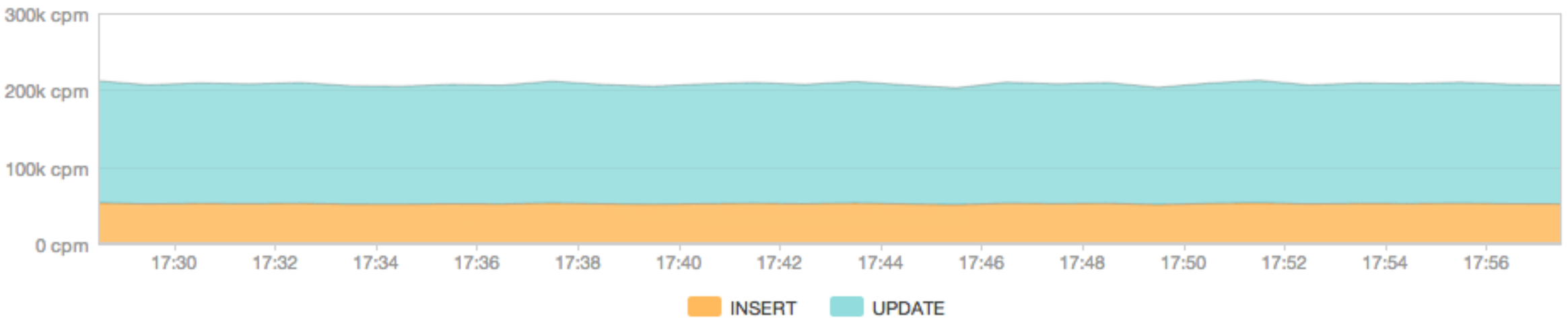
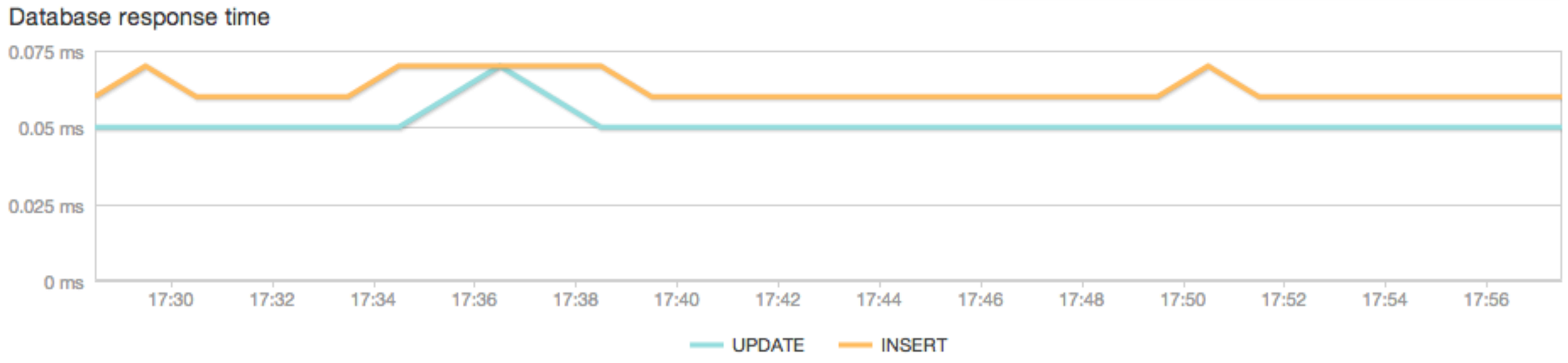
NoSQL Infrastructure

Database throughput



Database response time



~3333 write ops/s

0.07 - 0.05 ms response

David Mytton

Woop Japan!

MongoDB Dashboard

Databases	Total data size	Total index size	Total inserts	Queries	Updates	Deletes
8	74.96GB	25.51GB	876.70/s	729.27/s	3,515.47/s	0.00/s

Alerts

! rs2a - System: Disk usage (/mongodbdata): 80%

pb1 - pb1a — 42% | 0.00 | 0.31GB | 1.42GB | 1.66GB | 7,794 | 0 / 0

pb1arbiter	pb1b
- ♥1s	2s ♥2s

set1 - rs1b — 40% | 0.00 | 13.00MB | 40.00MB | 7.55GB | 7,168 | 0 / 0

rs1a	rs1arbiter	rs1c	rs1d
0s ♥0s	- ♥1s	3s ♥1s	2h ♥0s

set2 - rs2a — 95% | 0.00 | 24.42GB | 72.10GB | -8.98GB | 7,165 | 0 / 0

rs2arbiter	rs2b	rs2c	rs2d
- ♥0s	1s ♥1s	1s ♥1s	2h ♥1s

set3 - rs3a — 19% | 0.00 | 0.77GB | 1.40GB | 6.79GB | 7,064 | 0 / 0

rs3arbiter	rs3b	rs3c	rs3d
- ♥1s	1s ♥1s	1s ♥1s	2h ♥0s

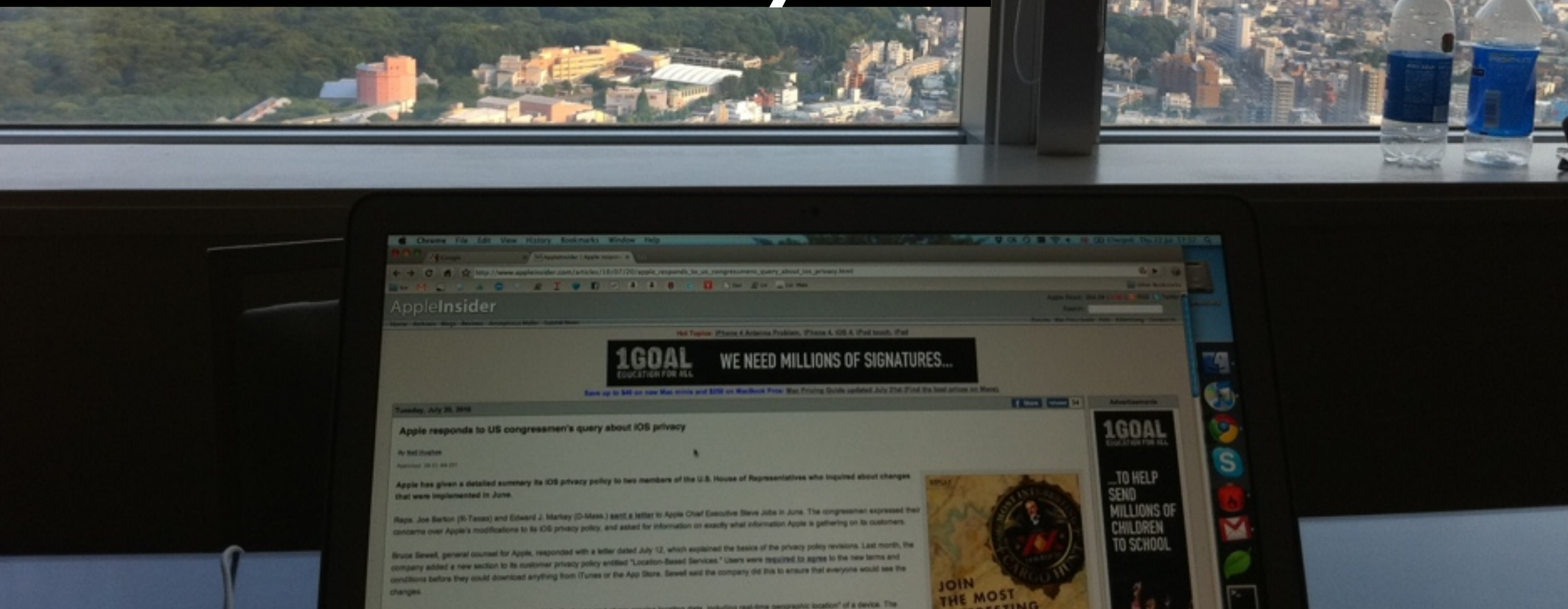
MongoDB at Server Density

MongoDB at Server Density

- 27 nodes

MongoDB at Server Density

- 27 nodes
- June 2009 - 4yrs



MongoDB at Server Density

- 27 nodes
- June 2009 - 4yrs
- MySQL -> MongoDB

MongoDB at Server Density

- 27 nodes
- June 2009 - 4yrs
- MySQL -> MongoDB
- 20TB data per month

Why?

ビュービュー
かの部屋で火事です

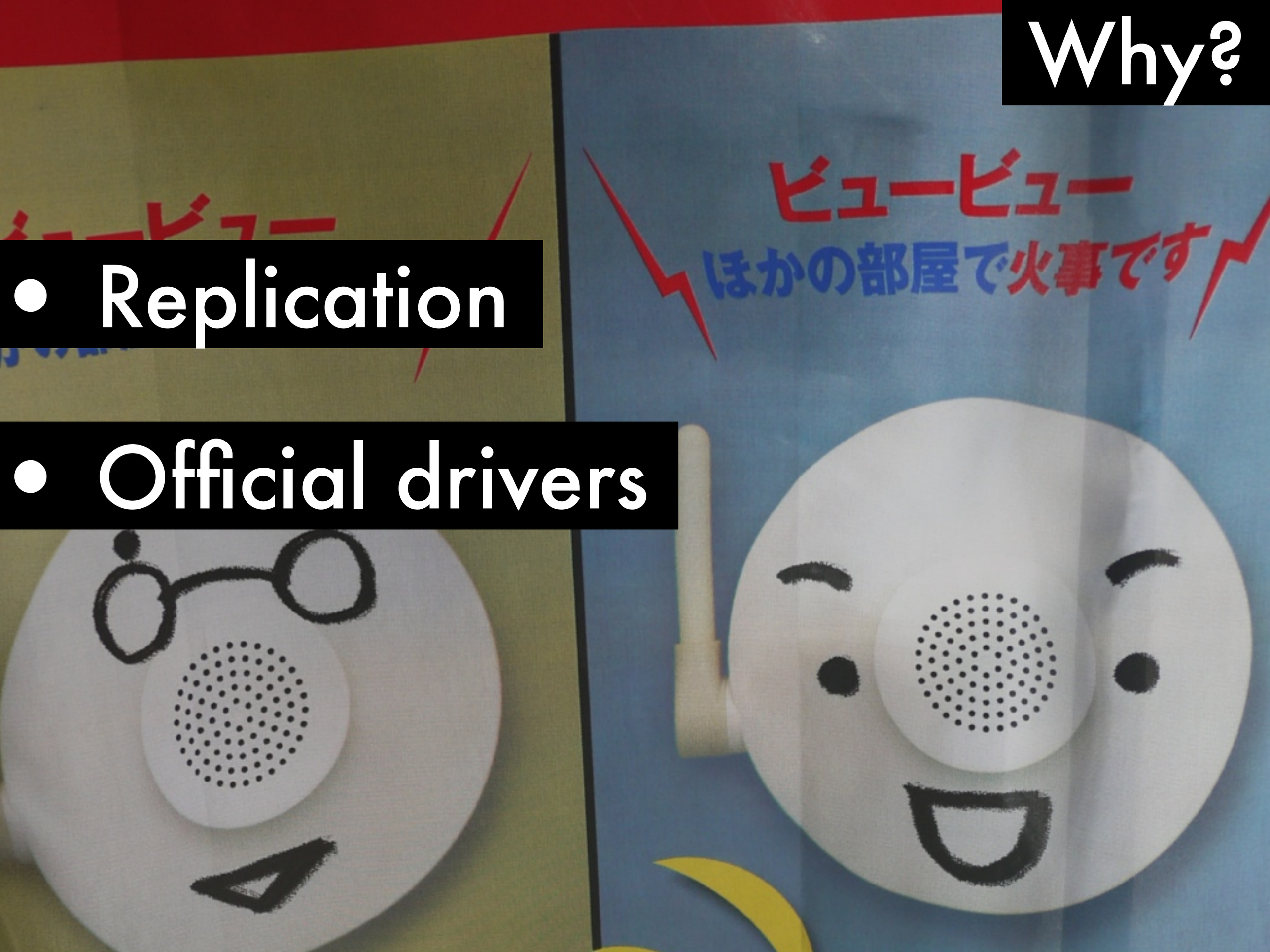
ビュービュー
ほかの部屋で火事です

Why?

- Replication

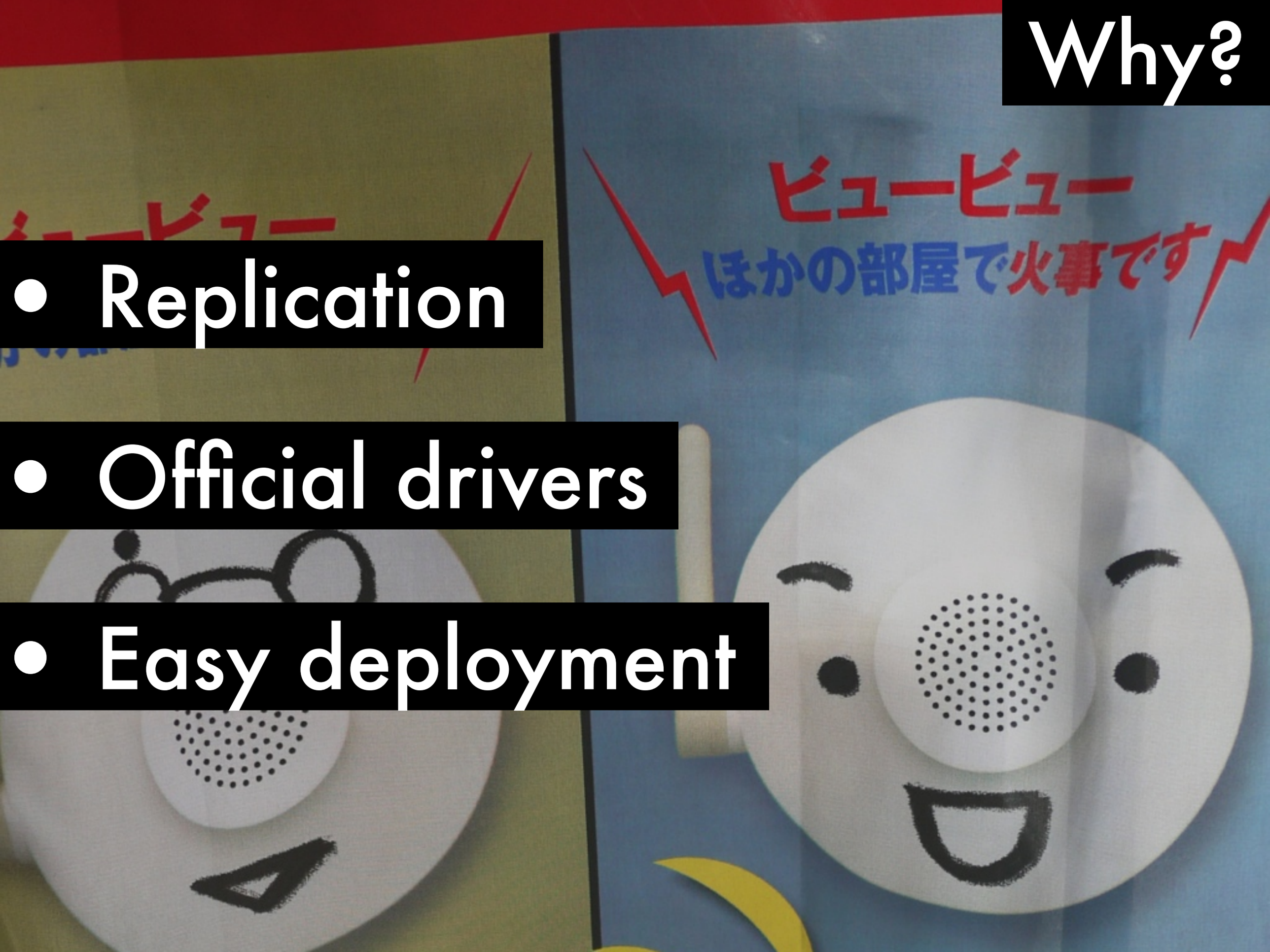
Why?

- Replication
- Official drivers



Why?

- Replication
- Official drivers
- Easy deployment



Why?

- Replication
- Official drivers
- Easy deployment
- Fast out of the box

It's a little different.

Performance

- Fast network

Picture is unrelated! Mmm, ice cream.

Performance

- Fast network

EC2 10 Gigabit Ethernet

- Cluster Compute
- High Memory Cluster
- Cluster GPU
- High I/O
- High Storage

Single Public and Private Ports (2 ports total)	
10 Mbps Ports	$0.00
100 Mbps Ports	$0.00
1 Gbps Ports	$20.00
10Gbps Ports*	$100.00

* Where available/Must meet hardware requirements

Dual Public and Private Ports (4 ports total)	
10 Mbps Dual Ports (20 Mbps max. throughput)*	$10.00
100 Mbps Dual Ports (200 Mbps max. throughput)*	$20.00
1 Gbps Dual Ports (2 Gbps max. throughput)*	$40.00

* Where available/Must meet hardware requirements

Performance

- Fast network

Workload: Read/Write?

What is being stored?

Result set size

Performance

- Fast network

Use	Network Throughput
Normal	0-100Mb/s
Replication (Initial Sync)	Burst +100Mb/s
Replication (Oplog)	0-100Mb/s
Backup	Initial Sync + Oplog

Performance

- Fast network

Inter-DC LAN

```
15/05/13 07:42:23 david@mtx2-md1a ~: ping mtx2-md2a
PING mtx2-md2a.wdc.sl.serverdensity.net (10.56.164.87) 56(84) bytes of data.
64 bytes from mtx2-md2a.wdc.sl.serverdensity.net (10.56.164.87): icmp_req=1 ttl=64 time=0.472 ms
64 bytes from mtx2-md2a.wdc.sl.serverdensity.net (10.56.164.87): icmp_req=2 ttl=64 time=0.570 ms
64 bytes from mtx2-md2a.wdc.sl.serverdensity.net (10.56.164.87): icmp_req=3 ttl=64 time=0.425 ms
64 bytes from mtx2-md2a.wdc.sl.serverdensity.net (10.56.164.87): icmp_req=4 ttl=64 time=0.398 ms
64 bytes from mtx2-md2a.wdc.sl.serverdensity.net (10.56.164.87): icmp_req=5 ttl=64 time=0.623 ms
64 bytes from mtx2-md2a.wdc.sl.serverdensity.net (10.56.164.87): icmp_req=6 ttl=64 time=0.444 ms
64 bytes from mtx2-md2a.wdc.sl.serverdensity.net (10.56.164.87): icmp_req=7 ttl=64 time=0.297 ms
^C
--- mtx2-md2a.wdc.sl.serverdensity.net ping statistics ---
7 packets transmitted, 7 received, 0% packet loss, time 5997ms
rtt min/avg/max/mdev = 0.297/0.461/0.623/0.101 ms
```

Performance

- Fast network

Inter-DC LAN

```
15/05/13 07:42:23 david@mtx2-md1a ~: ping mtx2-md2a
PING mtx2-md2a.wdc.sl.serverdensity.net (10.56.164.87) 56(84) bytes of data.
64 bytes from mtx2-md2a.wdc.sl.serverdensity.net (10.56.164.87): icmp_req=1 ttl=64 time=0.472 ms
64 bytes from mtx2-md2a.wdc.sl.serverdensity.net (10.56.164.87): icmp_req=2 ttl=64 time=0.570 ms
64 bytes from mtx2-md2a.wdc.sl.serverdensity.net (10.56.164.87): icmp_req=3 ttl=64 time=0.425 ms
64 bytes from mtx2-md2a.wdc.sl.serverdensity.net (10.56.164.87): icmp_req=4 ttl=64 time=0.398 ms
64 bytes from mtx2-md2a.wdc.sl.serverdensity.net (10.56.164.87): icmp_req=5 ttl=64 time=0.623 ms
64 bytes from mtx2-md2a.wdc.sl.serverdensity.net (10.56.164.87): icmp_req=6 ttl=64 time=0.444 ms
64 bytes from mtx2-md2a.wdc.sl.serverdensity.net (10.56.164.87): icmp_req=7 ttl=64 time=0.297 ms
^C
--- mtx2-md2a.wdc.sl.serverdensity.net ping statistics ---
7 packets transmitted, 7 received, 0% packet loss, time 5997ms
rtt min/avg/max/mdev = 0.297/0.461/0.623/0.101 ms
```

```
15/05/13 08:13:27 david@mtx2-md1a ~: ping mtx2-md1b
PING mtx2-md1b.sjc.sl.serverdensity.net (10.52.8.160) 56(84) bytes of data.
64 bytes from mtx2-md1b.sjc.sl.serverdensity.net (10.52.8.160): icmp_req=1 ttl=54 time=71.7 ms
64 bytes from mtx2-md1b.sjc.sl.serverdensity.net (10.52.8.160): icmp_req=2 ttl=54 time=72.1 ms
64 bytes from mtx2-md1b.sjc.sl.serverdensity.net (10.52.8.160): icmp_req=3 ttl=54 time=71.9 ms
64 bytes from mtx2-md1b.sjc.sl.serverdensity.net (10.52.8.160): icmp_req=4 ttl=54 time=72.1 ms
64 bytes from mtx2-md1b.sjc.sl.serverdensity.net (10.52.8.160): icmp_req=5 ttl=54 time=71.9 ms
64 bytes from mtx2-md1b.sjc.sl.serverdensity.net (10.52.8.160): icmp_req=6 ttl=54 time=72.0 ms
64 bytes from mtx2-md1b.sjc.sl.serverdensity.net (10.52.8.160): icmp_req=7 ttl=54 time=71.9 ms
^C
--- mtx2-md1b.sjc.sl.serverdensity.net ping statistics ---
7 packets transmitted, 7 received, 0% packet loss, time 6007ms
rtt min/avg/max/mdev = 71.780/72.000/72.158/0.235 ms
```

Cross USA Washington, DC - San Jose, CA

Performance

- Fast network

Location	Ping RTT Latency
Within USA	40-80ms
Trans-Atlantic	100ms
Trans-Pacific	150ms
Europe - Japan	300ms

Failover

- Replica sets

Failover

- Replica sets
- Master/slave

Failover

- Replica sets
- Master/slave
- Min 3 nodes

Failover

- Replica sets
- Master/slave
- Min 3 nodes
- Automatic failover

Performance

- Replication lag

Location	Ping RTT Latency
Within USA	40-80ms
Trans-Atlantic	100ms
Trans-Pacific	150ms
Europe - Japan	300ms

Replication Lag

1. Reads: eventual consistency

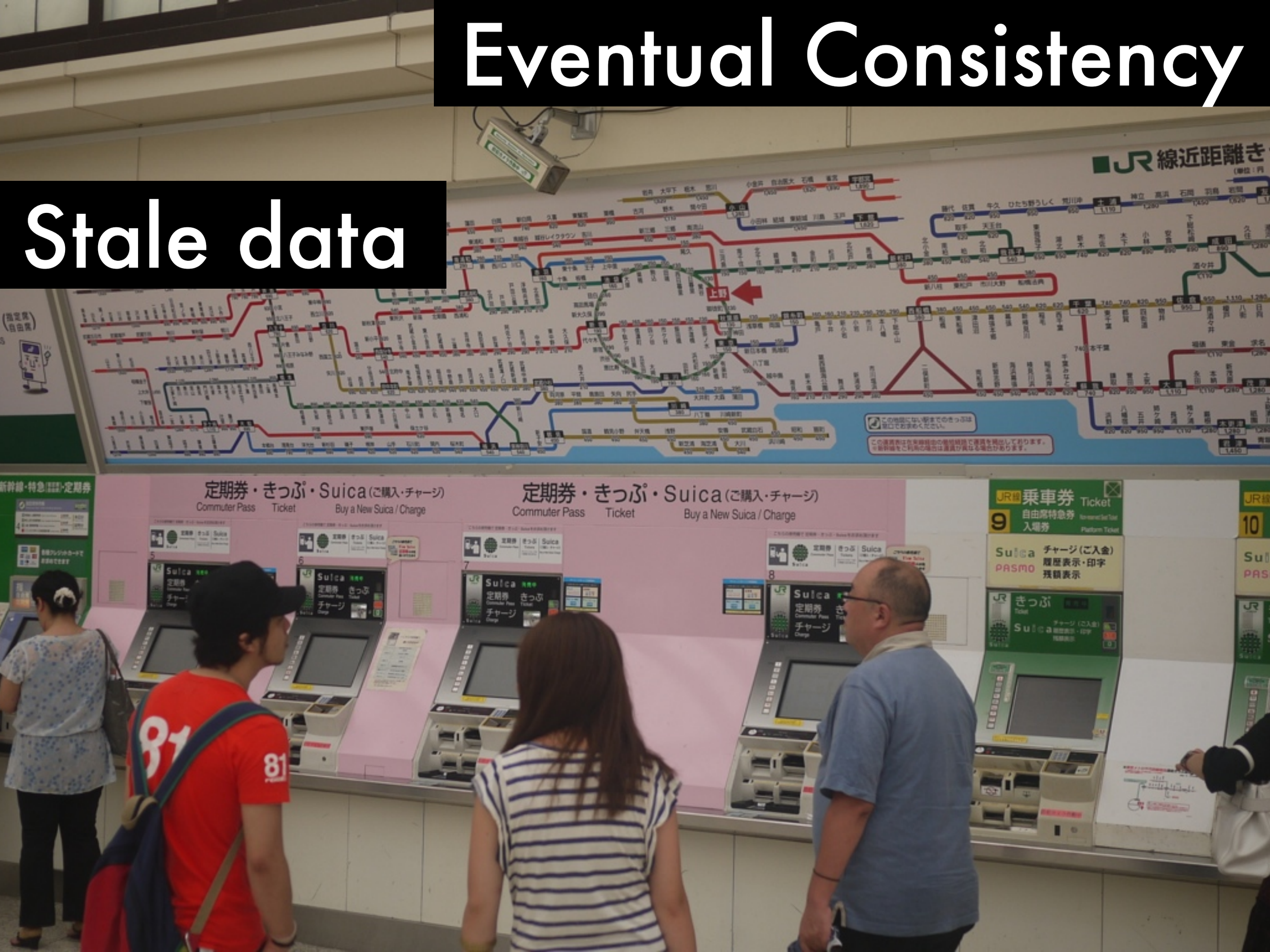
Replication Lag

1. Reads: eventual consistency

2. Failover: slave behind

Eventual Consistency

Stale data



Eventual Consistency

Stale data

Inconsistent data

Eventual Consistency

Stale data

Inconsistent data

Changing data

Eventual Consistency

Use Case	Needs consistency?
Graphs	No
User profile	Yes
Statistics	Depends
Alert config	Yes

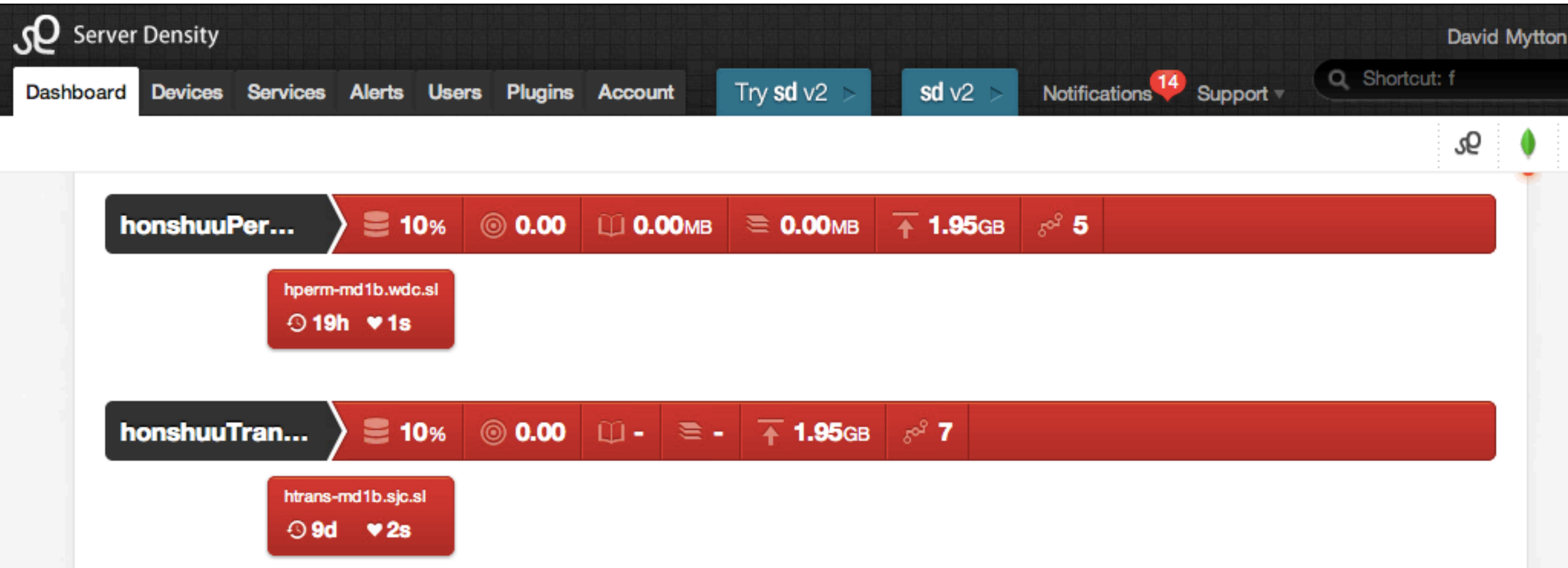
Replication Lag

1. Reads: eventual consistency

2. Failover: slave behind

Slave behind

Failover: out of date master



MongoDB WriteConcern

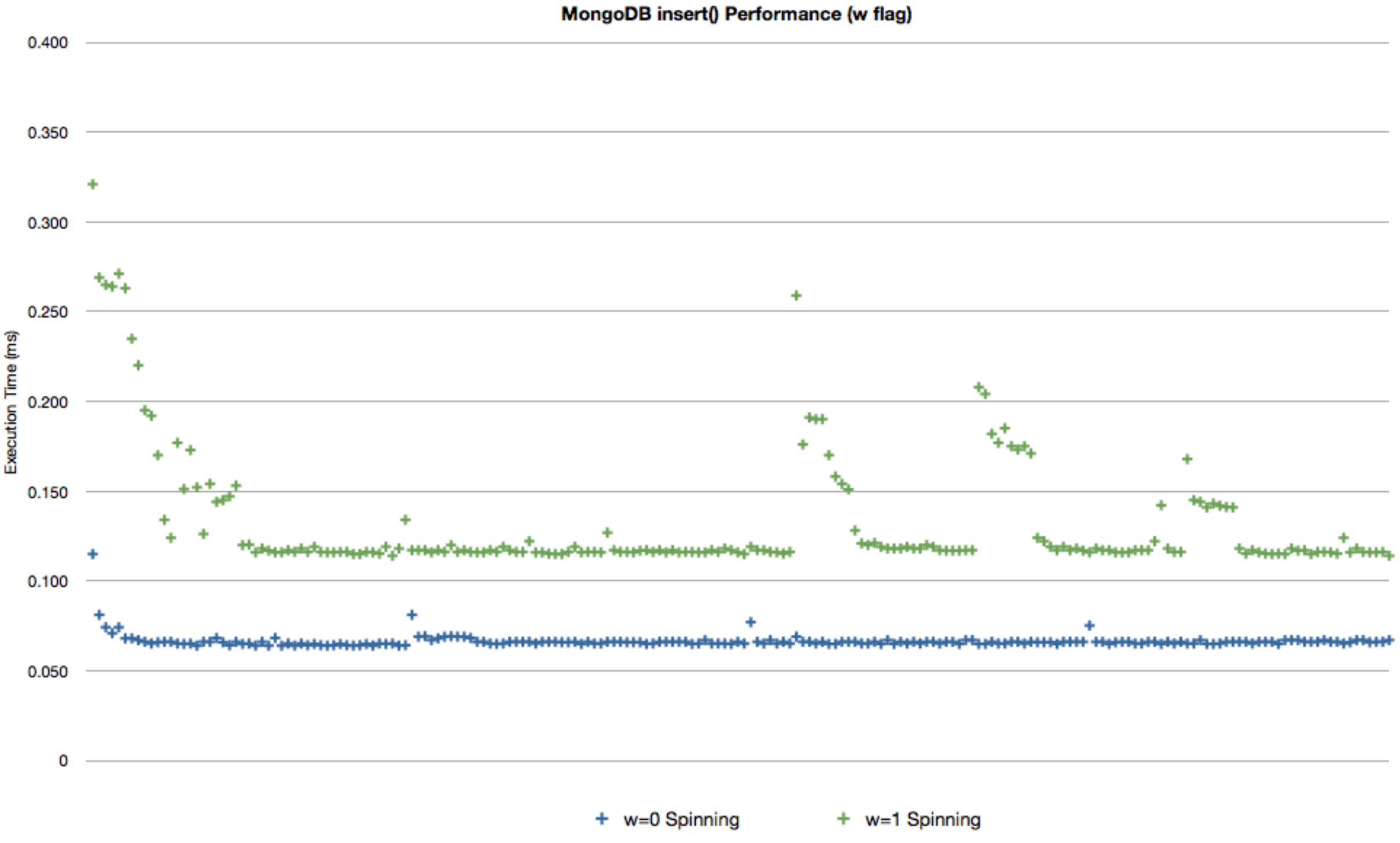
Changed Nov 27 2012

MongoDB WriteConcern

- Safe by default

```
>>> from pymongo import MongoClient
>>> connection = MongoClient(w=int/str)
```

Value	Meaning
0	Unsafe
1	Primary
2	Primary + x1 secondary
3	Primary + x2 secondaries


MongoDB insert() Performance (w flag)
0.400
0.350
0.300
0.250
0.200
0.150
0.100
0.050
0
Execution Time (ms)
w=0 Spinning
w=1 Spinning

Tags

```
{
    _id : "someSet",
    members : [
        {_id : 0, host : "A", tags : {"dc": "ny"}},
        {_id : 1, host : "B", tags : {"dc": "ny"}},
        {_id : 2, host : "C", tags : {"dc": "sf"}},
        {_id : 3, host : "D", tags : {"dc": "sf"}},
        {_id : 4, host : "E", tags : {"dc": "cloud"}}
    ]
    settings : {
        getLastErrorModes : {
            veryImportant : {"dc" : 3},
            sortOfImportant : {"dc" : 2}
        }
    }
}
> db.foo.insert({x:1})
> db.runCommand({getLastError : 1, w : "veryImportant"})
```

Tags

(A or B) + (C or D) + E

```
{
    _id : "someSet",
    members : [
        {_id : 0, host : "A", tags : {"dc": "ny"}},
        {_id : 1, host : "B", tags : {"dc": "ny"}},
        {_id : 2, host : "C", tags : {"dc": "sf"}},
        {_id : 3, host : "D", tags : {"dc": "sf"}},
        {_id : 4, host : "E", tags : {"dc": "cloud"}}
    ]
    settings : {
        getLastErrorModes : {
            veryImportant : {"dc" : 3},
            sortOfImportant : {"dc" : 2}
        }
    }
}
> db.foo.insert({x:1})
> db.runCommand({getLastError : 1, w : "veryImportant"})
```

Tags

(A + C) or (D + E) ...

```
{
    _id : "someSet",
    members : [
        {_id : 0, host : "A", tags : {"dc": "ny"}},
        {_id : 1, host : "B", tags : {"dc": "ny"}},
        {_id : 2, host : "C", tags : {"dc": "sf"}},
        {_id : 3, host : "D", tags : {"dc": "sf"}},
        {_id : 4, host : "E", tags : {"dc": "cloud"}}
    ]
    settings : {
        getLastErrorModes : {
            veryImportant : {"dc" : 3},
            sortOfImportant : {"dc" : 2}
        }
    }
}
> db.foo.insert({x:1})
> db.runCommand({getLastError : 1, w : "sortOfImportant"})
```

Performance

- Fast network
- More RAM

Picture is unrelated! Mmm, ice cream.

Amazon EC2 Instance Types

Standard On-Demand Instances	Linux/UNIX Usage
~~Small (Default)~~	~~$0.085 per hour~~
Large	$0.34 per hour
Extra Large	$0.68 per hour
Micro On-Demand Instances	
Micro	$0.02 per hour
High-Memory On-Demand Instances	
Extra Large	$0.50 per hour
Double Extra Large	$1.00 per hour
Quadruple Extra Large	$2.00 per hour
High-CPU On-Demand Instances	
~~Medium~~	~~$0.17 per hour~~
Extra Large	$0.68 per hour
Cluster Compute Instances	
Quadruple Extra Large	$1.60 per hour
Cluster GPU Instances	
Quadruple Extra Large	$2.10 per hour

32-bit = Don't Use

Typical MongoD

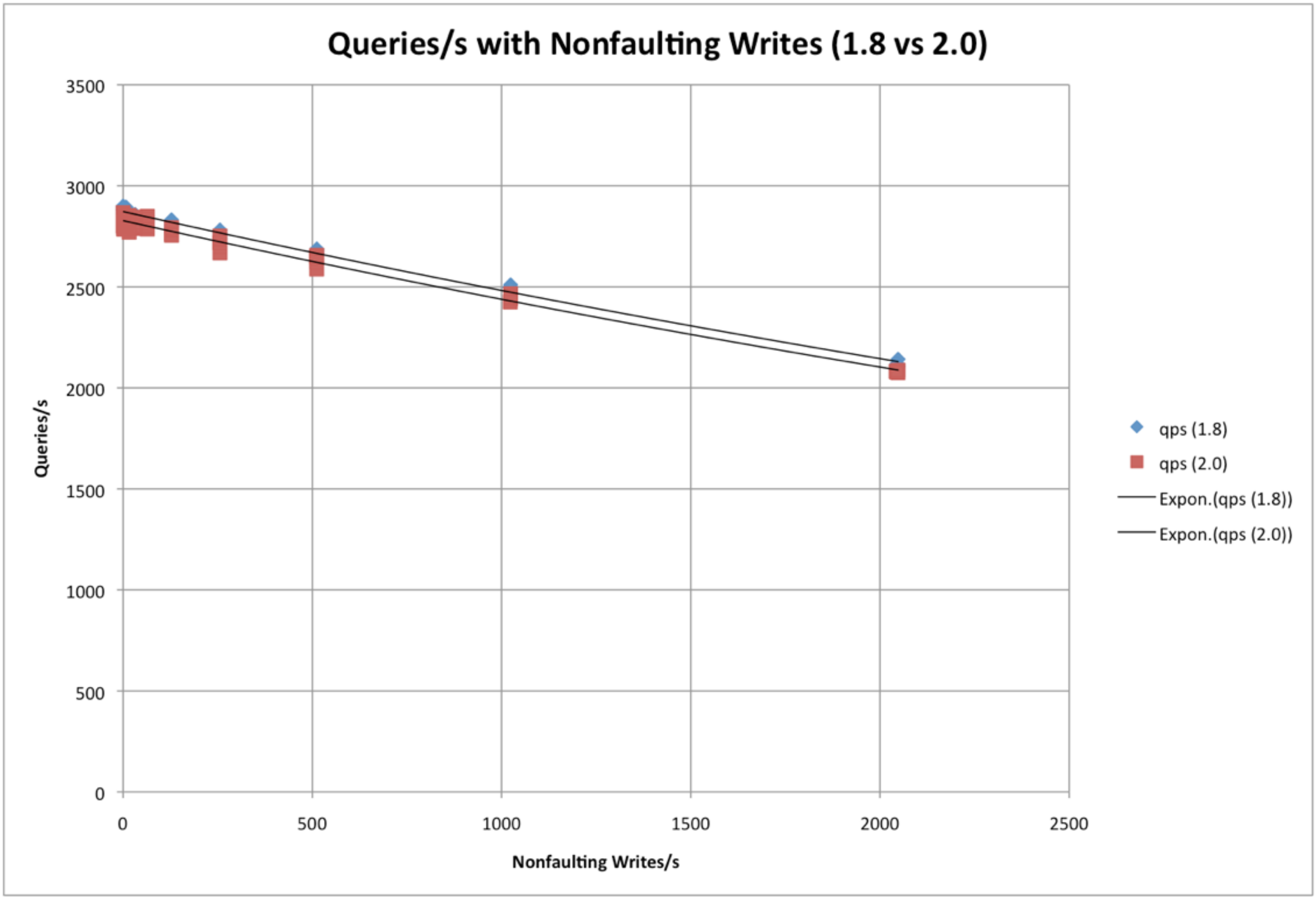
ConfigD / Arbiter

Big MongoD

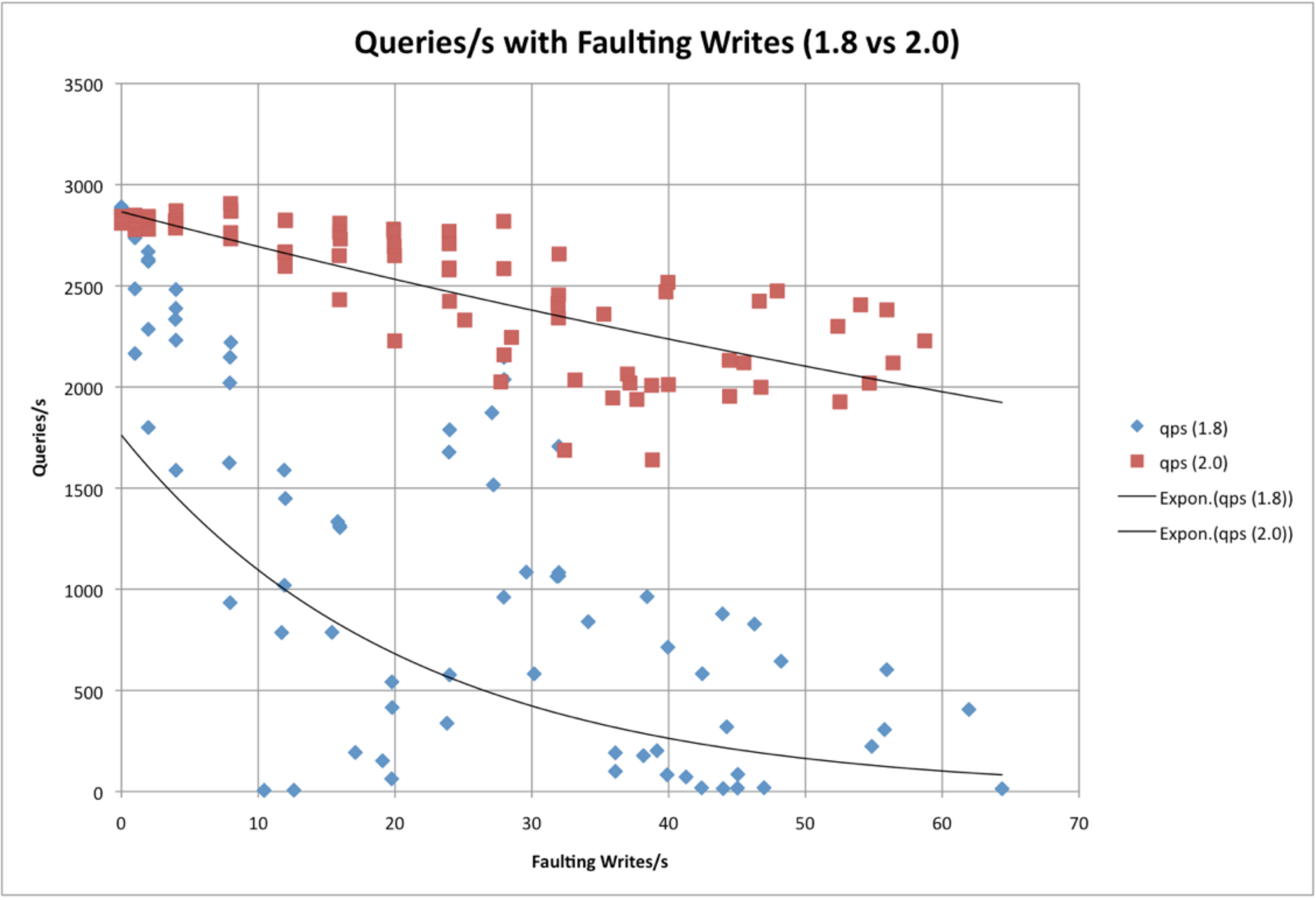
32-bit = Don't Use

High CPU not necessary

http://www.slideshare.net/jrosoff/mongodb-on-ec2-and-ebs



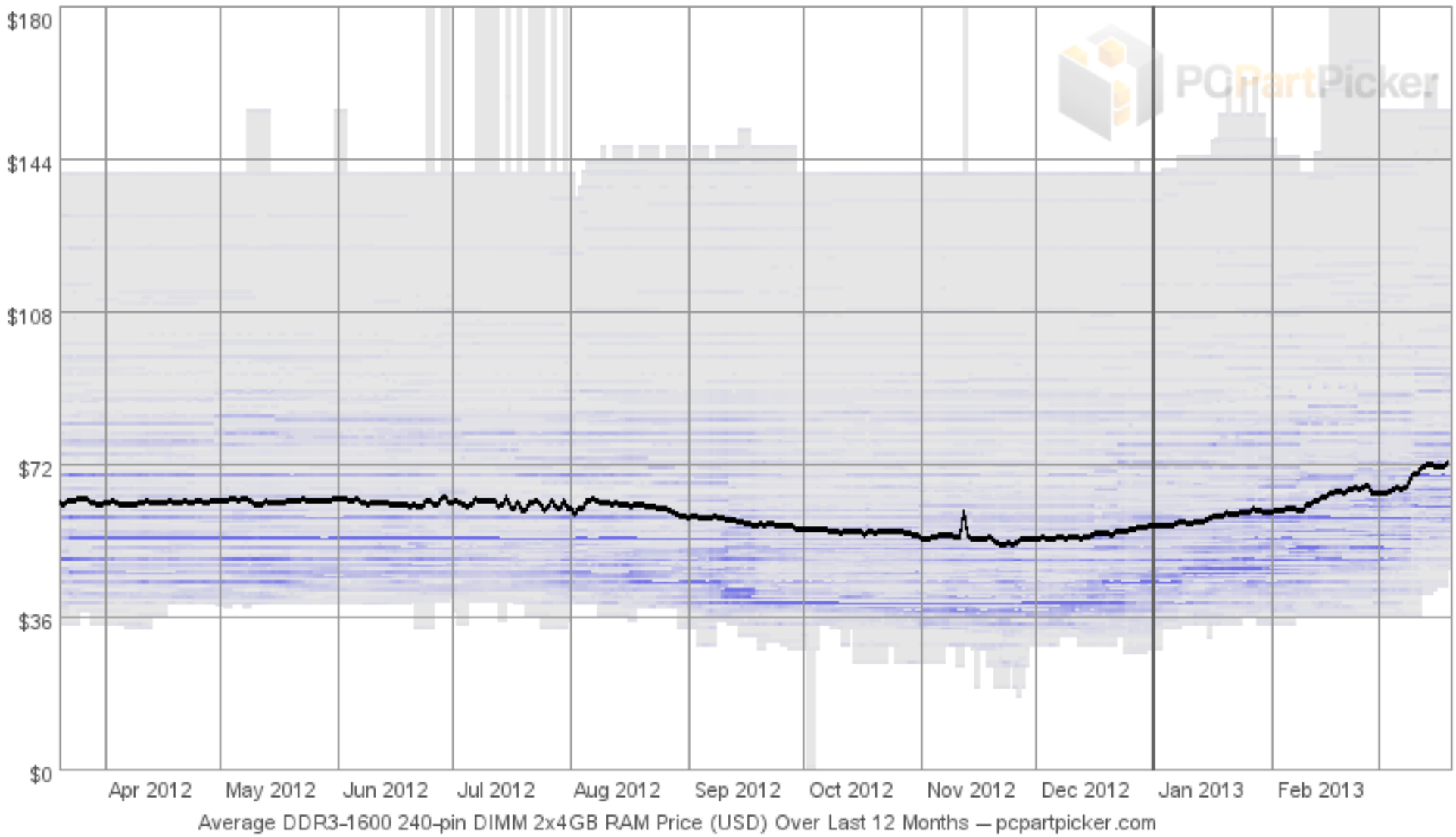
http://blog.pythonisito.com/2011/12/mongodbs-write-lock.html



http://blog.pythonisito.com/2011/12/mongodbs-write-lock.html

Performance

More RAM = expensive



x2 4GB RAM 12 month Prices

RAM

SSDs

Spinning disk

Cost

Speed

Performance

Softlayer disk pricing

Unit	Monthly
32GB SSD	$25.00
50GB SSD	$30.00
64GB SSD	$40.00
100GB SSD	$100.00
200GB SSD	$125.00
400GB SSD	$200.00
800GB SSD	$300.00

Unit	Monthly
250GB - SATA II Hard Drive	$20.00
500GB - SATA II Hard Drive	$30.00
750GB - SATA II Hard Drive	$40.00
1.00TB - SATA II Hard Drive	$50.00
2.00TB - SATA II Hard Drive	$60.00
3.00TB - SATA III Hard Drive	$80.00
4.00TB - SATA III Hard Drive	$100.00

Unit	Monthly
73GB SA-SCSI 10K Hard Drive	$30.00
73GB SA-SCSI 15K Hard Drive	$50.00
147GB SA-SCSI 10K Hard Drive	$50.00
147GB SA-SCSI 15K Hard Drive	$75.00
300GB SA-SCSI 10K Hard Drive	$75.00
300GB SA-SCSI 15K Hard Drive	$100.00
450GB SA-SCSI 15K Hard Drive	$125.00
600GB SA-SCSI 15K Hard Drive	$150.00

Performance

EC2 disk/RAM pricing

Standard On-Demand Instances	
Small (Default)	$0.060 per Hour
Medium	$0.120 per Hour
Large	$0.240 per Hour
Extra Large	$0.480 per Hour

$43/m

High-Memory On-Demand Instances	
Extra Large	$0.410 per Hour
Double Extra Large	$0.820 per Hour
Quadruple Extra Large	$1.640 per Hour

$295/m

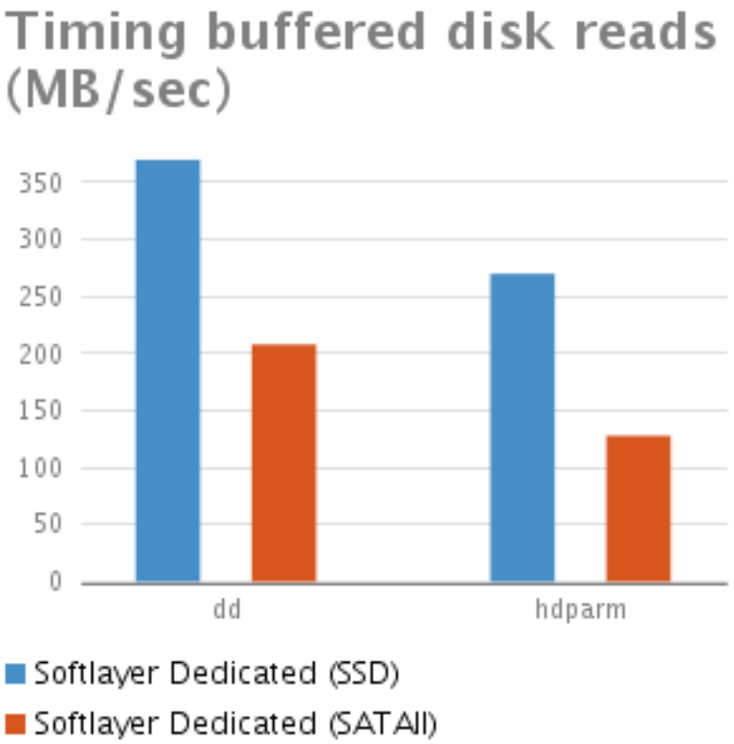
High-Memory Cluster On-Demand Instances	
Eight Extra Large	$3.500 per Hour
Cluster GPU Instances	
Quadruple Extra Large	$2.100 per Hour
High-I/O On-Demand Instances	
Quadruple Extra Large	$3.100 per Hour

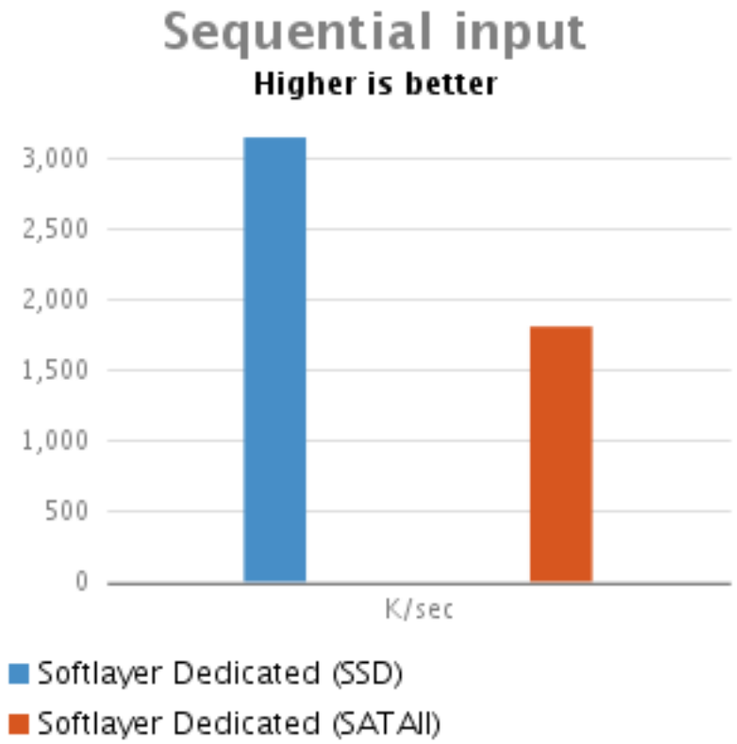
$2520/m

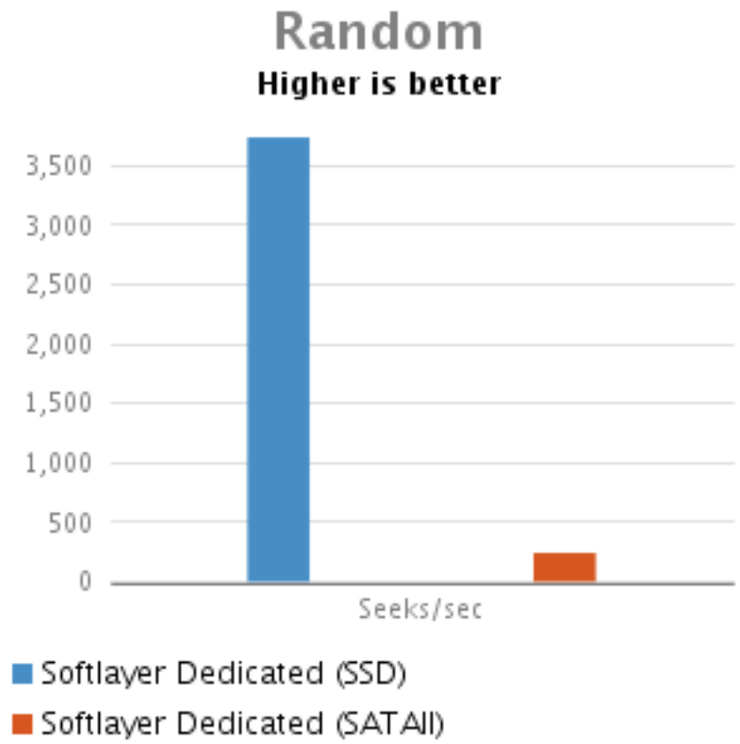
$2232/m

Performance

SSD vs Spinning

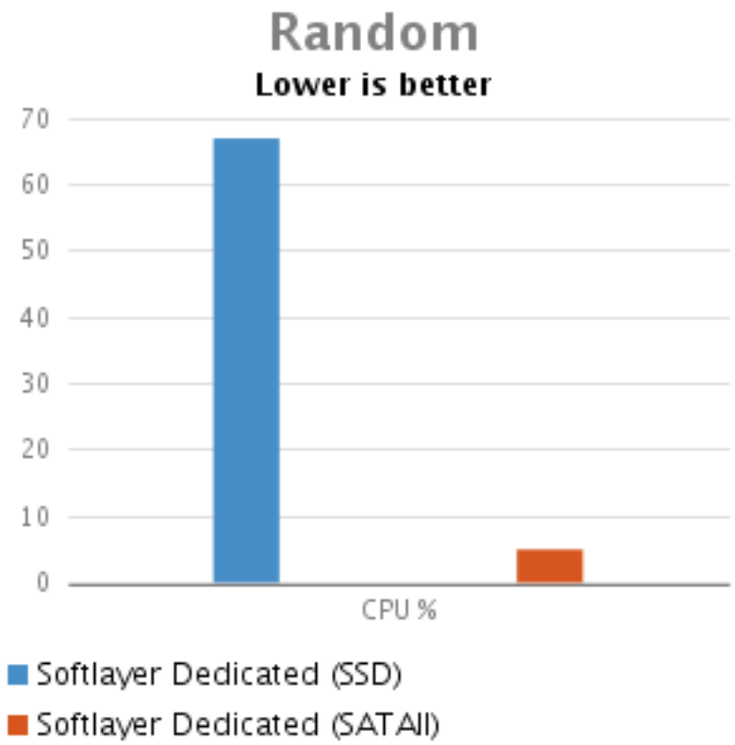


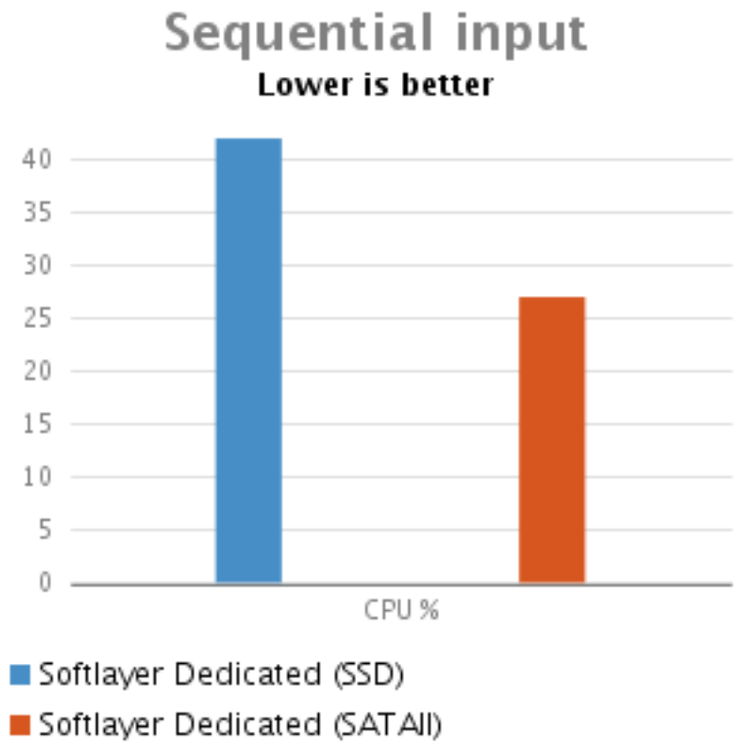




Performance

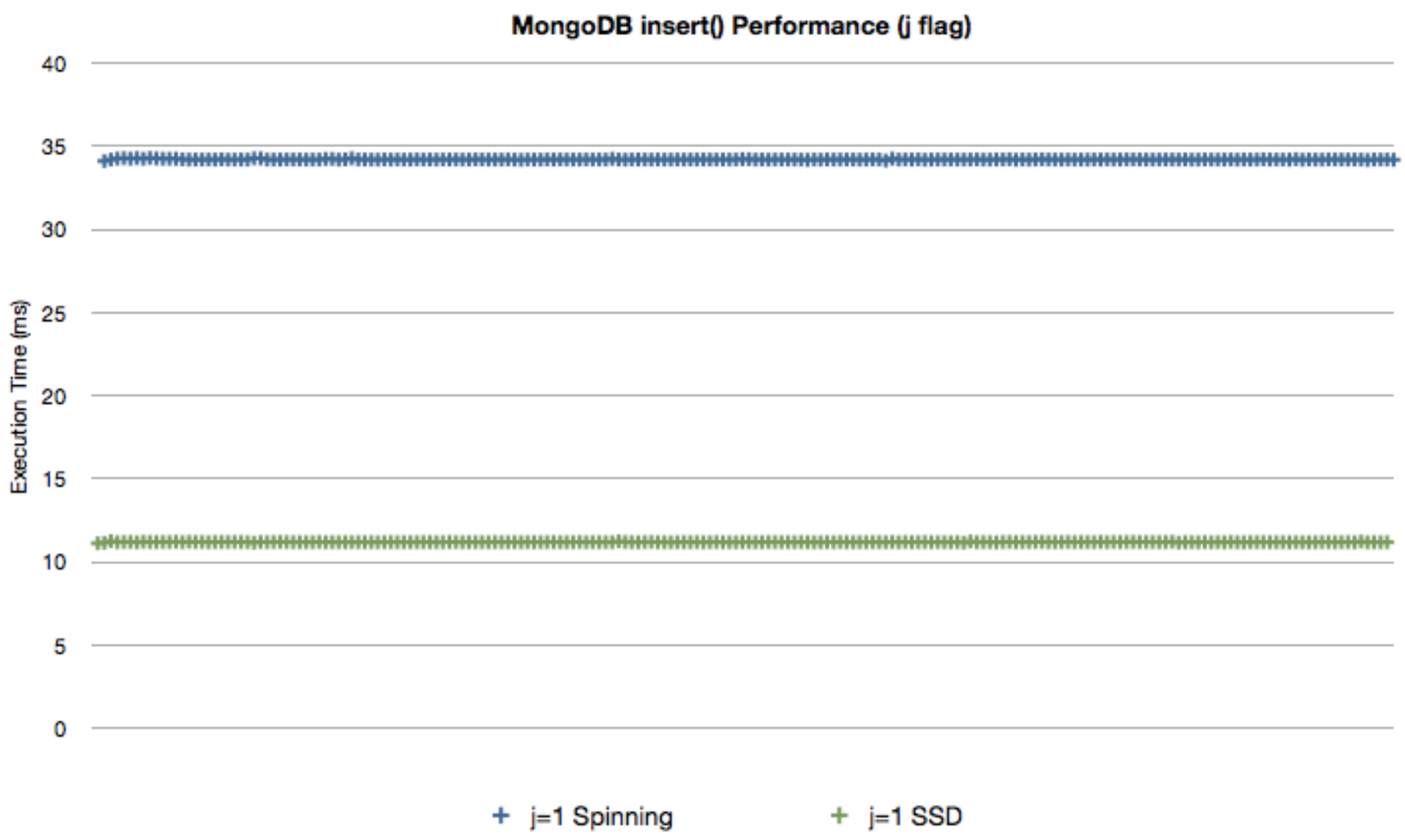
SSD vs Spinning





Performance

SSD vs Spinning



Tips: rand()

- Field names

Tips: rand()

- Field names
- Covered indexes

Tips: rand()

- Field names
- Covered indexes
- Collections / databases

Backups

What is the use case?

Backups

What is the use case?

Fixing user errors?

Point in time restore?

Disaster recovery?

Backups

- Disaster recovery

Offsite

ど せきりゅう き けん けいりゅう

井田大川水系井田大川

土石流が発生する恐れがありますので
大雨の時などは十分注意して下さい。

静岡県
Shizuoka Pref

Attention

Because debris flows are likely to occur, please take great caution in the event of a heavy rain.

注 意

由于有可能发生泥石流、在下大雨等时、请充分注意。

Atenção

Devido a possibilidade da ocorrência de inundações e avalanches, em dias de chuvas torrenciais tome maiores precauções.

주 목

산사태가 발생할 우려가 있으니 큰 비가을 때 등은 충분히조심하십시오.

Backups

- Disaster recovery

Offsite

Age

ど　せきりゅう　き　けん　けいりゅう

井田大川水系井田大川

土石流が発生する恐れがありますので
大雨の時などは十分注意して下さい。

静岡県
Shizuoka Pref

Attention
Because debris flows are likely to occur, please take great caution in the event of a heavy rain.

注 意
由于有可能发生泥石流、在下大雨等时、请充分注意。

Atenção
Devido a possibilidade da ocorrência de inundações e avalanches, em dias de chuvas torrenciais tome maiores precauções.

주 목
산사태가 발생할 우려가 있으니 큰 비가을 때 등은 충분히조심하십시오.

Backups

- Disaster recovery

Offsite

Age

Restore time

Backups

Frequency

Consistency

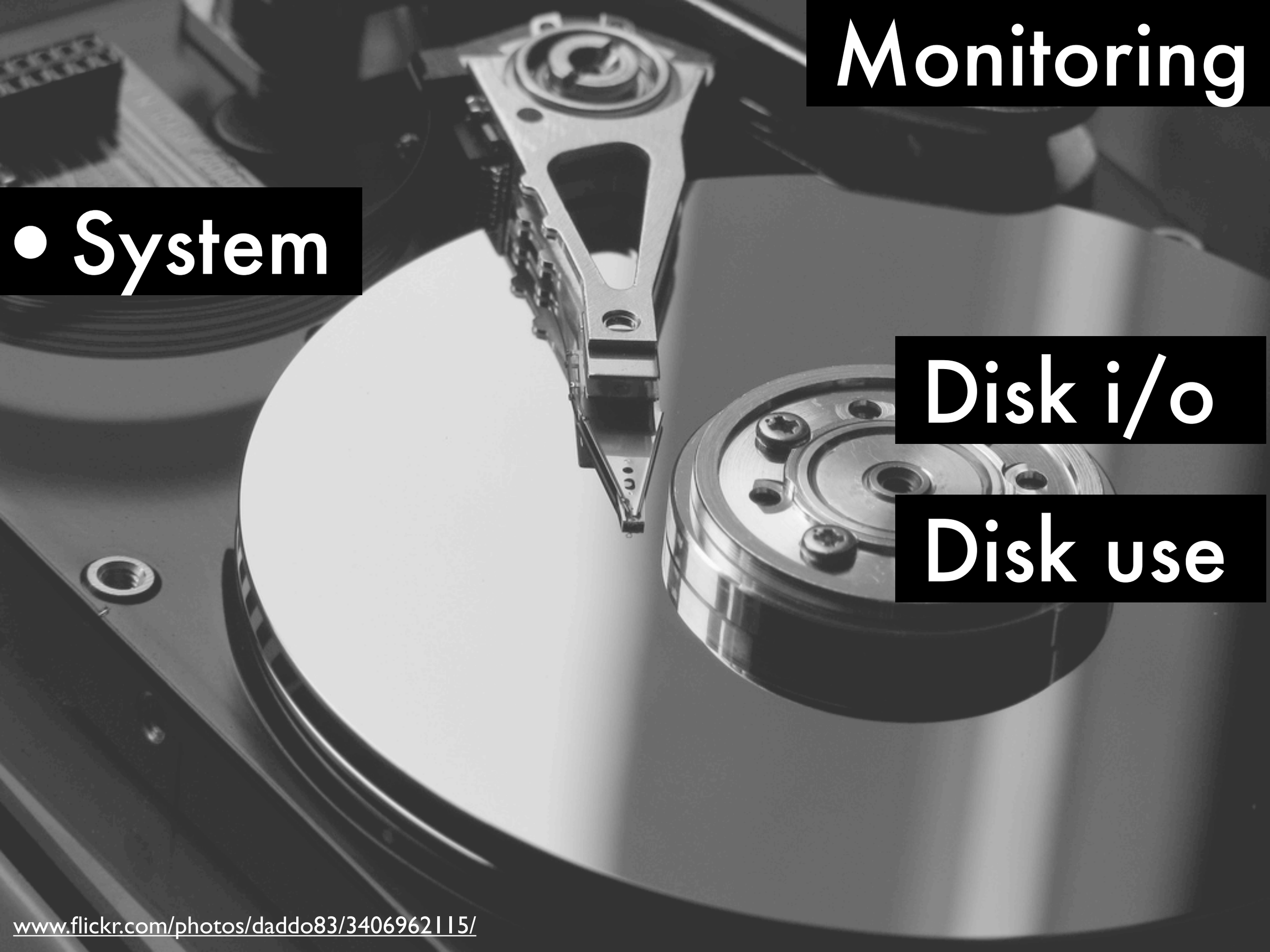
Verification

Monitoring

- System

Disk i/o

Disk use



www.flickr.com/photos/daddo83/3406962115/

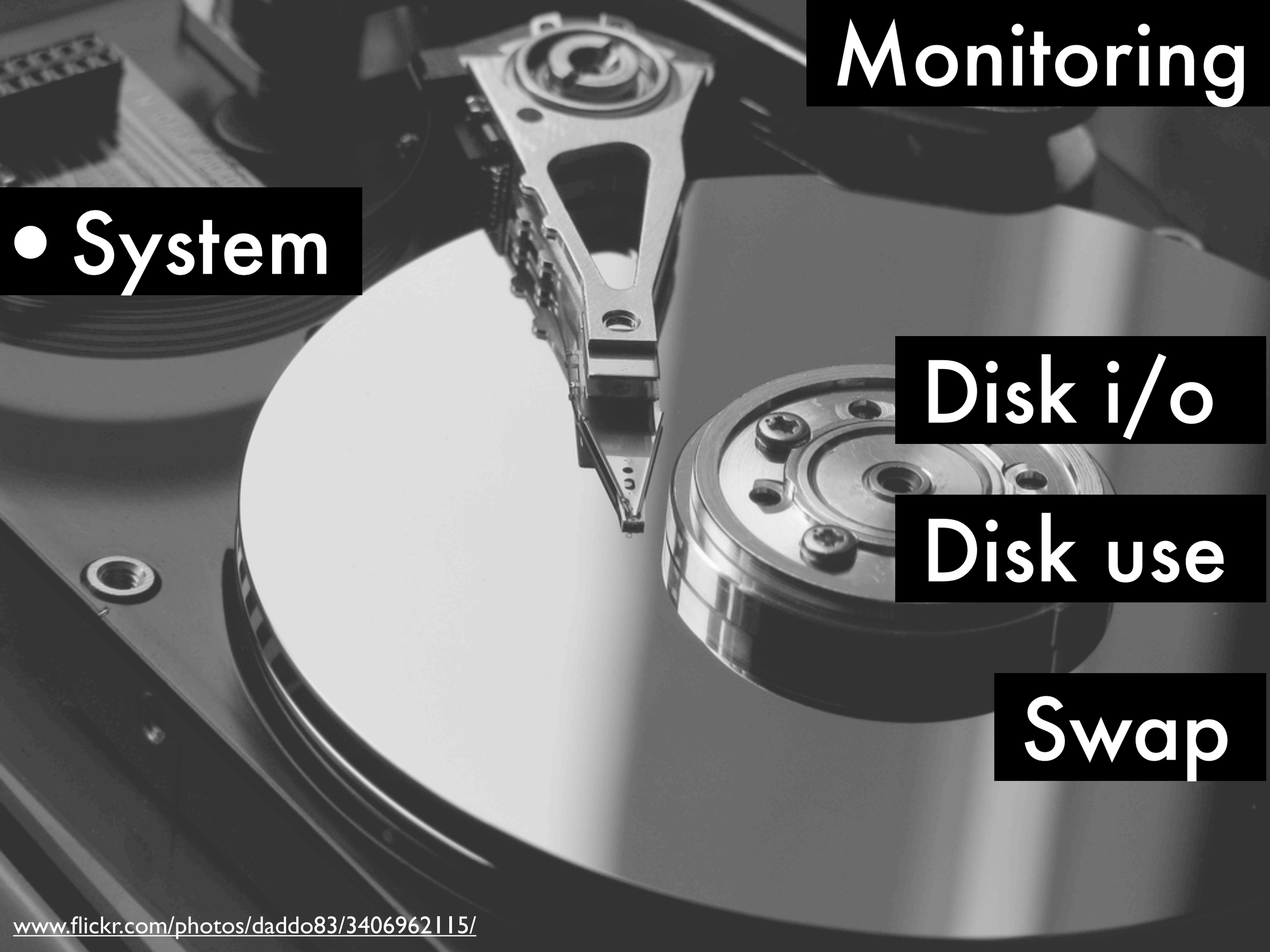
Monitoring

- System

Disk i/o

Disk use

Swap



www.flickr.com/photos/daddo83/3406962115/

Monitoring

- Replication

Optime

State

www.flickr.com/photos/daddo83/3406962115/

mongostat

```
                insert  query update delete getmore command flushes mapped  vsize    res faults    locked db idx miss %     qr|qw   ar|aw  netIn netOut  conn  set repl
mtx-md1a             5      1     55      0       5      10       0   116g   234g  14.1g      0 metrics:2.1%          0      27|2     0|2    30k    70k   106 mtx1  PRI
mtx-md2a            67      0    335      0      16      11       0   124g   250g  13.8g      2 metrics:16.9%         0       0|0     0|0   284k   460k   144 mtx2  PRI
mtx-md3a           129      0    529      0      20      11       0   122g   245g  14.1g      2 metrics:28.9%         0       0|0     0|0   429k   632k   137 mtx3  PRI
mtx-md4a           188      0    528      0      22       3       0   120g   241g    14g      1 metrics:41.6%         0       0|0     0|1   773k     1m   129 mtx4  PRI

mtx-md1a           550      5    996      0       9     365       1   116g   234g  14.1g      2       .:93.1%         0       0|0     0|0     1m     1m   106 mtx1  PRI
mtx-md2a           195      4    697      0      19      21       0   124g   250g  13.8g      1 metrics:40.4%         0       0|0     0|1   645k   902k   144 mtx2  PRI
mtx-md3a           225      0    723      0      23       6       0   122g   245g  14.1g      2 metrics:48.2%         0       0|0     0|1   798k     1m   137 mtx3  PRI
mtx-md4a           162      1    597      0      23      14       0   120g   241g    14g      0 metrics:37.3%         0       0|0     0|0   584k   833k   129 mtx4  PRI

mtx-md1a           236      2    522      0      22      89       0   116g   234g  14.1g      1 metrics:29.9%         0       7|0     0|0   499k   660k   106 mtx1  PRI
mtx-md2a           146      0    694      0      21      14       0   124g   250g  13.8g      2 metrics:45.2%         0       0|0     0|1   668k   926k   144 mtx2  PRI
mtx-md3a           122      0    502      0      22      16       0   122g   245g  14.1g      0 metrics:29.3%         0       0|0     0|0   516k   715k   137 mtx3  PRI
mtx-md4a           202      1    558      0      16      15       0   120g   241g    14g      0 metrics:36.7%         0       0|0     0|0   632k   869k   129 mtx4  PRI
```

Monitoring tools

Nagios®

Run yourself

MUNIN

Ganglia

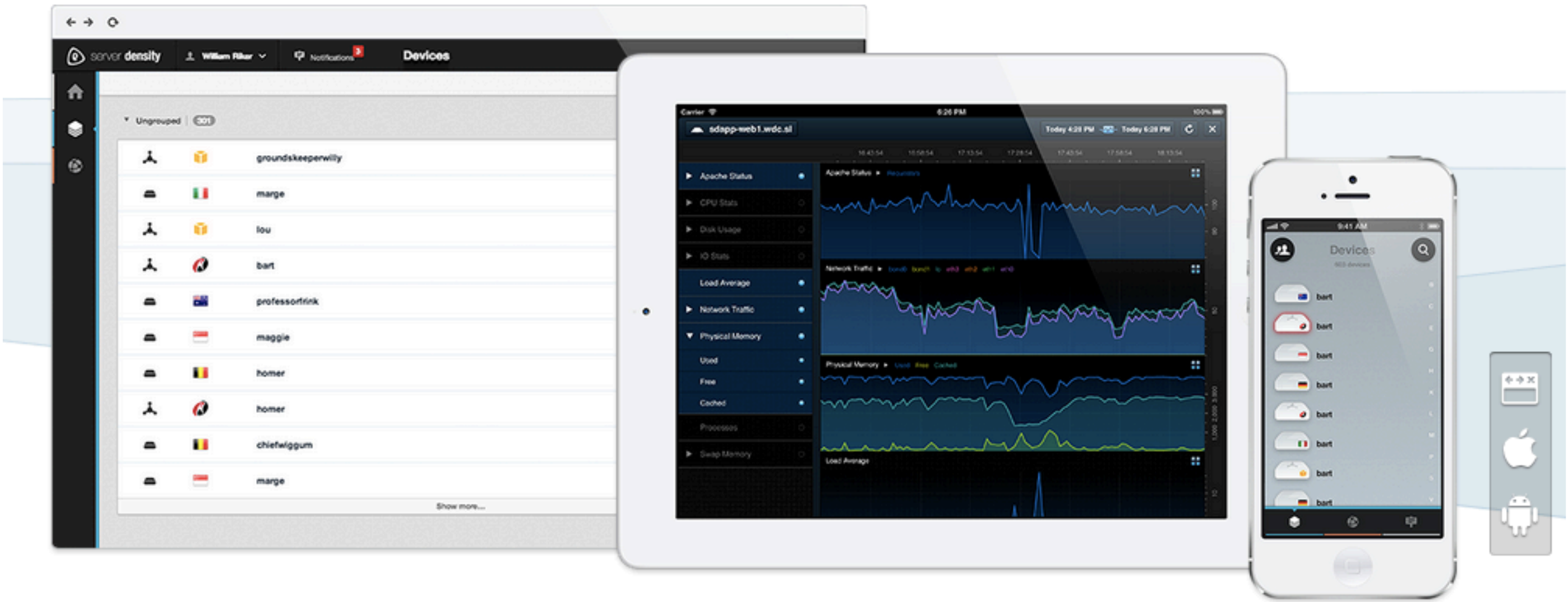
Monitoring tools

server density

cloud management

server monitoring

website monitoring



www.serverdensity.com/nosql

www.serverdensity.com/nosql

David Mytton

@davidmytton

david@serverdensity.com

www.serverdensity.com/nosql

Woop Japan!