

2024 | Industry Report Breakdown

The State of Kubernetes Efficiency: An Insider's Look

I Executive Summary

Since its March 2023 launch, PerfectScale's Kubernetes Optimization and Governance Platform has redefined and simplified how DevOps and Platform teams efficiently manage their resources with data-driven intelligence and automation. Designed to empower organizations to maintain lean Kubernetes environments that seamlessly adapt to fluctuating demands, our platform has optimized tens of thousands of nodes and hundreds of thousands of workloads and containers in its first year.

In reviewing the trends throughout our user base, and their environments, we have uncovered many interesting, and shocking, statistics. To highlight a few:

- Node-Autoscalers solution trends are changing, with Cluster Autoscaler found in only 17% of clusters, while Karpenter is found in 75% of AWS-based clusters
- Over 50% of every dollar spent on Kubernetes is waste due to over-provisioning
- Yet with all these additional resources, over 25% of workloads are at risk of performance issues caused by under-provisioning or misconfigurations.

This report analyzes the significant data collected, providing insights into cost optimization and performance metrics. We invite you to explore the critical trends and metrics that define success in today's dynamic technological landscape.

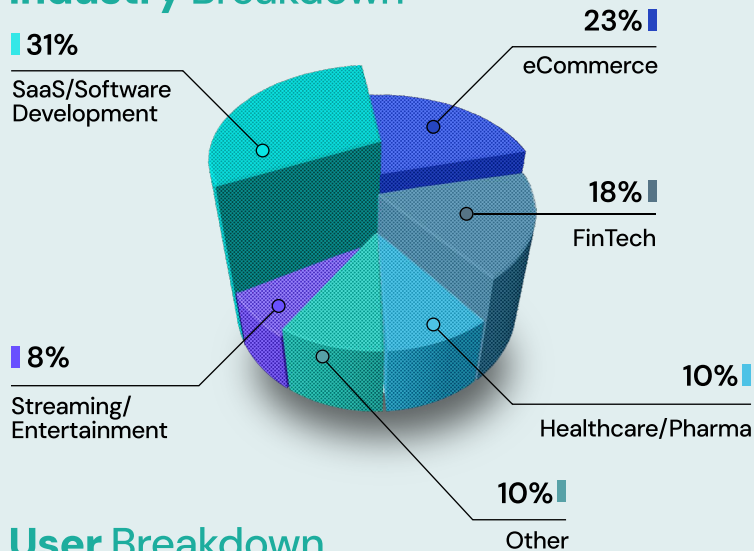
I Table Of Content

- User Demographics and Priorities
- Cluster Trends
- Cost Optimization Trends
- Performance and Resilience Trends
- Final Thoughts

User Demographics and Priorities

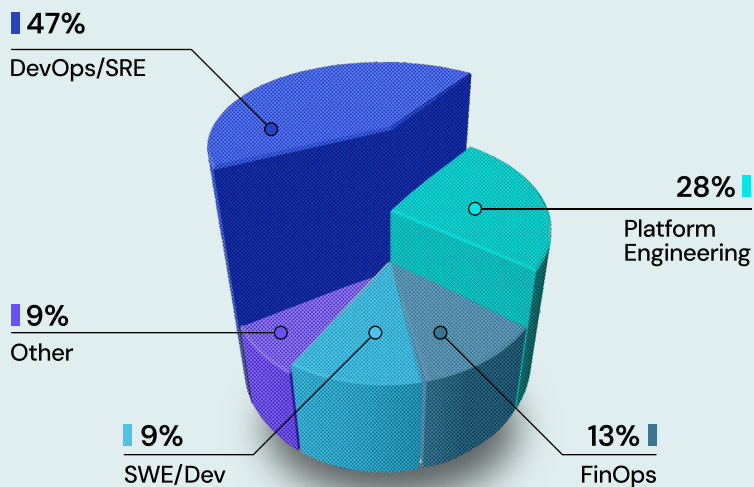
Before taking a deeper look into the statistics, let's dig a bit more into trends associated with our users.

Industry Breakdown



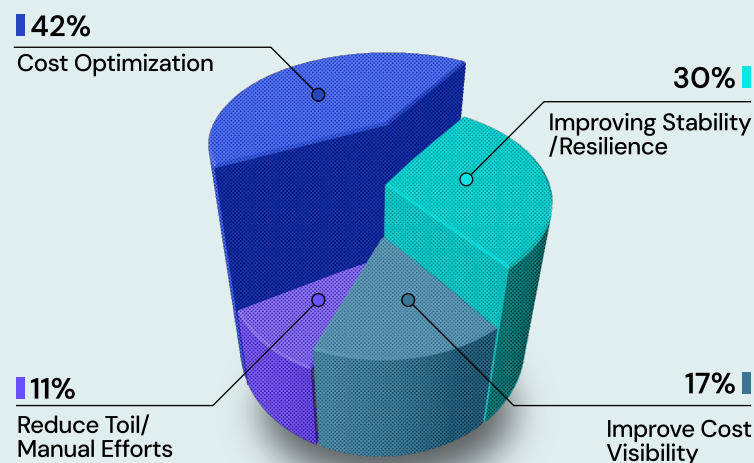
Another common theme among our users is organizations that have entered Day 2 Kubernetes operations, now concentrating on optimizing performance and cost in large-scale production environments.

User Breakdown



While DevOps and SREs remain the primary users, we have observed a significant increase in Platform and FinOps teams utilizing our data and insights. Additionally, many organizations are expanding access to our tool for their development teams, integrating optimization activities into their "shift left" strategies.

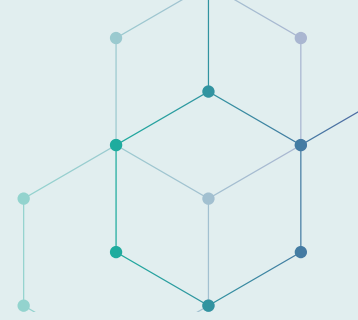
Top K8s Optimization Priorities



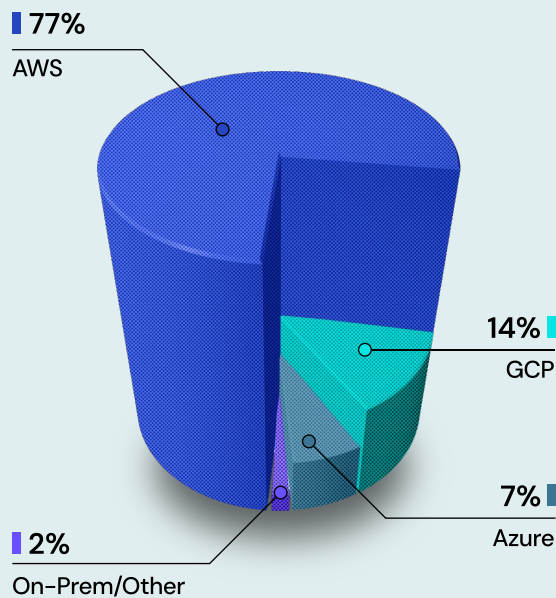
While most users prioritize reducing cloud costs and ensuring their applications are performant and resilient, many companies now view autonomous optimization capabilities as essential in selecting a solution to minimize the labor involved in optimization. Additionally, many companies still lack accurate cost visibility across their environments, making it difficult to control costs and enforce proper governance as they scale.

Cluster Trends

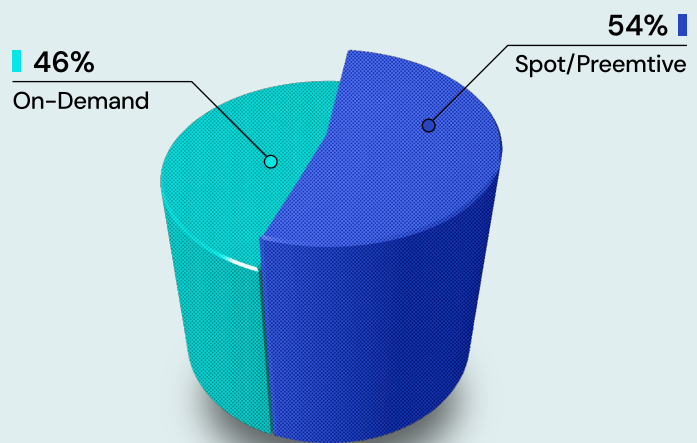
In our analysis of various user clusters, we've identified compelling trends in infrastructure and technology stacks. This section will explore these trends in detail, focusing on infrastructure developments, autoscaling practices, and Kubernetes add-on adoption across our diverse user base.



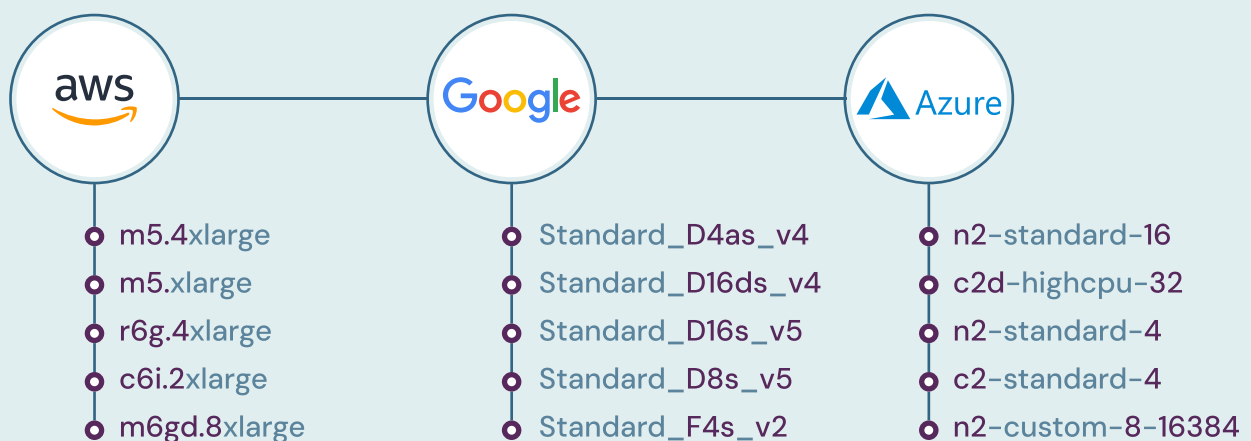
Cloud and Infrastructure Trends



Spot Vs On-Demand Instances

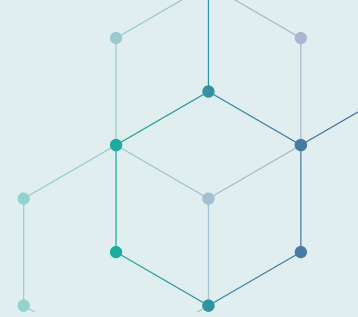


Most common Node Types



Kubernetes Add-on Trends

Besides the obvious Kubernetes add-ons – such as coredns, metrics-server, and node-driver-registrar – we see the following trends:



Prometheus is present in around 72% of the clusters – making it the most popular CNCF tool (not directly related to K8s) to run on K8s.



Grafana comes in a very close second, with almost 70% of the cluster running at least one instance –to visualize the metrics collected by Prometheus.



Interestingly, the full **kube-prometheus-stack** is present only in about 45% of all clusters. Which means the rest are doing some custom deployment of the Prometheus–Grafana combination.



External-secrets operator is also widely present with around 44% of the clusters integrating it with secret managers. In only 19% of the clusters, we are observing **Hashicorp's Vault**, which means that the rest are probably using a managed service such as AWS Secret Manager.



cert-manager is present in 36% of the clusters even though most of the cert-manager we found runs on cloud-managed clusters and can use the native certificate integration. Such a presence means that in 36% of the cases, organizations prefer to save some money on certificates most probably generated from **Let's Encrypt**.

Autoscaling

Horizontal Pod Autoscaling: Scaling pods up and down to meet demand



• Horizontal Pod Autoscaler (HPA) is the predominate autoscaler found in 97% of clusters.



• KEDA is providing event-driven autoscaling in 36% of clusters

Cluster Autoscaler: Scaling the number of machines up and down to meet demand



- Karpenter is slowly but surely replacing cluster-autoscaler and is currently present in 75% of all AWS-based clusters
- Cluster-autoscaler is present in around 17% of clusters, with some clusters appearing to have both

I Policy Controllers



We found instances of [Kyverno](#) in 21% of clusters, but we didn't find OPA used alone for admission control. These findings suggest that most clusters still need to enforce policy standards.

I Databases, caches, message queues



- [Redis](#) is the most popular cache/queue/in-memory-db solution, appearing in almost 60% of the clusters



- [RabbitMQ](#) is far behind Redis, appearing in only 24.5% of the clusters



- [Kafka](#) with zookeeper seen running in 21% of the clusters

I Service Mesh

[Istio](#) is present in 23% of the clusters. [Cilium](#) is present in around 10% of all clusters. We're guessing it's more for observability than service mesh management in some enterprises. We didn't find any instances of [Linkerd](#) in the clusters we manage.



I GitOps

While it may seem every enterprise is doing [GitOps](#) these days, we only see [ArgoCD](#) running in 6% of clusters, although one instance of Argo can serve multiple clusters in an environment, and barely any [Flux](#) instances. The reality is that GitOps is still the domain of early adopters.



Cost Optimization Trends

We have identified a prevalent pattern of excessive provisioning of Kubernetes resources across various environments. A frequent rationale for this inefficiency is fear—many organizations worry that pushing for optimization might lead to performance latency and availability issues, overshadowing the potential cost savings.

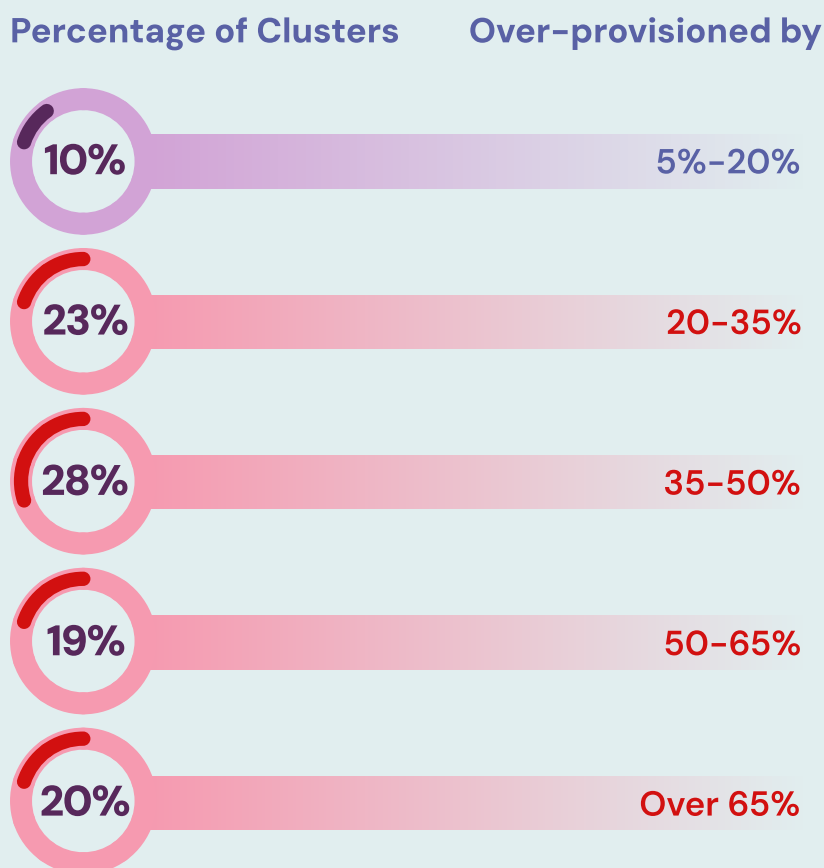
PerfectScale addresses these concerns head-on. Leveraging AI-driven insights and automated processes, our platform minimizes risks while enhancing efficiency. This allows organizations to confidently streamline their operations, secure in the knowledge that they can reduce costs without sacrificing performance or reliability.

Let's explore further into how these cost optimization strategies are reshaping industry standards.

Cluster Over-Provisioning

More than 90% of clusters are over-provisioned, meaning that at least 20% of their CPU and Memory resources are provisioned but not used.

Percentage of Over-provisioned Clusters



When you look at the total average cost of cloud overprovisioning, **51 cents of every dollar** spent on cloud services is wasted by overprovisioned resources!

To break this down further, roughly 38 cents out of every dollar is spent on over-provisioned CPU, while 13 cents is spent on over-provisioned memory.

Wasted Resources vs. Idle Resources

Over-provisioning can be categorized into two main types:

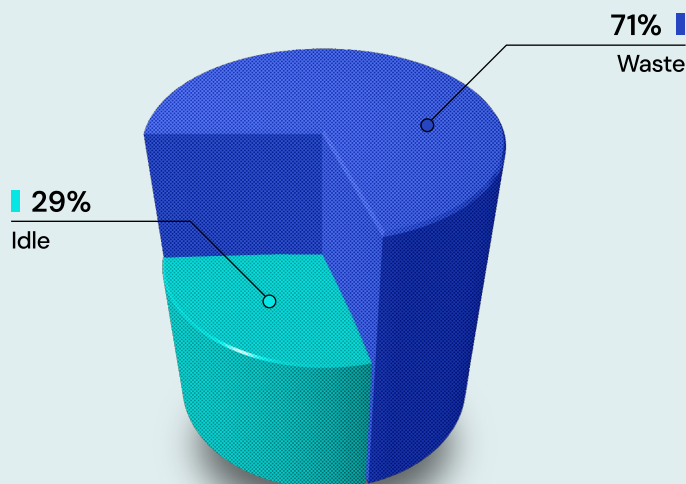


Wasted Resources This refers to resources allocated to workloads that aren't actively in use. Waste often stems from developers requesting more resources than necessary to support the usage load of their deployments. This not only causes excessive costs, but impacts the efficiency of auto-scaling configurations, exacerbating waste when scaling up during peak loads.

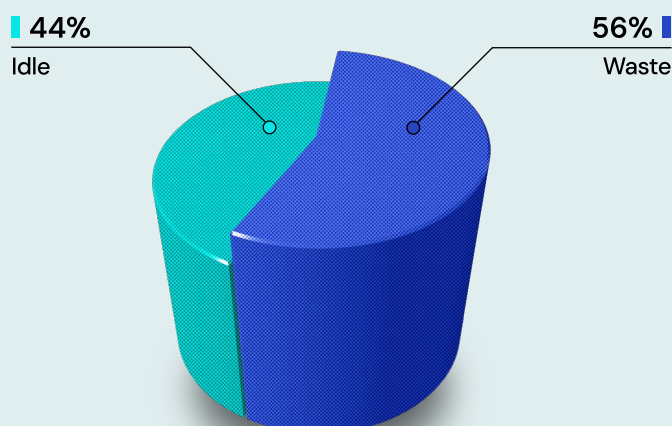


Idle Resources This involves purchasing node space that remains unused. While having some idle space in Kubernetes clusters is acceptable as it allows for horizontal pod scaling, excessive idle space leads to unnecessary budget wastage. Selecting the right instance types to support dynamic Kubernetes environments remains a major challenge across the industry.

Over-Provisioned CPU



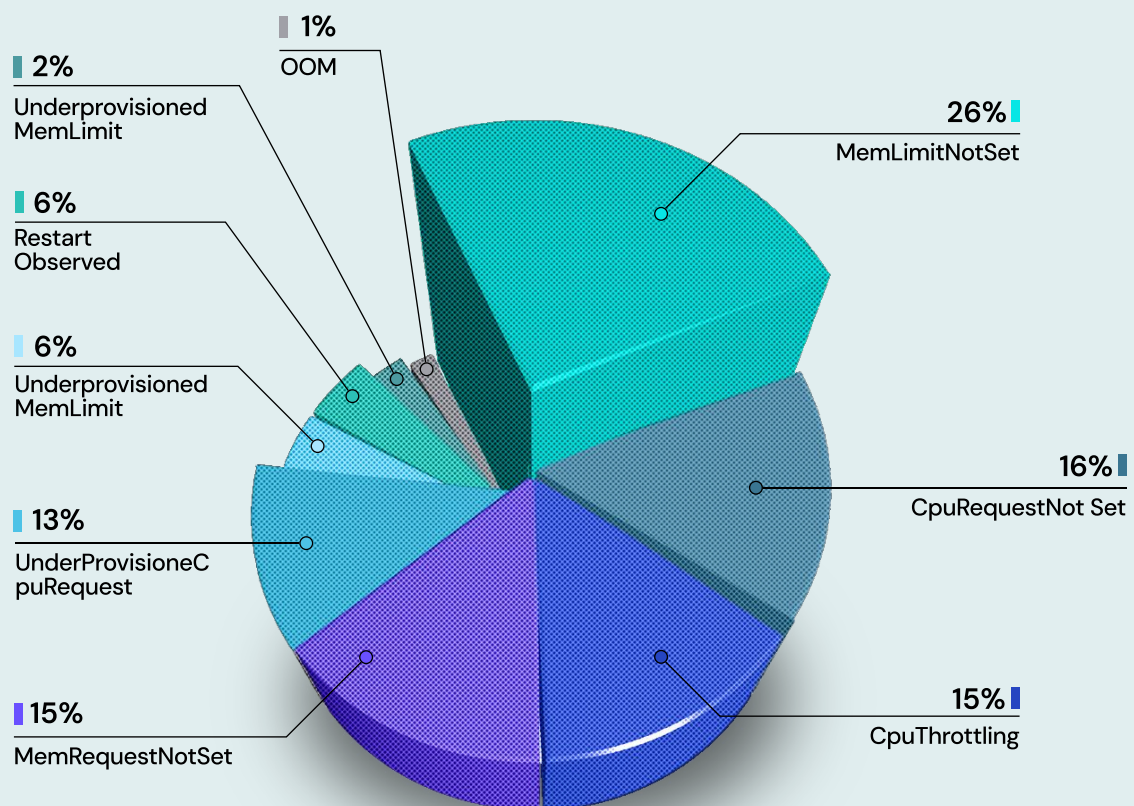
Over-Provisioned Memory



Performance and Resilience Trends

Despite many developers over-provisioning to guarantee deployment reliability, a startling trend has emerged: about one-quarter of workloads are either under-provisioned or suffer from resource configuration errors. These issues can cause significant latency, and in more severe cases, lead to evictions that result in availability outages, drastically affecting crucial business metrics.

Next, we will take a look at the most prevalent issues observed across our users' environments, exploring their implications and the strategies to mitigate them.



Misconfigurations continue to plague Kubernetes clusters, leading to issues including:

- Mem Request not set
- Mem Limit not set
- Cpu Request not set

There are also performance issues to monitor, especially:

- OOM
- CPU Throttling
- Evictions

In a world where teams constantly strive for 99.99 percent availability, resiliency issues can severely impact their ability to maintain optimal service levels. With constant environmental changes, optimizing for cost and resilience is essential.

Final Thoughts

Monitoring and remediating Kubernetes efficiency is still a growing practice in cloud and FinOps teams. Enterprises relying on Kubernetes must be proactive and make optimization practices part of their operational playbooks. Here are some practices that we recommend you integrate into your Kubernetes operations if you haven't already done so:

- **Continuous Optimization** – To reduce risk and proactively optimize environments, leverage tools with AI-guided recommendations and automation to continuously tune the cost and performance of your environment
- **Policy Control** – Introduce policy controllers such as OPA Gatekeeper or Kyverno to ensure container resource defaults comply with the business goals.
- **Revision Awareness** – Continuously review and adapt resource allocations for new application revisions and demand fluctuations.
- **Pod2Node Alignment** – Align pod allocations with provisioned node resources to optimize application performance. Detailed scheduling rules based on container requirements can help in efficient resource allocation.
- **Smart Node Selection** – Balance the number of nodes against workload performance goals and availability requirements. The type and capacity of nodes play a crucial role in this optimization.
- **Placement Constraints** – Utilize taints and tolerations to control node placement, ensuring pods are assigned to the most suitable nodes.
- **Optimize Autoscaling** – Fine-tune the autoscaling configurations for just-in-time node provisioning and node density control.

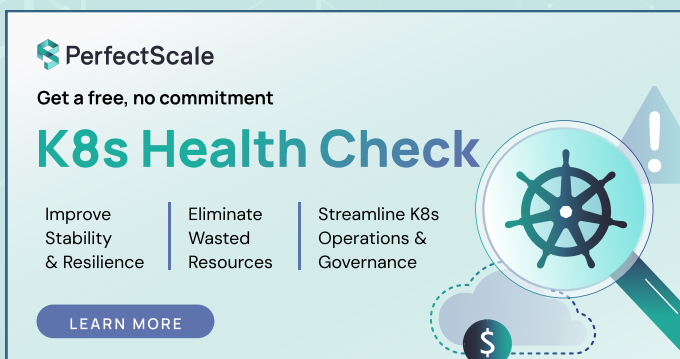
The PerfectScale Approach:

PerfectScale allows our users to continuously optimize their environments safely and autonomously. Our data-driven approach analyzes the dynamic usage patterns of the environment and provides the best possible configuration to ensure peak performance and constant availability while reducing cloud costs.

One last common trend we find is many organizations don't know where to start when it comes to optimizing their environments. To help address this, PerfectScale is offering a complementary Cluster Health Check. During this free, 30-day program, our Optimization Specialists will help you leverage our solution to:

- Improve your environment's stability and resilience
- Eliminate wasted and idle resources
- Streamline your K8s Operations and Governance practices

Learn more about our Cluster Health Checks, or book an initial consultation, [here](#).



The banner features the PerfectScale logo at the top left, followed by the text 'Get a free, no commitment'. Below this is the main heading 'K8s Health Check' in large, bold letters. To the right of the heading is a magnifying glass icon with a ship's steering wheel inside it. Below the heading, there are three columns of text: 'Improve Stability & Resilience', 'Eliminate Wasted Resources', and 'Streamline K8s Operations & Governance'. At the bottom left is a 'LEARN MORE' button. At the bottom right is a circular icon with a dollar sign and a cloud.

PerfectScale
Get a free, no commitment

K8s Health Check

Improve Stability & Resilience | Eliminate Wasted Resources | Streamline K8s Operations & Governance

LEARN MORE

We hope you found the State of Kubernetes Efficiency report useful and enlightening. As we continue to push the boundaries of Kubernetes optimization and Day-2 Operations, we will keep sharing our insights into the developing and shifting trends impacting the Cloud Native ecosystem.

Thanks, The PerfectScale Team