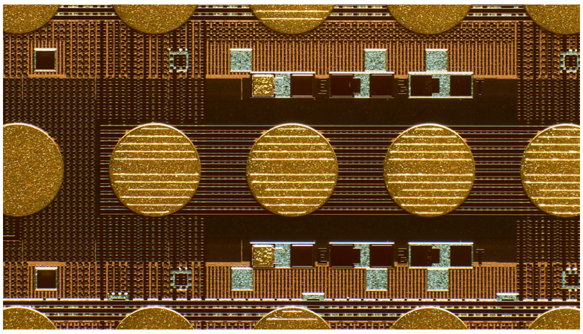
NICOLE KOBIE LONG READS 17.06.2021 10:10 AM

NVIDIA and the battle for the future of AI chips

NVIDIA's GPUs dominate AI chips. But a raft of startups say new architecture is needed for the fast-evolving AI field



SUN LEE

THERE'S AN APOCRYPHAL story about how NVIDIA pivoted from games and graphics hardware to dominate AI chips – and it involves cats. Back in 2010, Bill

Dally, now chief scientist at NVIDIA, was having breakfast with a former colleague from Stanford University, the computer scientist Andrew Ng, who was working on a project with Google. "He was trying to find cats on the internet – he didn't put it that way, but that's what he was doing," Dally says.

Ng was working at the Google X lab on a project to build a neural network that could learn on its own. The neural network was shown ten million YouTube videos and learned how to pick out human faces, bodies and cats – but to do so accurately, the system required thousands of CPUs (central processing units), the workhorse processors that power computers. “I said, ‘I bet we could do it with just a few GPUs,’” Dally says. GPUs (graphics processing units) are specialised for more intense workloads such as 3D rendering – and that makes them better than CPUs at powering AI.

Dally turned to Bryan Catanzaro, who now leads deep learning research at NVIDIA, to make it happen. And he did – with just 12 GPUs – proving that the parallel processing offered by GPUs was faster and more efficient at training Ng’s cat-recognition model than CPUs.

But Catanzaro wants it known that NVIDIA didn’t begin its efforts with AI just because of that chance breakfast. Indeed, he had been developing GPUs for AI while still a grad student at Berkeley, before joining NVIDIA in 2008. “NVIDIA’s position in this market is not an accident,” he says.

The when and how of it all seems unimportant now that NVIDIA dominates AI chips. Co-founded in 1993 by CEO Jensen Huang, NVIDIA’s major revenue stream is still graphics and gaming, but for the last financial year its sales of GPUs for use in data centres climbed to $6.7 billion. In 2019, NVIDIA GPUs were deployed in 97.4 per cent of AI accelerator instances – hardware used to boost processing speeds – at the top four cloud providers: AWS, Google, Alibaba and Azure. It commands “nearly 100 per cent” of the market for training AI algorithms, says Karl Freund, analyst at Cambrian AI Research. Nearly 70 per cent of the top 500 supercomputers use its GPUs. Virtually all AI milestones have happened on NVIDIA hardware. Ng’s YouTube cat finder, DeepMind’s board game champion AlphaGo, OpenAI’s language prediction model GPT-3 all run on NVIDIA hardware. It’s the ground AI researchers stand upon.

Despite this success, Catanzaro is annoyed by the persistent suggestion that NVIDIA stumbled blindly into AI from gaming. “I swear, pretty much every story that I read, the narrative is that GPUs randomly happen to be excellent at AI, and NVIDIA lucked into a temporary windfall by selling existing chips to a new market, and soon they’re going to be displaced by startups,” Catanzaro says. “But NVIDIA has been very strategic about how it approaches the AI market for a decade now.”

A decade in, that market is ripe for disruption. AI is beginning to be used by more and more businesses to make sense of the oceans of data they collect, while governments pump money into deep learning research to keep ahead of one another. The race between the US and China is particularly hot; Deloitte analyst Costi Perricos says AI will become the “next kind of superpower” for nations to compete over. At the same time, deep learning models are increasing in size and complexity, requiring ever more computing power.

OpenAI’s GPT-3, a deep learning system that can write paragraphs of sensible text, is the extreme example, made up of 175 billion parameters, the variables that make up models. It cost an estimated $4.6 million to compute, and that’s since been topped by a Google language model with 1.6 trillion parameters. More efficient hardware is necessary to chew through more parameters and more data for increased accuracy, but also to keep AI from becoming even more of an environmental disaster – Danish researchers calculated that the energy required to train GPT-3 could have the carbon footprint of driving 700,000km.

We need more AI chips and we need better AI chips. While NVIDIA’s early work has given the GPU maker a head start, challengers are racing to catch up. Google started making its own chips in 2015; Amazon last year began shifting Alexa's brains to its own Inferentia chips, after buying Annapurna Labs in 2016; Baidu has Kunlun, recently valued at $2 billion; Qualcomm has its Cloud AI 100; and IBM is working on an energy-efficient design. AMD acquired Xilinx for AI data centre work, and Intel added AI acceleration to its Xeon data centre CPUs in 2019; it has also bought two startups, Nervana in 2016 for $408 million and Habana Labs in 2019 for $2 billion. The startups that haven't yet been snapped up have released their own hardware, with the past few years seeing AI chips released or trialled by the likes of Graphcore, SambaNova, Cerebras, Mythic AI, Blaize and TensTorrent.

We are still in the early days of AI. Those cats were only calculated ten years ago; most of these startups are no more than a few years old. With more data set to flow as smart Internet of Things devices begin a machine-to-machine revolution, all have their view set on the same thing: owning the future of AI chips.

Bryan Catanzaro, vice president, applied deep learning at NVIDIA WINNI WINTERMEYER

MACHINE LEARNING IS a computing workload unlike any other, requiring a lot of maths using not very precise figures. Traditional high-performance computing (HPC), where multiple systems are linked together to build supercomputers to process complex workloads such as scientific simulations or financial modelling, requires high-precision maths, using 64-bit numbers if not higher. AI computing also requires massive computing infrastructure, but the maths used is less precise, with numbers that are16-bit or even 8-bit – it's akin to the difference between hyper-realistic graphics and pixelated games from the 80s. "The math is mostly easy, but there's a lot of it," says Andrew Feldman, CEO of AI chip startup Cerebras.

An AI chip is any processor that has been optimised to run machine learning workloads, via programming frameworks such as Google's TensorFlow and Facebook's PyTorch. AI chips don't necessarily do all the work when training or running a deep-learning model, but operate as accelerators by quickly churning through the most intense workloads. For example, NVIDIA's AI-system-in-a-box, the DGX A100, uses eight of its own A100 "Ampere" GPUs as accelerators, but also features a 128-core AMD CPU.

AI isn't new, but we previously lacked the computing power to make deep learning models possible, leaving researchers waiting on the hardware to catch up to their ideas. "GPUs came in and opened the doors," says Rodrigo Liang, co-founder and CEO of SambaNova, another startup making AI chips.

In 2012, a researcher at the University of Toronto, Alex Krizhevsky, walloped other competitors in the annual ImageNet computer vision challenge, which pits researchers against each other to develop algorithms that can identify images or objects within them. Krizhevsky used deep learning powered by GPUs to beat hand-coded efforts for the first time. By 2015, all the top results at ImageNet contests were using GPUs.

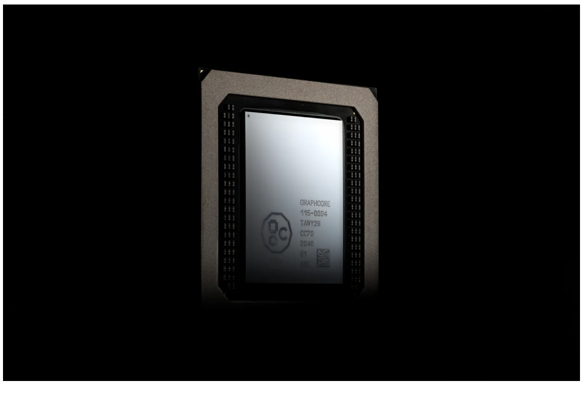
Deep learning research exploded. Offering 20x or more performance boosts, NVIDIA's technology worked so well that when British chip startup Graphcore's co-founders set up shop, they couldn't get a meeting with investors. "What we heard from VCs was: 'what's AI?'" says co-founder and CTO Simon Knowles, recalling a trip to California to seek funding in 2015. "It was really surprising." A few months later, at the beginning of 2016, that had all changed. "Then, everyone was hot for AI," Knowles says. "However, they were not hot for chips." A new chip architecture wasn't deemed necessary; NVIDIA had the industry covered.

But, in May 2016, Google changed everything, with what Cerebras' Feldman calls a "swashbuckling strategic decision", announcing it had developed its own chips for AI applications. These were called Tensor Processing Units (TPUs), and designed to work with the company's TensorFlow machine learning programming framework. Knowles says the move sent a signal to investors that perhaps there was a market for new processor designs. "Suddenly all the VCs were like: where are those crazy Brits?" he says. Since then, Graphcore has raised $710 million (£515 million).

What's in a name?

GPU, IPU, RPU – they're all used to churn through datasets for deep learning, but the names do reflect differences in architecture.

Graphcore



SUN LEE

Graphcore's Colossus MK2 IPU is massively parallel with processors operated independently, a technique called multiple instruction, multiple data. Software is written sequentially, but neural network algorithms need to do everything at once. To address this, one solution is to lay out all the data and its constraints, like declaring the structure of the problem, says Graphcore CTO Simon Knowles. It's a graph – hence the name of his company.

NVIDIA's rivals argue that GPUs were designed for graphics rather than machine learning, and that though their massive processing capabilities mean they work better than CPUs for AI tasks, their market dominance has only lasted this long due to careful optimisation and complex layers of software. "NVIDIA has done a fabulous job hiding the complexity of a GPU," says Graphcore co-founder and CEO Nigel Toon. "It works because of the software libraries they've created, the frameworks and the optimisations that allow the complexity to be hidden. It's a really heavy lifting job that NVIDIA has undertaken there."

But forget GPUs, the argument goes, and you might design an AI chip from scratch that has an entirely new architecture. There are plenty to choose from. Google's TPUs are application-specific integrated circuits (ASICs), designed for specific workloads; Cerebras makes a Wafer-Scale Engine, a behemoth chip 56 times larger than any other; IBM and BrainChip make neuromorphic chips, modelled on the human brain; and Mythic and Graphcore both make Intelligence Processing Units (IPU), though their designs differ. There are plenty more.

But Catanzaro argues the many chips are simply variations of AI accelerators – the name given to any hardware that boosts AI. "We talk about a GPU or TPU or an IPU or whatever, but people get too attached to those letters," he says. "We call our GPU that because of the history of what we've done... but the GPU has always been about accelerated computing, and the nature of the workloads people care about is in flux."

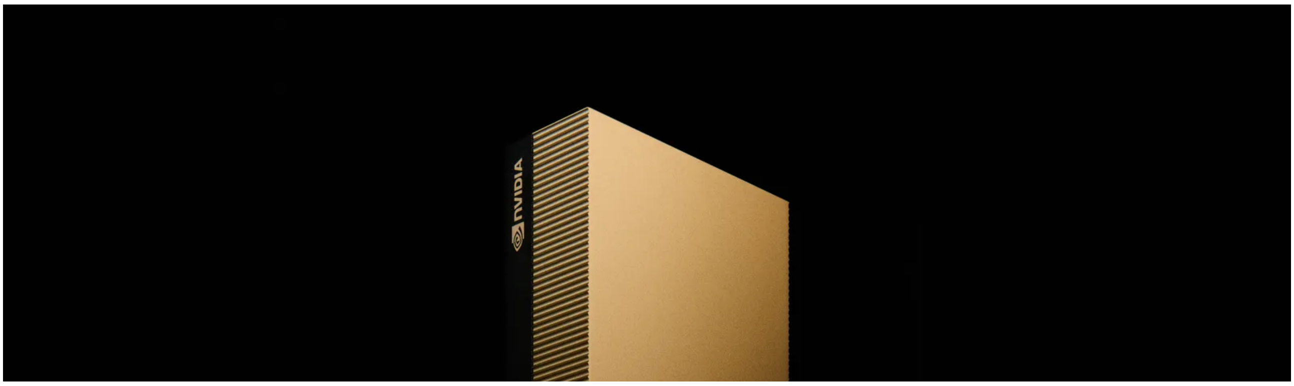
Can anyone compete? NVIDIA dominates the core benchmark, MLPerf, which is the gold standard for deep-learning chips, though benchmarks are tricky beasts.

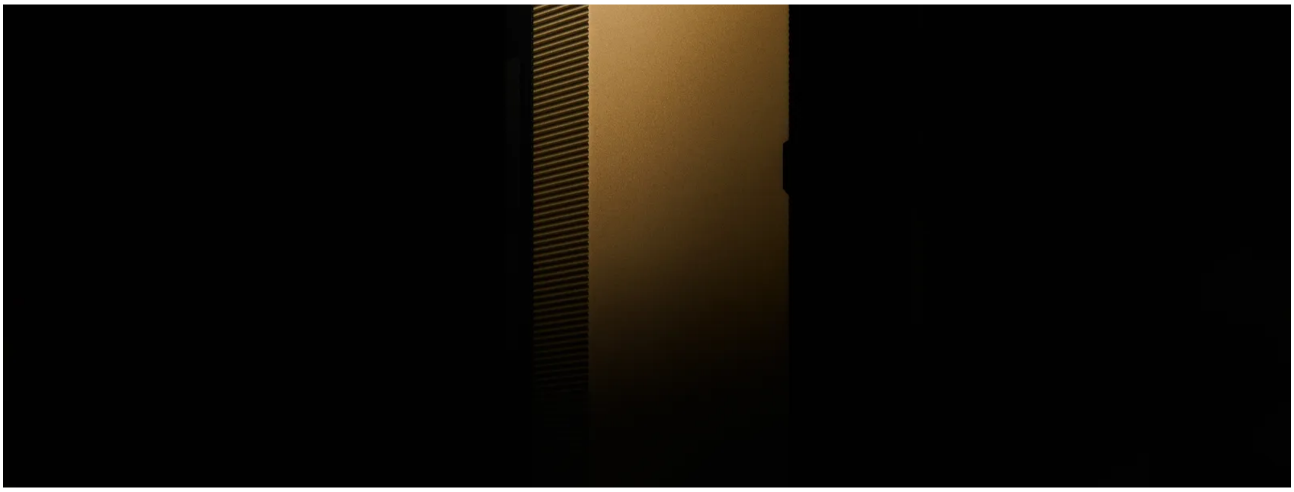
Analyst Karl Freund of Cambrian AI Research notes that MLPerf, a benchmarking tool designed by academics and industry players including Google, is dominated by Google and NVIDIA, but that startups usually don't bother to complete all of it because the costs of setting up a system are better spent elsewhere.

NVIDIA does bother – and annually bests Google's TPU. "Google invented MLPerf to show how good their TPU was," says Marc Hamilton, head of solutions architecture and engineering at NVIDIA "Jensen [Huang] said it would be really nice if we show Google every time they ran the MLPerf benchmark how our GPUs were just a little bit faster than the TPU."

To ensure it came out on top for one version of the benchmark, NVIDIA upgraded an in-house supercomputer from 36 DGX boxes to a whopping 96. That required recabling the entire system. To do it quickly enough, they simply cut through the cables – which Hamilton says was about a million dollars worth of kit – and had new equipment shipped in. This may serve to highlight the bonkers behaviour driven by benchmarks, but it also inspired a redesign of DGX: the current-generation blocks can now be combined in groups of 20 without any rewiring.

When it comes to benchmarks and supercomputers, you can always add more chips. But for the other side of AI computing – something called inference at the edge – it's a different story.





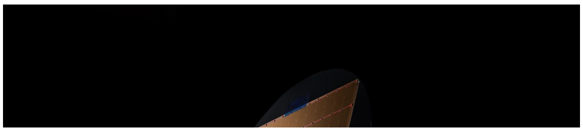
An NVIDIA SuperPOD, racks of which make up the Cambridge-1 supercomputer
SUN LEE

NVIDIA GRABBED THE world’s attention in 2020 when it bid $40 billion for ARM, the British chip designer whose architecture powers 95 per cent of the world’s smartphones. But the response wasn’t entirely positive. ARM co-founder Hermann Hauser, who no longer works at the company but still retains shares, has called it a “disaster” that may destroy ARM’s neutrality in the market. Regulators around the world – in the EU, UK, China and US – are closely studying the deal.

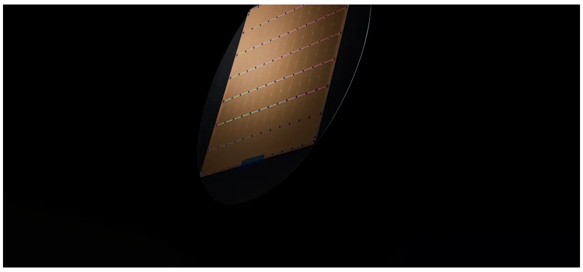
ARM designs chips, licensing the intellectual property out to companies to use as they see fit. If an AI chip maker needs a CPU for a system, they can license a chip design from ARM and have it made to their specifications. Rivals are concerned that NVIDIA taking control of ARM could limit those partnerships, though Huang has said “unequivocally” that NVIDIA would respect ARM's open model. The UK government is reportedly considering any national security implications, though ARM is currently owned by Japan’s SoftBank, and there are concerns in China that ARM being owned by an American company could mean its designs are banned from export to blacklisted Chinese companies under existing restrictions.

ARM is a major designer of the chips that will apply deep learning in the real world – so-called inference at the edge. This means the deal could have a huge

Cerebras



impact on the shape of the market; NVIDIA could dominate the data centre side with its GPUs and the edge with help from ARM.



SUN LEE

At Cerebras, CEO Andrew Feldman realised that communications on-chip are fast, but the slowdown happens between them – so why not just build a really big chip, so your data never has to leave? The Cerebras Wafer Scale Engine crams 400,000 cores onto 46.225 square millimetres. “GPUs have the right cores, but the wrong communication architecture,” he says.

What is inference at the edge? NVIDIA’s beefy, high-performance systems churn through data in order to train and apply models, but there’s another AI workload known as inference, which is the more lightweight task of using a trained model to then interpret something – such as a driverless car understanding what its cameras see, a smartphone app finding the edges of your face to apply cat ears to your selfie, or a medical imaging model spotting signs of cancer in a scan. Because of the huge amounts of computing power required, training is done in a data centre, but inference can be found in two places.

The first is also in the data centre: when you ask Alexa or Siri a question, it’s sent back to servers at Amazon and Apple for transcription and a response. The second place inference happens is in end-user devices, such as cameras, cars and smartphones – this is called edge computing. This requires less processing power, but it needs to be fast (no one wants to wait for their driverless car to think before deciding whether to brake).

NVIDIA currently dominates the data centre side. Its A100 churns through data for training, while inference is virtualised into smaller mini-servers, allowing 50 or more inference workloads to happen at the same time on the same hardware. That’s helpful for tech giants like AWS that offer AI as a service, as multiple companies can use the same hardware without risk of data leaking. At the edge, NVIDIA has DRIVE for driverless cars and EGX for on-location inference, but low-power chips aren’t its traditional speciality – if you’ve ever used a gaming laptop, you’ll have noticed it needs to be plugged in more regularly than a Chromebook.

Low-power chips are the domain of ARM, which is why NVIDIA has dropped $40 billion to acquire the company.

When it comes to AI, ARM's efforts centre on two areas. First, it is fitting software frameworks onto its existing CPUs. For more intense workloads, it has developed a neural processing unit (NPU) called Ethos to be used as an accelerator. Rene Haas, president of ARM's IP Products Group, says that devices using the Ethos-U55 should be arriving soon, as companies that licensed the design already have silicon produced.

With AI on the edge, voice assistants would no longer need to upload speech to AWS or Apple servers for processing, but could respond based on local intelligence. "It allows the work to be done close to the source, which helps in many ways in terms of efficiency," Haas says, noting that sending data back and forth to the cloud chews through battery power.

"We've talked about IoT for a long time, but the vision's never been realised until now," says David Hogan, vice-president of EMEA at NVIDIA. "It's this transformation that's at the heart of our plans to acquire ARM."

A technician inside the controlled environment of the Cambridge-1 supercomputer WINNI WINTERMEYER

WHILE THE REST of us baked banana bread and binged Netflix, Marc Hamilton, head of solutions architecture and engineering at NVIDIA, spent much of the last year building a £40 million supercomputer, navigating shortages caused by the pandemic to assemble the Cambridge-1 mostly on time. The build was made easier by NVIDIA's LEGO-style system. Eight A100 chips make up the heart of the computing system it calls DGX – it's the same relationship between the Intel or AMD chip running your laptop. Costing $199,000, the DGX is a full AI computer, with memory and networking and everything else, designed to be relatively plug-and-play. Cambridge-1 consists of racks upon racks of gold boxes in premade sets of 20 DGXs, known as a SuperPod.

Cambridge-1 will be the largest and most powerful AI supercomputer in the UK, and about 30th in the world, Hamilton says (though that ranking is likely to shift) – but it will only be the fifth largest in NVIDIA's own collection. Cambridge-1 was built using 80 DGX A100 boxes versus 560 for Selene, its largest.

NVIDIA built Cambridge-1 in the UK partially because of the ARM acquisition, as the buyout would mean the company gained employees in the UK. While it's not the overall fastest nor the biggest, Cambridge-1 claims two firsts. Hamilton calls it the world's first cloud-native supercomputer, as it features compartmentalisation akin to AWS, letting companies use the same hardware without risk of security breaches or data leaking. And that lets Cambridge-1 have its second first: this is the only supercomputer that NVIDIA will open up to external partners, letting universities and healthcare giants AstraZeneca, Oxford Nanopore and GSK run their own deep learning models.

Why does NVIDIA build its own supercomputers? One reason is that it needs toys to attract the best people. Back in 2016, NVIDIA didn't have a supercomputer, and Facebook and Google were snapping up the best AI researchers. "It's not because they pay them more," Hamilton says. "It's because Google and Facebook have thousands of GPUs that they use to run their business, and they make those accessible to their AI researchers."

SambaNova

SUN LEE

SambaNova Systems' software-defined approach puts data to the fore, replacing integers such as add and subtract with instructions to filter and reduce. SambaNova calls its design a reconfigurable dataflow, and that's achieved with 1.5TB of memory per "Cardinal" chip, with eight of those in each of its DataScale SN10-8R systems.

Now, NVIDIA's supercomputer Selene is fifth largest in the world, after one in Japan, one in China and two owned by the US government. That means, Hamilton says, that if you're a researcher who wants access to the fastest AI hardware, you can work for China, the US, or NVIDIA. China aims to be a global leader in AI by 2030, while the US wants to maintain its lead in the technology; there was already tension on the AI front, but the recent trade war between the two countries may turn it into something of an arms race. As a US company, NVIDIA doesn't completely avoid such issues.

Researchers in Catanzaro's 40-person lab develop AI to be used inside NVIDIA's own systems, but the lab also acts as a "terrarium" for systems architects to peek in and see how deep-learning models may work in the future. "If you want to build a chip for the future, you want it to be useful for the future, you have to have skill with forecasting what are the most important workloads of the future – what they look like computationally," says

Catanzaro. “If you mess it up, you build the wrong chip.” Chips take years to design and build, so such foresight is necessary.

What happens if models are developed that no longer work on GPUs, or at least not as well? NVIDIA’s Dally admits it’s a possibility, but with most researchers working on GPUs, he thinks it’s unlikely. “Before a new model takes off, we have generally heard about it and had a chance to kick its tyres and make sure it runs well on our GPUs,” he says.

Others disagree – and believe GPUs may be holding back deep learning models from their full potential. “Everybody bends their models to today’s technology,” says Cerebras’ Feldman. “One of the things we are happiest and most excited about are a group of customers who are writing entirely new models.” He says this year Cerebras will show examples of what it calls “GPU impossible work” – work that simply can’t be done on GPUs.

Graphcore’s Toon says researchers have long told him they’re held back by today’s hardware; his partner Knowles compares it to Orwell’s Newspeak, simple language that prevents people thinking more complicated thoughts. “There are ideas, such as probabilistic machine learning, which is still being held back because today’s hardware like GPUs just doesn’t allow that to go forward,” Toon says. “The race will be how fast NVIDIA can evolve the GPU, or will it be something new that allows that?”

Neil Thompson, a researcher at MIT’s Computer Science and Artificial Intelligence Lab, noticed a trend at AI conferences of researchers hinting that computational limits were holding back their models, limiting their choices and datasets, and compelling some to leave mistakes in their work because they couldn’t afford to re-run a model to fix the problem. “It’s really widespread and it’s a really big problem in terms of the future of deep learning if we’re going to practise it as we have been so far,” he says.

Thompson and colleagues analysed 1,058 AI papers, and found that the computing demands of machine learning were far outstripping hardware improvements or model training efficiencies. On this path, systems will one day cost hundreds of millions or even billions of dollars to train – and have other costs. “The problem with chucking more GPUs at it is every time you double the number of GPUs, you double the cost, you double the environmental footprint, carbon and pollution,” Thompson says.

He believes that hardware solutions alone – be they from NVIDIA or challengers – won’t be enough to prevent AI innovation from stumbling. Instead, we need to build more efficient models and make better use of what we already have. Ideas such as sparsity – ignoring the zeros in a data set to save on calculations – can help, as can being more methodical about data, only putting it against related parameters. Another idea is distilling what we learn from models into more lightweight equations, running only a relevant section of a model rather than a massive universal one.

Without such efforts, we’ll need bigger data centres. But AI shouldn’t be limited just to those who can afford a supercomputer. “Universities with less computer power are already becoming a smaller proportion” of those doing top-end deep-learning work, says Thompson. “There’s still quite a few people who can play in the game, but the number of players is getting smaller as the computation burden goes up. And we’ve already gotten to the point where some people have been excluded.”

Costs can be cut, which may be one way for startups to win customers against incumbents. AWS added chips from Habana Labs to its cloud last year, saying the Intel-owned Israeli designer was 40 per cent cheaper to run. “For AI to reach everyone and not just the rich, you really need to improve price performance,” says Eitan Medina, chief business officer at Habana Labs.

AI already has a bias problem, and that is exacerbated by unequal access to hardware. “It means we’ll only be looking at one side of the coin,” says Kate Kallot, head of emerging areas at NVIDIA. “If you leave out a large chunk of the population of the world... how are we going to be able to solve challenges everywhere in the world?” She points to the UN’s sustainable development goals: plenty of AI researchers are turning their work to address challenges such as poverty and the climate crisis, but these are issues that will largely impact emerging markets.

There are other challenges to add to the mix. Manufacturing of processors has been constrained during the pandemic, while last year's trade skirmish between the US and China raised concerns that the world's chip factories are predominately in Asia, with the EU recently pledging to produce a fifth of the world's top-end chips by 2030. Chip designers largely outsource manufacturing – NVIDIA's are made by Taiwan's TSMC – though Intel has its own foundries. In March, Intel announced plans to open two new factories in the US to make chips for external designers for the first time, perhaps giving the US more control over manufacturing.

As these hurdles are overcome, and chips continue to evolve, AI will expand to touch everything, akin to the wave of connectivity that saw wi-fi support and apps added to objects from toasters to fridges. But in the future, smart won't just mean internet-connected, but embedded with AI. "It will be everywhere," ARM's Haas says. "It will be ubiquitous in every single computing application in the next few years."

Nicole Kobie is a WIRED contributing editor. She writes Work Smarter, WIRED's weekly newsletter about the trends and technologies shaping the way you work.

CONTRIBUTING EDITOR

TOPICS ARTIFICIAL INTELLIGENCE TECHNOLOGY GAMING

JULY / AUGUST 2021 ISSUE