SNIA DEVELOPER CONFERENCE
SDC 25
By Developers FOR Developers

Hyatt Regency Santa Clara, CA
September 15-17, 2025

Gen6 is coming, but what is needed from NV Storage?

Suresh Rajgopal and Brent Byron

www.sniadeveloper.org

Outline

- Divergence in requirements for SSD Storage
- Near-GPU SSDs
  - AI Workload Drivers
  - SSD Requirements
- Capacity SSDs
  - Needs from AI Infrastructure
  - SSD Requirements

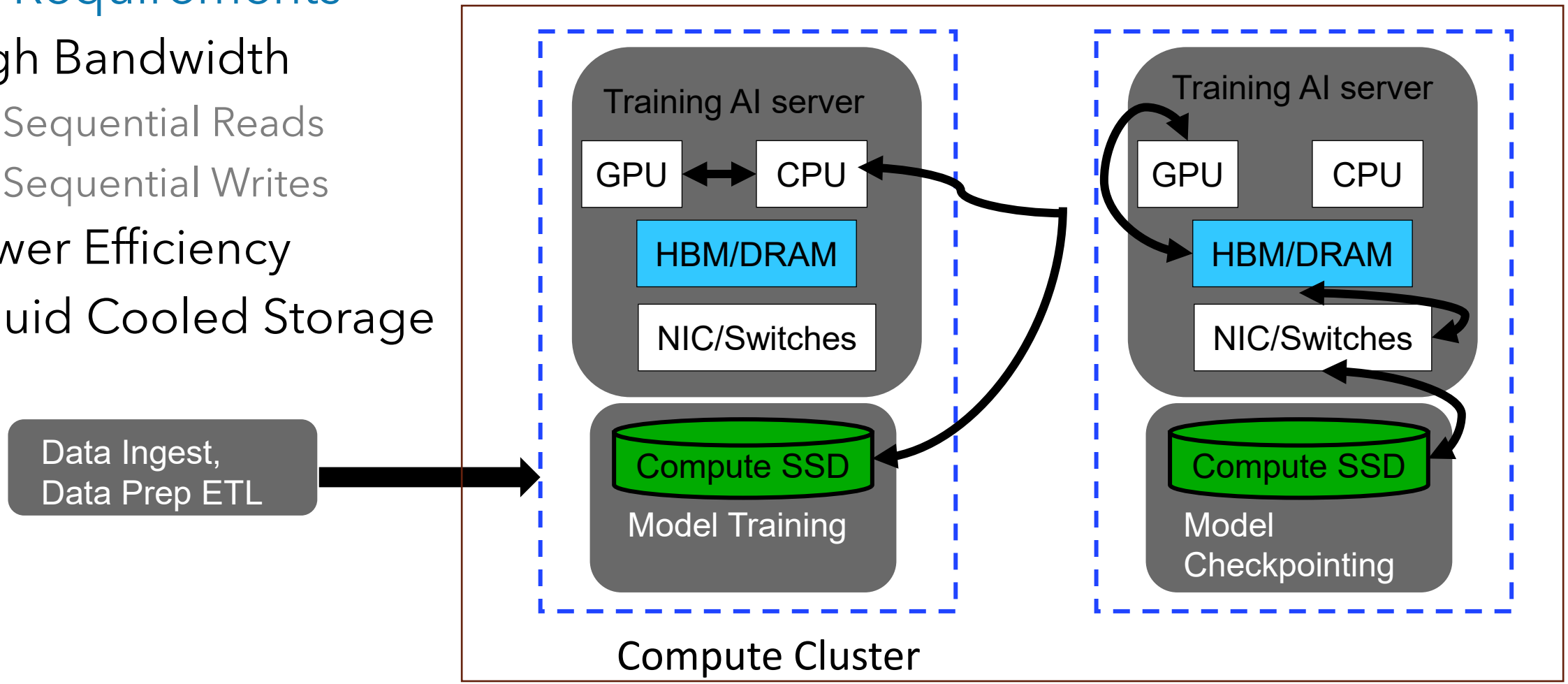
micron

SDC 25

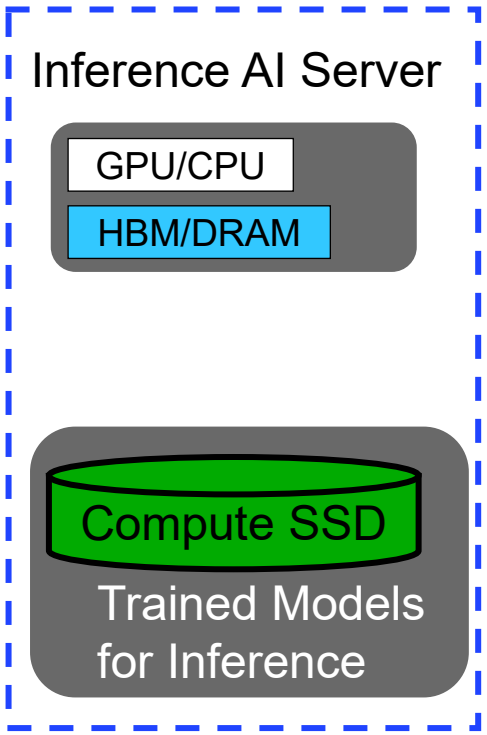
Near GPU Storage – Training and Checkpointing

Requirements

- High Bandwidth
  - Sequential Reads
  - Sequential Writes
- Power Efficiency
- Liquid Cooled Storage



Near GPU Storage- KV Cache Reuse Tiering (during Inference)



KV Cache Reuse

- Tradeoff between Recompute or Reuse from Storage
- Recompute threshold increases as GPUs get faster
- If KV Cache can be reused, GPU cycles can be used elsewhere
- Batch Sizes, I/O Sequence lengths can drive extent of reuse

Workload Characteristics

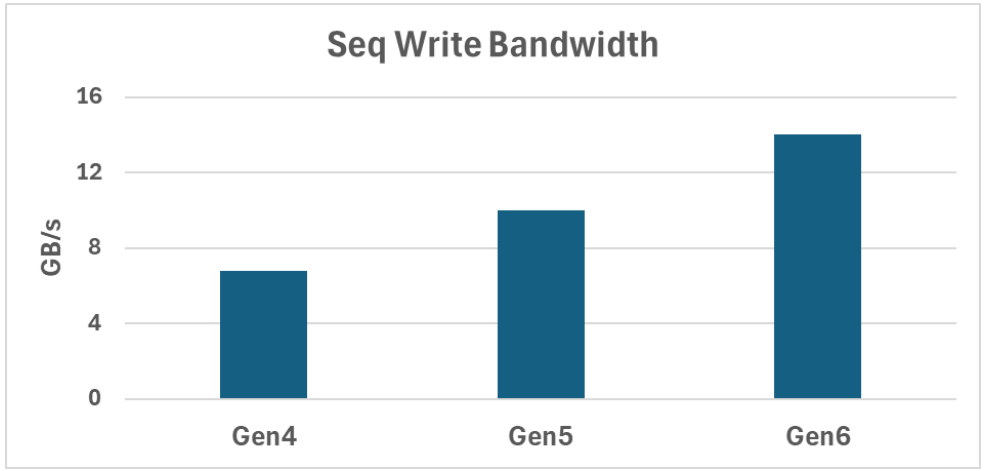
- Large transfers (2MB+ writes,28-32KB+ reads) and growing
- R/W ratio depends on cache hit rate
- Discard sizes match write sizes (entire caches discarded at once)

KV Cache Reuse tiering also driving high sequential read/write bandwidth

micron

SDC 25

Checkpointing Requirements-Sequential Write BW



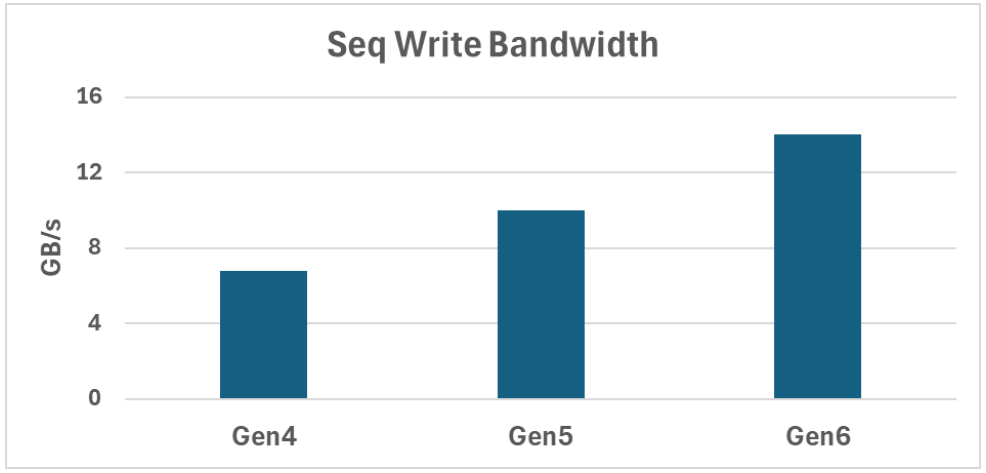
- Checkpoint Size – f(Param, Optimizer States)
- 405B param model ~11TB
- Checkpoint (CP) frequency – GPU Utilization
  - Taking too long can be impactful
  - Distributed across multiple checkpoint writers

Checkpoint completion Time(s)		
Gen4	Gen5	Gen6
72	49	35
180	122	87
360	245	175
540	367	262
720	490	350

Checkpoint – every 2hrs (Illustration)

Checkpoint Completion	Checkpoint Time		# of Drives needed		
Target Time(s)	GB/s	%age	Gen4	Gen5	Gen6
72	153	1%	22	15	11
180	61	3%	9	6	4
360	31	5%	4	3	2
540	20	75%	3	2	1
720	15	10%	2	2	1

Checkpointing Requirements-Sequential Write BW

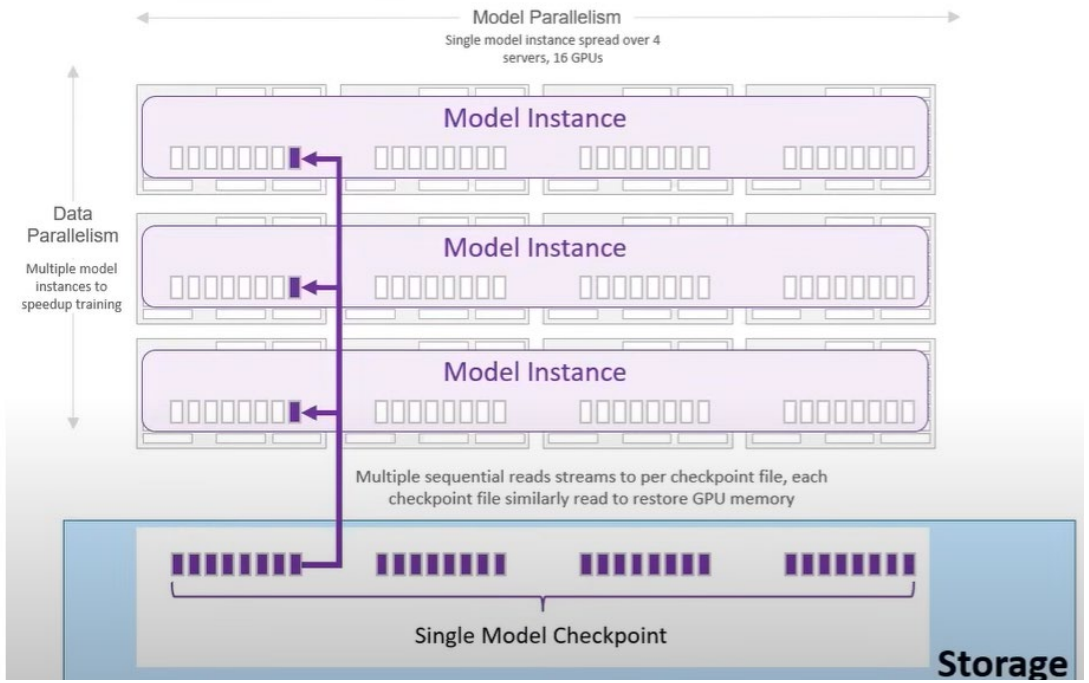


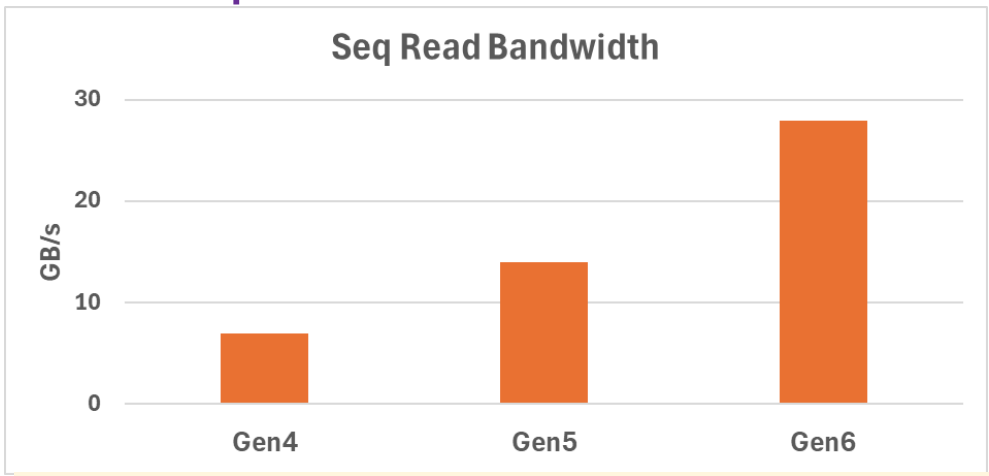
- Checkpoint Size – f(Param, Optimizer States)
- 405B param model ~11TB
- Checkpoint (CP) frequency – GPU Utilization
  - Taking too long can be impactful
  - Distributed across multiple checkpoint writers

Checkpoint – every 2hrs (Illustration)

Checkpoint Completion	Checkpoint Time		# of Drives needed		
Target Time(s)	GB/s	%age	Gen4	Gen5	Gen6
72	153	1%	22	15	11
180	61	3%	9	6	4
360	31	5%	4	3	2
540	20	75%	3	2	1
720	15	10%	2	2	1

Checkpoint Restore Requirements-Sequential Read BW





- Data parallelism- Multiple readers to restore checkpoint data to all GPUs
- Training cannot restart till checkpoints are restored

11TB Checkpoint Restore in 5mins (Illustration)

Data Parallelism Instances	BW GB/s	PCIe SSDs Gen4	Gen5	Gen6
1	7	1	1	0
8	59	8	4	2
16	117	17	8	4
32	235	34	**17**	8
64	469	67	34	**17**
128	939	134	67	34

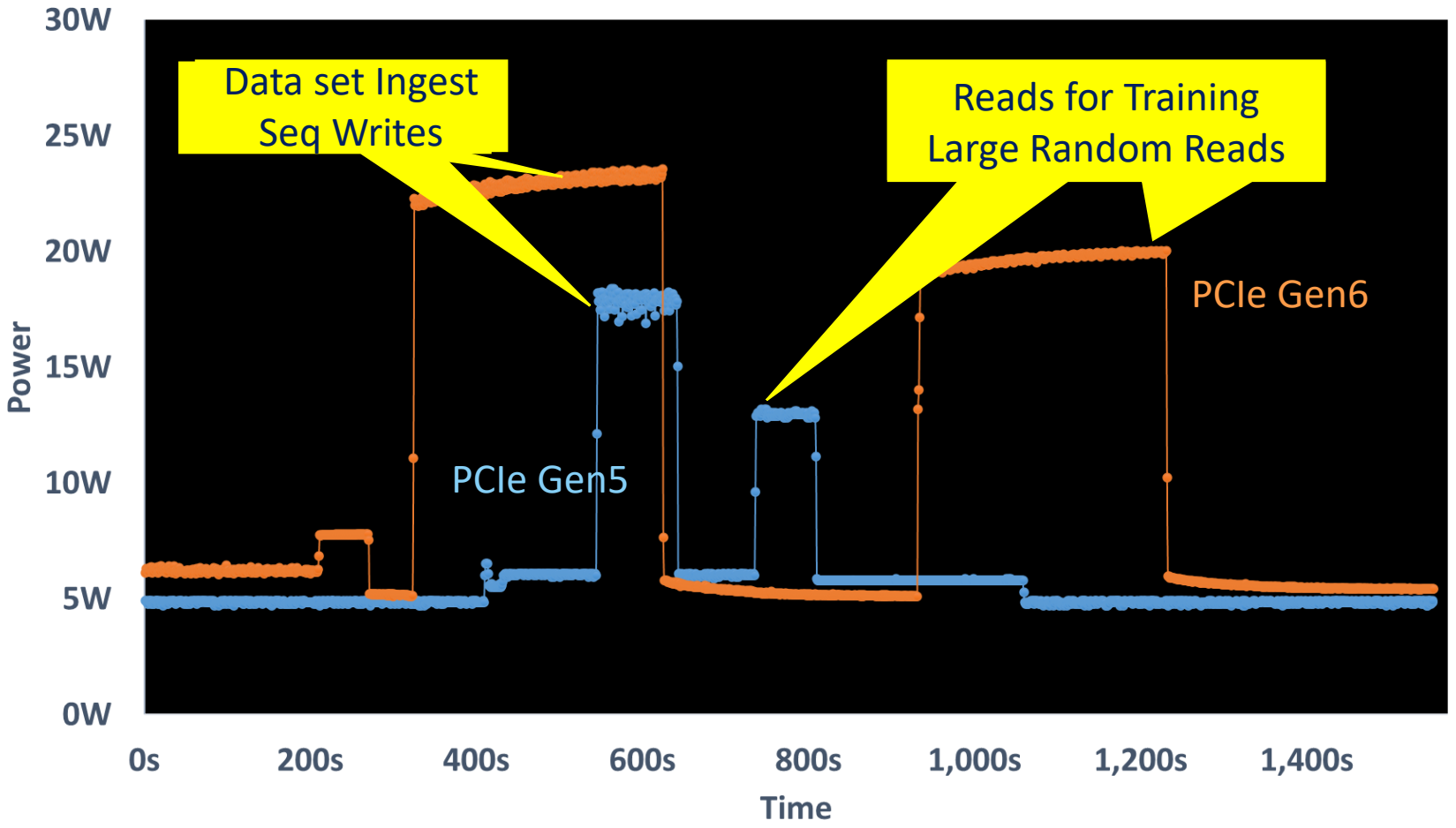
Near GPU SSD Requirements – Power Efficiency

- UNET3D (3D Medical Imaging) – ML Perf Storage Benchmark
  - A 3D medical imaging model
  - Reads large image files into GPU memory
  - Training is performed to generate dense volumetric segments.
- From the storage perspective
- Large Sequential Writes to Ingest the files
- Large Random Reads to enable training – storage bound

micron

SDC 25

Near GPU SSD Requirements – Power Efficiency



Ingest			
PCIe Gen	Gen5	Gen6	
Data(TB)	1.1	3.9	3.6X
Perf/W	0.6	0.6	

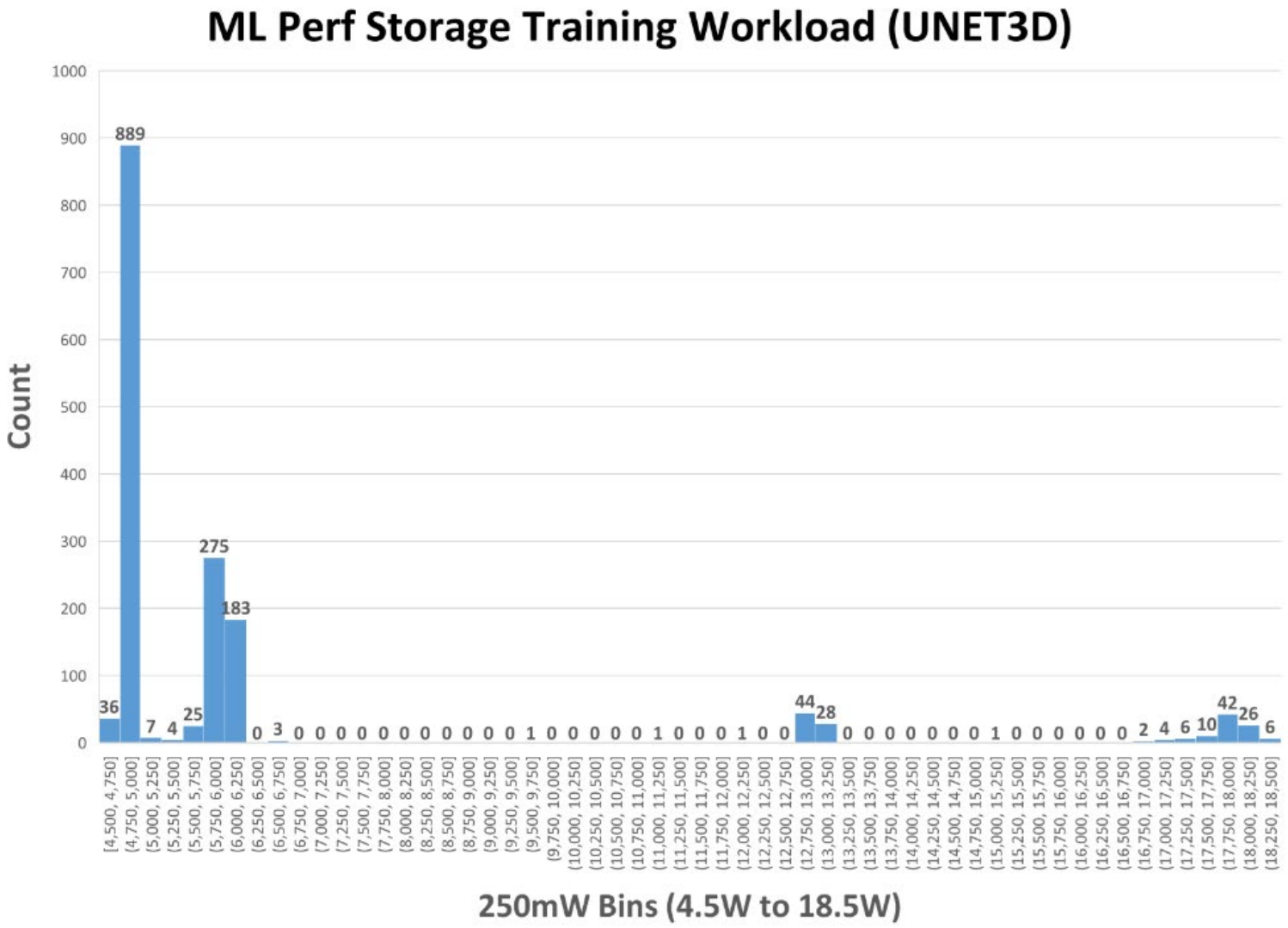
Read for Training			
PCIe Gen	Gen5	Gen6	
Data(TB)	1.1	8.0	7.5X
Perf/W	1.1	1.3	

- **Higher Bandwidth and same or lower power – Improved Power Efficiency**

micron

SDC 25

Near GPU SSDs – Self-Reported Power



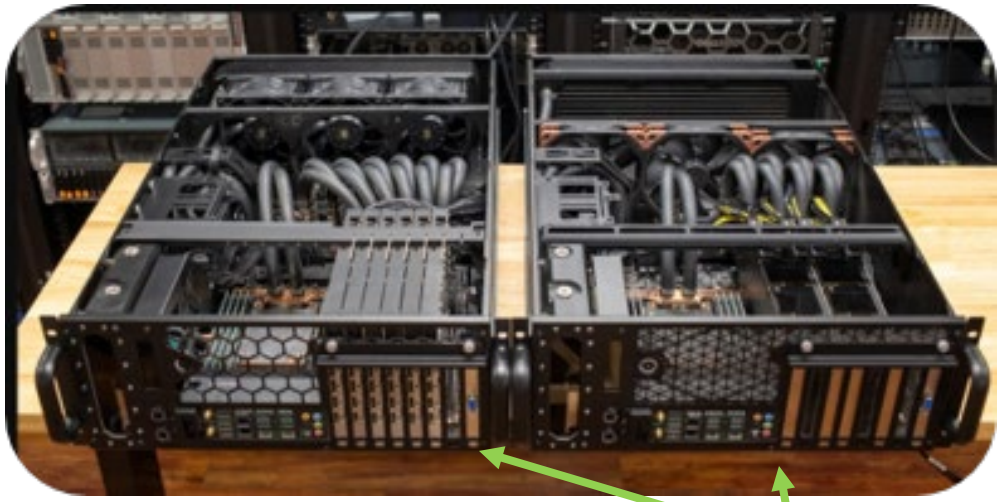
- TP4199 (NVMe)
  - Total Power Consumed
  - Track Energy Consumption
  - Power over time – Histogram
  - Power Threshold Reporting

Tracking and Optimizing SSD Power
Helps optimize AI Server Energy Efficiency

micron

SDC 25

Near GPU SSD Requirements – Liquid Cooling



2U Server-Liquid Cooled GPU/Mem but air-cooled storage

- Benefits of Liquid Cooled Storage for near-GPU compute trays
  - Rack density, Denser (1U/0.5U) server configs
  - Less energy and Less noise: >20W Fans vs liquid cooled
  - Greater efficiency: higher thermal conductivity/ specific heat of liquid vs air

1U liquid cooled server with cold-plate LC storage bays

Direct 2 Chip Cold Plate Liquid Cooling

E1.S 9.5mm

D2C Location	Both Sides	Single Side	Connector End
Thermals			
SSD Design Complexity			
Host Design Complexity / Serviceability			

- E1.S 5.9mm not very conducive
- Standards considerations
  - Cold plate contact area
  - Label placement area
  - Degree of flatness/roughness
  - TIM or no-TIM (Thermal Interface Material)

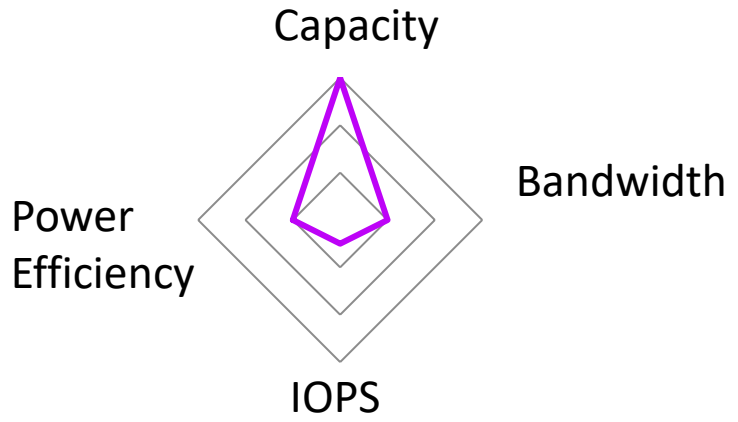
E3.S 1T

DLC Location	Both Sides	Single Side	Edge
Thermals			
SSD Design Complexity			
Host Design Complexity / Serviceability			

micron

SDC 25

AI Pipeline and the role of Storage – Ingest



Data Ingest,
Data Prep ETL

Relational datastore

Data lake/ object store

Network file system

Capacity SSD

Capacity SSD

Capacity SSD

Capacity SSD

Capacity SSD

Capacity SSD

Storage Cluster

Network Attached

Compute Cluster

micron

Capacity Storage Requirements and SSD Considerations

Requirements

- High Capacity
- Power
- Cost

Architectural Considerations

- Enabling SSD Capacities >128TB to 1PB
  - NAND Placements
- Power
- Cost – DRAM with 4K IU

micron

SDC 25

DRAM Budget Challenge with Drive Capacity Growth

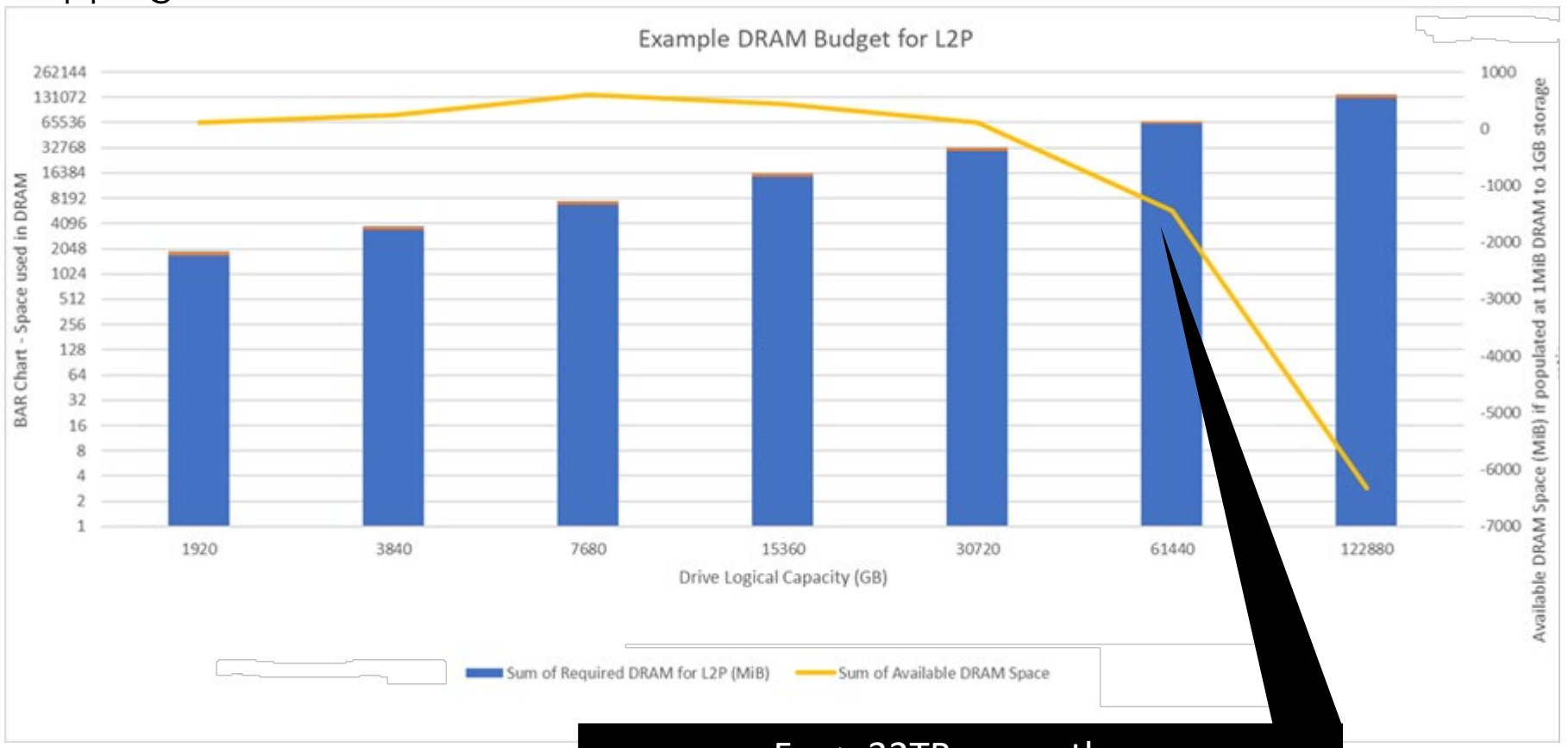
For a standard 4K IU based mapping table

As drive capacity increases, L2P table size in DRAM increases:

1. Number of L2P entries - doubles with capacity
2. L2P entry size – more NAND to address

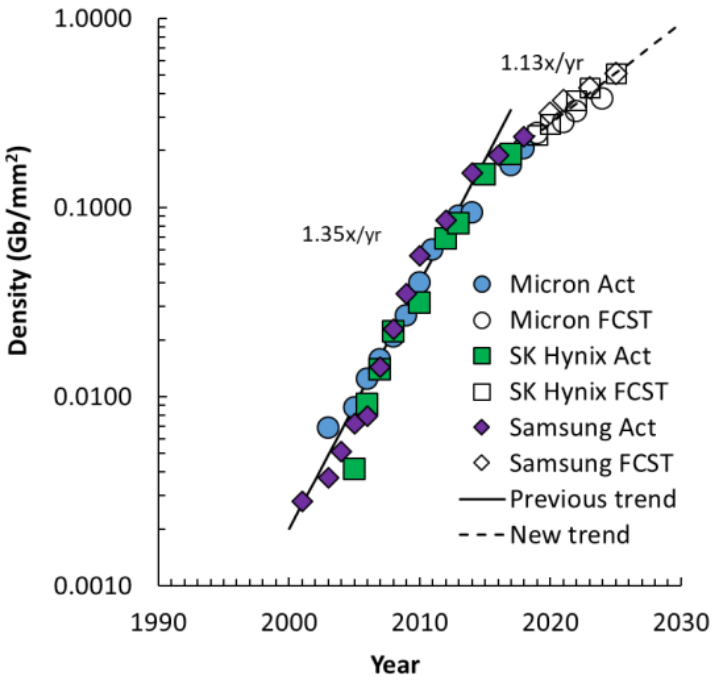
Increased DRAM challenges:

- PCB Space
- Less power budget for NAND
- SI may reduce DRAM speed
- Cost

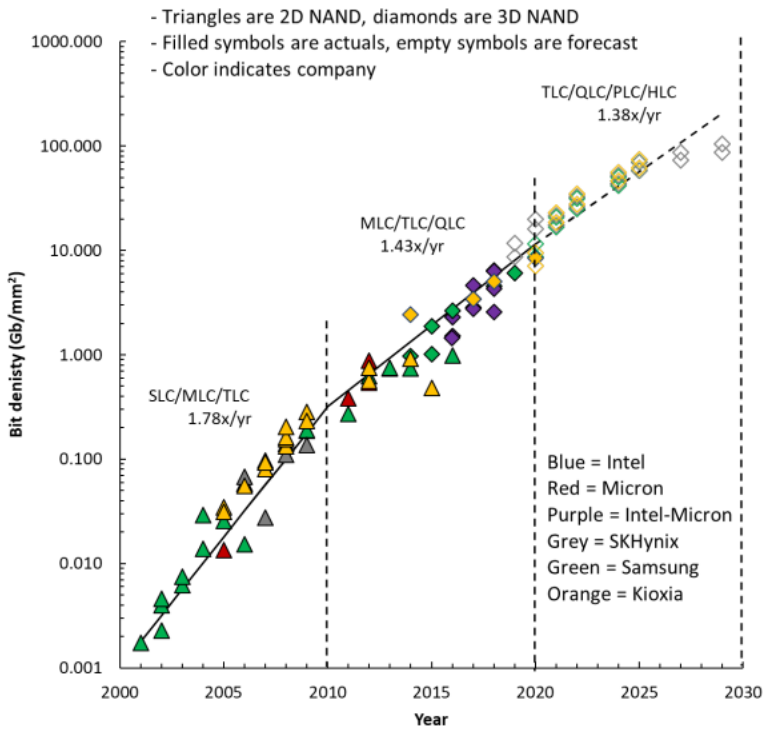


L2P Entry = Physical address on NAND, including IU within a page

DRAM density scaling vs. NAND scaling



DRAM Bit Density [1]



NAND Bit Density Versus Company and Year [1]

NAND Scaling ~ 1.38x/yr
DRAM Scaling ~ 1.13x/yr

Reference: 2020 LithoVision "Economics in the 3D Era", Scotten W. Jones
https://semiwiki.com/wp-content/uploads/2020/03/Lithovision-2020.pdf

Fitting L2P tables in DRAM will become increasingly difficult in next 10 years
May need up to 64K IU

Solutions to Reduce DRAM Needs

Store L2P in NAND:

- Inconsistent QoS for random workloads

New L2P structures:

- Complexity and may not solve PCB space up to 512TB/1PB

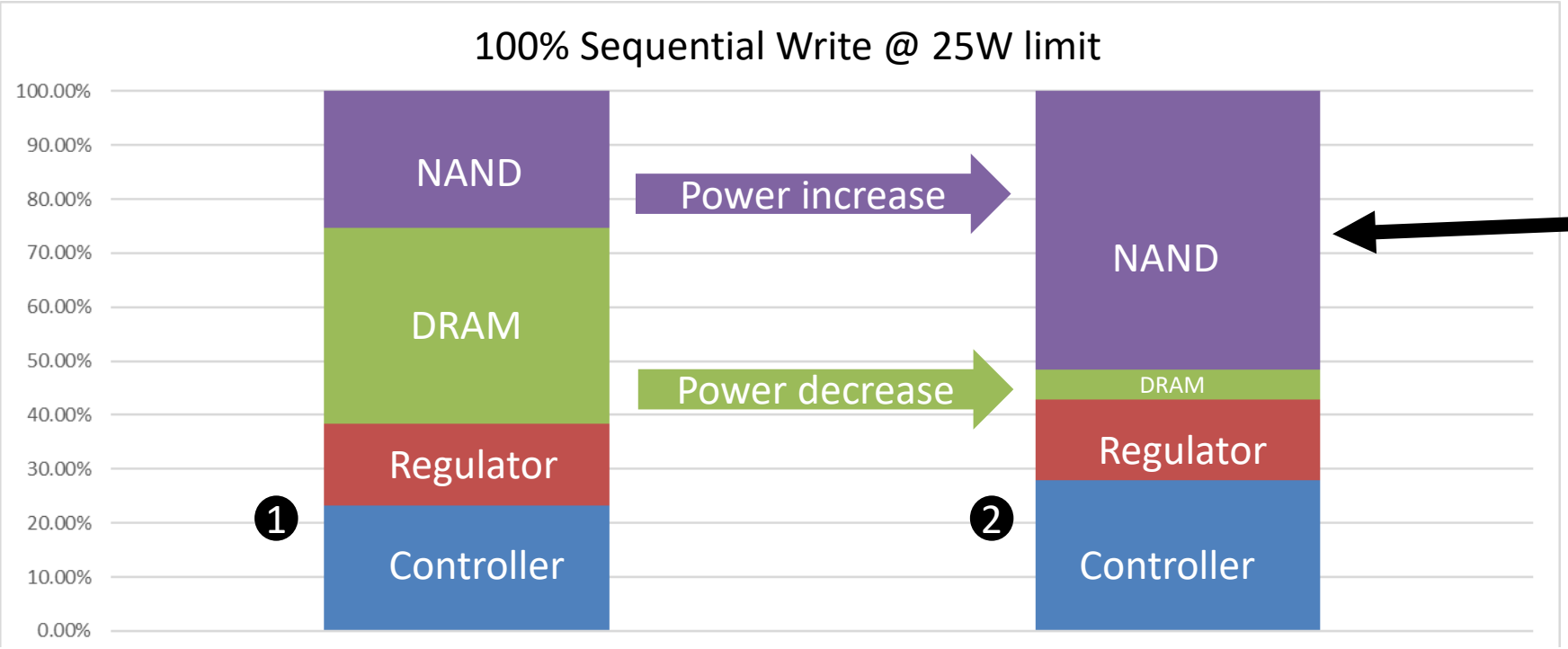
Increase Indirection Unit Size

- Increases write amp if transfer size < IU size
- May want to spec endurance in terms of 4K and IU size

Increased IU gaining acceptance in industry

Representative Power Benefit from Larger IU

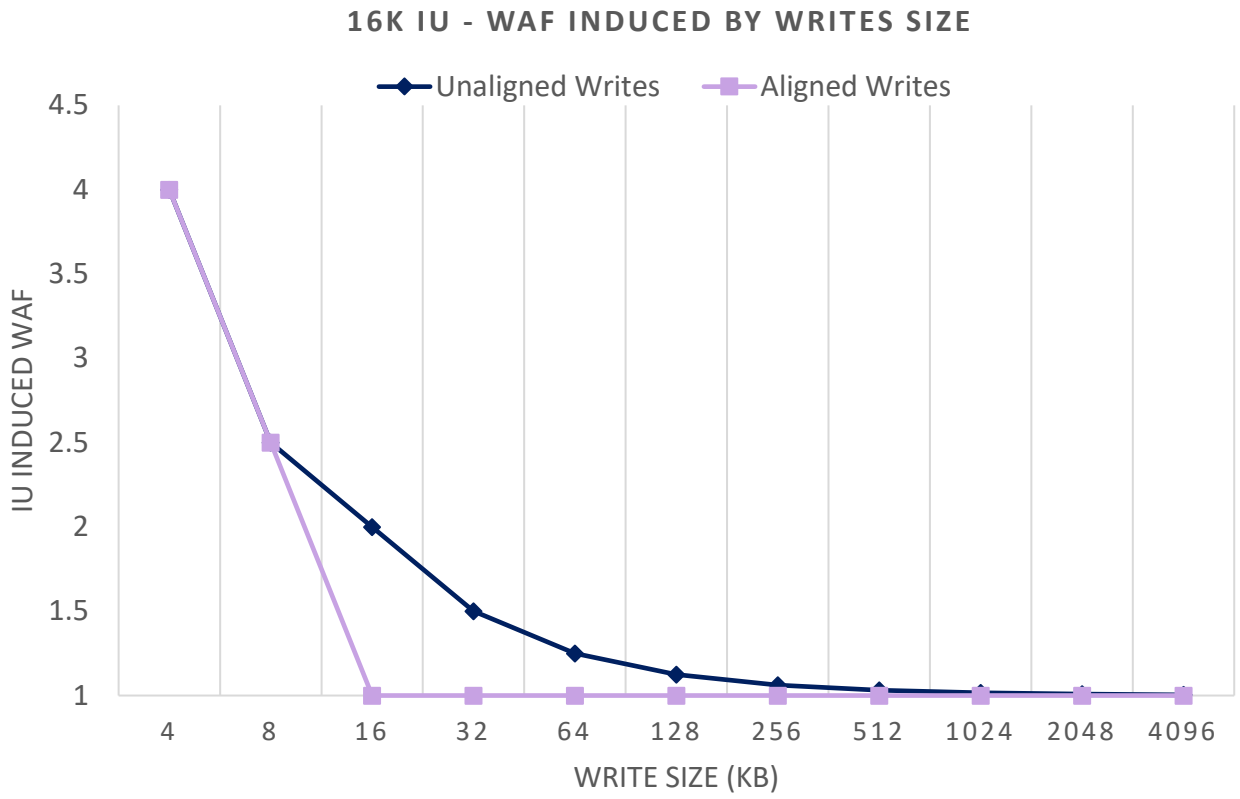
Less DRAM -> Additional power for NAND



Drive	SS Sequential Writes MB/s	SS Seq Reads MB/s
1 64TB w/ 4K IU, 128GB DRAM	4.6 GB/s	8.8 GB/s
2 64TB w/ 16K IU, 16GB DRAM	9.6 GB/s	14.3 GB/s

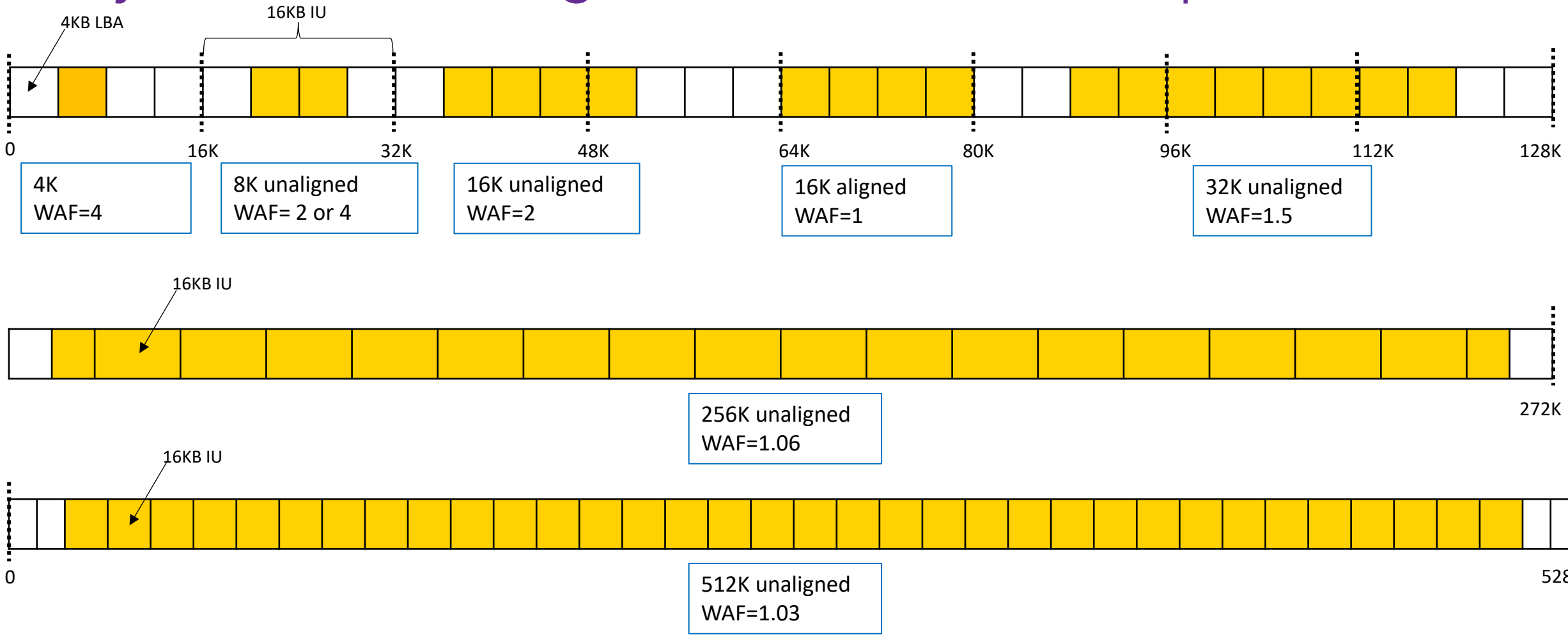
16K IU induced WAF: larger IOs have smaller impact

- Impacts only writes
  - Reads impacted if IU > NAND page
- Impacts only 16KB-boundary misaligned writes
  - 16KB-boundary aligned Writes do not introduce any WAF
- Impact decreases exponentially with increased Write Size
- Aggregating and issuing large Writes is a good idea



Aligned transfers >= IU size do not induce extra WAF

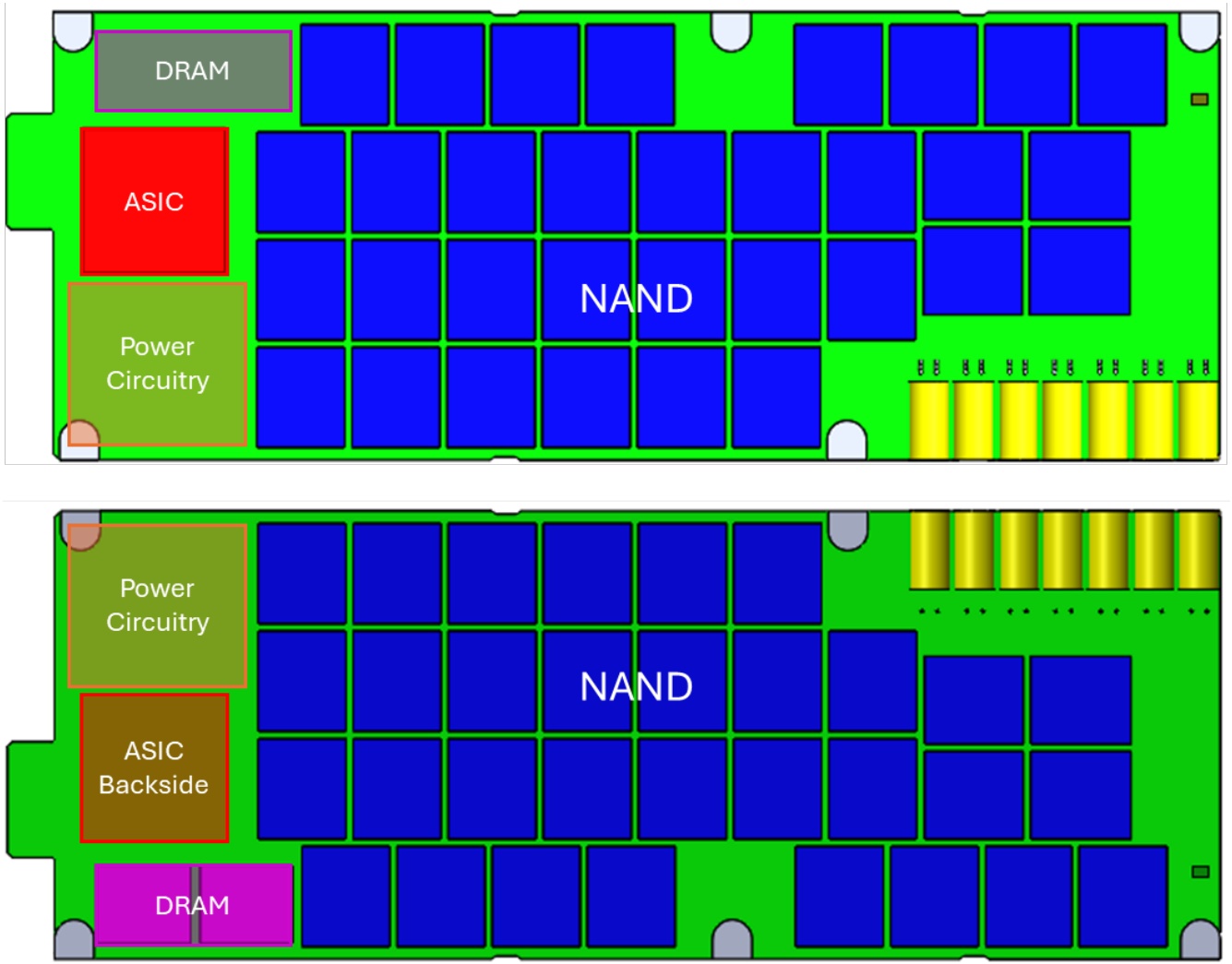
Why size matters: larger IOs have smaller impact



Head and Tail IU are the only ones affected. All others are “aligned” by definition

Form Factors for High-Capacity SSD

- NAND up to thousands of die
  - NAND package height not growing, so die stackups may be limited
- DRAM and holdup capacitors also increasing with drive capacity
- New form factors may need to be considered for component space, thermals, and server fit



Example 512TB E2 SSD

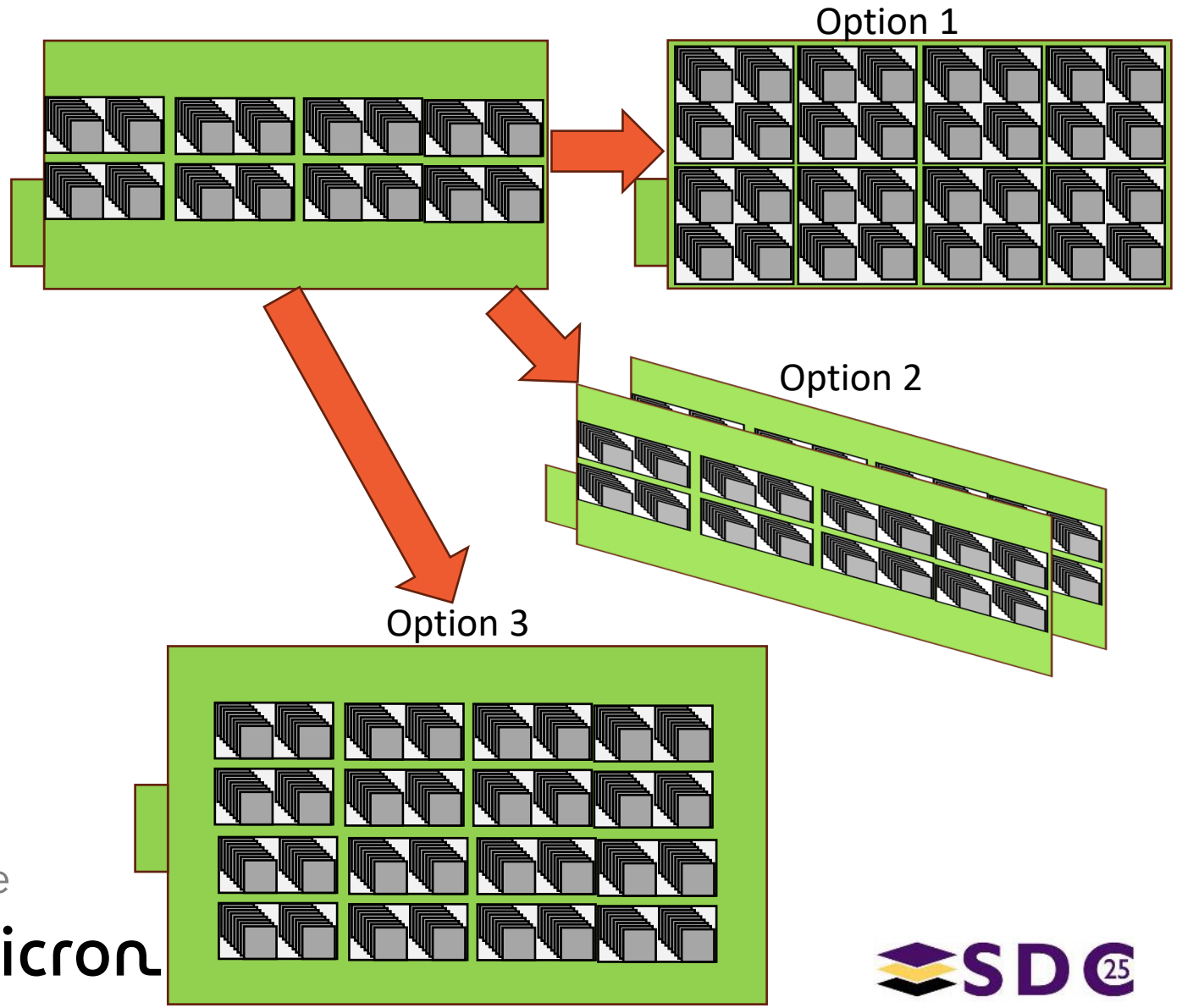
Refer to Anthony Constantine's SDC presentation for more about E2 form factor

micron

SDC 25

Challenge: Cost optimized capacity

- How to get 1024+ NAND dies, more holdup caps, and more DRAM in a single SSD?
- Option 1: Use current SSD form factors and Double the NAND dies/package
  - 32 NAND packages x 32 dies/package
- Option 2: Use current SSD form factors but make them thicker.
  - 64 NAND packages x 16 dies/package
- Option 3: New form factor
  - 64 NAND packages x 16+ dies/package



micron

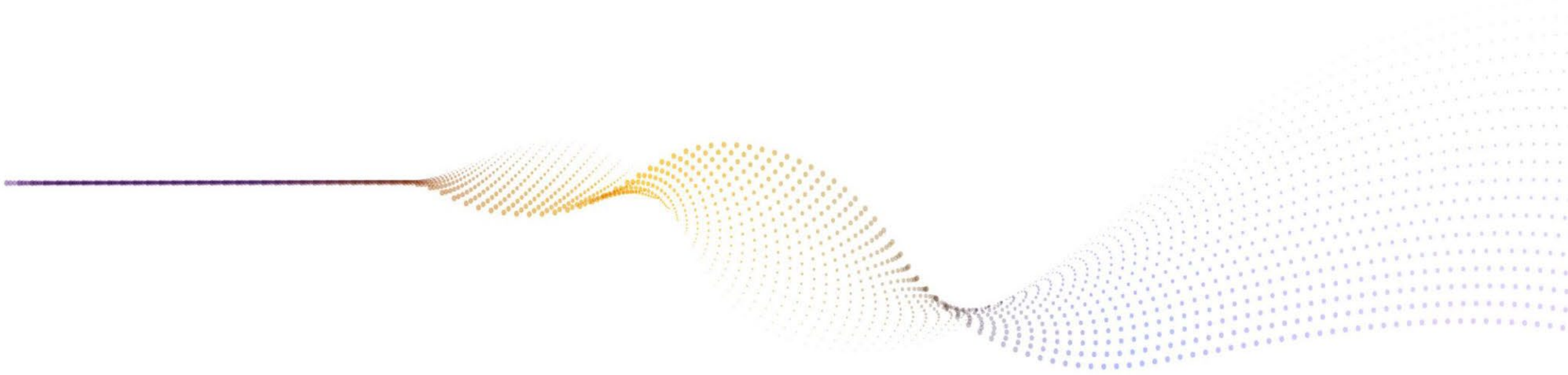
SDC 25

Conclusion

- AI Use-cases are driving diverging storage requirements
- Near GPU SSDs – Need High Seq R/W BW, support liquid cooling
- Capacity SSDs – Need to optimize density and cost
- Both need to be power-efficient

micron

SDC 25



Thank you for attending!

Please remember to rate this session. You can get access to the presentations at
http://sniadeveloper.org/conference

