SCALITY

SCALITY ADI

SOLUTION BRIEF

Scality AI Inference Factory

Open-weight models on your own infrastructure, at a cost you set

Scality AI Inference Factory gives enterprises, government agencies and neo-cloud providers an open-code stack for running open-weight AI models on infrastructure they own. Built on Scality ADI (AI Data Infrastructure), it is validated, shipped and maintained by Scality. It is available as a software license or, for sites that can be managed remotely, as a fully managed service.

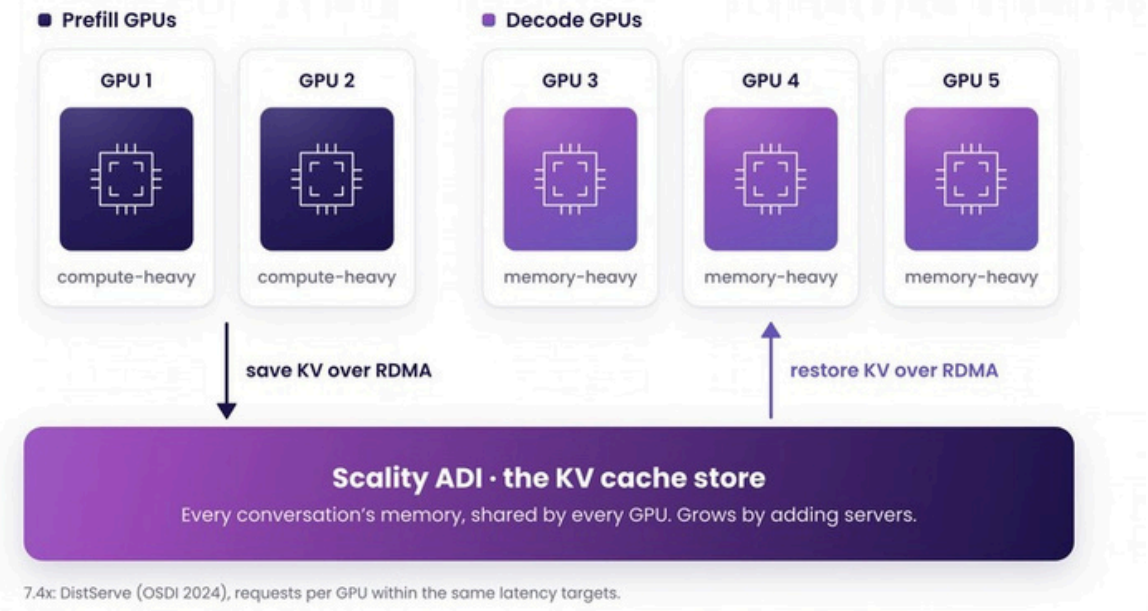
Agents re-read the same context all day

In measured agentic workloads, 84.6 to 99.5 percent of input tokens are reused from one turn to the next. The key-value (KV) cache that holds that context outgrows GPU memory fast, and whatever does not fit is recomputed on the GPU.

> **One KV cache every GPU can read: Don't rebuild the context. Read it back.**
> Scality AI Inference Factory keeps every conversation's KV cache on Scality ADI, so any GPU in your cluster can read it back instead of recomputing it.

Split serving: up to 7.4x more requests

Prefill and decode split over one shared KV cache; on Scality ADI, that cache has no size limit.



7.4x: DistServe (OSDI 2024), requests per GPU within the same latency targets.

14x to 72x
faster to read the context back than to recompute it
14x to 18x at 14K tokens, 72x at 439K tokens

8 TB
of KV cache held on Scality ADI
More than 80x one GPU's memory, with first-token time flat as it grew

97%
of network line rate with the Scality AI Connector
Between GPU memory and Scality ADI, over RDMA

The inference stack is new, and the storage underneath it holds more than 12 EB in production across 70+ countries and has run continuously since 2009.

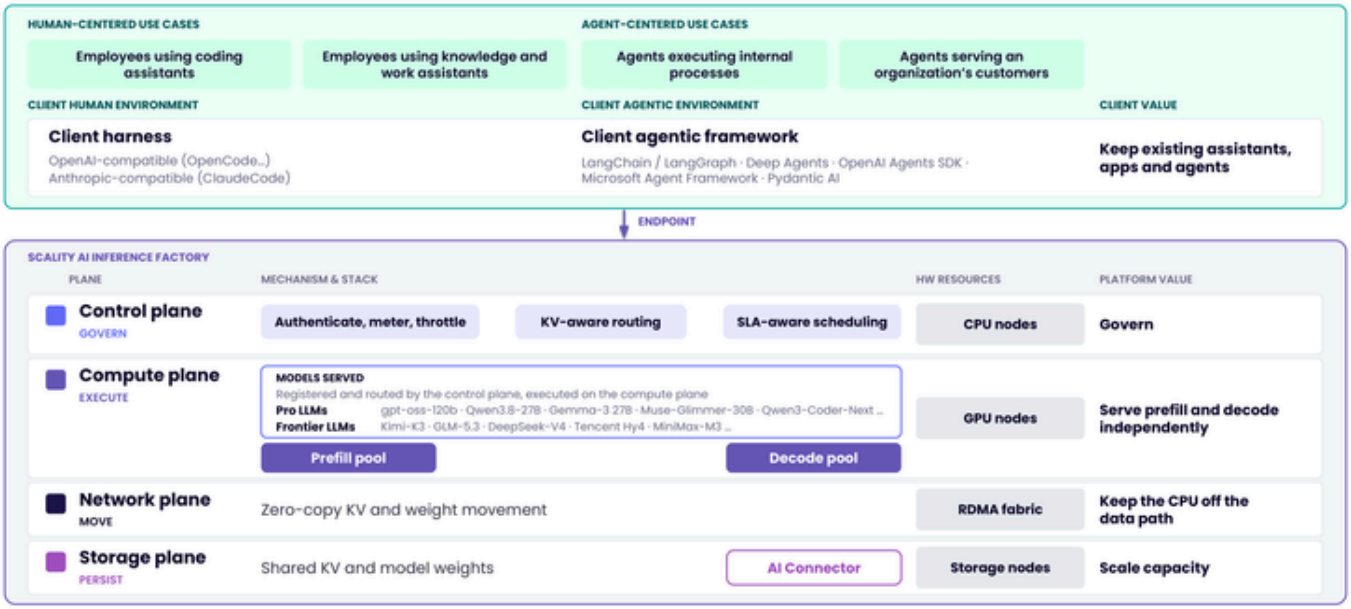
Source: Y. Yuan et al., "Agentic AI Workload Characteristics", arXiv 2605.26297, 2026. Lab figures: Scality lab, August 2026, one server with eight GPUs, five storage servers a few generations old and four 100 Gbit/s RDMA links. Network was the limit in the lab; product runs on an 800 Gbit/s fabric. Results depend on model, context length and traffic.

© 2026 Scality · v1 · 2026-10-07

> **"As AI adoption expands and AI agents become more prevalent, organizations are evaluating which IT environments are best suited to run inference. On-premises deployments are coming into focus for a variety of reasons, including cost optimization, data security and privacy, and regulatory requirements.**
>
> **IDC research shows that running inference on-premises is a credible option for a growing set of enterprise and public sector workloads, provided the infrastructure can hold and serve model state efficiently at scale. Scality is among the vendors addressing that requirement across the inference and storage layers."**
>
> Natalya Yezhkova, vice president, storage and data management, enterprise infrastructure, IDC

What you deploy: How it works

One endpoint over the whole serving stack



Models are served by vLLM, with no Scality fork, and we keep every layer current as models and serving engines change. The stack runs validated open-weight families, including Mistral, Gemma, gpt-oss, Qwen, Kimi, GLM and DeepSeek, on standard servers from Dell, HPE, Lenovo and Supermicro. On SWE-rebench's fresh coding tasks, GLM-5.2 resolves 62.9% against 64.5% for the best closed model, and other open-weight models score 15 to 25 points lower, so the model is chosen with you during sizing.

Source: SWE-rebench (Nebius), 111 fresh tasks from 65 repositories, 2026-05-15 to 2026-07-01, read 2026-09-29.

Sovereignty is no longer only about data

Running inference on infrastructure you own means owning three things.

DATA SOVEREIGNTY

Your context stays on infrastructure you own, under your jurisdiction.

Every request is authenticated by the control plane.

COST SOVEREIGNTY

You set what each request costs, and so your margin.

You meter each request and schedule it against your own SLA targets.

INFERENCE SOVEREIGNTY

The model, and how it is served, change only when you decide.

Version and quantization stay as you validated them.

Start from one of the four use cases above. A Scality AI architect sizes the GPUs, the fabric and the cache against your traffic, and your team compares output quality on its own workload before anyone switches.

SCALITY
RELIABLE. SECURE. SUSTAINABLE.

Size it against your own traffic and code.

Talk to a Scality AI architect about your workload · scality.com

© 2026 Scality · v1 · 2026-10-07