Building Extreme-Scale File Services in the Oracle Public Cloud

Ed Beauvais, Director Product Management

Ed.Beauvais@Oracle.com

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.

File Storage

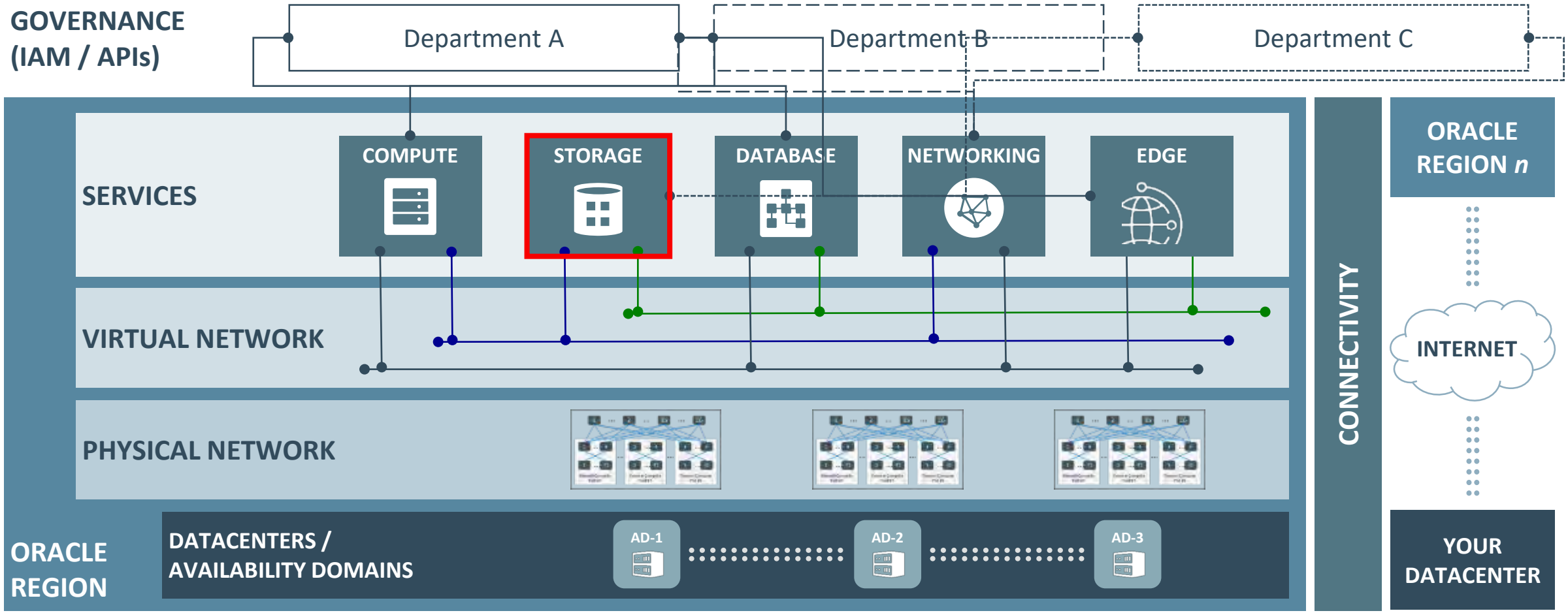
Agenda

- Oracle Cloud Infrastructure Overview
- What is the File Storage Service (FSS)?
- Architecture
- Use Cases
- Q&A

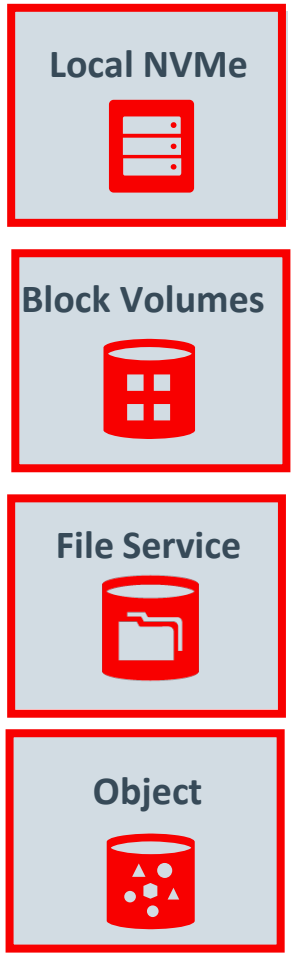
OCI Overview

Oracle Cloud Infrastructure Overview

High performance compute, storage, database, edge on the same flexible virtual network



Overview of Storage Options



Lowest Latency

Highest Durability

- **High performance** NVMe SSD storage
- Local to a 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
- Volumes and file shares can be detached and attached to different instances

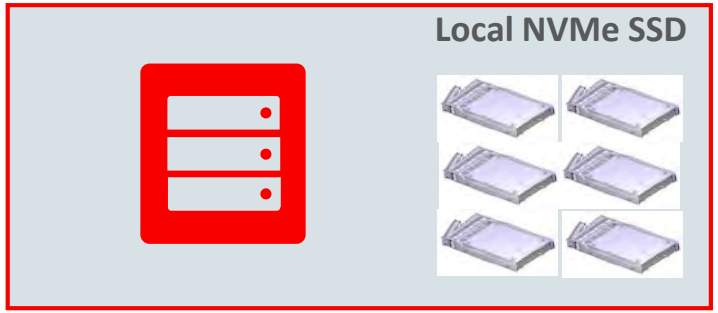
- **Regional network accessible, durable storage**
- Data is replicated regionally for high availability and durability
- Designed for big data, backup and unstructured content

Local NVMe Storage

- High performance local storage available with bare metal compute instances

Boot Volume

- 50GB Boot Volume exposed via iSCSI



NVMe storage with bare metal compute

- Local storage with Dense IO Compute
  - CPU: 36-Cores; RAM: 512 GB; Local SSD**: NVMe (28.8 TB total)**
  - CPU: 52-Cores; RAM: 768 GB; Local SSD**: NVMe (51.2 TB total)**
- NVMe performance
  - Millions of IOPS
  - 10-100 Microsecond latencies

Block Volume Service Characteristics

Block Volumes

Metric or Feature	Block Volume Service Characteristics
Flexibility	Configurable: 50GB to 16TB (1GB increments) *All NVMe-based*
Perf: IOPS/Volume	60 IOPS/GB - up to 25K IOPS*
Perf: Throughput/Volume	480 KBPS/GB - up to 320 MBPS**
Perf: Latency/Volume (P95)	<1 msec
Perf: Per-instance Limits	• 32 attachments/instance, up to ½ PB • Up to 620K IOPS, near line rate throughput
Volume Durability	Multiple replicas across AD
Restore from Backups (RTO)	<1 minute, regardless of size
Backup Performance (RPO)	~15 minutes per TB, via point-in-time snapshot
Cost per GB/month	Still 4.25 cents! Still simple model, 1 option!

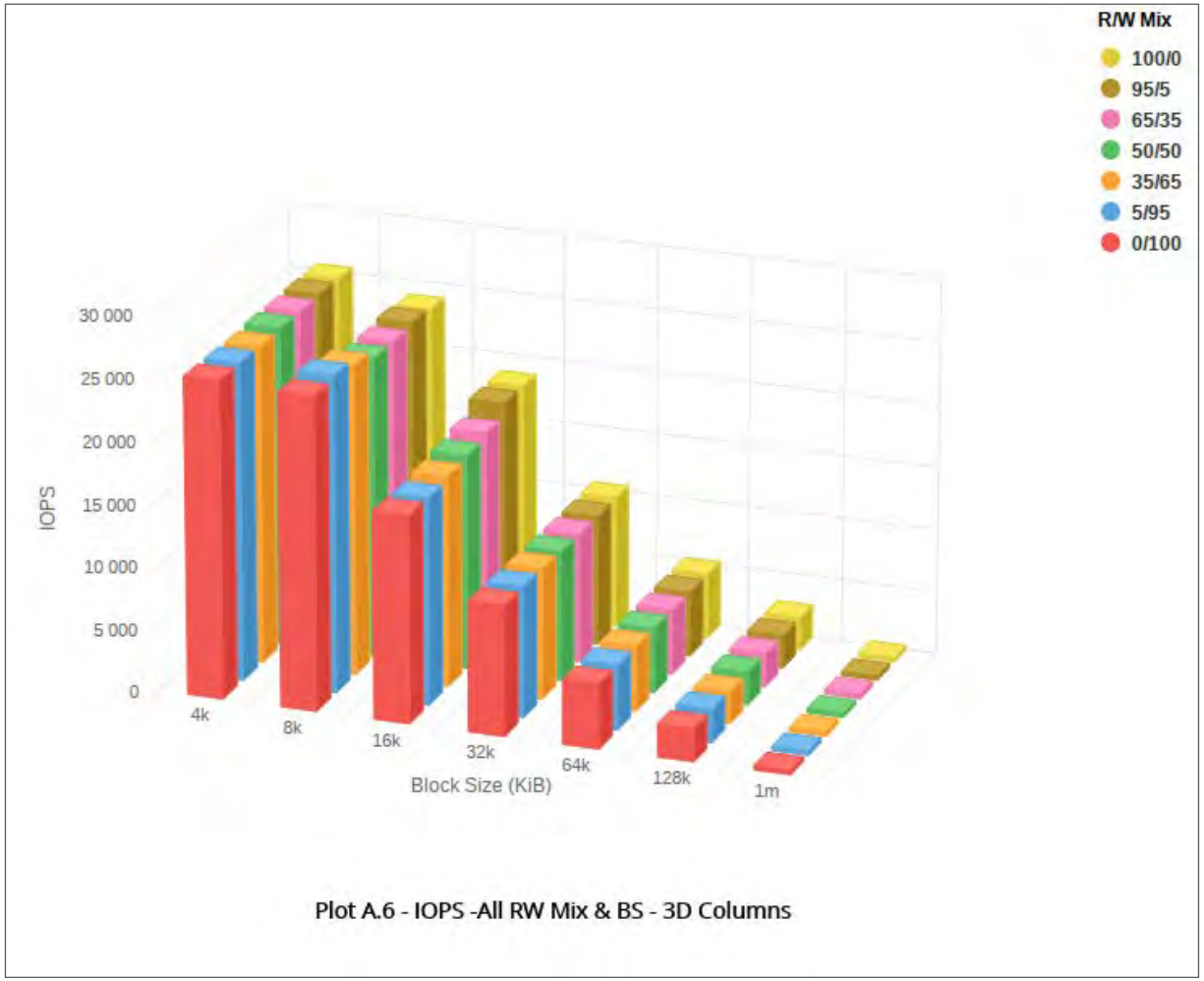
** For Bare Metal or 8-core+ VM compute instance, using 4KB blocks. VM perf is limited by VM network bandwidth.*
*** At 256 KB block size*

ORACLE

Block Volumes

Block Volume Performance Analysis

- 25,000 IOPS @ 4k, < 1ms, 1TB volume
- Consistent performance measured across all block sizes and R/W mixes
- Simple test reproduction steps
  - Provision & attach >417GB volume
  - Measure with Gartner Cloud Harmony test suite:
    - https://github.com/cloudharmony/block-storage
- Blog and analysis
  - https://community.oracle.com/community/cloud_computing/bare-metal/blog/2017/05/16/block-volume-performance-analysis

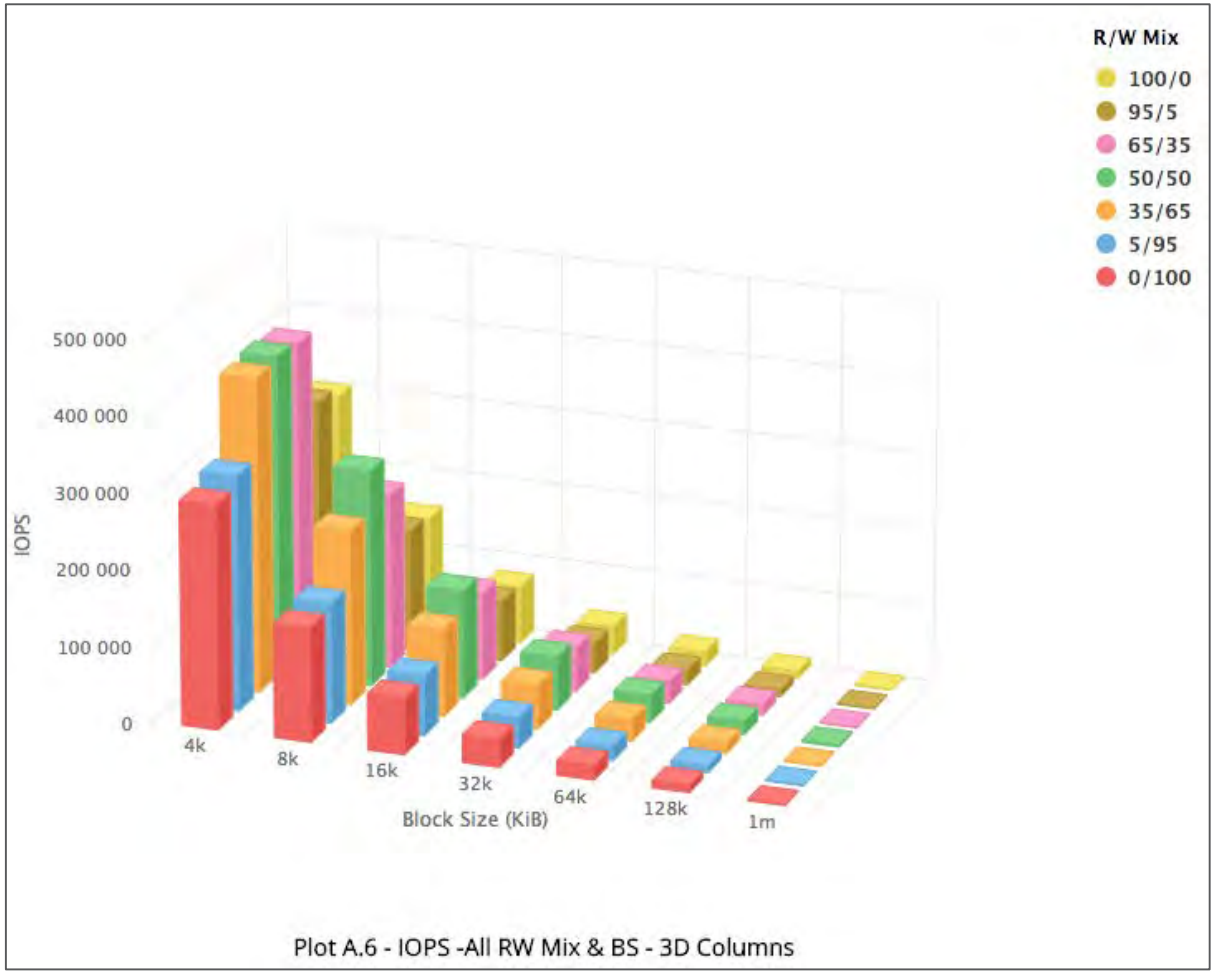


Plot A.6 - IOPS -All RW Mix & BS - 3D Columns

Block Volumes

Host Block Volume Performance Analysis

- >400,000 IOPS @ 4k, < 1ms, 20 x 1TB vol
- Consistent performance measured across all block sizes and R/W mixes
- Simple test reproduction steps
  - Provision & attach >417GB volumes
  - Measure with Gartner Cloud Harmony test suite:
    - https://github.com/cloudharmony/block-storage
- Blog and analysis
  - https://community.oracle.com/community/cloud_computing/bare-metal/blog/2017/05/16/block-volume-performance-analysis



Plot A.6 - IOPS -All RW Mix & BS - 3D Columns

Object Storage Features

Object Storage

Feature	Description
Multipart Upload	Upload object in part, each part can be 10 MiB - 50 GiB in size Pause and resume upload
Pre-Authenticated Requests (PAR)	Server side Temporary URLs, used to share data Defined on Objects and/or Buckets Support for listing and deleting previously generated PARs
Large Object Support	Support for large 10 TiB objects
Audit Service support	Audit support for bucket operations
Public Buckets	Anonymous public access to data stored on object storage Read and/or List privileges supported
Custom Object/Bucket metadata	Define custom metadata (~2kb) per Object or Bucket
Tagging	Tag bucket resources for chargeback or resource management
Compartment Management	Move buckets between compartments Designate default compartments for use with Amazon S3 API
Bucket ACLs	Define IAM policy at the granularity of a bucket in a given compartment

ORACLE

File Storage Service Overview

File Storage Service (FSS)

FSS Benefits	Description
Elastic Growth	• No minimum capacity, or upfront provisioning required. • Start with Kilobytes, scales up to 8 Exabytes. • Pay only for what you use (your capacity stored).
Enterprise-Grade File Storage	• A dynamic, enterprise-grade file storage service that scales up to meet the storage needs of enterprise customers.
Fully Managed	• Lower Operational Expense (OPEX) Oracle manages: capacity growth planning, software upgrades, failed components, etc.
Ease of Deployment	• With a few clicks from OCI Console, create a file system and a mount target in your network that can be accessed by thousands of compute resources within a region.
Access	• Simplify cloud file share management with a range of tools provided (OCI Console, APIs, CLIs, Terraform, and data-path commands).
Data Protection	• Highly available. • Multi-way replication for your data and metadata.

ORACLE

File Storage Service

File Storage

FSS Features

- NFSv3
- Network Lock Management
- Data Protection: Space efficient snapshots
- Enterprise Defaults:
  - 100 file systems
  - 2 mount targets per AD
  - Up to 10,000 Snapshots per file system
- Security: Data-at-rest encryption for all file systems
- AD-local service, in all OCI regions & ADs!
- Cost: Simple $0.0425 GB/Month,
- Performance: 150 MB/s per TB*

Use Cases

General Purpose
Archive

Big Data
Analytics

HPC
Scale Out Apps

Oracle Applications
Lift and Shift

Test / Dev
Databases

Micro Services
Containers

File Storage Service – Access Options

File Storage

AD-local service, accessible from all ADs in the same region, and by thousands of OCI resources concurrently over OCI Console, APIs, CLI, Terraform, and data-path commands.

Scenarios	Recommendations
Local AD Access	Mount from local VM or BM compute instances*
Remote access from another AD* (within a region)	Configure your network connectivity VCN to allow that traffic, and open all required NFS ports. Mount from remote VM or Bare Metal compute instances*
Remote access from another region or from customer data center	Use VCN Peering, FastConnect or VPN.
Remote access over the internet	Install our secure S3 gateway (Instructions & Terraform template) to enable ingesting files/file systems or sharing externally via S3/URL

***For best performance we recommend mounting locally, due to latency across AD's**

ORACLE

File Storage Service - Data Protection Options

File Storage

- Snapshots provide a consistent point-in-time view of your entire file system. 10,000 snapshots/file system.
- Copy-on-write; space efficient

Scenarios	Recommendations
AD data protection	Today it is a manual customer driven process. Create a cron job to copy your file system or snapshots asynchronously to another file system in a different AD, using rsync*
Regional data protection	Copy your file system or snapshot data asynchronously to another region, using rsync* Copy to local or remote Object Storage, using tar; or zip your file system or snapshot data*
3rd Party data protection	Use 3rd party software to protect application and file system data in another AD or region. Support for NFSv3 is required

*** Parallelizing your jobs will speed up data transfer**

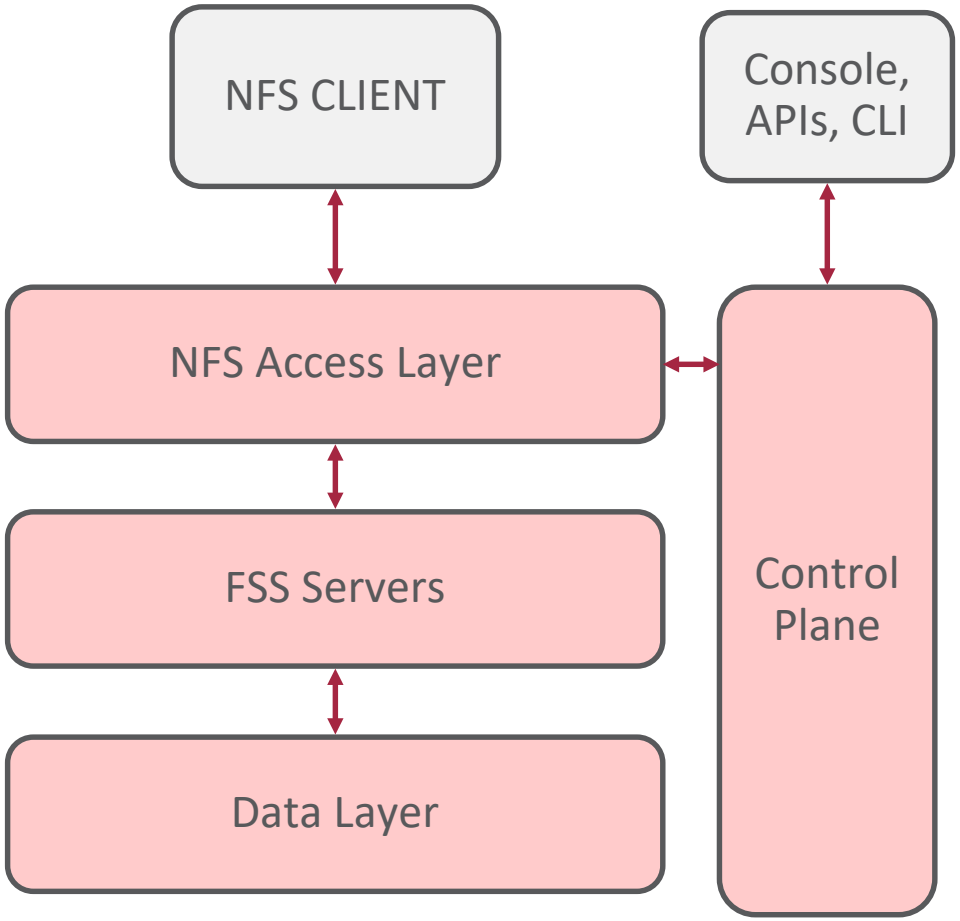
ORACLE

FSS Architecture

Scalability, Security, Performance

File Storage Service Architecture

File Storage



NFS Client

A customer installs an NFSv3 client to connect to the File Storage Service (FSS) data path.

Access Layer

These servers present customer file systems, respond to client NFS IO requests, and provide high availability.

FSS Servers

These servers communicate via a proprietary protocol to manage metadata, provide snapshots and manage data storage.

Data Layer

This layer is where customers file data is protected and stored. Data and metadata protection is delivered using erasure encoding.

Control Plane

These servers provide the ability to manage the service from the console, CLI and APIs requests.

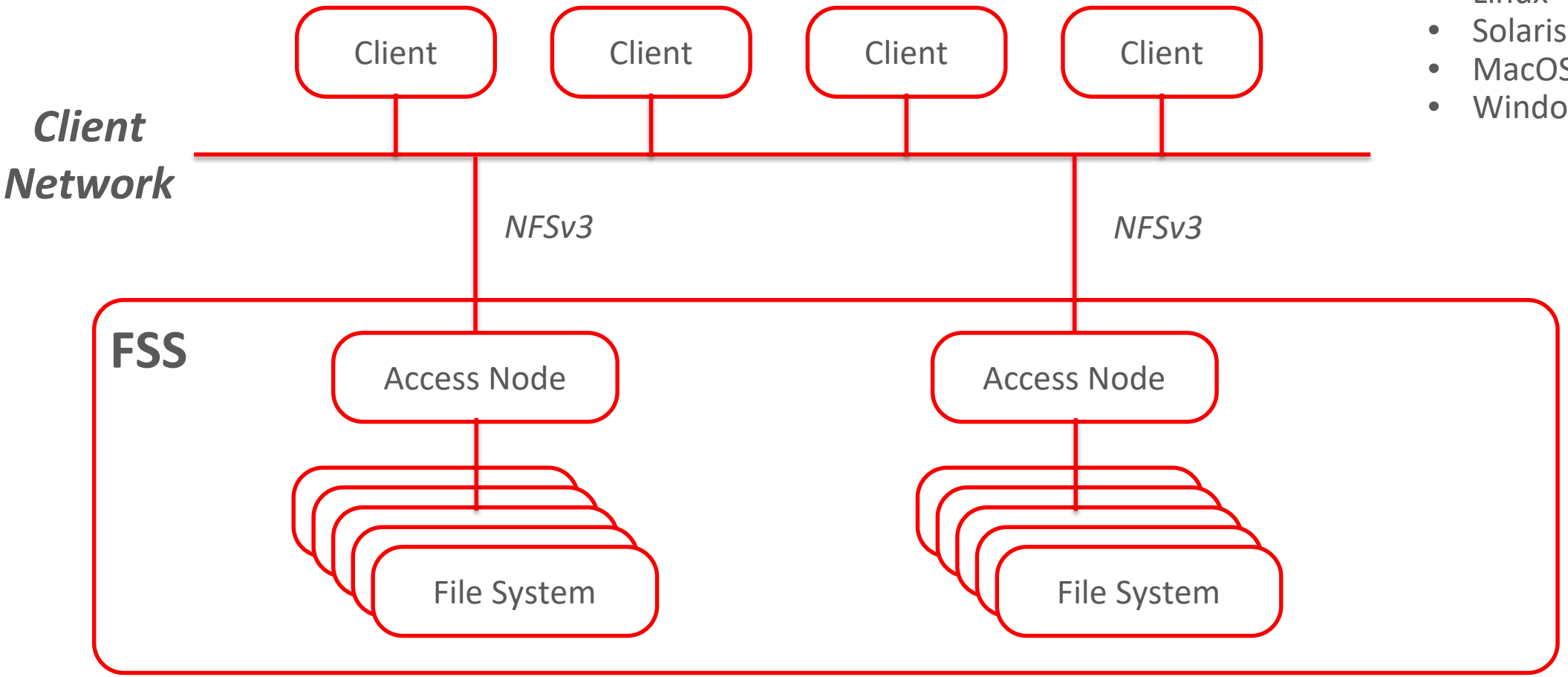
ORACLE

FSS Architecture: Bird's Eye View

File Storage

Supported NFS Clients:

- Linux
- Solaris
- MacOS
- Windows*



ORACLE

File Storage

FSS Architecture: Scalability

Capacity Scaling

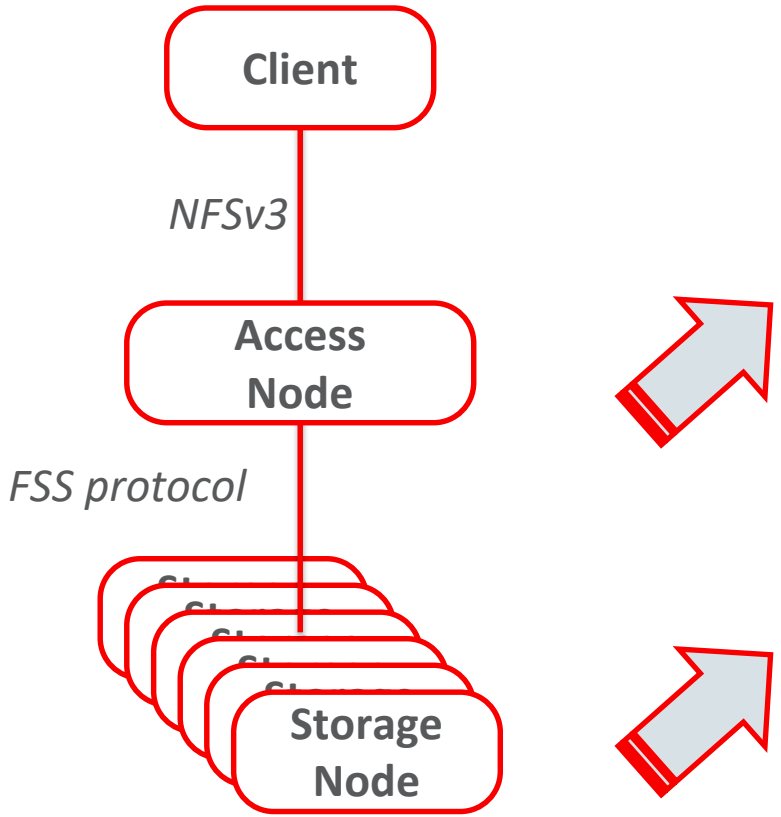
Scale by adding Storage nodes

Performance Scaling

Scale by adding Access or Storage nodes

Enterprise Defaults

- 100 file systems
- 8 EB capacity per file system
- 2 Mount Targets
- 10,000 Snapshots



ORACLE

FSS Architecture: Security

File Storage

Encryption

User data and metadata encrypted at entry (by Presentation Layer servers)

Unique file-system-level keys, unique file-level keys

Network Isolation

Client networks and FSS internal network are isolated

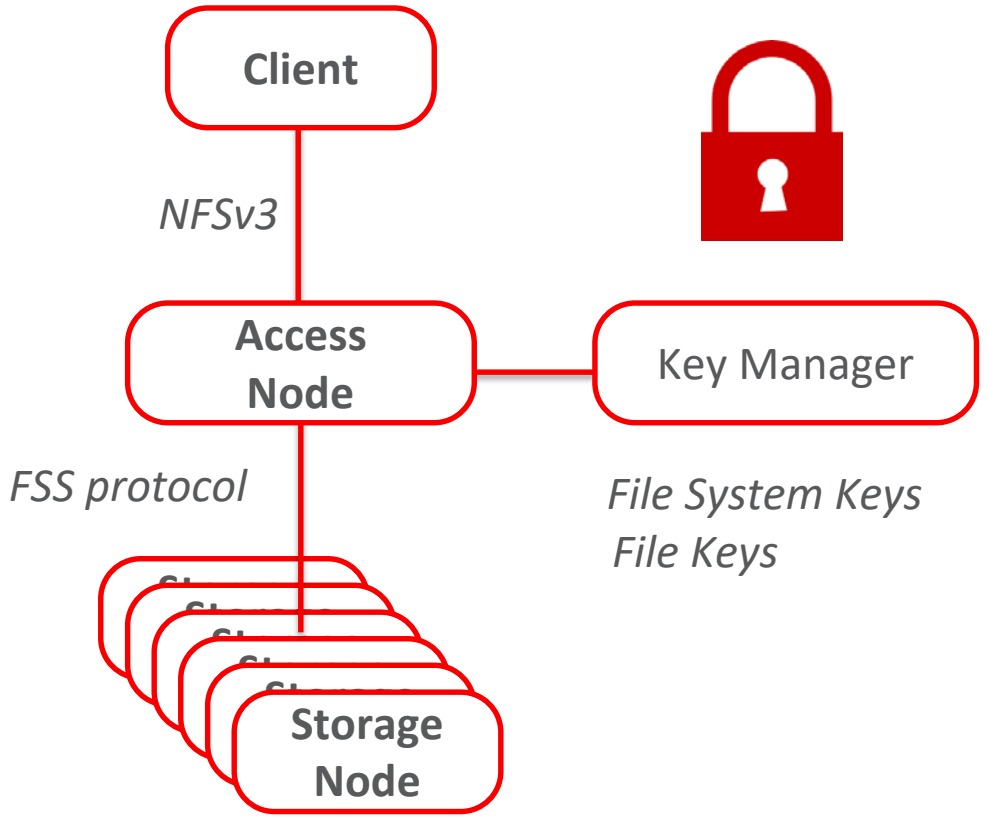
Network isolation is provided by Virtual Cloud Network (VCN)

Improved Access Controls (Export Options)*

Limit file system access by host IP or CIDR

Define Read-only file systems

Unix user permissions via Root squash



Workloads

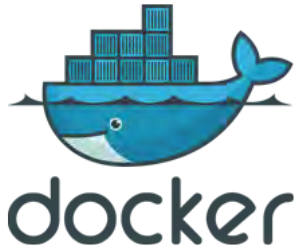
Why Containers?

- Ease of development
- Maximize resource utilization
- Faster more agile infrastructure
- Simplicity and lower environment constraints
- Early wins -> Stateless applications

Why Persistence?

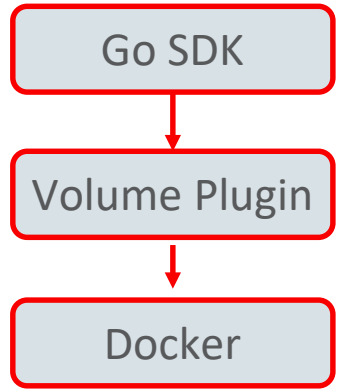
- Want to run in the enterprise?
- Existing apps are not built for container native environments
- Exponential growth of elements
- How do I move data? What about other hosts?
- Many storage environments are not easily integrated with new tools

File Storage: Docker Volume Plugin



File Storage

- Not Released yet!
- Commitment to Docker, Containers, and Open Source
- Low Maintenance!
- OCI driver for File Storage Service
- With 10,000 snapshots per filesystem new models are possible



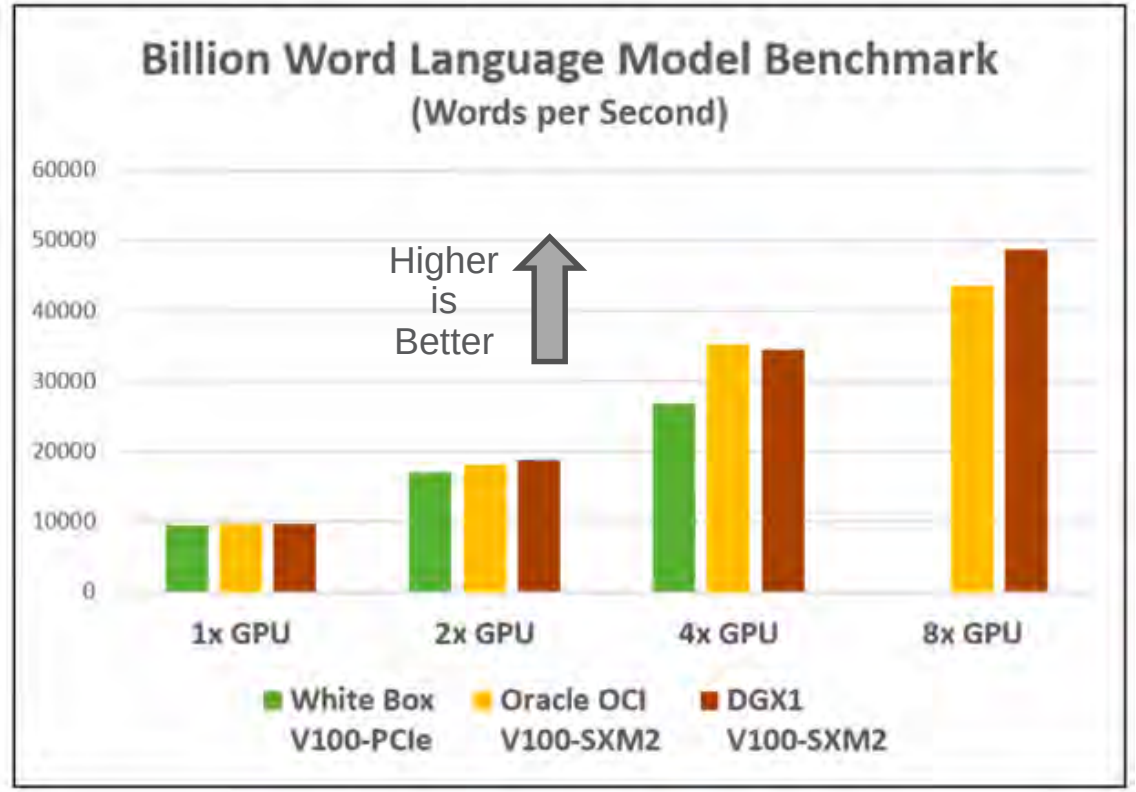

```
sudo docker volume create --driver oci demovol

sudo docker run --rm -it -p 80:8080 -v demovol:/app/appdir --volume-driver=oci application/appname
```

ORACLE

FSS & GPUs on OCI

- Used for ML, AI, Deep Learning, and engineering visualization
- 8x V100's (Volta's) recently released
- 32x V100's coming soon
- Providing on-premises perf and exceeding DGI performance
- Developing deep relationship with NVIDIA



Demonstrations
Inferenceing link

FSS Performance Comparison, HPC Workloads

- Computational Fluid Dynamics
- Fluent 71M Model – Standard Engineering Benchmark for CFD. Solver running on FSS performed slightly faster than on standalone file servers

Oracle Cloud Infrastructure

- New Public Cloud
- Enterprise focus
- New file architecture

Questions?


ORACLE®

FSS Architecture: Servers and Roles

Presentation Layer

Presentation Layer - presents NFSv3 as defined in RFC 1813

User data and metadata encryption/decryption

Storage

Data Store

Metadata Store - B-tree

Control Plane (multiple node types)

Control API (REST), Resource Management, Metering, Garbage Collection, etc

File Storage Service - Components

File System

All data is stored and organized in a file system. File systems consist of both data and metadata. File systems are organized with directories or folders.

Mount Target

Mount target is an NFS endpoint that lives in your subnet of choice and it is highly available. It takes a user specified DNS name & IP address that will be used in the mount command when connecting your NFS client to File Storage Service.

Protocol

File Storage Service (FSS) supports Network File System (NFSv3), this protocol allows clients to access file systems and data on remote systems as if it is a local resource.

FSS Architecture: Snapshots

File Storage

Read-only

Point-in-time view of entire file system

Instantaneous

Snapshot creation is a matter of a small fixed number of meta-data transactions

Incremental

Snapshots keep track of changed blocks, so no need to copy/store full image

ORACLE

File Storage

FSS Architecture: Pages, Blocks, Extents

Page

Page - I/O Unit

8KiB or 32KiB

Block

Allocation Unit, consists of one or more Pages

8KiB, 32KiB, 256KiB, 2MiB

Extent

Replication Unit, consists of multiple Blocks of the same size

Extents vary by intended use (data vs metadata)

Data extents vary by block size

ORACLE

File Storage

FSS Architecture: Metadata

B-Tree

- Transactional distributed key-value store
- Supports multi-key transactions

Keys

- One or more 64-bit values
- Ordered lexicographically

Values

- Small number of bytes
- Each metadata page can accommodate several key-value pairs

ORACLE

Appendix


ORACLE®