The great AI chip rush

Friday, July 10, 2026

Today's top story, from reporter Aili McConnon:

In recent months, a growing number of AI companies have accelerated efforts to develop custom silicon tailored to their own models and workloads.

OpenAI recently unveiled its first inference chip, Jalapeño, developed with Broadcom. IBM introduced the first **sub-1 nanometer chip** designed to lower the cost of running AI systems and improve efficiency. Meta continues to invest in custom AI accelerators, and Microsoft is developing its Maia family of chips to support AI services across its cloud infrastructure. Anthropic, meanwhile, is reportedly exploring its own chip initiative as it looks for ways to optimize compute and diversify its infrastructure options.

The result is that everyone—from model makers to cloud providers to enterprise AI vendors—is trying to control more of the hardware stack. As Rynne Whitnah, Technical Lead for AI Ecosystem at IBM, put it on this week's episode of the Mixture of Experts podcast, vertical integration allows companies to "take over more of the supply chain." In an industry where access to compute can determine who trains the next generation of models first, owning more of that stack is becoming a strategic advantage, she said.

That strategy is also a response to a simple market reality. "Compute is in high demand; it's hard to come by," Gabe Goodhart, Chief Architect of AI Open Innovation at IBM, said on the podcast. "There's a supply-demand imbalance." Building proprietary hardware gives AI companies a way to secure more compute while reducing their exposure to supply constraints and competitors.

There's also a technical rationale behind the move. Mihai Criveti, Distinguished Engineer for Agentic AI at IBM, believes the industry is heading toward a mix of general-purpose and specialized hardware, and most companies will likely use both custom chips and more generic ones depending on the use case. For instance, NVIDIA chips will continue to power a wide range of AI workloads, but dedicated accelerators can be more efficient when optimized for a specific model, he said. "It runs just this one specific model, but it runs it really, really well," Criveti said, noting that custom chips can improve inference performance and speed up "time to first token."

And so, what began as a race to build the best models is increasingly becoming a race to control the infrastructure that powers them.