hosted·ai

making AI cloud simple + profitable for service providers

- Build, manage + sell AI cloud infrastructure
- Software-defined, multi-tenant GPUaaS platform
- 100% GPU utilization for maximum efficiency
- GPU overcommit for maximum margins

Host AI model training + tuning + inference + traditional VMs

100% GPU UTILIZATION

hosted·ai enables 100% utilization by fully virtualizing GPU resources, delivering huge efficiency gains vs GPU passthrough or instancing

5x MARGIN

With **hosted·ai**'s software-defined GPU you can serve more customers per card and maximize returns from your GPU infrastructure investment

welcome to the future of AI hosting

build your AI cloud in 24h

hosted·ai is a complete AI cloud stack for your datacenter. GPU orchestration, multi-tenancy, pricing, packaging, consumption metering, app library and self-service are built in and ready to go.

designed for profitability

Sell GPU just like CPU, with GPU resource pooling and overcommit. **hosted·ai** maximizes utilization and margins with smart deployment of workloads across all GPUs in a cluster.

so it's easy to compete

hosted·ai's super-efficient platform reduces the CAPEX and OPEX of GPU cloud. You can offer lower prices to customers, or match competitor pricing and invest the extra margins in growth, or find the sweet spot in the middle.

Platform overview

Hyperconverged infra

hosted·ai is a hyperconverged CPU/GPU virtualization stack with an extremely efficient type 1 hypervisor and ultra-fast software-defined storage and networking.

Software-defined GPU

hosted·ai pools GPUs in a cluster and schedules user tasks across all available GPU resources. Each user process has isolated access to the full resources of a GPU.

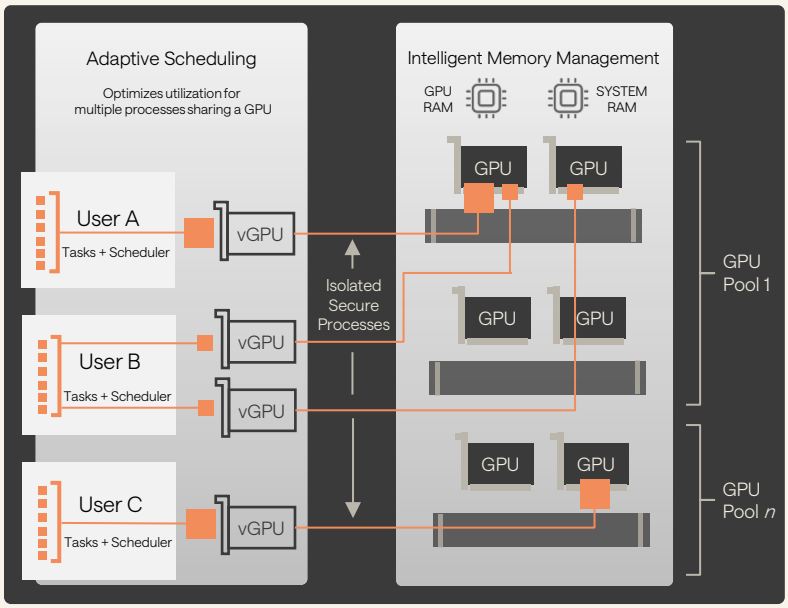
GPU overcommit

hosted·ai enables sharing ratios to be configured for each GPU pool. Resources are allocated intelligently based on priority: system RAM is used if insufficient VRAM is available.

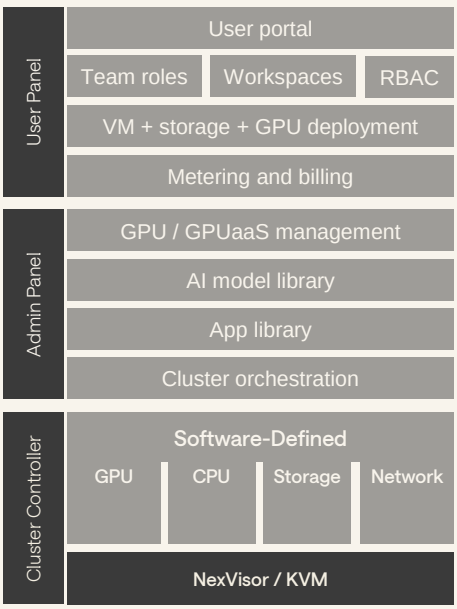
GPU and GPUaaS

hosted·ai enables private GPU resources and GPU passthrough as well as multi-tenant GPUaaS and IaaS cloud.

Software-defined GPU



Full cloud stack



Monetizing your AI cloud

- **GPU**: bill for consumption (TFLOPs/VRAM), or fixed resources per hour or month; set pricing for individual GPU pools or cards
- **CPU**: bill for vCPU cores
- **Storage**: bill for capacity
- **Network**: bill for throughput or fixed access
- **Regions**: set global or local prices for different datacenter locations
- **Packages**: combine resources into easy-to-consume packages (GPU, CPU, storage, network)
- **Applications**: combine applications with resources and bill for one-click installs

Pricing + deployment

- **hosted·ai** runs on commodity servers, supports a wide range of Nvidia GPUs, and different storage and network types
- **Full REST API** and integration with billing engines including WHMCS
- **Pricing** is based on the VRAM managed by the platform and consumed by your customers
- **24x7 support** with a 15-minute SLA is included as standard

hosted·ai

For a demo or more information:

hello@hosted.ai https://hosted.ai