



Open source, dynamic AI GPU sharing, scheduling and orchestration that accelerates AI throughput, delivers seamless auto scaling, and maximizes GPU utilization.

10x

GPU Availability

10x

Workloads Running

50%

GPU Utilization

0

Manual Intervention

WHY HAMi?

01 Maximize GPU Utilization

GPU on-demand sharing, scheduling optimization, and priority strategies eliminate waste, reduce costs, drive efficiency, and accelerate AI initiatives.

02 Seamless Auto Scaling

GPU on-demand auto-scaling: Seamless VPA scaling for AI workloads without restarts during GPU consumption surges.

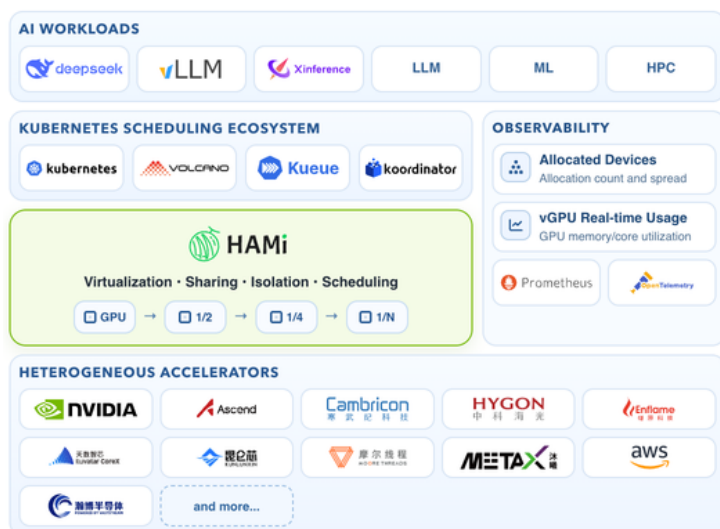
03 Unified Memory Out of the Box

Use HAMi to integrate NVIDIA unified memory in Kubernetes and enable GPU memory oversubscription.

04 Flexible Integration

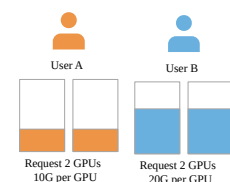
The open architecture integrates seamlessly with any toolchain (e.g., Kueue, Koordinator, Volcano) and infrastructure across public clouds, private clouds, or on-premises data centers.

ARCHITECTURE

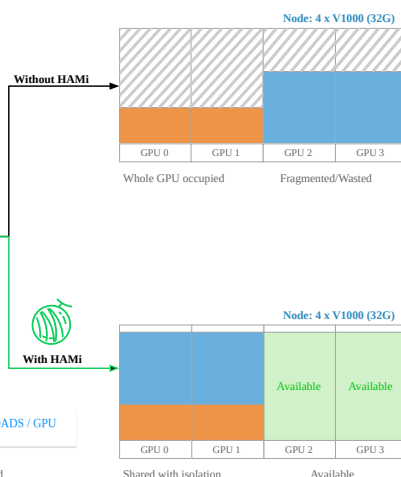


Before and After using HAMi

Compare traditional whole-GPU allocation with HAMi GPU sharing under the same workloads.



HAMi packs fragmented AI workloads onto fewer GPUs while preserving isolation.



Learn how HAMi improves GPU utilization in Kubernetes → <https://project-hami.io>

USE CASES

OCR OCR and Translation

Scientific research (e.g., PyTorch, TensorFlow) with co-located, fragmented algorithms sharing GPUs

Colleges and Labs

Multiple small/medium models share GPUs seamlessly and flexibly to maximize utilization.

AI Hybrid Cloud AI Workloads

A lightweight, non-intrusive, centralized solution is needed to maximize GPU utilization across public clouds, private clouds, and on-premises data centers.

AI Training and Inference Platform

Colocation of training and inference tasks, flexible GPU sharing, seamless auto scale, over-allocation of GPU memory.

Cost Effective GPU Service for Startups

GPU sharing, oversubscription, and task priority enable flexible sales and leasing models, reduce costs.

Finance

Model fine-tuning, GPU virtualization, and precise scheduling.

ADOPTERS

 DYNAMIA

 DaoCloud

 范式
PHANCY

 Rise Union

 HUAWEI

 LinkedIn

 SAP

 (SF) TECHNOLOGY
顺丰科技

 viettel

 SNOW

 NIO

 贝壳
让家更美好

 PREP

 招商银行
CHINA MERCHANTS BANK

 科大讯飞
IFLYTEK

 PPIO

 OpenCSG
开放传神

 H3C

 BONC
东方国信

 百度智能云

 马上消费
WWW.MSXF.COM

 中国东信
China-ASEAN
Information Harbor

 中国移动
China Mobile

 中国联通
China unicom

 赛力斯

 哈尔滨工业大学
HARBIN INSTITUTE OF TECHNOLOGY

 COOCCA 酷开

 基石智算
CoresHub

 SANGFOR
深信服科技

 ZStack

AND MORE...

GROWING GLOBAL COMMUNITY

500+

Contributors

300+

Adopters

4K

Stars

1.4K+

Commits

323K

Image Downloads

 <https://project-hami.io>

 <https://github.com/project-hami/hami>



GitHub



Discord