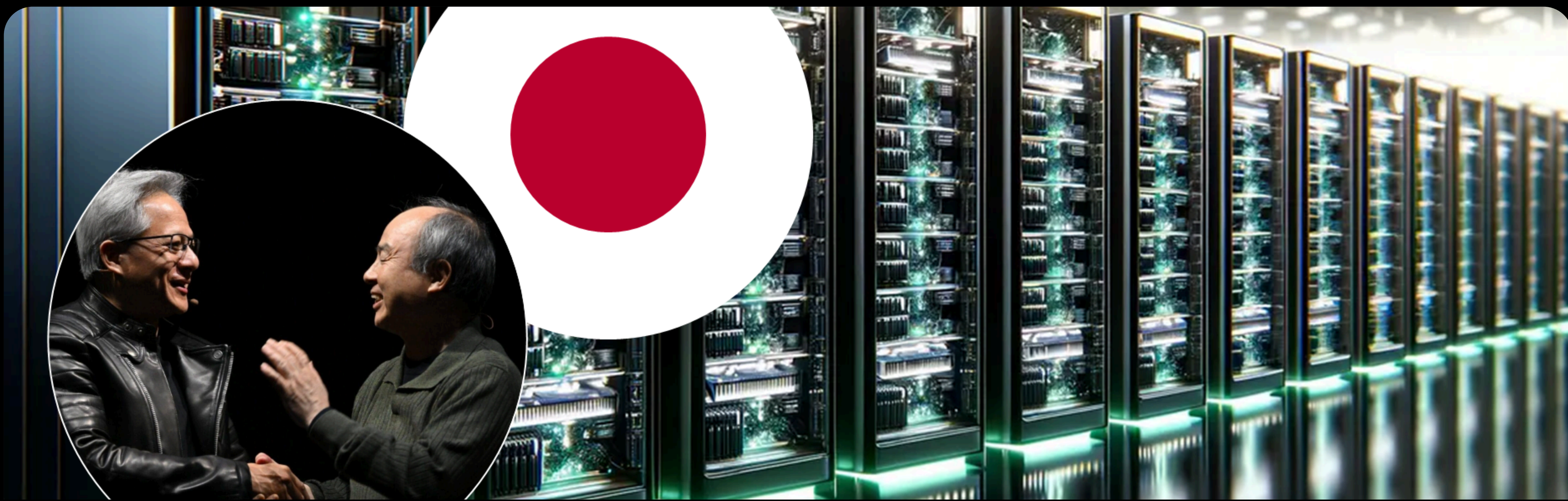
AI Pulse

Top AI & GenAI Headlines This Week

CW 46

Looking to catch up. The first of Nvidia's newest chips goes to SoftBank and Japan



Summary

- SoftBank has become the first customer for NVIDIA's next-generation Blackwell chips, collaborating to construct Japan's most powerful AI supercomputer. The supercomputer will initially be based on NVIDIA's DGX B200 and in a follow-up effort be equipped with Grace Blackwell chips, the more advanced version
- This strategic partnership aims to accelerate AI adoption across sectors such as telecommunications, robotics, and healthcare in Japan
- The two companies also announced deals to provide AI services on 5G networks and to create an AI marketplace for local compute

Why it matters

- SoftBank's initiative aligns with founder Masayoshi Son's ambition to make bold AI investments. It positions the firm as a leader in leveraging advanced AI hardware and Son said he's preparing to 'swing for the fences' in AI bets
- For Japan this is a big step to regaining a lead in AI. The country struggles with an aging society, labour shortages & overwork and sees AI as a big opportunity. The initiative thus aligns well with Japan's recently announced plan to invest $65B in its chip and AI industries, signalling a nationwide push to increase its global competitiveness

Strengthening US and its allies: OpenAI calls for North American AI compact

Summary

- OpenAI proposed “North American Compact for AI” urging US and neighbours to streamline access to talent, financing and supply chains for AI
- The proposal was included in new policy blueprint unveiled by OpenAI in Washington and details suggestions for how the US can maintain its lead in AI. OpenAI said the compact could later expand to include “global network of US allies and partners”
- The plan includes creating “AI Economic Zones” to expedite permitting processes, the government backstopping energy infrastructure projects as well as revitalizing nuclear energy capacity

Why it matters

- OpenAI’s plan underscores the increasing geopolitical dimension of AI, a technology seen by world leaders as crucial to future national competitiveness. As OpenAI said: It “presents a national security imperative to protect our nation and our allies against a surging China by offering an AI shaped by democratic values”
- The proposal of course also comes as the US prepares for a change in administration and could signal a shift among Silicon Valley firms looking to realign with anticipated policy priorities of the incoming Trump administration

Agent race accelerating. OpenAI to launch AI agent that will do work for users

Summary

- In a staff meeting on Wednesday, OpenAI announced plans to release AI agent codenamed "Operator" designed to perform tasks like writing code or booking travel in January 2025

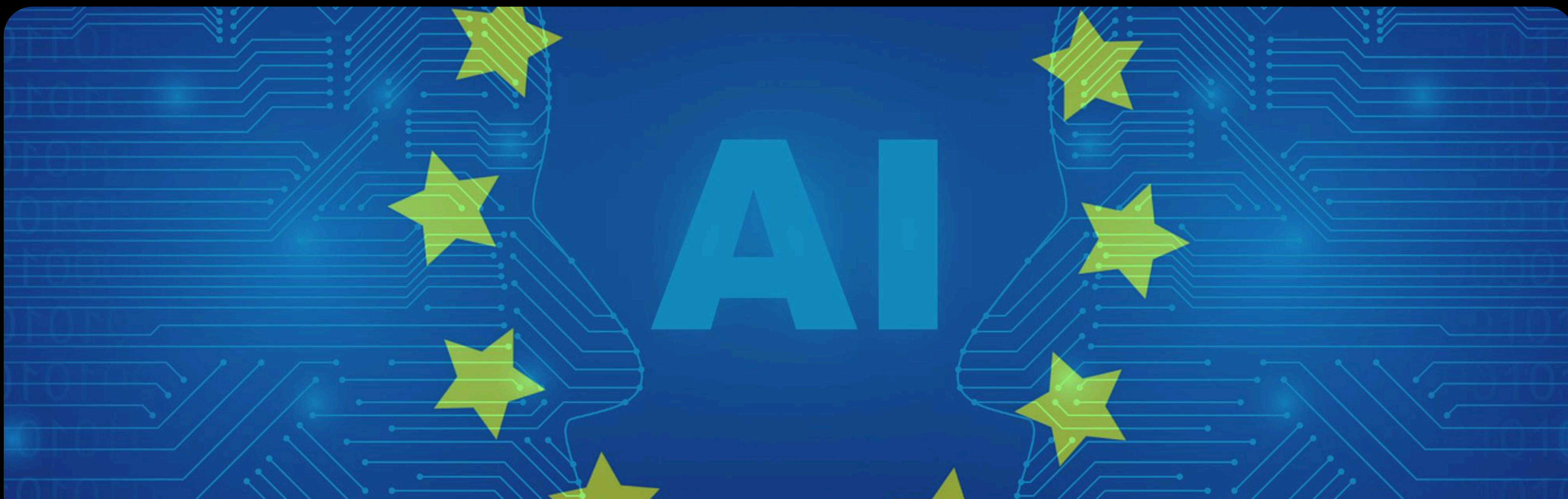
- The tool will initially launch as a research preview and through an API for developers. Operator is part of OpenAI’s broader effort to develop AI agents and positioning itself in the competitive landscape of AI-powered tools

Why it matters

- The announcement is part of a broader industry push toward AI agents that can manage multi-step tasks with minimal supervision. This is the next frontier for automation and human-machine collaboration set to unlock a new level of productivity for people and companies alike

- With the announcement OpenAI pulls level with other tech firms that recently announced AI agents. Anthropic, for example, launched “computer use” and Microsoft unveiled platform for enterprise ready, customizable agents (see also Newsletter CW 43)

EU AI Act taking shape. First guidance for general-purpose AI released



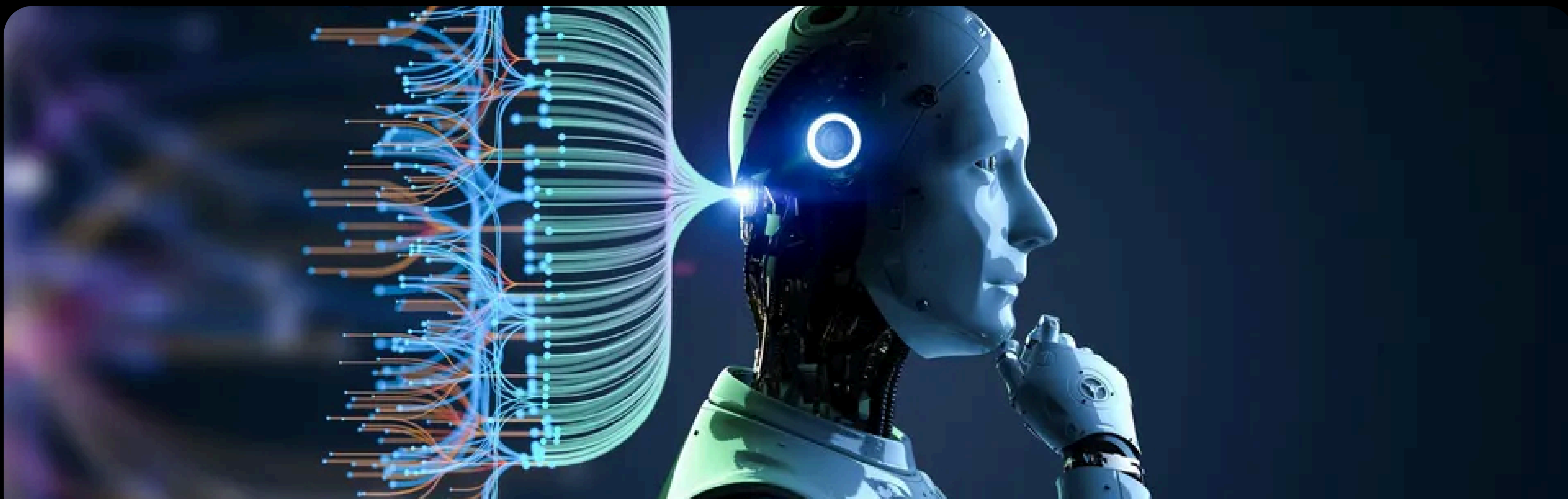
Summary

- The EU released first draft of the General-Purpose AI Code of Practice. The guidance applies to general-purpose AI systems (GPAIs) under the AI Act and importantly covers Generative AI
- The draft covers transparency and accountability obligations, risk management requirements and a safety and security framework
- The current version of the draft runs to 36 pages and is still relatively open and vague. The EU commission is inviting feedback until Nov. 28 and the drafting process will continue into next year

Why it matters

- The AI Act is the EU's risk-based regulatory approach to AI. It entered into force Aug. 1 2024. Its elements take effect in a staged approach. Rule for GPAIs apply from Aug. 2 2025 for new GPAI models. Models that entered the market previously have until 2027 to comply
- This is the first comprehensive law of its kind and widely watched by regulators globally. Moreover, the law has extraterritorial reach as it applies to all providers, even those based outside the EU, which place models on the EU market. Thus it has an impact of AI & GenAI firms globally

Two futures. Anthropic's CEO underlines his view on AGI timeline and impact



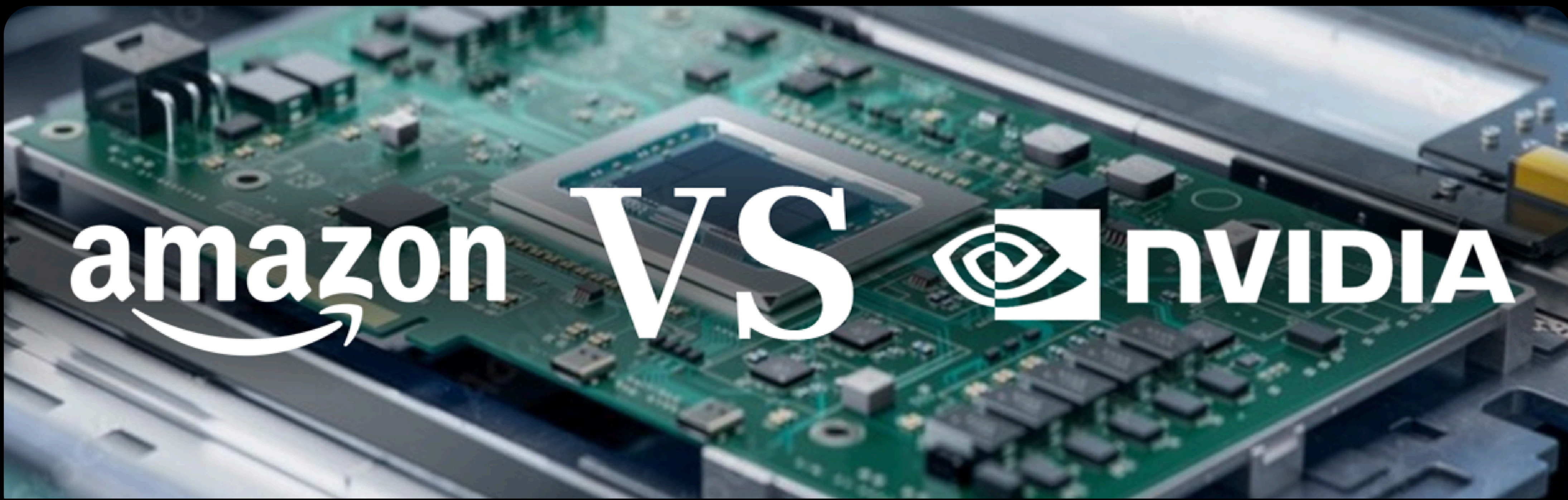
Summary

- In a Lex Friedman interview Anthropic CEO Amodei said he expects Artificial General Intelligence (AGI), AI systems surpassing humans across a wide array of domains, to occur by 2026 or 2027
- He reasons that based on (Gen)AI's current rate of improvement, having gone in the last two years from the level of a high school student to that of an undergrad to now PhD level, this is the most likely scenario
- Limiting factors connected to data and compute might delay this slightly. Yet he is adamant that truly convincing reasons for AGI not to happen in the next few years are rapidly decreasing

Why it matters

- In his assessment Amodei notably disagrees with OpenAI's Sam Altman. The latter is expecting AGI to take another five years to occur (still an incredibly short timeline!)
- Moreover, they also disagree on the impact this will have. Altman anticipates that "the AGI moment came and went" i.e. with surprisingly little societal change. Amodei on the other hand expects that AGI will kick-off a chain of scientific and economic breakthroughs yet also pose substantial new political challenges

Chip wars. Amazon steps up effort to build AI chips that can rival Nvidia



Summary

- Amazon increases competition with Nvidia, unveiling Trainium 2 AI Chips. Amazon is set to roll out Trainium 2, its newest AI chip developed by Annapurna Labs, next month. The move is part of its strategy to reduce reliance on Nvidia and optimize AI infrastructure
- Further upping the ante, Amazon also announced it will provide $110M in free computing credits to AI researchers wanting to use its custom Trainium chips as opposed to Nvidia's GPUs. The company plans to make 40'000 of the first-generation Trainium chips available for the program

Why it matters

- Several being seeing Nvidia's dominance in GPU market as challenging bottleneck. In August, Nvidia sparked worries that Blackwell chips would be delayed, pushing shipments to the first quarter of 2025 instead of later this year, sending shockwaves through the industry
- Amazon is not alone in seeking to diversify its chip supply. Last year November, for instance, Microsoft unveiled two custom-designed chips. Also OpenAI is set to create first own chip in partnership with Broadcom (see also CW 44 newsletter)