

OPENSTACK DAYS
CHINA

Topic: ITRI OpenStack Distribution

Speaker: Yuh-Jye (EJ) Chang 張裕杰



About Myself

- 1984-1988 NTU ME BS
- 1994-1999 Syracuse CS PhD
- 1998-2006 Lucent/Bell Labs
- 2006-2011 Alcatel-Lucent/Bell Labs
- 2011-Present ITRI/CCMA S Division
- 2015-Present ITRI/ICL F Division



Agenda

- About ITRI OpenStack Distribution
- BAMPI
- High Availability
- Disco (Cinder Plugin)
- SOFA (All flash storage)
- Peregrine (Neutron Plugin)
- PDCM (Monitoring)



Why ITRI OpenStack?

Because we need

- Scalable and comprehensive bare metal provisioning
- HA support for every OpenStack system component
- Standard operating procedures (SOPs) and tools for change management
- Scalability for Internet-facing packet processing
- Overhead-minimizing network virtualization
- Physical data center administration tool
- PMLS (HaaS): Physical Machine (Hardware) Leasing Service)



What's inside IOD?

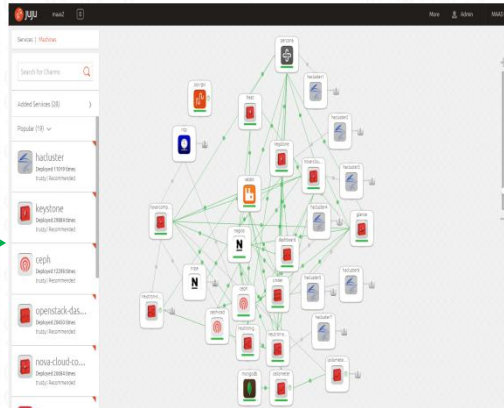
- Auto Deployment from Bare Metal
- ITRI OpenStack Components High Availability
- Dual Switch Protection
- Physical Data Center Monitor
- Cinder Plugin - DISCO
- Neutron Plugin - Peregrine
- Compute Node Failover
 - Move VMs in the broken Host to another healthy Host



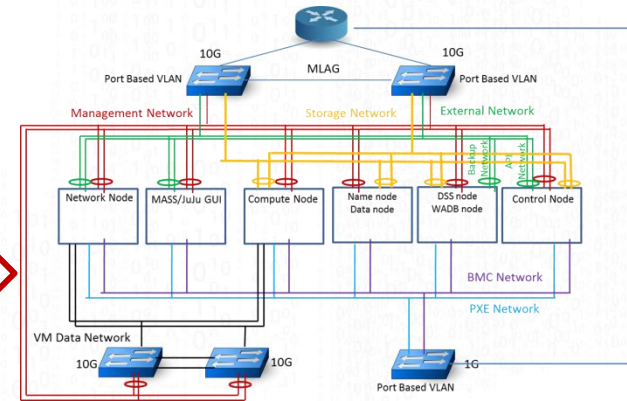
IOD Deployment Procedure



Deploy from
Bare-metal



ITRI OpenStack Distribution



ITRI OpenStack Distribution
Network Architecture



BAMPI

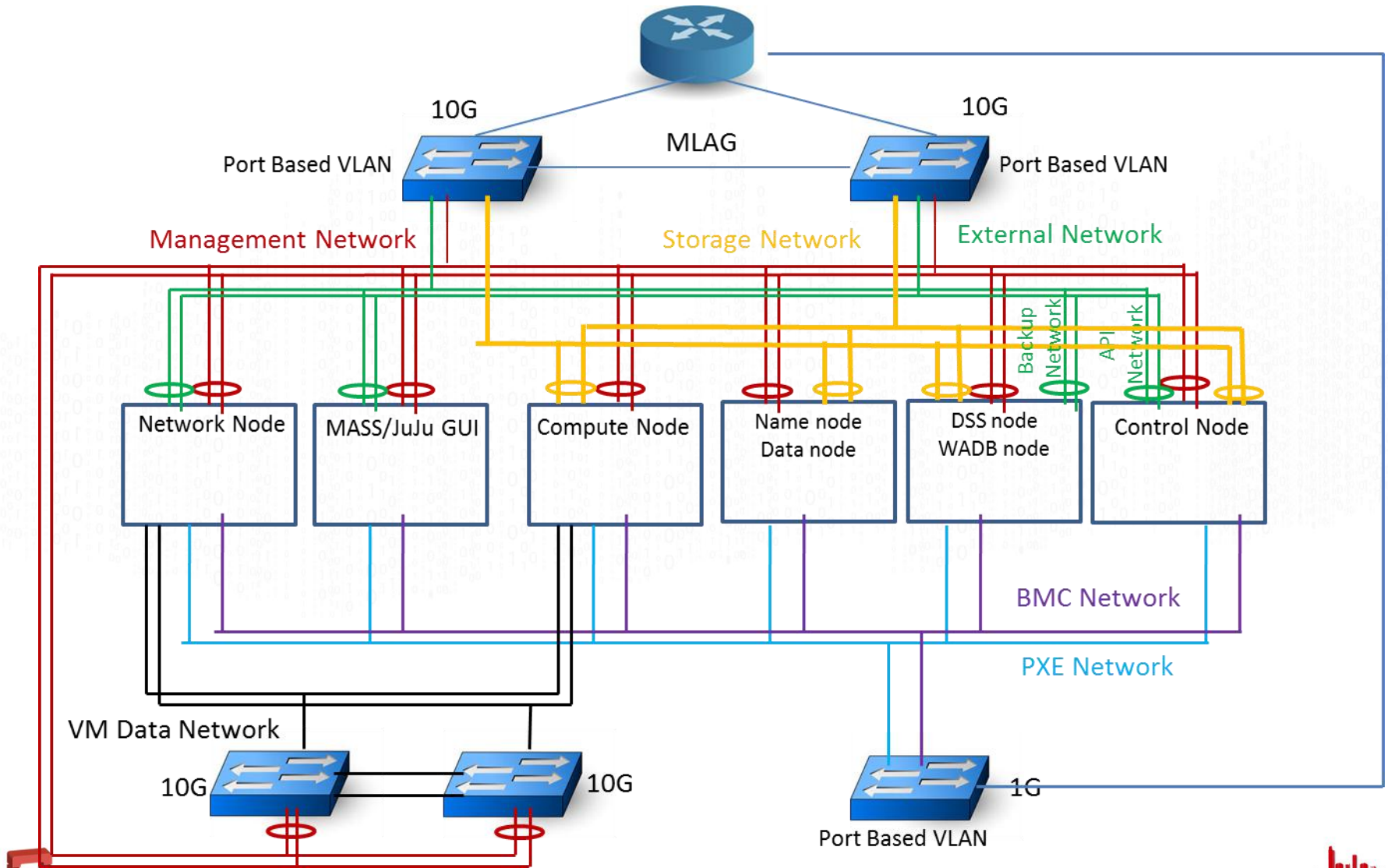


- BAMPI is an infrastructure software application used in data centers to deploy servers from bare metal.
- BAMPI can be used to remotely configure BIOS, BMC, RAID ,OS and restore operating systems on servers.
- In addition, BAMPI can take care of hardware-specific tasks such as firmware upgrades, check BIOS, BMC, RAID and OS.

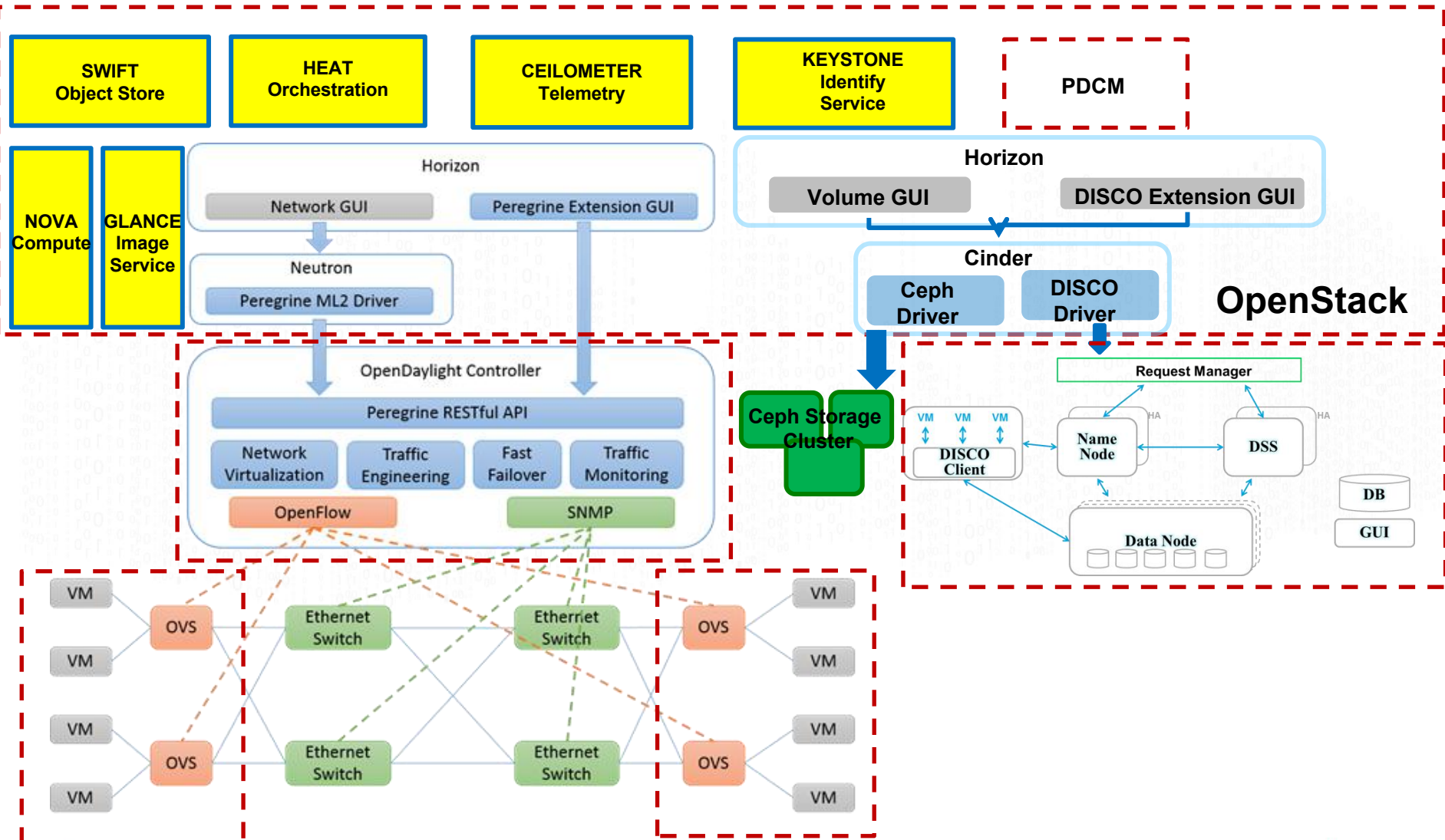
	Manpower	BAMPI
Initialize BMC Network	※ Time of Completion for 80 servers: 288 man-hours	※ Time of Completion for 80 servers: 1.5 man-hours
Find the MAC Address of Server		
Upgrade BIOS / BMC / RAID Firmware		
Configure BIOS / BMC / RAID / OS		
Check BIOS / BMC / RAID / OS		
Restore OS		
Configure OS		
Check Service Connectivity		
Delete Kitting VMkernel		



Typical IOD Deployment



IOD Stack



High Availability

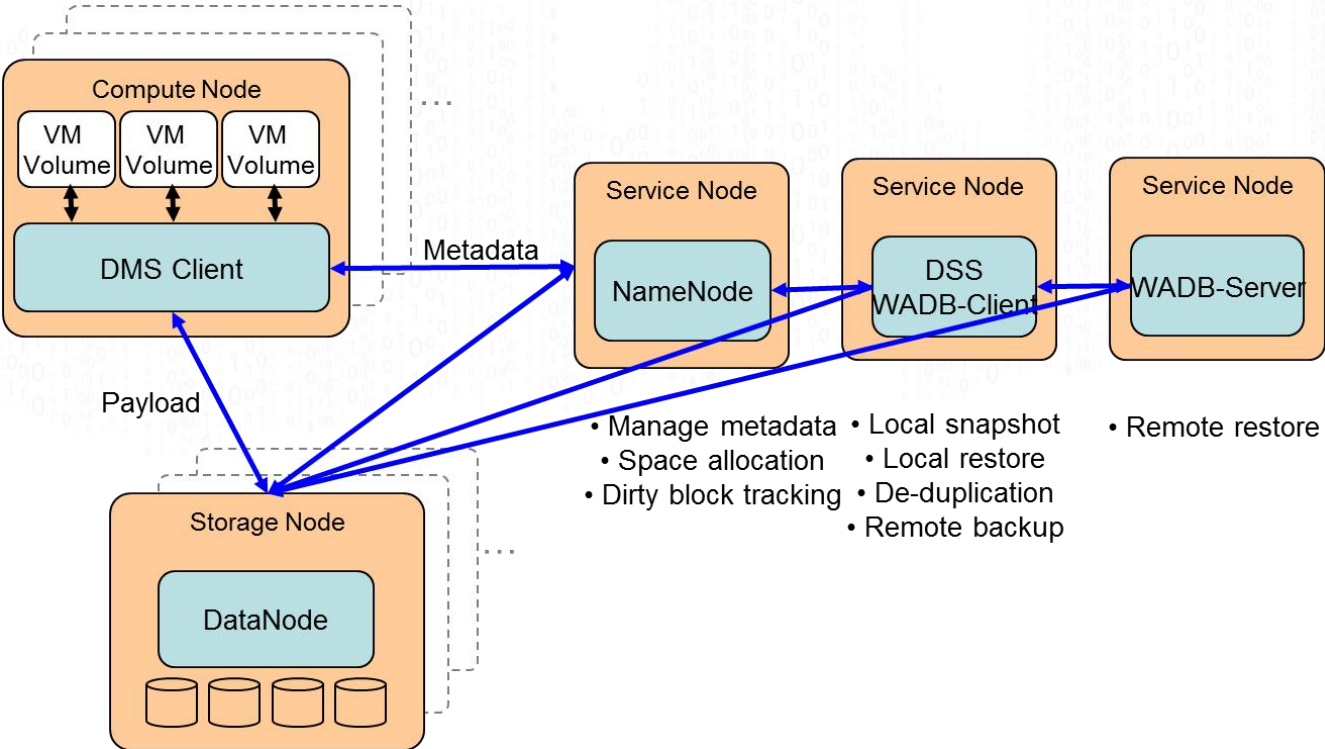
- Dual switch protection
- VM SDN: Peregrine redundant switch fast failover
- MySQL Galera cluster
- RabbitMQ server cluster
- API end points (Nova, Keystone, Glance,)
HA (Haproxy + Heartbeat)
- Multiple Agent instance (Nova, Keystone,)
- Neutron layer 3 HA



DISCO

Distributed I Integrated S Storage with C Comprehensive Data PrOtection

A storage abstraction on a large number of JBOD (just a bunch of disks) in storage servers



DISCO Characteristics

Thin provisioning

Just use what you need,
Physical space is
allocated dynamically for
better efficiency.

Transparent data protection

DISCO keeps your data
safe through its N-way
replication & self-healing
mechanisms.



HA support

Data integrity is always
preserved no matter what
disaster occurs.

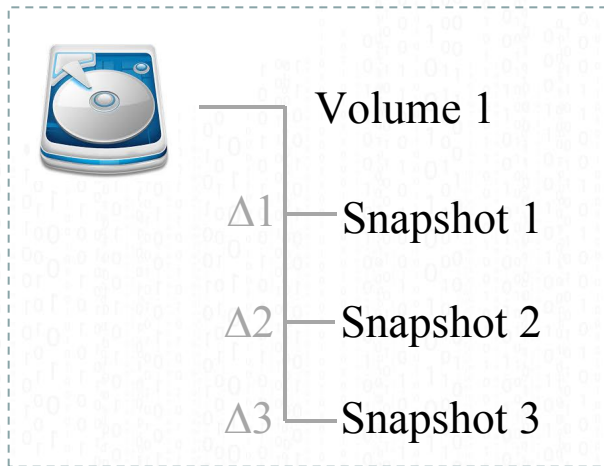
Fast volume cloning

No copy of metadata nor
data while cloning a
volume.

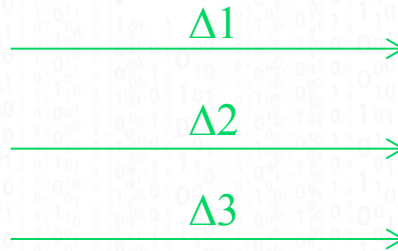
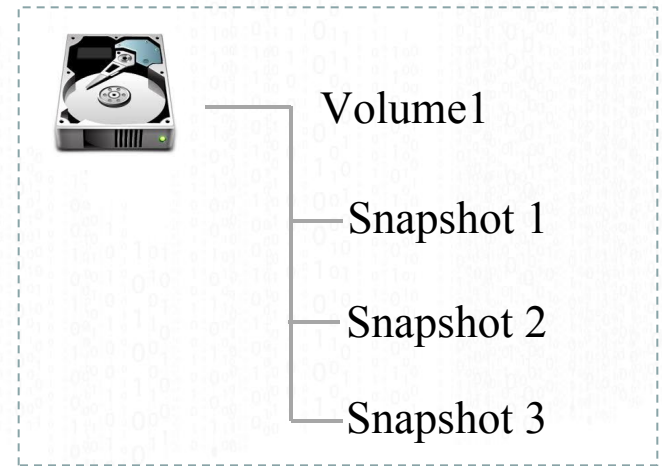



WADB – Wide Area Data Backup

Zone A (Ex: Taipei)



Zone B (Ex: HsinChu)

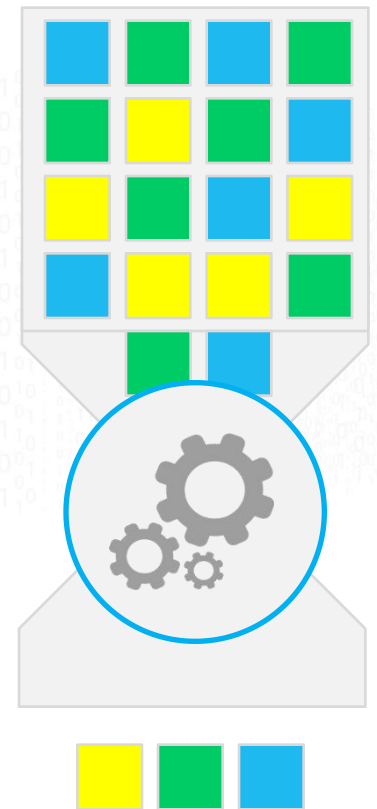


 Copy of the volume + its snapshots



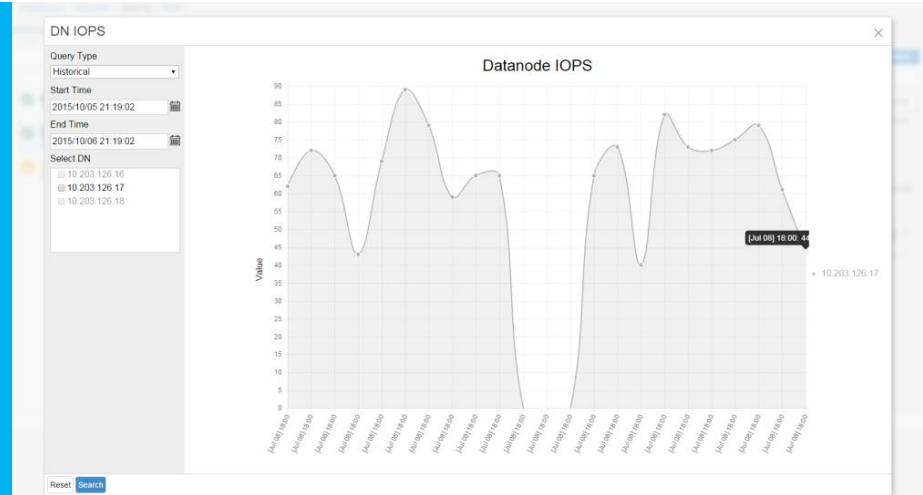
De-duplication

- Process the dirty blocks when taking the snapshot
- Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data
- Background process without performance impact



DISCO UI

Monitor service & hardware
Volume to component mapping
Component performance
OpenStack integration



Dashboard | Volumes | Service | RAID

System Status

- Cluster Health: OK
- Data Space Usage: 20.06 / 40.96 TB (48.97%)
- Dedup: 14.42 / 40.96 TB (35.21%)

Namesnode

Service Status	IP	HA Status	Space Usage	NN service rate	Actions
OK	10.203.126.11	OK	8.45 TB of 10 TB Used (84.5%)		Detail

DSS

Service Status	IP	HA Status	Space Usage	Actions
Degraded	10.203.126.13	OK	5.39 TB of 10 TB Used (53.9%)	Detail

Datanode

Service Status	Hardware	Space Usage	DN IOPS	Actions
94 OK 4 9	97 OK 8 0	3.8 TB of 10 TB Used (38%)		Detail

Clientnode

Service Status	# of Clientnodes	Queue activity to NN	Queue activity to DN	Actions
1 OK 2	3			Detail

WADB

Service Status	IP	Host Name	Actions
Degraded	140.110.7.8	Chris's Computer	Detail

Dashboard | Volumes | Service | RAID

Namesnode | Datanode | Clientnode | DSS | WADB

Summary

IP	HA Status	Master Server	Dedup Rate	Actions
10.203.126.13	OK	10.203.126.21	10.15 TB / 40.96 TB (24.78%)	

Detail

DSS	RAID	Disk	RAID Information
10.203.126.21 (Master) 8.73 TB of 10 TB Used (87.3%)	RAID-I-04 (2%)	disk01 (51 °C - 23err)	Name: RAID-P-05 Vendor: Proxmox VE Status: OK Battery: 14% Description: Hardware Raid, RAID5
10.203.126.22 (Slave) 5.06 TB of 10 TB Used (50.6%)	RAID-P-05 (14%)	disk02 (76 °C - 84err)	
		disk02 (76 °C - 84err)	
		disk04 (71 °C - 82err)	



Instances

Instance Name Filter Filter [Launch Instance](#) [Terminate Instances](#) More Actions

<input type="checkbox"/>	Instance Name	Image Name	IP Address	Size	Key Pair	Status	Availability Zone	Task	Power State	Time since created	Actions
<input type="checkbox"/>	ubuntu-demo-2	-	10.10.10.12 Floating IPs: 10.214.169.8	m1.medium	mykey	Active	nova	None	Running	2 hours, 2 minutes	Create Snapshot
<input type="checkbox"/>	ubuntu-demo-1	-	10.10.10.10 Floating IPs: 10.214.169.7	m1.medium	mykey	Active	nova	None	Running	2 hours, 13 minutes	Create Snapshot

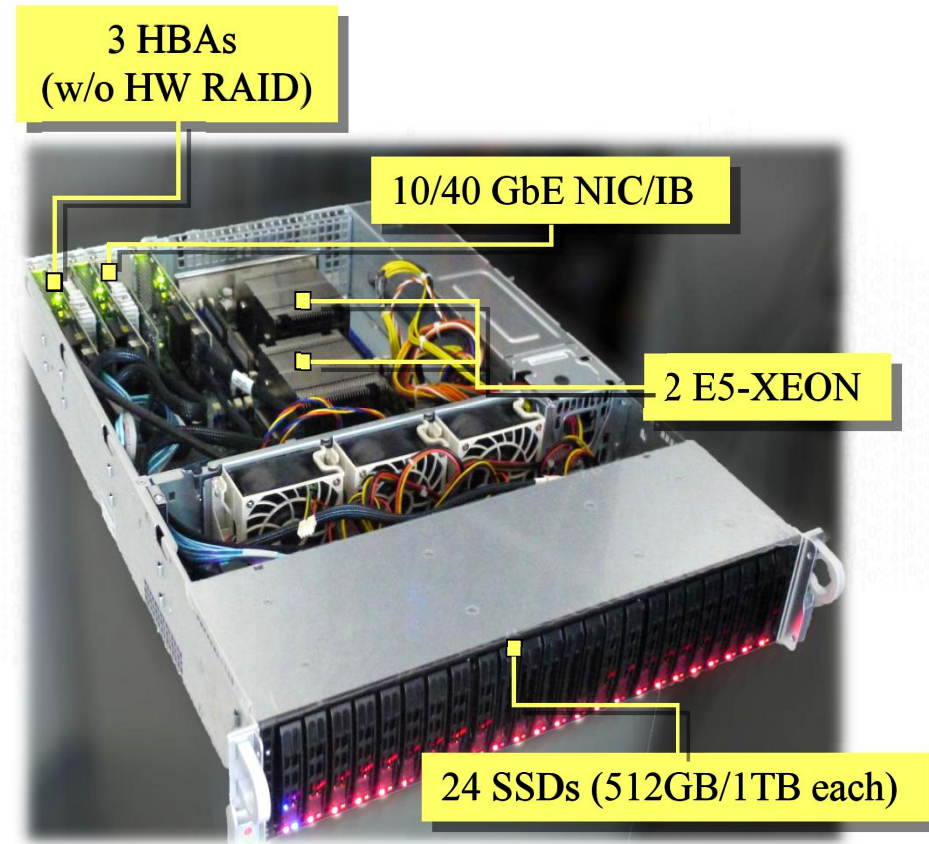
Displaying 2 items



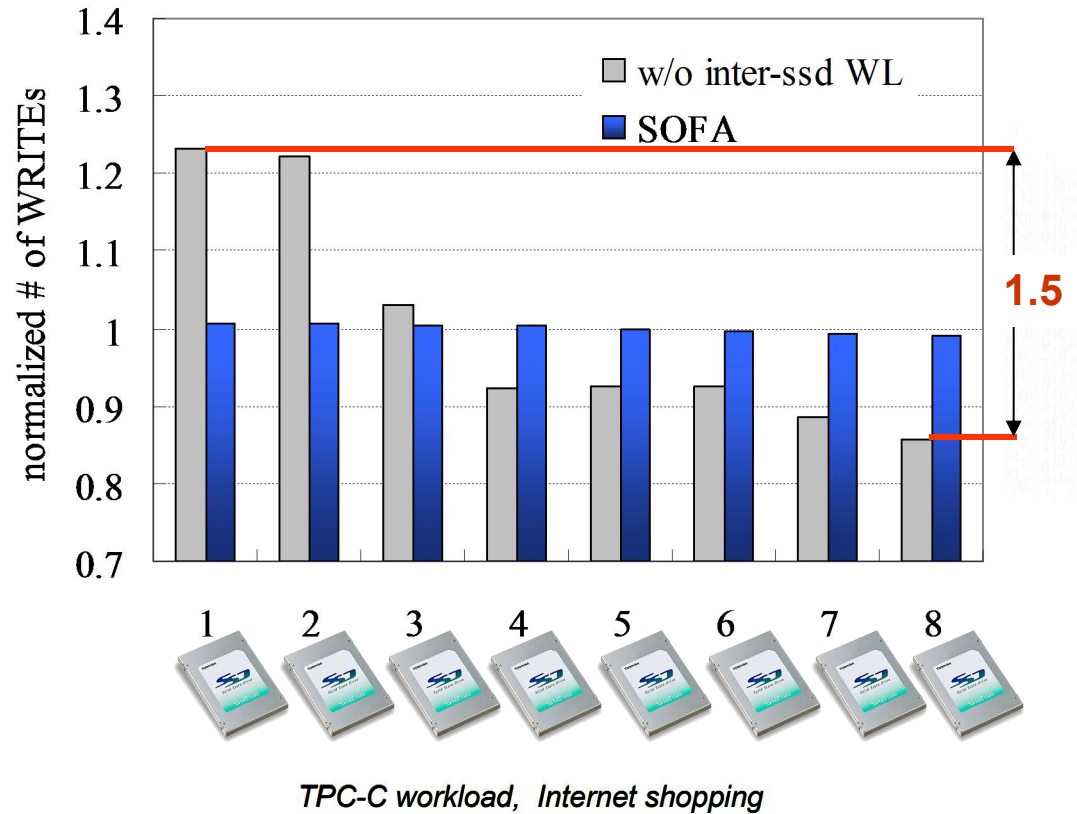
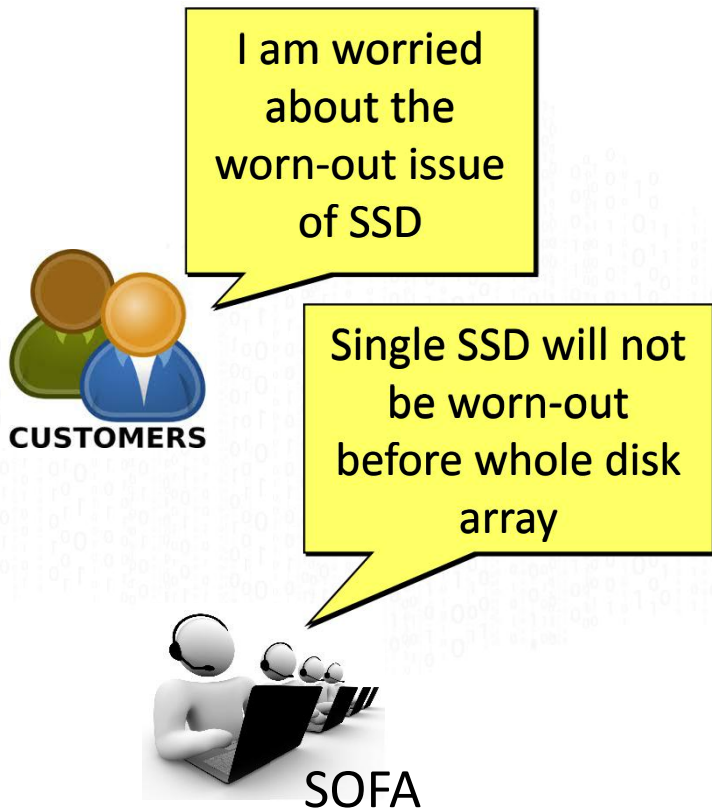
SOFA

- **Key Features :**

- Commodity hardware
- 1 M Random 4KB IOPS
- Proprietary RAID protection (w/o IOPS and lifetime penalty)
- Global hot spare for SSD failure
- Global Wear Leveling
- QoS : minimum IOPS guaranteed
- Fast Volume Clone
- Fast full snapshot and incremental snapshot
- Optimized network protocol
- Self-adaptive mechanism compatible with all kinds of platforms

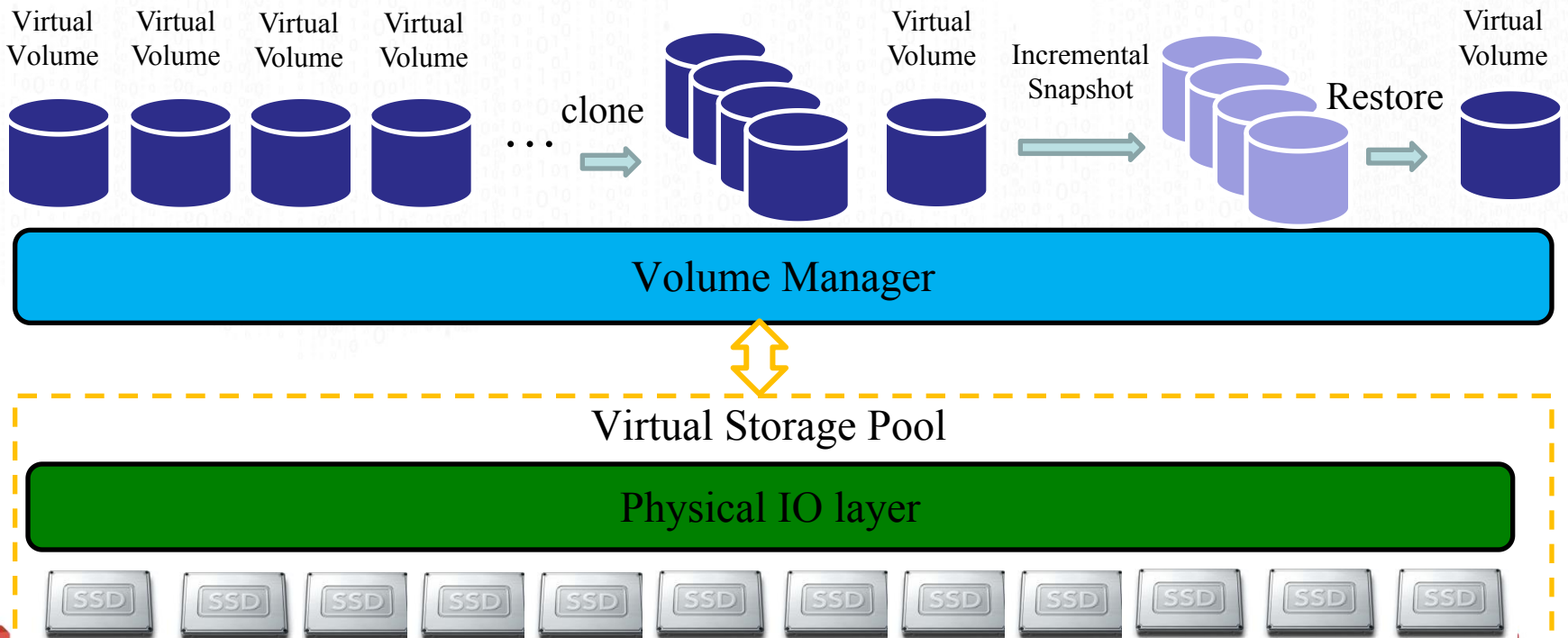


Global Wear Leveling



Volume Manager

- Main features
- Thin Provisioning
- Fast Clone Volume
- Incremental Snapshot



QoS

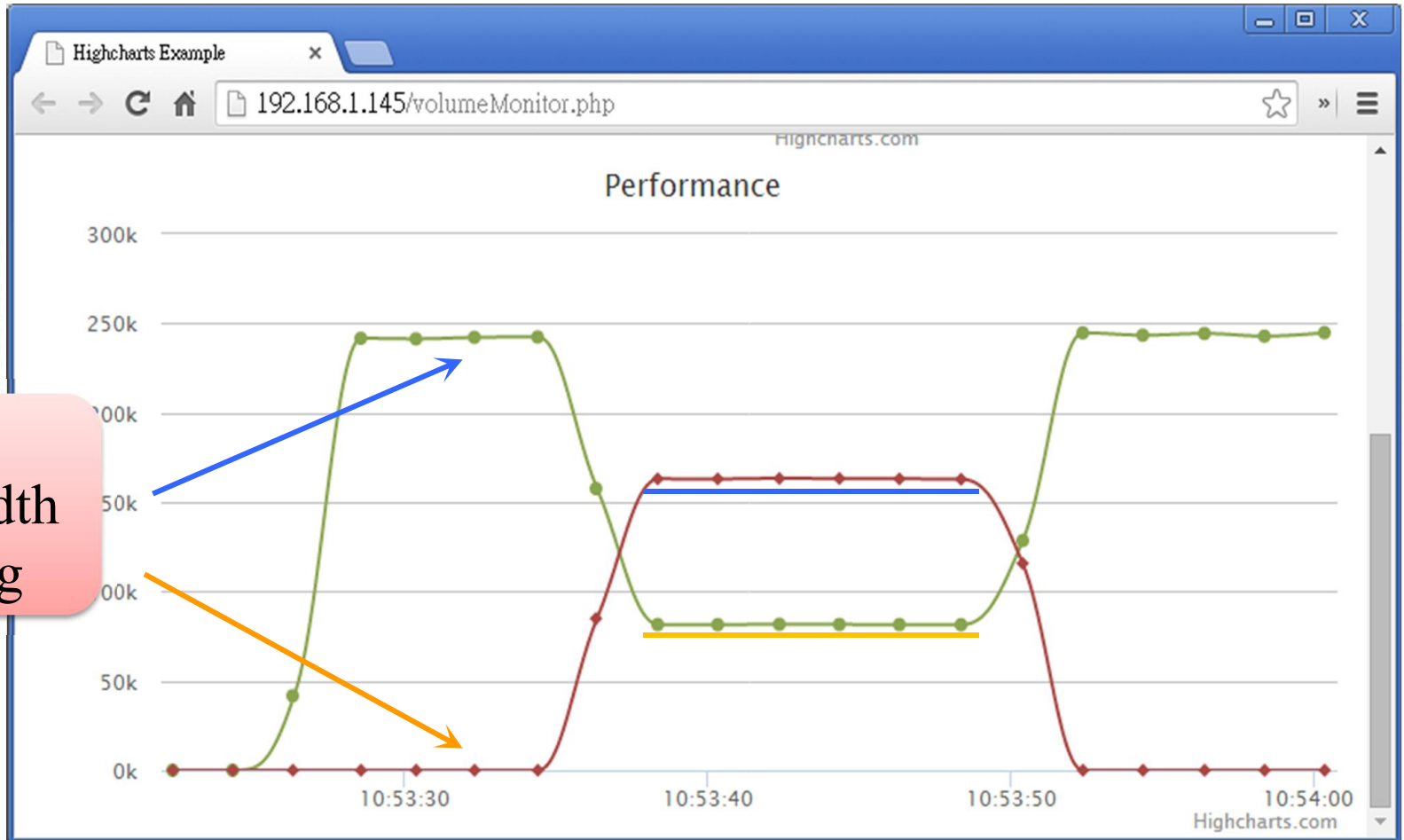
- Minimum IOPS guaranteed
- Maximum IOPS bound: for better pricing strategy

Minimum IOPS



QoS

- High utilization: Idle bandwidth sharing



Idle bandwidth sharing



1M 4KB Random IOPS

- 1 million 4KB random read / write IOPS

SOFA



```
[root@TEST3]# fio write.fio
```

```
rand-write: (g=0): rw=randwrite, bs=4K-4K/4K-4K, ioengine=libaio, iodepth=256000  
rand-write: (g=0): rw=randwrite, bs=4K-4K/4K-4K, ioengine=libaio, iodepth=256000  
Starting 2 processes  
rand-write: (groupid=0, jobs=2): err= 0: pid=3643
```

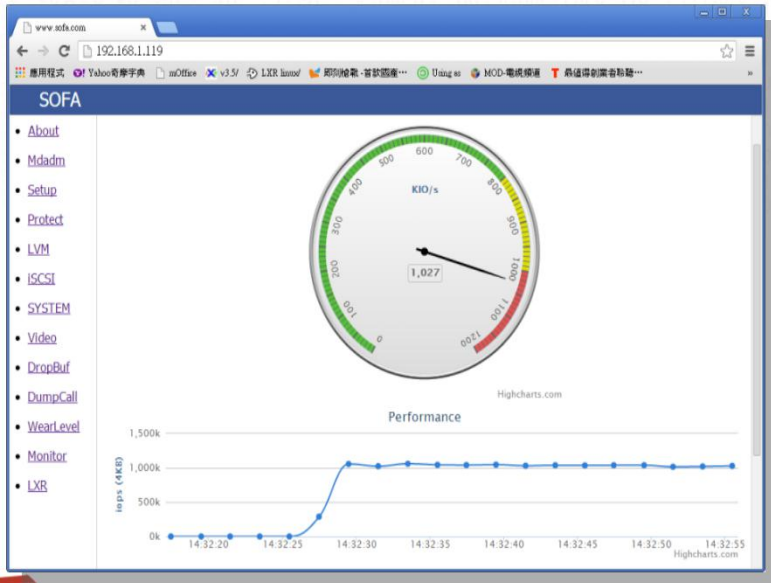
```
3, bw=4,084MB/s, iops=1,029K, runt= 30487msec  
%, sys=86.66%, ctx=361123, majf=0, minf=60  
=0/31358747, short=0/0
```

SRP @ Mellanox 40Gb InfiniBand

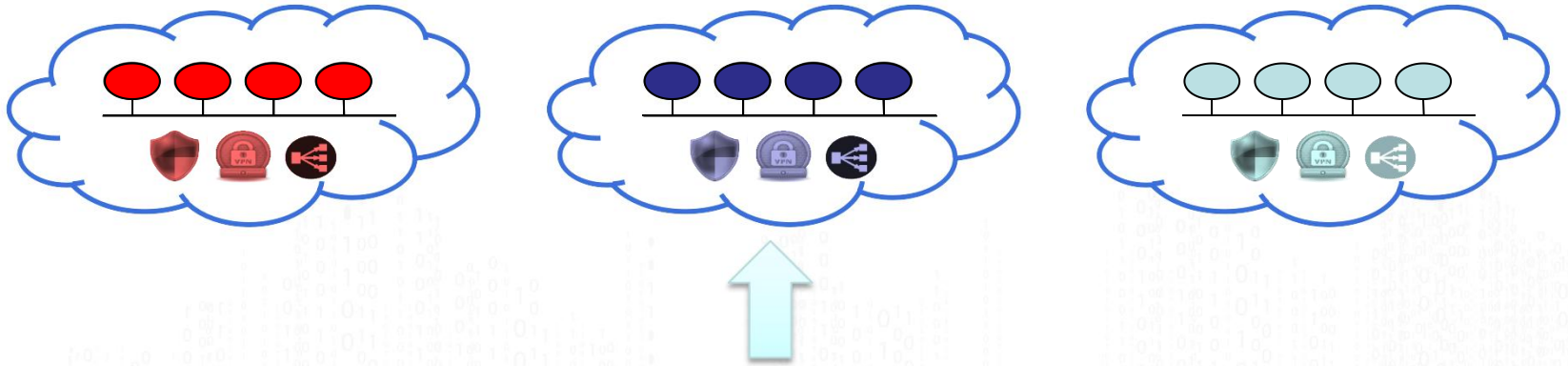
Fio - Flexible I/O Tester



SRP : SCSI RDMA Protocol



Peregrine



Peregrine hybrid SDN solution

ITRI contributes SNMP4SDN plugin to OpenDaylight, the plugin use SNMP and CLI to control Ethernet switches

Commodity Ethernet Switch

No vendor lock-in and no need to spend money in expensive hardware



Virtual OpenFlow Switch (OVS)

Provide powerful edge intelligence



Peregrine Characteristics

Commodity Ethernet Switch

Use OVS and Ethernet Switch provide SDN feature make it cost efficiency.

Traffic Engineering

Dynamically calculate the packet transmission path and balance the traffic load on each physical link.



Fast Failover

Pre-calculate backup path and immediately deploy it when error occurs.

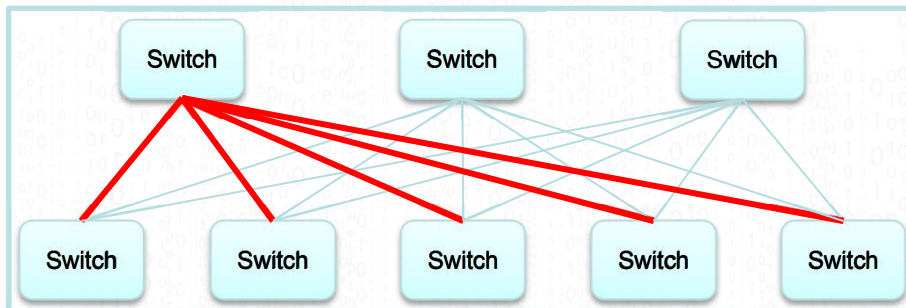
Diagnostic UI

Provide Physical / virtual topology and traffic load, VM traffic load and traffic analysis.

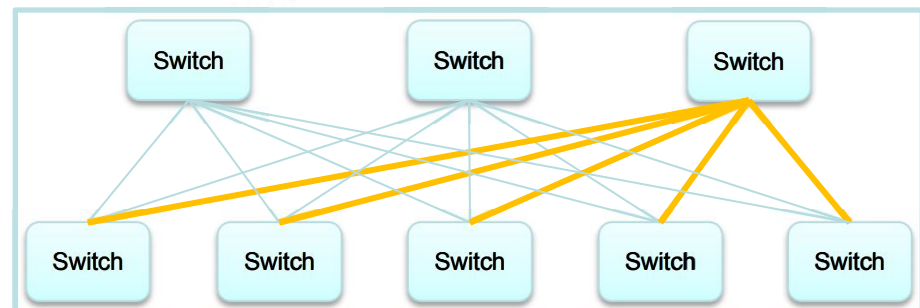
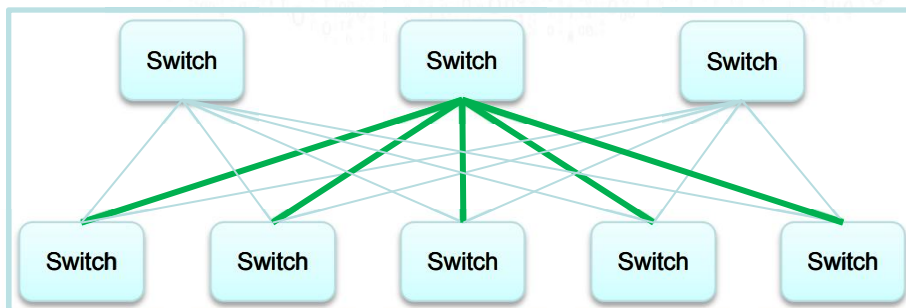


Traffic Optimization

- Peregrine is L2 fabric architecture and able to achieve optimal load-balanced of all the physical networks by dynamically calculates the packet transmission path.

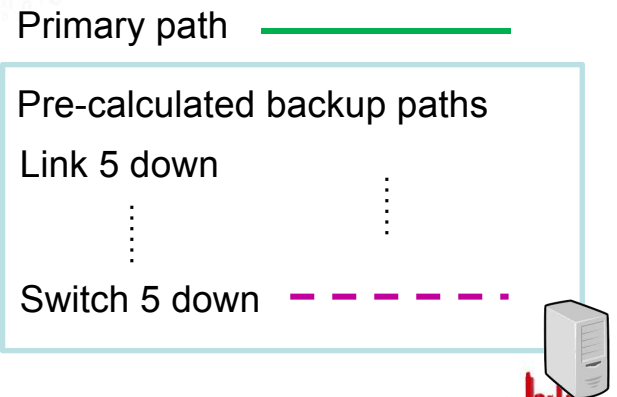
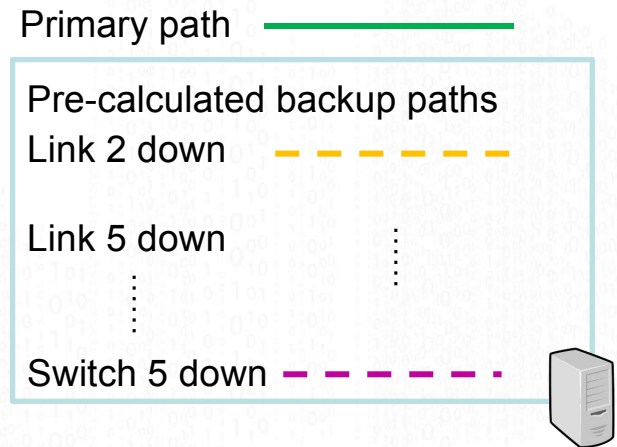
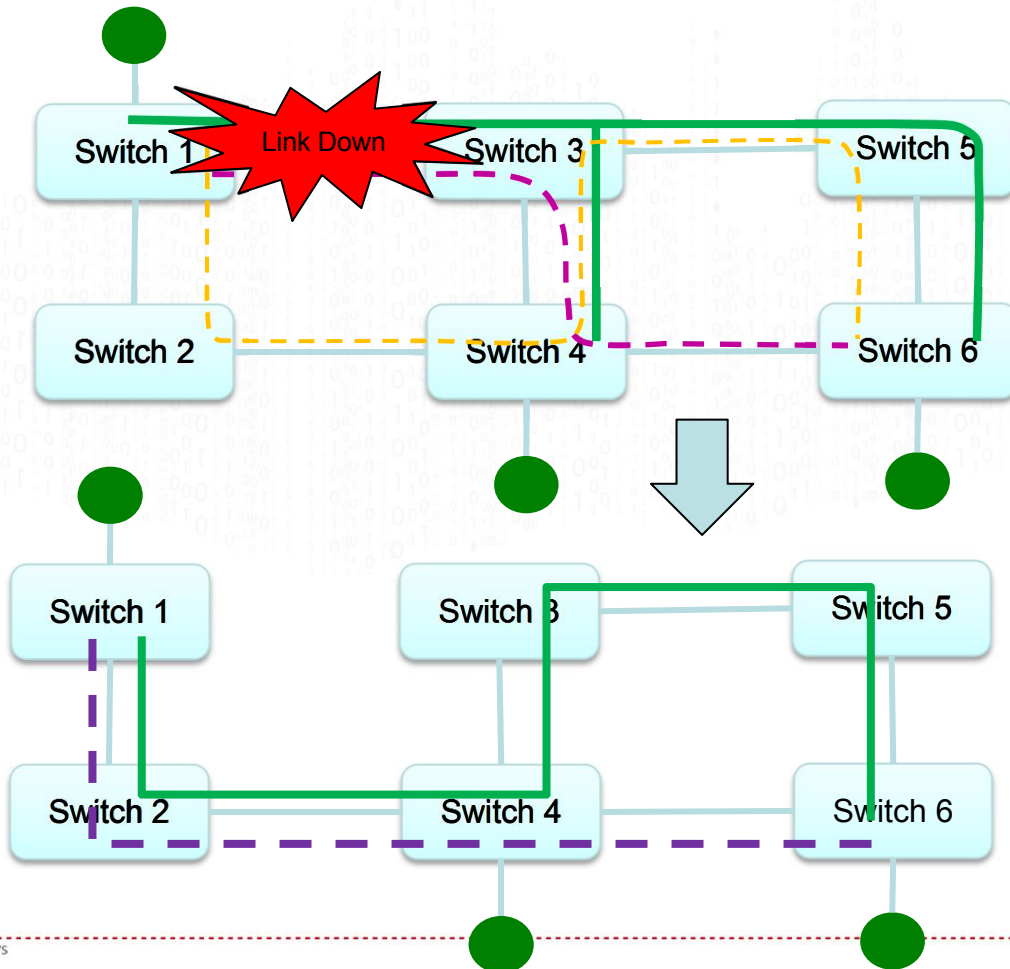


— VLAN: 10
— VLAN: 20
— VLAN: 30



Fast Failover

- Peregrine is able to re-deploy packet transmission path when any of link or device is failed by applying centralization control architecture in Fast Failover.



Peregrine UI

Physical & virtual topology
Physical & virtual traffic load
VM traffic analysis
User defined data path
OpenStack integration

ITRI OpenStack admin

Switch Network

Device Topology Trace

Links Virtual Networks VMs Packets

LINK 10.214.0.33 P15 - 10.214.0.32 P16 > VLAN 304 > VM 10.10.50.5 - 10.10.50.6

Time	Source	Destination	Protocol	Id	Flags	Seq	Ack	Win	Length	Other
05:34:28.624552	10.10.50.5.57959	10.10.50.6.5001							1470	UDP
05:34:28.624361	10.10.50.5.57959	10.10.50.6.5001							1470	UDP
05:34:28.624244	10.10.50.5.57959	10.10.50.6.5001							1470	UDP
05:34:28.624197	10.10.50.5.58228	10.10.50.6.5001			[]	146260950	1	229	66160	options [nop, nop, TS val: 605485415, ec: 605470130]
05:34:28.624055	10.10.50.5.57959	10.10.50.6.5001							1470	UDP
05:34:28.623901	10.10.50.5.57959	10.10.50.6.5001							1470	UDP
05:34:28.623615	10.10.50.5.57959	10.10.50.6.5001							1470	UDP
05:34:28.623513	10.10.50.5.57959	10.10.50.6.5001							1470	UDP
05:34:28.623460	10.10.50.5.58228	10.10.50.6.5001			[]	146195400	1	229	66160	options [nop, nop, TS val: 605485415, ec: 605470130]
05:34:28.623242	10.10.50.5.57959	10.10.50.6.5001							1470	UDP
05:34:28.623121	10.10.50.5.57959	10.10.50.6.5001							1470	UDP

ITRI OpenStack admin

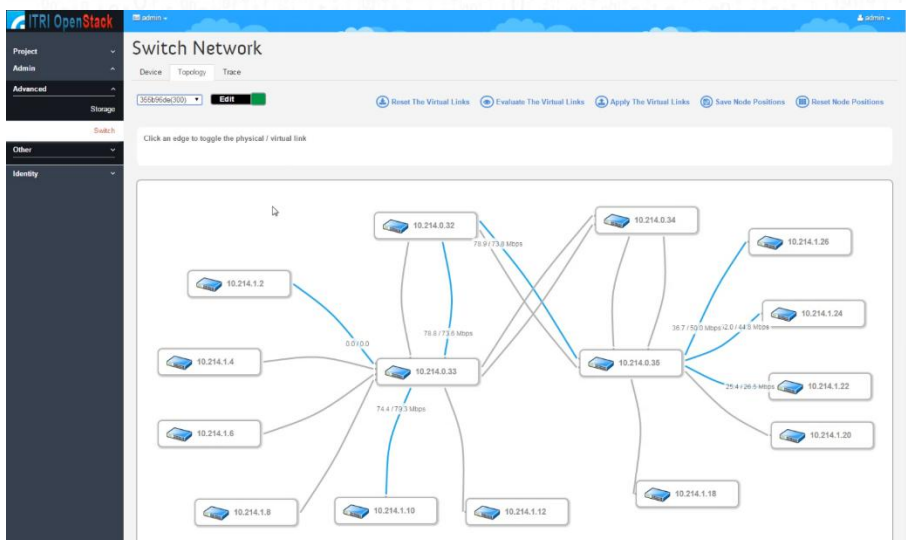
Switch Network

Device Topology Trace

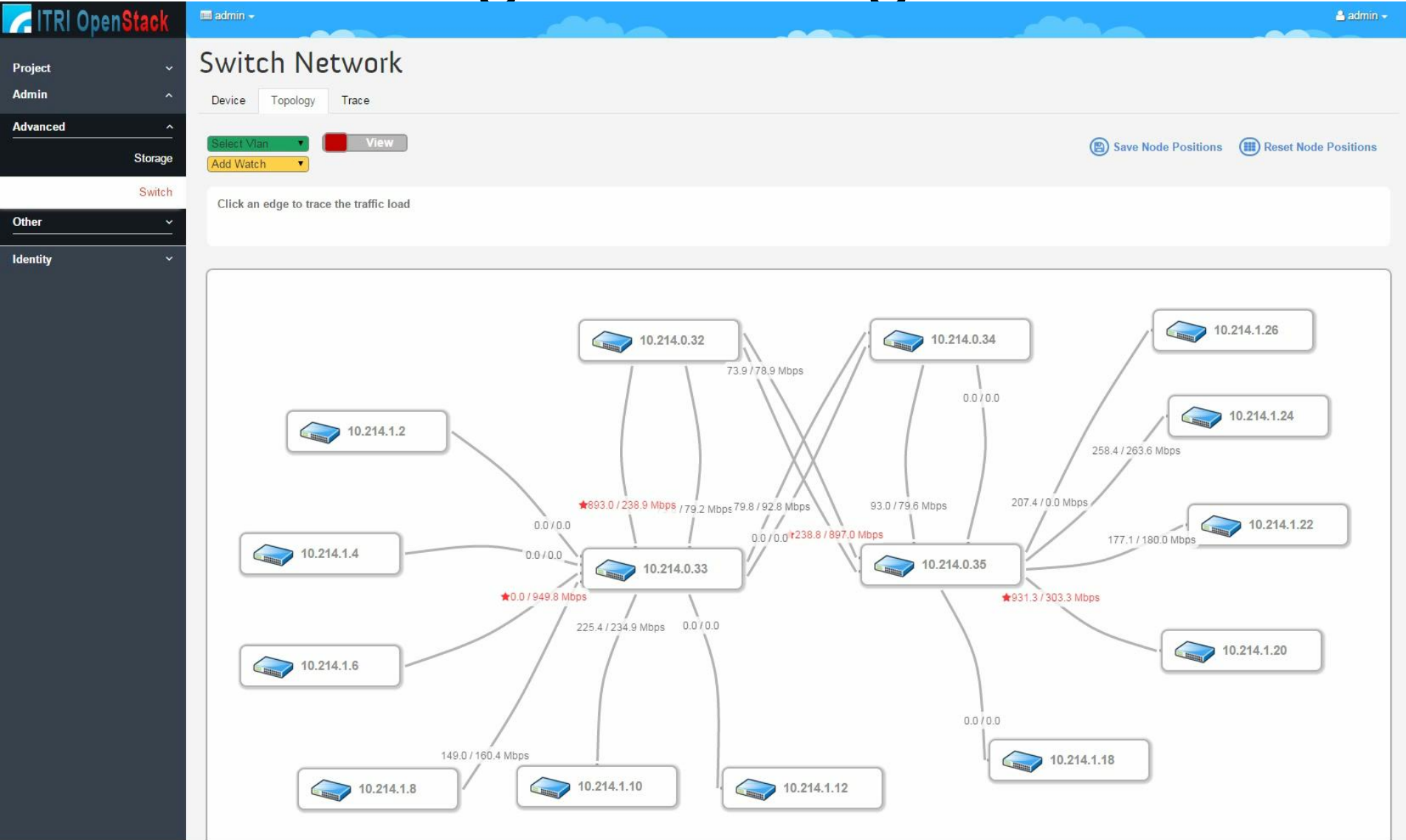
Links Virtual Networks VMs Packets

LINK 10.214.0.33 P15 - 10.214.0.32 P16 > VLAN 304

Links	Virtual Networks	VM Pairs
10.214.0.33 P15 - 10.214.0.32 P16	VLAN 304	10.10.50.5 - 10.10.50.6
10.214.0.32 P16 - 10.214.0.33 P15	VLAN 303	10.10.50.5 - 10.10.50.2
		10.10.50.5 - 10.10.50.4
		10.10.50.1 - 10.10.50.6
		10.10.50.1 - 10.10.50.2
		10.10.50.1 - 10.10.50.4
		10.10.50.3 - 10.10.50.6
		10.10.50.3 - 10.10.50.2
		10.10.50.3 - 10.10.50.4



Traffic Congestion Diagnosis Video



Link Failover Video

ITRI OpenStack

admin

Switch Network

Device Topology Trace

Select Vlan Edit

Reset The Virtual Links Evaluate The Virtual Links Apply The Virtual Links Save Node Positions Reset Node Positions

Click an edge to toggle the physical / virtual link

The diagram illustrates a network topology with several switches and their interconnections. The switches are represented by icons and labeled with IP addresses. The connections between switches are shown as lines, with bandwidth usage data displayed on each link. The switches and their connections are as follows:

- Switch 10.214.0.32 is connected to 10.214.0.34 (73.8 / 78.8 Mbps) and 10.214.0.33 (0.0 / 0.0 Mbps).
- Switch 10.214.0.34 is connected to 10.214.0.33 (92.8 / 79.6 Mbps) and 10.214.0.35 (91.0 / 127.1 Mbps).
- Switch 10.214.0.33 is connected to 10.214.0.35 (121.2 / 127.4 Mbps) and 10.214.1.18 (0.0 / 0.0 Mbps).
- Switch 10.214.0.35 is connected to 10.214.1.18 (0.0 / 0.0 Mbps).
- Switch 10.214.1.2 is connected to 10.214.1.4 (0.0 / 0.0 Mbps).
- Switch 10.214.1.4 is connected to 10.214.1.6 (0.0 / 0.0 Mbps).
- Switch 10.214.1.6 is connected to 10.214.1.8 (332.7 / 314.1 Mbps).
- Switch 10.214.1.8 is connected to 10.214.1.10 (148.9 / 160.3 Mbps).
- Switch 10.214.1.10 is connected to 10.214.1.12 (212.1 / 234.5 Mbps).
- Switch 10.214.1.12 is connected to 10.214.1.18 (0.0 / 0.0 Mbps).
- Switch 10.214.1.26 is connected to 10.214.1.24 (217.2 / 217.0 Mbps).
- Switch 10.214.1.24 is connected to 10.214.1.22 (7.7 / 248.9 Mbps).
- Switch 10.214.1.22 is connected to 10.214.1.20 (274.6 / 167.2 Mbps).



PDCM

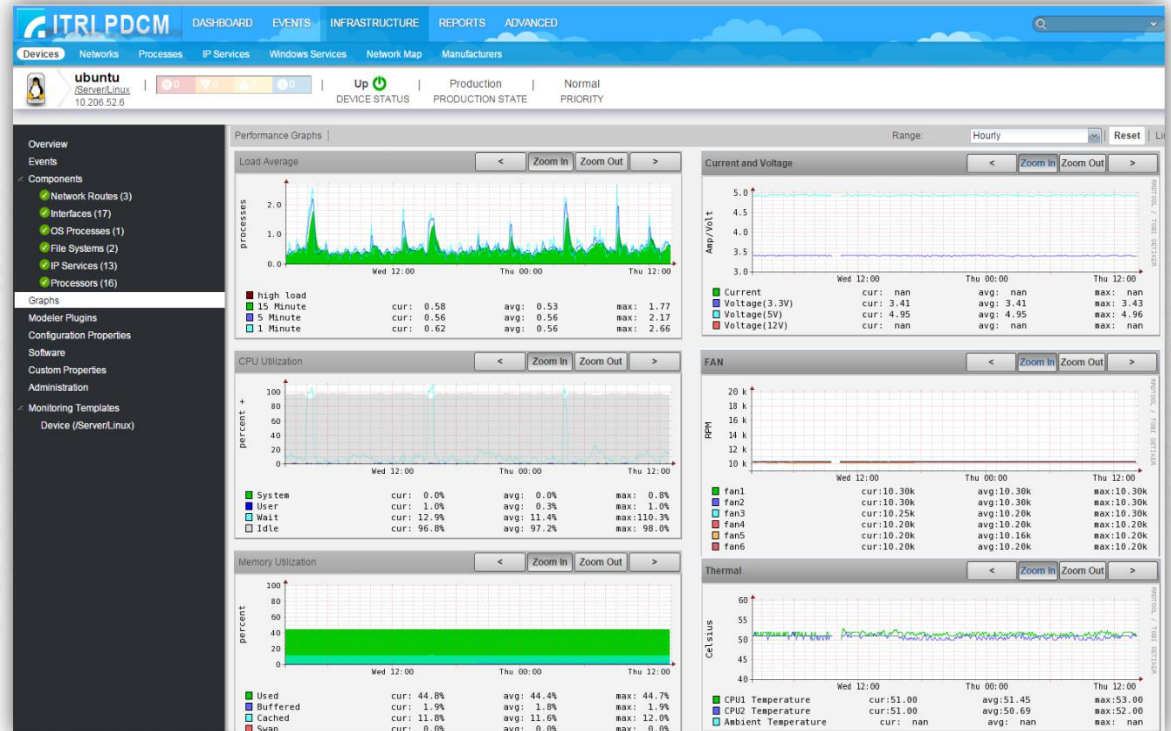
- PDCM stands for **Physical Data Center Management**.
- It is a hardware monitor system and a service management system.
- **Features:**
 - ✓ Health monitoring of physical devices
 - ✓ Health monitoring of OpenStack system components
 - ✓ Traffic load and resource usage reporting
 - ✓ Event and alerting system

PDCM provides a comprehensive solution for monitoring OpenStack cloud, including hardware devices and OpenStack services.



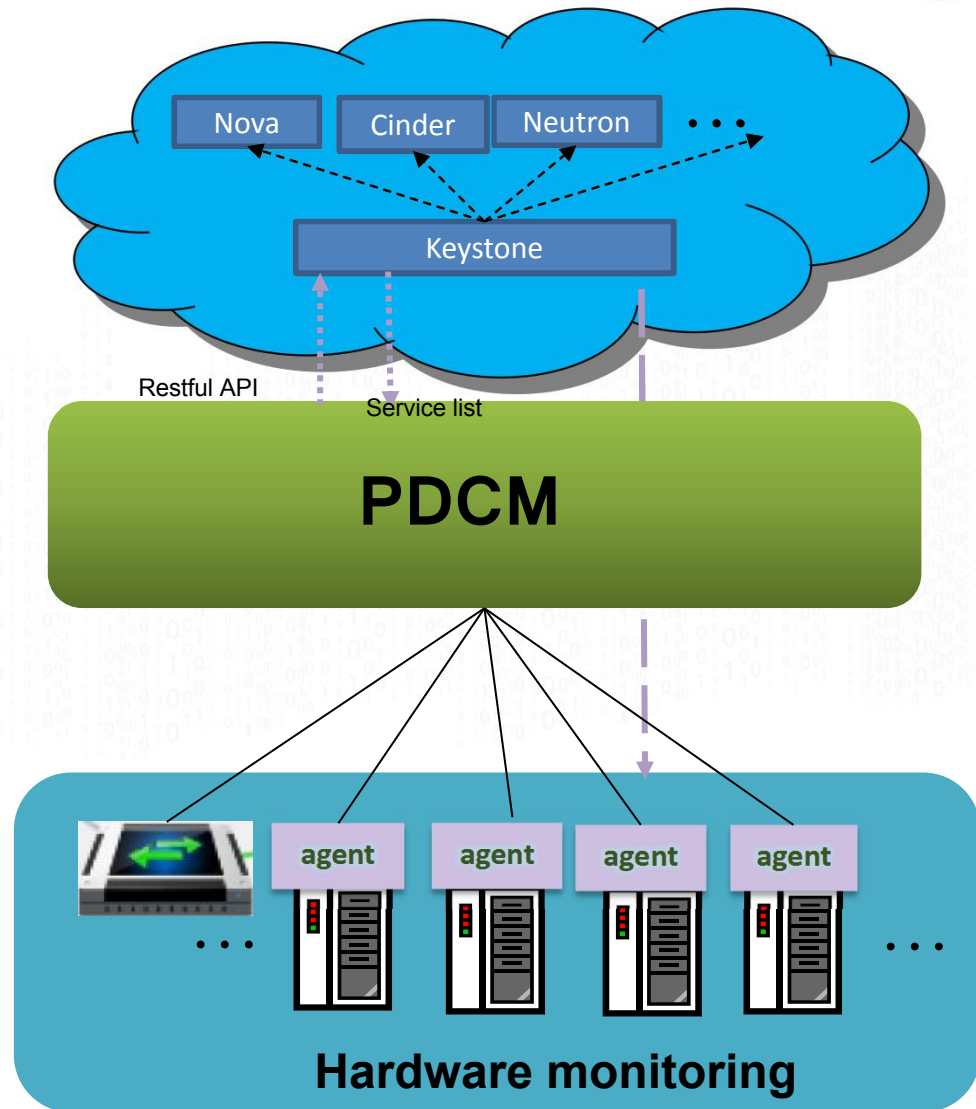
Device Hardware Monitoring

- ✓ CPU Utilization
- ✓ Memory Utilization
- ✓ Power Usage
- ✓ Network Routes
- ✓ Interfaces
- ✓ File Systems
- ✓ Current and Voltage
- ✓ Fans
- ✓ Thermal
- ✓ Hard Disk
- ✓ Raid Card
- ✓ ...



OpenStack Services Monitoring

- ✓ Nova Services
- ✓ Neutron Agents
- ✓ Cinder Services
- ✓ Regions
- ✓ Availability Zones
- ✓ Instances
- ✓ Hosts
- ✓ Hypervisors
- ✓ Flavors
- ✓ Images
- ✓ Networks
- ✓ Subnets
- ✓ Routers
- ✓ Ports
- ✓ Floating IPs.
- ✓ PM-VM mapping



Thanks

