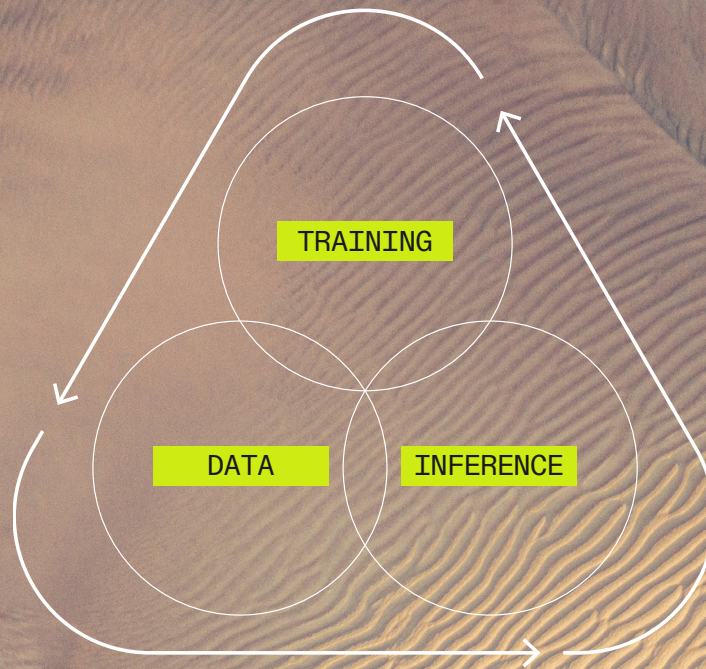




Building for Physical AI on Crusoe Cloud:

How Crusoe's AI Platform and Managed AI services support the physical AI lifecycle, from pretraining and synthetic data to fine-tuning, inference, and fleet operations at the edge.

CONTENTS

02	Introduction: The infrastructure problem
04	Physical AI lifecycle
05	Physical AI usecases
06	How Crusoe helps your physical AI development
09	Crusoe Cloud for Physical AI : AI Platform
10	Crusoe Cloud for Physical AI : Managed AI
11	A reference architecture: Cloud training, edge inference
13	Appendix: How to evaluate infrastructure for physical AI
14	Crusoe's evidence, independently measured
15	Sources and notes

Scaling physical AI is an infrastructure problem

Physical AI is AI that perceives, reasons, and acts in the real world. It has already left the lab and started being implemented across industries. Waymo runs roughly 500,000 paid autonomous rides a week across 10 U.S. cities ([Source](#), March 2026), on its way to a public target of one million by year's end. Aurora hauls freight driverless in Texas today ([Source](#), May 2025). And the money is following the miles: Goldman Sachs' humanoid robotics forecast has already been revised from \$38 billion ([Source](#), Feb 2024) to roughly \$138 billion ([Source](#), Sep 2026) by 2035, a number that's moved twice in under two years. Every dollar behind that curve needs GPU infrastructure to generate, train, simulate, and serve.

Scaling is where teams stall. Cisco found 61% of industrial organizations are deploying physical AI, while only 20% have scaled it successfully ([Source](#), April 2026). The gap is rarely the model; it's infrastructure built for text-and-image AI meeting a workload that's structurally different in five ways:

1. Data pipeline demands:

Handling petabytes of synchronized sensor streams, video, and joint states that require repeated ingestion across every training epoch, evaluation pass, and replay.

2. Diverse compute patterns:

Balancing three distinct workload profiles on a single infrastructure stack—finite, steady long-horizon pretraining; bursty, on-demand simulation and synthetic data generation; and continuous, perpetual 24/7 cloud inference serving active fleets.

3. Strict latency constraints:

Meeting 1–10 millisecond real-time control deadlines by distilling models and deploying them directly at the edge, a constraint absent in traditional LLM architectures.

4. Emergence of synthetic generation:

Replacing costly teleoperated collection (\$118/hour) with world models, introducing a compute-heavy, elastic synthetic data workload alongside training and inference. This shift was highlighted by NVIDIA's Jensen Huang at CES 2026 "Cosmos turns compute into data" ([Source](#), Jan 2026), and is already utilized by Agility Robotics ([Source](#), Mar 2025) to scale training datasets beyond real-world limits.

5. Closed-loop feedback:

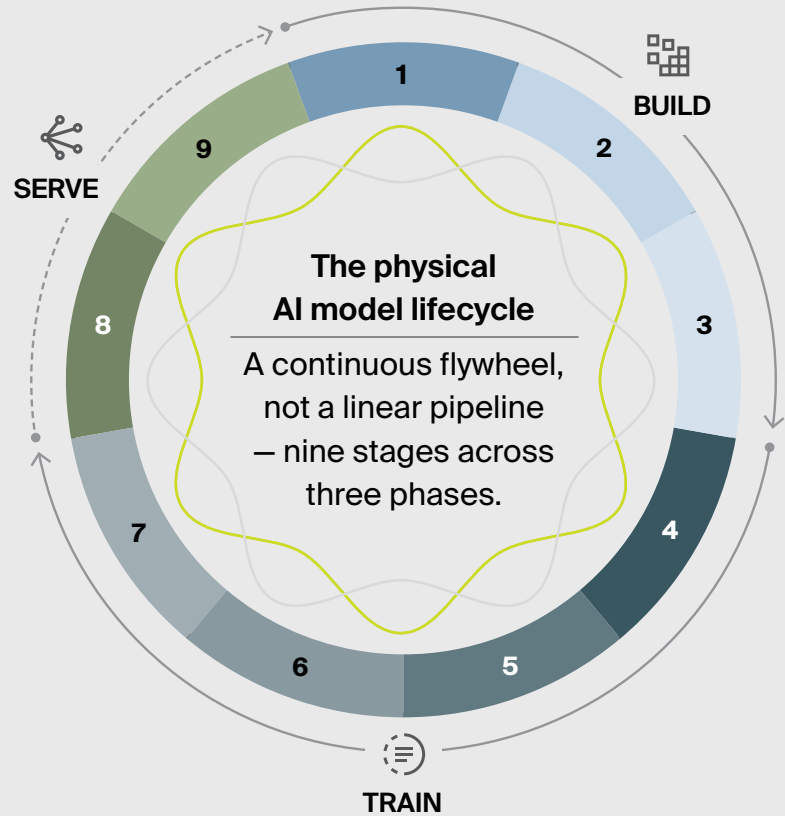
Continuously ingesting operational telemetry back into future training runs, turning rapid iteration velocity into the key market differentiator.

That fourth shift is moving fast. World model investment jumped from roughly \$1.3B in 2024 to \$6.5B in 2025 as revealed in the RAISE summit ([Source](#), Feb 2026), and robotic foundation model startups alone pulled in \$3.92B across 9 deals by mid-2026 ([Source](#), July 2026), now over 77% of all disclosed physical AI capital. Infrastructure sized only for training and inference will hit a wall the moment a customer's data strategy shifts toward synthetic generation at scale, because that workload didn't exist in most capacity plans 18 months ago. The platforms built to absorb it now are the ones physical AI teams will need next.



Physical AI lifecycle

Many physical AI programs run a version of the same nine-stage loop: data collection, ingestion and processing, synthetic data generation, pretraining, fine-tuning, distillation, simulation and evaluation, deployment, and monitoring and feedback feeding back into the stages. What differs by customer is which stages dominate the spend, and what “done” looks like:



BUILD

1. Data collection:

Capturing raw sensor streams and robot data like video, LiDAR, joint states, force/torque, teleoperation traces.

2. Ingestion & processing:

Reformatting, captioning, embedding, and curating raw capture into usable datasets.

3. Synthetic data generation:

World models generate video and trajectory variations, unique synthetic scenarios and edge cases that real-world collection can't reach cheaply.



TRAIN

4. Pretraining:

Base model (LLM, VLA, or world model) learns from the full multimodal dataset.

5. Fine-tuning:

Adapting the base model to a specific embodiment, task, or customer environment.

6. Distillation:

Compressing the model to run onboard, in real time.

7. Simulation & closed-loop eval:

Testing in a digital twin or simulated environment before anything touches a real robot or vehicle.



SERVE

8. Deployment:

Shipping the model to production from cloud to fleet via OTA update or direct provisioning at the edge.

9. Monitoring & feedback:

Deployed telemetry streams back to feed the next dataset, closing the loop.

Physical AI use cases

We’ve seen seven distinct scenarios throughout industries and that fits in the ops pattern. Where each industry might have a different focus on which stage of the loop carries the weight, and how fast a customer needs to move through it.

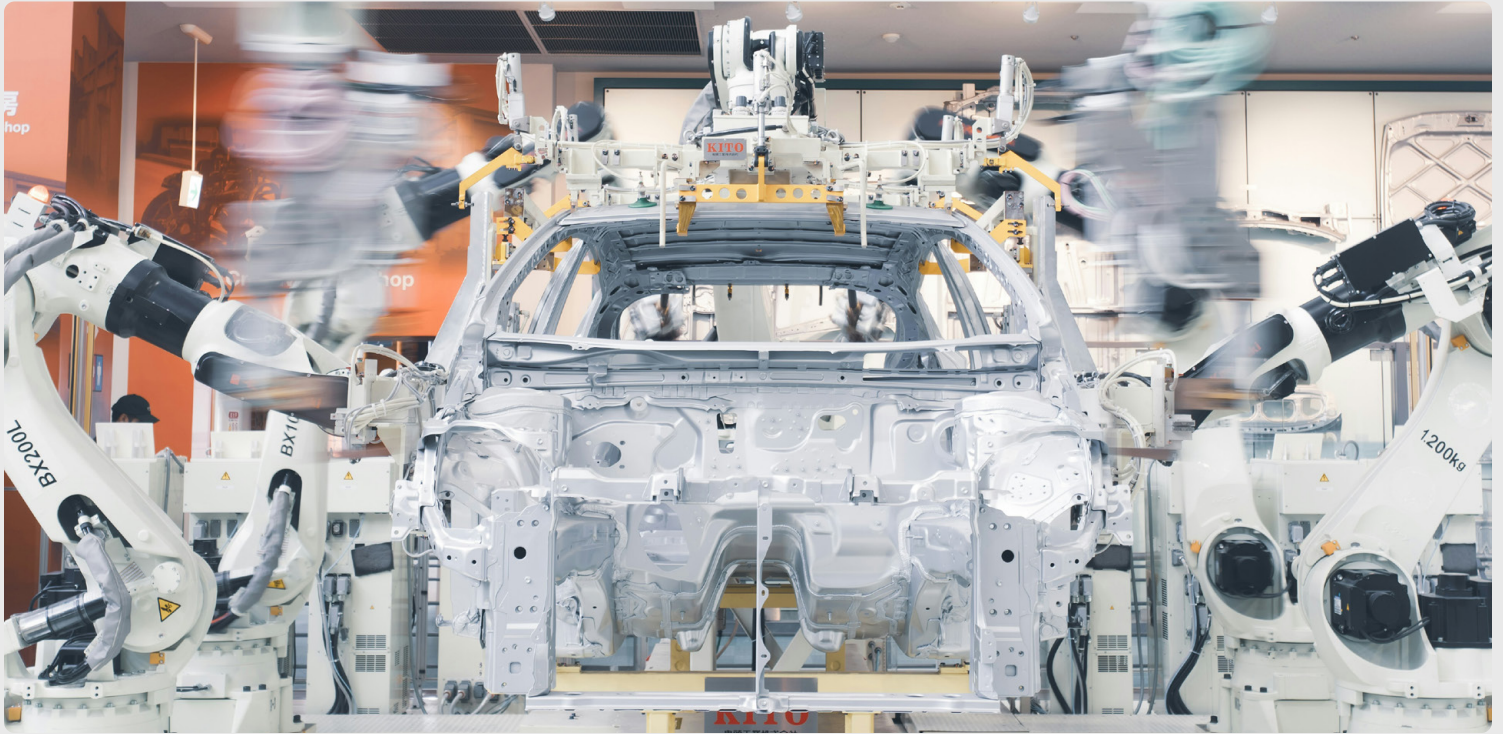
Customer scenario	What they’re trying to achieve	Where Crusoe fits
Model labs	Complete frontier pretraining runs at maximum goodput, then adapt models for licensees	Built with frontier labs: exascale GPU clusters, managed orchestration, tuned for goodput
Autonomous driving	Turn petabytes of fleet captured logs and simulated data into validated, in-vehicle policies	Built for continuous training and inference: zero-ingress and egress fee storage, simulation GPUs
Humanoid robotics	Stretch scarce teleoperation data with synthetic generation; adapt policies per embodiment	Tuned with humanoid teams: elastic pretraining and synthetic data generation
Industrial	Validate every change in a digital twin and keep factory data local	Built for data sovereignty: cloud simulation and inference + on-site Crusoe Edge Zones
Medical	Produce model versions that survive regulatory review, with a full audit trail	Built for pay-per-run fine-tuning: with traceable storage for data store
Logistics	Coordinate robot fleets in real time at a cost per pick the P&L can absorb	Built for cost effectiveness: priced per million tokens with no platform management overhead
Service robotics (IoT)	Serve continuous inference across thousands of devices at workable per-device economics	Built for fleet-scale economics: Fully bespoke Tailored Deployments engineered to adapt to your specialized pipeline requirements

Seven scenarios, one cloud platform built to fit whatever workload you’re running.



How Crusoe helps your physical AI development

Crusoe's AI Platform is the infrastructure layer beneath physical AI's build and train stages. We provide GPU and CPU compute, storage, networking, orchestration, and edge footprint, all in one platform. Crusoe is the industry's first vertically integrated AI infrastructure company, bringing energy, data centers, and cloud together. This brings predictable capacity for your future growth, flexible deployment options for any workload you run, and deep partnership with your team to bring the latest technology to life in physical AI space, with new capacity online in months, not years.



Why teams choose Crusoe for physical AI

- **Versatility:** Purpose-built AI infrastructure on NVIDIA and AMD's reference architectures, from gigawatt-scale data centers to edge deployments with Crusoe Edge Zones powered by Crusoe Spark modular data centers. Because Crusoe owns the full stack, you run training through inference on one platform, with predictable capacity expansion, competitive pricing, and stronger tokenomics from deep integration with OSS partners.
- **Composability:** Enter at any lifecycle stage, keep as much control as you want, and hand off the rest to Crusoe. Whether you're training or serving, working with proprietary or open models, we take on the hard work of building an easy-to-use platform that fully validates the frameworks you already use such as PyTorch for training, SGLang and vLLM for inference, Kubernetes and Slurm for orchestration, so when you move to Crusoe Cloud, you don't need any rework or being locked-in, while keeping the freedom to customize every layer you touch.
- **Partnership:** Physical AI infrastructure evolves fast, and Crusoe treats that as an advantage. We partner closely with clients: engineering support tailored to your pipeline, a roadmap aligned to your future requirements, and a commitment to co-developing custom solutions alongside leading physical AI innovators.

Crusoe Cloud for Physical AI		Color legend	Crusoe Services	Crusoe Service Features
Build and process datasets Setup your pipeline to process real world data or generate synthetic data for physical AI scenarios.				Crusoe Edge Zones + Crusoe Spark™ Predictable capacity expansion Low-latency GPU fleet access Sovereign deployment
CPU	High-performance data processing			
Object storage	High throughput storage for real and synthetic data			
File storage				
Direct Connect	Streamline data transfer to Crusoe with zero ingress & egress fees			
Pre-and post-training Train, fine-tune, and evaluate LLMs, Vision-Language-Action (VLA) models, or world models.				
Managed Kubernetes	Orchestrate training and simulation jobs.			
Managed Slurm				
GPU	Train models, generate synthetic data, and run simulations.			
NVIDIA InfiniBand / RoCE				
VPC	Secure fleet-to-cloud isolation			
Container registry	Rapid image pull and workload scaling			
Serverless Fine-Tuning	Fine-tune models on custom datasets			
Deploy and serve Serve inference from cloud to edge with flexible deployment options and superior performance.				
Managed inference	Serve inference in the cloud and at the edge.			
Serverless inference	Deploy via simple API integrations.			
Self-serve deployments	Optimize for latency- or throughput-sensitive apps with predictable cost.			
Tailored deployments				
Load balancers	Managing fleet traffic.			

Crusoe Cloud for data processing, training and serving physical AI workloads.

Crusoe Cloud for Physical AI: AI Platform

Compute: the right accelerator for each stage

- **Large-scale GPU clusters:** Heterogeneous GPUs from NVIDIA and AMD let teams match accelerator memory and economics to the job, from large language model (LLM) pretraining, to frontier Vision-Language-Action (VLA) model training, to world-model synthetic data generation.
- **Purpose built simulation pools:** L40S GPUs for parallel simulation and rendering campaigns that need scale-out capacity.
- **Early access to what's next:** As an Reference Platform NVIDIA Cloud Partner, Crusoe brings next-generation platforms to customers as they become available, through deep collaboration with NVIDIA across hardware, software, and models including Nemotron, Cosmos, the OSMO framework, and the broader NVIDIA software stack.

Orchestration: managed service for scheduling, training and serving

- **Crusoe Managed Slurm (CMS):** The scheduler most training teams already know, delivered as a managed service for long-horizon pretraining jobs and bursty simulation queues on the same fabric.
- **Crusoe Managed Kubernetes (CMK):** For teams standardizing on cloud-native tooling across training, data processing, evaluation, and serving.

Data: high-performance storage and networking built for multimodal scale

- **Across conversations with physical AI teams, Crusoe sees storage requirements approaching 100 TB per GPU** as fleet data and synthetic pipelines grow, CPU-to-GPU ratios in data processing running from 12:1 up to 69:1, and streaming video to the GPUs taking two to three times longer than the training step itself. In physical AI, storage is often the bottleneck.
- **Direct Connect:** Physical AI workloads generate immense data processing requirements. Streamline your data migration to Crusoe using direct transfer tools with zero ingress fees.
- **High-throughput object storage with zero-ingress and egress fee for data movement:** Physical AI re-reads its data constantly; on per-GB-egress clouds, iteration itself carries a toll. On Crusoe, replaying a petabyte of fleet logs carries no data-movement tax.
- **RDMA networking:** Distributed training is only as fast as its slowest link; Crusoe's fabric is engineered so nominal cluster size translates into usable training throughput.

Edge: compute where the fleet operates

- **Crusoe Edge Zones, built on Crusoe Spark modular data centers:** Place inference and data processing at or near customer facilities and deploy in months. Latency-critical serving and locality-sensitive raw data stay close to the robots while elastic training stays in the cloud.

Crusoe Cloud for Physical AI: Managed AI

Crusoe's Managed AI portfolio for teams that want outcomes without operating the machinery. Bring a model and data, and get back adapted models and production serving. For physical AI teams, this matters most in two places: the recurring adaptation work, and the serving economics that decide whether a fleet's unit costs pencil out.

Managed Inference, powered by MemoryAlloy™ technology

Once models serve fleets, inference becomes the dominant cost: continuous serving in the cloud across hundreds or thousands of endpoints powering the different devices and robots or self-run inference at the edge, inference usually runs 24/7, while training runs only occasionally.

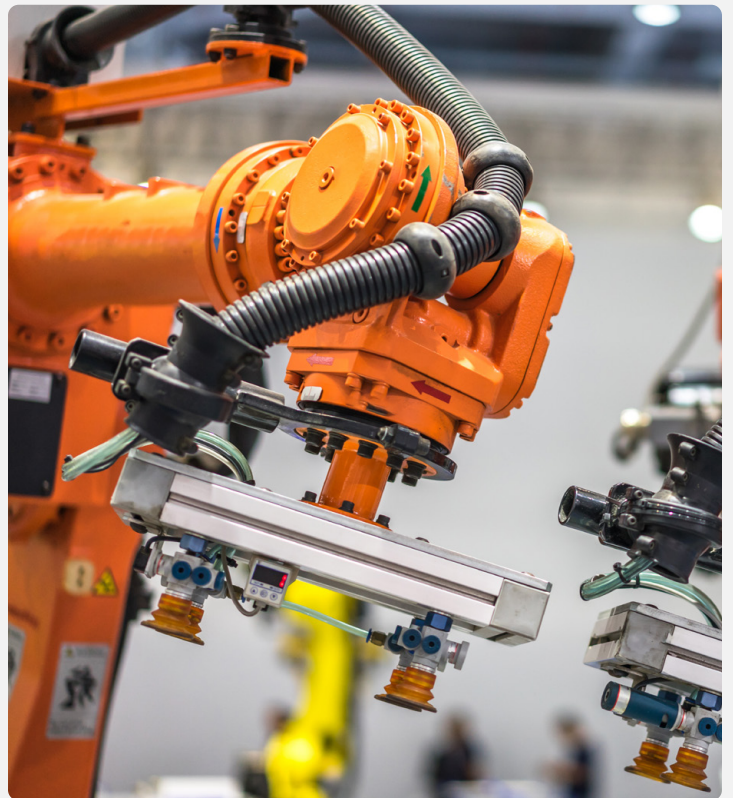
Crusoe Managed Inference runs in the cloud, powered by MemoryAlloy™ technology, and Crusoe's open-source contribution, fastokens, has been adopted into vLLM, SGLang, and NVIDIA Dynamo, the mainstream serving engines.

Crusoe Managed Inference delivers the throughput economics that fleet-scale serving demands, with deployment options to match: Serverless Inference for instant access to pre-deployed AI models, Self-Serve Deployments for teams optimizing a specific latency or throughput profile, and fully bespoke Tailored Deployments backed by Crusoe engineering support. We meet you where you are, with an onboarding experience built for complex physical AI serving workloads.



Where the portfolio is headed

Crusoe aims to bring more physical AI customers onboard, and we want your feedback to shape what we build next.



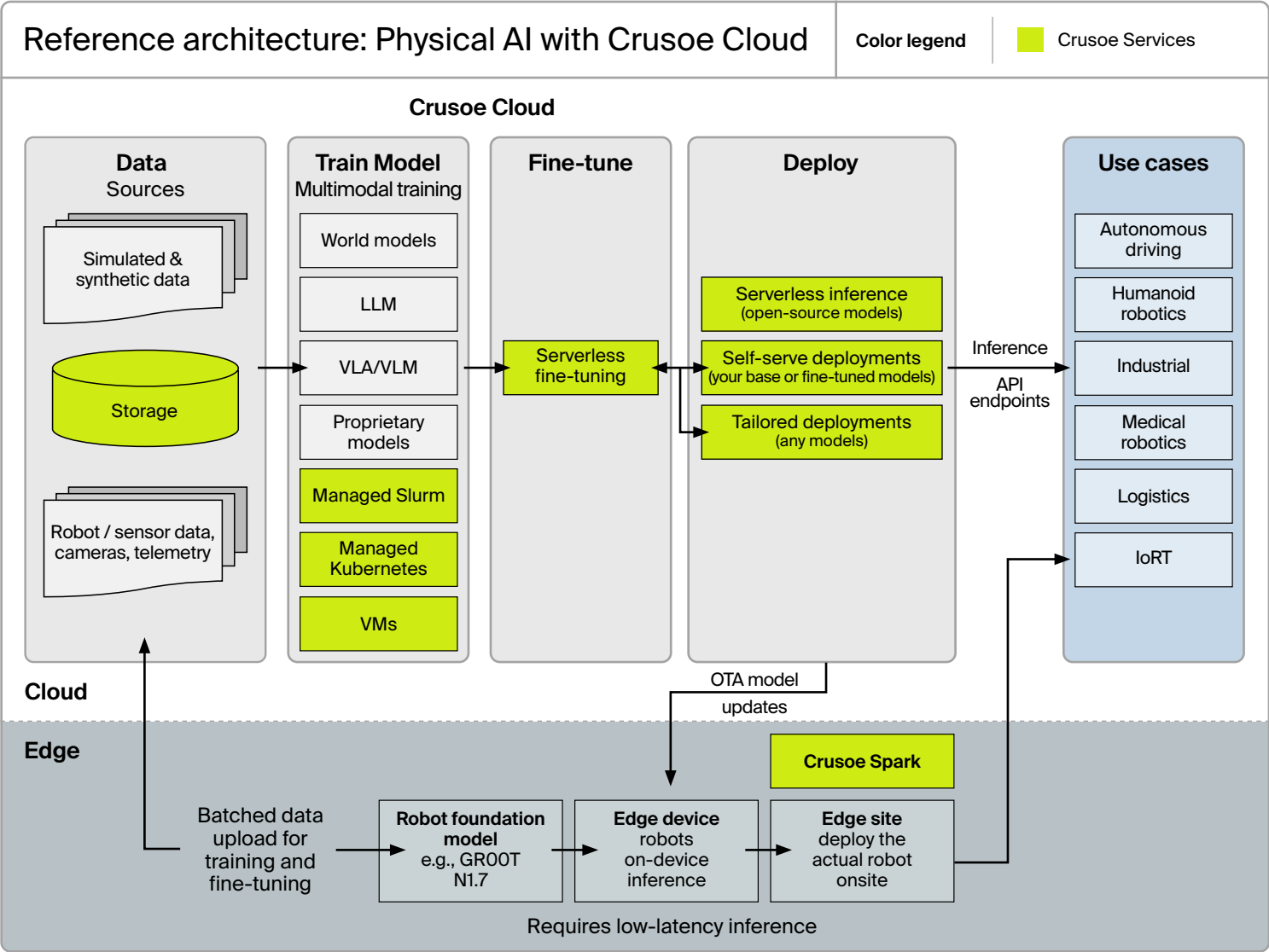
A reference architecture: Cloud training, edge inference

Most physical AI deployments are split into two environments: Cloud for training, fine-tuning, and synthetic data generation, and edge for the low-latency inference a live fleet needs. Crusoe operates both as one platform.

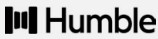
Crusoe Cloud powers model training and inference at scale. Crusoe Edge Zones power on-site low-latency serving and data movement. And the two run as a single loop: model updates flow from Crusoe Cloud to robots over the air, and fleet telemetry flows back into the next training cycle.

Crusoe

Try OSMO on Crusoe Cloud →





	BRAID ROBOTICS	
		
		
		
	Veeda AI	

Reach out to Crusoe team →

Try the OSMO solution skills →

Built on Crusoe today

Physical AI companies already building on Crusoe include 1X, Figure, SpAltial, RoboFlow, Pantheon, Veeda Innovation and other robotics teams running training, fine-tuning, and simulation workloads on Crusoe infrastructure today.

Getting started

Talk to Crusoe about your physical AI workload

Reach out to your Crusoe account team or visit crusoe.ai/solutions/physical-ai for engineering support on your specific pipeline.

- Physical AI infrastructure decisions compound: your workload shapes your architecture, and that leads to what’s possible at training, fine-tuning, and inference workflow your fine-tuning approach shapes what’s deployable at the edge. A first conversation with Crusoe’s physical AI solutions engineering team typically covers:
 - Which scenario matches your roadmap, and which stages carry your heaviest spend
 - Where your pipeline bottlenecks first: storage throughput, simulation capacity, or edge latency
 - Your unit economics, cost per validated mile, per policy iteration, per pick, per device-month, and what your infrastructure spend looks like against them
 - Which stages you run yourself on the AI Platform, and which you hand to Managed AI

Appendix: How to evaluate infrastructure for physical AI

Standard public clouds frequently struggle under the specialized demands of robotics, such as constrained storage throughput, network oversubscription, latency bottlenecks that degrade real-time performance, or unreliable OTA streaming that impedes dataset updates. Regardless of whether you choose Crusoe, evaluating the following core criteria helps distinguish infrastructure that accelerates physical AI development from infrastructure that hinders progress:

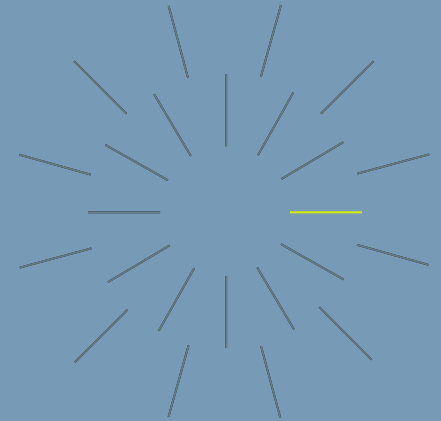
- **GPU availability & scalability:** Guaranteed access to modern accelerators alongside elastic burst capacity to prevent scheduling bottlenecks during critical launch windows.
- **Interconnect & network architecture:** High-performance fabric designed to preserve training efficiency across large cluster sizes rather than just raw GPU count.
- **Data throughput & ingress-egress economics:** High sustained read speeds for multimodal datasets, paired with predictable pricing structures when re-reading data across epochs, replays, and evaluations.
- **Proven cluster reliability:** Transparent metrics for time-between-failures and time-to-recovery, as rapid recovery directly determines total cost per completed training run.
- **First-class simulation integration:** Native support for parallel simulation and digital-twin environments without requiring custom integration overhead.
- **Cloud-to-edge deployment path:** Capabilities to run low-latency inference near your fleet and a clear path bridging cloud training to on-site serving.
- **Custom engineering flexibility:** A provider willing to build and adapt infrastructure to your exact workload specifications rather than forcing a rigid, standardized menu.
- **Ecosystem & stack compatibility:** Seamless execution of key frameworks—including PyTorch, Isaac, LeRobot, SGLang, and vLLM—allowing your software stack to transfer without costly refactoring.

Crusoe's evidence, independently measured

- Crusoe-sponsored benchmark report - “The True Engineering Cost of Open Model Training: A TCO and ROI Analysis of GPU Infrastructure” by Surf Report, shows Crusoe delivered the lowest cost per unit of compute at \$0.01010 per TFLOP-hour and the lowest training-token cost at \$2.86 per million tokens, with one-year TCO 10% below the competing neocloud and 35% below AWS. The goodput advantage alone compounds into roughly 12 extra days of productive training per year.
- Artificial Analysis: As of June 2026, Crusoe's inference engine ranked first for output speed on its Kimi K2.6 and K2.7 deployments, reaching 430+ output tokens per second and topping the provider leaderboard for response latency.
- InferenceMAX: Crusoe is a compute partner behind SemiAnalysis's open-source InferenceMAX benchmark, supporting transparent, continuously updated inference benchmarking across NVIDIA and AMD hardware.
- Gartner: Crusoe was named a Visionary in the 2026 Gartner Magic Quadrant for Cloud AI Infrastructure.
- NVIDIA: Crusoe Cloud earned an NVIDIA Exemplar Cloud achievement for large-scale AI training on NVIDIA HGX B200 systems
- Crusoe Spark: An edge deployment solution extending Crusoe's cloud infrastructure capabilities with Crusoe Edge Zones, engineered to support advanced chipsets and rapidly deploy high-performance NVIDIA hardware to fulfill stringent data sovereignty and low-latency requirements.

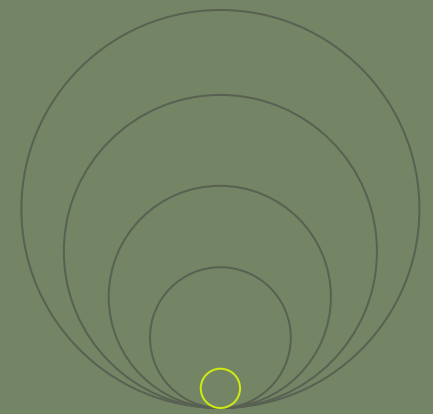
\$0.01010

Crusoe delivered the lowest cost per unit of compute per TFLOP-hour



\$2.86

Crusoe had the lowest training-token cost per million tokens



Sources and notes

1. Waymo, 500,000 weekly paid rides across 10 U.S. cities: TechCrunch, “Waymo’s skyrocketing ridership in one chart,” March 27, 2026, techcrunch.com/2026/03/27/waymo-skyrocketing-ridership-in-one-chart. Primary source: Waymo announcement, March 2026.
2. Waymo, nearly 200 million fully autonomous miles: Waymo, sixth-generation Waymo Driver announcement, February 12, 2026, as reported by Electrek, “Waymo begins fully autonomous ops with 6th-gen Driver, targets 1M weekly rides,” electrek.co/2026/02/12.
3. Waymo, one million weekly rides target by end of 2026: Takeda Mawakana, Waymo co-CEO, interview with Bloomberg Tech, February 10, 2026, as reported by Automotive World, “Waymo’s metric for 2026 success: one million weekly rides,” February 12, 2026.
4. Aurora, nearly 440,000 driverless miles through June 2026: Aurora Innovation, second-generation driverless truck announcement, as reported by electrive, “Aurora deploys second-generation driverless trucks,” July 24, 2026, electrive.com/2026/07/24/aurora-deploys-second-generation-driverless-trucks.
5. Aurora, more than 200 driverless trucks targeted by end of 2026: Aurora Innovation, Q4 2025 earnings commentary and 2026 commercial plan, March 2026, ir.aurora.tech.
6. \$55.8 billion raised by robotics companies in 2026 year to date: Dealroom data, as reported by CNBC, “Humanoid robotics company raises up to \$1.4 billion from Nvidia, Amazon and others,” June 10, 2026, cnbc.com/2026/06/10/neura-robotics-funding-ai-humanoid-robots.html.
7. Skild AI, approximately \$1.4 billion Series C: Skild AI, “Announcing Series C,” January 14, 2026, skild.ai/blogs/series-c.
8. NEURA Robotics, Series C of up to \$1.4 billion: CNBC, June 10, 2026 (see source 6).
9. Figure, more than \$1 billion raised at a \$39 billion post-money valuation: Figure Series C announcement, September 2025.
10. Cisco, 61% of industrial organizations deploying physical AI, 20% scaled successfully: Cisco physical AI research, April 2026. [State of Industrial AI Report.]
11. Artificial Analysis output speed and latency rankings: Artificial Analysis provider benchmarks, artificialanalysis.ai, as published in Crusoe announcements, July 2026.
12. InferenceMAX compute partnership: SemiAnalysis, “InferenceMAX: Open Source Inference Benchmarking,” October 2025, inferencemax.ai.
13. Gartner: Gartner, “Magic Quadrant for Cloud AI Infrastructure,” Analysts: Dennis Smith, Alessandro Galimberti, Ed Anderson, Carolin Zhou, Wataru Katsurashima, Ashish Banerjee. Published: 06 July 2026. External use of Magic Quadrant content is governed by Gartner’s Copyright and Quote Policy.
14. NVIDIA Exemplar Cloud achievement for large-scale AI training on NVIDIA HGX B200: Crusoe announcement, 2026, crusoe.ai.
15. Surf Report benchmark results: The True Engineering Cost of Open Model Training: A TCO and ROI Analysis of GPU Infrastructure Measuring Real-World Performance, Operational Efficiency, and Full-Lifecycle business impact in Multi-Node LLM Training Comparison of Crusoe Cloud, a Gold Tier Neocloud, and a Hyperscaler” by Surf Report : William McKnight and Jake Dolezal, September 2026.”
16. Physical AI workload characteristics (storage per GPU, CPU-to-GPU ratios, video streaming versus training step time): Crusoe customer and prospect engagements, 2026.
17. NVIDIA’s Jensen Huang put the shift plainly at CES 2026: “Cosmos turns compute into data.” (Jan 2026).
18. Agility Robotics has already adopted it to scale synthetic training data beyond what it could feasibly collect in the real world (Mar 2025).