SPONSORED CONTENT | WHITE PAPER

Oracle Cloud Infrastructure: Unrivaled price/performance leader with proven expertise **for Kubernetes and serverless**

CIO

SPONSORED BY

ORACLE | AMD

As cloud computing has become mainstream for workloads of all sizes, organizations are increasingly looking to go beyond the basics of lifting and shifting on-premises workloads to building new applications with cloud-native constructs. By 2025, Gartner estimates, over 95% of new digital workloads will be deployed on cloud-native platforms, up from 30% in 2021.

Cloud-native computing leverages software containers, microservices, and serverless functions to support applications built from loosely connected independent components. This modularity enables the development process to shift from coding to a high-speed software development life cycle while supporting more accessible updates, maintenance, and high availability. Containerized applications can run anywhere and scale automatically across clusters of machines.

In the realm of cloud-native computing, Kubernetes reigns supreme as the dominant orchestrator of software containers. A staggering 84% of companies are currently evaluating Kubernetes, with 66% having already adopted it, according to the Cloud Native Computing Foundation 2023 survey.[1] Kubernetes automates the deployment, scaling, and operation of containers across clusters of servers, ensuring high availability and efficient resource management. It's a powerful tool for deploying cloud-native applications at scale, all managed through a single pane of glass.

Remote work leader GoTo shifted to Oracle Cloud Infrastructure (OCI) to accommodate a spike in activity during the COVID-19 pandemic. GoTo improved fault tolerance, availability, and workload resilience by moving 70% of its products to OCI and adopting a multicloud strategy. OCI's global regions reduced customer latency, while Oracle Cloud Infrastructure Kubernetes Engine streamlined infrastructure management and optimized resource scaling. The company also achieved a 50% reduction in infrastructure costs and enhanced developer productivity through automation and better tooling. The move boosted innovation, customer satisfaction, and operational efficiency while delivering massive advantages in automatic scaling for dynamic workloads across multiple cloud deployments.

1 Cloud Native Computing Foundation, CNCF 2023 Annual Survey

8x8, a provider of integrated customer engagement and communication services, also moved its video meeting services and core workloads to OCI in 2020, achieving significant cost savings and performance improvements. More than 300 microservices orchestrated by OCI Kubernetes Engine and OCI Functions streamlined development and management. The Kubernetes-based infrastructure enabled 8x8 to enhance scalability and efficiency, and OCI Observability and Management Platform improved monitoring and troubleshooting capabilities, further ensuring optimal customer experiences while enhancing cost efficiency.

What to look for in a Kubernetes cloud platform

Organizations looking to scale up cloud-native development should seek cloud providers that can accommodate known and unanticipated needs. Among the essential capabilities to consider:

- **Global reach**. Organizations should have the flexibility to shift workloads across regions for reasons such as performance, localization, and compliance. Look for providers with regional pricing consistency to make the shifting of workloads economically seamless.
- **Multicloud compatibility**. Organizations should be able to use a multicloud approach to maintain maximum flexibility. Look for prebuilt integrations that interoperate with major cloud infrastructure and platform services to access best-of-breed tools.
- **Hybrid cloud option**. More than 70% of organizations use a combination of public and private cloud resources. One of the benefits of using Kubernetes — and containerized workloads in general — is the ability to deploy in heterogeneous environments.

A range of options

Over time, most organizations want various deployment options to accommodate their needs for security, control, cost, and other factors. A robust cloud solution should provide self-managed, fully managed, and serverless deployment options.

Managed nodes, a feature of some cloud-based Kubernetes services, are useful in scenarios that need to accommodate various use cases. The cloud provider assumes responsibility for ensuring continuous availability, applying updates, maintaining application program interface availability, and scaling as needed.

Some providers offer **self-managed** nodes, which allow for custom configurations and specialized infrastructure for applications such as high-performance computing and distributed training of large language models (LLMs). They reduce the operational burden, lower the total cost of ownership, and decrease time to productivity.

A **serverless Kubernetes option** abstracts infrastructure management and automatically scales Kubernetes nodes and capacity, based on workload requirements. Users don't need to size, provision, or upgrade nodes manually; service providers handle these details transparently. Customers pay only for the resources they consume, and much of the burden of patching, securing, and upgrading servers is eliminated.

Oracle's managed Kubernetes offering

Oracle Cloud Infrastructure Kubernetes Engine (OKE) is a managed Kubernetes service that simplifies the operations of enterprise-grade Kubernetes at scale. It uses virtual nodes to deliver a serverless Kubernetes experience that eliminates the need for customers to manage infrastructure.

OKE virtual nodes are ideal for batch-processing applications. They support continuous integration/deployment workflows for automatically building, testing, and deploying containerized applications. They are also well suited for hosting web applications, as variations in activity can be handled automatically.

OKE provides customers access to multiple value-added extensions.

- Cluster autoscaling and the Istio service mesh for distributed microservices-based applications provide optimized traffic management, security, observability, and policy enforcement.
- Enhanced cluster security features include workload identity and access controls that limit the number of Kubernetes pods that can make application programming interface (API) calls and access cloud resources.
- Network security groups for Kubernetes clusters limit cluster access to authorized users. Optional private clusters can restrict access to the Kubernetes API endpoint.
- OKE applies 256-bit encryption to block volumes, boot volumes, and volume backups at rest and in motion on an internal and highly secure network.

Container image scanning, signing, and verification ensure that application images are free of serious vulnerabilities. Administrators can sign images in the Oracle Cloud Infrastructure Registry before deployment.

Scalable worker nodes adjust the number of nodes, based on workload requirements. Node cycling provides for automated updates of worker nodes without any application downtime.

"The use of OKE streamlined our computational resources as well as spending on servers, antivirus, and licenses, among other items," says Bruno Lopes, CIO and CTO at JSL. "Consequently, the company achieved a 70% cost reduction, and these saved funds were redirected toward new technology projects, undoubtedly fueling our growth."

Kubernetes' value in AI

With investments in artificial intelligence (AI) surging,[2] Kubernetes has emerged as a critical tool for supporting AI and machine learning (ML) model training and inferencing. Kubernetes excels at coordinating numerous containerized applications across widespread systems, an architecture widely used in ML. It also supports a vibrant community of open source contributors, commercial providers, and users who are evolving the AI ecosystem.

For example, dynamic resource allocation ensures that applications get the resources they need when they need them. Automatic scaling reduces the risk of overprovisioning. The Kueue cloud-native job queueing system was built specifically for ML tasks. Upper-stack tools such as Kubeflow manage ML workloads.

AI training often requires significant computational resources, but they can vary during different stages of training and inference. Kubernetes can automatically scale resources up or down, depending on demand. Containerized AI applications ensure consistency across different environments, making it easier to deploy, update, and roll back applications. Kubernetes also supports distributed training frameworks such as TensorFlow, PyTorch, and Horovod to enable seamless scaling of training jobs for parallel processing across a cluster. By automatically detecting and recovering from node failures, it ensures that AI workloads run smoothly without manual intervention on a consistent deployment platform that spans different cloud providers and on-premises environments.

2 Goldman Sachs, AI investment forecast to approach $200 billion globally by 2025, August 1, 2023

When selecting a cloud-based Kubernetes platform for AI and ML development, look for access to specialized hardware such as NVIDIA H100, A100, and A10 Tensor Core GPUs with full compute capabilities, cross-cloud integration, and access to powerful data management platforms.

AI and ML development can also be accelerated by using intelligent coding tools. Oracle Code Assist, an AI code companion that boosts developer velocity and enhances code consistency, is optimized for Java, SQL, and OCI. Powered by LLMs running on OCI, Oracle Code Assist provides developers with context-specific suggestions that can be tailored to an organization's best practices and codebase.

What's unique about Oracle Cloud Infrastructure

Oracle is the unrivaled price/performance leader. For example, Oracle's pod/node pricing per virtual CPU hour costs between one-third and one-quarter of what other major public cloud providers charge. Oracle virtual serverless nodes cost more than 60% less than other hyperscalers.[3]

OCI provides consistent, predictable, and competitive pricing across all regions globally. It supports the most demanding workloads and offers flexible shapes that enable customers to allocate only what's needed for each workload. Oracle offers the smallest minimum serverless pod size of any cloud provider. Pods can scale up to 120 virtual CPUs, with both x86 and Arm options available. The maximum serverless pod size is between eight and 32 times as large as that of other hyperscalers.

OCI's Kubernetes expertise is on par with that of other major cloud providers. Oracle was an early adopter of Kubernetes as its platform for building, training, and deploying AI services. It has migrated existing AI services and deployed new AI services on OKE, demonstrating that it has the experience and the expertise to deploy AI infrastructure on Kubernetes at scale. OKE was used in all stages of building and delivering the newly released OCI Generative AI.

OKE is the platform of choice for thousands of enterprises of all sizes across all industries globally. Oracle runs more than 100 mission-critical cloud services and software-as-a-service (SaaS) applications on OKE.

3 Cloud Economics, Oracle.com

Better with AMD

OCI and AMD have a long history of partnership, and AMD EPYC processors are fully supported by OCI's Kubernetes offerings. Customers using OCI and AMD for cloud-native deployments include Wiz, one of the fastest-growing SaaS security companies, and JSL, Brazil's leading road logistics company.

To learn more about OCI, visit **oracle.com/cloud** and **oracle.com/cloud/cloud-native**

CIO

SPONSORED BY

ORACLE | AMD