AI & HPC Infrastructure on Oracle Cloud Infrastructure

Marcin Zablocki
Solutions Architect

https://www.linkedin.com/in/marcinzablocki/
@mz_oracle

Rajan Panchapakesan
PM, HPC and Big Compute
Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle’s products remains at the sole discretion of Oracle.
Our Journey

BARE METAL CLOUD
Oracle Open World 2016
Our Journey

ORACLE CLOUD INFRASTRUCTURE
Rebranding & Launch October 2017
Our Journey

NVIDIA TESLA
P100
INSTANCES

GENERAL AVAILABILITY
Oracle Open World 2017
Our Journey

Bare Metal
V100 Volta
instances

GENERAL AVAILABILITY
March 2018
Our Journey

P100 and V100 GPU Virtual Machines.

TOKYO
September 2018
Our Journey

FIRST CLOUD PROVIDER
WITH HGX-2 ARCHITECTURE

#GTC18 MUNICH
October 2018
THE METAL IN BARE METAL
Oracle Cloud Infrastructure BM and VM

Bare Metal Standard
52 Cores, 768 GB RAM, up to 1PB Block Storage
2x 25Gbe Network Interfaces

Bare Metal DenseIO
51.2 TB of local NVMe SSD
up to 1PB Block Storage
2x 25Gbe Network Interfaces

Virtual Machine Standard
1-16 and 1-24 Cores
7 – 320 GB RAM
up to 1PB Block Storage
Up to 25Gb/s Ethernet

Virtual Machine Dense
4 – 24 Cores
60 – 320 GB RAM
3.2 TB – 25.6 TB local NVMe SSD
up to 1PB Block Storage

$ 0.0638 / core/h
$ 0.1275 / core/h
$0.0638 / core/h
$0.1275 / core/h
Virtual Machine GPU
1 P100 GPU
12 OCPU | 104 GB RAM
up to 1PB Block Storage
$1.275 GPU/h

Virtual Machine GPU v2
1 – 4 V100 GPU
6 – 24 OCPU
104 to 360 GB RAM
up to 1PB Block Storage
NVLINK
$2.25 GPU/h

Bare Metal GPU
28 Cores, 192 GB RAM,
2x Tesla P100 GPUs
$1.275 GPU/h

Bare Metal GPU V2
52 Cores, 768 GB RAM,
8x Tesla V100 GPUs
NVLINK
$2.25 GPU/h

Oracle Cloud Infrastructure GPU
COMING SOON

Bare Metal GPU v3
48 OCPU – 3.5GHZ all core turbo
768 GB RAM
HGX-2
8 * V100 32 GB SXM3 GPU
NVLINK
Up to 1PB Block Storage

Virtual Machine GPU v3
5 - 22 OCPU – 3.5GHZ all core turbo
90 - 360 GB RAM
HGX-2
1 - 4 * V100 32 GB SXM3 GPU
NVLINK
Up to 1PB Block Storage
Oracle Cloud Infrastructure Storage

File Storage Service
- Up to 8 Exabytes of storage
- Managed distributed file service
- NFSv3 mount point
- Pay for what you use

Block Storage
- 50 GB-16 TB volumes
- Up to 25K IOPS per volume
- 400K IOPS per host
- 320MB/S per volume
- Pay for Storage not IOPS

Local NVMe
- 51.2 TB of local NVMe SSD
- 6.4-25.6 TB for VM
- $0.0425 GB/month

Object Storage
- Up to 10TB per object
- Highly durable
- $0.0255 GB/month
- $0.0034 10,000 req

Pay for Storage not IOPS
Oracle Cloud Infrastructure Services

PaaS Services

- Autonomous Analytics Cloud
- Autonomous API Platform Cloud Service
- Autonomous Data Warehouse Cloud
- Autonomous Integration Cloud
- Autonomous Mobile Cloud Enterprise
- Autonomous Visual Builder Cloud Service
- Big Data Cloud
- Content and Experience Cloud
- Data Hub Cloud Service
- Data Integration Platform Cloud
- Database Cloud Service
- Developer Cloud Service
- Event Hub Cloud Service
- Java Cloud Service
- MySQL Cloud Service
- Oracle SOA Cloud Service
Our Architecture

- Multiple fault-domains per availability domain
- Multiple availability domains per region
- Global Hyper-Scale Regions with Backbone Network
- Encrypted interconnect between Availability Domains
- Encrypted backbone between regions
HPC is at our Core
Datacenters & Networking

- Global Hyper-Scale Regions with Backbone Network
- Truly Non-Over Subscribed Flat Networking meaning Flat, Fast & Predictable
- Predictable and low (10s of micro-second) latency within Datacenter
- 100s of micro-seconds across Datacenters within Regions
- Millisecond latency across Regions
HPC is at our Core
Compute & Storage

- Multi-Tenant Bare-Metal Cloud Instances
- Highly configurable private overlay networks – moves management and IO out of the hypervisor and enables lower overhead and bare metal instances
- Each host providing a total of 50Gbps of Bandwidth & low latency within a Datacenter (2x25Gb/s)
- Best-in-Class Storage Performance with up to 512TB of NVMe Block Storage per Host
- Managed Distributed File System Service
Myth: Cloud means I **HAVE** to run virtualized

- **Truth** OCI Bare Metal runs HPC workloads up to 18% faster than a VM

- Faster turn-around time for a simulation model leads to increased license ROI

- 15% can mean hours and days when running large simulations or AI workloads
Myth: HPC in the cloud is slower than on-premises

• **Truth**: OCI Bare Metal instances perform inline with on-premises HPC clusters

• Bare Metal instances scale workload performance linearly, similar to on-premises

• Oracle helps maximize utilization of ISV licenses while providing linear scaleup on performance – “Best of both worlds”
Myth: GPUs are too expensive on the Cloud

- **Truth**: Cost of an OCI GPU instance is cheaper on a per job basis

- Hardware refresh is faster and consistent on OCI. Jobs complete faster with each generation and the price performance increases

- Cost of full GPU server on-premises is over $150,000 plus power, space, maintenance and support

---

![Billion Word Language Model Benchmark](chart)

- White Box
- Oracle OCI
- DGX1
Myth: Managing storage on the cloud is difficult & expensive

- **Truth**: FSS on removes the burden for managing file servers while costing same as block storage ($0.04)

- ANSYS Fluent runs some models faster on OCI with the model stored in FSS, even though the R/W takes slightly longer

- Read/Write happens 1 time, solve happens 5000 times
Up to **52%** less for virtual machine instances

Up to **79x** less for high performance block storage

Up to **13x** less for internet data egress
Openness in Oracle

- Chef
- Ansible
- Fn Project
- GraphPipe
- Kubernetes
- Docker
- Terraform
- GraalVM
- Oracle Linux
An enterprise data science platform that brings everything together.
What is it?

**GraphPipe** is a protocol and collection of software designed to simplify machine learning model deployment and decouple it from framework-specific model implementations.

Why did we make it?

We found existing solutions for model serving to be inconsistent and/or inefficient. Without a consistent protocol for communicating with different model servers, it is often necessary to build custom clients for each workload.

GraphPipe solves these problems by standardizing on an efficient communication protocol and providing simple model servers for the major ML frameworks.

We hope that open sourcing GraphPipe makes the model serving landscape a friendlier place.

Features

- A minimalist machine learning transport specification based on flatbuffers
- Simple, efficient reference model servers for Tensorflow, Caffe2, and ONNX.
- Efficient client implementations in Go, Python, and Java.
Serving Before GraphPipe

- autogenerated client
  - protocol buffers
    - tensorflow serving
  - custom protocol
    - custom server
  - standard json
    - mxnet server

Frameworks:
- TensorFlow
- PyTorch
- Caffe2
- ONNX
- mxnet
Serving After GraphPipe

GraphPipe

graphpipe-tf

graphpipe-onnx

Caffe2

ONNX

PyTorch

mxnet

TensorFlow
NGC Deployment

• General Availability
• Pascal & Volta Bare Metal and Virtual Machine Instance Shapes supported
• Deploy Deep Learning Frameworks, HPC Applications and Visualization Applications seamlessly
• Pre-Configured NGC Image available to deploy on OCI
• Flexibility to run dev/test on Virtual Machine GPU Instances and run production workloads on Bare-Metal Instances

https://cloud.oracle.com/iaas/gpu
How can I get started with Oracle Cloud?

- Get started with up to 235 GPU hours or 4700 OCPU hours.
- Start with storage, dev / test, or analytics
- More than 30 services available via trial

Test Drive Oracle Cloud: Cloud.Oracle.com/TryIt