Futurum

QumulusAI

The Off Ramp From Per-Token Pricing:

How Enterprises Regain AI Cost Control With Reserved Bare Metal

SEPTEMBER 2026

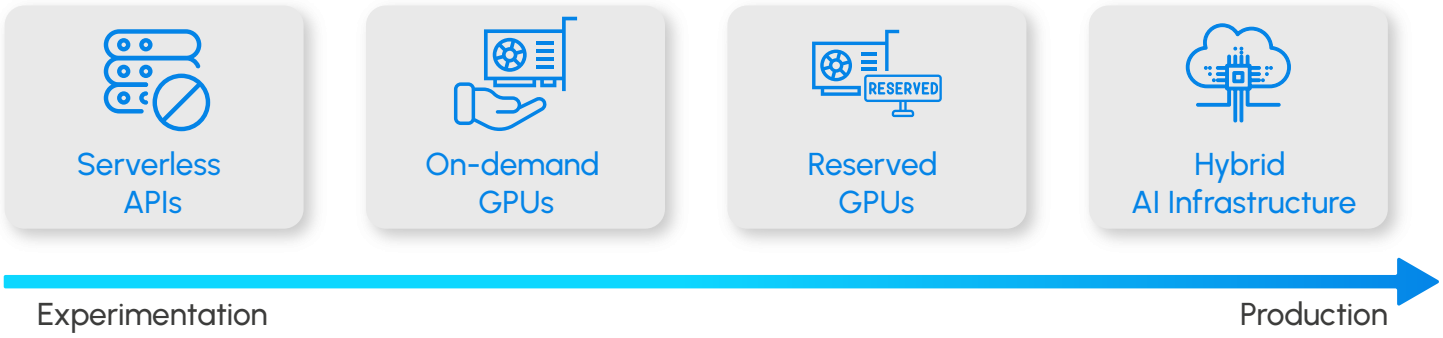
Executive Summary

Per-token serverless pricing is often the fastest path to experimentation and early production, but sustained high-volume inference creates a cost control problem that can be solved with reserved infrastructure. As agentic AI multiplies token consumption per task by up to 100x, it exposes production workloads to costs that compound exponentially rather than linearly.

Infrastructure tier selection is the primary lever organizations control over that cost exposure. AI workload deployment has already gone hybrid: 41% primarily runs in public cloud, 36% in organizations' own data centers, 13% in colocation, and 6% on bare metal or HPC providers.[3]

Consumption types show an even greater emphasis on model control via reserved instances. Reserved and owned XPUs together account for 66% of all AI compute consumption, while on-demand sits at 19%.[3] Our interviews with three compute leaders at Amberd.ai, Runpod, and Qubrid AI find that AI customers progress toward reserved instances as their AI applications mature. The end result is that customers can route application workloads to the right infrastructure to balance cost, model control, privacy, and resource flexibility (see Figure 1).

Figure 1: The Off Ramp From API Experimentation to Hybrid AI Infrastructure



Inference is the workload driving the hybrid expansion, growing from $120 billion in CY25 to $885 billion by CY30, with agent and reasoning inference alone rising 219% year over year in CY26.[2] AI-first cloud, the non-hyperscaler tier, is growing fastest at 107.1% year over year in 2026, heading from $45.2 billion in CY25 to $375.2 billion by CY30.[1]

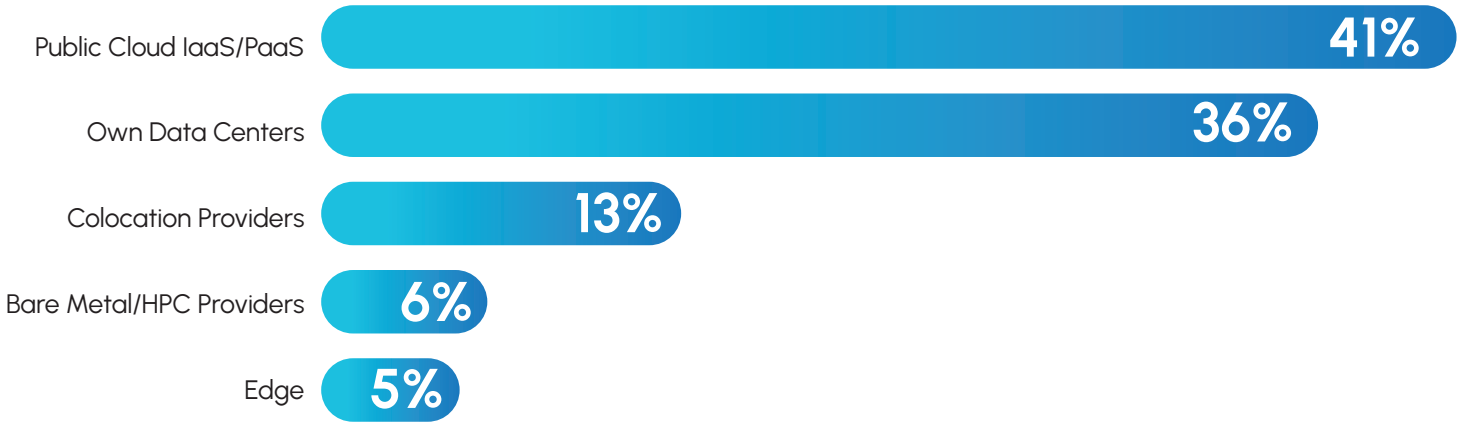
QumulusAI provides reserved bare metal capacity inside the AI-first cloud tier, where economics track hardware utilization rather than token consumption, and improvements in the inference software stack lower the customer's own cost per token over time.

Hybrid Cloud Is Already the Default for AI Inference

No Single Tier Wins the Deployment Decision

The choice between cloud and on-premises for AI hasn't produced a single answer, even with the AI Cloud Platform market exceeding $100 billion in revenue in 2025. Organizations have chosen both based on their unique needs. Futurum's enterprise decision-maker research shows inference distributed across three primary tiers (see Figure 2).

Figure 2: AI Workload Deployment by Infrastructure Tier



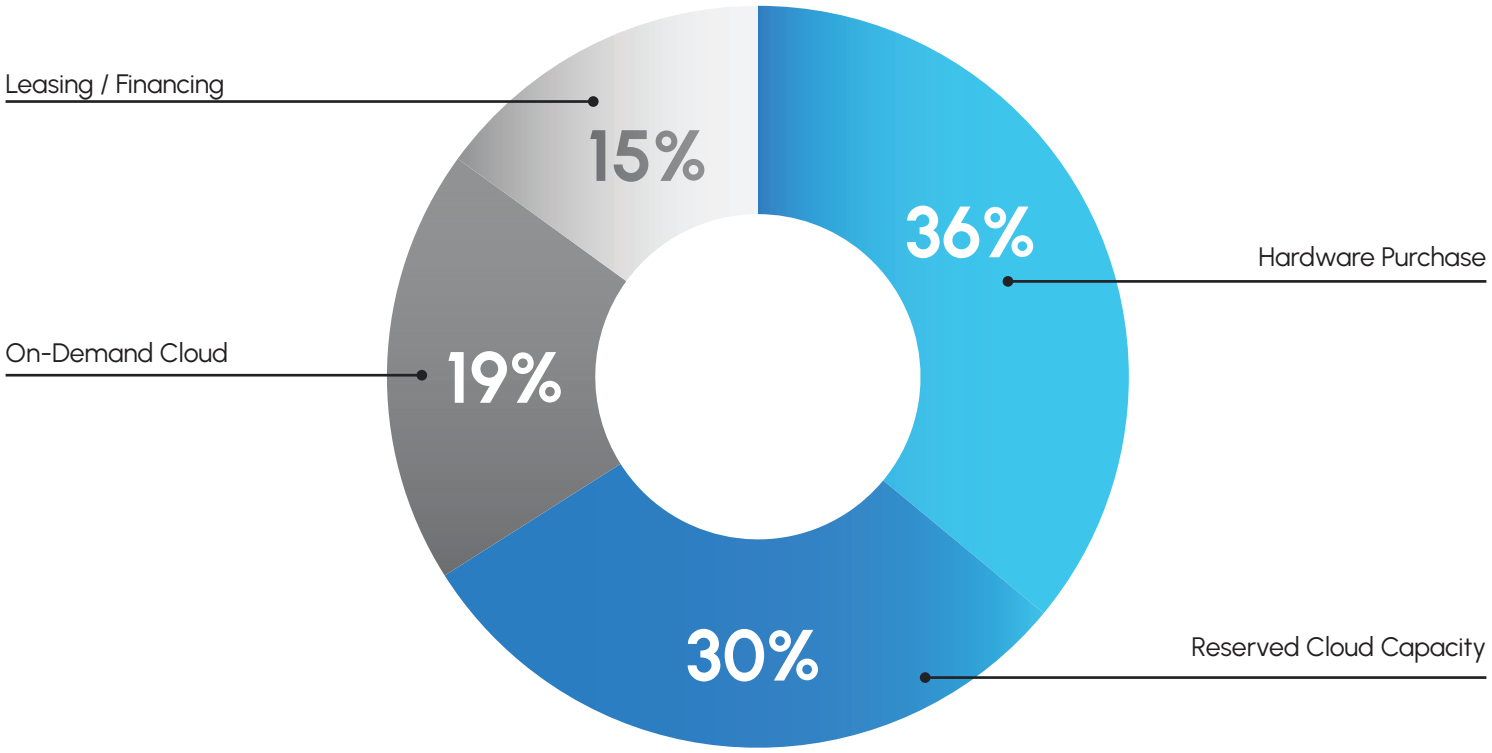
Source: Futurum Research, AI Chipsets Decision Maker Survey, 1H 2026 (n=824)

Among AI decision-makers, 59% already run primarily outside hyperscaler public cloud.3 This distribution reflects a strategic priority: organizations are actively optimizing for flexibility and scalability. By diversifying infrastructure, enterprises ensure seamless integration across on-premises and cloud environments, allowing workloads to adapt dynamically to specific latency sensitivity, data residency, cost profiles, and customization requirements.

The consumption model tells the same story (see Figure 3). Committed models, capital purchase plus reserved capacity, account for two-thirds of AI compute consumption.[3] On-demand, the model most associated with hyperscaler flexibility, represents less than 20%.

Figure 3: The Hybrid Cloud Reality for AI Inference:

66% of Compute Decision Makers Primarily Consume AI Compute via Reserved or Owned Models



Buyers have already chosen predictability. The on-demand category itself blends serverless inference APIs and on-demand GPU compute. Serverless APIs are where most organizations begin, long before utilization data justifies reserved infrastructure.

AI Cloud & Inference Platform provider Qubrid AI mirrors that progression. CEO Pranay Prakash frames the path in plain economic terms.

"We call ourselves a $5 to $5 billion infrastructure provider. When you start off, you start using APIs and try different models, so you're paying for API tokens per hour. Now you've outgrown that, which is where bare metal infrastructure comes into play."

Pranay Prakash, CEO, Qubrid AI]

Qubrid AI guides customers from token-metered APIs toward dedicated systems as scale and performance demands rise. Prakash notes that reserved capacity serves the larger deployments that run very large models and high token volumes inside the private cloud environment enterprise buyers require. The saving extends past the infrastructure line item to the surrounding tooling and the security exposure that variable, multi-tenant access carries.

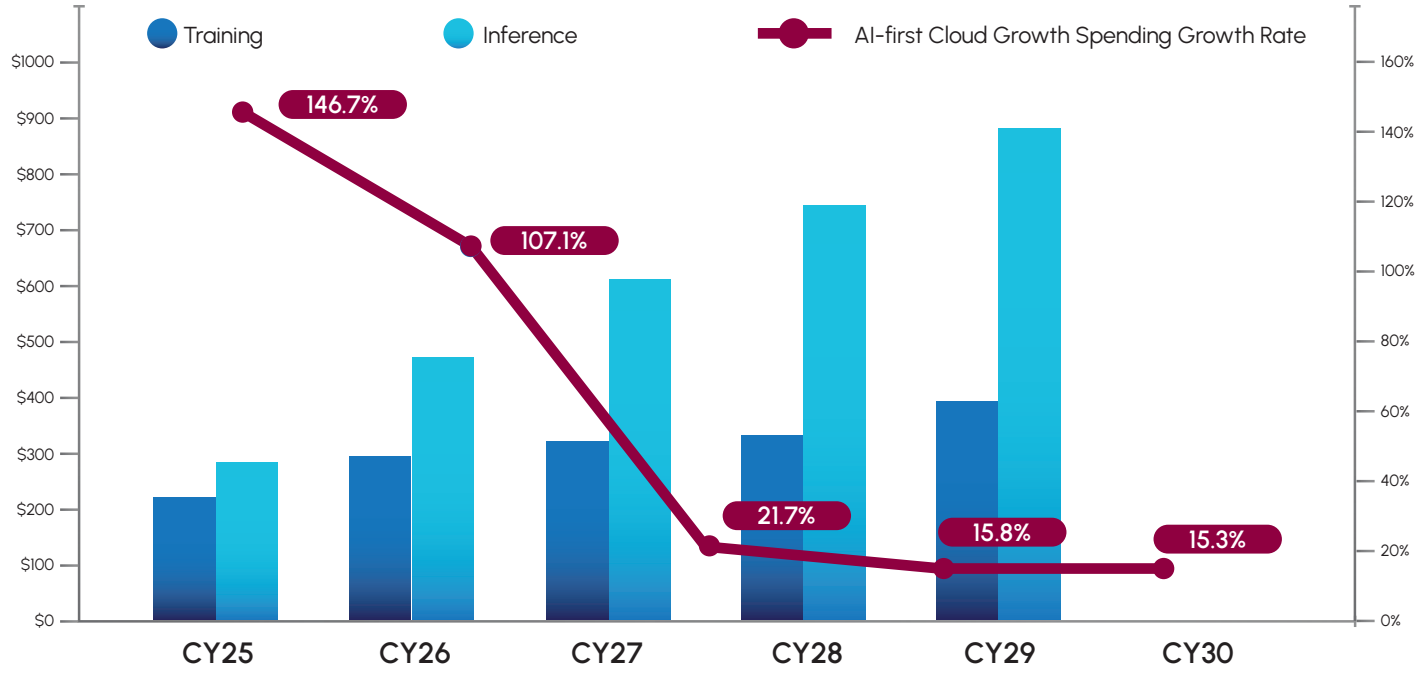
Inference Economics, Not Training Scale, Drive the Shift to Reserved Instances

Training concentrated in on-demand hyperscaler clusters because it needed scale for defined time periods. Inference behaves differently. It is distributed, sustained, cost-sensitive, and often latency-bound, and those characteristics favor hybrid deployment.

- **Sustained Utilization:** Inference runs around the clock at production scale, favoring reserved capacity.
- **Cost Sensitivity:** Per-token margins matter at volume, rewarding lower-cost infrastructure.
- **Latency SLAs:** Many workloads need proximity to users or data, pushing inference toward edge and on-premises.
- **Model and Stack Control:** For the latest GPUs, workload optimization can drive exponential gains in tokens-per-watt. Inference teams increasingly need to tune serving engines, batching, quantization, KV cache offloading, observability, and routing.

The market reflects the shift. Futurum forecasts total inference spending to grow 7.4x by CY30, with agent and reasoning inference growing 15x, rising 219% year over year in CY26.[2]. The fastest-growing workload is the one that benefits most from hybrid deployment, and the fastest-growing deployment tier, AI-first cloud, at 107.1% year over year in 2026,[1] is built for it (see Figure 4). Agentic workloads change inference tokenomics directly as the per-task token footprint grows 10x to 100x above a simple inference call, and on per-token serverless pricing that growth lands straight on the bill.

Figure 4: AI Data Center Semiconductor Market Size by Lifecycle Stage, CY25–CY30 ($B) vs. AI-First Cloud Growth Rate



Source: Futurum Research, Data Center Semiconductors Deployment Forecast, 1H 2026.

AI Maturation Runs From Serverless APIs to Reserved Infrastructure

Mature organizations end up with a hybrid architecture, but a few start there. Inference workloads follow a predictable progression as volume grows and utilization becomes easier to forecast:

- **Stage 1:** serverless inference APIs – fastest to start, priced per token, suited to experimentation.
- **Stage 2:** on-demand GPU compute – more control, still variable in cost and availability.
- **Stage 3:** reserved infrastructure – committed capacity for predictable, sustained workloads.
- **Stage 4:** hybrid multi-tier architecture – distributes workloads across deployment environments and consumption models by their characteristics.

Agentic AI is compressing this curve. Organizations that run prompt-and-response workloads comfortably on serverless APIs find that the same models, deployed as agents, multiply per-task token consumption by an order of magnitude, bringing the tokenomics inflection point forward.

The Inference Cloud Opportunity Sits Outside the Hyperscalers

AI-First Cloud Is the Fastest-Growing Infrastructure Tier

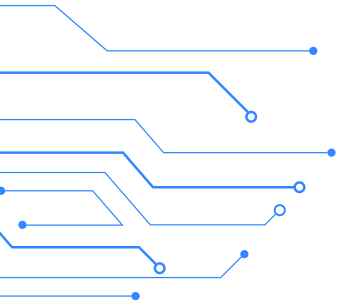
AI-first cloud, the category spanning bare metal GPU providers, specialized AI clouds, and GPU marketplaces, is growing faster than any other deployment tier. By CY30, Futurum forecasts it will reach 62% the size of the hyperscaler tier, moving from alternative to co-equal layer (see Figure 5).[1] Hybrid inference demands it, because organizations need:

- Cost optimization below hyperscaler pricing floors
- Dedicated hardware for sustained inference, without multi-tenant variability
- Procurement flexibility through GPU marketplaces, without long-term lock-in
- Custom configuration for model-serving stacks tuned to specific architectures
- Compounding gains from software optimization and observability.

QumulusAI sits in this growth lane, providing bare metal reserved instances for the high-utilization, cost-optimized allocation within a hybrid strategy. Runpod, the AI developer cloud serving more than 1 million developers and hosts some customers on reserved QumulusAI capacity, sees the same migration patterns appearing from AI developers.

Brennen Smith, the company's chief technology officer, describes three recurring customer profiles:

- Teams that outgrew homegrown tooling and no longer want to run infrastructure themselves
- Teams graduating off hyperscalers after hitting GPU-availability ceilings and reliability issues
- Teams that find generic hyperscalers were never built for AI and want a managed alternative to self-managed Kubernetes.

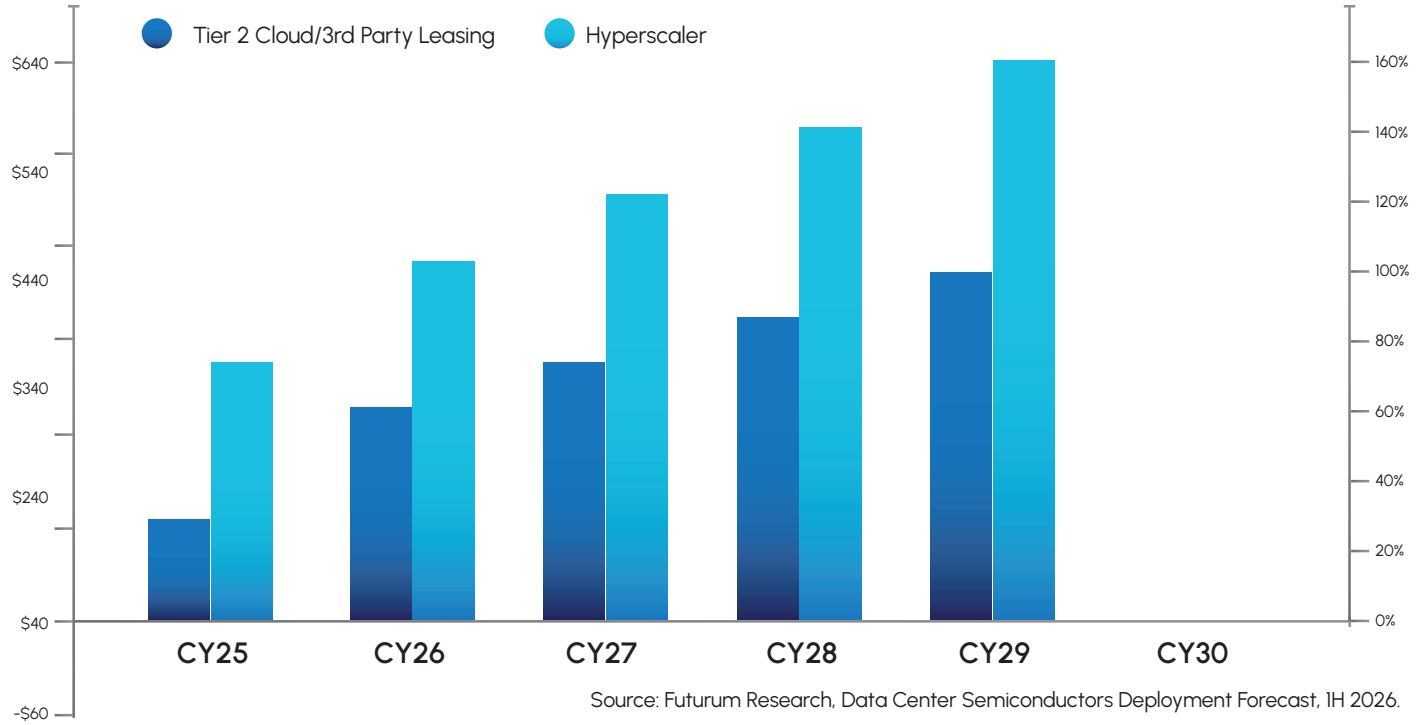


"We want a consistent experience for our customers. And that is where having that capacity on hand, dedicated, available to us for whatever need we may have, that is incredibly important. So by doing that capacity reservation deal, by being a customer of QumulusAI, we were able to guarantee we will always have that available."

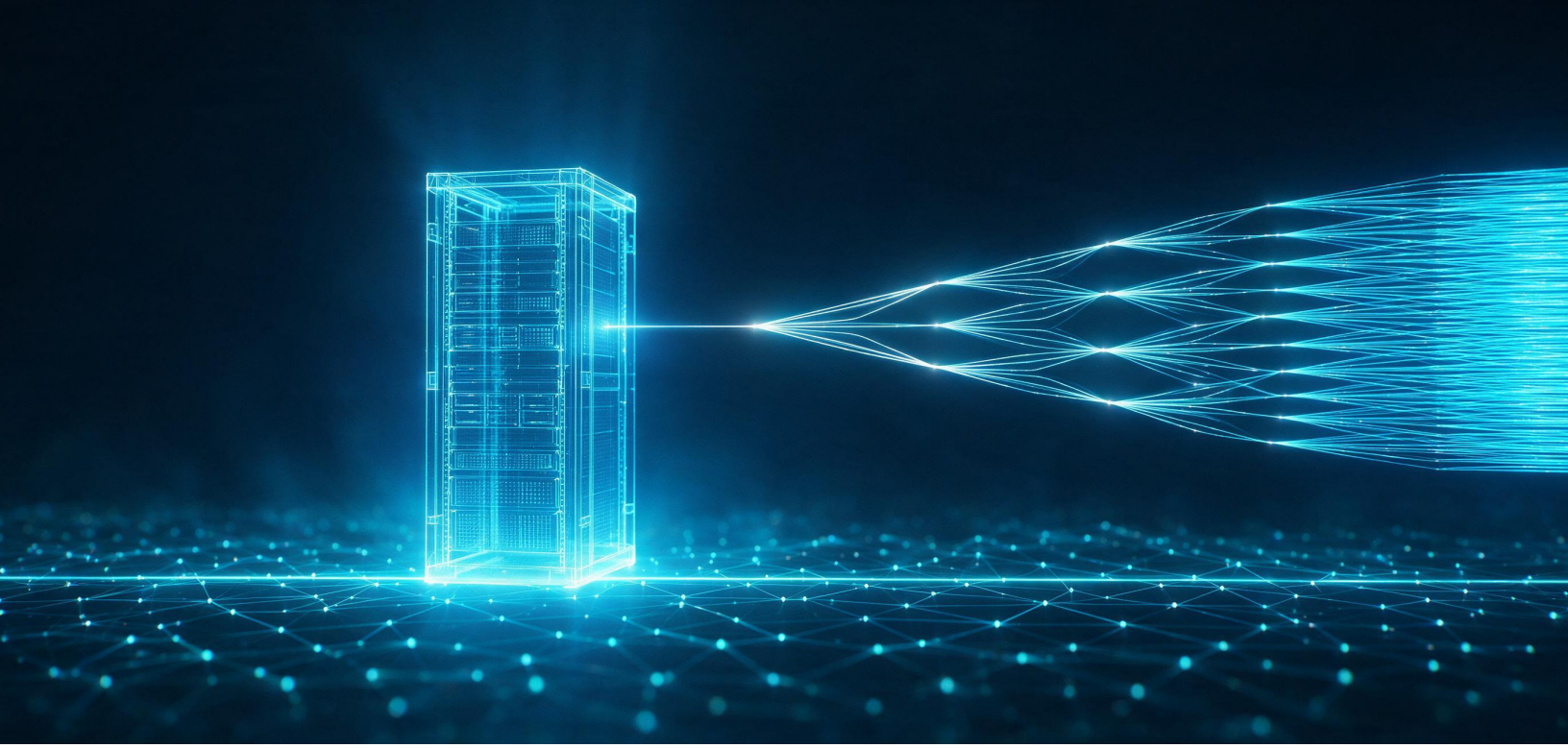
– Brennen Smith, CTO, Runpod

Reserved capacity is where inference unit economics improve most. The workloads that migrate first are sustained inference, fine-tuning pipelines, and batch processing, where dedicated hardware delivers a materially different cost structure than hyperscaler on-demand equivalents. These operations require more custom engineering, limiting the operating margin gains for teams without the hardware expertise to make use of the resources.

Figure 5: Tier 2 Cloud Data Center vs. Hyperscaler Semiconductor Spending Trajectory (CY25–CY30)



Source: Futurum Research, Data Center Semiconductors Deployment Forecast, 1H 2026.



QumulusAI in Production: How Amberd.ai Builds Fixed-Cost AI on Reserved Bare Metal

Amberd.ai Architected Around Reserved Capacity From the Start

The market data points one way: hybrid is the default, AI-first cloud is growing fastest, and reserved capacity dominates consumption. AI application deployment specialist Amberd.ai shows what that looks like in a live production environment. The software provider built its product on private, open-source large language models running on QumulusAI bare metal, and sold enterprises a fixed monthly price in place of token metering.

Two constraints drove the design: data security and governance, and predictable pricing. Mazda Marvasti, the company's chief executive, describes what pushes customers toward that model.

> *"When they start deploying it throughout the organization, the cost starts skyrocketing because it's a useful tool that somebody built, but it's now priced on a variable basis. It starts getting the attention of the CFO and the CIO in terms of how much am I exactly spending to run this tool, and is it worth it."*
>
> – Mazda Marvasti, CEO, Amberd.ai

Amberd.ai has seen customers stop using internally built automation tools outright because the variable bill could not be predicted or justified, and others search for alternative ways to deliver the same function at a fixed cost.

Bare Metal Virtualization Turns Utilization Into Margin: Two Customers Cover the Server Cost of Ownership

Amberd.ai built a custom GPU virtualization layer on top of QumulusAI bare metal. At the lowest level, an eight-GPU H200 server is partitioned into four virtual environments of two H200s each, and customers are tiered across those layers by the latency each can accept. That structure lets Amberd.ai charge different rates by tier while covering infrastructure costs and monetizing its own software on top.

"We take an 8-GPU H200, break it up into four virtual environments with two H200s each. It's one physical server. We deploy varying numbers of customers on a per-layer basis, and that's how we optimize GPU not just on a utilization basis but also on the tiering structure that allows us to charge differently."

– Mazda Marvasti, CEO, Amberd.ai

On a single eight-GPU H200 server, the first two customers cover the full cost of the server, and every customer added beyond the second is margin. Amberd.ai runs models as large as 200 billion parameters on this structure and fits roughly 30-35 customers per server.

"With one 8x H200 server, two customers pay for the entire server, and I can have probably about 30 to 35 customers running on that one server. So after the second customer, the server is free to me, and any customer that comes on after that, it's profit."

– Mazda Marvasti, CEO, Amberd.ai

The economics hold because reserved bare metal prices on hardware utilization, not per-token consumption. The same server cost supports two customers or 35 customers, so each additional unit of utilization converts directly to margin.

QumulusAI designates its new capacity for NVIDIA Blackwell systems, specifically the B200 SXM and B300 SXM6 models, shifting the per-server math Amberd.ai built on H200. The binding constraint in each two-GPU environment is HBM capacity, which determines how large a model fits and how much room is left for the KV cache that caps concurrent customers. A two-GPU H200 slice holds 282 GB, which leaves a 200-billion-parameter model served at FP8 with limited concurrency headroom. A two-GPU B300 slice holds 576 GB and turns that same model from a near-single-tenant load into a shared one, while native FP4 halves the weight footprint again and frees even more room for concurrency.

That capacity roughly doubles customer density per server, moving the ceiling from 30 to 35 customers toward 60 or more. Blackwell server cost runs higher than an H200 box, so the break-even point rises from two customers toward three or four, but total capacity scales faster than the break-even count, and the absolute margin pool per server grows rather than shrinks. Blackwell also offers model sizes and latency tiers that an H200 slice cannot physically host, allowing customers to charge up at the premium end.

The Compounding Case for Reserved Inference

Reserved Bare Metal Gains Value as the Software Stack Matures

Within hybrid cloud, reserved bare metal AI-first cloud carries the strongest growth trajectory and the best unit economics for sustained workloads. Organizations that commit capacity now do so in a market where inference pricing has risen due to more sophisticated models and scarce resources. Beyond hedging against future price increases, reserved bare metal improves economically as the inference software stack beneath it matures. In managed serverless environments such as Amazon Bedrock, OpenAI API, and Google Vertex AI, those optimization gains accrue to the platform provider, not the customer.

"By entering into a committed agreement, companies are able to hedge against where the futures are going and lock in their cost basis. There's nothing worse than having to go to the CFO and say, all those models we did, well, turns out we're having to throw them out the window. The assumptions were fine except the on-demand pricing was off."

– Brennen Smith, CTO, Runpod

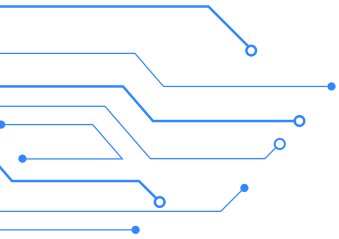
A Decision Framework for Allocating the Hybrid Estate

Futurum recommends placing a workload on reserved bare metal when it requires:

- Sustained inference with predictable utilization above roughly 60% baseline
- Peak customer token demand expands from thousands of tokens to millions
- Custom model-serving configuration or hardware tuning
- Data residency or privacy constraints that limit hyperscaler options.

Keep a workload on hyperscaler infrastructure when it is:

- Bursty, with unpredictable utilization
- Tightly integrated with hyperscaler-native AI services
- Small enough that management overhead exceeds the cost saving.



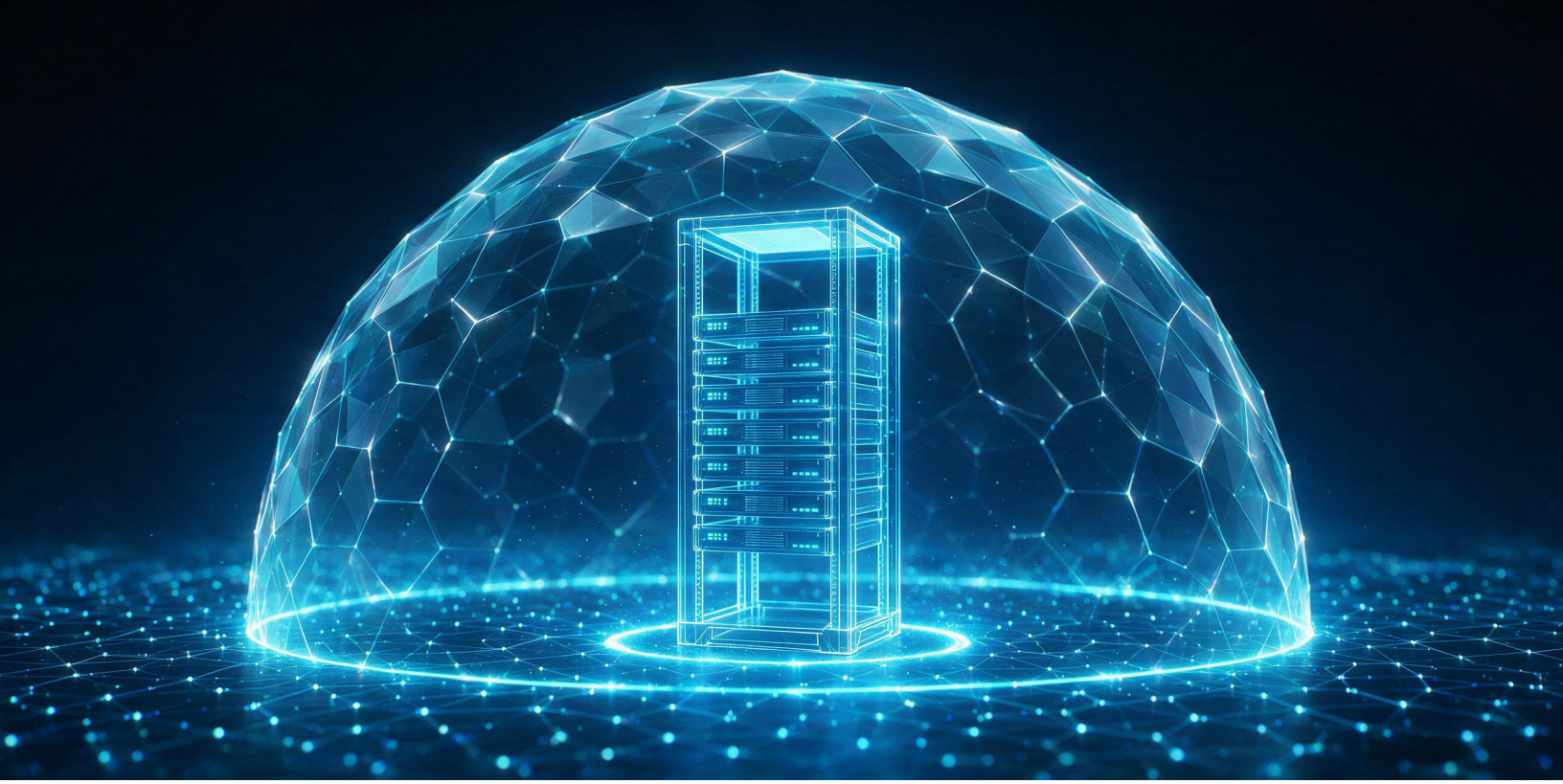
"If you have a predictable workload, like a very defined business operation, a fixed lease contract is the way to go. However, if it's experimentation, if it's scaling, if it's variable velocity, that's when you need to go on demand. Talking to CFOs, what I recommend is budgeting for both."

– Brennen Smith, CTO, Runpod

Across the access layers, portfolio allocation options include:

- Serverless inference APIs: experimentation and low, unpredictable volume.
- Hyperscaler on-demand GPU: burst capacity and integration with native services.
- GPU marketplaces: multi-provider optionality and standardized deployment.
- Reserved bare metal, direct (QumulusAI): production inference baseline, fine-tuning, and custom deployments. These workloads are not subject to per-token pricing; economics are governed by hardware utilization. QumulusAI's distributed compute network adds proximity to end users and data sources, reducing inference latency across geographies as agentic orchestration compounds round-trip delays.
- Edge and on-premises: latency-critical inference, regulated data, and offline requirements.

Predictable, high-utilization production inference belongs on reserved capacity, where hardware utilization rather than per-token pricing governs cost. Experimentation, bursts, and variable velocity belong on demand. The winning posture will be a portfolio based on the discipline to know when a workload has stabilized enough to migrate down the stack toward reserved economics.



The Cost Curve Favors Locking In Capacity Now

Organizations locked into single-tier hyperscaler pricing face exponential cost exposure as agentic workloads multiply token consumption. On reserved bare metal, inference is not subject to per-token pricing; cost tracks hardware utilization, the same whether a request generates 1,000 tokens or 100,000. Teams that establish hybrid control now, with reserved bare metal as the cost-optimized production tier, gain compounding advantage with each year of growth. Planning the next capacity cycle ahead of contract expiry protects that advantage, because GPU availability is not guaranteed when it is time to scale.

"With dedicated hardware, you have more control, and thereby you serve more users, and you generate more revenue. If you reserve an instance, let's say 6 months before your contract ends, you've got to start thinking about the next generation. That way, you ensure that you have continuity, and your revenue continues to grow. There may not be available GPUs because somebody else consumed it when you needed to scale."

– Pranay Prakash, CEO, Qubrid AI]

Methodology

This report draws on Futurum Research market forecast and decision-maker survey data from the first half of 2026, cited below, and on analyst-led interviews conducted in June 2026 with executives at Runpod, Qubrid AI, and Amberd.ai. Each interviewee consented to attribution by name and company. Quoted material has been edited only to remove filler and correct grammar slips; wording and substance are unchanged.

Data Citations:

1 Futurum Research, Data Center Semiconductors Deployment Forecast, 1H 2026.

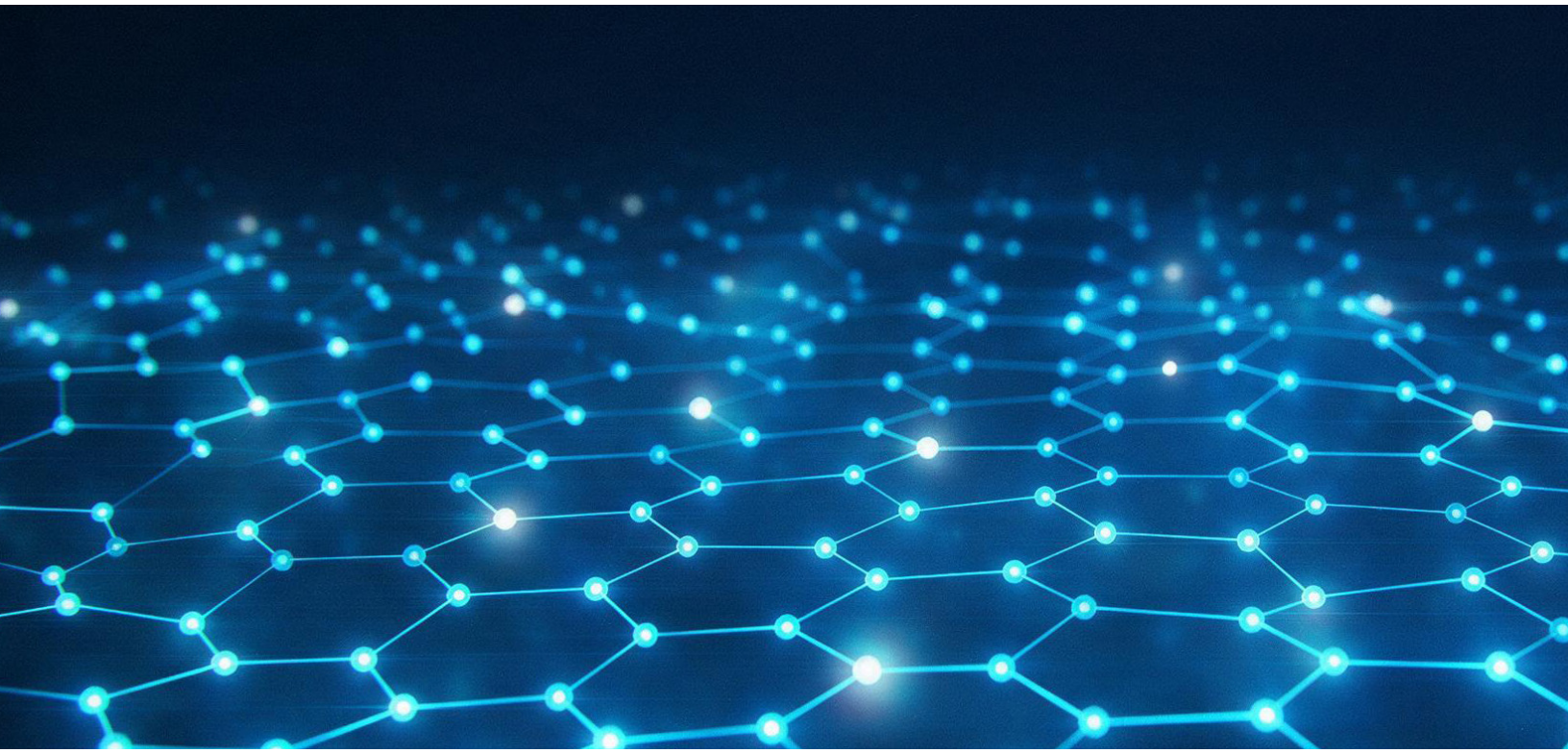
2 Futurum Research, Data Center Semiconductors Workload Forecast, 1H 2026.

3 Futurum Research, AI Chipsets Decision Maker Survey – Deployment Method, 1H 2026.

4 Futurum Research, AI Chipsets Decision Maker Survey – Decision Drivers, 1H 2026.

5 Futurum Research, AI Platforms Decision Maker Survey, 1H 2026.

6 Futurum Research, AI Platforms Market Forecast, 1H 2026.



Important Information About This Report

Authors

Brendan Burke
Research Director, Semiconductors,
Supply Chain & Emerging Tech | The Futurum Group

Publisher

Futurum Research

Inquiries

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

Citations

This paper can be cited by accredited press and analysts, but must be cited in context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

Licensing

This document, including any supporting materials, is owned by The Futurum Group. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of The Futurum Group.

Disclosures

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

QumulusAI

ABOUT QUMULUSAI

QumulusAI is a GPU cloud and AI infrastructure provider headquartered in Atlanta, Georgia, with additional data center sites in Oklahoma, Missouri, and Texas. An NVIDIA Cloud Partner, the company offers shared, dedicated, and bare-metal GPU compute built on NVIDIA's B300, B200, H200, H100, and RTX Pro 6000 chips, supporting workloads that range from single-GPU inference to large-scale model training. That compute runs on infrastructure QumulusAI controls end to end, including its own power generation, Tier 3+ data centers, and low-latency networking and storage. The company positions itself as a more cost-transparent alternative to hyperscale cloud providers, with all-inclusive pricing and direct operational control across the stack. QumulusAI serves both enterprise customers with dedicated infrastructure needs and startups seeking flexible, on-demand GPU access.

Futurum

About The Futurum Group

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

Futurum

CONTACT INFORMATION: The Futurum Group LLC | futurumgroup.com | (833) 722-5337

© 2026 The Futurum Group. All rights reserved.