# Future of Computing: Moore's Law, The End of Frequency Scaling, Domain-specific accelerators

15-213: Introduction to Computer Systems
27th Lecture, Apr. 28, 2022

**Instructor:**
David Andersen

# Moore's Law Origins


**April 19, 1965**

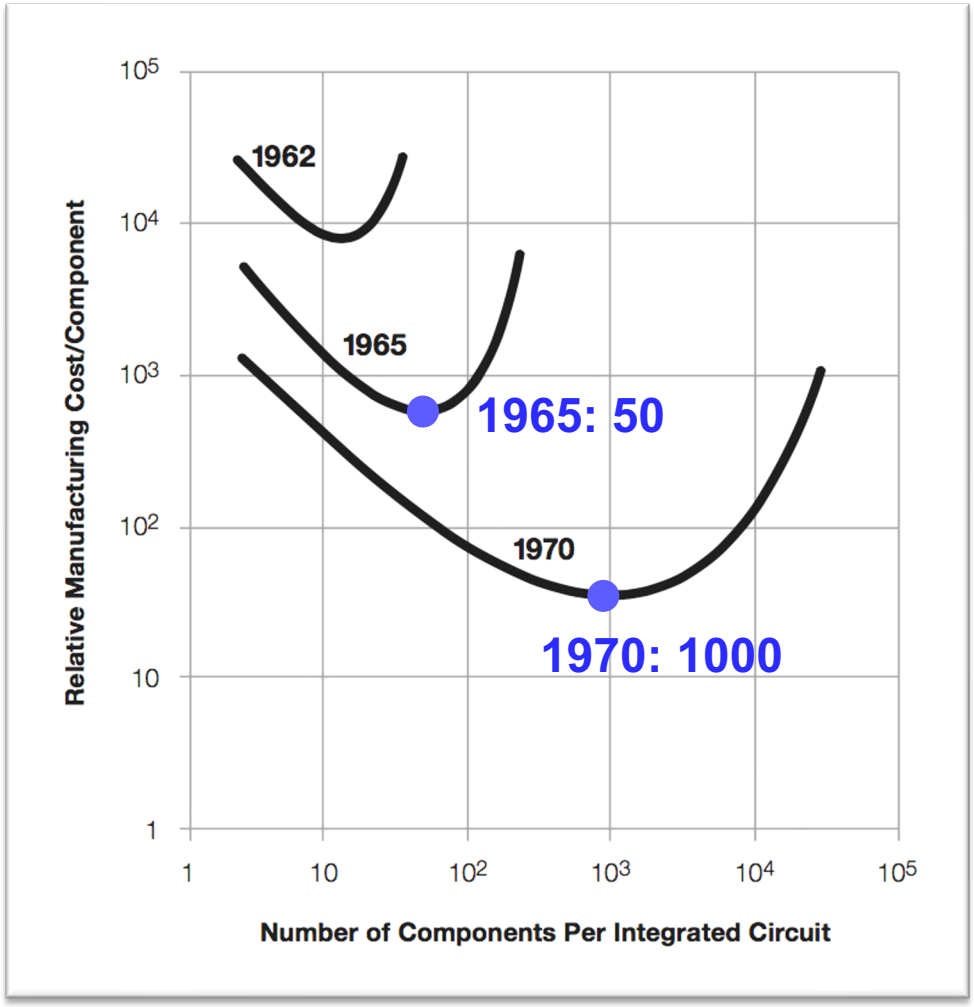
## Cramming more components onto integrated circuits

**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



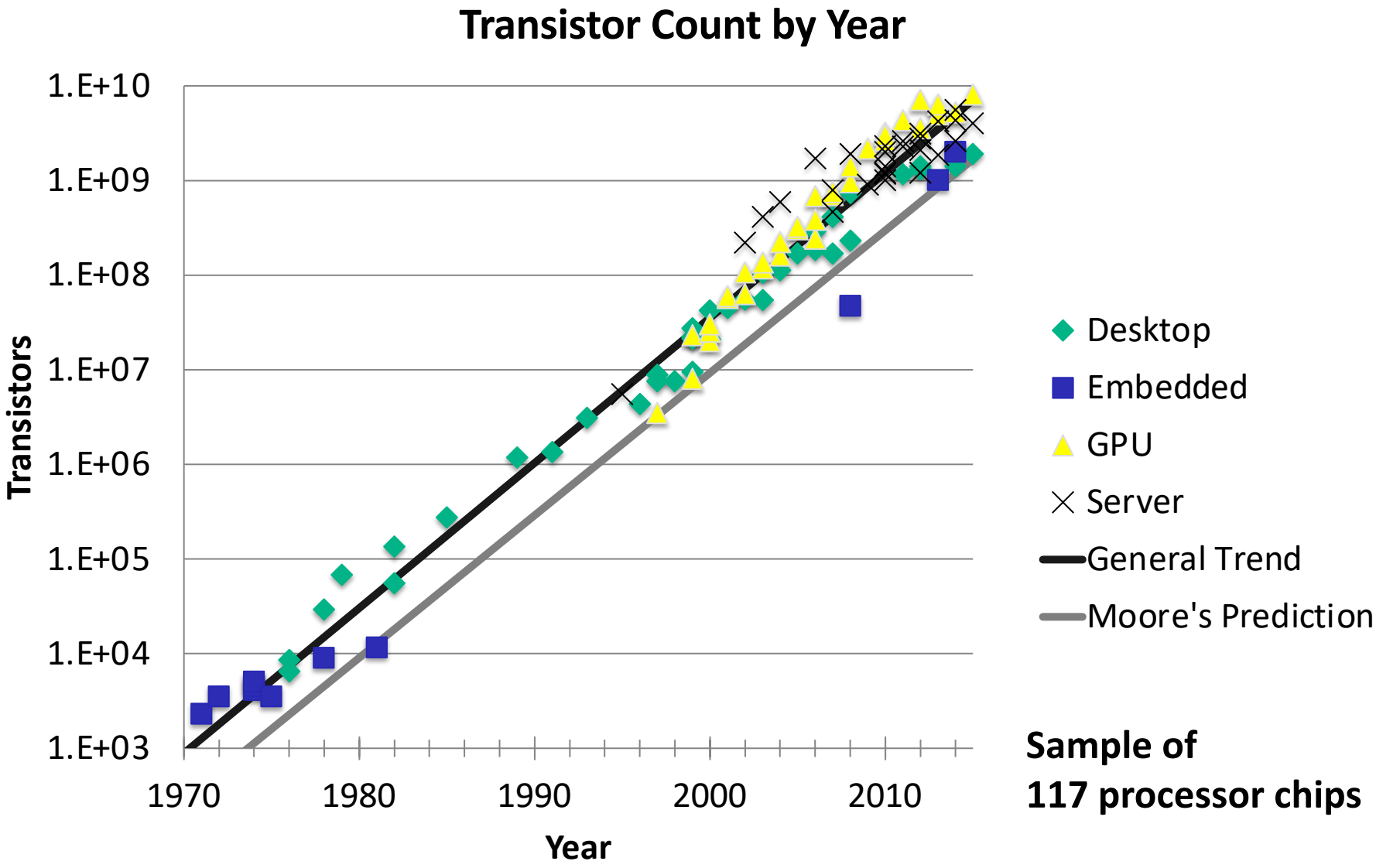

## Moore's Thesis

- Minimize price per device
- Optimum number of devices / chip increasing 2x / year

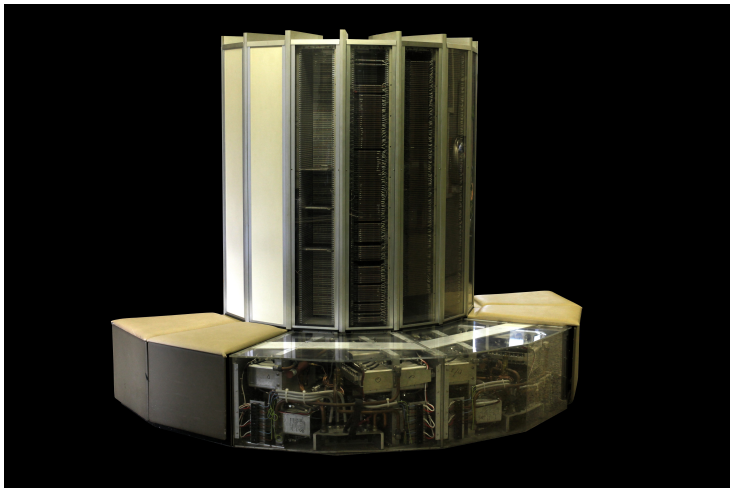
## Later

- 2x / 2 years
- "Moore's Prediction"

# Moore's Law: 50 Years



# What Moore's Law Has Meant



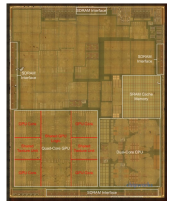
- **1976 Cray 1**
  - 250 M Ops/second
  - ~170,000 chips
  - 0.5B transistors
  - 5,000 kg, 115 KW
  - $9M
  - 80 manufactured

- **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

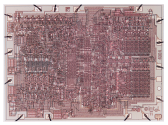


Apple A8 Processor
2 B transistors

# Visualizing Moore's Law to Date

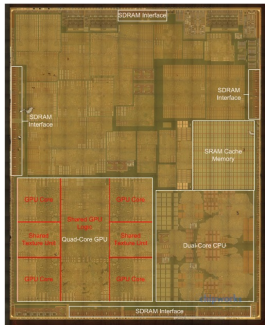
*If transistors were the size of a grain of sand*

**Intel 4004**
**1970**
**2,300 transistors**



**0.1 g**

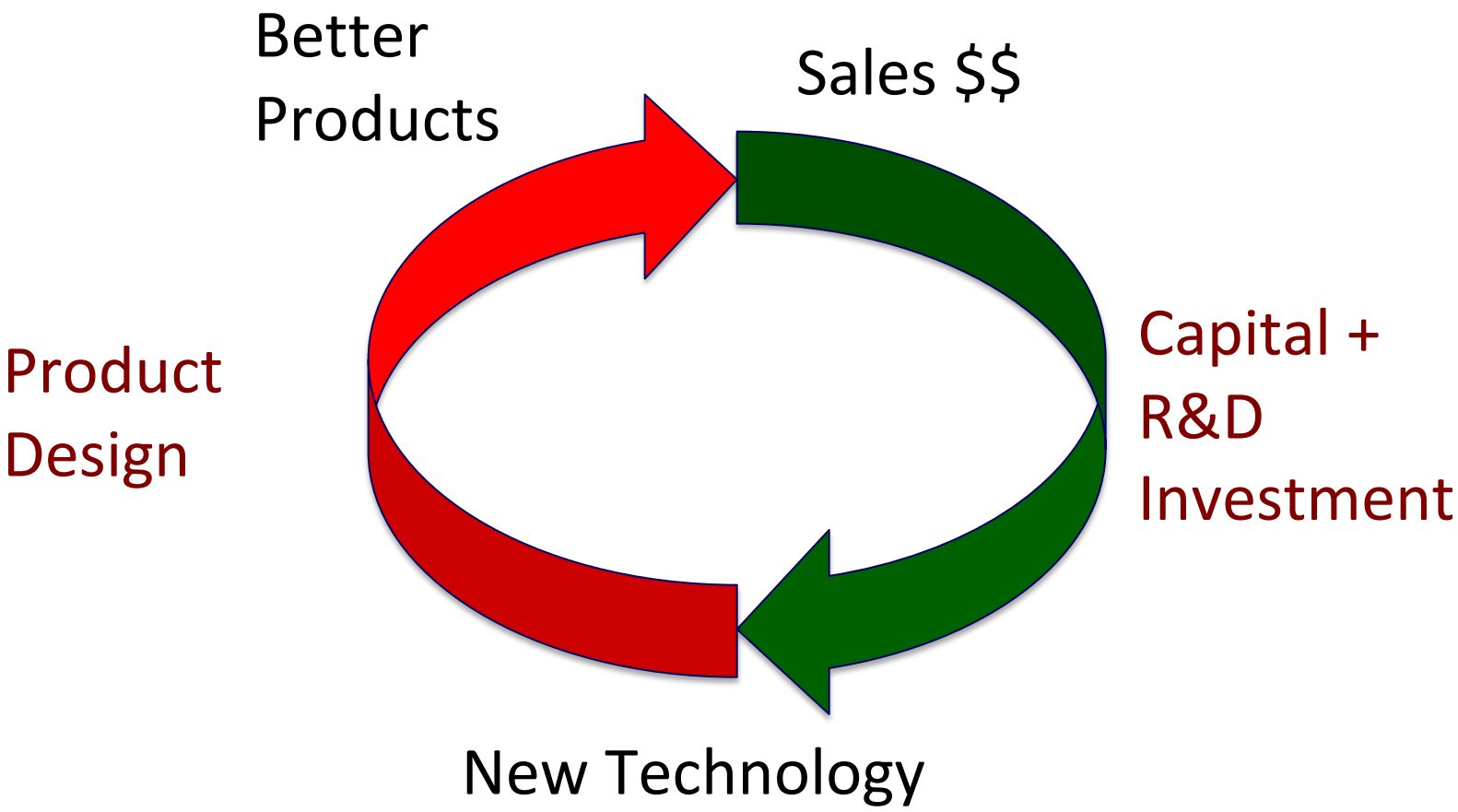
**Apple A8**
**2014**
**2 B transistors**





**88 kg**

# Moore's "Law" drove itself




**Consumer products sustain the $300B semiconductor industry**

# 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

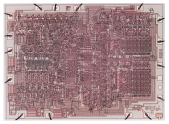
# Increasing Transistor Counts

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

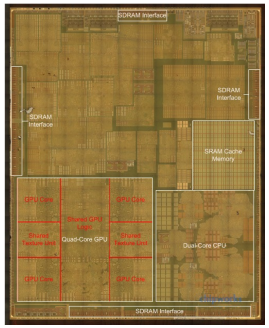
***Will these trends continue?***

# Chips Have Gotten Bigger

**Intel 4004**
**1970**
**2,300 transistors**
**12 mm²**

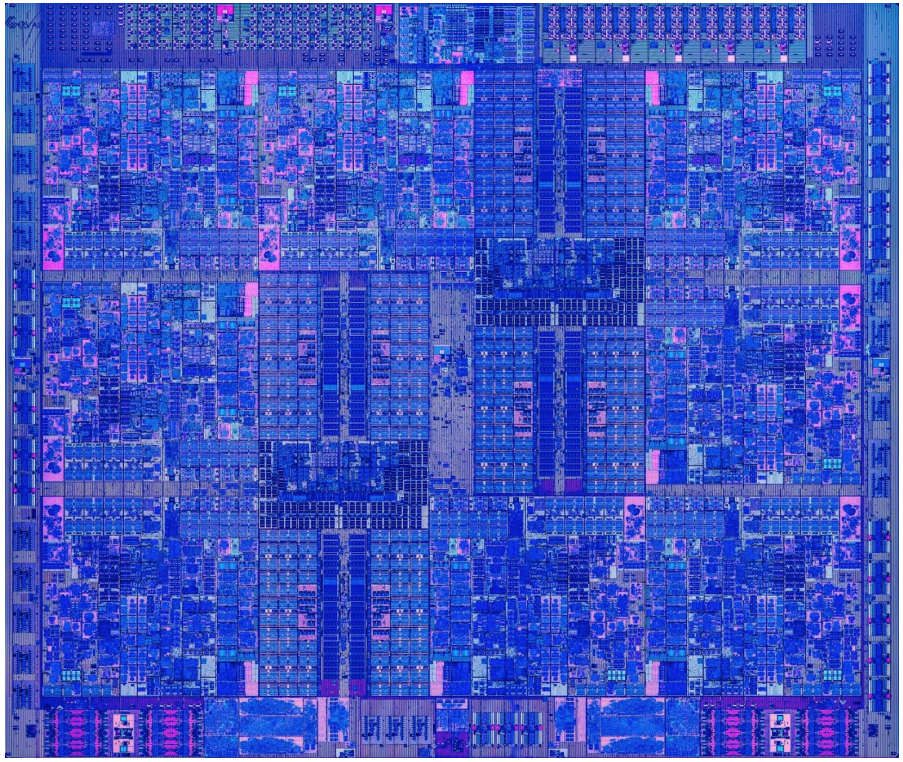


**Apple A8**
**2014**
**2 B transistors**
**89 mm²**

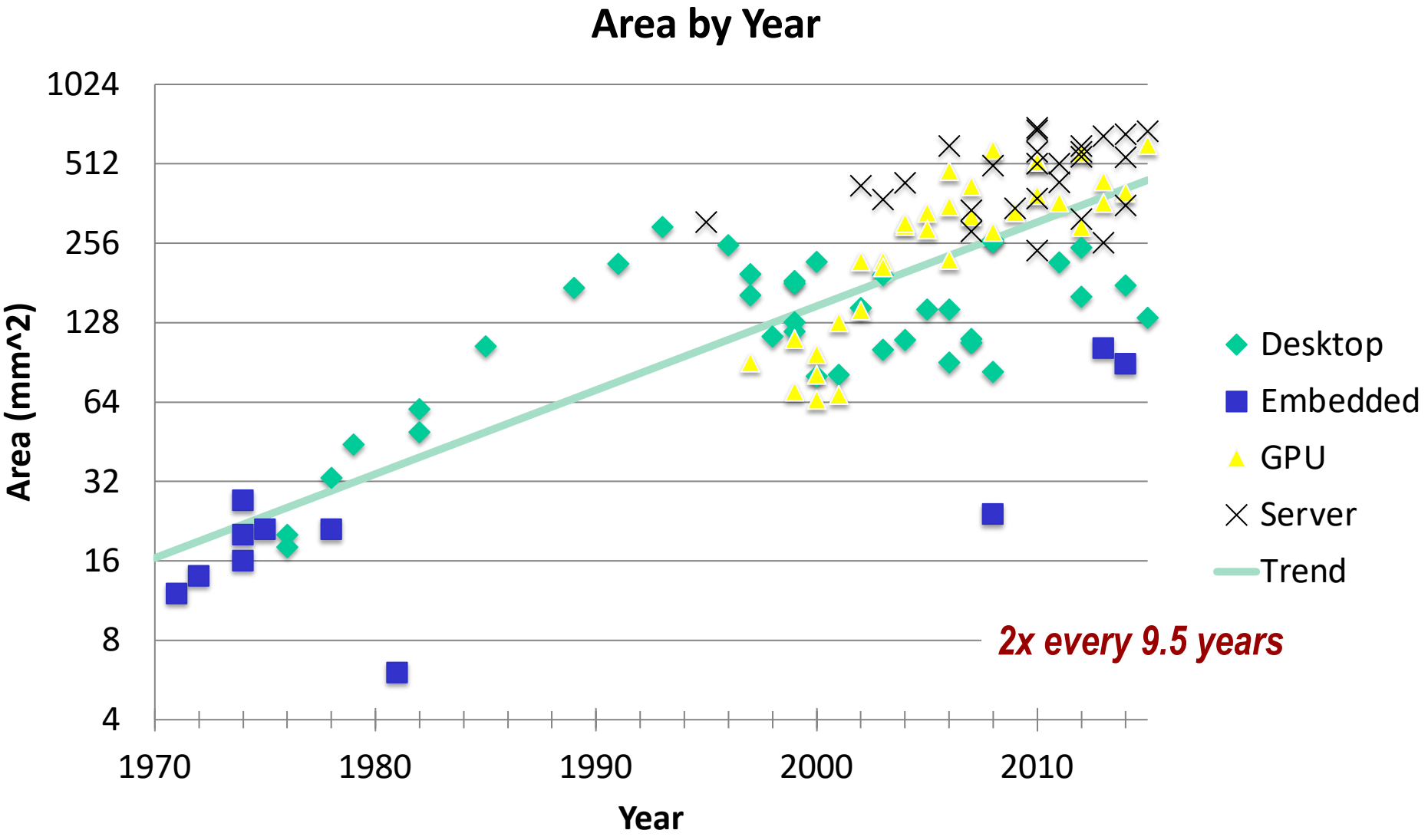




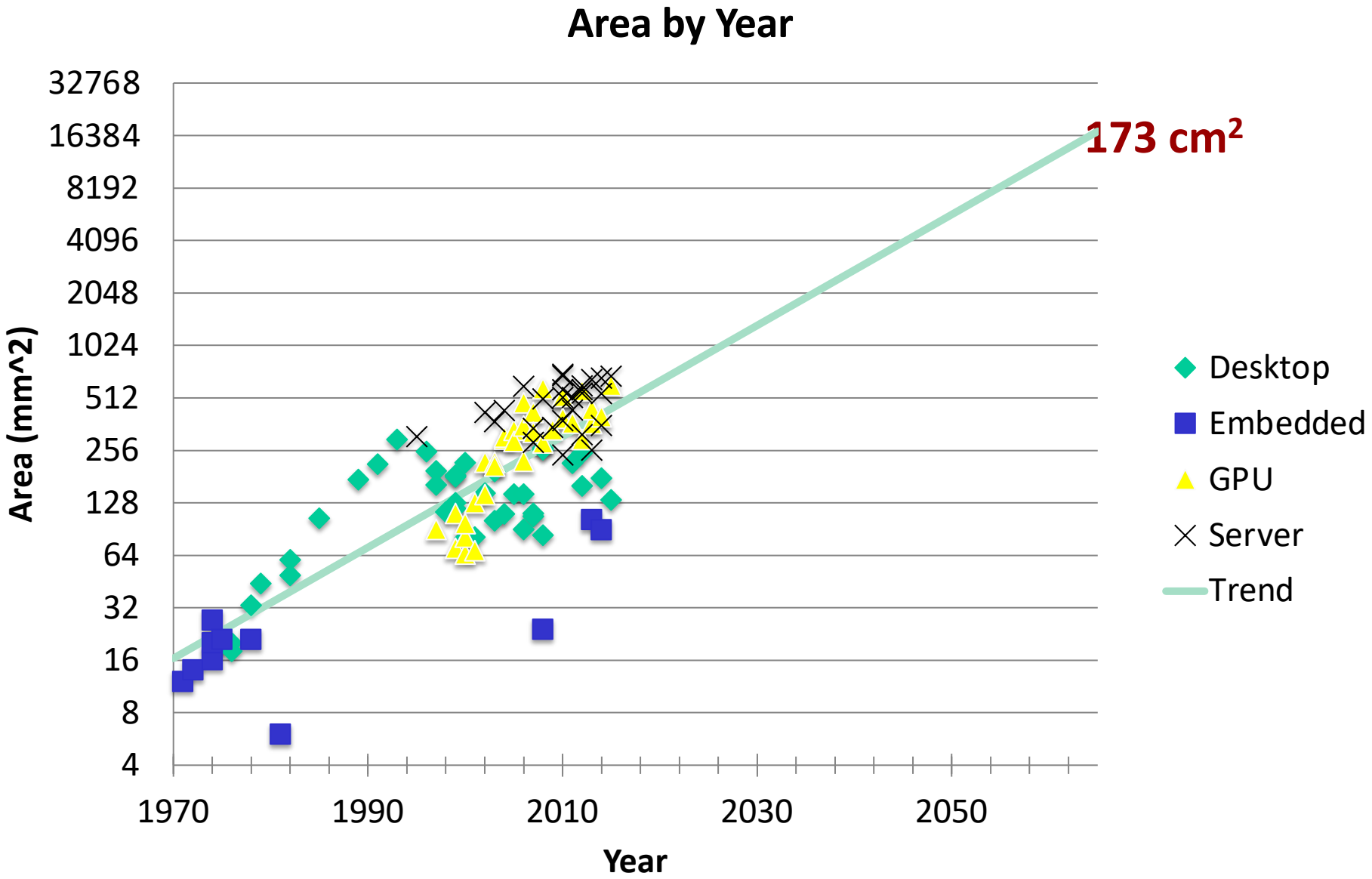
**IBM z13**
**205**
**4 B transistors**
**678 mm²**



# Chip Size Trend



# Chip Size Extrapolation



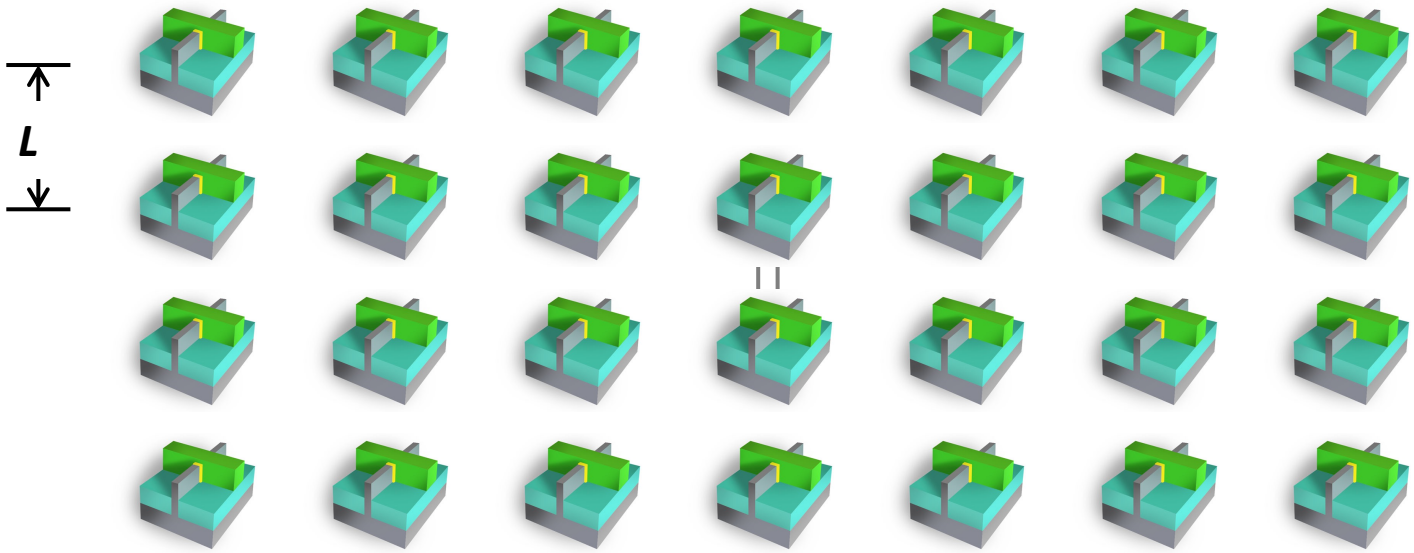
# Extrapolation: The iPhone 31s

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

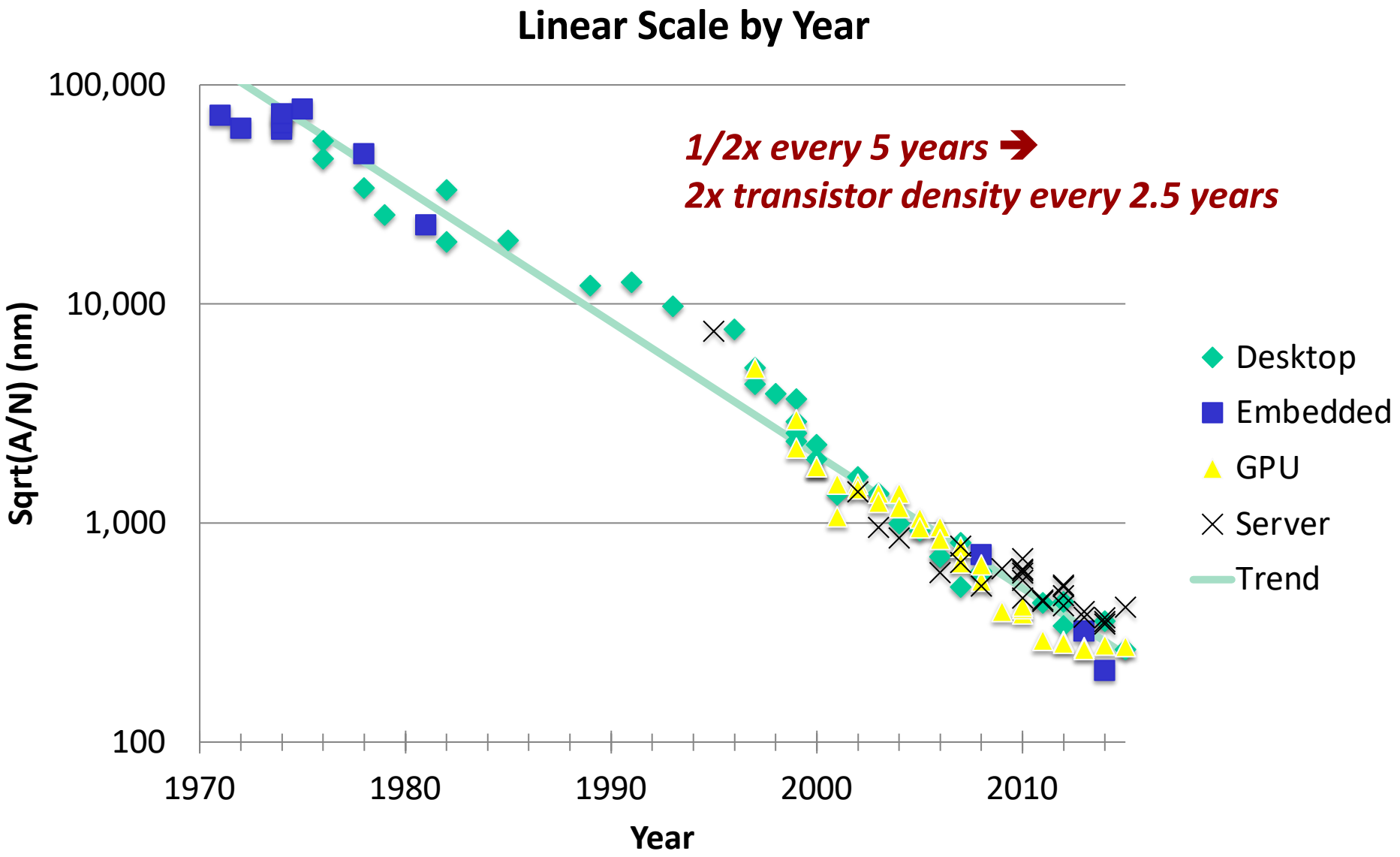

# Transistors Have Gotten Smaller

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

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

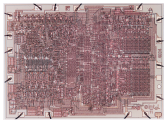


# Linear Scaling Trend

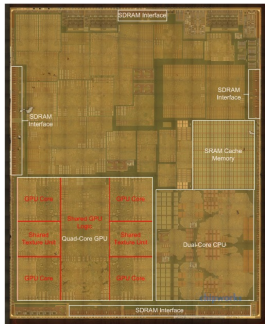


# Decreasing Feature Sizes

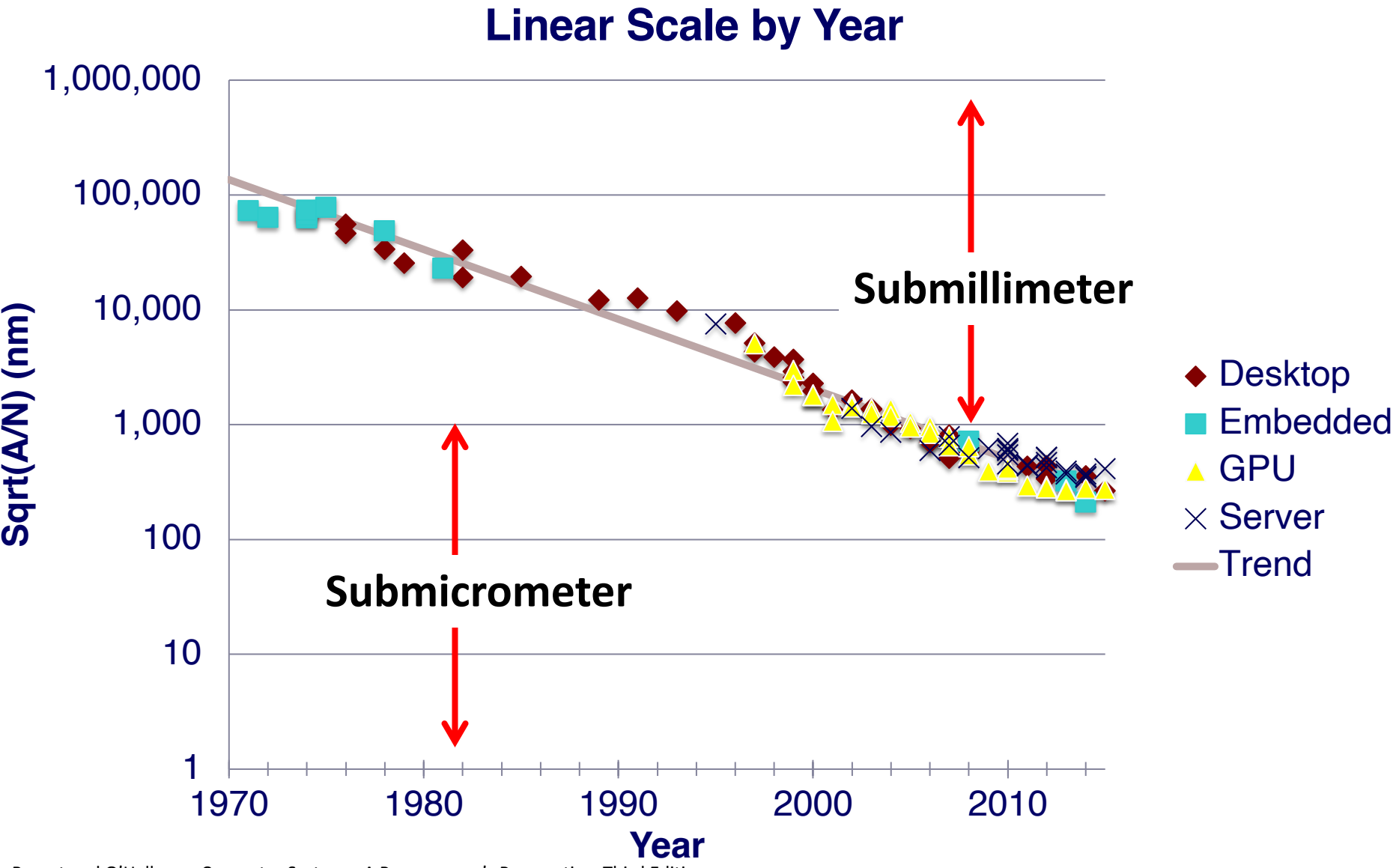
**Intel 4004**
**1970**
**2,300 transistors**
***L* = 72,000 nm**



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



# Linear Scaling Trend

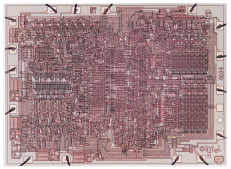


# Submillimeter Dimensions

$10^{-3}$ 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

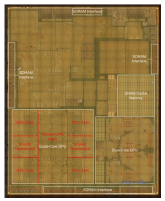
2µm: E coli bacterium length

$10^{-6}$ 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

$10^{-8}$

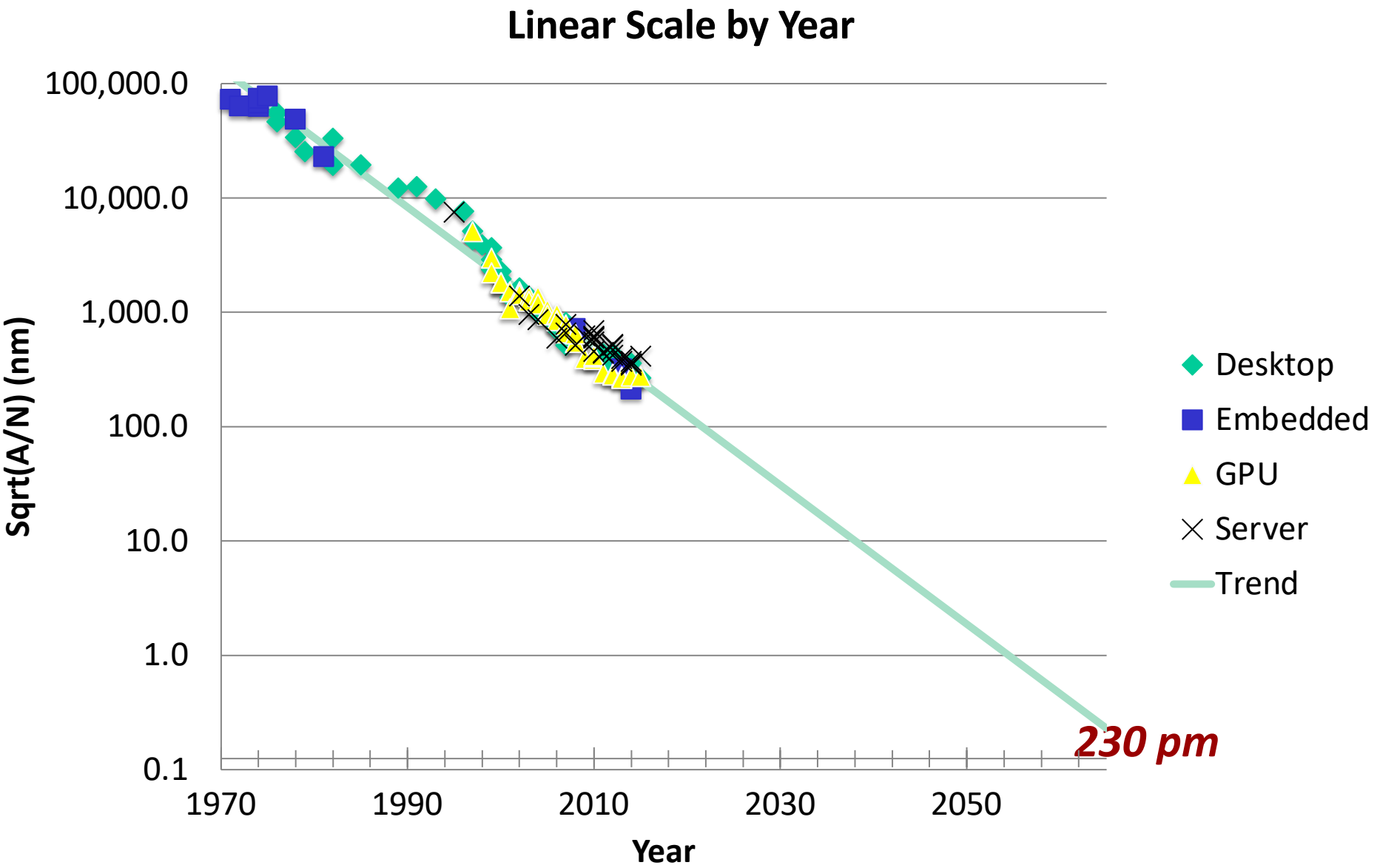
9nm: Cell membrane thickness

2nm: DNA helix diameter

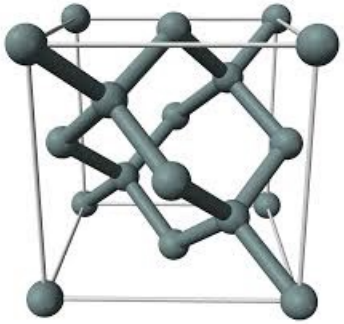
$10^{-9}$ 1 nanometer (nm)

1nm: Carbon nanotube diameter

# Linear Scaling Extrapolation



# Subnanometer Dimensions



$10^{-9}$ 1 nanometer (nm)

1nm: Carbon nanotube diameter

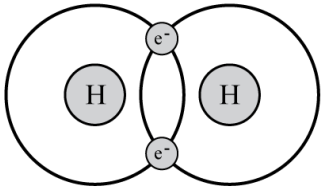
543pm: Silicon crystal lattice spacing

230pm: 2065 linear scale projection

$10^{-10}$

74pm: Spacing between atoms in hydrogen molecule

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




$10^{-11}$

2.4pm: Electron wavelength (Compton wavelength)

$10^{-12}$ 1 picometer (pm)

# Reaching 2065 Goal

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



- **Is this possible?**

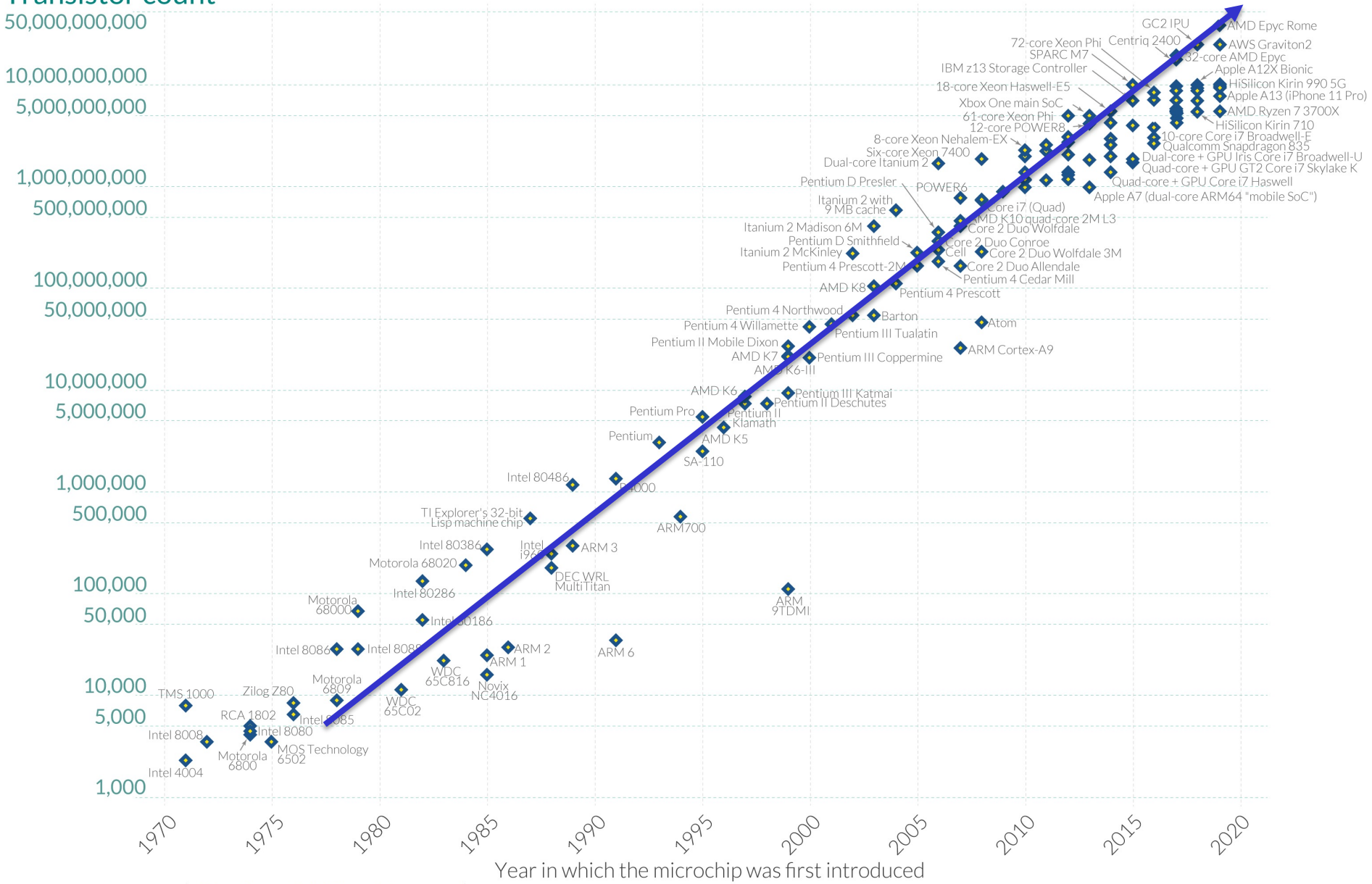
**No!**

**Not with 2-d fabrication**

# Moore's Law: The number of transistors on microchips has doubled every two years

Our World in Data

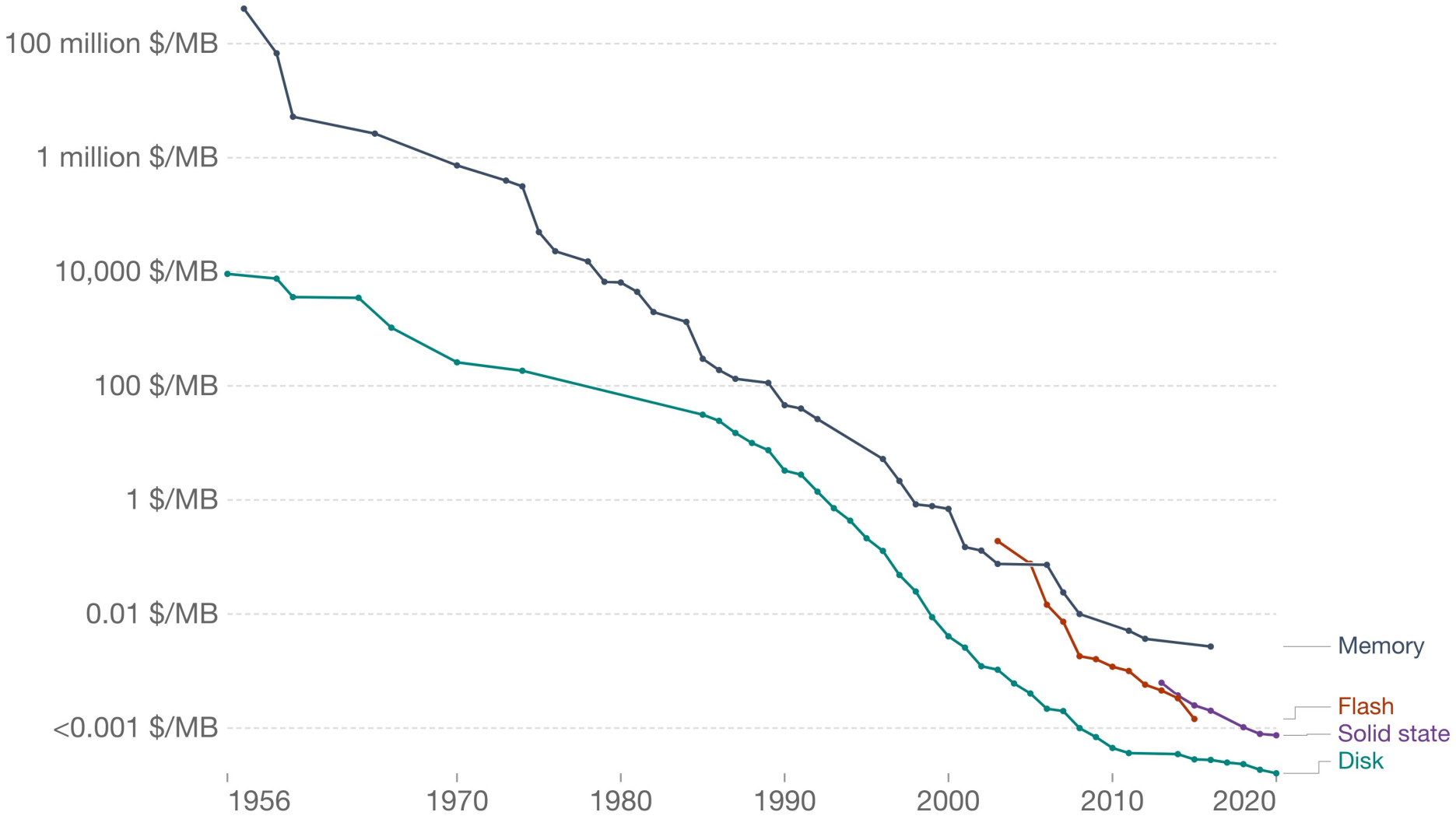
Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.




Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

# Historical cost of computer memory and storage

Measured in US dollars per megabyte.

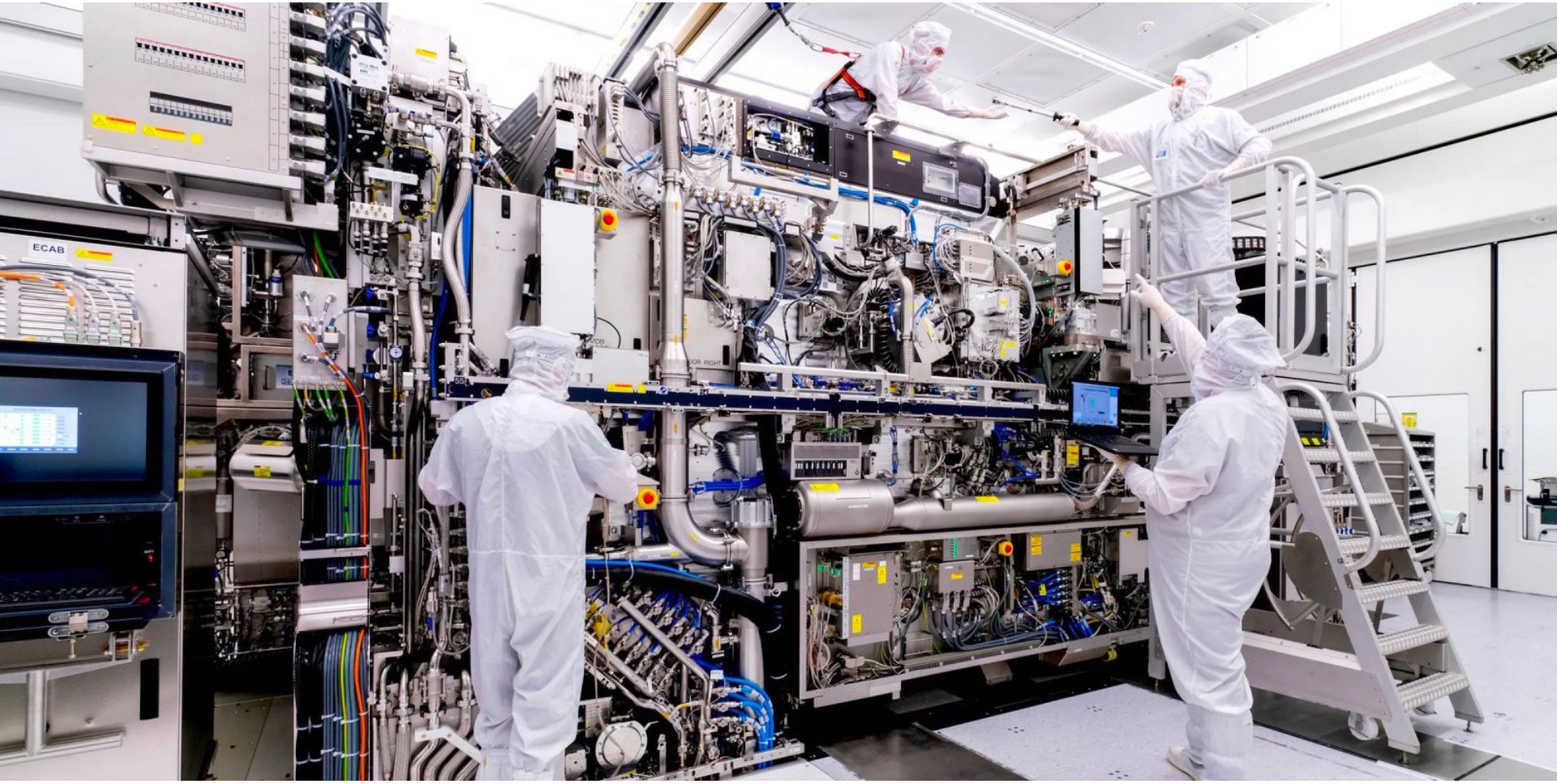



Source: John C. McCallum (2022)

Note: For each year the time series shows the cheapest historical price recorded until that year.

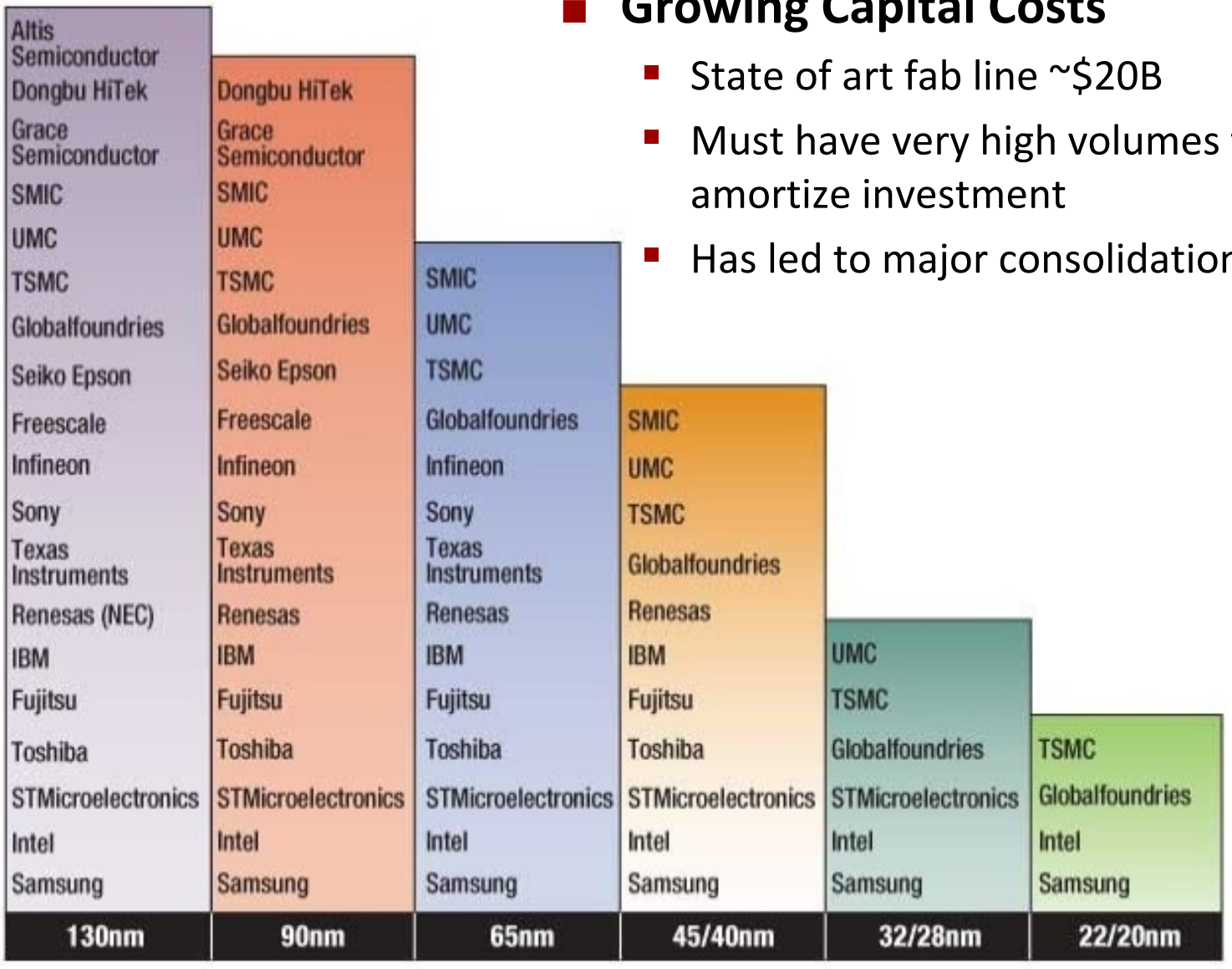
# Fabrication Economics

- **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


ECAB

# Challenges 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



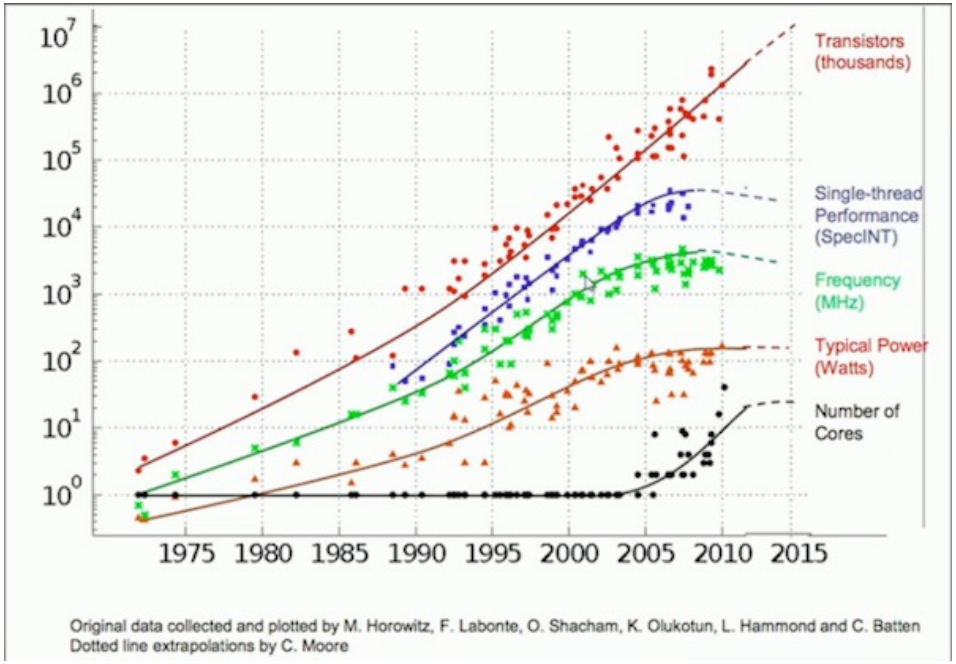
# Meeting Power Constraints

- 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

# End of Dennard Scaling

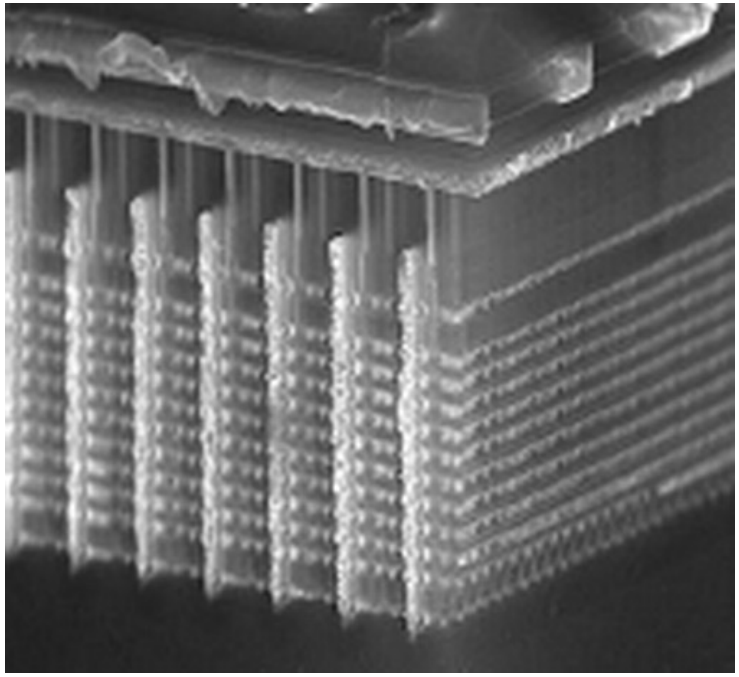


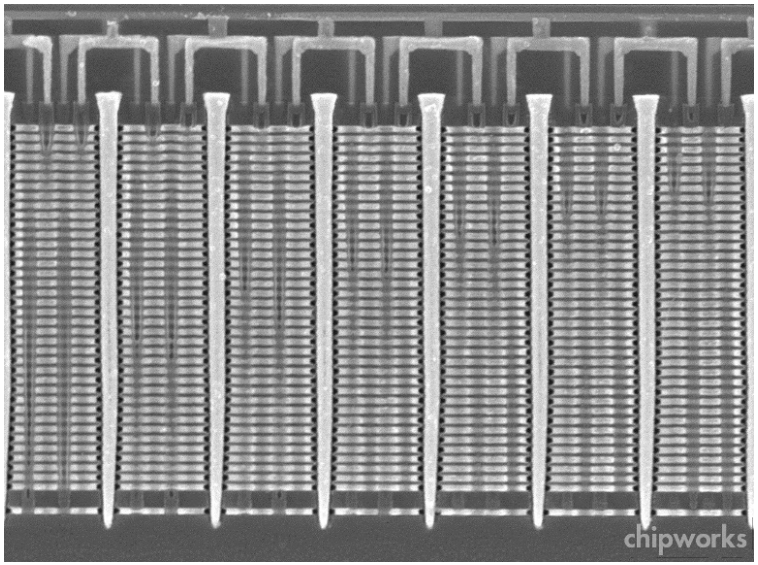

- **What Happened?**
  - Old: Reduce voltage as we reduced transistor size
  - New: Can't drop voltage below ~1V
  - Reached limit of power / chip in 2004
  - More logic on chip (Moore's Law), but can't make them run faster
    - Response has been to increase cores / chip

# End of Easy Gains – Fun Times Ahead

- We can't just crank up the clock
- We won't be able to infinitely add cores
- We *want* to keep improving perf/$, perf/W, etc.
- But we're going to have to be even more clever at all layers of the computation stack…

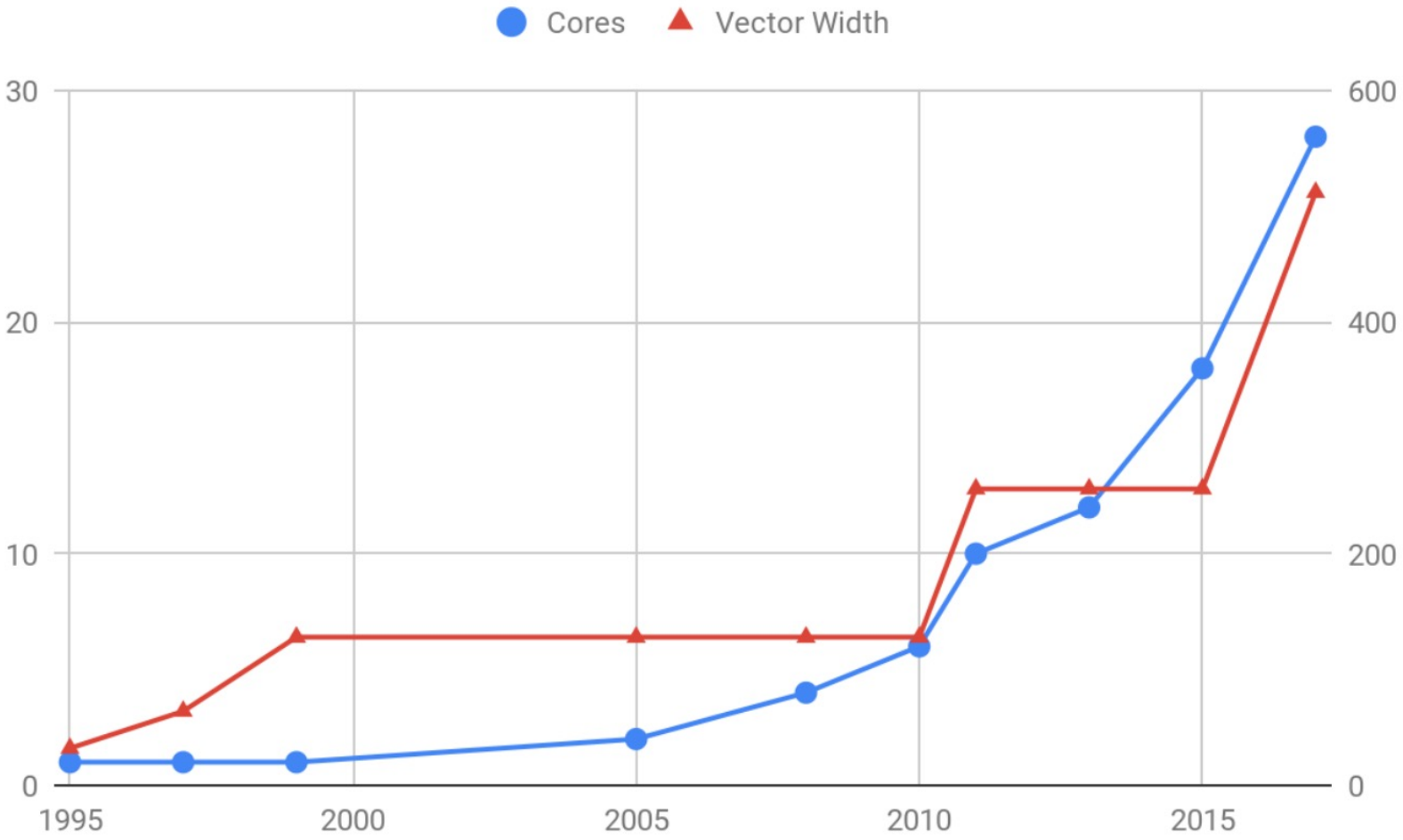
# Samsung V-Nand Flash Example






- 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

# Putting Parallelism on the Programmer

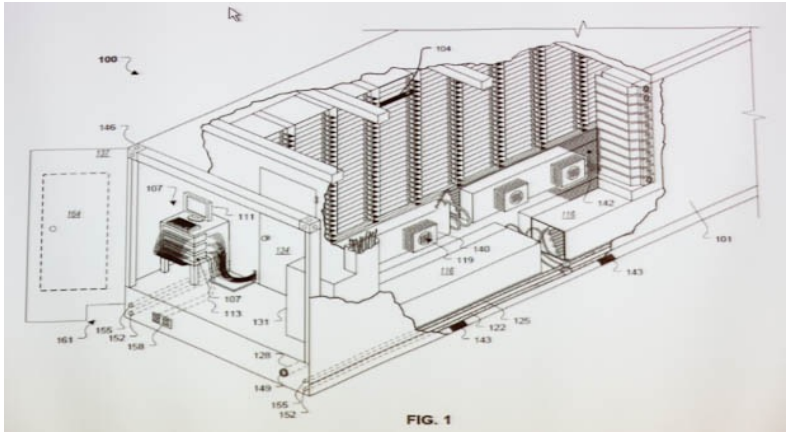



448x

# A lot of parallelism…


**At 350W, a lot of power, too**

# Google Data Centers



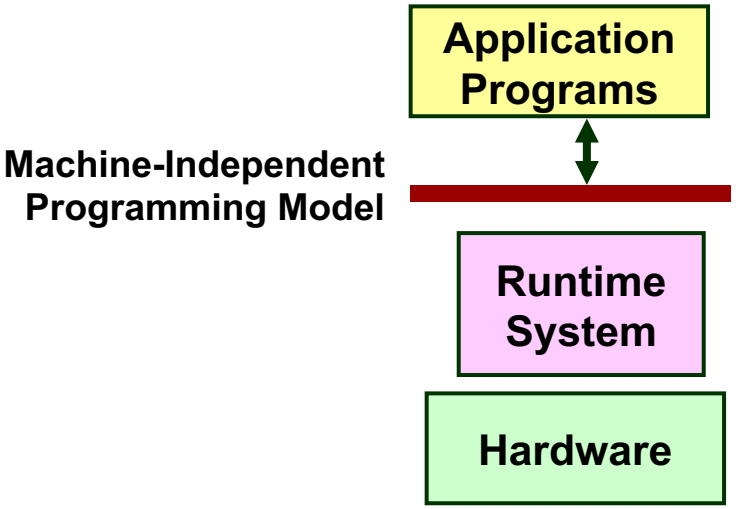

- **Dalles, Oregon**
  - Hydroelectric power @ 2¢ / KW Hr
  - 50 Megawatts
  - Enough to power 60,000 homes

- Engineered for low cost, modularity & power efficiency
- Container: 1160 server nodes, 250KW

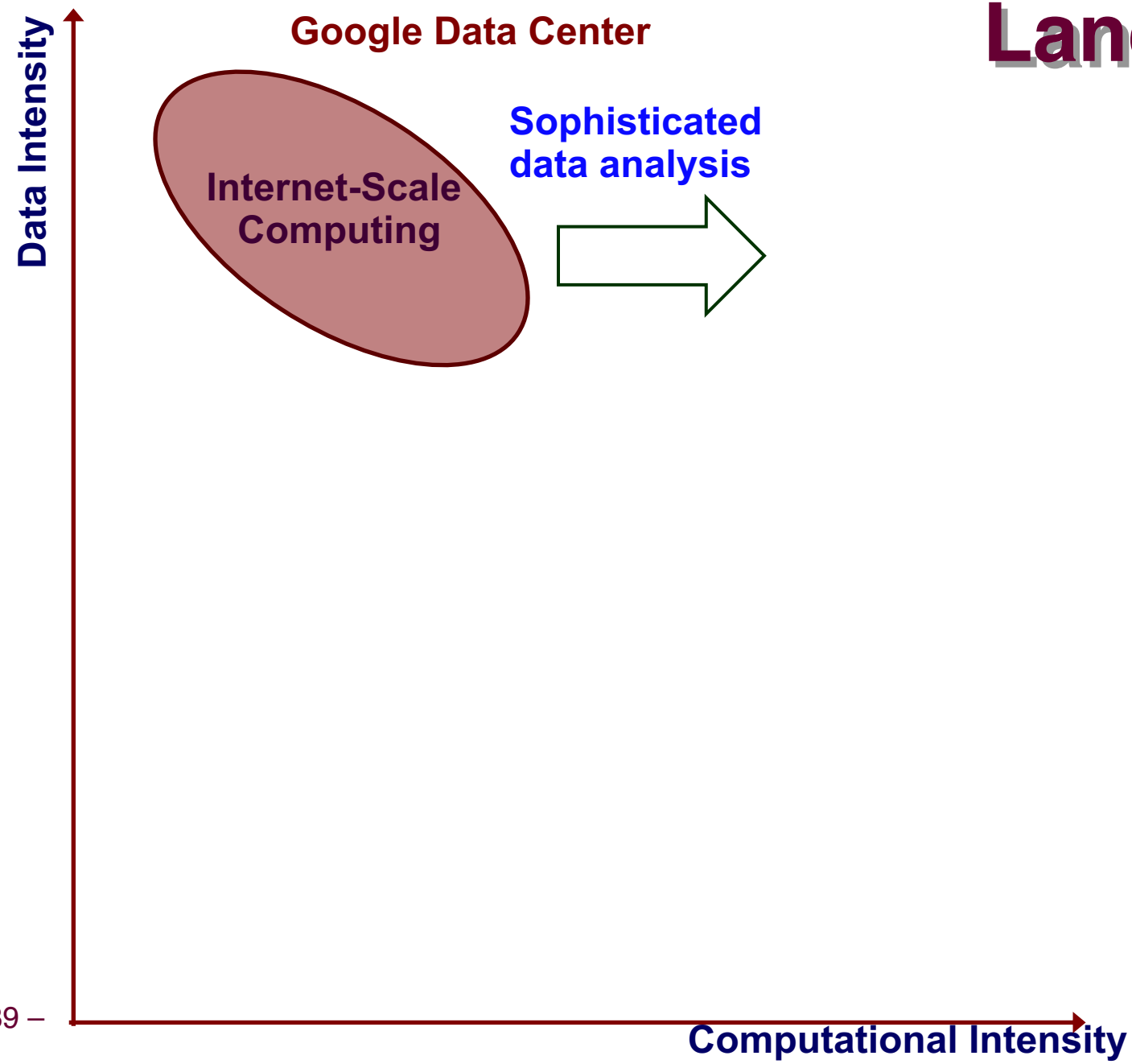
# Cluster Programming Model

- Application programs written in terms of high-level operations on data
- Runtime system controls scheduling, load balancing, …

- **Scaling Challenges**
  - Centralized scheduler forms bottleneck
  - Copying to/from disk very costly
  - Hard to limit data movement
    - Significant performance factor

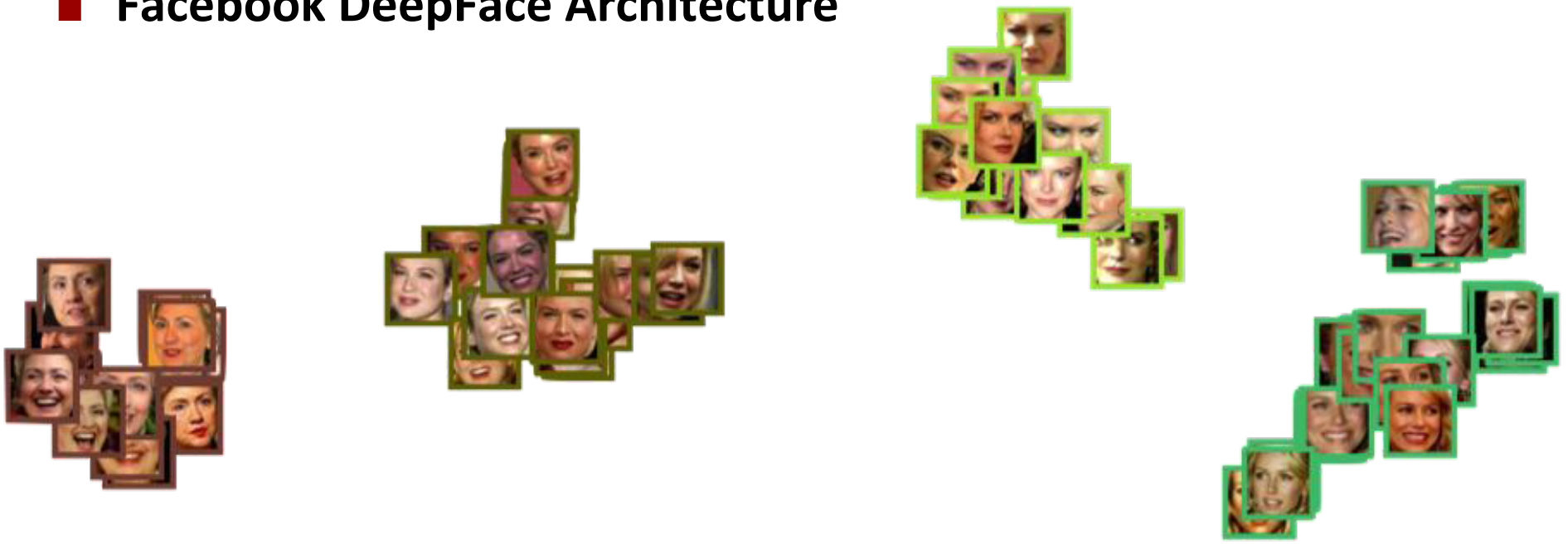


# Computing Landscape Trends



# DNN Application Example

- **Facebook DeepFace Architecture**



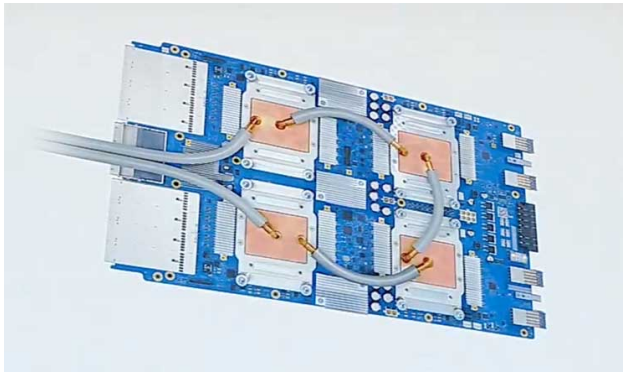
# The "More than Moore" approach

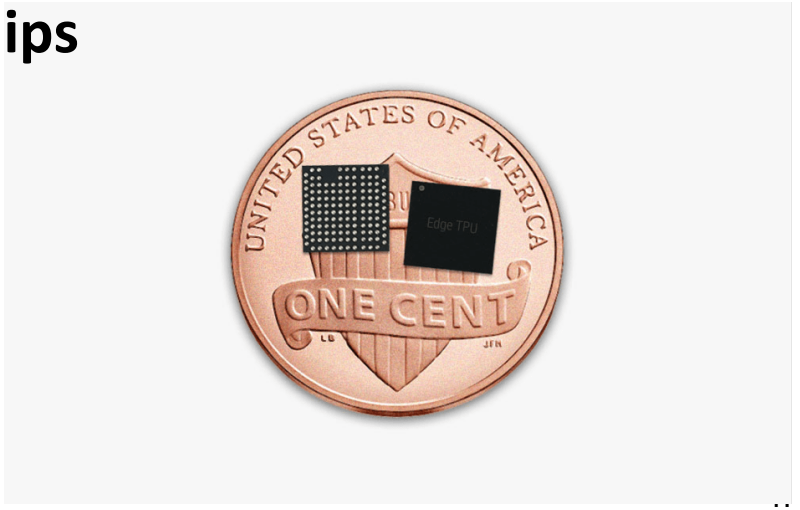
- **"Functional diversification of semiconductor-based devices"**
  - Integration of sensors, RF, MEMS, quantum?, storage

GlobalFoundries Stops All 7nm Development: Opts To Focus on Specialized Processes

by Anton Shilov & Ian Cutress on August 27, 2018 4:01 PM EST

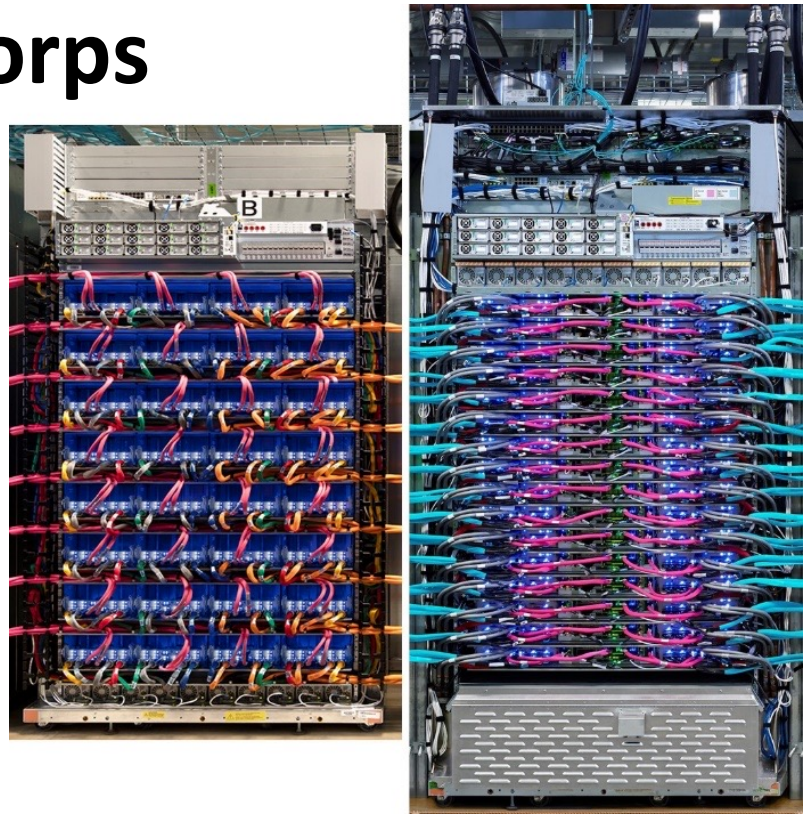
- **Flowering of application-specific chips**



# ASICs in the wild at gigacorps

- Google: TPUv1, v2, v3, custom network interface card, video transcoder,
- Amazon: AI chip, custom VM controller, custom switching chip
- Apple: Its own ARM chips, M1, custom AI chip

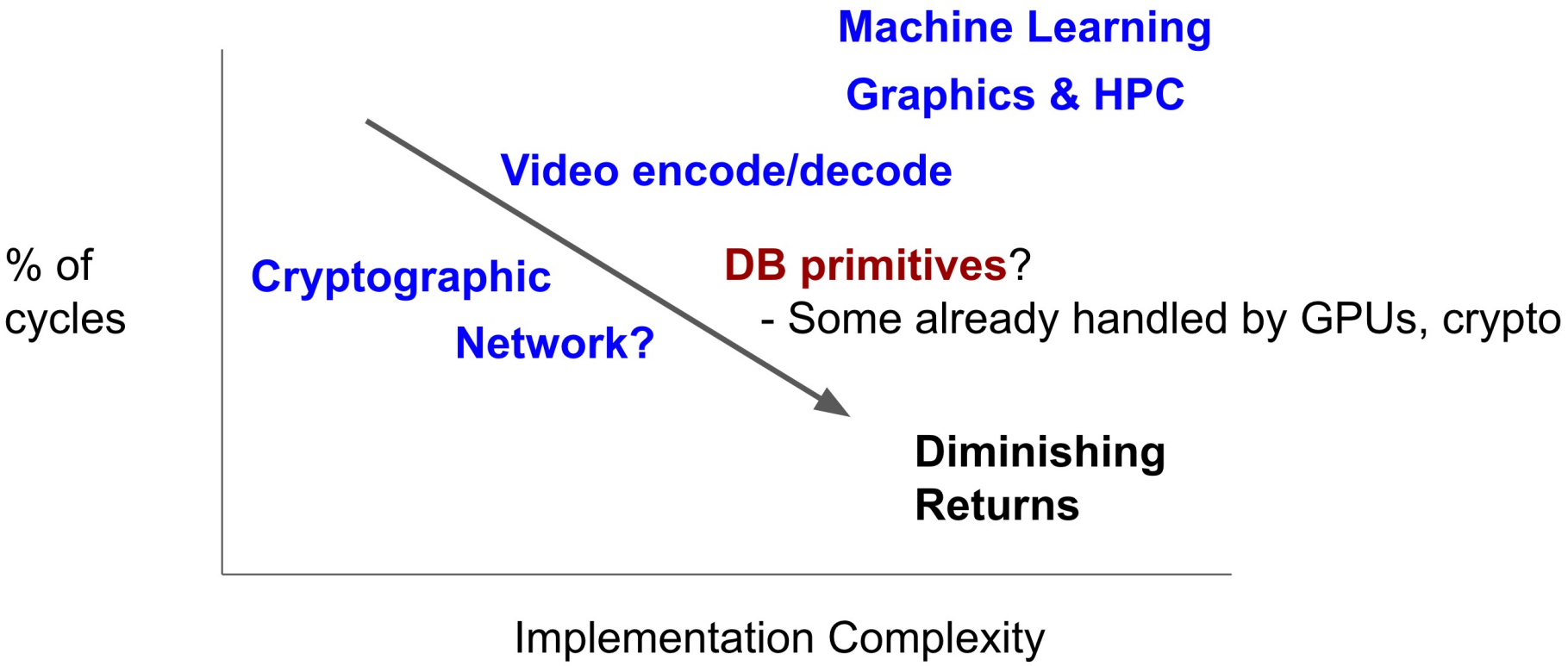



6 core CPU,
5 core GPU,
16 core Neural Engine,
ISP image processor,
Hardware video codec

# We will have more advances

- But they're bumpy!
- Optical interconnects give us a one time big bump
- Persistent memory might still – but bumpy
- We have more lithographic advances – but really hard
- We have yield improvements for EUV – but slowing over time…

# Adding accelerators



# Can’t outrun Amdahl

- **Moore:**
  - Most CPU functions got faster simultaneously
  - Memory density scaled too!
  - I/O (& mem latency) was the primary bottleneck to work around
- **Multicore:**
  - Parallelization bottleneck
- **GPUs / SIMD**
  - Vectorization & parallelization
- **Post-Moore:**
  - Specialization bottleneck

# Future speedups will need it all…

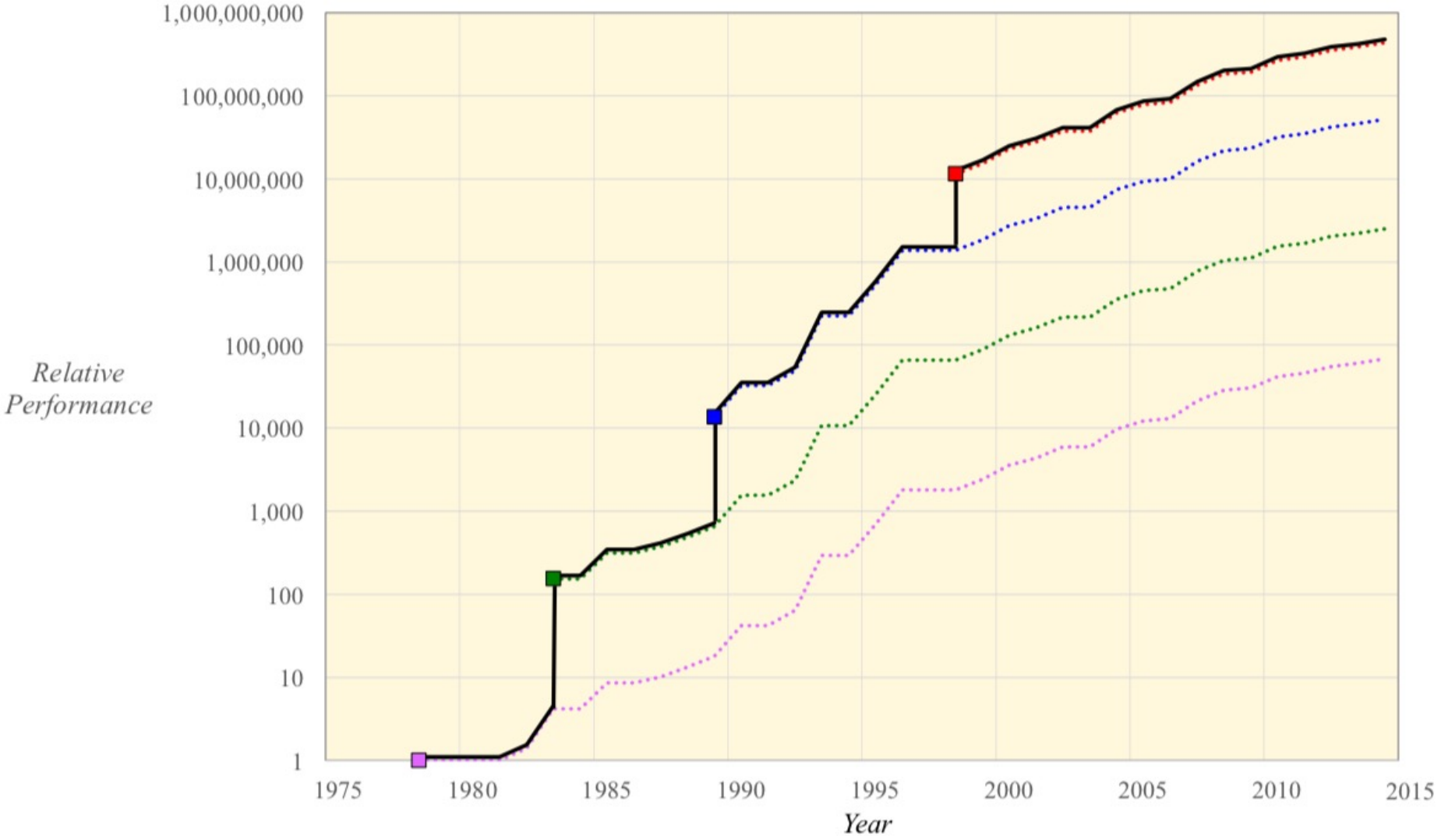
- Hardware improvements
- Specialized hardware/software co-design
- Software implementation improvements
- Algorithmic improvements

# 4096 x 4096 Matrix Multiply

| *Implementation* | *Running time (s)* | *GFLOPS* | *Absolute speedup* |
|---|---|---|---|
| Python | 25,552.48 | 0.005 | 1 |
| Java | 2,372.68 | 0.058 | 11 |
| C | 542.67 | 0.253 | 47 |
| Parallel loops | 69.80 | 1.969 | 366 |
| Parallel divide-and-conquer | 3.80 | 36.180 | 6,727 |
| + vectorization | 1.10 | 124.914 | 23,224 |
| + AVX intrinsics | 0.41 | 337.812 | 62,806 |

- **(Leiserson et al., "There's Plenty of Room At The Top")**

# MaxFlow over time (Leiserson et al)



# 2018--?: Putting Heterogeneity on the programmer

- **Trend in architecture over last decade: Increasingly shift pain to the programmer**
  - Parallelization, vectorization, massive concurrency, etc.
- **This will get worse. (No alternative yet)**

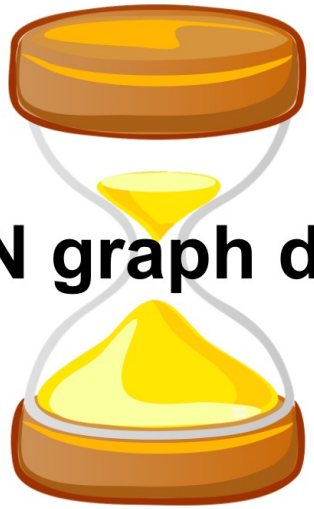
# Applications

Narrow Waists

Heterogenous, experts-only Hardware

# Waists are emerging: ML example

Applications



**DNN graph definition**

TensorFlow TensorRT Android NNAPI Apple CoreML

TPUs v1-3, EdgeTPU, Neural Compute Stick, A12 Bionic, Intel FPGA DLIA, GPUs, x86, ARM, ....

# How do we manage?

- By making sure we combine systems knowledge + domain knowledge to craft frameworks that help programmers get things done in this increasingly complex hardware world

- Go forth and build awesome systems!