Avesha

Innovating the Cloud by automating infrastructure, workflow and GPU/CPU management

Smart Scaler

Introduction

Smart Scaler is the world's first AI based, predictive and fully-automated horizontal pod scaler for Kubernetes workloads. It is the only application-aware scaler that improves application performance (thus APDEX scores) and reduces costs by using fewer nodes.

Key Features

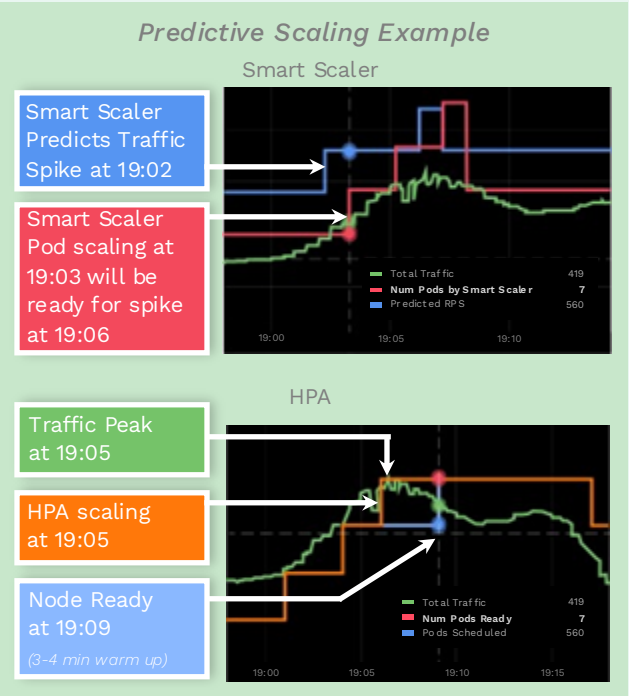
1. **Traffic-Based Scaling:** Utilizes algorithms to predict demand and adjust pod capacities in real-time.
2. **Event-Based Scaling:** Proactively scales pods ahead of scheduled events for seamless operations.
3. **Application-Metrics Based Scaling:** This is the only scaling automation that uses application behaviors such as service graphs, service-to-service latencies, error rates, RPS and CPU for accurately predicting scaling ratios for microservices.
4. **Lights-On/Lights-Off Use Case:** Automates pod startup and shutdown, aligning with working hours and schedule variations.
5. **Weekend and Weeknight Optimization:** Reduces resources during low-traffic periods to cut costs without affecting readiness.
6. **Karpenter Enhancements:** Enhances node scaling performance by offering predictability, reducing node latency. Karpenter plus Smart Scaler is our new product Smart Karpenter.

Benefits

1. **Resource Efficiency:** Addresses underutilization, with significant CPU savings in cloud-native applications.
2. **Cost Reduction:** Dynamically adjusts resources to avoid over-provisioning and idle costs.
3. **Reduces DevOps Burden:** To accurately scale based on application behaviors would take a DevOps person 2 weeks to set up for a set of microservices. Smart Scaler automates all that.
4. **Adaptability and Flexibility:** Easily adapts to schedule changes, maintaining operational efficiency.
5. **Sustainability:** Reduces resource wastage, aligning with environmental sustainability practices.

How It Works

Deployed as a SaaS, Smart Scaler analyzes application behaviors and service graphs to predict the pod count needed to maintain desired SLO. It uses AI for pod scaling, adapting to traffic patterns and microservice behaviors. In addition, it can scale up for events in a node-efficient manner, automatically.



Predictive Scaling

Note how in HPA (in bottom figure), the scaling is reactive and suffers from node startup delays. In Smart Scaler (in top figure), the scaling is predictive and hence able to mask node startup delays.

Conclusion

Smart Scaler represents a shift in K8s scaling leveraging application behaviors, predictive analytics and AI/Reinforcement Learning to enhance node efficiency and reduce costs while improving application performance and APDEX scores.