OXMIQ + CORESPAN SYSTEMS

SC26 · CHICAGO · BOOTH 5003
NOVEMBER 17 TO 19, 2026

SOVEREIGN AI ON PREMISES

Enterprise AI on *your* infrastructure.

Your models, your data and your GPUs, in one rack.

Electrons in. Tokens out.

~490B
tokens a year on Kimi K2

2.3 TB
GPU memory at 43 TB/s

~20 kW
workload power budget per rack

1 to 8 racks
with no additional networking hardware

Expanded telemetry

Rich observability from the application layer down to deep hardware health.

RACK PERFORMANCE BASED ON 32 × NVIDIA® RTX PRO 5000 72 GB



WHAT'S IN THE RACK?

THE MACHINE

ETM-N32

Electron to Token Machine. Complete and self contained: compute, control and storage fully integrated.

GPUs	**Any vendor, up to 4 nodes**
Power	**~20 kW workload budget**
Scale	**1 to 8 racks, no extra networking**
Spec	**33 PFLOPS FP4 · 2,304 GB**

THE CHASSIS

Corespan

Chassis engineered and built by Corespan.

Fabric	**Photonic PCIe Gen5**
Servers	**Off-the-shelf servers of choice**
Resources	**GPU (NVIDIA, AMD, Intel), Storage and Networking**

THE SOFTWARE

Composer + OxCapsule

Bare metal to model endpoints. Provision, schedule, monitor and tune for tokens per dollar.

Composer	**Up to 32 GPU shared pool**
OxCapsule	**Runs models across your fleet[1]**
Telemetry	**Agents to Atoms**

Pricing under NDA. [1]oxmiq.ai/oxcapsule

Booth 5003. Bring your model.

Scan to book a session with OXMIQ and Corespan engineers.

oxmiq.ai · corespan.ai

THE COMPOSABLE INFERENCE SOFTWARE STACK

Every GPU composed. Every model unchanged.

Your models keep growing. Your servers do not. OXMIQ and Corespan let you build a machine that scales to your model needs. Run any GPU of your choice and keep the runtimes and libraries you already use. One rack can serve many tenants, and each sees only the machines you give it.

OXMIQ

oxcapsule

Scalable orchestration, fleet management and inference software for heterogeneous AI compute.

CAPSULE Queue → CAPSULE Serve → CAPSULE Model

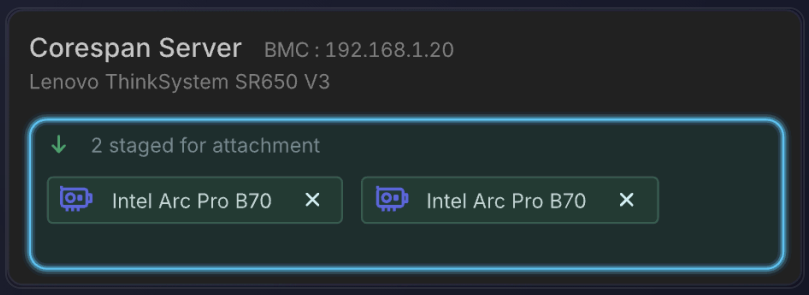
```
$ capsule orbit # live view of the fleet
$ capsule list --filter # by vendor or VRAM
```

CORESPAN

Composer

Composes 1 to 32 GPUs per host from shared pools.

- **A machine sized to the model.** Large models get one domain; small jobs stop leaving GPUs idle.
- **No recabling.** Nobody touches a cable.



THE GPU POOL

Any GPU: NVIDIA, AMD and Intel, pooled in PRU 2500 chassis.



HOW THE STACK FITS TOGETHER — A LOOP, EVERY TIME DEMAND SHIFTS

01 Compose — CORESPAN

Composer attaches the GPUs a model needs to a host from pooled PRU 2500 chassis, with NVMe storage close to the GPUs.

02 Cluster — OXMIQ

OxCapsule finds GPUs that can serve the model and can run it across multiple systems when it does not fit on one.[1]

03 Deploy — OXMIQ

OxCapsule uses the serving engines your team already knows, such as vLLM and SGLang, to deploy models onto the cluster. It handles capacity, bring-up and endpoints underneath.

04 Scale — CORESPAN

When demand shifts, Composer returns GPUs to the pool and recomposes for the next workload. OxCapsule sees the new fleet.

[1] oxmiq.ai/oxcapsule. Names are trademarks of their owners.

Booth 5003. Bring your model.

See a preview of Composer 4.0. Join the OxCapsule beta at oxmiq.ai/oxcapsule.[1]

OXMIQ Labs | oxmiq.ai
Corespan Systems | corespan.ai