MONASH University

Information Technology

New Generation Mass Storage Technology for Big Data Applications

FIT Seminar, 11th June 2014

Dr Carlo Kopp, Senior Member IEEE, Associate Fellow AIAA, Fellow LSS, PEng
Carlo.Kopp@monash.edu
Clayton School of Information Technology,
Monash University
Australia

Author Background in Mass Storage / Bussing

- 1983-1984 Emitter Coupled Logic (ECL) and optical high speed design and measurement, including high data rate copper and optical cable characterisation, measurement, and modelling;
- 1985-1986 Contributed ECL bus arbitration circuit to Wallace password-capability multiprocessor backplane bus design;
- 1990-1993 SCSI bus systems integration, design of high speed SPARC/MBus motherboard bussing/clocking, SCSI drive enclosure design including thermal management; introduction of first generation Ciprico RAID controllers;
- 1990-1996 MSc under C.S. Wallace dealing with WalNUT password-capability virtual memory system / architecture;
- 1994-1999 Performance measurement and modelling consultancies for PHA Pty Ltd, including SCSI storage subsystems measurement and characterisation;
- 2000-2014 Teaching CSE2324/3324 / FIT2069 Comp Arch.

What is Mass Storage?

- **Wikipedia Definition:**
- "In computing, mass storage refers to the storage of large amounts of data in a persisting and machine-readable fashion."
- "Devices and/or systems that have been described as mass storage include tape libraries, RAID systems, hard disk drives, magnetic tape drives, optical disc drives, magneto-optical disc drives, drum memory (historic), floppy disk drives (historic), punched tape (historic) and holographic memory (experimental)."
- "Mass storage includes devices with removable and non-removable media."
- "It does not include random access memory (RAM), which is volatile in that it loses its contents after power loss."
- **Punchline:** *Persistent / Non-volatile storage of machine readable data.*

Evolution of Mass Storage

- 1950s – IBM RAMAC patent – first ever HDD (Hard Disk Drive);
- 1960s – Drum and early HDD mass storage technologies in mainframes, memory primarily based on non-volatile magnetic core storage;
- 1970s – Static RAM (Random Access Memory) displaces magnetic core storage; parallel Mass Storage Busses proliferate;
- 1980s – Density of SRAM and HDD improve, magnetic floppy disks and DRAM (Dynamic Random Access Memory) appear;
- 1990s – Exponential growth continues in RAM and magnetic disk technology;
- 2000s - Exponential growth continues in RAM and magnetic disk technology, Flash non-volatile RAM appears and proliferates in consumer applications;
- 2010s – Exponential growth continues in RAM and magnetic disk technology, Flash RAM commodified and appearing in Solid State Disks, designed to emulate form factor and interfaces in commodity HDD technology;

WD Velociraptor 10,000 RPM 3.5 inch Disk

Context: Mass Storage vs. Big Data Applications

- The growth in “Big Data” applications, whether analytics or management, is a direct consequence of “Kryder’s Law”, an empirical observation of exponential growth in magnetic data storage capacity per dollar expended;
- Current trends are Petabyte scale storage, and increasingly, Exabyte scale storage capacities;
- While “Kryder’s Law” will continue for some time yet, before it hits a plateau, “Big Data” is here to stay, whether we like it or not;
- The major long term challenge in “Big Data” applications of all types is achieving reasonable *time complexity* in working with massive datasets, which will only further grow in size over time;
- While some datasets can be readily partitioned, and worked on small chunks at a time, this is not generally the case;
- Time complexity models always make “cost” assumptions.

Cost Assumptions in Time Complexity Models

- When we model the time complexity of an algorithm, we always make assumptions about the cost of an operation, such as a *read*, *write*, *modify* or *insertion* operation;
- The simplest and most commonly used assumption is that the cost per operation is constant, and usually “unitary”, to simplify the model;
- The “constant cost” assumption is reasonable and valid, for many smaller datasets, especially if they can fit into the hardware cache, or hardware main memory of the machine;
- *With very large datasets, and complex multi-tiered storage architectures, or cloud storage, the assumption is invalid;*
- Advances in mass storage technology are now further complicating this problem, as Flash RAM technology adds yet another tier to storage architectures;
- *This presentation will explore advances in mass storage hardware.*

Storage Access Time vs. Transfer Rate

- The term “Access Time” or “Latency” [sec] describes the time elapsed, for a given storage device and architecture, between the start of a data transfer of a given size or type, and its completion;
- The term “Transfer Rate” or “Bandwidth” [bytes/sec] describes the rate at which data can be transferred to or from a device, or between devices, or between locations within a device;
- Access times and Transfer rates vary widely between storage technologies, and may depend on the location of the data within the device, the quantity of data, and how the data is arranged within the storage device – e.g. filesystems;
- In general, there is no rule which says that a storage technology with a high transfer rate has a short access time, or vice versa, or that high transfer rates and short access times are mutually exclusive, or vice versa;
- *System level integration, especially due to queuing effects, can further impact performance of mass storage once integrated into a computer system, and must be carefully considered when assessing application performance.*

Challenges: Mass Storage vs. Big Data Applications

- "Kryder's Law" will see big data sets get bigger and bigger over time, as the technology will exist to support ongoing accretion of extant big data sets, and aggregation of extant big data sets to form even bigger datasets;
- The industry push to migrate computing, and especially storage, to *Clouds* will see an increasing tendency to use distributed storage, with high latencies;
- Extant problems in distributed storage arising from network performance bottlenecks and maintenance of coherency / consistency will grow over time;
- Disruptive technologies such as *Solid State Disks* will influence computer architectures, and have already demonstrated immense potential for big data applications in HPC – case study is the NSF funded SDSC Gordon system;
- ***Future challenge*** *– migrate ideas from the Gordon effort into smaller computer systems to enable similar data intensive computing performance gains;*
- ***Future challenge*** *– algorithms must evolve further to exploit advanced storage and overcome known performance bottlenecks arising from hardware.*

EXPONENTIAL GROWTH IN MASS STORAGE

What is Exponential Growth?

- **Wikipedia Definition:** "Exponential growth occurs when the growth rate of the value of a mathematical function is proportional to the function's current value."

$$x_t = x_0 (1 + r)^t$$

- Where x_0 is the value at initial time $t=0$;
- Commonly used forms to represent exponential growth functions include:

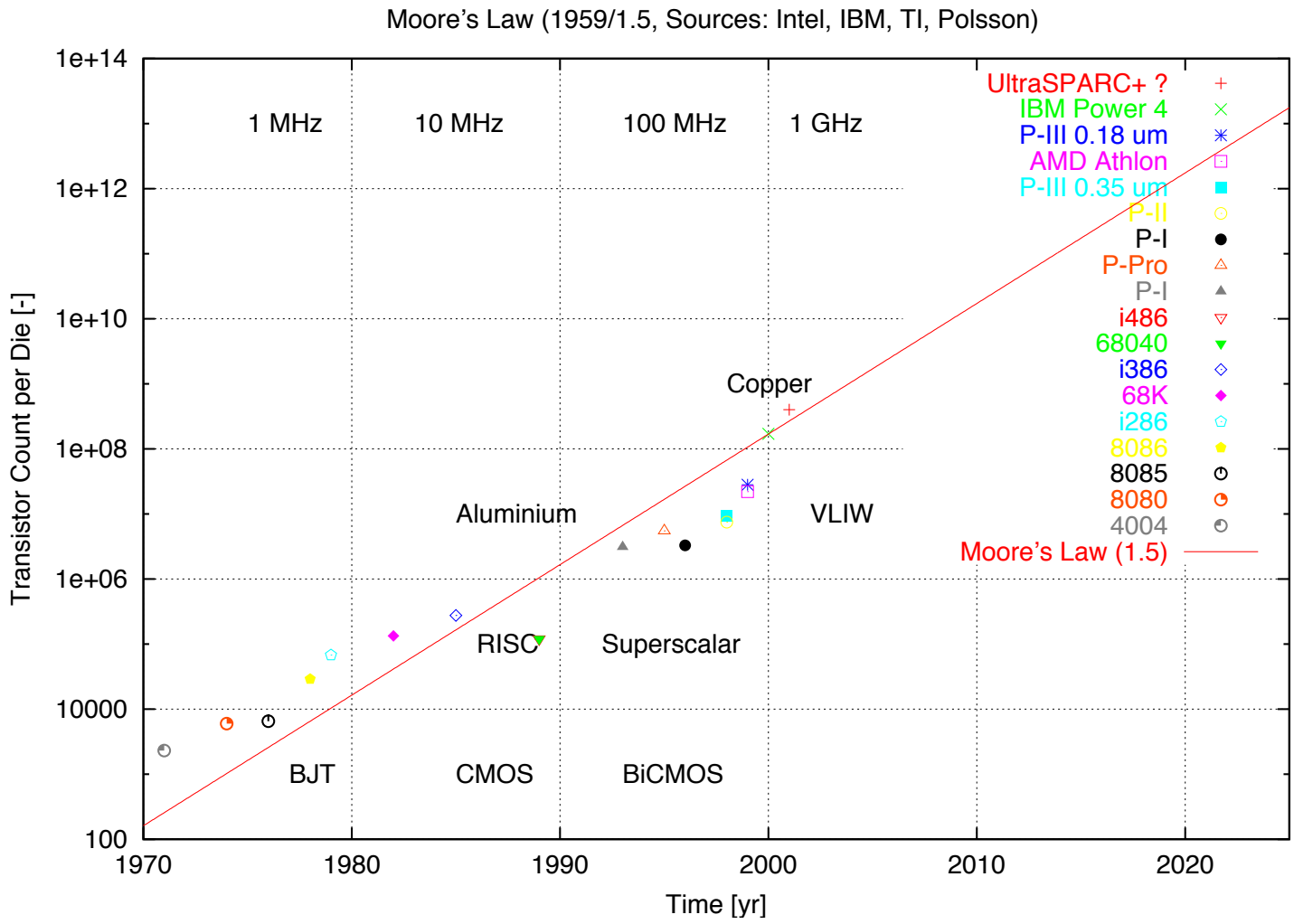
$$x(t) = x(0).e^{kt} = x(0).e^{t/\tau}$$

- Where k and τ are constants constraining the rate of growth over time;
- Often graphically represented as a straight line function on a logarithmic scale;
- **Punchline:** *Empirically observed effect in semiconductor monolithic technologies, and magnetic storage technologies – always bounded by the physics of the technology employed, and usually very "noisy" functions.*

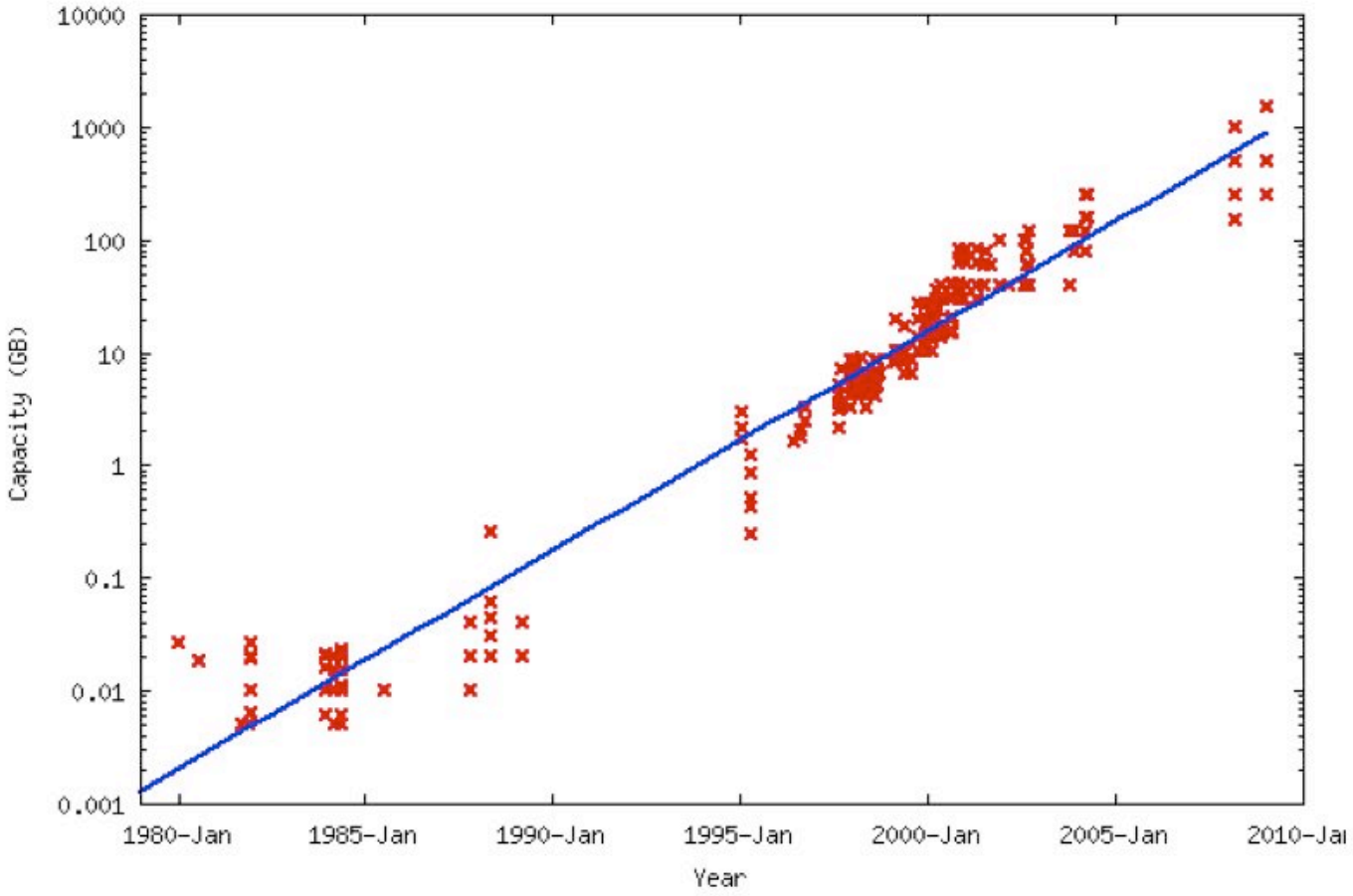
Exponentially Growing Mass Storage Technology

- Rotating “hard disks” continue to grow in surface storage density following “Kryder’s Law”, exponentially;
- While the density of rotating disks and thus transfer rate grow exponentially, improvements in access times / latency are limited due to mechanical rotation and quasi-linear head movements;
- New technologies will extend Kryder’s Law, although physics bounds are being approached;
- Solid State Disks (SSD) using “Flash RAM” technology, common to USB thumbdrives and SDHC modules, are now becoming available at affordable prices of < $1/Gigabyte;
- Within the physics bounds of Flash RAM technology, SSDs will follow the “Moore’s Law” exponential growth curve;
- *SSD access times are ~10^2 x shorter than Hard Disk access times ➔ dramatic changes to application performance.*

Moore's Law – Processor Chip Density

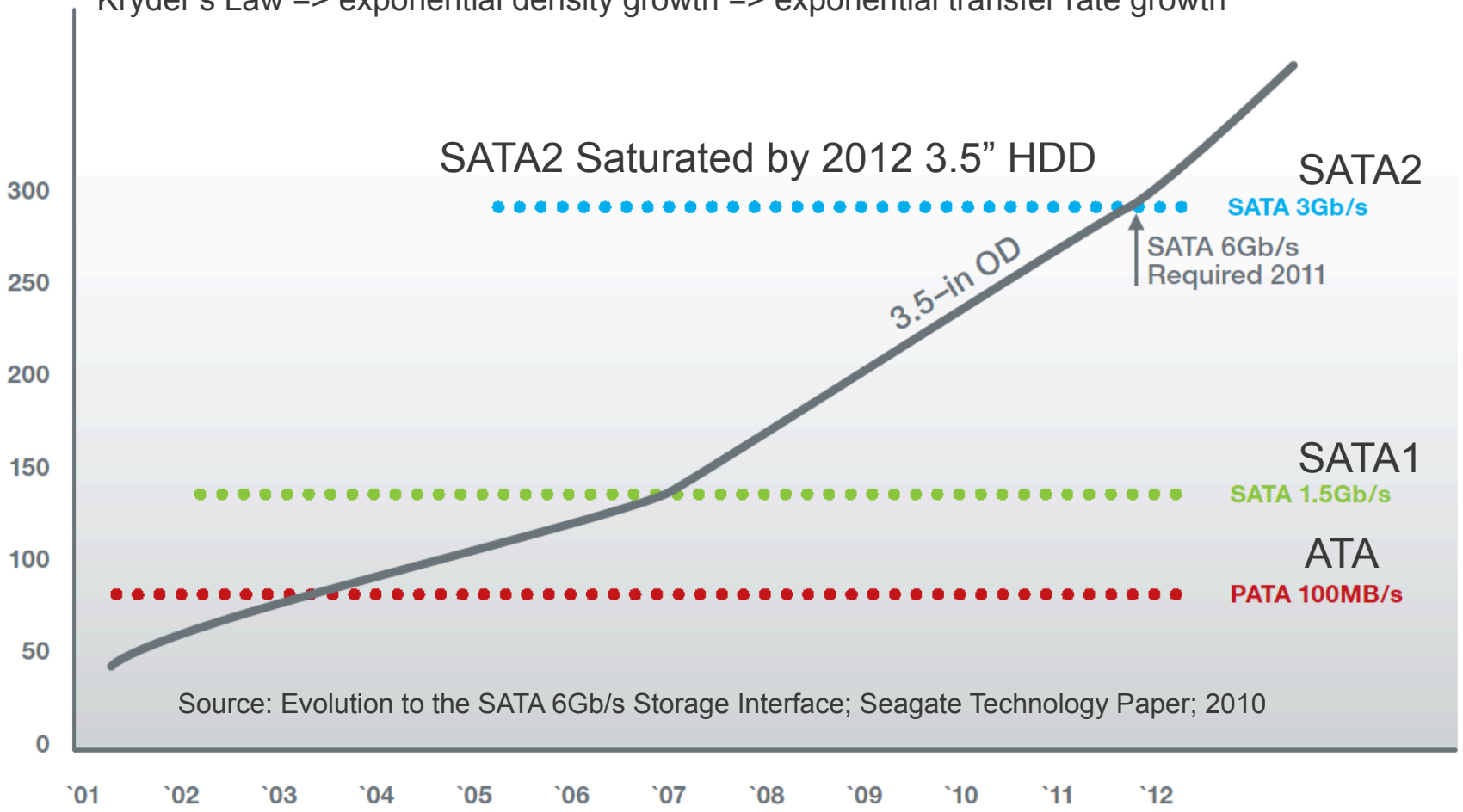


Kryder's Law (Storage Capacity vs. Time)



SATA Bus Variant Comparison / Kryder's Law

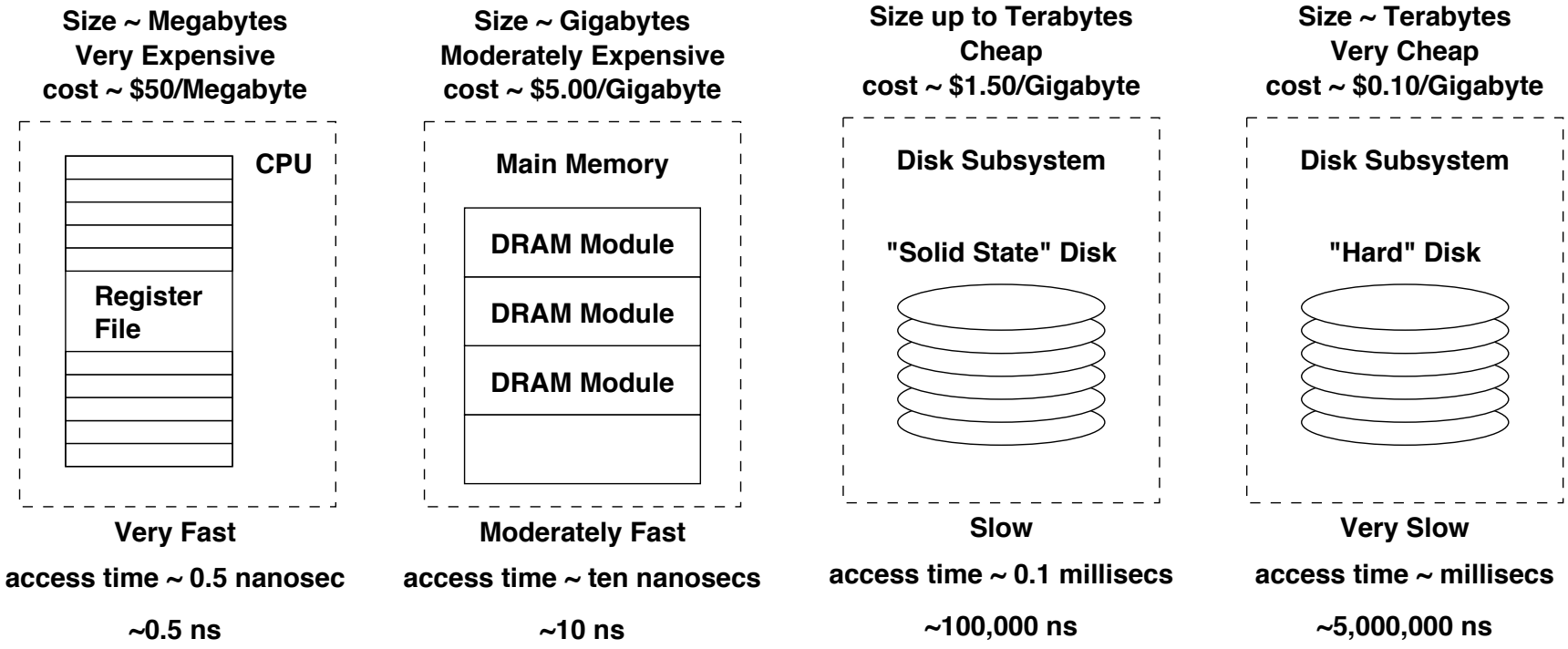
Kryder's Law => exponential density growth => exponential transfer rate growth



Source: Evolution to the SATA 6Gb/s Storage Interface; Seagate Technology Paper; 2010

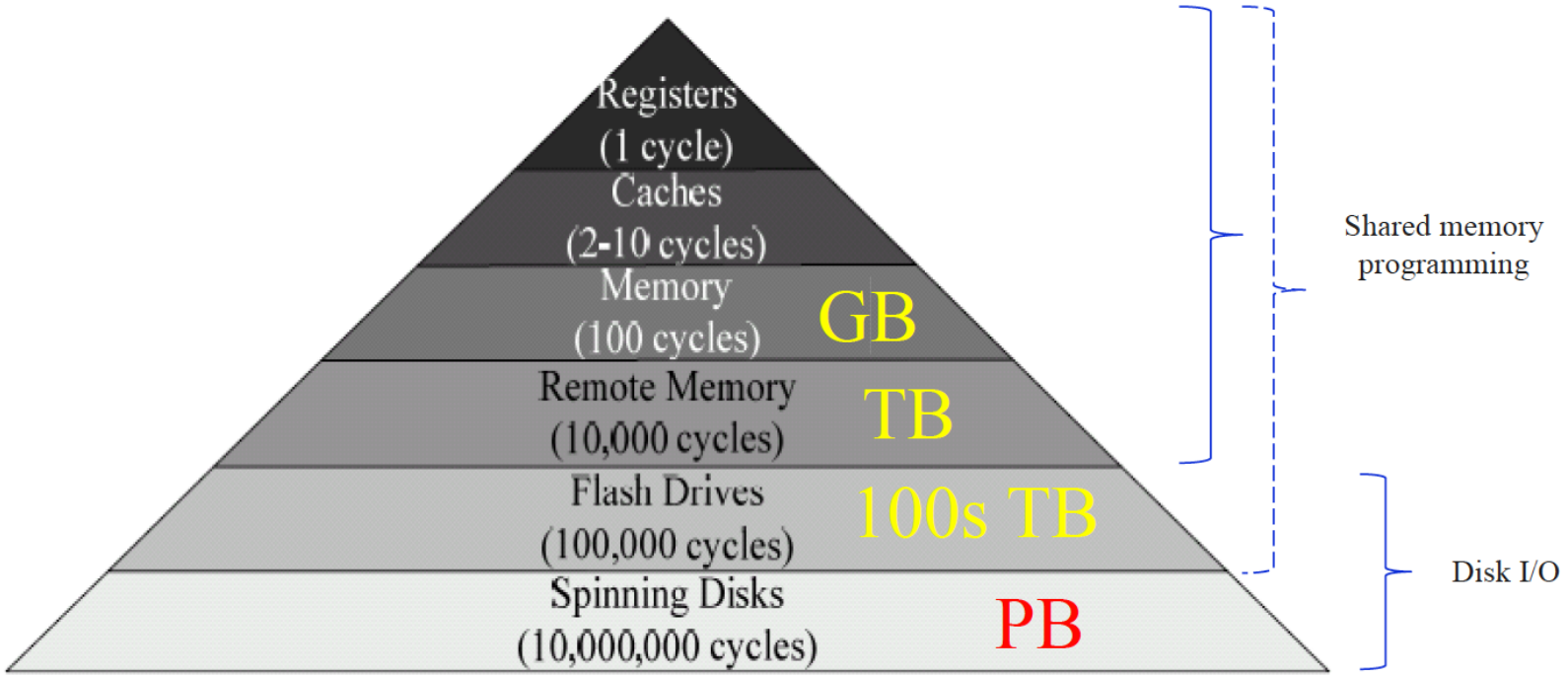
Figure 1. Sustained Data Transfer Rate Estimates

Storage Hierarchies - Today



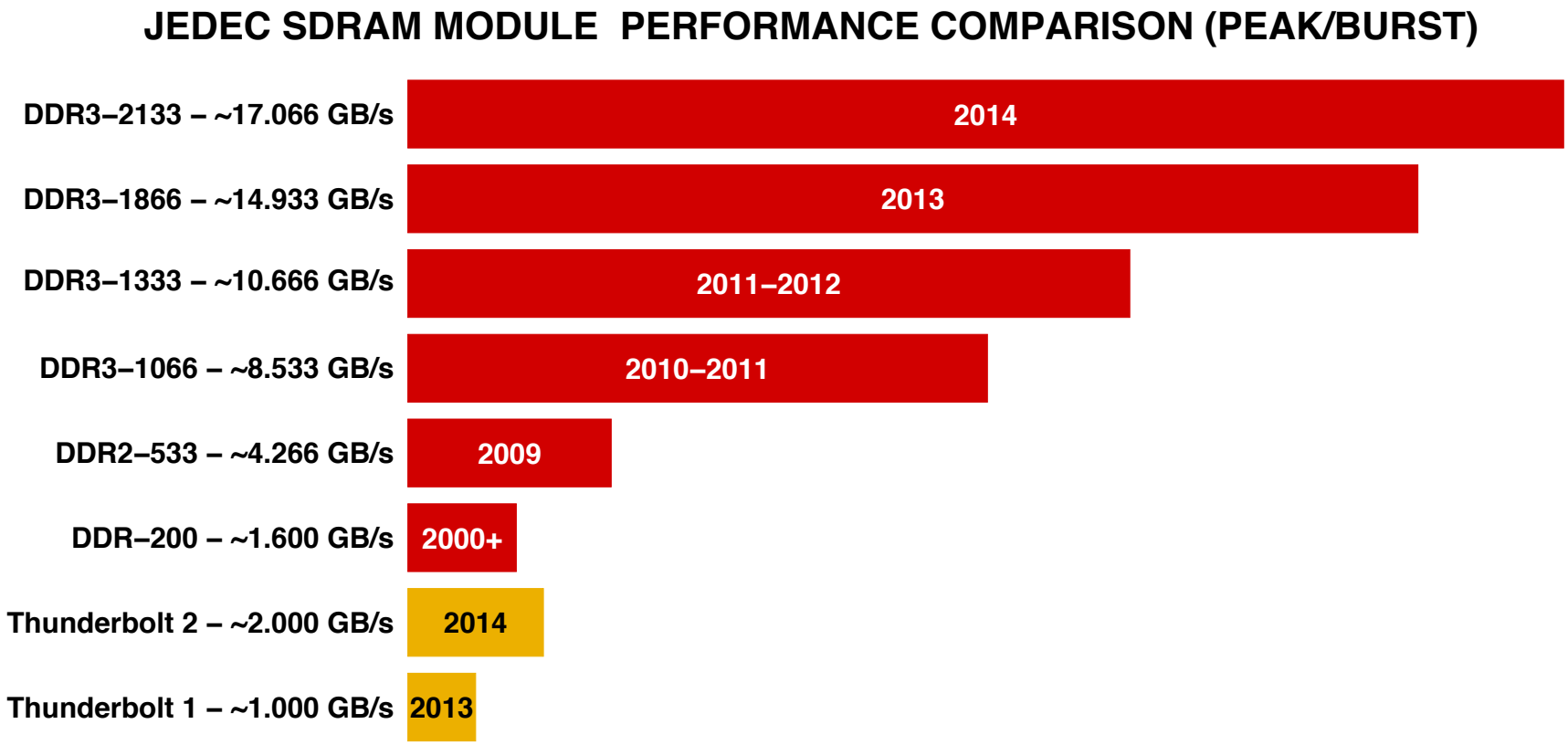
- Note the trade between access time, capacity and cost per capacity;
- This tradeoff evolves over time due to Moore's and Kryder's Laws;
- The nett effect is concurrent exponential growth in capacity and transfer rates, but not necessarily access times.

Latency in Storage Heirarchies (Norman 2013)



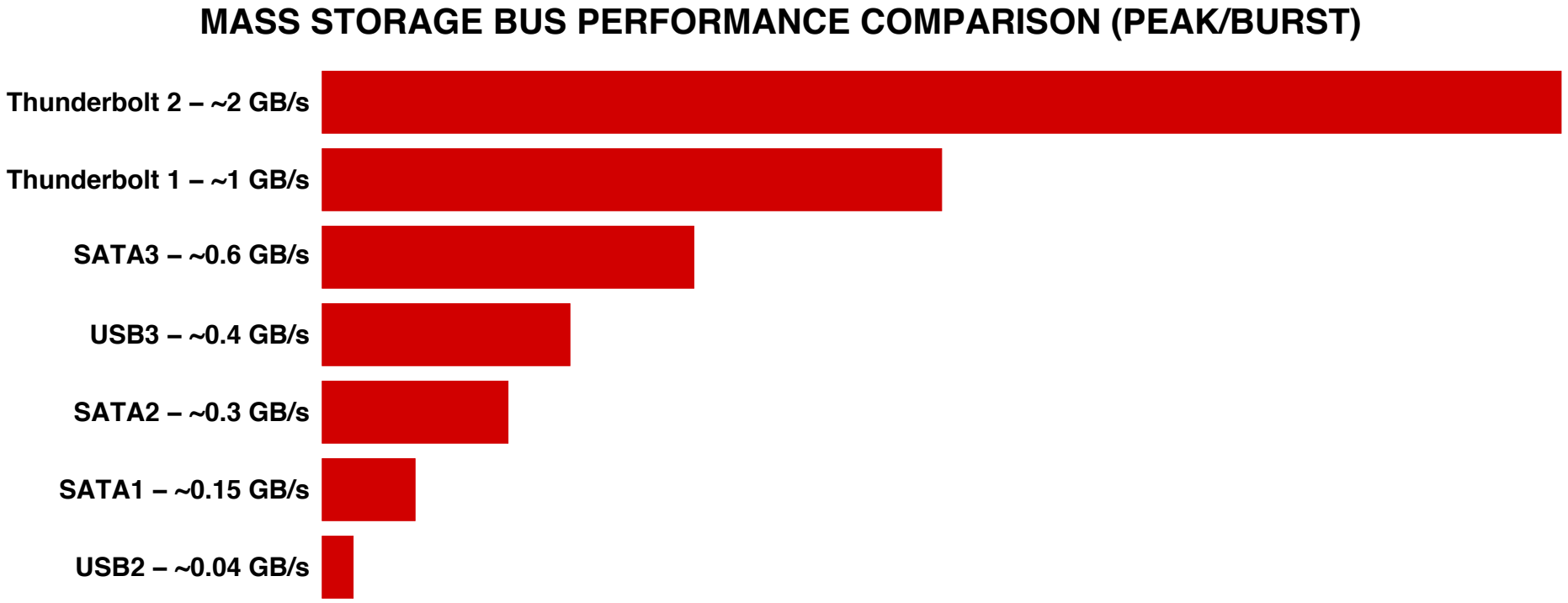
- Contemporary example: Latency vs. storage type and capacity in SDSC Gordon supercomputer (Michael L. Norman, *Application experiences with Gordon – a flash-based HPC system*, MURPA presentation, 2013);
- Desktop heirarchies are shallower, comprising Registers/Cache/Memory/Disk.

Storage Hierarchies – SDRAM Bandwidth



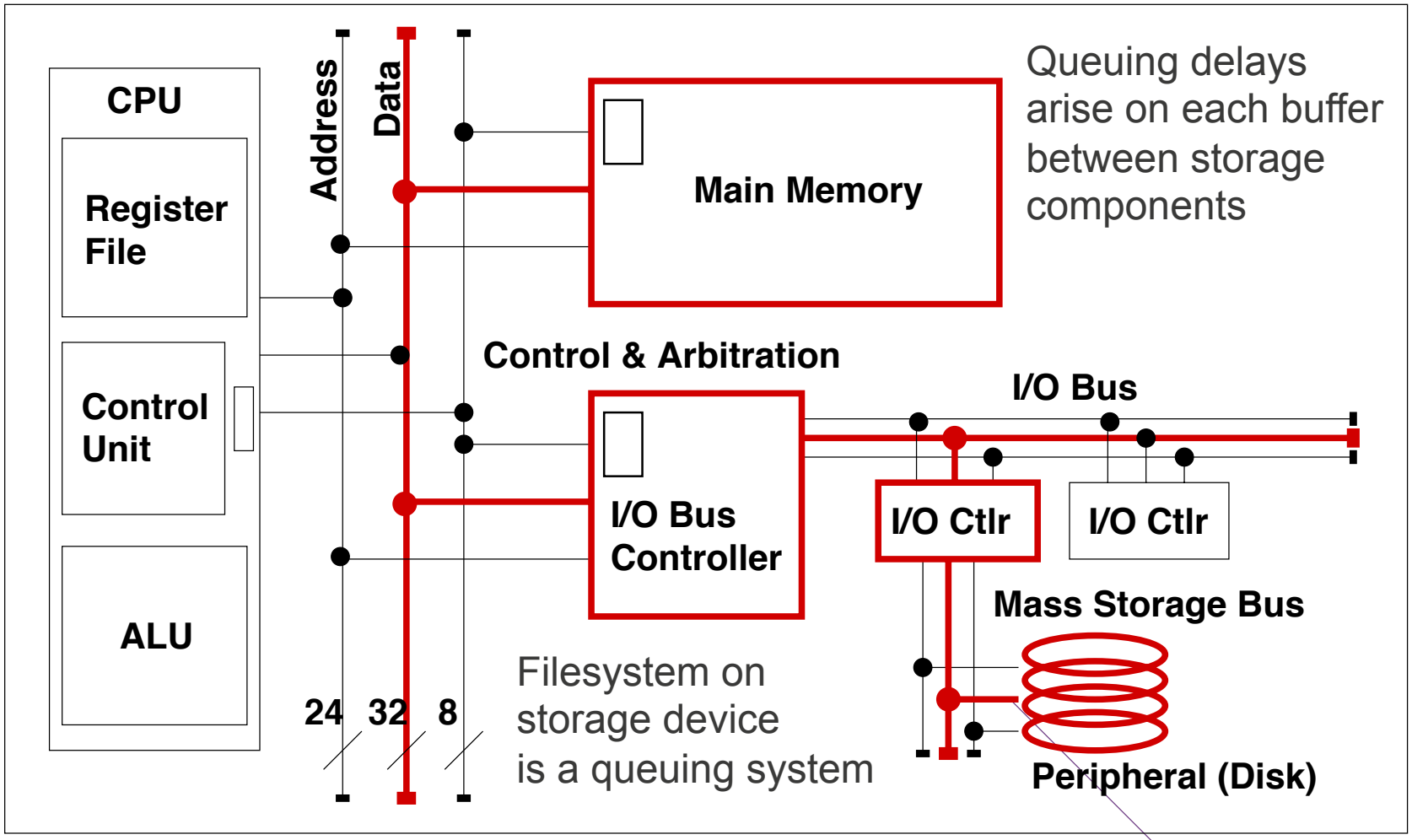
Note that JEDEC standard SDRAM module availability overlaps between DDR, DDR2 and DDR3, with all three formats available in 2013. GPU SRAM – 25 GB/s+

Storage Hierarchies – Mass Storage Bandwidth



Note that achievable throughput performance varies with implementation, and in practice may fall between 50% and 90% of nominal peak/burst transfer rates

Bus Heirarchies – A Serial Queuing System



MASS STORAGE BUS ADVANCES

Mass Storage System Transfer Rate Performance

- *How "fast" is a mass storage subsystem?*
- We need to consider disk access time, disk transfer rates, mass storage bus transfer rates, interface chips in machines, and the operating system being used;
- Operating systems also matter – filesystem design strongly impacts access times and transfer rates in rotating storage devices;
- Benchmarks performed with real hardware reflect the aggregate impacts of all components in the mass storage subsystem;
- *To get maximum performance, good choices must be made in what disk is used, what bus is used, and what operating system is run, and how it is configured, on a given item of hardware;*
- Example: an SSD with a SATA transfer rate of 600 Megabytes/s will only deliver ~80 Megabytes/s on a Firewire 800 bus – the bus is the "*bottleneck*" in overall transfer rate performance.

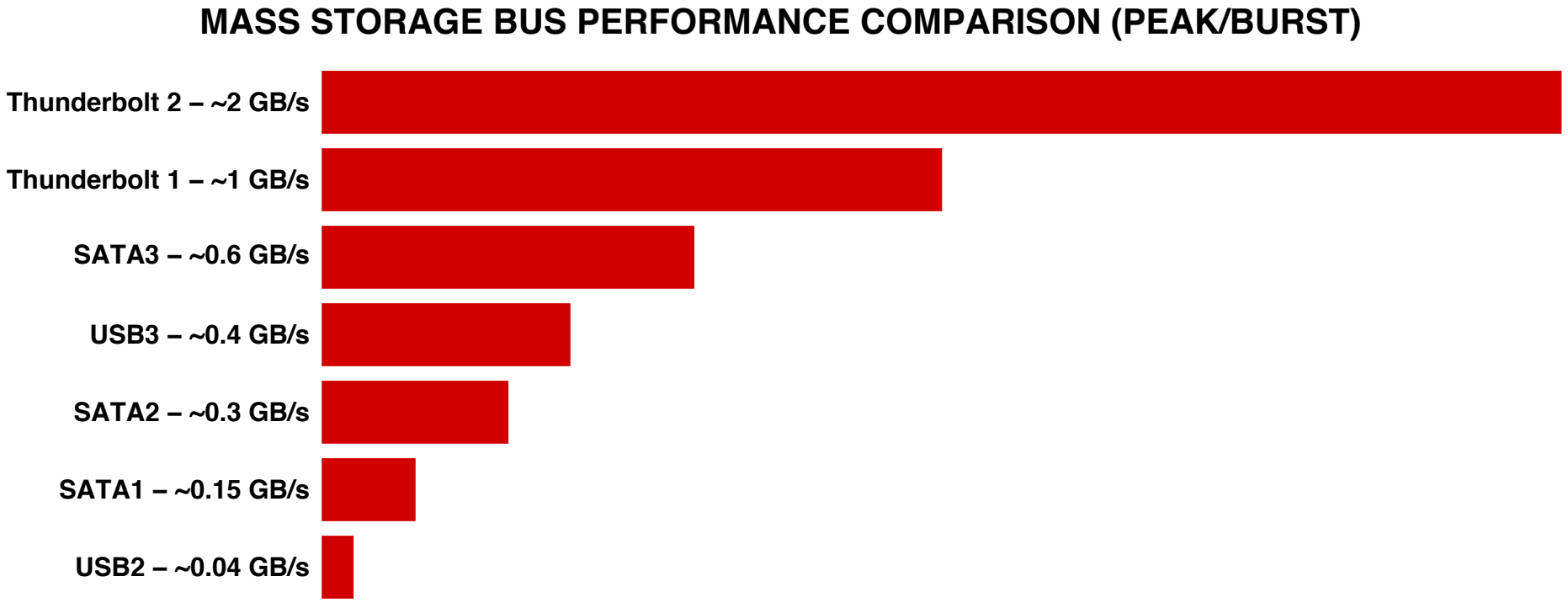
Mass Storage Busses

- Specialised busses for internal or external mass storage, e.g hard disks, magnetic tapes, optical drives;
- Designed for “bulk” block transfers rather than byte transfers;
- Usually based on widely used industry standards, with specific signals, connectors and cables – “plug and play” model;
- Until a decade ago were mostly parallel busses, more recently serial bus designs – e.g. IDE/EIDE/ATA evolved to SATA (Serial ATA) standards to overcome signal skew in cables/tracks;
- *Internal SATA 3.0 at 6 Gigabits/s ≈ 480 Megabytes/s; external variant is eSATA;*
- *External USB 3.0 at 5 Gigabits/s ≈ 450 Megabytes/s;*
- *External IEEE 1394b “Firewire 800” at 0. 786432 Gigabits/s ≈ 98.3 Megabytes/s; S3200 to 3.2 Gigabits/s ≈ 400 Megabytes/s;*
- *SAS or “Serial Attached SCSI” at 6–12 Gigabits/s. 480-960 MB/s*

Mass Storage Busses - Thunderbolt

- *Thunderbolt* is an Apple/Intel led initiative to provide a cabled derivative of the PCI-Express (PCIe) serial protocol for use over 3 metre copper cables, or 100 metre optical fibre cables, with combined support for VESA *DisplayPort* 8.64 Gigabit/s protocol in the same interface / connector;
- PCIe is now the dominant interconnect embedded in motherboards, to connect to various I/O adaptors, usually plugged in boards;
- PCIe available with 1, 2, 4, 8, 16 or 32 lanes, each lane comprising two twisted pairs for bidirectional 8B/10B encoded serial traffic at 2.525 Gigabits/s;
- Thunderbolt 1 multiplexes a 4 lane PCIe channel into a single serial stream at 10 Gigabits/s ≈ 1 Gigabytes/s throughput, Thunderbolt 2 at 20 Gigabits/s ≈ 2 Gigabytes/s throughput;
- Actual Thunderbolt 1 & 2 performance depends strongly on hardware design.

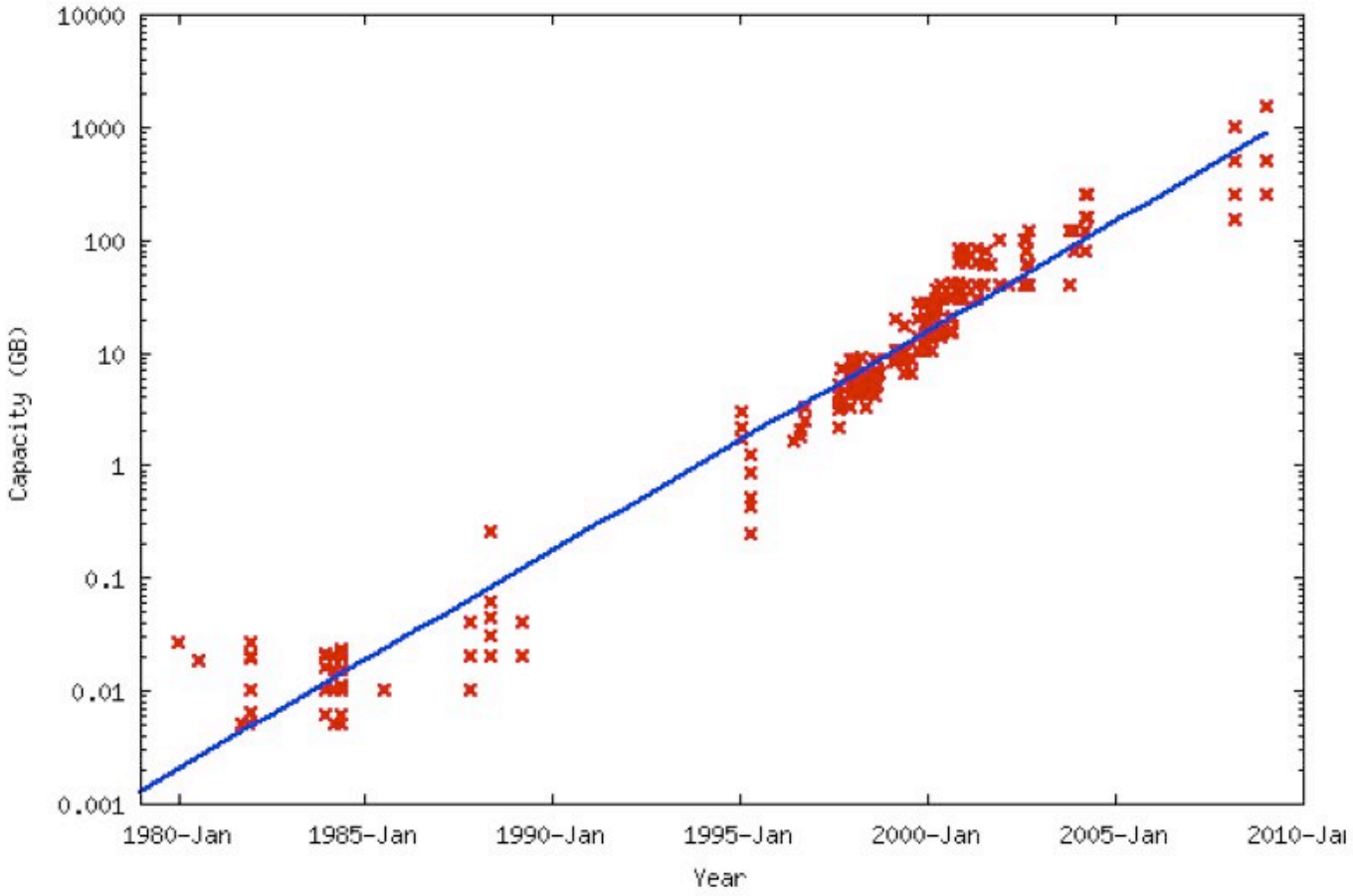
Storage Hierarchies – Mass Storage Bandwidth



Note that achievable throughput performance varies with implementation, and in practice may fall between 50% and 90% of nominal peak/burst transfer rates

MAGNETIC STORAGE ADVANCES

Kryder's Law (Storage Capacity vs. Time)



Rotating Disk Performance

- The stack of disk platters rotates at a constant RPM, in older disks 3,600 or 4,500 RPM, in newer types, 5,400, 7,200, 10,000 or 15,000 RPM.
- The time to access any item of data on the disk surfaces therefore depends on two parameters:

1. Rotational Latency - the time it takes for platter to rotate into position under the head.
2. Seek Time - the time it takes to move the head over the cylinder holding the data.

<Access Time> = <Seek Time> + <Rotational Latency> [ms]

<… Time> - denotes mean or arithmetic average time

Rotating Disk Access Time Performance

- The average rotational latency is 50% of the time it takes for a disk to rotate through 360 degrees:
  - 3600 RPM - 8.33 milliseconds
  - 7200 RPM - 4.17 milliseconds
  - 15,000 RPM – 2.0 milliseconds
- The average seek time between two adjacent cylinders (tracks) depends on the mechanical design of the head system:
  - Typically 1.5-5 milliseconds
- *<Access Time>* was usually 10 -> 13 [ms], now ~3.5-7 [ms]
- NB HDD access time improvement over a ~20 year period of ~2-3X!
- *Extremely slow compared to main memory or SSD!*
- Manufacturer specifications may be biased.

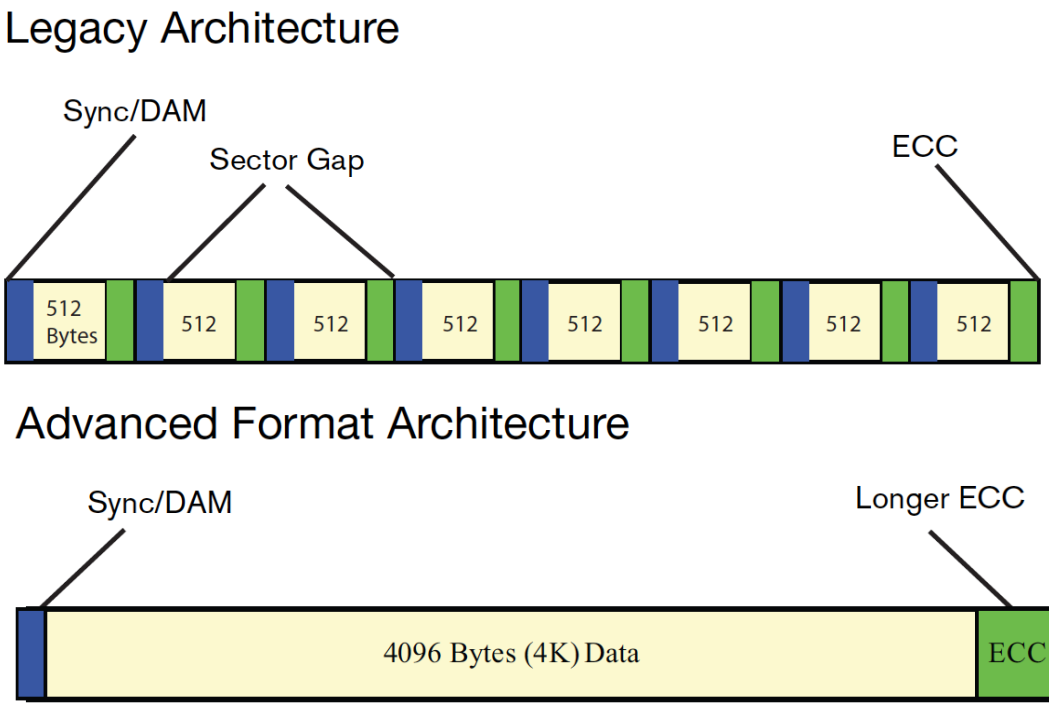
Improving Disk Access Performance

- Because mechanical disk hardware is very slow, we need to find ways of improving access performance:
  - Data on disks is stored in blocks of 512 byte - 16 kbyte, to minimise the overheads of accessing it e.g. 4 kbyte "AF" disks.
- Caching is a very useful technique for this purpose, and is used in two ways:
  - A hardware "disk block cache" is built into the disk drive assembly, typically using 8 - 64 Megabyte of SRAM or SDRAM.
  - "Hybrid" drives with 4 – 8 Gigabyte of DRAM cache within the drive enclosure (eg Seagate *Momentus XT* series – 2011).
  - The operating system uses portions of the main memory, ie "buffer cache" to cache disk blocks;
  - Multiple GB of on-disk and main memory cache are common now.

Rotating Disk Transfer Rate Performance

- Disk transfer rate performance is often called “bandwidth” in industry, and is a measure of the number of Megabytes per second transferred between the drive head and disk surface during a read or write operation;
- This parameter is the ultimate performance limit for reading or writing data to a rotating disk;
- It is *mostly* determined by the linear density of the data written on the disk surface, and the RPM of the disk;
- Increasing capacity (TB) ➔ increasing density ➔ increasing transfer rate or “bandwidth”;
- *Transfer rate is often confused with access time by laymen – NB common error in popular literature;*
- Other impacts are cache and storage bus “bandwidth”.

Disk Access Performance - Blocks



Advanced Format Technology White Paper, Western Digital, 2006

ECC – Error Correction Code Bit Pattern – cf network packets with headers and ECC

Advancing Magnetic Storage Technology

- *Giant Magneto-Resistive Effect (GMR)* heads – IBM a decade ago, now all manufacturers using licenced GMR technology;
- *Helium filled drive enclosures* – Western Digital; Helium permits cooler operation, allowing for more drive platters ~50-100% increased capacity;
- *Heat Assisted Magnetic Recording (HAMR)* – Seagate; HAMR uses a laser to preheat the point at which the magnetic head records, to change material properties and defeat "super-paramagnetic limit" – 100 x density gain;
- *Self-Ordered Magnetic Arrays (SOMA)* – Seagate; self-assembled, ordered and uniform nano-magnet arrays using FePt and CoPt compounds as recording medium;
- *It is expected these technologies will extend Kryder's Law for at least another decade;*

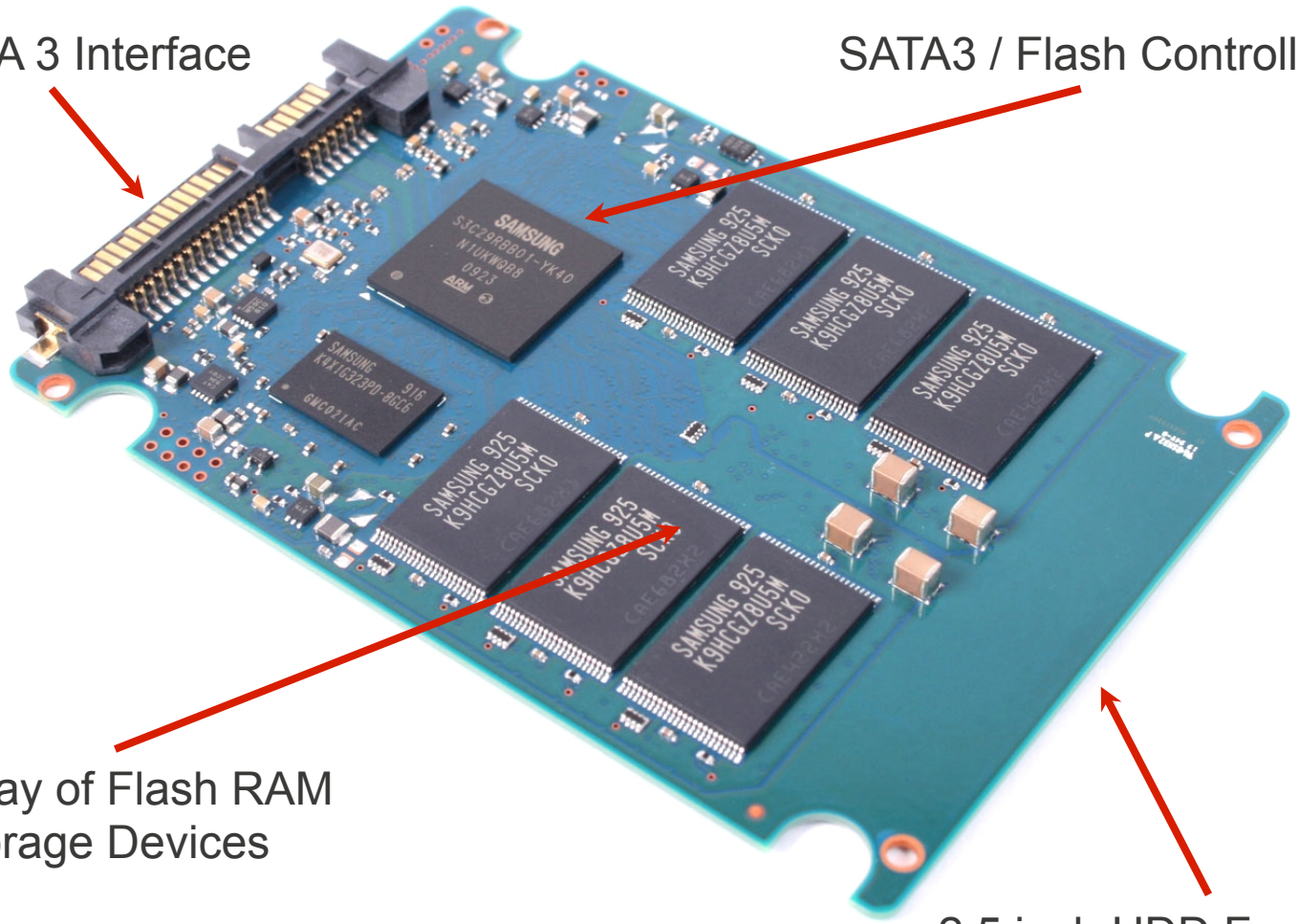
Seagate Cheetah 15,000 RPM 3.5 inch Disk

SOLID STATE DISK STORAGE ADVANCES

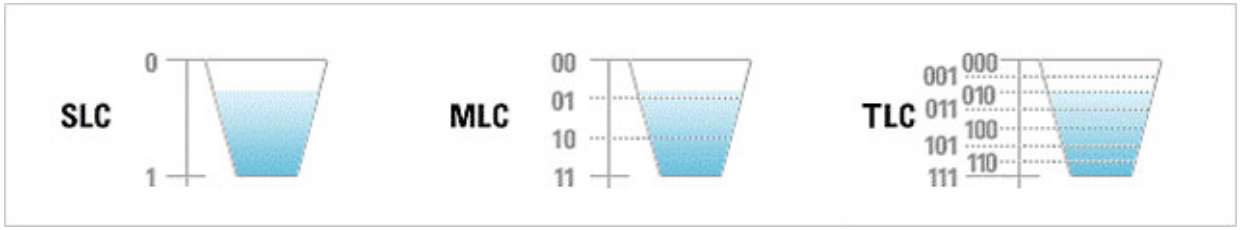
Solid State Disks

- SSDs employ arrays of Flash RAM chips for nonvolatile mass storage of data – latency typically *1/100* of a magnetic HDD;
- Flash RAM chips are used in commodity USB and SDHC media;
- An SSD device is accessed via a specialised embedded SSD controller chip, or chipset, with an industry standard SATA3 storage bus interface, and are packaged into the same “form factor” as industry standard “2.5 inch hard disks”;
- In most respects, an SSD qualifies as a lower power and quicker “drop in” replacement for a traditional rotating hard disk in the same physical and electrical format;
- Flash RAM employs a *write-read-erase-write-read* operating cycle, unlike magnetic media which use a *write-read-write-read* operating cycle – this has important ramifications.

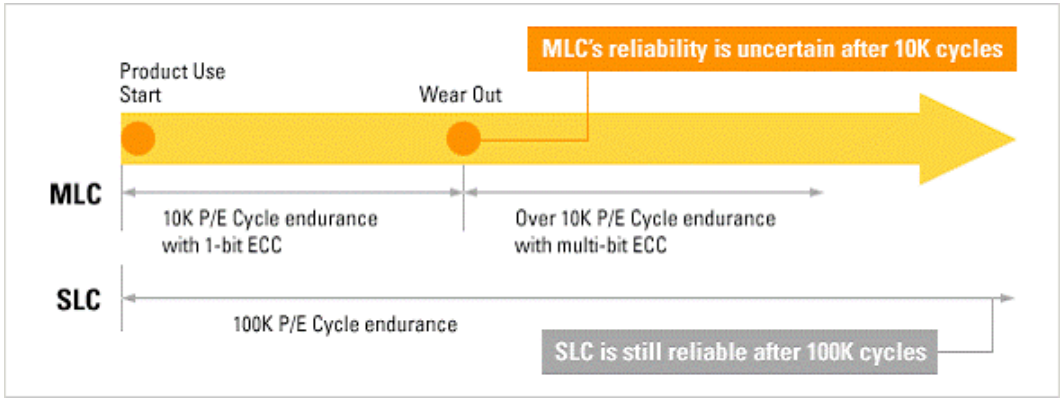
Samsung 830 Series SSD – 2.5 inch SATA3

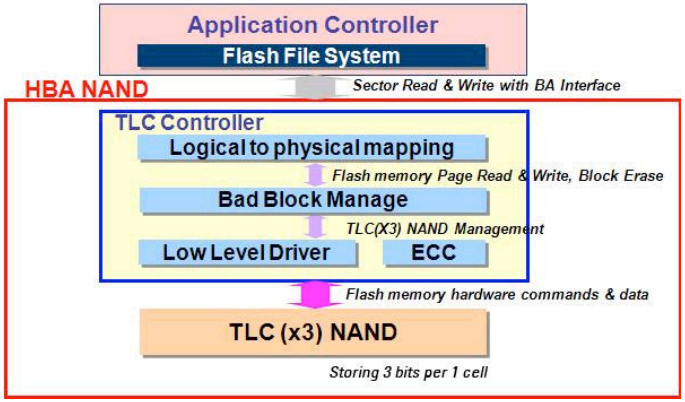


SSD Flash RAM Technology



SLC- Single Layer Cell (Enterprise/HPC)
MLC- Multi Layer Cell (Consumer/Pro)
TLC- Three Layer Cell (Consumer Only)





Flash RAM Write/Read/Erase Cycles

- SSD Flash RAM is typically constructed using NAND floating gate transistor technology, in which a logical “1” or “0” must be written into a cell using a unique “write pulse”, upon which the cell can be read for an almost unlimited number of times (the “read disturb” effect imposes some limits);
- The cell must be erased using a unique and much slower “erase signal” before it can be rewritten with new data;
- If a page of data in an SSD must be rewritten, the hardware must locate a previously unwritten, or erased, free page, and write the data to this new page, enqueuing the original page to be erased later;
- *The SSD controller chip hides this functionality, and presents to an operating system the illusion of a magnetic technology mass storage disk, with logically addressable read/write pages.*

Flash RAM Cycle Impacts on SSD Designs

- While an SSD may appear as a conventional read/write disk, internally the controller must maintain a continuously updated mapping table, in which logical page addresses point to the internal (hidden) physical address of the Flash RAM page being used at that time;
- A Flash device can be read and written by page, but only erased by block, each block containing multiple pages;
- Each block has a finite "write endurance", before it wears out and can no longer be erased and rewritten – typically between 1-5 x 10^3 and 10^6 write/erase cycles depending on which specific Flash technology is employed (SLC/MLC/TLC);
- Example: *100 GB SSD using TLC @ 10^3 writes, has a potential useful life, if wholly overwritten every day, of 2.74 years, assuming all writes are evenly distributed across all blocks.*

SSD Internal Management Functions

- The SSD controller chip performs a number of internal tasks to manage the use of physical blocks:

1. Garbage collection during “idle time” between writes, of blocks to be erased, since the contents of these blocks have been overwritten and put into a new block;
2. Wear levelling, to ensure that writes are as evenly as possible distributed across all blocks in the drive;
3. Management of “bad blocks”, detected by ECC (Error Correction Codes) which have failed due to defects or wearout;
4. Management of “over-provisioning” i.e. spare blocks, used to extent life of SSD –e.g. 2X overprovisioning in IBM 910;

- *NB controller algorithms are usually proprietary, and none are “ideal” e.g. in wear levelling and garbage collection.*

SSD Design Limitations and Problems

- Limitations in controller algorithms or hardware performance can impair performance or lifetime of the SSD device;

1. Poor management of writes can produce “write amplification” where repeated write/erase cycles might occur, while written data is shuffled and consolidated for block/page writes – this reduces the life of the SSD if arising frequently;
2. Poor wear levelling will see some blocks wear out faster due to overuse, reducing usable SSD capacity through its life;
3. Inadequate controller speed can impair write performance, as the controller cannot keep up with write requests, and manage blocks at the same time;

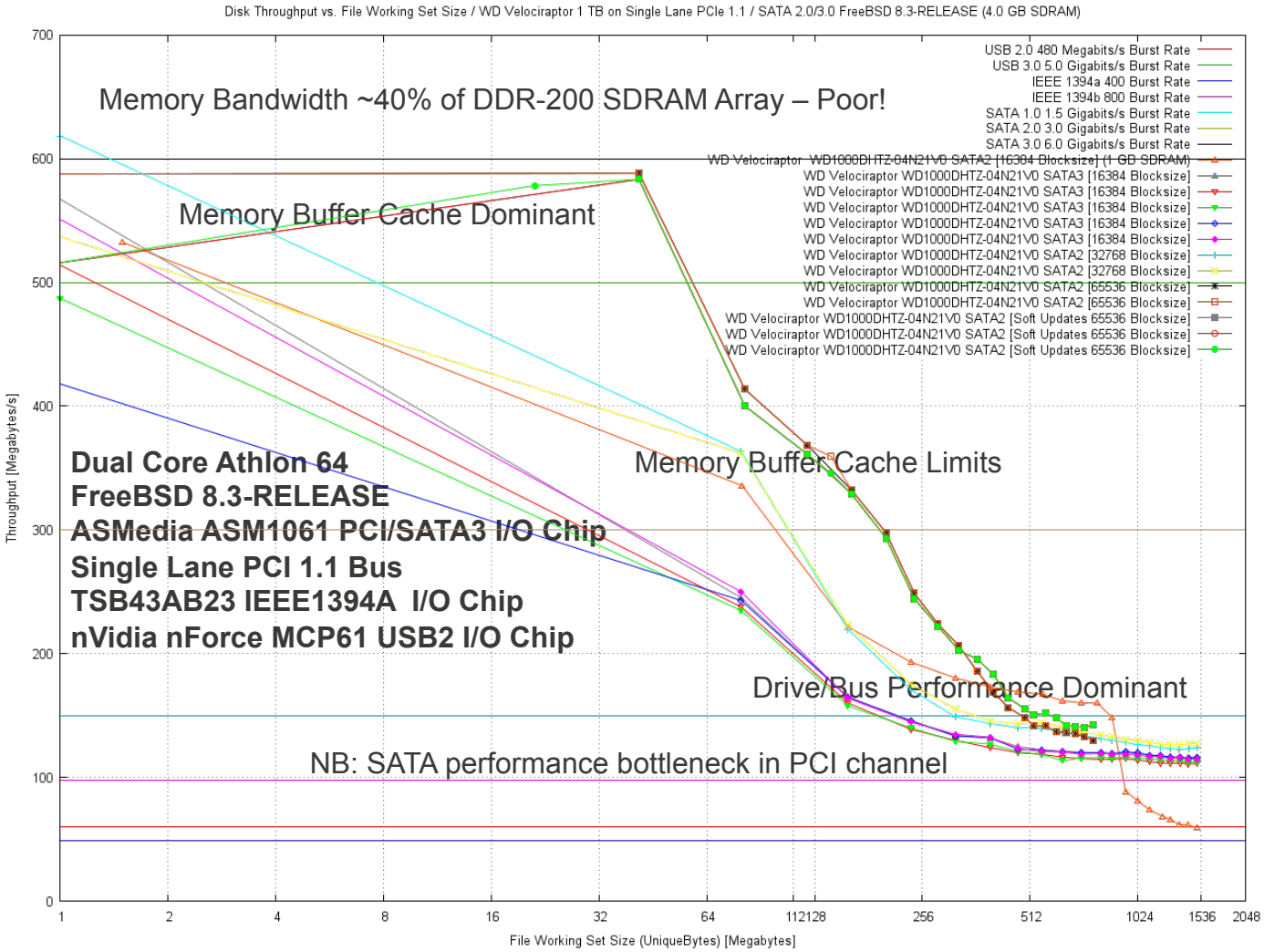
- *These problems are exacerbated when SSDs have few free blocks available!*

SSD SATA TRIM Command

- Modern SSDs with SATA interfaces mostly support the SATA “TRIM” command, which enables the filesystem to tell the SSD that a disk block (page) has been freed on a file delete operation, and can thus be erased and reused at any time by the SSD internal controller;
- Without TRIM, the SSD cannot know the filesystem has freed a block in a deleted file, until the filesystem attempts to write the freed block again – this causes a write operation fail and slows down write performance as the controller must immediately find a free block to write to;
- Newer operating systems support the TRIM SATA command, but older operating systems do not;
- TRIM capability is not critical in applications with low write frequencies – e.g. operating system or “boot” disks.

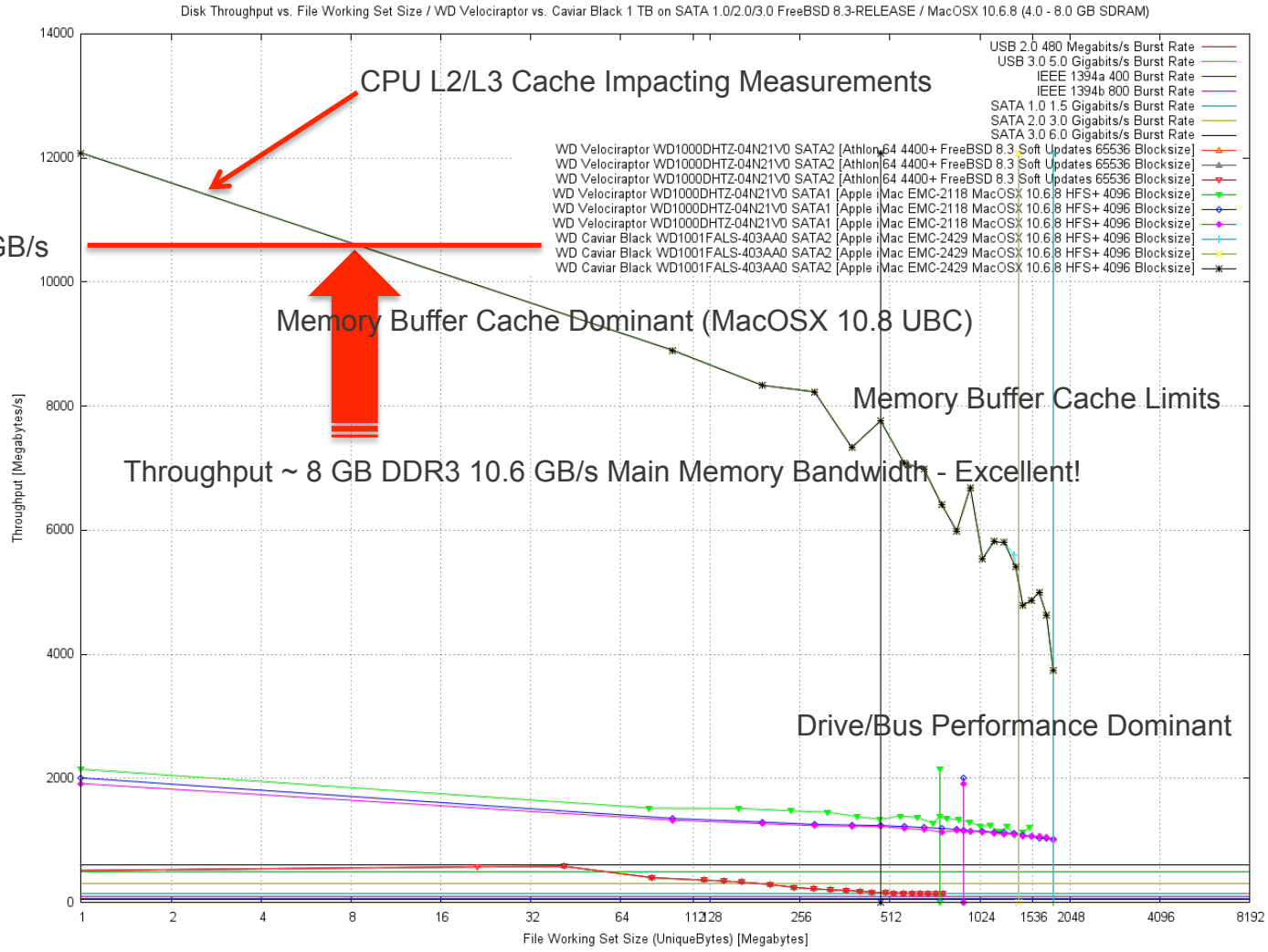
BENCHMARKING SSD PERFORMANCE

HDD Self Scaling I/O Benchmark (Chen 1992/Kopp 2013)

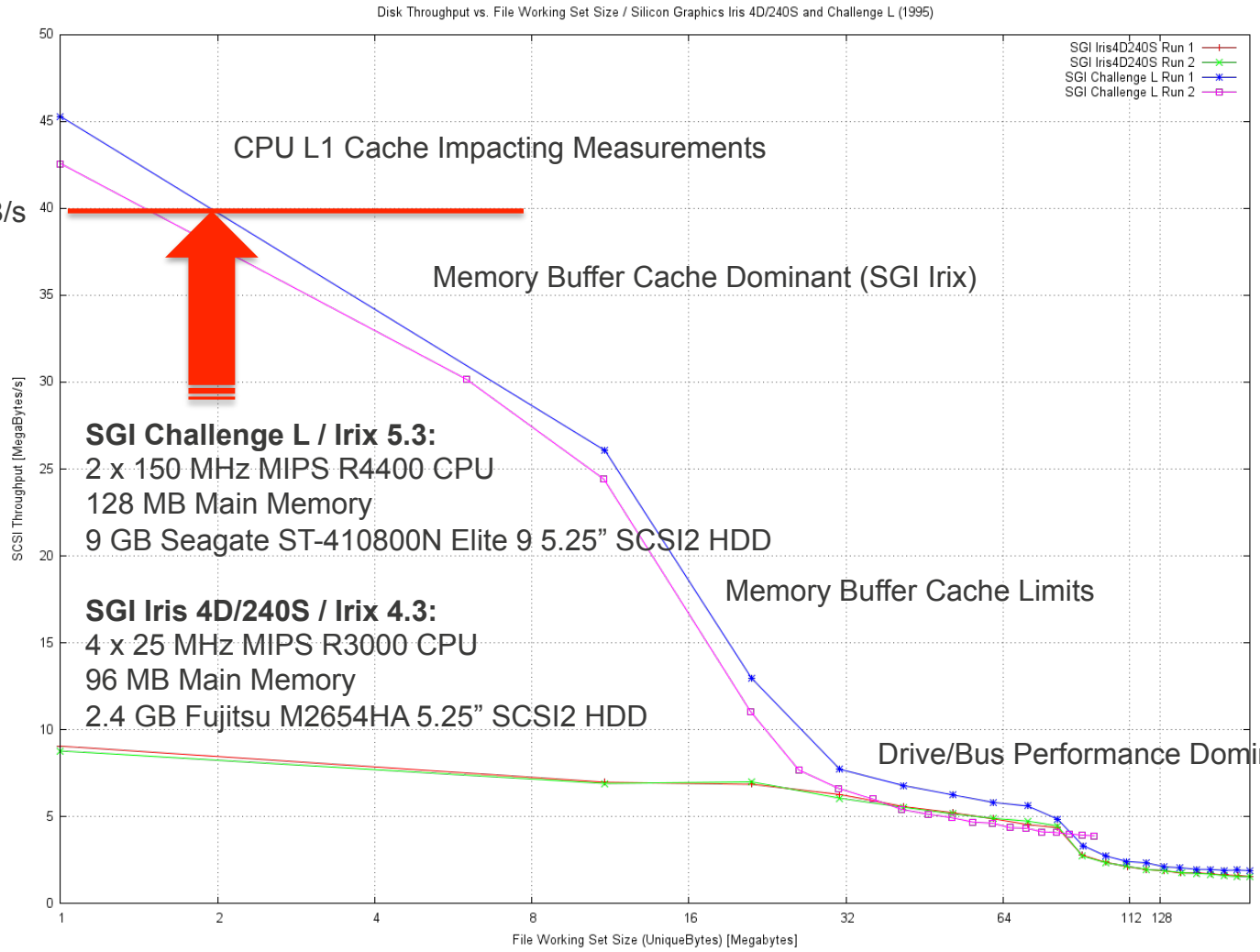


HDD Self Scaling I/O Benchmark (Chen 1992/Kopp 2013)

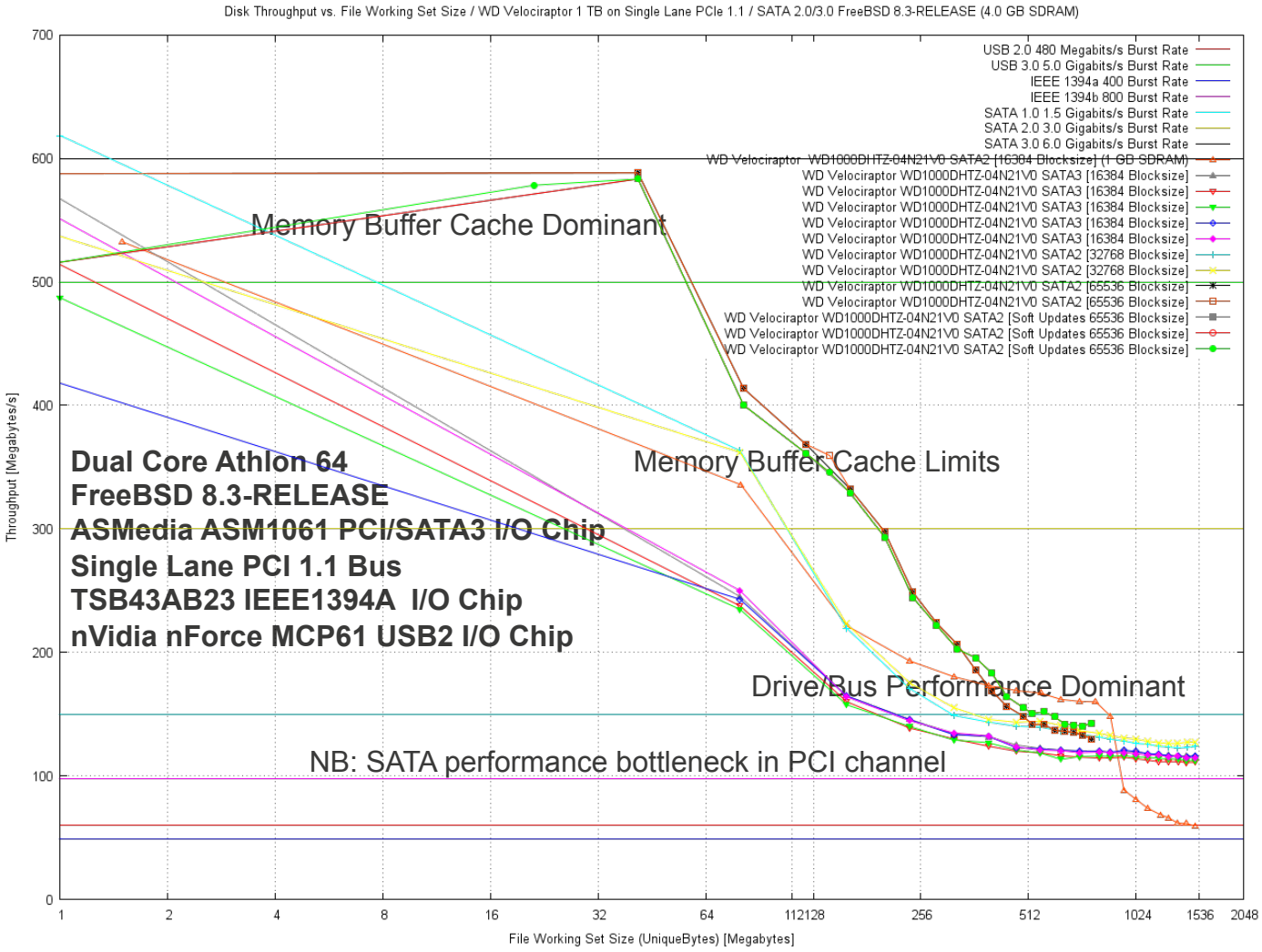
DDR3 10.6 GB/s



HDD Self Scaling I/O Benchmark (Chen 1992/Kopp 1995)

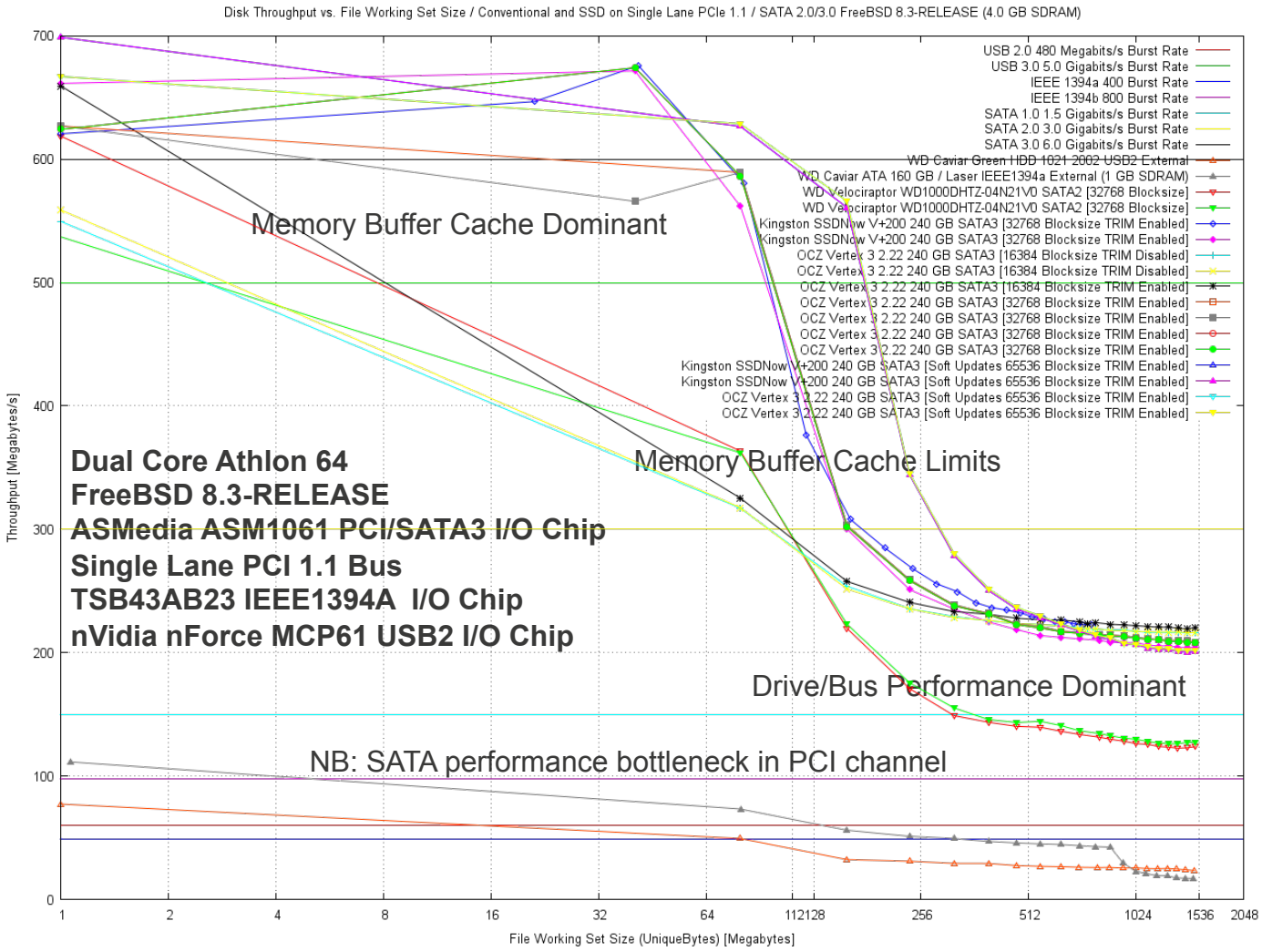


HDD Self Scaling I/O Benchmark (Chen 1992/Kopp 2013)



Not Examinable

SSD Self Scaling I/O Benchmark (Chen 1992/Kopp 2013)



Not Examinable

Samsung 840 Pro SSD / Seagate GoFlex Test (1)

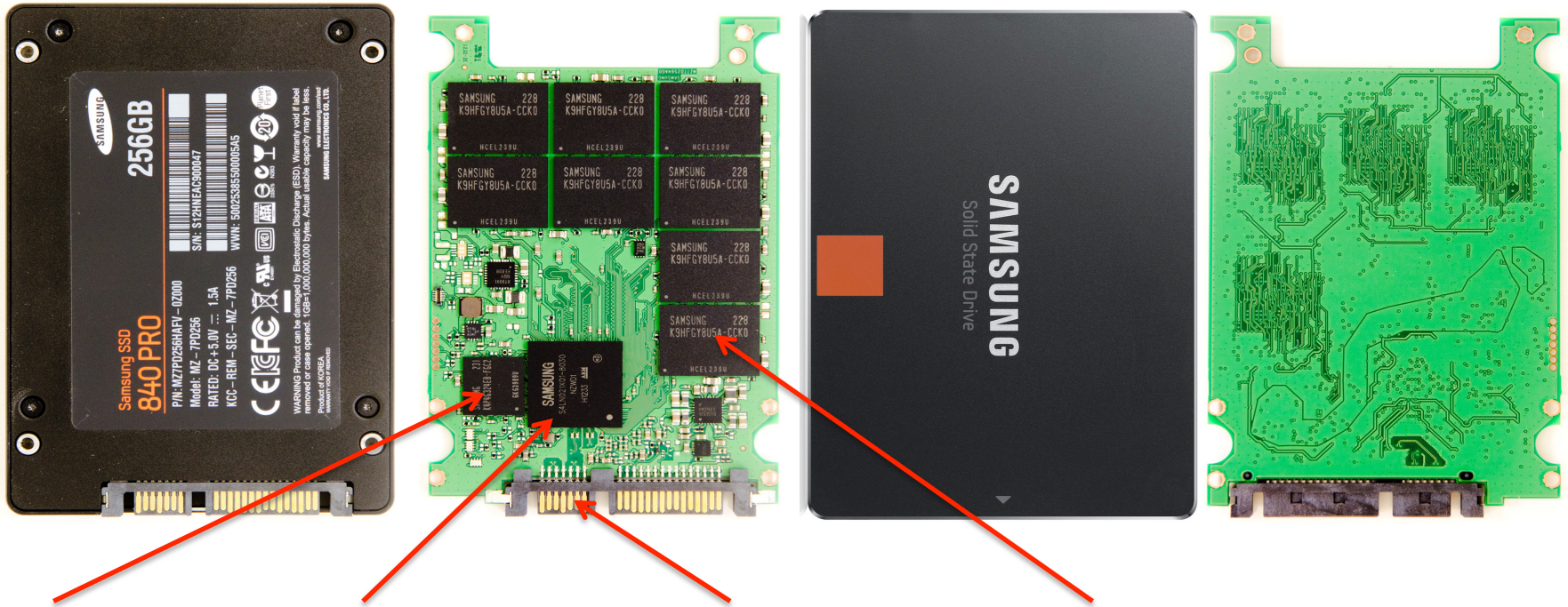
- Samsung 840 Pro 256 GB SSD nominally rated at 540 Megabytes/s in Read and 520 Megabytes/s in Write operations;
- Seagate STAE128 Thunderbolt / SATA3 Adapter nominally rated at 1,000 Megabytes/s throughput;
- Total installation cost including cables and mounting brackets, excluding enclosure chassis of ~ AU$350.00;
- Measured transfer rates – Read: ~390MB/s / Write: ~360MB/s (BMD DST)

Seagate STAE128 Thunderbolt / SATA3 Adapter

- Low Cost Single Drive Adaptor;
- Host Powered Thunderbolt Port to SATA3 Drive Interface;
- Intended for use with Seagate proprietary packaged “GoFlex” 2.5 inch Hard Disk Drives;
- Recently used by enthusiasts for use with 2.5 inch SSDs;
- Known limitations – dependent on host power supply to Thunderbolt port which may sag under load;
- Known to run hot under load.

Samsung 840 Pro SSD 256 Gigabytes

- High performance “Professional grade” SSD developed for desktop systems;
- Lacks write endurance of much more expensive “Enterprise grade” SSDs;



What Next in SSDs?

- Commonly used SATA3 and mini-SATA3 I/O is now at its performance limits for commodity SSDs, and borderline for higher performing SSDs;
- PCIe 2 and 4 Lane SSDs for internal use inside chassis are now available in the market, supported by proprietary drivers mostly for MS Windows;
- DDR3 DIMM main memory bus SSDs – *Sandisk ULLtraDIMM* uses standard DDR3 SDRAM packaging and bus interfaces, supported by proprietary drivers for Linux, Windows, VMs;
- As with previous advances in hardware, uptake is slow as new hardware is poorly supported by software due to the absence of a common interface standard;
- *There is a significant gap between the operating systems and applications base, and advancing SSD / bussing technology!*

Software vs SSDs?

- What gaps currently exist between SSD hardware and software?
- *Operating systems* currently treat an SSD as a rotating disk storage device – filesystems, drivers, utilities cannot exploit unique low latency properties of the SSD;
- *Log-structured FileSystems* (LFS – Ousterhout / Seltzer) are optimal for SSDs, as read fragmentation is irrelevant in SSDs, and write behaviour will minimise write/erase cycles – there are no widely used LFS at this time;
- While SATA and Thunderbolt SSDs can exploit existing OS/Driver/FS support for legacy HDD hardware, PCIe and DDR3 SSDs remain "orphans" due to scarce proprietary software, and absence of common standards for interface behaviour;
- As with GPUs, applications mostly built around legacy H/W.

APPLICATION DESIGN TO EXPLOIT SSD

Application Design to Exploit SSD Performance

- SSDs offer affordable hundreds of Gigabytes up to Terabytes of storage with latency performance 10^2 times lower than rotating mass storage, and transfer rates competitive against older DDR SDRAM main memory hardware;
- The future will see lower latency and higher transfer rates, as Flash based storage overlaps performance of slower DDR style main memory;
- Most current applications assume mass storage using high latency rotating media, and fast main memory which is expensive and difficult to deploy above 8-16 GB;
- Data intensive and many memory intensive applications could gain enormously from optimisation for SSD hardware;
- The best case study is the NSF funded Gordon HPC / supercomputer at SDSC in the US (Norman, 2013).

SDSC Gordon Performance Highlights (Norman 2012)

Data intensive but “storage-latency-bound” / “storage latency limited” applications running on Gordon exhibited significant performance gains:

- *RCSB Protein Data Bank* (PDB): 3.8 to 5.8 X faster processing;
- NSF funded *OpenTopography Facility* (LIDAR Datasets): ~20 X faster processing;
- *IntegromeDB* biomedical database (50 TB PostgreSQL): 50 X faster database operations, 10 X faster file read / write operations;
- *National Center for Microscopy and Imaging Research* (NCMIR): 12 X faster image processing;
- All of these applications were run on the Gordon system’s “I/O Nodes”;
- Each of the 32 x I/O Node has 9.6 Terabytes (2 x 4.8 TB) of SSD Flash RAM, yielding an aggregated ~300 TB of SSD Flash RAM;
- Observed performance gains between 3.8:1 up to 50:1, compared to previous.

Optimising for SSD Performance

- Several steps necessary:

1. Understand data access patterns to disk and to main memory, identify extant optimisations for legacy hardware;
2. Identify most critical time complexity behaviours in mass storage access;
3. Modify and adapt to exploit SSD;

- During legacy application design, implicit assignment of per operation costs in time complexity modelling - assumed high latency for all mass storage;
- Some memory resident structures may be migrated to SSD from main memory without major performance penalties;
- Many mass storage resident structures will display dramatically different time complexity on SSD storage;

Conclusions

- SSDs are a “disruptive technology”, which challenges most “traditional” design and performance assumptions made for mass storage devices;
- SSDs are sufficiently mature in the niche of high performance rotating media replacement for general use;
- High performance SSD interfaces such as PCIe and DDR3 are still poorly supported by operating systems, due to a lack of industry standards, and hardware remains expensive due to slow uptake;
- Extant software application base crafted for decades around the implicit latency, and often transfer rate, limitations of rotating mass storage media;
- *Significant research opportunities for improving performance of data intensive applications, e.g. simulations, modelling and data mining.*

BACKUP SLIDES

Aggregated Storage Subsystem Behaviour

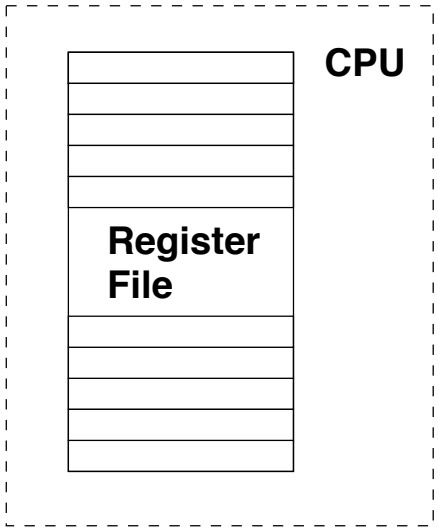
- Storage subsystems exhibit often complex queuing behaviours, as they behave as serial chains of queuing systems;
- Hard Disk -> Hard Disk Cache -> Filesystem -> Storage Bus -> I/O Controller -> Main Bus -> Memory Controller -> Memory Bus -> Memory -> Main Memory Buffer Cache;
- Measurements and modelling should account for queuing behaviours, including asymptotic saturation effects under heavy loads;
- Simple measurements of individual device and component performance are useful, but cannot represent performance of storage device as part of a storage subsystem, or storage subsystem as part of a computer system; Distributed storage using networked storage components also exhibits queuing behaviours;
- Conclusion is that benchmarking at system level required for accuracy.

Impacts of Exponential Growth in Storage

- Commodity hard disks of 2-4 TB capacity now in $100-$400 retail price range;
- Commodity SSDs of up to 1 TB now discounted at retail prices below $1,000 – *good buys are < $1/Gigabyte*;
- Commodity hardware now supplied with USB3 busses delivering ~400 MB/s, SATA3 busses ~ 500 MB/s; Thunderbolt (PCIe / PCI Express) busses ~1,000 – 2,000 MB/s;
- Main Memory DDR3 SDRAM performance ~ 17-34 GB/s;
- Critical Considerations: Bandwidth vs. Access Time vs. Cost / Capacity;
- *The SSD is now a "disruptive technology" in mass storage;*
- *Impacts similar to solid state RAM vs. magnetic core memory 40 years ago.*
- *Non-impact: Programming practice and application design has yet to align with exponentially growing storage density and bus performance;*

Storage Hierarchies – One Decade Ago

Size ~ hundreds of bytes
Very Expensive
cost ~ $2000/Megabyte

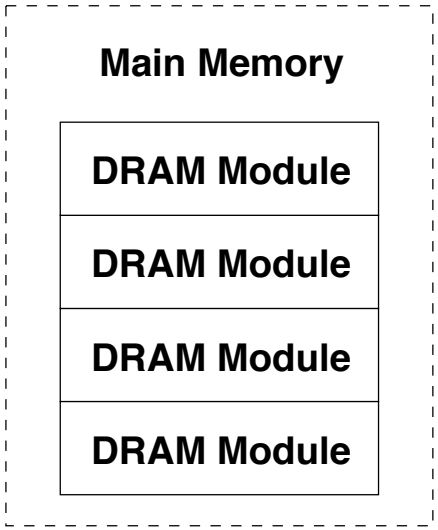


Very Fast

access time ~ nanosecs

~1 ns

Size ~ hundreds of Megabytes
Moderately Expensive
cost ~ $1.50/Megabyte

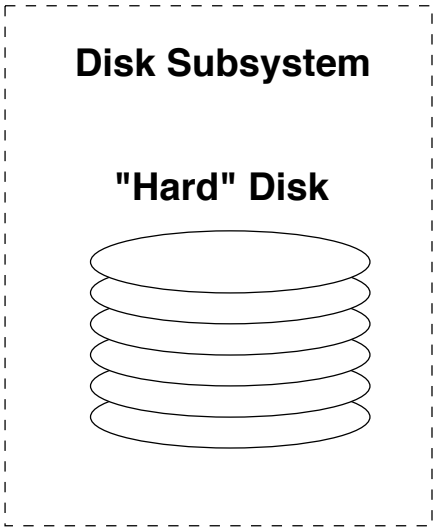


Moderately Fast

access time ~ tens of nanosecs

~50 ns

Size ~ tens of Gigabytes
Very Cheap
cost ~ $0.15/Megabyte

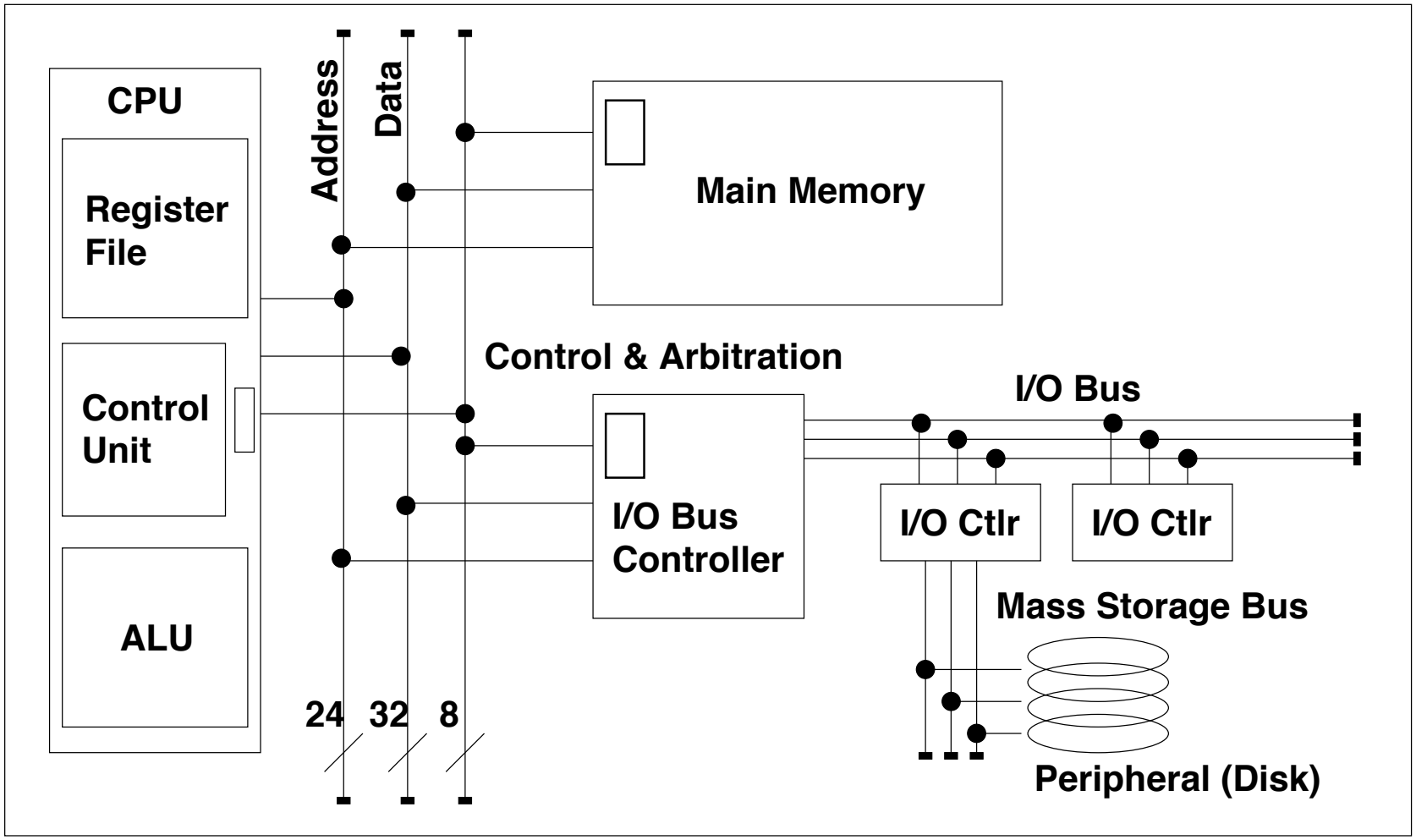


Very Slow

access time ~ millisecs

~5,000,000 ns

Bus Heirarchies



WD Velociraptor 10,000 RPM 3.5 inch Disk

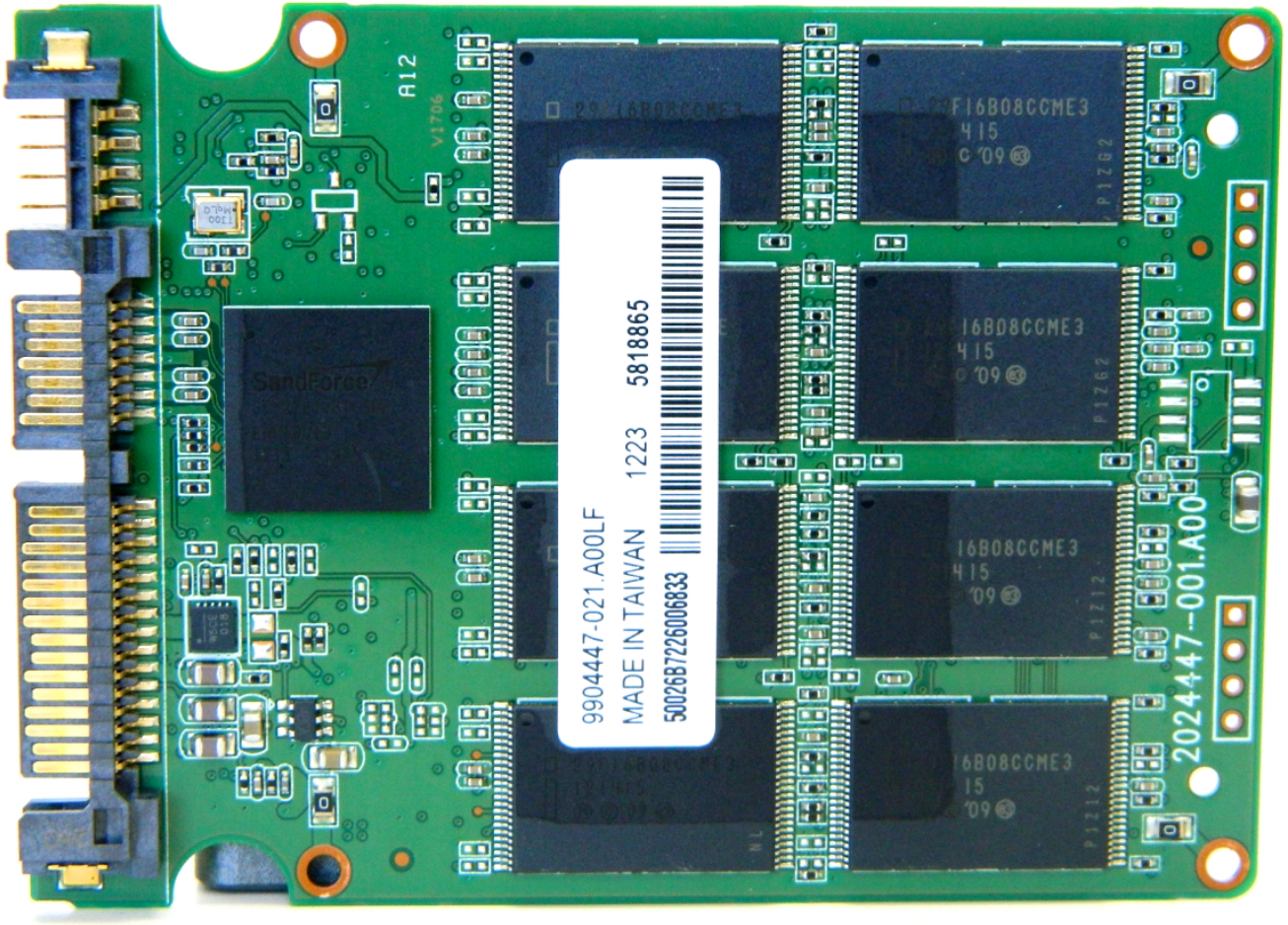
Seagate Cheetah 15,000 RPM 3.5 inch Disk

Benchmarking HDD Transfer Rate Performance

- *Consider a 2012 built 2.5" 10,000 RPM 1TB SATA 3 Disk Drive:*
- **Theoretical SATA 3 limit ≈ 480 MB/s; Measured = 400.8 MB/s;**
- Measured Best Drive Transfer Rate = 209.1 MB/s;
- Measured Worst Drive Transfer Rate = 114.7 MB/s;
- Theoretical SATA 1 limit ≈ 160 MB/s;
- Measured Best DTR (iMac 5,1/Snow Leopard) = 121.8 MB/s;
- *Consider a 2012 built 3.5" 7,200 RPM 1TB SATA 3 Disk Drive:*
- Measured Best DTR (iMac 12,2/Snow Leopard) = 113.6 MB/s;
- *Consider SATA 2/3 Disk Drives in USB2 Enclosures (≈ 40 MB/s):*
- Measured Best DTR (iMac 5,1/Snow Leopard) = 36.4 MB/s;
- Measured Best DTR (iMac 12,2/Snow Leopard) = 31.7 MB/s;
- Measured Best DTR (MM 4,1/Mountain Lion) = 39.1 MB/s;
- Comparison: SATA 3 SSD Measured DTRs ≈ 200 – 500 MB/s

Kingston SSDNow V+200 SSD – 2.5 inch SATA3

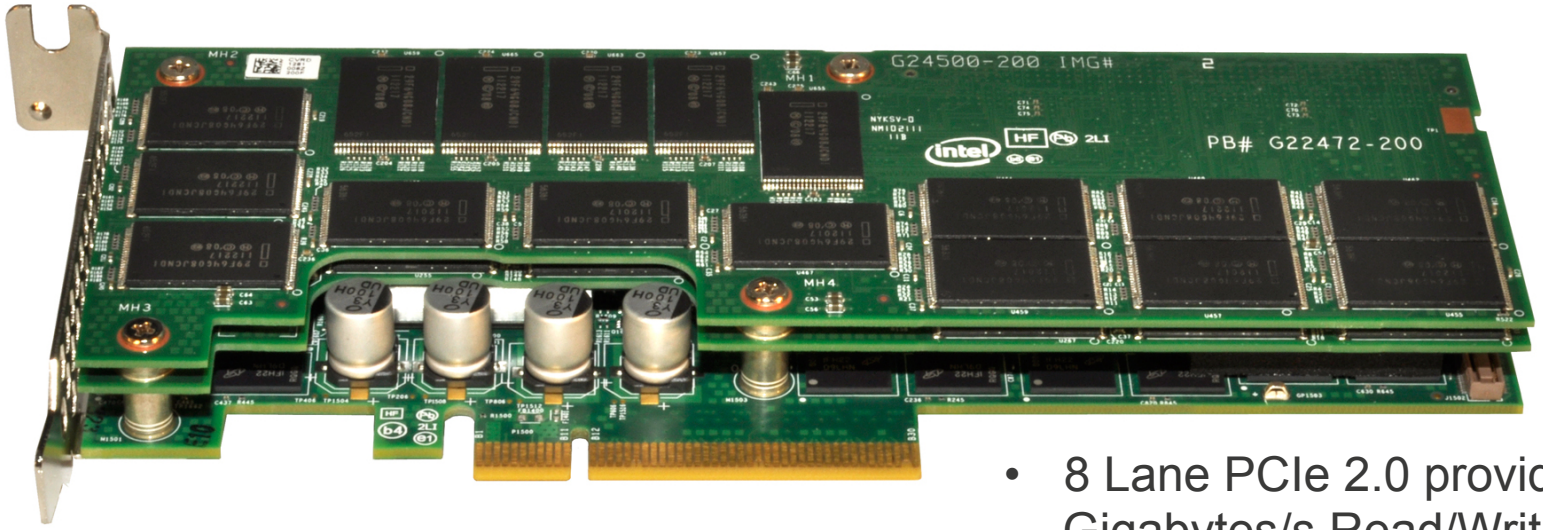
Kingston SSDNow V+200 SSD – 2.5 inch SATA3

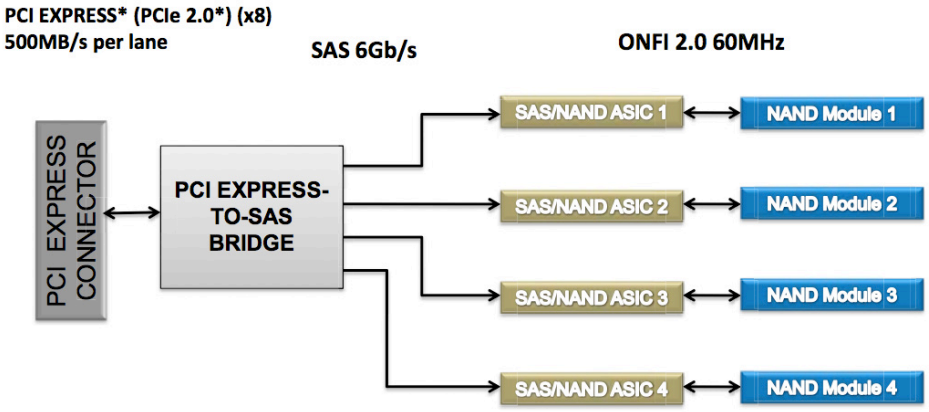


DDR3 I/O to SSD Storage - SMART ULLtraDIMM

- DDR3 memory bus providing 1 and 0.76 Gigabytes/s Read/Write aggregate transfer rates for sequential data – limited by Flash RAM/Controller hardware;
- Capacity of 200 GB – 400 GB, using RAID-like “Flexible Redundant Array of Memory Elements” overprovisioning;
- Very loow access latency due to parallel I/O via DDR3 memory bus;
- Intended for Enterprise server systems with large numbers of DDR3 slots;
- Limitations: drivers available only for Windows Server 2008/12, Centos, RedHat, SUSE Linux, and VMware;
- First DDR3 packaged SSD to appear in the market.

PCIe I/O to SSD Storage – Intel 910 Series





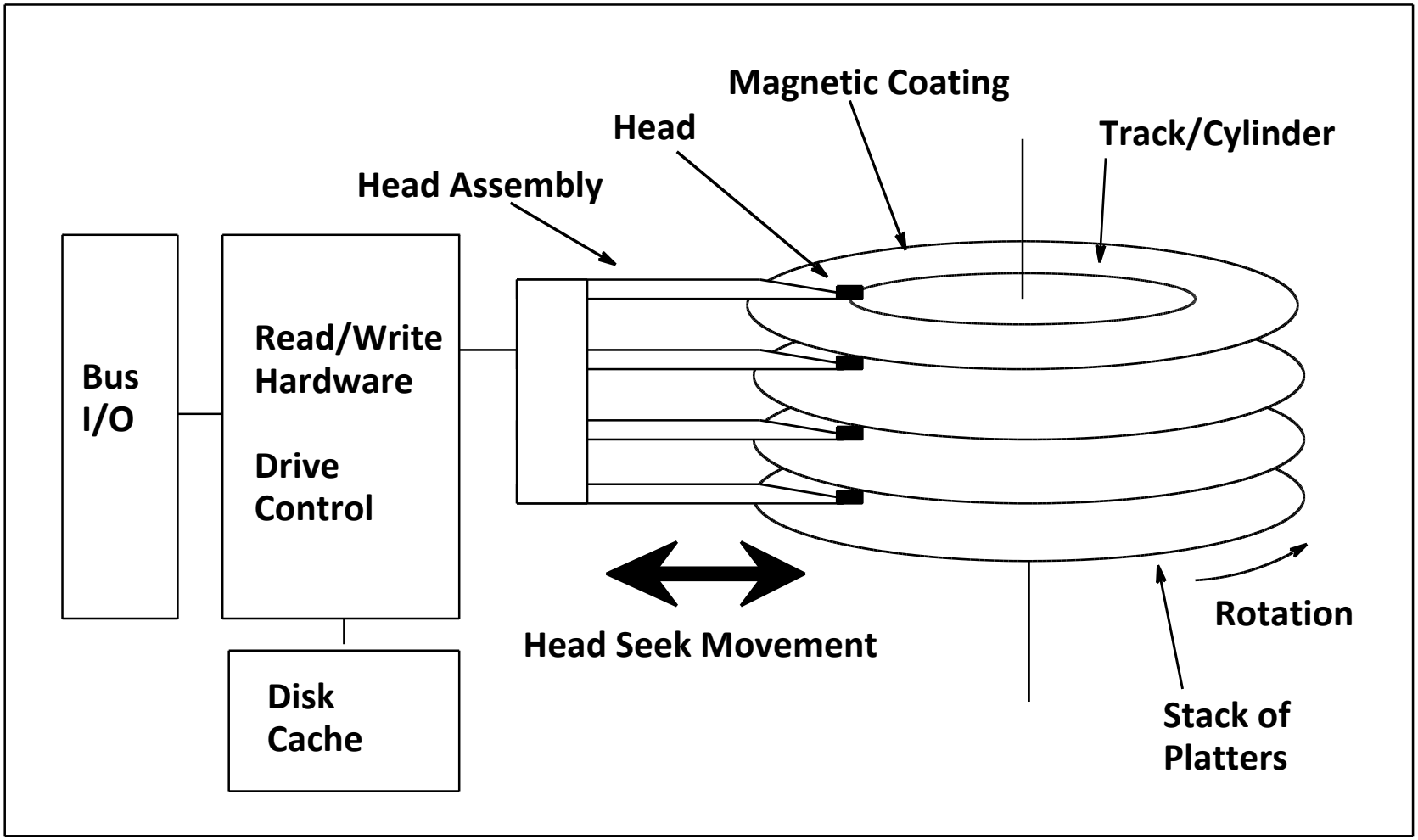
- 8 Lane PCIe 2.0 providing 1 – 2 Gigabytes/s Read/Write aggregate transfer rates;
- Capacity of 400 GB – 800 GB with overprovisioning (896GB / 1792GB) to extend SSD life;
- Pricing is ~5.6 X commodity SSDs due to enterprise grade storage devices, redundancy;

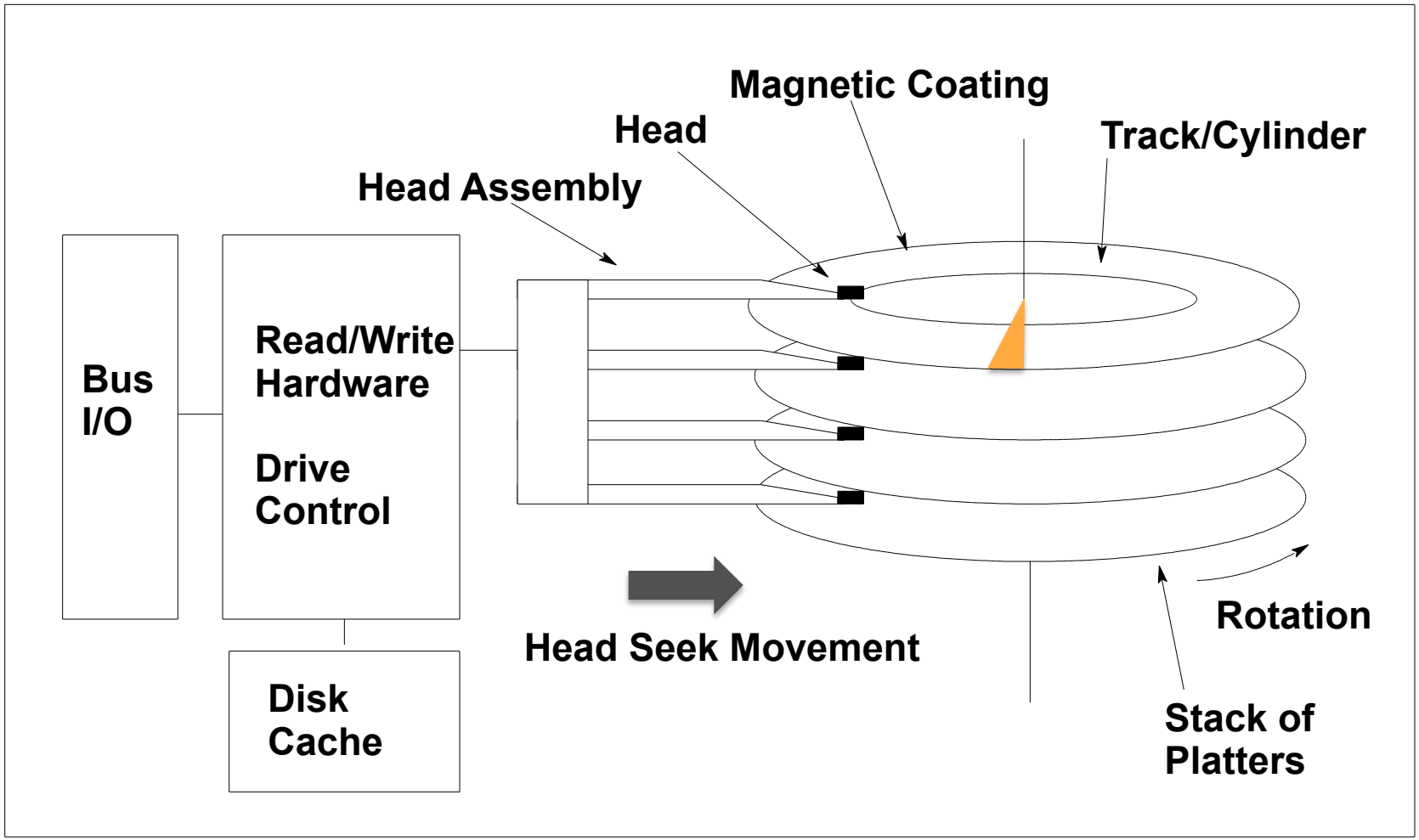
PCIe I/O to SSD Storage – OCZ RevoDrive 3 X2

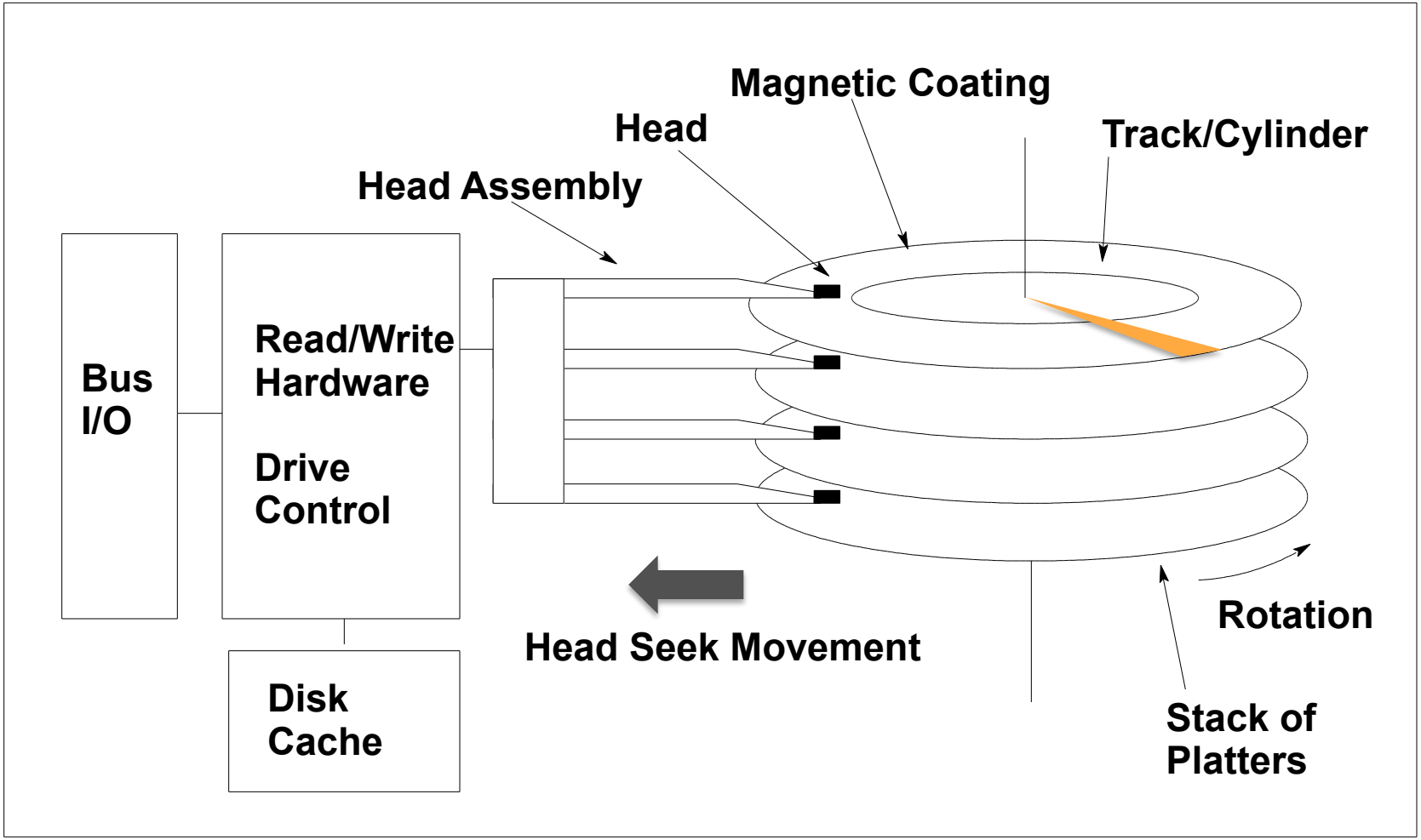
- 4 Lane PCIe 2.0;
- ~1.5/1.3 GB/s R/W;
- Intended for gaming, multimedia users;
- Capacities available:
  - 240 GB;
  - 480 GB;
  - 960 GB;
- Embedded ECC with 55 bits correctable per 512-byte sector (BCH);
- Driver support only for MS Windows;

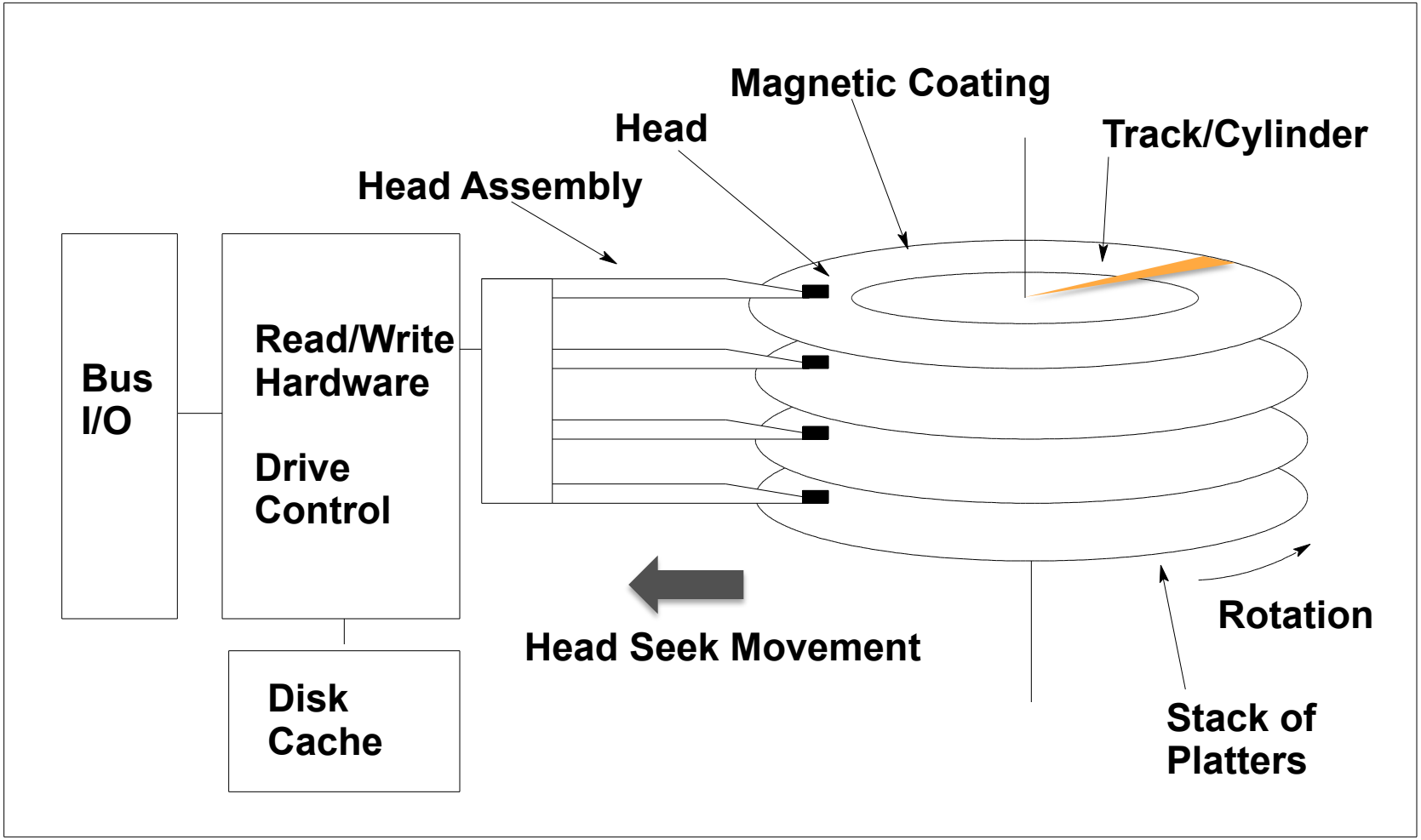
Rotating Disk Operating Principles

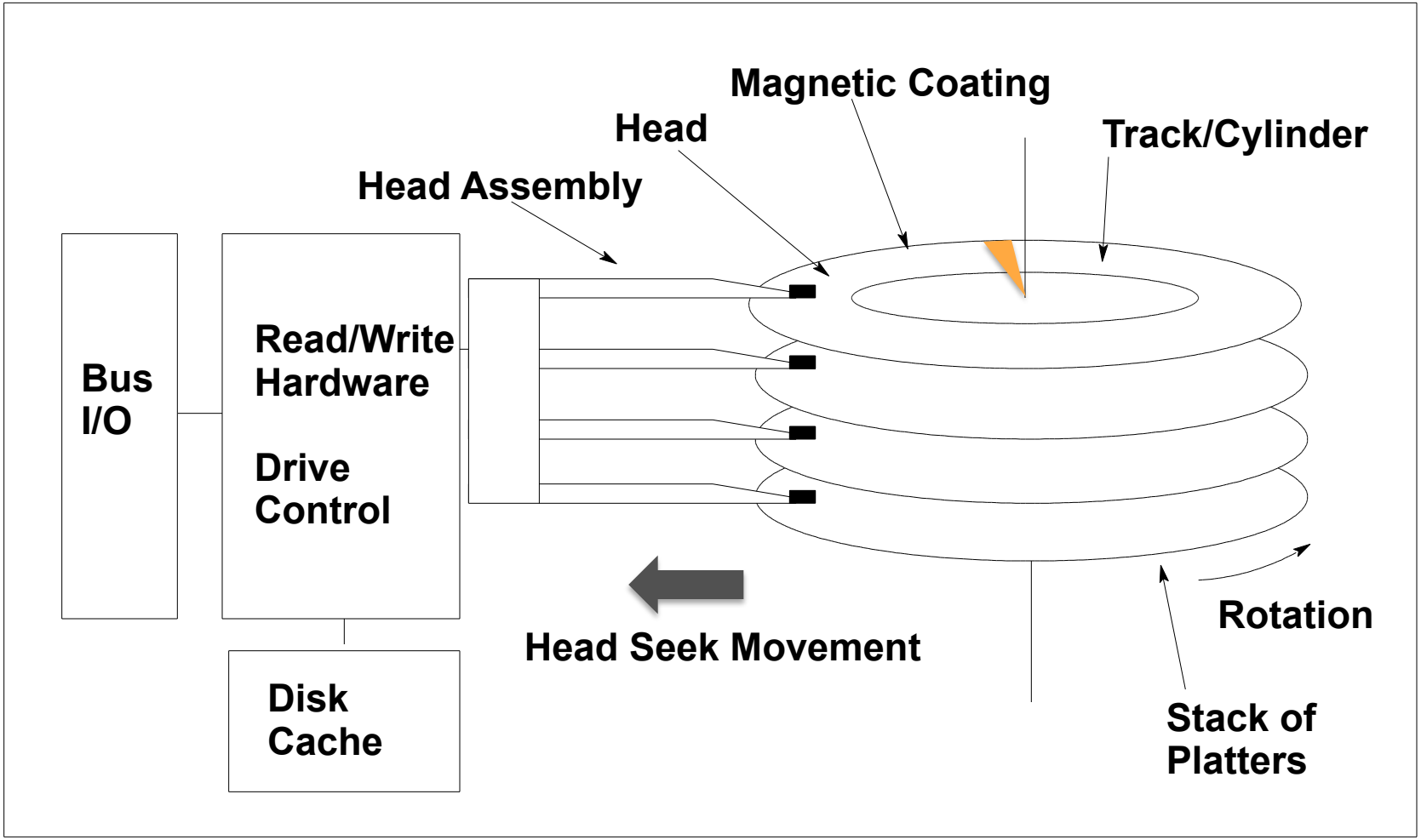
- Disk drives use a magnetic coating on a rotating disk platter to store information;
- Data is written serially in “cylinders”;
- Each cylinder comprises a very large number of consecutive data “blocks”, 512 or 4096 bytes each;
- An electromagnetic transducer, termed a “head” is used to write or read from the disk surface;
- A pivoting mechanical arm or translating mechanical finger is used to position the head over the part of the disk where data is to be read or written;

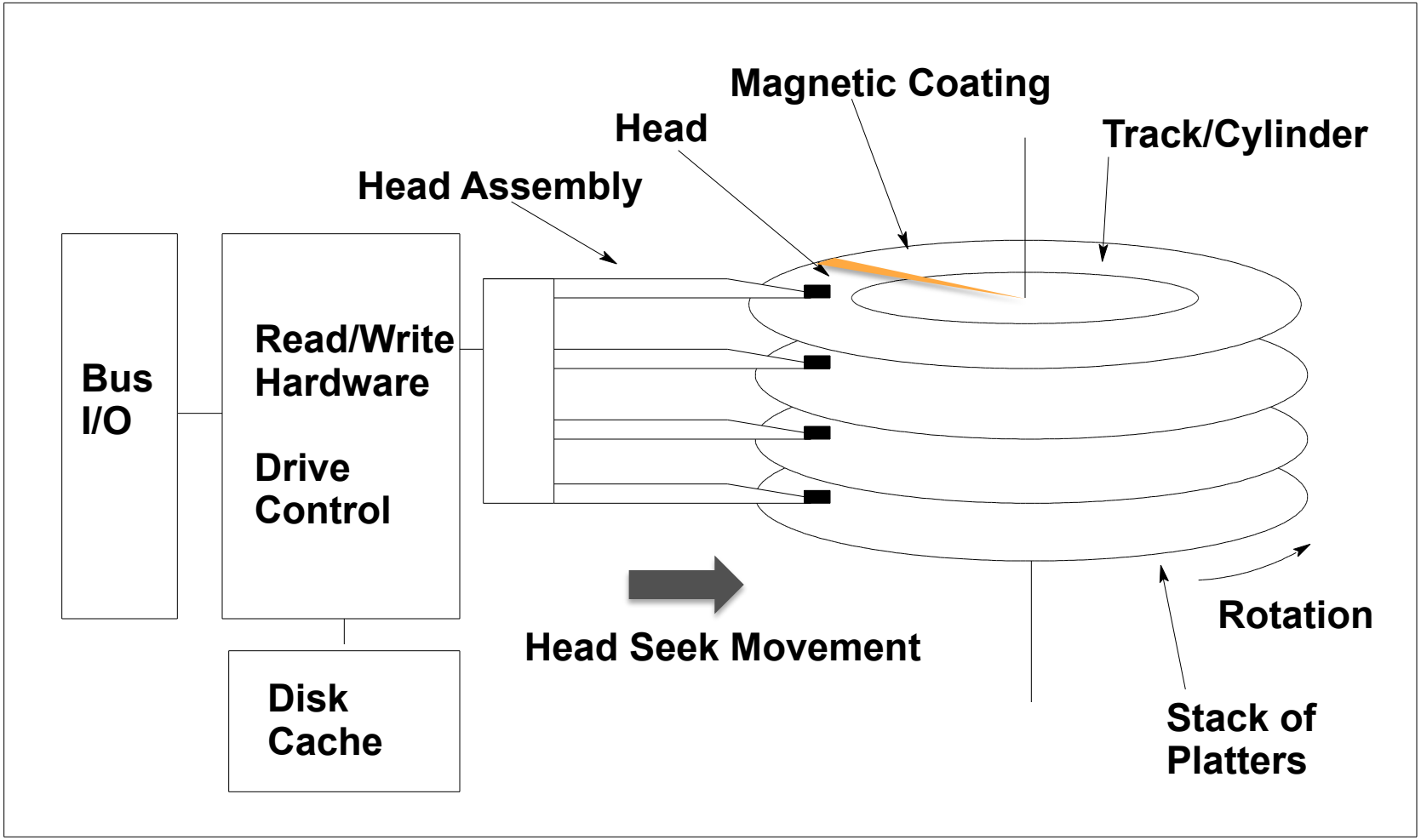

Magnetic Coating
Head
Track/Cylinder
Head Assembly
Bus
I/O
Read/Write
Hardware
Drive
Control
Head Seek Movement
Rotation
Disk
Cache
Stack of
Platters

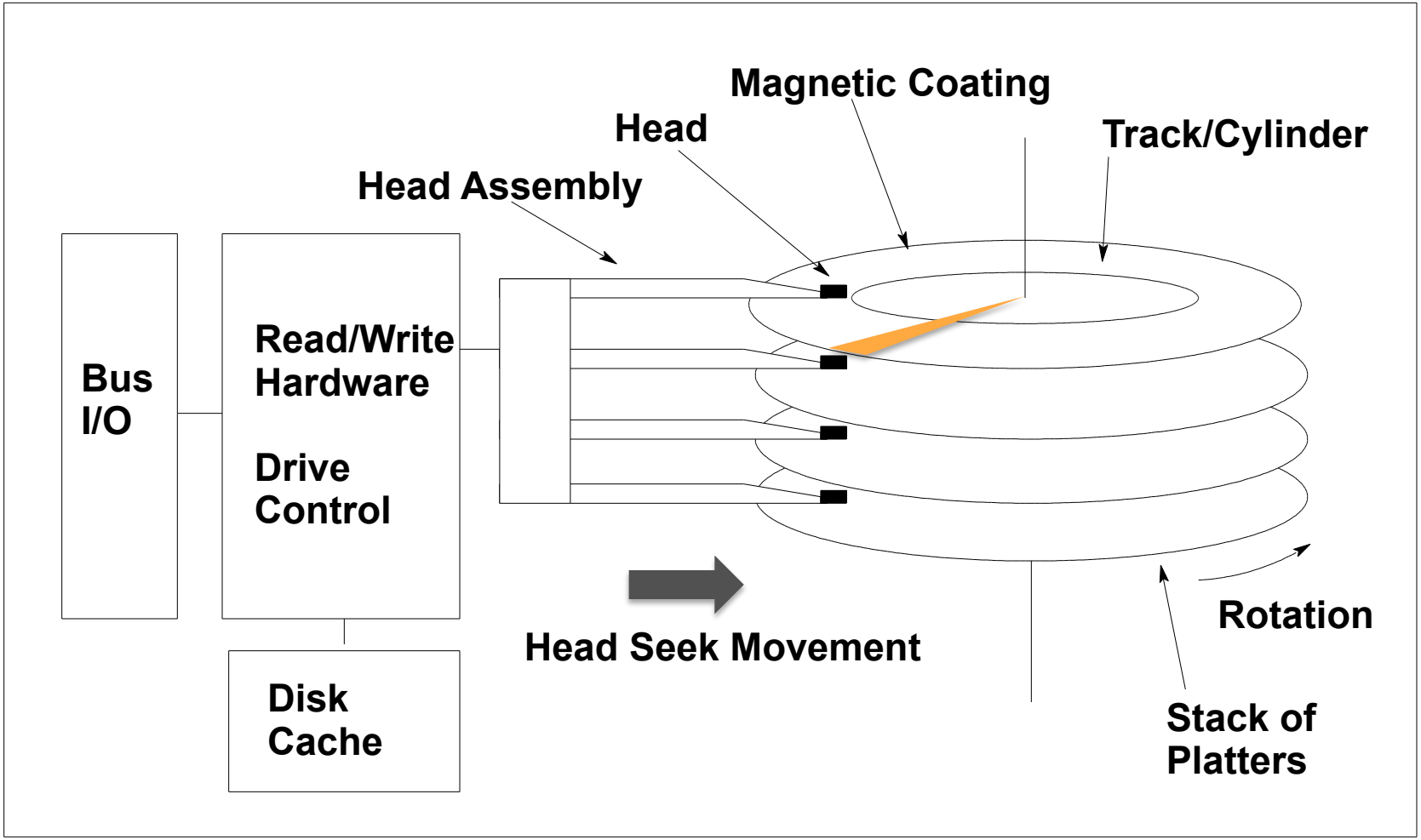

Magnetic Coating
Head
Track/Cylinder
Head Assembly
Bus
I/O
Read/Write
Hardware
Drive
Control
Disk
Cache
Head Seek Movement
Rotation
Stack of
Platters

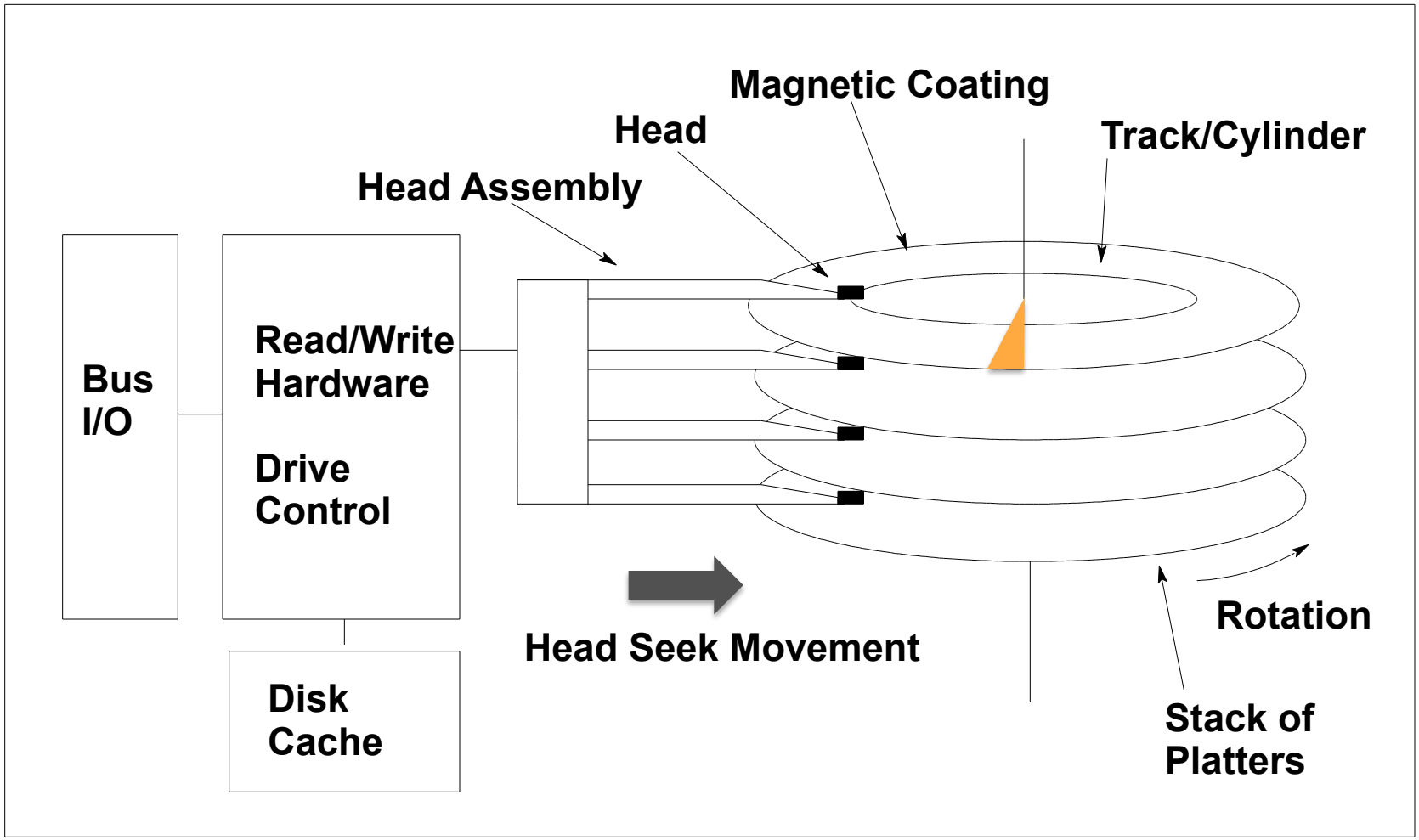

Magnetic Coating
Head
Track/Cylinder
Head Assembly
Bus
I/O
Read/Write
Hardware
Drive
Control
Disk
Cache
Head Seek Movement
Rotation
Stack of
Platters

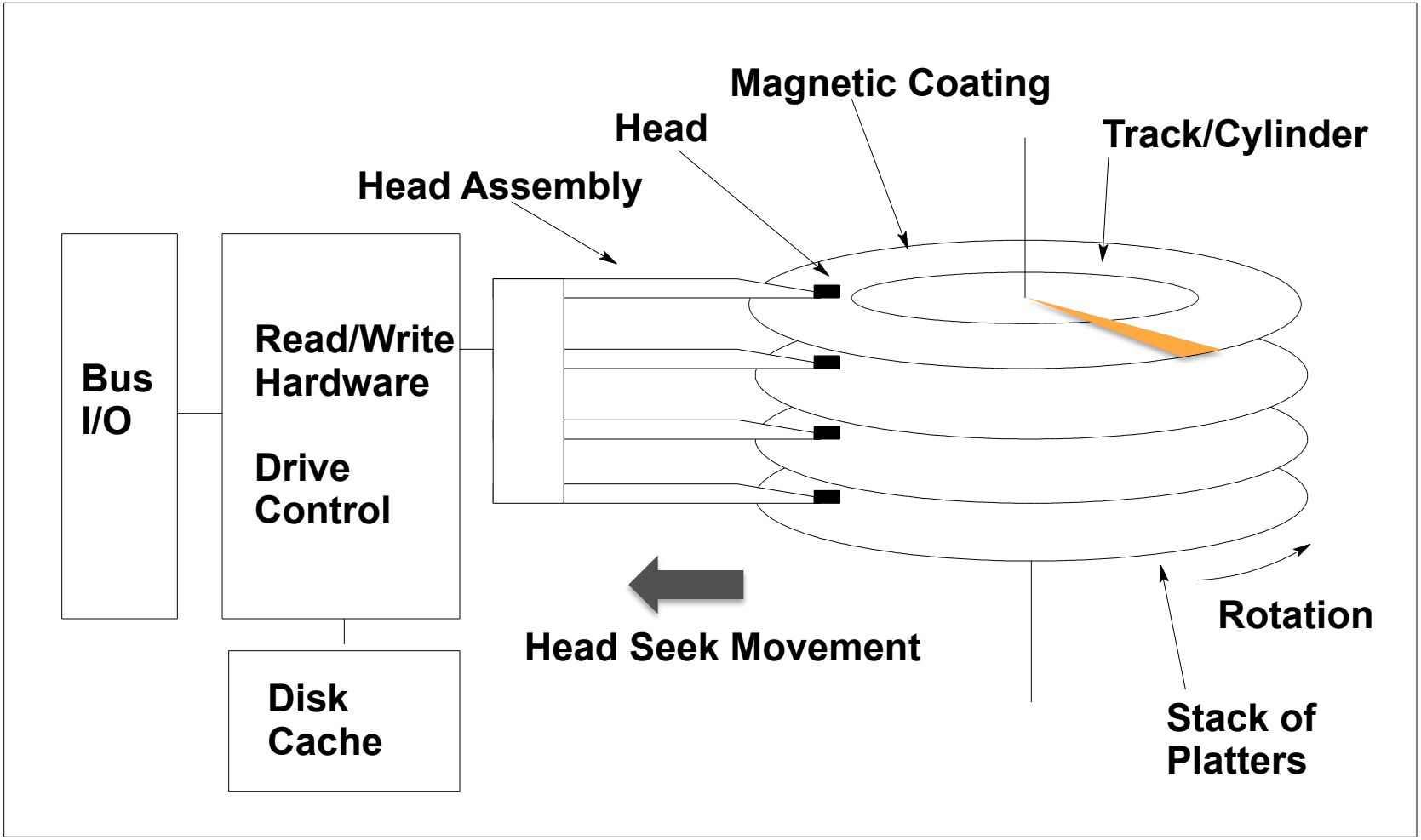

Magnetic Coating
Head
Track/Cylinder
Head Assembly
Bus
I/O
Read/Write
Hardware
Drive
Control
Disk
Cache
Head Seek Movement
Rotation
Stack of
Platters

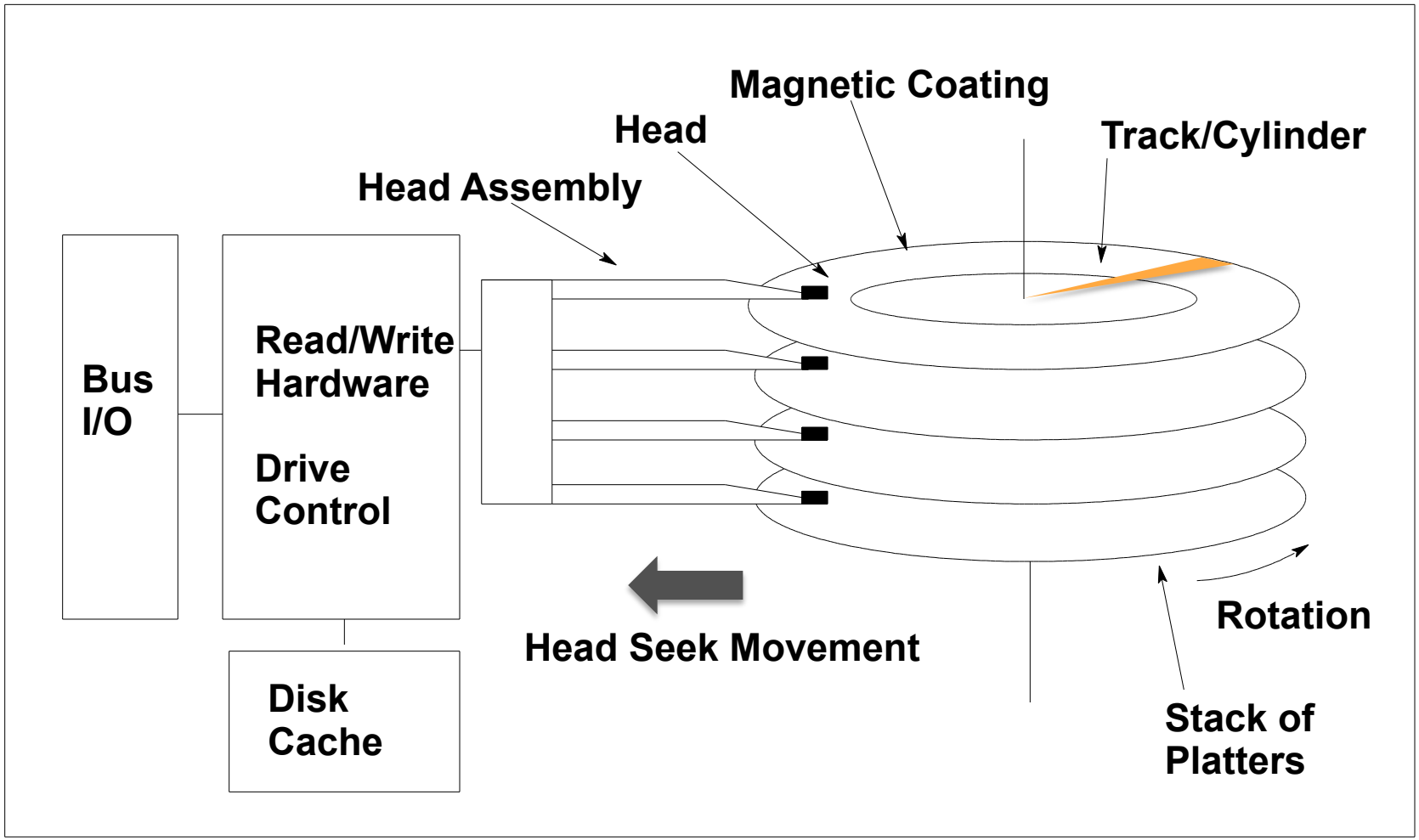

Magnetic Coating
Head
Track/Cylinder
Head Assembly
Bus
I/O
Read/Write
Hardware
Drive
Control
Head Seek Movement
Rotation
Disk
Cache
Stack of
Platters

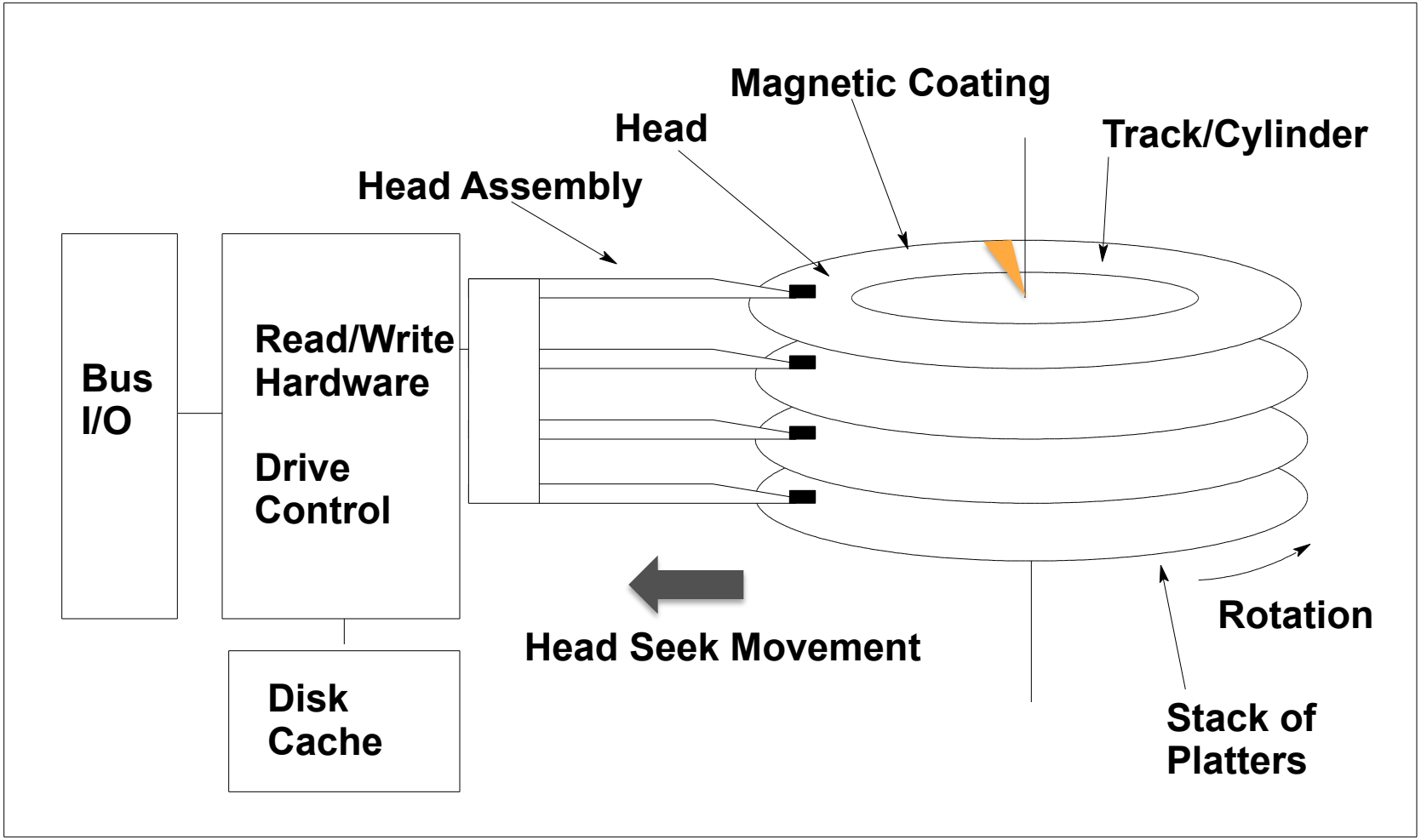

Magnetic Coating
Head
Track/Cylinder
Head Assembly
Bus
I/O
Read/Write
Hardware
Drive
Control
Disk
Cache
Head Seek Movement
Rotation
Stack of
Platters

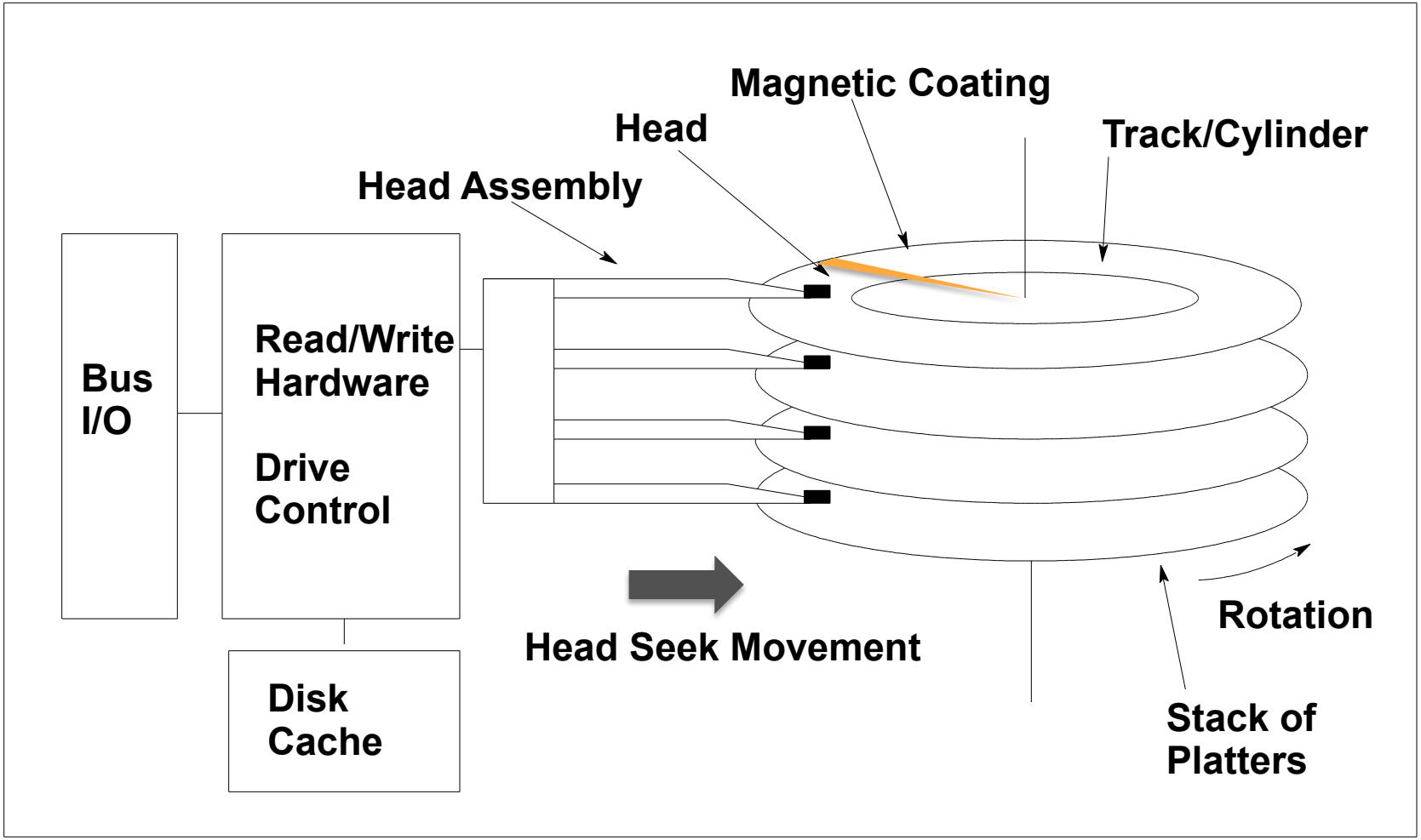

Magnetic Coating
Head
Track/Cylinder
Head Assembly
Bus
I/O
Read/Write
Hardware
Drive
Control
Disk
Cache
Head Seek Movement
Rotation
Stack of
Platters

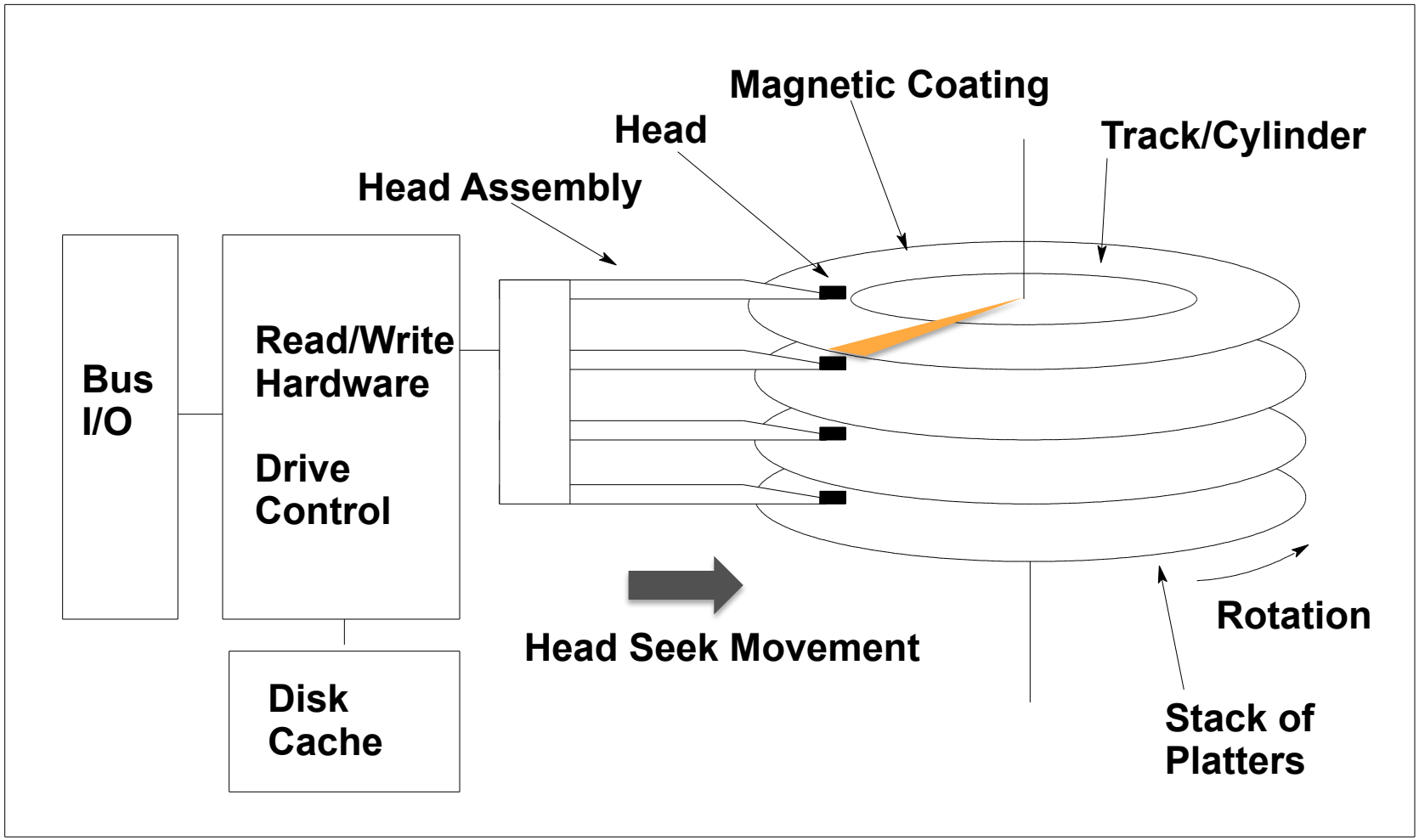

Magnetic Coating
Head
Track/Cylinder
Head Assembly
Bus
I/O
Read/Write
Hardware
Drive
Control
Disk
Cache
Head Seek Movement
Rotation
Stack of
Platters


Magnetic Coating
Head
Track/Cylinder
Head Assembly
Bus
I/O
Read/Write
Hardware
Drive
Control
Disk
Cache
Head Seek Movement
Rotation
Stack of
Platters


Magnetic Coating
Head
Track/Cylinder
Head Assembly
Bus
I/O
Read/Write
Hardware
Drive
Control
Head Seek Movement
Rotation
Disk
Cache
Stack of
Platters


Magnetic Coating
Head
Track/Cylinder
Head Assembly
Bus
I/O
Read/Write
Hardware
Drive
Control
Head Seek Movement
Rotation
Disk
Cache
Stack of
Platters


Magnetic Coating
Head
Track/Cylinder
Head Assembly
Bus
I/O
Read/Write
Hardware
Drive
Control
Disk
Cache
Head Seek Movement
Rotation
Stack of
Platters


Magnetic Coating
Head
Track/Cylinder
Head Assembly
Bus
I/O
Read/Write
Hardware
Drive
Control
Disk
Cache
Head Seek Movement
Rotation
Stack of
Platters