STORAGE DEVELOPER CONFERENCE

SDC 23

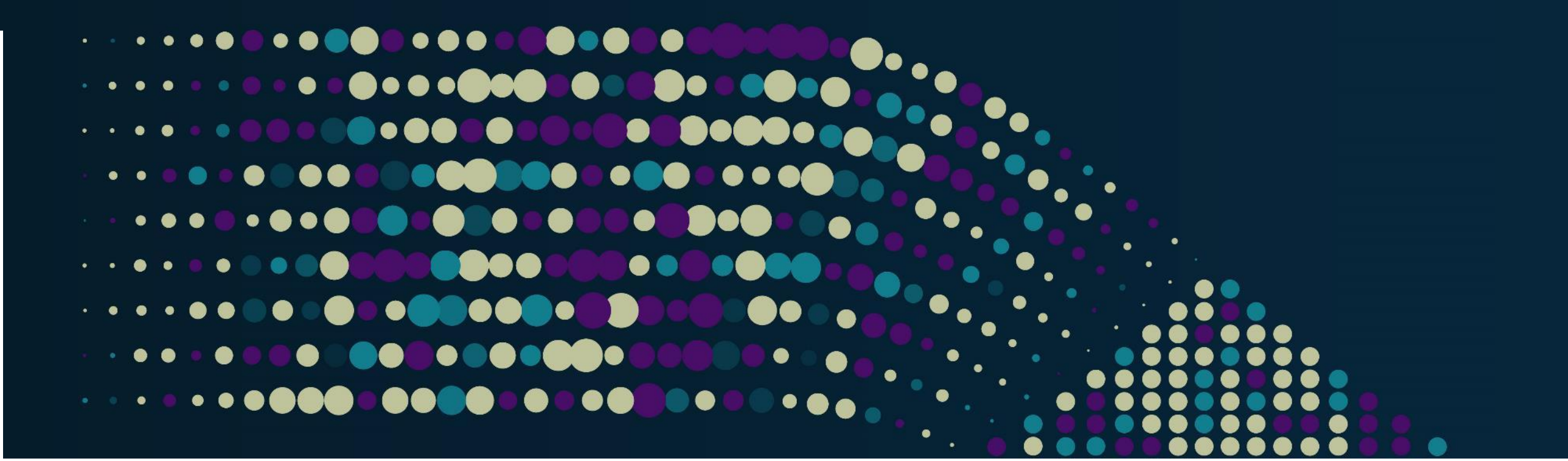
BY Developers FOR Developers

Maximizing EDSFF E3 SSD Design

Trent Johnson
FlashCore Hardware Architect
trejo@us.ibm.com

Trent Johnson is a Hardware Architect at IBM, with a focus on the IBM FlashCore Module. He joined IBM as part of the Cleversafe Acquisition where he was the System Hardware Architect of exabyte-scale Object Storage. Prior to Cleversafe, he developed system-level manufacturing and test solutions for AMD CPUs and GPUs where he was awarded the AMD Corporate Technical Achievement Award.

He has 24 years of industry experience, holds 7 US patents and has presented at the Burn-in and Test Socket Workshop, Flash Memory Summit as well as the Conference for Consumer Electronics. He earned BSEE and MSEE degrees from The University of Texas at Austin in Electrical Engineering with a focus on Manufacturing System Engineering.



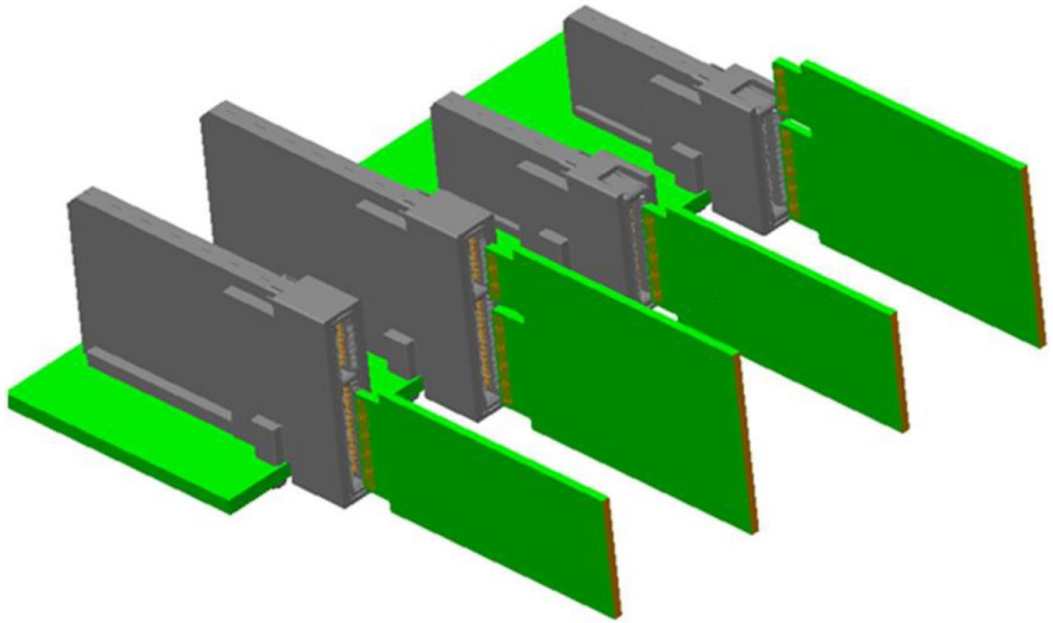
Why migrate to EDSFF?

Key Benefits of Enterprise Datacenter Standard Form Factor (EDSFF)

- Connector
  - Good for high amperage
  - Low cost on card
  - High Speed by design
    - PCIe® 5.0/6.0
    - Very high lane counts (up to 16)

- Thermal Architecture
  - Each Form Factor has realistic and achievable thermal performance targets
  - Density is foremost in the standard
  - Generous device power budgets

- Device Flexibility
  - It's a "Standard" form factor, not SSD form factor
  - Peripherals of almost any kind may be used



Relevant Specs:
SFF-TA-1002 (Connector)
SFF-TA-1008 (Mechanical)
SFF-TA-1009 (Electrical)
SFF-TA-1023 (Thermal)

IBM FlashSystem

High-End: FlashSystem 9500

Mid-Range: FlashSystem 7300

Entry-Level: FlashSystem 5200

IBM FlashCore™ Module

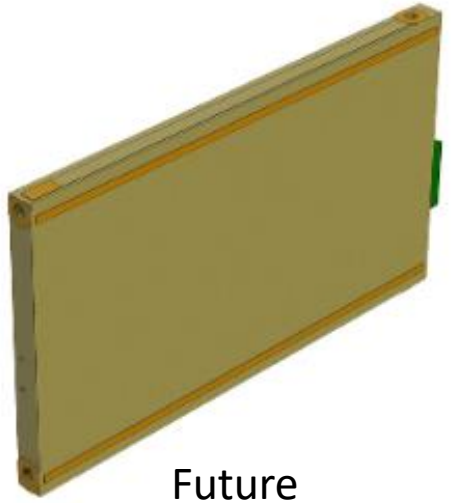
- Enterprise QLC storage
- Compression at speed
- Encryption at speed
- RAID assists

NVMe SSD

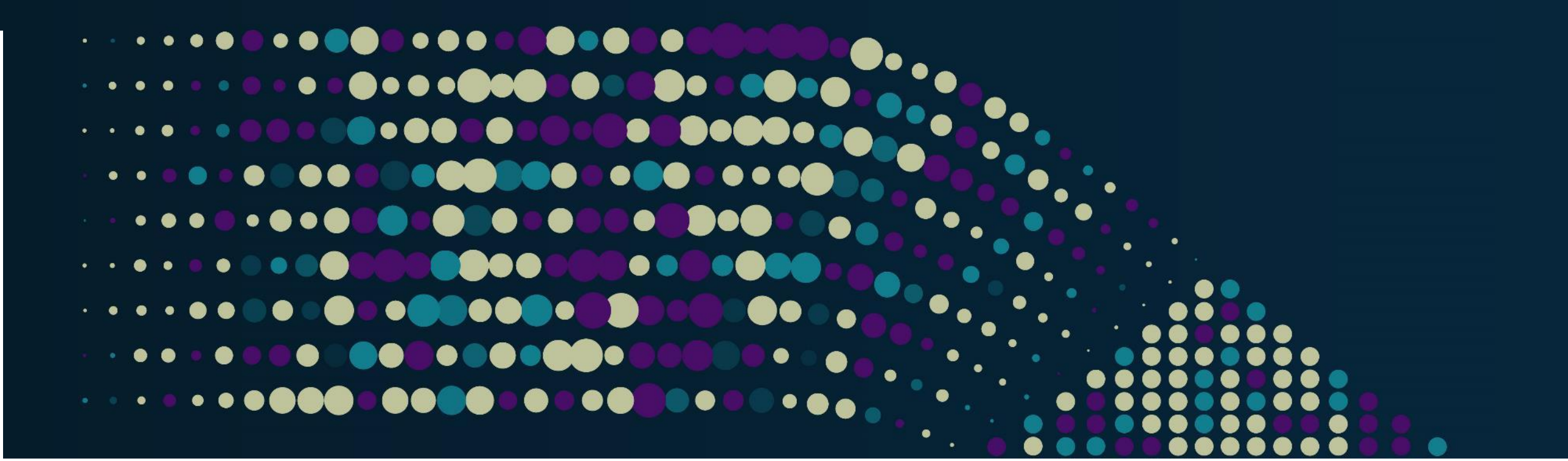
Benefits of EDSFF for IBM FlashSystem

- **Greater drive power envelope allows for more flexible design**
- **More efficient use of space allows greater enclosure density**
- **Fewer lane connections enabled by PCIe® 5**
- **Increased switch bandwidth enabled by PCIe® 5**
- **Simpler connector**
- **Slots are useful for more than just SSDs**
- **Increased adoption of E3 is driving volumes from U.2 to E3**

Today



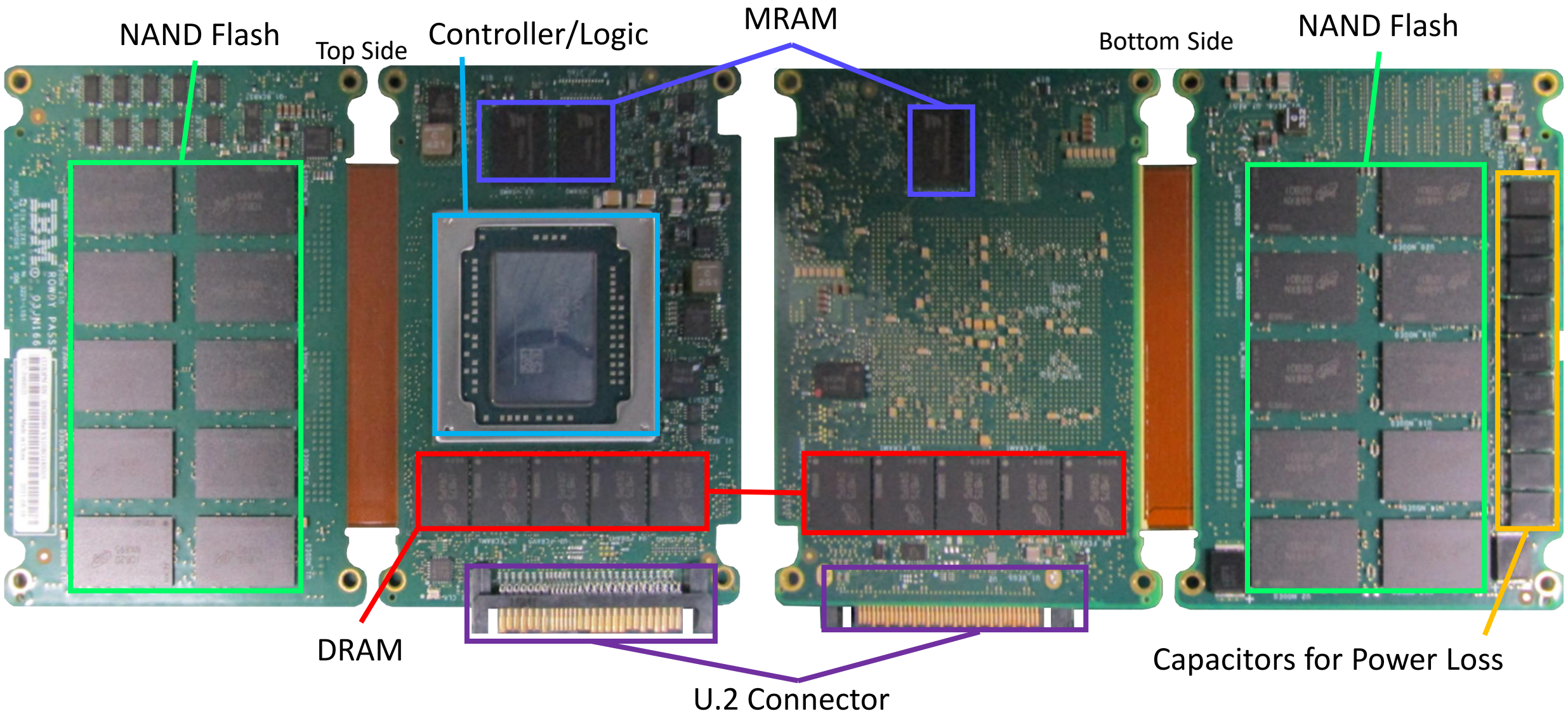
Future



SSD Design Methodology

FlashCore™ Module

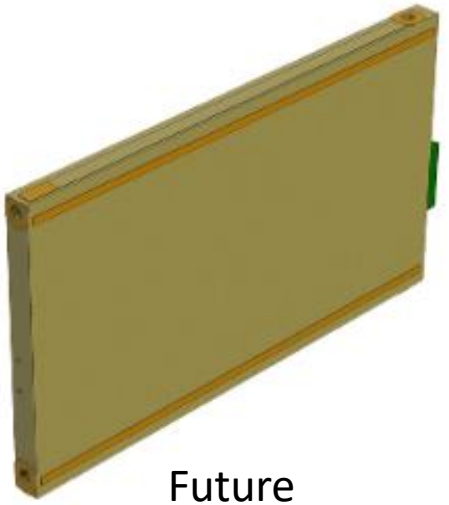
The Layout of Today's 3rd Gen IBM FlashCore™ Module



FCM Hardware Design Goals

- Migrate to a new industry standard, EDSFF E3
- Maximize the overall Terabytes per rack unit
- Minimize cost per Terabyte
- High Quality & Reliability
- Utilize an FPGA for advanced computational storage techniques like inline compression, encryption, RAID assists and future features

Today



Future

Exploring E3.S

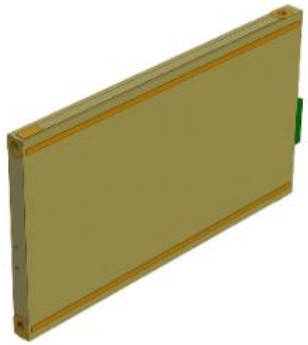
- E3.S: Targeted to NVMe SSDs
- Power budget up to 25W

- E3.S 2T: Targeted to be high performance SSDs, SCM, CXL
- Power budget up to 40W

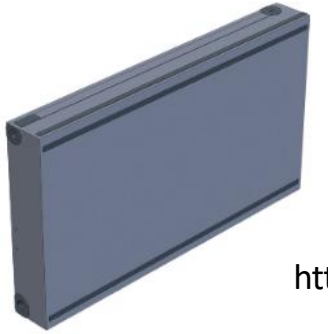
https://americas.kioxia.com/content/dam/kioxia/en-us/business/ssd/data-center-ssd/asset/KIOXIA_EDSFF_E3_Intro_White_Paper.pdf

- The FCM's FPGA and the goal of high capacity don't fit well in E3.S
  - Footprint too small
  - Power budget too small
- E3.S 2T might work for FCM
  - Close to U.2 dimensions
  - Good power budget

Exploring E3.L



- E3.L: Targeted to be a primary form factor for storage subsystems and server platforms requiring **maximum capacity** for each ‘U’
- Power budget up to 40W

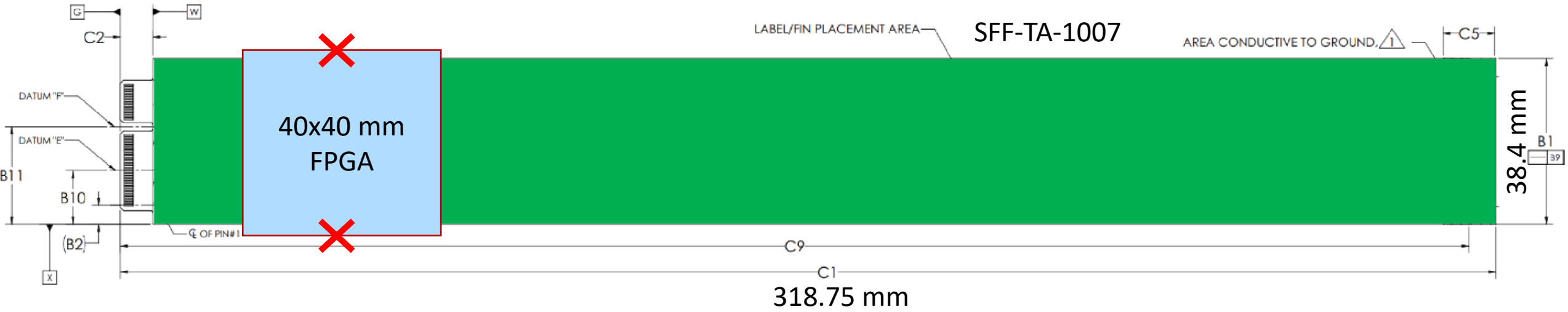


- E3.L **2T**: Targeted to **FPGAs** or accelerators (Computational Storage)
- Power budget up to 70W

https://americas.kioxia.com/content/dam/kioxia/en-us/business/ssd/data-center-ssd/asset/KIOXIA_EDSFF_E3_Intro_White_Paper.pdf

- What if your product is a mix of both use cases?
- If you don’t need 70W of power, you lose 50% of your density by using E3.L **2T** vs E3.L
- E3.L offers better enclosure density than E3.S 2T

FPGA on E1.L?

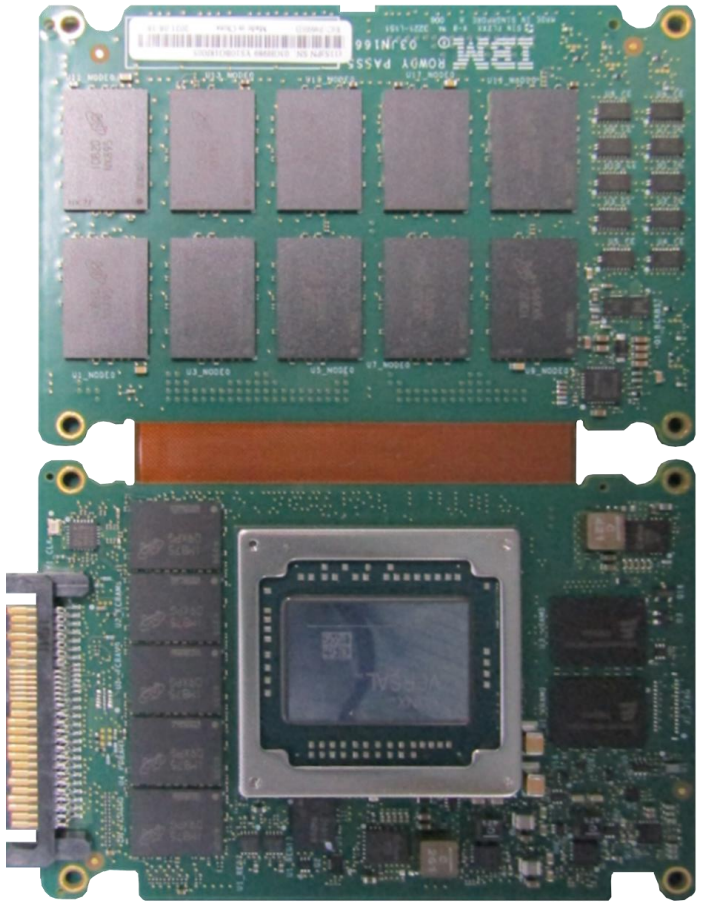


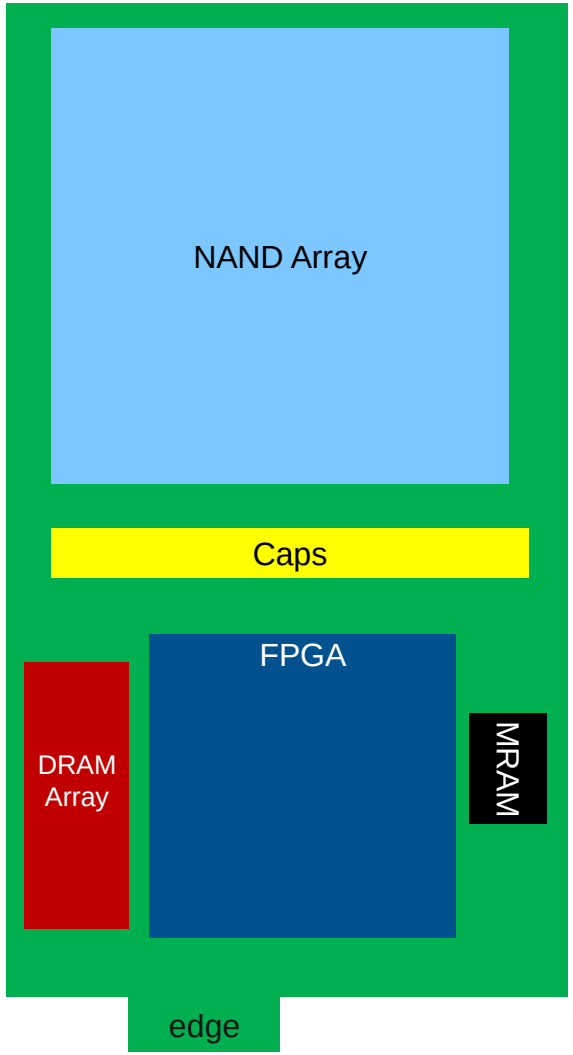
- E1.L comes in two sizes:
  - 9.5mm: up to 25W
  - 18mm: up to 40W
- 40mm x 40mm FPGAs clearly do not fit on a 38.4mm wide board
- 35mm x 35mm FPGAs (or smaller) are plausible, but signal exits can only go east/west
  - Complex routing
  - Reduced I/O for Flash

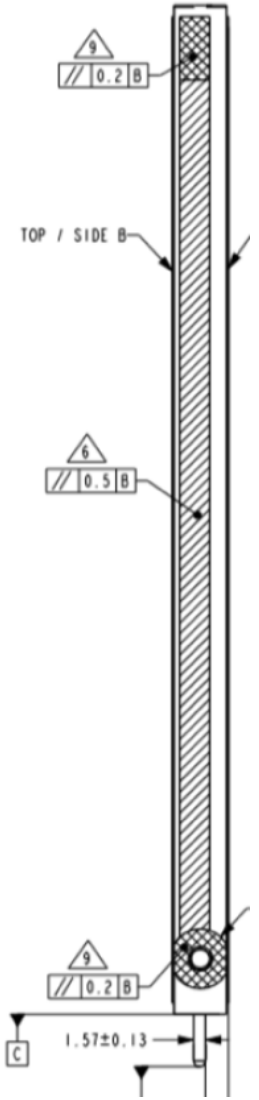
Let’s not forget E1.S

- Optimized for 1U
- 5 sizes:
  - 31.5 x 111.49 x 5.9 mm: 12W max
  - 31.5 x 111.49 x 8 mm: 16W max
  - 33.75 x 118.75 x 9.5 mm: 20W max
  - 33.75 x 118.75 x 15 mm: 25W max
  - 33.75 x 118.75 x 25 mm: 25W max
- E1.S use cases
  - Boot Media
  - Cache
  - Blade servers
  - Edge servers
  - AI/ML
  - High Performance
- Rear-plug devices may operate in HT-LF thermal space (> 50°C)
- Not really optimal for high capacity FCM

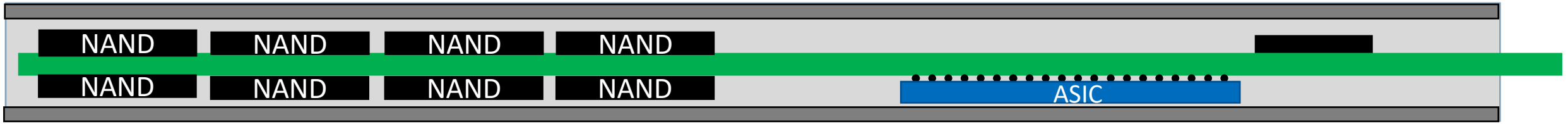
FlashCore™ Module E3.L Floor Planning Concept





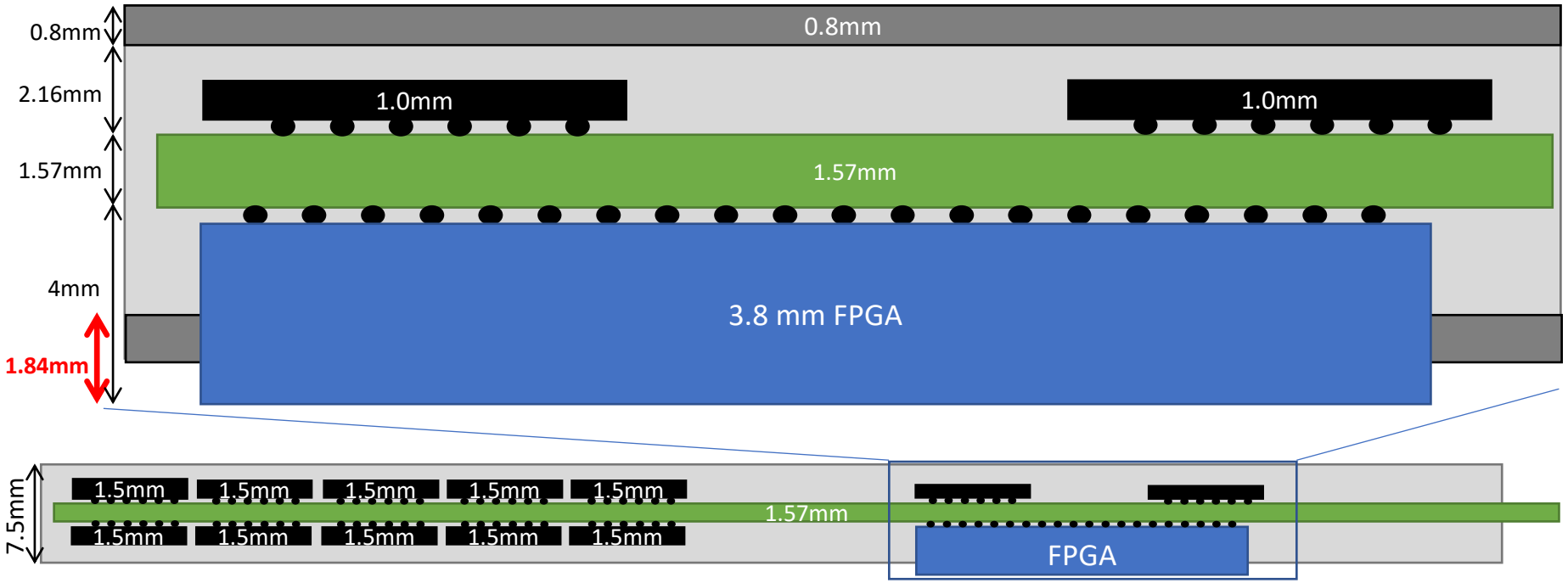


EDSFF E3.L Challenge With FPGAs



- 7.5mm Z-height
- The E3 spec is optimal for thin ASICs & flash memory
- The card edge is very close to the center of the stack-up

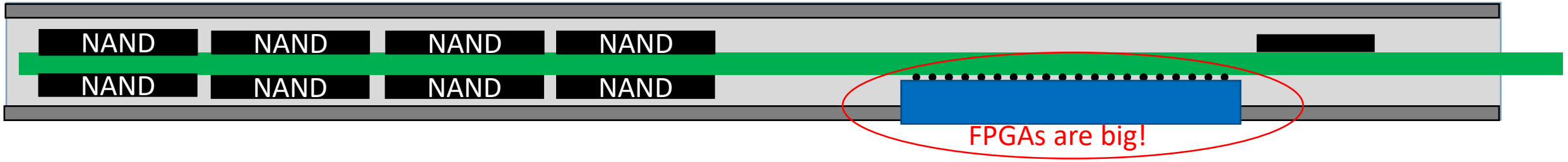
E3 Problem: FPGAs Are Big!

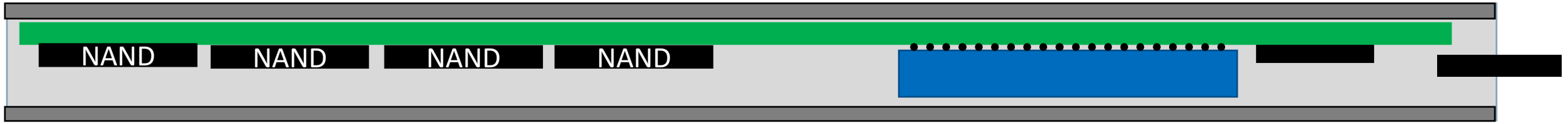


Not just FPGAs!

- Regulators
- Capacitors
- Inductors

One Potential Solution: Offset The PCB

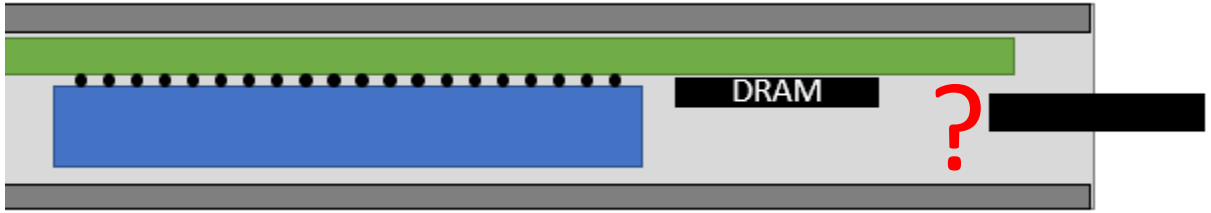




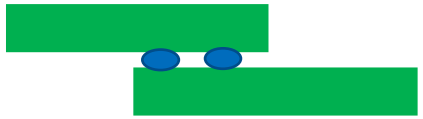
Challenges:

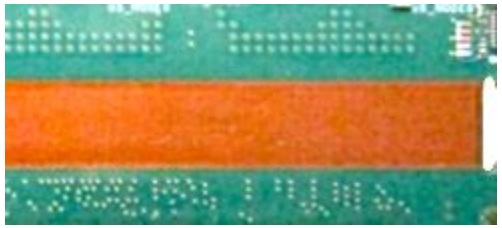
- Card edge alignment
- LED alignment
- Back-side components
- Thermal design

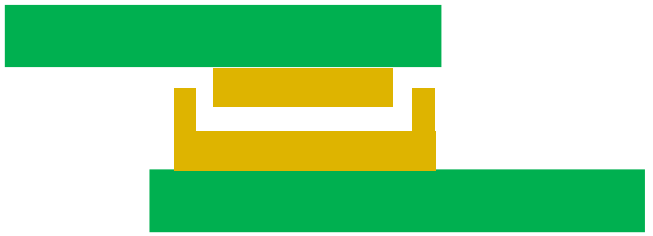
Shifting Your PCB And Maintain The Card Edge

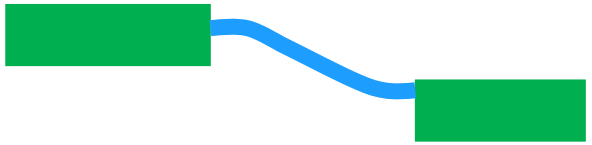


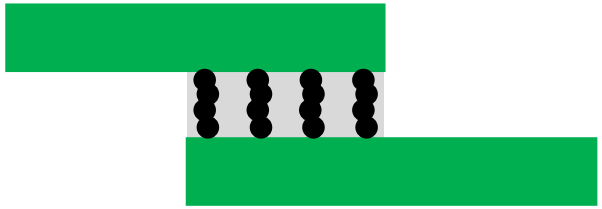
- Methods
  - Mezzanine soldering
  - Rigid flex
  - Plug/socket connectors
  - Elastomer connections
- Challenges:
  - Signal Integrity
  - Mechanical Stability
  - Tiny Dimensions
  - Reliability
  - Cost





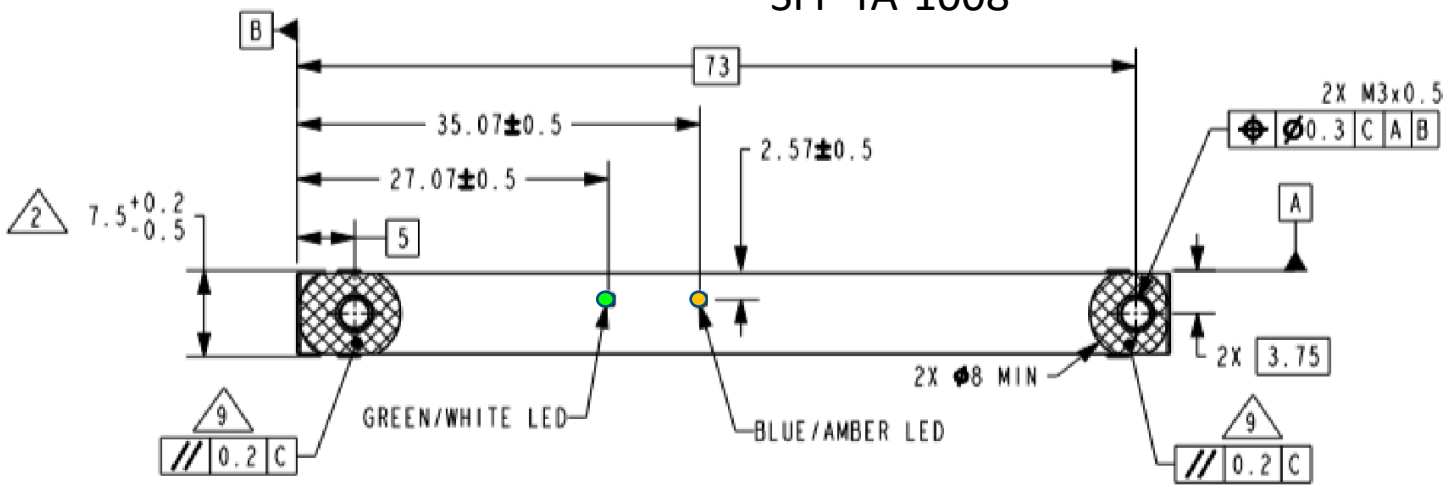


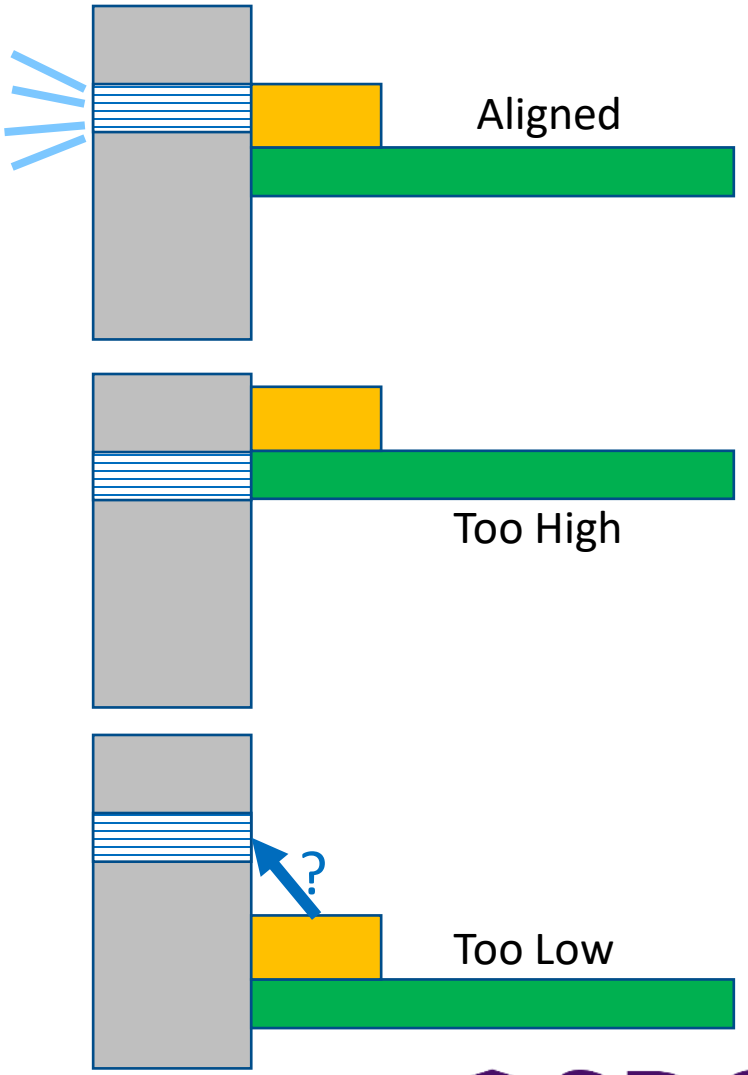




Aligning LEDs

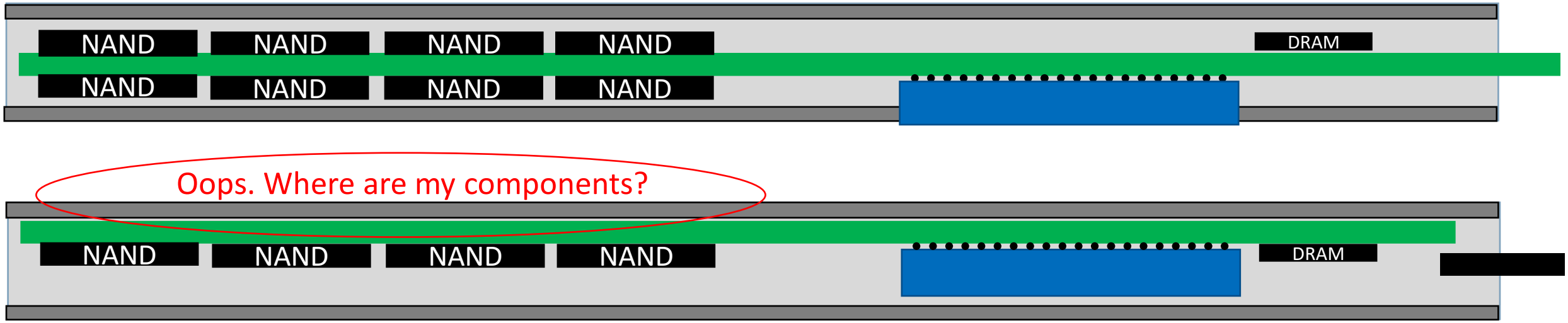
SFF-TA-1008





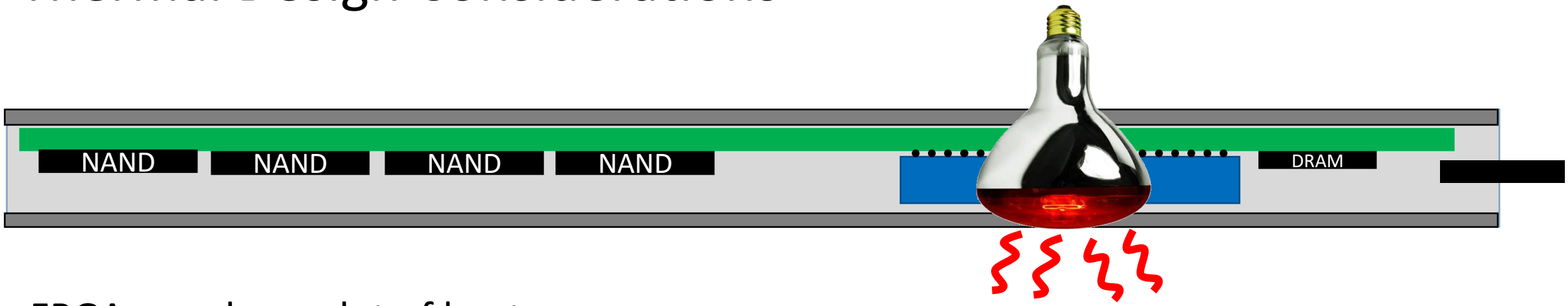
- LED hole is ideal for SMT LEDs on nominal board height
- Too high, and your LED holes are covered
- Too low, and your LED won't be visible
- Mitigations:
  - Light Pipes
  - Custom LED
  - PCB Mezzanine
  - Flare the hole

Back-Side Components

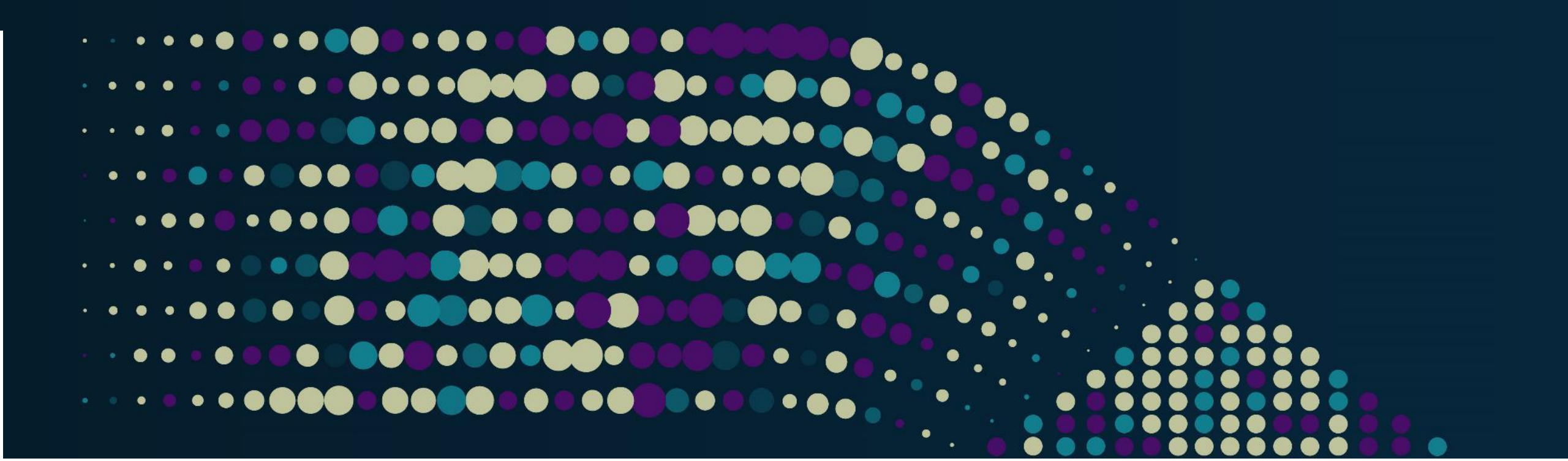


- You may find you run out of space when you adjust your Z height
- As with anything, it's a tradeoff
- Mitigations:
  - Move tall components to the taller side
  - Thin your shell
  - Use a mezzanine to put the PCB in the center again

Thermal Design Considerations



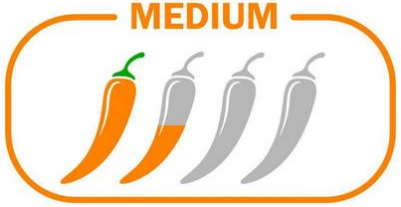
- FPGAs produce a lot of heat
  - Much of the power envelope will come from the FPGA
- Case design is very important
  - Material: Copper, Aluminum, alloys
  - Consider heat spreaders
  - Thermal Interface Material
  - Fin design and aerodynamics
- Thermal simulations are a must
- Test using SFF-TA-1023 methodology



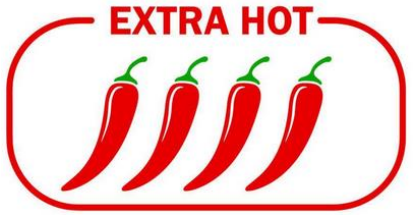
Enclosure Design Methodology

Acknowledgement: Brent Yardley

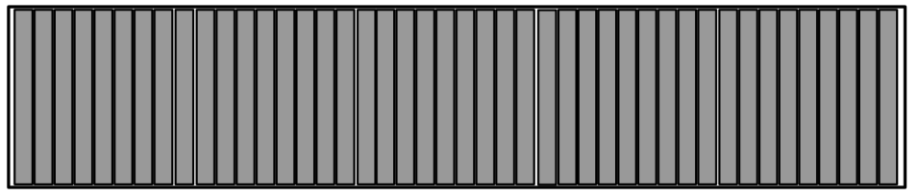
Keeping SSDs Cool



A typical 2U U.2 Storage Server:
24 SSDs at 25W max each
600W of drive power

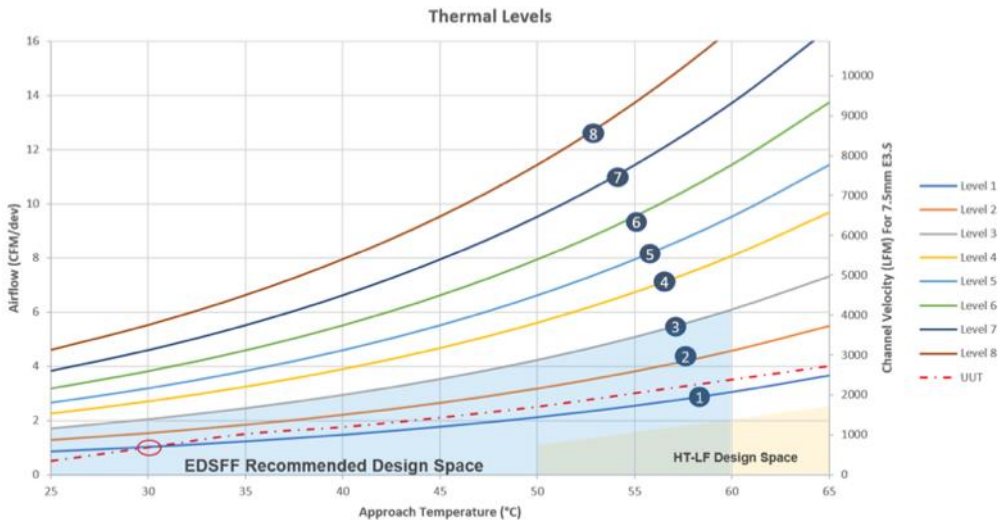


A 2U E3.L Storage Server:
44 SSDs at 40W max each
1760W of drive power

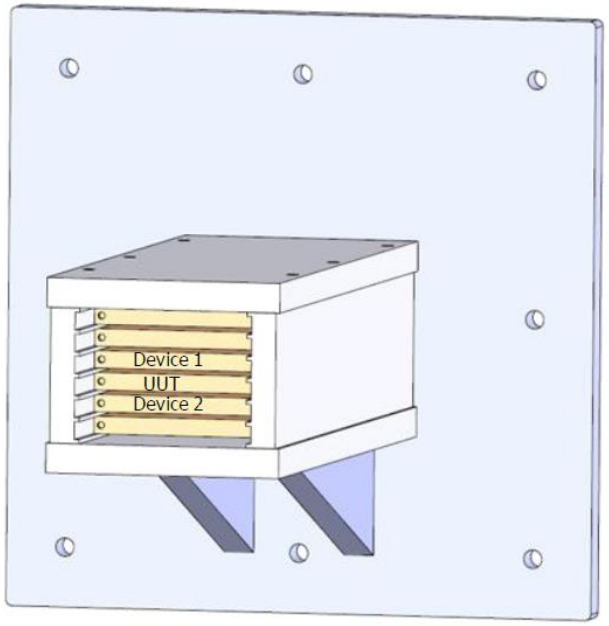


Concept from 2023 EDSFF whitepaper, B. Lynn, P. Kaler, and J. Geldman

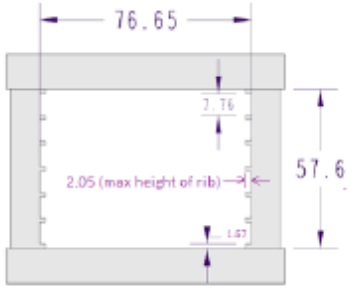
- SFF-TA-1023 defines thermal design criteria for both Enclosures and EDSFF devices
- Spec recommends operation in the blue region
- In a nutshell, characterize your enclosure to perform at or better than the level of your device



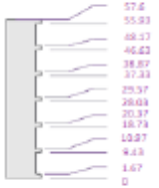
SFF-TA-1023 device test environment setup



- Set up an airflow chamber with various temperature and airflow setpoints
- Build a test box as suggested by the spec
- Run various workloads on different NVMe power states and collect data

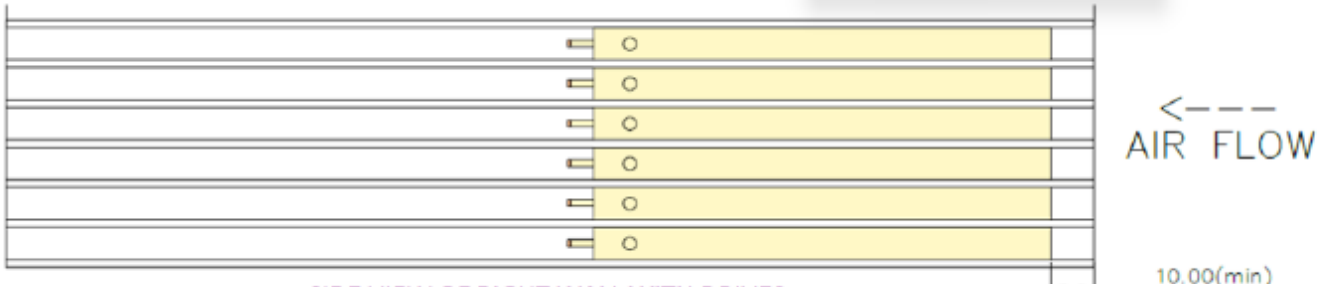


FRONT VIEW OF ASSEMBLY



INSIDE WALL DIMENSIONS

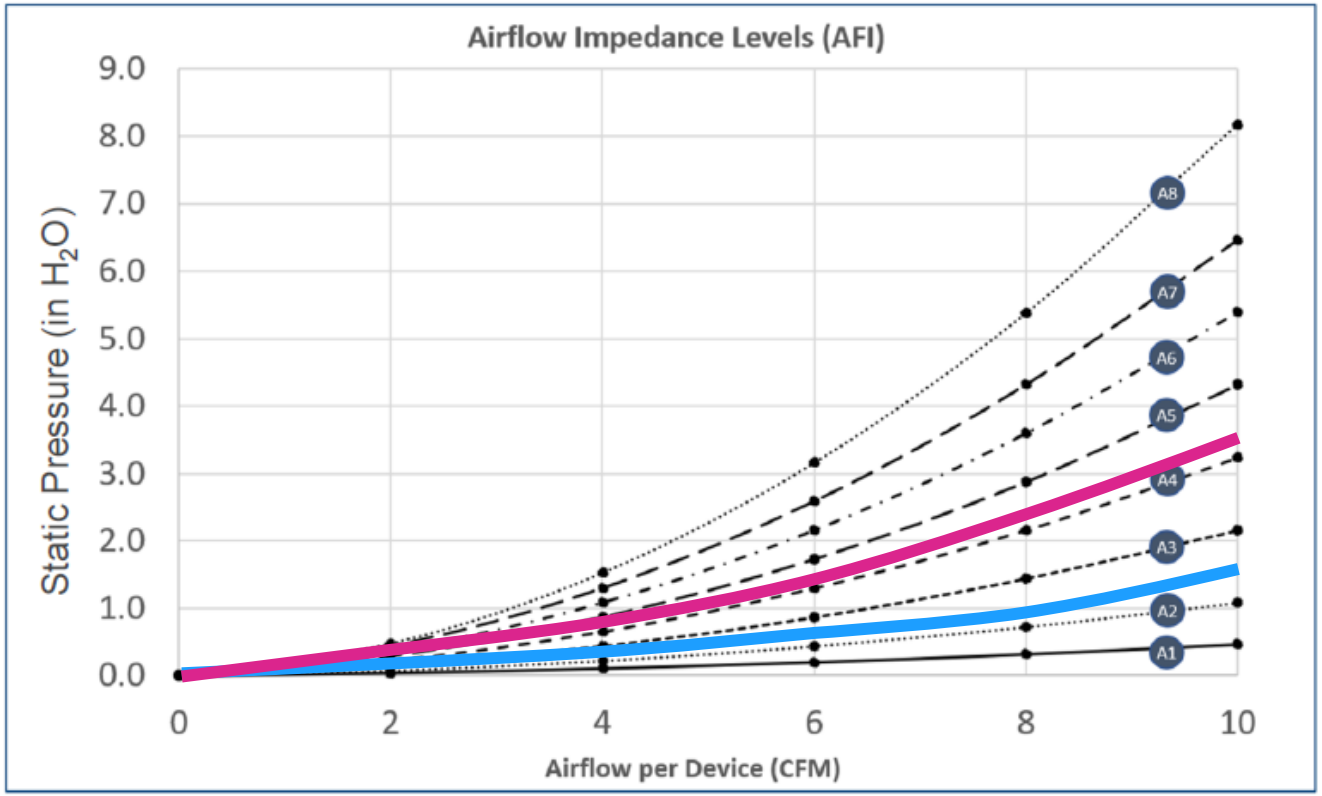
SFF-TA 1023 test environment s...



SIDE VIEW OF RIGHT WALL WITH DRIVES

SFF-TA-1023 Airflow Impedance (AFI) levels

- 8 AFI levels are defined
- Enclosures are tested with devices producing different AFI levels to determine which levels they can support
- Devices are tested to characterize their AFI level
  - AFI impacted by device shape (heat-sink fins, vents, etc)
- The device AFI must not exceed the enclosure AFI capability
  - Fan efficiency loss may occur

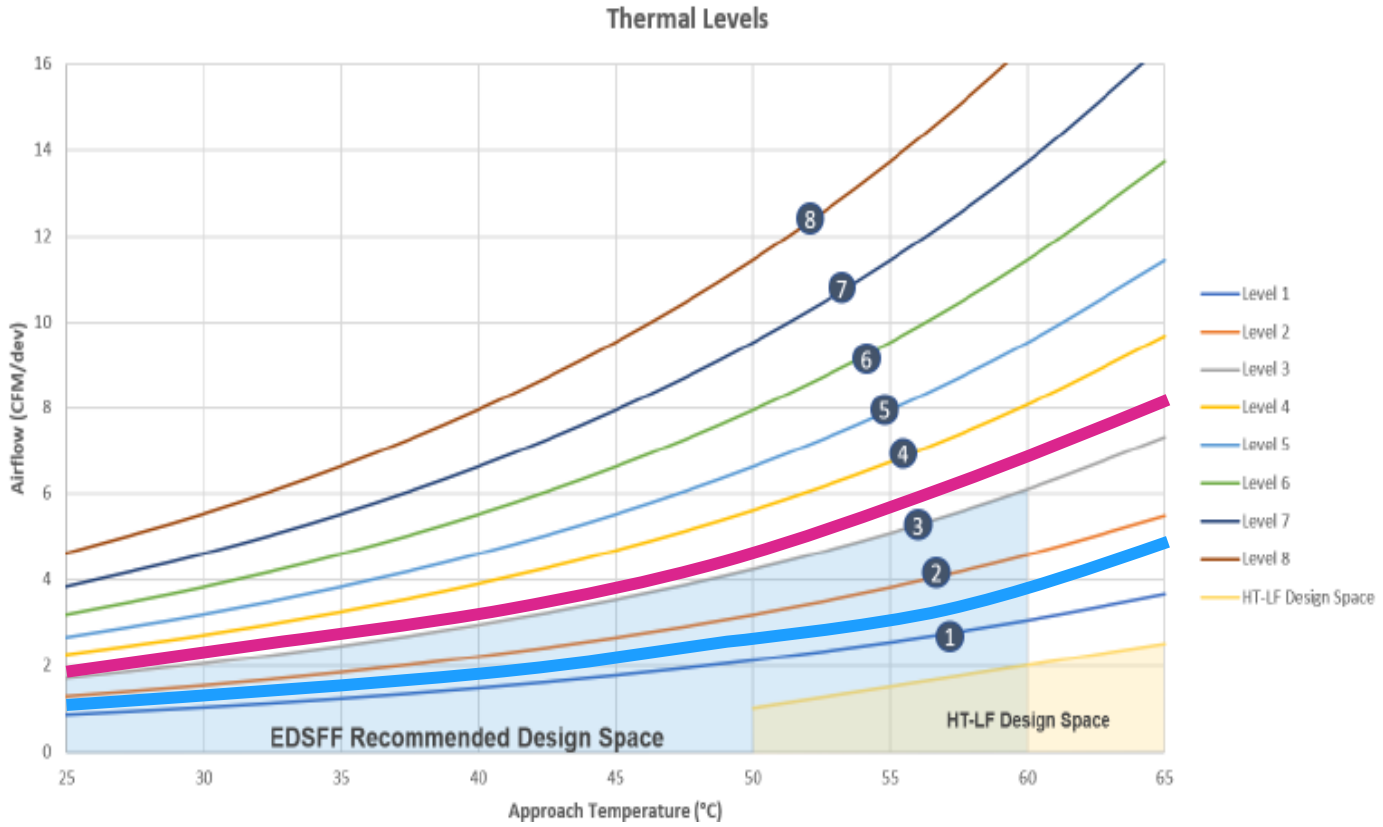


Example Enclosure Capability

Example Device Characteristic

SFF-TA-1023 MaxTherm Levels

- Minimum airflow required at Thermal Design Power
- 8 MaxTherm Levels
- Devices are tested to determine their MaxTherm Level
- Enclosure supports devices at or below its MaxTherm level

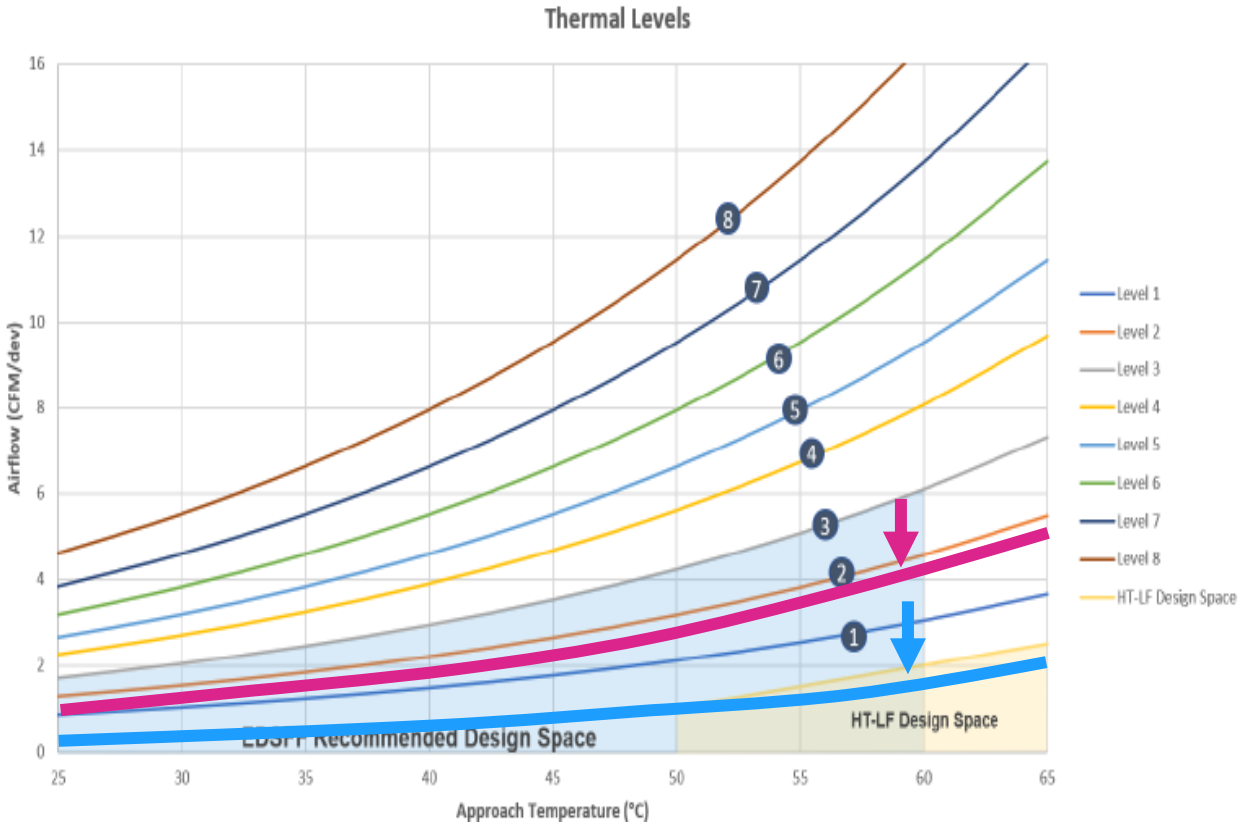


Example Enclosure Capability

Example Device Characteristic

SFF-TA-1023 DTherm Levels

- Your device may need to throttle in the case of a fan degradation or service action
- Same MaxTherm Level definitions
- Drive can self manage or be told to downgrade by the host
- Typically use NVMe Power States
- Quality of Service is impacted

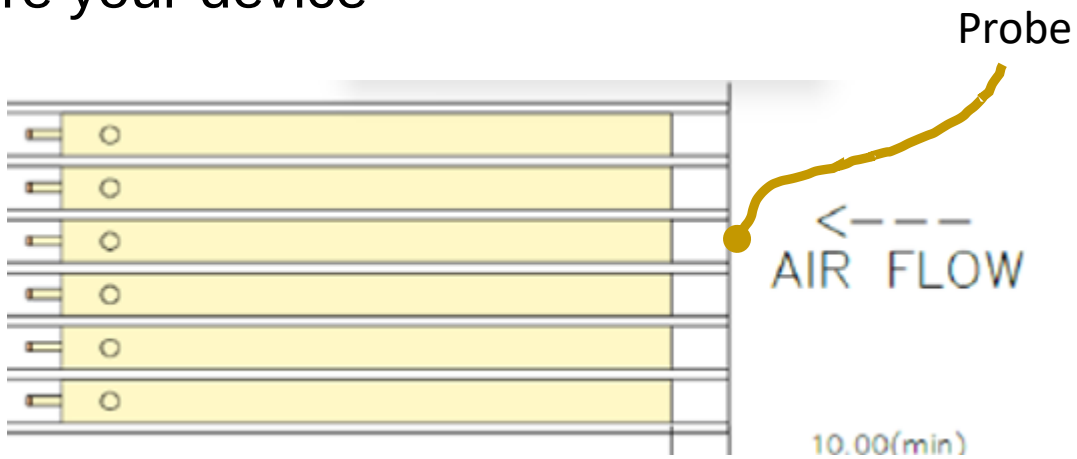


Example Enclosure Capability

Example Device Characteristic

SFF-TA-1023 MinAmbient and MaxAmbient

- Documented approach temperature limits, outside which your device may have problems operating
- MinAmbient is the lowest approach temperature your device supports
  - The thermal level curves have a minimum temperature of 25C
    - Testing may be impractical below 25C
  - Some components may not be qualified for less than room temperature
  - Change MinAmbient if your device has a specific minimum operating temperature
- MaxAmbient is the highest approach temperature your device supports
  - Allowed ranges from 50°C to 65°C
  - MaxAmbient defines the point above which, throttling to DTherm levels may occur



SFF-TA-1023 MaxAmbient vs ASHRAE

- MaxAmbient ranges from 50°C to 65°C per SFF-TA-1023

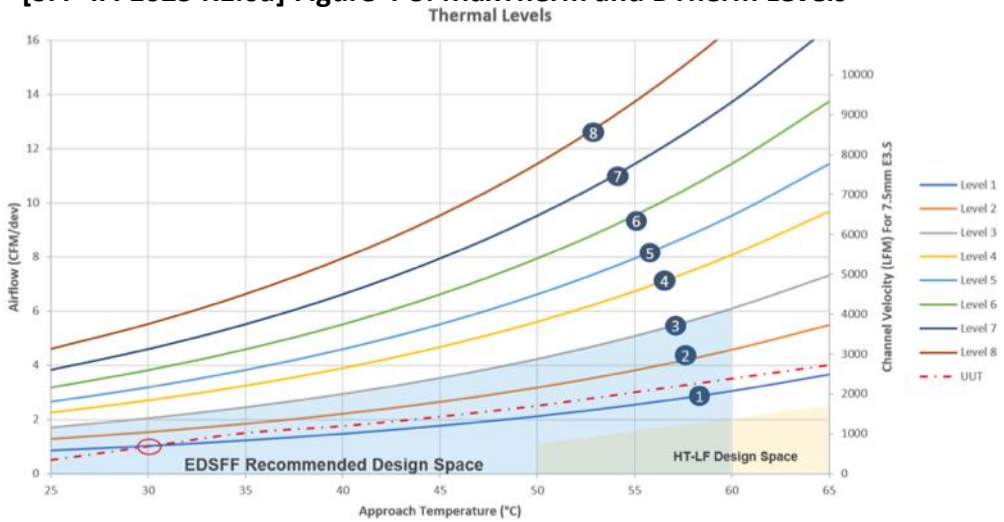
- ASHRAE standard max temperatures are 32°C, 35°C, 40°C, 45°C
  - Many servers are designed to ASHRAE standards
  - ASHRAE standards are designed for optimal SSD cooling
- IBM would like to see MaxAmbient allowed below 50°C

ASHRAE = **American Society of Heating, Refrigerating and Air-Conditioning Engineers**

EDSFF Gap Design Considerations

- SFF-TA-1023 recommends E3 device to operate within the blue shaded design space envelope (MaxTherm Level3 in the left picture)
  - Spec defines the SSD-SSD gap for E3 to be 1.8mm
- Example: Ta 35C, 2.5 CFM can be translated to ~1700 LFM (8.6 m/s) per equation 5-1
  - Compared to U.2 7mmT (pitch 12.5mm, gap 5.5mm), LFM is much higher from smaller gap (5.5mm vs. 1.8mm) But regarding U.2 15mmT (pitch 16.5mm, gap 1.5mm), we will see little differences vs 1.8mm
- Channel velocity will vary based on chassis design SSD to SSD pitch but can be calculated easily
- AFI is not easily calculated, but you can decrease your device AFI with a wider gap

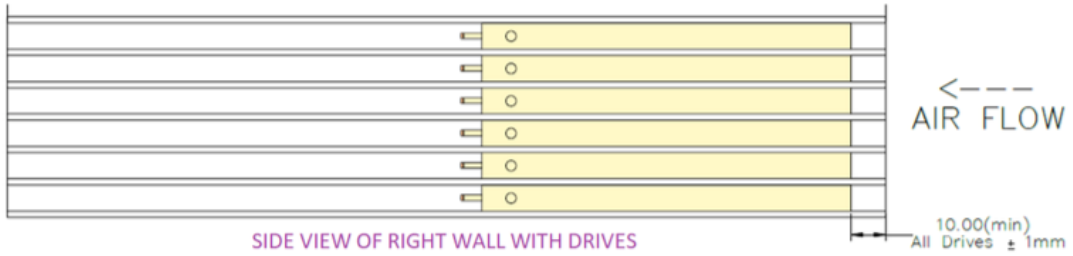
[SFF-TA-1023 R1.0a] Figure 4-3: MaxTherm and DTherm Levels

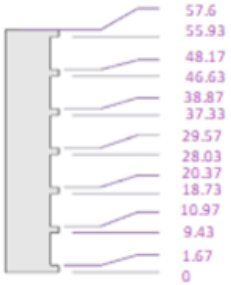


Equation 5-1: $Channel\ Velocity\ (LFM) = \frac{\frac{CFM}{Dev} * 92\,903}{(Drive\ Pitch\ (mm) - Drive\ Thickness\ (mm)) * Drive\ Width\ (mm)}$

Example: $Channel\ Velocity\ (LFM) = \frac{2.5\ CFM/dev * 92\,903}{(9.3mm - 7.5mm) * 76mm} = 1\,697\ LFM$

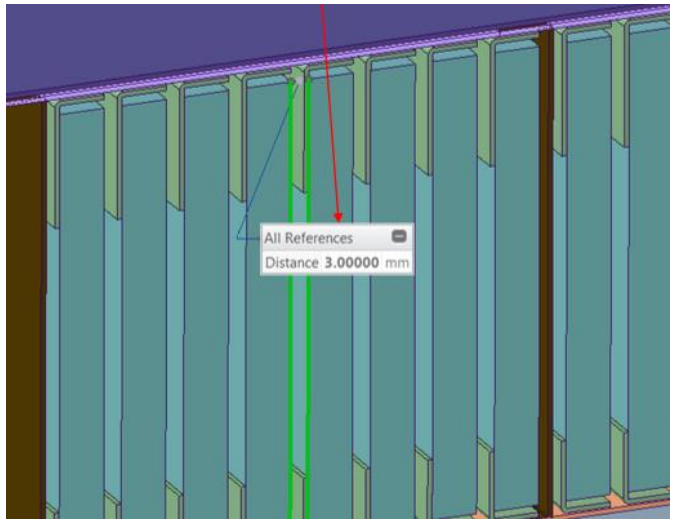
[SFF-TA-1023 R1.0a] Figure 5-5: Required Dimensions of E3 1T Test Fixture

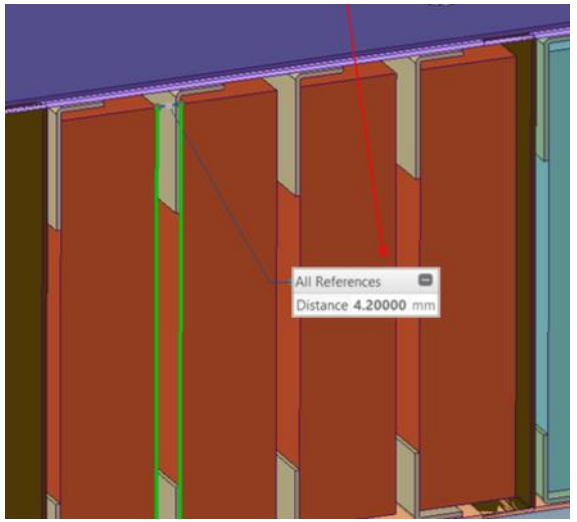




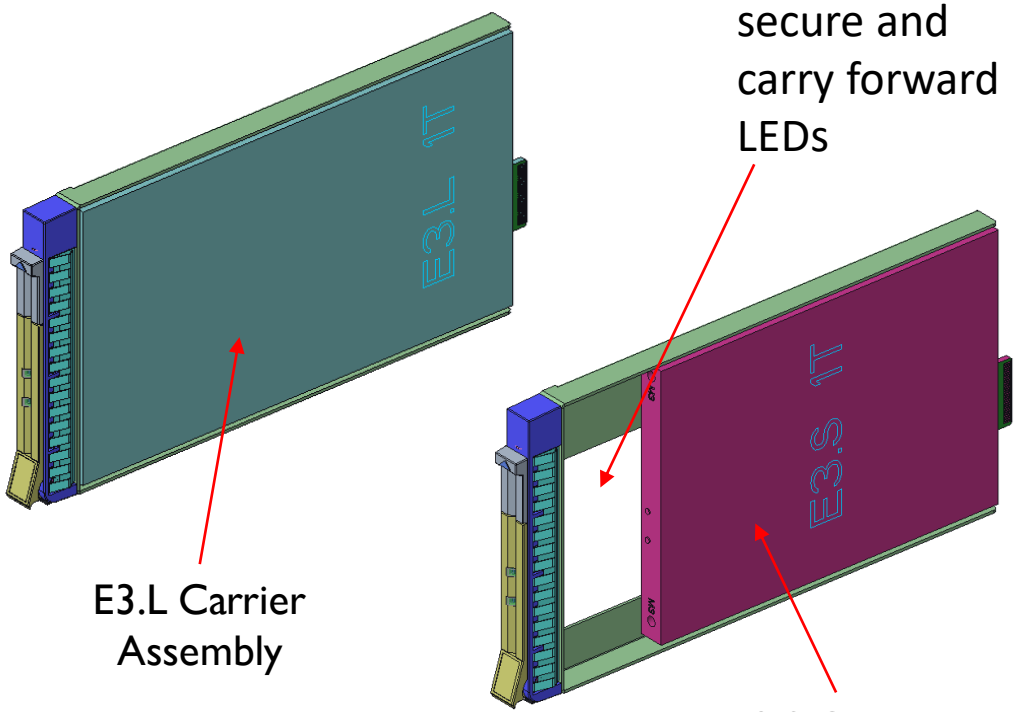
Mechanical Enclosure Slot Design

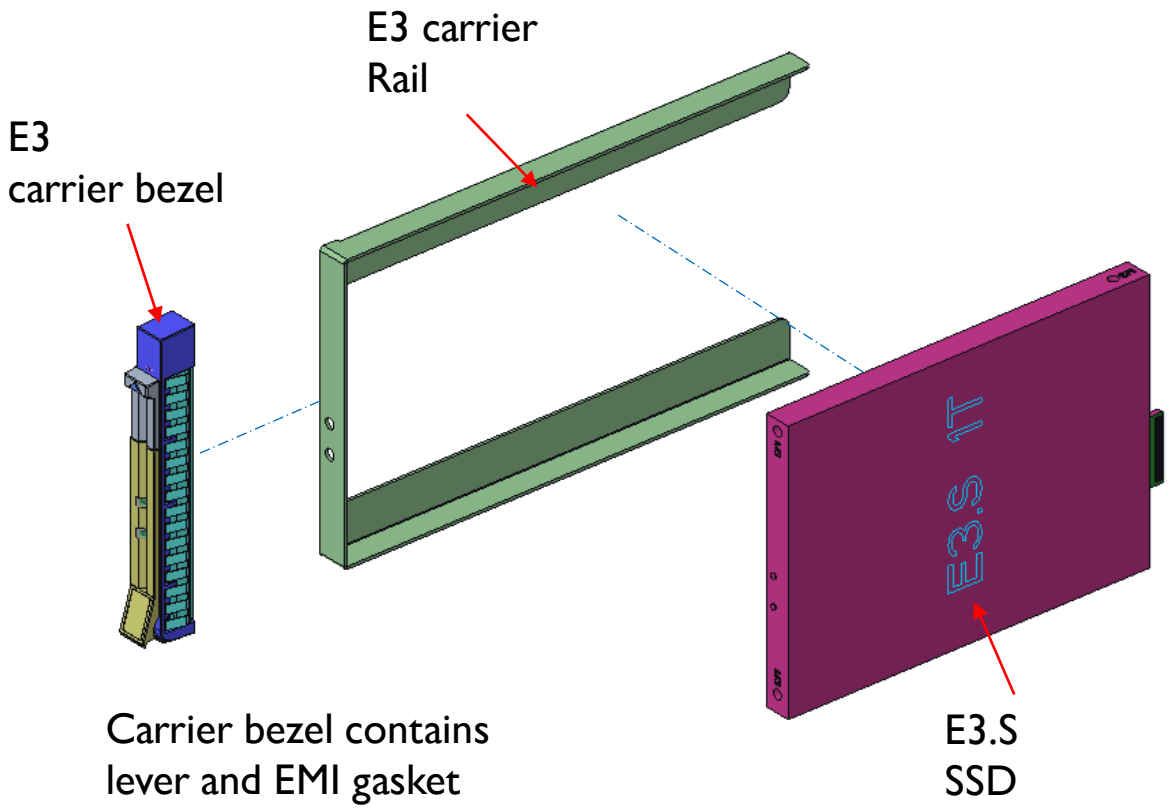
- To meet the thermal envelope requirements, a design is proposed using an approx. 3mm gap between E3 devices
- E3 2T devices in a pair of the same E3 slots will have a slightly larger gap of 4.2mm



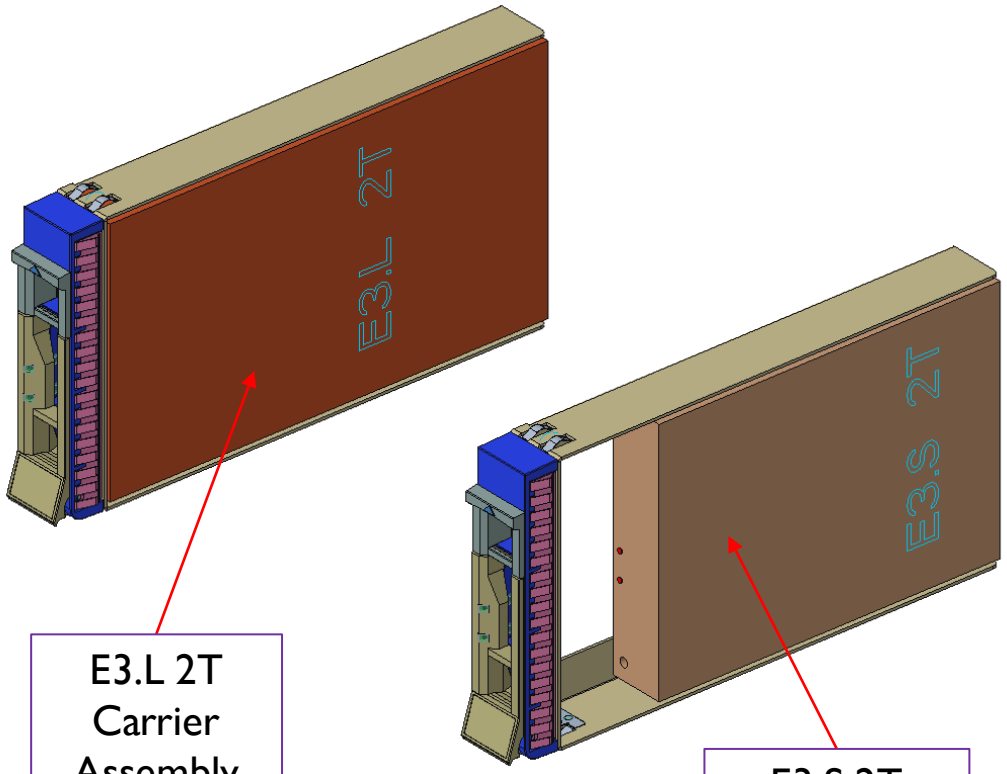


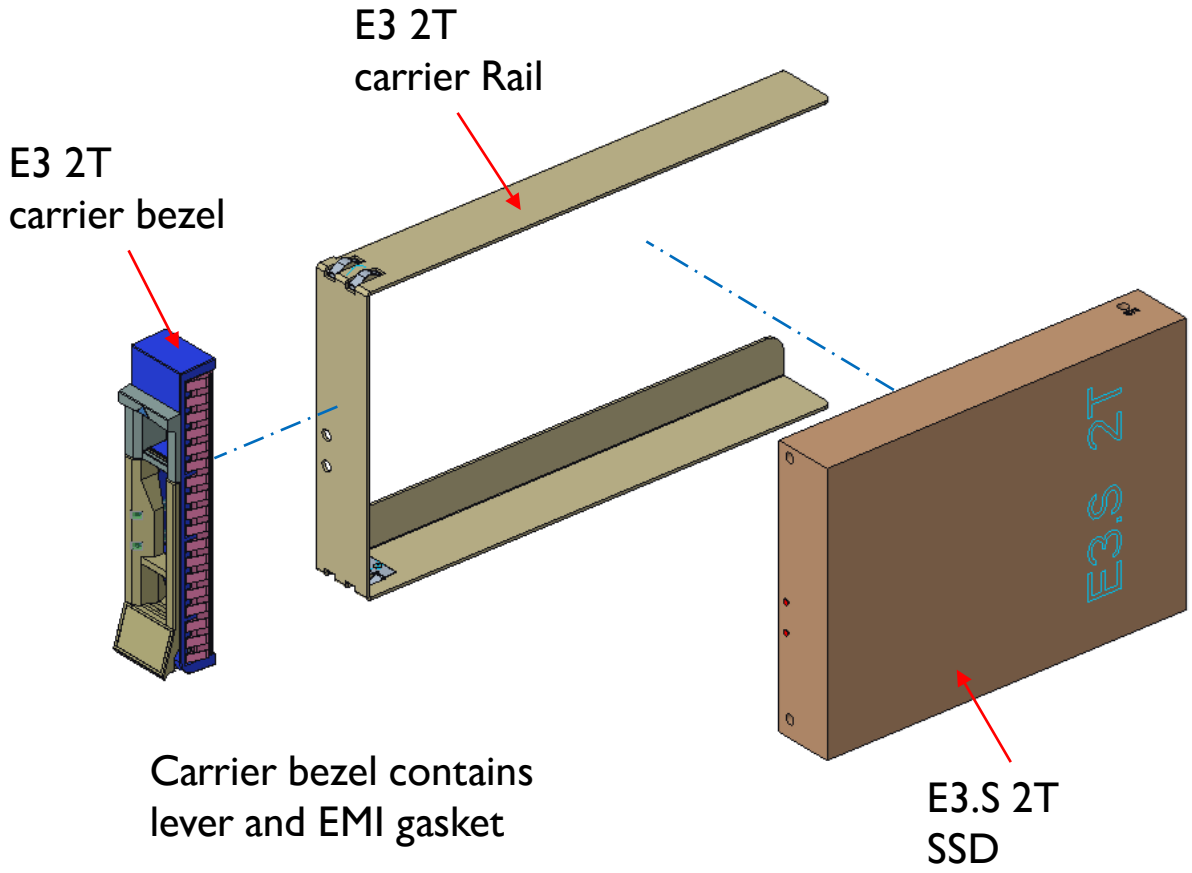
E3 Carrier Designs





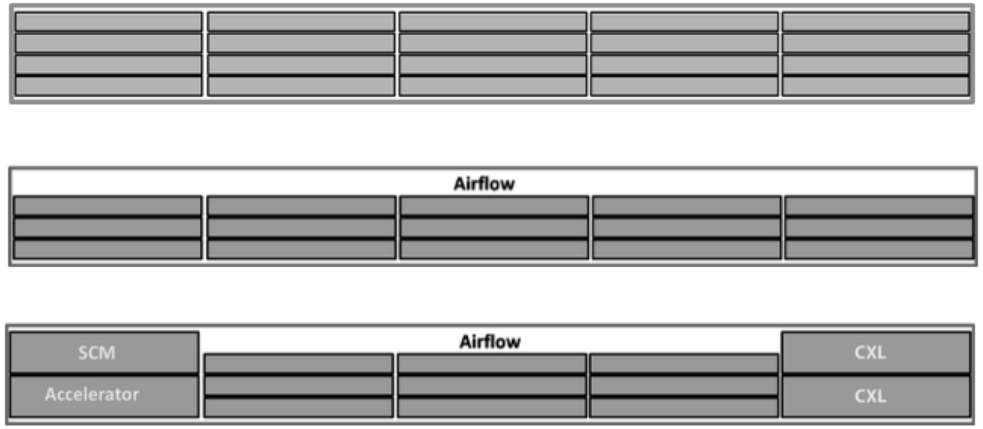
E3 2T Carrier Design

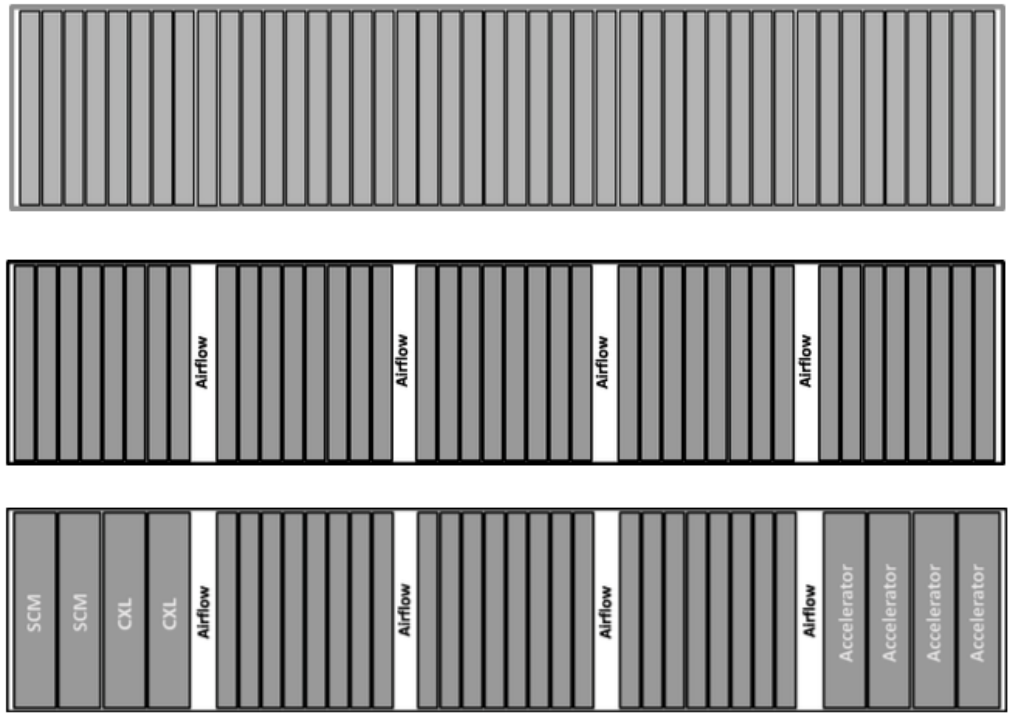




Enclosure Configurations

- In a white paper, B. Lynn, P. Kaler, and J. Geldman, several enclosure configuration options were explored





"The flexibility of the E3 form factor gives platform architects a wide range of options when it comes to supporting multiple system use cases. The ability to optimize around either density, host bandwidth, system power or device type makes the E3 form factor an ideal choice for platform architects and system designers. "

https://business.kioxia.com/content/dam/kioxia/ncsa/en-us/business/asset/KIOXIA_EDSFF_Intro_White_Paper.pdf

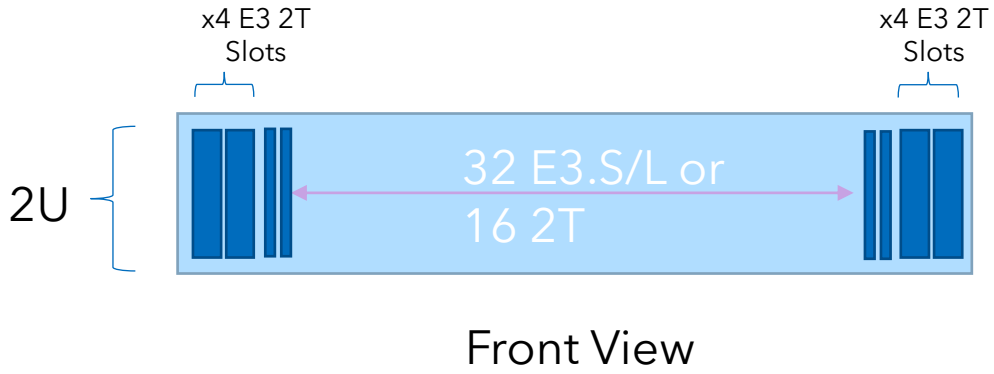
Enclosure Configuration Considerations

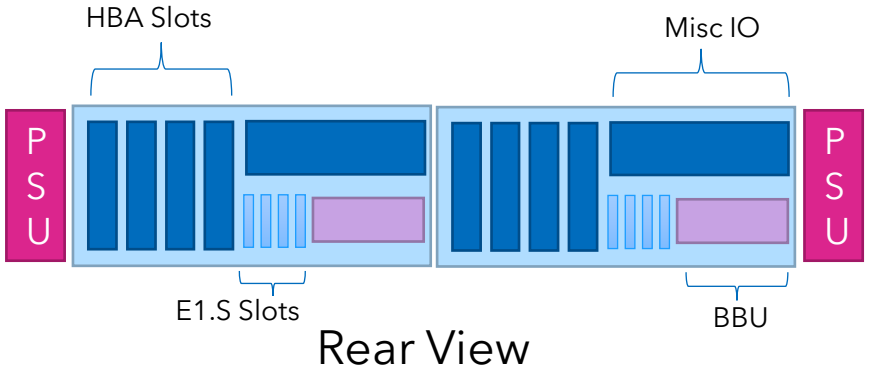
- We are looking at specifically 3 different enclosure types
  - 1U – Entry and Expansion
  - 2U – Midrange
  - 4U – High End
- Want to maintain a common design for IBM storage brands
  - E3.S/L is primary design point
- Flexibility to support E3 and E3 2T in the same enclosure
- Meet the thermal design requirements based on anticipated slot to slot pitch

1U Mechanical Concept

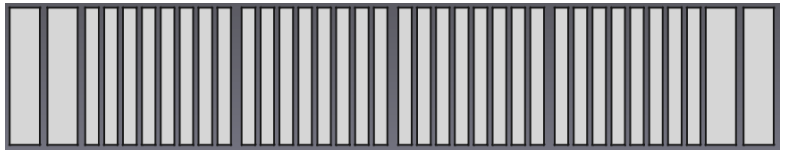
- Support up to 20 is possible, however, the ability to maintain a 3mm is not possible for all slots. To support a true even gap, between all devices, a gap spacing of 2.4mm would be needed
- 2T would be possible, however, uniform spacing gaps is not maintained. The gap spacing would likely lower: 2.4mm, middle: 3mm, and top: 3mm

2U Physical Mockup Concept

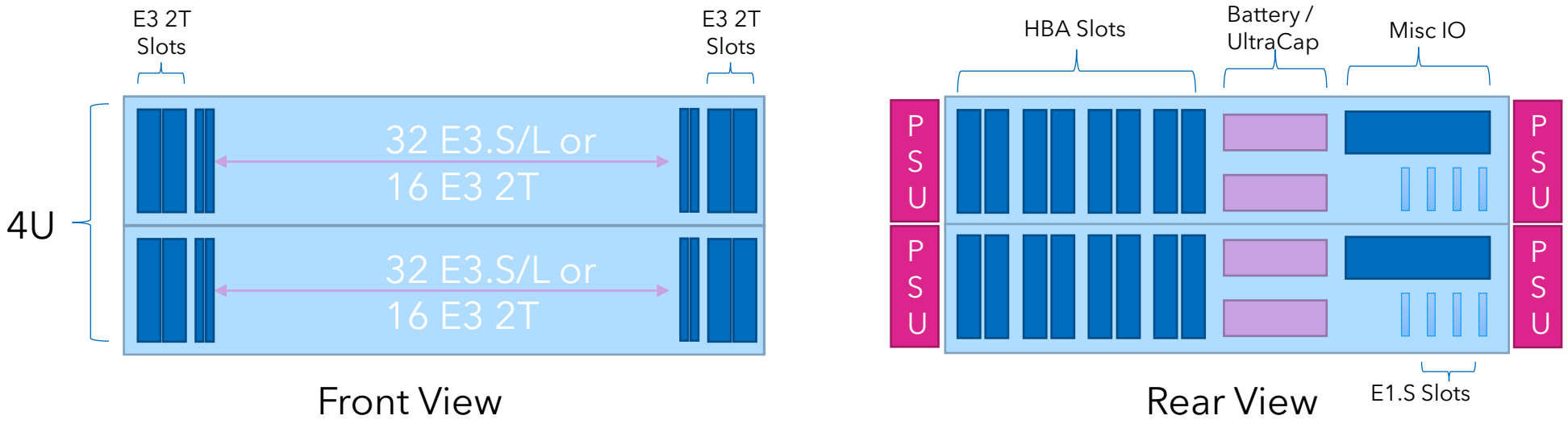




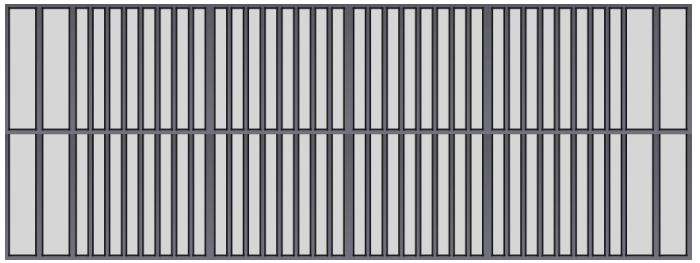
- 2U enclosure with 2U half wide canisters configured in a side-by-side configuration
- Each canister supports single socket w/ up to 12 DIMM slots per socket
- Up to 4 PCIe® 5.0 slots per canister
  - All mechanically x16, supporting a mix of x8 and x16 electrical
- Hot plug/replicable, E1.S boot drives
- Hot plug/replicable power loss protection
- 1+1 Common Redundant Power Supplies
- Configuration using up to 32 E3 2x2 slots and 4 E3 2T 2x4 or 1x8 slots



4U Physical Mockup Concept

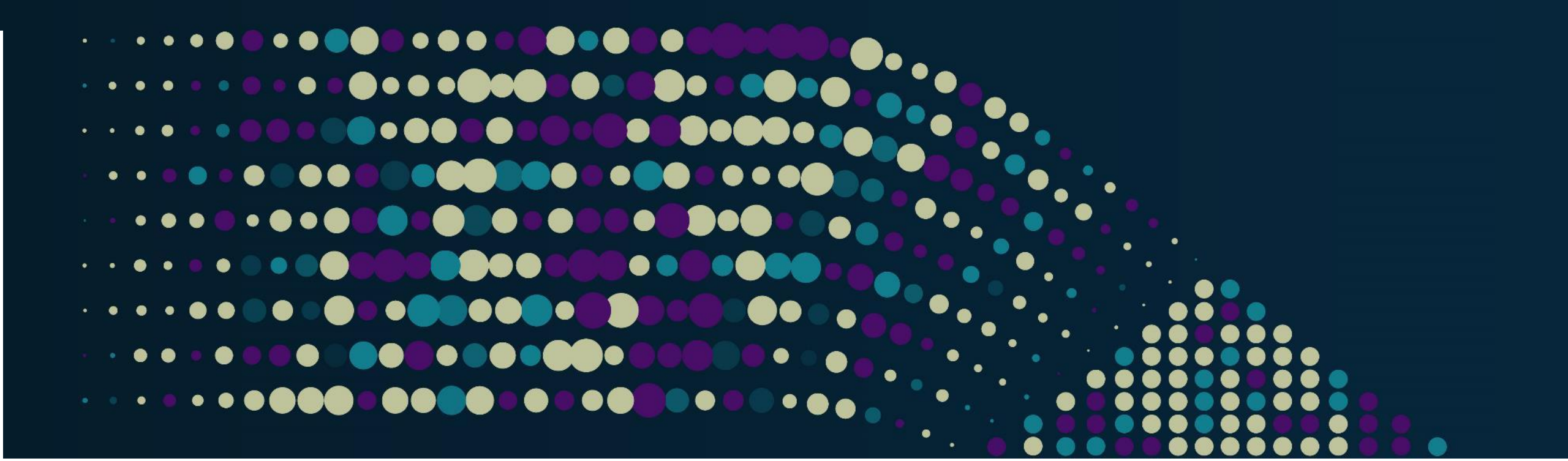


- 4U enclosure with 2U canisters configured in a top/bottom configuration
- Each canister supports dual socket w/ up to 12 DIMM slots per socket
- Up to 8 PCIe® 5.0 slots per canister
  - All mechanically x16, supporting a mix of x8 and x16 electrical
- Hot plug/replicable, E1.S boot drives
- Hot plug/replicable power loss protection
- 2+2 Common Redundant Power Supplies
- Configuration using up to 64 E3 2x2 slots and 8 E3 2T 2x4 or 1x8 slots



Summary

- SSD and storage server suppliers need an EDSFF roadmap to stay competitive
- Design concepts for SSDs and enclosures were presented
- An FPGA can fit into an E3.L form factor to maximize density
- The 1.8mm device gap on the enclosure is an example, but not ideal for all use cases
  - Not everyone is going for maximum enclosure density
  - AFI can be adjusted by gap



Please take a moment to rate this session.

Your feedback is important to us.

