Crusoe

Five examples of AI infrastructure done right

CONTENTS

Infrastructure shouldn't be the hard part.

We are currently witnessing a shift in how technology is built. For AI startups and enterprise innovators, the “generalist” cloud model is breaking. Waiting months for capacity, deciphering opaque pricing, and wrestling with legacy hardware that wasn’t designed for high-performance computing (HPC) is no longer a viable roadmap.

At Crusoe, we think like mountaineers. We believe that audacious goals require intense preparation and the right tools. Speed and efficiency are not just metrics — they are the new competitive moats.

We’ve gathered five real-world stories from the frontier of AI development. From the creators of the world’s fastest coding assistant to the architects of PyTorch, these builders faced critical roadblocks in cost, latency, and control.

They didn’t scale by compromising. They scaled by choosing an infrastructure built specifically for the climb. Here is how they did it.

Scaling an AI coding assistant to 800,000 developers

The benefits

Windsurf

50%
cost savings compared to traditional cloud providers.

99.98%
cluster uptime, maintaining reliability for enterprise clients.

800,000+
developers supported on the new infrastructure.

> **Windsurf's NVIDIA H100 GPUs on Crusoe have been incredibly reliable... This reliability, combined with the significant cost savings, has enabled us to scale our infrastructure confidently while maintaining healthy unit economics.**
>
> Varun Mohan
> CEO & Co-founder
> Windsurf

The context:
Windsurf (creators of the Codeium IDE) built a product that developers loved too much, too quickly. In roughly one month, they hit over 100,000 weekly active users. Their tool, which keeps developers in a "flow state" via multi-step reasoning, required massive, instant GPU availability.

The bottleneck:
Success created a crisis of scale. As their user base exploded to 800,000 developers, Windsurf's existing infrastructure strained. Traditional hyperscalers offered a grim choice: wait months for capacity or accept pricing that hinders their ROI. They needed to scale immediately without breaking the bank.

The solution:
Windsurf refused to let infrastructure slow them down. To break the bottleneck, they migrated their GPU workloads to Crusoe Cloud, securing immediate access to NVIDIA H100 GPUs. Leveraging Crusoe's streamlined onboarding experience, Windsurf's engineering team executed the migration and was fully operational in just one day. They now capitalize on VM boot times of under 90 seconds (the industry "gold standard") to dynamically scale capacity the moment their users need it.

Achieving sub-120ms latency for real-time voice

Cartesia

The benefits

Sub-120ms
latency, establishing Sonic as the fastest model in its class.

3x
cluster expansion within the same NVIDIA Quantum InfiniBand fabric.

Custom
Slurm integration deployed specifically for Cartesia's workload.

> **So much of the competitive advantage of AI models depends on time to market… relationships like this one; rooted in nimbleness, high-performance, and customer success are crucial.**
>
> Cartesia Team

The context:
Cartesia is redefining how AI sounds. Their Sonic model is designed to generate speech so lifelike and fast that it feels like a human conversation. To achieve this "multimodal intelligence," they utilize State Space Models (SSMs), which require a specialized, high-throughput inference stack.

The bottleneck:
For Cartesia, latency is the enemy. To make a voice assistant feel "present," the model must react in real-time. They needed an infrastructure partner capable of supporting sub-120ms latency while allowing them to expand their cluster size without fragmented networking slowing them down. They also required complex job scheduling that standard cloud support teams couldn't handle.

The solution:
Cartesia leveraged NVIDIA H100 GPU clusters on Crusoe Cloud for optimal price-performance. Crucially, Crusoe didn't just provide servers; we integrated into Cartesia's engineering workflow. When Cartesia needed to implement Slurm for job scheduling, Crusoe brought in a subject matter expert to configure and deploy a custom Slurm cluster, ensuring their SSMs could train and infer without friction.

Turning hardware telemetry into efficiency gains

The benefits

Neuralwatt

33%
increase in AI inference throughput.

40%
reduction in idle GPU power draw.

33%
improvement in density, running 8 GPUs.

> **On Crusoe, we've had access to everything we needed. We have access to all of the NVIDIA System Management Interface... nothing has been restricted or unavailable to us, which isn't the case with a lot of the other providers we've tried.**
>
> Chad Gibson
> Co-founder
> Neuralwatt

The context:
Neuralwatt is tackling one of AI's biggest elephants in the room: energy consumption. Their software optimizes power usage at the hardware level, increasing server density and reducing waste. But to prove their software worked, they needed to run a data-driven demo on NVIDIA H100 GPUs.

The bottleneck:
You can't optimize what you can't touch. Neuralwatt needed deep access to the NVIDIA System Management Interface (SMI) to manipulate power states. Most major cloud providers lock this down, restricting the very telemetry Neuralwatt needed to access. As a bootstrapped startup, they also faced the "chicken-and-egg" problem of needing expensive compute to prove the value of their solution to investors.

The solution:
Through the Climate Collective accelerator, Crusoe provided Neuralwatt with credits and, more importantly, technical freedom. Neuralwatt gained unrestricted access to the bare-metal telemetry of NVIDIA H100 GPU clusters running in Crusoe's renewable-powered Iceland data center. This allowed them to test a critical hypothesis: that granular power management could allow them to run 8 GPUs in the power envelope typically reserved for 6. The validation was successful, turning an abstract theory into a tangible breakthrough in density and efficiency.

Pushing the boundaries of training speed

Meta

The benefits

34-43% **training acceleration** using Float8 rowwise.	85% **reduction in checkpoint time.**	Validated **2,000 NVIDIA H200 GPUs,** performance at enterprise scale.

The context:

As models grow larger, training times become the primary bottleneck for innovation. The PyTorch team, in collaboration with Meta, is constantly developing new techniques to make training faster and more efficient. They needed a massive, stable sandbox to validate these next-gen features.

The bottleneck:

As models grow larger, training times become the primary bottleneck for innovation. The PyTorch team, in collaboration with Meta, is constantly developing new techniques to make training faster and more efficient. They needed a massive, stable sandbox to validate these next-gen features.

The solution:

The team utilized a 2,000 NVIDIA H200 GPU cluster on Crusoe Cloud to validate next-gen training techniques at scale. Beyond proving that Float8 rowwise training could maintain convergence while accelerating the process, they tackled the “checkpointing” bottleneck. By deploying PyTorch Distributed Checkpointing (DCP) with TorchTitan, the team validated asynchronous checkpointing on the cluster. This approach allows training to continue while the model state saves in the background, drastically minimizing GPU downtime.

Key takeaway:

This research highlighted how asynchronous checkpointing minimizes GPU downtime… Crusoe’s infrastructure played a key role in validating these improvements and their potential to save valuable training time.

Solving the orchestration “cost conundrum”

union

The benefits

Zero

infrastructure management overhead when using Union Self-Managed.

100%

Kubernetes-native, ensuring reproducibility and scalability.

Predictable

pricing model, removing the volatility of hyperscaler billing.

The context:

Union.ai is the driving force behind Flyte, the open-source platform defining the standard for reliable, reproducible machine learning orchestration. Yet, even with the right tools, building the model is only half the battle. Engineering teams often struggle with the “infrastructure tax” — the time spent managing fragile pipelines and MLOps instead of improving the model itself.

The bottleneck:

Teams were previously trapped in a dilemma: shoulder the heavy maintenance burden of self-hosting open-source tools, or accept the unpredictable, skyrocketing bills often associated with traditional hyperscaler managed services. They needed a way to orchestrate complex workloads that offered enterprise-grade reliability without the volatility of legacy cloud pricing.

The solution:

Union.ai partnered with Crusoe to create a unified path. Teams can now deploy open-source Flyte directly on Crusoe for maximum control, or choose Union Self-Managed. This enterprise-grade solution handles the complex orchestration control plane for you, while keeping your data and compute securely within your Crusoe environment. This hybrid approach eliminates infrastructure management overhead while capitalizing on Crusoe’s cost-efficient compute.

Key takeaway:

Teams need more than just access to the latest generation chips... they need strategic partners and a reliable system that can handle all the messy details without complex cost structures.

Crusoe

Build the future faster

Explore Crusoe Cloud →

At Crusoe, we believe that you can only climb fast if you trust your gear. We've built a cloud designed with the specific rigors of AI in mind, considering factors such as reliability, energy sourcing, and cost. This results in a foundation of resilience that can help turn your infrastructure from a constraint into a competitive advantage.

The bottleneck isn't your code. Don't let it be your cloud.

Ready for your own breakthrough?

Privacy Policy

Crusoe is an NVIDIA Preferred Partner