Tales of a Daemontown Performance Peddler

Why “It Depends” and What You Can Do About It

Nick Principe
iXsystems

Twitter: @nickprincipe
Github: @powernap
Email: nap@ixsystems.com

Ask Any Performance Person a Question…

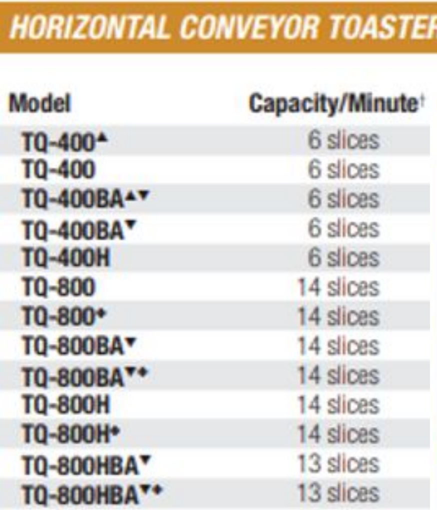
Go ahead… ask me a performance question… :-)

… You’ll Usually Get the Same Answer…

It Depends!

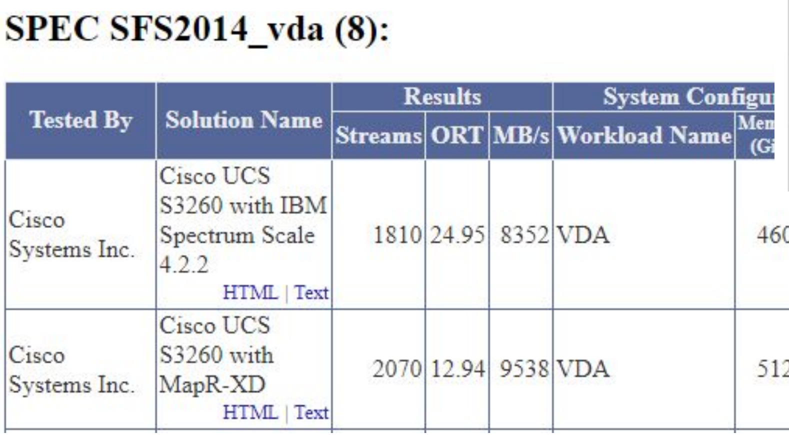
“What’s the Performance of __ ?”

- Want to quantify how something performs
  - Usually for comparison leading to purchase
- Level of detail depends on purpose and value of item



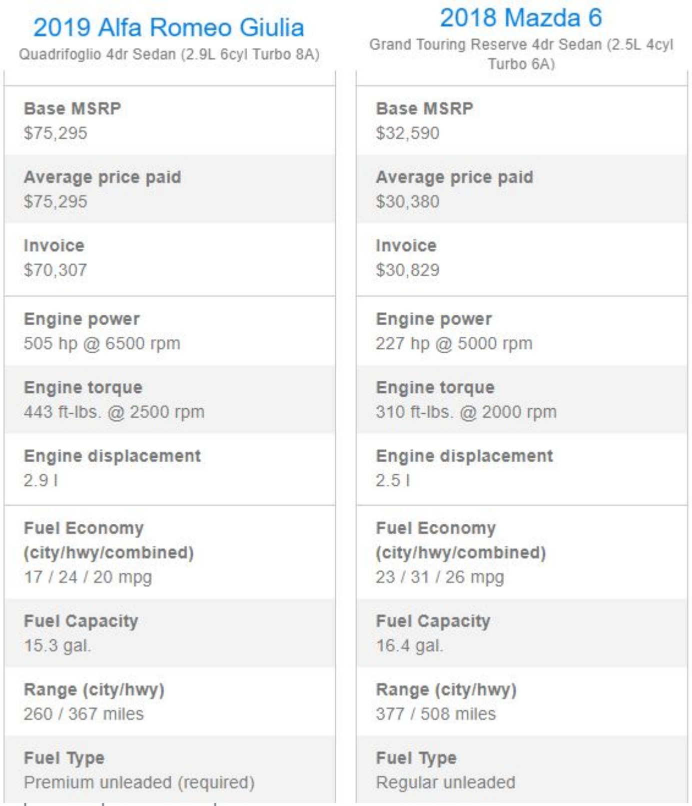
HORIZONTAL CONVEYOR TOASTERS

Model	Capacity/Minute
TQ-400▲	6 slices
TQ-400	6 slices
TQ-400BA▲▼	6 slices
TQ-400BA▼	6 slices
TQ-400H	6 slices
TQ-800	14 slices
TQ-800•	14 slices
TQ-800BA▼	14 slices
TQ-800BA▼•	14 slices
TQ-800H	14 slices
TQ-800H•	14 slices
TQ-800HBA▼	13 slices
TQ-800HBA▼•	13 slices



SPEC SFS2014_vda (8):

Tested By	Solution Name	Results			System Configu			
		Streams	ORT	MB/s	Workload Name	Mem (Gi		
Cisco Systems Inc.	Cisco UCS S3260 with IBM Spectrum Scale 4.2.2 HTML | Text	1810	24.95	8352	VDA	4608	1.1 PiB	Aug 30, 2017
Cisco Systems Inc.	Cisco UCS S3260 with MapR-XD HTML | Text	2070	12.94	9538	VDA	5120	1.5 PiB	Nov 21, 2017

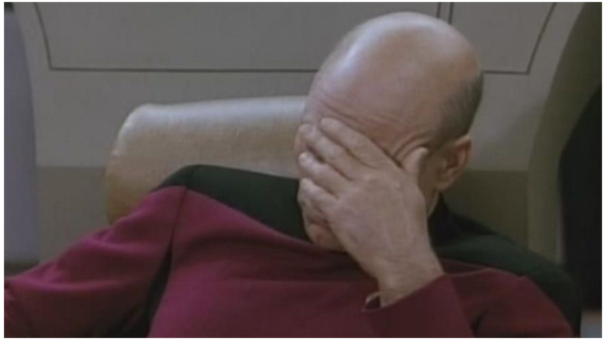


2019 Alfa Romeo Giulia Quadrifoglio 4dr Sedan (2.9L 6cyl Turbo 8A)	2018 Mazda 6 Grand Touring Reserve 4dr Sedan (2.5L 4cyl Turbo 6A)
Base MSRP $75,295	Base MSRP $32,590
Average price paid $75,295	Average price paid $30,380
Invoice $70,307	Invoice $30,829
Engine power 505 hp @ 6500 rpm	Engine power 227 hp @ 5000 rpm
Engine torque 443 ft-lbs. @ 2500 rpm	Engine torque 310 ft-lbs. @ 2000 rpm
Engine displacement 2.9 l	Engine displacement 2.5 l
Fuel Economy (city/hwy/combined) 17 / 24 / 20 mpg	Fuel Economy (city/hwy/combined) 23 / 31 / 26 mpg
Fuel Capacity 15.3 gal.	Fuel Capacity 16.4 gal.
Range (city/hwy) 260 / 367 miles	Range (city/hwy) 377 / 508 miles
Fuel Type Premium unleaded (required)	Fuel Type Regular unleaded

What Performance Do You Need?

- Performance needs are balanced with budget size and capacity needs
- Marketing and spec-sheet numbers can determine basic suitability
  - Applications and environments have their own quirks
- Everyone has different:
  - Minimum performance requirements
    - “Response time must not exceed 5ms”
  - Maximum performance needs
    - “Expected peak load in 3 years is 180k Ops/sec”

And this is where everyone fires up good ol’ dd…

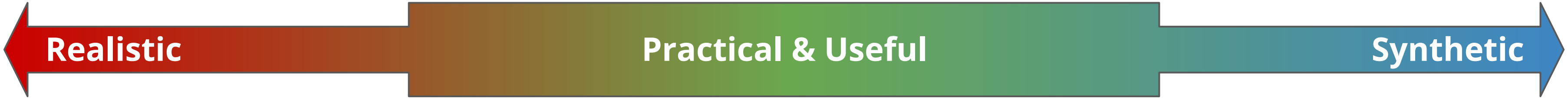


The Nice Thing About Performance Tools is...

- Performance folks like to badmouth dd, but...
  - Everyone does it!
  - First test run on new storage is usually a dd or manual file copy
- In some cases, this is actually perfectly fine!
  - USB flash drive? I **really** care about dd write performance!
  - Home NAS? I **really** care about the performance of a few manual file copies!
- But, for enterprise or ZFS environments, there are better tools out there
  - Alas, there are no perfect tools

The Perfect Performance Benchmark

- The perfect performance benchmark is your production environment
- Rarely is this practical or even possible
  - Hardware and software expense, setup time, execution time, and expertise required
- Therefore, shortcuts are taken
  - Every shortcut slides testing farther toward the synthetic
  - The more synthetic the test, the more “magic” is required to make results useful
- Let’s work together to create more realistic and standardized tests
  - I suggest the SPEC Open Systems Group’s Storage subcommittee as the venue
  - Participation is easy if your company is a SPEC OSG member
    - If not, get in touch with me (nap@ixsystems.com)



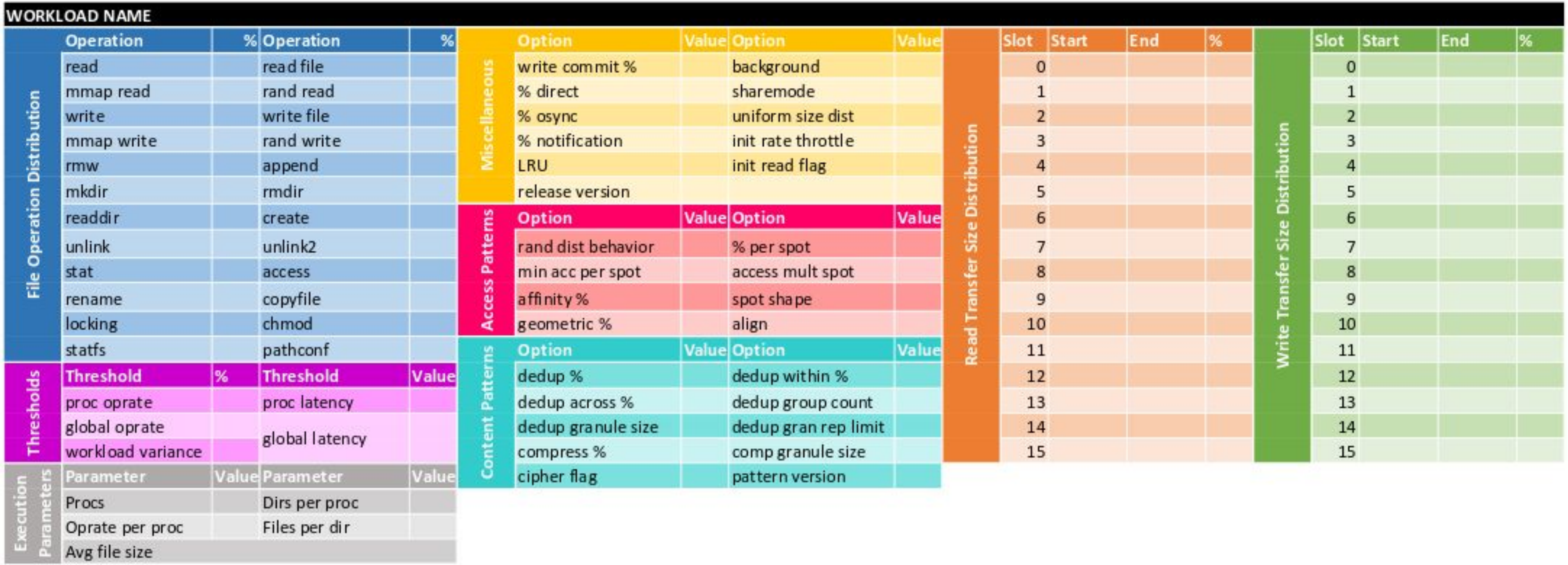
Performance Load Generators I Use

	vdbench	fio	netmist
Cost	free*	free	$2k* (SPEC SFS 2014)
License	Oracle	GPLv2	SPEC
Freedom of Speech	Clearly Restricted	Yes* (Moral License)	Unclearly Restricted
Source Available	Yes	Yes	Yes
Platform Support	Most, except BSD	Most	Most
Multi-host Coordination	Yes	Yes, except Windows	Yes
Flexible Workload Definition	Yes	Yes	Yes
Flexible dataset layout	Yes	Yes	No

More About the Load Generators…

	vdbench	fio	netmist
File/Metadata Operations	Yes	No	Yes
Best for…	“Four corners” or advanced synthetic block tests	“Four corners” or advanced synthetic block tests	Complex workloads involving file-access testing
Cool Thing #1	Easy to iterate over multiple test factors	Awesome easily accessible access pattern distributions	Very repeatable results
Cool Thing #2	Config files very flexible	pkg install fio	Advanced dataset fill pattern and access parameters
Uncool Thing	java	File testing parameters can be confounding	Have to buy it

You Get What You Measure: Workload Parameters



WORKLOAD NAME

	Operation	%	Operation	%
File Operation Distribution	read		read file	
	mmap read		rand read	
	write		write file	
	mmap write		rand write	
	rmw		append	
	mkdir		rmdir	
	readdir		create	
	unlink		unlink2	
	stat		access	
	rename		copyfile	
	locking		chmod	
	statfs		pathconf	

	Threshold	%	Threshold	Value
Thresholds	proc oprate		proc latency	
	global oprate		global latency	
	workload variance			

	Parameter	Value	Parameter	Value
Execution Parameters	Procs		Dirs per proc	
	Oprate per proc		Files per dir	
	Avg file size			

	Option	Value	Option	Value
Miscellaneous	write commit %		background	
	% direct		sharemode	
	% osync		uniform size dist	
	% notification		init rate throttle	
	LRU		init read flag	
	release version			

	Option	Value	Option	Value
Access Patterns	rand dist behavior		% per spot	
	min acc per spot		access mult spot	
	affinity %		spot shape	
	geometric %		align	

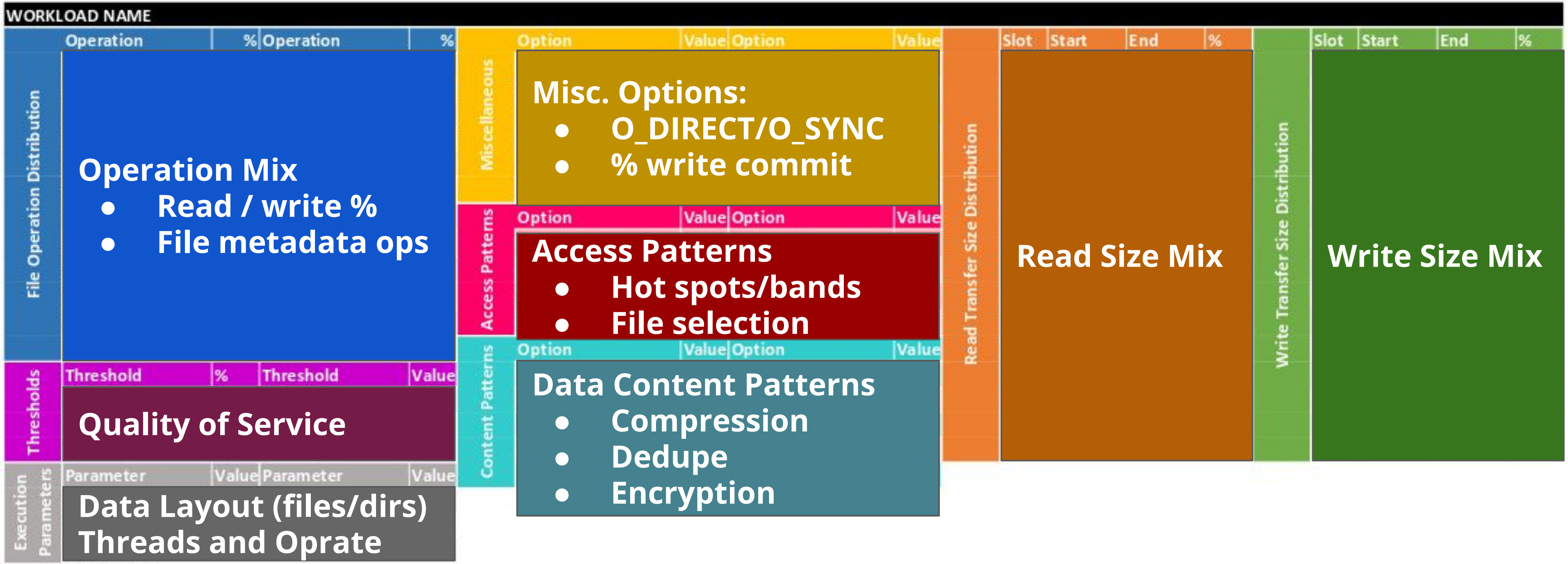
	Option	Value	Option	Value
Content Patterns	dedup %		dedup within %	
	dedup across %		dedup group count	
	dedup granule size		dedup gran rep limit	
	compress %		comp granule size	
	cipher flag		pattern version	

	Slot	Start	End	%
Read Transfer Size Distribution	0			
	1			
	2			
	3			
	4			
	5			
	6			
	7			
	8			
	9			
	10			
	11			
	12			
	13			
	14			
	15			

	Slot	Start	End	%
Write Transfer Size Distribution	0			
	1			
	2			
	3			
	4			
	5			
	6			
	7			
	8			
	9			
	10			
	11			
	12			
	13			
	14			
	15			

SFS 2014 Workload Template: https://spec.org/sfs2014/docs/usersguide.pdf (page 71)

You Get What You Measure: Workload Parameters



Operation Mix

- Read / write %
- File metadata ops

Quality of Service

Data Layout (files/dirs)
Threads and Oprate

Misc. Options:

- O_DIRECT/O_SYNC
- % write commit

Access Patterns

- Hot spots/bands
- File selection

Data Content Patterns

- Compression
- Dedupe
- Encryption

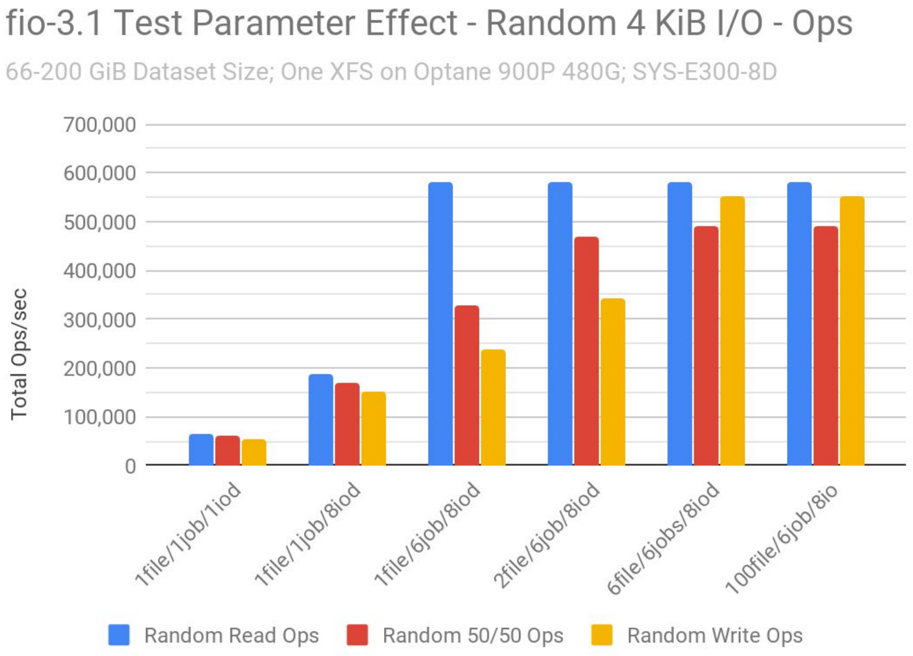
Read Size Mix

Write Size Mix

SFS 2014 Workload Template: https://spec.org/sfs2014/docs/usersguide.pdf (page 71)

You Get What You Measure: Workload Parameter Variation

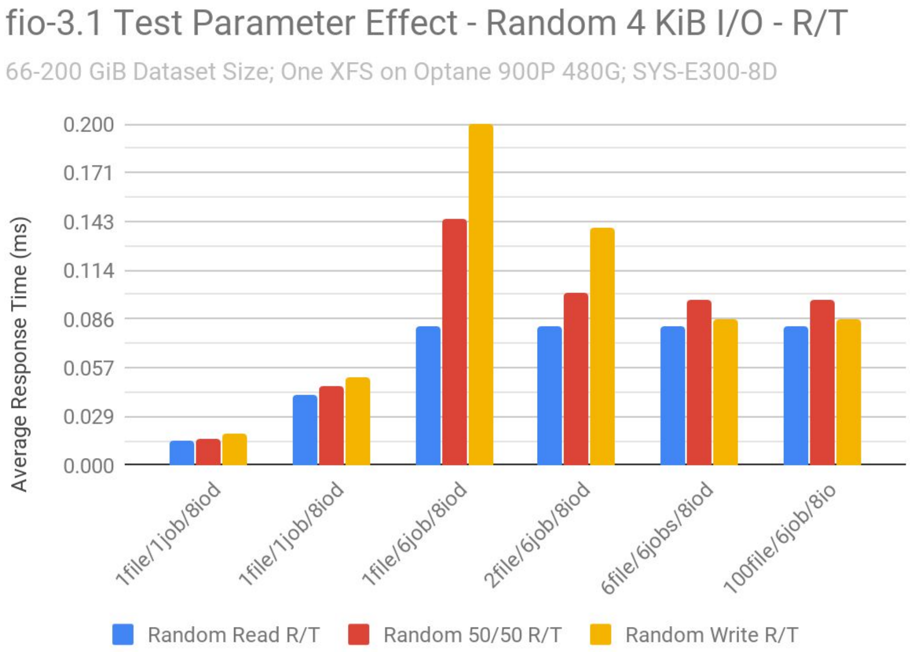
- Test multiple parameters before settling on a methodology
- Seemingly small things like file count can have a large effect
- Queue depth and an async I/O engine are important



iod = I/O depth, a.k.a. queue depth

You Get What You Measure: Workload Parameter Variation

- Don’t forget about response time!
- High write latencies as queue length stacks up against a single file
  - Uncovers a bottleneck in OS, file system, benchmark tool, etc.
  - Solution: use multiple files!



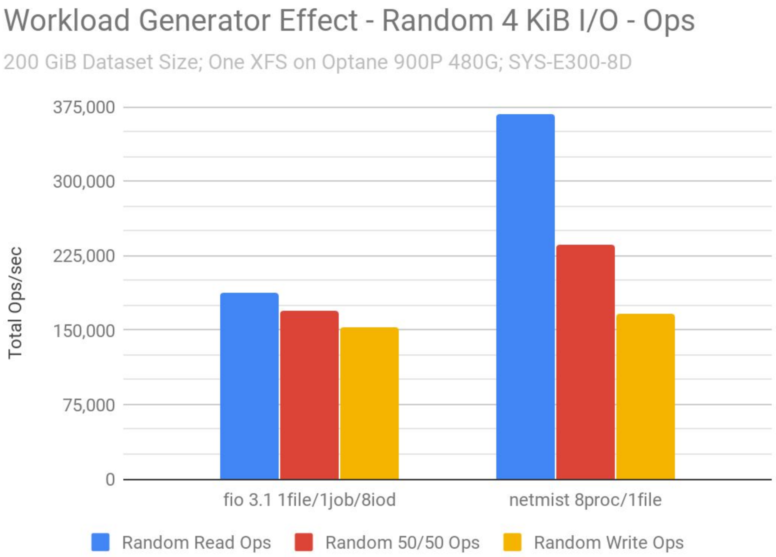
fio-3.1 Test Parameter Effect - Random 4 KiB I/O - R/T

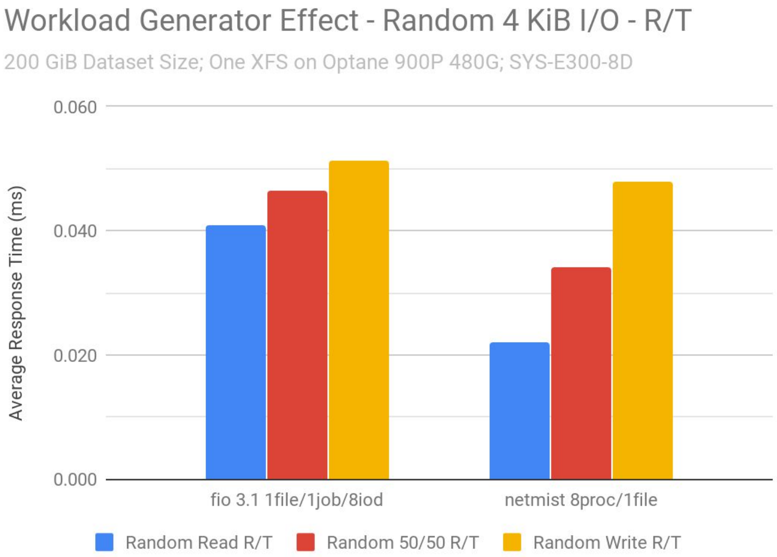
66-200 GiB Dataset Size; One XFS on Optane 900P 480G; SYS-E300-8D

iod = I/O depth, a.k.a. queue depth

You Get What You Measure: Load Generator Variation

- Different load generators can give different answers
- May converge with tweaks to test parameters





iod = I/O depth, a.k.a. queue depth

You Get What You Measure: Environmental Design Rules

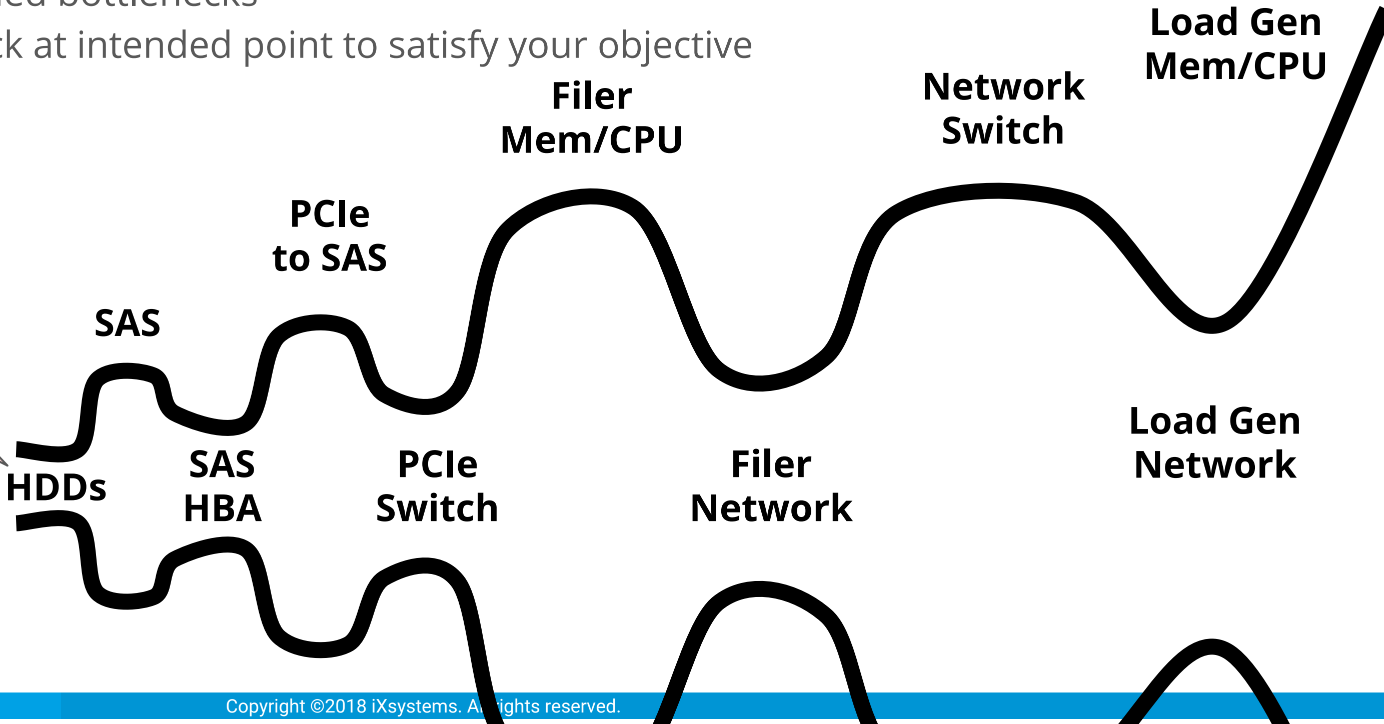
- Load Generating Clients
    - Using VMs? It can work!
        - Disable Hyperthreading
        - Total vCPUs <= Total Real Cores
        - Total vMem < Total Phys. Mem
    - Total network bandwidth of all clients > Total network bandwidth of filer
    - Avoid LAGs
    - Be aware of memory
        - Too much can hurt or help, depending on
            - Workload
            - Goals of testing

- Network
    - Avoid switch hops, but if you must...
        - Ensure sufficient ISL bandwidth
    - Consistent MTU - don't fragment!
- Filer
    - Ensure sufficient network and storage bandwidth
    - Be aware of NIC/HBA controller limits
        - Dual-port 100GbE != 200Gb
        - PCIe speed and width limits
        - SAS expander oversubscription
- Beware PCIe switches!
    - Always check server block diagrams

You Get What You Measure: Bottleneck Placement

- Testing parameters and the environment must be designed carefully
  - Avoid unintended bottlenecks
  - Place bottleneck at intended point to satisfy your objective

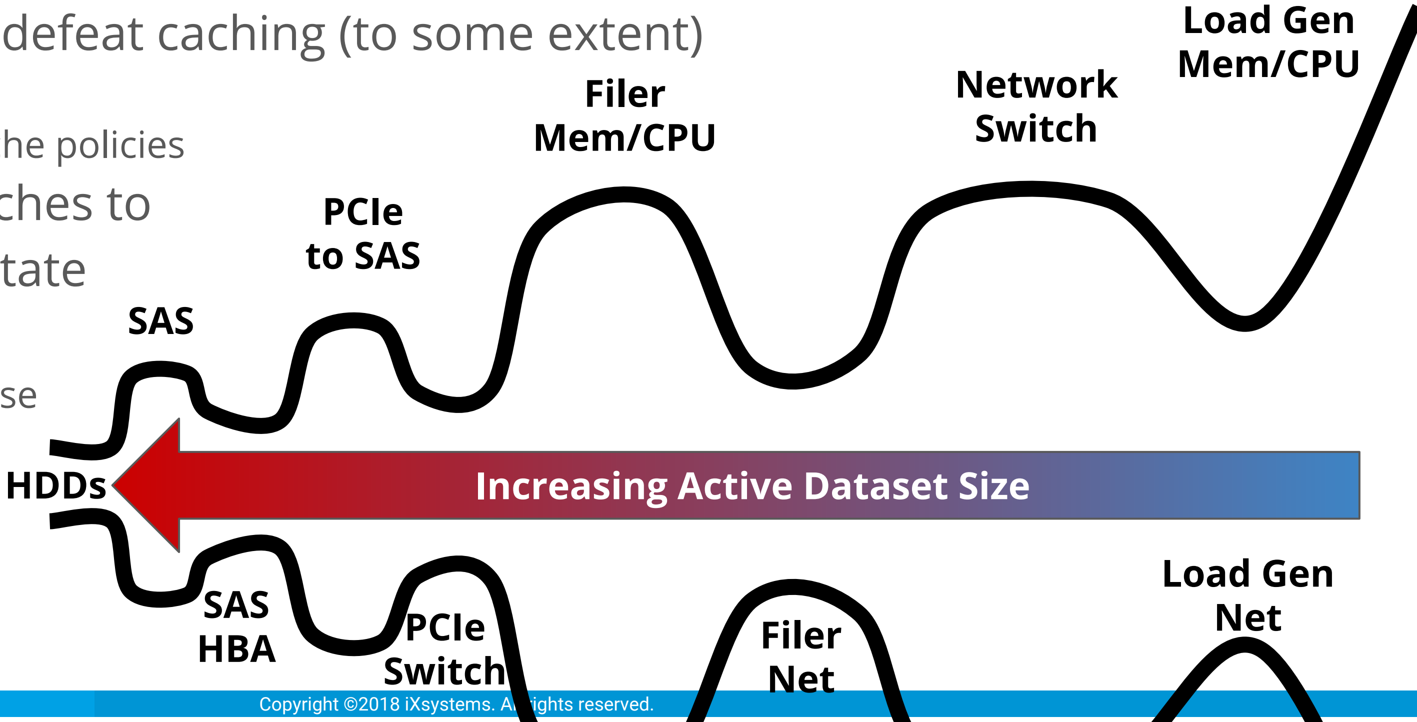
This solution is designed to bottleneck on the performance of its HDDs



Storage systems make strangely shaped bottles

You Get What You Measure: Active Dataset Size

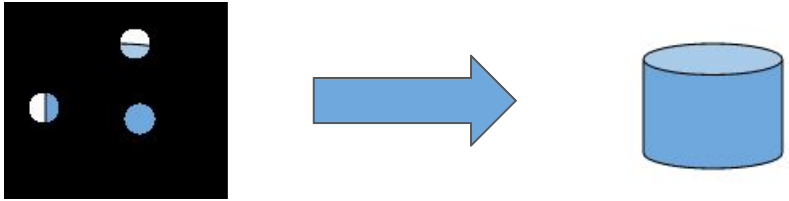
- Increasing active dataset size defeats caching
  - Moves I/O deeper into storage solution
- Other ways to defeat caching (to some extent)
  - Direct I/O
  - Changing cache policies
- Must warm caches to reach steady-state condition
  - Warmup Phase
  - Ramp Time
  - Pre-conditioning



So… Why Does “It Depend”?

- Workloads used to characterize systems are an approximation of reality
  - Load generation tools vary in fidelity and behavior
- Environments for testing don’t exactly reflect production environments
  - Must be carefully designed to place bottlenecks at desired target
- Active dataset size is one of the biggest drivers of performance variability
  - Affected by compression and dedupe if data is reducible
  - Both active dataset size and data reducibility are generally difficult to measure and not well known
- There is a desire for “the number” - or perhaps up to three numbers
  - Performance is a shape, not a curve or a point

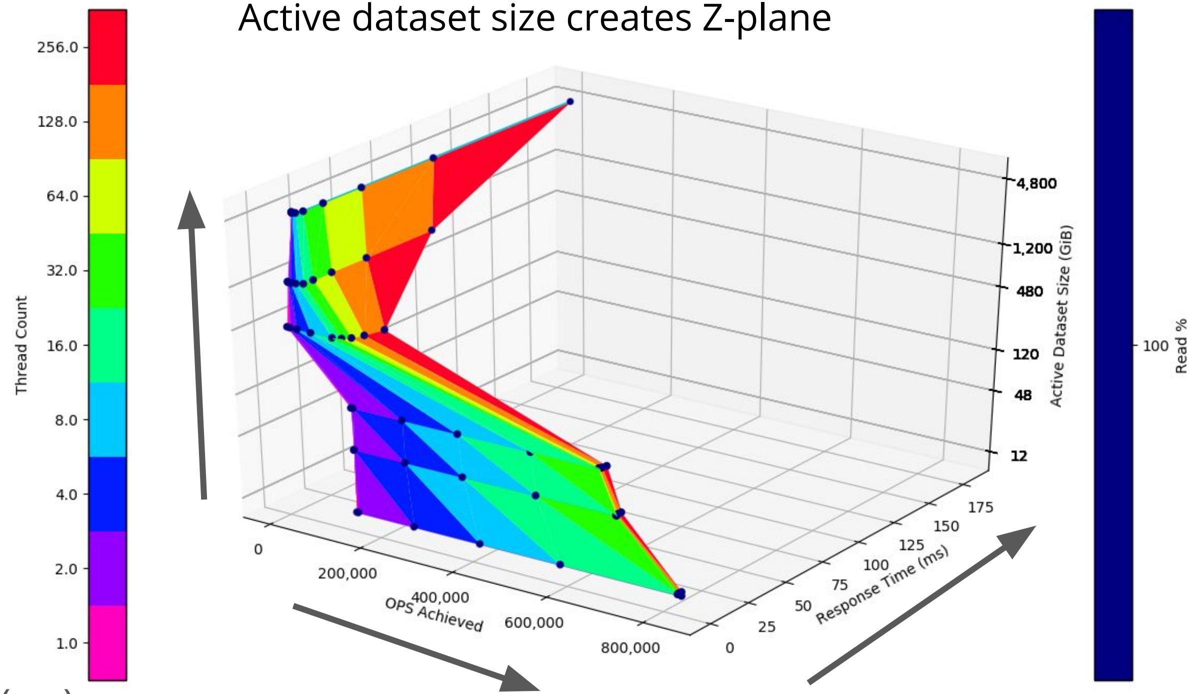
What shape is this?



Performance is a Shape

Shout out to Ryan McKenzie @ iXsystems
for data and 3D visualizations

- Random 4k reads
- OpenZFS storage
  - 256GB RAM
  - 1.6TB L2ARC
  - 142 HDDs (mirrors)
- Scaling up by:
  - Thread count
  - Active dataset size
- Each combination of {thd_cnt,act_data_sz} provides both:
  - Achieved Ops/sec
  - Average Response Time (ms)

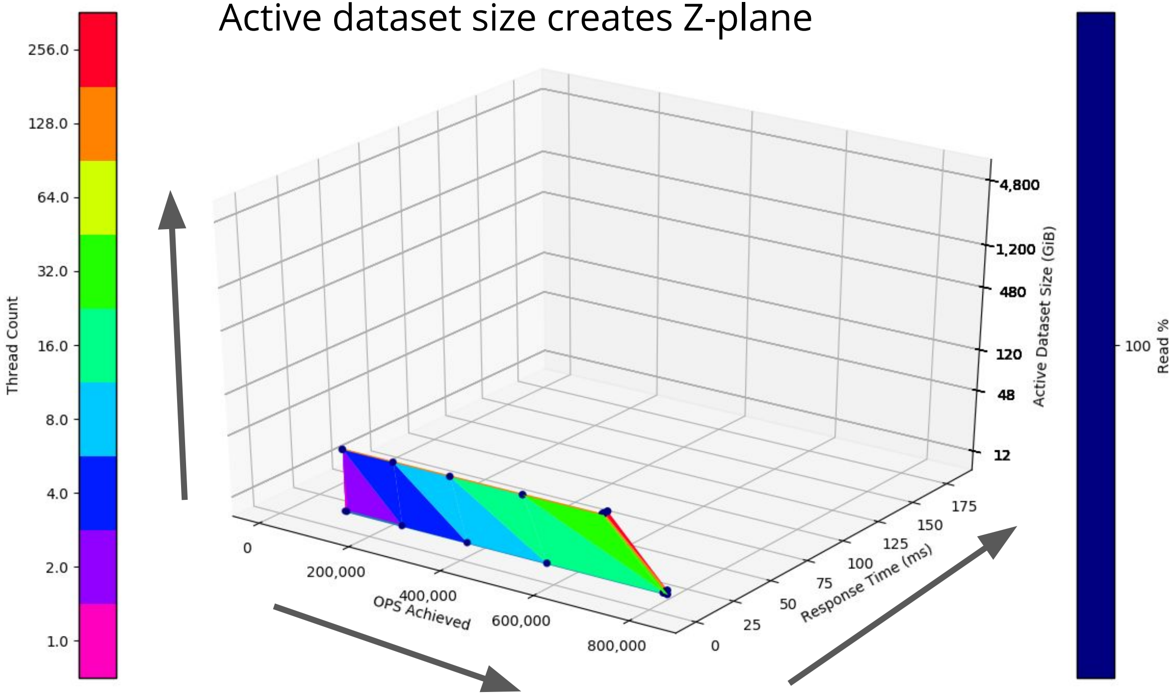


Active dataset size creates Z-plane

Thread count drives ops/sec and response time up in X-Y plane

Performance is a Shape

- Small active dataset size
  - Firmly in ARC hit zone



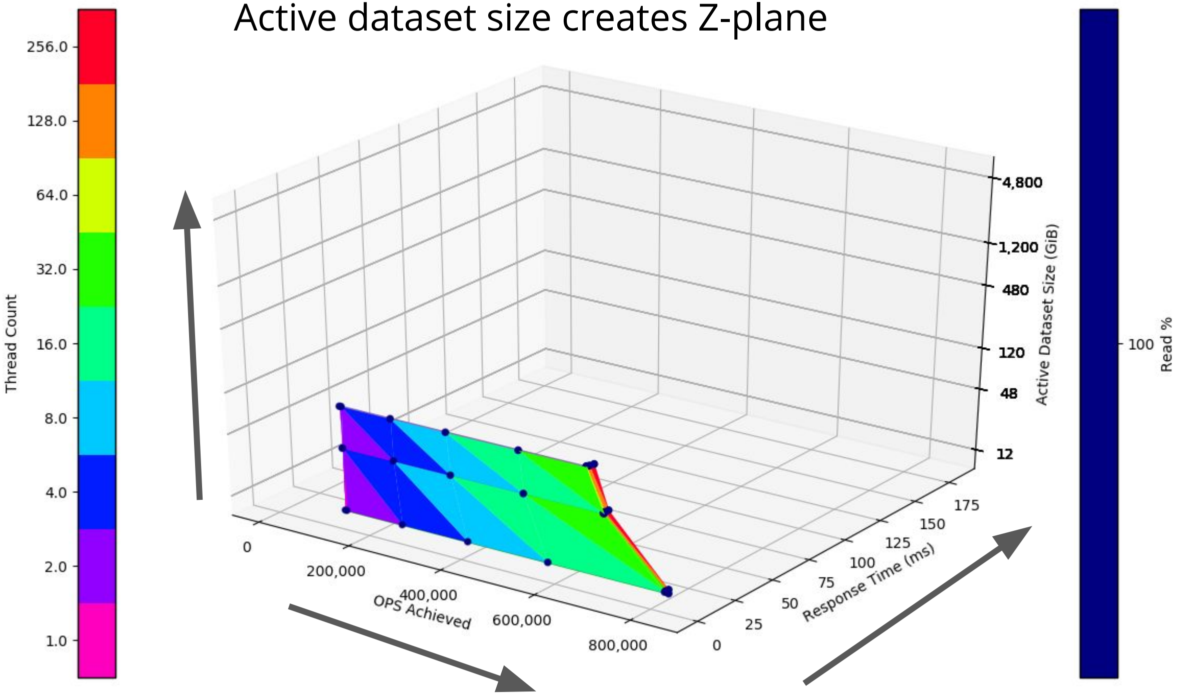
Active dataset size creates Z-plane

Thread count drives ops/sec and response time up in X-Y plane

Performance is a Shape

- Small active dataset size
  - Reaching end of ARC hit zone

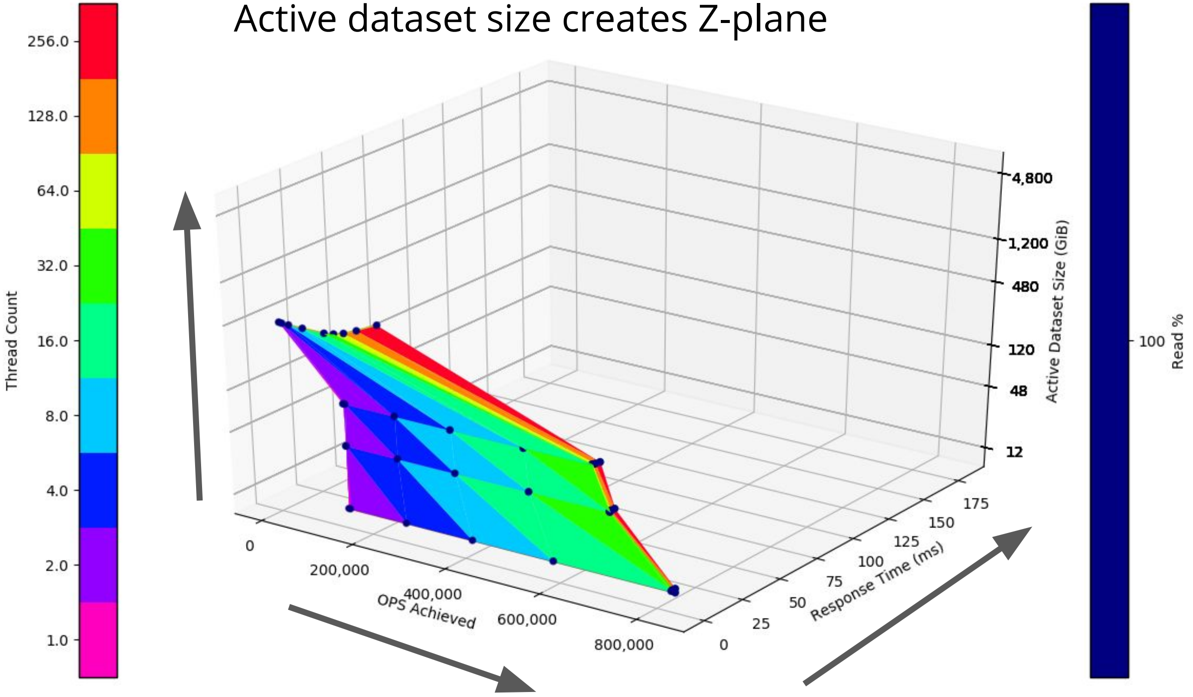
Active dataset size creates Z-plane



Thread count drives ops/sec and response time up in X-Y plane

Performance is a Shape

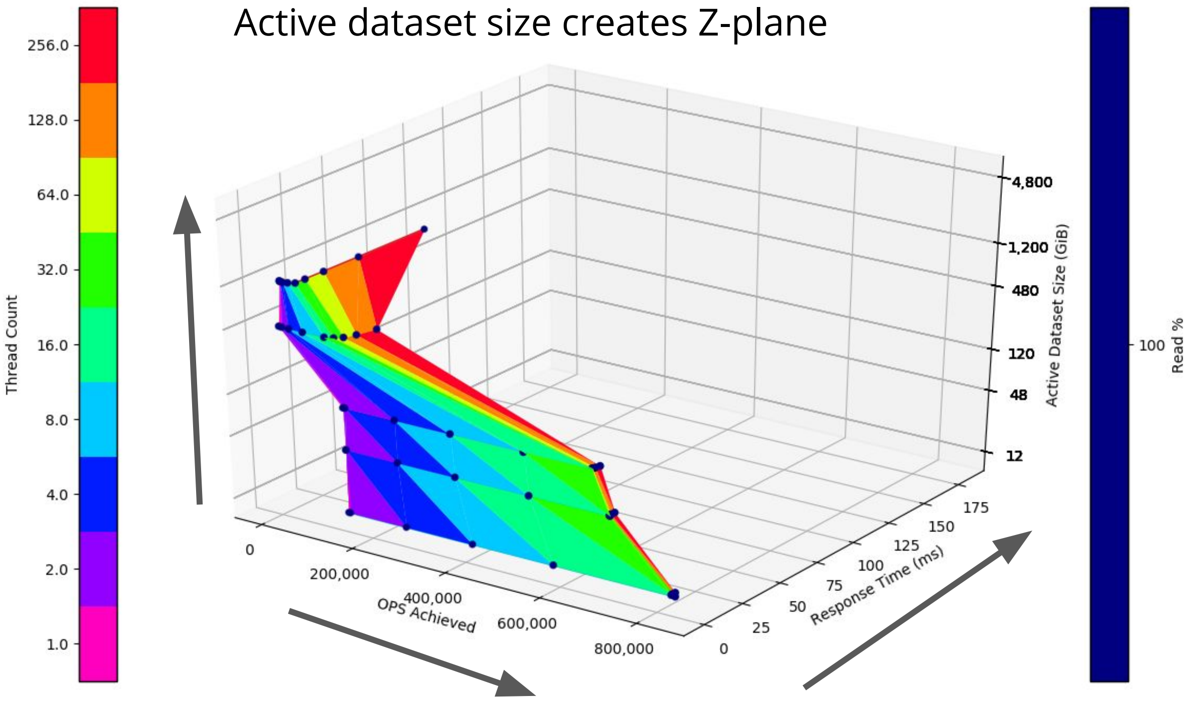
- Medium active dataset size
  - Transition from ARC hit to the L2ARC hit zone



Thread count drives ops/sec and response time up in X-Y plane

Performance is a Shape

- Large active dataset size
  - Starting to exit L2ARC hit zone
  - Pushing bottleneck down to drives

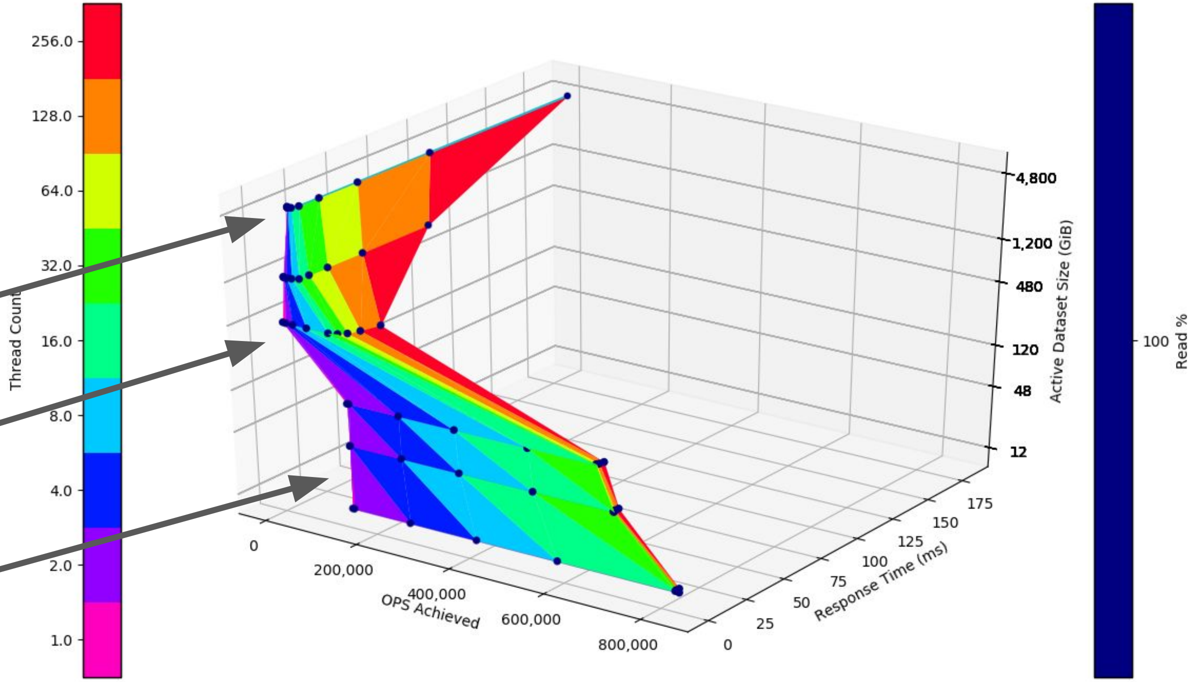


Active dataset size creates Z-plane

Thread count drives ops/sec and response time up in X-Y plane

Performance is a Shape

- As active dataset size changes, scaling up thread count has different behavior
  - ○ Increases latency
  - ○ Increases ops/sec and latency
  - ○ Increases ops/sec



Final Thoughts

- Next time you reach for dd to test performance, give fio a shot
- Experiment with different test parameters, like iodepth, and number of files before you decide on a methodology
- Architect your environment to place the bottleneck at the desired point
- Consider your active dataset size - perhaps test multiple sizes!
- Performance is a shape!
- Let’s work together to make better standardized workloads and tests
    - I suggest SPEC Open Systems Group’s Storage subcommittee

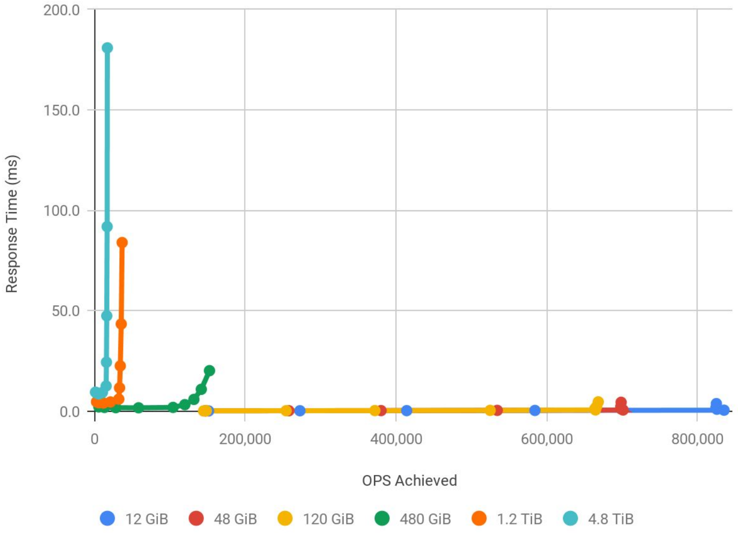
Thank You! Questions?

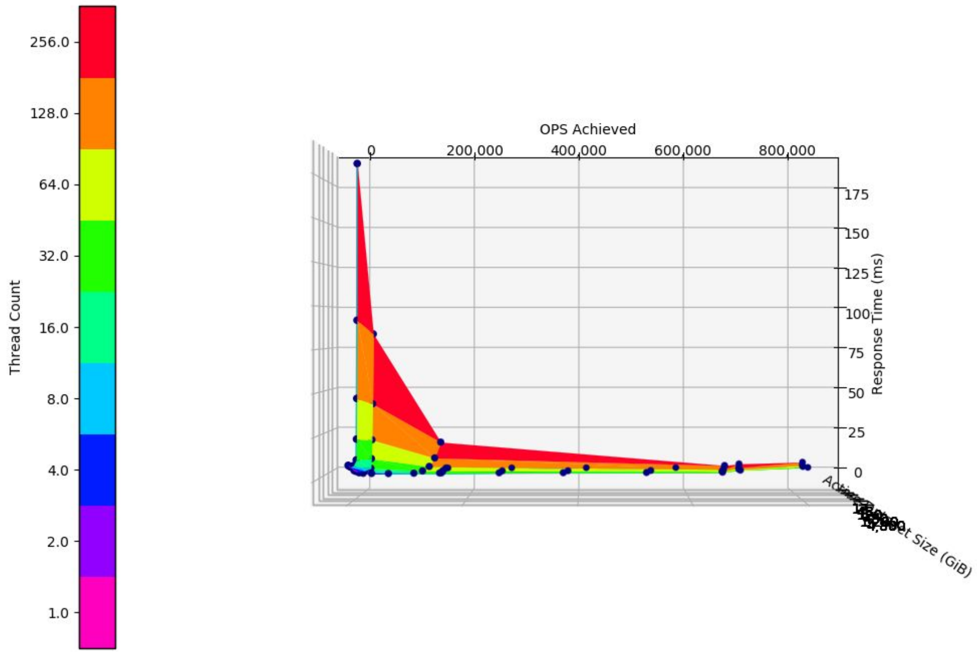
Nick Principe
iXsystems

Twitter:	@nickprincipe
Github:	@powernap
Email:	nap@ixsystems.com

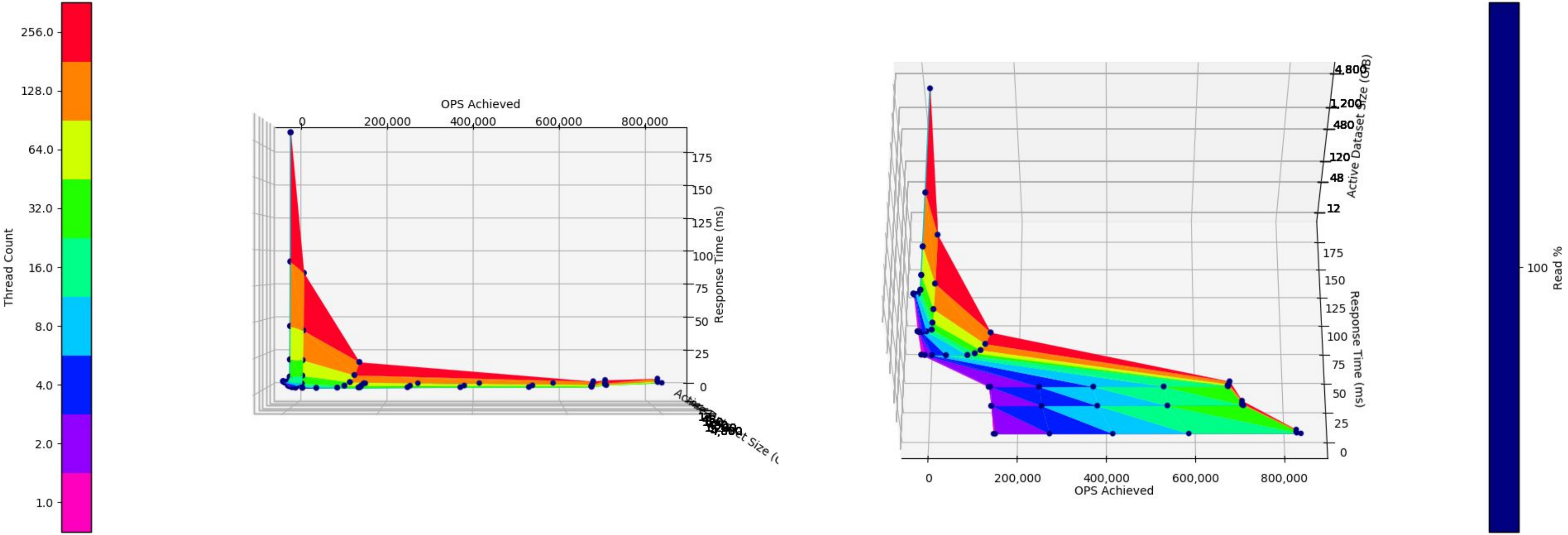
Backup Slides

Building the Shape





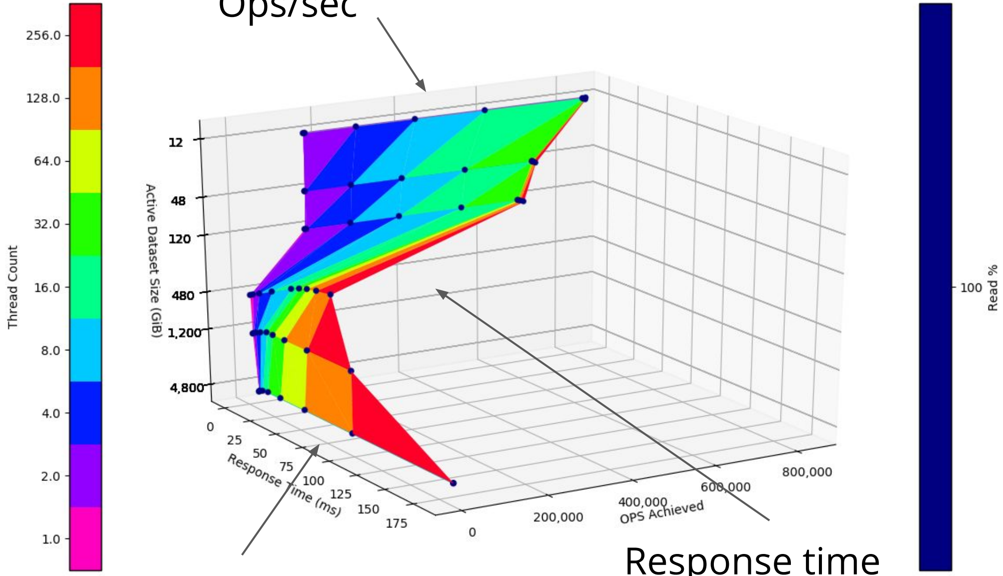
Building the Shape II



Rotate to reveal Z-axis

Some Different Views

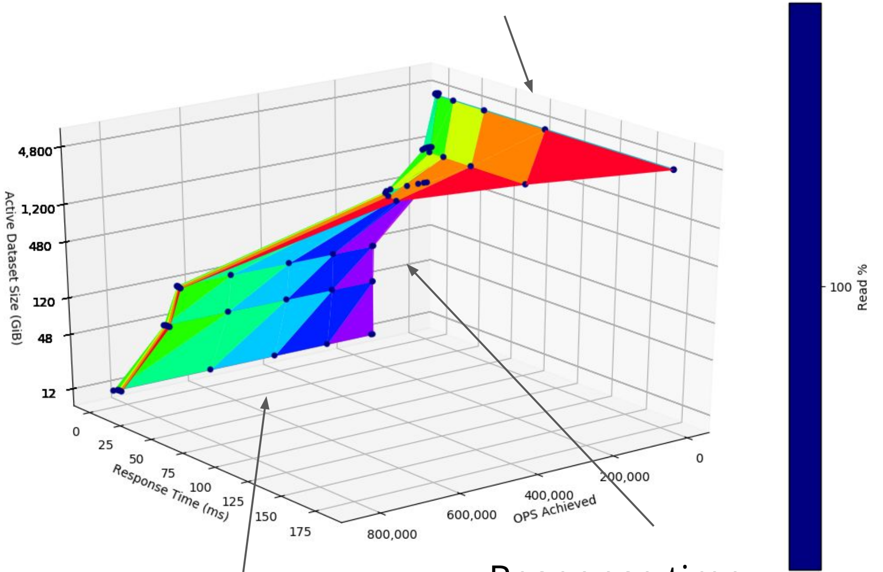
Adding threads increases...



Ops/sec

Response time

Response time and Ops/sec



Response time

Ops/sec

Response time and Ops/sec