SDC

STORAGE DEVELOPER CONFERENCE

SNIA SANTA CLARA, 2017

Fine-Grained Latency Measurement

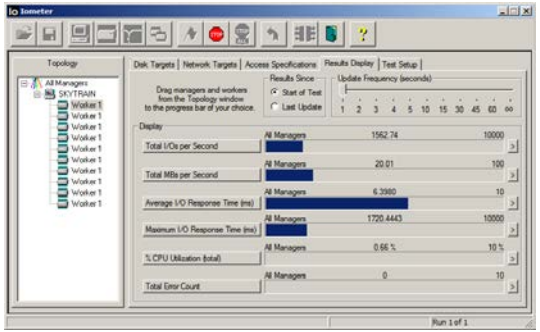
David J. Cuddihy
ATTO Technology, Inc.

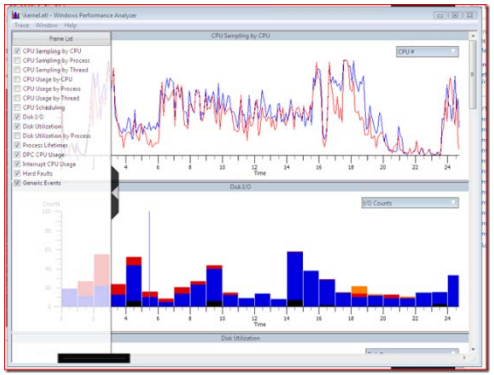
Performance Measuring Tools (Block Storage)


```
Starting 4 processes
Jobs: 4 (f=4): [WWWW] [100.0% done] [0K/174.3M /s] [0 /340K iops] [eta 00m:00s]
writeiops: (groupid=0, jobs=4): err= 0: pid=3289
  write: io=5074.9MB, bw=173216KB/s, iops=346427 , runt= 30001msec
    slat (usec): min=41 , max=8527 , avg=139.96, stdev=42.22
    clat (usec): min=2 , max=8618 , avg=221.28, stdev=88.00
     lat (usec): min=104 , max=8737 , avg=360.29, stdev=88.22
    clat percentiles (usec):
     |  1.00th=[  113],  5.00th=[  133], 10.00th=[  143], 20.00th=[  173],
     | 30.00th=[  179], 40.00th=[  185], 50.00th=[  191], 60.00th=[  199],
     | 70.00th=[  217], 80.00th=[  294], 90.00th=[  350], 95.00th=[  414],
     | 99.00th=[  462], 99.50th=[  470], 99.90th=[  510], 99.95th=[  580],
     | 99.99th=[ 1448]
    bw (KB/s)  : min=    1, max=60400, per=24.55%, avg=42527.75, stdev=8020.48
    lat (usec) : 4=0.01%, 10=0.01%, 20=0.01%, 50=0.01%, 100=0.04%
    lat (usec) : 250=76.16%, 500=23.69%, 750=0.07%, 1000=0.01%
    lat (msec) : 2=0.03%, 4=0.01%, 10=0.01%
  cpu          : usr=7.46%, sys=69.82%, ctx=10920620, majf=0, minf=117
  IO depths    : 1=0.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.1%, 32=117.1%, >=64=0.0%
     submit    : 0=0.0%, 4=0.0%, 8=0.0%, 16=100.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=0.0%, 8=0.0%, 16=100.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     issued    : total=r=0/w=10393184/d=0, short=r=0/w=0/d=0

Run status group 0 (all jobs):
  WRITE: io=5074.9MB, aggrb=173216KB/s, minb=173216KB/s, maxb=173216KB/s, mint=30001msec, maxt=30001msec

Disk stats (read/write):
  fiob: ios=12/12139609, merge=0/0, ticks=0/1671657, in_queue=1667116, util=99.78%
```





How Benchmarks Lie

- Report minimum/maximum/average
  - IOPS or MB/s
- Bimodal systems
  - Fast at times, slow at others
  - Min/Max/Average is a poor representation
- Multiple initiator aggregates
  - Averaging multi-server systems skews numbers further

Benchmarks In a RAID System

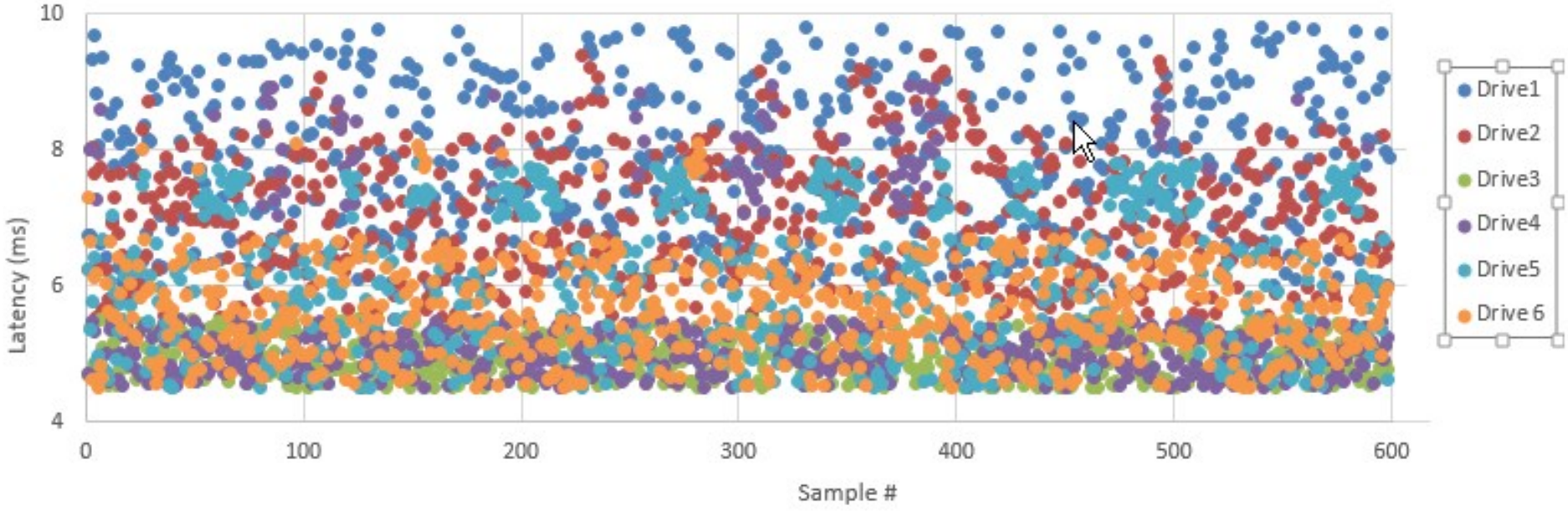
- Cache decisions create noise
  - Direct read vs read from cache (bimodal)
  - Early status vs wait-for-completion on writes
- Rebuilds create noise
  - Smaller system I/Os can be stuck behind large reads/writes used for rebuilds (drive prioritization)
- Noisy systems are harder to analyze
  - Averages hide smaller signals in the noise

Measure Latency Alongside IOPS

- Software tools provide latency measurement
  - Linux FIO – include latency bar graph
  - xPerf – disk latency output
- Latency measurements give visibility
  - Spot bimodal systems
  - Weed out hiccups

IOPS vs. Latency

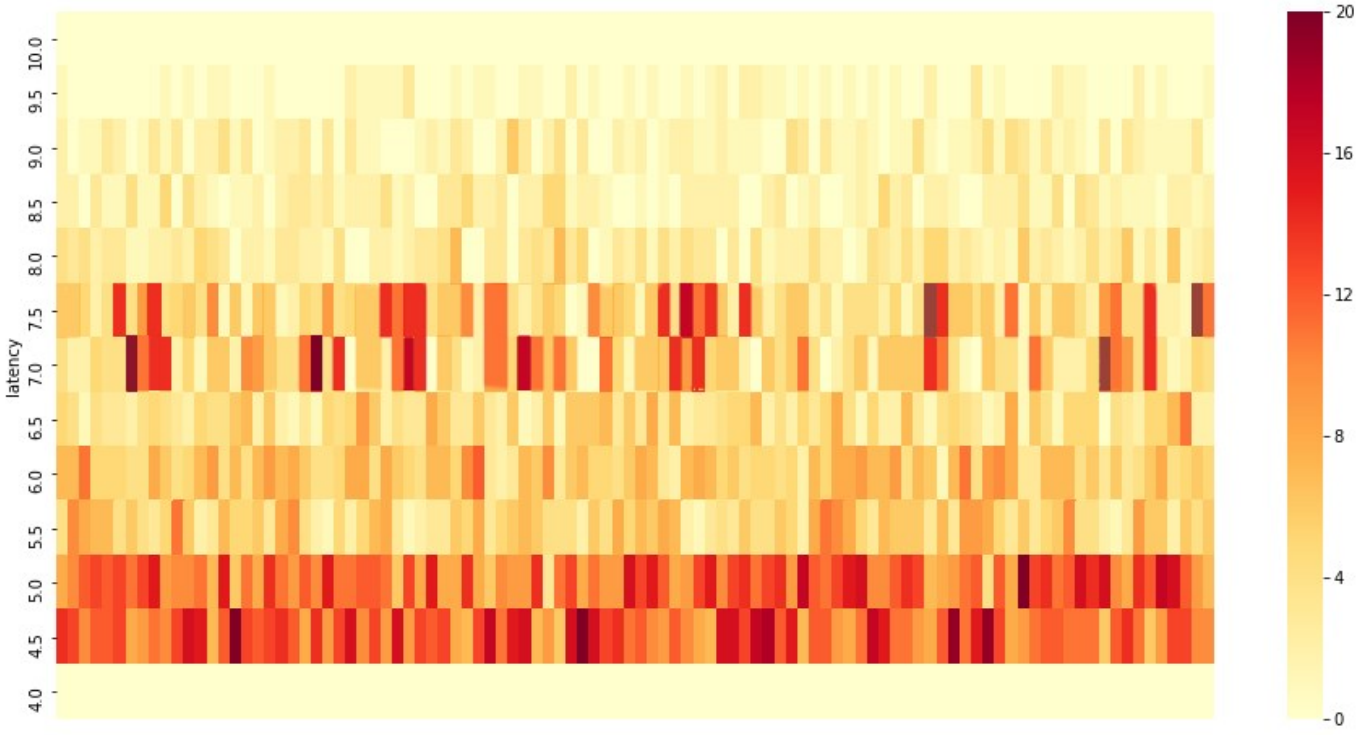
6 HDD
Random 4K Read
15K IOPS



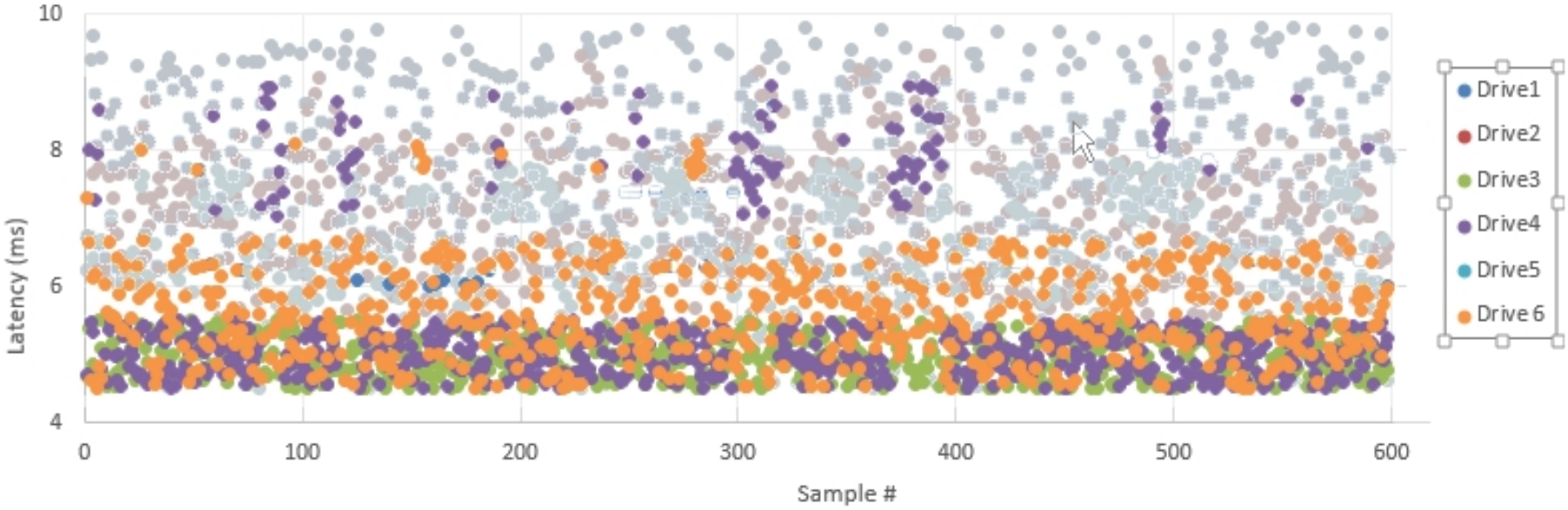
A Better Visualization: Latency Heat Map

- Better than scatter plot due to time compression
  - Colors show recurrence within a sample period
- 3-dimensional:
  - X & Y
  - Color intensity is the third dimension
- Credits to Bryan Cantrill & Brendan Gregg
  - Sun Microsystems ZFS storage appliance, 2007
  - CACM July 2010 : ‘Visualizing System Latency”
  - www.brendangregg.com

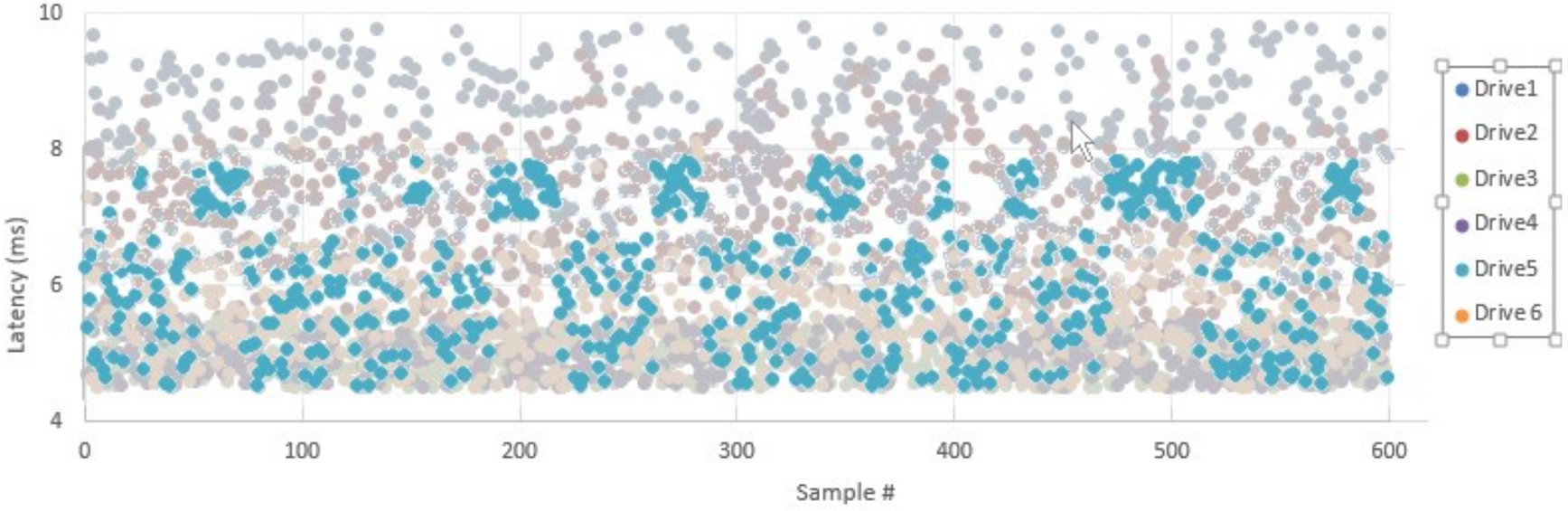
Heat Map Example



Band at 4.5 to 5 ms



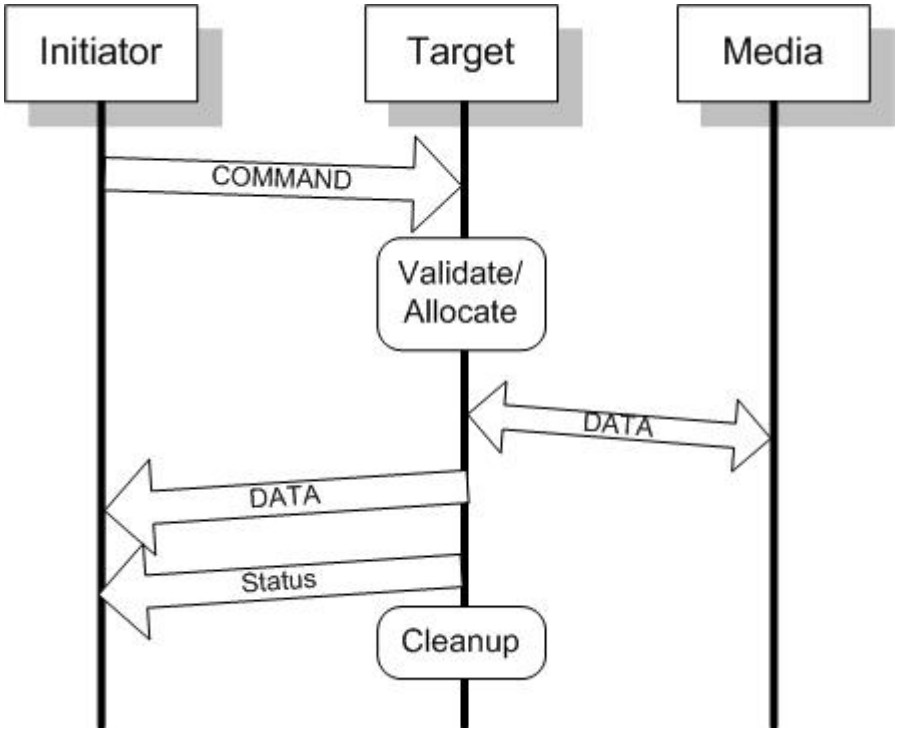
What's with Drive 5?



Storage Controller Overview

- Some storage controller basics
- Latency measurement system

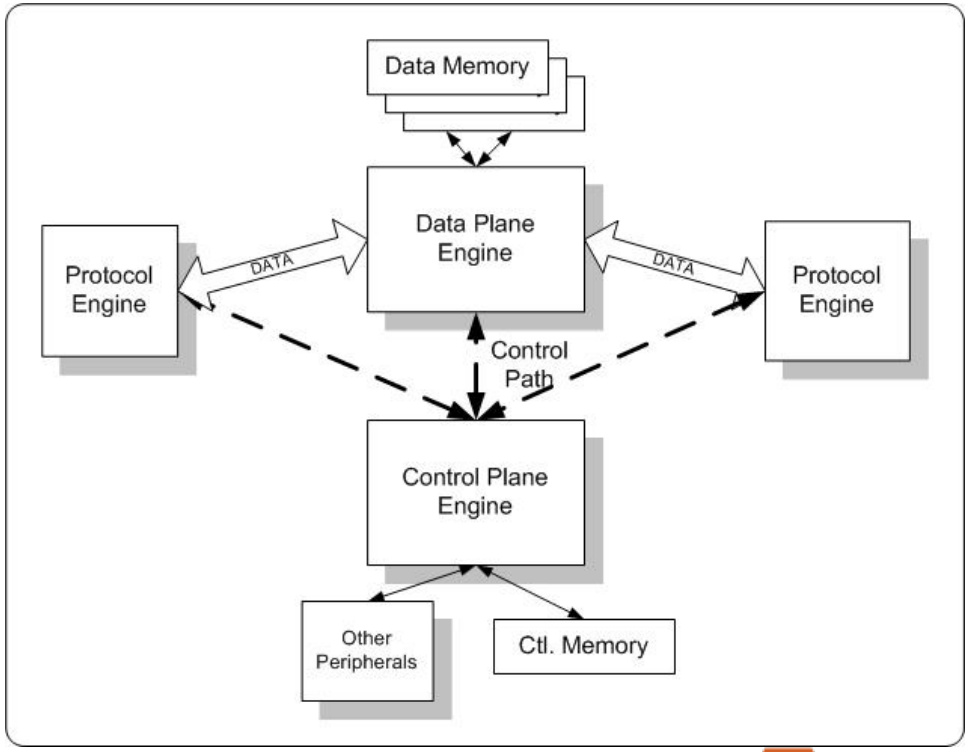
Anatomy of Storage Algorithm (Data In)



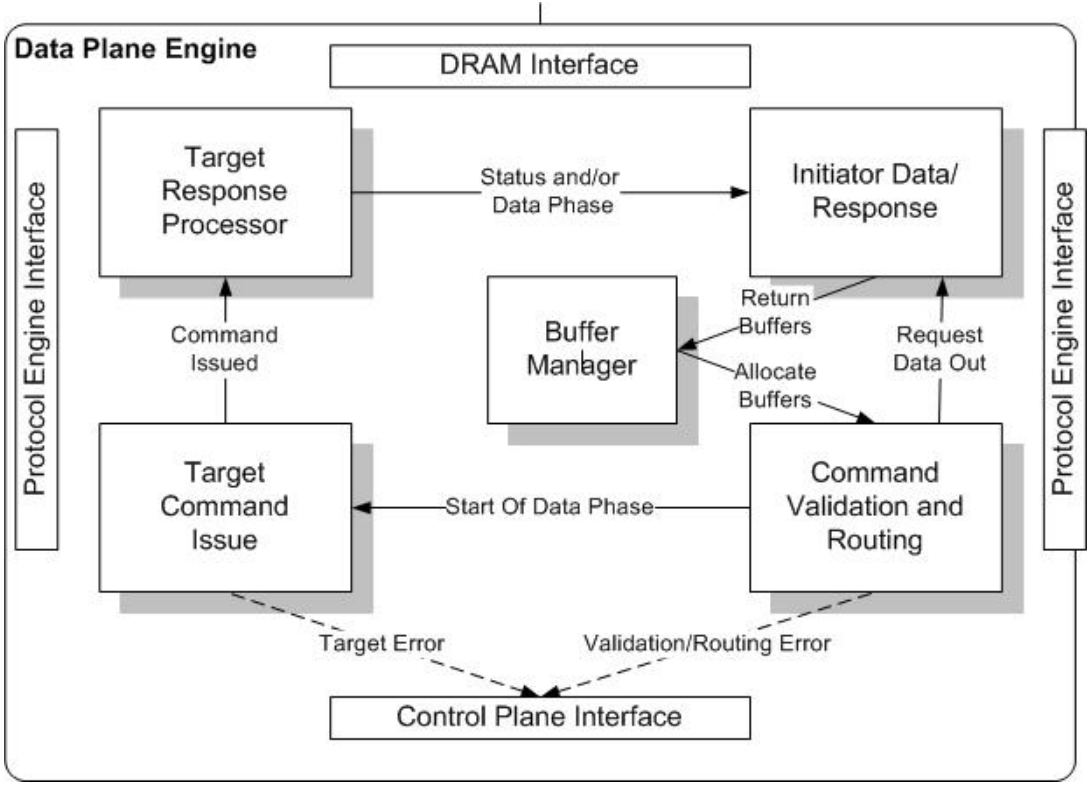
- Command Phase
  - Command validation
- Allocation Phase
  - Allocate command structures
  - Allocate buffer(s) for data
  - Create scatter/gather
- Data Phase
  - Data transfer
- Status Phase
- Completion/Cleanup
  - Return buffers/structures

Anatomy of a Storage Controller

- System on chip
- Protocol engine(s)
- Data plane engine
- Control plane engine



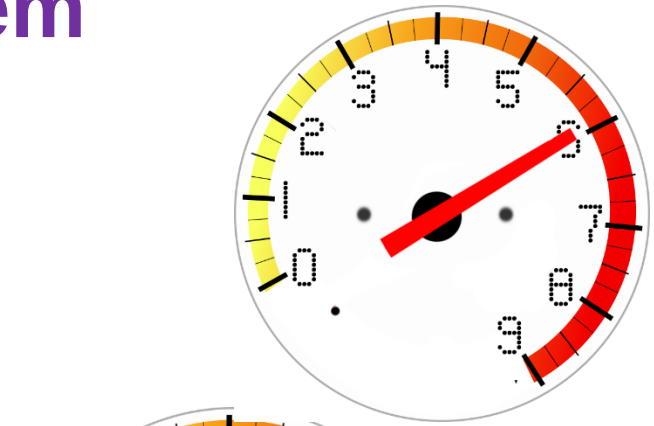
Data Plane Engine

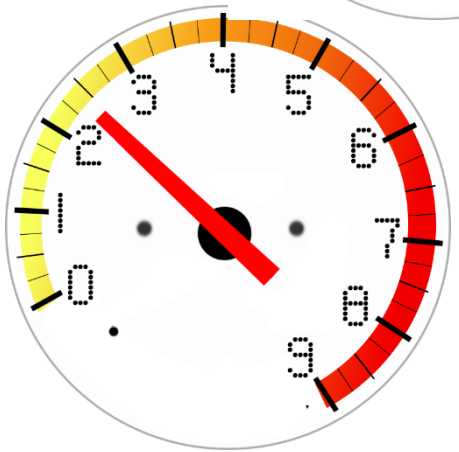


- Hardware engines
  - Handle each stage of transfer
- NonData commands handed to control plane
- Error states handed to control plane

Accelerating Storage System

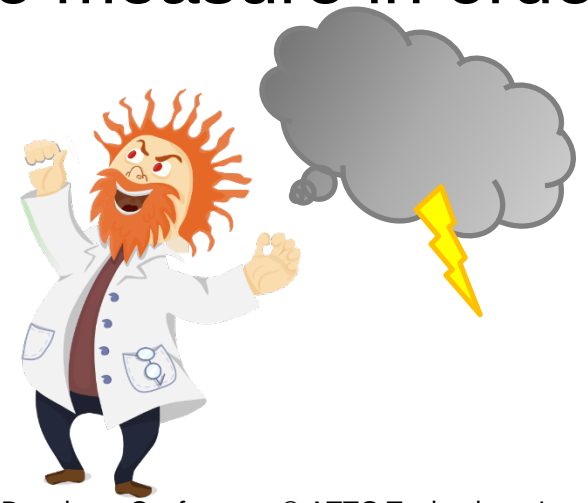
- Use fine grained latency measurement
- Transfer split into subcomponents
- One component can throw latency out of whack
- Measure to optimize system performance
  - Iterative
  - Identify bottleneck; mitigate;
  - lather-rinse-repeat





What Does This Have To Do With Latency?

- Control plane processor gathers statistics
- Data plane processor is opaque
  - We need to measure in order to optimize

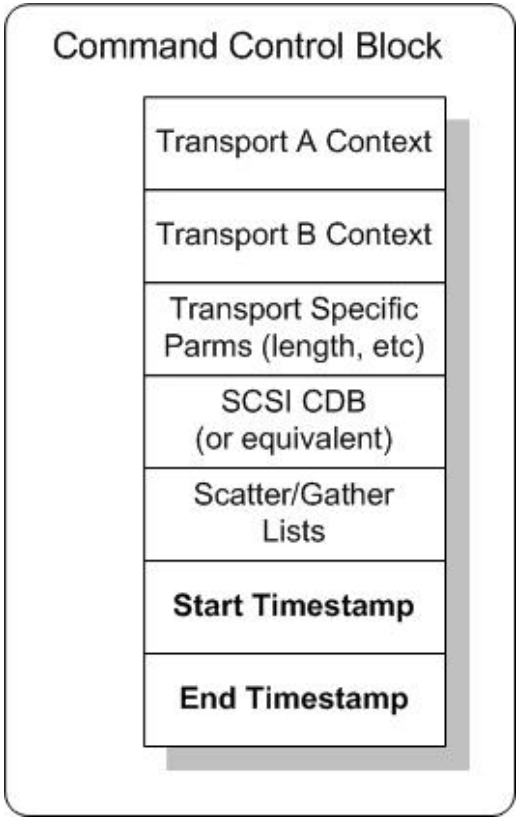


Data Plane Measurement - Timestamps

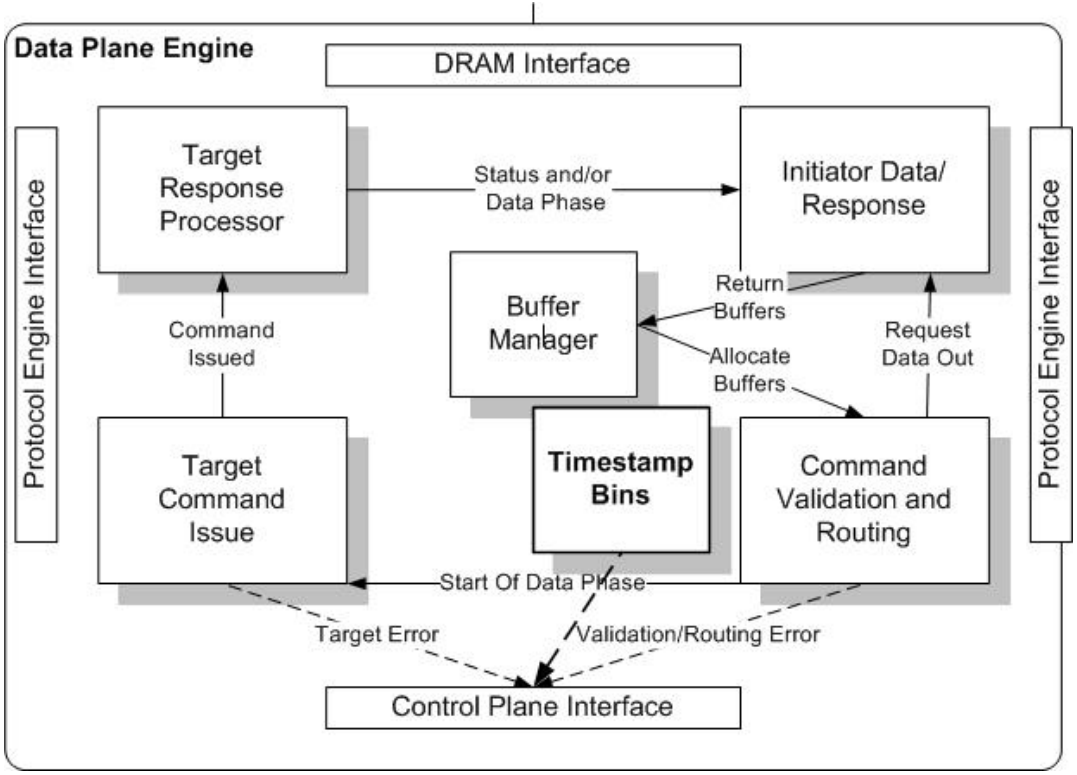
- Two timestamps in each command context
- Configure engines for latency window
  - Specify start and end of latency measurement
- Start engine logs timestamp
- End engine logs timestamp
- Cleanup engine calc's delta and bins result

Command Timestamps

- Start and end timestamp added to control block
- Engines configured to be at start, middle, or end
- Timestamps are binned at command completion
  - 8 Programmable bins
  - 1 us/20 bit resolution



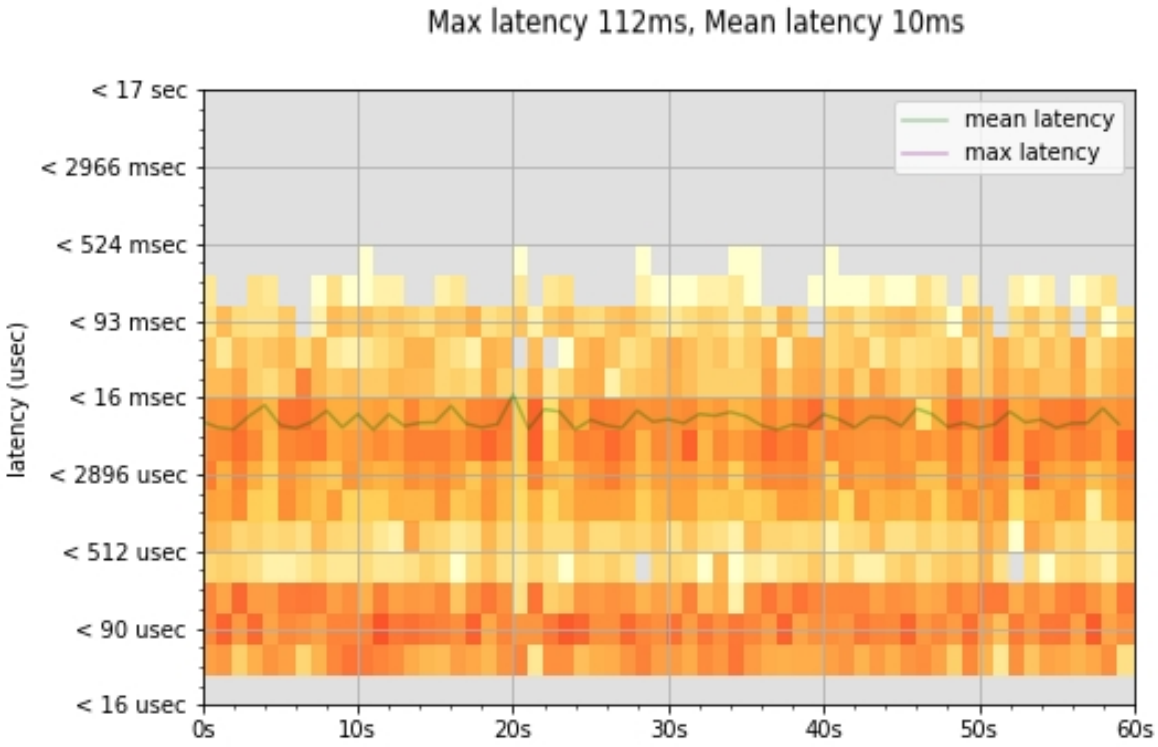
Command Timestamps: Bin Component

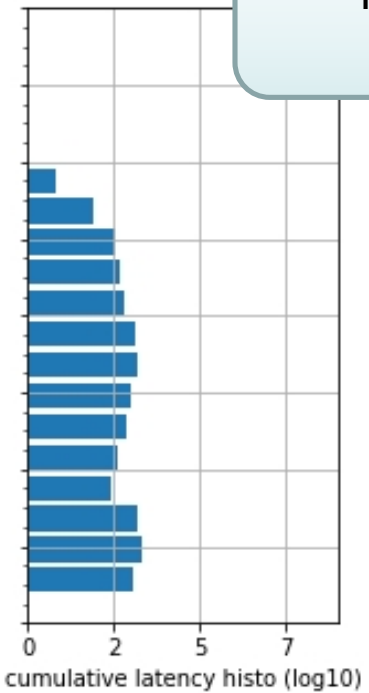


How Did We Get Here?

- Latency measurements
- Heat maps
- Data plane measurements

Example: Bimodal System

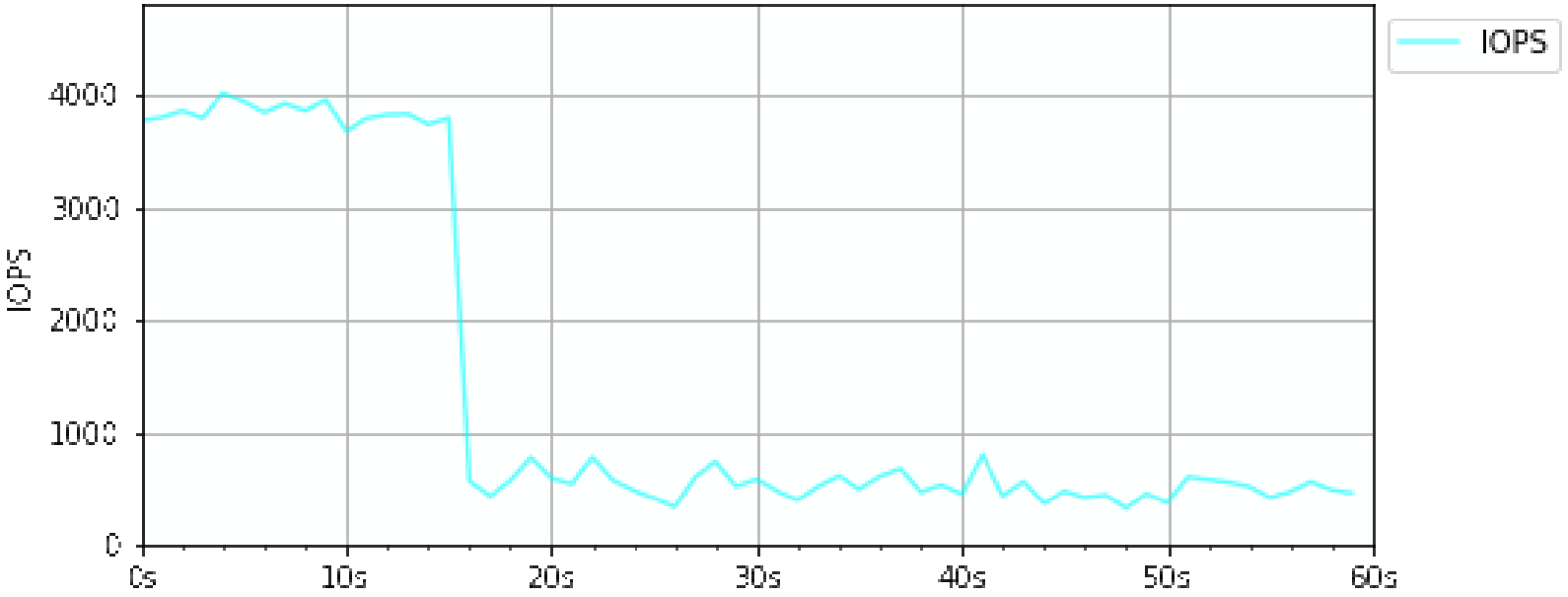




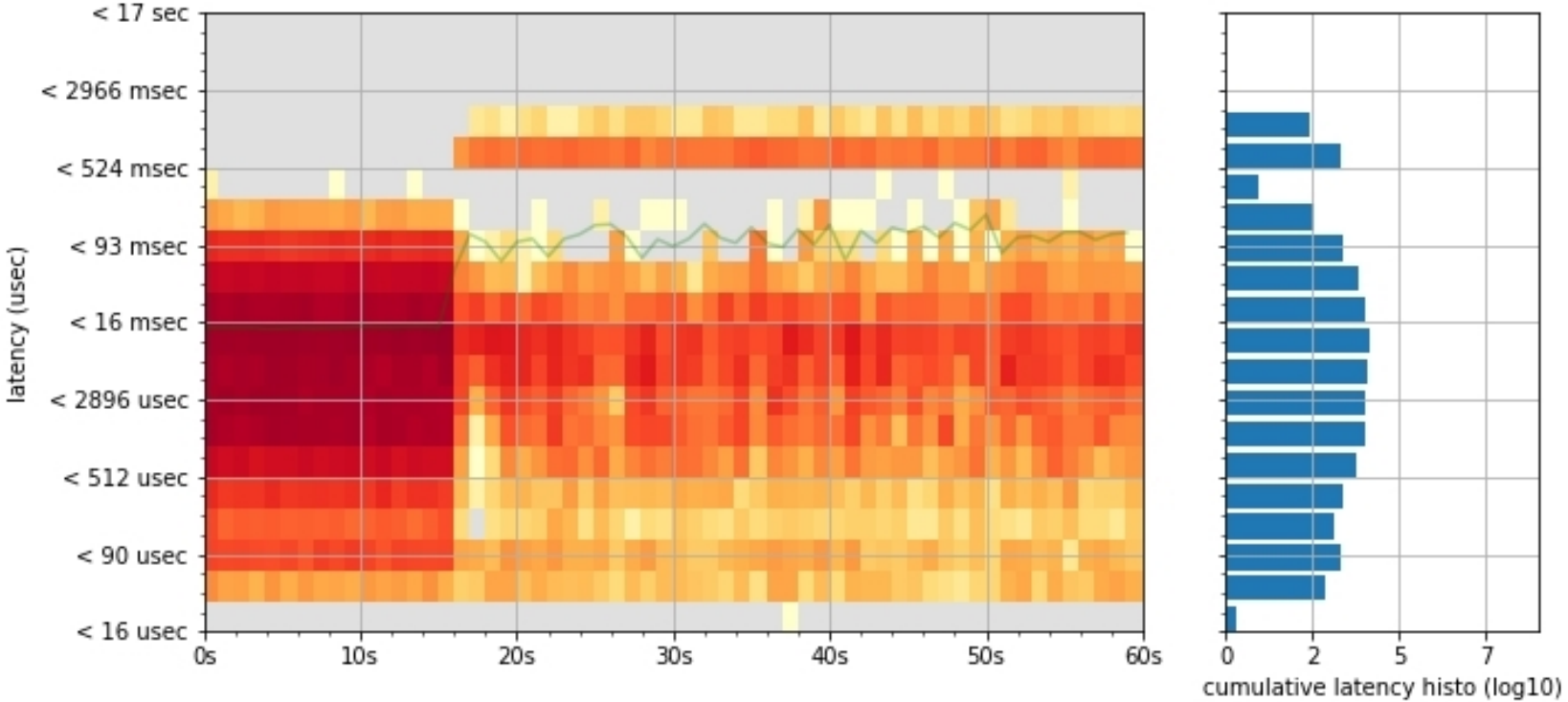
RAID 5
4K Writes
QD 64

Example: IOPS Shelf

RAID 5
15 Drives
4K Random Reads
QD 64

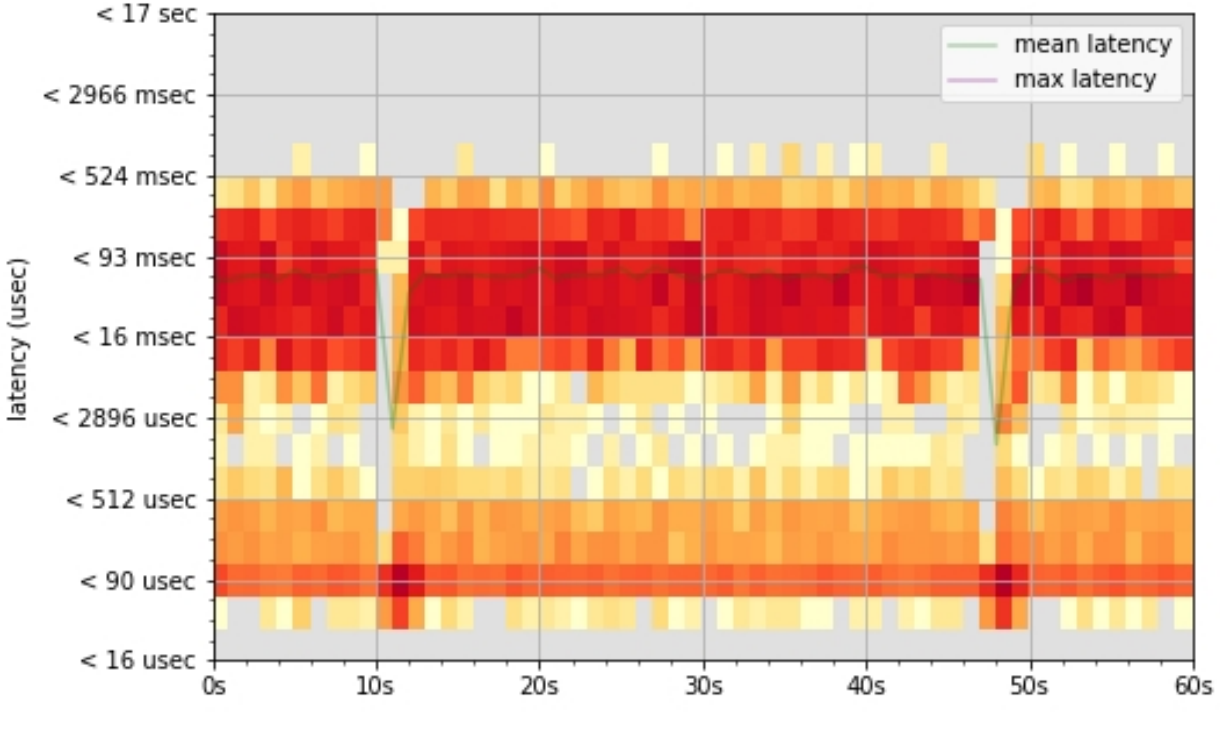


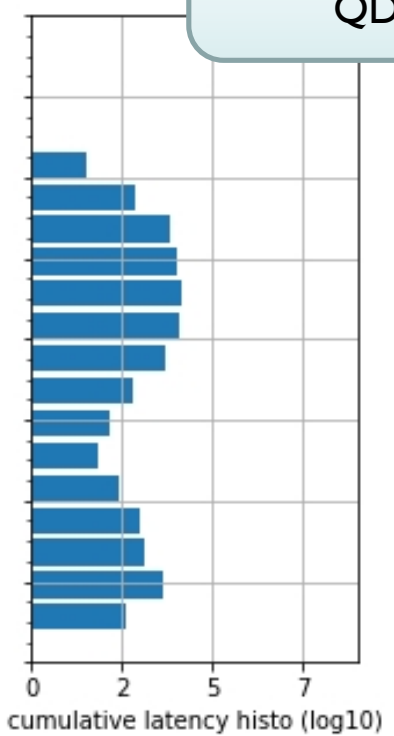
Heat Map : IOPS Shelf (Single Drive Fail)



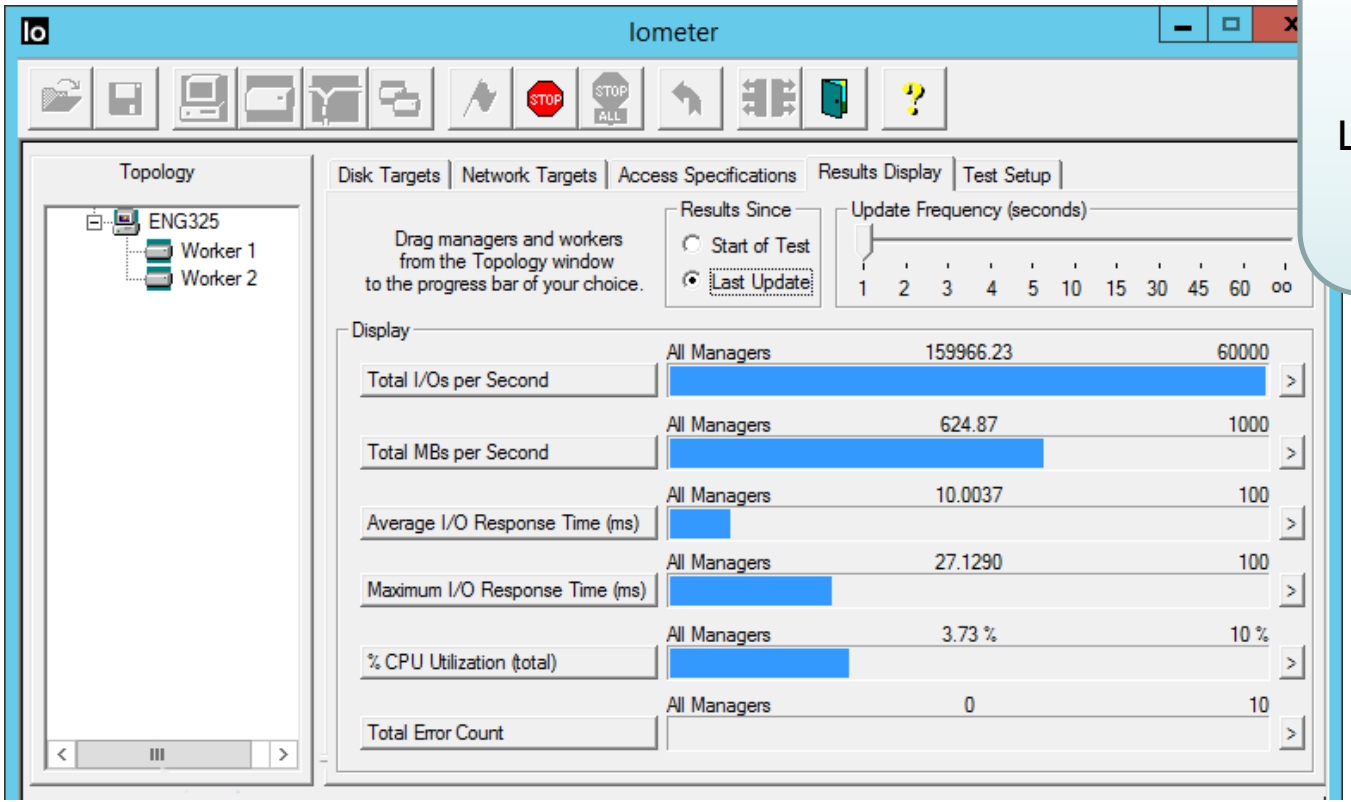
Example: Lun Reset

RAID 5
4K Writes
QD 64



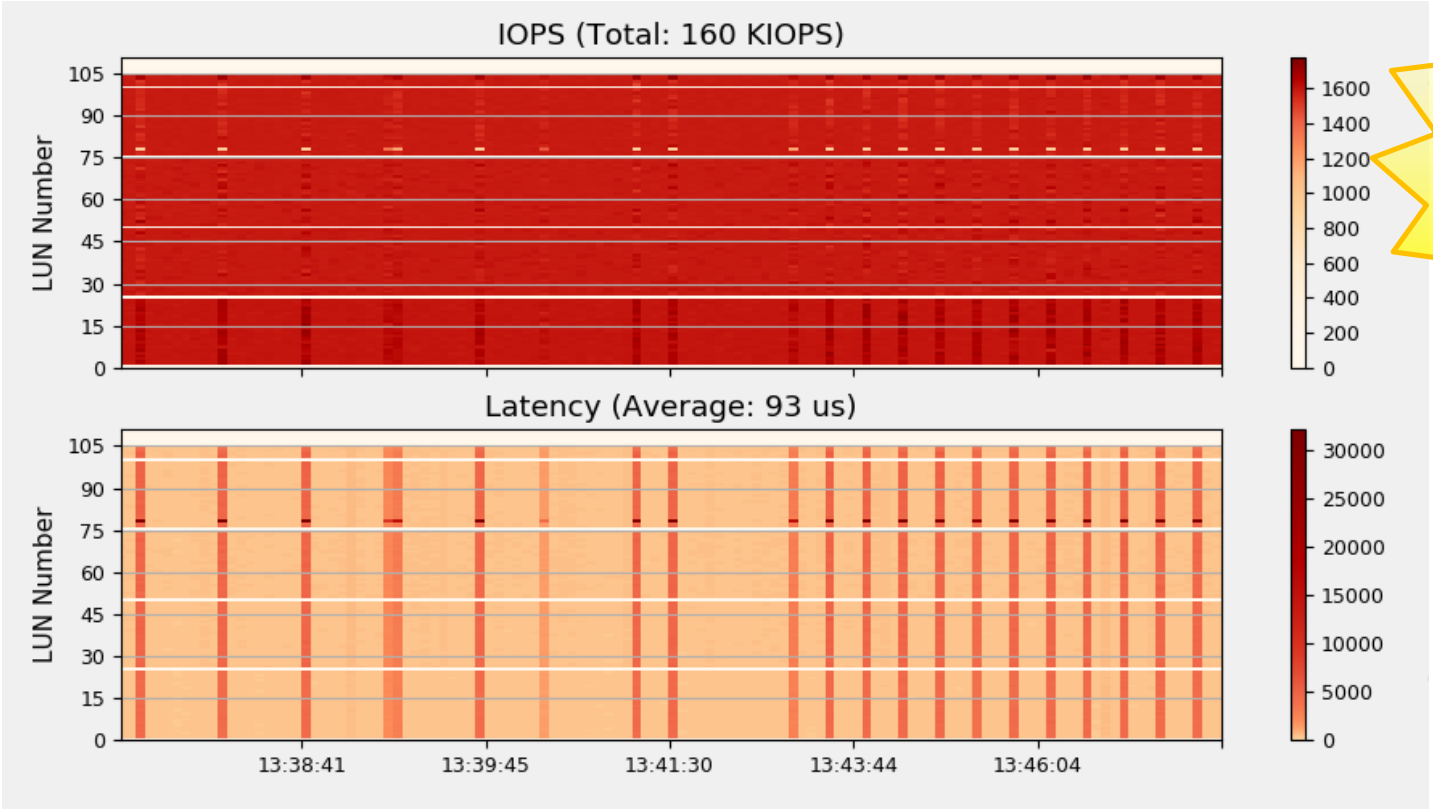


105 Drive IOMETER Workload



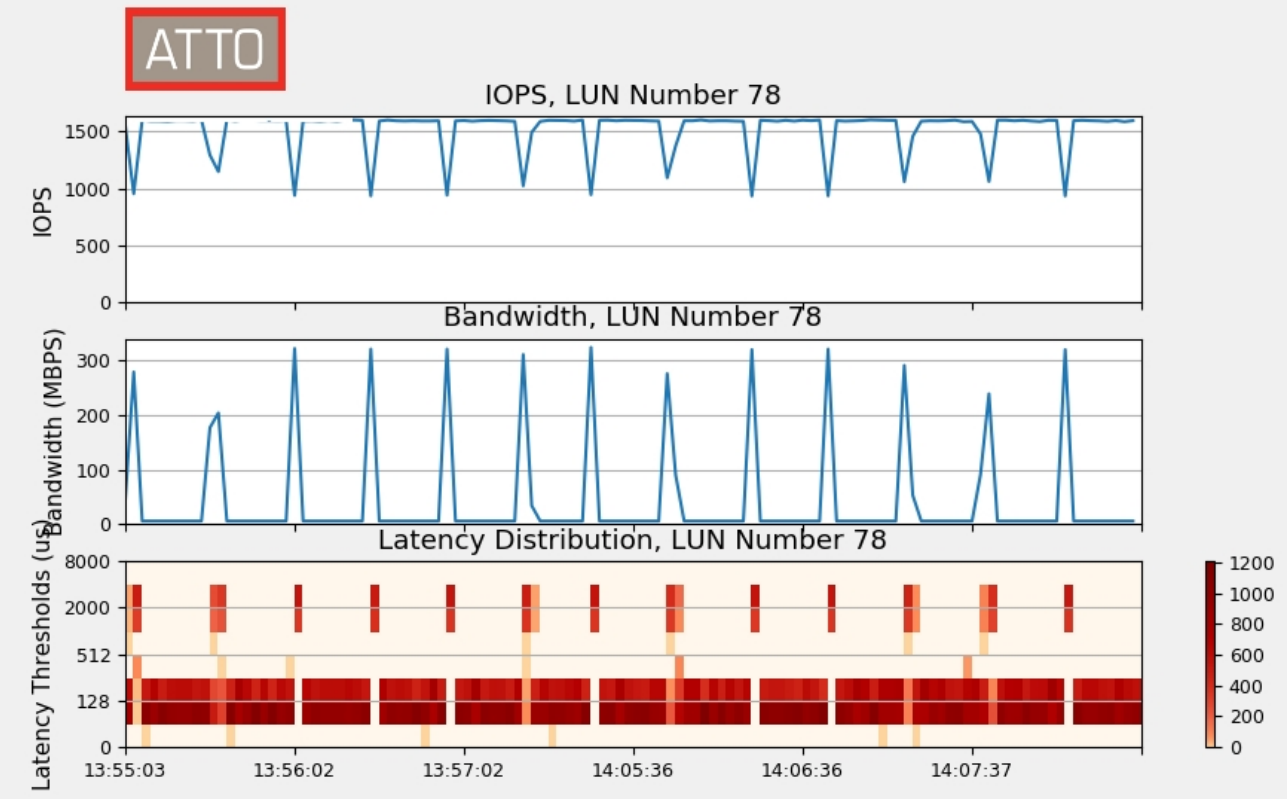
JBOD
120 Drives
4K Sequential Reads
LUN 78: 1 MB Rnd Read
QD 32
2 Workers

Banding in IOPS Heat Map

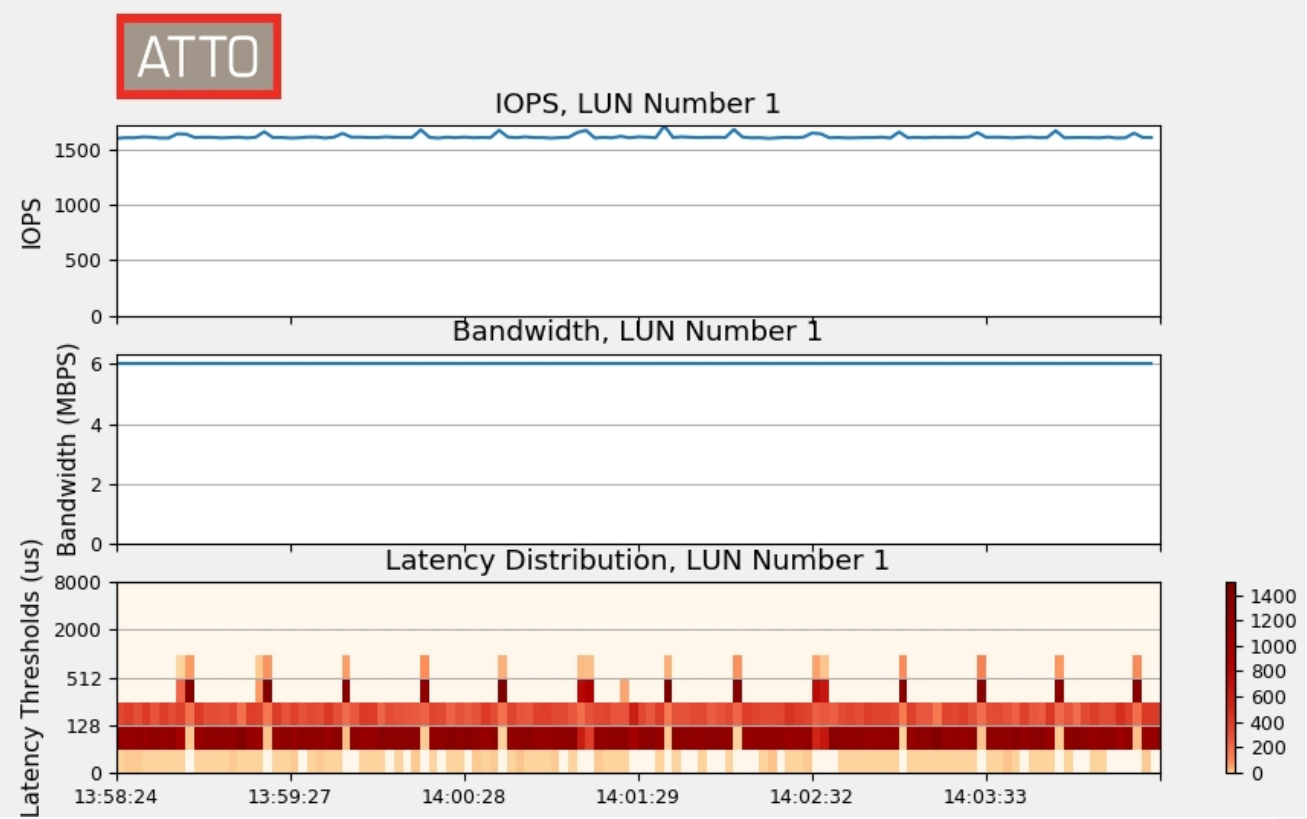


Note: Y-Axis Is LUN

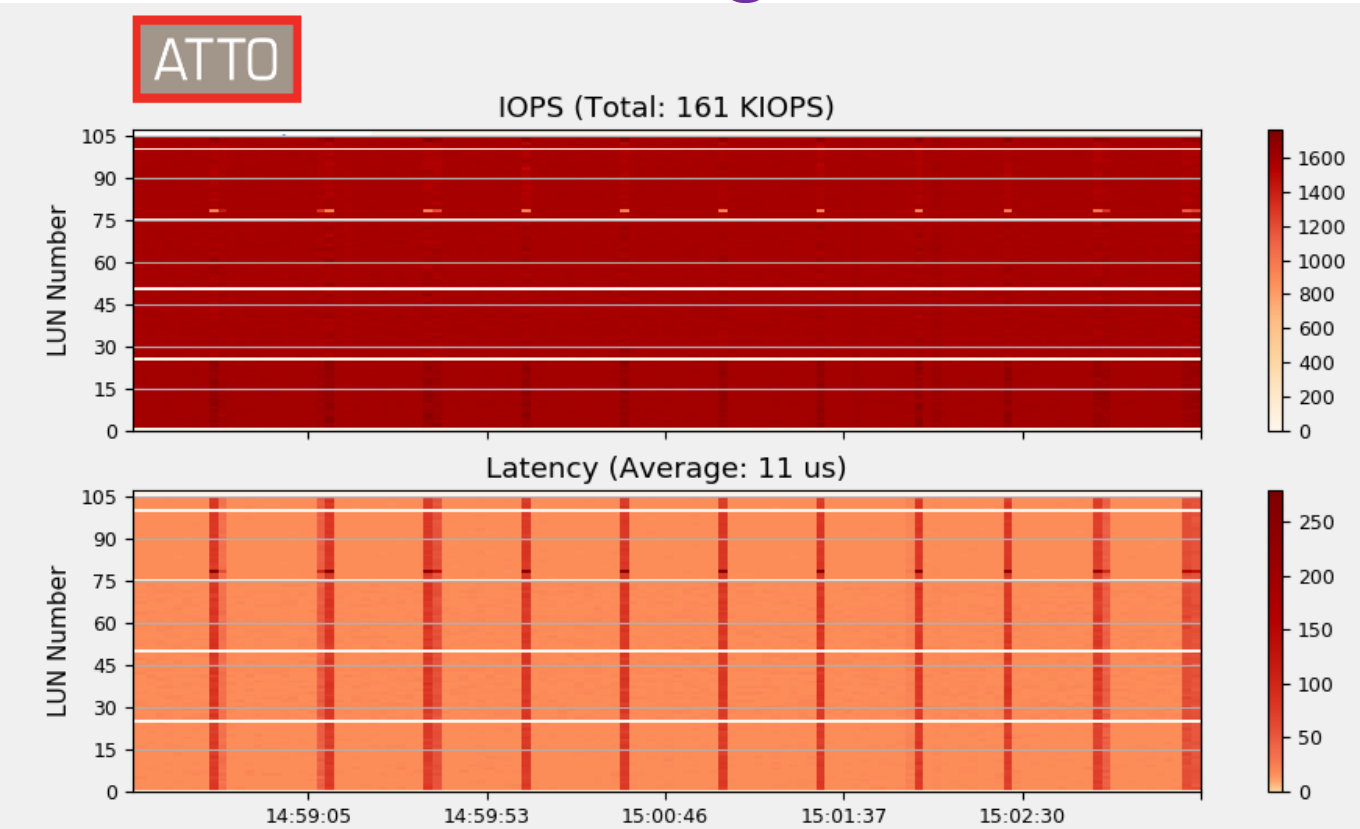
Lun with random reads



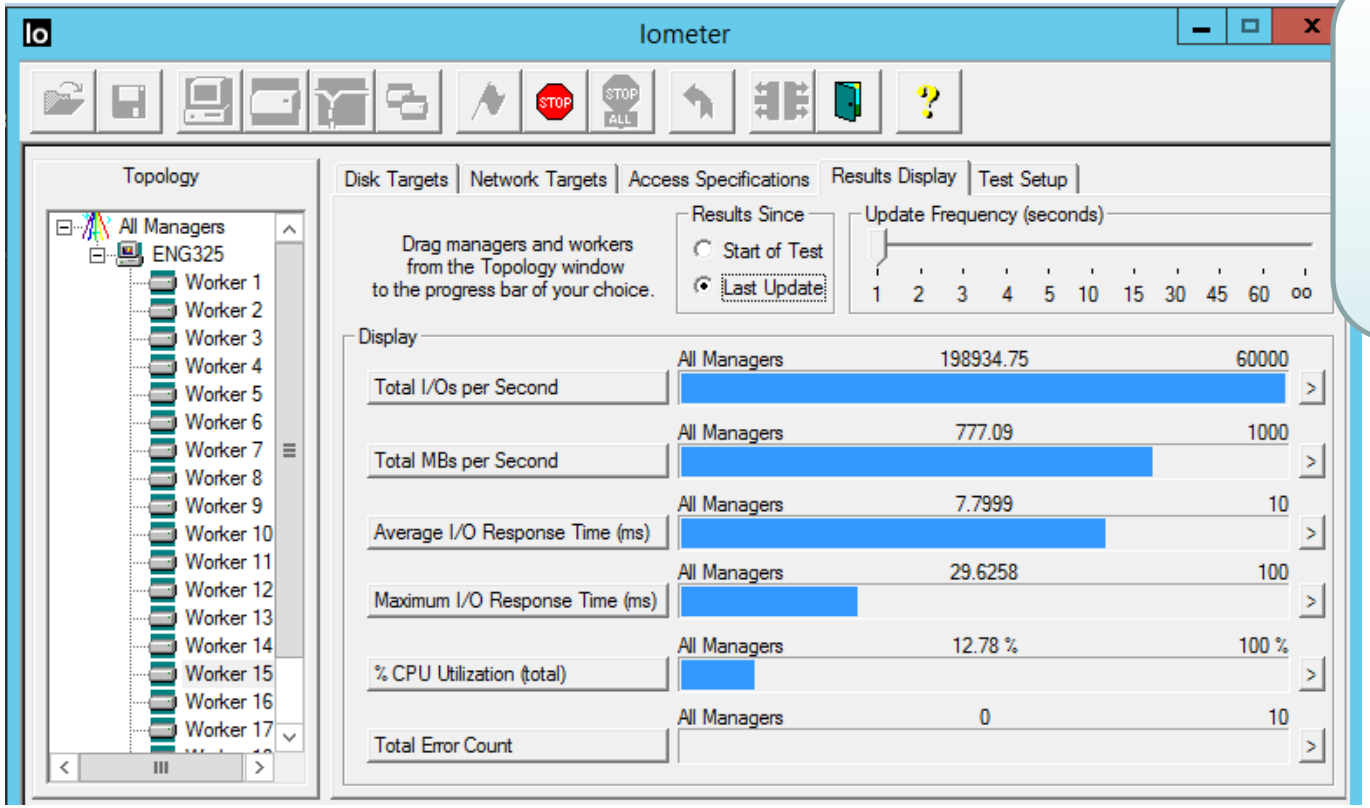
Tick in IOPS – single lun measurement



Measure Command Ingress ONLY

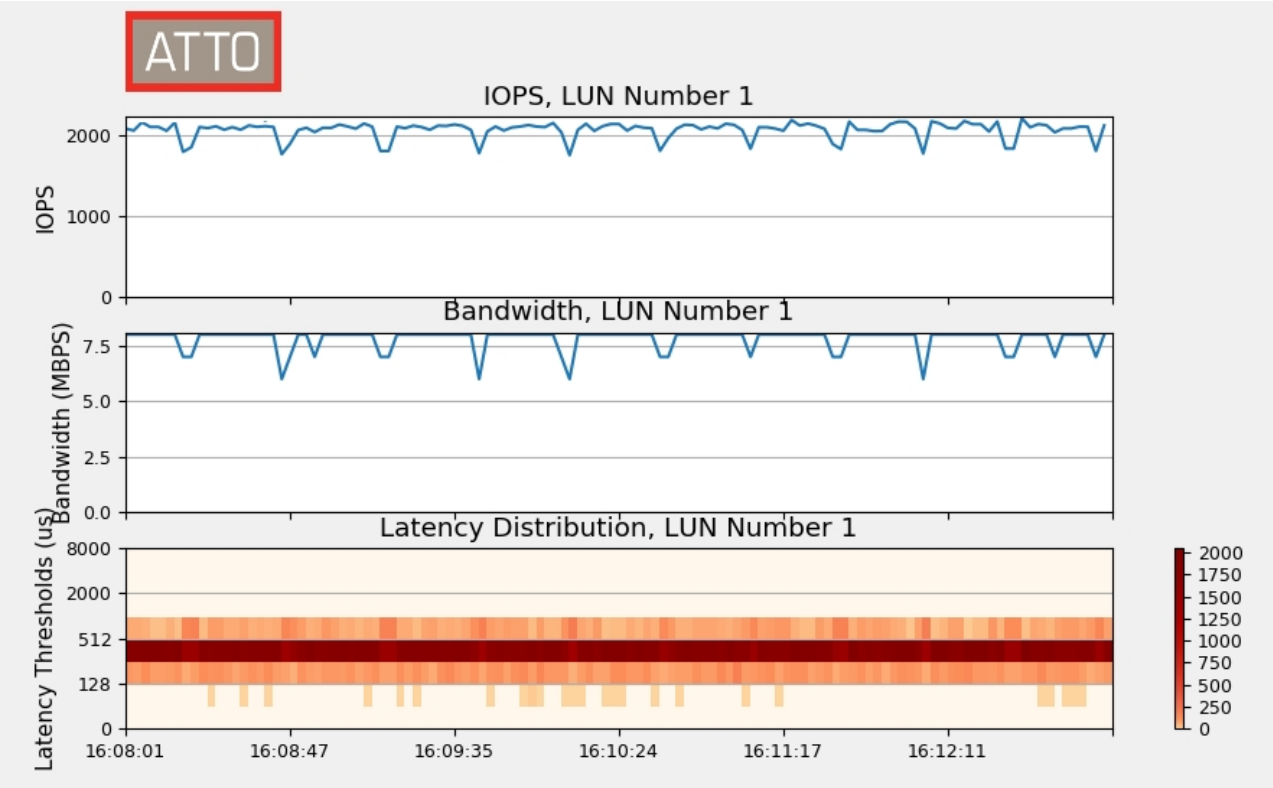


Use 20 Workers Instead of 2

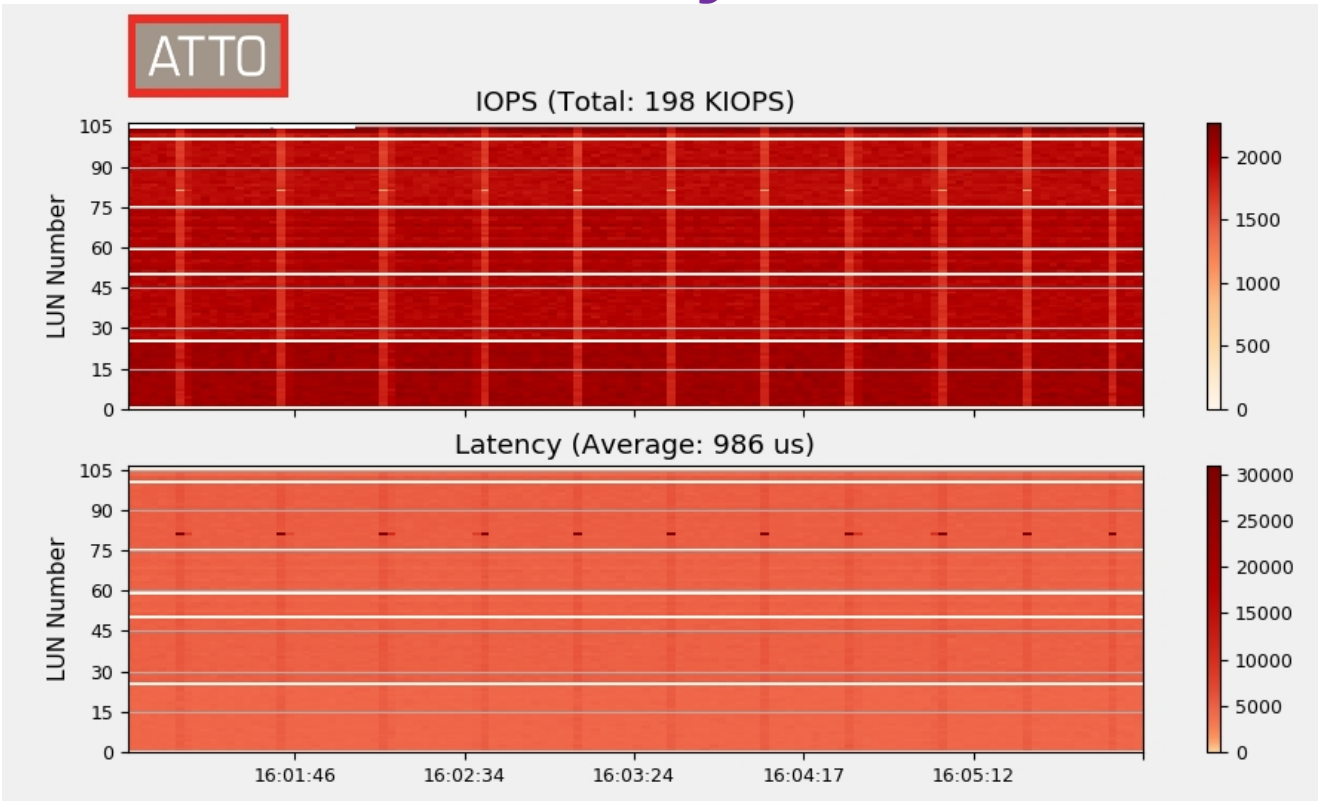


JBOD
120 Drives
4K Sequential Reads
LUN 78: 1 MB Rnd Read
QD 32
20 Workers

What Does LUN 1 Look Like?



Higher Overall Latency? Better IOPS?



Questions

Acknowledgements

- Special thanks to Scott Snowden and Barry Debbins (ATTO Technology, Inc.)
- Images on slides 15,16 & 33 downloaded from OpenClipArt.org