



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



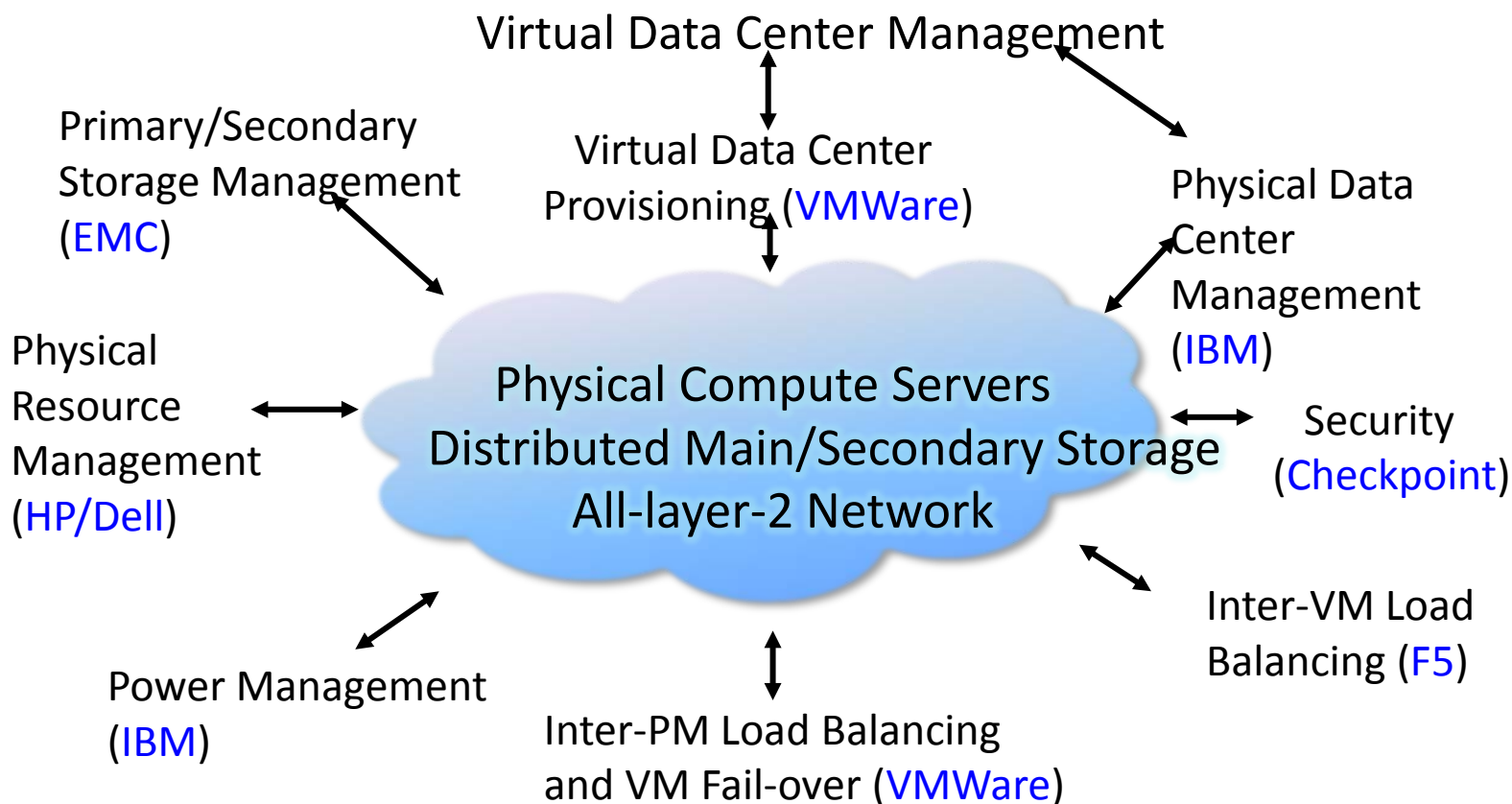
Overview

- Data is growing exponentially and moving on to Cloud. 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)



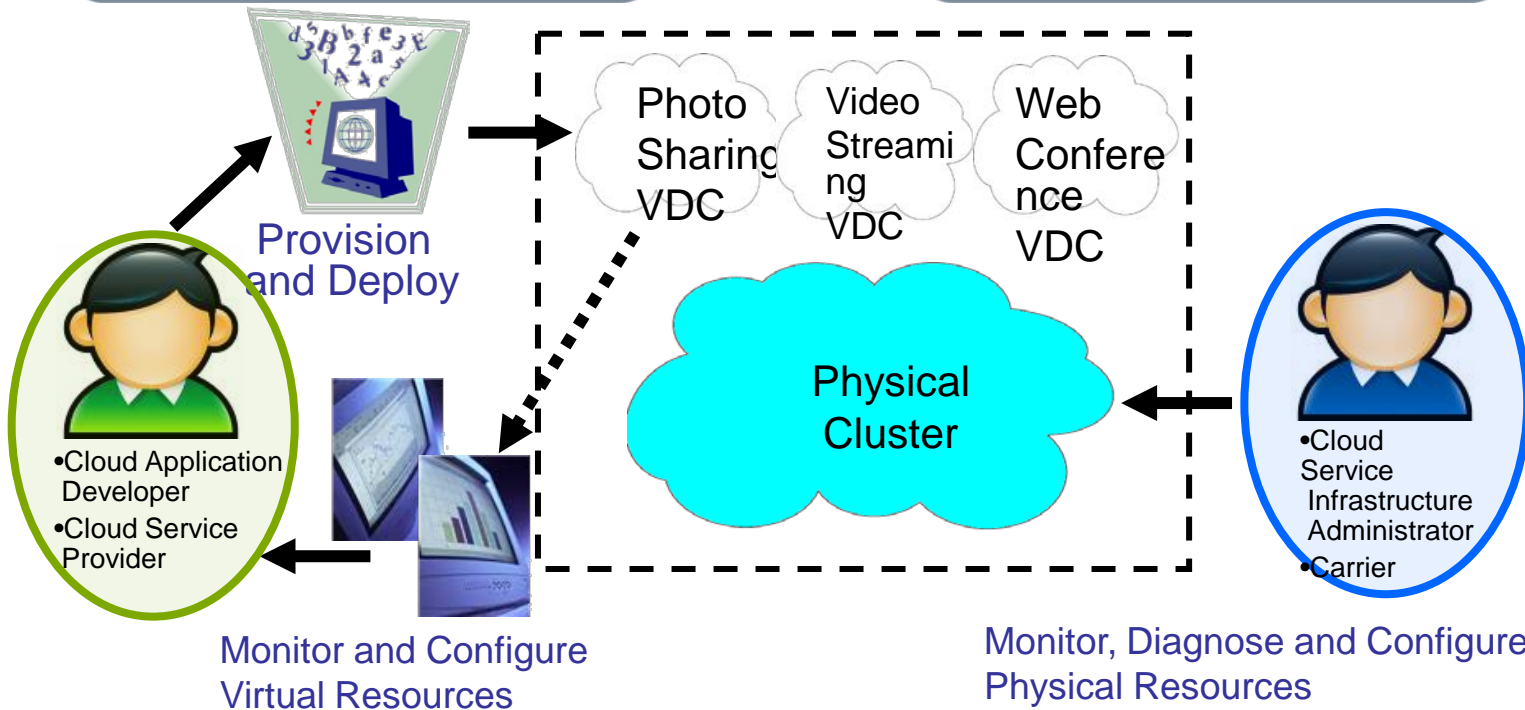
*Blue name in bracket indicates the leading company of the component

ITRI Cloud OS Service Interface

Multiplexing VDC's in a Physical DC

Virtual Data Center Management

Physical Data Center Management



VDCM – Assets (VDC, VC, VM)

CCMA VDCM 1.0

My Account History: test11VC Logout

Main Monitor Composition Usage Account

Refresh Save As Template Unset Proxy Cluster

Assets

- RogerDC
- RogerLDAP
- EricDC
- test11
 - test11VC
 - test11VM1
 - test22VC
 - test22VM1

Image

Volume

Composition: Assets > test11 > test11VC

Name: test11VC

Create Time: Thu Jan 05 15:21:29 GMT+800 2012

Public IP: 172.106.210.10

DNS: 172.106.210.10

SSH/RDP Proxy:

Platform: Linux

Firewall: Default Firewall Policy

Load Balancer: Default Load Balancing

Auto Scaling: None Linux Policy

Refresh Create Delete Run Stop Resume Suspend Instances

Name	Instance Type	Volume Type	Instance			Physical Machine	Private IP	Status	Auto Gen	Cre
			Image	Kernel	Ramdisk					
test11VM1	Standard-CPU, Sm	Persistent in ISI	CentOS5.5	kernel_2.6.	ramdisk_2.	CWISR6S48	172.106.222.229	Running	false	Thu

VDCM – System Images

CCMA VDCM 1.0

Main

Refresh

Image Name	Description	Type	Platform
CentOS5.5	centos.5-5	Machine	Linux CentOS
kernel_2.6.18-194.el5xen	vmlinuz-2.6.18-194.el5xen	Kernel	Linux CentOS
ramdisk_2.6.18-194.el5xen	initrd-2.6.18-194.el5xen	Ramdisk	Linux CentOS
Windows 2003	windows	Machine	Windows 2003
CentOS5.5-withSEC&SLB	CentOS5.5-withSEC&SLB	Machine	Linux CentOS
		Machine	Linux CentOS

Page 1 of 1

Refresh | + Upload | - Delete

Image Name	Description	Type	Platform
CentOS5.5-yousign	CentOS5.5-yousign	Machine	Linux CentOS
kernel_2.6.18-164.6.1.el5...	vmlinuz-2.6.18-164.6.1.el5xen	Kernel	Linux CentOS
ramdisk_2.6.18-164.6.1.e...	initrd-2.6.18-164.6.1.el5xen	Ramdisk	Linux CentOS
Windows XP	Windows XP	Machine	Windows 2003



VDCM - Volumes

The screenshot shows the VDCM interface with a list of volumes and a 'Create Volume' dialog box. The dialog box is open, showing the 'Create Volume' form. The 'Name' field is empty, 'Volume Size (MB)' is empty, and 'Volume Type' is set to 'Persistent in DMS'. The 'Backup Policy' dropdown menu is open, showing options: None, Sample Backup Policy, back2, backupPolicy165522, backupPolicy171931, backupPolicy172411, backupPolicy172451, backupPolicy172623, and backupPolicy174751.

Name	Volume Size (MB)	Volume Type	Status	Attached	Read-only	System Volume	WADB Volume
vol10020	1000	Persistent in DMS	Success	false	false	false	false
vol172411	1000	Persistent in DMS	Success	false	false	false	false
vol172451	1000	Persistent in DMS	Success	false	false	false	false
vol172623	1000	Persistent in DMS	Success	false	false	false	false
vol174751	1000	Persistent in DMS	Success	false	false	false	false
vol1	100000	Persistent in DMS	Success	false	false	false	false
vol2	10000000	Persistent in DMS	Success	false	false	false	false
i-6ABAC4CD-system-volume	4461	Persistent in DMS	Success	false	false	false	false
testVol	10	Persistent in DMS	Success	false	false	false	false
testVolISCSI	200	Persistent in iSCSI	Success	false	false	false	false



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

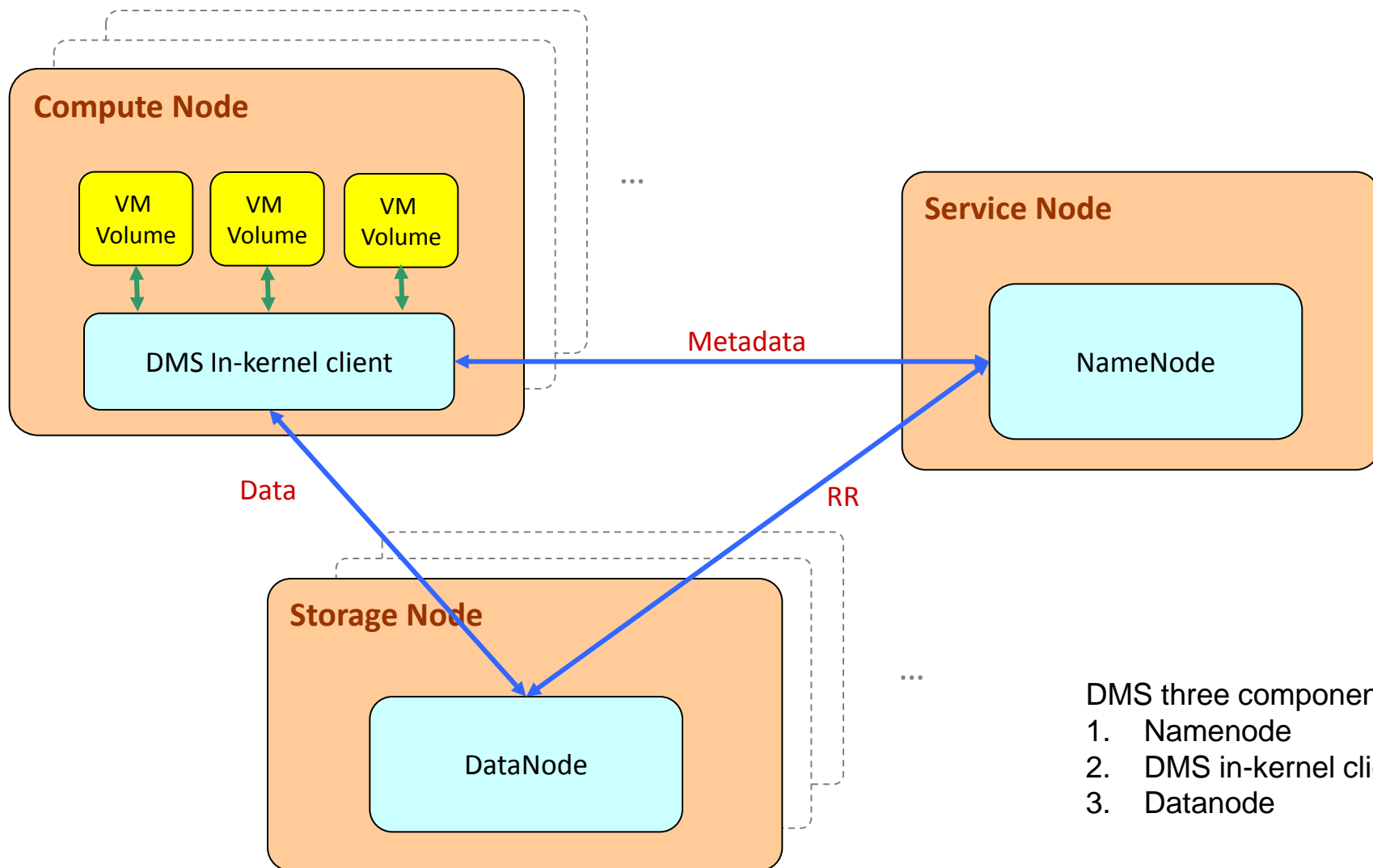


DMS in Cloud OS

- **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**

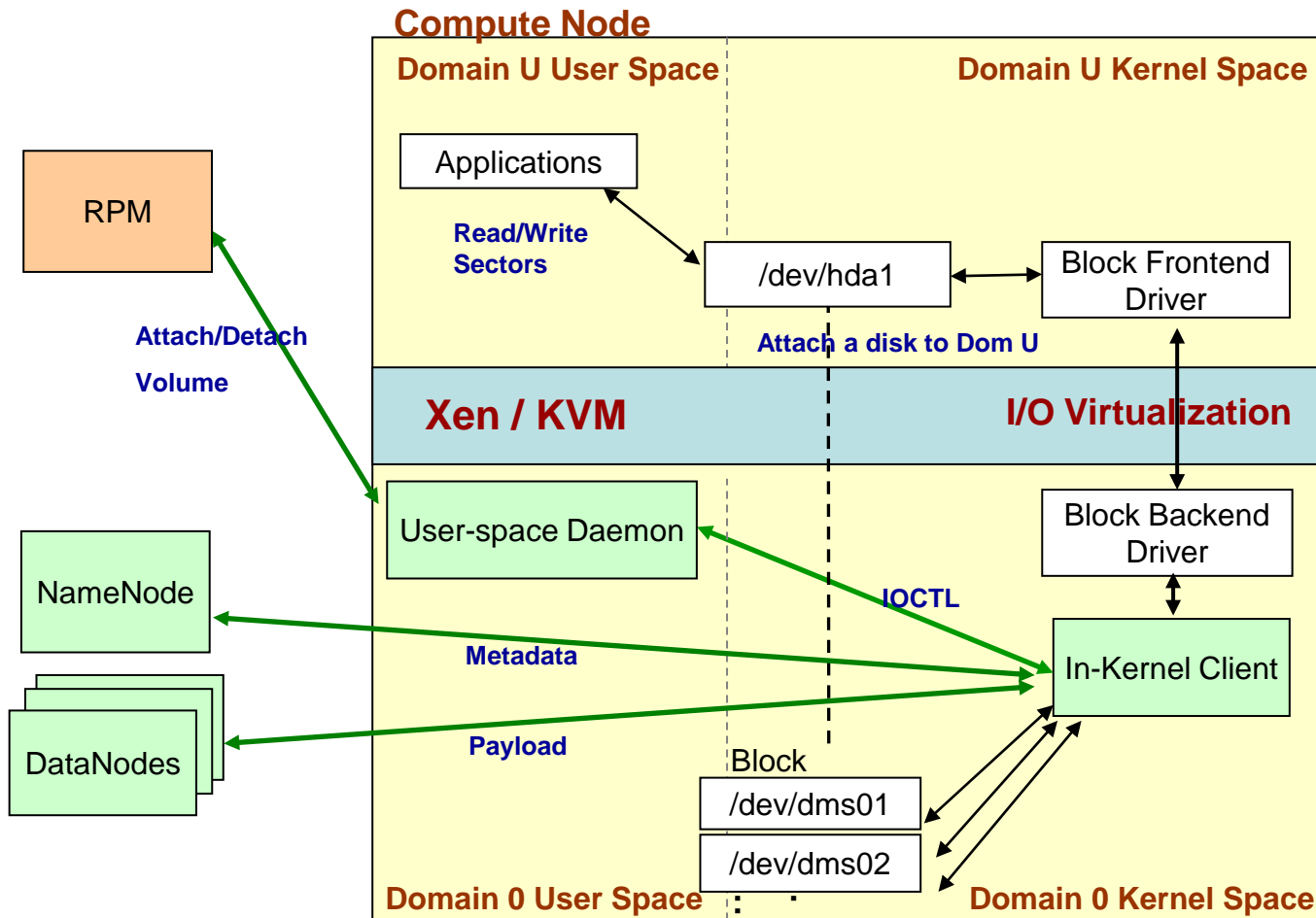


DMS System Architecture

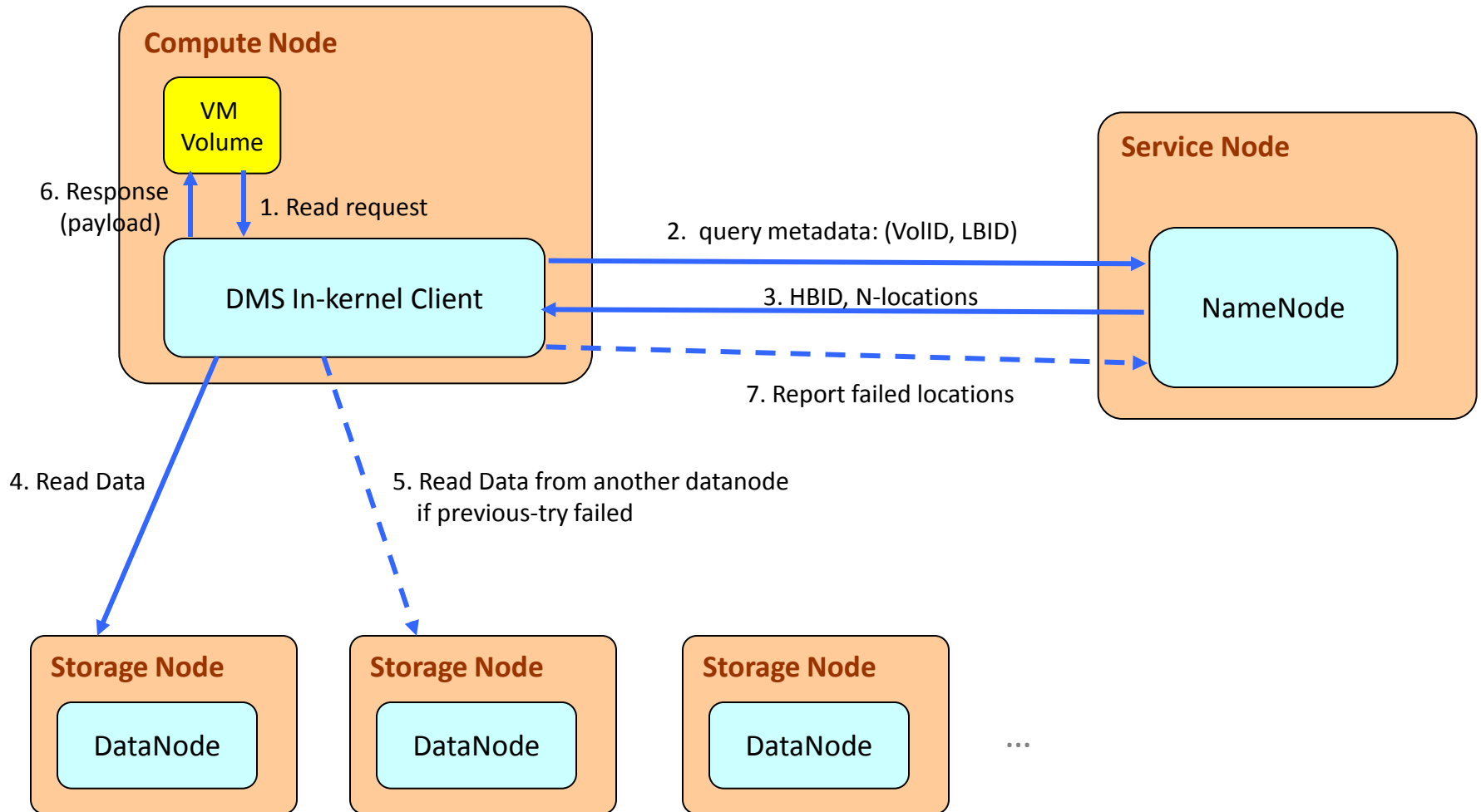


- DMS three components:
1. Namenode
 2. DMS in-kernel client
 3. Datanode

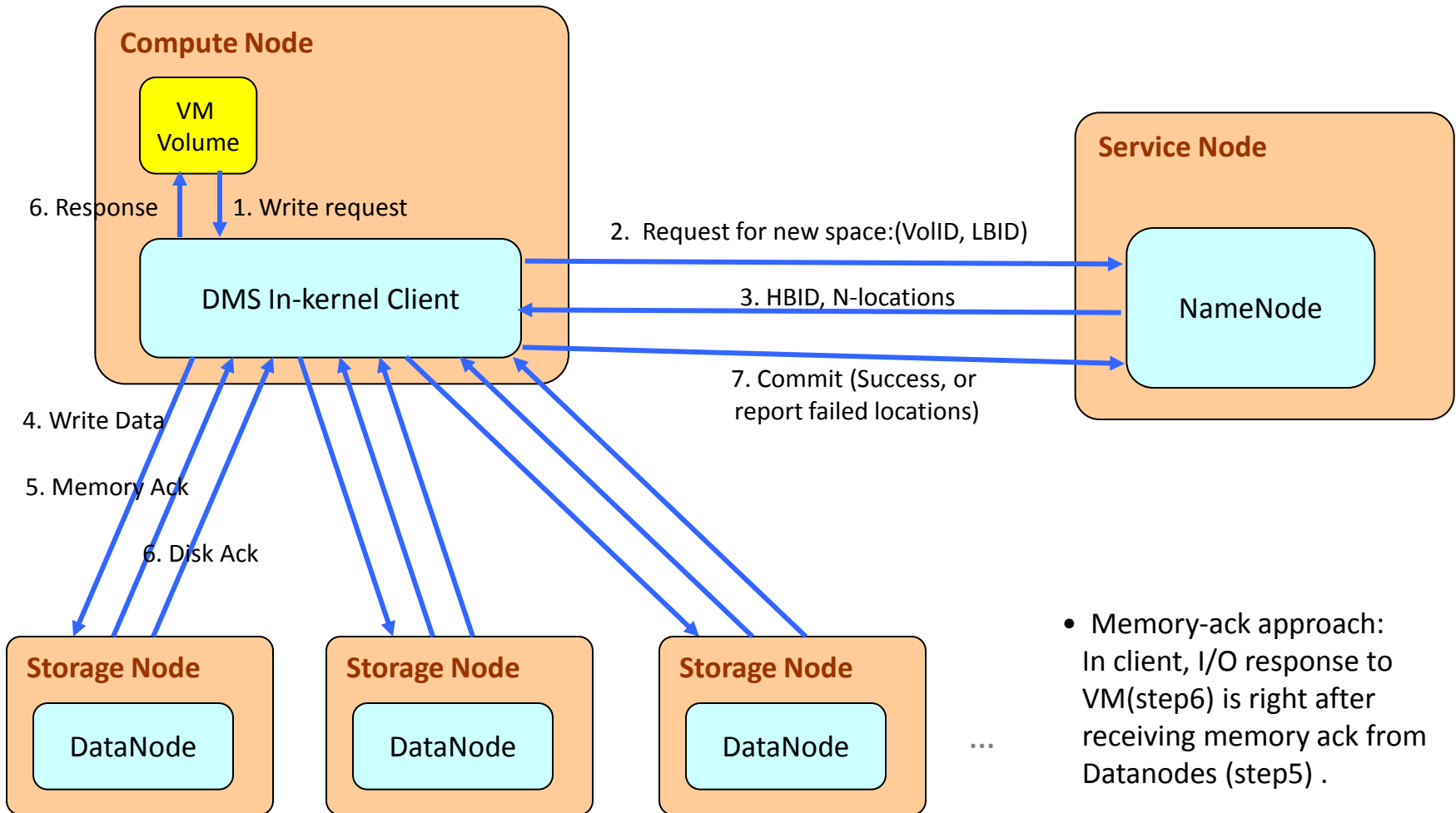
DMS Compute Node Architecture



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

Create backup Policy

Name: Description:

Enable WADB Remote Zone:

Start Date: Start Time(Hour): Frequency(Day): Retention Window(Day):

Start Date	Start Time(Hour)	Frequency(Day)	Retention Window(Day)
Fri Apr 06 00:00:00 GMT+800 2012	00:00	15	30



DSS – Assign Backup Policy

Create Volume

Use template?: No Yes

Name:

Volume Size (MB):

Volume Type:

Policy

Backup Policy:

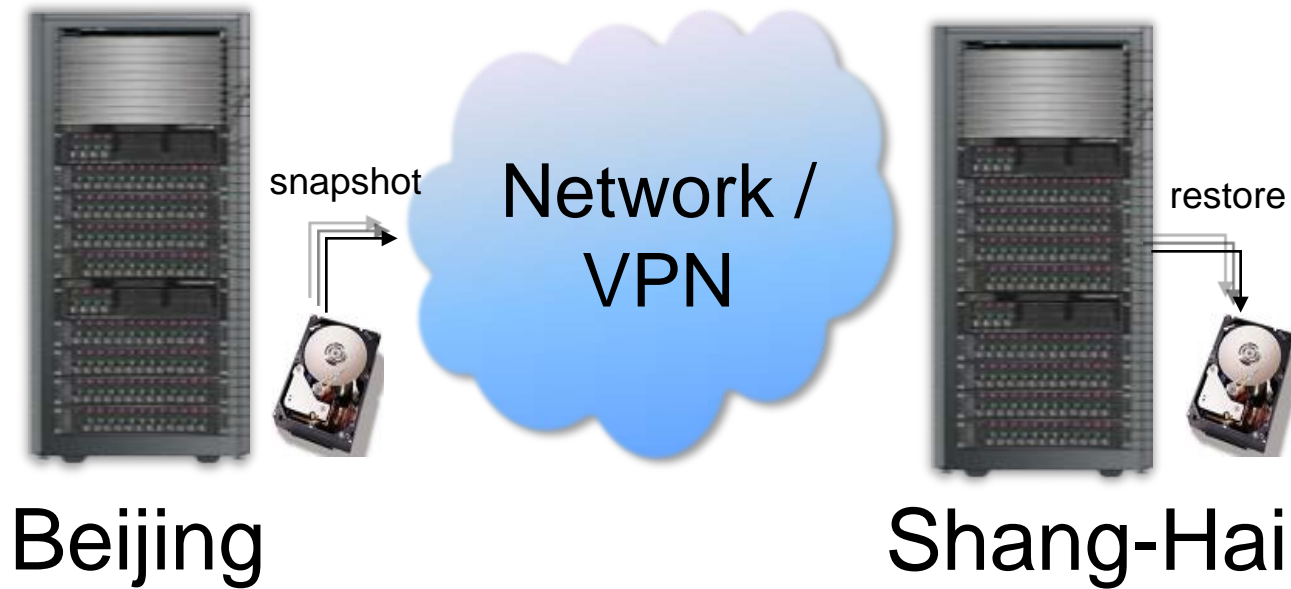
- None
- None
- Sample Backup Policy
- BackupPolicy1**

YES Cancel



DSS – WADB

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





DSS – Enabling WADB

Create backup Policy

Name: BackupPolicy1 Description: Backup

Enable WADB Remote Zone: Yang-Mei

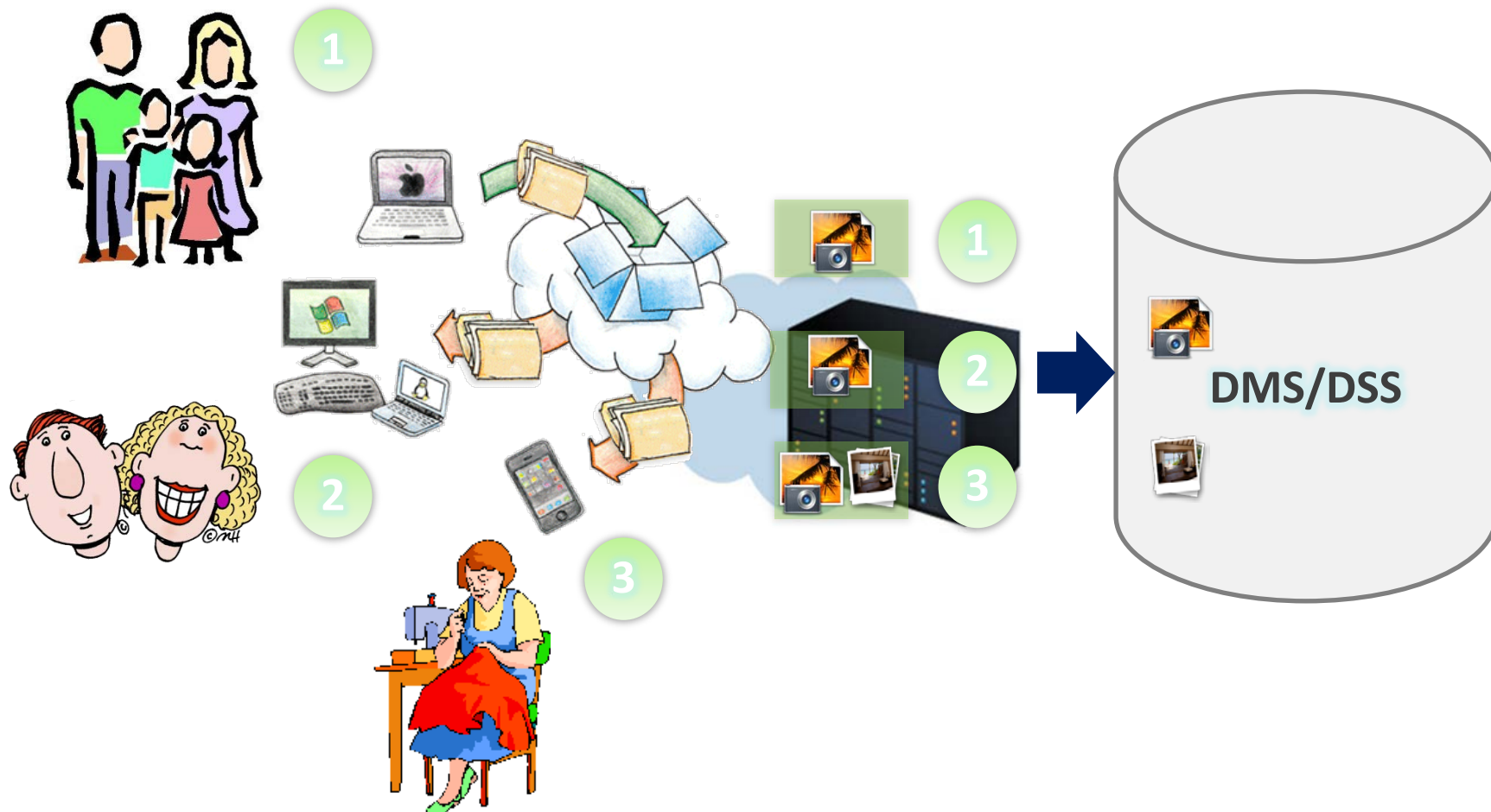
Start Date: 2012-04-05 Start Time(Hour): 00:00 Frequency(Day): 15 Retention Window(Day): 30

Add Rule Remove Rule

Start Date	Start Time(Hour)	Frequency(Day)	Retention Window(Day)
Fri Apr 06 00:00:00 GMT+800 2012	00:00	15	30

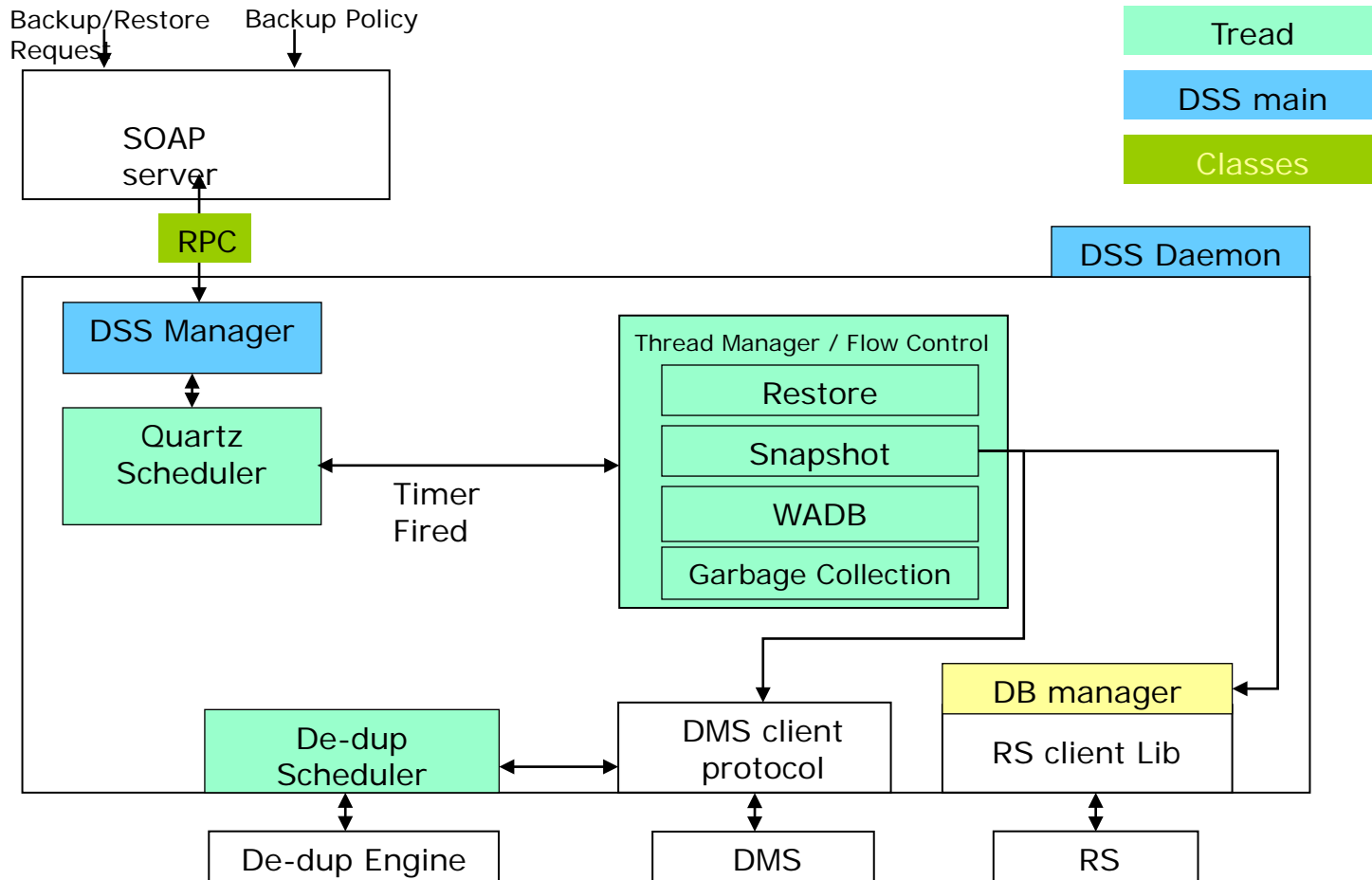
Yes Cancel

De-duplication in DropBox



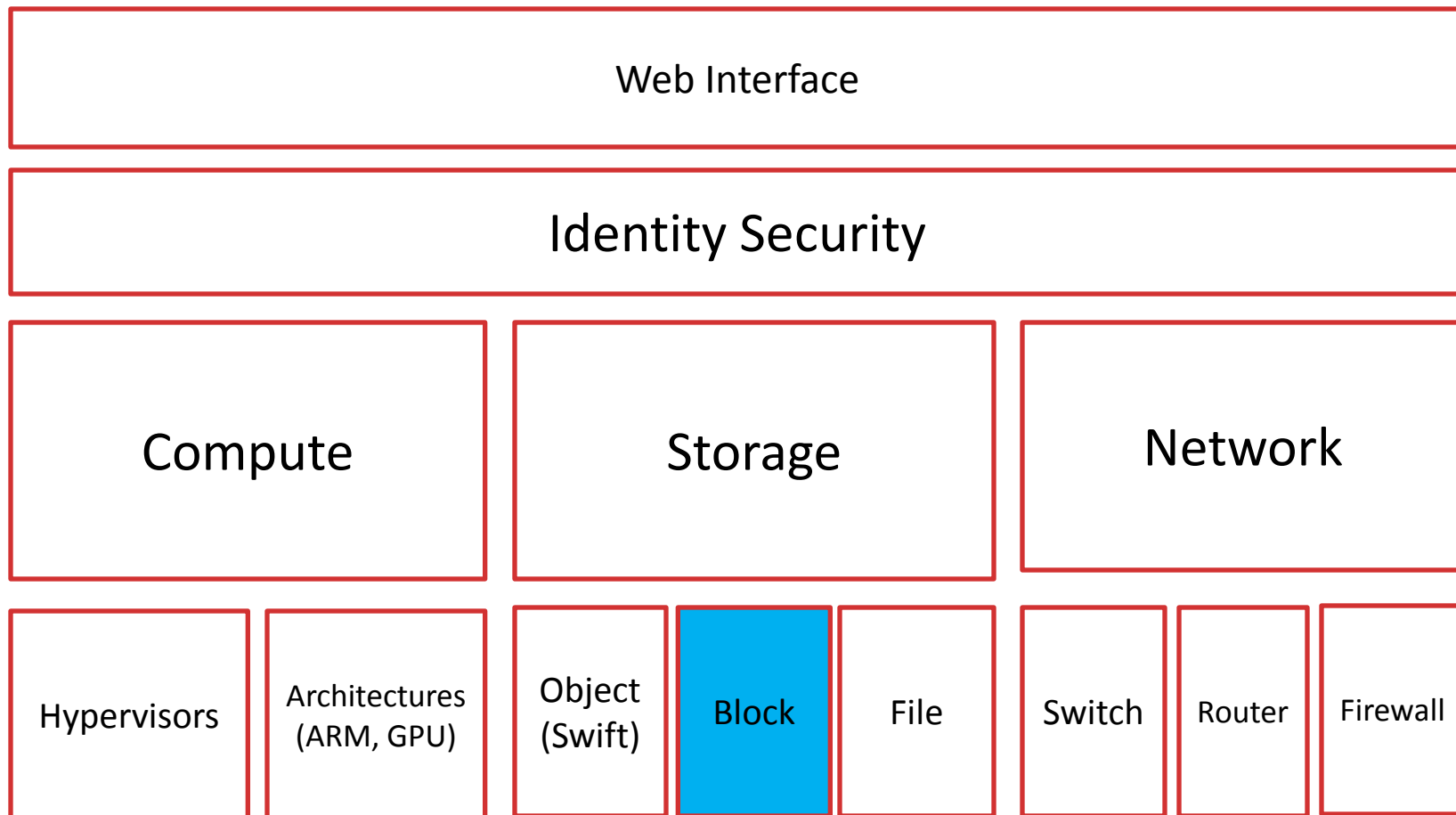


DSS Software Architecture Diagram

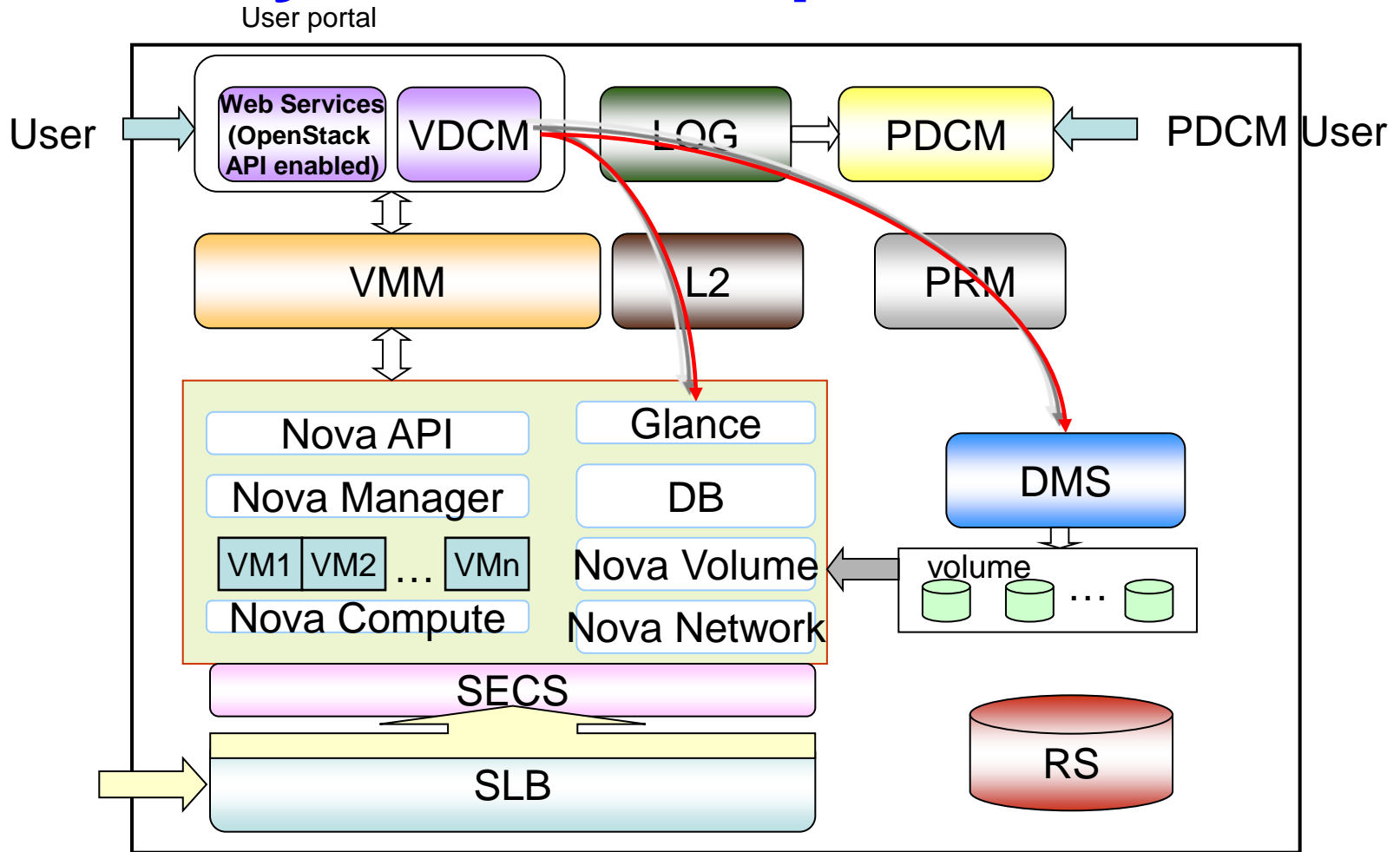




OpenStack



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)
<u>Used as system volume</u>	Save Image, Fast Boot, etc.
<u>Thin Provisioning</u>	Optimize utilization of available storage
<u>Highly available (HA) storage servers (DMS/DSS)</u>	Server failover for high available
Snapshot and backup for block volumes	Data protection and recovery for VM data
<u>Policy based backup</u>	Free of worry – set it and forget about it
<u>Wide-area data backup</u>	For disaster recovering
<u>Block-level data de-duplication</u>	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 !