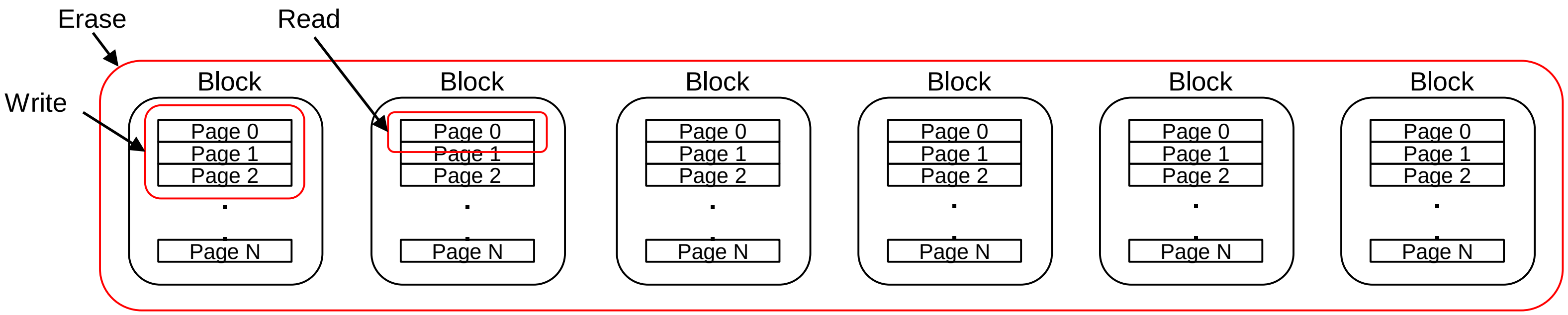
Tunable (and Flexible) Flash Translation Layer Improves Storage System ~~Performance~~ Behavior

Chris Bergman - Burlywood

Overview

- Intended Audience:
  - **Data center and cloud storage system designers and architects**
- Current Configuration Options
  - Capacity, OP%, and Endurance
- Why not have more options?
  - GC and Wear Leveling Schemes => Where and how data is placed on the media.
  - Performance Optimizations
  - Data Integrity
- How you evaluate your options is important!
  - Benchmarks, standard tests, and data sheets can be misleading
- There is opportunity!
  - Lower Total Cost of Ownership, Improved drive life
  - Better and more consistent performance

Why do we need an FTL?



- Media Granularities (Erase >> Write >= Read), Sequential Programming
- Endurance, read disturb, retention, power loss handling, defect handling
- It’s not just flash => Shingled Magnetic Recording HDD, Storage Class Memory

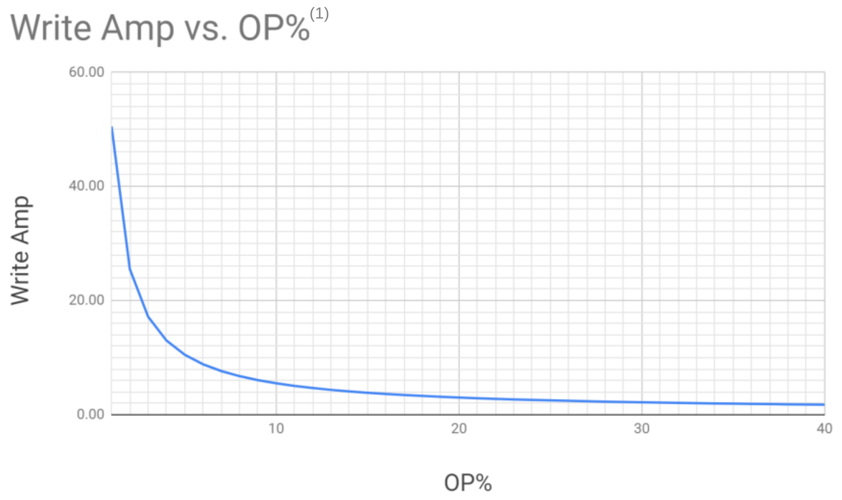
It’s the Storage Media Properties!

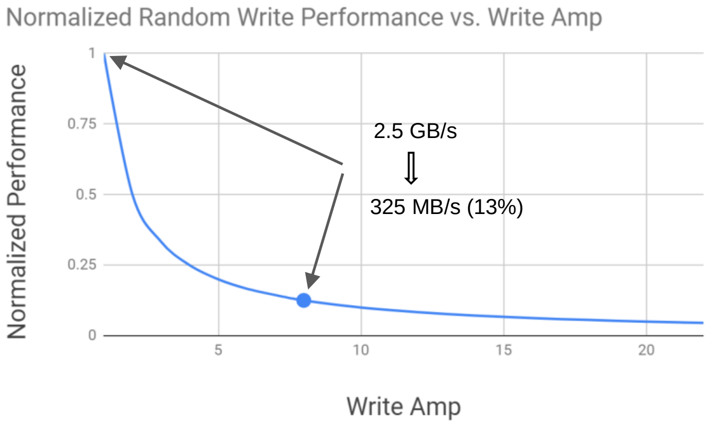
What does an FTL do about these issues?

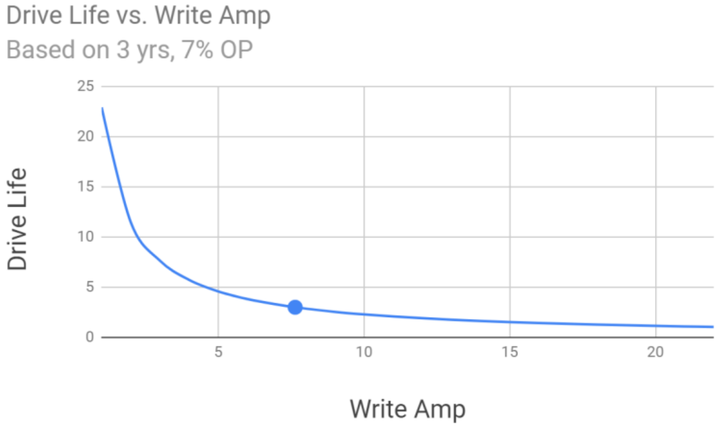
Feature	Side Effect
Translation Tables	Memory Cost (1:1000), Performance
Garbage Collection	Performance, Drive Life [Write Amp]
Wear Leveling	Performance, Drive Life [Write Amp]
Data Integrity (ECC/RAID)	Capacity, Performance, Drive Life
Background Scanning	Data Integrity, Performance
Read/write priority (QoS)	Performance
Overprovisioning	Media Cost, Performance

Necessary evils!

How is it all related?







Lower cost (OP) = **lower** performance and **shorter** drive life.

What do we mean by tunable?

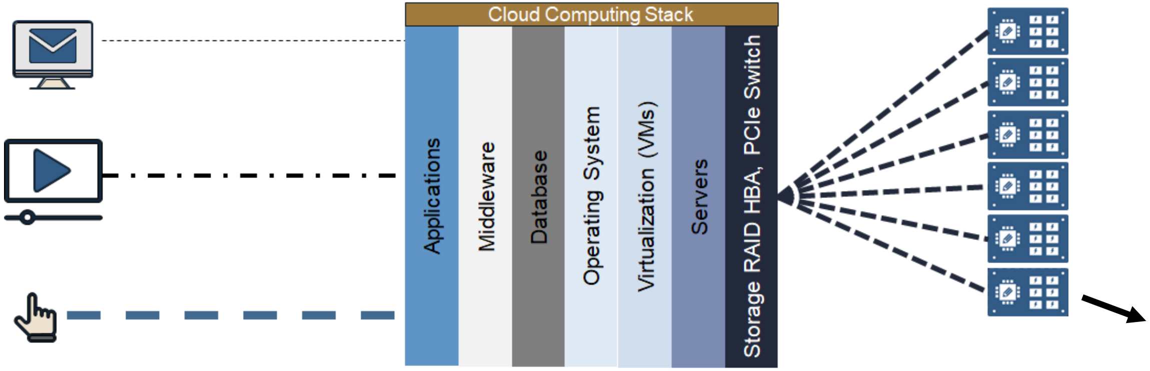
- Imagine the flexibility to optimize based on the application and use model
- Requires knowledge of the workload at the drive
- Traditional FTL’s are statically configured, one size fits all
  - Pick a point on the graphs and that’s what you get
  - Designed for least common denominator (4K random write)
  - One Firmware update away from trouble

If you’re not the least common denominator you’re sacrificing something!

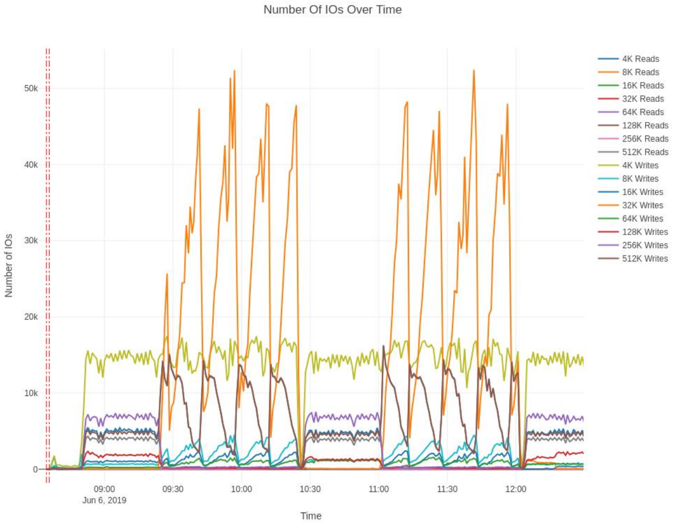
Examples

- Workload complexity
- Read, Write Mixed Workloads, Consistency and QoS
- Good intentions = Not so good results
- Data integrity, ECC optimization

Workloads Are Complex

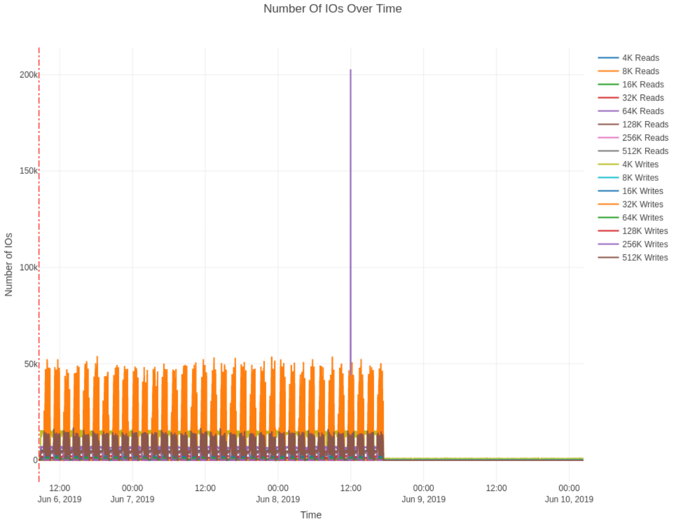


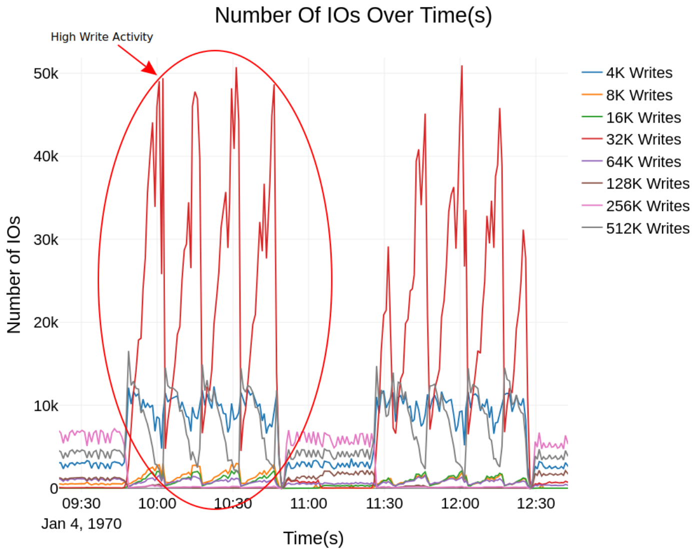
Jetstress[2] + virtualization, RAID, snapshots, etc.



Even the simplest scenario can be complex.

Visualize the workload

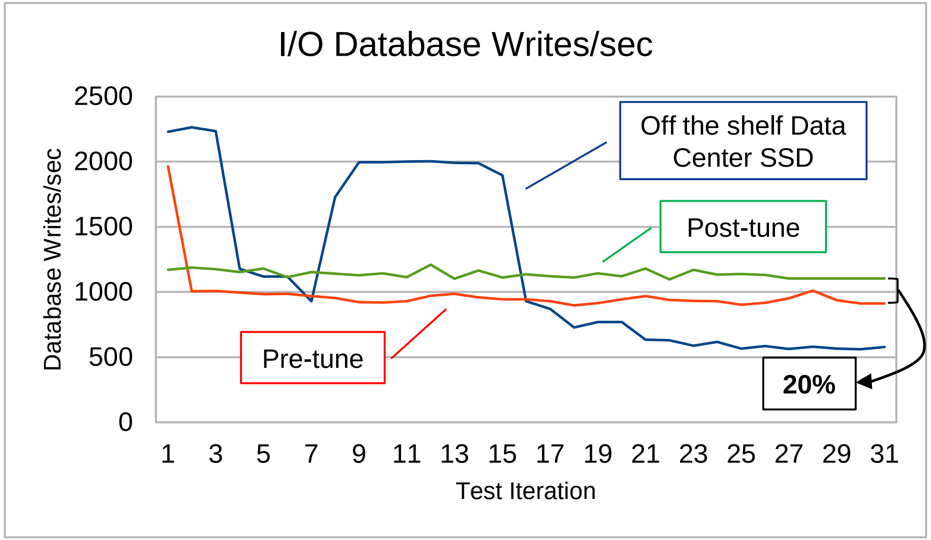


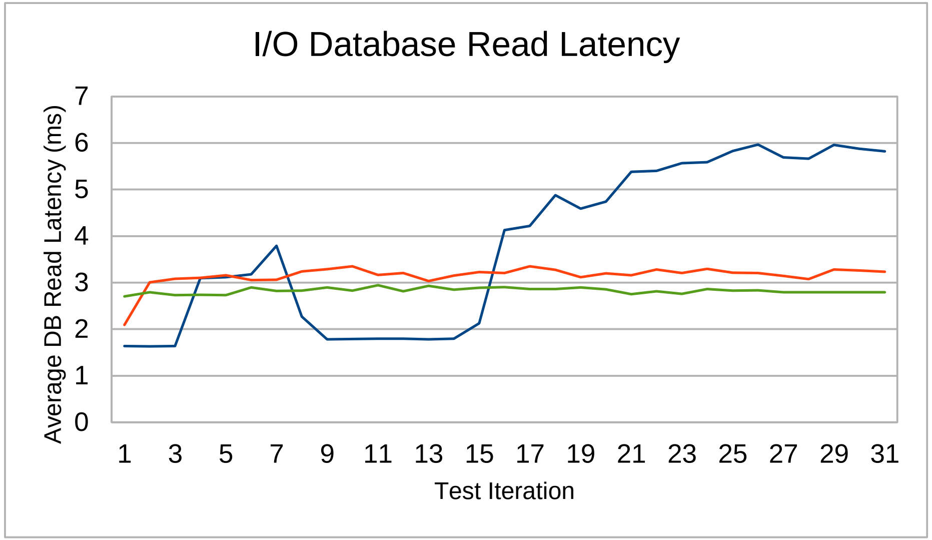


Periodic, heavy write activity followed by very little write activity. Always mixed read/write.

Tuning Results

- Scenario: Optimize garbage collection selection and timing

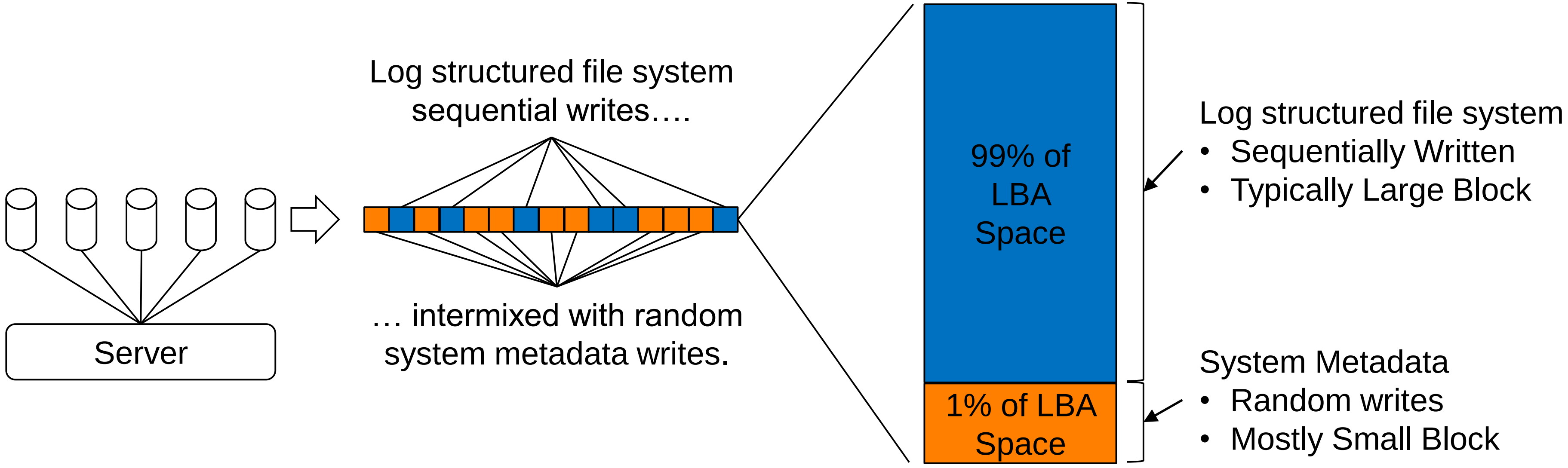




Drives under test, equally pre-conditioned, SATA Data Center quality drives.

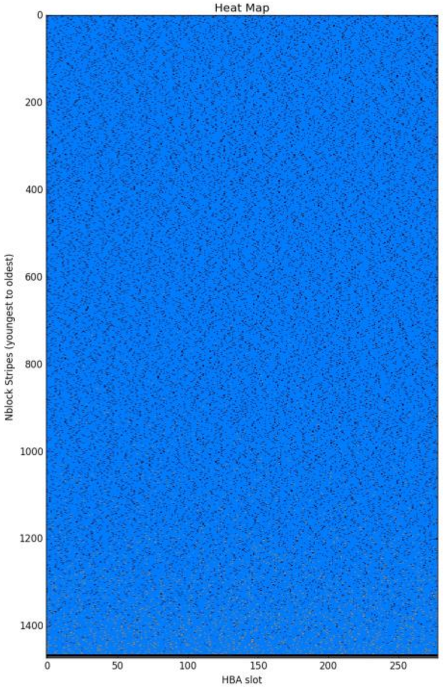
Tuning improved overall performance and consistency.

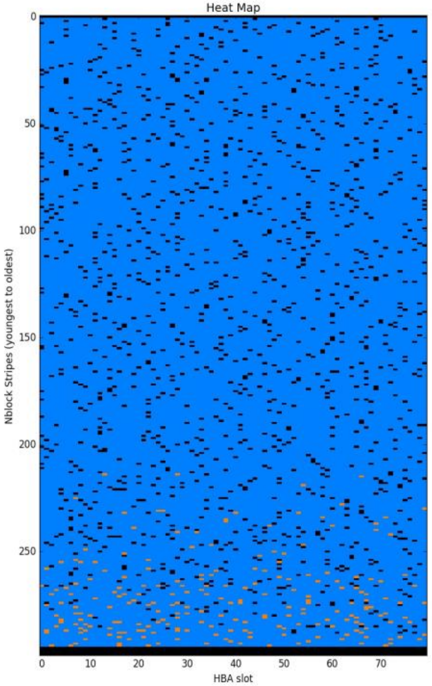
Good Intentions



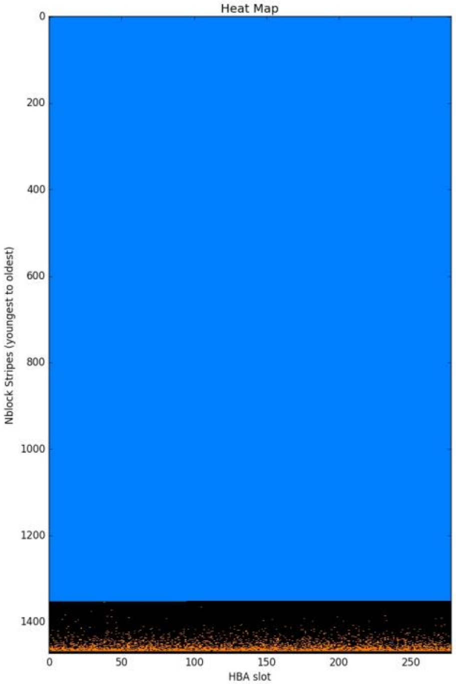
Highly sequential workload may not lead to better behavior.

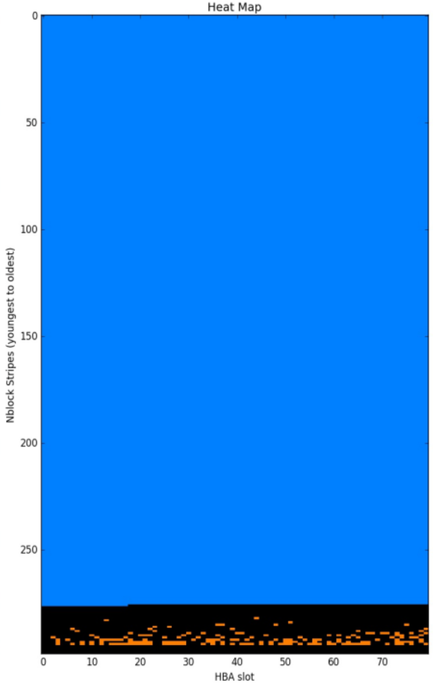
Drive Snapshots





Standard FTL





Optimized FTL

Tuning Results

- Scenario: 16 TB drive, 7% nominal OP, 3 year drive life
- Tuned: Weight OP to small random area + table optimization

Feature	Std FTL	Tuned FTL
Effective WA	4-7	~1
Performance	14%-25% of FOB*	~100% of FOB
Drive Life†	up to 3.3-5.7 years	up to 20+ years
DRAM	16 GB	< 1 GB (or used for other purposes)

* Fresh Out of Box, †Relative to 100% Random Write Workload

Tuned = significant benefit. Standard configuration is highly susceptible to design choices.

Summary

- One size fits all is likely costing you something
- Knowledge is **KEY** and using that knowledge can lead to
  - Lower Total Cost of Ownership
  - Better and more consistent performance
  - Improved drive life
  - Proper evaluation of your storage solution
- Benchmarks and “standard” workloads don’t tell the whole story
- This is even more important in data center applications where inefficiencies can be amplified by 100x, 1000x, 10000x, …

References & Contributors

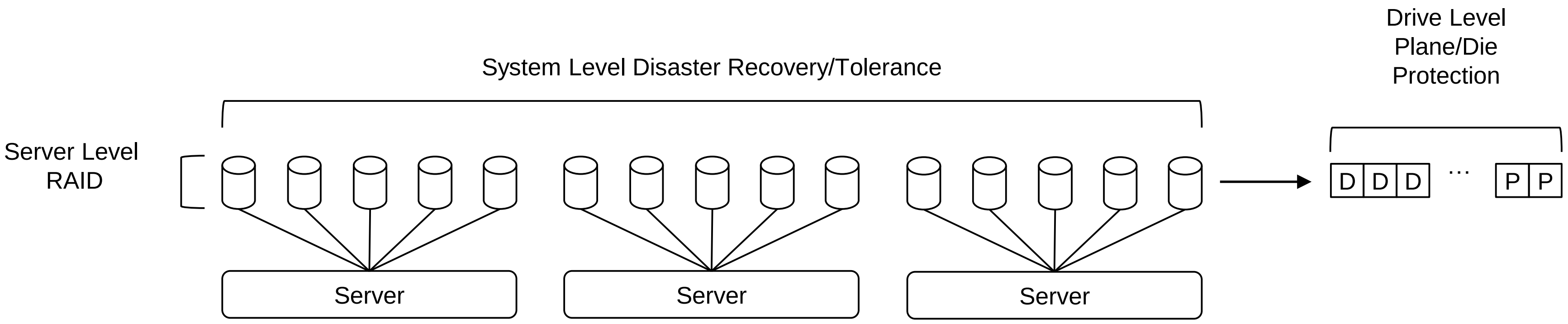
1) Write Amplification Calculation
   - http://www.ece.neu.edu/groups/nucar/NUCARTALKS/WriteAmplification.pdf
2) Jetstress Workload Emulation
   - https://www.microsoft.com/en-us/download/details.aspx?id=36849

Thanks to those who contributed time and effort to this presentation:

Nate Koch, Ed Daelli, John Slattery, Mike Tomky, Tod Earhart, John Murphy, & the entire Burlywood team!

Backup/Reference Slides

Understand the system requirements



Redundancy at many levels

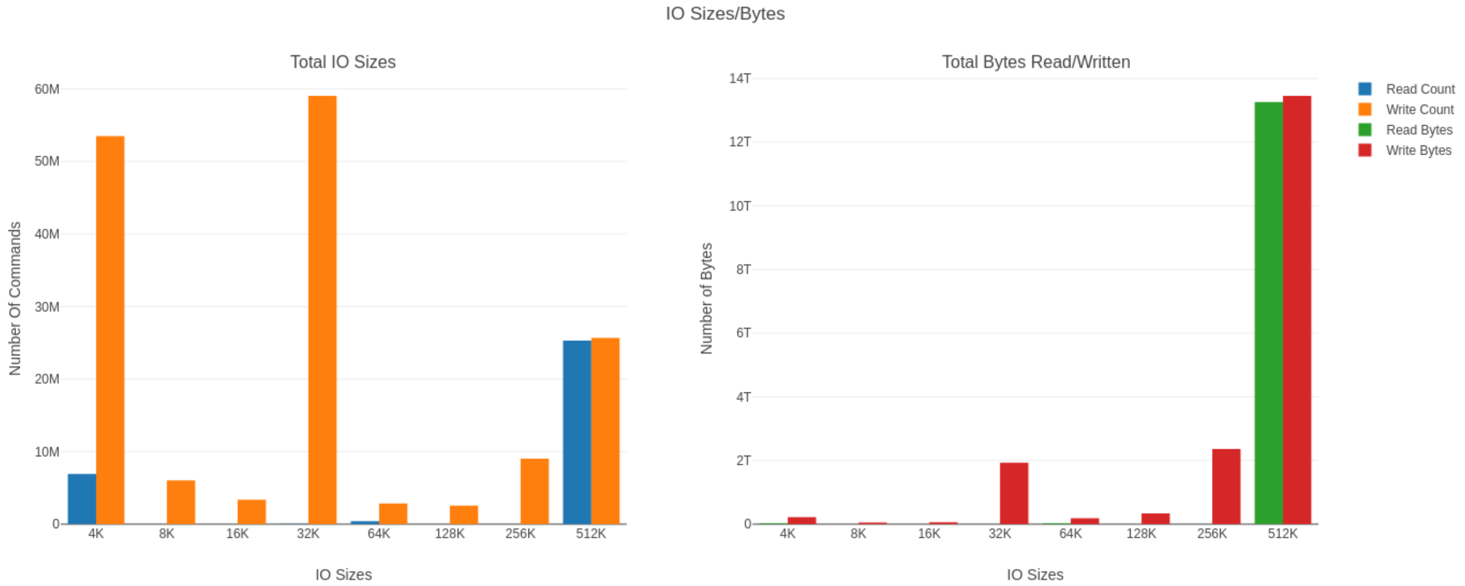
Tune accordingly

- Scenario: 4 TB drive, 7% nominal OP
- Eliminate or reduce protection on drive, redirect to OP

Config	Metric	LUN Protection	Single Plane Protection	None
128 planes/stripe [quad plane]	Perf	13% of FOB*	20% of FOB	21% of FOB
	Life†	3 years	4.0 years	4.2 years
128 planes/stripe [dual plane]	Perf	13% of FOB	16% of FOB	17% of FOB
	Life	3 years	3.4 years	3.6 years

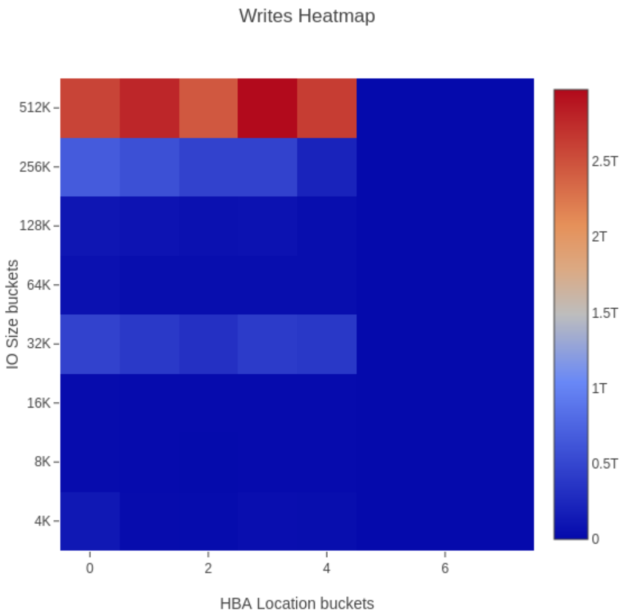
* Fresh Out of Box, †Relative to 100% Random Write Workload

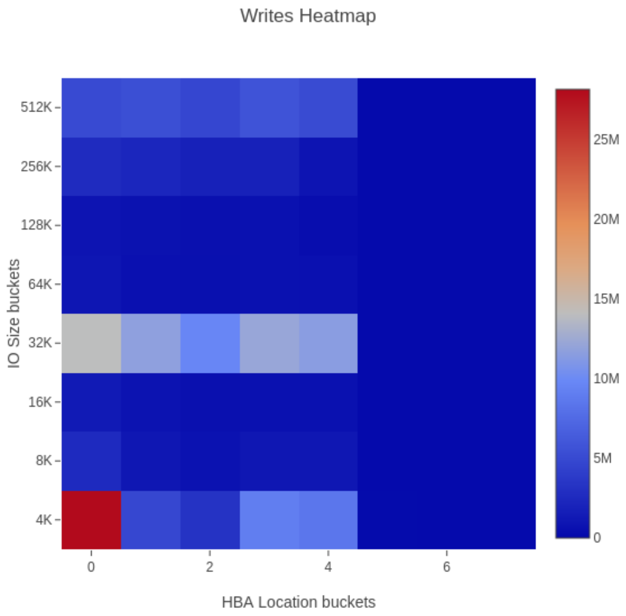
Jetstress IO Breakdown



<=32K WRs: 75% of the traffic, but only ~12% of capacity

Jetstress IO Breakdown





Heatmap view of the same data