FTL on the Host

Doug Dumitru

CTO WildFire Storage

doug@wildfire-storage.com

SSDs and Local Arrays

- This talk is about how you can optimize SSDs in local arrays.
- Specifically, you might want to …
  - Have drive level redundancy.
  - Optimize the lifespan of your SSDs.
  - Optimize the performance of your Array.
  - Store more data.
- The best solution if you want all of these at the same time is to …

Use a Host FTL

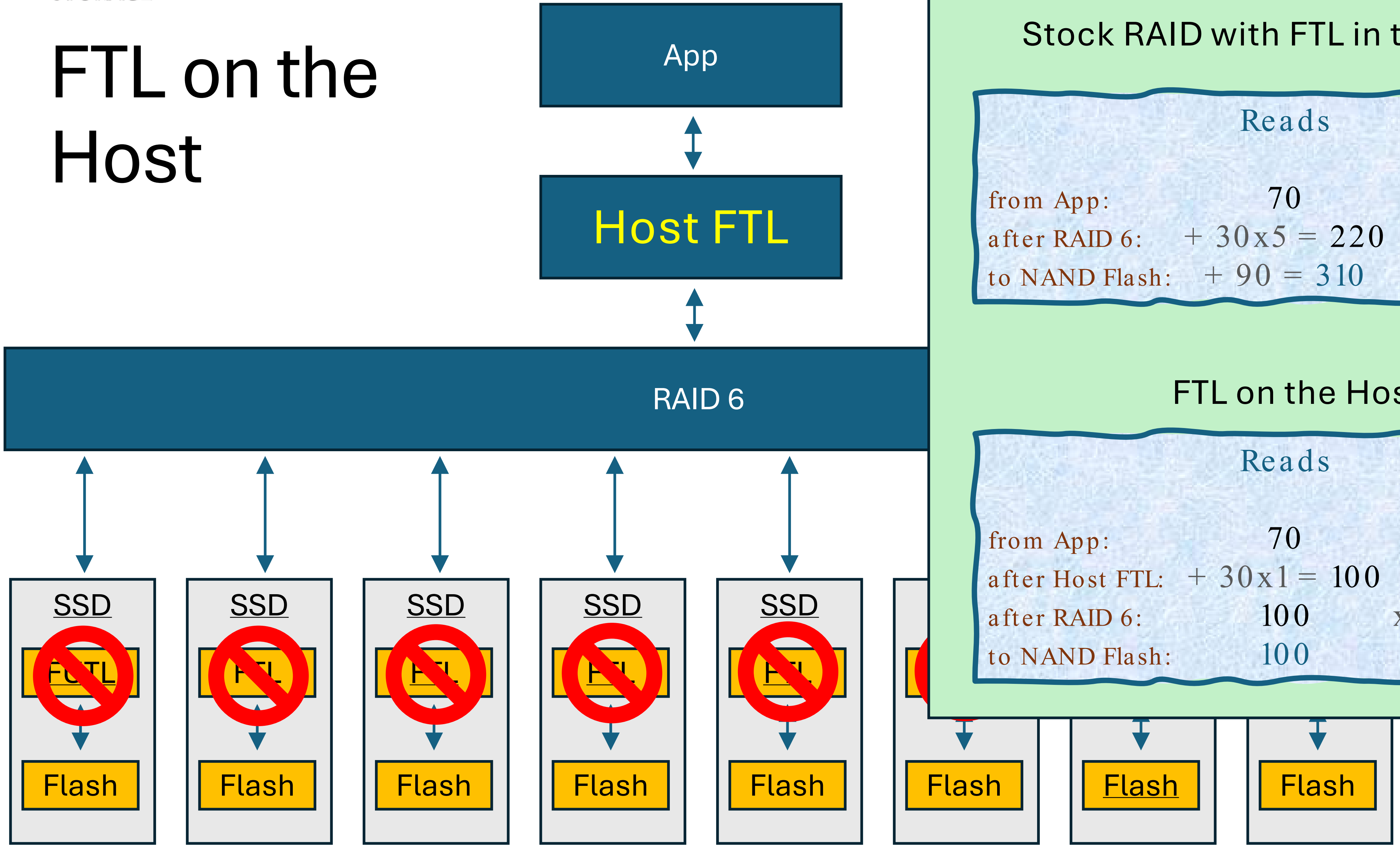
What is a “Host FTL”

- A transparent software layer that creates a block device that your application uses.
  - No special coding.
    - File System, LVM, Virtuals, all just work.
- Sits below the application.
- Sits above RAID
- Linearizes writes, so RAID, and the SSDs get a linear write workload.
  - Linear write workloads are “better” for both parity RAID and for the SSDs.

A Host FTL is Better than the FTL in the SSD

- A pretty bold claim, but the host has more.
  - Resources
  - Time
- Most SSD FTLs are compromises
  - Insufficient RAM
  - Insufficient processor capabilities
  - Requirements for fast mount

FTL on the Host



Stock RAID with FTL in the SSDs

	Reads	Writes
from App:	70	30
after RAID 6:	+ 30x5 = 220	x 3 = 90
to NAND Flash:	+ 90 = 310	x 2 = 180

FTL on the Host

	Reads	Writes
from App:	70	30
after Host FTL:	+ 30x1 = 100	x 2 = 60
after RAID 6:	100	x (8/6) = 80
to NAND Flash:	100	80

How can the overhead be so different

- Parity RAID hates random writes
  - RAID becomes an “IO Amplifier”
    - Writes are 2X – 3X
    - Reads are 2X – 10X or more
  - … and this is before the SSD FTL
    - 2X writes
    - Each write needs a read (for GC)
- Trying to maintain high write IOPS is impossible
  - 1M IOPS can become 8M+ total OPs across the bus.
    - … for an 8 drive array

Host FTLs are just getting started

- A host FTL can compress blocks
  - Compressed blocks use less space.
    - Less space is lower “write amp”
    - Less space is less space.
  - Many workloads end up with under 1:1 write amp even after parity RAID

Where the FTL is Located Matters: --- A LOT

	Stock RAID-6	Host FTL		
		0% Comp	25% Comp	60% Comp
From App	70/30			
After Comp			52/22	28/12
FTL Write Amp		2:1	1.5:1	1.2:1
From FTL		100/60	63/33	30/14
From RAID	220/90	100/80	63/44	30/19
SSD FTL WA	2:1	1:1		
NAND IOs	310/180	100/80	63/44	30/19
Array Writes Per Day	0.4	1.0	1.9	3.8

Where the FTL is Located Matters:
--- Even RAID-5 gets a huge boost

	Stock RAID-5 (fast)	Stock RAID-5 (safe)	Host FTL		
			0% Comp	25% Comp	60% Comp
From App	70/30				
After Comp				52/22	28/12
FTL Write Amp			2:1	1.5:1	1.2:1
From FTL			100/60	63/33	30/14
From RAID	130/60	250/60	100/69	63/38	30/16
SSD FTL WA	2:1	2:1	1:1		
NAND IOs	190/120	310/120	100/69	63/38	30/16
Array Writes Per Day	0.6	0.6	1.0	1.9	4.5

So can a host FTL be fast

- It turns out, “blindingly so”.
  - Lower write amp and less data is less traffic to and from the SSDs.
- Write transfers are longer
  - This is less bus chatter which means fewer system interrupts and their associated overhead.
- To see how fast, follow along …

So on to some Benchmarks

- All of these benchmarks are run on an AWS i4i.metal instance
  - “Bare Metal” rentable server in AWS
    - No hypervisor
    - Direct, actual NVMe SSDs
  - CPU
    - Dual Intel 8357C (Ice Lake) Scalable Xeon Platinum
      - 32 cores x 2 (HT) x 2 (dual socket)
  - Memory
    - 1 TB
  - SSD
    - 8 x 3750 GB NVMe (presumably gen-3) + EBS boot volume
  - Cost
    - < $1.20/hour spot if you shop regions

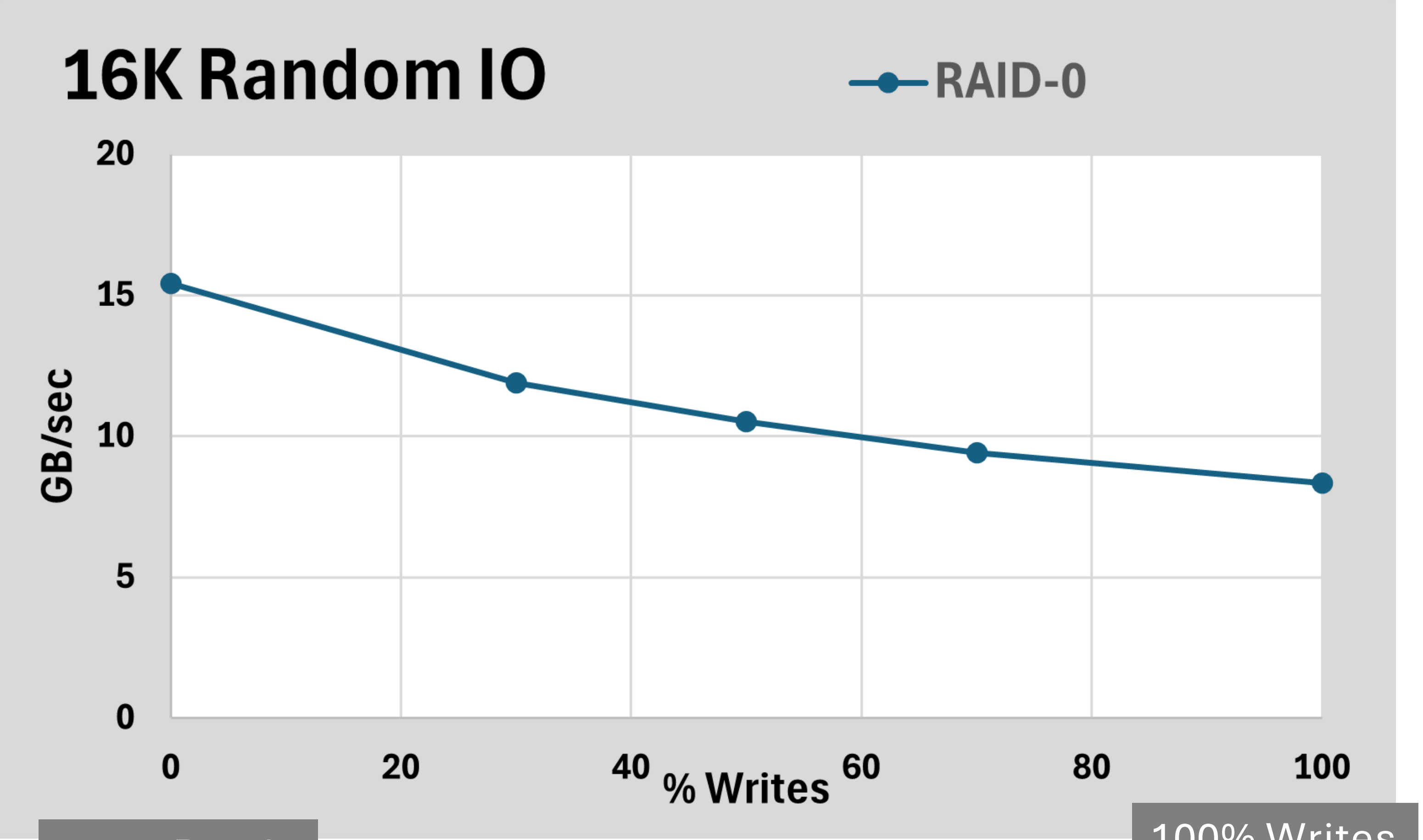
Why Benchmark on AWS

- Easy and Low Cost
- Results can be reproduced by anyone
- Less appearance of “cheating on the test” with unrealistic hardware.

... So on to the benchmarks

Host FTL joins the Performance Race

- All tests are:
  - FIO
    - 16K random blocks
    - Jobs=120
    - Queue=16
- RHEL 9.3
  - Stock Kernel
  - Rocky Linux
- 100/0, 70/30, 50/50 30/70, 0/100 RW



100% Reads

100% Writes

Host FTLs in Linux

- Nothing In-Box
- Actual Host FTL
  - Enterprise Compressed RAID – WildFire Storage
- Almost Host FTL
  - XDP RAIDplus – Pliops
    - … uses co-processor board
- Not a Host FTL
  - Xinnor, GRAID, MD-raid, Megaraid
  - VDO - Redhat

When to use a Host FTL

- Replace RAID-0
  - Faster, longer life, drive redundancy
- Replace RAID-10
  - Faster, larger, longer life
- Replace RAID-5/6
  - Stupidly faster, longer life.

Thank you

... questions

WILDFIRE™
STORAGE

Enterprise **C**ompressed **R**AID

Visit us in booth 1050

WildFire Storage

https://wildfire-storage.com

sales@wildfire-storage.com

+1 610 237-2000

+1 888 473-7866

FMS