Smart Karpenter : Smart Scaler + Karpenter

Challenges

Addressing Kubernetes Scaling Complexities, Inefficiencies and High Costs

In today's rapidly evolving financial landscape, small and mid-sized businesses (SMBs) are increasingly adopting Generative AI (Gen AI) technologies to enhance operational efficiency. Specifically, sectors like insurance claim processing and bank loan underwriting are leveraging AI-driven solutions to automate document processing, improving accuracy and speed. However, one of the biggest challenges faced by these companies is the underutilization of GPU resources, which typically operate at only 15% efficiency. Enter Elastic GPU Service (EGS) – a groundbreaking solution designed to optimize GPU utilization and drive cost efficiency.

Solution

Achieving the highest K8s efficiency at the lowest costs

The Smart Karpenter solution, putting AWS Karpenter with Avesha Smart Scaler, addresses these challenges by offering a predictive, application-aware scaling mechanism. It combines the rapid node provisioning capabilities of Karpenter with the AI driven predictive pod scaling of Smart Scaler. Smart Scaler takes into account application behavior metrics from APM tools like Datadog etc. to accurately predict the number of pods. Smart Scaler feeds into Karpenter the predicted number of pods for all microservices, which enables Karpenter to provision the right-size of the nodes needed for the forthcoming pods. Also with Smart Scaler's intelligent HPA, there is no need to set a threshold on CPU utilization per pod. These features of Smart Scaler when combined with Karpenter, ensure the "highest" resource utilization, cost-effectiveness, and performance by accurately predicting application traffic and pod count needs in advance, leading to provisioning the right-sized nodes, thus more efficient and effective scaling decisions.

Benefits

1. **Predictive Scaling Efficiency**
   Smart Karpenter utilizes AI to accurately predict and manage application demands, ensuring the right number of pods and the right-sized nodes are optimally allocated before they are needed, enhancing efficiency and reducing waste. Since application metrics are used for scaling, it's far more accurate.
2. **Cost Reduction**
   With fully automating the K8s autoscaling process and optimizing resource utilization, Smart Karpenter significantly lowers operational costs, eliminating over- provisioning and underutilization.
3. **Completely Automated**
   Smart Karpenter simplifies the integration of Aidriven scaling into existing Kubernetes environments, making scaling decisions more accurate and less reliant on manual intervention.
4. **Enhanced Node Utilization**
   Smart Karpenter intelligently provisions the right type and number of nodes based on predictive analytics, ensuring high node utilization and efficiency, which translates into better application performance and lower costs.

Smart Karpenter

Smart Karpenter on AWS

Smart Karpenter is an innovative solution for Kubernetes autoscaling, combining AWS Karpenter's rapid node provisioning capabilities with Avesha's AI-driven Smart Scaler for predictive pod scaling. This innovative approach not only anticipates application demands in advance but also ensures optimal resource utilization (accurate pod count prediction) and cost-efficiency. Smart Scaler ‘s AI works on application behavior metrics for complexi modern microservices architectures to achieve the most accurate predictions for the application as a whole. Smart Karpenter paves the way for businesses to achieve higher operational efficiency and performance in their cloud environments.

Features

1. **AI-Driven Predictive Scaling**
   Smart Karpenter leverages advanced AI (Reinforcement Learning, Proximal Policy Optimization algorithms) to predict application traffic and pod scaling requirements ahead of time. The AI works on app metrics such as Request Per Second, Latency, service to service relationships, CPU, Memory etc. to arrive at the best scaling factor for each microservice) This feature allows for proactive resource allocation, ensuring that applications have the necessary resources before they are actually needed, thus avoiding performance bottlenecks and enhancing user experience.

2. **Optimal Node Provisioning**
   With its intelligent node provisioning mechanism, Smart Karpenter analyzes the requirements of “predicted” pending pods to select the most suitable instance types and quantities. This predictive approach ensures that the cloud environment is not only responsive to current demands in real - time, but also cost-effective, by avoiding over-provisioning and under- utilization of resources.

Case Study: theScore

1. **Challenges**
   theSCore struggled with the absence of a solution for dynamically scaling their sports-betting apps in sync with sports events' schedules and locations. Manual HPA configurations were labor-intensive and ineffective against traffic spikes, leading to application failures. They required a smarter solution to predict and manage demand surges without any manual input.

2. **Solution**
   Smart Scaler, with its AI-driven predictive scaling, was deployed at theScore to tackle their scaling challenges. It accurately predicted demand from the event calendar, automatically adjusting resources before and after events to manage traffic spikes and optimize costs.

3. **Results**
   Implementing Smart Scaler revolutionized theScore's operations, boosting both efficiency and reliability.The event-driven scaling eradicated service interruptions during high-traffic periods, drastically improving user satisfaction and ensures continuous service availability. Additionally, theScore saw a significant decrease in manual scaling tasks and resource overuse, resulting in significant cost reductions.

Get started with Smart Scaler on AWS

Visit Avesha on AWS Marketplace to purchase Smart Karpenter Today. For more information, go to https://avesha.io/products/smart-scaler