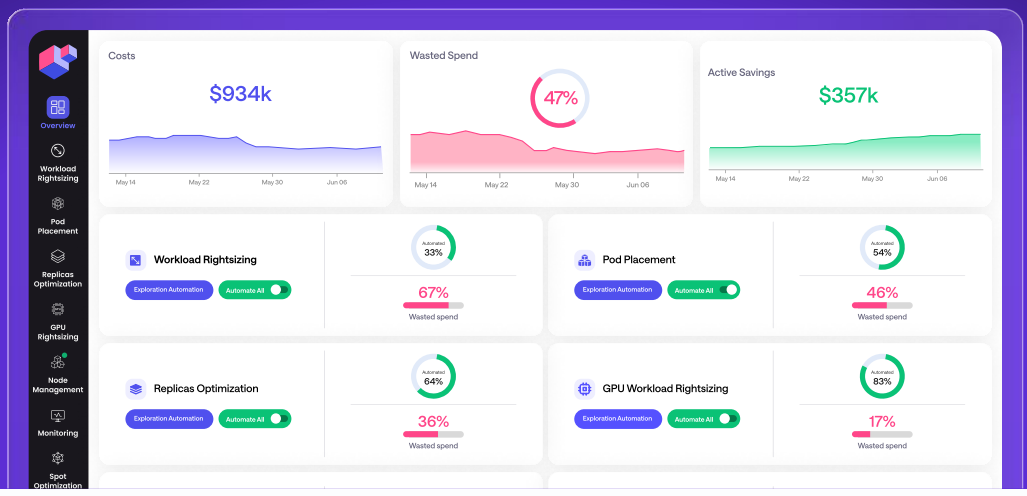
ScaleOps

Real-Time Autonomous Kubernetes Resource Management

Fully self-hosted, real-time, context-aware automation

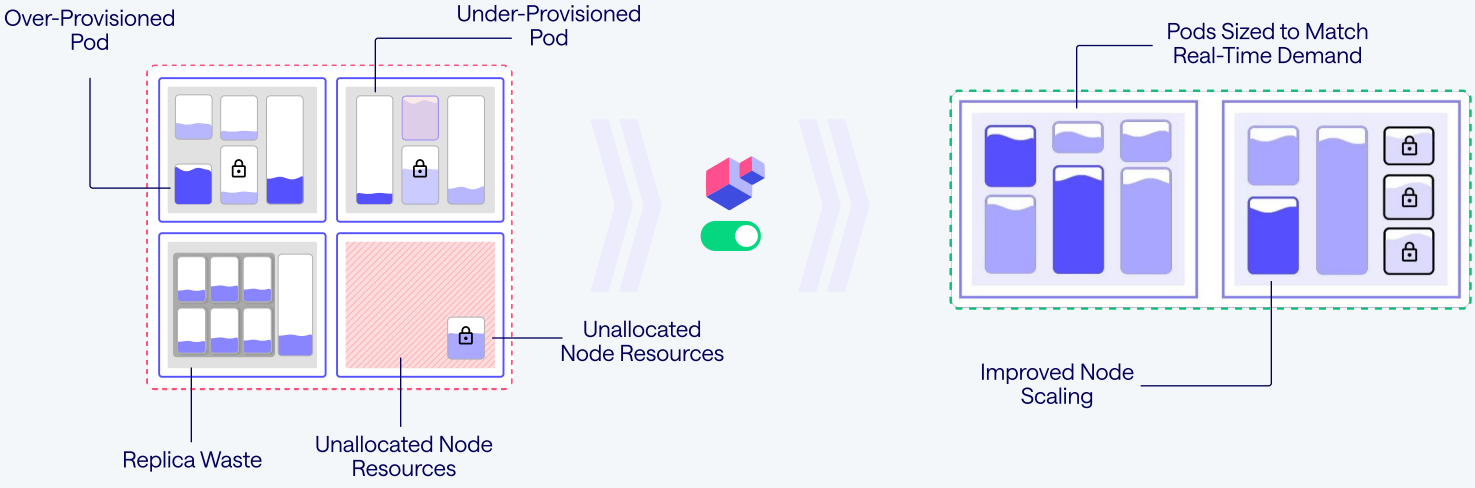
- Maximize Performance & Reliability
- Cut Costs by up to 80%
- Free Engineers from Manual Tuning



Industry Leaders Using **ScaleOps Automation in Production**

salesforce · WIZ · docusign · H&M · Freddie Mac · Adobe · coupa

Application Context-Aware Resource Optimization Platform Built for the Most Critical Production Environments



Run ScaleOps on any Kubernetes Environment

aws · Google Cloud · Azure · ORACLE · OPENSHIFT · IBM Cloud · Alibaba Cloud · On-Prem · Air-Gapped

ScaleOps

CORE INFRA

Real-Time Pod Rightsizing

Automatically rightsize CPU and memory requests based on workload behavior and live cluster conditions

Karpenter Optimization

Optimize Karpenter instance selection, disruption budgets, and identify inefficiencies

Smart Pod Placement

Eliminate waste from unevictable pods that block efficient binpacking and keep nodes underutilized

Node Optimization

Continuously optimize nodes while ensuring alignment with the needs of running workloads

Replica Optimization

Proactively scale ahead of demand by automating the management of min and max replica counts and triggers

Java Optimization

Automatically manage Java heap allocation and memory requests to match real application demand

Spot Optimization

Automatically shift replicas to Spot instances while maintaining workload reliability

AI INFRA

Manage and optimize **AI infrastructure** at scale with peak performance and zero GPU waste

Autonomous GPU Workload Rightsizing

Maximize GPU utilization with dynamic GPU sharing and automated workload rightsizing that ensures each workload receives the resources it needs based on real-time demand

AI Replica Optimization

Maximize model performance while minimizing replica overhead with intelligent scaling that eliminates cold starts and automatically adjusts replica counts based on real-time demand

Improved GPU Availability

Optimized GPU selection that cuts costs, boosts performance, and guarantees availability in your cloud region

OBSERVABILITY

Kubernetes Cost Monitoring

Visibility into your actual Kubernetes costs and accurately attribute spend at any level

Cluster and Workload Troubleshooting

Troubleshoot critical performance issues like OOM kills, CPU throttling and failed health checks

Cloud Cost Integration

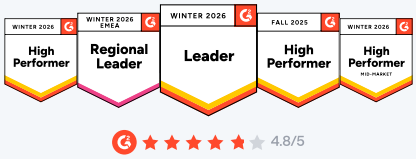
Integrate with cloud provider billing and cost data to get the full picture into your cloud costs

Ranked #1 by Customers in Autonomous Cloud Optimization

ScaleOps dramatically reduces our cloud resource costs. Automation in production freed teams from dealing with ongoing configurations, critical in our rapidly growing environment

Ron Tzrouya
Director of Cloud Financial Strategy

WIZ



We came in looking to save costs, but not at the expense of performance. With ScaleOps, we got both, and saw dramatically high cost savings without compromising on reliability.

Jeff Burger
Lead Platform Engineer

Vantor

Install in 2 minutes with a single Helm command. That's it