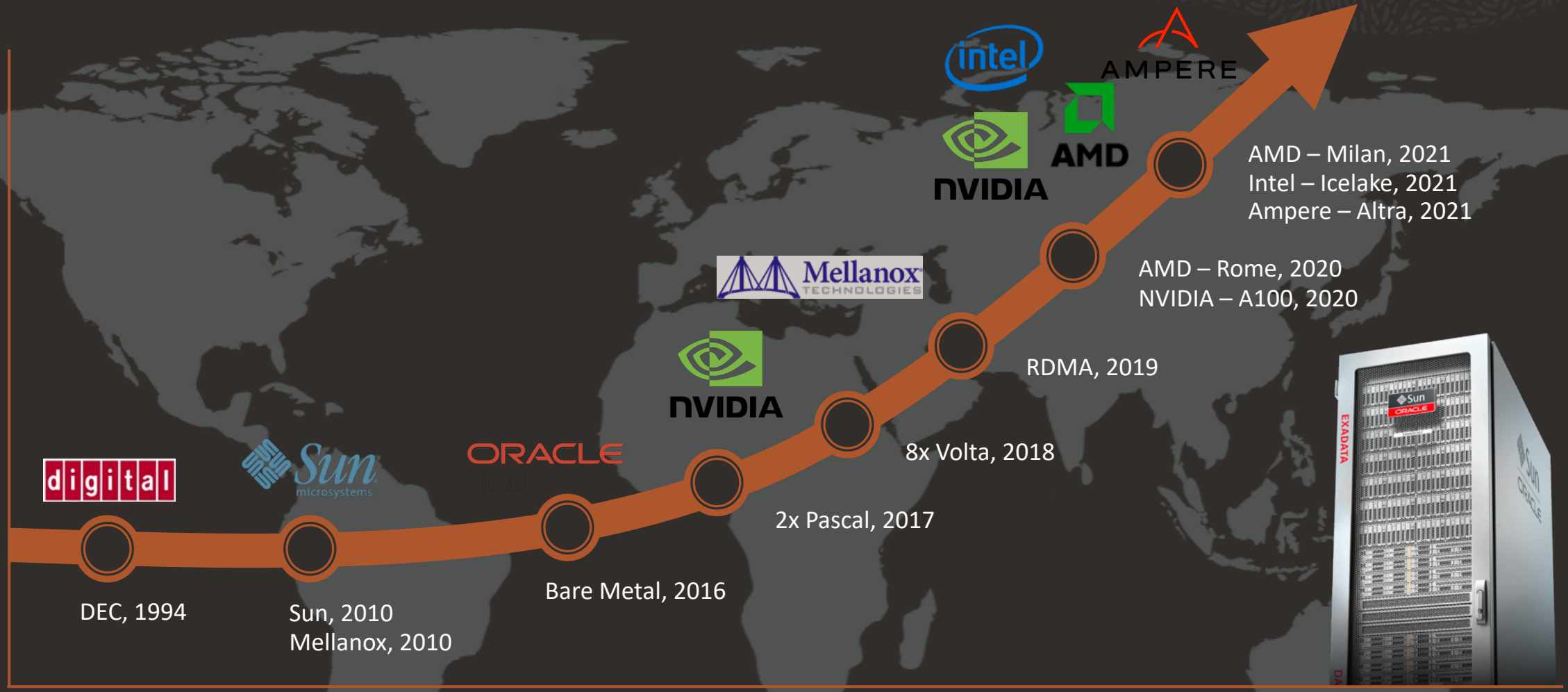
Tutorial

Running scalable clusters on Oracle Cloud Infrastructure

Agenda

1. Introduction to Oracle Cloud Infrastructure
2. HPC on Oracle Cloud (OCI). Compute, Storage, Networking
3. Automated deployment
4. Autoscaling
5. Demo
6. Questions

Oracle’s HPC journey



HPC is at Our Core

- Global **hyperscale regions** with backbone network
- Truly **non-over subscribed, fast, and predictable** networking
- Latency of **1,5 to 3.5µs** between nodes

Oracle Cloud Infrastructure – Global footprint

36 Oracle regions

Why HPC is better on Oracle than other clouds

Bare Metal Compute

- First true bare metal offering
- Better price performance than other cloud providers
- Similar or better performance with on-premises environments

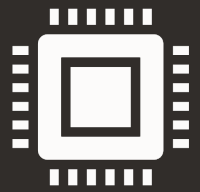
Scalable, Independent Storage

- More NVME local storage than any other provider
- Block storage delivers millions of IOPS at the lowest cost in the industry
- Faster parallel file system performance than any other cloud provider

Specialized Network

- Non-over subscribed, flat 100 Gbps bandwidth
- Only cloud with Network performance SLA
- 100G RDMA networking

Why HPC on the cloud is better than on-premises



Run on the latest generation

- CPUs, GPUs, ARM, NVIDIA, Intel, AMD, Oracle Cloud will be running the latest generation hardware
- Eliminate the cost of lost performance
- Eliminate refresh cycles

Pay only for what you use

- Scale your compute cluster for your current workload
- Pay for the correct hardware, at the correct time, at the correct scale
- Use cloud credits for any type of infrastructure or Oracle Cloud service offering

Get out of the DC business

- No more real estate, co-lo's, electricity bills, etc. Focus on what you do best
- Leverage Oracle's public cloud infrastructure, monitoring, maintenance, and management

Compute choices

Bare Metal

Direct hardware access without hypervisor

Single tenant server

No jitter, no noisy neighbor, no performance loss

Virtual Machine

Multi tenant VMs

Lightweight hypervisor

Full access to hardware through SRIOV

Available also paravirtualized and emulated for legacy compatibility

Dedicated VM Host

Dedicated hypervisor running on bare metal

Single tenant VMs

Best of both worlds when full isolation is required

The bare metal for HPC

True High Performance and control

High Performance Computing Clusters

BM.Optimized3.36
Intel Xeon Gold 6354 (Ice Lake) 3.6Ghz
512GB RAM | 3.84TB NVME
2x50Gb/s vNIC

BM.HPC2.36
Intel Xeon Gold 6154 (Skylake) 3.7GHz
384GB RAM | 6.4TB NVME
1x25Gb/s vNIC

100 Gb/s RDMA (RoCEv2)

Hyperscale Environments

BM.Standard.E4.128
2 x AMD EPYC 7J13 3.5 GHz
2TB RAM
2x50Gb/s vNIC

BM.Standard.E3.128
2 x AMD EPYC 7742 3.4 GHz
2TB RAM
2x50Gb/s vNIC

ARM Compute Instances

BM.Standard.A1.160

3.0 GHz Ampere® Altra™
1TB RAM
2x50Gb/s vNIC

NVIDIA GPU Cloud Platform

BM.GPU4.8

Highest GPU performance in the cloud

8 x NVIDIA A100
2 TB Mem | 27.2 TB NVMe

16 * 100 Gb/s RDMA (RoCEv2)

New era of Flexibility

Flexible sizing of cores and memory

Flexible High Performance

VM.Optimized3.FLEX
Intel Xeon Gold 6354 (Ice Lake) 3.6Ghz

FLEX	MIN	MAX
Memory (GB)	1	256
OCPU	1	18

Flexible RAM and core density

VM.Standard.E4.FLEX
AMD EPYC 7J13 3.5 GHz

VM.Standard.E3.FLEX
AMD EPYC 7742 3.4 GHz

FLEX	MIN	MAX
Memory (GB)	1	1024
OCPU	1	64

Flexible ARM instances

VM.Standard.A1.FLEX

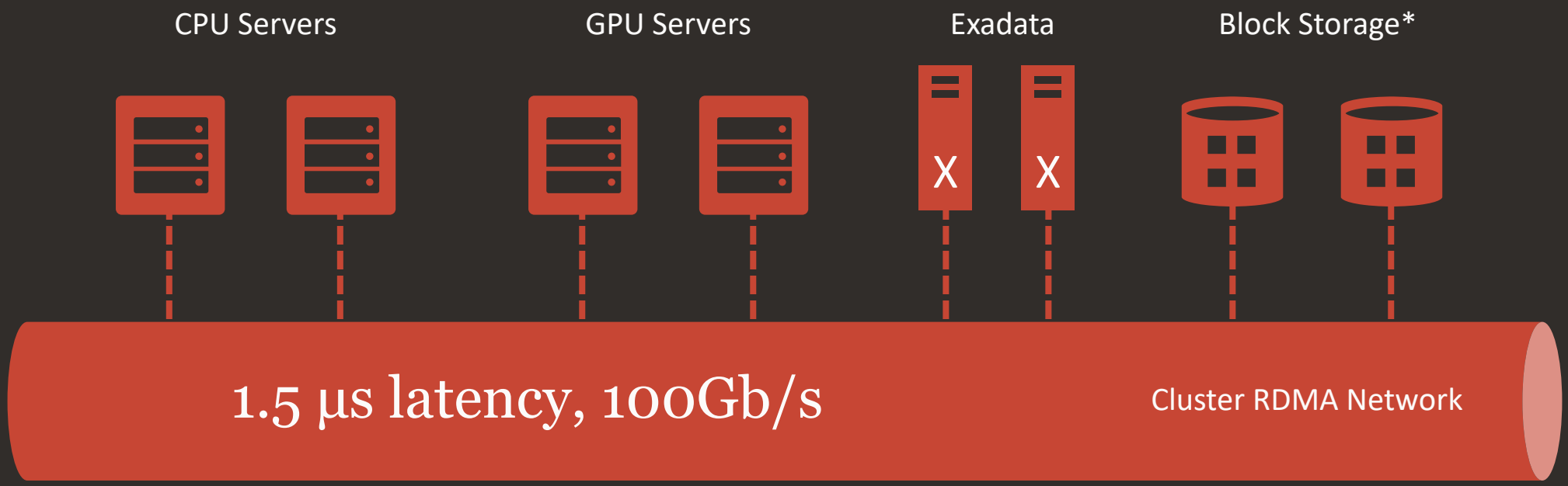
3.0 GHz Ampere® Altra™

FLEX	MIN	MAX
Memory (GB)	1	512
OCPU	1	80

Cluster networking

Low latency
High Bandwidth
Predictable performance
Smart host placement within data center

20,000 core clusters
RDMA over converged Ethernet
True hyperscale HPC cloud



For high performance workloads (HPC, Database, Big Data, AI) including the hardest product development workloads like CFD, Crash Simulations, Reservoir Modelling, DNA Sequencing

Storage Options

Local NVMe

Block Volume Service

File Service

Object

Fastest

Most Durable

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.
- Perfect for distributed file systems. Scalable size and performance.
- Volume can be attached to multiple 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
- API driven access

Shapes with Local NVMe

Compute Shapes with Local NVMe:

- BM.HPC2.36 – one 6.4 TB disk
- BM.Optimized3.36 – one 3.4TB disk
- BM.DenseIO2.52 – Eight 6.4 TB disks. Total: 51.2 TB
- VM.DenseIO2.x - 6.4-25.6 TB for VMs. 1 to 4 disk based on Shape selected.

Block Volumes

- Network attached Block Storage
- Consistent High Performance, persistent & durable storage
- 32 Block volumes per instance
- Each volume can be of 50GB – 32TB size
- Block volumes can scale to 1 PB per compute instance
- Highly reliable - built-in durability and run on redundant hardware
- Attach Block Volume to multiple compute instance in sharable read or read/write mode (like SAN storage, but without SAN like cost)
- Typical workloads include NoSQL databases, Hadoop/HDFS applications, Internet of Things (IoT), and ecommerce applications.

Block Volumes – Performance Tiers

Block Volume Elastic Performance tiers

- Ultra High Performance
- Higher Performance
- Balanced
- Lower Cost

Dynamically change the performance and cost characteristics of block storage and boot volumes instantaneously

Elastic Performance Level	Volume Performance Units (VPUs)	IOPS per GB	Max IOPS per Volume	Size for Max IOPS (GB)	KBPS per GB	Max MBPS per Volume
Lower Cost	0	2	3,000	1,500	240	480
Balanced	10	60	25,000	417	480	480
Higher Performance	20	75	50,000	667	600	680
Ultra High Performance	30	90	75,000	833	720	880
ADJUSTABLE UP TO						
Ultra High Performance	**120**	**225**	**300,000**	**1,333**	**1,800**	**2,680**

File Storage Service (FSS)

- Oracle Managed service
- Pay as you go – storage cost
- NFSv3 based file system
- General workloads which require a shared/distributed file system
- Add more mount targets to get higher performance, each mount target is limited to about 600 MB/s of read or write traffic.

Connectivity

VPN service

High availability (HA) support
Both static and dynamic routing (BGP over IPSec connections)

Both policy-based and route-based configuration

Fast Connect

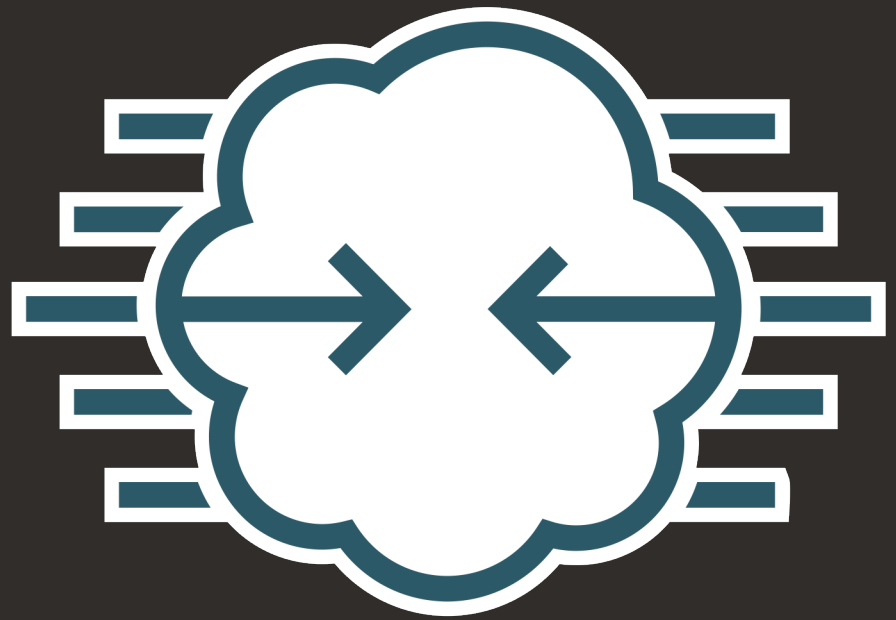
Don’t pay more for accessing your data

Dedicated connection to OCI

1/10 Gb/s (with 1Gbps increment)

Pay only for port

No ingress/egress charges



Product Overview and Features

Compute

- Both VM and Bare Metal (BM) compute instances with the latest NVIDIA GPUs and AMD and Intel CPUs, including high core count and high core-frequency options and up to 52 TB local NVMe storage

Deploy high-performance file systems in a single click

- BeeGFS, Lustre, Gluster, IBM Spectrum Scale, Quobyte
- Achieve throughput of 60 -140GB/sec

Networking

- 2 X 50 Gbps network interfaces
- 100 Gbps for RDMA cluster networking (RoCE v2)

Block Volume

- Block storage volumes up to 1PB

Deployment Strategy

Marketplace

Use Marketplace for deploying partner provided infrastructure and applications

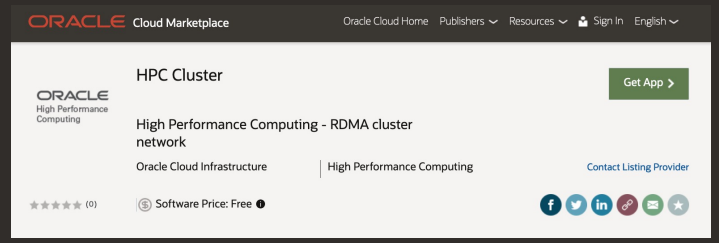
Command Line

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

Stacks

Use Resource Manager stacks for interface based Terraform Deployment

Oracle HPC Solutions



OCI HPC

Automate your HPC cluster deployment

Everything you need to start. Graphical wizard available, API and CLI

Project started early 2019. Used by customers in production.

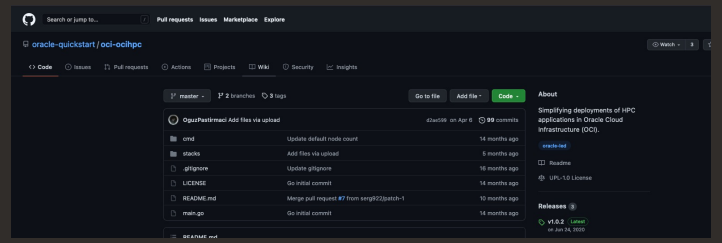
HPC Cluster on Oracle Marketplace

OCI HPC File Systems (HFS)

Peta-byte scale filesystem

Deploy your choice of parallel file system – BeeGFS, Lustre, Gluster, IBM Spectrum Scale

OCI HPC File System (HFS) on Oracle Cloud Marketplace



Easy HPC

Simple command line to deploy HPC clusters of any size on dedicated bare metal HPC compute

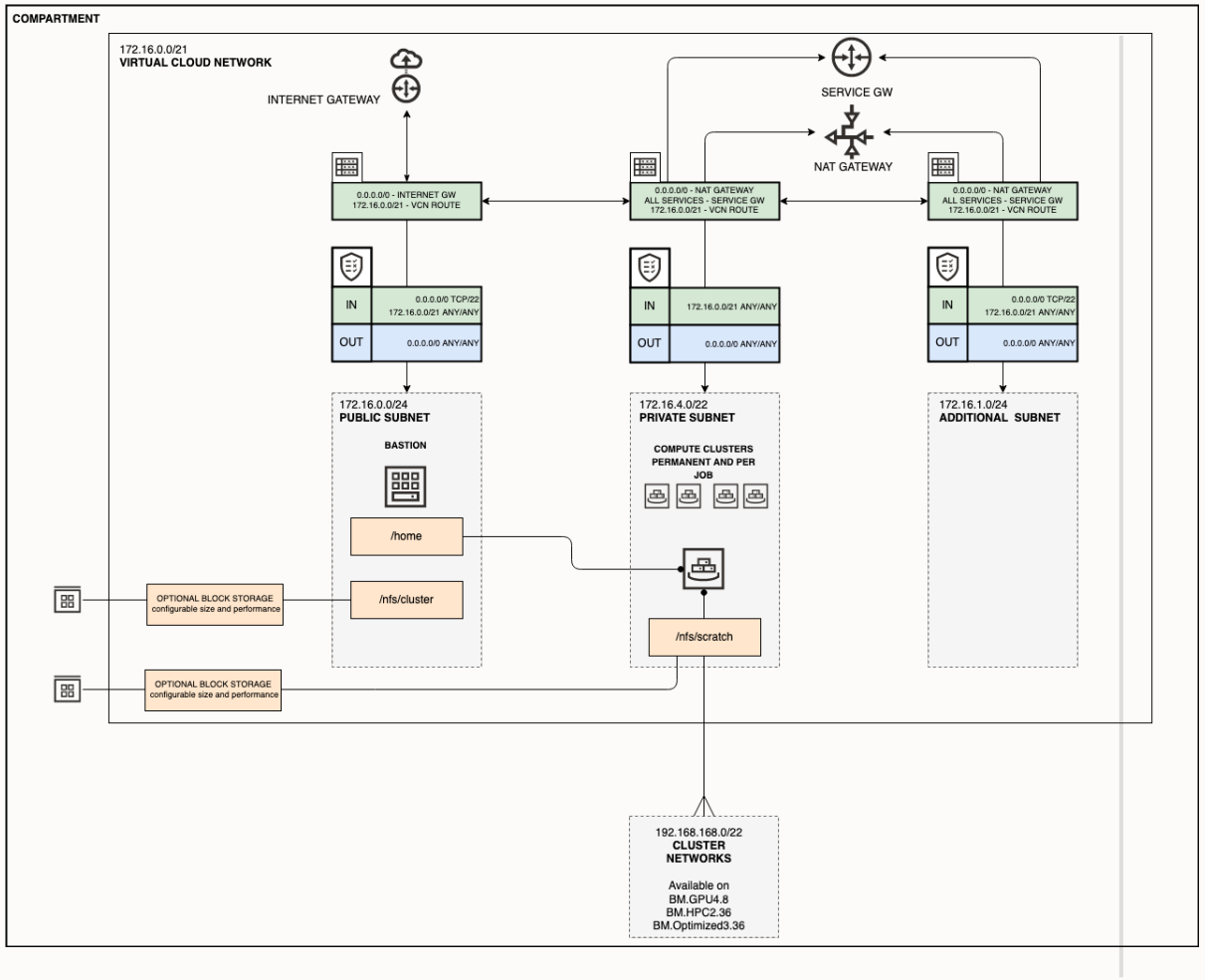
No expertise of Terraform or OCI resource manager required to launch network clusters

Deployment includes a complete set of software packages for running parallel processing with RDMA

Customizable to execute your own terraform scripts

Oracle quickstart on GithHub

Cluster Deployment from Marketplace



- Open Source and actively developed
- Available in Oracle Cloud marketplace and GitHub
- Open technology (terraform and ansible) no proprietary stack

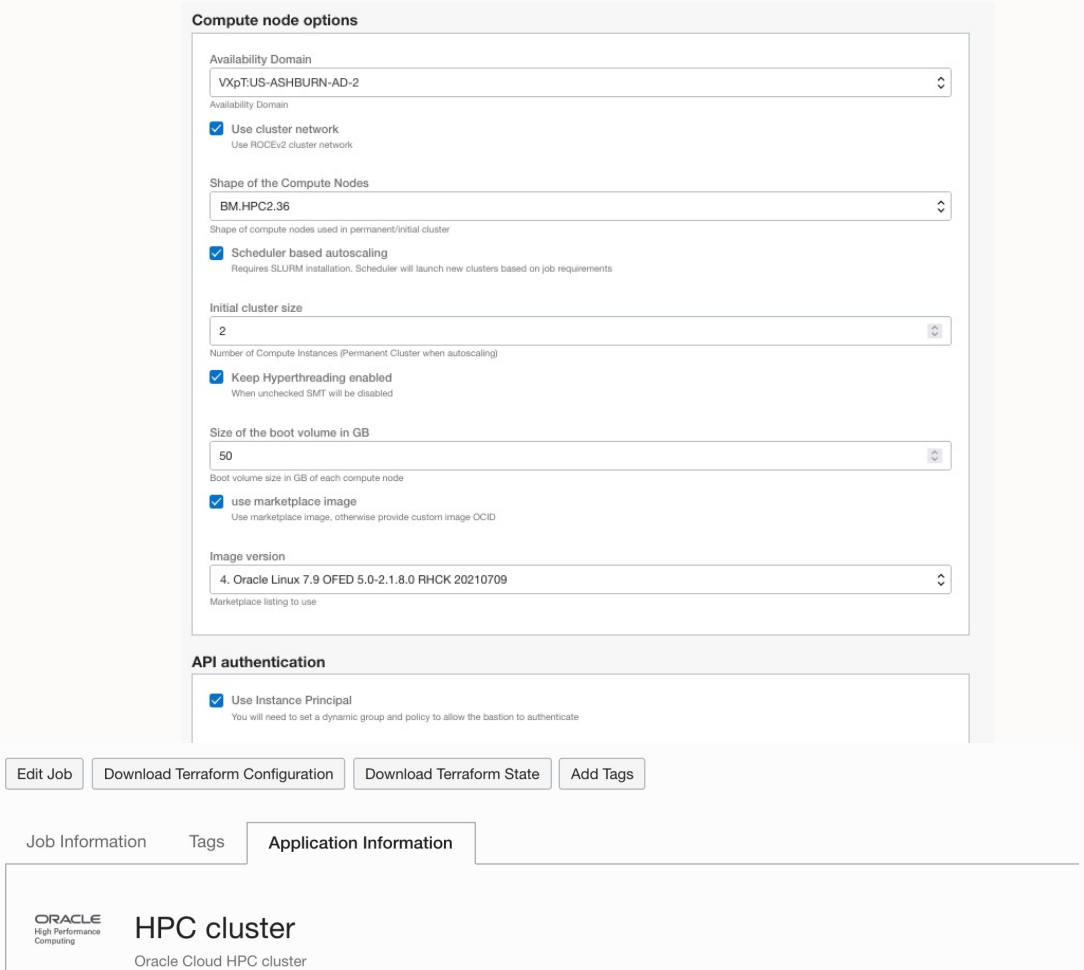
- Graphical wizard, API and CLI deployment
- Configurable storage options
- Autoscaling with SLURM or PBS Pro
- LDAP server
- Spack installation
- Monitoring portal
- Free ksplice for live patching

ORACLE LINUX

CentOS

ubuntu

Cluster Deployment from Marketplace



- Open Source and actively developed
- Available in Oracle Cloud marketplace and GitHub
- Open technology (terraform and ansible) no proprietary stack

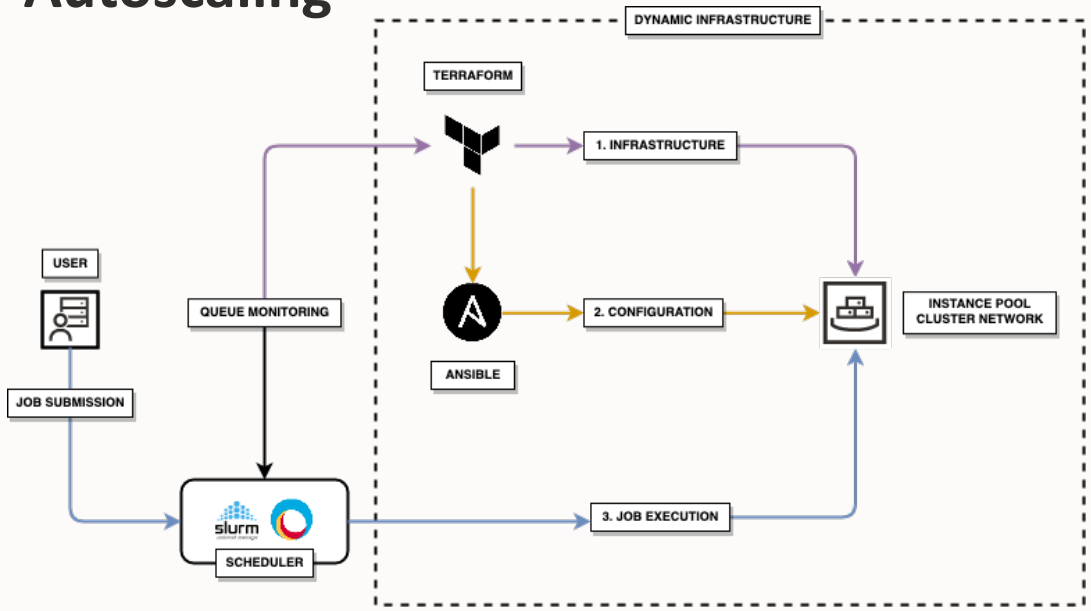
- Graphical wizard, API and CLI deployment
- Configurable storage options
- Autoscaling with SLURM or PBS Pro
- LDAP server
- Spack installation
- Monitoring portal
- Free ksplice for live patching

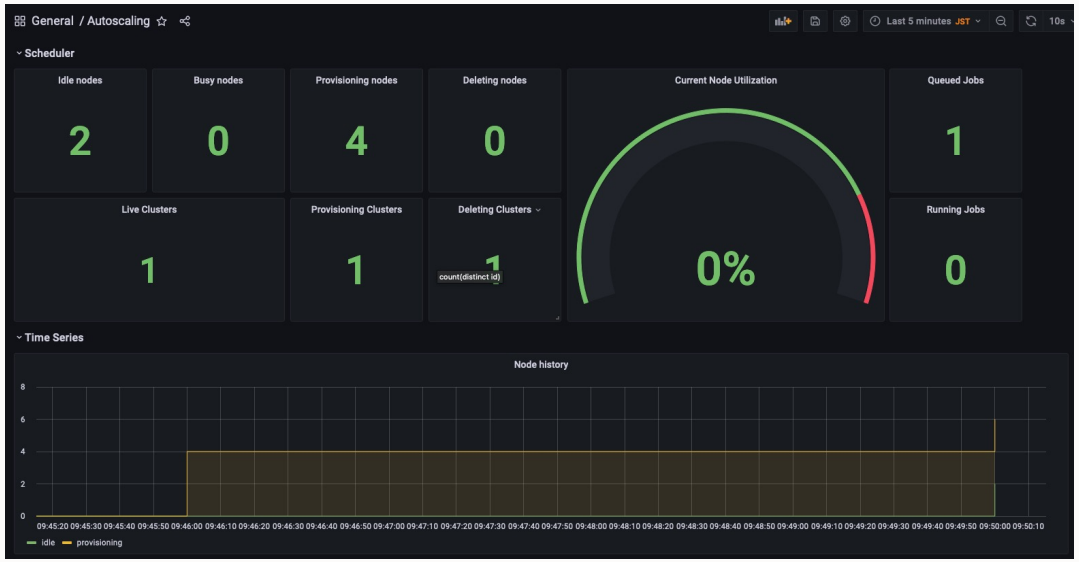
ORACLE LINUX

CentOS

ubuntu

Autoscaling





- Multiple queues
- Multiple shapes per queue
- Configurable limits
- Simple YAML based configuration

- Support for all shapes including RDMA, Ampere, GPU and FLEX configurations
- Infrastructure as a code deployment using open-source tools
- Supports both persistent capacity and bursting

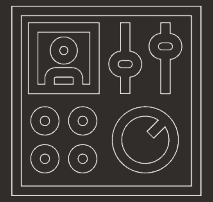
Nissan supports massive simulation needs with Oracle Cloud

- Scalable, high performance and cost-effective solution
- Required powerful HPC cloud vendor to run complex crash and CFD simulations
- Latency sensitive workloads
- Uses VDI on GPUs to postprocess directly on Oracle Cloud

Japan
Automotive | CFD & Crash Simulations

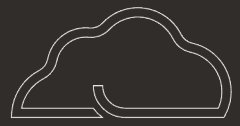
"We selected Oracle Cloud Infrastructure's HPC solutions as a part of our multi-cloud strategy to meet the challenges of increased simulation demand under constant cost savings pressure. I believe Oracle will bring significant ROI to Nissan."

DEMO



CLUSTER CONFIGURATION

JOB SUBMISSION



AUTOSCALING

MONITORING

Thank You