



| **SCALITYDAY26**

Scality ADI for AI: Vision, Roadmap & Demo



Erwan Girard
CPO
Scality



Baptiste Caruana
Product Manager RING & ADI
Scality



SCALITY

RELIABLE. SECURE. SUSTAINABLE.

LATEST RELEASE

Scality ADI

Built for GPU hunger, deep cold,
and everything in between



AUTONOMOUS DATA INFRASTRUCTURE

Vision

Complete end-to-end storage infrastructure
Exa-scale ready, cross-temperature, and sustainable



Outcome / SLA based pricing



Guardian Intelligent Operational Layer



SCALITY
RING

DISAGGREGATED
OBJECT ENGINE



S3

OBJECT ACCESS



GPU Direct

AI THROUGHPUT



File

ENTERPRISE WORKFLOWS



ScalityOS



Infrastructure layer across media classes

Four founding design choices

Software only

- Standard x86 servers, NICs, and drives.
- No proprietary controller or appliance.

Infinite Scalability

- Scale in storage and performance
- Scale in accounts, users, policies

No single point of failure

- Any connector, node, drive, or server can fail.
- Copies land in separate failure domains.

Self-healing

- Repair runs continuously, without operator action.
- Protection returns to its configured level.

SCALITY ADI PERFORMANCE

The right tier for every workload



EXTREME

GPU-Direct

USE CASE
KV cache, model weights

SOFTWARE
Alconnector

MEDIA
TLC

ARCHITECTURE
Aggregated, single-site

THROUGHPUT
97 GB/s / U (97% line rate)

HOT

Performance + resilience

USE CASE
Fast backup, demanding analytics

SOFTWARE
**Standard S3 + RDMA fabric
S3 over RDMA**

MEDIA
QLC / NL-SSD

ARCHITECTURE
**Aggregated or
disaggregated, single site**

THROUGHPUT
Multi-TB/s

WARM

Performance + economics

USE CASE
Sovereign cloud, backup at scale

SOFTWARE
**Standard S3 + RDMA fabric
S3 over RDMA**

MEDIA
NL-SSD / NL-HDD / HDD

ARCHITECTURE
**Disaggregated,
geo-distributed**

PROFILE
Ultra-scale, geo-resilient

COLD

Maximum economics

USE CASE
Long-term archive and retention

SOFTWARE
XDM lifecycle sidecar

MEDIA
Tape and cloud targets

ARCHITECTURE
Policy-driven tiering

PROFILE
Capacity-optimized

One platform. Four tiers. One operational model. One S3-compatible namespace, with policy-driven lifecycle across all tiers. No new silos.



Three connector families, five protocols

S3

- AWS S3 API with IAM, versioning, lifecycle, Object Lock.
- RDMA data path through NVIDIA cuObject.

Native REST

- Key-based REST API (sproxyd), no bucket or IAM lookup.
- The S3 connector stores its data through it.

NFS

- NFSv3 and NFSv4.0 through NFS-Ganesha.
- NFSv3 through sfused, with Kerberos.

SMB

- Samba up to SMB3, with a RING VFS module.
- CTDB moves public IPs on failure.

FUSE

- A POSIX mount through sfused.
- Samba shares sit on this mount.

Where erasure coding 8+4 reaches 14 nines

Each column: drive MTBF and repair time after a drive fails

Class of service

and placement

MTBF 1,000,000 h

repair 48 h

0.9% of drives fail a year

MTBF 1,000,000 h

repair 24 h

0.9% of drives fail a year

MTBF 1,500,000 h

repair 24 h

0.6% of drives fail a year

MTBF 2,000,000 h

repair 24 h

0.4% of drives fail a year

3 copies

CoS 2, on 3 servers

9.1 nines

MTTDL 21,000 years

9.5 nines

MTTDL 56,000 years

9.9 nines

MTTDL 150,000 years

10.2 nines

MTTDL 290,000 years

Erasure coding 9+3

12 servers, 1 fragment each

9.8 nines

MTTDL 21,000 years

10.5 nines

MTTDL 88,000 years

11.0 nines

MTTDL 310,000 years

11.4 nines

MTTDL 740,000 years

Erasure coding 8+4

12 servers, 1 fragment each

13.1 nines

MTTDL 1.6 million years

95%: 13.0 to 13.2, 192 events

13.9 nines

MTTDL 9.1 million years

95%: 13.8 to 13.9, 110 events

14.5 nines

MTTDL 40 million years

95%: 14.3 to 14.8, 15 events

15.2 nines

MTTDL 200 million years

95%: 14.7 to 15.9, 3 events

Model output, not a measurement. Engine: SprovGX, the in-house GPU reimplementation of sprov's data-safety model; 300 million to 1 billion simulated ring-years per column.

Drive failures only, Weibull shape 0.7 (failures cluster); latent sector errors off; 1 MB mean object, 70% full. No server, site, or batch-correlated failures.

Nines: $-\log_{10}(\text{user objects lost per year} / \text{user objects stored})$. Rebuild rate: 14 TB (a 20 TB drive, 70% full) over the repair time. Failures/yr = $1 - \exp(-8,760 \text{ h} / \text{MTBF})$.

Modeled annual durability per object and mean time to data loss (MTTDL). Shipped sprov keyspace for 12 servers, each 24 x 20 TB drives (5.8 PB raw). Columns: drive MTBF and repair time.

Across sites, erasure coding passes 16 nines

Each column: drive MTBF and repair time after a drive fails

Sites and protection	Overhead	Survives declared in sprov	MTBF 1,000,000 h repair 24 h, 0.9% of drives a year	MTBF 600,000 h repair 48 h, 1.5% of drives a year
1 site, 12 servers Erasure coding 8+4	+50%	3 servers	13.9 nines 95%: 13.8 to 13.9	12.2 nines 95%: 12.2 to 12.3
3 sites, stretched Erasure coding 7+5	+71%	3 servers, or 1 site + 1 server	16.5+ nines no loss in 600M ring-years	15.2 nines 95%: 15.0 to 15.4
2 sites, stretched Erasure coding 5+7	+140%	3 servers, or 1 site + 1 server	17.5+ nines no loss in 600M ring-years	17.0+ nines no loss in 200M ring-years
2 sites, stretched 4 copies	+300%	3 servers, or 1 site + 1 server	13.3 nines 95%: 13.1 to 13.4	11.8 nines 95%: 11.8 to 11.9

Model output, drive failures only (Weibull shape 0.7, latent sector errors off). A whole-site outage is covered by the declared tolerance, not by these nines.

Two ADI systems with asynchronous replication are not modeled: data is lost only if both systems lose the same object before the replication lag closes.

Modeled annual durability per object, shipped sprov keyspaces for 12 servers of 24 x 20 TB (5.8 PB raw). "+" marks a 95% lower bound from zero loss events.

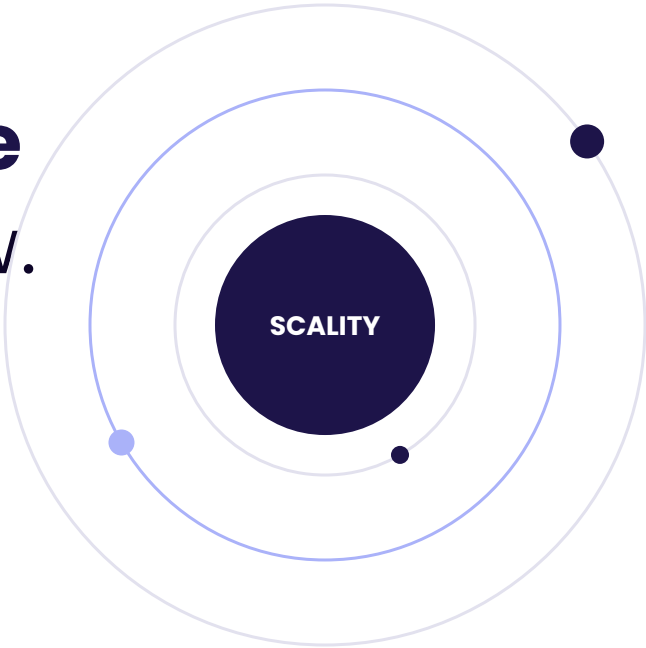
CUSTOMER VOICE



Scality is the **most reliable system** [we run] right now.

INFRASTRUCTURE LEADERSHIP

Leading commercial space-flight company





Conversation Dashboard

+ Start new conversation

Search...

Untitled conversation



Good morning, Erwan Girard.

Learn more about Scality products, anticipate capacity-related issues, and plan corrective actions through predictive models and anomaly detection.

Select Agent...

Select deployment - 1552 avail. ↑

Type your question or request...



The AI-generated and may contain errors, verify before acting. Your conversations are not used to train our models.

Analytics

Metrics, vitals, alerts and projections for a deployment: health, capacity and performance read from its live monitoring data.

Documentation

Answers about Scality products (RING, ARTESCA, ADI, S3C, Maestro) from product documentation, TSKB articles and release notes.

Troubleshooting

Answers from resolved Scality support cases: closed RING and ARTESCA tickets summarised into problem, root cause and resolution steps, each precedent cited by ticket number.

OPAL Agent

Internal CS agent that answers from the OPAL RAG corpus for Scality RING and Artesca products.



Conversation Dashboard

ARTESCA

ARTESCA Overview

RING / ADI

RING Home Dashboard

RING Monitoring Service

RING Platform System

RING S3 Connectors

RING SOFS Volumes

RING Storage Service

Customer [redacted] [icon]

Deployment [redacted] [icon]

RING Home Dashboard

✓ ApplyExport

Period Last week [icon]

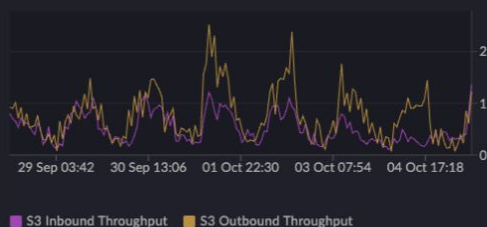
Warning Alerts

258

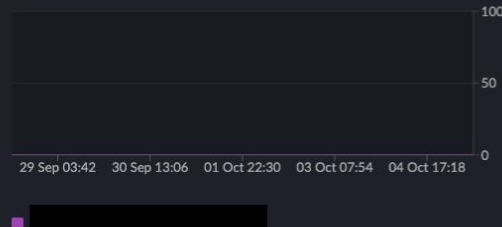
Critical Alerts

0

S3 Throughput — in vs out (GiB/s)



S3 5xx Error Rate (%)



high count worth investigating, though these are platform-level warnings (e.g. disk usage or low-storage forecasts), not necessarily service outages.

- **Critical Alerts:** 0 — no critical native alerts currently active.

S3 Throughput [\(chart\)](#)

- Inbound traffic ran roughly 180–500 MB/s for most of the window, then spiked sharply at the end: **920 MB/s at 2026-10-05 09:00** and **1.46 GB/s at 2026-10-05 10:00** (about 1 hour before window end) — a clear, large upload surge.
- Outbound traffic fluctuated between about 90 MB/s and 1.55 GB/s, with a notable peak of 1.55 GB/s at 2026-10-04 18:00 and another rise to 1.3 GB/s at 2026-10-05 10:00, mirroring the inbound spike. Worth checking what workload drove this late surge.

S3 Error Rate [\(chart\)](#)

- Essentially flat at 0%, with a few negligible blips (~0.0001%) at 14:00, 19:00, 22:00, and 23:00 on 2026-10-04. Not significant — S3 service health looks clean, consistent with the [S3 Services status heatmap](#) showing "OK" across the entire window.



ISSUE

S3 throughput collapsed 0 MiB/s (base 144)

POSSIBLE CAUSE

not yet identified — see full signal list

Opal check **Opal: P1** **disagrees with P2G** **Accept Opal's P1** asked 2026-10-05 07:05

S3 throughput at 0 MiB/s for 24h means no object traffic is being served.

MECHANISM — no single mechanism — could be a real S3 outage or a monitoring/metrics-path break reporting 0 while S3 is actually up.

CHECK FIRST (ON THE PLATFORM) — Supervisor GUI "Storage Services" widget + S3 healthcheck ('ansible -m uri -a 'url=http://127.0.0.1:8000/_/healthcheck' runners_s3').

► Full Opal answer (diagnostic procedure + sources)

Platform history Customer360 · 730d

No major incident or service change recorded on this account in the last 730 days — this looks like a first occurrence.

History arbiter **UPGRADE P1** **Accept upgrade's P1** asked 2026-10-05 07:06

► Full answer

[Detail](#)

ISSUE

ring latency tail 16ms (>7)

POSSIBLE CAUSE

not yet identified — see full signal list

Opal check **Opal: P2** asked 2026-10-05 07:05

ring still serving but tail latency sustained 24h above threshold; degraded, not down.

MECHANISM — no single mechanism confirmed — likely a slow/outlier disk or server triggering read reconstruction (Read Latency Predictability), but connector-side RPC saturation cannot be excluded without live data.

CHECK FIRST (ON THE PLATFORM) — RING Administration Interface → Performance tab (Latency chart) for the DATA ring, then 'ringsh supervisor ringStatus DATA'.

► Full Opal answer (diagnostic procedure + sources)

Platform history Customer360 · 730d

0 major incidents · 1 major service change

History arbiter **UPGRADE P1** **Accept upgrade's P1** asked 2026-10-05 07:06

► Full answer

[Detail](#)



Who we work with, by use case

Backup & data protection



Compliance archiving & records



Analytics & AI data platforms



Email & messaging



Virtualization & private cloud



Open-source data & AI frameworks



File & unstructured data



Hammerspace

Video streaming & delivery



Data security & encryption



Medical imaging (PACS, VNA)



Media asset mgmt & production



Observability, logging & SIEM



Video management systems



Customer engagement



SCALITY ADI PERFORMANCE

The right tier for every workload



EXTREME

GPU-Direct

USE CASE
KV cache, model weights

SOFTWARE
Alconnector

MEDIA
TLC

ARCHITECTURE
Aggregated, single-site

THROUGHPUT
97 GB/s / U (97% line rate)

HOT

Performance + resilience

USE CASE
Fast backup, demanding analytics

SOFTWARE
**Standard S3 + RDMA fabric
S3 over RDMA**

MEDIA
QLC / NL-SSD

ARCHITECTURE
**Aggregated or
disaggregated, single site**

THROUGHPUT
Multi-TB/s

WARM

Performance + economics

USE CASE
Sovereign cloud, backup at scale

SOFTWARE
**Standard S3 + RDMA fabric
S3 over RDMA**

MEDIA
NL-SSD / NL-HDD / HDD

ARCHITECTURE
**Disaggregated,
geo-distributed**

PROFILE
Ultra-scale, geo-resilient

COLD

Maximum economics

USE CASE
Long-term archive and retention

SOFTWARE
XDM lifecycle sidecar

MEDIA
Tape and cloud targets

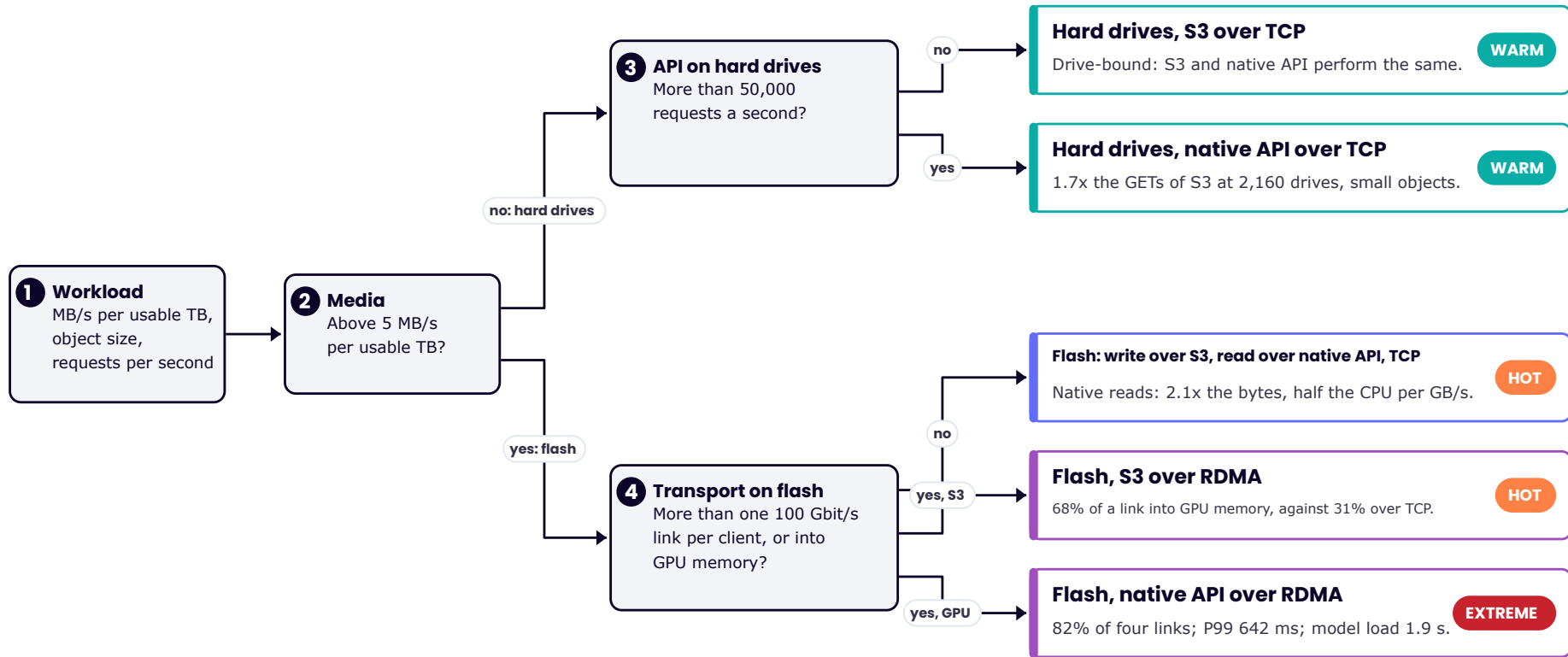
ARCHITECTURE
Policy-driven tiering

PROFILE
Capacity-optimized

One platform. Four tiers. One operational model. One S3-compatible namespace, with policy-driven lifecycle across all tiers. No new silos.



Four questions select the configuration



One S3 metadata cluster (50K operations/s). HDD at 40 IOPS per drive for sizing. Flash results: Comcast campaign, 2 GiB objects.

Exascale sustainable by design

POWER SUSTAINABILITY

Workload-aligned tiering delivers up to 20x power efficiency overall

GPU-Direct: highest draw per TB, full power for AI and HPC workloads

Hot: full draw, 5% of your data estate

Warm: up to 1/25th the draw per TB, 30% of your data estate

Cold: near-zero at rest, 65% of your data estate

OPERATIONS SUSTAINABILITY

Scale without adding operational complexity

One cross-temperature lifecycle, no silos, no separate policies

Exabyte-scale failure domains keep resilience built in

Appliance-like experience with non-event upgrades means no extra planning, time, or manpower needed

COST SUSTAINABILITY

The right economics at every tier, not averaged across all of them

Object storage economics apply everywhere, with data reduction.

Performance TCO is optimized per workload

One platform means one procurement motion and no infrastructure constraints that force you to overbuy



ADI

Available for all

including the one line migration from existing RINGs

New Customers

Available now

New RINGs

Available now for new projects
at Scalify current customers

Existing RINGs

Q1 2027 to migrate RINGs to
Scalify ADI in without forklift
upgrade or service interruption



ADI

A constantly evolving product

We never stop improving - 3 examples

Scality Guardian

A fully integrated customer experience.

Continuous Data Protection

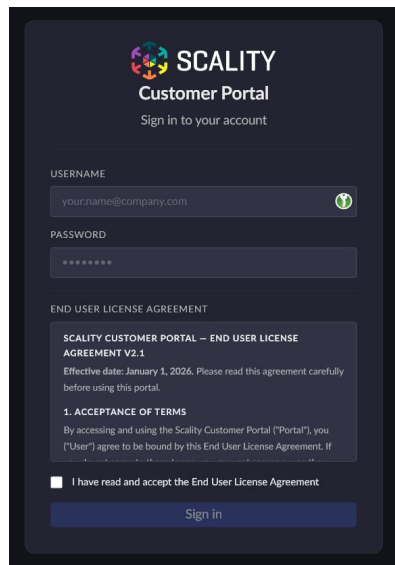
ADI ability to recover any point in time data state

AI

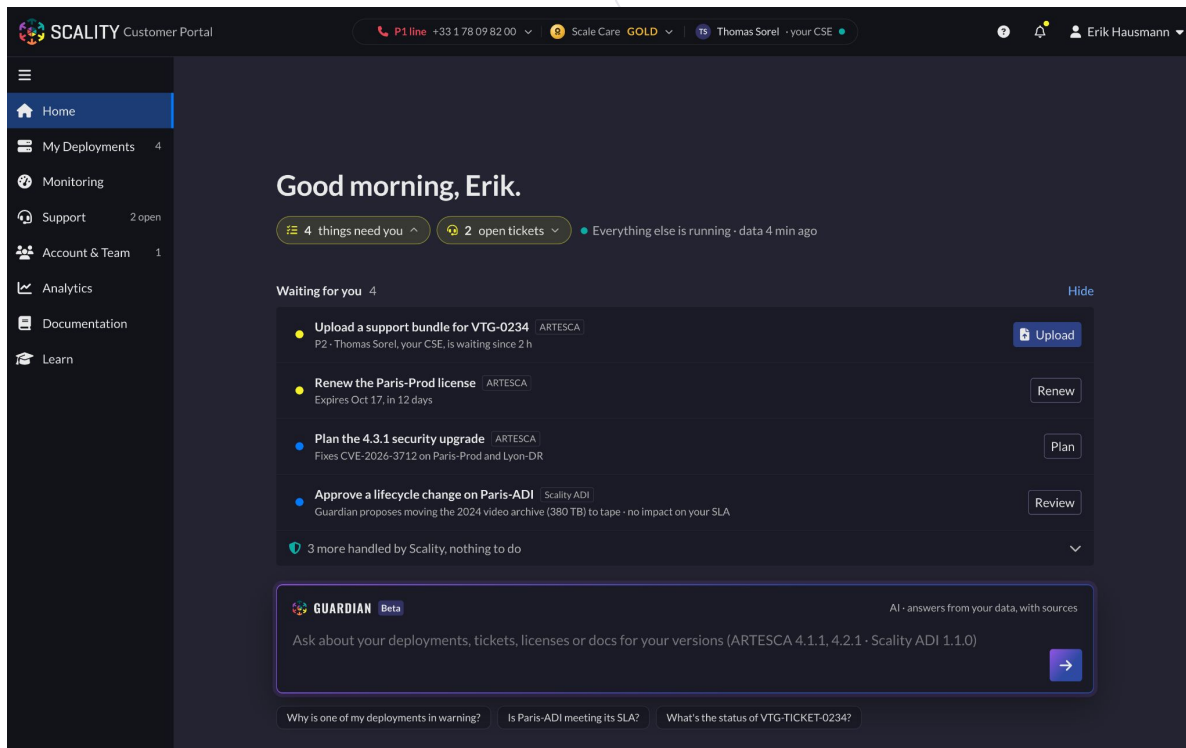
AI Datalakes and Inference Factory



Everything, one login, one place.



The login screen for the Scality Customer Portal. It features the Scality logo and the text "Customer Portal" and "Sign in to your account". There are input fields for "USERNAME" (with a placeholder "your.name@company.com") and "PASSWORD" (with a masked field "*****"). Below these is a section for the "END USER LICENSE AGREEMENT" which includes the "SCALITY CUSTOMER PORTAL – END USER LICENSE AGREEMENT V2.1" and a "1. ACCEPTANCE OF TERMS" section. A checkbox labeled "I have read and accept the End User License Agreement" is present, followed by a "Sign in" button.

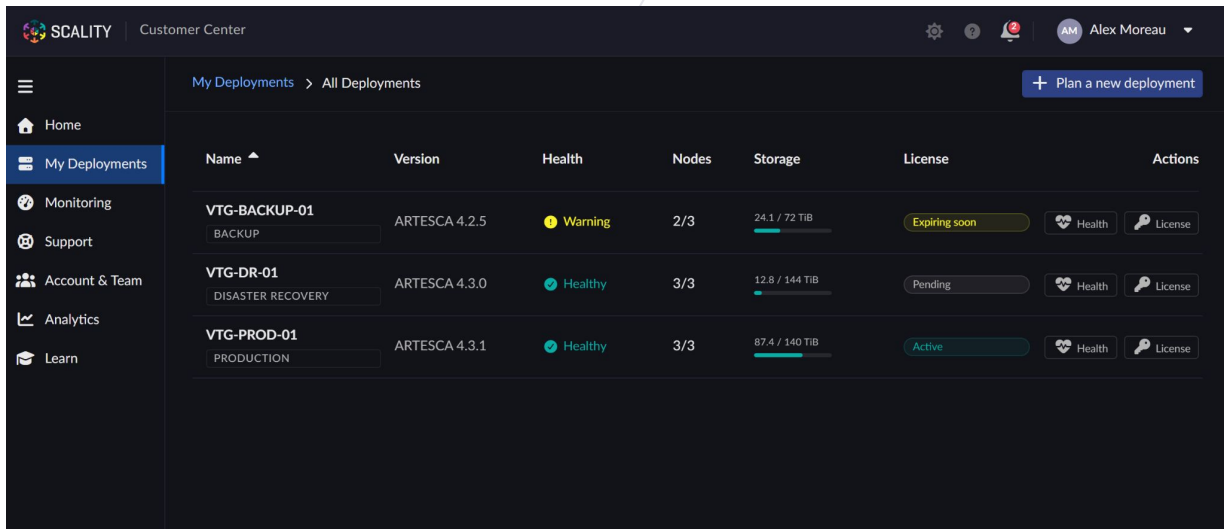


The dashboard for the Scality Customer Portal. It features a dark theme with a sidebar on the left containing navigation links: Home, My Deployments (4), Monitoring, Support (2 open), Account & Team (1), Analytics, Documentation, and Learn. The main content area displays a greeting "Good morning, Erik." followed by a summary of "4 things need you" and "2 open tickets". Below this is a section "Waiting for you" with four items: "Upload a support bundle for VTG-0234" (ARTESCA), "Renew the Paris-Prod license" (ARTESCA), "Plan the 4.3.1 security upgrade" (ARTESCA), and "Approve a lifecycle change on Paris-ADI" (Scality ADI). Each item has a corresponding button (Upload, Renew, Plan, Review). At the bottom, there is a "GUARDIAN Beta" section with a text input field for asking questions about deployments, tickets, licenses, or docs, and a "→" button. The top of the dashboard shows a header with the Scality logo, "Customer Portal", and user information: "P1 line +33 1 78 09 82 00", "Scale Care GOLD", "TS Thomas Sorel · your CSE", and "Erik Hausmann".

OPERATE

See all your deployments

- Every ARTESCA and ADI deployment you own.
- Quick overview:
 - Platform health
 - Storage consumption
 - License status



The screenshot shows the SCALITY Customer Center interface. The top navigation bar includes the SCALITY logo, 'Customer Center', and user information 'AM Alex Moreau'. A sidebar on the left contains navigation links: Home, My Deployments (selected), Monitoring, Support, Account & Team, Analytics, and Learn. The main content area is titled 'My Deployments > All Deployments' and features a '+ Plan a new deployment' button. Below this is a table with columns: Name, Version, Health, Nodes, Storage, License, and Actions. The table lists three deployments: VTG-BACKUP-01 (Warning), VTG-DR-01 (Healthy), and VTG-PROD-01 (Healthy). Each row includes a progress bar for storage and a license status indicator.

Name	Version	Health	Nodes	Storage	License	Actions
VTG-BACKUP-01 BACKUP	ARTESCA 4.2.5	Warning	2/3	24.1 / 72 TiB	Expiring soon	Health License
VTG-DR-01 DISASTER RECOVERY	ARTESCA 4.3.0	Healthy	3/3	12.8 / 144 TiB	Pending	Health License
VTG-PROD-01 PRODUCTION	ARTESCA 4.3.1	Healthy	3/3	87.4 / 140 TiB	Active	Health License

OPERATE

Monitor each deployment

Drill into a deployment to see its health and the status of each service.



GET HELP

Ask Guardian

- Guardian helps you describe the issue and opens a support ticket with the right context.
- Support can answer faster, with less back-and-forth.

The screenshot displays the Scalify Customer Portal interface. At the top, the header includes the Scalify logo, 'Customer Portal', and user information for Thomas Sorel. A navigation sidebar on the left contains icons for home, settings, and other portal features. The main content area is titled 'GUARDIAN' and provides a search bar for asking questions about ARTESCA and Scalify ADI deployments. Below this, a 'What's new' section lists recent updates, including guides for lifecycle policies, upgrade guides for ARTESCA, and new preview versions. A 'Quick overview and actions' sidebar on the right shows a summary of tickets (4 Deployments, 3 Licenses, 4 Team) and offers quick links to open a ticket, download resources, and renew licenses.

SCALIFY Customer Portal

P1 line +33 1 78 09 82 00 Scale Care GOLD Thomas Sorel - your CSE Erik Hausmann

GUARDIAN Ask about your ARTESCA and Scalify ADI deployments, tickets or docs

What's new (last 30 days)

- Scalify ADI guide: lifecycle policies across storage classes** 2026-10-02
How Guardian moves data between flash, disk and tape on Paris-ADI, and when it asks for your approval.
[Read the guide →](#)
- Upgrade guide: ARTESCA 4.2.1 to 4.3.1** 2026-09-28
Step-by-step guide with pre-checks, expected downtime per node, and rollback plan. Applies to Paris-Prod and Lyon-DR.
[Read the guide →](#)
- ARTESCA 4.4.0-preview.11 available for testing** 2026-09-21
New preview includes Guardian AI integration, Maestro orchestration v2, and automated failover.
[Try preview →](#)
- Webinar replay: Ransomware readiness with Object Lock** 2026-09-17
45 min · How to set up immutable buckets on ARTESCA and run a recovery test without disrupting production.
[Watch the replay →](#)
- ARTESCA 4.3.1 released** 2026-09-14
Security patch for CVE-2026-3712, improved S3 authentication, and 15% read performance boost on HDD profiles.
[View release notes →](#)

Quick overview and actions

Open a ticket →

4 Deployments 1 warning	3 Licenses 1 expiring	4 Team 1 pending
--------------------------------------	------------------------------------	-------------------------------

[Downloads](#) **New** >

[Get / Renew license](#) **New** >

[Book an expert](#) **New** >

[Try ARTESCA 4.4.0](#) >

GET HELP

Book Scality services

- New deployment
- Upgrade

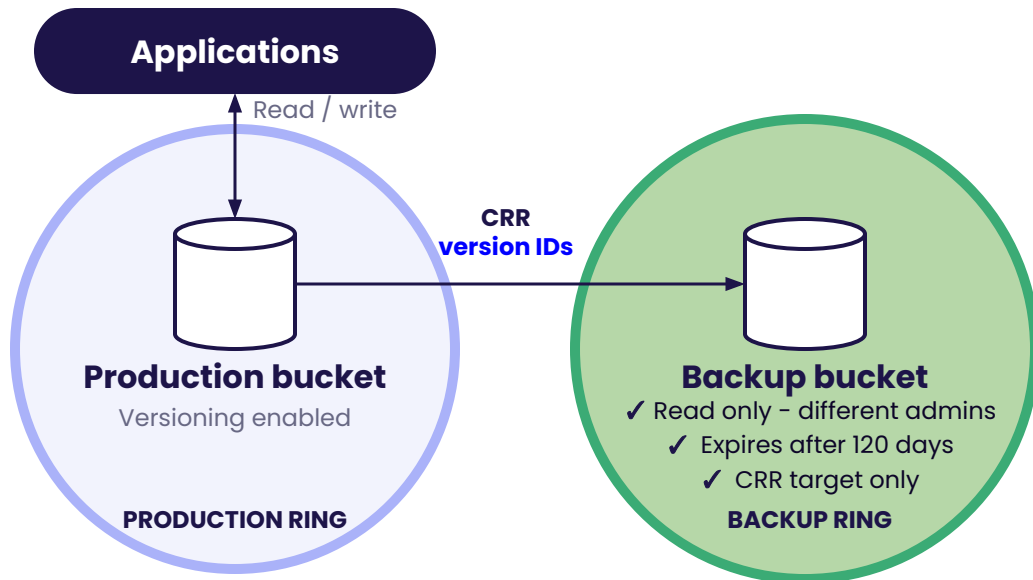
AND MUCH MORE...

Generate your licenses

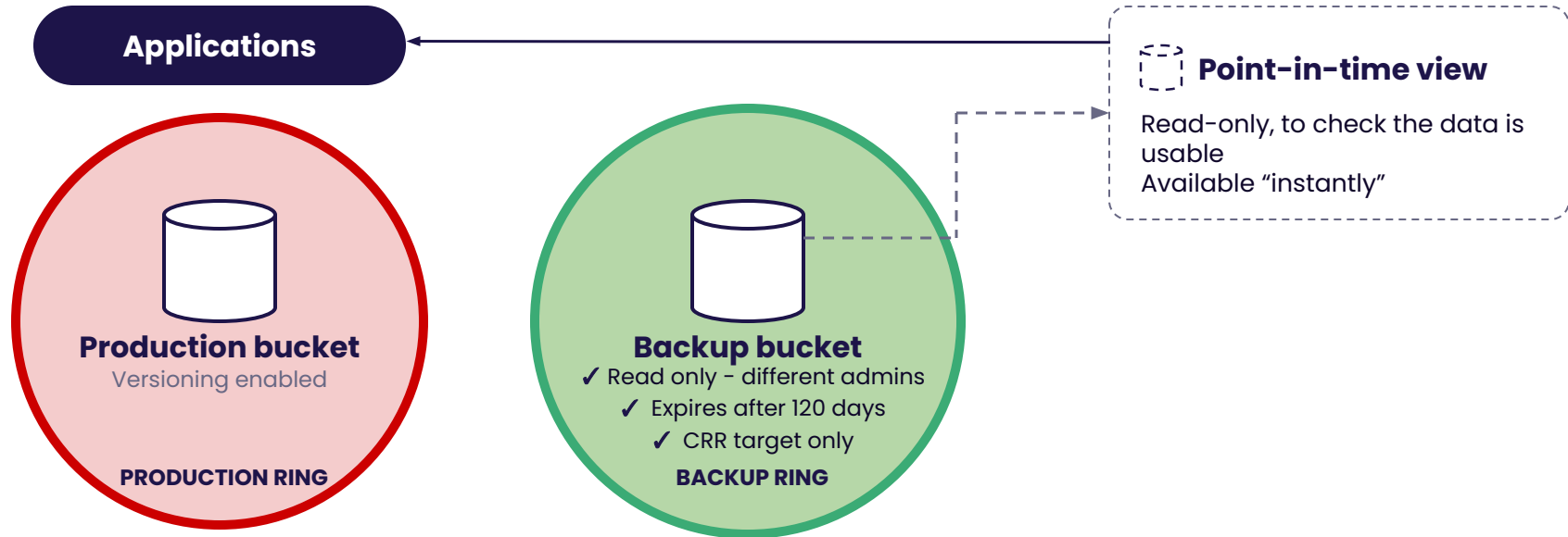
Delegate access

Download packages

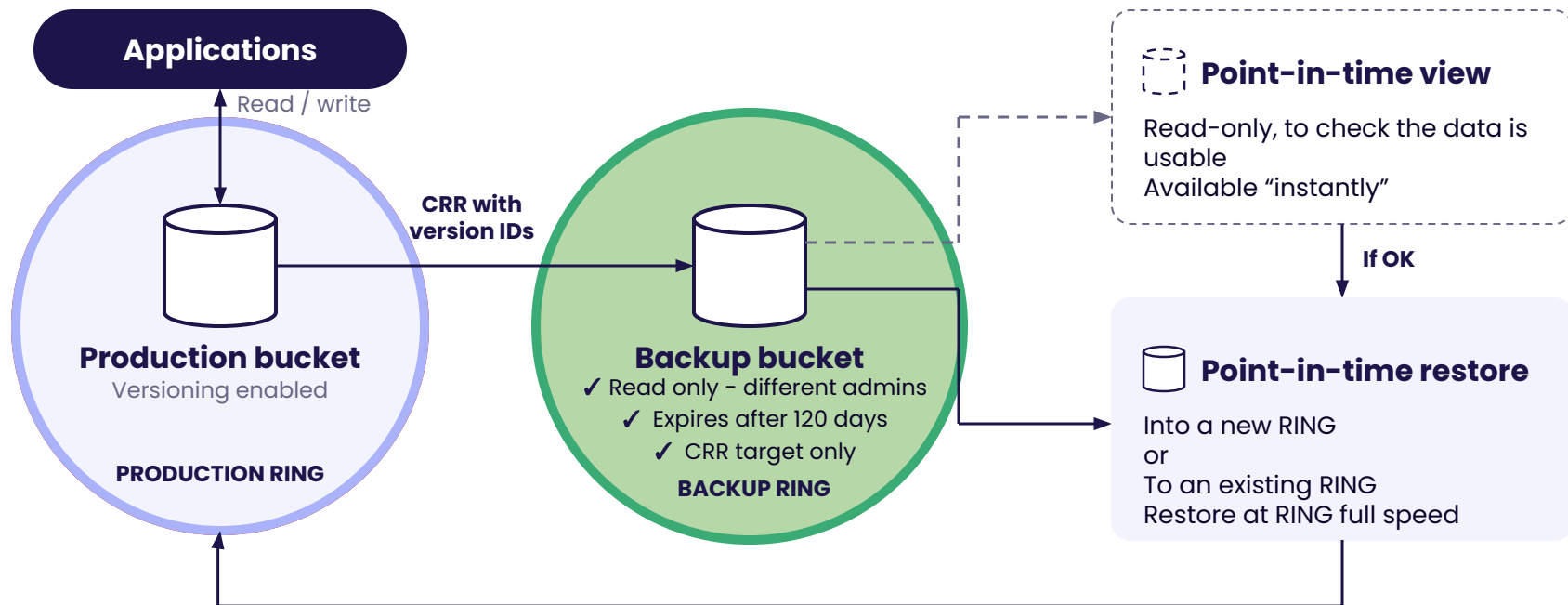
Recover to any point in time at any scale



Recover to any point in time at any scale



Recover to any point in time at any scale



SCALITY ADI PERFORMANCE

The right tier for every workload



EXTREME <i>GPU-Direct</i>	HOT <i>Performance + resilience</i>	WARM <i>Performance + economics</i>	COLD <i>Maximum economics</i>
USE CASE KV cache, model weights	USE CASE Fast backup, demanding analytics	USE CASE Sovereign cloud, backup at scale	USE CASE Long-term archive and retention
SOFTWARE Alconnector	SOFTWARE Standard S3 + RDMA fabric S3 over RDMA	SOFTWARE Standard S3 + RDMA fabric S3 over RDMA	SOFTWARE XDM lifecycle sidecar
MEDIA TLC	MEDIA QLC / NL-SSD	MEDIA NL-SSD / NL-HDD / HDD	MEDIA Tape and cloud targets
ARCHITECTURE Aggregated, single-site	ARCHITECTURE Aggregated or disaggregated, single site	ARCHITECTURE Disaggregated, geo-distributed	ARCHITECTURE Policy-driven tiering
THROUGHPUT 97 GB/s / U (97% line rate)	THROUGHPUT Multi-TB/s	PROFILE Ultra-scale, geo-resilient	PROFILE Capacity-optimized

One platform. Four tiers. One operational model. One S3-compatible namespace, with policy-driven lifecycle across all tiers. No new silos.



Two control planes, one cluster.

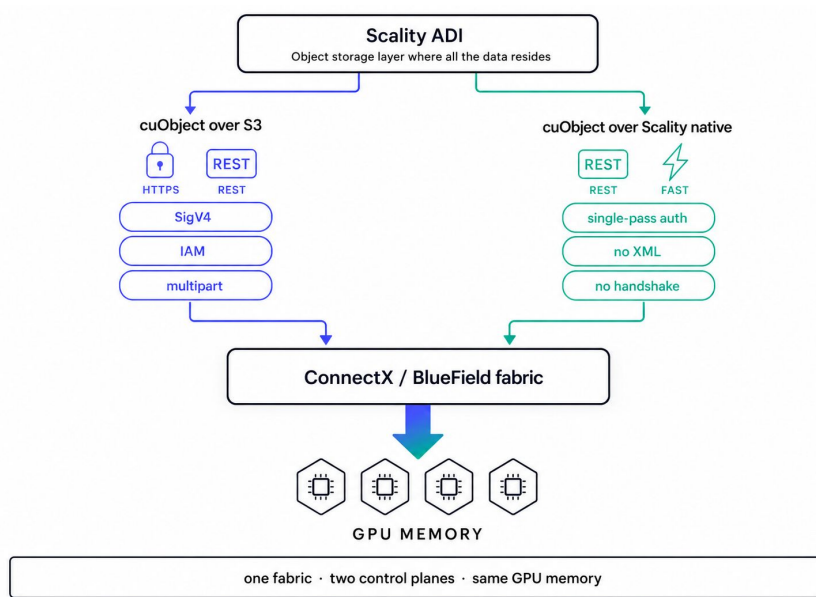
Same fabric. Same GPU memory. The control plane is the choice.

CUOBJECT OVER S3

- Drop-in for any framework that speaks S3.
- Standard SigV4, IAM, multipart.
- Training ingest, bulk reads, multi-vendor tooling.

CUOBJECT OVER SCALITY NATIVE

- Inference, tier-zero training, KV-cache rehydrate.
- Lightweight REST. Single-pass auth, no XML.
- Where microseconds on the control path matter.



SCALITY ADI PERFORMANCE

The right tier for every workload



EXTREME <i>GPU-Direct</i>	HOT <i>Performance + resilience</i>	WARM <i>Performance + economics</i>	COLD <i>Maximum economics</i>
USE CASE KV cache, model weights	USE CASE Fast backup, demanding analytics	USE CASE Sovereign cloud, backup at scale	USE CASE Long-term archive and retention
SOFTWARE Alconnector	SOFTWARE Standard S3 + RDMA fabric S3 over RDMA	SOFTWARE Standard S3 + RDMA fabric S3 over RDMA	SOFTWARE XDM lifecycle sidecar
MEDIA TLC	MEDIA QLC / NL-SSD	MEDIA NL-SSD / NL-HDD / HDD	MEDIA Tape and cloud targets
ARCHITECTURE Aggregated, single-site	ARCHITECTURE Aggregated or disaggregated, single site	ARCHITECTURE Disaggregated, geo-distributed	ARCHITECTURE Policy-driven tiering
THROUGHPUT 97 GB/s / U (97% line rate)	THROUGHPUT Multi-TB/s	PROFILE Ultra-scale, geo-resilient	PROFILE Capacity-optimized

One platform. Four tiers. One operational model. One S3-compatible namespace, with policy-driven lifecycle across all tiers. No new silos.



WHAT ABOUT HDD?

S3 over RDMA stops at flash.

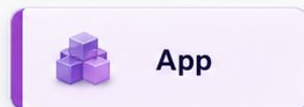
HDDs sit on SAS or SATA. Not PCIe.



Inside an ADI storage server

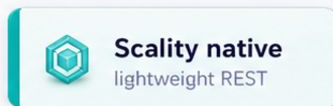
One software stack: connector to storage node to IO Daemon to drive.

CLIENTS



on other servers, across the network

CONNECTOR LAYER

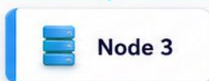
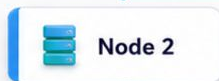


S3 or Scality native, across the network

NETWORK FABRIC (RoCE / TCP)

ONE ADI STORAGE SERVER

STORAGE NODES (6) · Chord P2P



each store node connects to each IO Daemon

IO DAEMONS · 1 process per physical storage device



...

DRIVES · 1 per IO Daemon · HDD or Flash (QLC / TLC)



HDD



Flash · QLC



HDD



Flash · TLC



Flash · QLC

Inside an ADI storage server

With RDMA: the IO Daemon is the endpoint for HDDs, GPUDirect straight from flash.

CLIENTS

RDMA: IO Daemon is the endpoint

HDD-backed: drive isn't on PCIe

App

GPU

on other servers, across the network

GPUDirect: straight from flash

Flash-backed: device is on PCIe

CONNECTOR LAYER

S3 connector

SigV4 / REST

Scality native

lightweight REST

S3 or Scality native, across the network

NETWORK FABRIC (RoCE / TCP)

ONE ADI STORAGE SERVER

STORAGE NODES (6) · Chord P2P

Node 1

Node 2

Node 3

Node 4

Node 5

Node 6

each store node connects to each IO Daemon

IO DAEMONS · 1 process per physical storage device

IOD

IOD

IOD

IOD

IOD

DRIVES · 1 per IO Daemon · HDD or Flash (QLC / TLC)

HDD

Flash · QLC

HDD

Flash · TLC

Flash · QLC



SCALITY

RELIABLE. SECURE. SUSTAINABLE.

Overview

AI Inference Factory

End-to-end stack for on-premise
AI inferencing

BAPTISTE CARUANA

Product Manager

SCALITY



WHERE IT FITS

Three workloads, three sovereignty gaps

The first workloads to bring on-premise

Coding agents for thousands of developers

Developers delegate coding tasks from the IDE or the terminal: agents read, edit and test code across repositories, chaining several model calls per task.

Assistants and agents over internal data

Employees ask questions and delegate tasks over policies, contracts, tickets and knowledge bases, through retrieval and internal tools.

AI inside your own products

Your customers use the AI features you ship: in-app copilots, search, summarization, document extraction, served under your product's SLA.



THE STAKE

Control cost, data and inference

Cost sovereignty · Data sovereignty · Inference sovereignty

COST SOVEREIGNTY

~40% LOWER TCO

Deterministic cost

Fixed and provisioned before usage

API pricing affects adoption

+1% API cost is associated with -0.29% usage

Removing visible per-token cost supports broader adoption — and productivity.

DATA SOVEREIGNTY

External APIs can receive the full working context

Codebase · documents · retrieved data · tool outputs

Moderation can extend retention

Flagged inputs / outputs: up to 2 years
Safety scores: up to 7 years

Jurisdiction remains external

U.S. providers are subject to the CLOUD Act

Legal conflicts are possible

Third-party or NDA-protected IP may not be shareable with external LLMs

INFERENCE SOVEREIGNTY

Model availability remains provider-controlled

Access can be tied to new provider conditions — e.g., Fable 5 retention requirements

Provider-side quantization is not customer-controlled

Load-driven quantization is opaque to the customer

Safety routing can change serving behavior

Fallback policies can change which model serves a request



WHERE IT FITS

Three workloads, three sovereignty gaps

The first workloads to bring on-premise

Coding agents for thousands of developers

Developers delegate coding tasks from the IDE or the terminal: agents read, edit and test code across repositories, chaining several model calls per task.

Assistants and agents over internal data

Employees ask questions and delegate tasks over policies, contracts, tickets and knowledge bases, through retrieval and internal tools.

AI inside your own products

Your customers use the AI features you ship: in-app copilots, search, summarization, document extraction, served under your product's SLA.



DEVELOPER EXPERIENCE

Cloud-grade experience, on-premises

What it takes to move thousands of developers off cloud APIs

Frontier level model quality

"It's not as good as before."

AA CODING INDEX · CODING AGENTS

GPT-6 Astra · cloud



Kimi K3 · open weights



Meta Muse Spark 1.3 · cloud



Serving performance on par at scale

"It's slower than before."

SERVING METRICS · QWEN3.8-27B

~120 tok/s

per developer

~440

137k-token sessions live per HGX B300 node

166 ms

to first token

Native in your tools and harness

"My tools don't work anymore."

Same harness, same habits

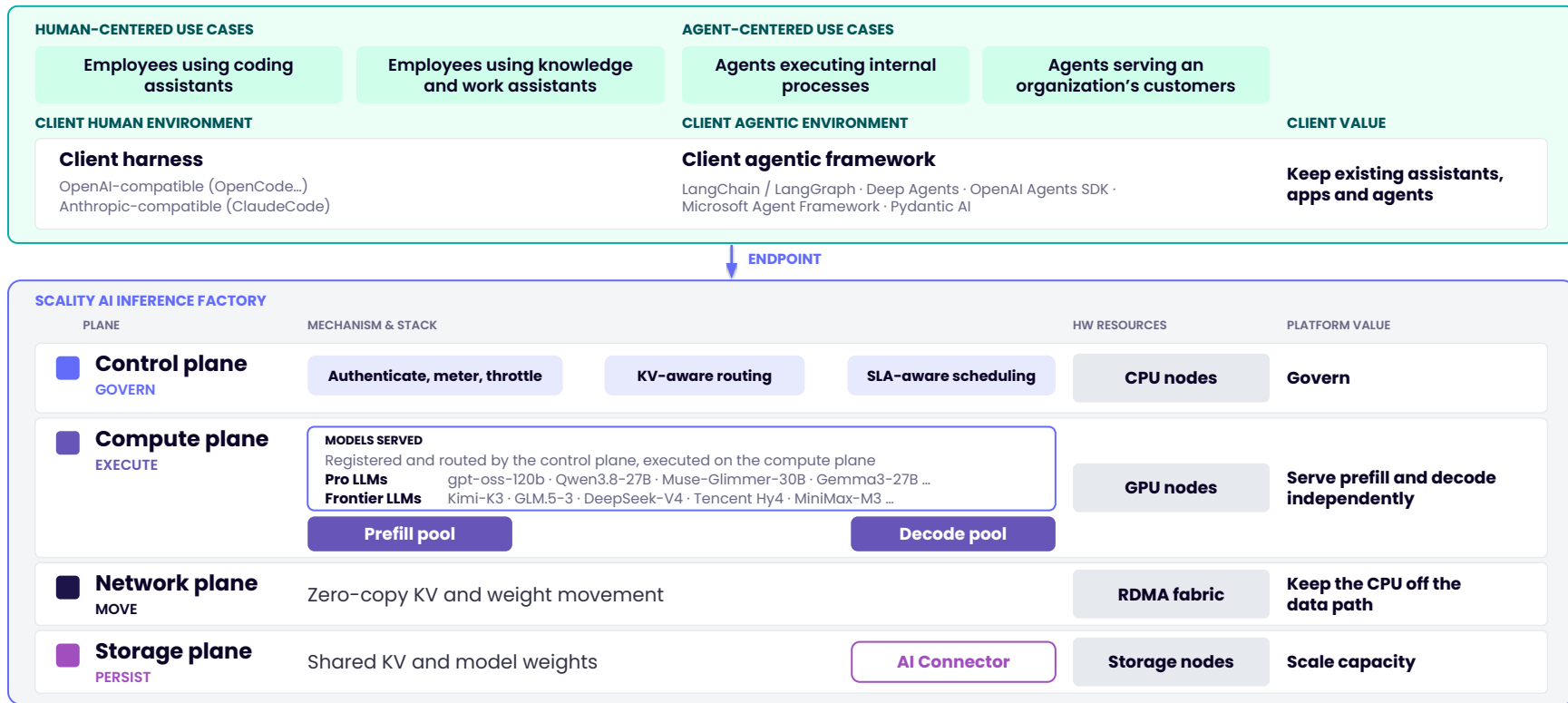
Zero setup for developers

Validated before cutover



HOW IT WORKS

One endpoint over the whole serving stack



10,000 developers: workload assumptions

Hybrid serving; 70% Qwen3.8-27B; 30% Kimi-K3

80,000

work sessions / day

320,000

developer requests / day

2.56M

model calls / day

Usage per developer	8 sessions × 4 requests per day, i.e. 32 requests Each request triggers 8 successive model calls
Work profile	About 700 tokens generated per call, reasoning included Median pause of 25 s between calls; history reused from cache
Projected concurrency	1,370 calls on average; 2,080 at P99* Peaks in the morning and afternoon, dip at lunchtime
Serving and models	Target: ≥ 40 tokens/s per active developer (median) Planned mix: 70% Qwen3.8-27B / 30% Kimi-K3



Equivalent cloud cost

Anthropic API: 70% Sonnet 5 / 30% Opus 5

\$34.7M / year

for 220 active days

\$2.89M

per month, over 12 months

\$289

per developer per month

Billed item	Annual cost
Re-reading cached history	\$21.83M
Writing new content to cache	\$7.15M
Generation, reasoning included	\$5.72M

Re-reading the history accounts for 63% of the bill



On-premise cost with Scality

Deployment assumption for 10,000 developers

60 nodes

HGX B300 + storage + control + network

480 GPUs

8 GPUs per HGX B300

≈ \$165 / month / dev

hardware + software + management + hosting

Item	Value	Ownership
Software + management	Priced per GPU, provide beyond SOTA technologies and 24/7 service	Scality
Hardware	Hyperscaler grade bill of material optimization	Partner via Scality
Hosting	AI specialized datacenters able to meet extreme requirements	Partner via Scality



HOW WE ENGAGE

From first call to production

Each step is validated on your workload before you commit to the next



STEP 1 · ASSESS

Tell us about your use case and workload

We study your goals, use case and workload: users, models, context lengths, peak hours...

OUTCOME

Preliminary architecture, sizing and end-to-end cost projection



STEP 2 · VALIDATE · 2 WEEKS

Run your workload on our HGX B300 lab

We deploy the product on 2× DGX B300 and you run your real workload for two weeks, with our measurement tooling.

OUTCOME

Sizing measured and confirmed on your own workload



STEP 3 · DEPLOY

Go to production with our support

We bring in the right hardware and hosting partners with BOMs optimized for your sizing, and go live as each rack lands.

OUTCOME

Production at the performance and cost you validated in the lab





THANK YOU

SCALITY
DAY26