SNIA ESF | ETHERNET STORAGE

Storage Performance Benchmarking: Part 3 – Block Components

Ken Cantrell, NetApp

Mark Rogov, EMC

David Fair, SNIA ESF Chair, Intel

March 8, 2016

SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

  NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

About The Speakers

SNIA ESF | ETHERNET STORAGE

Ken Cantrell

NetApp

Manager Perf Engineering

@kencantrelljr

Mark Rogov

EMC

Advisory Systems Engineer

@rogovmark

Dr. David Fair

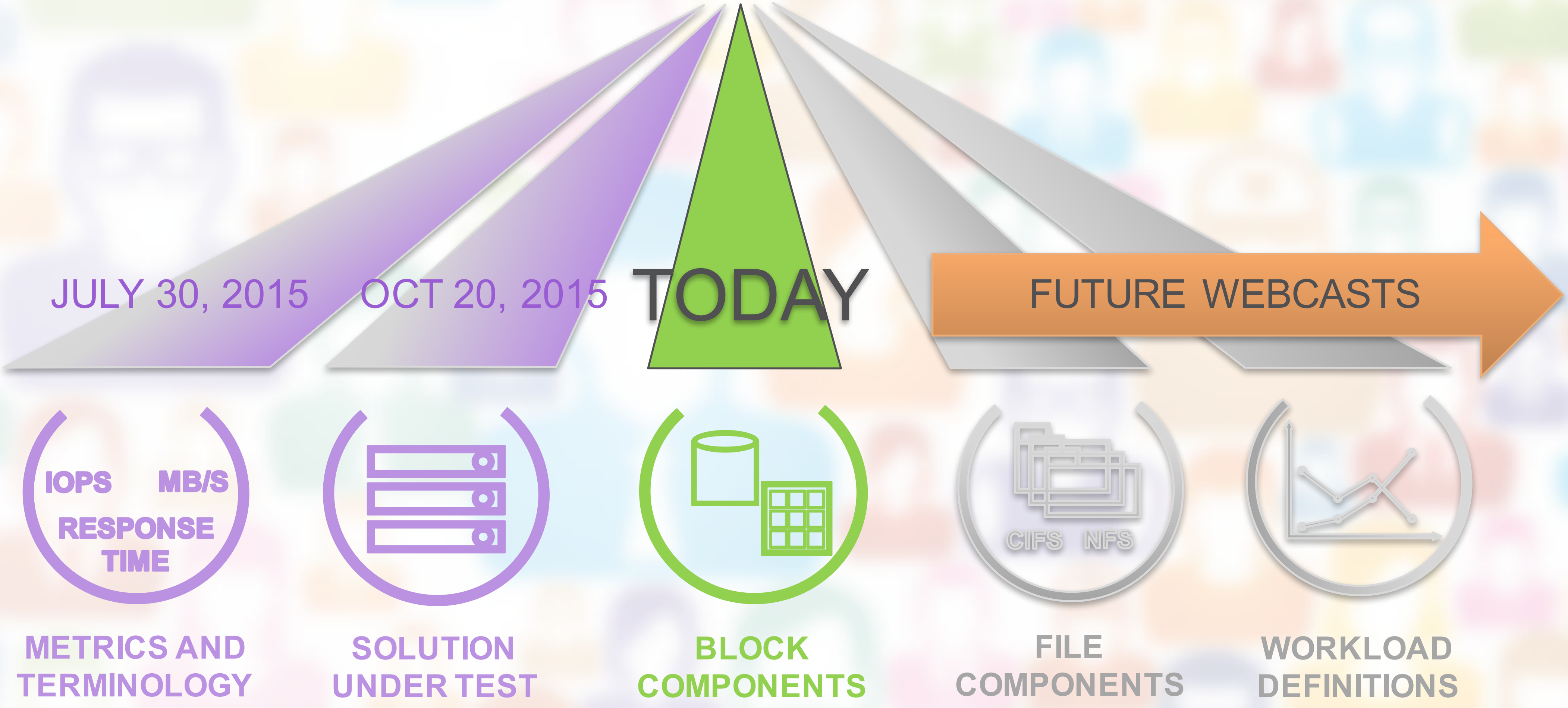
SNIA ESF Chair

& Intel

Ethernet Networking Marketing Manager

Storage Performance Benchmarking

SNIA ESF | ETHERNET STORAGE



JULY 30, 2015

OCT 20, 2015

TODAY

FUTURE WEBCASTS

METRICS AND TERMINOLOGY

SOLUTION UNDER TEST

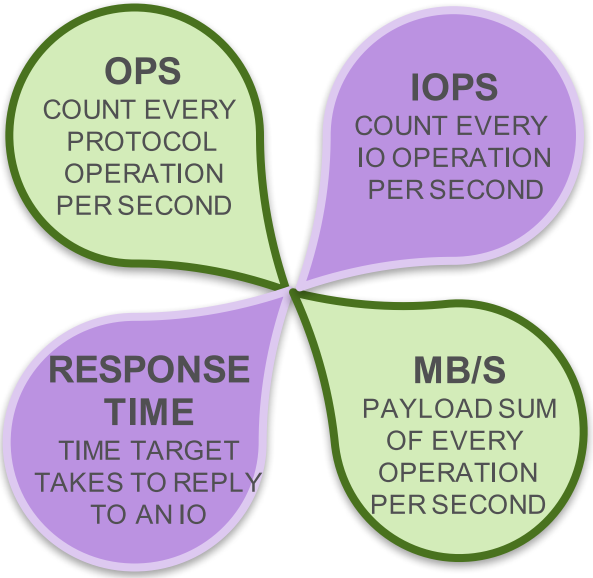
BLOCK COMPONENTS

FILE COMPONENTS

WORKLOAD DEFINITIONS

INTRO > R/W > TECH > RAID > FUN > END

Session 1 – Terminology and Context



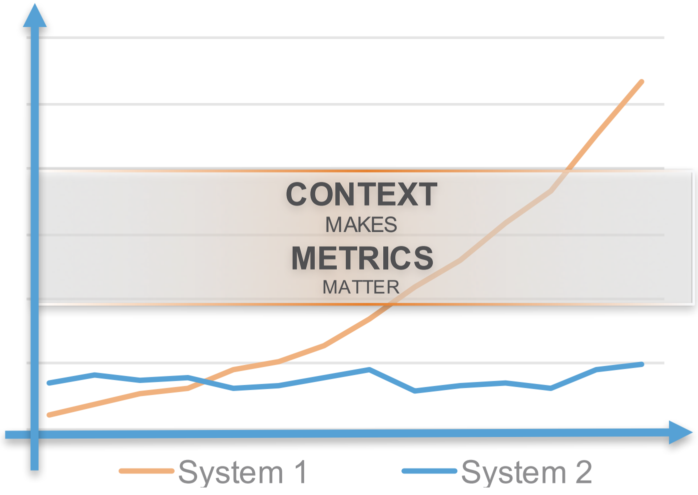
OPS
COUNT EVERY PROTOCOL OPERATION PER SECOND

IOPS
COUNT EVERY IO OPERATION PER SECOND

RESPONSE TIME
TIME TARGET TAKES TO REPLY TO AN IO

MB/S
PAYLOAD SUM OF EVERY OPERATION PER SECOND

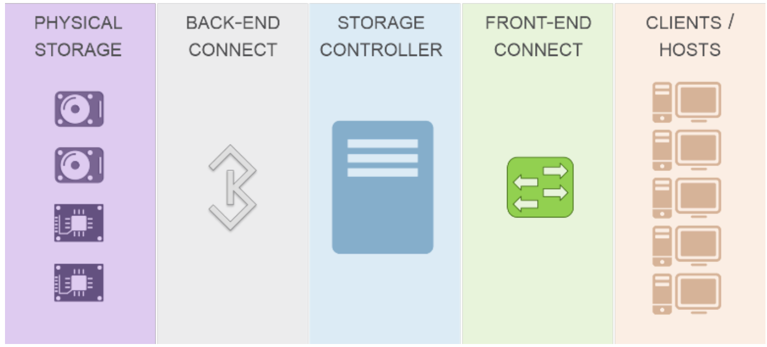
TERMINOLOGY



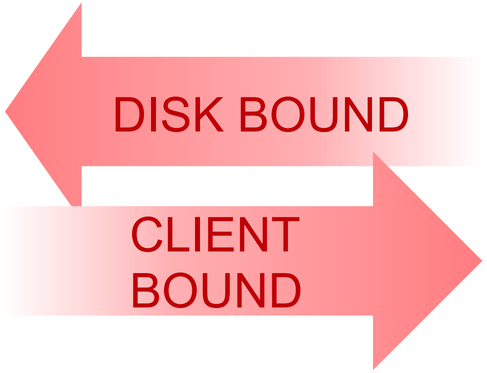
CONTEXT
MAKES
METRICS
MATTER

GRAPH FUN

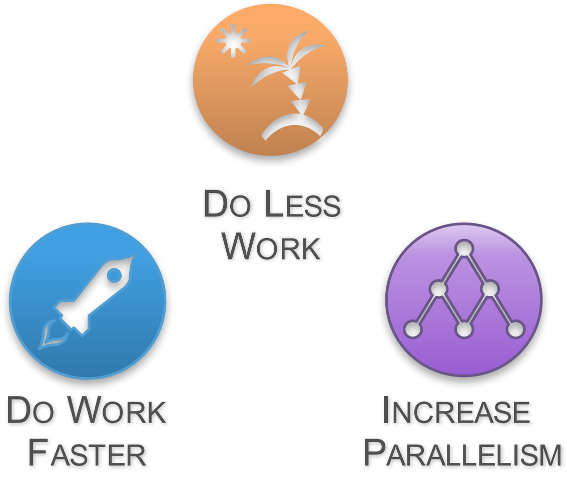
Session 2 – The Slowest Component Matters Most



SLOW COMPONENT MATTERS MOST



BOTTLENECKS ALWAYS EXIST



3 PERFORMANCE PRINCIPLES

Enterprise Storage Capacity Shipped In 3Q'15

33.1 EXABYTES

WORLD POPULATION

IOPS
MB/S
RESPONSE TIME

7.4 BILLION

SPLIT EQUALLY

OR

METRICS AND TERMINOLOGY

4.5

GIGABYTES PER INDIVIDUAL

1541

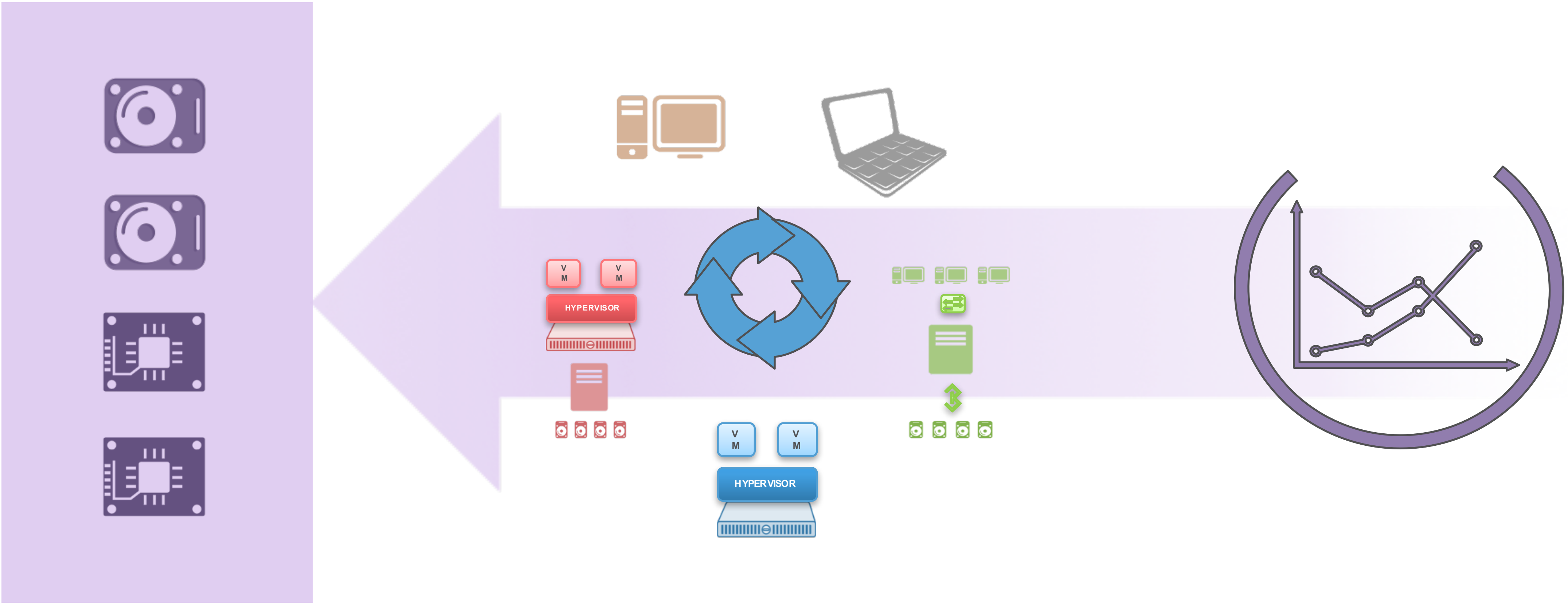
COPIES OF OUR FIRST WEBCAST POWERPOINT

Eventually, All Data Goes To Block Storage

BLOCK STORAGE

SOLUTIONS UNDER TEST

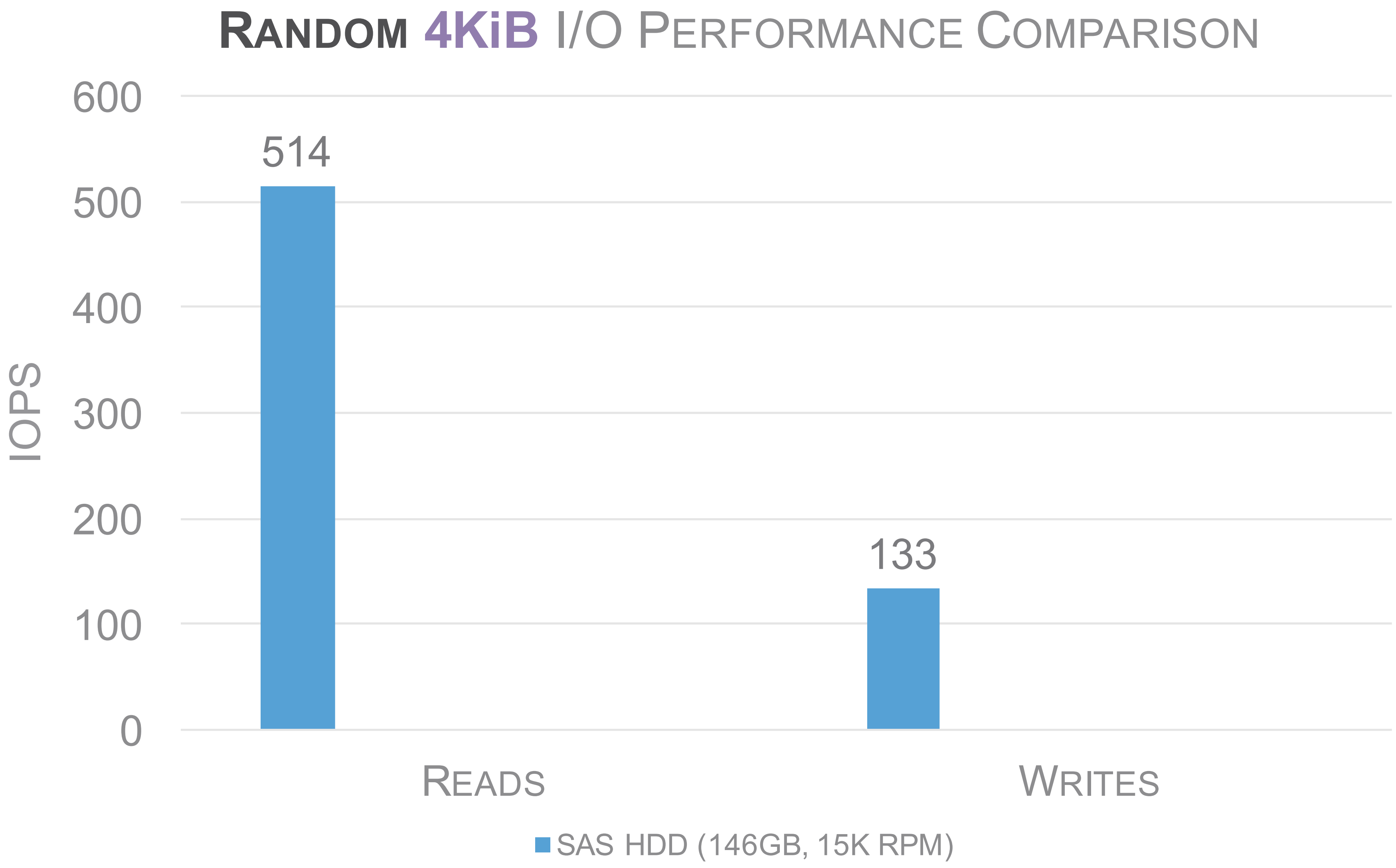
WORKLOADS



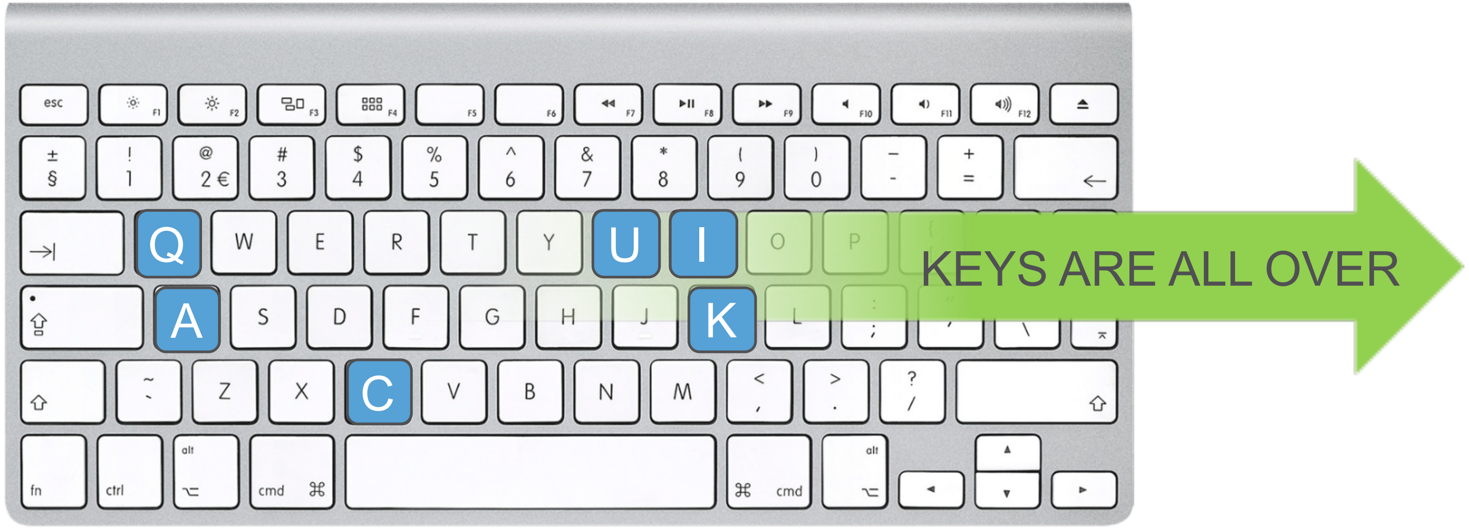
Agenda

INTRO	Introduction
R/W	Reading, Writing; What is the Difference?
TECH	How does this tech work anyway?
RAID	What if you need more than one?
FUN	Performance?
END	Summary

Let’s Take A Drive… And Test It!



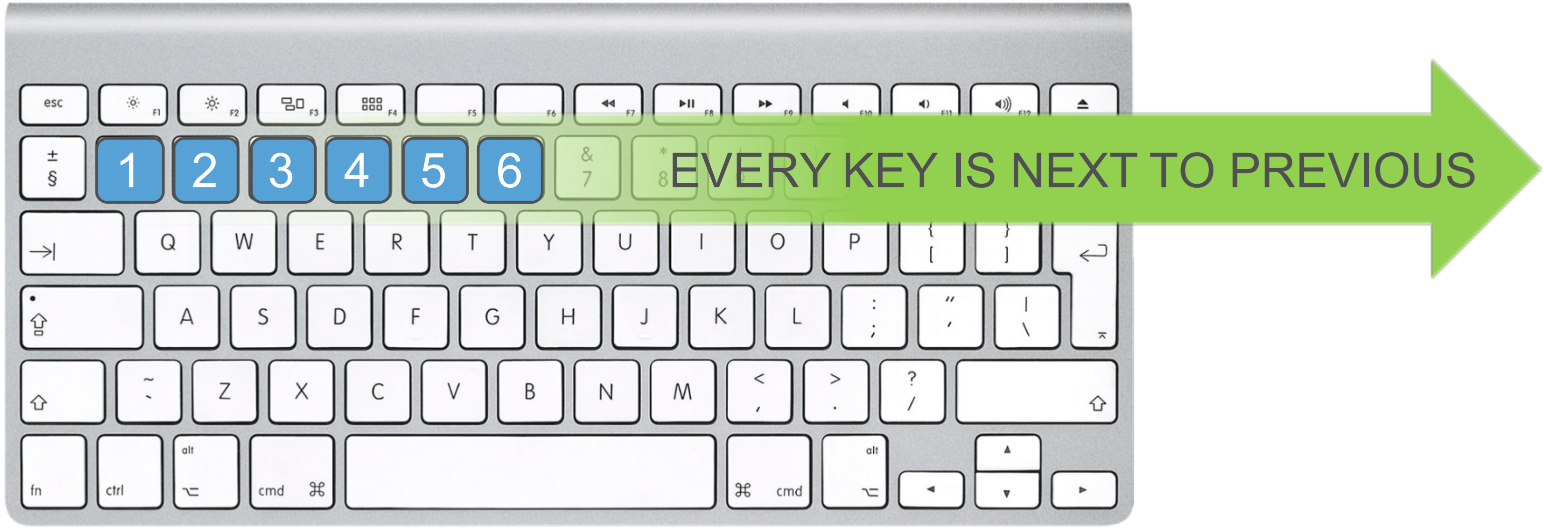
Detour! What Does "Random" Mean?



A QUICK BROWN FOX JUMPED OVER A LAZY DOG

IMAGINE THAT THE KEYBOARD IS A DISK DRIVE

What Does "Sequential" Mean?



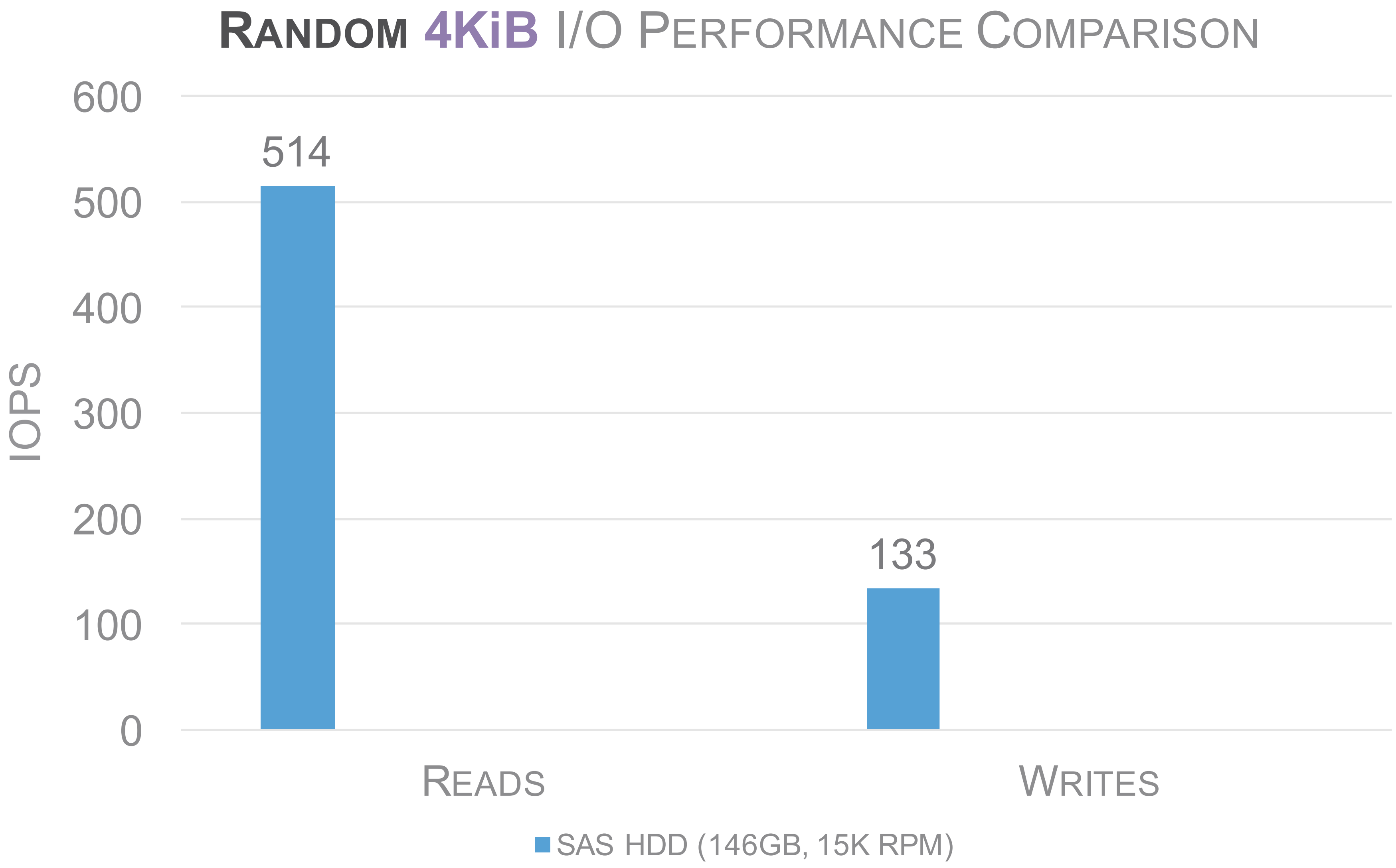
1 2 3 4 5 6

IMAGINE THAT THE KEYBOARD
IS A DISK DRIVE

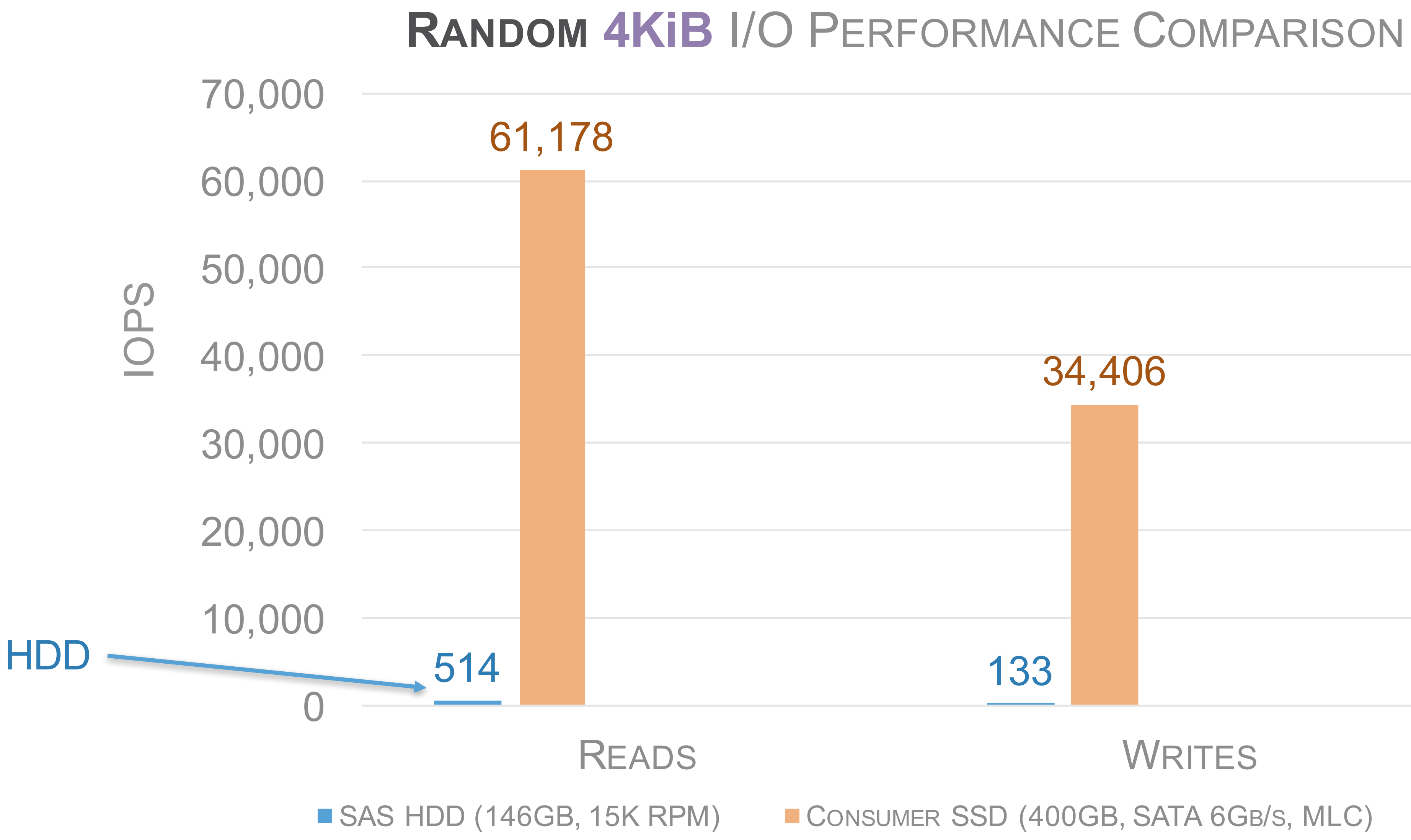
“Sequential Read” Example

“SEQUENTIAL READ”

Let's Take A Drive… And Test It!

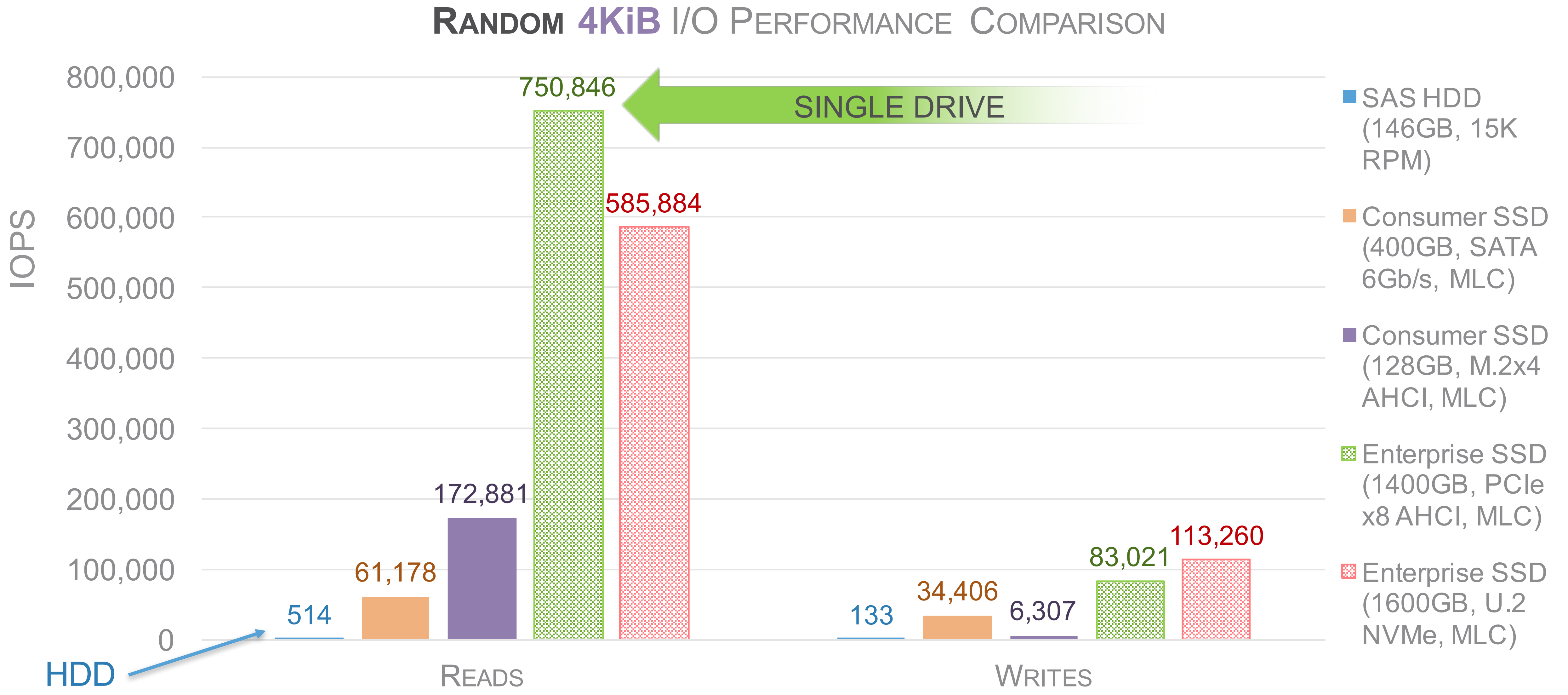


Let's Take Two Drives… And Test Them!



And Add More SSDs

SNIA ESF | ETHERNET STORAGE

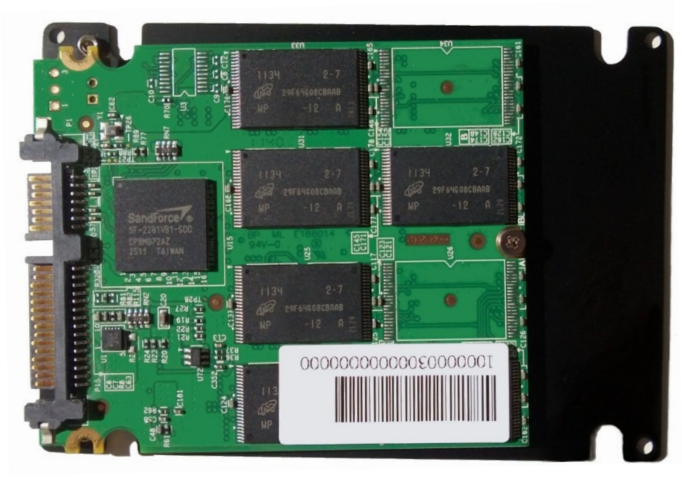


Agenda

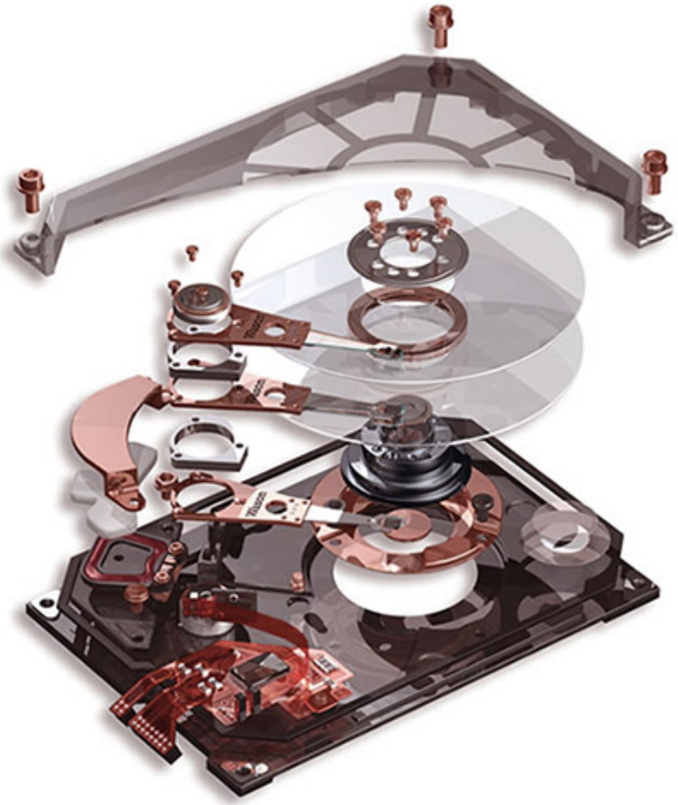
INTRO	Introduction
R/W	Reading, Writing; What is the Difference?
TECH	How does this tech work anyway?
RAID	What if you need more than one?
FUN	Performance?
END	Summary

How Does This Tech Work?

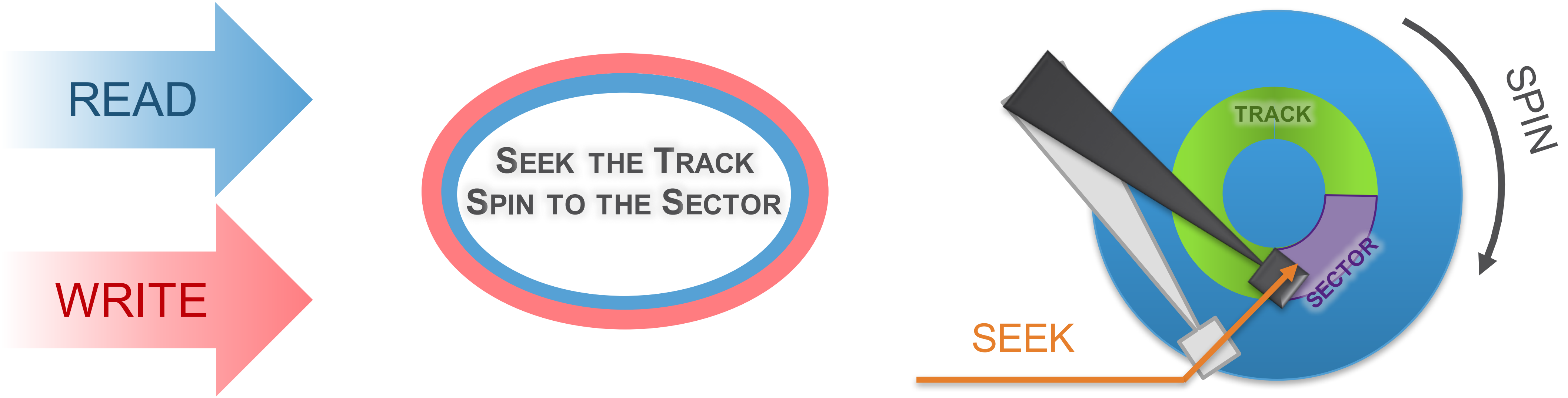
Flash



HDD or Disk Drive



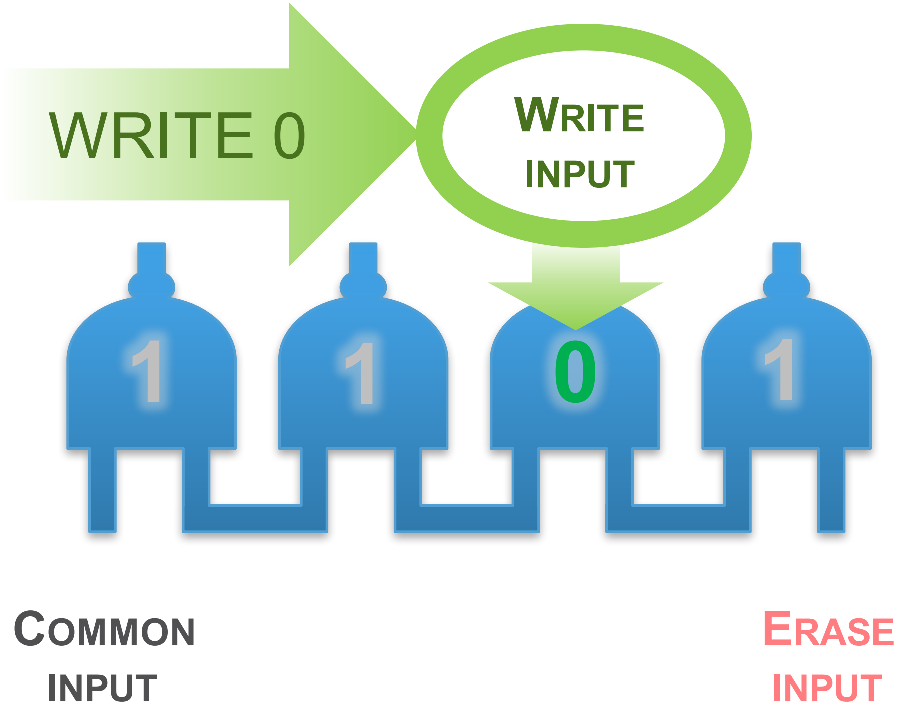
Spinning Drives And Sectors



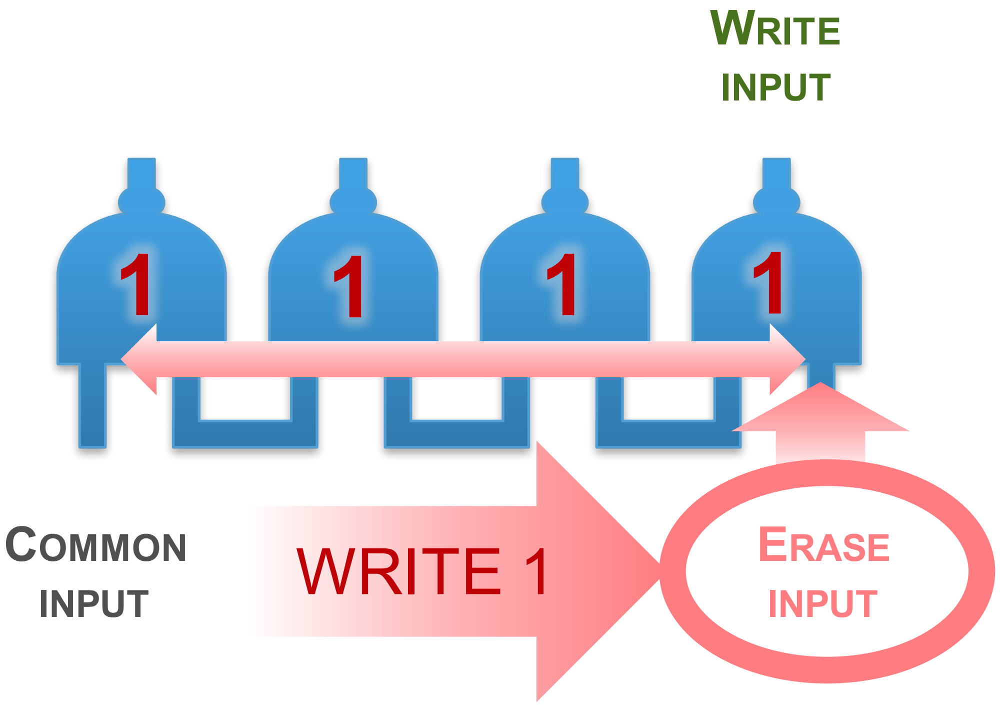
Seek the Track
Spin to the Sector

Flash And NAND Gates

Every NAND can be set to 0 individually

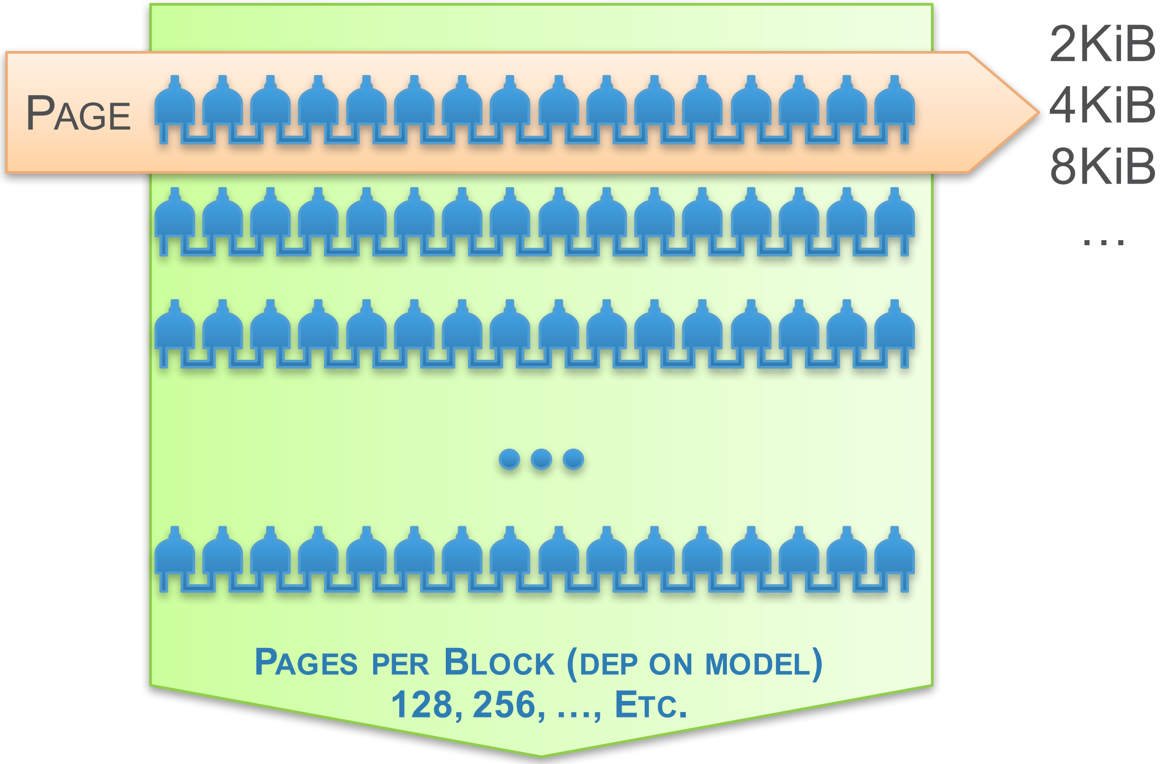


To set back to 1, an entire group needs to be reset

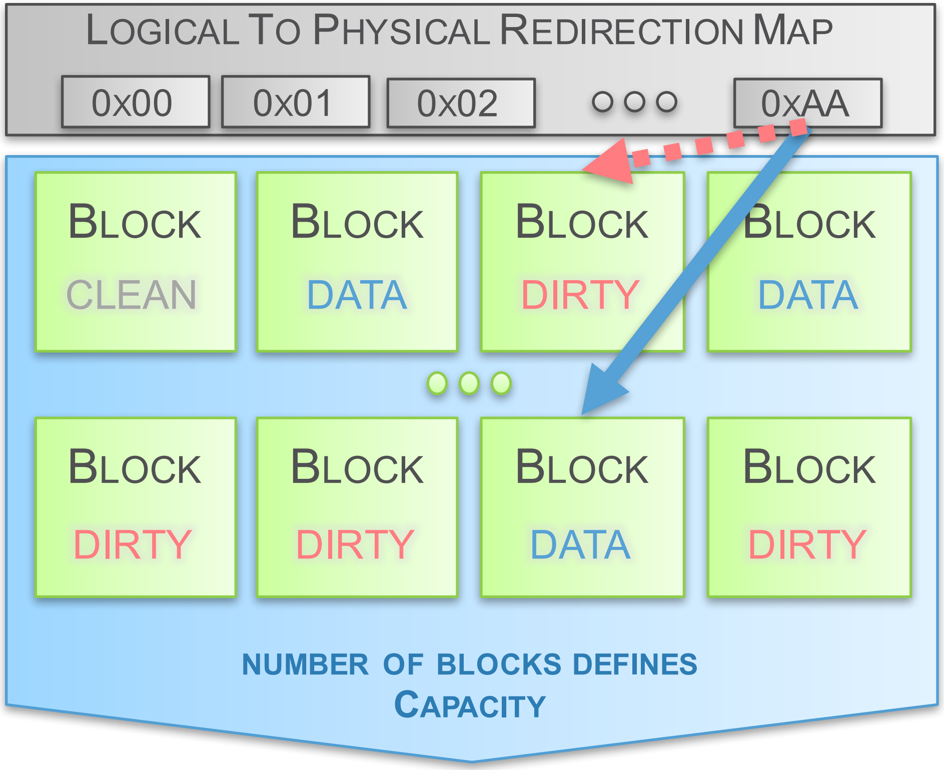


Flash Construction

Flash BLOCK



Flash Device



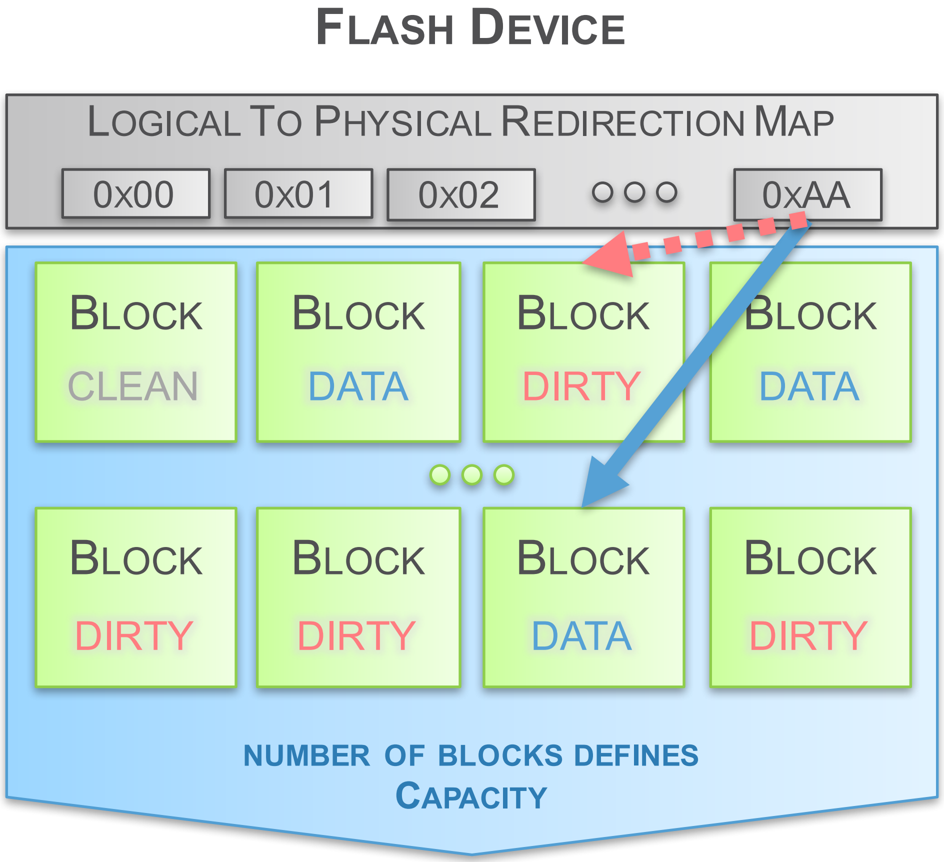
Most Flash

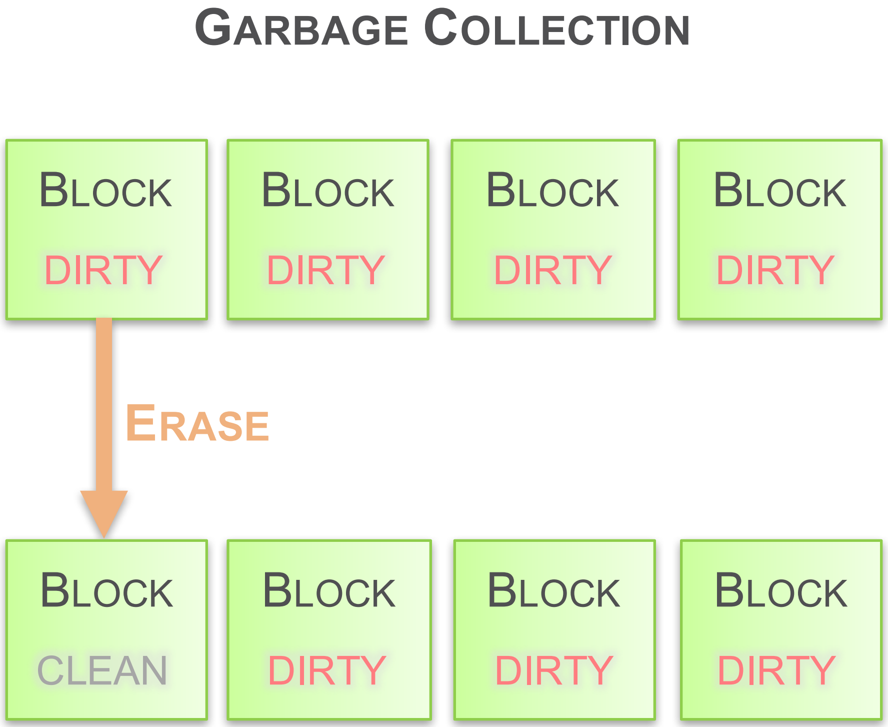
Write—1 page at a time

Redirect On Over-Write

An IO is redirected to a CLEAN block/page
Leaving old block/page DIRTY

Garbage Collection



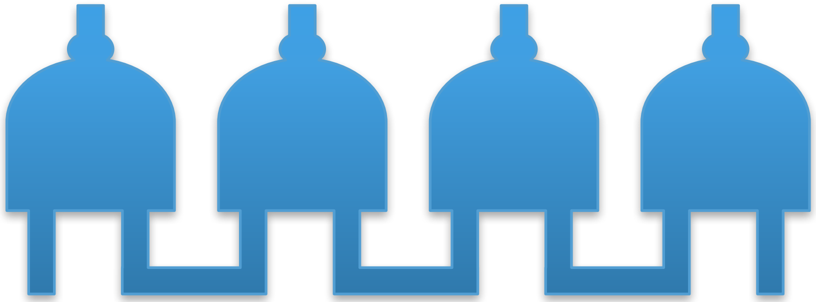


ERASE—1 DIRTY BLOCK AT A TIME

(WHEN NUMBER OF CLEAN BLOCKS IS LOW)

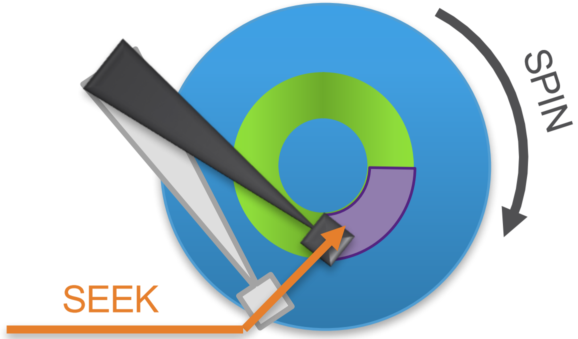
Sequential Vs. Random

SSD or Flash



	Write	Read
Everything is random IO for Flash	Erase + Write	Read

HDD or Disk Drive



	Write	Read
Sequential	Write	Read
Random	Seek/Spin + Write	Seek/Spin + Read

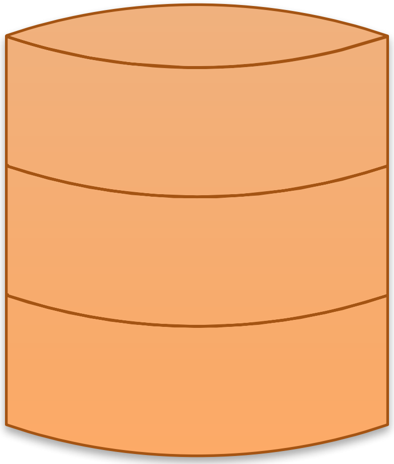
SLOWER PERFORMANCE

Agenda

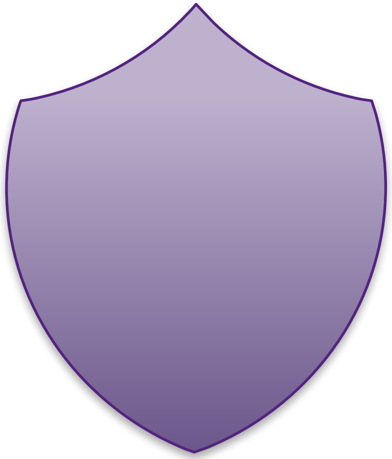
INTRO	Introduction
R/W	Reading, Writing; What is the Difference?
TECH	How does this tech work anyway?
RAID	What if you need more than one?
FUN	Performance?
END	Summary

Just One?

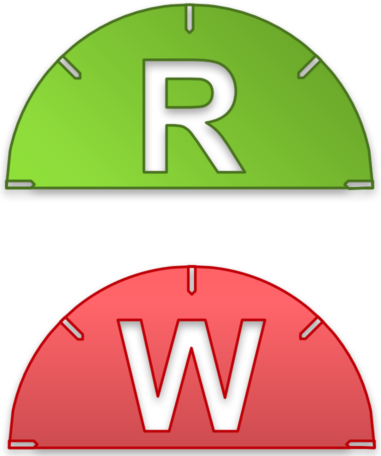
CAPACITY



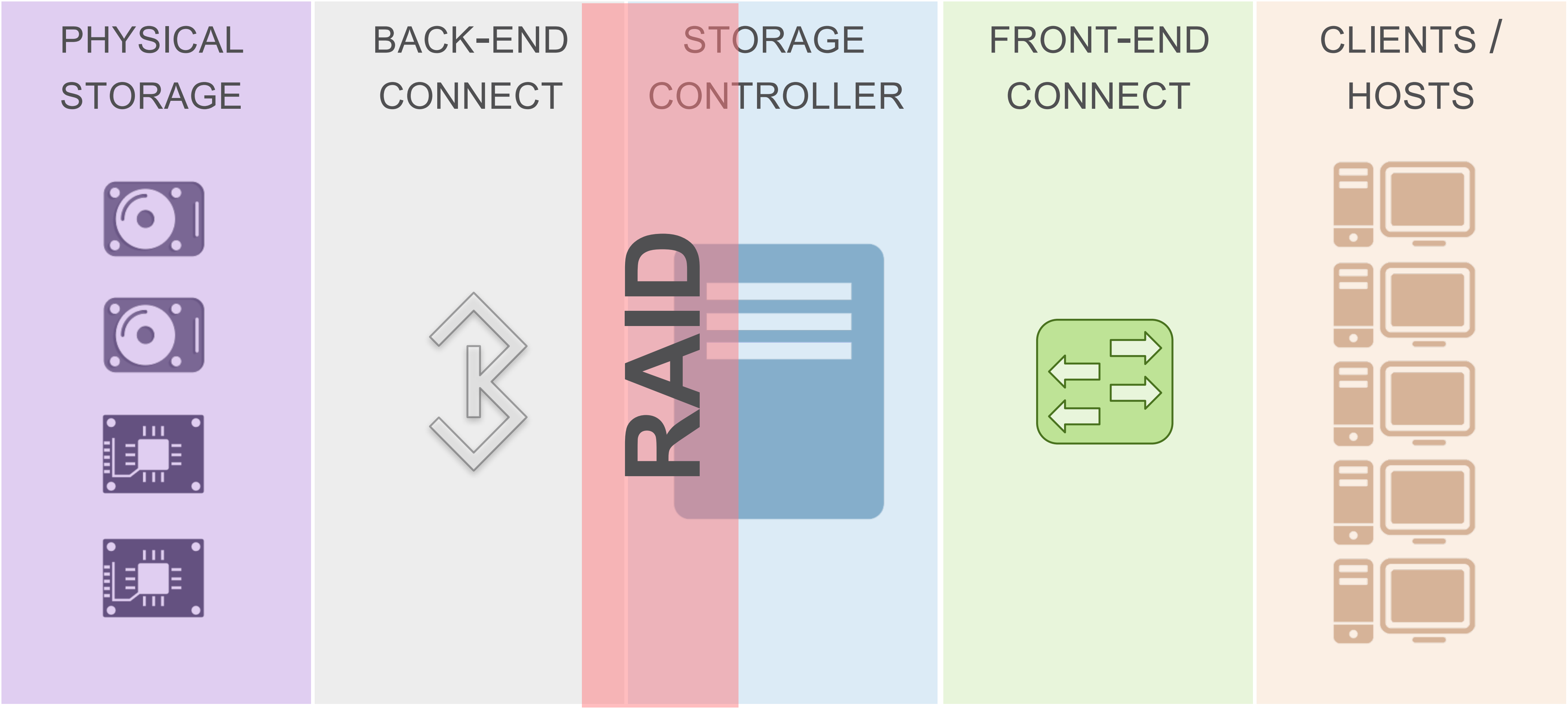
PROTECTION



PERFORMANCE



RAID—Redundant Array Of Inexpensive Disks

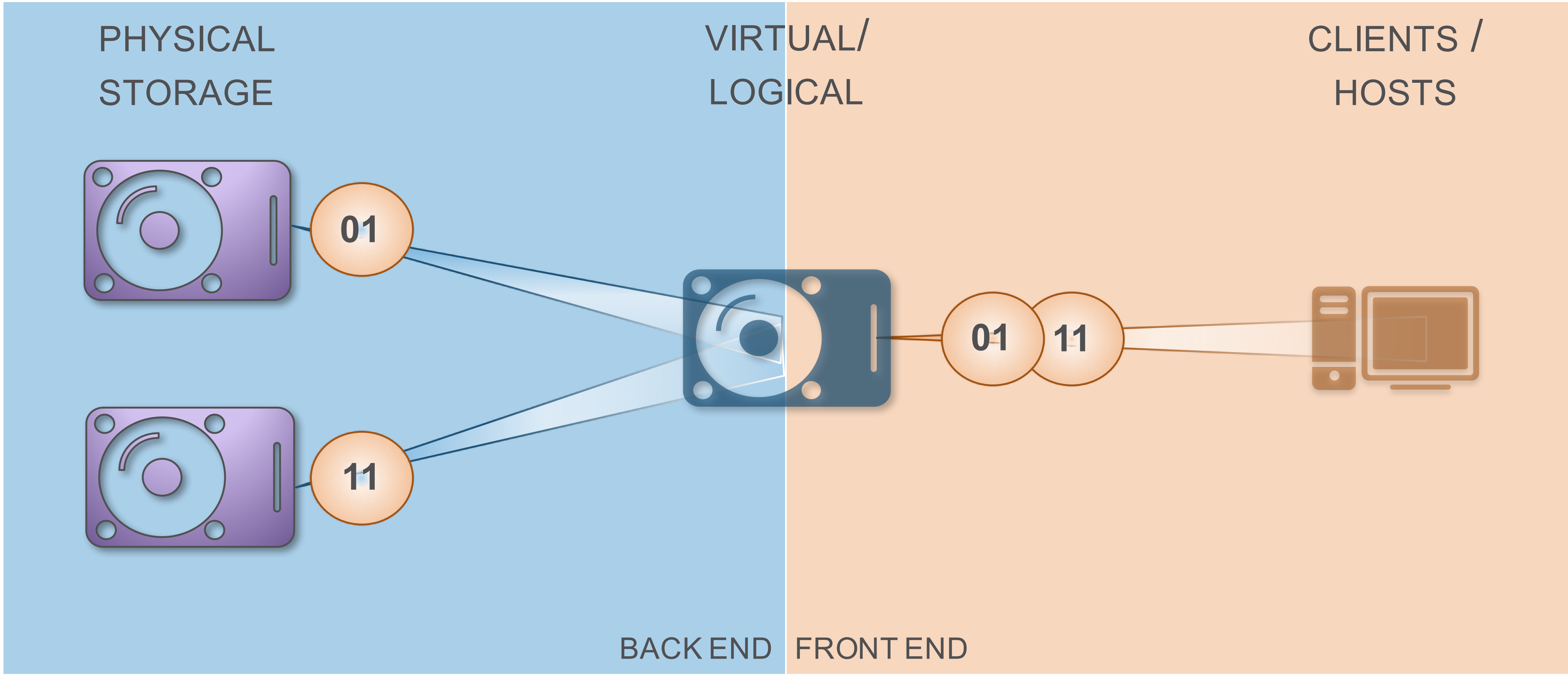


RAID-0 (Striping Without Parity)

CAPACITY 100%

PROTECTION NONE

PERFORMANCE R 100% W 100%

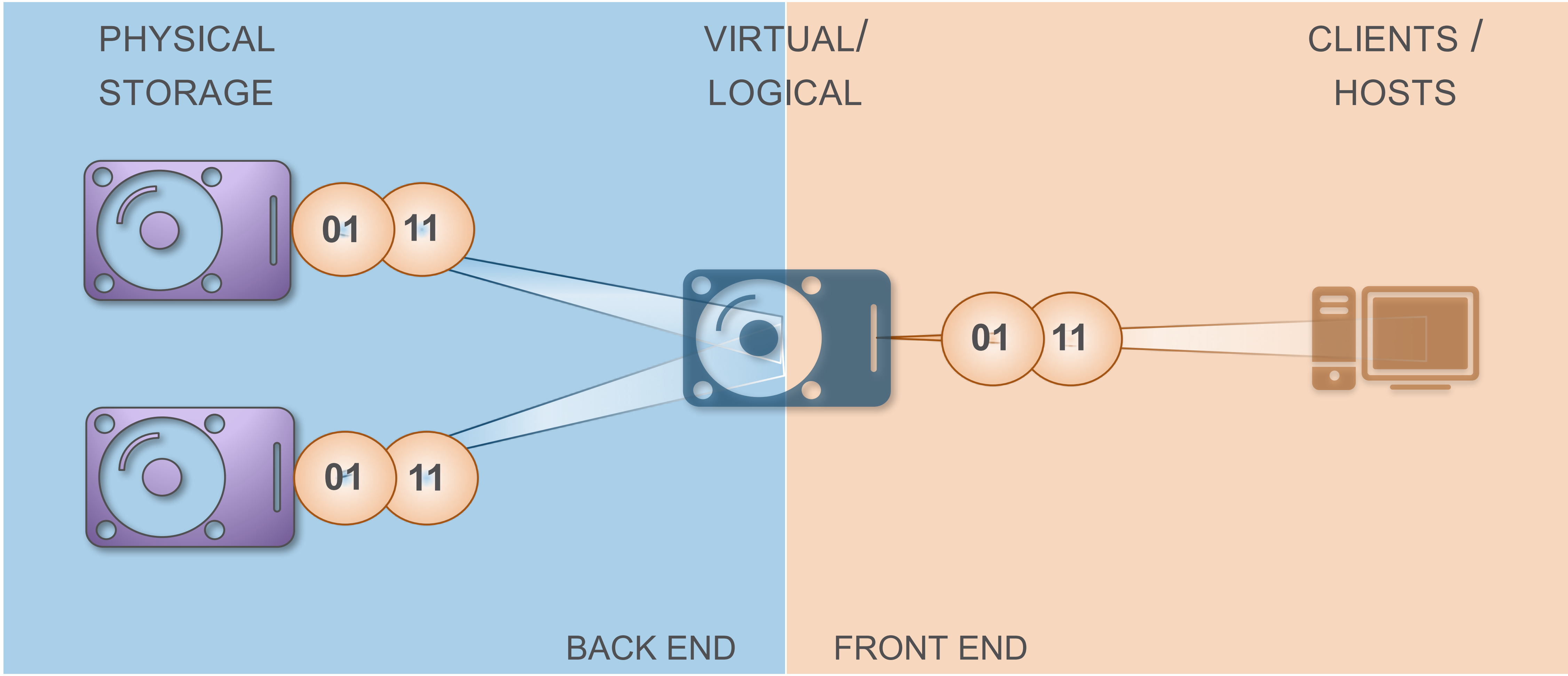


RAID-1 (Mirroring)

CAPACITY 50%

PROTECTION 1 DRIVE

PERFORMANCE R 100% W 50%



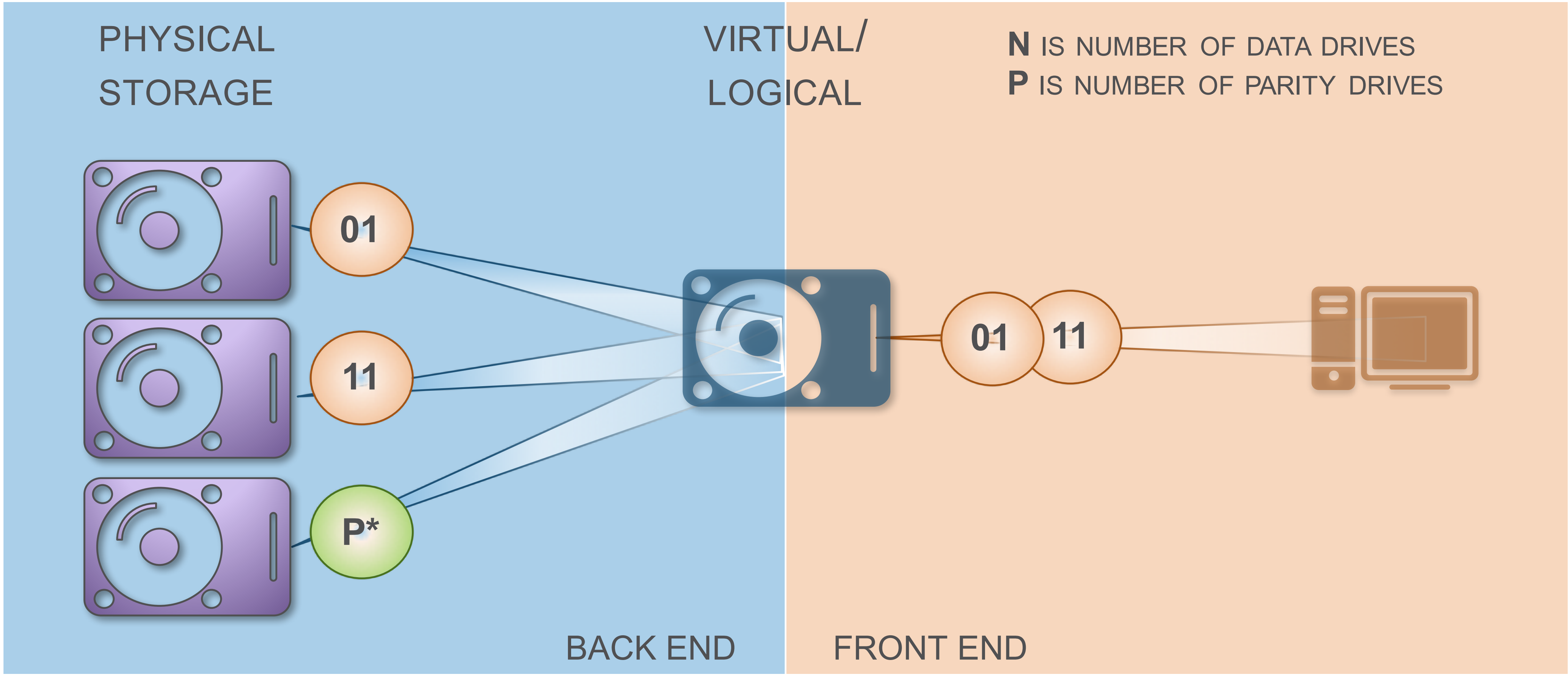
RAID-3, -4, -5 [-6, -DP]*

Striping With Parity*

CAPACITY N

PROTECTION P

PERFORMANCE R N W "IT DEPENDS"



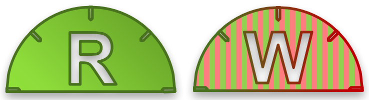
N IS NUMBER OF DATA DRIVES
P IS NUMBER OF PARITY DRIVES

* RAID-6/-DP requires more than one parity

RAID Partial Writes

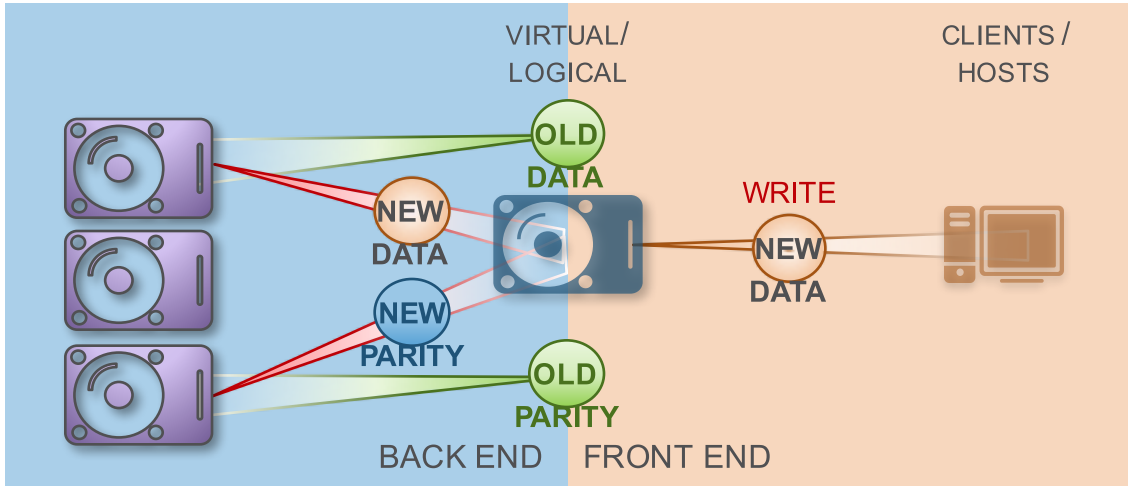
All Single Parity RAID: RAID-3, -4, -5, and etc.

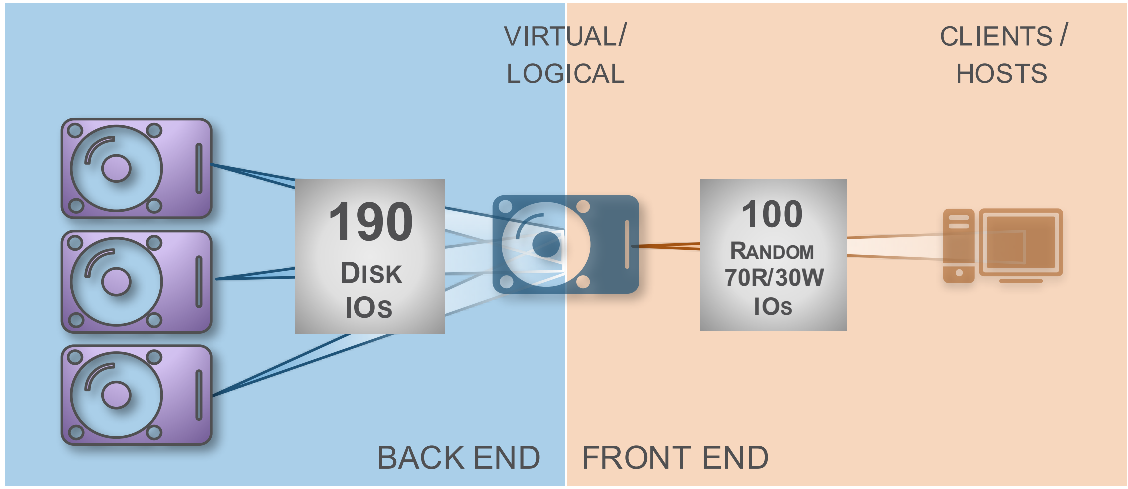
PERFORMANCE



1 READ

2 READS
2 WRITES





SINGLE PARTIAL WRITE:

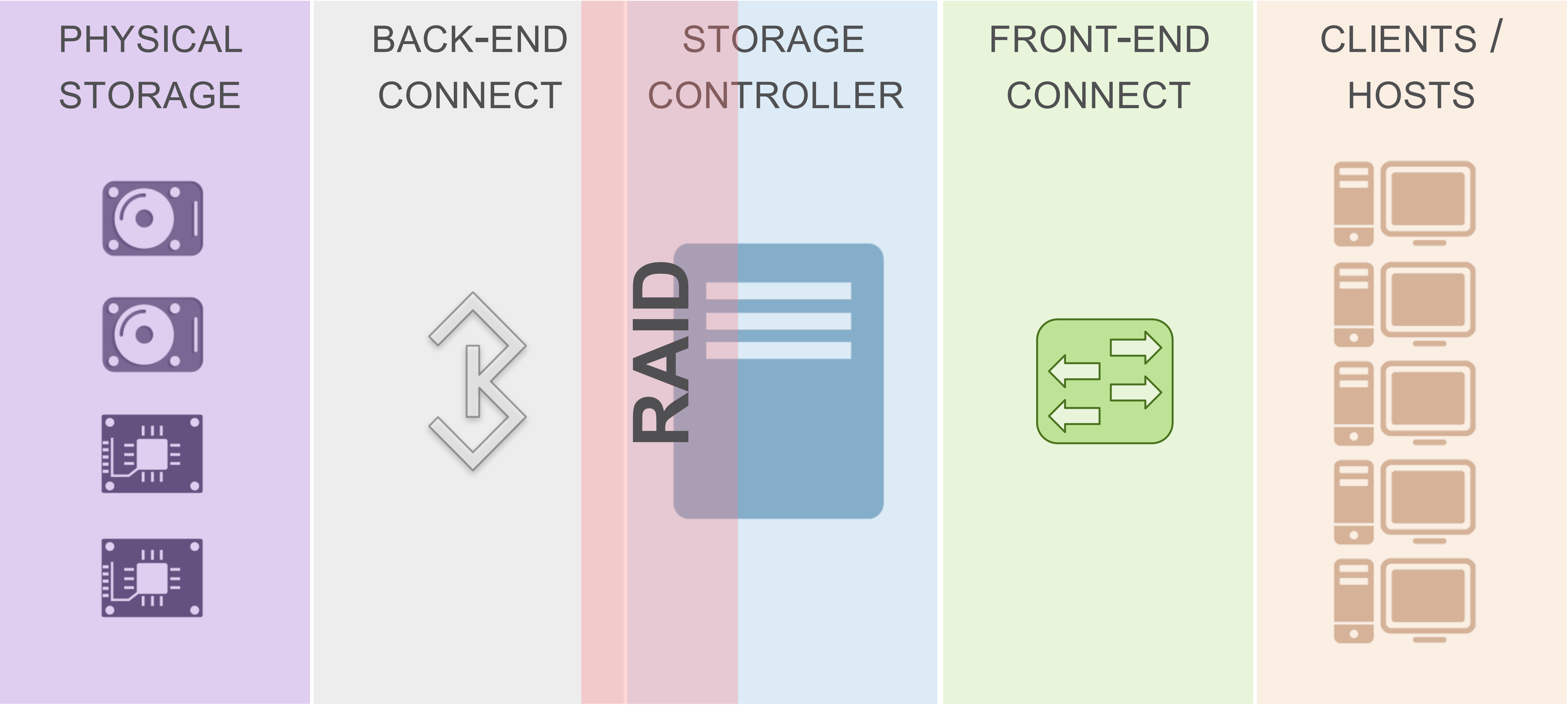
- READ OLD DATA
- READ OLD PARITY
- CALCULATE NEW PARITY
- WRITE NEW DATA
- WRITE NEW PARITY

} 2 READS
2 WRITES

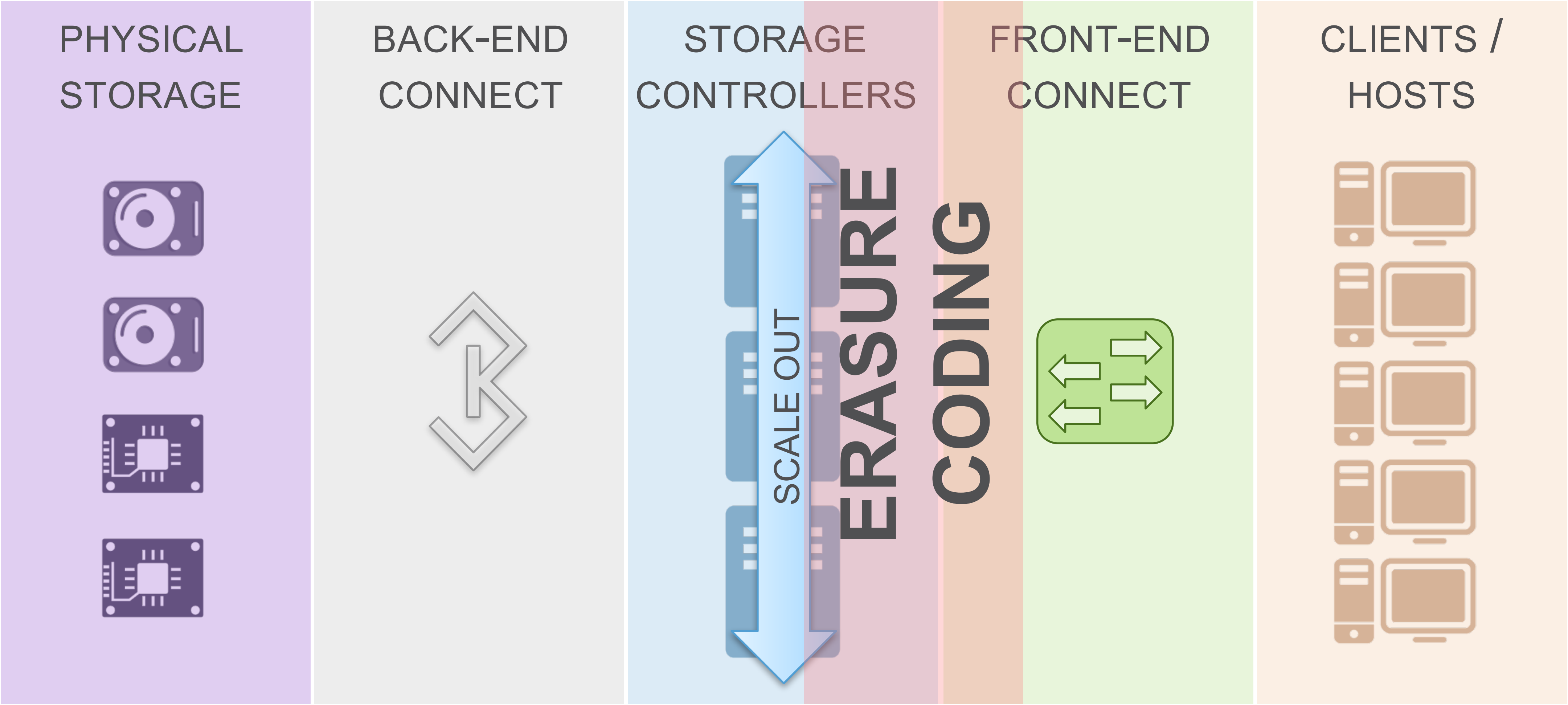
100 IOS 70R/30W = 70 READ + 30 WRITE IOS

$$\text{BACKEND} = (70R + 30 * \underbrace{(2W + 2R)}_{\text{RAID PENALTY}}) = 190\ \text{IOS}$$

RAID Implementation

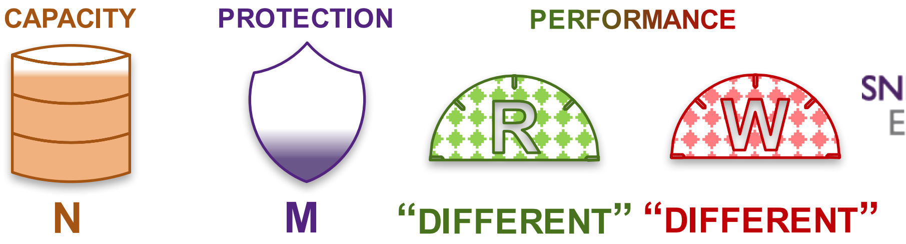


Erasure Coding Implementation

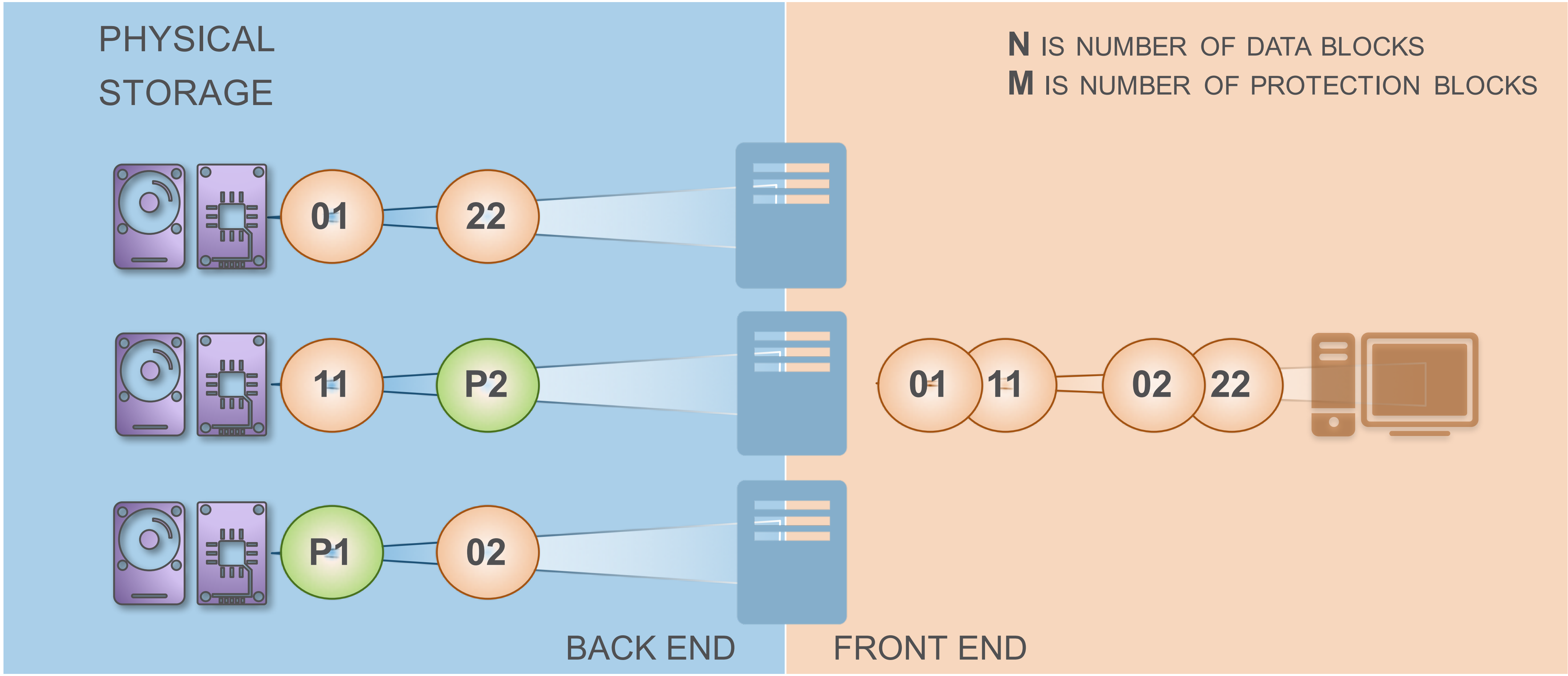


Erasure Coding

N + M = 2 + 1



SNIA ESF | ETHERNET STORAGE



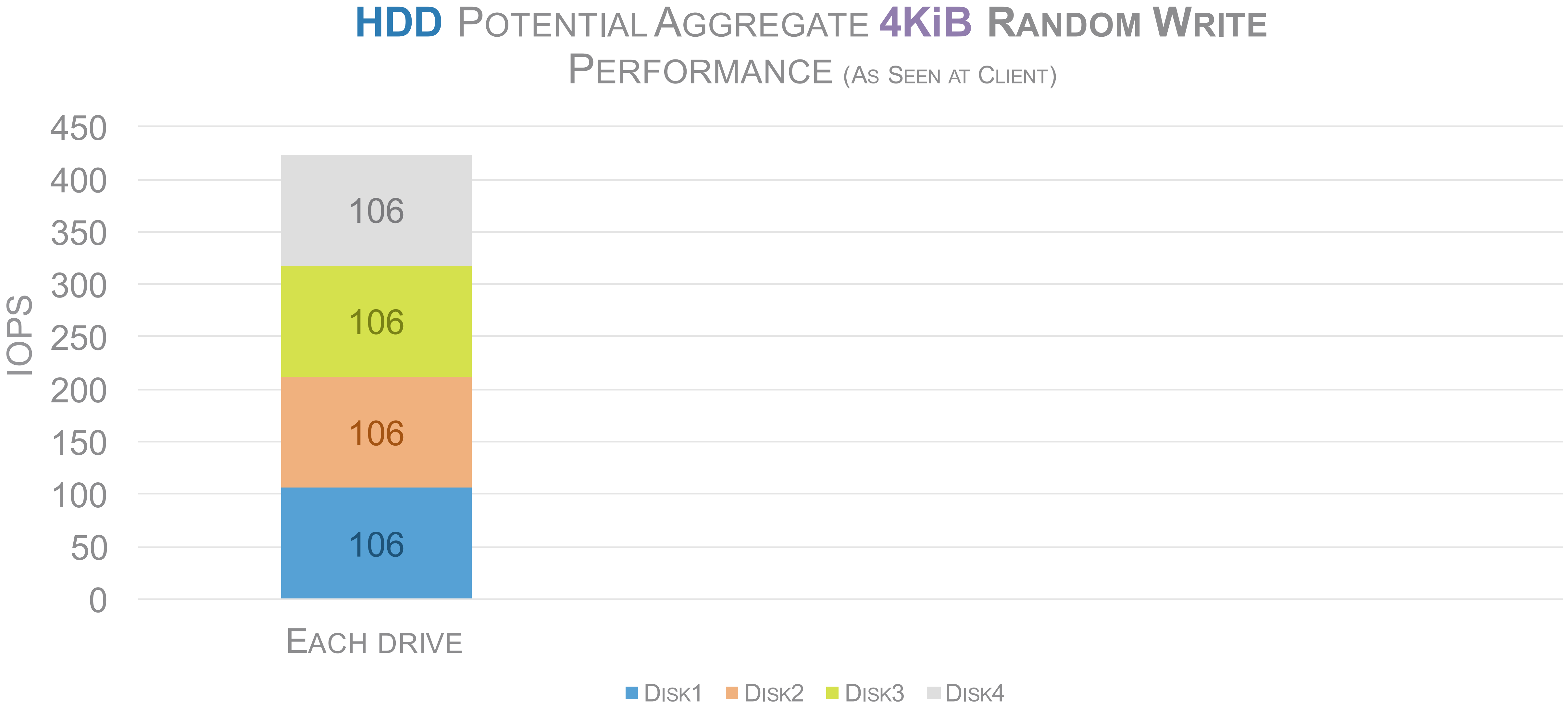
N IS NUMBER OF DATA BLOCKS

M IS NUMBER OF PROTECTION BLOCKS

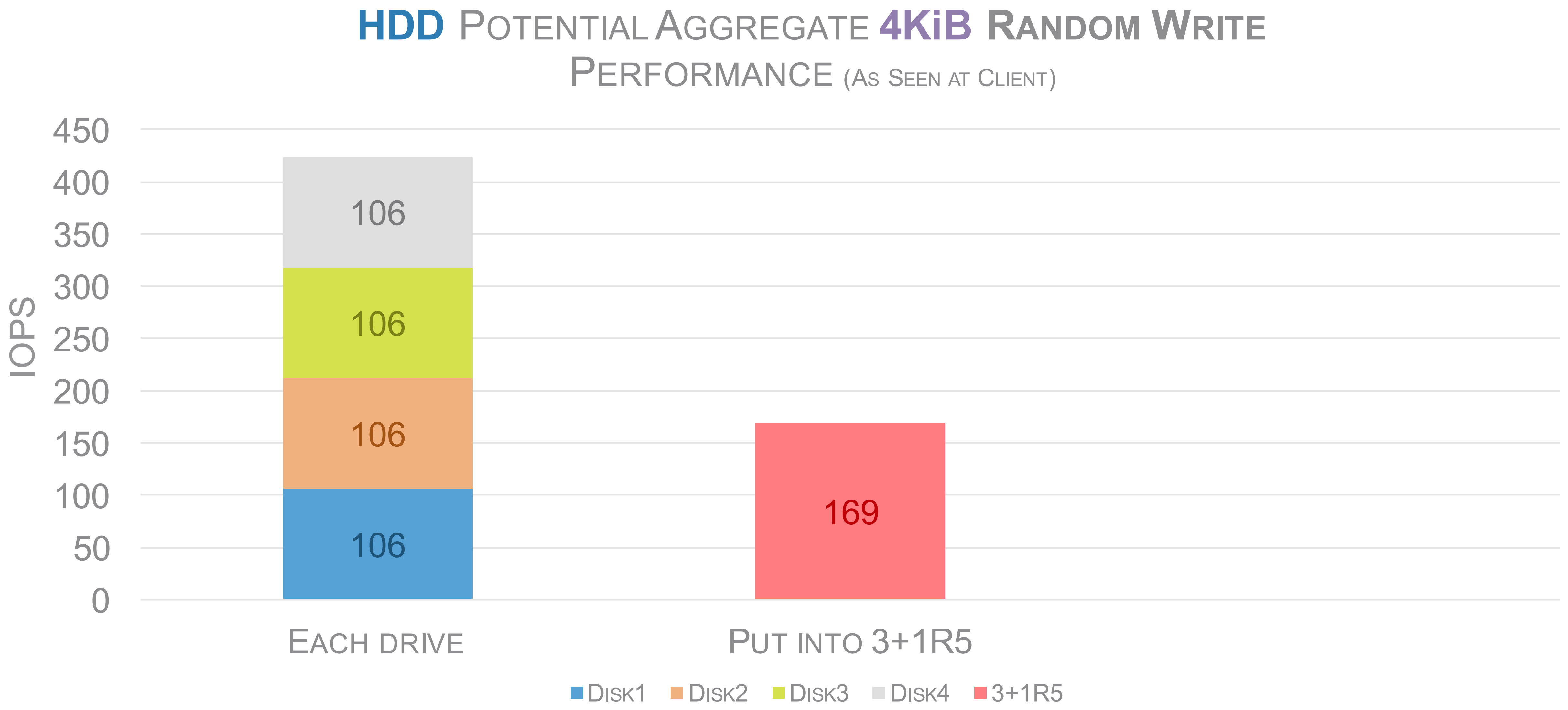
Agenda

INTRO	Introduction
R/W	Reading, Writing; What is the Difference?
TECH	How does this tech work anyway?
RAID	What if you need more than one?
FUN	Performance?
END	Summary

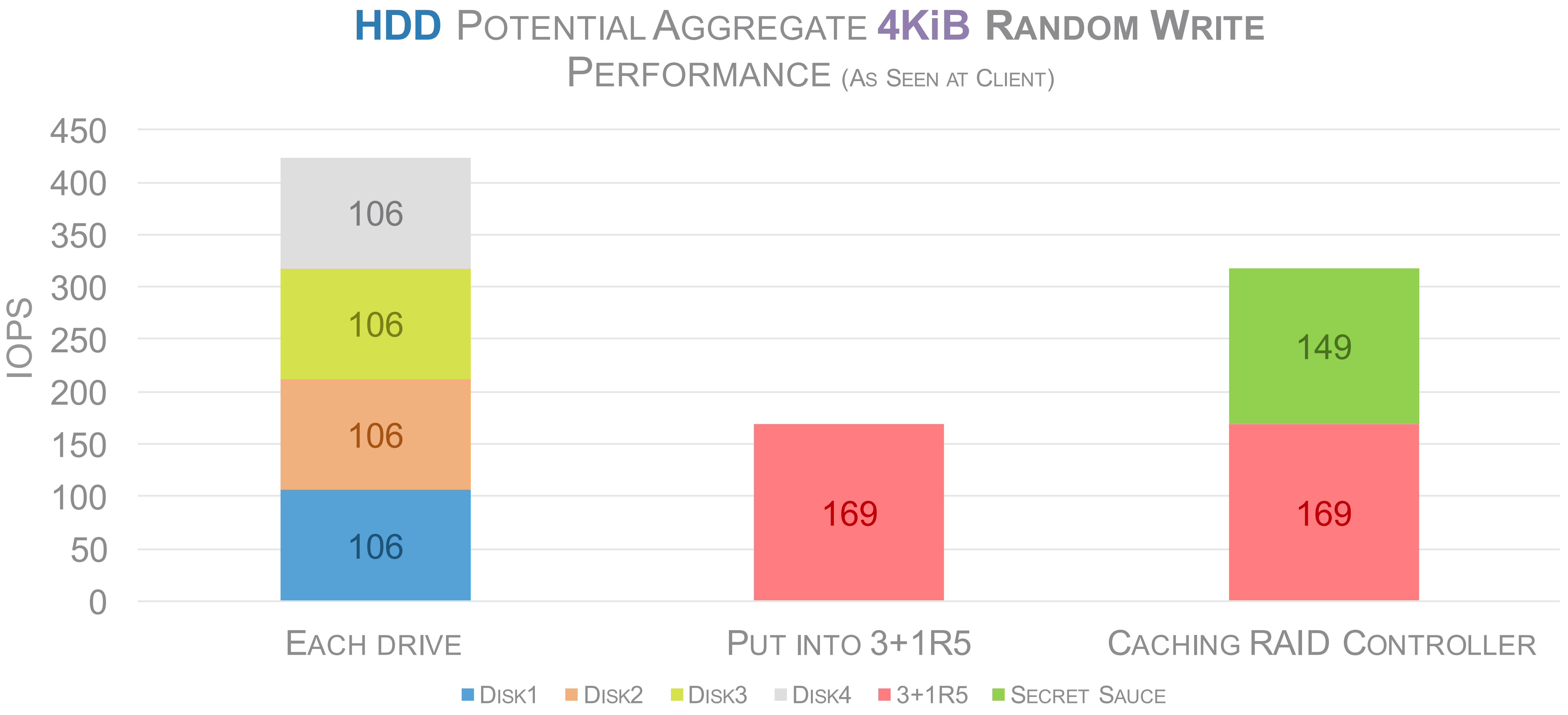
What “Really” Happens With RAID-5?



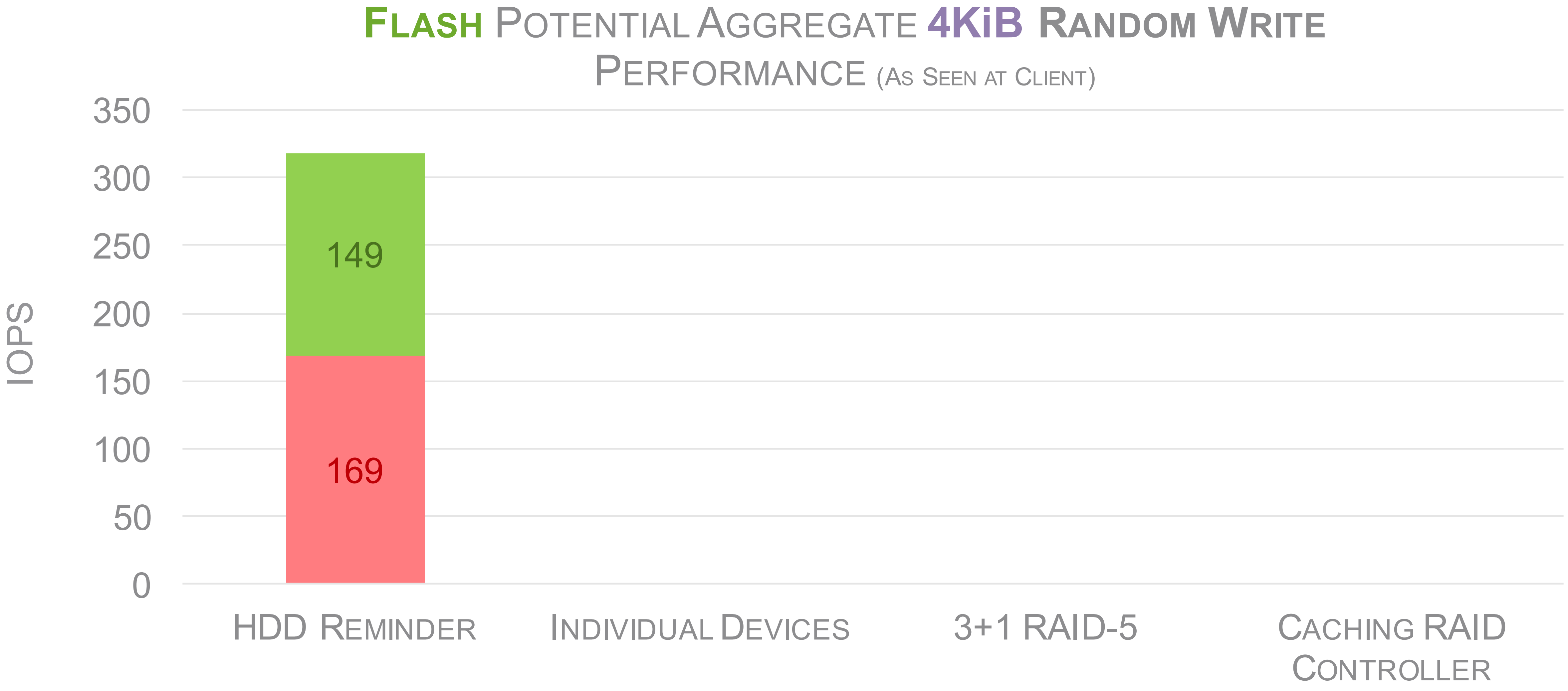
What "Really" Happens With RAID-5?



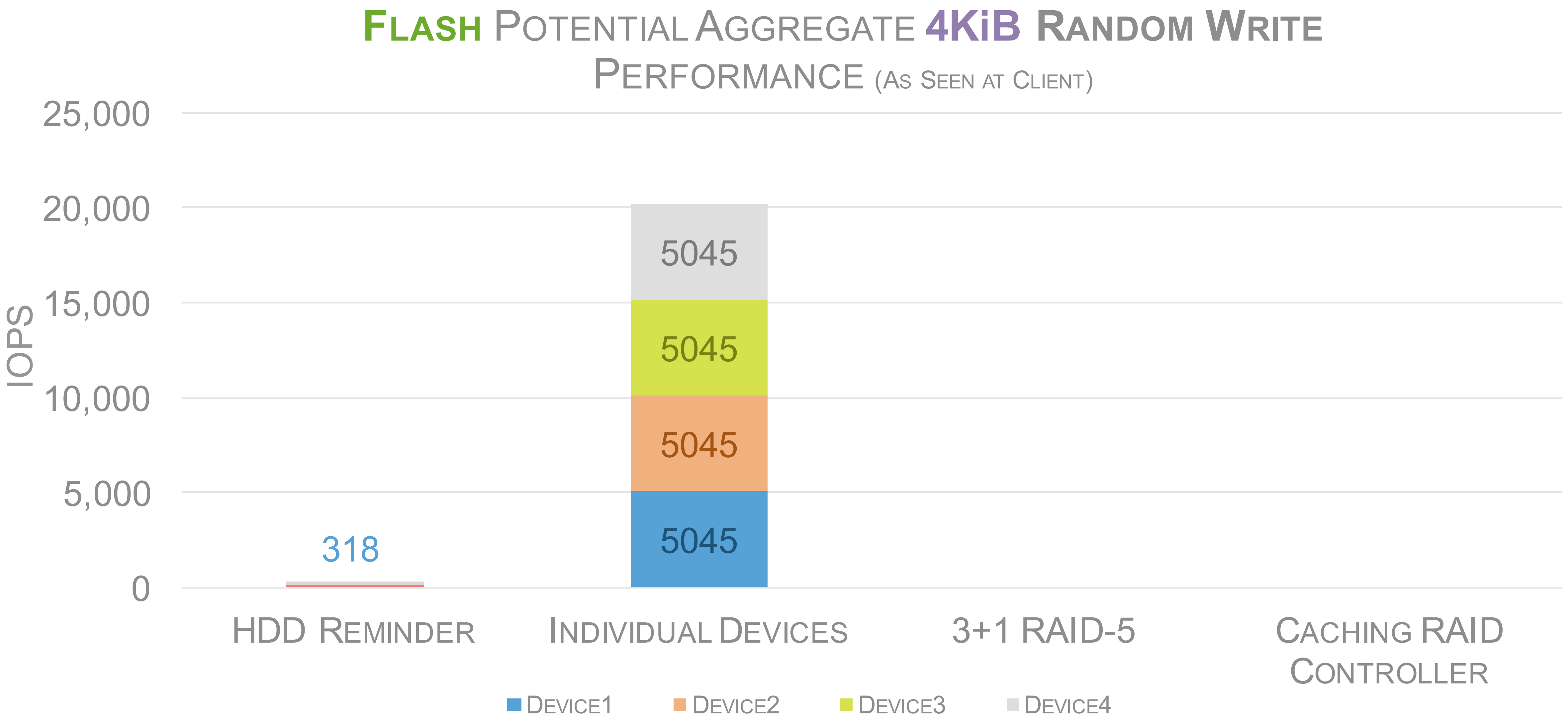
What “Really” Happens With RAID-5?



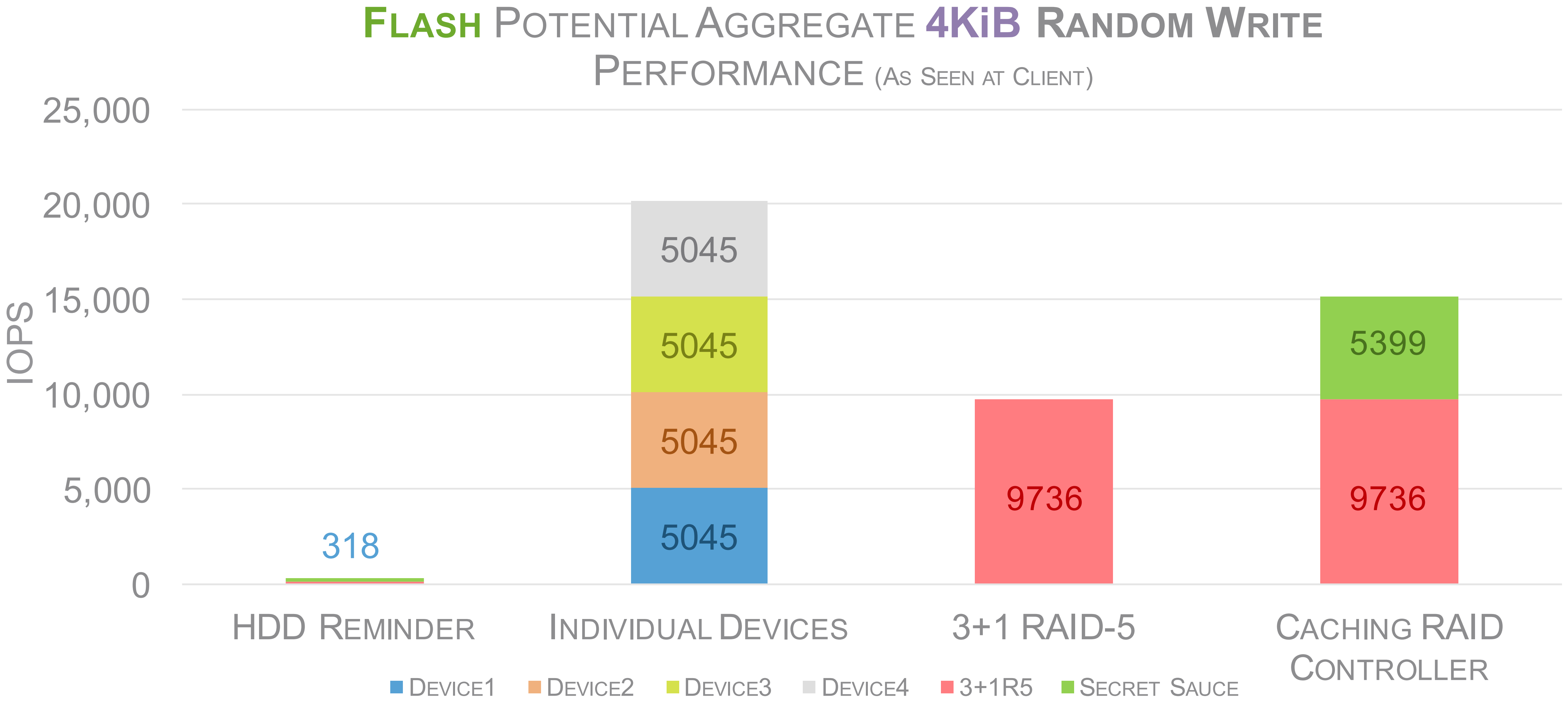
What “Really” Happens With RAID-5?



What "Really" Happens With RAID-5?

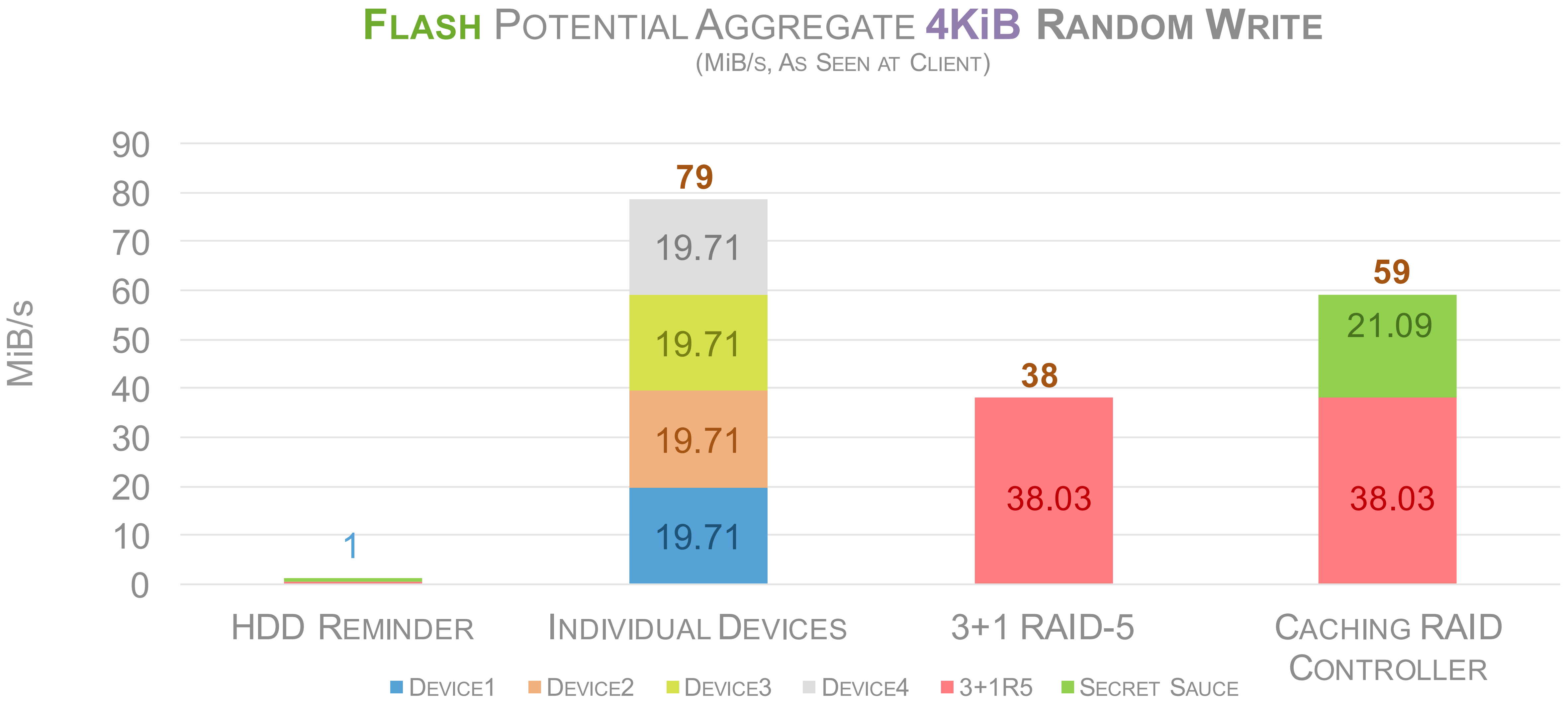


What “Really” Happens With RAID-5?

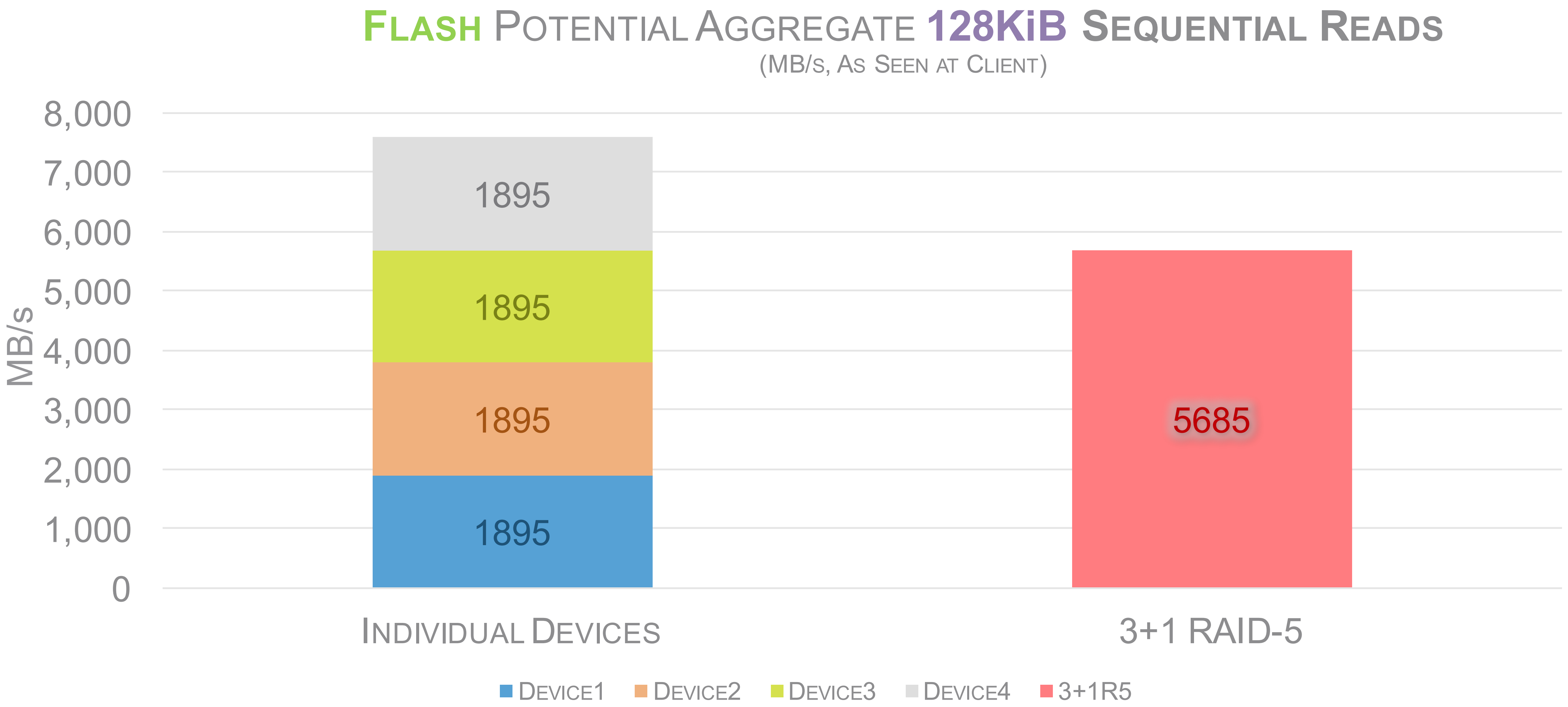


What “Really” Happens With RAID-5?

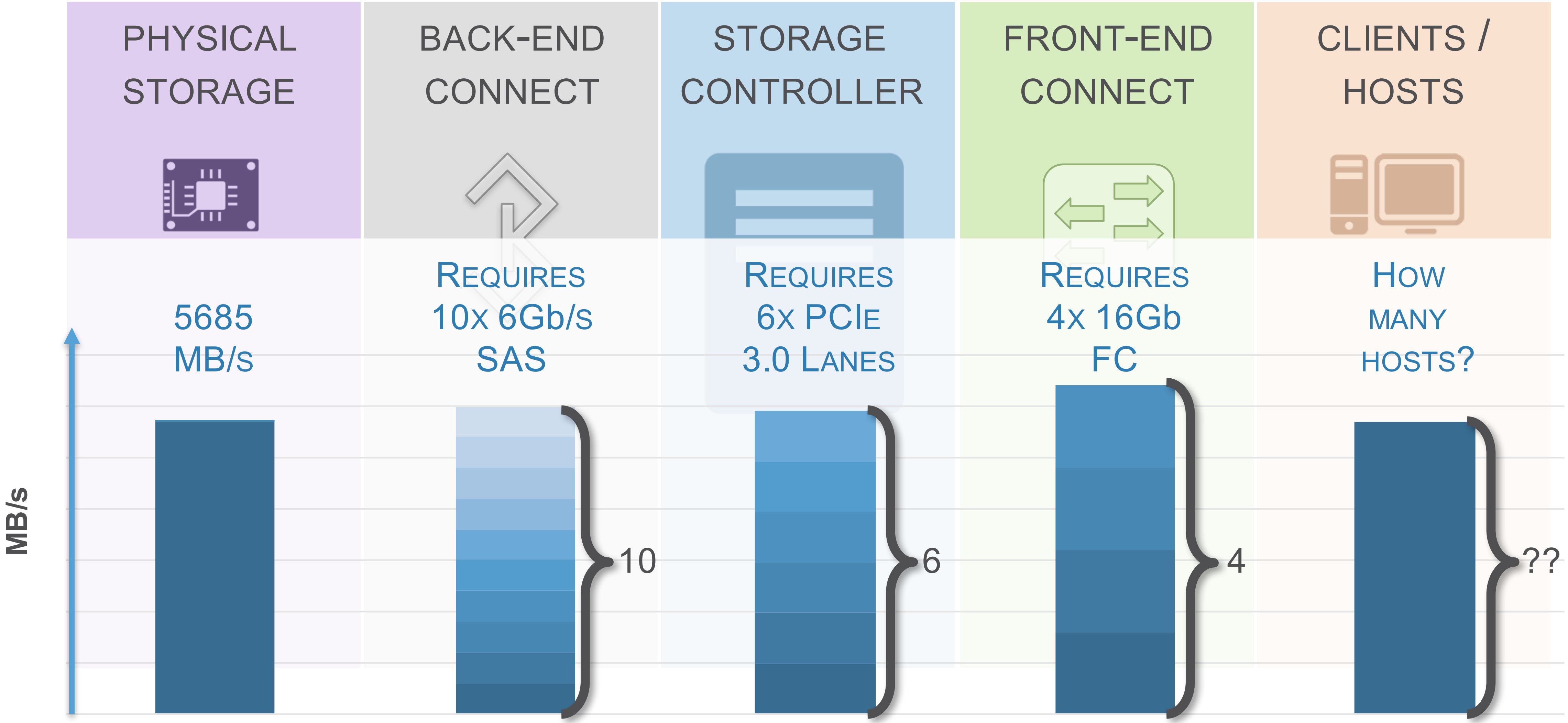
SNIA ESF | ETHERNET STORAGE



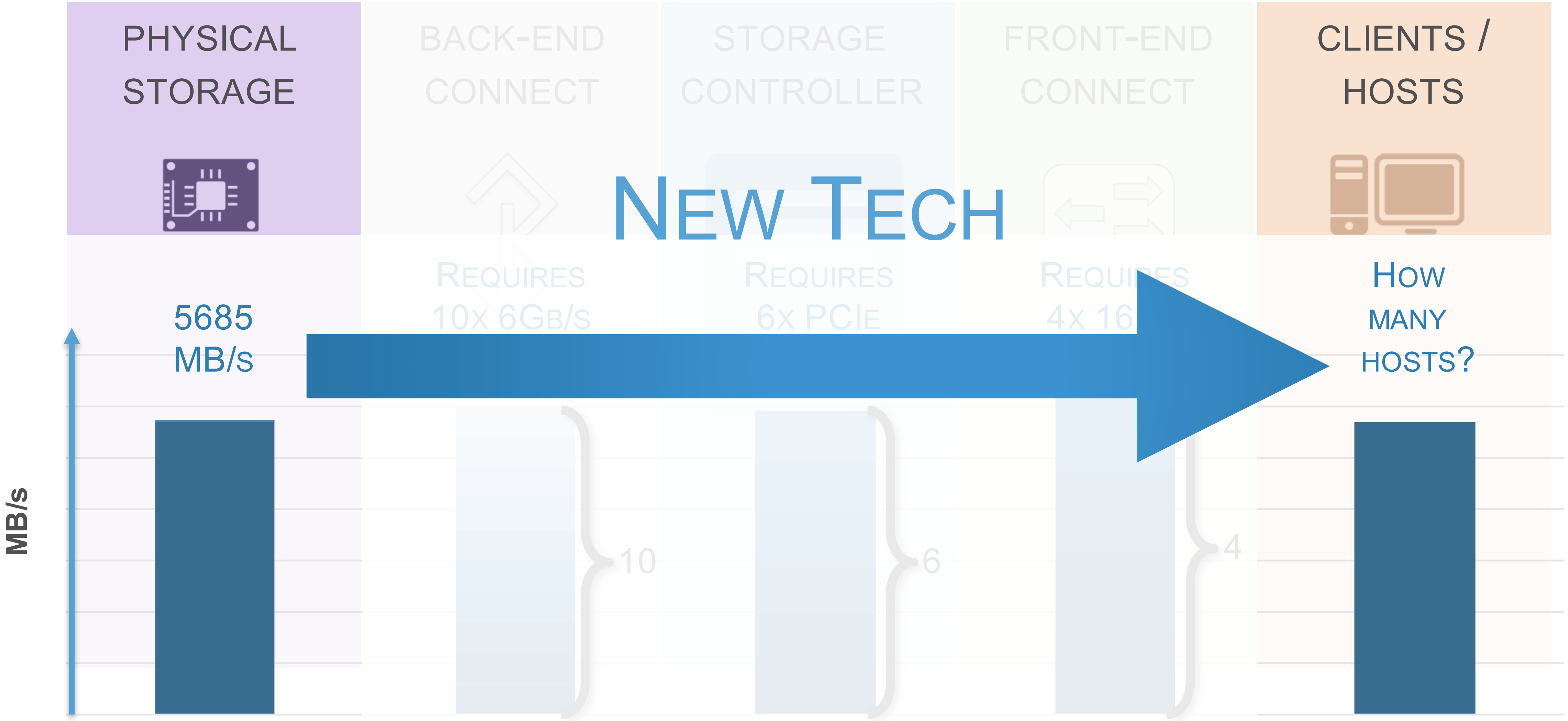
What “Really” Happens With RAID-5?



Flash In The Real World



Flash In The Real World



PHYSICAL STORAGE

BACK-END CONNECT

STORAGE CONTROLLER

FRONT-END CONNECT

CLIENTS / HOSTS

New Tech

5685 MB/s

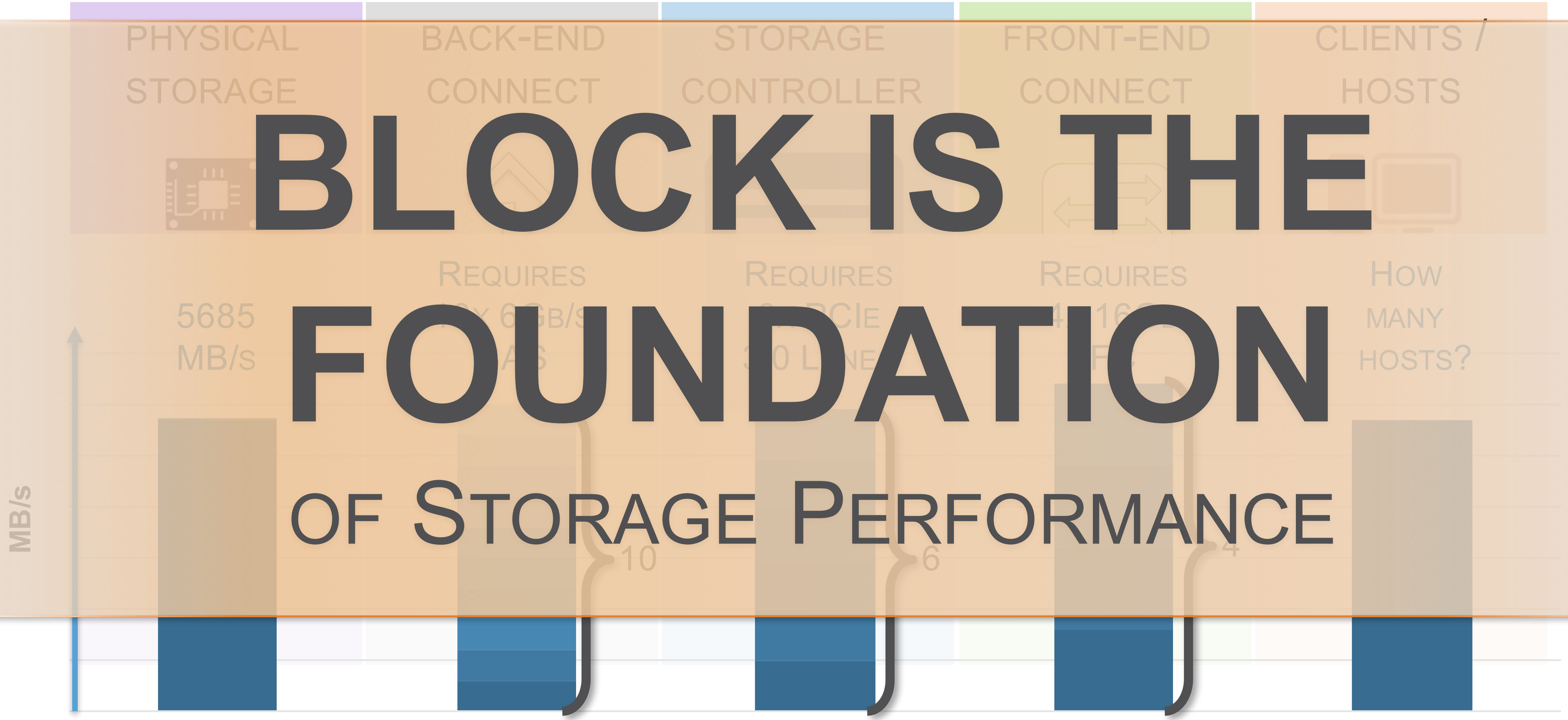
Requires 10x 6Gb/s

Requires 6x PCIe

Requires 4x 16

How many hosts?

Flash In The Real World



BLOCK IS THE FOUNDATION

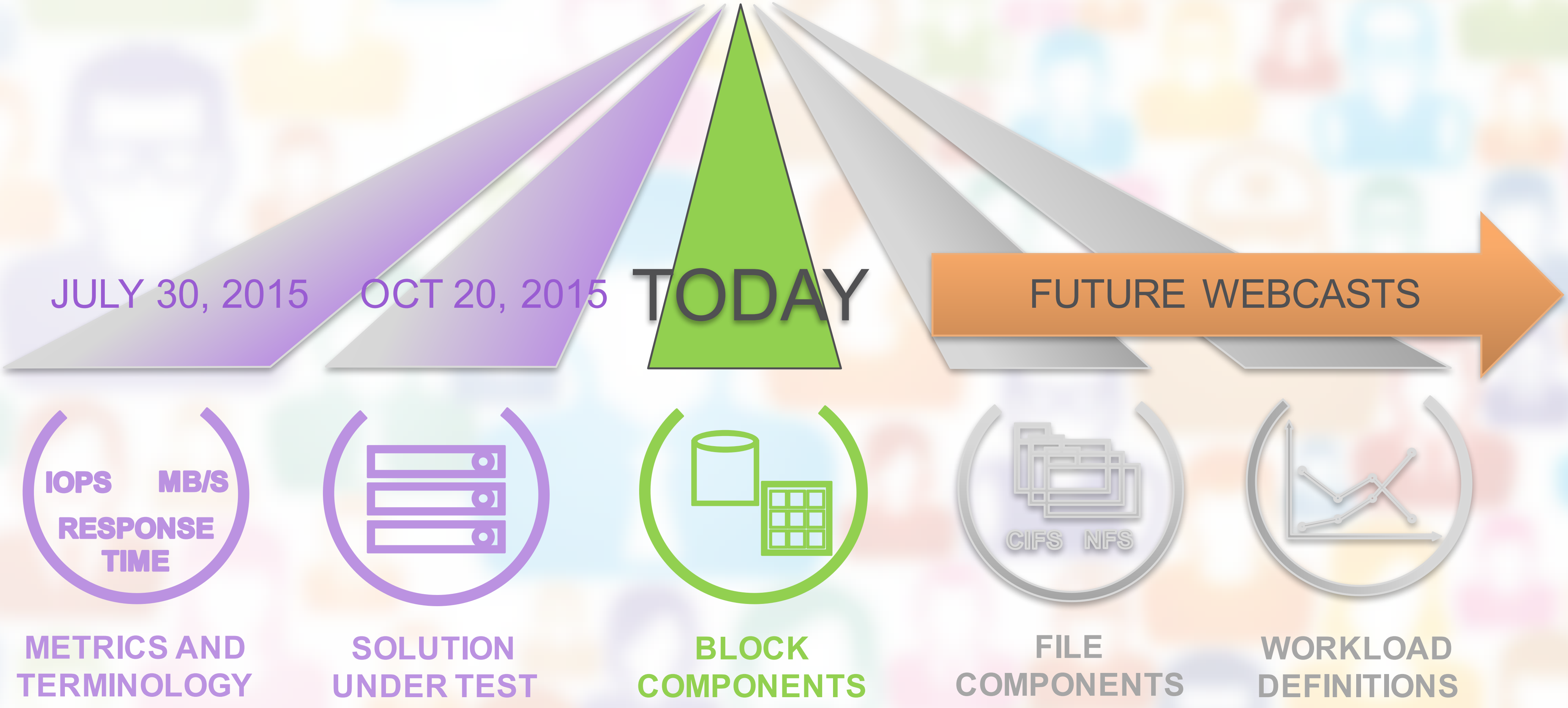
OF STORAGE PERFORMANCE

Agenda

SNIA ESF | ETHERNET STORAGE

INTRO	Introduction
R/W	Reading, Writing; What is the Difference?
TECH	How does this tech work anyway?
RAID	What if you need more than one?
FUN	Performance?
END	Summary

Storage Performance Benchmarking



METRICS AND TERMINOLOGY

SOLUTION UNDER TEST

BLOCK COMPONENTS

FILE COMPONENTS

WORKLOAD DEFINITIONS

INTRO > R/W > TECH > RAID > FUN > END

After This Webcast

- A PDF and a PPT of the slides for this and all previous parts of this Webcast series will be posted to the SNIA Ethernet Storage Forum (ESF) website and available on-demand
  - PPT and PDF: http://www.snia.org/forums/esf/knowledge/webcasts
  - Presentation Recording: https://www.brighttalk.com/webcast/663/164323
- A full Q&A from this webcast, including answers to questions we couldn't get to today, will be posted to the SNIA-ESF blog
  - http://sniaesfblog.org/
- Follow us on Twitter @SNIAESF, @RogovMark, @KenCantrellJr, @DrJMetz
- Next Webcast – Second Half of 2016
  - "Storage Performance Benchmarking: Part 4"

QUESTIONS?

SNIA ESF | ETHERNET STORAGE

THANK YOU!

Appendix – Additional Reading

Appendix – More Reading

- SNIA S3 TWG Guide to SSD Performance: http://www.snia.org/sites/default/files/UnderstandingSSDPerformance.Jan12.web_.pdf
- SNIA S3 TWG SSD Performance Primer, 2013: http://www.snia.org/sites/default/files/SNIASSSI.SSDPerformance-APrimer2013.pdf
- Benchmarking methods for randomly sampling from a file, and why random seeks can (usually) hurt performance: http://simpsonlab.github.io/2015/05/19/io-performance/
- Excellent hard drive overview: https://www.backblaze.com/hard-drive.html
- SSD Performance results: http://www.tomshardware.com/charts/ssd-charts-2014/benchmarks,129.html
- SSD Performance results: http://www.anandtech.com/show/6433/intel-ssd-dc-s3700-200gb-review/3
- Intel Performance Benchmarking for PCIe* and NVMe* Enterprise Solid-State Drives: http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/performance-pcie-nvme-enterprise-ssds-white-paper.pdf
- SSD M.2 Interface: http://arstechnica.com/gadgets/2015/02/understanding-m-2-the-interface-that-will-speed-up-your-next-ssd/
- More complete SSD interface article, covering NVMe, U.2 and M.2: http://blog.ocz.com/ssd-interfaces-sata-m2-u2-nvme/
- SSD vs HDD performance characteristics: http://www.tomshardware.com/reviews/ssd-gaming-performance,2991-3.html

- RAID
  - http://www.raid-recovery-guide.com/raid5-parity.aspx
  - http://rickardnobel.se/how-raid5-works/
  - http://igoro.com/archive/how-raid-6-dual-parity-calculation-works/
  - RAID Perf Calculator: http://wintelguy.com/raidperf.pl
  - RAID Reliability Calculator: http://wintelguy.com/raidmttdl.pl
  - RAID Failure Calculator: http://raid-failure.com/raid10-50-60-failure.aspx
  - RAID Survival Rate Simulation: https://linustechtips.com/main/topic/103179-lets-talk-about-raid-survival-rates/