CASE STUDY · FORTUNE 500 FINANCIAL SERVICES

A Fortune 500 financial services company releases $329K a year in idle Kubernetes capacity

Using Kyverno scale-down policies managed through Nirmata, the platform team scaled down 1,945 idle deployments across two clusters and released 579 CPU cores and 4.5 TiB of memory.

$329K	1,945	579.1	4,539.5
ESTIMATED ANNUAL INFRASTRUCTURE COST RELEASED	DEPLOYMENTS SCALED DOWN	CPU CORES RELEASED	GIB MEMORY RELEASED

CUSTOMER	SCOPE	NAMESPACES	POLICIES	REPORTS
Fortune 500 financial services company (name withheld)	Two production clusters	649	Kyverno scale-down policies, managed through Nirmata	Policy reports dated 23 September 2026

§01 CHALLENGE

The challenge

The company runs large, shared Kubernetes clusters serving hundreds of application teams. Over time, deployments that were no longer in active use kept holding their CPU and memory requests. That reserved capacity kept nodes running and made clusters look full, even though much of it was doing no work.

§02 SOLUTION

How the scale-down policies work

The platform team used Kyverno, the CNCF policy engine, with Nirmata to enforce scale-down as policy. Rules apply the same way across every namespace, so no team has to chase owners or run manual clean-up projects, and every action is recorded in a policy report.

01. **Identify.** Kyverno evaluates deployments against the agreed scale-down criteria across every namespace.

02. **Scale to zero.** Matching deployments are scaled down, releasing their CPU and memory requests back to the cluster.
03. **Flag.** Workloads that stay unused are labeled allow-delete for owner review.
04. **Report.** A per-cluster report tracks deployments, namespaces and capacity released.

Scaled-down workloads keep their configuration, so owners can restore them in seconds. Nothing is deleted without the allow-delete marker.

§03 RESULTS

Results

Policy reports dated 23 September 2026. Monthly and annual cost use the standard pricing described below.

SCALE-DOWN ACTIVITY

CLUSTER	DEPLOYMENTS SCALED DOWN	MARKED ALLOW-DELETE	NAMESPACES AFFECTED
Cluster 1	1,121	14	393
Cluster 2	824	7	256
Total	**1,945**	**21**	**649**

CAPACITY RELEASED AND COST

CLUSTER	CPU (CORES)	MEMORY (GIB)	MONTHLY	ANNUAL
Cluster 1	301.7	2,931.8	$16,030	$192,361
Cluster 2	277.4	1,607.7	$11,374	$136,487
Total	**579.1**	**4,539.5**	**$27,404**	**$328,849**

WHERE THE SAVINGS COME FROM

MONTHLY COST RELEASED	CPU	MEMORY	TOTAL
Cluster 1	$6,962	$9,068	$16,030
Cluster 2	$6,401	$4,973	$11,374
Total	**$13,363 (49%)**	**$14,041 (51%)**	**$27,404**

Memory accounts for slightly more than half of the released cost. Cluster 1's workloads were especially memory-heavy, at about 9.7 GiB per released core.

§04 NEXT STEPS

What comes next

The team plans to roll the same policies out to the rest of the fleet, turn the 21 allow-delete candidates into permanent clean-up, and pair scale-down with right-sizing so active workloads request only what they use.

§05 METHOD

Pricing method

- CPU at $0.031611 per vCPU-hour and memory at $0.004237 per GiB-hour: standard on-demand list rates, and the default rates used by OpenCost and Kubecost.
- 730 hours per month; annual figures are monthly × 12.
- Figures reflect resource requests released by scaled-down deployments. Savings are realized as the cluster autoscaler removes the freed node capacity.
- List prices before any committed-use or enterprise discounts. Only two clusters are included, so fleet-wide savings will be higher.

nirmata

Nirmata created Kyverno, the CNCF graduated policy engine for Kubernetes. We help platform and security teams govern clusters, workloads, and AI agents with policy as code.

TALK TO THE KYVERNO MAINTAINERS

nirmata.com/request-a-demo

Customer name withheld at the customer's request. Prepared by Nirmata · September 2026. © 2026 Nirmata, Inc. Kyverno is a CNCF graduated project. Kubernetes is a registered trademark of The Linux Foundation.