

今回は79,200パターン！
OpenStack&Cephによる
Hyper Converged Infrastructure検証結果

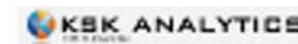
OSCA技術検討会 OpenStack分科会
<http://www.osca-jp.com/>

デル株式会社 日比野 正慶
レッドハット株式会社 佐々木 宏忠
株式会社日立ソリューションズ 工藤 雄大



OSCA のご紹介

2012年2月にインテルとデルが中心となり発足、2016年6月時点で23社が参加。
先進的且つ標準的な技術の情報発信を通じ、参画企業のビジネス活性化を目的とする。
2016年、OSCA v2.0 で、ソリューションスコープを拡張、IOT実現に向けた取り組みを強化。



五十音順



OSCA の活動

- 各分科会による検証と、ブログ、ホワイトペーパーを通じた**情報発信**
- ビジネスにつなげるための、**各種セミナー主催やイベント参加**
- お客様やパートナー様との関係を強化する**技術交流会やパーティーイベント**の開催



OSCA OpenStack分科会



redhat

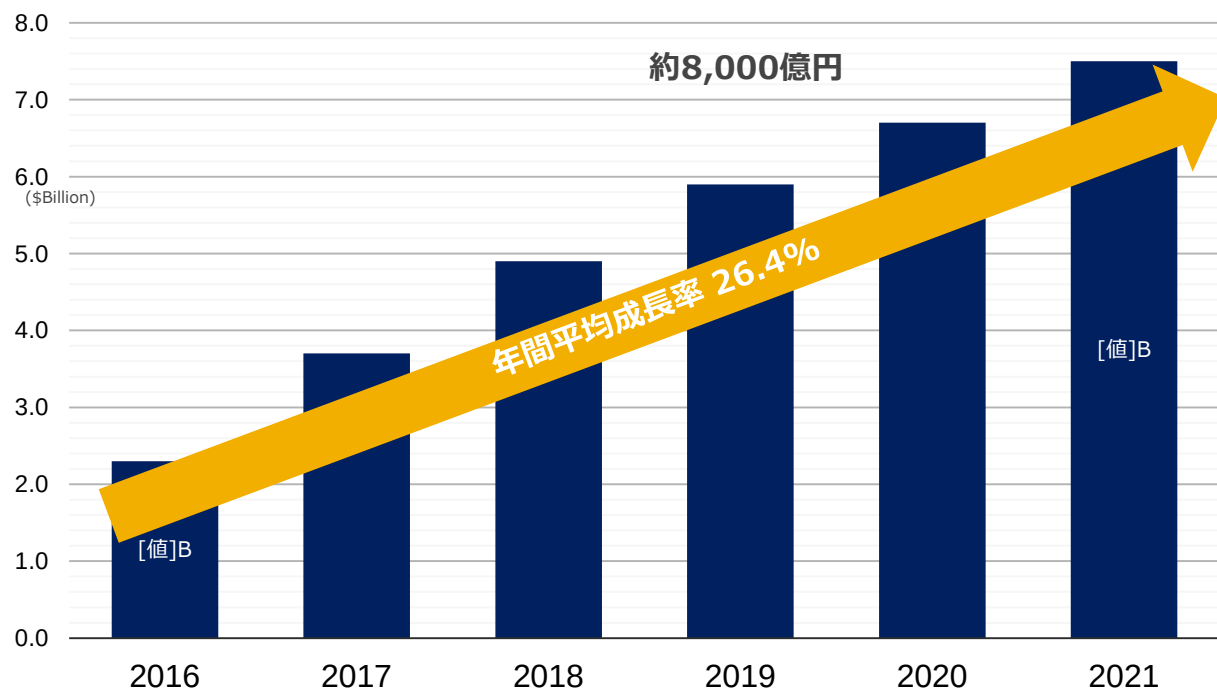
DELL EMC

OSCA OpenStack分科会 活動履歴

- 2012年「サーバ 1 台で構築する Red Hat OpenStack Preview Step by Step Guide」
- 2012年「サーバ 1 台で構築する OpenStack Swift Step by Step Guide」
- 2014年「Red Hat OpenStack 3.0 運用検証報告書 (VM レンタルサービス編)」
- 2014年「OpenStack大規模環境構築におけるネットワーク設計のポイント」
- 2016年「OpenStack on Cephストレージ設計のポイント」
- 2016年「OpenStack on Ceph性能評価」
- 2017年「OpenStack HCIのポイントとベンチマーク結果考察」

なぜOpenStackでHCIか

HCIハードウェア市場予測 (グローバル)



- HCIの注目度が高い
- HCIのマーケットが確実に拡大している
- OpenStackでもHW費用は削減したい
- Red Hat OpenStack Platform 11で正式にOpenStack HCI構成をサポート

HCIは豊富なOpenStackストレージオプションのひとつ

出典: IDC Hyperconverged Systems 2016-2021 Forecast,
IDC Worldwide Converged Systems Tracker Forecast 2016Q4

Dell EMC 最新 OpenStack Solution



Dell EMC Ready Bundle for Red Hat OpenStack Platform

- 完全検証済み Dell EMC Ready Bundleソリューション
- Red Hat 共同開発 安心のOpenStackプラットフォーム
- 信頼のHWと冗長構成によるEnterpriseクラスのOpenStack
- Red Hat OSP OME化による単一保守窓口
- リファレンスガイドやツールにより導入までの時間を大幅削減
- OpenStack 導入自動化ツール Jet Packで迅速・確実な導入
- OpenShift、CloudForms導入ガイドなどの技術情報の提供

RH OpenStack HCI構成のポイント

レッドハット株式会社
ソリューションアーキテクト

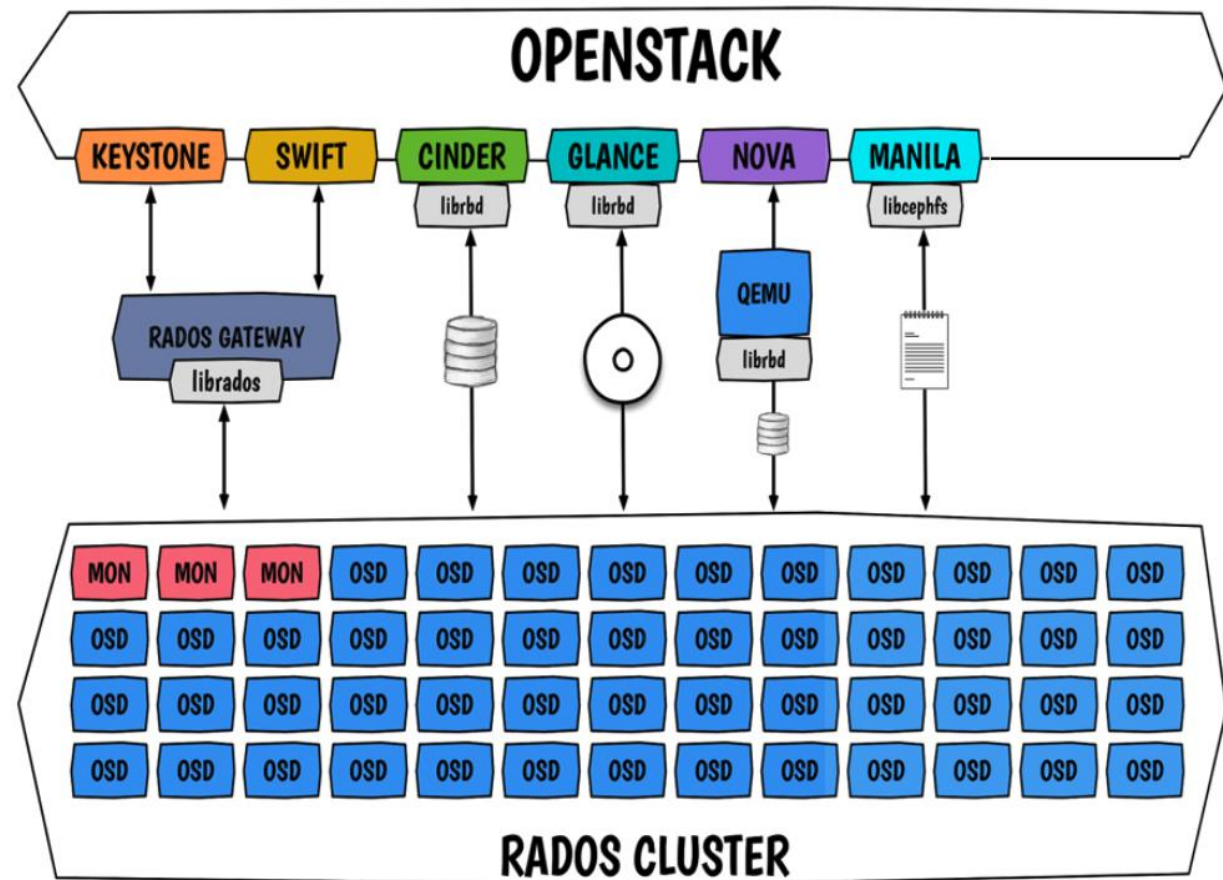
佐々木 宏忠



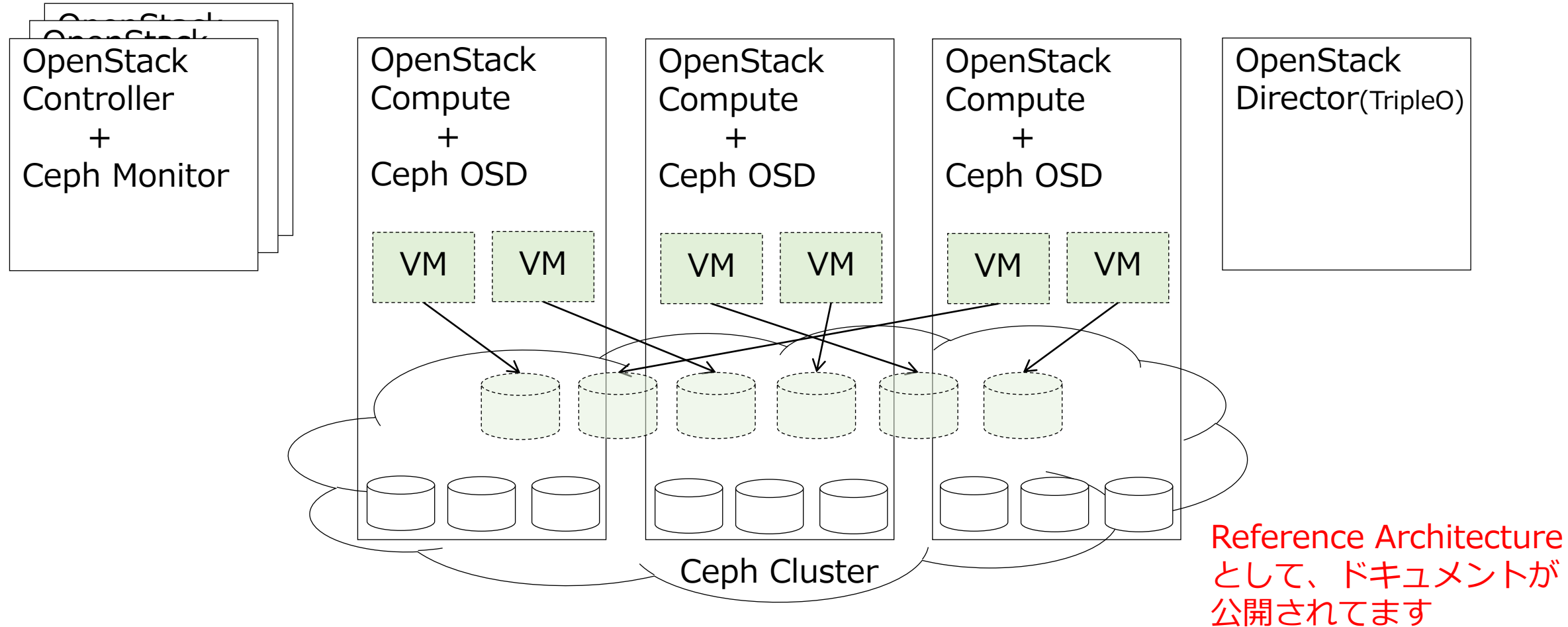
Open
Standard
Cloud
Association

Ceph

- Cephとは
 - オープンソースのSDS
 - 大規模スケールアウト
 - オブジェクトストレージ
+
フロントエンド
 - OpenStackとの統合
- 最近の話題
 - NFS GW for RGW (RHCS 2.3 GA)
 - iSCSI (RHCS 2.1 TP →もう少しでGA)
 - CephFS (TP →もう少しでGA)
 - BlueStore (RHCS2.0 TP →もう少しでGA)



OpenStack & Ceph の HCI



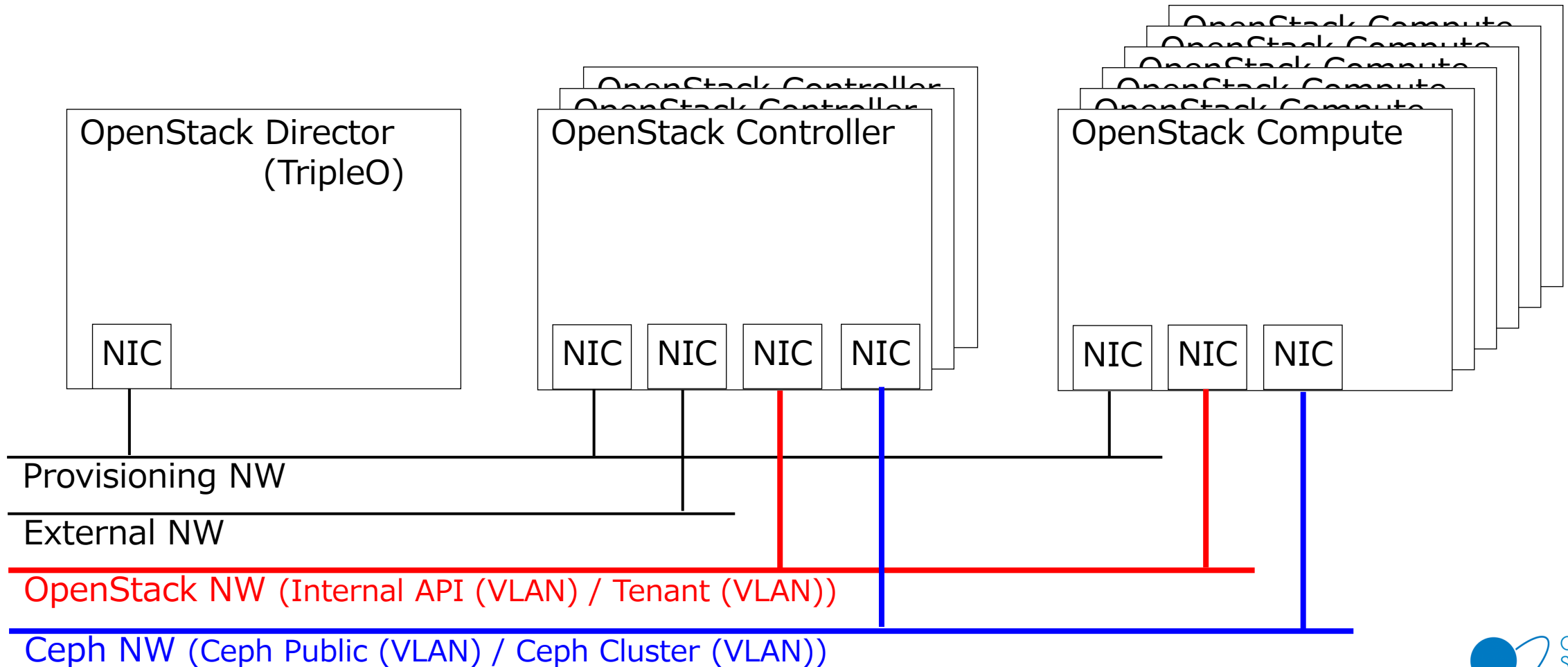
https://access.redhat.com/documentation/en-us/reference_architectures/2017/html-single/hyper-converged_red_hat_openstack_platform_10_and_red_hat_ceph_storage_2/

ComputeノードにOSDを置く時の注意点

- Network
 - Ceph用のネットワークとOpenStack用のネットワークの分離
- Resource Isolation
 - ComputeノードにOSD用のリソース(CPU、メモリ)を確保
 - reserved_host_memory_mb
 - cpu_allocation_ratio
- NUMA pinning
 - OSDデーモンをNICと同じNUMAで動かす
 - lstopo
 - numactl

ComputeノードにOSDを置く時の注意点

- Network
 - Ceph用のネットワークとOpenStack用のネットワークの分離



ComputeノードにOSDを置く時の注意点

- Resource Isolation
 - ComputeノードにOSD用のリソースを確保

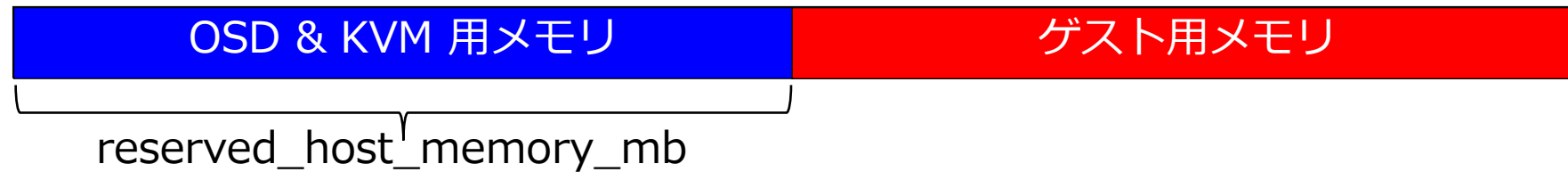
OpenStack Compute + Ceph OSD

CPU:

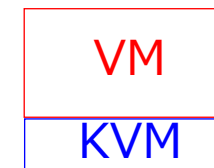
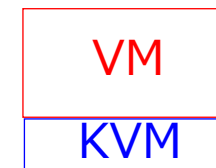
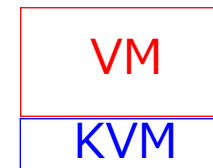
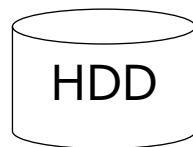


$\text{cpu_allocation_ratio} = (\text{ゲスト用コア} \div \text{平均ゲストCPU使用率}) \div \text{実際のコア数}$

Memory:

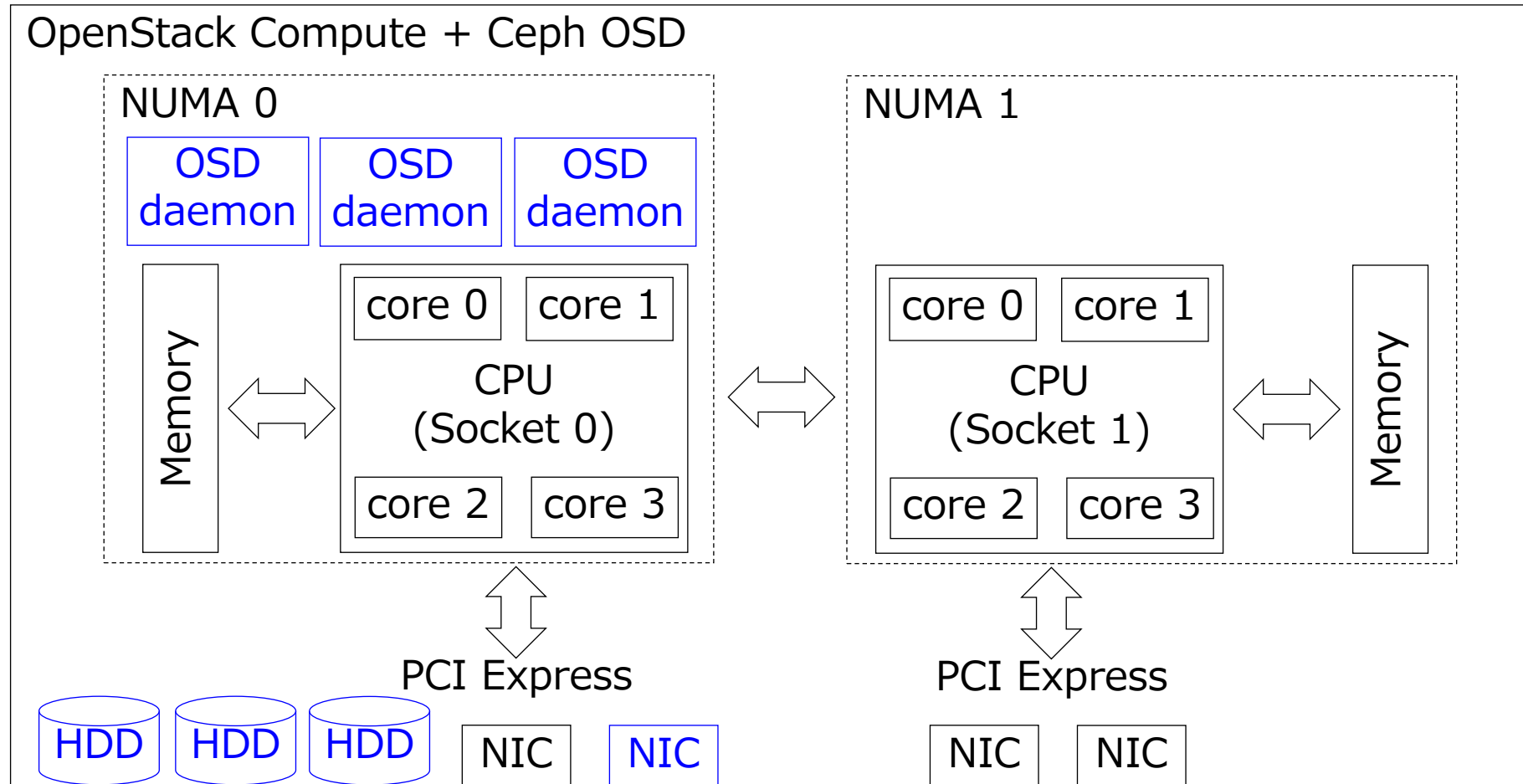


OSD daemon OSD daemon OSD daemon



ComputeノードにOSDを置く時の注意点

- NUMA pinning
 - OSDデーモンをNIC/ストレージコントローラと同じNUMAで動かす



Directorを使ったデプロイ・増設

- Ceph部分も含めてOpenStack Directorでインストール可能
Red Hat OpenStack Platform: Director
Open Source Project: TripleO
- コマンド一発でOSインストールを含めて、OpenStack + Ceph のデプロイが完了
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
- 増設時は設定ファイルを変更し、同じコマンドで増設完了
OsdComputeCount: 3 → 4

ベンチマーク考察と結果

株式会社日立ソリューションズ
技術革新本部 研究開発部
技師

工藤 雄大

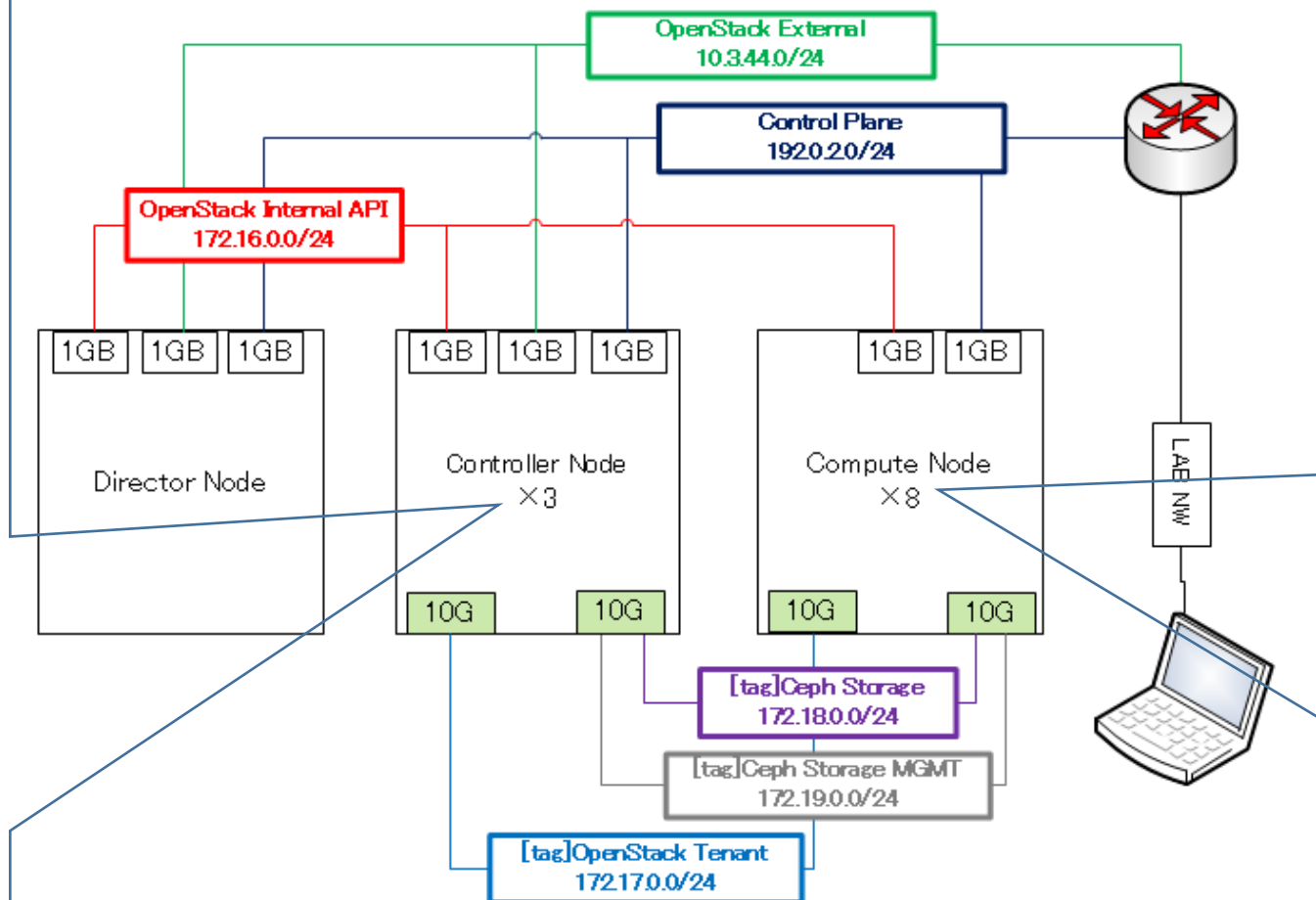


Open
Standard
Cloud
Association

検証環境

■ Controller + Ceph Mon

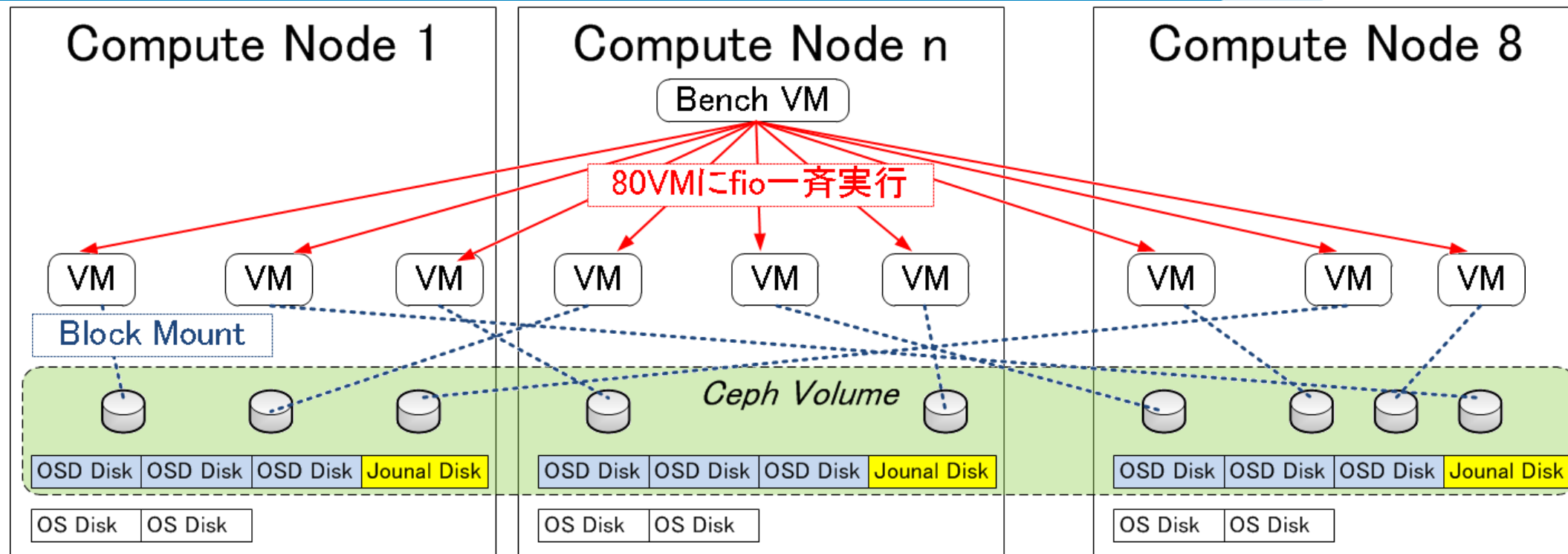
CephMon, CephExternal, CephRgw, **Cinder**Api, CinderScheduler, CinderVolume, **Keystone**, **Glance**Api, GlanceRegistry, **Heat**Api, HeatApiCfn, HeatApiCloudwatch, HeatEngine, NeutronDhcpAgent, NeutronL3Agent, NeutronMetadataAgent, **Neutron**Api, NeutronCorePlugin, NeutronOvsAgent, NovaConductor, **Nova**Api, NovaMetadata, NovaScheduler, NovaConsoleauth, NovaVncProxy, **Ceilometer**Api, **Horizon**, **Ironic**Api, etc.....



■ Nova+ Ceph OSD

CephOSD, CephClient, CephExternal, **NovaCompute**, NovaLibvirt, ComputeNeutronCorePlugin, ComputeNeutronOvsAgent, ComputeCeilometerAgent, **ComputeNeutronL3Agent**, ComputeNeutronMetadataAgent, NeutronSriovAgent etc.....

ベンチマーク概要



■ Compute Node

- 4、5、6、7、8台時のパターンで計測
- OS : 300GB SAS 10Krpm x2(Raid1)
- OSD Disk : 600GB SAS 10Krpm x3 (JBOD)
- Journal : 320GB SSD
- (各OSD用に80GBのVolumeを準備)

■ インスタンス(VM)

- 80 VM作成(Compute Node数によらず)
- CinderからBoot from Volume
- RHEL 7.3上に、fio-2.1.10をインストール
- 上記とは別でベンチマークリモート実行用VMから一斉にfio実施

ベンチマーク方法

BenchVMから、80台のVMへ向けてfioを実行

```
ssh (user@instance IP) 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) & ssh...
```

・下記を変えながら2分ベンチマーク実行、2分sleep

- VM: 80
- [read/write]: randread | randwrite
- [BlockSize]: 4 | 16 | 32 | 64 | 128
- [Job数]: 1 | 4 | 8 <=多くなると負荷高
- 3回繰り返し
- Compute Node: 4 | 5 | 6 | 7 | 8 (台)
- VMキャッシュクリア機構の有無

80

× 2

× 5

× 3

× 3

× 5

× 2

=72,000

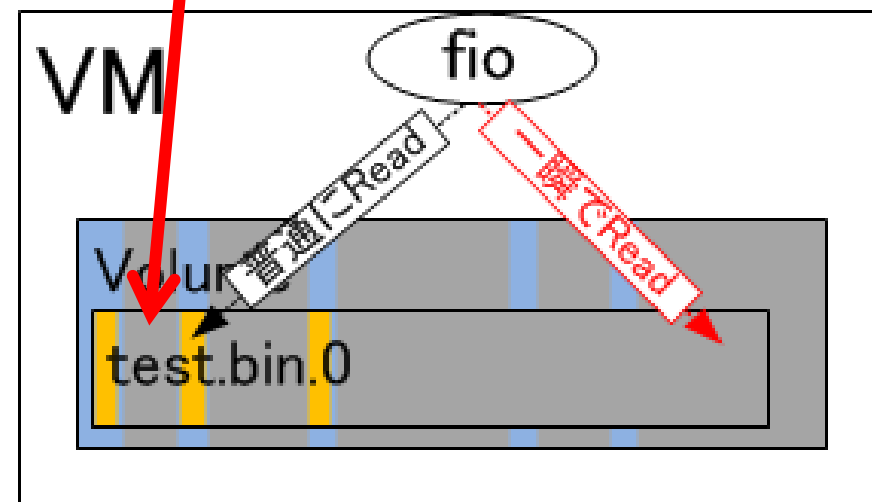
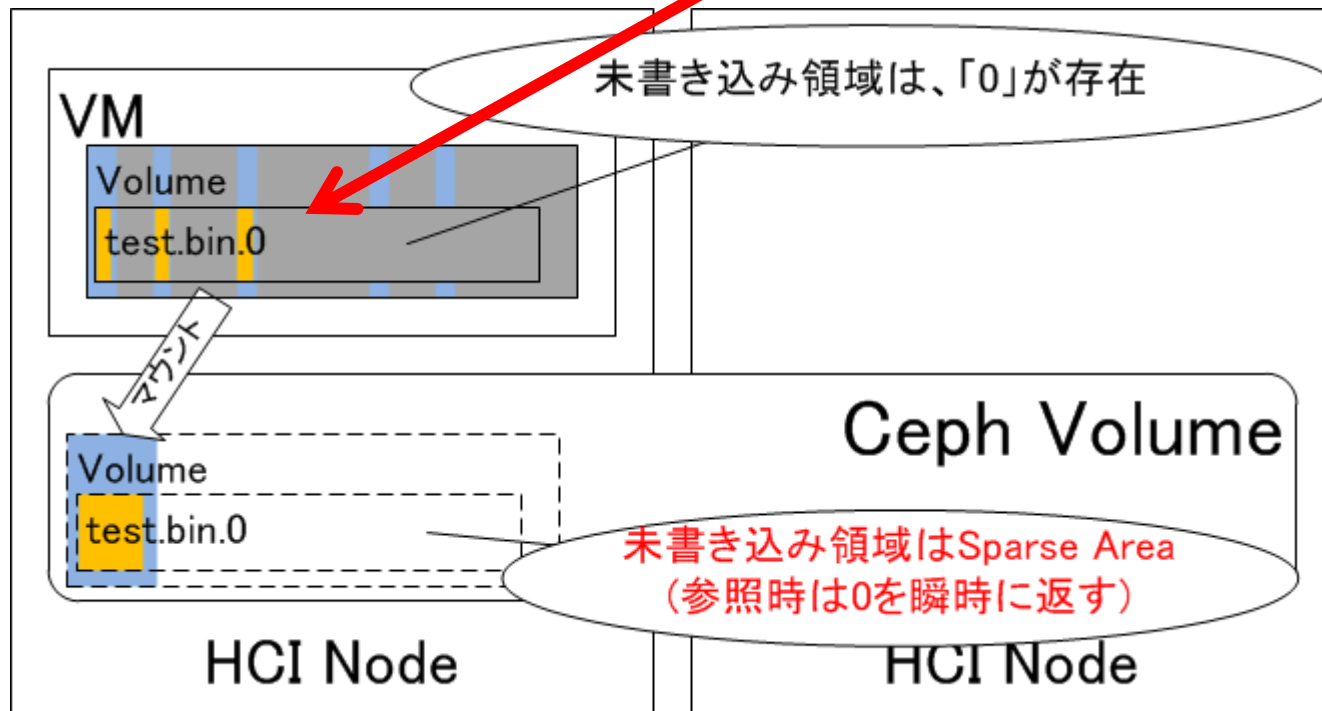
79,200 個の項目

更に追加検証
が7,200件

少し脱線(PoC時注意！)： fio実行順序

■ fioをRandom Write→Random Readの順で実施すると正しく計測できない

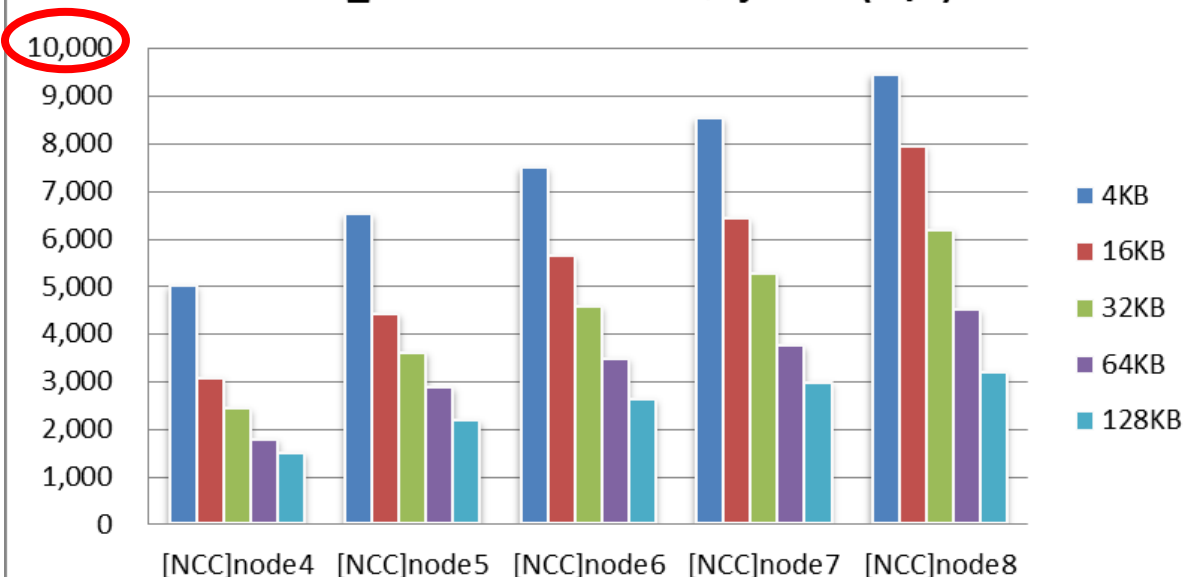
	ベンチマークファイル無し	ベンチマークファイル有り
Random Write	・ベンチマークファイルを作成しながらベンチ実行 ・ベンチ時間不足の場合、不完全なファイルを作成	・既存ファイルを上書きしながら、ベンチマーク実施
Random Read	・正常なベンチマークファイルを自動生成後、ベンチマーク実施	・既存ファイルをそのまま流用



少し脱線：不完全ファイル使用時

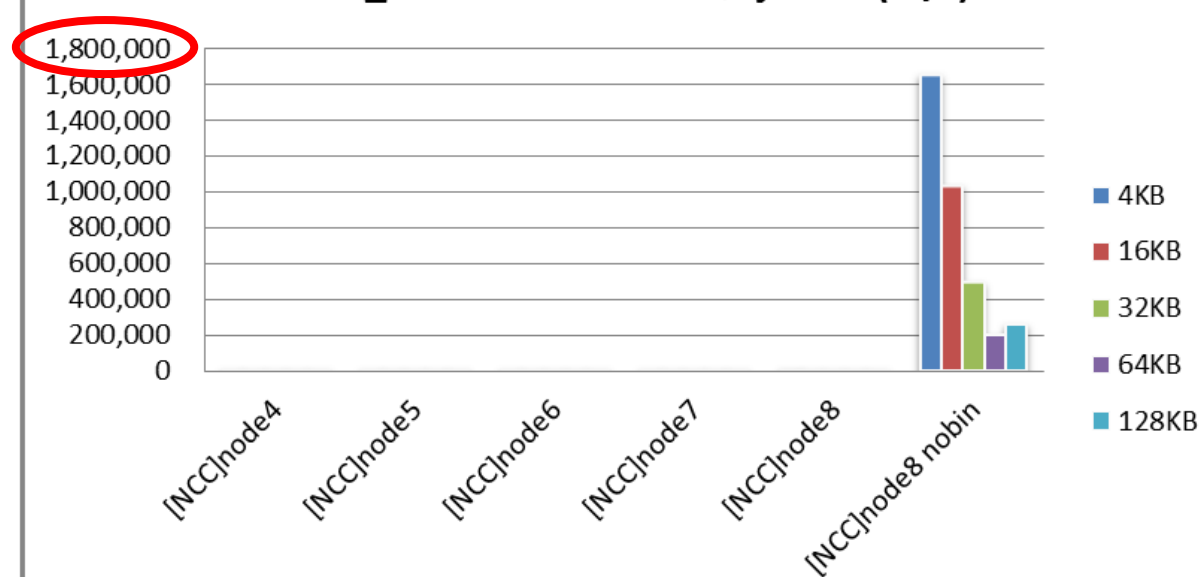
正常なファイル利用

IOPS_R : No Cacheclear、job8 (io/s)



不完全ファイルパターンを右端に追加

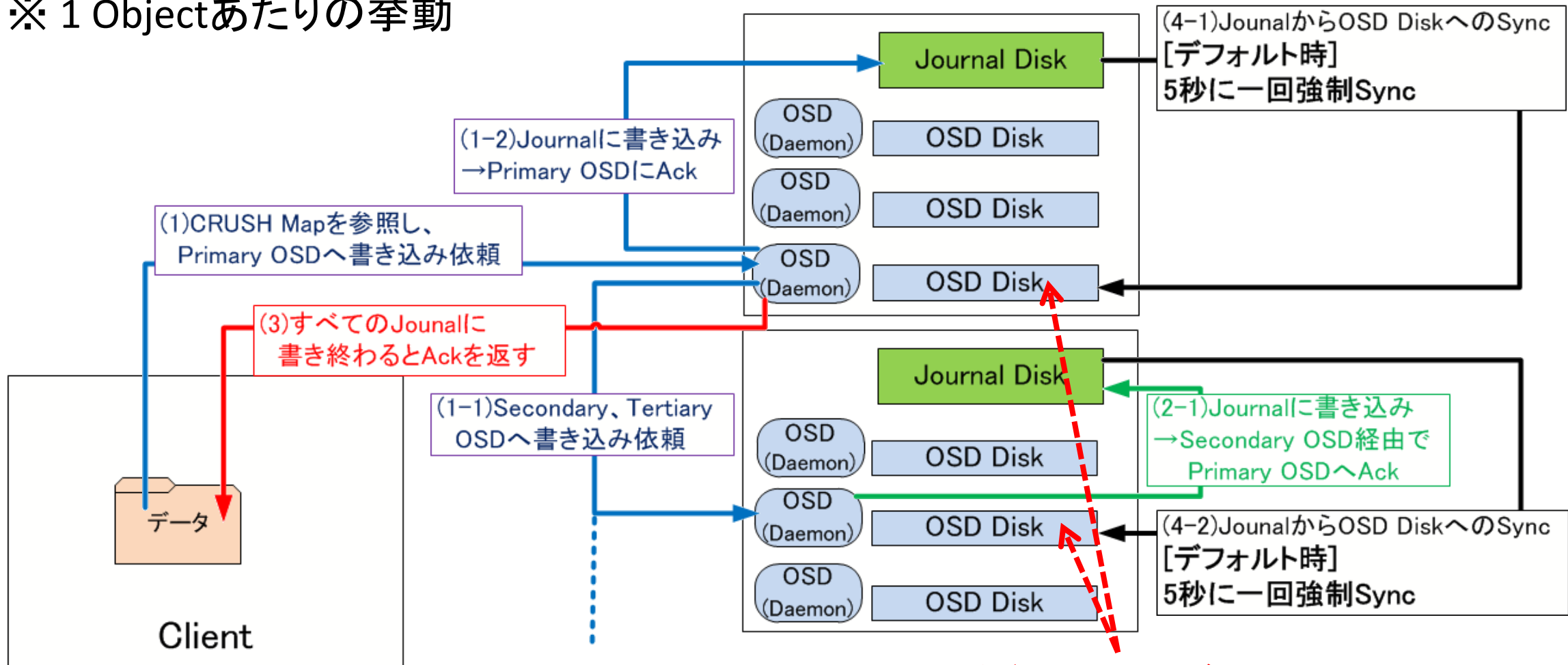
IOPS_R : No Cacheclear、job8 (io/s)



Write先 だめ 絶対

Writeの仕組み

※ 1 Objectあたりの挙動

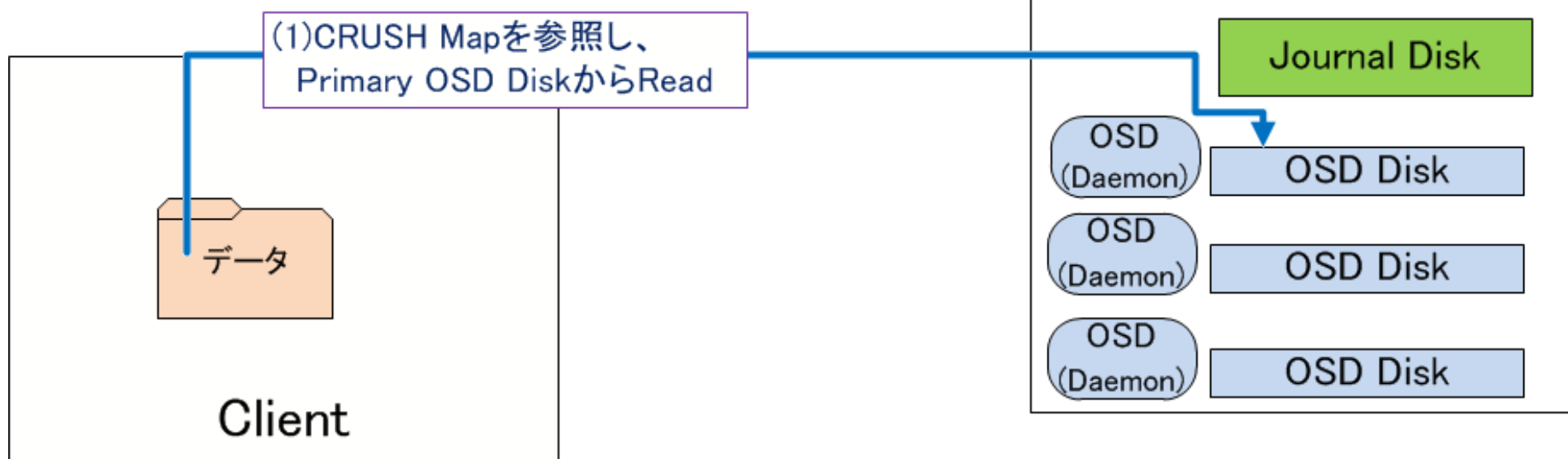


**Journalを経由するが、
最終的にHDDの速度に依存**

Readの仕組み

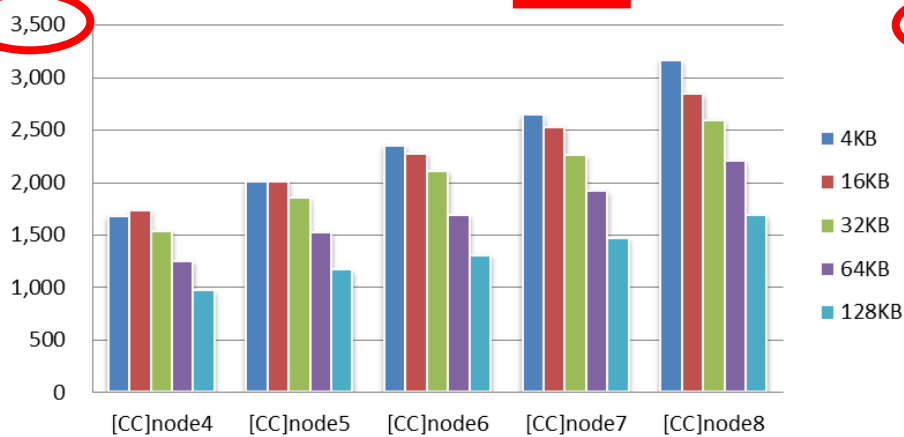
※ 1 Objectあたりの挙動

OSD Diskから直接Read !

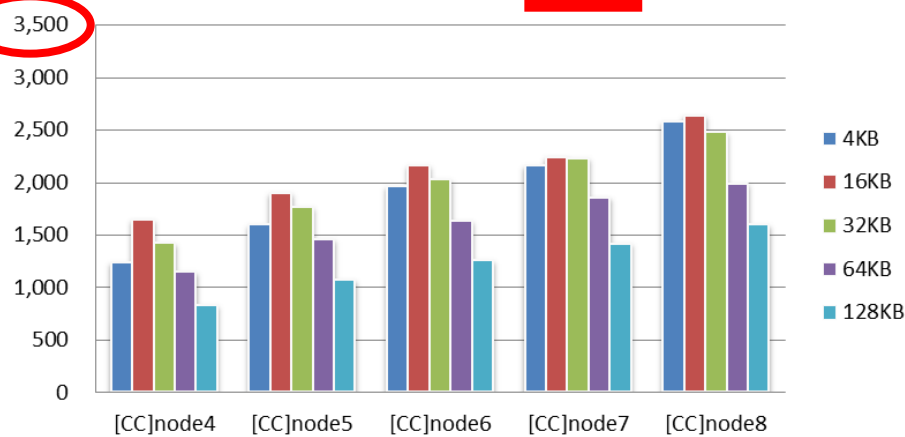


Node増加時 Write

IOPS_W : Cacheclear **job1** (io/s)

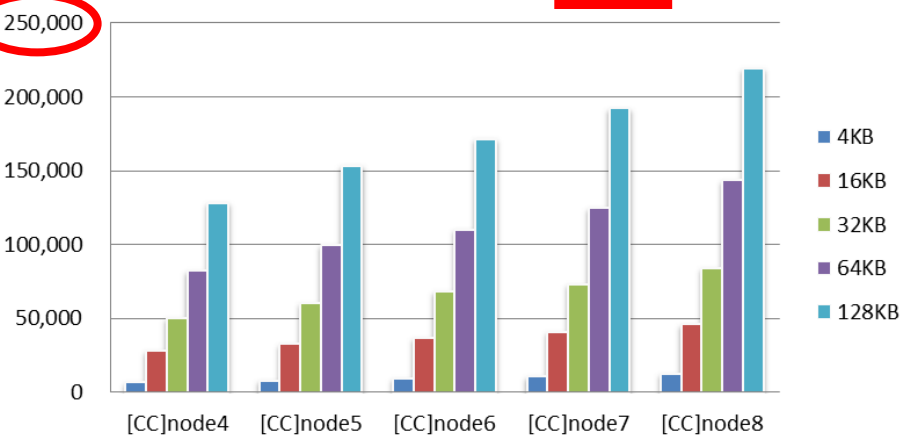


IOPS_W : Cacheclear **job8** (io/s)

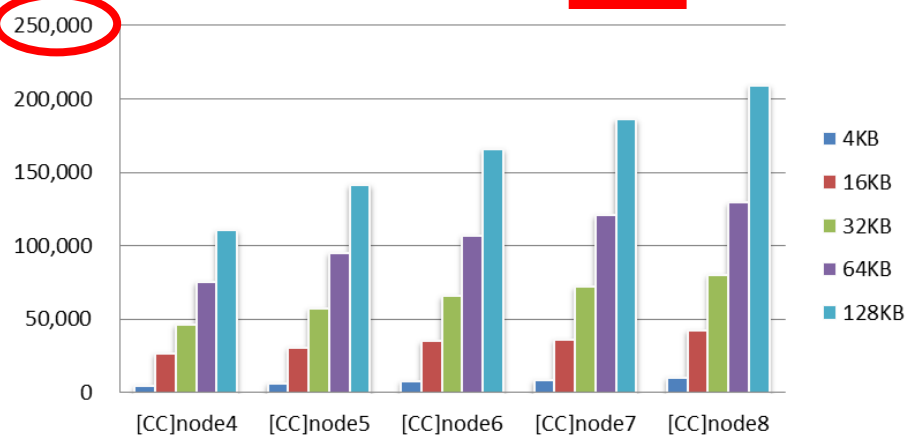


✓ Node数増加に伴い
性能向上

Throughput_W : Cacheclear **job1** (KB/s)



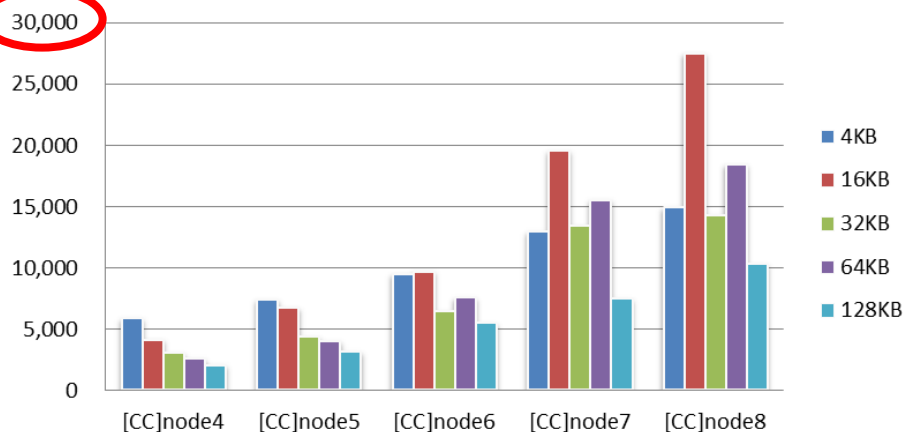
Throughput_W : Cacheclear **job8** (KB/s)



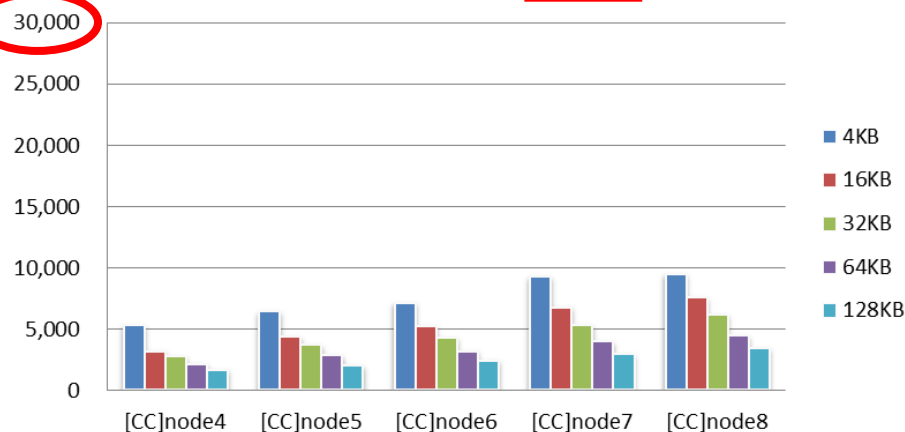
✓ Job数(負荷)を高くしても
あまり変化無し

Node増加時 Read

IOPS_R : Cacheclear、**job1** (io/s)



IOPS_R : Cacheclear、**job8** (io/s)

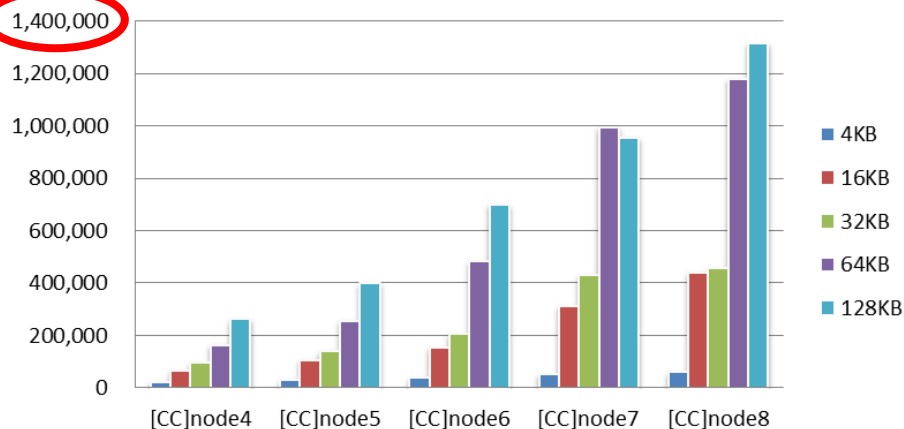


✓ Writeと比較して高性能
- 高負荷時でも3倍
(HDD自体のR/W性能差?)
(Replication数?)

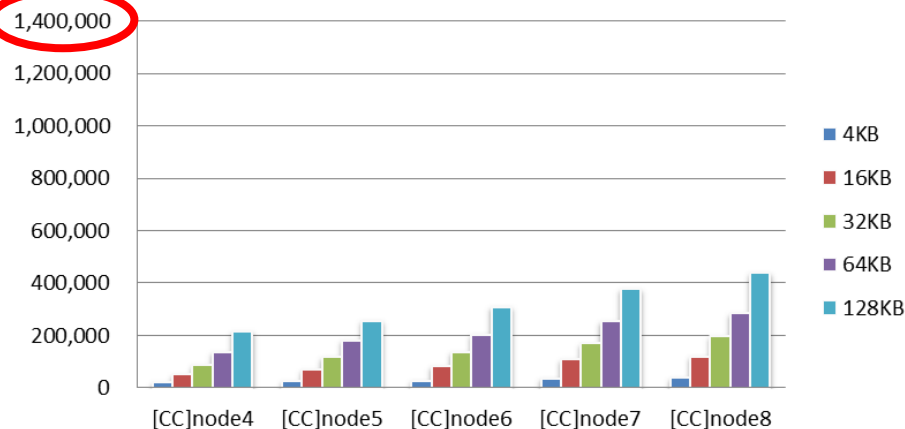
- 低負荷時で7倍程度
(なんらかのキャッシュ?)

✓ Node数増加に伴い
性能向上

Throughput_R : Cacheclear、**job1** (KB/s)



Throughput_R : Cacheclear、**job8** (KB/s)



サイジング指針

■ 高負荷時のサイジング指針 (ノード8、Job8時の値から算出)

IOPS / Throughput	Block Size	Write / Read	全体	1ノードあたり	1 OSD Disk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)

まとめ

- HCI はOpenStack においてもコストを削減できるソリューション
- コンピュートとストレージのリソースを考慮したサイジングが必要
- ローカルキャッシュ機能を有していないため性能要件が高い場合、OSDにSSDを利用するのが効果的
- ホワイトペーパーもご覧ください

<http://ja.community.dell.com/techcenter/m/mediagallery/3798>

今回は79,200パターン！ OpenStack & Cephによる Hyper Converged Infrastructure検証結果

OSCA技術検討会 OpenStack分科会
<http://www.osca-jp.com/>

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。
デル株式会社、レッドハット株式会社、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。
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。

本講演は、情報提供のみを目的としており、誤字脱字、技術上の誤りには一切責任を負いません。
本講演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。
本講演の内容は検証時のものであり、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。



Open
Standard
Cloud
Association