https://doi.org/10.1038/s41928-024-01164-6

AI success relies on access

As the scale and application of artificial intelligence technologies continues to grow, addressing challenges related to the wider accessibility of the underlying technology becomes increasingly important.

In an Editorial back in April 2018, we asked if hardware innovation would be able to meet the growing demands of artificial intelligence (AI)[1]. At that time, AlphaGo – a powerful machine learning program created to compete with human players in the game Go – illustrated the forefront of AI capabilities. Six years later, and with the widespread use of large language models and generative AI, the answer to that question appears to be yes. At least for now.

Central to this success has been the continuing development of graphics processing units (GPUs). Machine learning models examine (or train on) large volumes of data and try to find patterns, allowing them to, for example, classify images and generate sequences of words. Much of this is performed in data centres using GPUs – sometimes with up to tens of thousands of them. The architecture of GPUs, which have highly parallel structures and close coupling between memory and processing units, can provide a particularly efficient approach to processing machine learning data.

The growing size of machine learning models – which can have up to trillions of parameters – combined with their widespread use has led to a surge in demand for AI chips, and the chips have become one of the main sources of revenue growth in the semiconductor industry. Leading the way in terms of hardware for AI is Nvidia and their A100 and H100 GPU products. These products power the AI offerings of many global technology companies, including Google, Microsoft, Meta and OpenAI. And last month, Nvidia announced the arrival of its Blackwell GPU platform, which is designed to run trillion-parameter large language models at lower energy than previous platforms[2].

The reliance on Nvidia's hardware has pushed some companies to design their own AI chips. Google has been doing this for some time with their tensor processing unit series of chips, and Microsoft and Meta have also been developing their own training- and inference-focused hardware. But Nvidia's success is also based on the software platform they offer engineers and developers to build AI tools and applications. This platform, which is known as CUDA, ties its user base to Nvidia hardware – a point that has not gone unnoticed[3].

As a result, several companies have recently come together to form the UXL Foundation, a consortium that plans to develop open-source tools and software in order to create a standardized approach to AI programming – and thus allow a greater choice in hardware. Given the size of CUDA's user base, and the maturity and all-encompassing nature of the platform, the UXL Foundation faces a considerable challenge. However, their principles of openness and accessibility are undoubtedly important for the field.

Nvidia's market ascendancy has parallels elsewhere in the semiconductor industry. The manufacture of advanced node integrated circuits is currently dominated by the Taiwan Semiconductor Manufacturing Company (TSMC), and ASML is the only supplier of the lithography tools needed to manufacture the most advanced chips. Having a handful of companies dominate key aspects of a technology is potentially problematic, particularly when governments restrict, due to geopolitical tensions, who the companies can sell their products to[4]. Legitimate security concerns may drive these restrictions, but a strong semiconductor industry – required for the success of AI – relies on access to global markets to finance and support expansion and drive research and development. And broader and more balanced access to semiconductor technology will be important if AI is to help solve some of the biggest computing problems we face this century such as planetary-scale weather modelling and real-time, brain-scale modelling[5].

Whether hardware innovation continues to meet the demands of AI over the coming years – and helps deliver advances in capabilities at a similar rate to what we have seen over the last few years – remains an open question. But history has shown that the semiconductor community can deliver continuous improvements in hardware and performance, and can adapt to changing computational needs. Thus, the question we should perhaps be asking now, instead, is how do we ensure that the most powerful AI technologies can be developed and applied in the most equitable ways possible.

Published online: 29 April 2024

References

1. *Nat. Electron.* **1**, 205 (2018).
2. NVIDIA Blackwell platform arrives to power a new era of computing. *NVIDIA News* (18 March 2024); https://go.nature.com/3vTIJBx
3. Cherney, M. A. Behind the plot to break Nvidia's grip on AI by targeting software. *Reuters* (26 March 2024); https://go.nature.com/3xwHE32
4. Baazil, D., Koc, C., Hawkins, M. & Nienaber, M. US urges allies to squeeze China further on chip technology. *Bloomberg* (6 March 2024); https://go.nature.com/443j2uT
5. Conklin, A. A. & Kumar, S. *Nat. Electron.* **6**, 464–466 (2023).