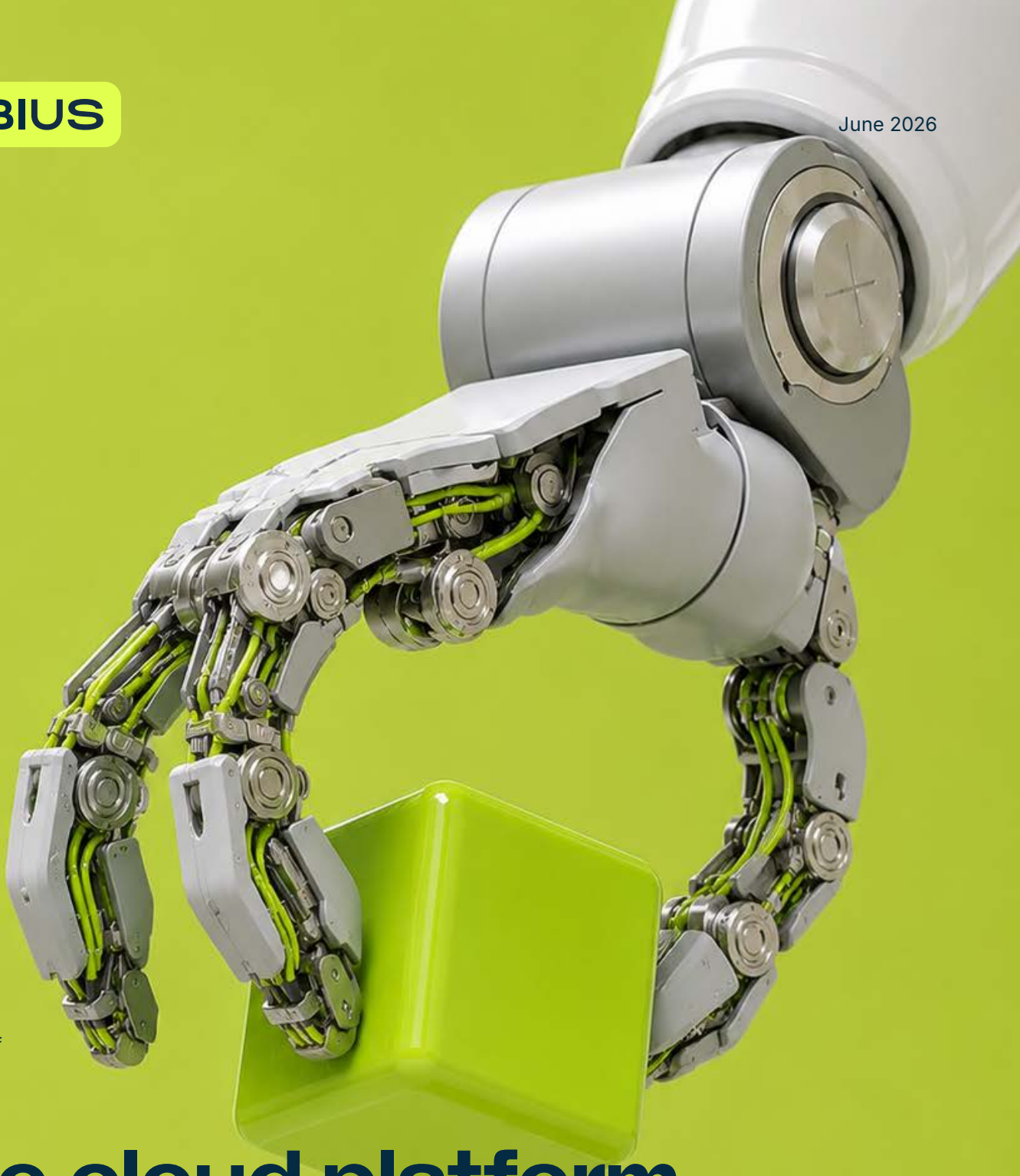
NEBIUS

June 2026



Solution brief

The cloud platform for physical AI: How Nebius accelerates robot learning and deployment

Introduction

Physical AI refers to autonomous systems that perceive, reason and act in the real world. This includes robotics, autonomous vehicles, drones, cameras and other forms of embodied intelligence. Physical AI must learn from and operate within messy and dynamic high-stakes environments where data is multimodal, evaluation depends on simulation and real-world feedback, and every improvement must eventually translate into physical action.

This paper analyzes the infrastructure requirements for physical AI, the drivers of cloud adoption in robotics and why AI-native architectures are better aligned with the physical AI development cycle than general-purpose cloud alternatives. Building and deploying these systems requires large-scale AI infrastructure, optimized simulation environments, high-throughput data pipelines, flexible orchestration services and physical AI building blocks that work together as part of one continuous development loop, not as isolated steps in a sequence.

What makes physical AI unique?

In physical AI, a robot's actions change the environment actions change the environment, and that change feeds back into the next decision. This continuous cycle makes physical AI technically distinct and its infrastructure requirements materially different from purely digital workloads, such as generative AI chatbots.

Representative systems and drivers

- **Diverse form factors:** Humanoids, warehouse robots, agricultural drones, autonomous vehicles, industrial inspection platforms and other AI-enabled machines.
- **Technological convergence:** Enabled by matured deep learning, multimodal foundation models, world models, high-quality sensors and accessible AI infrastructure including GPUs.
- **Operational stakes:** Models must manage latency, sensor noise and collision risk; poor control decisions create immediate hardware damage or safety risks.

The physical AI learning loop

The infrastructure needed for physical AI is not localized to one phase, but instead accumulates across a continuous data flywheel. This learning loop transforms raw environmental interaction into refined behavioral models through seven discrete stages.

Data collection. Captures logs from real robots, teleoperation sessions and synthetic rollouts generated by world models.

Ingestion and organization. Centralizes multimodal telemetry — including sensor data and joint states — into high-performance object storage.

Indexing and curation. Mines long-tail scenarios by tagging critical events, such as collisions, failures and human interventions.

Model training. Develops perception layers, control policies through imitation learning/reinforcement learning and vision-language-action (VLA) architectures.

Evaluation. Executes regression suites across large-scale simulation farms and real-world test environments.

Deployment. Manages model distribution via over-the-air (OTA) updates, canary rollouts and telemetry streaming.

Observation and feedback. Monitors field performance to identify behavioral drift; new failure cases are then used to seed subsequent datasets.

The complexity with physical AI

The transition to physical AI introduces significant complexity across the technical stack, requiring new approaches to data management, model architecture and compute orchestration. Unlike AI for digital interfaces, these systems must solve for five distinct technical shifts:

- **New data profiles:** Infrastructure must handle high-concurrency, multimodal streams (including video, LiDAR and 3D spatial geometry) that differ structurally from text-based datasets.
- **New model architectures:** Engineering teams are increasingly adopting vision-language-action (VLA) models, world models and video diffusion, requiring tight integration between reasoning and physical actuation.
- **New workload demands:** Standard compute requirements are expanded by parallel simulation, edge inference and the need for continuous evaluation across diverse environments.
- **New system architectures:** Development requires specialized frameworks for data collection and management that can support heterogeneous compute across the entire model lifecycle.
- **New tooling requirements:** Teams must address substantial product gaps in synthetic data generation, fleet management and model observability, to bridge the gap between the robot and cloud.

At the same time, robot developers must quickly transform raw environmental observations into validated behavioral improvements. To do this, they must accelerate the physical AI learning loop that comprises data ingestion, simulation, model training, evaluation and deployment.

In physical AI, iteration speed is a competitive advantage. It's not about training bigger models; it's about accelerating the learning loop.

The infrastructure required for physical AI

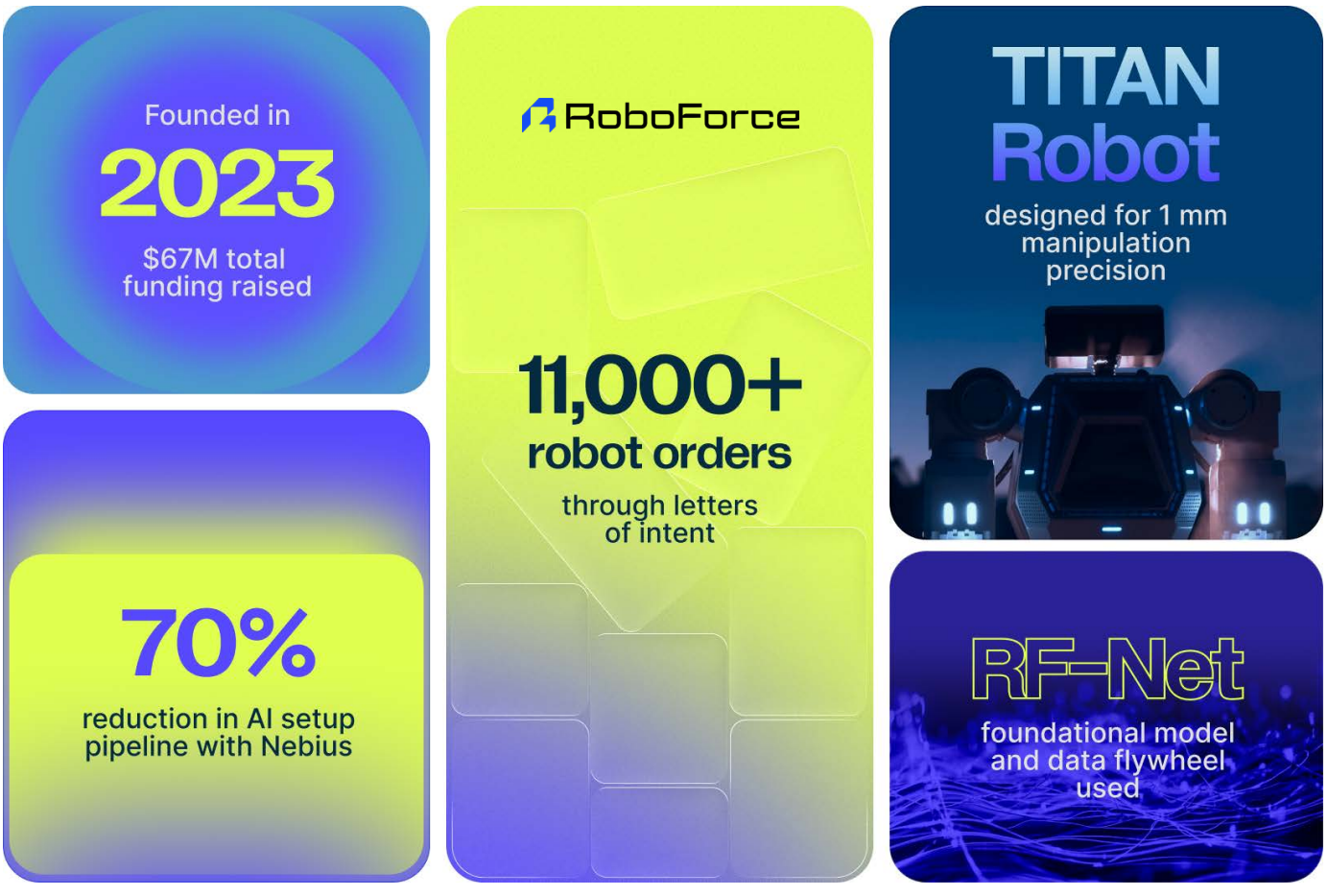
Given the importance of iteration speed, physical AI systems require large-scale accelerated compute, high-throughput storage, fast interconnects, reliable orchestration and simulation platforms that can operate in parallel with training jobs. This combination is what makes robotics one of the most infrastructure-intensive extensions of the AI market.

Core infrastructure requirements

- **Accelerated compute clusters:** Large-scale GPU environments are required for foundation model training, perception systems and high-batch policy experimentation.
- **Parallel simulation engines:** Physics simulation and synthetic data generation require CPU and GPU resources to operate in tandem, supporting photorealistic rendering and actuator modeling. Simulations can stress CPU, GPU, memory and storage simultaneously, especially when rendering, logging and training are happening together.
- **High-throughput data fabric:** Storage systems must support the rapid ingestion and versioning of multimodal datasets, including video streams, joint states and action traces.
- **Low-latency interconnects:** High-speed networking is a first-order requirement for distributed training. Weak interconnect performance creates immediate scaling inefficiency and bottlenecks when moving large datasets or checkpoints.
- **Extremely reliable infrastructure:** Large-scale deployments demand infrastructure that is optimized for parameters such as Mean Time Between Failures (MTBF) and Mean Time To Recovery (MTTR).

RoboForce is scaling physical AI data with NVIDIA Cosmos and OSMO on Nebius

Explore the RoboForce story

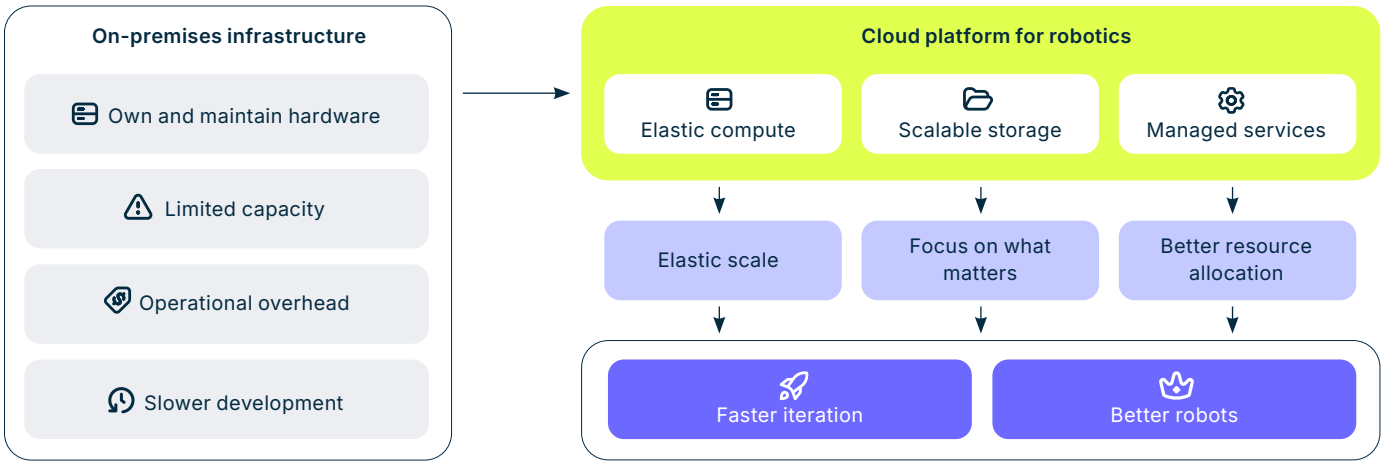


The emergence of cloud platforms for robotics

Building and maintaining a GPU cluster can create cost and operational overhead. It requires capital expenditure before a single model is trained, and demands engineering attention that compounds as the cluster grows. For robotics teams, especially startups and growth-stage companies, the constraint on progress is rarely raw hardware ownership but the speed of the development loop. Instead of investing in talent and systems for building infrastructure, it's better to do the same for building robot capabilities.

Cloud infrastructure changes the operating model for robotics. Here's how:

- **Elastic capacity for burst demand:** Simulation and training workloads arrive in sharp bursts; cloud infrastructure allows compute to scale with demand, rather than with prior procurement cycles.
- **Operational focus:** Robotics teams can focus on operations to improve robot behavior instead of cluster reliability and performance.
- **Resource allocation:** Moving to the cloud shifts capital allocation from GPUs and data centers to robot platforms, world models and physical AI innovations.



RoboForce, a Silicon Valley robotics company, leveraged Nebius to reduce AI pipeline setup time by 70% and cut iteration cycles from months to days.

Evaluating cloud platforms for physical AI

Many cloud platforms designed for general enterprise workloads fail under robotics-specific failure modes, such as weak storage throughput or oversubscribed networking. Consider these technical criteria for evaluating your cloud platform:

GPU availability and scalability

Evaluation must focus on access to current-generation accelerators and the ability to scale into large distributed clusters without scheduling delays.

Interconnect performance

High-speed networking is essential. Weak interconnects turn nominal cluster size into unusable cluster efficiency for distributed training.

Data throughput

Platforms must support high-speed ingestion and retrieval of large multimodal datasets (including image streams and sensor logs) across active experiments.

Simulation integration

Infrastructure should support large-scale parallel simulation and digital twin workflows, without forcing teams into custom integration layers.

Workload economics

Efficiency is dictated by utilization and failure recovery rather than hourly pricing alone.

Ecosystem fit

Easy access to machine learning frameworks, simulation toolchains, data pipelines and increasingly open robotics stacks.

Why AI-native cloud platforms have an advantage

Physical AI workloads are structurally different from general cloud compute workloads. They combine large-scale GPU training, parallel simulation, high-throughput storage and distributed coordination in ways that stress infrastructure at every layer simultaneously. Cloud infrastructure originally optimized around web applications, enterprise software or broad-purpose infrastructure services is a structural mismatch for robotics.

AI-native cloud platforms start from a different architectural assumption. They are designed around training-heavy workloads, high-performance networking, accelerated storage paths and operational patterns shaped by long-running AI jobs rather than by conventional enterprise traffic. This difference matters because in physical AI the development loop is only as fast as its slowest infrastructure dependency. There are a few key advantages of an AI-native cloud platform:

- **Cluster design:** Distributed model training and simulation-scale workloads depend on efficient GPU-to-GPU communication, fast node-to-node transport, and predictable performance across jobs. A provider that treats these requirements as core architecture rather than optional add-ons is better aligned with robotics development.
- **Workflow fit:** Physical AI teams increasingly need one environment that can simultaneously support simulation, data curation, training and deployment preparation. When those stages sit across disconnected tools and inconsistent infrastructure layers, engineering time shifts away from robot performance and toward integration work. That overhead is not theoretical.

Engineering teams routinely spend 30% to 40% of their time on integration work rather than improving robot behavior, framing integration friction as a measurable productivity drag in physical AI development.

Nebius for physical AI

Nebius is a purpose-built NVIDIA Exemplar Cloud for AI workloads. It is a Reference Platform NVIDIA Cloud Partner that combines supercomputing performance with hyperscale flexibility. As a vertically integrated provider, Nebius delivers the full stack required for physical AI, from optimized hardware to platform-level orchestration. With 3.0 GW of contracted power and major agreements with Microsoft and Meta, the platform is architected for long-term AI scaling. Physical AI customers on Nebius include 1X, World Labs, Voxel51, Milestone Systems and Rhoda.

Compute reliability that accelerates the learning loop

In physical AI, reliability is a direct function of iteration speed. Distributed training runs are long-horizon jobs where a single node failure can stall the entire development pipeline. Nebius minimizes this risk through a self-healing infrastructure — a single failure doesn't just cost a GPU hour, it can stall the entire pipeline.

An anonymous customer who ran multiple LLM training jobs on a 3,000-GPU (375-node) cluster. This system achieved a peak Mean Time Between Failure (MTBF) of 56.6 hours (169,800 GPU hours), with an average of 33.0 hours over the past several weeks. When it came to the cluster's ability to restore its state, Nebius achieved an average Mean Time to Recovery (MTTR) of 12 minutes across most of our installations. This impressive result is possible through end-to-end automation of the recovery process: from early-stage fault diagnosis to spinning up replacement nodes without human intervention. Further details are available in the The Economics of AI Clusters paper.

For physical AI teams running world models, policy learning and synthetic data pipelines in parallel, those time savings compound rapidly to enable much lower cost per learning cycle. The reliability and cost-effectiveness of Nebius is substantiated by SemiAnalysis, an independent research and analysis company focused on accelerated computing.

Total cost of ownership index — Nebius = 1.0x baseline (lower is better)

Scenario	Nebius	Hyperscaler	Silver-tier cloud
Large LLM pretrain 5,184 GB300 NVL72 GPUs	1.0x (baseline)	1.09x (+9%)	1.08x (+8%)
Multimodal RL research 2,048 B200 GPUs	1.0x (baseline)	1.43x (+43%)	1.08x (+8%)
Inference endpoints 512 H200 GPUs	1.0x (baseline)	2.13x (+113%)	1.04x (+4%)

Source: Calculating the Total Cost of a GPU Cluster, SemiAnalysis

Storage for multimodal data flywheels

Physical AI generates multimodal "torrents" — including 4K video, LiDAR point clouds and high-frequency joint states — that overwhelm standard cloud storage. Nebius employs a tiered architecture to maintain GPU saturation:

- **All-Flash parallel filesystem:** Delivers throughput exceeding 2 TB/s to prevent data-starvation of GPUs.
- **Enhanced Object Storage:** Provides up to 2 GB/s of throughput per GPU for high-concurrency data-set hydration.
- **Cost-optimized simulation storage:** Non-replicated NVMe (NRD) provides high IOPS per dollar for transient simulation workloads (e.g., Isaac Lab, MuJoCo), significantly reducing simulation farm overhead. The result is a storage layer that actively fuels the physical AI data flywheel rather than acting as a passive archive or a bottleneck.

Managed orchestration and ecosystem integration

To solve the fragmentation between simulation, training and edge environments, Nebius provides managed orchestration and other physical AI capabilities:

- **NVIDIA OSMO Managed by Nebius:** Enables single-click deployment of NVIDIA's physical AI workflow orchestrator, allowing teams to define "simulation → training → evaluation" pipelines via YAML.
- **NVIDIA physical AI Data Factory:** Accelerates the data flywheel, comprising curation, search, augmentation and evaluation, with an open reference architecture for massive data generation and evaluation deployable on Nebius global AI infrastructure.
- **Serverless GPU compute:** Allows for episodic simulation bursts without always-on infrastructure costs; early customers report cutting setup time by 70%.
- **Nebius physical AI Workbench:** Turns NVIDIA Cosmos 3, NVIDIA Isaac Sim, NVIDIA Isaac GR00T and other physical AI tools into composable building blocks that agents can wire together.

Conclusion

Physical AI is expanding AI from digital interfaces into machines that perceive, decide and act in the real world. That shift creates a new infrastructure category defined by continuous interaction between simulation, data pipelines, model training and deployment. Robotics is therefore emerging as a major cloud workload not because it resembles prior AI demand, but because it combines several of the most demanding AI workload types into a single development loop.

Cloud platforms purpose-built for AI are better aligned with this shift than general-purpose infrastructure environments. Among that group, Nebius has credible positioning where the evidence is strongest today: purpose-built AI cloud for robotics, reliable compute, high performance storage and flexible orchestration options. For robotics teams building embodied systems, the platform decision is not only about access to AI infrastructure, it's also about how much infrastructure friction stands between an experiment and an improved robot. Less friction improves iteration speed, leading faster to better robots.

NEBIUS

Nebius is the ultimate AI cloud. We combine custom hardware, proprietary software and energy-efficient data centers to deliver unmatched speed, scale and lower costs — on your terms. Whether you're building foundation models, fine-tuning or scaling inference globally, Nebius gives you the performance of a supercomputer with the flexibility of a hyperscaler.

Learn more at nebius.com and nebius.com/solutions/phy.