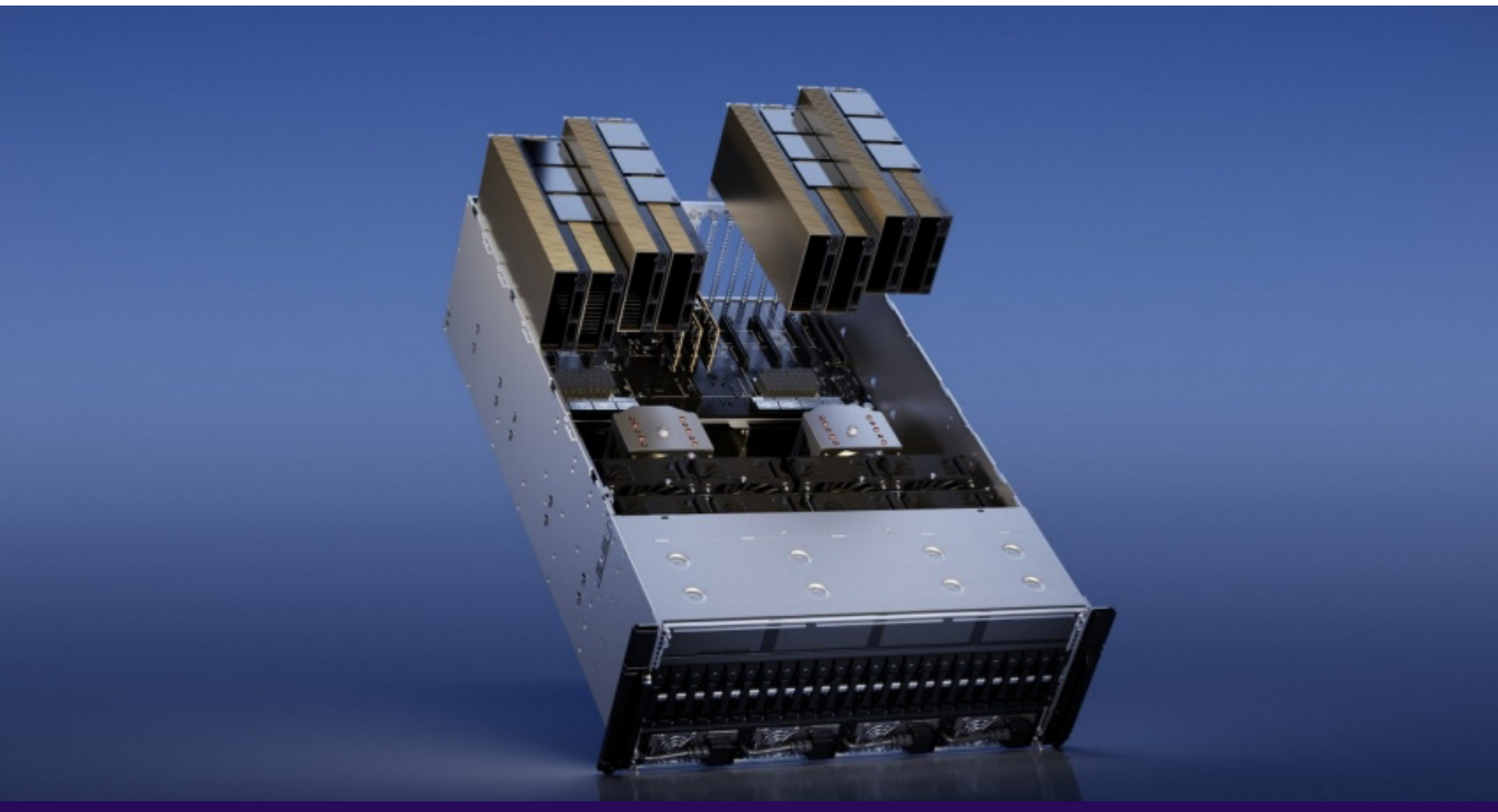
Key example:

The University of Southampton upgraded to the HPE Iridis 6 cluster, powered by:

- NVIDIA H100 GPUs
- AMD EPYC processors

Benefits:

- Handles massive data sets in AI research (e.g., space debris analysis)
- Uses 70% less power than its predecessor, ensuring sustainable AI workload scaling



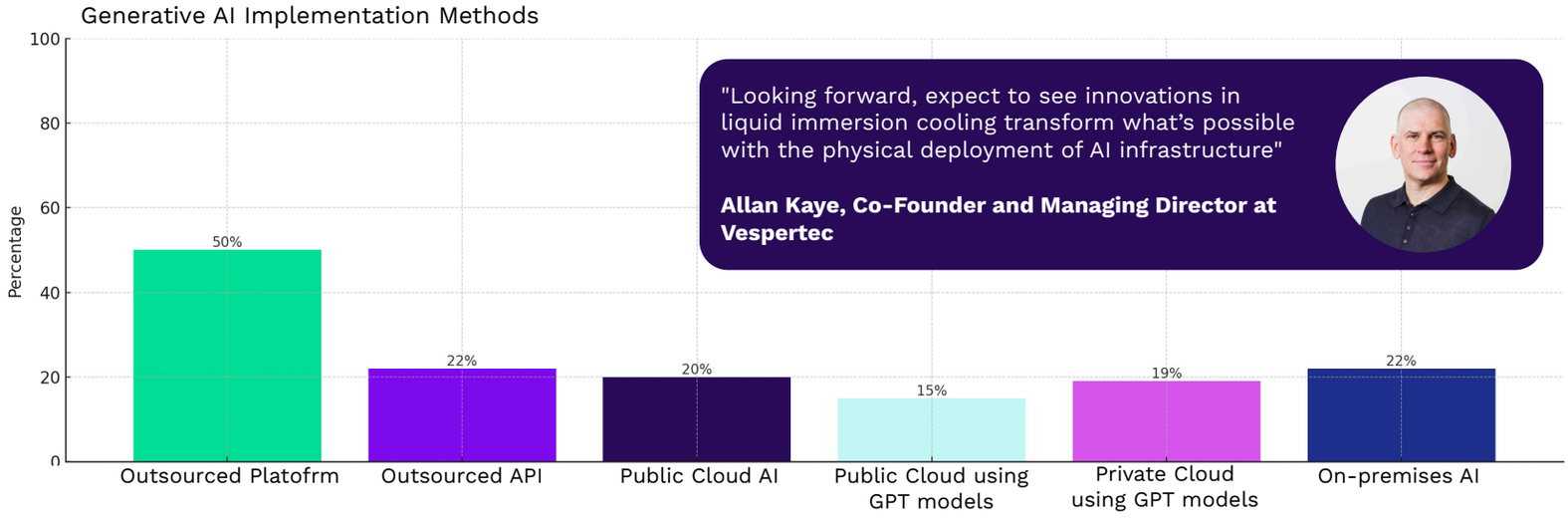
The following further illustrates why enterprises are choosing NVIDIA over its competitors for AI workloads:

- MLPerf benchmark dominance: NVIDIA H100 GPUs set new records across all eight tests in the 2023 MLPerf AI benchmarks, showcasing the hardware's ability to handle diverse AI tasks like language generation, computer vision, and speech recognition with ease.
- Industry backing: Nearly a dozen major companies, including ASUS, Dell, and Lenovo, submitted benchmark results using NVIDIA platforms, demonstrating the wide industry reliance on NVIDIA's technology.
- Increased production: NVIDIA plans to at least triple the output of its compute GPUs in 2024. The projected H100 shipments for 2024 are estimated to range between 1.5 million and 2 million units, up from an anticipated 500,000 units in 2023.

Future trends in AI hardware investment

As 70% of enterprises prepare to run AI models on their own hardware, choosing the right infrastructure has never been more important. Looking forward, expect to see innovations in liquid immersion cooling transform what's possible with the physical deployment of AI infrastructure.

OVHcloud's liquid cooling system, implemented at Data4's Marcoussis campus has already resulted in a 25% reduction in power consumption compared to standard air cooling. This allowed OVHcloud to increase server density without sacrificing performance, even under high AI and HPC workloads. The setup has also significantly improved the site's water usage efficiency, achieving a WUE of 0.06 l/kWh IT, making it one of the most sustainable data centers in Europe. This level of efficiency has been crucial in meeting the increased compute demands without escalating energy costs or environmental impact.



"Looking forward, expect to see innovations in liquid immersion cooling transform what's possible with the physical deployment of AI infrastructure"

Allan Kaye, Co-Founder and Managing Director at Vespertec

vespertec

Chapter 2: Enterprise Adoption of High-Performance AI Systems

Our data shows that 56% of enterprise IT decision-makers are using AI to enhance their products or services, while 50% are using it to improve decision-making. The demand for hyper-efficient, scalable AI systems has led enterprises to adopt solutions like the NVIDIA DGX H100 and GH200 Grace Hopper Superchip, pushing the boundaries of AI performance.

Sector-specific adoption

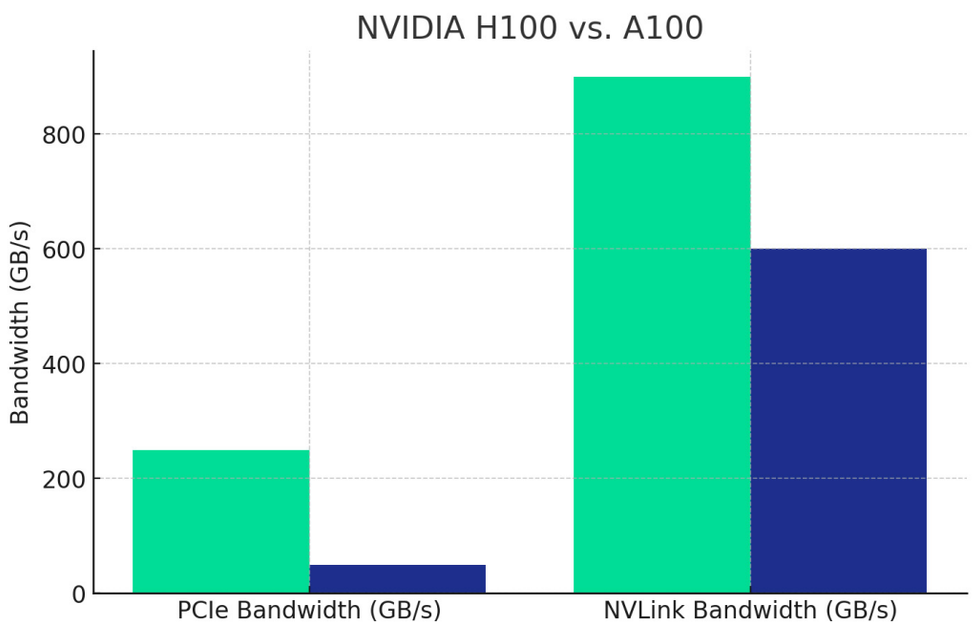
Three key sectors are leading the charge for performance-focused AI and HPC workloads: financial services, manufacturing, and healthcare.

Financial services

Case studies
AI-driven chatbots and virtual assistants—like Bank of America's Erica, which handled 50 million requests in 2019—are becoming central to customer experience.

Similarly, OCBC Bank's AI chatbot trial demonstrated a 50% increase in efficiency, showcasing how generative AI is redefining service delivery at scale.

Deployment
Systems like the NVIDIA DGX H100, with 8 H100 Tensor Core GPUs, provide the compute density and low-latency required for high-frequency trading and fraud detection. With 640GB of GPU memory and fourth-generation NVLink's 900 GB/s bidirectional bandwidth, this system ensures microsecond-level response times, essential for financial algorithms.



Manufacturing

Case studies
Siemens plans to achieve a 30% reduction in machine downtime using AI-driven predictive maintenance, while Foxconn used NVIDIA Omniverse for real-time factory simulations and robot training, speeding up production cycles and improving productivity.

Deployment
NVIDIA Jetson modules, such as the Orin NX, with up to 100 TOPS of AI performance, can optimise edge applications like predictive maintenance and quality inspections, while DGX systems can handle the complex simulations and large-scale data analysis required in tasks like creating digital twins.

Healthcare

Case studies
Mount Sinai, in collaboration with IBM, is leveraging AI and behavioural data to personalise mental health care for young people, predicting outcomes like treatment dropouts and hospitalisations. Meanwhile, Mayo Clinic is using IBM's Watson to match cancer patients with clinical trials more quickly, enhancing treatment options and accelerating research outcomes.

Deployment
NVIDIA's Clara is a GPU-accelerated platform designed for healthcare AI. It supports massive parallelism for deep learning in medical imaging, real-time edge processing for medical devices, and ultra-fast genomics pipelines using CUDA cores and TensorRT. Clara's optimised CUDA-X libraries and pretrained models slash latency in data-heavy tasks, making it perfect for AI-driven diagnostics, drug discovery, and genomics.

AI's Strategic Role in Driving Enterprise Efficiency

Enterprises are also adopting the NVIDIA GH200 Grace Hopper Superchip, combining CPU and GPU capabilities with 900GB/s coherent memory bandwidth and 480GB of LPDDR5X CPU memory and 96GB of HBM3 GPU memory to handle highly complex simulations.

These systems deliver up to 10x performance gains for large-scale analytics, transforming workloads in risk management, compliance, and environmental monitoring into real-time, precision-driven tasks.

Chapter 3: Competitive Pressures Add Fuel to AI Adoption Fire

86% of enterprises believe their competitors are already leveraging AI, so decision-makers are ramping up investments in AI infrastructure to avoid being outpaced. In fact, 83% of organisations facing competitive threats report accelerating their AI initiatives to maintain market position and operational efficiency.

This all contributes to an environment where global AI spending is projected to reach $279.22 billion in 2024, driven by firms like Microsoft and BlackRock, which launched a $100bn fund in September 2024 to invest in AI data centres and the energy infrastructure needed to power AI workloads.

Strategies to stay competitive

To remain competitive, enterprises are adopting a mix of in-house AI infrastructure and outsourced AI platforms

- 52% of companies use external platforms like OpenAI and Google Cloud to achieve scalability without major CAPEX investments.
- 24% of enterprises opt for on-prem AI infrastructure to maintain tighter control over data and compliance, particularly in sectors with stringent regulations such as finance and healthcare.

Public cloud AI services, used by 21% of enterprises, provide flexibility, but private cloud solutions—chosen by 19%—offer the security and customisation needed for critical workloads.

Looking ahead

NVIDIA's DGX H100 and GH200 Superchip solutions, along with TensorRT and CUDA-X AI optimisations, are driving the performance enterprises need to process petabytes of data in real-time. For industries like manufacturing, predictive maintenance systems built on Jetson modules are already reducing downtime by up to 30%, showcasing AI's capacity to transform operational efficiency.

Conclusion

As AI workloads grow, enterprises are optimising their hardware spend to ensure a precise balance of compute power, memory efficiency, and scalability to handle complex tasks like fraud detection and deep learning diagnostics.

NVIDIA's DGX H100 and GH200 Grace Hopper Superchip are leading solutions, built to support large-scale AI models while minimising latency and maximising throughput. But it's not just about processing power—advanced cooling systems, such as immersion or direct liquid cooling (DLC), are critical for maintaining performance under load, especially for AI-driven operations requiring near-overclocked hardware.

The shift to in-house AI infrastructure reflects the need for tighter control over data, compliance, and performance, particularly in sectors like finance and healthcare.

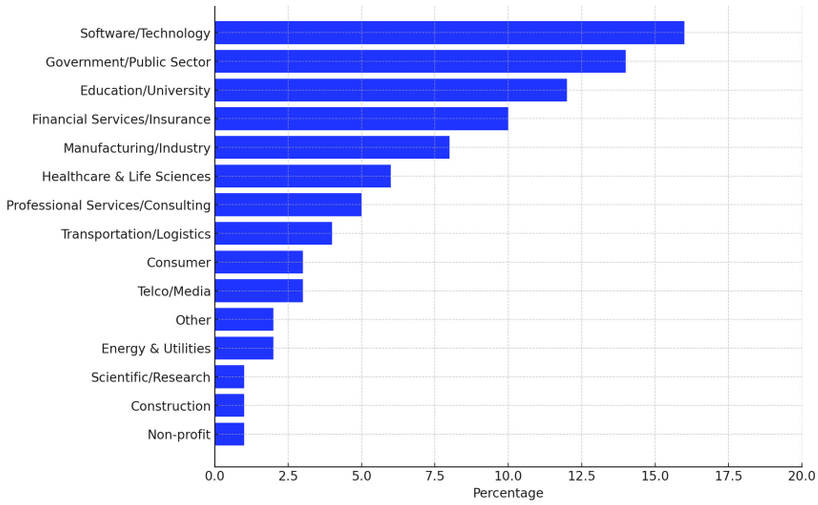
Open infrastructure standards provide flexibility, reducing the risks of vendor lock-in, while custom ASICs and AI-accelerated hardware will increasingly dominate high-efficiency tasks.

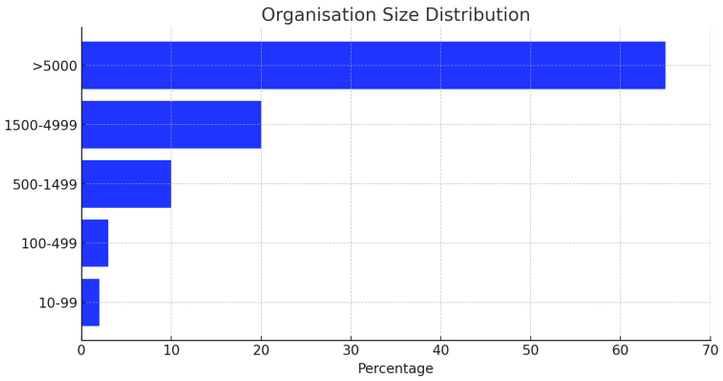
The competitive landscape is clear—86% of enterprises believe their competitors are already leveraging AI. This is driving rapid adoption of AI hardware, with 70% prioritising private AI systems. The key to staying ahead lies in precision engineering—building infrastructure capable of scaling with AI's growing demands while maintaining energy efficiency and cost control.

Enterprises that invest in these cutting-edge systems now will ensure their AI capabilities remain competitive, adaptable, and efficient in the long term.

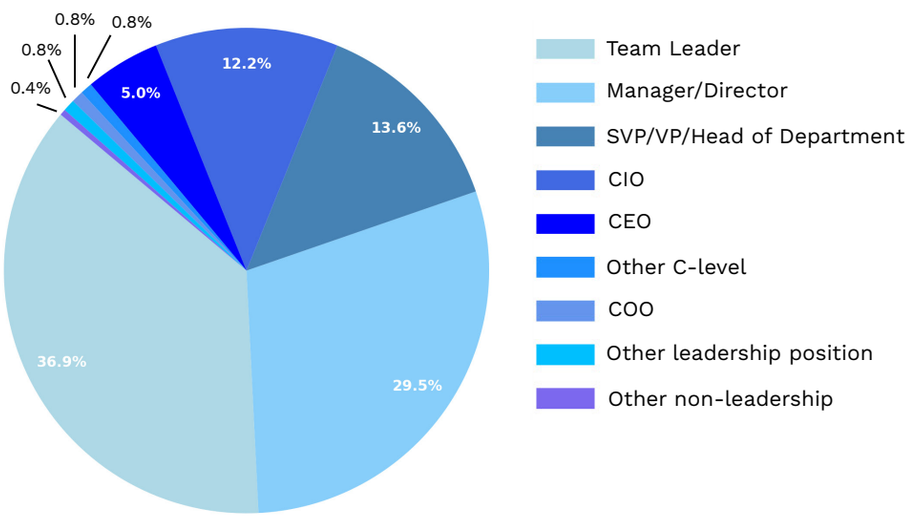
About the Whitepaper

This survey was run on a representative sample of 502 senior IT decision-makers across various industries, including software/technology, financial services, healthcare, manufacturing, and the public sector.





Seniority breakdown



The sample included both large enterprises with over 5,000 employees and smaller organisations, ensuring a balanced perspective on the challenges and opportunities presented by AI.

Respondents included Chief Technology Officers, Chief Information Officers, and other senior IT decision-makers, providing a comprehensive view of current AI adoption trends and the strategic considerations influencing these decisions.

vespertec

Unit 5 Rugby Park, Bletchley Road,
Stockport, SK4 3EJ, United Kingdom

+44 (0) 161 947 4321

info@vespertec.com

Find out more today

For more information, including benchmarking test details, please contact Vespertec

Call +44 (0) 161 947 4321
or email info@vespertec.com now.