Moore's Law · Bent Not Broken

D. Ferguson, IEEE Life Senior Member

Presentation Goals

- Review Moore's Law from three perspectives
  - observations on classic transistor scaling
  - scaling impact on chip building
  - the economic impact of improved productivity
- Describe near·future trends that will maintain the status quo.
- Consider how the industry may unfold in the foreseeable future (20 years hence) when Moore's Law is not viable.

Amara's Law: we tend to overestimate progress in the near term and underestimate progress in the far term.

From left, Andrew Grove, Robert Noyce, and Gordon Moore

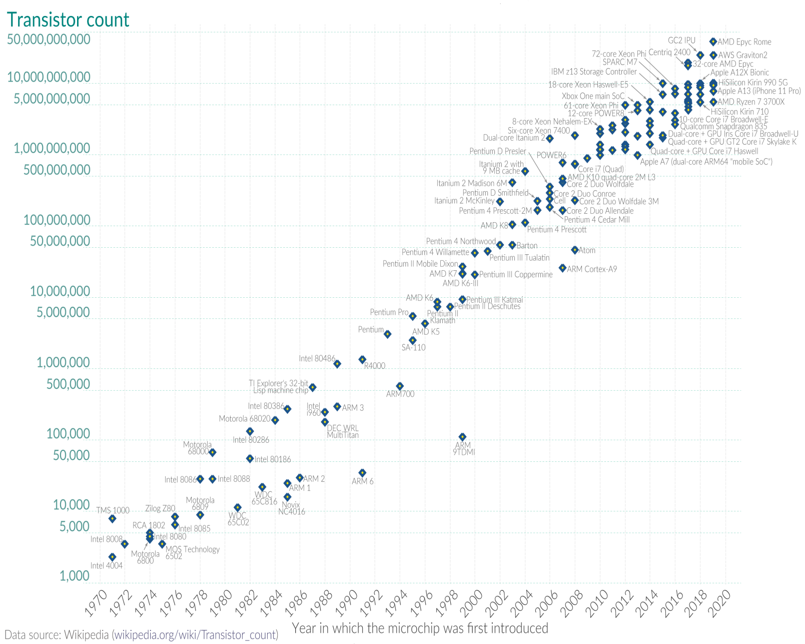
Disclaimer

- There is a huge body of literature on Moore's Law (over 15,939 IEEE conference papers, journal papers, magazine articles, and books since 2000 alone). I am "cherry picking" issues and examples to present a view of where I think we are going. There is no way to be all inclusive and if I've missed your favorite perspective, I apologize.
- My perspective:
  - Involved in telecommunications and RF/mixed signal chip design since the late 70's
  - Majority of my career is at Honeywell's research facility and startups

Moore's Law · Certainly Bent if Not Broken

Moore’s Law

In 2015, Gordon Moore reflected on his 1965 paper, **“The message I was trying to get across was that integrated circuits were the road to less-expensive electronics. It really evolved from being a measure of what goes on in the industry to something that more or less drives the industry.”** [1]

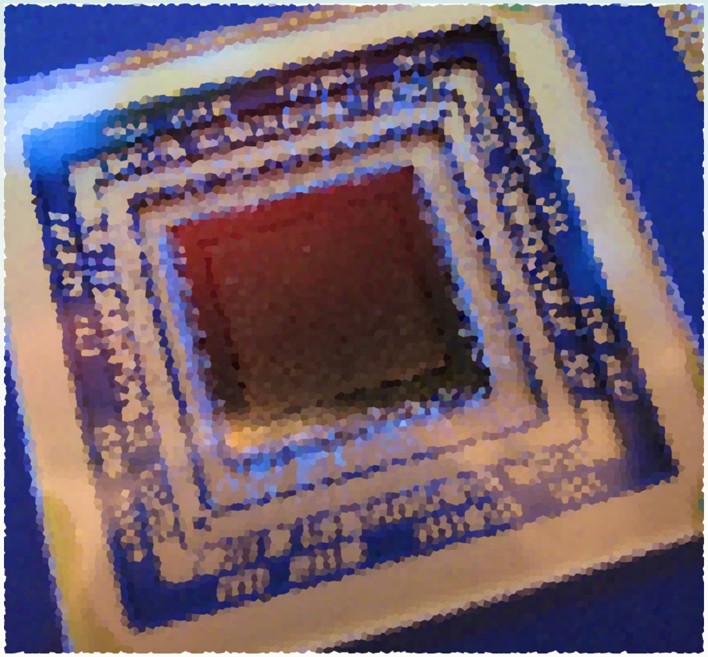


Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)
OurWorldinData.org – Research and data to make progress against the world’s largest problems.

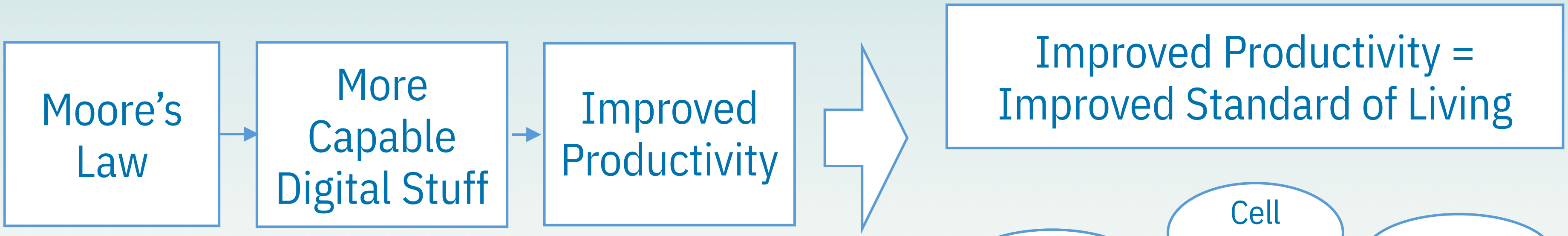
In popular culture, Moore’s Law has become a shorthand for better processors and generically, greater productivity through digital stuff.

Moore’s Law · Observations 1

- It impacts almost everything (as the pandemic has recently demonstrated). The impact of Moore’s Law on the world economy cannot be overstated.
- Note that only the highest volume products (e.g. cell phones, laptop processors, servers) can afford to use the latest (smallest) feature processes. Smaller volume products (e.g. TV remotes, key fobs) still use older processes (e.g. 180/130nm or larger).



Moore's Law Impacts Productivity



Moore's Law → More Capable Digital Stuff → Improved Productivity

Improved Productivity = Improved Standard of Living

Cell Phones

Internet

Spreadsheets

Robots

DSP

Uber

Inexpensive Video

GPS

Amazon

Industrial Automation

Engine Controls

Zoom

Bar Code Readers

Remote Work

US population engaged in agricultural employment:

1900: ~17%

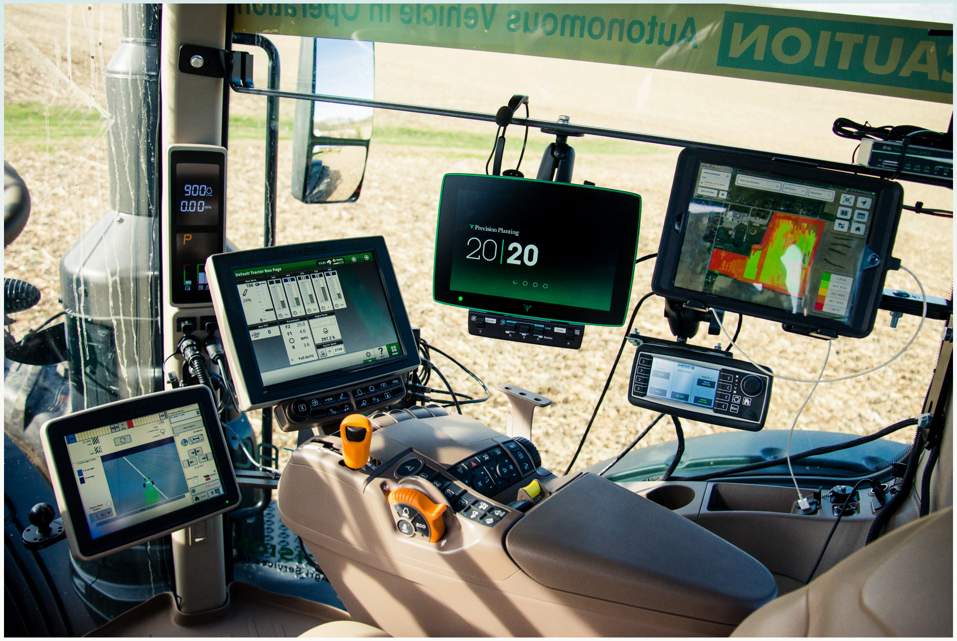
2020: <2%*

*Caution, there are 4.3x as many people in the US in 2020 compared to 1900. Nevertheless, fewer people grow more far more agricultural product.

I will use agribusiness as a stand-in for industrial or commercial productivity.

Agribusiness · A Modern Tractor

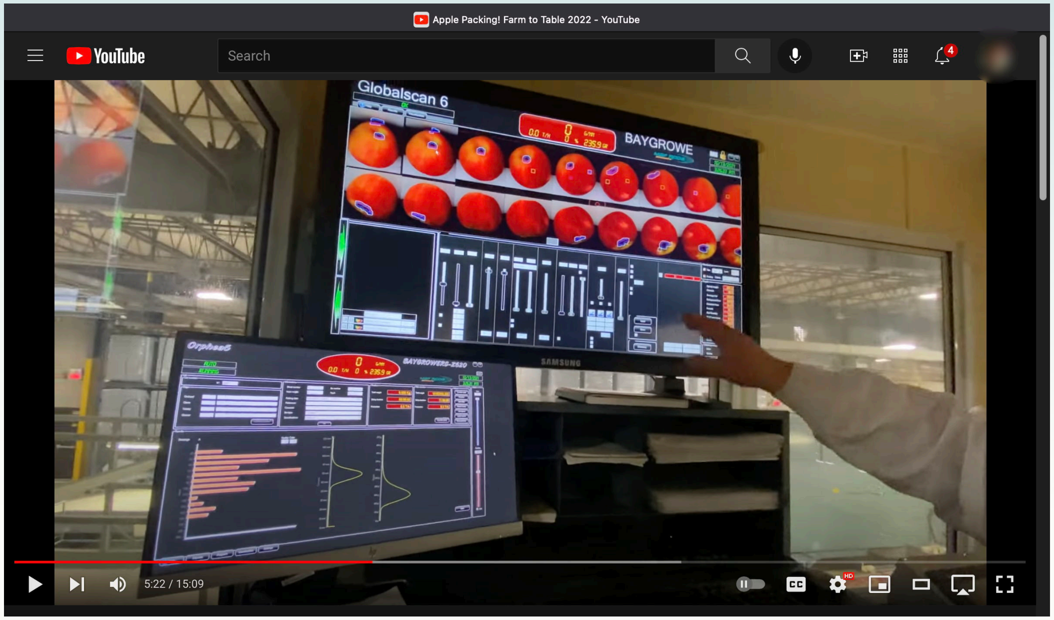
- Multiple screens for tracking progress
- GPS autonomous driving
- Precision planting (tracks amount of fertilizer/pesticide needed at a given point in the field)
- Crop yield precision tracking
- Crop moisture content
- Low down time, tells John Deere its health status and lets driver know when an "implement" experiences failure



The Millennial Farmer on YouTube gives a good perspective on farming automation

Agribusiness · Product Preparation

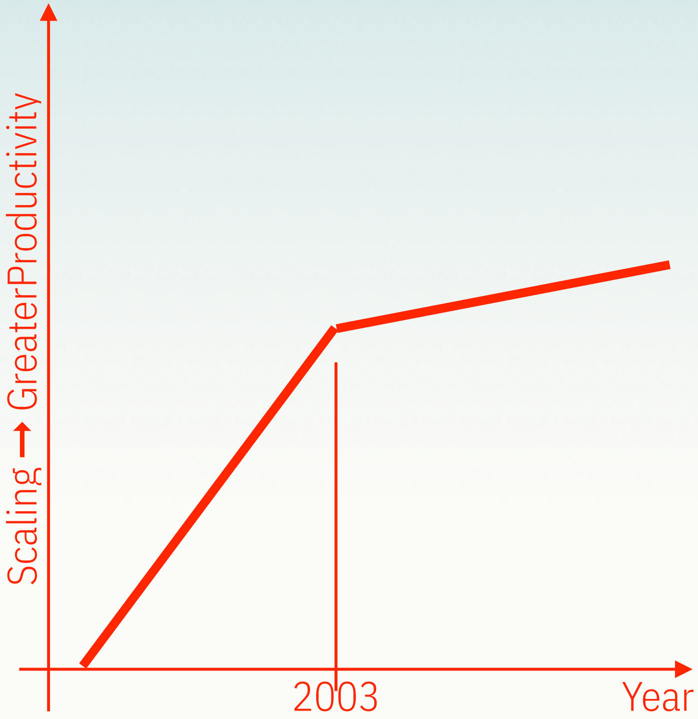
- This Kate's Ag video illustrates the automation that goes into picking apples and getting them to your store. It includes inspection of each apple for blemishes and field-to-store tracking.
- African farmers are using cell phones to contract and then ship their crop as well as gain weather info, crop insurance, etc.



https://www.youtube.com/watch?v=rKHDWTRyU54

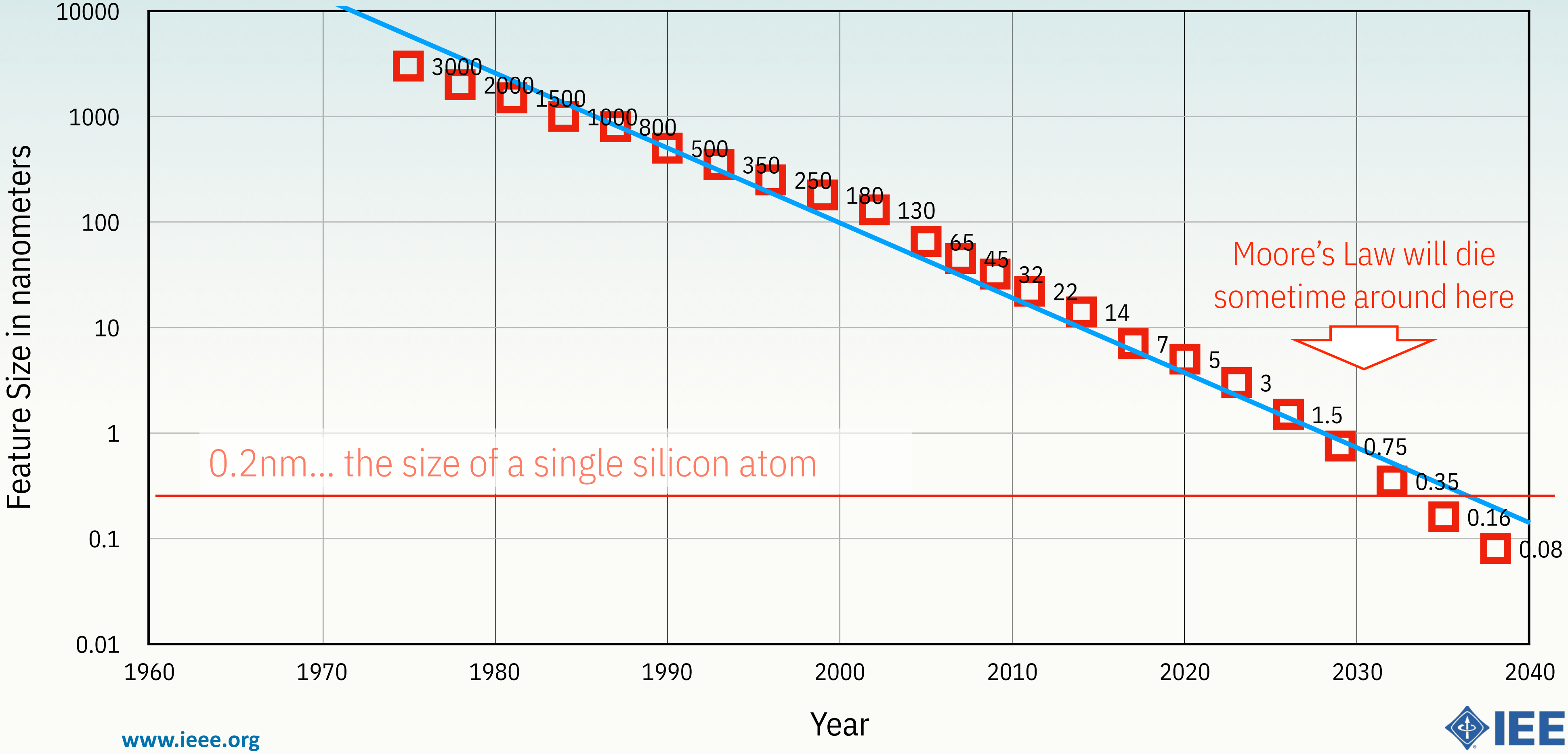
Moore's Law · Observations 2

- A modern laptop has 10^{15} times as much raw computing power as the discrete transistor IBM 7090 commonly used in 1961.[2]
- In addition to raw transistor count, this includes architecture improvements, multiple cores, cache, faster memory access, etc.
- Gate length has become meaningless as a driving performance measure. Circuit innovation (in its broadest sense) is needed to take advantage of all those transistors.
- Finally, if transistors don't get smaller, the chips need to get bigger.



Traditional Scaling Died ≈2003

Semiconductor Process Node



Moore’s Law · Observations 3

- Moore’s Law is dead! This really only impacts a small part of the overall electronics industry.
- There are still billions of transistors available for circuit innovation and new chip architectures. Process price will decline and then level out.
- As of early March 2022, AMD had a greater market value than Intel and AMD is fabless.

Chip Building or Meeting Moore's Law Required More than Just Shrinking Feature Size

Thermal Management 1 · Clock Frequency

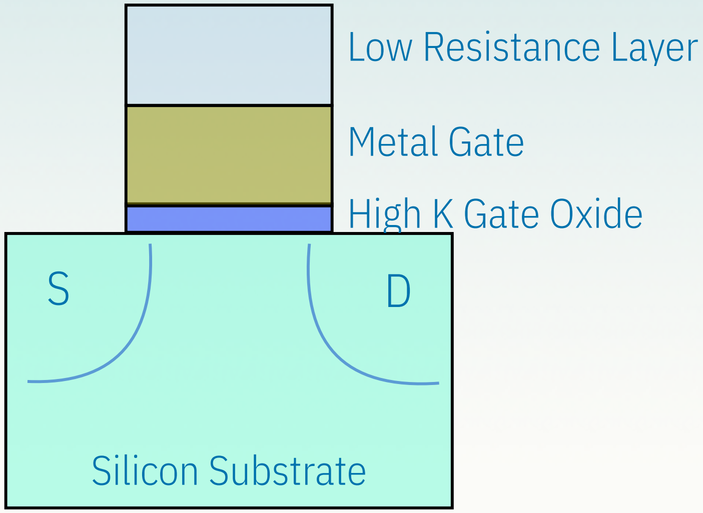
- CMOS power consumption increases with clock rate:

$$\text{Power} = \frac{1}{2}\text{CFV}^2$$

- Once processors reached about 75W (~4GHz) in the late 1990's, extraordinary measures were used to cool processors.
- So called 'over·clockers' used the disparity between the "safe" clock speed set by the manufacturer and what the CPU could actually do to get much faster clock rates. This led to "turbo" modes.

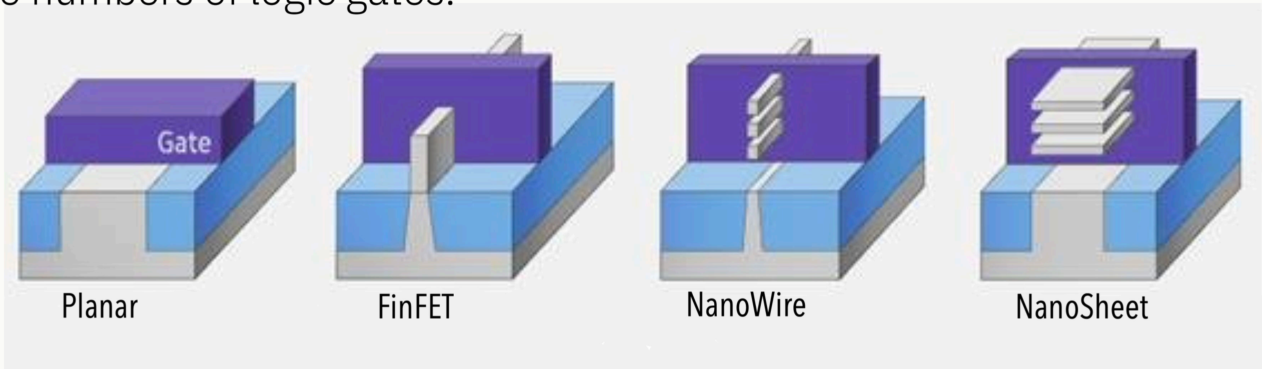
Thermal Management 2· Current Leakage

- With shrinking gate thickness needed for reduced gate length, quantum effects began to dominate. Specifically gate current leakage started to add a significant thermal load.
- This led to exotic gate materials such as hafnium oxide. This is the first break in Moore’s Law since the period of “traditional” scaling was broke with Intel’s 90nm process in 2000·2003.



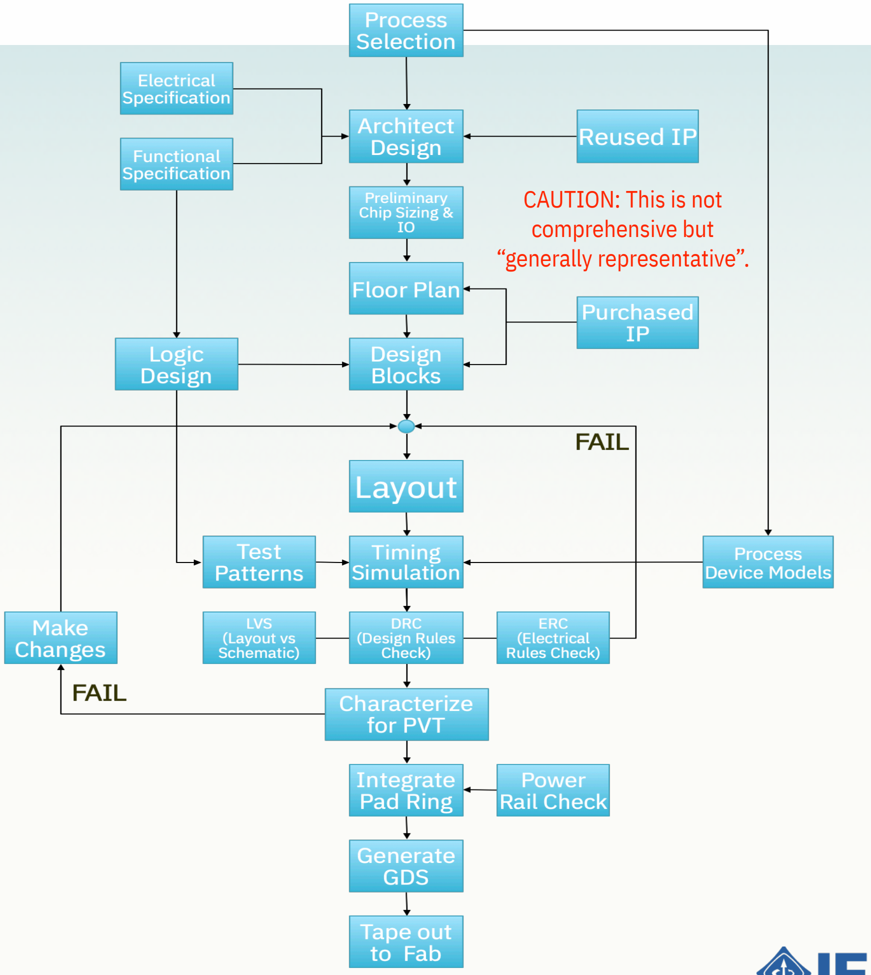
Device Models

- Ideally, physics based models are created CMOS logic gate FET switching time. As FET size shrinks, the switching time delay variation can be quite large.
  - By illustration, imagine designing a car where every year the engine got smaller, but the mpg varied per engine. In the initial years, the engine got 30-32 mpg, then 20-40, to 10-50. Planning a trip with such wide variation in engine mpg would become increasingly difficult.
- Statistical fluctuations get worse as gate size (no. of atoms, etc.) shrinks. The single most gating process (yes… bad pun) in digital IC development is "timing closure".
- The device model must also be tractable in the sense that it can be used to model the operation of (very) large numbers of logic gates.



Timing Closure is Becoming Very Difficult

- Timing closure normally done under PVT (process, voltage, temperature). This is very difficult to do with good process control. It is even more difficult if there is large variation in device parameters.
- Historically, the "device architect" threw the design over the wall to the layout team. This can't be done anymore.
- Some companies are using AI/ML to meet timing closure.



Process Selection

Electrical Specification

Functional Specification

Architect Design

Reused IP

Preliminary Chip Sizing & IO

Floor Plan

Purchased IP

Logic Design

Design Blocks

FAIL

Layout

Test Patterns

Timing Simulation

Process Device Models

Make Changes

LVS (Layout vs Schematic)

DRC (Design Rules Check)

ERC (Electrical Rules Check)

FAIL

Characterize for PVT

Integrate Pad Ring

Power Rail Check

Generate GDS

Tape out to Fab

CAUTION: This is not comprehensive but "generally representative".

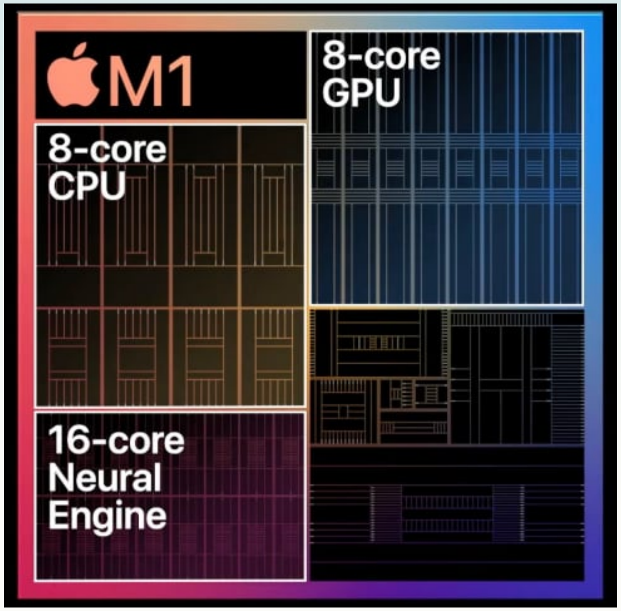
Near Future · Moore's Law Productivity Improvement Trends

Where Does the Industry Stand Now?

- Process fabs (e.g. TSMC) are shipping 5nm devices and claim they'll ship 3nm in 2023. This is "qualified" in that they do not claim to ship in "high volume" anytime soon.
- Device builders are shipping product with anywhere from eight CPU cores and eight GPU cores and higher.
- Gaming and bitcoin mining are the two drivers of high performance processors. Cell phones and laptops are beginning trade performance/efficiency for battery life.

Coprocessors Used Instead of 128 Bit Processors

- Coprocessors have become a viable technique to take advantage of the large number of transistors available.
- For example, the Apple M1 (5nm) processor has:
  - 8 CPUs, 4 for high performance & 4 for low power
  - 8 GPUs, 4 for high performance & 4 for low power
  - 16 core neural engine
  - Plus cache, DRAM, and communications functions
- Expect additional coprocessor "types" (e.g. SDR, AI/ML, media or security engine).



There is diminishing return in how many coprocessor cores one can implement due to intra·core communication, software, and physical size limitations.

Cloud Based Computing

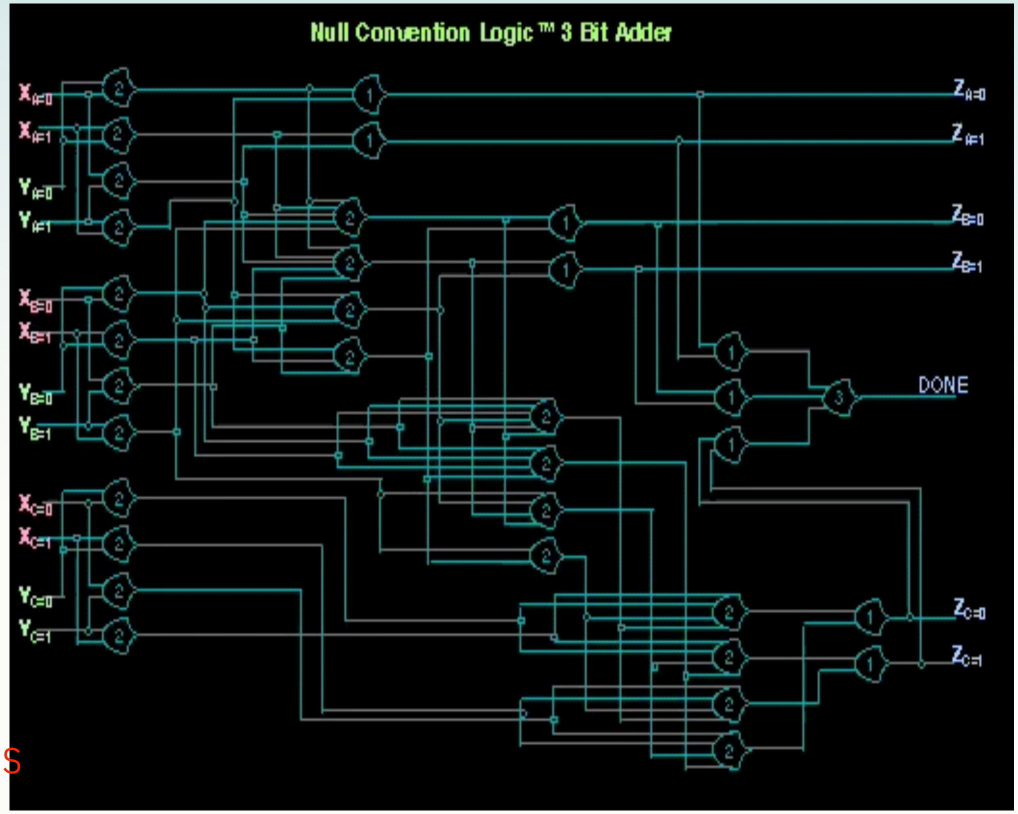
- Cloud based computing is attractive because it can be subscription based and the average user is unlikely to need the most powerful processors
- While the mobile/remote user can be relatively low CPU intensive (e.g. Chromebook's), the server farms can employ the latest multicore processors in a very controlled environment.
- It depends upon ubiquitous communications capability (e.g. 5G wireless) and fiber optics.

The weak link in cloud based computing is the communications link.

A Resurgence of Asynchronous Logic

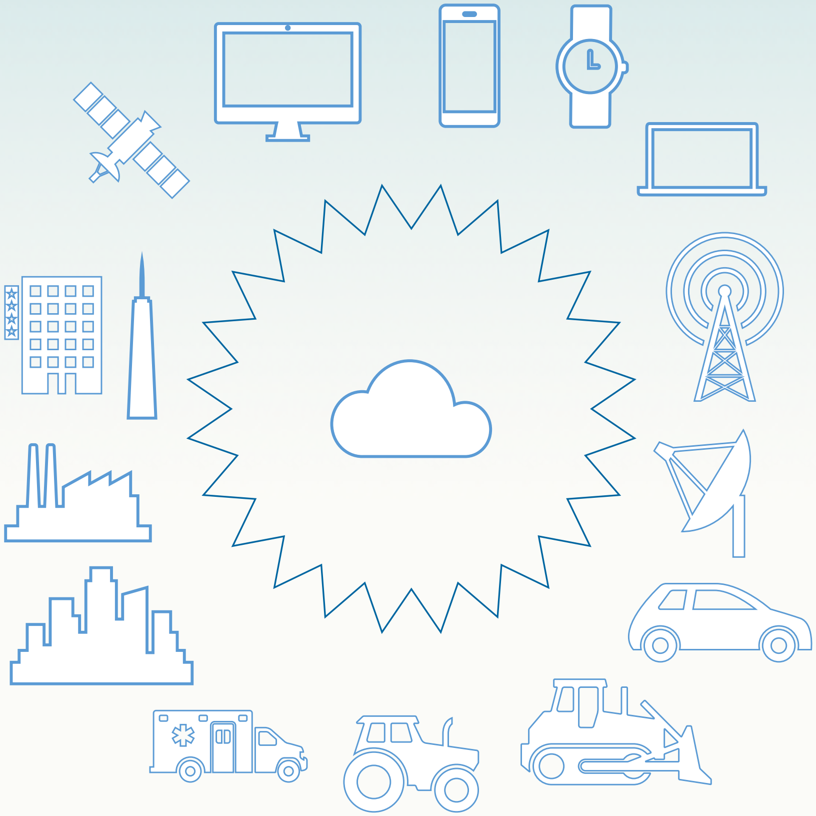
- Asynchronous or clockless logic does not require a clock tree or clock distribution. It does require more gates.
- Clockless logic uses a “hand shake” between registration levels to verify the calculation is complete. By design, there are no race conditions.
- The clock is so embedded in modern logic design flows, it is very very hard to get designers to consider clockless logic

There is reason to believe that some fruit named company is using a variant of clockless logic for their processors at 7nm and 5nm.



Internet of Things

- The goal here is to move processing to the edge of the internet. Think of this as an extreme version of Chrome books
- IoT also depends upon ubiquitous communications capability (e.g. 5/6G)
- The amount of data generated is massive. It will keep data scientists busy for decades.



Most IoT applications do not need high performance, they need very low power consumption.

Software Defined Everything

- Originally, there was the software defined radio (SDR) which took shape in the 1980's. [4] This was a radio that took advantage of digital signal processing to change radio functionality on the fly.
- The concept of developing generic digital hardware and "personalizing" it through software has taken hold and is now being considered for many different functionalities such as networking, storage, etc.

```
##################################################################################
#             Coplanar Waveguide Calculation
#       (c) copyright 2014-2022, Sailvolt LLC, All rights reserved,
#                 d2ferguson@sailvolt.com
#
##################################################################################
import math

π = math.pi
Er = 2.2              # dielectric constant
S = 5.0               # mils spacing between conductor & ground
H = 31.25             # mils substrate height
T = 0.1               # mils conductor thickness
W = 157.0             # mils conductor width

def ellipk(k):
   a0 = 1.38629436112;              b0 = 0.5
   a1 = 0.09666344259;              b1 = 0.12498593597
   a2 = 0.03590092383;              b2 = 0.06880248576
   a3 = 0.03742563713;              b3 = 0.03328355346
   a4 = 0.01451196212;              b4 = 0.00441787012
   z = 1-k*k
   K = (a0+a1*z+a2*(z**2)+a3*(z**3)+a4*(z**4))+(b0+b1*z+b2*(z**2)+b3*(z**3)+b4*(z**4))*math.log(1.0/z)
   return K

# ===============Coplanar WG w & w/o Gnd Plane =====================
k1 = W/(W+2.0*S)                                      # correction for split dielectric
k1P = math.sqrt(1.0-k1*k1)

k2 = math.sinh(π*W/(4.0*H))/math.sinh(π*(W + 2.0*S)/(4.0*H))      # correction for finite dielectric height
k2P = math.sqrt(1.0-k2*k2)

k3 = math.tanh(π*W/(4.0*H))/math.tanh(π*(W + 2.0*S)/(4.0*H))      # correction for finite dielectric height & gnd
k3P = math.sqrt(1.0-k3*k3)

Del = (1.25*T/π)*(1.0+math.log(4.0*π*W/T))   # correction for metalization thickness
Se = S-Del                                   # correction for metalization thickness
We = W+Del                                   # correction for metalization thickness
k1e = We/(We+2.0*Se)                         # k1 correction for metalization thickness
k1eP = math.sqrt(1.0 - k1e*k1e)

Ereff = 1.0 + ((Er - 1.0)/2.0)*(ellipk(k2)/ellipk(k2P))*(ellipk(k1P)/ellipk(k1))      # correction finite dielectric
q = (ellipk(k3)/ellipk(k3P))/((ellipk(k1)/ellipk(k1P))+(ellipk(k3)/ellipk(k3P)))      # filling factor for gnd plane
Ere = 1.0+q*(Er - 1.0)                                  # correction for finite dielectric
Eret = Ere - (0.7*(Ere - 1.0)*T/S)/(ellipk(k1)/ellipk(k1P) + 0.7*T/S)

Z0_NoGe = (30.0*π/math.sqrt(Ereff))*(ellipk(k1eP)/ellipk(k1e))                        # no gnd tested
#Z0_NoG1 = (30.0*π/math.sqrt(Eret))*(ellipk(k1P)/ellipk(k1))                          # no gnd tested
#Z0_G = (60.0*π/math.sqrt(Eret))*1.0/((ellipk(k1)/ellipk(k1P))+(ellipk(k3)/ellipk(k3P)))    # gnd
Z0_Ge = (60.0*π/math.sqrt(Eret))*1.0/((ellipk(k1e)/ellipk(k1eP))+(ellipk(k3)/ellipk(k3P)))  # gnd & thickness
```

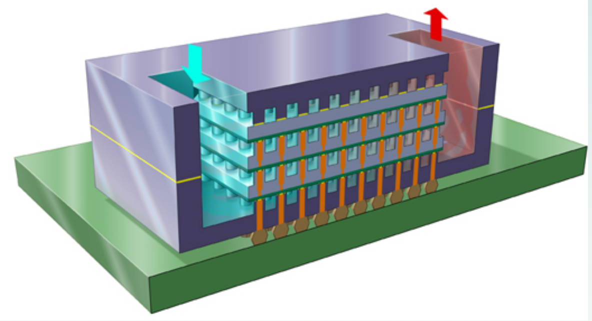
Device As A Service · Subscription Based Hardware

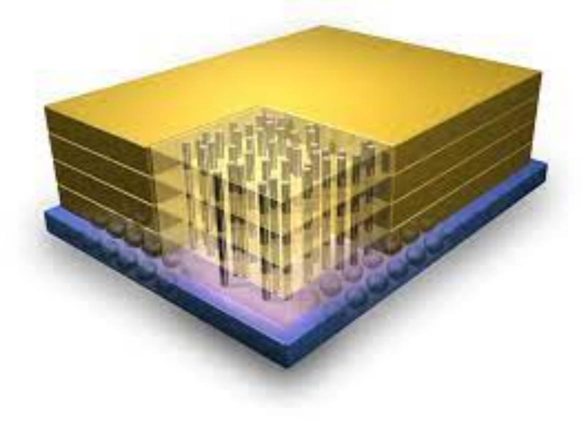
- There is a growing movement to make hardware subscription serviced.
- Intel has announced its plan to develop subscription based processors. For additional money, you can activate additional coprocessors or other functionality in your computer.
- IBM used this technique in the 1960’s.

Far Future · Moore's Law Productivity Improvement Trends

3D Processors

- Make devices physically larger:
  - Integrate several layers of devices. There is initial progress being made with memory.
  - New packaging technology needed.



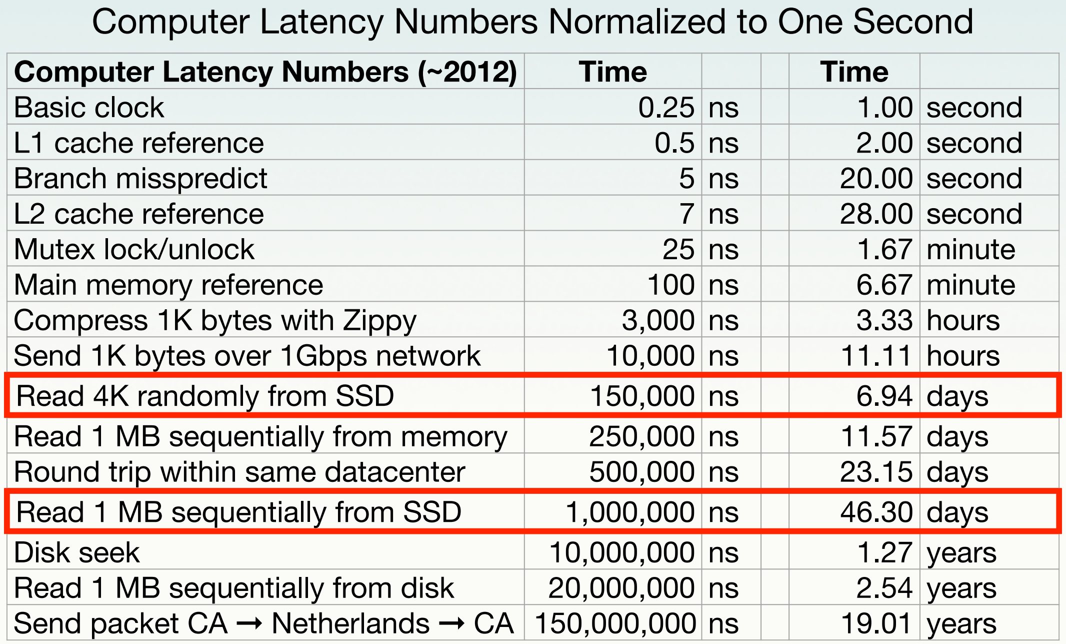


While this increases the number of components, it does not address fundamental thermal and signal wiring limitations.

Integrated Non-Volatile Memory

- Improve CPU performance by making all memory cache.
- Have a single N Terabyte non-volatile memory for both cache and long term storage. Differentiation is logical, not electrical.

This could dramatically improve computer performance.



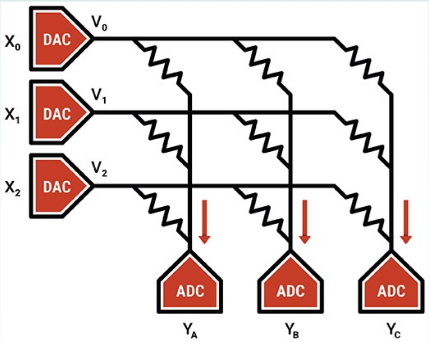
Computer Latency Numbers Normalized to One Second

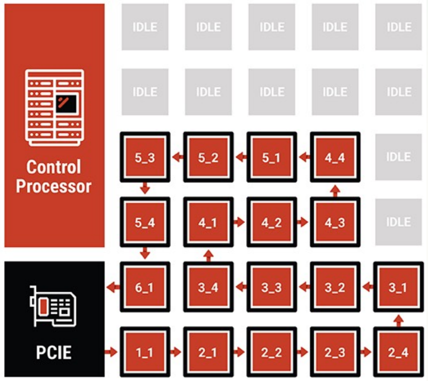
Computer Latency Numbers (~2012)	Time			Time	
Basic clock	0.25	ns		1.00	second
L1 cache reference	0.5	ns		2.00	second
Branch misspredict	5	ns		20.00	second
L2 cache reference	7	ns		28.00	second
Mutex lock/unlock	25	ns		1.67	minute
Main memory reference	100	ns		6.67	minute
Compress 1K bytes with Zippy	3,000	ns		3.33	hours
Send 1K bytes over 1Gbps network	10,000	ns		11.11	hours
Read 4K randomly from SSD	150,000	ns		6.94	days
Read 1 MB sequentially from memory	250,000	ns		11.57	days
Round trip within same datacenter	500,000	ns		23.15	days
Read 1 MB sequentially from SSD	1,000,000	ns		46.30	days
Disk seek	10,000,000	ns		1.27	years
Read 1 MB sequentially from disk	20,000,000	ns		2.54	years
Send packet CA → Netherlands → CA	150,000,000	ns		19.01	years

NonConventional Computing · Analog

- Analog computing traces its roots to the era before digital computing. Opamp based analog computers were still available at UT in the mid 70’s.
- Austin startup Mythic AI is using non·volatile memory to build analog computing for AI/ML. See Veritasium video in references for an excellent explanation.
- Analog computing will be special purpose and likely to emerge as coprocessors.

Analog computing will dominate a certain class of problems due to its low power and high speed.





NonConventional Computing · DNA

- DNA Computing ➞ massively parallel computer which holds potential of integration with living organisms ➞ babble fish a reality?
  - uses four nucleic acids instead of binary (0,1 ➞ A,T,C,G)
  - Extremely high data density ➞ >1 petabits/in^2 in 2D
- DNA as computer memory has been demonstrated.

It is both interesting and possibly disturbing to contemplate where DNA computing could lead.

NonConventional Computing · Other

- Optical Computing ➞ still in very early stages, potential unknown
- Quantum Computing ➞ useful for specialty problems, particularly those amenable to parallel computing
- Molecular Computing?

All of the non·conventional computing approaches except DNA computing are not likely to be general purpose.

Past Economic Models

- One example is steel. There was a period of great innovation and then it became embedded in everything. Transportation, buildings, energy, food production. Everything was impacted.
- The steel industry is somewhat static and electronics is likely to follow a comparable path.
- That said, with electronics, there is the potential for continued innovation that doesn't exist with steel.

Inexpensive, high quality & specialty steels are fundamental to an industrial society. ICs have become comparable.

What Does the Far Future Look Like?

- Computers and embedded processors will perform marginally better than they do today. The human interface will be more sophisticated.
- Coprocessors will be available to accelerate solutions to certain types of problems
- The chip industry will be stable with fewer players. Every economic block will have their "domestic" suppliers.
- There will continue to be innovation in circuit design, processor architecture, communication and control interfaces, etc.

Summary | Conclusions

- For Moore’s Law to continue in a productivity sense, either transistors must continue to shrink, or chips must get larger, or both. Continuing to shrink transistors will ultimately fail. It seems the path forward is bigger chips.
  - Apple’s recent M1 Ultra reveal does just that by connecting two M1 chips through a 10,000 high speed line “interposer”. The M1 Ultra uses 114 billion transistors.
- There could be some final barrier lurking in the tall grass.
- Innovations in architecture and circuits are needed to take advantage of all the transistors available.
- The focus will change from huge investments in building transistors to huge investments in making everything “smart”.

References | Resources

References:

(1) Gordon Moore, “50 Years of Moore’s Law”, Intel.com video, April, 2015 http://www.intel.com/content/www/us/en/silicon-innovations/50-years-of-moores- law-video.html.

(2) “So Much Moore”, IEEE Spectrum, January 2022, p23

(3) “Moore’s Law: past, present, and future”, IEEE Spectrum, June 1997, p5

(4) I’m an eyewitness. We originally called it “synthetic radio”.

(5) “The Summer Intel Fell Behind”, The Verge, 29 July 2021

Resources:

- Kate’s Ag: https://www.youtube.com/watch?v=rKHDWTRyU54
- Millennial Farmer: https://www.youtube.com/channel/UCp0rRUsMDlJ1meYAQ6_37Dw
- Veritasium: https://www.youtube.com/watch?v=GVsUOuSjvcg