PULSEMETER

Techstrong Research

Kubernetes Infrastructure Optimization Strategies

SPONSORED BY NetApp

RISING CLOUD COSTS and demanding business conditions are putting additional pressure on software and cloud infrastructure teams to find new ways to optimize and improve the efficiency of how Kubernetes clusters are deployed and consumed. As cloud, DevOps and platform engineers work to further reduce redundancies and inefficiencies, the complexities of Kubernetes environments make these tasks that much harder.

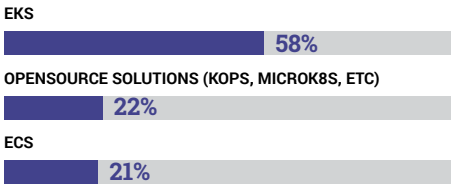
With the right solutions designed for Kubernetes-specific needs, a number of resources can be controlled and consumed in a more efficient and reliable way within the cluster. For example, the pods, or containers within pods, can be adjusted to consume a maximum amount of CPU and memory. Namespace control enables cluster management or resource management to be more equitably distributed among different users by granting privileges. Other methods include resource quotas and scaling to the right size for pods and clusters in order to avoid over-provisioning.

The Kubernetes API serves as a main conduit for automating resource management. With it, pods, services, applications, and more can be created, deleted, and managed as a way to optimize Kubernetes resources, both on the cloud and on-premises. But the Kubernetes API is just the starting point and not every organization is prepared to invest the time or their talent in creating one-off solutions to their resource management challenges, nor dedicate the staff to support those internally developed tools.

During 2024, Techstrong Research polled our cloud, cloud-native, platform and DevOps communities about the practices and challenges of Kubernetes management and optimization. The respondents expressed a range of challenges and approaches. Nearly half (47%) of respondents indicated 50-100% of their applications are containerized while 58% indicate they use AWS

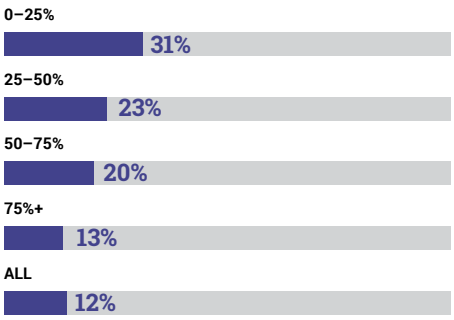
What is the primary K8s/ container orchestration or management service you use on AWS?

No surprise here, as the majority (58%), relies on AWS's EKS designed for Kubernetes environments. To a lesser extent, 22% use open-source options to manage Kubernetes and containers, including kOps, MicroK8s, etc.



Approximately what percentage of your AWS workloads are containerized?

AWS workloads are containerized to different degrees: 31% of respondents use containers for 0 to 25% of their workloads while 13% use containers for 75% or more of their workloads and 12% for all their workloads.



EKS, followed by AWS ECS (21%) and open source kubernetes (22%) options. Kubernetes complexity and skill requirements were rated as the biggest challenge by respondents.

TECHSTRONG RESEARCH ANALYST VIEW

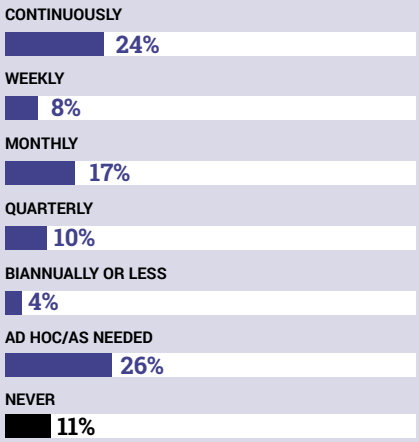
Implementing and operating applications using a cloud-native architecture, which includes microservices, containers, and Kubernetes container orchestration, accelerates software delivery.

Microservices and Kubernetes provide greater flexibility to distribute, scale and optimize workloads across clusters, cloud providers, and data centers. With these capabilities comes increased complexity and many more variables and decisions required to optimize Kubernetes. Costs can easily skyrocket due to overprovisioning, changing application needs and limited skill sets within the organization.

Our research shows that a third of respondents automate provisioning, which can aid in standardizing platforms and configurations. A modest few (17% or below) automate rightsizing and software updates, leaving significant gaps and opportunities for optimization and further automation.

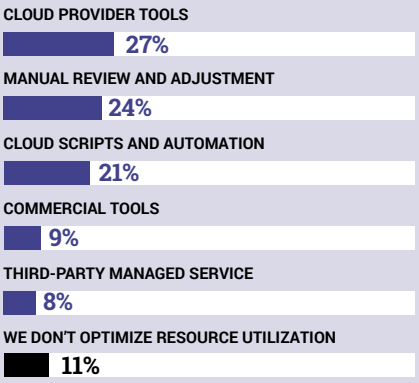
Approximately how often does your team optimize Kubernetes infrastructure resources?

Organizations optimize Kubernetes infrastructure resources with varying frequencies, ranging from continuously (24%) to never (11%), while 26% initiate the process on an as-needed basis.



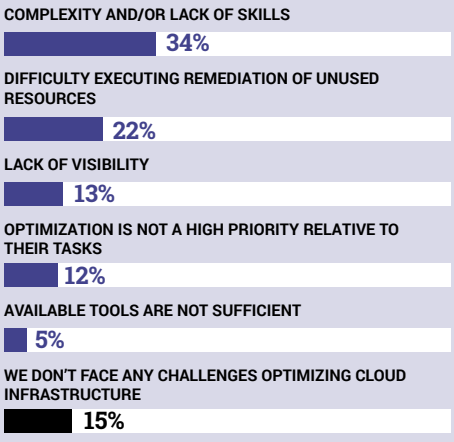
How do you optimize Kubernetes resource utilization?

Kubernetes resource utilization with cloud-provider tools is the top choice (27%), but more collectively rely on manual reviews and processes (24%) and custom scripts and automation (21%).



What is the biggest challenge you face in optimizing Kubernetes infrastructure?

Kubernetes complexity issues represent a key concern, including challenges associated with both complexity and lack of skills available (34%) and difficulties in remediation of unused resources (22%).



With the high demand for Kubernetes skills and rapid advancements in AI, operations teams need to align with key vendors across four vectors; resource management and optimization (including scaling), AI/ML/GenAI workload optimization, GPU/IPU hardware scheduling and comprehensive cost containment.

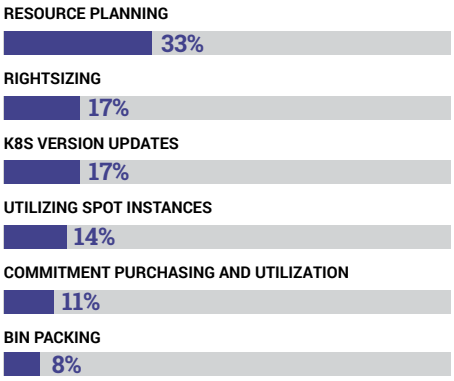
Our study also showed that 24% of our respondents continuously optimize Kubernetes while the measures taken by another 26% are ad-hoc. As cloud workloads become increasingly dynamic and as AI/ML use as part of our applications grows, operations teams are advised to shift from ad-hoc processes to continuous optimization through automation, continuous resource management technologies, and ongoing improvements offered as AI/ML/GenAI matures in this market space.

KEY TAKEAWAYS

1. Kubernetes resource management is critical amid rising cloud costs and budget pressures.
2. Most businesses are optimizing K8s infrastructure resources infrequently or on an ad hoc basis, presenting significant opportunities to reduce costs.
3. Infrastructure complexity and the difficulty of manually optimizing resources are challenging many teams' efforts to drive resource efficiency.
4. Adoption of automation is patchy, and is lagging for infrastructure optimization tasks, heaping manual work and pressure on operations teams.
5. AI/ML-driven automation offers a reliable route to continuous resource optimization, reducing costs and improving efficiencies while cutting through complexity.

What K8s infrastructure-related tasks has your team automated?

Resource provisioning, rightsizing, and Kubernetes version updates represent the most utilized automation process for Kubernetes infrastructure management: resource provisioning (33%), rightsizing (17%), and Kubernetes version updates 17%.



What best describes your role?