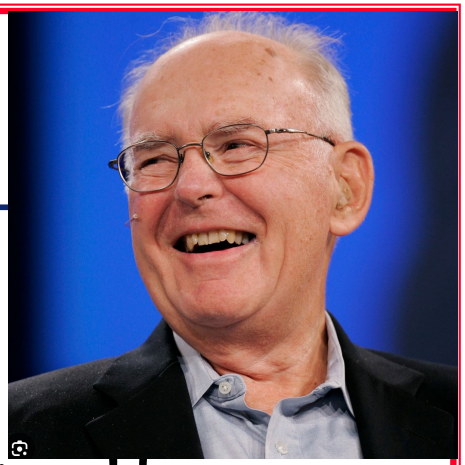
Domain Specific Architectures
CSCE 4013/5013

David Andrews
Rm 527 JBHT
dandrews@uark.edu
CSCE University of Arkansas

Agenda

- Moore's Law
- Dennard Scaling
- Power, Energy

Moore's Law

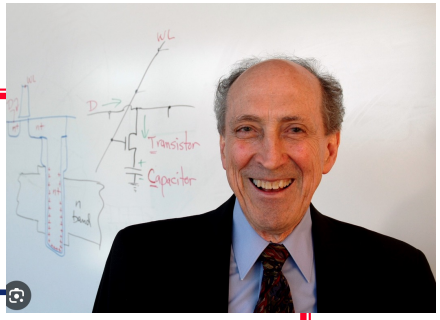


- Gordon Moore looked at several generational chips and did quick math
- "The complexity for minimum component costs has increased at a rate of roughly a factor of two per year"……
- Interpretation – the number of components that can be fabricated in a chip is doubling every year…….

Moore's Law

- Number of Transistors in an IC doubles every year (later 18 months) because:
  - The advent of metal-oxide-semiconductor (MOS) technology
  - The exponential rate of increase in die sizes, coupled with a decrease in defective densities, with the result that semiconductor manufacturers could work with larger areas without losing reduction yields
  - Finer minimum dimensions
  - What Moore called "circuit and device cleverness"

Moore's Law Secret Sauce: Dennard Scaling



- Dennard observed that transistor dimensions could be scaled by 30% (0.7x) every technology generation, thus reducing their area by 50%.

reduce circuit delays by 30% (0.7x)

increase frequency by ~ 40% (1.4x)

voltage is reduced by 30%, reducing energy by 65% and power (at 1.4x frequency) by 50%

$$\text{Power} = CV^2f$$

If the transistor density doubles, the circuit becomes 40% faster, and power consumption (with twice the number of transistors) stays the same! What ????

Carver Mead Explains the Physics

https://www.youtube.com/watch?v=UFa_tk3K5oY

Moore's Law Enabled the "Virtuous Cycle"

Transistor Scaling

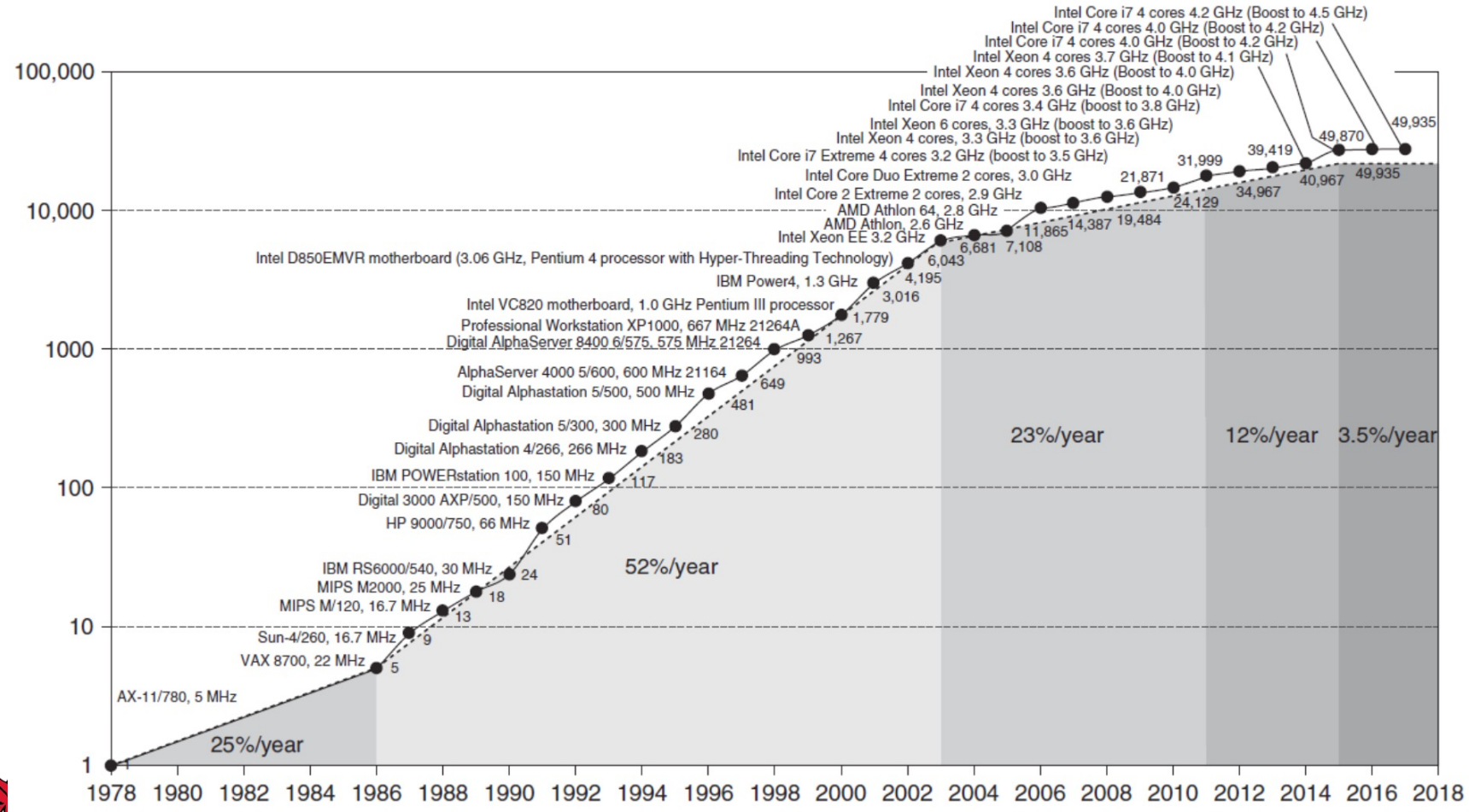
Investment

Better Performance/Cost

Market Growth

From ITRS.

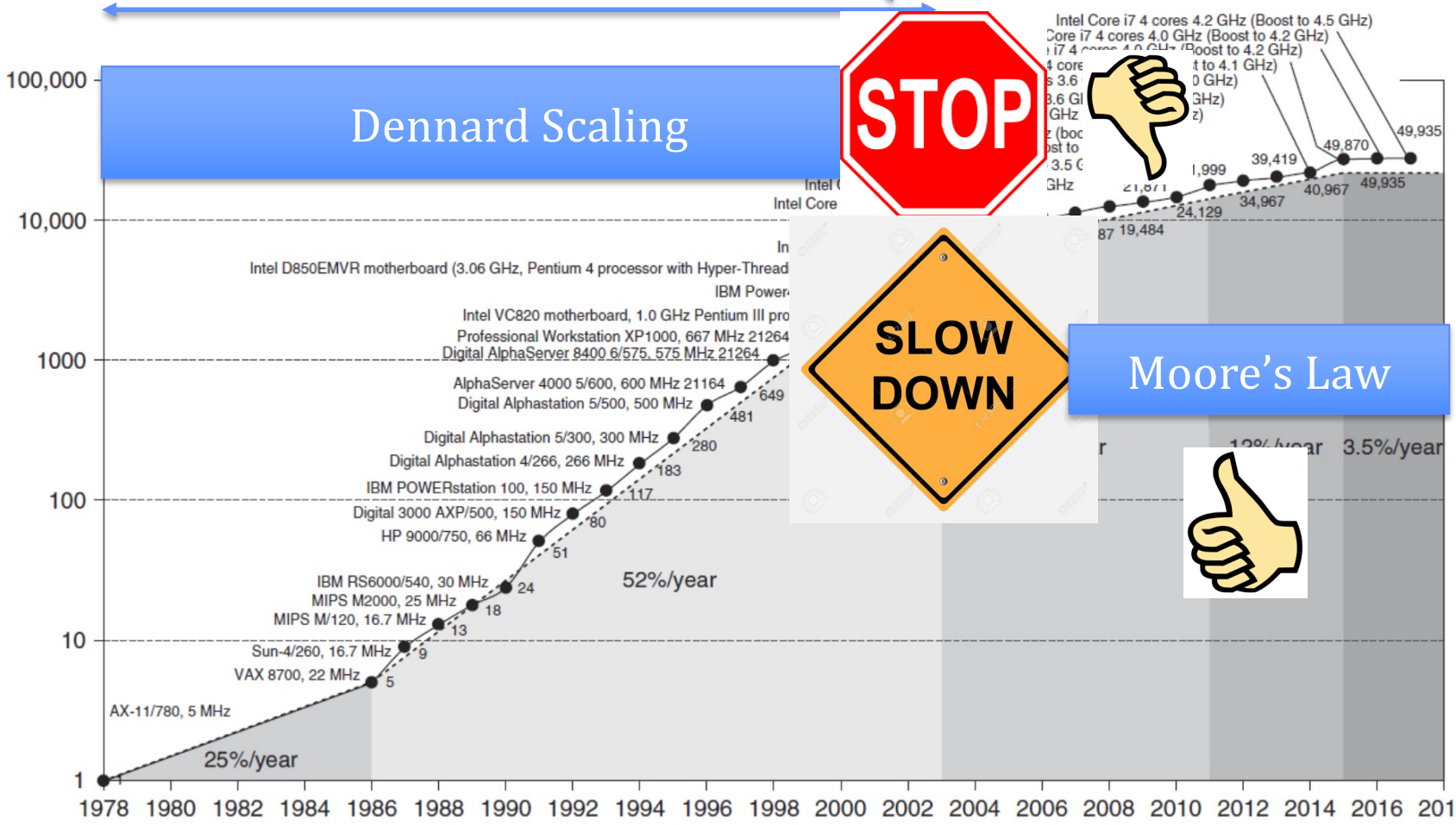
Performance versus VAX-11/780 (look it up ☺)



Rapid Changes You Say ?? : *Cause & Effect*

Era of 'LIFE IS GOOD"

Era of 'Manycore"



Power and Energy

Power: How fast energy is transmitted $P = \frac{\Delta E}{\Delta \tau}$ Watt = joule/sec

Energy: Ability to create a change Joules = watt-second

Energy Can Be Stored, Power Cannot

https://energyeducation.ca/encyclopedia/Energy_vs_power

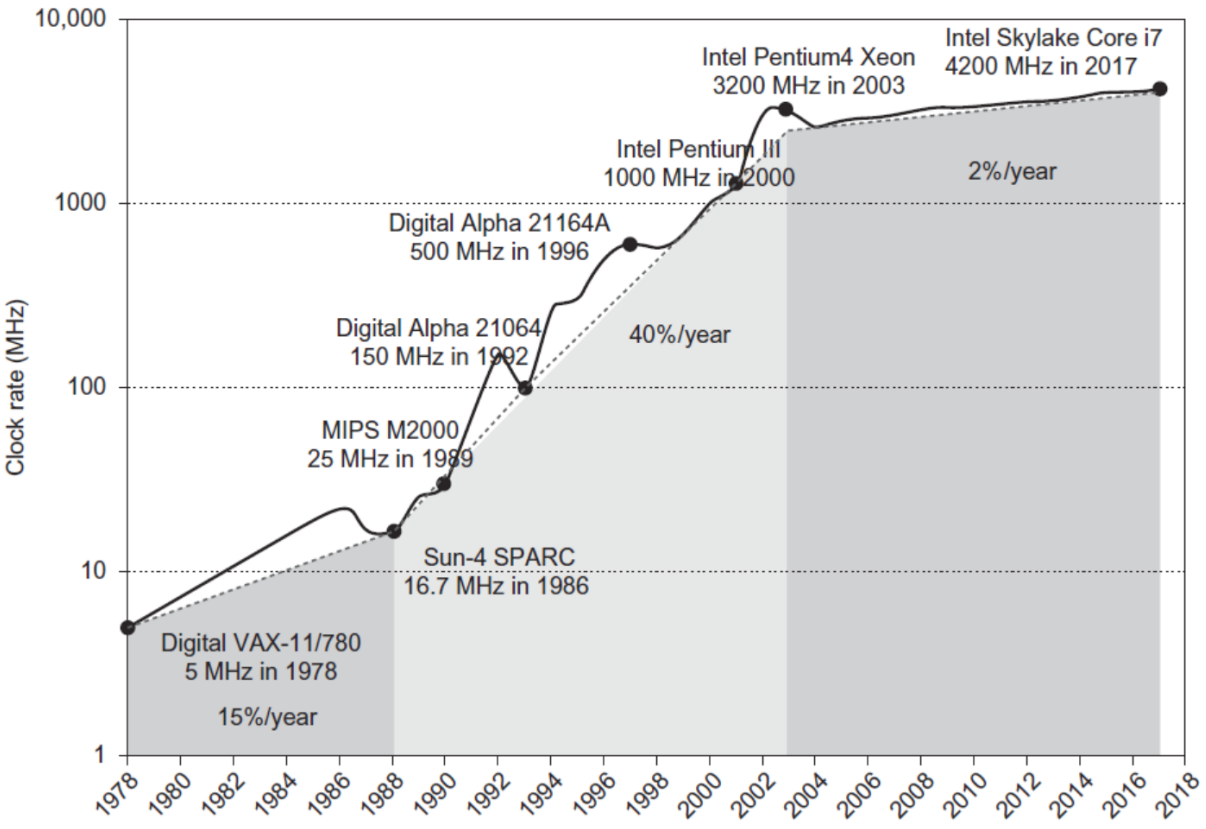
http://www.tecategroup.com/ultracapacitors-supercapacitors/ultracapacitor-FAQ.php]]

What Happened ?

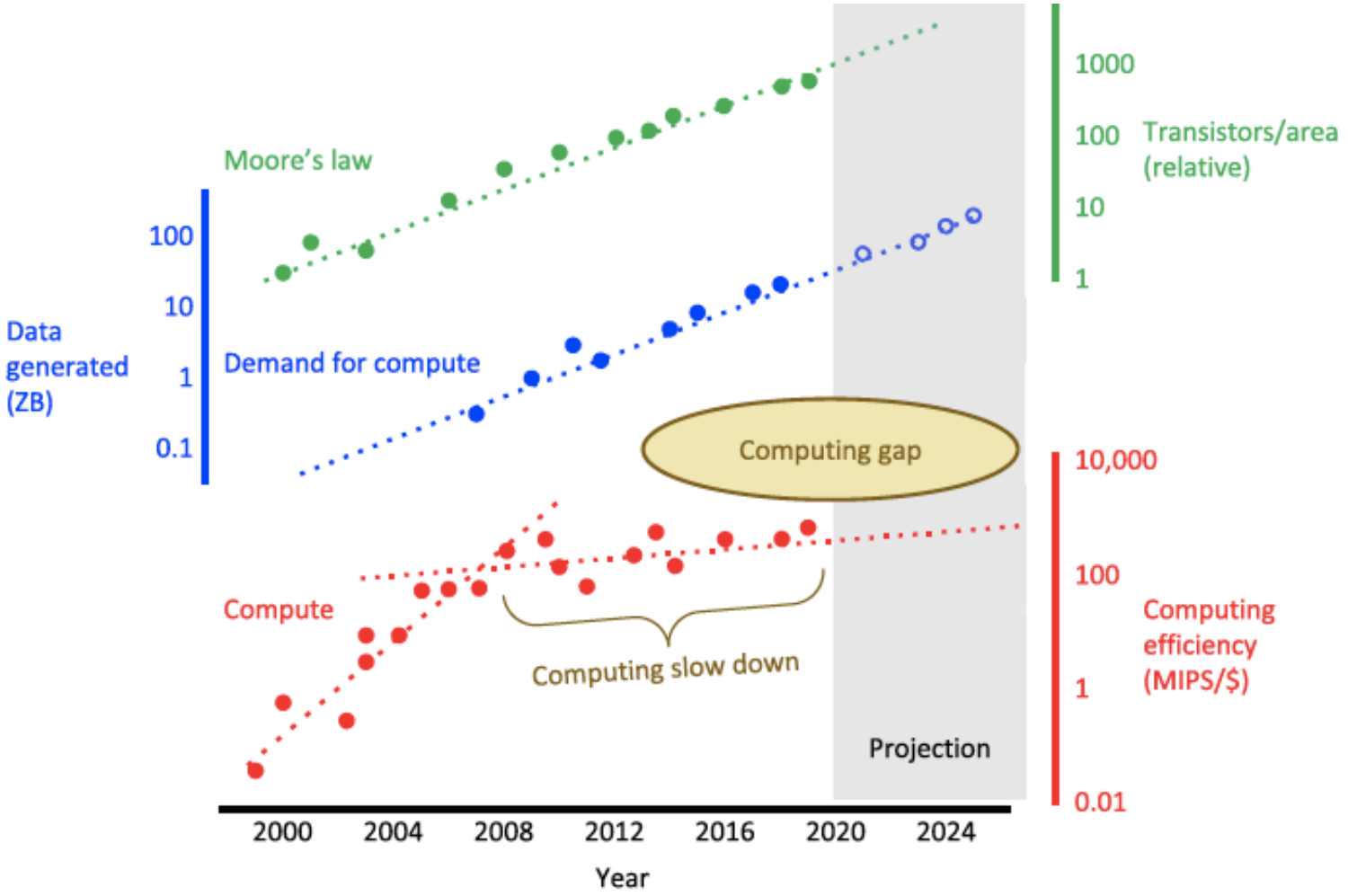
- The dynamic (switching) power consumption of CMOS circuits is proportional to frequency ($P = CV^2f$) .
- Historically, the transistor power reduction afforded by Dennard scaling allowed raising clock frequencies from one generation to the next without significantly increasing overall circuit power consumption.
- breakdown of Dennard scaling resulted in the inability to increase clock frequencies. CPU manufacturers switched to multicore processors as an alternative way to improve performance.

Power

- Intel 80386 consumed ~ 2 W
- 3.3 GHz Intel Core i7 consume 130 W
- Heat must be dissipated from 1.5 x 1.5 cm chip
- This is the limit of what can be cooled by air



We Created a Monster That Needs Continual Feeding !!!



J. D. Kendall and S. Kumar, "The building blocks of a brain-inspired computer," Applied Physics Reviews, vol. 7, no. 1, 2020.

Performance Via Parallelism

- Cannot Clock Faster so Do More In Parallel
  - Apply Transistors to Exploit Parallelism
  - Parallelism Exists at Different "Granularities"
  - Circuit, Data, Instruction, Procedural, Program….
- Implicit Parallelism within a Processor
  - Out of Order Instruction-Level parallelism (ILP)
  - Speculation
- From the Application Program
  - Data-level parallelism (DLP)
  - Thread-level parallelism (TLP)
  - Domain Specific Acceleration (DSA)

Further Fun…

Gordon Moore & Carver Mead: *Moore's Law 40th Anniversary with Gordon Moore*

https://www.youtube.com/watch?v=MH6jUSjpr-Q