SSD Benchmarking at CERN

Liviu Vâlsan

HEPiX Fall 2014 Workshop, 16th of Oct 2014

SSD Benchmarking at CERN

- **Technology overview**
- **SSD performance**
  - Synthetic benchmarks
  - Simulations of Real Work applications
- **SSD endurance**
- **Hurdles for SSD adoption**
- **Next steps**
- **Conclusion**

SATA: A 10+ Years Old Interface for HDDs

- **2003**: SATA revision 1.0 - 1.5 Gbit/s - **150 MB/s**
- **2004**: SATA revision 2.0 - 3 Gbit/s - **300 MB/s**
- **2008**: SATA revision 3.0 - 6 Gbit/s - **600 MB/s**
- **2011**: SATA revision 3.1
  - No speed increase
  - Just extra form factors and features (TRIM)
- **2013**: SATA revision 3.2 (**SATA Express**) - 16 Gbit/s - **1969 MB/s**

NVMe Allows Standardization of PCIe SSDs

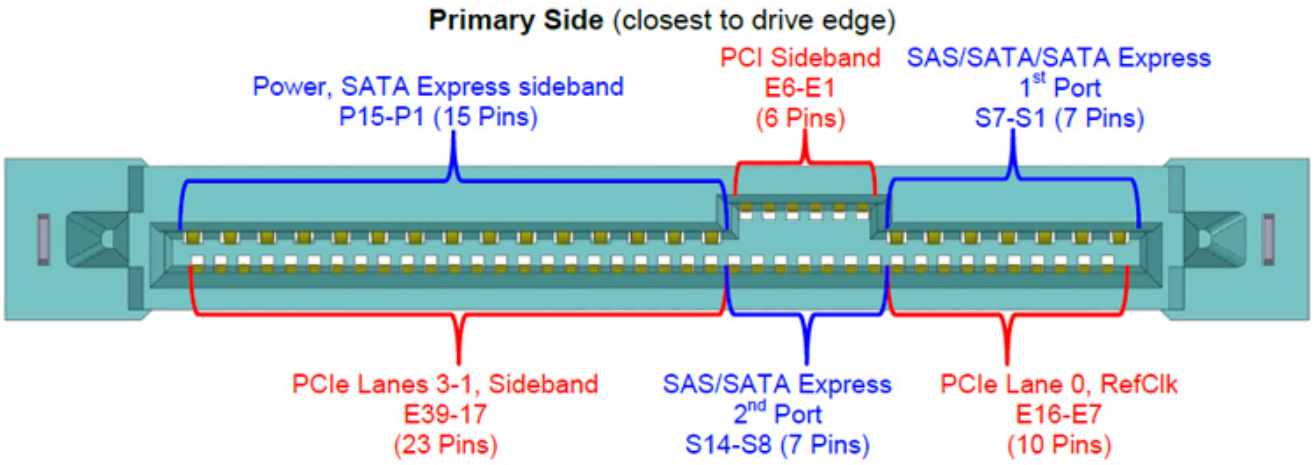
- **A specification for accessing PCIe attached SSDs**
- **PCIe removes controller latency**
- **NVMe reduces software latency**
- **OS support**
  - Linux support since kernel 3.3, stable since kernel 3.10
  - RHEL / SL 6.5, RHEL / CentOS / SL 7
  - Microsoft Windows Server 2012 R2, Windows 8.1
  - UEFI, QEMU, FreeBSD, Solaris

SATA Express

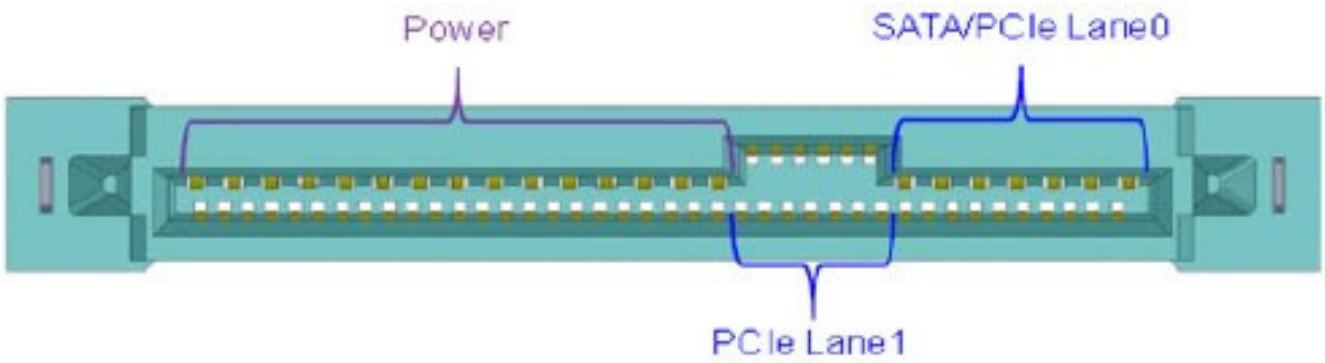
- Defines form factors / connectors that support either SATA or PCIe based drives
- Does not define the software interface
  - AHCI or NVMe can be used
- Exposes multiple PCIe lanes and two SATA gen 3.0 6 Gbit/s ports through the same host-side SATA Express connector
- Exposed PCIe lanes provide a pure PCIe connection to the SSD, without any additional layers of abstraction

Two Main Connectors

SFF-8639



SATA Express



AHCI vs NVMe

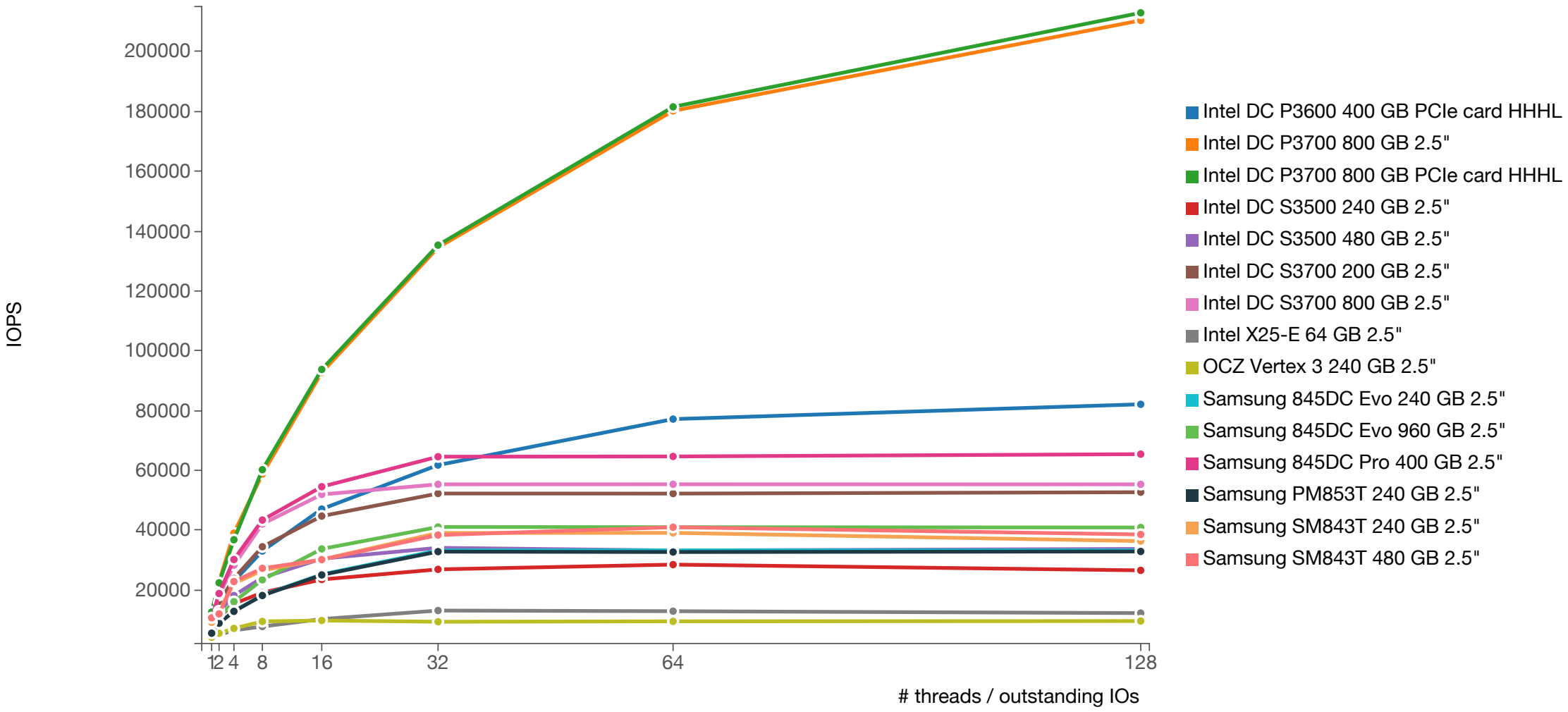
	AHCI	NVMe
Maximum queue depth	1 command queue 32 commands per queue	65536 queues 65536 commands per queue
Uncacheable register accesses (each consumes 2000 CPU cycles)	4 per command - 8000 cycles ~ 2.5 μs	0 per command
MSI-X and interrupt steering	single interrupt no steering	2048 MSI-X interrupts
Parallelism and multiple threads	requires synchronization lock to issue a command	no locking
Efficiency for 4 KB commands	command parameters require two serialized host DRAM fetches	gets command parameters in one 64 Bytes fetch

List of Benchmarked SSDs

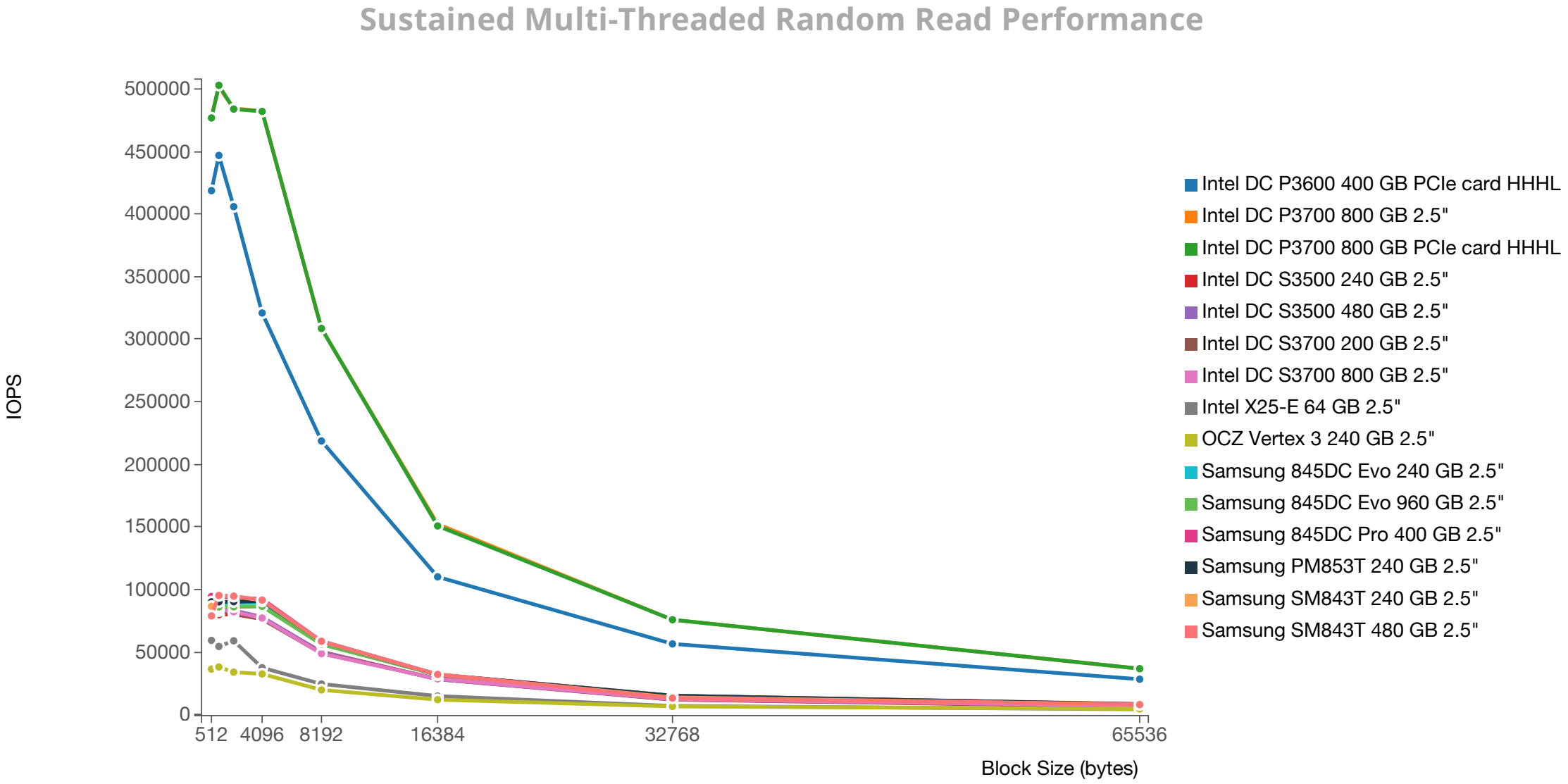
Manufacturer	Family	Available capacities (GB)	Tested capacities (GB)	Interface	Flash Type	Endurance (DWPD)
Intel	DC S3500	80, 120, 160, 240, 300, 480, 600, 800	240, 480	SATA rev. 3.0	20 nm MLC	0.3
Intel	DC S3700	100, 200, 400, 800	200, 800	SATA rev. 3.0	25 nm MLC	10
Intel	DC P3600	400, 800, 1200, 1600, 2000	400	PCIe gen 3 x4	20 nm MLC	3
Intel	DC P3700	400, 800, 1600, 2000	800	PCIe gen 3 x4	20 nm MLC	10
Intel	X25-E	32, 64	64	SATA rev. 2.0	50 nm SLC	18
Samsung	845DC Evo	240, 480, 960	240, 960	SATA rev. 3.0	19 nm TLC	0.35
Samsung	845DC Pro	400, 800	400	SATA rev. 3.0	40 nm MLC V-NAND	10
Samsung	SM843T	120, 240, 480, 960	240, 480	SATA rev. 3.0	20 nm MLC	2
Samsung	PM853T	240, 480, 960	240	SATA rev. 3.0	19 nm TLC	0.3
OCZ	Vertex 3	60, 90, 120, 240, 480	240	SATA rev. 3.0	25 nm MLC	0.3

4 KB Random Mixed 70% Read / 30% Write

Sustained 4 KB Random Mixed 70% Reads / 30% Writes Performance by # of Threads

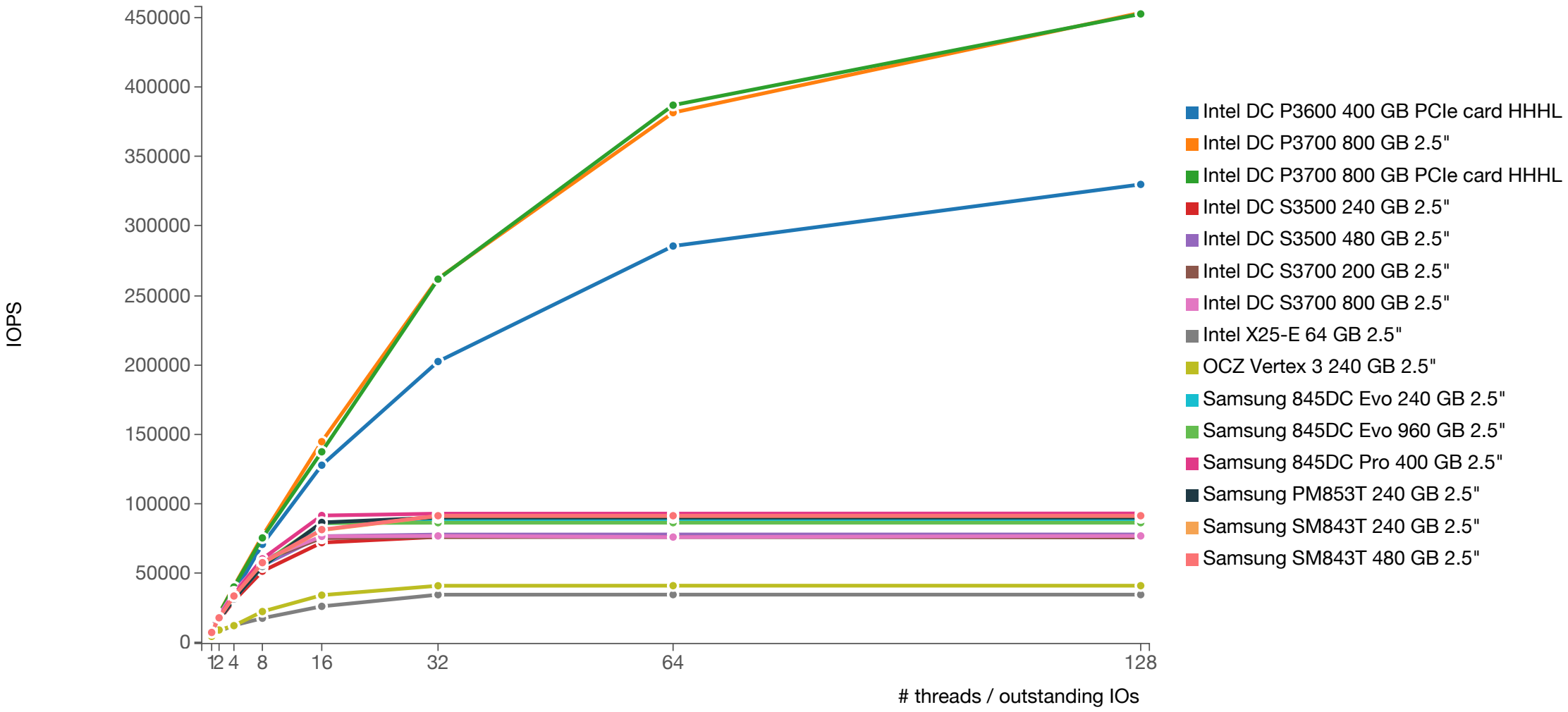


Sustained Multi-Threaded Random Read Performance



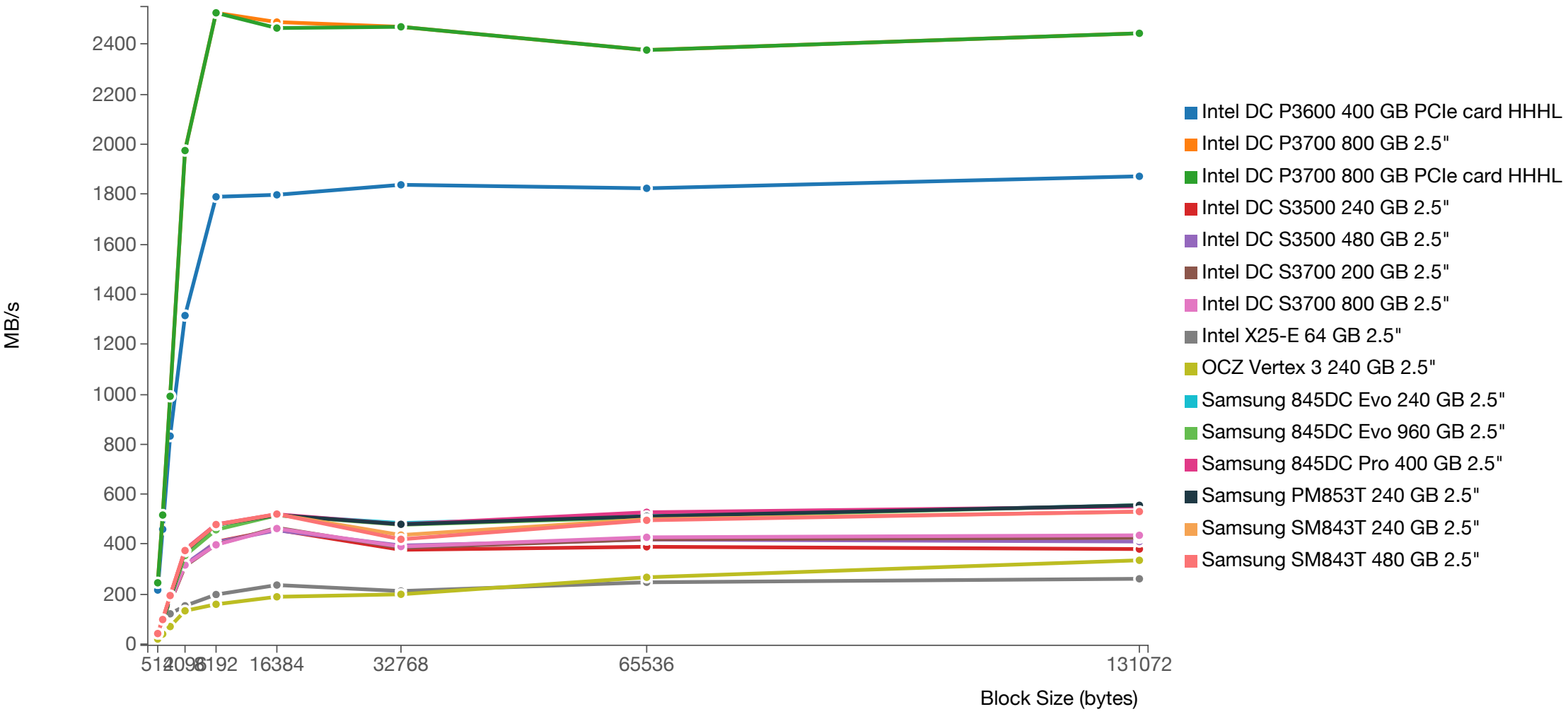
Sustained 4 KB Random Read Performance

Sustained 4 KB Random Read Performance by # of Threads

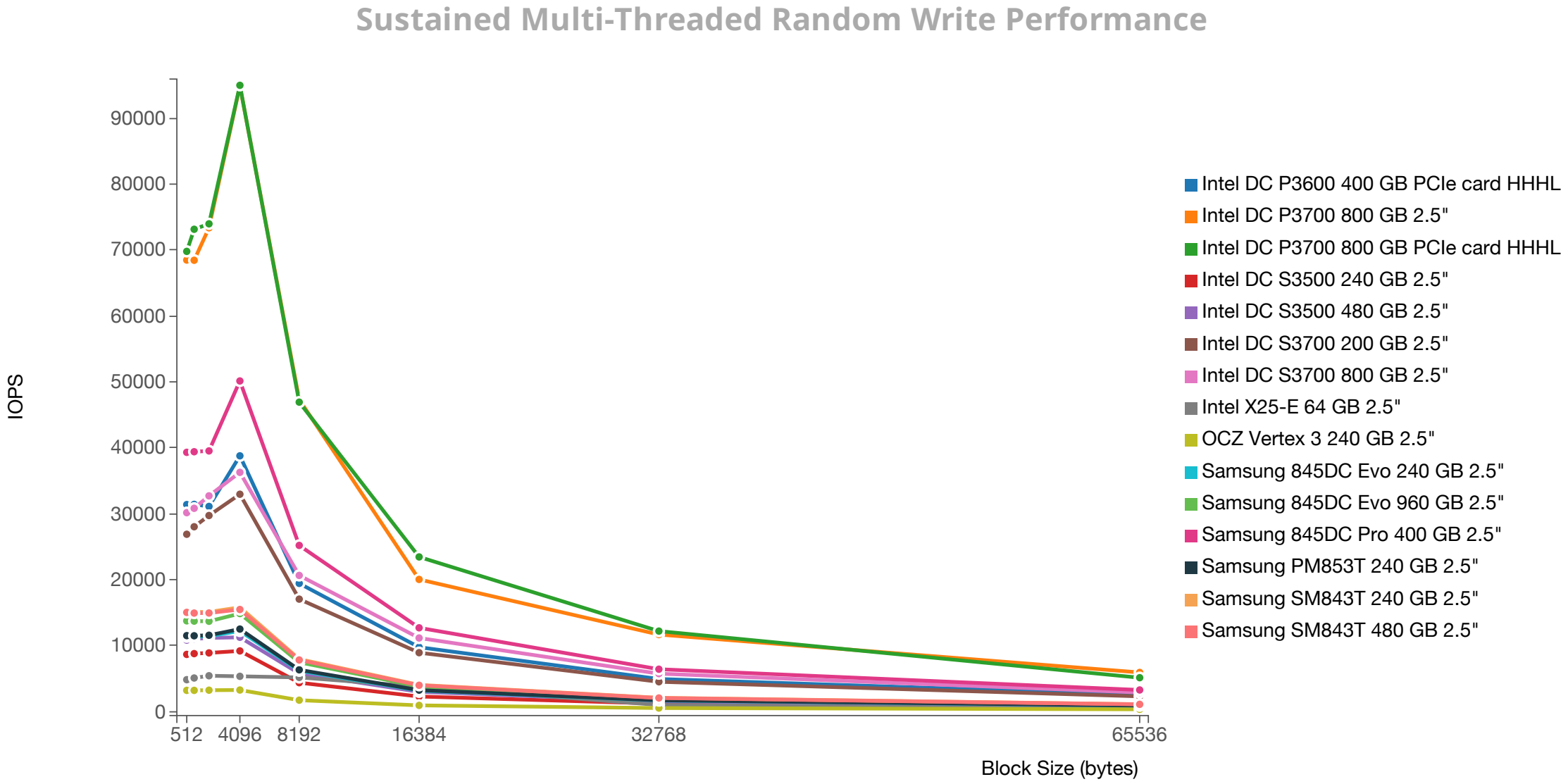


Sustained Multi-Threaded Random Read Throughput

Sustained Multi-Threaded Random Read Bandwidth

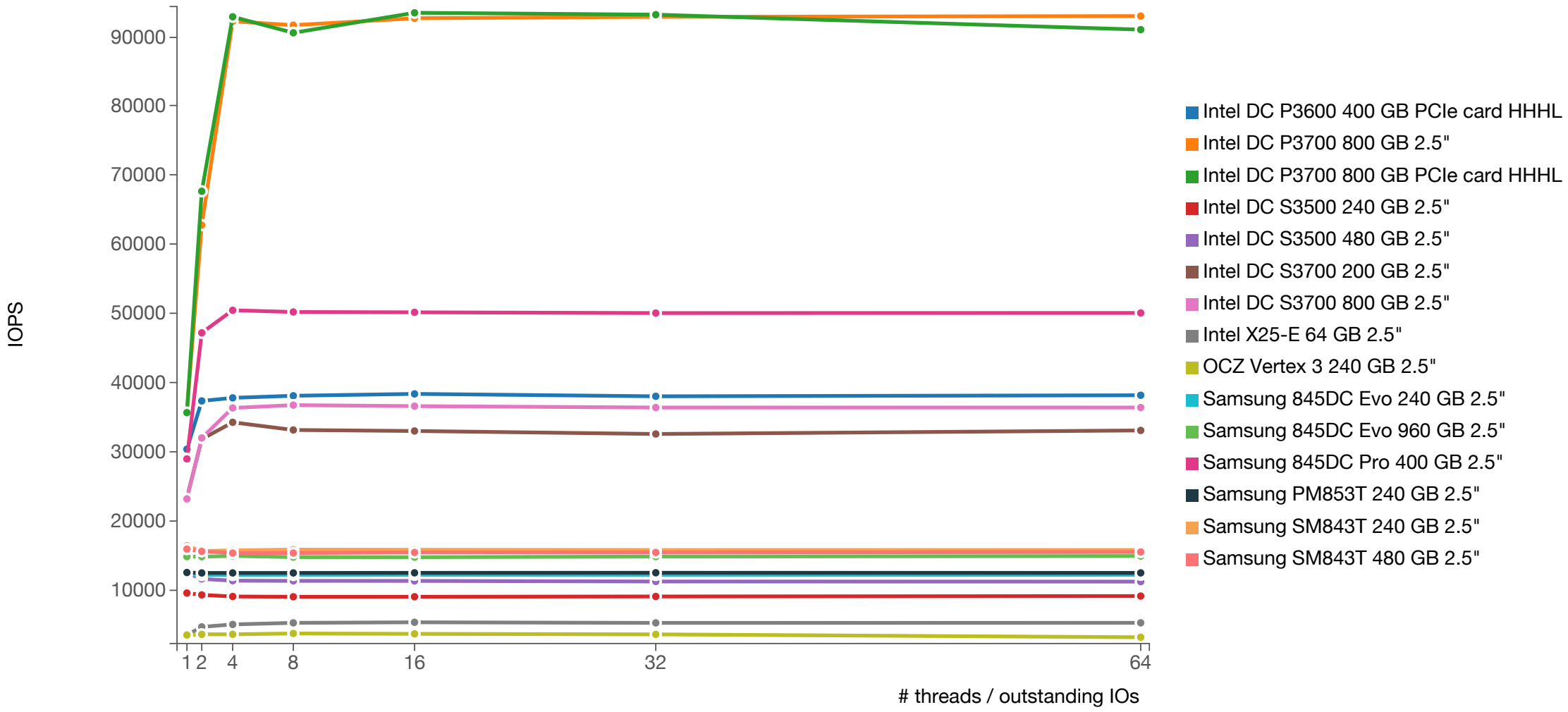


Sustained Multi-Threaded Random Write Performance

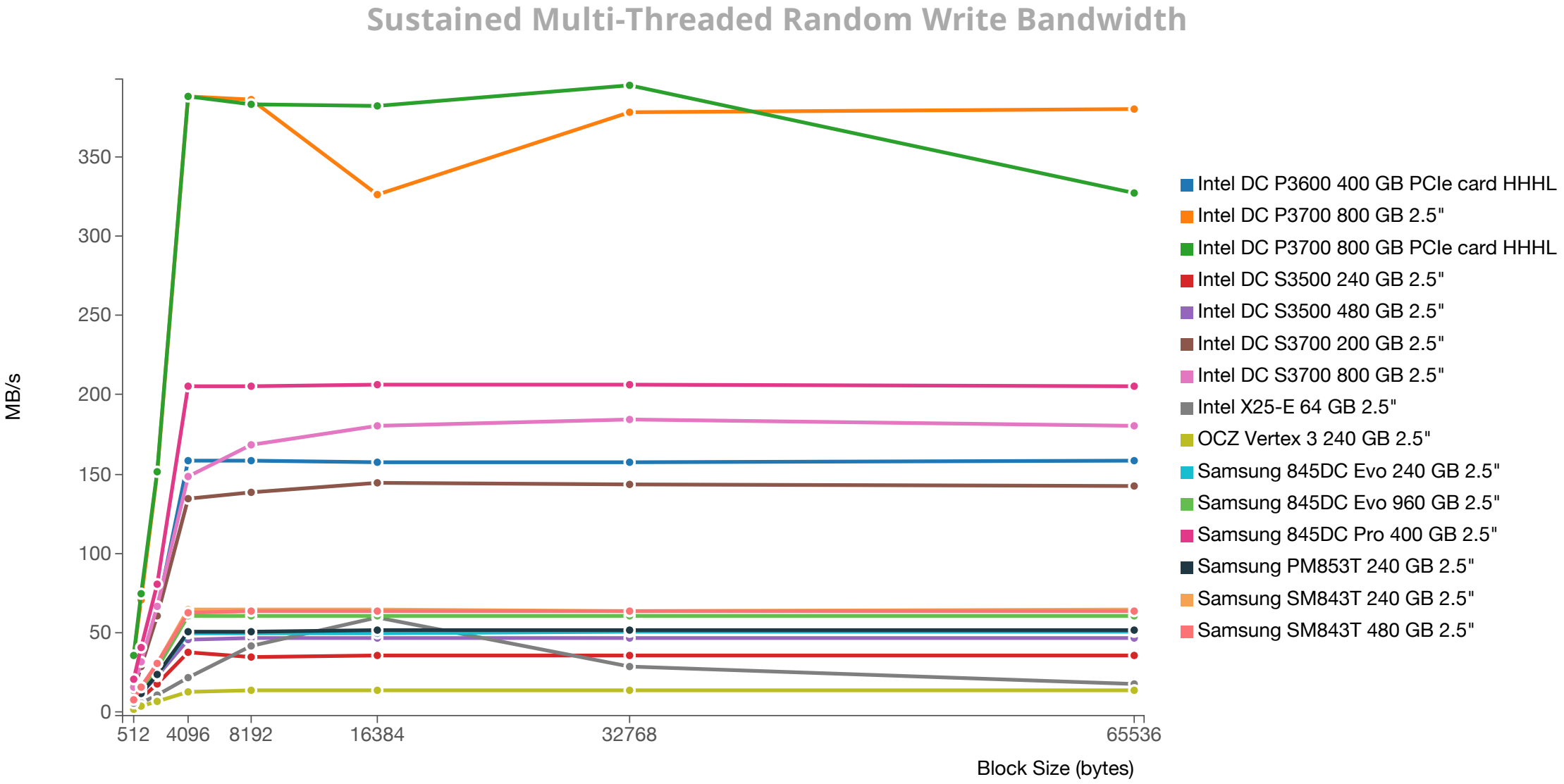


Sustained 4 KB Random Write Performance

Sustained 4 KB Random Write Performance by # of Threads

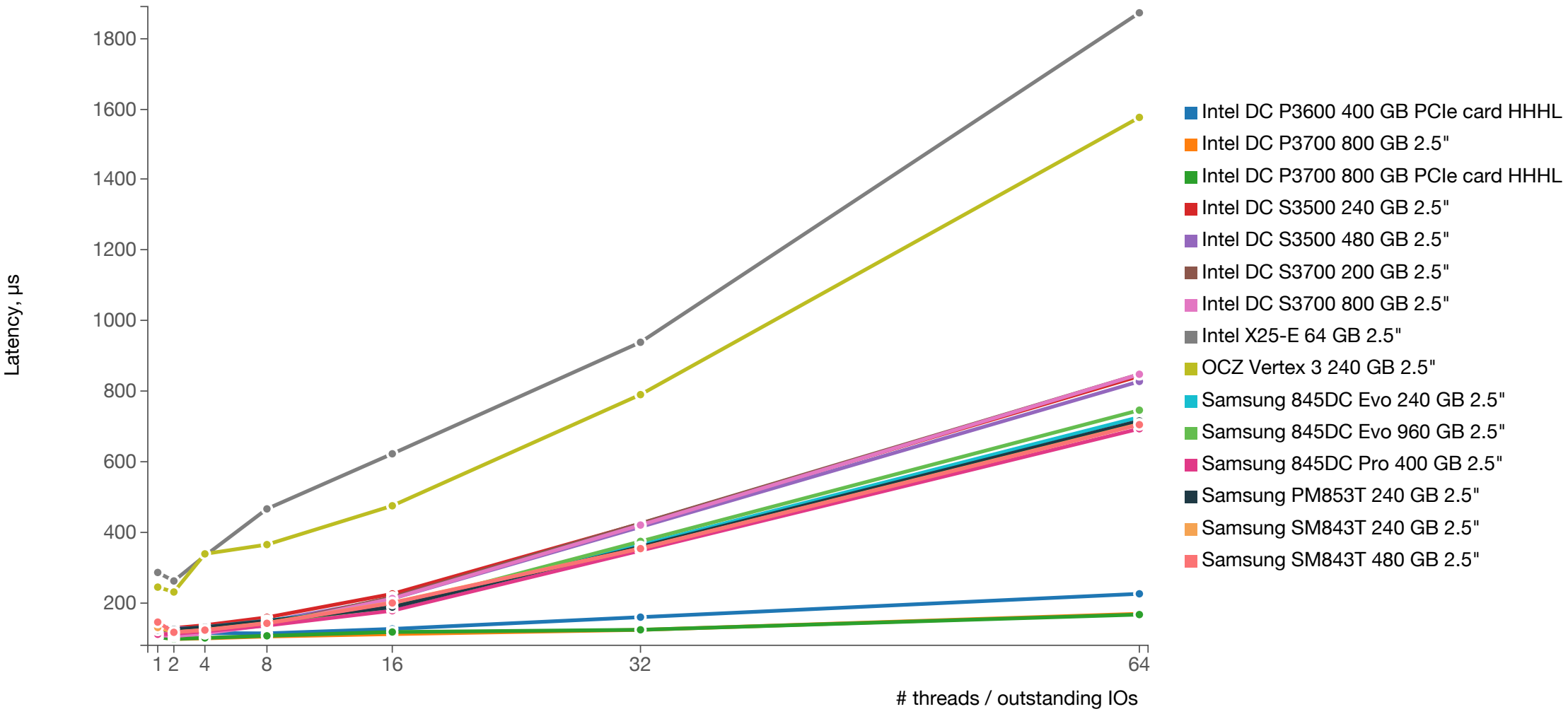


Sustained Multi-Threaded Random Write Bandwidth



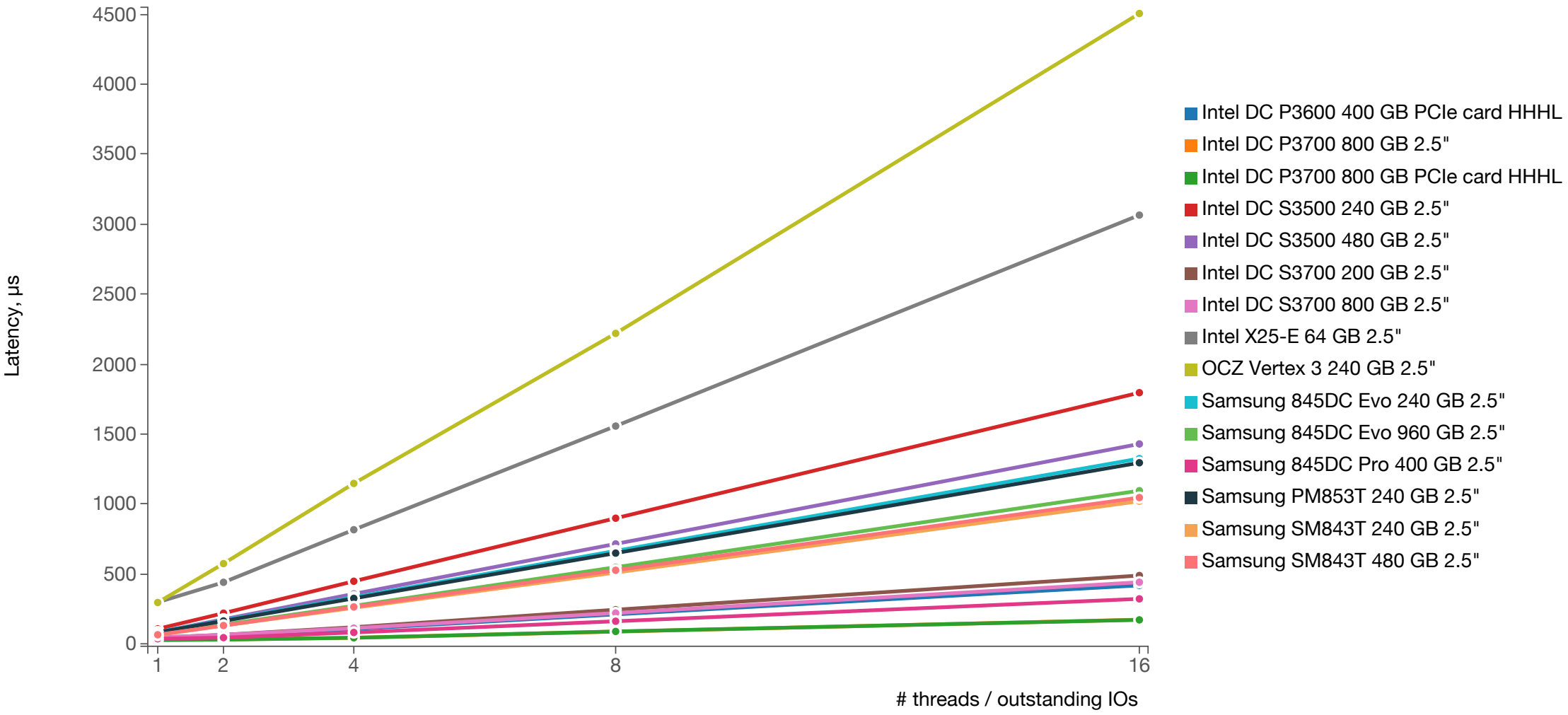
Sustained 4 KB Random Read Latencies

Sustained 4 KB Random Read Latencies by # of Threads



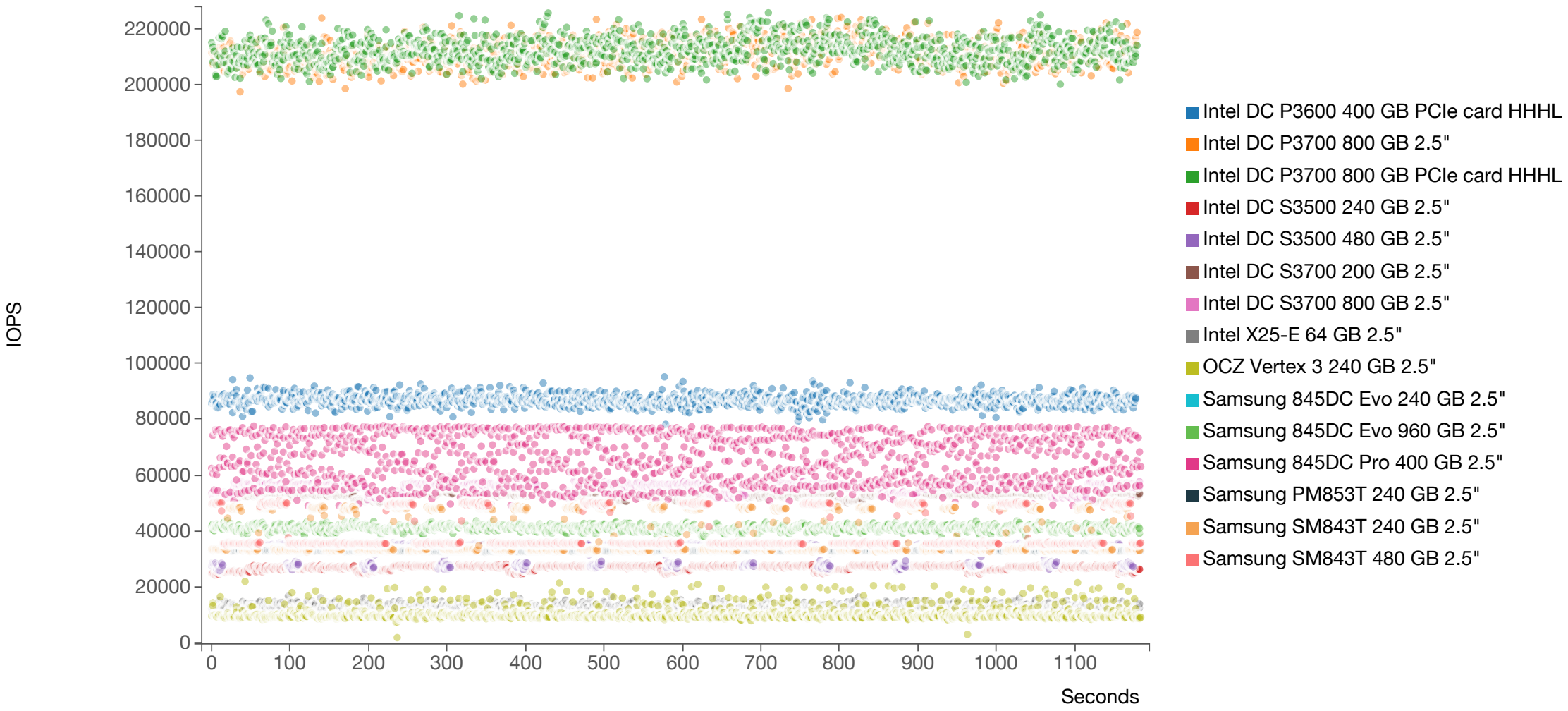
Sustained 4 KB Random Write Latencies

Sustained 4 KB Random Write Latencies by # of Threads

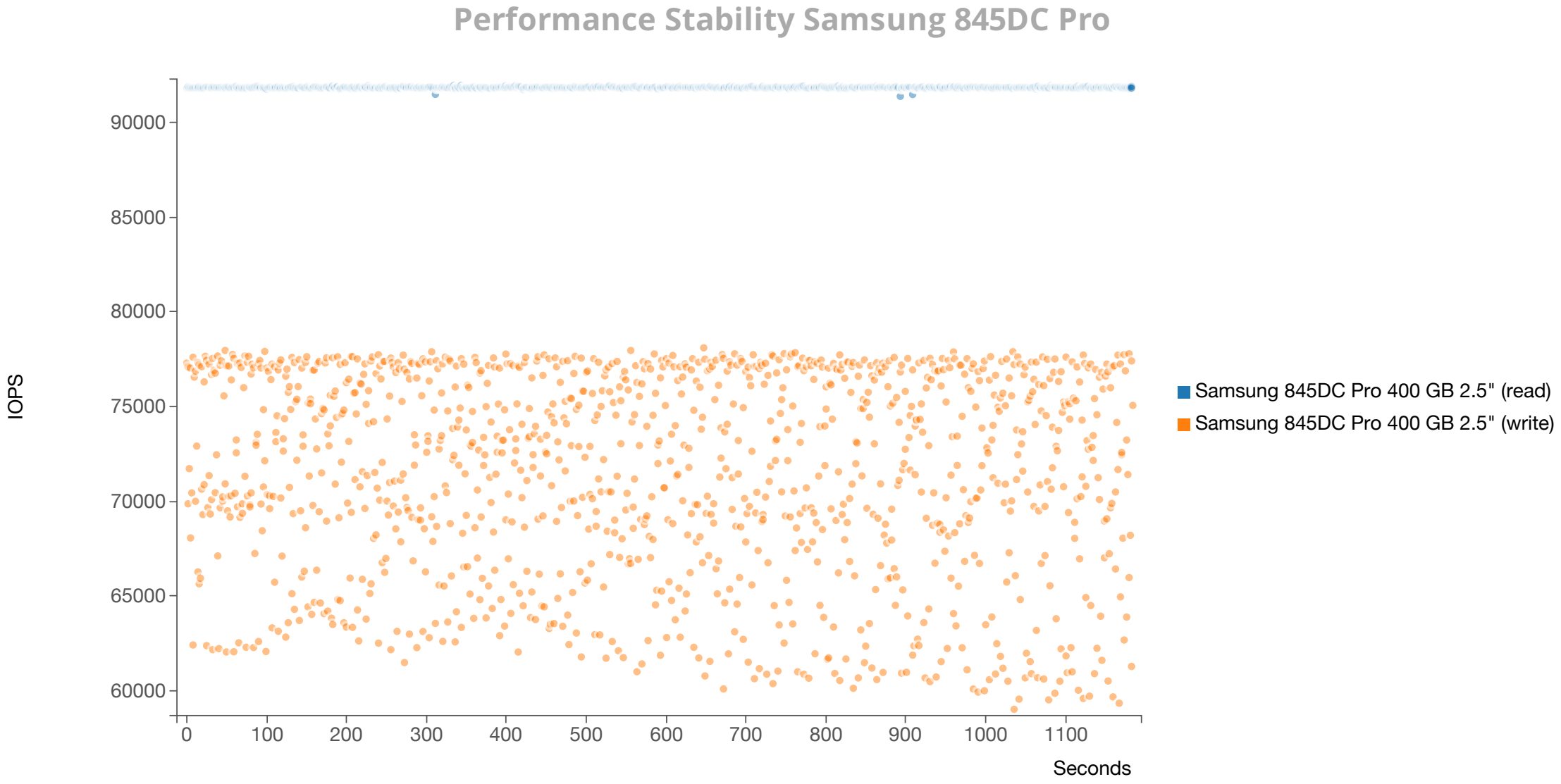


Performance Stability -- 4 KB Random Mixed 70% Read / 30% Write

Performance Stability -- Sustained 4 KB Random Mixed 70% Reads / 30% Reads Performance



Performance Stability Samsung 845DC Pro



Real-Work Application Use Cases

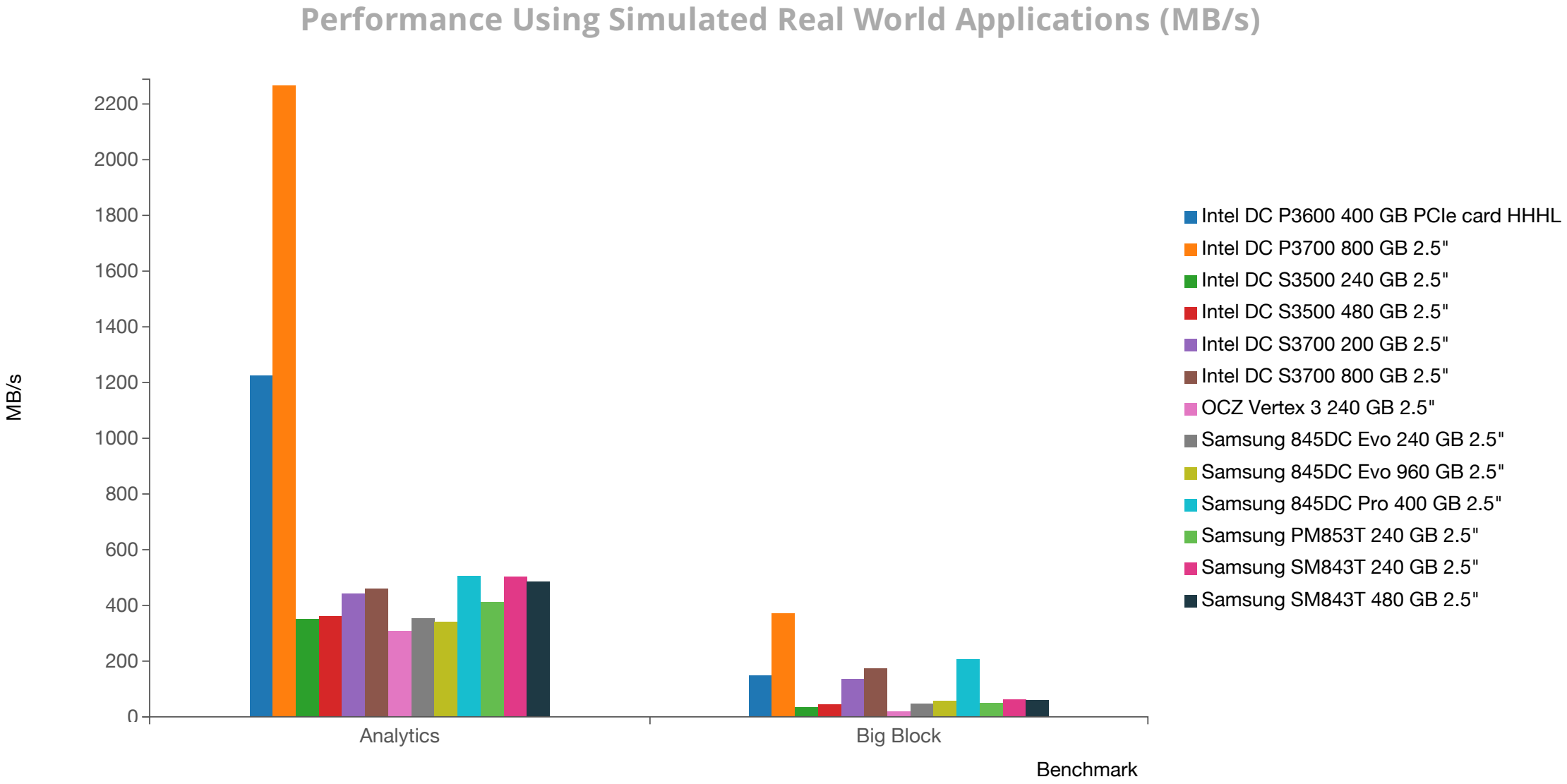
Large Blocks Analytics Engine

- Simulates a large-block analytics engine that is capable of streaming in data at a very high rate, sequentially scanning with multiple readers and a low (but non zero) update rate.
- Block size: 128 KB
- Number of jobs: 1
- Writes: 10%
- IO depth: 129
- Access pattern: sequential writes

Big Block

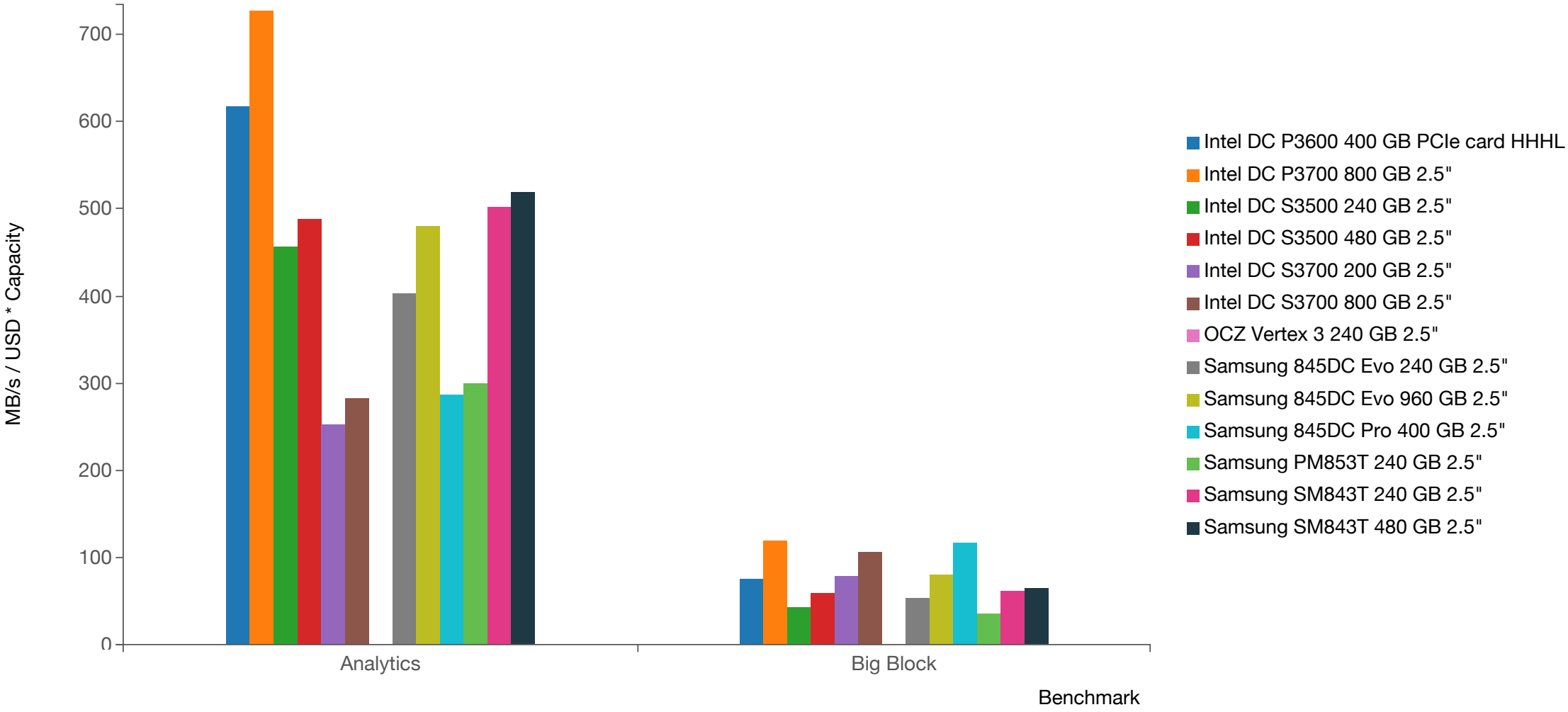
- Simulates a large-block, aggregated checkpointing application where sequential checkpoint writes from many different systems in a HPC cluster are aggregated at a single node (turning those nice sequential streams into an effectively random write setup).
- Block size: 512 KB
- Number of jobs: 16
- Writes: 100%
- IO depth: 16

Performance using simulated Real Work applications (MB/s)



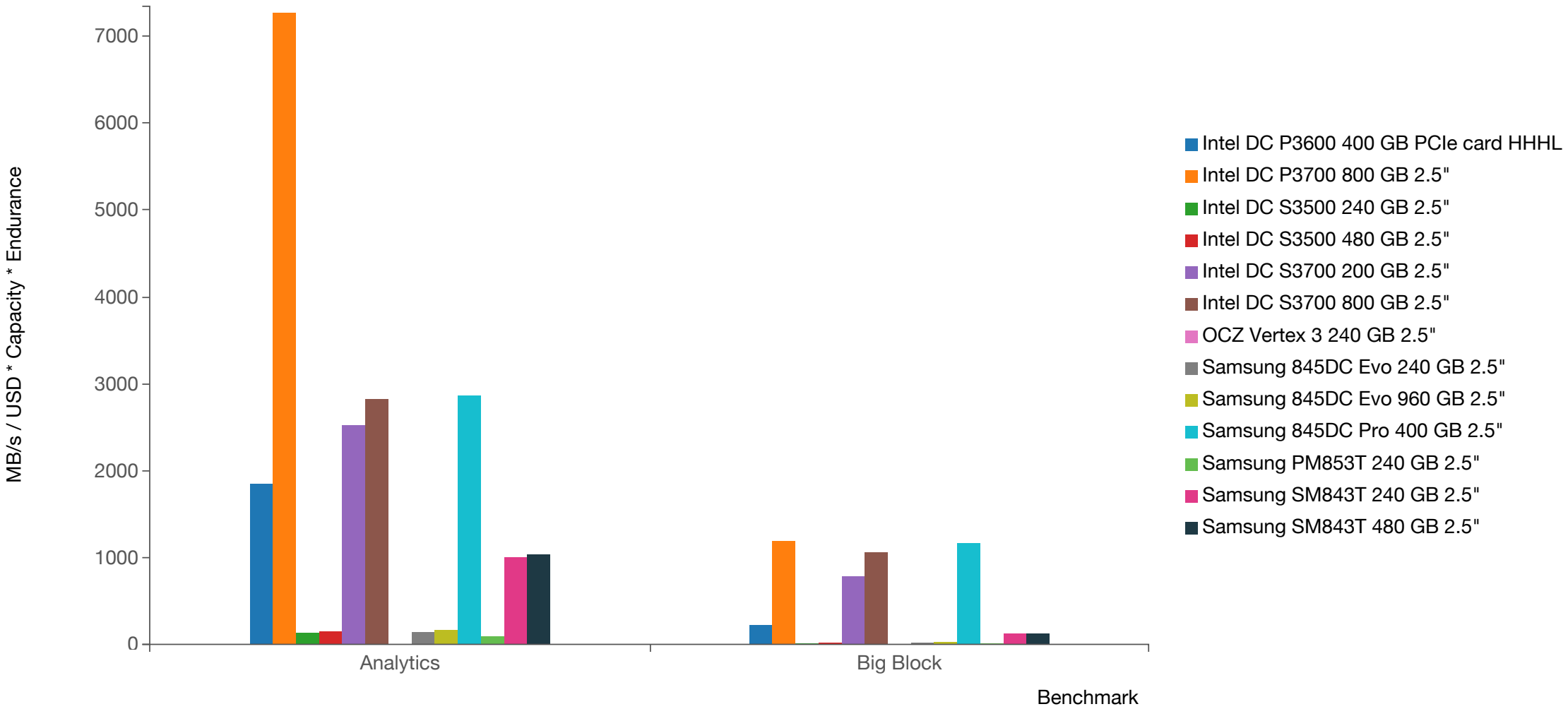
Throughput / USD * Capacity Using Simulated Real Work Applications

Throughput / USD * Capacity Using Simulated Real World Applications (MB/s / USD * Capacity)



Throughput / USD * Capacity * Endurance Using Simulated Real Work Applications

Throughput / USD * Capacity * Endurance Using Simulated Real World Applications (MB/s / USD * Capacity * Endurance)



Checkpointing

- Similar to Big Block, but this time with small, 4K checkpoint chunks.
- Block size: 4 KB
- Number of jobs: 16
- Writes: 100%
- IO depth: 16
- Access pattern: random writes

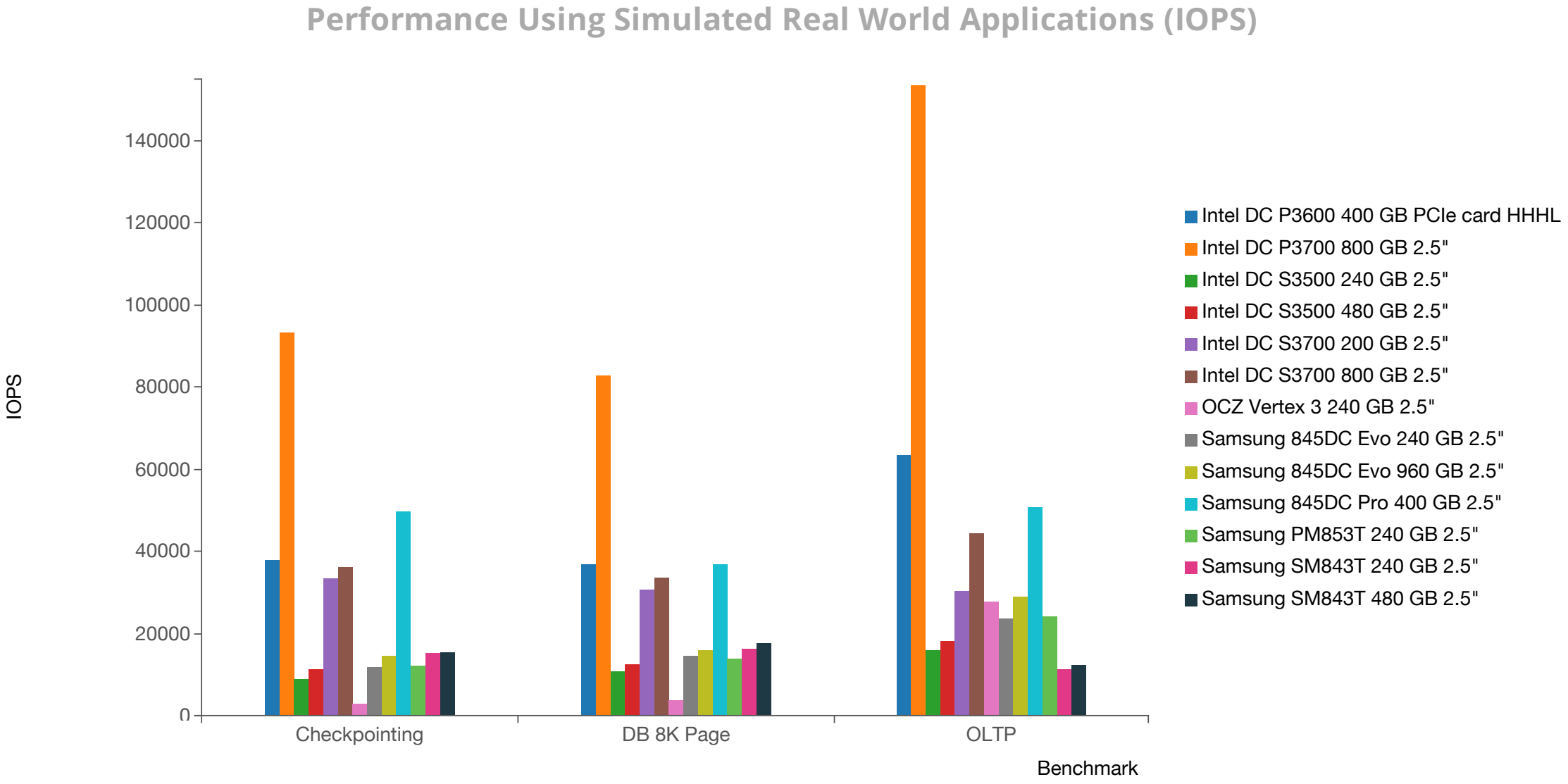
DB 8 KB

- Simulates a DB backend which uses 8 KB pages, without any logging or think time between reads/writes. This test stresses the IO subsystem without taking into account the locking, think times, etc. of a real DB.
- Block size: 8 KB
- Number of jobs: 8
- Writes: 40%
- IO depth: 4

OLTP

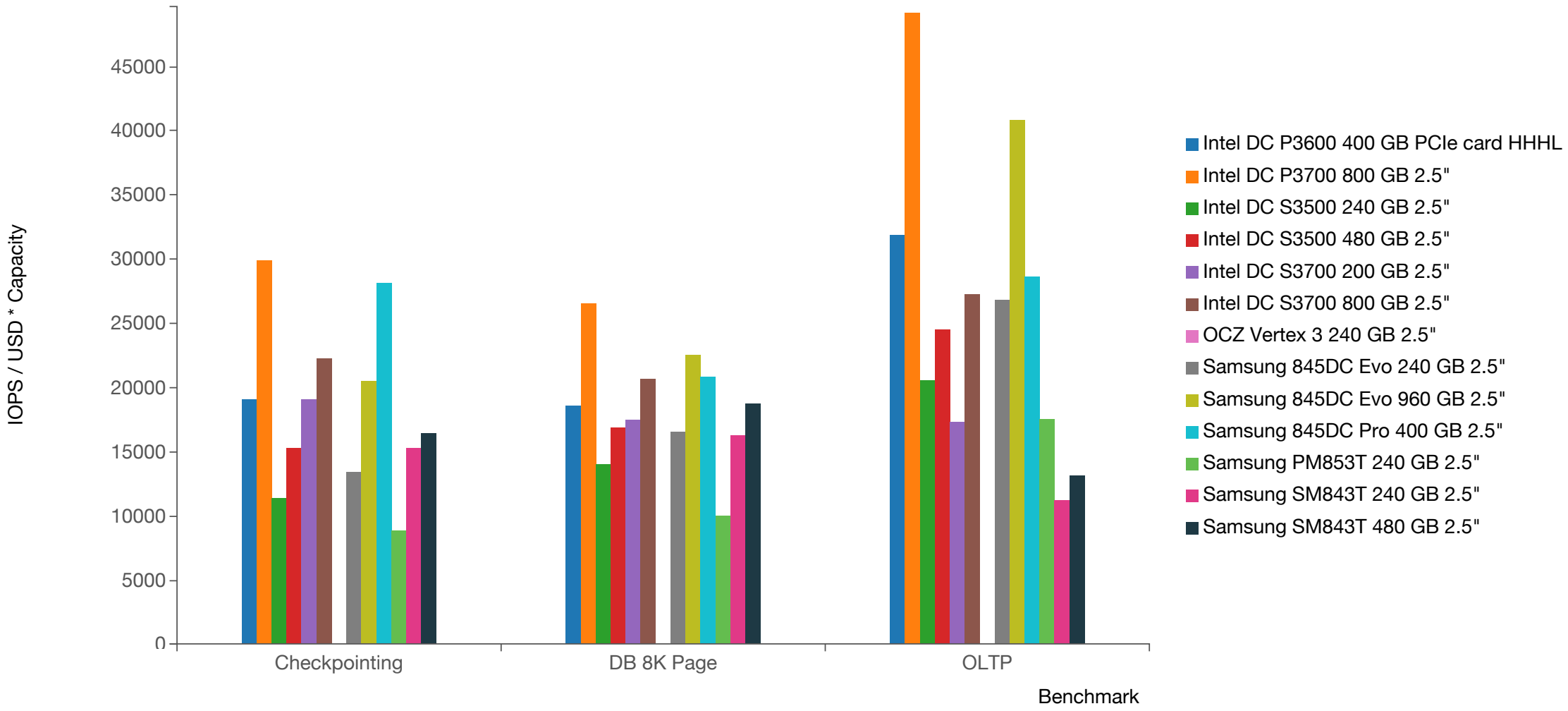
- Simulates an OLTP database pattern where the DB logs and data files are stored on the same SSD.
- Innosim is an InnoDB IO simulator that tests the disk IO capacity. Innosim mimics the workflow of a real DB (i.e. log writes are included and block transaction completions and data file updates).
- XFS file system used with a 4 KB block size

Performance Using Simulated Real Work Applications (IOPS)



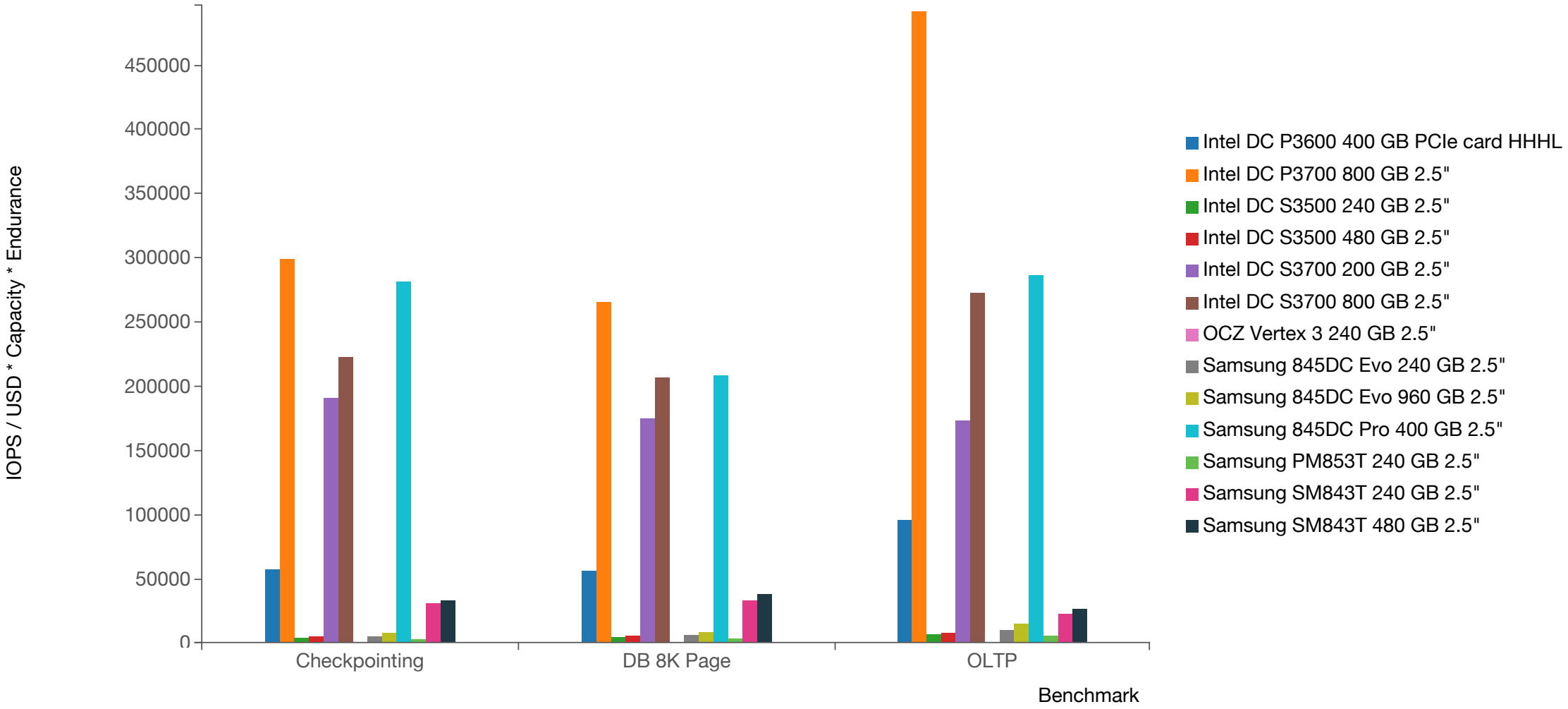
IOPS / USD * Capacity Using Simulated Real Work Applications

Performance / USD * Capacity Using Simulated Real World Applications (IOPS / USD * Capacity)



IOPS / USD * Endurance Using Simulated Real Work Applications

Performance / USD * Capacity * Endurance Using Simulated Real World Applications (IOPS / USD * Capacity * Endurance)



SSD Endurance

Intel S.M.A.R.T. Attributes

ID	Name	Interpretation
233	Media_Wearout_Indicator	**Normalized value:** reports the number of cycles the NAND media has undergone. Declines linearly from 100 to 1 as the average erase cycle count increases from 0 to the maximum rated cycles. Once the normalized value reaches 1, the number will not decrease, although it is likely that significant additional wear can be put on the device. **Raw value:** always 0.
241	Host_Writes_32MiB	**Normalized value:** always 100. **Raw value:** reports the total number of sectors written by the host system. The raw value is increased by 1 for every 65,536 sectors (32MB) written by the host.
226	Workld_Media_Wear_Indic	**Normalized value:** always 100. **Raw value:** measures the wear seen by the SSD (since reset of the workload timer, attribute E4h), as a percentage of the maximum rated cycles. Divide the raw value by 1024 to derive the percentage with 3 decimal points.
228	Workload_Minutes	**Normalized value:** always 100. **Raw value:** measures the elapsed time (number of minutes since starting this workload timer).

SSD Endurance

Samsung S.M.A.R.T. Attributes

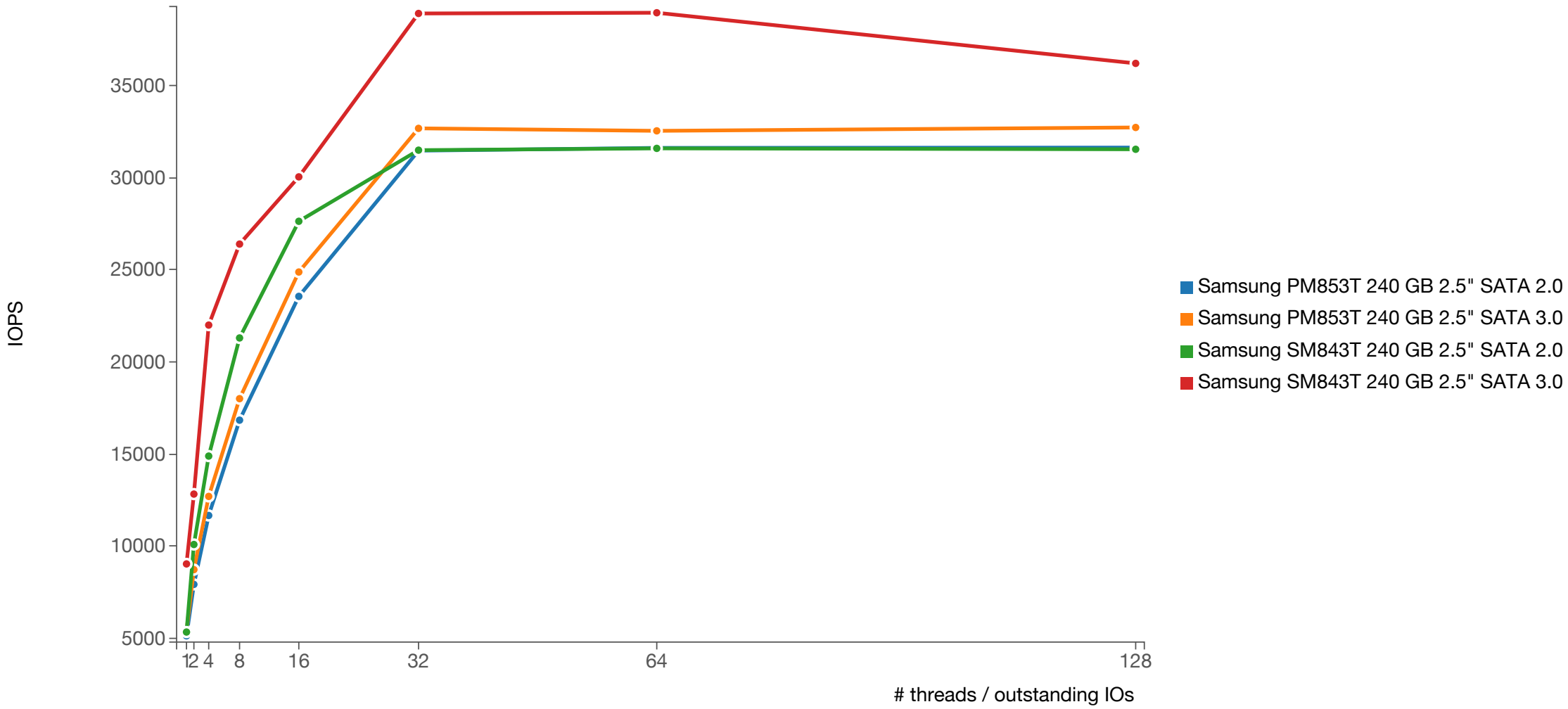
ID	Name	Interpretation
177	Wear_Leveling_Count	**Normalized value:** reports the number of cycles the NAND media has undergone. Declines linearly from 99 to 1 as the average erase cycle count increases from 0 to the maximum rated cycles. Once the normalized value reaches 1, the number will not decrease, although it is likely that significant additional wear can be put on the device. **Raw value:** the total count of P/E cycles.
241	Total_LBAs_Written	**Normalized value:** always 100. **Raw value:** the total size of all LBAs (Logical Block Address) required for all of the write requests sent to the SSD from the OS. To calculate the total size (in Bytes), multiply the raw value of this attribute by 512.

Hurdles for SSD adoption

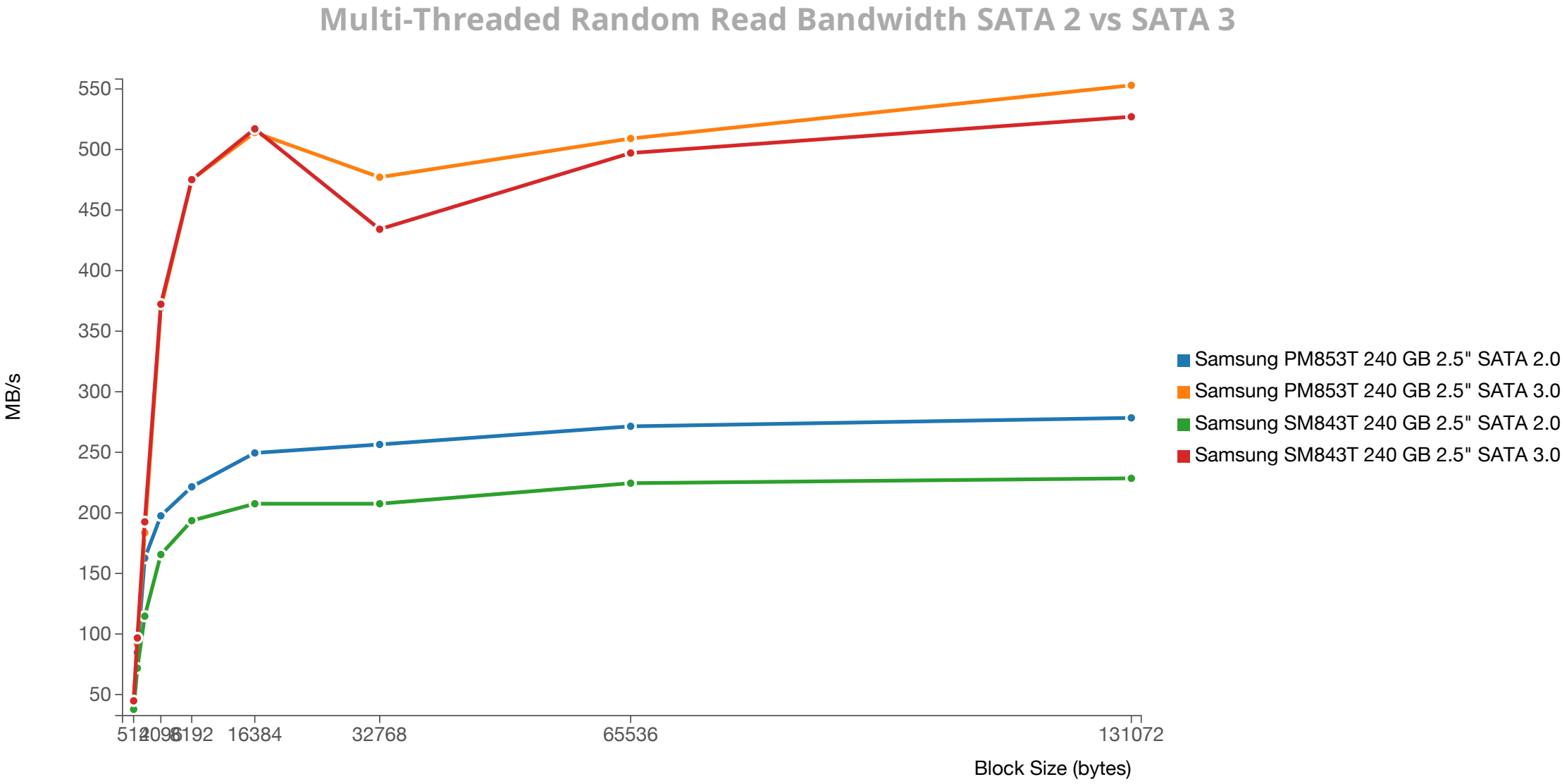
- Still a significant price difference compared to HDDs
- Still a gap in capacities compared to HDDs
- Almost no servers on the market with SATA Express / SFF-8639
- Currently no NVMe support in smartmontools, hdparm
- Many servers are still only SATA revision 2.0
- Ancient smartmontools version in SL 6 makes monitoring challenging

4 KB Random Mixed 70% Read / 30% Write - SATA 2 vs SATA 3

Sustained 4 KB Random Mixed 70% Reads / 30% Writes Performance by # of Threads



Multi-Threaded Random Read Bandwidth - SATA 2 vs SATA 3



Bogus smartmontools attribute names on SL 6

```
smartctl 5.43 2012-06-30 r3573 [x86_64-linux-3.14.5-1.el6.elrepo.x86_64] (local build)
Vendor Specific SMART Attributes with Thresholds:
ID# ATTRIBUTE_NAME          FLAG     VALUE WORST THRESH TYPE      UPDATED  WHEN_FAILED RAW_VALUE
170 Unknown_Attribute       0x0033   100   100   010    Pre-fail  Always       -       0
171 Unknown_Attribute       0x0032   100   100   000    Old_age   Always       -       0
172 Unknown_Attribute       0x0032   100   100   000    Old_age   Always       -       0
174 Unknown_Attribute       0x0032   100   100   000    Old_age   Always       -       1
175 Program_Fail_Count_Chip 0x0033   100   100   010    Pre-fail  Always       -       82173035128
183 Runtime_Bad_Block       0x0032   100   100   000    Old_age   Always       -       0
190 Airflow_Temperature_Cel 0x0022   079   075   000    Old_age   Always       -       21 (Min/Max 15/29)
192 Power-Off_Retract_Count 0x0032   100   100   000    Old_age   Always       -       1
194 Temperature_Celsius     0x0022   100   100   000    Old_age   Always       -       21
199 UDMA_CRC_Error_Count    0x003e   100   100   000    Old_age   Always       -       0
225 Load_Cycle_Count        0x0032   100   100   000    Old_age   Always       -       284832
226 Load-in_Time            0x0032   100   100   000    Old_age   Always       -       65535
227 Torq-amp_Count          0x0032   100   100   000    Old_age   Always       -       4294967295
228 Power-off_Retract_Count 0x0032   100   100   000    Old_age   Always       -       65535
232 Available_Reservd_Space 0x0033   100   100   010    Pre-fail  Always       -       0
233 Media_Wearout_Indicator 0x0032   099   099   000    Old_age   Always       -       0
234 Unknown_Attribute       0x0032   100   100   000    Old_age   Always       -       0
241 Total_LBAs_Written      0x0032   100   100   000    Old_age   Always       -       284832
242 Total_LBAs_Read         0x0032   100   100   000    Old_age   Always       -       486496
```

Recent smartmontools version

```
smartctl 6.3 2014-07-26 r3976 [x86_64-linux-3.14.5-1.el6.elrepo.x86_64] (local build)
Vendor Specific SMART Attributes with Thresholds:
ID# ATTRIBUTE_NAME          FLAG     VALUE WORST THRESH TYPE      UPDATED  WHEN_FAILED RAW_VALUE
170 Available_Reservd_Space 0x0033   100   100   010    Pre-fail  Always       -       0
171 Program_Fail_Count      0x0032   100   100   000    Old_age   Always       -       0
172 Erase_Fail_Count        0x0032   100   100   000    Old_age   Always       -       0
174 Unsafe_Shutdown_Count   0x0032   100   100   000    Old_age   Always       -       1
175 Power_Loss_Cap_Test     0x0033   100   100   010    Pre-fail  Always       -       632 (19 8681)
183 SATA_Downshift_Count    0x0032   100   100   000    Old_age   Always       -       0
190 Temperature_Case        0x0022   080   075   000    Old_age   Always       -       20 (Min/Max 15/29)
192 Unsafe_Shutdown_Count   0x0032   100   100   000    Old_age   Always       -       1
194 Temperature_Internal    0x0022   100   100   000    Old_age   Always       -       20
199 CRC_Error_Count         0x003e   100   100   000    Old_age   Always       -       0
225 Host_Writes_32MiB       0x0032   100   100   000    Old_age   Always       -       284834
226 Workld_Media_Wear_Indic 0x0032   100   100   000    Old_age   Always       -       65535
227 Workld_Host_Reads_Perc  0x0032   100   100   000    Old_age   Always       -       4294967295
228 Workload_Minutes        0x0032   100   100   000    Old_age   Always       -       65535
232 Available_Reservd_Space 0x0033   100   100   010    Pre-fail  Always       -       0
233 Media_Wearout_Indicator 0x0032   099   099   000    Old_age   Always       -       0
234 Thermal_Throttle        0x0032   100   100   000    Old_age   Always       -       0/0
241 Host_Writes_32MiB       0x0032   100   100   000    Old_age   Always       -       284834
242 Host_Reads_32MiB        0x0032   100   100   000    Old_age   Always       -       486499
```

Next Steps

- **Long term reliability measurements**
  - Performance degradation over time
- **Power measurements**
- **Finer grained monitoring of maximum latencies and standard deviations**
- **Improve monitoring of SSDs**
- **Evaluation of filesystem performance**
- **Evaluation of consumer drives**

Conclusions

- **Know your requirements**
  - Performance
  - Endurance
- **Know the limits of available SSDs**
- **Samsung SSDs provide better performance than Intel ones**
- **NVMe - significant performance improvements**
- **Think of developing SSD aware / friendly software**
- **Monitor the health of the drives**

One more thing…

Results available online at:

http://cern.ch/go/7x8W