Corey Osman
Logic Minds Corp
corey@logicminds.biz

# ENTERPRISE SSD

Boost your infrastructure's performance with SSD!

LOGIC Minds

# I.T. Consulting Services

- Remote Sys Admin Services
- Datacenter Automation Services
- Green I.T. Services

# Preface

- Only tested with RHEL 4 and OS X
- Did not use Solaris or ZFS
- Some numbers are approximations
- Always be sure to do your own research and due diligence before implementing any solution

# The IO Bottleneck Problem

- Oracle DB was saturating a RAID 1+0 of 6 SAS drives (3G) during peak
- IO wait times averaged10-24ms
- Disk Percent Usage reached 100%
- Memory/CPU was never an issue

IOSTAT was used to measure these metrics

# Possible Solutions

- Add expensive SSD
- Add Tray of Disks and bigger controller
- Add 2nd server
- Use different filesystem (XFS, ASM)
- Use direct access
- Tune filesystem

# Getting your Team on Board

- Consider the lifetime of card after careful calculation
- Concentrate on latency factor and increased performance
- Consider a hybrid approach with current disk infrastructure if cost is a factor
  - Flash cache
  - Separate data

# How Fast is SSD?

- Let’s do a quick test…

# Data Usage

- Oracle DBA said 100GB for daily change write
- Actual usage is approximately 2 TB daily
- SSD card went from lasting 109 years to 5.47 years

# Power Utilization

- Fusion IO card
  - Min 5W – Max 25W
- Intel 510 SSD
  - 380 mW (active)
  - 100 mW (idle)

- Consider a PCI solution vs. 6 SSD drives and Raid controller

# SSD vs. Magnetic

| Metric | SSD | 15K SAS Magnetic |
|---|---|---|
| Latency | 26 microseconds | 2000 microsecs |
| Price | $2 - $20 / GB | $0.05 - $0.20 / GB |
| Bandwidth | 250 – 800 MB/s | 200 MB/s |
| IO | 30K – 511K IOPS | 210 IOPS |
| Seek time | 26 microseconds | 3400 microsecs |

# SSD vs. Magnetic: Differences

- No moving parts
- Lower power usage by 33%-50%
- Not affected by Magnets as much
- Potentially longer life span
- Much higher IO rate
- Low latency
- SSD doesn’t get carsick (provided it’s removed from PCI slot)

# SSD vs. Magnetic:  Similarities

- SSD and Magnetic media do not perform well in excessive heat (over 170 degrees)
- Both have drivers to configure (PCI version)
- Both can be monitored via snmp or S.M.A.R.T.

# Enterprise vs. Consumer

- Wear limit
- Number of sensors
- Warranty period
- Auto shutdown on heat issues
- Read-only mode after wear limit

# HP SSD

| Size | Type | Cycles | Cost | Time | RW |
|---|---|---|---|---|---|
| 200GB | SLC | 21 PB | $4,800 | 12.9 years | 350/160 |
| 400GB | SLC | 42 PB | $6,200 | 20.1 years | 415/180 |
| 200 GB | MLC | 2.23 PB | $2,400 | 3.4 years | 320/100 |
| 400 GB | MLC | 4.5 PB | $4,200 | 4.7 years | 310/110 |
| 800 GB | MLC | 6.8 PB | $11,000 | 6.8 years | 400/130 |

No block size was given for these calculations

# Value Enterprise SSD

- Axium Memory  - difficult to see the value

Block size at 128K

| Size | Type | Cycles | Cost | Time | RW |
|---|---|---|---|---|---|
| 60 GB | MLC | 22 TB | $250 | | 250/170 |
| 120 GB | MLC | 22 TB | $450 | | 250/220 |
| 240 GB | MLC | 22 TB | $800 | | 250/220 |
| 32 GB | SLC | 1 PB | $600 | | 250/170 |

**60GB**
Up to 31K IOPS – Random Read @ 4K blocks
Up to 11K IOPS – Random Write @ 4K blocks
**120GB & 240GB**
Up to 31K IOPS – Random Read @ 4K blocks
Up to 20K IOPS – Random Write @ 4K blocks

# SSD vendors

| Vendor | Common Use |
|---|---|
| Fusion IO | Enterprise Grade |
| OCZ | Enterprise, Consumer |
| OWC | Consumer Mac |
| Intel | Consumer |

# Life expectancy

- SSD have a max cell cycle count (wear limit)
- Easy to estimate when SSD will die

```
Media status: Healthy; Reserves: 100.00%, warn at 10.00%

Lifetime data volumes:
            Physical bytes written:  52,811,538,026,560
            Physical bytes read:     51,021,930,005,880
```

# Types of Flash

- MLC
  - Cheaper
  - Older technology
- SLC
  - Expensive 2x cost
  - 2x performance
  - 10x wear limit

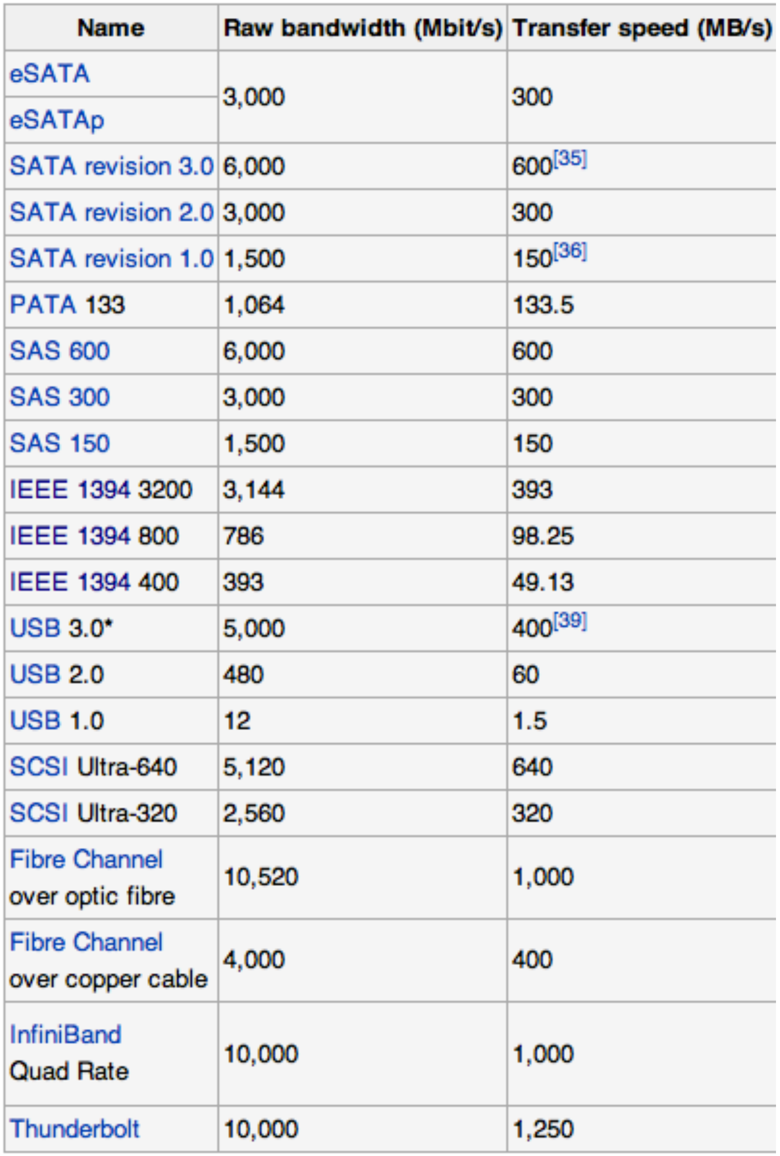
# SSD Form Factors

- PCI express
- Standard drive form factors (3.5”, 2.5”)

# Interface Bottleneck

- SSDs can easily saturate an Interface because of their speed
  - Sata 3G is not enough
  - Controllers may be a bottleneck too
    - P400 has max 2GB/s

- PCI express cards resolve the bottleneck issue
  - Require at least 4 lanes
  - Sits directly on the PCI Bus (lower latency)
  - 32 total lanes for 16GB/s total throughput



| Name | Raw bandwidth (Mbit/s) | Transfer speed (MB/s) |
|---|---|---|
| eSATA | 3,000 | 300 |
| eSATAp | | |
| SATA revision 3.0 | 6,000 | 600[35] |
| SATA revision 2.0 | 3,000 | 300 |
| SATA revision 1.0 | 1,500 | 150[36] |
| PATA 133 | 1,064 | 133.5 |
| SAS 600 | 6,000 | 600 |
| SAS 300 | 3,000 | 300 |
| SAS 150 | 1,500 | 150 |
| IEEE 1394 3200 | 3,144 | 393 |
| IEEE 1394 800 | 786 | 98.25 |
| IEEE 1394 400 | 393 | 49.13 |
| USB 3.0* | 5,000 | 400[39] |
| USB 2.0 | 480 | 60 |
| USB 1.0 | 12 | 1.5 |
| SCSI Ultra-640 | 5,120 | 640 |
| SCSI Ultra-320 | 2,560 | 320 |
| Fibre Channel over optic fibre | 10,520 | 1,000 |
| Fibre Channel over copper cable | 4,000 | 400 |
| InfiniBand Quad Rate | 10,000 | 1,000 |
| Thunderbolt | 10,000 | 1,250 |

# Reliability

- Slows down when gets hot
- Goes into read only mode when heat threshold is met
- Goes into Read-only when wear limit is reached
- No moving parts

# Fusion IO Fault Tolerance

- Status LEDs on PCI slot to show when the SSD has issues
- Sends alerts via snmp traps
- Intelligently marks bad cells
- Provides health status
- Great monitoring tools provided by Fusion IO

# Latency

- One of the biggest advantages of SSD over magnetic media
- 26 microseconds vs. 3000 microseconds
- Reaches the data 115x faster!

# Longevity

- Varies from product to product
- Measured in total bytes written or cycle counts

| Product | Limit |
|---|---|
| Fusion IO MLC | 4-8 PB |
| Fusion IO SLC | 50 PB |
| Fujitsu SSD (MLC) | 30 TB |

# Cost

| Product | Cost |
|---|---|
| Fusion IO 320 GB MLC | $7,000 |
| Fusion IO 320 GB SLC | $15,000 |
| OWC Extreme Pro 240GB | $529 |
| OCZ Revodrive 2 | $600 |
| OCZ 1TB VeloDrive PCI | $5,500 |

# Choosing the Right SSD

1. PCI or SATA
2. Consumer or Enterprise grade
3. MLC vs. SLC
4. Total cycle count
5. Supported Operating Systems
6. On board controller( sandforce)
7. Fusion IO or rebranded (HP, Dell, …)
8. Total IOPS
9. Total Bandwidth
10. Latency

# Benchmarking

- IOSTAT (part of Sysstat package)
- FIO tool (Raw Disk performance)
- Watch IOWait times
- System commands are not good to test RAW performance

# Problems with SSD

- Difficult to understand all of the differences
- Some people want to use SSD just like a magnetic disk
- Most people are quick to blame the unknown

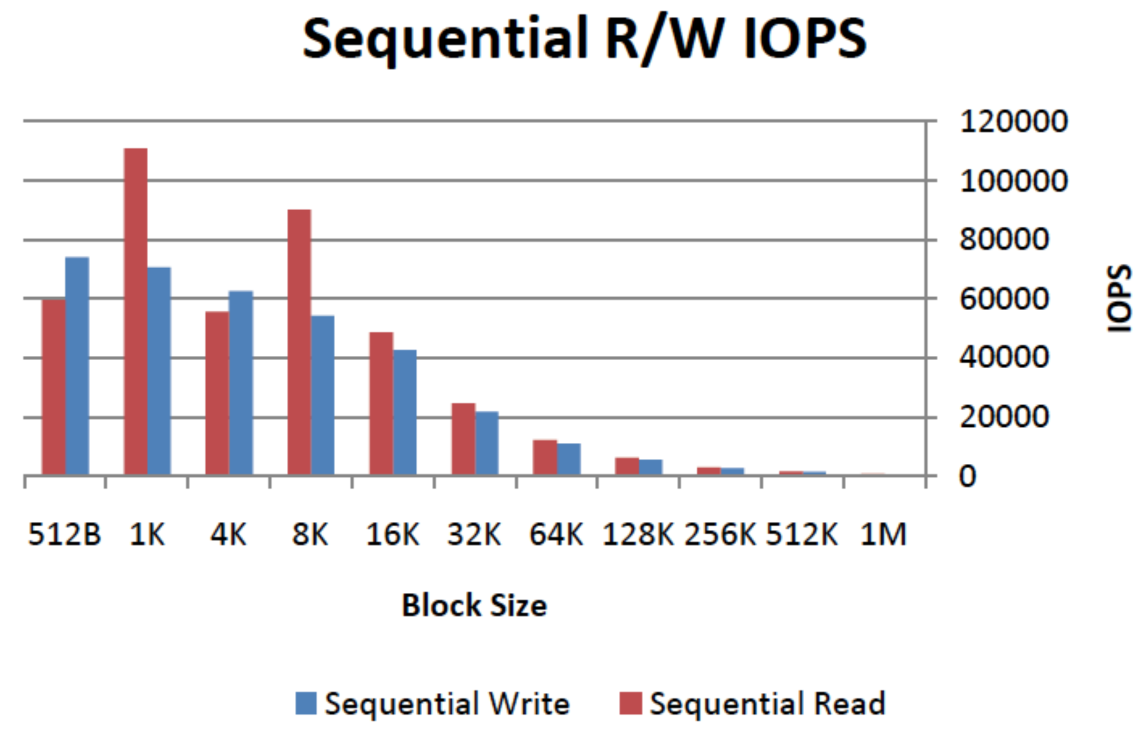
# Block Size

- Size that the filesystem will use to read and write data
- Larger block sizes will help improve disk IO performance when using large files, such as databases
- Smaller block sizes require more IOPS to push the same amount of data
- More IOPS require additional CPU, Memory

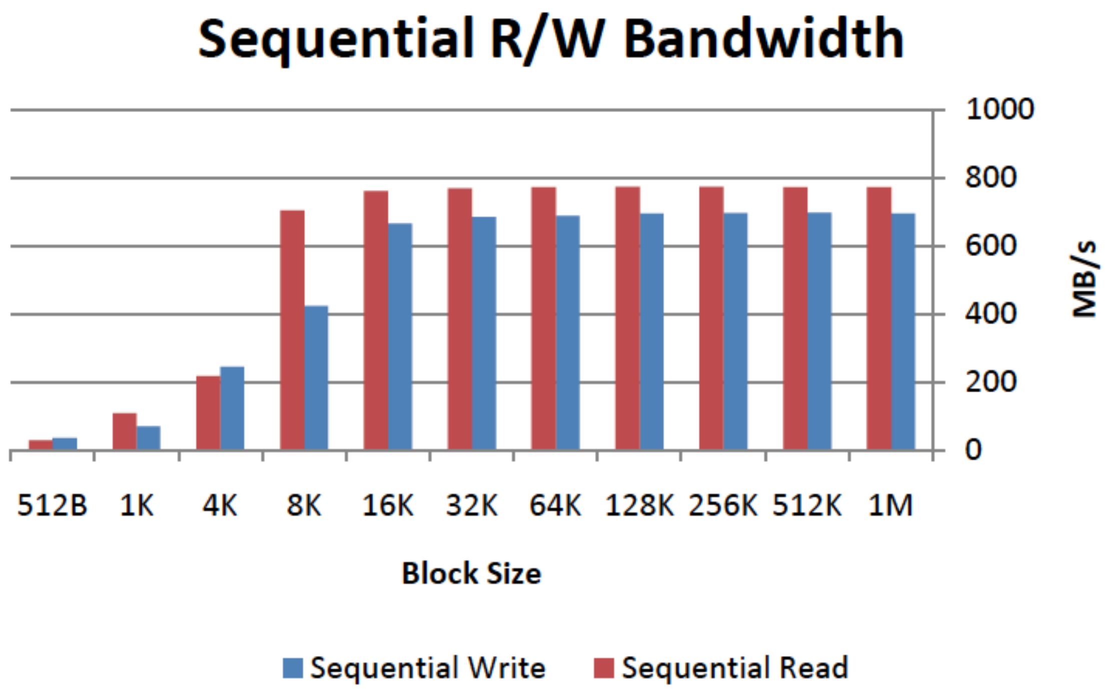
# Block Size is Important

- A SAS 15K drive can do 300MB/s too at 1MB block size
- An SSD can do 300MB/s at virtually any block size (4K plus)
- SSDs can perform great because of the high number of IOPS.  40,000 IOPS vs 300!

# IOPS Chart


Sequential R/W IOPS
120000
100000
80000
60000
40000
20000
0
IOPS
512B
1K
4K
8K
16K
32K
64K
128K
256K
512K
1M
Block Size
Sequential Write
Sequential Read

# Bandwidth Chart



# Page Cache

- Filesystem pre disk cache used to speed up access to files on disk
- Block size cannot be bigger than the disk page cache

# Filesystem Bottlenecks

- Block size can be a limiting factor
  - Max of 4K on EXT3
  - XFS gives better options
- Number of IOPS
- EXT4 allows up to 64K block size for x86_64 systems

# RAID and SSD

- Striping is good
- Redundancy decreases performance
- Redundancy removes TRIM support
- Avoid using Volume Manager
- Integrated controllers are a better option but add latency
- All drives in RAID will die at the same time

# Usage with Oracle

- Use oracle ASM for direct control of SSD drives
- Use XFS filesystem for improved performance
- Use raw devices with Oracle (within 3%)
- Use DirectIO with Oracle with EXT3
- Try different sized block sizes within Oracle
  - At least 32K

```
raw /dev/raw/raw1 /dev/sda
raw /dev/raw/raw2 /dev/md1
raw /dev/raw/raw3 /dev/vol1
```

# Need IO (not a blank check)

- The cost of a new server far exceeds the cost of an SSD card
  - Physical Server (most costly)
  - OS license
  - Oracle license
  - Shipping
  - Adds environment cost (space, power, networking, cooling)

# Problems Solved

- Low latency, IOWAIT averages less than 1ms
- Programmers can continue to write bad code (to some extent)
- No need to purchase additional server with additional licenses

# Flash Cache in front of SATA

- Cost-effective way to add high-performance storage
- Add SSD flash drive in Netapp filer
- Acts as a huge cache for SATA drives
- Can be as fast as FC 15K drives
- This method can also work in a server

# Hypervisor Use

- Greatly speed up VM storage
- Use SSD as swap storage for VMs

# Questions?

Contact:

corey@logicminds.biz

LinkedIn: Corey Osman