APNET '26

SINGAPORE | 7 AUGUST 2026

RAG on the Decentralized Web

Akash and Golem · two decentralized compute markets

Suting Chen | Matteo Varvello | Aleksandar Kuzmanovic | Yunming Xiao

Northwestern University | Nokia Bell Labs | The Chinese University of Hong Kong, Shenzhen

Decentralized compute markets

IDLE SUPPLY

Spare compute

⟷

Compute dApp

Connects requestors and providers

WORKLOAD DEMAND

Application workloads

> **Can low prices deliver useful application performance?**

Decentralized compute markets

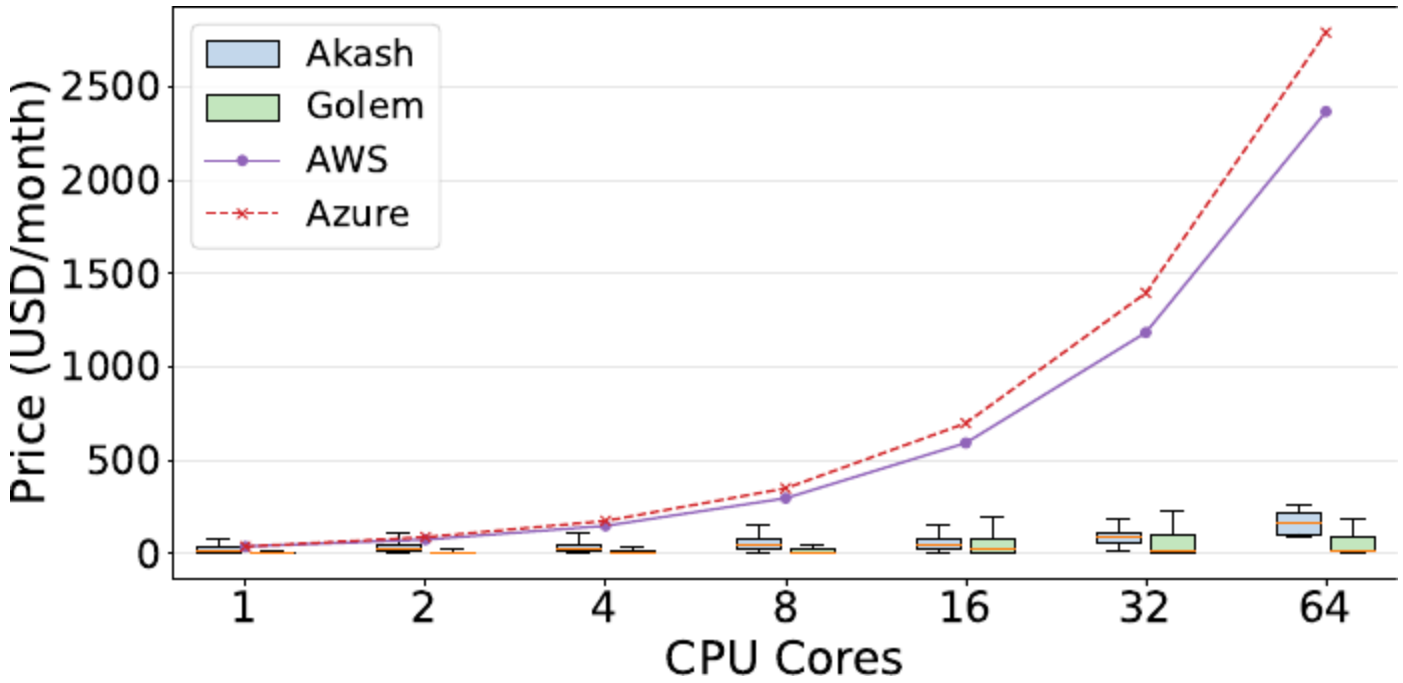
From request to running job

The market coordinates bidding and payment onchain; providers execute jobs off-chain.

1	REQUESTOR **Workload description**	4 CPU cores · 8 GiB RAM · at least 200 GiB disk
2	PROVIDERS **Auction bidding**	A: $3.10/day · Ohio | B: $3.60/day · Singapore
3	REQUESTOR + PROVIDER **Agreements onchain**	Provider B · $3.60/day · $15 allocated
4	REQUESTOR → PROVIDER **Send configuration**	Docker configuration sent off-chain

Decentralized compute markets

Compare listed prices

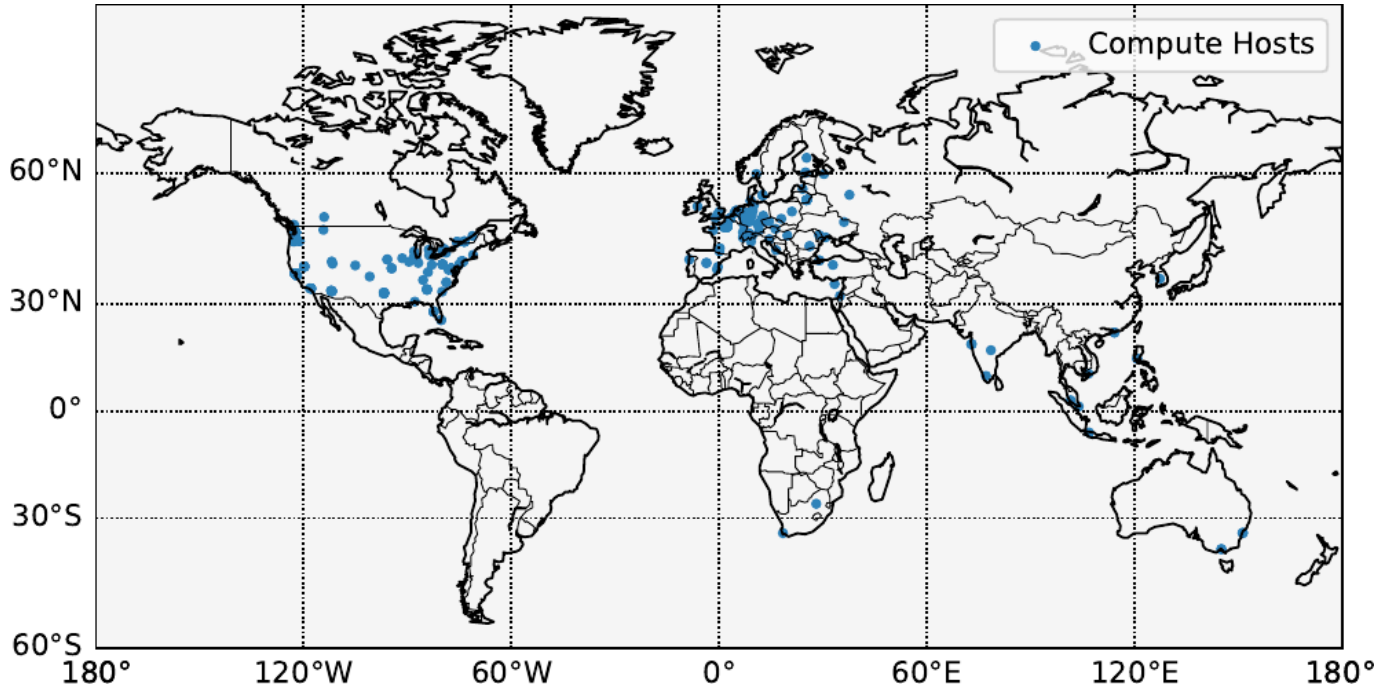


Akash and Golem list lower prices at the same nominal core count.

Nominal cores are not performance-normalized.

Decentralized compute markets

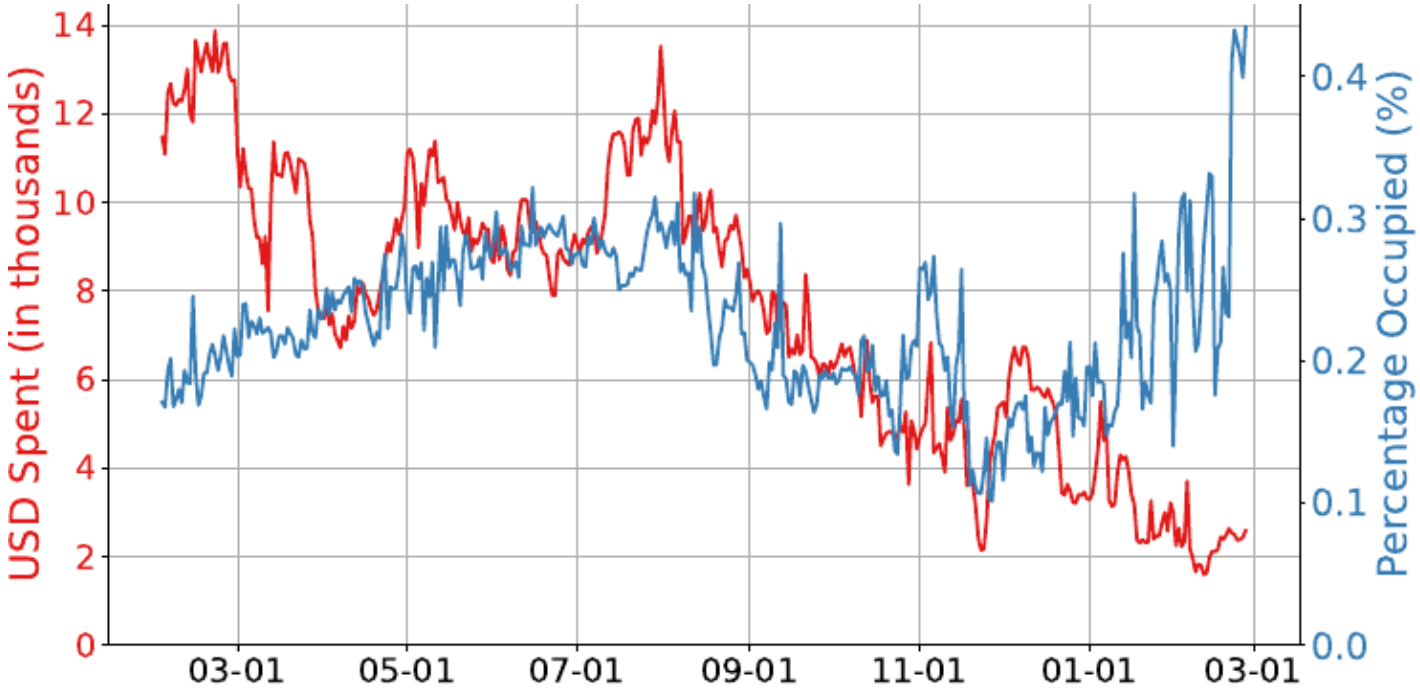
Map provider locations



Most measured hosts are in North America and Europe.

Decentralized compute markets

Measure market activity



Akash averaged 0.30% occupied CPU; Golem activity remained negligible.

RAG case study

One pipeline, two compute stages

STAGE 1

Build the index

Chunk → embed → store

Batch

FIXED

Vector service

Storage + retrieval

STAGE 2

Answer a query

Retrieve → infer → return

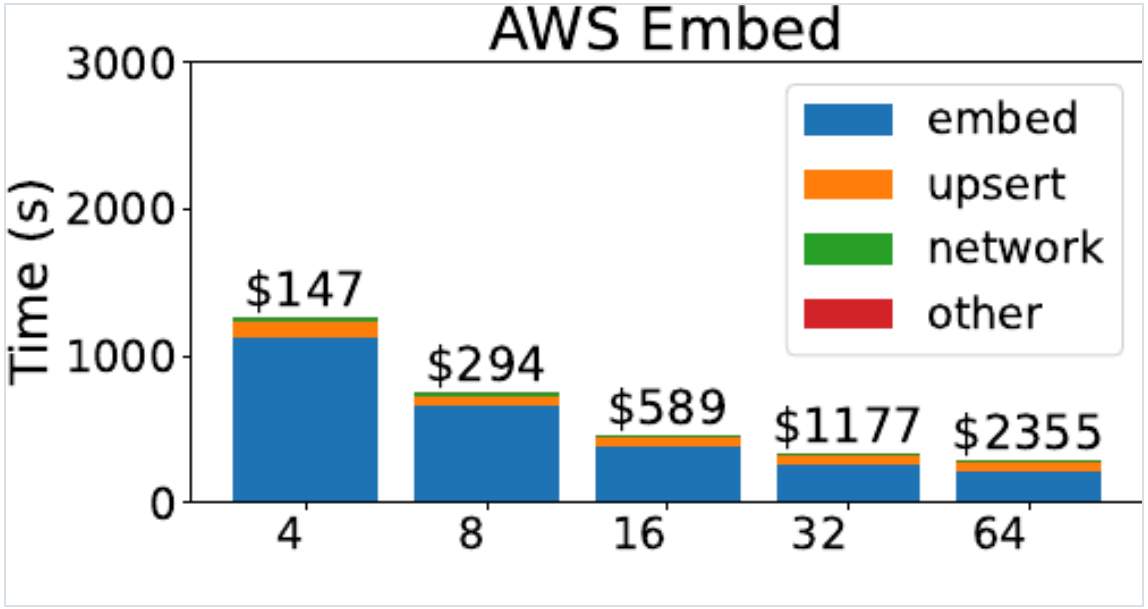
Online

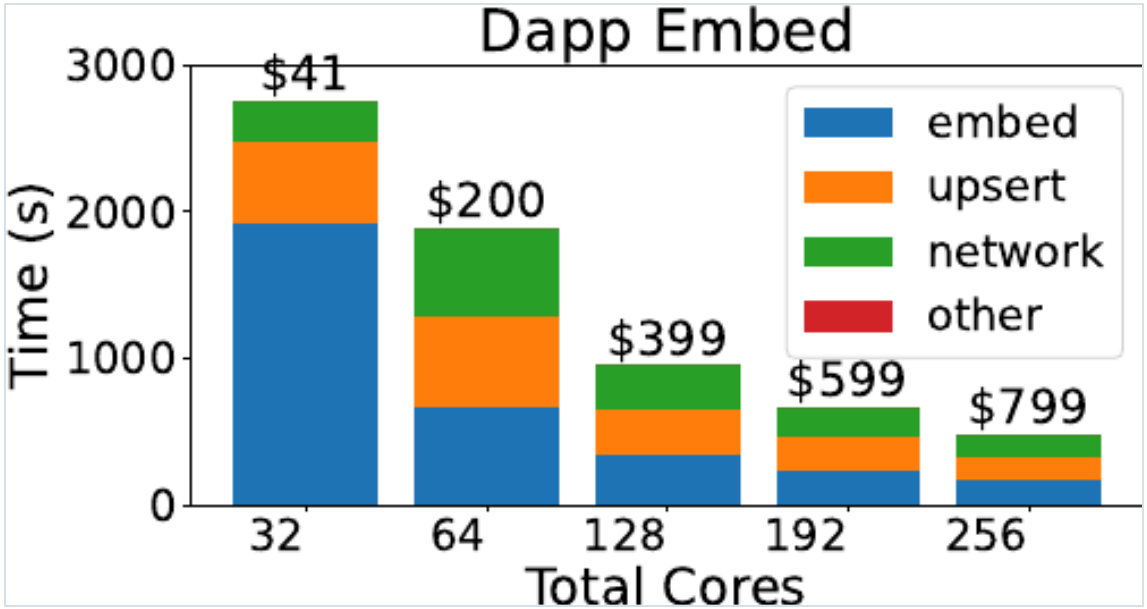
Embedding and generation move between AWS and dApp; one centralized vector service stays fixed.

RAG case study - Stage 1

Compare AWS and dApp embedding

Chunk → compute embeddings → store





AWS 64 cores: ~270 s · $2,355/mo | dApp 256 cores: ~480 s · $799/mo

RAG case study - Stage 2

Compare cost and delivered latency

Retrieve → run inference → return

Compute dominates

Heavier inference makes network a smaller share.

Provider variance

Up to 4x runtime at the same price.

Future work

Toward GPU-ready, efficient compute markets

1 **GPU deployments** Evaluate GPU-backed deployments as viable options emerge.

2 **Machine-aware tuning** Tune placement and configuration across machine types.

3 **Economic modeling** Model price, demand, user cost, and provider income.

Takeaways

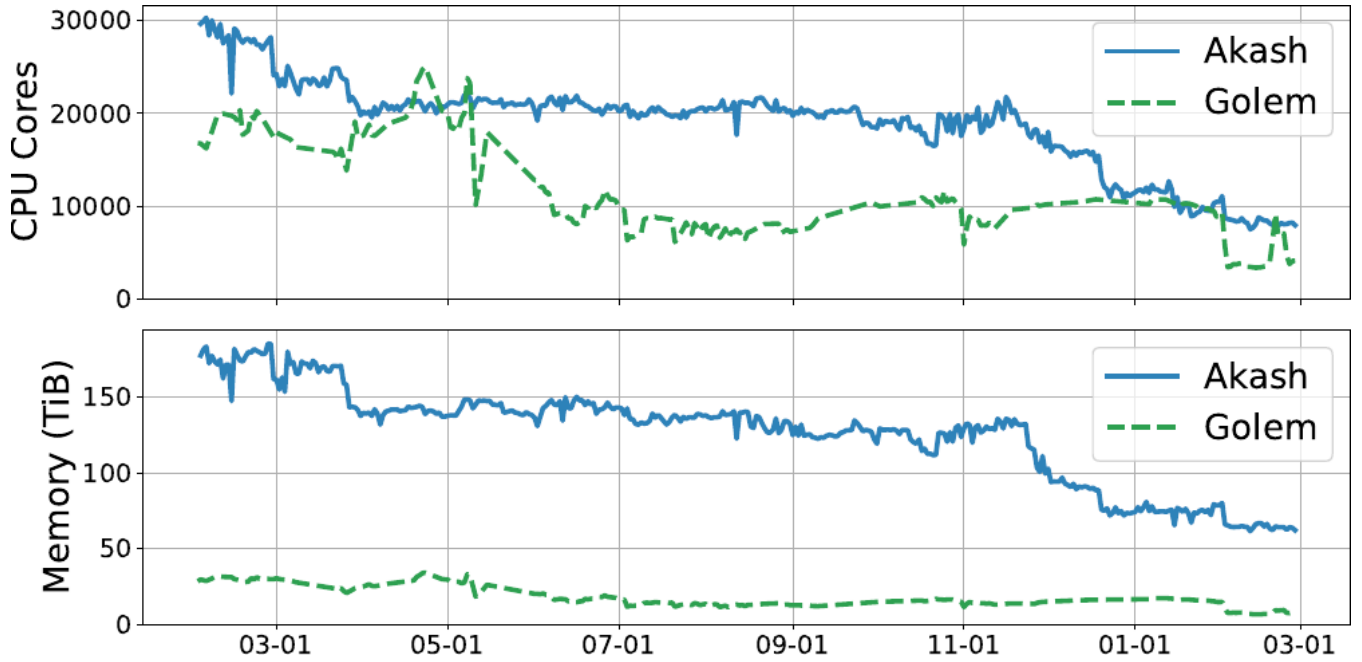
1. Abundant capacity, limited activity.
2. Lower cost at comparable RAG performance.
3. Provider variance: careful choice required.

Questions?

Decentralized compute markets

Backup · Track available capacity

Nov. 2025 update Outdated Akash nodes gradually left the network.



Capacity fell after the update; remaining nodes absorbed the load.