# Enterprises set to save millions as edge AI overtakes cloud by 2025



The transition from cloud-centric models to edge AI deployment is becoming increasingly prominent in enterprise settings, offering substantial potential for cost savings and enhanced scalability. An economic analysis by Latent AI underscores this shift, particularly as 2025 approaches—a year predicted to be pivotal in the evolution of enterprise AI. Factors contributing to this trend include the rising costs associated with cloud computing, persistent shortages of GPUs, and escalating energy prices. These elements make the allure of edge AI—where data is processed locally—more attractive to organisations aiming to remain competitive.

Latent AI, a company renowned for developing edge AI solutions specifically tailored for national security and defence, performed a detailed study titled "From Cloud-First to Edge-First: The Future of Enterprise AI". It reveals that organisations must now analytically compare performance gains against the infrastructure investments associated with cloud and edge strategies. Jags Kandasamy, CEO and Co-founder of Latent AI, commented on this challenge, stating, "As enterprise leaders scale AI deployments, they must weigh performance gains against infrastructure investments. With tighter budgets and growing demands for real-time processing, organisations can no longer afford the heavy computational costs of cloud-only solutions." This perspective aligns with broader industry sentiments regarding the financial burdens of traditional cloud models.

The analysis provided concrete examples, including a manufacturing firm dependent on a cloud-based AI system that utilised 50 GPUs for processing image streams for defect detection. This firm faced exorbitant hardware costs—USD $224,000 per site—which hindered its expansion plans across multiple sites. By switching to edge AI optimisation and advanced quantisation techniques, the GPU requirement plummeted from 50 to just four per site. This drastic reduction resulted in hardware expenditures dropping to USD $18,000 per deployment, translating to savings of USD $207,000 per site, or a staggering total of USD $2.07 million if replicated across ten locations. Such dramatic cost reductions highlight the transformative potential of edge technology.

Moreover, the benefits of this shift extend beyond hardware savings. The latency in inference speed improved dramatically, with defect detection time decreasing from 55.2 milliseconds to just 14.7 milliseconds—an impressive 73% boost. Memory usage also saw a notable drop of 73%, illustrating edge AI’s capability to operate efficiently within constrained resources. This efficiency not only contributes to operational cost savings but also enhances overall performance, allowing firms to implement AI solutions across a broader array of applications.

The drive towards adopting edge AI technology is also fuelled by optimisation technologies like quantisation and pruning, which enhance model performance while reducing resource requirements. Companies such as NVIDIA are developing AI accelerators tailored for edge devices, facilitating powerful AI processing in resource-constrained environments. These innovations are essential in industries where real-time analysis is vital, especially in mission-critical applications where reliability is paramount.

Another key driver of this transition is the evolving landscape of data privacy regulations. New guidelines, reminiscent of GDPR, mandate local data processing to mitigate compliance issues related to data transmission. Such regulations are prompting organisations to reassess their cloud strategies, enabling a more streamlined approach that aligns with both legal requirements and operational efficiency.

Kandasamy emphasises the ongoing nature of this technological shift: "Technology shifts don't happen overnight. They build momentum until a tipping point emerges." He posits that 2025 will mark a significant inflection point for edge AI, paralleling the rise of cloud computing in the early 2000s. Such momentum is indicative of a broader convergence of technological maturity, economic pressures, and market demands, driving rapid adoption of edge AI solutions.

For organisations keen to capitalise on the cost advantages of edge AI, the report offers strategic recommendations. Investment in infrastructure that is specifically tailored for edge deployments, such as servers and IoT devices capable of scaling efficiently, is crucial. Moreover, optimising AI models for low-latency, energy-efficient performance at the edge is essential. This balanced approach allows firms to leverage both cloud and edge resources effectively—utilising the cloud for training while conducting inference at the edge to manage ongoing costs.

In an era where businesses are increasingly seeking sustainable and efficient technological solutions, embracing edge AI could be fundamental not only for savings but also for maintaining a competitive edge in the market. As organisations navigate this transition, the emphasis on maximising operational efficiencies and adhering to evolving regulations will likely shape the future landscape of enterprise AI.

## Reference Map:

* Paragraph 1 – [[1]](https://cfotech.asia/story/edge-ai-deployment-enables-enterprises-to-save-2-07-million), [[2]](https://builtin.com/artificial-intelligence/edge-ai-energy-solution)
* Paragraph 2 – [[1]](https://cfotech.asia/story/edge-ai-deployment-enables-enterprises-to-save-2-07-million), [[3]](https://www.barbara.tech/blog/edge-ai-business-models-drive-tangible-value)
* Paragraph 3 – [[5]](https://www.restack.io/p/ai-for-edge-computing-answer-cost-analysis-cat-ai), [[6]](https://www.wevolver.com/article/2023-edge-ai-technology-report-chapter-ii-advantages-of-edge-ai)
* Paragraph 4 – [[4]](https://developer.nvidia.com/blog/building-an-edge-strategy-cost-factors), [[6]](https://www.wevolver.com/article/2023-edge-ai-technology-report-chapter-ii-advantages-of-edge-ai)
* Paragraph 5 – [[7]](https://www.wevolver.com/article/2023-edge-ai-technology-report-chapter-ix-challenges-of-edge-ai)
* Paragraph 6 – [[1]](https://cfotech.asia/story/edge-ai-deployment-enables-enterprises-to-save-2-07-million), [[3]](https://www.barbara.tech/blog/edge-ai-business-models-drive-tangible-value)
* Paragraph 7 – [[1]](https://cfotech.asia/story/edge-ai-deployment-enables-enterprises-to-save-2-07-million)

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://cfotech.asia/story/edge-ai-deployment-enables-enterprises-to-save-2-07-million> - Please view link - unable to able to access data
2. <https://builtin.com/artificial-intelligence/edge-ai-energy-solution> - This article discusses how edge AI can address the energy challenges associated with cloud computing. By processing data locally, edge AI reduces the need for data transmission, leading to energy savings of 65% to 80% compared to cloud solutions. Techniques like model quantization and pruning enable edge AI to run on resource-constrained devices, further enhancing efficiency. The article also highlights a case study where a manufacturing company reduced memory usage by 73% and improved inference speed by 73% while maintaining model accuracy, demonstrating the practical benefits of edge AI in real-world applications.
3. <https://www.barbara.tech/blog/edge-ai-business-models-drive-tangible-value> - This blog post explores how edge AI can drive tangible value for businesses by reducing costs and increasing profit margins. It highlights that organizations investing in cloud computing are now exploring cost-saving opportunities by partially shifting applications to edge computing. The article provides examples, such as a manufacturing company that experienced a 30% reduction in operating costs by implementing edge AI for inspection processes. Additionally, it discusses how edge AI enables faster and more accurate analysis in various industries, leading to improved results and forecasts.
4. <https://developer.nvidia.com/blog/building-an-edge-strategy-cost-factors> - This blog post from NVIDIA discusses the cost factors associated with building an edge strategy. It emphasizes that while edge computing requires an upfront investment, it can lead to significant cost savings in the long term. The article provides examples, such as a manufacturer that reduced inspection costs from 30% of total manufacturing costs by using AI optical inspection in the factory. It also discusses the value of edge computing in scenarios requiring real-time responses or deployment in remote locations with limited bandwidth, highlighting the importance of edge AI in such contexts.
5. <https://www.restack.io/p/ai-for-edge-computing-answer-cost-analysis-cat-ai> - This article provides a cost analysis of AI for edge computing, highlighting the financial benefits of implementing edge AI solutions. It discusses considerations such as infrastructure investment, scalability, and security measures when adopting edge AI. The article presents a case study of a manufacturing company that implemented edge AI for predictive maintenance, resulting in a 30% reduction in downtime and an estimated annual savings of $500,000. This demonstrates the tangible financial benefits that can be achieved through effective edge computing strategies.
6. <https://www.wevolver.com/article/2023-edge-ai-technology-report-chapter-ii-advantages-of-edge-ai> - This article outlines the advantages of edge AI, including reduced latency, power efficiency, and cost-effectiveness. It explains that edge AI applications limit data transfers over wide area networks by processing data close to the source, resulting in faster data processing and reduced latency. Additionally, edge AI reduces the number of I/O operations and processes data within edge devices or edge data centers, leading to improved overall CO2 footprint for AI applications. The article also highlights how edge AI economizes on network bandwidth and computing resources, making AI applications more cost-effective.
7. <https://www.wevolver.com/article/2023-edge-ai-technology-report-chapter-ix-challenges-of-edge-ai> - This article discusses the challenges associated with edge AI, including hardware and software costs, data storage and management, and network connectivity. It highlights that edge devices often require specialized hardware components, such as specific CPU and GPU hardware, which can increase costs. The article also mentions the need for robust network infrastructure to support real-time processing and analysis, adding to infrastructure costs. Additionally, it emphasizes the importance of data privacy and security, as edge devices collect sensitive data that must be protected from breaches.