



Boost Kubernetes efficiency at every layer

Leverage predictive horizontal and vertical autoscaling, with a suite of solutions designed for performance and cost-efficiency across stateless and stateful applications.

The problem

Optimizing costs while ensuring 100% availability

Managing and optimizing a dynamic K8s environment is a complex task.

DevOps teams' main challenge is to ensure application stability, maintain SLAs, and guarantee performance.

Limitations in compute and storage resource scaling push teams to over-provision, leading to unnecessary expenses and missed opportunities for maximizing cost savings.

Zesty's solution

Kubernetes optimization platform

Our K8s optimization platform provides a holistic suite of automated stack that seamlessly adapts to workload changes across every layer of your K8s environment, maintaining perfect SLAs while achieving unmatched cost reduction.

With a multidimensional automation approach and unique technologies designed for faster, more precise resource scaling and right-sizing, Zesty Kompass ensures tight alignment of Kubernetes resources with real-time application demand.

Benefits



Reduce Kubernetes cost

Cut K8s costs by up to 70% through automated resource optimization.



Enhance performance & resiliency

Ensure application resiliency and maintain SLAs with real-time autoscaling to meet any workload spike.



Unify Kubernetes optimization

Centralize all K8s optimization for compute, storage, and memory—in one platform with automation and insights.



Boost operational efficiency

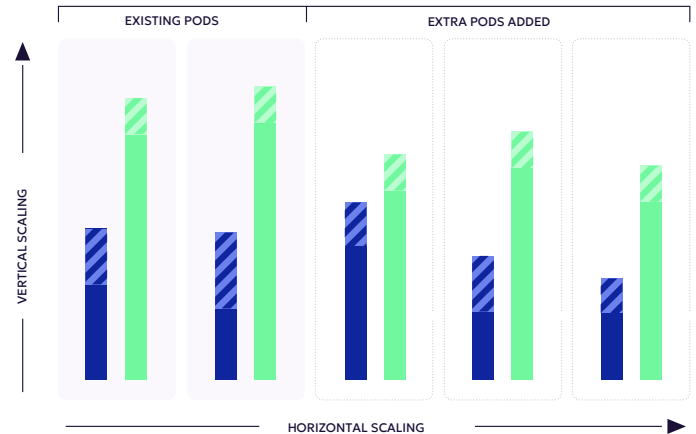
with automation that reduces manual intervention and optimizes workloads.



What makes our platform different?

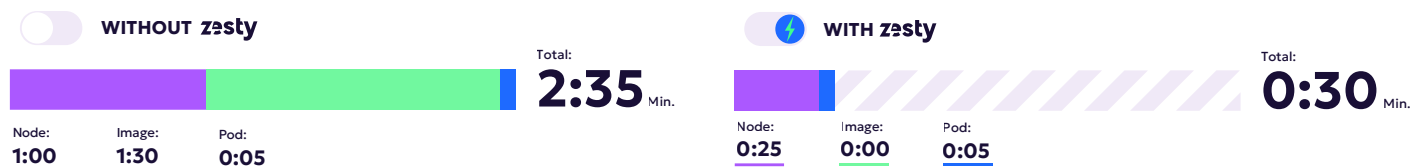
Multi-dimensional Automation

Zesty's multi-dimensional approach combines horizontal & vertical autoscaling strategies and compute & storage resource optimization, through tailored recommendations, node headroom reduction, Pod Rightsizing, Spot Protection, and PV autoscaling. This comprehensive approach ensures continuous, effortless, and durable optimization, across your entire Kubernetes environment, maintaining performance and SLAs while dramatically reducing costs.



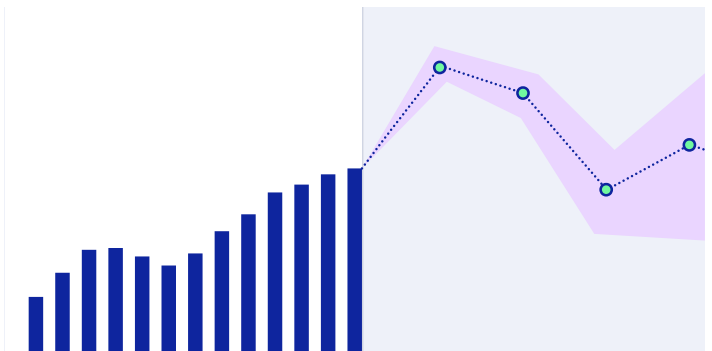
HiberScale™ technology

Zesty's HiberScale™ technology enables large-scale node hibernation, and their instant reactivation within 30 sec. to handle any spike. This unique capability enables faster scaling and maintains efficiency without risking SLAs.



Predictive Scaling

AI-powered algorithms analyze historical and real-time utilization patterns to accurately forecast workload demand, proactively adjusting resources before usage spikes occur.



Gain workload visibility & insights

Challenge

Lack of cost & utilization visibility in dynamic and complex Kubernetes environments.

Solution

Insight provides granular, real-time visibility over clusters and workloads utilization. It monitors costs and potential savings for fast, data-driven optimization.

Benefits



Get a granular view of your usage & costs



Effortlessly identify potential savings



Receive actionable recommendations

How it works

- 1 Zesty is granted IAM role access permission
- 2 Zesty installs a Kubernetes agent with a read-only-permission
- 3 You connect your CUR to Zesty platform
- 4 Zesty Insights collects your workload usage and pattern histories
- 5 Real-time data is analyzed by AI algorithms
- 6 Zesty Insights generates actionable recommendations for potential savings

WORKLOADS RECOMMENDATIONS

NAME	RAM UTILIZATION	CLUSTER	COST	RECOMMENDATIONS
search-api	<div><div></div></div>	Prod	\$184	<div><div></div><div></div><div></div></div>
item-worker	<div><div></div></div>	Staging	\$162	<div><div></div><div></div><div></div></div>
prometheu...	<div><div></div></div>	Staging	\$157	<div><div></div><div></div></div>
config-man...				

CPU



HEADROOM REDUCTION



304 with Headroom reduced
36 not suitable for Headroom reduction
909 Headroom reduction candidates

[EXPLORE SAVINGS CANDIDATES >](#)

Potential savings: \$5,010/month



Headroom reduction

zesty

Kubernetes Optimization
Platform

Scale nodes faster to enhance efficiency

Challenge

Deploying a new node in a Kubernetes environment can take several minutes. To overcome this limitation and ensure application availability during traffic peaks, DevOps teams often over-provision nodes, leading to inefficiencies and unnecessary costs.

Solution

Zesty's HiberScale™ technology accelerates Karpenter and Cluster Autoscaler scaling to under 30 seconds, eliminating the need for CPU overprovisioning and reducing costs, while maintaining SLAs during unexpected traffic spikes.

Benefits



Reduce node headroom



Cut cluster costs by 70%



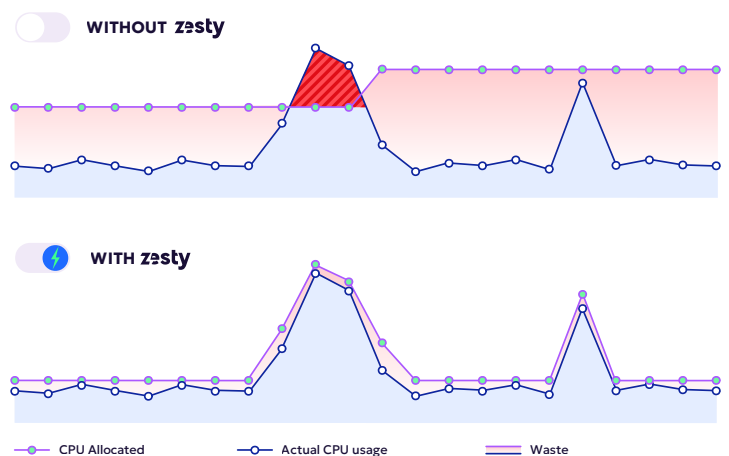
Eliminate manual operations



Preserve SLAs

How it works

- 1 Zesty is granted IAM role access permission
- 2 Zesty installs a Kubernetes agent with a read-only-permission
- 3 You connect your CUR to the Zesty agent, which starts collecting workload usage and pattern histories
- 4 Zesty analyzes your system's scaling performance, identifies unnecessary headroom
- 5 You install the K8s scaler, which automatically creates a pool of hibernated nodes, minimizing headroom
- 6 In your Kubernetes environment, you reduce your replicas to the minimum recommended level
- 7 The hibernated nodes are re-activated within 30 seconds when needed, in response to spikes or increased CPU requests



zesty supports  

Contact Us!  www.zesty.co  info@zesty.co



Automate pod rightsizing at scale




Challenge

While VPA adoption is low due to pod restarts, imprecise scaling, and HPA incompatibility, manual vertical scaling requires continuous monitoring, and can cause CPU throttling, downtime, OOM errors, or inflated costs in case of overprovisioning.

Solution

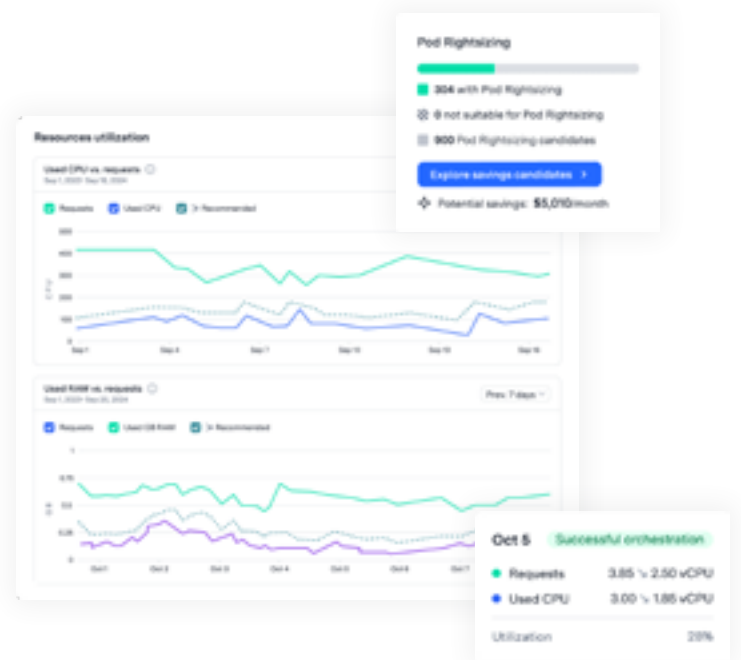
Zesty's Pod Rightsizing automatically adjusts CPU & memory resources at the container level to match precise, real-time workload usage, maximizing stability & performance while reducing costs.

Benefits

-  Ensure performance & stability
-  Reduce costs
-  Eliminate manual monitoring
-  Container-level VPA
-  No pod restarts required
-  Compatible with HPA

How it works

- 1 Zesty is granted IAM role access permission
- 2 Zesty installs a Kubernetes agent with a read-only-permission
- 3 You connect your CUR to Zesty platform & install Pod Rightsizing
- 4 Zesty collects and analyzes workload utilization and pattern histories
- 5 Zesty Pod Rightsizing generates recommendations to optimize resource allocation and generate savings
- 6 Once you activate the recommendation, Zesty scales up or down the container's request



Boost Spot instances utilization with confidence

Challenge

Spot instances utilization is risky for most applications, with AWS interrupting Spot Instances with only 2 minutes' notice, while it takes on average 5 minutes to replace an instance.

Solution

Zesty's HiberScale™ technology enables Spot instance deployment in less than one minute. You can safely and automatically extend Spot instances coverage and unlock maximum savings, with no risk of app disruptions.

Benefits



Maximize savings



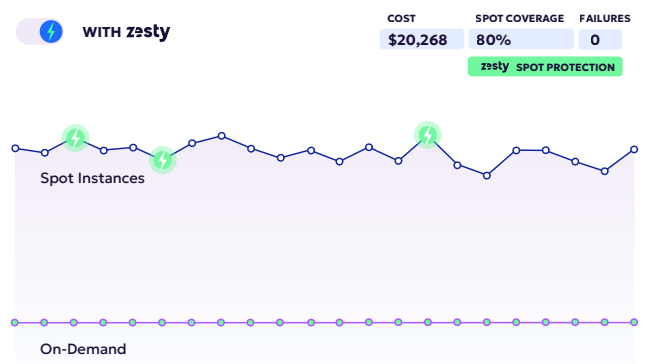
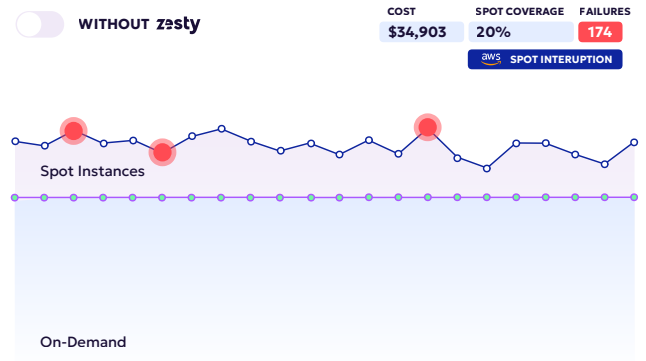
Avoid service disruptions



Eliminate manual effort

How it works

- 1 Zesty is granted IAM role access permission
- 2 Zesty installs a Kubernetes agent with a read-only permission
- 3 You connect your CUR to the Zesty agent, which starts collecting workload usage and pattern histories
- 4 You install the K8s scaler, which automatically creates a pool of hibernated nodes
- 5 The platform identifies Spot candidates' workloads and Zesty modifies Karpenter configuration so that the identified workloads run on Spot Instances
- 6 When AWS sends a 2-minute notification for a Spot instance termination, a hibernated node is re-activated within 30 seconds, and a new Spot instance is scheduled before the current Spot instance terminates.
- 7 At any moment you can migrate back your workload to On-Demand.





PV autoscaling

zesty

Kubernetes Optimization
Platform

Make your PVs scalable with no risk of downtime

Challenge

Persistent volumes in Kubernetes environments lack elasticity and scalability. To prevent storage capacity failures, DevOps teams often over-provision, which results in significant unnecessary costs.

Solution

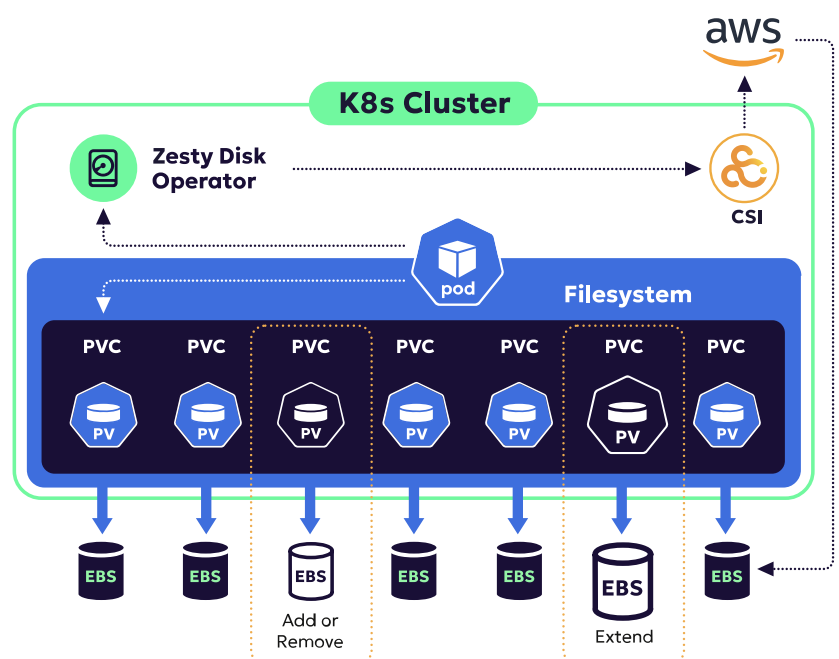
Zesty automatically scales up or down persistent volumes based on your real-time cluster needs. Our unique solution helps you save up to 70% with no risk of downtime.

Benefits

-  **Reduce** storage over-provisioning
-  **Cut** storage costs by up to 70%
-  **Eliminate** manual operations
-  **Prevent** storage capacity failures

How it works

- 1 Zesty creates a virtual filesystem which consists of several small storage volumes.
- 2 Usage metrics, instance, and disk metadata are continuously tracked and sent to Zesty's backend.
- 3 Zesty adds, removes, or extends persistent volumes without disrupting running pods, ensuring continuous operation and dynamic scalability.



zesty supports  

Contact Us!  www.zesty.co  info@zesty.co





Maximize EC2 savings with full flexibility

Challenge

As workloads fluctuate, managing discount plans is complex and financially risky. Teams must commit 1–3 years in advance, risking overprovisioning and paying for unused commitments. When underprovisioned, they rely on On-Demand, driving up spend.

Solution

Zesty's Commitment Manager maximizes coverage with dynamic Savings Plans that constantly adjust to fit your changing workloads' demand, maximizing EC2 savings while eliminating commitment risk.

Benefits



Cut EC2 Costs by 50%



Maximize commitment flexibility

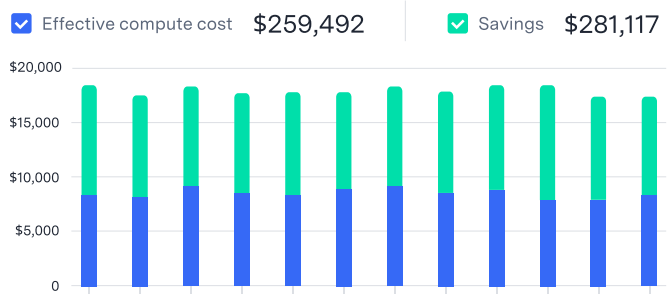


Eliminate manual effort

How it works

- 1 Zesty is granted limited IAM role permissions.
- 2 Zesty collects data on your workload usage, pattern histories, and commitment portfolio.
- 3 You define your strategy: expected usage growth, target coverage, and risk level (3Y/1Y ratio).
- 4 Our AI-based algorithm begins purchasing small units of Savings Plans, and gradually ramps up to build a flexible portfolio aligned with your strategy.
- 5 As usage increases, micro Savings Plans are added. As usage decreases, they are simply allowed to expire, ensuring maximum flexibility.

Cost over time



Coverage type breakdown

