18-600 Foundations of Computer Systems

Lecture 26: "Future of Computing Systems"

John P. Shen (with contribution from Randy Bryant)

December 4, 2017

Introduction to "Neuromorphic Computing"

John P. Shen (with contribution from Mikko Lipasti)

December 4, 2017

Electrical & Computer ENGINEERING

Carnegie Mellon University

Moore's Law Origins

April 19, 1965

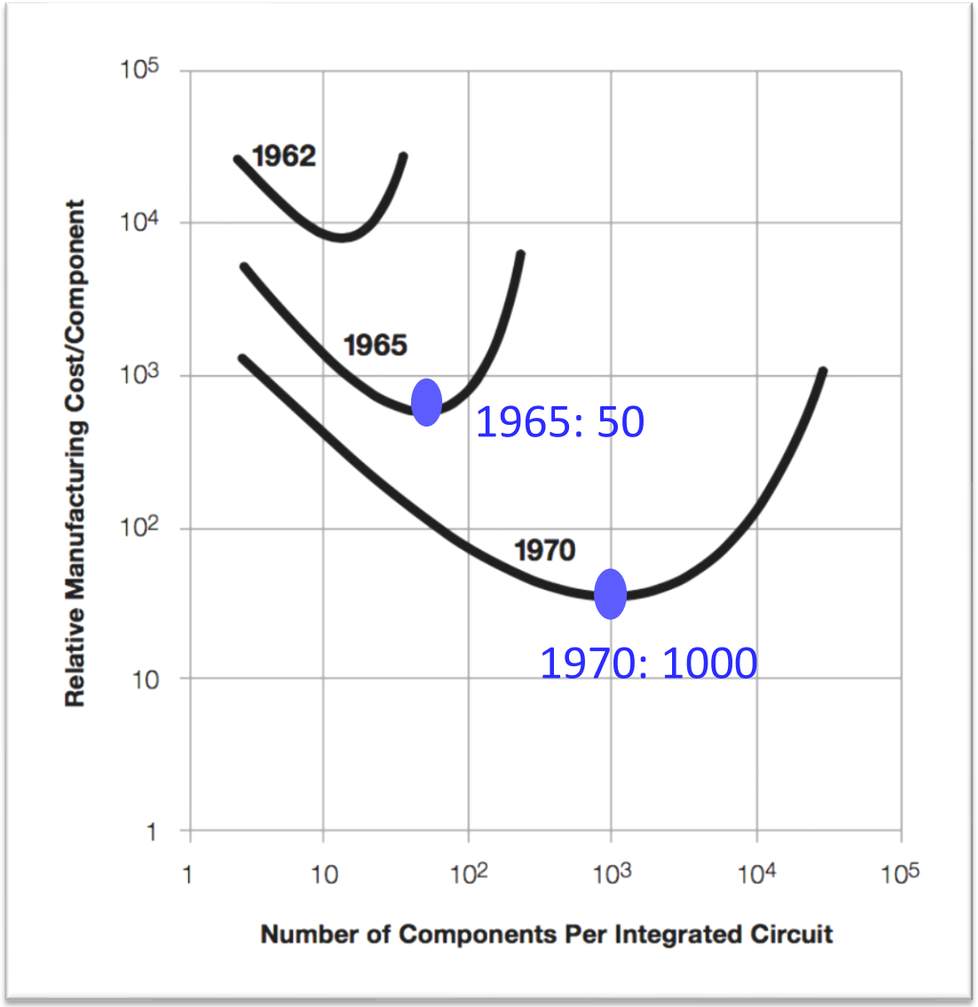
Cramming more components onto integrated circuits

With unit cost falling as the number of components per circuit rises, by 1975 economics may dictate squeezing as many as 65,000 components on a single silicon chip

By Gordon E. Moore

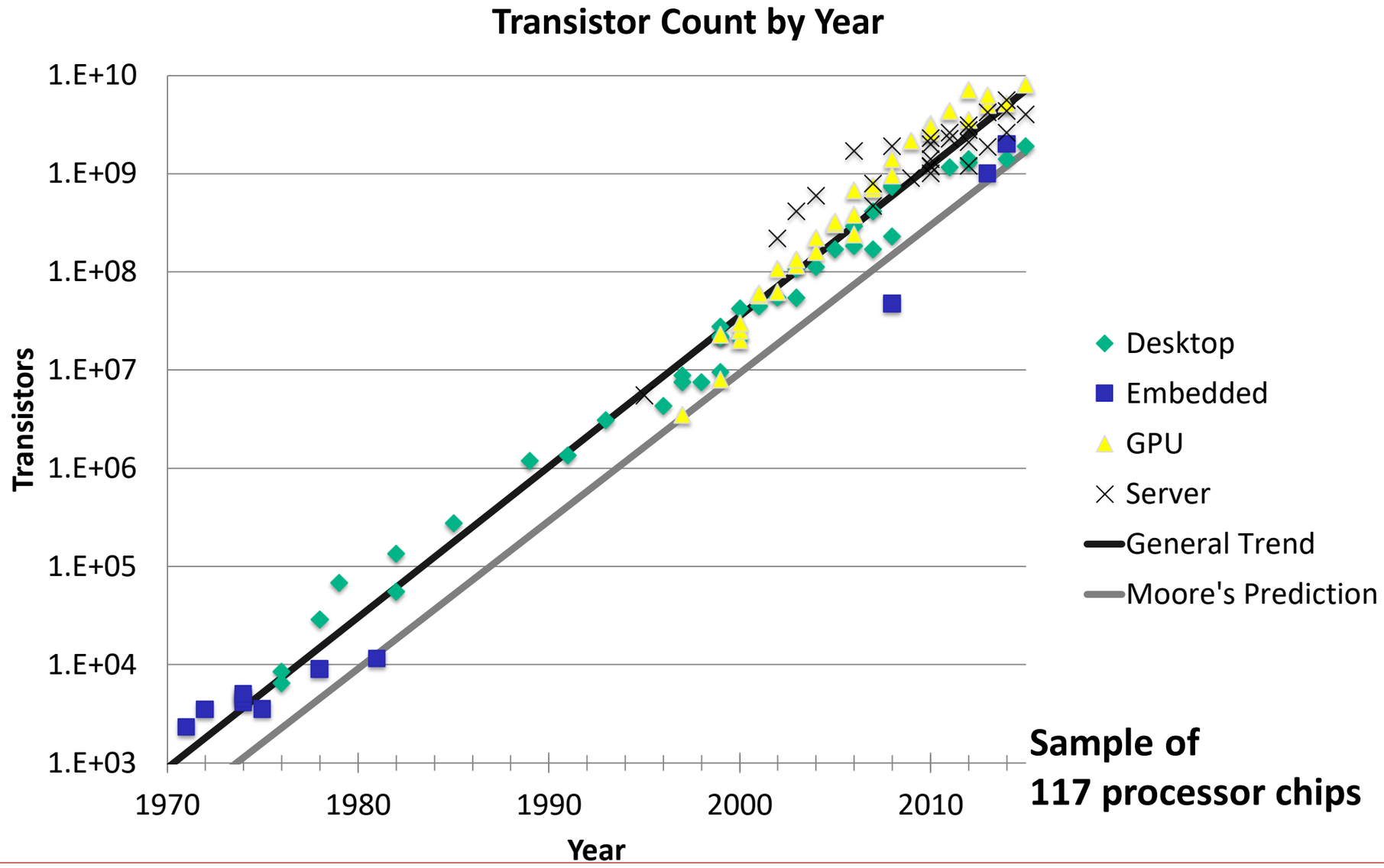
Director, Research and Development Laboratories, Fairchild Semiconductor division of Fairchild Camera and Instrument Corp.

Moore's Law Origins

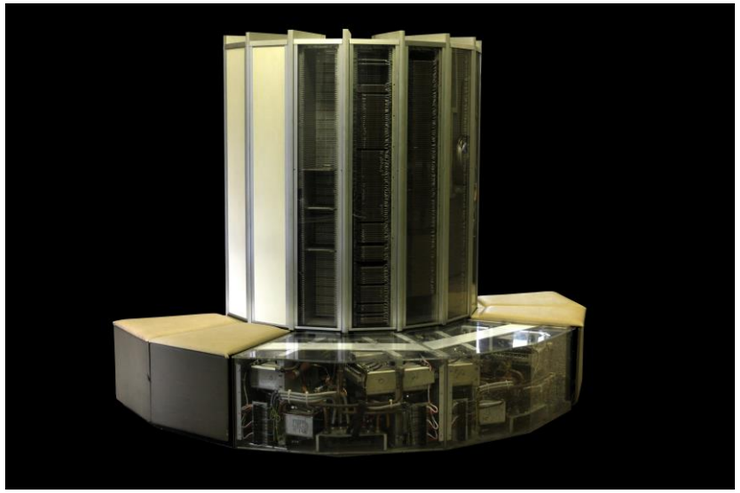


- **Moore's Thesis**
  - Minimize price per device
  - Optimum number of devices / chip increasing 2x / year
- **Later**
  - 2x / 2 years
  - "Moore's Prediction"

Moore's Law: 50 Years



What Moore's Law Has Meant



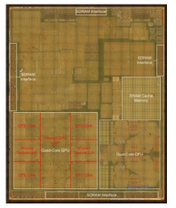
- **1976 Cray 1**
  - 250 M Ops/second
  - ~170,000 chips
  - 0.5B transistors
  - 5,000 kg, 115 KW
  - $9M
  - 80 manufactured

- **2014 iPhone 6**
  - > 4 B Ops/second
  - ~10 chips
  - > 3B transistors
  - 120 g, < 5 W
  - $649
  - 10 million sold in first 3 days

What Moore's Law Has Meant

- **1965 Consumer Product**

- **2015 Consumer Product**

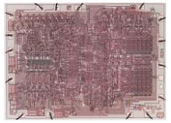


Apple A8 Processor
2 B transistors

Visualizing Moore's Law to Date

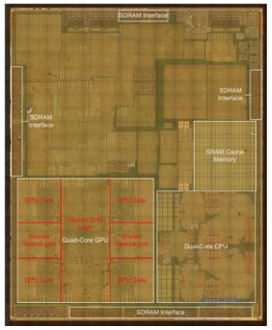
If transistors were the size of a grain of sand

Intel 4004
1970
2,300 transistors



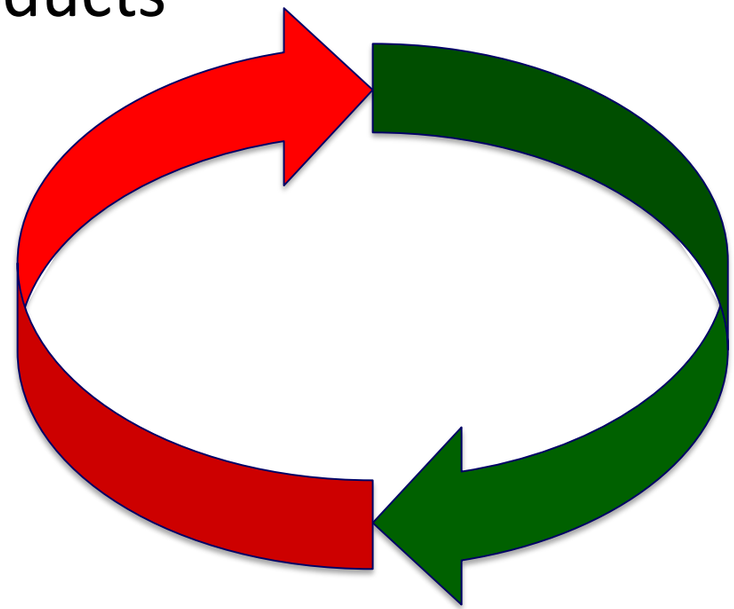
0.1 g

Apple A8
2014
2 B transistors



88 kg

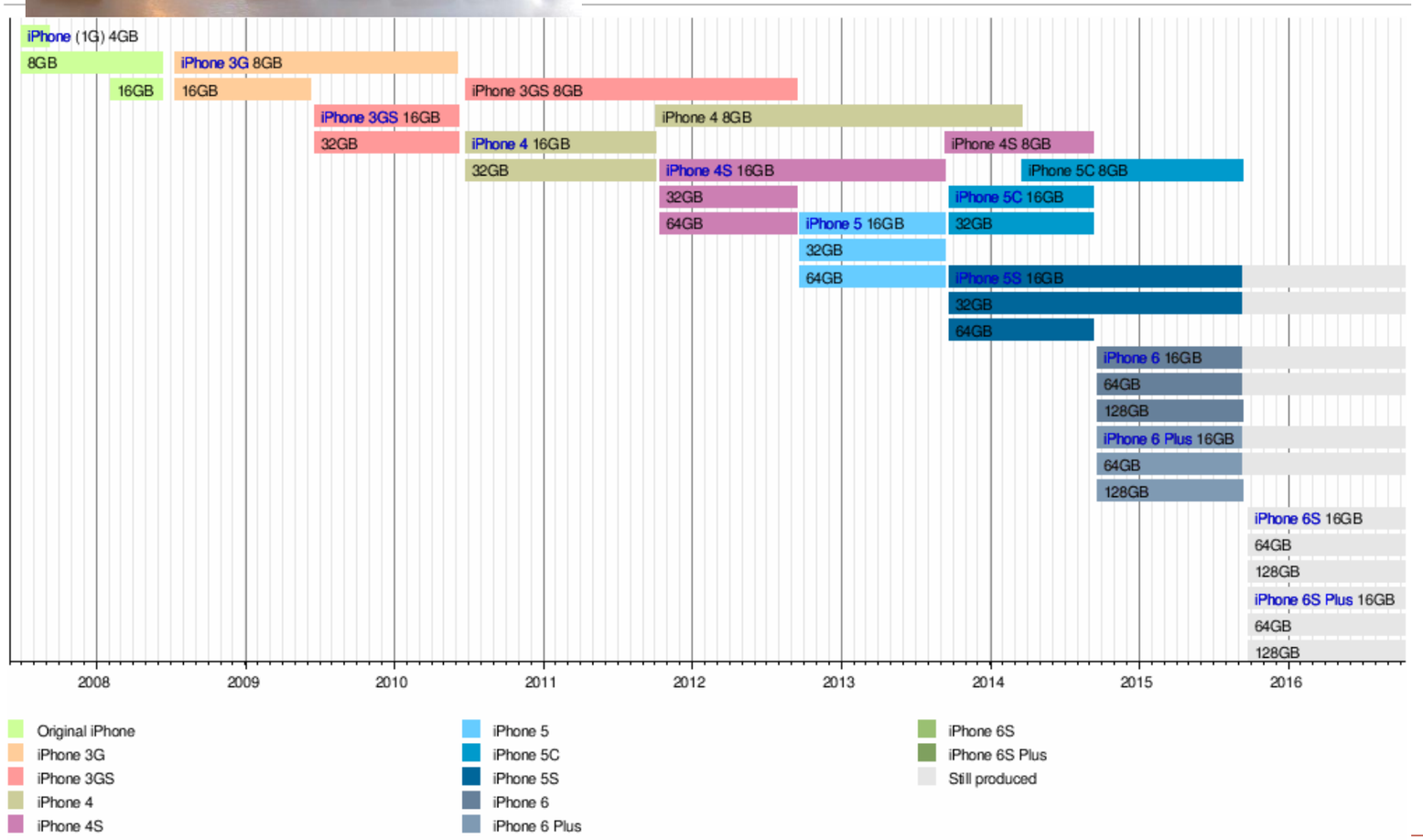
Moore's Law Economics



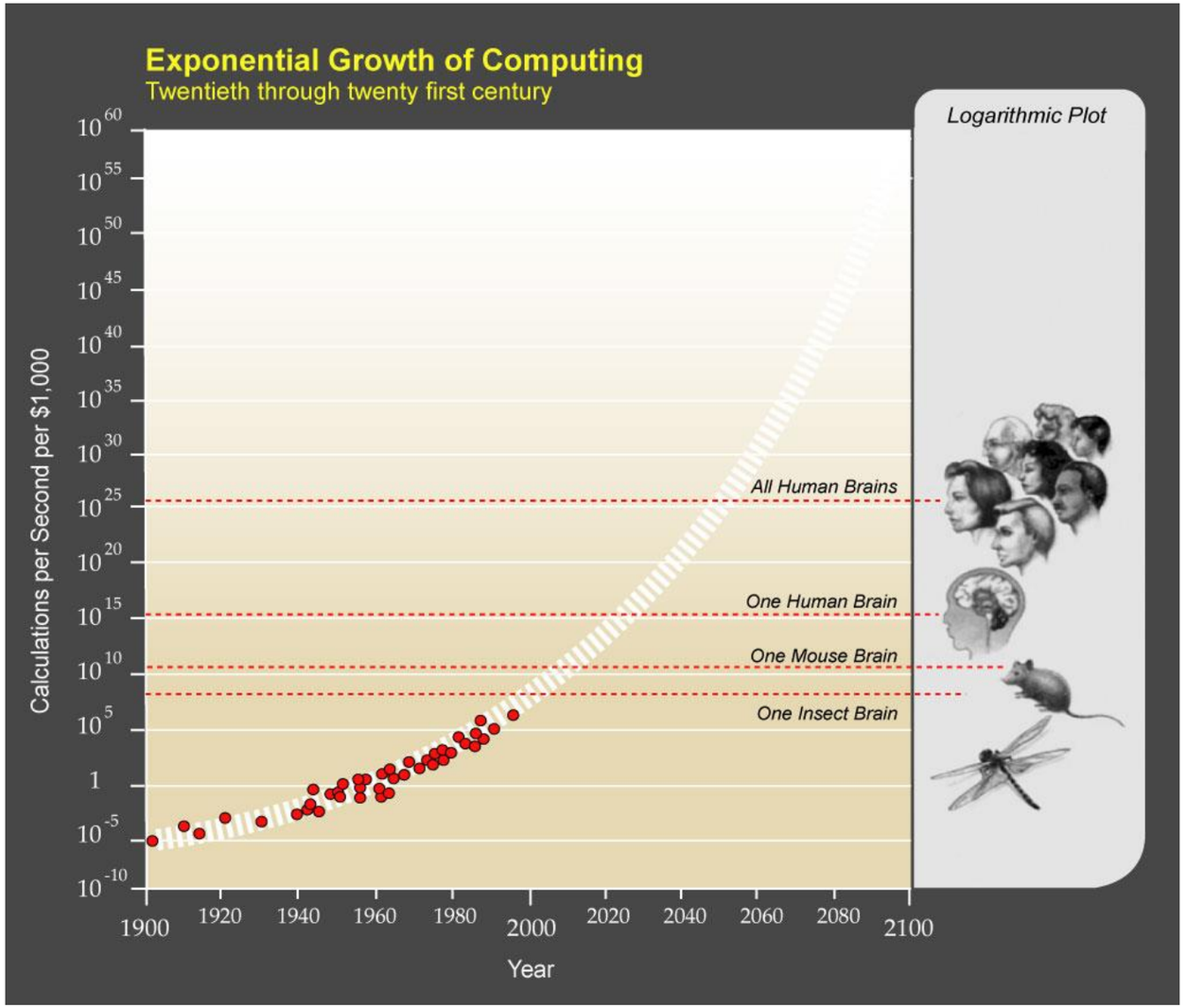
Consumer products sustain the $300B semiconductor industry

What Moore's Law Has Meant

9 generations of iPhone since 2007



What Moore's Law Could Mean



Kurzweil, *The Singularity is Near*, 2005

What Moore's Law Could Mean

- **2015 Consumer Product**

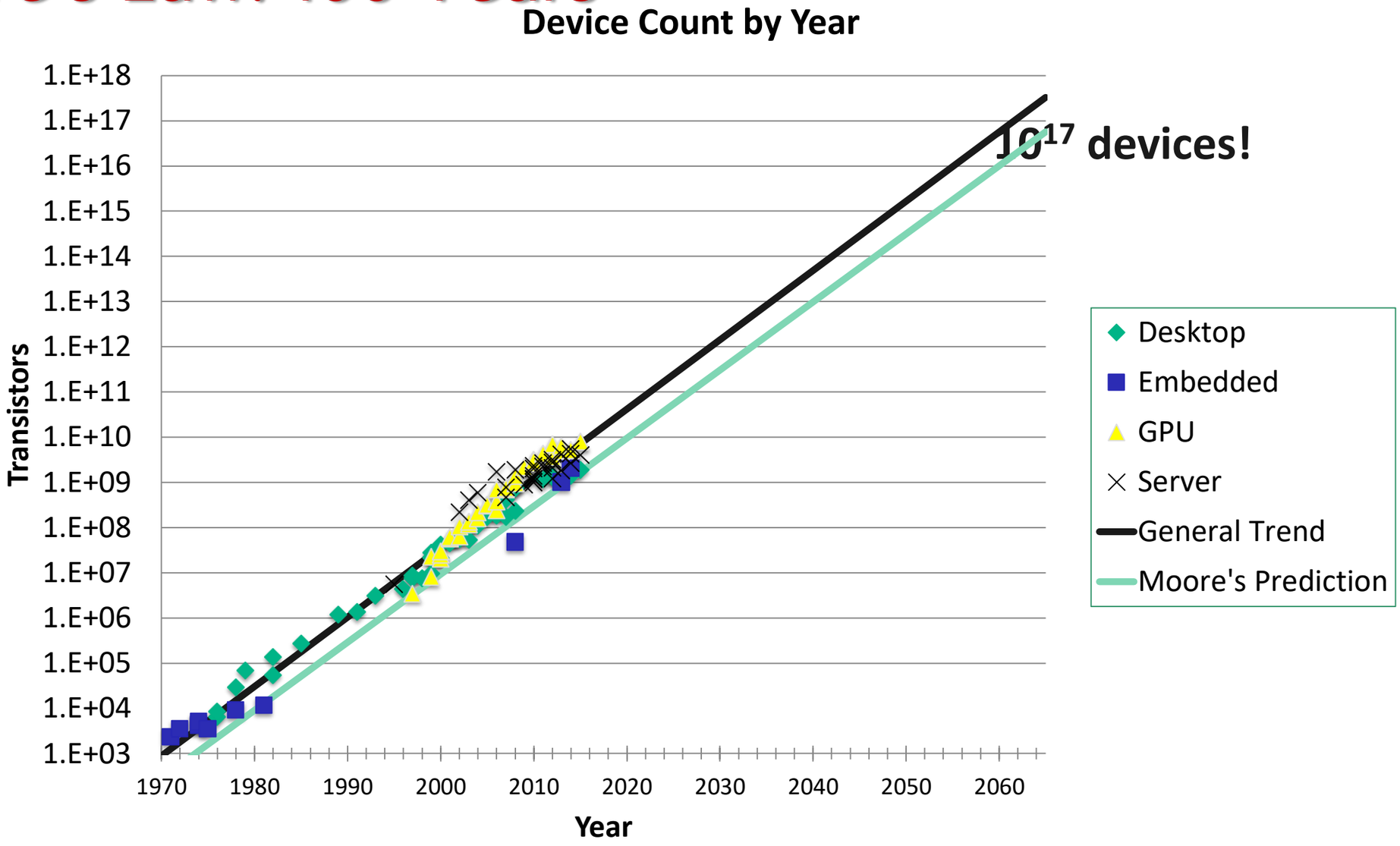
- **2065 Consumer Product**

- Portable
- Low power
- Will drive markets & innovation

Requirements for Future Technology

- **Must be suitable for portable, low-power operation**
  - Consumer products
  - Internet of Things components
  - Not cryogenic, not quantum
- **Must be inexpensive to manufacture**
  - Comparable to current semiconductor technology
    - O(1) cost to make chip with O(*N*) devices
- **Need not be based on transistors**
  - Memristors, carbon nanotubes, DNA transcription, ...
  - Possibly new models of computation
  - But, still want lots of devices in an integrated system

Moore's Law: 100 Years



Visualizing 10^{17} Devices

If devices were the size of a grain of sand

0.1 m^3

3.5 X 10^9 grains

1 million m^3

0.35 X 10^{17} grains

Increasing Transistor Counts

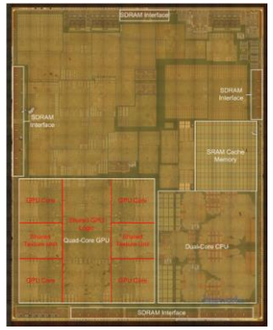
1. **Chips have gotten bigger**
   - 1 area doubling / 10 years
2. **Transistors have gotten smaller**
   - 4 density doublings / 10 years

Will these trends continue?

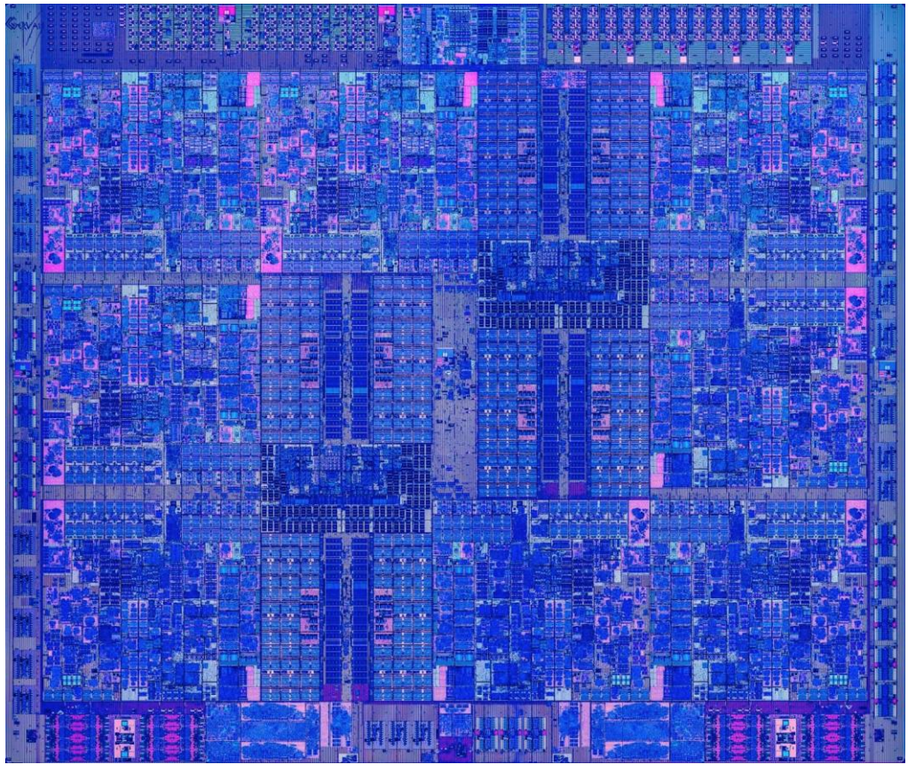
Chips Have Gotten Bigger

Intel 4004
1970
2,300 transistors
12 mm²

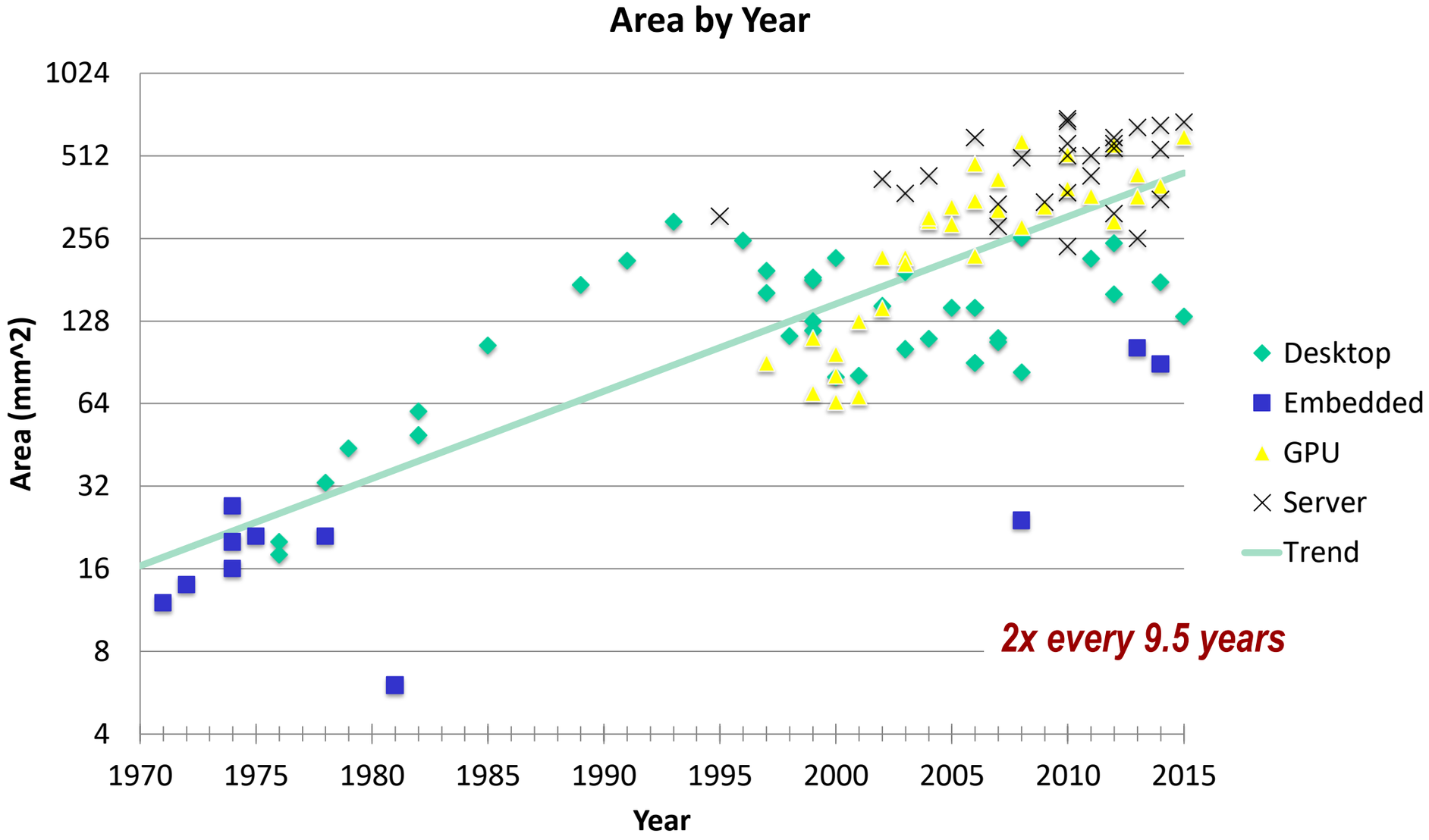
Apple A8
2014
2 B transistors
89 mm²



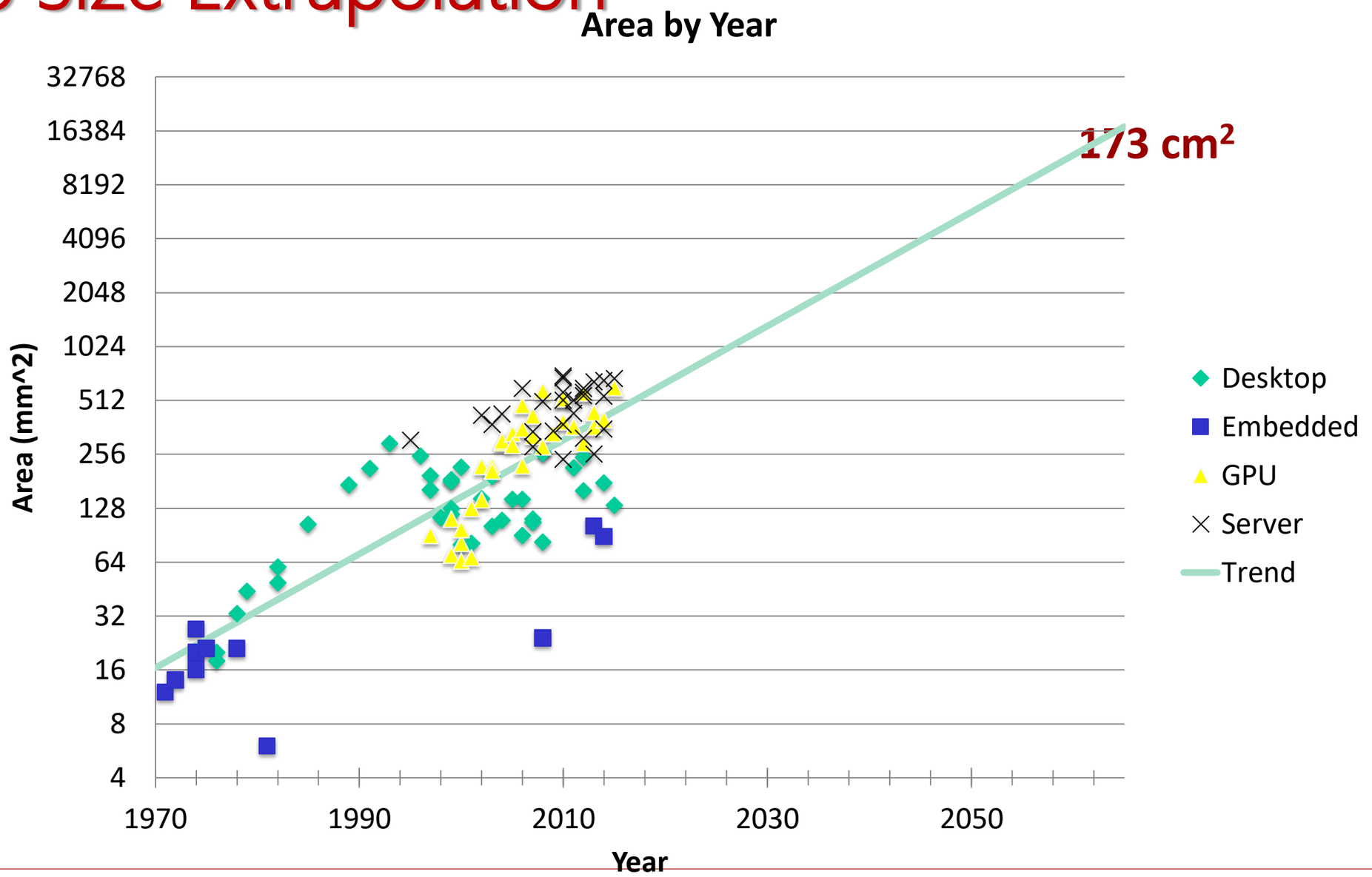
IBM z13
2015
4 B transistors
678 mm²



Chip Size Trend



Chip Size Extrapolation



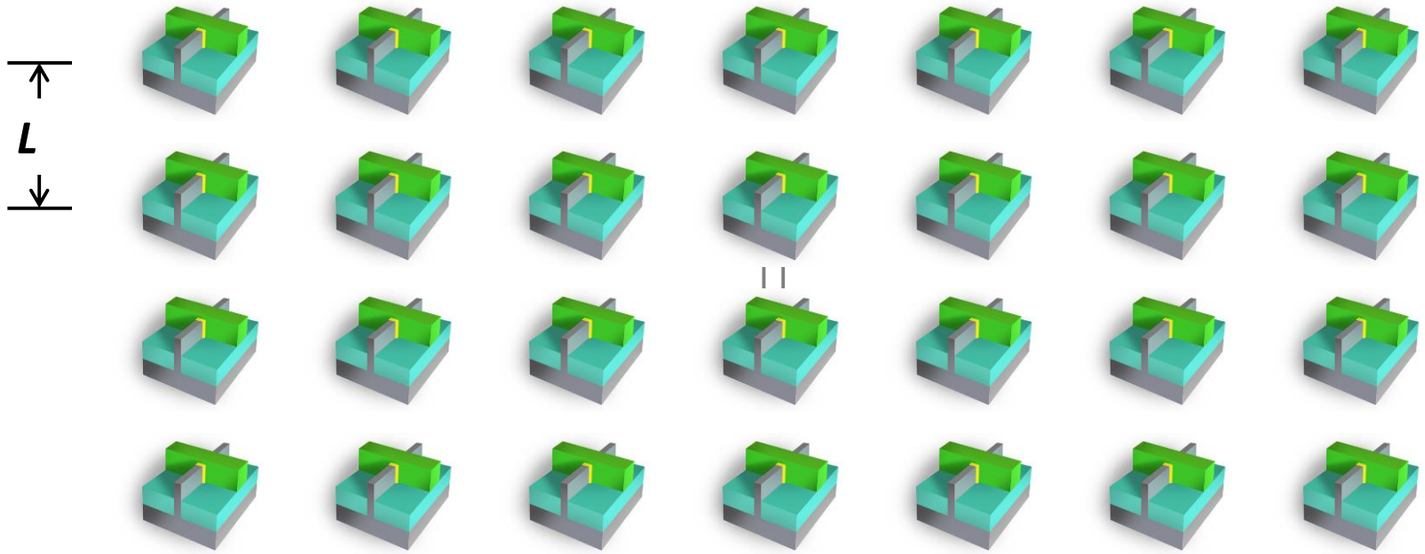
Extrapolation: The iPhone 31s

Apple A59
2065
10^{17} transistors
173 cm^2

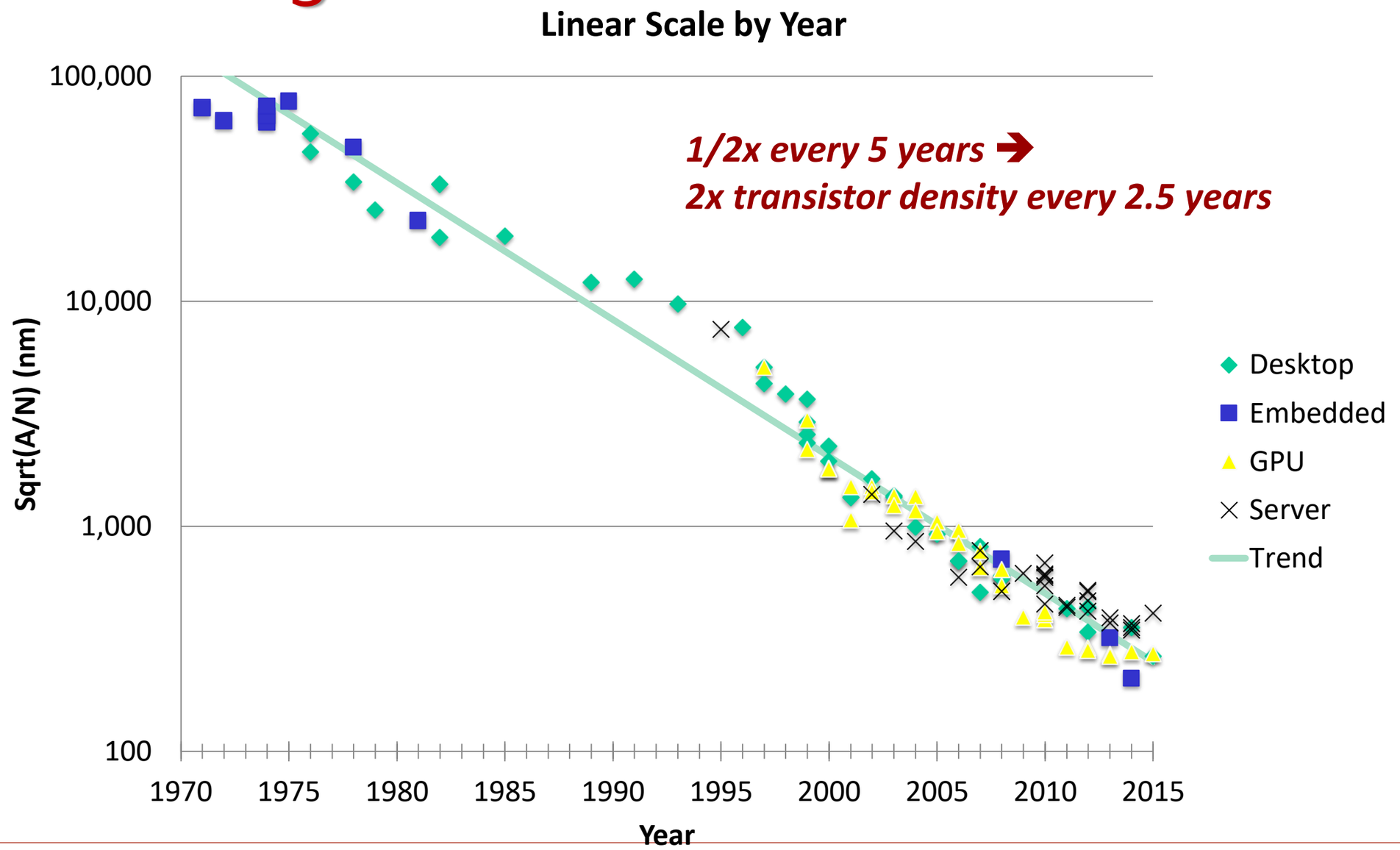
Transistors Have Gotten Smaller

- Area *A*
- *N* devices
- Linear Scale *L*

$$L = \sqrt{A / N}$$



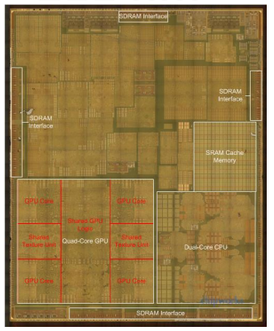
Linear Scaling Trend



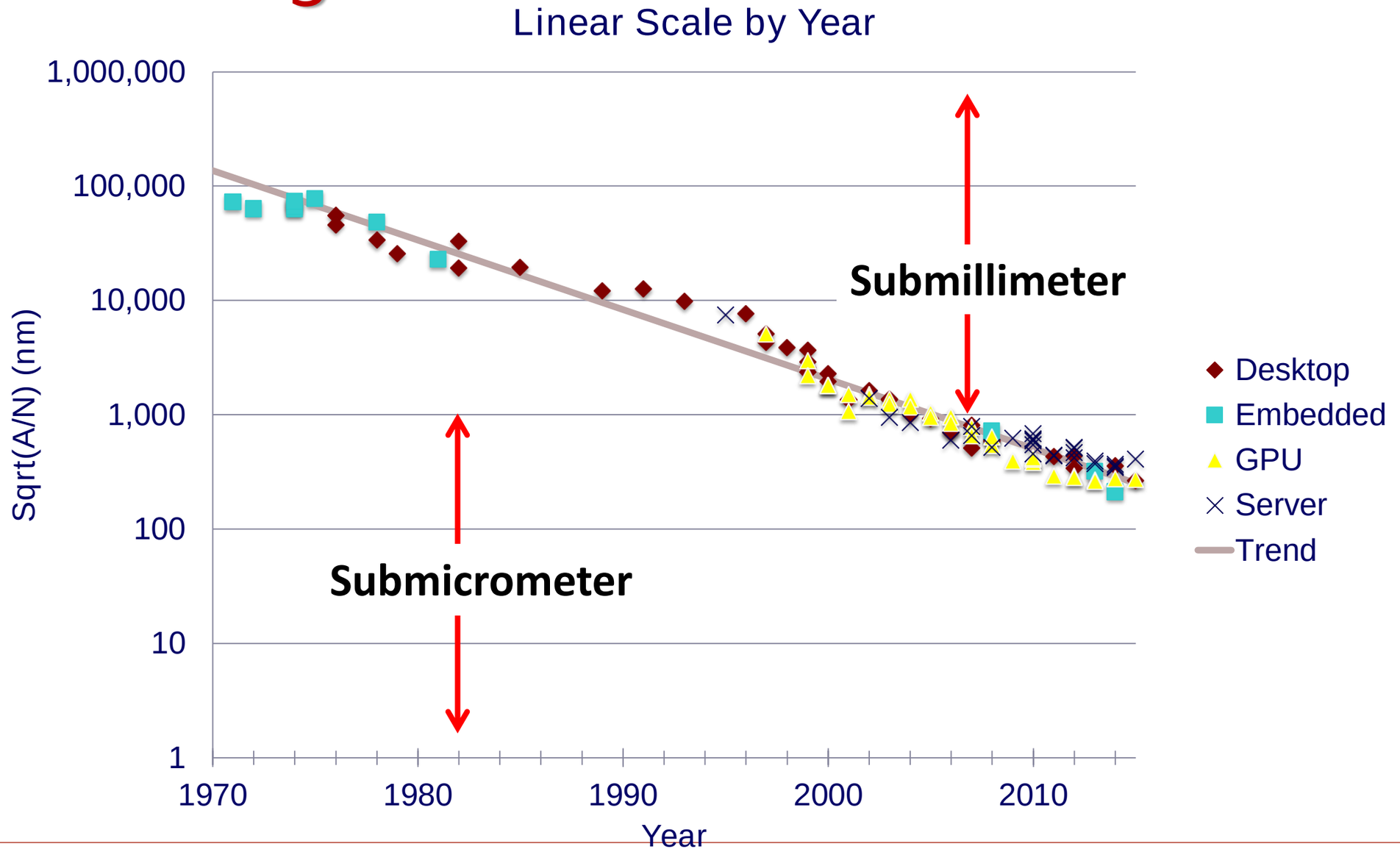
Decreasing Feature Sizes

Intel 4004
1970
2,300 transistors
***L* = 72,000 nm**

Apple A8
2014
2 B transistors
***L* = 211 nm**



Linear Scaling Trend

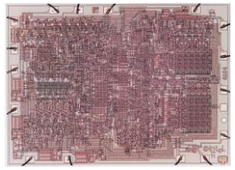


Submillimeter Dimensions

10^{-3} 1 millimeter (mm)

500µm: Length of amoeba

10^{-4}



72µm: Intel 4004 linear scale

50µm: Average size of cell in human body

10^{-5}

10µm: Thickness of sheet of plastic food wrap

5µm: Spider silk thickness

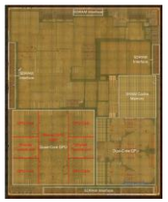
2µm: E coli bacterium length

10^{-6} 1 micrometer (µm)

Submicrometer Dimensions

10^{-6} 1 micrometer

400-700nm: Visible light wavelengths



211nm: Apple A8 linear scale

10^{-7}

30nm: Minimum cooking oil smoke particle diameter

10^{-8}

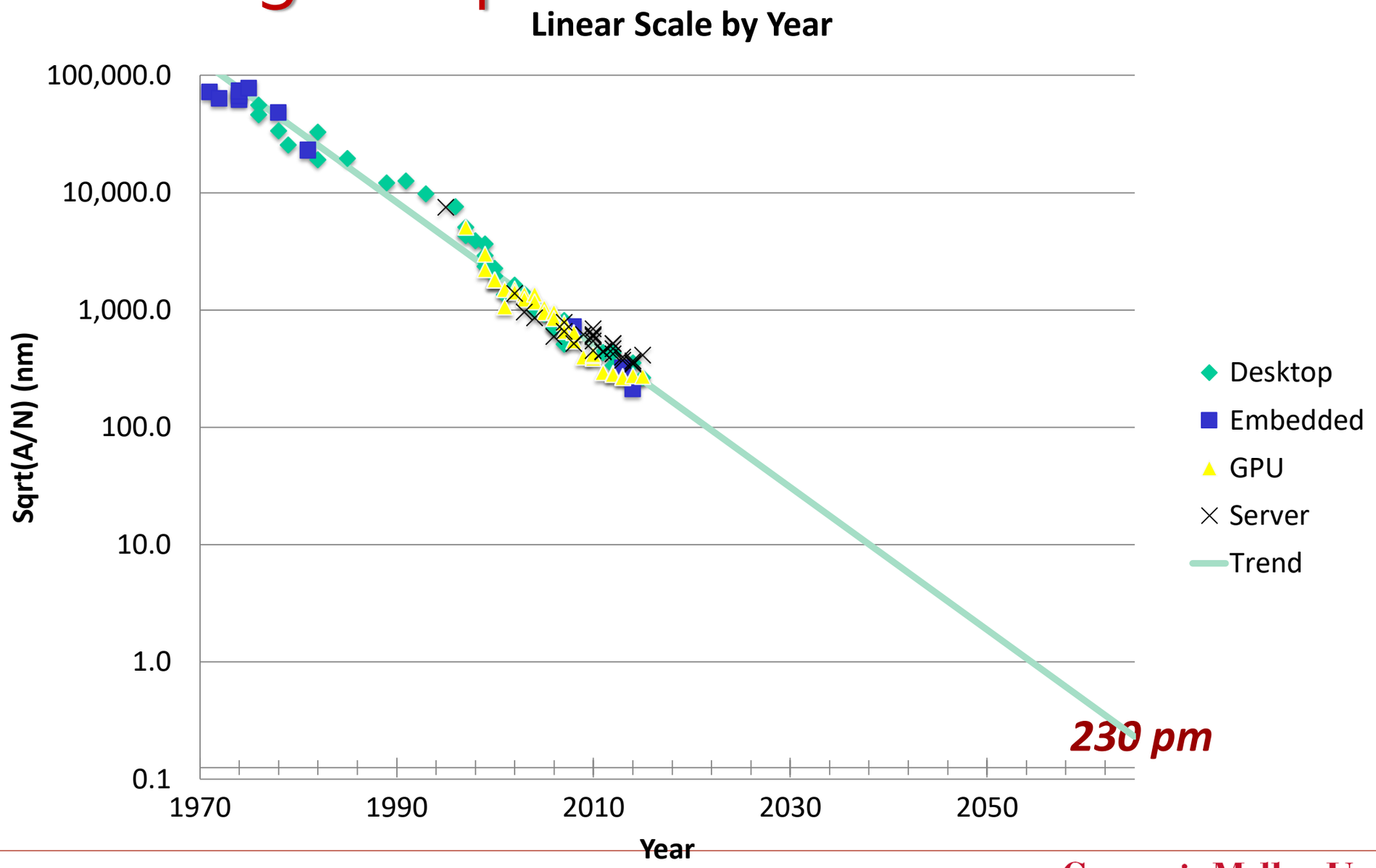
9nm: Cell membrane thickness

2nm: DNA helix diameter

10^{-9} 1 nanometer (nm)

1nm: Carbon nanotube diameter

Linear Scaling Extrapolation



Subnanometer Dimensions

10^{-9} 1 nanometer (nm)

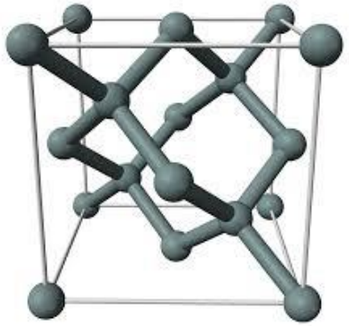
10^{-10}

10^{-11}

10^{-12} 1 picometer (pm)

1nm: Carbon nanotube diameter

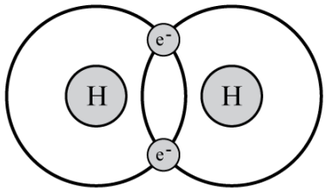
543pm: Silicon crystal lattice spacing



230pm: 2065 linear scale projection

74pm: Spacing between atoms in hydrogen molecule

53pm: Electron-proton spacing in hydrogen (Bohr radius)



2.4pm: Electron wavelength (Compton wavelength)

Reaching 2065 Goal

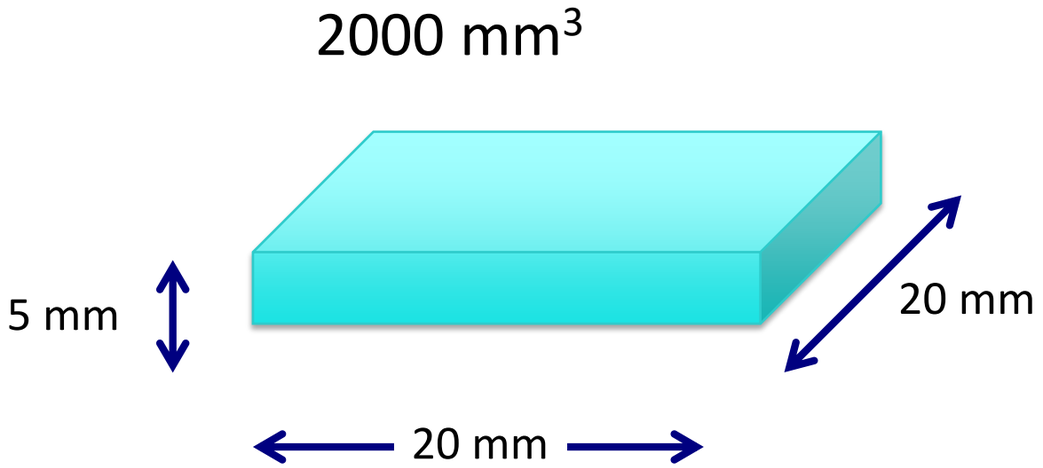
- **Target**
  - 10^{17} devices
  - 400 mm^2
  - L = 63 pm

- **Is this possible?**

No!

Not with 2-d fabrication

Fabricating in 3 Dimensions



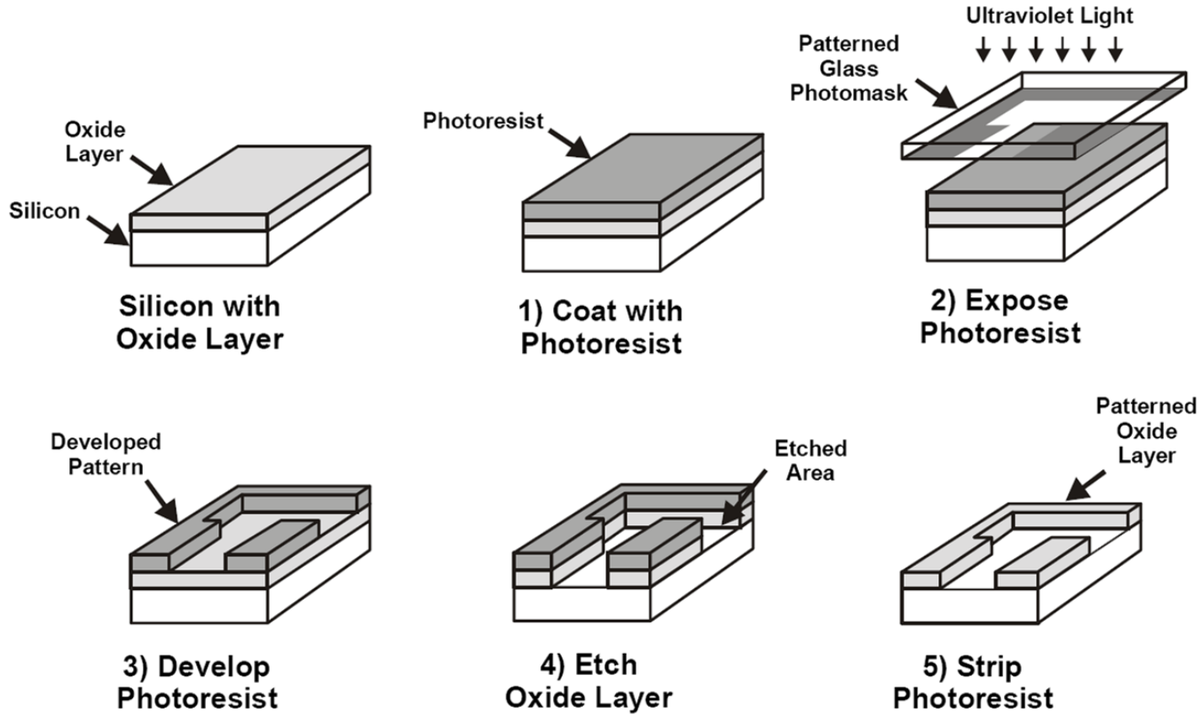
- **Parameters**
  - 10^{17} devices
  - 100,000 logical layers
    - Each 50 nm thick
    - ~1,000,000 physical layers
      - To provide wiring and isolation
  - *L* = 20 nm
    - 10x smaller than today

2065 mm³

3D Fabrication Challenges

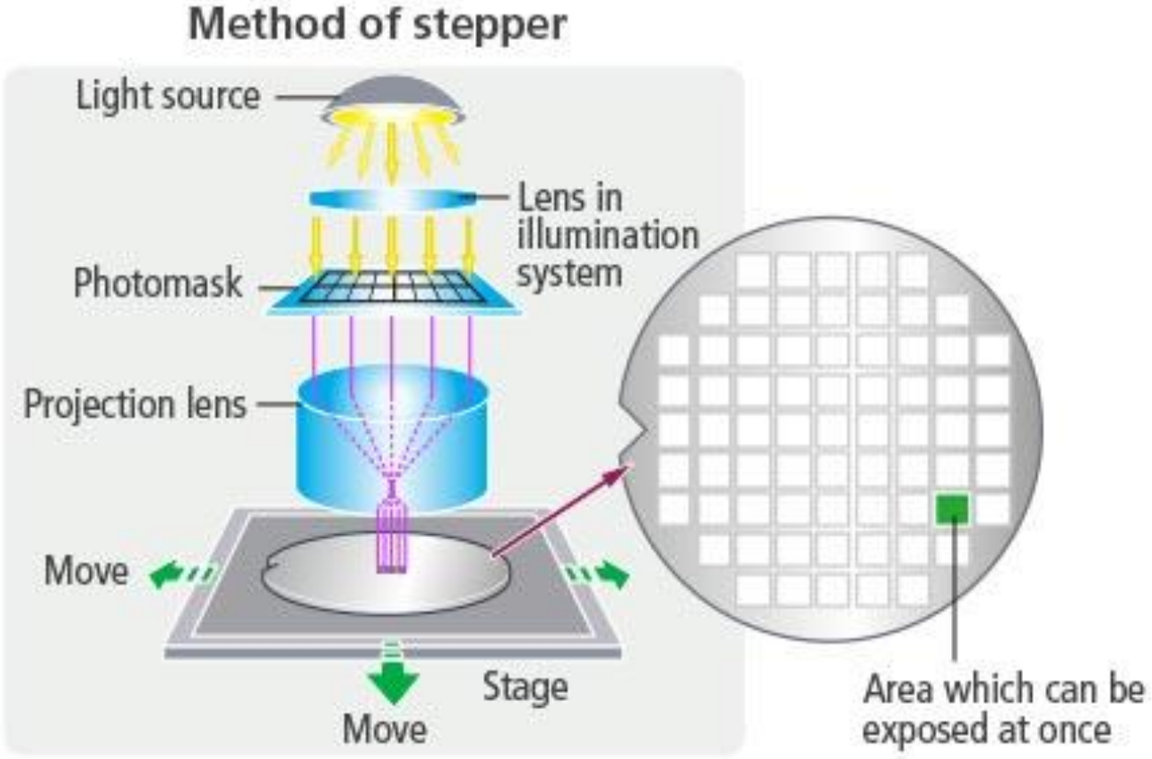
- **Yield**
  - How to avoid or tolerate flaws
- **Cost**
  - High cost of lithography
- **Power**
  - Keep power consumption within acceptable limits
  - Limited energy available
  - Limited ability to dissipate heat

Photolithography



- Pattern entire chip in one step
- Modern chips require ~60 lithography steps
- Fabricate *N* transistor system with O(1) steps

Fabrication Costs

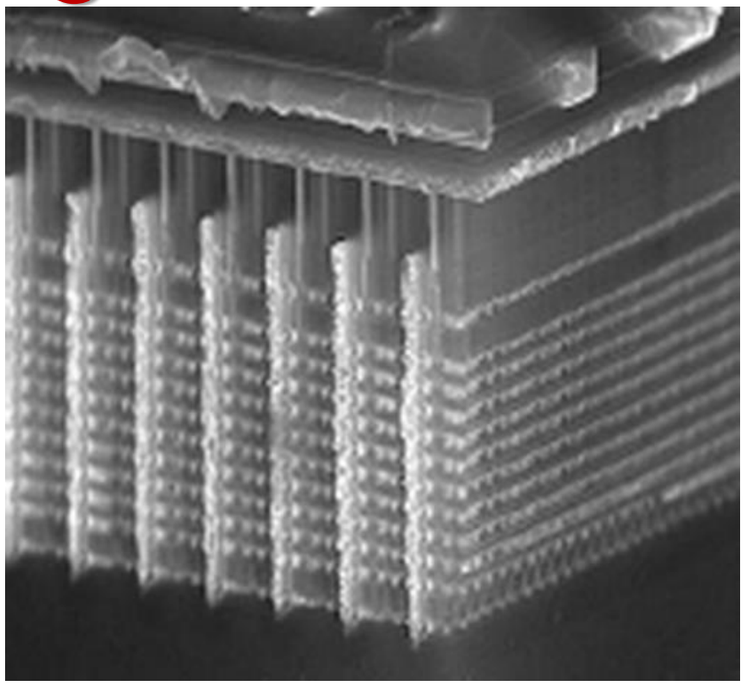


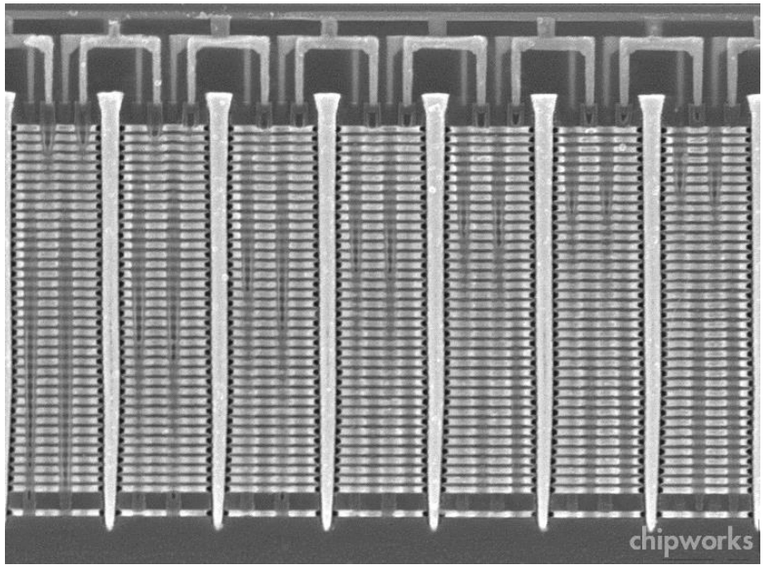
- **Stepper**
  - Most expensive equipment in fabrication facility
  - Rate limiting process step
    - 18s / wafer
  - Expose 858 mm^2 per step
    - 1.2% of chip area

Fabrication Economics

- **Currently**
  - Fixed number of lithography steps
  - Manufacturing cost $10–$20 / chip
    - Including amortization of facility
- **Fabricating 1,000,000 physical layers**
  - Cannot do lithography on every step
- **Options**
  - Chemical self assembly
    - Devices generate themselves via chemical processes
  - Pattern multiple layers at once

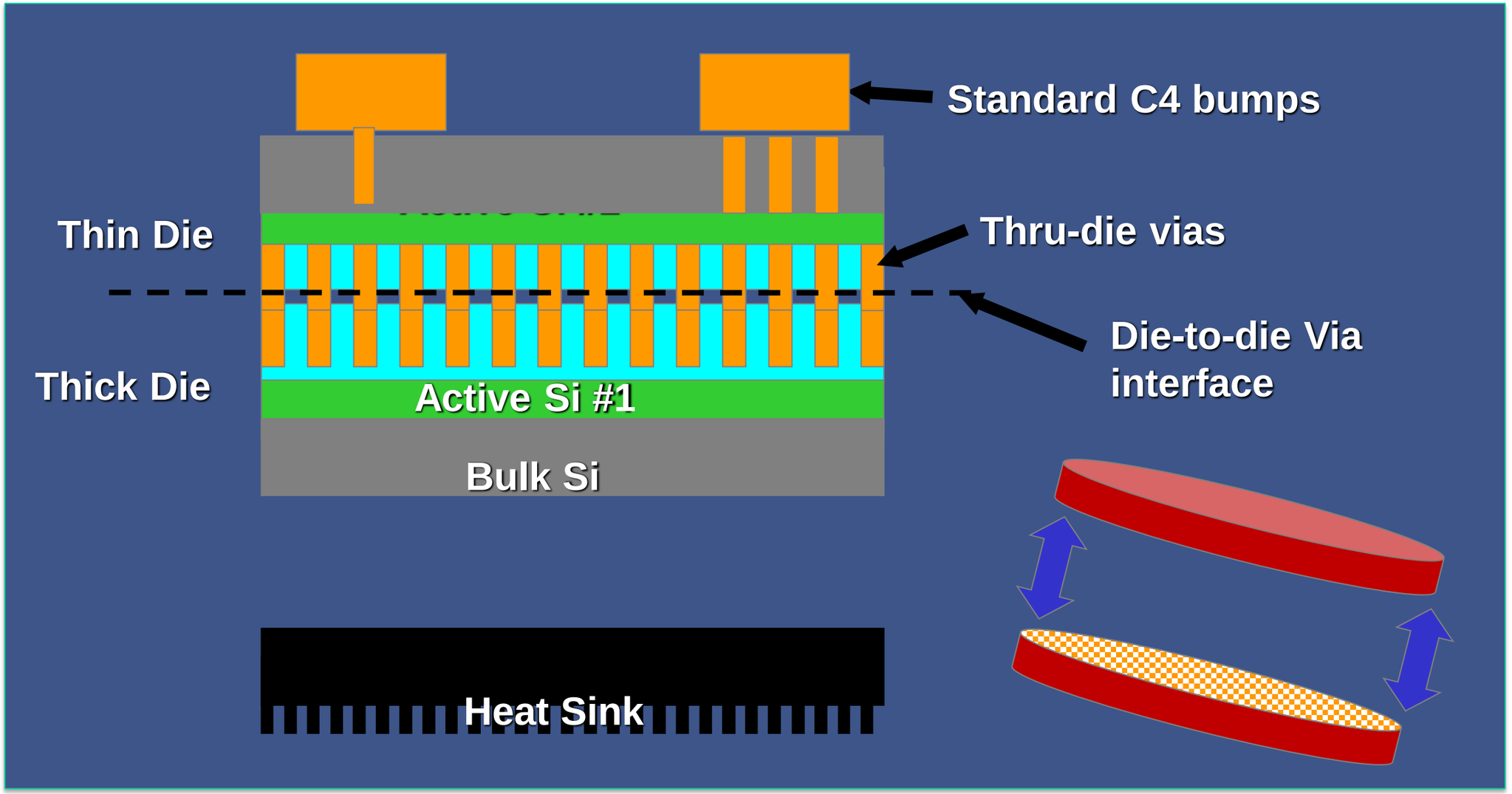
Samsung V-Nand Flash Example



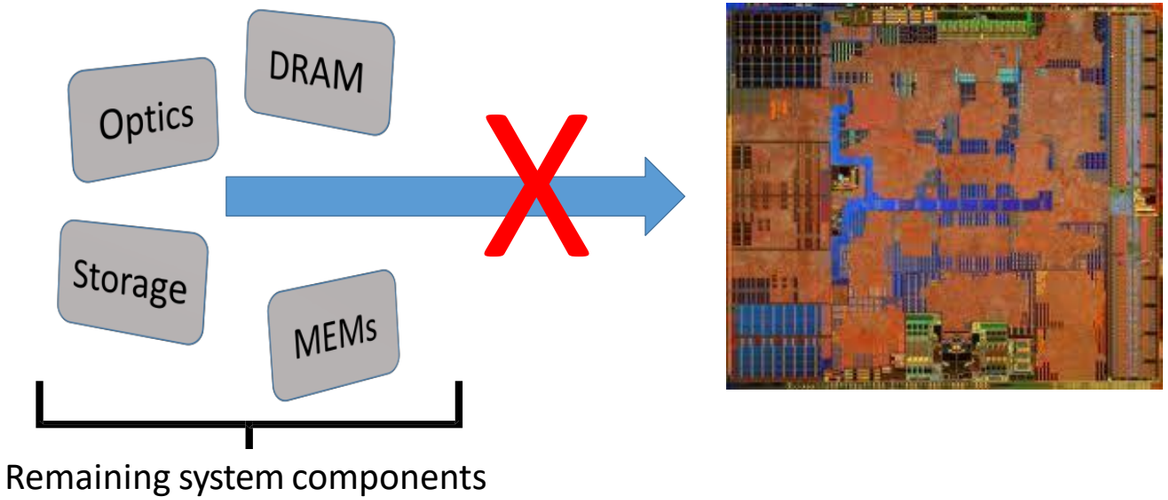


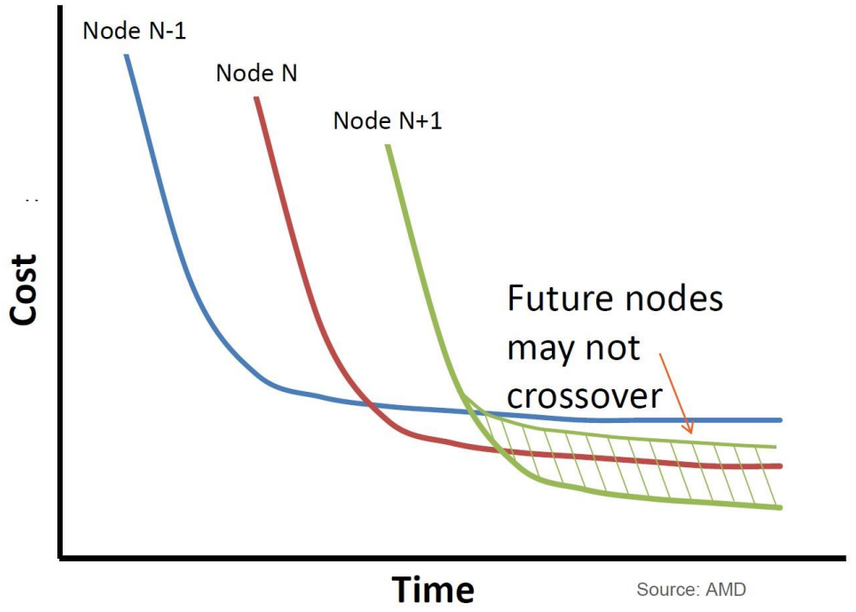
- Build up layers of unpatterned material
- Then use lithography to slice, drill, etch, and deposit material across all layers
- ~30 total masking steps
- Up to 48 layers of memory cells
- Exploits particular structure of flash memory circuits

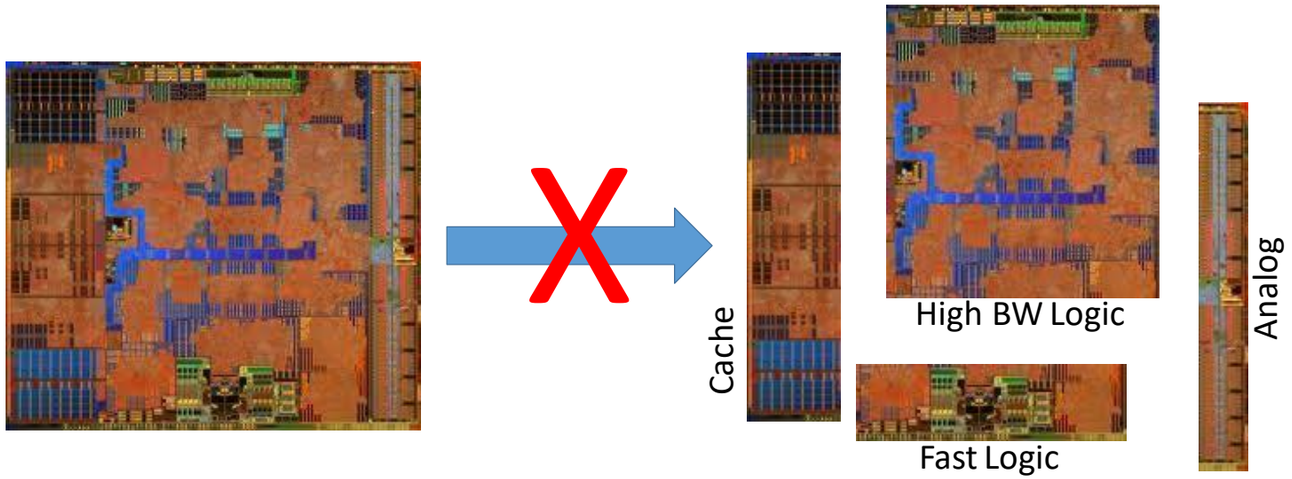
Potentials of 3D Die Stacking



Three Limitations to Moore's Law



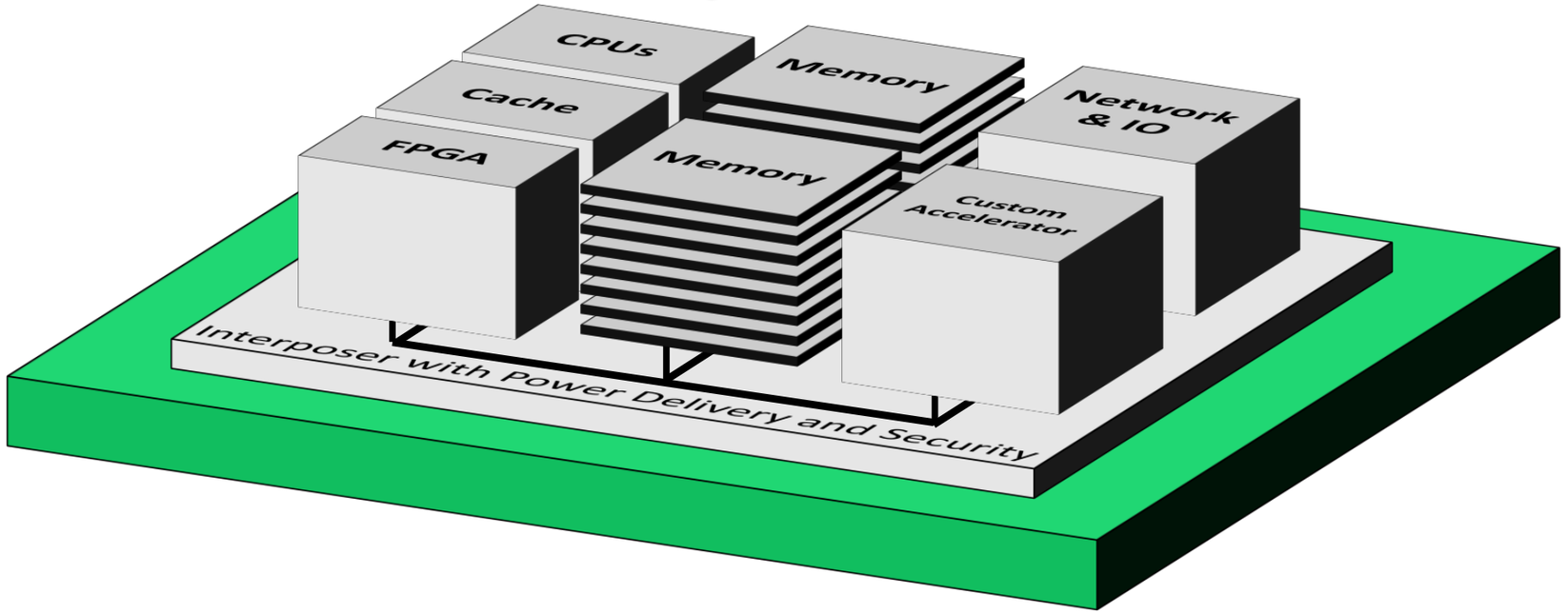




These limitations will make it very challenging to continue integrating systems

[Bryan Black, 2015, AMD]

Strategic Vision (Stacked System)



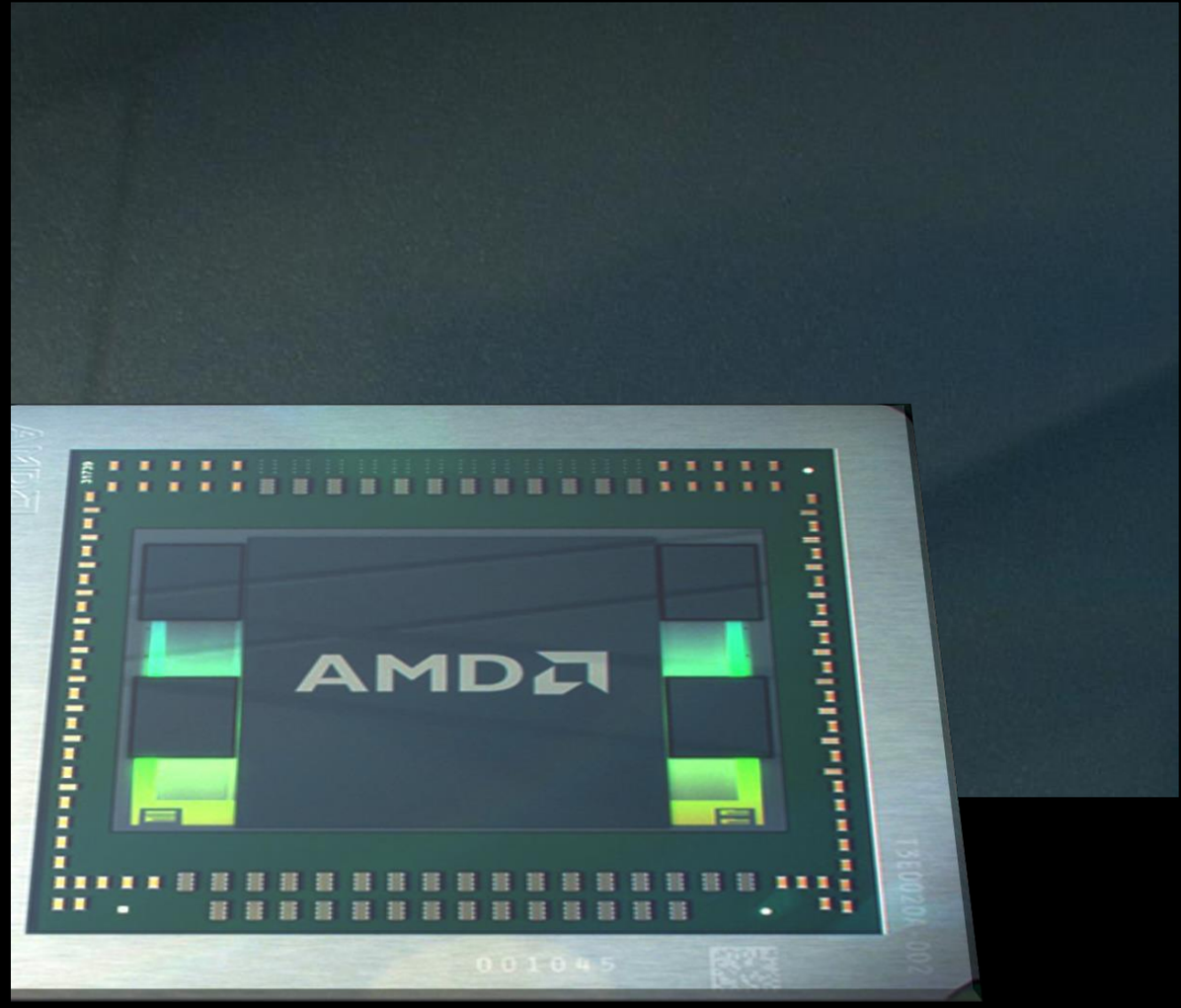
- The Stacked System model integrates dies from disparate technologies using a combination of 2.5D and 3D technology
- This construction model enables:
  - Disparate die integration to improve form factors and reduce system overheads
  - Die splitting to reduce process node complexity and cost
- Results in an interesting business model opportunity

[Bryan Black, 2015, AMD]

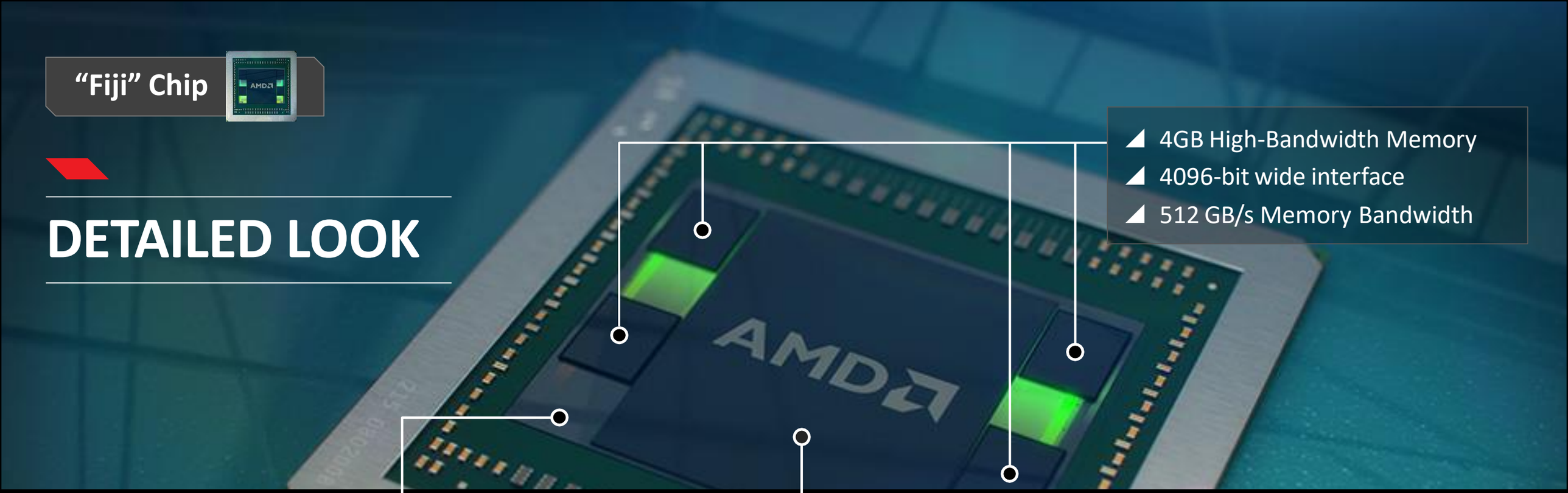
The Road to

"FIJI"

Featuring Die Stacking and HBM Technology

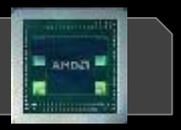


[Bryan Black, 2015, AMD]



[Bryan Black, 2015, AMD]

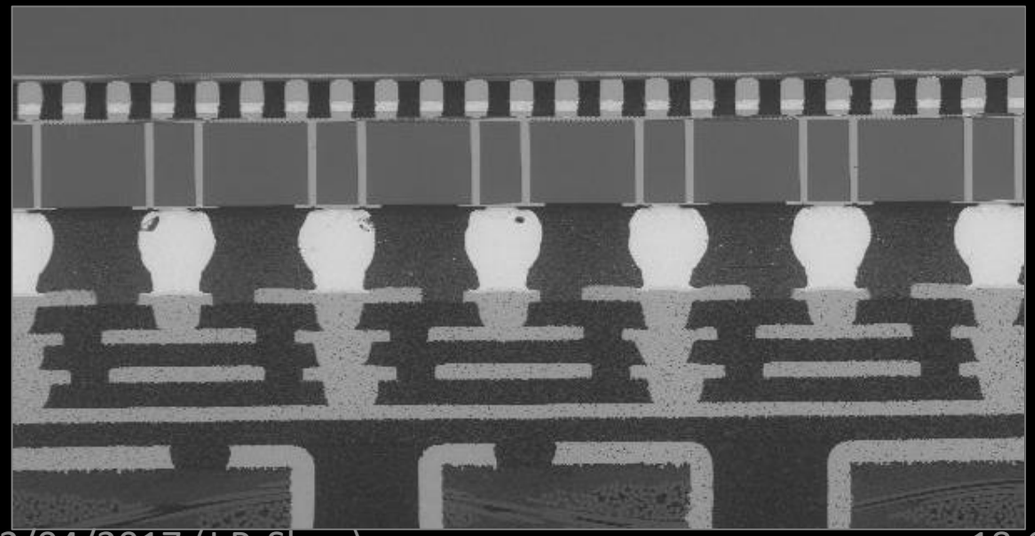
"Fiji" Chip

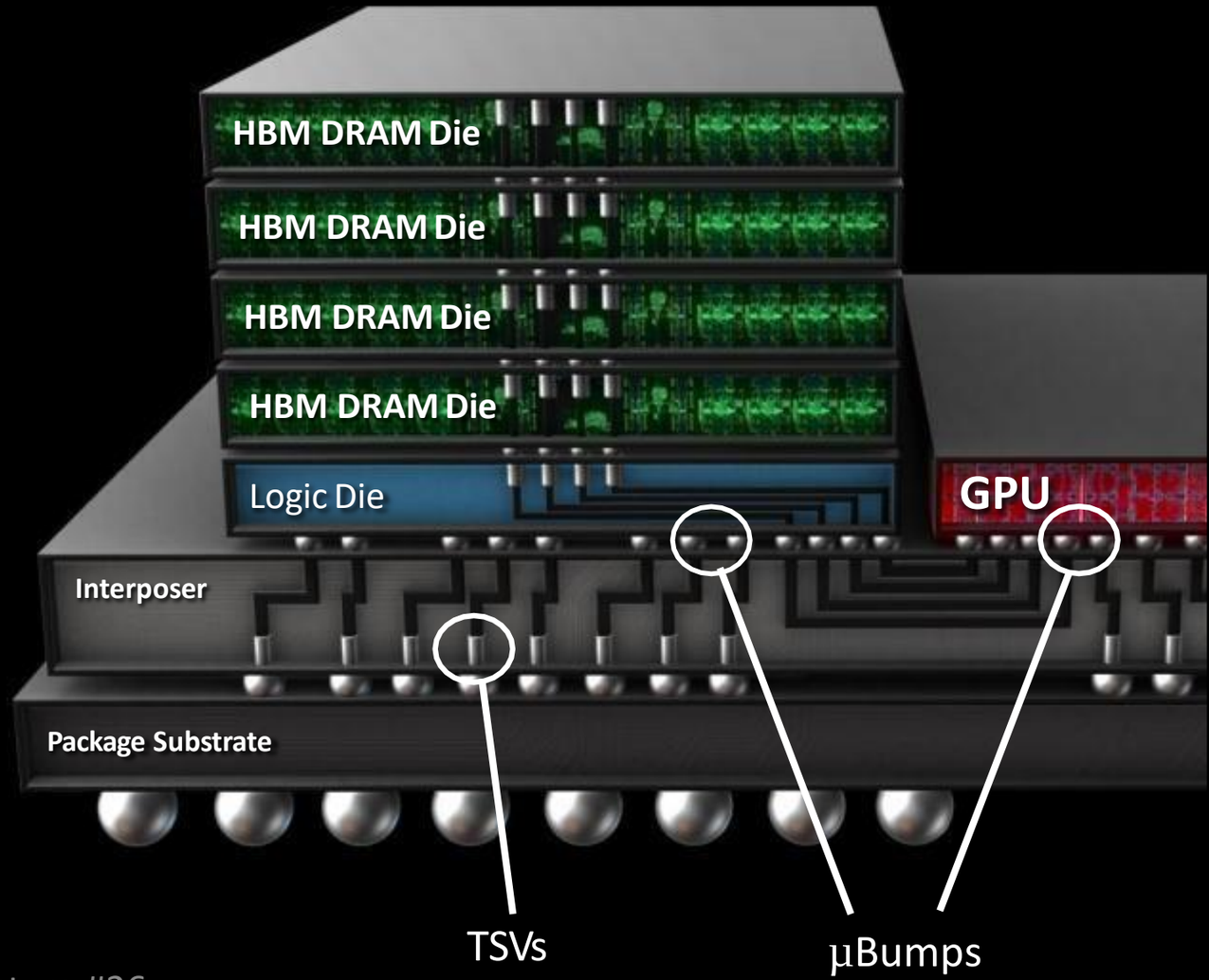


AMD

DIE STACKING TECHNOLOGY

- Die stacking facilitates the integration of discrete dies
- 8.5 years of development by AMD and its technology partners

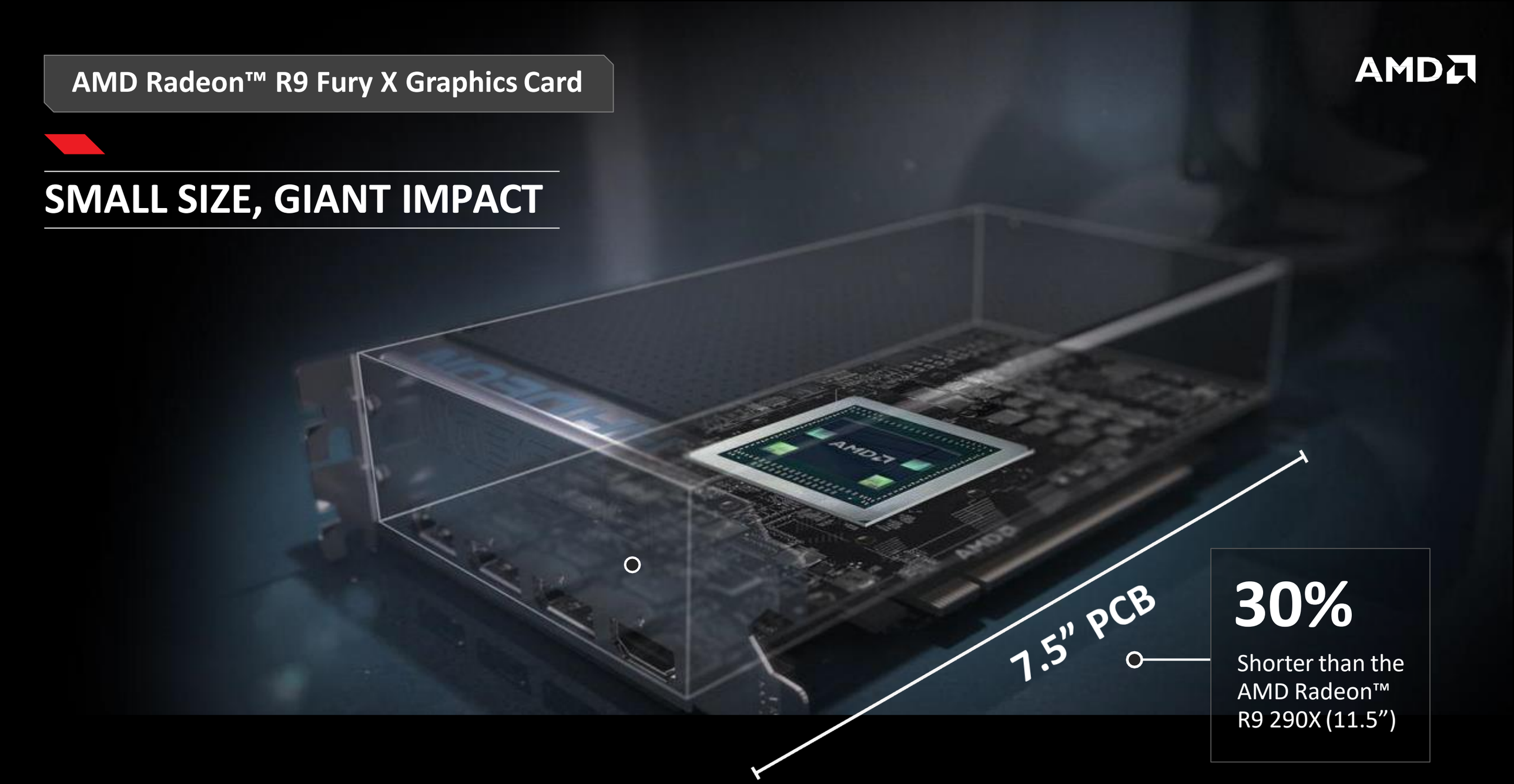




[Bryan Black, 2015, AMD]

AMD Radeon™ R9 Fury X Graphics Card

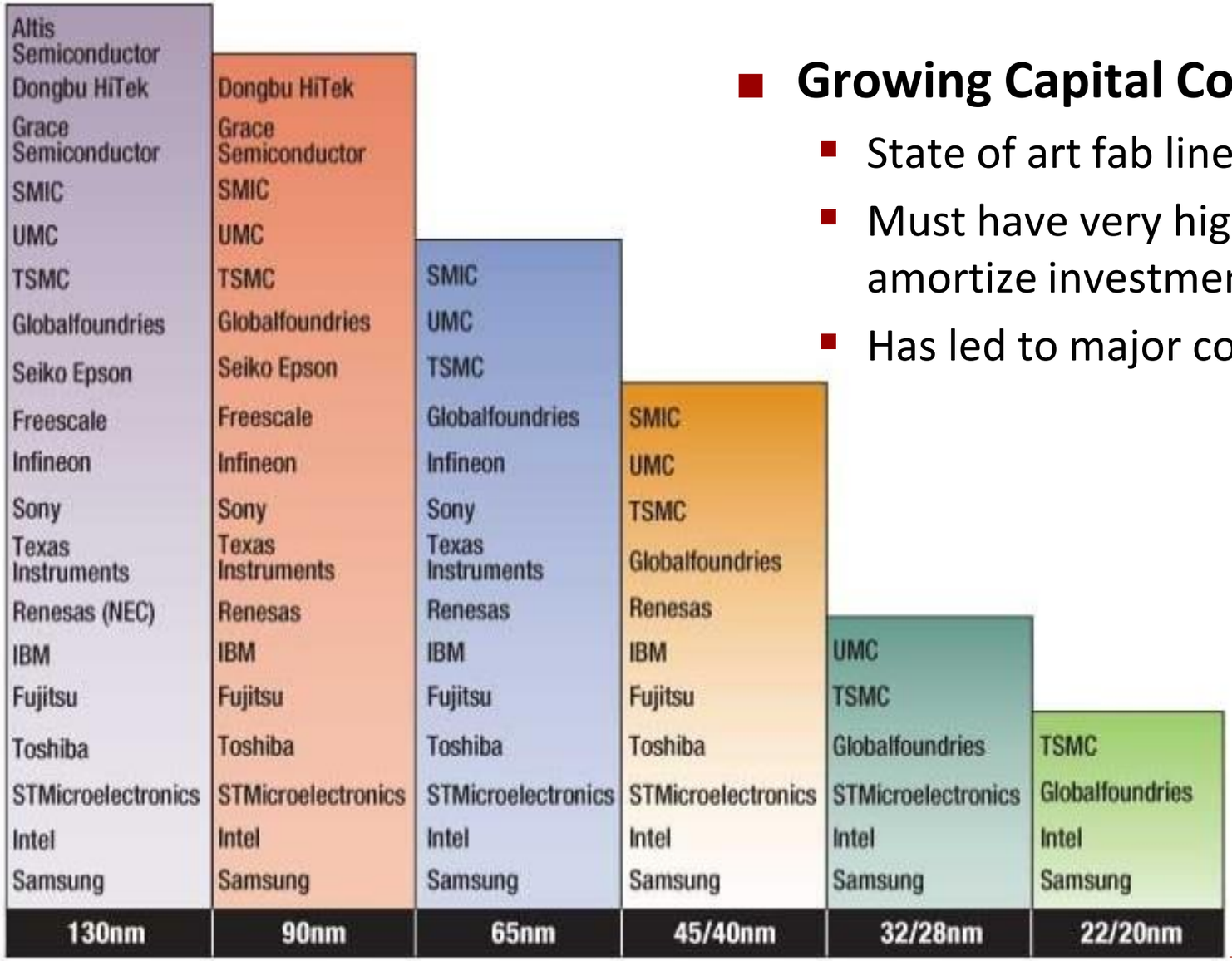
SMALL SIZE, GIANT IMPACT



Board shot shown for illustration purposes only. Final board design may differ.

[Bryan Black, 2015, AMD]

Key Challenge to Moore's Law: Economic



- **Growing Capital Costs**
  - State of art fab line ~$20B
  - Must have very high volumes to amortize investment
  - Has led to major consolidations

Meeting Power Constraints

- 2 B transistors
- 2 GHz operation
- 1—5 W

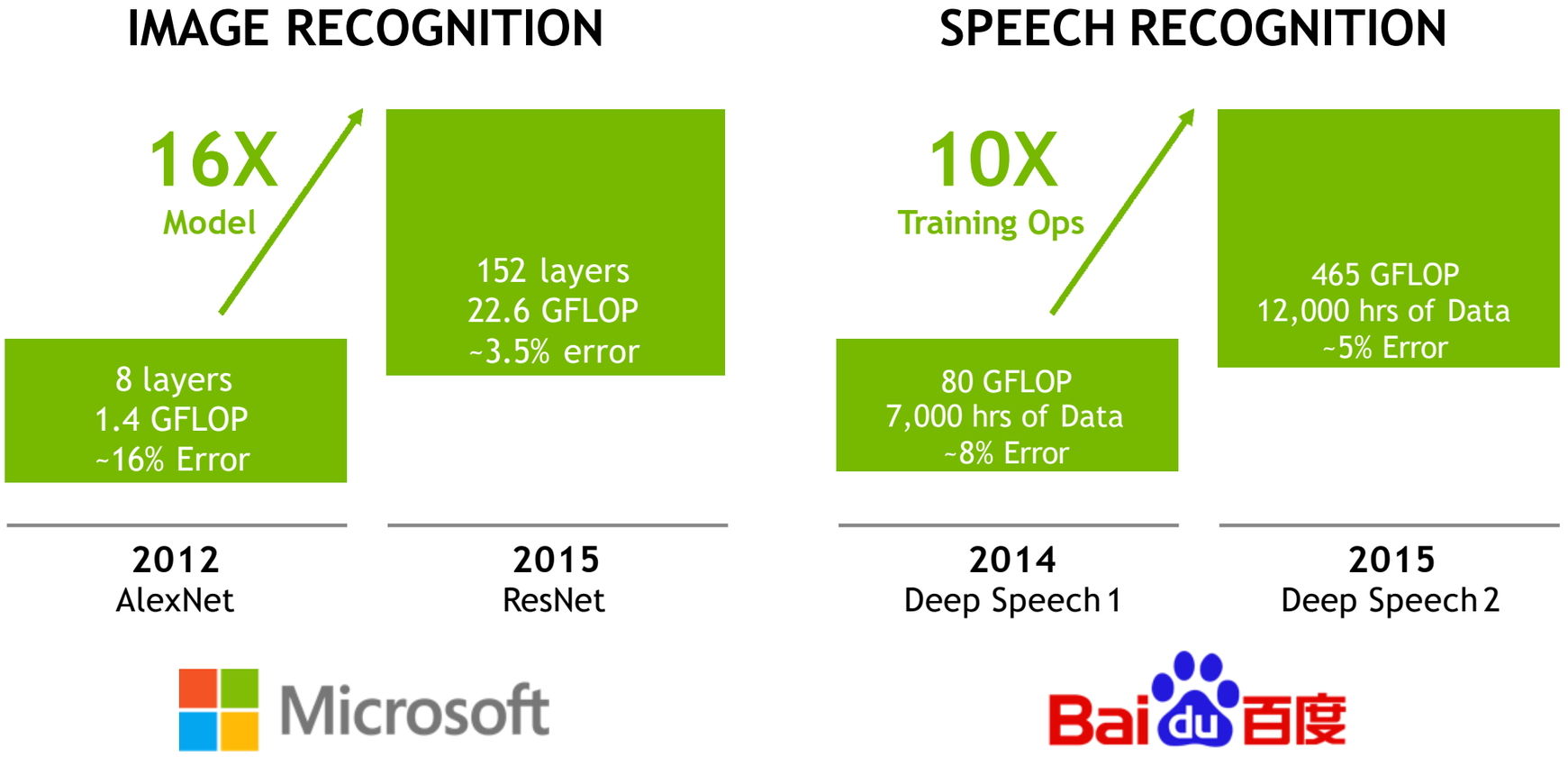
Can we increase number of devices by 500,000x without increasing power requirement?

- 64 B neurons
- 100 Hz operation
- 15—25 W
  - Liquid cooling
  - Up to 25% body's total energy consumption

Final Thoughts

- **Compared to future, past 50 years will seem fairly straightforward**
  - 50 years of using photolithography to pattern transistors on two-dimensional surface
- **Questions about future integrated systems**
  - Can we build them?
  - What will be the technology?
  - Are they commercially viable?
  - Can we keep power consumption low?
  - What will we do with them?
  - How will we program / customize them?

Deep Learning Models are Getting Larger



Dally, NIPS'2016 workshop on Efficient Methods for Deep Neural Networks

CPUs for DNN Training

Intel Knights Landing (2016)

- 7 TFLOPS FP32
- 16GB MCDRAM– 400 GB/s
- 245W TDP
- 29 GFLOPS/W (FP32)
- 14nm process

Knights Mill: next gen Xeon Phi "optimized for deep learning"

Intel announced the addition of new vector instructions for deep learning (AVX512-4VNNIW and AVX512-4FMAPS), October 2016

Slide Source: Sze et al Survey of DNN Hardware, MICRO'16 Tutorial.
Image Source: Intel, Data Source: Next Platform

GPUs for DNN Training

Nvidia PASCAL GP100 (2016)

- 10/20 TFLOPS FP32/FP16
- 16GB HBM – 750 GB/s
- 300W TDP
- 67 GFLOPS/W (FP16)
- 16nm process
- 160GB/s NV Link

Slide Source: Sze et al Survey of DNN Hardware, MICRO'16 Tutorial.
Data Source: NVIDIA

GPU Systems for DNN Training

Nvidia DGX-1 (2016)

- 170 TFLOPS
- 8× Tesla P100, Dual Xeon
- NVLink Hybrid Cube Mesh
- Optimized DL Software
- 7 TB SSD Cache
- Dual 10GbE, Quad IB 100Gb
- 3RU – 3200W

Slide Source: Sze et al Survey of DNN Hardware, MICRO'16 Tutorial.
Data Source: NVIDIA

WHAT IS IT?

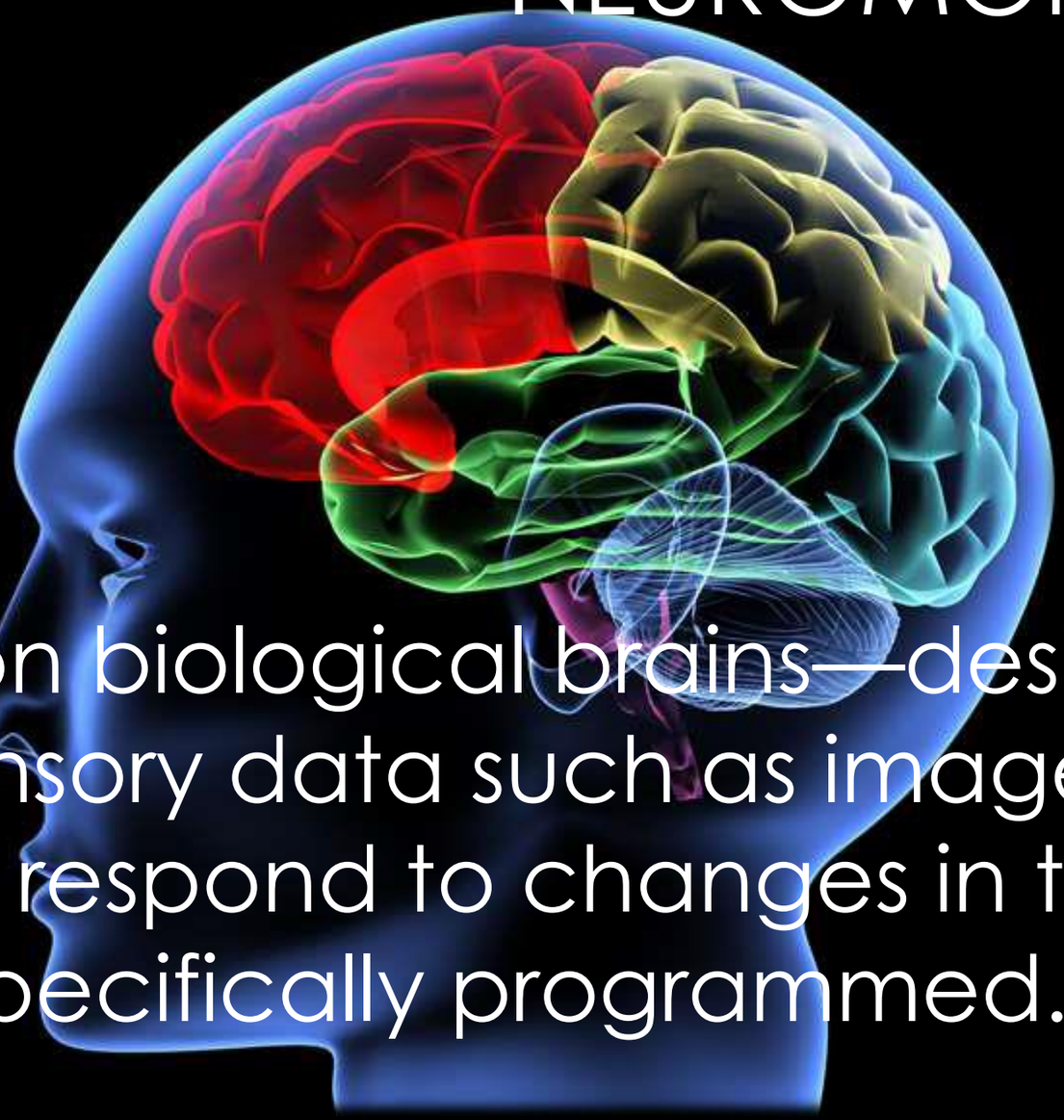
"Neuromorphic engineering, also known as neuromorphic computing started as a concept developed by Carver Mead in the late 1980s, describing the use of very-large-scale integration (VLSI) systems containing electronic analogue circuits to mimic neurobiological architectures present in the nervous system."

- Easton, 2015

WHAT IS IT? (SIMPLER)

"Neuromorphic engineering/computing is a new emerging interdisciplinary field which takes inspiration from biology, physics, mathematics, computer science and engineering to design hardware/physical models of neural and sensory systems."

NEUROMORPHIC CHIPS



- Modeled on biological brains—designed to process sensory data such as images and sound and respond to changes in that data in ways not specifically programmed.

MORAVEC'S PARADOX

- Sensory information processing is extremely easy for brains but extremely hard for modern computers; whereas symbolic information processing is comparably hard for brains but extremely easy for modern computers.

18-600 Foundations of Computer Systems

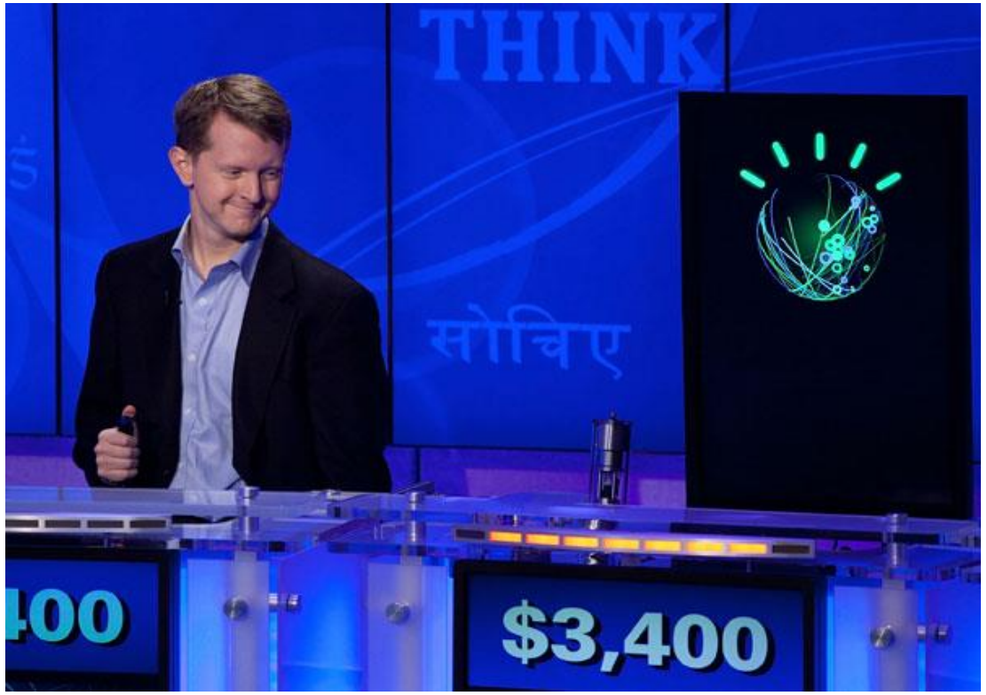
Introduction to "Neuromorphic Computing"

John P. Shen (with contribution from Mikko Lipasti)
December 4, 2017

Mikko H. Lipasti
Professor, Electrical and Computer Engineering
University of Wisconsin – Madison

Electrical & Computer ENGINEERING

Ken Jennings vs. IBM Watson

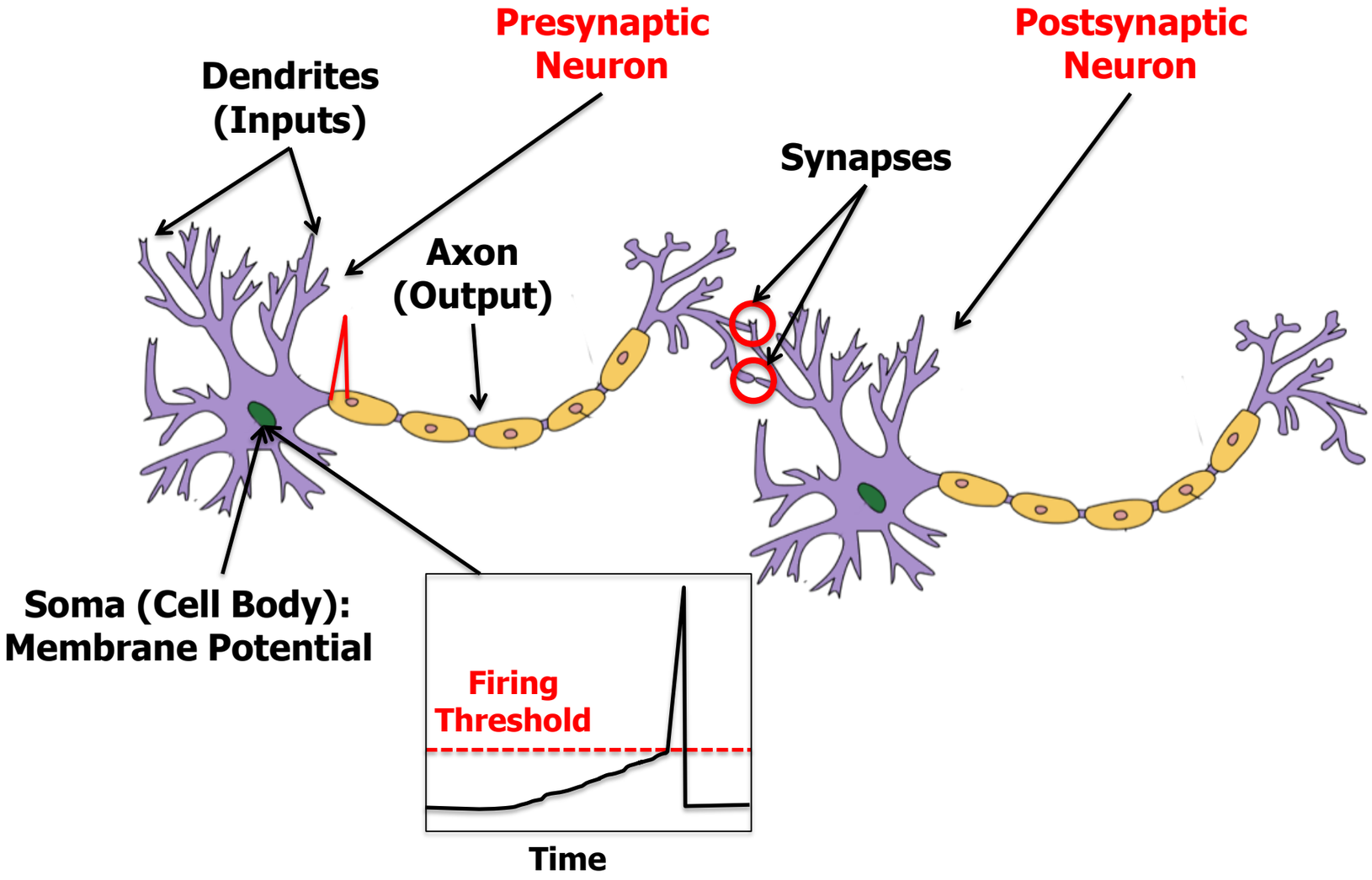


Ken ("baseline")	Watson
Pretty good at Jeopardy (also, life)	Pretty good at Jeopardy
400g gray matter	10 racks, 15TB DRAM, 2880 CPU cores, 80 TFLOPs
20W	200KW
1 lifetime of experience	100 person-years to develop

DNNs vs. BNNs

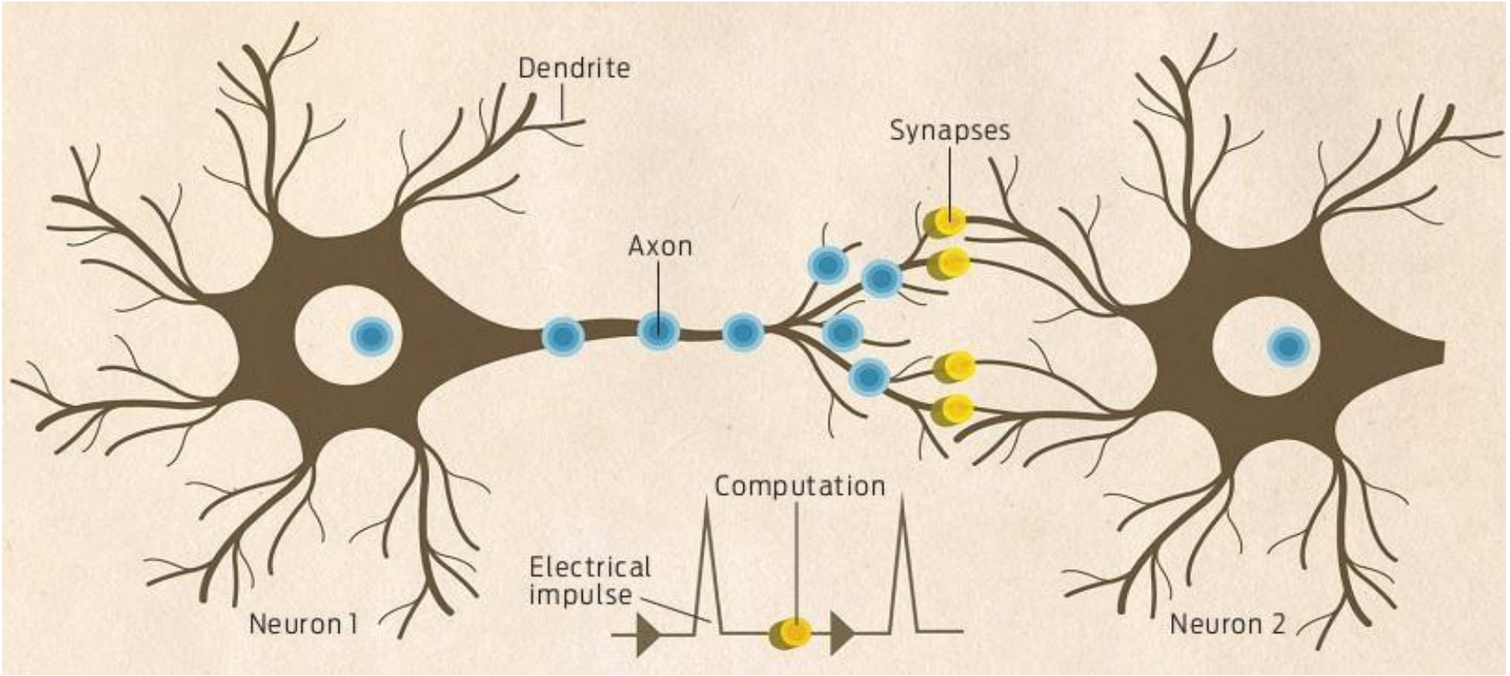
- Deep NNs (usually convolutional)
  - Rely on artificial perceptron neuron model
  - Trained using stochastic gradient descent (backprop): no biological bases
  - Must "experience" everything (training cost)
  - Can have superhuman performance
- Biological (spiking) NNs
  - Energy efficient, powerful
  - Training/learning/development poorly understood
  - Currently not very useful or practical…yet

Neurons - Introduction



Neuromorphic Architectures

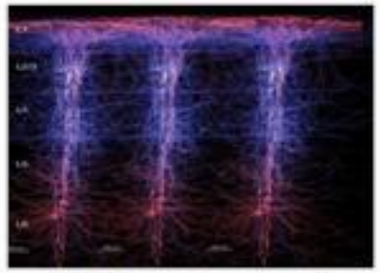
- Computer architectures that are similar to biological brains; computer architectures that implement artificial neural networks in hardware.
- Functional units are composed of neurons, axons, synapses, and dendrites.
- Synapses are connections between two neurons
  - Remembers previous state, updates to a new state, holds the weight of the connection
- Axons and dendrites connect to many neurons/synapses, like long range bus.



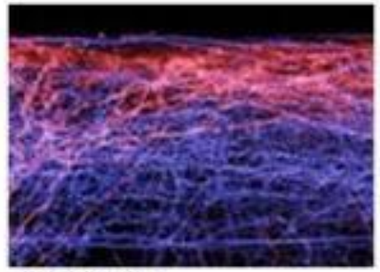
Biological vs. Hardware

~10^{10} synapses/cm^2

~10^{6} neurons/cm^2



~10^{4} neurons/ cortical column

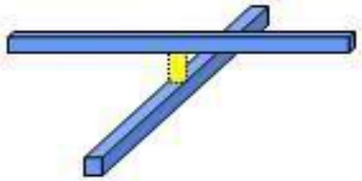


~5X10^{8} long range axons @ 1 Hz

Biological Brain

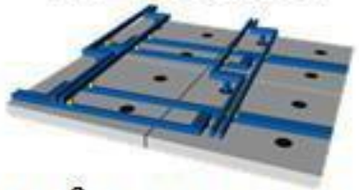
Synapse | Neurons | Microcircuit | Long-range interconnects | Brain

CROSSBAR JUNCTION



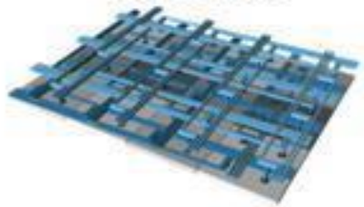
~10^{10} intersections/ cm^2 @ 100 nm pitch

CMOS SUBSTRATE



5X10^{8} transistors/ cm^2 @ 500 transistors/ neuron

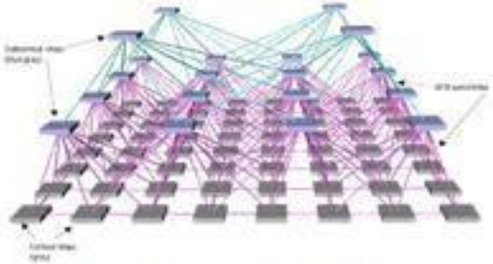
LAMINAR CIRCUIT



Layered cortical circuits with ~10^{6} neurons/cm^2

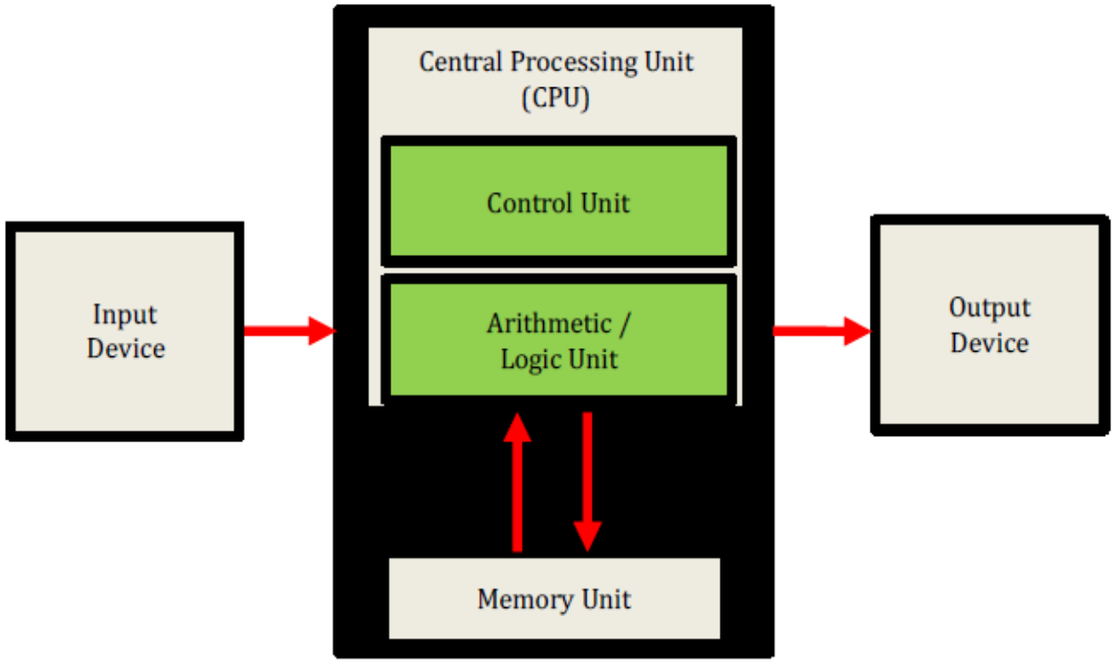
HIGH SPEED BUS

Multi-Gbit/sec digital comms

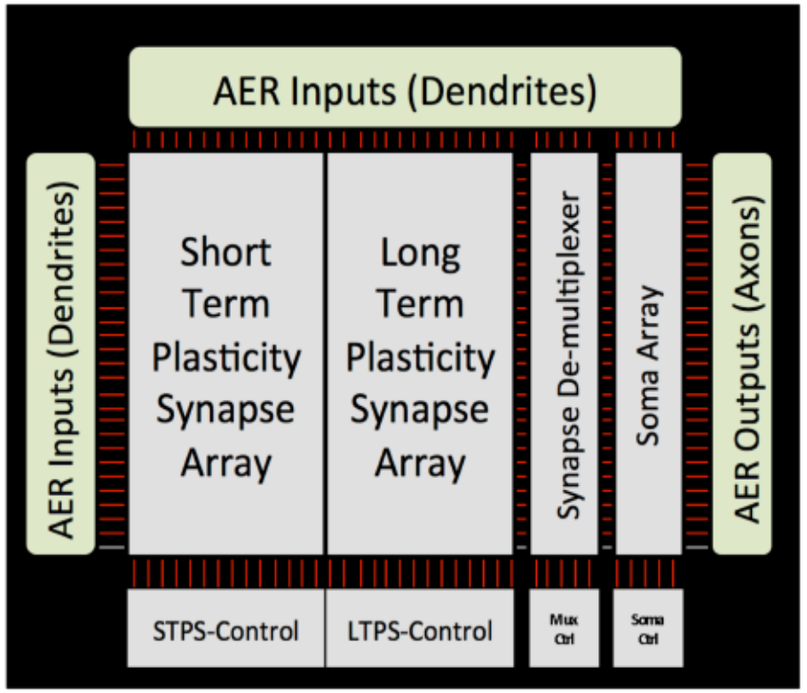


Electronic Brain

von Neumann Architecture



Neuromorphic Architecture



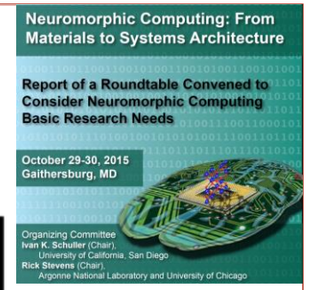


Figure 1. Comparison of high-level conventional and neuromorphic computer architectures. The so-called "von Neumann bottleneck" is the data path between the CPU and the memory unit. In contrast, a neural network based architecture combines synapses and neurons into a fine grain distributed structure that scales both memory (synapse) and compute (soma) elements as the systems increase in scale and capability, thus avoiding the bottleneck between computing and memory.

	Biology	Silicon	Advantage
Speed	1 msec	1 nsec	1,000,000x
Size	1μm - 10μm	10nm - 100nm	1,000x
Voltage	~ 0.1V	V_{dd} ~1.0V	10x
Neuron Density	100K/mm^2	5k/mm^2	20x
Reliability	80%	< 99.9999%	1,000,000x
Synaptic Error Rate	75%	~ 0%	>10^9
Fan-out (-in)	10^3-10^4	3-4	10,000x
Dimensions	Pseudo 3D	Pseudo 3D	Similar
Synaptic Op Energy	~ 2 fJ	~10pJ	5000x
Total Energy	10 Watt	>>10^3 Watt	100,000x
Temperature	36C - 38C	5C - 60C	Wider Op Range
Noise effect	Stochastic Resonance	Bad	
Criticality	Edge	Far	

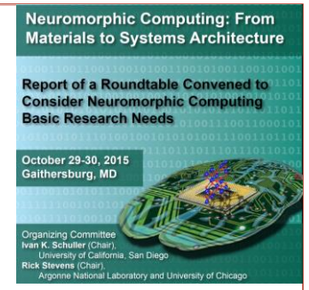
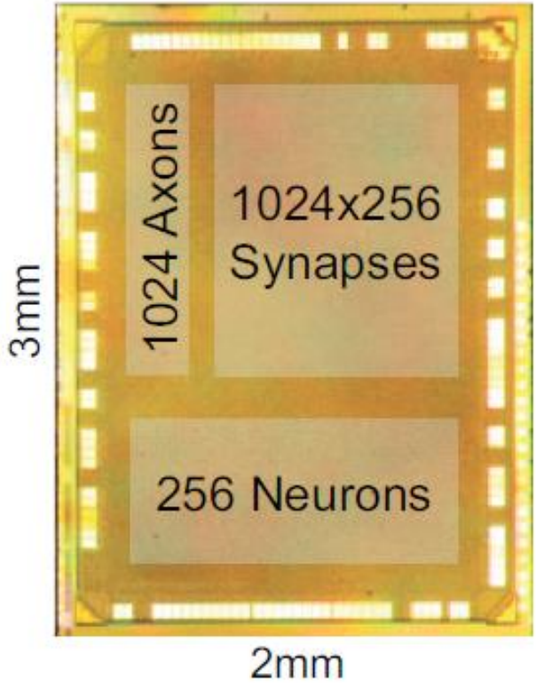
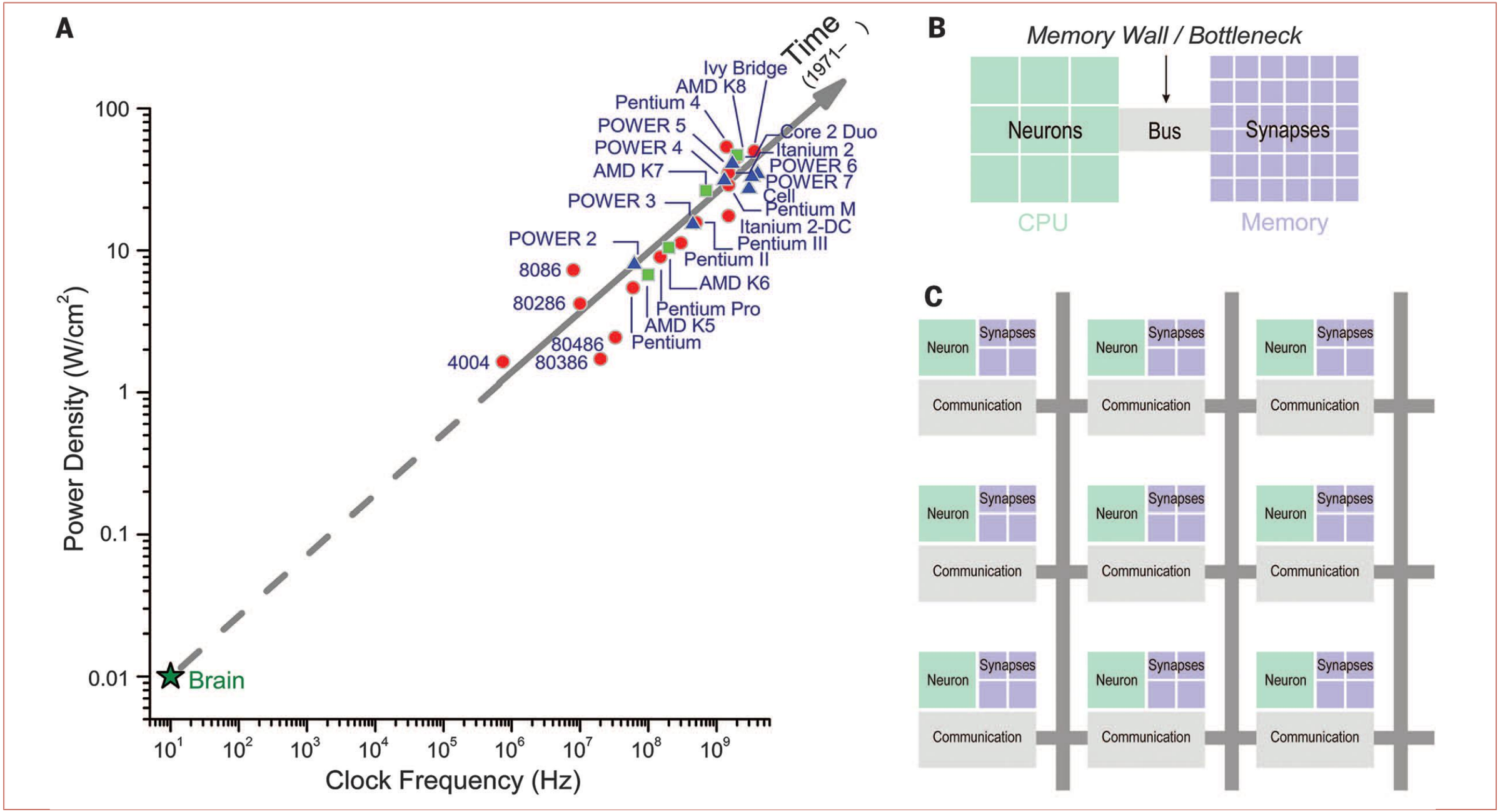


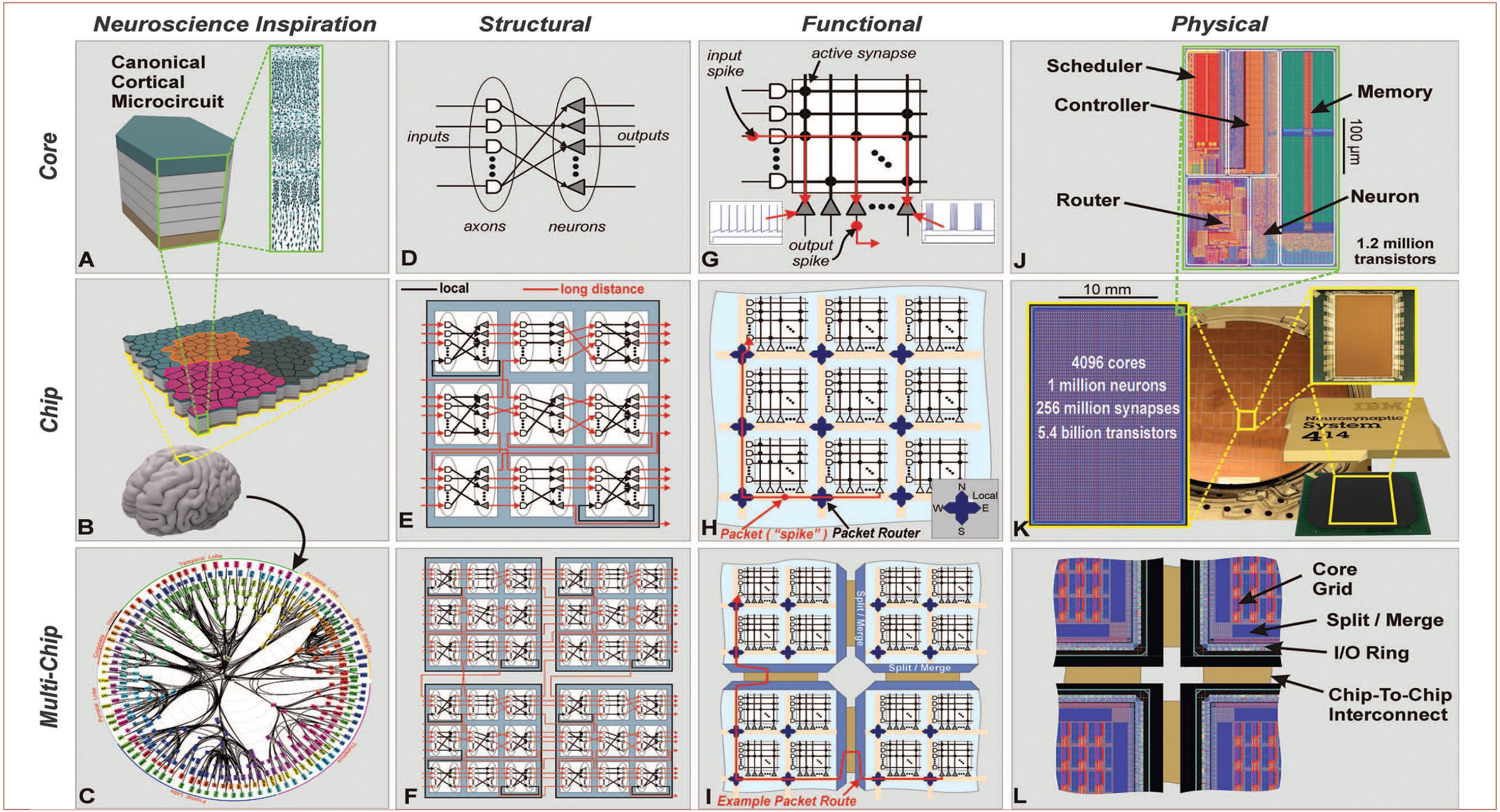
Table 1. Comparison of biological and silicon based systems. This table shows a comparison of neurons built with biology to equivalent structures built with silicon. Red is where biology is ahead; black is where silicon is ahead. The opportunity lies in combining the best of biology and silicon.

IBM's TrueNorth

- Digital spiking Neurosynaptic Core Neurons (NCNs)
  - LP CMOS, standard digital logic
  - 256 neurons/core on 4.2mm^2
- "Biologically competitive" energy
  - Few parameters/neuron
  - Binary synapses
  - Linear, no transcendental functions
  - 1kHz operating frequency of NCNs
  - 45pJ/spike
- Scaled to 1M neurons/chip




A
Power Density (W/cm^2)
100
10
1
0.1
0.01
Brain
10^1 10^2 10^3 10^4 10^5 10^6 10^7 10^8 10^9
Clock Frequency (Hz)
Time (1971–)
Ivy Bridge
AMD K8
Pentium 4
POWER 5
POWER 4
AMD K7
POWER 3
POWER 2
8086
80286
4004
80486
80386
Pentium
AMD K5
Pentium Pro
AMD K6
Pentium II
Pentium III
Itanium 2-DC
Pentium M
Cell
POWER 7
POWER 6
Itanium 2
Core 2 Duo
B
Memory Wall / Bottleneck
Neurons
Bus
Synapses
CPU
Memory
C
Neuron
Synapses
Communication


Neuroscience Inspiration
Structural
Functional
Physical
Core
Chip
Multi-Chip
Canonical Cortical Microcircuit
A
inputs
outputs
axons
neurons
D
input spike
active synapse
output spike
G
Scheduler
Controller
Memory
100 μm
Router
Neuron
1.2 million transistors
J
B
local
long distance
E
10 mm
4096 cores
1 million neurons
256 million synapses
5.4 billion transistors
IBM
Neurosynaptic System 4x4
K
N
Local
W
E
S
H
Packet ("spike")
Packet Router
C
F
Split / Merge
Split / Merge
I
Example Packet Route
Core Grid
Split / Merge
I/O Ring
Chip-To-Chip Interconnect
L

A Very Simple Digital Neuron

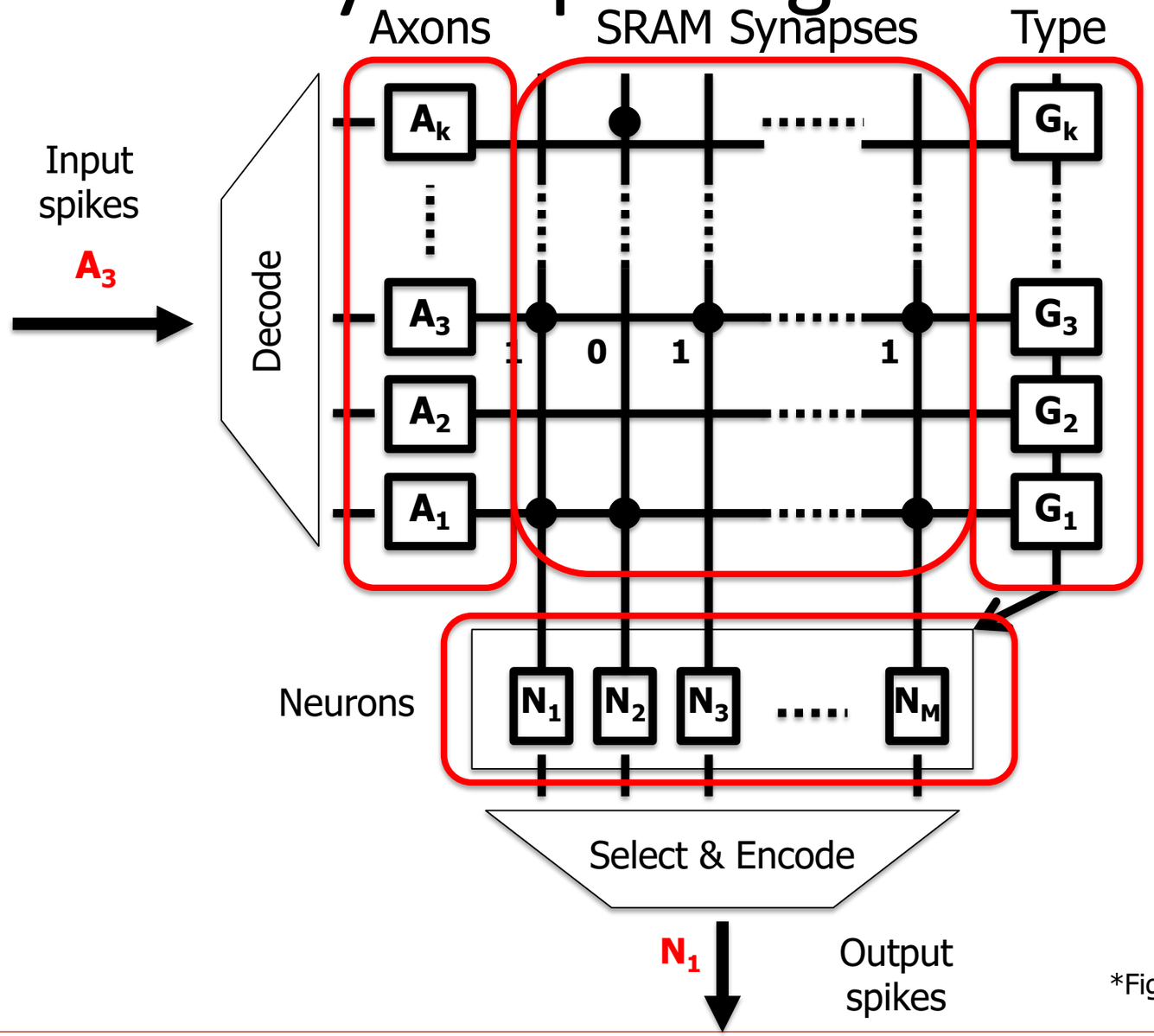
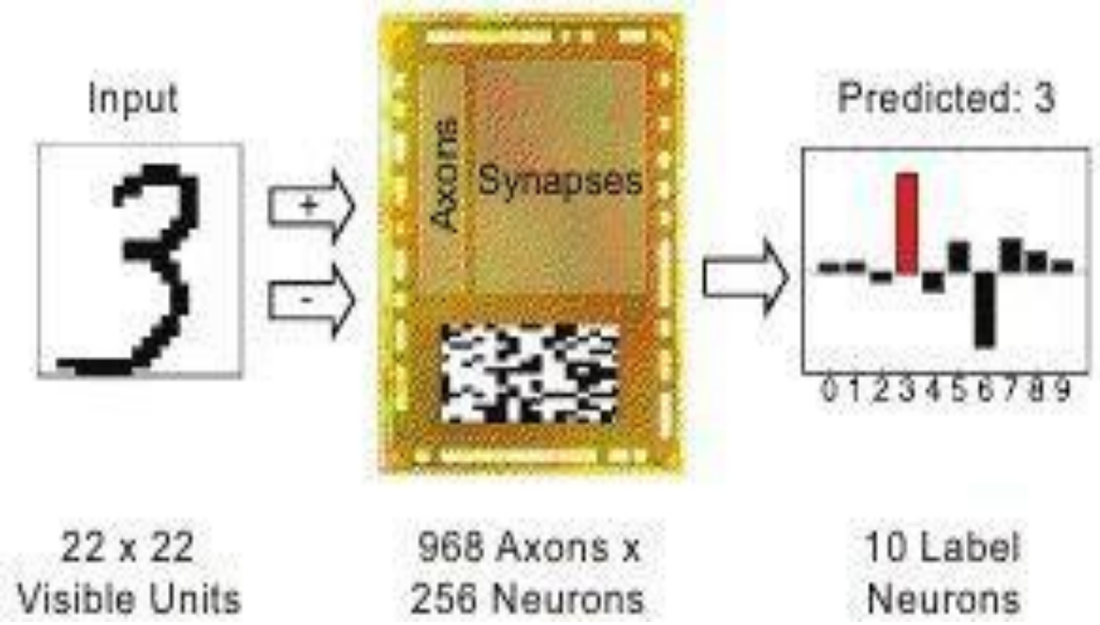


*Figure adapted from Merolla et al.

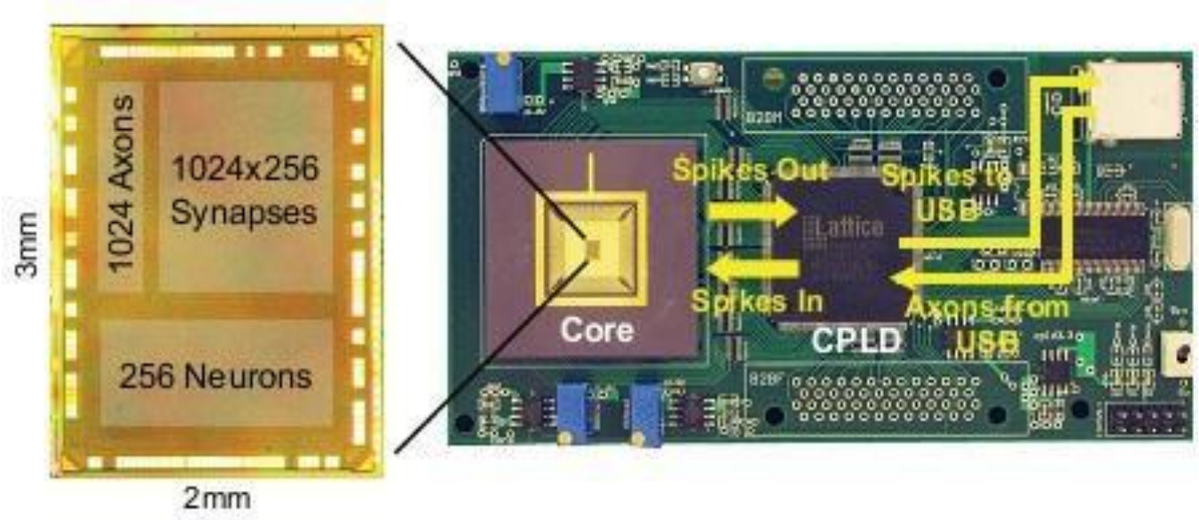
Example Pattern Recognition



- 10 networks each trained to identify one digit as its "pattern"
  - Each receives the input in parallel, and reaches output at the same time
- Every neuron fires proportional to how strong it thinks the input matches the internalized pattern
- The strongest output wins, so "3" is the result

IBM's TrueNorth Chip

- IBM along with DARPA funding, created 2 brain-like chips
- These chips are not as complex as the human, in fact they have less neurons and synapses than a snail
- Scaling is tough, as increasing # of neurons exponentially increases # of synapses
  - at 500 transistors/synapse, this gets large fast
- What sets these chips apart is that they were done on current CMOS technology



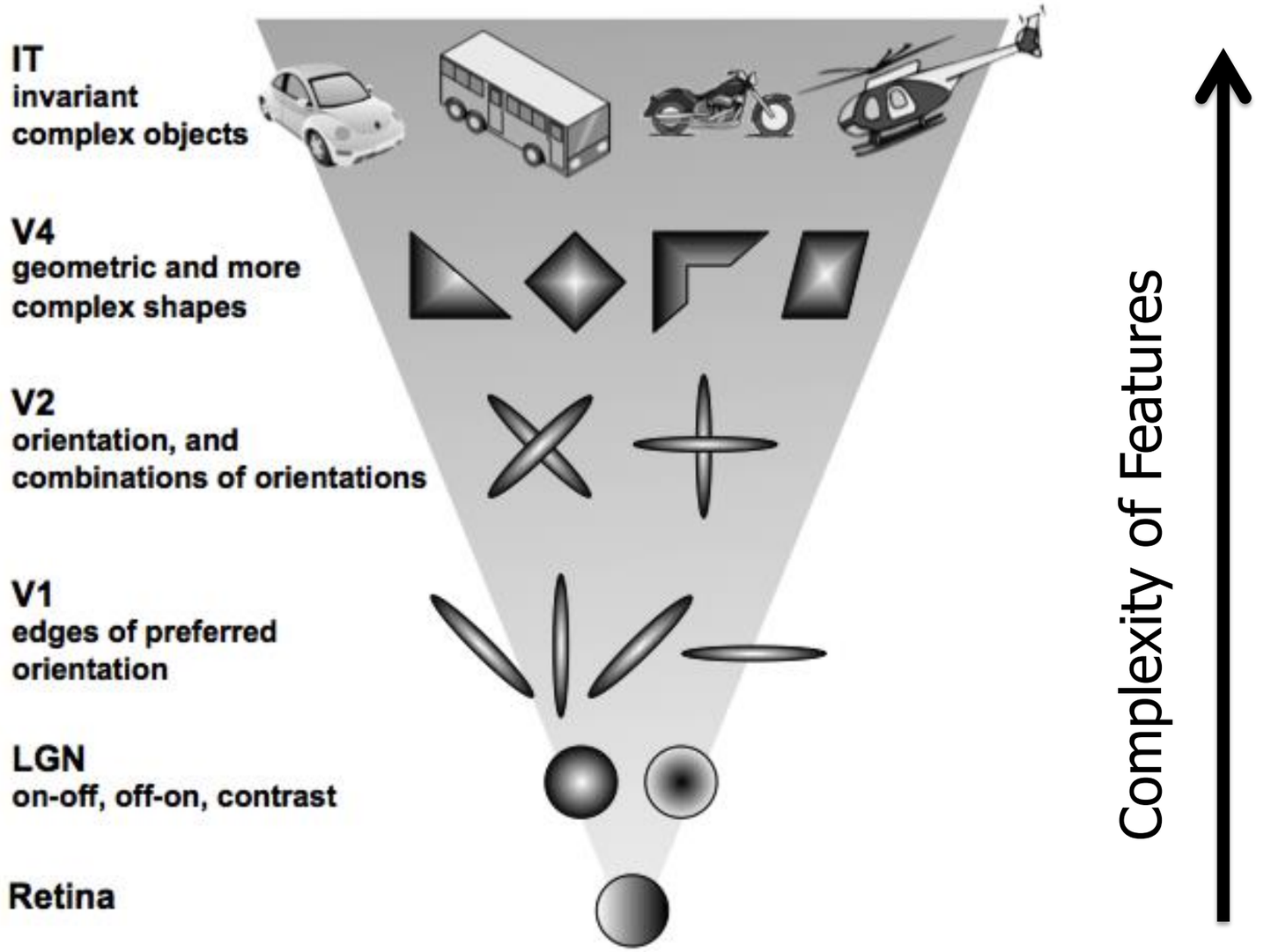
Spiking Neural Network Hardware: TrueNorth

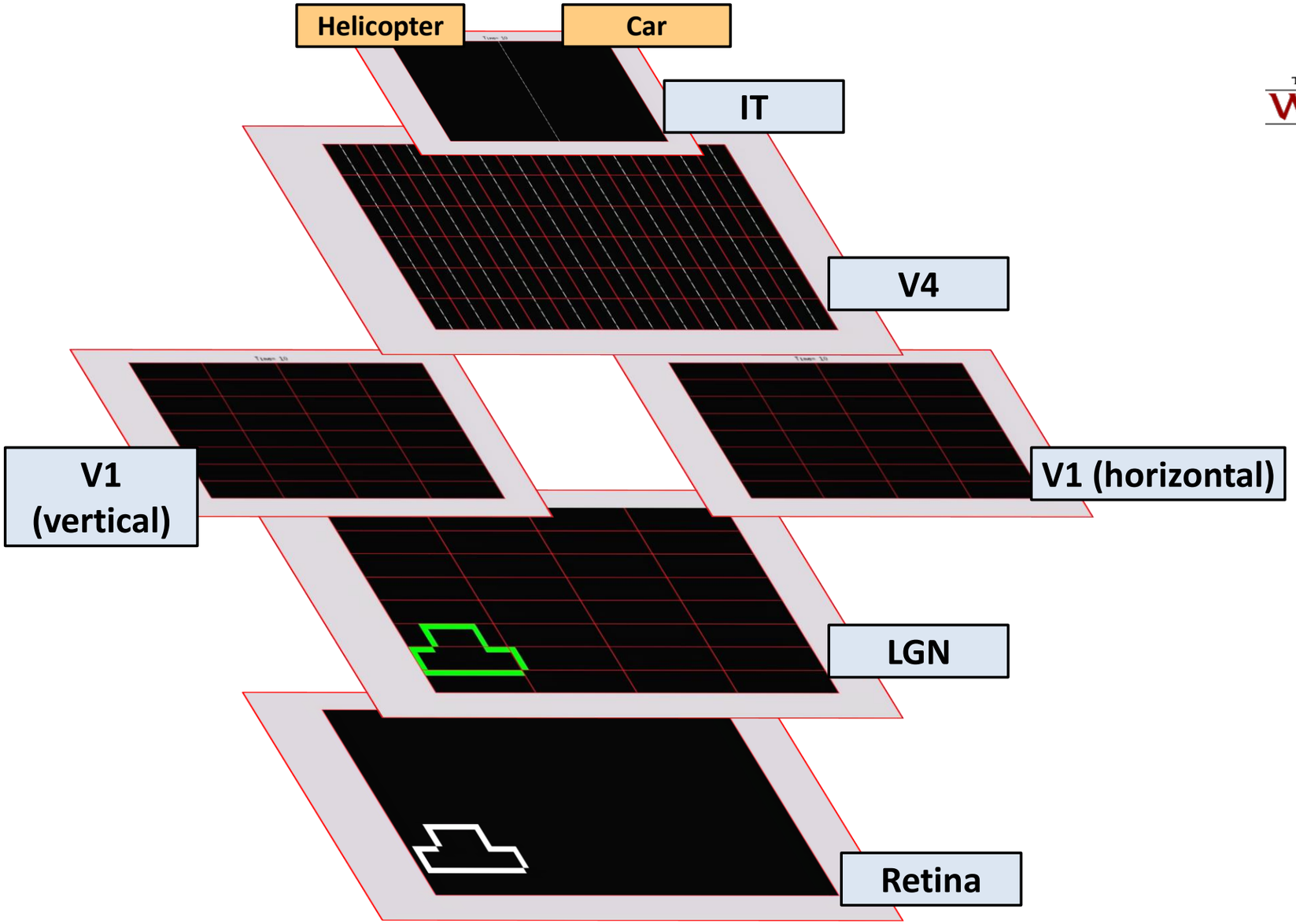
- Biologically inspired neurons where computations happen via spikes.
- Goal: Hardware substrate that mimic this computation scheme and be energy efficient
- IBM TrueNorth: Abstracts away relevant computing features of biological neurons and only keeps them.
- TrueNorth architecture can be scaled up.
- ns1e (1-Chip system) -> ns16e (16-chip system)
- Real hardware (ns1e) prototype.

ns1e: UW-Madison

ns16e: courtsey LLNL

Visual Cortex: Possible?

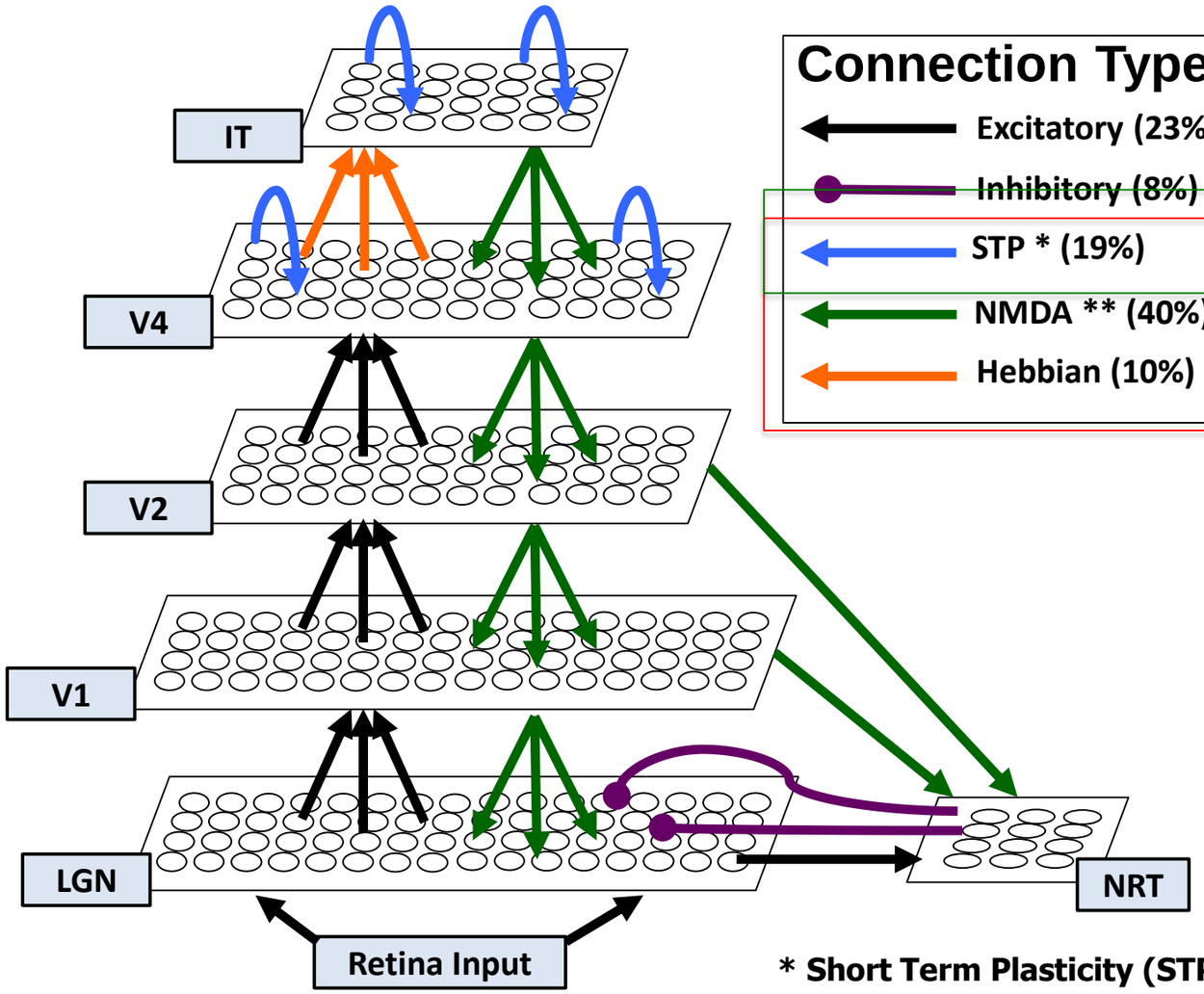



Helicopter
Car
IT
V4
V1
(vertical)
V1 (horizontal)
LGN
Retina

Visual System NNet (VSNN)

- 100,000 modeled neurons
- Applications
  - Invariant object recognition
  - Pattern completion
  - Motion detection/tracking/prediction
  - Noise filtering
- Requires complex neuronal behaviors
  - Hebbian learning, voltage-dependent behavior, long-timescale neuroreceptors (NMDA) etc. [Self et al., *Proc. NAS*, 2012]
  - Not implemented in IBM TrueNorth!

VSNN Architecture



Connection Type

- Excitatory (23%)
- Inhibitory (8%)
- STP * (19%)
- NMDA ** (40%)
- Hebbian (10%)

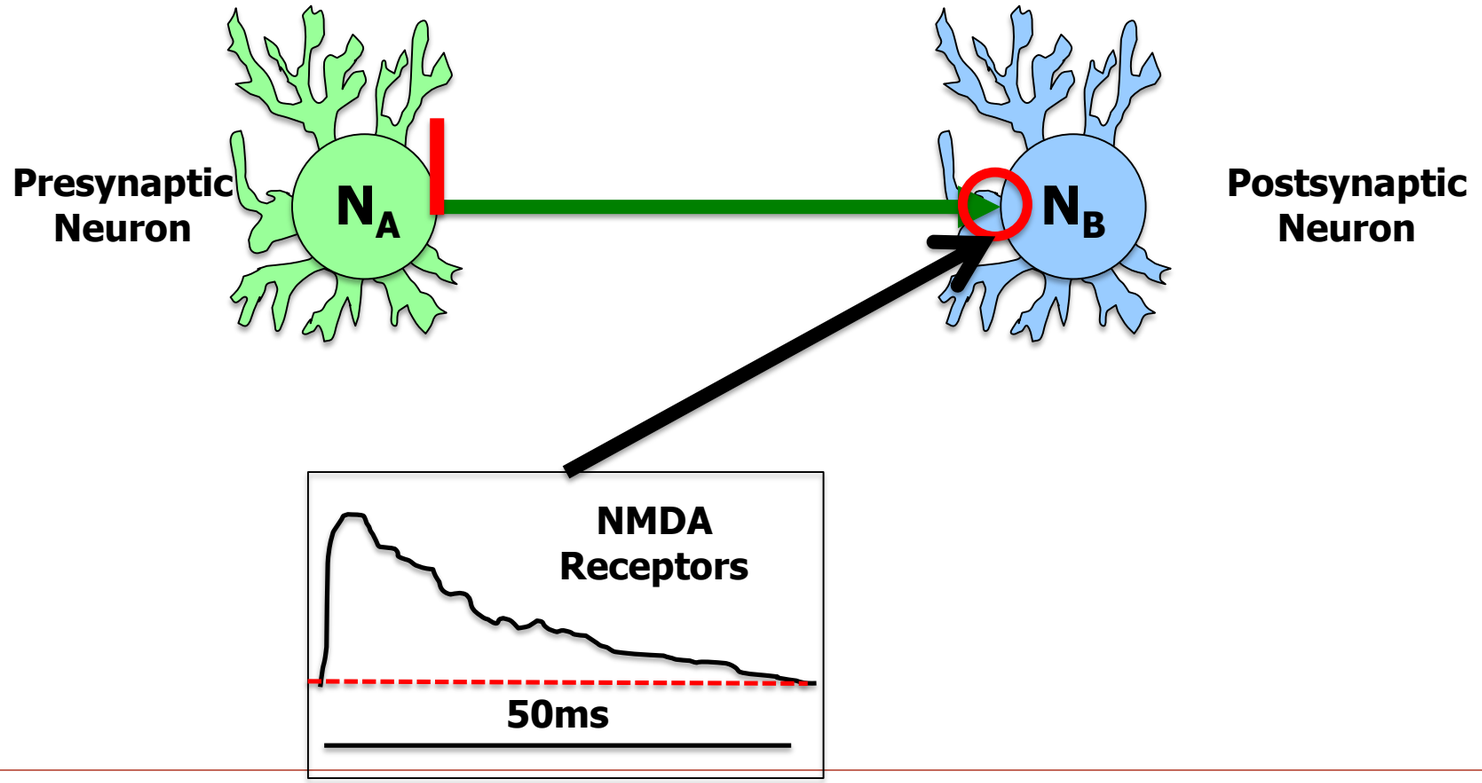
TN Compatible Complex Behaviors!

* Short Term Plasticity (STP) modulated synapse

** N-methyl D-aspartate (NMDA) modulated synapses

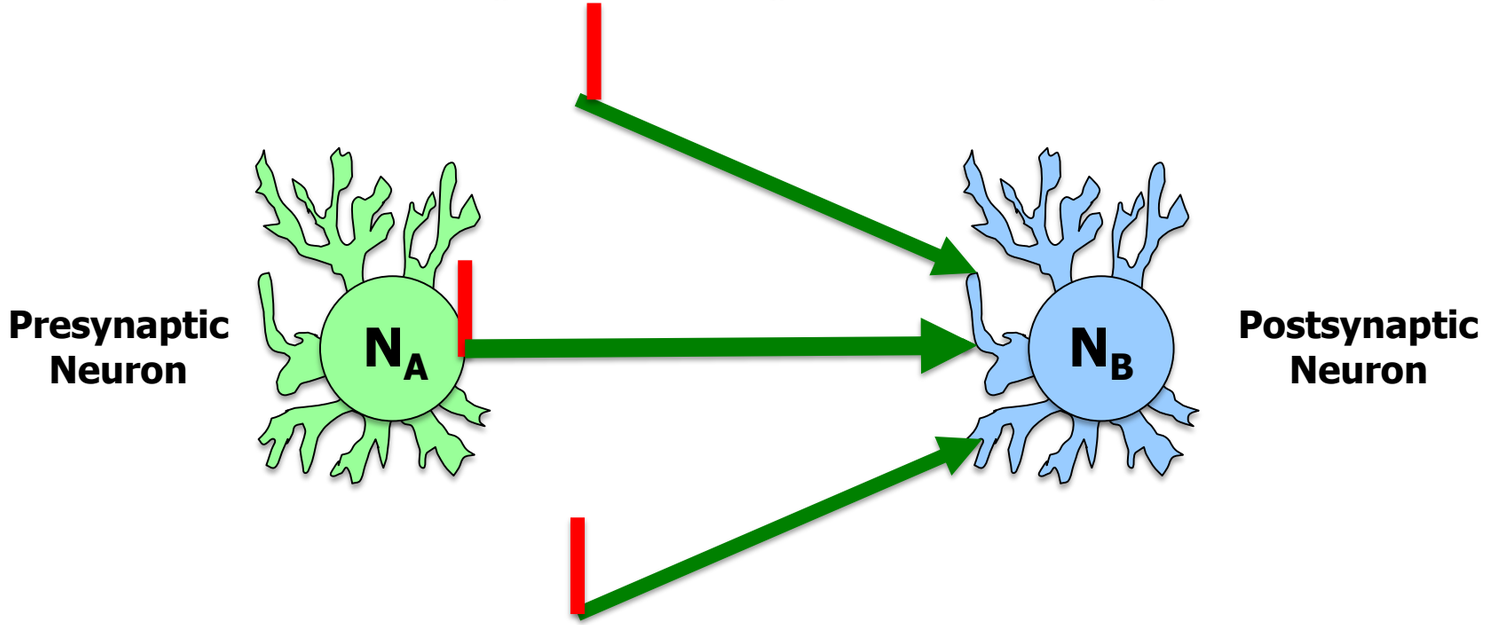
Neuromorphic Semantic Gap

- IBM NCNs are very simple (for efficiency)
- Biology incorporates numerous complex behaviors
  - NMDA receptor effects last much longer than 1ms



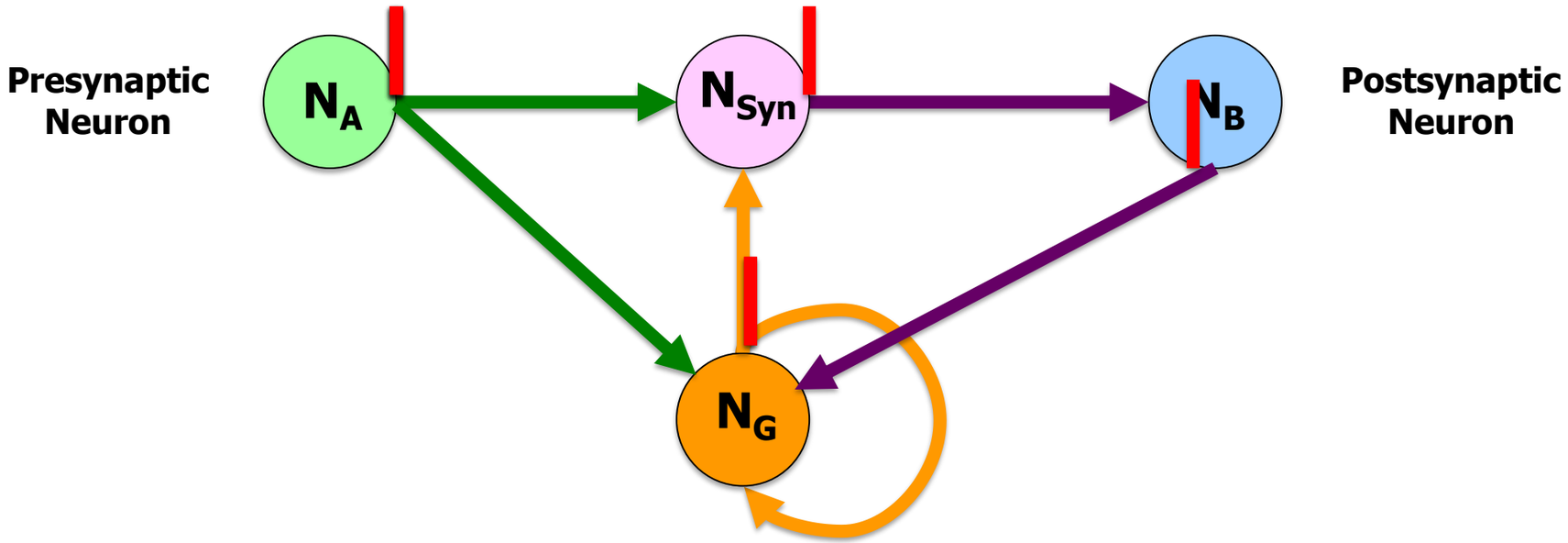
Semantic Gap – Plasticity

- IBM NCN does not support synaptic plasticity*
- Hebbian learning – "fire together, wire together"



*Seo et al. design features two simple online learning rules

Hebbian Learning Assembly

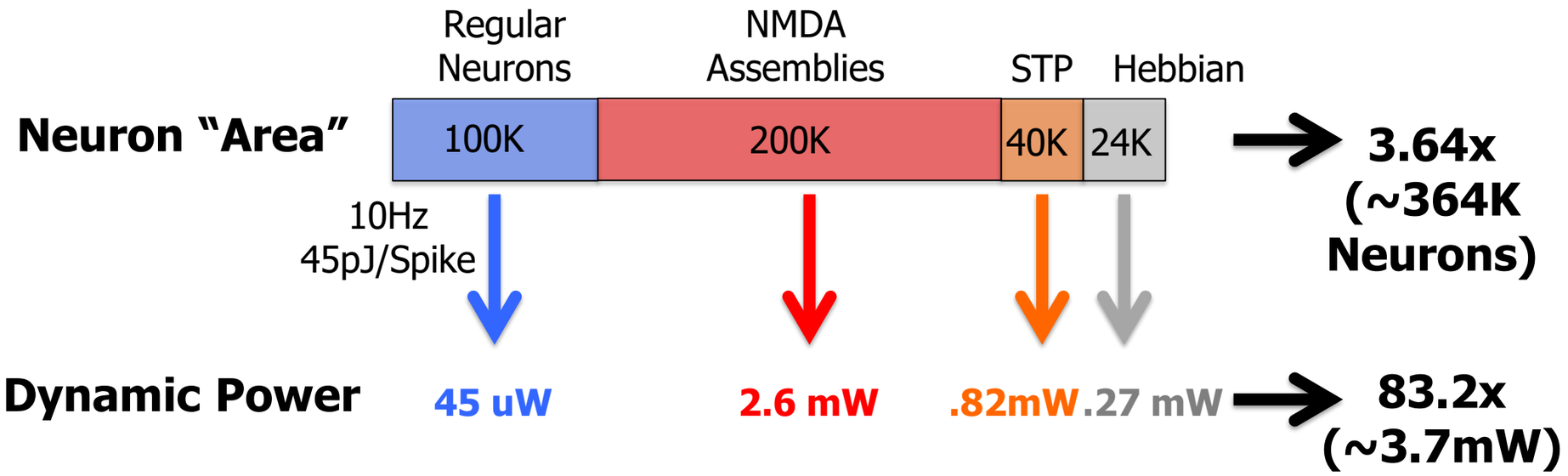


- 2 extra NCNs/synapse
- ~1000*45pJ power overhead/learned synapse

VSNN on Neurosynaptic Core

- "Compiler" replaces complex neurons/synapses with NCN assemblies
  - Deployable on Neurosynaptic Core hardware

- VSNN System Overheads



IBM TrueNorth Summary

- Complex behaviors can be emulated with simple LIF neurons
  - Cost is considerable
  - Minor HW changes could improve this (“latch”)
- Not the final answer on neuron complexity!
- Still at very small scale (100K biological neurons)
- Communication overhead already significant
  - Will dominate in larger networks

[Details in Nere et al., HPCA 2013]

Algorithmic Gap

- Neural substrates well suited for algorithms that were developed with them in mind:
  - CNNs, DNNs, etc.
  - Mainly targeted towards vision, classification
- Vast majority of existing algorithms developed for Von Neumann machines
  - Can we find a way to map at least some interesting subset of these algorithms?
  - Can we do so in a way that avoids *ad hoc* approaches for correctness?

Mapping to Hardware

- Mathematical formulation is very useful
  - Can prototype in e.g. Matlab
  - Can reason about it, prove properties, etc.
  - Often this work (algorithm development) has already been done by others
- Mapping to hardware-constrained neural substrate
  - Precision, range of weights
  - Precision, range of spike trains
  - Handling +/- values
  - Implementing arithmetic operators
  - Implementing state (memory)

How do Neural Circuits Emerge?

- Genetically programmed
  - Structure, interconnect, neuron parameters
    - Billions of them! Often specialized, tailored to particular task
  - Way too much information for human genome
- So, how then?
  - Is cortex a *blank slate?*
  - Structure, function emerges in response to stimulus
  - Relatively simple set of developmental rules are genetically programmed
  - Everything else is a product of experience

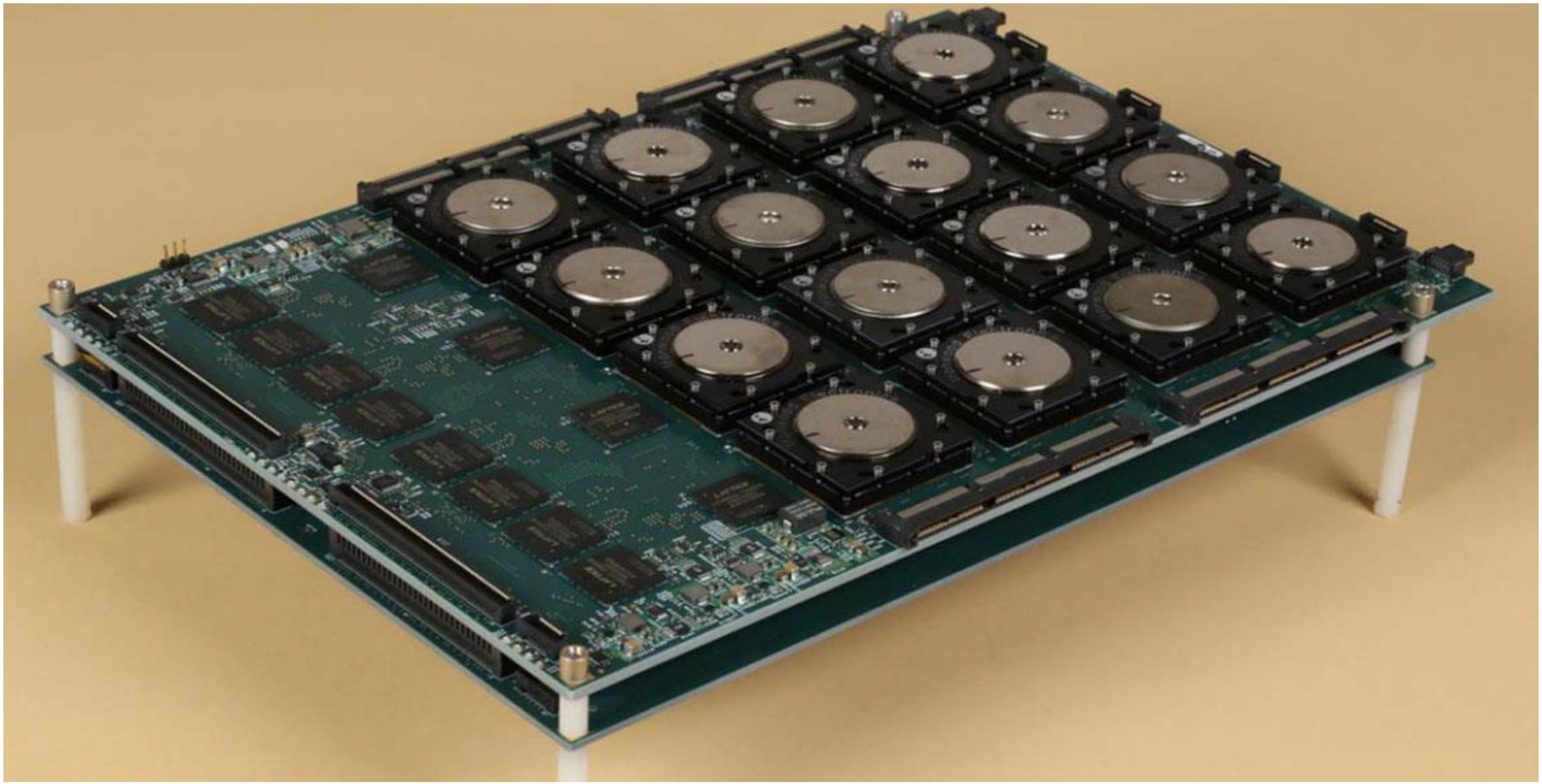
Conclusions

- We <u>can</u> deploy linear solvers on hardware-constrained neural substrates
  - IBM TrueNorth
- Looks like we can do interesting things
  - Mimic not just vision, but rewards, planning, recall
- Many practical challenges remain
- Many open questions

IBM TrueNorth Chip [2014-2015]

- TrueNorth is a neuromorphic CMOS chip produced by IBM in 2014.
- It is a manycore processor network on a chip, with 4096 cores,
- Each core simulating 256 programmable silicon "neurons" for a total of just over a million neurons.
- Each neuron has 256 programmable "synapses" that convey the signals between neurons.
- The total number of 268 million (2^{28}) programmable synapses.
- In terms of basic chip building blocks, its transistor count is 5.4 billion.
- TrueNorth is very energy-efficient, consuming 70 milliwatts and a power density that is 1/10,000th of conventional microprocessors.

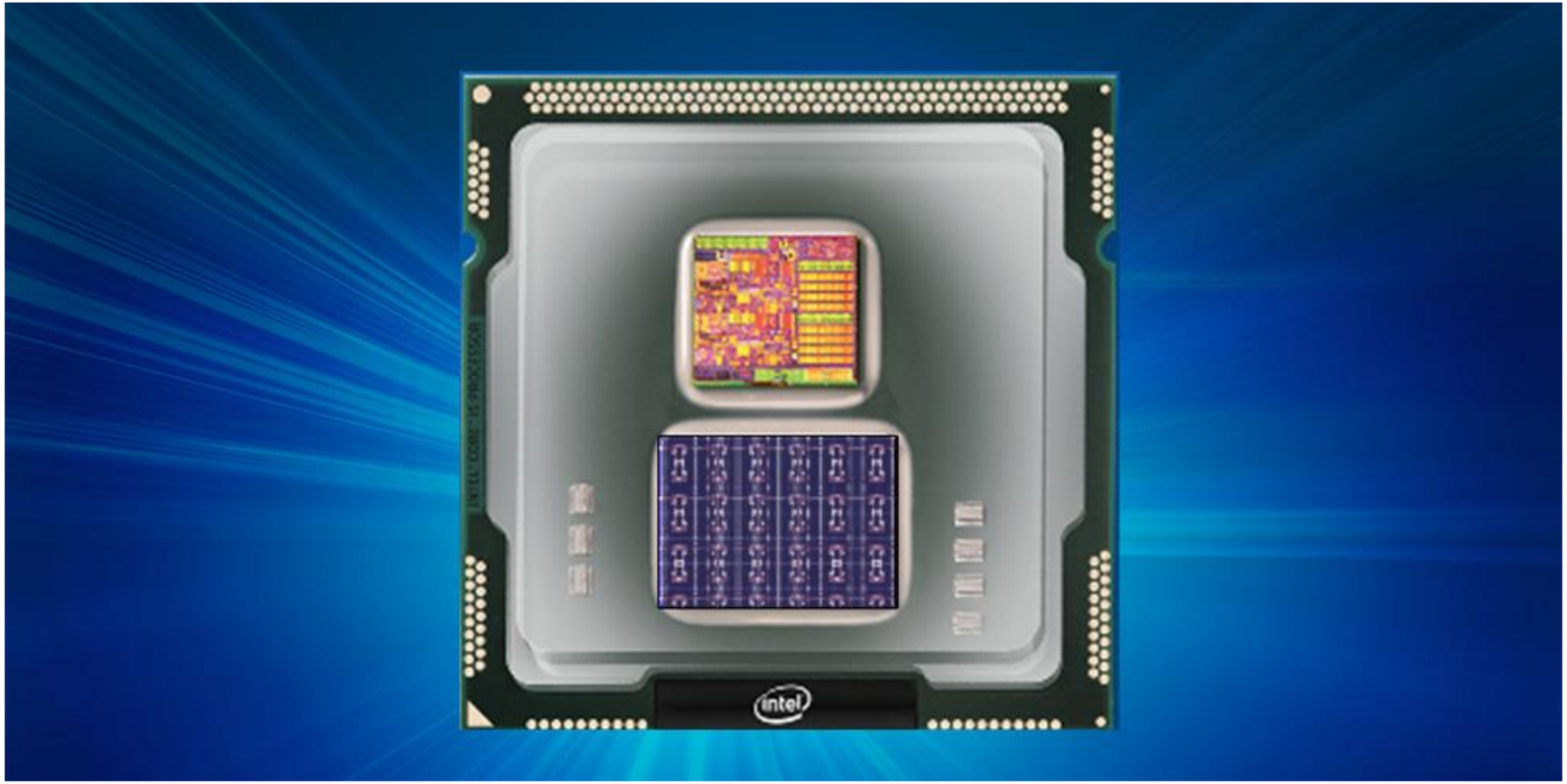
DARPA SyNAPSE 16 chip board with IBM TrueNorth



http://www.darpa.mil/NewsEvents/Releases/2014/08/07.aspx

Intel's Loihi "AI" Chip [2017-2018]

- Intel's Loihi chip has 1,024 artificial neurons, or 130,000 simulated neurons, with 130 million possible synaptic connections.



THE POSSIBILITIES ARE ENDLESS. . .