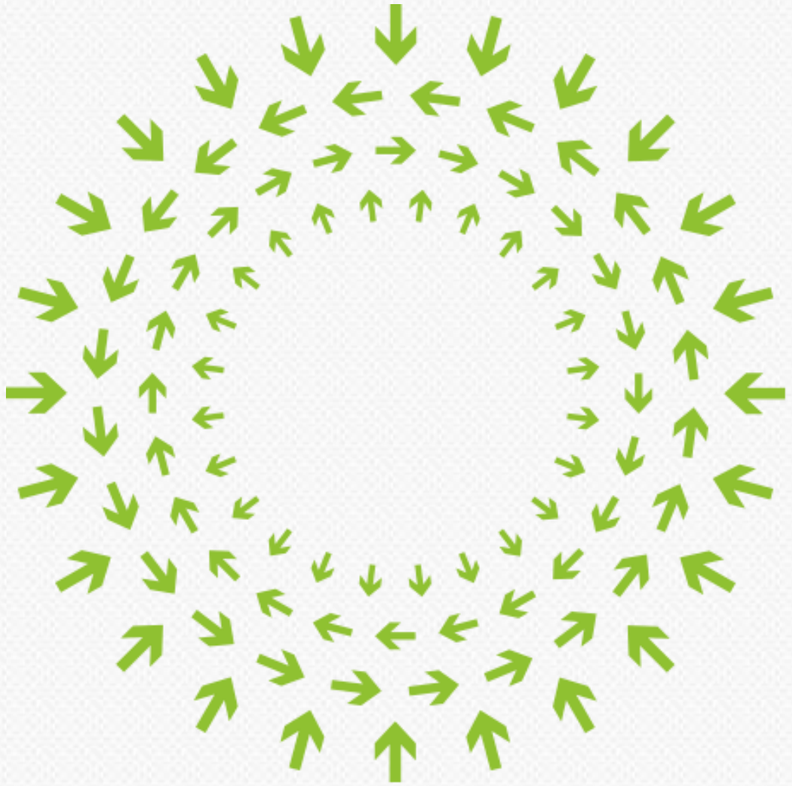


OPEN Compute Summit

March 10–11, 2015

San Jose

OPEN Compute Summit

Microsoft OCS Cloud M.2 SSD

Optimizing flash storage for hyperscale

Laura Caulfield
Microsoft
Software Developer II

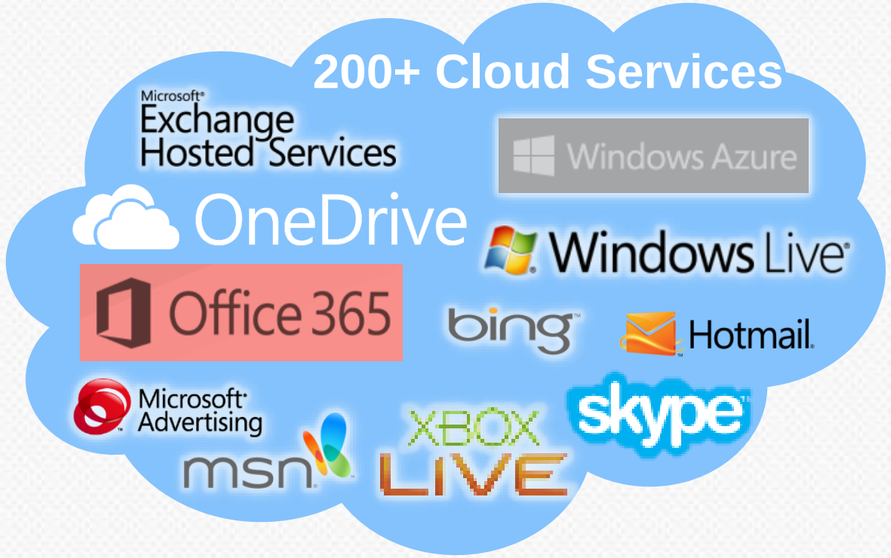
Priorities for Cloud Storage

Fast is good, but not for infinite cost

Design once, buy many

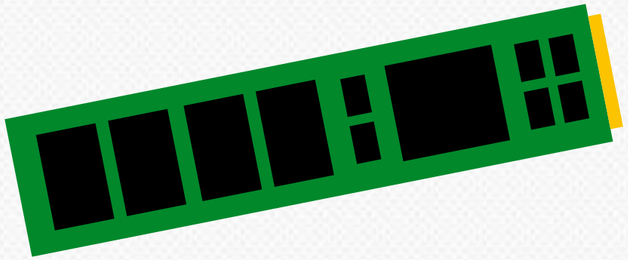
Wide variety of applications

M.2 is best suited
accomplish these goals



Outline

M.2 Background



Advantages to M.2

Cloud-Specific Requirements

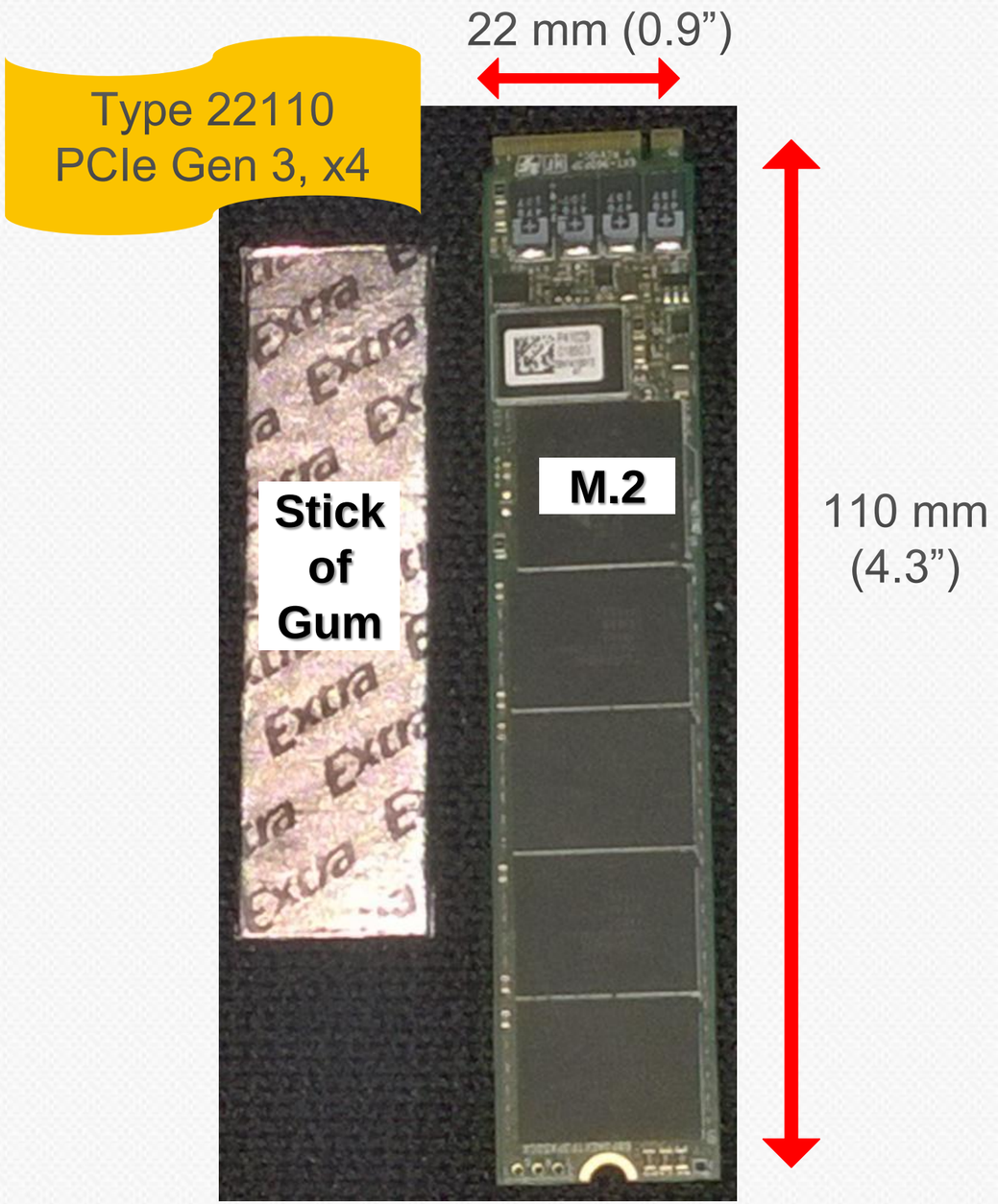
What is M.2?

M.2 is a small form factor card

- Specification from PCI-SIG
- Not just for SSDs (eg: WWAN)
- For SSDs, Typically SATA or PCIe
- Well suited to mobile devices (laptops, tablets)

M.2 SSDs in our blade design

- Hardware: Minimally PCI-Express,
  Ideally Gen3 x4
- Protocol: Can be “AHCI over PCIe”
  Ideally NVMe
- Mechanical: 60mm, 80mm or 110mm



M.2 in the Blade

M.2 Riser

M.2
PCIe x4
Connector

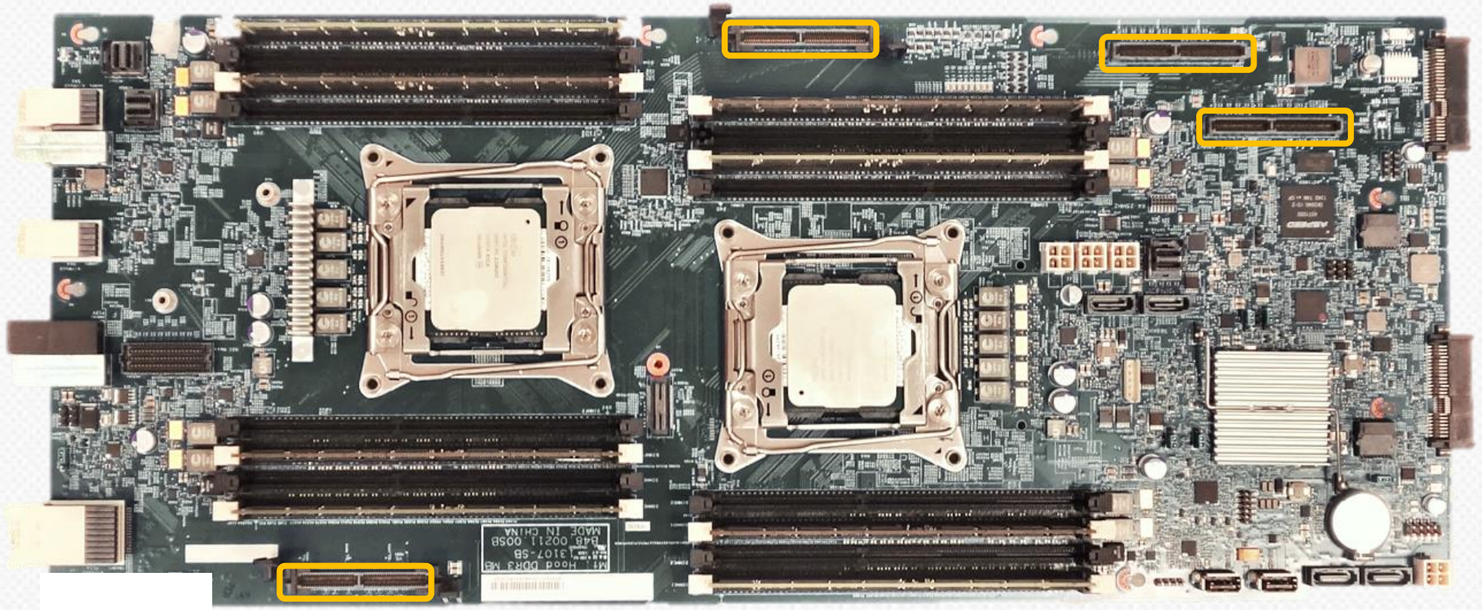
PCIe x8 Gen3

60mm, 80mm, 110 mm

M.2

22 mm
(0.9”)

Type 22110 - 110 mm (4.3”)

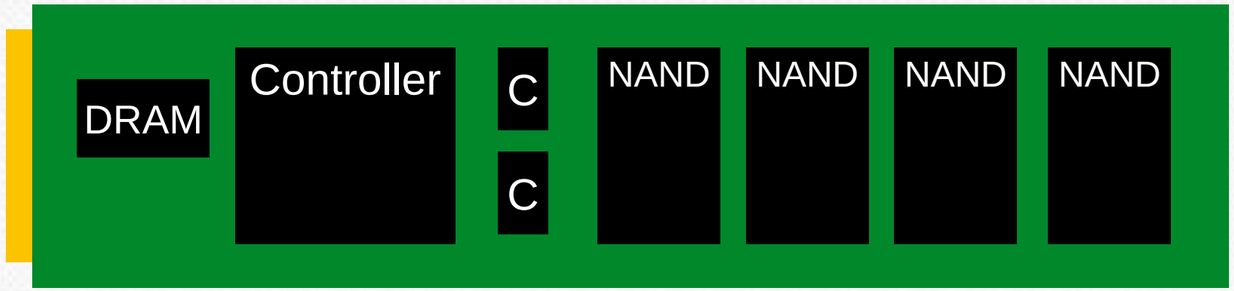


[reference designs shown]

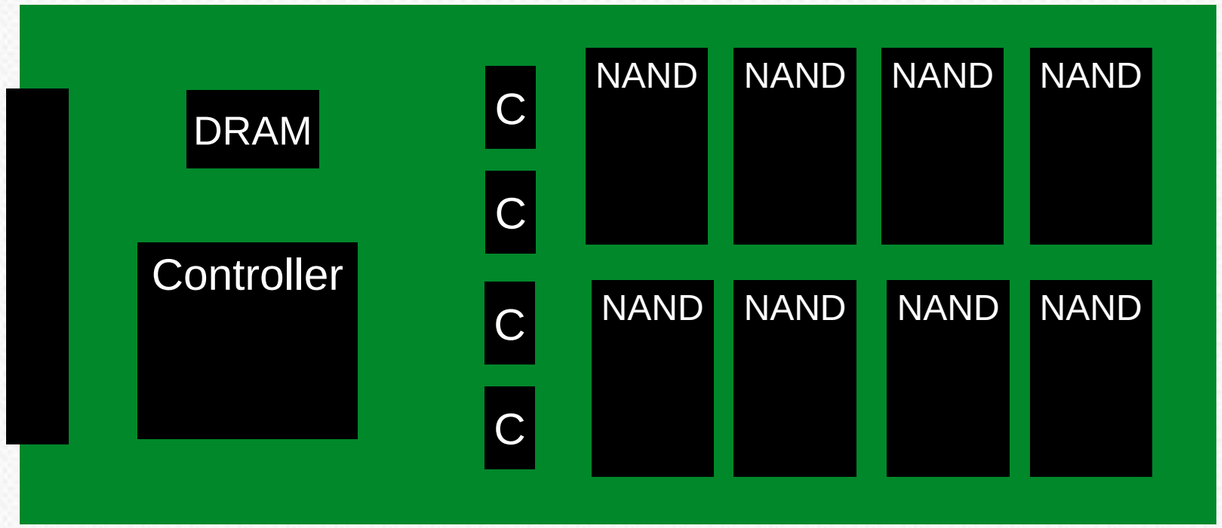
PCIe x8 /
M.2 Risers

Components of the M.2

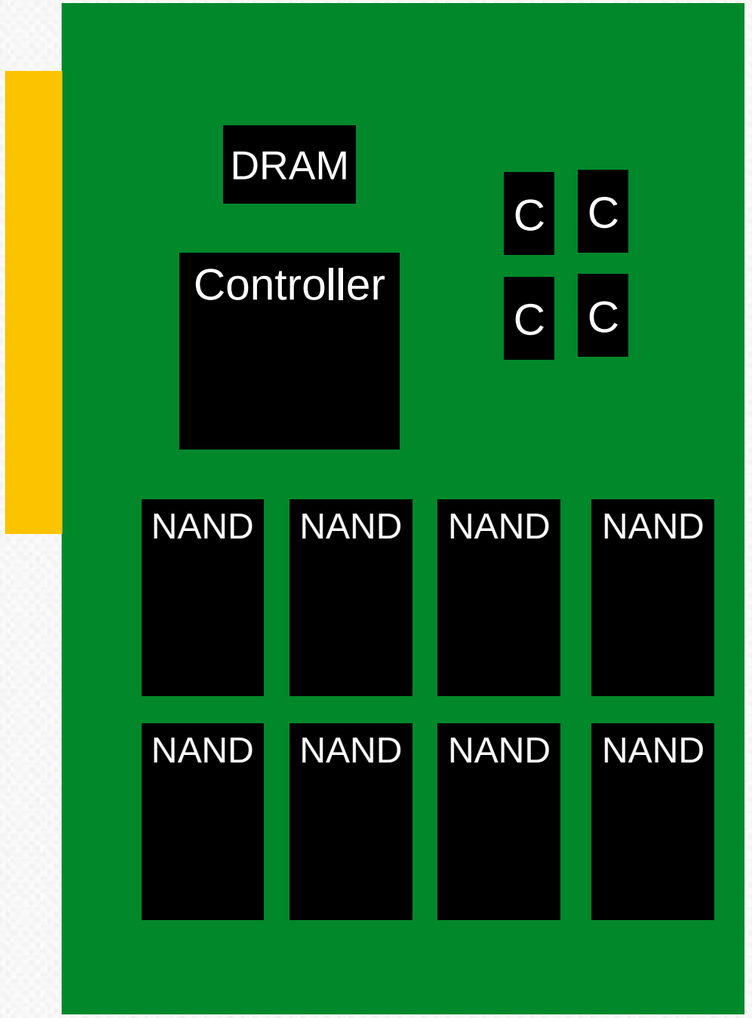
M.2



2.5” Drive



Large Form Factor PCIe Card

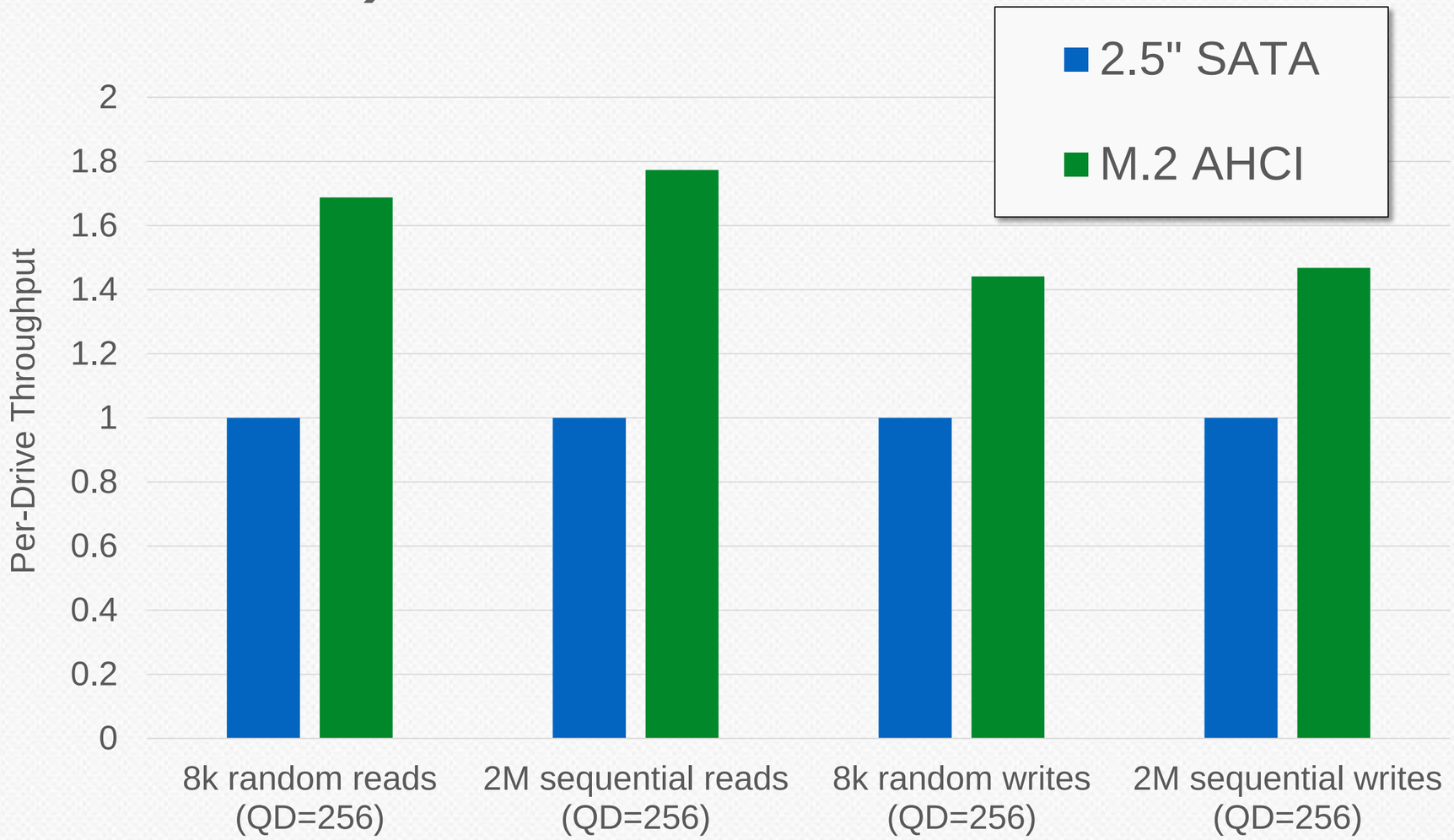


These Diagrams are Not to Scale

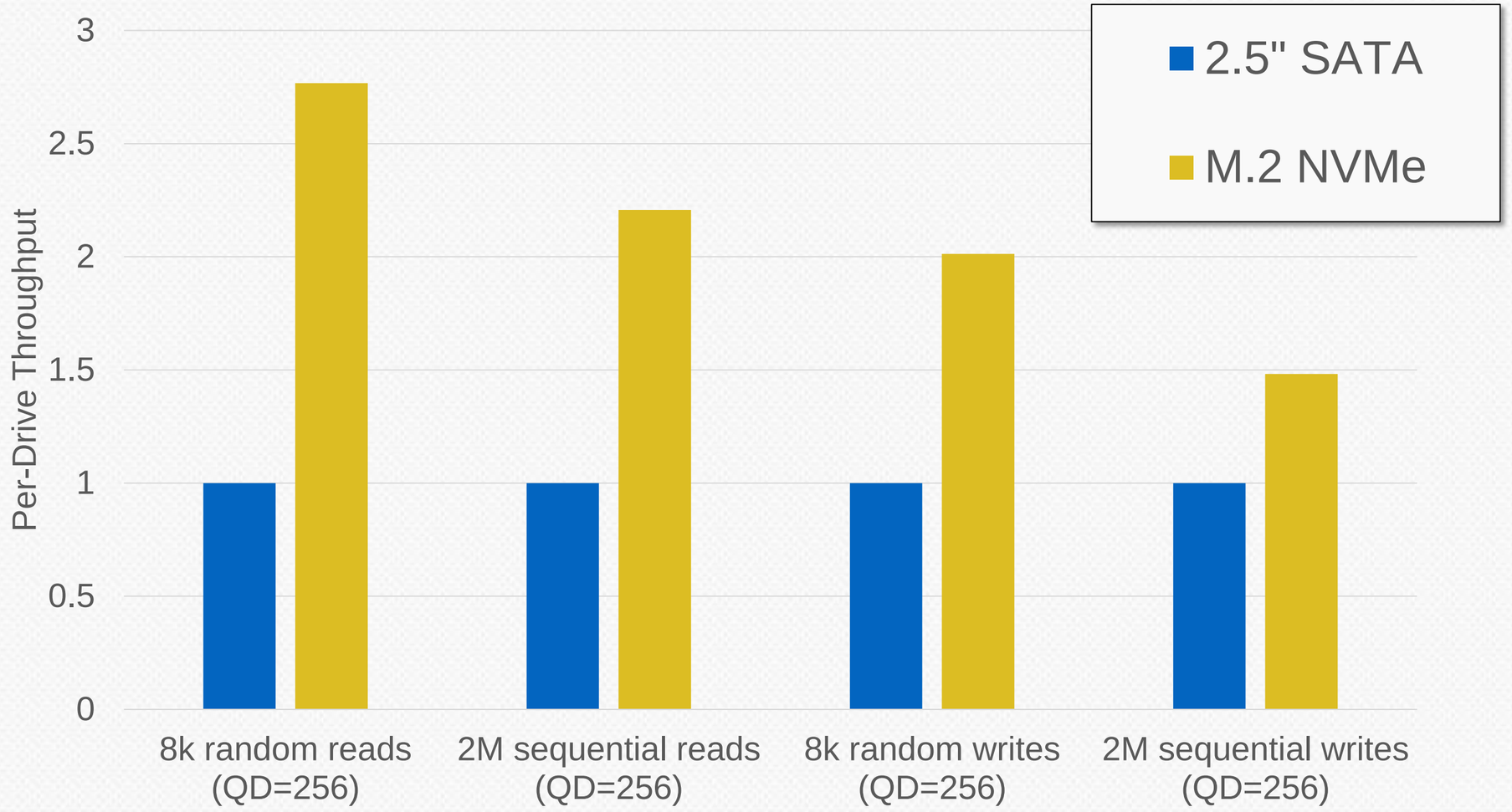
Step 1 (Hardware): AHCI over PCIe

Hardware advances to PCIe, Software remains backward compatible

Removing the bottleneck on SATA hardware creates at least 1.4x throughput improvement for all workloads



Step 2 (Software): NVMe over PCIe



NVMe reduces the protocol overhead, creating average of 43% better throughput for all workloads except large sequential writes.

Outline

M.2 Background

The M.2 Advantage

Cloud-Specific Requirements

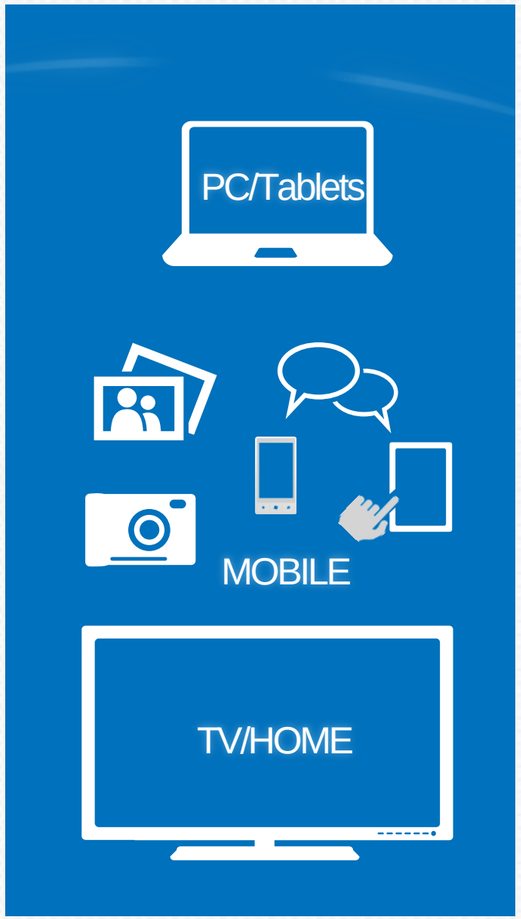
Why PCIe M.2 for Cloud Storage?

Leverage the Economy of Scale in Client Market

- PCIe M.2s are emerging as the "go to" for mobile applications

- Some differences for cloud applications
  - Save cached data during power failure (PFail)
  - Endurance and Retention

Can address these differences with minor changes to the FW and BOM (more on this later)



Many PCIe Form Factors: The Trade-offs

	M.2	2.5”	Large Card
Capacity	960 GB	1.6 TB	1.6 TB
Power	8 W	25 W	25 W
4k Random Read (kIOPS)	440	740	750
4k Random Write (kIOPS)	40	115	95
Sequential Read (GB/s)	1.5	3	3.3
Sequential Write (MB/s)	750	1,400	630
Cabling	No	Yes	No
Drives Per Server	8	4	1

Overprovisioning (OP)
7% vs. 28%
(2x rand. write performance)

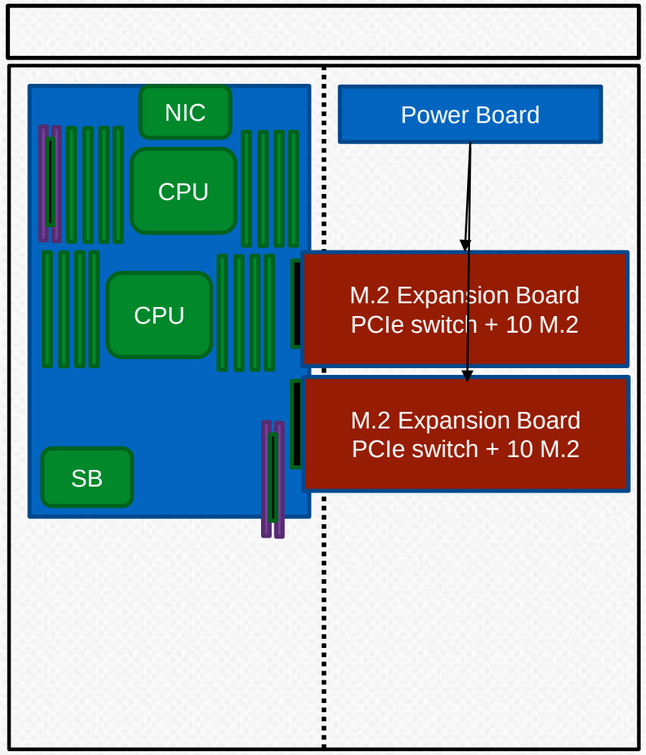
Large card has no cabling,
but inflexible physical size

36% less power
per server with M.2
(same raw capacity)

Same performance per GB.
Better performance per watt.

M.2 is an Efficient Element

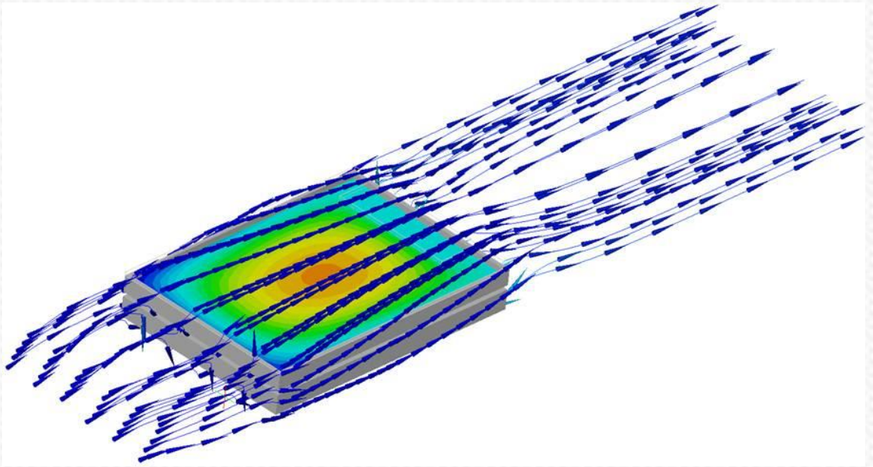
- M.2 unit is dense (GB/volume)
- Cost and power scale with capacity
- One blade
  - Down to 480GB (minimal flash, low cost)
  - Up to 24 TB (lots of flash, high perf. storage)
- Many adapters to add capacity
  - Adapters to load M.2 onto HHHL PCIe Card
  - Custom design to expand into other ½ U



Thermal Characteristics: M.2 vs. 2.5”

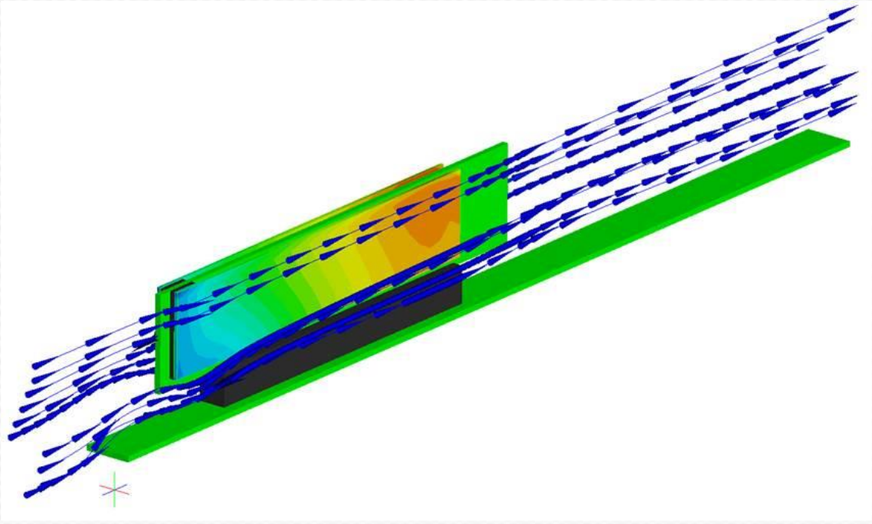
36% less power per server

- 8 x 8 W
- 4 x 25 W



Better airflow

- located in parallel with the DIMMs
- not in main path to processors



Outline

M.2 Background

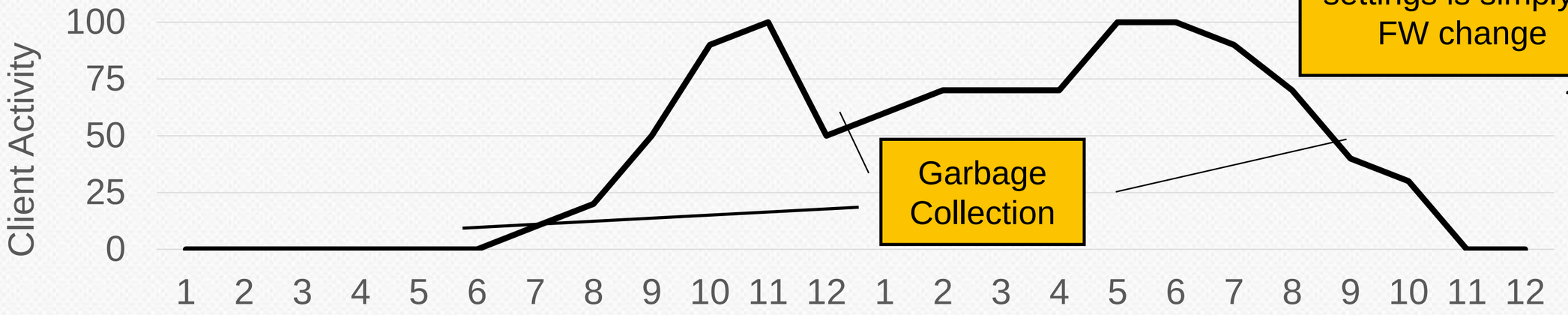
The M.2 Advantage

- Economy of Scale
- Form Factors in Brief
- Modularity, Density and Scalability
- Thermals: Airflow, Power and Density

Cloud-Specific Requirements

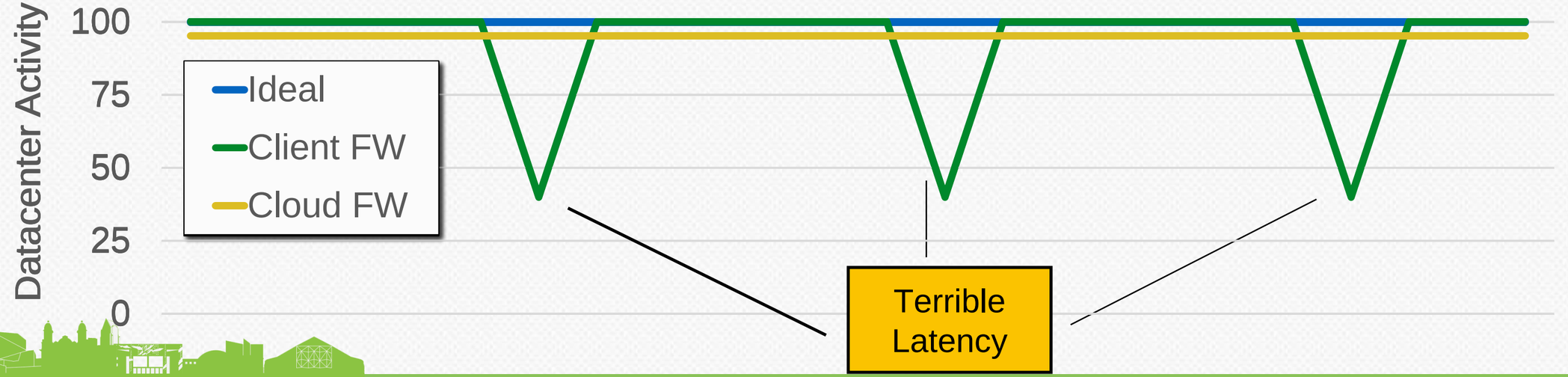
Duty Cycle and Latency Consistency

Cloud Workload != Client Workload

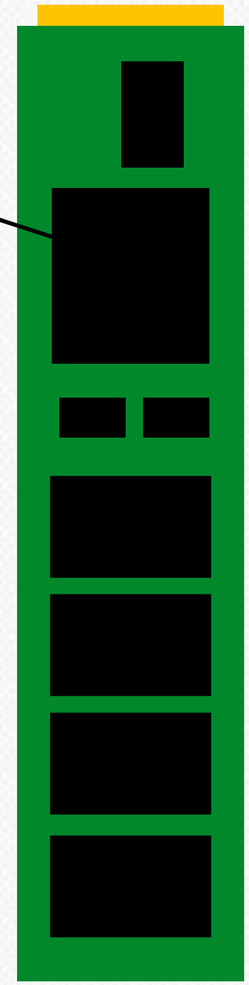


Adjusting these GC settings is simply a FW change

Garbage Collection



Terrible Latency



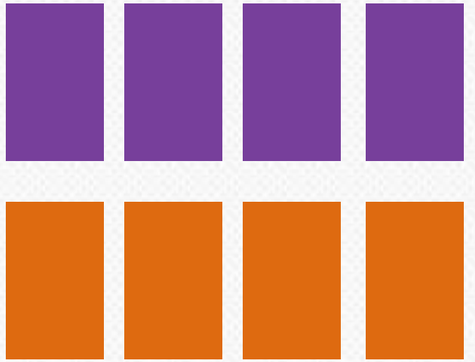
Endurance

Two Endurance Levels, Same Hardware Design

2 Application classes:

(DWPD == “Drive Writes Per Day”)

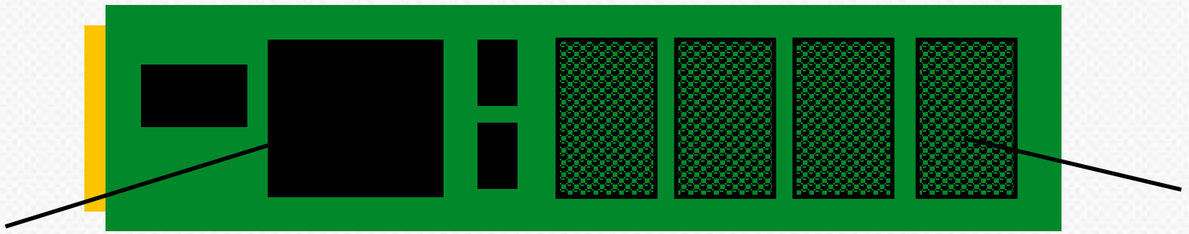
- Low Cost: 0.5 DWPD for 3 years
- High Endurance: 3.0 DWPD for 3 years



3k P/E Cycle NAND
$

15k P/E Cycle NAND
$$$

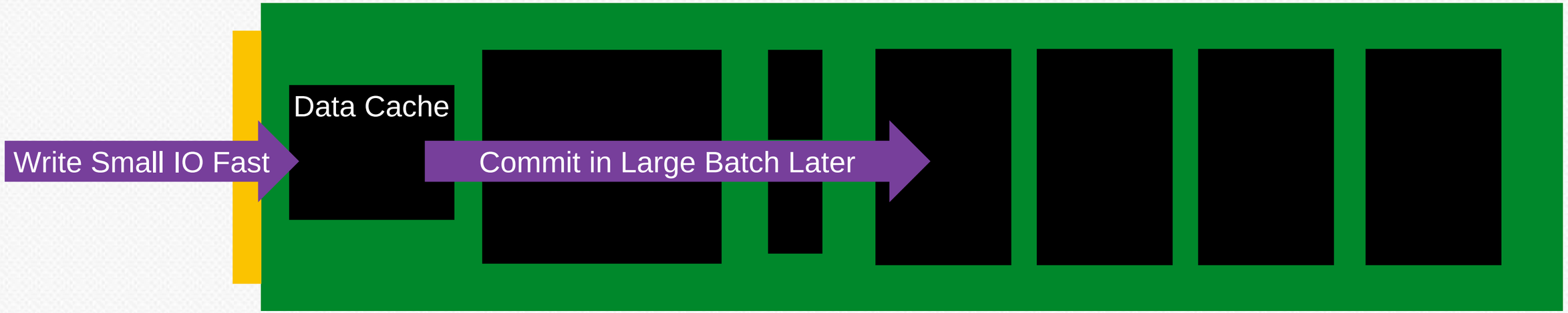
Each NAND type needs its own qualification and FW settings



NAND of different endurance levels are pin-compatible

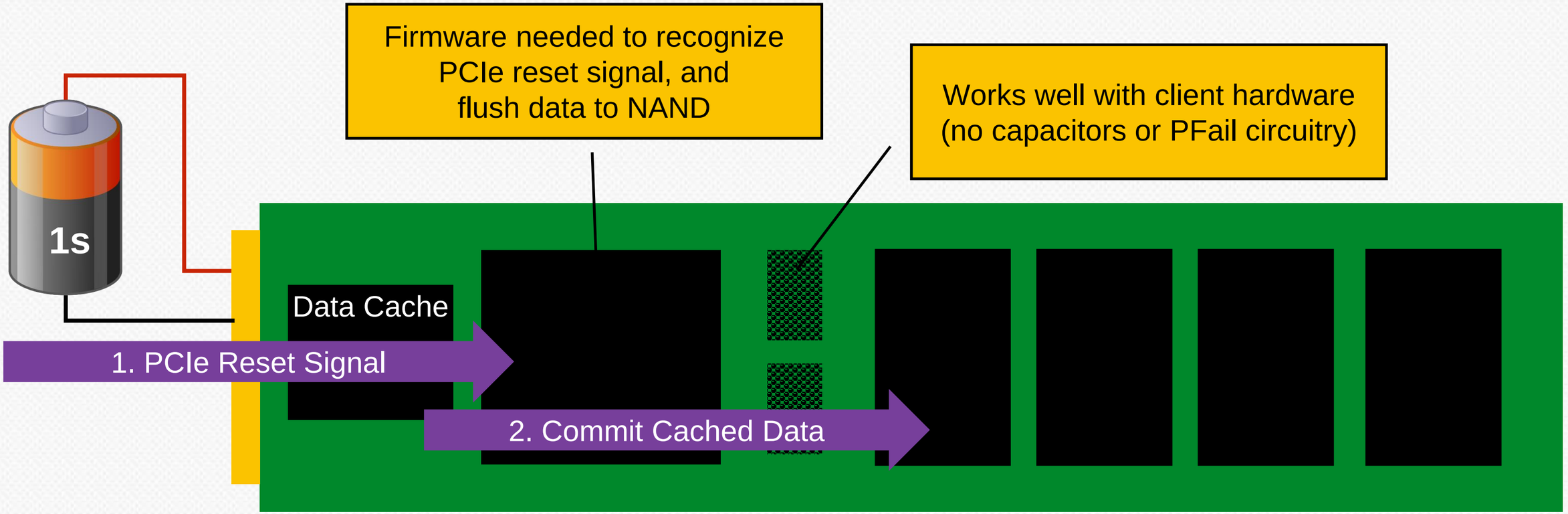
Protecting Data During Power Loss

Typical Operation (Consistent Power Supply)



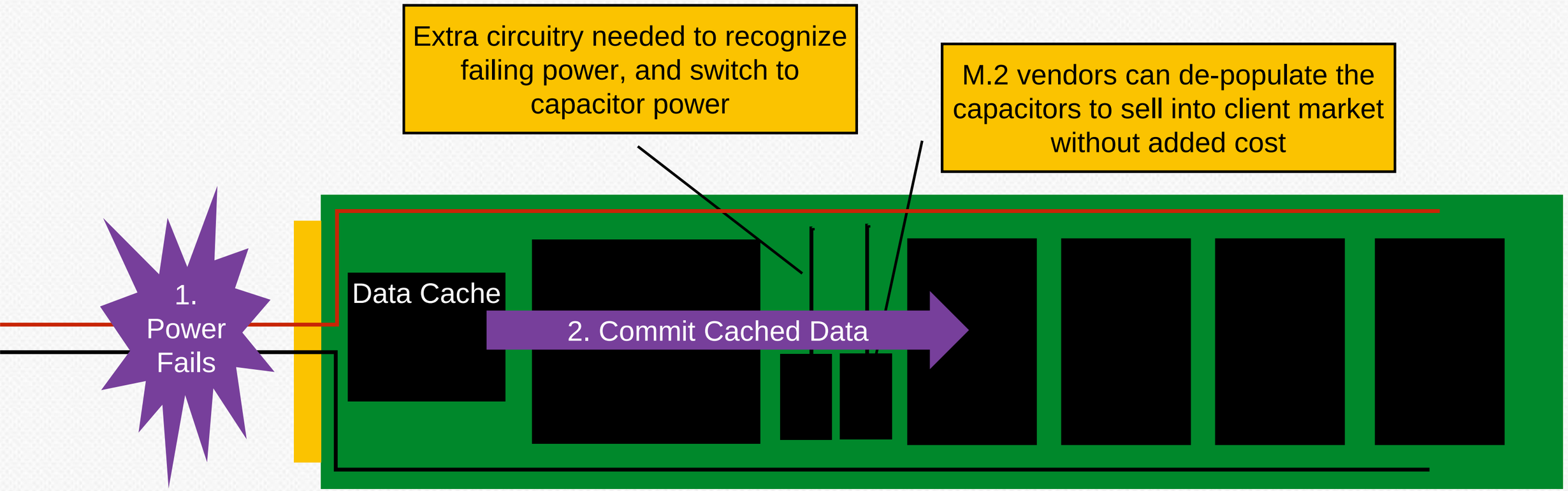
Protecting Data During Power Loss

Power Failure Option 1: Hold-up Power from System Battery



Protecting Data During Power Loss

Power Failure Option 2: Hold-up Power from Capacitors



Assorted Other Relaxations

- Lower retention (2 weeks vs. 3 mo., 1 yr)
  - Can help increase the NAND endurance
  - Or reduce the refresh rate in high temp. environment

- Top-side z-height: allow for PFail capacitors

Conclusion

The M.2 Advantage

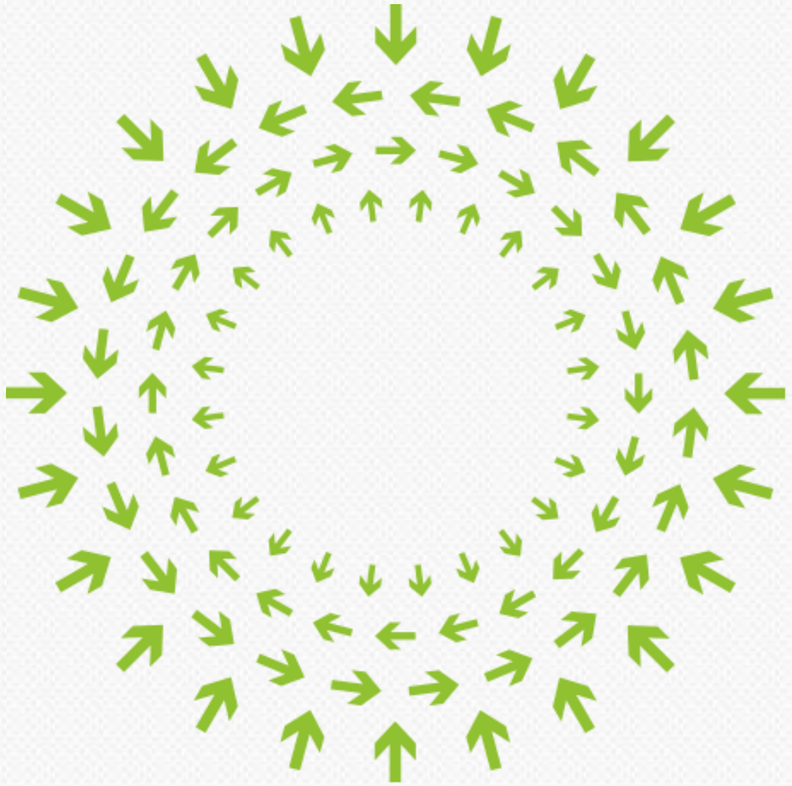
- Economy of Scale
- Form Factors in Brief
- Modularity, Scalability, Density
- Thermals: Airflow, Power, Density

Cloud-Specific Requirements

- Minimal impact to design and manufacturing
- Power Failure – FW Change, or BOM loading
- NAND characteristics – pin-compatible NAND & FW change

Visit our Booth for a Demo

Microsoft



OPEN Compute Summit

March 10–11, 2015

San Jose