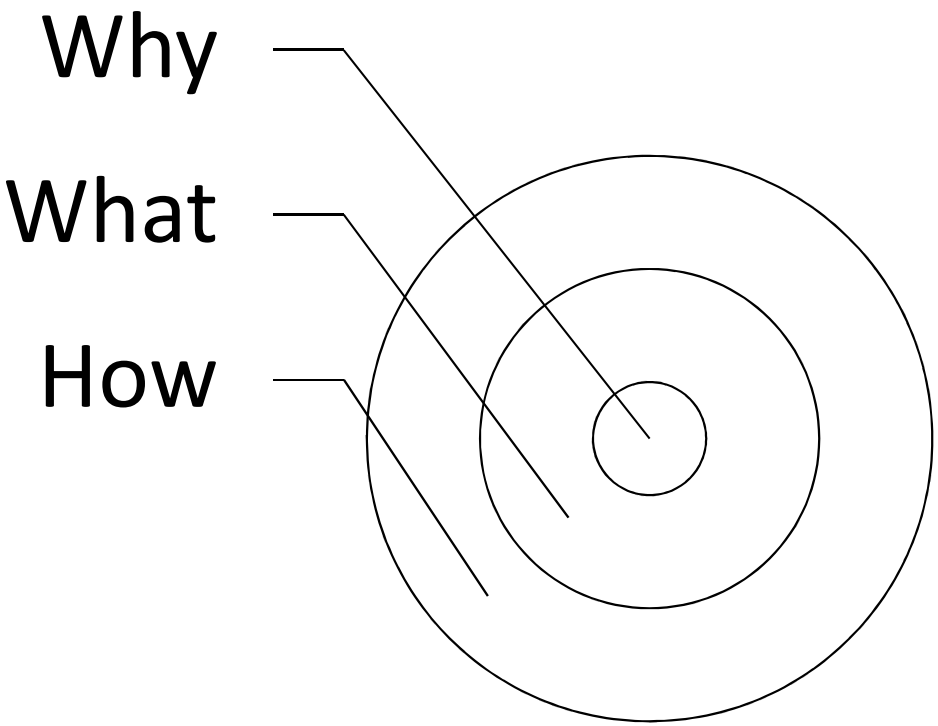
# CS249r

# Goal Toward Domain Specific Architectures for Autonomous Machines



# Reading Feedback

- 2) Why is a **constant budget of power** important? Is it mainly a thermal issue where the processor might degrade? Or does it have to do with the economic cost of power consumption?
- 1) The author seems to suggest that Dennardian scaling is vanishing, but assume that Moore's law is more or less around to stay for the near future. For **how long will the transistor density increase?**
- The background for the 2nd section was a bit lacking. **Would it be possible to do a quick summary of Dennardian and Post-Dennardian scaling?**
- Not much, but I think the **explanation of why Dennard scaling has broken down** in the early part of the paper is somewhat confusing to follow and would benefit from some sort of graph.
- Two sections that I would have liked more detail on (mostly because of my own lack of knowledge, not because of the fault of the author) were using dark silicon for bigger on-chip caches and using it for **CGRAs**. I was slightly confused about the cross-over point for bandwidth-limitation and power-limitation for on-chip cache. And also, about how the datapaths for computation for CGRAs would be integrated onto dark Silicon.
- Why the analysis scope is these four approaches, is there any other **promising potential possibilities** that we can leverage in dark silicon dominated future?
- The paper helps me understand why different companies / industries may try to build their own custom silicon. However, **the relation between the hardware and the associated software / programming languages is hard to follow**. Why couldn't a middle-man be added which would automatically optimize code for the silicon it would run on?
- Terms, definitions, and concepts that were never introduced properly. **What is multicore and what does it mean to scale it? Transistor switching?** Some of the quantitative examples of tradeoffs between metrics like frequency and area/size were confusing, mostly because I don't have a solid understanding of how chips work. Didn't understand NTV processor section.
- Considering this paper was published in the DAC, I'm not sure if this is a weakness or not, but I would have liked to see some **systems-level implications of dark silicon**. I was left asking if there's any way the hardware-software interface can make the best of dark silicon, and support application-level programmers.

# The Glory of Moore's Law

The experts look ahead

**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.

**The future of integrated electronics** is the future of electronics itself. The advantages of integration will bring about a proliferation of electronics, pushing this science into many new areas.

Integrated circuits will lead to such wonders as home computers—or at least terminals connected to a central computer—automatic controls for automobiles, and personal portable communications equipment. The electronic wristwatch needs only a display to be feasible today.

But the biggest potential lies in the production of large systems. In telephone communications, integrated circuits in digital filters will separate channels on multiplex equipment. Integrated circuits will also switch telephone circuits and perform data processing.

Computers will be more powerful, and will be organized in completely different ways. For example, memories built of integrated electronics may be distributed throughout the machine instead of being concentrated in a central unit. In addition, the improved reliability made possible by integrated circuits will allow the construction of larger processing units. Machines similar to those in existence today will be built at lower costs and with faster turn-around.

**Present and future**

By integrated electronics, I mean all the various technologies which are referred to as microelectronics today as well as any additional ones that result in electronics functions supplied to the user as irreducible units. These technologies were first investigated in the late 1950's. The object was to miniaturize electronics equipment to include increasingly complex electronic functions in limited space with minimum weight. Several approaches evolved, including microassembly techniques for individual components, thin-film structures and semiconductor integrated circuits.

Each approach evolved rapidly and converged so that each borrowed techniques from another. Many researchers believe the way of the future to be a combination of the various approaches.

The advocates of semiconductor integrated circuitry are already using the improved characteristics of thin-film resistors by applying such films directly to an active semiconductor substrate. Those advocating a technology based upon films are developing sophisticated techniques for the attachment of active semiconductor devices to the passive film arrays.

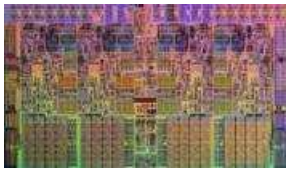
Both approaches have worked well and are being used in equipment today.

The author

Dr. Gordon E. Moore is one of the new breed of electronic engineers, schooled in the physical sciences rather than in electronics. He earned a B.S. degree in chemistry from the University of California and a Ph.D. degree in physical chemistry from the California Institute of Technology. He was one of the founders of Fairchild Semiconductor and has been director of the research and development laboratories since 1959.

Electronics, Volume 38, Number 8, April 19, 1965

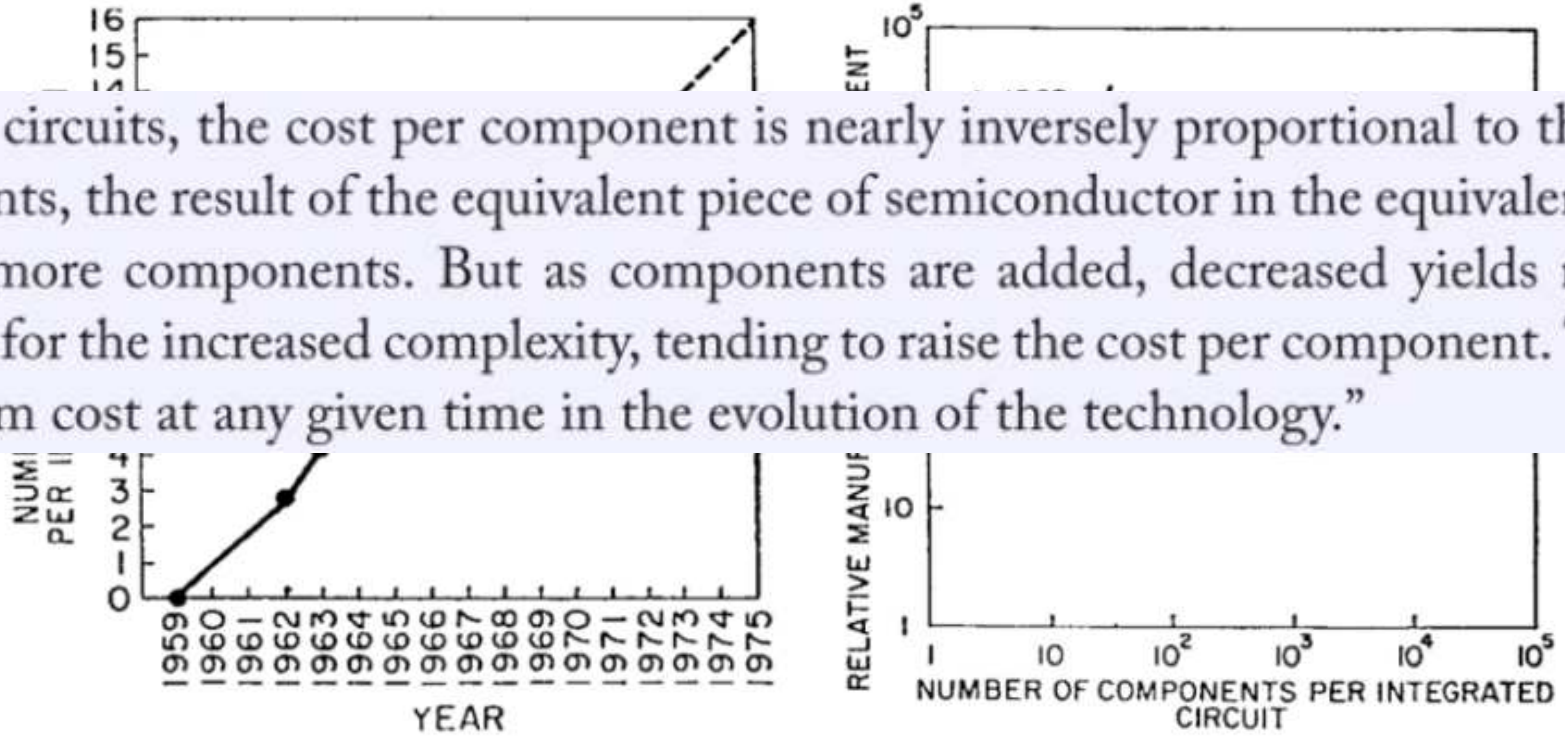
Intel 4004
2300 transistors
740 kHz clock
10um process
10.8 usec/inst



Intel Core i7 980X
1.17B transistors
3.33 GHz clock
32nm process
73.4 psec/inst

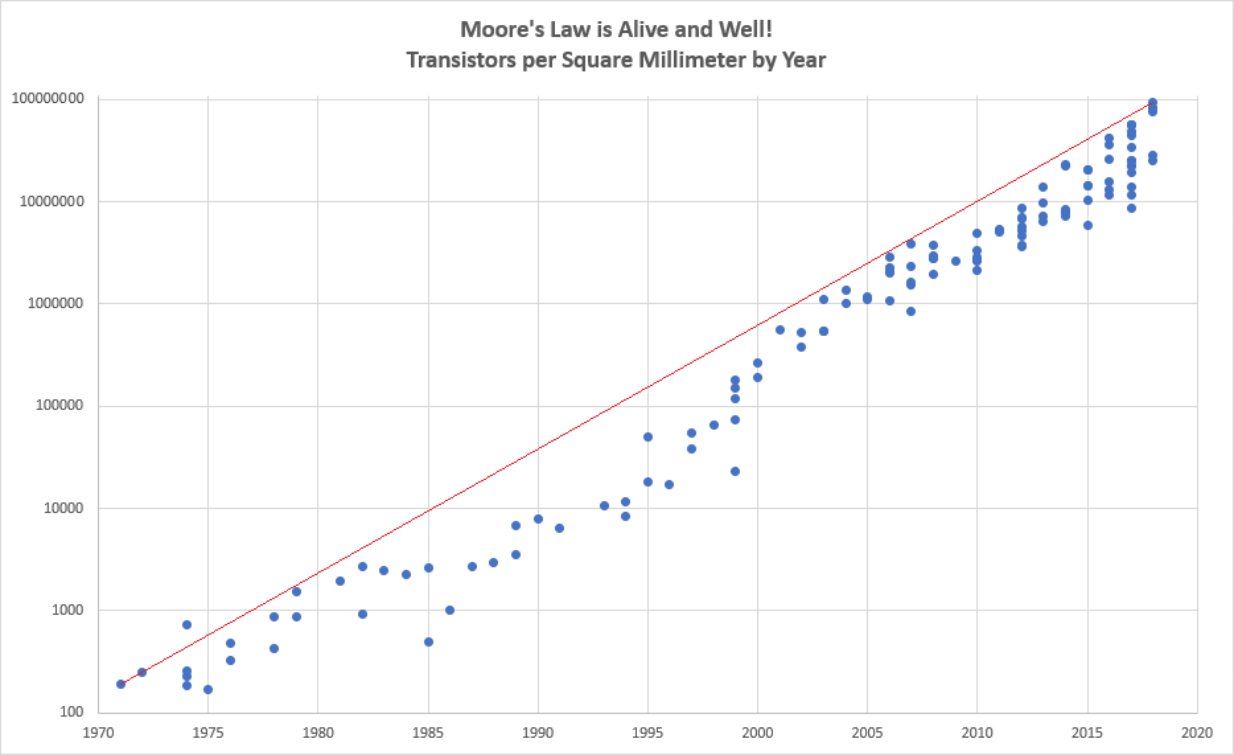
%/year, Ratios:
38%, 508000
23%, 4450
15%, 312
34%, 147000

# The Driving Factor -- Cost



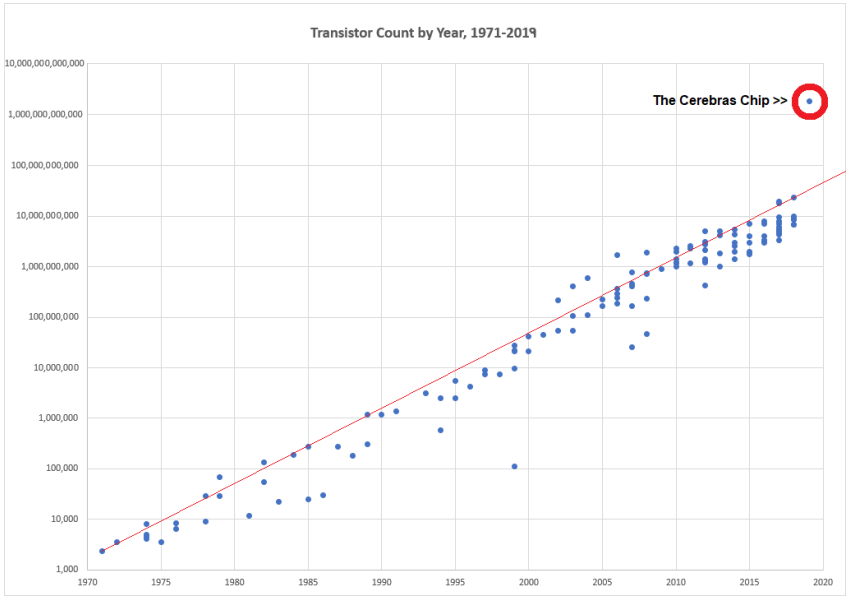

"For simple circuits, the cost per component is nearly inversely proportional to the number of components, the result of the equivalent piece of semiconductor in the equivalent package containing more components. But as components are added, decreased yields more than compensate for the increased complexity, tending to raise the cost per component. Thus there is a minimum cost at any given time in the evolution of the technology."

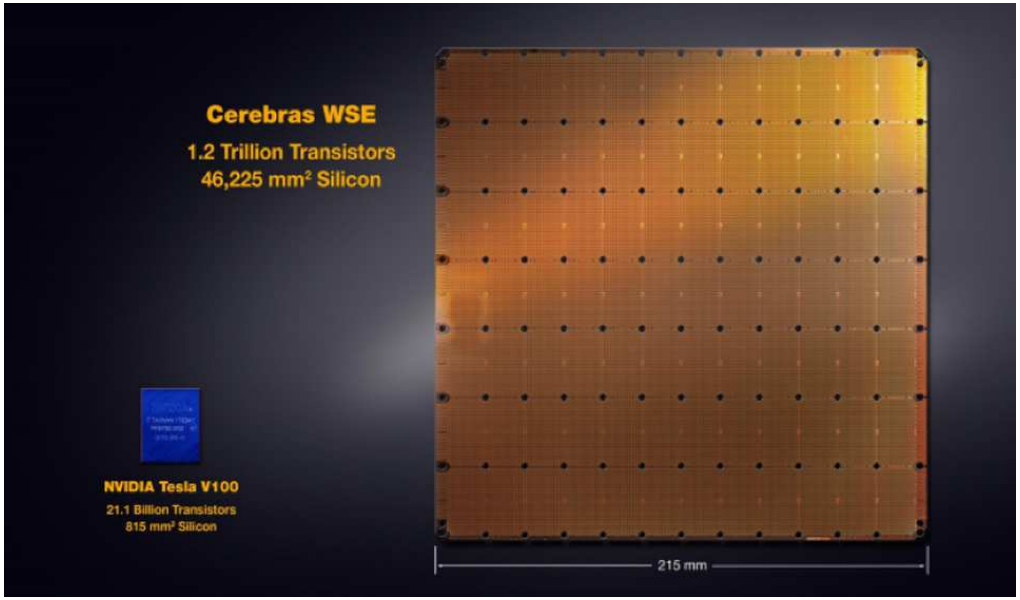
# Moore’s Law is Alive and Well




https://medium.com/predict/moores-law-is-alive-and-well-adc010ea7a63

# Moore’s Law is Alive and Well




https://medium.com/predict/moores-law-is-alive-and-well-adc010ea7a63

# Moore's Law

- Typically cast as:
  "Performance doubles every X months"
- Actually closer to:
  "Number of transistors per unit cost doubles every two years"

> The complexity for minimum component costs has increased at a rate of roughly a factor of two per year.
> […] Over the longer term, the rate of increase is a bit more uncertain, although there is no reason to believe it will not remain nearly constant for at least 10 years.
>
> -- Gordon Moore, Electronics, 1965

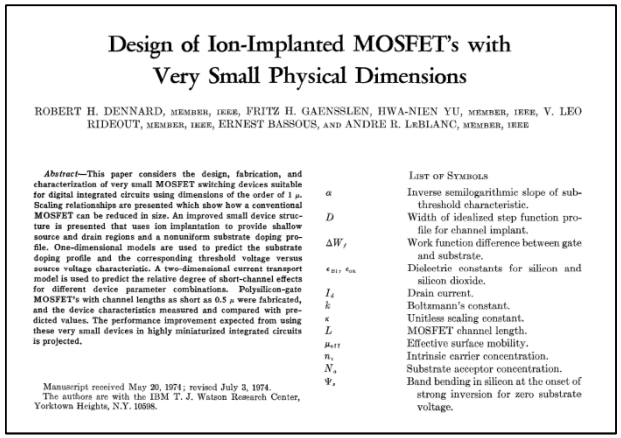
Why is Moore's Law conflated with processor performance?

# Moore's Secret Sauce: Dennard Scaling

[Dennard, Gaensslen, Yu, Rideout, Bassous, Leblanc, **IEEE JSSC**, 1974]

| Device or Circuit Parameter | Scaling Factor |
|---|---|
| Dimension, Tox, L, W | 1/k |
| Doping Concentration Na | k |
| Voltage (V) | 1/k |
| Current (I) | 1/k |
| Capacitance (eA/t) | 1/k |
| Delay time/circuit (VC/I) | 1/k |
| Power dissipation/circuit (VI) | 1/k^2 |
| Power density (VI/A) | 1 |

Historically, k ~= 1.4



Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions

ROBERT H. DENNARD, MEMBER, IEEE, FRITZ H. GAENSSLEN, HWA-NIEN YU, MEMBER, IEEE, V. LEO RIDEOUT, MEMBER, IEEE, ERNEST BASSOUS, AND ANDRE R. LEBLANC, MEMBER, IEEE

*Abstract*—This paper considers the design, fabrication, and characterization of very small MOSFET switching devices suitable for digital integrated circuits using dimensions of the order of 1 μ. Scaling relationships are presented which show how a conventional MOSFET can be reduced in size. An improved small device structure is presented that uses ion implantation to provide shallow source and drain regions and a nonuniform substrate doping profile. One-dimensional models are used to predict the substrate doping profile and the corresponding threshold voltage versus source voltage characteristic. A two-dimensional current transport model is used to predict the relative degree of short-channel effects for different device parameter combinations. Polysilicon-gate MOSFET's with channel lengths as short as 0.5 μ were fabricated, and the device characteristics measured and compared with predicted values. The performance improvement expected from using these very small devices in highly miniaturized integrated circuits is projected.

Manuscript received May 20, 1974; revised July 3, 1974.
The authors are with the IBM T. J. Watson Research Center, Yorktown Heights, N.Y. 10598.

List of Symbols

| | |
|---|---|
| $\alpha$ | Inverse semilogarithmic slope of subthreshold characteristic. |
| $D$ | Width of idealized step function profile for channel implant. |
| $\Delta W_s$ | Work function difference between gate and substrate. |
| $\epsilon_{Si}, \epsilon_{ox}$ | Dielectric constants for silicon and silicon dioxide. |
| $I_d$ | Drain current. |
| $k$ | Boltzmann's constant. |
| $\kappa$ | Unitless scaling constant. |
| $L$ | MOSFET channel length. |
| $\mu_{eff}$ | Effective surface mobility. |
| $n_i$ | Intrinsic carrier concentration. |
| $N_a$ | Substrate acceptor concentration. |
| $\Psi_s$ | Band bending in silicon at the onset of strong inversion for zero substrate voltage. |

2x transistor count

40% faster

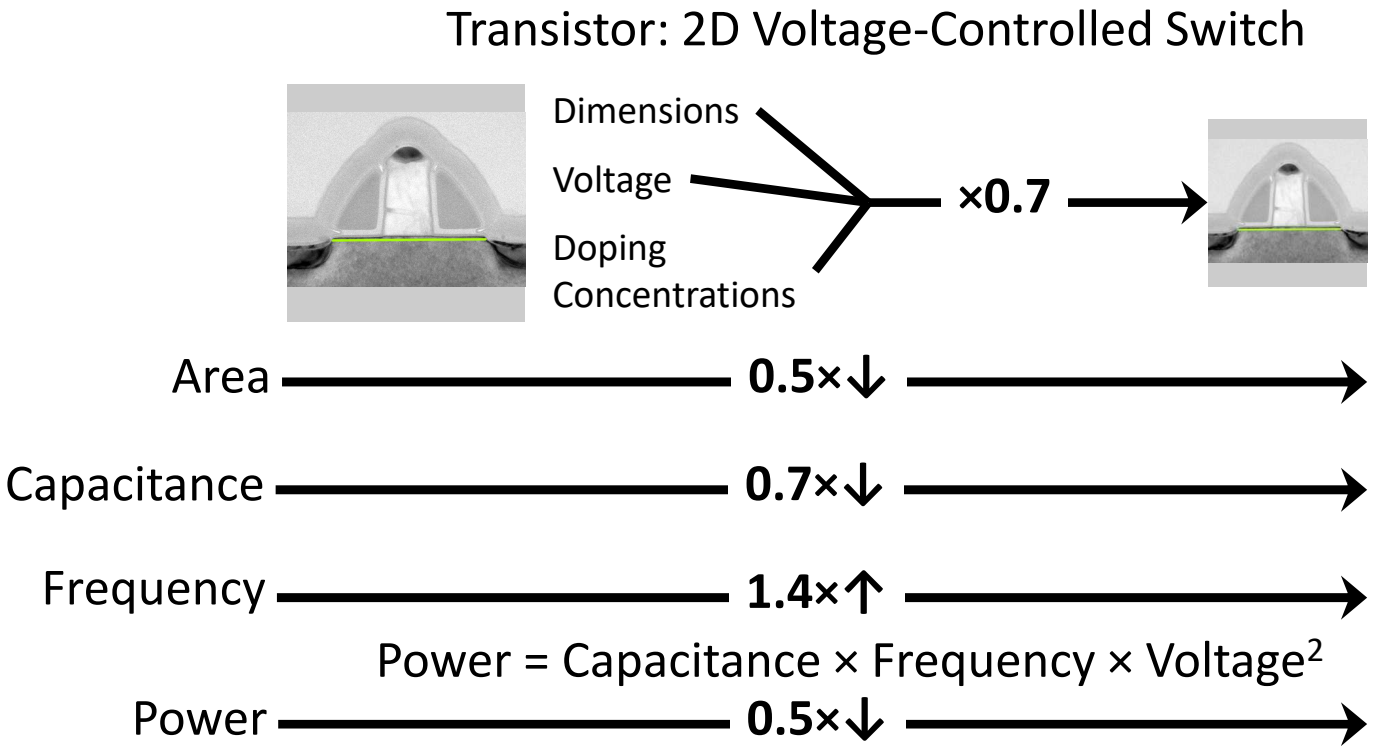
50% more efficient

# Dennard Scaling in Lay Man's Terms

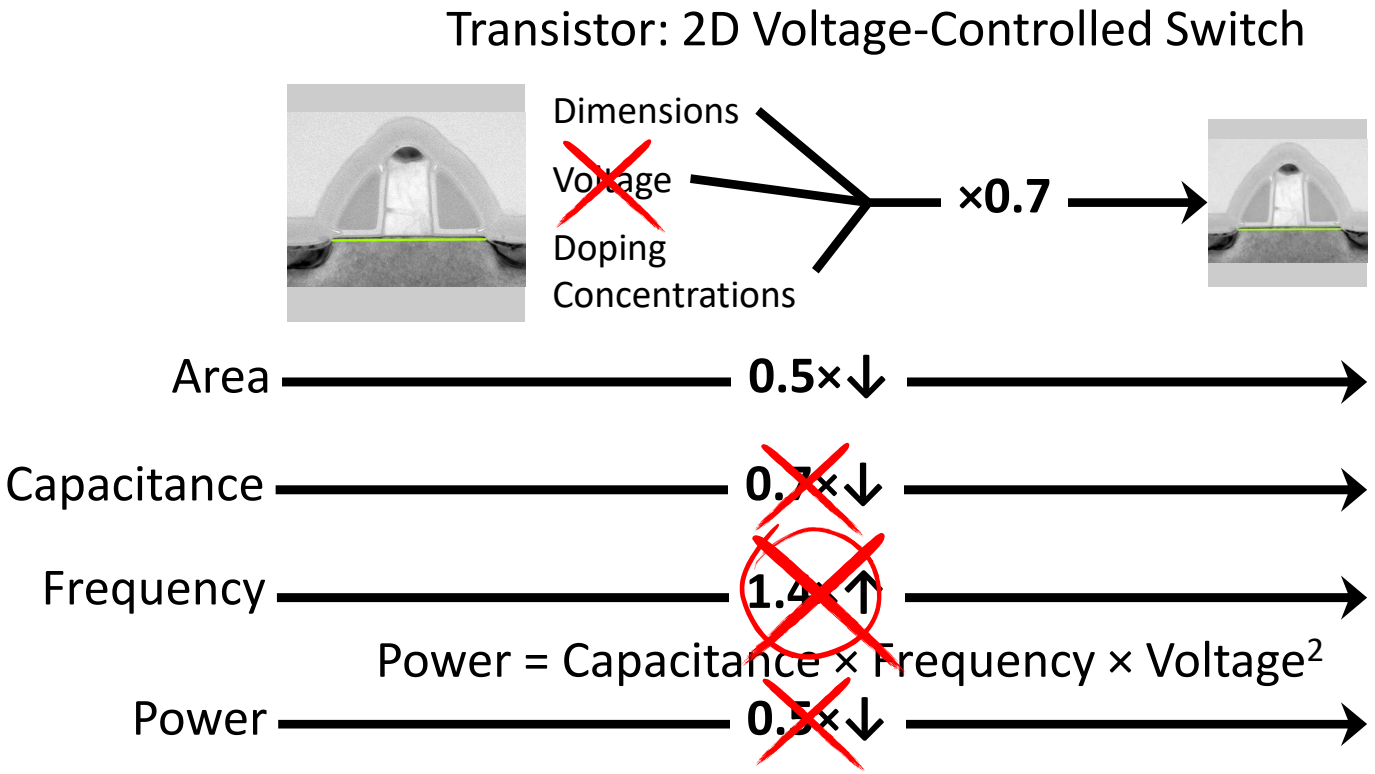
- "Power density stays constant as transistors get smaller"
- Intuitively
  - Smaller transistors -> shorter propagation delays -> faster frequency
  - Smaller transistors -> smaller capacitance -> lower voltage

  - Power ∝ Capacitance x Voltage^2 x Frequency
  - Moore's Law -> Faster performance @ constant power

# Dennard Scaling:
# Doubling the transistors; scale their power down



# Dennard Scaling Broke:
# Double the transistors; still scale their power down



# What Happened?

Only knob left to turn

Gate-oxide stopped scaling

Stopped scaling due to leakage

$$Power = (ActiveGateRatio \times Capacitance \times Voltage^2 \times Frequency)$$

Dynamic power

Stopped scaling due to leakage

$$+ (Voltage \times LeakageCurrent)$$

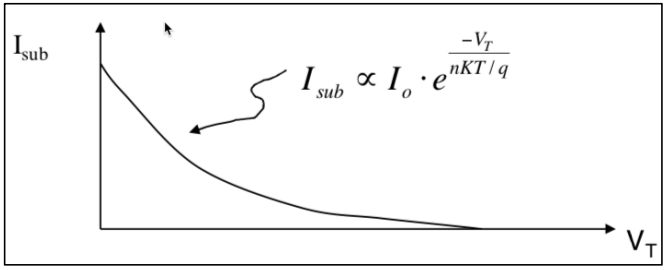
Static power

# Deeper Look At Leakage Current

$$I_{sub} = K_1 W e^{-V_{th}/nV_\theta}\left(1 - e^{-V/V_\theta}\right)$$

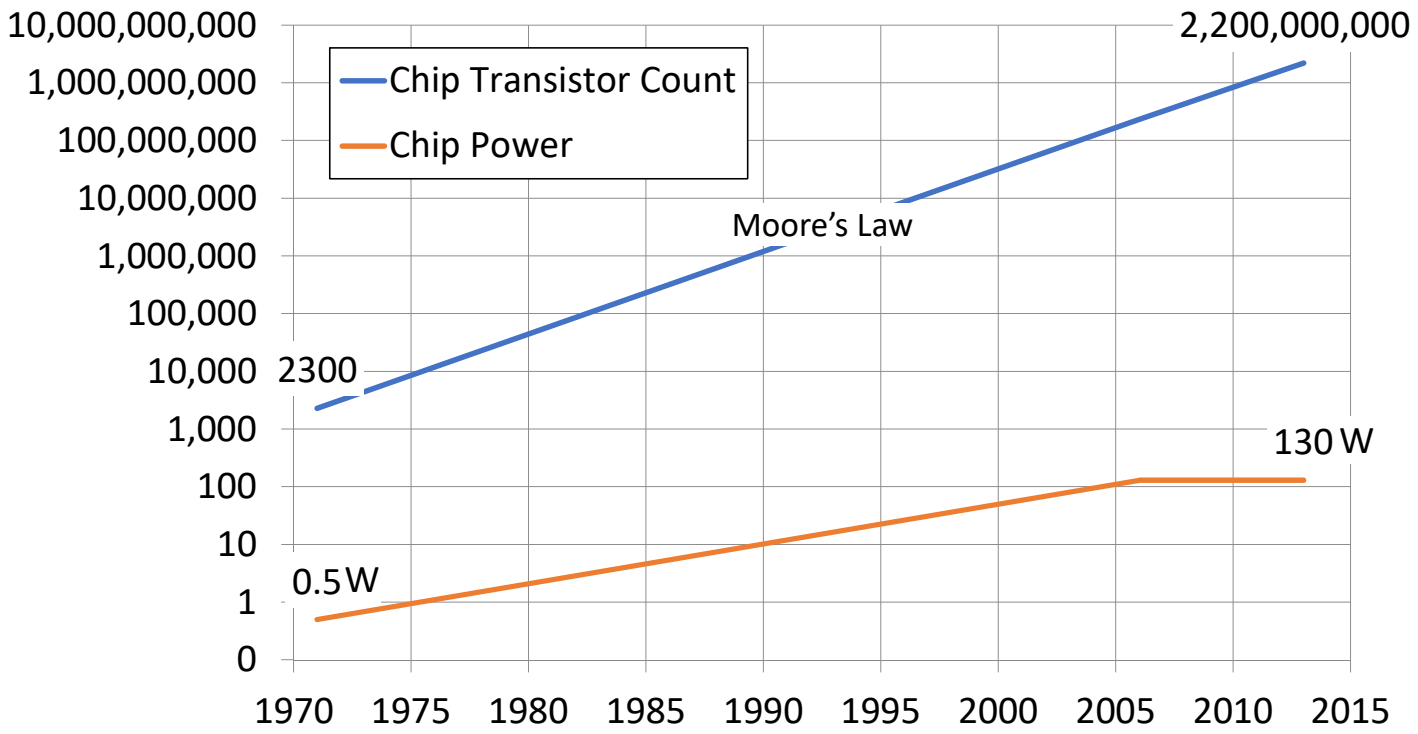
Constants: $K_1 W$, $e$, $nV_\theta$, $V_\theta$

- Subthreshold leakage: power leaked before voltage reaches threshold
  - Can be reduced by increasing threshold voltage ($V_{th}$) or decreasing voltage (V)
  - Lower voltage at the same threshold voltage -> unstable circuit
  - Threshold voltage does not scale at low size (Without reducing frequency … long story)
  - -> ***Voltage cannot scale!***

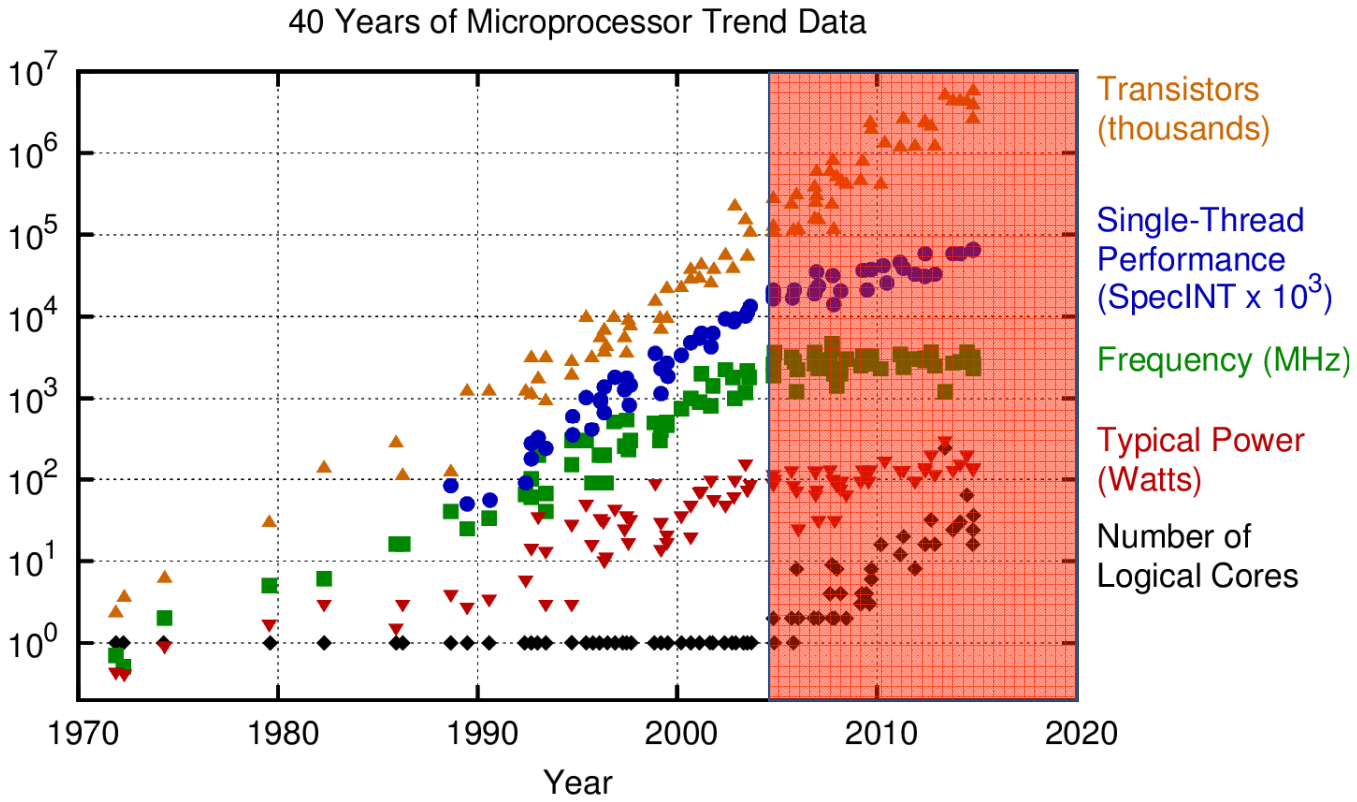


# So What's the Catch with Moore's Law?

## Powering the transistors without melting the chip



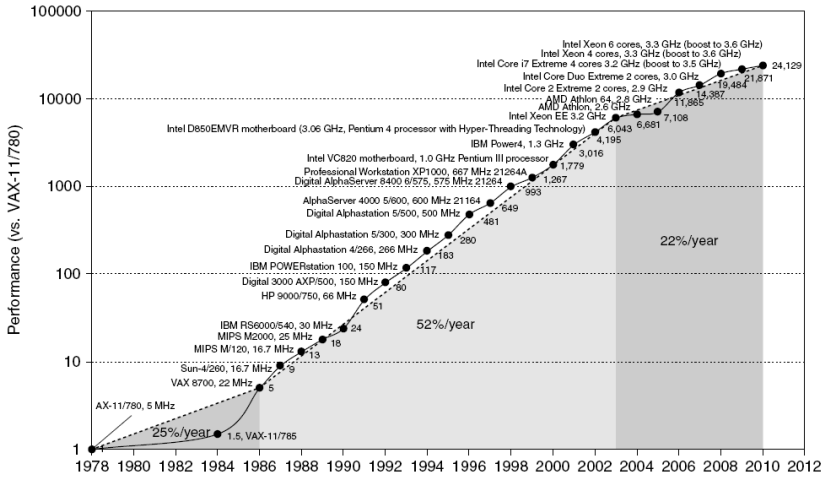
# Dennard Scaling Effects are Real




Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

# Historical Data

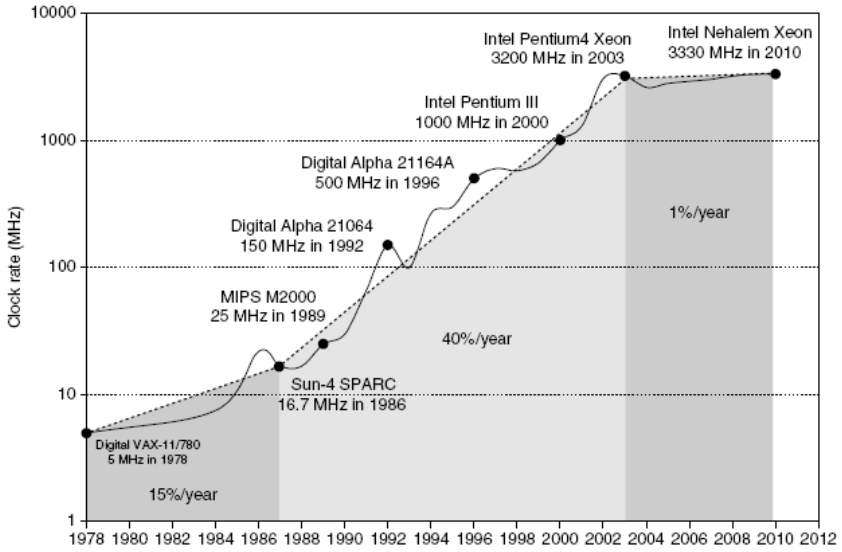
**Performance Improvement**




- The 52% growth per year is because of faster clock speeds and architectural innovations (led to 25x higher speed)
- The 22% growth includes the parallelization from multiple cores

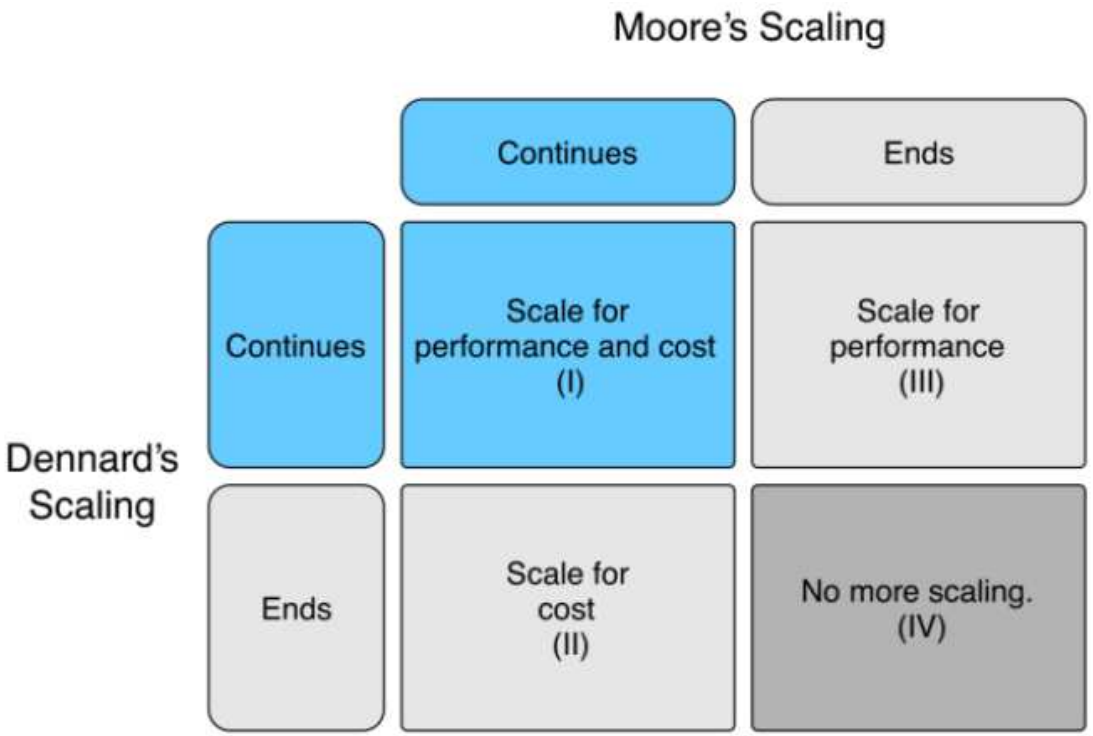
# Historical Data (2)

**Clock Frequency Improvement**

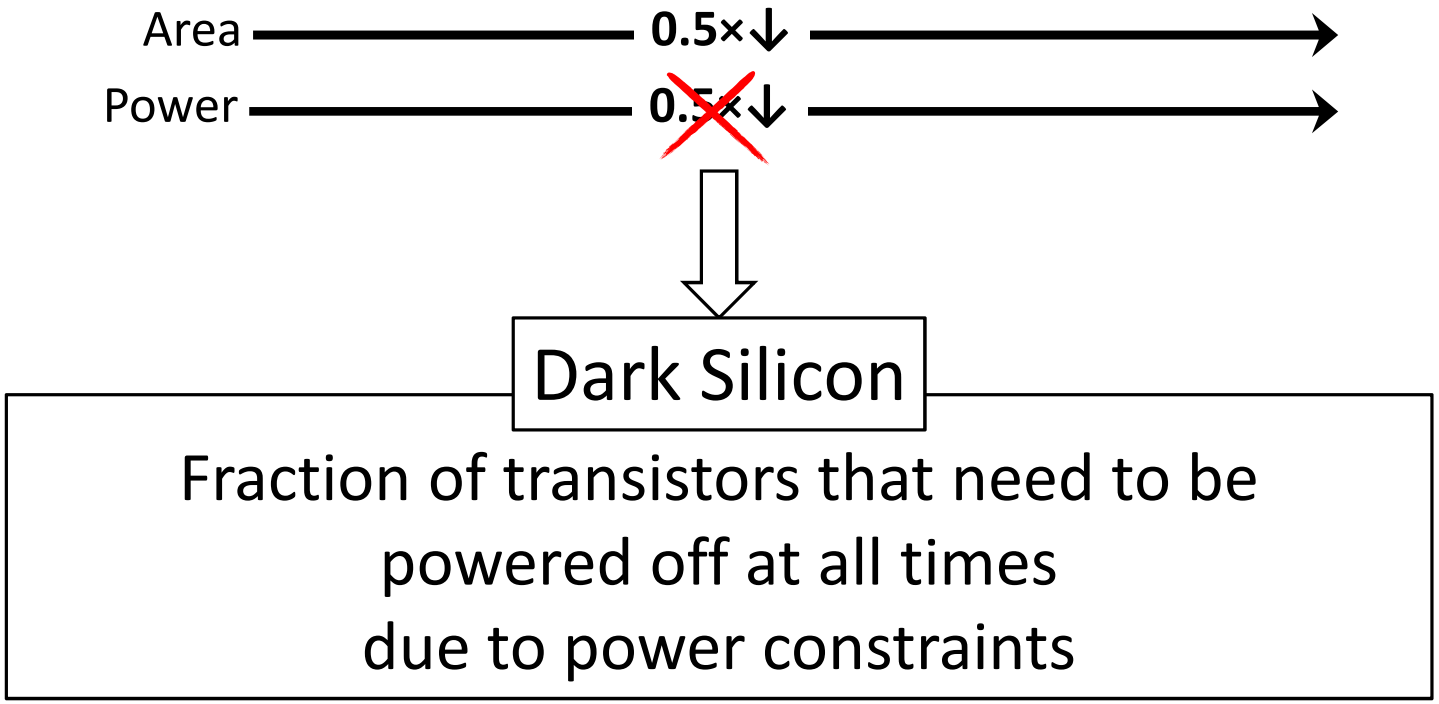



- The 52% growth per year is because of faster clock speeds and architectural innovations (led to 25x higher speed)
- The 22% growth includes the parallelization from multiple cores
- Clock speed increases have dropped to 1% per year in recent years

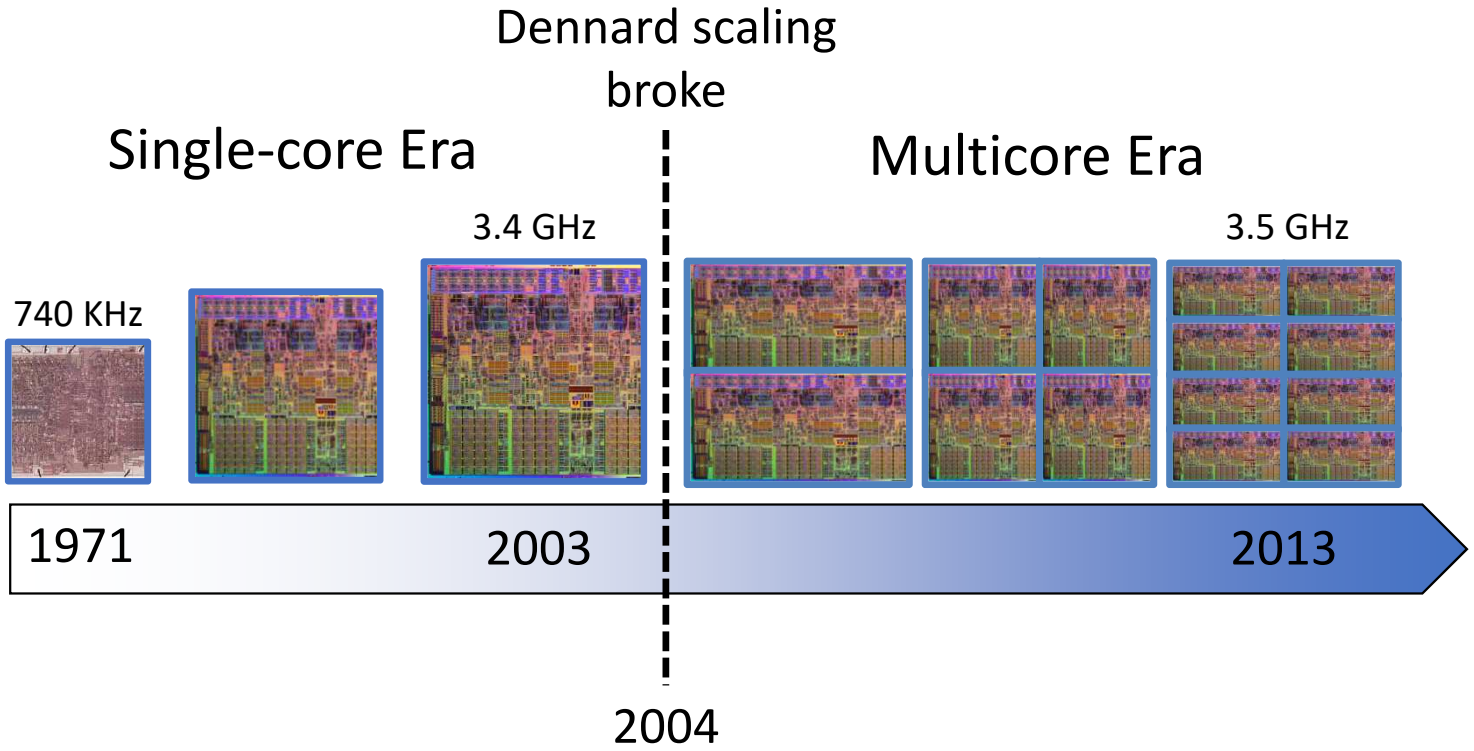
# Summary of Relationship between Dennard Scaling and Moore's Law



# Dark Silicon:
# If you cannot power them, why make them?



# Looking back
# Evolution of processors



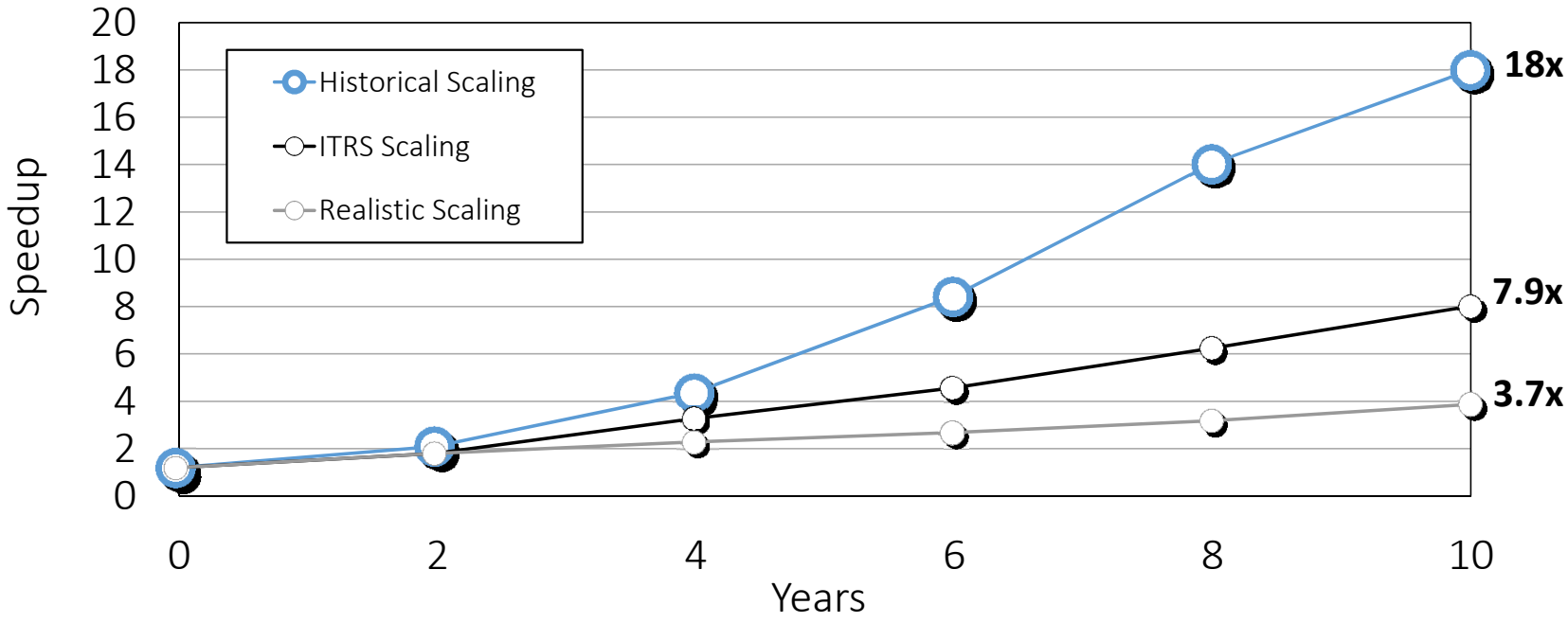
# Multicore model (Amdahl's Law)

$$\text{Speedup} = \frac{1}{\frac{1-f_{Parallel}}{\text{Serial Speedup}} + \frac{f_{Parallel}}{\text{Parallel Speedup}}}$$

$$\text{Serial Speedup} = 1 \times \text{Core Performance}$$

$$\text{Parallel Speedup} = \text{N} \times \text{Core Performance}$$

# Multicore to the rescue?

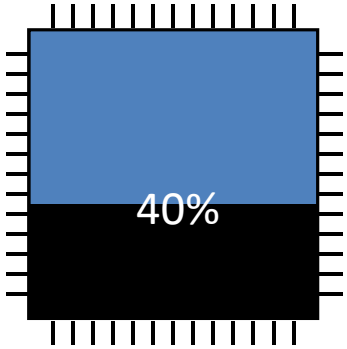



[Esmaeilzadeh, Blem, St. Amant, Sankaralingam, Burger, ISCA 2011]

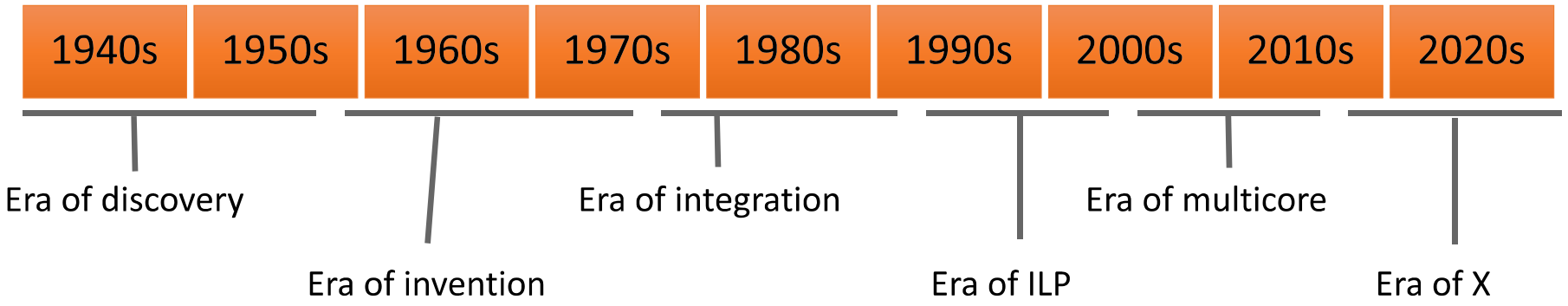
# Dark silicon

$$N_{Core} = min(\frac{Area\ Budget}{Area_{Core}}, \frac{Power\ Budget}{Power_{Core}})$$

$$Dark\ Silicon = 1 - \frac{N_{Core} \times Area_{Core}}{AreaBudget}$$

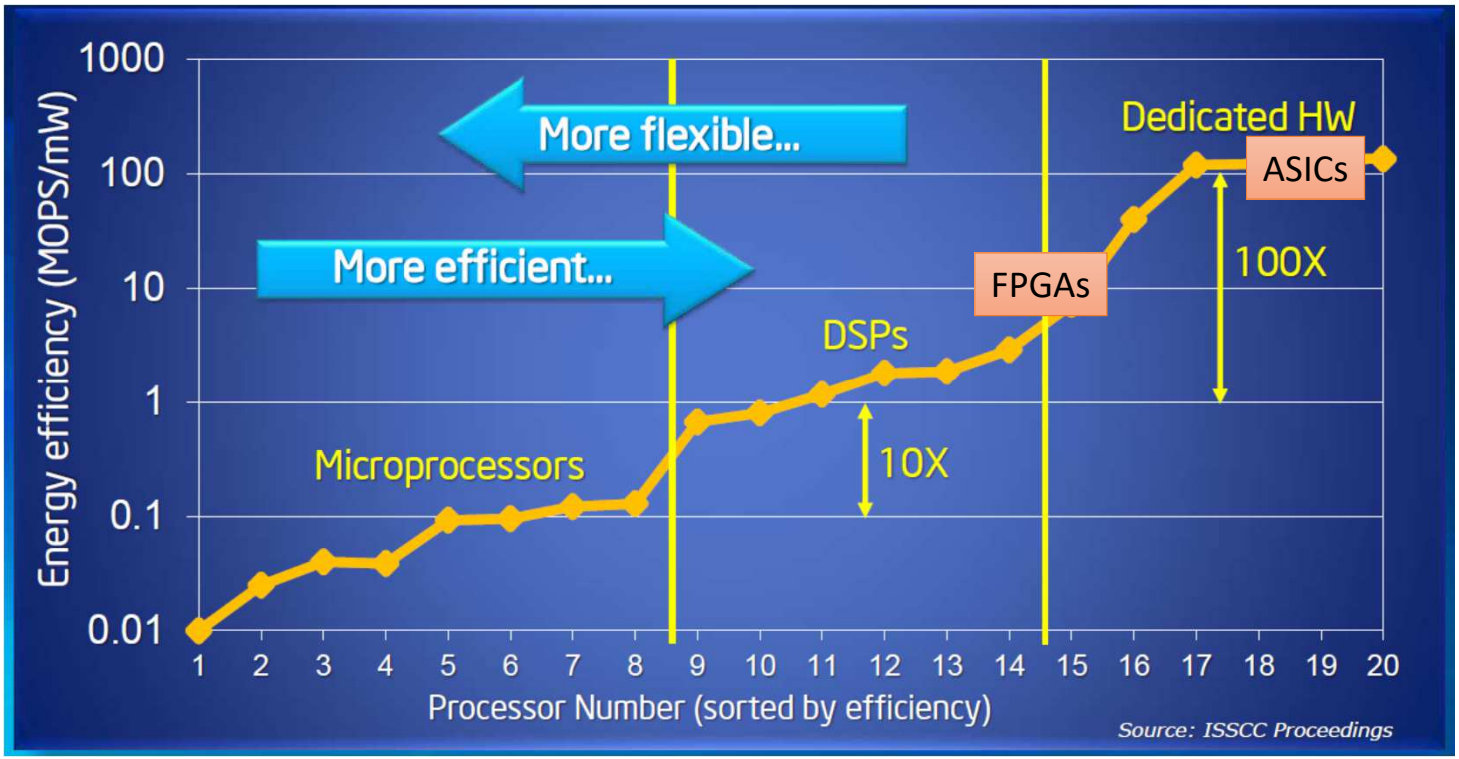


# A brief history of computer architecture

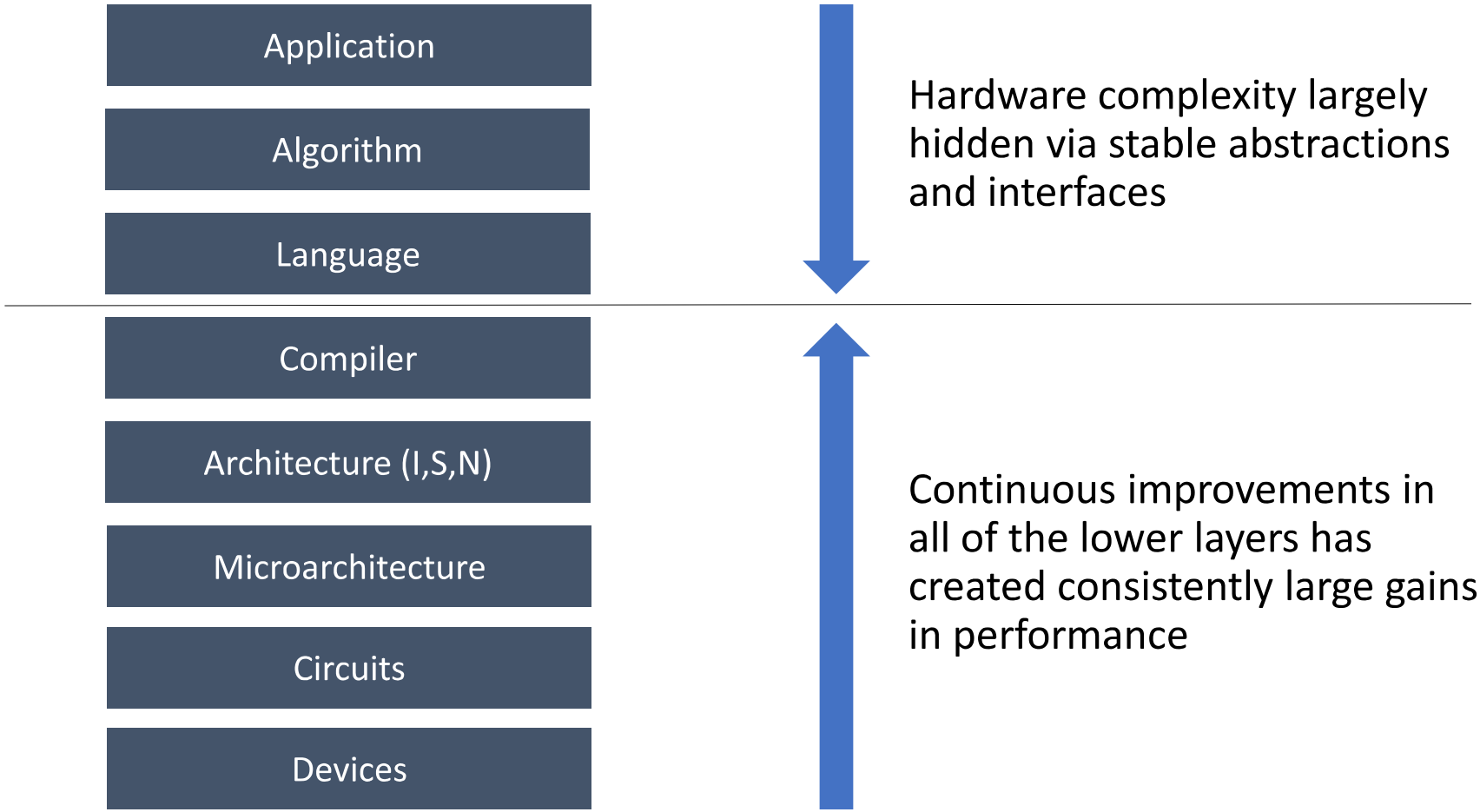


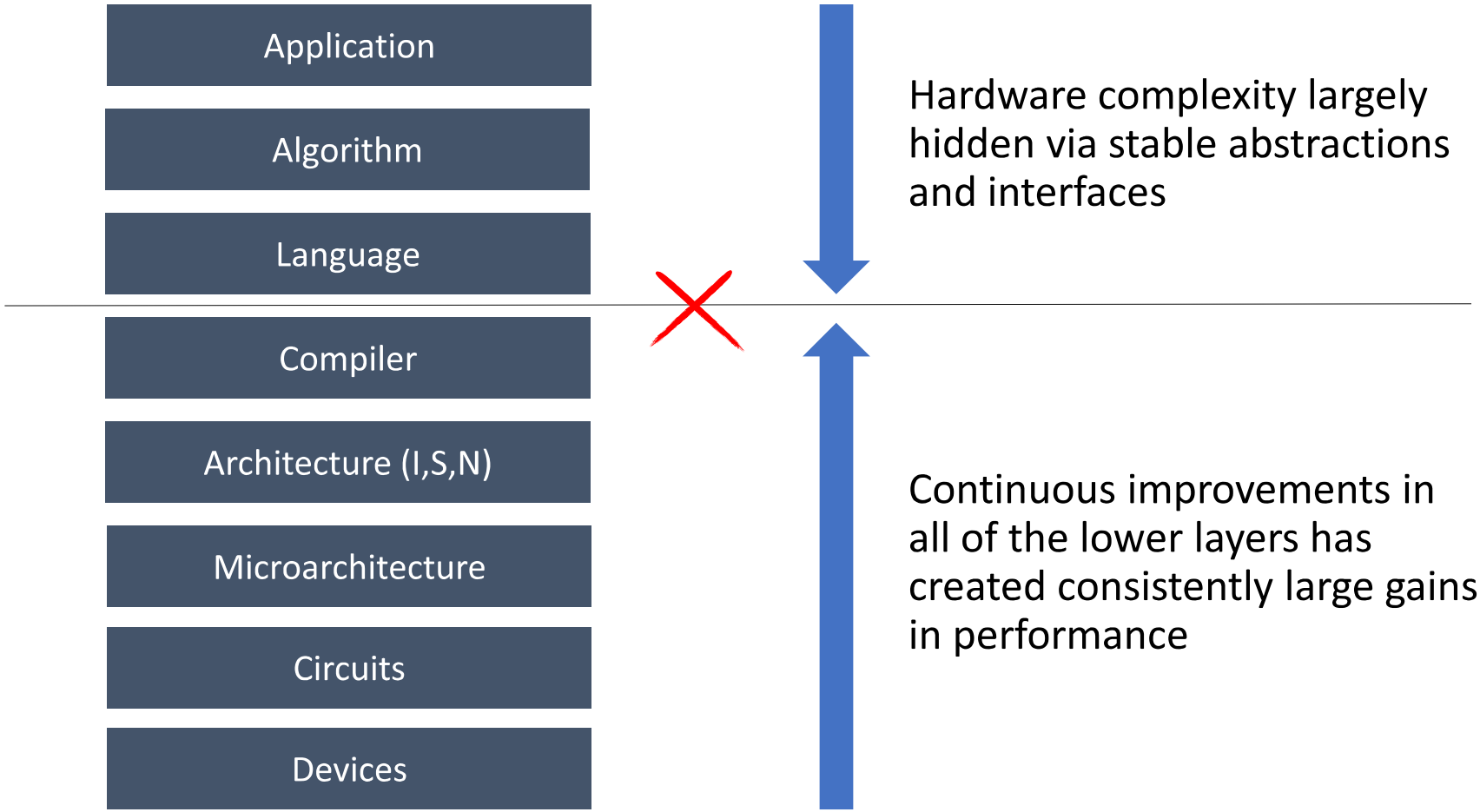

X = {Logic specialization, neural computing, cold computing, ?}

# Specialization: A path forward (?)

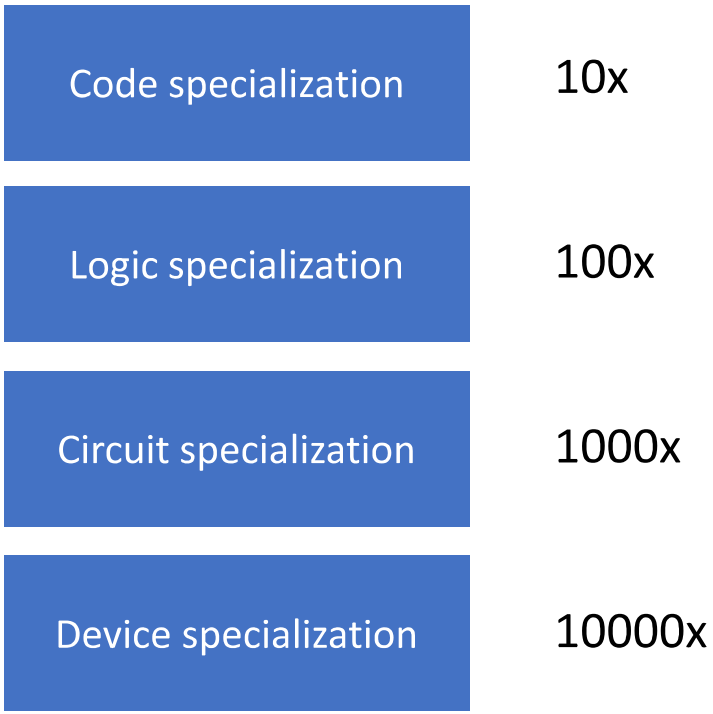



Source: Bob Broderson, Berkeley Wireless group


Application
Algorithm
Language
Compiler
Architecture (I,S,N)
Microarchitecture
Circuits
Devices
Hardware complexity largely hidden via stable abstractions and interfaces
Continuous improvements in all of the lower layers has created consistently large gains in performance


Application
Algorithm
Language
Compiler
Architecture (I,S,N)
Microarchitecture
Circuits
Devices
Hardware complexity largely hidden via stable abstractions and interfaces
Continuous improvements in all of the lower layers has created consistently large gains in performance

# More gains the lower you go



# So… What's Next?