Crusoe

From electrons to tokens
Expert insights on building the AI factory

CONTENTS

The decisions that define AI

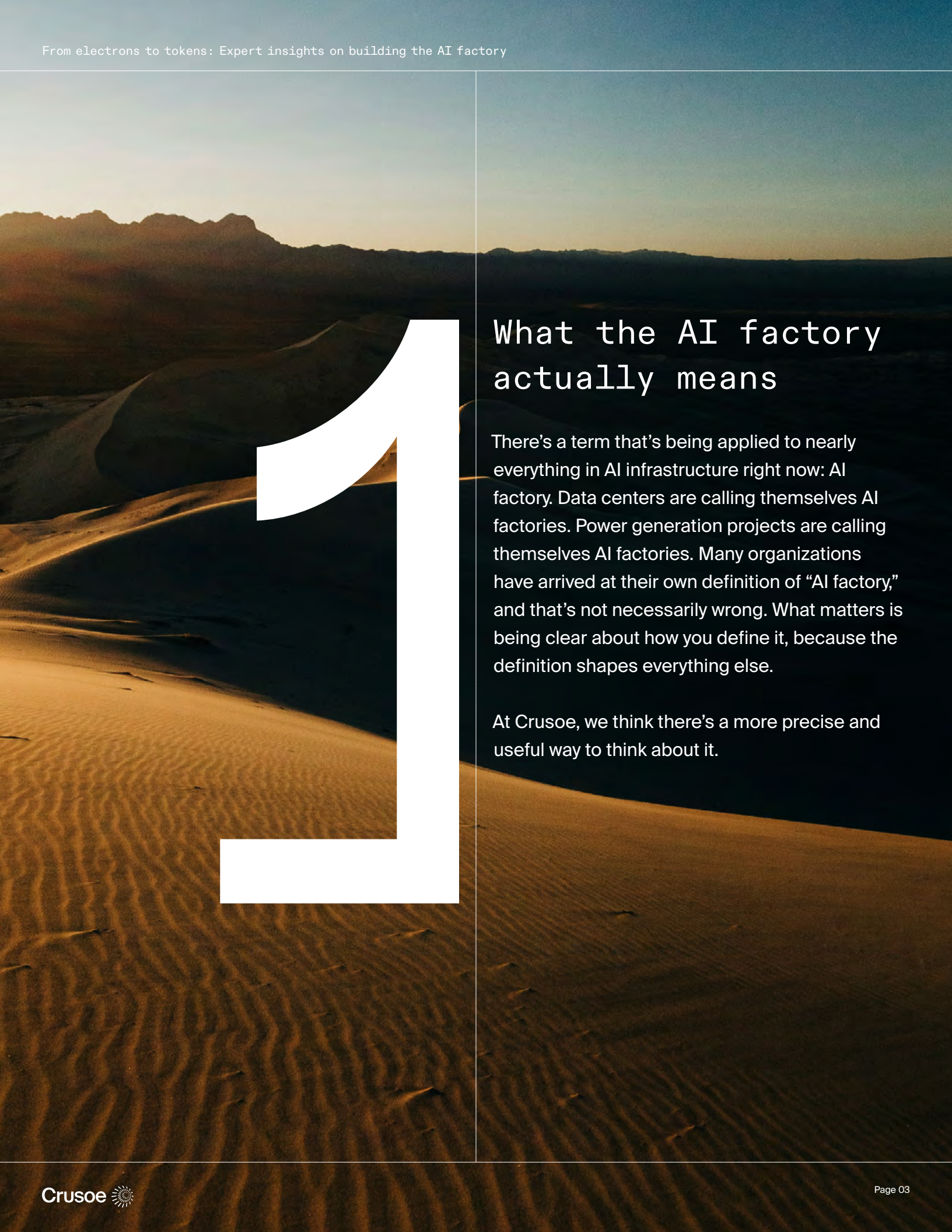
Intelligence is now manufactured. Every token produced is the result of electrons moving through infrastructure built specifically to convert energy into computation. And that chain, from power source to model output, is the new competitive advantage in AI.

That's not a metaphor. It's an engineering reality, and it has consequences for every team building on AI today.

The infrastructure decisions being made right now — where to source power, how to integrate the stack, what to manage versus what to hand off — will determine who can scale and who gets squeezed out. The ones that build on a foundation uniquely designed for the workloads they're running will prevail.

We sat down with five Crusoe leaders across engineering, product, infrastructure, and strategy to understand what's actually changed, what it means for the teams building AI now, and where the biggest opportunities and risks lie in the years ahead. What follows are their honest perspectives.

One idea surfaced in almost every conversation, unprompted: the factory analogy. Not as a reference to a data center for high-performance compute, but as a way of thinking about what it actually means to build AI infrastructure. A factory is defined by its output. The output here is intelligence. Everything else, including the energy, the hardware, and the software, exists to serve that end.



1

What the AI factory actually means

There's a term that's being applied to nearly everything in AI infrastructure right now: AI factory. Data centers are calling themselves AI factories. Power generation projects are calling themselves AI factories. Many organizations have arrived at their own definition of "AI factory," and that's not necessarily wrong. What matters is being clear about how you define it, because the definition shapes everything else.

At Crusoe, we think there's a more precise and useful way to think about it.

“An AI factory converts energy and data into intelligence. That’s the output. Everything else is in service of that.”

Cully Cavness
Co-founder, President & Chief Strategy Officer

That definition cuts through a lot of noise. A facility that generates megawatts is an energy facility. A building full of GPUs optimized for compute throughput is a data center. Neither of those is an AI factory unless intelligence is what comes out the other side. Think of it like a factory that makes ball bearings: it’s not a car factory, even though ball bearings go into every car. The end product is what defines the factory. An AI factory is defined by intelligence as its output; everything else is inputs and processes.

The implications of this definition go beyond semantics. If intelligence is the output, then every layer of the stack has to be evaluated by how well it serves that end.

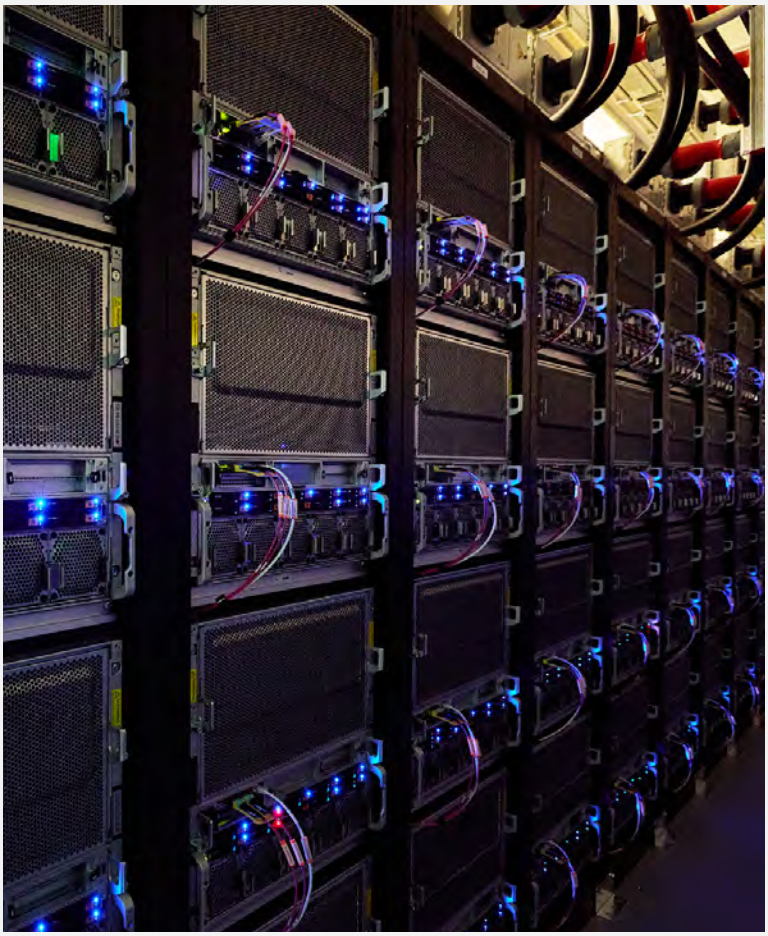
“It’s an Intelligence Factory. You plug something in and it outputs intelligence. Electrons to tokens.”

Kyle Sosnowski
VP of Software Engineering

You can’t optimize a system you haven’t defined. And you can’t define the system without first being clear about what it produces.

For teams evaluating AI infrastructure today, this definition is a useful filter. If your infrastructure isn’t purpose-built to produce intelligence (if it’s compute that happens to run models, rather than a system designed around that output) the gap will show up in your costs, your reliability, and your ability to scale.

This is what Crusoe means by vertical integration. We source and develop our own energy, build data centers engineered specifically for the demanding requirements of frontier computing, and operate a commercially-available Cloud platform that provides the infrastructure backbone for AI workloads. The full chain from electrons to tokens.



2

Energy is the constraint

For most of the history of data center infrastructure, power was a cost line. You found your location, you built your building, and you negotiated with the utility. The power showed up. That model no longer works in the era of the AI factory.

Grid constraints are real and permitting timelines are long. Costs are rising in markets where AI demand has concentrated. The infrastructure teams that are moving fastest right now are the ones who recognized early that energy isn't just an input: it's the constraint that determines where you can build, how fast you can grow, and what your economics look like at scale.

Crusoe's approach inverts the traditional model entirely. Rather than bringing power to infrastructure, the infrastructure goes to where power is most abundant, often in places where energy is being generated but not fully utilized, selling onto the grid at a loss or curtailed entirely.

"Instead of the old model of bringing power to AI data centers, Crusoe has flipped the script and moved data centers to where energy is most abundant. It is fundamentally easier to move data than to move power."

Tamanna Sait
VP of Engineering Cloud

What that looks like in practice on the data center side is described by Chris Dolan, our Chief Data Center Officer, who has been building this model from the ground up.

This is what it means to treat energy as a first-class variable rather than a background assumption. The wind producer gets a committed buyer, which stabilizes the economics of their project. Crusoe gets power at a cost that reflects its actual abundance rather than its scarcity in constrained markets. The climate case and the business case point to the same answer.

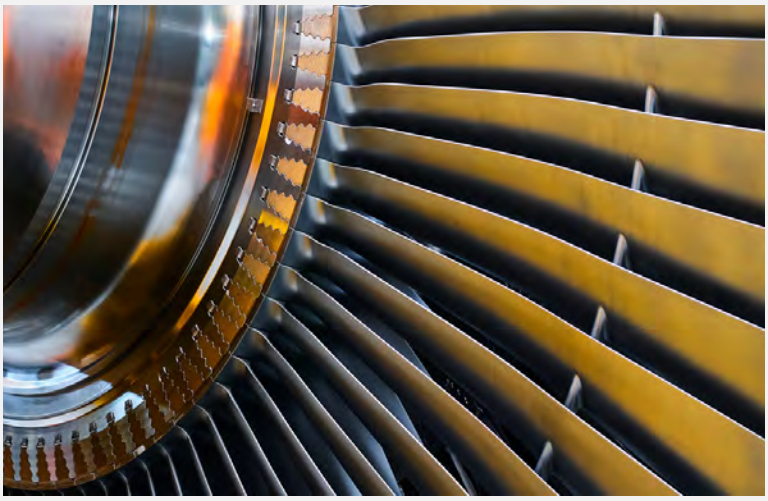
“By contracting for power directly behind the meter, before it reached the grid, we were able to support a wind producer facing losses from negative electricity pricing while securing the capacity our operations required.”

Chris Dolan
Chief Data Center Officer

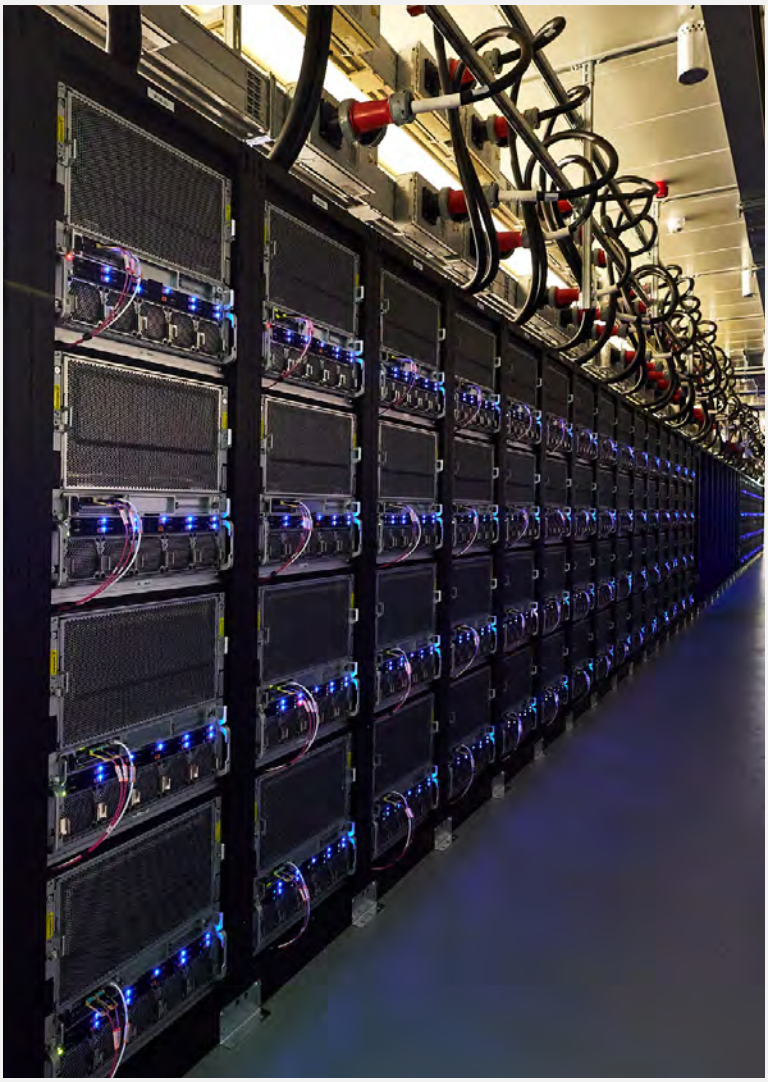
What makes this possible isn't just strategy; it's the depth of expertise on the team. A meaningful portion of Crusoe's leadership comes from power generation, upstream oil and gas, and energy development. That breadth of background is what allows Crusoe to structure deals that companies without that experience can't.

"We have a deep bench of energy expertise. In fact, a significant portion of our leadership comes from energy backgrounds, and that lets us make bets others can't."

Cully Cavness
Co-founder, President &
Chief Strategy Officer

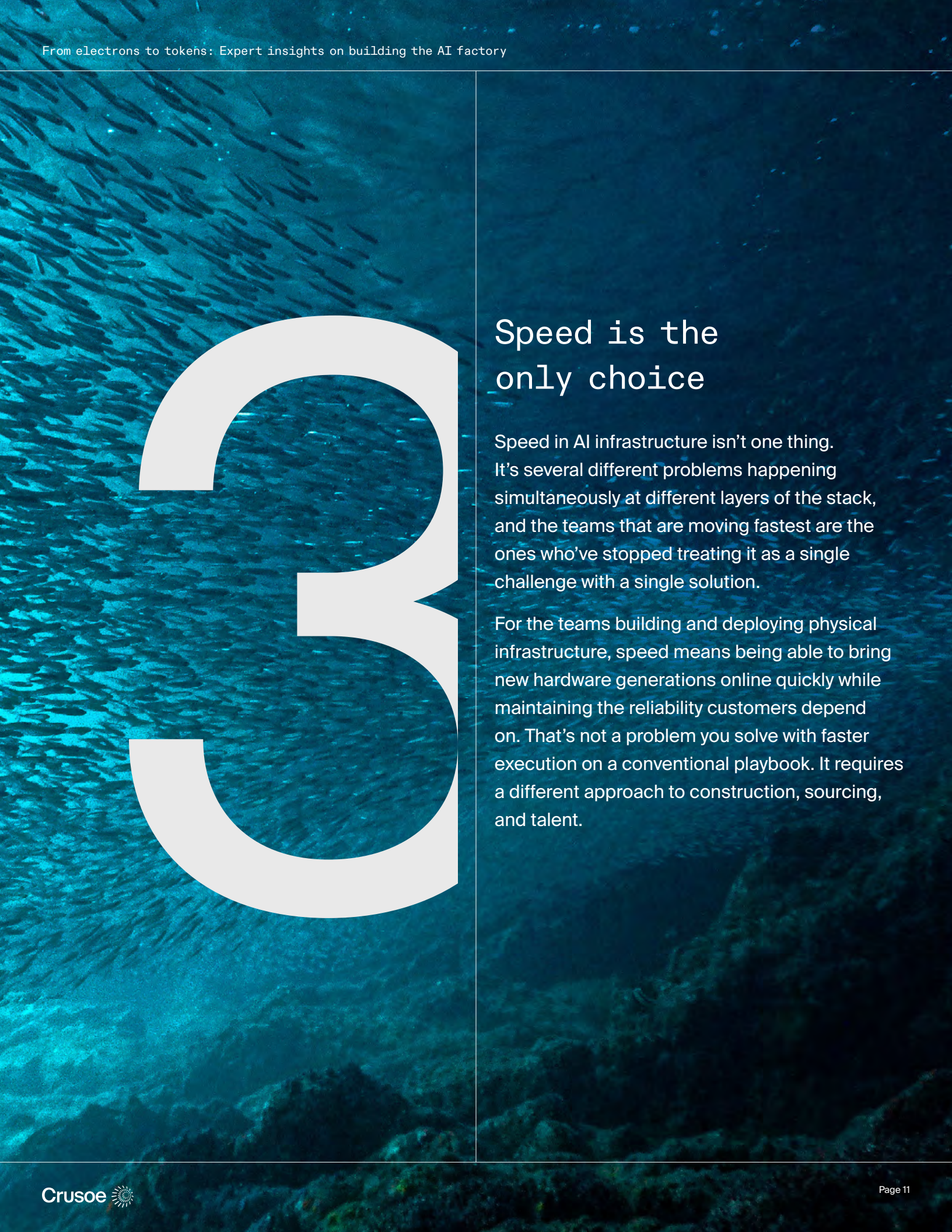


Erwan, our SVP of Product Management, offers a frame for understanding why that cross-domain expertise matters so much in practice.



“Energy people think of projects in terms of years. Data center people think in terms of quarters. Software engineers think in terms of days to weeks. Aligning those timescales, and building a team that speaks all three languages, is a big part of what makes this model work”.

Erwan Menard
SVP of Product Management



03

Speed is the only choice

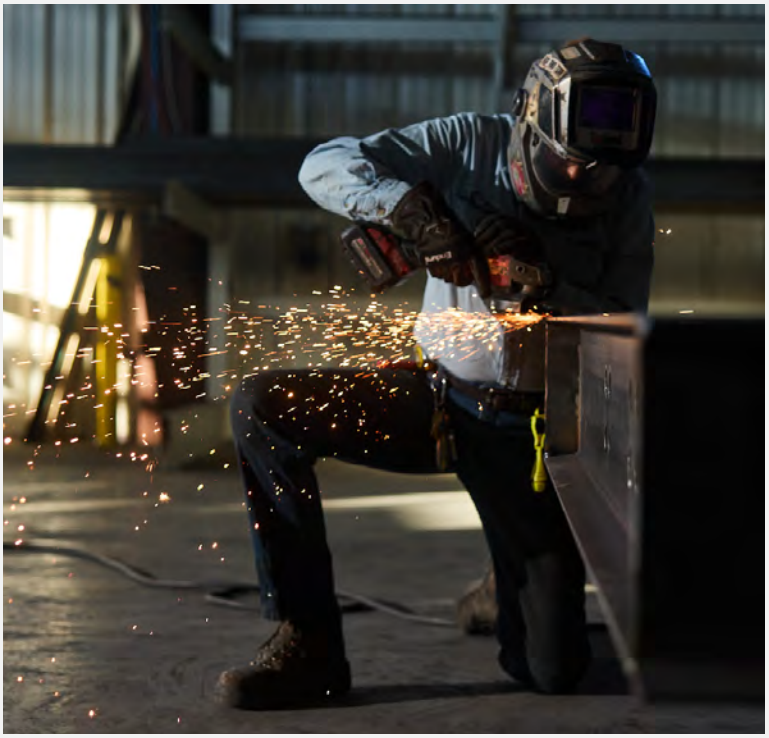
Speed in AI infrastructure isn't one thing. It's several different problems happening simultaneously at different layers of the stack, and the teams that are moving fastest are the ones who've stopped treating it as a single challenge with a single solution.

For the teams building and deploying physical infrastructure, speed means being able to bring new hardware generations online quickly while maintaining the reliability customers depend on. That's not a problem you solve with faster execution on a conventional playbook. It requires a different approach to construction, sourcing, and talent.

Crusoe's approach on the physical infrastructure side combines joint ventures with non-traditional contractors, vertical manufacturing of critical long lead data center components, a multi-vendor supply chain strategy, and offsite assembly of power infrastructure. The ultimate goal is to remove the bottlenecks that slow down conventional data center builds, not by pushing harder on the same constraints, but by removing those constraints.

"What keeps me up at night and wakes me up in the morning are two sides of the same coin: this insatiable need for power, and the opportunity it creates for us. Not just to keep up, but to lead."

Tamanna Sait
VP of Engineering Cloud



"Vertical integration is a critical success factor for speed and cost effectiveness. This is not talked about enough. Among all neoclouds, this is the one aspect that will pay off the most a few years down the road."

Erwan Menard
SVP of Product Management

For product and engineering teams, speed means something different: not having to manage infrastructure they didn't sign up to manage. Every hour an AI team spends on cluster maintenance is an hour not spent on the actual problem they're trying to solve.

The two definitions of speed aren't separate. They're connected at the architecture level. Infrastructure designed specifically for AI workloads allows for the kind of optimization that general-purpose infrastructure can't support. That optimization is what makes both kinds of speed possible.

CASE STUDY

Odyssey

Company	Odyssey
About	An AI lab pioneering general-purpose world models
Headquarters	Palo Alto, CA, USA

Odyssey needed infrastructure that could grow with their research, from NVIDIA HGX H100 GPUs to Blackwell, without rigid hyperscaler contracts. As an early design partner on Crusoe's Managed Kubernetes service, they burst capacity on demand instead of provisioning for peak load. Explorer launched to hundreds of thousands of users on schedule, with infrastructure costs up to 4.4x lower. Built to remove bottlenecks and scale without friction.

4

What people are still getting wrong

The biggest obstacle now isn't technical. It's the outdated assumptions people are still carrying about the market, about the technology, about their own teams.

"People often say AI is a bubble, making a reference to the internet bubble. But they forget that the demand here is massive. The internet bubble came from overprovisioning infrastructure ahead of unproven demand. Here, demand overwhelms supply in a magnitude we didn't think would exist anymore."

Erwan Menard
SVP of Product Management

The bubble question

The comparison to the dot-com bubble keeps coming up. It's worth addressing directly, because the people making that comparison are often confusing two very different situations.

Erwan, our SVP of Product Management, has been watching the technology industry for 25 years. His read on the demand side is clear:

That view carries an important nuance. The demand thesis is real, but infrastructure investment will ultimately track value creation. The companies that emerge strongest won't be the ones that simply scaled compute, but the ones that built infrastructure designed around actual workloads and measurable customer outcomes. The returns from this era won't be evenly distributed. They'll flow to the builders who got the fundamentals right.

Both things are true. The demand is real. The returns will not be evenly distributed. The companies that will look back on this period well are the ones building on infrastructure designed for the actual workload, not retrofitted for it, and the ones focused on the value they're actually creating for the customer, not just how much compute they're consuming.

"I've been around the block for a couple of decades and I've never seen this pace of innovation sustained the way it has been. People may be expecting a plateau, but we haven't reached it yet. This could very well be the new pace."

Erwan Menard
SVP of Product Management

The complacency risk

There's a quieter failure mode that doesn't get enough attention, and it may be doing more damage than the bubble debate.

Tools and use cases that genuinely didn't work six months ago may work now. The models have improved that fast. Teams that evaluated an AI-assisted workflow last year, decided it wasn't ready, and moved on, without building in a process to re-evaluate, are already operating on stale information. That's not a technology problem. It's a discipline problem.

The implication is uncomfortable but worth sitting with: if you're not actively re-challenging your assumptions about what AI can and can't do, your competitive position is quietly eroding. Not because a competitor deployed more compute, but because they kept testing while you stopped.

5

The engineer's job just changed completely

Something fundamental shifted for software engineers in the last year, and the people still debating whether to accept it are already behind.

Code generation is no longer the main job. It's a starting point. Kyle, our VP of Software Engineering, has been writing code for 20 years. He watched it become something anyone can do in just 12 months.

What Kyle is describing as "keeping things on the rails" is something that resists easy definition but most experienced engineers recognize immediately. It's the judgment that comes from knowing not just how to build a system, but why it might fail. What the edge cases are. When the elegant solution is actually the fragile one. That hasn't been automated away.

"Coding is now a commodity. It's no longer the best use of an engineer's time to handwrite code. Effort will shift to system design, execution patterns, and keeping the thing on the rails."

Kyle Sosnowski
VP of Software Engineering

"AI can write the code, but it can't stop us from thinking. My engineers aren't worried about generating code. They're still worried about pulling together the right architecture and the right design. No AI can do that for you yet."

Tamanna Sait
VP of Engineering Cloud

Tamanna, our VP of Cloud Engineering, reinforces this from the management side. Her team isn't worried about generating code. They're spending their time on architecture and design: the decisions that require context, experience, and taste.

The engineers who are thriving in this environment are the ones who have made the shift from writer to orchestrator — moving from authoring code to directing it, evaluating it, correcting it, and integrating it into something that holds together under load. That's what "engineer as agent orchestrator" means in practice. It's not a metaphor for a future state. It's a description of what the best engineering teams are doing right now.

The result is that fluency with AI tools is no longer a differentiator. It's a baseline. Kyle put it directly: sending an engineer into the world without AI fluency is like graduating from medical school only knowing how to do surgery with a chainsaw. The tools are too fundamental to the job to treat as optional.

CASE STUDY Cartesia

Company	Cartesia
About	On a mission to build the next generation of AI: ubiquitous, interactive intelligence that runs wherever you are.
Headquarters	San Francisco, California, USA

Cartesia's team had one goal: stay focused on the state space models behind their Sonic voice AI, not the infrastructure underneath it. When they needed Slurm for job scheduling, Crusoe brought in a specialist and built the integration alongside them, rather than leaving Cartesia to figure it out alone. The team kept building. Sonic now delivers sub-120ms latency at scale, the fastest in its class, and Cartesia has tripled its cluster size since launch.

6

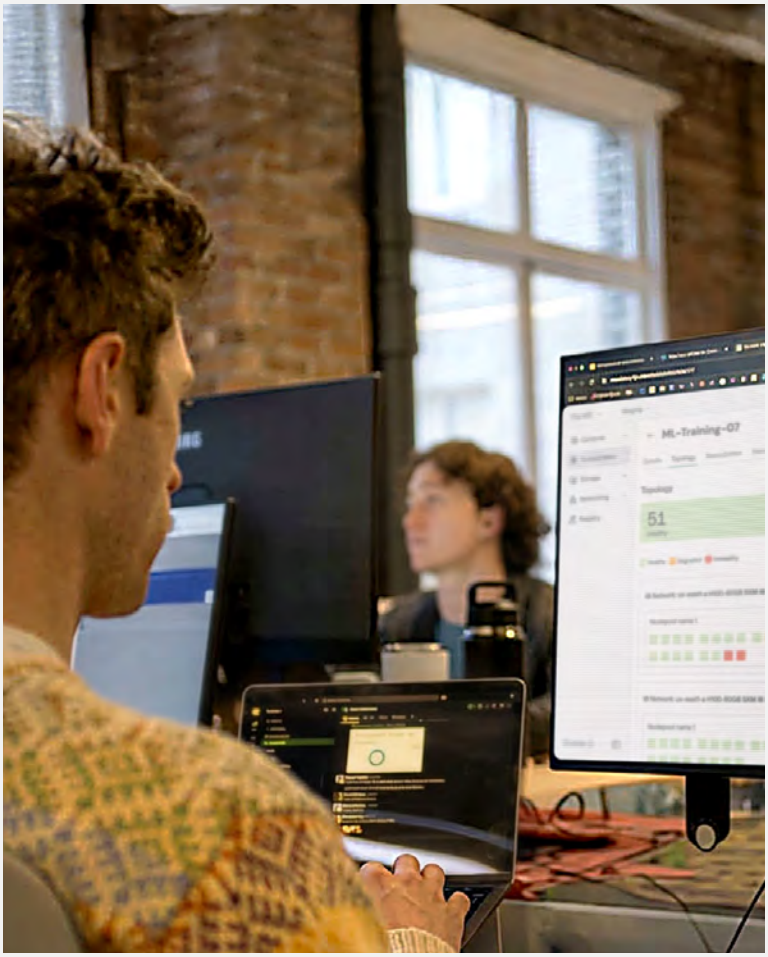
What customers actually need from AI infrastructure

The companies winning with AI right now aren't distinguished by how much compute they have. They're distinguished by what they're doing with it, and more specifically, by how little of their time they're spending managing the infrastructure underneath it.

The managed services stigma that existed two or three years ago is gone. The assumption that sophisticated AI teams would always prefer to own and operate their own infrastructure rather than depend on a cloud provider contradicts current market demand. Companies that offer only managed services are seeing surging demand. Customers want to run their AI. They don't want to babysit their clusters.

"The assertion that customers didn't want dependency on a cloud provider has been largely disproven. You can see it in the demand we're getting on Managed Inference, Managed Kubernetes, and in companies like Fireworks and Baseten, who offer only managed services and are getting very high traction."

Erwan Menard
SVP of Product Management



There's a specific gap that isn't being filled by the largest cloud providers: fine-tuned open models need a place to live. The major platforms are built around their own closed-source models. That works well for customers who want to use those models. It doesn't work for customers who have invested in fine-tuning their own, who also want the simplicity and quality of a managed inference endpoint without being forced onto a model they didn't choose.

The next step in this evolution is already visible in Crusoe's roadmap. The agent engine, managed infrastructure purpose-built for deploying AI agents in production, closes the loop on the electrons-to-tokens story. Energy becomes compute. Compute becomes a model. The model becomes an agent. The whole chain, managed.

"Customers should be focused on adopting managed AI solutions that simplify deployment and remove the complex infrastructure overhead. The goal is for them to focus on building, not on managing what they're building on."

Tamanna Sait
VP of Engineering Cloud

There's an important distinction at the heart of this shift. Today's AI infrastructure was largely built for researchers. These are the people pushing the boundaries of what technology can do. The engineers whose job is to take that technology and make it deliver real impact in the world have been underserved. That's the gap Crusoe aims to close. Kyle, our VP of Software Engineering, captures what that means in practice:

"Reliability. It doesn't matter what fancy automation you build if the stuff can't stay stable. People don't want to manage their infrastructure anymore. They want hands-off management so they can focus on their application."

Kyle Sosnowski
VP of Software Engineering

CASE STUDY Windsurf

Company	Windsurf
About	An "AI-native" Integrated Development Environment (IDE) that acts as an intelligent coding partner, designed to automate software development tasks through a highly contextual agentic system named Cascade.
Headquarters	Mountain View, CA, USA

Windsurf's Cascade is a collaborative AI coding agent serving over 800,000 developers. Running it in production required infrastructure reliable enough for enterprise customers and cost-efficient enough to support healthy unit economics; a combination hyperscalers couldn't offer. On Crusoe's H100 infrastructure, GPU costs dropped 50% and cluster uptime held at 99.98%. Windsurf onboarded in a single day and hit 100,000 weekly active users within a month of launching their IDE, one of the fastest growth curves for a developer tool in recent memory.

“We are taking an energy-first approach to AI. We bring a unique skill set around energy, a lot of passion for it, and we are looking to partner with new energy technologies as we build our infrastructure. That’s part of what makes Crusoe genuinely different.”

Cully Cavness
Co-founder, President & Chief Strategy Officer

Build beyond the horizon

The five leaders interviewed bring unique perspectives rooted in their expertise. Erwan sees AI demand being consistently underestimated and cautions against complacency. Kyle sees the role of the engineer changing faster than most teams have adapted to. Tamanna sees that change as clarifying rather than concerning. The broader view in the room: the infrastructure hockey stick will eventually rationalize, and value will not be created uniformly.

What runs through all of it is a single, operational truth. Producing intelligence at volume requires a system designed for that purpose from the source of energy to the point of delivery. General-purpose infrastructure, assembled from whatever pieces happen to be available, will not get you there. Every layer has to be optimized in service of the output, and the output is intelligence.

The teams that understand this now will be the ones who look back on this period well. Not because they consumed more compute, but because they built on a foundation designed for what they were actually trying to produce.

See what it looks like to build on infrastructure designed for this moment.

Meet the team

Ready to build
the future faster?

Schedule a consultation
with our AI experts today.

Cully Cavness
Co-founder, President &
Chief Strategy Officer

Chris Dolan
Chief Data Center Officer

Erwan Menard
SVP of Product
Management

Tamanna Sait
VP of Engineering Cloud

Kyle Sosnowski
VP of Software
Engineering