How savvy trillion-dollar chipmaker Nvidia is powering the AI goldrush

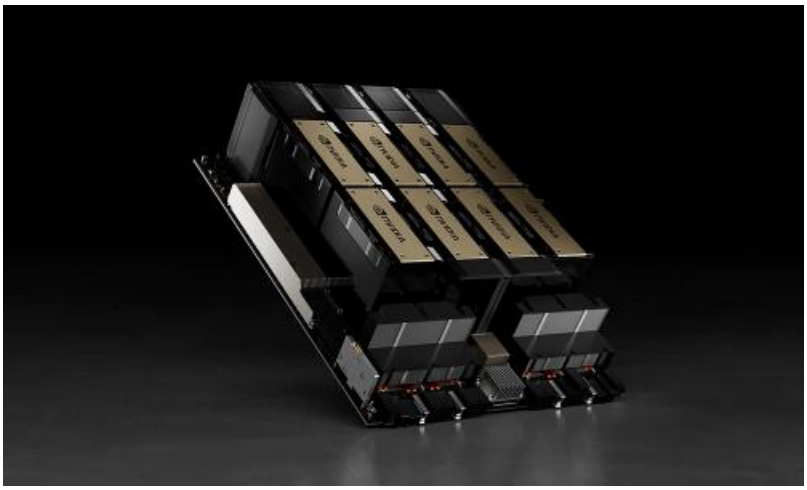
Publication Date: 2023-09-09

Author: John Naughton

Section: Technology

Tags: Computing, The networker, Artificial intelligence (AI), comment

Article URL: https://www.theguardian.com/technology/commentisfree/2023/sep/09/nvidia-processors-ai-artificial-intelligence-chip-goldrush-intel



It's not often that the jaws of Wall Street analysts drop to the floor but late last month it happened: Nvidia, a company that makes computer chips, issued sales figures that blew the street's collective mind. It had pulled in $13.5bn in revenue in the last quarter, which was at least $2bn more than the aforementioned financial geniuses had predicted. Suddenly, the surge in the company's share price in May that had turned it into a trillion-dollar company made sense. Well, up to a point, anyway. But how had a company that since 1998 – when it released the revolutionary Riva TNT video and graphics accelerator chip – had been the lodestone of gamers become worth a trillion dollars, almost overnight? The answer, oddly enough, can be found in the folk wisdom that emerged in the California gold rush of the mid-19th century, when it became clear that while few prospectors made fortunes panning for gold, the suppliers who sold them picks and shovels prospered nicely. We're now in another gold rush – this time centred on artificial intelligence (AI) – and Nvidia's A100 and H100 graphical processing units (GPUs) are the picks and shovels. Immediately, everyone wants them – not just tech companies but also petro states such as Saudi Arabia and the United Arab Emirates. Thus demand wildly exceeds supply. And just to make the squeeze really exquisite, Nvidia had astutely prebooked scarce (4-nanometre) production capacity at the Taiwan Semiconductor Manufacturing Company, the only chip-fabrication outfit in the world that can make them, when demand was slack during the Covid-19 pandemic. So, for the time being at least, if you want to get into the AI business, you need Nvidia GPUs. What's so special about GPUs? Well, here's where the video gaming connection comes in. In gaming, graphics images are made up of polygons (mostly tiny triangles) – rather as the images produced by a digital camera are composed of rectangular pixels. The more triangles you have, the higher the resolution of the resulting image. For gaming, polygons are defined as the coordinates of their vertices, so each object becomes a large matrix of numbers. But most objects in a video game are dynamic, not static: they move and change shape, and for each change, the matrix has to be recalculated. Underpinning a video game, therefore, is a fiendish amount of continuous computation. And for the game to be realistic, this computation has to be done very quickly. Which basically means that conventional central processing units – which do things serially, one step at a time – are not up to the job. What makes GPUs special is their ability to do thousands or even millions of mathematical operations in parallel – which is why, when you're playing Grand Theft Auto V, the goodies and baddies move swiftly and smoothly, and roam around a convincingly rendered fictional version of Los Angeles in real time. As interest in machine learning and neural networks surged in the 00s, and especially after 2017, when Google introduced the "transformer" model on which most generative AI is now based, AI researchers realised that they needed the parallel processing capabilities offered by

GPUs. At which point it became clear that Nvidia was the outfit that had the head start on everyone else. And since then the company has wisely capitalised on that advantage and consolidated its lead by building a software ecosystem around its hardware that is like catnip for AI developers. So is Nvidia set to become the next Apple, or at least the next Intel? For the next few years, its dominance seems pretty secure, partly because its revenues are coming more from cloud-computing companies anxious to kit out their datacentres not just with conventional servers but increasingly with parallel-processing kit that will address the anticipated needs of the AI gold rush. They are good customers that pay on time and it'll take them a couple of years at minimum to reconfigure their cloud infrastructures. But nothing lasts for ever. After all, it's not that long ago that Intel's dominance of the semiconductor industry seemed total. And now it's a shadow of its former self. Curiously enough, though, when Nvidia passed the trillion-dollar milestone, the thought on everyone's mind was not of Intel but of Cisco, a famous manufacturer of networking and telecoms equipment that once happened to be in the right place at the right time, when the first internet boom kicked off in the mid-1990s. Its revenues tripled between 1997 and 2000 as demand for routers and other networking equipment soared. Then came the bust and by 2001 Cisco's share price (and consequent market valuation) had dropped by 70%. Could something such as this happen to Nvidia? The key question, says Ben Thompson, the shrewdest tech guru around, is: what will the eventual market for AI be when the frenzy has abated? Nobody knows the answer to that. Whatever happens, though, Nvidia's picks and shovels will have made some people an awful lot of money. What I've been reading Definite article Consciousness Is a Great Mystery. Its Definition Isn't is an interesting post by Erik Hoel on his Intrinsic Perspective blog. Intelligence test In his typically laconic and thoughtful essay Generative AI and Intellectual Property on his website, Benedict Evans addresses an as yet unresolved problem. Foreseen consequences How Misreading Adam Smith Helped Spawn Deaths of Despair is a wonderful lecture in the Boston Review by Nobel economics laureate Angus Deaton.