Design Intel SSDs Into Datacenters

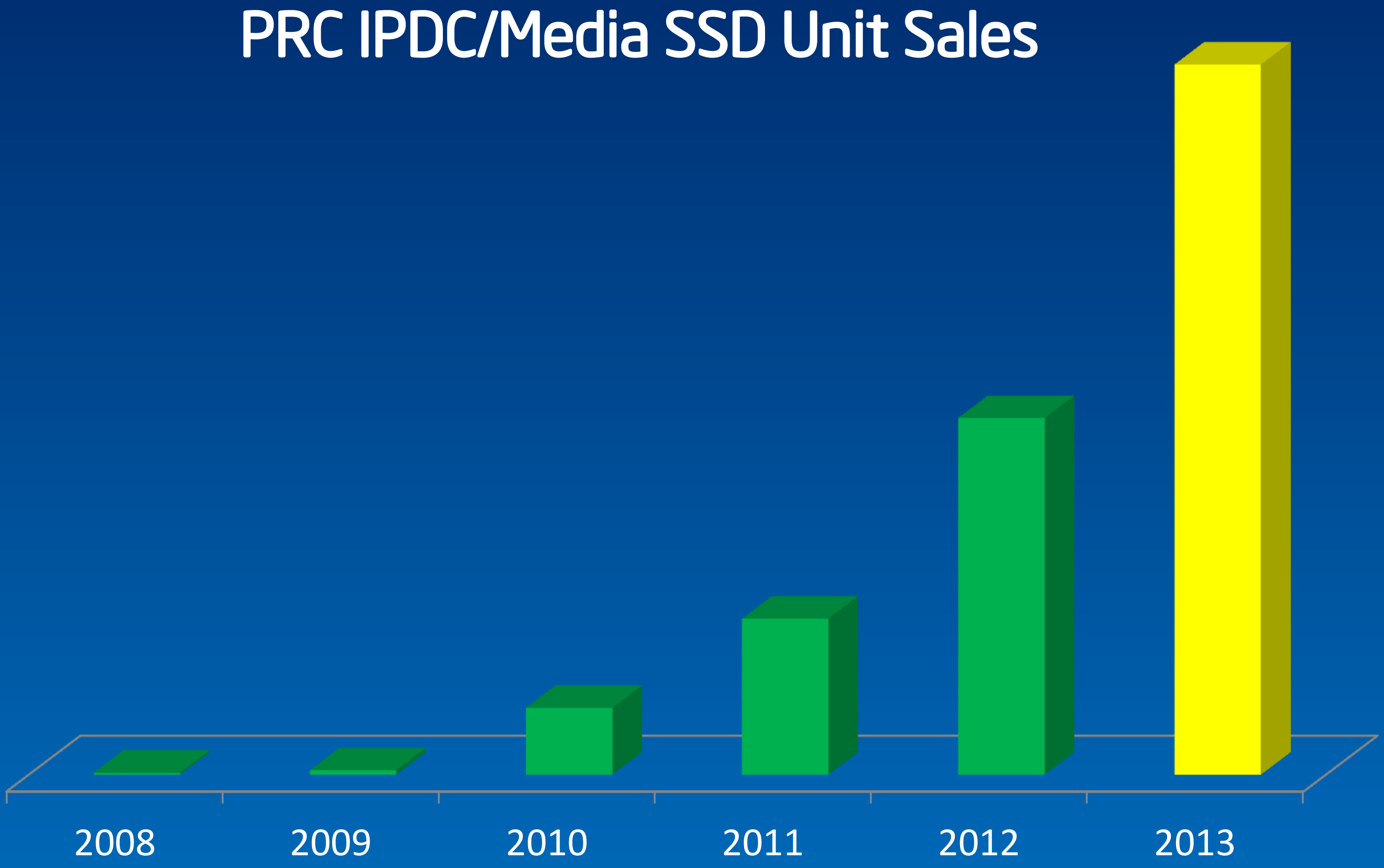
Benny NI

Business Development Manager

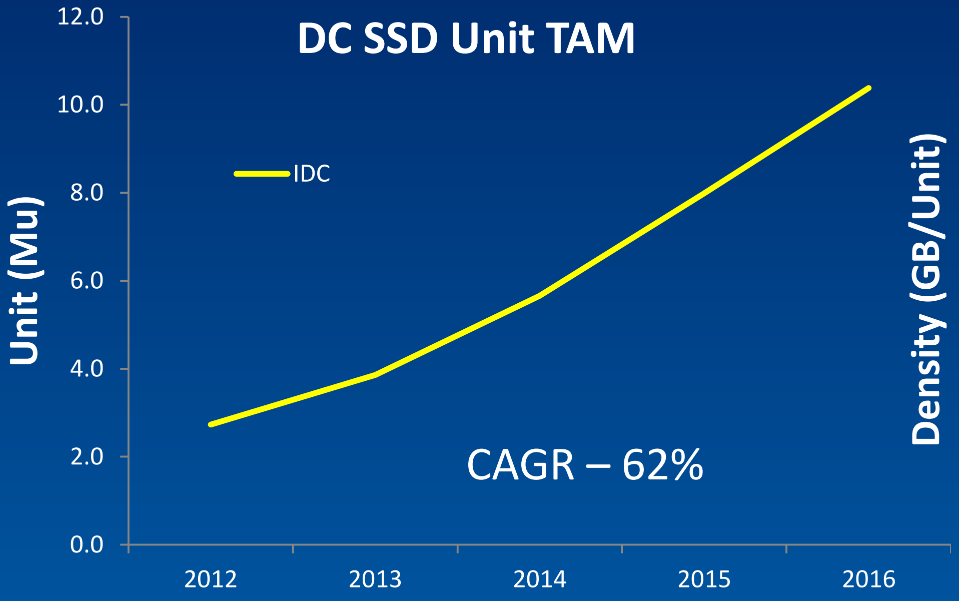
NVM Solutions Group, Intel

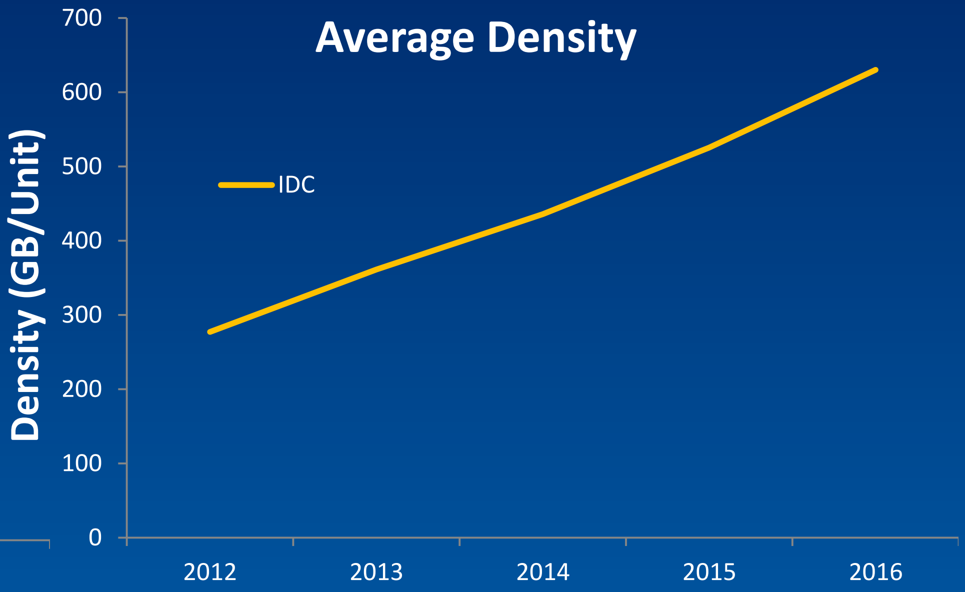
July, 2013

Thank You … for Being SSD Champions



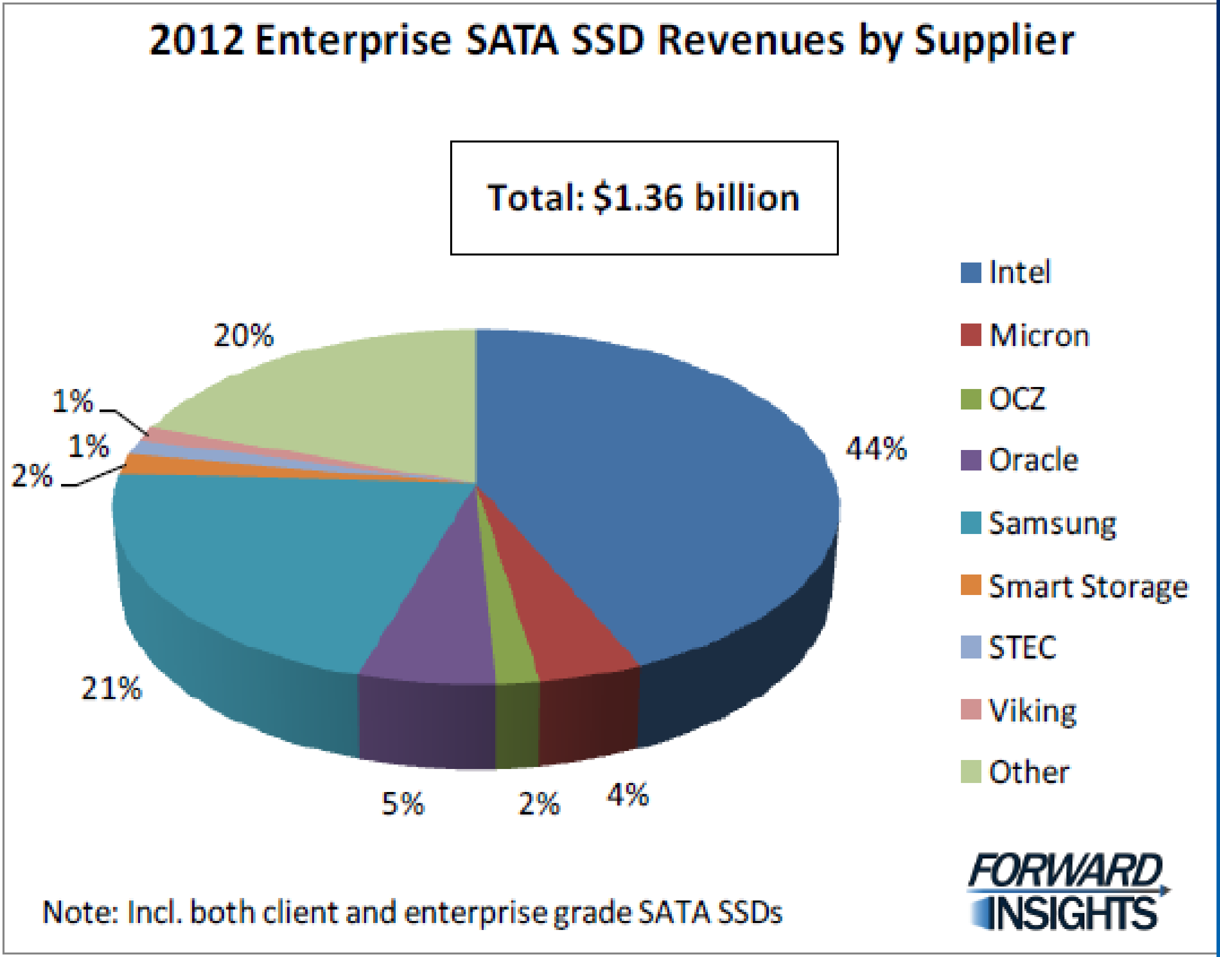
Data Center SSD Market Trend





Every High-Performance HDD Will Be Replaced by a SSD!

Enterprise SSD Market Analysis - SATA

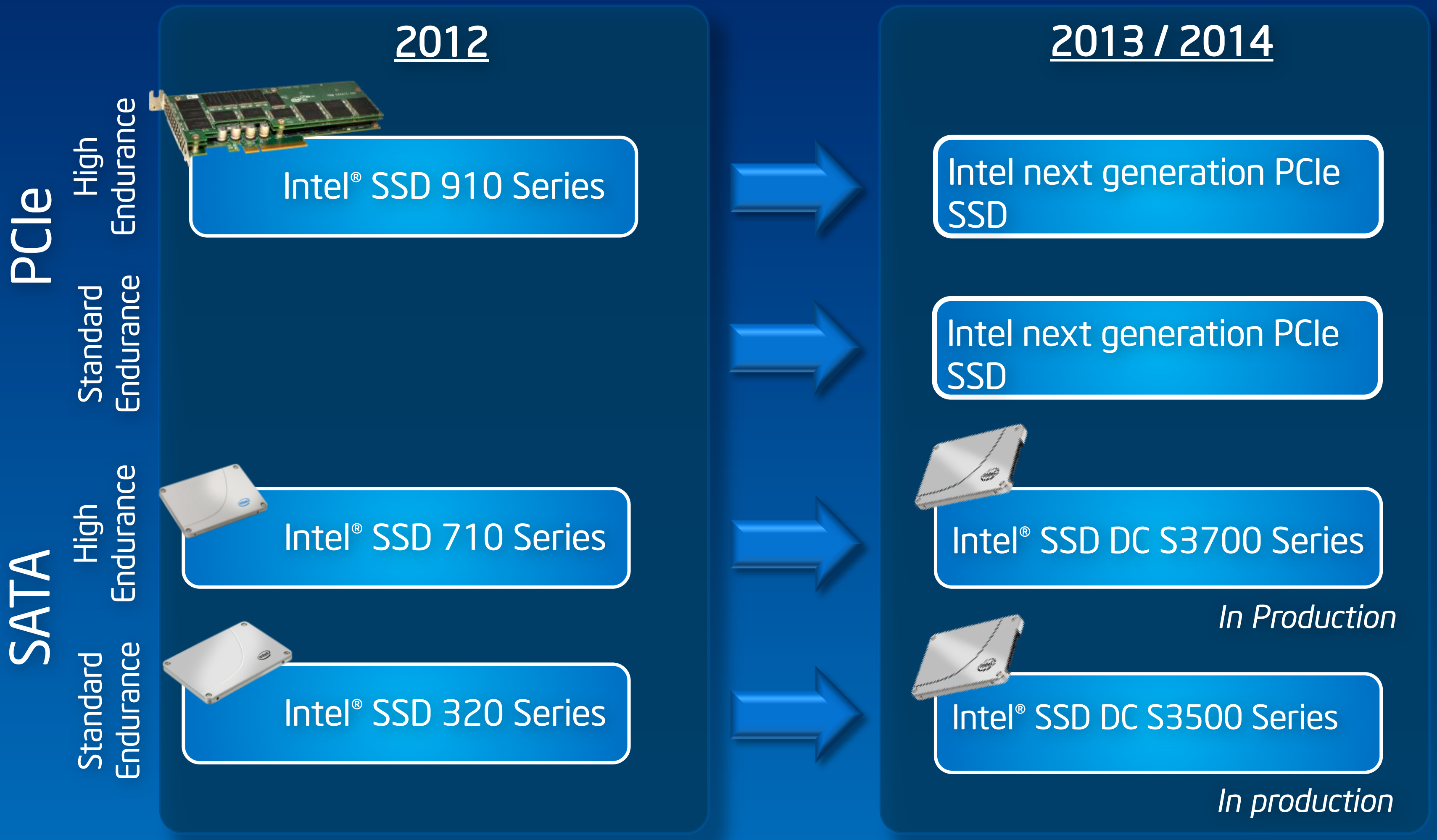


2012 Enterprise SATA SSD Revenues by Supplier

Note: Incl. both client and enterprise grade SATA SSDs

Intel is leading the market growth!

Intel Data Center SSDs



2012

2013 / 2014

PCIe

High Endurance

Intel® SSD 910 Series → Intel next generation PCIe SSD

Standard Endurance

→ Intel next generation PCIe SSD

SATA

High Endurance

Intel® SSD 710 Series → Intel® SSD DC S3700 Series

In Production

Standard Endurance

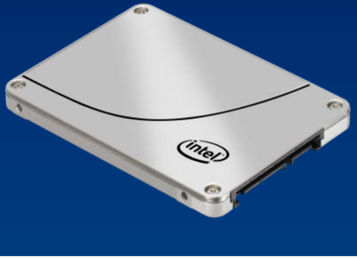
Intel® SSD 320 Series → Intel® SSD DC S3500 Series

In production

Product Feature Differences

Improvement across the board

	Intel® SSD 710 Series[1]	DC S3700 Series[2]
Capacity	100/200/300GB	2.5" 100/200/400/800GB 1.8-inch 200/400GB
Interface	SATA 3Gbps (ATA8)	SATA 6Gbps (ATA8)
Performance Transfer Rate (Read/Write)	270/210MB	500/460MB
IOPS (4K Random Read/Write)	38.5K/2.7K IPOS	75K/36K IPOS
Latency Average (Read/Write)	75/85μs	50/65μs
Features Encryption	128-bit AES	256-bit AES
Data Integrity	LBA Tag Checking	End-to-end data protection
Warranty	Three years	Five years
Endurance	4.5 drive writes per day	10 drive writes per day
Power Loss Protection	Yes	Yes plus Self Test



Increased capacities

Improved performance, latencies, and endurance

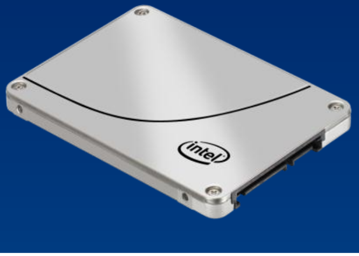
2X the endurance

[1] Data based on Intel® SSD 710 Series data sheet.
[2] DC S3700 data is preliminary.

Product Feature Differences

Improvement across the board

	Intel® SSD 320 Series	DC S3500 Series
Capacity	80/120/160/300/600GB	2.5” 80/120/160/240/300/480/800 1.8” 80/240/400
Interface	SATA 3Gbps (ATA8)	SATA 6Gbps (ATA8)
Performance Transfer Rate (Read/Write)	270/220MB	500/450MB
IOPS (4K Random Read/Write)	39.5K/600 IPOS	75K/11.5K IPOS
Latency Average (Read/Write)	75/95µs	50/65µs
Features Encryption	128-bit AES	256-bit AES
Data Integrity	LBA Tag Checking	End-to-end data protection
Warranty	Five years	Five years
Endurance (4k full span)	0.06 drive writes per day	0.3 drive writes per day
Power Loss Protection	Yes	Yes plus Self Test



Increased capacities

Improved performance, latencies, and endurance

>5X the endurance

Transition to the DC S3500 Series

	Intel® SSD 320 Series	Intel® SSD 520 Series	Intel® SSD DC S3500 Series	Benefit
Full Data Path protection	◒	◒ Data Path only	● Data + Non Data Path	Protects against unexpected data corruption throughout the drive
Power Loss Data Protection	◒ PLI	○	● PLI + PLI check	Protects data against unexpected power loss
Intel Developed Controller	●	○	●	Intel Quality & Reliability
Consistent Performance	○	◒ 18% better than 320	● 50% better than 520	Tighter IOPS and lower max latencies for consistent and predictable performance
AES 256b encryption	◒ 128b	◒ 128b	● 256b	Enhanced data protection for data at rest
High Capacities	◒ 600GB	◒ 480GB	● 800GB	Increased capacities for growing storage needs
NAND Technology	◒ 25nm	◒ 25nm	● 20nm	Leading edge NAND technology provides a better cost structure

Migrate to DC S3500 to gain and save!

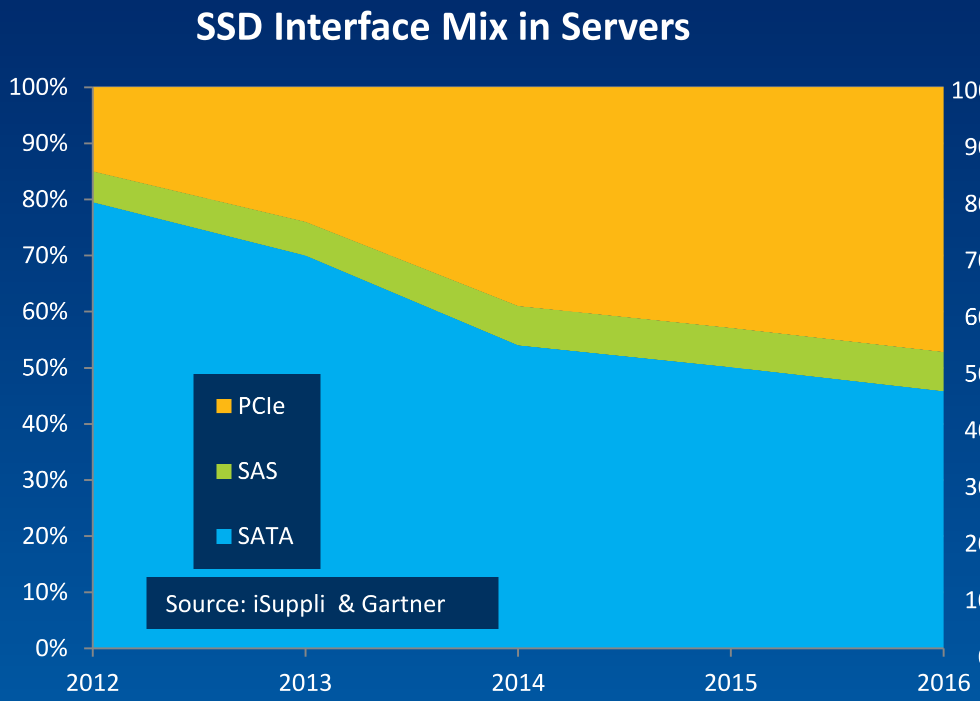
Intel® SSDs Enhance Corp IT Efficiency Microsoft Exchange

- Intel IT – Server + 40 HDD > Server + 14 DC S3700 SSD
  - 50% per user infrastructure cost reduction

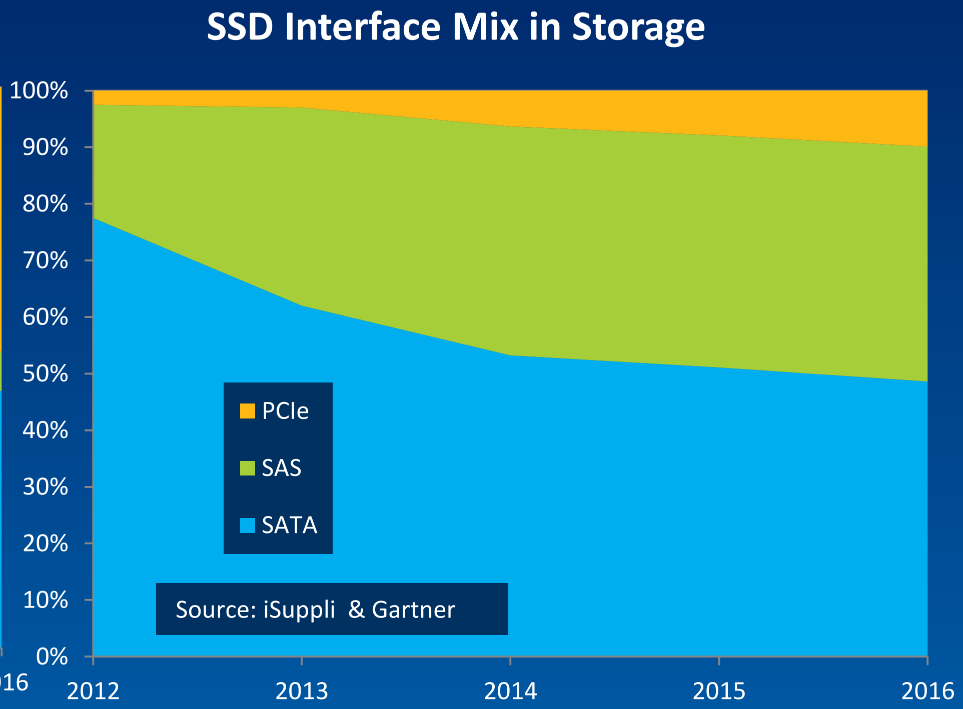
80% Utilization of All Assets
99% SLA in T1 Apps
95% SLA in T2+
10% Y-o-Y Cost Reductions

Parameter	10k + 7k HDD Config	DC S3700 SSD Config	Delta
Active Users	6K Users	12K > 18k Users	2x More Users
LDAP look up	1x	2x	2x Faster
Mail Submission[1]	1x	6x	6x Faster Outbox
CPU Headroom	NA	2x Available CPU	Room to Grow Predictable Performance
System Configuration	Server + 2x JBOD (spindles for IOPS not TB)	Server only	Less Management & Complexity
Size	6U ($120/Yr. @ $105/SqFt)	2U ($40/Yr. @ $ 105/SqFt)	60% Space Reduction
Total Power & Cooling (Server + 1.25*Server)	1780 Watts* ($1080/Yr. @ $.07KWh)	370Watts* ($230/Yr. @ $.07KWh)	79% Power Reduction*
Cost Server & JBOD	~$20k Total Server + 2x JBOD	~$30K Server Only	33% Increase in BoM Cost
$/user	3.33$/user	2.5$/user， low to 1.67$/user	25%-50%↓

SSD Interface Mix Trend in Data Center



Source: iSuppli & Gartner



Source: iSuppli & Gartner

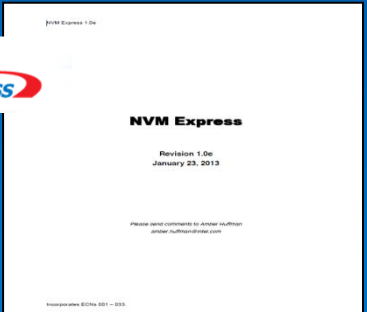
SATA continues to take >50% share while PCIe is taking off!

NVM Express (NVMe) Overview

- NVM Express is a high performance, scalable host controller interface designed for Enterprise and client systems that use PCI Express* SSDs
- NVMe developed by industry consortium of 80+ members and is directed by a 13-company Promoter Group
- NVMe 1.0 published March, 2011
- NVMe 1.1 published October, 2012 adding Enterprise and Client capabilities
  - Enterprise: Multi-path I/O and namespace sharing
  - Client: Lower power through autonomous transitions during idle
- Reference drivers available for Microsoft* Windows and Linux*, others in development
- The first UNH-IOL NVMe plugfest held on May 13-16, 2013 in Durham, NH to enable an interoperable ecosystem.
- Additional information at NVMExpress.org website

http://www.nvmexpress.org/resources/

NVMe command structures and specs found here



NVM Express 1.0e

NVM Express

Revision 1.0e
January 23, 2013

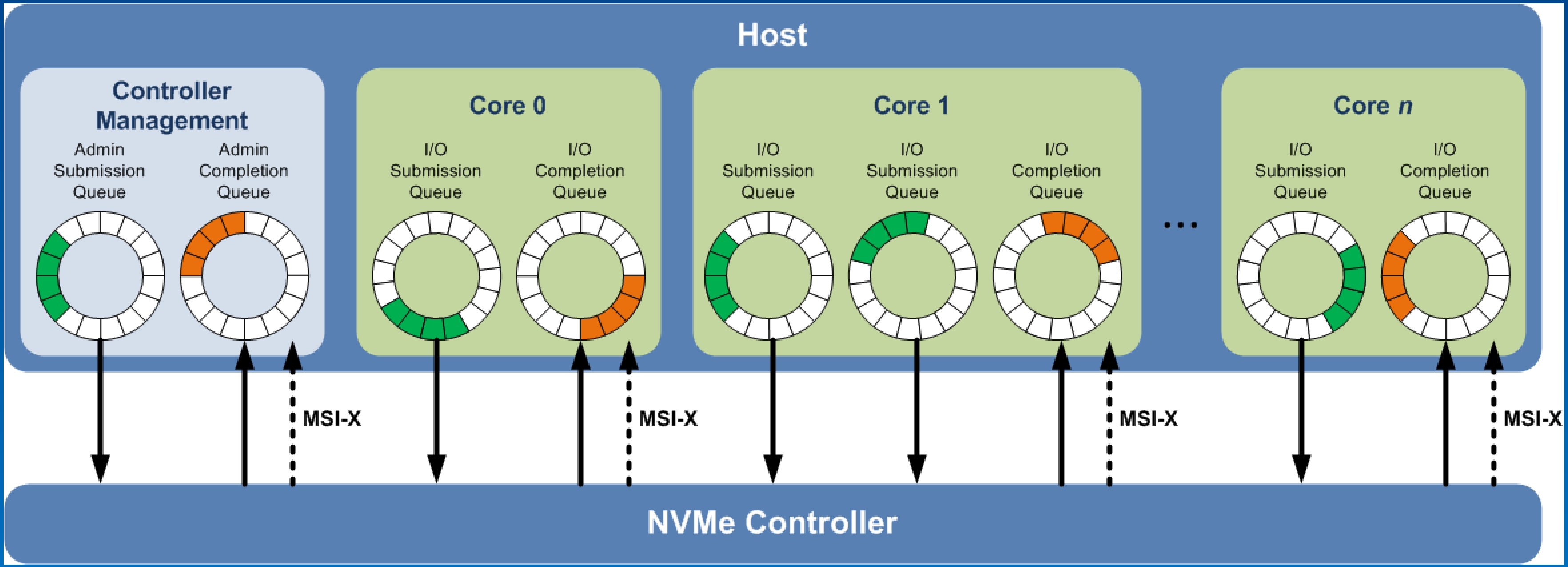
Please send comments to Amber Huffman
amber.huffman@intel.com

Incorporates ECNs 001 – 033

*Other names and brands may be claimed as the property of others.

NVM Express (NVMe) Technical Basics

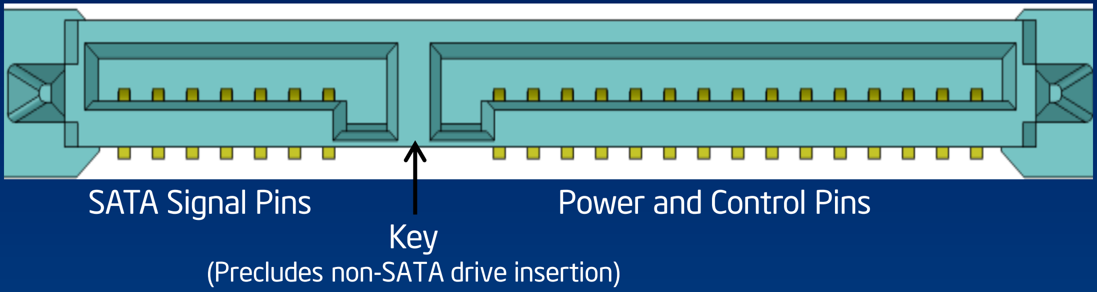
- The focus of the effort is efficiency, scalability and performance
  - All parameters for 4KB command in single 64B DMA fetch
  - Supports deep queues (64K commands per Q, up to 64K queues)
  - Supports MSI-X and interrupt steering
  - Streamlined command set optimized for NVM (6 I/O commands)
  - Enterprise: Support for end-to-end data protection (i.e., DIF/DIX)
  - NVM technology agnostic



2.5" SFF PCIe Drive:
From SATA, to SAS, to SFF 8639

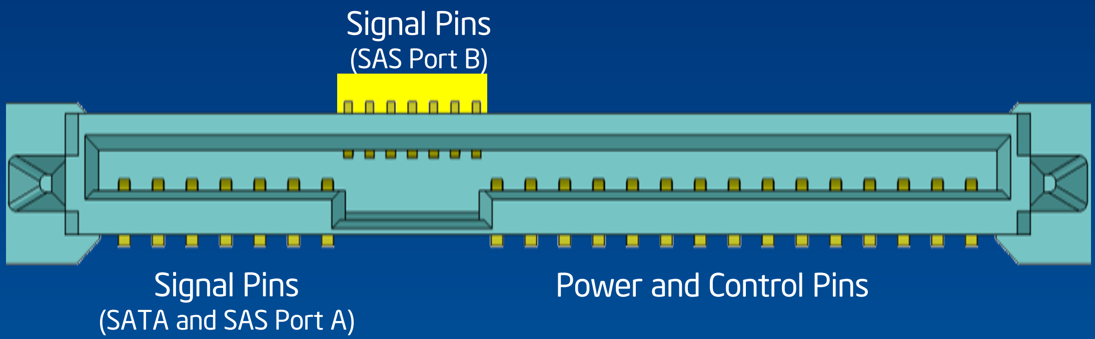
Current SATA Connector

- Uses legacy SATA pin pitch
- Keyed to preclude the insertion of a non-SATA drive



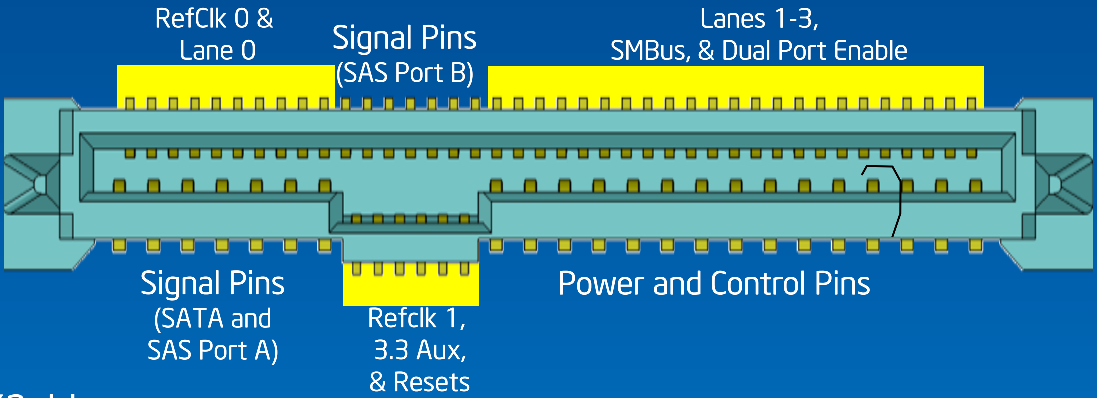
Current SAS Connector

- Added additional signaling pins for a secondary port option at with a tighter, modern, pin pitch
- Supports both SATA and SAS drives



SFF 8639 Connector

- Fills out all remaining pin capacity of the legacy form factor
- Designed to support many protocols
- Enterprise mapping supports legacy SATA, SAS, and modern PCIe drives simultaneously
  - Both single port X4 and dual port X2 drives

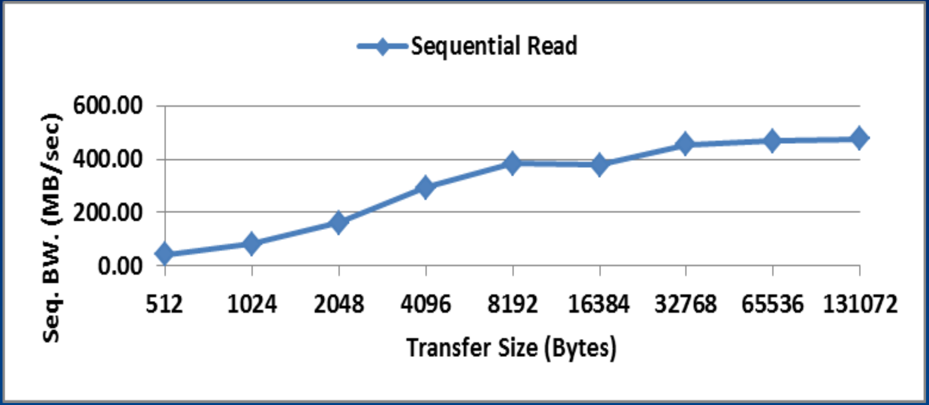


SFF 8639 Drives will support OOB Management

Specs can be found here-
http://www.ssdformfactor.org/docs/SSD_Form_Factor_Version1_00.pdf

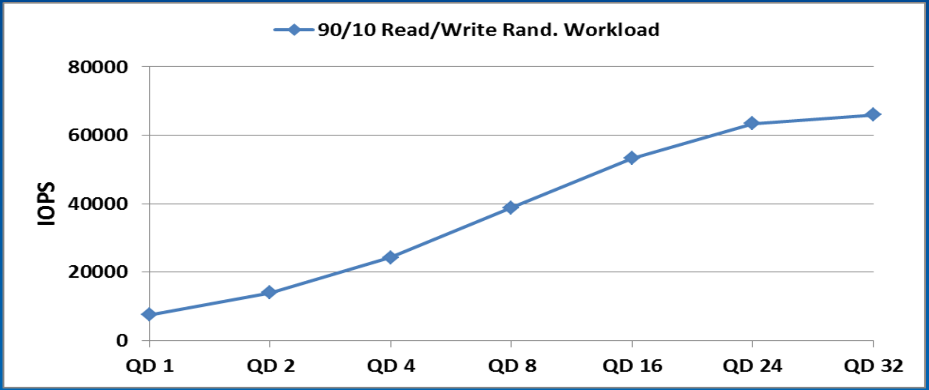
Parameters Effecting Performance –
Request Size, Queue Depth

- Request Size
  - Bandwidth Increases from smaller transfer size to bigger transfer size
  - Why: Fix command processing overhead



DC S3700 data

- Queue Depth
  - By operating at high queue depth, you increase performance. (More on random reads)
  - Why: We can assign work to multiple flash in parallel



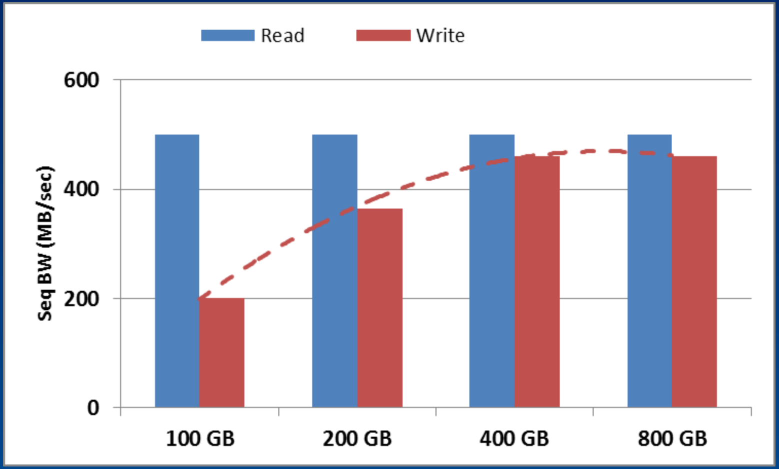
DC S3700 data

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.
Configurations: Measurements made usinng Intel i5-2400S CPU at 2.50 GHZ CPU and 4GB of DDR3 PC3-10600 Memory. Intel® DC S3700 Series, 800GB used for analysis.

Parameters Effecting Performance –
Density, Read/Write Mix

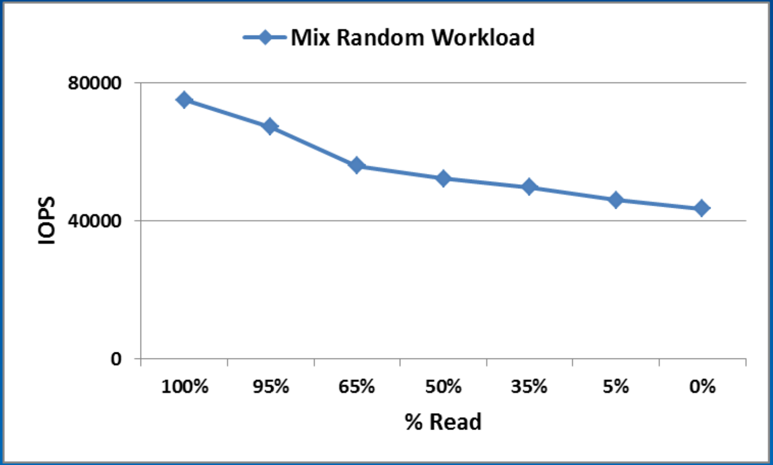
- Performance vs. Density
  - Density
    - Lower density → higher density increases performance
    - Why: More flash devices means more concurrent work possible

DC S3700 data



- Read/Write Mix
  - Moving from more writes to more reads increases performance
  - Why: Reads process faster than writes on NAND plus less “housekeeping”

DC S3700 data

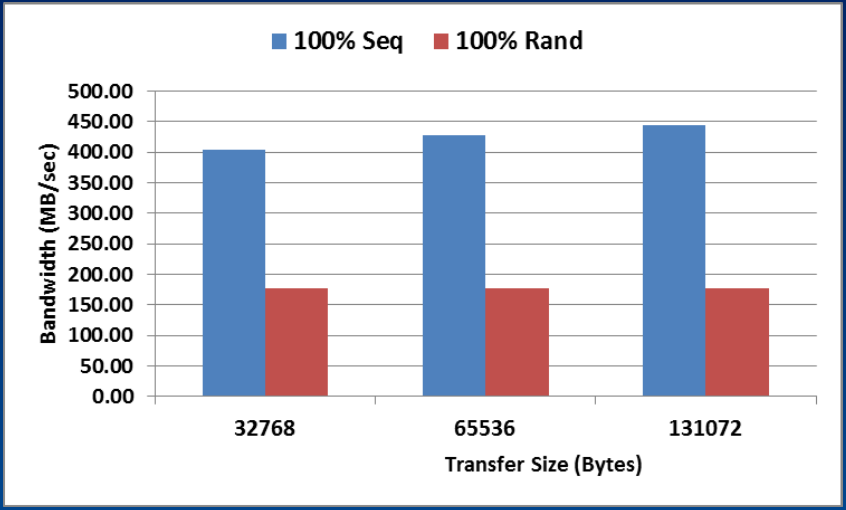


Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. Configurations: Measurements made usinng Intel i5-2400S CPU at 2.50 GHZ CPU and 4GB of DDR3 PC3-10600 Memory. Intel® DC S3700 Series, 800GB used for analysis.

Parameters Effecting Performance – Randomness, Over-provisioning

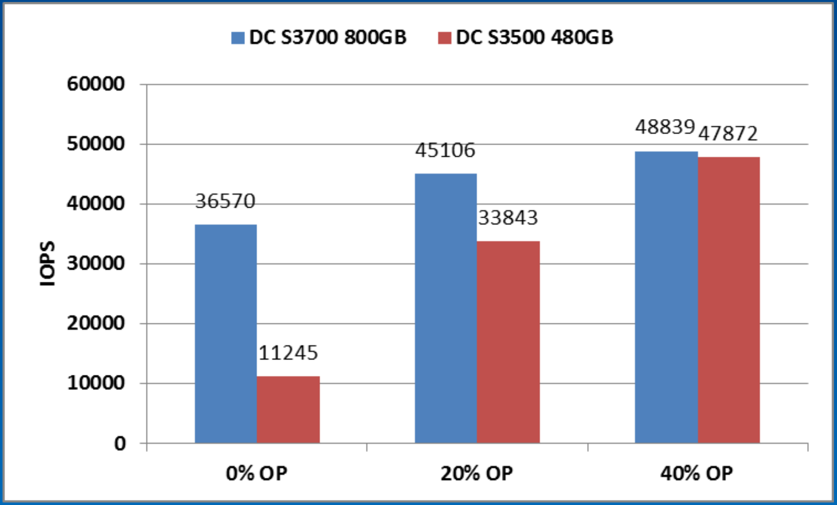
- % Random access
  - If application uses sequential accesses instead of random, it will improve performance and QoS
  - Why: Pre fetch on reads, reduced channel collisions, less NAND "housekeeping"

DC S3700 data



- Over-provisioning
  - Go from full LBA access to limited LBA access will improve performance, endurance and QoS
  - Why: Additional spare capacity allows "housekeeping" algorithms to run more efficiently

DC S3700/S3500 data

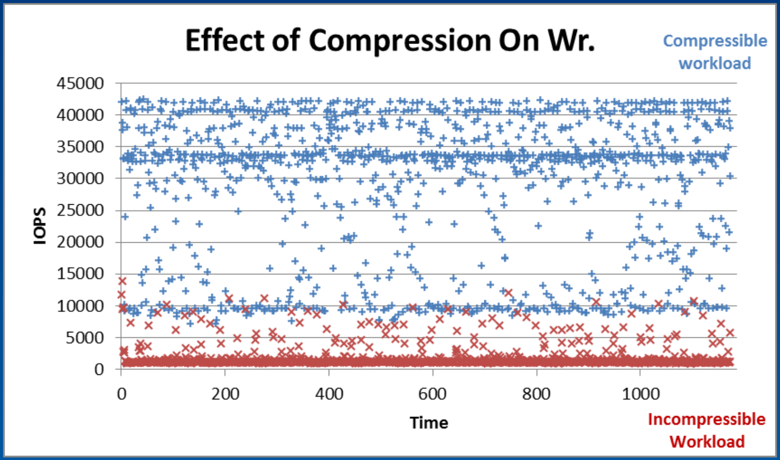


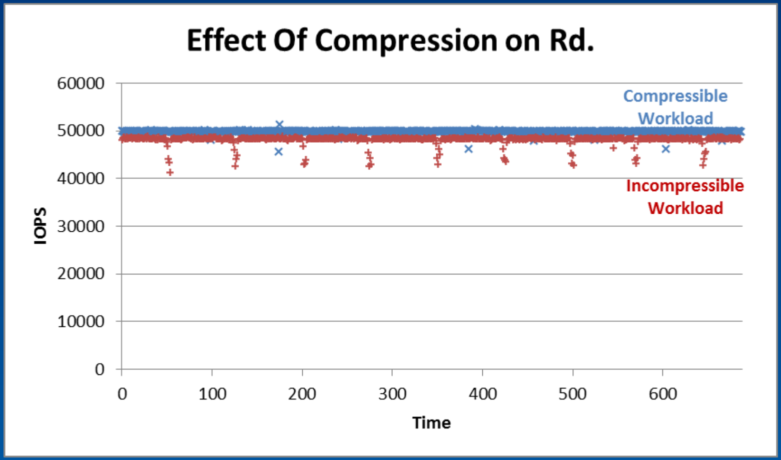
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.
Configurations: Measurements made usinng Intel i5-2400S CPU at 2.50 GHZ CPU and 4GB of DDR3 PC3-10600 Memory. Intel® DC S3700 Series, 800GB used for analysis.

Parameters Effecting Performance –
Compressibility, State of Drive

- Data Compressibility
  - Uncompressible data → compressible data → improved performance, improved endurance, QoS
  - Why: Less data read/written to NAND and increased spare capacity same value as short stroking

Intel SSD 520 Series Data



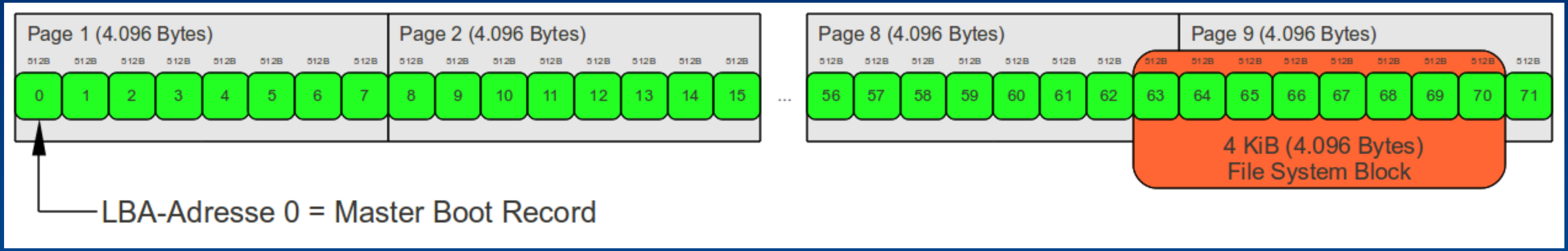


- Prior State of the Drive
  - Full and random drive → sequential writes and/or TRIM → higher performance
  - Why: the housekeeping algorithms need to work harder

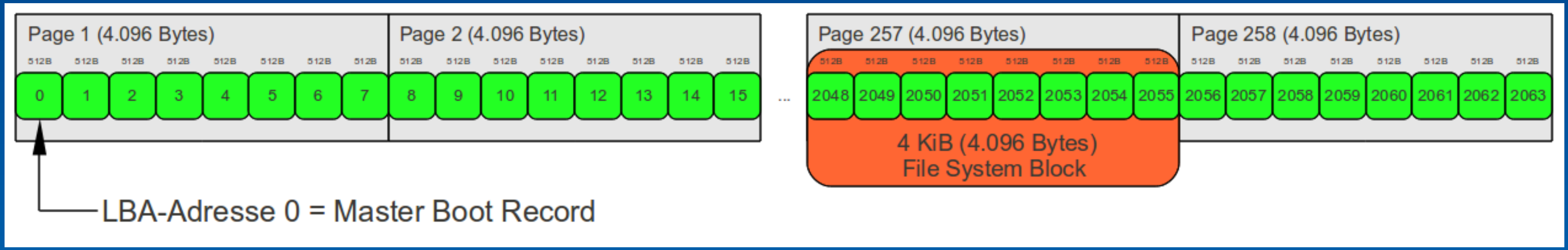
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.
Configurations: Measurements made usinng Intel i5-2400S CPU at 2.50 GHZ CPU and 4GB of DDR3 PC3-10600 Memory.Intel® DC S3700 Series, 800GB used for analysis.

LBA (4K-bytes) alignments

- Improper alignment, first partition starts with LBA address 63, it will hurt SSD performance due to RMW



- Proper 4Kbytes aligned partition



- Typical example at Linux partition

```
>> fdisk –u –c –b 4096 /dev/sdX
```

QoS (Quality of Service) 101

What Impacts QoS

- Drop in Bandwidth or IOPS from regular range
  - Background NAND management for reliability
  - Host versus housekeeping activity
- Latency outlier
  - move from usecond to milliseconds
- High frequency of latency outliers
  - Moving from 99.9999% availability to 99% availability

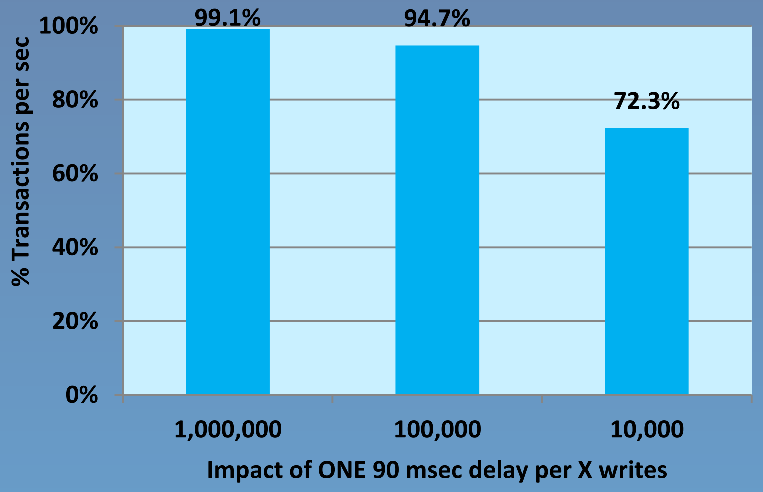
How to Benchmark QoS

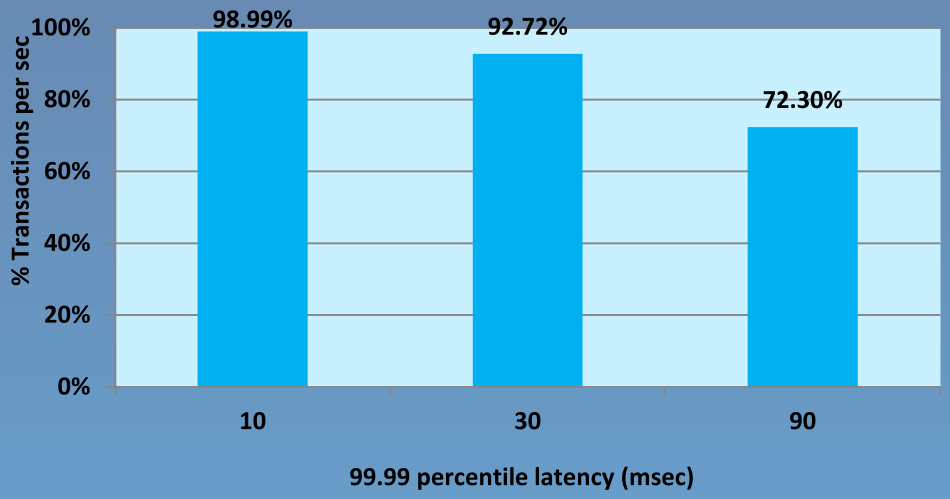
- Look at the tightness of IOPS spread
  - Measure average to min value, set to <20% variation for HE
- Look at the max latencies at low and high QD
  - Measure max latency with a high 9s availability (99.9999%)
  - 99.9999% means 1 outlier in 100 million data points

Importance of Latency QoS
TPCC* Random Workload

Transaction processing requires dense IO (Higher IOPS/GB)

No Mercy for latency outliers and occasional drops of IOPS





Non consistent performance impacts transaction processing

*Source: Intel simulated data based on Transaction Processing Performance Council (TPCC) workload

New Quality of Service Specification[1]

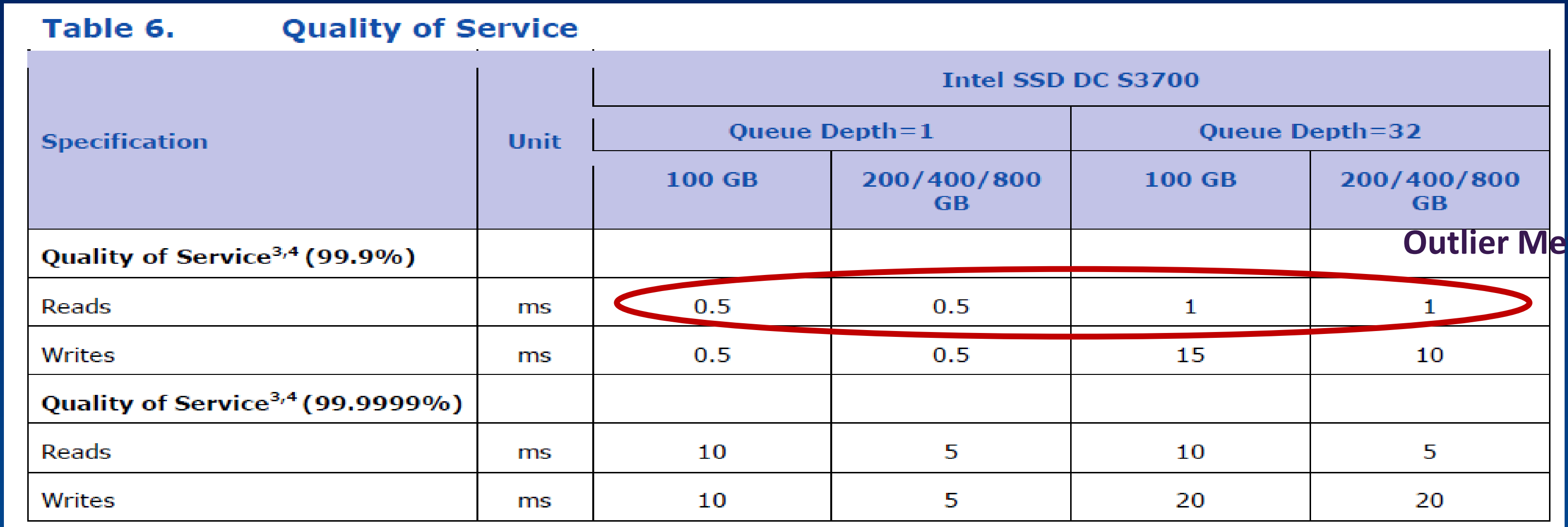


Table 6. Quality of Service

Specification	Unit	Intel SSD DC S3700			
		Queue Depth=1		Queue Depth=32	
		100 GB	200/400/800 GB	100 GB	200/400/800 GB
Quality of Service[3,4] (99.9%)					
Reads	ms	0.5	0.5	1	1
Writes	ms	0.5	0.5	15	10
Quality of Service[3,4] (99.9999%)					
Reads	ms	10	5	10	5
Writes	ms	10	5	20	20

Outlier Metric

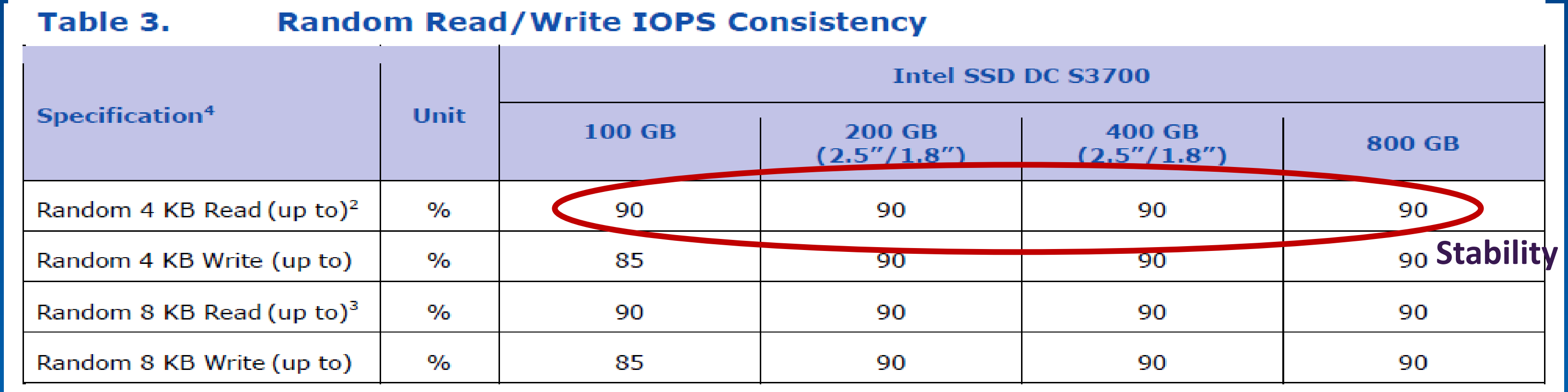


Table 3. Random Read/Write IOPS Consistency

Specification[4]	Unit	Intel SSD DC S3700			
		100 GB	200 GB (2.5"/1.8")	400 GB (2.5"/1.8")	800 GB
Random 4 KB Read (up to)[2]	%	90	90	90	90
Random 4 KB Write (up to)	%	85	90	90	90
Random 8 KB Read (up to)[3]	%	90	90	90	90
Random 8 KB Write (up to)	%	85	90	90	90

Stability Metric

Max Latency & IOP Consistency Specified

Source: http://www.anandtech.com/show/6433/intel-固态硬盘-dc-s3700-200gb-review/3

[1] Source: Intel® SSD DC S3700 Datasheet

New Quality of Service Specification[1]

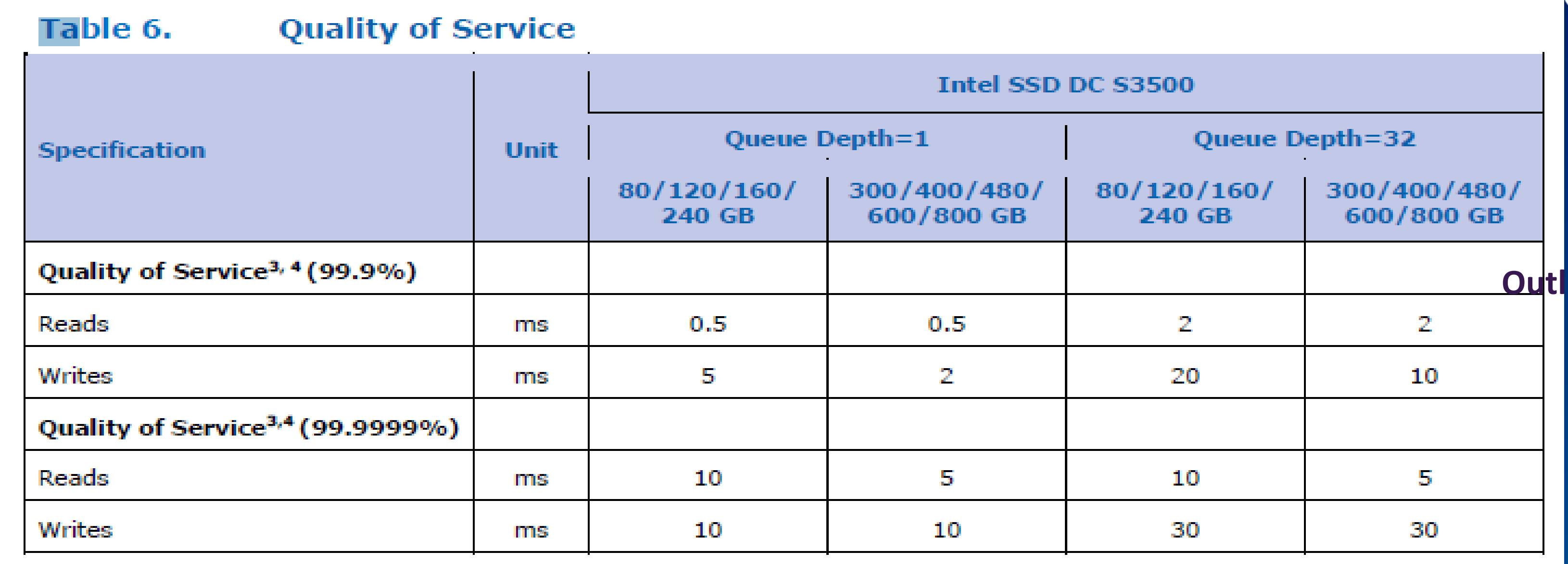


Table 6. Quality of Service

Specification	Unit	Intel SSD DC S3500			
		Queue Depth=1		Queue Depth=32	
		80/120/160/ 240 GB	300/400/480/ 600/800 GB	80/120/160/ 240 GB	300/400/480/ 600/800 GB
Quality of Service[3, 4] (99.9%)					
Reads	ms	0.5	0.5	2	2
Writes	ms	5	2	20	10
Quality of Service[3,4] (99.9999%)					
Reads	ms	10	5	10	5
Writes	ms	10	10	30	30

Outlier Metric

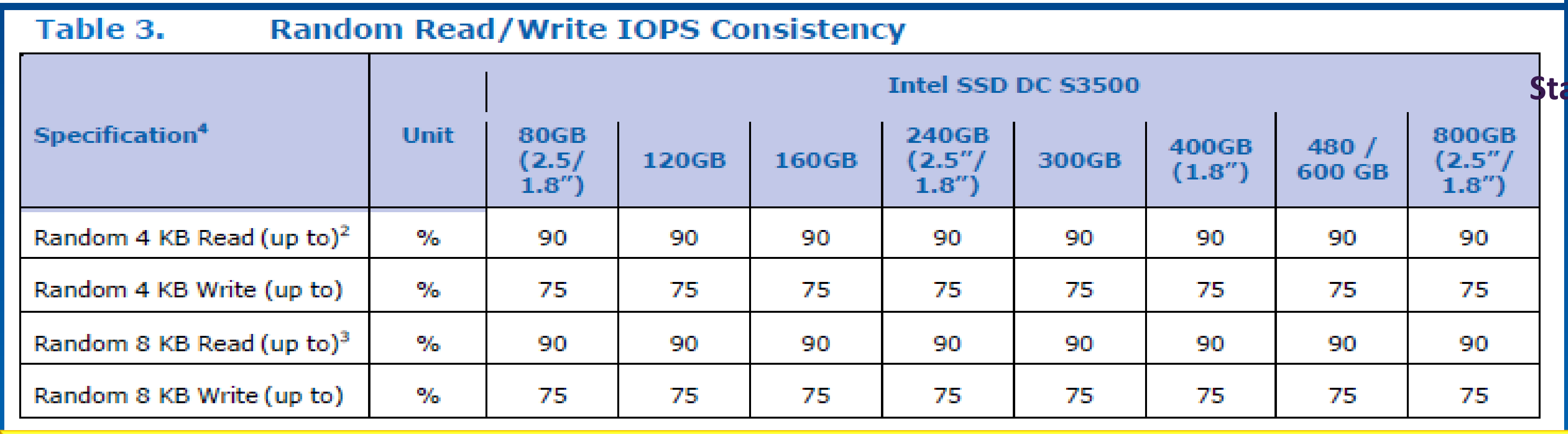


Table 3. Random Read/Write IOPS Consistency

Specification[4]	Unit	Intel SSD DC S3500							
		80GB (2.5/ 1.8")	120GB	160GB	240GB (2.5"/ 1.8")	300GB	400GB (1.8")	480 / 600 GB	800GB (2.5"/ 1.8")
Random 4 KB Read (up to)[2]	%	90	90	90	90	90	90	90	90
Random 4 KB Write (up to)	%	75	75	75	75	75	75	75	75
Random 8 KB Read (up to)[3]	%	90	90	90	90	90	90	90	90
Random 8 KB Write (up to)	%	75	75	75	75	75	75	75	75

Stability Metri

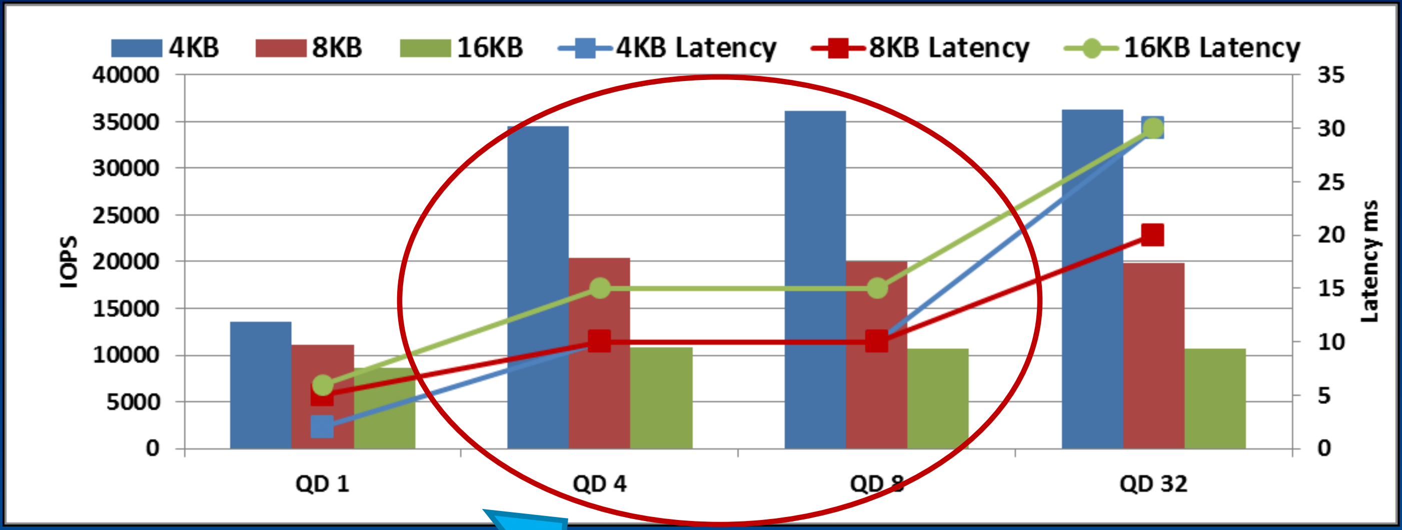
Max Latency & IOP Consistency Specified

http://www.anandtech.com/show/7065/intel-ssd-dc-s3500-review-480gb-part-1

[1] Source: Intel® SSD DC S3700 Datasheet

Data Center Performance Optimization

Example Intel® SSD DC S3700



100% Random Write workload on DC S3700
Latency measured at 99.999% outlier

IOPS Saturation Already Happened here? What's the point of incurring more latency?

- Add more drives or over-provision to gain higher IOPS
- Limit QD per drive to meet the max latency requirement of the system
- QD/drive and IOPS/drive will help size your database without hitting high latency events

Minimize Latency by Optimizing for QD/Drive

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.
Configurations: Measurements made usinng Intel i5-2400S CPU at 2.50 GHZ CPU and 4GB of DDR3 PC3-10600 Memory. Intel® DC S3700 Series, 800GB used for analysis.

Configuring for Raid/HBA performance

- Use latest RAID/HBA SSD-friendly firmware which simplifies previous HDD software stacks, such as called FastPath* IO

– Disable RAID Read Caching

– Application stacks IO queues/threads

  - Use max queue depth on each striped drives times the number of stripped drives for maximum **read** performance
  - Use proper queue depth (4 to 8) on each striped drives times number of striped drives for better **write** performance and lower latency

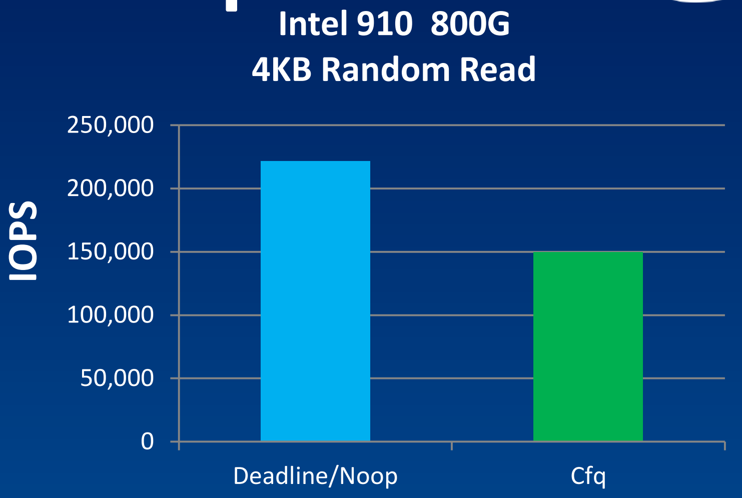
– RAID parameters: **wt nora direct…strpszM sz**

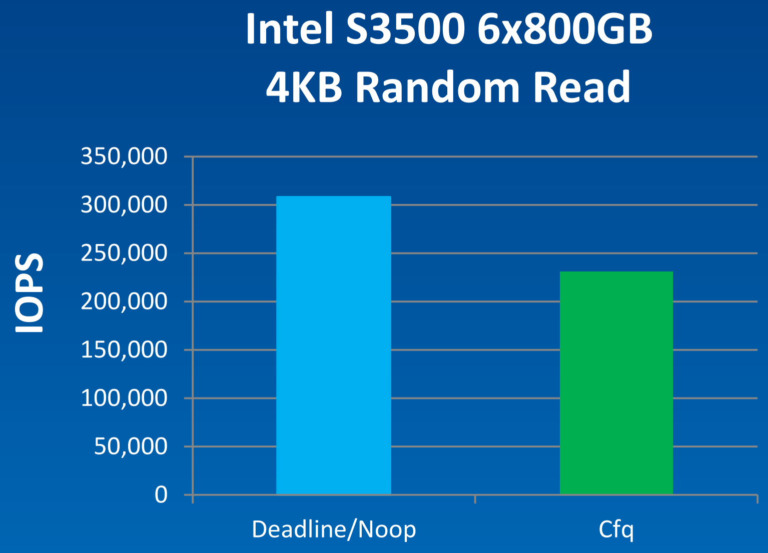
  - "sz" equals to over-provision for all SSDs in RAID (MUST do security erase SSDs before "sz")

Configuring Linux IO for better performance

- Use noop/deadline (default is cfq)
  - /sys/block/sdX/queue/scheduler
- Turn rotational=0
- Turn off read_ahead_kb=0
- Adjust nr_requests value based on number of drives
- Disable I/O barrier on all Intel data center SSDs (all have power protection feature)
  - barrier=0 (ext3, ext4) or nobarrier (XFS)

– Check rq_affinity (use 2, RHL6.4 default is 1)



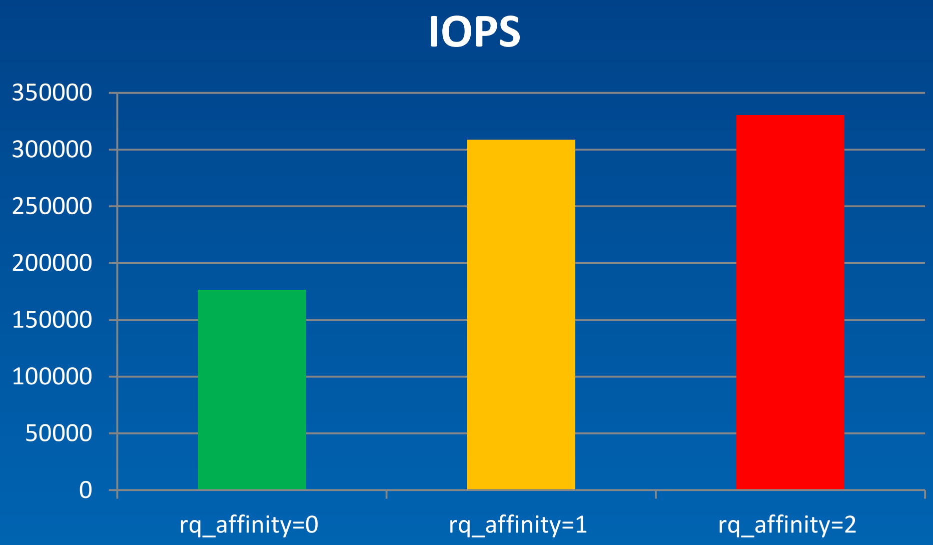


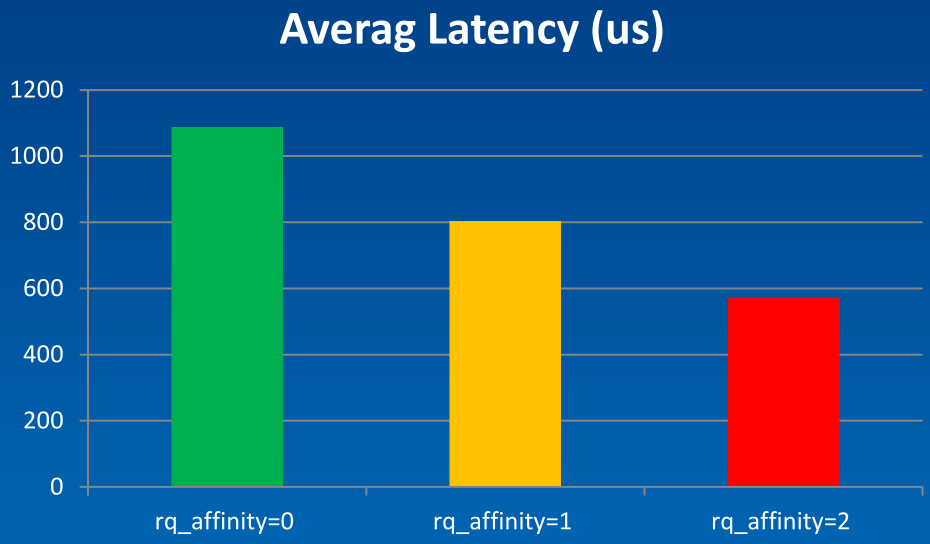
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.
Configurations: Measurements made using Hanlan Creek (Intel S5520HC) system with two Intel® Xeon X5560@ 2.93GHz and 12GB (per CPU) Mem running RHE6.1 O/S

Regarding softirq rq_affinity

- **Strict** rq_affinity distributes soft interrupts to different CPU cores (rq_affinity=2 is available at RHE 6.4 release)
- Demo on S3500x6 with LSI HBA 2008
- Example case: 6 x Intel® 3500 800GB SSD behind HBA controller

4K 100% random read (threads =32x6)


IOPS


Averag Latency (us)

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.
Configurations: Measurements made using Hanlan Creek (Intel S5520HC) system with two Intel® Xeon X5560@ 2.93GHz and 12GB (per CPU) Mem running RHE6.1 O/S, Strict affinity is available after kernel patches, LSI*2008 HBA controller

SSD life measurement and monitor

- S.M.A.R.T provides SSD health info
- SMART info can be retrived on most Raid controllers now
- At pre-production, use E2/3/4 to measure SSD wearing status under timed workloads, then estimate SSD life time
  - E2(226) Timed Workload Media Wear out Indicator
    - Reports % of wear during a test period not less than 60 mins
    - Raw value needs to be divided by 1024 to get the % #
  - E3(227) Timed Workload Read/Write Ratio
    - Reports the raw value of the ratio
  - E4(228) Workload Timer
    - Reports out the raw value of time during a run
- Monitoring E9 at your regular maintenance job,

  when E9 reaches to 1, backup data and change SSDs

THANK YOU