Oracle on SSD for Performance

FusionIO and EMC SSD performance for Oracle databases

Presenter: Steve Fluge
Bank of America

Bank of America Higher Standards

Background

- Steve Fluge
  - Oracle Engineering team at Bank of America (formerly Merrill Lynch)
    - Oracle builds, POCs for new technologies (SSD,etc), establish best practices for Oracle database architectures
  - 15 years experience with Oracle databases as DBA, consultant, engineer

Introduction

What is the compelling argument for SSD?

PERFORMANCE!

The goal:

How SSD can improve performance of Oracle databases, especially for high demand, I/O bound environments.

Bank of America Higher Standards

Introduction

- Applications which are high demand Tier-0
  - Trading applications
  - Online data services
  - Data marts
  - Etc..
- SSDs can address Tier-0 requirements for performance

SSD storage tested

- Solid State Disk (SSD) technology
  - Tier 0 storage
  - Uses NAND based flash memory
  - Available in single level cell (SLC) and multilevel cell (MLC)
    - **SLC is less dense and faster than MLC**
    - **MLC supports more capacity**
  - Capacity
    - **FusionIO cards 80G,160G,320G,640G**
    - **EMC SSD STEC 200G (tested)**

SSD storage tested

- FusionIO iodrive card
  - Connects to PCIe bus on the server
    - **Limited number of cards depend on # PCI slots available**
  - PCI configuration enables very high speed access
    - **100k and higher IOPS on reads and similar numbers for writes**
    - **Latency is in the microsecond range (vs milliseconds in HDD)**
  - Configurable reserve area
    - **Results in better performance for writes**
    - **Reduces amount of useable storage**
  - NOTE: Available on Windows 64bit, RHEL,SLES only

SSD storage tested

- FusionIO iodrive specs

ioDrive capacity	80GB	160GB	320GB
NAND Type	Single Level Cell (SLC)	Single Level Cell (SLC)	Multi Level Cell (MLC)
Write Bandwidth	500 MB/s (32K packet size)	670 MB/s (32K packet size)	490 MB/s (64K packet size)
Read Bandwidth	750 MB/s (32K packet size)	750 MB/s (32K packet size)	700 MB/s (64K packet size)
IOPS*	119,790 (4K read packet size)89,549 (75/25 r/w mix 4k packet size)	116,046 (4k read packet size)93,199 (75/25 r/w mix 4k packet size)	71,256 (4K read packet size)67,659 (75/25 r/w mix 4k packet size)
Access Latency	50µs Read	50 µs Read	80 µs Read
Operating Systems	Microsoft Windows**, Solaris 10***, RHEL 4 & 5; SLES 10 & 11	Microsoft Windows**, Solaris 10***, RHEL 4 & 5; SLES 10 & 11	Microsoft Windows**, Solaris 10***, RHEL 4 & 5; SLES 10 & 11

*160GB tested

SSD storage tested

- FusionIO drives are configurable
- Command line utilities include:
  - Fio-status : status of fio devices
  - Fio-format : format usable storage area of device
  - Fio-attach : attach device
  - Fio-detach: detach device
  - Fio-update-iodrive: update drive software

  (more)

Bank of America Higher Standards

SSD storage tested

- EMC SSD STEC ST0200
  - SAN Array based storage device
    - Symmetrix DMX4 (tested configuration)
    - Can have lots of drives up to 32 per quadrant
    - Can be used in SRDF configuration
    - Device capacity tested is 200GB

SSD storage tested

- EMC STEC specs

Zeus SSD capacity	Up to 512GB
NAND Type	SLC/MLC
Write Bandwidth	115MB/sec
Read Bandwidth	220MB/sec
IOPS*	Random Reads 45,000/sec Random Writes 16,000/sec
Transfer Rate	FC 4G/sec (dual port) SAS 3G/sec (single port) SATA 3G/sec (single port)
Interfaces	FC/SATA/SAS

*200GB FC tested

Bank of America Higher Standards

Test configuration

- Platform
  - Dell 24 core Nehalem 32GB memory, 2G HBA cards
  - Linux RHEL5
- Oracle11gR1 with ASM
- Tools
  - Oracle IO Numbers : Orion
    - **Simulates Oracle RBMS disk I/O usage and records performance data**
  - **Benchmark Factory**
    - **TPC-C and Hardware Scalability tests**

Test configuration (cont)

- Orion
  - Oracle database IO simulation tool
  - I/O workload options configured
    - Small IO size = 8kb
    - Large IO size = 1024kb
    - Storage Array simulated type = Raid 0
    - Cache size = 90 GB (SAN Array cache size accounted for)
    - Stripe depth = 1024kb
    - IO types tested = Small Random IOs, Large Random IOs
    - Write = 0 for read intensive tests, write=100 for write intensive test

Bank of America Higher Standards

Test configuration (cont)

- Benchmark Factory

Workload replay and scalability test tool from Quest includes standard industry benchmarks

- TPC-C standard benchmark test (OLTP)
- Scalable Hardware benchmark test
- Test load up to 800 concurrent users max load
  - 100GB database placed on SSD
  - ASM used with 2 disk groups

Test configuration (cont)

- Sample operational timed tests
  - RMAN backup/restore
  - Exp/Imp
  - Index create

Test Results - Orion

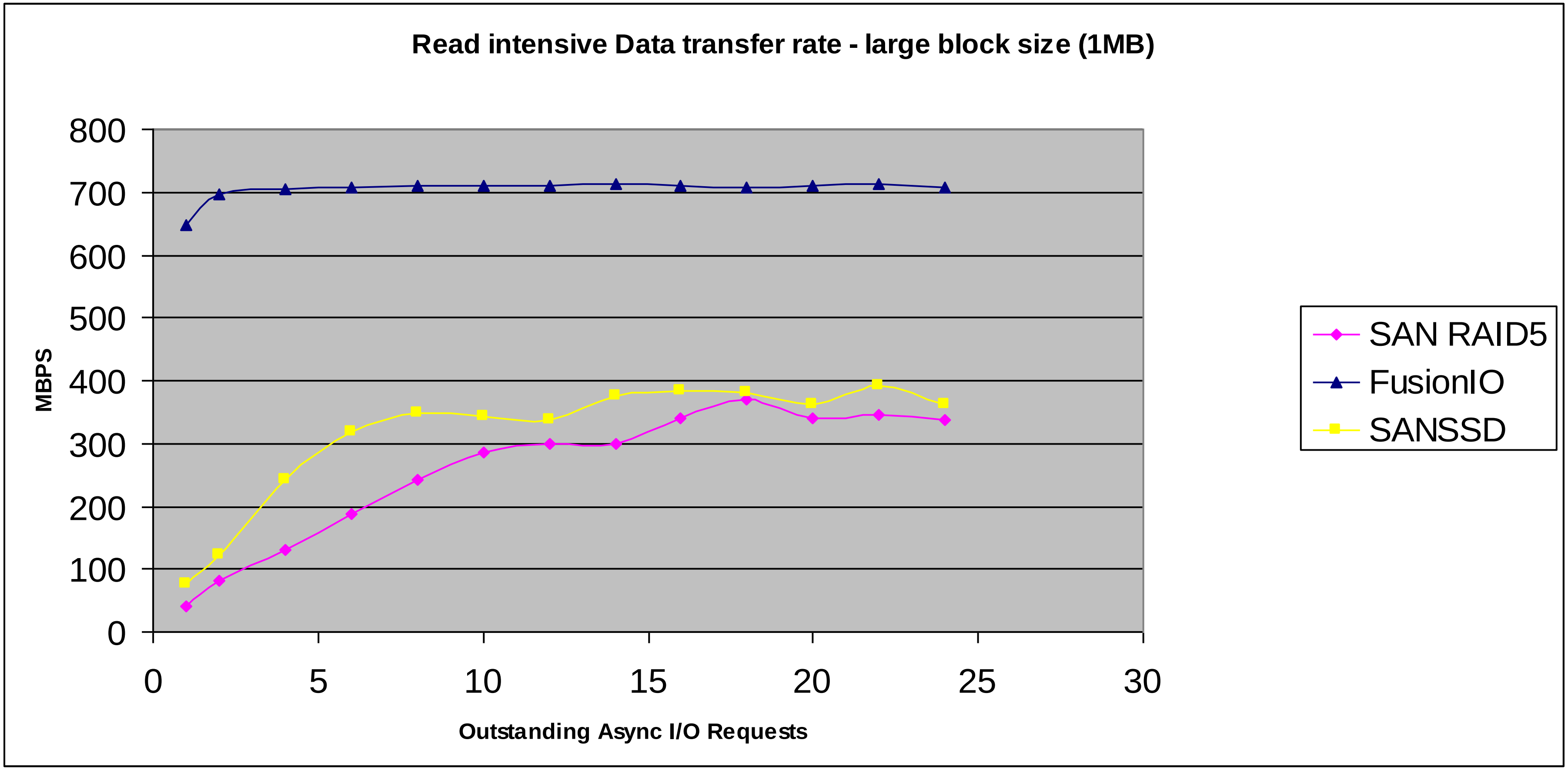
- Graphs will show significant performance increase for SSD technology compared to SAN HDD storage for data transfer and IOPS rates
- Most notable results are Read Intensive test results for large random reads.
- FusionIO cards optimized for random write performance by increasing reserve area

Bank of America Higher Standards

Read intensive test results
Large random reads MBPS

FusionIO shows highest transfer rate 700MBPS

SANSSD performance is limited due to the HBA card speed (2Gbps)

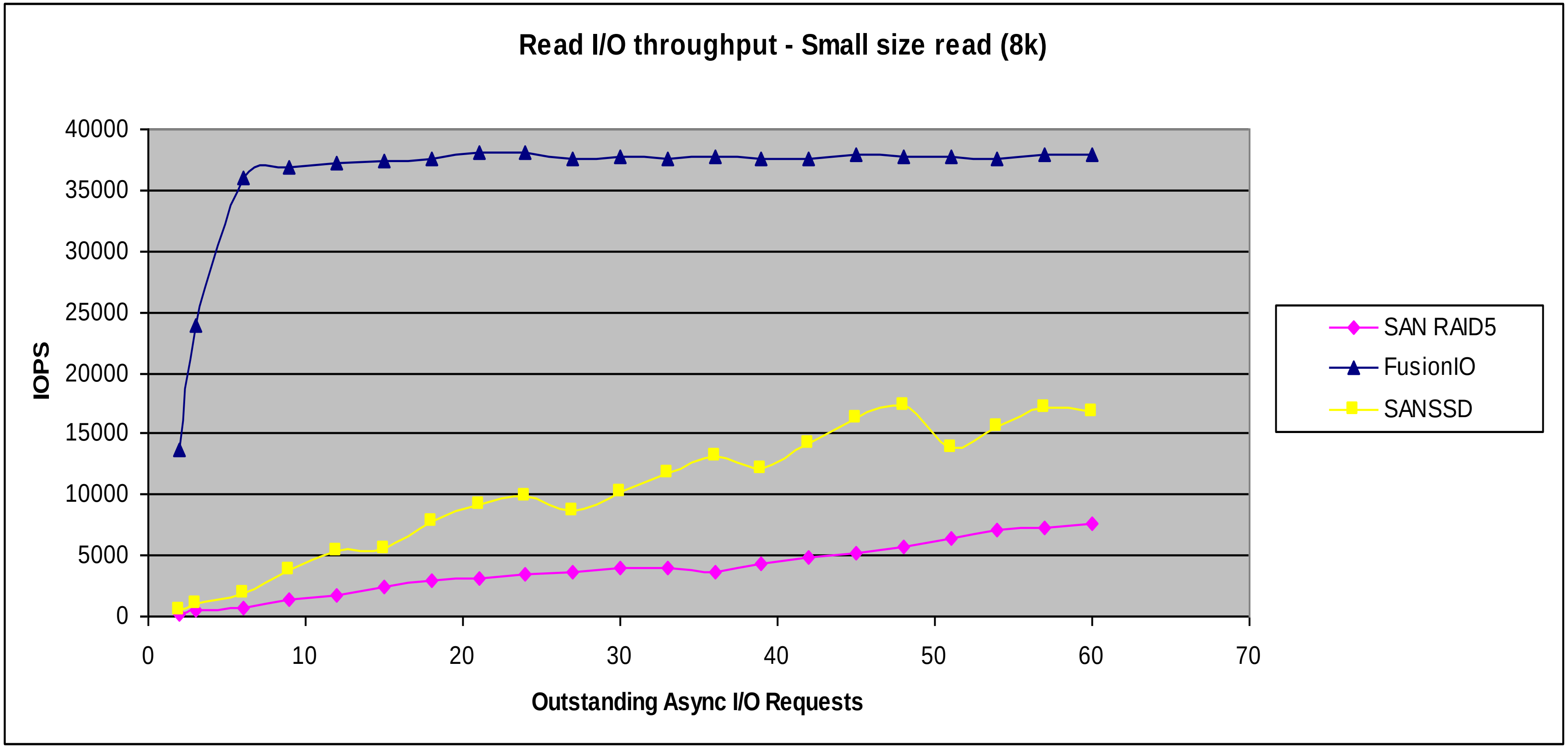


Bank of America Higher Standards

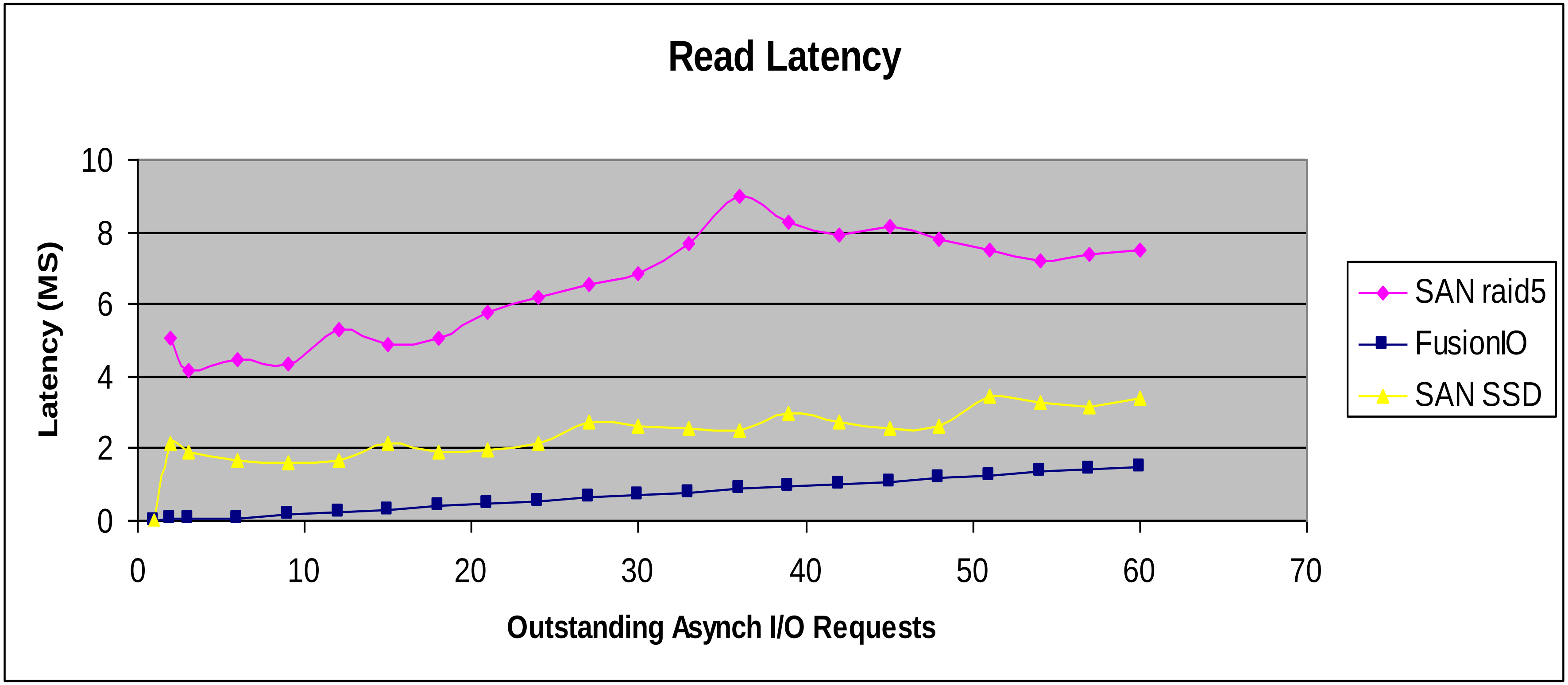
Read intensive test Results Small sequential reads IOPS

FusionIO shows highest transfer rate 39,000 IOPS

SANSSD performance is limited due to the HBA card speed (2Gbps)



Read intensive test Results
Large Random reads - Latency

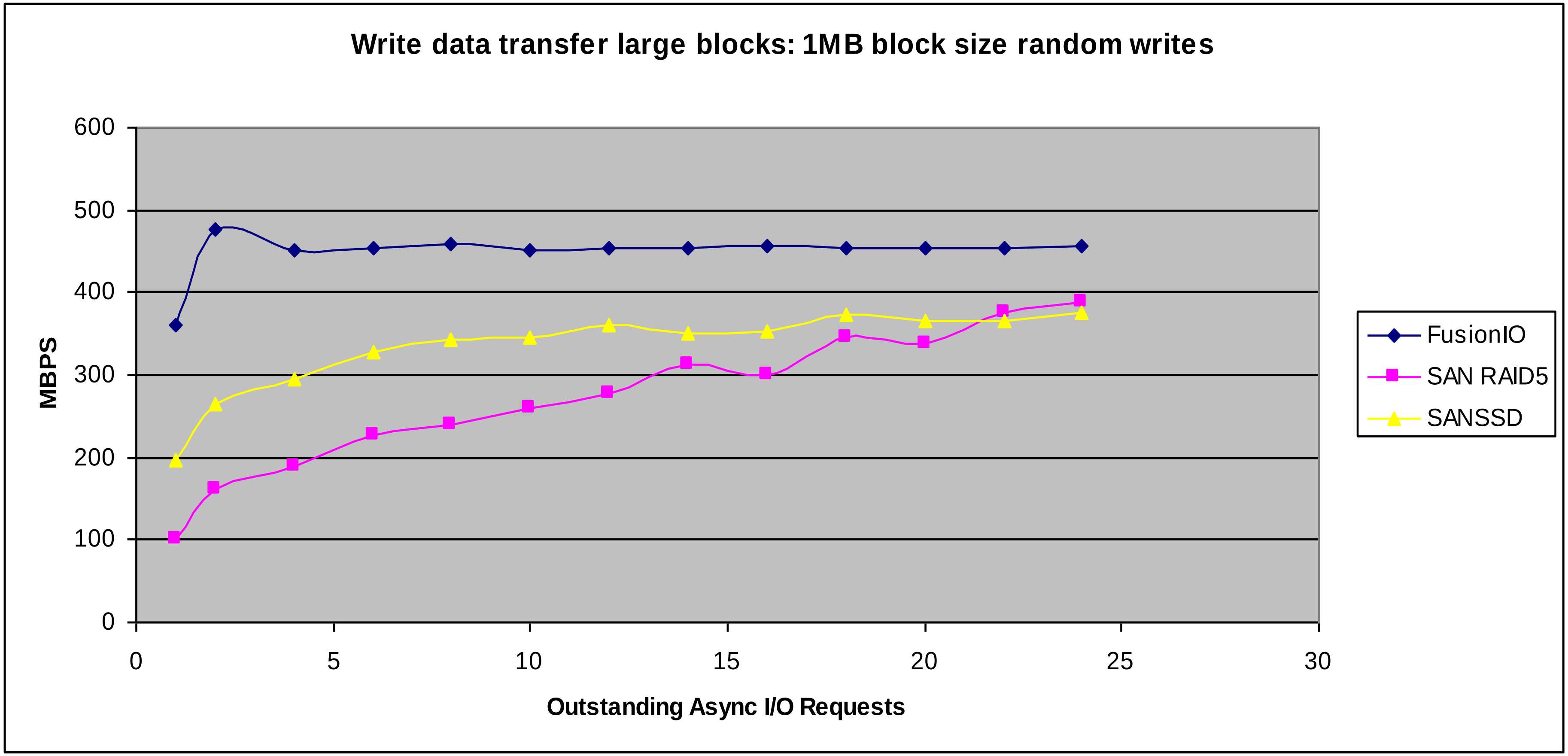


Bank of America Higher Standards

Write intensive test Results
Large random writes MBPS

FusionIO shows highest transfer rate 470MBPS

SANSSD performance is limited due to the HBA card speed (2Gbps)

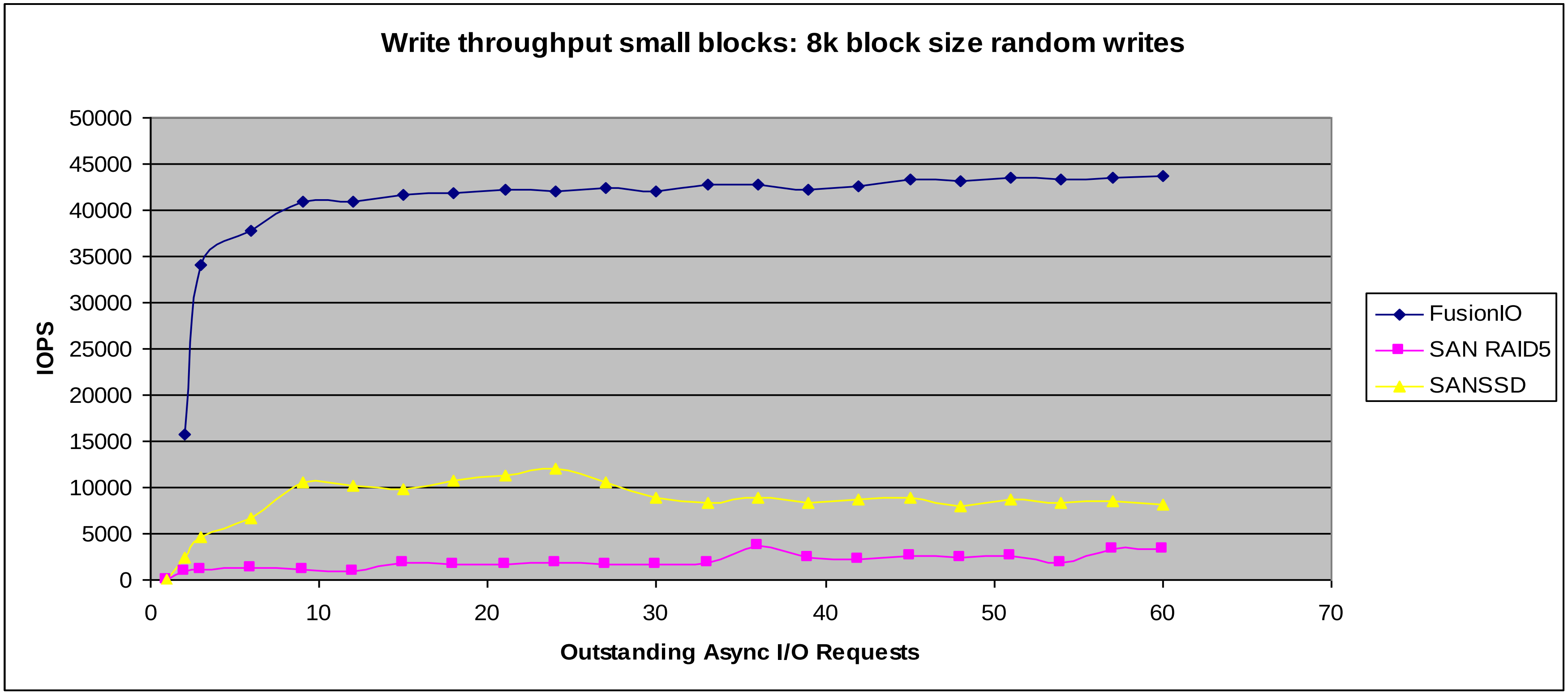


Bank of America Higher Standards

Write intensive test Results
Small random writes IOPS

FusionIO shows highest transfer rate 43,000 IOPS

SANSSD performance is limited due to the HBA card speed (2Gbps)

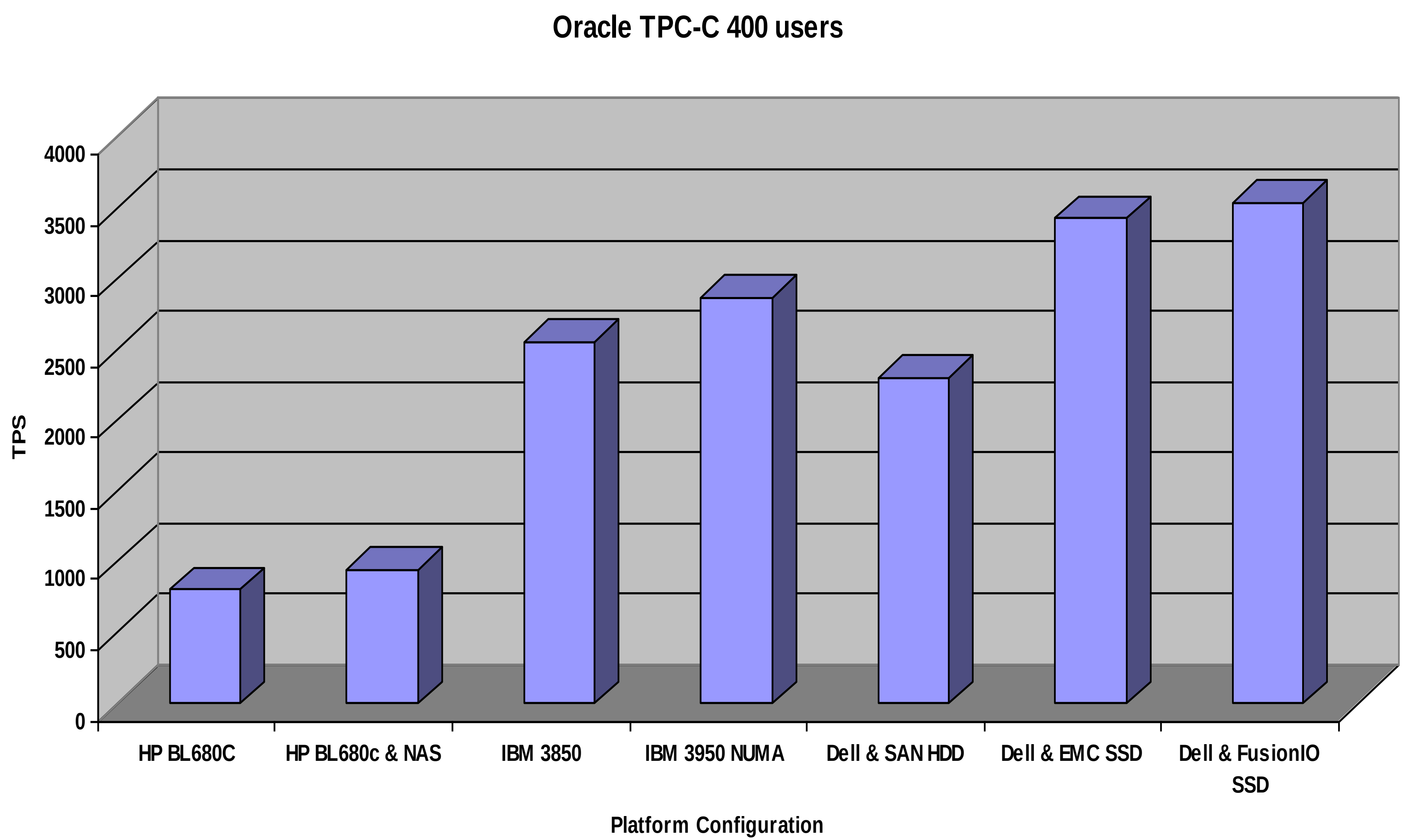


Bank of America Higher Standards

Test Results for Benchmark Factory

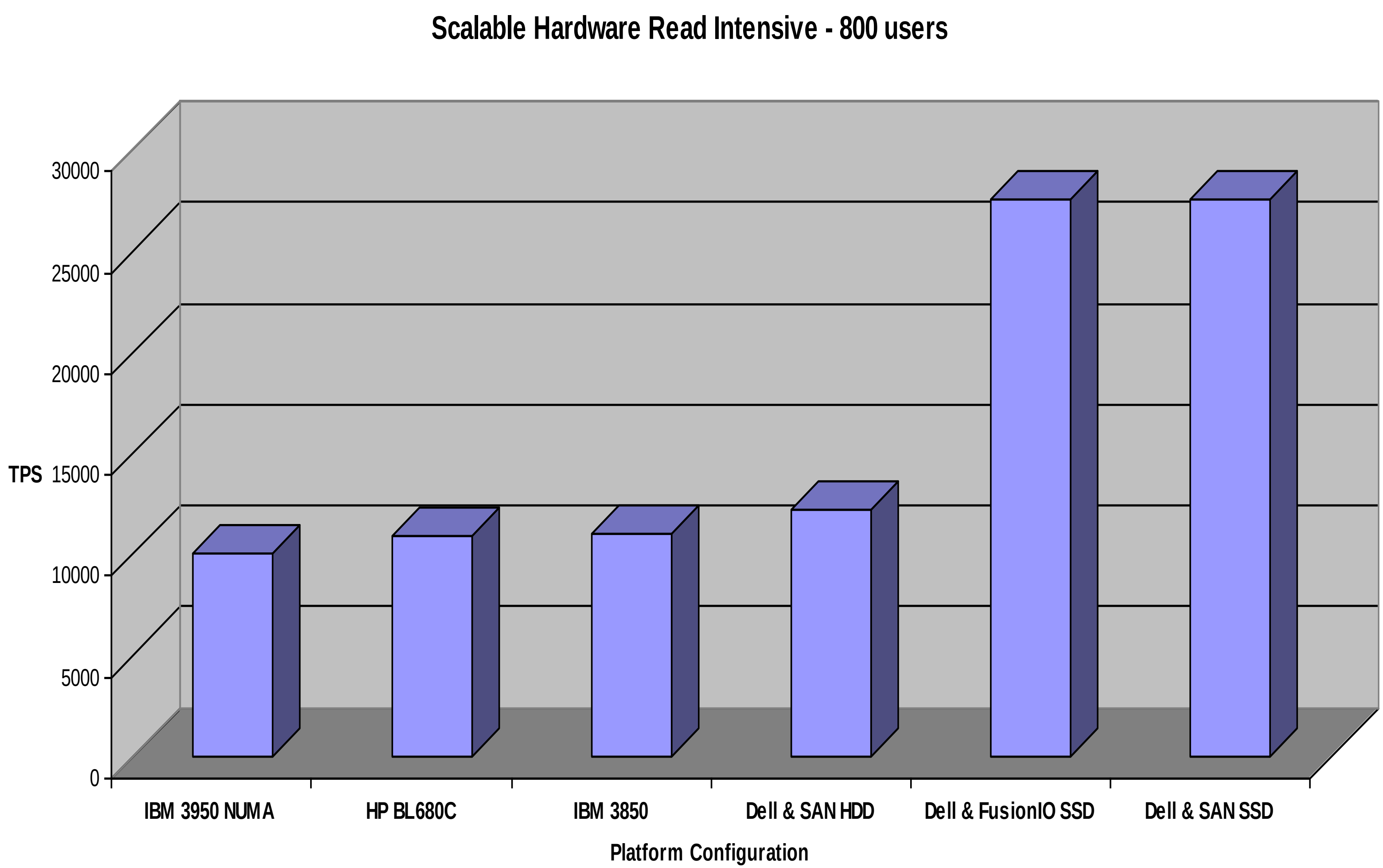
– Graphs will show major difference between SSD results and SAN HDD results
– Includes test data from prior performance tests for other platforms for comparison

Test results – Benchmark Factory TPC-C comparative results



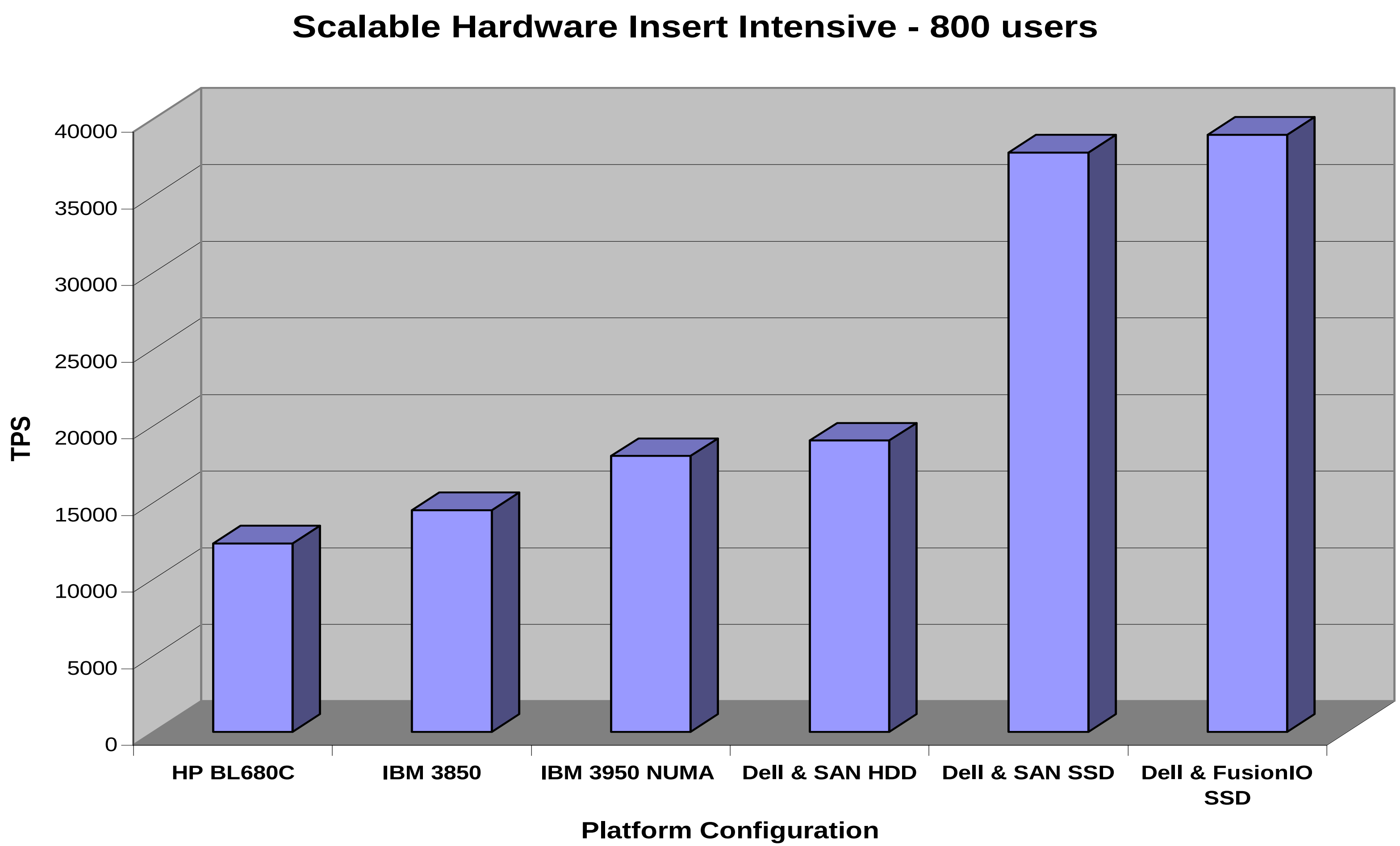
Bank of America Higher Standards

Test results – Benchmark Factory Scalable Hardware comparative results



Bank of America Higher Standards

Test results – Benchmark Factory Scalable Hardware comparative results



Bank of America Higher Standards

SSD reserve area and write operations

Question:

Why are the write operations so much slower than read operations?

Answer:

Write intensive operations need to do this when pages are updated:

1. Copy data to reserve area
2. Erase the page(s)
3. Copy the original data plus new data back to the original page(s)

Bank of America Higher Standards

Configuring FusionIO card Reserve area

- SSD devices can experience degraded write performance over time
- Most SSD devices have reserve area, typically 10%
- FusionIO SSD can be configured to increase performance of high random write activity
  - The tradeoff is less available storage

Use a command line utility to format the drive.

Steps:

1. fio-detach /dev/fct0

2. fio-format –s 100G /dev/fct0

3. fio-attach /dev/fct0

Configuring FusionIO card Reserve area

Example format session:

[root]# fio-detach /dev/fct0

Detaching: [====================] (100%) /

[root]# fio-format -s 100G /dev/fct0

WARNING: formatting will destroy any existing data on the device!

Do you wish to continue [y/n]? y

data channel: geometry: 4096x512x189056 (25 pads, 2 planes, 4 banks)

Creating a device of size 100.00GiB (107.37GB)

Formatting: [====================] (100%) -

Format successful.

[root@]# fio-attach /dev/fct0

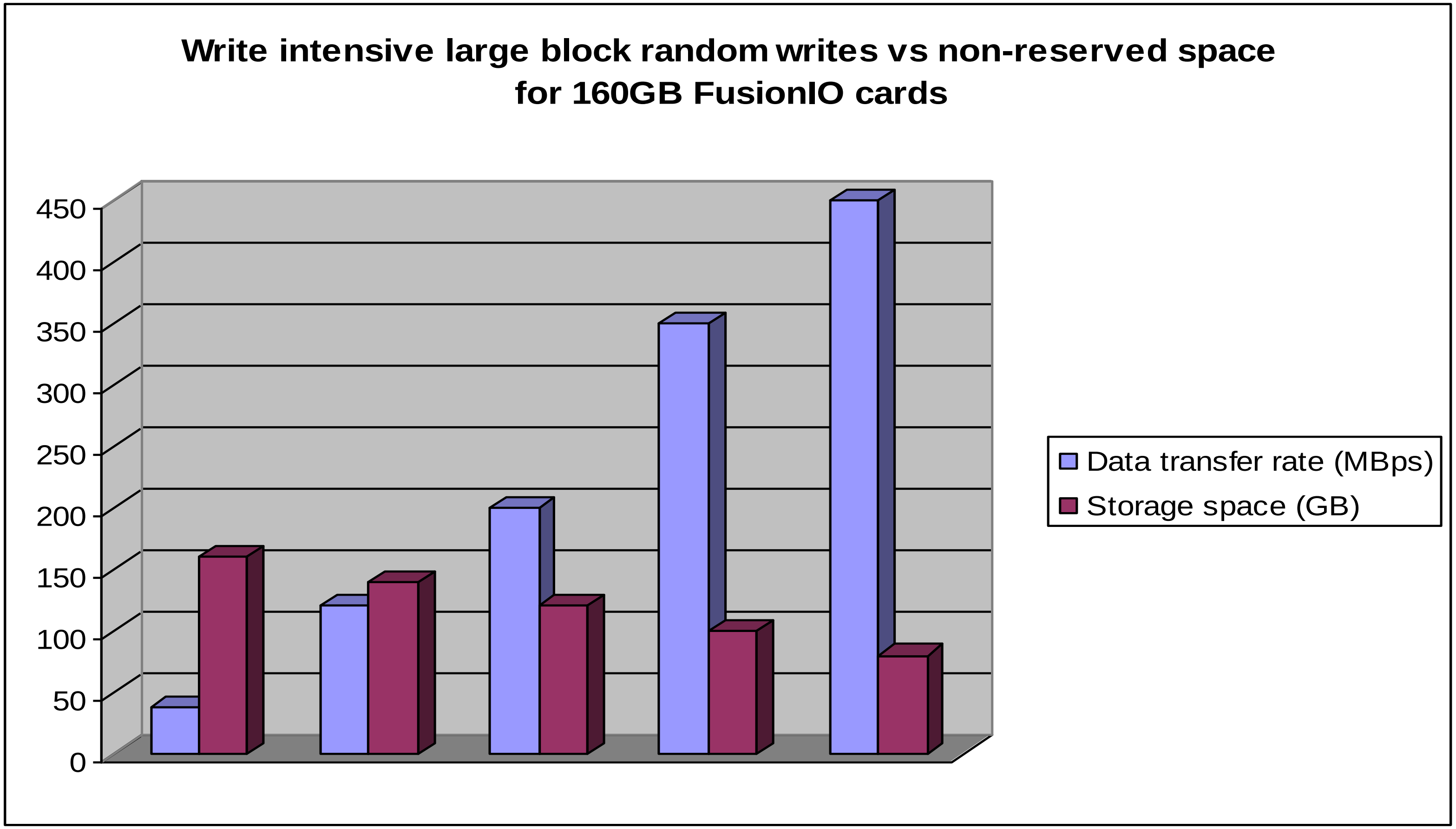
Attaching: [====================] (100%) -

fioa

Bank of America Higher Standards

Configuring FusionIO card Reserve area test results

The following chart shows the relationship between write intensive data transfer rates and the amount of storage in the available (non-reserved) area of FusionIO card (160GB card)



Bank of America Higher Standards

Test results – Operational Samples

Sample database operations tested and timed. For Export, Import, RMAN backup and restore, disk storage was used to hold the dump files and RMAN backups.

Operation	FusionIO	SANSSD	SAN RAID
Export	65minutes	70minutes	80minutes
Import	40minutes	60minutes	120minutes
Index creation	20minutes	30minutes	45minutes
RMAN backup	30minutes	40minutes	60minutes
RMAN restore	65minutes	90minutes	130minutes

Bank of America Higher Standards

Database components for SSD

Oracle databases can either be wholly on SSD or have some I/O intensive components placed there.

Use AWR report information to identify objects with the most physical reads/writes.

In addition, candidates for locating to SSD:

- Temporary tablespace
- Undo tablespace
- Redo logs
- Flash recovery area

Bank of America Higher Standards

Database components for SSD

- FusionIO tests
  - Temporary and Undo tablespaces placed on SSD showed 20% improvement in performance for index creation
  - Application Data load. Placing application schema on SSD showed 60% improvement over the original load time

Bank of America Higher Standards

Conclusions

- SSD performance makes it a prime candidate for Tier-0 storage requirements
- FusionIO is shown to perform best based on lab tests for all categories
- For write intensive databases tune FusionIO for better sustained performance (increase reserve area)
- When configuring SAN SSD consider all SAN components to avoid bottlenecks (HBA cards for example)
- Optimal configuration for SSD and Oracle should include identifying objects related to I/O bottlenecks in the database and relocating those to SSD

Bank of America Higher Standards

Conclusions

- Another key to storage in addition to performance is HA capability
  - **EMC SSD has advantages in that it is external storage which is critical in an HA architecture**
  - **FusionIO storage is local, however it can use Oracle DataGuard to address HA requirements**

Bank of America Higher Standards

Questions?

Bank of America Higher Standards

Thank you!

Steve Fluge

steve_fluge@ml.com

Bank of America Higher Standards