

PCIe SSDを活用したDB I/Oの簡単解決

2017年 3月

法人事業本部 ソリューション & マーケティング部
華為技術日本株式会社（ファーウェイ・ジャパン）

HUAWEI TECHNOLOGIES CO., LTD.

1

会社紹介・自己紹介

2

MySQL運用中の検討課題(レイテンシ・保守性・容量)

3

ES3600 v3のスペックの確認 (カタログ)

4

MySQLでの検討課題解決

5

ES3600 v3の一押し機能 (Atomic Write SR-IOV)

6

導入実績

自己紹介

- ✓ 水戸部 章生(ミトベ アキオ)
- ✓ 2016年3月に華為技術日本株式会社入社
サーバプロダクト担当
Huawei ES3600 V3のパフォーマンス検証
- ✓ 前職コナミ・デジタルエンタテインメントにてSE
モバイルアプリのインフラ周りのパフォーマンス監視
Perlでのゲーム開発
PHPでの統計システムの開発

会社紹介

華為技術って、どんな会社
英語表記だとHUAWEI



フォーチュングローバル 129位

実は、ご存じないかもしれませんが...

会社紹介

サーバ作ってます。SSDも作ってます。



会社紹介

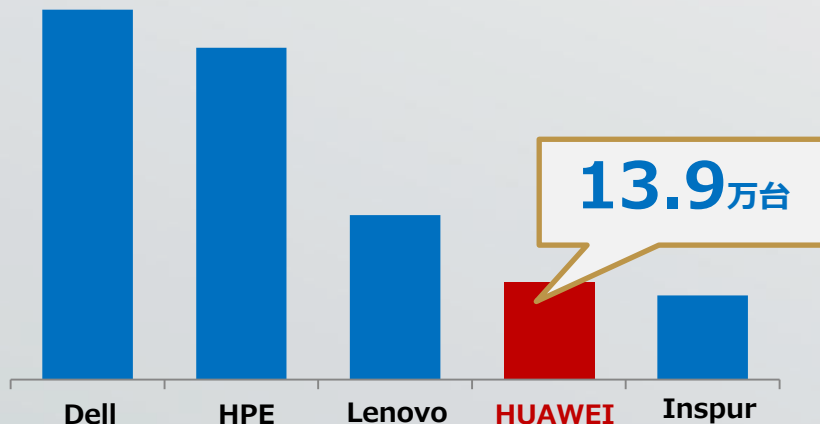
HUAWEI サーバ製品の市場シェア

サーバの出荷台数

12四半期連続

No.4

世界第4位の実績!!



HUAWEIは、2ソケット以上の製品ラインナップで展開！
エンタープライズ市場では、実質“No. 3”

Source: Servers Market Share WW by Region 2016Q2,
Gartner
<http://www.gartner.com/newsroom/id/3441917>

会社紹介

創業以来右肩上がりで成長中!



営業収益

7兆3,273億円

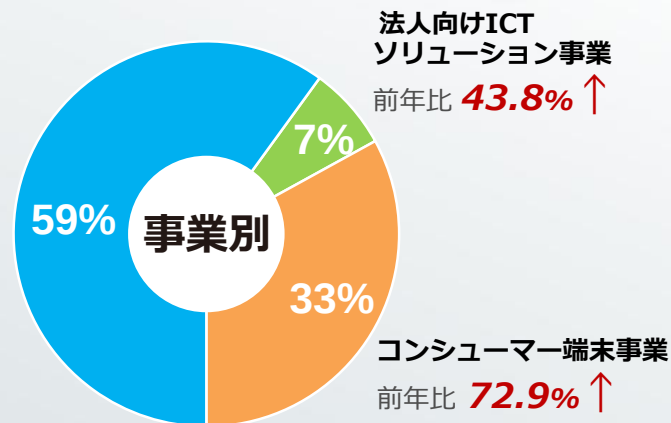
USD Billion



通信事業者向け
ネットワーク事業

前年比

21.4% ↑



法人向けICT
ソリューション事業

前年比 43.8% ↑

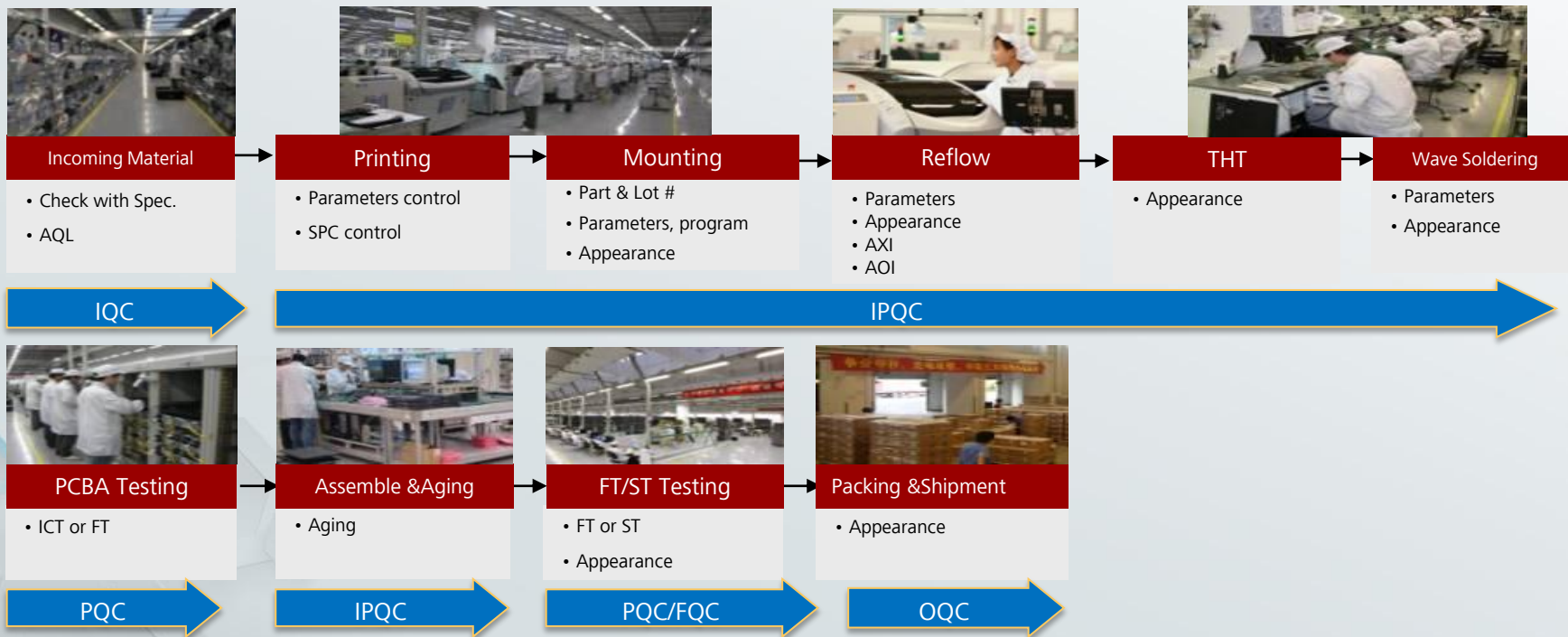
コンシューマー端末事業

前年比 72.9% ↑

	2015	前年比
通信事業者向け	4兆3,092億円	21.4%
法人向け	5,120億円	43.8%
コンシューマー向け	2兆3,948億円	72.9%
その他	1,113億円	
合計	7兆3,273億円	37.1%

会社紹介

しかも、自社製です！品質と価格に自信あります！



会社紹介

ホームページURL :

<http://e.huawei.com/ja/>

1 会社紹介・自己紹介

2 MySQL運用中の検討課題(レイテンシ・保守性・容量)

3 ES3600 v3のスペックの確認 (カタログ)

4 MySQLでの検討課題解決

5 ES3600 v3の一押し機能 (Atomic Write SR-IOV)

6 導入実績

MySQL運用中の検討課題

スロークエリーに泣かされたことがありますよね。

解決します！！

```
[root@localhost mysql]# mysqladmin -u root -p status
Enter password:
Uptime: 9184  Threads: 2  Questions: 57321816  Slow queries: 0  Opens: 528  Flush tables: 1  Open tables: 359  Queries per second avg: 6241.486
[root@localhost mysql]#
```

✓ **レイテンシ**

MySQLでSATA SSDとHuawei PCIe SSD製品 ES3600 V3 との比較

MySQL運用中の検討課題

PCIeカードのSSDが壊れて交換の為にサーバ止めた経験ありますよね。

解決します！！



✓ 保守性

ディスク交換方法と

MySQLから見たオンラインディスク交換時のパフォーマンス比較

MySQL運用中の検討課題

データが増え続けて夜なべして不要データの調査した事ありますよね。

解決します！！



✓ 容量

PCIe SSDが大容量であるメリットは？

1 会社紹介・自己紹介

2 検討課題(レイテンシ・保守性・容量)

3 ES3600 v3のスペックの確認 (カタログ)

4 MySQLでの検討課題解決

5 ES3600 v3の一押し機能 (Atomic Write SR-IOV)

6 導入実績

ES3600 v3のスペックの確認

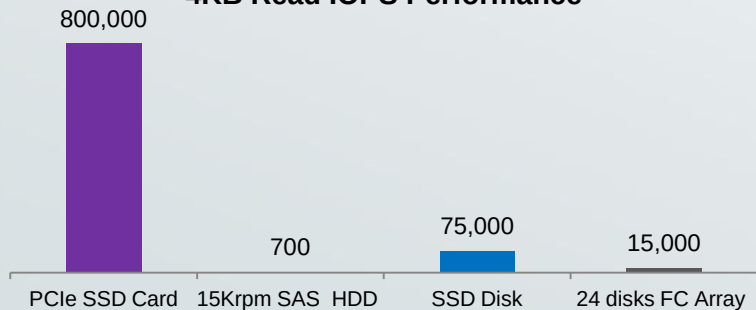
予習です。

ES3600 v3とその他ディスクの性能比較になります。

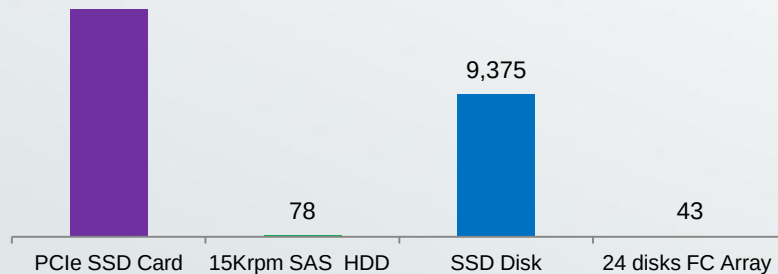
ES3600 v3のスペックの確認

4KB ブロックの IOPS 性能において PCIe SSD カードは 1000基の SAS HDD と同等性能
優れたエネルギー消費効率に加えて高い同時アクセス性能と帯域性能を両立

4KB Read IOPS Performance



Performance/Power Comparison (IOPS/W)



	PCIe SSD Card	15Krpm SAS HDD	SSD Disk	24 disks FC Array
Usable Capacity	3200 GB	600 GB	800 GB	14,400 GB
Stable Read IOPS (4KB)	800,000	700	75,000	15,000
Max. Bandwidth	3100MB/s	200MB/s	500MB/s	1,500MB/s
Max. Power Consumption (W)	22	9	8	350
Avg. Latency	10~100 μs	10~100 ms	0.1~1ms	1~10ms
IOPS/W	36363	78	9,375	43

※ DDR4-2400の転送速度は、19.2GB/s

1 会社紹介・自己紹介

2 MySQL運用中の検討課題(レイテンシ・保守性・容量)

3 ES3600 v3のスペックの確認 (カタログ)

4 MySQLでの検討課題解決

5 ES3600 v3の一押し機能 (Atomic Write SR-IOV)

6 導入実績

MySQLでの検討課題解決

✓レイテンシ ※ 単一クエリのレスポンス改善

環境情報

CentOS7.2

MySQL5.7(innodb、IO Direct指定、innodb_buffer_pool_size=0)

tpcc-mysql利用

SATA SSD(RAID1)とES3600P V3(RAID1)での検証

計測テーブル情報(tpcc-mysql stockテーブル)

テーブル情報

TABLE_SCH	TABLE_NAME	TABLE_TYPE	ENGINE	VERSION	ROW_FORMAT	TABLE_ROWS	AVG_ROW_LENGTH	DATA_LENGTH	MAX_DATA_LENGTH	INDEX_LENGTH	DATA_FREE
tpcc	stock	BASE TABLE	InnoDB	10	Dynamic	46986488	349	16434331648	0	962592768	5242880

実行計画

explain select s_dist_01 from stock;

id	select_type	table	partitions	type	possible_keys	key	key_len	ref	rows	filtered	Extra
1	SIMPLE	stock	NULL	ALL	NULL	NULL	NULL	NULL	46986488	100	NULL

レコード数5千万で全データで約15Gほどのテーブルに対して計測実施。。

※MySQL再起動後にどちらも実施

MySQLでの検討課題解決

✓実施結果

SATA SSD RAID1における結果

Status	Duration	CPU_user	CPU_system	Block_ops_in	Block_ops_out	Page_faults_major	Page_faults_minor	Swaps
starting	0.000046	0.000037	0.000007	0	0	0	0	0
checking permissions	0.000005	0.000004	0	0	0	0	0	0
Opening tables	0.000021	0.000017	0.000003	0	0	0	3	0
init	0.000012	0.00001	0.000002	0	0	0	0	0
System lock	0.00001	0.000008	0.000002	0	0	0	1	0
optimizing	0.000003	0.000002	0	0	0	0	0	0
statistics	0.000008	0.000008	0.000001	0	0	0	0	0
preparing	0.000007	0.000005	0.000001	0	0	0	0	0
executing	0.000002	0.000002	0.000001	0	0	0	0	0
Sending data	51.949	44.303258	9.119838	33367616	0	0	738	0
end	0.000023	0.000015	0.000003	0	0	0	5	0
query end	0.000014	0.000011	0.000002	0	0	0	4	0
closing tables	0.000022	0.000019	0.000004	0	0	0	8	0
freeing items	0.000018	0.000015	0.000003	0	0	0	1	0
cleaning up	0.000039	0	0.000883	0	0	0	11	0

結果

単一クエリによるSending Data処理が、
51.949043s → 38.098275sへ短縮

ES3600P V3 RAID1における結果

Status	Duration	CPU_user	CPU_system	Block_ops_in	Block_ops_out	Page_faults_major	Page_faults_minor	Swaps
starting	0.000071	0.00006	0.000009	0	0	0	0	0
checking permissions	0.000009	0.000007	0.000001	0	0	0	0	0
Opening tables	0.000058	0.000051	0.000008	0	0	0	5	0
init	0.000027	0.000023	0.000003	0	0	0	1	0
System lock	0.000022	0.000019	0.000003	0	0	0	1	0
optimizing	0.000005	0.000005	0.000001	0	0	0	0	0
statistics	0.000016	0.000013	0.000002	0	0	0	0	0
preparing	0.000014	0.000012	0.000002	0	0	0	0	0
executing	0.000003	0.000003	0	0	0	0	0	0
Sending data	38.0983	38.714461	5.813429	33367616	0	0	384	0
end	0.000019	0.000011	0.000002	0	0	0	5	0
query end	0.000013	0.000011	0.000002	0	0	0	4	0
closing tables	0.000018	0.000017	0.000002	0	0	0	8	0
freeing items	0.000016	0.000013	0.000002	0	0	0	1	0
cleaning up	0.002278	0.000203	0.002081	0	0	0	13	0

MySQLでの検討課題解決

✓レイテンシ ※ 並列クエリのレスポンス改善

SSD

tpcCによるデータ取得

```
<Raw Results>
[0] sc:0 lt:9629 rt:0 fl:0 avg_rt: 608.1 (5)
[1] sc:0 lt:9609 rt:0 fl:0 avg_rt: 145.9 (5)
[2] sc:0 lt:960 rt:0 fl:0 avg_rt: 83.0 (5)
[3] sc:0 lt:928 rt:0 fl:0 avg_rt: 11116.7 (80)
[4] sc:0 lt:963 rt:0 fl:0 avg_rt: 714.9 (20)
in 300 sec.
```

```
<Raw Results2(sum ver.)>
[0] sc:0 lt:9629 rt:0 fl:0
[1] sc:0 lt:9609 rt:0 fl:0
[2] sc:0 lt:960 rt:0 fl:0
[3] sc:0 lt:928 rt:0 fl:0
[4] sc:0 lt:963 rt:0 fl:0
```

```
<Constraint Check> (all must be [OK])
[transaction percentage]
Payment: 43.50% (>=43.0%) [OK]
Order-Status: 4.35% (>= 4.0%) [OK]
Delivery: 4.20% (>= 4.0%) [OK]
Stock-Level: 4.36% (>= 4.0%) [OK]
[response time (at least 90% passed)]
New-Order: 0.00% [NG] *
Payment: 0.00% [NG] *
Order-Status: 0.00% [NG] *
Delivery: 0.00% [NG] *
Stock-Level: 0.00% [NG] *
```

<TpmC>

1925.800 TpmC

PCIe SSD

tpcCによるデータ取得

```
<Raw Results>
[0] sc:0 lt:20247 rt:0 fl:0 avg_rt: 161.2 (5)
[1] sc:0 lt:20230 rt:0 fl:0 avg_rt: 40.3 (5)
[2] sc:30 lt:1993 rt:0 fl:0 avg_rt: 24.2 (5)
[3] sc:0 lt:2004 rt:0 fl:0 avg_rt: 6835.0 (80)
[4] sc:0 lt:2024 rt:0 fl:0 avg_rt: 241.4 (20)
in 300 sec.
```

```
<Raw Results2(sum ver.)>
[0] sc:0 lt:20247 rt:0 fl:0
[1] sc:0 lt:20231 rt:0 fl:0
[2] sc:30 lt:1993 rt:0 fl:0
[3] sc:0 lt:2004 rt:0 fl:0
[4] sc:0 lt:2024 rt:0 fl:0
```

```
<Constraint Check> (all must be [OK])
[transaction percentage]
Payment: 43.48% (>=43.0%) [OK]
Order-Status: 4.35% (>= 4.0%) [OK]
Delivery: 4.31% (>= 4.0%) [OK]
Stock-Level: 4.35% (>= 4.0%) [OK]
[response time (at least 90% passed)]
New-Order: 0.00% [NG] *
Payment: 0.00% [NG] *
Order-Status: 1.48% [NG] *
Delivery: 0.00% [NG] *
Stock-Level: 0.00% [NG] *
```

<TpmC>

4049.400 TpmC

結果

並列クエリによるTpmCが、
1925.800TpmC→ 4049.400TpmCへ向上

結果サマリ

PCIe SSDへ換装しただけで

IO時間短縮により

- ・ 単体クエリ速度は、**20%以上**のディスクからのデータ読み取り(Sending data)処理の向上
- ・ 並列クエリ数は、**100%以上**のトランザクション(tpmC)処理数の向上

解決！！

※ 注意：それでも設計は、大事です。

設計しないとPCIe SSDで早くなってもスロークエリーは出ます。

MySQLでの検討課題解決

✓保守性

ホットプラグ対応！！
オンラインでのPCIe SSD交換が可能！！

U.2ディスクであれば、更にディスクの取り出しがスムーズに出来ます。



サーバー筐体を空けなくても、ディスク交換が可能に！

MySQLでの検討課題解決

ディスク1本故障時、リビルド時のパフォーマンスチェック

PCIe SSD

※ 通常時

```
<Raw Results>
[0] sc:0 lt:20247 rt:0 fi:0 avg_rt: 161.2 (5)
[1] sc:0 lt:20230 rt:0 fi:0 avg_rt: 40.3 (5)
[2] sc:30 lt:1993 rt:0 fi:0 avg_rt: 24.2 (5)
[3] sc:0 lt:2004 rt:0 fi:0 avg_rt: 6835.0 (80)
[4] sc:0 lt:2024 rt:0 fi:0 avg_rt: 241.4 (20)
in 300 sec.
```

```
<Raw Results2(sum ver.)>
[0] sc:0 lt:20247 rt:0 fi:0
[1] sc:0 lt:20231 rt:0 fi:0
[2] sc:30 lt:1993 rt:0 fi:0
[3] sc:0 lt:2004 rt:0 fi:0
[4] sc:0 lt:2024 rt:0 fi:0
```

```
<Constraint Check> (all must be [OK])
[transaction percentage]
  Payment: 43.48% (>=43.0%) [OK]
Order-Status: 4.35% (>= 4.0%) [OK]
  Delivery: 4.31% (>= 4.0%) [OK]
  Stock-Level: 4.35% (>= 4.0%) [OK]
[response time (at least 90% passed)]
  New-Order: 0.00% [NG] *
  Payment: 0.00% [NG] *
  Order-Status: 1.48% [NG] *
  Delivery: 0.00% [NG] *
  Stock-Level: 0.00% [NG] *
```

<TpmC>

4049.400 TpmC

PCIe SSD

※ ディスク1本故障時

```
<Raw Results>
[0] sc:0 lt:20247 rt:0 fi:0 avg_rt: 161.2 (5)
[1] sc:0 lt:20230 rt:0 fi:0 avg_rt: 40.3 (5)
[2] sc:30 lt:1993 rt:0 fi:0 avg_rt: 24.2 (5)
[3] sc:0 lt:2004 rt:0 fi:0 avg_rt: 6835.0 (80)
[4] sc:0 lt:2024 rt:0 fi:0 avg_rt: 241.4 (20)
in 300 sec.
```

```
<Raw Results2(sum ver.)>
[0] sc:0 lt:20247 rt:0 fi:0
[1] sc:0 lt:20231 rt:0 fi:0
[2] sc:30 lt:1993 rt:0 fi:0
[3] sc:0 lt:2004 rt:0 fi:0
[4] sc:0 lt:2024 rt:0 fi:0
```

```
<Constraint Check> (all must be [OK])
[transaction percentage]
  Payment: 43.48% (>=43.0%) [OK]
Order-Status: 4.35% (>= 4.0%) [OK]
  Delivery: 4.31% (>= 4.0%) [OK]
  Stock-Level: 4.35% (>= 4.0%) [OK]
[response time (at least 90% passed)]
  New-Order: 0.00% [NG] *
  Payment: 0.00% [NG] *
  Order-Status: 1.48% [NG] *
  Delivery: 0.00% [NG] *
  Stock-Level: 0.00% [NG] *
```

<TpmC>

4090.300 TpmC

PCIe SSD

※ リビルド時

```
<Raw Results>
[0] sc:0 lt:20801 rt:0 fi:0 avg_rt: 133.2 (5)
[1] sc:0 lt:20789 rt:0 fi:0 avg_rt: 35.0 (5)
[2] sc:77 lt:2002 rt:0 fi:0 avg_rt: 20.8 (5)
[3] sc:0 lt:2062 rt:0 fi:0 avg_rt: 6954.2 (80)
[4] sc:0 lt:2079 rt:0 fi:0 avg_rt: 210.5 (20)
in 300 sec.
```

```
<Raw Results2(sum ver.)>
[0] sc:0 lt:20801 rt:0 fi:0
[1] sc:0 lt:20789 rt:0 fi:0
[2] sc:77 lt:2002 rt:0 fi:0
[3] sc:0 lt:2062 rt:0 fi:0
[4] sc:0 lt:2079 rt:0 fi:0
```

```
<Constraint Check> (all must be [OK])
[transaction percentage]
  Payment: 43.48% (>=43.0%) [OK]
Order-Status: 4.35% (>= 4.0%) [OK]
  Delivery: 4.31% (>= 4.0%) [OK]
  Stock-Level: 4.35% (>= 4.0%) [OK]
[response time (at least 90% passed)]
  New-Order: 0.00% [NG] *
  Payment: 0.00% [NG] *
  Order-Status: 3.70% [NG] *
  Delivery: 0.00% [NG] *
  Stock-Level: 0.00% [NG] *
```

<TpmC>

4160.200 TpmC

結果

パフォーマンスに影響なし

結果サマリ

ホットプラグが対応により、サーバの停止を行わなくて良くなり保守性がUP
また、ソフトウェアRAIDを組んでいれば、万が一の故障時でも、パフォーマンスに影響なく交換が可能

解決！！

※ 注意：あくまで、今回のテストケースにおける結果です。
利用時には、必ずテストを行ってください。

MySQLでの検討課題解決

✓ 容量

6.4TのPCIe SSDが発売されます。

しかも、2Uのサーバに最大12本搭載できるサーバもあります。

RAID0なら、76T以上。。

解決！！



RH228H V3 NVMeモデル
4基 ~ 12基

MySQLでの検討課題解決

ラインナップ

ES3x00P V3 NVMe Drive	モデル		フォームファクタ	接続規格	容量タイプ	寿命
		ES3500P V3	2.5インチドライブ	PCIe Gen3 4レーン NVMe SFF8639	1600GB	1 DWPD, 5年
					2000GB	
					3200GB	
		ES3600P V3	2.5インチドライブ	PCIe Gen3 4レーン NVMe SFF8639	1200GB	3 DWPD, 5年
					1600GB	
					3200GB	
		ES3600C V3 ロープロファイル	ハーフハイト ハーフレングス	PCIe Gen3 4レーン	1600GB	3 DWPD, 5年
					3200GB	

1 会社紹介・自己紹介

2 MySQL運用中の検討課題(レイテンシ・保守性・容量)

3 ES3600 v3のスペックの確認 (カタログ)

4 MySQLでの検討課題解決

5 ES3600 v3の一押し機能 (Atomic Write SR-IOV)

6 導入実績

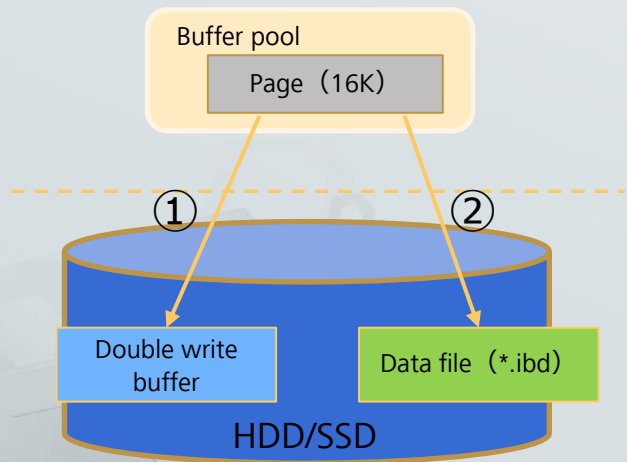
ES3600 v3の一押し機能

Atomic Write機能

Double Writeによるデータ書込み

- ① Buffer poolからDouble write Bufferへの書込み
- ② Buffer poolからData fileへの書込み

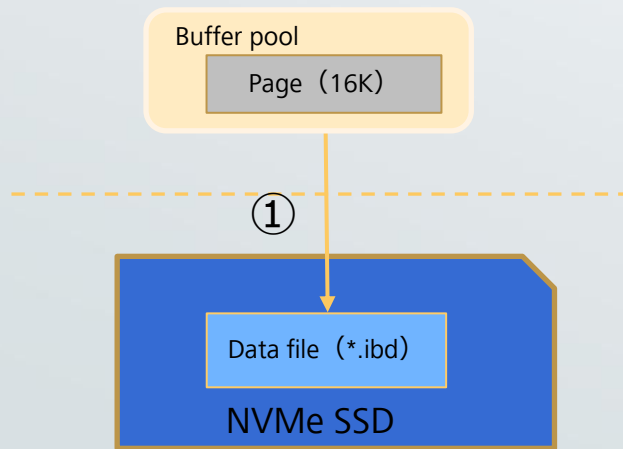
電源ロス時などでのデータロス対策



AtomicWriteによるデータ書込み

- ① Buffer poolからData fileへの書込み

OSとアプリで違うブロックサイズをNVMEでデータロスを担保



効果

- MySQL TPMC **7%**向上
- SSD寿命 **40%**向上

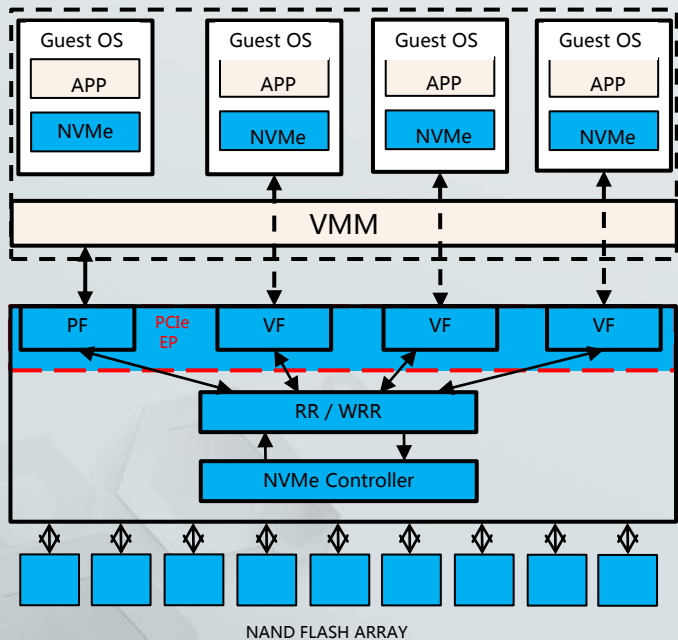
年内実装予定

MySQLに最適

ES3600 v3の一押し機能

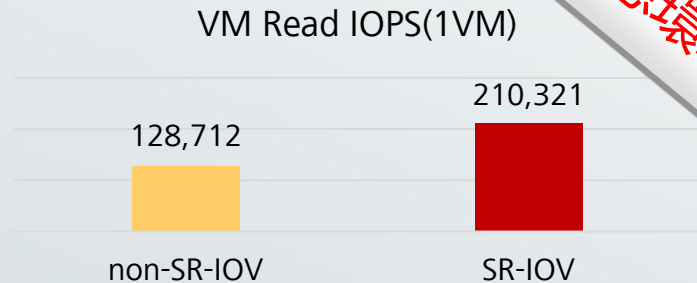
SR-IOV機能

VMサーバからVMMを透過してハードウェアにアクセス



效果：

各VMI/O处理 40%向上



定予发表

仮想環境に最適

1 会社紹介・自己紹介

2 MySQL運用中の検討課題(レイテンシ・保守性・容量)

3 ES3600 v3のスペックの確認 (カタログ)

4 MySQLでの検討課題解決

5 ES3600 v3の一押し機能 (Atomic Write SR-IOV)

6 導入実績

実績

株式会社 サイバーエージェント様

- ✓ CDN配信システムで利用中
- ✓ DBでの利用も検討中

株式会社サイバー
エージェント様

大手ゲーム系S社様

DBで利用中

大手ゲーム系S,C社

数十社様のお客様にご利用頂いております。

Etc....

うれしい事に皆様から品質が良いと高評価頂いてます。

1 会社紹介・自己紹介

2 MySQL運用中の検討課題(レイテンシ・保守性・容量)

3 ES3600 v3のスペックの確認 (カタログ)

4 MySQLでの検討課題解決

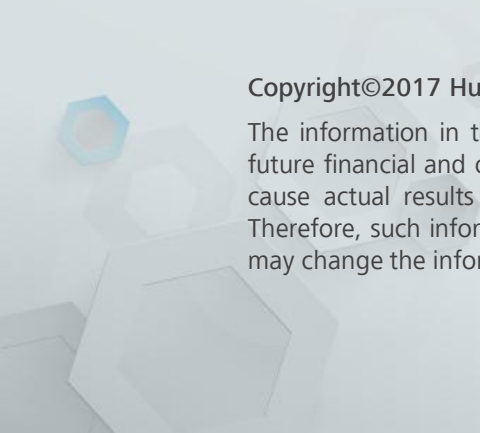
5 ES3600 v3の一押し機能 (Atomic Write SR-IOV)

6 導入実績



THANK YOU

www.huawei.com



Copyright©2017 Huawei Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.