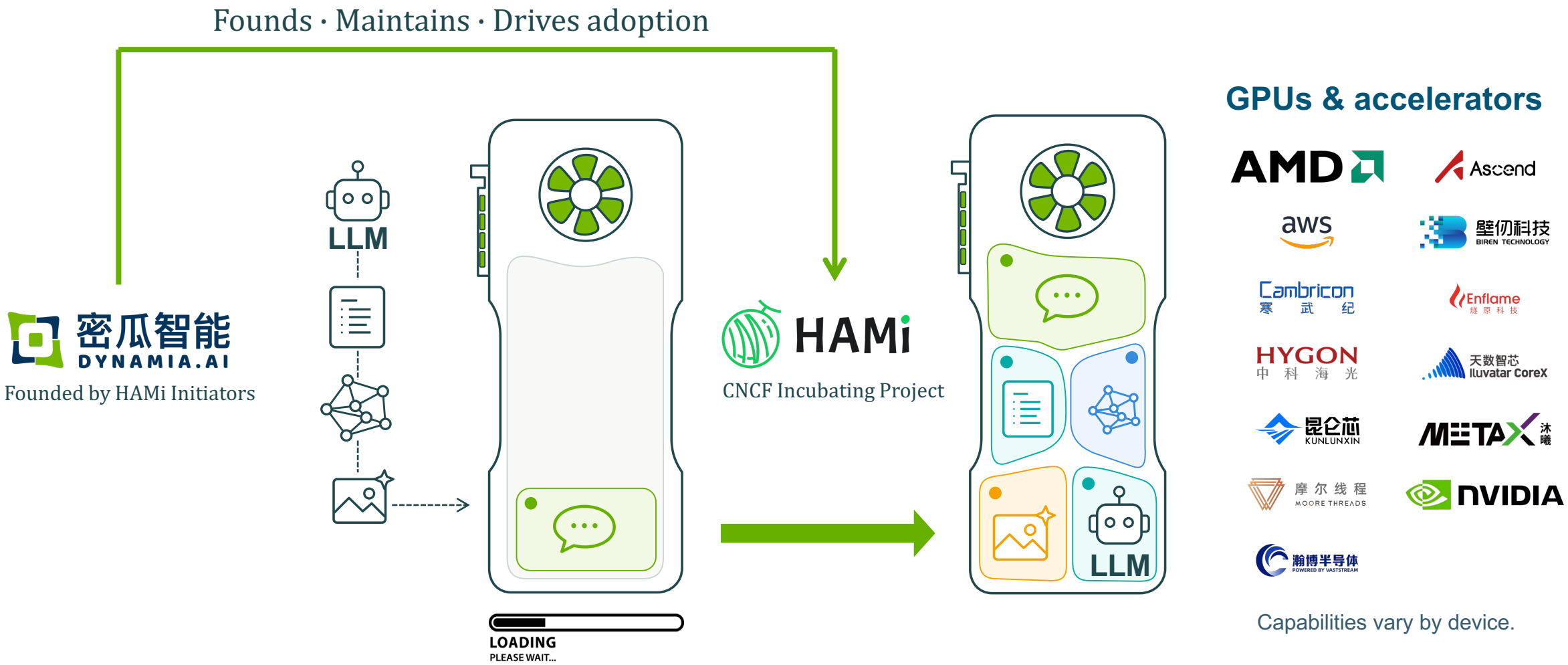
GPU sharing with HAMi



Xiao Zhang · CEO, Dynamia · HAMi Maintainer

HAMi in the cloud native ecosystem

密瓜智能 DYNAMIA.AI | HAMi

Contributors	Organizations	Countries and regions	Incubating since July 2026
723	1075	44	CLOUD NATIVE COMPUTING FOUNDATION

KAI Scheduler → HAMi-core

Per-container GPU memory limits

Schedulers

VOLCANO koordinator Kueue

Cloud & AI platforms

UCLOUD DaoCloud 金山云

HAMi in production

招商銀行 CHINA MERCHANTS BANK

+46.7%

Single-card inference throughput

ResNet 50 · batch 32

SNOW

30% fewer GPU hours

Driving simulation

NIO

50% fewer GPUs

Train-to-inference pipelines

PREP

90% of GPU infrastructure

Managed by HAMi

HAMi adopters

SF TECHNOLOGY 顺丰科技 · 平安银行 PING AN BANK · 平安证券 PING AN SECURITIES · XW 新网银行 · 贝壳 让家更美好 · 新浪微博 weibo.com

DaoCloud · 基石智算 CoresHub · SERES 赛力斯 · 元戎启行 · 同程旅行 · BONC 东方国信 · SAP

LinkedIn · H3C · 科大讯飞 iFLYTEK · Technische Universität München · viettel · InferVision

Dynamia’s engineering contributions

llm-d

MetaX integration

Prefill / decode compatibility

Under review

Lupine

vLLM compatibility

Per-client GPU metrics

Merged upstream

Later today · 8 Sep (UTC+8)

09:59–10:04	PD Disaggregation vLLM Deployment on Alternative AI Accelerators Using llm-d
11:14–11:19	Project Lightning Talk: From Static Slices to Elastic GPUs: Dynamic MIG with HAMI
14:30–15:00	How Intsig Serves Billions of Document Scans: GPU Virtualization at Scale with HAMi

HAMi community

WeChat assistant