Building an All Flash Server

What's the big deal? Isn't it all just *"plug and play"*?

Doug Rollins

Micron Technology

What we'll cover

- Industry Secrets (shhhhh……. 

)
- Example Platform
  - Key features
- Example Workloads:
  - Local boot, data storage, IO acceleration
    - Random? Sequential? Small or large transfers?
    - Synthetic and 'real'
  - How many workloads can there be in one server?
  - Are they different?
- Let's Go Shopping!
  - Are drives optimized – or all they all the same (*and just really, really fast*) ?

Industry Secret

- There's a poorly kept secret in the flash and SSD world……..

Industry Secret

- There's a poorly kept secret in the flash and SSD world…….all flash is NOT the same

Industry Secret

- There's a poorly kept secret in the flash and SSD world…….all flash is NOT the same

- Corollary: All SSDs are not the same

Industry Secret #2

- There's also a poorly kept secret in the datacenter……..

Industry Secret #2

- There's also a poorly kept secret in the datacenter……..all workloads are NOT the same

Industry Secret #2

- There's also a poorly kept secret in the datacenter……..all workloads are NOT the same

- Corollary: Choose your SSDs carefully

Industry Secret #2

- There's also a poorly kept secret in the datacenter……..all workloads are NOT the same
- Corollary: Choose your SSDs carefully
- Corollary: It isn't all just "plug and play" (at least not for optimal designs)

Example Workloads: Overview of a Self-Contained, Small dB Server*

*We will use a **gross** synthetic approximation of a database-like workload

Workloads Overview

- Get it running
- Store some “stuff”
- Access the “stuff” (quickly)

Workloads Overview

- Get it running
- Store some “stuff”
- Access the “stuff” (quickly)

Workload 1: Getting it running (boot)

- Workload: OS boot / Application load
  - Assume virtual memory is not needed
- Data pattern: Sequential (mostly)
- Traffic type: READ (mostly)
- Drive size: Small/moderate capacity
  - 200GB to 256GB or so should be enough
- Latency: Low on READ, WRITE less important
- Cost: Low as possible, but reliable

Workload 1: Boot – what's important

- Interface: SATA is fine
- Capacity: Small-ish is fine (enough for OS, application, and patches)
- Focus: READ performance
- Flash: MLC (typically)

Workload 1: Boot Sidebar

- Q: Why SATA?
- Q: Why “small-ish”?
- Q: Why focus on READ performance?
- Q: Why MLC?

Workloads Overview

- Get it running
- Store some “stuff”
- Access the “stuff” (quickly)

Workload 2: Store some stuff (database-like workload)

- Workload: Database main storage*
- Data pattern: Random
- Traffic type: Small transfers
  - ~8KiB
  - READ 2 for ever 1 WRITE
- Drive size: Large-"ish" capacity (assume we need ~1 TB total)
- Latency: Balanced READ/WRITE, lower (but not lowest)
- Cost: Moderate, but MUST have *robust* data protection

*67R/33W 8KiB RND is a **gross** synthetic approximation of a database-like workload (w/o logging, etc)

Workload 2: dB – what's important

- Interface: SAS is best, SATA is OK
- Capacity: Large-ish
- Focus: Mixed-mode, random performance
- Flash: eMLC (good balance price/performance)

Workload 2: dB Sidebar

- Why is SAS preferred? Why is SATA OK?
- What drives the capacity requirement? Should you use more “smaller” drives, or fewer “larger” drives?
- Why is “mixed” performance paramount?
- Is “special” eMLC really good enough? How can eMLC be made better?

Workloads Overview

- Get it running
- Store some “stuff”
- Access the “stuff” (quickly)

Step 3: Speed it up! (IO acceleration workload)

- Workload: Accelerate the IO from the main array
- Data pattern: Random
- Traffic type: Small transfers, R/W ‘mix’ may vary
- Drive size: Smaller than main array (~10% to 20% of the main array size)
- Latency: Lowest possible for all traffic
- Cost: Tends to be high(er), may have robust data protection*

*write cache => more robust, read cache less so

Workload 3: I/O Acceleration – what's important

- Interface: PCIe is best (NVMe later?)
- Capacity: Midsized
- Focus: FAST, mixed-mode random R/W
- Flash: SLC or eMLC

Let’s build a server…..

Example Platform

- Key features:
  - Supports SAS, SATA, and PCIe* storage
  - Flexible CPU options (2P capable)
  - Rack mount
- Flexible DRAM ‘footprint’

*Card FF is good enough, but 2.5” is far better

Example Platform: Today's focus

- Key features:
  - Supports SAS, SATA, and PCIe* storage
  - Flexible CPU options (2P capable)
  - Rack mount
- Flexible DRAM 'footprint'

*Card FF is good enough, but 2.5" is far better

3 Workloads…..1 Server……
Let’s go SSD shopping!

Preview: 3 Sample SSDs

- 3 drive interfaces
  - SATA, SAS, and PCIe
- 3 media types
  - Standard MLC, “special” MLC, and SLC
- 3 design targets
  - Read-centric, Balanced, and Premium/Accelerator

3 Workloads……3 Drive Options

Drive 1	
Capacity[1]	175GB, 350GB
Interface	x4 PCIe Gen2
Connector	SATA/SAS/PCIe combination
Sequential read/write performance[2]	Up to 1.75/1.1 GB/s
Random read/write performance[3]	
Latency	
Active power consumption	
Idle power consumption	
MTTF	
Form factor	

Drive 2	
Capacity[1]	50GB, 100GB, 200GB, 400GB
Interface	SATA 6 Gb/s; backward-compatible to SATA 3 Gb/s
Sequential read/write	Up to 350/140 MB/s
Random read/write	Up to 50,000/7000 IOPS
Active average power consumption[2]	
Idle power consumption	
Operating shock	
Operating vibration	
MTTF	
Endurance	

Drive 3	
Capacity[1]	100GB, 200GB, 400GB
Interface	SAS 6 Gb/s
Sequential read/write performance	100GB: 410/235 MB/s 200GB: 410/345 MB/s 400GB: 410/345 MB/s
Random read/write performance	100GB: 50,000/20,000 IOPS 200GB: 50,000/30,000 IOPS 400GB: 50,000/30,000 IOPS
Active average power consumption[2]	<9W
Operating shock	1000G/0.5ms
Operating vibration	10–500Hz at 3.1G
MTTF	2 million device hours
Endurance	Up to 7PB total bytes written

3 Workloads……Let's go SSD shopping!

- We need to: Boot, store, accelerate
- Drive 1:

Drive 1	
Capacity[1]	175GB, 350GB
Interface	x4 PCIe Gen2
Connector	SATA/SAS/PCIe combination
Sequential read/write performance[2]	Up to 1.75/1.1 GB/s
Random read/write performance[3]	Up to 415,000/145,000 IOPS
Latency	<50µs
Active power consumption	25W (MAX)
Idle power consumption	6.1mW
MTTF	2 million device hours
Form factor	2.5in

3 Workloads……Let's go SSD shopping!

- We need to: Boot, store, accelerate
- Drive 1:

Drive 1	
Capacity[1]	175GB
Interface	x4 PCIe Gen2
Connector	SATA/SAS/PCIe combination
Sequential read/write performance[2]	Up to 1.75/1.1 GB/s
Random read/write performance[3]	Up to 415,000/145,000 IOPS
Latency	<50µs
Active power consumption	25W (MAX)
Idle power consumption	6.1mW
MTTF	2 million device hours
Form factor	2.5in

<u>Key Features of Drive 1</u>

14% of planned main array size (28% available)

Low latency drive/host interface

Flexible implementation

Excellent random performance for READ and WRITE

Low latency

This drive uses SLC NAND

3 Workloads……Let's go SSD shopping!

- We need to: Boot, store, accelerate
- Drive 2:

Drive 2	
Capacity[1]	50/64/100/128/200/256/400/512GB
Interface	SATA 6 Gb/s; backward-compatible to SATA 3 Gb/s
Sequential read/write	Up to 350/140 MB/s
Random read/write	Up to 50,000/7000 IOPS
Active average power consumption[2]	50GB: 2.5W 400GB: Up to 5W
Idle power consumption	0.95W (MAX)
Operating shock	1500G/0.5ms
Operating vibration	5–500Hz at 3.1G
MTTF	1.2 million device hours
Endurance	Up to 175TB lifetime data written

3 Workloads……Let's go SSD shopping!

- We need to: Boot, store, accelerate
- Drive 2:

Drive 2	
Capacity[1]	50/64/100/128/200/256/400/512GB
Interface	SATA 6 Gb/s; backward-compatible to SATA 3 Gb/s
Sequential read/write	Up to 350/140 MB/s
Random read/write	Up to 50,000/7000 IOPS
Active average power consumption[2]	50GB: 2.5W 400GB: Up to 5W
Idle power consumption	0.95W (MAX)
Operating shock	1500G/0.5ms
Operating vibration	5–500Hz at 3.1G
MTTF	1.2 million device hours
Endurance	Up to 175TB lifetime data written

<u>Key Features of Drive 2</u>

Widest range of user capacities

Moderate latency drive/host interface

Fast READ performance (either random or sequential)

This drive uses MLC NAND

3 Workloads……Let's go SSD shopping!

- We need to: Boot, store, accelerate
- Drive 3:

Drive 3	
Capacity[1]	100GB, 200GB, 400GB
Interface	SAS 6 Gb/s
Sequential read/write performance	100GB: 410/235 MB/s 200GB: 410/345 MB/s 400GB: 410/345 MB/s
Random read/write performance	100GB: 50,000/20,000 IOPS 200GB: 50,000/30,000 IOPS 400GB: 50,000/30,000 IOPS
Active average power consumption[2]	<9W
Operating shock	1000G/0.5ms
Operating vibration	10–500Hz at 3.1G
MTTF	2 million device hours
Endurance	Up to 7PB total bytes written

3 Workloads……Let's go SSD shopping!

- We need to: Boot, store, accelerate
- Drive 3:

Drive 3	
Capacity[1]	100GB, 200GB, 400GB
Interface	SAS 6 Gb/s
Sequential read/write performance	100GB: 410/235 MB/s 200GB: 410/345 MB/s 400GB: 410/345 MB/s
Random read/write performance	100GB: 50,000/20,000 IOPS 200GB: 50,000/30,000 IOPS 400GB: 50,000/30,000 IOPS
Active average power consumption[2]	<9W
Operating shock	1000G/0.5ms
Operating vibration	10–500Hz at 3.1G
MTTF	2 million device hours
Endurance	Up to 7PB total bytes written

<u>Key Features of Drive 3</u>

Wide range of user capacities

Moderate latency drive/host interface

Fast READ and WRITE performance (both random and sequential)

This drive uses "eMLC" NAND

Matching Workloads to SSDs

Matching Workloads to SSDs

Drive 1	
Capacity[1]	175GB, 350GB
Interface	x4 PCIe Gen2
Connector	SATA/SAS/PCIe combination
Sequential read/write performance[2]	Up to 1.75/1.1 GB/s
Random read/write performance[3]	
Latency	
Active power consumption	
Idle power consumption	
MTTF	
Form factor	

Drive 2	
Capacity[1]	50/64/100/128/200/256/400/512GB
Interface	SATA 6 Gb/s; backward-compatible to SATA 3 Gb/s
Sequential read/write	Up to 350/140 MB/s
Random read/write	Up to 50,000/7000 IOPS
Active average power consumption[2]	
Idle power consumption	
Operating shock	
Operating vibration	
MTTF	
Endurance	

Drive 3	
Capacity[1]	100GB, 200GB, 400GB
Interface	SAS 6 Gb/s
Sequential read/write performance	100GB: 410/235 MB/s 200GB: 410/345 MB/s 400GB: 410/345 MB/s
Random read/write performance	100GB: 50,000/20,000 IOPS 200GB: 50,000/30,000 IOPS 400GB: 50,000/30,000 IOPS
Active average power consumption[2]	<9W
Operating shock	1000G/0.5ms
Operating vibration	10–500Hz at 3.1G
MTTF	2 million device hours
Endurance	Up to 7PB total bytes written

Matching Workloads to SSDs: Boot

- Interface: SATA is fine
- Capacity: Small-ish is fine (enough for OS, application, and patches)
- Focus: READ performance, low price

Drive 2	
Capacity[1]	50/64/100/128/200/256/400/512GB
Interface	SATA 6 Gb/s; backward-compatible to SATA 3 Gb/s
Sequential read/write	Up to 350/140 MB/s
Random read/write	Up to 50,000/7000 IOPS
Active average power consumption[2]	50GB: 2.5W 400GB: Up to 5W
Idle power consumption	0.95W (MAX)
Operating shock	1500G/0.5ms
Operating vibration	5–500Hz at 3.1G
MTTF	1.2 million device hours
Endurance	Up to 175TB lifetime data written

Matching Workloads to SSDs: dB

- Interface: SAS is best, SATA is OK
- Capacity: Large-ish
- Focus: Mixed-mode, random performance, moderate price

Drive 3	
Capacity[1]	100GB, 200GB, 400GB
Interface	SAS 6 Gb/s
Sequential read/write performance	100GB: 410/235 MB/s 200GB: 410/345 MB/s 400GB: 410/345 MB/s
Random read/write performance	100GB: 50,000/20,000 IOPS 200GB: 50,000/30,000 IOPS 400GB: 50,000/30,000 IOPS
Active average power consumption[2]	<9W
Operating shock	1000G/0.5ms
Operating vibration	10–500Hz at 3.1G
MTTF	2 million device hours
Endurance	Up to 7PB total bytes written

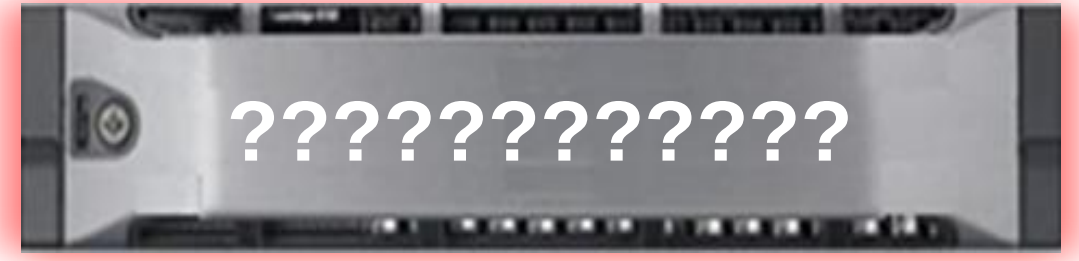
Matching Workloads to SSDs: IO Acceleration

- Interface: PCIe is best (NVMe later?)
- Capacity: Midsized
- Focus: FAST, mixed-mode random R/W, price/GR less of a concern

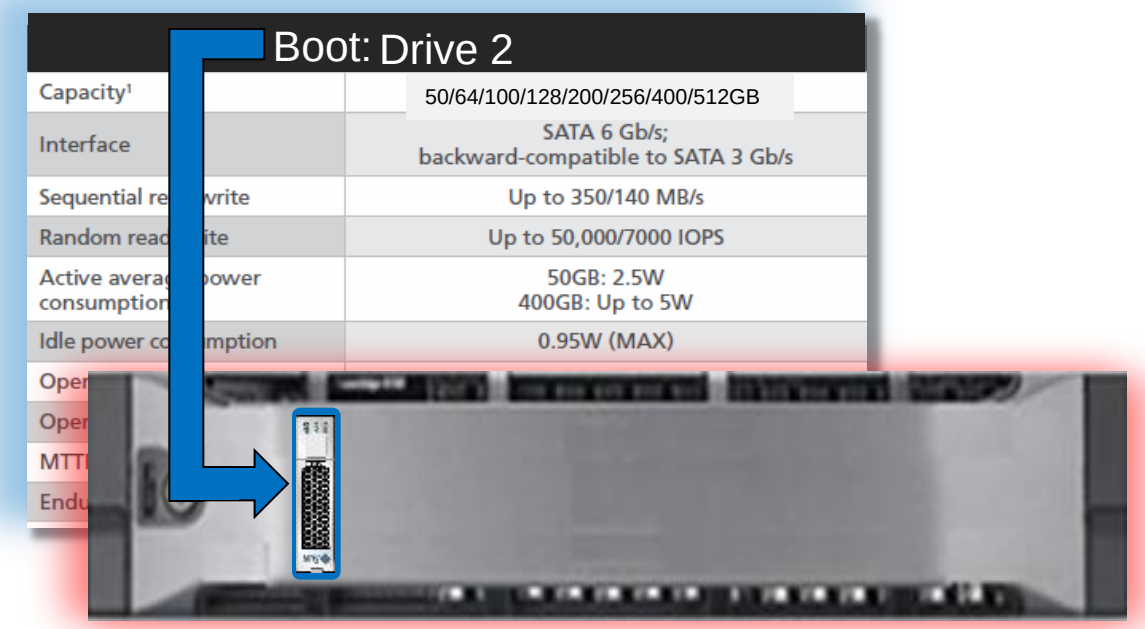
Drive 1	
Capacity[1]	175GB, 350GB
Interface	x4 PCIe Gen2
Connector	SATA/SAS/PCIe combination
Sequential read/write performance[2]	Up to 1.75/1.1 GB/s
Random read/write performance[3]	Up to 415,000/145,000 IOPS
Latency	<50µs
Active power consumption	25W (MAX)
Idle power consumption	6.1mW
MTTF	2 million device hours
Form factor	2.5in

Flash Memory SUMMIT

Putting them all together…..



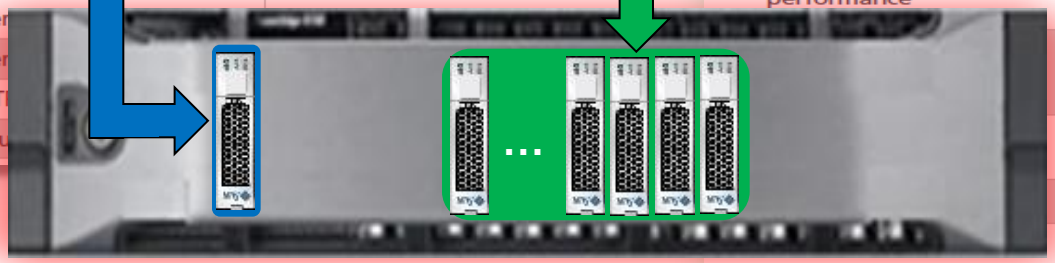
Putting them all together…..



Putting them all together…..

Boot: Drive 2

Capacity[1]	50/64/100/128/200/256/400/512GB
Interface	SATA 6 Gb/s; backward-compatible to SATA 3 Gb/
Sequential re… vrite	Up to 350/140 MB/
Random read… ite	Up to 50,000/7000 IO
Active averag… power consumption	50GB: 2.5W 400GB: Up to 5W
Idle power co… mption	0.95W (MAX)
Oper…	
Oper…	
MTTF	
Endu…	



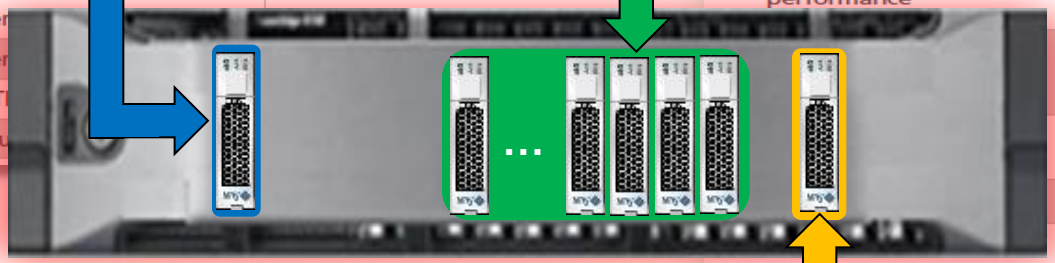
Store: Drive 3

Capacity[1]	100GB, 200GB, 400GB
Interface	SAS 6 Gb/s
Sequential read/write performance	100GB: 410/235 MB/s 200GB: 410/345 MB/s 400GB: 410/345 MB/s
	100GB: 50,000/20,000 IOPS 200GB: 50,000/30,000 IOPS 400GB: 50,000/30,000 IOPS
	<9W
	1000G/0.5ms
	10–500Hz at 3.1G
MTTF	2 million device hours
Endurance	Up to 7PB total bytes written

Putting them all together…..

Boot: Drive 2

Capacity[1]	50/64/100/128/200/256/400/512GB
Interface	SATA 6 Gb/s; backward-compatible to SATA 3 Gb/
Sequential re…write	Up to 350/140 MB/
Random read…ite	Up to 50,000/7000 IO
Active averag…power consumption	50GB: 2.5W 400GB: Up to 5W
Idle power co…mption	0.95W (MAX)
Oper…	
Oper…	
MTT…	
Endu…	



Store: Drive 3

Capacity[1]	100GB, 200GB, 400GB
Interface	SAS 6 Gb/s
Sequential read/write performance	100GB: 410/235 MB/s 200GB: 410/345 MB/s 400GB: 410/345 MB/s
	100GB: 50,000/20,000 IOPS 200GB: 50,000/30,000 IOPS 400GB: 50,000/30,000 IOPS
	<9W
	1000G/0.5ms
	10–500Hz at 3.1G
MT…	2 million device hours
…rance	Up to 7PB total bytes written

Accelerate: Drive 1

Capacity[1]	175GB, 350GB
Interface	x4 PCIe Gen2
Connector	SATA/SAS/PCIe combination
Sequential read/write performance[2]	Up to 1.75/1.1 GB/s
Random read/write performance[3]	Up to 415,000/145,000 IOPS
Latency	<50µs
Active power consumption	25W (MAX)
Idle power consumption	6.1mW
MTTF	2 million device hours
Form factor	2.5in

….wrap up…..

- All flash is not created equal
  - Endurance and performances differentiate flash
- All flash management is not created equal
  - Good flash management differentiates drives
- All SSDs are not created equal
  - SSDs may be best suited for vastly different workloads: READ, Mixed-mode, Cache, etc
  - Media type and interface suggest, but don't determine what is "best"
- A good understanding of workloads enables optimal platform design and best value

…. Q and A…..