Selecting the Right Enterprise SSD: Highlighting the M.2 Form Factor Part II

Richard A. Mataya

Co-Founder and SVP

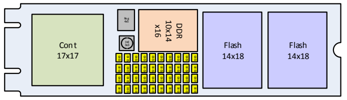
NxGn Data, Inc.

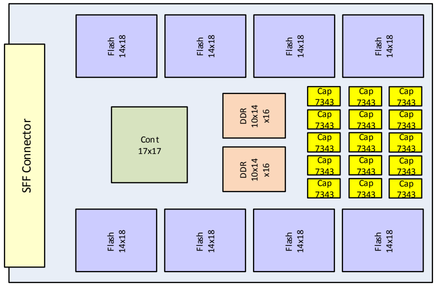
Enterprise SSD Options

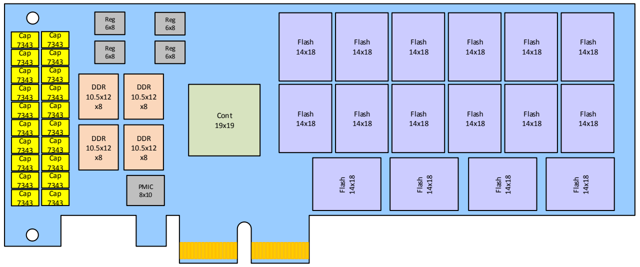
- Form Factor Options
  - Overview of M.2 and SFF
  - Cross section and volume overview
- Enterprise SSD Tradeoffs
  - Importance of “right-sizing” characteristics
- Thermal Management
  - Stack-up analysis
  - M.2 Challenges
  - System Example

Current Enterprise Options

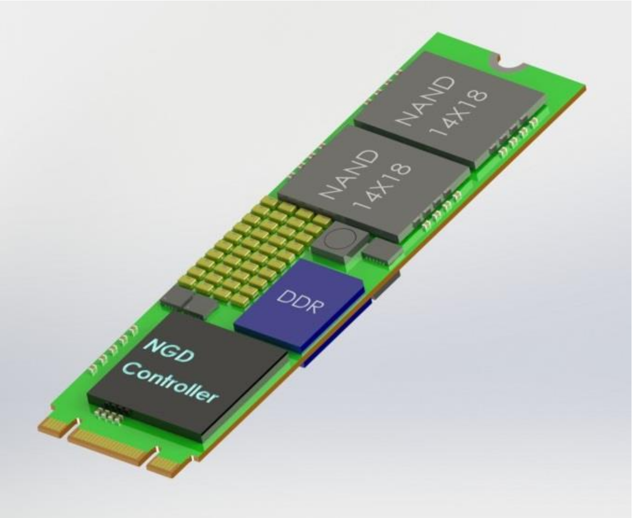
- M.2
  - PCIe or SATA Card format without frame
  - 22mm wide; 42, 60, 80 or 110mm long
  - 1.5mm clearance over 0.8mm PCB
- SFF
  - 2.5-inch with PCIe, SAS or SATA
  - 5mm, 7mm, 9.5mm or 15mm height
  - Metal or plastic enclosure
- PCIe Card
  - Half Height and Full Height
  - Heat sink – active or passive
- LFF
  - 3.5-inch and custom





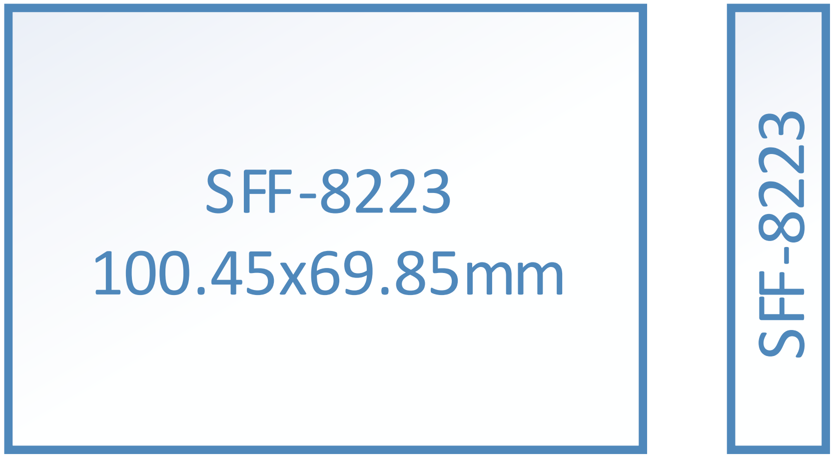


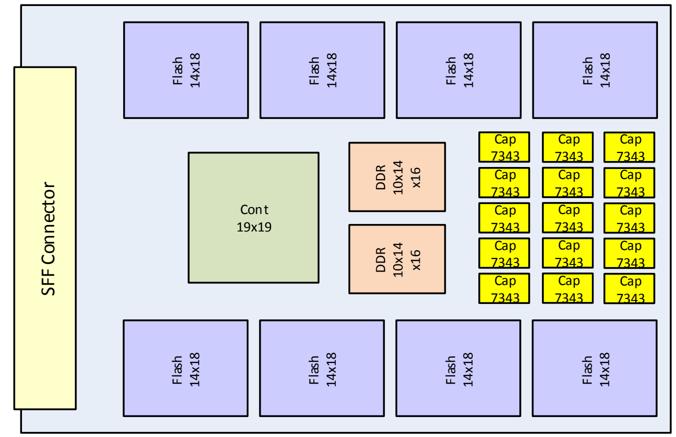
M.2 SSD Overview



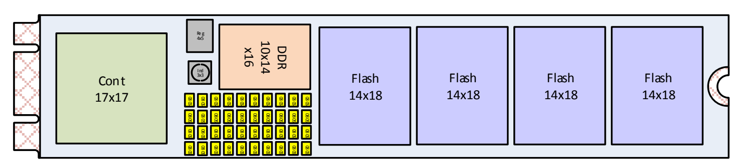
- PCIe or SATA
  - Up to 4-lanes of PCIe
- Frameless - 8.25W maximum power
- 22mm wide; various lengths
  - 42mm, 60mm, 80mm and 110mm
- 0.8mm PCB
  - Restricts signal layer count
- 1.5mm maximum component clearance on each side
  - Impacts controller, magnetics and backup
- Keying - two types (or both)
  - B-key – multifunction with SATA or 2-lane PCIe
  - M-key – 4-lane PCIe

Compare SFF and M.2





5mm to 15mm thick

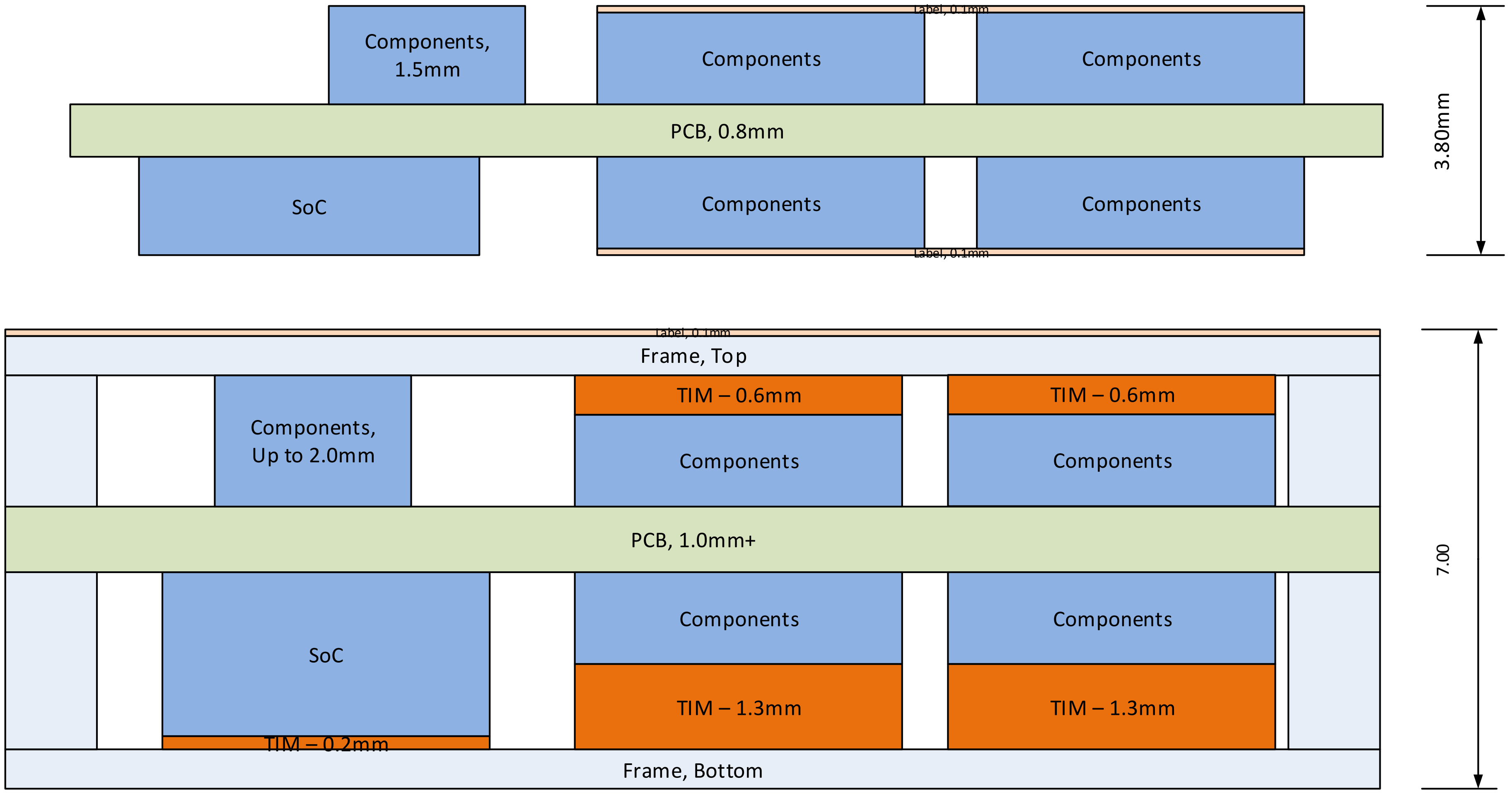


3.2mm or 3.8mm thick

A Note on SFF

- Combining SFF standards
  - 2.5-inch PCIe/NVMe (PCI-SIG)
    - SFF-8201 - Form Factor of 2.5" Disk Drives
    - SFF-8223 - 2.5" Drive Form Factor with Serial Conn
    - SFF-8639 - Multifunction 12 Gb/s 6X Unshielded Conn
  - 3.5-inch PCIe/NVMe – "LFF"
    - SFF 8301 – Form Factor of 3.5" Disk Drives
    - SFF 8323 – 3.5" Drive Form Factor with Serial Conn
    - SFF 8639 – Multifunction 12 Gb/s 6X Unshielded Conn

Stack-up - M.2 v. SFF 7mm

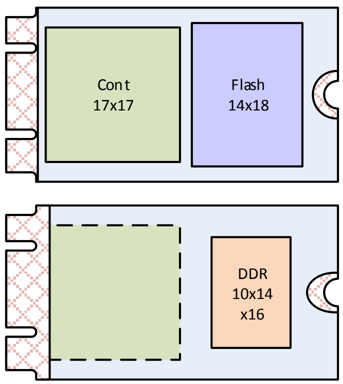


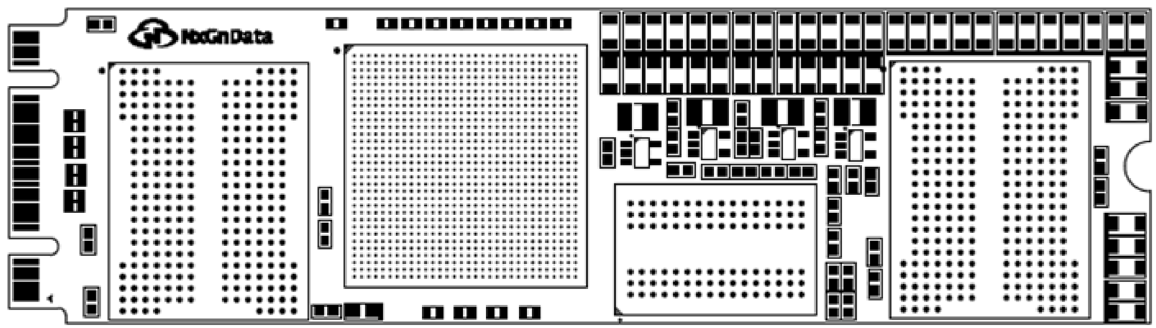
SFF SSD Volume

2.5-inch SFF Parameters				
	5mm	7mm	9.5mm	15mm
Volume (cm^3)	35.1	49.1	66.7	105.2
Power (W)[1]	Low	Med	Med	High
Placements	8-16	16	16+	32
Capacity[2] (TB)	4TB	4TB	4TB+	8TB

[1]Low=5-9W; Med=7-15W; High=15-25W
[2]256Gbit die - ODP

M.2 SSD Volume



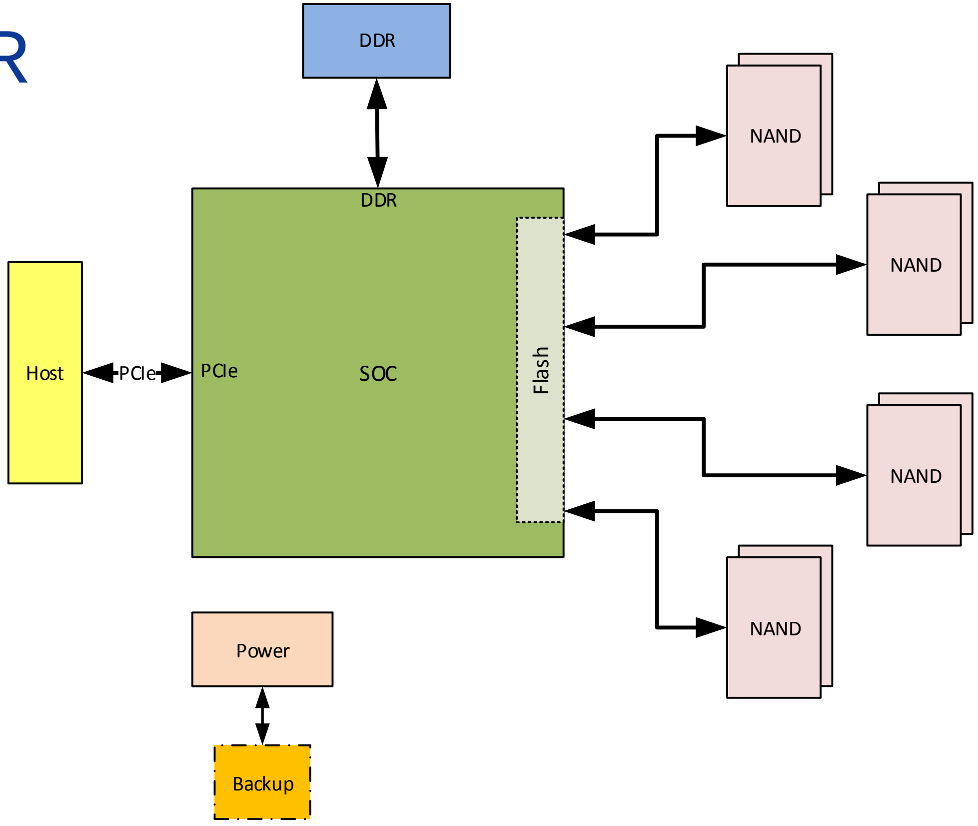


M.2 SSD Parameters				
	2242	2260	2280	22110
Volume (cm^3)	3.51	5.02	6.7	9.2
Power (W)	8.25W	8.25W	8.25W	8.25W
Placements	1-2	2-4	4-8	8-10
Capacity* (TB)	256GB	1TB	2TB	2TB+

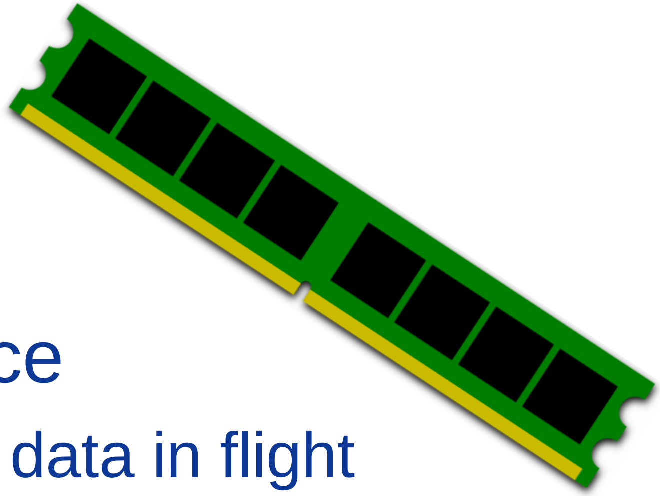
*256Gbit die - ODP

Enterprise SSD Decisions

- To DDR or not DDR
- Flash Implementation
- SoC Choices
  - Rightsizing
- Power Backup
- Power Regulation

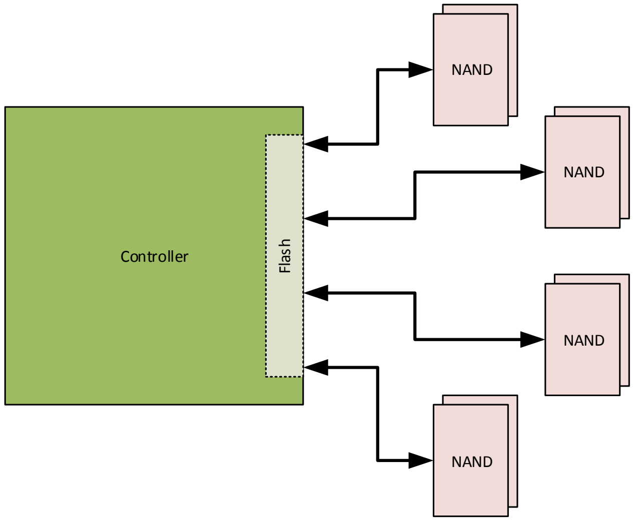


DDR on SSDs

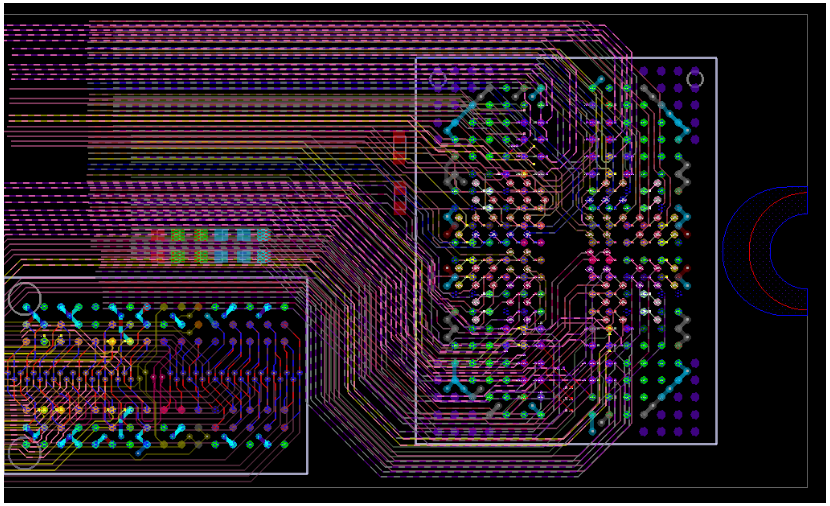


- Directly impacts performance
  - Typically used for tables, not data in flight
  - Power on to ready timing may increase
- Requires PCB space and power
  - May reduce final storage capacity
- Often necessitates power backup
- BOM cost impact

Flash



- Channel Count
  - Capacity
  - PCB routing
  - Controller I/O
- Die count on each channel
  - Available Placements
  - Interleaving
- Performance
  - Matched to SSD and controller requirements
  - Overall power impact
  - Page/Block Availability



SoC Selection/Design

- Host interface requirements
- Physical size requirements
  - PCB real estate (width)
  - Component height restrictions
- Power consumption
  - Product Requirements
- Matched Capabilities
  - Right-sizing the controller

Right Sizing the Controller

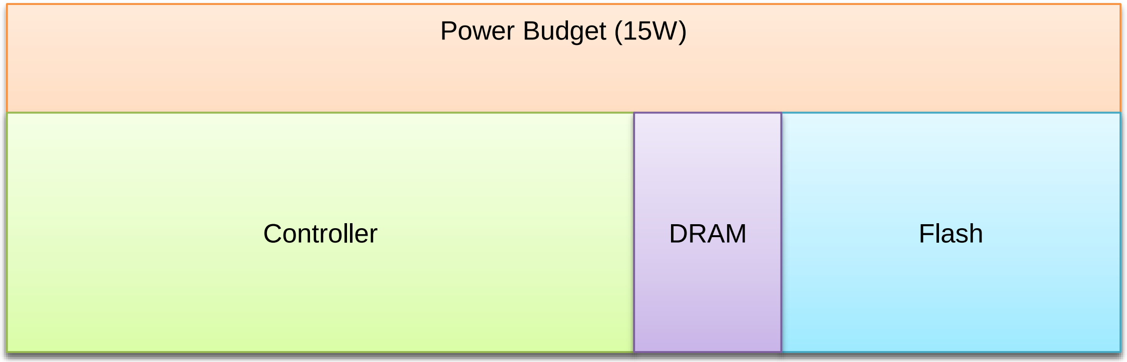
- Key selection criteria
  - Performance
    - Meets performance objective for throughput and latency
  - Power
    - Meets the power objective of the entire system for expected workloads
  - Endurance and reliability
  - Matches peripheral capabilities
  - Meets (or beats ☺) the cost objective

Example – Power Compromise

- Flash Basic Assumptions
  - Power consumption is directly related to performance
    - ~100mW per active flash die
    - For writes, 2,500 IOPs/die entitlement
    - For reads, 15,000 IOPs/die entitlement
- Flash die utilization depends on architecture
  - Controller capability
  - Number of flash channels
  - Channel performance limitations
  - Dice per channel - loading
  - FTL and GC

Overweight Controller

- Overdesigned/over power controller
  - Too little power left for flash to meet performance goal



Entitlements

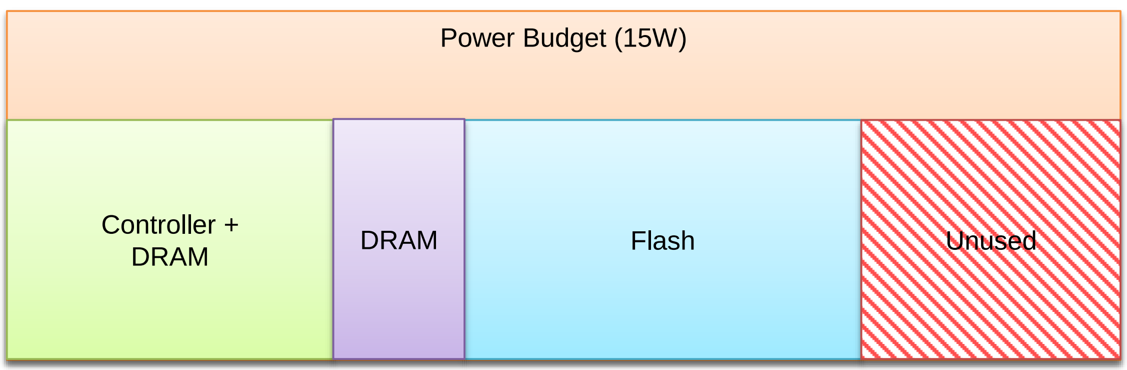
SoC – 1500kIOPs @ 15W

Flash – 750kIOPs @ 5W
1000kIOPs @ 7W
1500kIOPs @ 10W

Actual Performance
750kIOPs @ 15W

Underweight Controller

- Under designed/under power controller
  - Overall performance gated by the controller



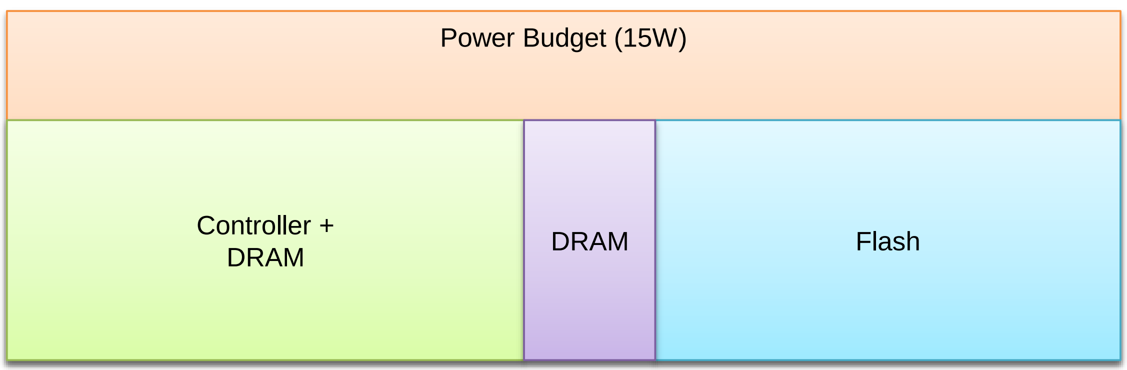
Entitlements
SoC – 600kIOPs @ 6W

Flash – 750kIOPs @ 5W
1000kIOPs @ 7W
1500kIOPs @ 10W

Actual Performance
600kIOPs @ 12W

Right-sized Controller

- Controller performance matches:
  - Flash entitlements
  - System power requirements



Entitlements

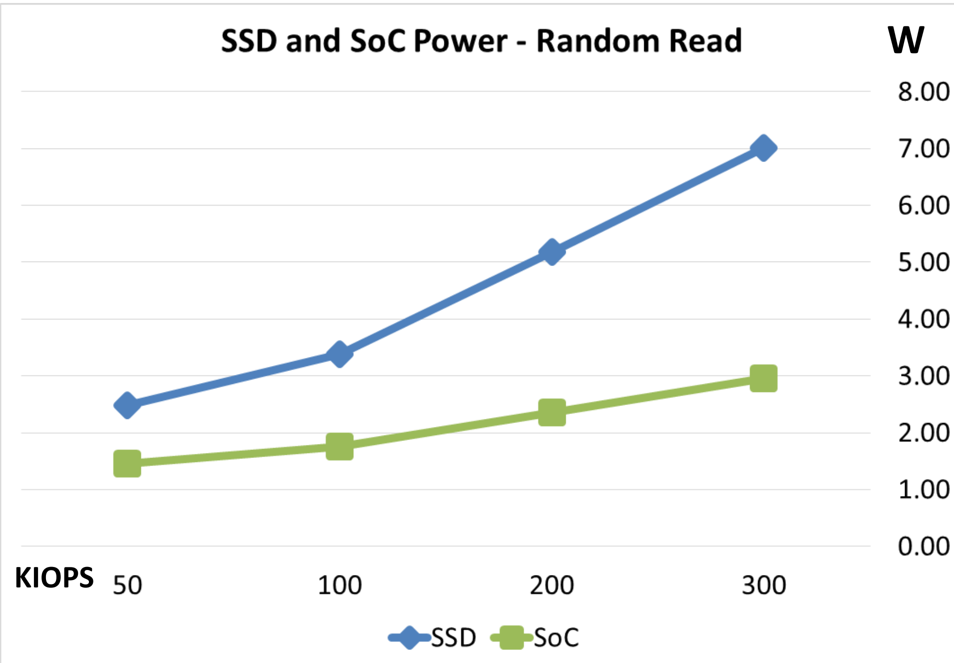
SoC – 1000KIOPs @ 7W

Flash – 750kIOPs @ 5W
1000kIOPs @ 7W
1500kIOPs @ 10W

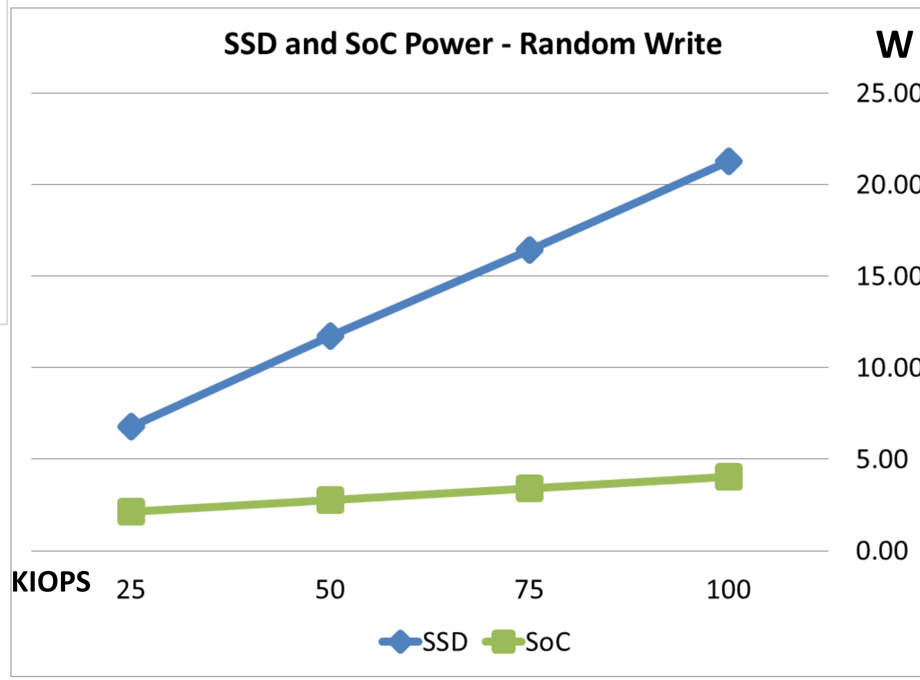
Actual Performance

1000kIOPs @ 15W

Power Estimation



Assumptions:
- Idealized MLC Flash specs
- Assumed WA of 3
- 4K aligned transactions



Analytical Model

Power Backup

- Needed as an enterprise requirement
  - Complete in-process flash commands (ECC)
  - User data in flight
  - Saving FTL
- Key Impact
  - Impacts available PCB placements
  - Increased BOM cost
  - Reliability (component selection)
  - Power-on ready timing

Regulator Efficiency Tables

- Power inefficiencies in regulator = lost performance
  - 25W example 5% = 160k Read IOPs
  - 10W example 5% = 60k Read IOPs
- Higher efficiency = higher $$$

Assumptions	IOPs	Power
Read Metric	12,000 IOPs	0.105W
Write Metric	2,400 IOPs	0.120W

Power	Efficiency	Reg Waste	Read Loss	Write Loss
25W	95%	1.25W	142,857 IOPs	24,000 IOPs
25W	94%	1.50W	171,429 IOPs	28,800 IOPs
25W	93%	1.75W	200,000 IOPs	33,600 IOPs
25W	92%	2.00W	228,571 IOPs	38,400 IOPs
25W	91%	2.25W	257,143 IOPs	43,200 IOPs
25W	90%	2.50W	285,714 IOPs	48,000 IOPs
25W	89%	2.75W	314,286 IOPs	52,800 IOPs
25W	88%	3.00W	342,857 IOPs	57,600 IOPs
25W	87%	3.25W	371,429 IOPs	62,400 IOPs
25W	86%	3.50W	400,000 IOPs	67,200 IOPs
25W	85%	3.75W	428,571 IOPs	72,000 IOPs

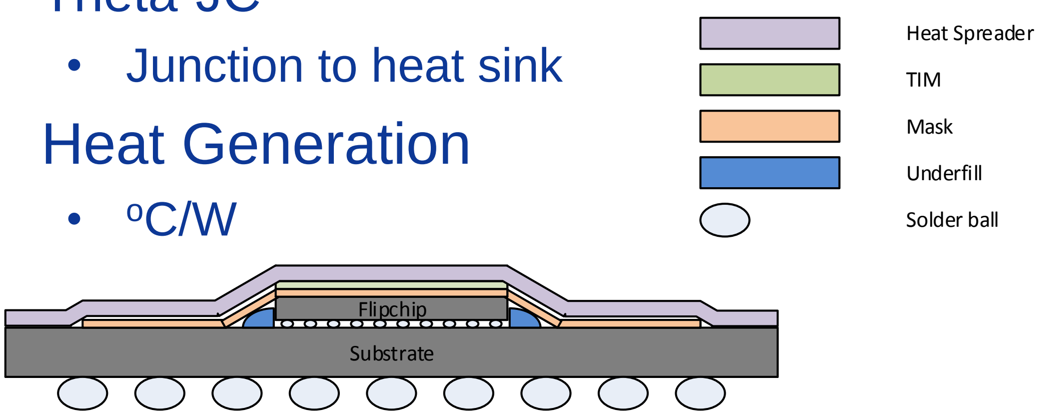
Power	Efficiency	Reg Waste	Read Loss	Write Loss
10W	95%	0.50W	57,143 IOPs	9,600 IOPs
10W	94%	0.60W	68,571 IOPs	11,520 IOPs
10W	93%	0.70W	80,000 IOPs	13,440 IOPs
10W	92%	0.80W	91,429 IOPs	15,360 IOPs
10W	91%	0.90W	102,857 IOPs	17,280 IOPs
10W	90%	1.00W	114,286 IOPs	19,200 IOPs
10W	89%	1.10W	125,714 IOPs	21,120 IOPs
10W	88%	1.20W	137,143 IOPs	23,040 IOPs
10W	87%	1.30W	148,571 IOPs	24,960 IOPs
10W	86%	1.40W	160,000 IOPs	26,880 IOPs
10W	85%	1.50W	171,429 IOPs	28,800 IOPs

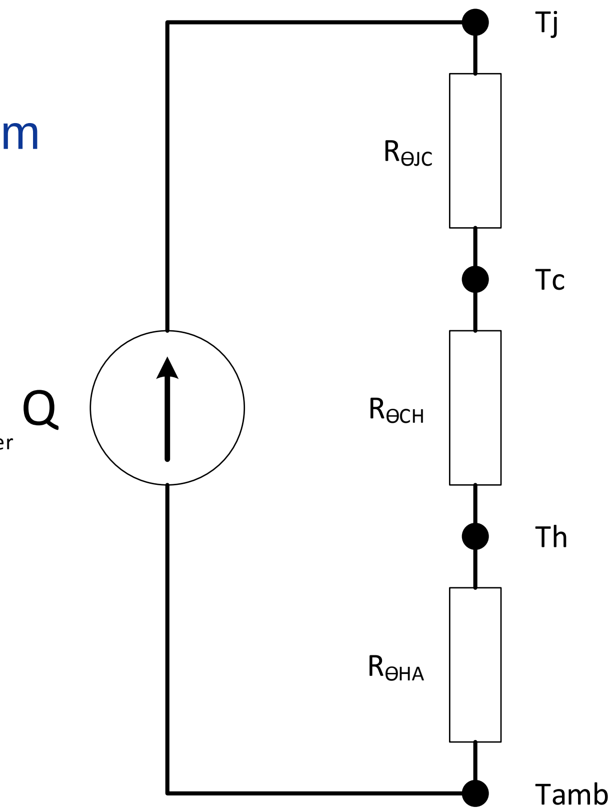
Thermal Management

- Requirements
  - Environmental considerations
    - Temperature
    - Airflow
  - Form Factor Restrictions
  - Performance Requirements
- Stack-up Analysis
  - Tolerance
  - Thermal Efficiency
- Tj of every component must be considered

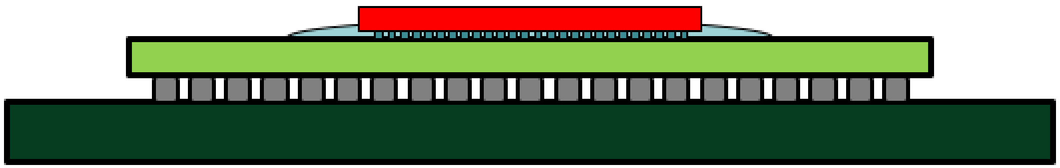
Thermal – Ɵ Measurements

- Theta-JA
  - Thermal resistance of a device package
  - Per JEDEC, measures temp difference from junction to ambient air
  - $8^{o}C/W$ to $50^{o}C/W$
  - Various PCB types
- Theta-JC
  - Junction to heat sink
- Heat Generation
  - $^{o}C/W$

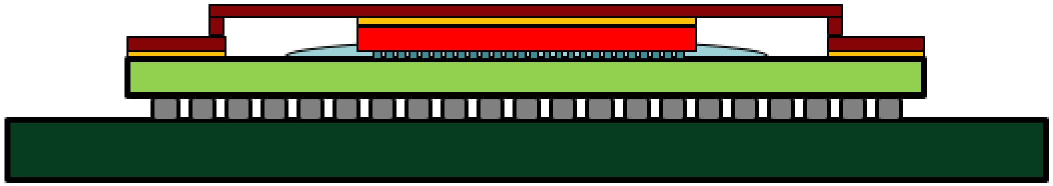




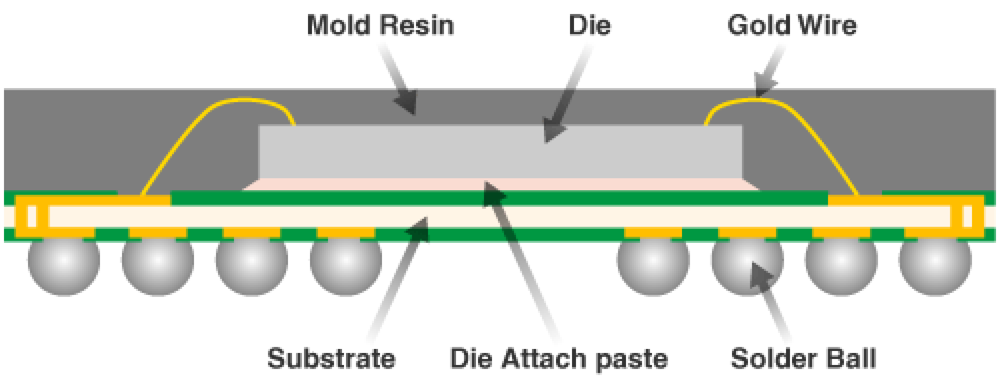
SoC package types



FCBGA



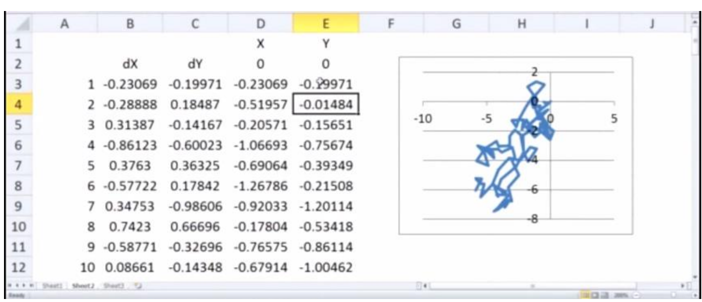
HFCBGA



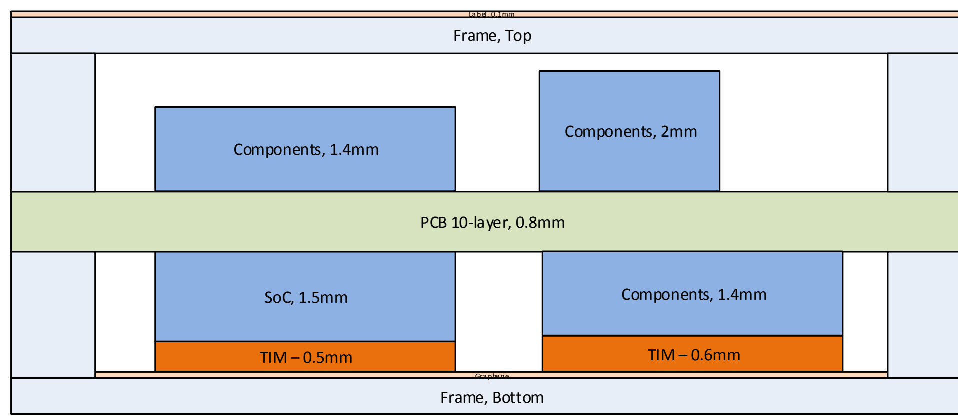
FBGA

Thermal Stack-up

- Critical to extracting heat
- Mechanical Tolerance must be accounted for
  - Monte Carlo Analysis
- Very challenging for M.2
  - System level Issue



	A	B	C	D	E
1				X	Y
2		dX	dY	0	0
3	1	-0.23069	-0.19971	-0.23069	-0.19971
4	2	-0.28888	0.18487	-0.51957	-0.01484
5	3	0.31387	-0.14167	-0.20571	-0.15651
6	4	-0.86123	-0.60023	-1.06693	-0.75674
7	5	0.3763	0.36325	-0.69064	-0.39349
8	6	-0.57722	0.17842	-1.26786	-0.21508
9	7	0.34753	-0.98606	-0.92033	-1.20114
10	8	0.7423	0.66696	-0.17804	-0.53418
11	9	-0.58771	-0.32696	-0.76575	-0.86114
12	10	0.08661	-0.14348	-0.67914	-1.00462

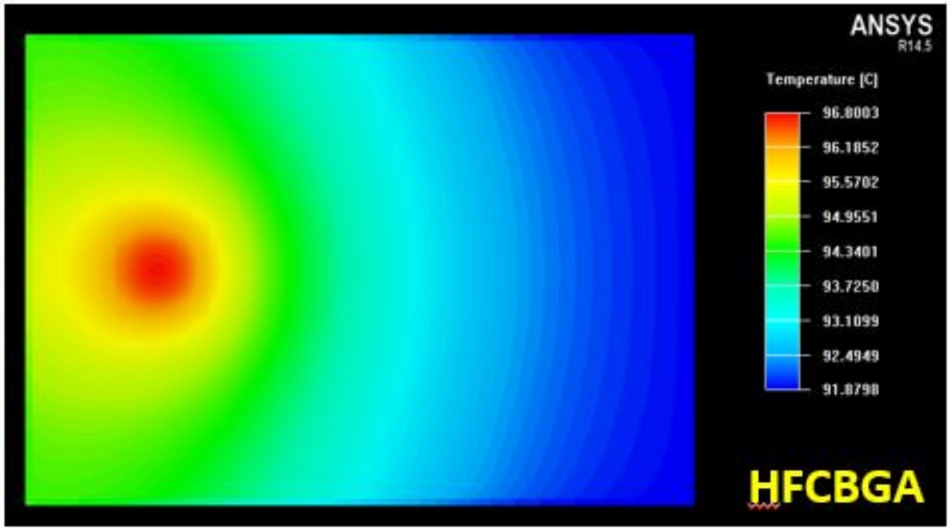


Stack-up Tolerance

- Tolerance
  - Understand component parametrics
  - Variation “adds up”
- TIM – Thermal Interface Material
  - Compression key to heat transfer
  - Without compression, it’s an insulator
  - Too much compression = mech deformation
- Thermal Conductivity (W/mk)
- Critical heat path
- Advanced materials

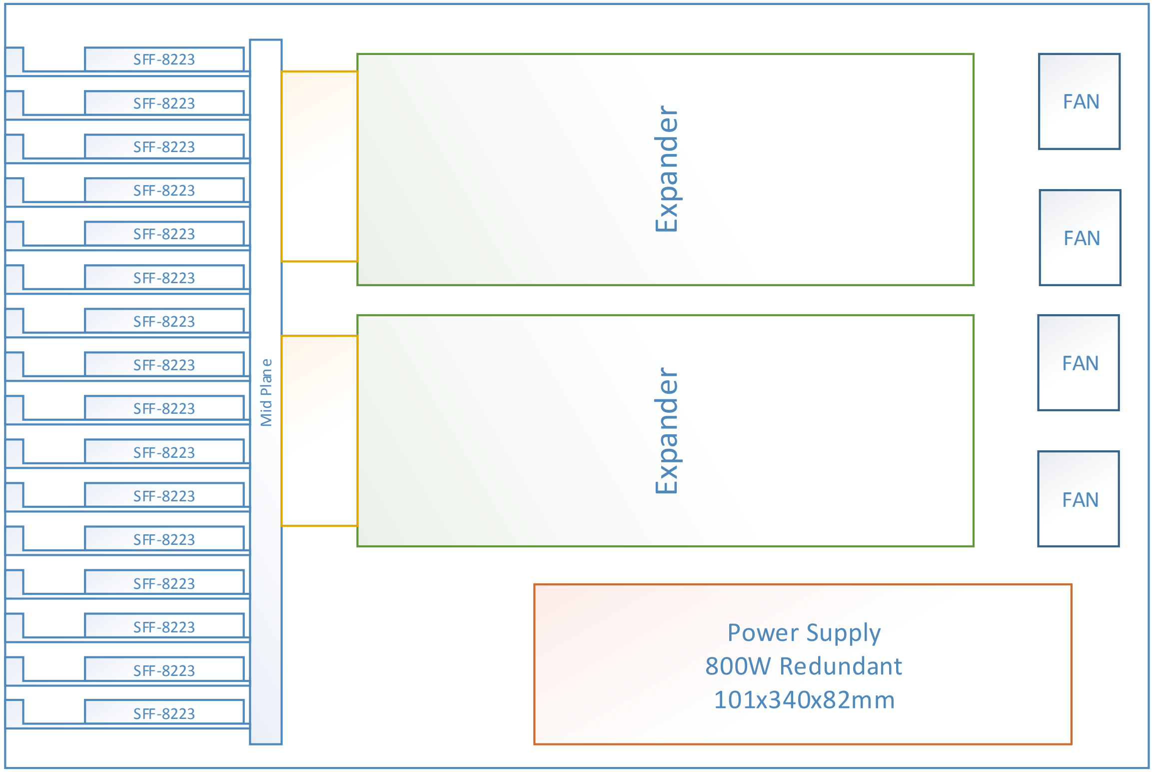
Real Answer – Thermal Modeling

- SoC
- Other components
- Housing (if any)
- PCB
- Airflow & Channel
- Ambient conditions
- System enclosure



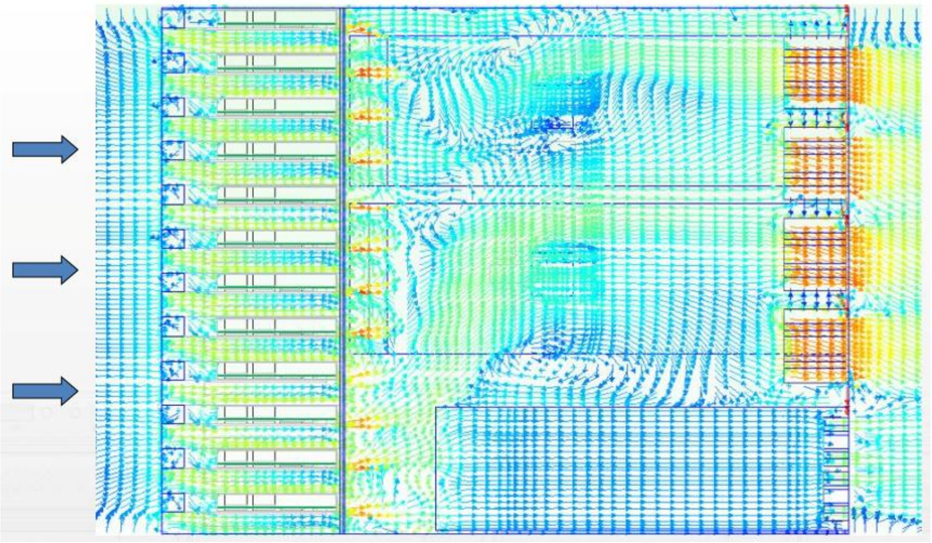
System Deployment

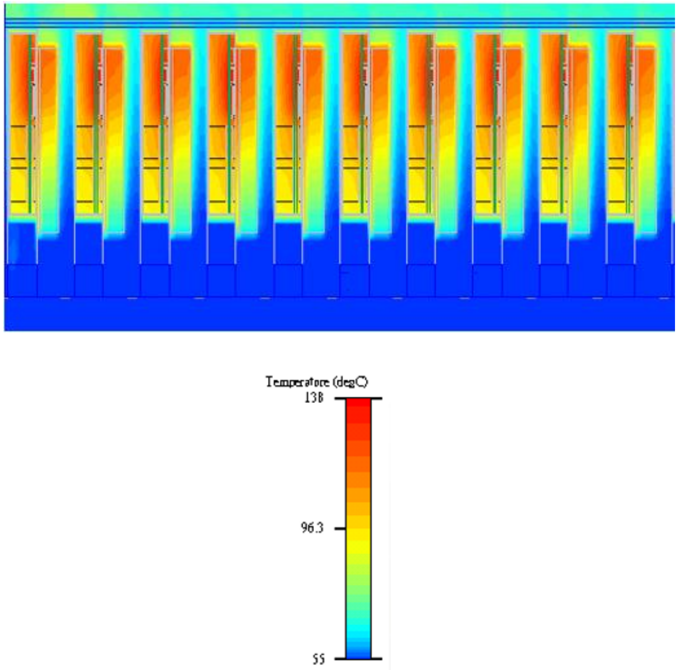
- 2U Rackable SSD Drive Array



System Thermal Results

- Simulations for airflow and heat





Key Takeaways

- Understand your requirements
  - It's not just about performance
  - Take a holistic approach
- Use leading indicators
  - Analyze and simulate
  - Do the heavy lifting early
  - Underpin everything
- M.2 suitable for enterprise applications

NxGn Data, Inc.

- Visit us at booth 707