WHITE PAPER

Beyond the Rack: Reinventing High-Efficiency Infrastructure to Power the Next Generation of Hosting and Online Gaming

MAY 2026

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Hosters and online gaming providers face a brutal economic reality: demand for compute power and immersive experiences is growing rapidly while demand for data center space and power is skyrocketing even faster. Lag, latency, and downtime are revenue killers. Today's typical "server sprawl" makes it difficult to meet efficiency standards. The ability to extract more value from every rack is now a competitive necessity.

Modern high-density architecture solves this dilemma by drastically reducing the physical footprint required for massive workloads. Supermicro MicroCloud takes this approach further, combining multi-node density with operational simplicity to maximize revenue per rack without expanding space or power.

The Supermicro MicroCloud platform's modular design empowers hosting and online gaming teams to scale performant resources where and when they are needed. The 10-node system maximizes core counts for functional hosting needs, while the specialized 5-node GPU configuration unlocks new possibilities for graphics-intensive workloads.

It's time to rethink rack economics and future-proof the data center for real-time workloads, rising player concurrency, and revenue-critical performance demands.

RETHINKING RACK ECONOMICS: FROM SQUARE FOOTAGE TO PERFORMANCE PER WATT

Why Rack Density Now Determines Who Can Scale

For hosting and gaming providers, growth depends on how much performance you can extract from the footprint you have. In metro data centers, space is limited, and rack level power caps are strict, yet proximity to users helps meet latency requirements. High density, high efficiency designs deliver more compute per rack, so capacity can grow without new real estate or larger power feeds. The result is better unit economics—more customers served, more revenue generated, more flexibility in how capacity is deployed—while staying close to customers.

At the same time, demand keeps climbing. Content is richer, cloud services are more pervasive, and player and user concurrency are rising, all of which drive more compute demand within the same or smaller footprint. Increasing density means consolidating more capability into each rack, keeping latency low, and pacing your expansion on your terms.

WHY DENSITY NOW?

- More compute per rack → lower cost to serve
- Metro friendly power envelopes → deploy where latency matters
- Fewer racks for the same capacity → delay buildouts and improve ROI

Density & Efficiency Comparison

High-density multi-node architectures deliver dramatic improvements in nodes per rack while staying within real-world metro data center power envelopes. As an example, MicroCloud delivers more than 3× the node density of traditional 1U servers while using ~40% less power per rack.

Platform	Form Factor	Nodes per Chassis	Nodes per Rack (typical)	Power Envelope (typical)	Nodes per kW	Density Advantage vs Traditional 1U
Traditional Single-Node Servers	1U	1	42	12–15 kW	2.8–3.5	Baseline
Typical Vendor Multi-Node	2U	4	80–84	10–13 kW	6.5–8.0	~2.3×
Supermicro MicroCloud	3U	8	96	9 kW	10.7	3.1×+

HOW TO BOOST REVENUE PER RACK WITH LEAN, FUTURE-READY PLATFORMS

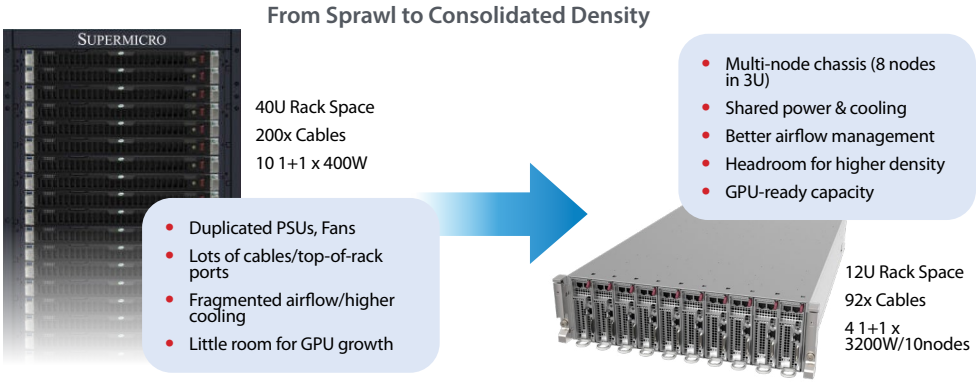
The mandate is simple: increase ROI, keep operations lean, and protect capital as workloads evolve. High-density, high-efficiency designs make this achievable within today's space and power constraints.

Infrastructure leaders should evaluate platform choices across three key requirements:

Requirement	Why It Matters	How To Implement It
Increase Revenue Per Rack	Spread fixed costs across more customers: • Lower cost to serve • Improve margins • Increase pricing flexibility	Consolidate into multi node, high density systems that raise usable node count within the same power envelope.
Simplify Operations	• Reduce downtime and costs • Improve serviceability, shorten maintenance windows, and reduce configuration drift • Reduce labor and incident risk	Use platforms that incorporate space/switches/cabling and hot swap serviceability; standardize on a repeatable chassis.
Protect infrastructure investments	Avoid forklift-style upgrades and pace spending in smaller, more predictable steps aligned to workload mix changes	Choose socket stable platforms; add GPU acceleration as needed without redesigning racks.

High density platforms increase throughput per rack, lower operating cost per customer, and defer new buildouts by extracting more capacity from existing footprints. At the same time, simpler day to day operations and investment protection keep teams focused on growth, not infrastructure churn.

ACT NOW TO UNLOCK RACK CAPACITY, SERVICEABILITY, AND GPU READINESS



Many server footprints were designed for a different era of scale and efficiency. Dozens of small, independent servers duplicate power supplies, fans, and cabling, driving up power and cooling in the rack while limiting the amount of usable compute that can fit. What might be simple at a small scale becomes a structural barrier, especially when space and power are tight or there's a need for GPU ready capacity.

LEAN, FUTURE-READY PLATFORMS IN PRACTICE

- **Hosting:** Standardize on a multi node chassis to add capacity in small, repeatable blocks. Service in minutes, not hours.
- **Gaming:** Choose platforms that deliver strong single core performance. Match tick rates and real time updates. Add GPUs only where features demand them and without a rack redesign.

Operational drag is just as real. Managing large fleets of single-node instances adds friction to provisioning, patching, and troubleshooting; increases the frequency of configuration drift; and slows response during peak periods or major launches. Older infrastructure wasn't designed for the thermal loads, higher rack power, faster interconnects, and mixed CPU/GPU growth demanded by today's AI-infused workloads. The result is a platform that limits scale and reliability, delays the introduction of new services, and increases the cost of serving each customer.

Delaying modernization increases cost and risk. Modern, modular, multi node chassis consolidate power and cooling, reduce cabling, and create the headroom needed for higher density and GPU assisted services—without requiring a proportional increase in space or operational burden. Upgrading converts stranded capacity into revenue per rack and puts teams on a path to scale efficiently to support growth.

NEW LEVERS FOR DENSITY: CPUS, GPUS, AND MODULAR CHASSIS

Modern processors are redefining what dense architectures can deliver within strict power envelopes. Supermicro MicroCloud shows how this works.

More work per watt: why CPU efficiency drives rack yield

New processor generations deliver more useful work from the same power budget, turning each rack into a higher yield asset. They're faster at both kinds of work that matter: single threaded tasks that keep experiences responsive and automation determinative, and multi threaded jobs that drive concurrency and scale.

The results are revealed in unit economics. There is more throughput without a larger power bill and more capacity within the same space. In real deployments, higher performance per watt has enabled higher node counts per rack while staying within common power caps. This supports growth in metro sites without chasing new space or larger feeds.

GPU-Ready from day one: add acceleration without a rack redesign

Adding acceleration to services that need it shouldn't require a rebuild or overprovisioning. GPU ready designs prepare the chassis, power, cooling, and networking from the start. This makes it possible to diversify workloads from CPU only to mixed CPU/GPU by adding cards, not redesigning racks or reworking power. That agility creates immediate business options.

For a game or streaming launch, a provider can only add GPUs to a necessary subset of servers to light up a premium graphics service—enabling upsell without the need for a new platform. For a spike in media or creator workflows, the provider can turn on GPU assisted processing for the event, then dial it back when demand eases. This is a faster time to market and smaller steps for capital expenditures, with higher utilization across the cycle. Because the underlying form factor remains the same, operations stay simple and predictable, and future upgrades don't require long maintenance windows or a platform swap.

Add nodes, not racks: a repeatable path to growth

Modern chassis concentrate capacity and cut complexity. With multiple independent nodes in an enclosure of only 3U, organizations can raise usable density per rack while consolidating power, cooling, and cabling, simplifying installations, and streamlining day to day operations. And scaling is more straightforward than ever. Add nodes or chassis in small, repeatable steps, more closely matching costs with demand. The result is higher revenue per rack, fewer failure points, easier service continuity management, and reduced demands on space, wiring, and maintenance.

Hosting and gaming providers evaluating high-density infrastructure typically compare several architectural approaches, each optimized for different tradeoffs in rack density, power utilization, GPU readiness, and operational simplicity. While many platforms can increase compute per rack, they behave differently when constrained by fixed rack-level power limits or when GPU acceleration is introduced over time.

The table below compares the approach taken by commonly deployed hosting platforms to address these challenges, highlighting the architectural choices that influence rack efficiency, modernization paths, and operational scalability.

Platform Architecture	Node Density	Rack Density	Power Utilization	GPU Expansion Model	Service Model	Design Tradeoffs
Supermicro MicroCloud	8–10 independent nodes	Up to ~96 nodes per 42U rack	Maximizes nodes within common rack level power caps	Dedicated GPU enabled configuration within the same chassis and power envelope	Hot swap, chassis level service optimized for high node counts	Optimized for rack-level density efficiency and modular growth rather than large single nodes
Dell PowerEdge C6520 / C6620	4 nodes	~80–84 nodes per 42U rack	Balanced performance and density, often favoring trading rack-level efficiency for higher per node power	GPU support available in specific configurations	Tool less, front serviceable components	Density achieved via fewer, higher capacity nodes
HPE ProLiant DL385 / Apollo	4–8 nodes (configuration dependent)	~70–88 nodes per 42U rack	Emphasizes performance flexibility, typically at higher per node power	GPU capable designs across select platforms	Enterprise grade serviceability and lifecycle tooling	Flexibility across workloads may reduce rack level density efficiency
Lenovo ThinkSystem SD550 V3	4 nodes	~80 nodes per 42U rack	Tuned for predictable power profiles and performance consistency	GPU enabled configurations available within defined configurations	Conventional multi node service model	Expansion paths are more fixed once deployed
Other, e.g., Gigabyte, QCT.	4–6 nodes	~70–90 nodes per 42U rack	Varies by configuration and integrator design choices	GPU support varies by model and integration	Service model dependent on partner and deployment approach	Outcomes depend heavily on customization and operational maturity

Table 3: Architectural Approaches to Rack Dense, Power Constrained Hosting

PROOF IN PRODUCTION: HOW MICROCLOUD ENABLED SCALE IN GAME HOSTING

A real-world deployment demonstrates how high-density, modular infrastructure translates into measurable performance and operational gains.

i3D.net, a Ubisoft company, needed to add capacity in metro facilities with tight per-rack power caps in order to host large multiplayer games. While changing the infrastructure, i3D.net had to preserve jitter free, fair, synchronized gameplay. The team prioritized single core speed and higher node density per rack to scale efficiently inside their existing space.

They chose Supermicro MicroCloud with AMD EPYC™ 4464P processors, with eight independent nodes in a 3U enclosure. The rollout delivered 52% higher single core performance and twice the node density (to 96 nodes/rack) while remaining within a 9 kW power envelope, a practical fit for crowded metro data centers. Operations also improved by consolidating hosting space, switches, and cabling, while selecting 12 core, 65-watt CPUs over 16 core, 170-watt alternatives reduced TCO per user.

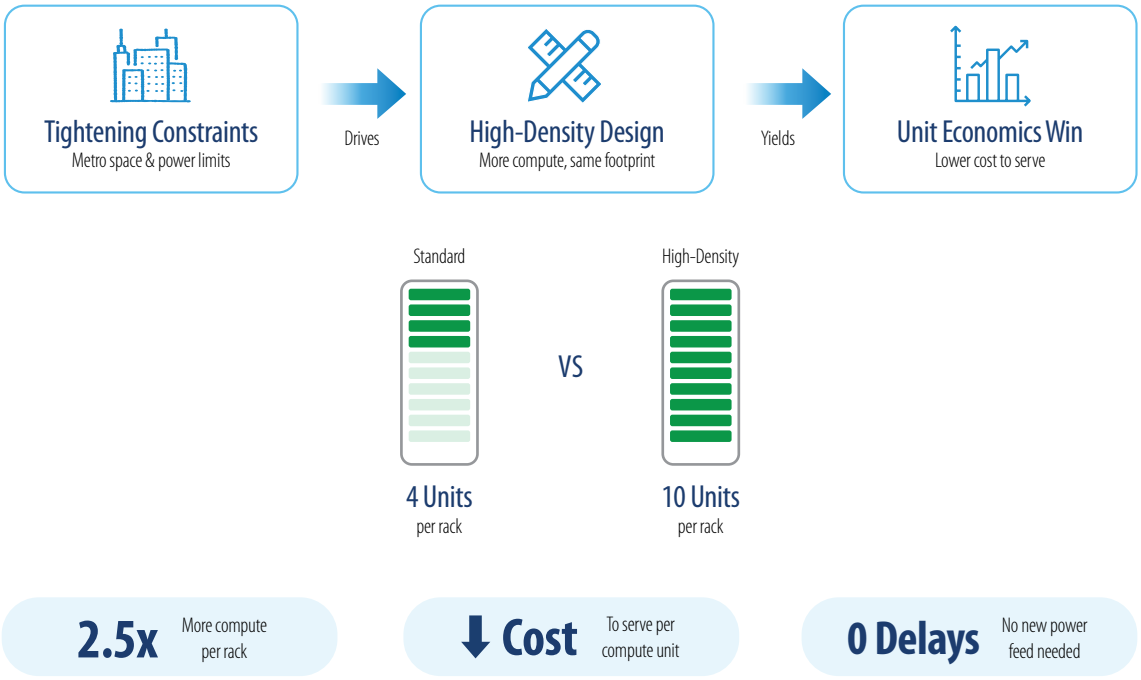
MicroCloud's generation-to-generation socket consistency provided a clean CPU refresh path within the same chassis. This extended the infrastructure lifecycle by enabling simple future upgrades, maintaining capacity close to users within existing real estate, and aligning builds with unit economics goals.

Challenge: Add capacity for large multiplayer game titles in metro data centers with strict per rack power limits, while ensuring fair, synchronized, jitter free gameplay (consistent tick rates, predictable latency).

Solution: Supermicro MicroCloud with AMD EPYC™ 4464P, eight independent nodes in a 3U chassis.

Results:

- 52% higher single core performance vs. the prior platform
- Gameplay experience protection: smoother, jitter-free gameplay with predictable latency
- 2× node density (up to ~96 nodes per rack) within 9 kW cap.
- Lower cost per user to serve and simplified operations.
- Investment protection with an extended platform lifecycle and a clean upgrade path



SUPERMICRO: AN OPEN ECOSYSTEM AND FAST SKUS, PROVEN AT SCALE

Supermicro MicroCloud pairs high density with operational simplicity in a form factor that fits the constraints many modern, innovative businesses face, enabling capacity increases without chasing more space or larger power feeds.

- **Density with range:** 3U MicroCloud chassis in 5 , 8 , and 10 node options lets teams tune for CPU only hosting or GPU assisted services without changing the rack design.
- **Simpler operations:** Consolidated space, switches, and cabling streamlines installation and day to day management while reducing failure points.
- **From box to rack:** Optional rack scale integration services, from design and cabling to burn in and delivery, reduce deployment time and operational risk.
- **Investment protection:** Socket consistency generation-to-generation extends the platform lifecycle.

Supermicro MicroCloud stands out with an open-ecosystem approach, open CPU, rapid SKU cadence, and rack scale integration services, enabling teams to deploy faster while maintaining flexibility as requirements evolve.

SUPERMICRO PRODUCTS COVERED

Supermicro MicroCloud Platform
3U chassis available in 10 node, 8 node, and 5 node configurations for CPU dense or GPU enabled workloads.

- AS -3015MR-H5TNR
- AS -3015MR-H8TNR
- AS-3015MR-H10TNR

AMD EPYC™ 4000 Series (e.g., 4464P)

- Efficient, right sized performance for hosting and gaming workloads; compatible across all MicroCloud node options.

SUPER MICRO COMPUTER, INC.

As a global leader in high-performance, high-efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs and enables us to build and deliver application-optimized solutions based on your requirements.

www.supermicro.com

Printed in USA ♻ Please Recycle 14_2026-BeyondTheRack_AMD_R02 **MAY 2026**