WEKApod. More AI. Same Data Center. No Compromise

How AI Cloud providers and enterprise teams get more AI out of the datacenter they already have.

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.

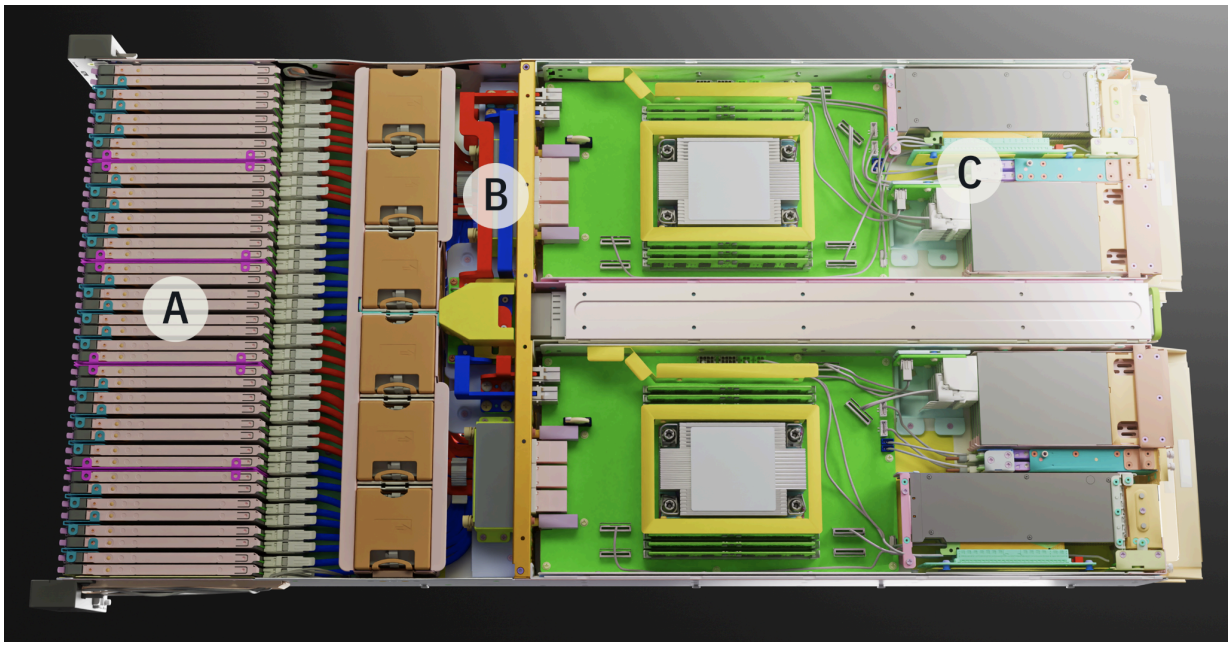
AI Cloud providers are running out of room. US datacenter construction fell in 2025 for the first time since 2020, even as AI demand accelerated. Grid connection queues in major markets now stretch 4 to 7 years. Every new GPU generation arriving into constrained physical infrastructure makes the efficiency of every other layer of the stack more consequential. Storage that consumes rack space at low density is directly competing with GPUs for the same constrained physical resources.

WEKApod was built to change that math. WEKA designed the chassis from scratch, with multiple patents pending on the core innovations inside it. A PCIe Gen 6 internal fabric gives every drive full bandwidth simultaneously. A cable-based, backplane-free drive interconnect creates four independent failure domains per 2U, so a single drive event stays a single drive event. WEKApod Prime Max packs 70 NVMe drives into 2U, QLC and TLC combined, producing 15.8 PB of raw capacity in that footprint. Across all three dimensions that determine whether a rack is working or wasting space — capacity, throughput, and IOPS per rack unit — no publicly available alternative comes close.

For AI Cloud providers, that density advantage is a margin decision, not an infrastructure decision. More storage capacity per rack unit means more floor space and power available for GPU compute. More GPUs per rack means more tenants per rack. More tenants per rack means more revenue from infrastructure already committed to.

What Happens When Hardware and Software Are Built Together

Custom hardware alone isn't the differentiator. Vendors from Everpure to IBM to VAST ship their own chassis designs. What none of them do is design that hardware around software efficient enough to spend the power and cooling budget on drives instead of compute. Every storage platform operates inside a fixed power and cooling envelope that has to cover compute, memory, networking, and drives. Kernel-based I/O and inline data reduction both burn CPU cycles, and that overhead costs power to run and heat to dissipate, budget that comes straight out of what could go to drives. NeuralMesh doesn't carry that overhead: kernel-bypass I/O in user space, Data Reduction entirely off the write path, and a lower memory footprint per server. That's the power and thermal headroom WEKApod spends on 70 drives (A) or up to 4 servers (C) per 2U. The density gap isn't about who owns a chassis design. It's about whose software earns the right to spend the power budget on capacity instead of overhead.



WEKApod's hardware is the product of a deliberate design process at WEKA, producing innovations that exist nowhere else in the market. The PCIe Gen 6 internal fabric (*B*) eliminates the internal bandwidth bottlenecks that limit drive count in commodity server platforms. The cable-based, backplane-free interconnect provides independent failure domains — 4 per 2U chassis — so drive events don't cascade to adjacent devices or require cluster-level intervention. The thermal architecture is engineered for 35°C ambient operation, the real conditions of production AI datacenters, not lab specifications. Under thermal stress, NeuralMesh reduces NVMe drive power in software rather than triggering shutdown. The system stays up. Tokens keep moving.

Serviceability was designed in from the start. Boot drives are hot-pluggable via a custom interposer with a GUI-guided replacement workflow. The NIC port layout keeps all drive bays accessible during maintenance without moving networking hardware. A drive replacement is a 10-minute self-service operation. At production inference scale, that's the difference between a routine maintenance event and an SLA clock running.

The Software That Makes Density Compound

The physical density of the WEKApod chassis is where the advantage starts. NeuralMesh data efficiency software is where it compounds. Three capabilities work together to push effective capacity and performance beyond what the raw spec delivers.

AlloyFlash steers writes intelligently across mixed TLC and QLC flash by I/O size on WEKApod Prime. Customers get QLC economics and capacity density without QLC performance penalties. It's what makes Prime's mixed-flash configuration production-viable, not just capacity-on-paper.

Drive Sharing lets multiple NeuralMesh clusters share physical NVMe drives, extracting more parallelism and utilization from each device. For AI Cloud providers running multi-tenant inference workloads, that means more active workloads per physical drive, and more revenue per rack unit.

Data Reduction runs entirely off the write path, background-first similarity-based compression and cross-filesystem deduplication, so writes commit at native NVMe speed with no sustained-write cliff during checkpoint-heavy training. WEKApod Prime Max's 1.1 Exabyte effective capacity figure is the result: 441.5 PB of raw capacity with data reduction running on top of it.

NeuralMesh was rebuilt alongside the WEKApod chassis. Memory footprint optimization reduces baseline memory consumption per server — a requirement for operating at this drive density. Core allocation templatization automatically configures the software layer based on drive count and capacity at order time, making configurable SKUs operationally clean across every configuration from sub-1 PB entry deployments to 100+ PB at the high end. No OEM appliance vendor ships this because none of them built hardware around software efficient enough to spend the power budget on drives instead of compute.

Ready to Deploy. Ready to Scale.

WEKApod ships with NeuralMesh software factory-installed and ready to run. Ubuntu pre-installed, BIOS pre-configured, iDRAC pre-set, full NeuralMesh software stack loaded. The engineering target was explicit: eliminate 10 to 15 manual deployment steps versus the previous generation. Each eliminated step is a risk surface removed and a delay avoided. No storage specialist required.