WEKApod™

Next-Generation AI Storage Appliances

Challenges

- Rack space, power, and cooling cap AI scale faster than GPU supply does.
- Legacy storage systems can't reach the density inference workloads require.
- Inference economics depend on tokens per rack unit, not just raw performance specs.

Solution

The new WEKApod family turns wasted rack space and power into more AI by fitting more storage into every rack unit than any available alternative.

Benefits

- More GPUs and more tenants fit in the same rack footprint.
- NeuralMesh data efficiency software compounds physical density into effective capacity.
- Augmented Memory Grid extends inference economics.

AI infrastructure is running out of data center, not out of ideas. Rack space, power, and cooling are constrained. GPU demand keeps climbing. WEKApod is the first AI storage platform built on hardware WEKA designed from scratch, a custom 2U chassis built around a PCIe Gen 6 internal fabric and a cable-based, backplane-free drive interconnect, with multiple patents pending. The hardware exists for one job: reclaim the rack space and power commodity storage wastes, and hand it back to compute.

WEKApod is a family of storage systems built on that chassis. WEKApod Prime Max delivers 1.1 Exabytes of effective capacity in a single 56U rack. WEKApod Nitro delivers 10.2 TB/s throughput and 210 million IOPS in a 56U rack. Each system is built for a different constraint.

For AI Cloud & Service Providers: For AI Cloud providers, storage density is a margin decision. Every rack unit not generating tokens is a black hole. WEKApod puts 267% more raw capacity per rack than the nearest alternative into the same footprint and the floor space and power that storage no longer consumes goes to GPUs and compute. More GPUs per rack means more tenants per rack. More tenants per rack means more revenue from existing infrastructure.

For AI & ML Engineering Teams: Deploy in hours with microsecond latency at scale—storage that never limits GPU performance, with full enterprise features (GPUDirect Storage, instant snapshots, multi-protocol access) and no storage engineering expertise required.

For CIOs and & Infrastructure Leaders: For enterprise AI teams, the move from training to production inference exposes infrastructure decisions that were invisible during pilots: storage latency that inflates

time-to-first-token, insufficient context memory that forces repeated prefill, and platforms not built for concurrent workloads at scale. WEKApod ships factory-installed and pre-validated with the full NeuralMesh feature set, including Augmented Memory Grid, with no dedicated storage team required.

The WEKApod Family

Three configurations cover every inference and training deployment from sub-1 PB entry deployments to over 1 EB at scale. Every configuration ships factory-installed: Ubuntu pre-installed, BIOS pre-configured, iDRAC pre-set, full NeuralMesh software stack loaded.

WEKApod Prime Max

The most ***capacity dense*** system.

Best for: Enterprises and AI Cloud providers operating at Exabyte scale where effective capacity per rack unit is the primary constraint.

WEKApod Prime

Industry-leading ***price-performance***.

Best for: Enterprise teams that need both performance and capacity economics from the same system — without managing two separate storage tiers.

WEKApod Nitro

Extreme performance for AI factories.

Best for: AI Cloud providers and AI factories where training throughput and GPU saturation are the primary workload requirements.

DCAI

"Space and power are the new limits of innovation in data centres. WEKApod's exceptional storage performance density allows us to deliver hyperscaler-level data throughput and efficiency within an optimised footprint—unlocking more AI capability per kilowatt and square metre. This efficiency directly improves economics and accelerates how we bring AI innovation to our customers."

– Nadia Carlsten, CEO, Danish Centre for AI Innovation (DCAI)

WEKApod Technical Specifications

Specification	Prime Max	Prime	Nitro
Form Factor	2U, 2 Servers	2U 4 Servers	2U, 4 Servers
NVMe Drive Type	QLC / TLC	QLC / TLC	TLC
SSD Slots	70	56	56
Raw Capacity (2U)	15.8 PB	12.8 PB	1.72 PB
Raw Capacity (56U)	441.5 PB	358.5 PB	48.2 PB
Usable Capacity (56U)	371.9 PB	302.2 PB	40.7 PB
Effective Capacity (56U)	1.1 EB	906.5 PB	122.0 PB
Throughput (56U)			10.2 TB/s
IOPs (56U)			210 M IOPs
Data Network	NVIDIA ConnectX-8 / 9, NDR / 800GbE (all configurations)		
Internal Fabric	PCIe Gen 6 (all configurations)		
Failure Domains	2 per 2U	4 per 2U	4 per 2U
Protocols	POSIX, NFS, SMB, S3, GPUDirect Storage (all configurations)		
Software	NeuralMesh 6.0, AlloyFlash, Drive Sharing, Data Reduction, Augmented Memory Grid		NeuralMesh 6.0, Drive Sharing, Augmented Memory Grid

🔗 **Discover how WEKApod can help you today**