Accelerating Time to Insight

Performance optimized emerging system architectures

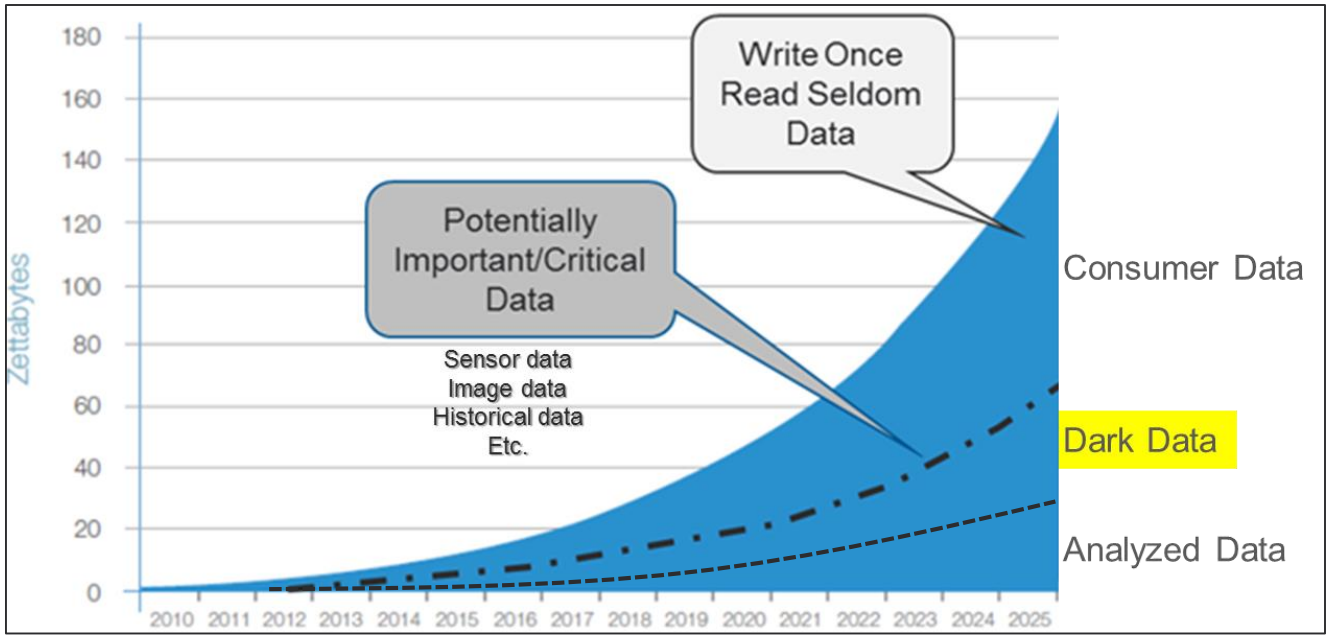
Balint Fleischer

Senior Director

Advanced Computing Solutions

September 25th, 2019

Rapidly growing *Dark Data*



Rapidly growing Data sets from many sources

+

Increasing complex algorithms

+

Need for faster Time to Insight

+

Affordability challenges

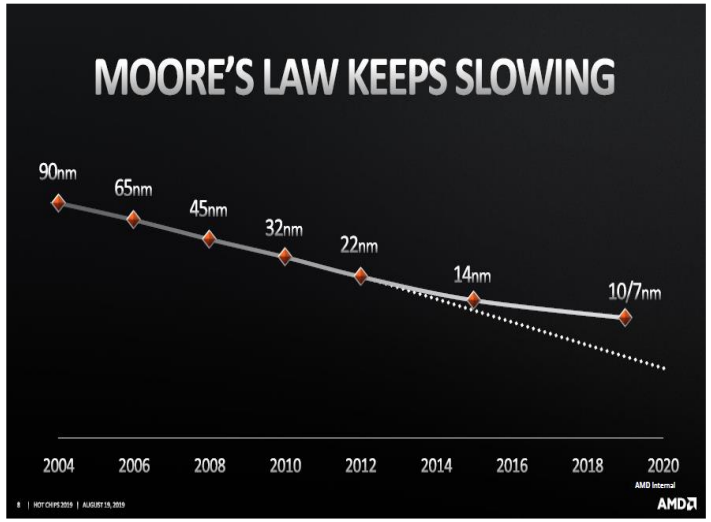
Dark Data

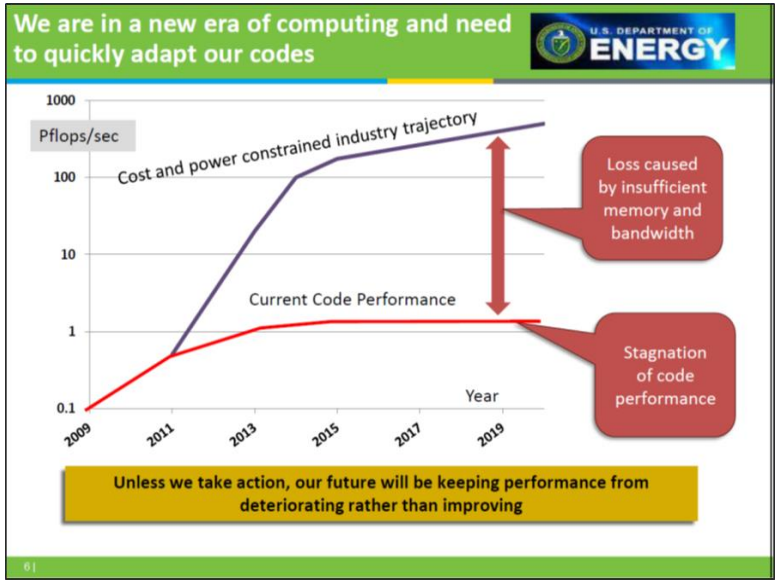
Estimated data[2]

1 Zettabyte = 10^21 (one billion, trillion) bytes

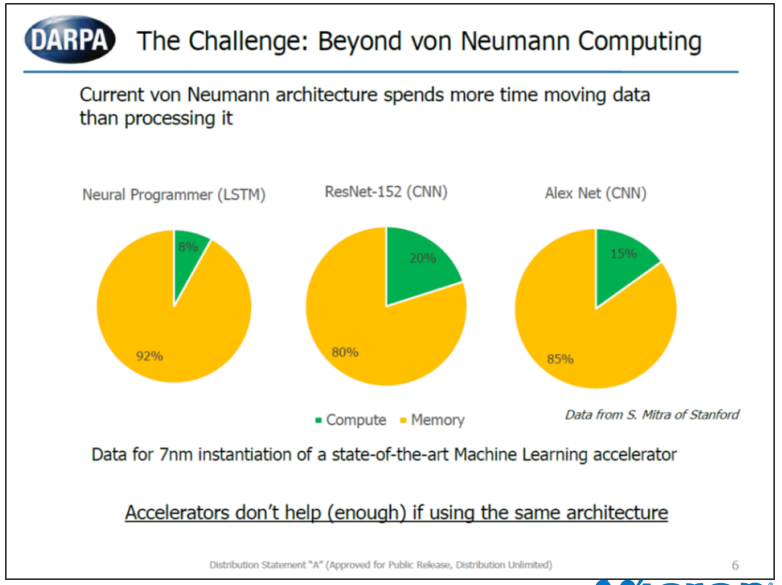
Perfect storm:

Moore's law is slowing, Dennard scaling is ending and von Neumann architecture became a bottleneck



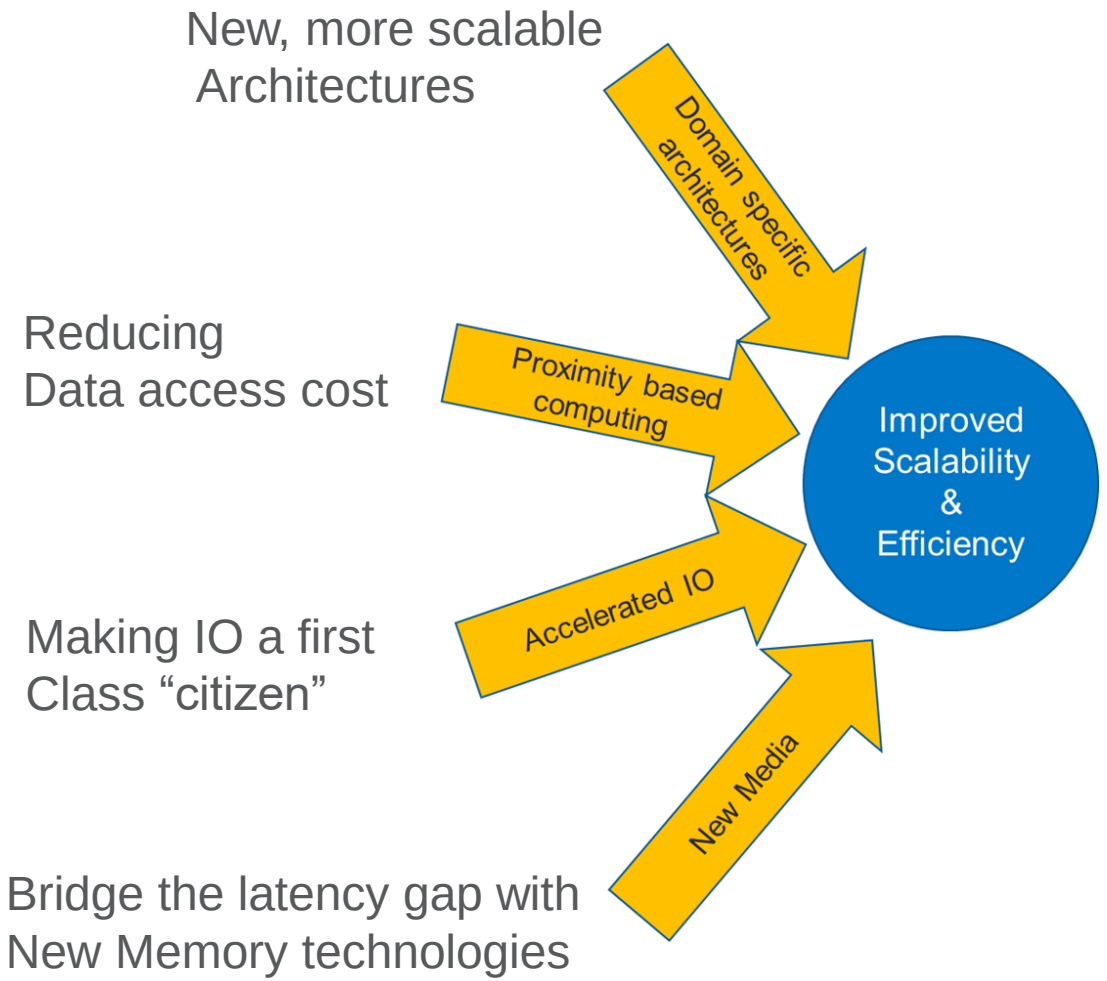


B. Meisner, The Bump in the road to ExaFlops and Rethinking LINPACK, HPC User Forum, June 2014

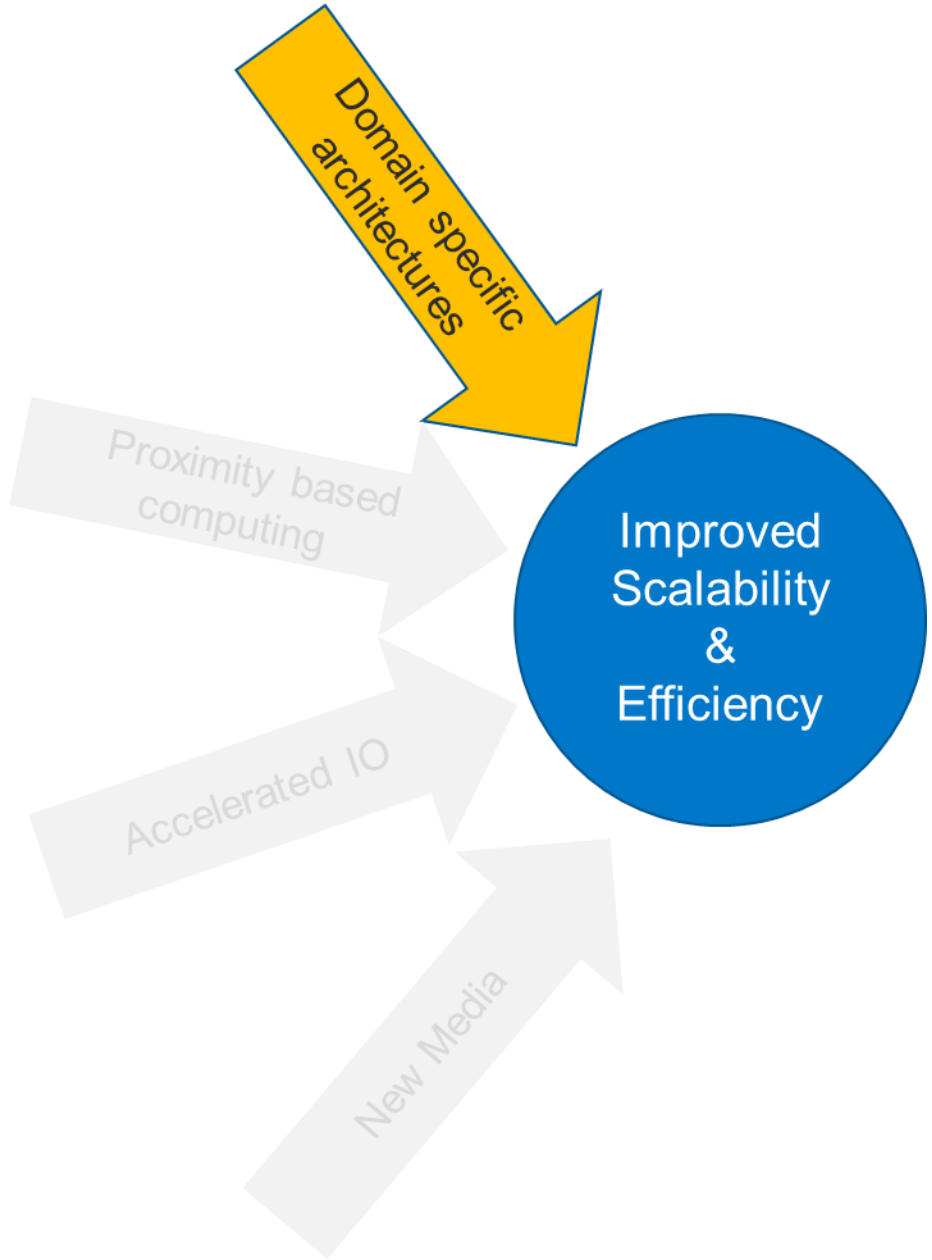


Micron

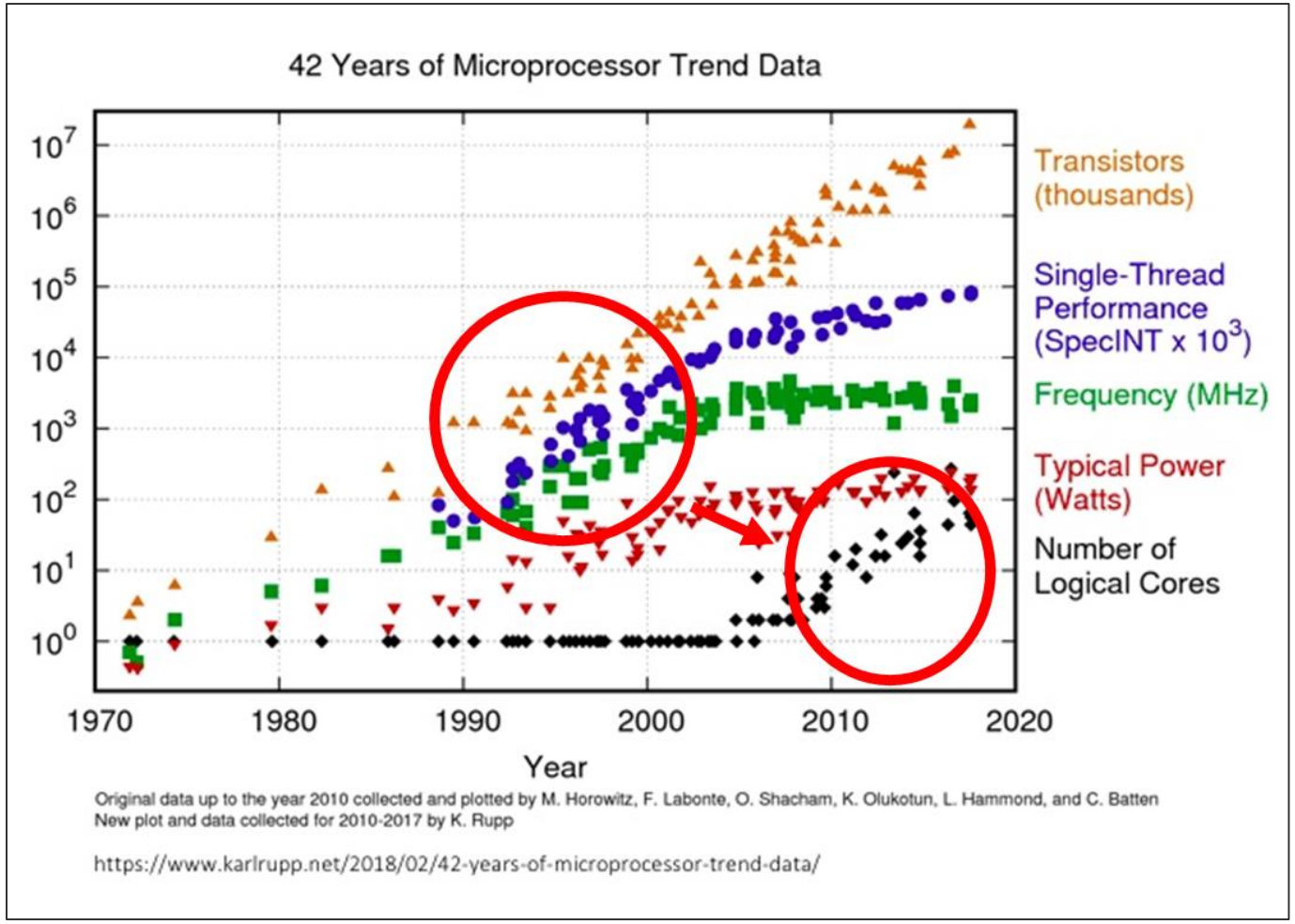
Large, industry wide approach but without a master plan to insure continued scaling



New, Non Von Neumann Architectures Scale well (for a while) with process technology improvements



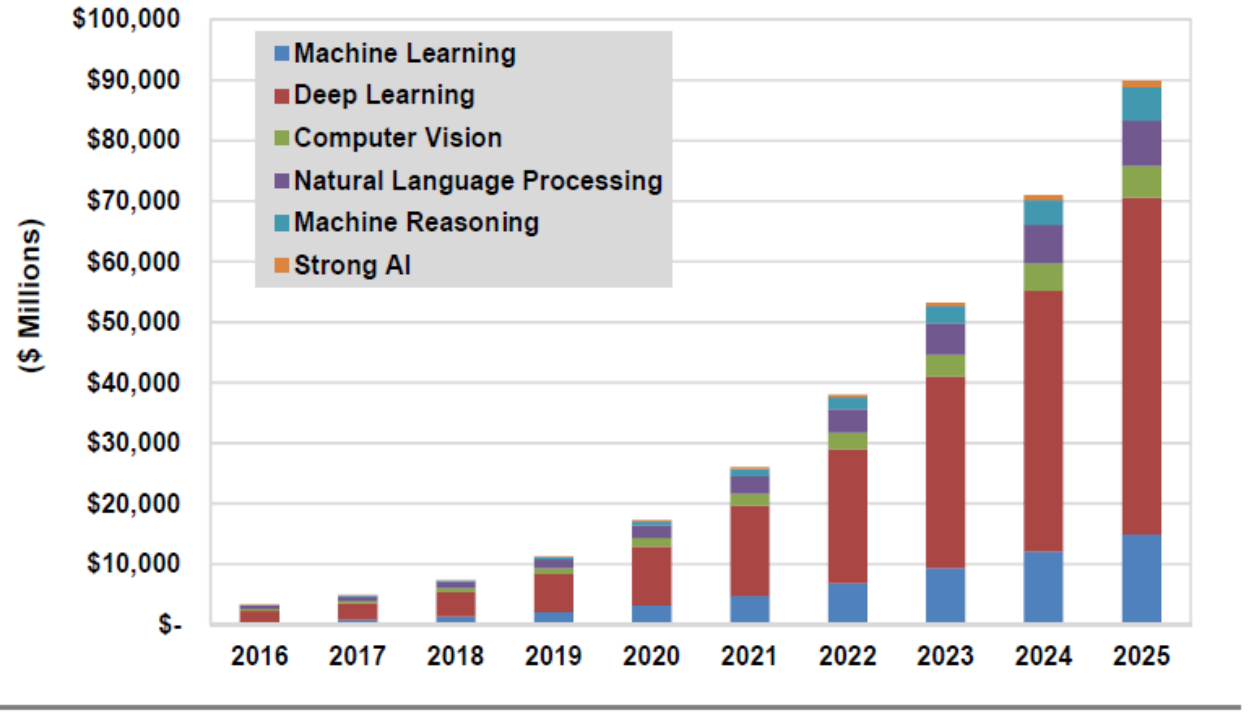
General purpose CPU performance CAGR is declining



Machine Learning
+
Deep Learning emerging as the key application for data analytics

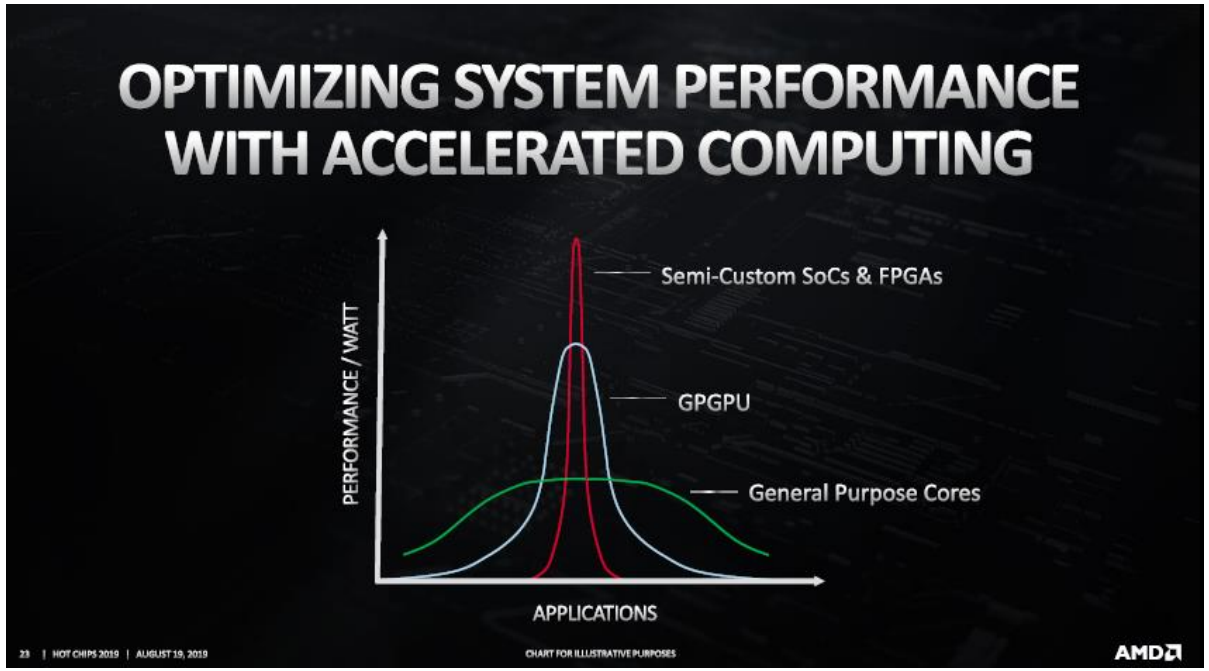
A rich target for Domain specific computing

Chart 3.5 Annual Artificial Intelligence Revenue by Technology, World Markets: 2016-2025



(Source: Tractica)

On a given technology node, Domain Specific Architectures deliver greater performance on targeted applications



Pattern for domain specific architectures

Simpler, lower performance, more efficient processing elements

High degree of parallelism (1000s of processing elements)

Highly optimized on die data movement

New, application specific ISA

Custom compiler

An example of Domain Specific Architecture delivering greater performance

Optimized AI processors deliver even better gain

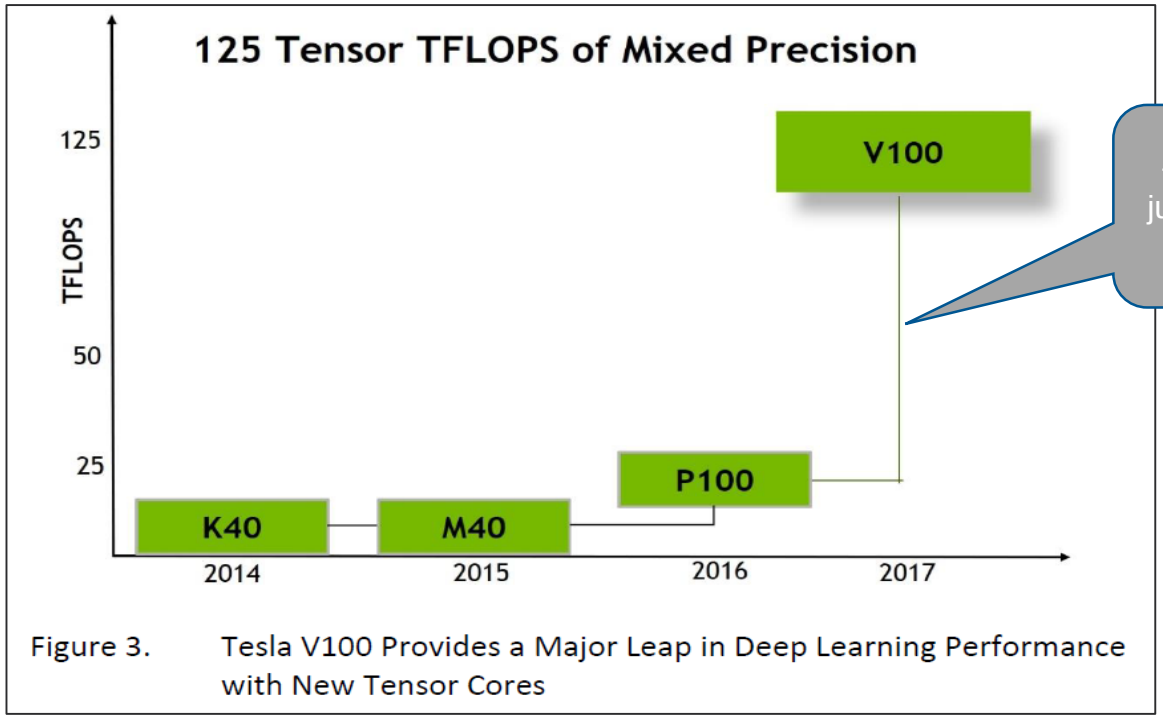


Figure 3. Tesla V100 Provides a Major Leap in Deep Learning Performance with New Tensor Cores

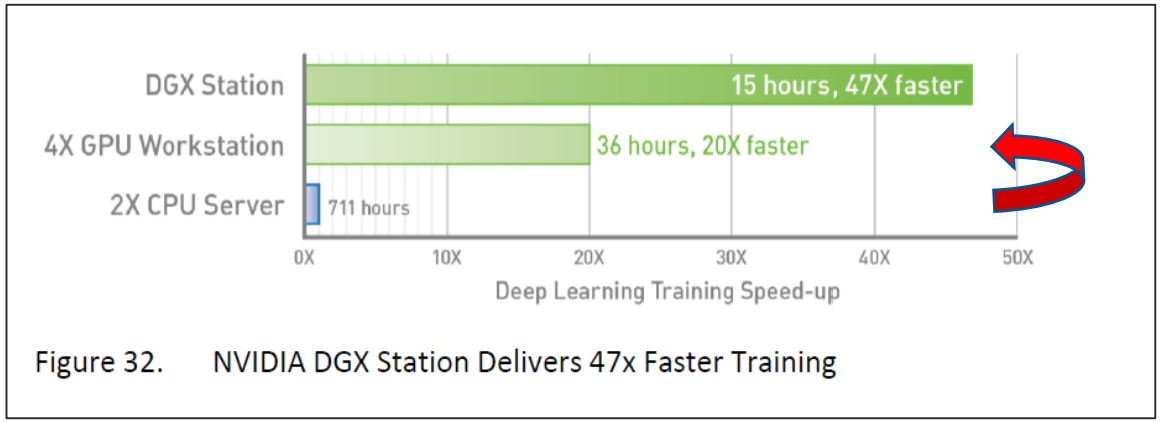


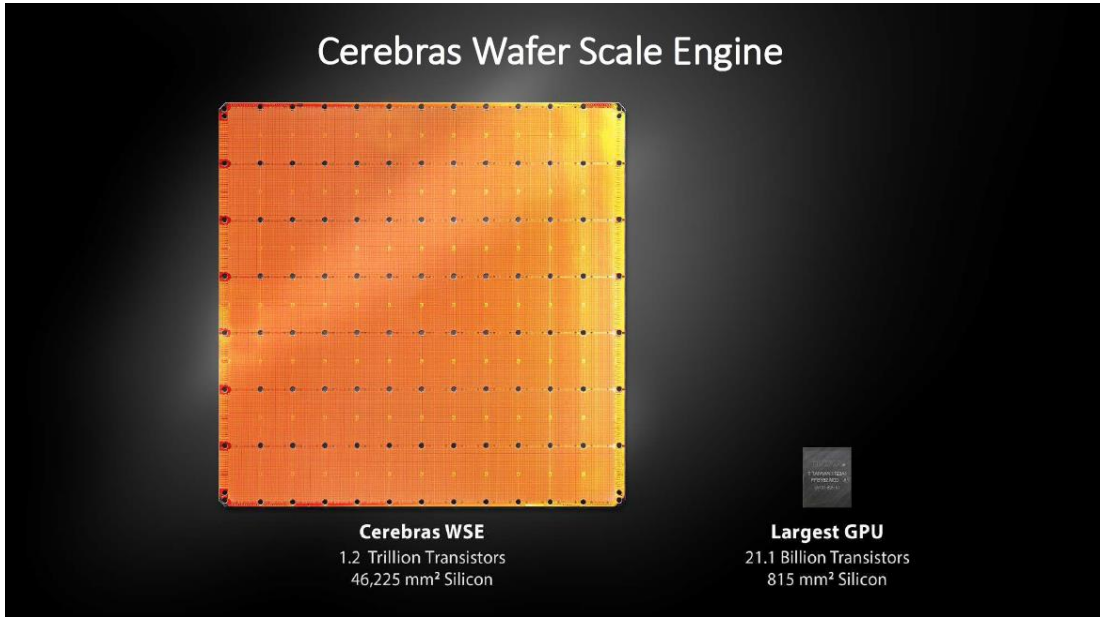
Figure 32. NVIDIA DGX Station Delivers 47x Faster Training

There are over 300 AI processor designs worldwide innovating from Low End to High End

The Hailo-8 achieves 26 TOPS (Image: Hailo)

EETimes 08.29.10 Details of Hailo AI Edge Accelerator Emerge

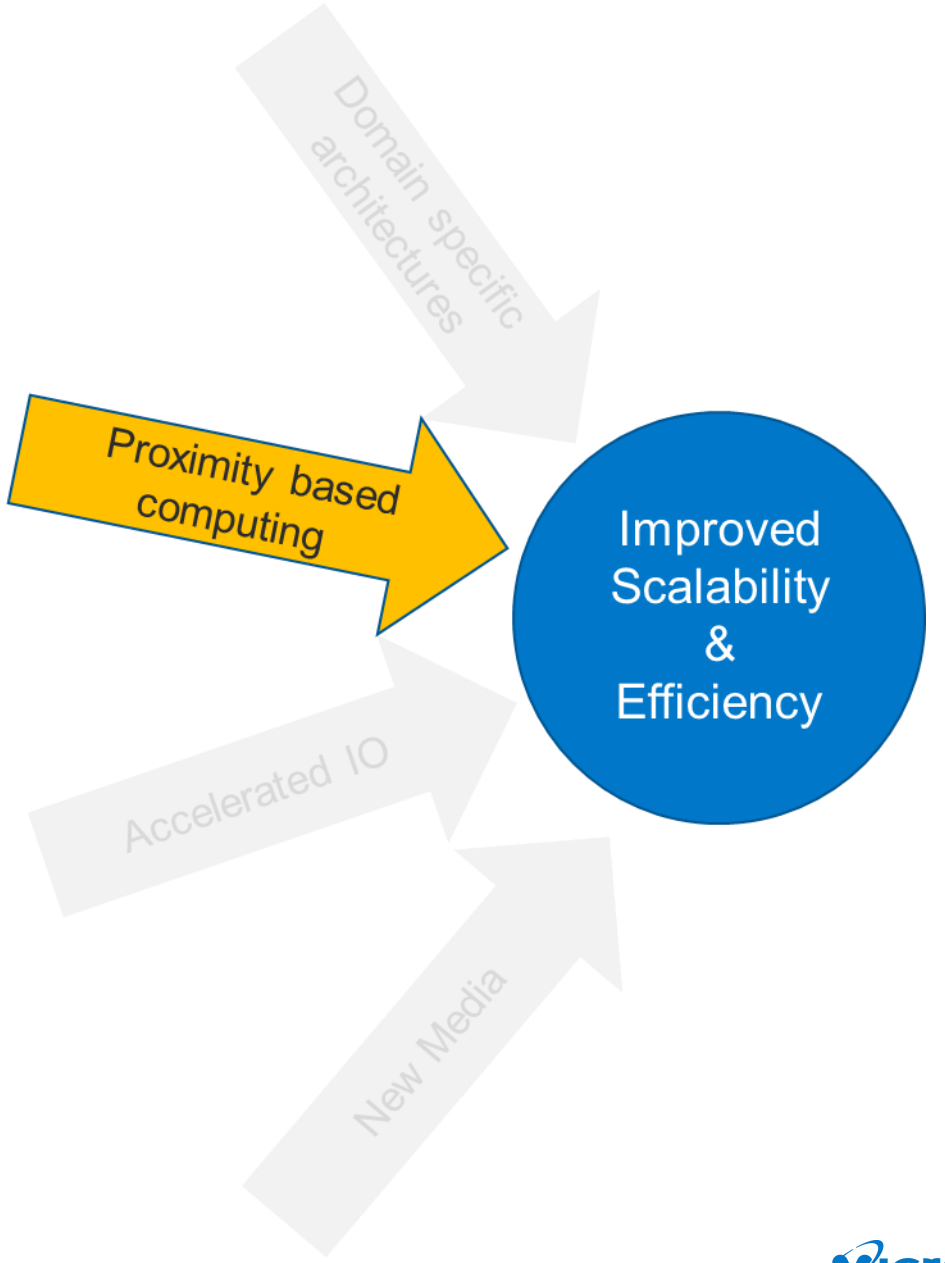
Claimed performance
2.8 TOPS @1.6W
~2x faster vs. CPU
1/15th power vs. CPU



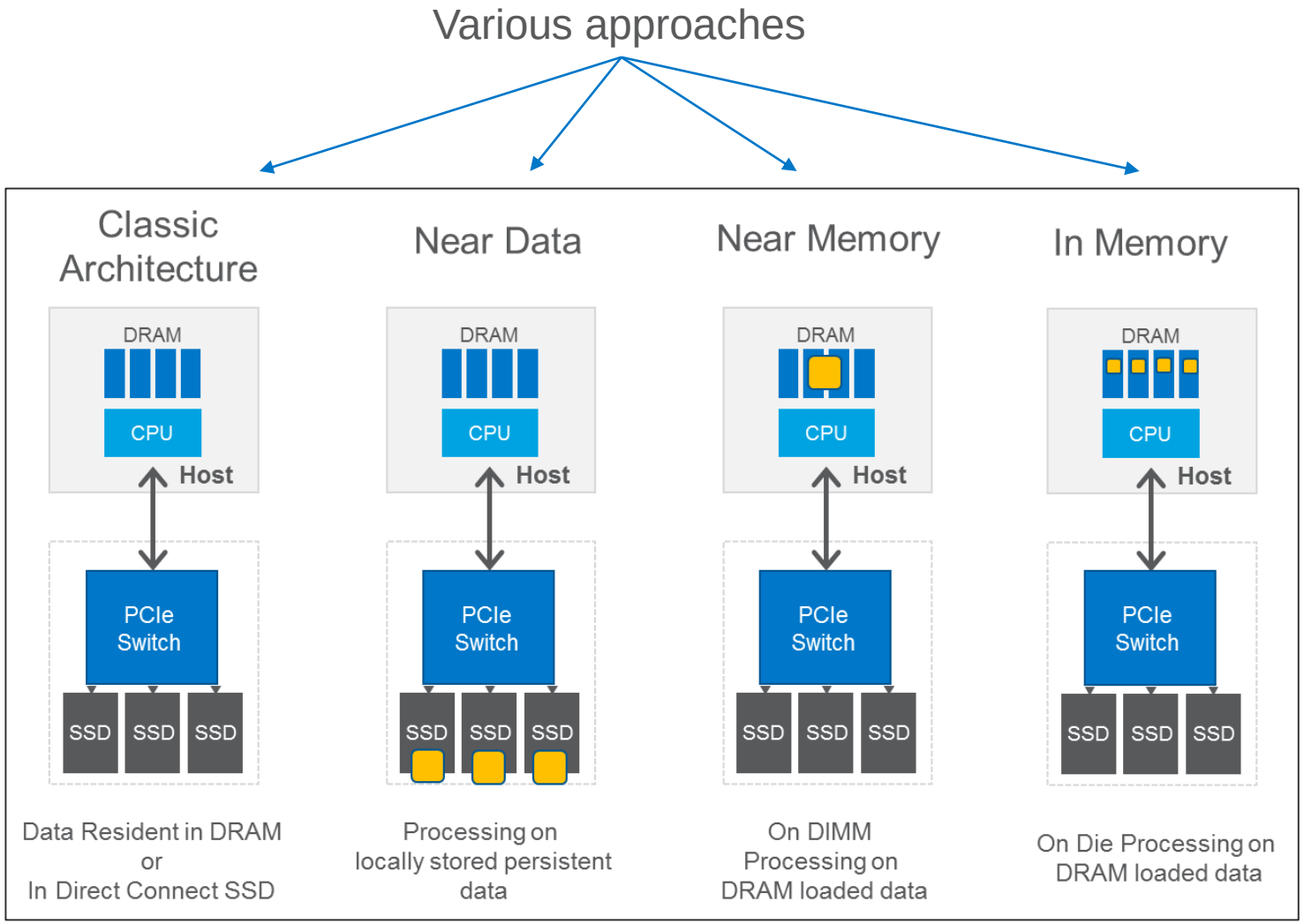
Hot Chips 2019, Cerebras Presentation

“Cost” of IO is critical to performance scaling and energy consumption

The goal is improving Bandwidth, reducing Latency & reducing Data movement

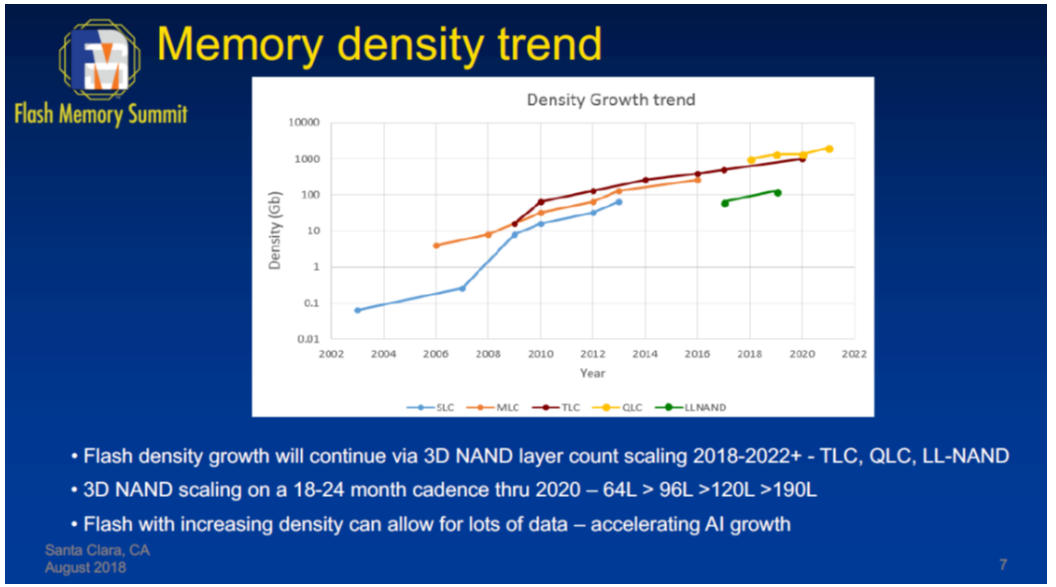


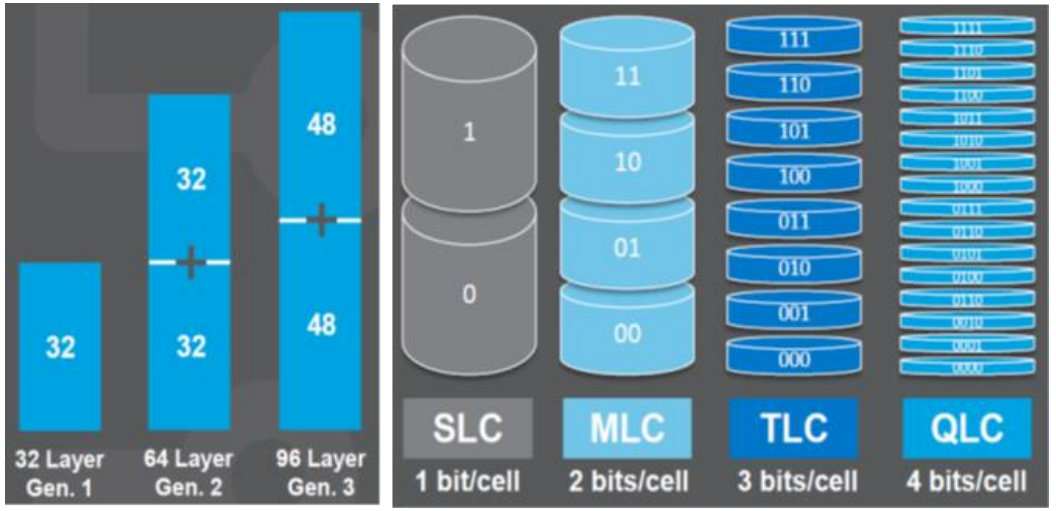
Improving Data Proximity is a vehicle to address IO cost



* For simplicity single socket CPU Hosts are shown

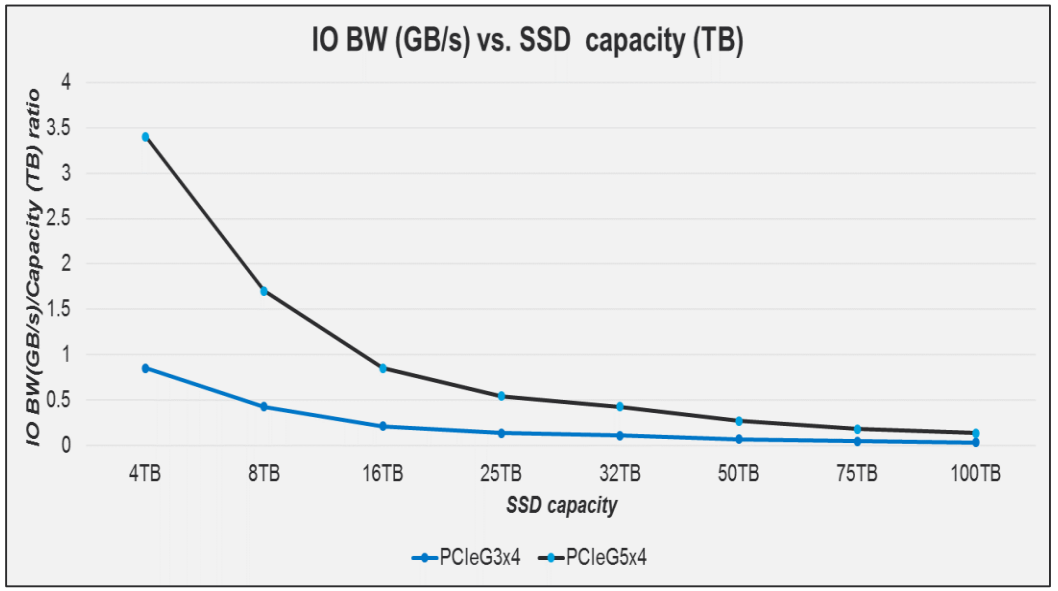
NAND die capacity continues to grow enabling more dense SSDs and storage solutions





Difficult to Beat the NAND Cost Structure

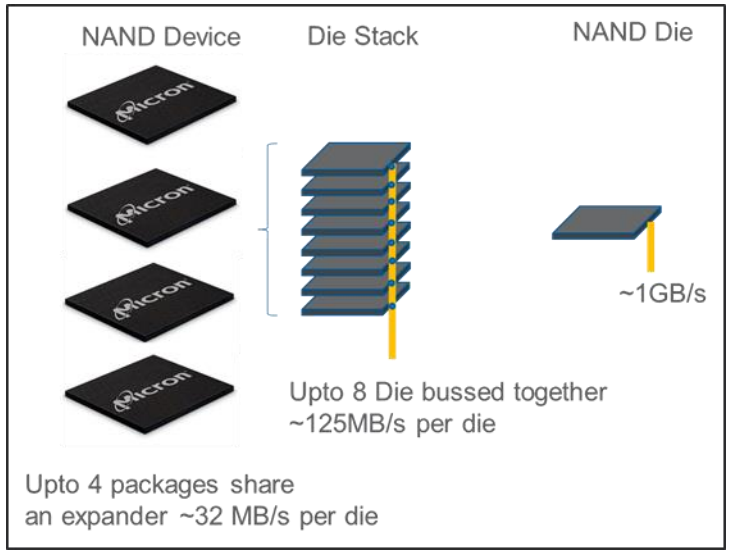
SSD IO Bandwidth is lagging SSD Capacity growth

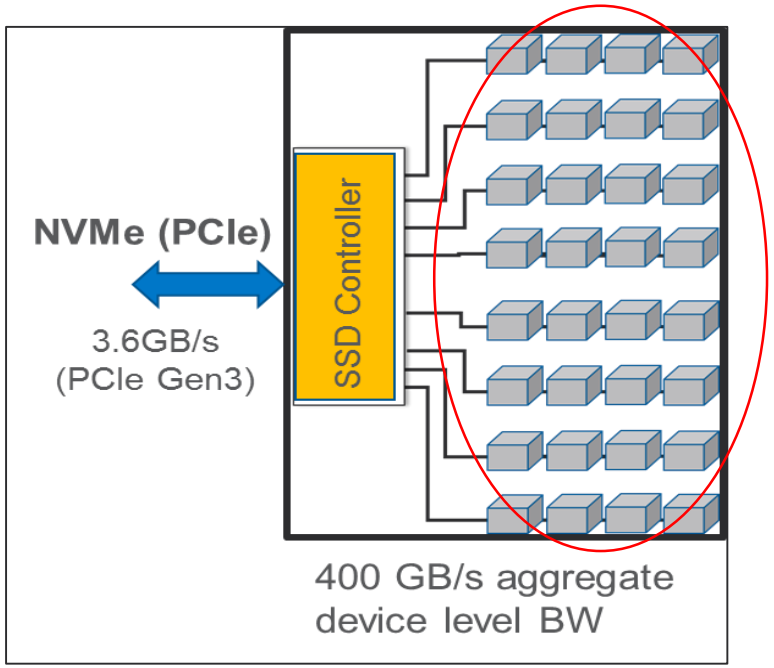


Query of very large data sets
Will take an increasingly long time

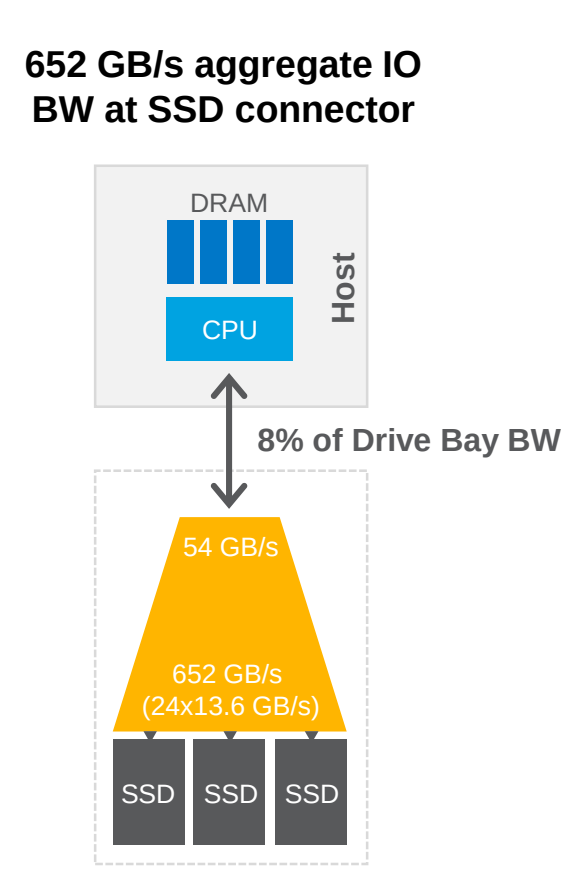
NAND stacks and SSDs are designed for capacity scaling

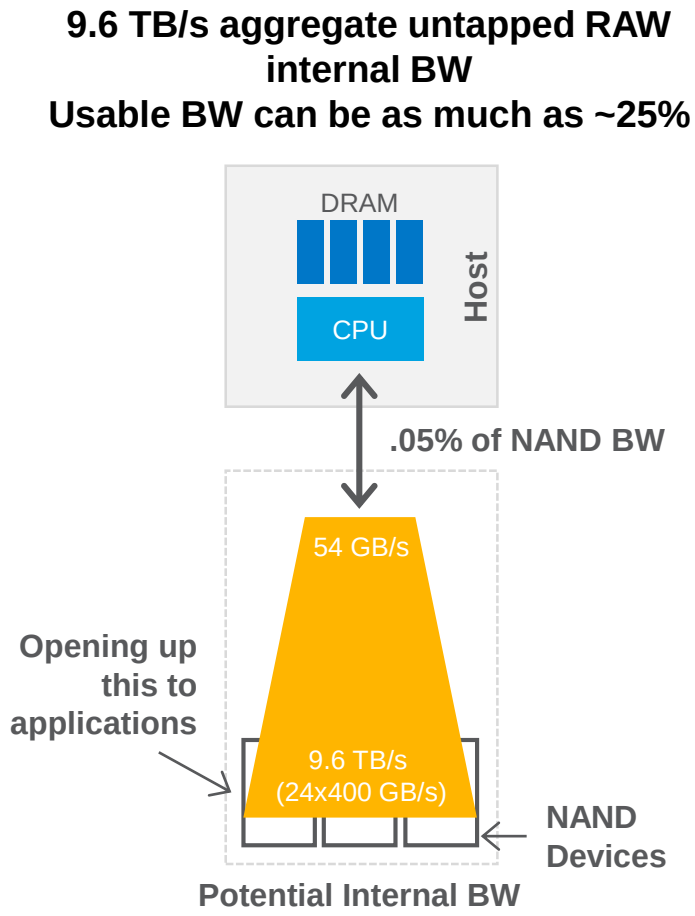
Device level aggregate BW is ~100x vs. SSD External IO Bandwidth Die level Bandwidth is ~1000x





Moving processing into SSD can benefit from the high Embedded Bandwidth
Time to Insight on very large shardable data will also benefit from Massive parallelism



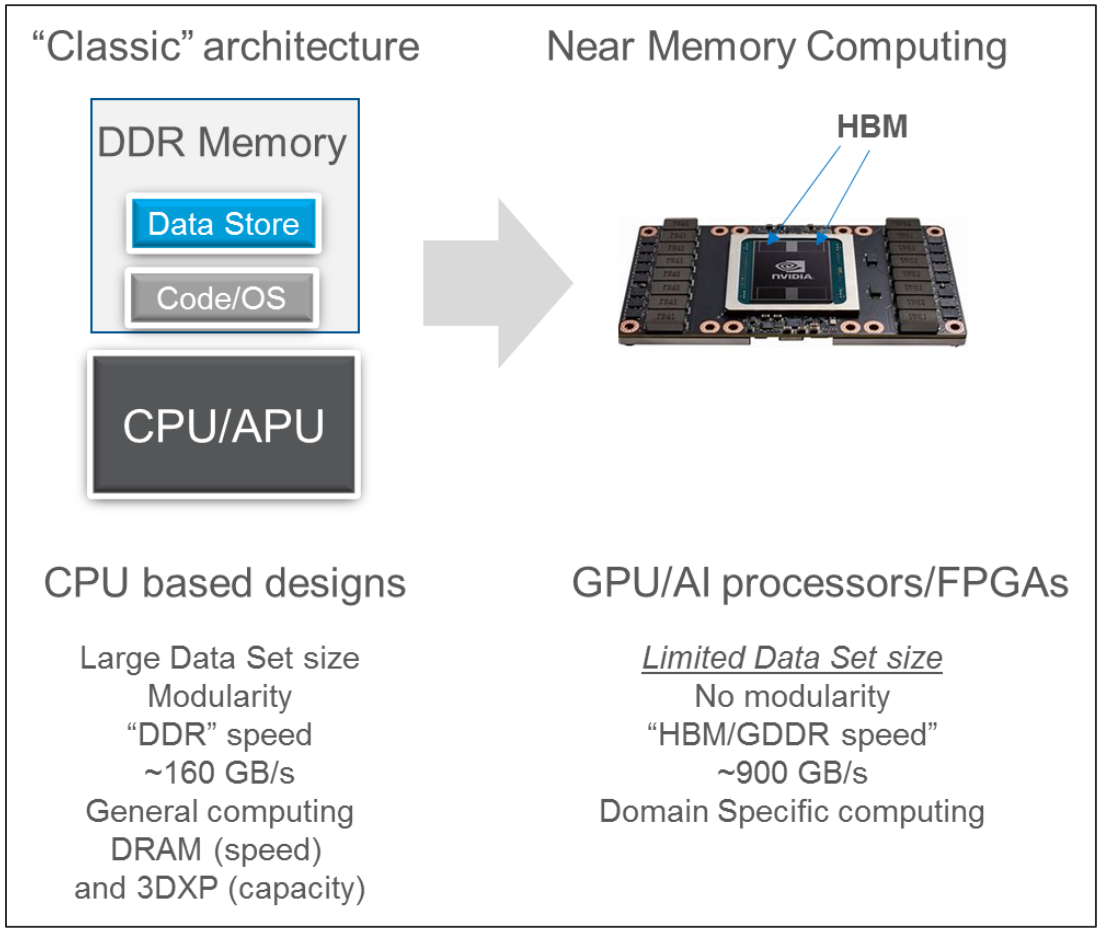


Parallel Processing coupled with Near Data Computing enables faster time to insight

* For simplicity single socket CPU Hosts are shown

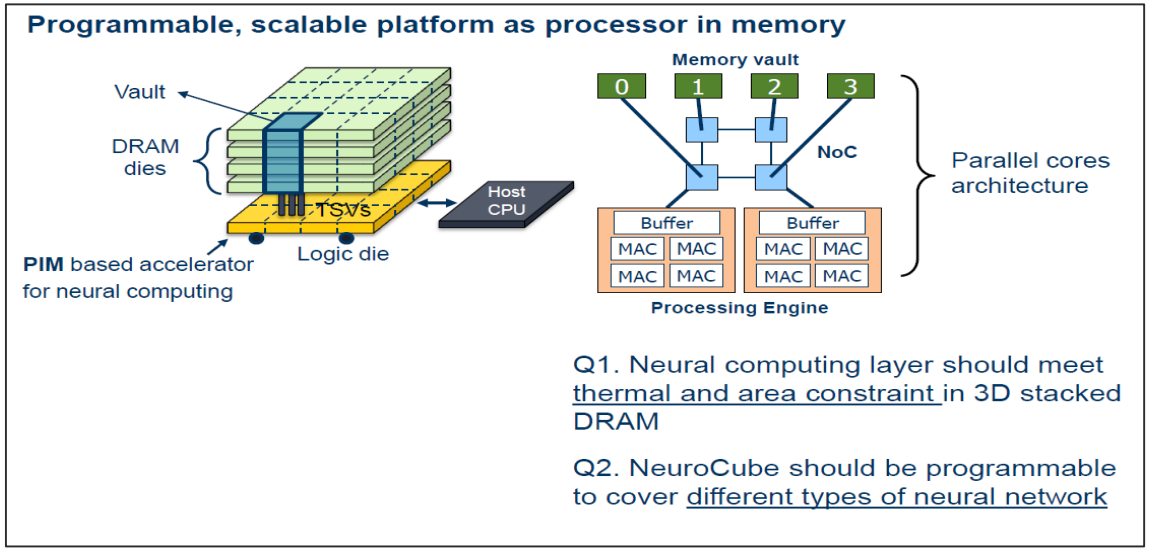
Note: In both cases there are 24 drives per JBOF, identical PCIe G5 NVMe interface.

Near Memory Computing can Deliver 5-10x better Bandwidth to applications in a 2.5D packaging technology



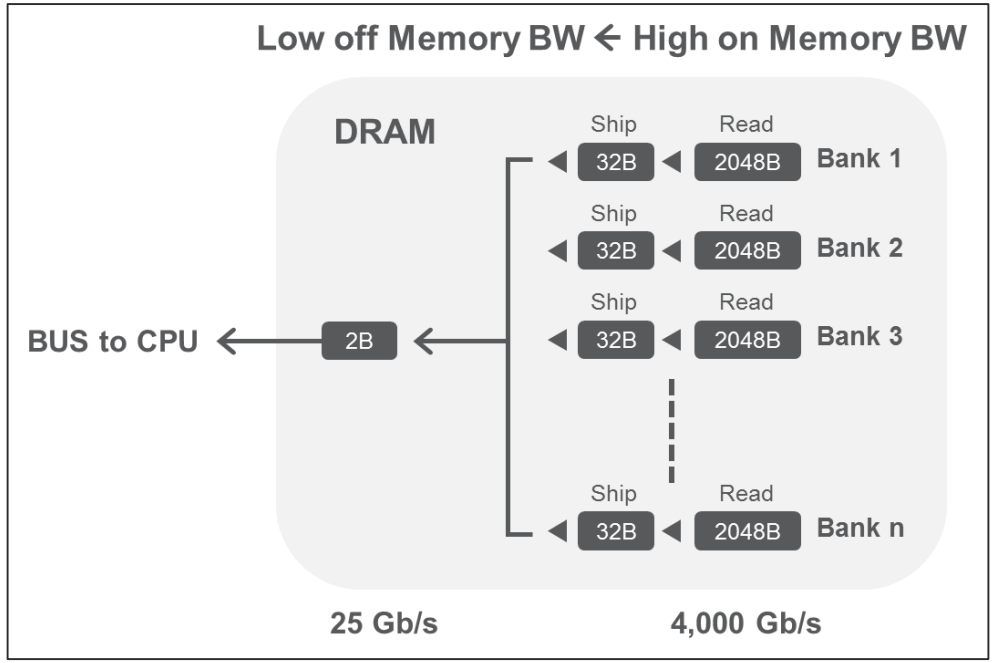
Tighter integration of processing and memory can lead to even greater Performance at lower Power

Near Memory Computing example



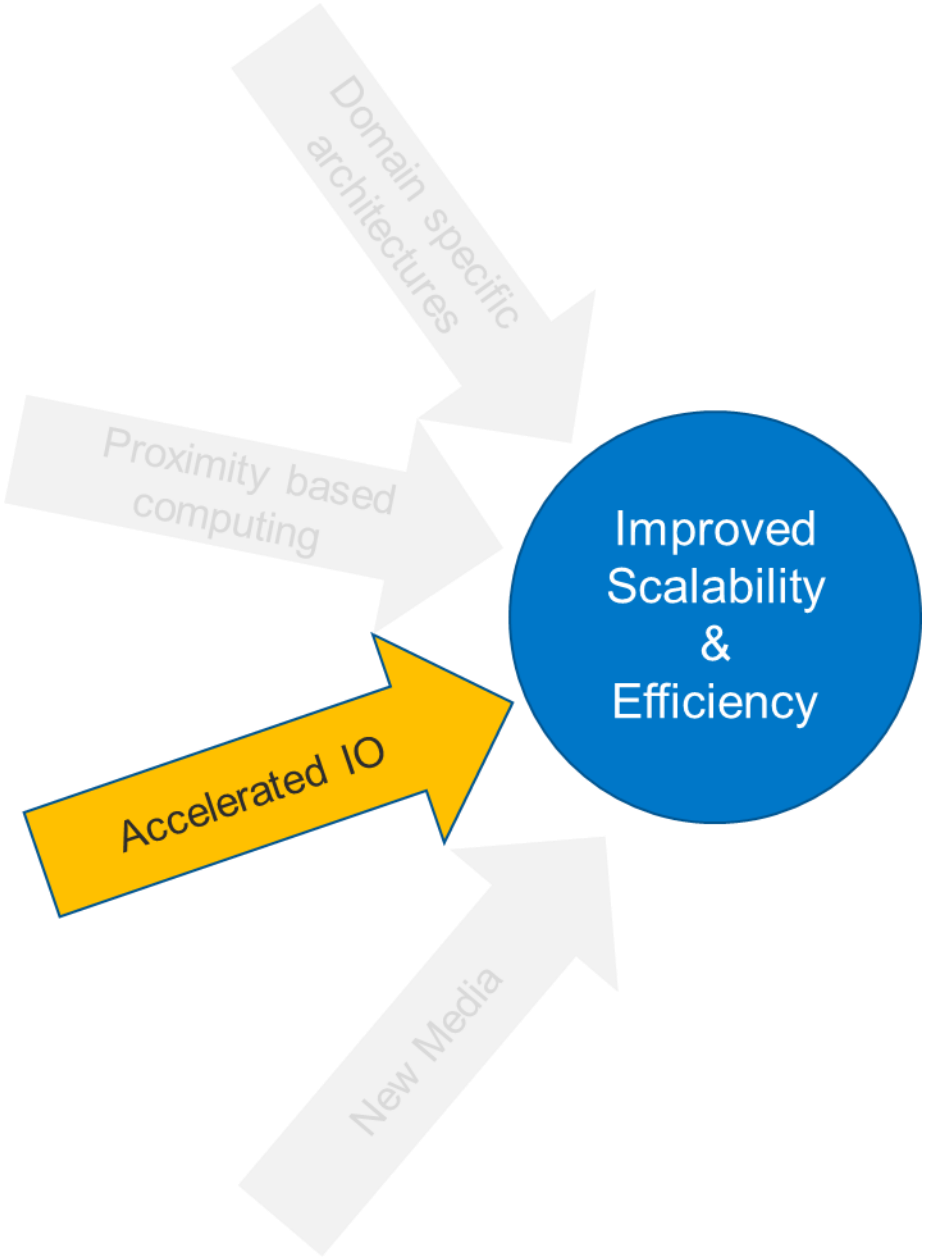
[Kim et al., NeuroCube, ISCA 2016]

On Die integration of compute and memory can take advantage of ~160x on Die bandwidth on smaller Data sets



Processing Element

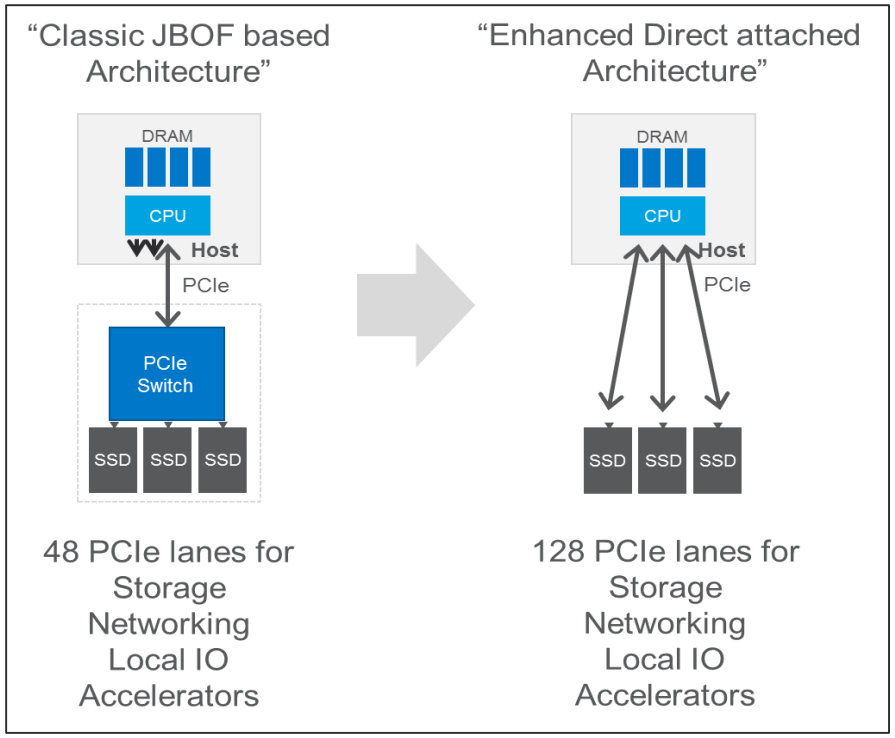
Large investment into improving Platform IO

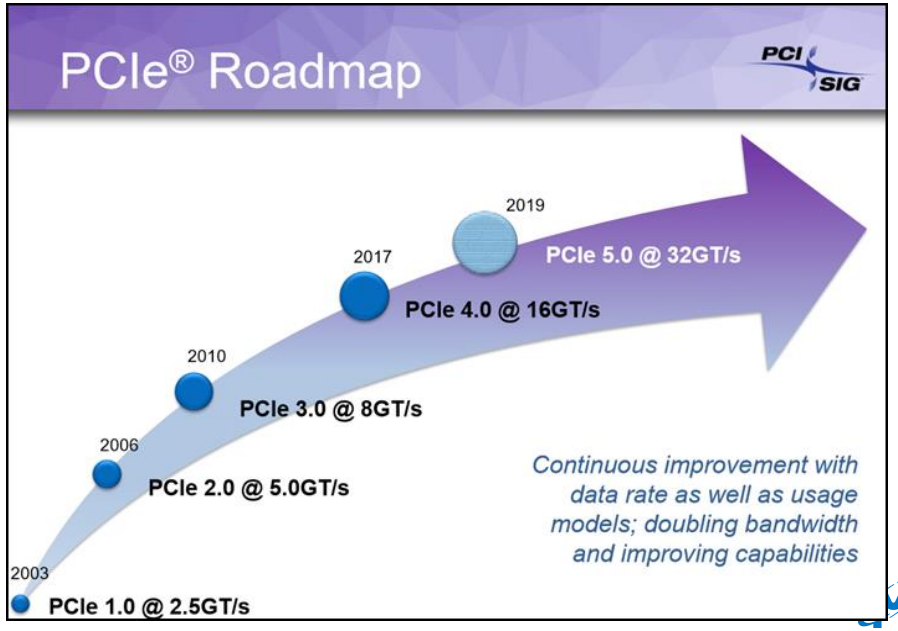


Increasing integrated IO lane count to grow connectivity

Faster PCIe will enable higher speed devices

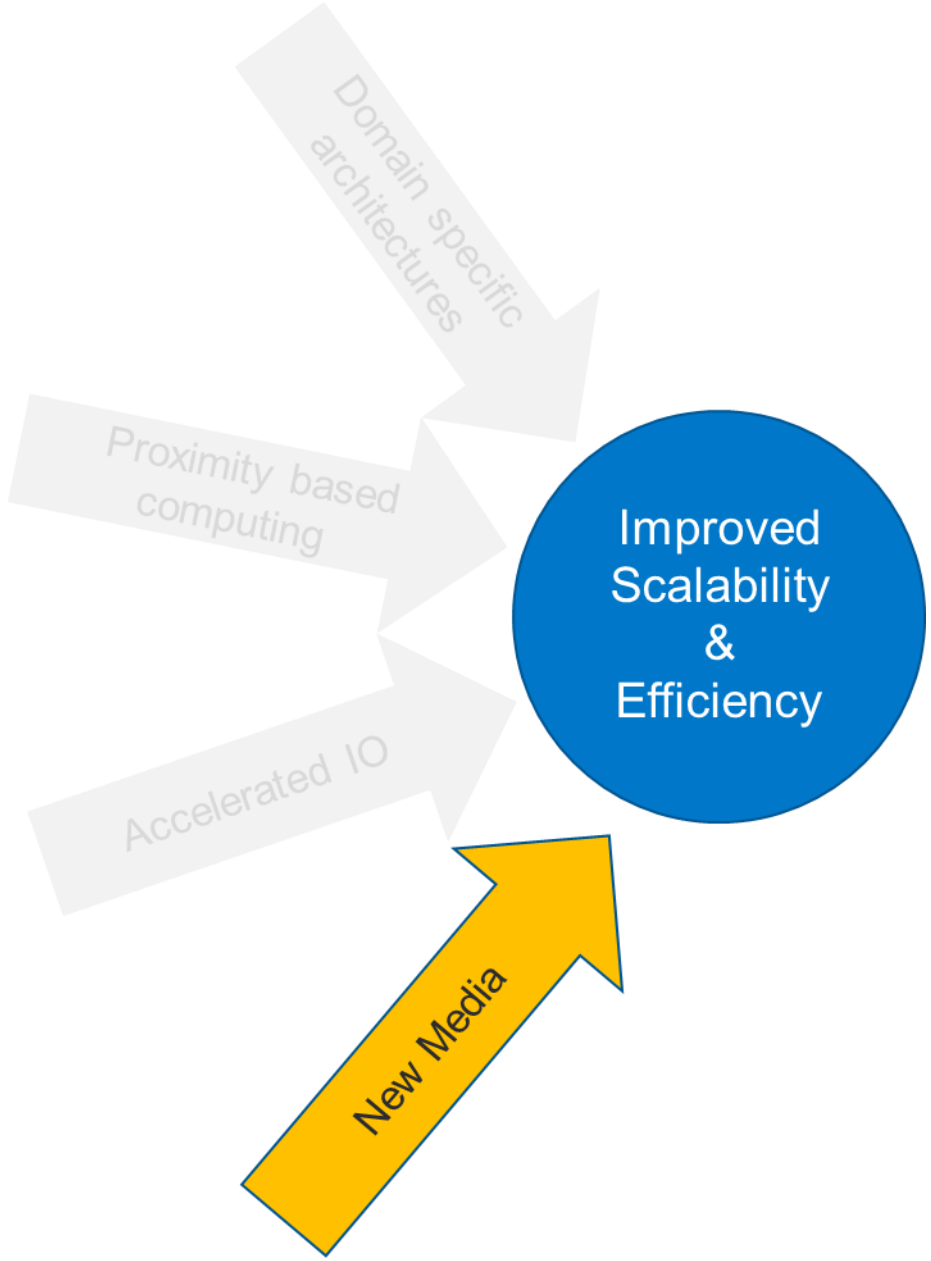
Silicon photonics to extend reach





* For simplicity single socket CPU Hosts are shown

Emerging Media technologies to improve capacity, latency and access methods

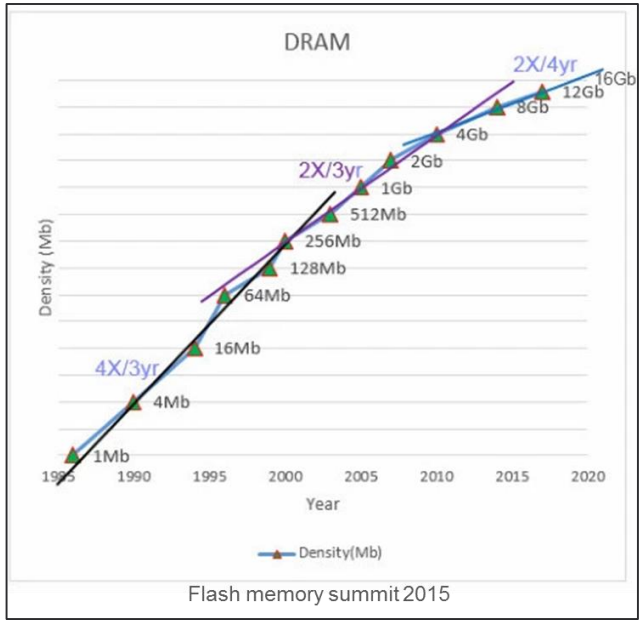


Emerging Memory research is intensifying

Difficult to beat DRAM Performance & Energy

But

DRAM density growth is slowing



Flash memory summit 2015

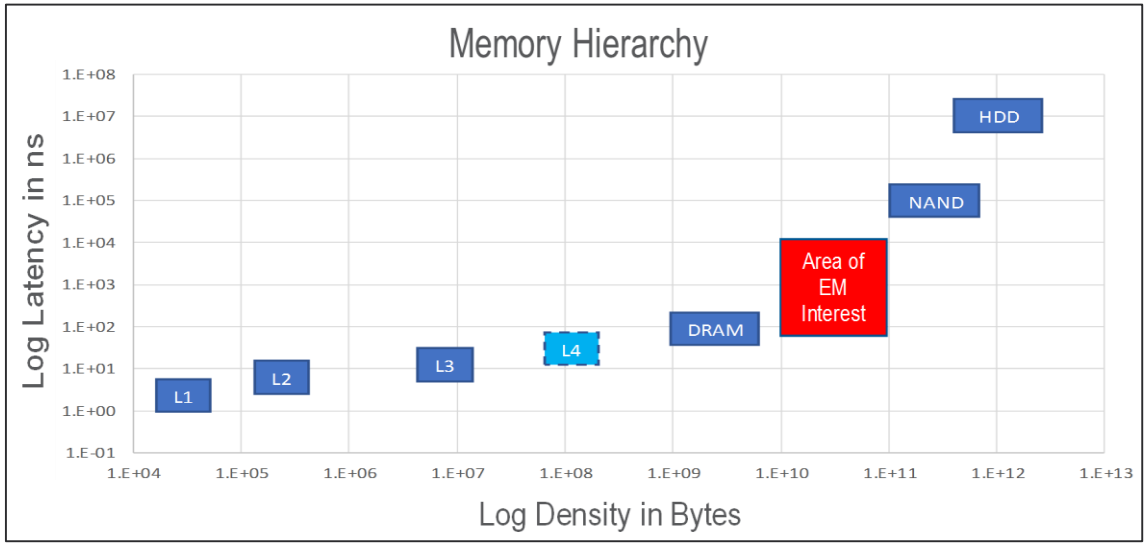
	DRAM	STTRAM	PCM/ 1T1R RRAM	Cross point RRAM	NAND
Read Latency	20ns	~ 50ns	~100ns-200ns	~100ns-200ns	~10us
Write Latency	20ns	~ 50ns	~1us	~1us	~10us
Read Endurance	>1e15	$>10^{11}$	$>10^{7}$	$>10^{7}$	$>10^{7}$
Write Endurance	>1e15	$>10^{11}$	$>10^{6}$	$>10^{6}$	2K-100K
Write/Read Energy/bit	<10pJ/bit	~25pJ/bit	~100-200 pJ/bit	~100-200 pJ/bit	> 100pJ/bit

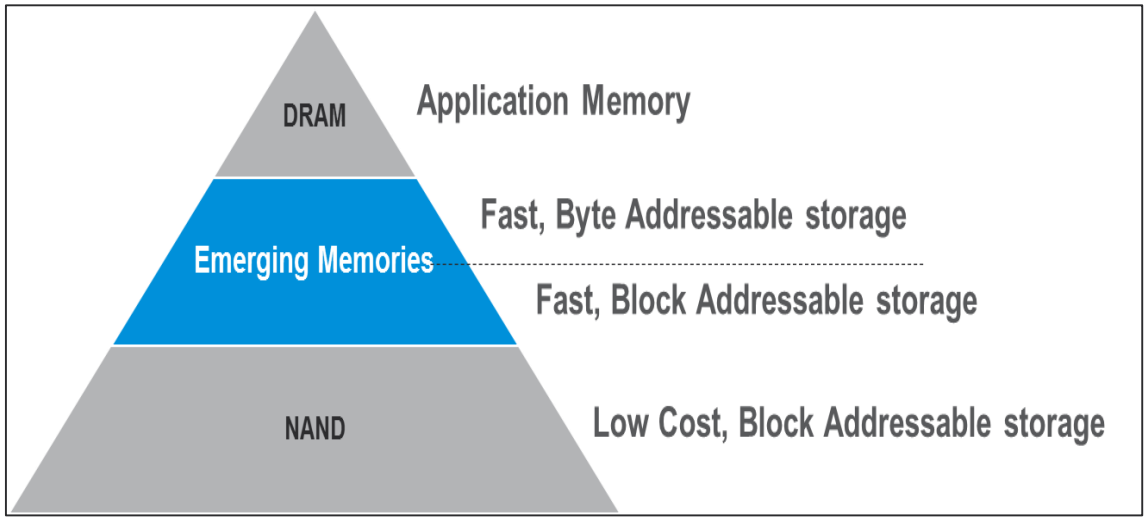
Emerging Memories

Lower Latency

Higher Density

Emerging Memories can fill data access latency gaps and enable new storage models

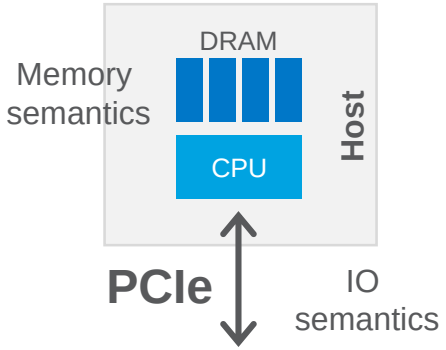




Attaching EM requires server architecture changes

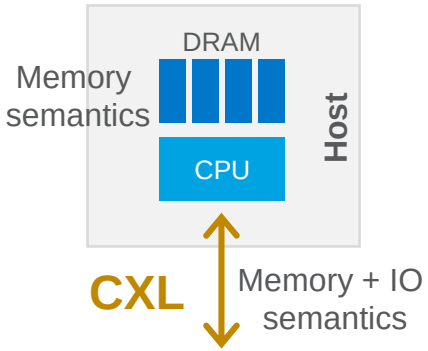
CXL is the emerging Standard for EM attach

Today



PCIe Gen4
IO devices
SSD
Accelerators

~2022

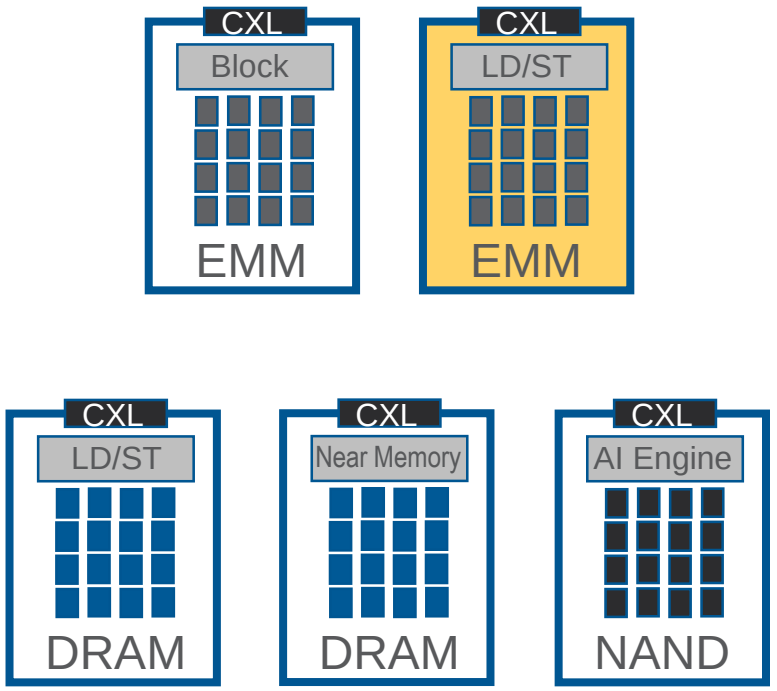


PCIe Gen5
2x IO Bandwidth
IO devices
SSD
Emerging Memory
Coherent Accelerators

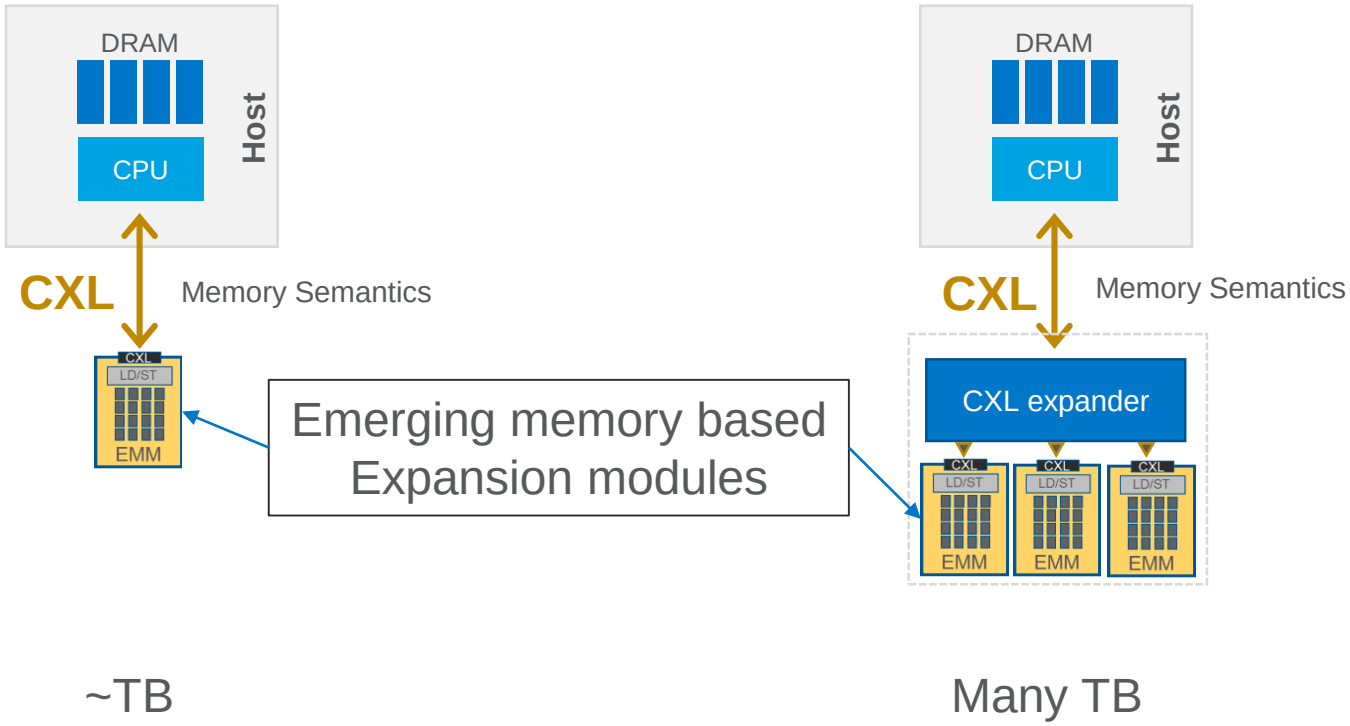
* For simplicity sake single CPU Hosts are shown

CXL enables innovations ranging from different memory types to heterogeneous computing

Some possible memory examples

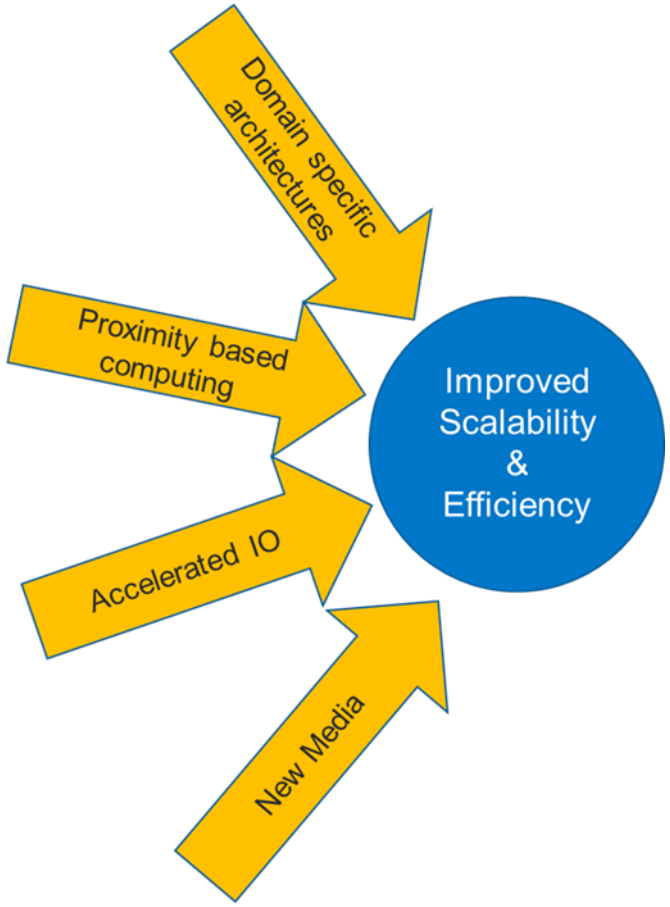


Scaling Emerging Memory capacity is critical to address use case requirements

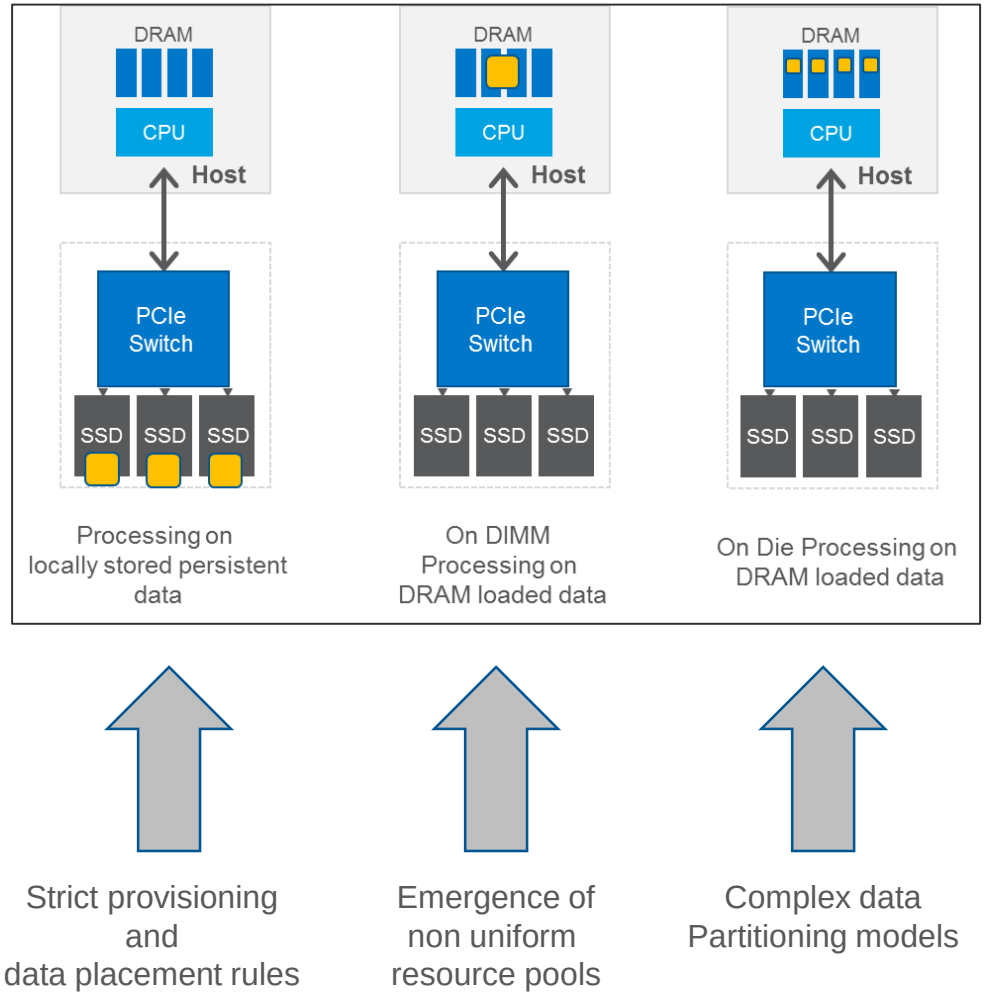


* For simplicity, single CPU Hosts are shown

The collection of these new technologies and architectures will impact how we build future Data Centers and Systems

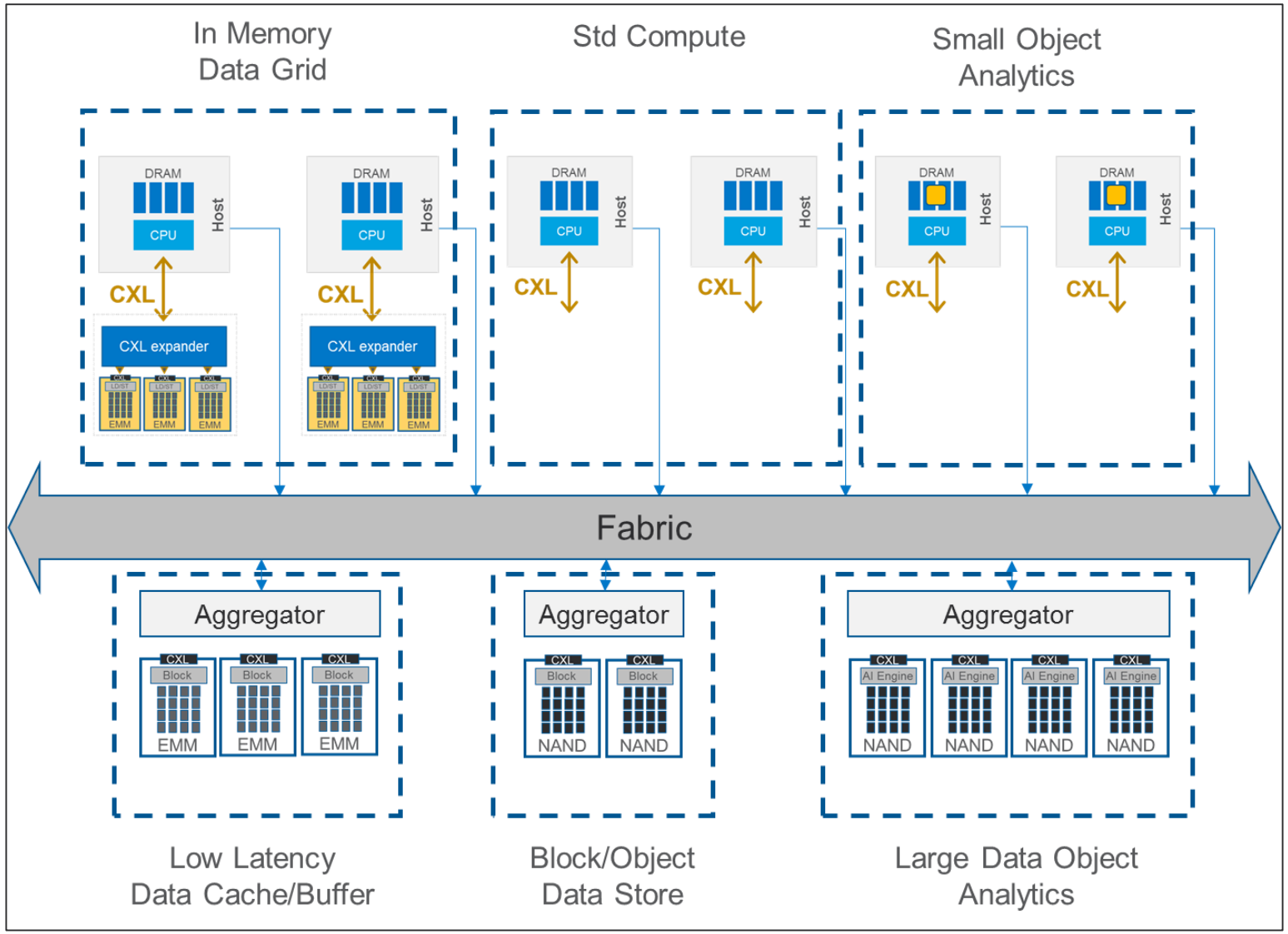


Co locating Compute and Data requires change in data placement, provisioning and load balancing strategies



* For simplicity single socket CPU Hosts are shown

Heterogeneous Resource Pools Example



* For simplicity single socket CPU Hosts are shown

Summary

Exciting times.

The Data Growth and the need for Faster Insight drives transitioning from decades old architectures to a new, emerging model utilizing breakthrough technologies

Buckle your seatbelt!

Thank you!

bfleischer@micron.com