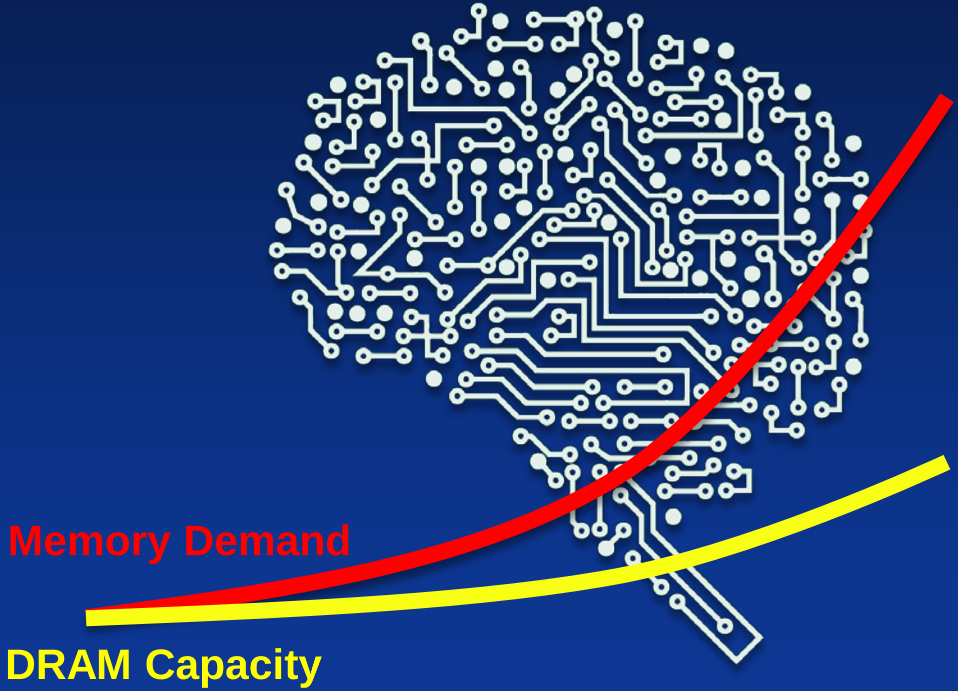
Persistent Memory for Artificial Intelligence

Bill Gervasi

Principal Systems Architect

bilge@Nantero.com

Demand Outpacing Capacity



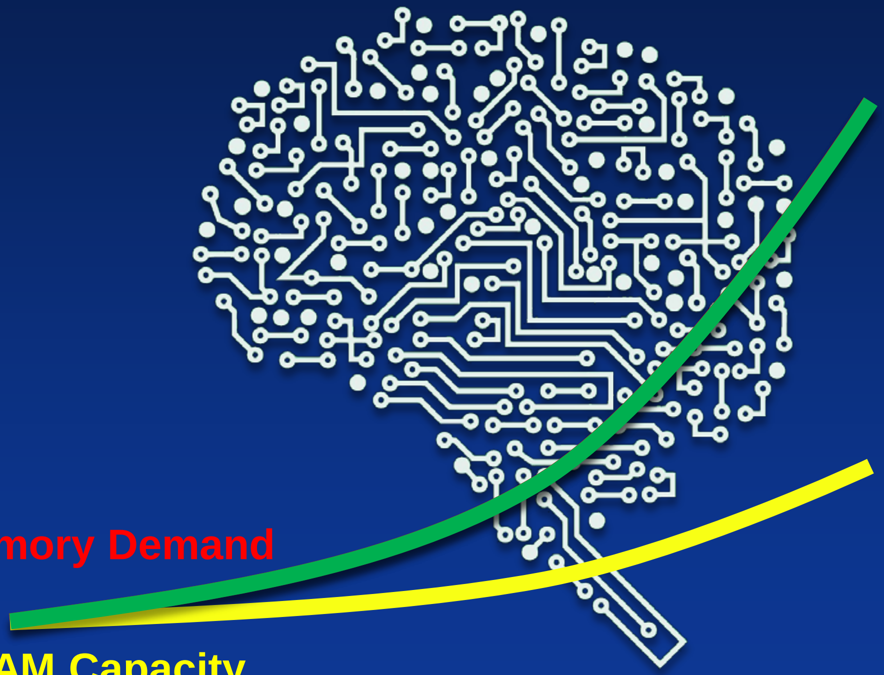
In-Memory Computing

Artificial Intelligence

Machine Learning

Deep Learning

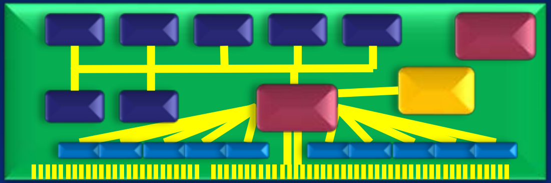
Driving New Capacity Models



Industry successfully snuggling large memories to the processors…

…but we can do oh! so much more

My Three Talks at FMS



NVDIMM Analysis

Memory Class Storage

Artificial Intelligence

History of Architectures

Let’s go back in time…

Historical Trends in Computing

Central Computing

Client Computing

Edge Computing

Central Processing

Distributed Processing

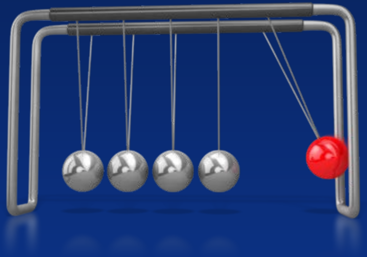
Co-Processing

Power Failure
Data Loss

Some Moments in History

Central Processing

Shared Processor
Dumb terminals



Peer-to-peer networks

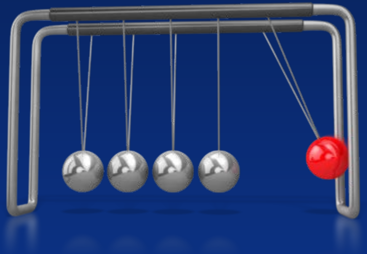
Distributed Processing

Processor per user

Some Moments in History

Central Processing

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



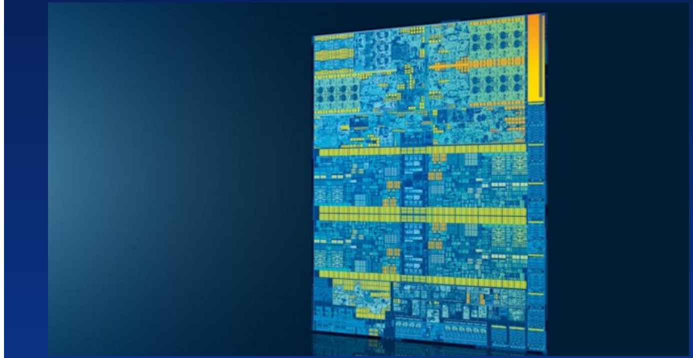
Tightly-coupled coprocessing

Distributed Processing

Hercules graphics
Sound Blaster audio
Rockwell modem
Ethernet DSP

The Lone Survivor…

Graphics add-in cards

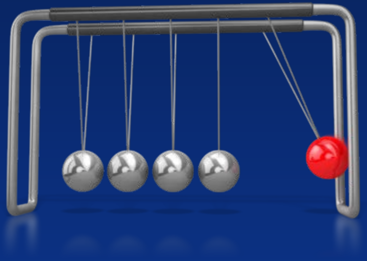


Integrated graphics

…survived the NSP war

Some Moments in History

Central Processing



Distributed Processing

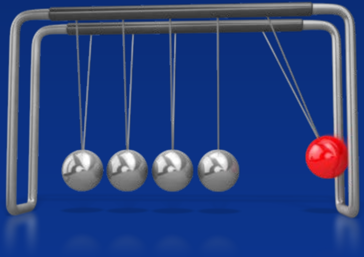
Phone providers controlled all data processing

Edge computing reduces latency

Phone apps provide local services

When the Playing Field Changes

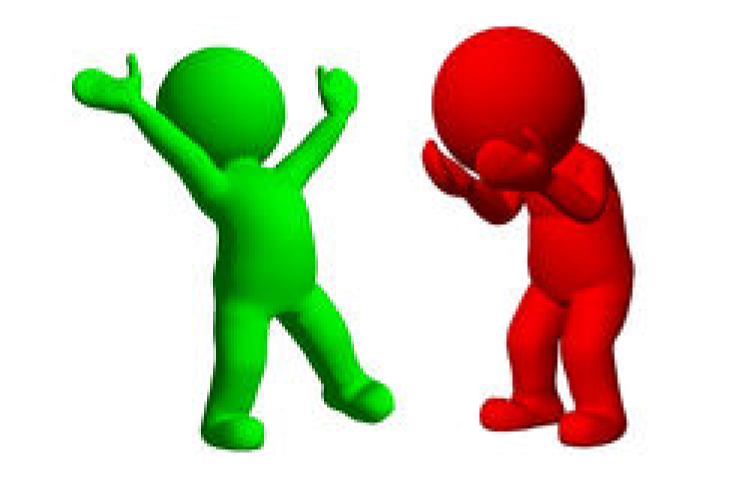
The speed of networking directly impacts the pendulum swing from centralized to distributed

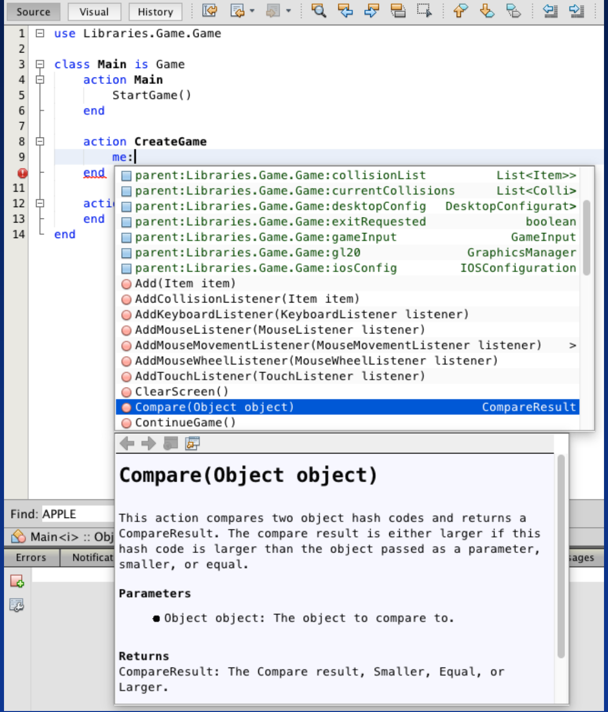


A faster network favors distributed computing

Winners and Losers

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





```
use Libraries.Game.Game

class Main is Game
    action Main
        StartGame()
    end

    action CreateGame
        me:
    end

    acti
    end
end
```

```
parent:Libraries.Game.Game:collisionList          List<Item>>
parent:Libraries.Game.Game:currentCollisions      List<Colli>
parent:Libraries.Game.Game:desktopConfig   DesktopConfigurat>
parent:Libraries.Game.Game:exitRequested              boolean
parent:Libraries.Game.Game:gameInput                GameInput
parent:Libraries.Game.Game:gl20               GraphicsManager
parent:Libraries.Game.Game:iosConfig         IOSConfiguration
Add(Item item)
AddCollisionListener(Item item)
AddKeyboardListener(KeyboardListener listener)
AddMouseListener(MouseListener listener)
AddMouseMovementListener(MouseMovementListener listener)    >
AddMouseWheelListener(MouseWheelListener listener)
AddTouchListener(TouchListener listener)
ClearScreen()
Compare(Object object)                          CompareResult
ContinueGame()
```

Compare(Object object)

This action compares two object hash codes and returns a CompareResult. The compare result is either larger if this hash code is larger than the object passed as a parameter, smaller, or equal.

Parameters

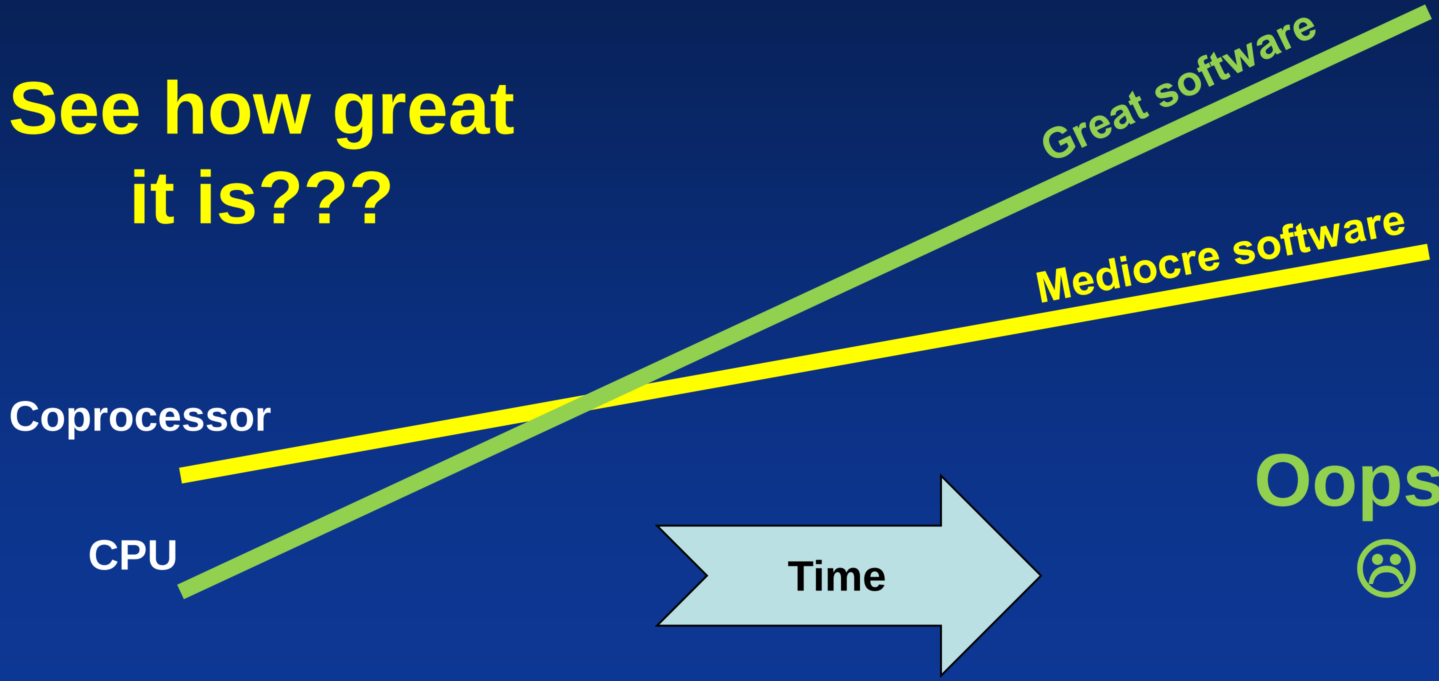
- Object object: The object to compare to.

Returns

CompareResult: The Compare result, Smaller, Equal, or Larger.

Maintaining an Edge

See how great it is???



Oops!!!
☹

The Tail Wagging the Dog

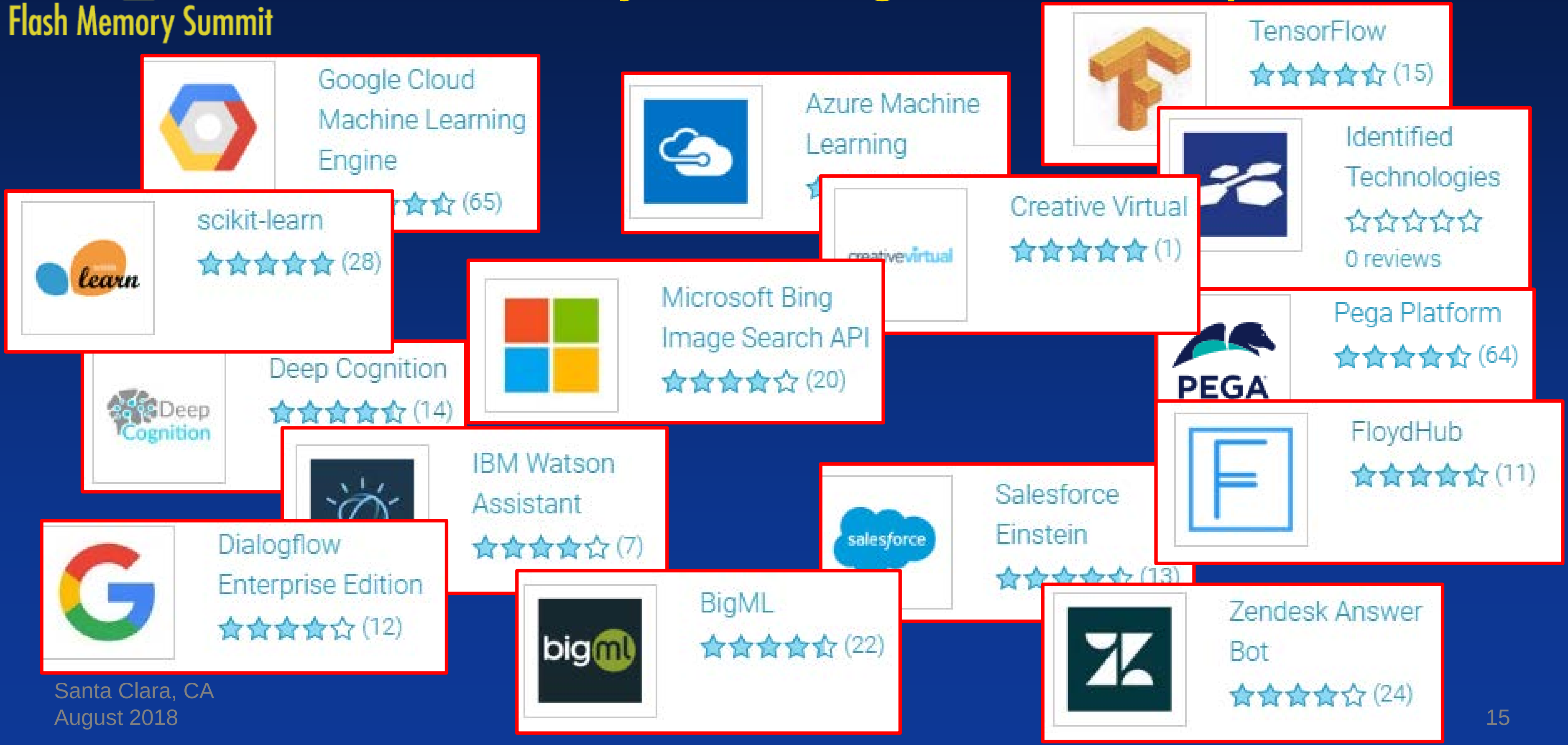
I won’t say “It’s the Software, Stupid” because I know you’re not stupid

however

To succeed, AI needs GREAT software infrastructure

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

Wild Array of Programmer Options



TensorFlow (15)

Google Cloud Machine Learning Engine (65)

Azure Machine Learning

Identified Technologies 0 reviews

scikit-learn (28)

Creative Virtual (1)

Microsoft Bing Image Search API (20)

Pega Platform (64)

Deep Cognition (14)

IBM Watson Assistant (7)

FloydHub (11)

Salesforce Einstein (13)

Dialogflow Enterprise Edition (12)

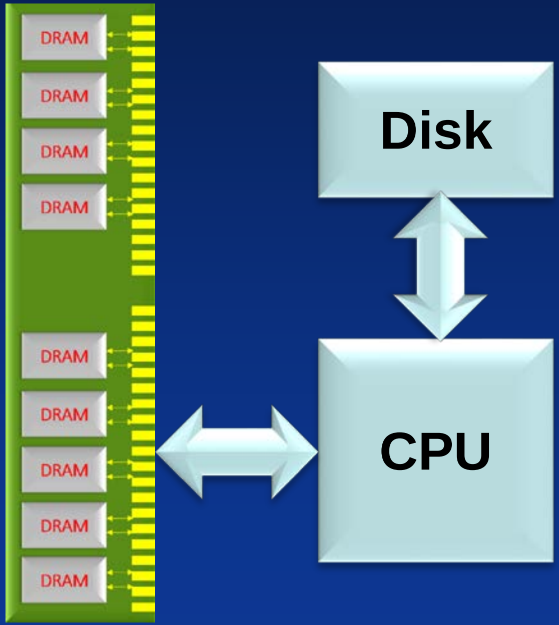
BigML (22)

Zendesk Answer Bot (24)

Flash Memory Summit

NANTERO

AI on Traditional Server



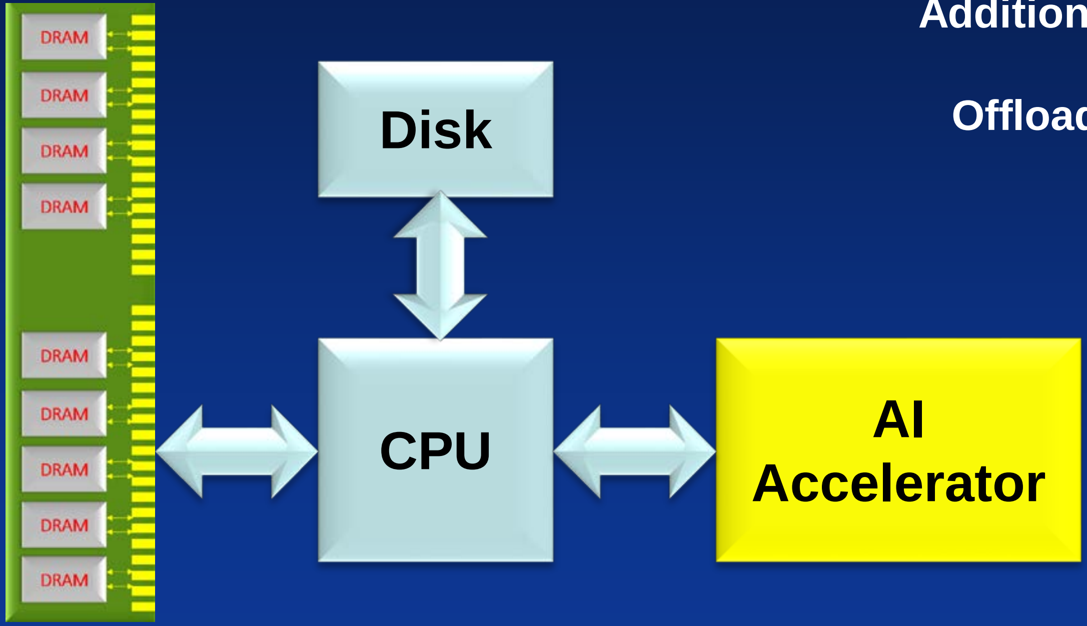
No magic

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

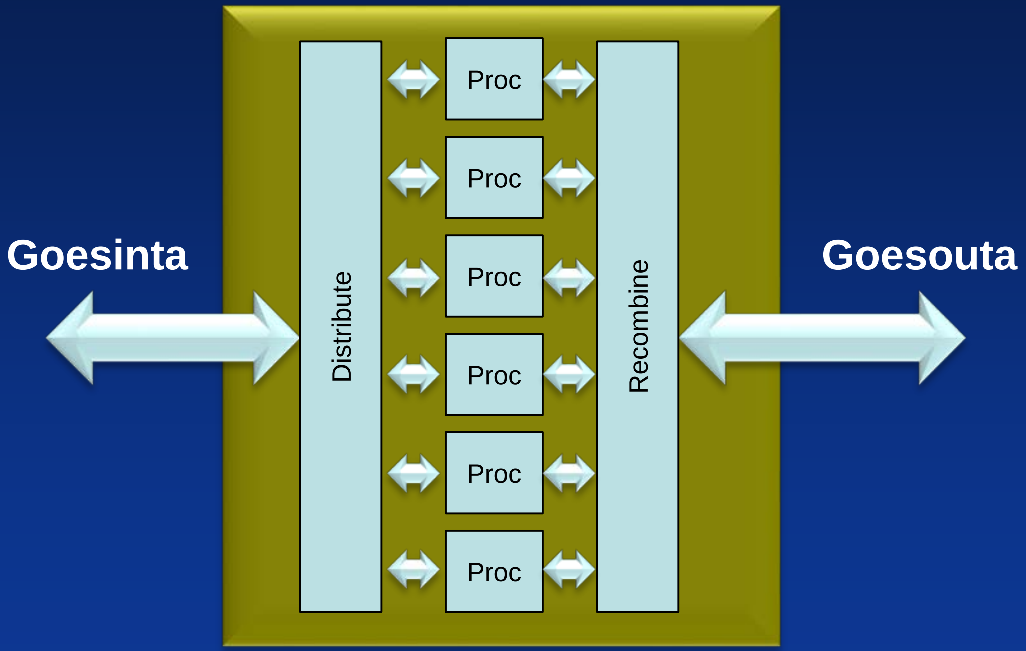
AI Evolution



Addition of AI Accelerator

Offloads main CPU for AI tasks

AI Evolution



AI Accelerator Characteristics

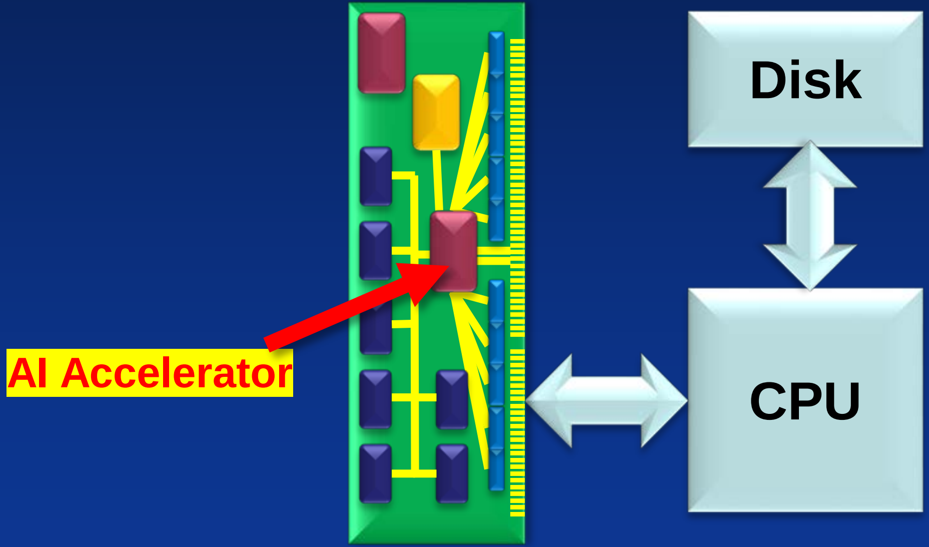
Wide array of simple processing elements

Reduced floating point precision

Tuned for matrix operations

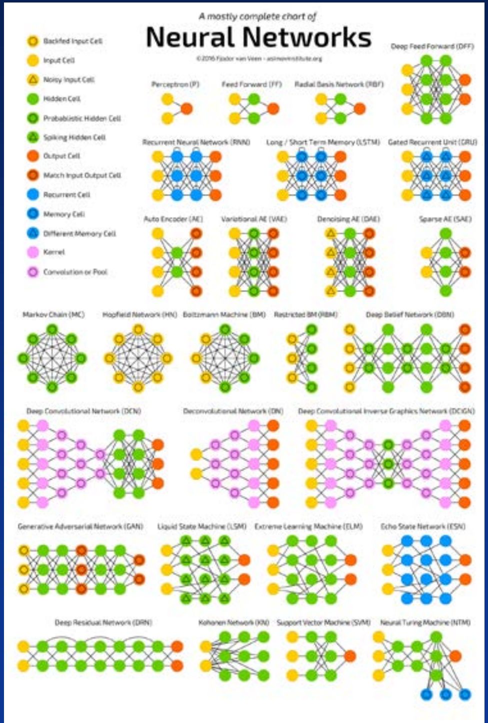
In-Memory Computing

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



Also great for encryption

Data Processing Paradigms


A mostly complete chart of
Neural Networks
©2016 Fjodor van Veen - asimovinstitute.org
Backfed Input Cell
Input Cell
Noisy Input Cell
Hidden Cell
Probablistic Hidden Cell
Spiking Hidden Cell
Output Cell
Match Input Output Cell
Recurrent Cell
Memory Cell
Different Memory Cell
Kernel
Convolution or Pool
Perceptron (P)
Feed Forward (FF)
Radial Basis Network (RBF)
Deep Feed Forward (DFF)
Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Auto Encoder (AE)
Variational AE (VAE)
Denoising AE (DAE)
Sparse AE (SAE)
Markov Chain (MC)
Hopfield Network (HN)
Boltzmann Machine (BM)
Restricted BM (RBM)
Deep Belief Network (DBN)
Deep Convolutional Network (DCN)
Deconvolutional Network (DN)
Deep Convolutional Inverse Graphics Network (DCIGN)
Generative Adversarial Network (GAN)
Liquid State Machine (LSM)
Extreme Learning Machine (ELM)
Echo State Network (ESN)
Deep Residual Network (DRN)
Kohonen Network (KN)
Support Vector Machine (SVM)
Neural Turing Machine (NTM)

Traditional database

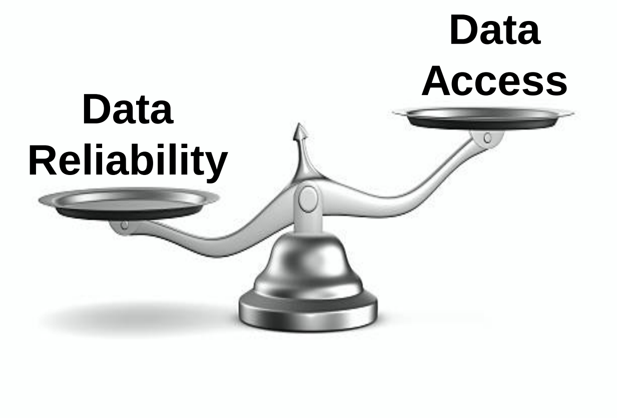
Data mining

Inferencing

Fuzzy logic

Recognition

etc

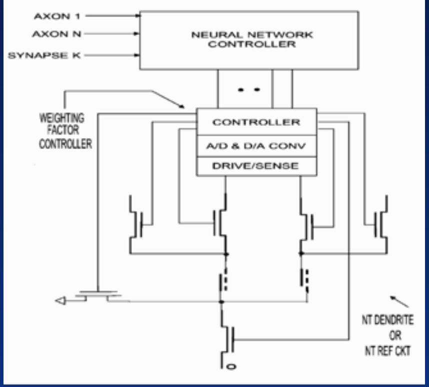


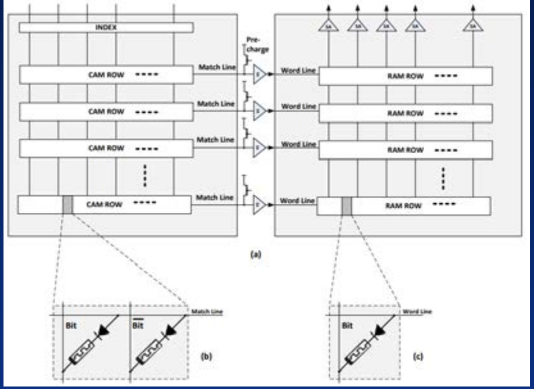
The Actualization Gap

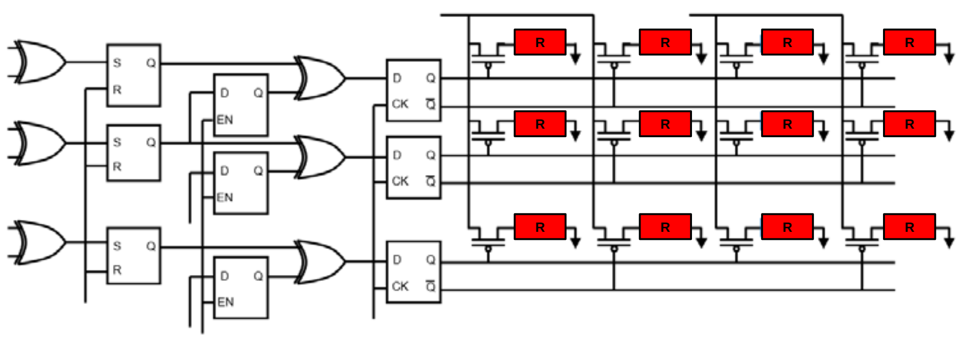
Research projects

Deployments

The “Research” Projects


AXON 1
AXON N
SYNAPSE K
NEURAL NETWORK CONTROLLER
WEIGHTING FACTOR CONTROLLER
CONTROLLER
A/D & D/A CONV
DRIVE/SENSE
NT DENDRITE OR NT REF CKT



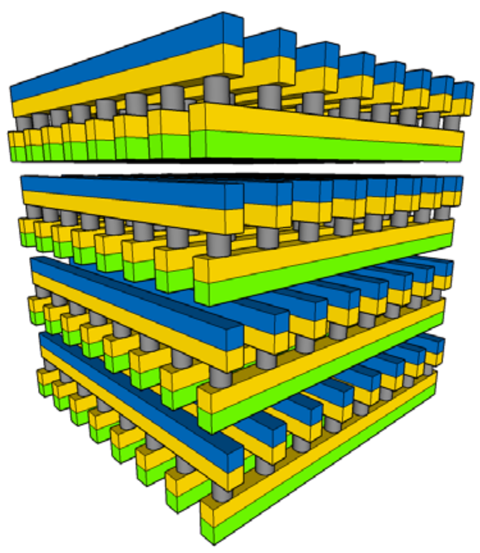


Many interconnects between storage elements and processing elements

Weighted calculations produce parallel possible results

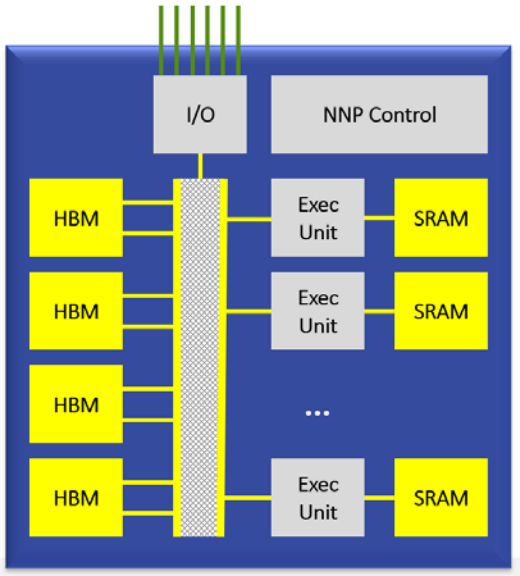
Focus for a number of startup companies

What Most People Mostly Building



Dense matrix memory for highest storage capacity

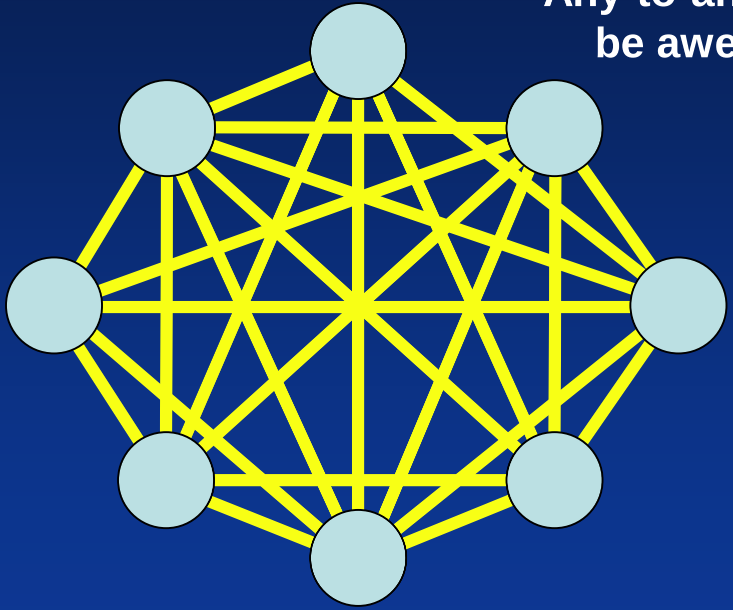
Pipes for networking



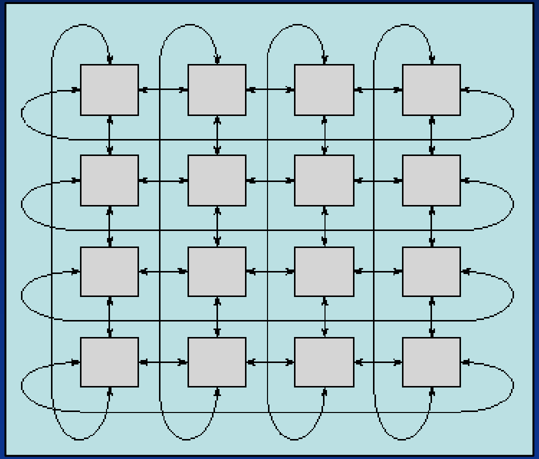
Shared memory controller for many execution units

Practical I/O Connection Limits

Any-to-any would be awesome

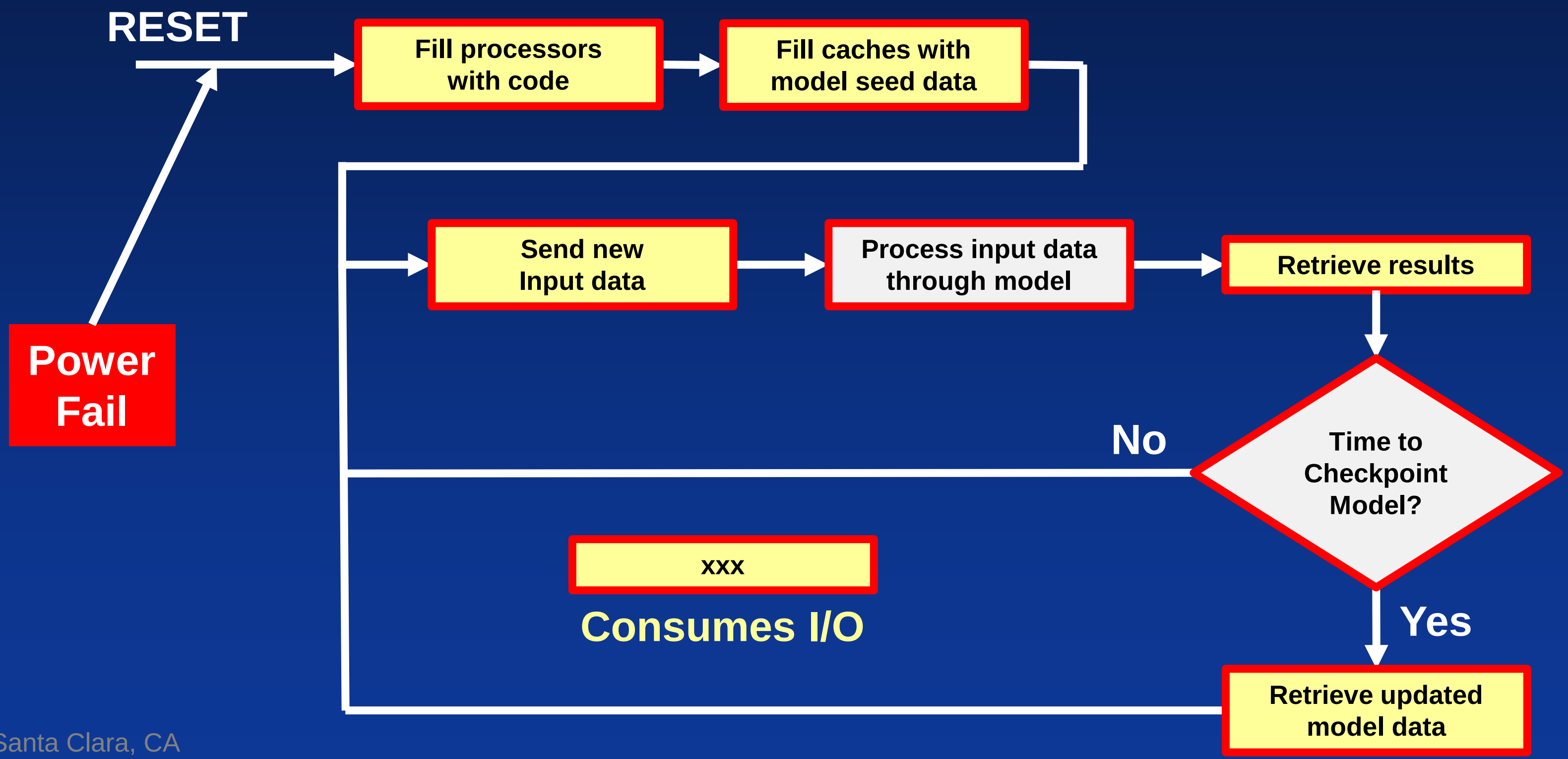


Toroid is a more practicable solution



Limits how quickly data can flow in and out

Network Utilization



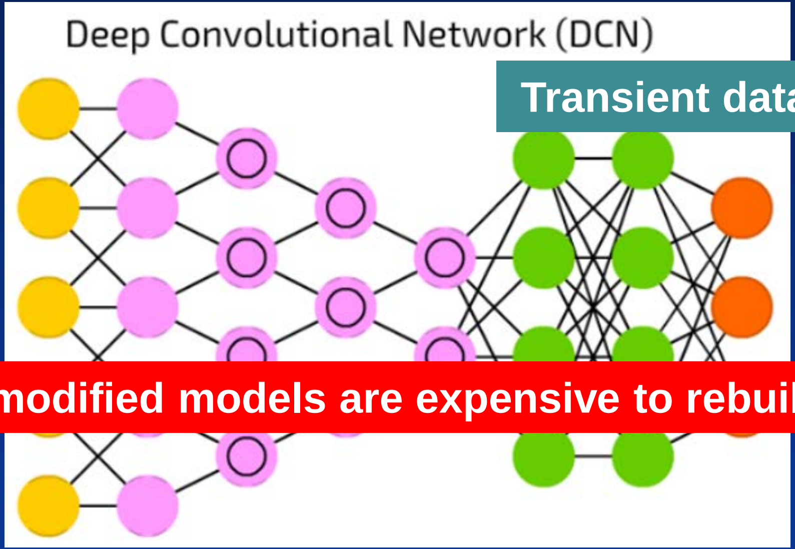
RESET
Fill processors with code
Fill caches with model seed data
Send new Input data
Process input data through model
Retrieve results
Time to Checkpoint Model?
No
Yes
Retrieve updated model data
Power Fail
xxx

Consumes I/O

Flash Memory Summit

Lossless Versus Lossy

Persistent data: reload needed



Transient data: reload, restart calculations

Accumulated data: modified models are expensive to rebuild

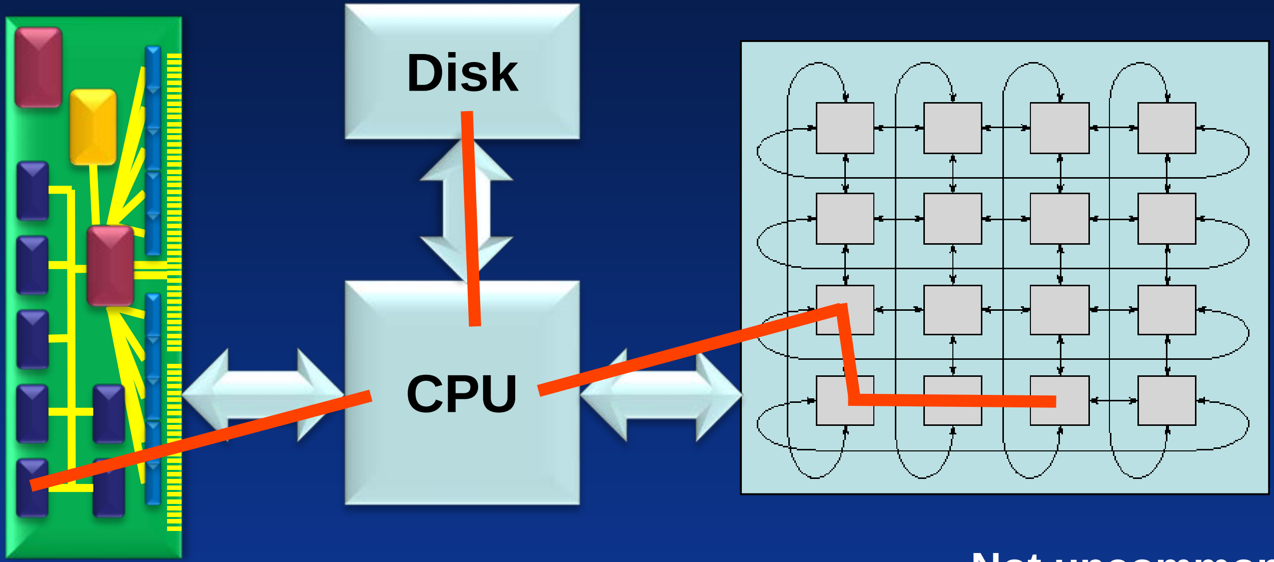
Time to reload is always an issue

Recovering From Power Fail

Data pulled from main memory

…or worse…

Backing store

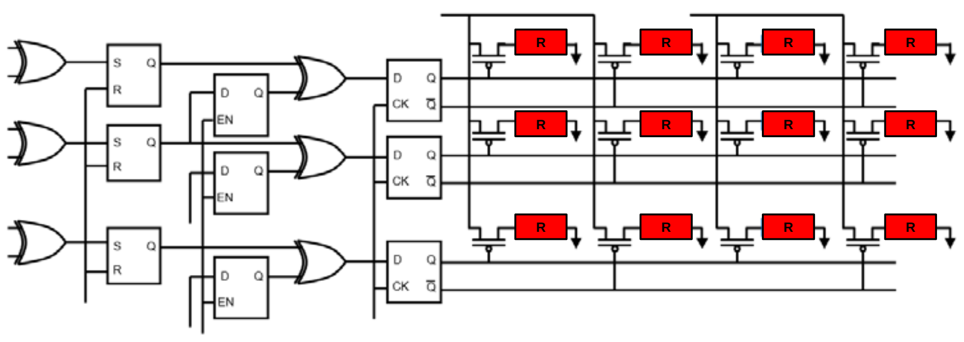


Data requires multiple hops through the interconnects

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

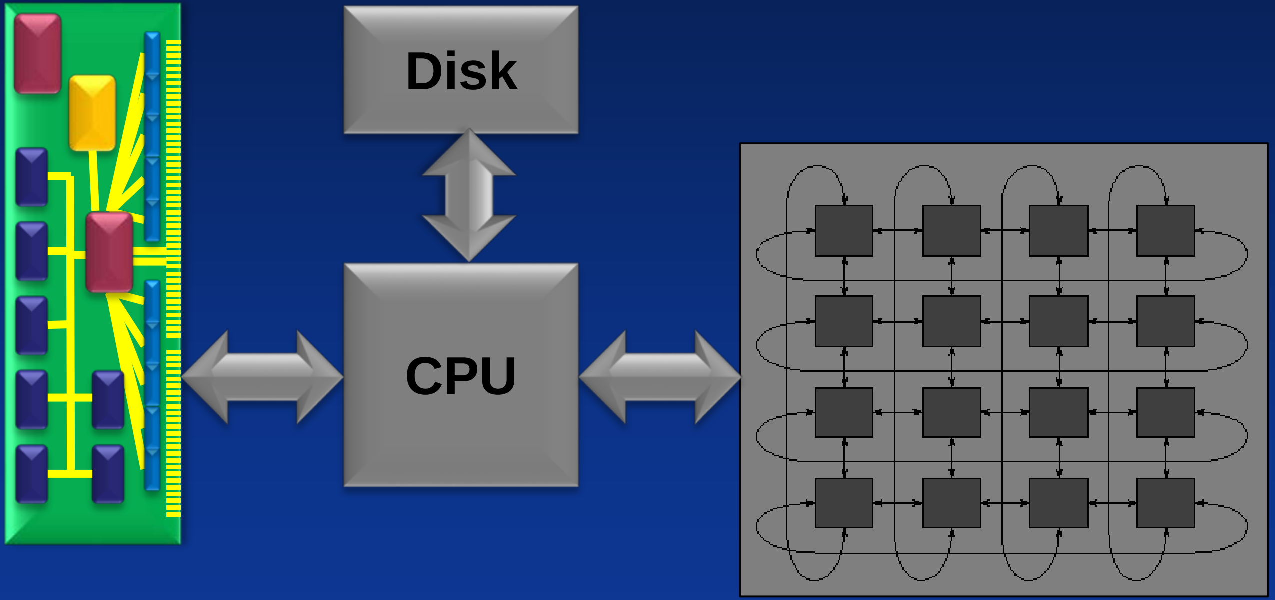
Distributed cells complicate download time into the arrays

Persistent Main Memory

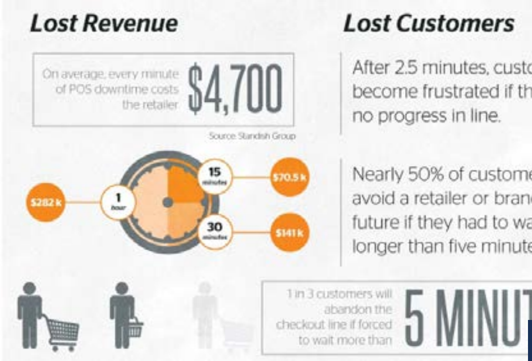
NVDIMMs are moving data persistence to the main memory bus

and in some cases increasing memory capacity

See my other talk later this week



Cost of Power Failure



Lost Revenue

On average, every minute of POS downtime costs the retailer $4,700

Source: Standish Group

Lost Customers

After 2.5 minutes, customers become frustrated if there is no progress in line.

Nearly 50% of customers avoid a retailer or brand in the future if they had to wait longer than five minutes.

1 in 3 customers will abandon the checkout line if forced to wait more than 5 MINUTES

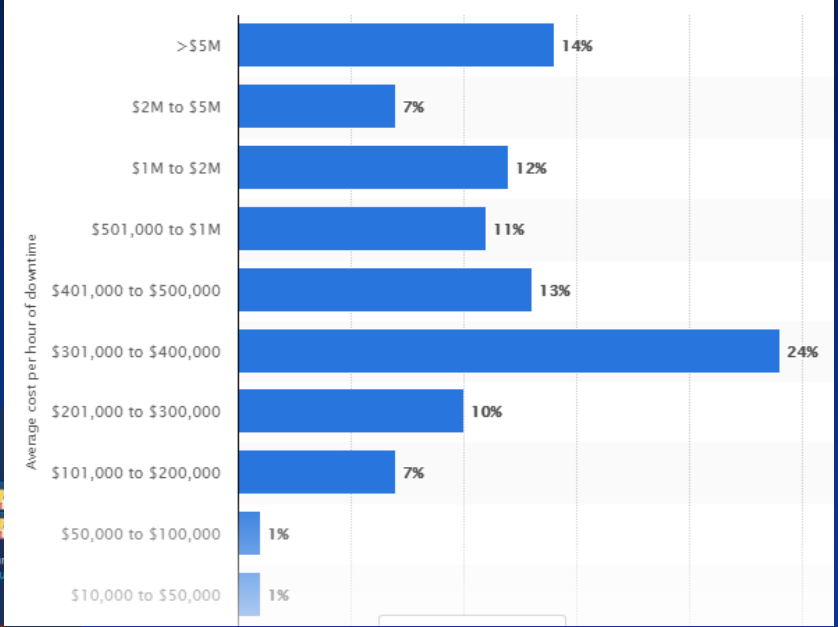
The Cost Of Internet Outages Beyond Revenue

The average downtime in the U.S. is 7.9 hours per outage.

Internet failure ca million person-hou

A company that suffers from worse than average downtime of 175 hours a year can lose more than $7 million per year.

The average cost of downtime per minute is $ 5600 per minute.



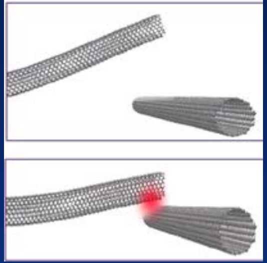
Statistics vary but all agree… downtime costs a LOT

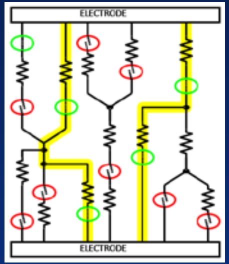
Persistent Memory

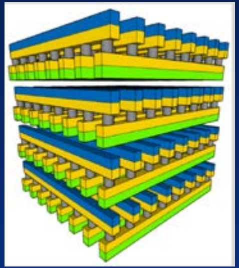
DRAM
Loses data
Must be refreshed
Can’t lose power

Persistent Memory
Holds data
forever, even
on power fail

Nantero NRAM™

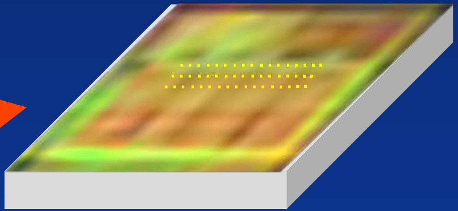






DDR4
DDR5

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



HBM

See my other talks later this week

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

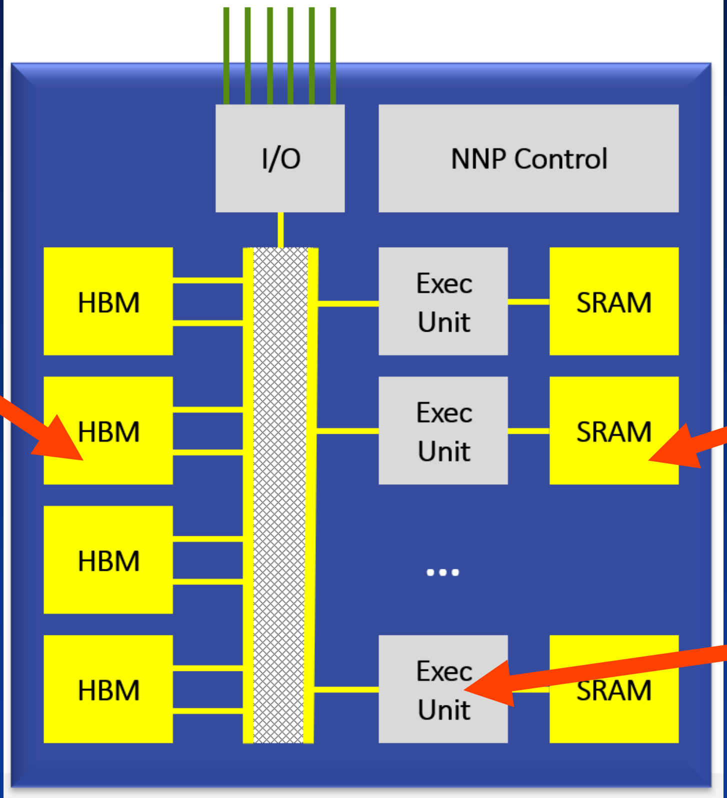
Flash Memory Summit

NANTERO

Applying Persistent Memory

Replace DRAM with Persistent Memory

Completely eliminates the need to reload on Power fail

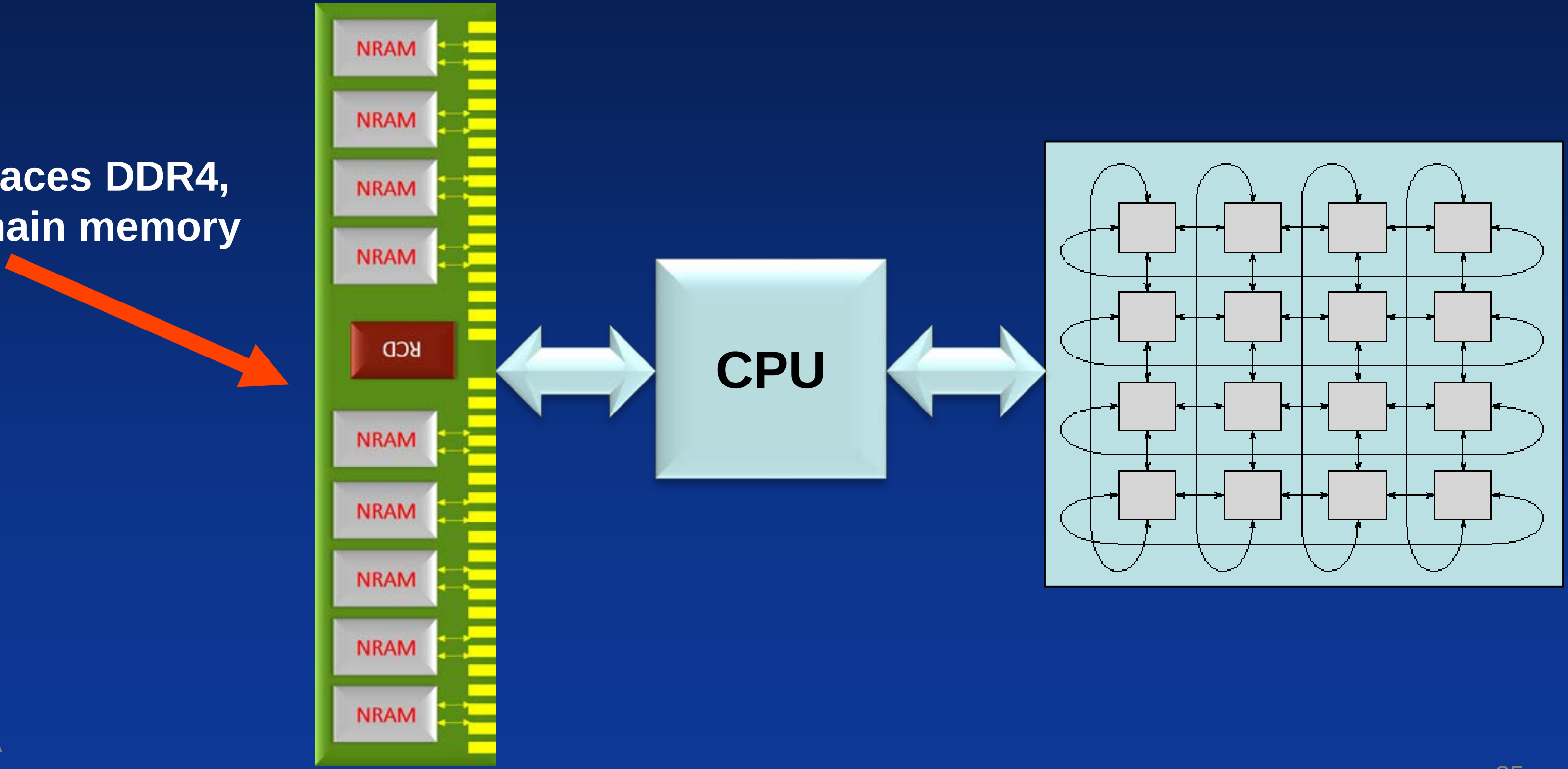


Next generation persistent memory will target SRAM, too

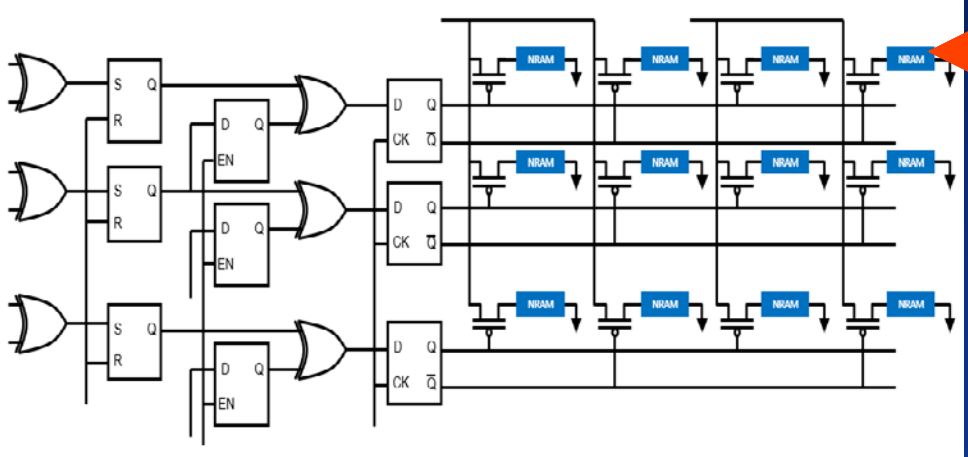
Persistent shadow registers aren't such a bad idea, either

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

Flash Memory Summit

NANTERO

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

Flash Memory Summit

NANTERO

Questions?

Bill Gervasi

Principal Systems Architect

bilge@Nantero.com