Perspective paper

POWERING AI, FOR TODAY AND TOMORROW

How ASUS server and infrastructure solutions are solving the biggest AI challenges facing organisations today – and those yet to arrive.

AI IS CHANGING THE GAME; BUT FOR SOME, INNOVATION IS STILTED

AI has fast become a part of our vocabulary, not just in business but in our everyday lives, too. It's unsurprising, given its magic-wand appeal to enable more output for less effort in less time. Streamlining tasks. Boosting productivity. Enabling more informed decisions based on accurate, predictive insights. It's personalising experiences. Bringing innovative ideas to life – at speed.

In manufacturing, anomalies can be detected on the production line before they affect the end customer. In transportation, AI can process sensor data and make critical decisions in real time for autonomous vehicles. And in healthcare we're seeing AI speed up the diagnosis processes, such as screening for lung cancer,[1] so patients receive life-saving treatment earlier.

But, with innovation comes responsibility.

And while AI can very nearly do it all, it does come with significant hurdles; to unleash the full potential of AI, huge processing power is required. Without a clear strategy and effective solution from the outset, organisations will be faced with AI deployments that can't fulfil their potential. This not only affects the organisation, but those its services reach, meaning end-users might not benefit from life-altering scientific advancements, for example, in time.

1 IN 6

UK organisations, totalling 432,000, have embraced at least one AI technology.[2]

68%

of large companies,

33%

of medium-sized companies, and

15%

of small companies in the UK have incorporated at least one AI technology.[3]

YET AI IS STILL EVOLVING, AND PROCESSING REQUIREMENTS ARE GROWING

AI is moving at a rate that's hard to keep pace with. The technology needs a complex infrastructure of software and servers to enable fast access to huge volumes of data to enable intricate tasks – from machine learning and deep learning (a machine learning technique that uses vast amounts of data to build layers of understanding), to large language models (LLMs).

Then there's the rise of generative AI (GenAI), with McKinsey Global Institute estimating that GenAI will add between $2.6 and $4.4 trillion in annual value to the global economy (increasing the economic impact of AI by 15-40%). The firm also projects that AI will automate half of all work by 2040 – 2060, with GenAI pushing that window a decade earlier than previous estimates.[4]

Open source LLMs are also coming into play, like Meta AI's Llama 3.2.[5] And while these increase transparency, customisation and cost-effectiveness compared with closed-source alternatives, they require increased expertise in model management, maintenance, governance and hardware infrastructure instead.

With all this (and more than we can fit here) happening in the AI arena, it's no surprise that the computing power required for AI is doubling every 100 days and is projected to increase by more than a million times over the next five years.[6] Keeping up with such rapid increases in computational capacity requirements is a challenge, and one that's seeing the rise of supercomputers, to help maximise the impact of AI.

The processing challenges that must be addressed to fully unlock the potential of AI for all organisations:

- Eliminating the latency that can compromise increasingly complex LLM, ML, and DL models and techniques, and addressing the need for exceptionally high-performance GPU compute
- Accommodating the move of data back on-premises as organsations consider the security implications of keeping confidential data in the cloud
- Managing high volumes of data and moving processing to the Edge as organisations move to decrease latency and bring better performance to the user
- Focus on resiliency of high-performance compute for the real-time decisioning nature of AI applications

HOW SUPERCOMPUTING IS REVOLUTIONISING AI

The supercomputer has become a critical solution for many industries shaping a future in AI – across research and academia, from healthcare to sciences and biotechnology. There are more of these hugely powerful devices than people realise, with the three most powerful based in the United States[7], but hundreds more existing and growing in number across the globe.

AI supercomputers mark a new era of computing – simplifying programming and supporting the larger memory requirements of AI tasks. Made up of countless processors and finely tuned hardware, AI supercomputers provide fast, scalable computing power, ample storage, and secure networking – easily overcoming the afore-mentioned challenges of AI.

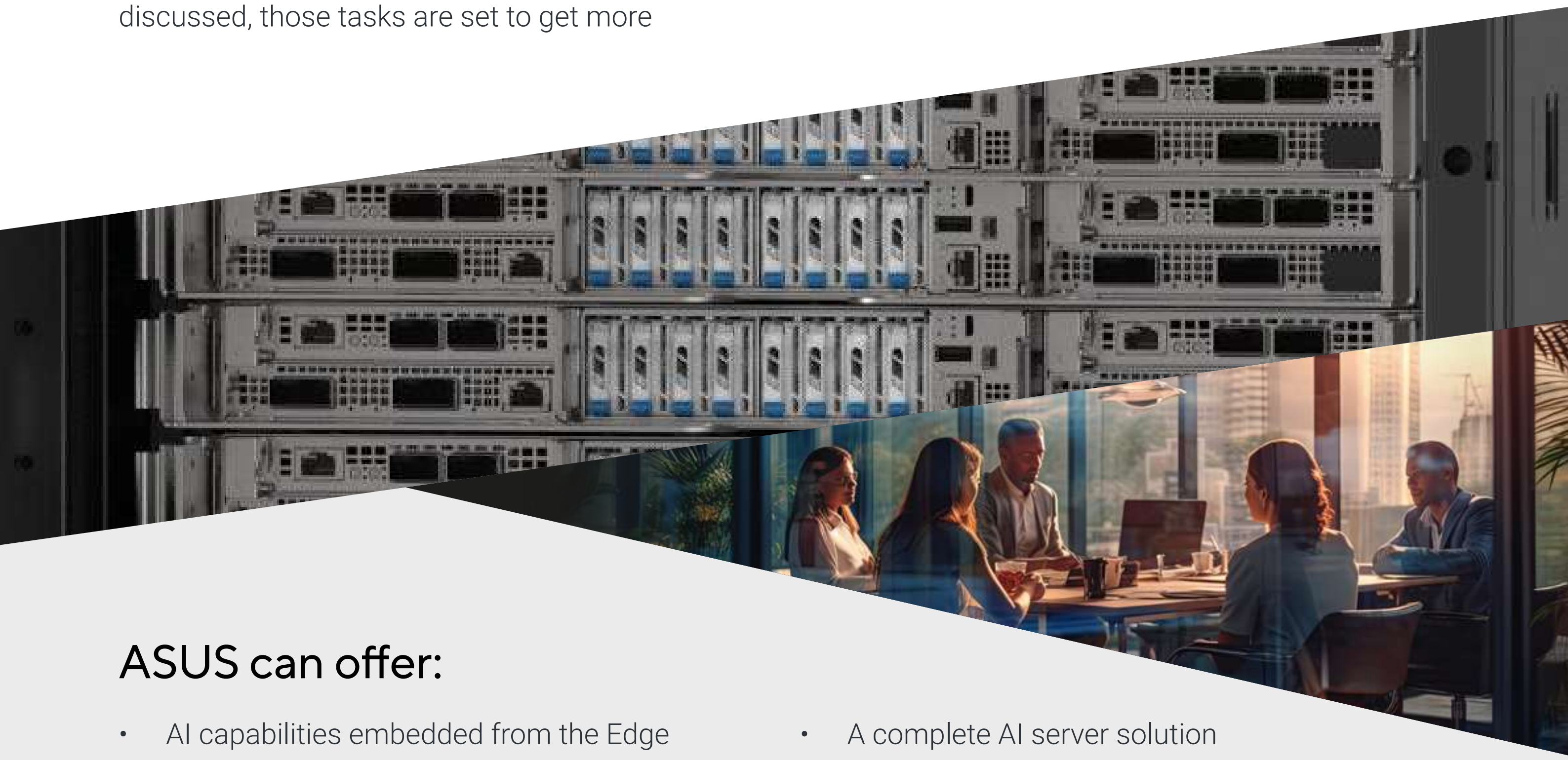
Businesses are already seeing these processing giants support the delivery of real-time and convenient computing services across industries; and by enabling heavy AI workloads the real-world applications are far reaching, from climate prediction and molecular model simulation all the way to engineering design and simulation.

BUILDING BESPOKE SOLUTIONS FIT FOR THE AI FUTURE

The power of supercomputers and what they can offer is obvious, overcoming many processing hurdles such as latency. With the right partner supporting the design of supercomputers, customers can access a comprehensive AI infrastructure built for diverse workloads and the dynamic needs that comes with deploying AI solutions – enabling even the most ambitious vision.

It's essential that both the hardware and software that organisations choose to adopt as part of their AI journey are purpose-built for handling intensive AI tasks. And as we've already discussed, those tasks are set to get more complicated and require increasing amounts of processing power, so they need to be ready for not just today, but tomorrow's AI needs too. This requires a software-driven approach that ensures seamless operations, speeding AI deployments so organisations can efficiently process vast datasets and execute complex computations without worry.

At ASUS we consider the needs of AI across every aspect of our portfolio, from the Edge to the server and beyond. Customers can feel confident that they'll get the very most out of their AI deployments with us.



ASUS can offer:

- AI capabilities embedded from the Edge to the server and beyond
- Quick fulfilment response to almost any requirement, with top-tier components, strong ecosystem partnerships, feature-rich designs, and superior in-house expertise
- A complete AI server solution from software to hardware
- AI expertise includes leveraging internal software whilst partnerships with software and cloud providers offer holistic solutions

ASUS AND NVIDIA: BUILDING A MORE SUCCESSFUL AI FUTURE TOGETHER

Together with NVIDIA®, ASUS is accelerating the potential of AI through powerful server and infrastructure solutions that eliminate latency and enable endless possibilities, supporting AI at the Edge, AI inference and fine tuning, AI training, Generative AI, and AI supercomputing. Creating bespoke solutions from the ground up to meet the most demanding computing needs; we're maximising General Processing Units (GPU) to accelerate and simplify even the most complex AI tasks, building supercomputers and LLMS, and innovating at every step to deliver pioneering value to our customers as they strive to meet the growing demands of AI workloads.

Over

72,000

awards won since 2001

Demonstrating our relentless innovation and focus on design and quality in everything we do.

KEY AWARD WINS

International Forum (iF) 315	Red Dot Award 268
Good Design Awards 227	Taiwan Excellence 793
Computex d&i Awards 153	CES Innovation Awards 144

ASUS ProArt GeForce RTX 4080 SUPER

The NVIDIA Omniverse™

As the AI market continues to grow traction, NVIDIA® believes the best is yet to come; an ethos demonstrated in the development of the NVIDIA Omniverse™. A platform of APIs, SDKs, and services that enable developers to easily integrate Universal Scene Description (OpenUSD) and NVIDIA RTX™ rendering technologies into existing software and simulation workflows for building AI systems. Designed to accelerate 3D design collaboration and simulation, ASUS servers, powered by NVIDIA AI accelerator solutions, offer seamless integration with Omniverse for unparalleled performance, reliability and scalability, from on-premise to hybrid cloud.

Operations Dashboard

7
Average Events

02:47
Average Completion Time

94%
Amount Complied to Date

ASUS Prime GeForce RTX 4070 SUPER

FROM STRENGTH TO STRENGTH: OUR MOST POWERFUL AI SERVER YET

The ASUS ESC N8-E11/E11V AI server is designed for generative AI with optimised server systems, data-centre infrastructure, and AI software-development capabilities.

Powered by 7U NVIDIA®'s HGX H100/H200 eight GPU server AI accelerators, the ESC N8-E11 and E11v are built to deliver robust AI computing capabilities where they're needed most. Giving organisations:

- The full power of NVIDIA GPUs, BlueField-3, NVLink, NVSwitch, and networking
- Efficient scaling with direct GPU-to-GPU interconnect via NVLink delivering 900GB/s bandwidth
- The highest throughput during computer-intensive workloads
- Modular design with reduced cable usage
- High-level power efficiency: 4 + 2 80 PLUS Titanium power supplies
- Optimised thermal design to support efficiency goals

ASUS ESC N8-E11V

NVIDIA HGX H100/H200

Clever cooling capabilities

Optimised power efficiency and thermal design means the ESC N8-E11/E11V feature both air cooling and direct-to-chip (D2C) liquid cooling solutions, in addition to dedicated CPU and GPU airflow tunnels to expel heat, reducing operational cost and ensuring peak performance at all times.

Liquid cooling

Traditional methods of cooling struggle to cope with the heat generated by dense clusters of CPUs and GPUs, impacting on data centre efficiency. Liquid cooling is the energy-efficient solution that can scale with cluster density and processing loads.

Direct Liquid Cooling (DLC) Direct to Chip (D2C)

A cold plate directly on top of CPUs and GPUs continually funnels heat through a liquid coolant network to a cooling distribution unit (CDU) in the rack. The CDU dissipates the heat, circulating the chilled coolant back in a closed-loop system.

Immersion cooling

This involves submerging an entire server into a thermally conductive dielectric fluid for maximum heat dissipation. It is regarded as the most energy-efficient form of liquid cooling on the market.

CASE STUDY: ADVANCING AI IN INDIA

ASUS design and innovation capabilities are helping many organisations to overcome the challenges of supercomputing. When managed data centre service provider Yotta developed Shakti Cloud (a comprehensive AI platform where users can build, train and deploy AI models) they needed to optimise performance to be able to cope with deep learning and other more complex AI workloads. Yotta came to ASUS to help create a scalable infrastructure ready for growth, and with the reliability needed for critical AI tasks where downtime isn't an option.

With ASUS ESC N8-E11, Shakti Cloud delivers an AI cloud platform with exceptional processing power, faster training times for complex AI models and smoother handling of large datasets, and a modular design for effortless scalability, supporting the advancement of AI in India.

READ THE FULL STORY

CASE STUDY: THE NATIONAL CENTER FOR HIGH-PERFORMANCE COMPUTING (NCHC)

The latest high-performance computing system built by the National Center for High-performance Computing (NCHC) in Taiwan, Forerunner 1 provides the resources for anything from research topics such as climate prediction, to astrophysics simulation, molecular model simulation, and more. ASUS was integral in the construction of Forerunner 1, from data centre construction to cabinet installation, testing and onboarding. To create a greener solution, ASUS refined the liquid-cooling setup, achieving a remarkable PUE of just 1.17 (surpassing the 1.28 accepted standard) and reaching #92 in the Green 500 in November 2023.

AI: THRIVE OR FALL BEHIND

Before long, AI will become instrumental in the success of businesses everywhere. Whether organisations have chosen to bring in chatbots to streamline customer services or build AI infrastructure to create their own 'AI analysts', businesses that don't jump on the bandwagon might just find themselves left behind.

The need is clear, with IDC forecasting that worldwide revenue for AI platforms software will grow to $153.0 billion in 2028 with a compound annual growth rate of 40.6% over the 2023-2028 forecast period.[9]

As more providers join the conversation, and more solutions and innovations flood the market, businesses need to be ready. AI strategy is something which needs to be properly built alongside known partners who are considering how to solve problems, such as processing, but also stay at the cutting edge of innovation.

Every day we see outputs from AI that 20 years ago would have been the stuff of science fiction. A strong partner will know this and provide an offering to help organisations deliver the future of AI, today.

ASUS | BUSINESS

NVIDIA

EXPERIENCE THE UNBELIEVABLE

With ASUS and NVIDIA® you'll find a partnership to elevate your AI computing infrastructure. Enabling your AI-driven solutions to flourish – and putting you firmly at the helm of the next generation of AI and all that is possible.

Power up your infrastructure with servers that are ready to meet the needs of your AI workloads today and tomorrow. There's no need to wait, with units of ASUS ESC N8-E11 available and ready to ship.

To experience it for yourself, you can access our remote PoC trial. Spin up an environment with us and test it over two weeks – without needing to leave your site – to get a taste of how it can work for your organisation's AI ambitions.

FIND OUT MORE

Sources:

1 https://thorax.bmj.com/content/early/2024/09/25/thorax-2024-221662

2 https://www.ons.gov.uk/businessindustryandtrade/itandinternetindustry/articles/understandingaiuptakeandsentimentamongpeopleandbusinessesintheuk/june2023

3 https://assets.publishing.service.gov.uk/media/61d87355e90e07037668e1bd/AI_Activity_in_UK_Businesses_Report__Capital_Economics_and_DCMS__January_2022__Web_accessible_.pdf

4 https://www.databricks.com/sites/default/files/2023-07/ebook_mit-cio-generative-ai-report.pdf

5 https://www.llama.com/

6 https://spj.science.org/doi/10.34133/icomputing.0006#:~:text=Challenges%20in%20computing&text=For%20example%2C%20the%20computing%20power,increase%20in%20computational%20capacity%20requirements

7 https://top500.org/lists/top500/2024/06/

8 https://servers.asus.com/stories/TAIWANIA-2-Supercomputer

9 https://www.idc.com/getdoc.jsp?containerId=prUS52472424