Is AI accelerating the edge opportunity?

Part one of a four-part series where we examine exactly how the advent of a new wave of Large Language Models (LLMs) will change the edge opportunity. We explore the AI use cases that edge computing could help to facillitate, the use of AI in edge orchestration, and whether there will be new entrants to the edge market.

David Gordon, Senior Consultant

July 2023

Will AI use cases fuel edge deployments?

With a lot of hype around AI since the launch of Chat GPT in 2022, we want to explore and evaluate the role that AI will play in accelerating the market for edge computing. AI, particularly through large language models (LLMs) and generative AI, is touted to have the capacity to revolutionise operations across various sectors, driving efficiency, and innovation.

The impact of AI can be profound within different industries, for instance:

- In healthcare – AI finds application in a myriad of ways, such as Medical Imaging Analysis, enabling precise disease diagnosis from medical scans, and Predictive Analytics for early disease detection and patient risk assessment.
- In finance – Financial institutions benefit from AI in areas like Fraud Detection, where the technology identifies and prevents fraudulent transactions, and Algorithmic Trading, automating data-driven investment decisions.
- In retail – Retail industries utilise AI to deliver Personalised Recommendations, enhancing customer shopping experiences, and employ AI-Powered Visual Search for efficient product discovery.

AI models enable enterprise to create smart outcomes by processing massive volumes of data. When combined with real-time response, processes can be automated, and outcomes made more efficient than traditional methods of human management and operation. The benefits of edge computing for use with AI can be understood in terms of the advanced computational power and speed offered by edge computing as well as the real-time and low latency speeds offered at the edge.

- **Computing power** – Implementing large language models and generative AI requires substantial data storage capabilities and increased computing power. This is where edge computing comes into play, offering higher bandwidth and ultra-low latency, ideal for the demands of generative AI.
- **Real-time** – Edge computing's proximity to data sources and real-time processing capabilities complements AI applications, providing a seamless and efficient experience.

Currently, the hype surrounding AI is centred on a handful of large scale, generalist LLMs. These are centrally controlled by a few organisations, with input data being added to the training data. As we are starting to see from major organisations (see Reuters, Goldman Sachs, etc), some enterprises are recognising the need to create their models, leveraging their extensive historical data to build more specific and secure algorithms.

As industries continue to explore the transformative potential of AI, the integration of edge computing becomes a vital factor in unlocking the full value of these advanced technologies. By combining AI use cases with the power of edge computing, organisations can embrace innovation, enhance operational efficiency, and create unique opportunities for growth and development.

What is AI's role in orchestrating edge platforms?

Advanced AI presents a myriad of possibilities for improving edge orchestration platforms across various fronts. Firstly, intelligent resource management becomes achievable through the analysis of real-time data from edge devices and cloud resources. New NLMs and AI algorithms intelligently allocate computing power, storage, and networking resources, dynamically adjusting allocations based on workload demands. This optimisation fosters better platform performance and cost efficiency.

AI enhances workload placement decisions by factoring in latency, bandwidth, and data sensitivity. Edge platforms can make informed choices on placing workloads on the most suitable edge devices or cloud resources for optimal execution.

AI-driven predictive maintenance and anomaly detection fortify edge platform performance. By proactively detecting potential issues and anomalies in edge devices. AI helps to minimise downtime, optimise device performance, and ensure uninterrupted operations.

Resource management, workload placement and predictive maintenance are just a handful of AI use cases within edge orchestration. AI can also play a role in dynamic network configuration, enhancing security, minimising reliance on a central cloud and the enforcement of policy for regulatory and organisational compliance.

Will growing demand for AI in enterprise augment the edge provider ecosystem?

The hype surrounding Generative AI and large language models has resulted in a significant surge in The recent surge in demand for Generative AI and large language models has led to a remarkable boom in the need for Graphics Processing Units (GPUs), propelling Nvidia's market capitalization to a staggering $3 trillion. As AI and NLMs require substantial computational power, GPUs have emerged as a critical component to support their data-intensive operations. Given this exponential growth and the continuous evolution of AI technologies, it raises a compelling question: Could Nvidia leverage its expertise in GPUs and embrace the edge infrastructure domain to capitalize on the burgeoning opportunities in processing data for AI?

Increasing their edge presence could open new avenues for Nvidia to enhance its role in supporting AI operations on a global scale. Edge computing's proximity to data sources ensures faster data processing and real-time insights, crucial for AI-driven applications in areas like autonomous vehicles, smart cities, and industrial automation. As the GPUs continue to be installed across a range of data centres, edge infrastructure will be integral to the development of the AI ecosystem.

We are already starting to see an increase in the demand for edge platforms as a result of the interest in AI. Many enterprise customers, having been investigating and testing edge solutions for many years, are now at the stage of deploying solutions. With AI, this ecosystem is becoming more and more diverse, involving SaaS players, security providers, and many more who are looking to develop their AI offering on fast, secure infrastructure.

We will continue to explore this topic in further articles. If you have any questions, please reach out to our team at edge@stlpartners.com.

David Gordon is a Senior Consultant at STL Partners, specialising in private networks, edge and B2B growth strategies.

Get in touch with the author to learn more.

edge@stlpartners.com

Or visit STL Partners' Edge Hub

www.stlpartners.com/edge-computing