

Less fighting over GPUs, more focus on the real AI work

Friday, May 15, 2026

Today's top story, from reporter Aili McConnon:

Over the past year, enterprise AI has been contending with a new constraint: access to GPUs. As demand for training and inference has surged, costs have spiked—often without close oversight—pushing teams within organizations and entire companies to compete for compute. Now, as enterprises take a closer look at their AI budgets, that bottleneck is beginning to ease and the conversation around AI is shifting with it.

At this week's Red Hat Summit in Atlanta, that shift was unmistakable, said Briana Frank, VP of Product and Design for IBM Cloud, speaking from the show floor on the latest episode of the Mixture of Experts podcast. "There's really no one here that I've run into that's afraid of AI. They're embracing it, and they're excited about what's possible and what's next." The conversation has evolved, said Frank, from proofs-of-concept to how to make AI work at scale inside large organizations.

One signal of that evolution is IBM's push to simplify inference through a new standalone Red Hat AI service that will be generally available on May 22. By streamlining the process of choosing and provisioning GPUs, the service allows companies to run production-grade AI models without the complexity of managing GPUs, infrastructure or AI platforms. "You only pay for what you use," Frank said on the podcast. "You can scale up if you need, and you don't have to fight your colleagues for the GPU."

But eliminating the GPU bottleneck doesn't necessarily solve the harder problems. And if anything, it exposes them. Across conversations at the conference, Frank said, the focus has shifted to the messy realities of operating AI in production. "How do you secure AI?" she asked. "How do you have consistent architecture on-prem and in the cloud?" The answers to these questions determine whether AI delivers real business value, she said.

That's part of why consulting and integration are suddenly back in focus. The barrier to progress isn't just technical, Frank said; it's organizational, and many companies are discovering that adopting AI requires rethinking how work gets done. That includes documenting processes more rigorously, creating transparency and designing workflows that AI agents can plug into.

In that sense, transformation begins with behavior, not technology.

There's also growing recognition that design will play a critical role in what comes next, Frank said. As AI becomes embedded in every product, users will expect systems that surface the right information and help complete tasks seamlessly.

"Each product we develop is going to have to be designed to get the best out of AI," she said.

The easing of GPU constraints doesn't mark the end of the AI race. It marks the beginning of a new stage—where success depends less on access to compute, and more on how effectively companies can redesign the way they work.