



OpenStack ハイパーコンバージドインフラ(HCI)の ポイントとベンチマーク結果考察

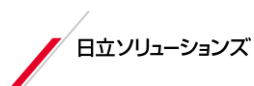
Version 1.1

2017 Jun.

Takehiro Kudou (Hitachi Solutions, Ltd.)

Hirofumi Sasaki (Red Hat K.K.)

Masayoshi Hibino (Dell Inc.)



目 次

1.	はじめに	4
1.1	概要	4
1.2	背景	4
1.3	目的	4
2.	調査環境	5
3.	調査方法	6
3.1	fio パラメータ	6
4.	構築・設定ポイント	8
4.1	構築手順概要	8
4.2	ノード追加・削除	10
4.3	NUMA 設定	10
4.4	Network 設計	10
4.5	VM イメージのフォーマット形式	10
5.	結果と考察	10
5.1	Block Size 別の傾向	12
5.2	Compute ノード台数別による傾向	15
5.3	Random Write	18
5.4	今回の検証における Random Read	18
5.5	前回の検証との比較における Random Read	18
6.	まとめ	22
6.1	構築・設定	22
6.2	ベンチマーク結果から得られた指針	22
6.3	懸念事項	23
7.	参考・関連文献	24
8.	付録	25
8.1	OpenStack 及び Ceph の構築に使用した設定ファイル	25

略号の説明

略号	意味
BS	<u>B</u> lock <u>S</u> ize
CoW	<u>C</u> opy <u>o</u> n <u>W</u> rite
HCI	<u>H</u> yper <u>C</u> onverged <u>I</u> nfrastructure
H/W	<u>H</u> ard <u>w</u> are
I/O	<u>I</u> nput / <u>O</u> utput
IOPS	<u>I</u> nput / <u>O</u> utput <u>P</u> er <u>S</u> econd
NIC	<u>N</u> etwork <u>I</u> nterface <u>C</u> ard
NUMA	<u>N</u> on- <u>U</u> niform <u>M</u> emory <u>A</u> ccess
OSP-d	Red Hat <u>O</u> pen <u>S</u> tack <u>P</u> latform <u>d</u> irector
RHEL	<u>R</u> ed <u>H</u> at <u>E</u> nterprise <u>L</u> inux
RH-Ceph	<u>R</u> ed <u>H</u> at <u>C</u> eph Storage
RH-OSP	<u>R</u> ed <u>H</u> at <u>O</u> penStack <u>P</u> latform
SDS	<u>S</u> oftware <u>D</u> efined <u>S</u> torage
VM	<u>V</u> irtual <u>M</u> achine(仮想マシン) 本書では OpenStack におけるインスタンスを示す。

1. はじめに

1.1 概要

Red Hat 社の商用 OpenStack ディストリビューションである Red Hat OpenStack Platform(以下 RH-OSP)と、同じく Red Hat 社の商用 Ceph ディストリビューションである Red Hat Ceph Storage(以下 RH-Ceph)を組み合わせ、Hyper Converged Infrastructure(以下 HCI)を構築する仕組みが Tech Preview として実装された。OSCA 技術検討会 OpenStack 分科会において、本組み合わせによる HCI について評価を実施した。

1.2 背景

OpenStack は成熟度向上と市場ニーズの高まりにより、検討フェーズから導入フェーズへと推移し始めている。しかしながら、その自由度の高さから設計・構築・運用には高度な技術力が必要であり、一から始めるには、少々敷居が高いという課題がある。ディストリビュータが出している各種 Deploy Tool を用いることで構築の敷居を下げることはできるが、やはり自由度の高さと対応 Hardware の多さから設計が難しい。

一方、Storage を Software で実装するという Software Defined Storage(以下 SDS)、Hypervisor と SDS を同一のノードで実装するという HCI がともに盛り上がりを見せている。前者は柔軟かつスケラブルなストレージを汎用ハードウェア上で実装できること、後者はノード拡張と管理対象ノードの削減が可能ながことがメリットである。

OSCA 技術検討会では前回 RH-OSP と RH-Ceph を組み合わせた OpenStack×SDS 検証(Ref.[1])を実施したが、今般 RH-OSP 10 の TechPreview 機能として、RH-Ceph を用いた HCI 構成が可能となったことから、これをさらに進化させた OpenStack×SDS による HCI 検証を実施した。

Hardware Storage、SDS、HCI の各構成の概念を図 1-1 に示す。

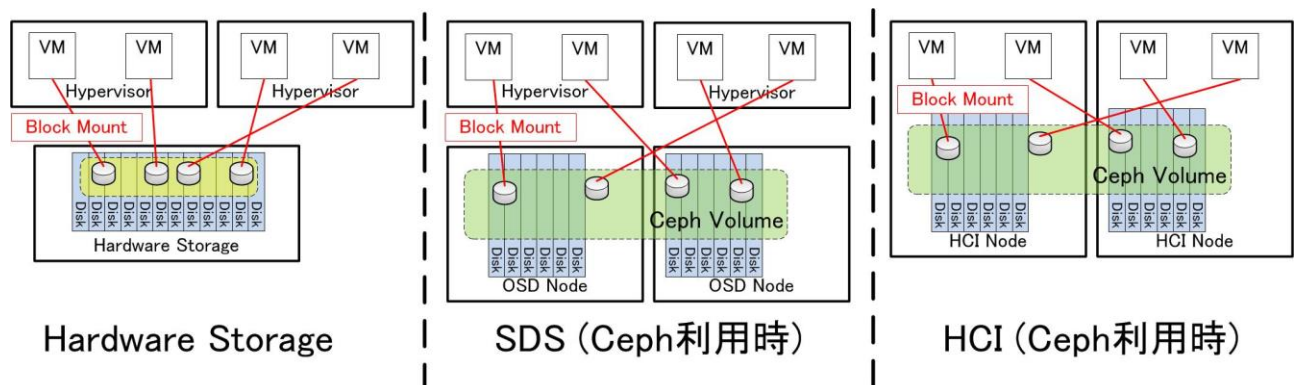


図 1-1 構成概念図

1.3 目的

本評価は、RH-OSP と RH-Ceph を組み合わせた HCI 環境について、構築手順、設計ポイント、ノード追加・削除、ストレージ性能等について検証し、H/W 設計、特に OSD Disk の設計指針を得ることを目的としている。

2. 調査環境

本評価では、表 2-1 に示すサーバ(DELL EMC PowerEdge C シリーズ)を 12 台用意し、検証を行った。なお、環境構築については Hyper-Converged Red Hat OpenStack Platform 10 and Red Hat Ceph Storage 2(Ref.[2])を参考とした。

表 2-1 使用機材

機材	仕様
	CPU : Intel Xeon E5-2620×2
	NIC : LOM 1Gbps×4、 Intel X520 dual port 10Gbps NIC×1
	OS : RHEL 7.3
	App : RH-OSP 10、Red Hat Ceph Storage 2.0

各ノードの役割別の詳細スペックを表 2-2 に示す。

表 2-2 ノード役割

#	役割	台数	MEM	HDD
1	Controller ノード - OpenStack Controller - Ceph Monitor	3	64GB	2TB SATA 7.2Krpm ×2(OS、App)
2	Compute ノード - OpenStack Compute - Ceph OSD	8	64GB	300GB SAS 10Krpm ×2 (OS、App) 600GB SAS 10Krpm ×3 (OSD) 300GB SSD (Journal) [各 OSD に 80GB の Journal Volume]
4	Director ノード - OpenStack、Ceph の Deploy	1	64GB	2TB SATA 7.2Krpm ×2(OS、App)

上記の各ノードについて、図 2-1 のとおりに構成した。環境構築には、RH-OSP director(以下 OSP-d)を用いて構築した。

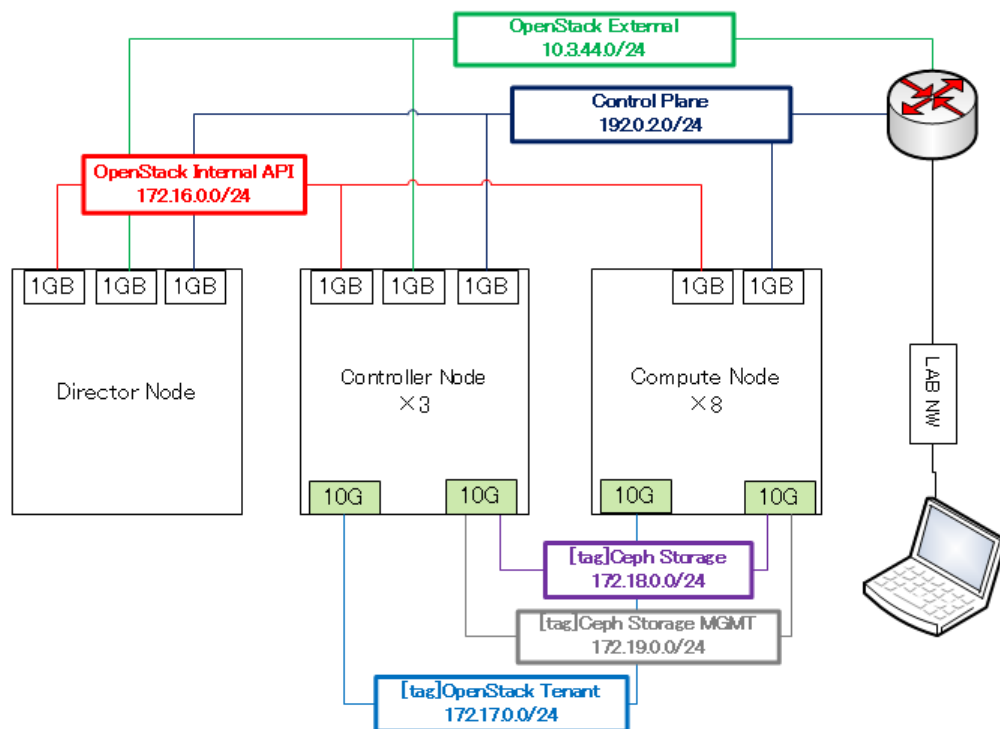


図 2-1 検証環境

3. 調査方法

設計ポイント、構築手順、ノード追加・削除については、(Ref.[2])記載の手順に沿って確認を実施した。IOPS・スループット性能は、下記手順を繰り返して調査した。

- (1) Compute ノード上に、合計 80 のインスタンス(以下 VM)を作成
- (2) 全 VM に fio-2.1.7 をインストール
- (3) ベンチマーク実行用 VM を作成し Floating IP を割り当て
- (4) ベンチマーク実行用 VM からパスワードレスで ssh 接続できるよう設定
- (5) ベンチマーク実行用 VM から ssh にて各 VM 上で fio を同時実行
- (6) ベンチマークデータ取得後、VM 及びボリュームを全て破棄。Compute ノードを 1 台減らし、(1)へ戻る

3.1 fio パラメータ

fio は、Compute ノードの台数、Read・Write、Block Size、job 数の値を変化させ、2 分間ベンチ実行、2 分間スリープを繰り返しながら各 3 回計測した。

fio 実行時の変数は以下の通りである。

```
ssh cloud-user@[instance] fio -rw=[read/write] -size=1G -ioengine=libaio -iodepth=4 -invalidate=1
-direct=1 -name=test.bin -runtime=120 -bs=[BlockSize]k -numjobs=[Job 数] -group_reporting > [file
name] & . . .
```

- [instance] : VM の IP Address。
- [read/write] : randomread | randomwrite 【×2】
- [BlockSize]: 4 | 16 | 32 | 64 | 128 【×5】
- [Job 数]: 1 | 4 | 8 【×3】
- [file name] : 出力ファイル名

また上記すべてに対し、下記パターン分のベンチマークを実施した。

- インスタンス数 : ノード数に関わらず 80 インスタンス 【×80】
- Compute ノード数 : 4 | 5 | 6 | 7 | 8 【×5】
- Loop 回数 : 3 回 【×3】
- VM のキャッシュクリア機構組み込み有無 : 有、無 【×2】

上記により、72,000 のベンチマーク結果を得た。

さらに後述の random read 時のベンチマーク性能異常値の切り分けのため、7,200 パターンのベンチマークを追加調査し、合計 79,200 の fio ベンチマーク結果を得た。

4. 構築・設定ポイント

以下に構築・設定について、検証から得られたことについて示す。

4.1 構築手順概要

OSP-d を用いた HCI 環境構築の流れは以下のとおりである。注意すべき部分を抽出しているため、完全な手順については Red Hat 社のドキュメント(Ref.[2]、[3]、[4])を参照いただきたい。

(1) Undercloud (OSP-d)の Deploy

下記ファイルを参考に `undercloud.conf` を作成する。

```
/usr/share/instack-undercloud/undercloud.conf.sample
```

下記コマンドを実行し、Undercloud (OSP-d)を Deploy する。

```
$ openstack undercloud install
```

(2) openstack baremetal コマンドによるノード登録

ノードの IPMI アクセスのために必要な情報を下記ファイルへ登録。ノードの役割(`osd-compute-x` 等)をファイル内の `capabilities` で設定。

```
~/instackenv.json
```

下記コマンドを実行し、ノードを Undercloud (OSP-d)に登録。

```
$ openstack baremetal import --json ~/instackenv.json
```

続いて、下記コマンドを実行し、各ノードの仕様詳細(CPU コア数やメモリ搭載量等)を取得。

```
$ openstack baremetal introspection bulk start
```

下記コマンドにより、パラメータ(Root Device 指定等)を実施。

(HDD を複数搭載している場合、どの HDD に OS 及び OpenStack をインストールするか指定する必要がある)

```
$ openstack baremetal node set --property root_device='{"(略)
```

(3) yaml ファイル作成 (custom role)

OSP-d を用いて HCI 環境を構築する場合、`custom-roles.yaml` で OpenStack Compute と Ceph OSD の役割を持つ Custom Role を定義することで実現する。今回の環境では、以下の `OsdCompute` という Role を新たに定義した。

```
~/custom-templates/custom-roles.yaml
```

```
(略)
- name: OsdCompute
  CountDefault: 0
  HostnameFormatDefault: '%stackname%-osd-compute-%index%'
  ServicesDefault:
    - OS::TripleO::Services::CephOSD
    - OS::TripleO::Services::CACerts
    - OS::TripleO::Services::CephClient
    - OS::TripleO::Services::CephExternal
    - OS::TripleO::Services::Timezone
    - OS::TripleO::Services::Ntp
    - OS::TripleO::Services::Snmp
    - OS::TripleO::Services::NovaCompute
    - OS::TripleO::Services::NovaLibvirt
    - OS::TripleO::Services::Kernel
    - OS::TripleO::Services::ComputeNeutronCorePlugin
    - OS::TripleO::Services::ComputeNeutronOvsAgent
    - OS::TripleO::Services::ComputeCeilometerAgent
    - OS::TripleO::Services::ComputeNeutronL3Agent
    - OS::TripleO::Services::ComputeNeutronMetadataAgent
    - OS::TripleO::Services::TripleoPackages
    - OS::TripleO::Services::TripleoFirewall
    - OS::TripleO::Services::NeutronSriovAgent
    - OS::TripleO::Services::OpenDaylightOvs
    - OS::TripleO::Services::SensuClient
    - OS::TripleO::Services::FluentdClient
    - OS::TripleO::Services::VipHosts
```

また、以下の設定ファイルを用いて、ノードと IP アドレスを紐づける。

```
~/custom-templates/layout.yaml
```

(4) Overcloud の Deploy

上記で作成した設定ファイル等を使い、Overcloud (OpenStack) を Deploy する。今回は下記のように実行した。

```
$ openstack overcloud deploy --templates ¥
-r ~/custom-templates/custom-roles.yaml ¥
-e /usr/share/openstack-tripleo-heat-templates/environments/puppet-pacemaker.yaml ¥
-e /usr/share/openstack-tripleo-heat-templates/environments/network-isolation.yaml ¥
-e /usr/share/openstack-tripleo-heat-templates/environments/storage-environment.yaml ¥
-e ~/custom-templates/network.yaml ¥
-e ~/custom-templates/ceph.yaml ¥
-e ~/custom-templates/compute.yaml ¥
-e ~/custom-templates/layout.yaml
```

(5) 外部ネットワークの設定

Deploy された Overcloud に外部ネットワークや Floating IP を設定する。今回は下記のように実行した。

```
$ source overcloudrc
$ openstack network create public --external --provider-network-type flat --provider-physical-network datacenter
$ openstack subnet create public --network public --dhcp --allocation-pool start=10.3.44.128,end=10.3.44.191 --gateway 10.3.44.254
--subnet-range 10.3.44.0/24
```

(6) フェンシングの設定

本来であれば冗長化された Controller ノードにフェンシングの設定を行う必要があるが、今回は検証環境ということで割愛した。

4.2 ノード追加・削除

ノードを追加する場合、事前準備としてネットワーク等の物理配線後、4.1(2)まで登録が終わっている場合、`layout.yaml` を修正し、4.1(4)を再実行するだけである。これは、他のノード及びインスタンスはオンライン状態のままで可能である。

ノードを削除する場合、該当ノード上で Ceph OSD の削除後、Overcloud から該当のノードを削除する必要がある。この際、該当ノードの Hostname(例: `overcloud-osd-compute-6`)は欠番となるため、再度ノード追加する際に該当の番号を使わないように、`instack.json` の `capability` や `layout.yaml` 上の設定において注意が必要である。

なお、追加・削除の詳細な手順は Red Hat 社のドキュメント(Ref.[2])を参照いただきたい。

4.3 NUMA 設定

HCI 構成の場合、NUMA を考慮して CPU Pinning をすることが望ましい。基本的には、Ceph OSD Daemon を、NIC と Storage Controller が紐づいている側の CPU に Pinning する必要がある。これは、Ceph OSD が非常に多くの Network Traffic を発生させるため、NUMA またがりによる性能劣化を防ぐためである。

Ceph OSD Daemon は `systemd` によって起動されるが、`numa-systemd-osd.sh` を使用して Ceph が使用する NIC が接続されている側の NUMA で Daemon が起動されるように変更している。

4.4 Network 設計

物理 Network 構成については、少なくとも Storage 系 Network(Ceph Storage、Ceph Storage Management)を、他の Network と物理スイッチを分けることを推奨する。例えば External Network でトラフィックが急増した場合、VM の処理が追いつかず 802.3x pause frames をスイッチへ送信し、そのスイッチのポートが使えなくなり、同一ポートの別 VLAN Network も停止してしまう。この状態でも上流からパケットが流れてくるという状態が連鎖し、スイッチのバッファを使い切る可能性がある。このスイッチに Storage 系 Network がぶら下がっている場合、Ceph が深刻な障害を受ける危険性がある。

4.5 VM イメージのフォーマット形式

OpenStack で Ceph を Cinder Backend として利用する場合、Copy on Write(CoW)により新規 VM 作成が高速に実施できるという特徴をもっている。ただし、このためには、Glance に格納されているイメージフォーマットを QCOW2 ではなく RAW に変更する必要がある。Ceph は VM の Disk として QCOW2 をサポートしない。(Ref.[5])。

QCOW2 イメージを Glance に登録してしまうと、QCOW2 から RAW へのフォーマット変更しながらのコピーが発生してしまうため、VM を作成するのに時間がかかってしまう。

5. 結果と考察

以下に、ベンチマークから得られた結果および考察について示す。

一度に 80VM で同時に fio を実行しており、その値を総和したものをベンチマーク結果として採用した。以前の非 HCI の OpenStack×Ceph の検証において、Random Read 性能が極端に高い現象が発生したことから、VM のキャッシュヒットを疑い、VM のキャッシュクリア機構¹をスクリプト内に組み込んだパターンでも検証を実施した。

全てのベンチマークは各 3 回計測しており、総和した値の中間値を採用した。

【重要:以降記載のデータは本検証における実測値であり保証値ではありません。OSCA 技術検討会、デル株式会社、レッドハット株式会社、株式会社日立ソリューションズは、一切の保証をしません】

¹ 毎回のベンチマークごとに、前後に 120 秒のクールタイムを設けて全 VM で下記を実行。

```
sudo echo 3 | sudo tee /proc/sys/vm/drop_caches && sudo sync
```

5.1 Block Size 別の傾向

Job が 1、4、8 の際の、Block Size(以下 BS)別のデータを以下に示す。なおグラフは、左側がキャッシュクリア機構なし、右側がキャッシュクリア機構あり、上の 4 グラフが Write 時の IOPS と Throughput 性能、下の 4 グラフが Read 時の IOPS と Throughput 性能である。

(1) Job1 の際の BlockSize 別性能

図 5-1 に示すように、BS 増加に伴い、IOPS は減少、Throughput は増加する。Random Read において、全体的に高い値が出ている上、さらに突出して高い値が点在している。

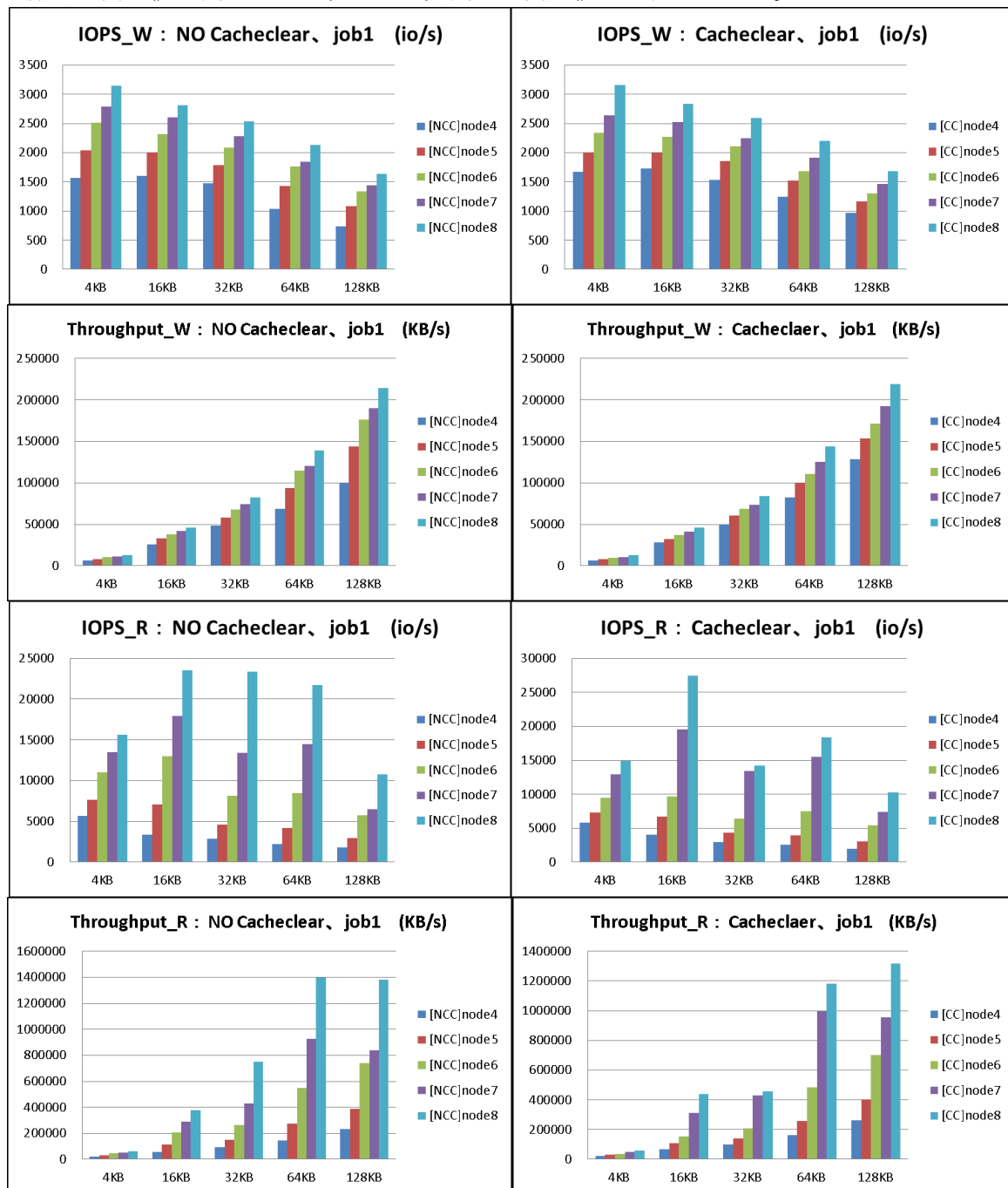


図 5-1 Job1 の際の BlockSize 別傾向

(2) Job4 の際の BlockSize 別性能

図 5-2 に示すように、Job1 と比較すると、Write についてはやや減少、Read は 1/3 程度に減少している。全体として突出した値がなく、綺麗なデータが取れている。

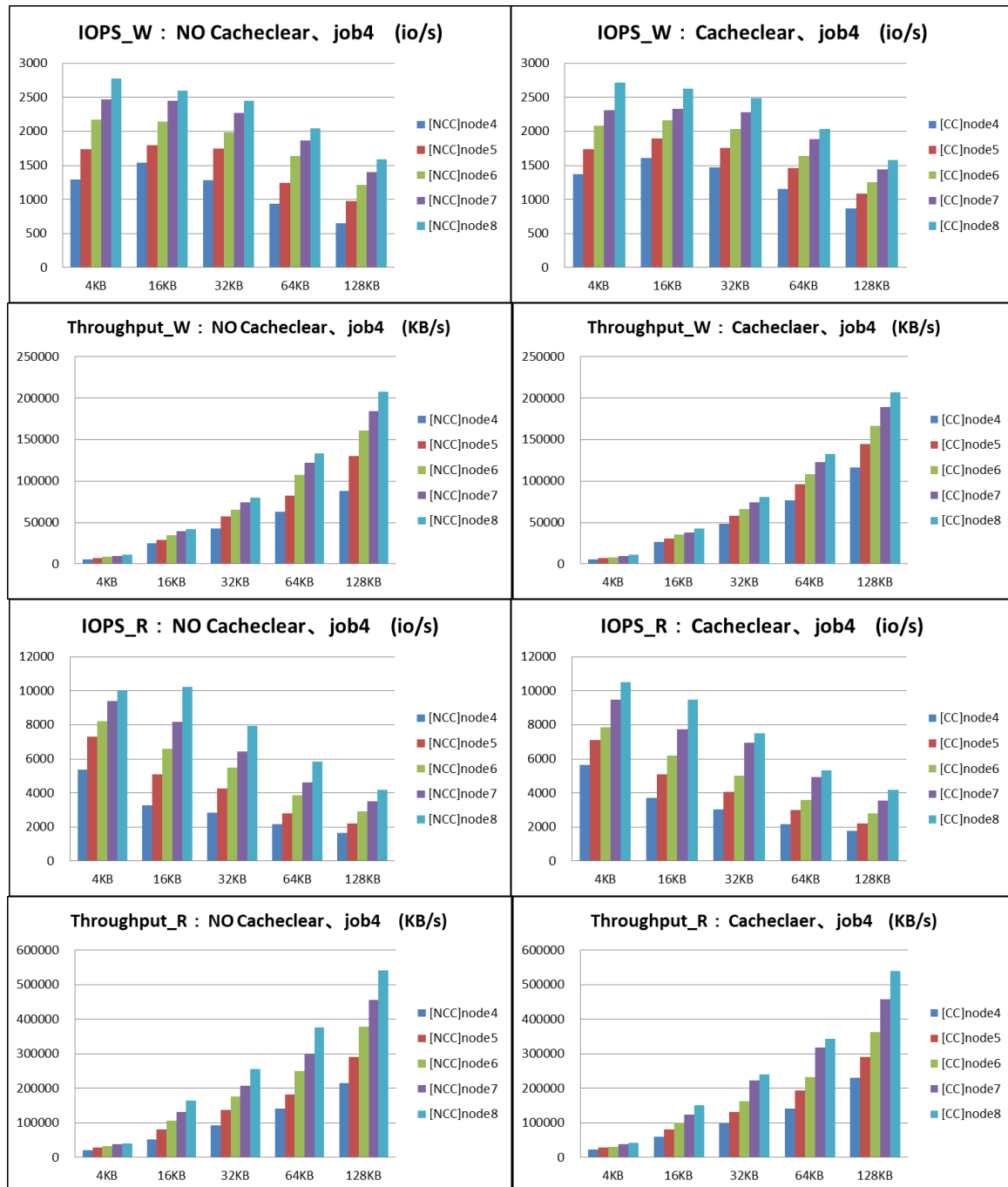


図 5-2 Job4 の際の BlockSize 別傾向

(3) Job8 の際の BlockSize 別性能

図 5-3 に示すように全体的に Job4 と傾向が似ており、やや性能が減少している。

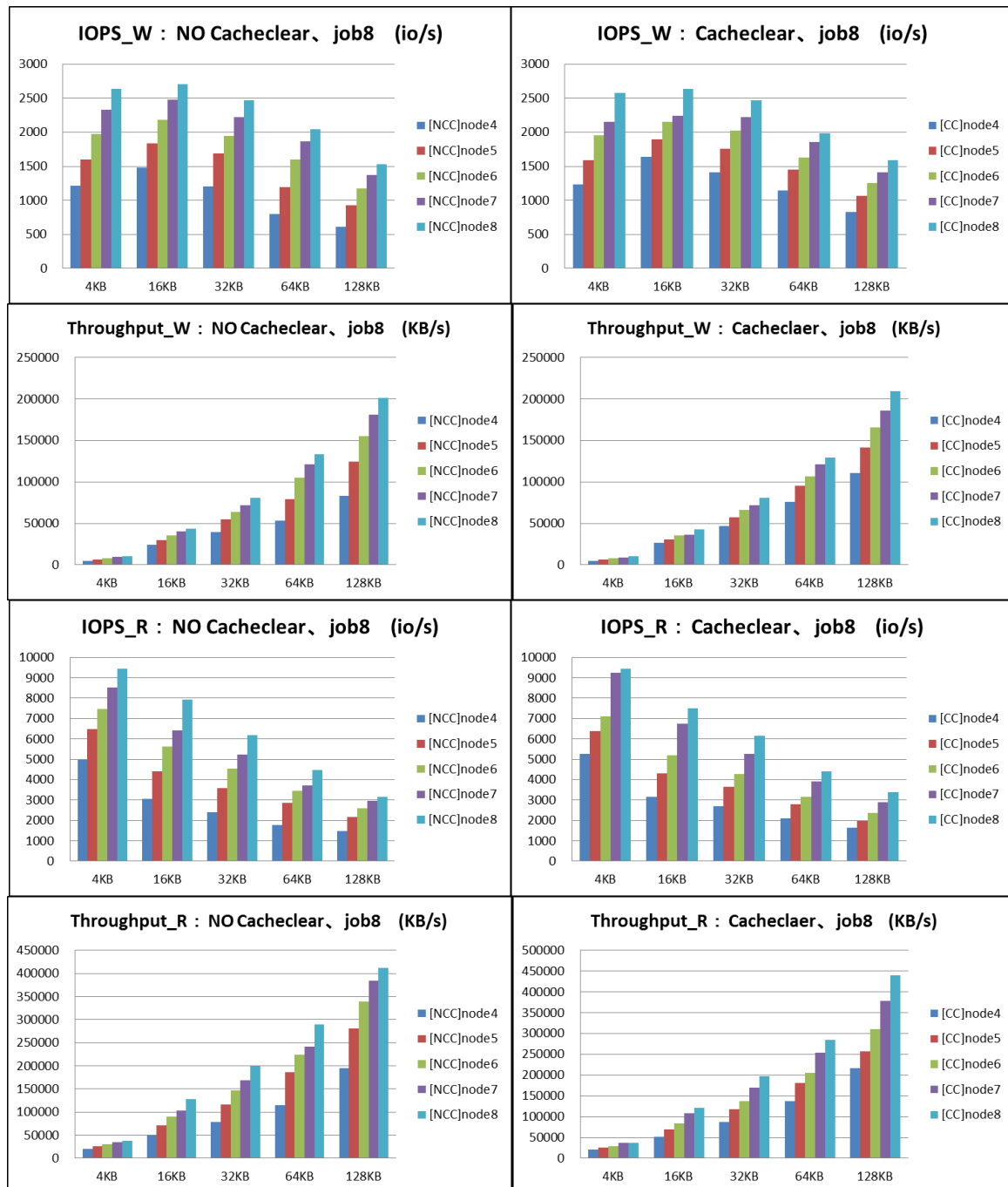


図 5-3 Job8 の際の BlockSize 別傾向

5.2 Compute ノード台数別による傾向

5.1 とは観点を変更し、Job1、4、8 時の、Compute ノード台数別のデータを以下に示す。なおグラフは、左側がキャッシュクリア機構なし、右側がキャッシュクリア機構あり、上の 4 グラフが Write 時の IOPS と Throughput 性能、下の 4 グラフが Read 時の IOPS と Throughput 性能である。

(1) Job1 の際の Compute ノード台数別性能

図 5-4 に示すように Compute ノード数に比例し、IOPS、Throughput 共に性能が向上する。Random Read において、全体的に高い値が出ている上、さらに突出して高い値が点在している。

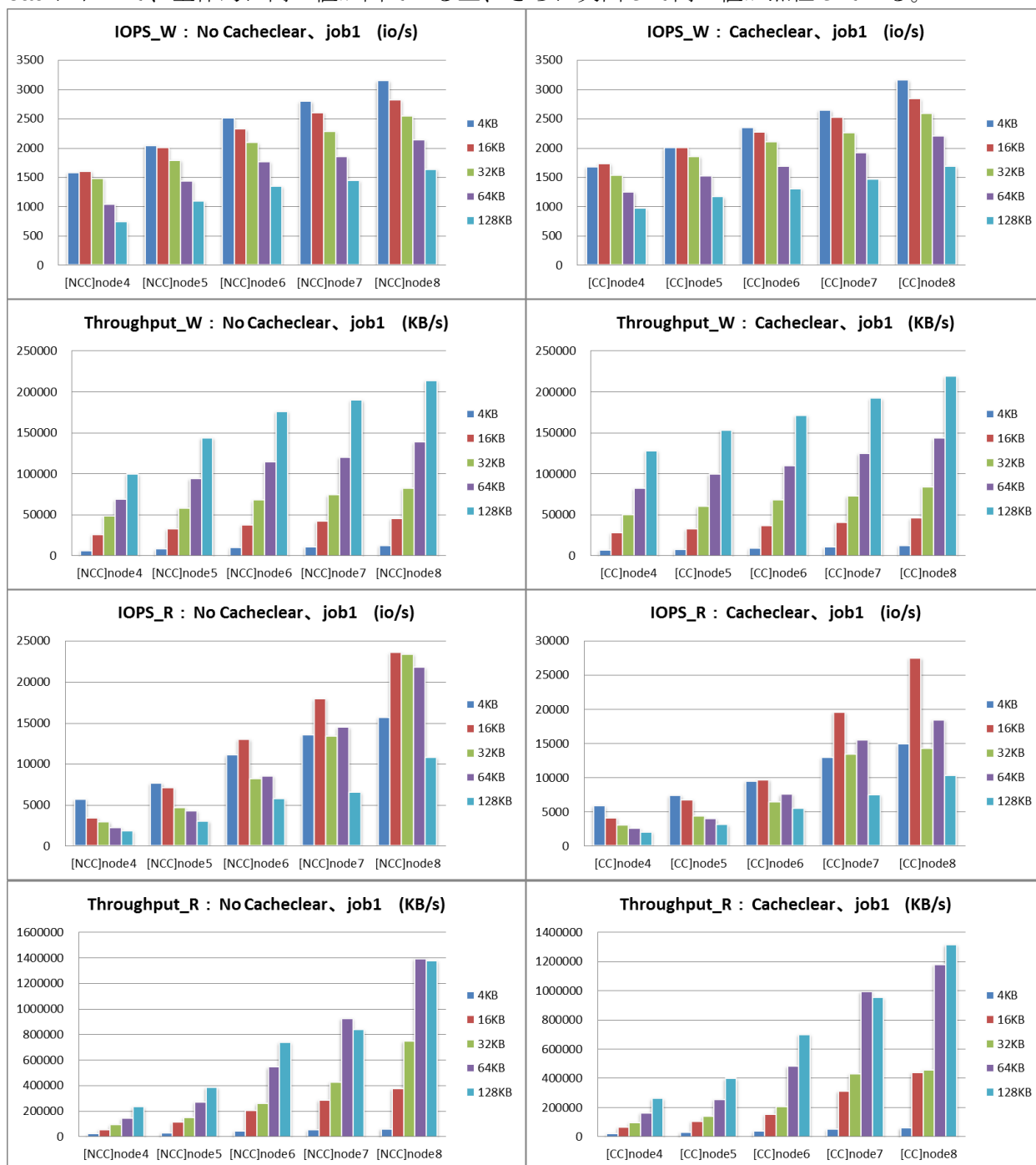


図 5-4 Job1 の際の Compute ノード別傾向

(2) Job4 の際の Compute ノード台数別性能

図 5-5 に示すように、Job1 と比較すると、Write についてはやや減少、Read は 1/3 程度に減少。全体として突出した値がなく、綺麗なデータが取れている。

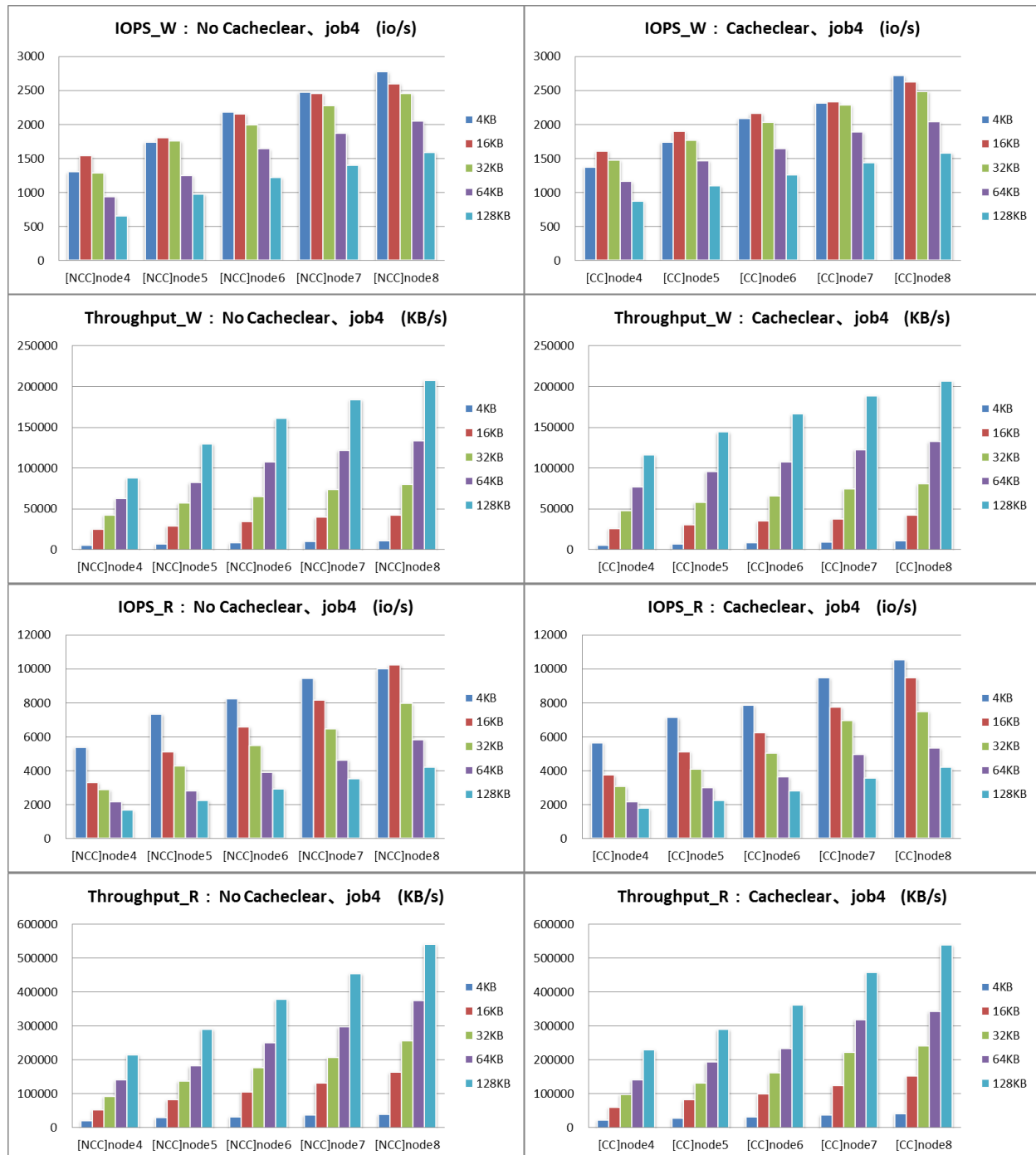


図 5-5 Job4 の際の Compute ノード別傾向

(3) Job8 の際の Compute ノード台数別性能

図 5-6 に示すように、全体的に Job4 と傾向が似ており、やや性能が減少している。

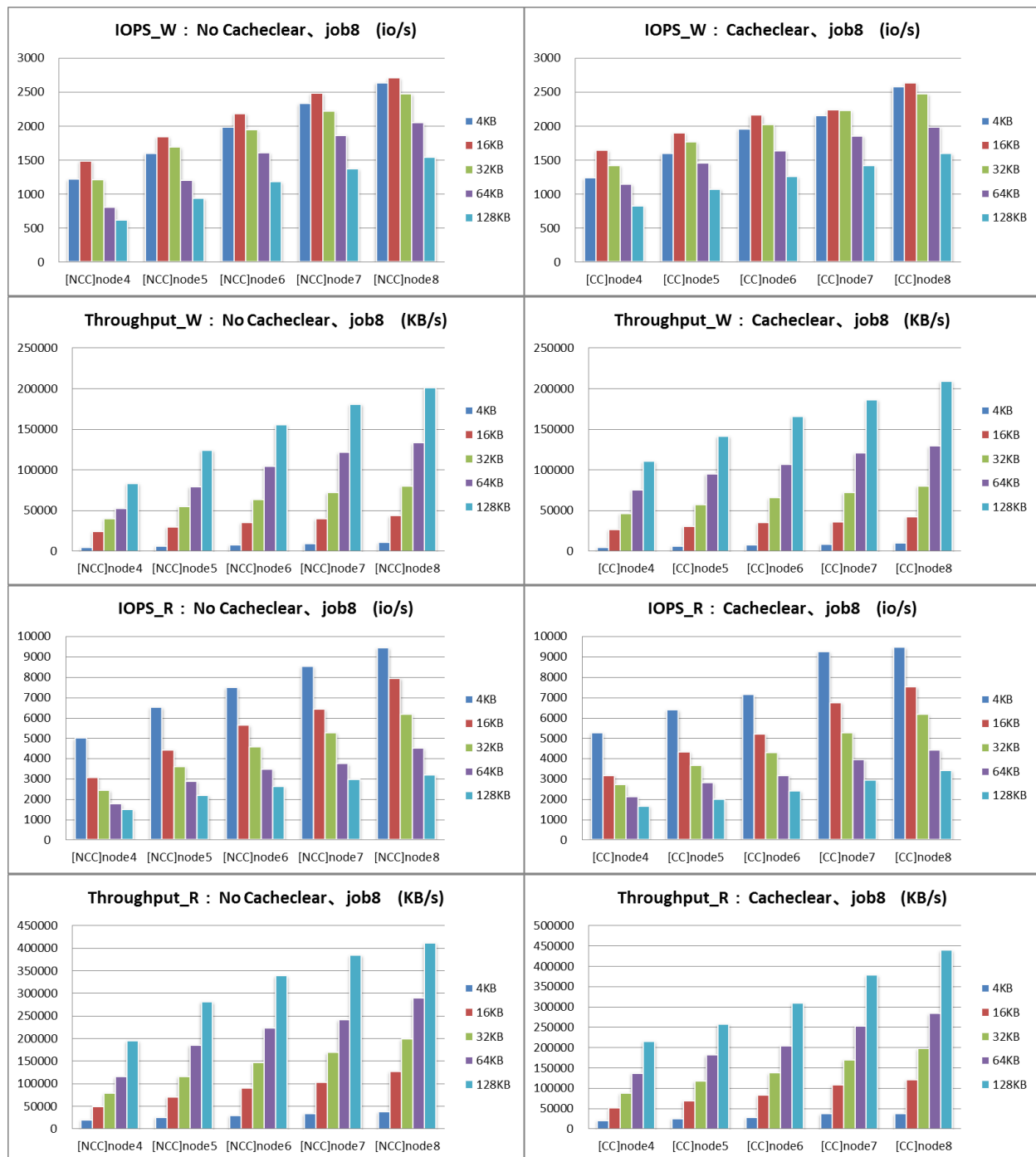


図 5-6 Job8 の際の Compute ノード別傾向

5.3 Random Write

Random Write に関しては、BS の観点では BS が増えるごとに IOPS は減少し、Throughput は増加するという傾向がみられた。これは、単体の HDD の性能と同じ傾向であり、特段おかしい挙動ではない。Compute ノード台数の観点では、OSD Disk が増加するごとに IOPS、Throughput 共に増加した。また、増加の比例定数は 1 に近く、OSD Disk 増加に伴い性能が単純増加する傾向がみられた。本結果は前回の検証(Ref.[1])と似た傾向であり、HCI ならではの性能の特徴は確認できなかった。

5.4 今回の検証における Random Read

今回の検証では、Cache による Random Read の高値を排除するため Cache Clear 機能をベンチマークに組み込んだが、その有無によりベンチマーク結果はそれほど変わらなかった。また、Block Size 別、Compute ノード台数別ともに、Job1 の場合だけ Random Read の値に極端な高性能値が見られた。これは多重度が低いため、全体の Write 負荷をかけきれずに、Ceph の Journal によるキャッシュ効果十分に動作する範囲内で Write が完了したためと考えられる。なお、キャッシュ効果については(Ref.[6])を参照いただきたい。

5.5 前回の検証との比較における Random Read

前回の検証の Random Read の数値に関しては、検証手順に問題があり、実態に近い値を計測できていなかったことが判明した。今回の検証では、その問題点を回避した上でベンチマークを実施しており、より実態に近い値を計測できた。以下にその根拠を示す。

今回の検証では、Job1 の時こそ Random Read で極端な高性能値が見られたが、Job4、8 の時は予想した範囲内(Random Read の 2~3 倍)の性能であった。一方、前回の検証ではすべてのパターンの Random Read で異常に高い値を示しており、なんらかのキャッシュ機能が働いたものと推測していた。前回の検証の Random Read の異常な高値がキャッシュ機能によるものだとすると、今回の検証の Job4、8 の場合でも高い値が出ると想定していたが、結果は異なっていた。

前回の検証と今回の検証の差は以下のとおりである。

- HCI 環境と非 HCI 環境の差
- ハードウェア構成の差
- ソフトウェア(RHEL、RH-OpenStack、RH-Ceph)の Version の差
- fio 実行時の、Random Write、Random Read の順序の差

本考察では Random Read の性能差が、HCI・非 HCI に起因するものか否かを調査した。具体的には、5.1、5.2 と同一ハードウェア、同一 Version のソフトウェアで HCI 環境を作り、fio 実行時の順序だけ変化させて追加調査した。

前回の検証、今回の検証共に、**fio** 用 **Script** として以下をもちいた。

```
s=120
t=120
for i in 1 2 3
do for j in 4 16 32 64 128
do for k in 1 4 8
do echo "***W_1G_${j}KB_job${k}_para80_loop${i}***"
ssh cloud-user@192.168.0.94 fio -rw=randwrite -size=1G -ioengine=libaio -iodepth=4 -invalidate=1 -direct=1 -name=test.bin
-runtime=${t} -bs=${j}k -numjobs=${k} -group_reporting > "hc8_W_1G_${j}KB_job${k}_para80_loop${i}_192.168.0.94.log" & ssh 【以
下インスタンス数分だけ繰り返し】

sleep ${s}s

echo "***R_1G_${j}KB_job${k}_para80_loop${i}***"
ssh cloud-user@192.168.0.94 fio -rw=randread -size=1G -ioengine=libaio -iodepth=4 -invalidate=1 -direct=1 -name=test.bin
-runtime=${t} -bs=${j}k -numjobs=${k} -group_reporting > "hc8_R_1G_${j}KB_job${k}_para80_loop${i}_192.168.0.94.log" & ssh 【以
下インスタンス数分だけ繰り返し】

sleep ${s}s

done
done
done
```

本 **Script** は、**fio** を **Random Write** から開始し、**Random Write**、**Random Read** を繰り返し実行している。今回の検証実施にあたり **fio** の挙動を調べた。

fio のベンチマークファイルは、指定したサイズ[**size**]のバイナリファイルを、指定した **Job** 数[**numjobs**]分作成する(**test.bin.0**、**test.bin.1** 等)。既にファイルがある場合は流用する。**Random Write**、**Random Read** 時の挙動は以下のとおりである。

(1) **Random Write** 実行時

(a) ベンチマークファイルがない場合

ベンチマークファイルを作成しながらベンチマーク実施。**fio** の引数で指定された時間[**runtime**]とサイズ[**size**]に応じて、中身が **0** で指定サイズのベンチマークファイルがまず作成され、そこに指定された時間書き込む。書き込み時間が足りない場合、指定されたファイルサイズだが、バイナリの中身が **0** 連続部分を含むファイルが作成される。

図 5-7 に示すように、これを **Ceph** 側の取り扱いも考慮すると、**VM** の **Volume** は **Thin Provisioning** で扱われており、上記ファイルも **Sparse File** で処理されているため、データが書き込まれていない部分は **Sparse Area** 扱いとなり、参照された場合 **0** を瞬時に返す部分が存在するファイルが作成される。

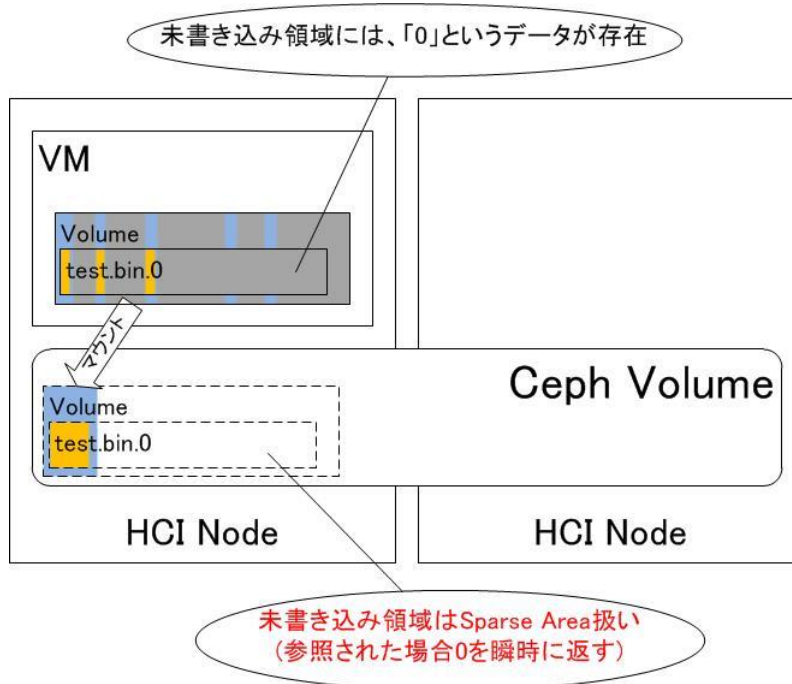


図 5-7 fio の Random Write におけるベンチマークファイルの処理

- (b) 既にベンチマークファイルがある場合
そのファイルを上書きしながらベンチマーク実施
- (2) Random Read 実行時
 - (a) ベンチマークファイルがない場合
fio の引数で指定したサイズ[size]のベンチマーク用ファイルを Job 数[numjobs]分だけ作成し、作成完了後にベンチマーク実施。
 - (b) 既にベンチマークファイルがある場合
そのファイルを利用してベンチマーク実施。

ここで問題になるのが、利用したスクリプトでは、(1)-(a)から(2)-(b)のルートを通ることである。つまり、ベンチマーク用としては不完全な 0 連続部分を含むファイルを、Random Read 用ファイルとして流用してしまう。先に説明した通り Ceph 側からは Sparse File 扱いとなるため、VM の fio がベンチマークファイルの 0 連続部分を Read する際、図 5-8 に示すように Ceph 側は実際に中身を返すのではなく、0 が連続しているという情報を瞬時に返すため一瞬で Read 処理が完了し、結果として極端に高性能を出す可能性がある。

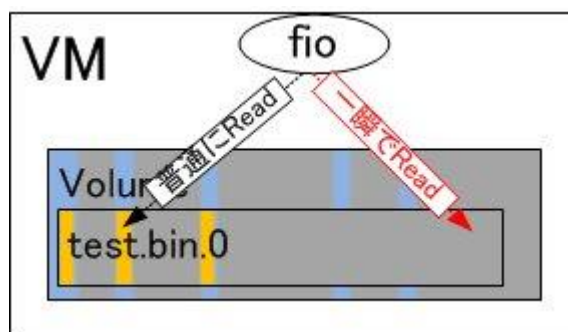


図 5-8 Sparse Area Read 時の処理

5.1、5.2 で得たデータは、本問題を回避するため、スクリプト実行前に(2)-(a)、つまり正しいベンチマークファイルの生成処理を別途実施し、適切なベンチマークファイルを作成した後でベンチマークを実施している。Job4、8 の Random Read の際に極端に高い値が出なかったのは、本回避策による効果と考えられ。

その裏付けのため、回避策なしの場合での追加検証を実施した。なお、追加調査については時間の関係上、キャッシュクリア機構なしのノード 8 台時のみ調査した。

図 5-3 のキャッシュクリア機構なしのパターンに、ノード 8 台時で(1)-(a)から(2)-(b)の処理、つまり不完全なベンチマークファイルを用いたパターンについて追加したデータを図 5-9 に示す。なお、[NCC]node4～8 とかかかれているのは正しいベンチマークファイル生成処理を別途実施したパターン、[NCC]node8 nobin は不完全なベンチマークファイルを用いたパターンの値を示している。

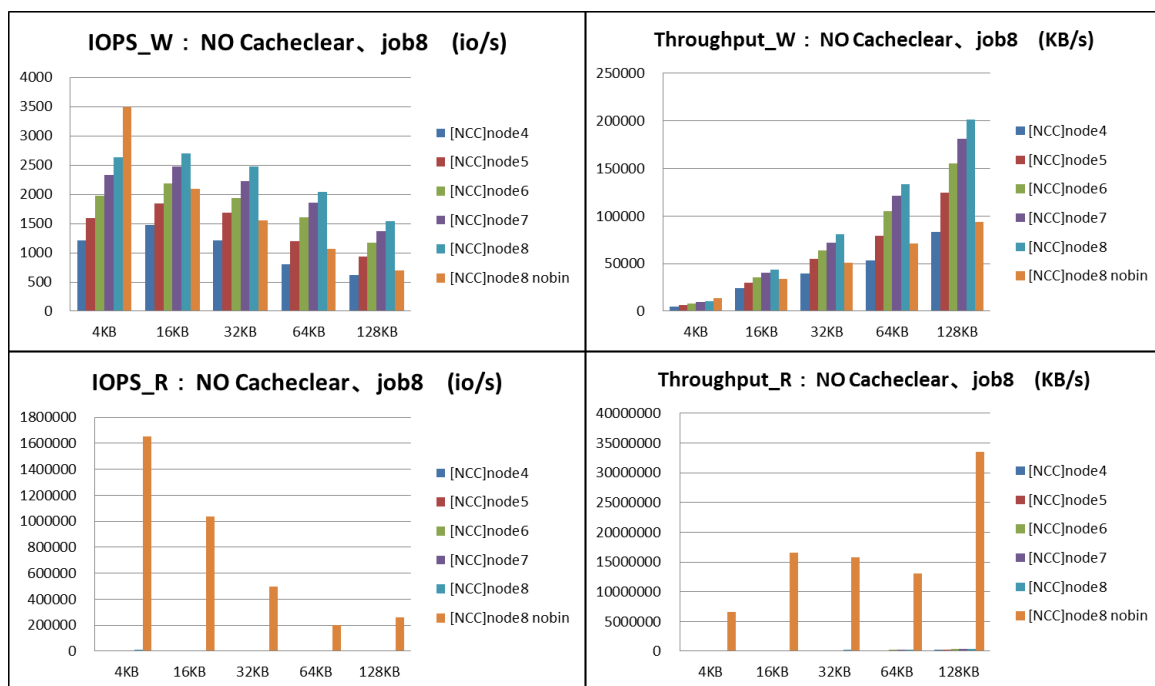


図 5-9 不完全なベンチマークファイルを用いたパターンとの比較

Write に関しては、正しいベンチマークファイルの生成処理を別途実施したパターンに比べて性能が減少している。これは VM から見た場合の 0 部分の書き込み、つまり Ceph の Sparse File の拡張処理により性能劣化が発生しているためと思われるが、裏付けはとれなかった。

一方 Read に関しては、正しいベンチマークファイルの生成処理を別途実施したパターンに比べて極端に高い値(数百倍)を出している。同一ハードウェア、同一 Version のソフトウェアの HCI 環境で、ベンチマークファイルの内容により性能が極端に変わることから、この差は不完全なベンチマークファイルを用いたことにより発生したと考える。つまり、前回の検証時に Random Read 性能が極端に高かった理由については、fio の Random Read、Random Write を実施する順序に起因して不完全なベンチマークファイルが生成されてしまったためと考える。

以上より、前回の検証の Random Read に関しては実態に近い値を計測できていなかったこと、今回の検証では問題点を回避しベンチマークを実施したため、より実態に近い値を計測できたことが示せた。

6. まとめ

6.1 構築・設定

以下に、構築・設定検証から得られた設定指針について示す。

(1) 構築手順

設定ファイルが多岐にわたるため設定が難しいが、そのファイルを元に **Operating System** から自動構築が可能のため、**Try & Error** が容易である。

(2) ノード追加・削除

ノード追加は容易であり、オンライン状態で実行可能である。ノード削除は手順及び注意点が存在する。

(3) NUMA

Ceph OSD Daemon と NIC、Storage Controller を同じ CPU に割り当てることを推奨する。

(4) Network

Storage 用 Network は、他の Network とは物理的に異なる Switch へ接続することを推奨する。

(5) VM イメージのフォーマット形式

RAW イメージで登録する必要がある。

6.2 ベンチマーク結果から得られた指針

以下に、ベンチマーク結果から得られた OSD Disk の設計指針について示す。

実際の HCI 環境においては、ノード数が十分で、負荷の並列度も高いと考えられることから、ノード数 8、Job8、Cache Clear 機構有効時のデータを参考にすべきと考える。その組み合わせにおける IOPS、Throughput の値は以下のとおりである。

表 6-1 BS 別の性能全体傾向

IOPS / Throughput	Block Size	Write / Read	全体	1 ノードあたり	1 OSD あたり
IOPS	4KB	Write	2,577	322	107
		Read	9,450	1,181	394
	32KB	Write	2473	309	103
		Read	6149	769	256
	128KB	Write	1,591	199	66
		Read	3,391	424	141
Throughput	4KB	Write	10,456(KB/s)	1,307(KB/s)	436(KB/s)
		Read	37,979(KB/s)	4,747(KB/s)	1,582(KB/s)
	32KB	Write	80,454(KB/s)	10,057(KB/s)	3,352(KB/s)
		Read	197,831(KB/s)	24,729(KB/s)	8,273(KB/s)
	128KB	Write	209,247(KB/s)	26,156(KB/s)	8,719(KB/s)
		Read	439,561(KB/s)	54,945(KB/s)	18,315(KB/s)

以上の結果より、HCI 環境において OSD Disk に 10,000rpm の SAS HDD を用いる場合、Disk 1 台

あたり 100 IOPS、Throughput 8MB/s 程度で試算し、必要な OSD Disk 数を見積もるべきであるとの結論を得た。

6.3 懸念事項

現時点では、ノード数が 42 台を超える場合、HCI ではなく Standalone の Ceph と Compute を用いることが推奨されている(Ref.[2])。これは、大規模構成の場合 Network ボトルネックを考慮する必要があること、Compute リソースと Storage リソースの必要量がずれ始める可能性があることからと考えられている。PoC 環境においては HCI 構成が適していると考えられるが、本番構成においては最大構成において上記制限にかかる可能性の有無を考慮すべきである。

7. 参考・関連文献

- [1] "OpenStack on Ceph" の性能評価と設計ポイントに関する 2 本のホワイトペーパーのご紹介。
URL = <http://ja.community.dell.com/techcenter/b/weblog/archive/2016/07/04/quot-openstack-on-ceph-quot-2-wp>
- [2] 2016 - Hyper-Converged Red Hat OpenStack Platform 10 and Red Hat Ceph Storage 2
URL = <https://access.redhat.com/articles/2861641>
- [3] Director Installation and Usage
URL = https://access.redhat.com/documentation/en-us/red_hat_openstack_platform/10/html/director_installation_and_usage/
- [4] Advanced Overcloud Customization
URL = https://access.redhat.com/documentation/en-us/red_hat_openstack_platform/10/html/advanced_overcloud_customization/
- [5] Block Devices and OpenStack — Ceph Documentation
URL = <http://docs.ceph.com/docs/master/rbd/rbd-openstack/>
- [6] 25,000 のベンチマーク結果からわかった！ OpenStack on Ceph の性能（OpenStack Days Tokyo 2016）
URL = http://openstackdays.com/archive/2016/wp-content/uploads/2016/10/O1_OSDT_OSCA20160705_3.pdf
- [7] Welcome to Ceph - Ceph Documentation
URL = <http://docs.ceph.com/docs/master/>
- [8] Product Documentation - Red Hat Customer Portal
URL = <https://access.redhat.com/documentation/en/>

本書は、情報提供のみを目的に執筆されており、誤字脱字、技術上の誤りには一切責任を負いません。

本書の内容は一般的な原則を記しており、すべての環境での動作を保証するものではありません。

本書の内容は執筆時現在のものであり、明示的、暗示的を問わず、いかなる内容も保証いたしません。

本書に掲載された文章、画像、図面等は、特に記載がない限り、OSCA、日立ソリューションズ、レッドハット、デルが保有しています。特に記載がない限り、複製、改変したものを無断で再配布することはできません。

Linux は、Linus Torvalds の米国およびその他の国における登録商標または商標です。

OpenStack®の文字表記と OpenStack のロゴは、米国とその他の国における OpenStack Foundation の登録商標/サービスマークまたは商標/サービスマークのいずれかであり、OpenStack Foundation の許諾を得て使用しています。日立製作所は OpenStack Foundation や OpenStack コミュニティの関連企業ではなく、また支援や出資を受けていません。デルは OpenStack Foundation または OpenStack コミュニティとは提携しておらず、公認や出資もを受けていません。Red Hat は、OpenStack Foundation と OpenStack コミュニティのいずれにも所属しておらず、公認や出資もを受けていません。

OSCA™（Open Standard Cloud Association）は、デル株式会社の商標です。

PowerEdge、DELL ロゴは、米国 Dell Inc.の米国およびその他の国における商標または登録商標です。

Red Hat および Red Hat をベースとしたすべての商標とロゴ、Ceph は、米国 Red Hat, Inc.の米国およびその他の国における商標または登録商標です。

その他、記載の商標やロゴは、各社の商標または登録商標です。

8. 付録

8.1 OpenStack 及び Ceph の構築に使用した設定ファイル

8.1.1 Undercloud.conf

```
[DEFAULT]

#
# From instack-undercloud
#

# Local file path to the necessary images. The path should be a
# directory readable by the current user that contains the full set of
# images. (string value)
image_path = /home/stack/images

# Fully qualified hostname (including domain) to set on the
# Undercloud. If left unset, the current hostname will be used, but
# the user is responsible for configuring all system hostname settings
# appropriately. If set, the undercloud install will configure all
# system hostname settings. (string value)
#undercloud_hostname = <None>

# IP information for the interface on the Undercloud that will be
# handling the PXE boots and DHCP for Overcloud instances. The IP
# portion of the value will be assigned to the network interface
# defined by local_interface, with the netmask defined by the prefix
# portion of the value. (string value)
local_ip = 192.0.2.1/24

# Network gateway for the Neutron-managed network for Overcloud
# instances. This should match the local_ip above when using
# masquerading. (string value)
network_gateway = 192.0.2.1

# Virtual IP address to use for the public endpoints of Undercloud
# services. Only used with SSL. (string value)
undercloud_public_vip = 192.0.2.2

# Virtual IP address to use for the admin endpoints of Undercloud
# services. Only used with SSL. (string value)
undercloud_admin_vip = 192.0.2.3

# Certificate file to use for OpenStack service SSL connections.
# Setting this enables SSL for the OpenStack API endpoints, leaving it
# unset disables SSL. (string value)
#undercloud_service_certificate =

# When set to True, an SSL certificate will be generated as part of
# the undercloud install and this certificate will be used in place of
# the value for undercloud_service_certificate. The resulting
# certificate will be written to
# /etc/pki/tls/certs/undercloud-[undercloud_public_vip].pem. This
# certificate is signed by CA selected by the
# "certificate_generation_ca" option. (boolean value)
#generate_service_certificate = false

# The certmonger nickname of the CA from which the certificate will be
# requested. This is used only if the generate_service_certificate
# option is set. Note that if the "local" CA is selected the
# certmonger's local CA certificate will be extracted to /etc/pki/ca-
# trust/source/anchors/cm-local-ca.pem and subsequently added to the
# trust chain. (string value)
#certificate_generation_ca = local

# The kerberos principal for the service that will use the
# certificate. This is only needed if your CA requires a kerberos
# principal. e.g. with FreeIPA. (string value)
#service_principal =

# Network interface on the Undercloud that will be handling the PXE
# boots and DHCP for Overcloud instances. (string value)
```

```
local_interface = em1

# MTU to use for the local_interface. (integer value)
#local_mtu = 1500

# Network CIDR for the Neutron-managed network for Overcloud
# instances. This should be the subnet used for PXE booting. The
# current default for this value is 192.0.2.0/24, but this is
# deprecated due to it being a non-routable CIDR under RFC 5737. The
# default value for this option will be changed in the Ocata release.
# A different, valid CIDR should be selected to avoid problems. If an
# overcloud has already been deployed with the 192.0.2.0/24 CIDR and
# therefore the CIDR cannot be changed, you must set this option to
# 192.0.2.0/24 explicitly to avoid it changing in future releases, and
# all other network options related to the CIDR (e.g. local_ip) must
# also be set to maintain a valid configuration. (string value)
network_cidr = 192.0.2.0/24

# Network that will be masqueraded for external access, if required.
# This should be the subnet used for PXE booting. (string value)
masquerade_network = 192.0.2.0/24

# Start of DHCP allocation range for PXE and DHCP of Overcloud
# instances. (string value)
dhcp_start = 192.0.2.5

# End of DHCP allocation range for PXE and DHCP of Overcloud
# instances. (string value)
dhcp_end = 192.0.2.24

# Path to hieradata override file. If set, the file will be copied
# under /etc/puppet/hieradata and set as the first file in the hiera
# hierarchy. This can be used to to custom configure services beyond
# what undercloud.conf provides (string value)
#hieradata_override =

# Path to network config override template. If set, this template will
# be used to configure the networking via os-net-config. Must be in
# json format. Templated tags can be used within the template, see
# instack-undercloud/elements/undercloud-stack-config/net-
# config.json.template for example tags (string value)
#net_config_override =

# Network interface on which inspection dnsmasq will listen. If in
# doubt, use the default value. (string value)
# Deprecated group/name - [DEFAULT]/discovery_interface
#inspection_interface = br-ctlplane

# Temporary IP range that will be given to nodes during the inspection
# process. Should not overlap with the range defined by dhcp_start
# and dhcp_end, but should be in the same network. (string value)
# Deprecated group/name - [DEFAULT]/discovery_iprange
#inspection_iprange = 192.0.2.100,192.0.2.120

# Whether to enable extra hardware collection during the inspection
# process. Requires python-hardware or python-hardware-detect package
# on the introspection image. (boolean value)
#inspection_extras = true

# Whether to run benchmarks when inspecting nodes. Requires
# inspection_extras set to True. (boolean value)
# Deprecated group/name - [DEFAULT]/discovery_runbench
#inspection_runbench = false

# Whether to support introspection of nodes that have UEFI-only
# firmware. (boolean value)
#inspection_enable_uefi = false

# Whether to enable the debug log level for Undercloud OpenStack
# services. (boolean value)
#undercloud_debug = true

# Whether to install Tempest in the Undercloud. (boolean value)
#enable_tempest = true

# Whether to install Mistral services in the Undercloud. (boolean
```

```
# value)
#enable_mistral = true

# Whether to install Zaqr services in the Undercloud. (boolean value)
#enable_zaqr = true

# Whether to install Telemetry services (ceilometer, aodh) in the
# Undercloud. (boolean value)
#enable_telemetry = true

# Whether to install the TripleO UI. (boolean value)
#enable_ui = true

# Whether to install requirements to run the TripleO validations.
# (boolean value)
#enable_validations = true

# Whether to use iPXE for deploy and inspection. (boolean value)
# Deprecated group/name - [DEFAULT]/ipxe_deploy
#ipxe_enabled = true

# Whether to store events in the Undercloud Ceilometer. (boolean
# value)
#store_events = false

# Maximum number of attempts the scheduler will make when deploying
# the instance. You should keep it greater or equal to the number of
# bare metal nodes you expect to deploy at once to work around
# potential race condition when scheduling. (integer value)
# Minimum value: 1
#scheduler_max_attempts = 30

# Whether to clean overcloud nodes (wipe the hard drive) between
# deployments and after the introspection. (boolean value)
#clean_nodes = false

[auth]

#
# From instack-undercloud
#

# Password used for MySQL databases. If left unset, one will be
# automatically generated. (string value)
#undercloud_db_password = <None>

# Keystone admin token. If left unset, one will be automatically
# generated. (string value)
#undercloud_admin_token = <None>

# Keystone admin password. If left unset, one will be automatically
# generated. (string value)
#undercloud_admin_password = <None>

# Glance service password. If left unset, one will be automatically
# generated. (string value)
#undercloud_glance_password = <None>

# Heat db encryption key(must be 16, 24, or 32 characters. If left
# unset, one will be automatically generated. (string value)
#undercloud_heat_encryption_key = <None>

# Heat service password. If left unset, one will be automatically
# generated. (string value)
#undercloud_heat_password = <None>

# Neutron service password. If left unset, one will be automatically
# generated. (string value)
#undercloud_neutron_password = <None>

# Nova service password. If left unset, one will be automatically
# generated. (string value)
#undercloud_nova_password = <None>
```

```
# Ironic service password. If left unset, one will be automatically
# generated. (string value)
#undercloud_ironic_password = <None>

# Aodh service password. If left unset, one will be automatically
# generated. (string value)
#undercloud_aodh_password = <None>

# Ceilometer service password. If left unset, one will be
# automatically generated. (string value)
#undercloud_ceilometer_password = <None>

# Ceilometer metering secret. If left unset, one will be automatically
# generated. (string value)
#undercloud_ceilometer_metering_secret = <None>

# Ceilometer snmpd read-only user. If this value is changed from the
# default, the new value must be passed in the overcloud environment
# as the parameter SnmpdReadOnlyUserName. This value must be between 1
# and 32 characters long. (string value)
#undercloud_ceilometer_snmpd_user = ro_snmp_user

# Ceilometer snmpd password. If left unset, one will be automatically
# generated. (string value)
#undercloud_ceilometer_snmpd_password = <None>

# Swift service password. If left unset, one will be automatically
# generated. (string value)
#undercloud_swift_password = <None>

# Mistral service password. If left unset, one will be automatically
# generated. (string value)
#undercloud_mistral_password = <None>

# Rabbitmq cookie. If left unset, one will be automatically generated.
# (string value)
#undercloud_rabbit_cookie = <None>

# Rabbitmq password. If left unset, one will be automatically
# generated. (string value)
#undercloud_rabbit_password = <None>

# Rabbitmq username. If left unset, one will be automatically
# generated. (string value)
#undercloud_rabbit_username = <None>

# Heat stack domain admin password. If left unset, one will be
# automatically generated. (string value)
#undercloud_heat_stack_domain_admin_password = <None>

# Swift hash suffix. If left unset, one will be automatically
# generated. (string value)
#undercloud_swift_hash_suffix = <None>

# Sensus service password. If left unset, one will be automatically
# generated. (string value)
#undercloud_sensu_password = <None>

# HAProxy stats password. If left unset, one will be automatically
# generated. (string value)
#undercloud_haproxy_stats_password = <None>

# Zaqr password. If left unset, one will be automatically generated.
# (string value)
#undercloud_zaqar_password = <None>

# Horizon secret key. If left unset, one will be automatically
# generated. (string value)
#undercloud_horizon_secret_key = <None>
```

8.1.2 instackenv.json

```

{
  "nodes": [
    {
      "_comment": "controller-0",
      "mac": [
        "xx:xx:xx:xx:xx:e6"
      ],
      "pm_type": "pxe_ipmitool",
      "pm_user": "xxxxxxx",
      "pm_password": "xxxxxxx",
      "pm_addr": "10.1.1.151",
      "capabilities": "node:controller-0,boot_option:local"
    },
    {
      "_comment": "controller-1",
      "mac": [
        "xx:xx:xx:xx:xx:3a"
      ],
      "pm_type": "pxe_ipmitool",
      "pm_user": "xxxxxxx",
      "pm_password": "xxxxxxx",
      "pm_addr": "10.1.1.152",
      "capabilities": "node:controller-1,boot_option:local"
    },
    {
      "_comment": "controller-2",
      "mac": [
        "xx:xx:xx:xx:xx:dc"
      ],
      "pm_type": "pxe_ipmitool",
      "pm_user": "xxxxxxx",
      "pm_password": "xxxxxxx",
      "pm_addr": "10.1.1.153",
      "capabilities": "node:controller-2,boot_option:local"
    },
    {
      "_comment": "osd-compute-0",
      "mac": [
        "xx:xx:xx:xx:xx:af"
      ],
      "pm_type": "pxe_ipmitool",
      "pm_user": "xxxxxxx",
      "pm_password": "xxxxxxx",
      "pm_addr": "10.1.1.162",
      "capabilities": "node:osd-compute-0,boot_option:local"
    },
    {
      "_comment": "osd-compute-1",
      "mac": [
        "xx:xx:xx:xx:xx:48"
      ],
      "pm_type": "pxe_ipmitool",
      "pm_user": "xxxxxxx",
      "pm_password": "xxxxxxx",
      "pm_addr": "10.1.1.163",
      "capabilities": "node:osd-compute-1,boot_option:local"
    },
    {
      "_comment": "osd-compute-2",
      "mac": [
        "xx:xx:xx:xx:xx:bd"
      ],
      "pm_type": "pxe_ipmitool",
      "pm_user": "xxxxxxx",
      "pm_password": "xxxxxxx",
      "pm_addr": "10.1.1.164",
      "capabilities": "node:osd-compute-2,boot_option:local"
    },
    {
      "_comment": "osd-compute-3",
      "mac": [
        "xx:xx:xx:xx:xx:34"
      ],
      "pm_type": "pxe_ipmitool",

```

```

        "pm_user": "xxxxxxx",
        "pm_password": "xxxxxxx",
        "pm_addr": "10.1.1.165",
        "capabilities": "node:osd-compute-3,boot_option:local"
    },
    {
        "_comment": "osd-compute-5",
        "mac": [
            "xx:xx:xx:xx:xx:7f"
        ],
        "pm_type": "pxe_ipmitool",
        "pm_user": "xxxxxxx",
        "pm_password": "xxxxxxx",
        "pm_addr": "10.1.1.167",
        "capabilities": "node:osd-compute-5,boot_option:local"
    },
    {
        "_comment": "osd-compute-6",
        "mac": [
            "xx:xx:xx:xx:xx:95"
        ],
        "pm_type": "pxe_ipmitool",
        "pm_user": "xxxxxxx",
        "pm_password": "xxxxxxx",
        "pm_addr": "10.1.1.168",
        "capabilities": "node:osd-compute-6,boot_option:local"
    },
    {
        "_comment": "osd-compute-7",
        "mac": [
            "xx:xx:xx:xx:xx:b0"
        ],
        "pm_type": "pxe_ipmitool",
        "pm_user": "xxxxxxx",
        "pm_password": "xxxxxxx",
        "pm_addr": "10.1.1.169",
        "capabilities": "node:osd-compute-7,boot_option:local"
    }
]
}

```

8.1.3 custom-roles.yaml

```

# Specifies which roles (groups of nodes) will be deployed
# Note this is used as an input to the various *.j2.yaml
# Jinja2 templates, so that they are converted into *.yaml
# during the plan creation (via a mistral action/workflow).
#
# The format is a list, with the following format:
#
# * name: (string) mandatory, name of the role, must be unique
#
# CountDefault: (number) optional, default number of nodes, defaults to 0
# sets the default for the {{role.name}}Count parameter in overcloud.yaml
#
# HostnameFormatDefault: (string) optional default format string for hostname
# defaults to '%stackname%{{role.name.lower()}}-%index%'
# sets the default for {{role.name}}HostnameFormat parameter in overcloud.yaml
#
# ServicesDefault: (list) optional default list of services to be deployed
# on the role, defaults to an empty list. Sets the default for the
# {{role.name}}Services parameter in overcloud.yaml
- name: Controller
  CountDefault: 1
  ServicesDefault:
    - OS::TripleO::Services::CACerts
    - OS::TripleO::Services::CephMon
    - OS::TripleO::Services::CephExternal
    - OS::TripleO::Services::CephRgw
    - OS::TripleO::Services::CinderApi

```

- OS::TripleO::Services::CinderBackup
- OS::TripleO::Services::CinderScheduler
- OS::TripleO::Services::CinderVolume
- OS::TripleO::Services::Core
- OS::TripleO::Services::Kernel
- OS::TripleO::Services::Keystone
- OS::TripleO::Services::GlanceApi
- OS::TripleO::Services::GlanceRegistry
- OS::TripleO::Services::HeatApi
- OS::TripleO::Services::HeatApiCfn
- OS::TripleO::Services::HeatApiCloudwatch
- OS::TripleO::Services::HeatEngine
- OS::TripleO::Services::MySQL
- OS::TripleO::Services::NeutronDhcpAgent
- OS::TripleO::Services::NeutronL3Agent
- OS::TripleO::Services::NeutronMetadataAgent
- OS::TripleO::Services::NeutronApi
- OS::TripleO::Services::NeutronCorePlugin
- OS::TripleO::Services::NeutronOvsAgent
- OS::TripleO::Services::RabbitMQ
- OS::TripleO::Services::HAproxy
- OS::TripleO::Services::Keepalived
- OS::TripleO::Services::Memcached
- OS::TripleO::Services::Pacemaker
- OS::TripleO::Services::Redis
- OS::TripleO::Services::NovaConductor
- OS::TripleO::Services::MongoDb
- OS::TripleO::Services::NovaApi
- OS::TripleO::Services::NovaMetadata
- OS::TripleO::Services::NovaScheduler
- OS::TripleO::Services::NovaConsoleauth
- OS::TripleO::Services::NovaVncProxy
- OS::TripleO::Services::Ntp
- OS::TripleO::Services::SwiftProxy
- OS::TripleO::Services::SwiftStorage
- OS::TripleO::Services::SwiftRingBuilder
- OS::TripleO::Services::Snmp
- OS::TripleO::Services::Timezone
- OS::TripleO::Services::CeilometerApi
- OS::TripleO::Services::CeilometerCollector
- OS::TripleO::Services::CeilometerExpirer
- OS::TripleO::Services::CeilometerAgentCentral
- OS::TripleO::Services::CeilometerAgentNotification
- OS::TripleO::Services::Horizon
- OS::TripleO::Services::GnocchiApi
- OS::TripleO::Services::GnocchiMetricd
- OS::TripleO::Services::GnocchiStatsd
- OS::TripleO::Services::ManilaApi
- OS::TripleO::Services::ManilaScheduler
- OS::TripleO::Services::ManilaBackendGeneric
- OS::TripleO::Services::ManilaBackendNetapp
- OS::TripleO::Services::ManilaBackendCephFs
- OS::TripleO::Services::ManilaShare
- OS::TripleO::Services::AodhApi
- OS::TripleO::Services::AodhEvaluator
- OS::TripleO::Services::AodhNotifier
- OS::TripleO::Services::AodhListener
- OS::TripleO::Services::SaharaApi
- OS::TripleO::Services::SaharaEngine
- OS::TripleO::Services::IronicApi
- OS::TripleO::Services::IronicConductor
- OS::TripleO::Services::Novalronic
- OS::TripleO::Services::TripleoPackages
- OS::TripleO::Services::TripleoFirewall
- OS::TripleO::Services::OpenDaylightApi
- OS::TripleO::Services::OpenDaylightOvs
- OS::TripleO::Services::SensuClient
- OS::TripleO::Services::FluentdClient
- OS::TripleO::Services::VipHosts

- name: Compute
CountDefault: 1
HostnameFormatDefault: '%stackname%-compute-%index%'
ServicesDefault:

- OS::TripleO::Services::CACerts
- OS::TripleO::Services::CephClient

```
- OS::TripleO::Services::CephExternal
- OS::TripleO::Services::Timezone
- OS::TripleO::Services::Ntp
- OS::TripleO::Services::Snmp
- OS::TripleO::Services::NovaCompute
- OS::TripleO::Services::NovaLibvirt
- OS::TripleO::Services::Kernel
- OS::TripleO::Services::ComputeNeutronCorePlugin
- OS::TripleO::Services::ComputeNeutronOvsAgent
- OS::TripleO::Services::ComputeCeilometerAgent
- OS::TripleO::Services::ComputeNeutronL3Agent
- OS::TripleO::Services::ComputeNeutronMetadataAgent
- OS::TripleO::Services::TripleoPackages
- OS::TripleO::Services::TripleoFirewall
- OS::TripleO::Services::NeutronSriovAgent
- OS::TripleO::Services::OpenDaylightOvs
- OS::TripleO::Services::SensuClient
- OS::TripleO::Services::FluentdClient
- OS::TripleO::Services::VipHosts

- name: BlockStorage
  ServicesDefault:
    - OS::TripleO::Services::CACerts
    - OS::TripleO::Services::BlockStorageCinderVolume
    - OS::TripleO::Services::Kernel
    - OS::TripleO::Services::Ntp
    - OS::TripleO::Services::Timezone
    - OS::TripleO::Services::Snmp
    - OS::TripleO::Services::TripleoPackages
    - OS::TripleO::Services::TripleoFirewall
    - OS::TripleO::Services::SensuClient
    - OS::TripleO::Services::FluentdClient
    - OS::TripleO::Services::VipHosts

- name: ObjectStorage
  ServicesDefault:
    - OS::TripleO::Services::CACerts
    - OS::TripleO::Services::Kernel
    - OS::TripleO::Services::Ntp
    - OS::TripleO::Services::SwiftStorage
    - OS::TripleO::Services::SwiftRingBuilder
    - OS::TripleO::Services::Snmp
    - OS::TripleO::Services::Timezone
    - OS::TripleO::Services::TripleoPackages
    - OS::TripleO::Services::TripleoFirewall
    - OS::TripleO::Services::SensuClient
    - OS::TripleO::Services::FluentdClient
    - OS::TripleO::Services::VipHosts

- name: CephStorage
  ServicesDefault:
    - OS::TripleO::Services::CACerts
    - OS::TripleO::Services::CephOSD
    - OS::TripleO::Services::Kernel
    - OS::TripleO::Services::Ntp
    - OS::TripleO::Services::Snmp
    - OS::TripleO::Services::Timezone
    - OS::TripleO::Services::TripleoPackages
    - OS::TripleO::Services::TripleoFirewall
    - OS::TripleO::Services::SensuClient
    - OS::TripleO::Services::FluentdClient
    - OS::TripleO::Services::VipHosts

- name: OsdCompute
  CountDefault: 0
  HostnameFormatDefault: '%stackname%-osd-compute-%index%'
  ServicesDefault:
    - OS::TripleO::Services::CephOSD
    - OS::TripleO::Services::CACerts
    - OS::TripleO::Services::CephClient
    - OS::TripleO::Services::CephExternal
    - OS::TripleO::Services::Timezone
    - OS::TripleO::Services::Ntp
    - OS::TripleO::Services::Snmp
    - OS::TripleO::Services::NovaCompute
    - OS::TripleO::Services::NovaLibvirt
```

```
- OS::TripleO::Services::Kernel
- OS::TripleO::Services::ComputeNeutronCorePlugin
- OS::TripleO::Services::ComputeNeutronOvsAgent
- OS::TripleO::Services::ComputeCeilometerAgent
- OS::TripleO::Services::ComputeNeutronL3Agent
- OS::TripleO::Services::ComputeNeutronMetadataAgent
- OS::TripleO::Services::TripleoPackages
- OS::TripleO::Services::TripleoFirewall
- OS::TripleO::Services::NeutronSriovAgent
- OS::TripleO::Services::OpenDaylightOvs
- OS::TripleO::Services::SensuClient
- OS::TripleO::Services::FluentdClient
- OS::TripleO::Services::VipHosts
```

8.1.4 layout.yaml

```
resource_registry:

  OS::TripleO::Controller::Ports::InternalApiPort:
    /usr/share/openstack-tripleo-heat-templates/network/ports/internal_api_from_pool.yaml
  OS::TripleO::Controller::Ports::TenantPort: /usr/share/openstack-tripleo-heat-templates/network/ports/tenant_from_pool.yaml
  OS::TripleO::Controller::Ports::StoragePort: /usr/share/openstack-tripleo-heat-templates/network/ports/storage_from_pool.yaml
  OS::TripleO::Controller::Ports::StorageMgmtPort:
    /usr/share/openstack-tripleo-heat-templates/network/ports/storage_mgmt_from_pool.yaml

  OS::TripleO::OsdCompute::Ports::InternalApiPort:
    /usr/share/openstack-tripleo-heat-templates/network/ports/internal_api_from_pool.yaml
  OS::TripleO::OsdCompute::Ports::TenantPort: /usr/share/openstack-tripleo-heat-templates/network/ports/tenant_from_pool.yaml
  OS::TripleO::OsdCompute::Ports::StoragePort:
    /usr/share/openstack-tripleo-heat-templates/network/ports/storage_from_pool.yaml
  OS::TripleO::OsdCompute::Ports::StorageMgmtPort:
    /usr/share/openstack-tripleo-heat-templates/network/ports/storage_mgmt_from_pool.yaml

parameter_defaults:
  NtpServer: x.x.x.x

  ControllerCount: 3
  ComputeCount: 0
  CephStorageCount: 0
  OsdComputeCount: 3

  ControllerSchedulerHints:
    'capabilities:node': 'controller-%index%'
  NovaComputeSchedulerHints:
    'capabilities:node': 'compute-%index%'
  CephStorageSchedulerHints:
    'capabilities:node': 'ceph-storage-%index%'
  OsdComputeSchedulerHints:
    'capabilities:node': 'osd-compute-%index%'

  ControllerIPs:
    internal_api:
      - 172.16.0.10
      - 172.16.0.11
      - 172.16.0.12
    tenant:
      - 172.17.0.10
      - 172.17.0.11
      - 172.17.0.12
    storage:
      - 172.18.0.10
      - 172.18.0.11
      - 172.18.0.12
    storage_mgmt:
      - 172.19.0.10
      - 172.19.0.11
      - 172.19.0.12

  OsdComputeIPs:
    internal_api:
      - 172.16.0.20
      - 172.16.0.21
      - 172.16.0.22
```

```
- 172.16.0.23
- 172.16.0.24
- 172.16.0.25
- 172.16.0.26
- 172.16.0.27
tenant:
- 172.17.0.20
- 172.17.0.21
- 172.17.0.22
- 172.17.0.23
- 172.17.0.24
- 172.17.0.25
- 172.17.0.26
- 172.17.0.27
storage:
- 172.18.0.20
- 172.18.0.21
- 172.18.0.22
- 172.18.0.23
- 172.18.0.24
- 172.18.0.25
- 172.18.0.26
- 172.18.0.27
storage_mgmt:
- 172.19.0.20
- 172.19.0.21
- 172.19.0.22
- 172.19.0.23
- 172.19.0.24
- 172.19.0.25
- 172.19.0.26
- 172.19.0.27
```

TimeZone: 'Japan'

8.1.5 network.yaml

```
resource_registry:
  OS::TripleO::OsdCompute::Net::SoftwareConfig: /home/stack/custom-templates/nic-configs/compute-nics.yaml
  OS::TripleO::Controller::Net::SoftwareConfig: /home/stack/custom-templates/nic-configs/controller-nics.yaml

parameter_defaults:
  NeutronBridgeMappings: 'datacentre:br-ex,tenant:br-tenant'
  NeutronNetworkType: 'vxlan'
  NeutronTunnelType: 'vxlan'
  NeutronExternalNetworkBridge: ""

# Internal API used for private OpenStack Traffic
InternalApiNetCidr: 172.16.0.0/24
InternalApiAllocationPools: [{'start': '172.16.0.100', 'end': '172.16.0.200'}]
InternalApiVirtualFixedIPs: [{'ip_address': '172.16.0.9'}]
RedisVirtualFixedIPs: [{'ip_address': '172.16.0.8'}]
InternalApiNetworkVlanID: 201

# Tenant Network Traffic - will be used for VXLAN over VLAN
TenantNetCidr: 172.17.0.0/24
TenantAllocationPools: [{'start': '172.17.0.100', 'end': '172.17.0.200'}]
TenantNetworkVlanID: 204

# Public Storage Access - e.g. Nova/Glance <--> Ceph
StorageNetCidr: 172.18.0.0/24
StorageAllocationPools: [{'start': '172.18.0.100', 'end': '172.18.0.200'}]
StorageVirtualFixedIPs: [{'ip_address': '172.18.0.9'}]
StorageNetworkVlanID: 202

# Private Storage Access - i.e. Ceph background cluster/replication
StorageMgmtNetCidr: 172.19.0.0/24
StorageMgmtAllocationPools: [{'start': '172.19.0.100', 'end': '172.19.0.200'}]
StorageMgmtVirtualFixedIPs: [{'ip_address': '172.19.0.9'}]
StorageMgmtNetworkVlanID: 203

# External Networking Access - Public API Access
ExternalNetCidr: 10.3.44.0/24
```

```
# Leave room for floating IPs in the External allocation pool (if required)
ExternalAllocationPools: [{'start': '10.3.44.128', 'end': '10.3.44.191'}]
PublicVirtualFixedIPs: [{'ip_address': '10.3.44.127'}]
# Set to the router gateway on the external network
ExternalInterfaceDefaultRoute: 10.3.44.254

# Gateway router for the provisioning network (or Undercloud IP)
ControlPlaneDefaultRoute: 192.0.2.1
# The IP address of the EC2 metadata server. Generally the IP of the Undercloud
EC2MetadataIp: 192.0.2.1
# Define the DNS servers (maximum 2) for the Overcloud nodes
DnsServers: ["10.1.3.1", "10.1.3.2"]
```

8.1.6 ceph.yaml

```
resource_registry:
  OS::TripleO::NodeUserData: /home/stack/custom-templates/first-boot-template.yaml
  OS::TripleO::NodeExtraConfigPost: /home/stack/custom-templates/post-deploy-template.yaml

parameter_defaults:
  ExtraConfig:
    ceph::profile::params::osd_pool_default_pg_num: 256
    ceph::profile::params::osd_pool_default_pgp_num: 256
    ceph::profile::params::osd_pool_default_size: 3
    ceph::profile::params::osd_pool_default_min_size: 2
    ceph::profile::params::osd_recovery_max_active: 3
    ceph::profile::params::osd_max_backfills: 1
    ceph::profile::params::osd_recovery_op_priority: 2
  OsdComputeExtraConfig:
    ceph::profile::params::osd_journal_size: 81920
    ceph::profile::params::osds:
      '/dev/sdc':
        journal: '/dev/sdb'
      '/dev/sdd':
        journal: '/dev/sdb'
      '/dev/sde':
        journal: '/dev/sdb'
```

8.1.7 first-boot-template.yaml

```
heat_template_version: 2014-10-16

description: >
  Wipe and convert all disks to GPT (except the disk containing the root file system)

resources:
  userdata:
    type: OS::Heat::MultipartMime
    properties:
      parts:
        - config: {get_resource: wipe_disk}

  wipe_disk:
    type: OS::Heat::SoftwareConfig
    properties:
      config: {get_file: wipe-disk.sh}

outputs:
  OS::stack_id:
    value: {get_resource: userdata}
```

8.1.8 wipe-disk.sh

```
#!/usr/bin/env bash
if [[ `hostname` = *"ceph"* ]] || [[ `hostname` = *"osd-compute"* ]]
then
```

```
echo "Number of disks detected: $(lsblk -no NAME,TYPE,MOUNTPOINT | grep "disk" | awk '{print $1}' | wc -l)"
for DEVICE in `lsblk -no NAME,TYPE,MOUNTPOINT | grep "disk" | awk '{print $1}'`
do
    ROOTFOUND=0
    echo "Checking /dev/$DEVICE..."
    echo "Number of partitions on /dev/$DEVICE: $(expr $(lsblk -n /dev/$DEVICE | awk '{print $7}' | wc -l) - 1)"
    for MOUNTS in `lsblk -n /dev/$DEVICE | awk '{print $7}'`
    do
        if [ "$MOUNTS" = "/" ]
        then
            ROOTFOUND=1
        fi
    done
    if [ $ROOTFOUND = 0 ]
    then
        echo "Root not found in /dev/${DEVICE}"
        echo "Wiping disk /dev/${DEVICE}"
        sgdisk -Z /dev/${DEVICE}
        sgdisk -g /dev/${DEVICE}
    else
        echo "Root found in /dev/${DEVICE}"
    fi
done
fi
```

8.1.9 post-deploy-template.yaml

```
heat_template_version: 2014-10-16

parameters:
  servers:
    type: json

resources:

  ExtraConfig:
    type: OS::Heat::SoftwareConfig
    properties:
      group: script
      inputs:
        - name: OSD_NUMA_INTERFACE
          config: {get_file: numa-systemmd-osd.sh}

  ExtraDeployments:
    type: OS::Heat::SoftwareDeployments
    properties:
      servers: {get_param: servers}
      config: {get_resource: ExtraConfig}
      input_values:
        OSD_NUMA_INTERFACE: 'p1p1'
      actions: ['CREATE']
```

8.1.10 numa-systemmd-osd.sh

```
#!/usr/bin/env bash
{
if [[ `hostname` = *"ceph"* ]] || [[ `hostname` = *"osd-compute"* ]]; then

    # Verify the passed network interface exists
    if [[ ! $(ip add show $OSD_NUMA_INTERFACE) ]]; then
        exit 1
    fi

    # If NUMA related packages are missing, then install them
    # If packages are baked into image, no install attempted
    for PKG in numactl hwloc; do
        if [[ ! $(rpm -q $PKG) ]]; then
            yum install -y $PKG
            if [[ ! $? ]]; then

```

```
        echo "Unable to install $PKG with yum"
        exit 1
    fi
fi
done

# Find the NUMA socket of the $OSD_NUMA_INTERFACE
declare -A NUMASOCKET
while read TYPE SOCKET_NUM NIC ; do
    if [[ "$TYPE" == "NUMANode" ]]; then
        NUMASOCKET=$(echo $SOCKET_NUM | sed s/L/g);
    fi
    if [[ "$NIC" == "$OSD_NUMA_INTERFACE" ]]; then
        # because $NIC is the $OSD_NUMA_INTERFACE,
        # the NUMASOCKET has been set correctly above
        break # so stop looking
    fi
done < <(lstopo-no-graphics | tr -d [:punct:] | egrep "NUMANode|$OSD_NUMA_INTERFACE")

if [[ -z $NUMASOCKET ]]; then
    echo "No NUMANode found for $OSD_NUMA_INTERFACE. Exiting."
    exit 1
fi

UNIT='/usr/lib/systemd/system/ceph-osd@.service'
# Preserve the original ceph-osd start command
CMD=$(crudini --get $UNIT Service ExecStart)

if [[ $(echo $CMD | grep numactl) ]]; then
    echo "numactl already in $UNIT. No changes required."
    exit 0
fi

# NUMA control options to append in front of $CMD
NUMA="/usr/bin/numactl -N $NUMASOCKET --preferred=$NUMASOCKET"

# Update the unit file to start with numactl
# TODO: why doesn't a copy of $UNIT in /etc/systemd/system work with numactl?
crudini --verbose --set $UNIT Service ExecStart "$NUMA $CMD"

# Reload so updated file is used
systemctl daemon-reload

# Restart OSDs with NUMA policy (print results for log)
OSD_IDS=$(ls /var/lib/ceph/osd | awk 'BEGIN { FS = "-" } ; { print $2 }')
for OSD_ID in $OSD_IDS; do
    echo -e "¥nStatus of OSD $OSD_ID before unit file update¥n"
    systemctl status ceph-osd@$OSD_ID
    echo -e "¥nRestarting OSD $OSD_ID¥n"
    systemctl restart ceph-osd@$OSD_ID
    echo -e "¥nStatus of OSD $OSD_ID after unit file update¥n"
    systemctl status ceph-osd@$OSD_ID
done
fi
} 2>&1 > /root/post_deploy_heat_output.txt
```

8.1.11 compute.yaml

```
parameter_defaults:
  ExtraConfig:
    nova::compute::reserved_host_memory: 20000
    nova::cpu_allocation_ratio: 0.875
    nova::block_device_allocate_retries: 90
    nova::block_device_allocate_retries_interval: 10
```

8.1.12 nic-configs/controller-nics.yaml

```
heat_template_version: 2015-04-30
```

```
description: >
  Software Config to drive os-net-config to configure VLANs for the
  controller role.

parameters:
  ControlPlaneIp:
    default: ""
    description: IP address/subnet on the ctlplane network
    type: string
  ExternalIpSubnet:
    default: ""
    description: IP address/subnet on the external network
    type: string
  InternalApiIpSubnet:
    default: ""
    description: IP address/subnet on the internal API network
    type: string
  StorageIpSubnet:
    default: ""
    description: IP address/subnet on the storage network
    type: string
  StorageMgmtIpSubnet:
    default: ""
    description: IP address/subnet on the storage mgmt network
    type: string
  TenantIpSubnet:
    default: ""
    description: IP address/subnet on the tenant network
    type: string
  ManagementIpSubnet: # Only populated when including environments/network-management.yaml
    default: ""
    description: IP address/subnet on the management network
    type: string
  ExternalNetworkVlanID:
    default: 10
    description: Vlan ID for the external network traffic.
    type: number
  InternalApiNetworkVlanID:
    default: 20
    description: Vlan ID for the internal_api network traffic.
    type: number
  StorageNetworkVlanID:
    default: 30
    description: Vlan ID for the storage network traffic.
    type: number
  StorageMgmtNetworkVlanID:
    default: 40
    description: Vlan ID for the storage mgmt network traffic.
    type: number
  TenantNetworkVlanID:
    default: 50
    description: Vlan ID for the tenant network traffic.
    type: number
  ManagementNetworkVlanID:
    default: 60
    description: Vlan ID for the management network traffic.
    type: number
  ControlPlaneDefaultRoute: # Override this via parameter_defaults
    description: The default route of the control plane network.
    type: string
  ExternalInterfaceDefaultRoute:
    default: '10.0.0.1'
    description: default route for the external network
    type: string
  ManagementInterfaceDefaultRoute: # Commented out by default in this template
    default: unset
    description: The default route of the management network.
    type: string
  ControlPlaneSubnetCidr: # Override this via parameter_defaults
    default: '24'
    description: The subnet CIDR of the control plane network.
    type: string
  DnsServers: # Override this via parameter_defaults
    default: []
    description: A list of DNS servers (2 max for some implementations) that will be added to resolv.conf.
    type: comma_delimited_list
```

```

EC2MetadataIp: # Override this via parameter_defaults
description: The IP address of the EC2 metadata server.
type: string

resources:
  OsNetConfigImpl:
    type: OS::Heat::StructuredConfig
    properties:
      group: os-apply-config
      config:
        os_net_config:
          network_config:
            -
              type: interface
              name: nic1
              use_dhcp: false
              dns_servers: {get_param: DnsServers}
              addresses:
                -
                  ip_netmask:
                    list_join:
                      - '/'
                      - - {get_param: ControlPlaneIp}
                      - {get_param: ControlPlaneSubnetCidr}
              routes:
                -
                  ip_netmask: 169.254.169.254/32
                  next_hop: {get_param: EC2MetadataIp}
            -
              type: ovs_bridge
              # Assuming you want to keep br-ex as external bridge name
              name: {get_input: bridge_name}
              use_dhcp: false
              addresses:
                -
                  ip_netmask: {get_param: ExternalIpSubnet}
              routes:
                -
                  ip_netmask: 0.0.0.0/0
                  next_hop: {get_param: ExternalInterfaceDefaultRoute}
              members:
                -
                  type: interface
                  name: p2p4
                  # force the MAC address of the bridge to this interface
                  primary: true
            -
              type: interface
              name: nic3
              use_dhcp: false
              mtu: 9000
            -
              type: vlan
              device: nic3
              mtu: 9000
              use_dhcp: false
              vlan_id: {get_param: StorageMgmtNetworkVlanID}
              addresses:
                -
                  ip_netmask: {get_param: StorageMgmtIpSubnet}
            -
              type: vlan
              device: nic3
              mtu: 9000
              use_dhcp: false
              vlan_id: {get_param: StorageNetworkVlanID}
              addresses:
                -
                  ip_netmask: {get_param: StorageIpSubnet}
            -
              # VLAN for VXLAN tenant networking
              type: ovs_bridge
              name: br-tenant
              mtu: 9000
              use_dhcp: false
              members:

```

```

-
  type: interface
  name: nic4
  mtu: 9000
  use_dhcp: false
  # force the MAC address of the bridge to this interface
  primary: true
-
  type: vlan
  mtu: 9000
  vlan_id: {get_param: TenantNetworkVlanID}
  addresses:
  -
    ip_netmask: {get_param: TenantIpSubnet}
-
  type: interface
  name: nic2
  use_dhcp: false
  addresses:
  -
    ip_netmask: {get_param: InternalApiIpSubnet}
-
  # Unused Interface
  type: interface
  name: p2p1
  use_dhcp: false
  defroute: false
-
  # Unused Interface
  type: interface
  name: p2p2
  use_dhcp: false
  defroute: false
-
  # Unused Interface
  type: interface
  name: p2p3
  use_dhcp: false
  defroute: false

```

outputs:

```

OS::stack_id:
  description: The OsNetConfigImpl resource.
  value: {get_resource: OsNetConfigImpl}

```

8.1.13 nic-configs/compute-nics.yaml

heat_template_version: 2015-04-30

description: >

Software Config to drive os-net-config to configure VLANs for the compute role.

parameters:

```

ControlPlaneIp:
  default: ""
  description: IP address/subnet on the ctlplane network
  type: string
ExternalIpSubnet:
  default: ""
  description: IP address/subnet on the external network
  type: string
InternalApiIpSubnet:
  default: ""
  description: IP address/subnet on the internal API network
  type: string
StorageIpSubnet:
  default: ""
  description: IP address/subnet on the storage network
  type: string
StorageMgmtIpSubnet:
  default: ""
  description: IP address/subnet on the storage mgmt network

```

```

type: string
TenantIpSubnet:
  default: ""
  description: IP address/subnet on the tenant network
  type: string
ManagementIpSubnet: # Only populated when including environments/network-management.yaml
  default: ""
  description: IP address/subnet on the management network
  type: string
InternalApiNetworkVlanID:
  default: 20
  description: Vlan ID for the internal_api network traffic.
  type: number
StorageNetworkVlanID:
  default: 30
  description: Vlan ID for the storage network traffic.
  type: number
StorageMgmtNetworkVlanID:
  default: 40
  description: Vlan ID for the storage mgmt network traffic.
  type: number
TenantNetworkVlanID:
  default: 50
  description: Vlan ID for the tenant network traffic.
  type: number
ManagementNetworkVlanID:
  default: 60
  description: Vlan ID for the management network traffic.
  type: number
ControlPlaneSubnetCidr: # Override this via parameter_defaults
  default: '24'
  description: The subnet CIDR of the control plane network.
  type: string
ControlPlaneDefaultRoute: # Override this via parameter_defaults
  description: The default route of the control plane network.
  type: string
ExternalInterfaceDefaultRoute: # Not used by default in this template
  default: '10.0.0.1'
  description: The default route of the external network.
  type: string
ManagementInterfaceDefaultRoute: # Commented out by default in this template
  default: unset
  description: The default route of the management network.
  type: string
DnsServers: # Override this via parameter_defaults
  default: []
  description: A list of DNS servers (2 max for some implementations) that will be added to resolv.conf.
  type: comma_delimited_list
EC2MetadataIp: # Override this via parameter_defaults
  description: The IP address of the EC2 metadata server.
  type: string

resources:
  OsNetConfigImpl:
    type: OS::Heat::StructuredConfig
    properties:
      group: os-apply-config
      config:
        os_net_config:
          network_config:
            -
              type: interface
              name: em1
              use_dhcp: false
              dns_servers: {get_param: DnsServers}
              addresses:
                -
                  ip_netmask:
                    list_join:
                      - '/'
                      - {get_param: ControlPlaneIp}
                      - {get_param: ControlPlaneSubnetCidr}
            routes:
              -
                ip_netmask: 169.254.169.254/32
                next_hop: {get_param: EC2MetadataIp}

```

```
-
  default: true
  next_hop: {get_param: ControlPlaneDefaultRoute}
-
type: interface
name: p1p1
use_dhcp: false
mtu: 9000
-
type: vlan
device: p1p1
mtu: 9000
use_dhcp: false
vlan_id: {get_param: StorageMgmtNetworkVlanID}
addresses:
-
  ip_netmask: {get_param: StorageMgmtIpSubnet}
-
type: vlan
device: p1p1
mtu: 9000
use_dhcp: false
vlan_id: {get_param: StorageNetworkVlanID}
addresses:
-
  ip_netmask: {get_param: StorageIpSubnet}
-
# VLAN for VXLAN tenant networking
type: ovs_bridge
name: br-tenant
mtu: 9000
use_dhcp: false
members:
-
  type: interface
  name: p1p2
  mtu: 9000
  use_dhcp: false
  # force the MAC address of the bridge to this interface
  primary: true
-
  type: vlan
  mtu: 9000
  vlan_id: {get_param: TenantNetworkVlanID}
  addresses:
  -
    ip_netmask: {get_param: TenantIpSubnet}
-
type: interface
name: em2
use_dhcp: false
addresses:
-
  ip_netmask: {get_param: InternalApiIpSubnet}
```

outputs:

```
OS::stack_id:
  description: The OsNetConfigImpl resource.
  value: {get_resource: OsNetConfigImpl}
```

以上