



## OpenStack on Ceph 性能評価

version 1.1

Jul. 2016 (Updated in Oct. 2016)

Writer

Takehiro Kudou (Hitachi Solutions, Ltd.)

Kazuho Hirahara (Hitachi Solutions, Ltd.)



## 目 次

1.	はじめに	4
1.1	概要	4
1.2	目的	4
2.	調査環境	4
3.	調査方法	6
3.1	fio パラメータ	6
4.	結果	7
4.1	Block Size 別の傾向	8
4.2	OSD サーバ台数別による傾向	18
5.	考察	28
5.1	Random Write における性能向上傾向	28
5.2	Random Read における性能向上傾向	29
6.	まとめ	31
6.1	懸念事項	31
7.	参考・関連文献	32

### 略号の説明

略号	意味
BS	<u>B</u> lock <u>S</u> ize
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
RHEL	<u>R</u> ed <u>H</u> at <u>E</u> nterprise <u>L</u> inux
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

OSCA 技術検討会 OpenStack 分科会において、Red Hat 社の商用 OpenStack ディストリビューションである Red Hat Enterprise Linux OpenStack Platform 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 を組み合わせた際の、挙動・ストレージ IOPS・スループット性能の評価を実施した。

### 1.2 目的

本評価は、OpenStack と Ceph を組み合わせた際の H/W 設計指針を得ることを目的とし、OpenStack のインスタンス(以下 VM)配置先として Ceph によるストレージ領域を選択し、ストレージ IOPS・スループット性能を調査した。

本書は主に性能評価結果について示している。挙動調査・H/W 設計指針については、別冊「[1]OpenStack on Ceph におけるストレージ設計のポイント」を参照いただきたい。

## 2. 調査環境

本評価では、表 2-1 に示すサーバ(1 シャーシあたり 4 サーバ)を 13 台用意し、検証を行った。

表 2-1 使用機材

機材	仕様
Dell PowerEdge C6220 	CPU : Intel Xeon E5-2620×2
	NIC : LOM 1Gbps×4、 Intel X520 dual port 10Gbps NIC×1
	OS : RHEL 7.2
	App : RHEL-OSP7、Red Hat Ceph Storage 1.3

各サーバの役割の詳細スペックを  
表 2-2 に示す。

表 2-2 サーバ役割

#	役割	台数	MEM	HDD
1	Controller Server - OpenStack Controller - Ceph Monitor	3	64G	2TB SATA 7.2Krpm ×2(OS、App)
2	Compute Server - OpenStack Compute	3	32G	300GB SAS 10Krpm ×2(OS、App)
3	OSD Server - Ceph OSD	6	32G	300GB SAS 10Krpm ×2 (OS、App) 600GB SAS 10Krpm ×3 (OSD) 320GB SSD (Journal) [各 OSD に 50GB の Journal Volume]
4	Director Server - OpenStack、Ceph の Deploy	1	64G	2TB SATA 7.2Krpm ×2(OS、App)

上記の各サーバについて、図 2-1 のとおりに構成した。環境構築には、RHEL-OSP director を用いて構築し、OpenStack Compute Node 上の VM が Ceph ストレージをブロックデバイスとしてマウントするよう設定した。

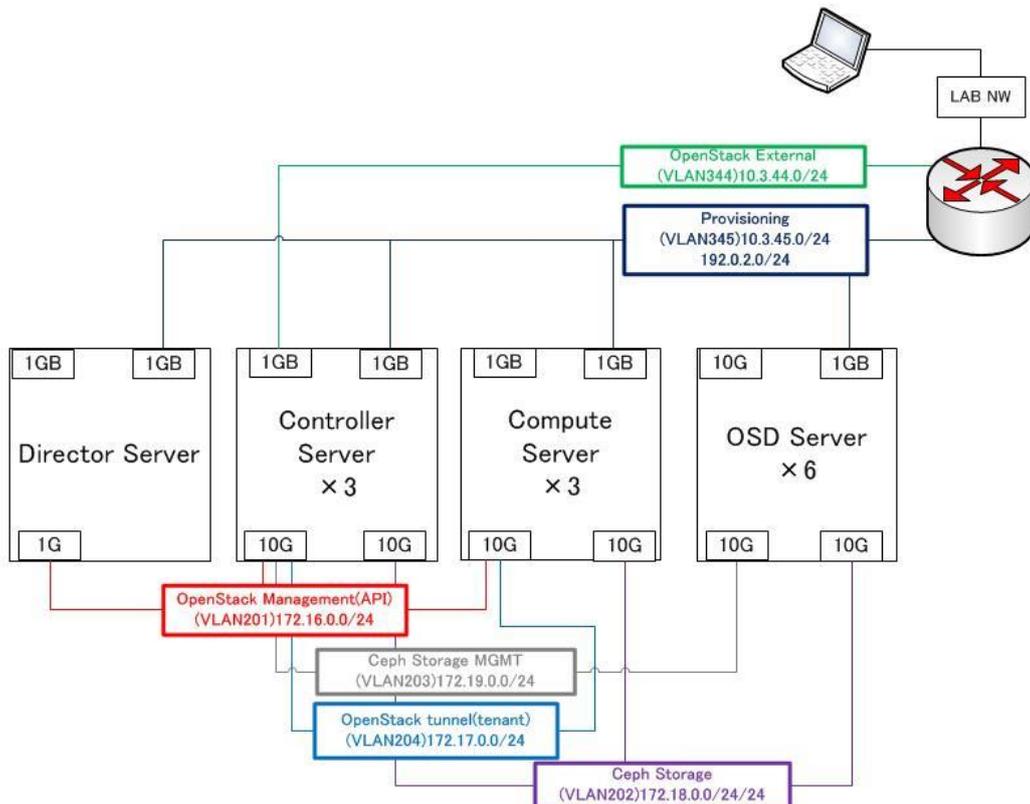


図 2-1 検証環境

なお、各 OSD(Disk)に対して、下記パラメータを投入した。

```
# ceph tell osd.[OSD No] injectargs --journal_max_write_entries 1000 --journal_max_write_bytes 1048576000
--journal_queue_max_ops 3000 --journal_queue_max_bytes 1048576000 --filestore_max_sync_interval 10
```

### 3. 調査方法

ストレージ IOPS・スループット性能は、下記手順にて調査した。

- Compute Server にそれぞれ 9 台の VM を作成(全体で 27 台)
- 全 VM に fio-2.1.7 をインストール
- 全 VM に Floating IP を割り当て
- Director Server からパスワードレスで ssh 接続できるよう設定
- Director Server から ssh にて各 VM 上で fio を実行

#### 3.1 fio パラメータ

fio は、OSD Server の台数、VM 数、Read/Write、Block Size、job 数の値を変化させ、2 分間ベンチ実行、2 分間待機を繰り返しながら各 3 回計測した。なお、複数 VM の場合、”&”でつないで並列実行した。

fio 実行時の変数は以下の通りである。

```
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) & . . . .
```

- OSD Server : 3|4|5|6
- VM : 1|2(2 Compute Server に各 1VM)|3(3 Compute Server に各 1VM)|  
9(3 Compute Server に各 3VM)|27(3 Compute Server に各 9VM)
- [instance] : ユーザ名@VM の IP Address。
- [read/write] : randomread|randomwrite
- [BlockSize]:4|16|32|64|128
- [Job 数]:1|4|8|16
- [file name] : ファイル名

これにより、計 20,160 のベンチマーク結果を得た。

## 4. 結果

複数 VM での実施結果は、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。また、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】**

得られた全データにおける、最大値は表 4-1 の通りである。

表 4-1 最大性能値

#	項目	条件	値
1	Write 時 最大 IOPS	OSD サーバ:6 台、Block Size:16KB、Job 数:4、VM:1	1,805 (io/s)
2	Write 時 最大 Throughput	OSD サーバ:6 台、Block Size:128KB、Job 数:1、VM:3	161,525(KB/s)
3	Read 時 最大 IOPS	OSD サーバ:3 台、Block Size:4KB、Job 数:16、VM:27	494,491(io/s)
4	Read 時 最大 Throughput	OSD サーバ:3 台、Block Size:32KB、Job 数:16、VM:27	28,111,873 (KB/s)

### 4.1 Block Size 別の傾向

VM1、2、3、9、27 台時の、Block Size(以下 BS)別のデータを以下に示す。

(1) Random Write 時の性能

(a) Write、VM1 台時

BS 増加に伴い IOPS は緩やかに低下するが、Throughput は大幅に向上する。Job 数観点では、job 数 4 をピークとし、job 数が増加すると性能が徐々に低下する。

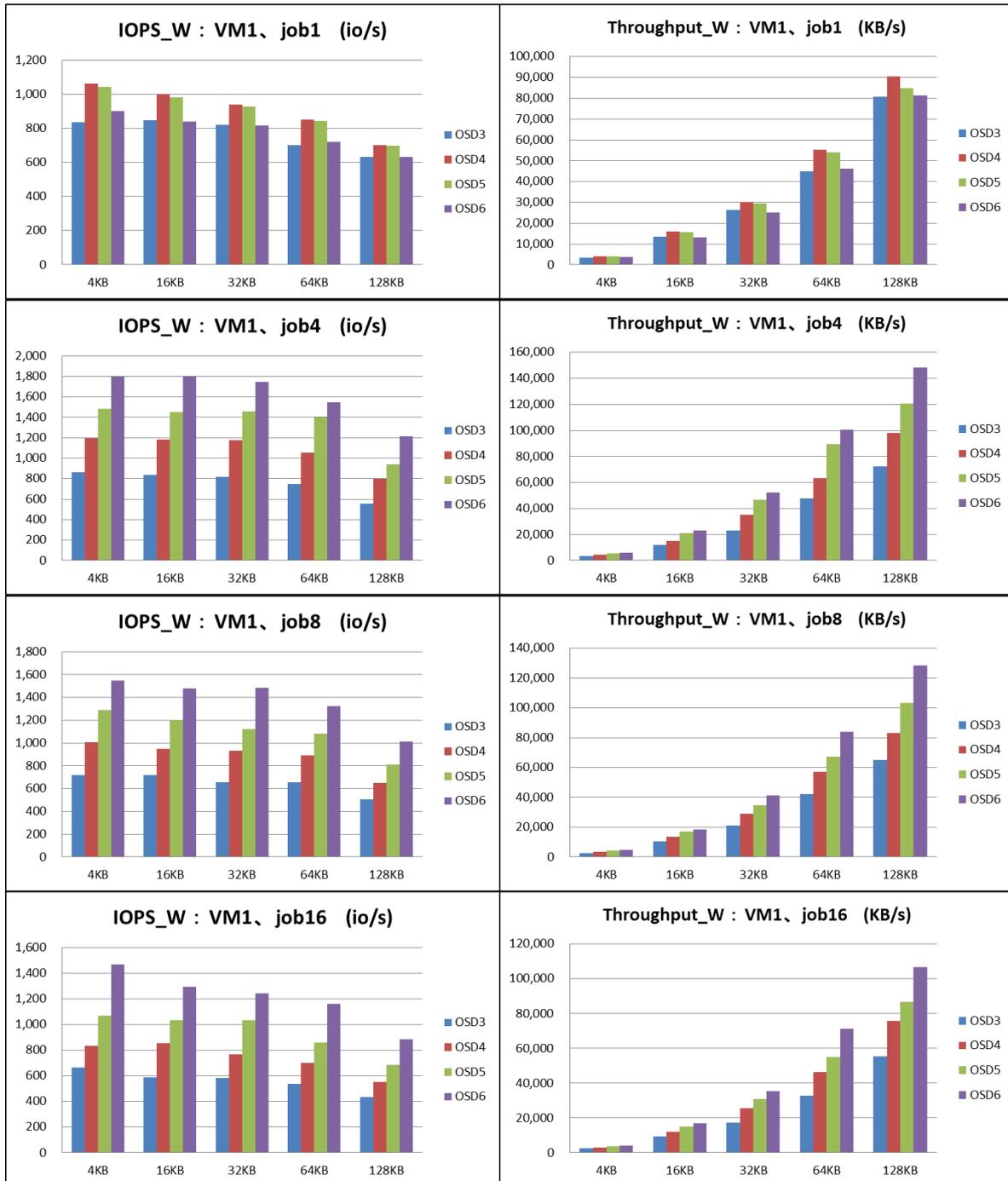


図 4-1 Write、VM1 台時の傾向

(b) Write、VM2 台時

BS 増加に伴い IOPS は緩やかに低下するが、Throughput は大幅に向上する。Job 数観点では、job 数 1 をピークとし、job 数が増加すると性能が徐々に低下する。

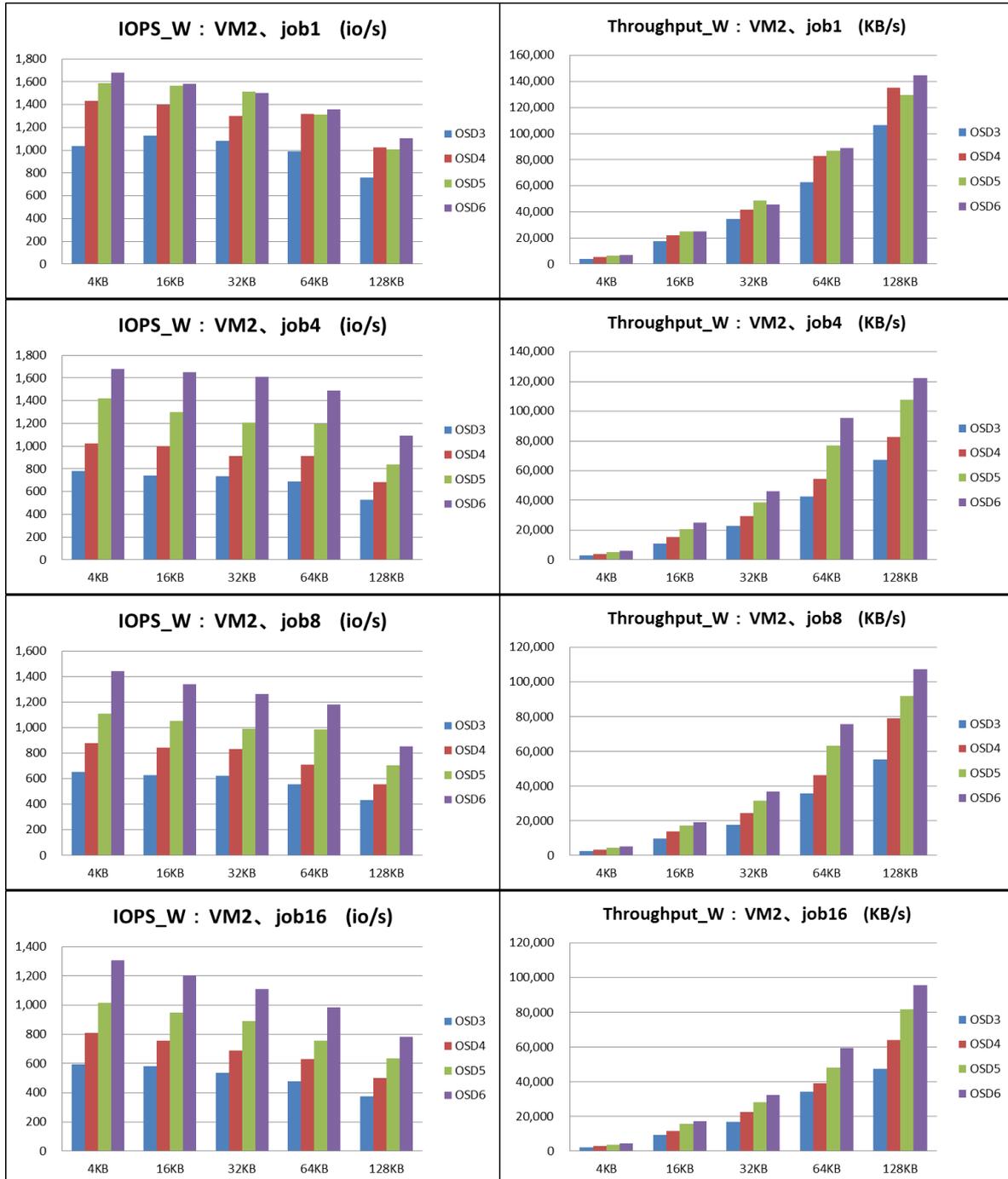


図 4-2 Write、VM2 台時の傾向

(c) Write、VM3 台時

BS 増加に伴い IOPS は緩やかに低下するが、Throughput は大幅に向上する。Job 数観点では、job 数 1 をピークとし、job 数が増加すると性能が徐々に低下する。

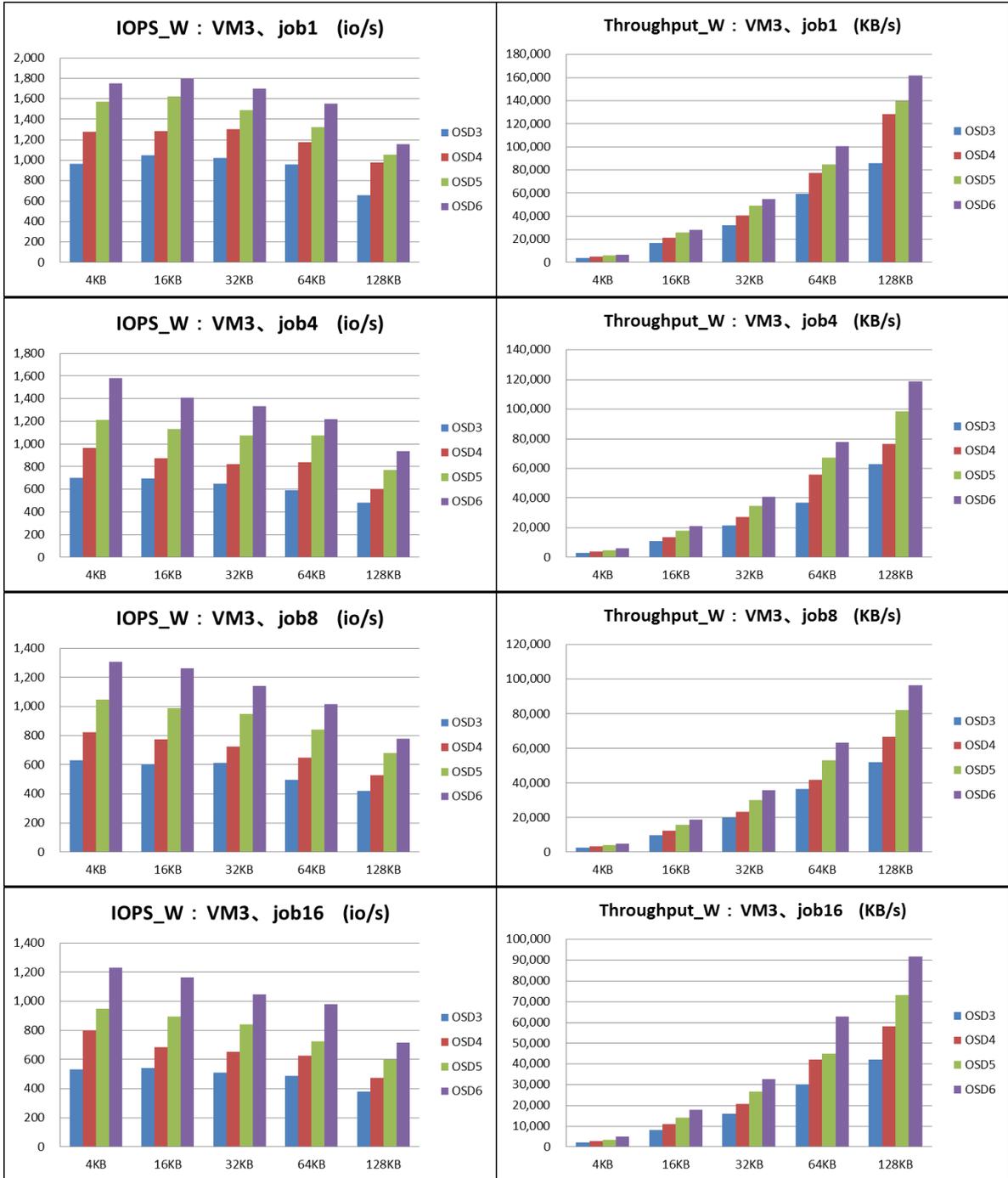


図 4-3 Write、VM3 台時の傾向

(d) Write、VM9 台時

BS 増加に伴い IOPS は緩やかに低下するが、Throughput は大幅に向上する。Job 数観点では、job 数 1 をピークとし、job 数が増加すると性能が徐々に低下する。

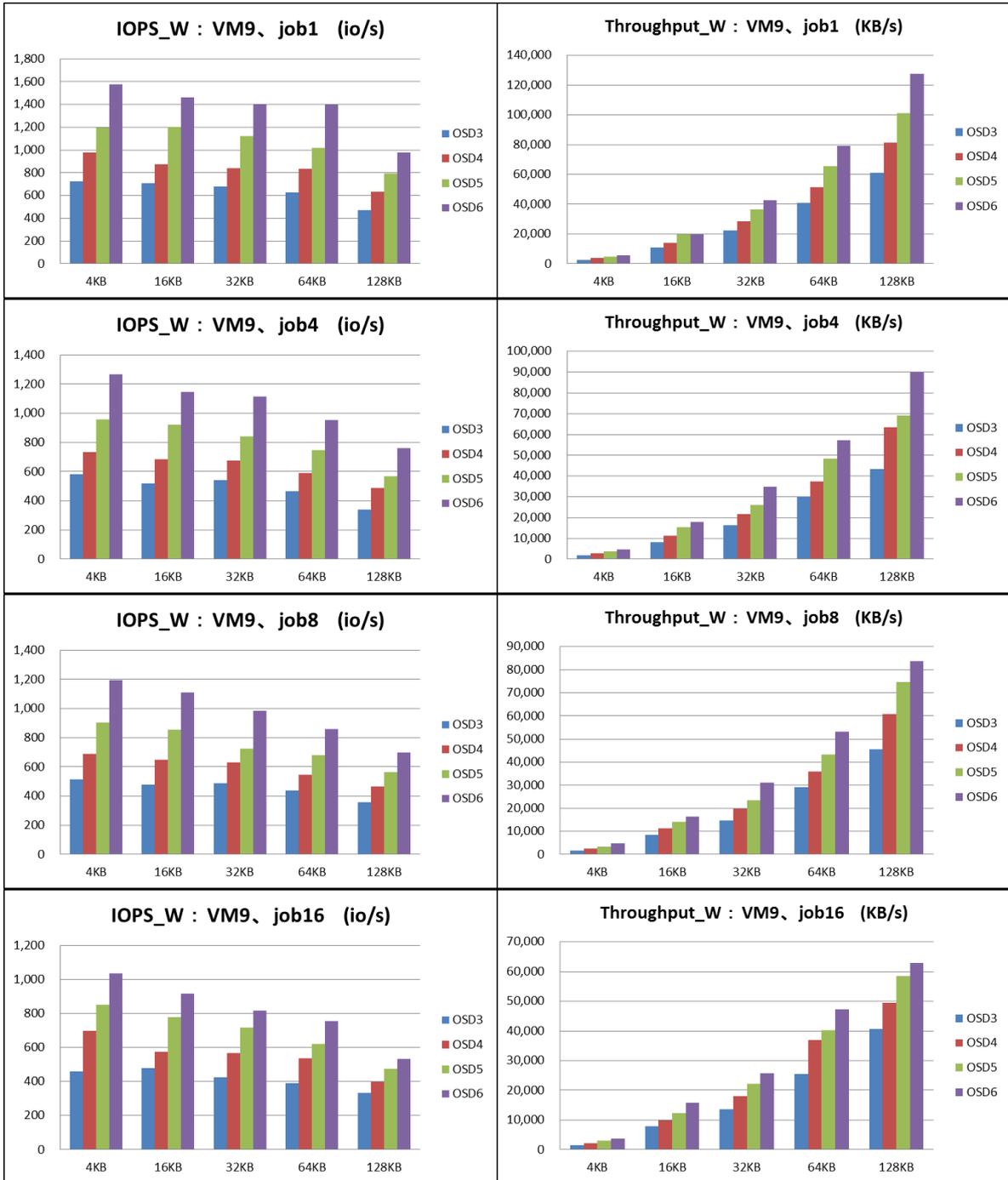


図 4-4 Write、VM9 台時の傾向

(e) Write、VM27 台時

BS 増加に伴い IOPS は緩やかに低下し、Throughput は大幅に向上する。Job 数観点では、job 数 1 をピークとし、job 数が増加すると性能が徐々に低下する。

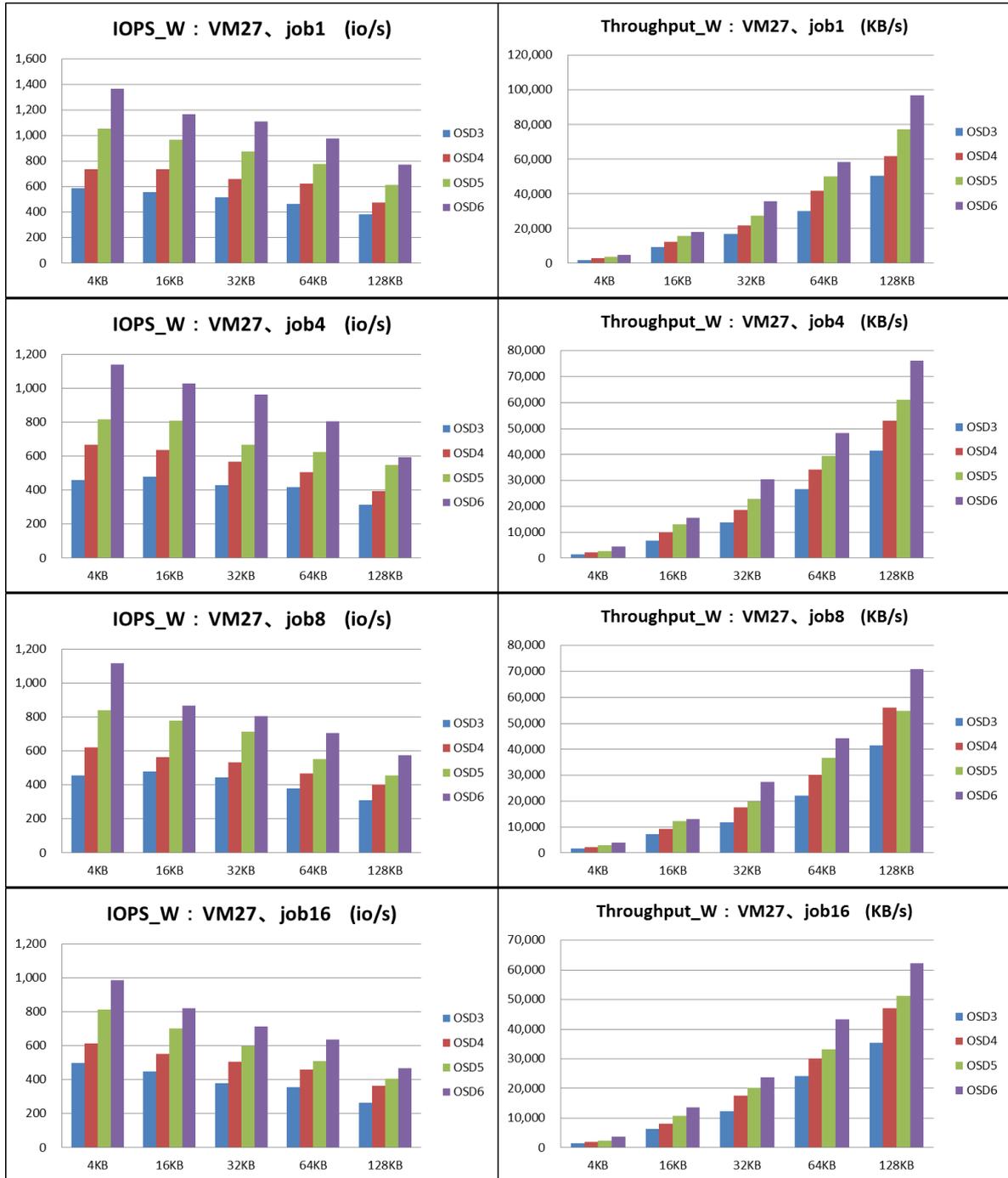


図 4-5 Write、VM27 台時の傾向

(2) Random Read 時の性能

(a) Read、VM1 台時

BS 増加に伴い IOPS は低下するが、Throughput は大幅に向上する。Job 数観点では、job 数が増加すると性能が向上する。

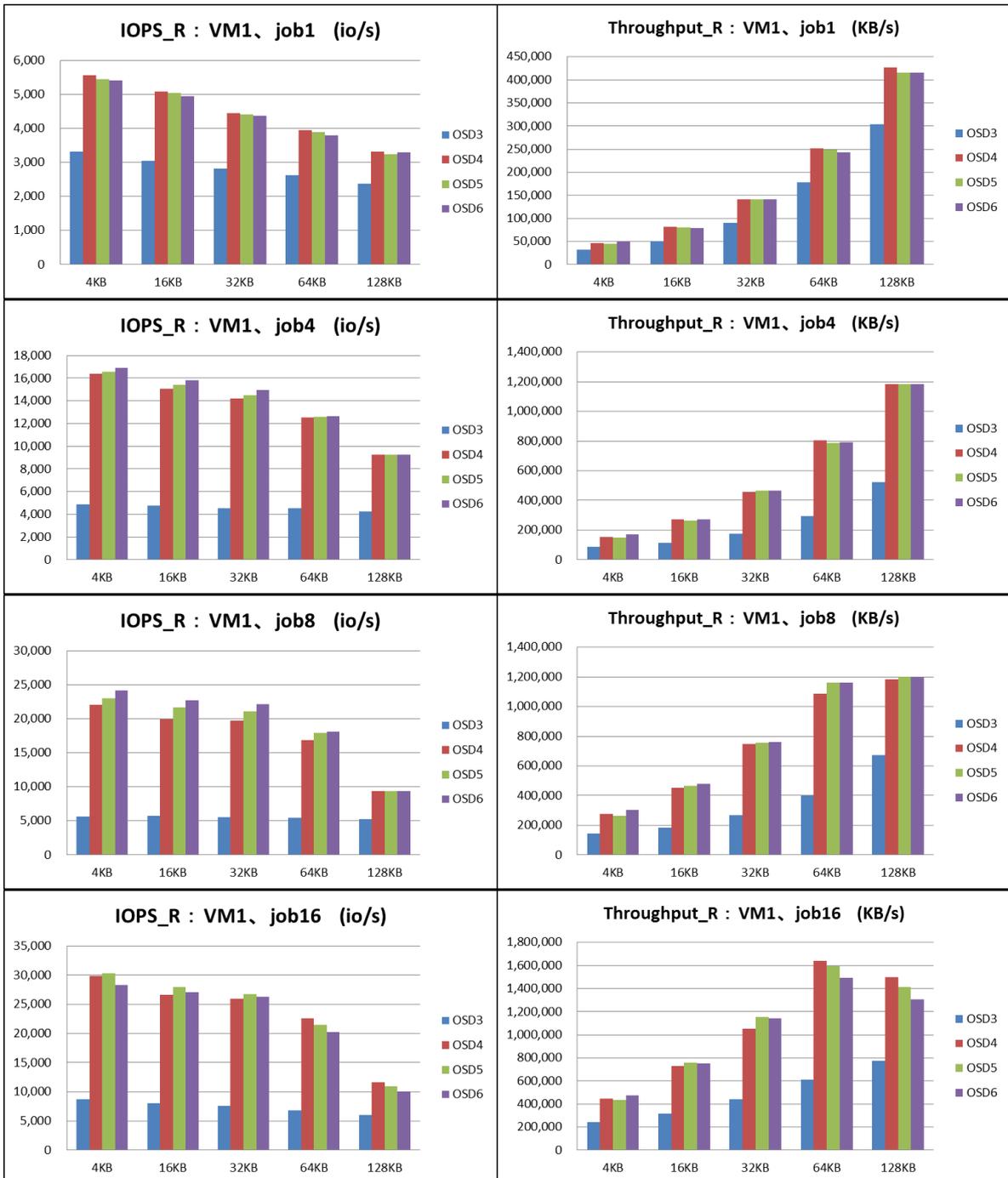


図 4-6 Read、VM1 台時の傾向

(b) Read、VM2 台時

BS 増加に伴い IOPS は低下するが、Throughput は大幅に向上する。Job 数観点では、job 数が増加すると性能が向上する。

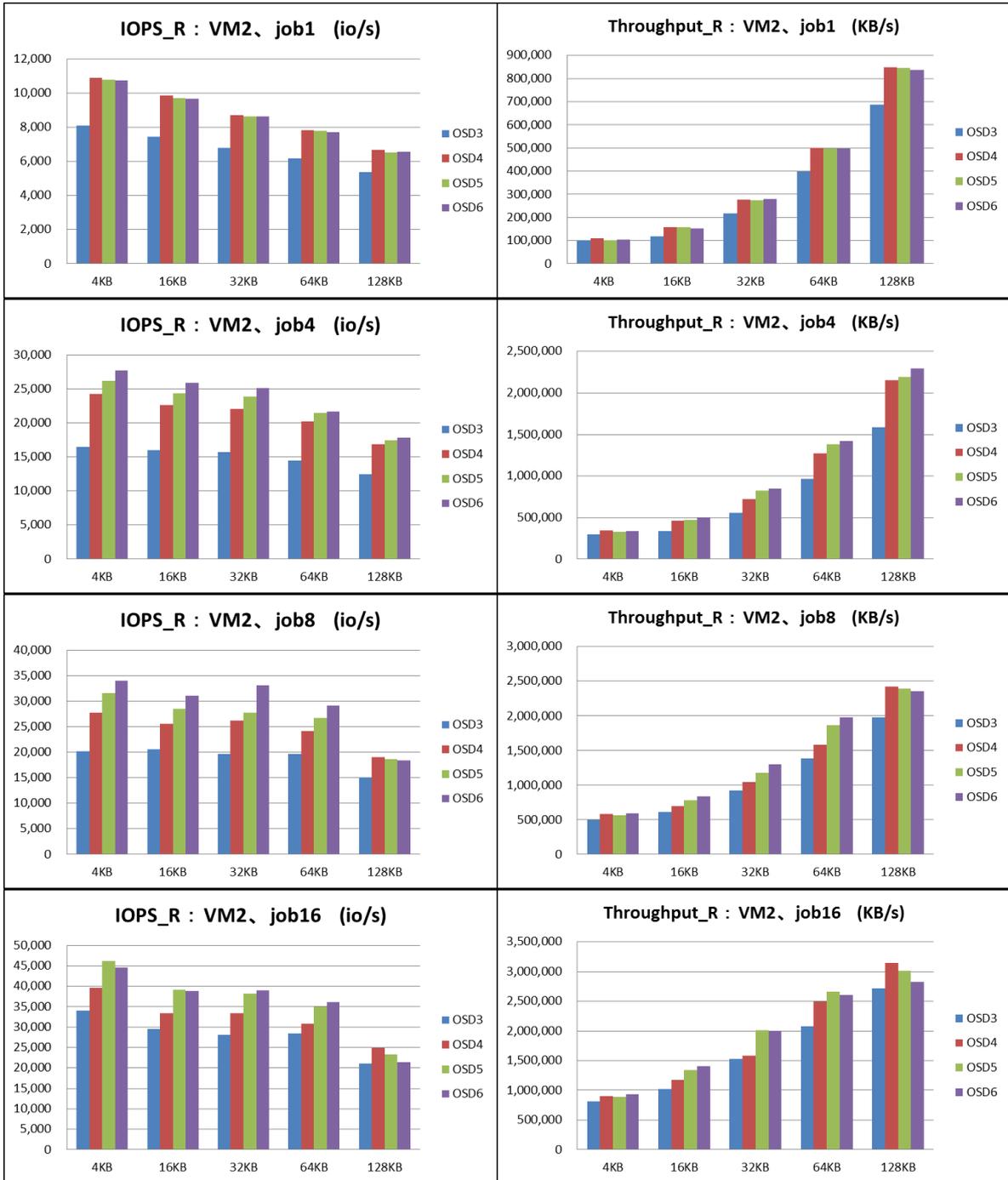


図 4-7 Read、VM2 台時の傾向

(c) Read、VM3 台時

BS 増加に伴い IOPS は低下するが、Throughput は大幅に向上する。Job 数観点では、job 数が増加すると性能が向上する。

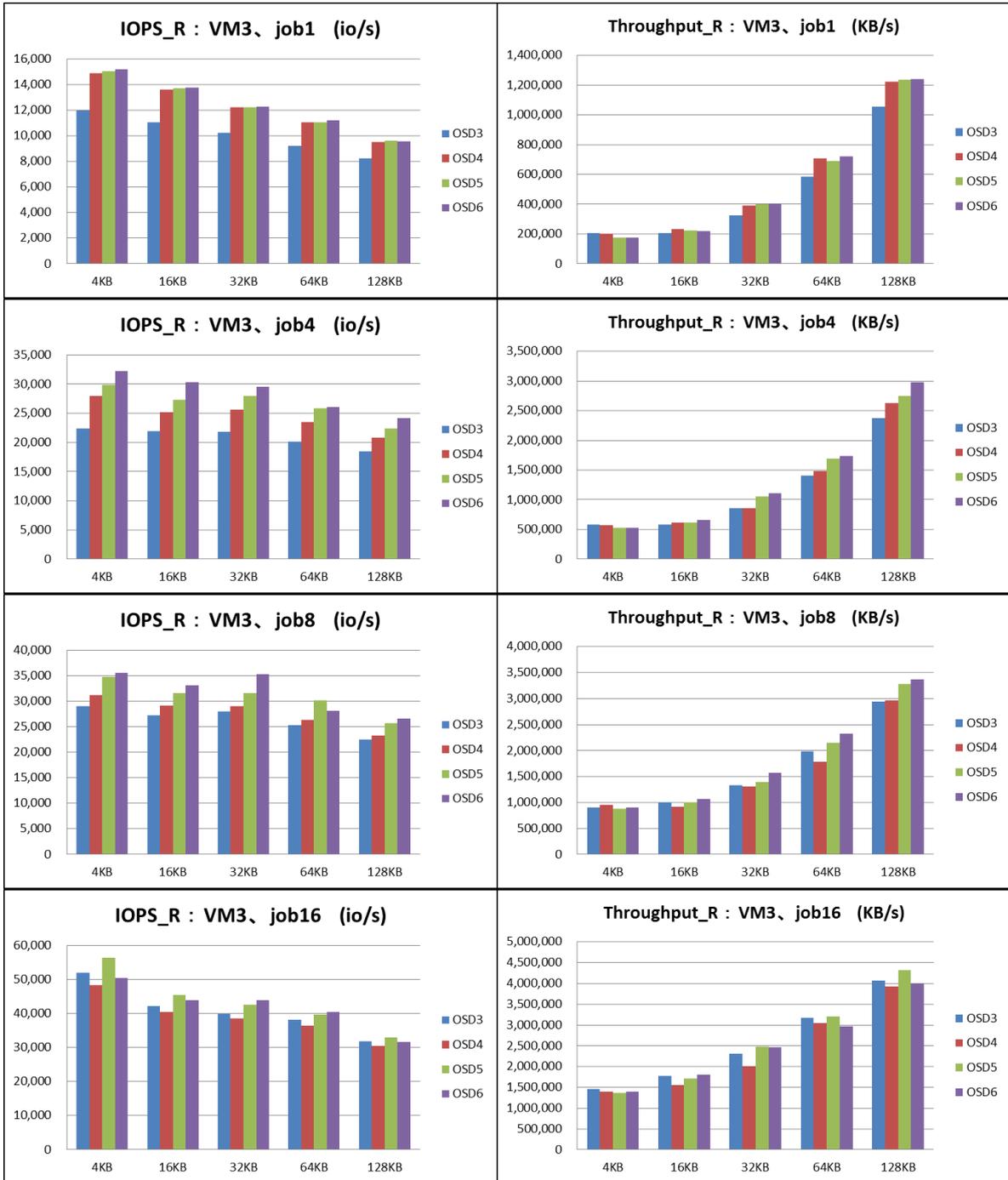


図 4-8 Read、VM3 台時の傾向

(d) Read、VM9 台時

BS 増加に伴い IOPS は低下し、Throughput の向上具合は 4.1(2)-(a)(b)(c)と比較し緩やかである。Job 数観点では、job 数が増加すると性能が向上する。

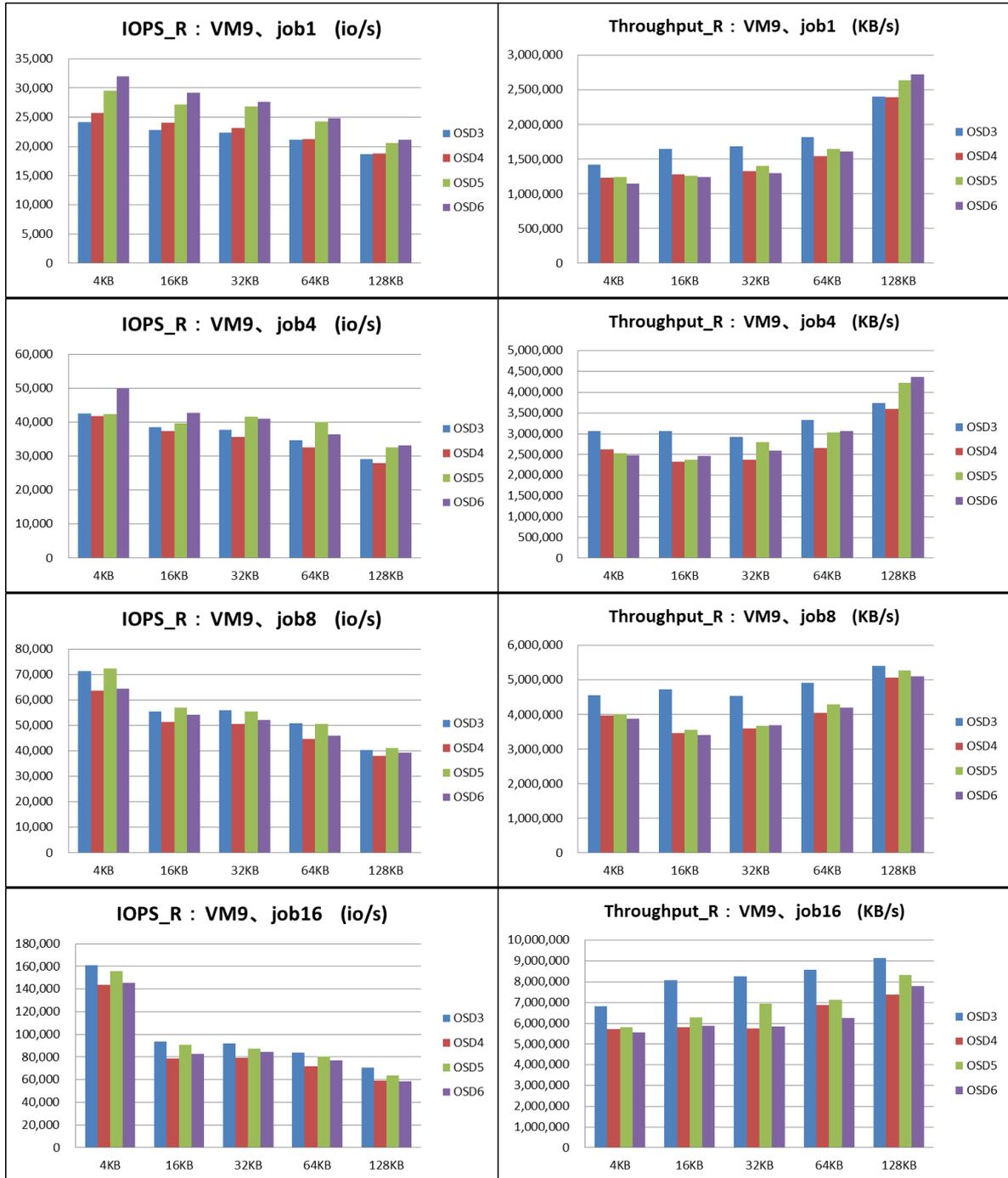


図 4-9 Read、VM9 台時の傾向

(e) Read、VM27 台時

BS 増加に伴い IOPS は低下するが、Throughput はあまり変化がみられない。Job 数観点では、job 数が増加すると性能が向上する。

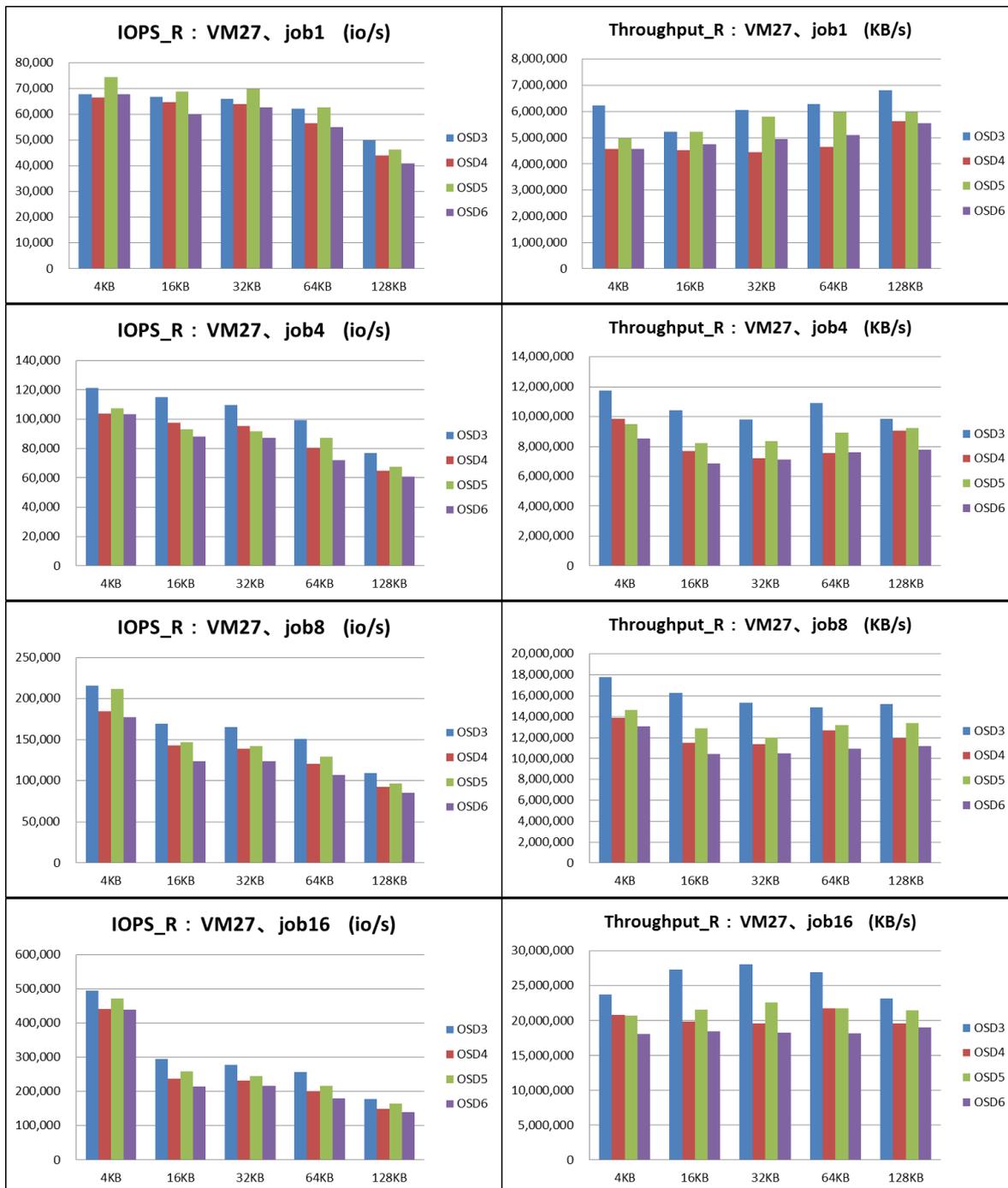


図 4-10 Read、VM27 台時の傾向

## 4.2 OSD サーバ台数別による傾向

4.1 とは観点を変更し、BS 4、16、32、64、128KB 時の、OSD サーバ台数別のデータを以下に示す。

### (1) Random Write 時性能

#### (a) Write、BS 4KB 時

OSD サーバ数に比例し性能が向上する。Job 数、VM 数観点では、job1 の場合を除き、job 数、VM 数の増加に伴い、性能が低下する。

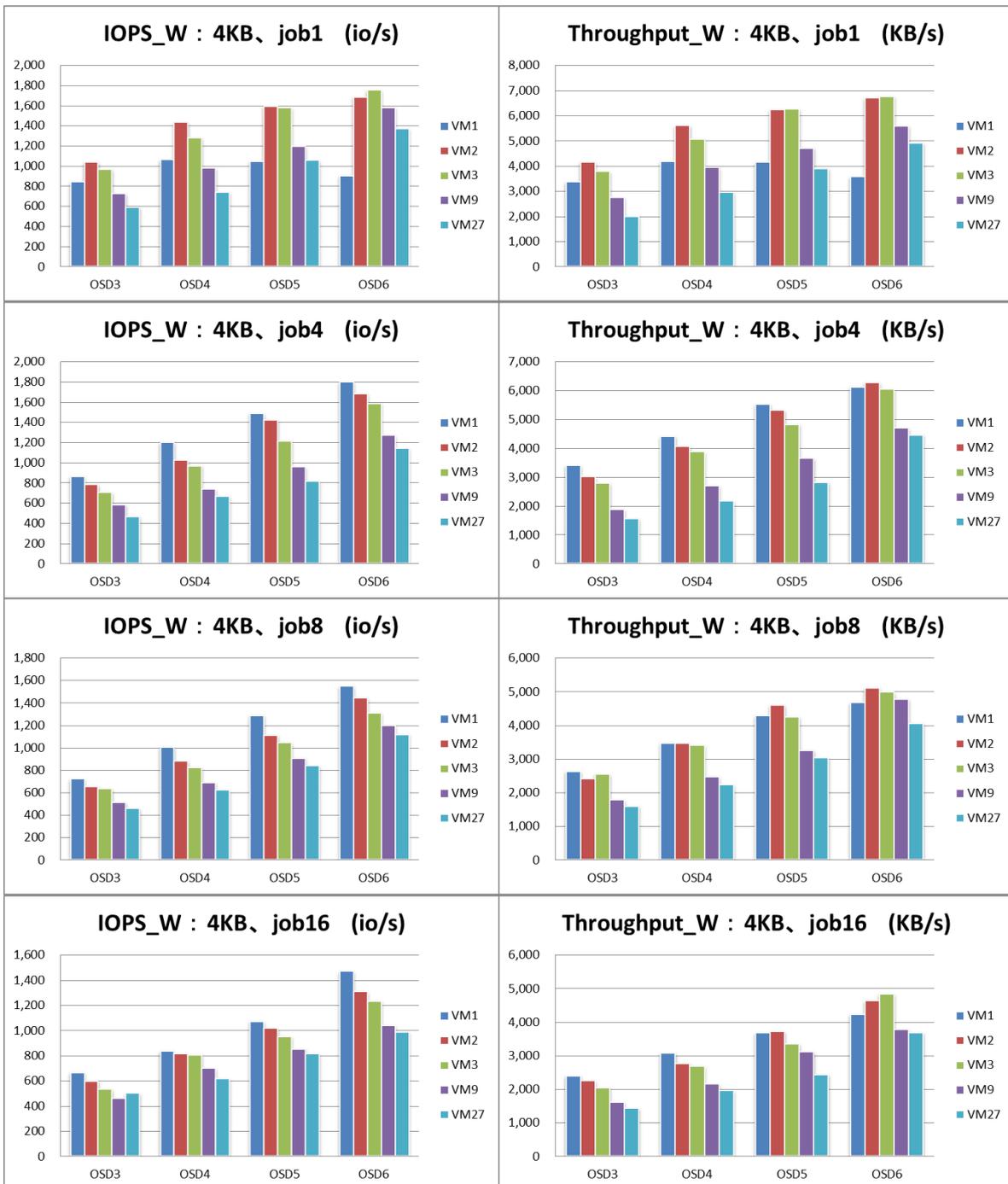


図 4-11 Write、BS 4KB 時の傾向

(b) Write、BS 16KB 時

OSD サーバ数に比例し性能が向上する。Job 数、VM 数観点では、job1 の場合を除き、job 数、VM 数の増加に伴い、性能が低下する。

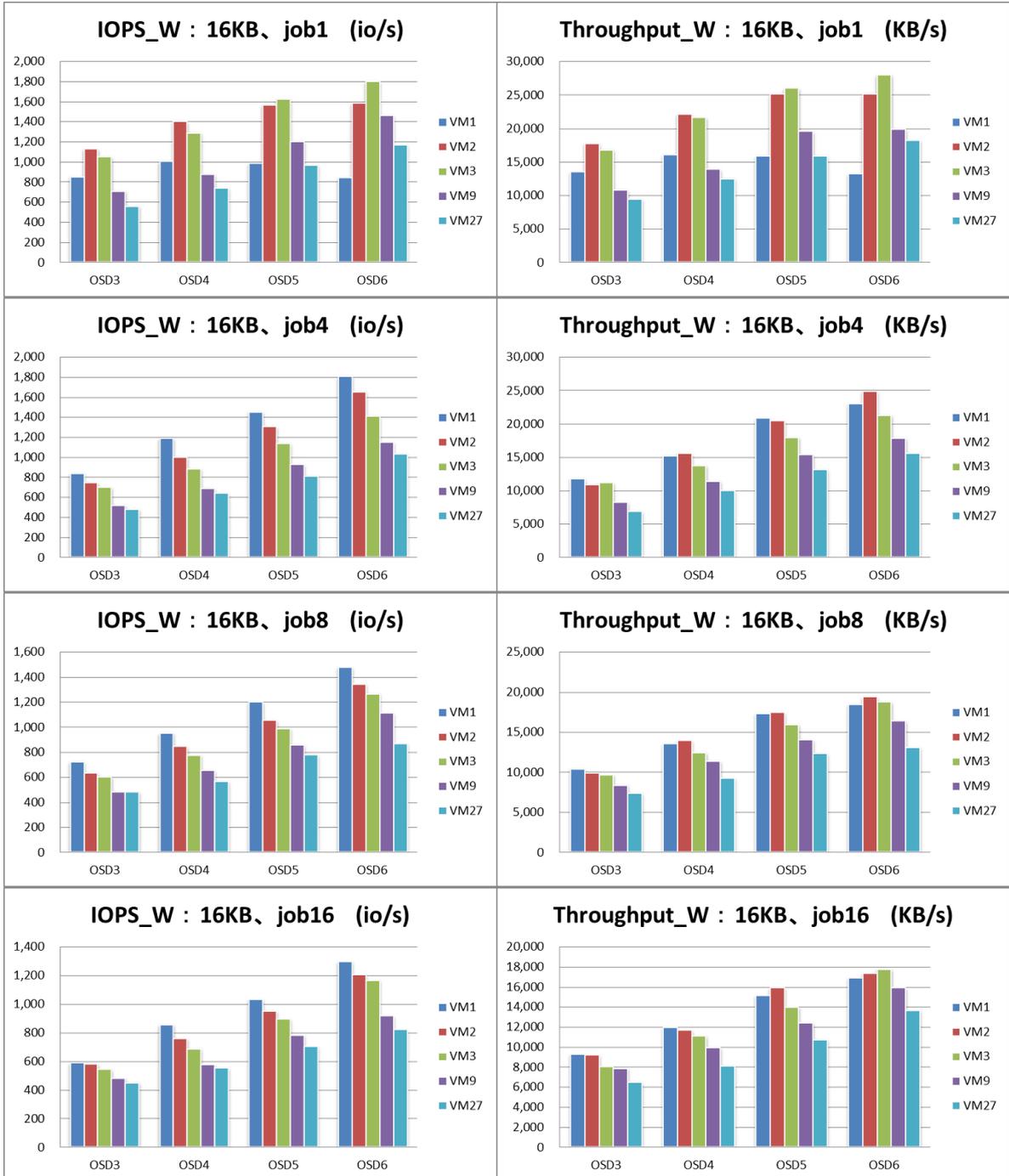


図 4-12 Write、BS 16KB 時の傾向

(c) Write、BS 32KB 時

OSD サーバ数に比例し性能が向上する。Job 数、VM 数観点では、job1 の場合を除き、job 数、VM 数の増加に伴い、性能が低下する。

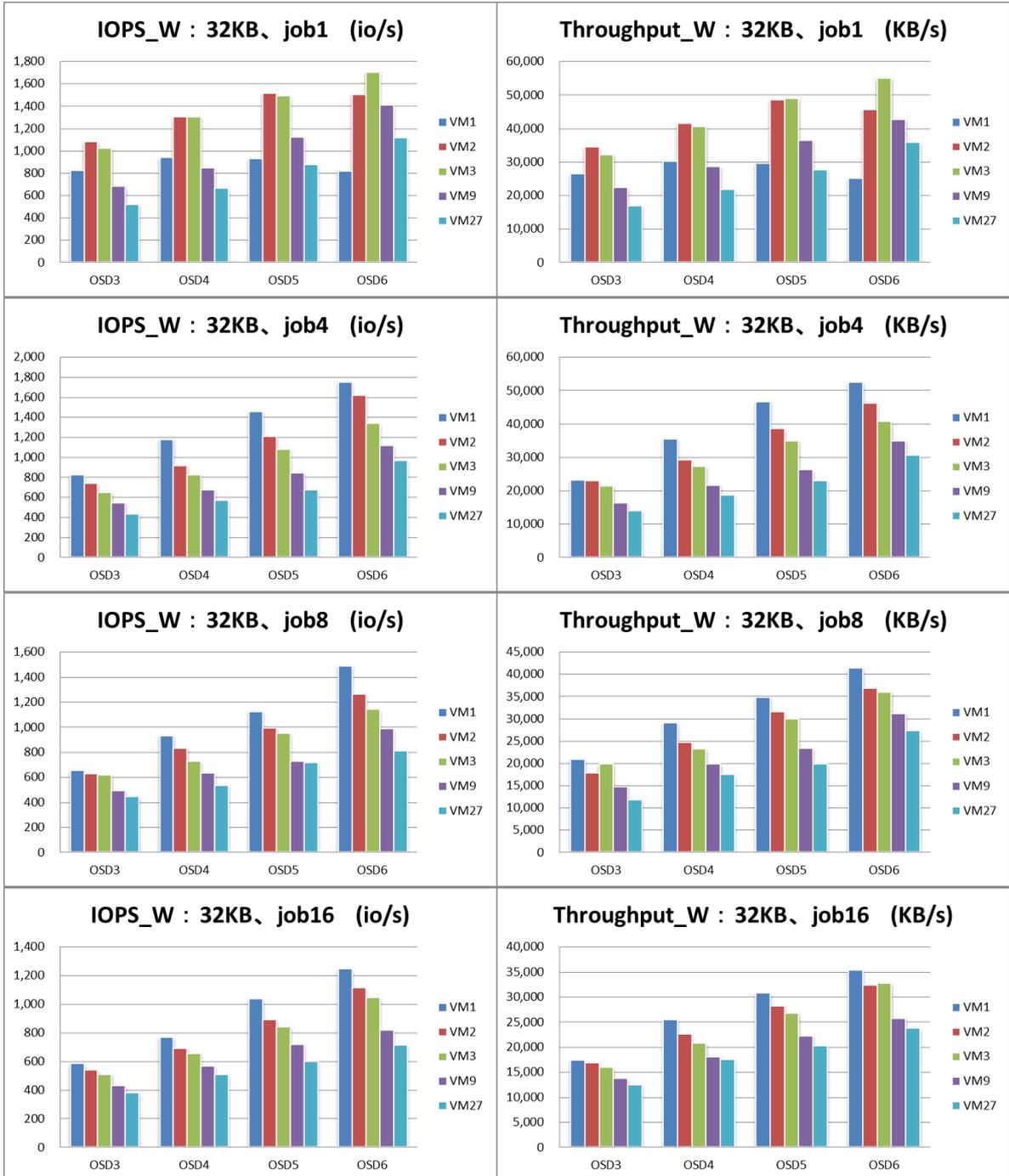


図 4-13 Write、BS 32KB 時の傾向

(d) Write、BS 64KB 時

OSD サーバ数に比例し性能が向上する。Job 数、VM 数観点では、job1 の場合を除き、job 数、VM 数の増加に伴い、性能が低下する。



図 4-14 Write、BS 64KB 時の傾向

(e) Write、BS 128KB 時

OSD サーバ数に比例し性能が向上する。job 数、VM 数観点では、job1 の場合を除き、job 数、VM 数の増加に伴い、性能が低下する。

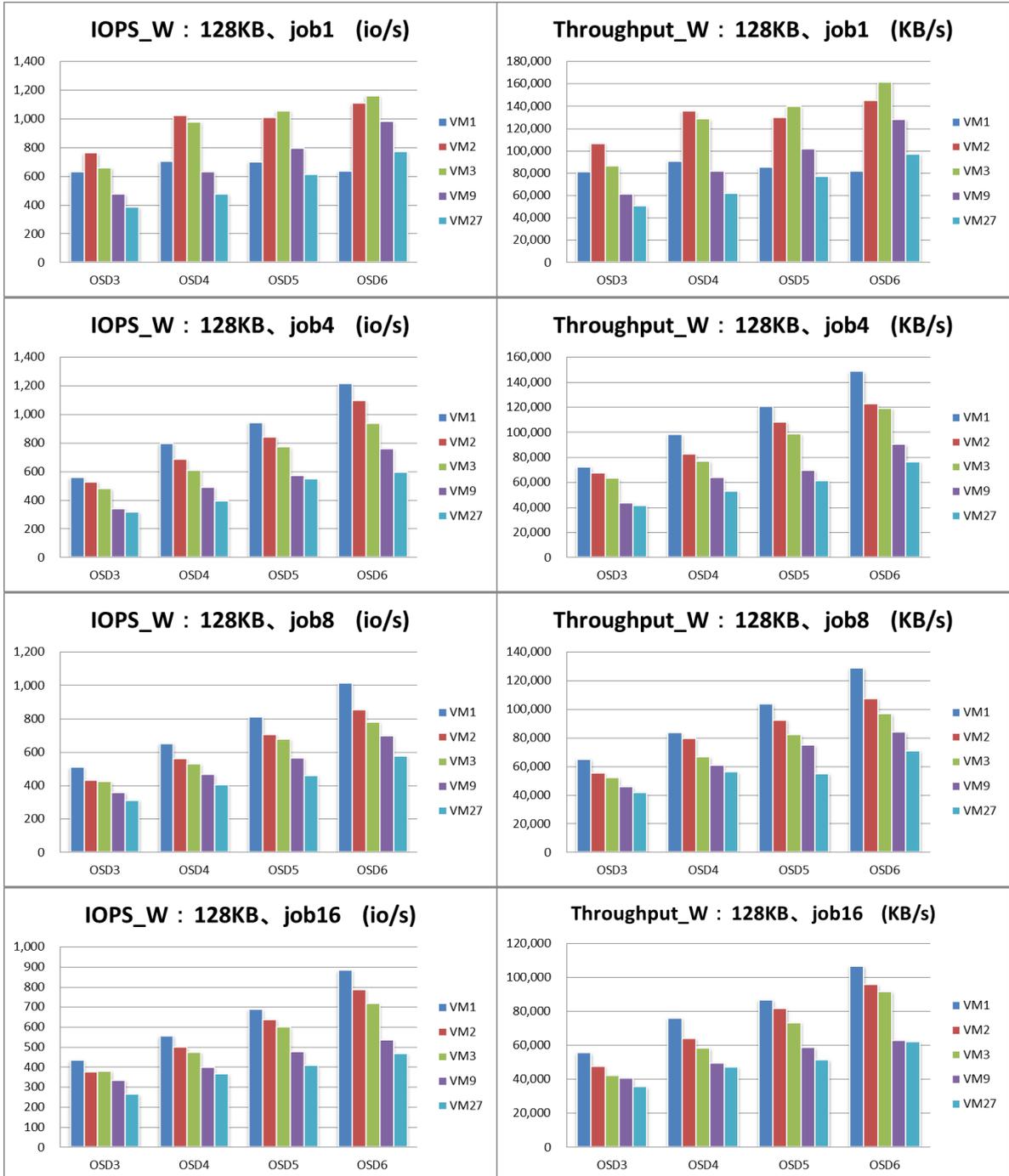


図 4-15 Write、BS 128KB 時の傾向

(2) Random Read 時性能

(a) Read、BS 4KB 時

OSD サーバ数と性能はあまり比例しない。job 数、VM 数観点では、job 数、VM 数の増加に伴い、性能が増加する。

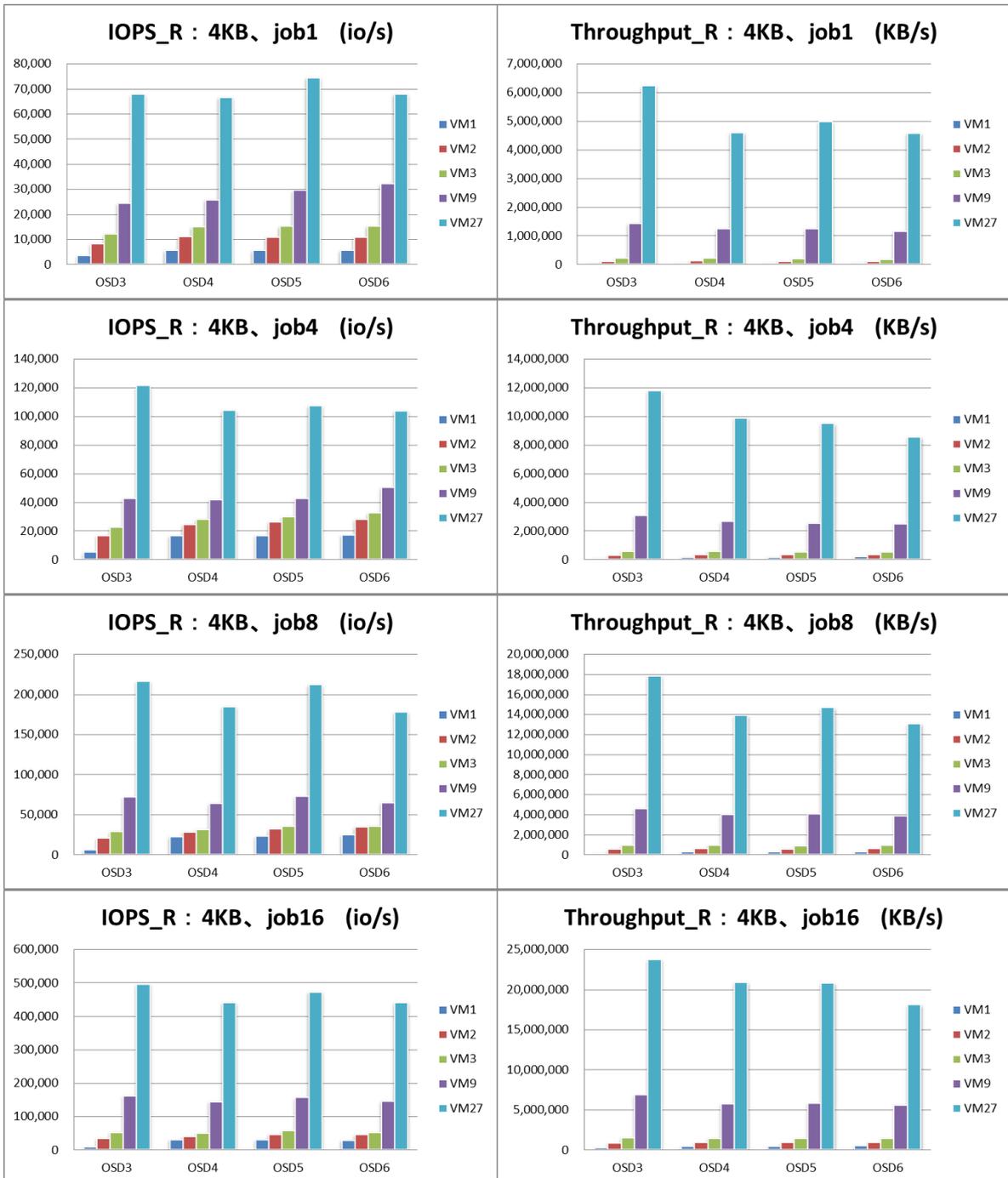


図 4-16 Read、BS 4KB 時の傾向

(b) Read、BS 16KB 時

OSD サーバ数と性能はあまり比例しない。job 数、VM 数観点では、job 数、VM 数の増加に伴い、性能が増加する。

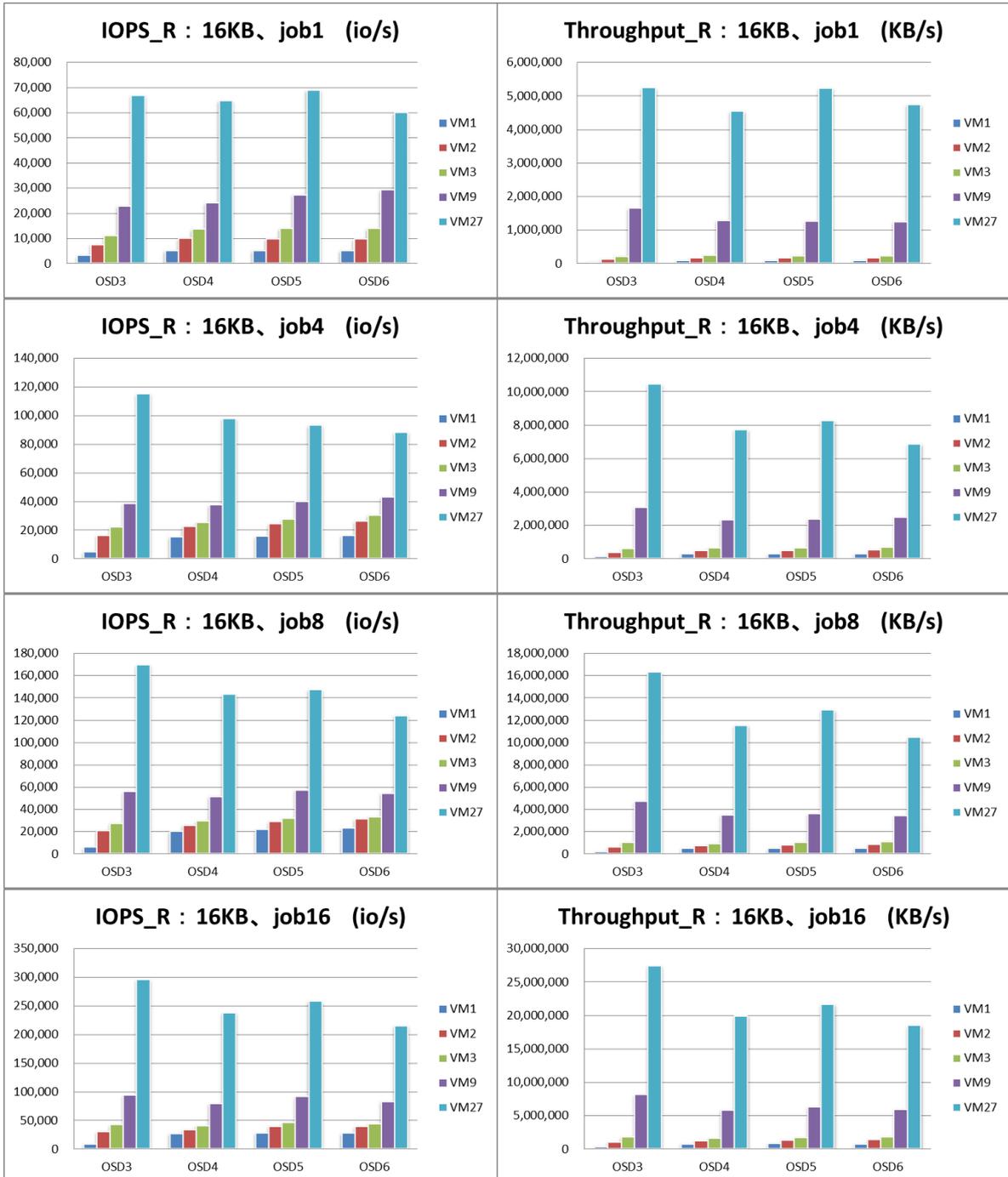


図 4-17 Read、BS 16KB 時の傾向

(c) Read、BS 32KB 時

OSD サーバ数と性能はあまり比例しない。job 数、VM 数観点では、job 数、VM 数の増加に伴い、性能が増加する。

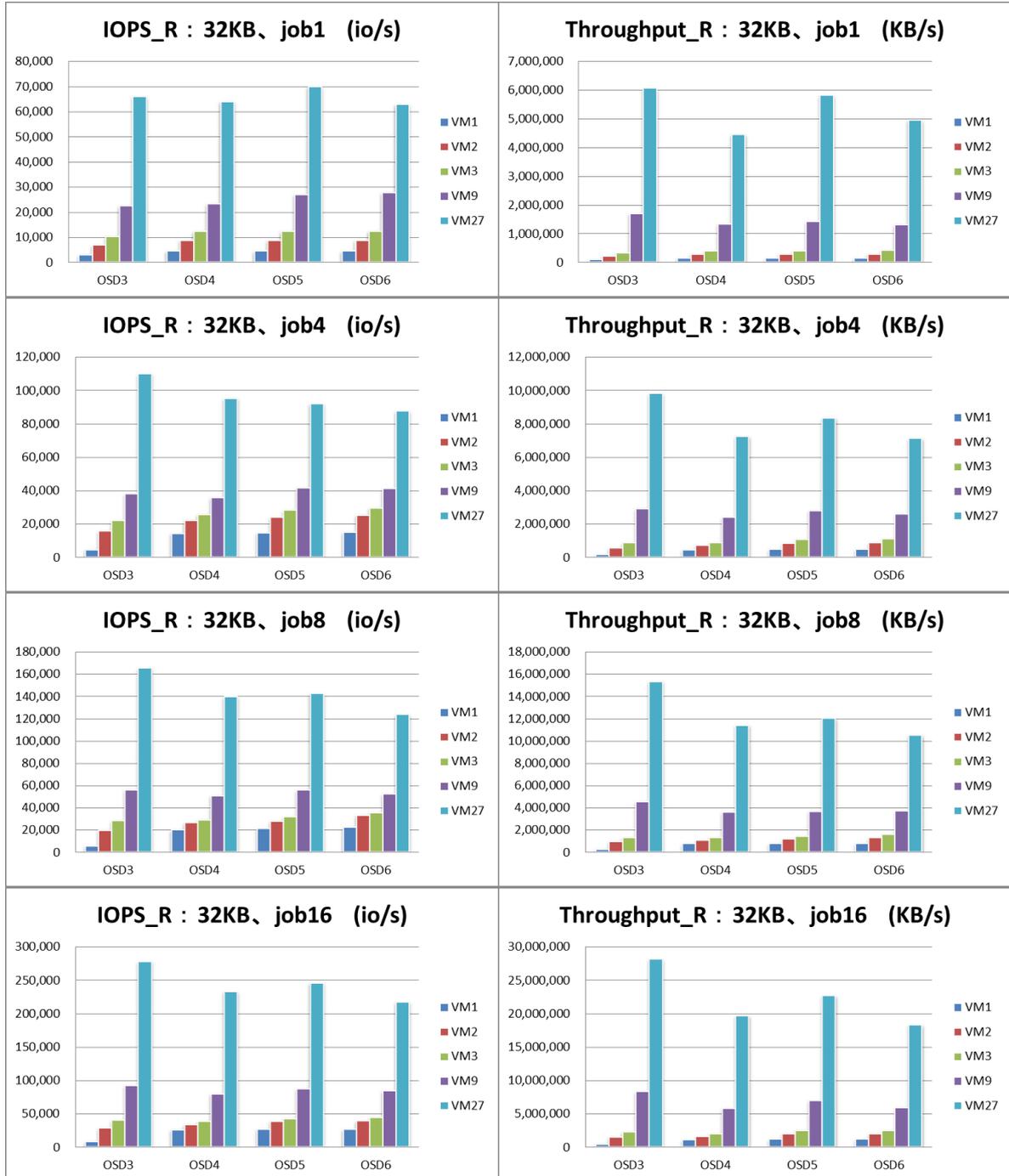


図 4-18 Read、BS 32KB 時の傾向

(d) Read、BS 64KB 時

OSD サーバ数と性能はあまり比例しない。job 数、VM 数観点では、job 数、VM 数の増加に伴い、性能が増加する。

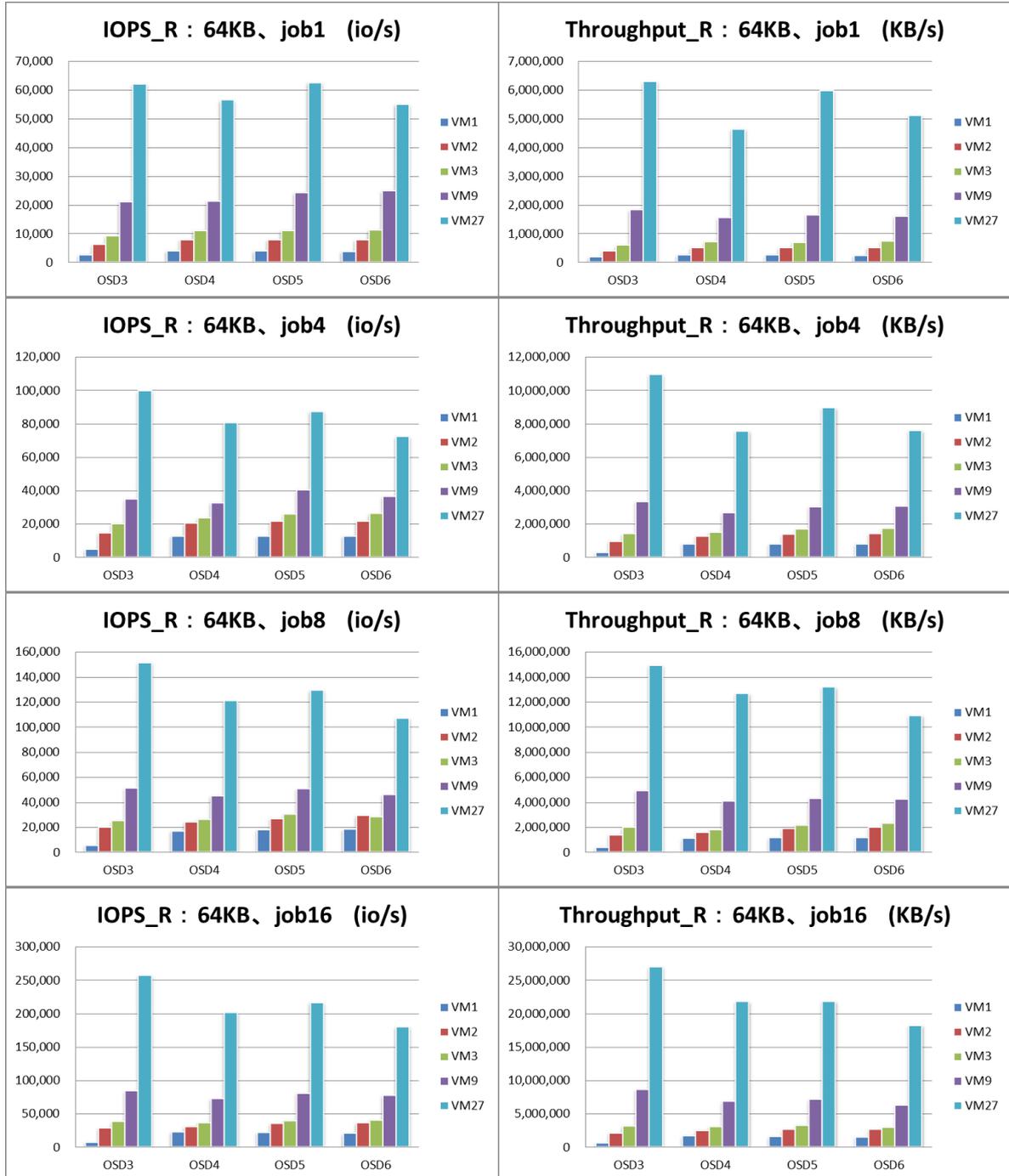


図 4-19 Read、BS 64KB 時の傾向

(e) Read、BS 128KB 時

OSD サーバ数と性能はあまり比例しない。job 数、VM 数観点では、job 数、VM 数の増加に伴い、性能が増加する。

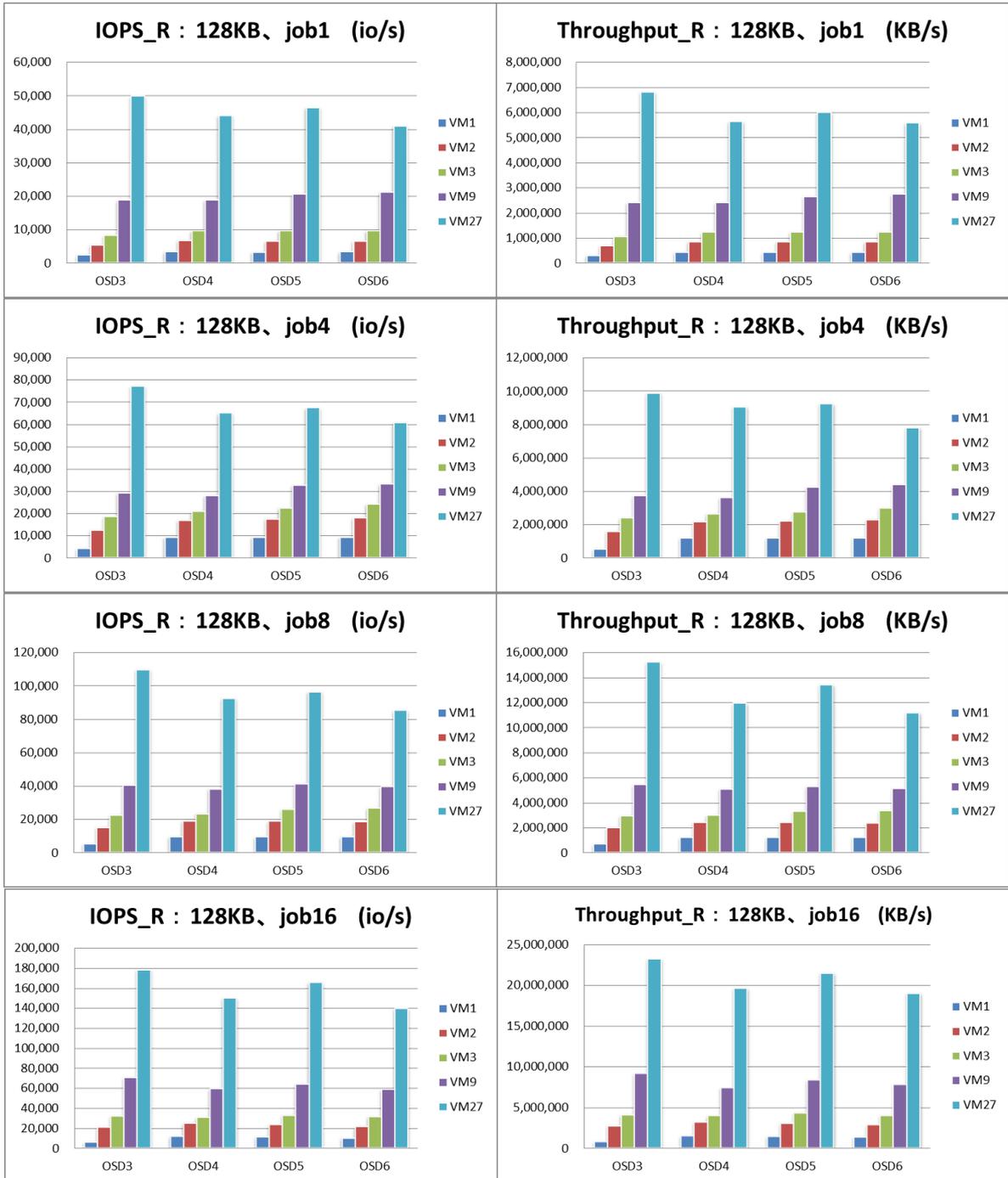


図 4-20 Read、BS 128KB 時の傾向

## 5. 考察

BS 別では、全体として表 5-1 の傾向が確認できた。Random Write に関しては一般的なストレージと同等の傾向が確認できたが、Random Read については限界性能に達していないものと考えられる。

表 5-1 BS 別の性能全体傾向

#	パターン	Random Write		Random Read	
		IOPS	Throughput	IOPS	Throughput
1	BS 増加	緩やかに低下	大幅に向上	緩やかに低下	大幅に向上
2	job 数増加	緩やかに低下		大幅に向上	

OSD サーバ台数別では、全体として表 5-2 の傾向が確認できた。Random Write については、OSD サーバ台数の増加に伴い性能が向上していることを確認できた。一方、Random Read については、OSD サーバ台数を増加しても変化が少ないことから、今回のベンチマークのかけ方では、OSD サーバ 3 台時の限界性能にも達しておらず、OSD サーバをスケールアウトすることによる性能向上効果が確認できなかったためと考える。

表 5-2 OSD サーバ別の性能全体傾向

#	パターン	Random Write		Random Read	
		IOPS	Throughput	IOPS	Throughput
1	OSD サーバ増加	比例的に向上		変化なし	
2	job 数、VM 数増加	低下		比例的に向上	

### 5.1 Random Write における性能向上傾向

Random Write においては、OSD サーバ増加による性能向上が確認できた。性能向上傾向を精査するため、IOPS の最大値をとったパターン(OSD サーバ 6、BS 16KB、job4、VM1)を代表として、VM1、job4 での IOPS 傾向を精査した。

表 5-3 VM1、job4 での IOPS 傾向

#	BS	OSD3	OSD4	OSD5	OSD6
1	4KB (OSD サーバ数で除算)	862 (287.3)	1,198 (299.5)	1,486 (297.2)	1,797 (299.5)
2	16KB (OSD サーバ数で除算)	835 (278.3)	1,185 (296.3)	1,449 (289.8)	1,805 (300.8)
3	32KB (OSD サーバ数で除算)	820 (273.3)	1,174 (293.5)	1,456 (291.2)	1,744 (290.7)
4	64KB (OSD サーバ数で除算)	748 (249.3)	1,053 (263.3)	1,400 (280.0)	1,549 (258.2)
5	128KB (OSD サーバ数で除算)	558 (186.0)	796 (199.0)	941 (188.2)	1,212 (202.0)

同様に、IOPS の最小値をとったパターン(OSD サーバ 3、BS 128KB、job16、VM27)から VM27、job16 での IOPS 傾向を精査した。

表 5-4 VM27、job16 での IOPS 傾向

#	BS	OSD3	OSD4	OSD5	OSD6
1	4KB (OSD サーバ数で除算)	499 (166.3)	614 (153.5)	813 (162.6)	985 (164.2)
2	16KB (OSD サーバ数で除算)	449 (149.7)	552 (138.0)	701 (140.2)	819 (136.5)
3	32KB (OSD サーバ数で除算)	378 (126.0)	506 (126.5)	599 (119.8)	713 (118.8)
4	64KB (OSD サーバ数で除算)	357 (119.0)	460 (115.0)	510 (102.0)	638 (106.3)
5	128KB (OSD サーバ数で除算)	263 (87.7)	365 (91.3)	406 (81.2)	467 (77.8)

いずれの場合も、OSD サーバ数で除算した値が近似値であることから、Random Write 性能は、OSD サーバ追加により比例的に性能があがることが確認できた。

また、各 OSD サーバは OSD Disk を 3 台搭載している。BS 4KB において、低負荷時は OSD Disk1 台あたり 100 IOPS、高負荷時は OSD Disk1 台あたり 55 IOPS 程度の性能であった。

## 5.2 Random Read における性能向上傾向

Random Read においては、OSD サーバ増加による性能向上がみられなかった。job 数、VM 数増加にともない、比例的に性能向上している傾向がみられた。性能向上傾向を精査するため、OSD サーバ 6 台時における、BS 最小値、最大値の観点でデータを精査した。

### (1) Read、OSD サーバ 6 台、BS 4KB 時

題記条件において、図 5-1 から全体傾向をみるかぎり、IOPS、Throughput とともに比例的に増加しているように見える。表 5-5 で精査したところ、IOPS では VM2、3 台時(1 ノード 1VM をホスト)は比例定数がほぼ 1、VM9 台以上時(1 ノード複数 VM をホスト)において比例定数が 1 未満であることから、Hypervisor 等でなんらかのオーバーヘッドが発生していると推測する。一方 Throughput では、表 5-6 より VM9 台以上において比例定数が 1 を大きく超えており、キャッシュ機能等が働いていると考えられる。

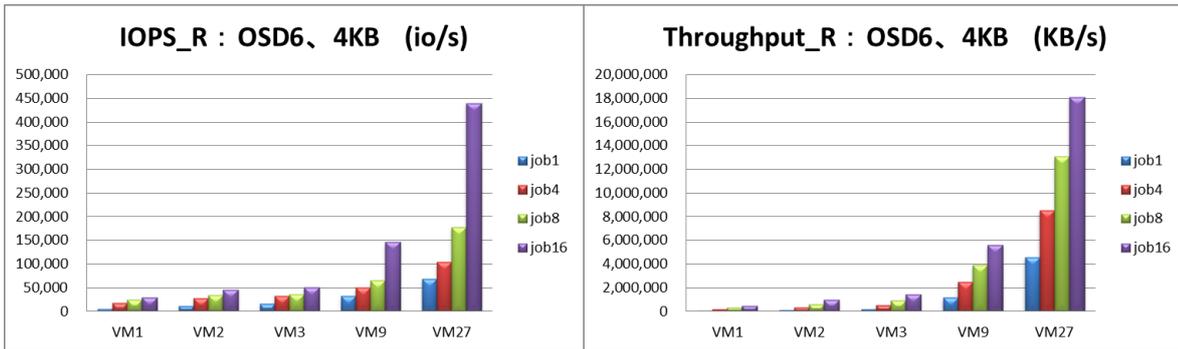


図 5-1 OSD サーバ 6 台、BS 4KB 時の全体傾向

表 5-5 OSD サーバ 6 台、BS 4KB 時の IOPS 傾向

#	job 数	VM1	VM2	VM3	VM9	VM27
1	job1 (対 VM1 比)	5,413	10,730 (2.0)	15,214 (2.8)	32,009 (5.9)	67,828 (12.5)
2	job16 (対 VM1 比)	28,336	44,632 (1.6)	50,373 (1.8)	145,427 (5.1)	439,865 (15.5)

表 5-6 OSD サーバ 6 台、BS 4KB 時の Throughput(KB/s)傾向

#	job 数	VM1	VM2	VM3	VM9	VM27
1	job1 (対 VM1 比)	50,519	102,800 (2.0)	173,806 (3.4)	1,146,594 (22.7)	4,563,478 (90.3)
2	job16 (対 VM1 比)	475,814	938,666 (2.0)	1,403,504 (2.9)	5,574,447 (11.7)	18,060,460 (38.0)

(2) Read、OSD サーバ 6 台、BS 128KB 時

題記条件において、図 5-2 から全体傾向をみるかぎり、IOPS、Throughput とともに比例的に増加しているように見える。表 5-7 表 5-8 より精査したところ、IOPS、Throughput 共に、VM2、3 台時(1 ノード 1VM をホスト)は比例定数がほぼ 1、VM9 台以上時(1 ノード複数 VM をホスト)において比例定数が 1 未満であることから、Hypervisor 等でなんらかのオーバーヘッドが発生していると推測する。

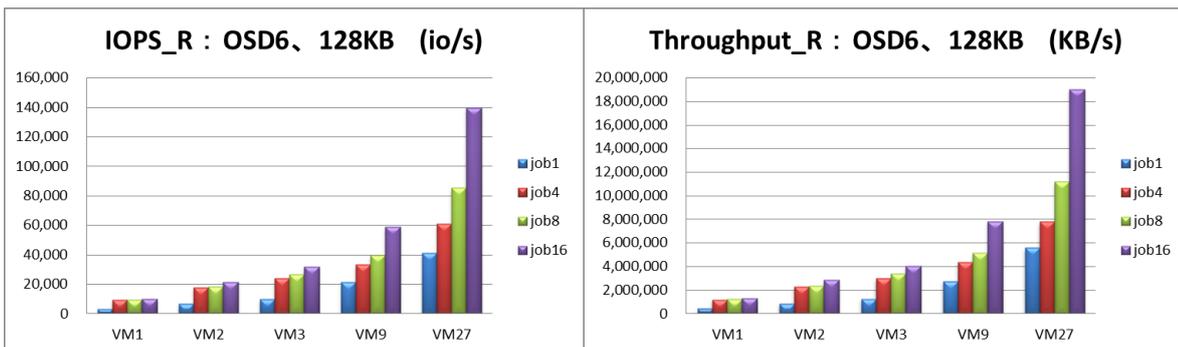


図 5-2 OSD サーバ 6 台、BS 128KB 時の全体傾向

表 5-7 OSD サーバ 6 台、BS 128KB 時の IOPS 傾向

#	job 数	VM1	VM2	VM3	VM9	VM27
1	job1 (対 VM1 比)	3,293	6,552 (2.0)	9,565 (2.9)	21,164 (6.4)	40,824 (12.4)
2	job16 (対 VM1 比)	10,066	21,428 (2.1)	31,583 (3.1)	58,606 (5.8)	139,422 (13.9)

表 5-8 OSD サーバ 6 台、BS 128KB 時の Throughput(KB/s)傾向

#	job 数	VM1	VM2	VM3	VM9	VM27
1	job1 (対 VM1 比)	416,267	836,033 (2.0)	1,241,203 (3.0)	2,728,548 (6.6)	5,566,290 (13.4)
2	job16 (対 VM1 比)	1,306,112	2,825,728 (2.2)	3,996,774 (3.1)	7,804,981 (6.0)	19,027,279 (14.6)

## 6. まとめ

今回の検証の範囲(連続的な負荷をかけた場合)では、Random Read に関しては、最大で 494,491 IOPS(OSD サーバ 3 台、Block Size:4KB、job 数:16、VM:27 時)、28Gbps のスループット(OSD サーバ 3 台、Block Size:32KB、job 数:16、VM:27 時)という高い性能を計測することができた。スループットに関しては、WireSpeed を超えるあたりをとっている事から、OpenStack Compute 側、もしくは Ceph 側でなんらかのキャッシュ機能が働いている可能性がある。

一方 Random Write に関しては、BS 4KB 時において OSD サーバ 1 台あたり、低負荷時で 300 IOPS、高負荷時で 165 IOPS 程度であった。Journal に SSD を使った効果はあまりみられず、HDD の性能に依存している可能性が高い。各々の OSD サーバは OSD Disk として HDD を 3 台搭載していることから、HDD 1 台あたり、低負荷時で 100 IOPS、高付加時で 55 IOPS 程度だと考えられる。

OSD Disk の台数が少ない場合、HDD では Write 性能に課題があると考えられる。Journal への SSD 採用だけでなく、OSD Disk への SSD 採用も検討すべきである。

### 6.1 懸念事項

本検証へ着手する前に、OSD Disk として容量違いの HDD を混在させたところ、ベンチマーク結果に偏りが見られた。OSD Disk の容量ベースでデータが分散され、アクセスが偏ったためと考えられるが、裏付けをとることができなかった。容量の差による挙動傾向や、OSD サーバ数をさらに増加させた場合、OSD サーバあたりの OSD Disk を増減させた際の挙動については、今回得られた結果と異なる可能性がある。

### 7. 参考・関連文献

- [1] OpenStack on Ceph におけるストレージ設計のポイント  
URL=<http://www.osca-jp.com/solution.html>
- [2] Welcome to Ceph - Ceph Documentation  
URL=<http://docs.ceph.com/docs/master/>
- [3] Product Documentation - Red Hat Customer Portal  
URL=<https://access.redhat.com/documentation/en/>
- [4] ビットアイル総合研究所ブログ  
URL=<http://blog.bit-isle.jp/bird/>

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。

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。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、公認や出資も受けていません。

PowerEdge、DELLロゴは、米国Dell Inc.の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 software,Inc.の登録商標です。

本文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。

以上