GPU Infrastructure

At Nvidia's GTC 2025 conference, CEO Jensen Huang unveiled a wave of new AI infrastructure built for what he calls the era of "AI factories", vast, industrial-scale computing setups designed to train and run the world's most powerful AI models.

The hardware may sound technical, but here's what you need to know: this is a massive leap in speed, efficiency, and scale. And if your organization uses or plans to use AI in a meaningful way, you'll want to pay attention.

What's New?

Smarter, faster chips: Nvidia introduced the new Blackwell Ultra architecture, which includes something called a **tensor core**—a part of the chip that acts like a turbo-charged calculator for the complex math used in AI. It also features a **transformer engine**, which is hardware designed to speed up the specific kind of AI model behind tools like ChatGPT (called transformers).

Smaller numbers, bigger speed: These chips can use **4-bit arithmetic**—a way of doing calculations using very small numbers. That may sound less powerful, but it actually makes the chip faster and more efficient, especially for AI tasks where absolute precision isn't always needed.

More memory: Each chip can hold up to **288 GB** (gigabytes) of high-speed memory. One GB equals about a billion bytes of data. More memory means the chip can handle bigger models and data sets at once.

Rack-scale power: The new **Vera Rubin Ultra SuperChip** combines an **88-core CPU** (think of it as 88 processors packed into one) with a high-performance GPU and stacks of memory. These chips can be bundled into an **NVL576 rack**, which holds 576 of them and is cooled by liquid, not fans. Each rack draws about **600 kilowatts (kW)** of power—that's like running 500+ microwaves at once.

Unprecedented speed: One rack can deliver up to **15 exaflops** of performance. An **exaflop** is a billion-billion math calculations per second. That's more than the fastest supercomputers on Earth just a few years ago.

Faster connections, lower power use: Nvidia also introduced **co-packaged optics** and **silicon photonics**—technologies that use light instead of electricity to move data. This kind of **optical networking** allows for lightning-fast communication between chips. These systems now reach **terabit** speeds—that's one trillion bits of data per second, reducing delays and energy use dramatically.

Why Should You Care?

These systems aren't just for scientists or big-tech labs. They mark a shift in how organizations will build and deploy AI. The improvements in speed, power, and energy use make it possible to train and use incredibly powerful AI models faster, cheaper, and more securely.

For business leaders, this isn't about buying a rack of GPUs. It's about understanding that access to next-generation AI power is becoming essential to compete—and that cloud providers will soon be offering it to anyone who's ready.

What Should You Do?

1. **Plan for scale.** If you run AI workloads today, start assessing whether your infrastructure partners can support high-density compute (and the power and cooling it requires). It might be time to think in racks, not servers.
2. **Talk to your cloud provider.** Nvidia's latest chips have begun appearing in major cloud platforms. Ask your provider when Blackwell Ultra and Vera Rubin will be available, and start benchmarking if they're worth adopting.
3. **Think security-first.** Blackwell includes hardware-based "confidential computing," which isolates sensitive data even from the host system. That's a big win for AI applications in regulated industries.
4. **Tie it to your energy goals.** These systems can deliver more performance per watt, but they still require serious power. Ensure your AI growth aligns with your company's carbon and sustainability goals.

AI infrastructure is rapidly moving toward an industrial model - factory-scale systems delivering real-time intelligence. That means companies who treat AI as a central capability—not just a side project will be best positioned to lead.

Hippification

When most people think of AI GPUs, one name comes to mind: Nvidia. Their GPUs have long been the industry standard for AI infrastructure. But there's a quiet shift happening with AMD GPUs and '***Hippification***'.

With the rise of open-source AI tools and growing demand for high-performance computing, AMD's new line of GPUs - the Instinct MI300X series - are becoming serious contenders in the AI race. So why does it matter? And what should you know?

Let's simplify it with a quick recap: A GPU (Graphics Processing Unit) is a chip originally designed to render complex images and video. AI workloads, especially training large language models, rely on similar types of linear algebra: performing billions of calculations quickly and in parallel. That's why GPUs, engines for high end visual effects have also become the engines of modern AI.

Nvidia dominates this space, but AMD's MI300X is one of its first high-end challengers. It includes 192 GB of HBM3 (High Bandwidth Memory), which enables it to process massive datasets at ultra-fast speeds—ideal for large AI models.

Most AI models today are trained using Nvidia GPUs and a proprietary software platform called CUDA, which stands for Compute Unified Device Architecture. CUDA lets you run code in parallel across thousands of GPU cores—making it ideal for large-scale AI computations.

AMD uses a different programming model called ROCm, short for Radeon Open Compute. Here's what sets ROCm apart: it's open-source, built on standard Linux operating system tooling, compatible with popular AI libraries like PyTorch and increasingly with TensorFlow, and improving rapidly.

But what's Hippification?
Hippification is the process of translating code from Nvidia's software platform called CUDA into AMD's software platform called ROCm - essentially "Nvidia-speak" to "AMD-speak."

If your workload is already written in Nvidia's platform and you want to consider moving to AMD's platform, the good news is that you don't have to rewrite everything. AMD offers a process called 'Hippification'—a whimsical term for portability between ecosystems.

HIP, short for Heterogeneous Interface for Portability, is the programming layer inside ROCm that developers write their code in. ROCm is the full software platform that supports AMD GPUs—like CUDA is for Nvidia. Think of ROCm as the toolbox, and HIP as the tool you write code with. If your code is written in HIP with portability in mind, it can often run on both AMD and Nvidia GPUs.

For organizations planning long-term AI infrastructure composed of a mix of Nvidia and AMD GPUs, this flexibility can be a game-changer.

Nvidia has long been the default choice. It still is for many. But we're entering a new phase of AI: more competitive, more cost-conscious, and more open. AMD, with its ROCm ecosystem, offers a real alternative to Nvidia's CUDA. Whether you're training AI models, choosing hardware, or designing cloud infrastructure, understanding both Nvidia and AMD ecosystems gives you more leverage and helps future-proof your decisions.

Takeaways for Leaders & Builders

1. Evaluate whether AMD + ROCm fits your use case—especially for new AI initiatives.
2. If you already use Nvidia's CUDA, explore 'Hippification' to assess portability.
3. For greenfield projects, consider starting with HIP to keep vendor options open.

Nvidia set the standard and AMD is now helping shape the future.