AKAMAS

The Kubernetes dilemma:
How to tune application performance while lowering cloud costs

A customer success story in the gaming and entertainment industry

Luca Forni - CEO

AKAMAS

Project background

- International customer, gaming & entertainment industry

- Application with more than 30M users

- Dozens of Kubernetes clusters in multi-cloud environment

- Service quality and E2E experience are a must

- Continuous focus on application performance tuning while minimizing cloud costs

ΛKΛMΛS

Kubernetes architecture comes with new challenges

Challenges

Kubernetes introduces new **operational and development** challenges

Kubernetes technology requires a new **resource management paradigm**, shifting responsibility from ops to dev teams, creating more opportunities for **failure**

Traditional APM and AIOps tools may identify potential optimization opportunities, but it is not easy to **safely address them**

Impacts

Teams spend **significant effort** for performance analysis and tuning to ensure **service quality**, with a **manual approach** leveraging observability tools

Kubernetes complex management introduces **cost issues**, with some applications requiring significant **infrastructure footprint**

Despite infrastructure investments, the ability to support increasing levels of **business growth** may also not be guaranteed nor managed

Goals and scope of the project

Akamas AI was tasked with finding the best configuration parameters in the Kubernetes architecture and JVM runtime, given these explicit goals:

- Improve cost efficiency of the cloud infrastructure, rightsizing the resource capacity of each K8s pod

- Maximize business throughput of the entire system

- Improve the overall quality of the application services
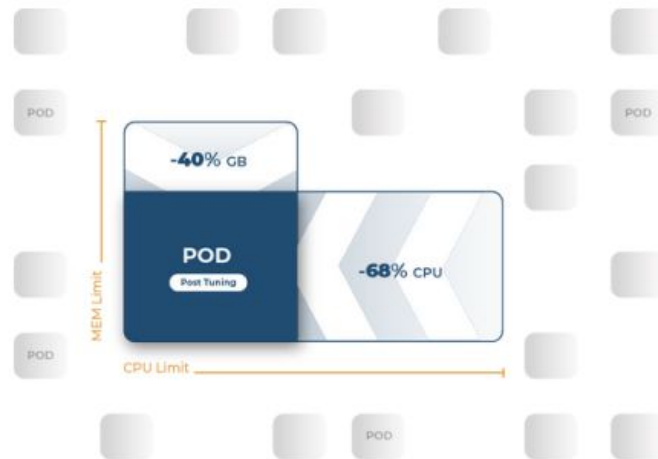
Total savings result from cumulative pod optimization resizing across farms

Additional project goal: automating the optimization processes to reduce **teams' effort** for manual performance analysis and configuration tuning

Akamas live optimization cycle

Akamas integrates in the customer technology ecosystem, **autonomously tuning** configuration parameters with an "application-aware" approach, to reach performance improvements and cost reduction goals, while preserving the end user experience quality. The optimization cycle is fully driven by Akamas patented AI.
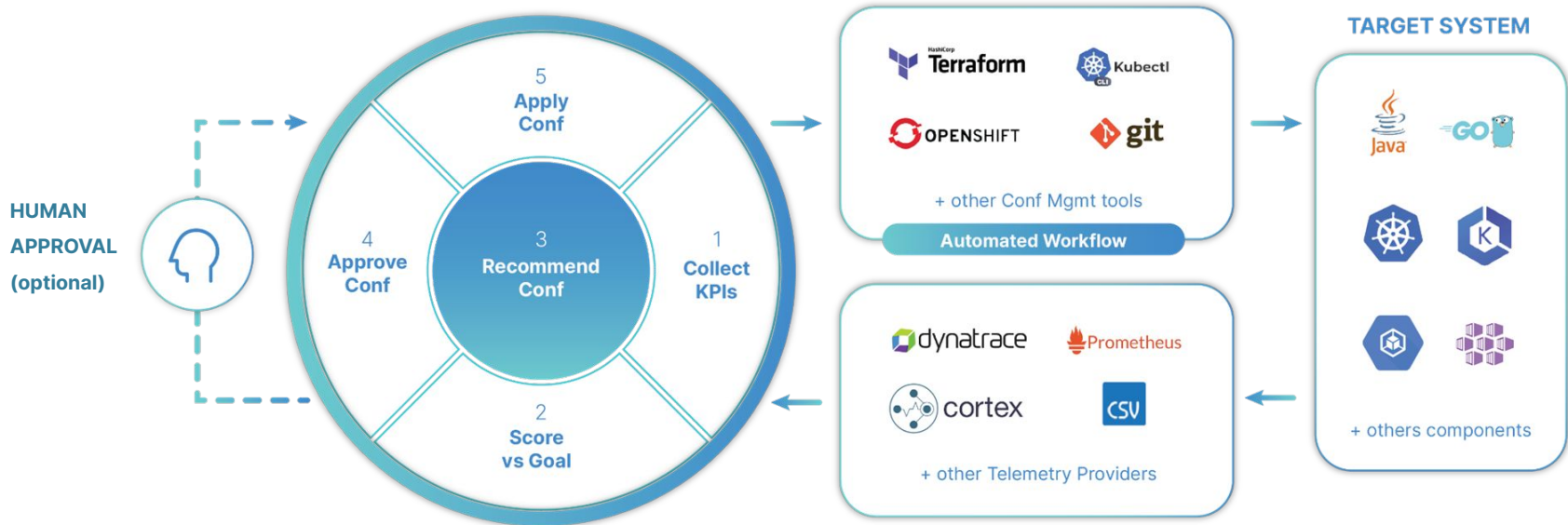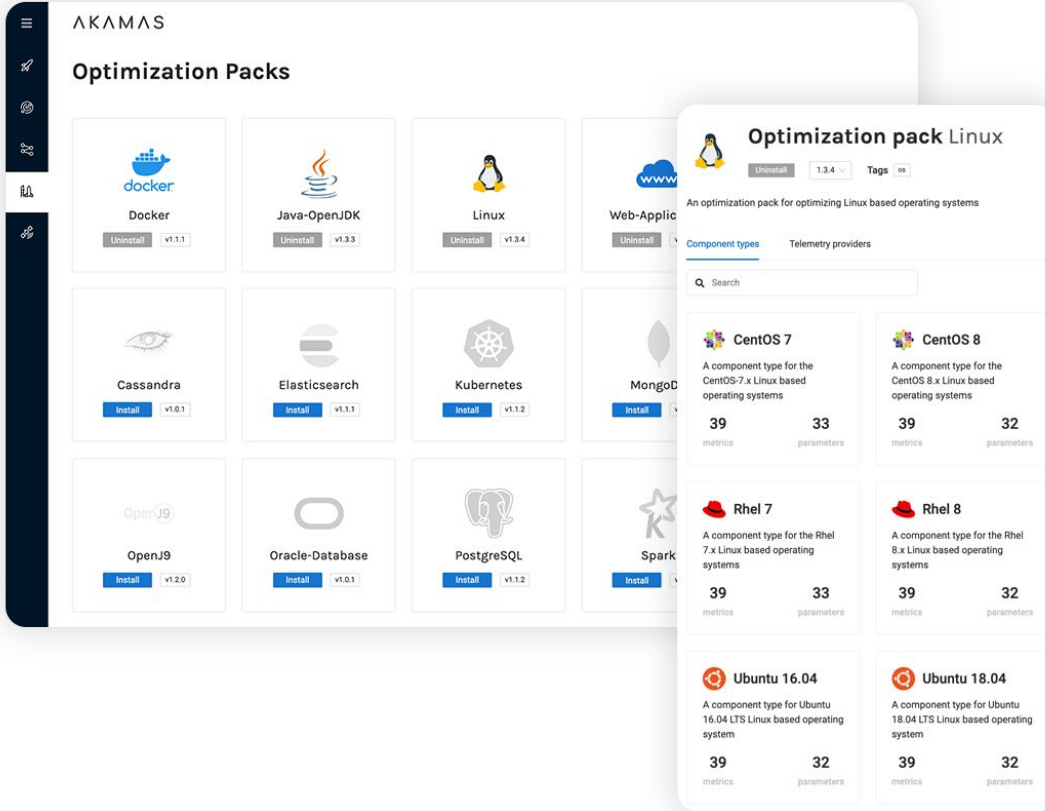
Autonomous optimization results

Akamas AI achieved and improvement in resource consumption (**-65% CPU, -40% memory**), while preserving end user performance and improving application resilience
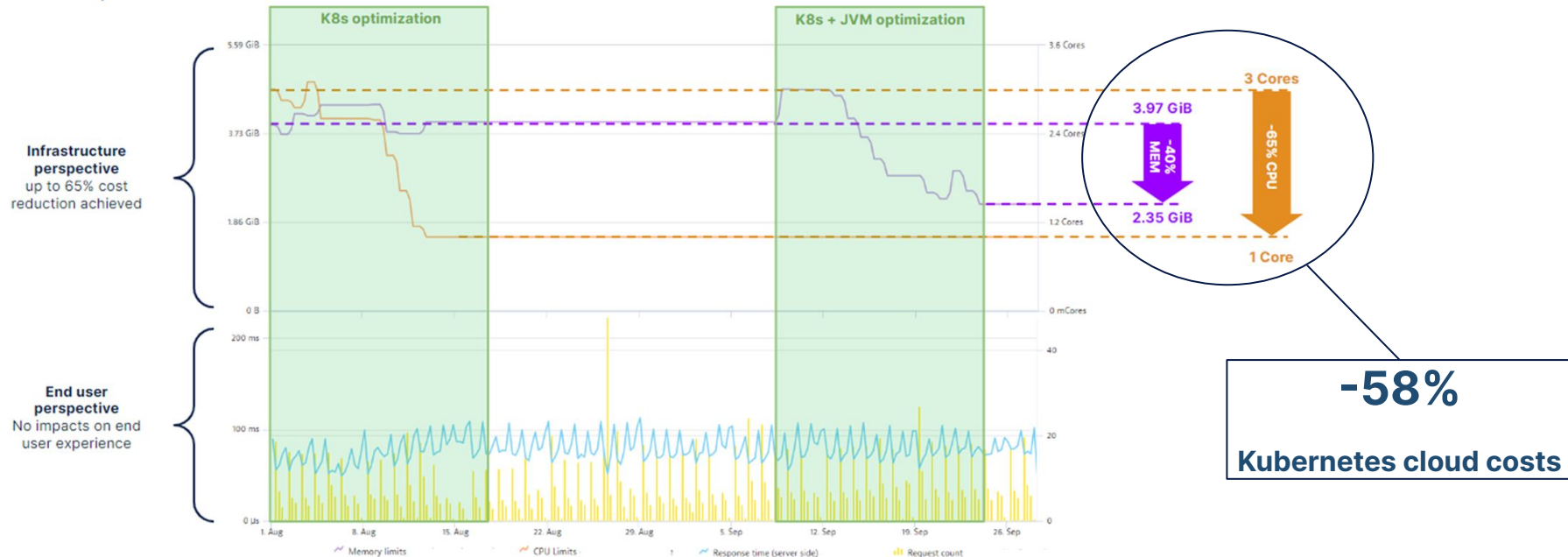
Technology overview

Akamas machine learning algorithms are able to optimize almost any technology.

The Akamas R&D lab, working jointly with partners and universities, enterprise customers, technology vendors and IT performance "gurus", trains the AI model on configuration parameters. The model learns how to (safely) modify parameters to reach the customer's optimization goals.

This knowledge is packaged in a set of libraries called "Optimization Packs".

Set of parameters to be tuned

Value after tuning (as recommended by Akamas AI)

Before tuning (original value)

Study Optimize konakart heap

● Finished   Started on 8 May 2021, 19:29   Duration 1

Tune the JVM to increase transaction throughput while keep...

Summary   Progress   Configuration analysis

Experiment Scores

| db postgres_checkpoint_completion_target | 0.75 distance (+50%) | 0.5 distance |
db postgres_effective_cache_size	1 GB (-75%)	4 GB
db postgres_effective_io_concurrency	2 iops (+100%)	1 iops
jvm8 jvm_gcType	G1	Parallel
jvm8 jvm_maxHeapSize	1.75 GB (-41.7%)	3 GB
jvm8 jvm_maxTenuringThreshold	4 (-73.3%)	15

Improvement

The Akamas AI explores the configuration parameters space, acting like a super-expert consultant, continuously analyzing data coming from monitoring, APM, observability and telemetry tool – often already in place within a customer's IT environment.

It recommends and implements configuration changes to get better and better results.

Every changes is tracked and shown in Akamas UI, letting human expert to review them and learn.
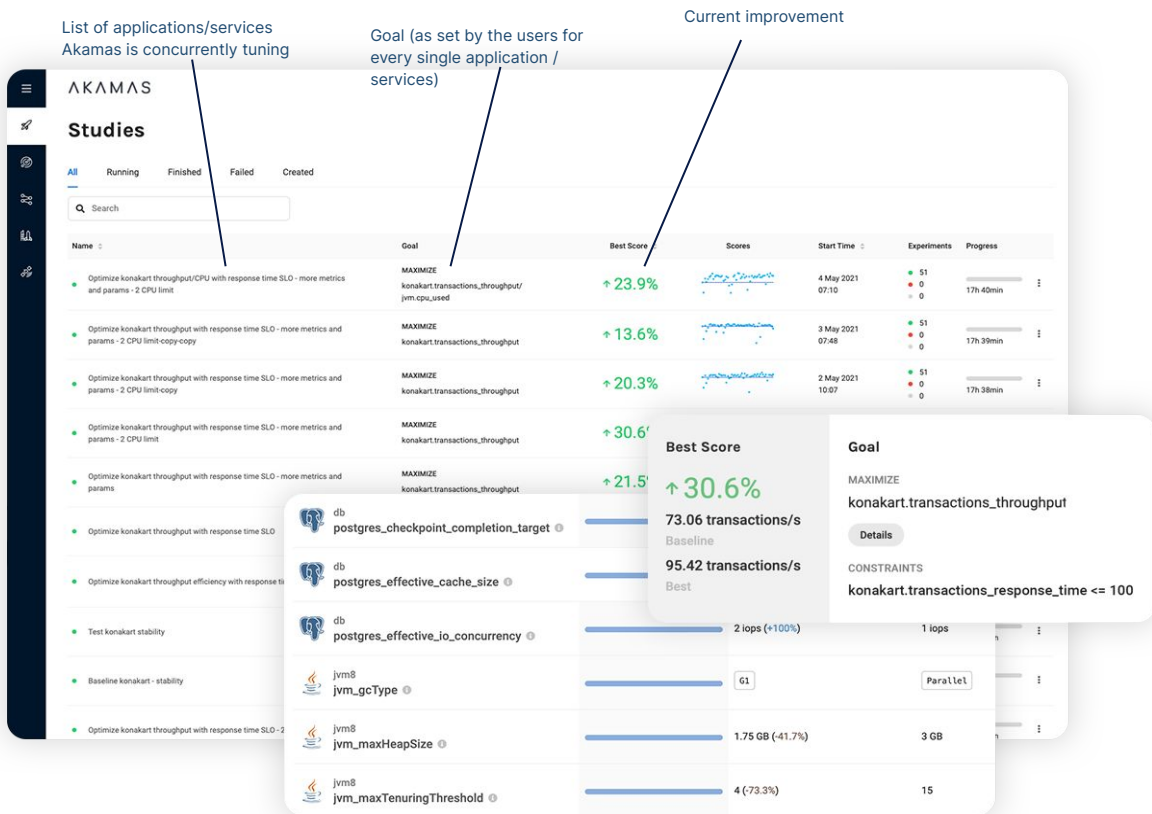
Let akamas run unattended in your IT environment, dramatically boosting your tuning process, for every application, continuously.

Free your valuable DevOps, SRE, Operations teams time, avoiding manual time-consuming and high-effort tuning activities.

Let the AI doing the job.