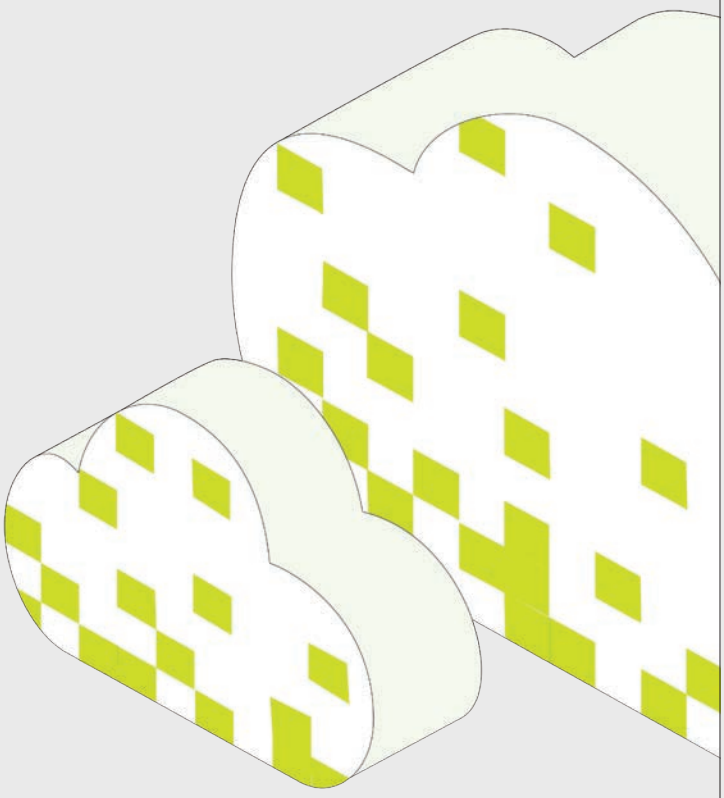
The ultimate guide:



Getting started with AI cloud

TABLE OF CONTENTS:

We are at a pivotal moment

For the first time in history, we are able to manufacture intelligence. The potential of AI is immense, offering unprecedented opportunities for innovation, but it also presents a significant challenge. For many builders — from developers and machine learning engineers to data scientists — the right infrastructure can be the difference between a breakthrough and a roadblock.

Crusoe exists to solve this problem. As a vertically integrated AI factory company, we aim to build the world's favorite AI cloud. To us, "favorite" means a reliable, feature-rich platform that is purpose-built to reduce complexity and accelerate your time to value. We are experts in AI workloads and have invested heavily in the usability of our platform, enabling you to get your projects running quickly and with consistent performance. Our team is also relentlessly focused on providing hands-on customer support and guidance to help you succeed. By freeing you from the constant orchestration of underlying infrastructure, you can focus on what matters most: building your products.

This guide is your compass for navigating the AI cloud journey. Choosing the right AI cloud isn't just technical; it's a strategic imperative dictating your speed, efficiency, and capacity to innovate. At Crusoe, we're building a future where energy is the foundation of intelligence, empowering you to build ambitiously and responsibly. This is about unlocking massive human prosperity gains, creating jobs, and powering a world where builders can truly reach their AI potential. Let's build the future of AI, together.

Chase Lochmiller
Co-Founder and CEO
Crusoe

Chapter

1 The AI imperative: Scaling your ambition

The scenarios that demand AI cloud

Every ambitious AI journey hits a point where local machines and conventional setups simply can't keep up. It's not a failure of your team; it's a sign of success. These are the crucial moments that call for a different kind of infrastructure — one built for the boundless potential of AI. Recognizing these signals is the first step toward unlocking true scale and impact.

Scenario 1:

Your AI product is exploding and you need to scale now

You've hit product-market fit. Your user base is surging, and your AI application is in high demand. That's incredible! But the thrill can quickly turn to dread if your infrastructure becomes a blocker. Meeting user expectations, ensuring seamless performance, and capturing market share requires **reliable, on-demand compute** that can flex instantly with your growth. Waiting for hardware, battling allocation queues, or managing an overwhelmed on-premise setup just isn't an option when your success depends on rapid, continuous delivery.

Scenario 2:

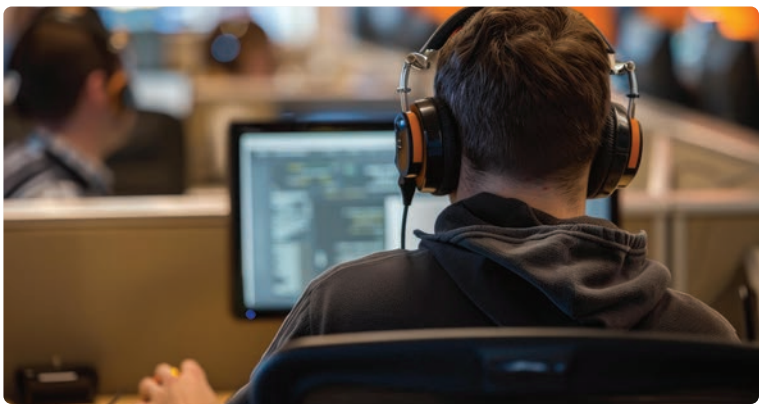
Moving from prototype to production with confidence

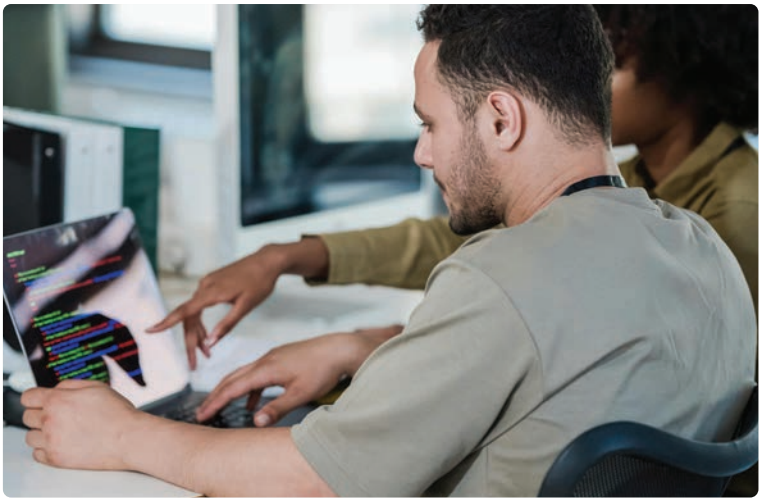
You've proven your AI concept, trained a compelling model, and showcased its value. Now, it's time to take it from a lab experiment to a resilient, production-grade service. This transition demands more than just raw compute. You need an environment that offers **robustness, security, consistent performance, and predictable costs**; foundational elements that most local or general-purpose cloud providers struggle to provide at scale. Moving to production isn't just about making it work; it's about making it work flawlessly, every time, for every user.

Scenario 3:

Accelerating internal AI initiatives across the enterprise

For enterprises, AI isn't confined to a single team anymore. It's a strategic lever for optimizing operations, enhancing decision-making, and driving innovation across departments. But often, siloed resources, inconsistent compute environments, and a lack of centralized GPU access slow progress to a crawl. To truly transform your organization with AI, you need **unified, high-performance infrastructure** that empowers every team to build, test, and deploy AI solutions with speed and agility.





Scenario 4:

The strategic pursuit of competitive AI advantage

In the race for AI leadership, iteration speed is king. Your ability to experiment, train new models, and refine existing ones faster than your rivals can define your market position. This isn't about incremental gains; it's about exponential acceleration. To outpace the competition, you need **immediate access to cutting-edge GPUs and flexible, optimized environments** that allow your researchers and developers to turn innovative ideas into tangible results without delay. The speed of your infrastructure directly translates to the speed of your innovation.

Scenario 5:

Unpacking the hidden costs of DIY AI infrastructure

You might have started with a few GPUs under a desk or a small server rack. But as your AI needs grow, the "DIY" approach reveals its true, often exorbitant, cost. Beyond the initial hardware purchase, there are the significant, unpredictable expenses of **power, cooling, real estate, maintenance, and the specialized talent** required to manage and optimize this complex infrastructure. These hidden costs can quickly spiral, diverting critical resources from core AI development. A purpose-built AI cloud offers predictable, cost-efficient scaling, freeing you to focus on innovation — not infrastructure.

Chapter

2

Charting your course: Navigating the AI cloud landscape

Hyperscaler vs. specialized neocloud

Once you recognize the imperative for an AI cloud, the next critical step is choosing the right partner. The landscape generally presents two primary paths: the broad, generalist hyperscalers and the emerging, AI-focused neoclouds. Each has its advantages and trade-offs.

Chapter 2:

Hyperscaler vs. specialized neocloud

The hyperscaler path (the broad approach):

The large and familiar cloud providers offer an expansive ecosystem of services for virtually any computing need, but they aren't built for AI.

Pros:

Familiarity, a vast array of general cloud services (storage, networking, databases, etc.), and often strong initial free tiers or credits. If your business needs a bit of everything and AI is just one small component, a hyperscaler might seem like the convenient choice.

Cons:

While they offer GPUs, their focus is broad. You often encounter a "Sea of Services" where navigating a huge number of tools and configurations can create difficulties in selecting the right ones, leading to a steep learning curve and longer implementation times. Crucially, this generalist approach means their AI expertise might be layered on top, rather than embedded deeply in the infrastructure itself. This results in less performance per dollar and a more complex, less optimized environment. Furthermore, navigating complex billing, unpredictable egress fees (costs for data leaving their network), and rigid contractual terms can be challenging. For serious AI development, securing large, contiguous blocks of the latest GPUs can feel like a constant battle, leading to serious resource constraints and delays.

The neocloud path (the AI-native approach):

This new category of cloud providers, like Crusoe Cloud, is purpose-built from the ground up to serve the unique, demanding needs of AI workloads.

Pros:

Purpose-Built for AI Workloads: Every aspect of their infrastructure, from GPU interconnects to networking, is optimized for AI training, inference, and model development. This means better **performance-per-dollar** and environments where your models can truly sing.

Optimized Developer Experience: They offer a streamlined, "out-of-the-box" experience with all the necessary drivers, tools, and scheduling software pre-installed. This removes the complexity and steep learning curves associated with generalist clouds, enabling teams to focus on innovation, not configuration.

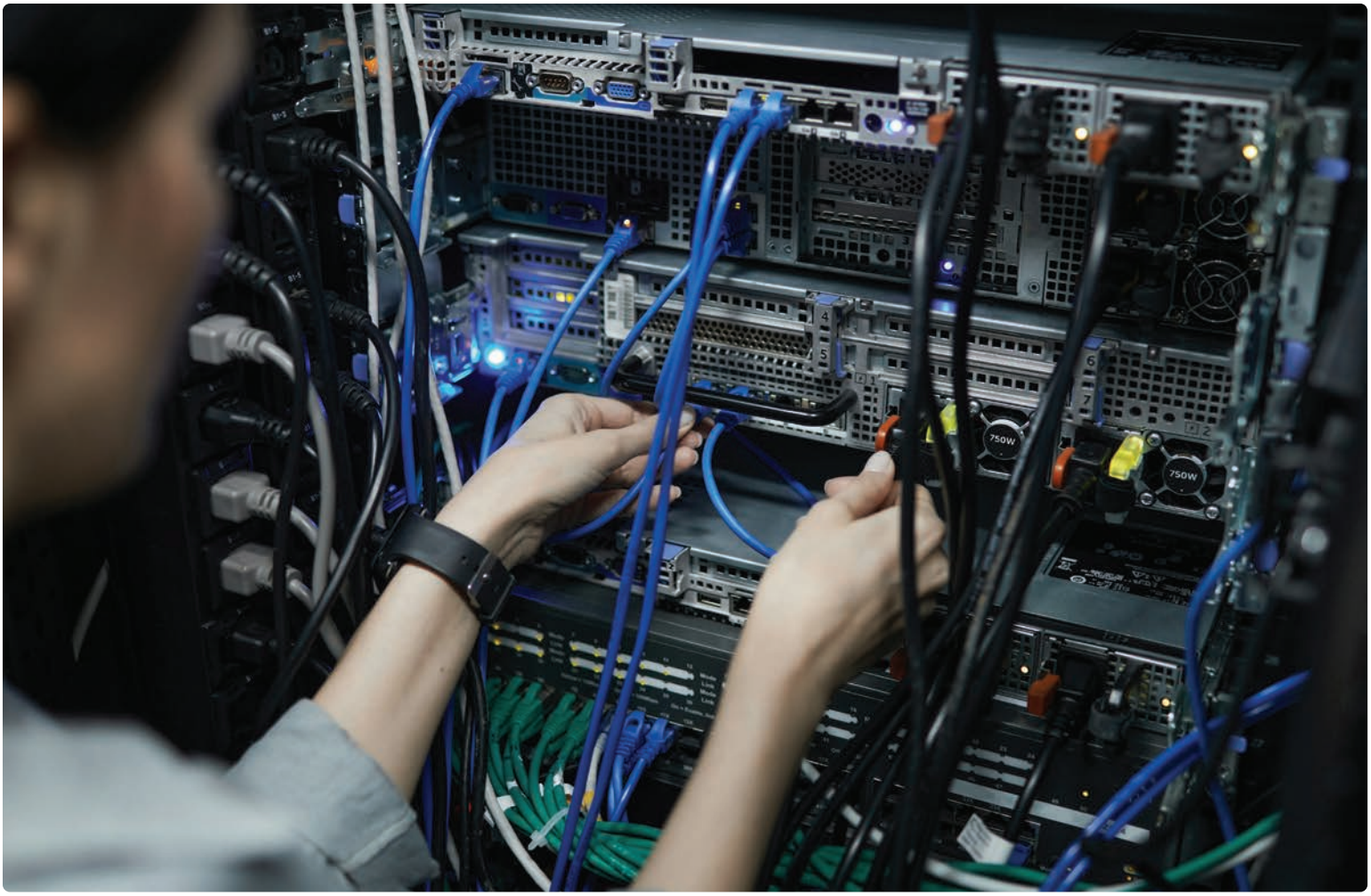
Predictable Pricing: Neoclouds typically offer transparent pricing models, eliminating hidden fees and making budget planning significantly simpler. You know what you're going to pay.

Direct Access to Infrastructure Experts: You're not talking to a general support desk. You're engaging directly with engineers who live and breathe AI infrastructure, understanding the nuances of distributed training, MLOps, and complex AI frameworks.

Focused Innovation: Neoclouds are constantly innovating specifically for AI, bringing the latest GPU technologies and specialized tooling to market faster.

Cons:

Neoclouds primarily focus on high-performance AI compute. They generally **do not provide generalized IT services** (like managed databases for non-AI applications, extensive serverless functions for general web apps, etc.) that hyperscalers do. If your needs are exclusively AI-driven, this isn't a "con," but it's important to understand the focused scope.



A Brief Note on Self-Hosting:

While some very large enterprises with deep pockets and specialized teams consider building their own mini-data centers and buying GPUs directly, this path demands significant upfront capital, extensive infrastructure expertise, and a long lead time for procurement, construction, and ongoing maintenance. It's a highly specialized endeavor, often proving impractical for most businesses aiming for rapid AI deployment and focusing on their core innovation. For most, the agility and expertise of an AI cloud make far more strategic sense.

How Crusoe Cloud fuels breakthroughs

Choosing a specialized AI cloud isn't just about getting GPUs; it's about strategically equipping your team with the right tools to accelerate discovery, scale ambition, and outmaneuver the competition. Here's why this path isn't just an option — it's an advantage.

Unmatched performance, purpose-built for AI:

At Crusoe, we don't just host GPUs; we architect environments where they thrive. The core difference lies in how the infrastructure is designed. For large-scale AI training and inference, the interconnect between GPUs (like NVIDIA NVLink™ and NVLink Switch) is as critical as the GPUs themselves. Specialized clouds prioritize this, building clusters with high-bandwidth, low-latency connections that enable GPUs to communicate seamlessly, often leveraging technologies like NVIDIA Quantum **InfiniBand or RoCE**. This translates to significantly faster training times, more efficient model development, and the ability to tackle larger, more complex AI problems without compromise. Our instance types are specifically configured for AI workloads, often delivering superior performance-per-dollar because every component is optimized for your compute-intensive tasks.

Lowered costs, unlocking budget clarity:

One of the biggest headaches with generalist clouds can be the unpredictable billing; the "death by a thousand cuts" of egress fees, mysterious service charges, and fluctuating spot pricing. Crusoe Cloud champions **transparent, pricing models.** You get clarity from day one, allowing you to plan your budget with confidence and allocate resources to core AI development rather than untangling complex invoices. This financial predictability is both convenient and a strategic enabler for long-term project planning and investment.

Scalability, on demand:

The AI race moves fast. Can you get 100 or 1,000 NVIDIA Hopper GPUs today? With hyperscalers, this often means fighting for allocations, waiting in queues, or settling for less powerful hardware. Crusoe Cloud is built to provide **rapid access** to large clusters of the latest GPUs. This means your projects never stall, your team's velocity remains high, and you can seize opportunities the second they arise. We ensure you have the compute you need, when you need it, to keep your ambition unbounded.

Expert support, in the trenches with you:

When you're pushing the boundaries of AI, you need more than just generic IT support. You need partners who speak your language, like engineers who intimately understand the complexities of distributed training frameworks like PyTorch Distributed, orchestrators like Kubernetes or SLURM, and the nuances of large language models. At Crusoe, our support isn't just responsive; it's proactive and deeply specialized. We're in the trenches with you, offering white-glove support and collaborating closely to architect scalable deployments aligned with industry best practices. This human layer of expertise is invaluable, helping you solve complex challenges faster and build with greater confidence. We are proud to share that we've achieved a 100% customer satisfaction rating for 13+ consecutive months, proving our dedication to superior customer service.

156GW

of projected AI-related data center capacity demand by 2030.

29%

of businesses say managing technical complexity is their top challenge, according to initial internal research.

The Crusoe difference: Integrated for impact

Crusoe stands apart by delivering a fully integrated AI factory — from sustainable energy sources, to purpose-built data centers, to our leading AI cloud platform. This vertical integration is our competitive moat, enabling unparalleled speed, cost-efficiency, and environmental responsibility. We're not just providing compute; we're built to solve the three core challenges that can stifle the AI revolution:

The power constraint:
According to a report by McKinsey & Company, AI's exponential growth will require a massive investment, with companies needing to invest $5.2 trillion into data centers by 2030 to meet worldwide demand. This is based on a projected 156 gigawatts (GW) of AI-related data center capacity demand by 2030, with 125 incremental GW added between 2025 and 2030. We are tackling this head-on by building our own AI-optimized data centers at sites with abundant clean energy and maximum Perf/W, providing a pathway for scalable and sustainable compute.

The deployment delay:
While the average time to develop a new data center is a multi-year effort, Crusoe's vertically integrated model allows us to achieve a record-breaking development time, turning over our hyperscale data centers in under 12 months. This speed delivers critical capacity to builders when they need it most.

The complexity of AI:
Initial internal research suggests that managing technical complexity is a top challenge for 29% of businesses. We eliminate this friction with an optimized, out-of-the-box experience, from pre-installed drivers to seamless tooling, allowing your team to get to work immediately. The 2025 ClusterMAX™ Rating System from SemiAnalysis reveals Crusoe Cloud VMs are able to boot up in under 90 seconds, turning a time-intensive process into a simple, rapid one.

Building your advantage

This energy-first approach is key to meeting AI's power demands without compromising our planet. With Crusoe, you can build ambitiously and responsibly with a partner that removes roadblocks to innovation and provides a direct path to a more intelligent, abundant future.

Chapter

3

Crusoe Cloud in action: Real-world impact

Seeing is believing. Here's how leading innovators are leveraging Crusoe Cloud to accelerate their AI breakthroughs and reshape their industries.

Crusoe has been an outstanding partner from the get-go — all of our in-house machine learning models have been trained on Crusoe Cloud. They provide a level of quality of service, responsiveness, and support for early access programs that we couldn’t find with any other cloud provider.

Prasanth Veerina
Co-Founder

Pixelcut

Windsurf

50%
cost savings with Crusoe Cloud

99.9%
cluster uptime delivered

1 day
to get up and running

Windsurf's NVIDIA H100 GPUs on Crusoe have been incredibly reliable with a cluster uptime of 99.98%. 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 challenge:
Windsurf rapidly became the AI code assistant platform of choice for over 800,000 developers, but this explosive growth strained their existing infrastructure. They needed reliable, cost-effective GPU capacity that could scale dramatically without impacting their healthy unit economics. Traditional cloud providers quoted prices that threatened their rapid expansion.

The Crusoe solution:
Windsurf turned to Crusoe Cloud for immediate access to NVIDIA H100 GPUs at significantly lower costs than hyperscalers. Our streamlined onboarding, praised for its "gold standard" out-of-the-box experience (with VMs booting in under 90 seconds), meant Windsurf was up and running in a single day. Beyond speed, Crusoe's infrastructure delivered the 99.98% cluster uptime critical for enterprise trust.

The results:
Crusoe Cloud enabled Windsurf to scale efficiently to millions of users while achieving 50% cost savings. This reliability and cost advantage empowered them to accelerate new feature deployment, like their AI-native IDE with "Cascade" agents, which quickly amassed over 100,000 weekly active users.

Decart

> Crusoe's infrastructure gave us the stability and performance we needed to focus fully on building MirageLSD and delivering a smooth launch.

Dean Leitersdorf
CEO & Co-founder
Decart

The challenge:

Decart embarked on an ambitious goal: to create MirageLSD, a video-to-video generative AI model that could seamlessly restyle a user's input video based on a prompt. This groundbreaking feat required a specialized AI infrastructure platform capable of handling immense datasets to train the model and deliver the performance needed for a flawless public demonstration. Decart needed a partner who could provide the scale and reliability for this high-stakes launch without the worry of infrastructure bottlenecks.

The Crusoe solution:

Crusoe was selected as the exclusive infrastructure partner. Our purpose-built AI environment, meticulously optimized for superior performance and reliability, provided the dynamic foundation required to train the model. This enabled Decart to focus entirely on its proprietary technology and deliver a powerful API to seamlessly restyle video files based on user input.

The results:

Decart successfully launched the MirageLSD demo to rave reviews, with Wired magazine publishing an article highlighting the impressive video-to-video technology. The successful demo proved Crusoe's ability to handle critical, market-defining workloads. With the model now exclusively available via API on Crusoe Cloud for B2B customers, Crusoe continues to be the trusted partner enabling pioneers like Decart to deliver next-generation AI applications to the world.

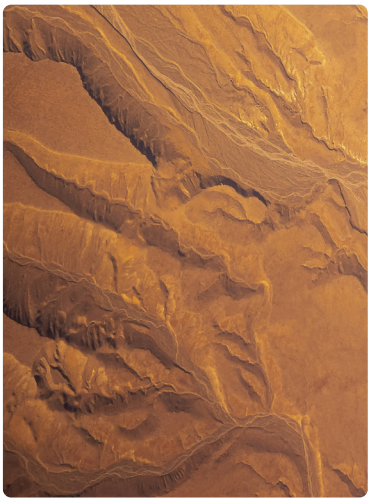
Chapter

4

Your first 30 days

A practical onboarding guide:

The leap to AI cloud might feel daunting, but it doesn't have to be. Here's a practical 30-day roadmap to guide your transition and ensure a smooth, powerful start with your new AI cloud partner.



Week 1:

Defining your pilot project

Don't try to move everything at once. Identify a single, high-impact workload or model that can serve as your pilot. This should be a project with clear objectives and measurable outcomes where you expect to see significant improvements in training time, cost, or scalability. This focused approach minimizes complexity and allows your team to get comfortable with the new environment without overwhelming them.

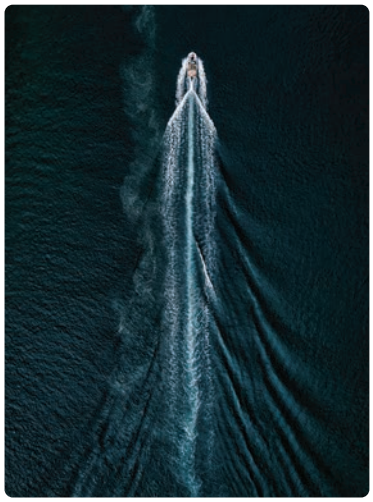
Week 2:

Data and environment setup

With your pilot project defined, it's time to prepare your data and configure your software environment on the cloud. Best practices include:

Data migration:
Develop a strategy for efficiently moving your datasets to the cloud. Consider data compression, secure transfer protocols, and storage solutions that align with your cloud provider's offerings.

Environment configuration:
Replicate your existing development environment as closely as possible. This involves setting up necessary libraries, frameworks (PyTorch, TensorFlow, etc.), and any custom tooling. Leverage your cloud provider's documentation and support to ensure a seamless setup. This is where a specialized cloud's deep expertise pays dividends.



Week 3:

Launching your first scaled training job

This is where the rubber meets the road. Take your pilot project and execute your first scaled training job on the AI cloud. Start with a "Hello World" at scale — a simple, well-understood model that you can run to validate the setup. Monitor performance closely, noting improvements in training speed compared to your previous setup. This initial success builds confidence and provides valuable baseline metrics.

Week 4:

Evaluating performance and cost

After your pilot run, it's time to dive into the numbers.

Analyze performance:
How much faster was your training? Did you hit any bottlenecks? What was the GPU utilization?

Evaluate cost:
Compare the actual costs of your pilot project on the AI cloud against your previous setup. Consider not just compute, but also power, cooling, and the efficiency gains from faster iteration.

Plan your next move:
Based on your findings, refine your strategy. What other workloads will you migrate next? What optimizations can you make? This continuous evaluation ensures you're always leveraging your AI cloud to its fullest potential.

Your infrastructure is your advantage

The future is being built with AI. And in this future, your infrastructure isn't just a cost center, it's **a competitive moat.** The right AI cloud partner enables you to build faster, smarter, and more responsibly than the competition. It frees your curious minds to focus on breakthrough innovation, rather than infrastructure headaches.

At Crusoe, we are pioneering the path to an abundant future where powerful AI meets sustainable energy. We deliver a fully integrated stack, purpose-built for AI, so you can scale ambitiously with speed, control, and integrity. This isn't just about providing compute; it's about accelerating human potential while bending the arc of energy toward sustainability.

About Crusoe

Crusoe is on a mission to accelerate the abundance of energy and intelligence. Crusoe provides a reliable, scalable, cost-effective, and environmentally friendly solution for AI infrastructure by harnessing large-scale clean energy, building AI-optimized data centers, and empowering builders to reach their AI potential. Crusoe empowers its customers to build the future faster.

Schedule a consultation with our AI infrastructure experts today.

semianalysis

Crusoe: A ClusterMAX™ Gold Standard

Leading industry analyst SemiAnalysis, in their comprehensive "GPU Cloud ClusterMAX™ Rating System," recognizes Crusoe as a top-tier provider, a testament to our commitment to performance, reliability, and value.

"Most customers are engaged in an AI arms race with their competitors, meaning that availability remains a critical differentiator... Crusoe excels in this domain, with extensive availability and the capability to provision substantial GPU clusters from initial contact to contract signing and provisioning within remarkably short timelines—often less than two days."

"Amongst GPU clouds that are highly competent on the technical front... Crusoe offers reasonable pricing and contract terms in addition to strong technical competency."

"Crusoe scores strongly as they are some of the few neoclouds that do multiple weeks-long burn in prior to handing over clusters to customers."

"Overall, Crusoe demonstrates significant responsiveness to user feedback and shows a strong commitment to rapidly evolving their product to meet customer needs and compete effectively in the GPU cloud marketplace."