Persistent Memory for Artificial Intelligence

Bill Gervasi

Principal Systems Architect

bilge@Nantero.com

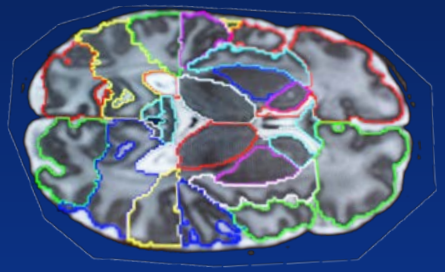
Agenda

- Memory for AI
- Challenges for Data Integrity
- NVRAM and AI
- New System Architectures
- Artificial Intelligence Variations

Flash Memory Summit

Every day…

…new applications for AI emerge…



…from speech recognition to medical analysis to self driving vehicles to purchasing tendencies to fraud detection to…

AI methods are getting smarter, too

Stage 7 – Singularity and Transcendence

Stage 6 – Artificial SuperIntelligence

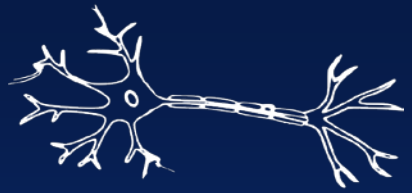
Stage 5 – Self Aware Systems / Artificial General Intelligence

Stage 2 – Context Awareness and Retention

Stage 1 – Rule Based Systems

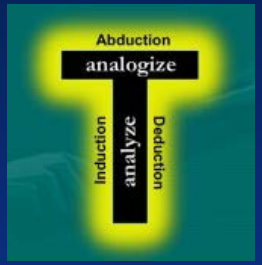
We're about mid-way into stage 4

*Em360tech.com



Neural

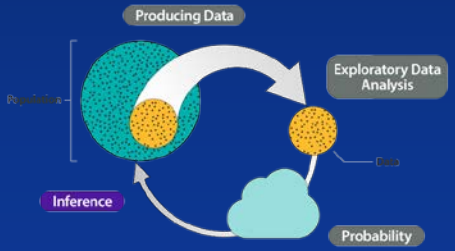
Fit



Symbolic

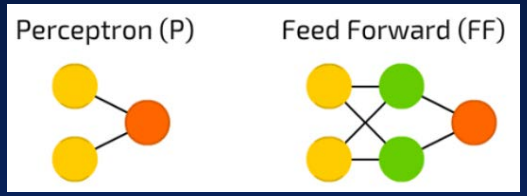
A wide variety of approaches

Analogy

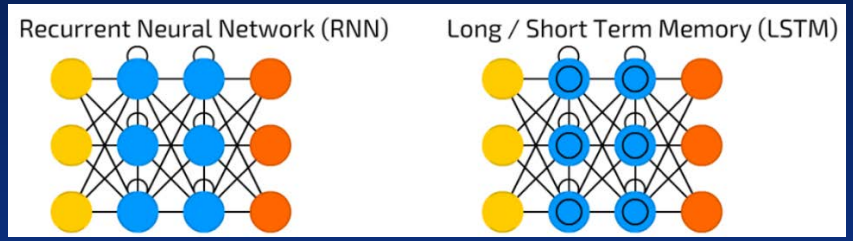


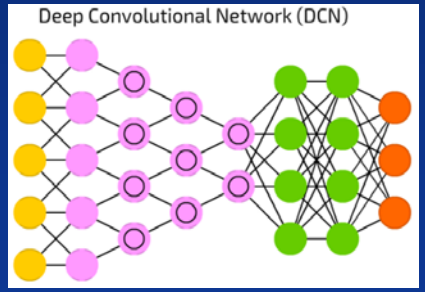
Inference

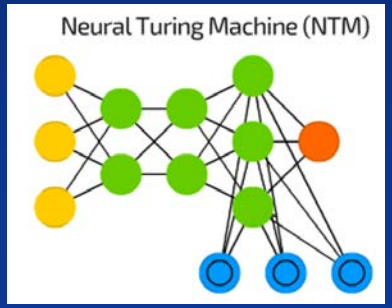
*Mueller/Massaron

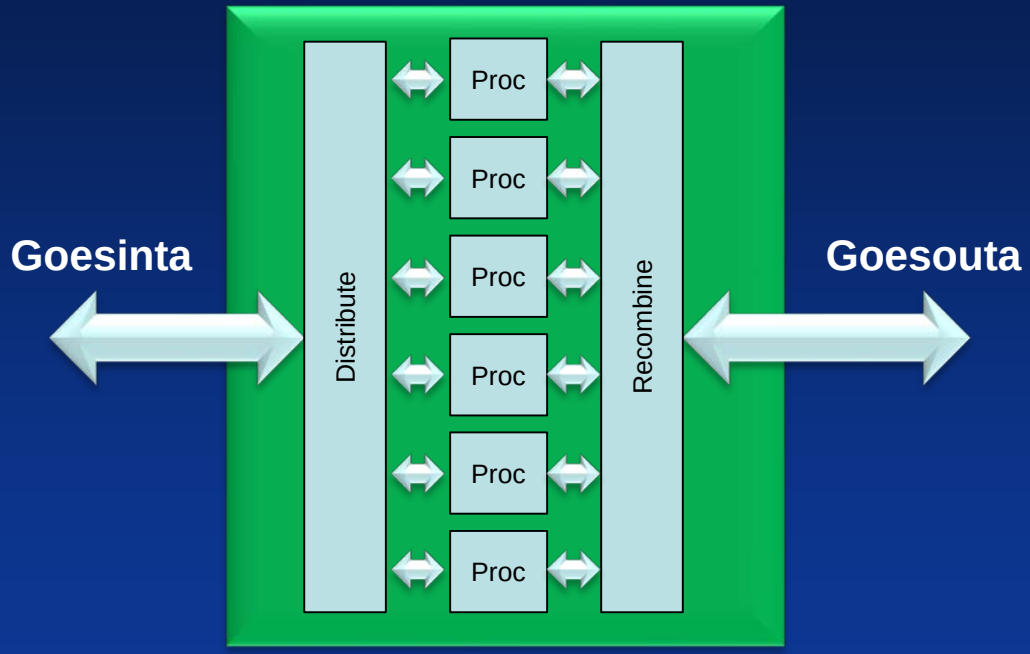


Each generation adds a new complexity in self modification









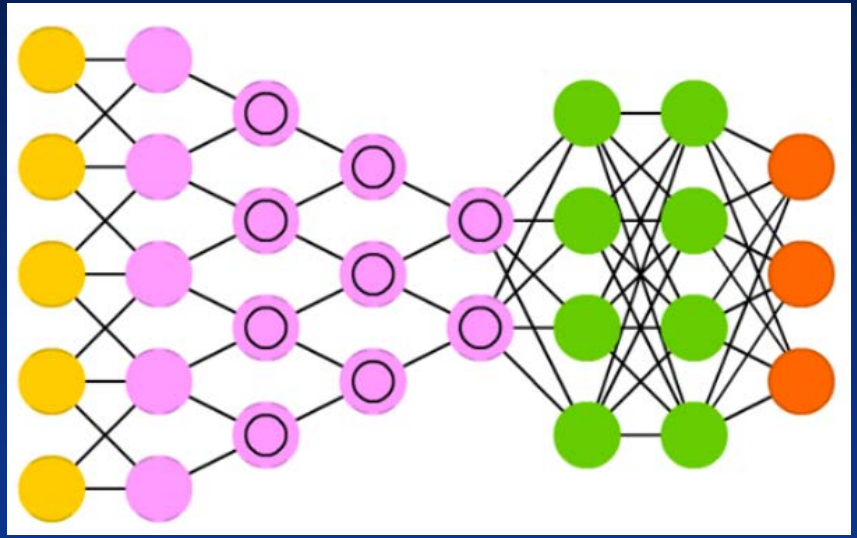
AI Accelerator Characteristics
SIMD (single instruction multiple data) rules

Wide array of simple processing elements

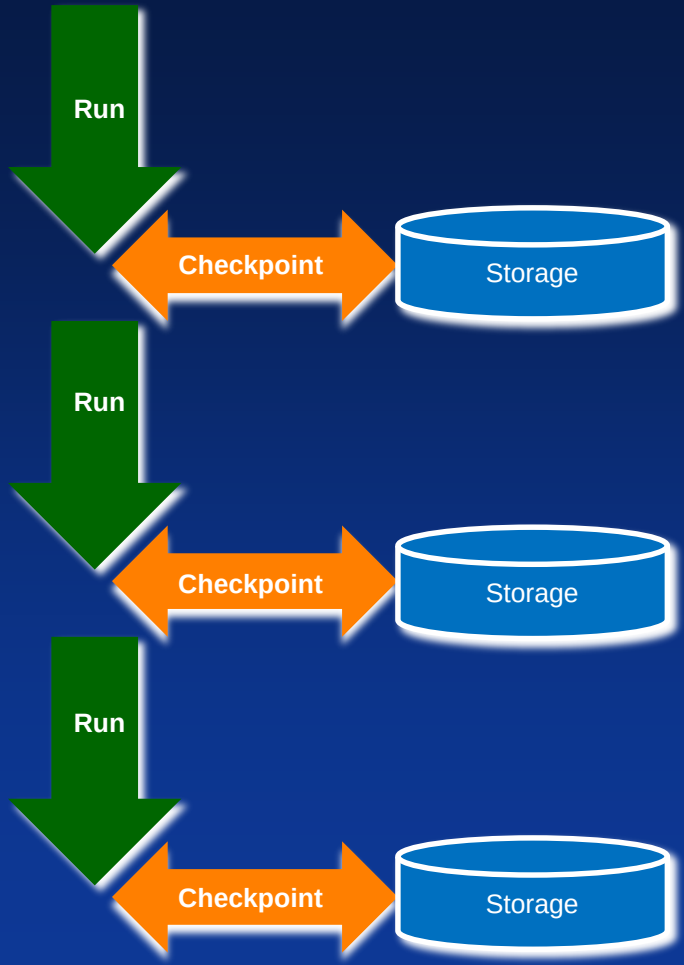
Reduced floating point precision

Tuned for matrix operations

All that trapped data

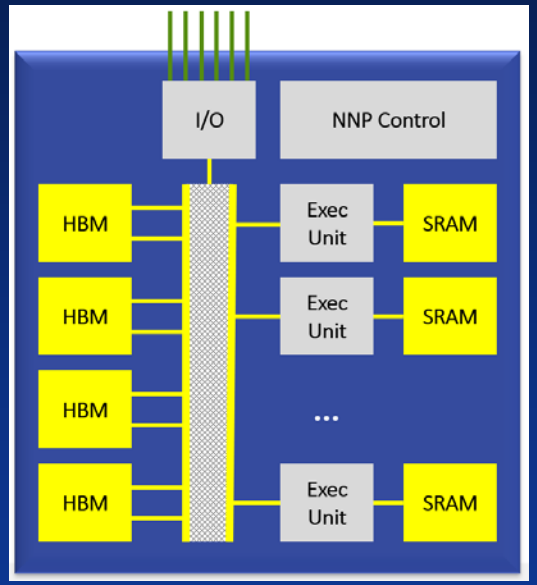


To avoid losing learned data, systems must periodically checkpoint



Ah, checkpointing!

Great way to burn lots of power and waste performance for no good reason



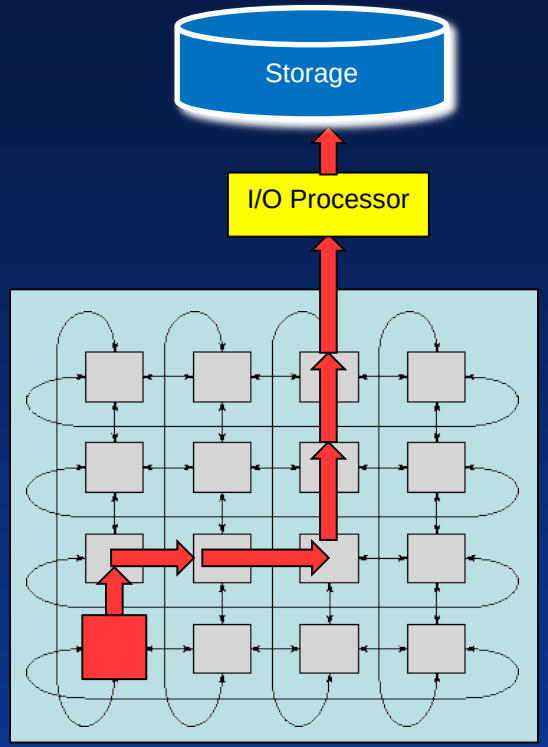
So now let's take our cute little AI engine and expand it into a mesh

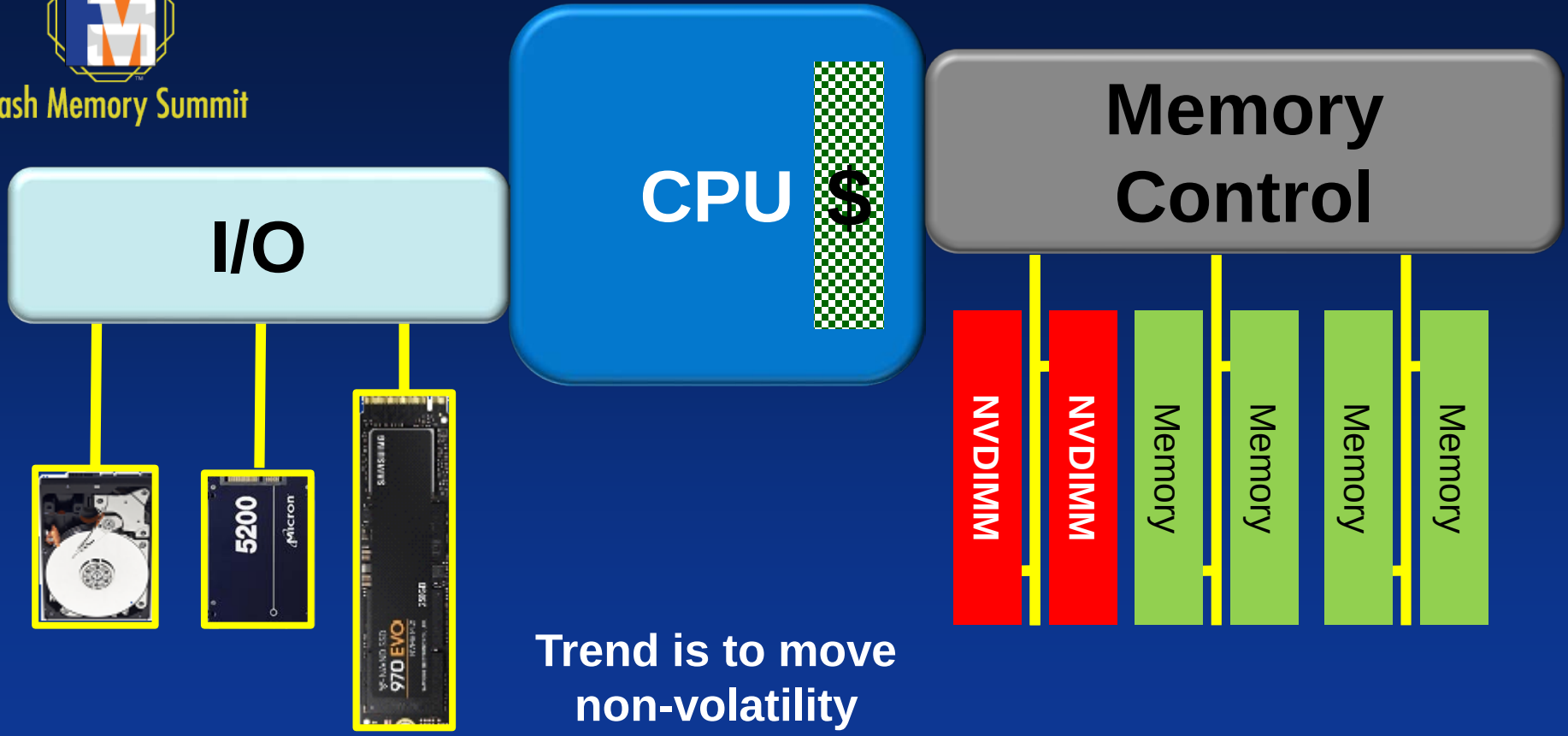
Oh, look, some data I've learned I don't want to lose!

Let's save it!

See how much fun it is to checkpoint?

Wasn't that great?





Trend is to move non-volatility closer to the CPU

Fortunately, a new wave of technology is coming

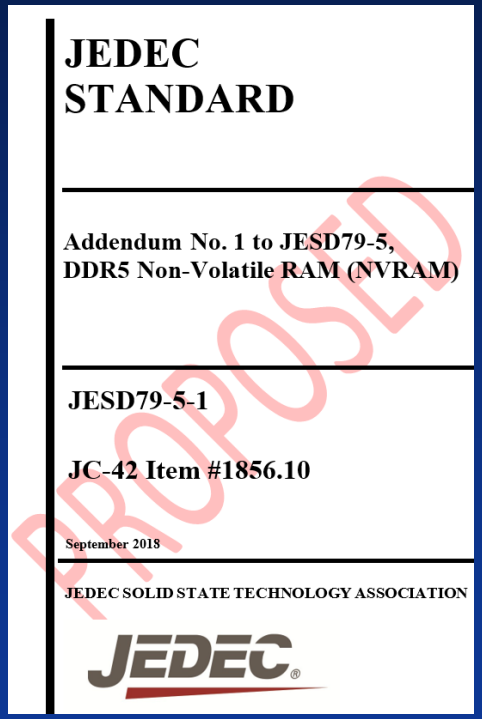
Speed of a DRAM

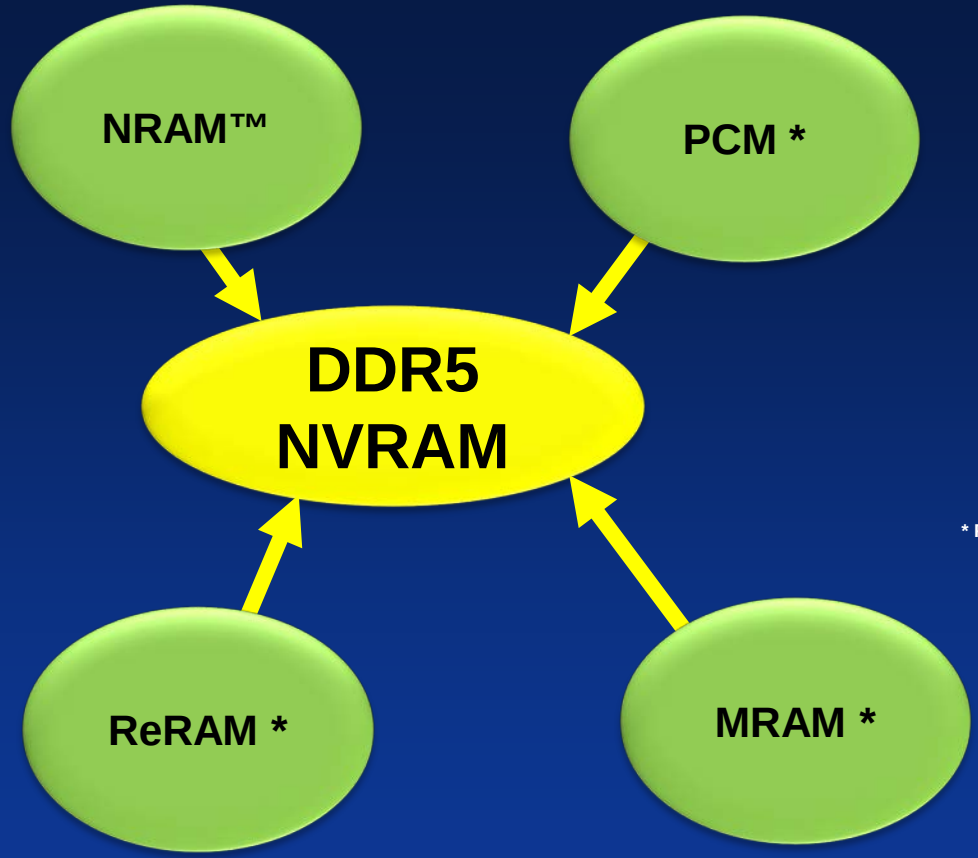
Power-off data persistence

Unlimited write endurance

Memory Class Storage

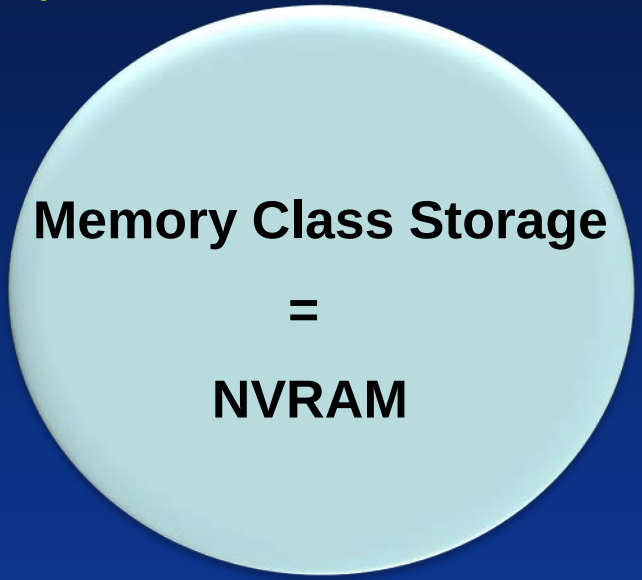
PERSISTENT DRAM DROP-IN REPLACEMENT





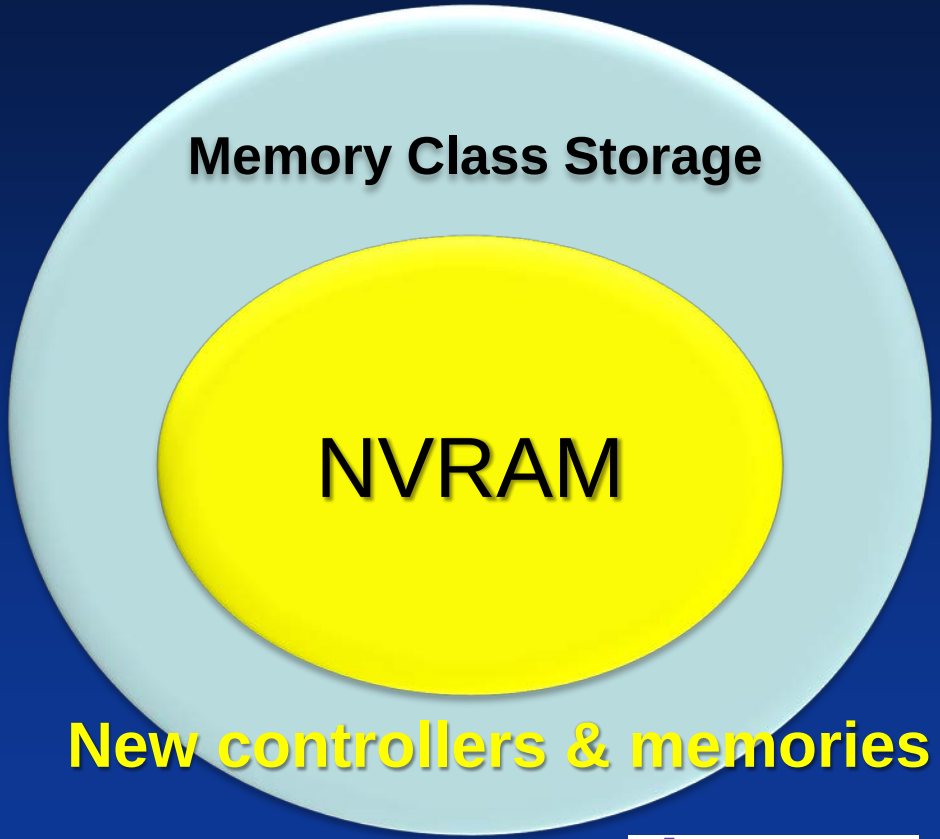
* Future generation devices

Details in my talk on Thursday 8:30



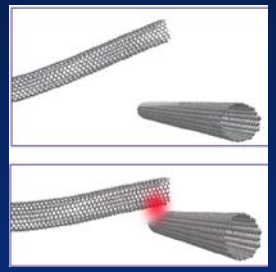
For now…

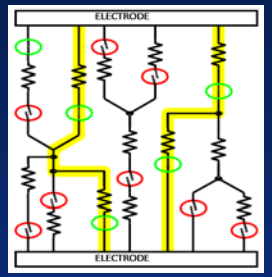
In the future

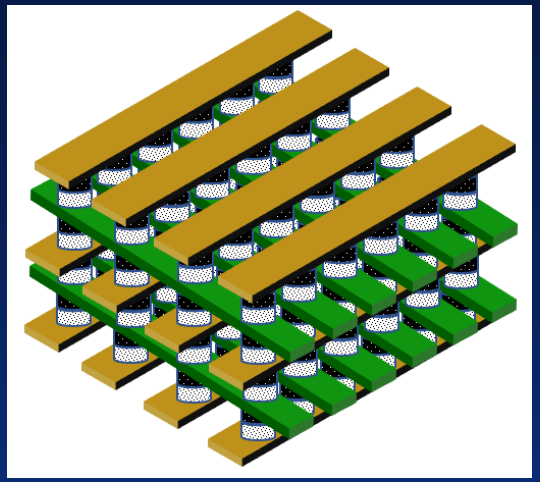


New controllers & memories

Flash Memory Summit

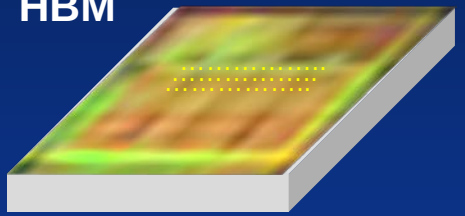




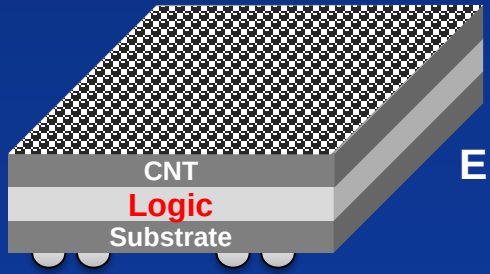


DDR4
DDR5

HBM

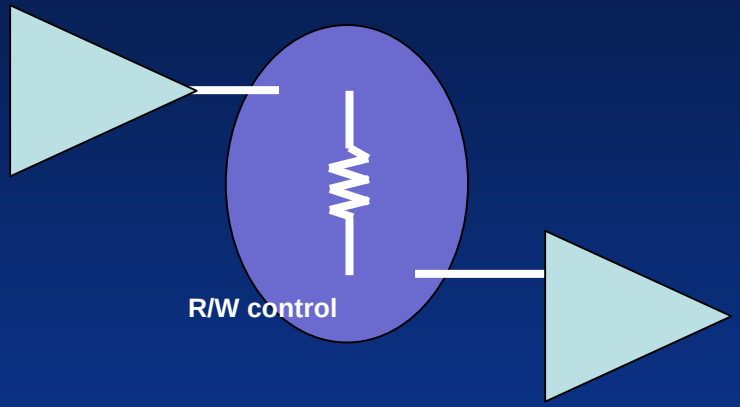


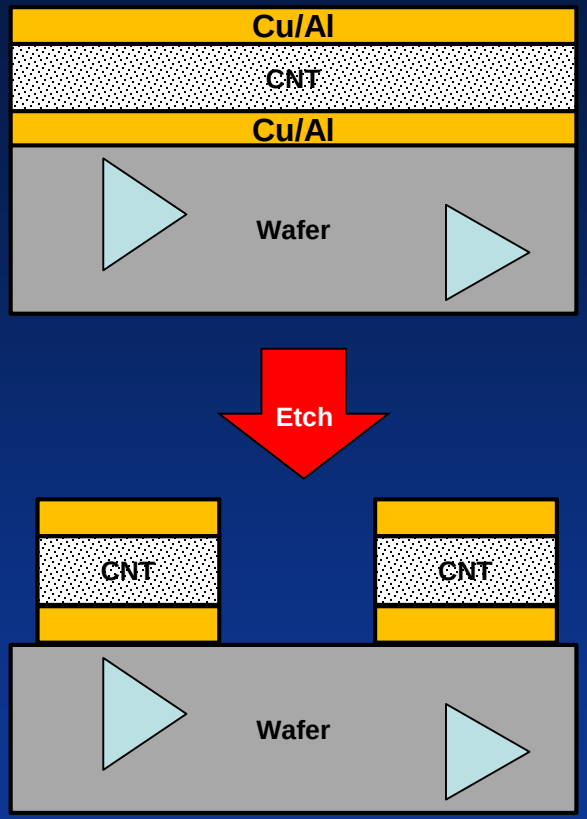
Nantero NRAM is a persistent memory using carbon nanotubes to build resistive arrays which can be arranged in a DRAM compatible device or deposited directly on circuits



Embedded

NANTERO

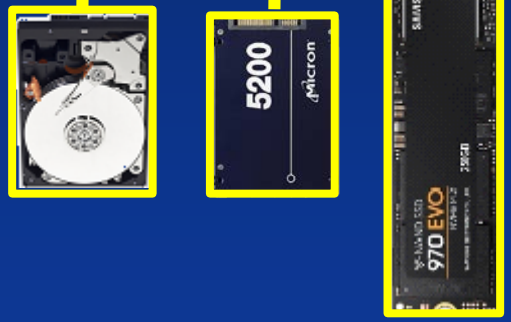




NRAM may be applied on top of driver circuits, integrated into the logic

I/O

Storage-less systems enabled



CPU $

Memory Control

Memory Class Storage

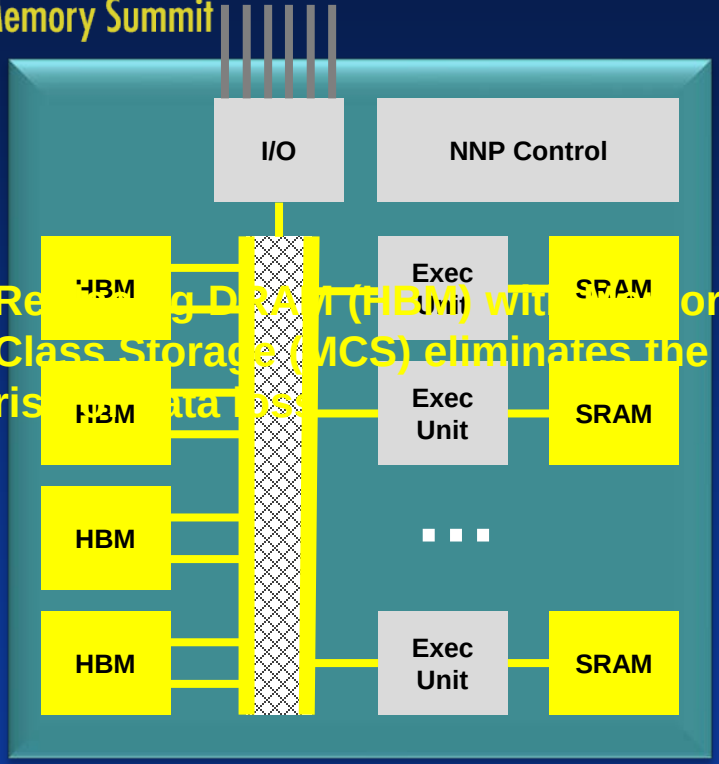
Memory Class Storage

Memory Class Storage

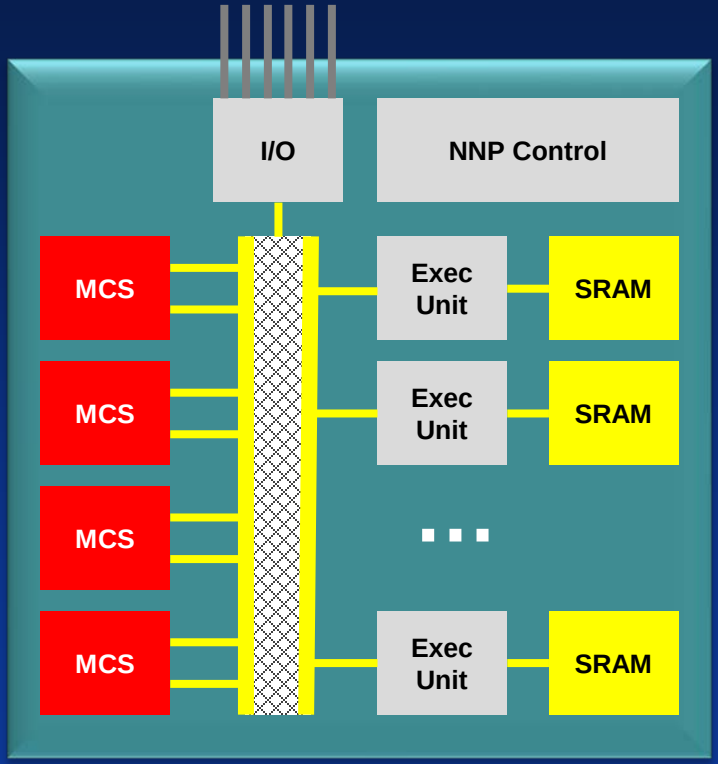
Memory Class Storage

Memory Class Storage

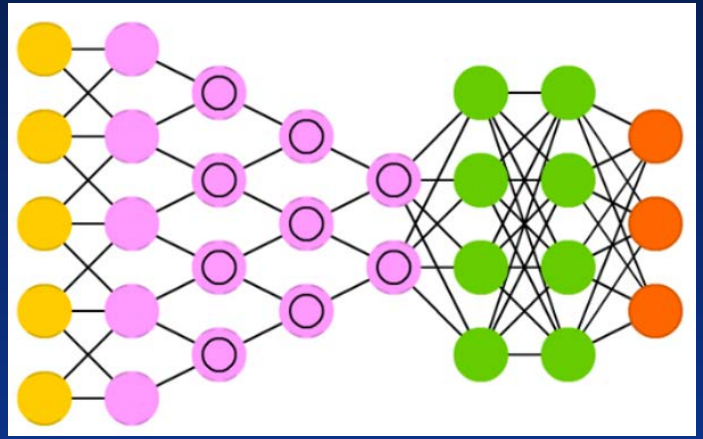
Memory Class Storage



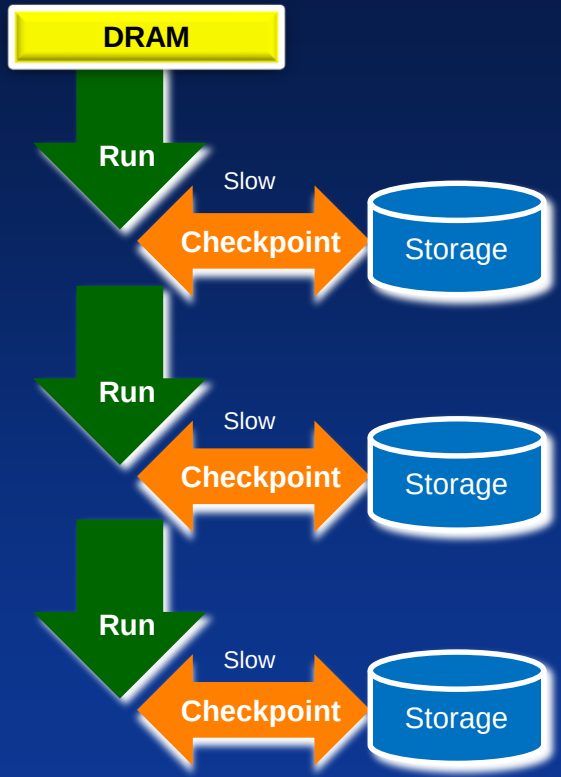
Re[illegible]g DRAM (HBM) wit[illegible]ory Class Storage (MCS) eliminates the ris[illegible]ata loss

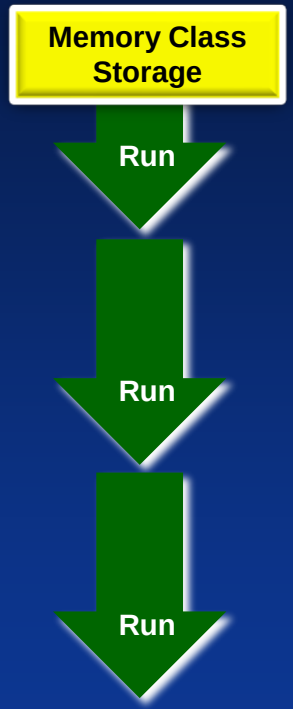


All that trapped data…

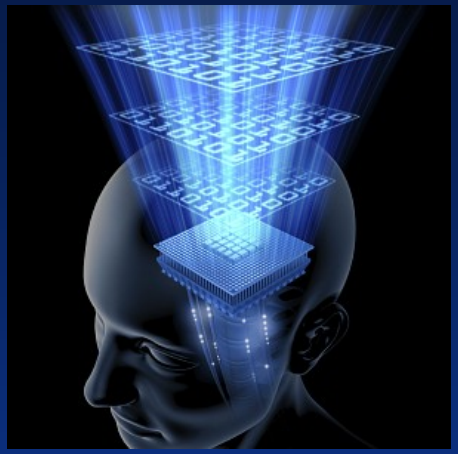


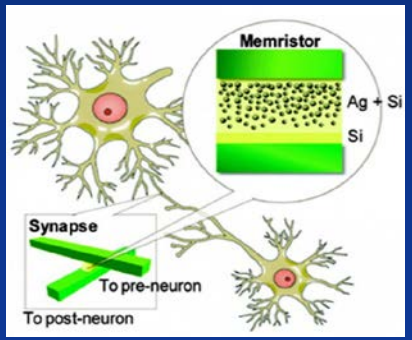
…can just stay there!


DRAM
Run
Slow
Checkpoint
Storage
Run
Slow
Checkpoint
Storage
Run
Slow
Checkpoint
Storage

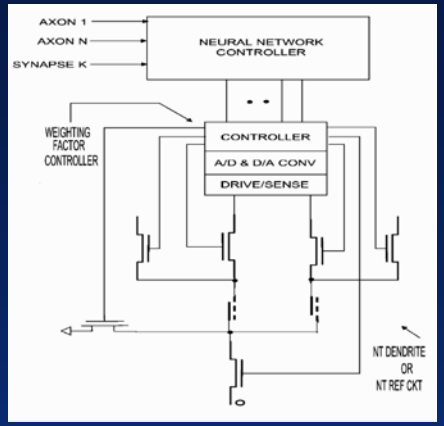

Memory Class Storage
Run
Run
Run

There is a HUGE gap between AI research work and what's being built

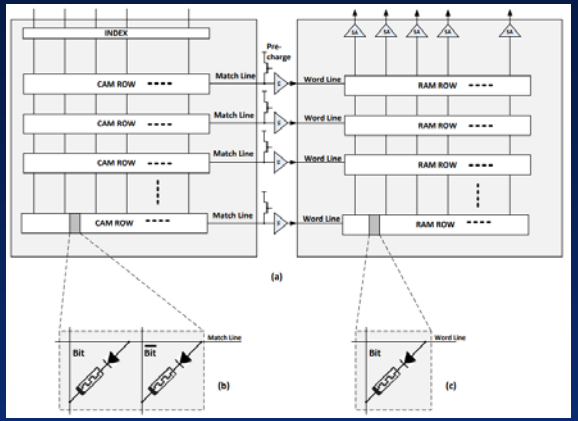


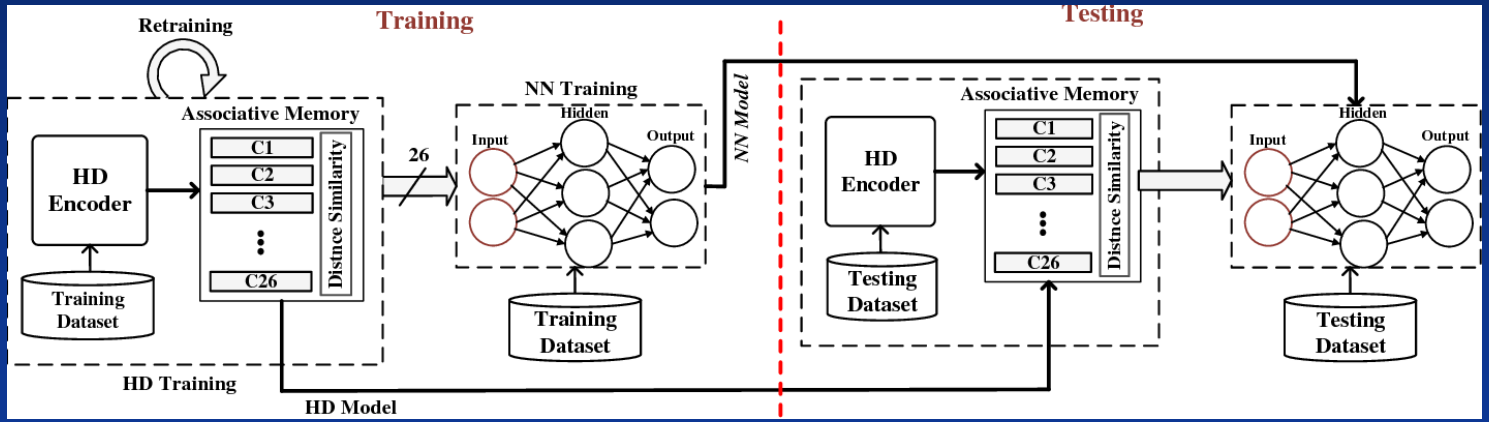


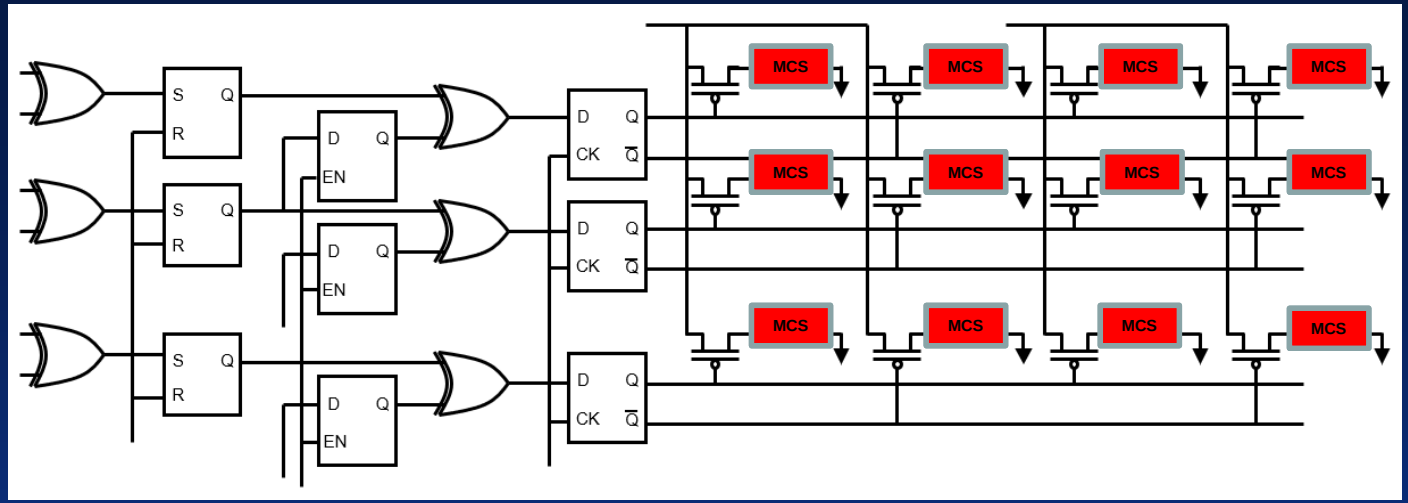
Modeling the brain leads to embedded memory



AI embedded memories are often not von Neumann





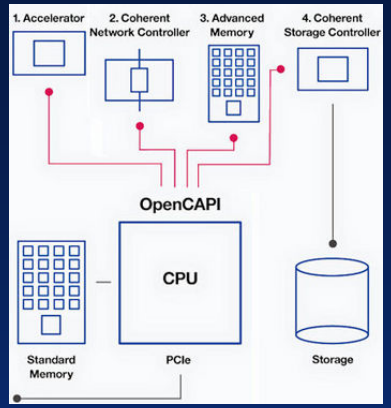


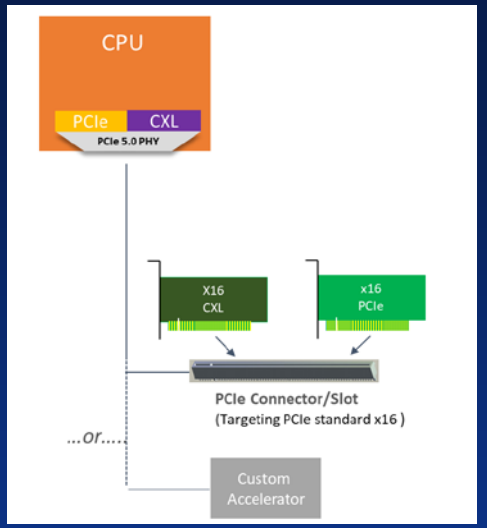
Embedded memories suffer from read/write time issues when used as random access (including checkpointing)

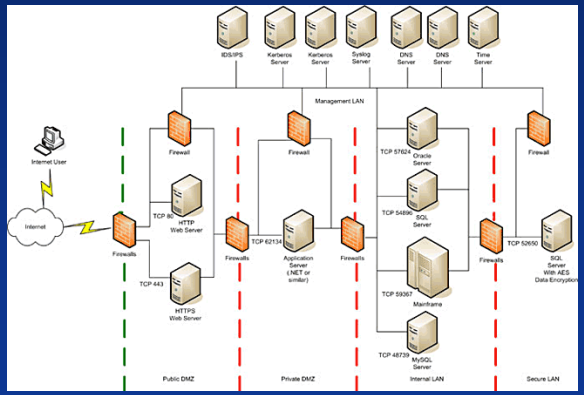
Replacing embedded memory with Memory Class Storage removes this barrier

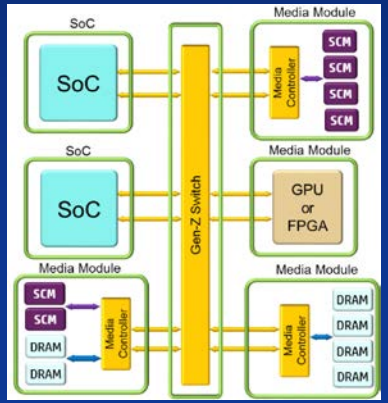
Flash Memory Summit

Computing networks are evolving to fabrics



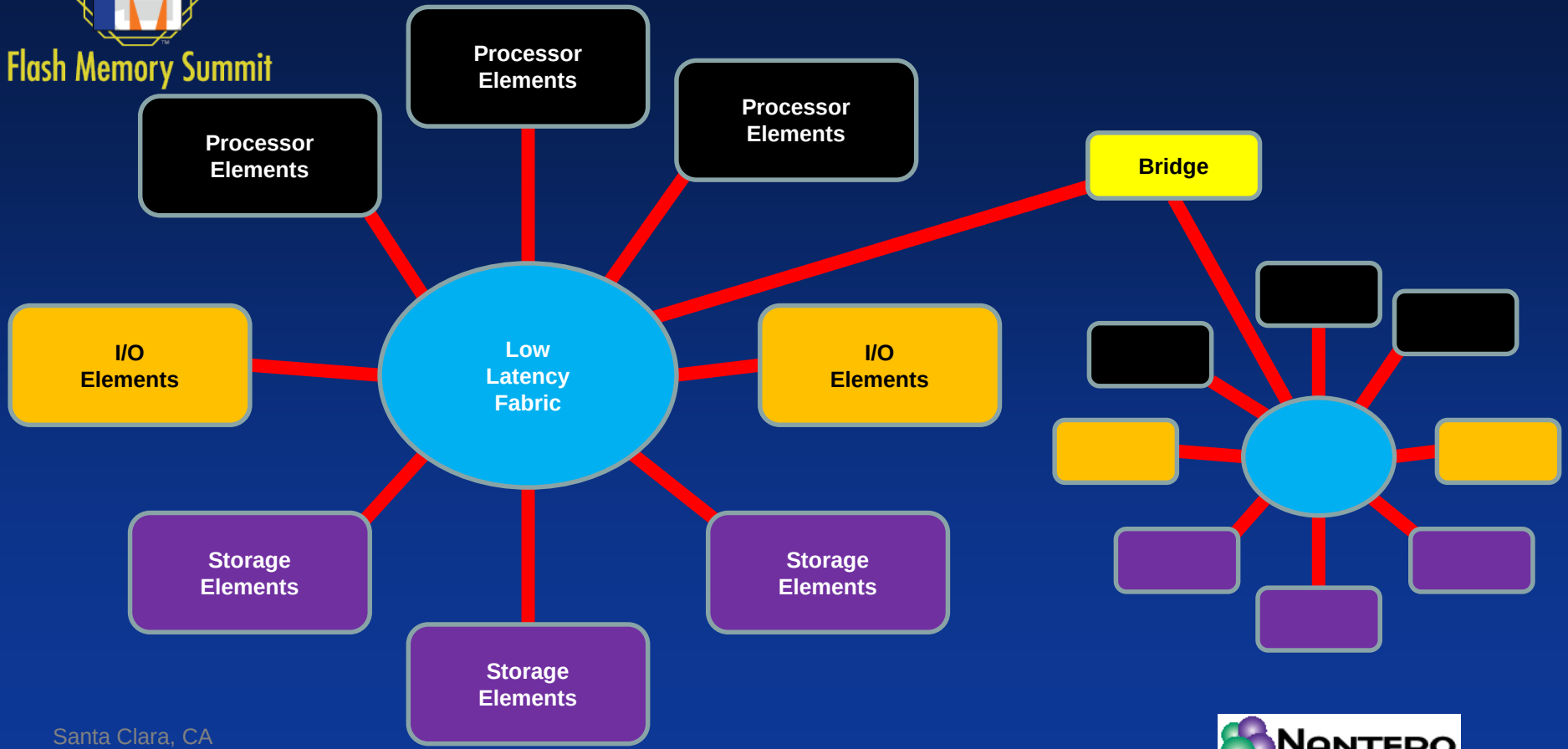


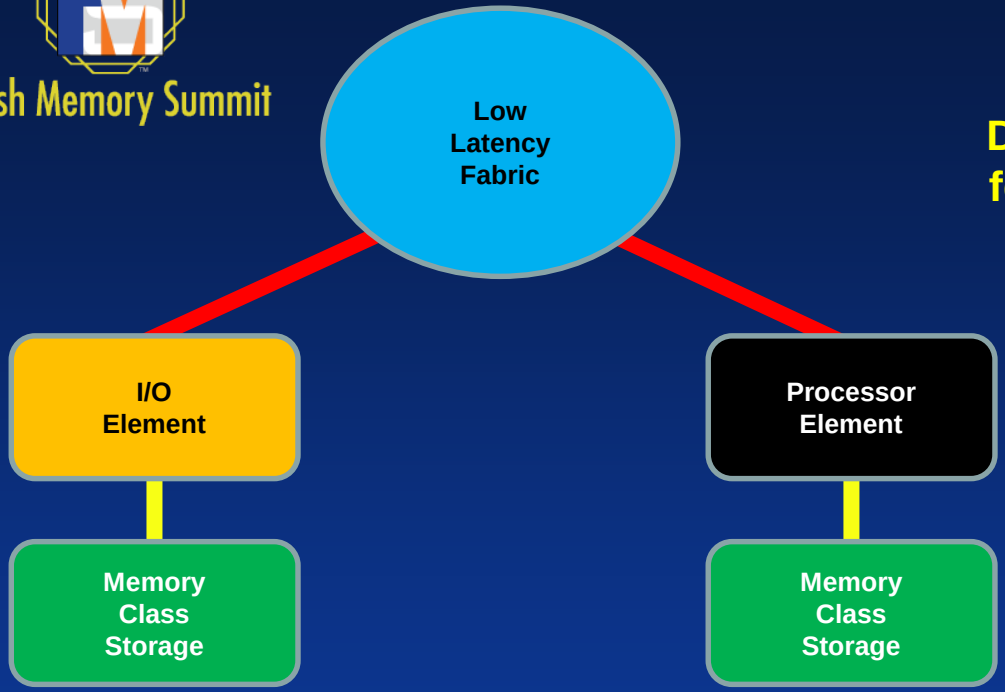




Heterogeneous processing in a unified space

NANTERO


Processor Elements
Processor Elements
Processor Elements
Bridge
I/O Elements
Low Latency Fabric
I/O Elements
Storage Elements
Storage Elements
Storage Elements



DDR DRAM protocols do nothing for these architectures but get in the way

Likely place for new Memory Class Storage interface protocols to emerge

Flash Memory Summit

Number of Devices Needed to Achieve Gen-Z Throughput

Gen-Z			DDR4-3200 x8	DDR5-6400 x8	LPDDR5-6400 x64	GDDR6-18000 x32	HBM3-4800 x1024
GT/s	Lanes	GB/s	3.2 GB/s	6.4 GB/s	51.2 GB/s	72 GB/s	615 GB/s
25	64	320	100x	50x	7x	5x	1x
25	128	640	200x	100x	13x	9x	2x
32	64	400	125x	63x	8x	6x	1x
32	128	800	250x	125x	16x	12x	2x
56	64	700	219x	110x	14x	10x	2x
56	128	1400	219x	110x	28x	20x	3x
112	64	1400	438x	220x	28x	20x	3x
112	128	2800	875x	438x	55x	39x	5x

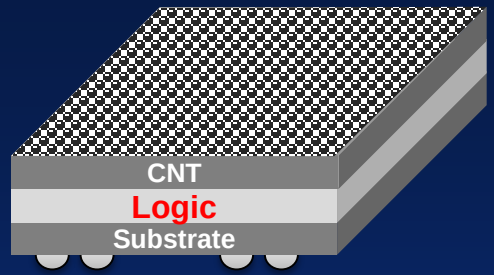
NANTERO

Problems with Standalone Memories

- Number of Devices Needed
- High System Power
- Little Control Over # GB
- Lots of Wasted Data Access

Gen-Z

GT/s	Lanes	GB/s	NRAM-6400 x4096 3200 GB/s*
25	64	320	1x
25	128	640	1x
32	64	400	1x
32	128	800	1x
56	64	700	1x
56	128	1400	1x
112	64	1400	1x
112	128	2800	1x



Single chip fabric storage possible

In-Memory computing enabled

Significantly lower power

Access granularity up to the controller:
Avoid unused data accesses

*No refresh: 15% higher throughput than DRAM

Flash Memory Summit

Embedded Persistent Memory changes how you view AI architecture

Embedded Persistent Memory enables new implementations

Common Memory Pools Used

AI Uses SIMD Data Flow

Data Checkpointing is Horrid for AI

Agenda

New Applications for Fabric Based Systems Coming

NVRAM Solves Many AI Memory Issues

Questions?

Bill Gervasi
Principal Systems Architect
bilge@Nantero.com