sedai

SOLUTION BRIEF

SEDAI FOR KUBERNETES

Autonomously Manage Kubernetes Services

Achieve the highest levels of cost efficiency, performance and availability for Kubernetes applications & infrastructure with the Sedai Autonomous Cloud platform.

Results

53% Cost reduction

30% Latency reduction

33% SRE workload reduction

Key Benefits

- **Autonomous systems take action safely in production, 24/7, without human intervention**
- **Application-aware optimization of performance, availability and cost**
- **Continuous improvement using reinforcement learning**
- **Boost productivity for SREs and Developers**
- **Accelerate innovation velocity**

Summary

The complexity of managing Kubernetes applications at scale has surpassed the ability of teams using manual optimization, resulting in sub-optimal results and coping strategies such as over-provisioning of resources. Autonomous systems such as Sedai use machine learning to continually optimize performance, cost and availability. Sedai autonomously detects, recommends, validates and executes improvements in production without human intervention, at both an application and infrastructure level, extending native Kubernetes capabilities. Sedai uses rigorous safety checks to safely operate in production. Autonomous operation reduces low value tasks (toil) by SREs & Operations teams, enabling them to accelerate delivery of higher level initiatives. Sedai also accelerates innovation velocity by providing immediate feedback on new releases to development teams. Sedai offers an attractive ROI (700% for recent customers) and a rapid payback (<3 months) supported by a rapid rollout period (10 minutes for initial setup, and a ~2 week learning period).

The Challenge of Kubernetes Optimization

Using Kubernetes with microservice architectures offers major advantages over traditional architectures and has become a strategic imperative. However, Kubernetes is difficult to optimize. Many Kubernetes users experience unexpected cost growth, performance and reliability problems. A key underlying challenge is the complexity of resource management. Users must specify a wide range of parameters (e.g., CPU, Memory) and other resources across large numbers of interrelated services in a dynamic environment that changes constantly as developers ship new code and user demand changes daily. Setting these parameters manually or using point-in-time optimization approaches fails to achieve optimal results on an ongoing basis and underdeliver on the potential of Kubernetes. Given these difficulties many development teams opt for a coping strategy involving significant over-provisioning of resources to minimize performance issues.

The Sedai Approach

Sedai manages Kubernetes applications autonomously using state-of-the-art machine learning and predictive engines that detect potential optimization opportunities and issues within your application and takes autonomous action to address them in production, optimizing cloud costs, improving performance and ensuring availability. To enable this model, Sedai automatically discovers your application's architecture, analyzes traffic patterns as well as performance and cost metrics. Sedai works with Kubernetes running on AWS EKS, Azure AKS, Google GKE and any Kubernetes distribution via the Kubernetes API.

Autonomous

Sedai's autonomous approach does not require manual thresholds or human intervention, unless desired by the user. This represents a substantial change to traditional tools (see below) - rather that just detecting and/or recommending changes, Sedai is able to validate changes to determine they are safe to make in production, as well as execute them in a production environment.

	Traditional Tool	Sedai Autonomous Platform
Detect	Partially Automated	Automated
Recommend	Partially Automated	Automated
Validate	Manual	Automated
Execute	Manual	Automated

For workloads where the user needs to use manual approval steps, Sedai can be set to manual mode, and integrations allow approval inside other workflow tools.

Application Aware

Unlike infrastructure-centric tools, Sedai understands the Kubernetes applications running and the dependencies between them through analysis of application behavior and high level objectives set by the user. This enables Sedai to optimize separately for latency-sensitivity applications that drive customer experience & revenue, and those that are cost-sensitive (e.g., batch workloads). After finding improvements that benefit both cost and latency, Sedai can prioritize cost alone, latency, or make a balanced tradeoff.

Continuous Learning

Sedai also evaluates the outcome to improve its ability to find future improvements, creating a continuous learning loop and providing more capabilities than one-time optimization tools. Sedai uses a reinforcement learning approach, providing its ML algorithms with reward signals as improvements are made in performance, cost and availability and as accurate predictions of application seasonality and application behavior are made.

Key Use Cases

Reduce Cloud Costs

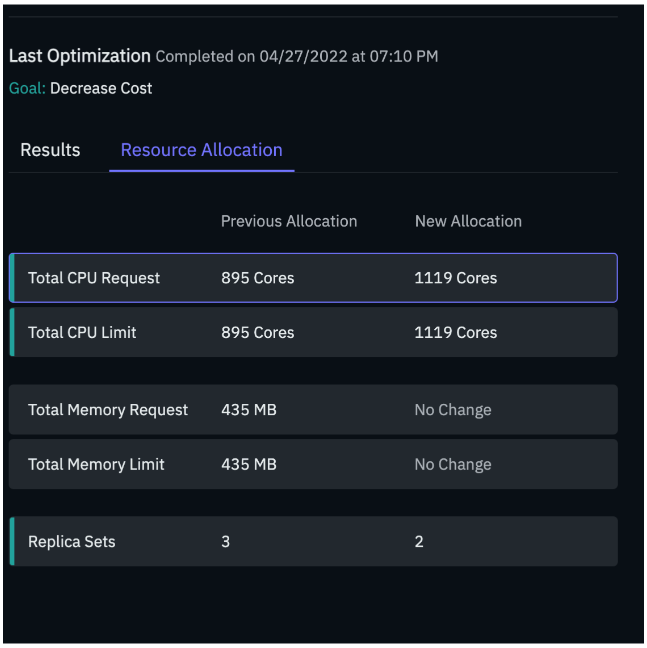
Sedai employs three sets of tactics to reduce Kubernetes costs: Application Optimization, Cluster Optimization and Purchase Optimization; each approach can generate 20-30% cost savings.

1) Application Optimization: Kubernetes services often run with low resource utilization which drives up cloud costs. The median Kubernetes deployment uses just 20-30% of requested CPU and 30-40% of requested memory[1]. Sedai continuously learns from traffic patterns (e.g., Monday morning peak, night time traffic lows) and uses this to match demand with resources needed. Where demand varies from predicted seasonality, Sedai will scale up or down. Sedai uses both horizontal and vertical scaling at the same time; most users today are not able to use horizontal and vertical together due to system limitations or without significant testing work. Sedai does this autonomously, allowing 24/7 operation and avoiding human error.

An autonomous approach avoids the heavy manual workload of a threshold approach which requires extensive analysis and the correct thresholds often change with each new release. Sedai's autonomous approach to cost management can resolve a common challenge for Kubernetes teams in which initial configurations may be set by development teams who have application reliability and performance as overriding concerns and are less focused on cost.

Other settings may not be used at all e.g., only 40% of organizations use Kubernetes HPA (horizontal pod autoscaling) and less than 1% use VPA (vertical pod autoscaling) [2]. Sedai acts as a controller for HPA and VPA and can choose smartly between both, wherever its available. Where HPA and VPA are not available, Sedai provides an autoscaling capability.

The example at right shows a Sedai application optimization updating CPU and memory settings as well as reducing node count, improving resource usage on behalf of the SRE team.



Last Optimization Completed on 04/27/2022 at 07:10 PM
Goal: Decrease Cost

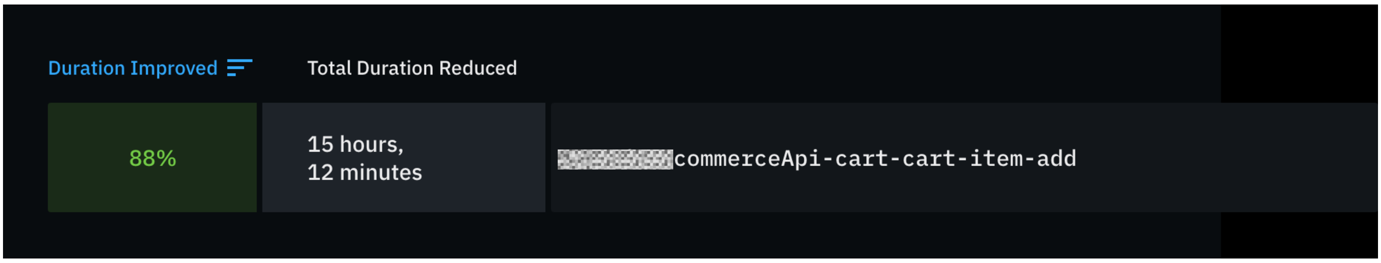
	Previous Allocation	New Allocation
Total CPU Request	895 Cores	1119 Cores
Total CPU Limit	895 Cores	1119 Cores
Total Memory Request	435 MB	No Change
Total Memory Limit	435 MB	No Change
Replica Sets	3	2

2) Cluster Optimization: Sedai optimizes the type and grouping of nodes to run applications. Sedai picks the optimal instances type to run applications (e.g., AWS M5 vs M6a). Sedai's node group optimization allows applications with similar needs to run on the optimal instance type (e.g., memory-intensive vs compute-intensive). Unlike many other cloud cost tools Sedai is application-aware and also uses the underlying application latency sensitivity to determine the optimal cluster configuration.

3) Purchase Optimization: Sedai optimizes the cost of compute resources based on the instance purchasing options provided by public cloud providers including spot and reserved instances.

Improve Performance & Customer Experience through Reduced Latency

For customer facing applications, latency is a top concern, driving customer experience and revenue. Sedai reduces latency by up to 95% at service level by finding the optimal parameters for memory and CPU, and placing the application in the optimal node group. In the example below the execution time of a service that adds a new item to an ecommerce shopping cart was reduced by 88%:



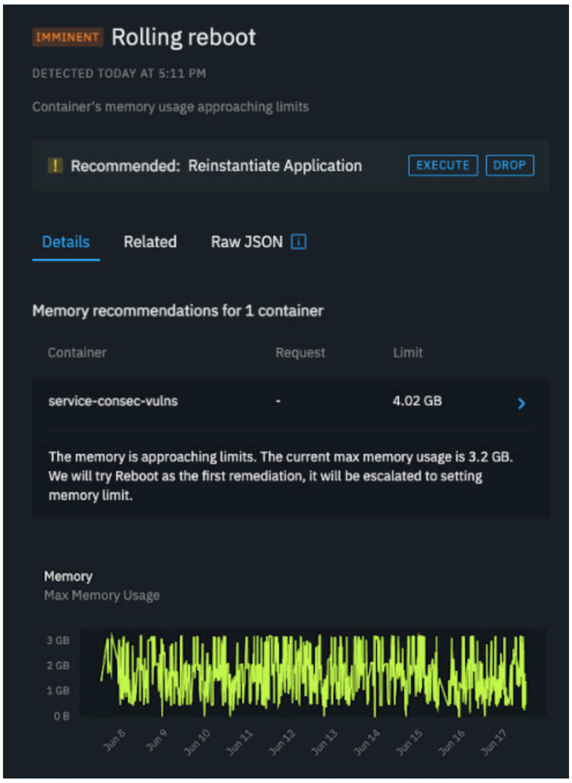
Duration Improved	Total Duration Reduced	
88%	15 hours, 12 minutes	[redacted]commerceApi-cart-cart-item-add

Sedai also reports on the number of days of cumulative latency reduction achieved and the cloud spend needed (where applicable) to achieve this.

Improve Availability

While application uptime has improved across the industry in recent years, outages remain a key risk with potential revenue and reputation impacts as well as absorbing significant team resources during the event. Sedai improves availability through scaling patterns based on seasonality to respond to demand surges. Sedai also watches to see if services are experiencing common problems such as high CPU usage resulting in throttling, or out of memory errors.

Sedai uses an escalation based remediation approach. Depending on the observed application context, Sedai attempts multiple remediations based on the learning from past application behavior. For example Sedai may first try to restart the service and if that does not address the issue Sedai may recommend changing the CPU or Memory limits (and execute those in autonomous mode after applying safety checks). The example shown at right shows the detection and proposed remediation of a memory issue.



IMMINENT Rolling reboot

DETECTED TODAY AT 5:11 PM

Container's memory usage approaching limits

! Recommended: Reinstantiate Application

Memory recommendations for 1 container

Container	Request	Limit
service-consec-vulns	-	4.02 GB

The memory is approaching limits. The current max memory usage is 3.2 GB. We will try Reboot as the first remediation, it will be escalated to setting memory limit.

Memory
Max Memory Usage

Hit SLOs

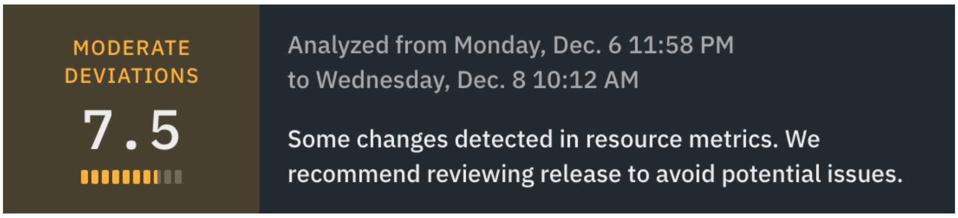
Sedai supports organizations with existing SLOs and can also automatically (or manually) set SLOs for organizations not yet using SLOs. Where SLOs are not yet in place, Sedai can auto-recommend SLOs based on analysis of application behavior and user-provided tags (e.g., key checkout services for ecommerce). These SLOs are then autonomously managed by Sedai, which ensures that performance for each service falls within the SLO. For example, if a service slows down (below the latency SLO) during a peak traffic period, Sedai will scale it up to compensate and make sure it stays within the SLO.

Improve SRE Productivity by Reducing Toil

Sedai enables you to support workload growth without needing to linearly scale up the SRE team. Sedai's autonomous detection and optimization means the work of detecting & resolving issues and manually finding optimization can now be managed by SREs with significant lower effort as the autonomous system takes care of many tasks on their behalf. For example, it is tedious for teams to configure HPA/VPA when they have multiple applications and each needs different thresholds. Sedai enables SREs to be less concerned with what is happening in the infrastructure and the chaos of alert fatigue and reactive troubleshooting, and instead focus on new deployments, site enhancements and other priorities. SREs use Sedai to check new applications and services are on-boarding effectively, that autonomous mode is on, and reviewing optimizations executed by Sedai.

Release with Confidence

Sedai provides early feedback on the performance of every Kubernetes service release with Release intelligence. Sedai will autonomously detect new releases entering production. Sedai analyzes the release compared to the prior release on latency and error count using models that take into account seasonality and traffic patterns under load. Each release is rated on a scale of 1-10 based on the amount of deviation from the prior release (example below).



MODERATE DEVIATIONS

7.5

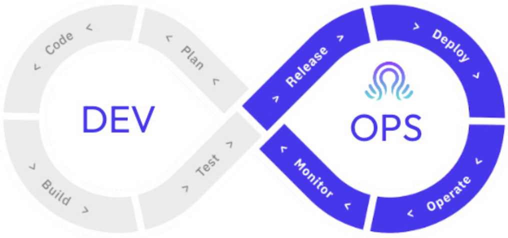
Analyzed from Monday, Dec. 6 11:58 PM to Wednesday, Dec. 8 10:12 AM

Some changes detected in resource metrics. We recommend reviewing release to avoid potential issues.

You can receive recommendations on whether to rollback a release if it doesn't meet certain criteria or parameters. In addition, scores provide signals to developers on where effort should be placed to optimize the underlying application code. You can either integrate with your CI/CD process to provide the optimal configuration established from prior releases, or when you redeploy Sedai can set your Kubernetes configuration to this optimal level.

Accelerate Innovation Velocity

Sedai users accelerate the speed at which their organizations operate the DevOps cycle. In Development stages, autonomous setting of Kubernetes configurations reduces the burden on Developers and frees them up to focus on core functionality. Release Intelligence supports both the Dev (feedback) and Ops part of the loop (faster releases). Autonomous operations reduce the time taken to perform actions in the Ops part of the cycle.



High ROI and Fast Time to Value

Sedai provides a high ROI due to a combination of:

- **Top-line benefits**. Sedai's latency and availability improvements and acceleration of the DevOps cycle contribute to customer experience and revenue for supported applications.
- **Cost savings**. Cloud cost reductions provide hard cost savings and improve gross margins for online businesses.
- **Productivity benefits**. Sedai reduces SRE/Operations workload as tasks are completed autonomously and helps developers focus on code where new releases underperform.
- **Low management effort**. Because Sedai is an autonomous system, it does not require constant monitoring by the team. Sedai executes improvements autonomously.

The Sedai team provides ROI assessments tailored for individual customer environments. Sedai also provides a fast time to value due to:

- **Start for free**. Sedai is a freemium service with a 25 pod free allowance.
- **Rapid setup.** Sedai takes about 10 minutes for initial setup[3], and has a ~2 week learning period.
- **Ease of integration**. Sedai integrates with popular tools including Monitoring and APM platforms (e.g., Datadog, Prometheus), notification providers (e.g., Slack, PagerDuty), ITSMs (e.g., ServiceNow), and runbook automations (e.g., Rundeck).

About Sedai

Sedai delivers the first autonomous cloud management platform that detects and proactively addresses potential issues in production, improving performance, ensuring availability, and managing cloud costs. Acting as an intelligent autopilot for SREs, Sedai eliminates significant toil for SREs so they can scale and increase innovation cycles. Sedai enhances your Observability and AIOps platforms by proactively preventing issues. Sedai's investors include Norwest Venture Partners, Sierra Ventures and Uncorrelated Ventures. Sedai was named a Gartner Cool Vendor in Observability and Monitoring for 2022.

Interested in learning more? Email contact@sedai.io and visit www.sedai.io for more information. Sign up for free at setup.sedai.app.

Endnotes

[1] Sedai internal analysis of Datadog Container Report https://www.datadoghq.com/container-report-2020

[2] Datadog Container Report, https://www.datadoghq.com/container-report/

[3] Longer setup may be required based on security configuration. Sedai supports VPC peering and other customized secured connectivity options which would require additional setup steps to be performed before on-boarding Sedai.

Gartner, 'Cool Vendors in Observability and Monitoring for Logging and Containers', Padraigh Byrne, Pankaj Prasad, Manjunath Bhat, Gregg Siegfried, April 27, 2022. GARTNER and COOL VENDORS are a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and are used herein with permission. Gartner does not endorse any vendor, product or service depicted in our research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.

Gartner
COOL
VENDOR
2022