



Flash Memory Summit



# Persistent Memory for Artificial Intelligence

Bill Gervasi

Principal Systems Architect

[bilge@Nantero.com](mailto:bilge@Nantero.com)

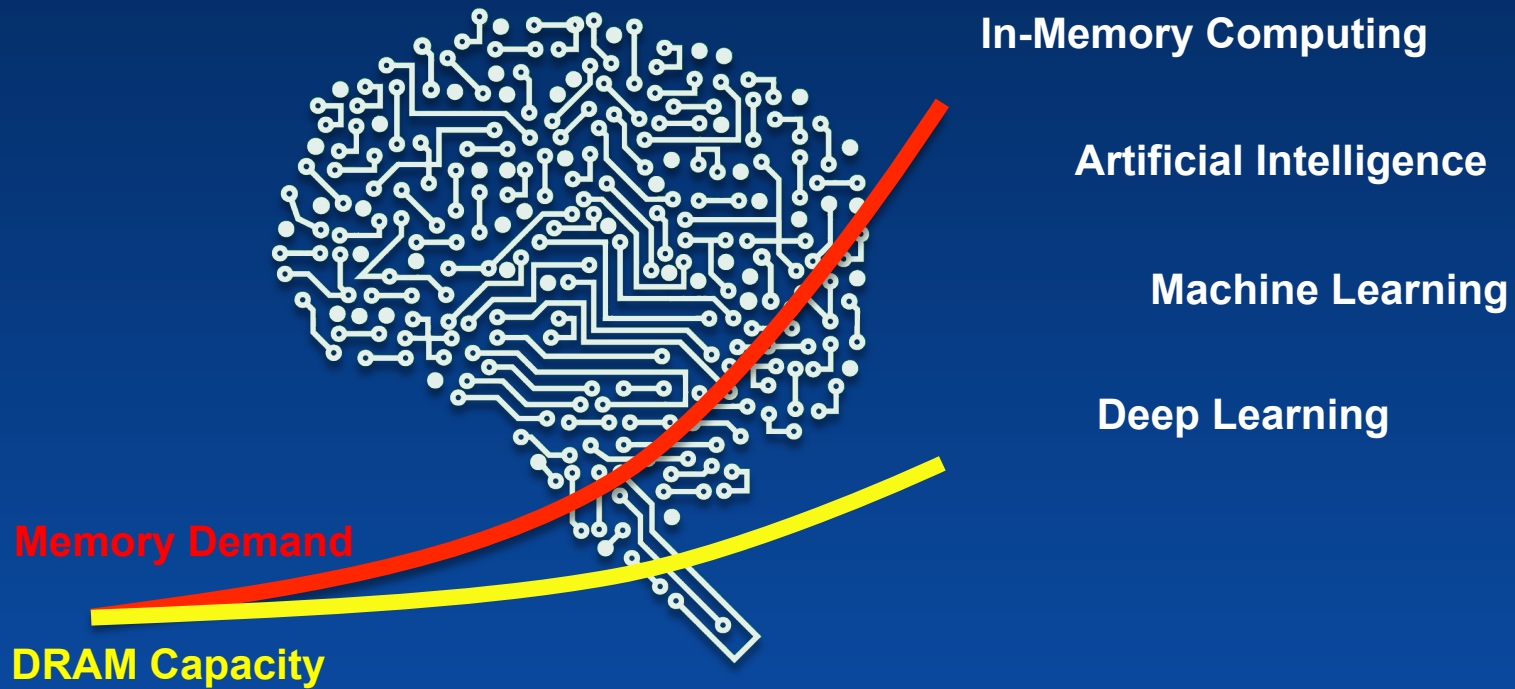
Santa Clara, CA  
August 2018



Flash Memory Summit



# Demand Outpacing Capacity



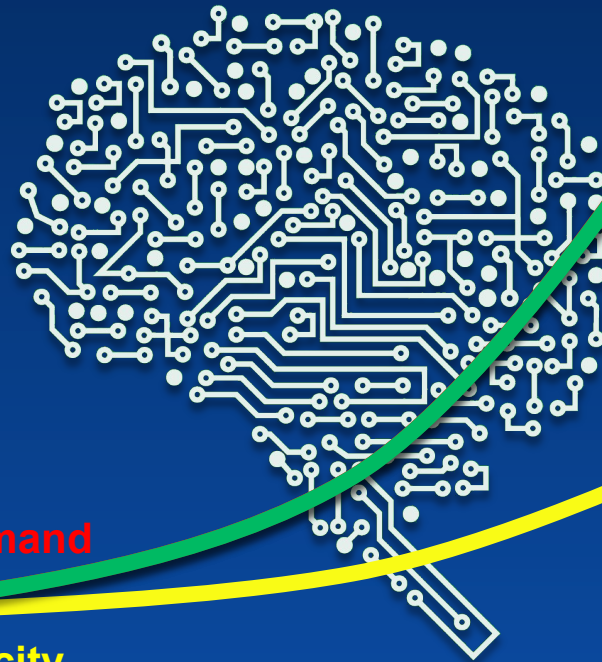
Santa Clara, CA  
August 2018



Flash Memory Summit



# Driving New Capacity Models



**Memory Demand**

**DRAM Capacity**

**Non-volatile memories**

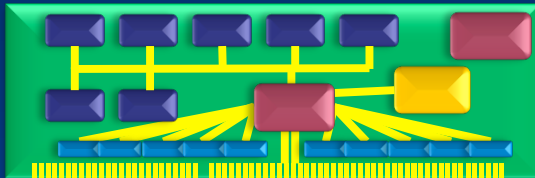
Industry successfully snuggling large memories to the processors...

...but we can do oh! so much more



Flash Memory Summit

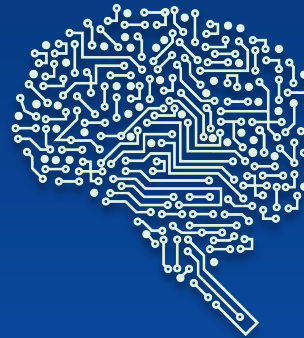
# My Three Talks at FMS



**NVDIMM Analysis**



**Memory Class Storage**



**Artificial Intelligence**



# History of Architectures



Let's go back in time...

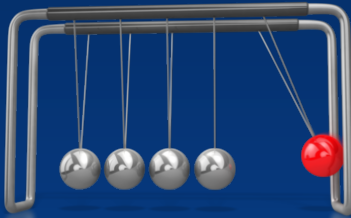


Flash Memory Summit



# Historical Trends in Computing

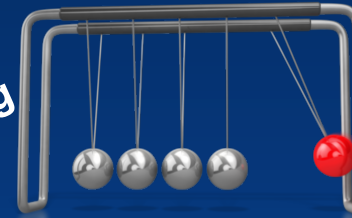
Central Computing



Client Computing

Edge Computing

Central Processing



Distributed Processing

Co-Processing



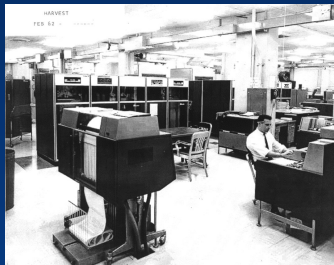
Power Failure  
Data Loss

Santa Clara, CA  
August 2018



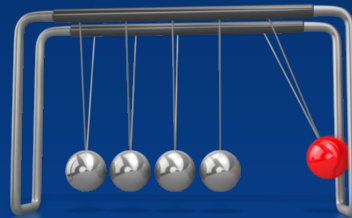
Flash Memory Summit

# Some Moments in History



**Central Processing**

**Shared Processor  
Dumb terminals**



**Distributed Processing**

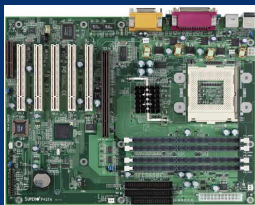
**Processor per user**



**Peer-to-peer networks**

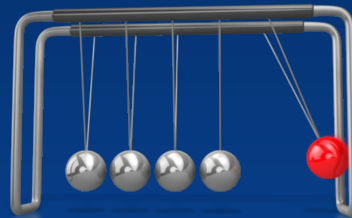


# Some Moments in History



**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



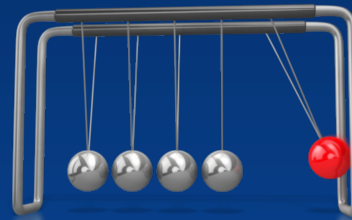
Flash Memory Summit

# Some Moments in History

**Central Processing**



**Phone providers controlled all data processing**



**Distributed Processing**



**Phone apps provide local services**

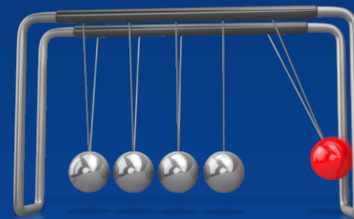


**Edge computing reduces latency**



# 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



```

1 use Libraries.Game.Game
2
3 class Main is Game
4   action Main
5     StartGame()
6   end
7
8   action CreateGame
9     me:
10    end
11  end
12  acti
13  end
14 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()

Find: APPLE  
 Main<> :: Obj  
 Errors Notificat

**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.



Flash Memory Summit



# Maintaining an Edge

See how great  
it is???

Coprocessor

CPU

Great software

Mediocre software



Time

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



Flash Memory Summit

# Wild Array of Programmer Options

Google Cloud Machine Learning Engine  
☆☆☆☆☆ (65)

Azure Machine Learning  
☆☆☆☆☆ (1)

TensorFlow  
☆☆☆☆☆ (15)

scikit-learn  
☆☆☆☆☆ (28)

Creative Virtual  
☆☆☆☆☆ (1)

Identified Technologies  
☆☆☆☆☆  
0 reviews

Deep Cognition  
☆☆☆☆☆ (14)

Microsoft Bing Image Search API  
☆☆☆☆☆ (20)

Pega Platform  
☆☆☆☆☆ (64)

IBM Watson Assistant  
☆☆☆☆☆ (7)

FloydHub  
☆☆☆☆☆ (11)

Salesforce Einstein  
☆☆☆☆☆ (13)

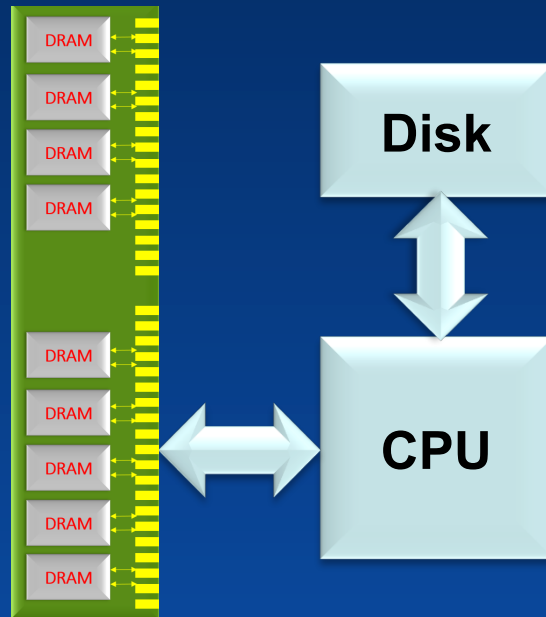
Dialogflow Enterprise Edition  
☆☆☆☆☆ (12)

BigML  
☆☆☆☆☆ (22)

Zendesk Answer Bot  
☆☆☆☆☆ (24)



# AI on Traditional Server



No magic

AI applications are like any other

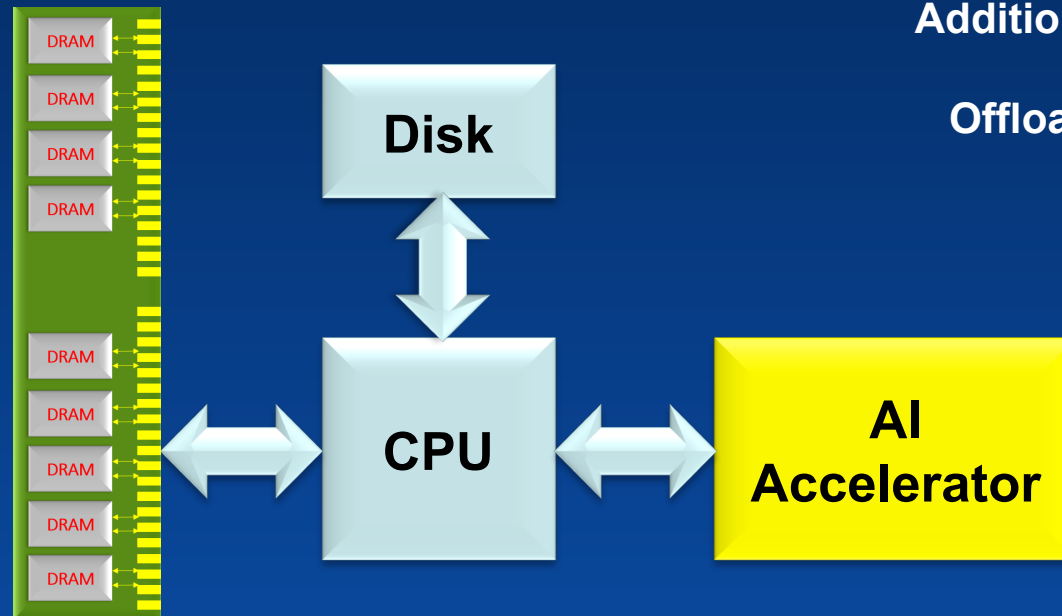
Data processing done on main CPU

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





# AI Evolution

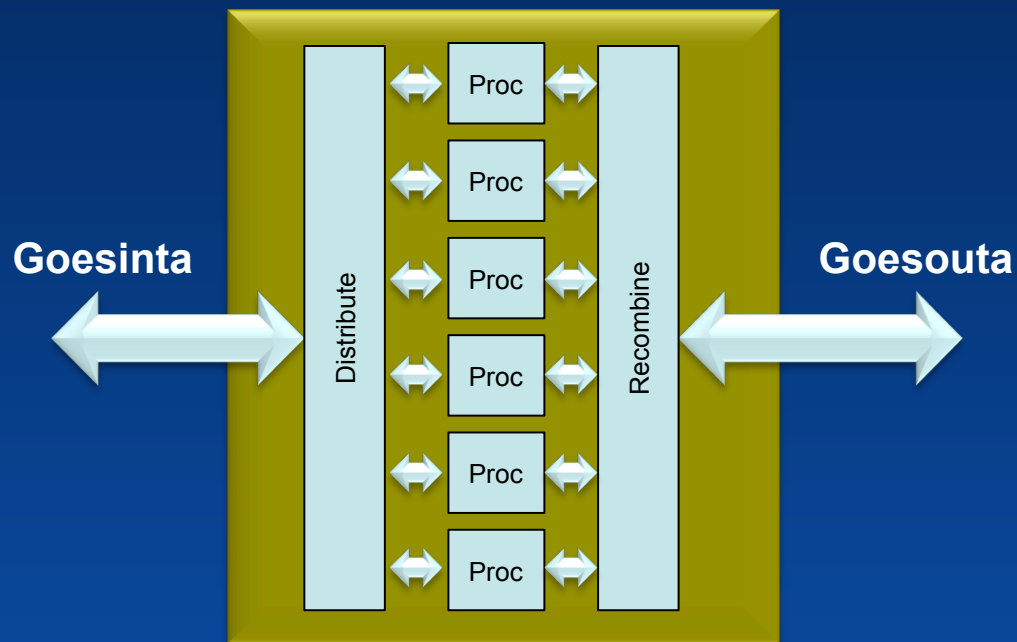


Addition of AI Accelerator

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



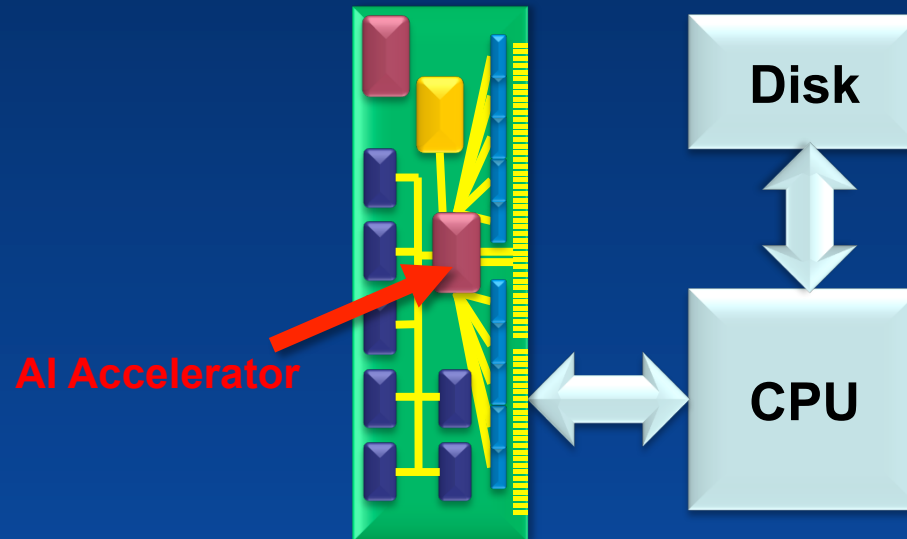
Flash Memory Summit



# In-Memory Computing

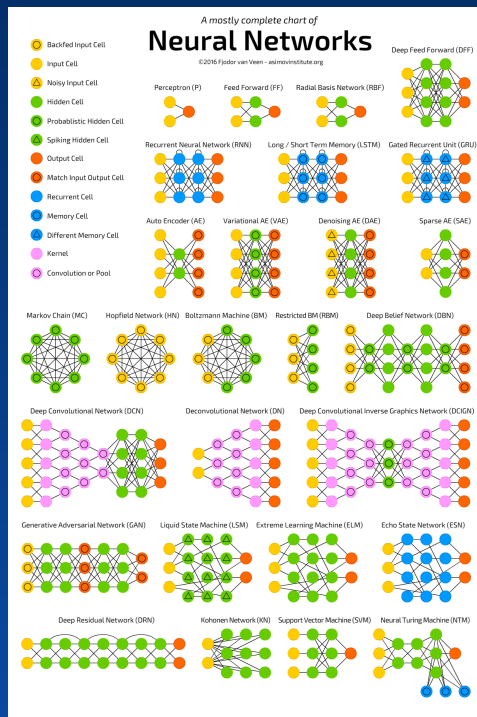
In-memory computing lets the AI accelerator control the memory directly

Also great for encryption





# Data Processing Paradigms



Traditional database

Data mining

Inferencing

Fuzzy logic

Recognition

etc





Flash Memory Summit

# The Actualization Gap



**Research projects**

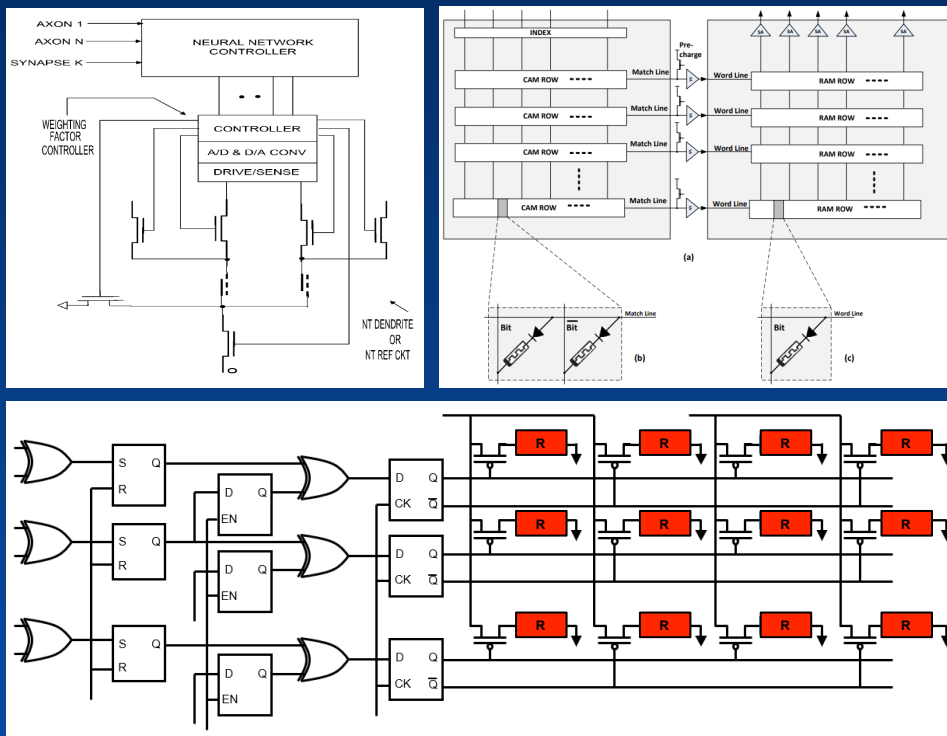


**Deployments**



Flash Memory Summit

# The "Research" Projects



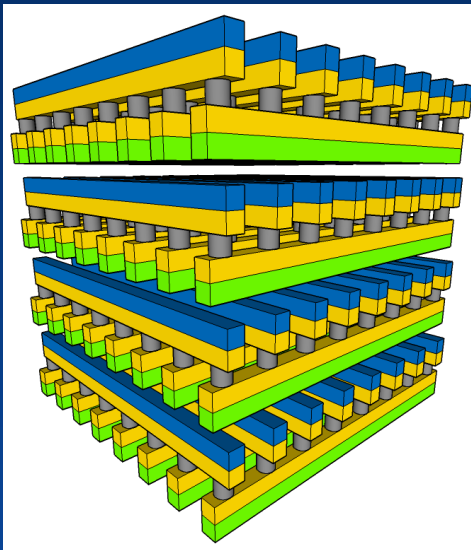
Many interconnects between storage elements and processing elements

Weighted calculations produce parallel possible results

Focus for a number of startup companies

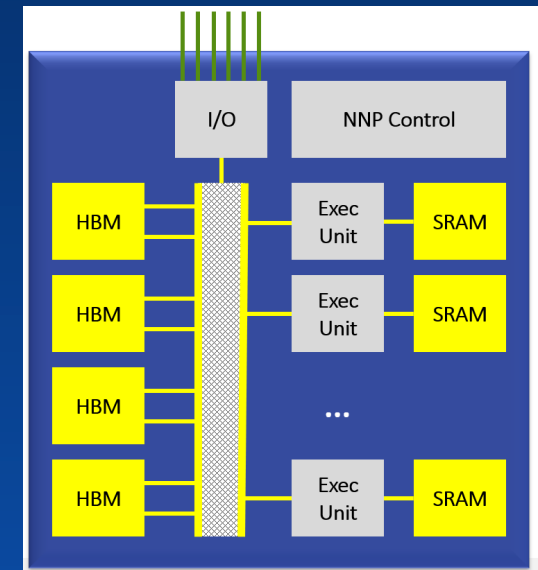


# What Most People Mostly Building



Dense matrix memory  
for highest storage  
capacity

Pipes for  
networking



Shared memory  
controller for many  
execution units

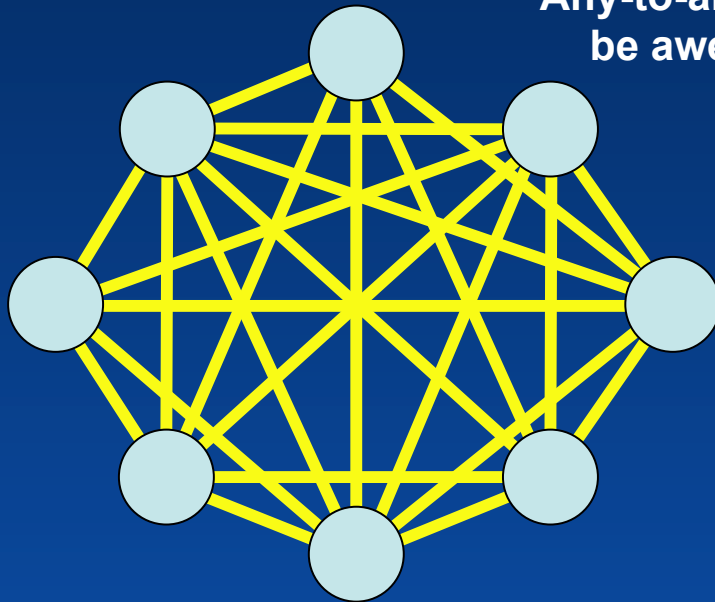


Flash Memory Summit

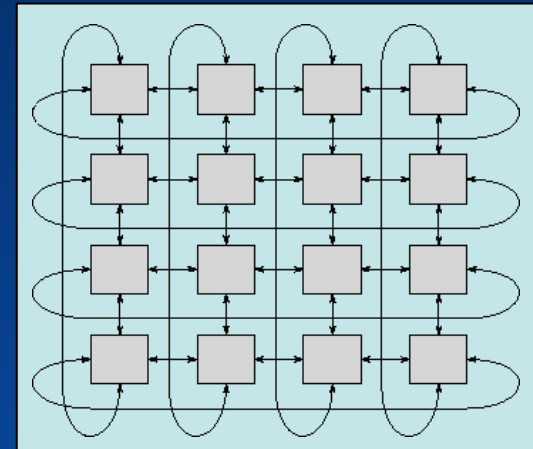


# Practical I/O Connection Limits

Any-to-any would be awesome



Toroid is a more practicable solution

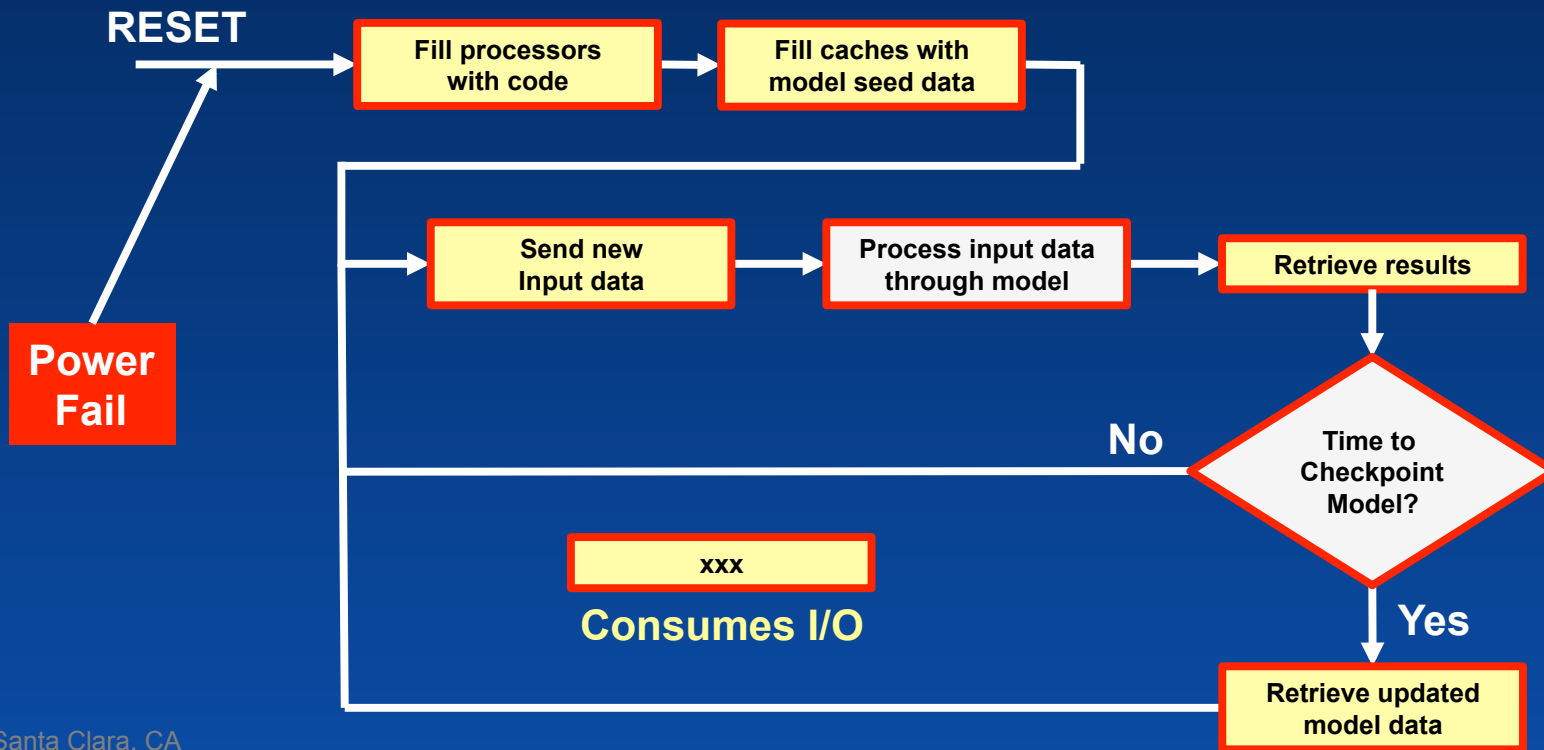


Limits how quickly data can flow in and out





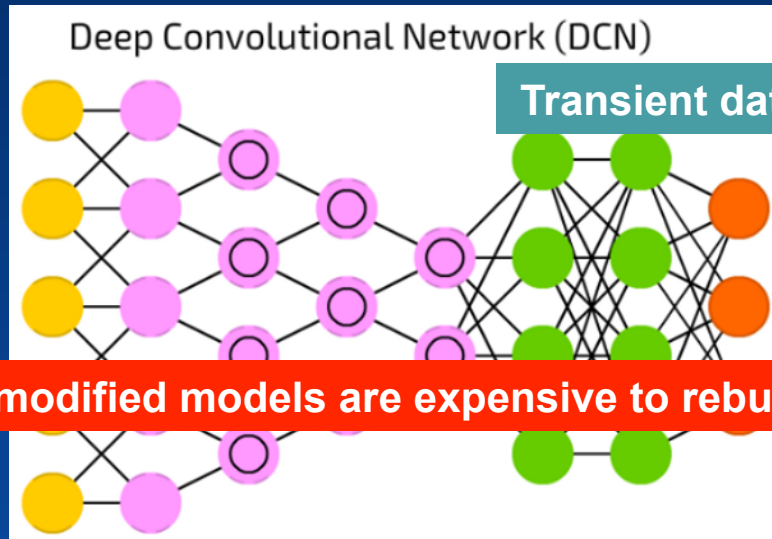
# Network Utilization





# 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



Flash Memory Summit

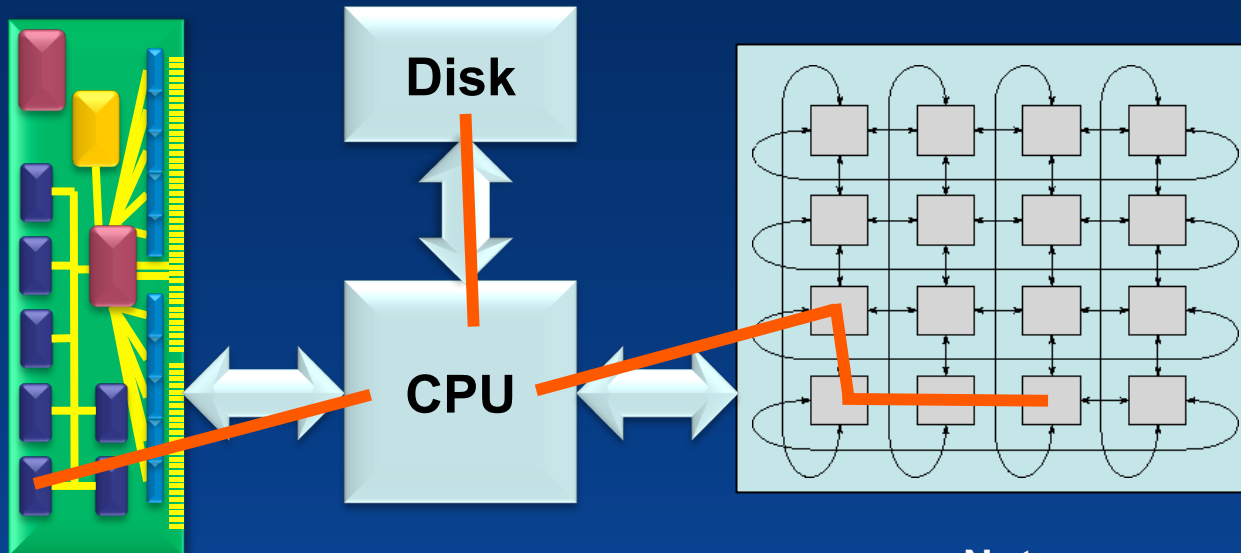


# Recovering From Power Fail

Data pulled from  
main memory

...or worse...

Backing store



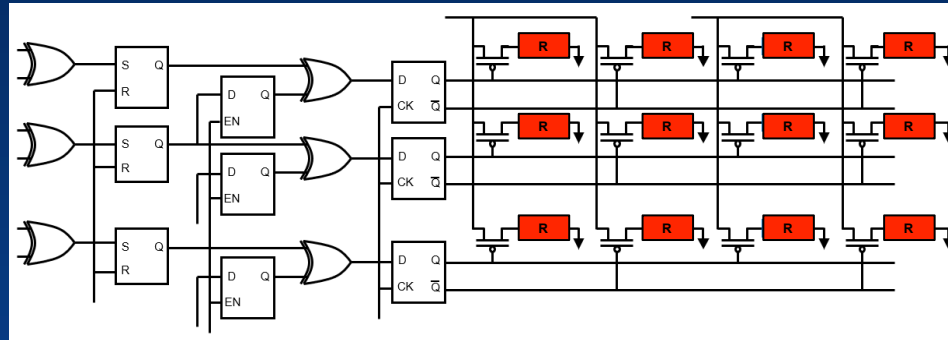
Data requires  
multiple hops  
through the  
interconnects

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**



Flash Memory Summit

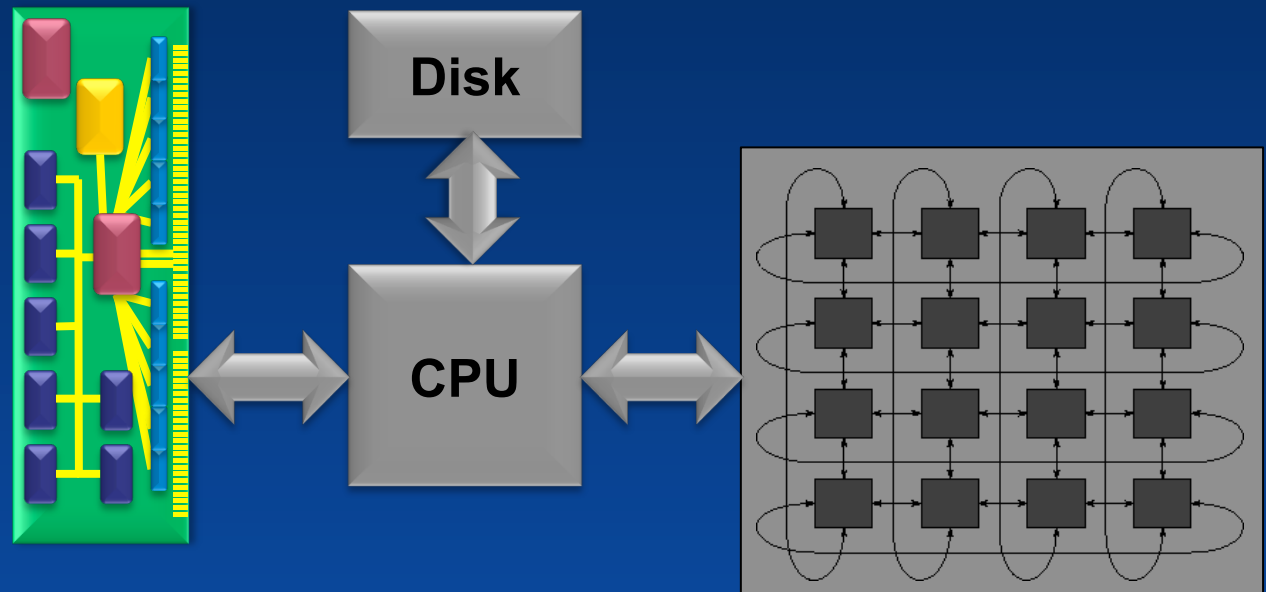


# Persistent Main Memory

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**

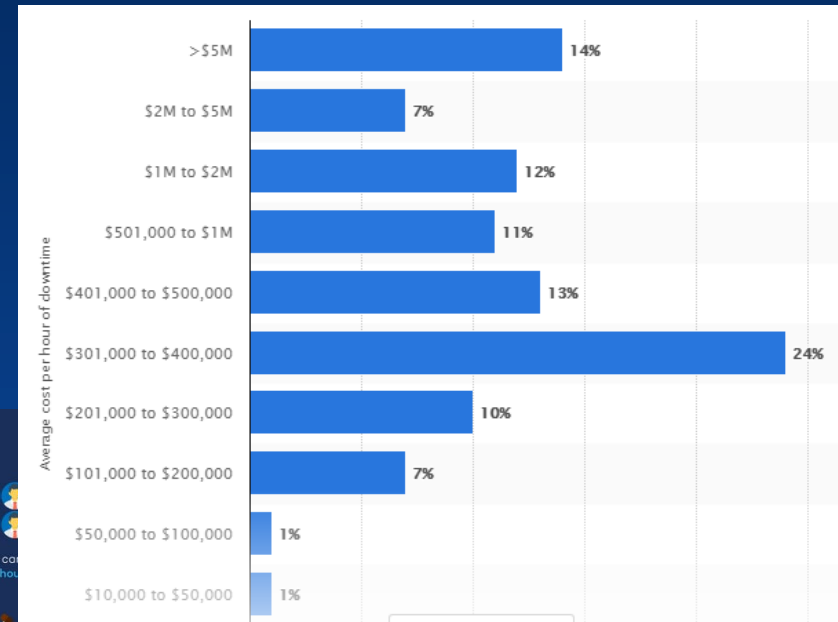
Source: Standish Group

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

**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.



## The Cost Of Internet Outages Beyond Revenue

The average downtime in the U.S. is 79 hours per outage.

Internet failure can cost a company 1 million person-hours per year.

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**



Flash Memory Summit



# Persistent Memory



## DRAM

Loses data  
Must be refreshed  
Can't lose power



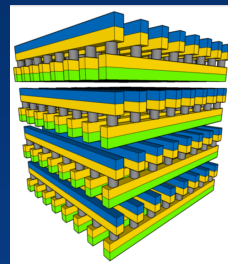
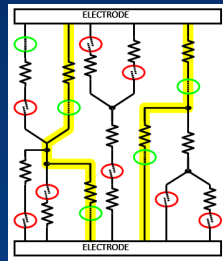
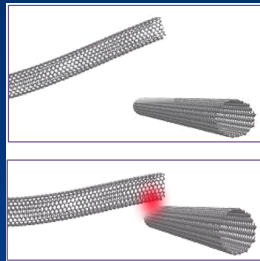
Persistent Memory  
Holds data  
forever, even  
on power fail

Santa Clara, CA  
August 2018



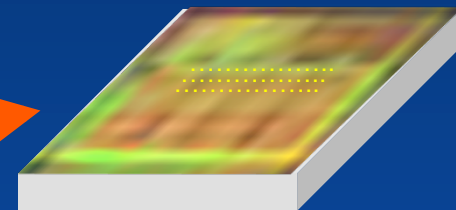
Flash Memory Summit

# Nantero NRAM™



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

DDR4  
DDR5



HBM

See my other talks  
later this week





# Classes of Persistent Memory

Non-volatility
Endurance
Read Time
Write Time

DRAM	NRAM
No	Yes
No limit	No limit
10 ns	10 ns
10 ns	10 ns

**Memory & Memory Class Storage**

MRAM	ReRAM	PCM / 3DXpoint	FeRAM
Yes	Yes	Yes	Yes
Limited	Limited	Limited	No limit
X	X	X	X
X	X	X	X

**Storage Class Memory**

Flash
Yes
10 <sup>3</sup>
50M ns
25M ns

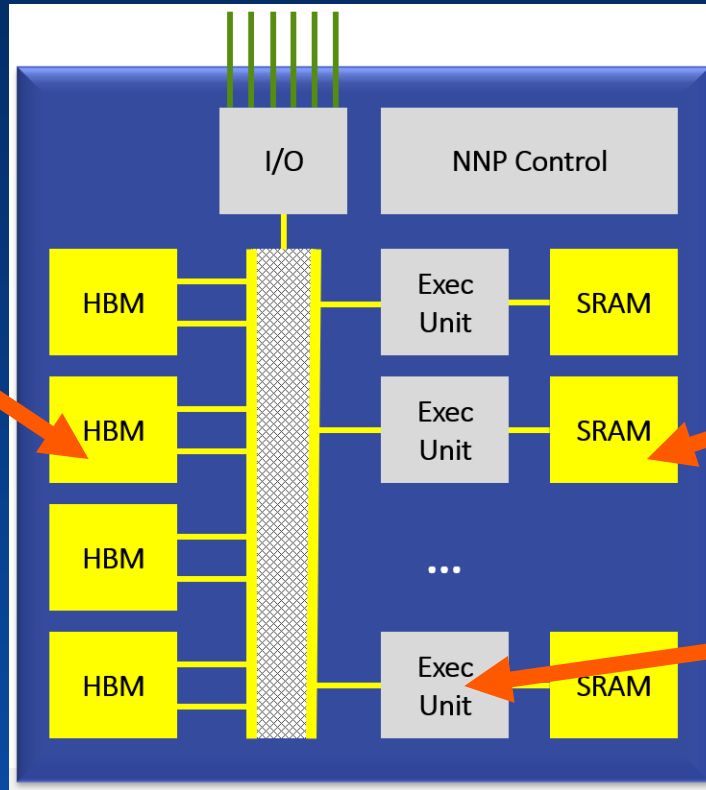
**Storage**



# 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

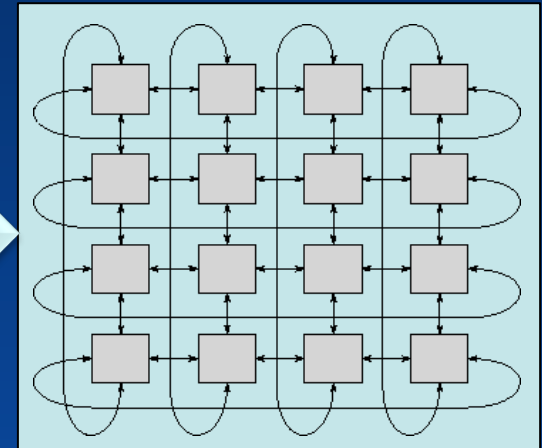
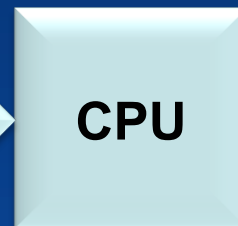
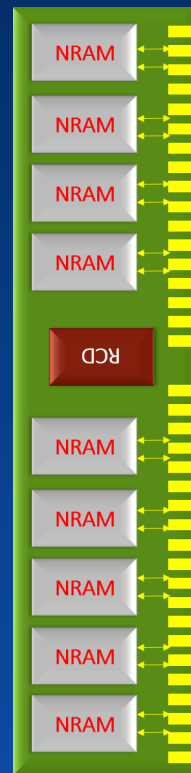


Flash Memory Summit



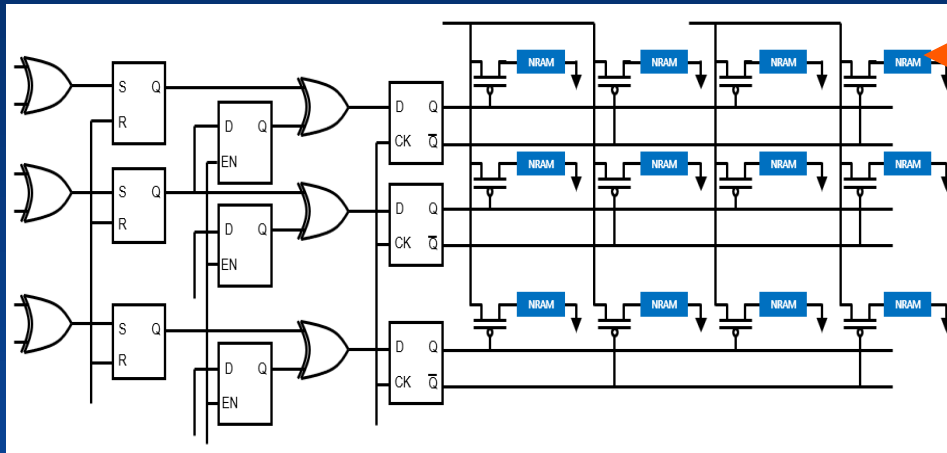
# NRAM for Main Memory

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



Flash Memory Summit



# NRAM Everywhere

Soon we will look back and say

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

and laugh

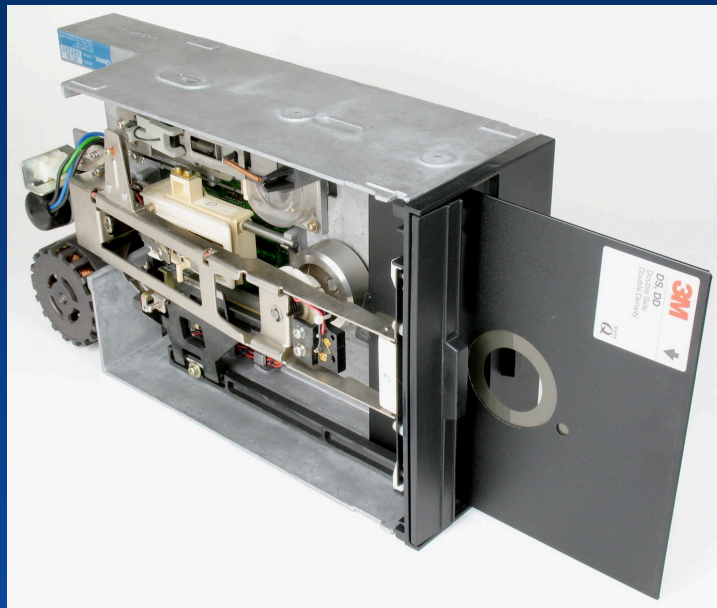


Santa Clara, CA  
August 2018



Flash Memory Summit

# 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](mailto:bilge@Nantero.com)