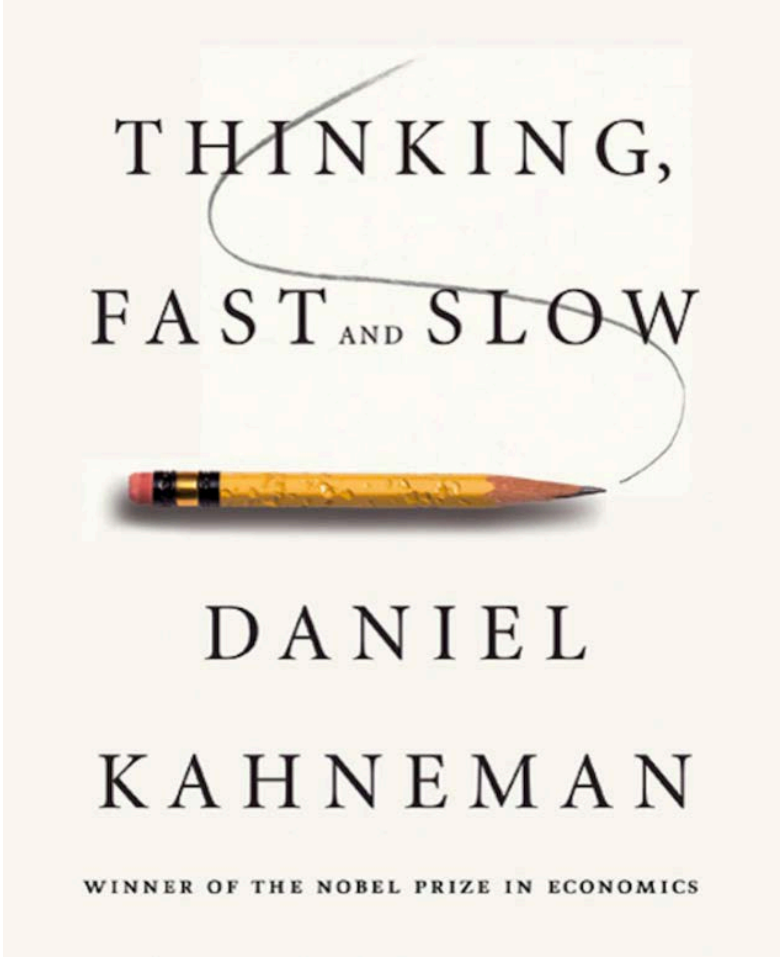
NVM Usage in the AI Era

Dave Eggleston

Intuitive Cognition Consulting

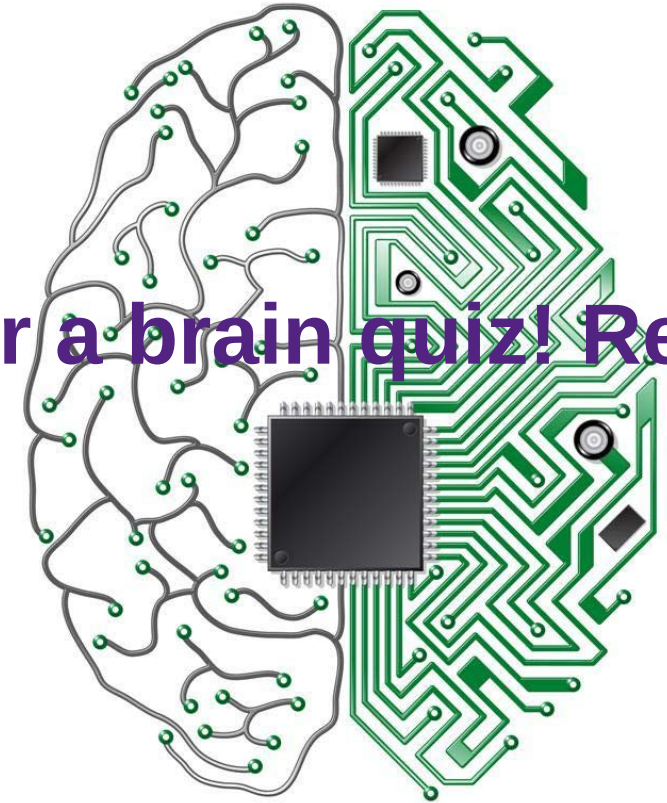


"A groundbreaking tour of the mind, and explains the two systems that drive the way we think."

"System 1 is fast, intuitive, and emotional; System 2 is slower, more deliberative, and more logical."

Daniel Kahneman is professor emeritus of psychology and public affairs at Princeton University.

Time for a brain quiz! Ready?



17 x 24 = ?

Intuition

17 x 24 = ?

Reasoning

Intuition

System 1

- Lightning fast
- Automatic
- Real time
- Effortless
- Approximate

Edge

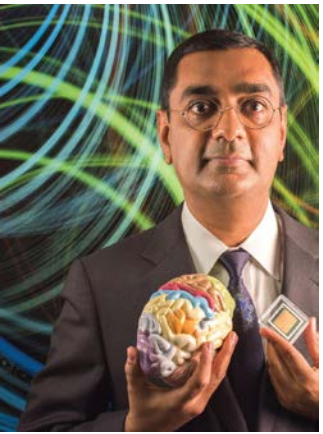
Reasoning

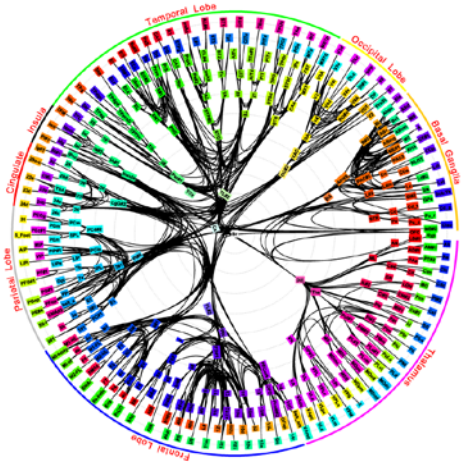
System 2

- Slow
- Interrupt driven
- Background
- Energy inefficient
- Precise

Datacenter

Edge

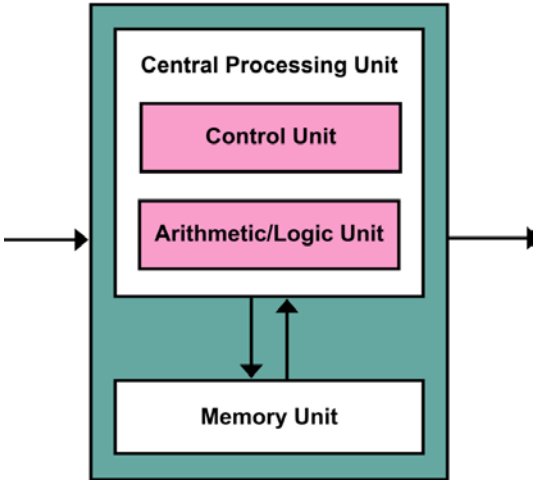




- Non-von Neumann architecture

Inference

Datacenter



- von Neumann architecture

Inference **Training**

AI Bottlenecks, Wants, and Solutions

Flash Memory Summit

Edge Inference	Datacenter Inference	Datacenter Training
		
MCU Bottleneck	Memory Bottleneck	I/O Bottleneck
Ultra low power & Speed	Parallelism	Fast & Coherent Checkpointing
NVM?	NVM?	NVM?

AI Bottlenecks, Wants, and Solutions

Edge Inference		DATACENTER Inference		DATACENTER TRAINING	
MCU Bottleneck		Memory Bottleneck		I/O Bottleneck	
Ultra low power & Speed		Parallelism		Fast & Coherent Checkpointing	
NVM?		NVM?		NVM?	

AI Bottlenecks, Wants, and Solutions

Edge Inference		Datacenter Inference		Datacenter Training	
MCU Bottleneck		Memory Bottleneck		I/O Bottleneck	
Ultra low power & Speed		Parallelism		Fast & Coherent Checkpointing	
NVM?		NVM?		NVM?	

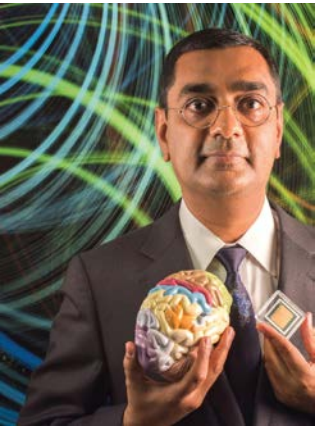
Flash Memory Summit

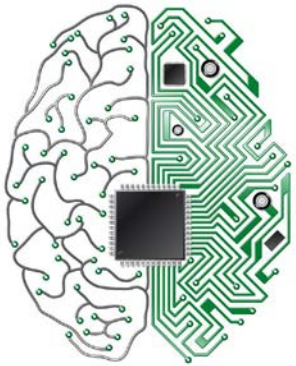
Edge Inference

MCU Bottleneck

Ultra low power & Speed

Analog NVM-based MAC acceleration



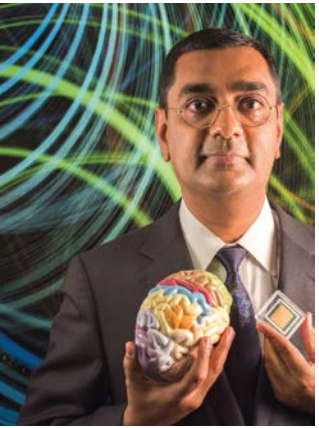


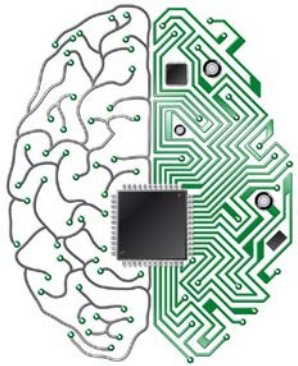
- Brain operates on <20 Watts
- Von Neumann inference >20 GigaWatts!
- Want non-von Neumann architecture (low power)
- Want real-time inference (speed)
- MCU lacks matrix math capability
- Want a MAC accelerator to do matrix math
- Use analog NVM for MAC acceleration!

Flash Memory Summit

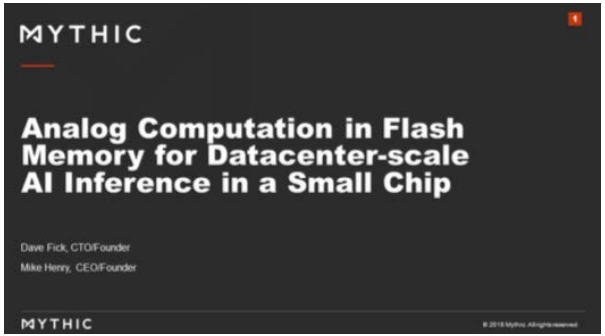
Edge Inference

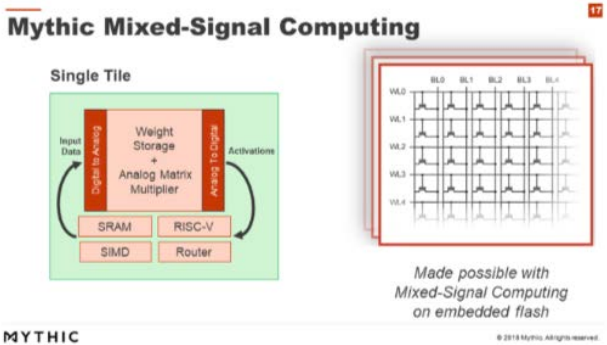
MCU Bottleneck
Ultra low power & Speed
Analog NVM-based MAC acceleration





- Brain operates on <20 Watts
- Von Neumann inference >20GigaWatts!
- Must have non-von Neumann architecture
- Want real-time inference (speed)
- MCU lacks matrix math capability
- Want a MAC accelerator to do matrix math
- Use analog NVM for MAC acceleration!





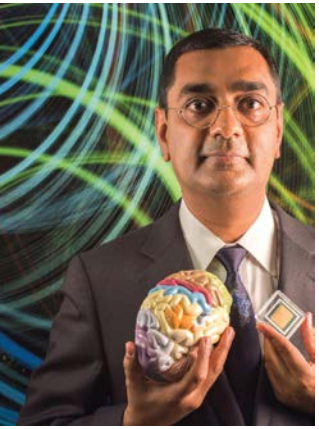
Flash Memory Summit

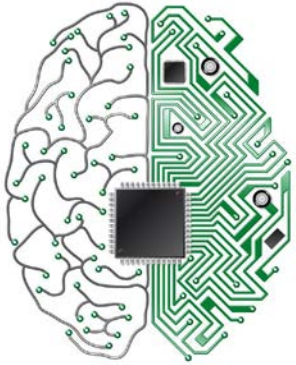
Edge Inference

MCU Bottleneck

Ultra low power & Speed

Analog NVM-based MAC acceleration



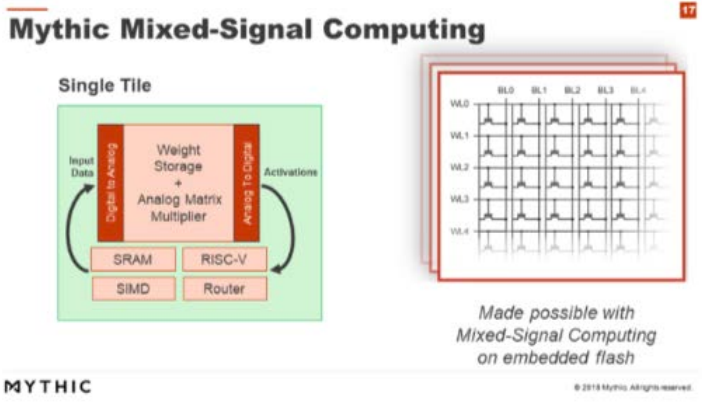


- Brain operates on <20 Watts
- Von Neumann inference >20GigaWatts!
- Must have non-von Neumann architecture
- Want real-time inference (speed)
- MCU lacks matrix math capability
- Want a MAC accelerator to do matrix math
- Use analog NVM for MAC acceleration!

SYNTIANT

ANAFLASH

MICROCHIP



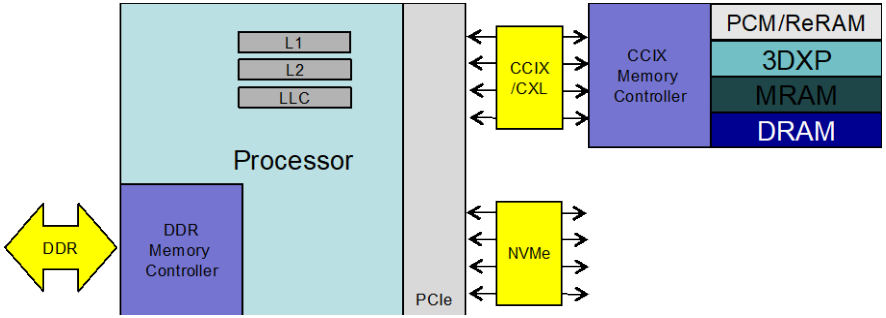
Datacenter Inference

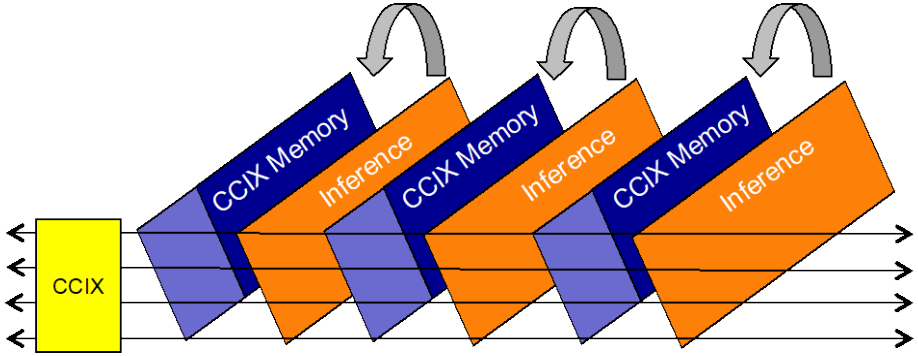
Memory Bottleneck

Parallelism

CCIX Memory with Inference engines

- Hold trained model in memory
- Inference bottlenecked by CPU-DDR
- Want parallelism for speed
- Multiple sets of CCIX Memory and Inference engines
- CCIX enables load/store peer-peer sharing
- Avoid CPU-DDR bottleneck
- Use DRAM or NVM based CCIX Memory!





Flash Memory Summit

Datacenter Training

I/O Bottleneck

Fast & Coherent Checkpointing

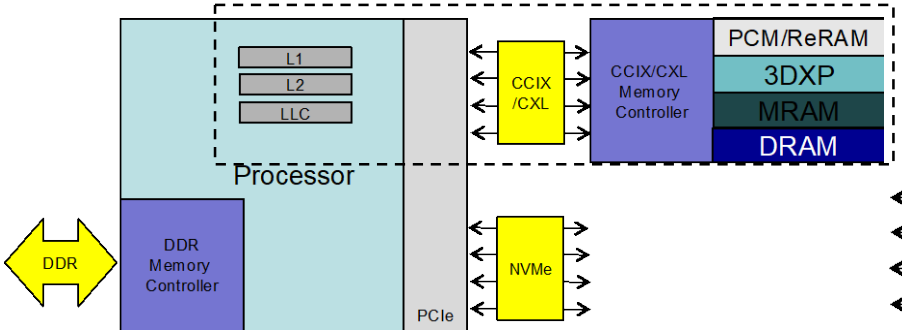
CCIX/CXL coherent NVM expansion

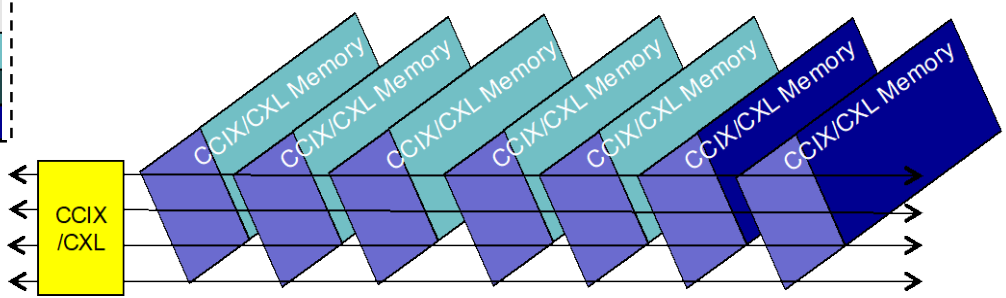
CXL Compute Express Link

CCIX

- Want more memory
- Want fast & coherent NVM checkpointing
- Want fast rebuild time
- Avoid 5us+ latency on I/O
- Avoid using precious DDR slots
- Coherent memory expansion on CCIX/CXL!
- Mix NVM and DRAM on CCIX/CXL!

Cache coherent memory expansion





AI Bottlenecks, Wants, and Solutions

Edge Inference		Datacenter Inference		Datacenter Training	
MCU Bottleneck		Memory Bottleneck		I/O Bottleneck	
Ultra low power & Speed		Parallelism		Fast & Coherent Checkpointing	
Analog NVM based MAC acceleration		CCIX Memory with Inference engines		CCIX/CXL coherent NVM expansion	

Now on to the AI NVM Experts!

Pyxeda

MICROCHIP

ANAFLASH

SYNTIANT

DPU

Qualcomm

Dave Eggleston
Intuitive Cognition Consulting
Technology & Business Strategy

Email: dave@in-cog.com
Twitter: @NVM_DaveE
LinkedIn:
linkedin.com/in/deggleston/