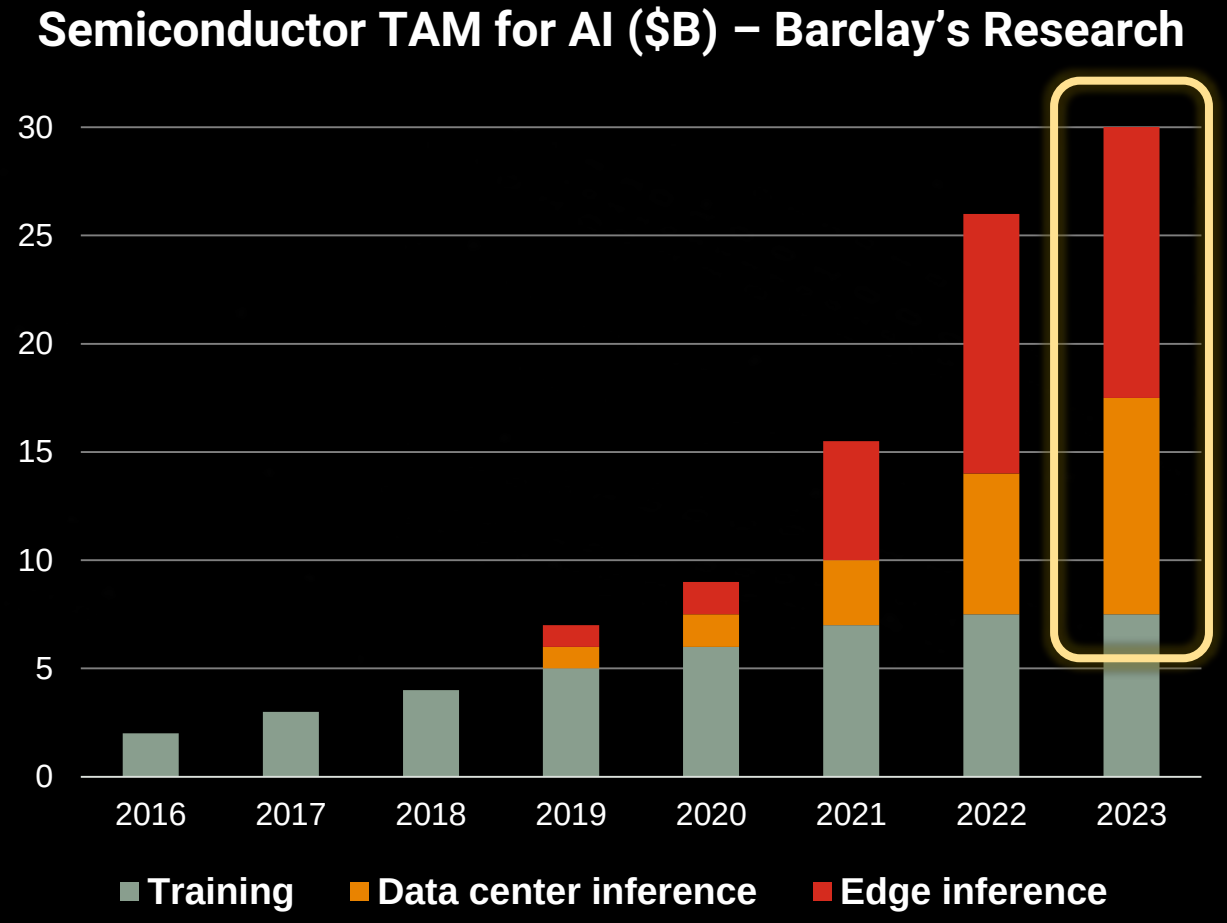
MYTHIC

Analog Compute-in-Memory for Inference

Mike Henry, CEO & Founder

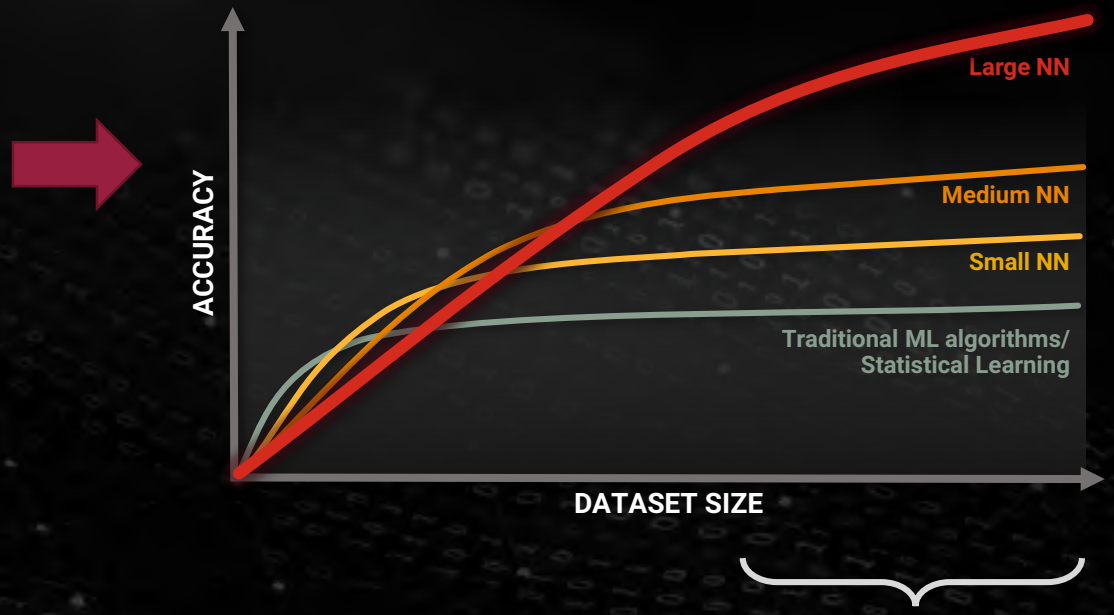
MYTHIC

Large Growth Opportunity for AI Inference



Tough Requirements for Inference

- Accuracy requirements dictate many millions of weights per model
- Trillions of operations per second
- Real-time, low-latency, with high levels of concurrency
- Affordable, scalable, low-power package

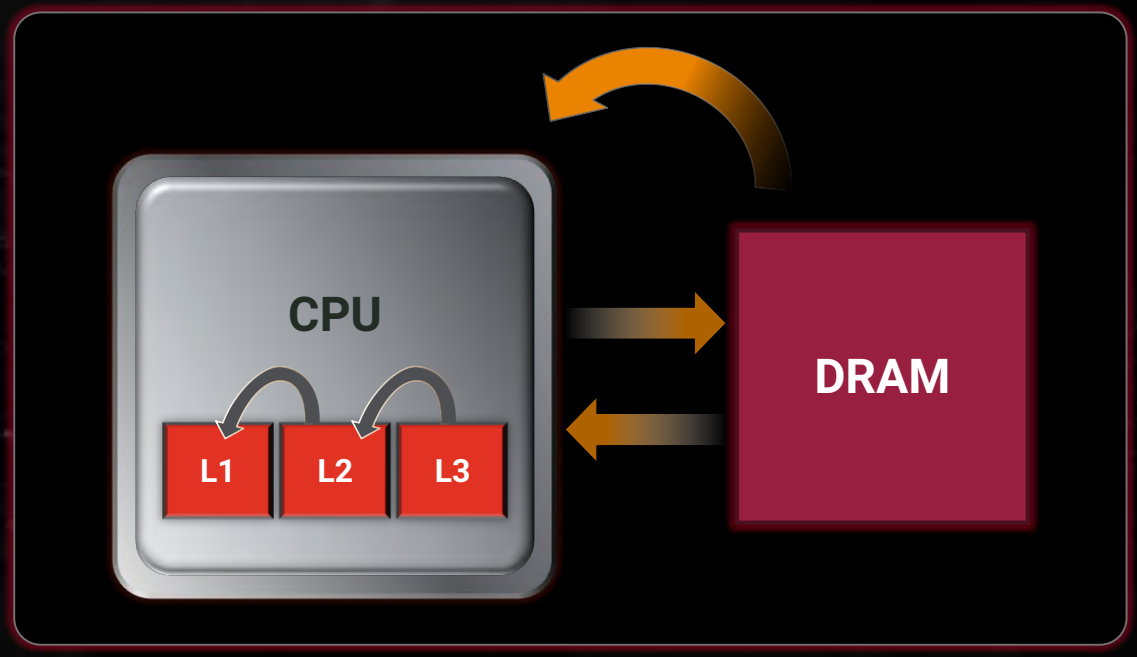


MYTHIC

On-Chip Compute-and-Memory

Principles of On-Chip Compute-and-Memory

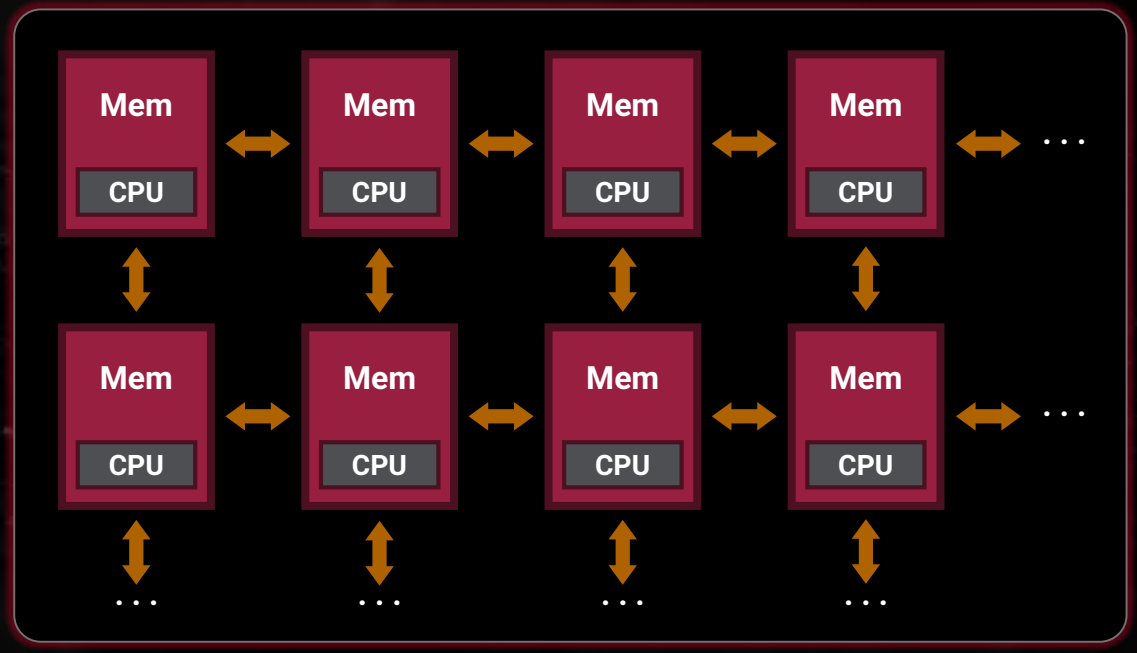
- Minimize journey from memory to processing units
- Many parallel compute units to maximize throughput and minimize latency
- Maximize memory bandwidth for entire set of DNNs



Traditional Architecture

Principles of On-Chip Compute-and-Memory

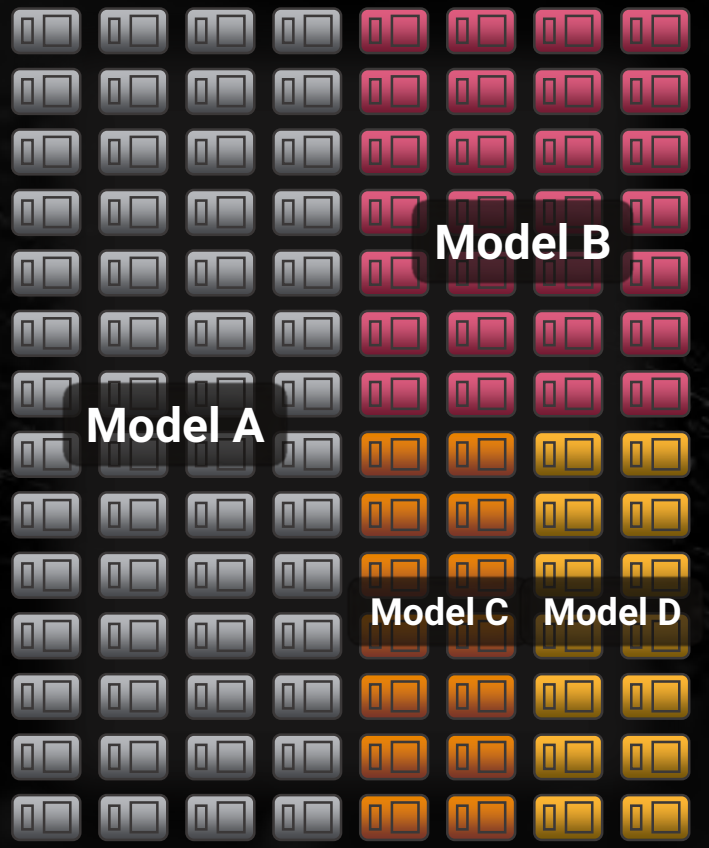
- Minimize journey from memory to processing units
- Many parallel compute units to maximize throughput and minimize latency
- Maximize memory bandwidth for entire set of DNNs



Compute-and-Memory Architecture

Principles of On-Chip Compute-and-Memory

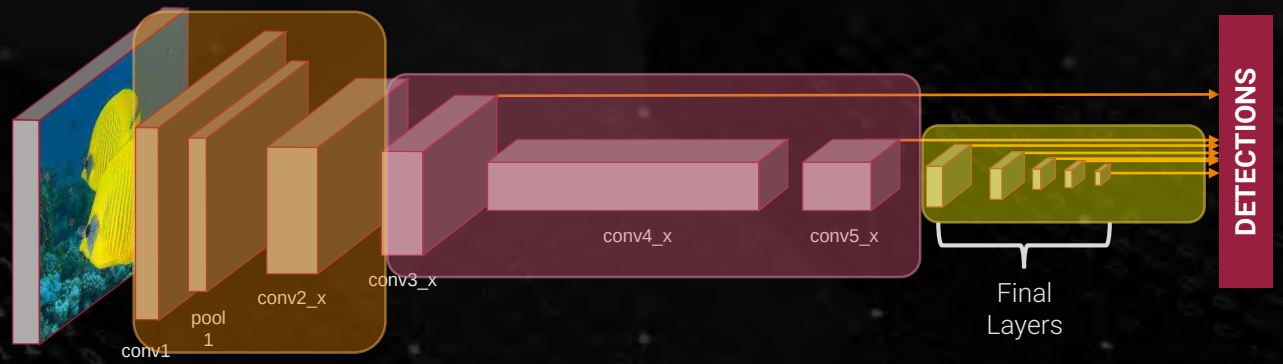
- Minimize journey from memory to processing units
- Many parallel compute units to maximize throughput and minimize latency
- Maximize memory bandwidth for entire set of DNNs



Weights are Stationary for Inference

MYTHIC

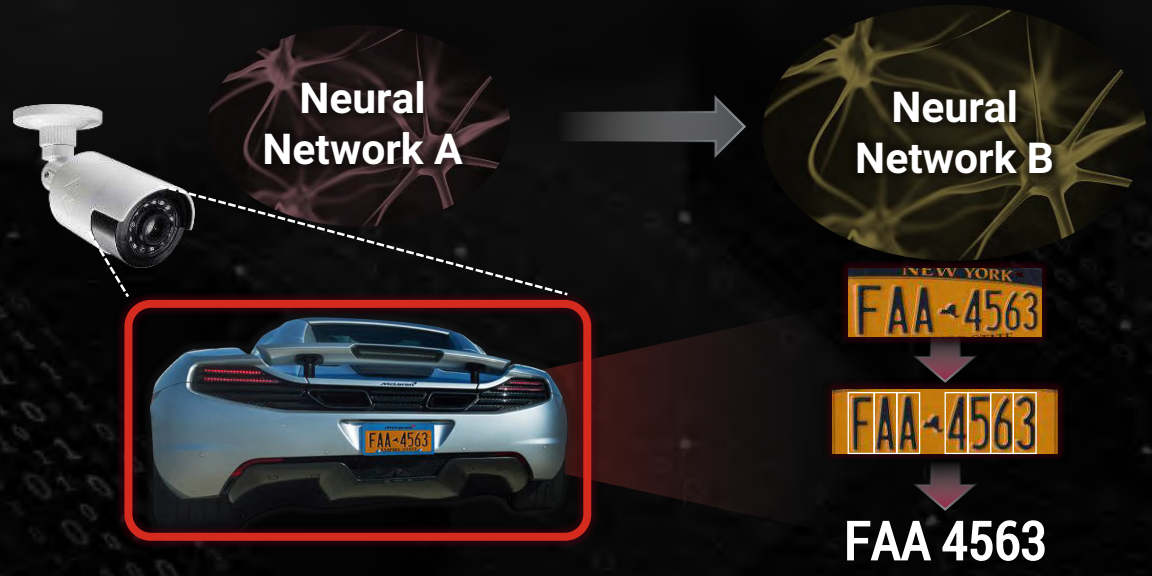
Innovate Faster and Easily Build Complex Applications

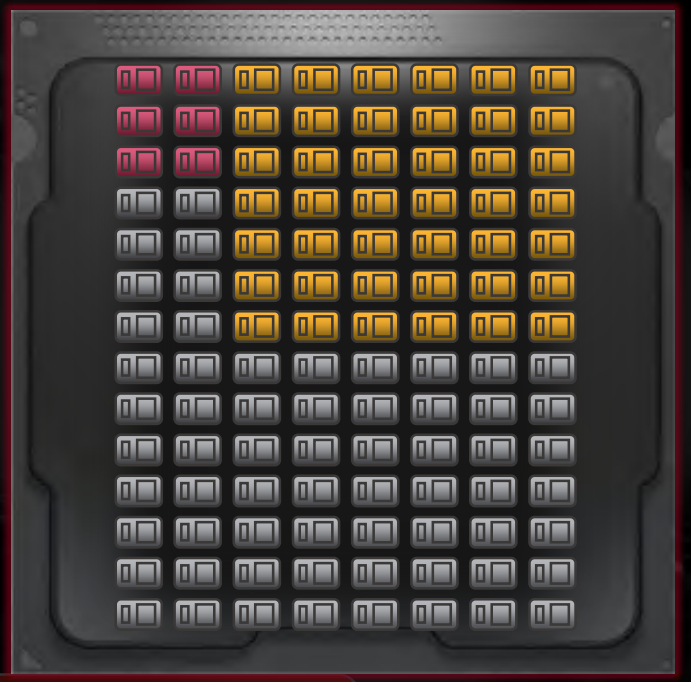


- Easily trade-off performance, power, latency, accuracy
- Enable concurrent and dynamic applications with deterministic execution
- Scalability from Very Large to Very Small

MYTHIC

Innovate Faster and Easily Build Complex Applications



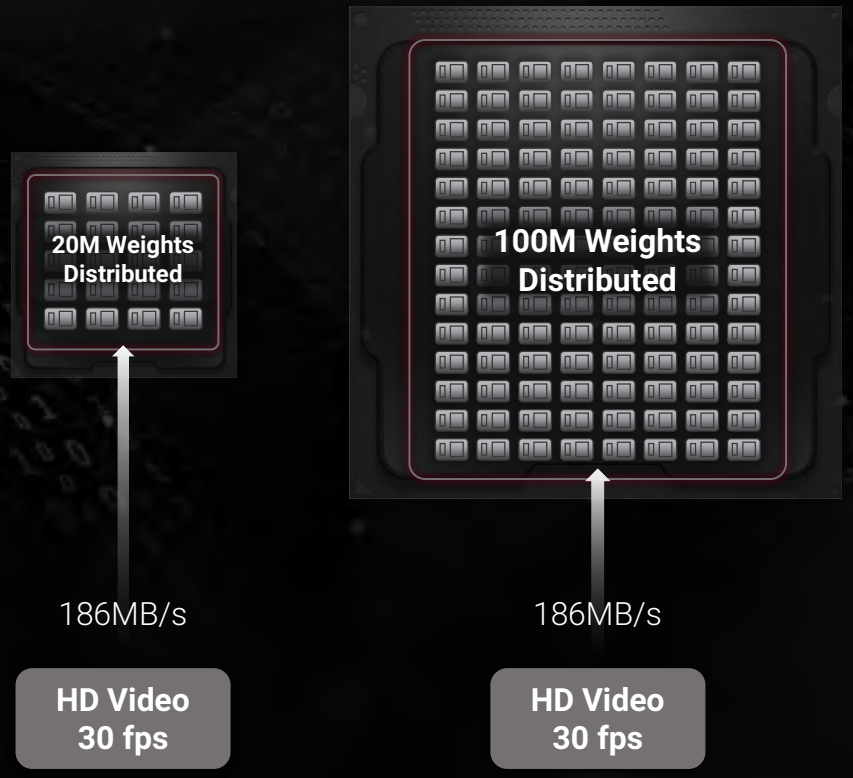


- Easily trade-off performance, power, latency, accuracy
- Enable concurrent and dynamic applications with deterministic execution
- Scalability from Very Large to Very Small

MYTHIC

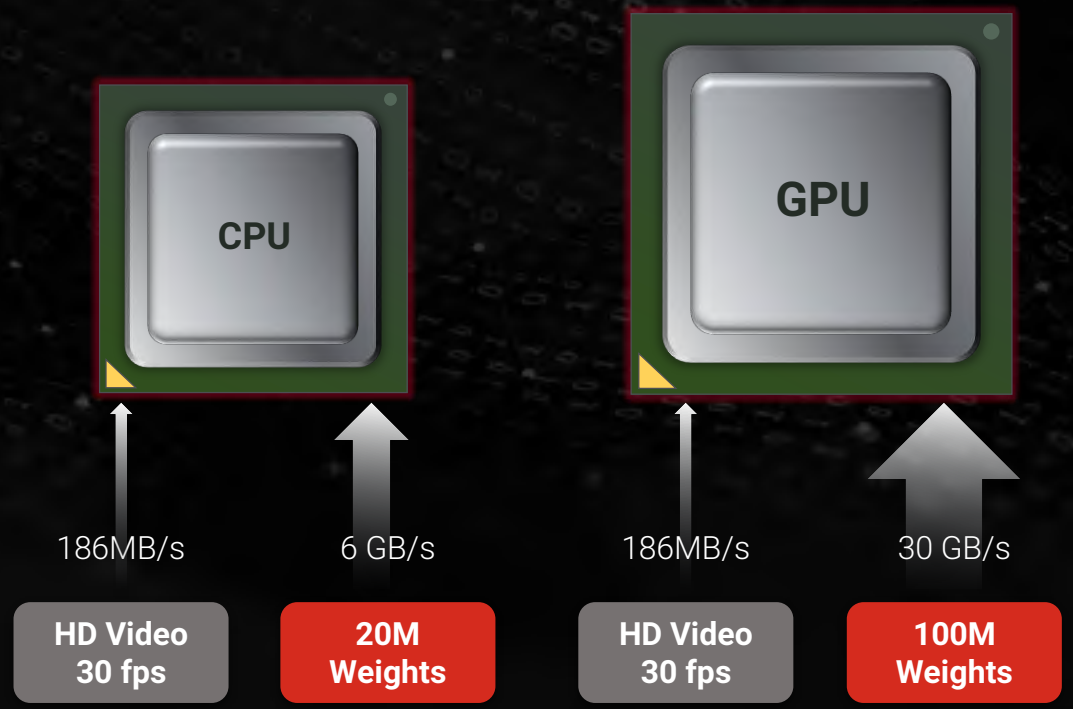
Scalability from Very Large to Very Small

On-Chip Stationary Weights = OK



Off-Chip Weights = Major Bottlenecks

Weight bandwidth = 10-20X data bandwidth



MYTHIC

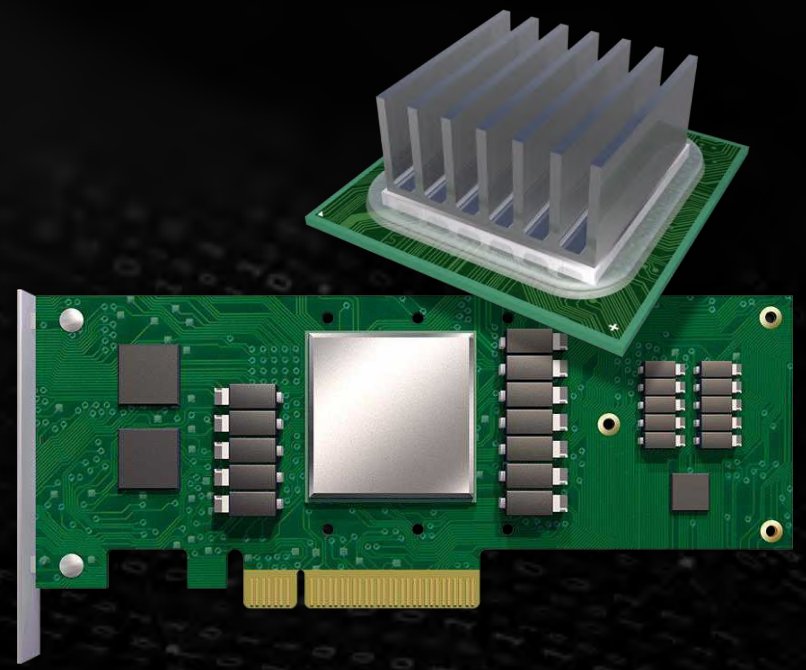
Best-in-Class Performance and Power for Inference

	Traditional CPU / GPU	Compute-and-Memory
Total model capacity	✓	
System Simplicity		✓
Low Latency		✓
Energy Efficiency		✓
High Performance		✓
Deterministic Execution		✓

MYTHIC

Implementation Challenges in Digital

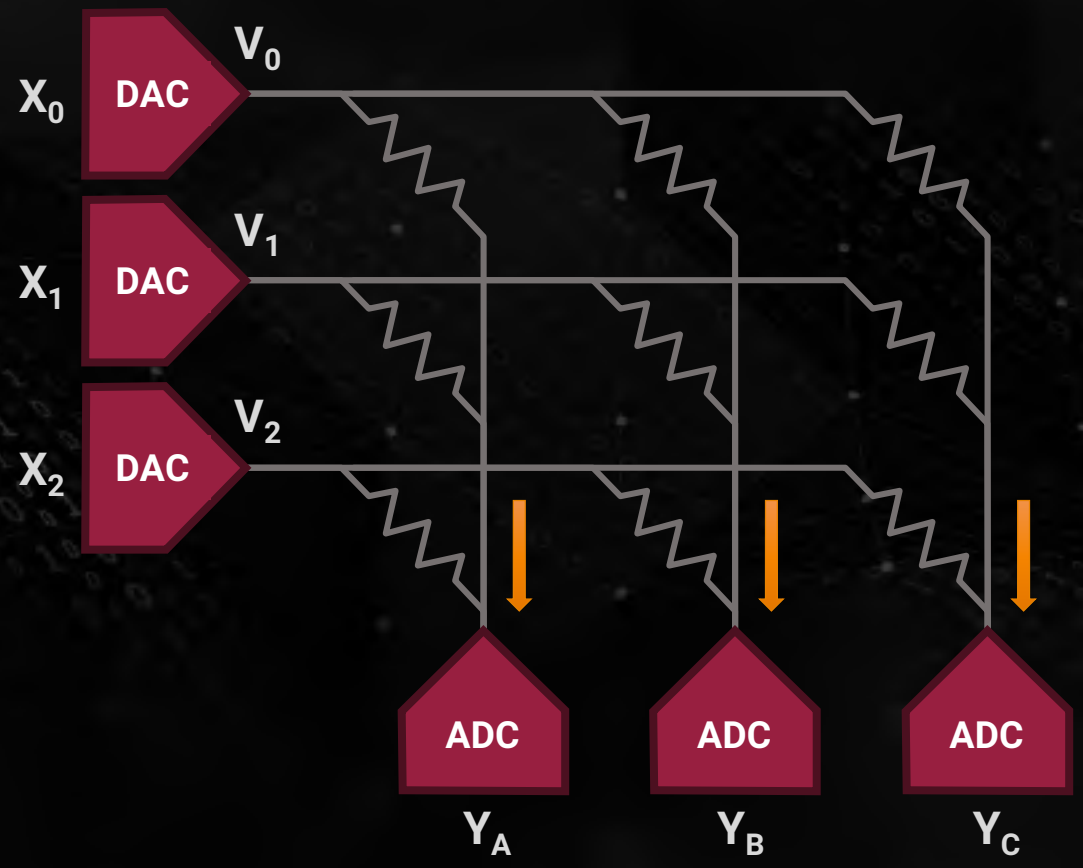
- True compute-in-memory not possible
- Large die-area from compute and memory
- High power-consumption
- High development costs due to state-of-the-art process nodes



MYTHIC

Enter Analog

Unique Analog Compute-in-Memory



- Mythic computes directly inside the flash memory array
  - Weights stored in flash
  - Inputs provided as voltages
  - Outputs collected as currents

- Result: High-performance, yet amazingly efficient processing

MYTHIC

Power, Performance, and Latency

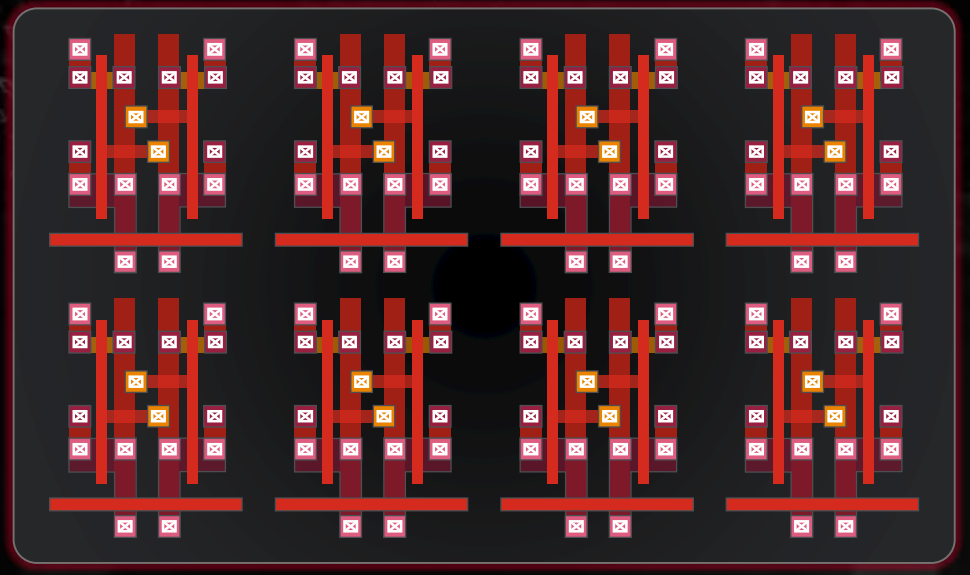
Incredibly high parallelization and “effective” memory bandwidth	→	1000 x 1000 vector-matrix product in a few microseconds
Energy efficiency of 4 TOPS/W for convolutional and dense networks	→	Accounts for all system aspects including processor control cores and PCI express

MYTHIC

Analog Compute Gives Us Model Capacity

- NOR Flash: Up to 50x denser weight storage
- No need to add more compute – the storage **is** the compute

SRAM = 48 transistors per 8-bit weight



Analog = 1 flash transistor per 8-bit weight

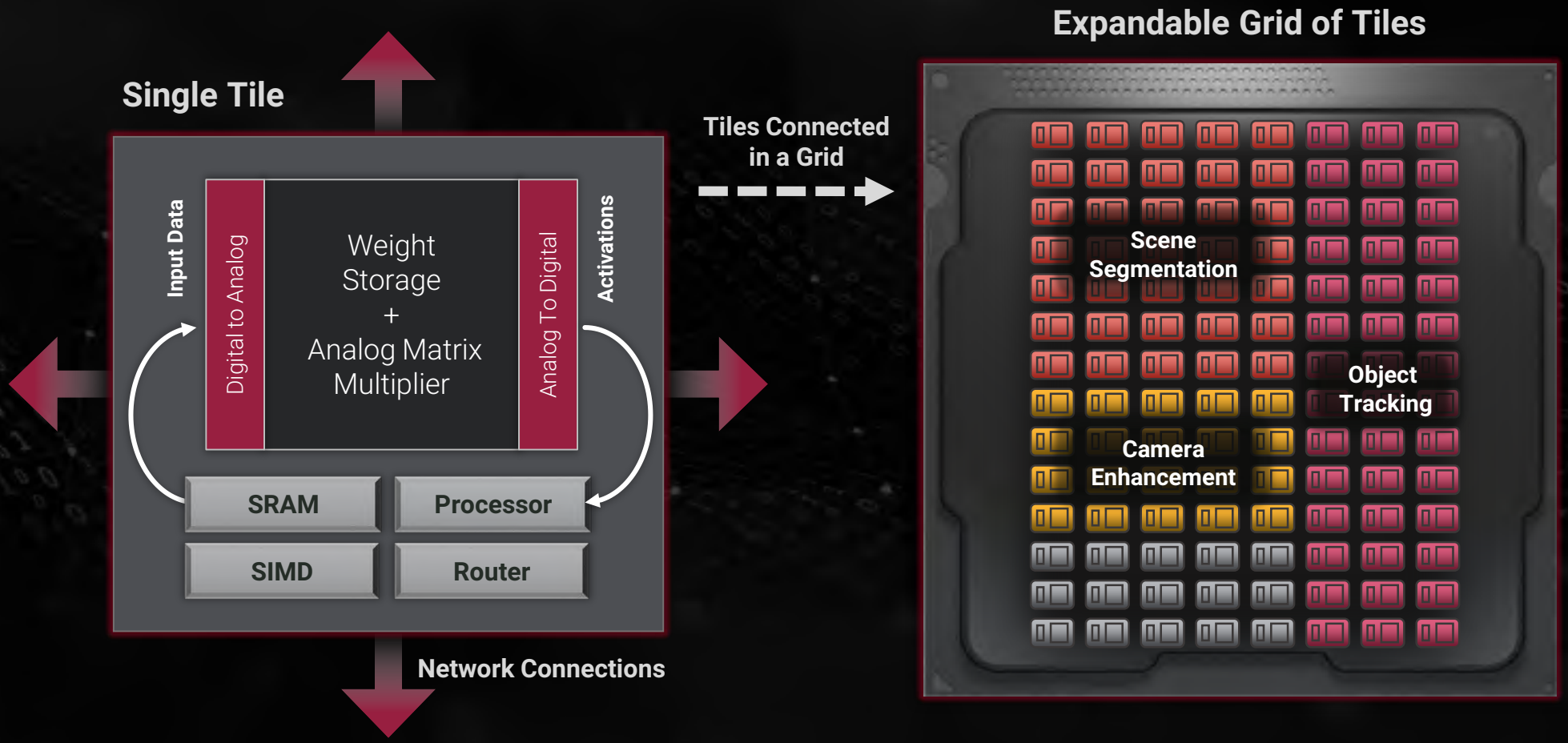
MYTHIC

Rich Roadmap for Analog Compute Improvements

- More on this later

MYTHIC

DNN Tile - Ultra-dense Storage + Matrix Multiplication



AI Inference Processor Landscape

Traditional Von-Neuman-like Architectures

SoC

Movidius

Ambarella™

cadence® (IP)

QUALCOMM®

CPU/GPU

intel Xeon inside™

NVIDIA

TPU

Google

habana

intel Nervana™

On-Chip Compute-and-Memory

Digital Compute-Near-Memory

cerebras

GRAPHCORE

HAILO

SYNTIANT

Analog Compute-In-Memory

MYTHIC

MYTHIC

The Mythic IPU

Mythic Initial Product Lineup

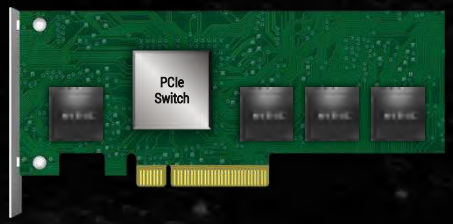
Mythic IPU

Up to 120M Weights

M.2 x4 2280

Mythic IPU x1

Mythic PCIe Cards



PCIe x4 HHHL

Mythic IPU x4

PCIe x16 FHFL

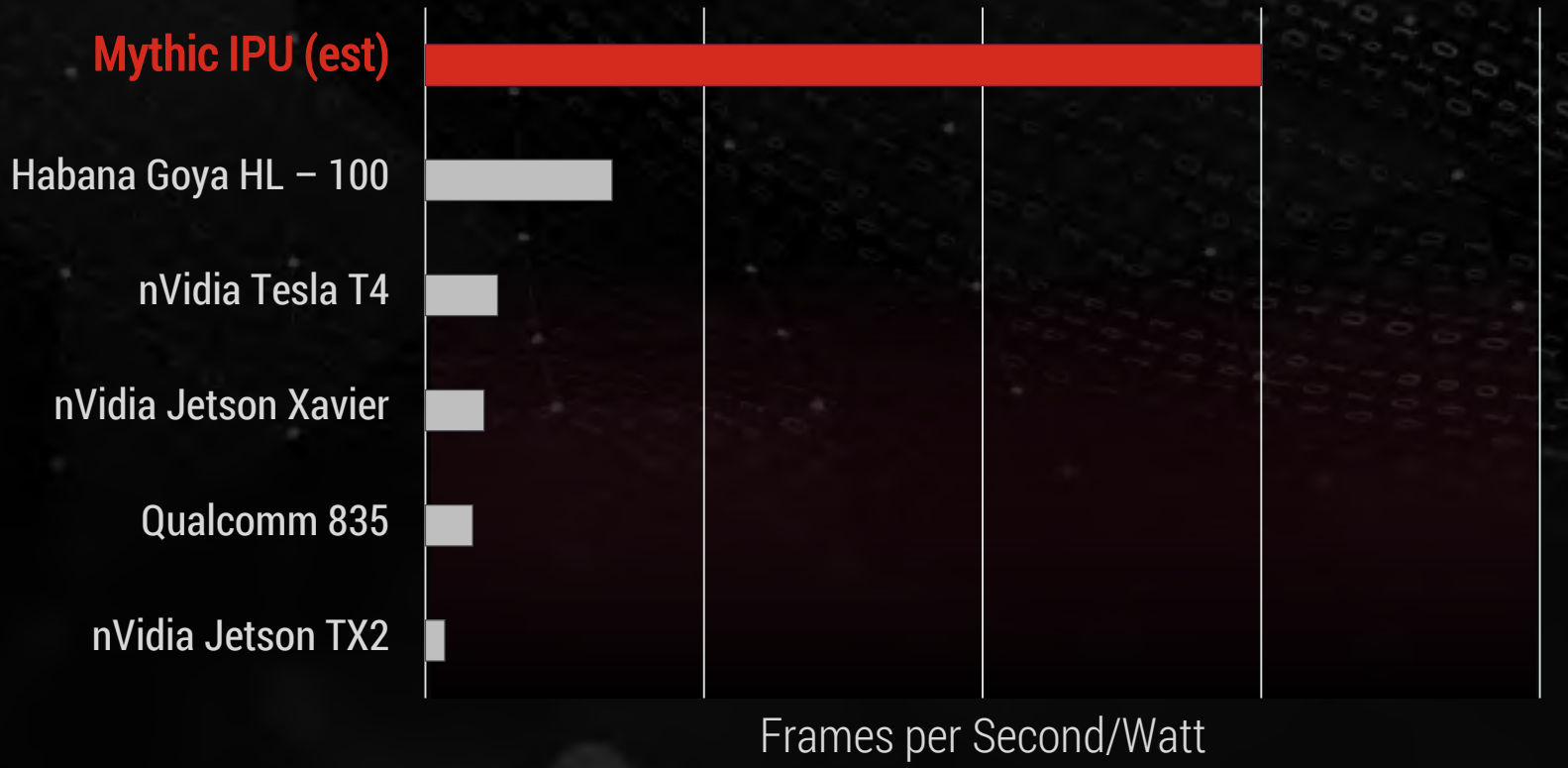
Mythic IPU x16

MYTHIC

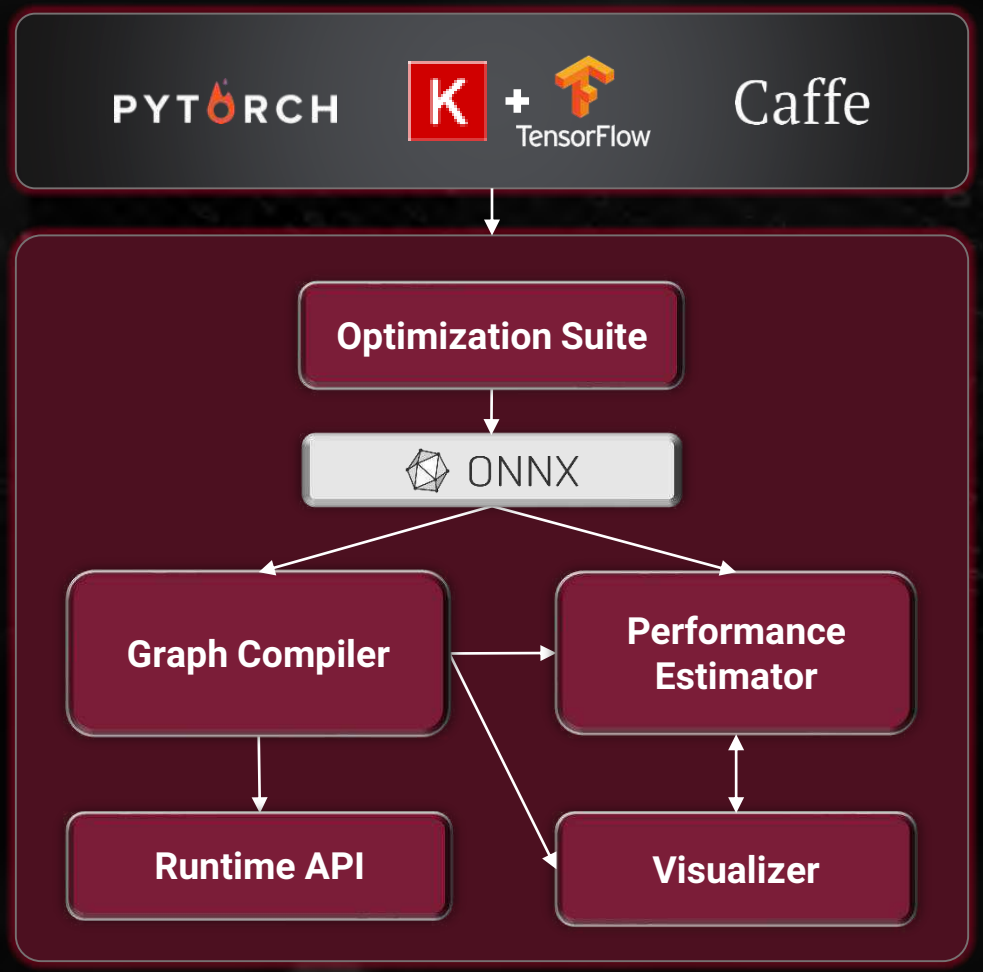
Low-Power DNN Inference Solution

Inference Capability per Watt for Common DNN Processors

ResNet-50, INT8, and Batch=1



Mythic SDK - Develop with Latest Networks / Frameworks



- Graph decomposition and mapping
- Code generation

- Host runtime API and OS drivers

- Post-training quantization
- Retraining libraries

- Annotated

- Power, speed and memory estimates

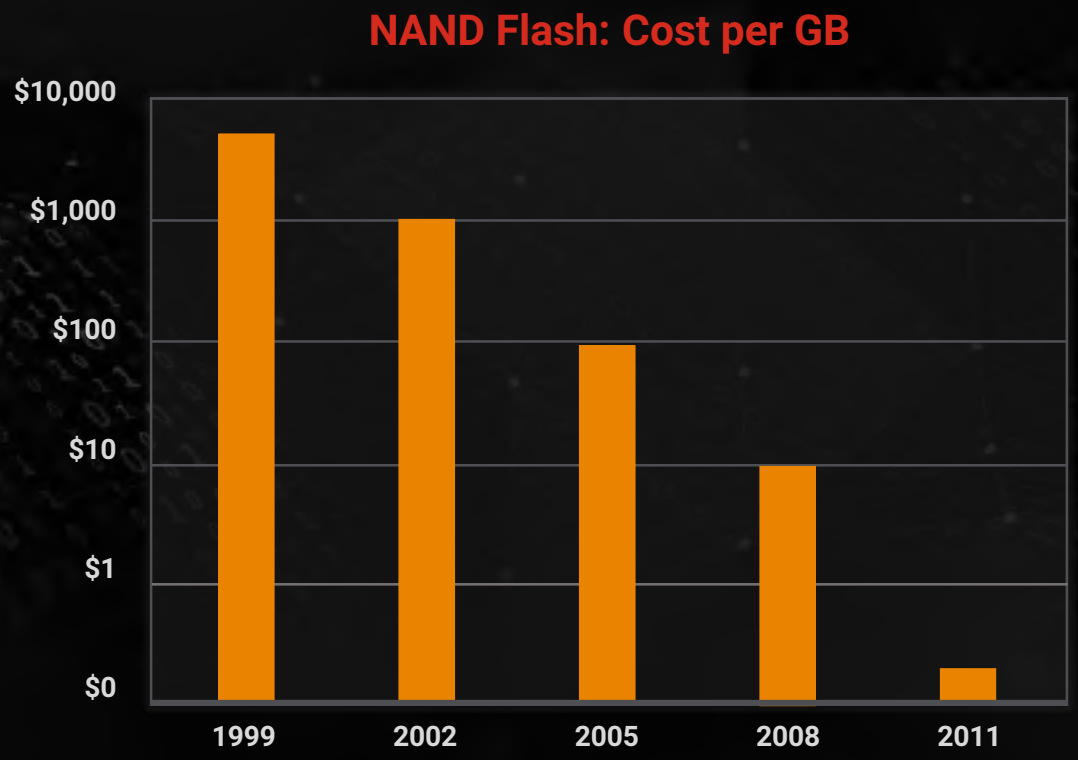
- Profiling and logging

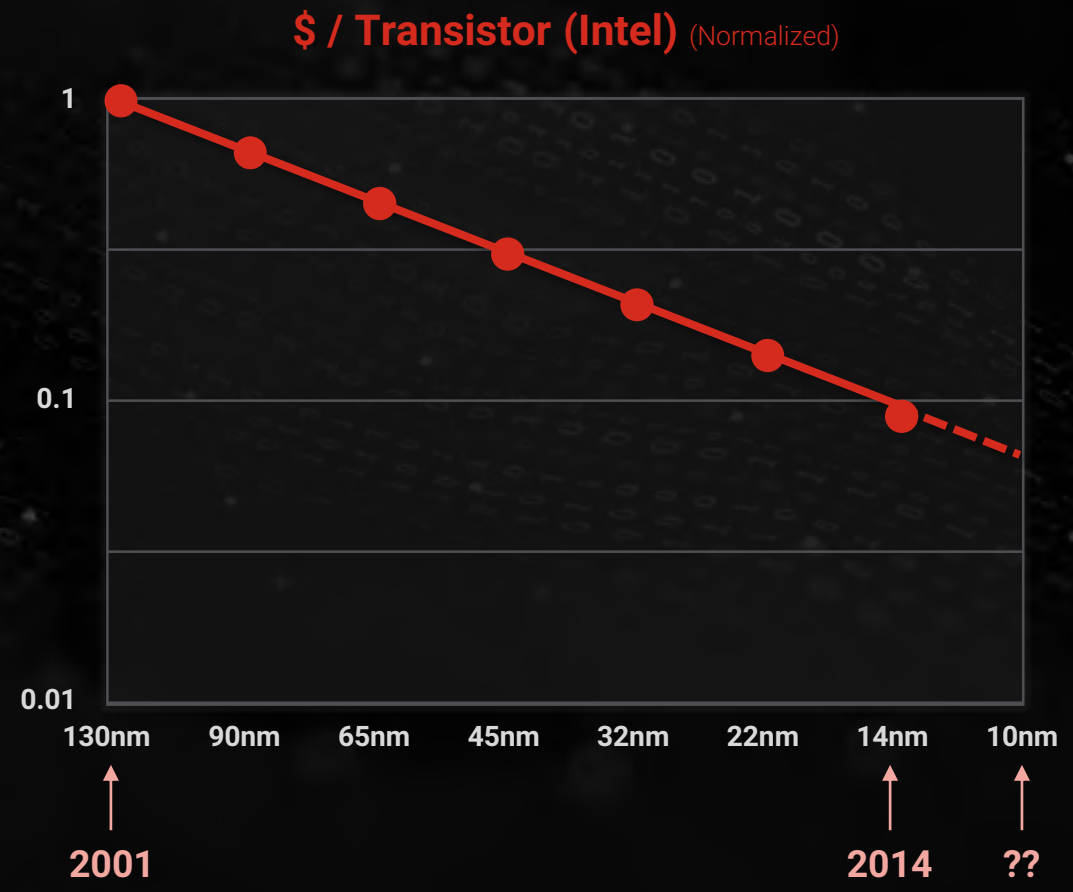
MYTHIC

The Next 10 years

MYTHIC

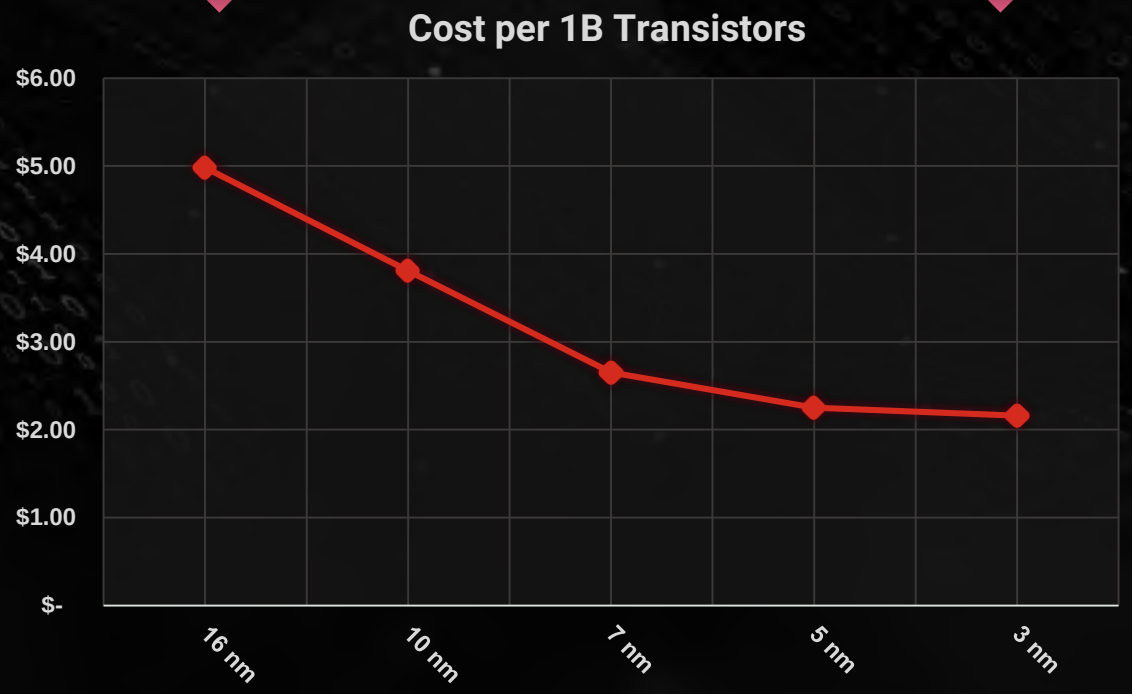
What the Hardware Industry is Used to





Reality: Cost Improvements are Flat-Lining

This should be a huge concern

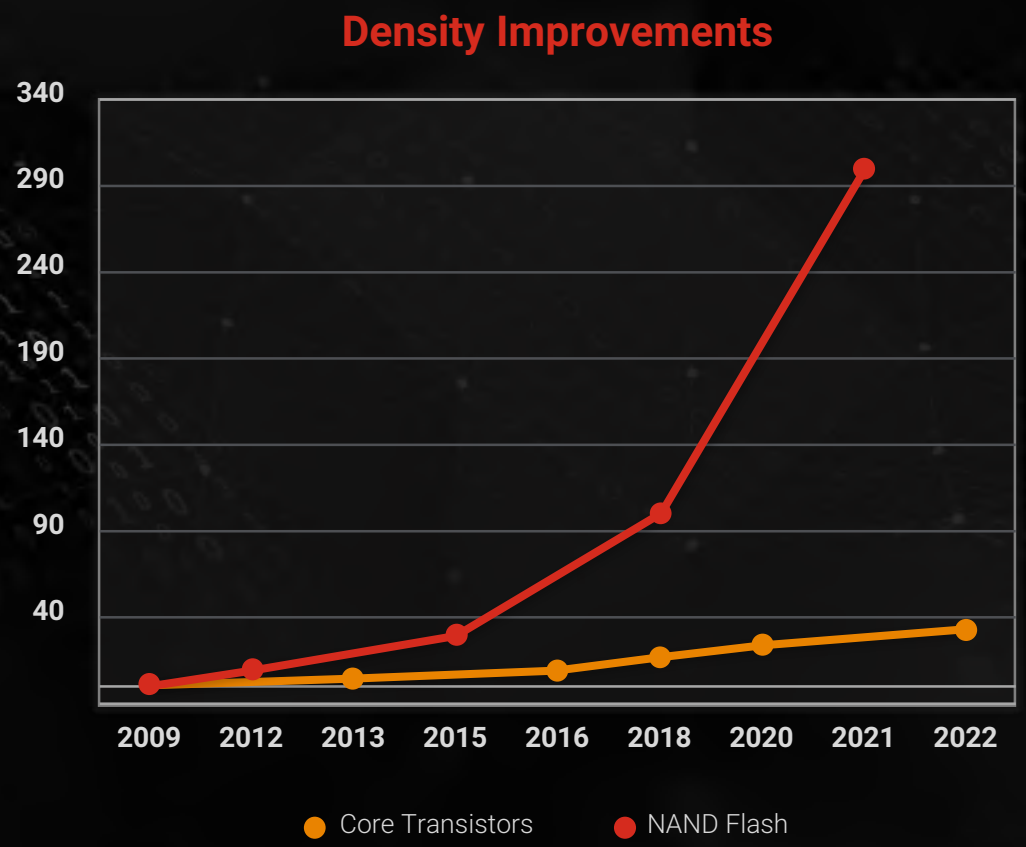


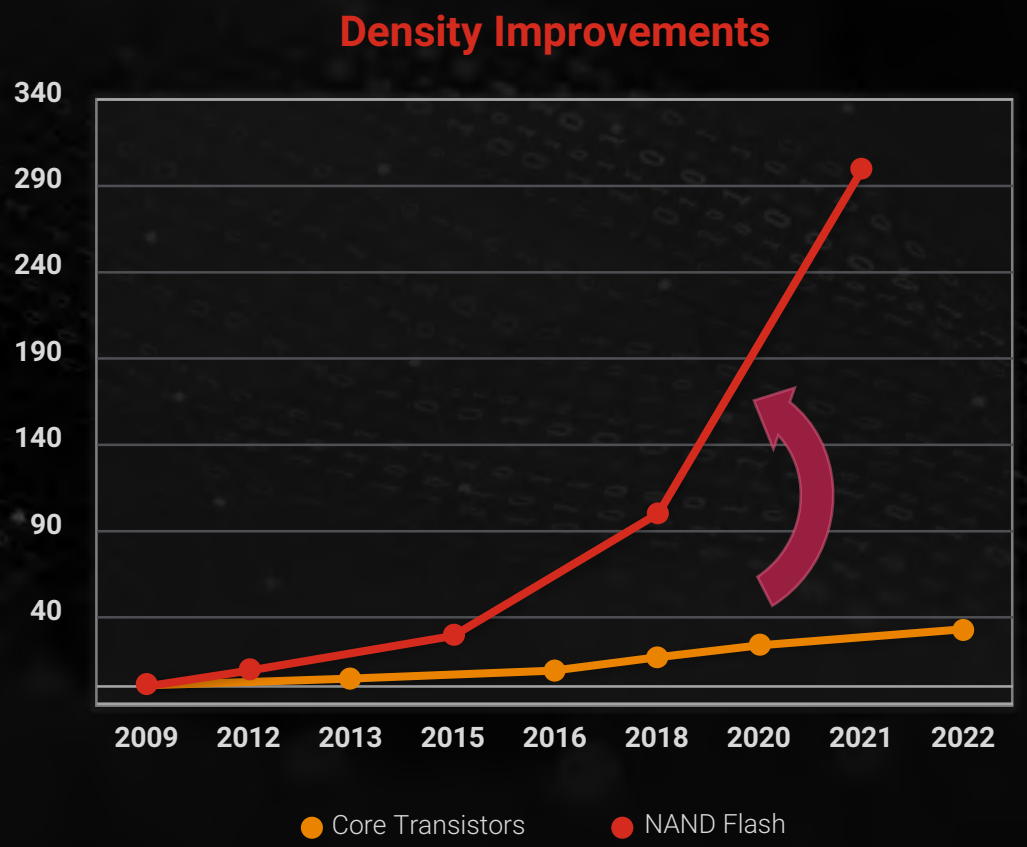
	16nm	10nm	7nm	5nm	3nm
Chip area (mm^2)	125.00	87.66	83.27	85.00	85.00
No. of transistors (BU)	3.3	4.3	6.9	10.5	14.1
Gross die per wafer	478	686	721	707	707
Net die per wafer	359.74	512.44	545.65	530.25	509.04
Wafer price ($)	5,912.00	8,389.00	9,965.00	12,500	15,500
Die cost ($)	16.43	16.37	18.26	23.57	30.45
Transistor cost per 1B transistors ($)	4.98	3.81	2.65	2.25	2.16

https://wccftech.com/apple-5nm-3nm-cost-transistors/

MYTHIC

NVM Scalability With Analog Compute-In-Memory

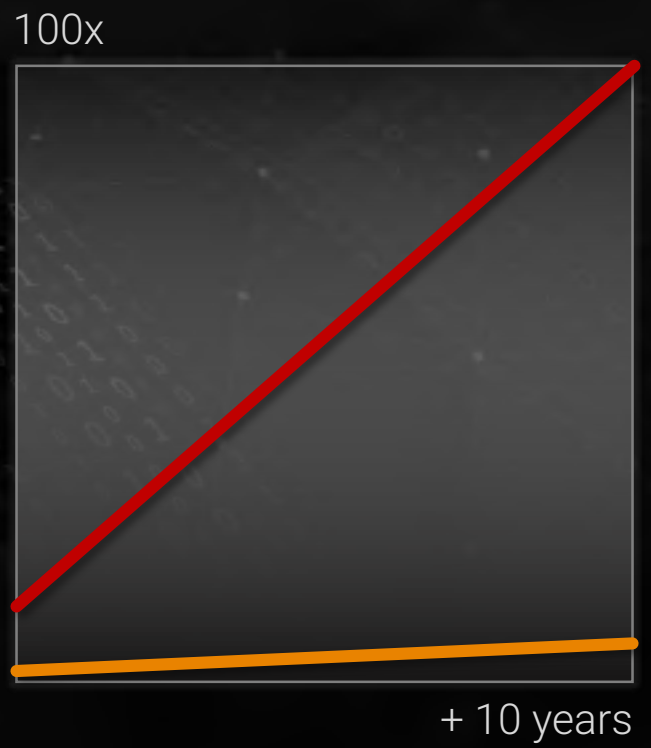




MYTHIC

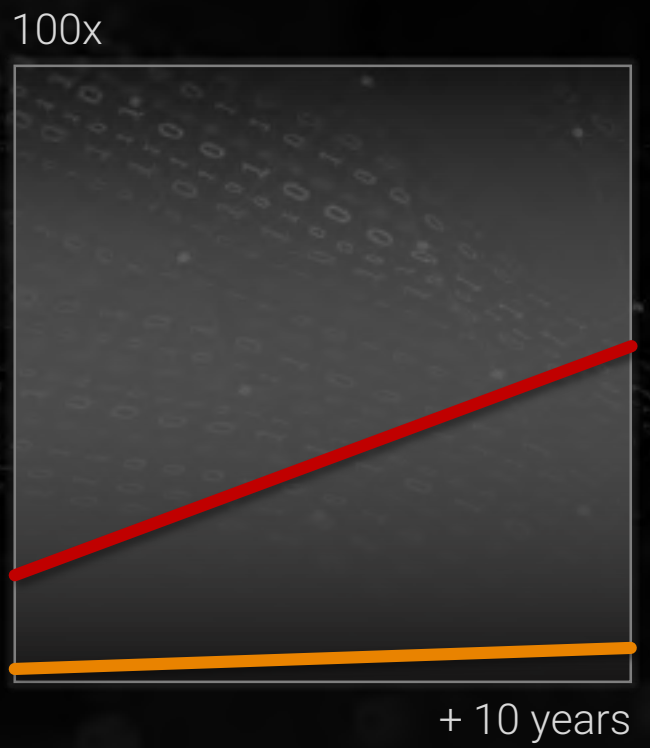
Mythic Analog Compute is the Clear Path Forward

Performance per Dollar



On-chip Model Capacity

Performance per Watt



Mythic Analog Compute-in-Memory Digital

Mythic Analog Compute is the Clear Path Forward