## ΛΚΛΜΛS

# Autonomous Optimization for Kubernetes applications





### Kubernetes is critical for app efficiency

Efficiency is one of **Kubernetes** top benefits, yet many companies often experience higher than expected **infrastructure/cloud costs** as well as **performance and stability issues**.

The **complexity of Kubernetes resource management** often leads developers, Performance Engineers and SREs to adopt very conservative configurations and **resource overprovisioning**.

The resulting **unnecessary infrastructure/could costs** may significantly affect the overall **cost efficiency** of delivered services, while not necessarily removing the **risks of missing SLOs**.



of companies mentioning lack of in-house skills and limited manpower as the #1 challenge with Kubernetes Source: Statista

of Kubernetes users with the unmet challenge of ensuring application performance and resilience Source: Akamas survey

customers facing increasing Kubernetes costs and missing/inadequate capabilities to avoid overspending Source: 2021 FinOps report

 $\Lambda K \Lambda M \Lambda S$ 

### **Tuning Kubernetes is very challenging**

Kubernetes cluster size is determined by **resource requests** set by developers in the container manifest for each application - so clusters can be full even if utilization is very low (e.g. 1%).

Sizing Kubernetes containers/pods by properly setting resource requests and resource limits/may require weeks or months for each single microservice - so overprovisioning is common.

Kubernetes autoscaling enable automatic scaling under load but can also magnify latency & efficiency issues when Kubernetes microservices applications are not well configured.

too low resource requests may cause slowdowns and reliability issues - too high requests may lead to wasted resources and overspending

too low resource limits impact application stability and performance as containers get killed or slowed down under resource pressure

### too high resource limits

allow runaway containers to act as "noisy neighbors" and degrade performance of other applications sharing the same nodes

ΛΚΛΜΛS

### **Autonomous Performance Optimization**

$ \begin{array}{c} & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & $	Parameter 🔶	Relevance \$	Best	Baseline
<ul> <li>Imits_opu ●</li> <li>Imits_memory ●</li> <li>Coutaliner requests_memory ●</li> <li>Imits_memory ●</li> <li>Cloud Bill ↓49.1%</li> <li>155 \$ → 79 \$</li> <li>Mym_mctiveProcessorCount ●</li> <li>Imits_crype ●</li> <li>Imits_memory ProcessorCount ●</li> <li>Imits_crype ●&lt;</li></ul>	requests_cpu		2.77 cores (+84.9%)	1.5 cores
<ul> <li>Imits_memory ●</li> <li>Baseline vs Best (Experiment #31)</li> <li>Cloud Bill ↓49.1%</li> <li>155 \$ → 79 \$</li> <li>ym_activeProcessorCount ●</li> <li>ym_gcType ●</li> <li>ym_minHeapSize ●</li> <li>Ym_minHeapSize ●</li> <li>Transactions/s ↓19.2%</li> <li>85 TPS → 104 TPS</li> </ul>	Container limits_cpu	-	3.67 cores (+83.3%)	2 cores
$ \begin{array}{c} \hline \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ $		-	Baseline vs Best (Expe	eriment #31)
$ \begin{array}{c} & \begin{tabular}{ c c } & tabu$		-		
$ \begin{array}{c} & \\ & \\ & \\ & \\ & \\ & \\ & \\ & \\ & \\ & $	jvm jvm_activeProcessorCount	-		
$ \begin{array}{c} \begin{array}{c} \begin{array}{c} \\ \end{array} \end{array} \\ \begin{array}{c} \\ \end{array} \end{array} \\ \begin{array}{c} \\ \end{array} \end{array} \\ \begin{array}{c} \\ \\ \end{array} \\ \begin{array}{c} \\ \\ \end{array} \end{array} \\ \begin{array}{c} \\ \\ \end{array} \\ \begin{array}{c} \\ \\ \end{array} \\ \end{array} \\ \begin{array}{c} \\ \\ \\ \end{array} \\ \begin{array}{c} \\ \\ \\ \end{array} \end{array} \\ \begin{array}{c} \\ \\ \\ \end{array} \\ \begin{array}{c} \\ \\ \\ \end{array} \\ \begin{array}{c} \\ \\ \\ \end{array} \end{array} \\ \begin{array}{c} \\ \\ \\ \\ \end{array} \\ \begin{array}{c} \\ \\ \\ \end{array} \\ \end{array} \\ \begin{array}{c} \\ \\ \\ \end{array} \\ \begin{array}{c} \\ \\ \\ \end{array} \\ \end{array} \\ \begin{array}{c} \\ \\ \\ \end{array} \\ \end{array} \\ \begin{array}{c} \\ \\ \\ \end{array} \\ \begin{array}{c} \\ \\ \\ \end{array} \\ \end{array} \\ \begin{array}{c} \\ \\ \\ \end{array} \\ \begin{array}{c} \\ \\ \\ \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \\ \\ \\ \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \\ \\ \\ \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \\ \\ \\ \end{array} \\ $	jvm ≝ jvm_gcType ❶	-		5
$ \int_{\text{jvm}_minHeapSize} \bullet $ $ \int_{\text{jvm}_min$	jvm jvm_maxHeapSize	-		5 V
- Exp 31, Trial 1, frontend_ jncoming_response_time_90_ms	jvm jvm_minHeapSize	-	↓ <b>19.2%</b>	500 ms 250 ms 0 ms 0 0 ms 0 0 0 0 20 0 0 0 16 40 0 0 25 00 0 0 25 00 0 0 33 20 Time from the start of the trial

akamas.io • © 2022 Akamas S.p.a. • All Rights Reserved

Akamas Al-powered optimization automatically identifies the best configurations for hundreds of parameters for any Kubernetes microservices applications, and any IT component, including runtimes, databases, middleware and cloud instances.

Custom optimization goals & constraints allow you to set the desired performance, resilience and cost objectives for each application to be optimized.

### **Akamas benefits for Kubernetes apps**

Akamas autonomous optimization guarantees the best levels of performance & resilience while also ensuring the **best cost efficiency** of your Kubernetes microservices applications, thus avoiding any resource overprovisioning and unnecessary infrastructure/cloud costs.

With Akamas, developers, performance engineering & SREs are freed from manual tuning activities and can focus on innovating and operating Kubernetes applications, by automatically receiving actionable insights on potential tradeoffs and recommended configurations.

This results into **higher operational efficiency**, business agility and competitive advantages.

-/0 -80%

decrease in response time with lower fluctuations and peaks



decrease in infrastructure/cloud costs with the same application performance

savings in engineering time spent for manual tuning tasks

ΛΚΛΜΛS

