ORACLE

ORACLE

HPC on OCI

Bare Metal Computing in a Cloud Environment

March 1st, 2020

Safe harbor statement

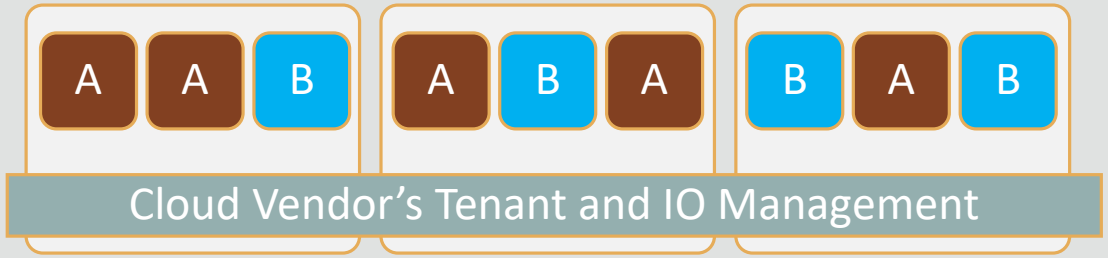
The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions.

The development, release, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.

Why is OCI different

Traditional Gen1 Cloud Infrastructure

Customer App & Data **share the same server** with the cloud vendor's multi tenant management

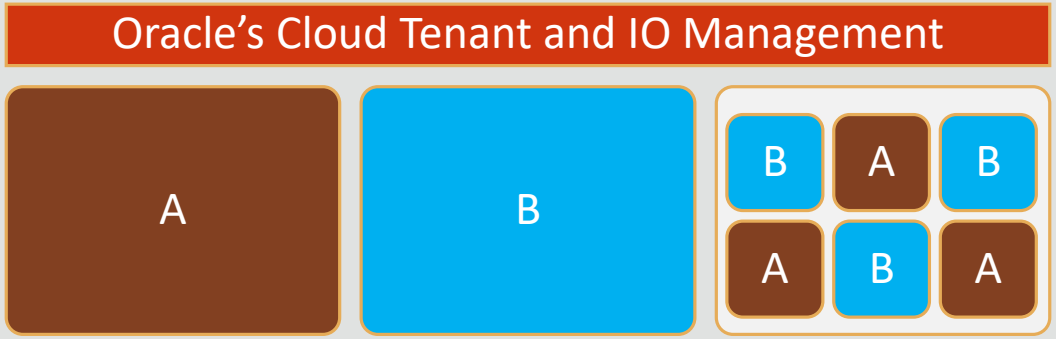


VMs

ORACLE Gen2 Cloud Infrastructure

Customer App & Data **are Fully Isolated** from Oracle's multi tenant management

No Oracle Cloud Software has to be present on the Customer's server!

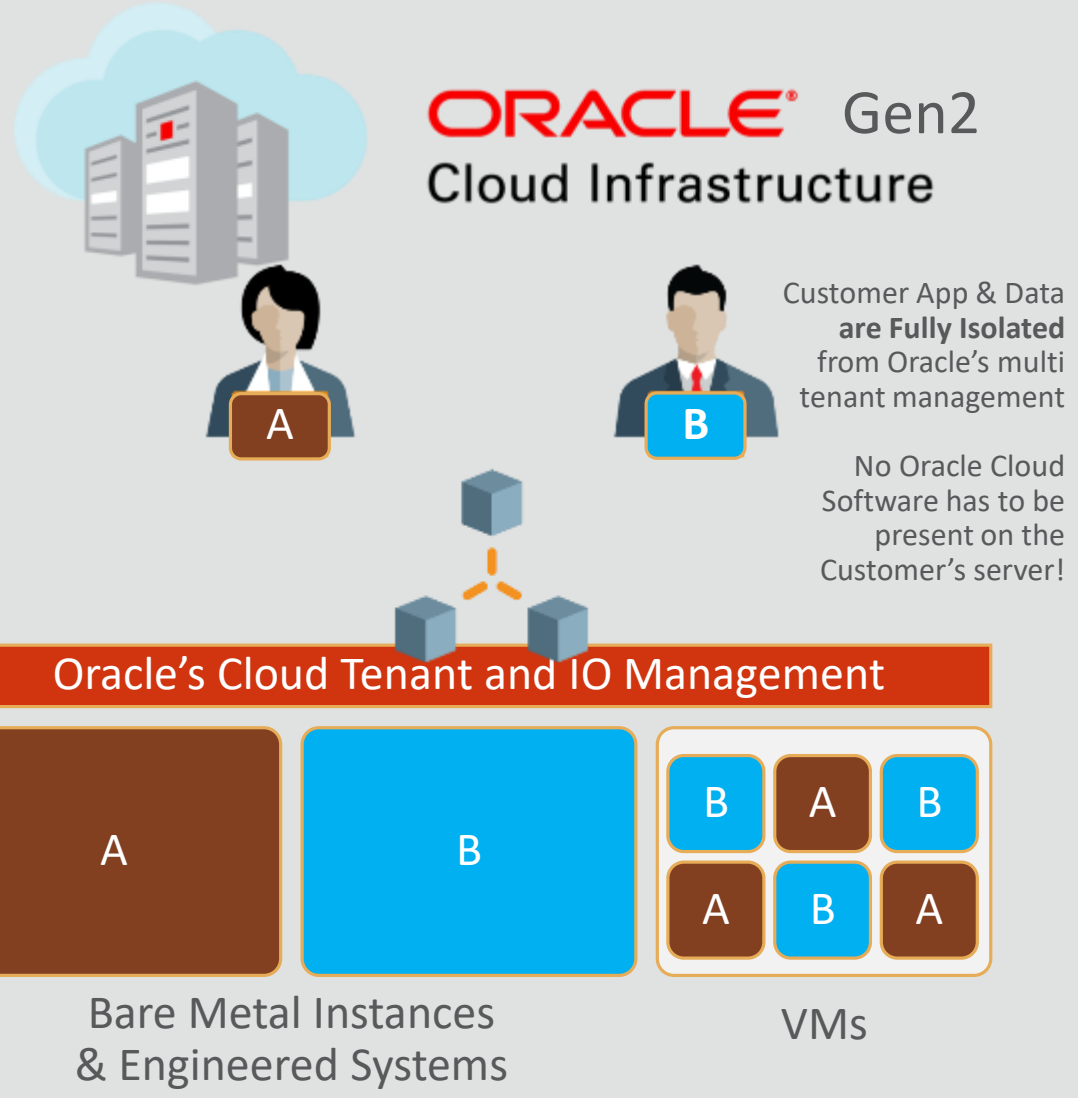


Bare Metal Instances & Engineered Systems

VMs

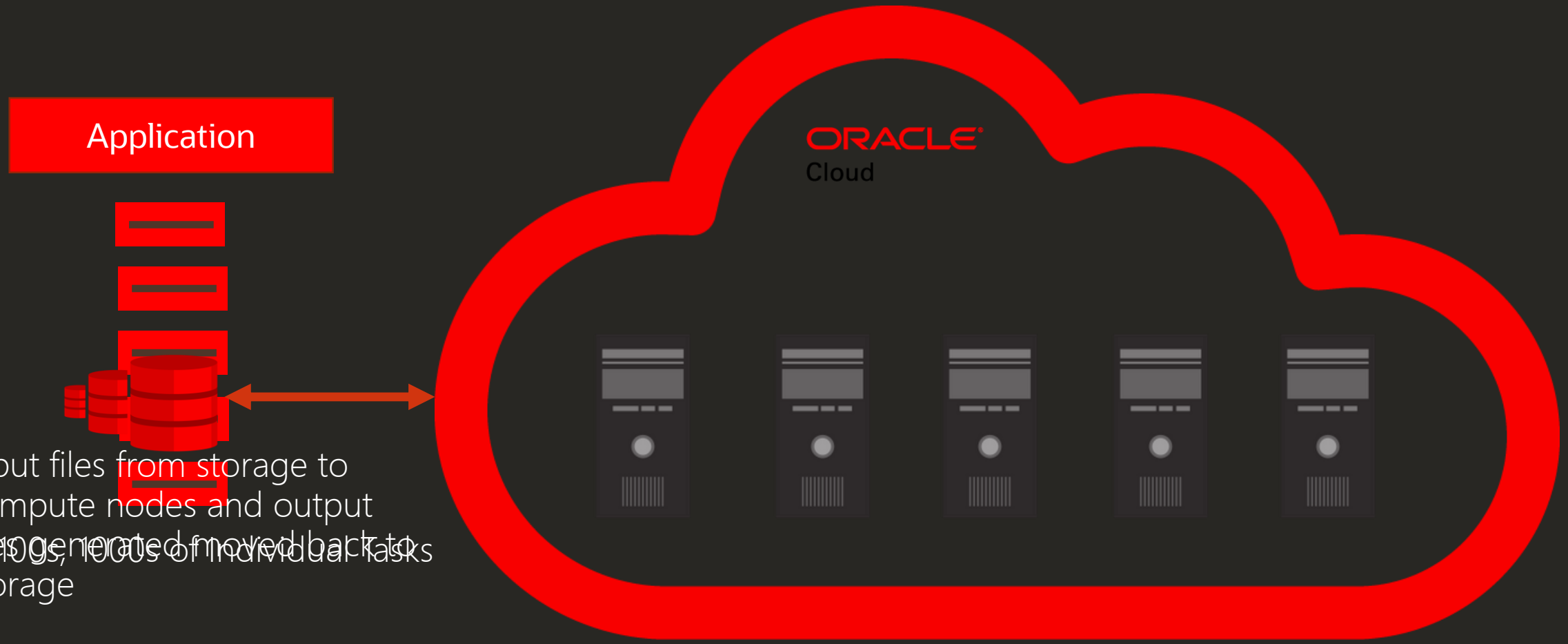
Why is OCI different

- ✓ Build for **Enterprise** with **Unique Multi-Tenant Isolation**
- ✓ Scalable **Ultra High Performance** at Predictable **Lower Cost**
- ✓ Deployment choices; VMs, BMs, Containers -> **Any Shape, Any Workload**
- ✓ Enterprise Edge Services; Expose any service **Secure and Scalable Globally**

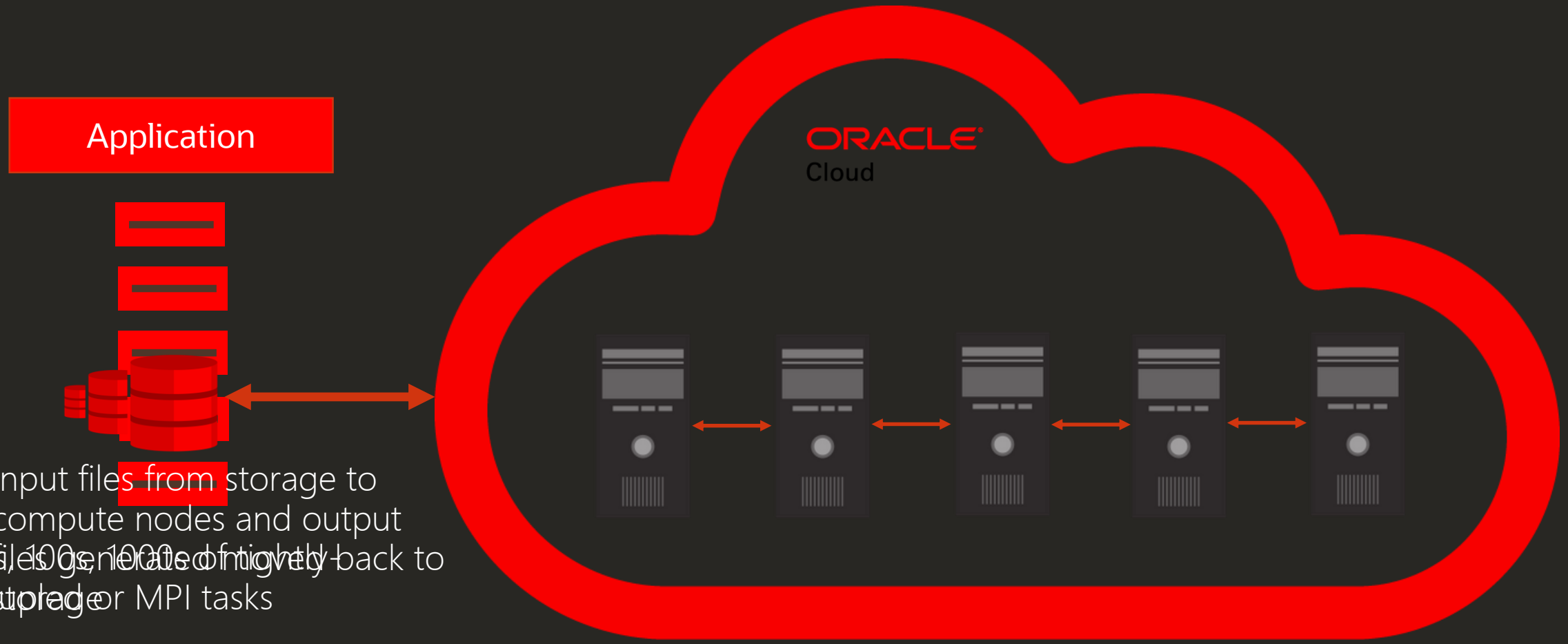


HPC Workloads

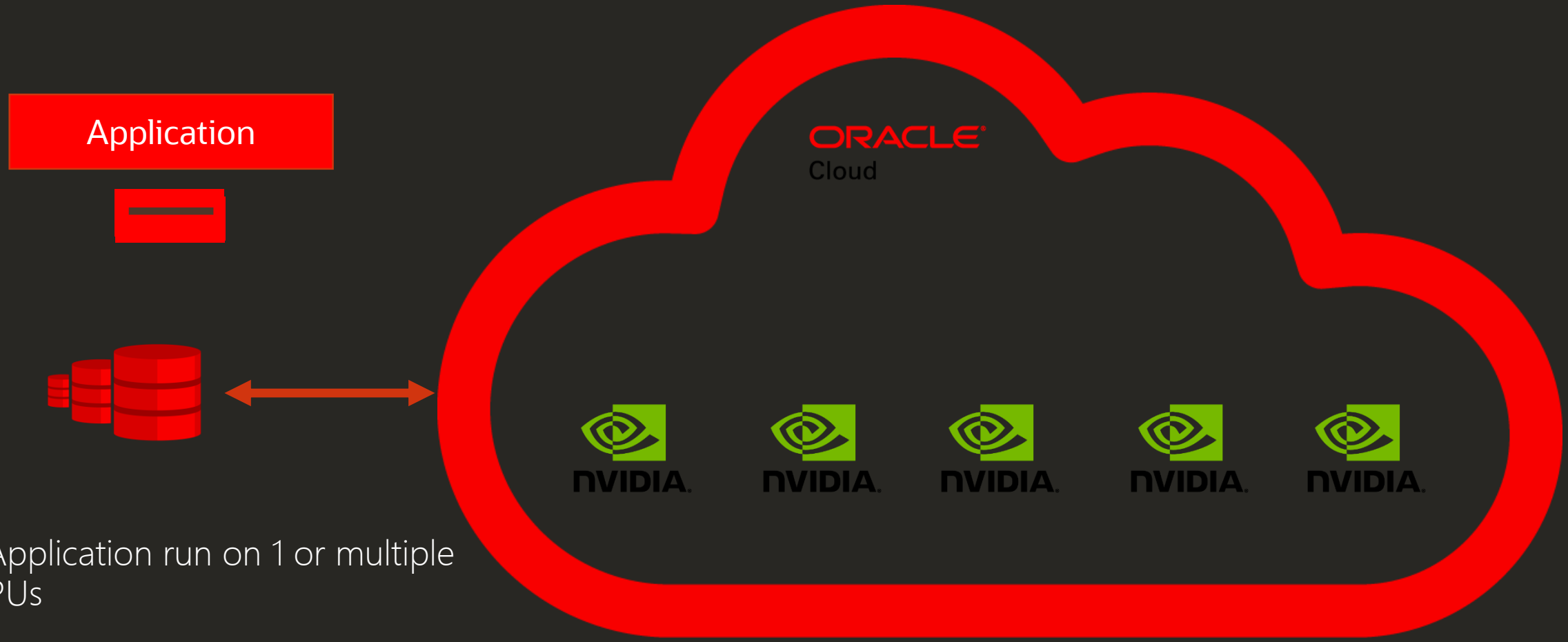
Embarrassingly Parallel



Tightly-Coupled (*or MPI)

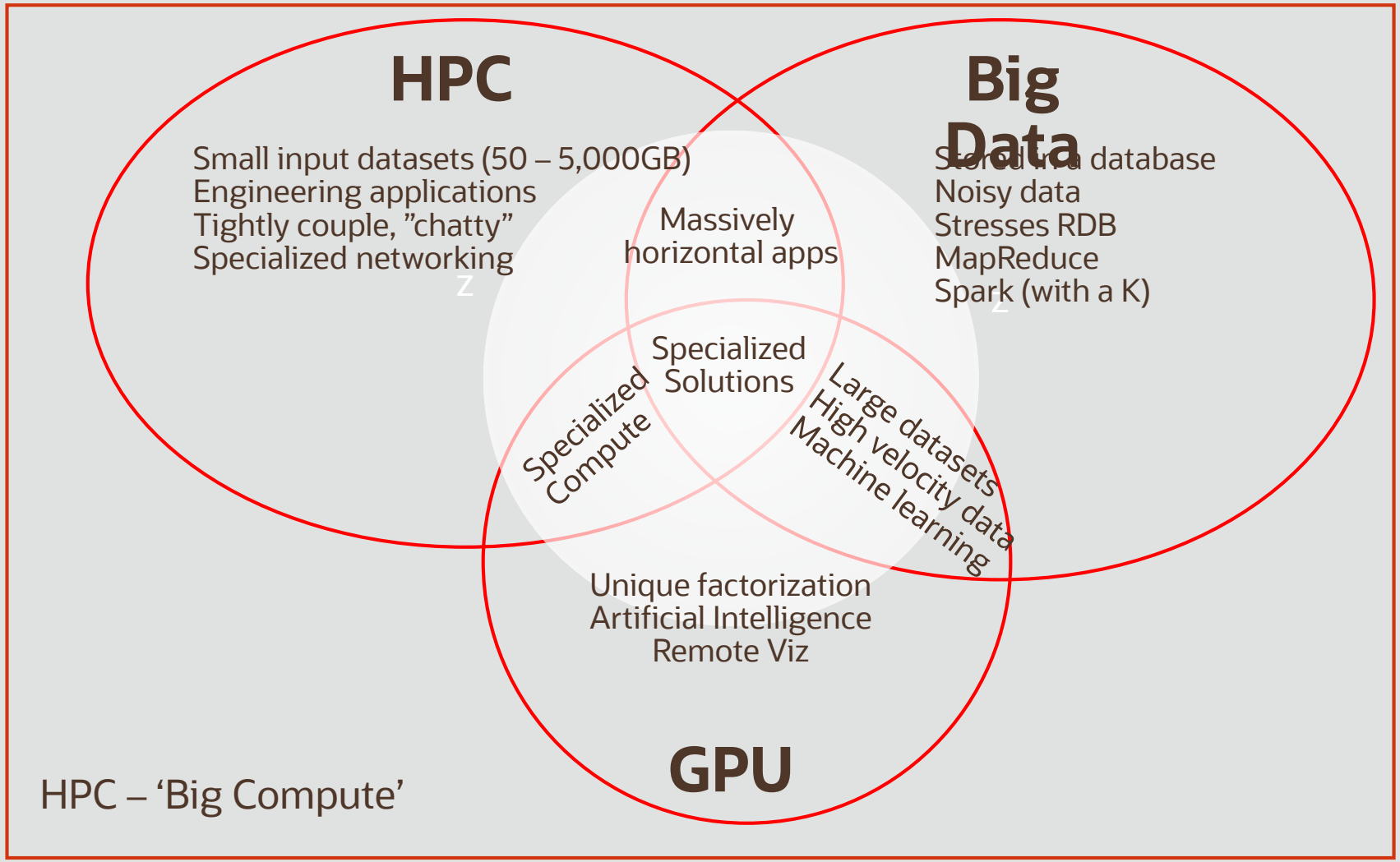


AI/ML (GPU)

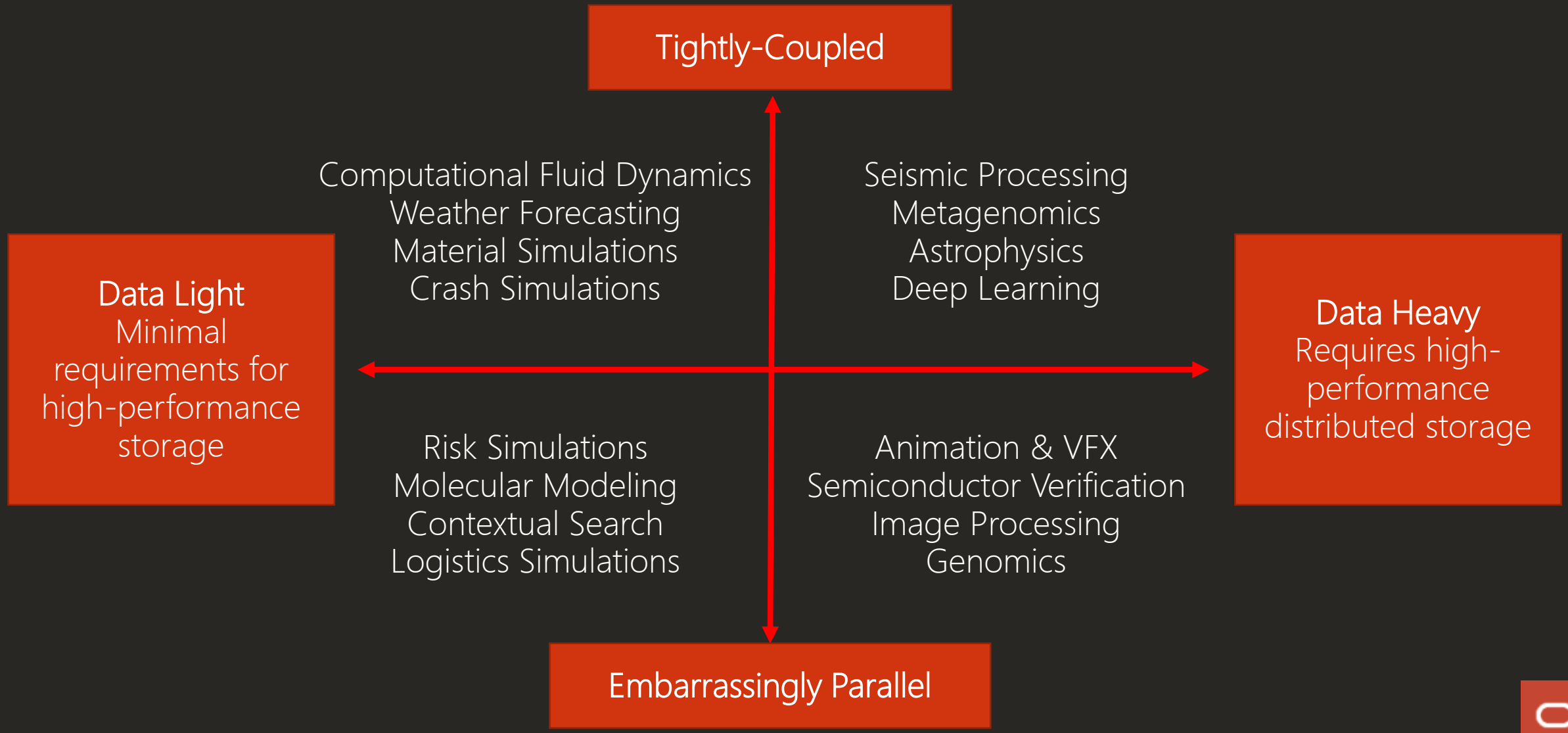


1 Application run on 1 or multiple GPUs

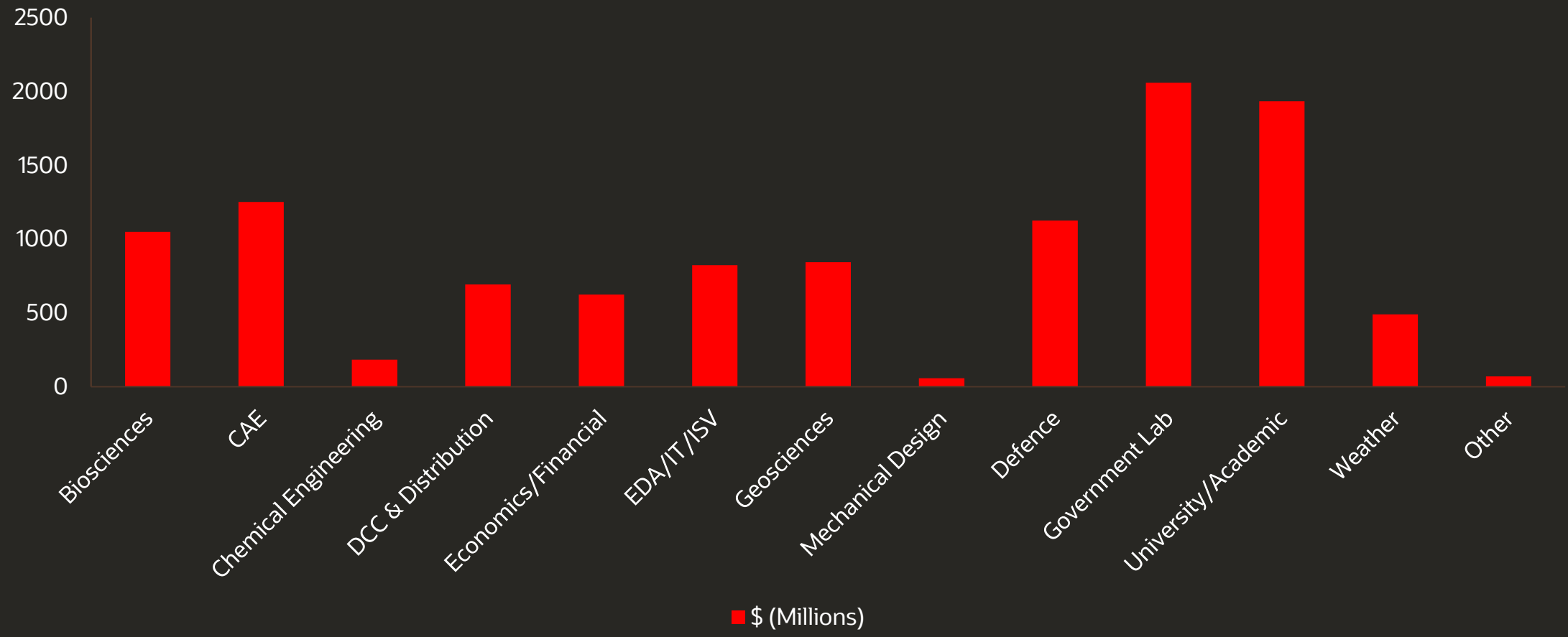
What do we define as HPC?



Mapping HPC across use-cases



Who is investing in HPC today?



Reference -* IDC 2016

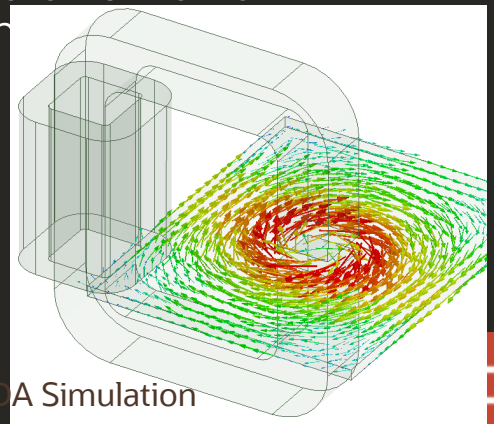
HPC in Manufacturing, Automotive & Aerospace

- Tightly Coupled
- 1,000 – 4,000 cores
- HPC, CPU, Storage

CAE (Computer Aided Engineering) workloads to develop commercial products including virtual testing for strength, reliability and operational capacity.

Environment may include finite element analysis (FEA), large-scale deformation, electric design automation (EDA), computational fluid dynamics (CFD), dynamic simulation and design optimization. Examples include Crash Analysis for automotive custom testing of different car materials.

These workloads are both parallel and tightly coupled.



EDA Simulation

HPC in Media & Entertainment

- Embarrassingly Parallel
- 1,000 – 2,000 cores
- GPU, CPU

Major workloads include Transcoding, 3D Rendering and Streaming

Typical GPUs today in OCI can support up to 30 concurrent HD 1080P streams for the Netflix like or gaming services.

Major compute utilization comes from Visual Effects Rendering. Workloads can range from 1000s of cores to upwards of 50k cores for single job. Each job can take over 12+ hours to complete with multiple passes of jobs needed. Think 60 minutes of visual effects = 3600 seconds = 216,000 frames (60 frames/second). Each frame could take hours to days to render. Frames are generally rendered multiple times.

Mostly ISV software is used in conjunction with a open-source or third party scheduler or service.

HPC in Financial Services

- Embarassingly Parallel
- 1 core, 10's of thousands of jobs
- CPU, very bursty

When we talk about OCI in Financial Services for HPC – we mostly mean Insurance, Banking and High-Frequency Trading.

There are various workloads such as Stress Testing, Risk Simulation, Derivatives Pricing, Actuarial Modelling, Catastrophe Modelling, Annuity Hedging and others.

Most workloads go to the cloud due to requirement for burst capacity.

There are ISVs involved however most workloads run proprietary code built by quants.

There are large networking needs to distribute data across various environments, Hybrid makes moving large data challenging.

There are also compliance issues that plague this industry.

HPC in Life Sciences

- Embarrassingly Parallel
- 4,000 – 8,000 cores
- GPU, CPU

DNA sequencing in Genomics is one of the largest HPC workloads utilizing 1000s of cores.

Recently a team including Oak Ridge National Laboratory & Rice University along with another cloud provider ran over 5300 genome samples over a cluster with 4000 cores in 50 days with a total core usage of ~5.2 million core hours. About 6TB of data transferred across.

GPUs can help accelerate some of these workloads almost by 5 – 10x over a single node with CPUs.

Other workloads include protein analysis, simulations to find new drugs, and cancer research.

AI & Deep Learning workloads using GPUs are taking off with predictive health symptom analysis and drug improvements using patient health history.

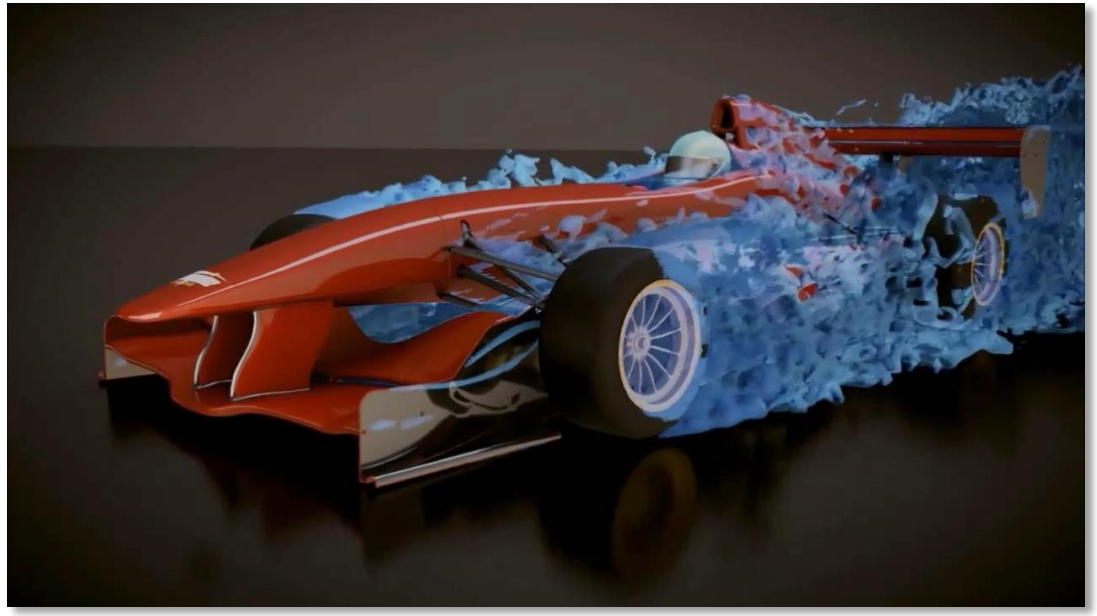
HPC Infrastructure: Compute

3.7 GHz, providing the highest performing processor in the cloud

36 Cores of Intel 6154 High Performance

384 GB of memory and 6.4 TB of local storage

20k over 20,000 cores in a single cluster



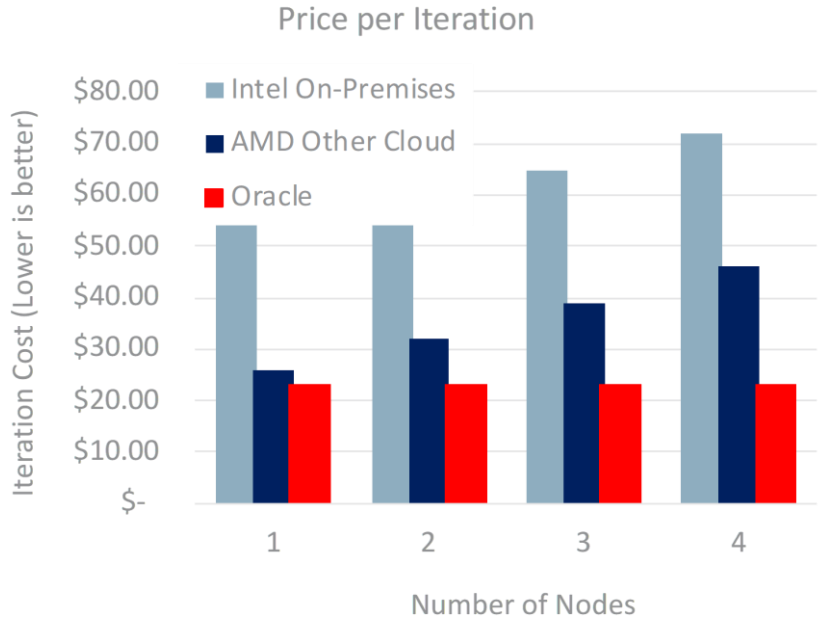
64 Physical threads on a single machine
128 total threads

3 ¢ per core per hour

most cost efficient physical processor

↑ Highest memory bandwidth for cloud

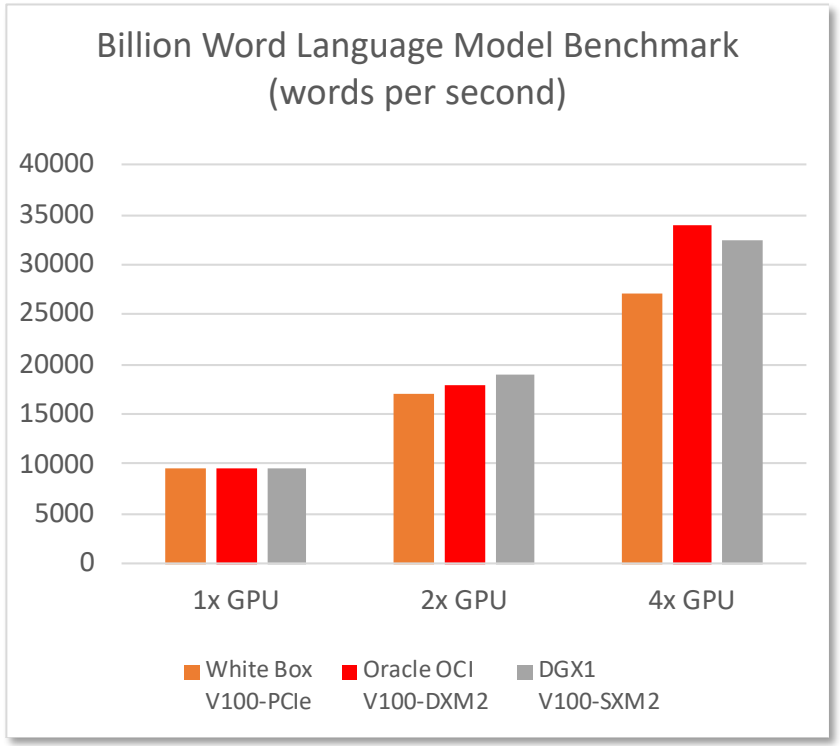
BM.Standard.E2.64 best price performance in the cloud



8 BM.GPU3.8 – 8 Way NVIDIA Volta SXM2, NVLINK, HGX-1 Open Compute 52 Skylake CPU cores, 50 Gbps

2 BM.GPU2.2 – 2 Way NVIDIA Pascal Perfect for Visualization or Inferencing

1 Attach up to 1 PB of NVME backed block storage to “Feed the Beast”



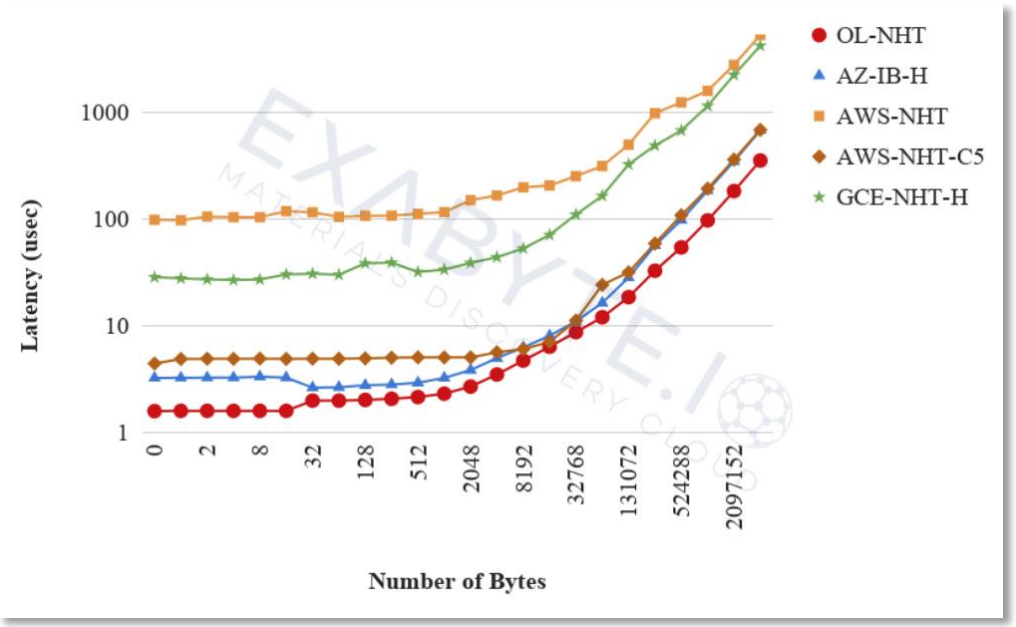
HPC Infrastructure: Network

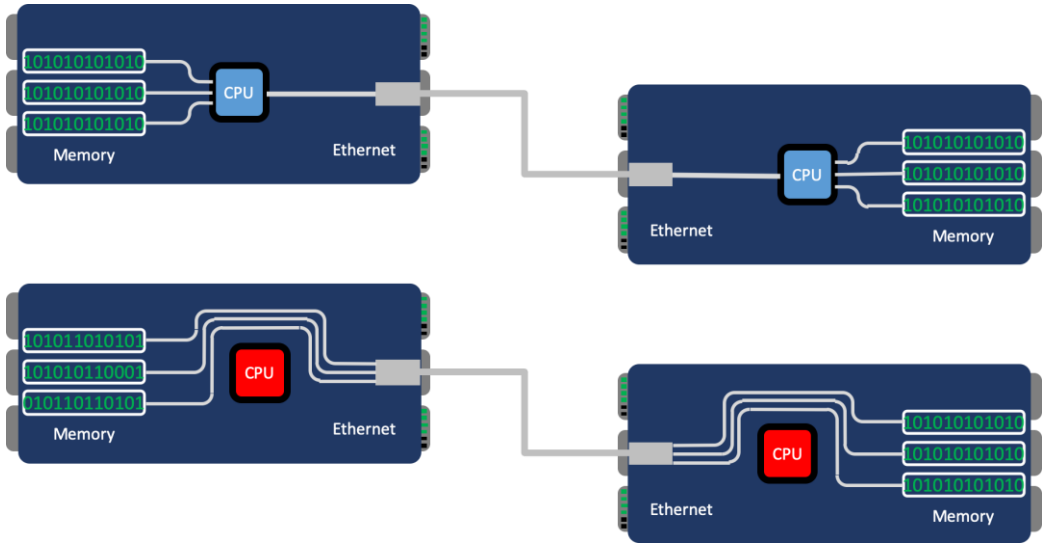
1.5 µs latency, over ROCEv2 - on-premises performance

12 GB/s transfer rate between nodes over the RDMA network, bare-metal application performance

125 Gbps full line rate connectivity

The only hardware point-to-point network connection in the cloud





- AMD V2
- GPU V3
- HPC V2
- ExaData is moving away from InfiniBand

Next ?

How are we different?

- Bare Metal HPC, on-premises like performance
- No 'jitter' in our HPC hosts or on our network
- Secure, isolated RDMA network with no oversubscription
- More than 20,000 cores in a single RDMA cluster
- Latency under 2μs vs. 15 μs on AWS

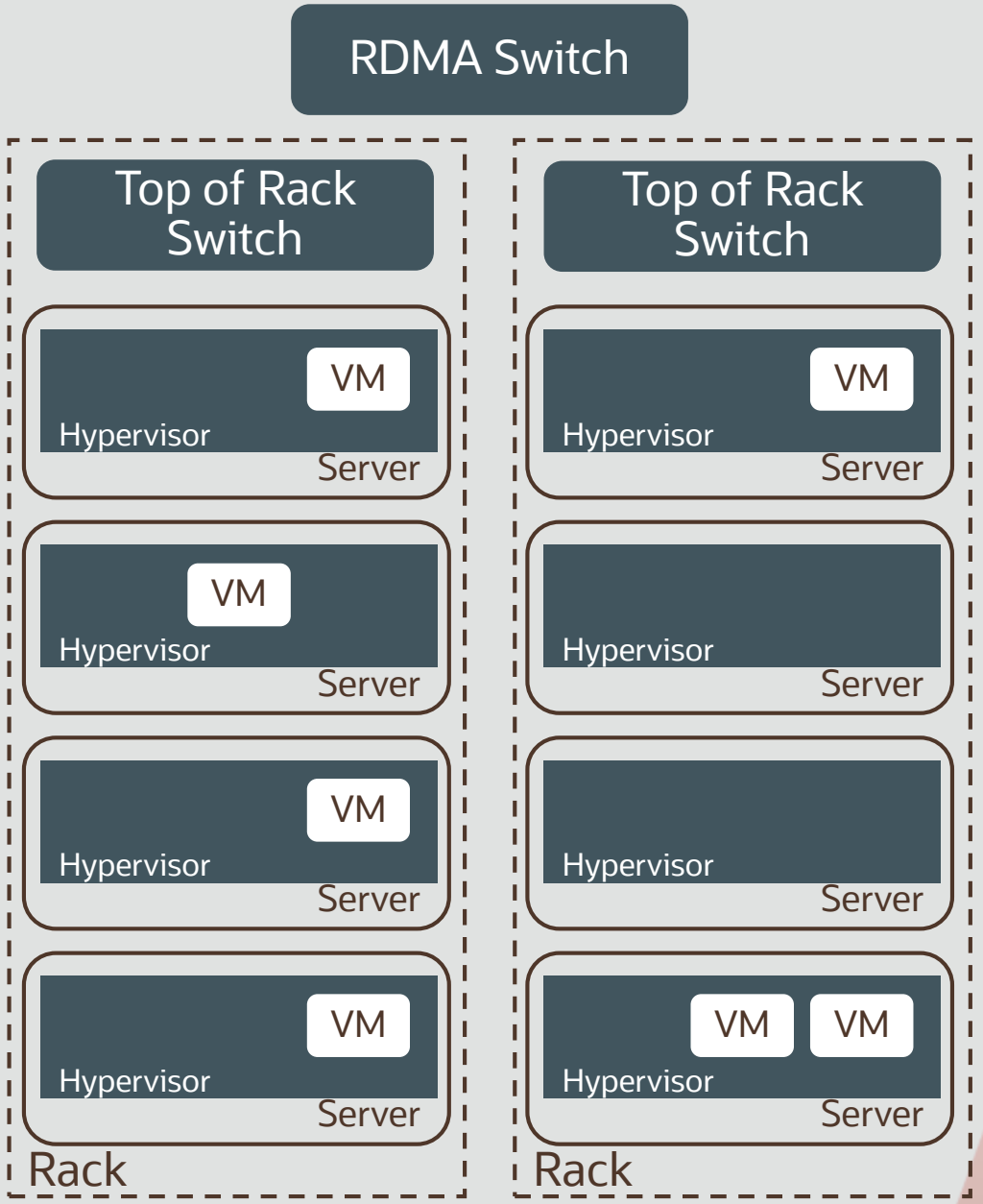
ORACLE

Why are we better?

In other clouds, data has to go through a VM and a hypervisor before it can access the underlying server hardware

Other clouds automatically distribute the VMs across the entire datacenter

This adds significant latency and un-predictability in performance

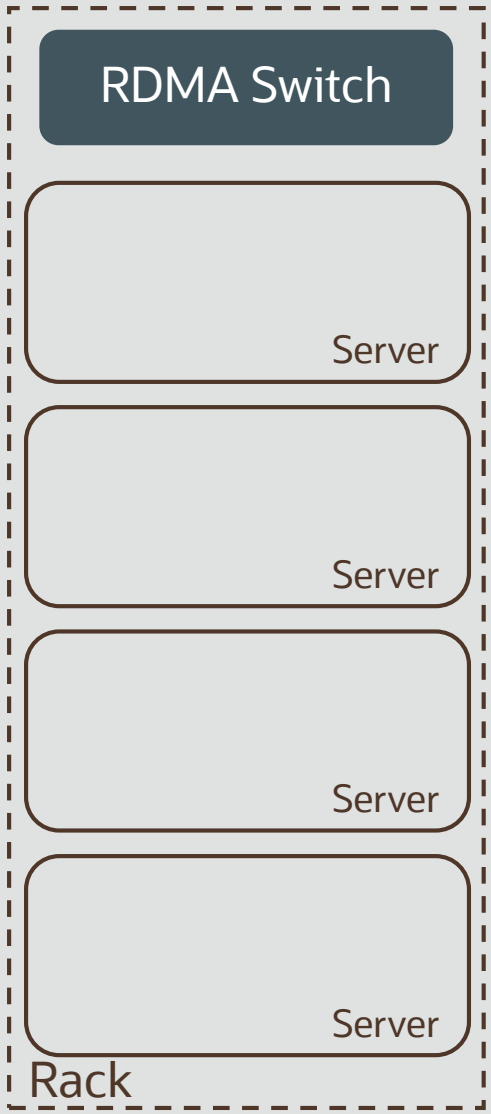


Why are we better?

Oracle connects the servers directly to the RDMA switch

No hypervisor, no virtualization, no jitter

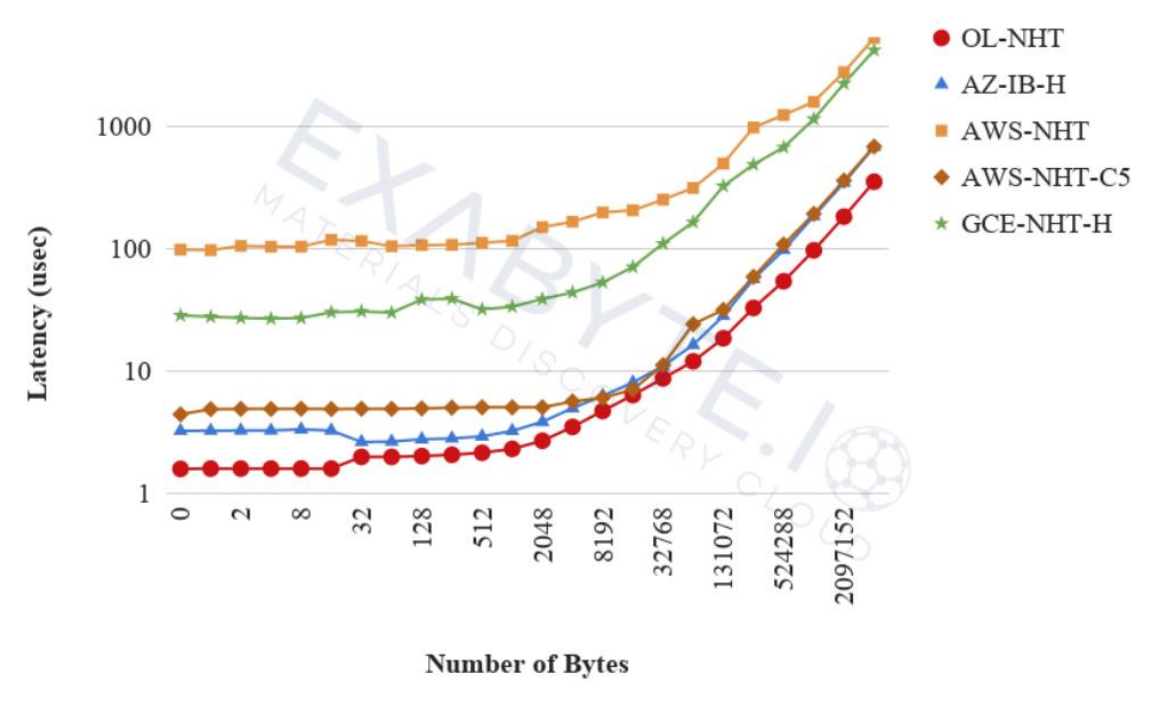
We allow customer to chose their own placement to maximize stability



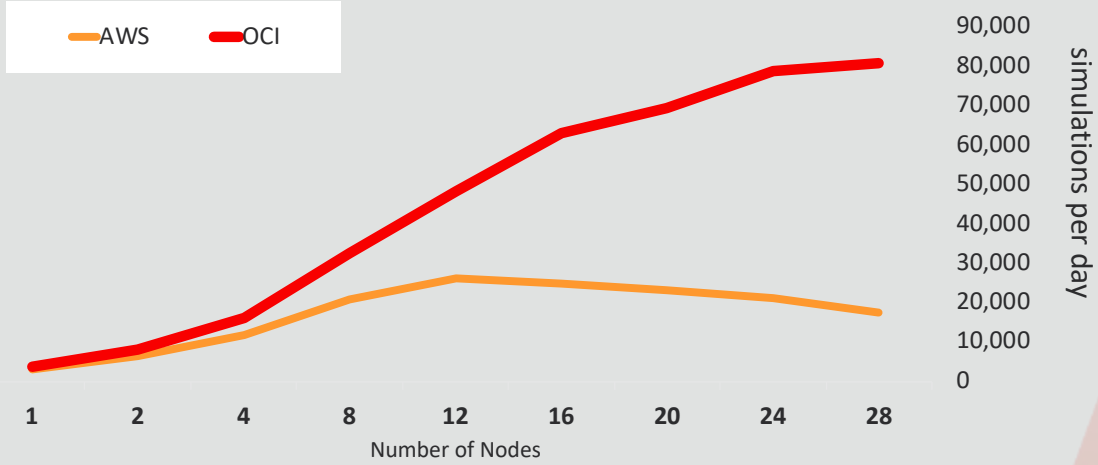
ORACLE

Performance

Latency

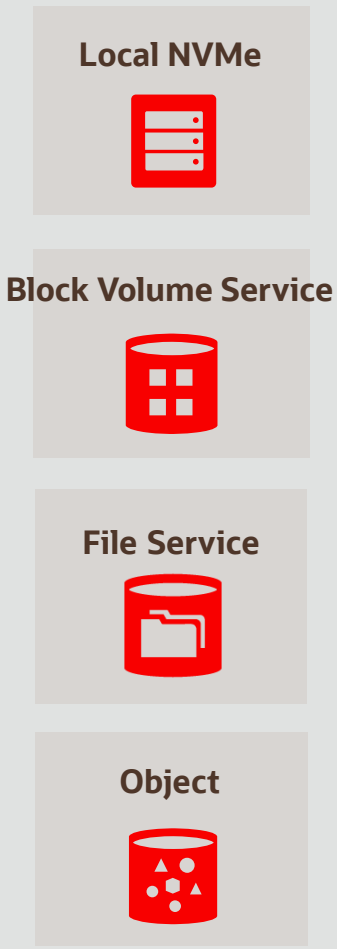


Application Performance



HPC Infrastructure: Storage

The Best Storage for Each Use Case



High performance NVMe SSD storage
- Local to a bare metal compute instance
- Non-resilient: Data doesn't survive beyond instance life

Resilient storage Data is persisted beyond instance life
- Volumes can be detached and attached to different instances

Shared storage Data is persisted beyond instance life
- Shared access or multi-attach with file semantics & scale-out performance

Regional network accessible, durable storage
- Data is replicated regionally for very high availability and durability
- Designed for big data, backup and unstructured content

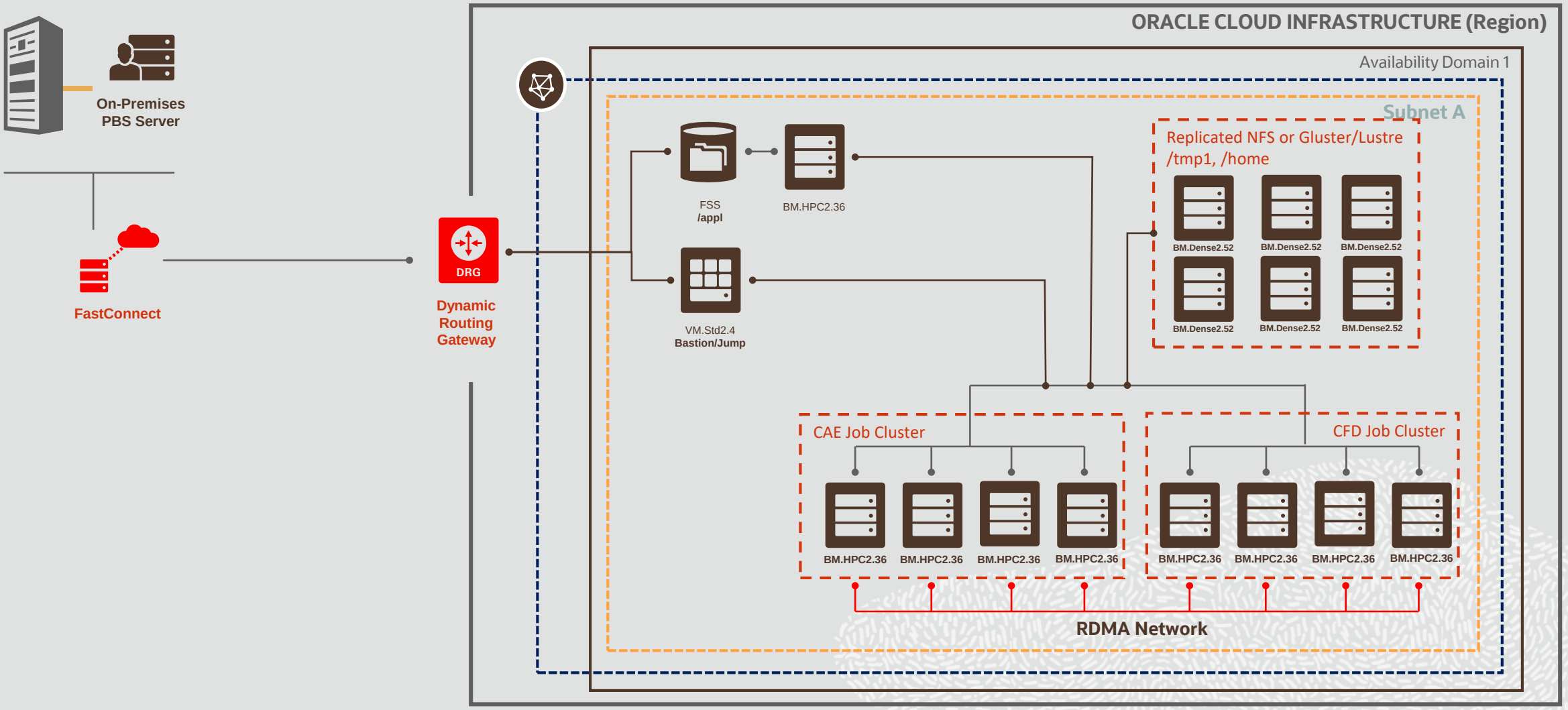
HPC Data Flow

- NVMe drive:
  - Scratch space on each node
  - Parallel file system for the model
- Block Storage:
  - Parallel System between nodes for model
- FSS:
  - Applications, easy to mount on multiple clusters
- Object Storage or Archive:
  - Persistent storage

Data Flow – Option 2

Tier	Description	Size	Technology
Home	Campaign storage file server	100 GB	
Applications	All application binaries, updated weekly	10 TB	FSS
Ingestion	Sync on-prem to campaign storage	5 TB/day	Multiplexer
Long Term Storage	None		
Campaign Storage	File Server created in OCI data sync'd with on prem and served to all jobs	300 TB	7 x BM.DenseIO2.52
Active Scratch (CFD)	Uses campaign storage as active scratch		
Active Scratch (CAE)	local share on each node (/TMP1)	6 TB x num nodes	BM.HPC2.36
Egress	Sync campaign storage to on-prem	5 TB/day	Multiplexer

Architecture Diagram – Option 2



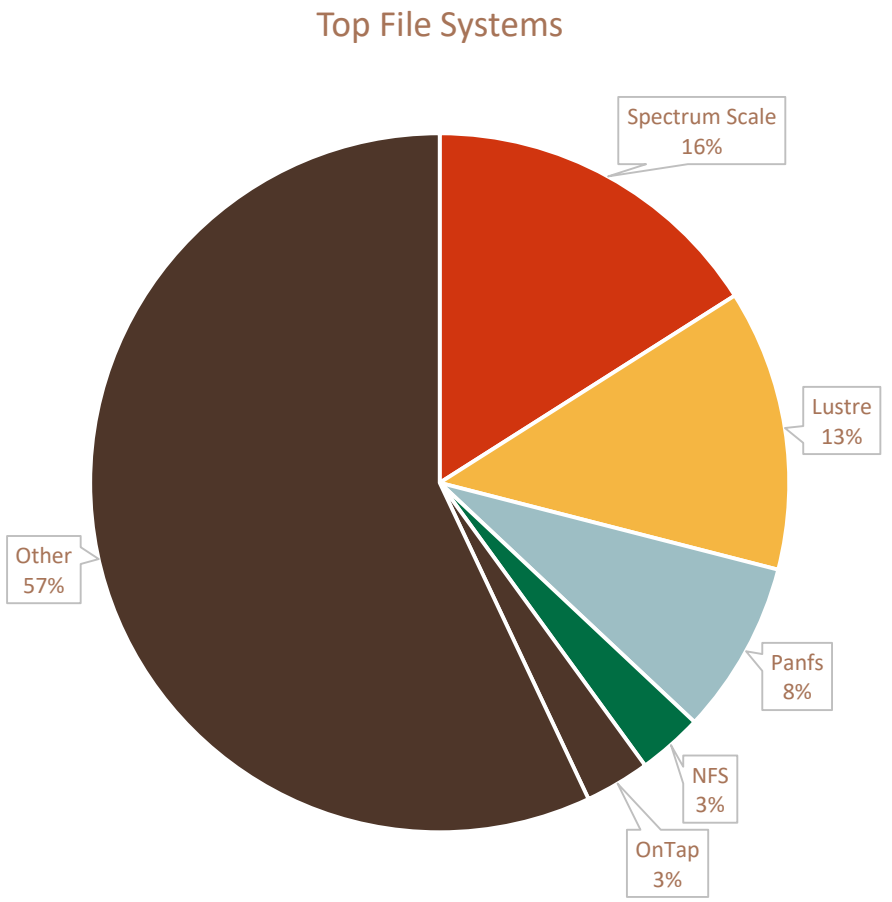
HPC Parallel File System

Parallel File System

- NFS (simple, used for 1 mounted drive)
- GlusterFS (Open-Source)
- BeeGFS (Open-Source)
- Lustre (Open-Source)
- Spectrum Scale (Licensed, high performance)

Market Share

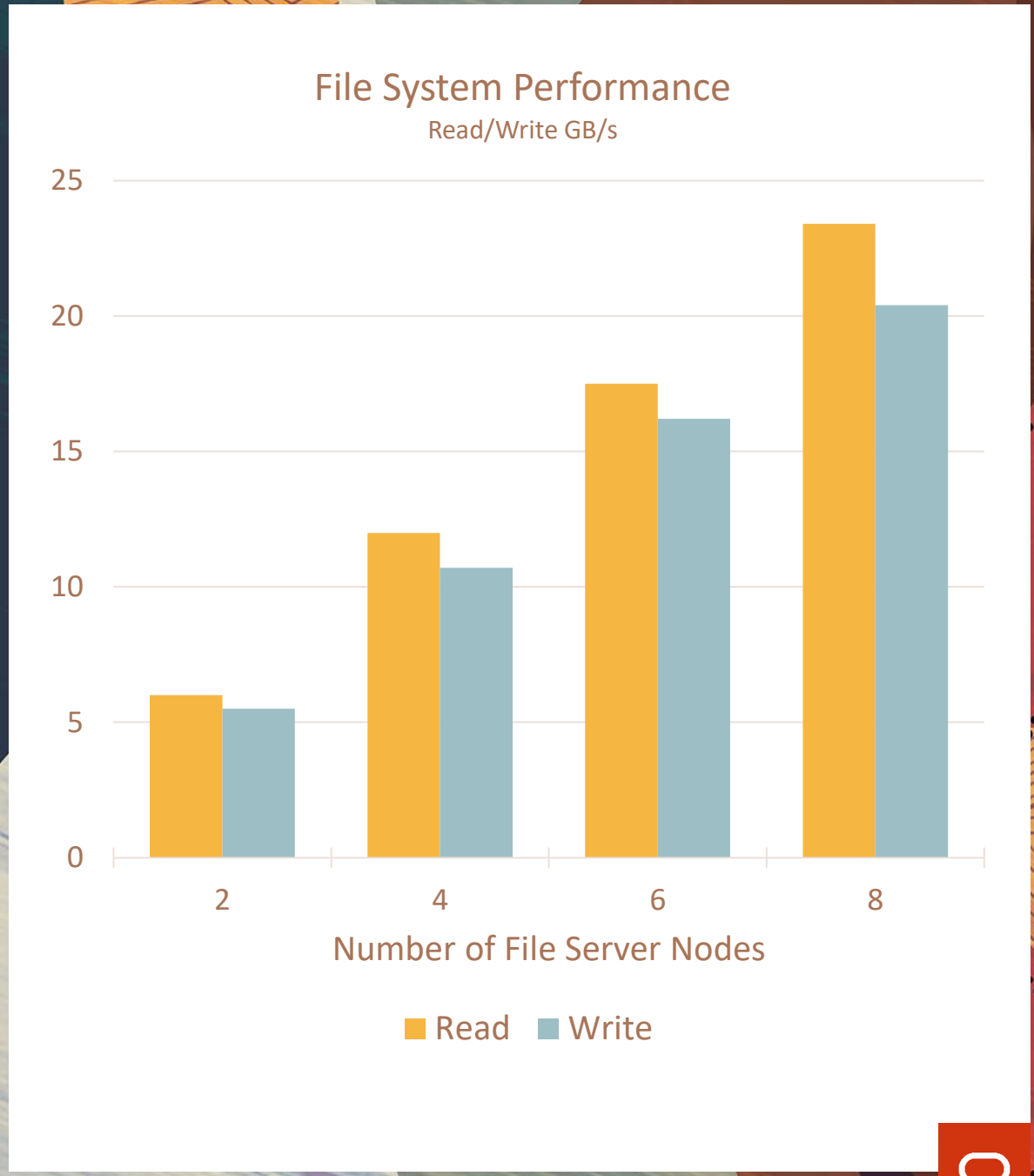
Spectrum Scale makes up the largest single file server provider in the HPC and big data markets



Intersect360, 2018

Spectrum Scale Performance

- On as small as two nodes, IBM Spectrum Scale on Oracle Cloud Infrastructure provides over 5 GB/s throughput
- By adding more building block, the throughput scales almost linearly



HPC Queueing System

Why is it difficult to schedule millions of jobs?

Queuing is horrible

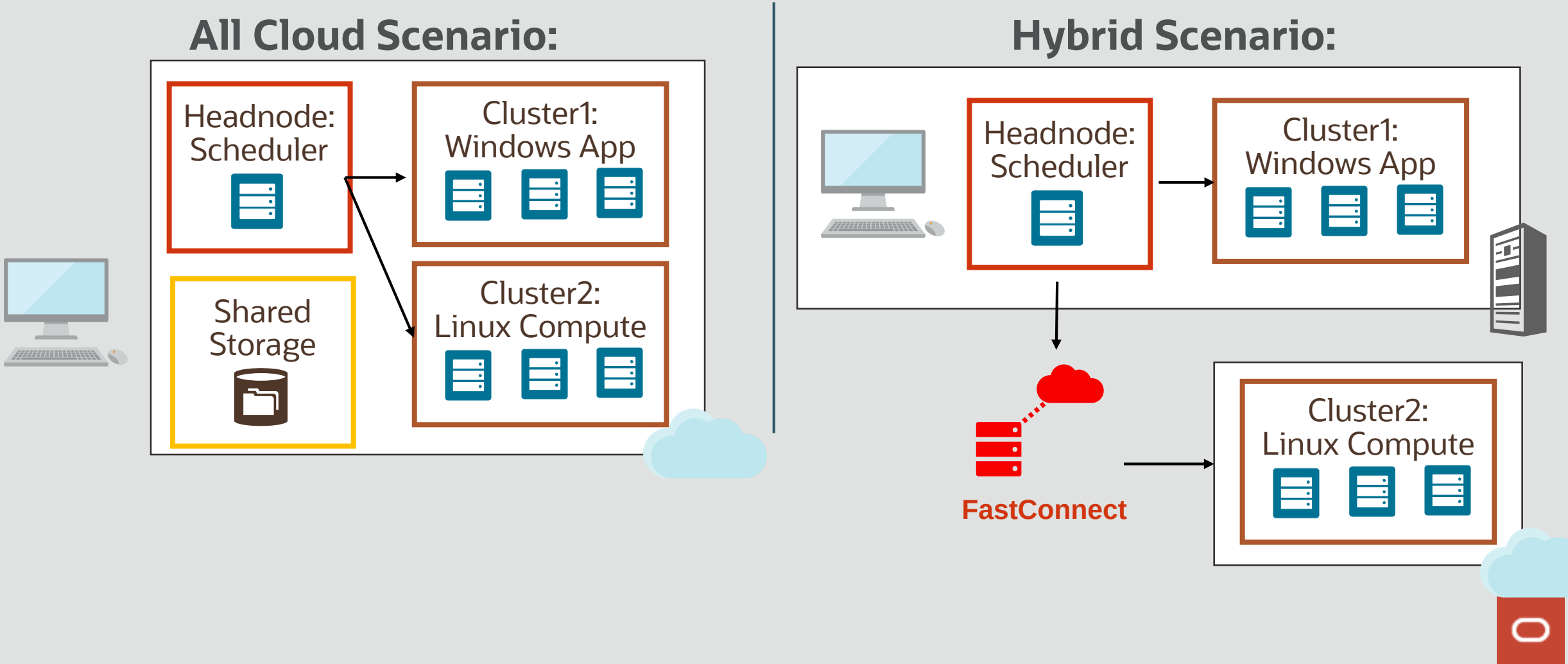
- Prioritization
- Utilization
- Runtimes
- Deployment

Paradigm shift

- On-premise:
  - Full cluster utilization → undersizing and queuing

- Cloud:
  - Size your cluster according to your needs
  - No queuing

Basic Architectures



Basic Architectures

- Scheduler: Slurm, PBS, LSF, Ansys RSM,…
- Workflow manager, optimizer,…

- From low cost archive storage to high performance HPC file system

- Virtualization on the cloud using GPUs instances.

Basic Architectures

- RDMA Network
- Elastic size

- Dedicated Network between on-premise and Oracle Cloud

Queuing systems

Dr. Chris Woods
Researcher, University of Bristol
Founder of Cluster-in-the-Cloud

Set of simulations from 90 days to 5 days

"With Oracle Cloud we were able to introduce a new paradigm into our research that allowed us to take advanced nicotinic research and process thousands of jobs within the same time that it previously took us to simulate a single job"

VAR Calculations

Computations

Simulations 100000

Batches 12

Stock Choice

Initial Date 01/01/2017

Stock Oracle Quantity 100

Stock Apple Quantity 100

Stock Cisco Quantity 100

Stock Intel Quantity 100

Submit

Available jobs

Running

Results Available

47718

HPC Applications

As Oracle Cloud **Infrastructure,** we provide a bare-metal box !

Anything you need can be installed and run !

ISV Ecosystem

Visual Effects Rendering

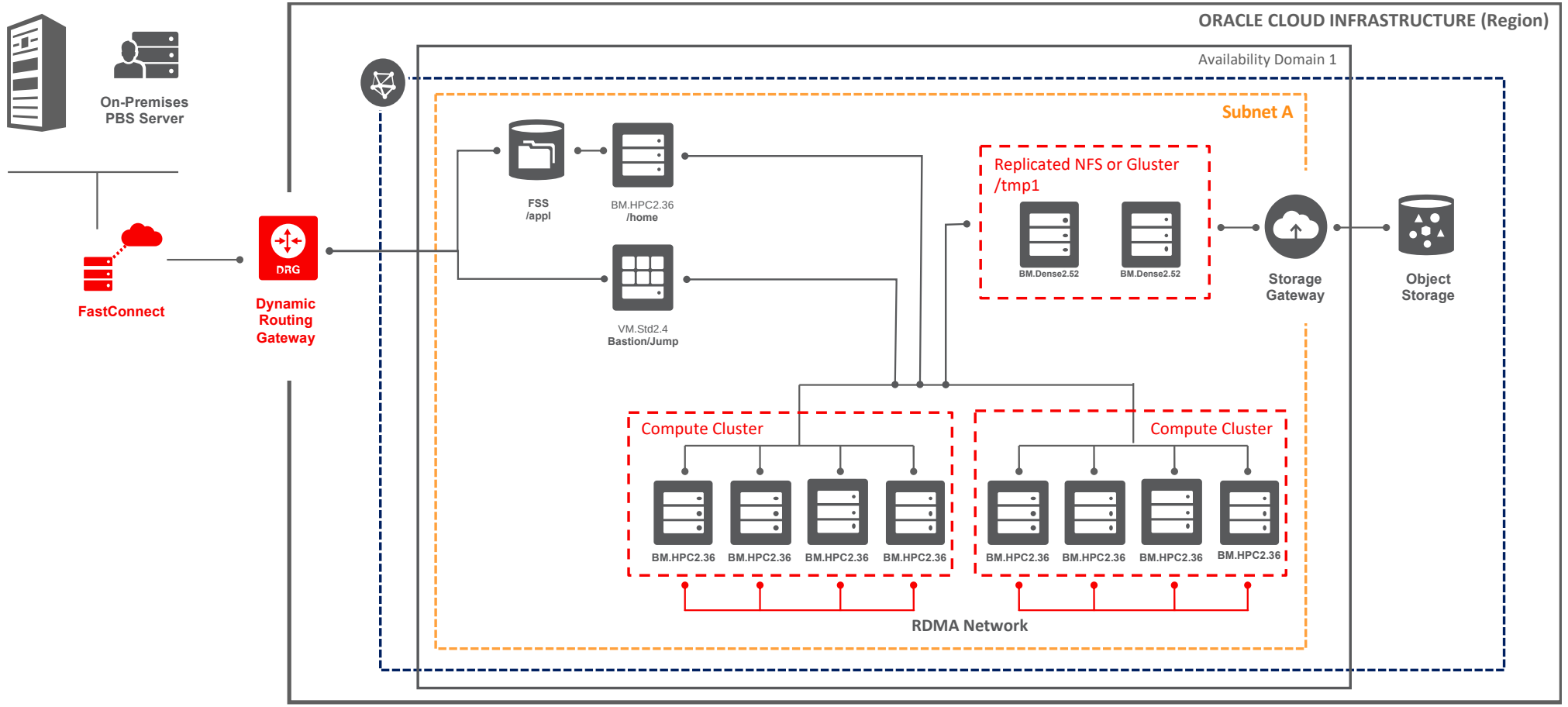
Artificial Intelligence & Deep Learning

Manufacturing Automotive Aerospace

Open Source HPC Applications

HPC Workflows

Common HPC Architecture



HPC Automation

Simple, Rapid Deployment

3 Deployment Models

Command Line

Use Command Line and API's for integration to automated workflows

Stacks

Use Resource Manager Stacks for interface based Terraform Deployment

Marketplace

Use Marketplace for deploying partner provided infrastructure and applications

HashiCorp Terraform

Deliver infrastructure as code with Terraform

Network, instances, storage,…. Anything you can do in the console can be scripted

Open-source

Multi cloud and on-prem

Documentation: https://www.terraform.io/docs/providers/oci/index.html

Oracle Resource Manager

GUI for your Terraform stack

Edit variables

Import or export state files.

Yaml file for GUI specifications.

Example with OpenFOAM: https://github.com/oci-hpc/oci-hpc-runbook-openfoam/blob/master/Documentation/ResourceManager.md

OpenFOAM webinar: https://go.oracle.com/LP=84085

Red Hat Ansible

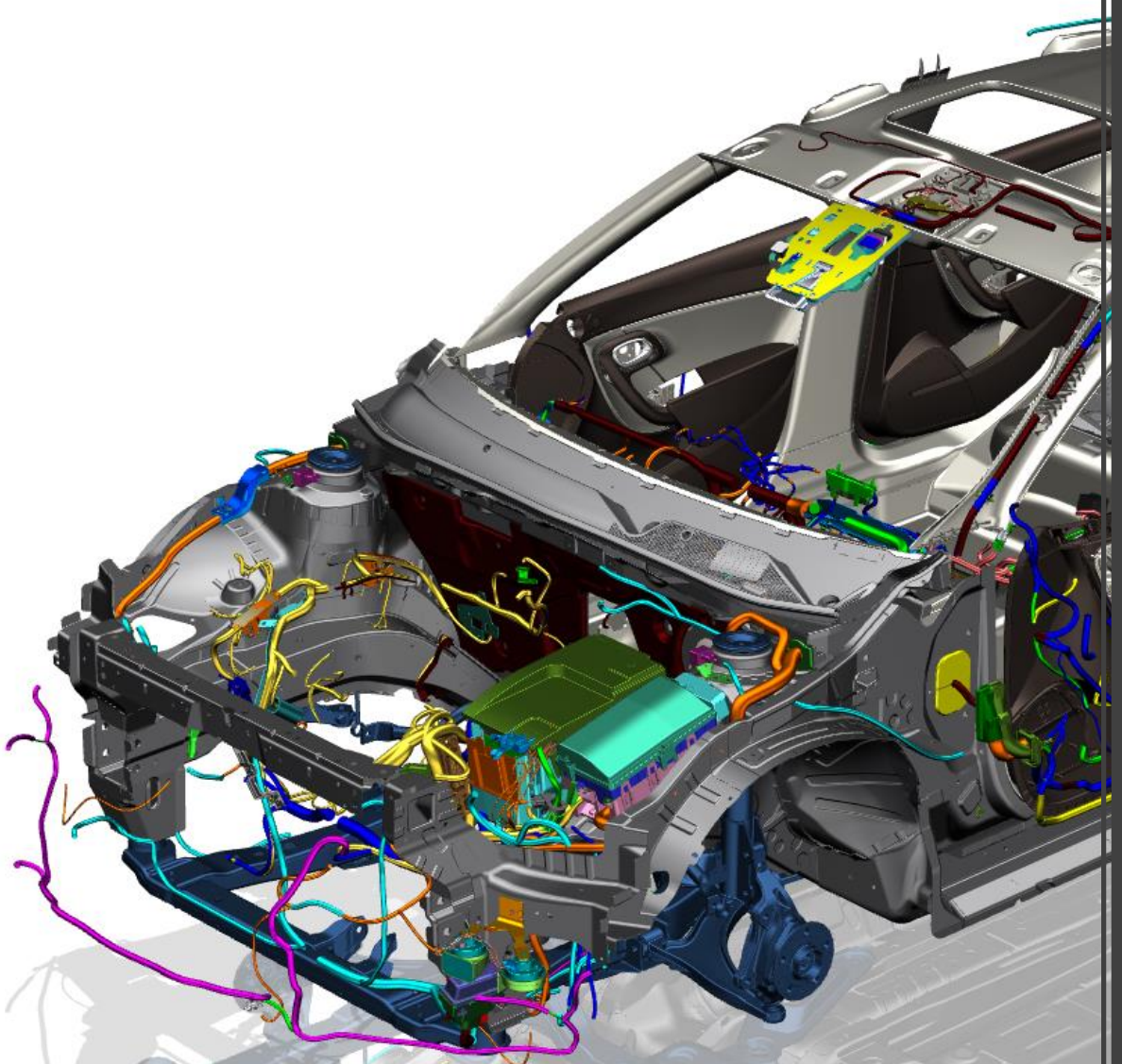
- Terraform is preferred, but some user are already using ansible
- Not available on Ansible Repository yet
- Get Started: https://docs.cloud.oracle.com/iaas/Content/API/SDKDocs/ansiblegetstarted.htm#prerequisites
- Github: https://github.com/oracle/oci-ansible-modules

HPC Virtualization


DECODING
DECODING

CITRIX® Features

- Provision the VMs through the hypervisor
- Handles the GPU distribution
- Creates your Virtual Desktop
- Handles the remote desktop session



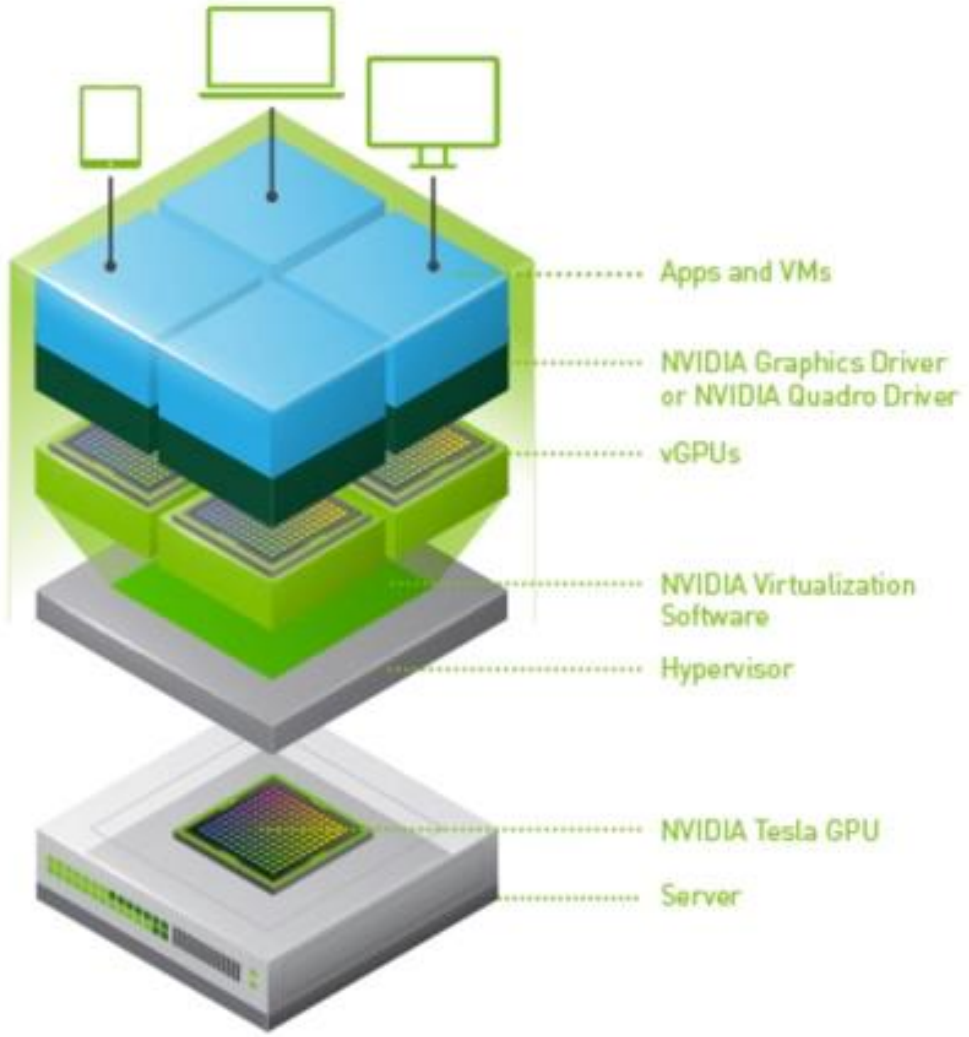
NVI) IA G1I) DC/ (I

- LimiHed AvaiAabiAHy
- v) 6 2 (VirHIaA) aHaceCHer 6 DrkGaHiDC DC 0aGcaADr VDAa baGed G0U lCGaCceG
- UGe 6 iCdDwG2erver 2012/2016 Dr aCy DHher LiCI L) iGHibI HiDC
- (iHiL H) X 3) 0rD G ppDrHed
- 3eradici (ADI d AcceGG2DfHware 2I ppDrHed

CITRIX®

teradici®

GPU Virtualization



- Hypervisor needs to be compatible with NVIDIA virtualization software and drivers.
- 2 Methods for GPU acceleration:
  - Passthrough: Give entire GPUs to VMs
  - vGPU: Divide each GPU in multiple smaller vGPUs
- GPU virtualization always requires some NVIDIA Grid License

CITRIX®
virtualization software

XenServer

- Citrix own Hypervizor
- Not supported on OCI

XenApp

- Virtualize applications

XenDesktop

- Virtualize entire desktop

To run CITRIX on GPUs, you need NVIDIA GRID as the “engine” of CITRIX

NVIDIA GRID™

NVIDIA virtualization software

vAPP Edition

- Virtualize Windows Apps

vPC Edition

- Virtualize entire desktop

Quadro® vDWS

- Virtualize entire desktop with high graphics (4K) and other Quadro features.

CITRIX® Supported Hypervisors

Generates Virtual Machines

XenServer

- Not supported on OCI
- Shared GPU: vGPU
- 1 GPU/VM

Microsoft Hyper-V

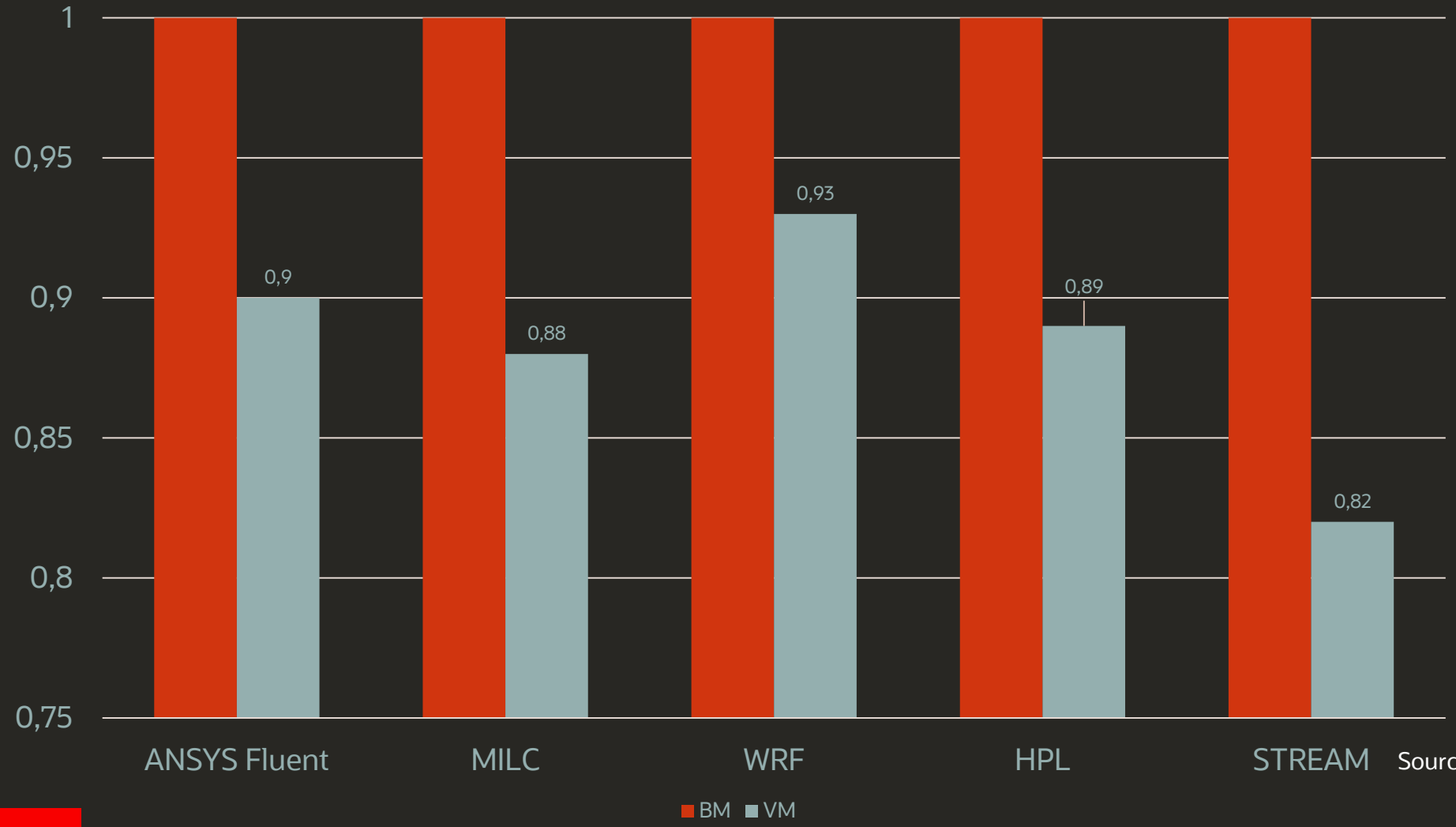
- Runs on Windows
- Shared GPU : RemoteFX
- 1 GPU/VM: DDA

VMWare vSphere

- **Coming Soon** to OCI
- Shared GPU : vSGA
- 1 GPU/VM : vDGA

HPC Performances

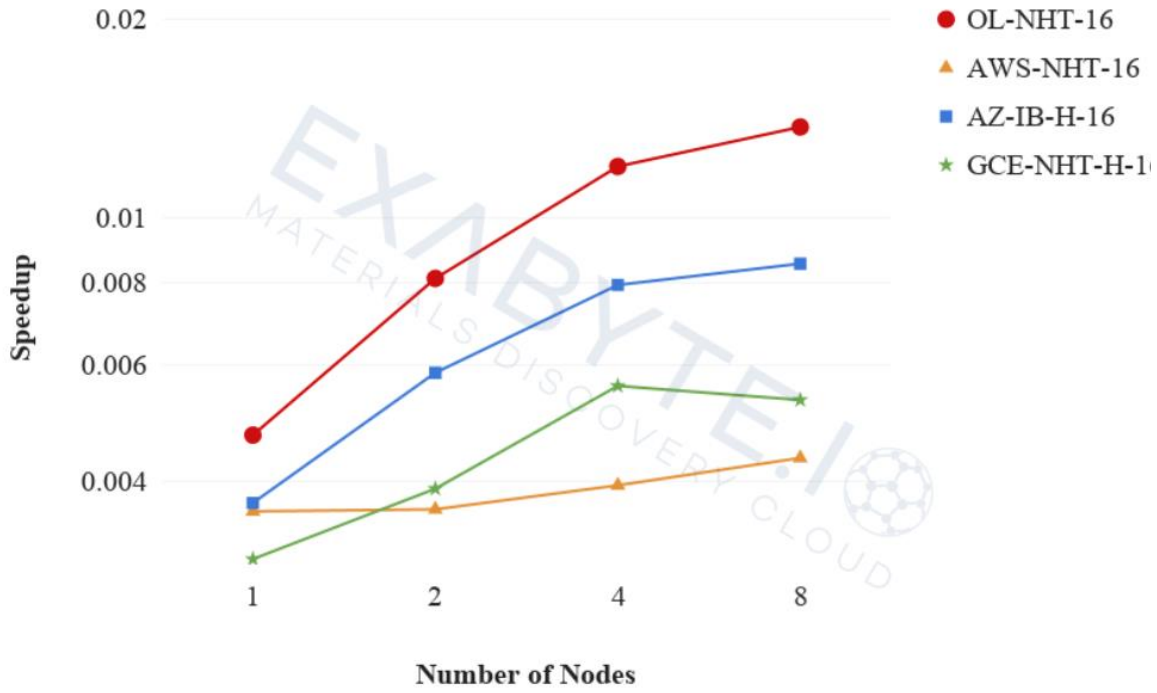
Normalized Application Performance Bare-Metal vs Virtual

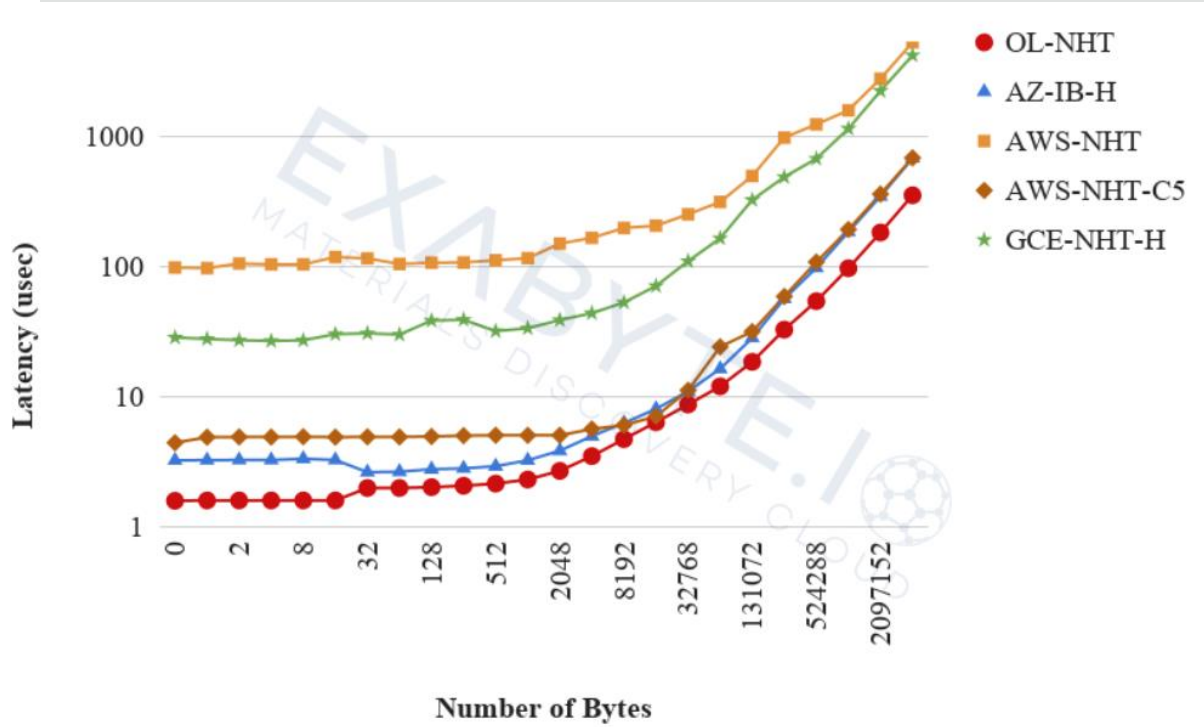


Source: Dell

ORACLE

GROMACS Performances by Exabyte.io





"We benchmark the performance of the latest cloud hardware with HPL, two VASP simulation cases, one GROMACS case and MPI Benchmarks. Our findings demonstrate that Oracle Cloud outperforms other cloud vendors due to the latest generation of the hardware and fast interconnect network." Exabyte.IO team

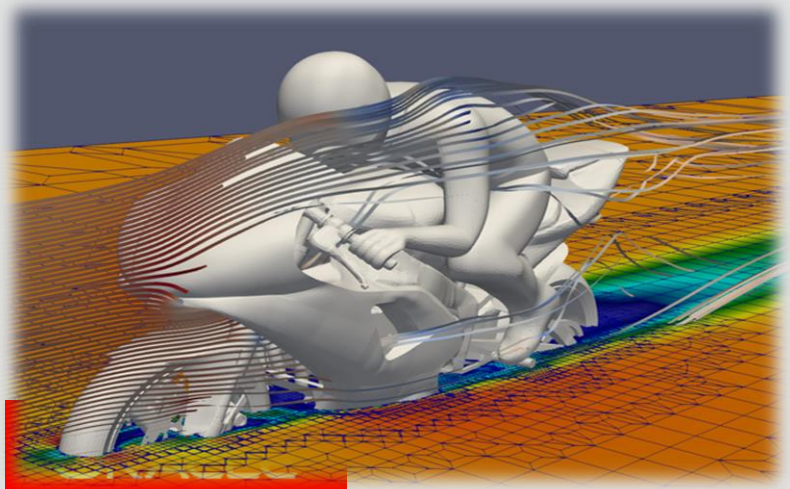
EXABYTE.IO

OpenFOAM

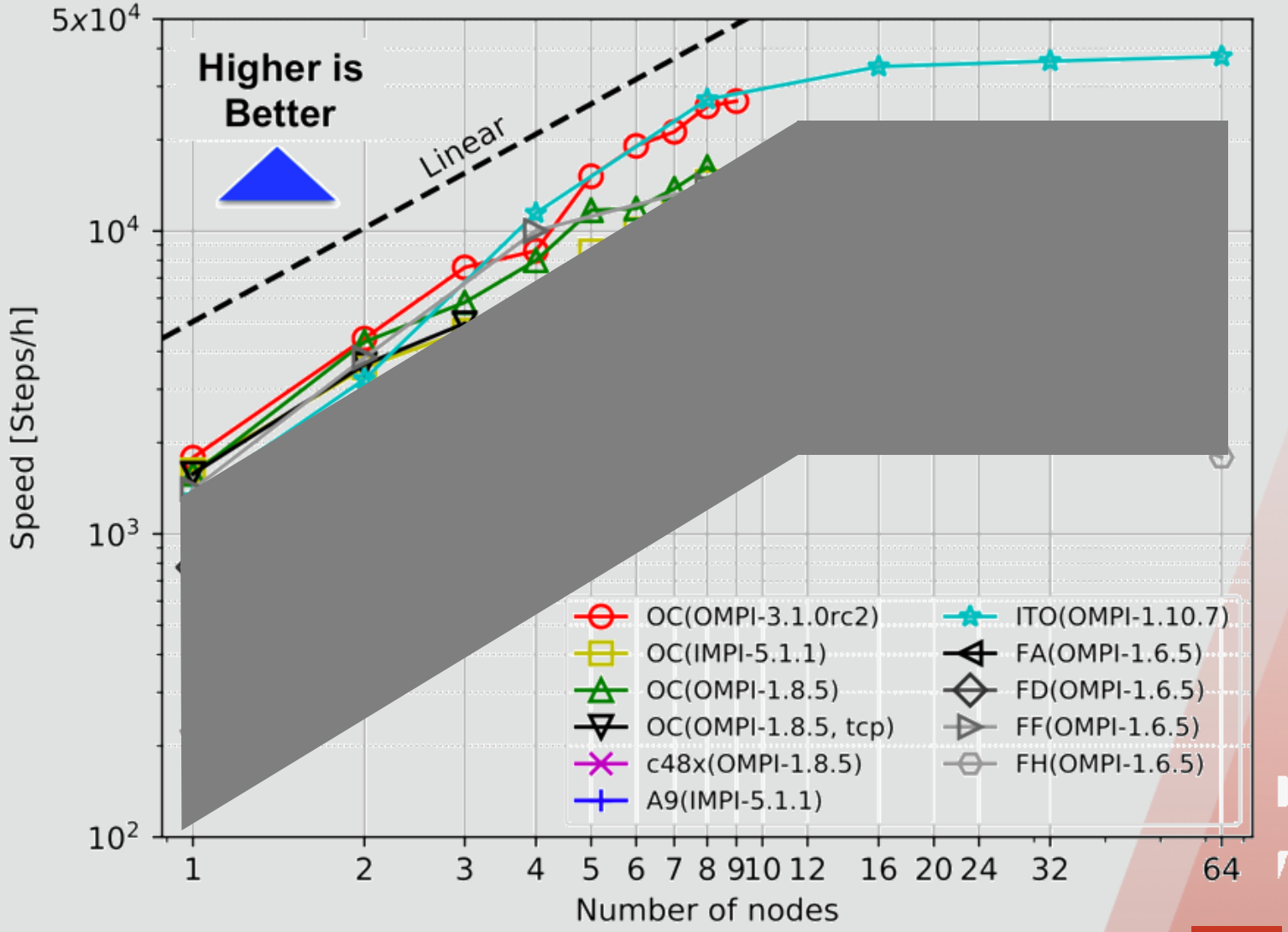
Same performance as an expensive on-premises system

List price is only $2.7/hr per node

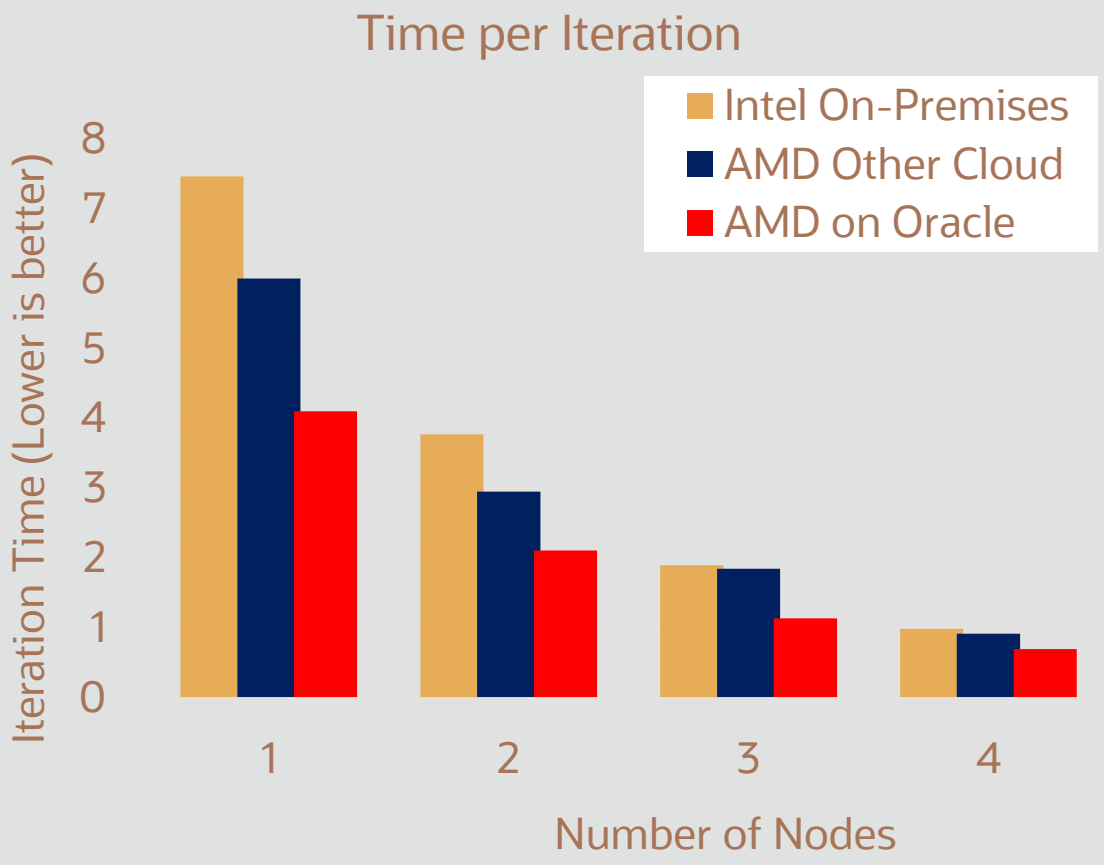
Scales linearly

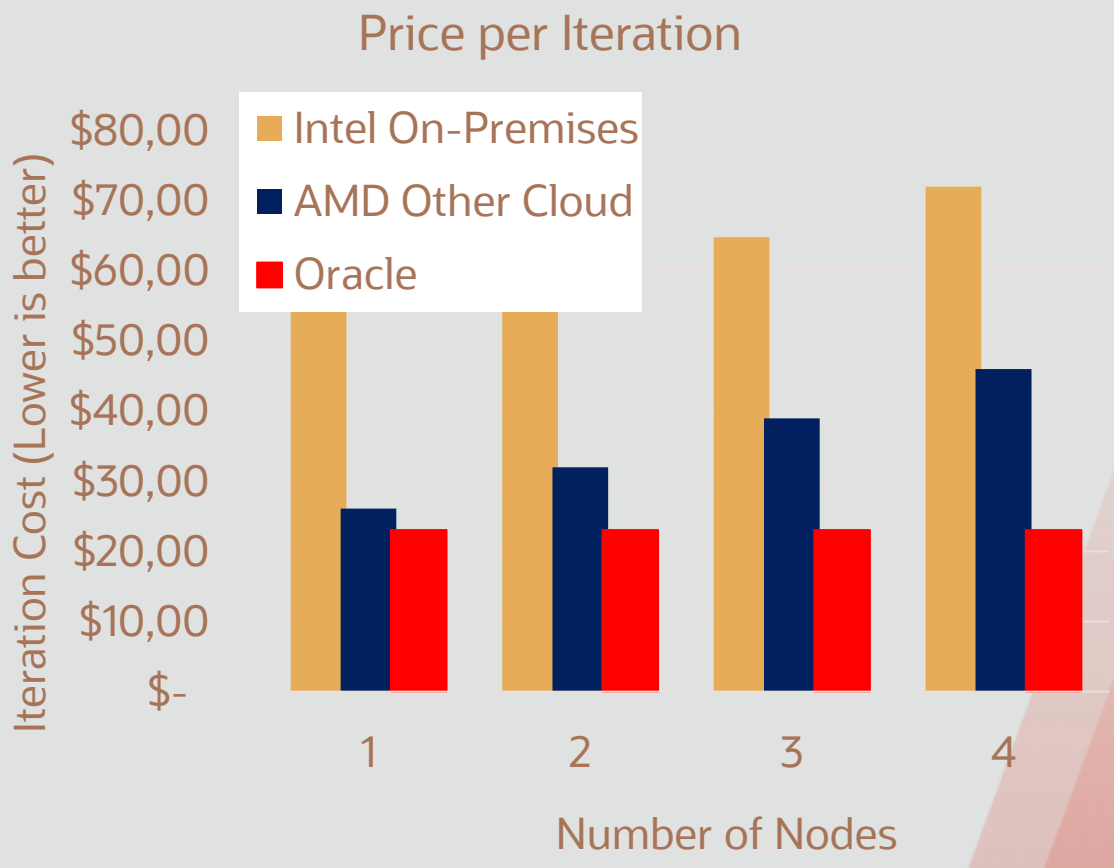


Performance – 3M mesh, use all cores



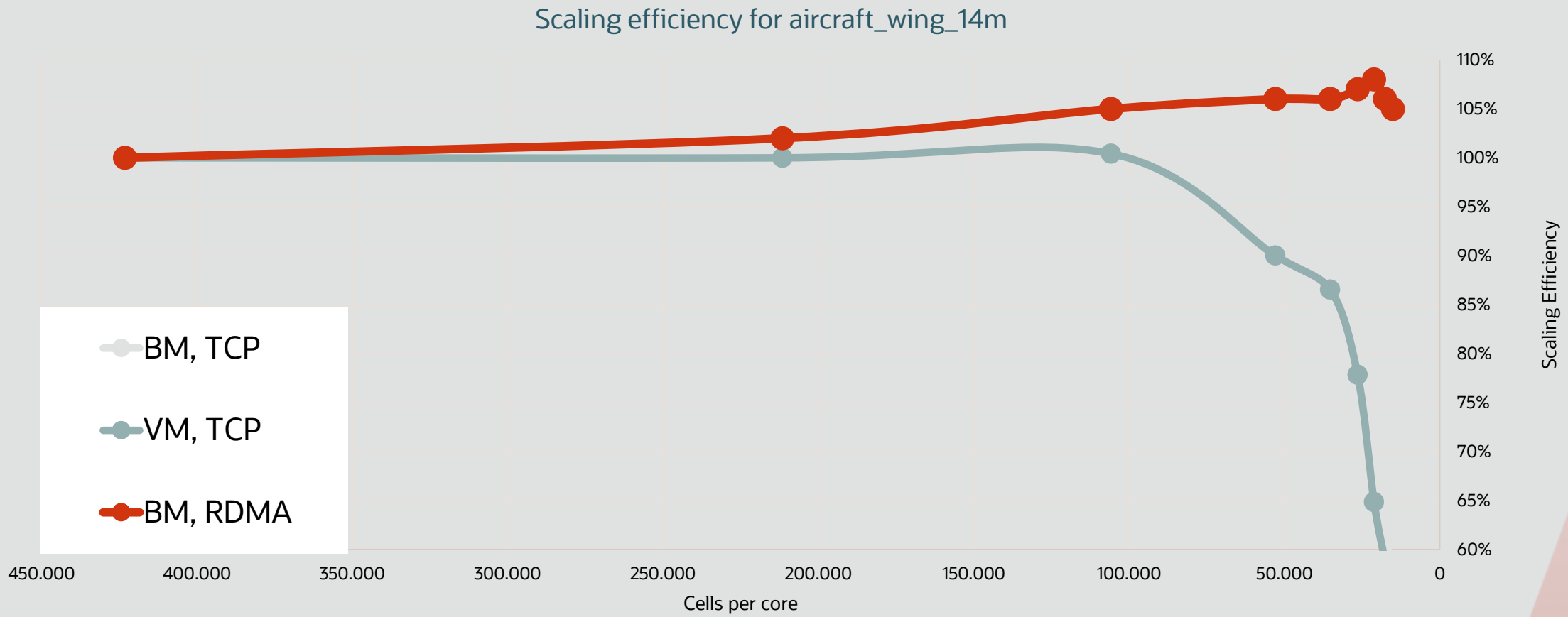
Price Performance StarCCM+ on AMD



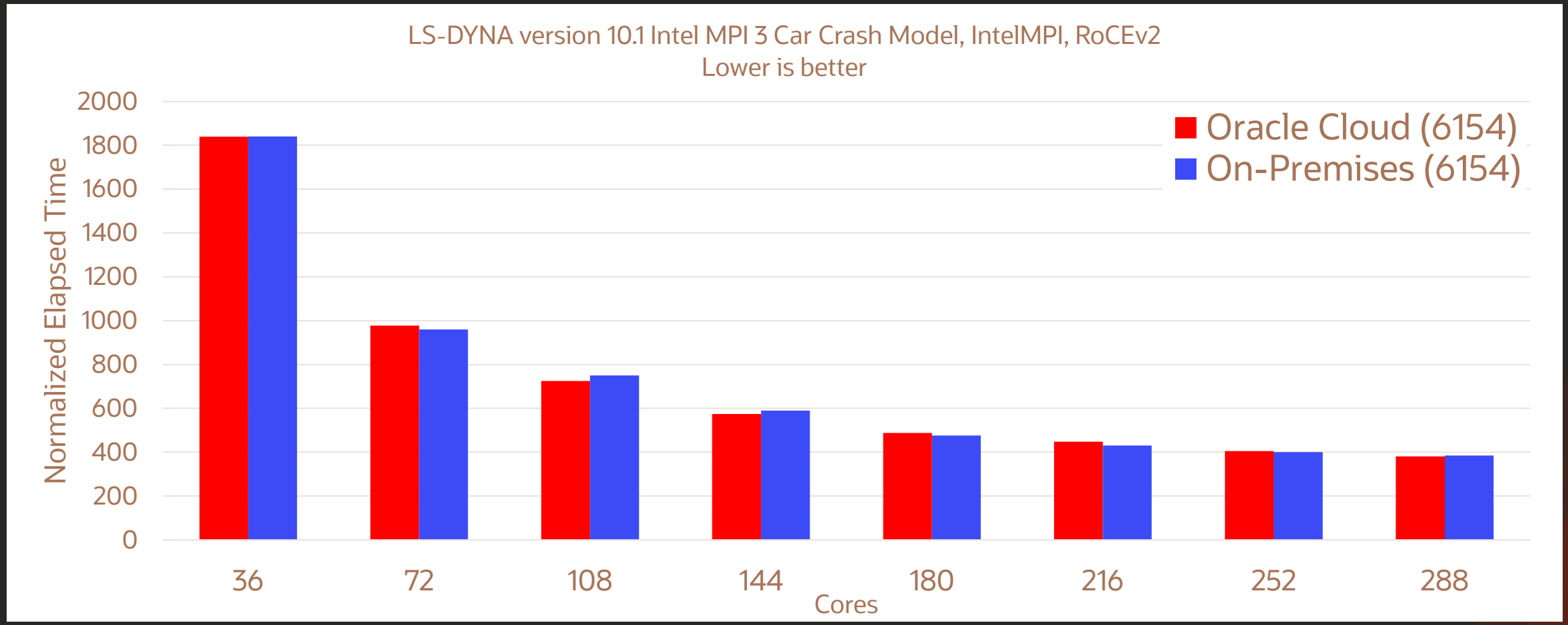


Fluent Performance

Fluent Scaling

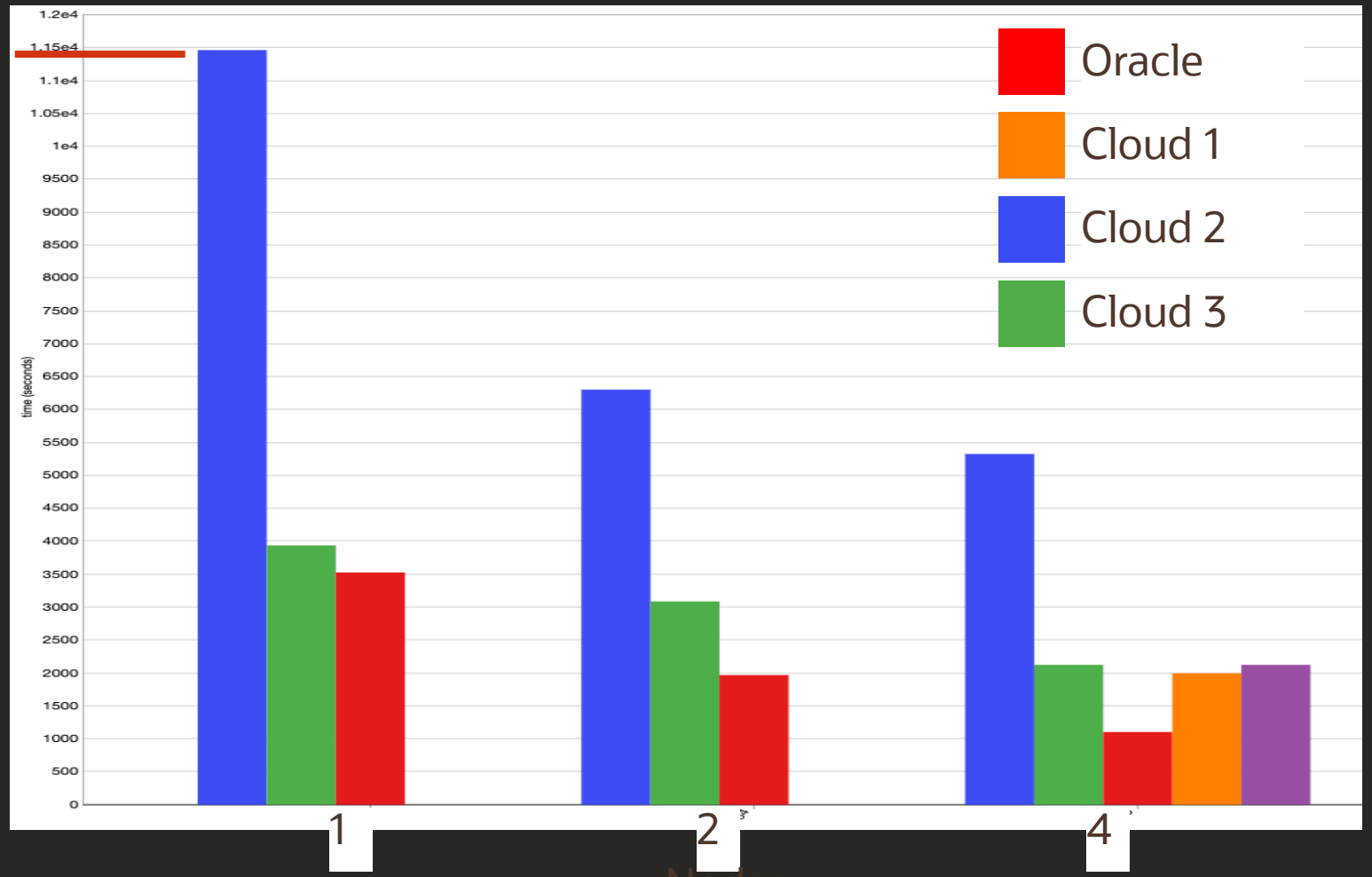


On-premises Bare Metal vs. Oracle Cloud Bare Metal



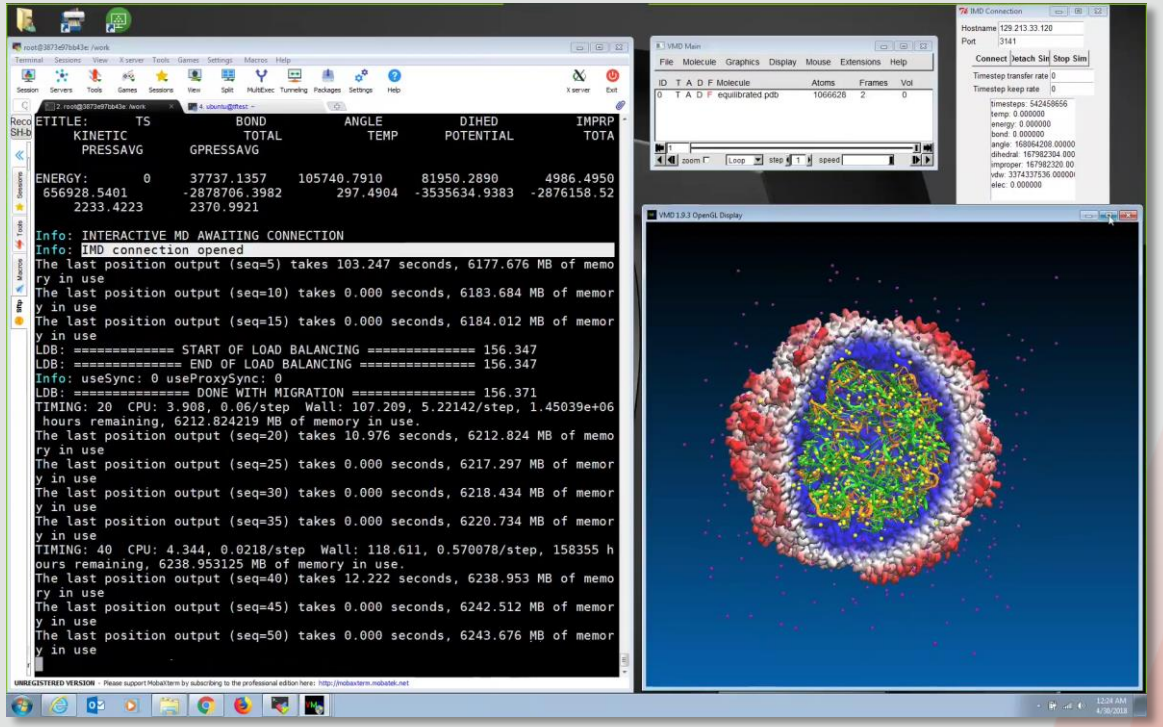
Competing cloud performance for LS-Dyna, prod model

3.1 Hours

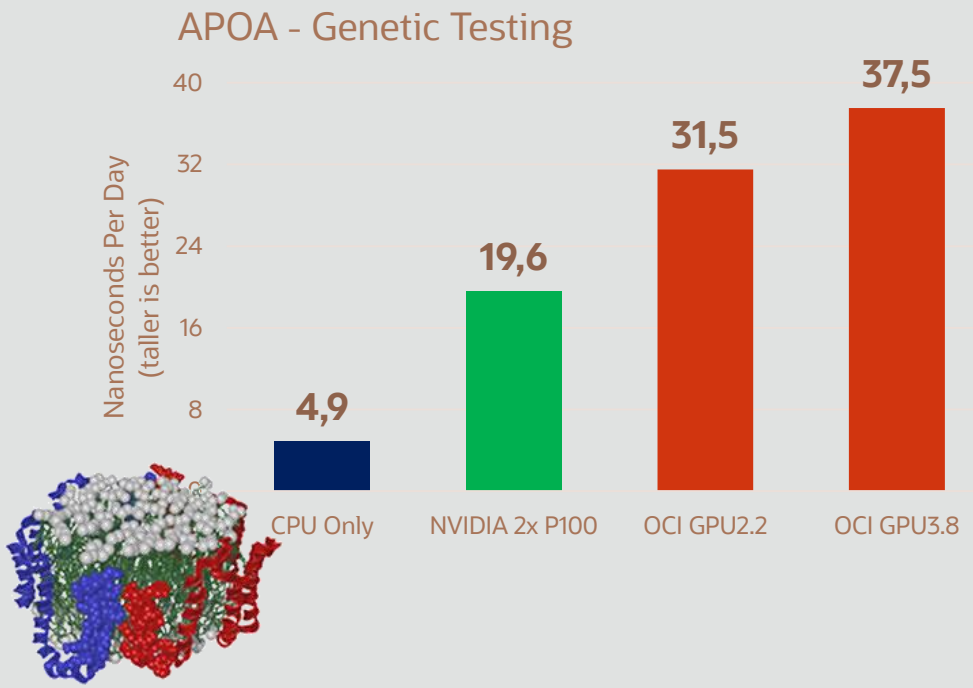


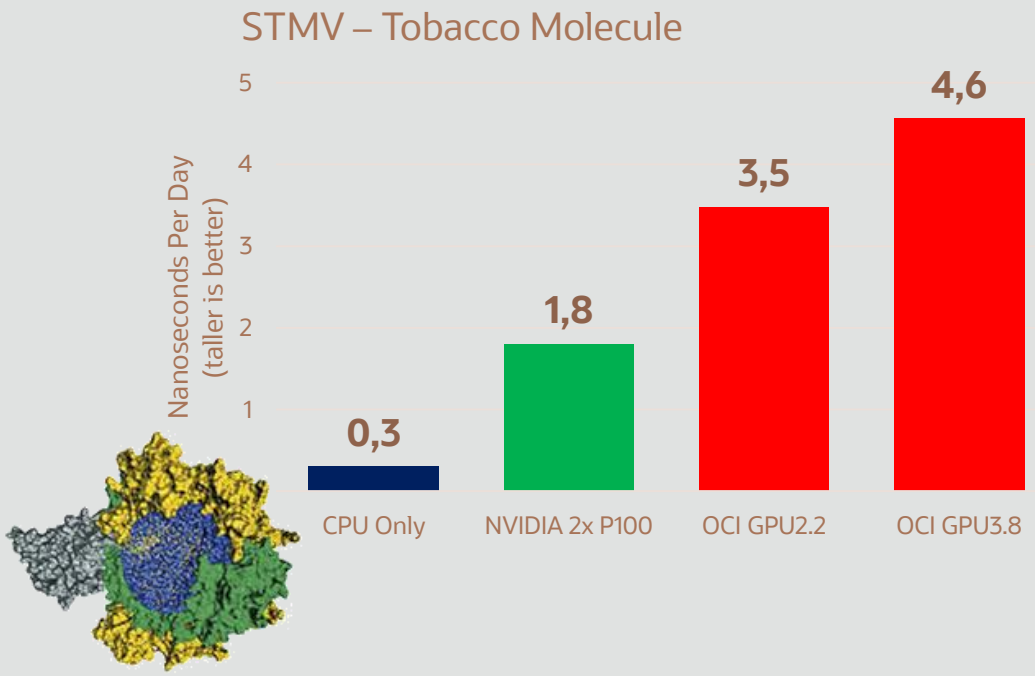
Molecular dynamics and NAMD

- MD to understand the structure and function of biomolecules
  - proteins, DNA, membranes
- NAMD is a production quality MD program
  - Active use by biophysicists (science publications)
  - Features and “accessories” such as
    - VMD: visualization and analysis
    - BioCoRE: collaboratory
    - Steered and Interactive Molecular Dynamics



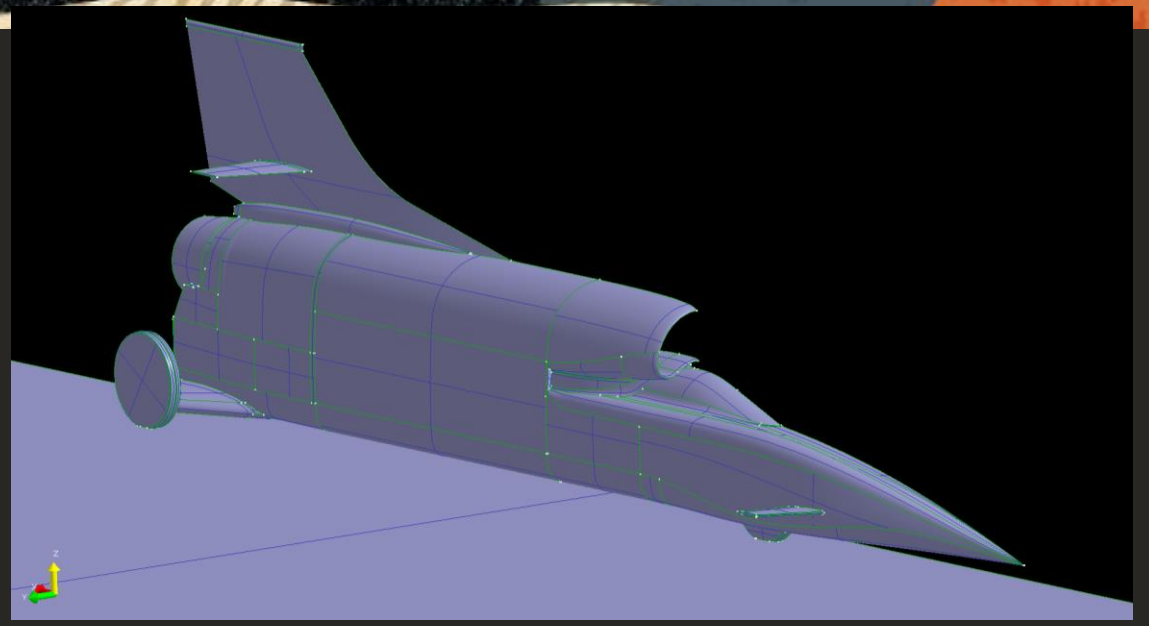
OCI Outperforms NVIDIA Benchmarks



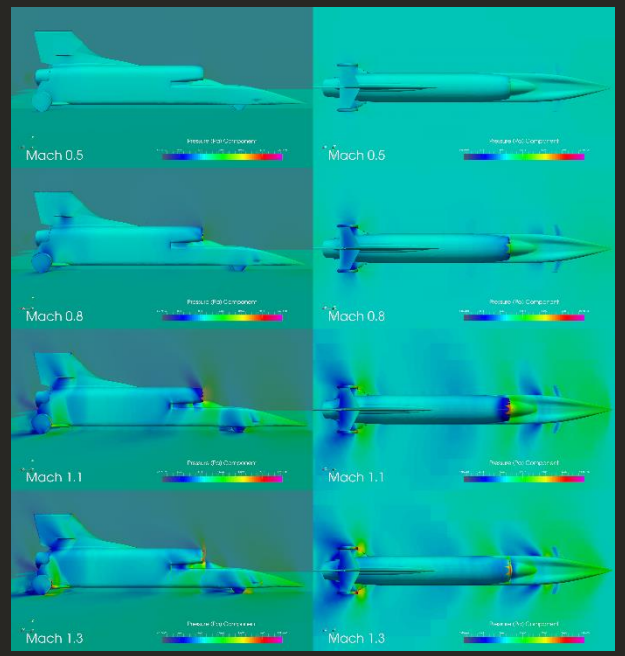


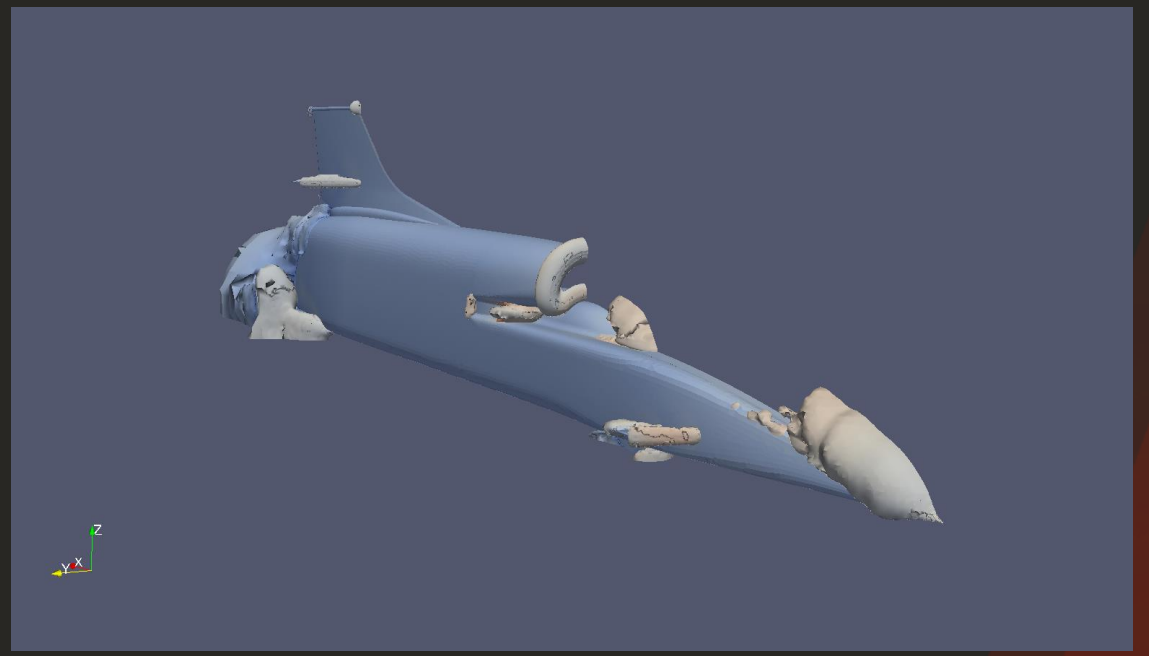
HPC

Zenotech



ZENOTECH
SIMULATION UNLIMITED





ORACLE

Zenotech

Engineering Simulation

High Performance Computing (HPC) for air flow simulation enables Zenotech to design racecars and airplanes faster, with greater fidelity and at a lower cost then using their own supercomputer

Solution

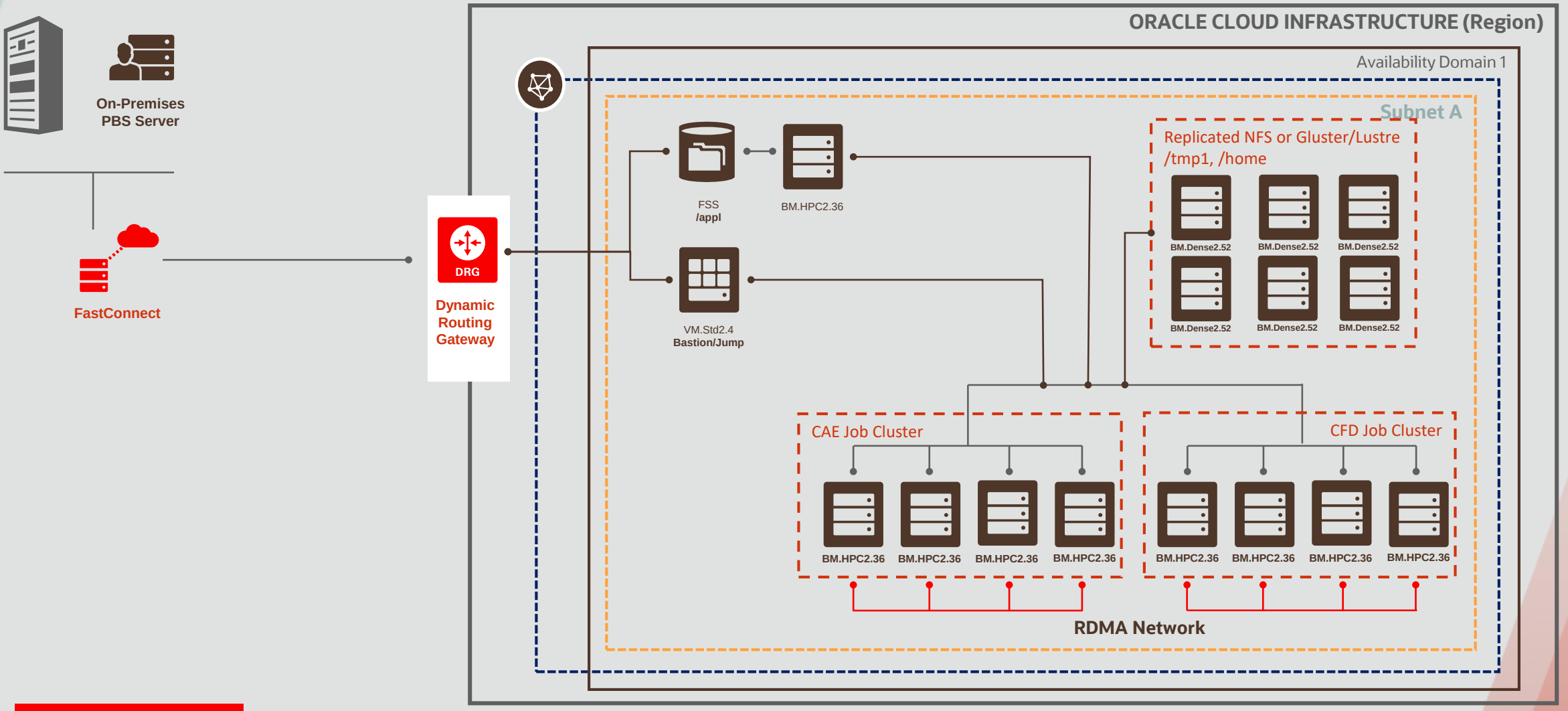
30 Oracle Cloud Infrastructure bare metal compute instances, each with 36 cores, 256 GB of RAM. Combined with a GPU for the visualization of the post-processing data, which enabled the large post data to stay in the cloud.

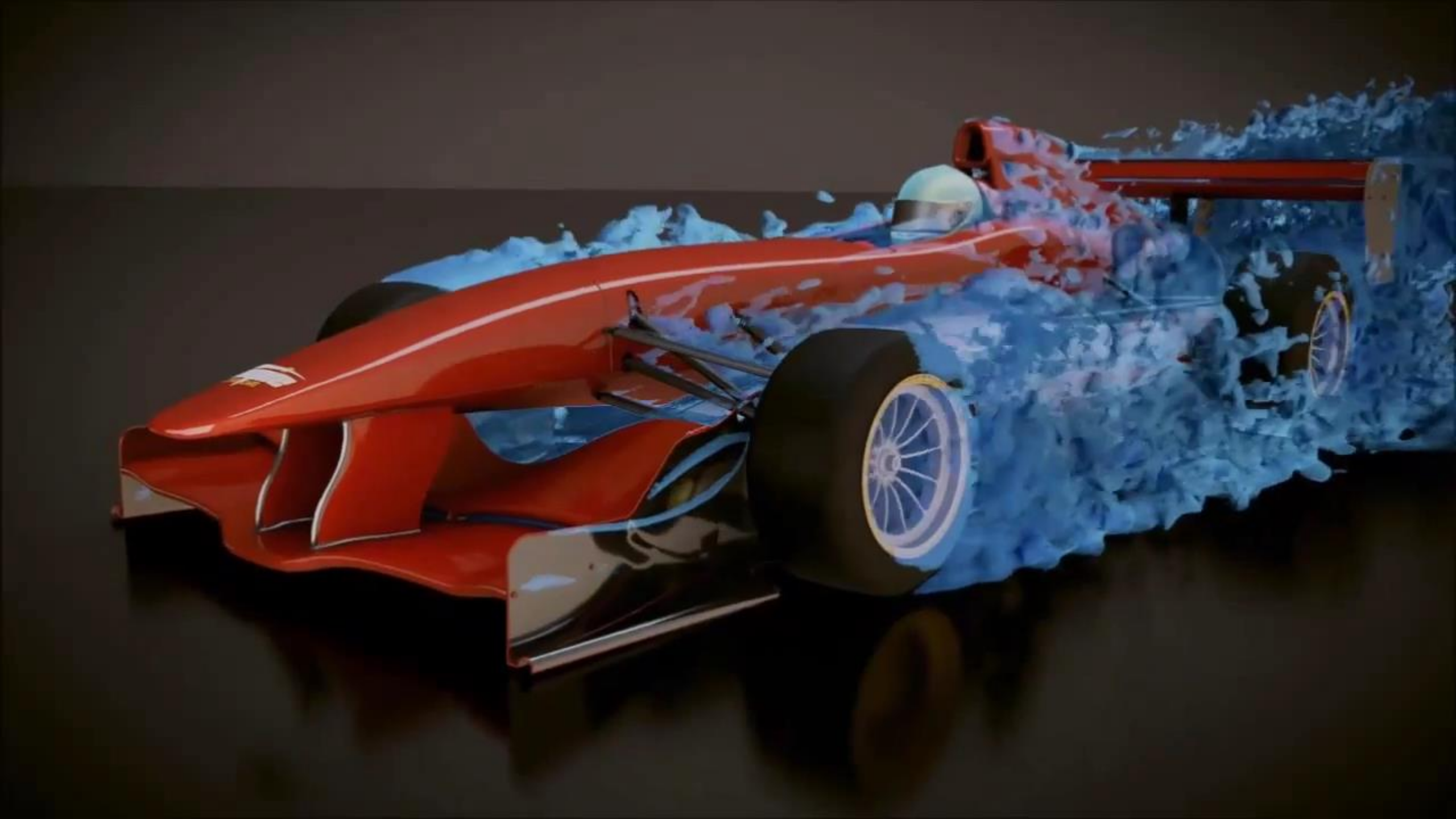
Scale

At 1296 cores, across three availability domains, CFD scaled at 76% similar to the Darwin cluster. OCI's flat network and high core count nodes enable this scaling performance.

ORACLE

Zenotech Architecture





Thank you