固态硬盘 I/O 特性测试

周大

众所周知，固态硬盘是一种由闪存作为存储介质的数据库存储设备。由于闪存和磁盘之间物理特性的巨大差异，现有的各种软件系统无法直接使用闪存芯片。为了提供对现有软件系统的支持，往往在闪存之上添加一个闪存转换层来实现此目的。固态硬盘就是在闪存上附加了闪存转换层从而提供和磁盘相同的访问接口的存储设备。一方面，闪存本身具有独特的访问特性。另外一方面，闪存转换层内置大量的算法来实现闪存和磁盘访问接口之间的转换。因此和磁盘相比，固态硬盘具有复杂的访问特性。本测试集中在展示各种固态硬盘在不同访问模式下的 IOPS 值，同时和磁盘加以比较。本测试中，采取的访问模式有：随机读、连续读、随机写，连续写，以及分别加入地址跳跃、时间延迟、批量提交、延迟+批量提交、读写混合等影响因数。通过在各种访问模式下的实验结果可知，固态硬盘普遍具有较好的读性能和连续写性能，在随机写表现出复杂的特性。

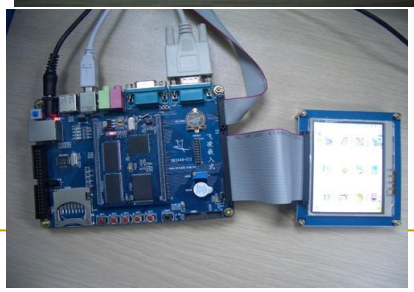
Understanding IO patterns of SSDs

Da Zhou

Outline

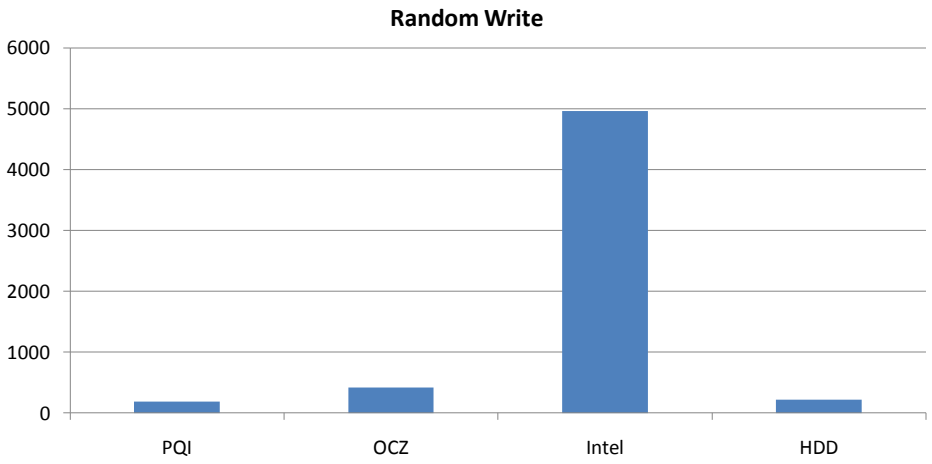
- Random/Sequential read/Write
- Address offset
- Delay
- Relay
- Burst
- Delay + Burst
- Semi Access
- Conclusion

Experiment Setup

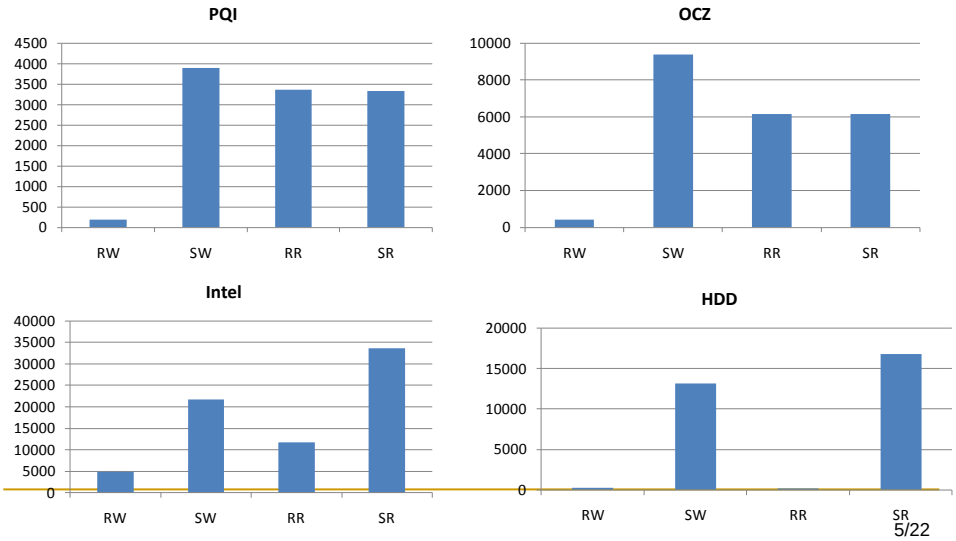


SSD:
Access granularity: 512 bytes
Test tool: IOMeter

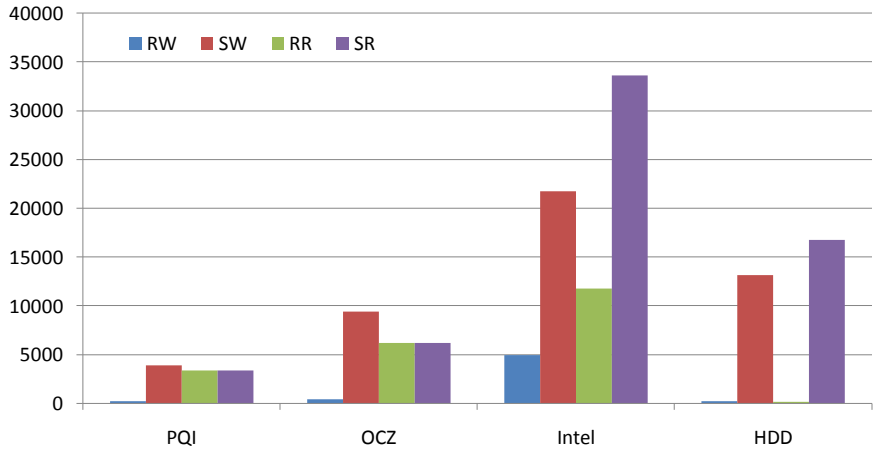
Random Write



Random Write



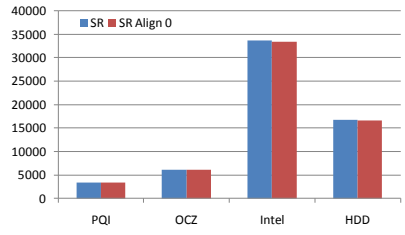
Random Write

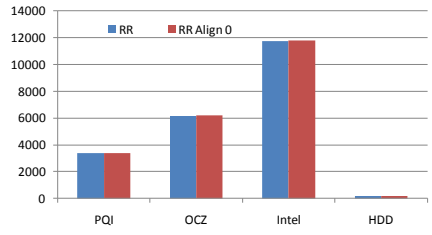


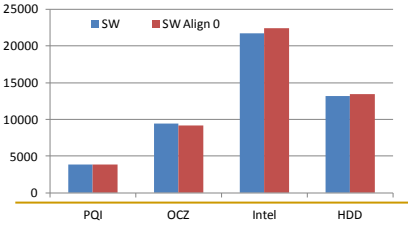
Hints

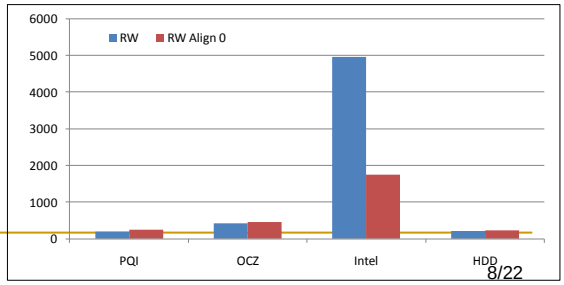
- Random write performance is low for most of SSDs
- New high-end SSD has high random write performance
- Random write is slower than sequential write.

Sequential Write VS Align Write

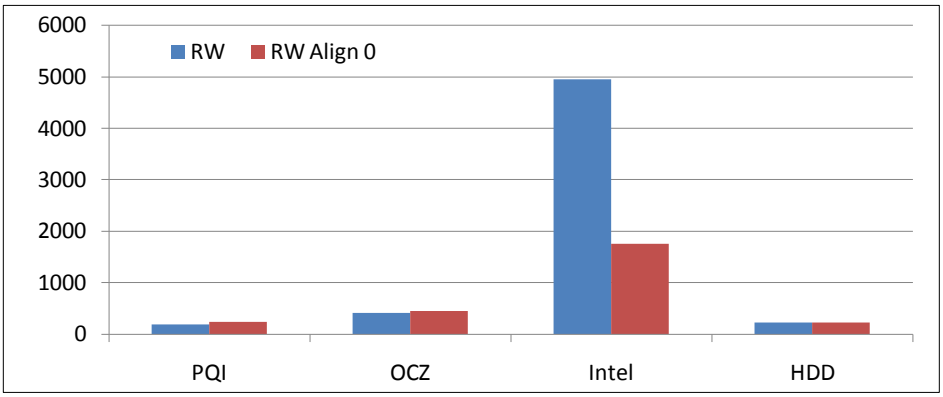






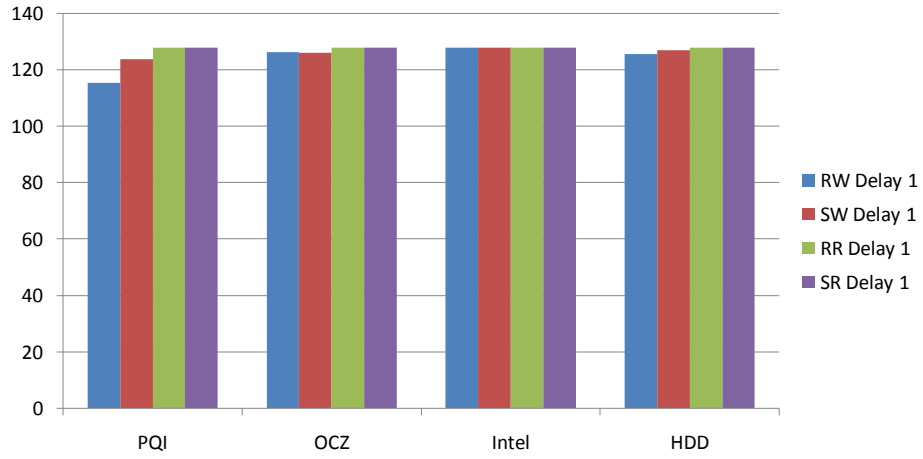


Random write VS Random write Align

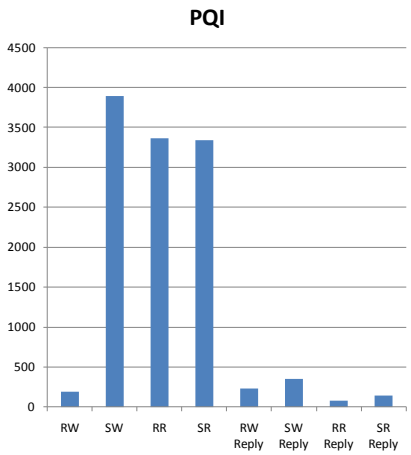


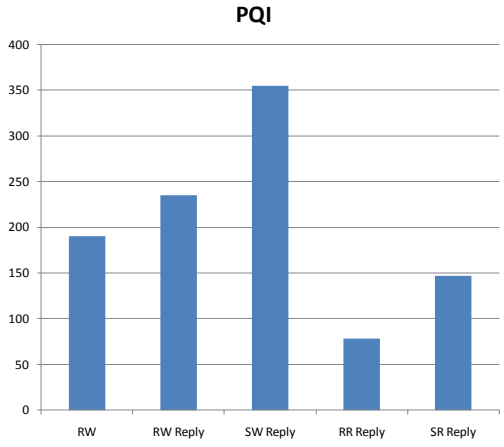
PQI, OCZ: directly written into flash memory chip
Intel: group written because of cache. Align leads to flush directly.

Delay

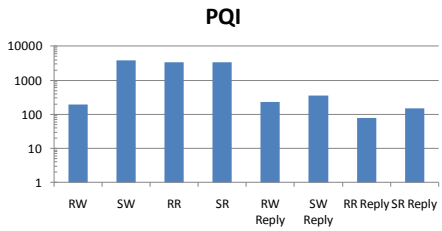


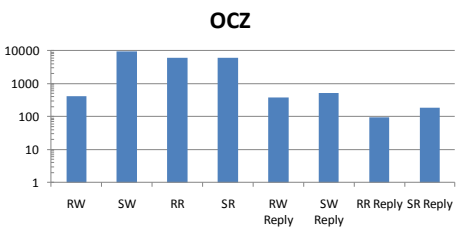
Replay

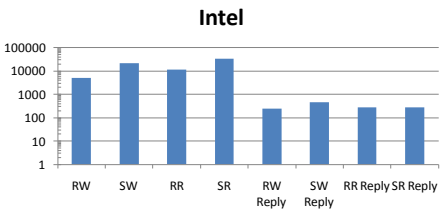


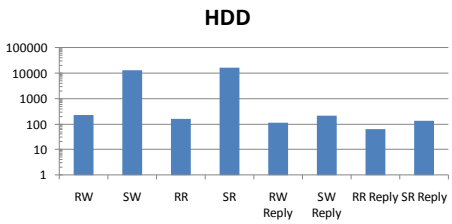


Write-Read

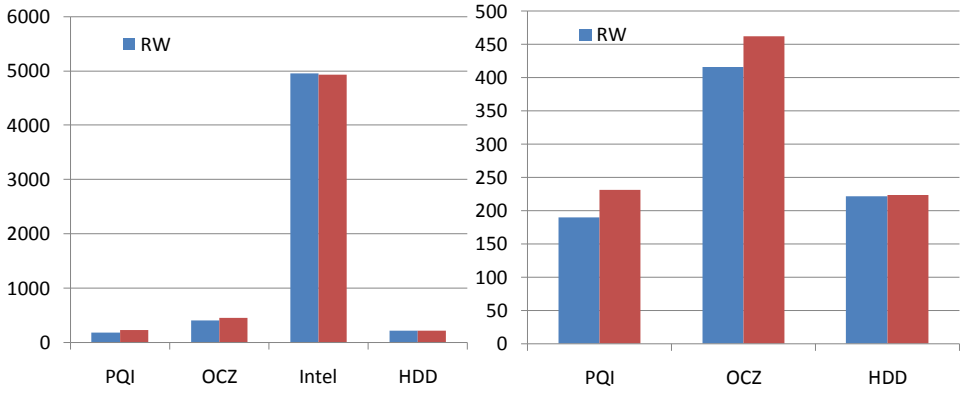








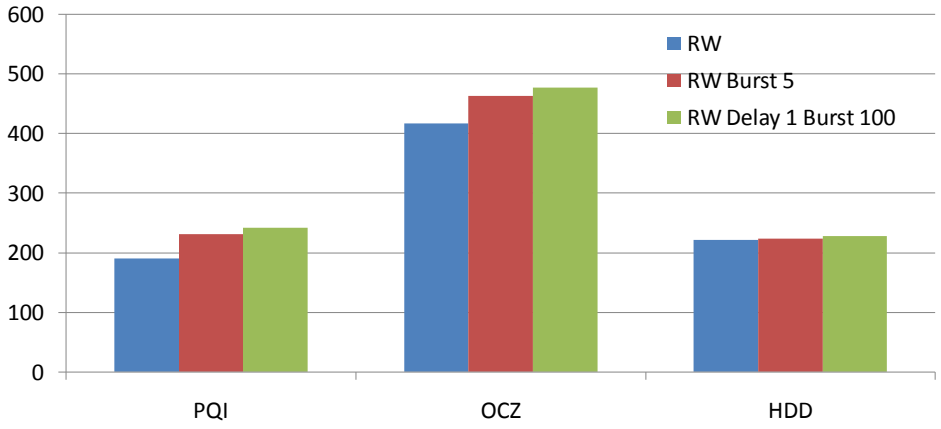
Burst 5



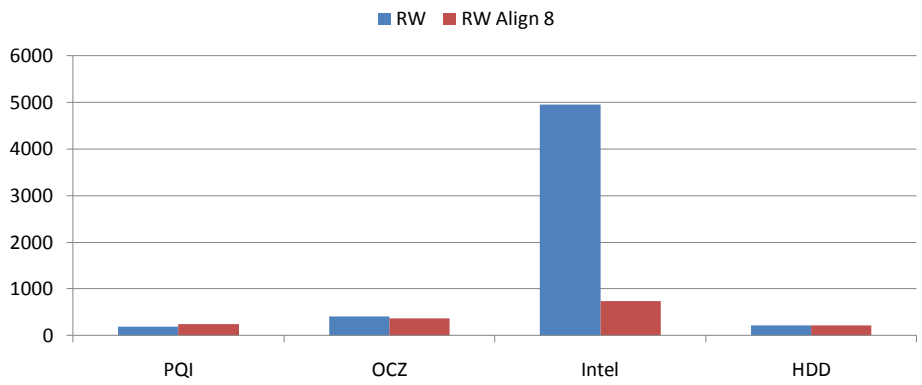
PQI, OCZ : Data are written into flash memory directly. While burst will improve the efficiency of Cache.

Intel, HDD: The capacity of cache is larger. More data are cached in RAM.

Burst and Delay

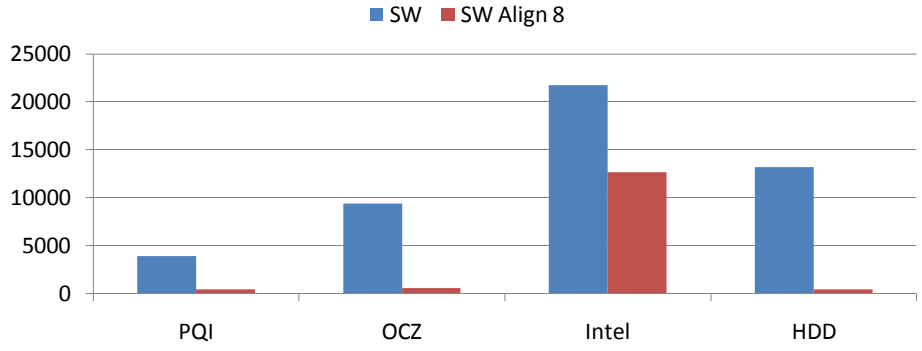


Random write alignment



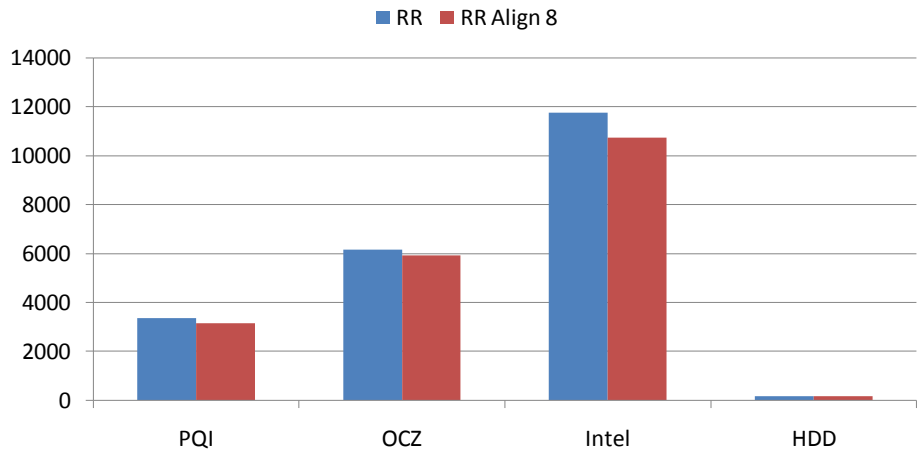
Intel: Data in the same 4kb area are flushed in the same time. When data do not locate in the same 4kb area, they will be flushed independently.
PQI, OCZ: data are flushed into SSD directly.

Sequential write alignment

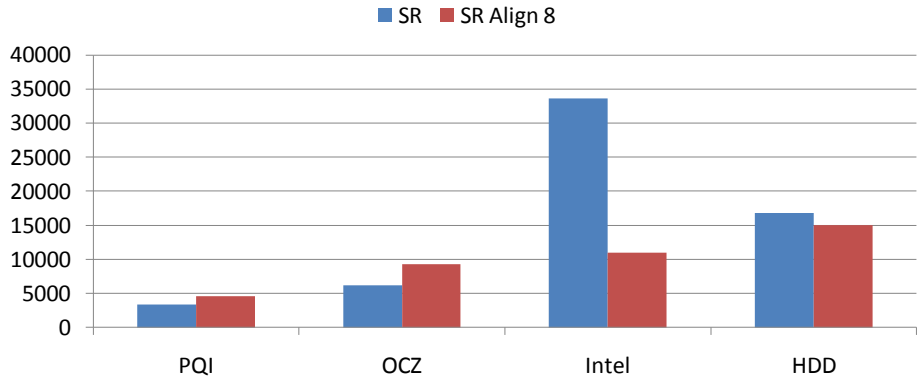


SSD: More erase operations will be triggered.
HDD: Sequential write ➔ random write

Random read alignment

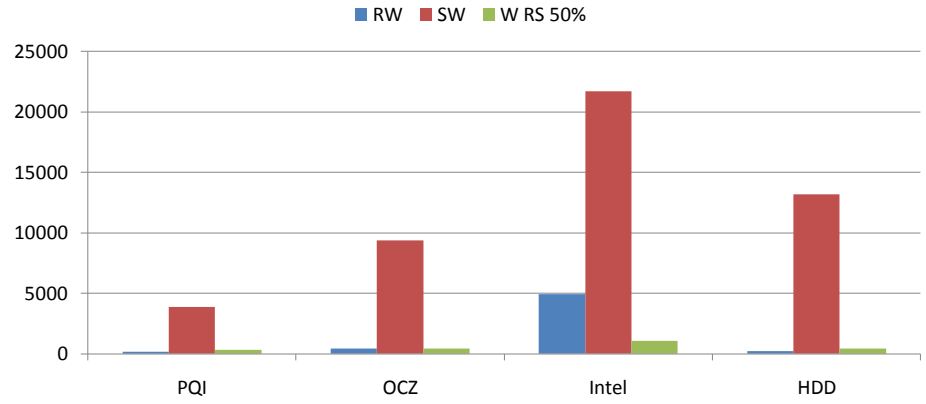


Sequential read alignment



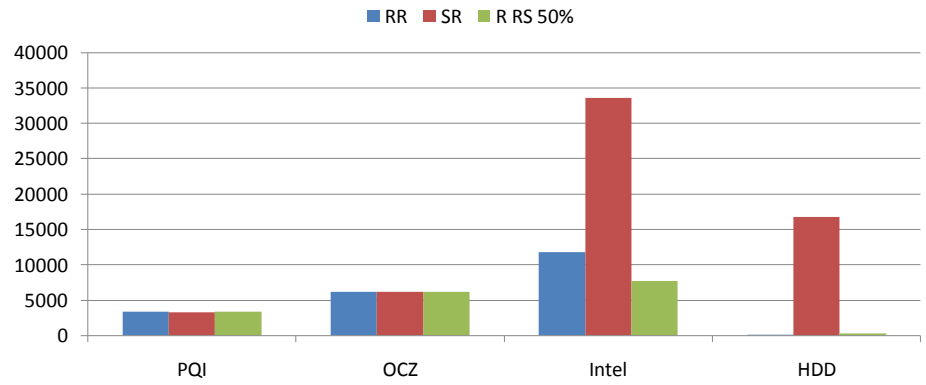
PQI, OCZ: ???
Intel: Prefetch is not utilized. Suppose the size is 4 KB, the remain data is needed.

Random Write VS Sequential Write VS Semi-Random write



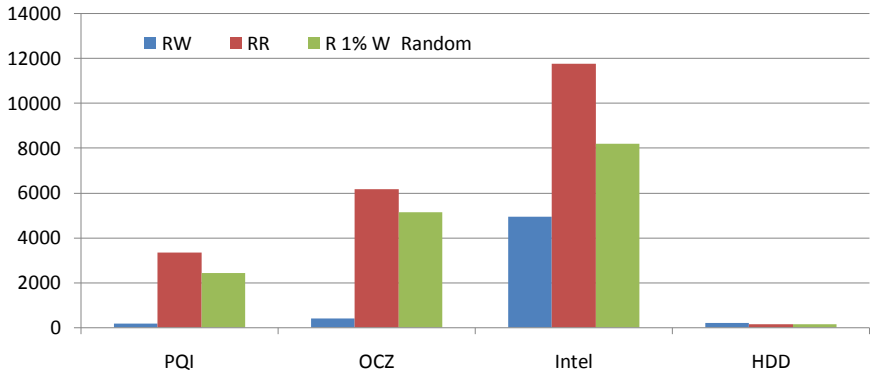
Intel: RW, SW: Cache; Read RS 50%: Data which is cached by RW is flushed by SW quickly. The reason maybe is the sequential write need a lot of cache. Or sequential write has higher priority than random write to use cache.

Random Read VS Sequential Read VS Semi-random Read



Intel: RR, SR: Prefetch; Read RS 50%: Data which is prefetched by RR is evicted by SR quickly. SR need more cache or high priority to use cache.

99% random Read, 1% random write



Write affects read

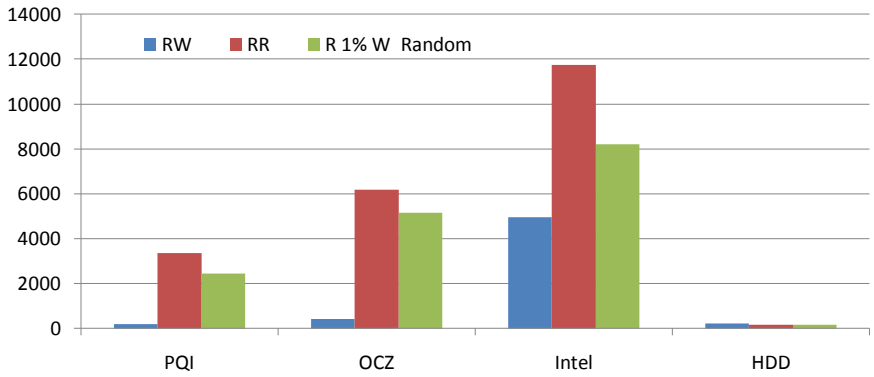
PQI	
RW	190.2775
RR	3360.399
R 1% W Random	2441.968

1 Write = 3360/190 reads= 18 reads

Ideal Read: 100/118= 84%

Real Read: 2441/3360= 72%

99% random Read, 1% random write



Problem is more obvious

Intel	
RW	4954.68
RR	11753.65
R 1% W Random	8200.149

1 Write = 11754/4954 reads= 2.4 reads

Ideal Read: 100/112.4= 97%

Real Read: 8200/11754= 69%