



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



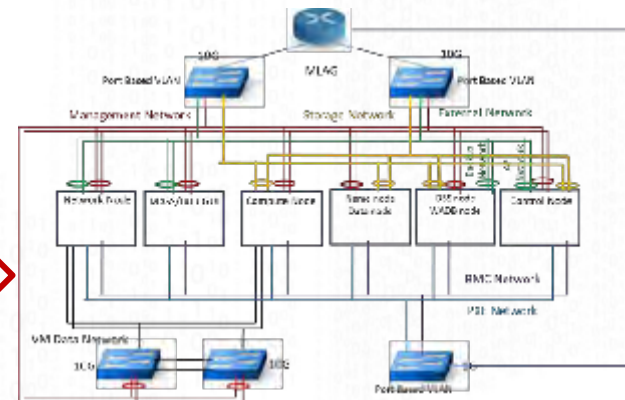
# IOD Deployment Procedure



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.



Designing The Future



	Manpower	BAMPI
Initialize BMC Network	※ Time of Completion for 80 servers: <b>288 man-hours</b>	※ Time of Completion for 80 servers: <b>1.5 man-hours</b>
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		

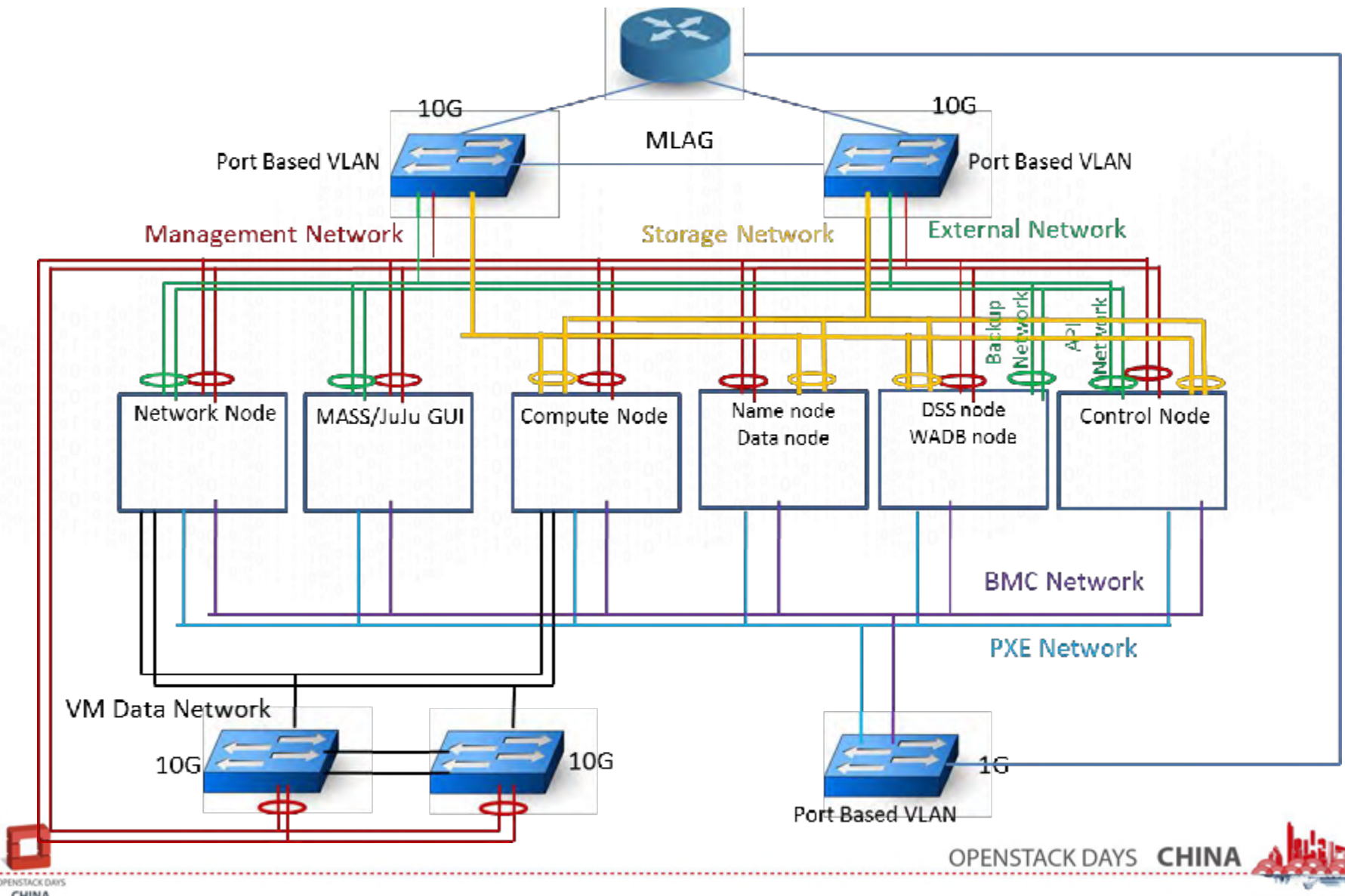


OPENSTACK DAYS  
CHINA

OPENSTACK DAYS CHINA

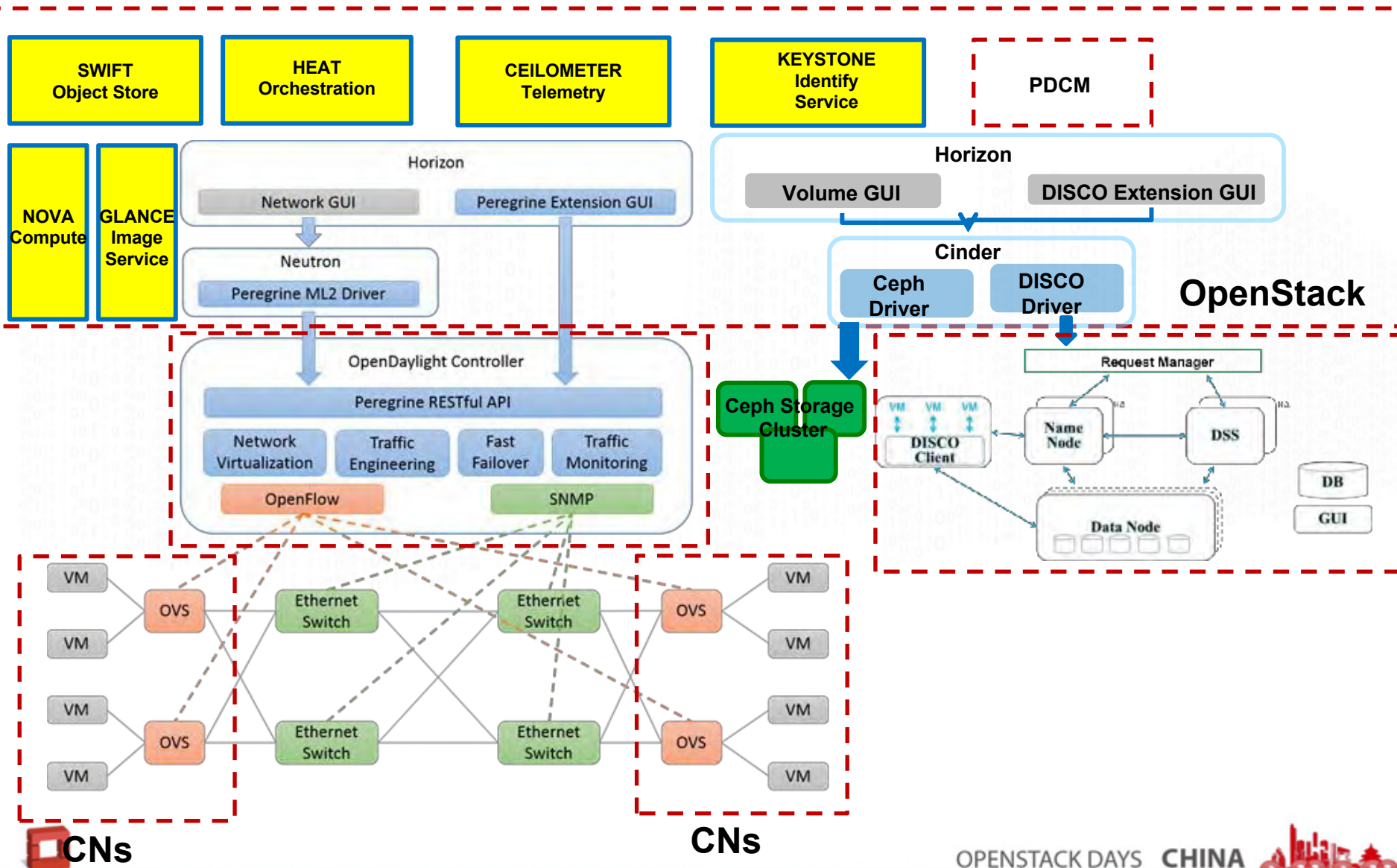


# Typical IOD Deployment





# IOD Stack



# High Availability

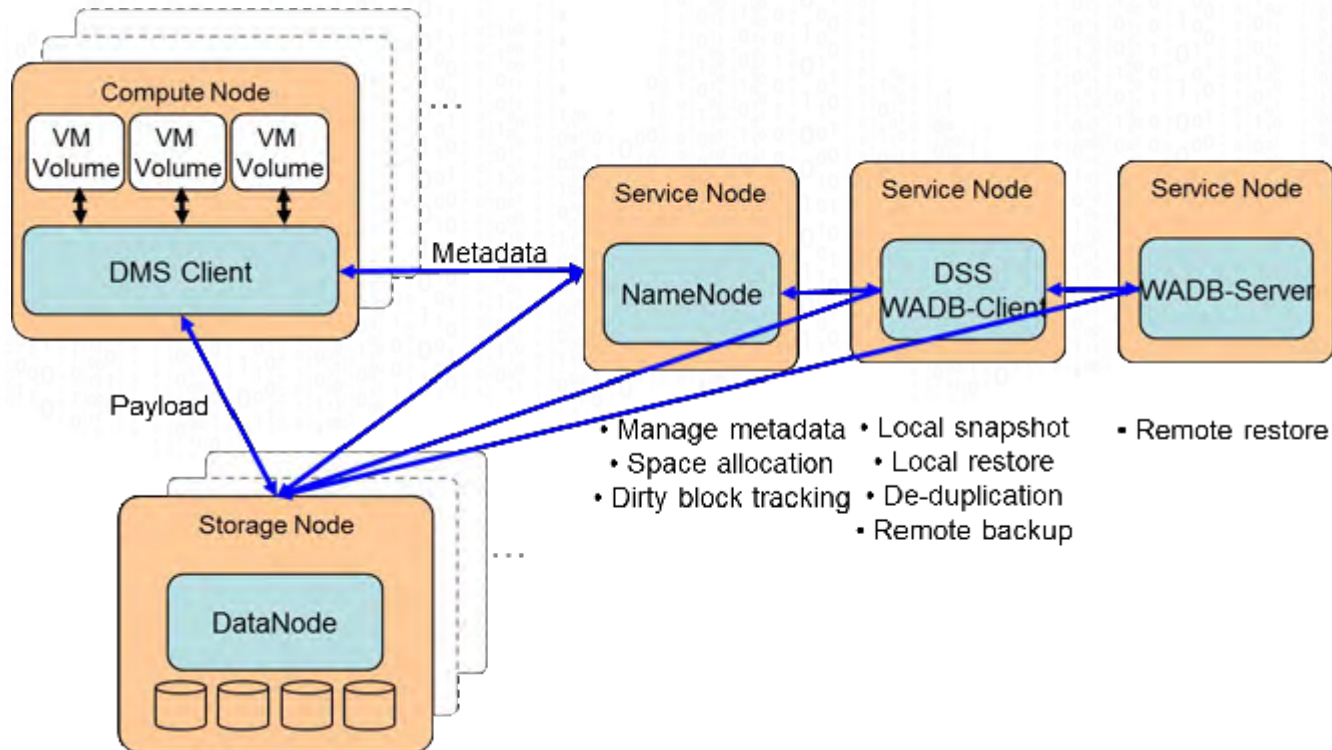
- Dual switch protection
- VM SDN: Peregrine redundant switch fast failover
- MySQL Galera cluster
- RabbitMQ server cluster
- API end points (Nova, Keystone, Glance, ....)  
HA (Haproxy + Heartbeat)
- Multiple Agent instance (Nova, Keystone, ....)
- Neutron layer 3 HA



# DISCO

Distributed I Integrated S Storage with C Comprehensive Data Pr Otection

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.

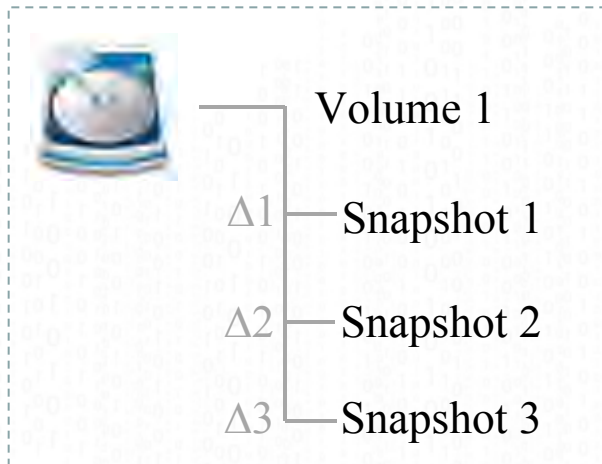
## Fast volume cloning

No copy of metadata nor  
data while cloning a  
volume.

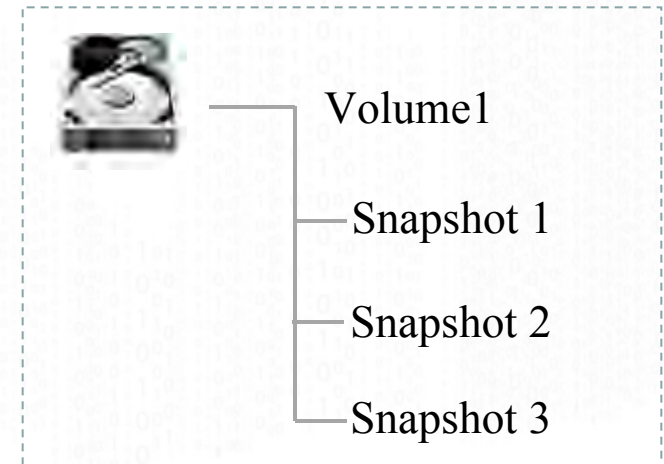


# WADB – Wide Area Data Backup

Zone A (Ex: Taipei)



Zone B (Ex: HsinChu)

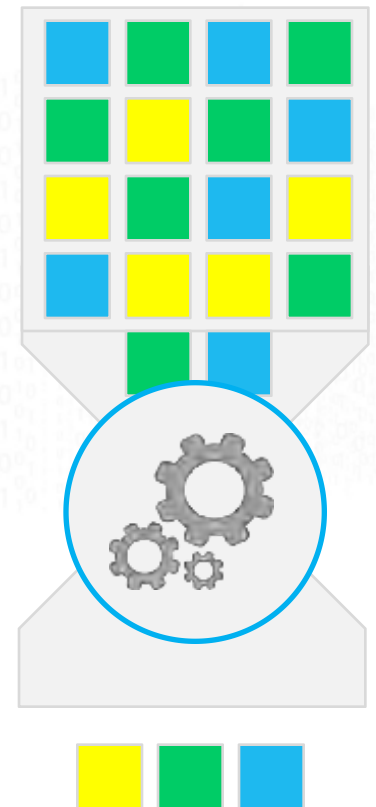


➡ Copy of the volume + its snapshots



# De-duplication

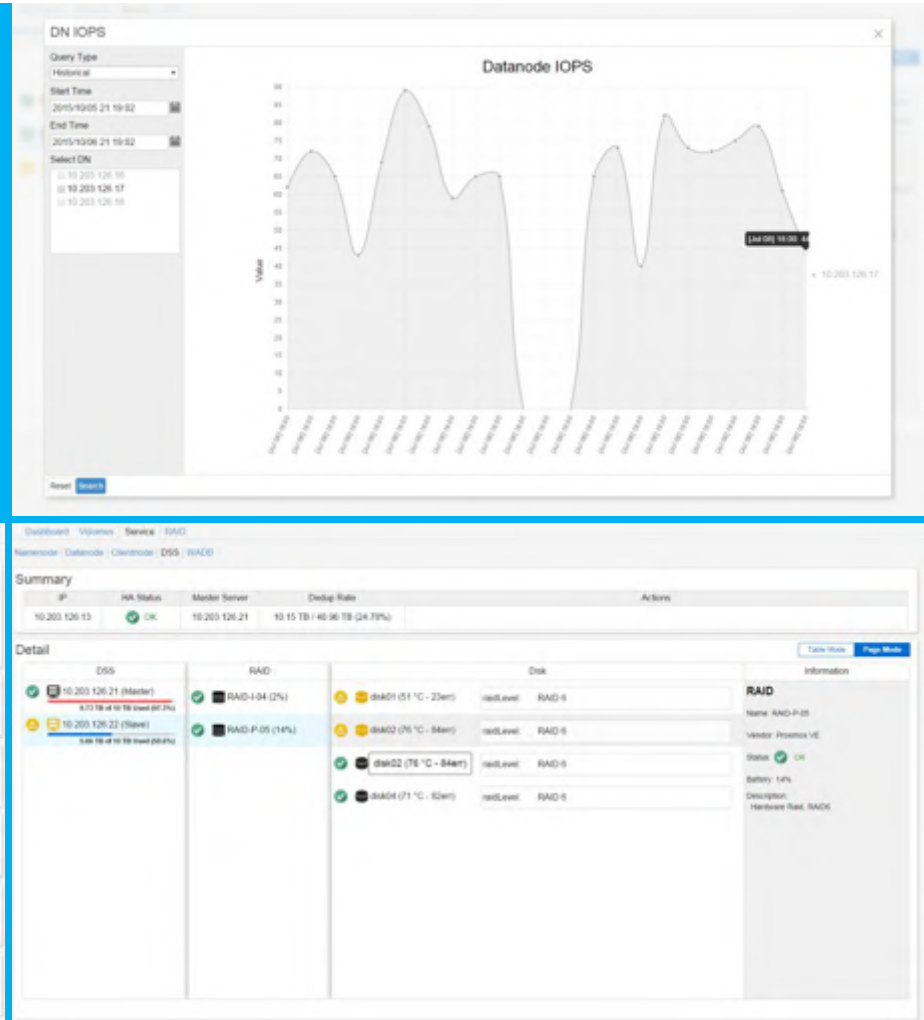
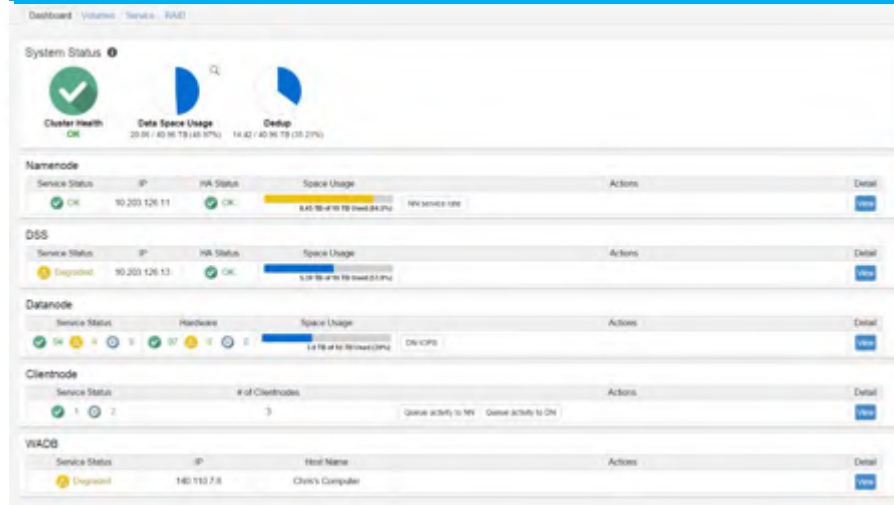
- Process the dirty blocks when taking the snapshot
- Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data
- Background process without performance impact





# DISCO UI

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



Instances - OpenStack D: X

10.214.128.4/horizon/project/instances/

应用程式 Bookmarks Dev ITRI GAGU [生活] 应用程式 SHAMBLES 哈佛大学公开课: ...

ITRI OpenStack admin

# Instances

Instance Name 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

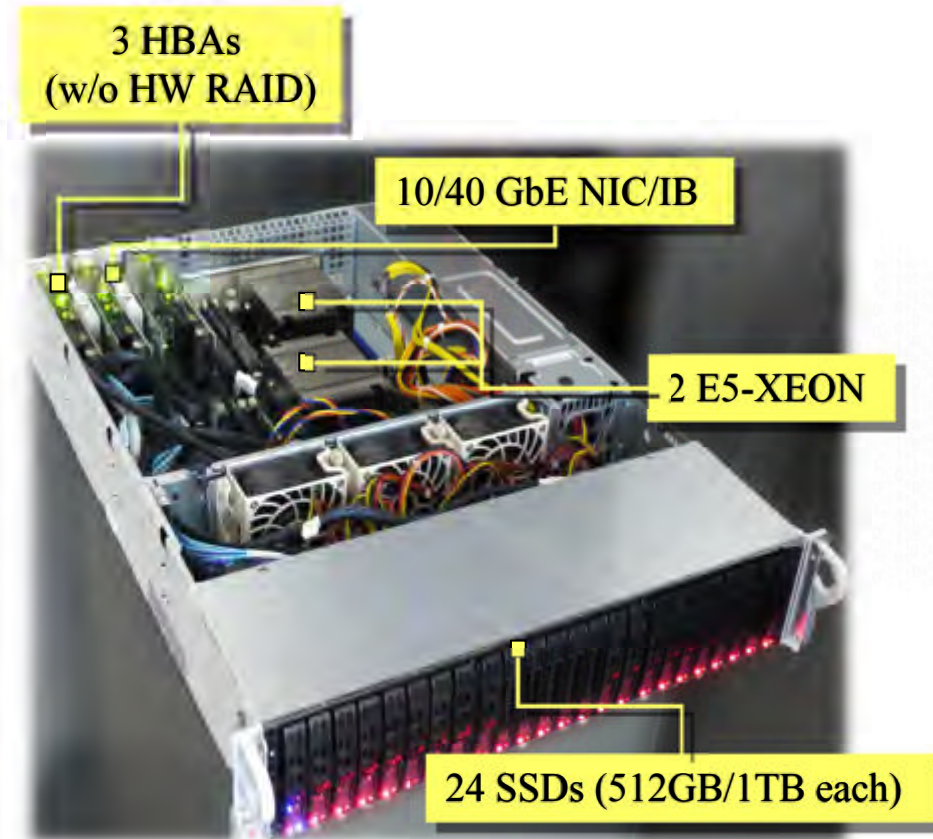
Displaying 2 items



# 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

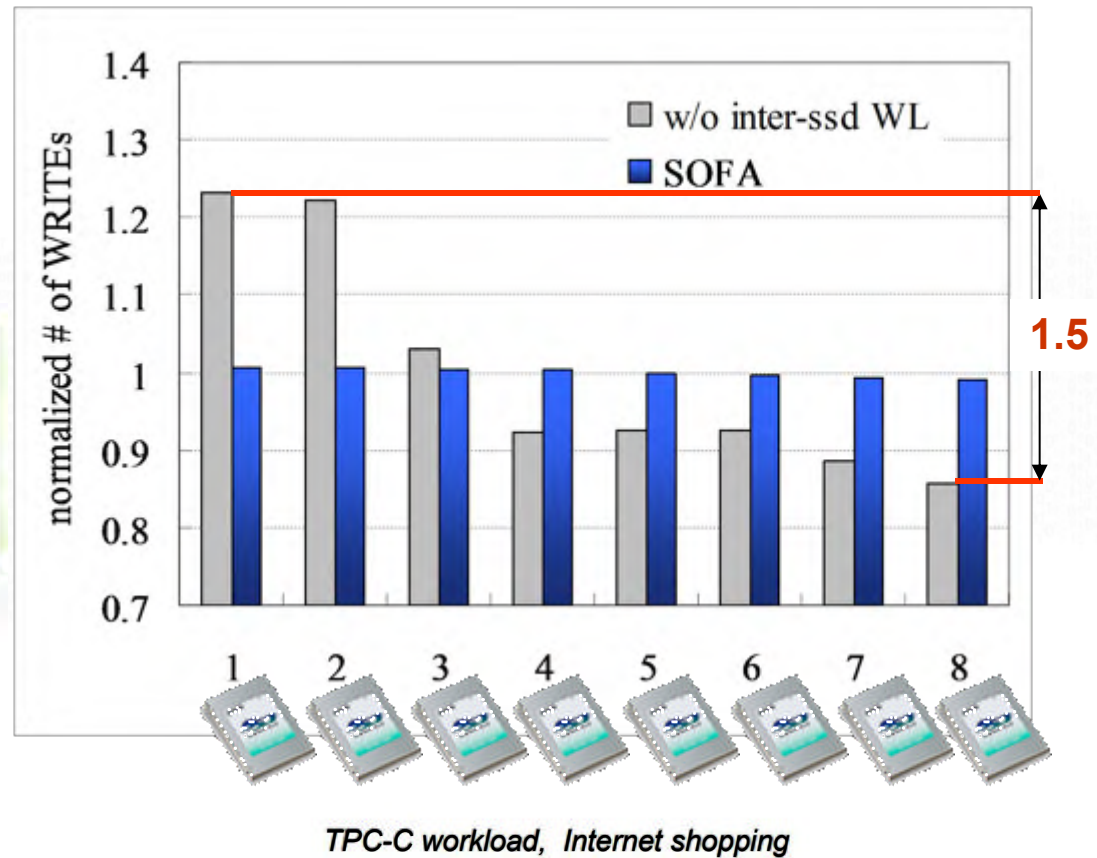


I am worried about the worn-out issue of SSD

Single SSD will not be worn-out before whole disk array

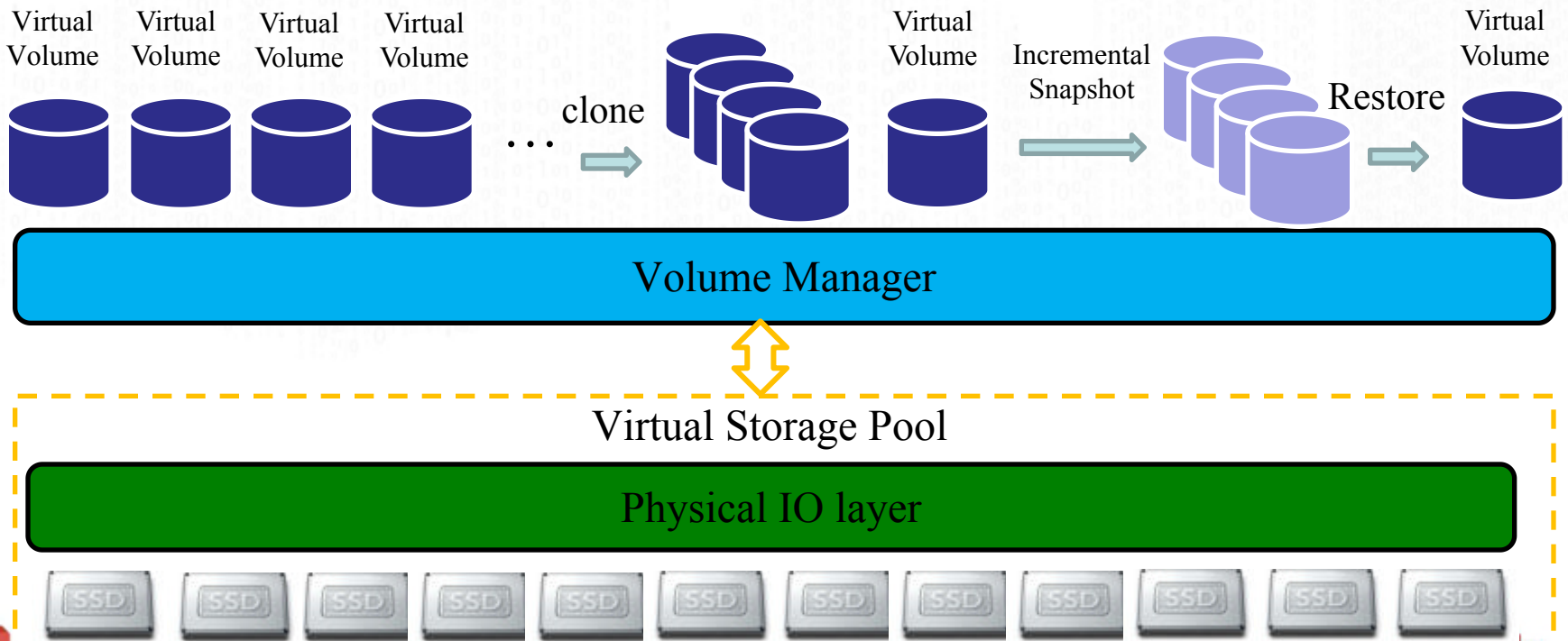


SOFA



# Volume Manager

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





# QoS

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

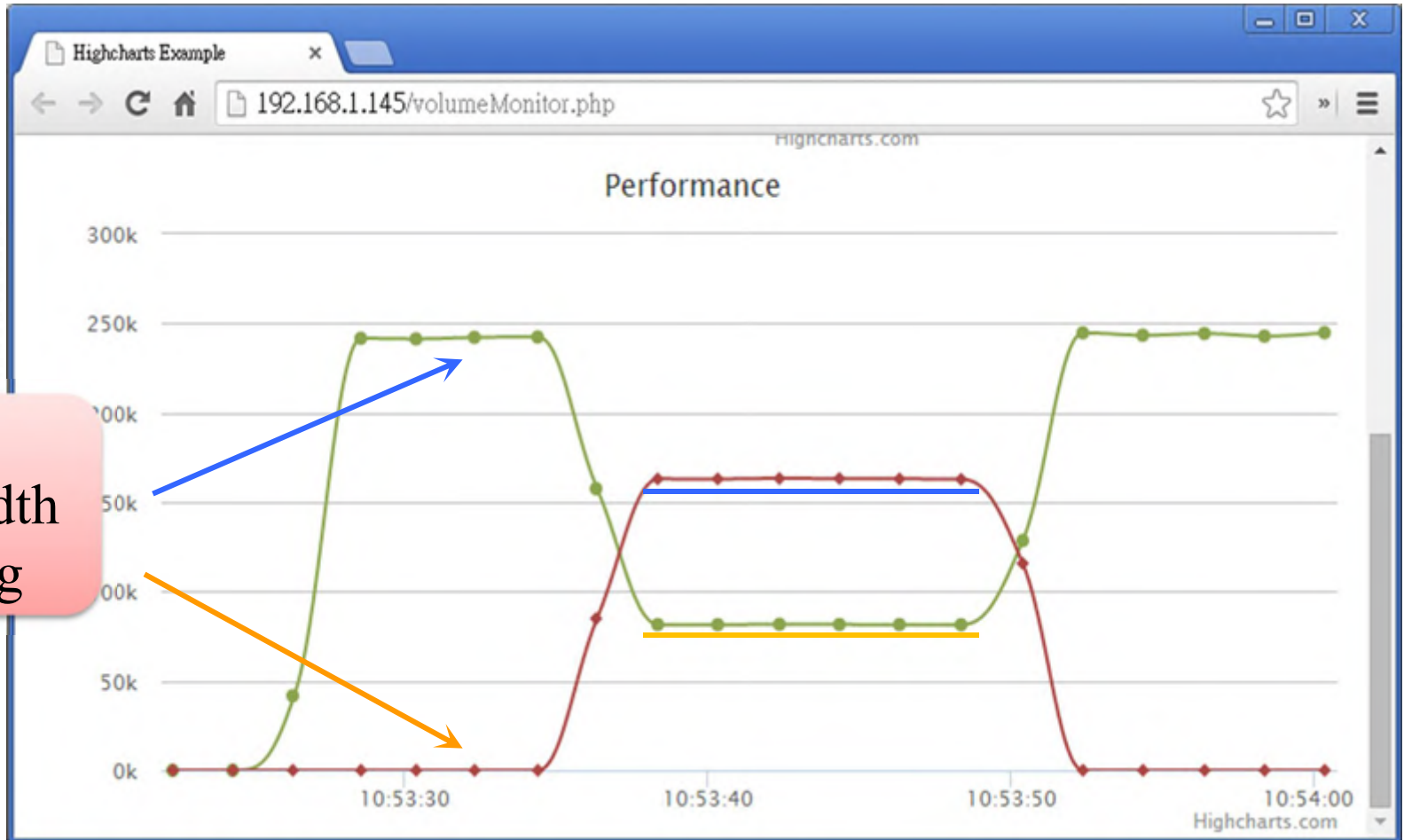
Minimum  
IOPS





# QoS

- High utilization: Idle bandwidth sharing



# 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



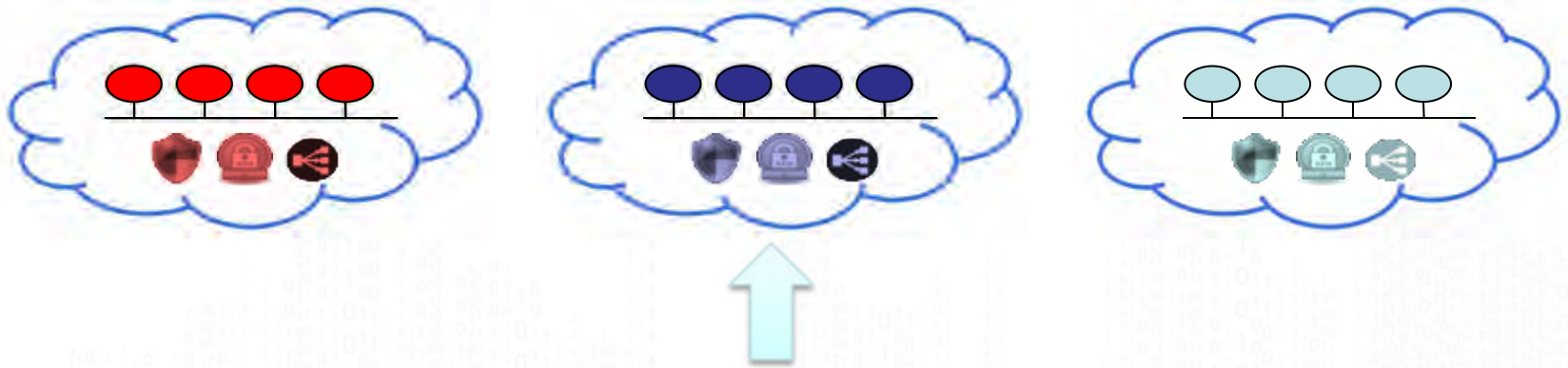
SRP : SCSI RDMA Protocol

OPENSTACK DAYS CHINA



OPENSTACK DAYS  
CHINA

# 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.

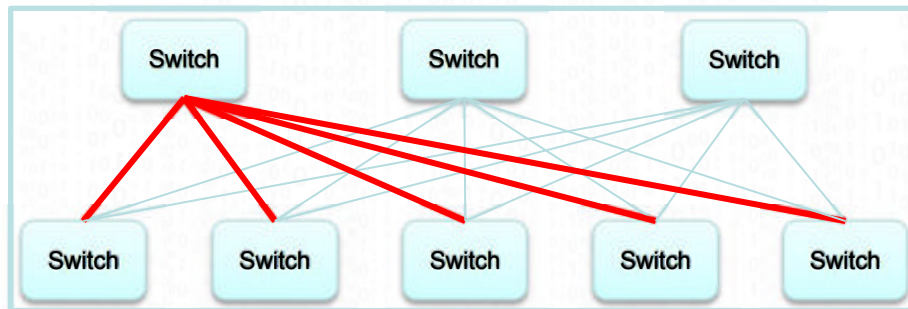
## Diagnostic UI

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

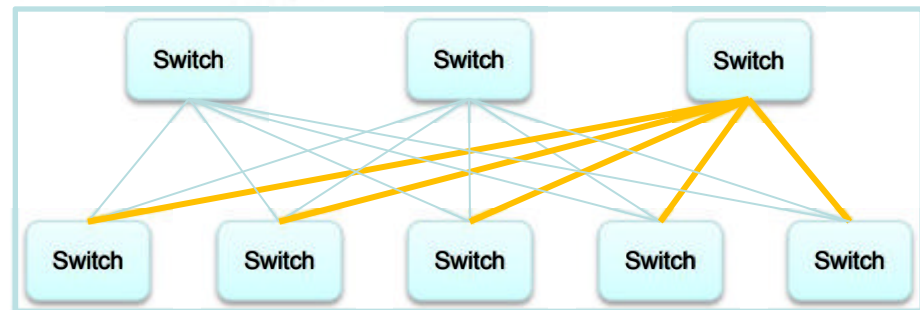
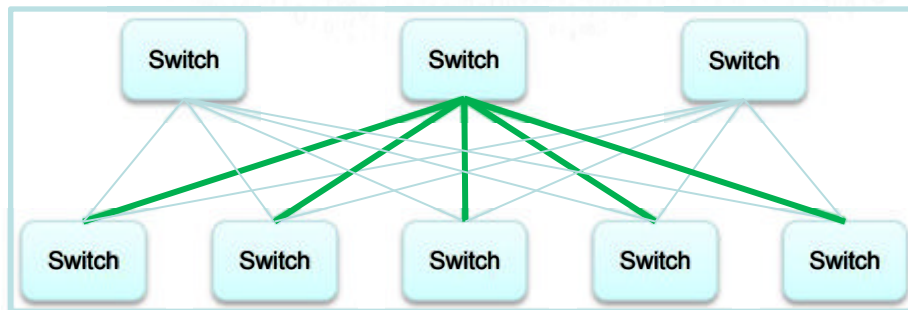


# Traffic Optimization

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



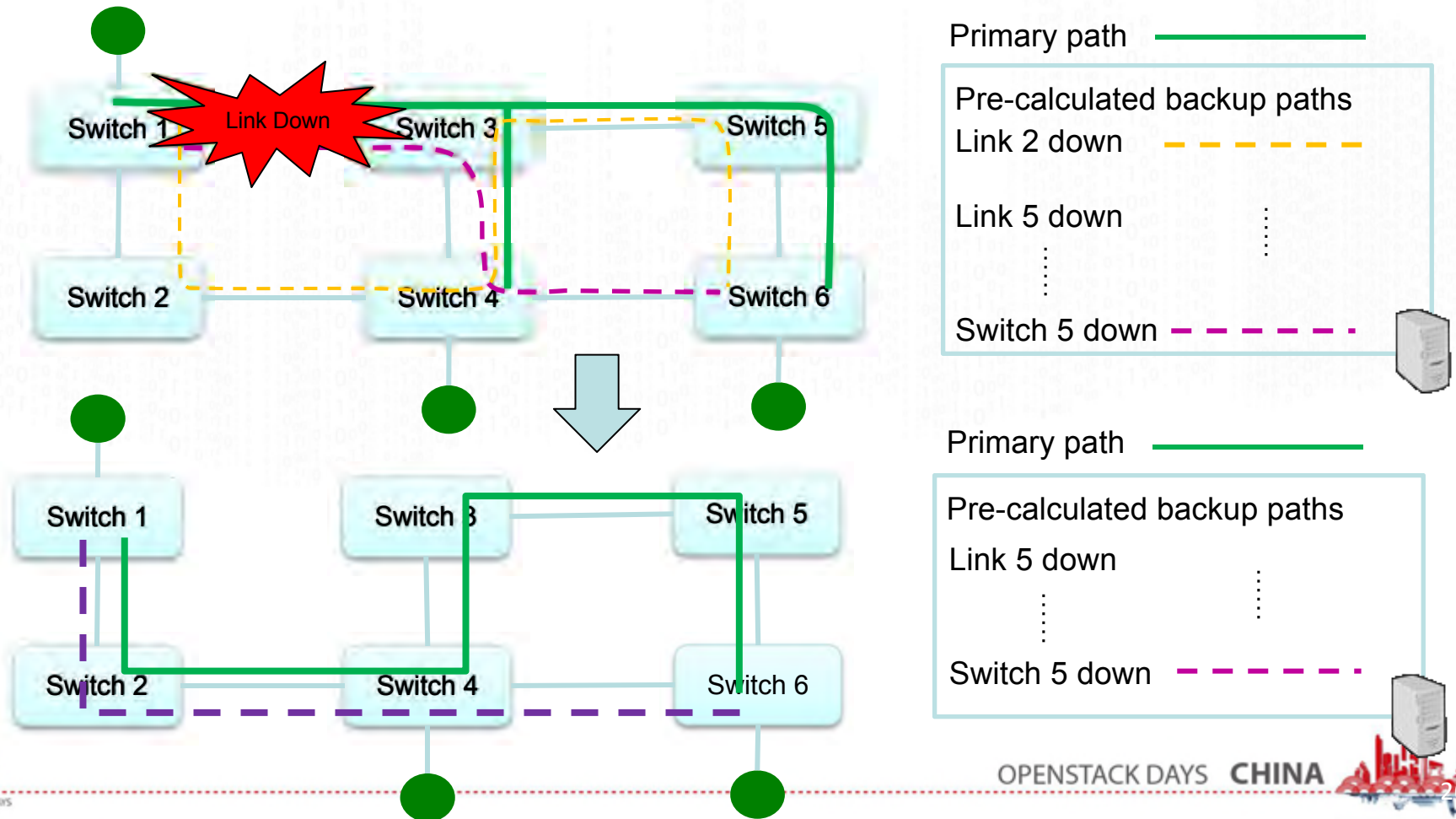
— VLAN: 10  
— VLAN: 20  
— VLAN: 30





# Fast Failover

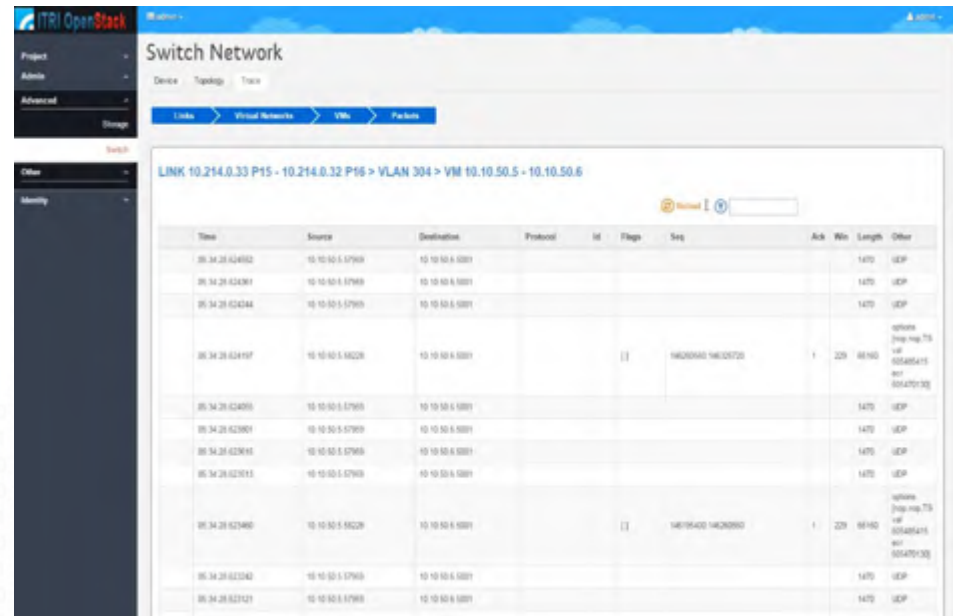
- Peregrine is able to re-deploy packet transmission path when any of link or device is failed by applying centralization control architecture in Fast Failover.





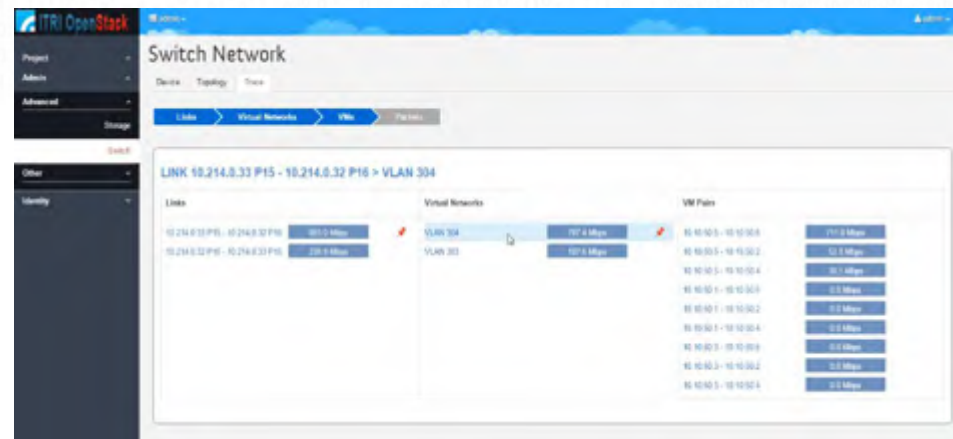
# Peregrine UI

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



The screenshot shows the 'Switch Network' page in the Peregrine UI. The breadcrumb trail is 'Links > Virtual Networks > VMs > Patches'. The selected link is 'LINK 10.214.0.33 P15 - 10.214.0.32 P16 > VLAN 304 > VM 10.10.50.5 - 10.10.50.6'. Below the breadcrumb, there is a table of traffic logs.

Time	Source	Destination	Protocol	Id	Flags	Seq	Ack	Win	Length	Other
00:34:20.624002	10.10.50.5.17900	10.10.50.5.5001							1470	UDP
00:34:20.623991	10.10.50.5.17900	10.10.50.5.5001							1470	UDP
00:34:20.624044	10.10.50.5.17900	10.10.50.5.5001							1470	UDP
00:34:20.624197	10.10.50.5.5002	10.10.50.5.5001			[ ]	146000000 146020720	1	228	60160	udp:seq 75 len 60160(415) win 601470(132)
00:34:20.624001	10.10.50.5.17900	10.10.50.5.5001							1470	UDP
00:34:20.623991	10.10.50.5.17900	10.10.50.5.5001							1470	UDP
00:34:20.623991	10.10.50.5.17900	10.10.50.5.5001							1470	UDP
00:34:20.623991	10.10.50.5.17900	10.10.50.5.5001							1470	UDP
00:34:20.627460	10.10.50.5.5002	10.10.50.5.5001			[ ]	146195400 146208800	1	228	60160	udp:seq 75 len 60160(415) win 601470(132)
00:34:20.623040	10.10.50.5.17900	10.10.50.5.5001							1470	UDP
00:34:20.623121	10.10.50.5.17900	10.10.50.5.5001							1470	UDP

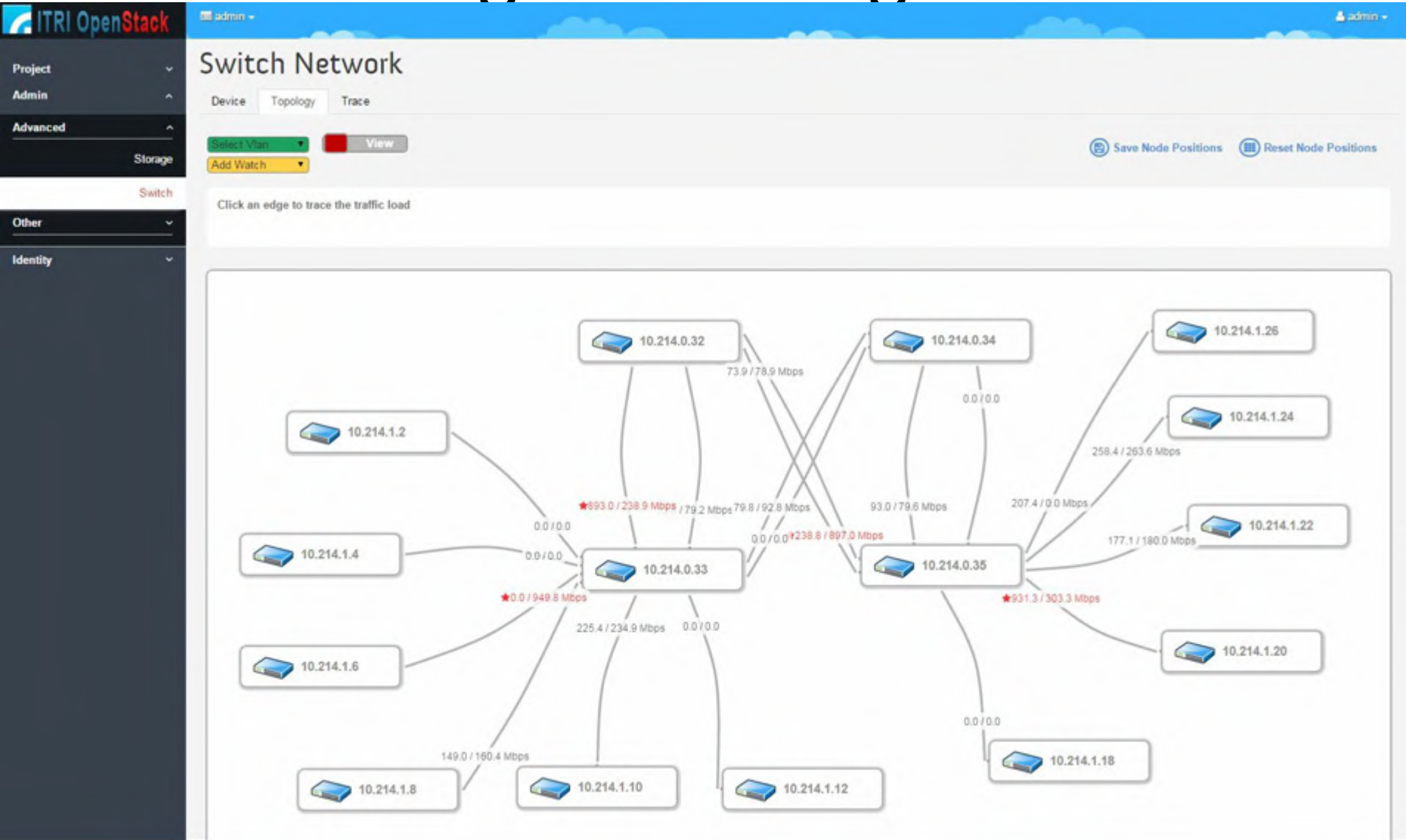


The screenshot shows the 'Switch Network' page in the Peregrine UI. The breadcrumb trail is 'Links > Virtual Networks > VMs > Patches'. The selected link is 'LINK 10.214.0.33 P15 - 10.214.0.32 P16 > VLAN 304'. Below the breadcrumb, there is a table of traffic logs.

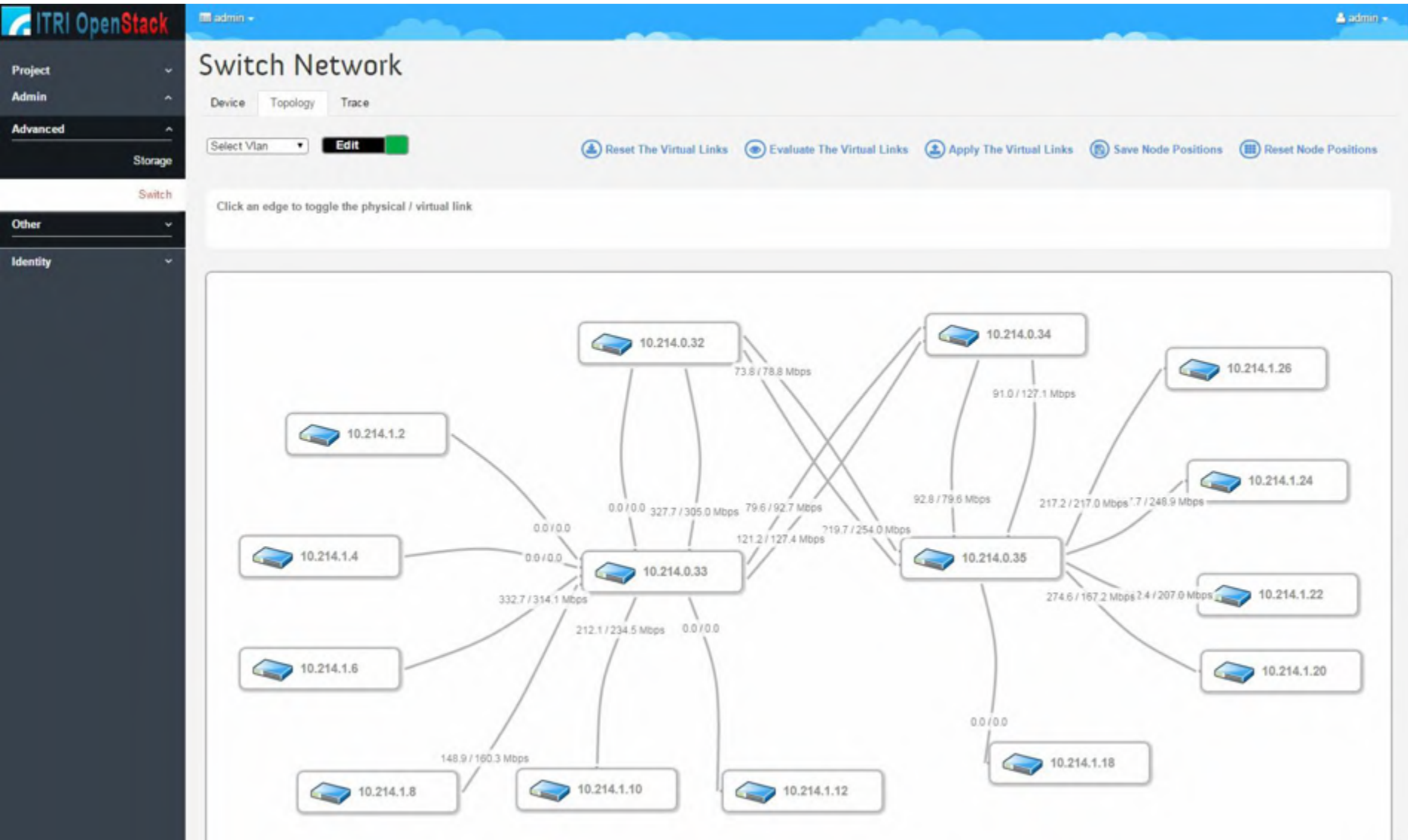
Time	Source	Destination	Protocol	Id	Flags	Seq	Ack	Win	Length	Other
00:34:20.624002	10.10.50.5.17900	10.10.50.5.5001							1470	UDP
00:34:20.623991	10.10.50.5.17900	10.10.50.5.5001							1470	UDP
00:34:20.624044	10.10.50.5.17900	10.10.50.5.5001							1470	UDP
00:34:20.624197	10.10.50.5.5002	10.10.50.5.5001			[ ]	146000000 146020720	1	228	60160	udp:seq 75 len 60160(415) win 601470(132)
00:34:20.624001	10.10.50.5.17900	10.10.50.5.5001							1470	UDP
00:34:20.623991	10.10.50.5.17900	10.10.50.5.5001							1470	UDP
00:34:20.623991	10.10.50.5.17900	10.10.50.5.5001							1470	UDP
00:34:20.623991	10.10.50.5.17900	10.10.50.5.5001							1470	UDP
00:34:20.627460	10.10.50.5.5002	10.10.50.5.5001			[ ]	146195400 146208800	1	228	60160	udp:seq 75 len 60160(415) win 601470(132)
00:34:20.623040	10.10.50.5.17900	10.10.50.5.5001							1470	UDP
00:34:20.623121	10.10.50.5.17900	10.10.50.5.5001							1470	UDP



# Traffic Congestion Diagnosis Video



# Link Failover Video



# PDCM

- PDCM stands for **P**hysical **D**ata **C**enter **M**anagement.
- 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

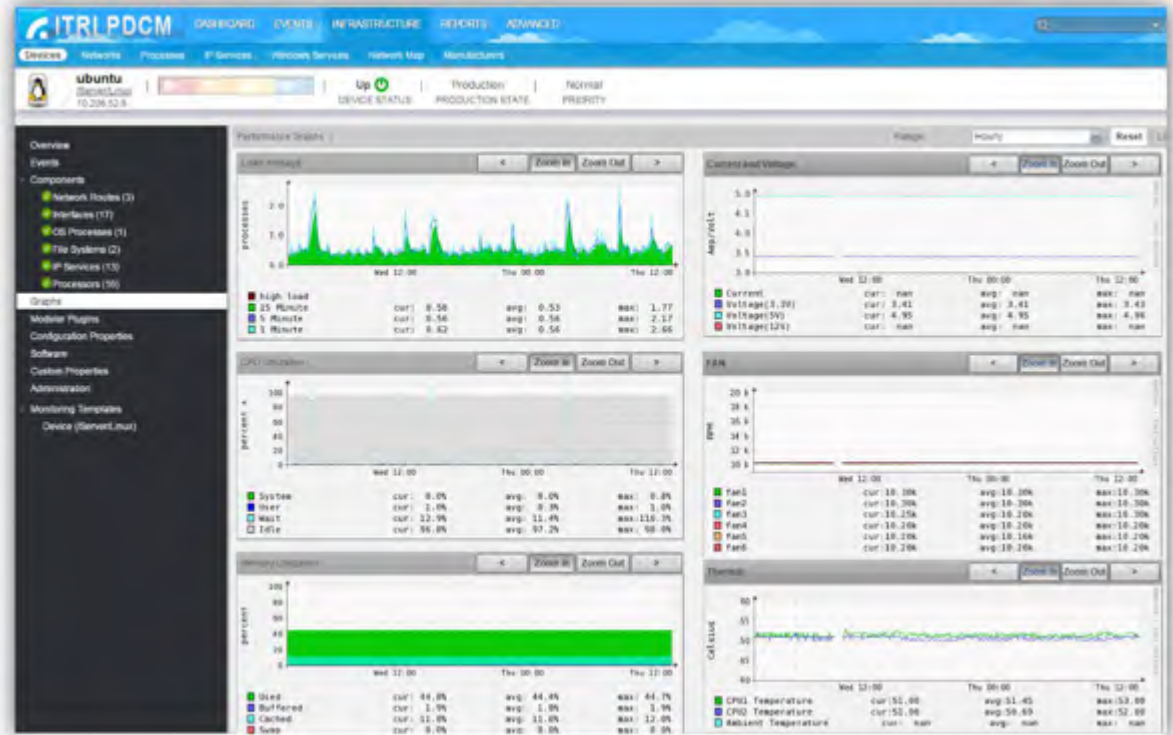
PDCM provides a comprehensive solution for monitoring OpenStack cloud, including hardware devices and OpenStack services.





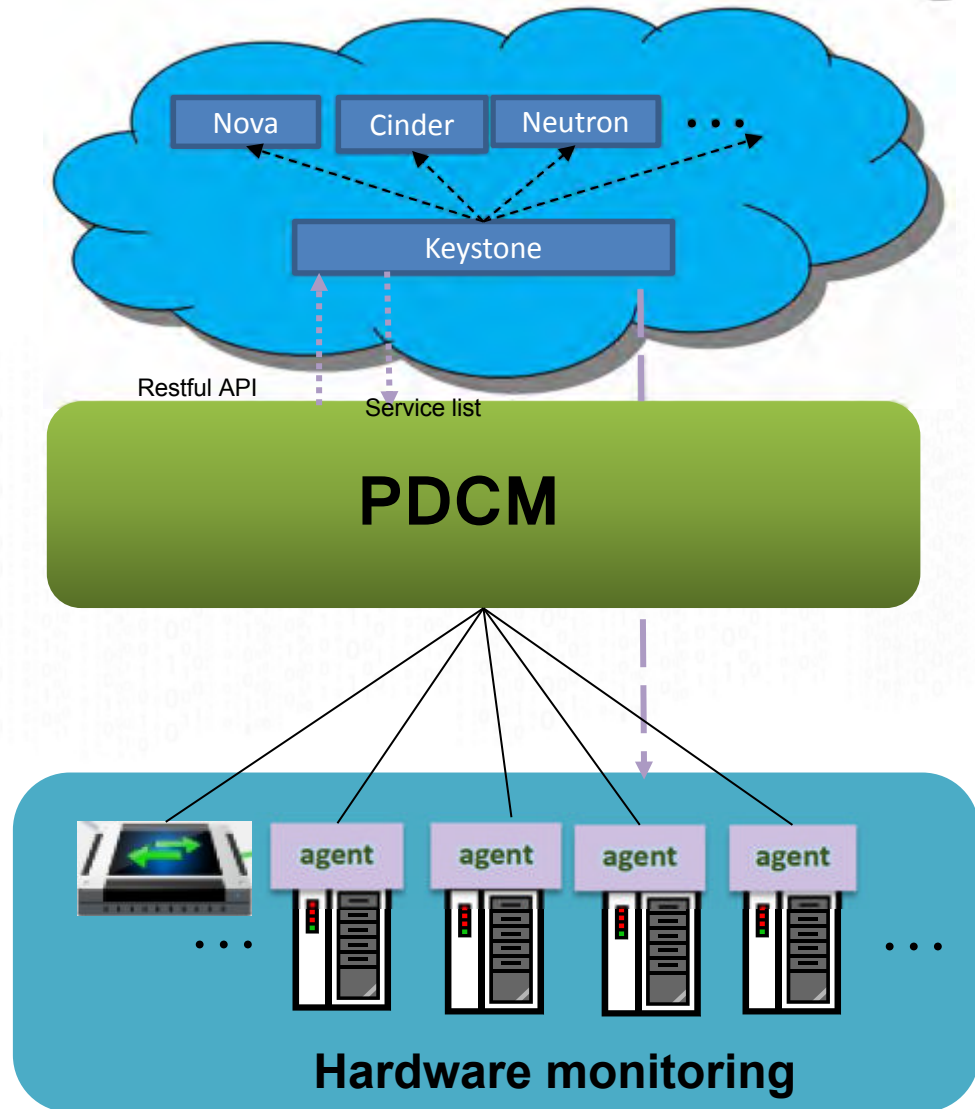
# Device Hardware Monitoring

- ✓ 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

