ORI

Scale into the *future of AI* inference

Ninety-six percent of AI spending goes towards inference workloads†, yet running them on virtual machines often leads to operational overhead, and dedicated inference platforms come with a steep price tag. Enter Ori Serverless Kubernetes. Ori Serverless Kubernetes is a game changer for AI inference, making it easy to deliver your ground-breaking AI models to a wider audience. With powerful scalability, easy management, and transformative cost-efficiency, Ori Serverless Kubernetes enables you to focus on innovation—not infrastructure.

Scale your models, *not your costs*

Scale seamlessly

Automatically scale your inference clusters up or down based on demand, without the hassles of managing nodes and node pools. Ori will manage your clusters and provision load balancers so you can focus on serving your models as widely as possible. Our pay-per-minute pricing helps you keep your costs predictable as you scale.

Powerful GPUs for higher throughput

Say goodbye to cold starts. Pick a high-performance NVIDIA® GPU—H100, L40S, or L4—and set up a cluster in under a minute. These datacenter grade GPUs deliver the performance you need to provide powerful concurrency and low-latency processing. No waiting for GPUs and no approvals needed.

Flexibility of Vanilla Kubernetes

Experience the flexibility of Vanilla Kubernetes with full access to the control plane via kubectl and multiple namespaces, unlike proprietary Kubernetes from other providers.

† https://menlovc.com/2023-the-state-of-generative-ai-in-the-enterprise-report/

Why Developers choose *Ori Serverless K8s*

> “Ori’s Serverless Kubernetes platform has been crucial in allowing us to dynamically scale our inference workloads while reducing costs. With Ori, we’ve been able to focus on growing our platform and making AI model inference accessible to everyone.”

Aditya Rajagopal
Co-Founder, nCompass Technologies | YC W24

Ori Serverless Kubernetes
Compared to other providers

Per minute billing
Compared to hourly billing.

Works *out-of-the-box*
Compared to proprietary K8s requiring refactoring.

Fast (<5s) startup times
Compared to slow, cold starts.

Pay only for what *you use*

With Ori Serverless Kubernetes, you’ll never have to worry about over- or under-provisioning virtual machines. Optimize your GPU budget by paying only for the resources you actually use.

GPU Resources	Price ($/Hour)
NVIDIA H100 80GB SXM	5.13
NVIDIA H100 80GB PCIE	4.17
NVIDIA L40S 48GB	2.16
NVIDIA L4 24GB	1.14

Other Resources	Price ($/Hour)
MEMORY (MB)	0.000005
CPU (1/100)	0.0001
LOAD BALANCER	0.01809

Run inference on Ori Serverless Kubernetes

Spin up a serverless GPU cluster on Ori today! If you’d like to have a conversation about using Ori for your AI business, contact our sales team.

ORI

https://www.ori.co/contact