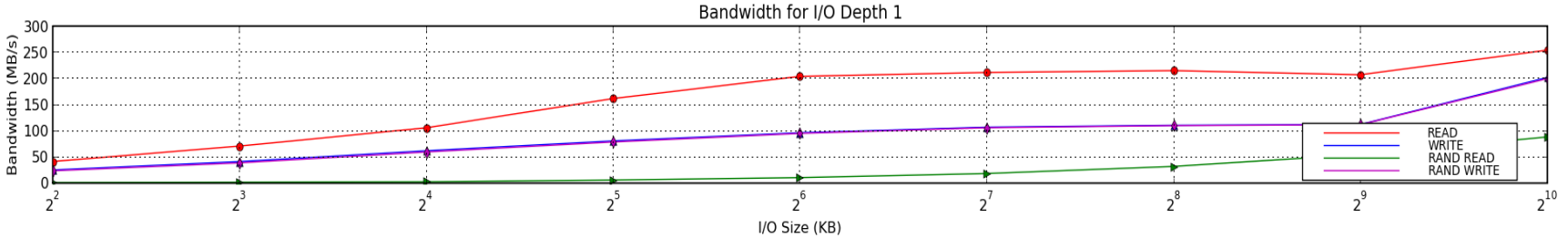
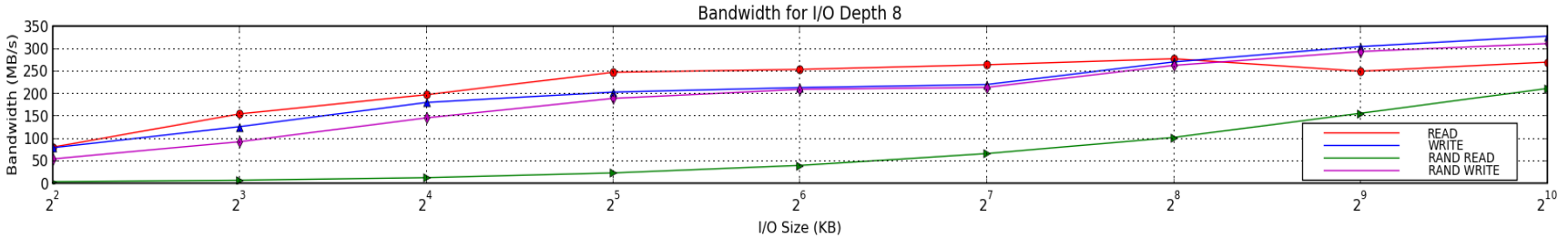
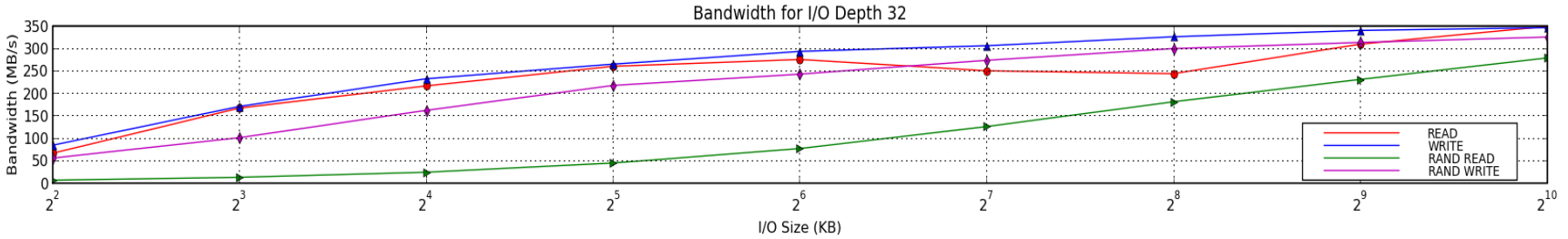
# A Complete Guide to SSDs

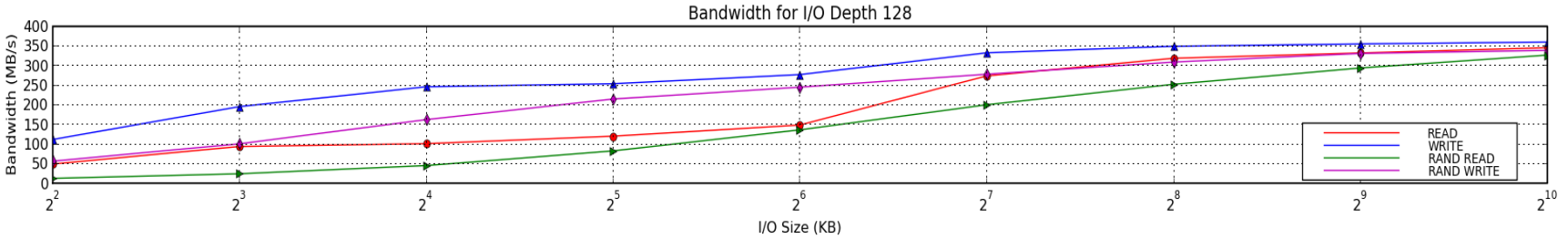
Jeff Moyer
Principal Software Engineer
Red Hat, Inc.
June 4th, 2010

aio-stress output for: 1 thread, 1 files

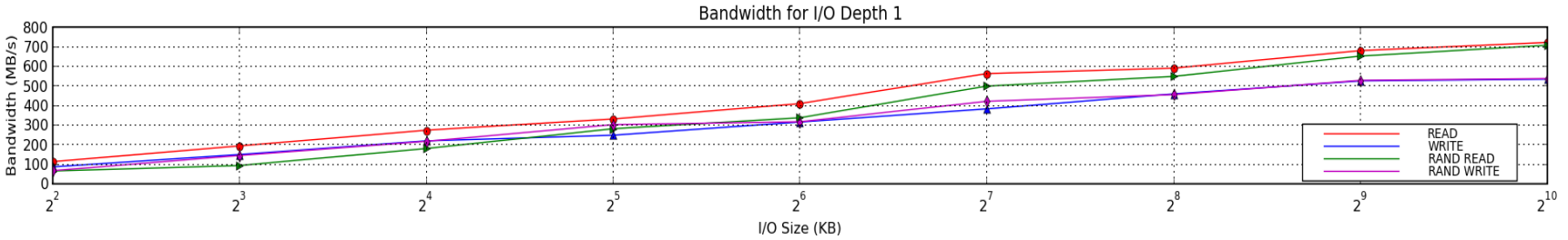


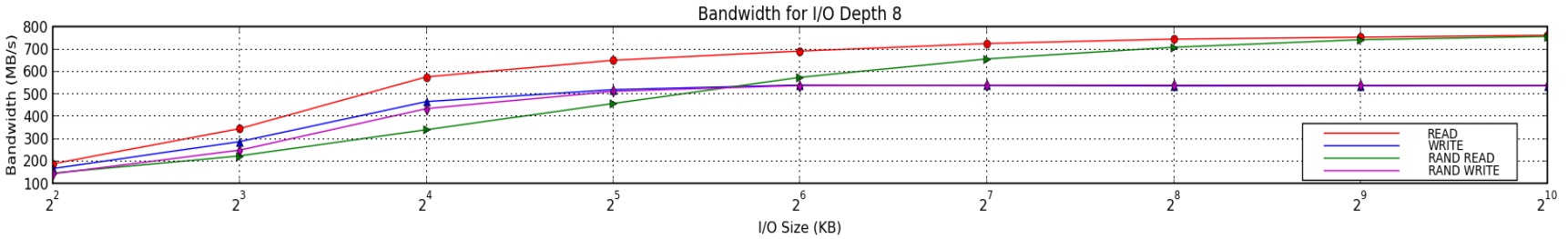



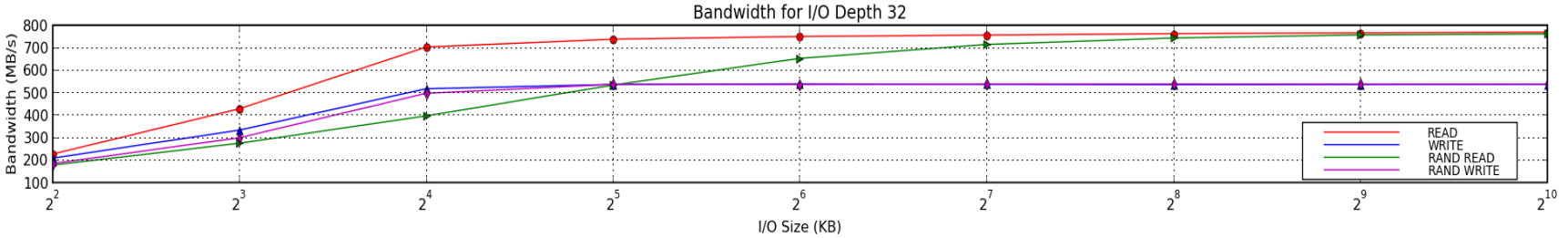



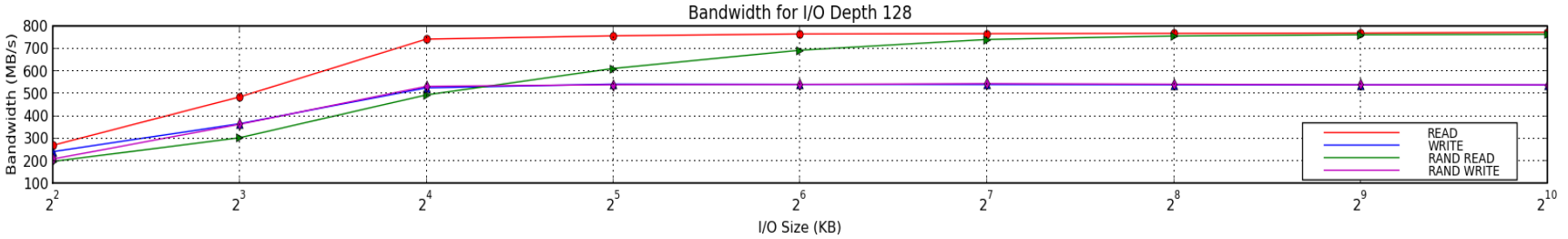



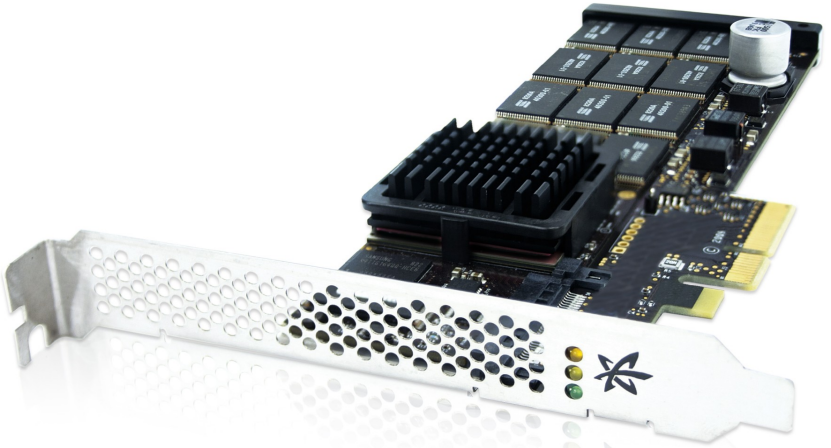
aio-stress output for: 1 thread, 1 files



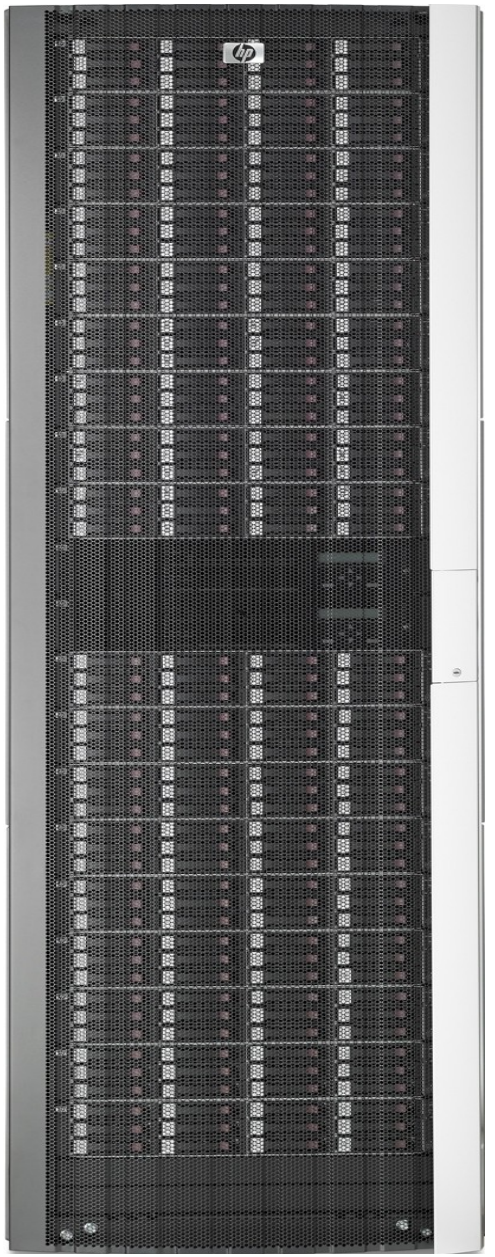



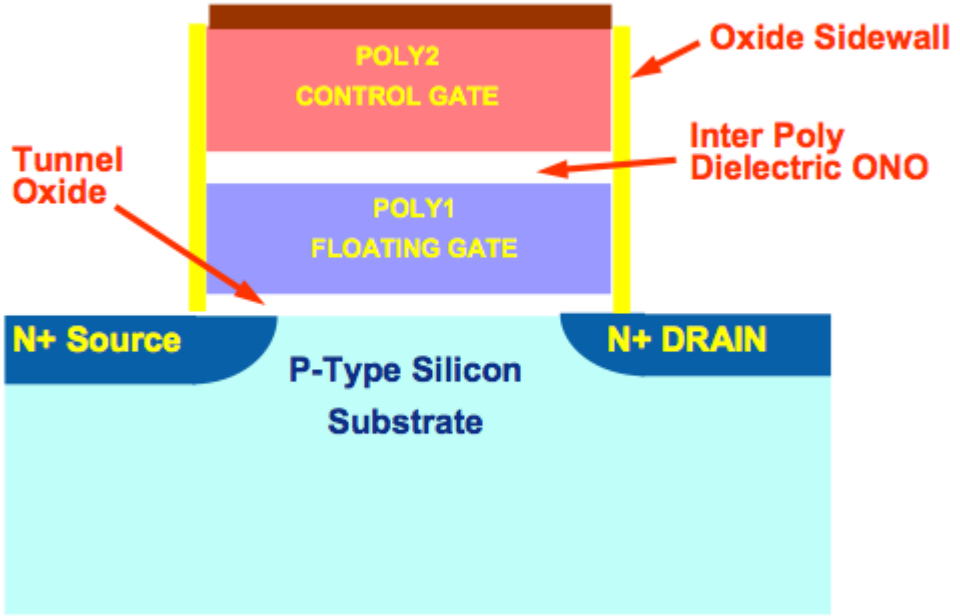





VS.



# nMOS Floating Gate Transistor








# NAND Flash Cell Array

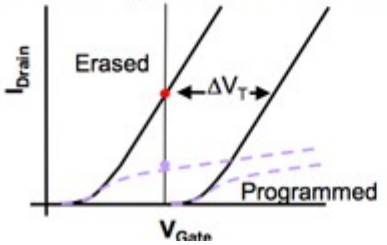
## Array Architecture & Read Operation

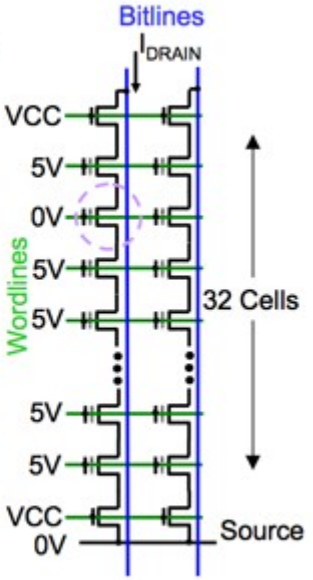
Array Architecture:

- Cells are in rectangular arrays of wordlines (gates), bitlines (drains), and common source
- Product = many "blocks" of ~1M cells each
- In NAND block, cells on a bitline are in *strings* of typ. 32 cells in series, as with NMOS transistors in a NAND gate
- NAND has NMOS select devices at source and drain, to isolate one block from another

Read:

- Cells are read by applying a gate voltage and testing the drain (bitline) current
- NAND: 31 cells are on, so the drain current is zero if the selected cell is OFF and is determined by the series resistance of the string if the selected cell is ON

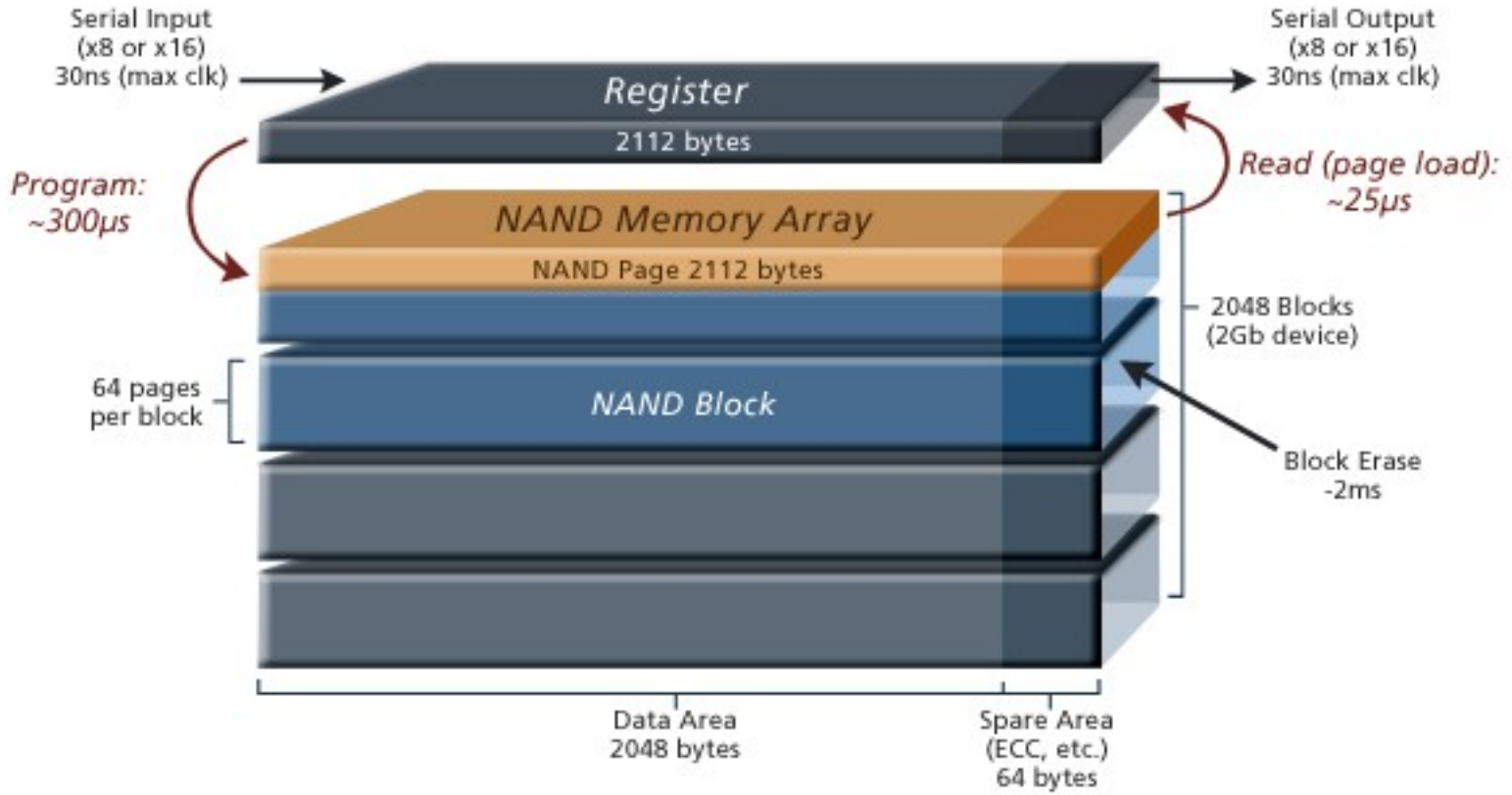






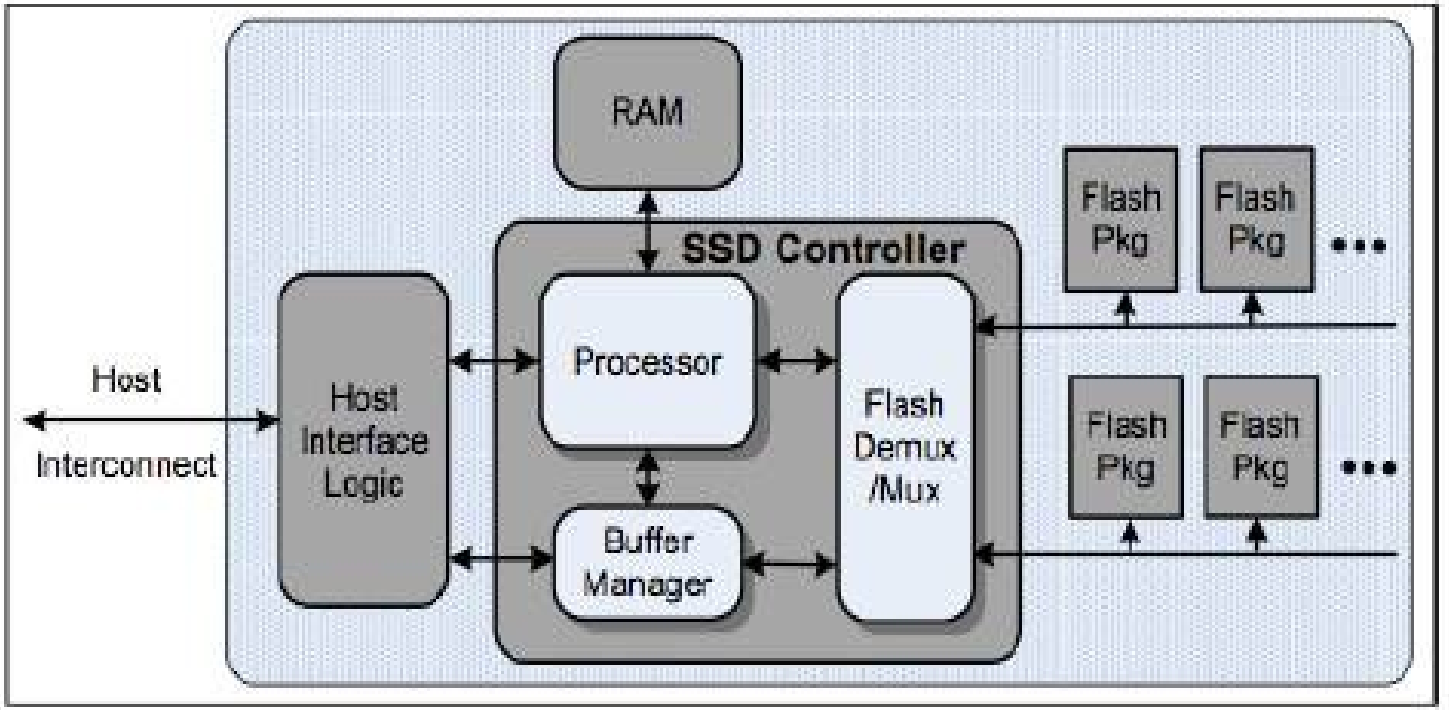

# Micron Flash Memory Plane








# Generalized SSD Block Diagram



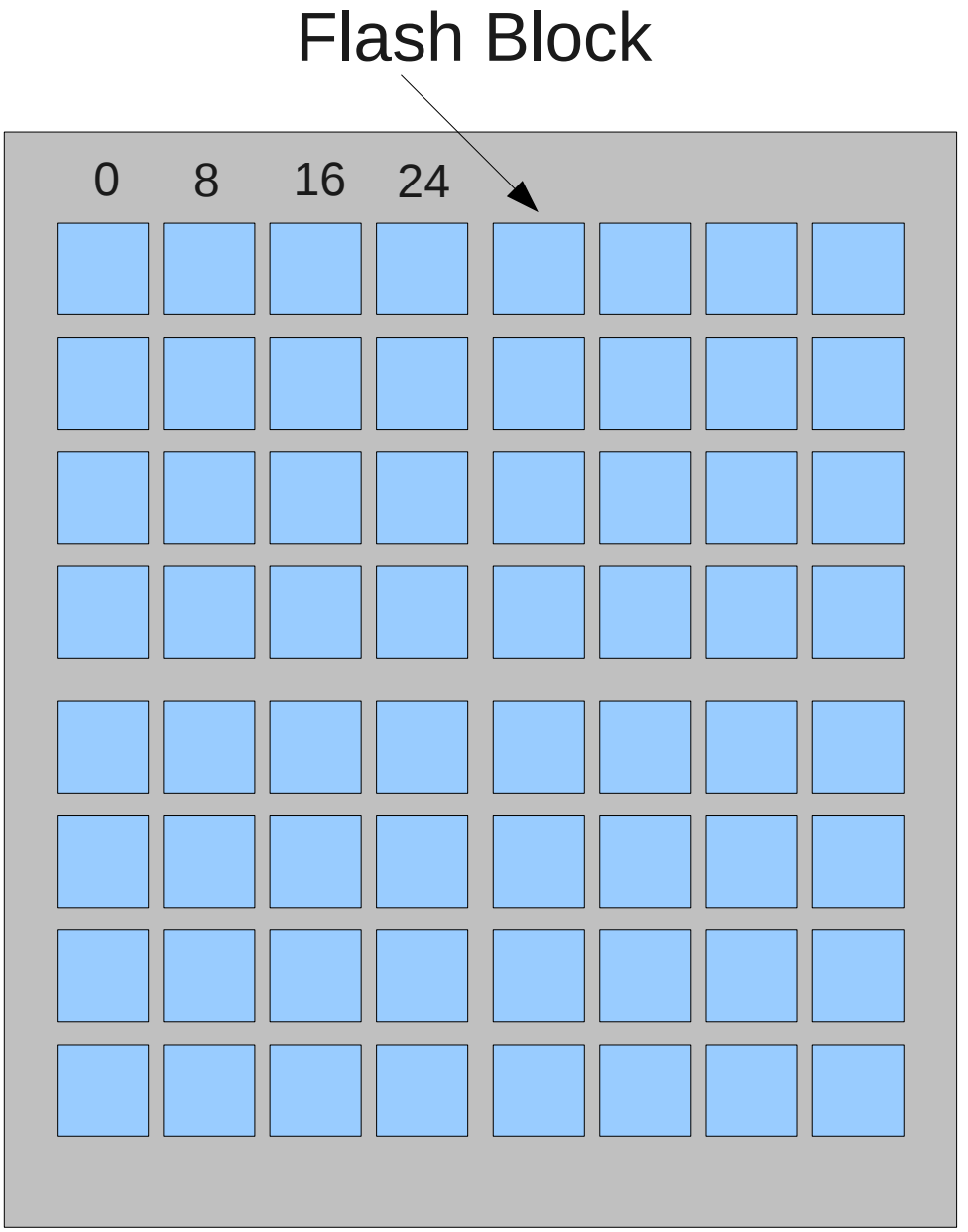

SSD Logic Components

Agrawal, Nitin, et. al. “Design Tradeoffs for SSD Performance”
Proceedings of the 2008 USENIX Technical Conference, June 2008





```
write(fd, buf, 4096);
write(fd, buf, 8192);
lseek(fd, 0, SEEK_SET);
write(fd, buf, 4096);
```







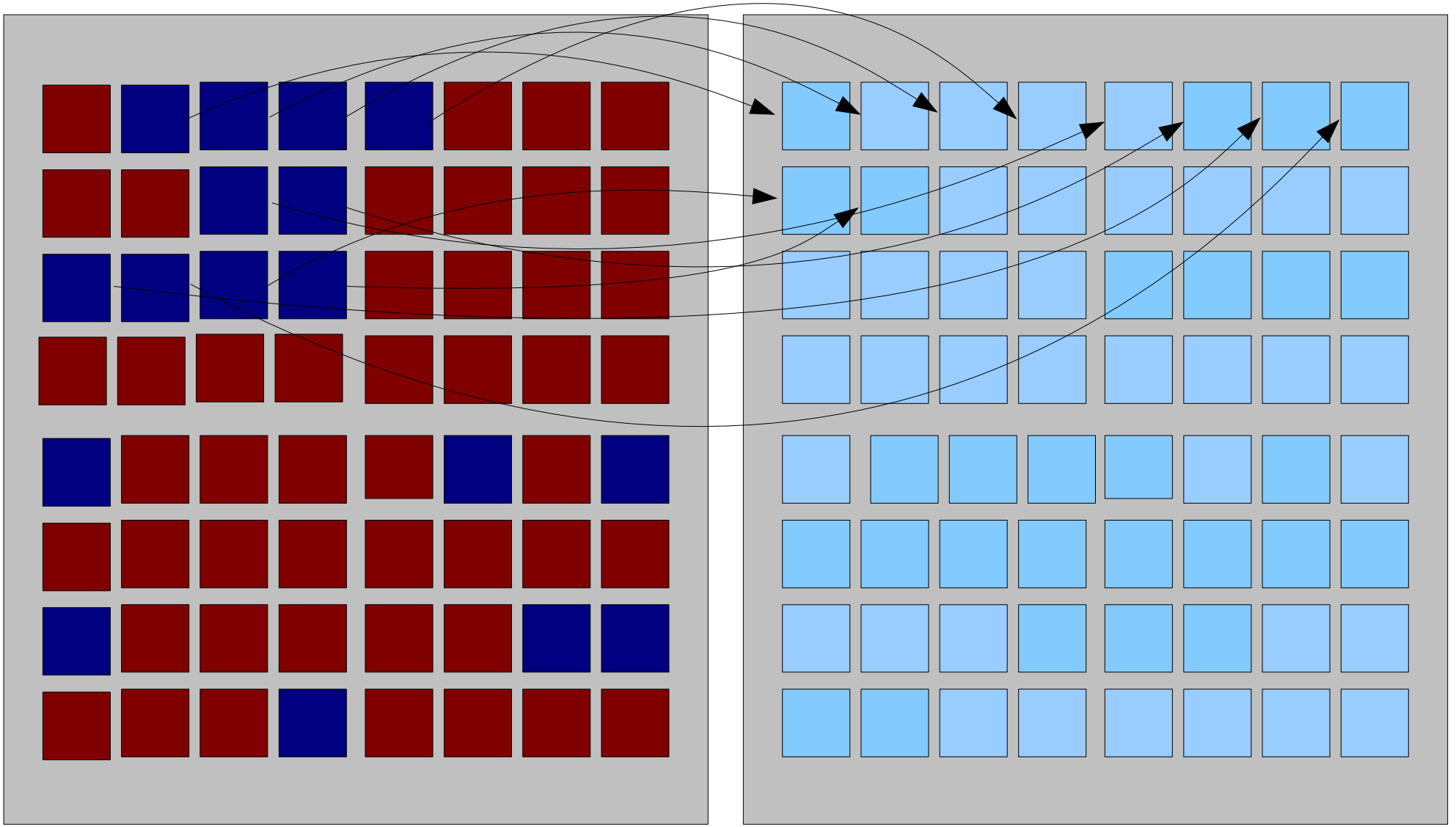

# Flash Translation Layer (FTL)

- LBA -> Physical Block Address
- Writes proceed sequentially within a block
- Re-writes are remapped (as space permits)
- **Requires garbage collection**





# Flash Block







# Write Amplification

- Formally, write amplification, due to garbage collection, is the average number of actual page writes per user page write. [Hu-SYSTOR-09]
- Always >1
- Intel advertises 1.1! (for certain workloads)
- Can be as bad a 3.5 or 4
- Upper bound on Program/Erase cycles (SLC: $10^5$, MLC $10^4$)
- Flash storage typically over-provisioned



# Garbage Collection

- Dynamic (most common)
- Static
- Background operations do affect performance




# TRIM

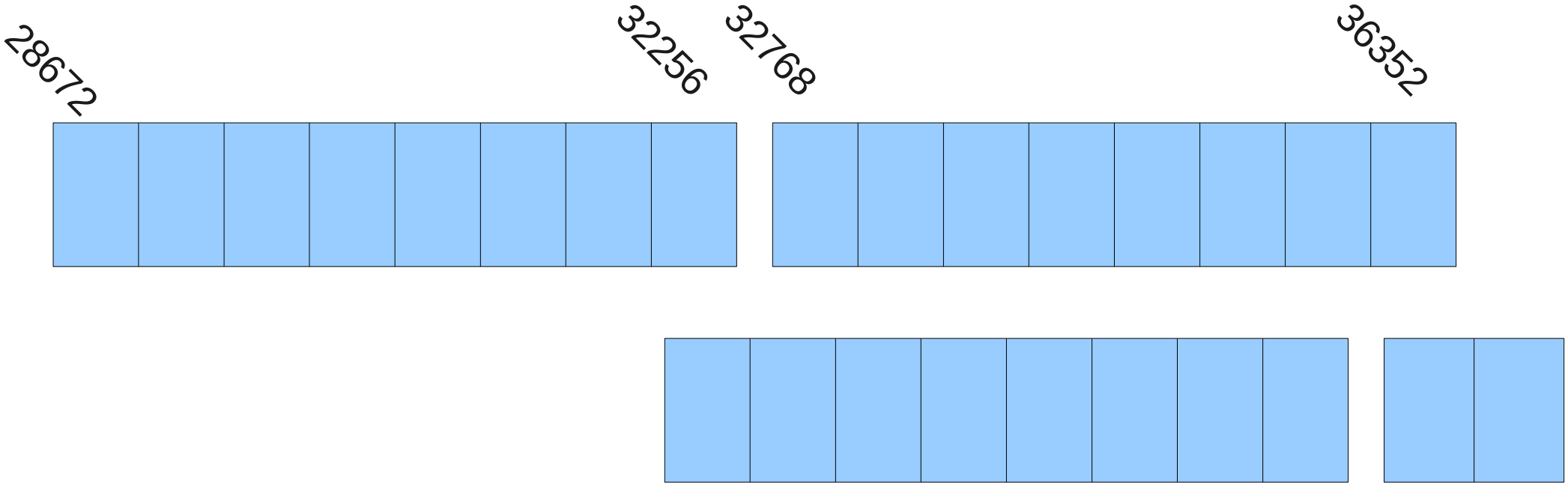
- Allows the file system to inform the disk about free blocks.
  - Unlink, truncate
- Not supported by all devices
- Some implementations are not standards compliant

- Why is this so important?

  Write performance can drop anywhere from 50-75% on a full disk!



# Alignment



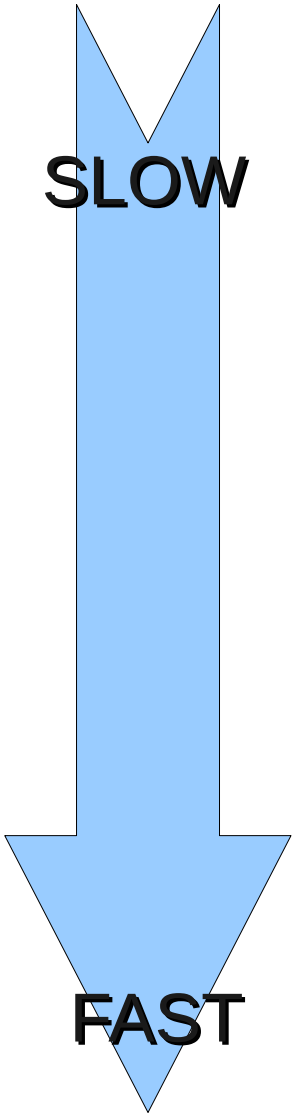

4KB = 8 512 byte blocks
Historically, partition 1 starts on sector 63.
63 * 512 = 32256

# Review

- Reads and Writes are done in units of 4KB
- Erases are done in flash block sizes (128KB-512KB)
- Re-writes require block remapping
- Garbage collection required to scrub mostly invalid blocks
- Flash requires wear leveling, as each cell is only capable of 10^5 or 10^6 program/erase cycles




# Classes of SSDs
# where the rubber meets the road




SATA/PATA Generation 0

Netbooks, etc

SATA Generation 1

jmicron

SATA Generation 2

Indilinx controller, intel

PCIe

Fusion I/O, TMS RAMSAN





**Next up**: The things Microsoft is doing to help us all out!

(no, seriously!)



# Windows 7 Storage Logo Proposal (1-3)

- Proposed Windows 7 logo requirements related to SSD
  - Storage devices complying with ATA8-ACS specification shall report their rotation speeds according to ATA8-ACS Identify Word 217: Nominal Media Rotation Rate
  - The performance of the storage device shall not degrade with any amount of data stored to the maximum capacity of the device

# Windows 7 Storage Logo Proposal (2-3)

- If “Trim” algorithm is applied, the “Trim” implementation must comply with ATA8-ACS2 proposal e07154r6 (Data Set Management Commands Proposal for ATA8-ACS2) section 5.3 and section 6.2. The completion time of Trim command should be less or equal to 20ms

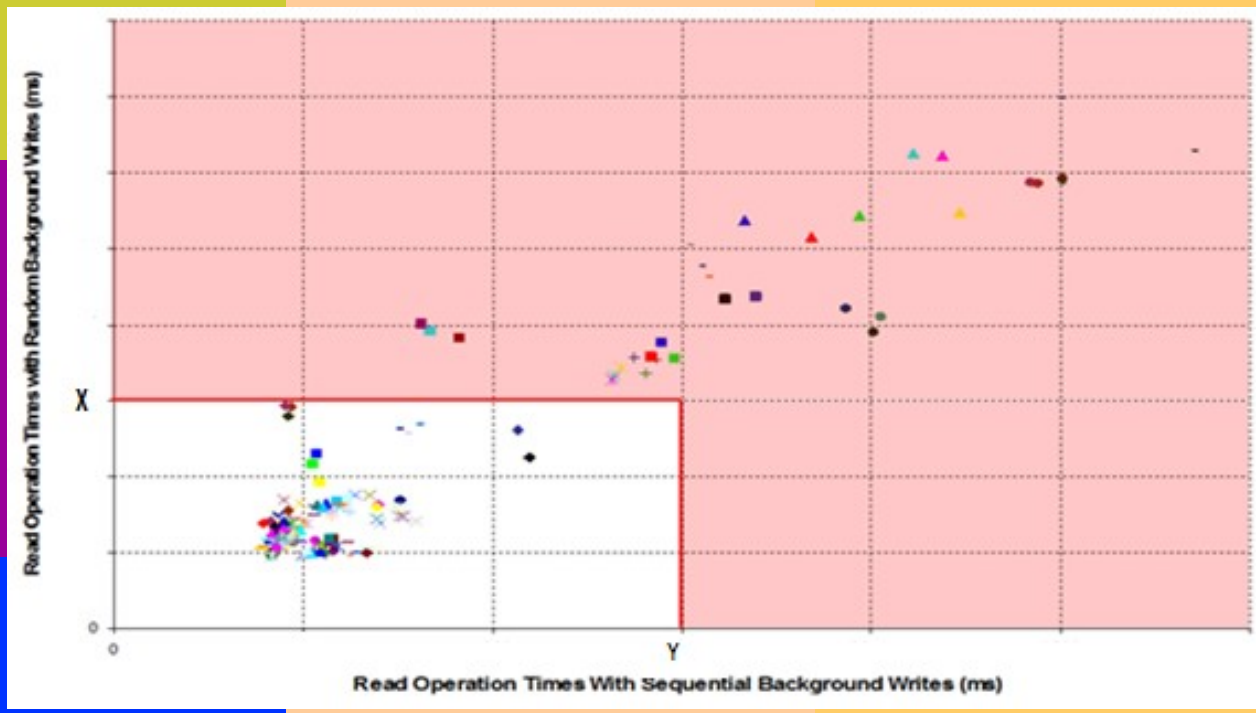
- SATA-IO certification is required for Solid State Drive (SSD) connected through SATA interface. More information on SATA-IO testing will be available on the SATA-IO Web site at: http://www.sata-io.org/testing.asp

# Windows 7 Storage Logo Proposal (3-3)

The read response time of storage device shall be less than or equal to the maximum response time required.

Giving read a priority can be important when there is a long queue of writes

The result is better user experience of system responsiveness



# Operating System Support

- Need to TRIM free blocks
  - Mkfs, unlink/truncate
- Need to align partitions properly
  - fdisk,parted,etc
  - Lvm tools
- Need to drive deep queue depths to exploit parallelism




# Linux Support

- Block Layer
  - Discard
  - Rotational flag
- File Systems
  - Ext4, fat, btrfs
    - Mkfs trims blocks
- Parted/fdisk align partitions based on exported toplogy
- Utilities such as hdparm support discard operations
  - wiper.sh




# Windows 7

- Disable defrag
- Align partitions
- Send trim where available for:
  - Format, delete, truncate, compression
  - o/s internal processes: snapshot, volume manager




# Are we done?

- Few devices support TRIM
- Ext4 TRIM usage not optimal
- TRIM support in the block layer not fully fleshed out
- TRIM is disabled by deafult
- No support in LVM (yet)
- Software RAID implementations need tweaking




# Deployment Recommendations

- File Systems:
  - To journal, or not to journal?
  - relatime (default for most distros by  now)
  - Discard support
- LVM is OK, so long as you don't plan on issue TRIM
- Align partitions to erase block boundary
- Deadline I/O scheduler
- RAID-0 OK
- Don't write to your disk!



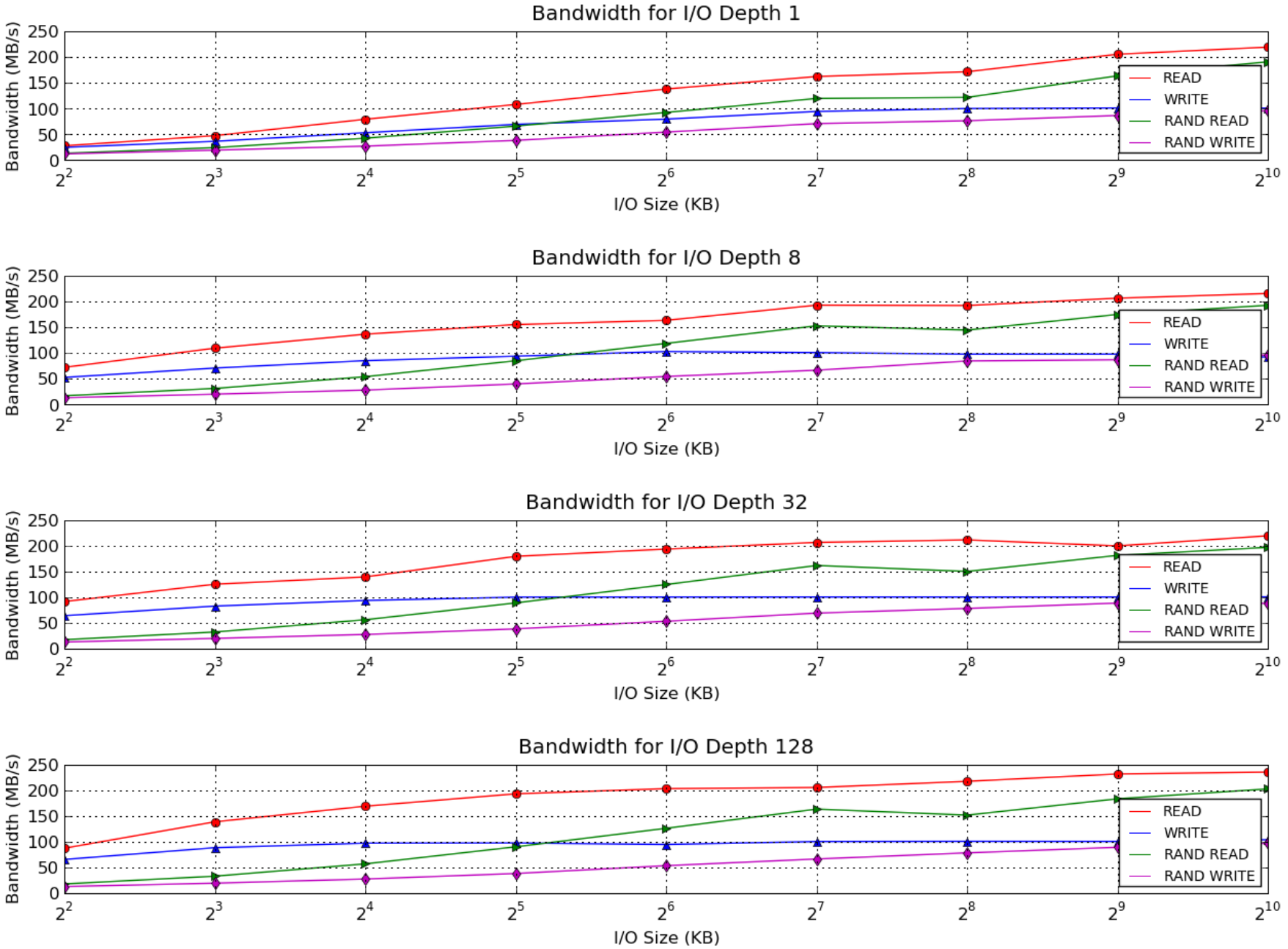

# SSD used for DB logs

# Frequently Asked Questions

- Is the MTBF for SSDs longer than that of spinning media?
- Can/Should I put swap on my SSD?
- Is an SSD worth the money?
- Can I use my SSD in a RAID set?




aio-stress output for: 1 thread, 1 files



# Further Reading


- [Hu-SYSTOR-09] Hu, Xiao-Yu, et. al., “Write Amplification Analysis in Flash-Based Solid State Drives.” Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference. 2009, Aritcle No. 10
- [Anand-Anthology] http://www.anandtech.com/show/2738/1
- http://www.linux-mag.com/cache/7590/1.html
- http://thunk.org/tytso/blog/2009/03/01/ssds-journaling-and-noatimerelatime/
- http://www.eeherald.com/section/design-guide/esmod16.html
- Chen, Feng, "Understanding Intrinsic Characteristics and System Implications of Flash Memory based Solid State Drives." Proceedings of the eleventh international joint conference on Measurement and modeling of computer systems. 2009, pp 181-192.
- Agrawal, Nitin, et. al., “Design Tradeoffs for SSD Performance.” Proceedings of the 2008 USENIX Technical Conference. June, 2008




# Further Reading (continued)

- Desnoyers, Peter, “Empirical Evaluation of NAND Flash Memory Performance.” ACM SIGOPS Operating Systems Review. January, 2010, pp 50-54.
- Narayanan, Dushyanth, “Migrating server storage to SSDs: analysis of tradeoffs” Proceedings of the 4th ACM European conference on Computer systems. 2009, pp 145-158
- http://download.microsoft.com/download/5/E/6/5E66B27B-988B-4F50-AF3A-C2F
- http://en.wikipedia.org/wiki/Flash_memory