Smart Karpenter

AI-driven Kubernetes autoscaling with Karpenter and Smart Scaler

Achieve the highest Kubernetes efficiency at the lowest cost

Smart Karpenter combines AWS Karpenter's just-in-time node provisioning with Avesha's Smart Scaler, a Generative AI-powered autoscaler that predicts pod counts and traffic in advance. It proactively adjusts both pod and node resources across Kubernetes environments—without manual tuning or thresholds.

Designed for real-time, event-driven microservices, Smart Karpenter enables high performance, cost efficiency, and fully automated scaling, empowering teams to meet SLOs while reducing cloud spend by up to 70%.

Finvi and GeneDX use Smart Karpenter to eliminate service degradation during peak usage while significantly reducing cloud cost and manual DevOps intervention.

Key Features

1. Predictive pod scaling using app-level metrics like latency, RPS, and service dependencies
2. Dynamic node provisioning with AWS Karpenter based on predicted pod demand
3. No manual CPU thresholds—AI-driven scaling decisions instead of HPA tuning
4. Continuous learning and optimization through Reinforcement Learning models
5. Observation and Optimize modes for safe rollout and gradual AI takeover

Benefits

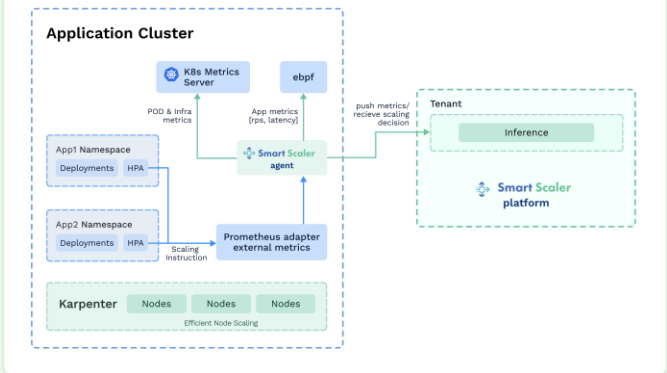
1. Up to 70% reduction in cloud costs through precise, just-in-time scaling
2. Right-sized pod and node provisioning, eliminating overprovisioning
3. Higher SLO compliance, especially during unpredictable traffic surges

How It Works

Smart Scaler is deployed via Helm in observation mode, monitoring real-time application metrics across environments. It maps out the service graph, forecasts traffic, and determines optimal pod and node counts.

Once in optimize(run) mode, Smart Scaler feeds predictions to Karpenter, which provisions the exact nodes needed—ensuring rapid startup and high resource utilization.

This two-layer AI + infrastructure approach enables fully autonomous autoscaling, removing the need for static node pools, dummy pods, or manual threshold configs.



Conclusion

Smart Karpenter delivers modern autoscaling built for modern apps—automating both application and infrastructure scaling with precision. It's the fastest way to optimize Kubernetes performance, reduce cloud spend, and remove DevOps scaling pain.