TechScape: With its trillion-dollar valuation, will Nvidia's reign last?

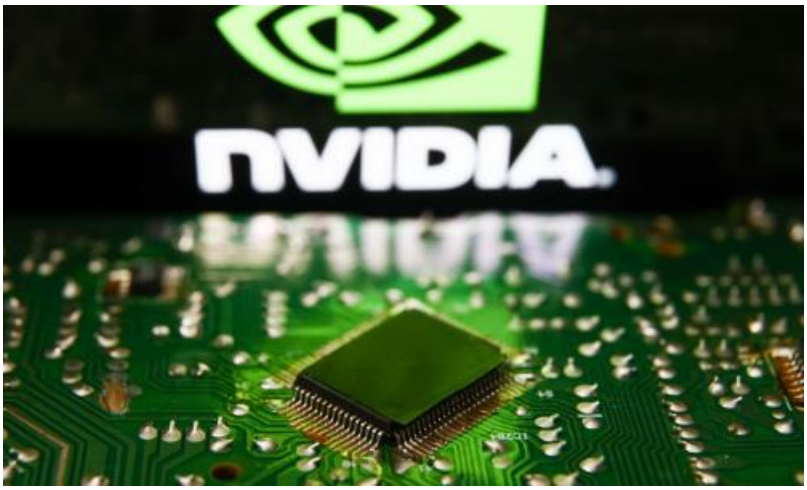
Publication Date: 2024-02-27

Author: Alex Hern

Section: Technology

Tags: Apple, TechScape newsletter, Computing, Artificial intelligence (AI), newsletters

Article URL: https://www.theguardian.com/technology/2024/feb/27/techscape-nvidia-apple-of-ai



Everyone wants to be like Apple. The largest publicly traded company in the world, with a flagship product that prints money, and a cultural footprint that has reached world-historical importance: the 21st-century Ford. On a surface level, the companies that get slapped with that comparison are obvious enough. If you pump out well-made, slickly designed consumer electronics that arrive in a nice box, someone somewhere will compare you to the Cupertino giant. Dig a bit deeper, and there's more meaningful comparisons to be made. Apple isn't defined only by its style but also by its focus. A small number of computers, phones and tablets, in a small number of configurations, comprises the bulk of its revenue. That focus allowed it to develop its reputation for quality, yes – but also contributed to its fearsome media strategy, ensuring that every product launch is an industry event in a way that few peers have been able to emulate. That's what I was thinking almost a decade ago when I called Blizzard, the gaming giant behind World of Warcraft and Diablo, "the Apple of gaming". (Now owned by Microsoft and racked by misconduct allegations, Blizzard's star has fallen since then.) But something else makes Apple what it is, and it's harder for upstarts to emulate: the Apple they see is just the latest evolution of a company that was a titan of industry before the latest generation of founders were even born. The Apple II, the Mac and the iMac all shaped computing in the 25 years before the iPod turned Apple into a consumer electronics company. And the iPod gave Apple a further decade of growth and refinement before the iPhone arrived and created the megacorp of today. Which brings us to Nvidia. A trillion dollars isn't cool On Friday, Nvidia became the fifth publicly traded company to ever surpass $2tn in valuation, just nine months after it became the ninth to ever break $1tn. Apple, Microsoft and Google took two years to cross that gap (although all three had the Covid pandemic in between the milestones). The simple reason for the explosion in valuation is that Nvidia is the company to bet on if you want to invest in AI. Google and Microsoft already have sky-high valuations, OpenAI isn't publicly traded, and while everyone and their dog is throwing a large language model up on a website and calling it a skunkworks project, it's hard to tell who has staying power. But Nvidia isn't making promises about future breakthroughs. The company is printing money now. Its chips, mostly shipped in the form of "GPUs", a special type of processor initially developed for the needs of PC gamers, are indispensable to companies on the cutting edge of the AI revolution. In January, Mark Zuckerberg spoke about Meta's plans to build the "world's fastest AI supercomputer" – and backed his words up with a commitment to buy 350,000 of Nvidia's fastest chips: The AI supercomputer, dubbed AI Research SuperCluster (RSC) by Zuckerberg's Meta business, is already the fifth fastest in the world, the company said. "The experiences we're building for the metaverse require enormous compute [sic] power (quintillions of operations/second!) and RSC will enable new AI models that can learn from trillions of examples, understand hundreds of languages, and more," wrote Zuckerberg. Just

one of the chips Zuckerberg committed to buy by the end of this year costs around $30,000. That $10bn investment in AI is pure cost for Meta, and pure revenue for Nvidia. The valuation starts to make sense. Those chips – Nvidia's H100 series – are so effective at training top-tier AI models that they're subject to export restrictions by the US government, which seeks to keep them out of the hands of Chinese companies and the Chinese government itself. The company had to spin up a separate, weaker, line of chips called the H800s to sell to mainland China as a result – until those, too, were blocked from export. Designing such hardware is an intensely specialist job. GPUs are distinct from CPUs – central processing units – by focusing purely on speed for the absolute simplest tasks possible, rather than devoting complex architecture to speeding up common, but slightly more convoluted, tasks. That's perfect for AI, the development of which involves a bunch of processes that simplify out to doing basic arithmetic in unimaginable quantities. It's also perfect for other uses. It is the general idea behind "mining" of cryptocurrencies, particularly those based on Ethereum. (Bitcoin and its derivatives use a mining process so streamlined that even Nvidia's processors can be beaten by ones designed specifically for mining bitcoin and only bitcoin.) That caused issues for the company during the most recent crypto bubble, when enthusiasts would outbid Nvidia's other customers for the latest GPUs – sparking the company's chief technology officer to dismiss the sector as "adding nothing useful to society". But Nvidia is an overnight success decades in the making thanks to the fortunate coincidence that another sector of computing also needs very, very simple calculations to be done very, very fast: gaming. The G in GPU initially stood for "graphics", and Nvidia's chips are still the best in the industry for people who want their video games to look really, really good on a PC screen. The strange quirk of history is that the only way to render 3D graphics fast enough to display on screen is to build an entire system for visualising 3D space that allows you to work out what any given part of the screen should be showing without knowing what any other part is already showing. That requires breaking the complex task down into a lot of much simpler tasks that can be done at the same time – like, for instance, working out what every single one of the 4,000 pixels on screen should be showing, simultaneously, 60 times a second. AI until I ain't That also points to the weaknesses in Nvidia's command of the AI industry, though. The company's sky-high valuation is based on three major assumptions about the future of AI: that training the best systems is going to be the way to stay on top of the industry; that the only way to do that is to be with the fastest possible chips; and that those systems are going to be kept in the cloud and run in massive data centres. But what if they aren't? There could be diminishing returns from training the best and biggest AI systems. If your goal is to produce a super-intelligence, you need to constantly improve things. But if your goal is to provide a compelling autocomplete for a programming tool, then it could be that you can beat a competitor with a more technically capable AI system by incorporating your own system into existing tools better, by making it run faster, and by updating it more frequently as the industry changes. That could devalue raw computing power of the sort that Nvidia sells. Or the "parallelisation" that enables GPUs to function could go further still: if you can split one task over 600,000 GPUs in a data centre, what's to stop you from splitting it over 60m GPUs in the laptops of all your customers around the world? From SETI@home, which took that approach to crunch data related to potential extraterrestrial intelligence, to cryptocurrency mining "pools", the idea of distributed computing isn't new, and there's no significant reason to believe it couldn't work in AI training, too. Or maybe the data centres that run the AI systems are what will crumble. A combination of privacy, environmental and cost concerns could push more and more AI models to the "edge", with your phone or laptop doing the final load of number crunching. Currently, the trade-offs of that are too weighty – who wants to wait a minute or two to receive a significantly worse version of the image that Dall-E could send you in a second? But as phones get ever faster and big businesses lose their appetite for subsidising AI usage, that could change. If any one of those assumptions is proved wrong, Nvidia can probably stay on top of its game. The company has already begun experimenting with the latter, for instance, releasing its "Chat with RTX" demo to let owners of its fastest graphics cards experiment with running a chatbot on their own computers. But if two or even three start to turn out differently to how the industry expects, then the company could have a very short-lived reign as the Apple of AI. • If you want to read the complete version of the newsletter please subscribe to receive TechScape in your inbox every Tuesday. • This article was amended on 27 February 2024. An earlier version incorrectly referred to the Nvidia chips made specifically for export to China as the A100, rather than the H800.