VIRTUSLAB

Success story

Increasing the efficiency of Kubernetes scaling

VirtusLab developed an AI-optimized Kubernetes operator for a Tel Aviv cloud provider, enhancing efficiency and cost savings.

Our Client

A Tel Aviv-based provider of cloud services, wanted to deliver to its customers a more efficient Kubernetes scaling. They were looking for a custom solution that could produce long-term cost savings without relying on allocating additional storage, memory or CPU.

VirtusLab helped the client build and integrate a custom Kubernetes operator with built-in optimization features, such as an AI-driven system for storing cache. The client implemented this solution for multiple customers and noted an increased reaction time to traffic spikes, better security of AWS Spot instances and cost savings.

The Challenge

Our client aimed to increase the efficiency of their Kubernetes scaling services, especially during unexpected traffic spikes. Typically, this would require the allocation of additional resources like memory, storage, and CPU to the Kubernetes cluster, which comes with a risk of overprovisioning. The client was looking for an alternative solution that would also cut an instance creation time to 30 seconds or less.

VirtusLab designed and implemented a Kubernetes operator that steered clear of overprovisioning. The goal was to improve the efficiency and cost-effectiveness of Kubernetes scaling, especially when it involves hundreds of nodes.

The solution

VirtusLab adopted two key Kubernetes performance optimization techniques, namely hibernation and internal scaling, to make sure the operator works efficiently and saves money.

Additionally, VirtusLab has integrated distributed tracing, powered by OpenTelemetry's set of standards and tools. This combination will help the client with future maintenance and system monitoring.

To enhance the efficiency of scaling up in Kubernetes, VirtusLab implemented an intelligent, AI-driven system that stores and rapidly accesses recently used container images, maintaining their cache. By doing so, the system quickly accesses these resources without having to retrieve them from a more distant source every time.

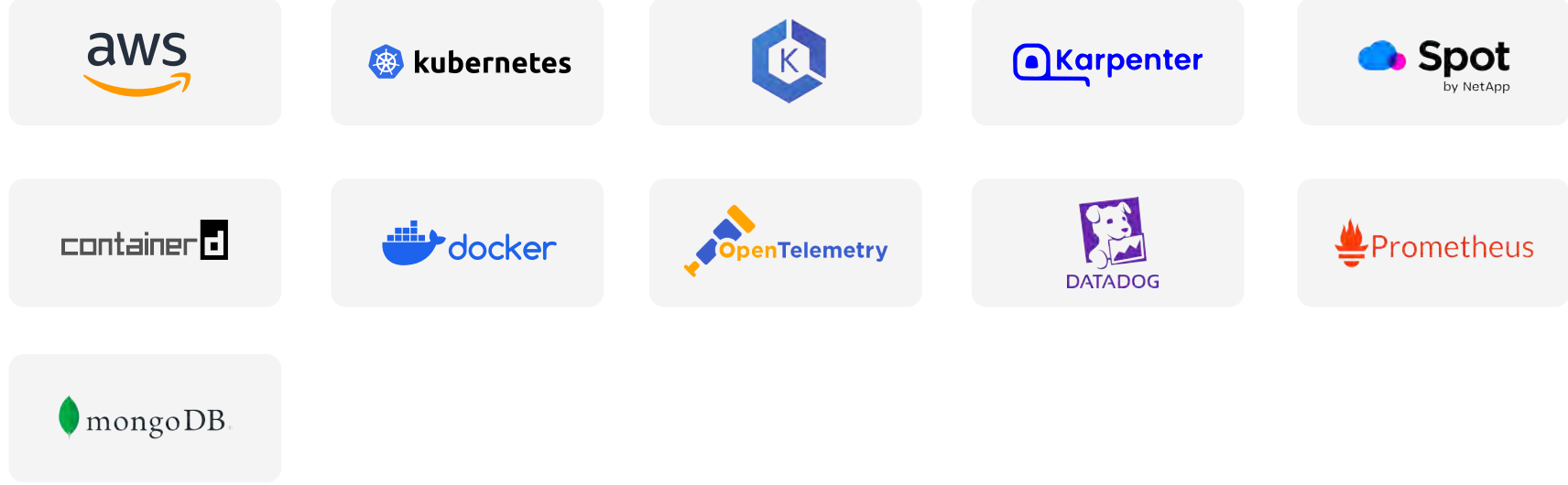
The results

The client successfully deployed the completed Kubernetes operator for multiple customers. This decreased the time required to create a single instance from 60 to 22 seconds, without exceeding the limitations of Kubernetes and the AWS API. Additionally, the client observed:

- **Reduced costs for each cluster**
- **Immediate response of cluster scaling to traffic surges**
- **Enhanced AWS Spot instance protection**

The tech stack

Programming language	Go
Platform	Kubernetes
Database	MongoDB
Cloud	Amazon Web Services (AWS)
Tooling	Various autoscaling products such as EKS Cluster AutoScaler, Karpenter and Spot.io, ContainerD and Docker, OpenTelemetry, DataDog, Prometheus, MongoDB



About VirtusLab

At VirtusLab, we aim to lead in software technology, working consistently to enhance efficiency. Our profound commitment to research and development and a dedicated focus on emerging trends and inspirations fuels an innovative culture. This ethos precisely guides advancing our cutting-edge solutions, inviting collaboration to expand the boundaries of software technology collectively. We welcome you to be a part of this transformative journey.

Let’s connect

Contact Details

info@virtuslab.com

POLAND

Kraków Headquarters

Virtus Lab Sp. z o.o.
ul. Szlak 49
31-153 Kraków

GERMANY

Berlin Office

+49 30 52014256

VirtusLab GmbH
Potsdamer Platz 10
10785 Berlin

UNITED KINGDOM

London Office

+44 (0)20 4577 1051

Virtuslab Ltd.
40 Bank Street HQ3
London E14 5NR