Understanding SSD Storage performance

Daniel Traynor, QMUL
HEPsysman June 2017, RAL

Aim

- Understand how different SSD storage hardware and technologies perform.

- Optimise hardware choice for environment (compute, bulk storage, cache).

- Optimise performance of OS for hardware.

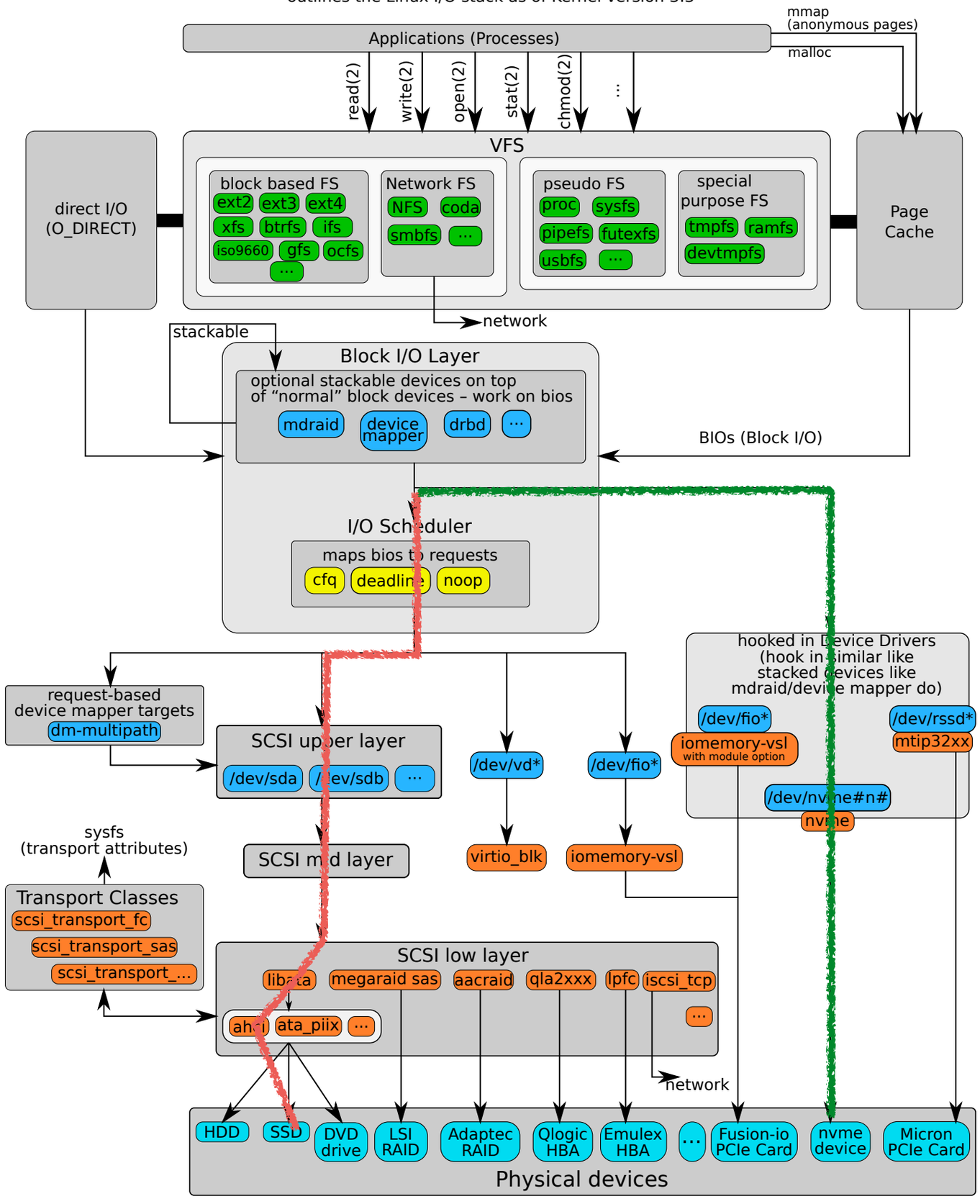
Talk Outline

- Understanding Linux IO (warning I’m no expert).
- SSD technology Overview (warning I’m no expert).
- Hardware and testing details (warning I’m supposed to be an expert).
- Interesting results.

Part 1

The Linux I/O Stack Diagram

version 1.0, 2012-06-20
outlines the Linux I/O stack as of Kernel version 3.3



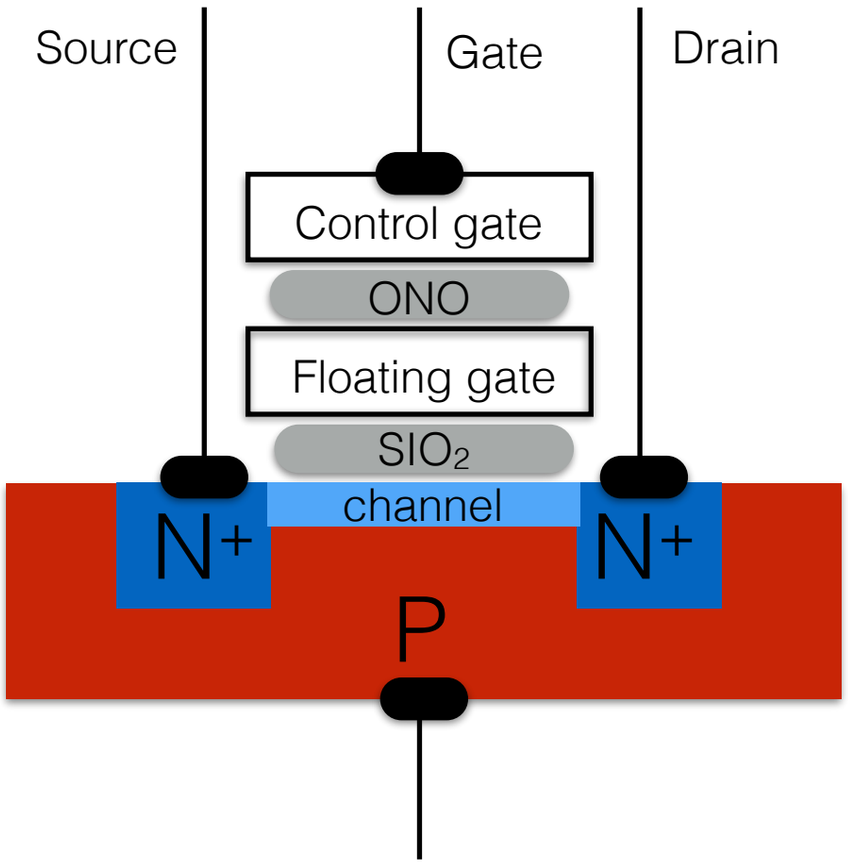
Page cache

SATA/SAS

NVMe

Disk buffer

Flash Cells



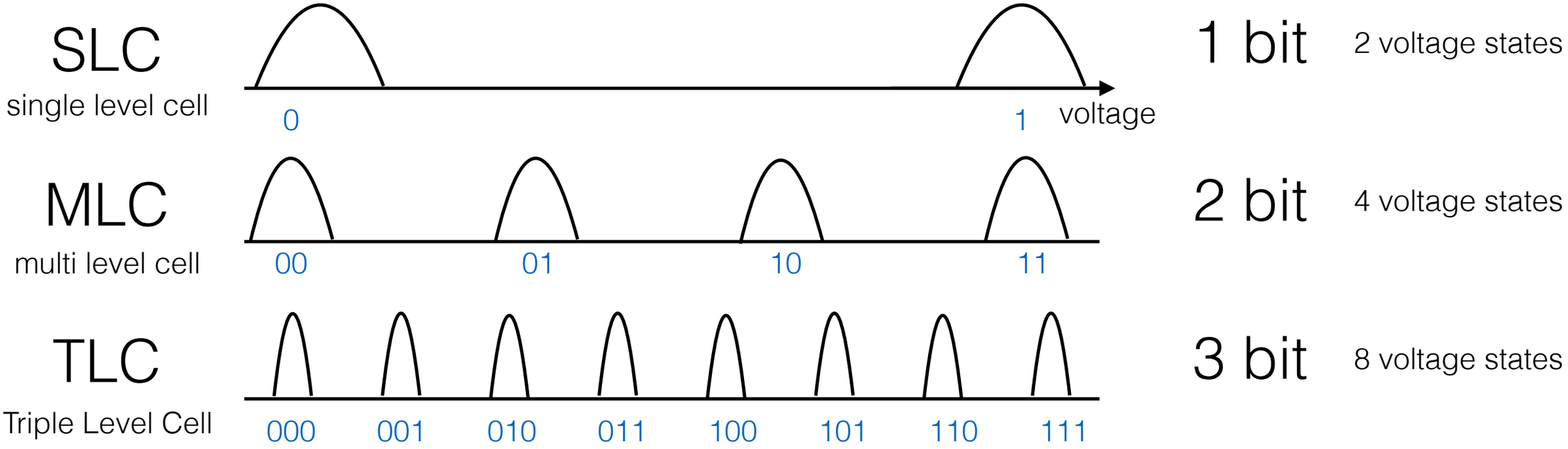
Floating Gate

MOSFET

Metal Oxide Semiconductor Field Effect Transistor

- A threshold Gate voltage is needed to make the channel conductive.
- When a correct voltage is applied to the control gate, electrons can “tunnel” to the floating gate. When the voltage is removed the electrons are isolated in the floating gate.
- The presence of the electrons create a screening effect which changes the threshold voltage needed to make the channel conductive.
- Electrons are removed by applying large voltage to the substrate (P).
- The floating gate isolation brakes down with erase cycles leading to a change in the screening effect, threshold voltage and wear.

SLC, MLC,TLC



- [x] Adding more voltage states adds more data per cell.
- [ ] However, it takes longer to process (read and write). Increased latency.
- [ ] Becomes more sensitive to voltage changes due to wear. Lower endurance.

Performance

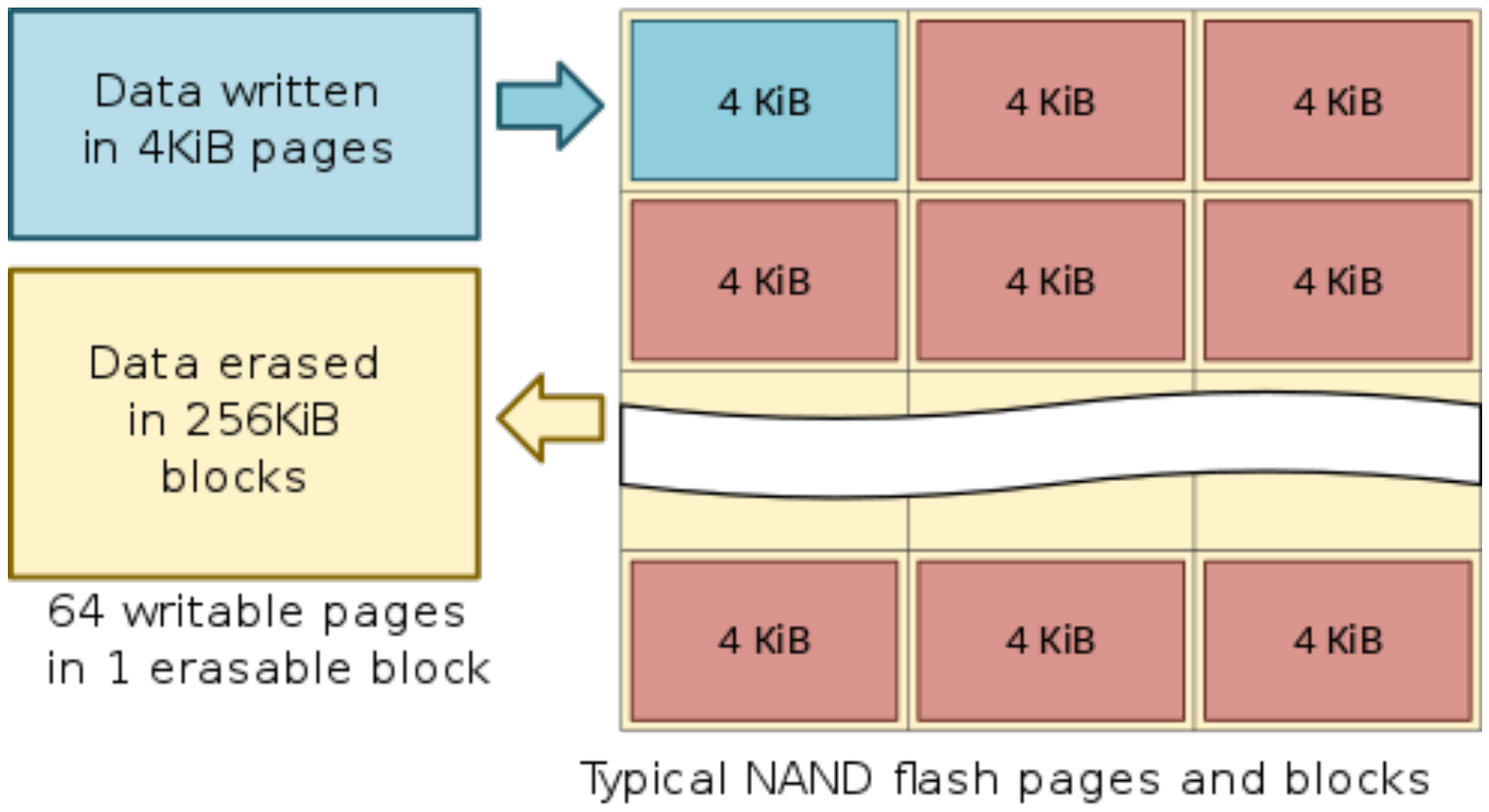
	SLC	MLC	TLC	HDD	RAM
P/E cycles	100k	10k	5k	*	*
Bits per cell	1	2	3	*	*
Seek latency (µs)	*	*	*	9000	*
Read latency (µs)	25	50	100	2000-7000	0.04-0.1
Write latency (µs)	250	900	1500	2000-7000	0.04-0.1
Erase latency (µs)	1500	3000	5000	*	*
Notes	* metric is not applicable for that type of memory				

https://www.extremetech.com/extreme/210492-extremetech-explains-how-do-ssds-work

SSD Operation

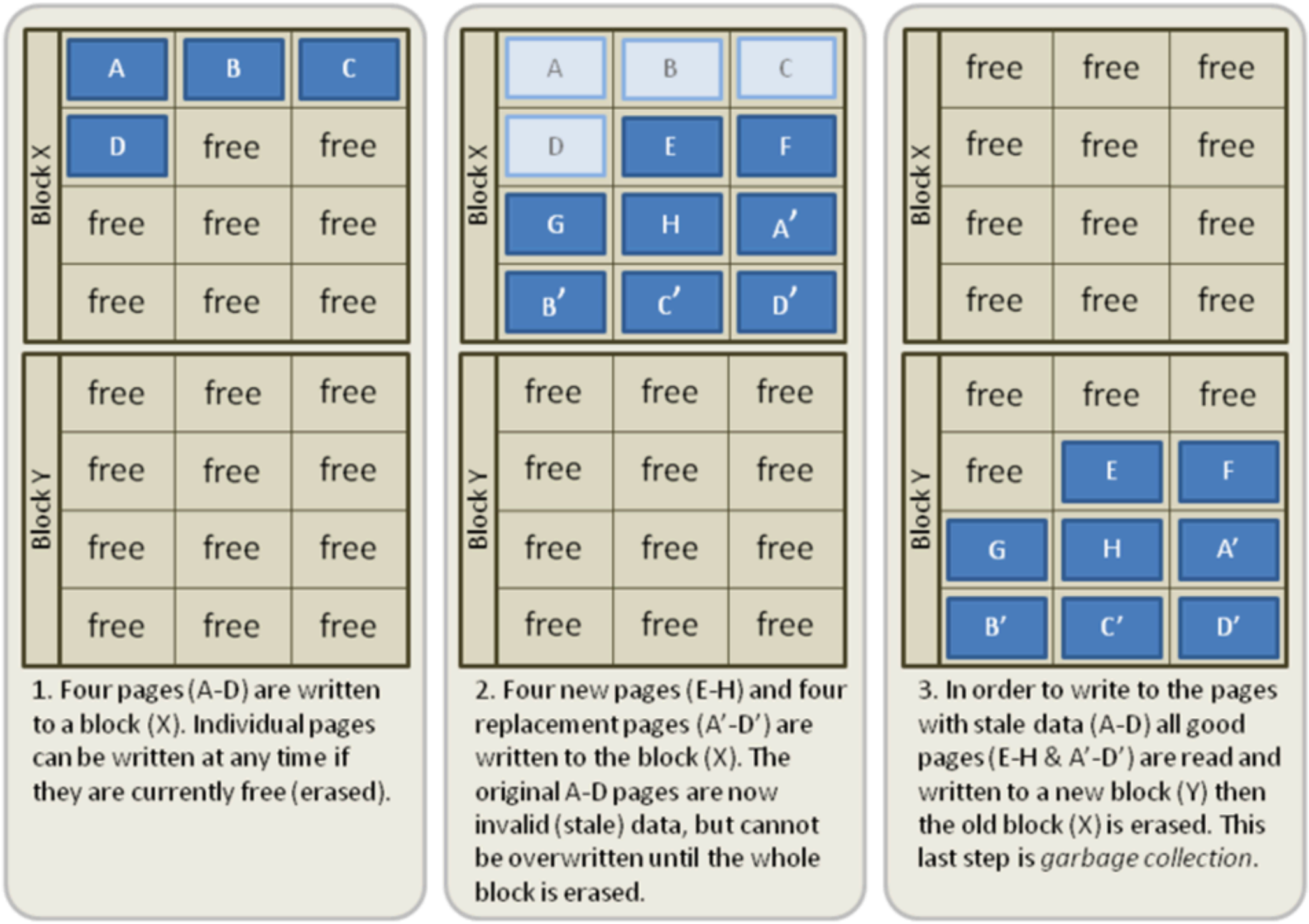
Cells organised into pages (2K-16K cells to a page)
pages into blocks (64-256 pages to a block)

SSDs read and write data at page level but erase data at block level (less stressful to cells)



https://en.wikipedia.org/wiki/Write_amplification

SSD Operation



https://en.wikipedia.org/wiki/Write_amplification

SSD Operation

- Write amplification
- Over provisioning
- Wear levelling
- Garbage collection
- TRIM
- Caches (RAM, SLC…)
- error correction
- ~~compression~~
- ……

All this means the controller ASIC/firmware much more sophisticated and performance impacting.

SSD Endurance

- Endurance limited by voltage changes due to program erase cycles (p/e) dominated by the erasure process. This needs to be converted into “English”.

- Tera bytes written (TBW)

- Drive writes per day (DWPD)

- Media Wear out Indicator (MWI) smart value used for SSD health monitoring. Not the same as TBW or DWPD.

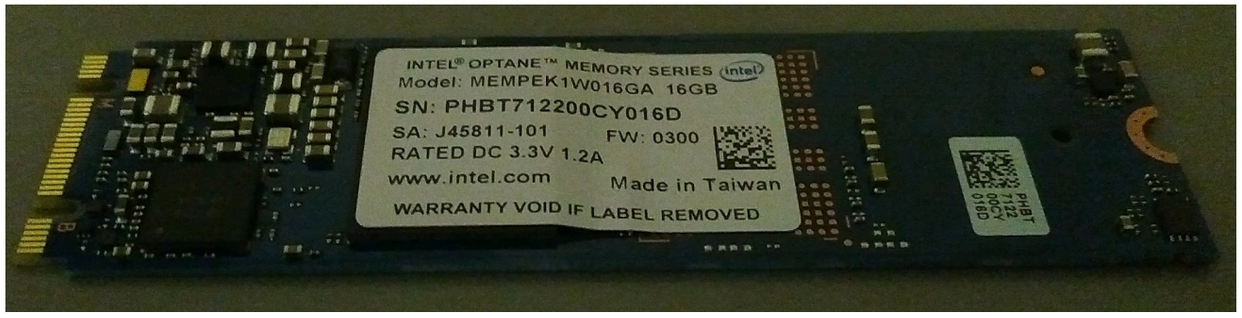
- Intel SSDs fail “hard” once they reach their MWI limit they go into readonly mode then on next reboot will not work. Have to mount as secondary disk to recover.

SSD Construction

NVMe M.2, PCI-E x 2 (B) / 4(M)

check notches

different lengths, single/double sided



NVMe U.2 hot swappable, with power loss protection (e.g capacitor)

https://rog.asus.com/articles/hands-on/easy-guide-to-ssds-sata-msata-m-2-and-u-2/

+ PCI-E card, SATA, SAS

Worker Node Requirements

Worker node disk writes ~ 75TB/year or 200GB/day

/sys/fs/ext4/dm-2/lifetime_write_kbytes

Typical 50GB of disk use (40 cores), rare to go above 250GB

Remember we use a network file system for data access (Lustre)

Almost all job IO which is transient and
data is deleted at the end of the job.
Mostly sequential writes
(expect little write amplification on SSD)

Requirements: size 512GB, endurance 400 TBW

Don’t worry about data loss form power or SSD failure,
no need for hot swap drives.

Bulk Storage Requirements

- Bulk storage, endurance not an issue,1 drive write per year. Expected to increase with changing write patterns.
  - Assume 10 writes per year. Still peanuts compared to lowest endurance SSDs (~1/2 DWPD for 7TB SSD).
- ZFS ZIL much higher level of writes. All data written to zpool written to ZIL doubles writes to bulk storage or dedicated ZIL.
  - 100TB zpool -> 5,000TBW or 3 dwpd (1TB SSD, 5 year).

CEPH Journal Requirements

- Typical example - per VM host 4 *1TB HD, Samsung SM863 120GB SSD - 5G for each journal on same SSD 30 VMs per server (endurance 1500TBW, 3.5DWPD over 10 years) seen 170TBW per year.

- 3* HPE DL360 HA PROXMOX VMhosts with CEPH (x2 copies) storage, on each host 6*1TB HDD + 2 * 400GB SSDs (2000TBW, 3 DWPD). brand new - no usage numbers

Part 2

Drive Specs

- Intel Optane 16GB M.2 NVMe: sequential read=900MB/s, sequential write=145MB/s, random read=190K IOPS, random write=35K IOPS, endurance = 182.5TBW (6DWPD). NOT based on FLASH technology!, PCI-E x2
- Intel 600p 512GB M.2 NVMe SSD: 17.5GB SLC write cache, sequential read=1,775 MB/s, sequential write=560MB/s, random Read=128.5K IOPs, random write =128K IOPS, endurance = 288TBW. PCI-E x4
- Crucial BX200 240GB SATA SSD: 3GB SLC write cache, sequential read=540MB/s, sequential write=490MB/s, random read=66K IOPS, random write=78K IOPS, endurance = 72TBW.
- HPE (Samsung PM1635) 400GB SAS SSD: 2GB RAM cache sequential read=950MB/s, sequential write=530MB/s, random read=190K IOPs, random write=50K IOPS, endurance 2,190TBW (3DWPD).(HPE DL360). Power loss protection for cache.
- Samsung 960 Pro 512GB NVMe SSD: 512MB RAM cache, sequential read=3,500 MB/s, sequential write=2,100MB/s, random read=330K IOPS, random write=330K IOPS, endurance = 400TBW. NO power loss protection for cache. PCI-E x4
- Samsung 850 Pro 256GB SATA SSD: 512MB RAM cache, sequential read=550MB/s, sequential write=520MB/s, random read=90K IOPS, random write=90K IOPS, endurance=150TBW. NO power loss protection for cache.
- HPE 500GB SATA 7.2H RPM HDD. expected sequential read/write ~150MB/s. expected random read/write~150 IOPS.
- Segate 8TB Archive SATA HDD. Max (avg) sequential read/write=190(150)MB/s, expected random read/write~150 IOPs.

Server Specs

- HPE DL60, E5-2650 V3, 20(40) cores(HTs), 128GB RAM, OS - 500GB 7.2K HDD. 3 additional 3.5 drive slots + 2 PCIe slots free. Balanced performance mode. SL6.7

- Note for the HPE SAS SSD tests a DL360 E52640 V4, 20(40) cores(HT), 128G RAM and a raid card (HPE P440ar) and with 2GB cache. Disks connected with SAS. SL6.7

- For many IO test the memory has been limited to 2GB to remove RAM cache effects. However, need to include RAM in any final evaluation as cache it is part of storage system.

Quick Studies: Raw Tests

- hd-parm -Tt [—direct] /dev/XXX
  - -direct, use kerne's O_DIRECT and bypass the page cache.
  - -T, read from the Linux cache for "cached disk read"
  - -t, read directly from the drive "buffered disk read"

- First number is cached read, second number is buffered read.

- The disk buffer is physically distinct from and is used differently than the page cache typically kept by the operating system in the computer's main memory. The disk buffer is controlled by the microcontroller in the hard disk drive, and the page cache is controlled by the computer to which that disk is attached. The disk buffer is usually quite small, and the page cache is generally all unused physical memory. While data in the page cache is reused multiple times, the data in the disk buffer is rarely reused.

Quick Studies: Raw Tests

- dd if=/dev/zero of=/dev/XXX bs=1M count=10240 conv=fdatasync

- dd if=/dev/XXX of=/dev/null bs=1M count=10240

- read/write 11GB of data to raw device, in 1MB chunks, fdatasync to make sure data is on disk.

Quick Studies: Raw Tests

MB/s	hdparm -Tt	hpdarm -Tt —direct	dd read	dd write
intel optane 16GB NVMe	10854/876	838/837	915	149
intel 600p 512GB NVMe	10,824/430	1,231/676	453	546
Crucial BX200 240GB SATA	10,853/500	537/388	516	55
HPE (PM1635) 400G SAS	11,977/510	862/686	464	410
Samsung 960 Pro 512GB NVMe	11,034/1,473	2,407/2,452	1,400	1,500
Samsung 850Pro 256GB SATA	10,973/495	462/495	540	522
HPE 7.2K 512GB SATA	10,860/137	439/137	142	141
Segate SMR 8TB SATA	10,896/191	459/195	15 to 200	133

Notes

- Crucial SSD seen to be slow at writes in external tests. Limitation of the 3GB SLC cache.

- Archive Disks read performance improves due to on disk caching 15MB/s - 31MB/s - 200MB/s over 3 runs, performance is reset by write test.

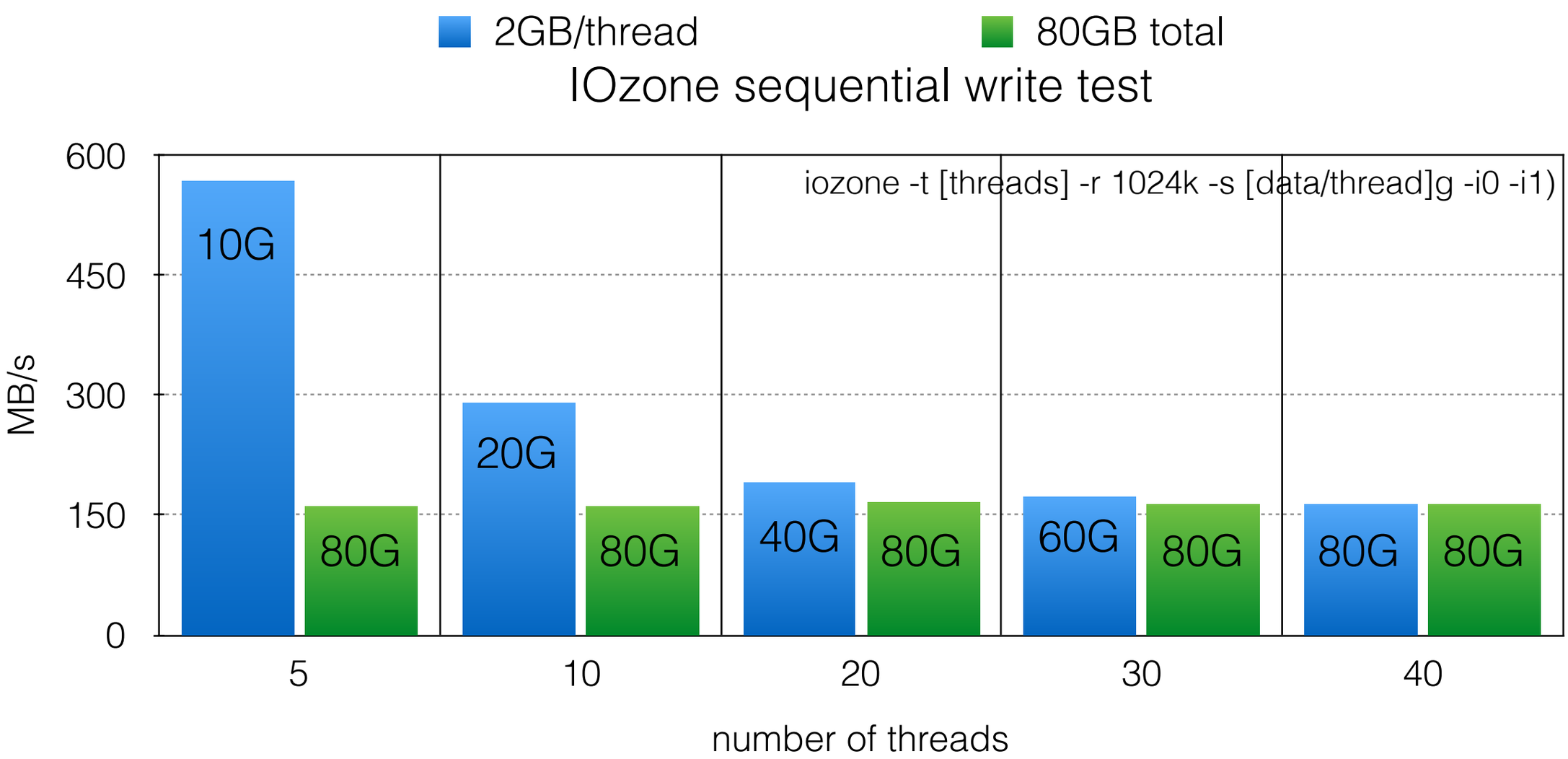
- No effect seen on BIOS “Power efficiency” mode on performance.

- NVMe dd read tests are cpu limited, Intel write tests has high cpu usage for NVMe process not seen for Samsung.

IOzone studies

- Ext4 format, noatime used when mounting, discard not used but fstrim run as cron job.
- iozone -e -+u -t 10 -r 64k -s 10g -i0 -i1 -i 2 -i 3 -i 5 -i 8
- iozone -e -+u -t 40 -r 1024k -s 2g -i0 -i1 -i 2 -i 3 -i 5 -i 8

IOzone test - NVMe

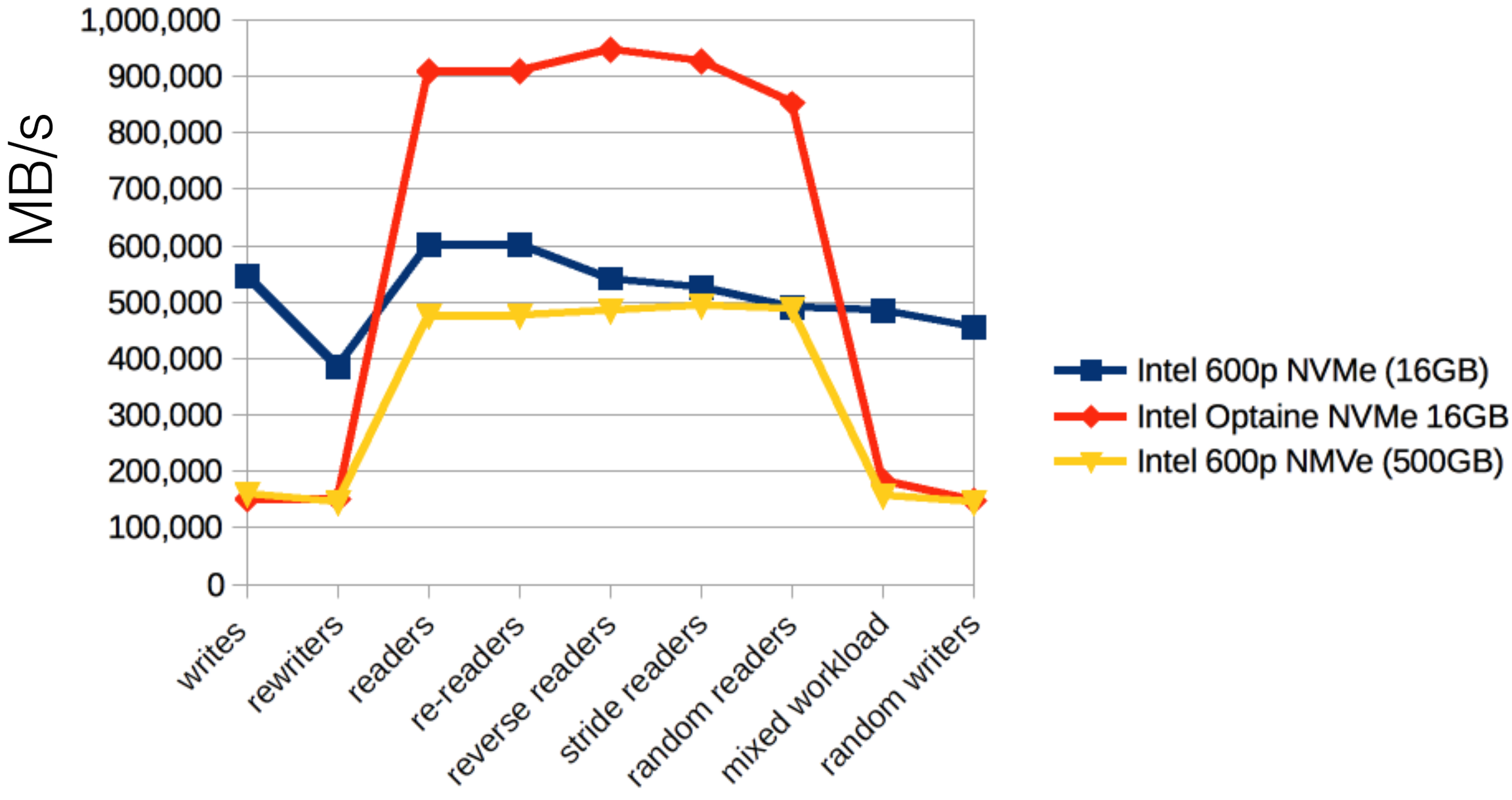


The effect of the 17.5GB of SLC write cache. Quoted sequential write speed of 560MB/s only seen for writes less than the size of the SLC cache!

No effect on throughput on number of threads.

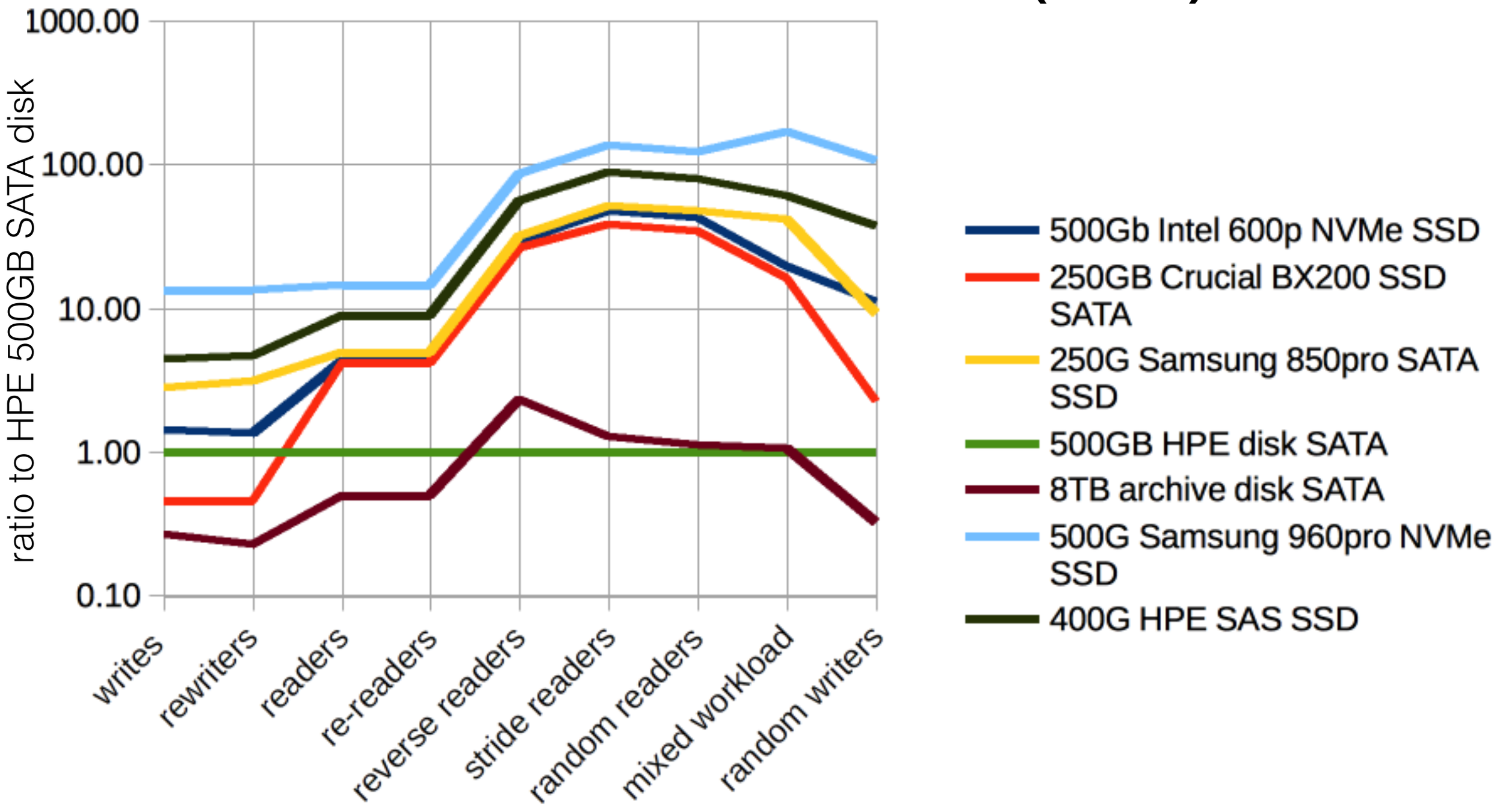
Optane vs SLC

read/write < 16GB



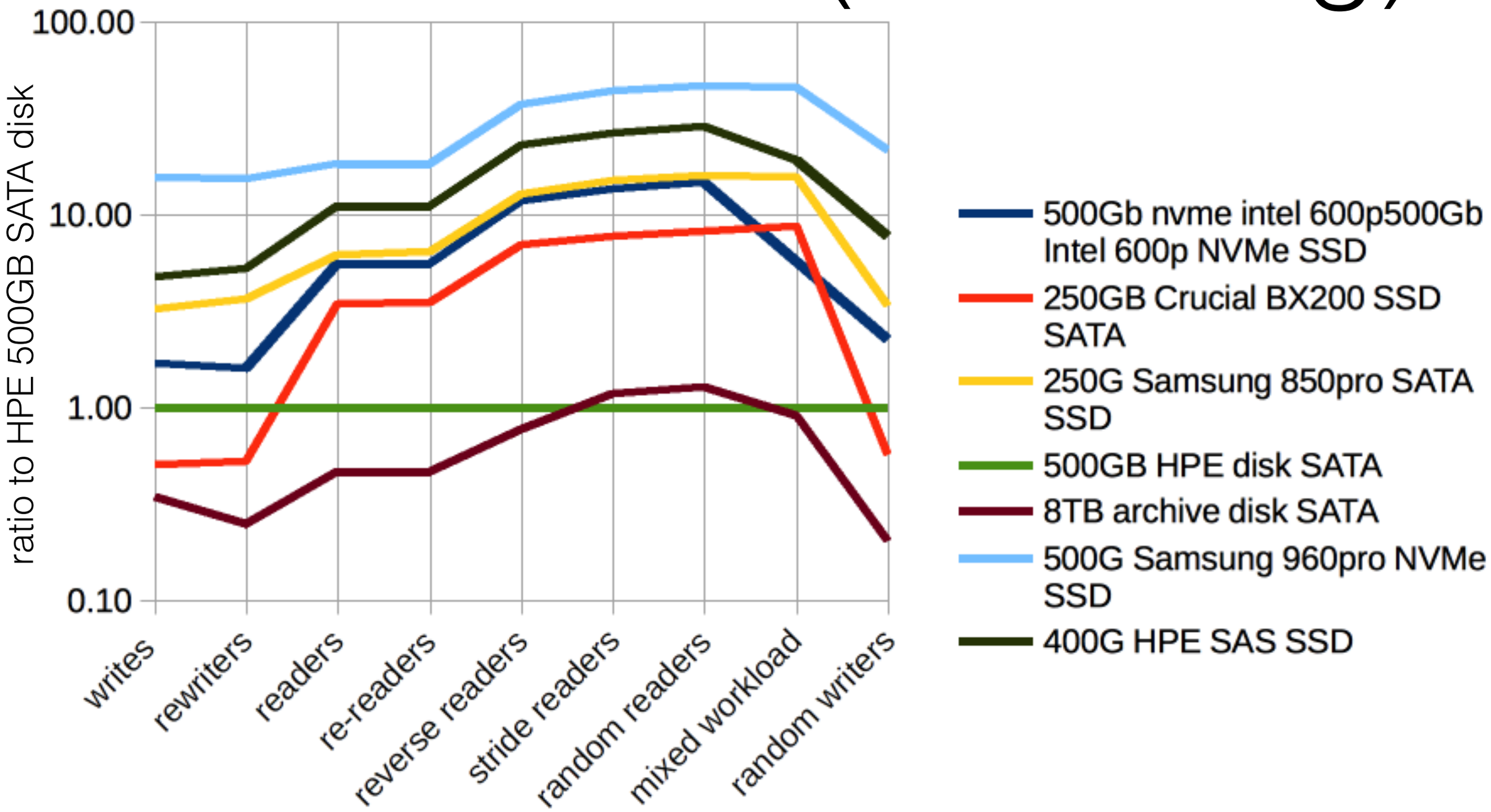
iozone -e -+u -t 6 -r 64k -s 2g -i0 -i1 -i 2 -i 3 -i 5 -i 8

IOzone test1(IO)



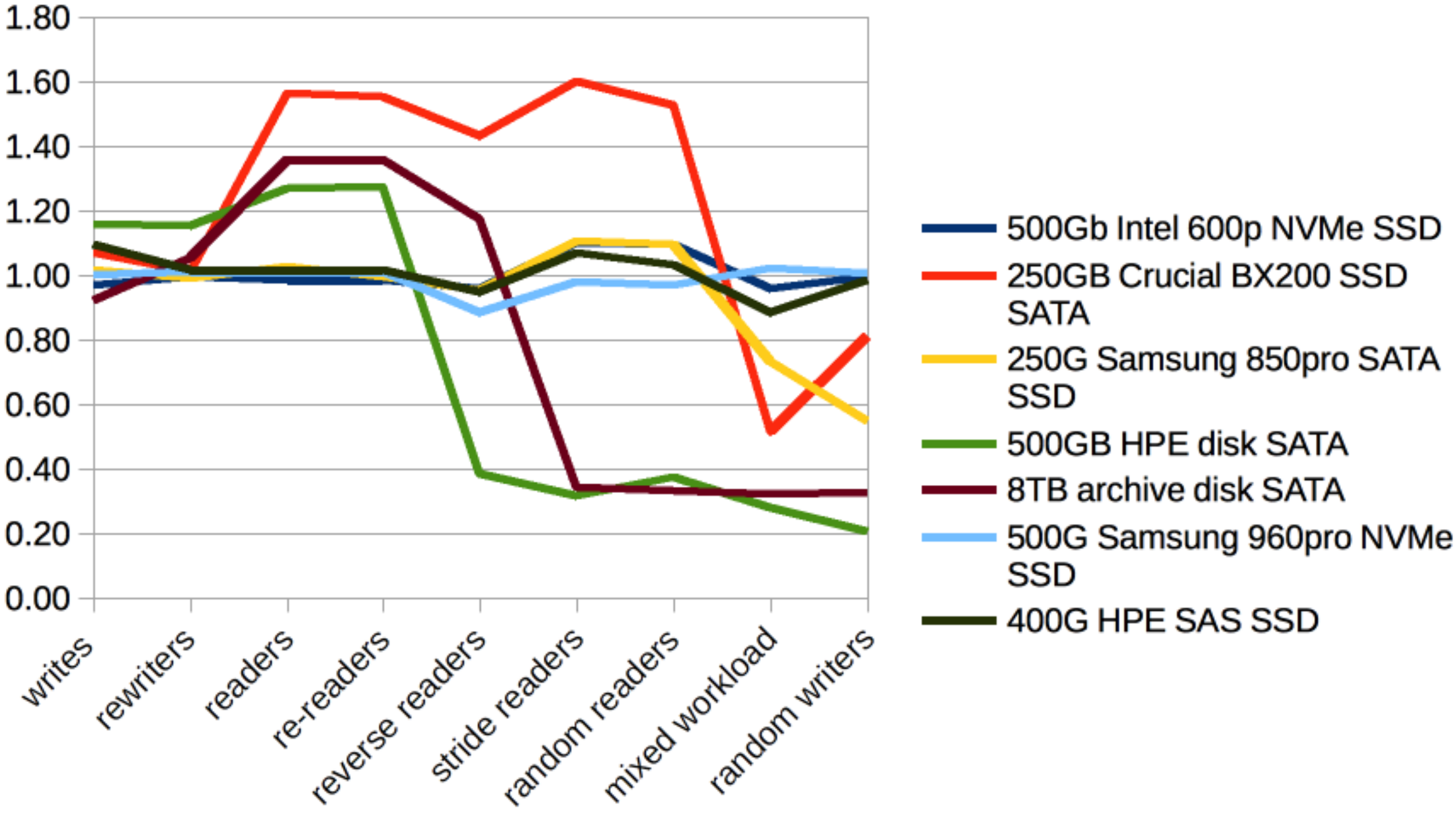
iozone -e -+u -t 10 -r 64k -s 10g -i0 -i1 -i 2 -i 3 -i 5 -i 8

IOzone test2 (streaming)



iozone -e -+u -t 40 -r 1024k -s 2g -i0 -i1 -i 2 -i 3 -i 5 -i 8

IOzone Ratio



test1/test2

Workload tests

- HEPSEPC - no impact of using SSD.
- make ROOT - no impact (-j40).
- make herwig - no impact (-j40).
- No benefit seen on workload tests when using an SSD.

Summary

- Important effects seen due to storage caches on performance results.

- Quoted performance numbers difficult to achieve.

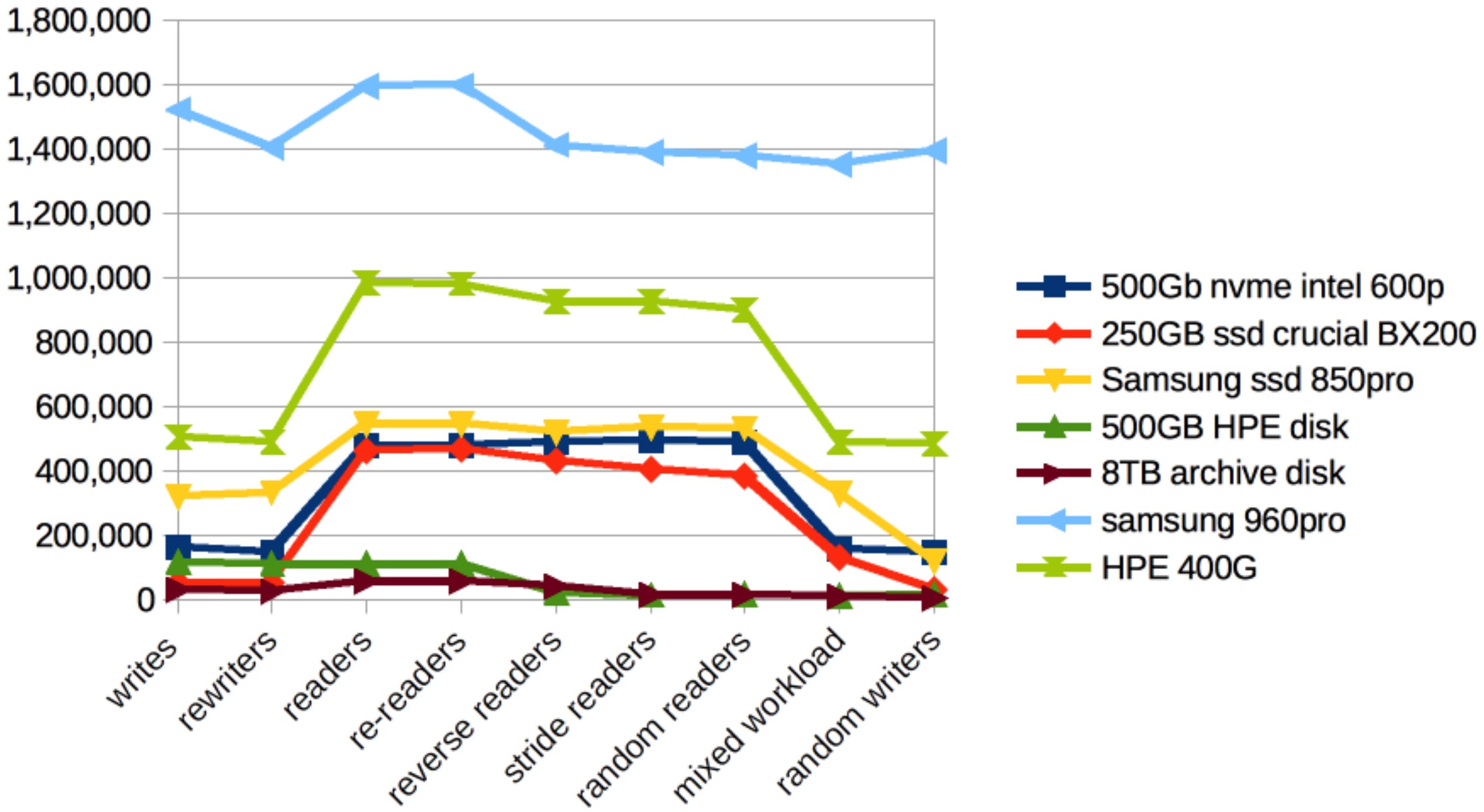
- NVMe > SAS > SATA > HDD (but not always see point 1 above).

- SLC>MLC>TLC. TLC for capacity, MCL for performance, SLC for caches.

- If you want performance you have to pay for it.

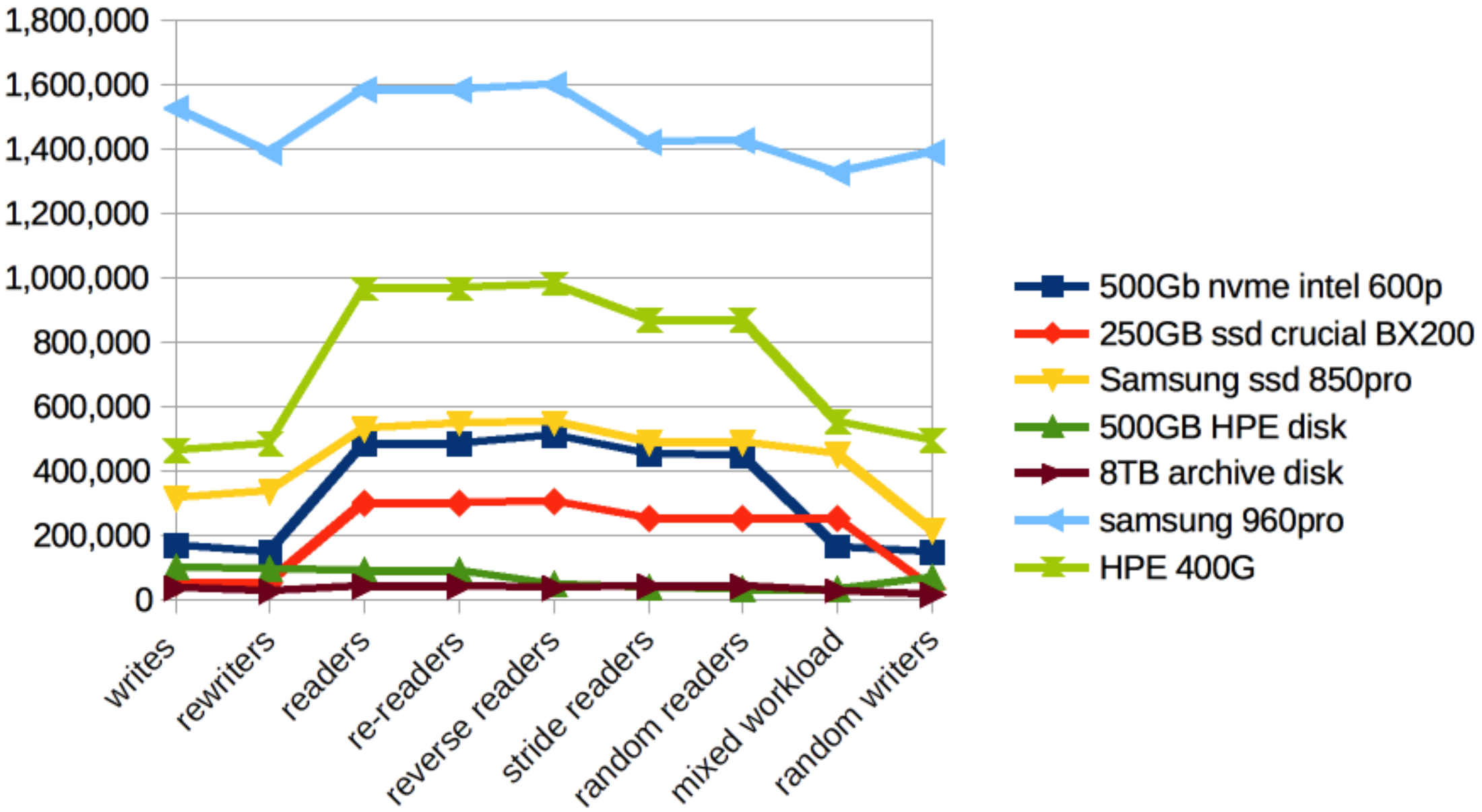
backup

IOzone test1(IO)



iozone -e -+u -t 10 -r 64k -s 10g -i0 -i1 -i 2 -i 3 -i 5 -i 8

IOzone test2 (streaming)



iozone -e -+u -t 40 -r 1024k -s 2g -i0 -i1 -i 2 -i 3 -i 5 -i 8