ORACLE
AI World

AI and Developer Hot Topics

THR1746

Kay Singh

Product Manager, OCI Kubernetes Engine (OKE)

Agenda

- Why Kubernetes for AI workloads?
- Customer use-case (Faire)
- Serverless Workloads

AI Models, Training, and Inferencing

- **AI models** are programs that learn from data to generate output
- Model **training** is the process of teaching a model to recognize patterns and make predictions by feeding it data and adjusting its parameters until it produces the desired behavior
- Trained models are used to make **inferences**: generating output on novel input data based on its training
- Examples of models include *Large Language Models (LLMs), Image and video recognition, Anomaly detection, Recommender systems, Predictive modeling and forecasting, Robotics and control systems*

Kubernetes is *the* Platform for AI/ML Workloads

SoundHound

Powering speech recognition for Mercedes and Pandora

FAIRE

Online wholesale marketplace connecting independent brands with retail stores globally.

pipefy

AI-powered workflow automation for business efficiency.

cohere

Infuse AI into apps using Cohere large language models

inworld

Embed LLMs, narratives, and non-playable characters that evolve with each action for interactive gaming experiences

Fireworks AI

AI platform to collaborate and share, fine-tune, and run large language models

Adept

Train large-scale AI/ML models faster and more economically

mosaicML

Reduce the cost of training neural networks

Kubernetes is *the* Platform for AI/ML Workloads

- Scalable resource orchestration
- Portability
- Resource management
- Ecosystem
- Community

Kubernetes is *the* Platform for AI/ML Workloads

Model Development

Model Training

Model Inferencing

Why OCI Kubernetes Engine (OKE) for your AI workloads?

Price Performance

Flexible Infrastructure

GPU Optimizations

Enterprise Ready

- Fully-managed control plane and simplified infrastructure management
- Integration with other OCI services
- Optimized for AI workloads
- Security and regulatory compliance
- Enterprise ready: deploy massive clusters of GPUs, CPUs
- Support for both Nvidia and AMD GPUs

FAIRE

The Future of Retail AI: Scaling with GPUs on OKE

What is Faire?

- Faire connects hundreds of thousands of retailers with 100K+ brands from over 120 countries
- Founded by former Square employees
- Raised $1.5B+ from top VCs like Sequoia and Y Combinator
- Ranked on 2025 Fortune Future 50 list
- Leader in B2B wholesale retail

Fond Shop | Lymm, UK

Where do they use AI/ML?

Classical ML

- Credit risk
- Shipping cost estimation
- Search & Discovery
  - Retrieval
  - Ranking
  - Personalization
- Ads

GenAI

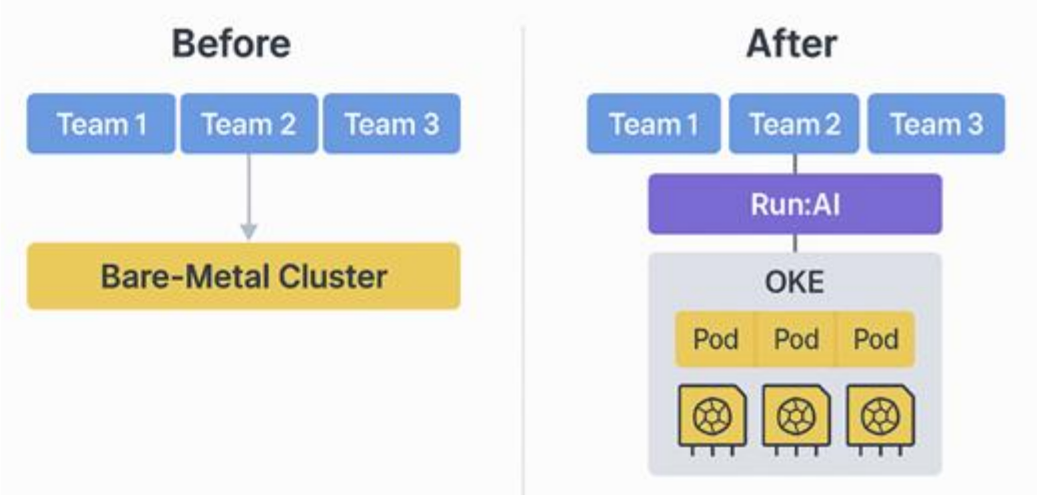
- LLMs
  - Translation
- Agents
  - Customer facing
  - Internal

Challenge： Unstable GPU Cluster

One job to take them all down

- Bare metal access to GPU cluster
- One bad job could take down the entire cluster
- Poor session management
- Submitting jobs to GPU cluster requires SSH
- Incidents are expensive
  - Productivity loss
  - 9 incidents over a 1-yr period

OKE + run:AI



- API/CLI access to GPU cluster
- Containerized workloads + failure isolation
- Elastic workloads
- Jobs can be monitored via UI/CLI
- 1 incident over a 9 month period

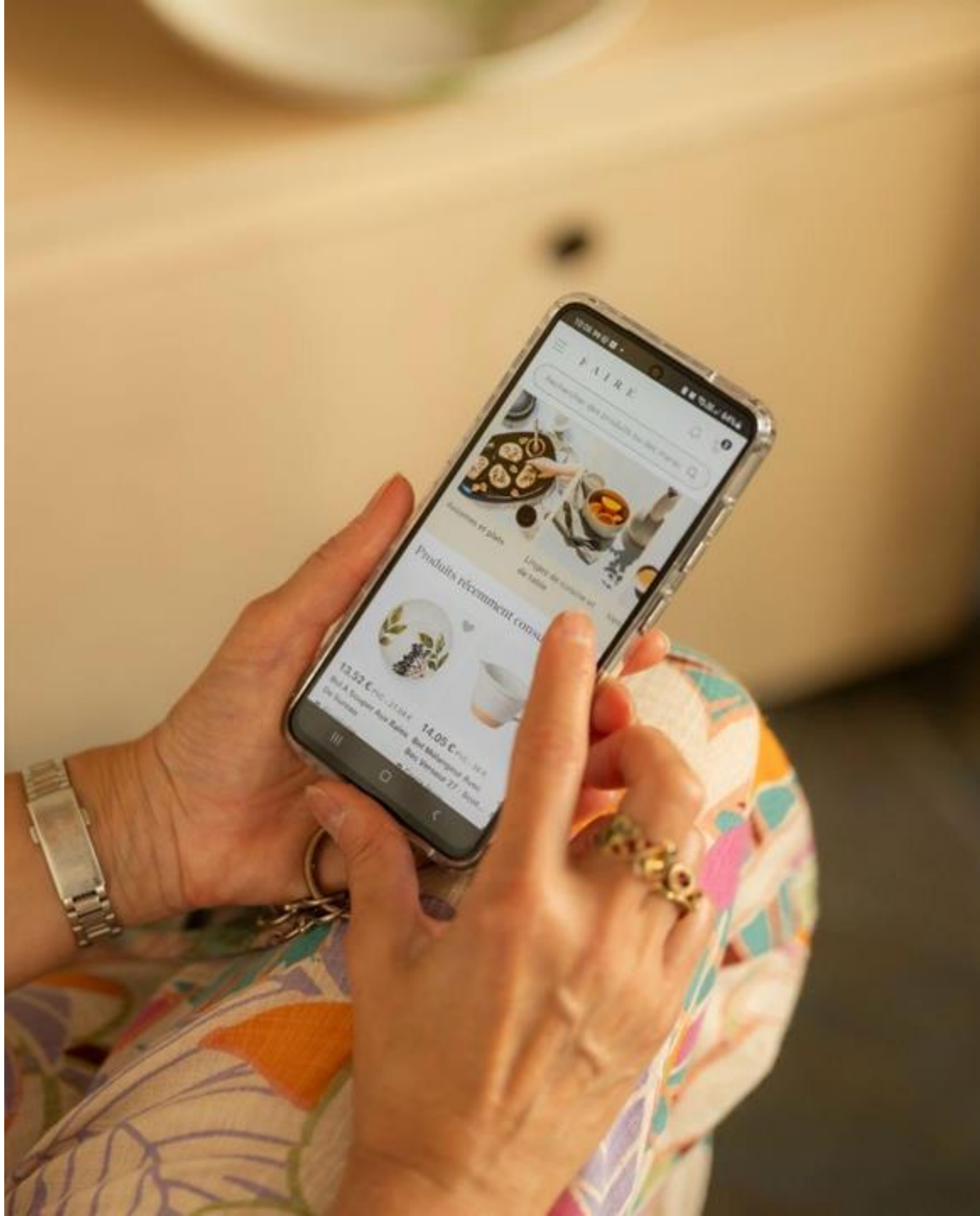
What's next?

Build workflows on top of OKE

- Batch inference infra for GPU bound workloads
- Pilot use case:
    - Backfill product catalog quality improvements on our entire catalog with a 10B param model using 8 A100s
    - Baseline: on-demand GPUs on another vendor
    - 86% runtime improvement with OKE + run:AI + GPU batch inference layer
        - 50 GPU days -> 7 GPU days
    - Save on incremental experiment costs with reserved cluster on OCI

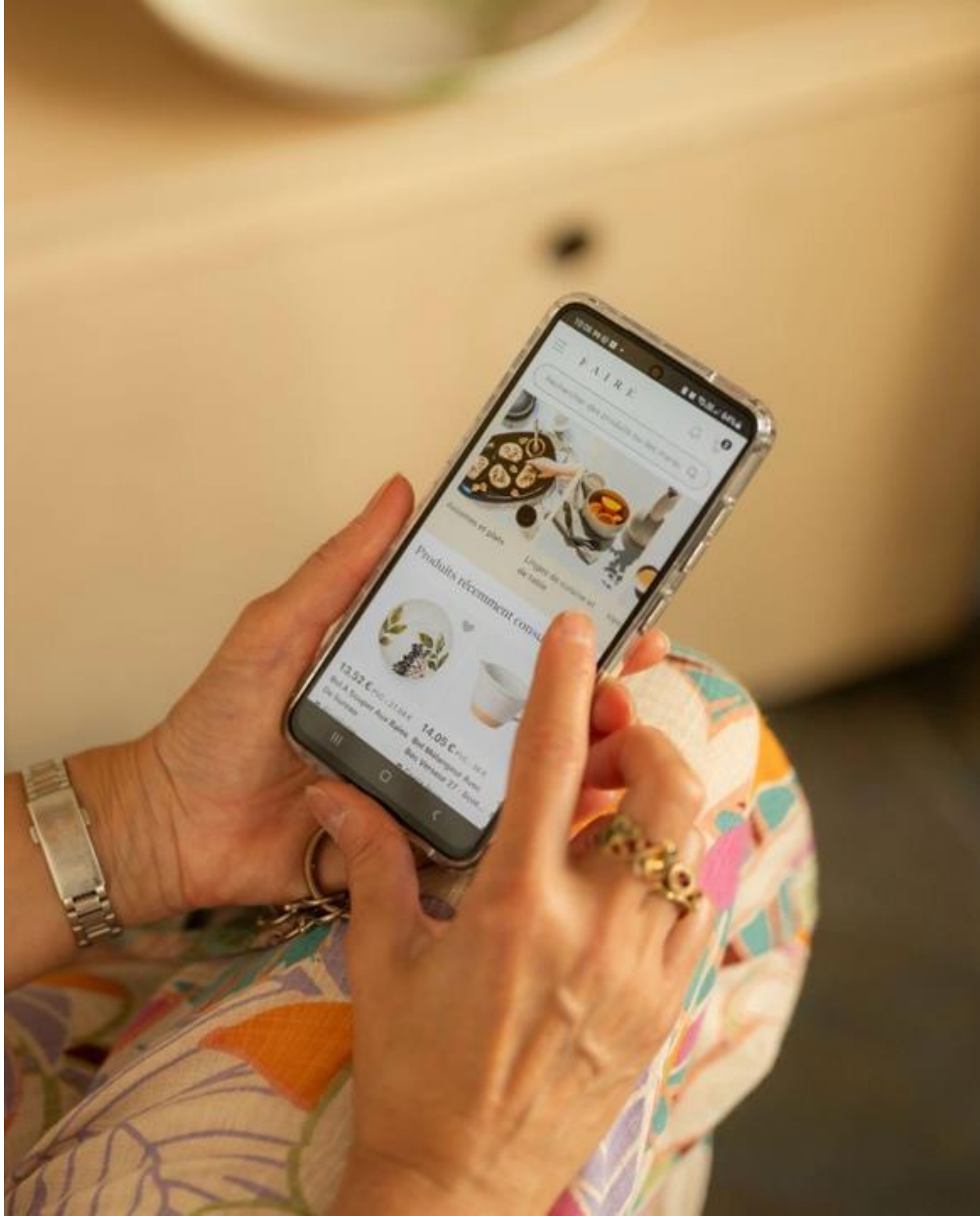
Relevance： 0 – 〉 1

- Search Relevance
- Improve relevance of search algorithms using ESCI framework
- First pass: Human annotators
  - One month delay to understand performance
- Enter GPT
  - Fine-tuned public LLM with text completion to assign ESCI rating to query -> product pairs
  - Reduced delay from one month to one hour



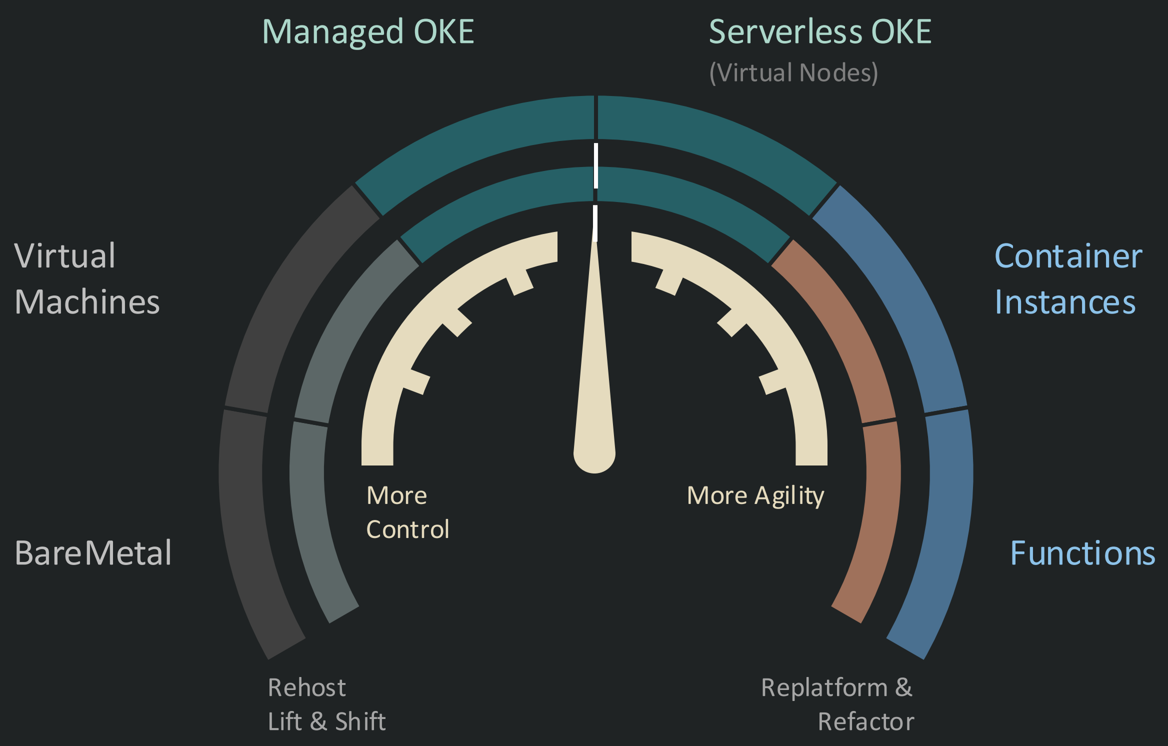
Relevance：　1　－＞　2

- Larger model is not always better
- With in-house fine-tuning, we get fine-grained control
- Cost and performance benefits from moving from LLM provider -> fine-tuned LLM
    - 28% jump in search relevance prediction accuracy
- Llama2 fine-tuned on our OKE GPU cluster (8 X A100s)



Serverless workloads

Control vs Agility, Pick What You Need



Flexible choice of hosting options

- Self Managed
- Managed OKE & Serverless OKE
- Containers as a platform

Simply K8s operations at scale

- Offload Kubernetes infrastructure management
- Automated management of common operational tasks such as upgrades
- Built-in security and governance controls

OCI Functions

Simple, Secure, OCI-Native

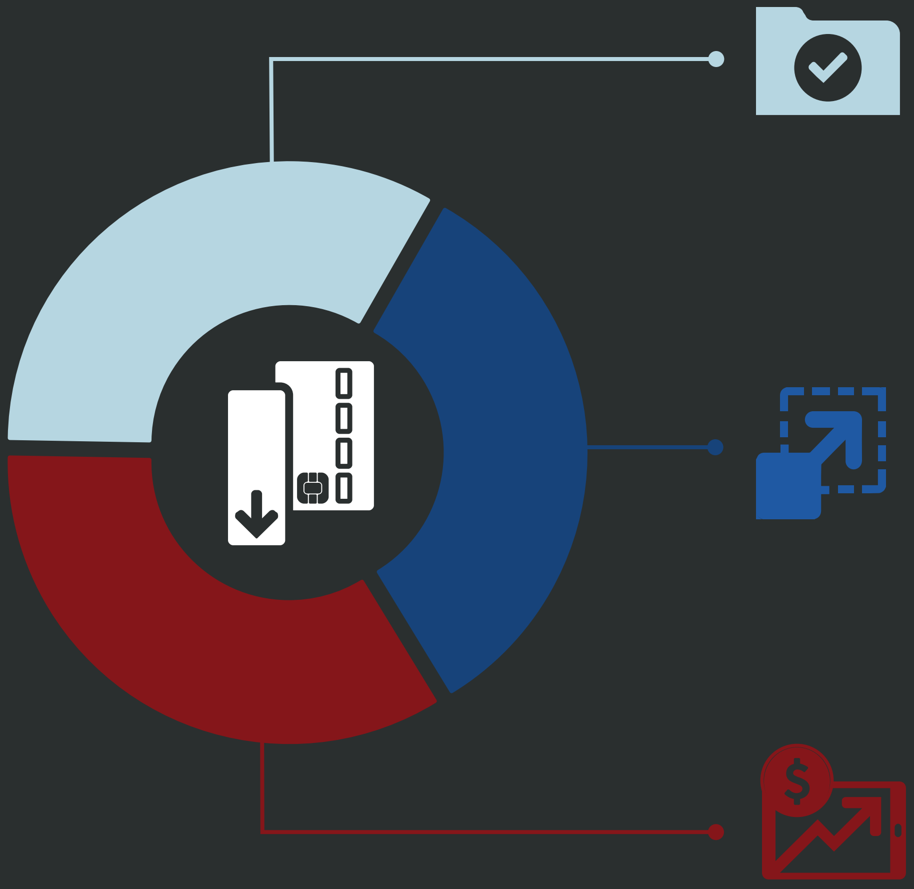
Functions-as-a-Service

Oracle Cloud Integrated

Container Native

Open Source

Secure



Pay per use

Pay for execution, not for idle time

Autonomous

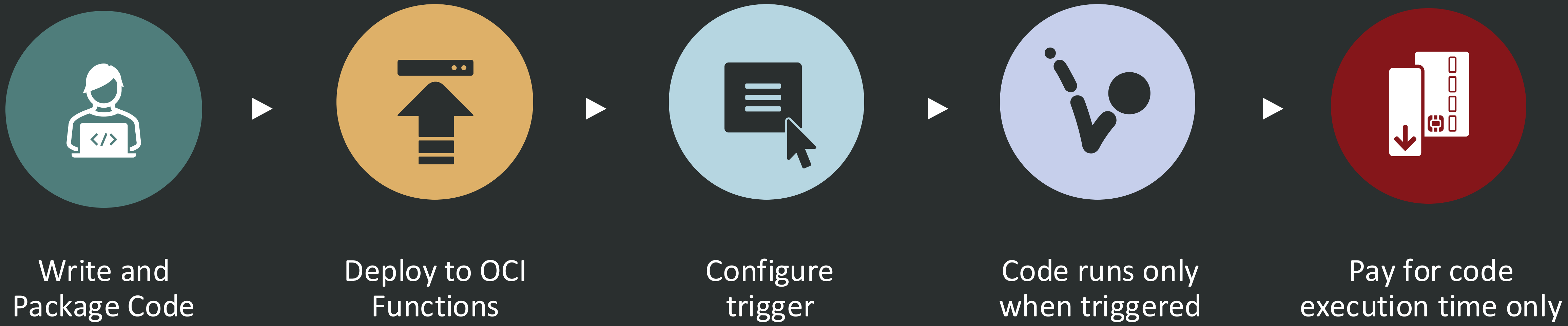
Platform auto-scales functions

No servers to provision, manage

Event-driven

Oracle Cloud Infrastructure triggers to run your code

How Does it Work



Write and Package Code

Deploy to OCI Functions

Configure trigger

Code runs only when triggered

Pay for code execution time only

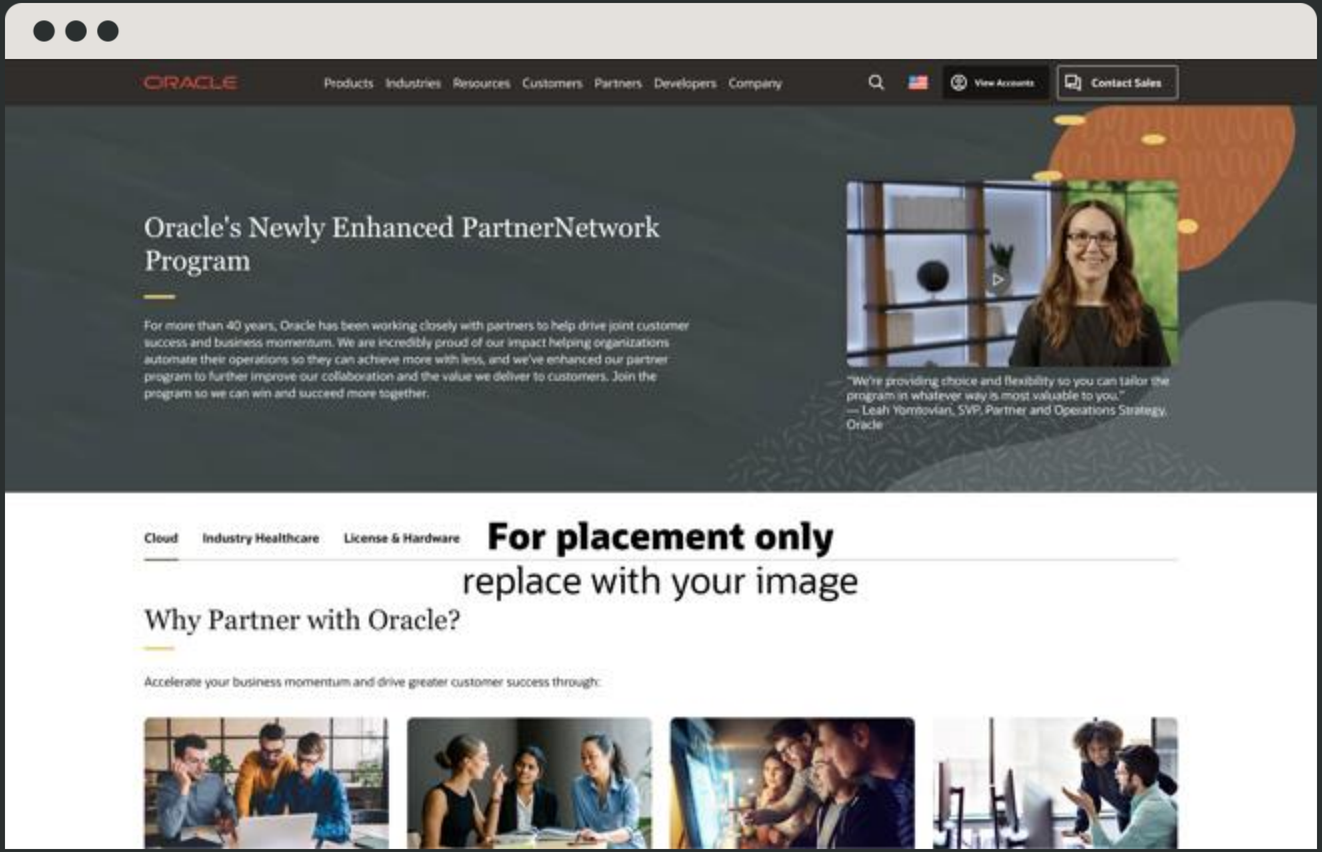
What’s New with OCI Functions?

Major Feature Releases Since Cloud World 2024

- 3GB Memory Functions
- Scheduled Functions
- Longer Running Functions
- Response Destinations

Scheduled Functions

- Run Functions on a defined schedule — hourly, daily, weekly, without external triggers.
- No cron servers, no scripts: native scheduling built into OCI Functions. Ideal for automated jobs like cleanup, report generation, and ETL pipelines
- Configurable directly in the Console or OCI CLI, with integrated logs and monitoring.



Oracle's Newly Enhanced PartnerNetwork Program

For more than 40 years, Oracle has been working closely with partners to help drive joint customer success and business momentum. We are incredibly proud of our impact helping organizations automate their operations so they can achieve more with less, and we've enhanced our partner program to further improve our collaboration and the value we deliver to customers. Join the program so we can win and succeed more together.

"We're providing choice and flexibility so you can tailor the program in whatever way is most valuable to you."
— Leah Yomtovian, SVP, Partner and Operations Strategy, Oracle

For placement only

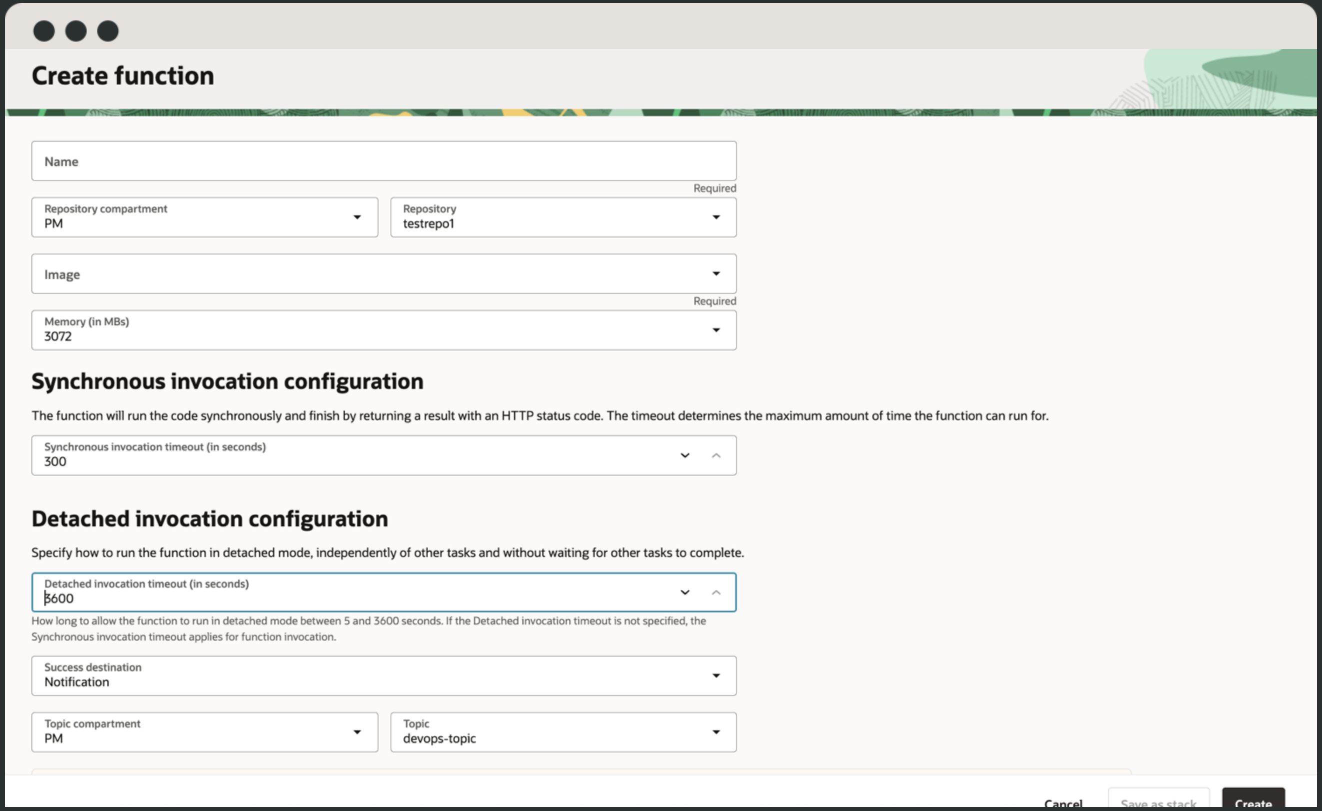
replace with your image

Why Partner with Oracle?

Accelerate your business momentum and drive greater customer success through:

Long-Runing Functions

- Execute for up to 60 minutes in detached mode (vs 5 min before)
- Ideal for AI/ML jobs, ETL pipelines, batch processing, and API integrations
- Extend runtime simply; no re-architecture
- All with serverless benefits: auto-scaling + pay-per-use



Create function

Name

Required

Repository compartment: PM

Repository: testrepo1

Image

Required

Memory (in MBs): 3072

Synchronous invocation configuration

The function will run the code synchronously and finish by returning a result with an HTTP status code. The timeout determines the maximum amount of time the function can run for.

Synchronous invocation timeout (in seconds): 300

Detached invocation configuration

Specify how to run the function in detached mode, independently of other tasks and without waiting for other tasks to complete.

Detached invocation timeout (in seconds): 3600

How long to allow the function to run in detached mode between 5 and 3600 seconds. If the Detached invocation timeout is not specified, the Synchronous invocation timeout applies for function invocation.

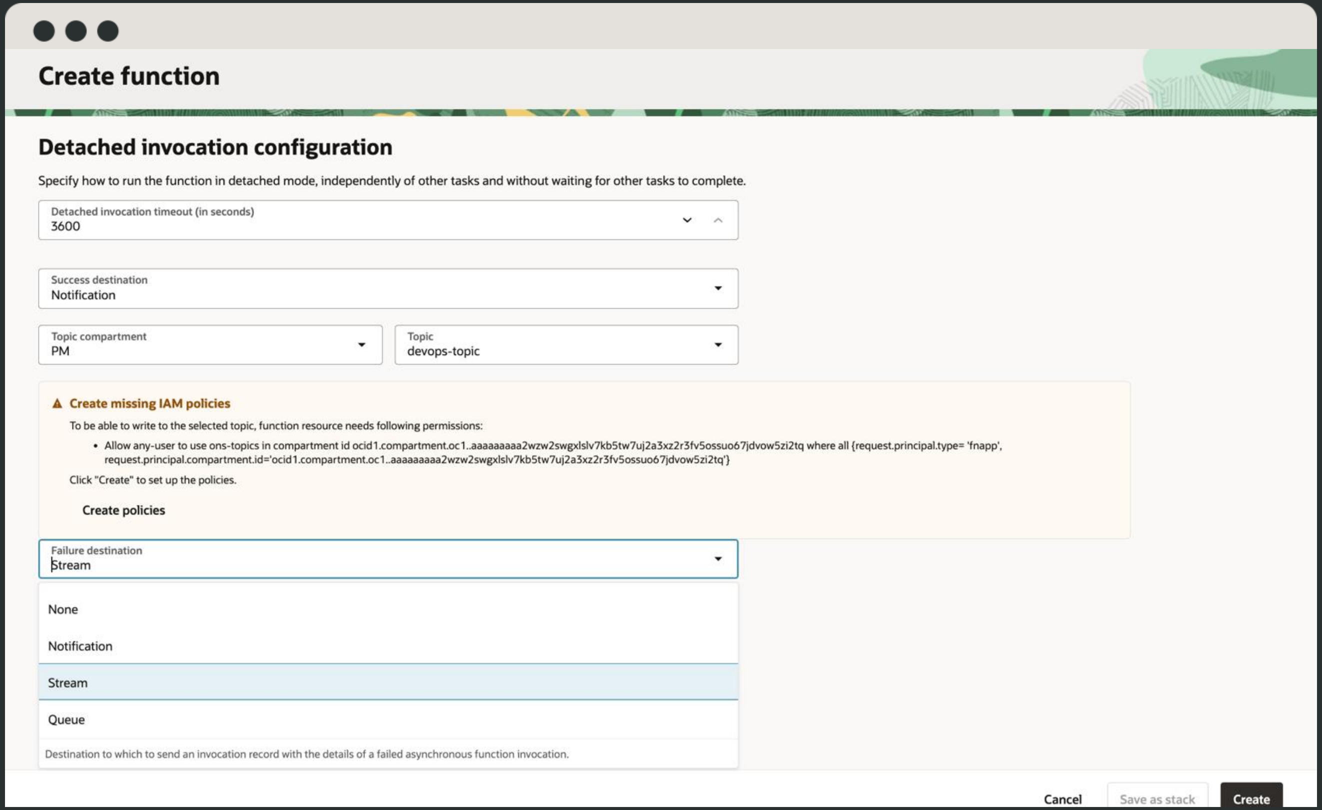
Success destination: Notification

Topic compartment: PM

Topic: devops-topic

Response Destinations

- Automatically route success or failure outcomes to Streaming, Queue, or Notifications
- Build clean, event-driven workflows without polling or custom code
- Simplifies error handling and visibility for asynchronous executions
- Perfect for alerting, pipelines, and chained automation



Create function

Detached invocation configuration

Specify how to run the function in detached mode, independently of other tasks and without waiting for other tasks to complete.

Detached invocation timeout (in seconds): 3600

Success destination: Notification

Topic compartment: PM

Topic: devops-topic

Create missing IAM policies

To be able to write to the selected topic, function resource needs following permissions:

- Allow any-user to use ons-topics in compartment id ocid1.compartment.oc1..aaaaaaaaa2wzw2swgxlslv7kb5tw7uj2a3xz2r3fv5ossuo67jdvow5zi2tq where all {request.principal.type= 'fnapp', request.principal.compartment.id='ocid1.compartment.oc1..aaaaaaaaa2wzw2swgxlslv7kb5tw7uj2a3xz2r3fv5ossuo67jdvow5zi2tq'}

Click "Create" to set up the policies.

Failure destination: Stream

Destination to which to send an invocation record with the details of a failed asynchronous function invocation.

Your feedback is important.

Scan this QR Code or use the Mobile App to share your thoughts on this session.