Distributed Block-level Storage Management for OpenStack

OpenStack APAC Conference 2012

Daniel Lee

CCMA/ITRI
Cloud Computing Center for Mobile Applications
Industrial Technology Research Institute
(雲端運算行動運用科技中心)

Outline

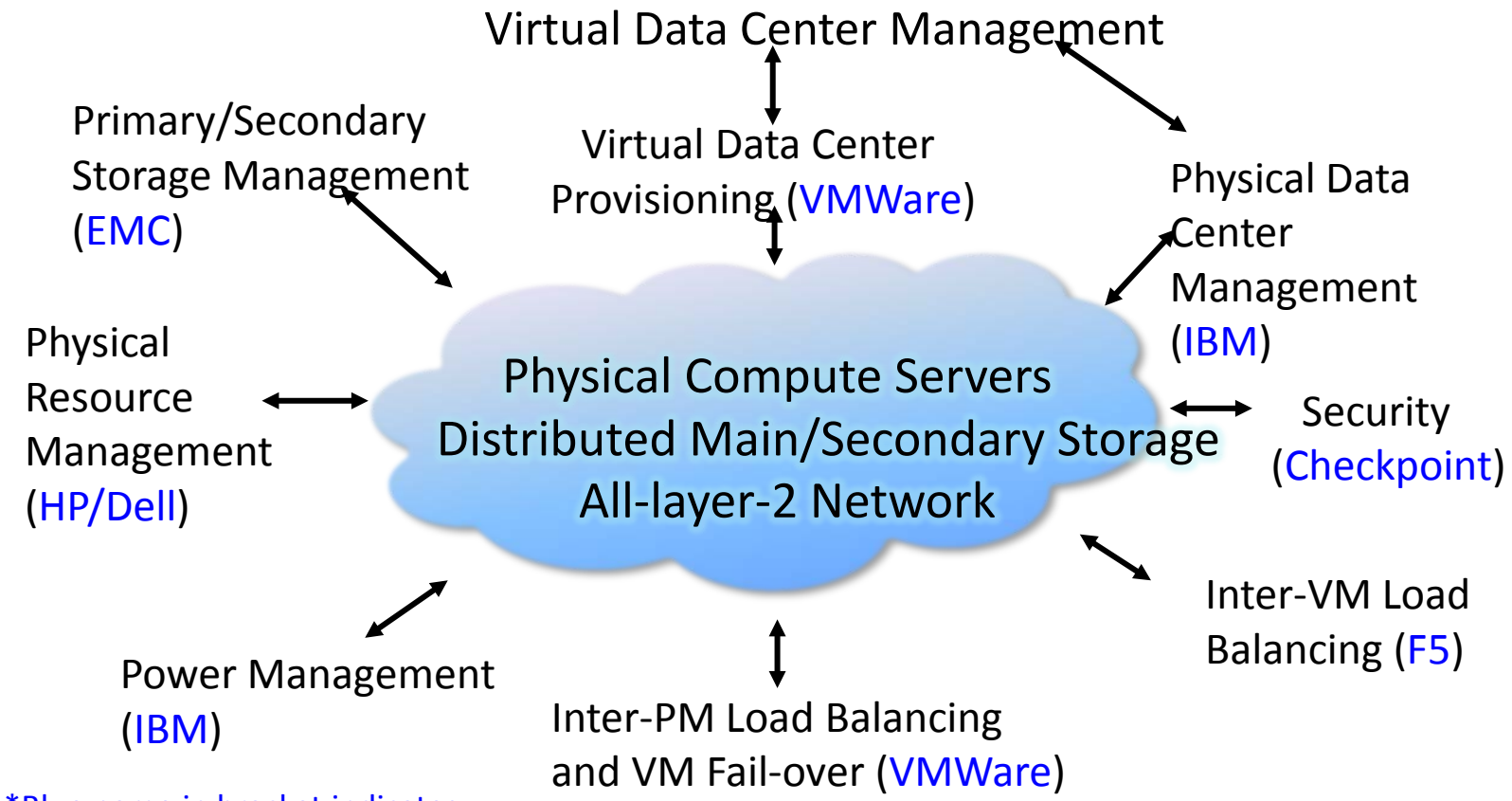
- Overview
- Brief on ITRI Cloud OS
- Cloud storage system in ITRI Cloud OS
- Distributed Main Storage System
- Distributed Secondary Storage System
- Integration of Cloud OS Storage System with OpenStack
- Summary
- Q&A

Overview

- Data is growing exponentially and moving on to Could. It needs to be available at all time, secured and protected.
- 911 in Twin Towers, 311 earthquake in Japan and 420 outage of Amazon datacenters are disasters.
  - Block-level and fully redundant
  - Backup / restore + wide-area
  - Thin provisioning
  - De-duplication
  - Others
- Cloud OS Storage System has these functionality and is ready to integrate to OpenStack

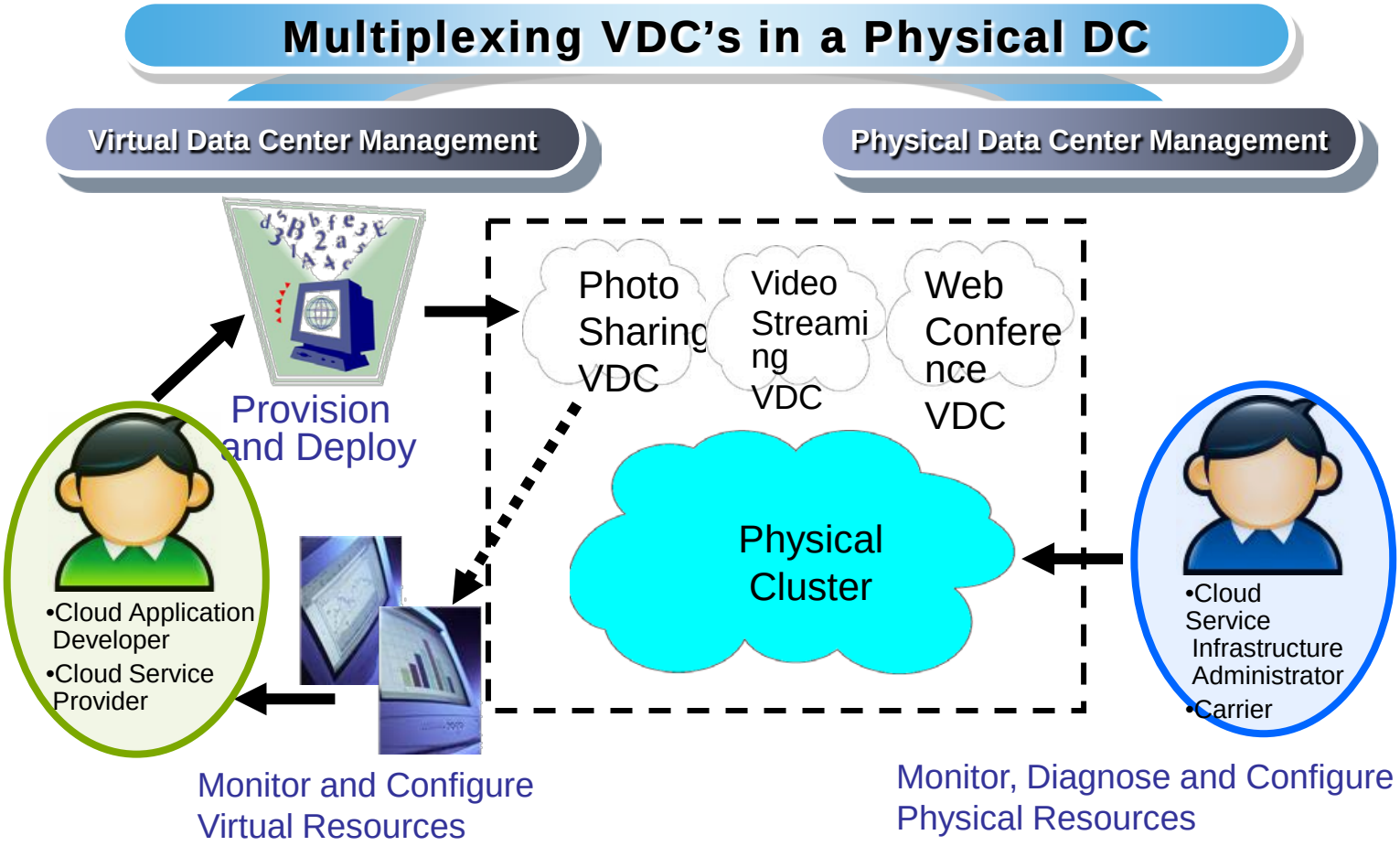
ITRI Cloud OS

(An All-in-one IaaS Total Solution)

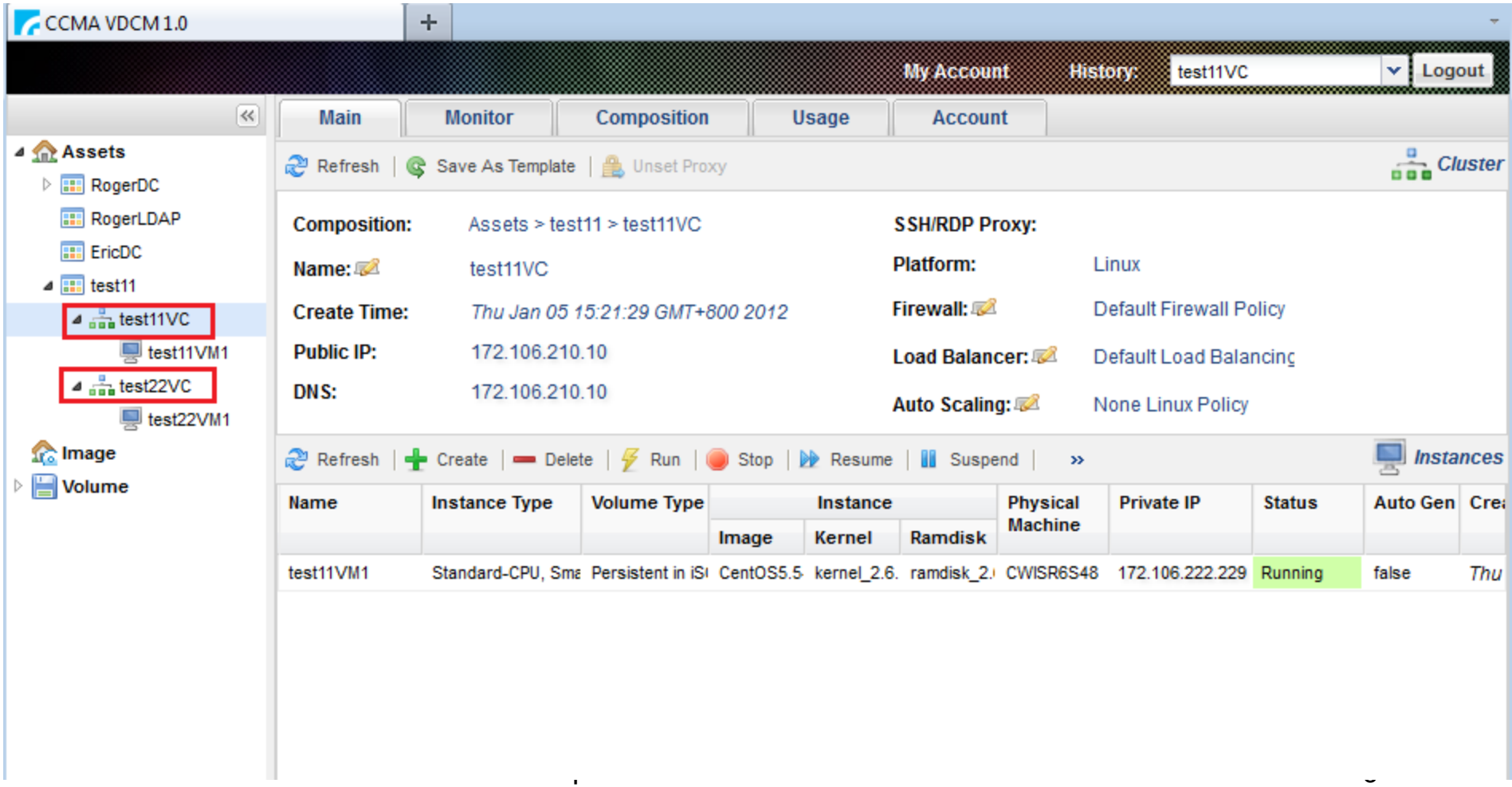


*Blue name in bracket indicates the leading company of the component

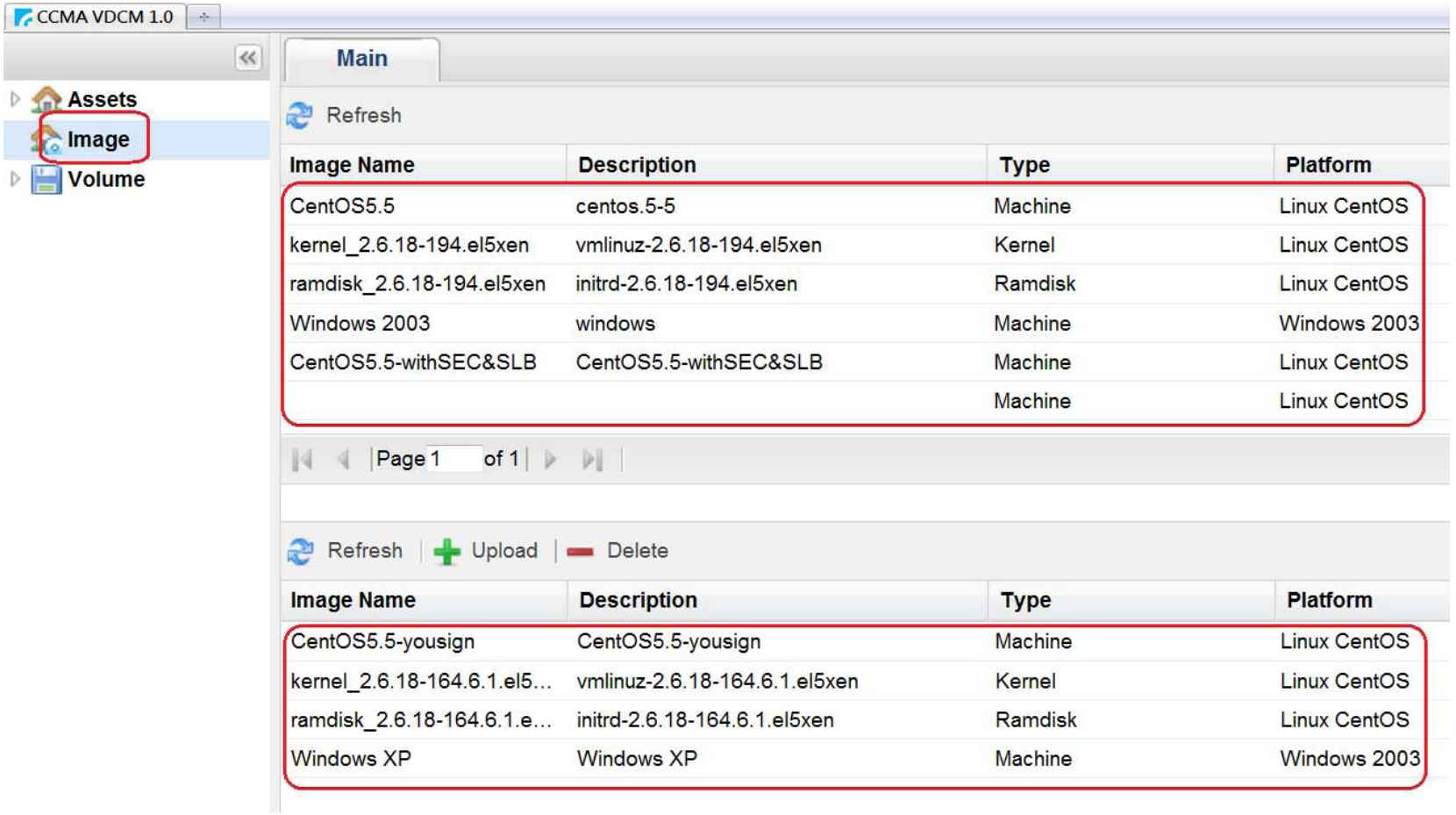
ITRI Cloud OS Service Interface



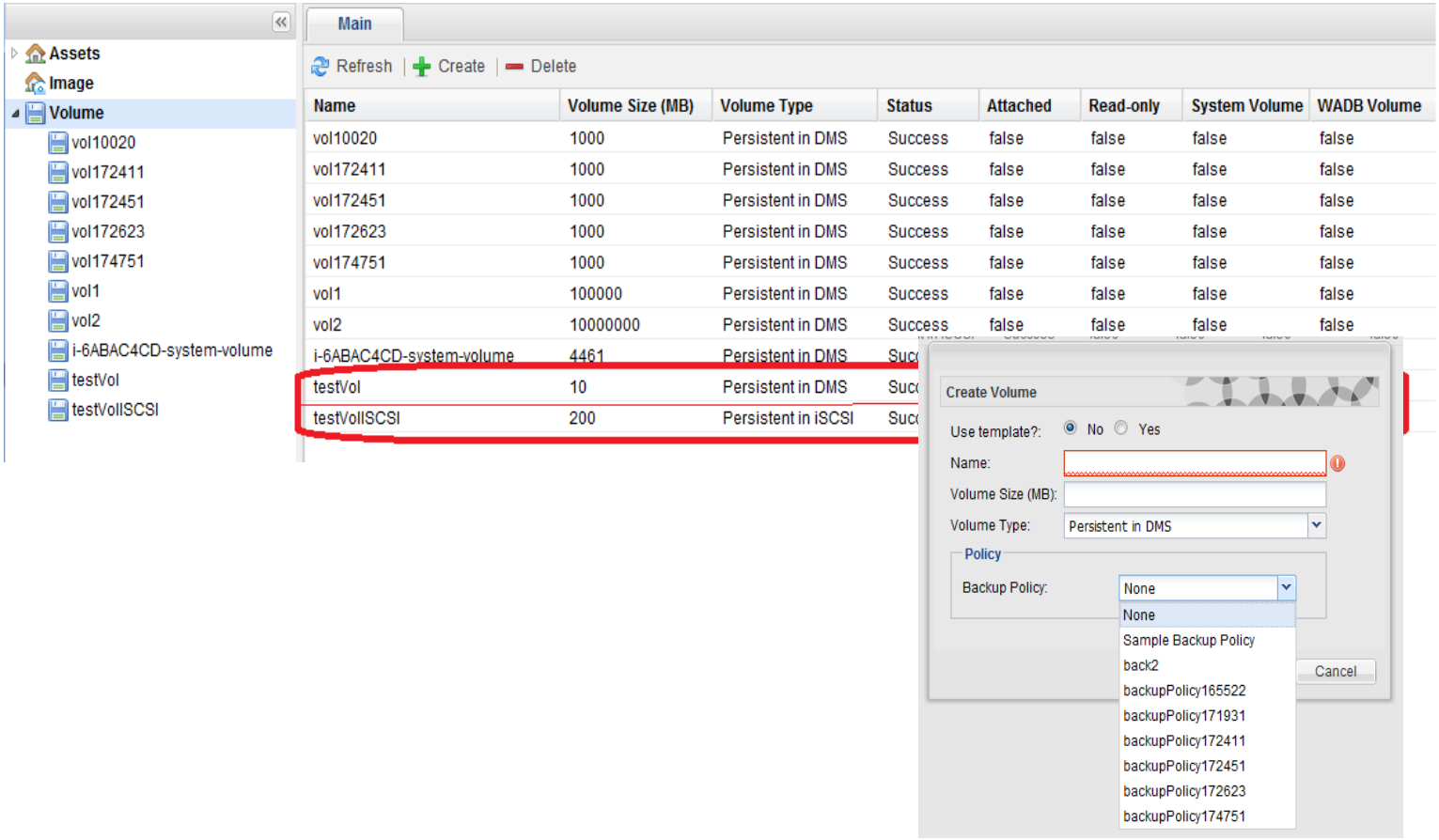
VDCM – Assets (VDC, VC, VM)



VDCM – System Images



VDCM - Volumes



Cloud Storage System in Cloud OS

- Cloud storage aims at **cloud-scale** data centers, and is designed to be **scalable, available and low-cost**.
- Main components
  - Distributed Main Storage subsystem – DMS
    - Provide current image I/O request processing
  - Distributed Secondary Storage subsystem – DSS
    - Block-level incremental metadata only backup system
    - Take volume snapshot / restore, de-duplication engine, garbage collection and WADB engine

Key Features of Cloud Storage System from Users' Perspective

- Use like **raw, unformatted** block devices
- Support volume size from **1 GB to 2 TB** (64TB architecture-wise)
- Can be **attached by different instances** (one at a time for now)
- **Multiple volumes** can be mounted to the same instance
- Data is **automatically replicated** on different physical storage server (N=3)
- Ability to **clone volumes** and/or create point-in-time **snapshots** of volumes
- **Wide-area backup** for disaster recovery
- Snapshot can be scheduled by assigning a **backup policy** with **time, frequency, retention**, and **WADB** option
- Save **system image** for starting multiple VMs using the same image

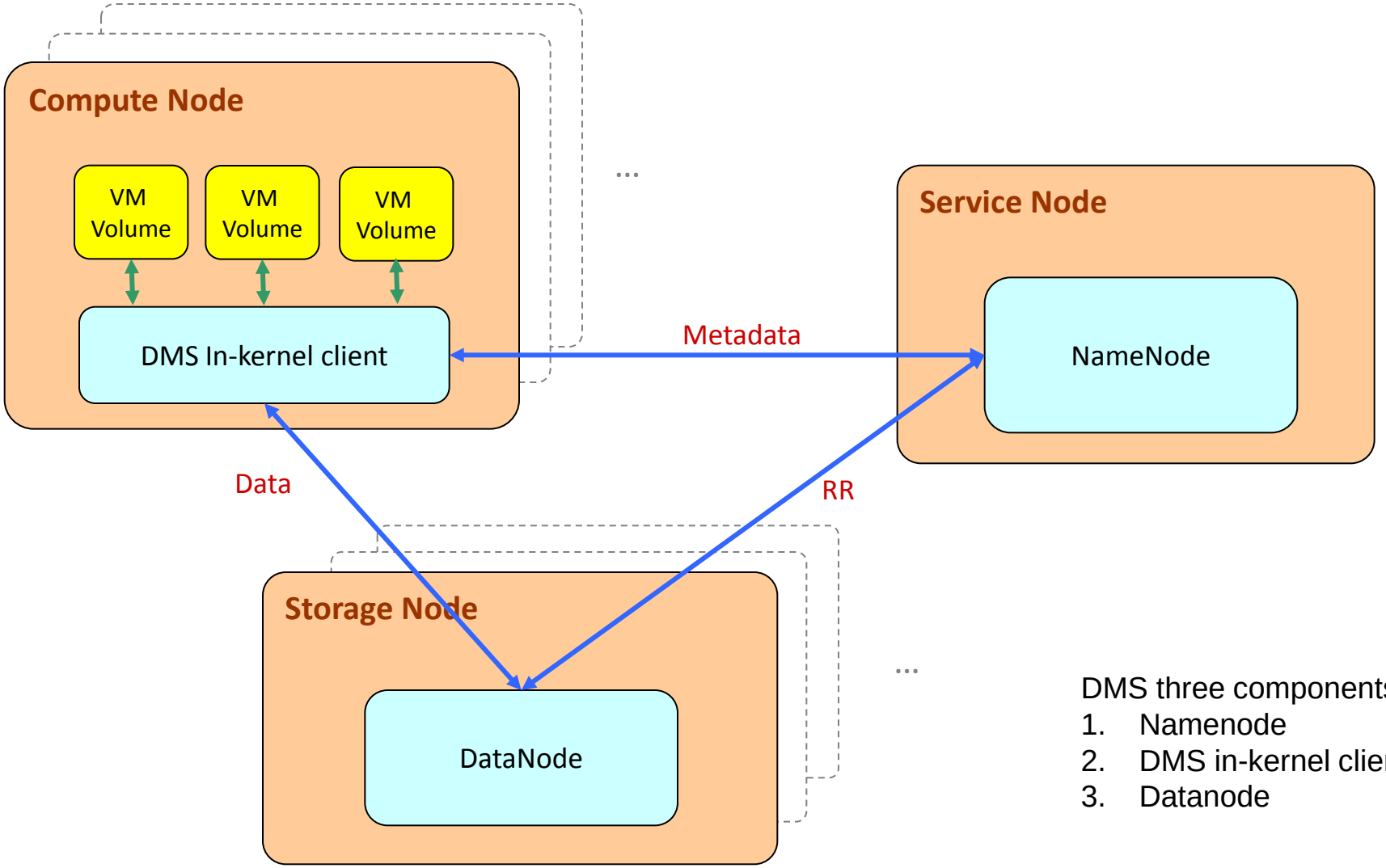
Key Features of Cloud Storage System from Providers' Perspective

- Use of **commodity hardware** – reducing cost (JBOD)
- Disk space management for up to **multiple petabytes**
- **Add disk, remove disk** without interruption
- **Thin provisioning**
- Enables you to provision a specific level of I/O performance if desired, by setting **I/O throttle**
- Block-level **de-duplication** for reducing space requirement

DMS in Cloud OS

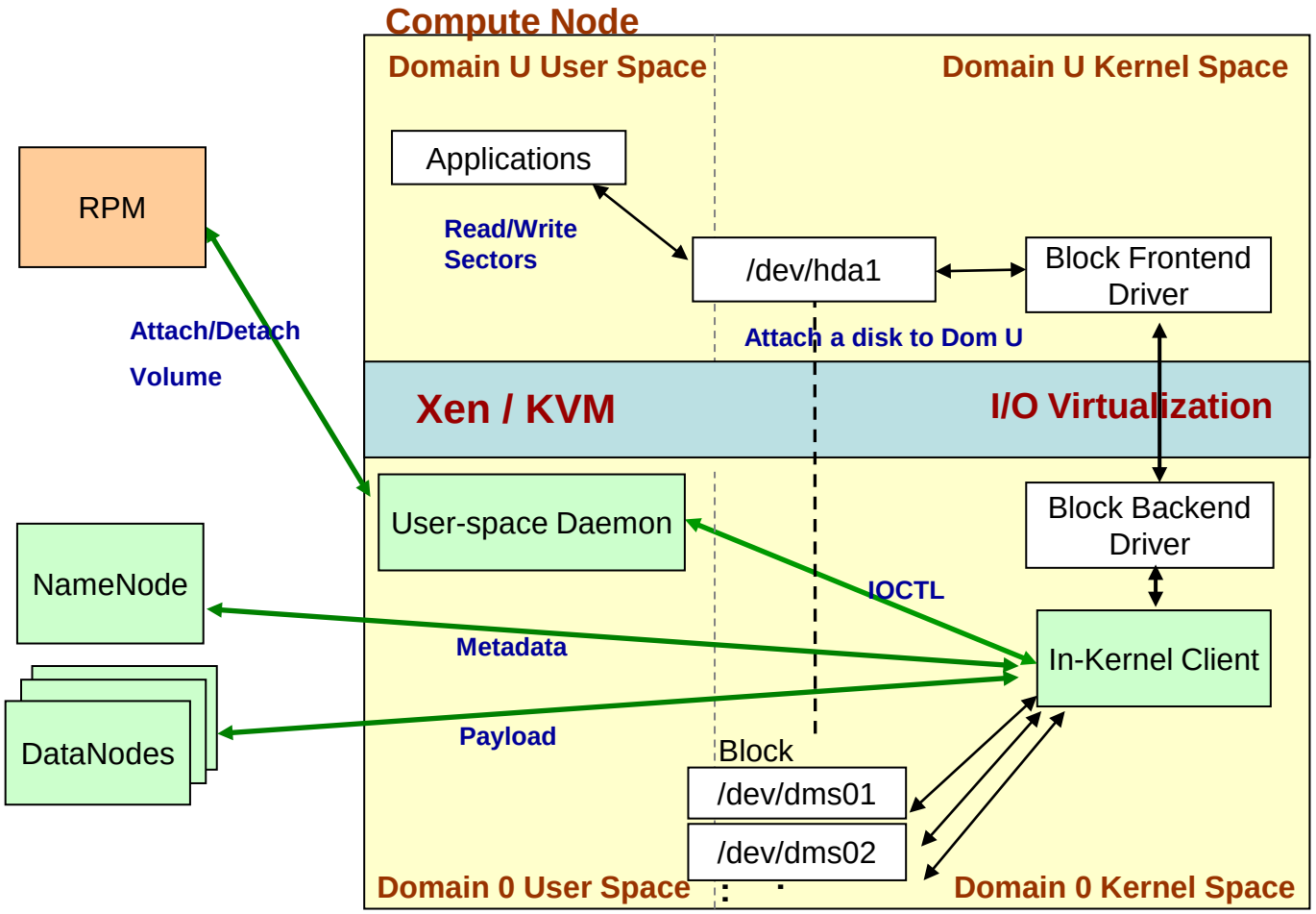
- **DMS** - Distributed Main Storage
- **Goal**: A **cloud scale network storage** solution that provide **high capacity, high reliability, high availability and high performance**
- **Features:**
  - Volume operations: **Create / Attach / Detach / Delete**
  - **High Reliability**: data replication (N=3)
  - **High Availability**: No single point of failure
  - **High Performance**: client-side metadata/data cache for performance
  - **Thin provisioning** is utilized to optimize utilization of available storage.
  - Fast **Volume Cloning**
  - **Save Image; FastBoot**
  - Disk **I/O Throttle**

DMS System Architecture

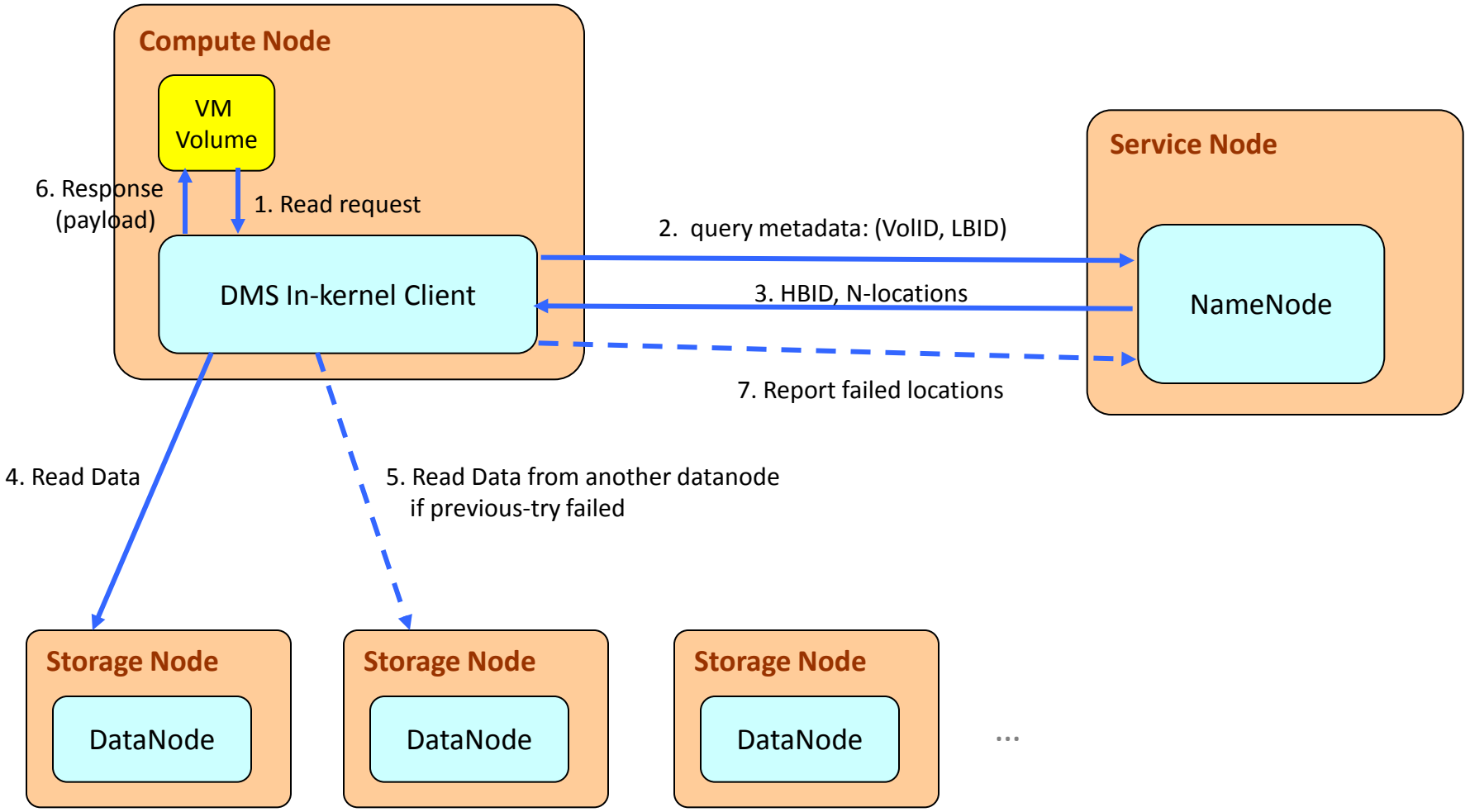


DMS three components:
1. Namenode
2. DMS in-kernel client
3. Datanode

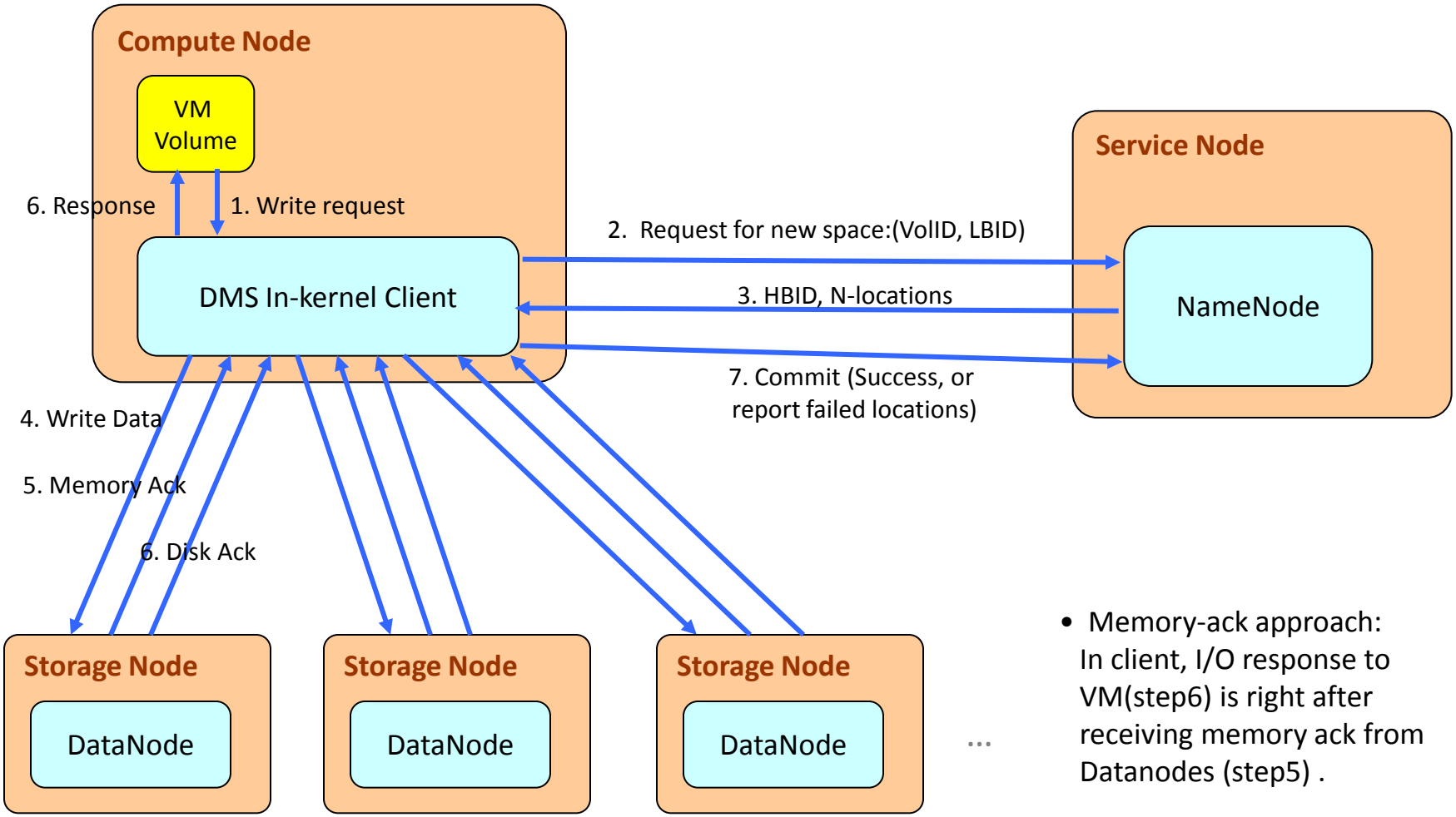
DMS Compute Node Architecture



Read Flow



Write Flow – Memory-Ack Approach

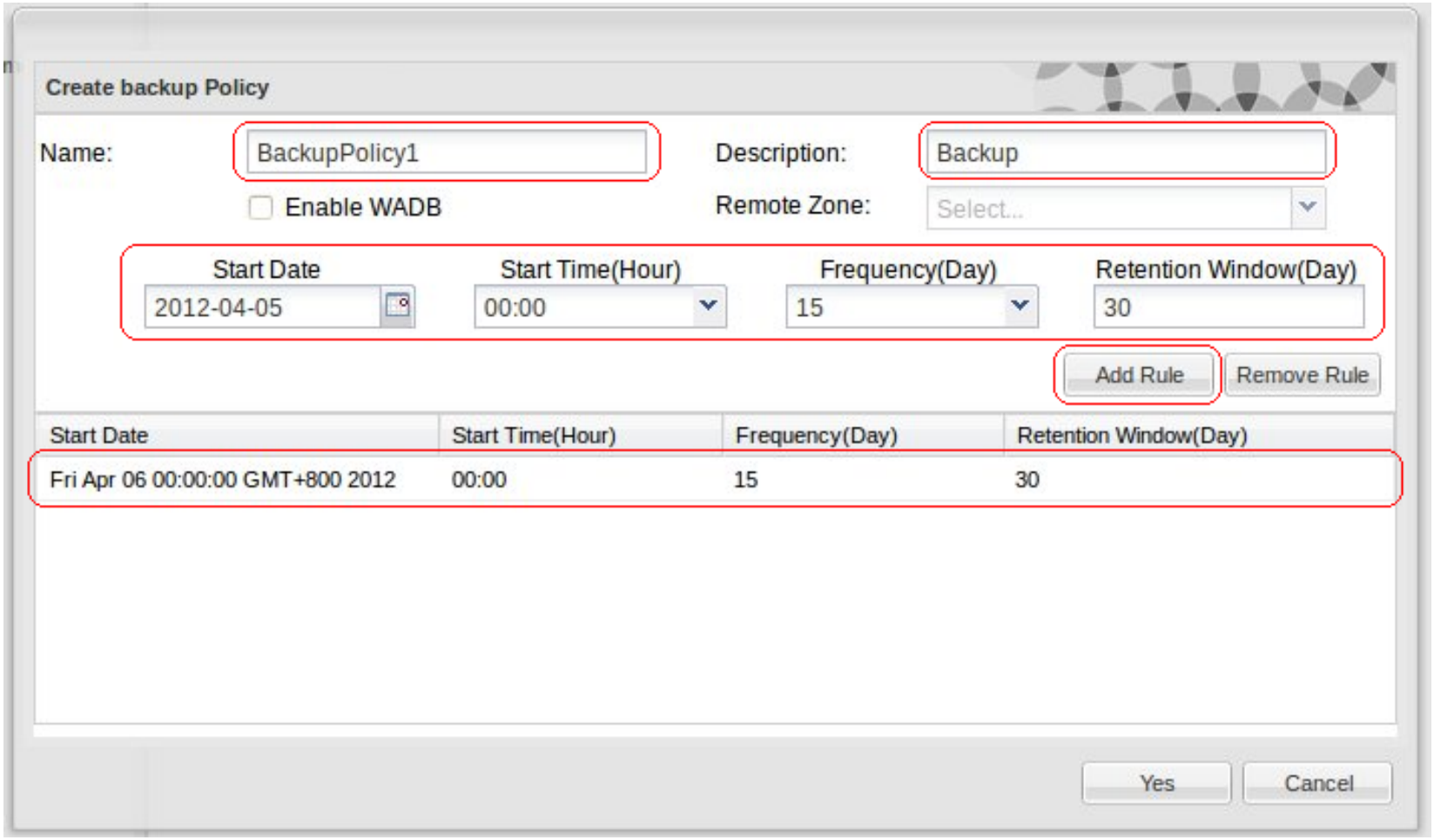


- Memory-ack approach:
  In client, I/O response to VM(step6) is right after receiving memory ack from Datanodes (step5) .

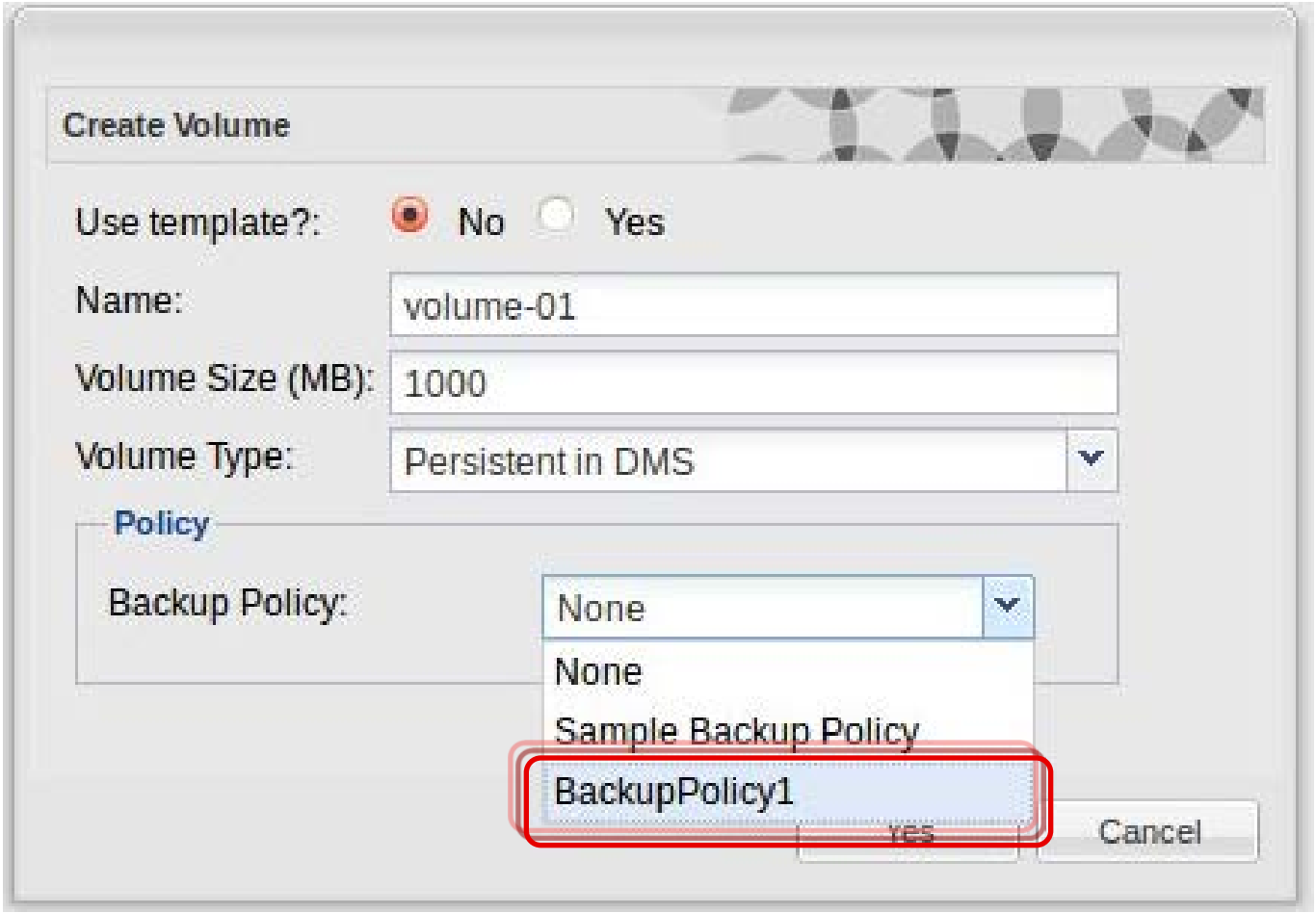
DSS in Cloud OS

- Share the **homogeneous space** with DMS
- Block-level Incremental backup
  - Copy-on-write (**COW**) and **volume cloning** techniques
  - Backup is just a **snapshot of meta-data**
- Support **volume-level restore**
- Block-level **de-duplication**
- **Garbage collection** of expired, un-used and redundant disk blocks
- **Wide-area data backup** / restore and near-realtime **replication**
- The secondary storage itself is failure-tolerant (**HA**)

DSS – Backup Policy

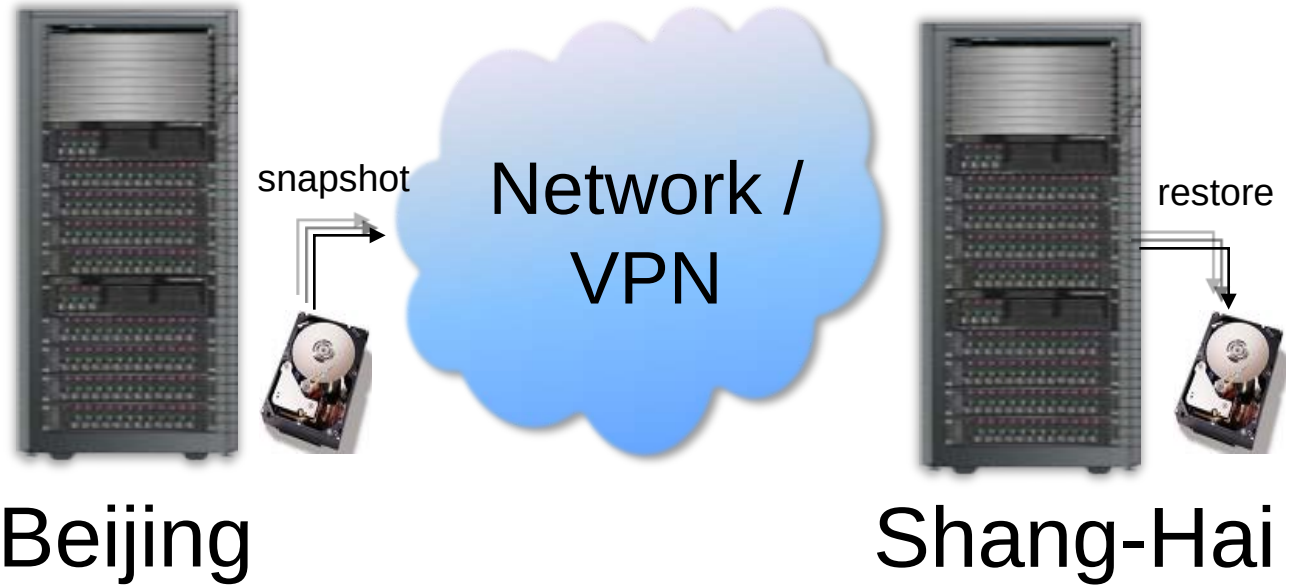


DSS – Assign Backup Policy

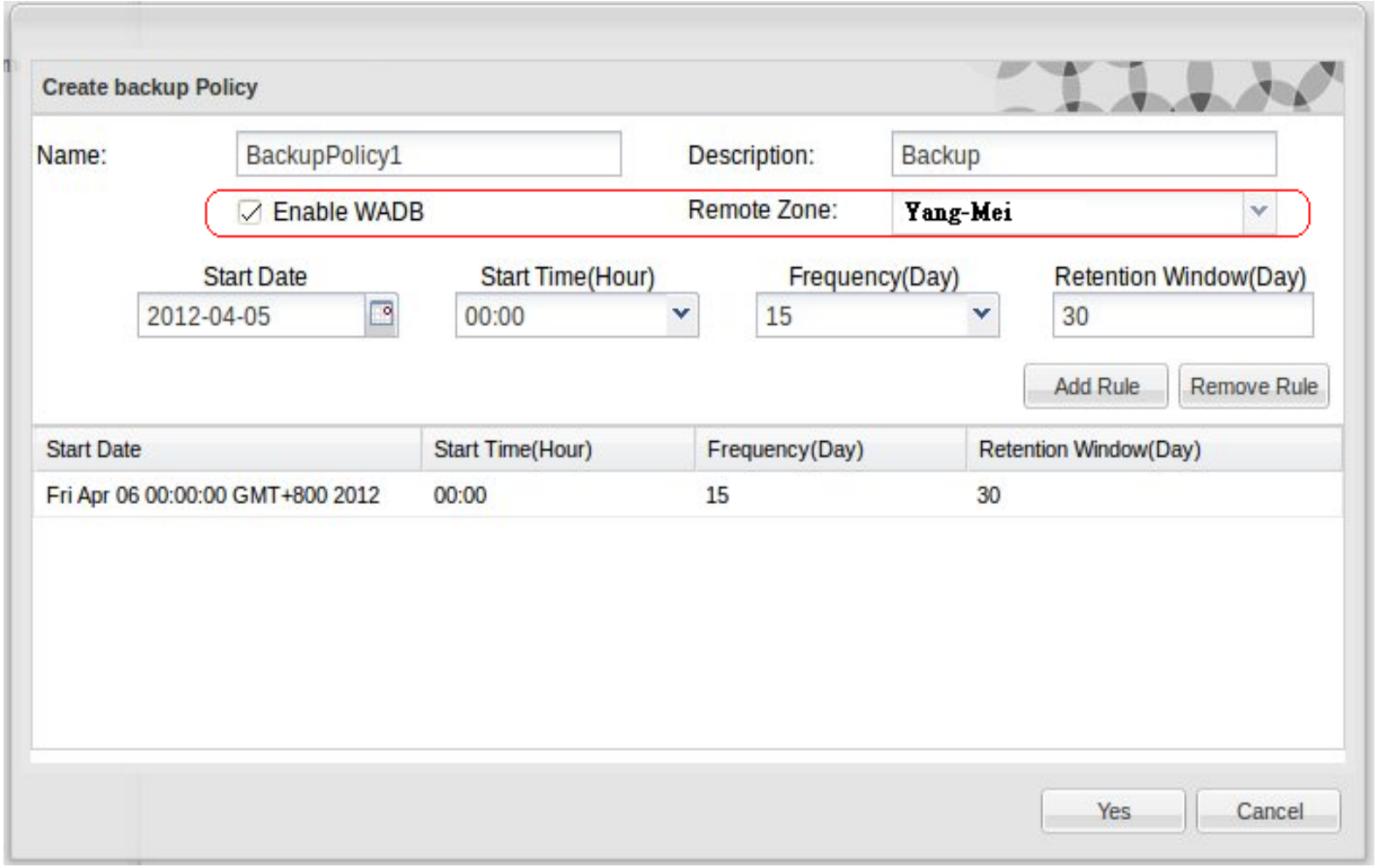


DSS – WADB

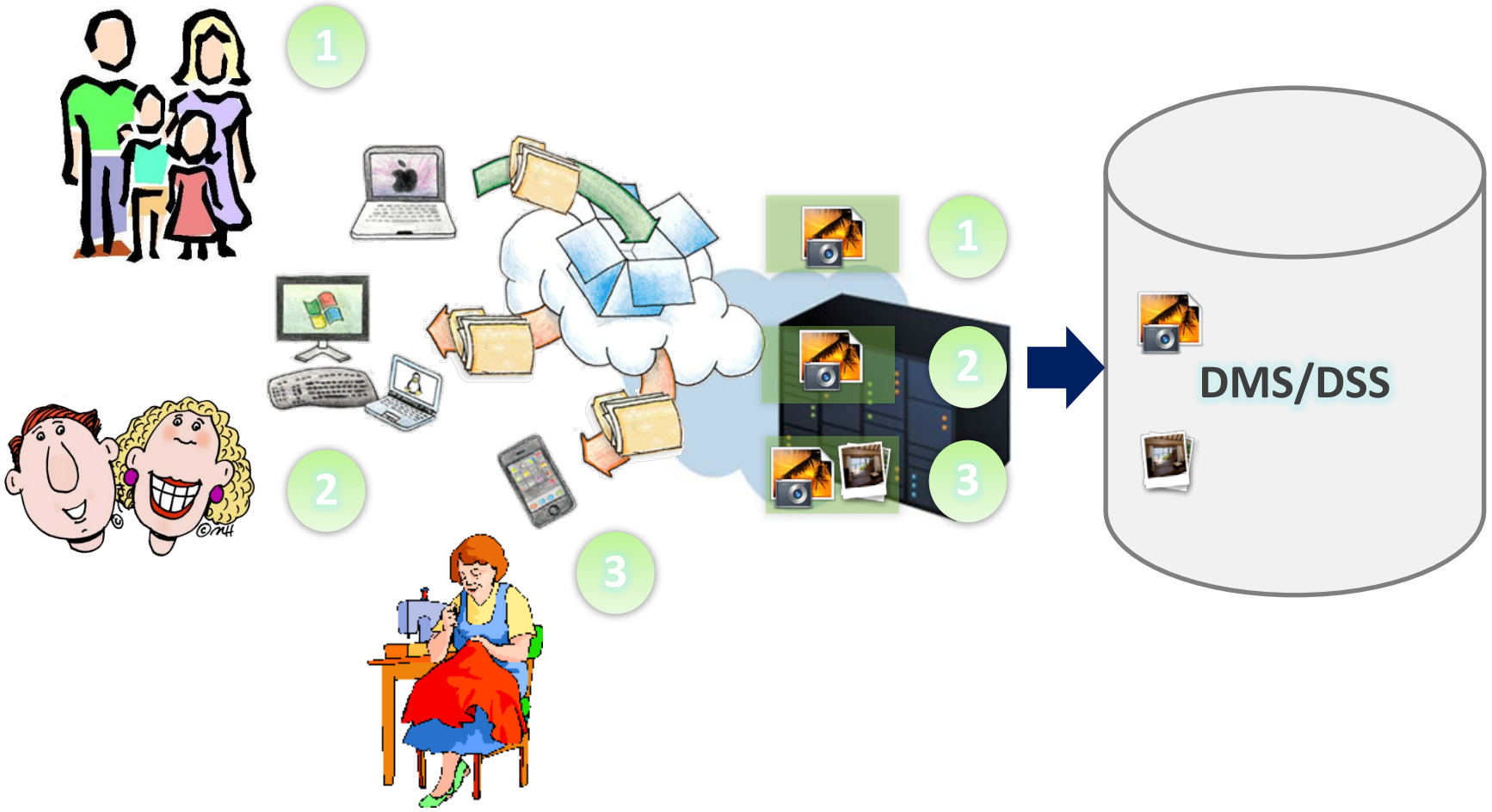
- Backing up the data to a remote location at the time volume snapshot is taken…



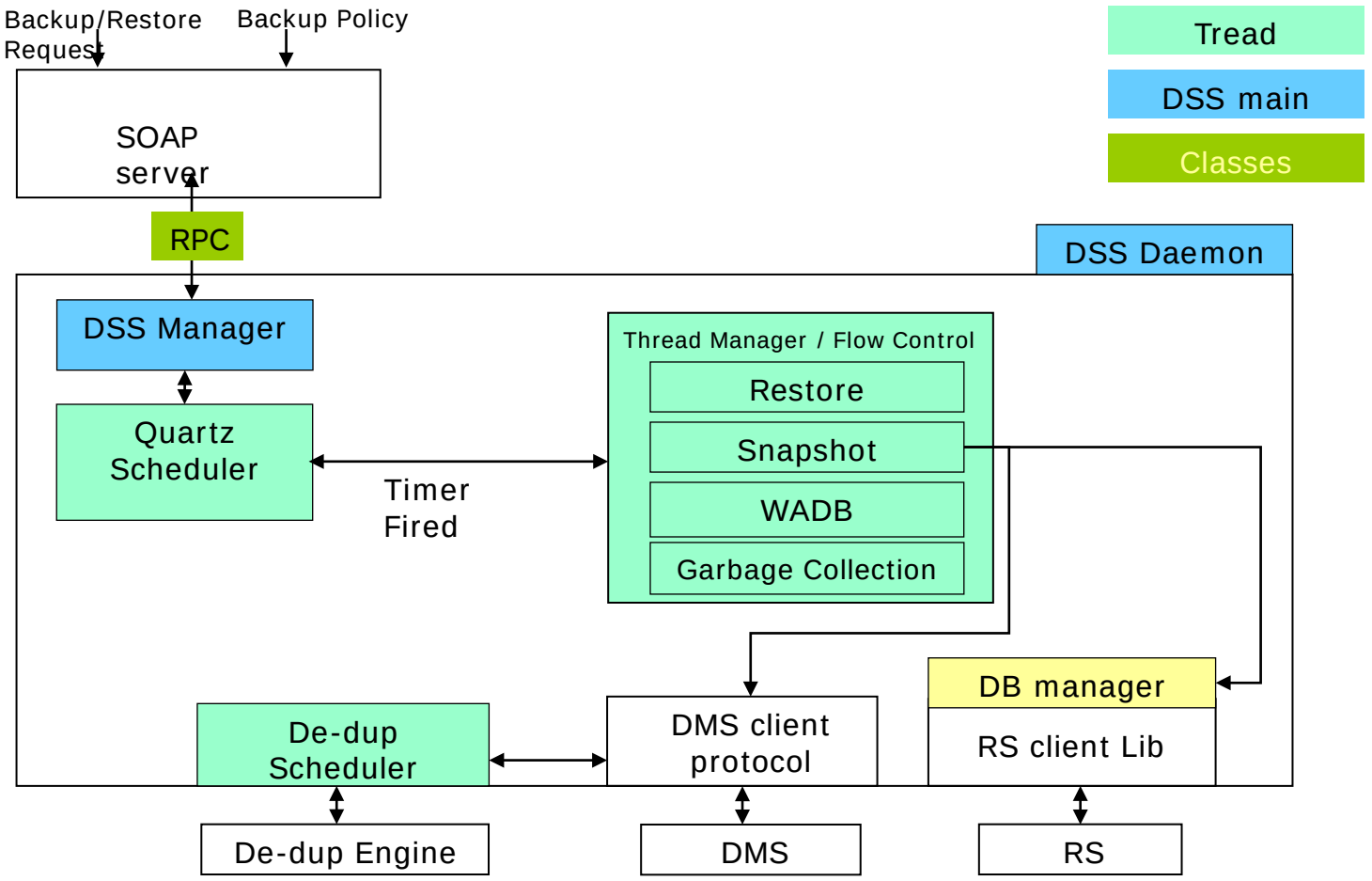
DSS – Enabling WADB



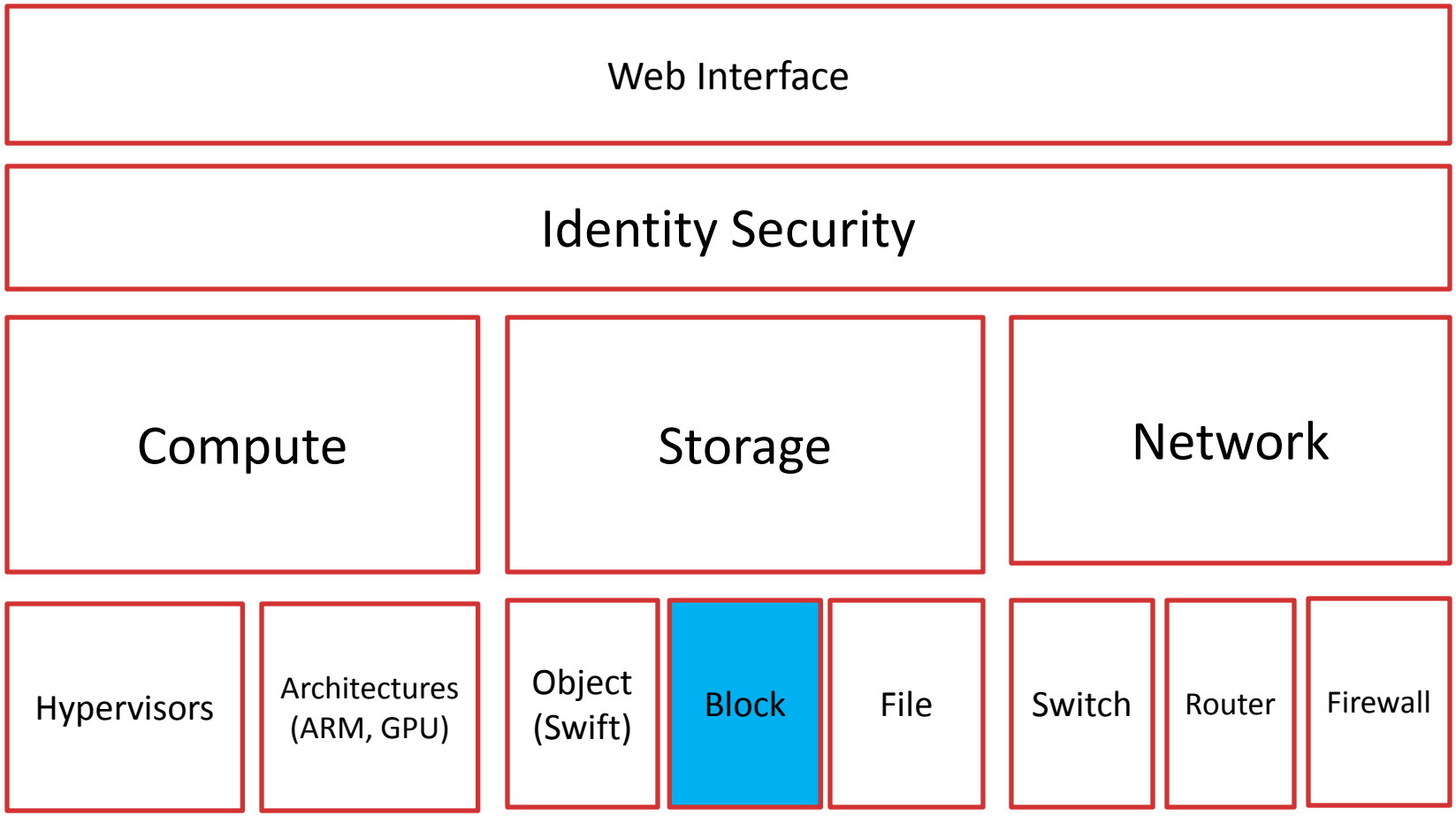
De-duplication in DropBox



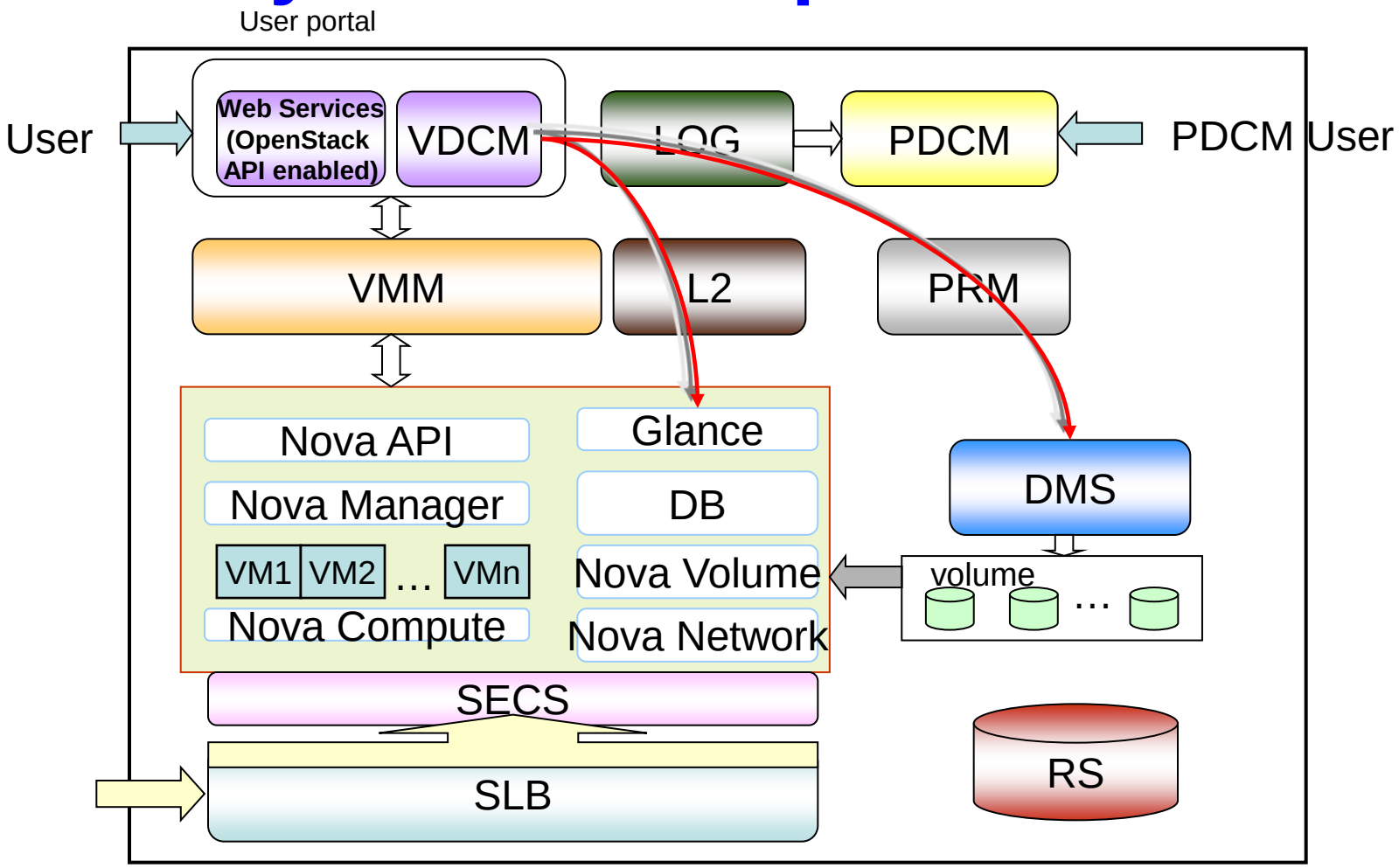
DSS Software Architecture Diagram



OpenStack



Integration of Cloud OS Storage System with OpenStack



Summary of Advantages

Feature	Benefits
Unlimited storage	highly scalable read/write access
Leverages commodity hardware (JBOD)	No lock-in, lower price/GB
Built-in replication	Fully redundant (default: N=3)
Used as system volume	Save Image, Fast Boot, etc.
Thin Provisioning	Optimize utilization of available storage
Highly available (HA) storage servers (DMS/DSS)	Server failover for high available
Snapshot and backup for block volumes	Data protection and recovery for VM data
Policy based backup	Free of worry – set it and forget about it
Wide-area data backup	For disaster recovering
Block-level data de-duplication	Store more data in the limited space
Standalone volume API, CLI available	Easy integration with other compute systems

More Feature in V2.0

Feature	Benefits
Volume Cloning	Fast cloning without copying data
Share virtual disk within cluster	Support sharing of volume and cluster file system
Disk bandwidth QoS guarantee	Dynamic adjustment at I/O request level
Storage space optimization	N-way replication for write-intensive blocks N-parity-group for read-intensive blocks
Wide-area data replication	Near real-time wide-area replication
Integration with Compute / Volume	Fully integrated to Compute for attaching block volumes and reporting on usage

Q&A

Thank You !