Comparison of control plane deployment architectures in the scope of hyperconverged OpenStack infrastructure

Miroslav Halas
Lenovo Cloud Technology Center
May 2017

2017 Lenovo. All rights reserved.

About me

- Director of SW Architecture, Cloud Infrastructure
- SW Engineer, Architect and Leader
    - Designed, implemented and operated Desktop, Mobile, SaaS, Cloud Applications and Large Private Cloud Platforms
    - 10 years of engineering leadership for information security, public and private cloud for one of top 3 financials in US
    - Holds patents in security, private and public cloud areas
- Carries picture of his home server in his pocket

linkedin.com/in/miroslavhalas/

2017 Lenovo. All rights reserved..

Outline

- Introduction to Lenovo Cloud Technology Center
- What It Takes to Deploy and Operate SDDC
- What Is Control Plane and Why Does It Matter
- Resilient Converged Infrastructure Control Plane
- Evaluation and Testing using Rally and Phoronix Test Suite

2017 Lenovo. All rights reserved..

Standing on the Shoulders of Giants

Lenovo DCG Research & Technology

- Focused research in Systems, Storage, Cloud, AI, Big Data, etc.

Lenovo System Technology Innovation Center

- Co-innovating with partners and customers to deliver Next-Next ideas such as Abstract hybrid data centers, Dynamic reconfiguration of IT and Deployment of IIoT

Lenovo Cloud Technology Center

- Pioneering cloud advancements, deployment and management experiences with community, partners and customers

Lenovo DCG Product & Development

- Engineer and support production ready enterprise solutions

2017 Lenovo. All rights reserved..

Lenovo Cloud Technology Center (LCTC)

Mission: *Align and Drive Lenovo entry and leadership in Enterprise Cloud Infrastructure, building core capabilities and partnerships for a complete portfolio of Solutions – focus on open source*

Key Activities

Customer and Market Engagement

- Deep customer engagement and support
- Business/Partner development, Technical Product evaluations
- Open source ecosystem visibility and credibility

Portfolio Development

- OpenStack and Open Systems development – distros, containers and integrations with solutions
- SDDC – development of software defined storage, networking
- Professional and Managed services

Center of Competence

- Cloud architecture and advisory board
- e2e Cloud roadmap

Long Term Goal

Credible Cloud Market Presence

Complete Cloud Solutions Portfolio

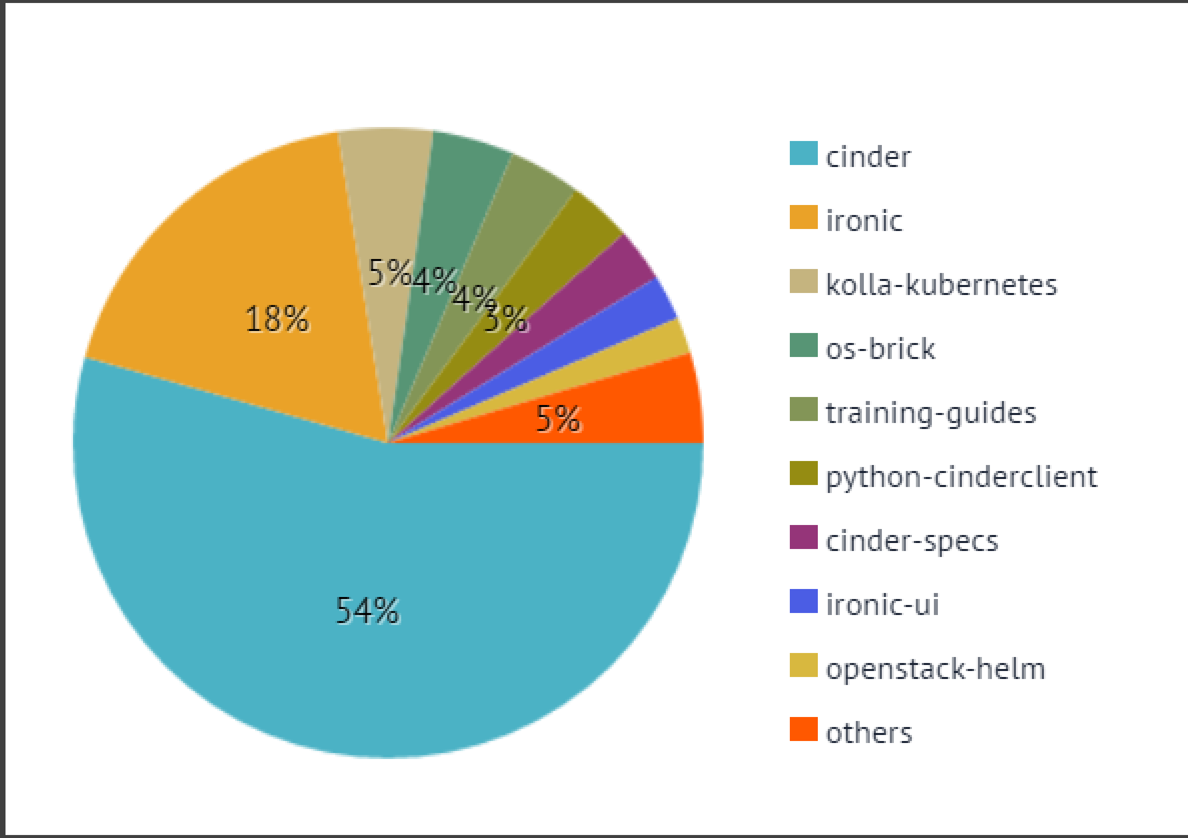
Depth and Alignment of Expertise

MAJOR RESEARCH CENTERS IN RALEIGH, BEIJING, SHANGHAI

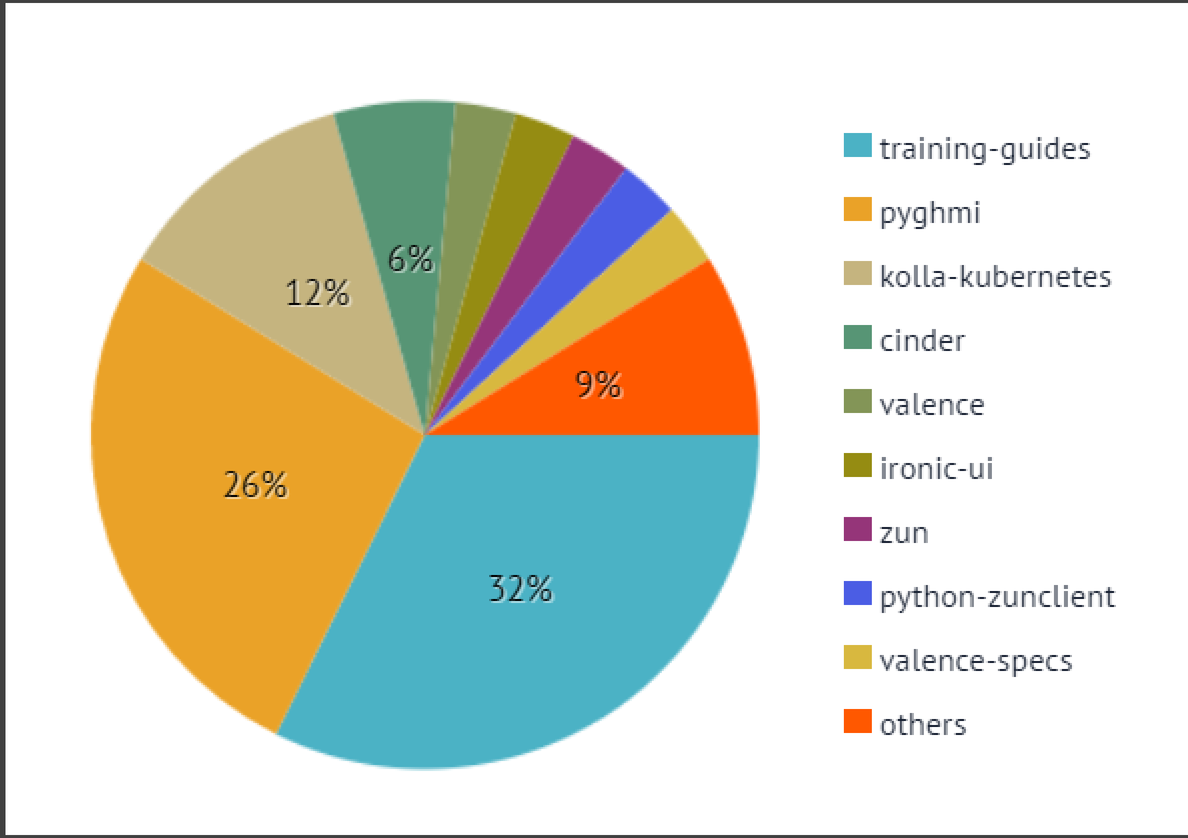
2017 Lenovo. All rights reserved.

OpenStack Upstream Engagement

Reviews



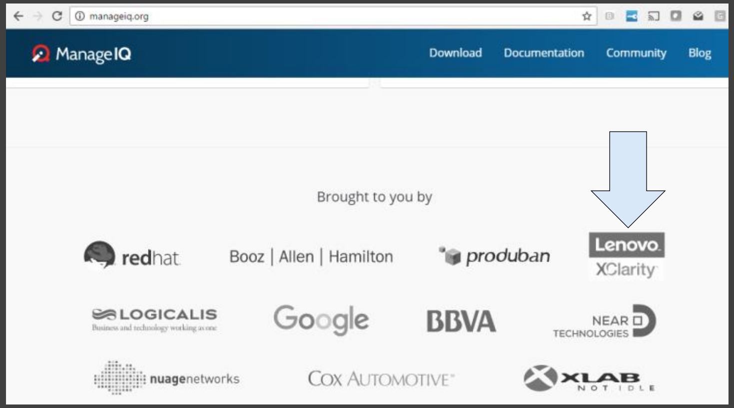
Commits

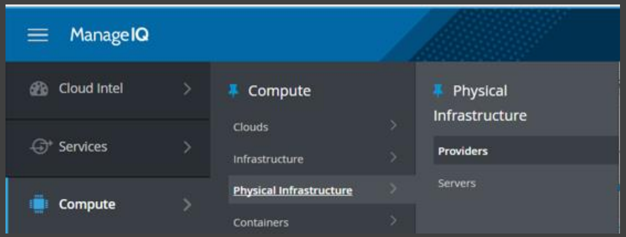


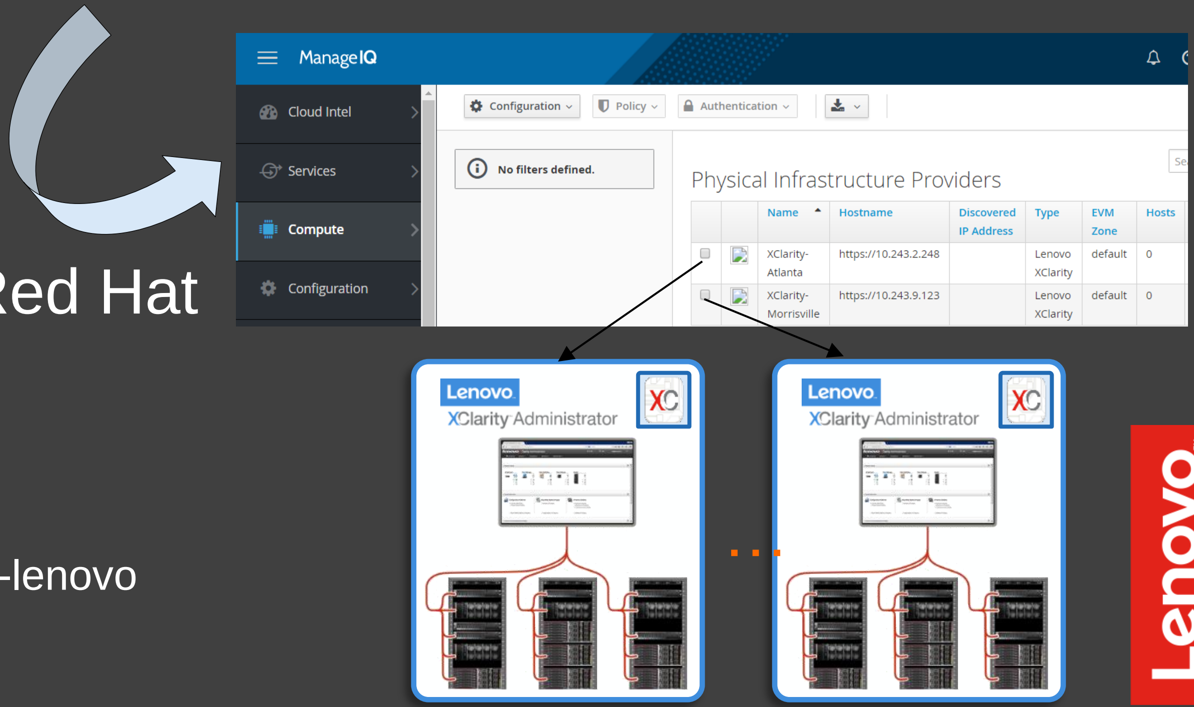
2017 Lenovo. All rights reserved..

ManageIQ Upstream Engagement

- Initial provider of ManageIQ Physical Infrastructure
- Features/Use cases
  - Physical Server Data Model
  - New XClarity Provider
  - XClarity Provider Summary View
  - Physical Server Inventory via REST
    - Vital Product Data (VPD)
    - Firmware Levels
    - Server to Platform Host Relationships



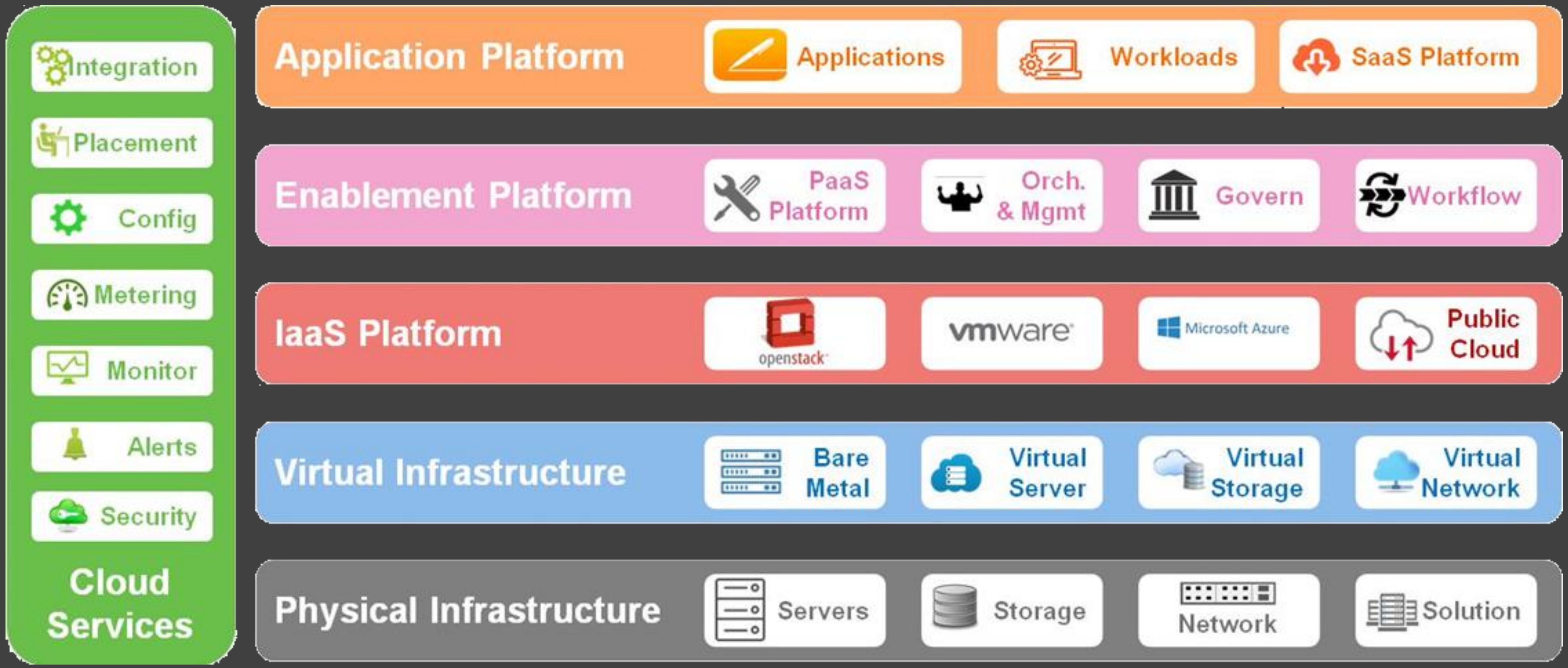




- Outstanding development support from Red Hat ManageIQ / CloudForms team
- Open Source Contributions Summary
  - **ManageIQ Lenovo Provider**
    - https://github.com/ManageIQ/manageiq-providers-lenovo
  - **Lenovo XClarity Client (Ruby)**
    - https://github.com/lenovo/xclarity_client

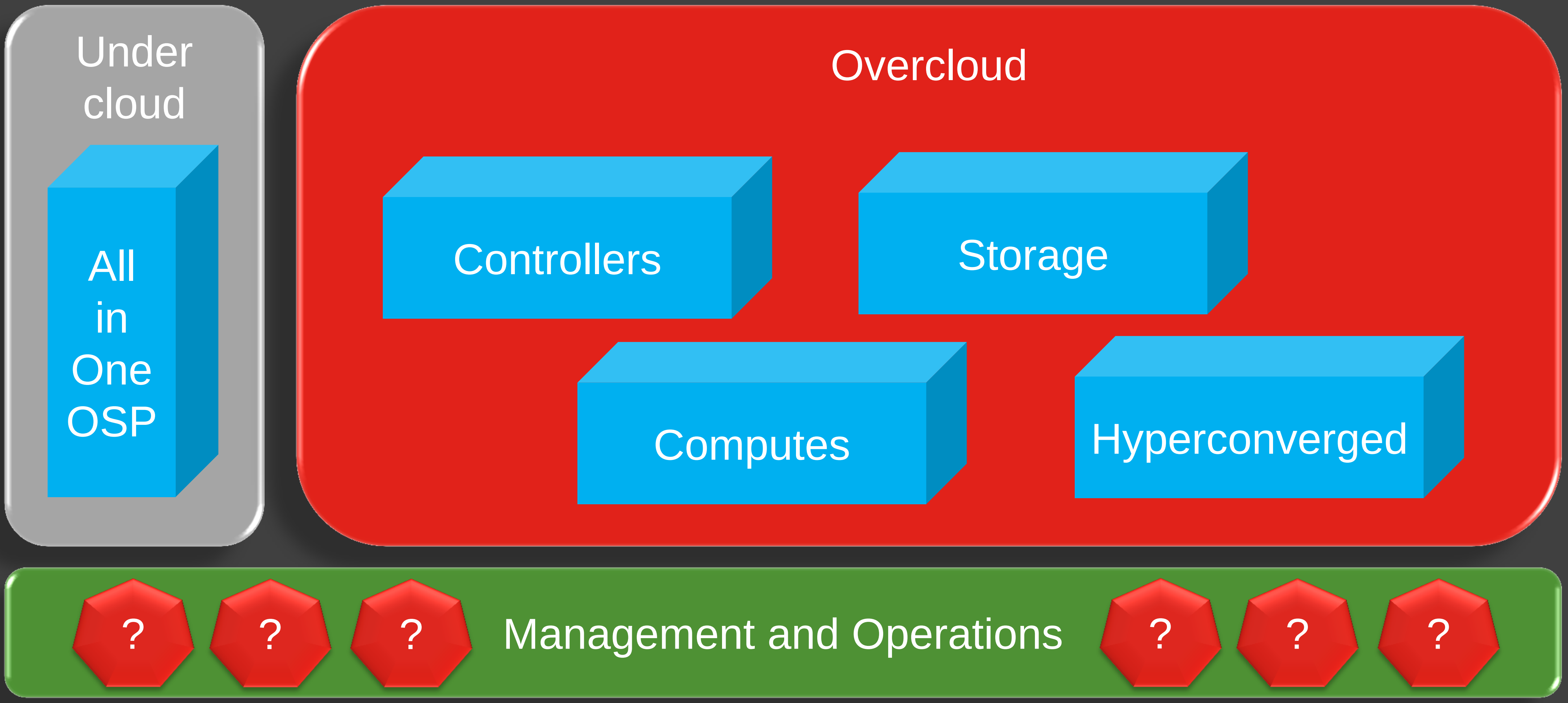
2017 Lenovo. All rights reserved..

What It Takes to Deploy and Operate SDDC



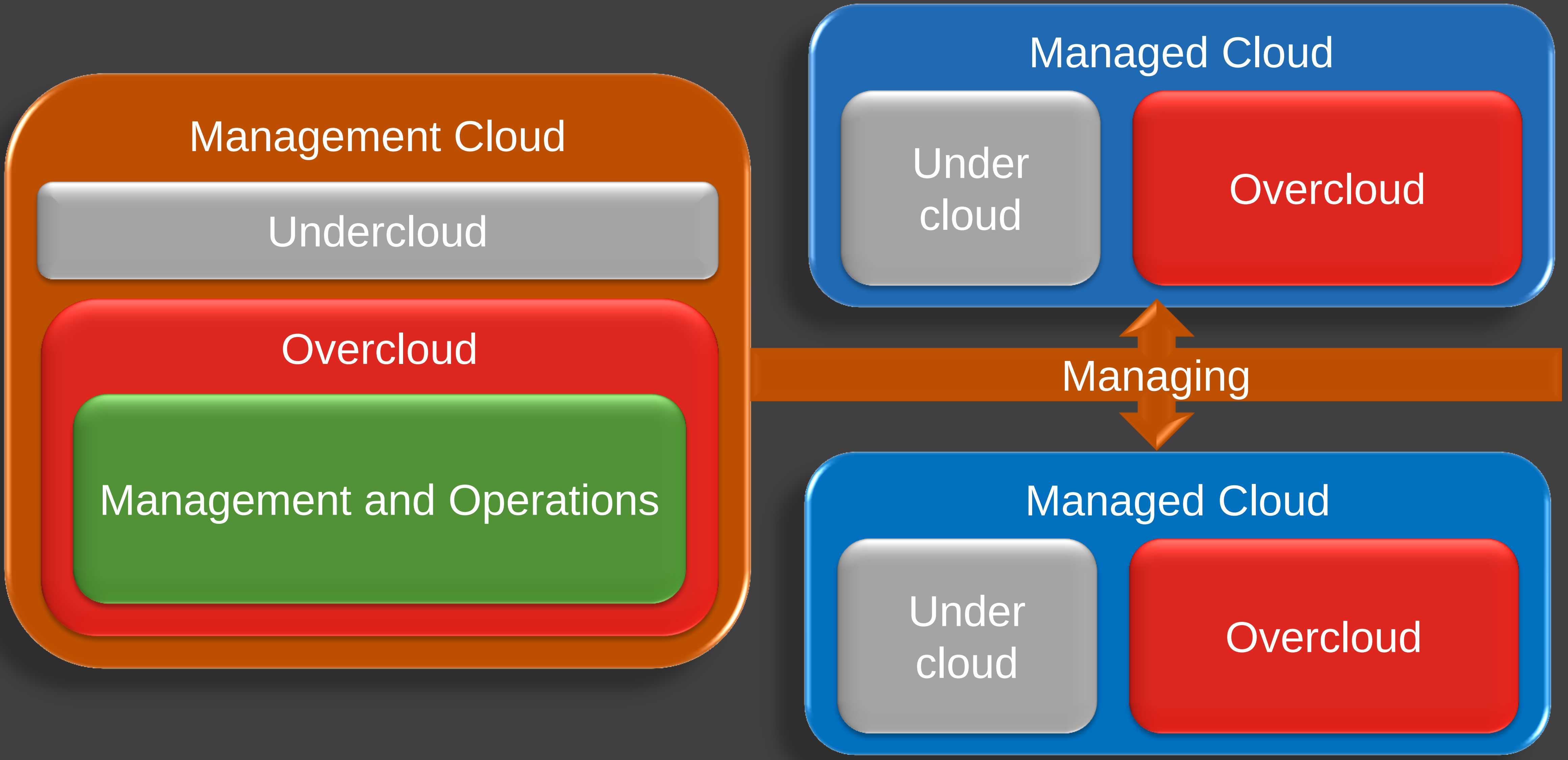
2017 Lenovo. All rights reserved.

Simple Deployment Pattern



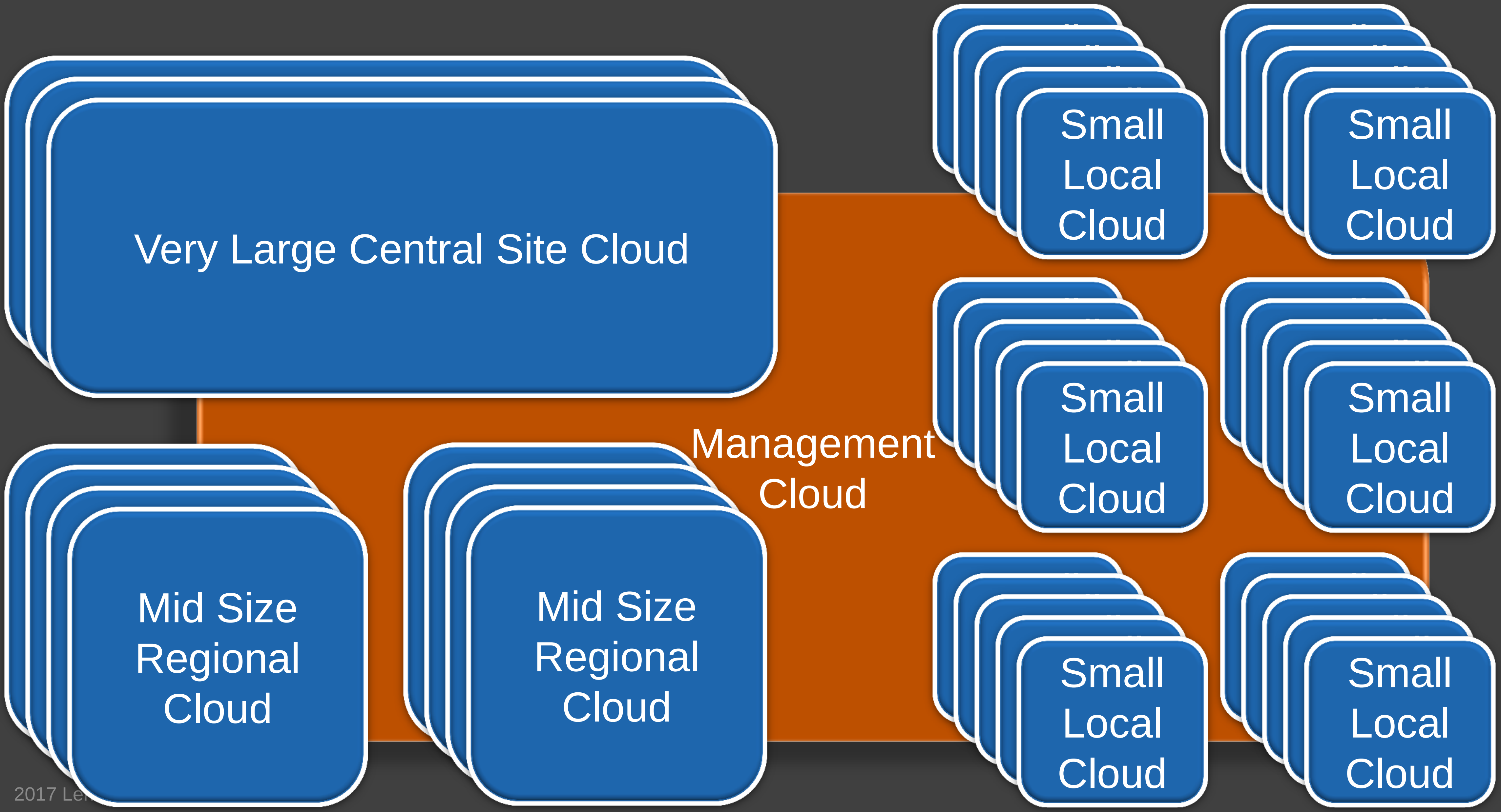
2017 Lenovo. All rights reserved..

Realistic Deployment Pattern



2017 Lenovo. All rights reserved..

Large Deployment Pattern

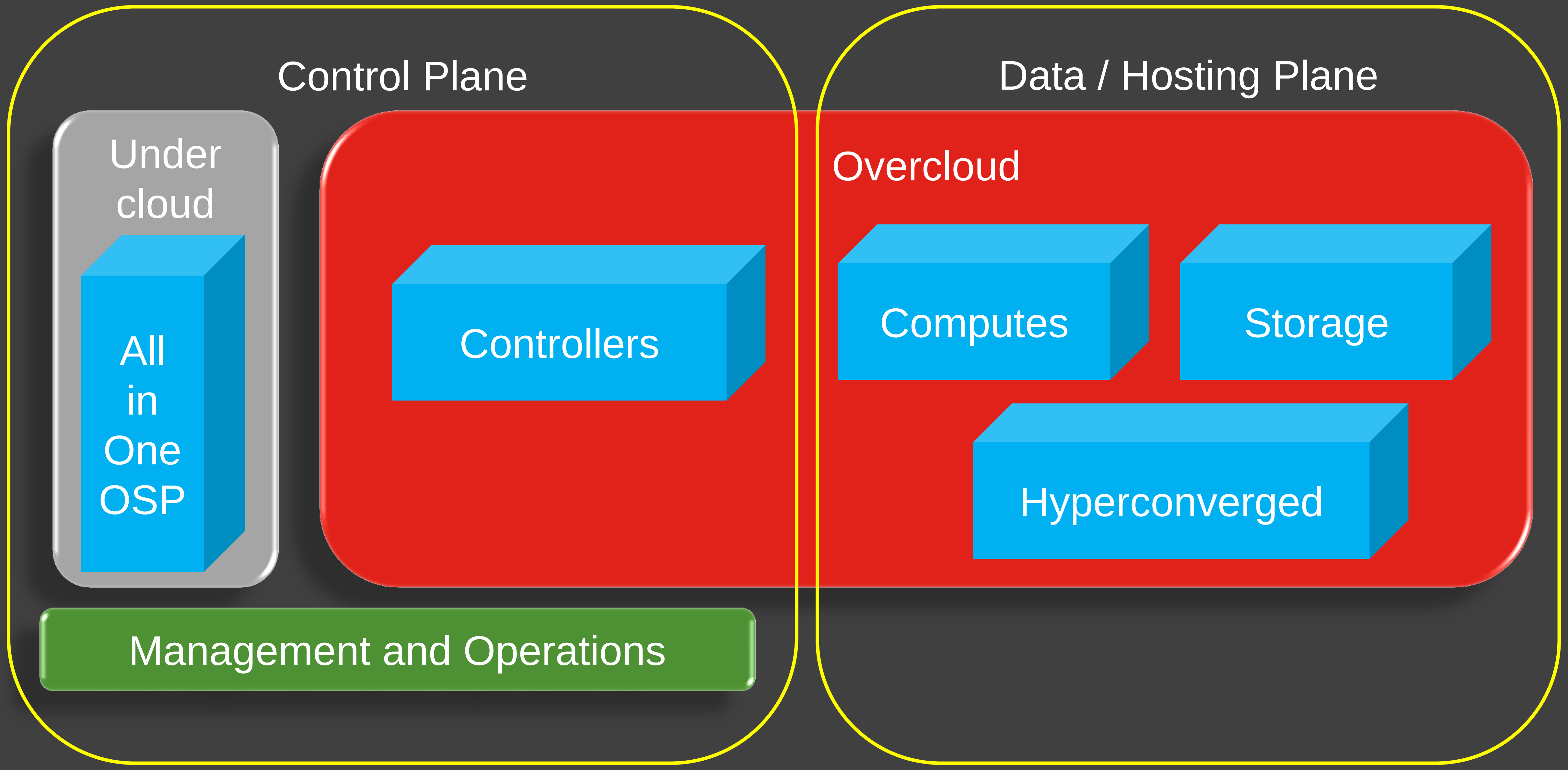


Size / Cost / Resource Optimized Cloud Infrastructure

- All deployment patterns have element that is sensitive to cost / space / power / skills / complexity
- Objectives achieved by compressing and collocating distinctive functions on the same set of HW and utilizing familiar concepts
- Converged Infrastructure Management Plane (CIMP)
  - OpenStack and non OpenStack management functions share the same HW resources
- Compressed Data Plane
  - Hyperconverged Compute Node collocating compute and storage on the same node
- Resource protection of shared resources

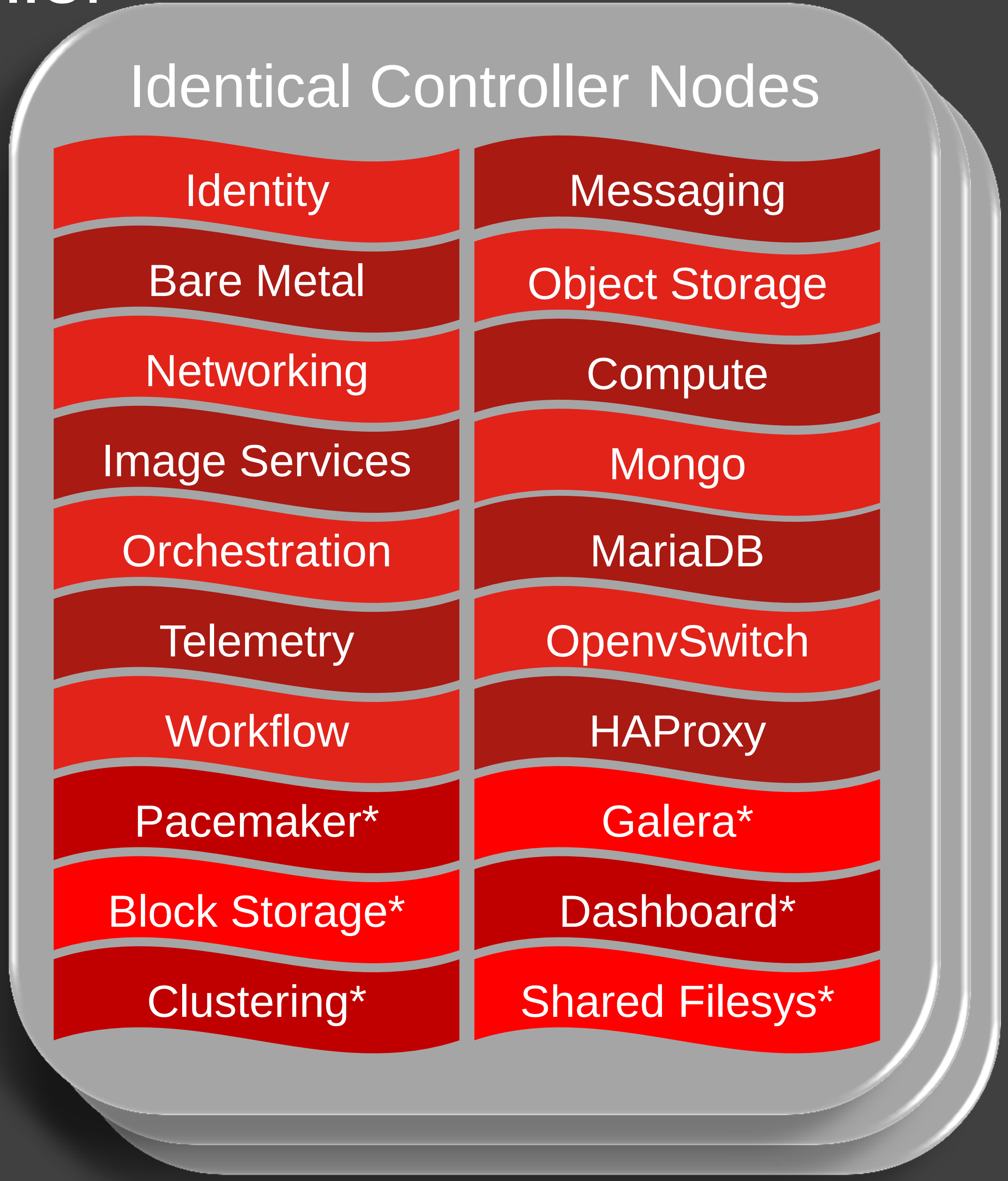
2017 Lenovo. All rights reserved..

What Is Control Plane and Why Does It Matter



2017 Lenovo. All rights reserved..

Monolithic Controller



2017 Lenovo. All rights reserved..

Control Plane Beyond OpenStack Services

- Required to support full set of SDDC features
- Provide resiliency and autonomy for remote locations

- SDN – Contrail, NSX, Nuage
- Configuration Management – Ansible Tower, Pupper, SaltStack
- Logging – Logstash, Splunk
- Analytics – Elasticsearch
- Visualization – Kibana, Graphana, Prometheus
- Monitoring – Nagios, Zabbix, DataDog
- Performance Monitoring – Telegraf, CollectD, Graphite, InfluxDB
- Security – PowerBroker, ESM, CyberArk
- Capacity planning and optimization – Cirba, ManageIQ,

2017 Lenovo. All rights reserved..

Disaggregated Controller

Generic Services Controller

Identity	Messaging
Bare Metal	Object Storage
Networking	Shared Filesys
Image Services	OpenvSwitch
Orchestration	HAProxy
Clustering	Workflow
Block Storage	Pacemaker

Database Controller

Telemetry	Mongo
HAProxy	MariaDB
Pacemaker	Galera

UI Services Controller

Dashboard	HAProxy

Compute Scheduler Controller

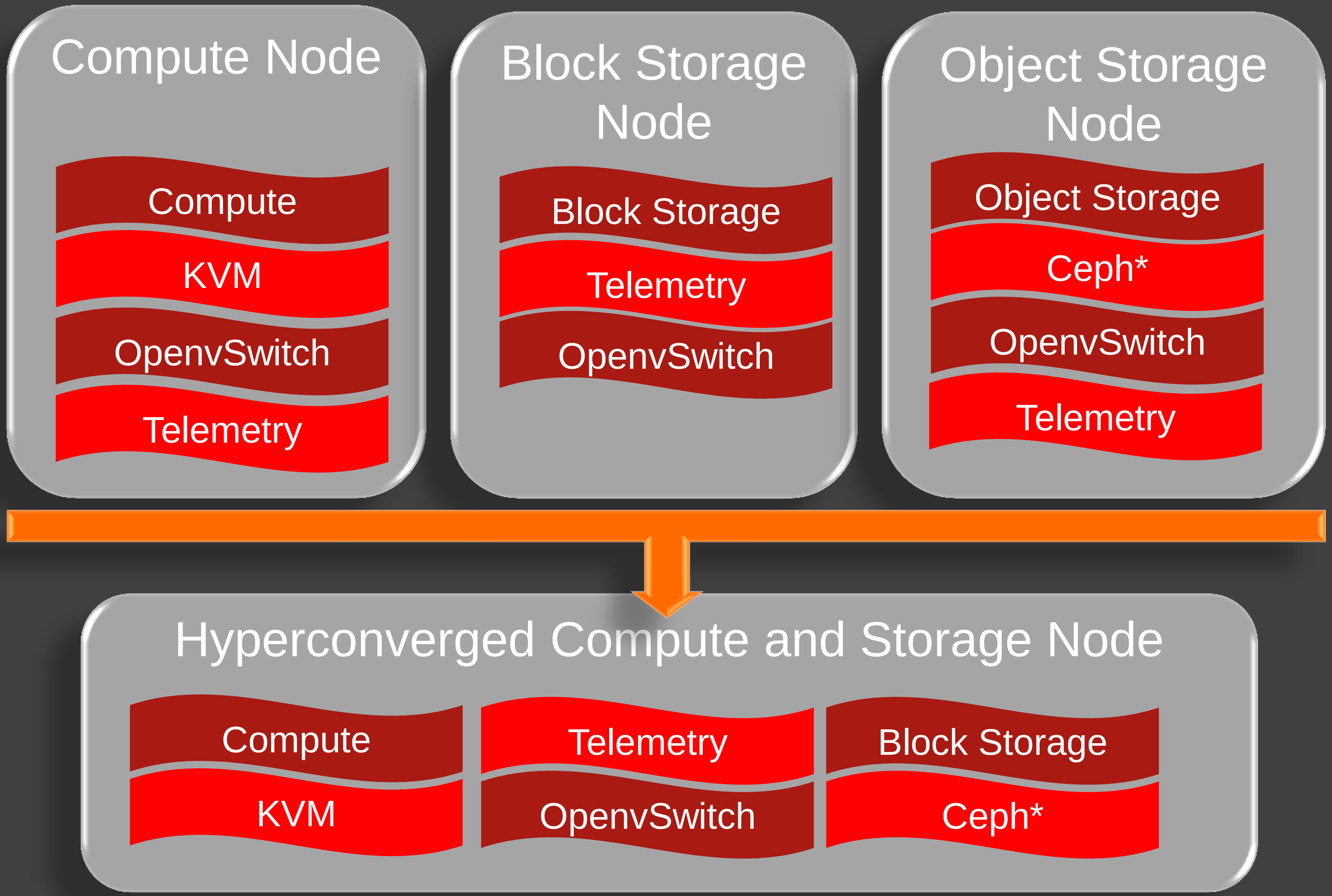
Compute	HAProxy

❖ OpenStack Performance Team, Barcelona Summit 2016

2017 Lenovo. All rights reserved..

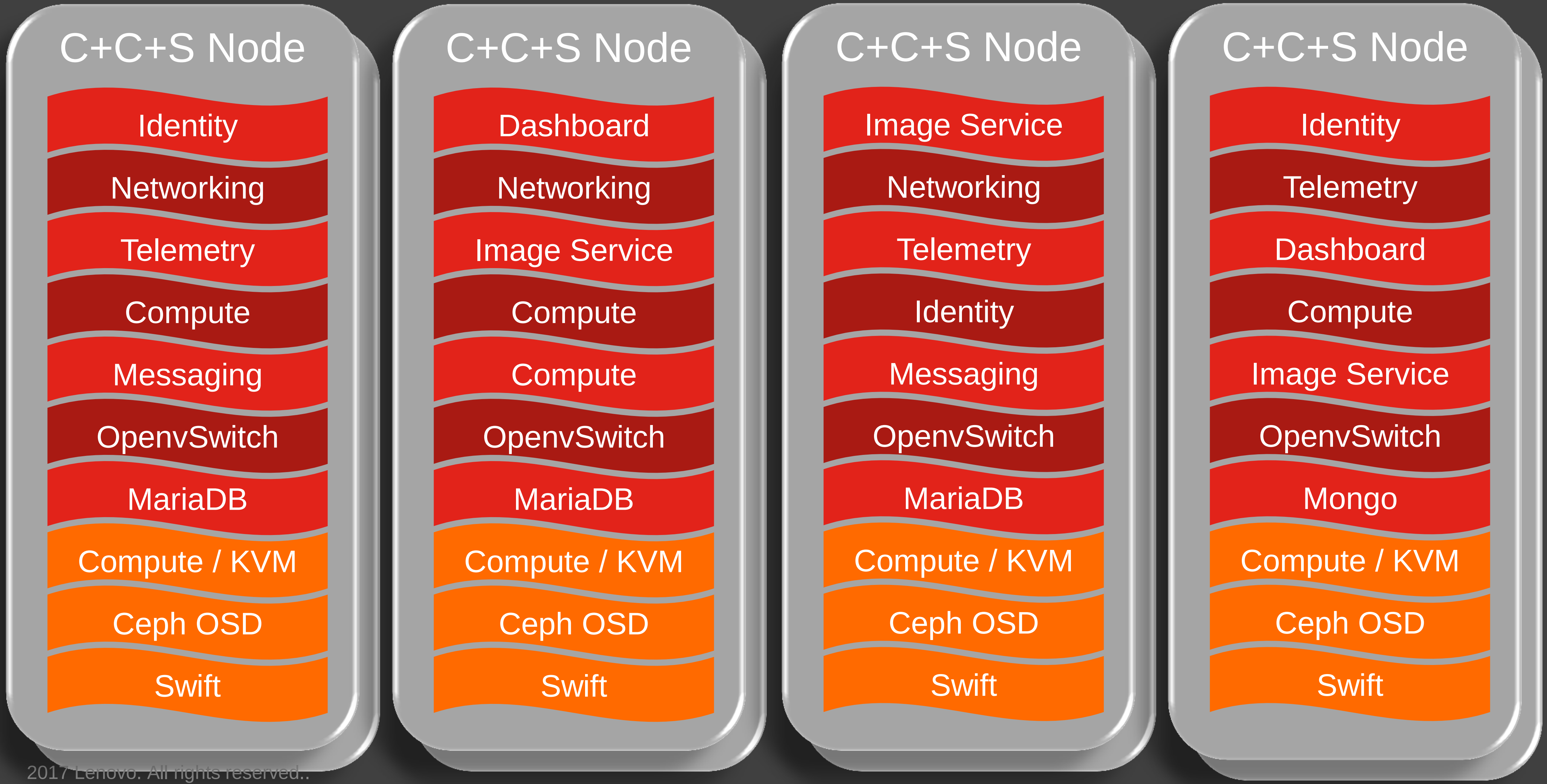
Lenovo

Hyperconverged Compute and Storage Node



2017 Lenovo. All rights reserved..

Hyperconverged Controller Compute and Storage Node



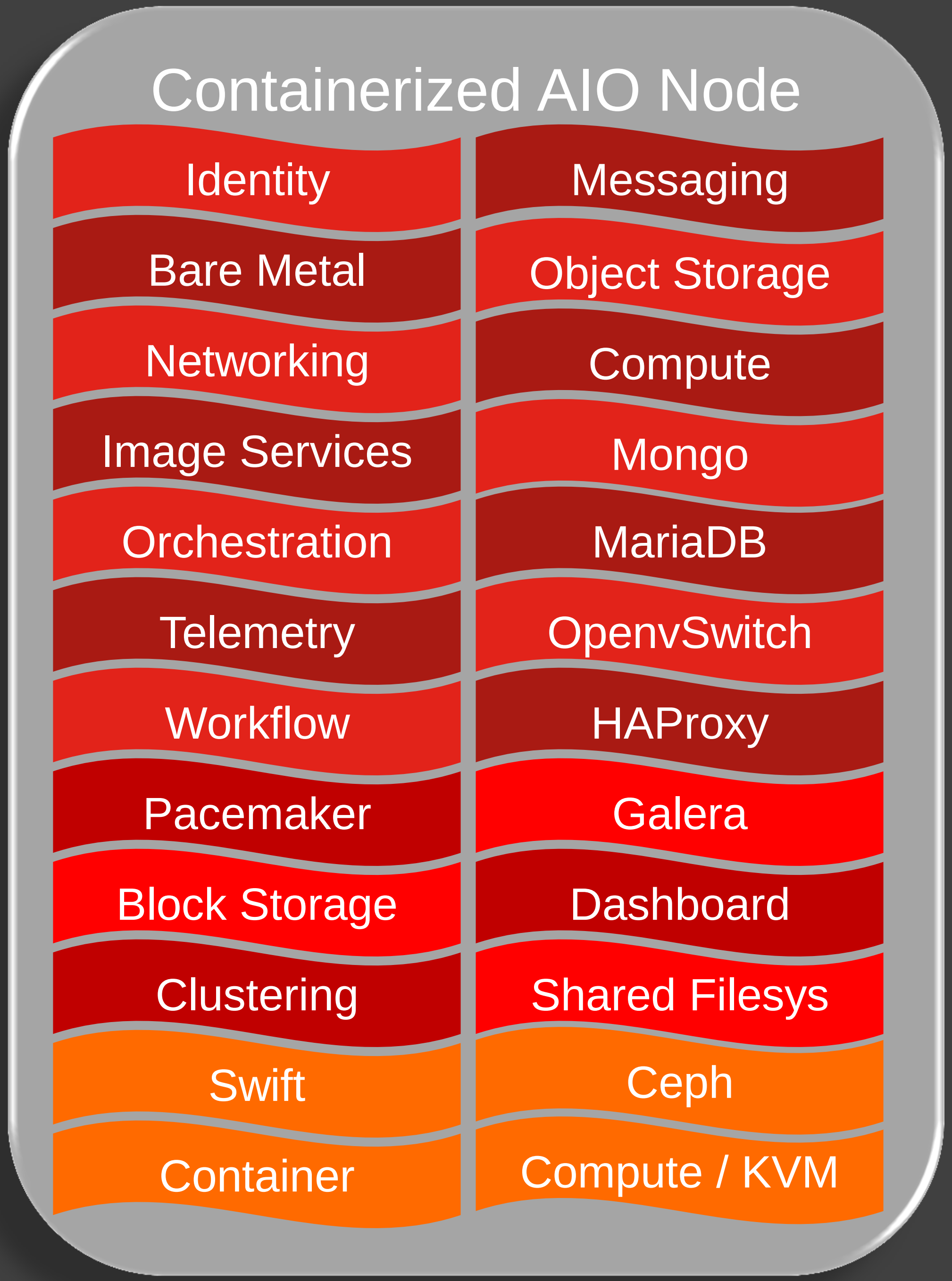
2017 Lenovo. All rights reserved..

Containerized Controller

❖Kolla, OpenStack Helm, OpenStack LOCI

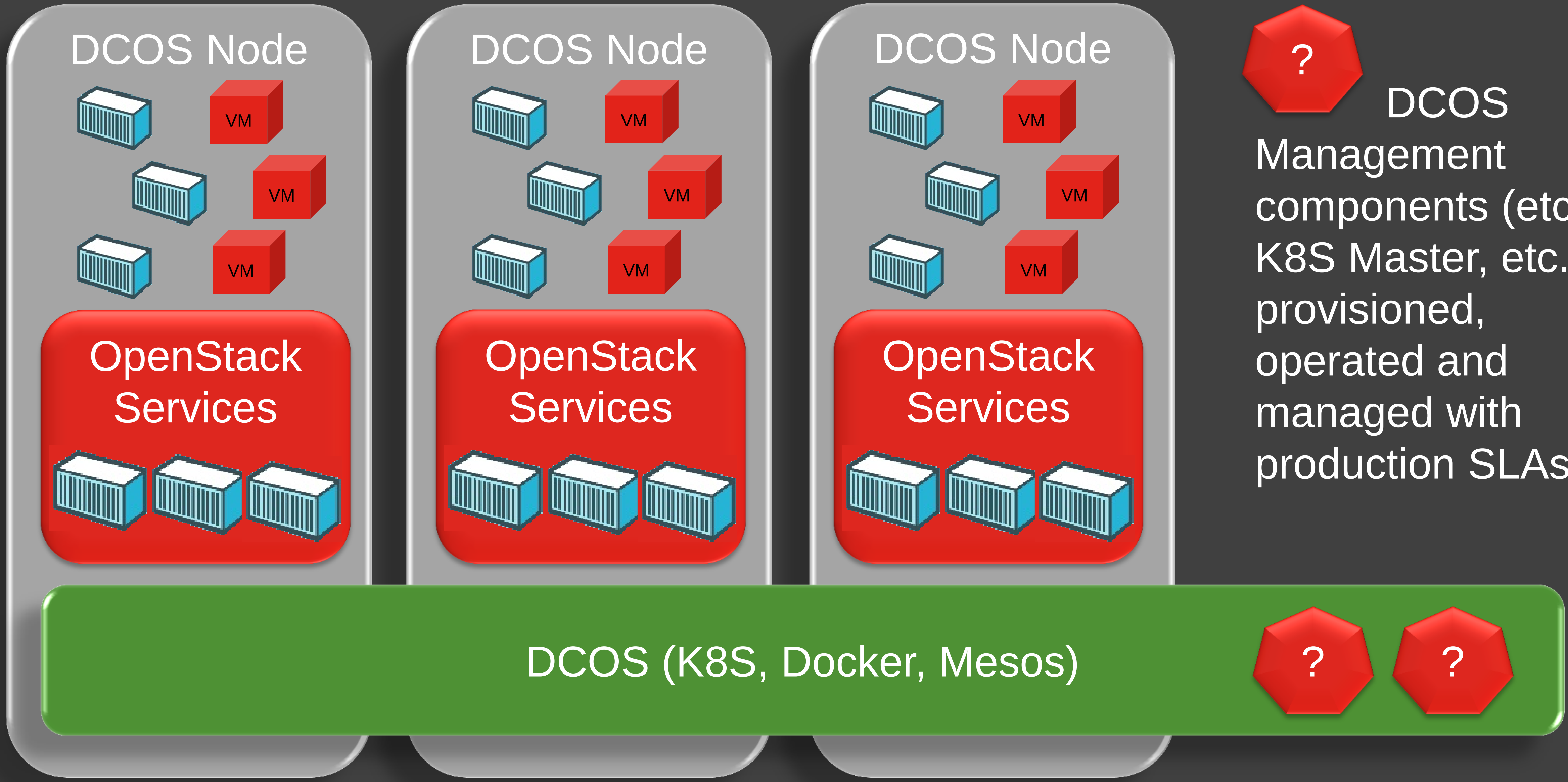
https://docs.openstack.org/developer/kolla-kubernetes/deployment-guide.html

```
helm install kolla-kubernetes/helm/service/mariadb --name mariadb
helm install kolla-kubernetes/helm/service/rabbitmq --name rabbitmq --values ./cloud.yaml
helm install kolla-kubernetes/helm/service/memcached --name memcached --values
./cloud.yaml
helm install kolla-kubernetes/helm/service/keystone --name keystone --values ./cloud.yaml
helm install kolla-kubernetes/helm/service/glance --name glance --values ./cloud.yaml
helm install kolla-kubernetes/helm/service/cinder-control --name cinder-control --values
./cloud.yaml
helm install kolla-kubernetes/helm/service/horizon --name horizon --values ./cloud.yaml
helm install kolla-kubernetes/helm/service/openvswitch --name openvswitch --values
./cloud.yaml
helm install kolla-kubernetes/helm/service/neutron --name neutron --values ./cloud.yaml
helm install kolla-kubernetes/helm/service/nova-control --name nova-control --values
./cloud.yaml
helm install kolla-kubernetes/helm/service/nova-compute --name nova-compute --values
./cloud.yaml
helm install kolla-kubernetes/helm/microservice/nova-cell0-create-db-job --name nova-cell0-
create-db-job --values ./cloud.yaml
helm install kolla-kubernetes/helm/microservice/nova-api-create-
helm install kolla-kubernetes/helm/service/cinder-volume-lvm
watch -d -n 5 -c kubectl get pods --all-namespaces
```



2017 Lenovo. All rights reserved..

Containerized OpenStack Infrastructure



DCOS Management components (etcd, K8S Master, etc.) provisioned, operated and managed with production SLAs

2017 Lenovo. All rights reserved..

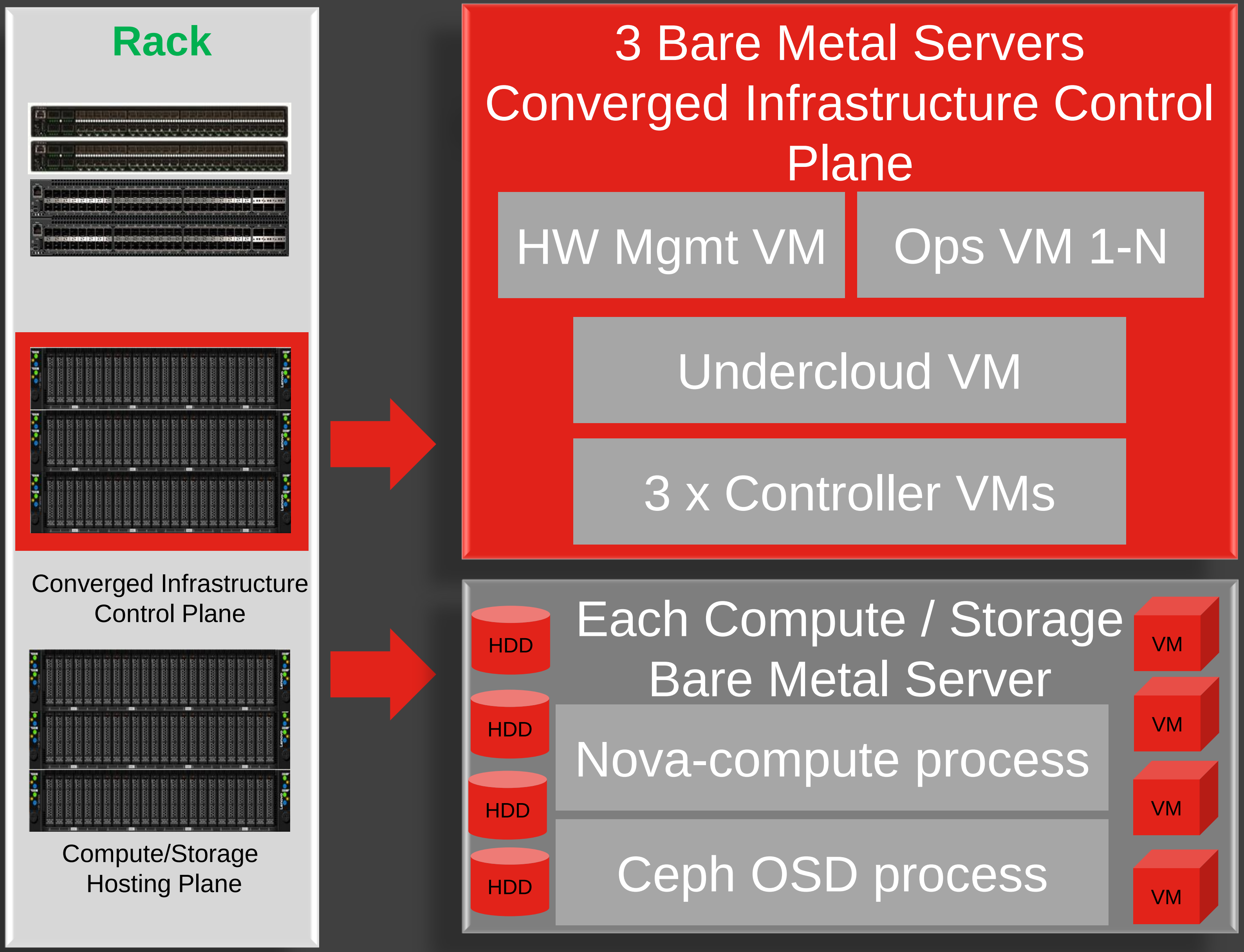
Resilient Converged Infrastructure Control Plane

- Use familiar and production ready technologies to provide cost / space / power / skills / complexity effective control plane required to support full set of SDDC features
- Provide resiliency at all levels for lights out operations
  - Network
    - Redundant management and data plane switches
    - Redundant dual port NICs
  - Servers and storage
    - Redundant PDUs and power supplies
    - 3+ node bare metal server cluster sizes
  - Software
    - Off the shelf, commodity software
    - Live Migration, Shared storage, Snapshots, Backup

2017 Lenovo. All rights reserved..

Control Plane for Resource Constrained Deployments

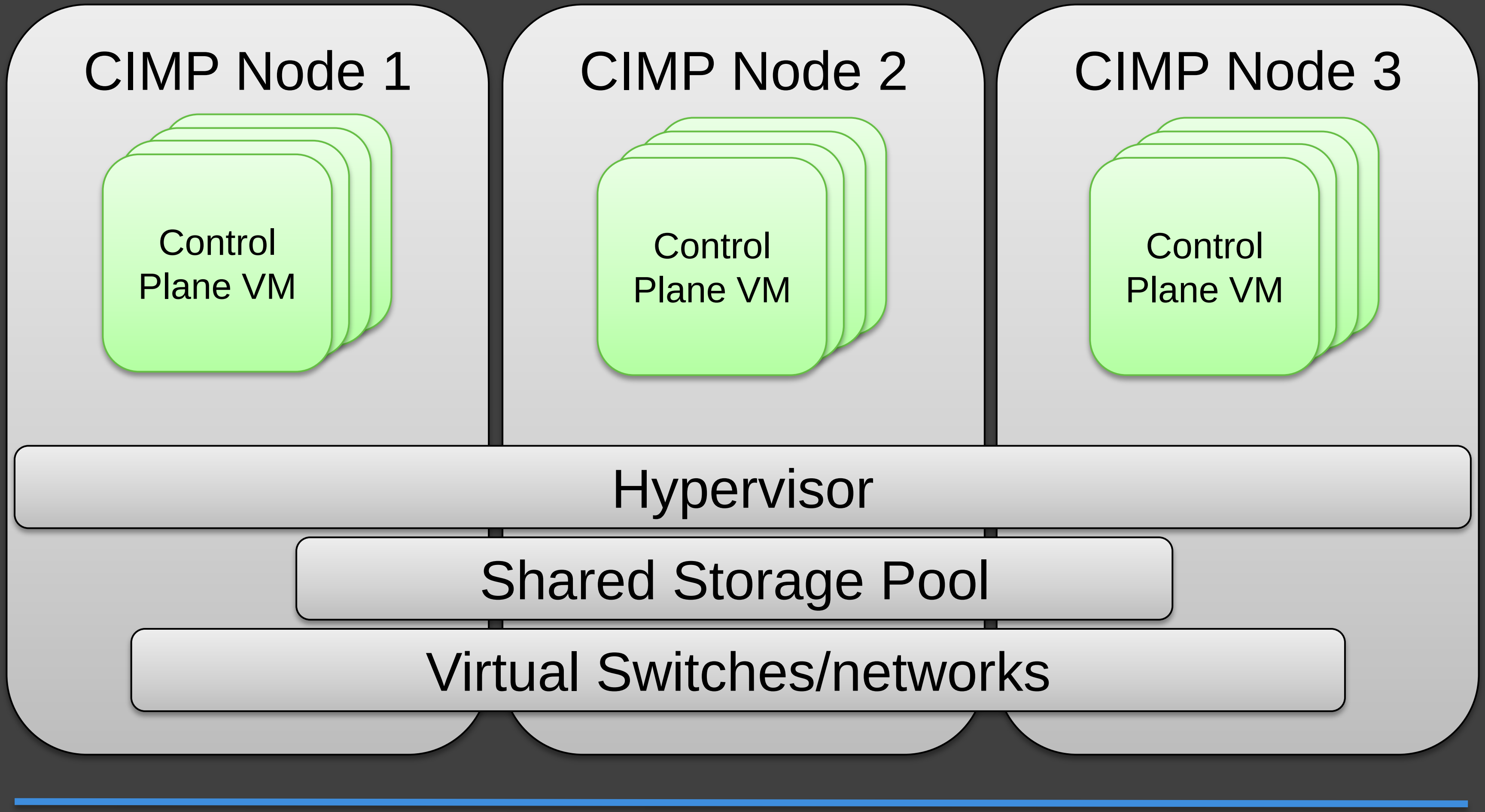
- Standard Lenovo servers and network switches
- Management components (XClarity, Undercloud, ManageIQ, etc.) deployed in virtualized fashion)
- Distributed storage provided by Ceph
- Ceph deployed in hyper-converged mode alongside the KVM hypervisor
- Easy expansibility to host future management and operations functions



2017 Lenovo. All rights reserved..

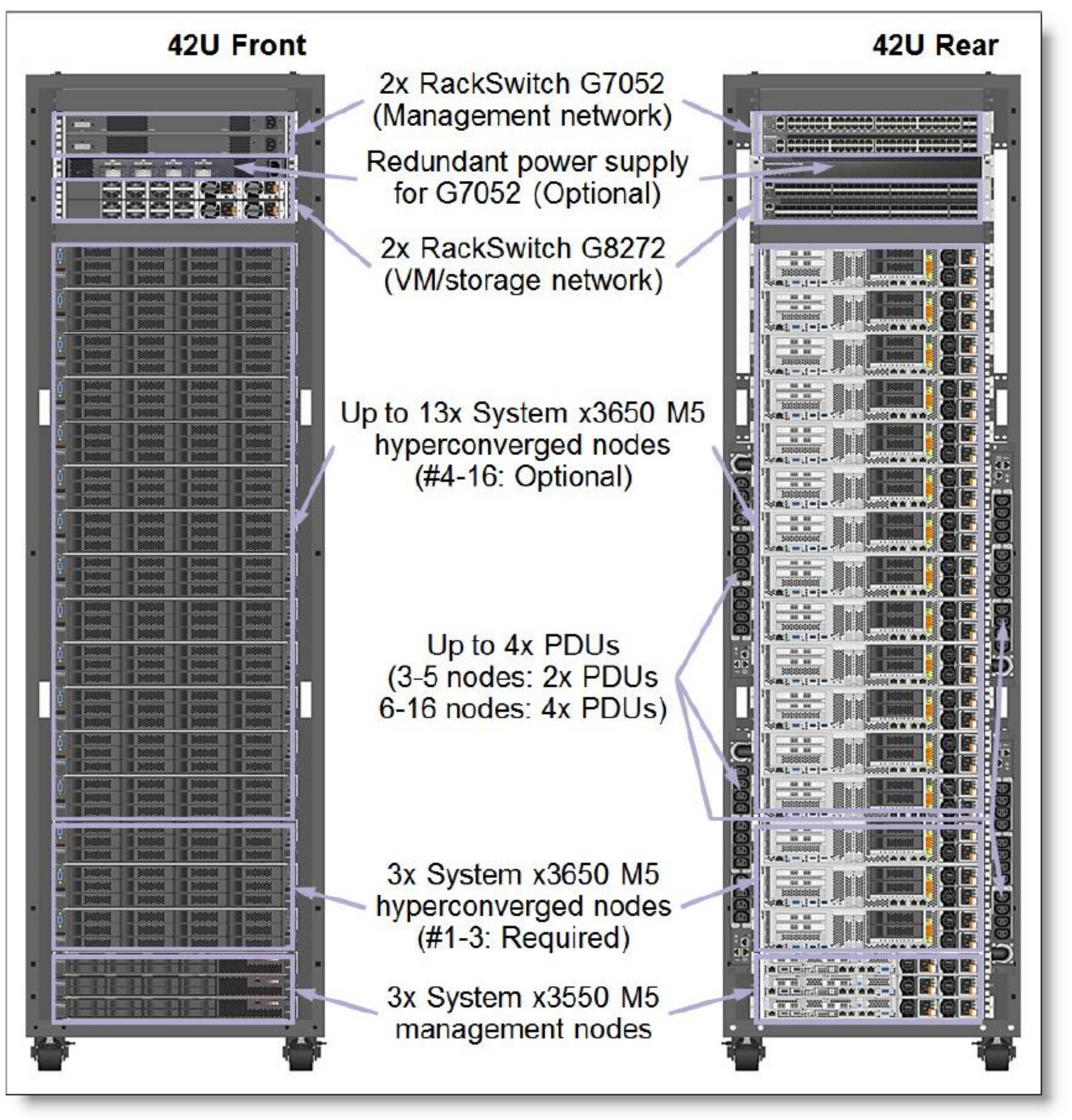
Virtualized Control Plane Structure

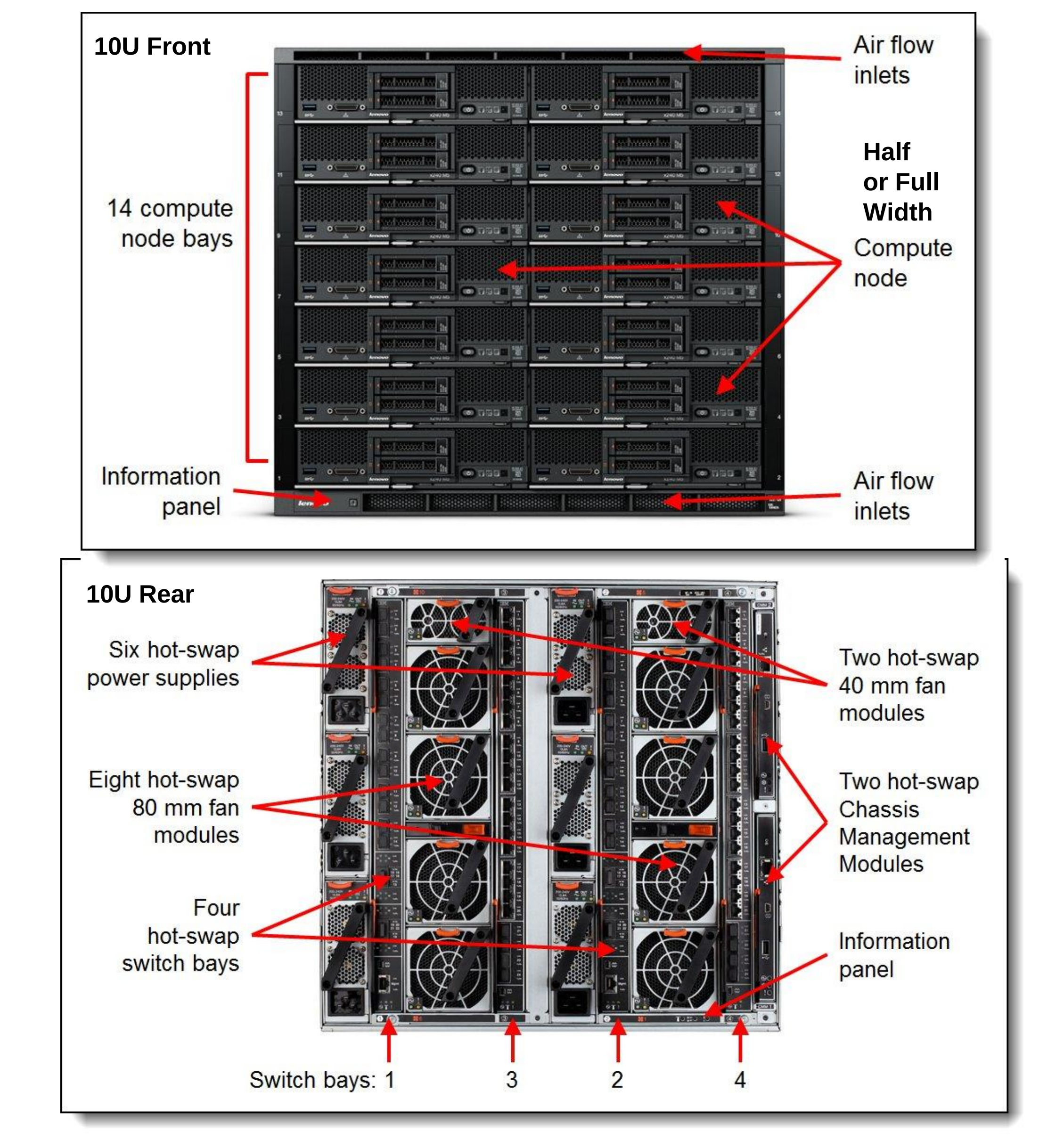
- Hypervisor
  - KVM, Libvirt
  - Live Migration
  - Snapshots
- Networking
  - Linux Bridge
  - VLANs
- Storage
  - GlusterFS
  - Shared storage
  - File system replication
  - HA



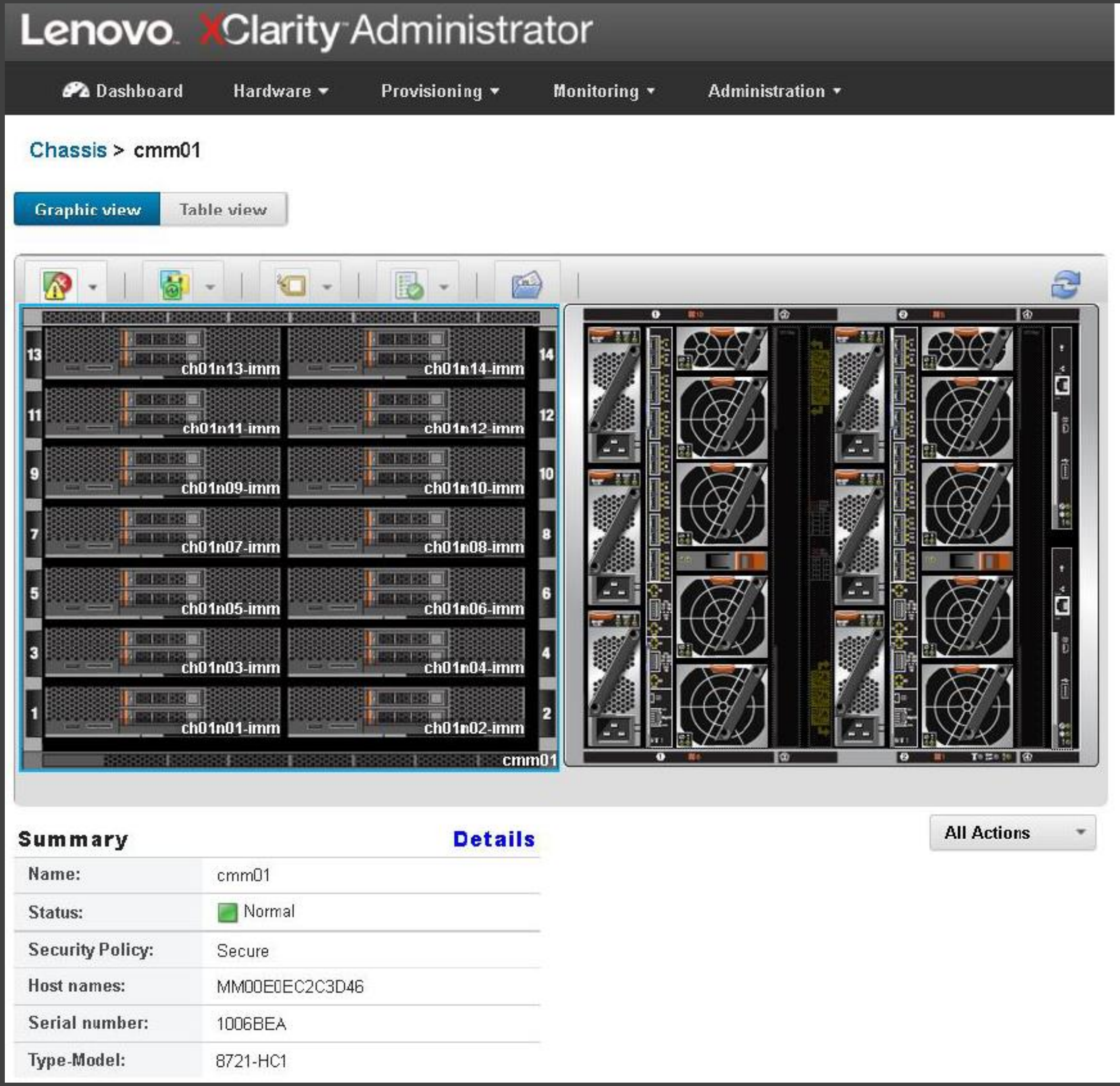
2017 Lenovo. All rights reserved..

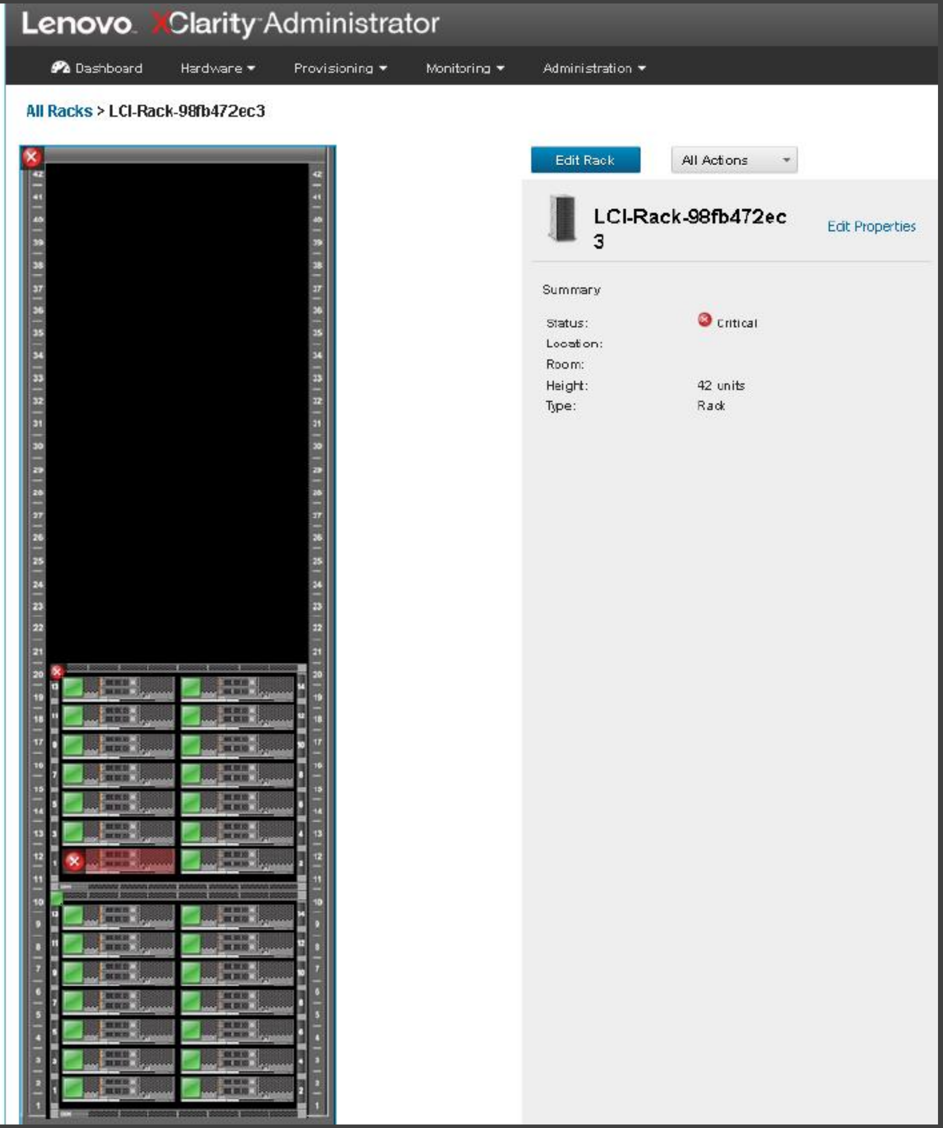
Lenovo Integrated HW Platform in 2 Form Factors





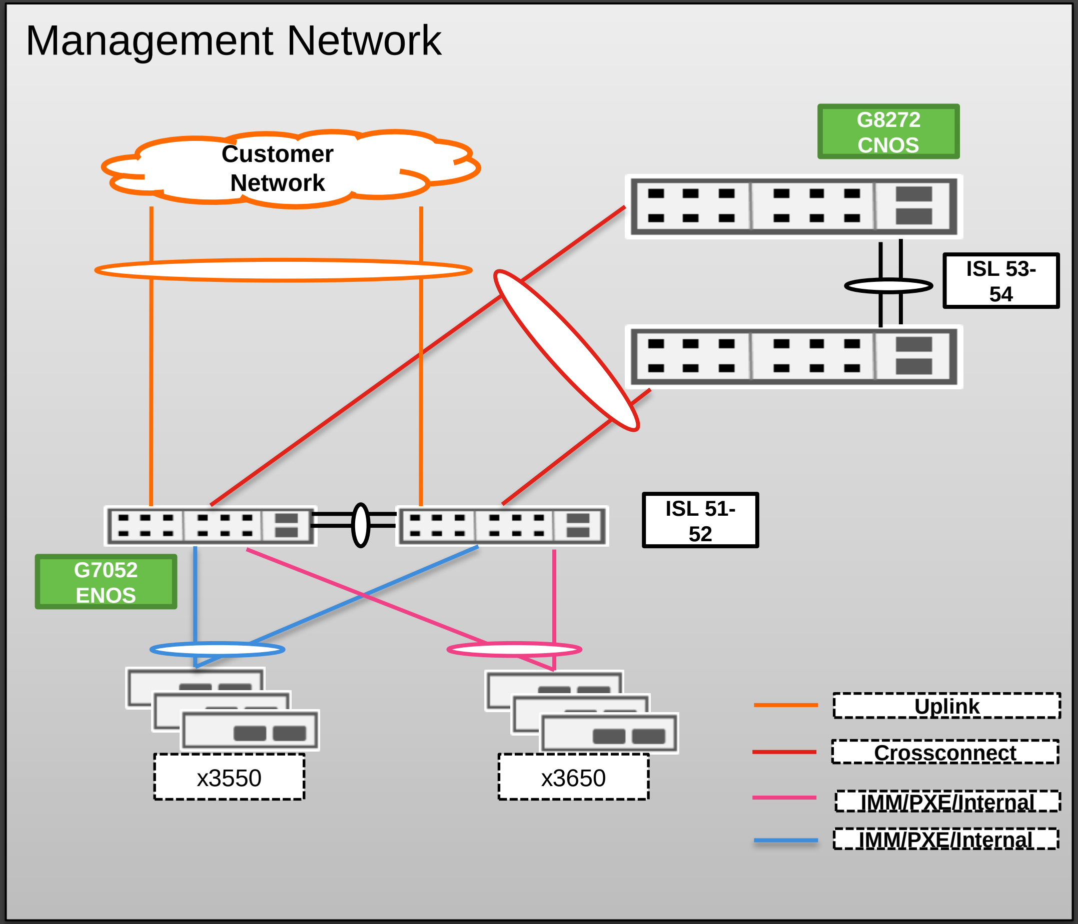
Lenovo HW Management Platform - XClarity

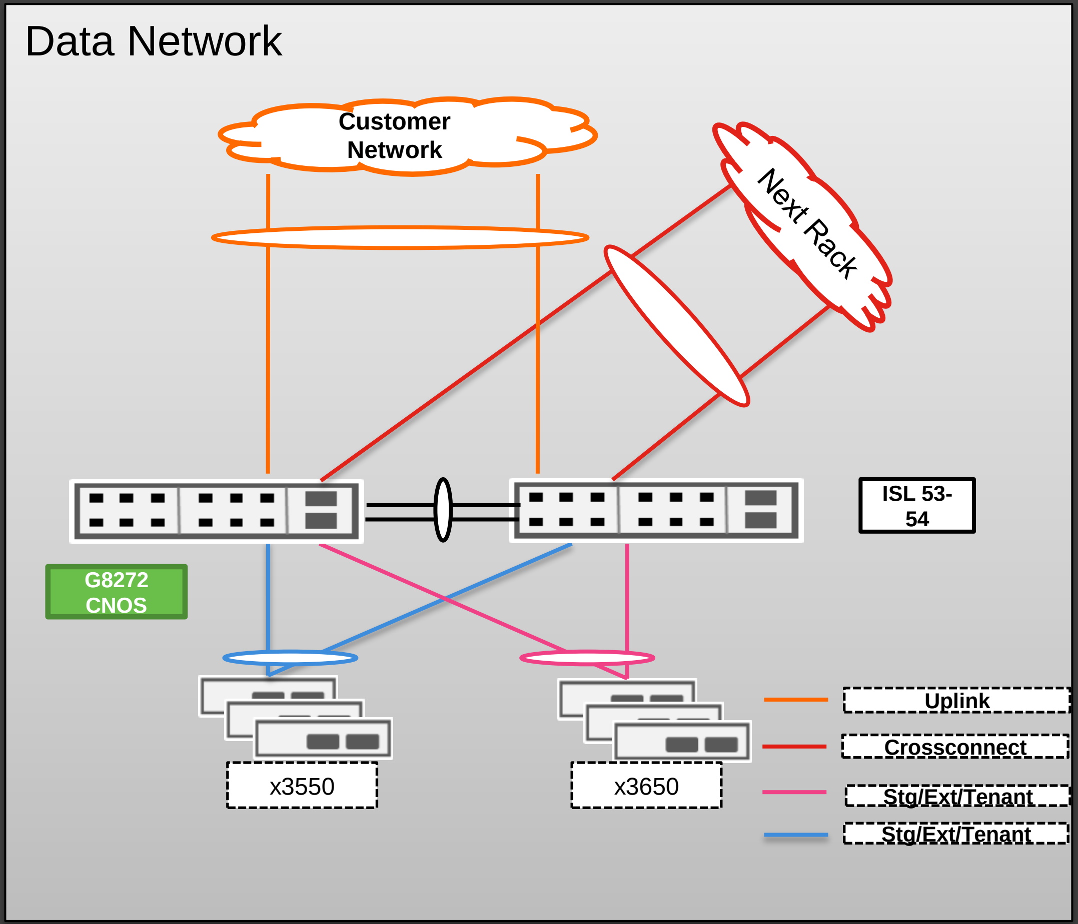




2017 Lenovo. All rights reserved..

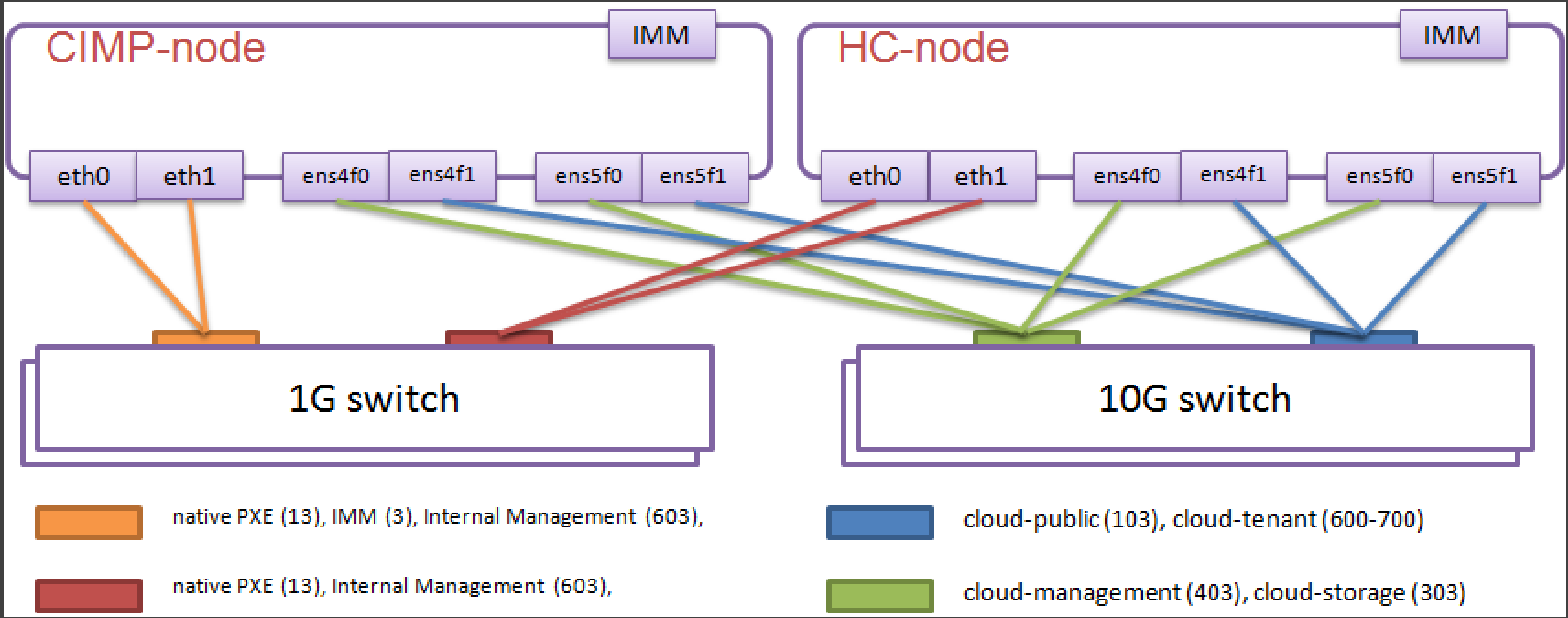
Network Topology





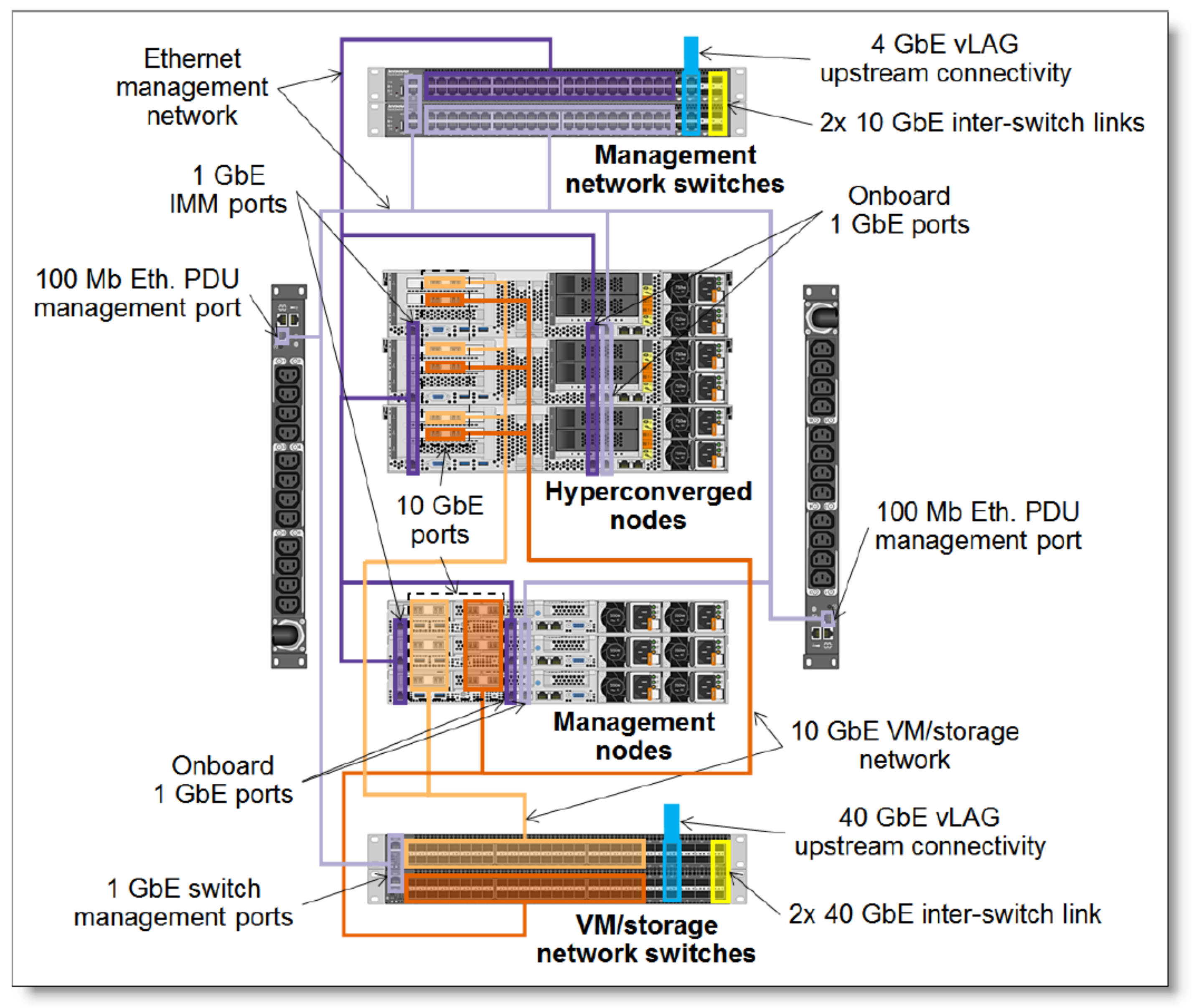
2017 Lenovo. All rights reserved..

Virtualized Control Plane Networking Resiliency



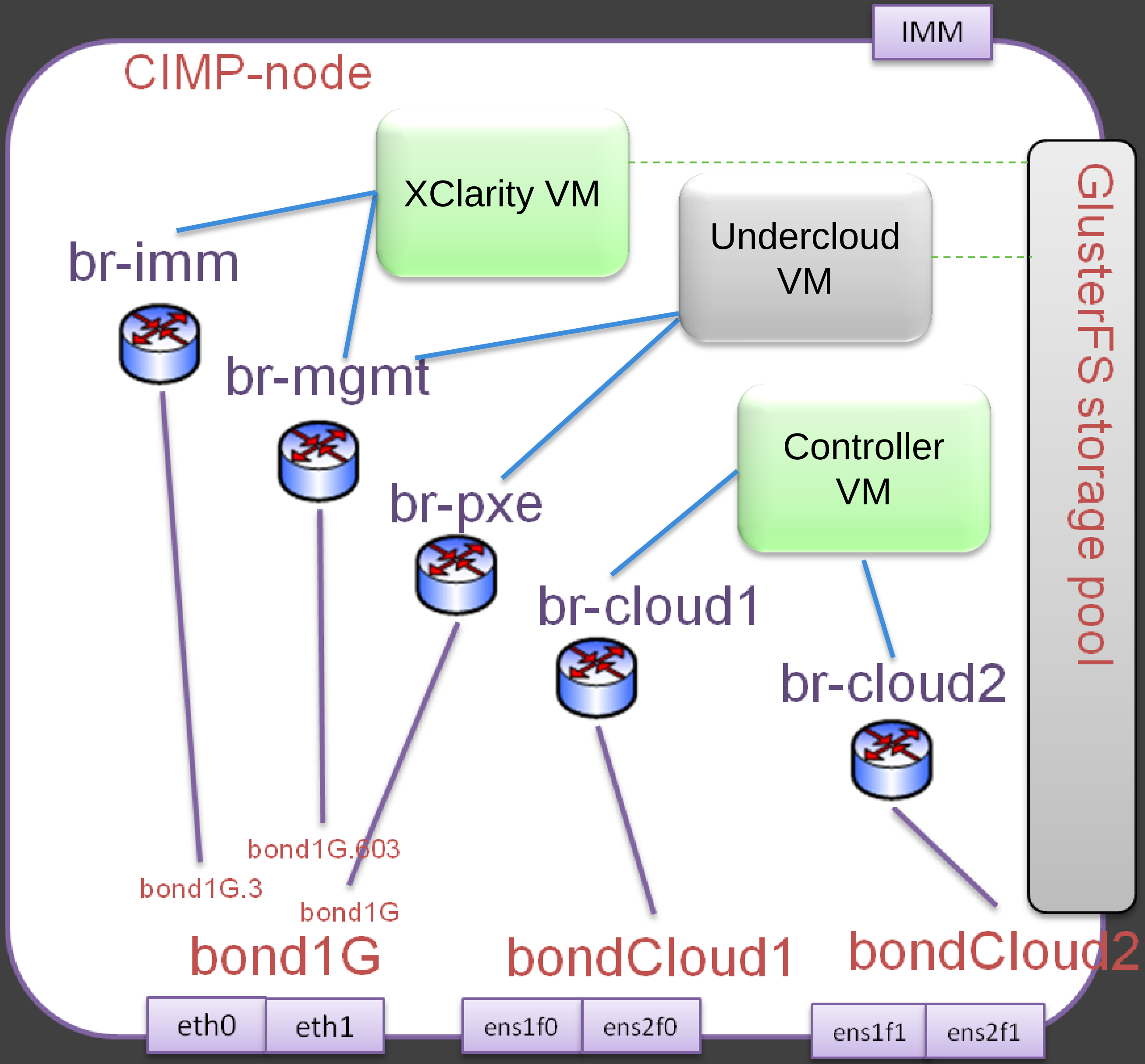
2017 Lenovo. All rights reserved..

Physical Network Setup and Resiliency



2017 Lenovo. All rights reserved..

Virtualized Control Plane Networking



2017 Lenovo. All rights reserved..

Virtualized Control Plane Storage

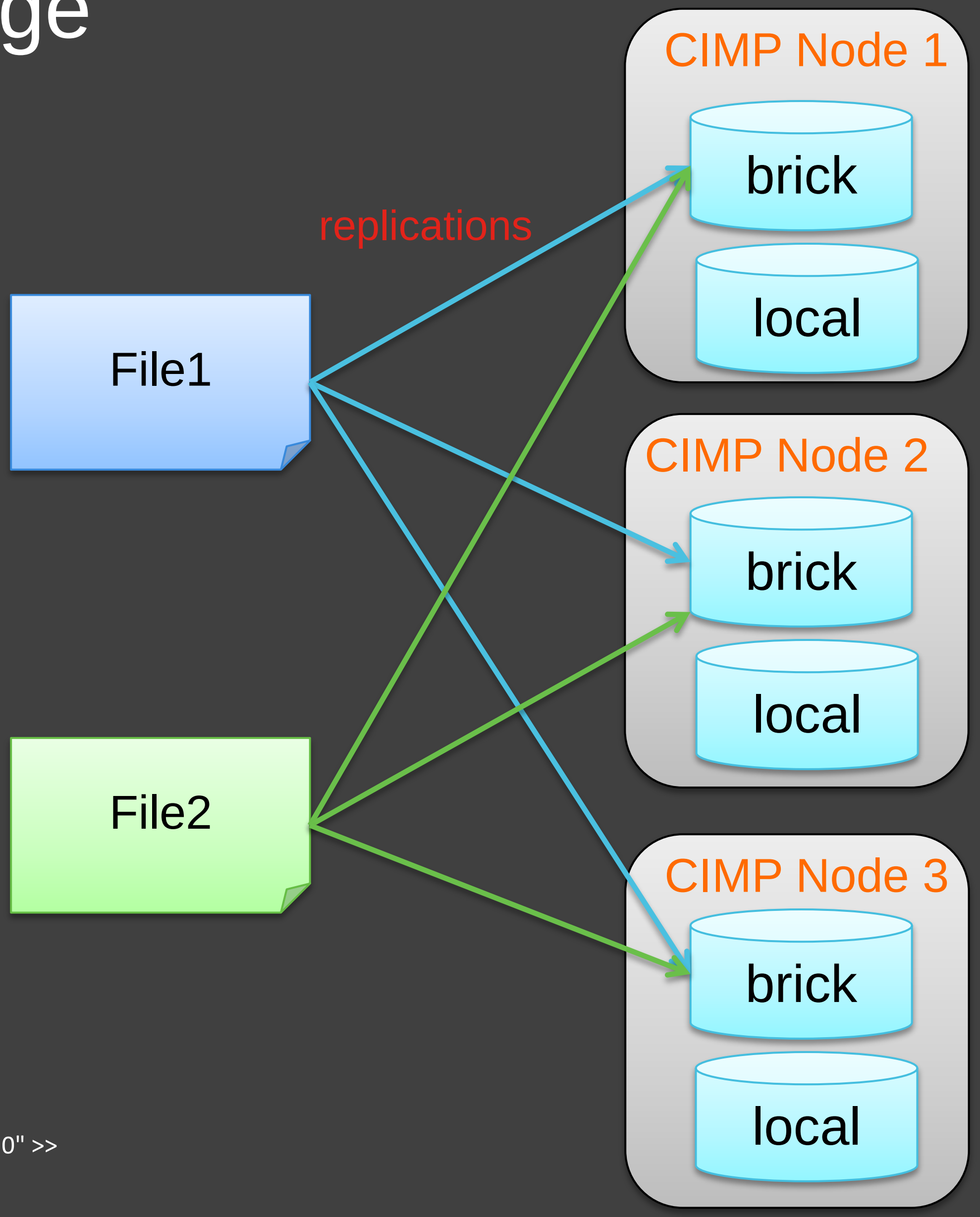
1. GlusterFS Installation
- # yum update –y
- # yum install glusterfs-server
- # systemctl enable glusterd
- # systemctl start glusterd
- # systemctl status glusterd
- Configure firewall to enable traffic on ports used by gluster

2. Build XFS bricks
- # pvcreate /dev/vdb
- # vgcreate vg_gluster /dev/vdb
- # lvcreate -L 1000G -n brick1 vg_gluster
- # mkfs.xfs /dev/vg_gluster/brick1
- # mkdir -p /bricks/brick1
- # mount /dev/vg_gluster/brick1 /var/bricks/images
- Add the following line at end of /etc/fstab:
- /dev/vg_gluster/brick1 /bricks/brick1 xfs defaults 0 0

3. Configure trusted pool
- **#gluster peer probe cimp-node2**
- **#gluster peer probe cimp-node3**
- **#gluster peer status**

4. Create GlusterFS volumes
- **# sudo gluster volume create vol-1 replica 3 cimp-node1:/var/bricks/running cimp-node2:/var/bricks/running cimp-node3:/var/bricks/running**
- **# sudo gluster volume start vol-1**
- **Confirm the Gluster volume running:**
- **# sudo gluster volume info all**

5. Use the GlusterFS volume as a shared storage pool
- # sudo mkdir –p /var/images/running
- # sudo mount -t glusterfs cimp-node1:/vol-1 /var/images/running
- # echo "127.0.0.1:vol-1 /var/images/running glusterfs defaults,_netdev,noauto,x-systemd.automount 0 0" >> /etc/fstab
- # sudo setsebool –P virt_use_fusefs 1



2017 Lenovo. All rights reserved..

Virtualized Controller VM Definition

```
<domain type="kvm" id="1">
 <name>controller1</name>
 <memory unit="GB">64</memory>
...
 <os>
  <type machine="pc" arch="x86_64">hvm</type>
  <boot dev="network"/>
  <boot dev="hd"/>
...
 <devices>
...
  <disk device="disk" type="file">
   <driver type="qcow2" name="qemu"/>
   <source file="/var/images/controller/controller1_os.qcow2"/>
...
  </disk>
  <disk device="disk" type="file">
   <driver type="qcow2" name="qemu"/>
   <source
file="/var/images/controller/controller1_mongo.qcow2"/>
...
  </disk>
<interface type="bridge">
   <source bridge="br-pxe"/>
   <target dev="vnet0"/>
...
  </interface>
  <interface type="bridge">
   <source bridge="br-cloud1"/>
   <target dev="vnet1"/>
...
  </interface>
  <interface type="bridge">
   <source bridge="br-cloud2"/>
   <target dev="vnet2"/>
...
  </interface>
 </devices>
</domain>
```

virsh create controller1.xml

2017 Lenovo. All rights reserved..

Virtualized OpenStack Controllers Considerations

- Power Control
  - Undercloud to use virsh to control power management of other nodes
  - Pre Pike release pxe_ipmitool => pxe_ssh
  - Starting from Pike transition to virtualbmc
- High Availability
  - Core services – Galera, RabbitMQ, Redis
  - Active-Passive services – Cinder-Volume service
  - SystemD services – independent and able to withstand service interruption
  - Isolating a faulty node to protect a cluster and its resources
  - Pacemaker + Shoot-The-Other-Node-In-The-Head
  - Fencing agent – fence_ipmilan / fence_xvm, fence_virt

2017 Lenovo. All rights reserved..

Lenovo

Hyperconverged Compute Node

- Deploy Ceph OSD alongside nova-compute on the same node
- Tuning - Compute shares memory and CPU with Ceph OSD
  - Nova.conf – set aside CPU and memory for Ceph OSD
    - cpu_allocation_ratio / reserved_host_memory_mb
  - Ceph.conf – balance system resources needed for ceph recovery and rebalancing and guest workloads
    - Osd_recovery_op_priority, osd_recovery_max_active, osd_max_backfilles
  - Tune overal system performance - throughput-performance
  - NUMA Pinning of Ceph OSD processes <=> NIC PCIe Slots
- Operations
  - Nova / Compute operational worklows must account for Ceph OSD processes and vice versa

2017 Lenovo. All rights reserved.

Testing using Rally

Tenant creation with user

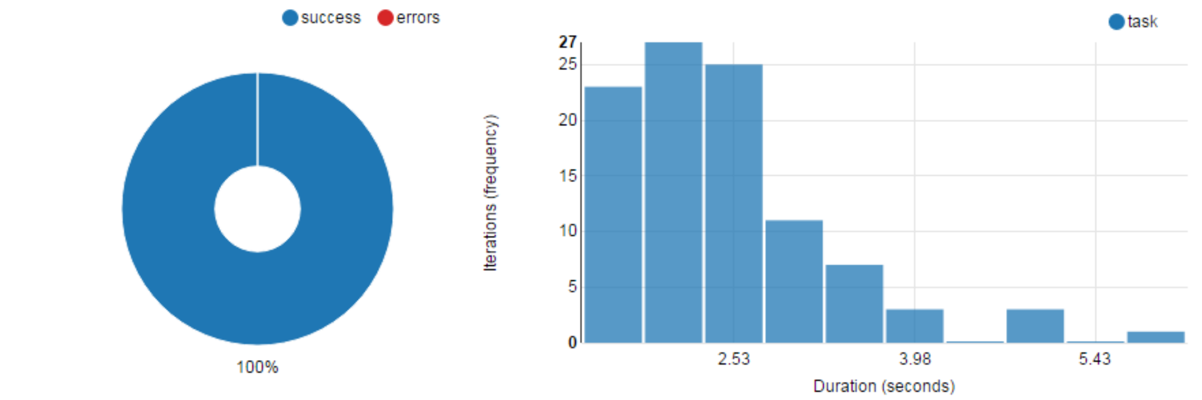
❖ Conditions

20-concurrency tests (create a tenant with user) and totally accomplish 100 times of tests

❖ Overall Results

All of the tasks were successfully finished. Majority of them accomplished the tasks within 4 seconds and the maximum consumed 6 seconds

Distribution



Total durations

Action	Min (sec)	Median (sec)	90%ile (sec)	95%ile (sec)	Max (sec)	Avg (sec)	Success	Count
keystone_v2.create_tenant	0.201	0.504	1.853	2.349	3.872	0.859	100.0%	100
keystone_v2.create_users	0.829	1.218	1.849	2.167	2.465	1.351	100.0%	100
keystone_v2.create_user	0.06	0.124	0.226	0.289	0.63	0.144	100.0%	100
keystone_v2.create_user (2)	0.056	0.114	0.275	0.336	0.679	0.149	100.0%	100
keystone_v2.create_user (3)	0.058	0.106	0.166	0.213	0.622	0.124	100.0%	100
keystone_v2.create_user (4)	0.061	0.114	0.223	0.259	0.529	0.135	100.0%	100
keystone_v2.create_user (5)	0.057	0.102	0.181	0.217	0.721	0.134	100.0%	100
keystone_v2.create_user (6)	0.054	0.107	0.179	0.2	0.626	0.125	100.0%	100
keystone_v2.create_user (7)	0.058	0.109	0.244	0.286	0.629	0.142	100.0%	100
keystone_v2.create_user (8)	0.057	0.111	0.193	0.22	0.64	0.128	100.0%	100
keystone_v2.create_user (9)	0.06	0.112	0.23	0.312	0.611	0.143	100.0%	100
keystone_v2.create_user (10)	0.059	0.103	0.187	0.264	0.621	0.126	100.0%	100
total	**1.079**	**2.048**	**3.099**	**3.771**	**5.918**	**2.21**	**100.0%**	**100**

2017 Lenovo. All rights reserved..

Testing using Rally

Snapshot creation and deletion

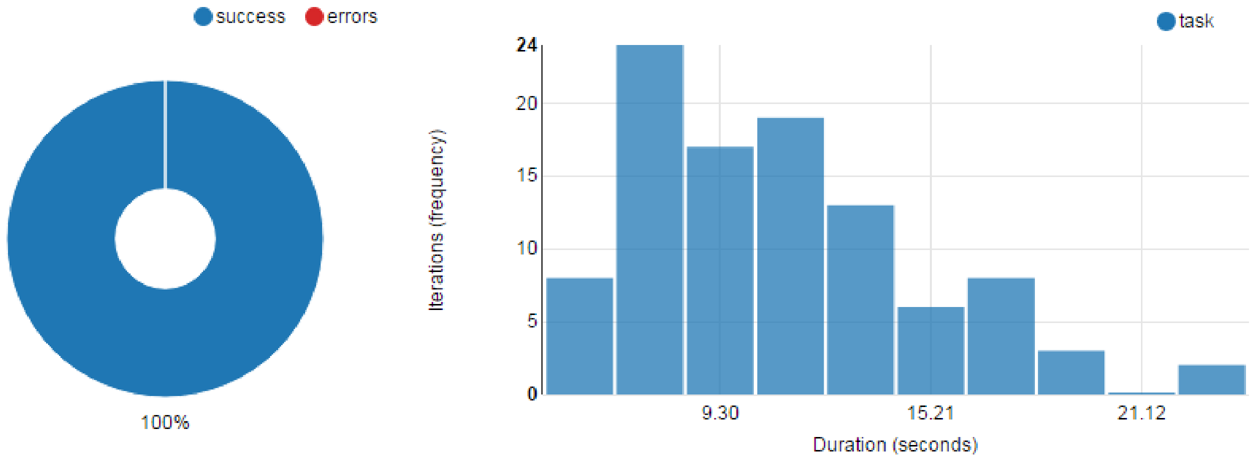
- ❖ **Conditions**

10-concurrency tests (create a snapshot and delete it, set it to 10 since the snapshot limitation is 10) and totally accomplish 100 times of tests

- ❖ **Overall Results**

All of the tasks were successfully finished. Majority of them accomplished the tasks within 17 seconds and the maximum consumed 23 seconds

Distribution



Total durations

Action	Min (sec)	Median (sec)	90%ile (sec)	95%ile (sec)	Max (sec)	Avg (sec)	Success	Count
cinder.create_snapshot	2.688	5.108	9.482	9.726	11.658	5.404	100.0%	100
cinder.delete_snapshot	2.341	4.872	11.55	15.623	19.782	6.415	100.0%	100
total	**3.397**	**9.899**	**16.465**	**16.786**	**23.089**	**9.819**	**100.0%**	**100**

2017 Lenovo. All rights reserved..

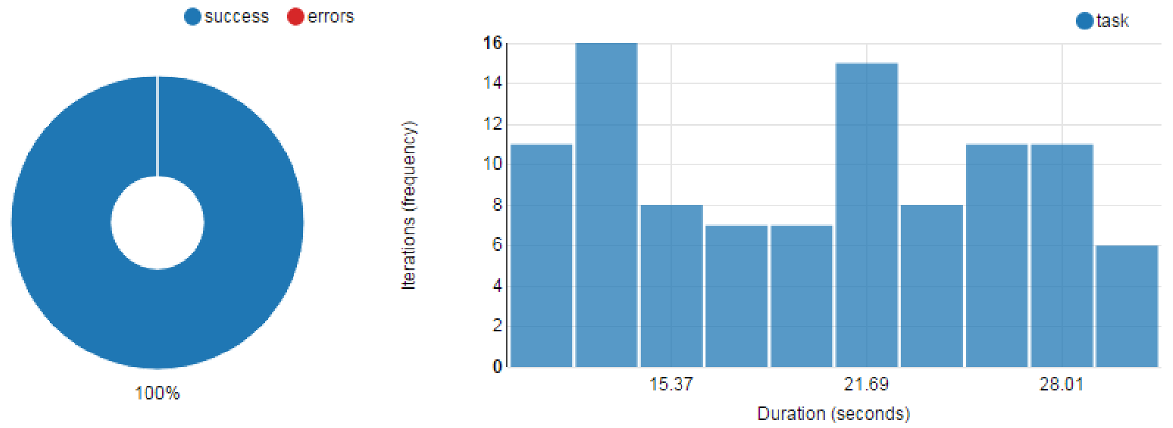
Testing using Rally

Small instance creation and deletion

- **Conditions**
  1. Image: TestVM (around 30M)
  2. Flavor: 1 core + 512M RAM+ 20G Disk
  3. 20-concurrency tests (create a instance and then delete it) and totally accomplish 100 times of tests
- **Overall Results**

All of the tasks were successfully finished. Majority of them accomplished the tasks within 28 seconds and the maximum consumed 30 seconds

Distribution



Total durations

Action	Min (sec)	Median (sec)	90%ile (sec)	95%ile (sec)	Max (sec)	Avg (sec)	Success	Count
nova.boot_server	6.896	14.285	20.262	21.993	24.495	14.207	100.0%	100
nova.delete_server	2.543	5.443	7.661	7.748	9.836	5.759	100.0%	100
total	**9.044**	**19.719**	**27.005**	**28.105**	**30.118**	**18.967**	**100.0%**	**100**

2017 Lenovo. All rights reserved..

Testing using Rally

Large instance creation and deletion

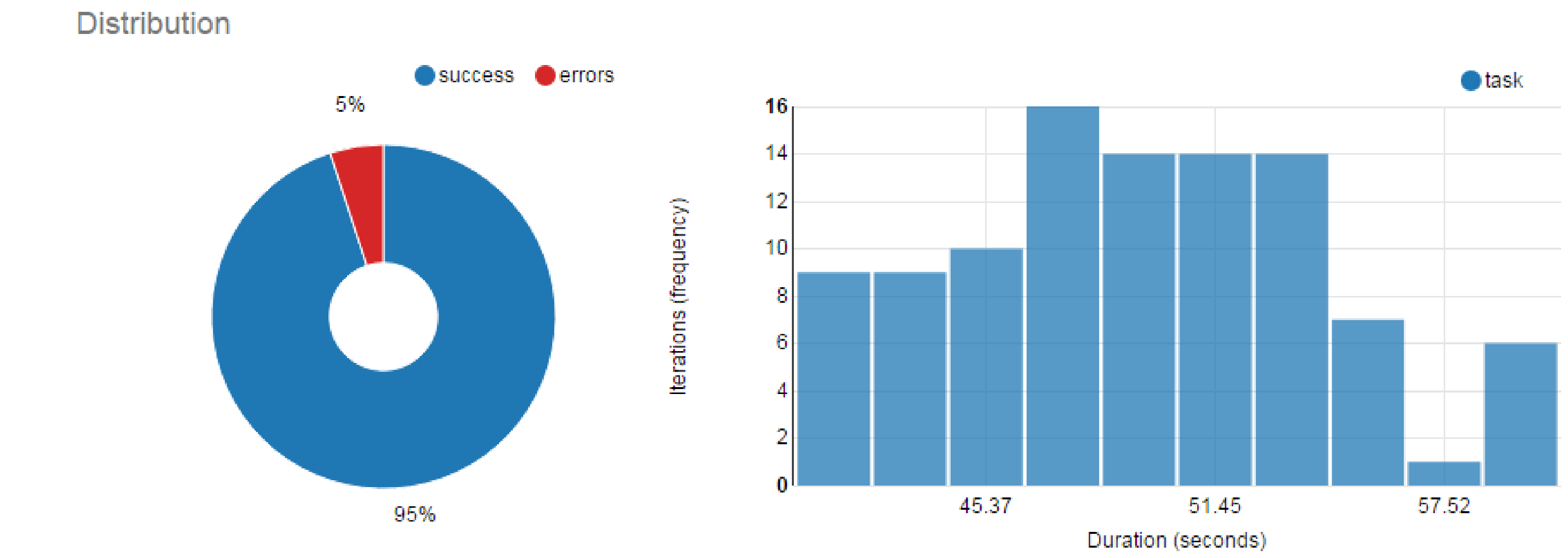
❖ Conditions:

1. Image: Redhat 7.2 (around 4G)
2. Flavor: 8 core + 19G RAM+ 100G Disk
3. 5-concurrency tests (create a instance and then delete it, set concurrency to 5 due to quota limitation) and totally accomplish 100 times of tests

❖ Overall Results

Almost all of the tasks were successfully finished, some of them failed because the RAM of coexisted instances exceeded the quota limitation.

Majority of them accomplished the tasks within 58 seconds and the maximum consumed almost 60 seconds



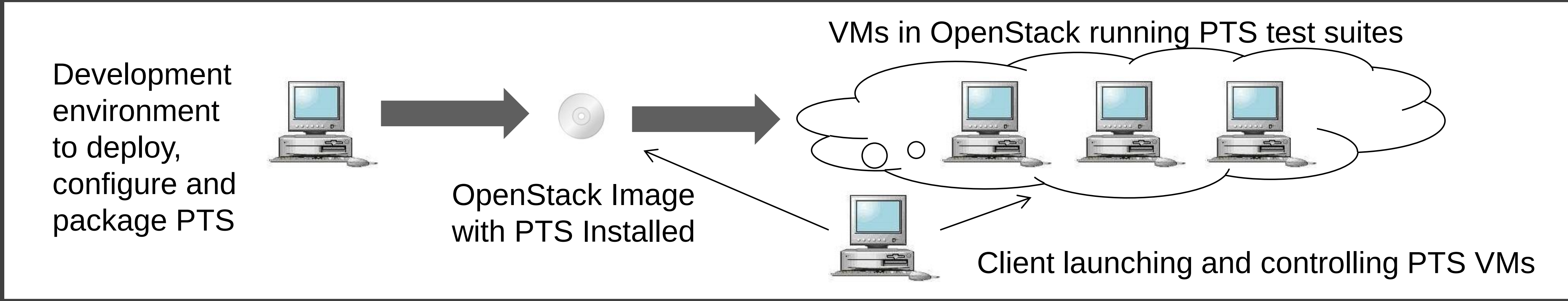
Total durations

Action	Min (sec)	Median (sec)	90%ile (sec)	95%ile (sec)	Max (sec)	Avg (sec)	Success	Count
nova.boot_server	34.452	39.671	42.408	43.279	44.686	39.623	95.0%	100
nova.delete_server	4.823	9.523	16.056	16.414	18.686	10.215	100.0%	95
total	**39.298**	**48.69**	**54.495**	**57.983**	**59.544**	**48.838**	**95.0%**	**100**

2017 Lenovo. All rights reserved..

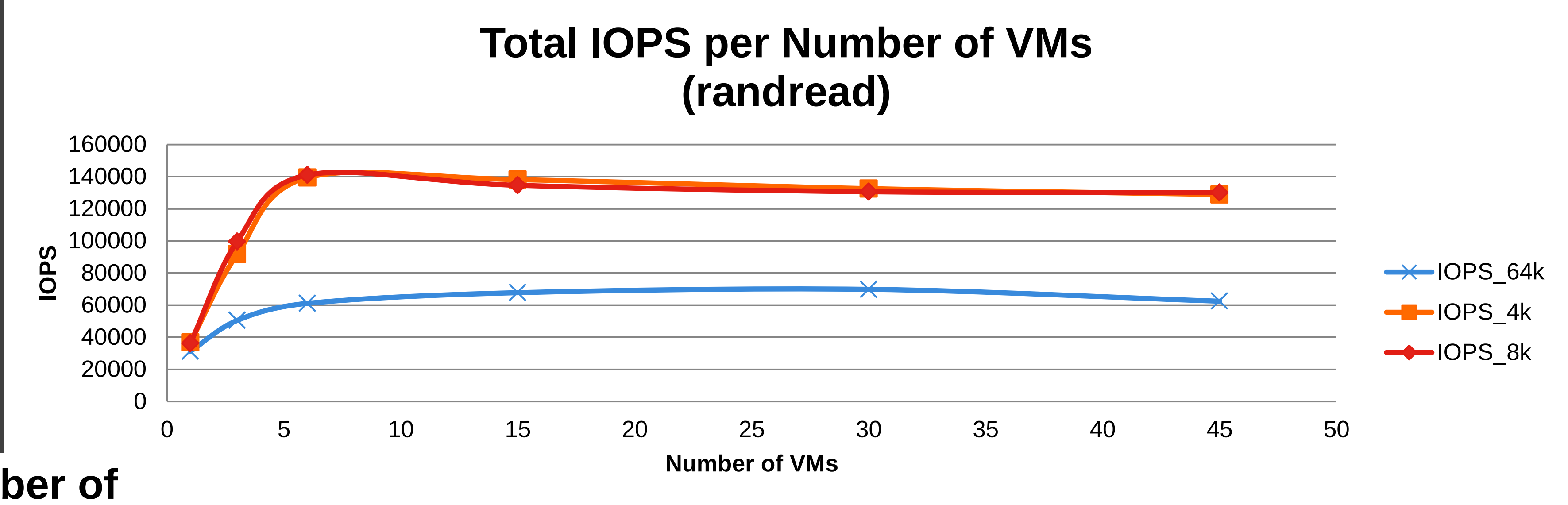
Testing using Phoronix Test Suite

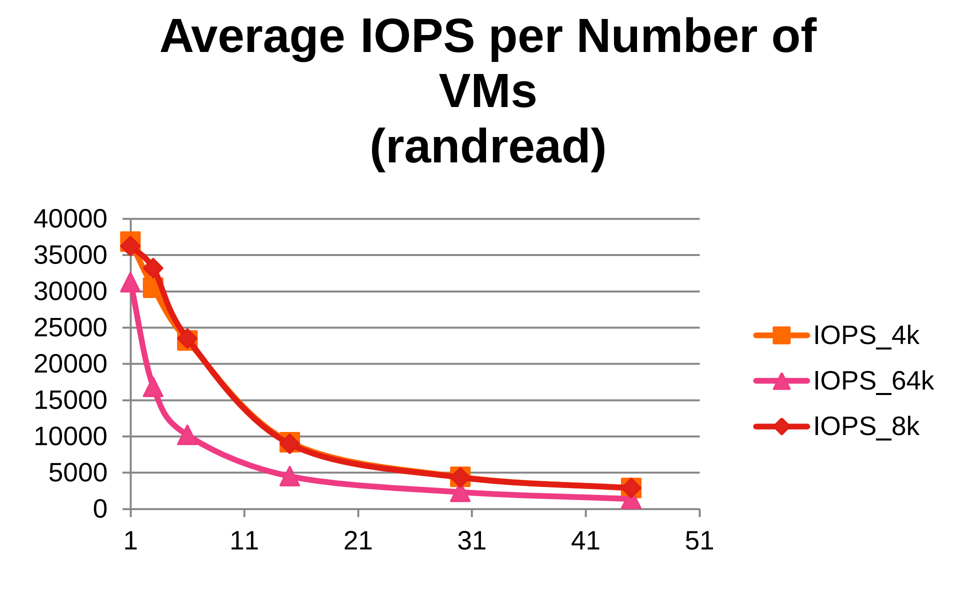
- Open source testing suites including test development framework, test runner, management and reporting
- Benchmark catalog at OpenBenchmarking.org is comprehensive 984 tests & suite
- Many real-world workloads wrapped for benchmarking
- Corpus of shared results for comparison / initial settings
- Easy to extend and share



2017 Lenovo. All rights reserved..

Testing using Phoronix Test Suite





Number of VMs

1 pts/fio-1.9.0 - randread libaio 0 1 8k 1 64 /

3 pts/fio-1.9.0 - randread libaio 0 1 8k 1 64 /

6 pts/fio-1.9.0 - randread libaio 0 1 8k 1 64 /

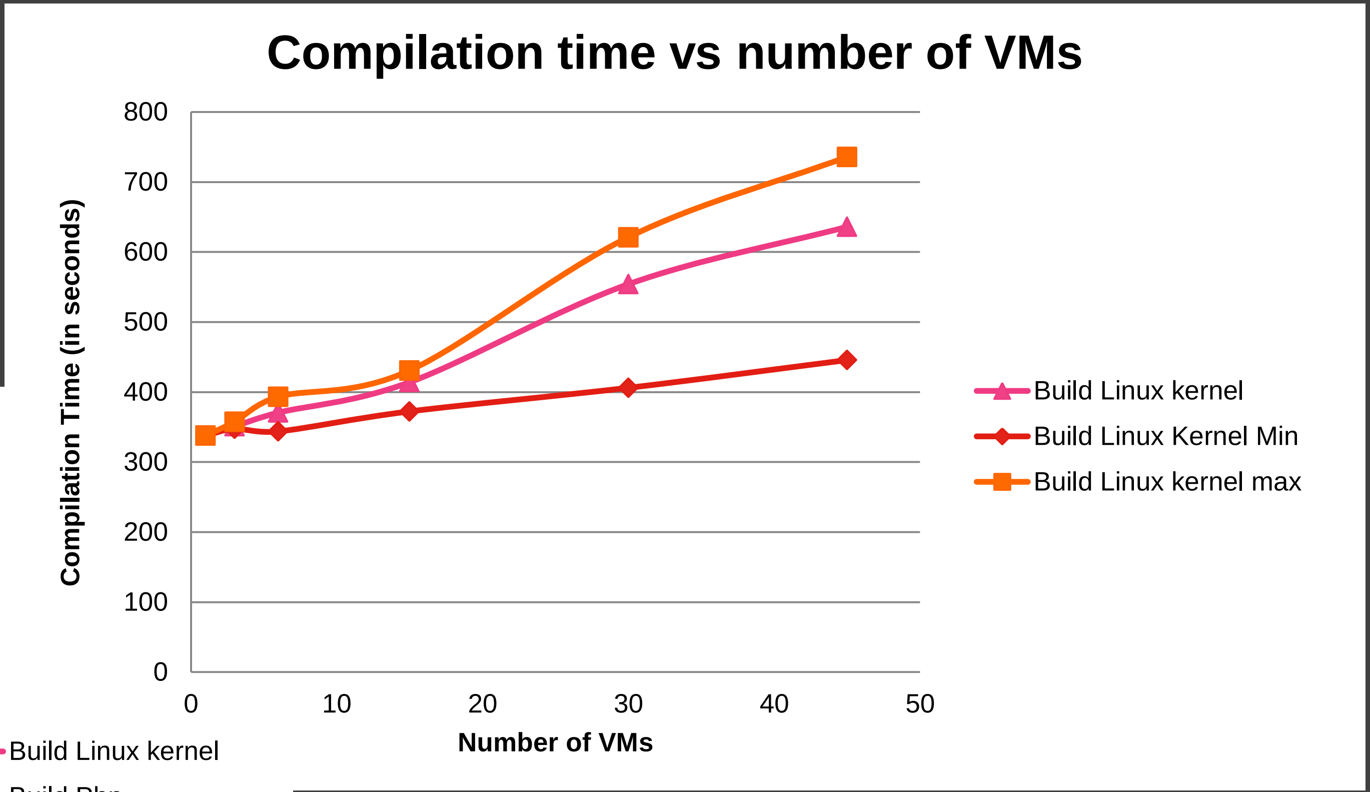
15 pts/fio-1.9.0 - randread libaio 0 1 8k 1 64 10g /

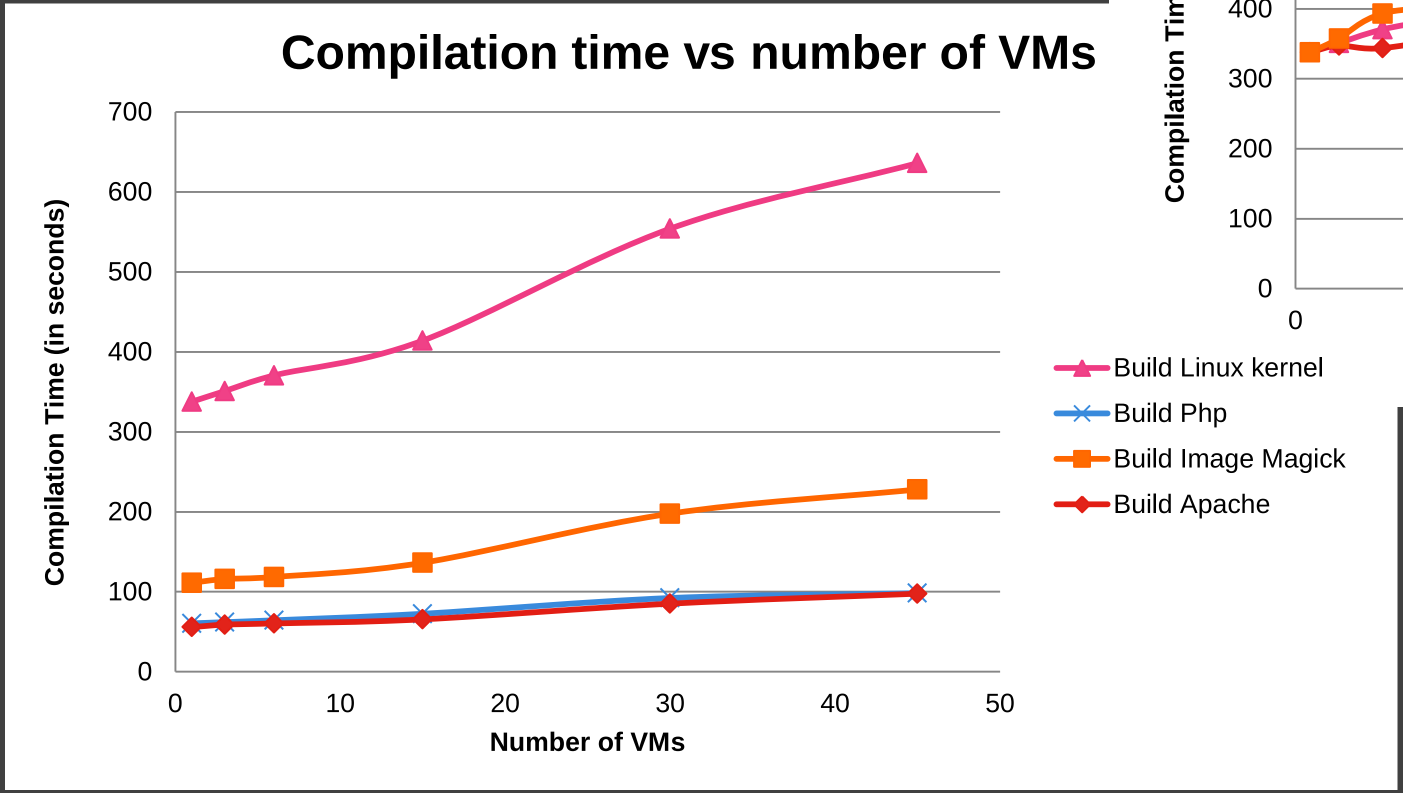
30 pts/fio-1.9.0 - randread libaio 0 1 8k 1 64 10g /

45 pts/fio-1.9.0 - randread libaio 0 1 8k 1 64 10g /

2017 Lenovo. All rights reserved..

Testing using Phoronix Test Suite





pts/build-apache-1.5.1 -

pts/build-imagemagick-1.7.2 -

pts/build-linux-kernel-1.7.0 -

pts/build-php-1.3.1 -

2017 Lenovo. All rights reserved..

Virtualized Control Plane Operations

- Similar concept as containers but ready to use today
- Reuse existing expertise in operating virtualized infrastructure
- Using the Host HA capabilities in addition to the app level HA
- Better resource utilization for resource limited deployments, can pack more components on the same HW
- VM snapshots before patching
- VM migration before HW upgrades
- VM resizing to provide additional resources
- VM restore after failure
  - Implication on VM structure and services deployments, “OS/services” stateless disk and data on separate disk
  - Should be combined with and is not replacement for backup strategy

2017 Lenovo. All rights reserved..

Potential Future Work

- Reusable VM images for Undercloud and Overcloud
- Containerized Control Plane

2017 Lenovo. All rights reserved..

Resources

- https://access.redhat.com/articles/2922421
- https://access.redhat.com/articles/2360321
- https://access.redhat.com/articles/2861641
- https://access.redhat.com/documentation/en-us/red_hat_openstack_platform/10/html-single/understanding_red_hat_openstack_platform_high_availability/
- http://docs.openstack.org/developer/performance-docs/
- http://tripleo.org/environments/virtualbmc.html
- https://docs.openstack.org/developer/kolla-kubernetes/deployment-guide.html
- https://www.openstack.org/assets/presentation-media/Chasing-1000-nodes-scale.pdf
- https://www.phoronix-test-suite.com/
- https://openbenchmarking.org

2017 Lenovo. All rights reserved..

thanks.