Ai2 Drives Open-Source AI Innovation with NVIDIA Hopper GPUs and Cirrascale Cloud Services

"The most capable open source AI model with visual abilities yet."[1]

This was not a headline Ai2 received when model development was fragmented across multiple data centers.

But it *was* their dream.

Ai2 is a nonprofit research institute dedicated to developing fully open source AI solutions that improve the work and lives of people around the world. They'd made incredible progress training increasingly powerful models, especially considering the limits of the resources at hand. What they needed next wasn't just more compute power but a scalable, forward-thinking strategy to launch Ai2 into the next chapter of open source research and innovation.

> "Ai2's scientists and engineers were ready to make big moves in the truly open AI space. We have expertise across the entire stack, but we were lacking the right infrastructure to meet our needs," says Michael Schmitz, Director of Engineering at Ai2.

They needed a company that could offer the flexibility to design a best-in-class setup for their NVIDIA Hopper GPU cluster—with the headroom to scale as they grew over time. They were not looking for the typical hands-off cloud provider who'd sell them hours, but provide little to no support. Rather, Ai2 needed a trusted partner in co-creating custom, highly performant AI infrastructure.

AT A GLANCE

Challenge
Deploy an initial cluster of 1,024 NVIDIA® Hopper GPUs with NVIDIA Quantum-2 InfiniBand efficiently and prepare a data center with significant room for growth for Ai2, a non-profit AI research institute.

Solution
Through a close managed services approach and a focus on proactive growth planning, Cirrascale delivered a customized, scalable infrastructure designed to support Ai2's long-term goals and reduce TCO.
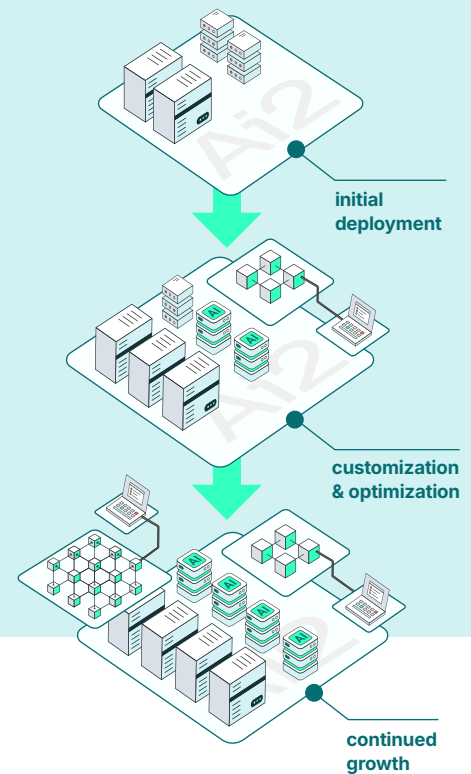
Business Value
- **Dramatic increase in compute power.** Enables Ai2 to train complex models quickly, driving innovation forward.
- **Custom setup with scalable infrastructure.** Offers high performance for Ai2's specific jobs and ensures Ai2 can expand to support future workloads.
- **Increased reliability and business continuity.** Delivers stability and efficiency for improved job completion rates.

Cirrascale Cloud Services was an easy choice

The Cirrascale team had an existing relationship with Ai2, having deployed an initial NVIDIA GPU cluster for them. They had witnessed firsthand the evolution of this research institute as they trained increasingly larger models and consistently pushed compute boundaries.

> **"We tried to anticipate where they're going to be in a year and make sure we took that into account when architecting a solution for them,"** says David Driggers, CTO of Cirrascale.

Key to Ai2's confidence in Cirrascale was their dedication to customize and optimize throughout the process, ensuring a successful deployment and a solution that would supercharge Ai2's capabilities for their current needs and future growth.

initial
deployment

customization
& optimization

continued
growth

Collaborating as true partners

Unique to Cirrascale was their flexibility as a boutique hybrid cloud provider to customize a solution tailored to their customer's needs. For Ai2, this meant enabling them to purchase a portion of their hardware while maintaining a longer-term relationship with Cirrascale for setup, configuration, and ongoing management and optimization.

> **"We also set up a live channel with Ai2 where we can maintain a regular dialogue, checking in to make sure everything is going well and supporting them if anything comes up,"** says Driggers.

This way, Cirrascale is ready to step in and manage technical needs, letting Ai2 focus on what they do best: building breakthrough AI aimed at solving the world's biggest challenges.

THE SPECS

The main AI training cluster: 1,024 NVIDIA Hopper GPUs

Networking: 3200 Gbps NVIDIA Quantum-2 InfiniBand Platform

Storage: Two-tiered multi-petabyte Weka.io storage cluster

Data center: State-of-the-art water-cooled facility in Austin, TX

Designing a complete solution for efficiency and reliability

Ai2 knew they wanted the NVIDIA accelerated computing platform and recognized that Cirrascale offered a much better storage option than the competition. Beyond that, they trusted Cirrascale to recommend the rest, from networking setup to infrastructure. Cirrascale ultimately proposed a comprehensive solution that prioritized business continuity and delivered as much performance as possible within the budget. Because of the way the team built the solution, selected materials, and set up the data centers, Ai2 is experiencing much higher job completion than they would have seen from traditional cloud providers.

> **"Cirrascale has been fantastic because of their willingness to customize every aspect of our setup,"** says Schmitz.

A key challenge for Ai2 was that they anticipated their workloads expanding considerably as they grew. Cirrascale addressed this by reserving space in their data center, ensuring Ai2 can seamlessly expand their compute capacity and adopt emerging technologies as their needs evolve.

The expansion has already begun, with Cirrascale deploying many more NVIDIA GPU clusters to augment the initial cluster for specific needs. As part of their ongoing efforts to remain at the cutting edge while reducing TCO, Ai2 is also pursuing an ultra-efficient NVIDIA HGX B200 cluster.

Empowering AI breakthroughs to solve the world's biggest problems

These days, headlines about "redefining LLM innovation"[2] and "driving 'critical shift' in AI development"[3] are the norm for Ai2.

> **"The new cluster is completely transformational for the company. And we have a long ramp for being able to grow our presence in the data center,"** says Schmitz.

They now have the infrastructure to power the training of large scale, truly open models with more parameters, more data, and in shorter amounts of time. This has enabled Ai2 to make leaps and bounds with their large language and multimodal models like Molmo, OLMo, and Tülu 3, cementing their position as a driving force in the open-source AI movement.

Enabling open-source AI initiatives to compete

The AI world moves fast. With a scalable, flexible infrastructure designed to grow alongside their workloads, Ai2 is equipped to keep up.

This partnership with Cirrascale ensures that, as Ai2 continues to push the boundaries of truly open AI, they'll have the resources and capacity to compete, and the infrastructure support to transform their bold visions into world-changing realities.

1 "The Most Capable Open Source AI Model Yet Could Supercharge AI Agents"

2 "How OLMo From Ai2 Redefines LLM Innovation"

3 "Allen Institute for AI releases 'truly open source' LLM to drive 'critical shift' in AI development"

Cirrascale Cloud Services

5775 Kearny Villa Road, San Diego, CA 92123 USA   **Phone** 888-942-3800   **Web** www.cirrascale.com