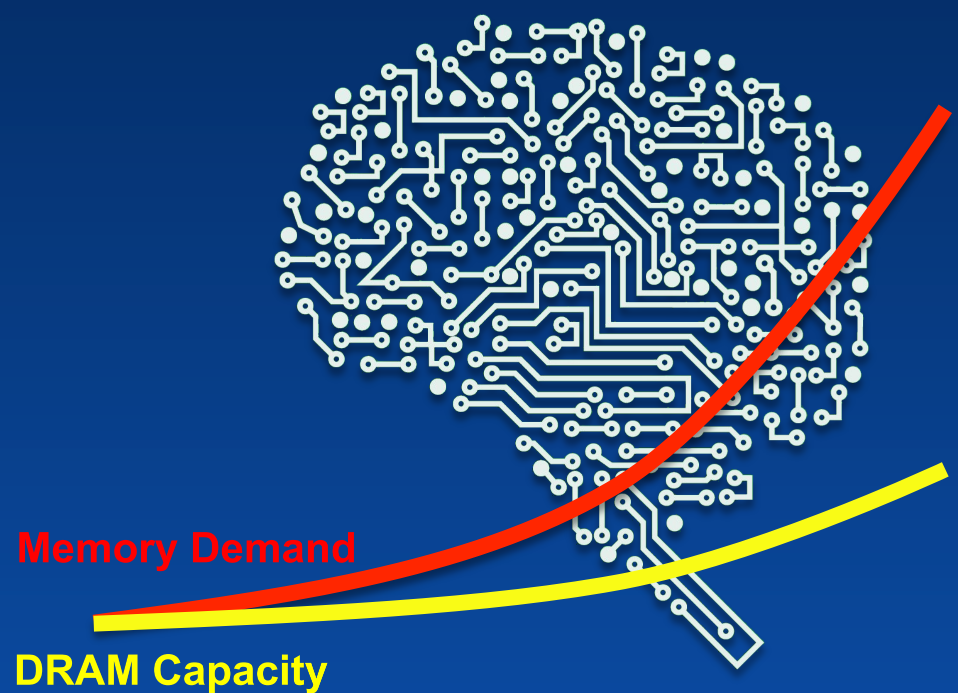
Persistent Memory for Artificial Intelligence

Bill Gervasi

Principal Systems Architect

bilge@Nantero.com

Demand Outpacing Capacity



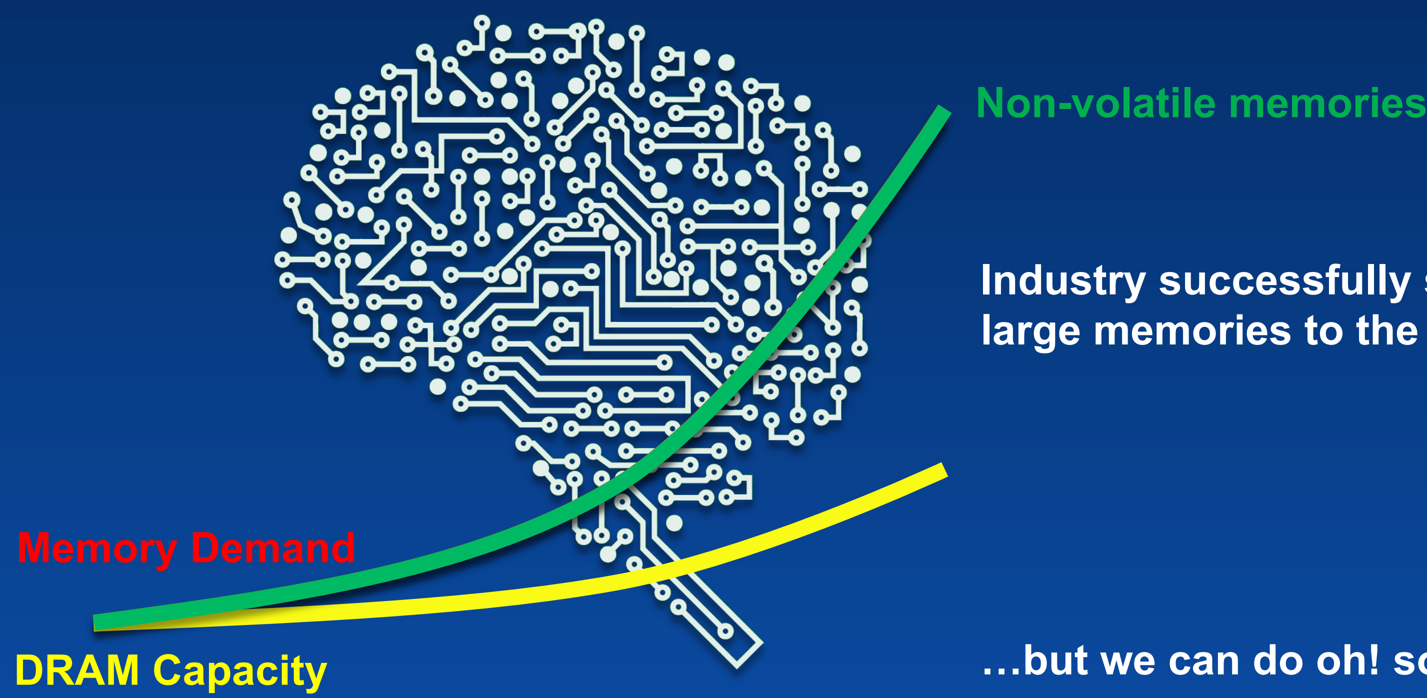
In-Memory Computing

Artificial Intelligence

Machine Learning

Deep Learning

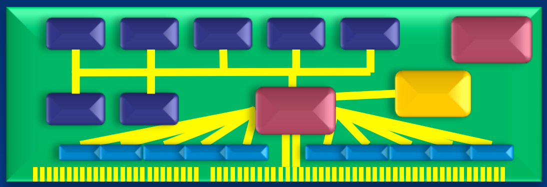
Driving New Capacity Models



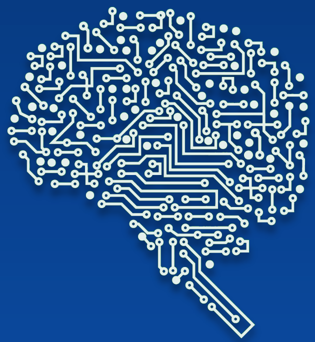
Industry successfully snuggling large memories to the processors…

…but we can do oh! so much more

My Three Talks at FMS



NVDIMM Analysis



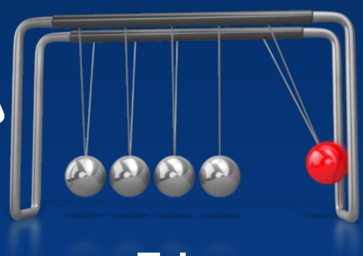
Artificial Intelligence

Memory Class Storage

History of Architectures

Historical Trends in Computing

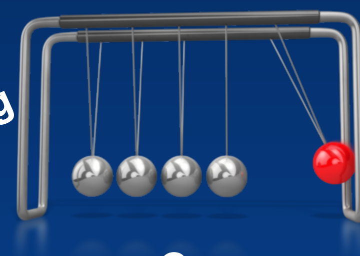
Central Computing



Client Computing

Edge Computing

Central Processing



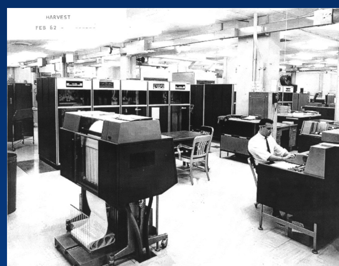
Distributed Processing

Co-Processing

Power Failure
Data Loss

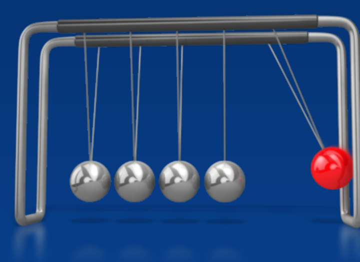
Flash Memory Summit

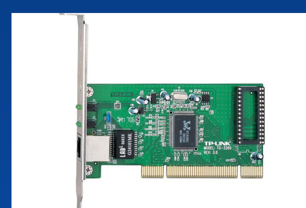
Some Moments in History



Central Processing

Shared Processor
Dumb terminals





Peer-to-peer networks

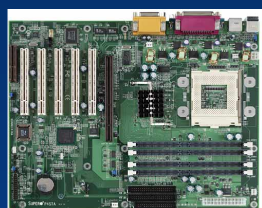
Distributed Processing

Processor per user

Flash Memory Summit

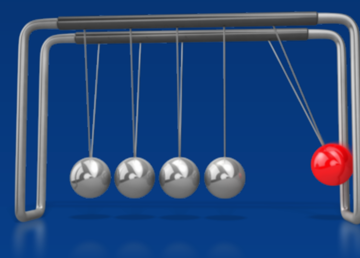
NANTERO

Some Moments in History



Central Processing

“Native Signal Processing”
Main CPU drivers
Cheap analog I/O



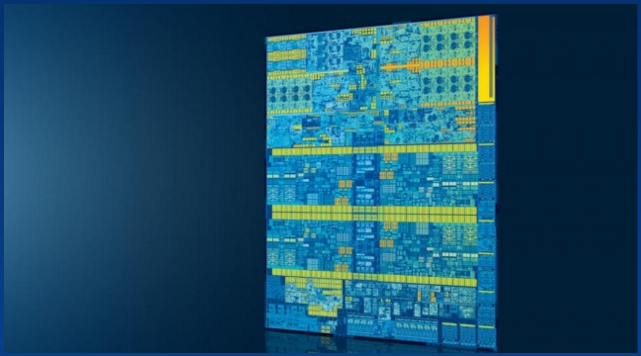
Tightly-coupled coprocessing

Distributed Processing

Hercules graphics
Sound Blaster audio
Rockwell modem
Ethernet DSP

The Lone Survivor…

Graphics add-in cards



Integrated graphics

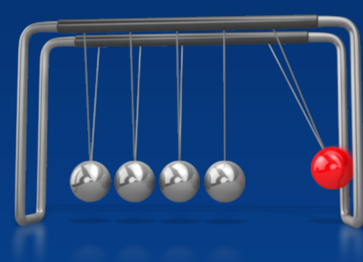
…survived the NSP war

Flash Memory Summit

NANTERO

Some Moments in History

Central Processing



Distributed Processing

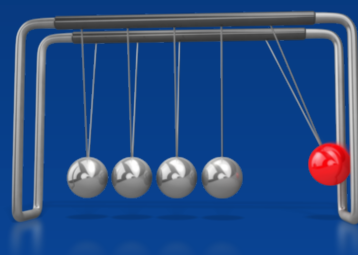
Phone providers controlled all data processing

Edge computing reduces latency

Phone apps provide local services

When the Playing Field Changes

The speed of networking directly impacts the pendulum swing from centralized to distributed

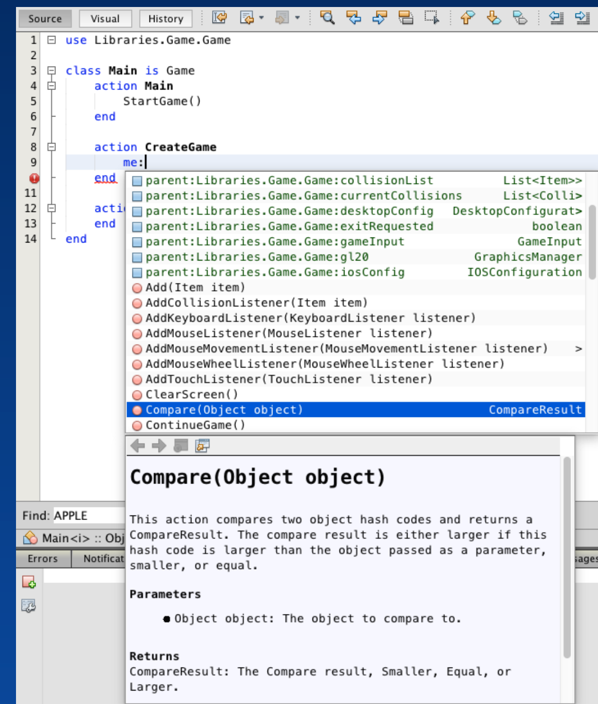


A faster network favors distributed computing

Flash Memory Summit

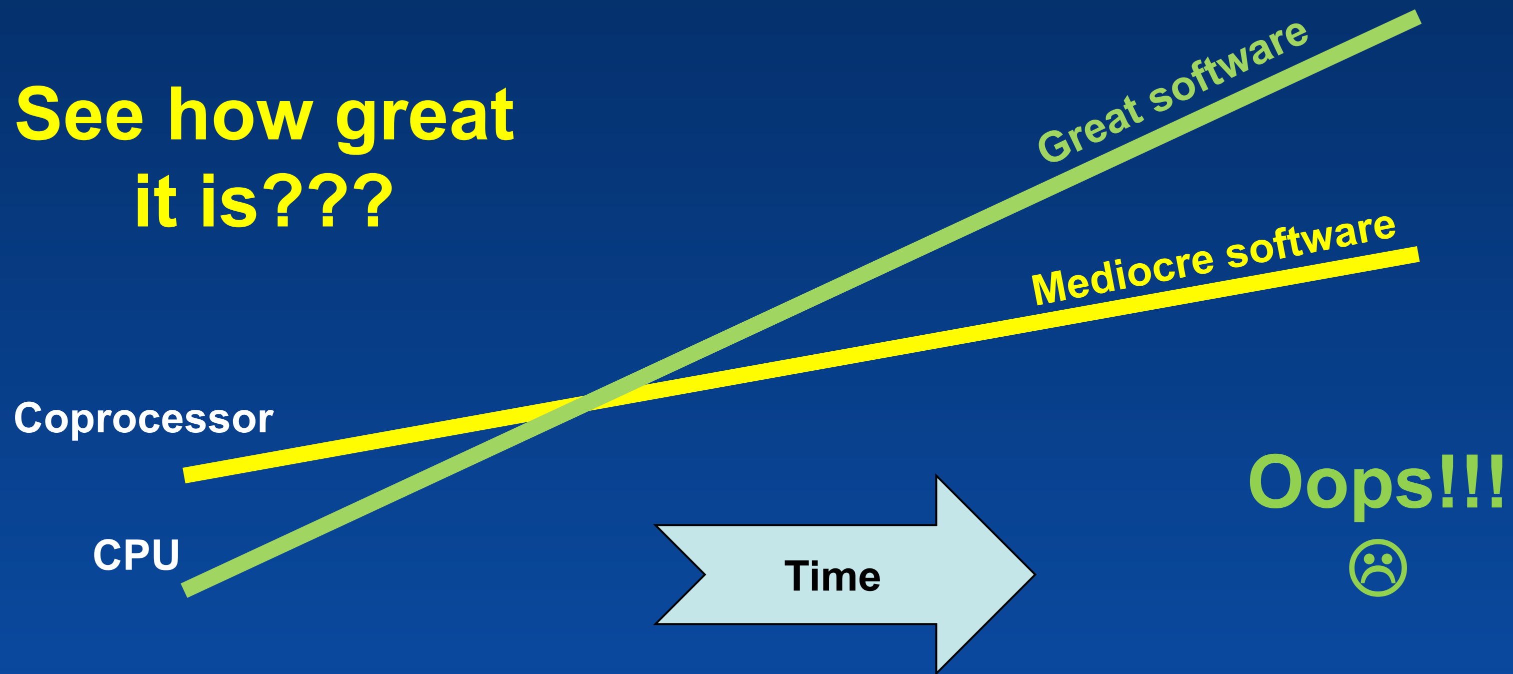
Winners and Losers

Often, the maturity of the software development environment determined who won and who lost



Maintaining an Edge

See how great it is???



The Tail Wagging the Dog

I won't say "It's the Software, Stupid"
because I know you're not stupid

however

To succeed, AI needs GREAT
software infrastructure

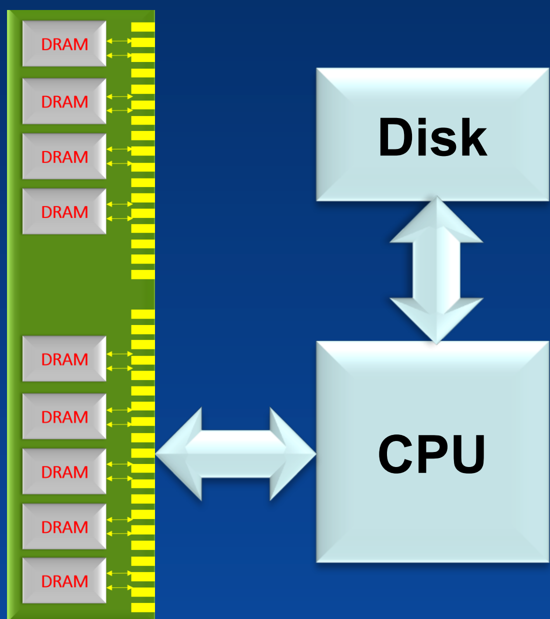
Driving some companies to design
hardware to the software
instead of software to the hardware

Flash Memory Summit

NANTERO

Wild Array of Programmer Options

AI on Traditional Server



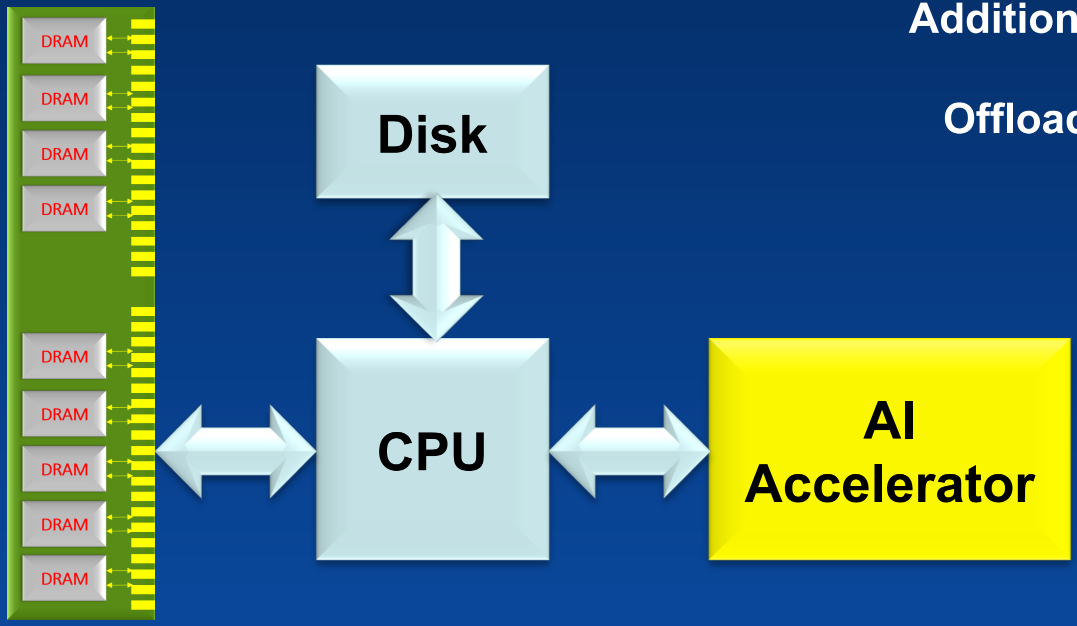
No magic

AI applications are like any other

Data processing done on main CPU

Downside is main CPU is overkill in floating point, and weak in parallelism

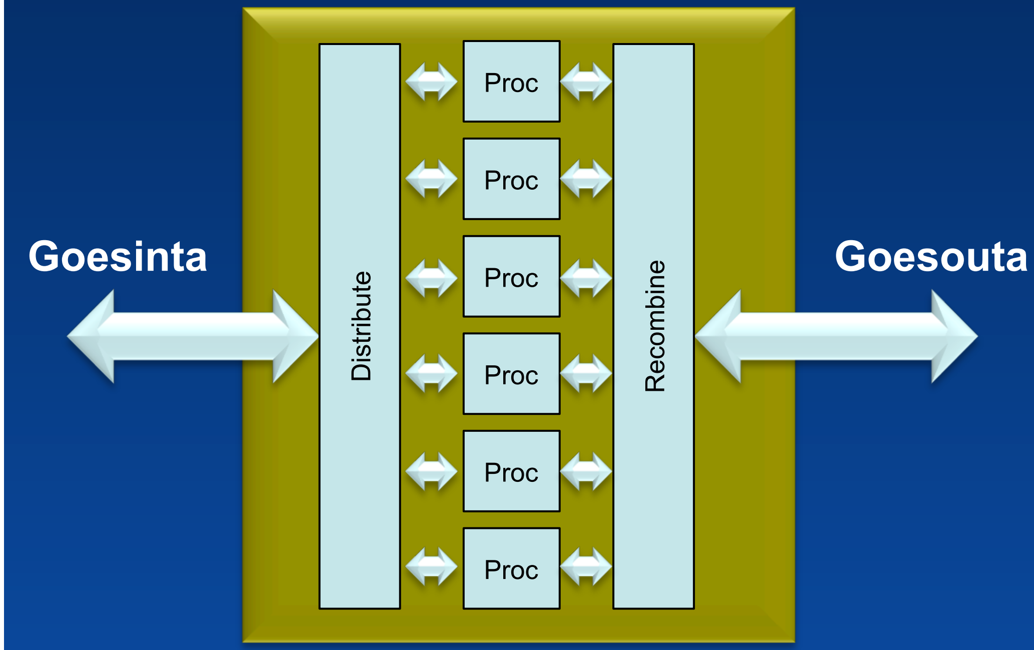
AI Evolution



Addition of AI Accelerator

Offloads main CPU for AI tasks

AI Evolution



AI Accelerator Characteristics

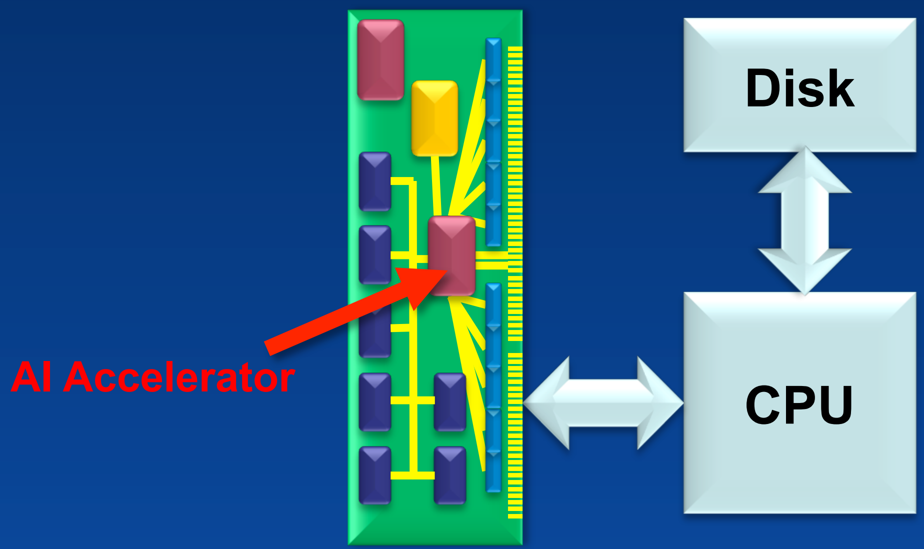
Wide array of simple processing elements

Reduced floating point precision

Tuned for matrix operations

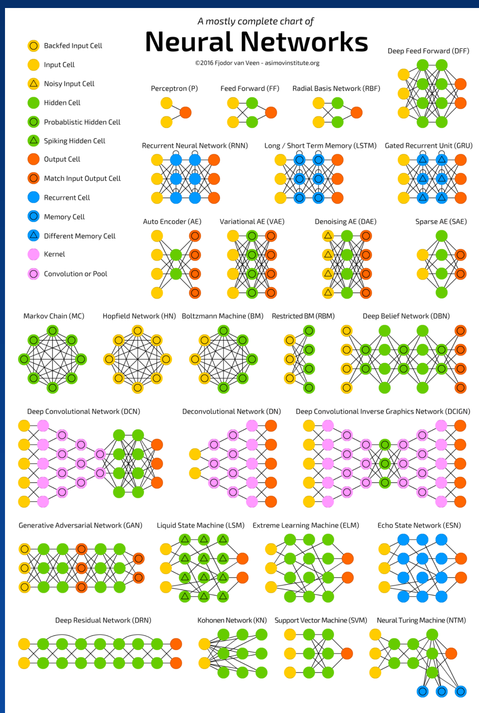
In-Memory Computing

In-memory computing lets the AI accelerator control the memory directly



Also great for encryption

Data Processing Paradigms



Traditional database

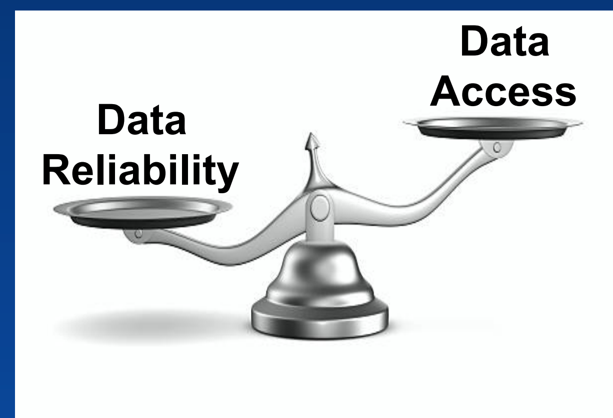
Data mining

Inferencing

Fuzzy logic

Recognition

etc

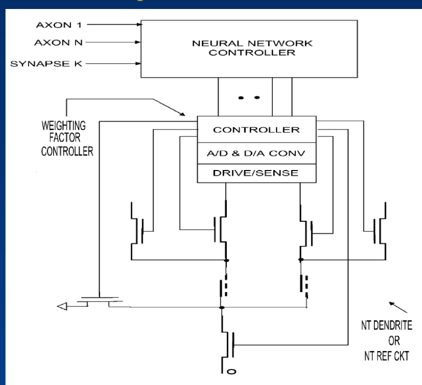


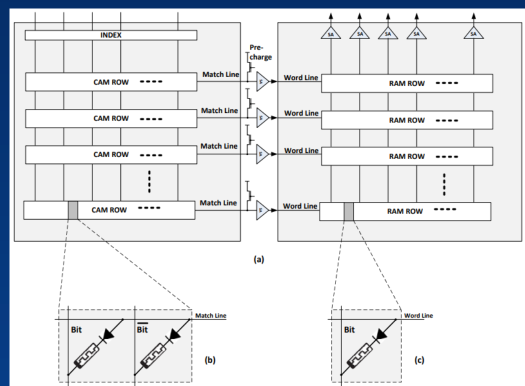
The Actualization Gap

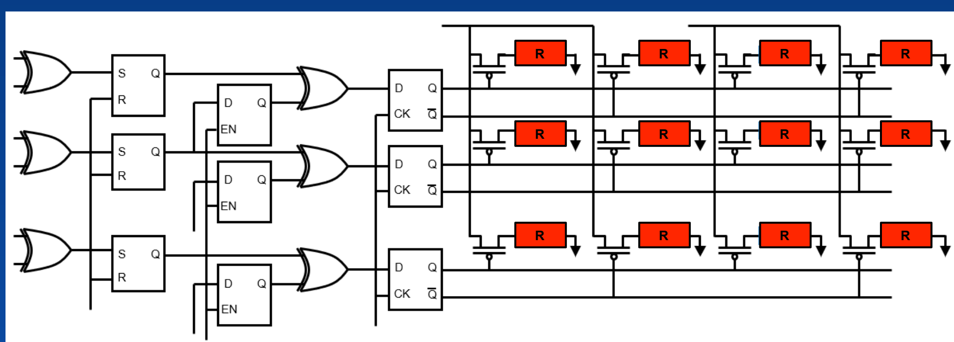
Research projects

Deployments

The “Research” Projects





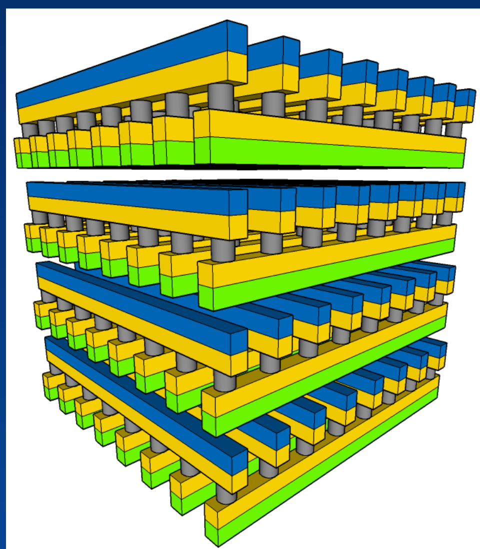


Many interconnects between storage elements and processing elements

Weighted calculations produce parallel possible results

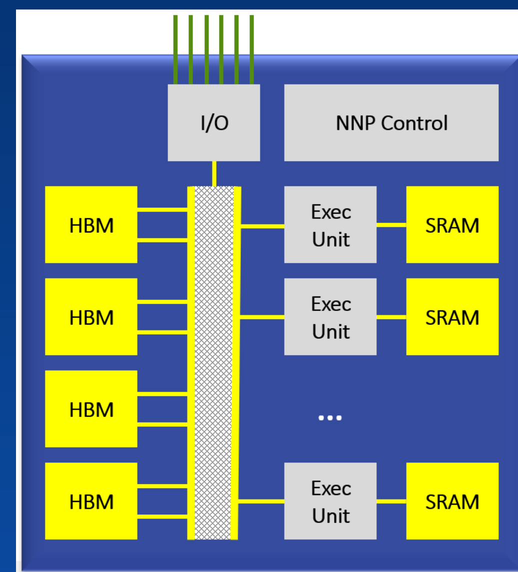
Focus for a number of startup companies

What Most People Mostly Building



Dense matrix memory for highest storage capacity

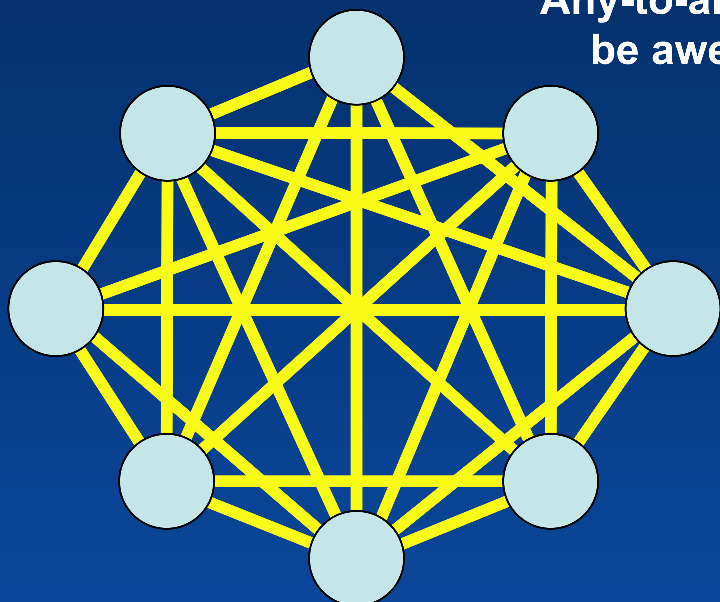
Pipes for networking



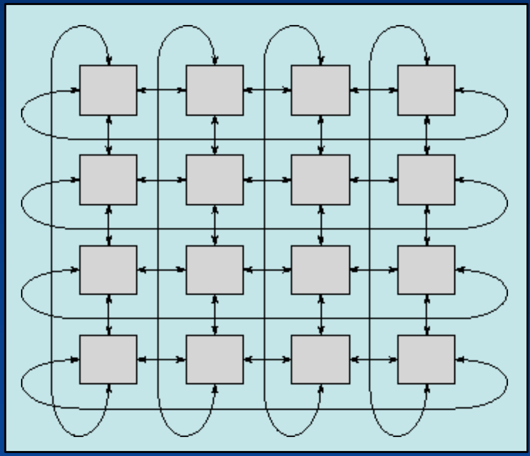
Shared memory controller for many execution units

Practical I/O Connection Limits

Any-to-any would be awesome

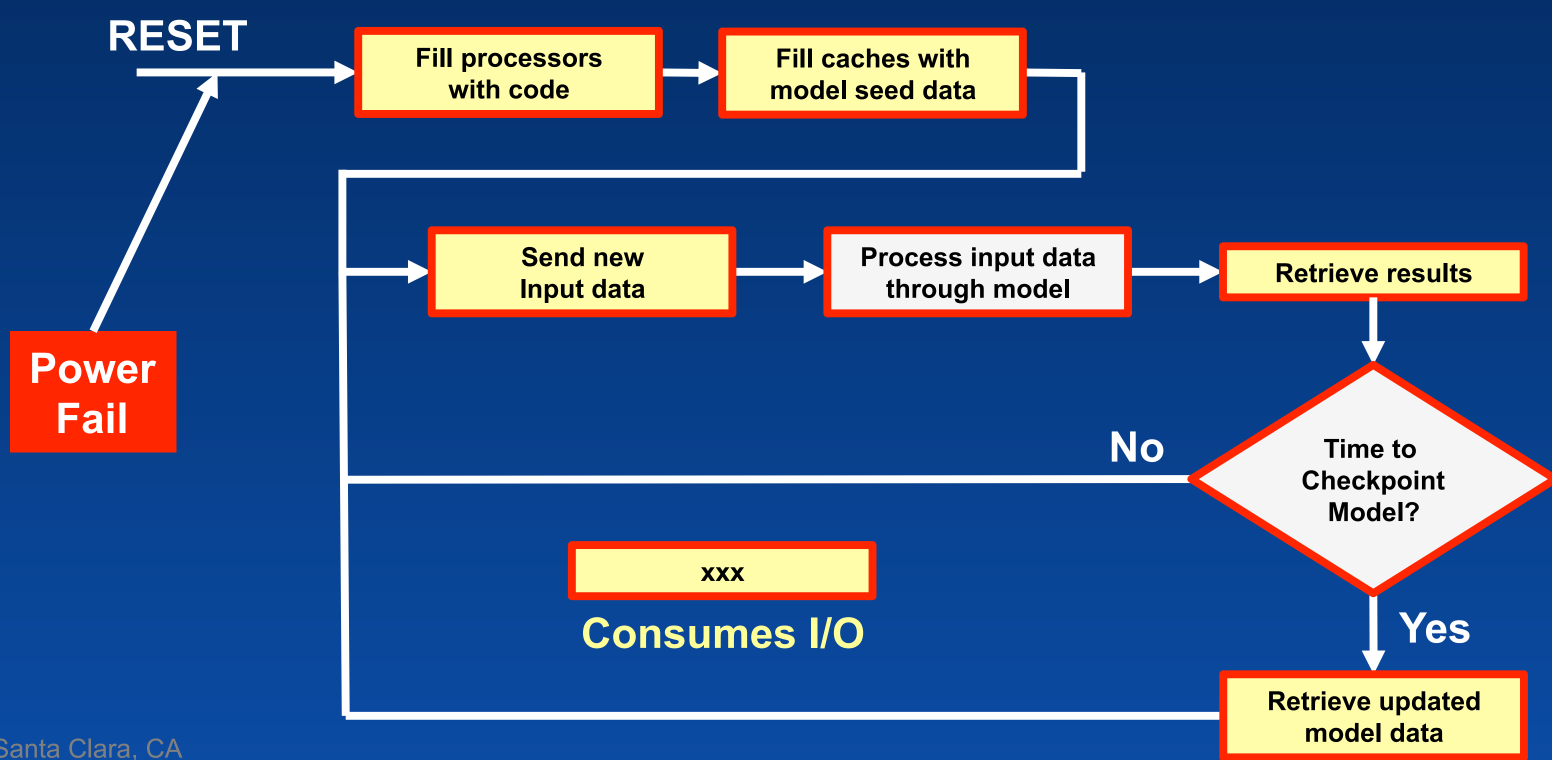


Toroid is a more practicable solution



Limits how quickly data can flow in and out

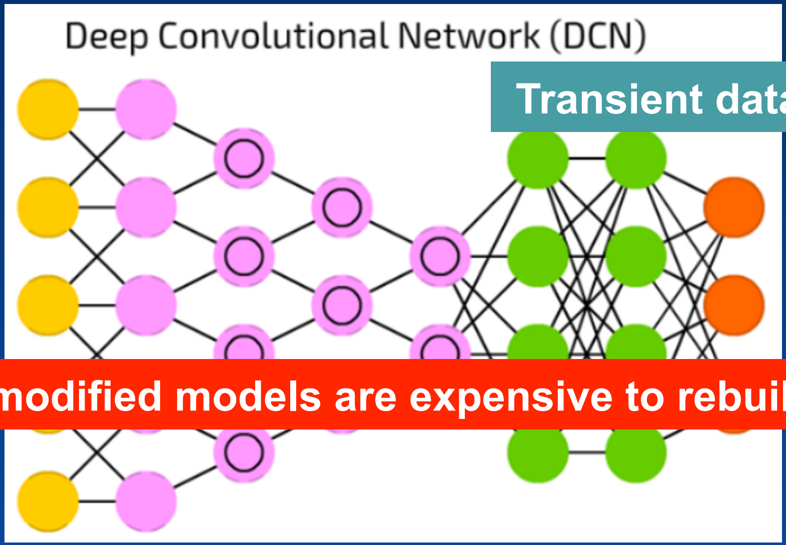
Network Utilization



Flash Memory Summit

Lossless Versus Lossy

Persistent data: reload needed



Transient data: reload, restart calculations

Accumulated data: modified models are expensive to rebuild

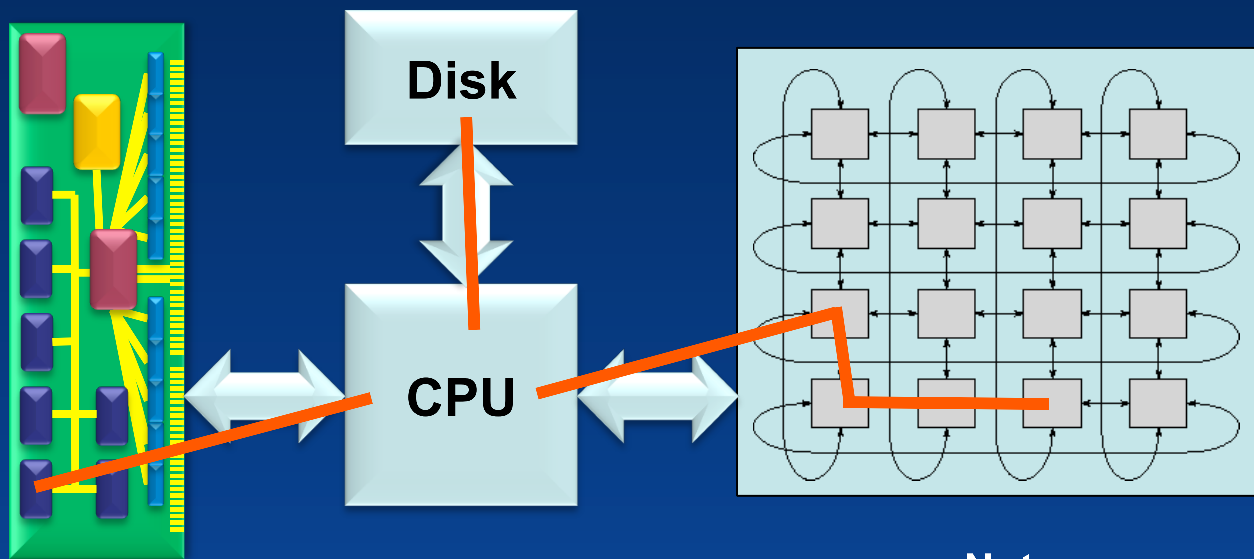
Time to reload is always an issue

Recovering From Power Fail

Data pulled from main memory

…or worse…

Backing store

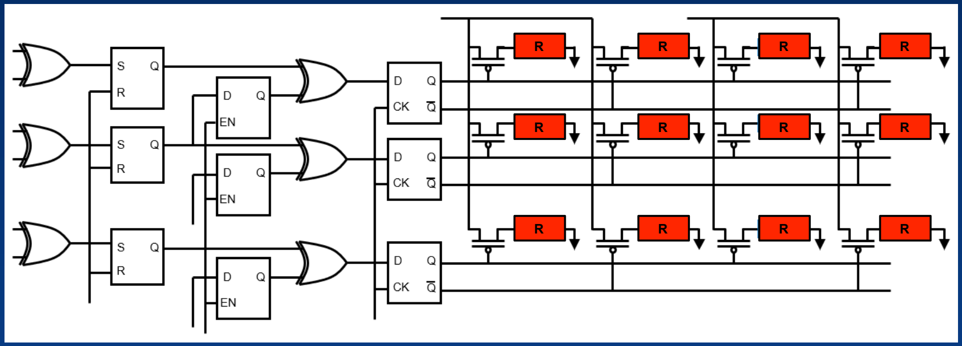


Data requires multiple hops through the interconnects

Not uncommon for data reload to take 3 minutes or more

Before recalculation can begin!

Distributed Memory Complications



This may help explain the gap between research projects and actual deployments

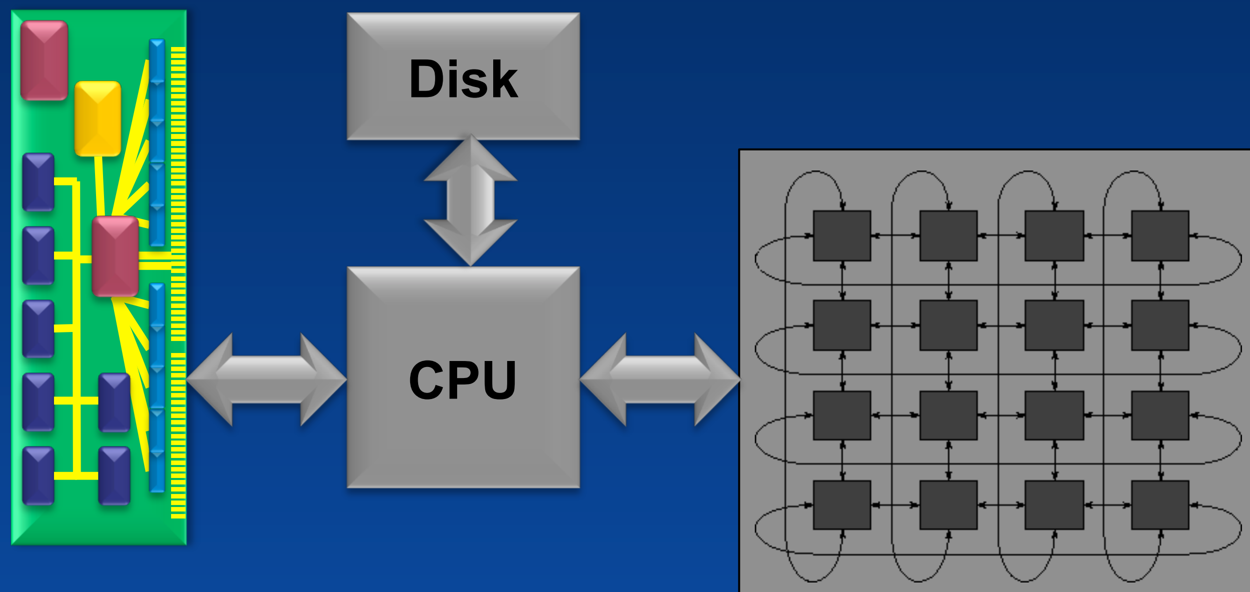
Distributed cells complicate download time into the arrays

Persistent Main Memory

NVDIMMs are moving data persistence to the main memory bus

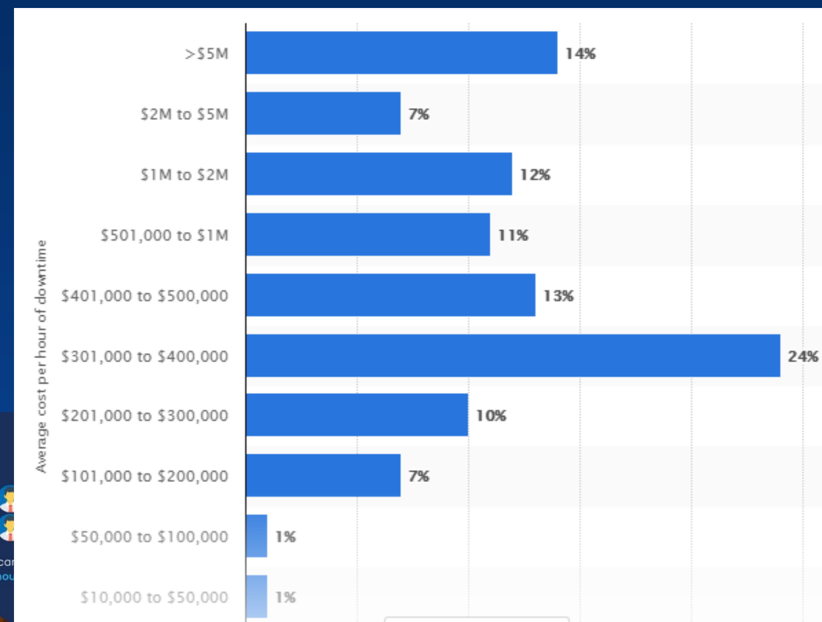
and in some cases increasing memory capacity

See my other talk later this week



Cost of Power Failure

The Cost Of Internet Outages Beyond Revenue



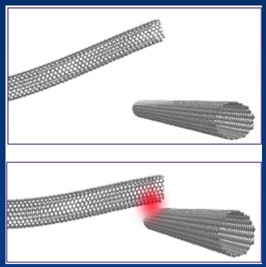
Statistics vary but all agree… downtime costs a LOT

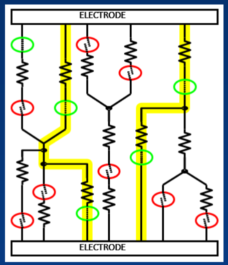
Persistent Memory

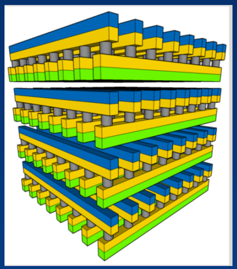
DRAM
Loses data
Must be refreshed
Can't lose power

Persistent Memory
Holds data
forever, even
on power fail

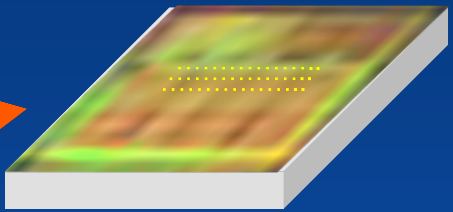
Nantero NRAM™







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

See my other talks later this week

Flash Memory Summit

NANTERO

Classes of Persistent Memory

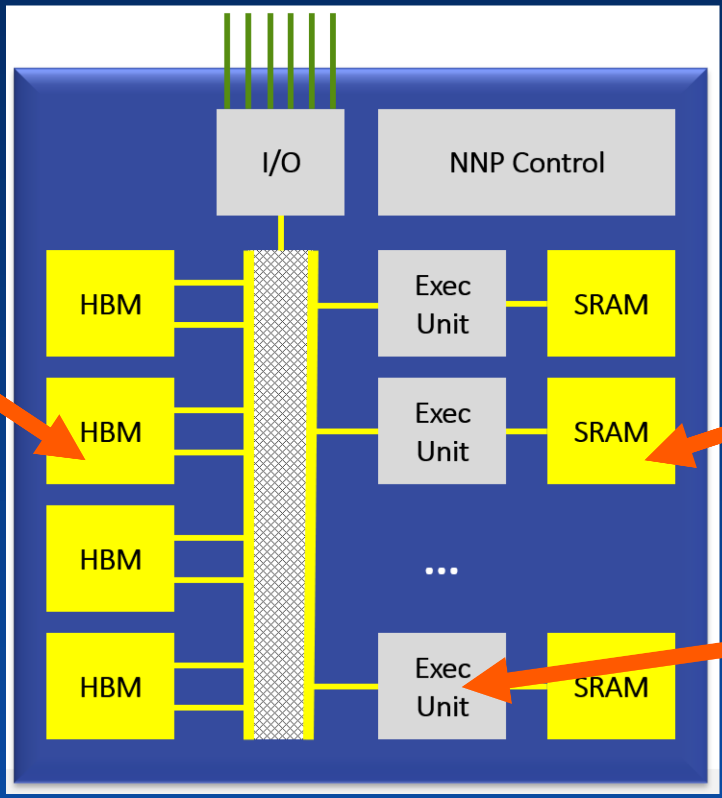
	DRAM	NRAM	MRAM	ReRAM	PCM / 3DXpoint	FeRAM	Flash
Non-volatility	No	Yes	Yes	Yes	Yes	Yes	Yes
Endurance	No limit	No limit	Limited	Limited	Limited	No limit	10^3
Read Time	10 ns	10 ns	X	X	X	X	50M ns
Write Time	10 ns	10 ns	X	X	X	X	25M ns
	Memory & Memory Class Storage		Storage Class Memory				Storage

See my other talks later this week

Applying Persistent Memory

Replace DRAM with Persistent Memory

Completely eliminates the need to reload on Power fail

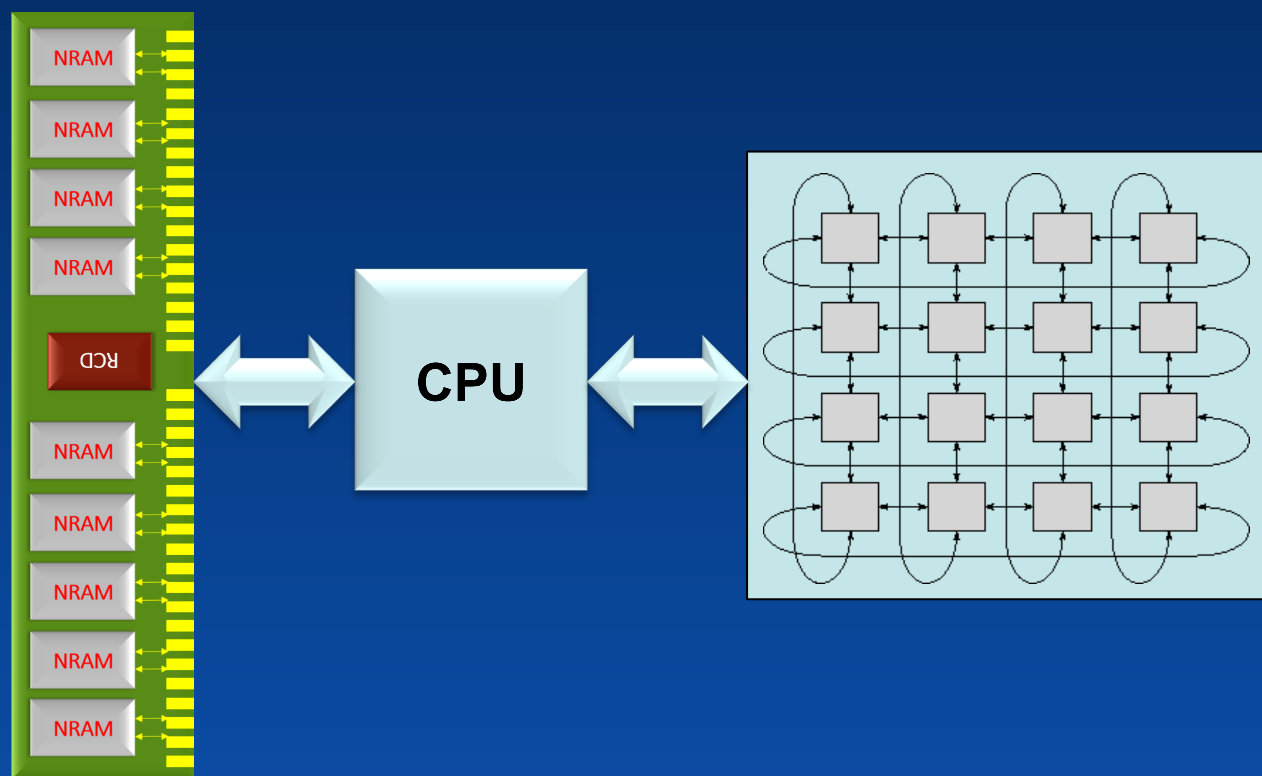


Next generation persistent memory will target SRAM, too

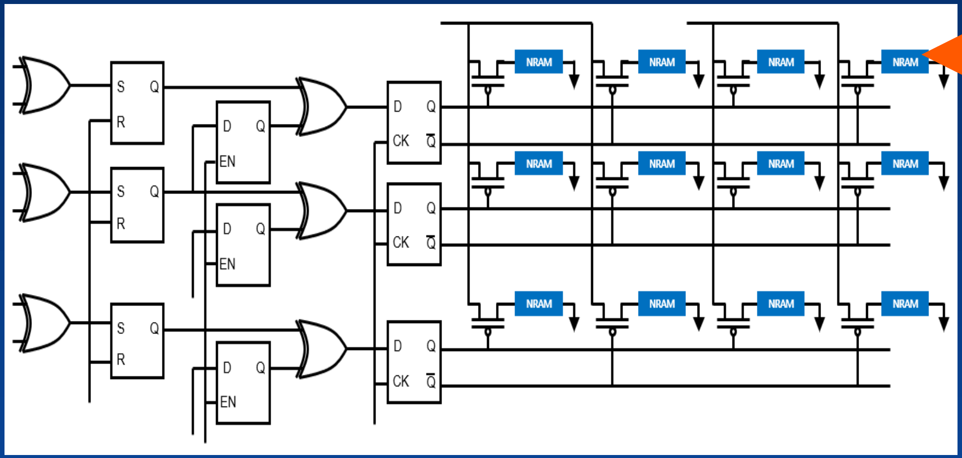
Persistent shadow registers aren't such a bad idea, either

NRAM for Main Memory

NRAM replaces DDR4, DDR5 for main memory



Enables the New Architectures



NRAM cells in the array

Permanent storage through power fail

Programmed once during manufacturing, no reload

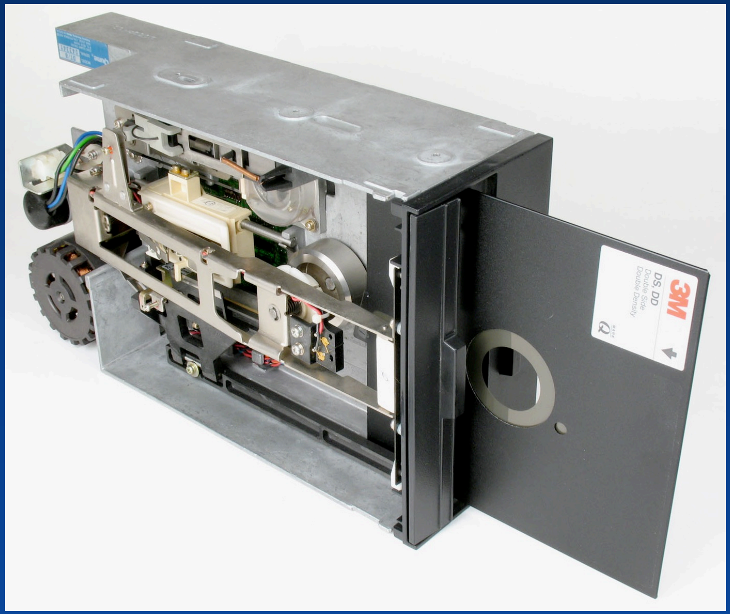
NRAM Everywhere

Soon we will look back and say

“Remember when data was lost when power went out?”

and laugh

Full Disclosure



My first home computer had an 8” floppy disk

I earned my gray hair

Summary

Centralized versus distributed computing is a long term cycle

Quality of software infrastructure typically determines the winner

Artificial intelligence accelerators are a recent co-processing addition

Data loss on power failure is worsened by AI architectures

Persistent memory in AI device solves major problems

Nantero NRAM addresses many usages of PM in AI systems

If you remember 8” floppies, you probably can’t read this screen

Questions?

Bill Gervasi

Principal Systems Architect

bilge@Nantero.com