The Decline of GPT's and Feasibility of Specialization

CSCE 4013/5012 Domain Specific Architectures
Professor David Andrews

Lecture materials drawn from the following two papers:

N. Thompson, S, Spanuth "*The Decline of Computers as a General Purpose Technology*" Communications of the ACM, March 2021

A. Fuchs, D. Wentzlaff, "*The Accelerator Wall: Limits of Chip Specialization*", HPCA 2019

The Decline of GPT's and Feasibility of Specialization

Recap:

1. Moore's Law is slowing and Dennard Scaling ended.
2. Manycores: Moore's Law applied to increasing GP processor cores.
3. Heterogeneous Manycores: Mix of cores to increase energy efficiency and exploit types of parallelism within an application.
3. Domain Specific Architectures: A new era of specialization.

Machine Learning Thundering Hurd

N. Thompson, S, Spanuth "*The Decline of Computers as a General Purpose Technology*" Communications of the ACM, March 2021

Key Insights

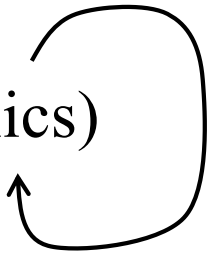
1. Moore's Law was driven by technical achievements and a "general purpose technology" (GPT) economic cycle where market growth and investments in technical progress reinforced each other. These created strong economic incentives for users to standardize to fast-improving CPUs, rather than designing their own specialized processors.

2. Today, the GPT cycle is unwinding, resulting in less market growth and slower technical progress.

3. As CPU improvement slows, economic incentives will push users toward specialized processors, which threatens to fragment computing. In such a computing landscape, some users will be in the 'fast lane,' benefiting from customized hardware, and others will be left in the 'slow lane,' stuck on CPUs whose progress fades

The end of GPT's

General Purpose Technologies (GPTs): Products with broad applicability. Programmable CPU's.
Lifecycle of GPTs



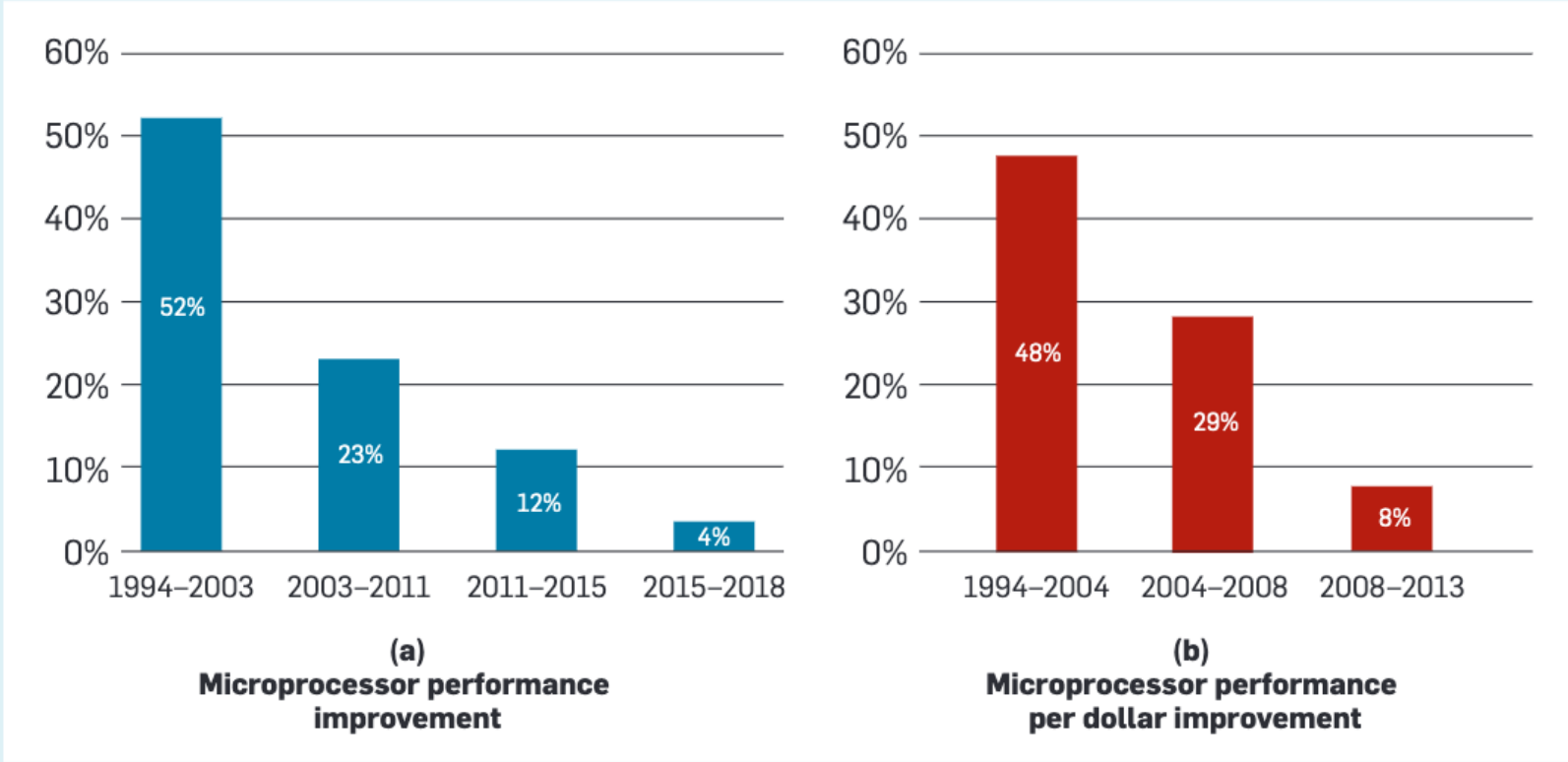
-Use case will grow (Expansion and growing Economics)

-Eventually progress slows
-Will be displaced in some niches
-Displacement will undermine economic model. When this happens applications will move to *specialized processors*.

Has this already happened ?

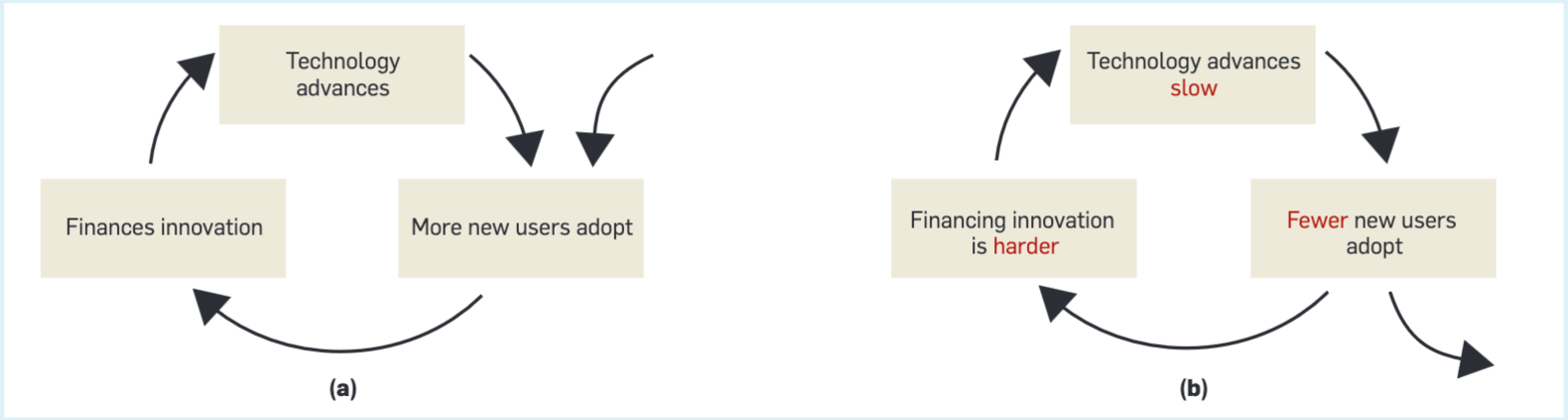
Rate of Improvements slowed

Figure 2. Rate of improvement in microprocessors, as measured by (a) Annual performance improvement on the SPECint benchmark,[7appx] and (b) Annual quality-adjusted price decline.[1appx]



From Virtuous to Fragmentation

Figure 1. The historical virtuous cycle of universal processers (a) is turning into a fragmentation cycle (b).



- PC replacement 4 -> 5-6 years
- Smartphones 23 months -> 31 months
- Some may skip generational replacements

"That's mathematics son! You can argue with me but you cant argue with figures"

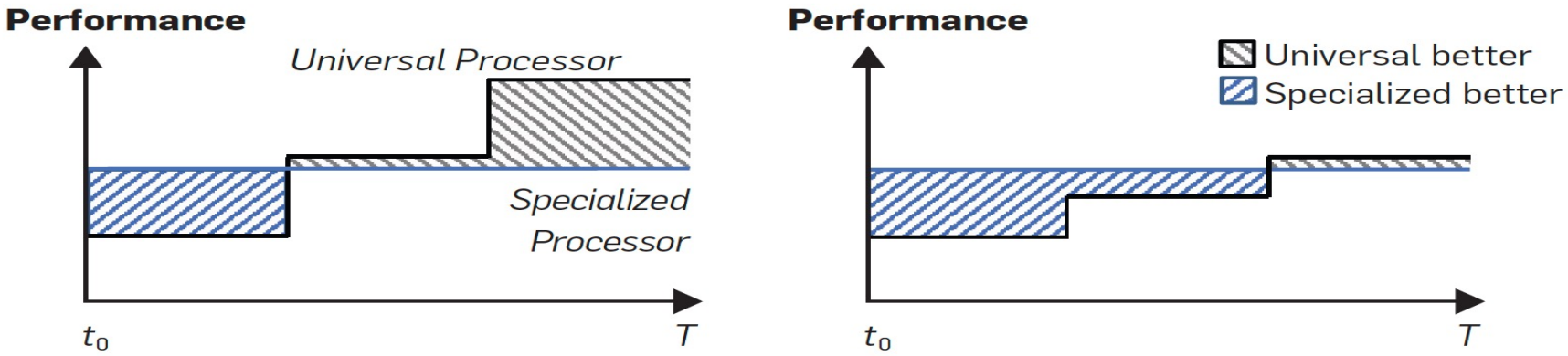


Illustration of processor choice trade-off when universal processor improvement rate is (a) fast or (b) slow

Peak Moore's Law

48% Improvement/year

$Sp = \frac{P_s}{P_u} = 100x \Rightarrow 83{,}000$

$Sp = \frac{P_s}{P_u} = 2x \Rightarrow 1{,}000{,}000$

2008-2013

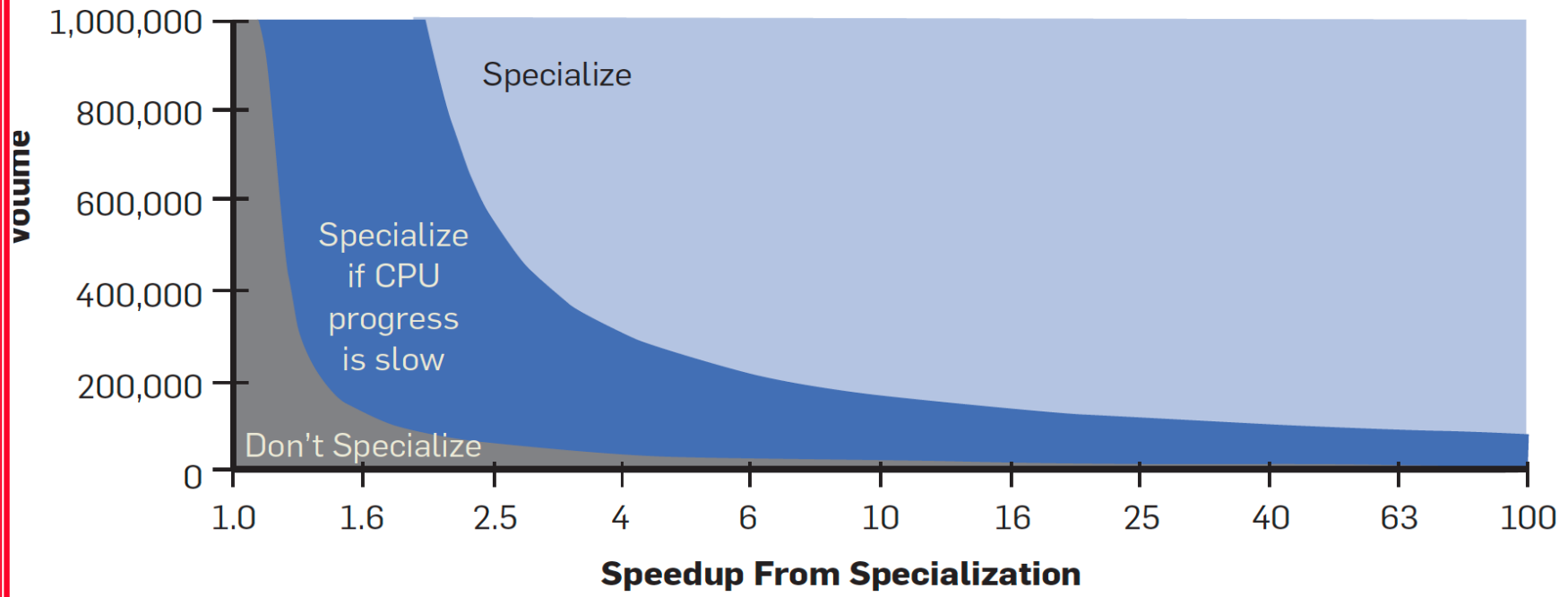
8% Improvement/year

$Sp = \frac{P_s}{P_u} = 100x \Rightarrow 15{,}000$

$Sp = \frac{P_s}{P_u} = 2x \Rightarrow 81{,}000$

N. Thompson, S. Spanuth, Decline of Computers as a General Purpose Technology, Comm of the ACM March 2021

"That's mathematics son! You can argue with me but you cant argue with figures"



Peak Moore's Law

48% Improvement/year

$Sp = \frac{P_s}{P_u} = 100 \Rightarrow 83{,}000$

$Sp = \frac{P_s}{P_u} = 2x \Rightarrow 1{,}000{,}000$

2008-2013

8% Improvement/year

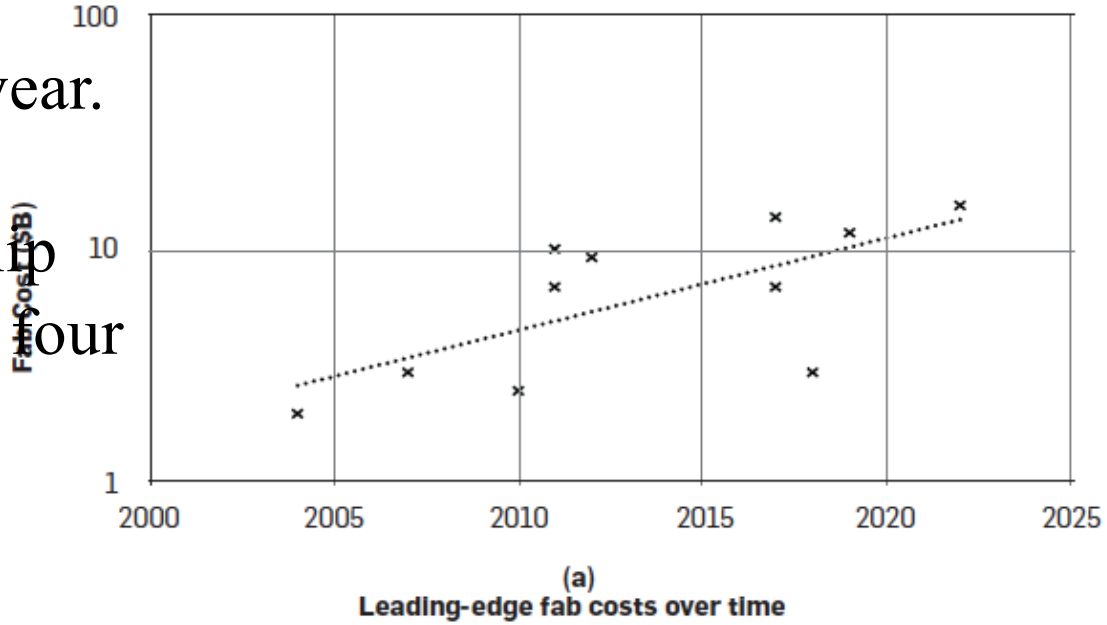
$Sp = \frac{P_s}{P_u} = 100 \Rightarrow 15{,}000$

$Sp = \frac{P_s}{P_u} = 2x \Rightarrow 81{,}000$

N. Thompson, S. Spanuth, Decline of Computers as a General Purpose Technology, Comm of the ACM March 2021

Its Economics

- Fab Costs 13%/year.
- Moore's "Second Law" Cost of a chip fab doubles every four years

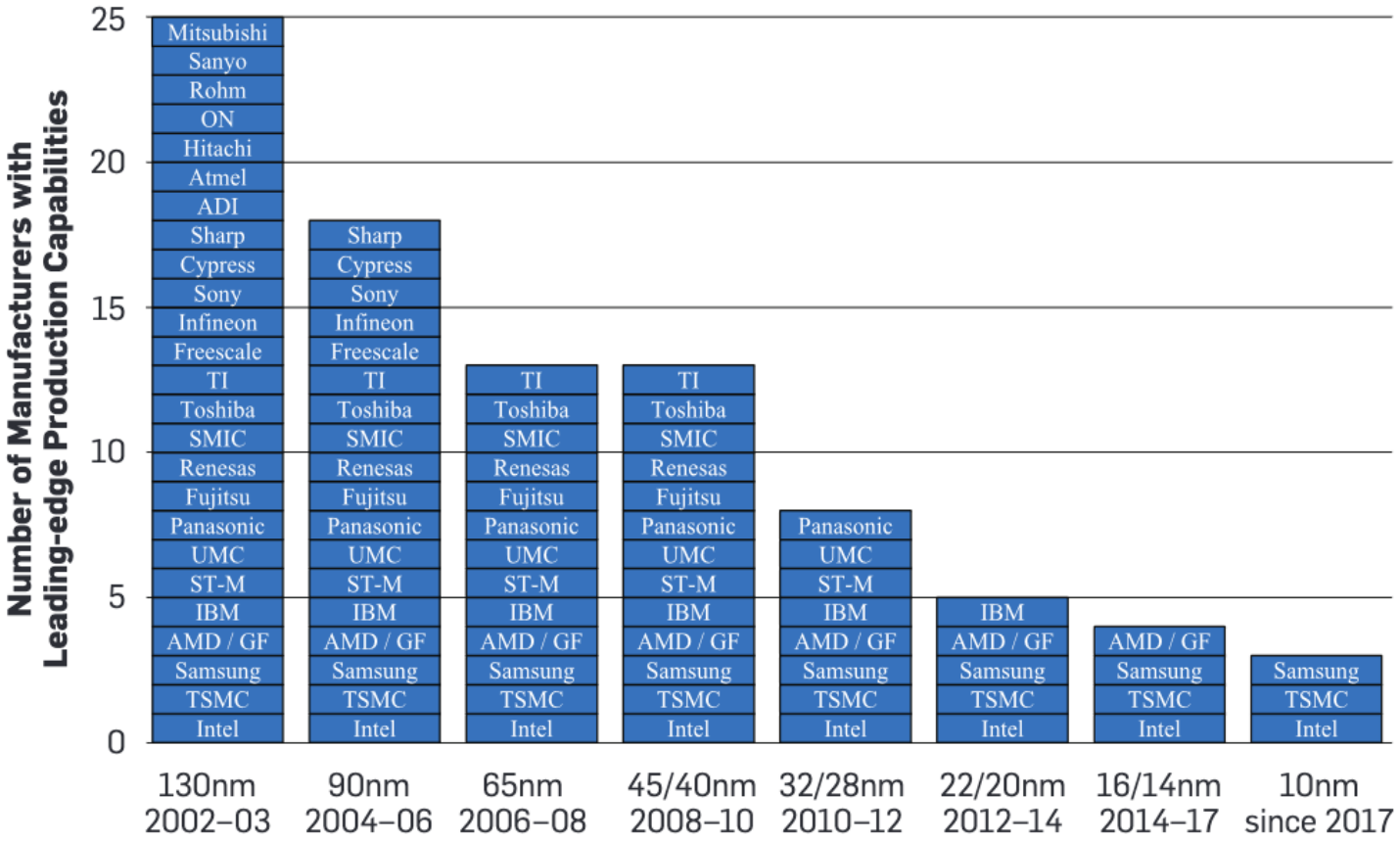


(a)
Leading-edge fab costs over time

- CAGR 5% | 1996-2016
- Less competitive players left the Market and remaining players amortized fixed cost over larger numbers of chips

N. Thompson, S. Spanuth, Decline of Computers as a General Purpose Technology, Comm of the ACM March 2021

The shrinking number of fabs

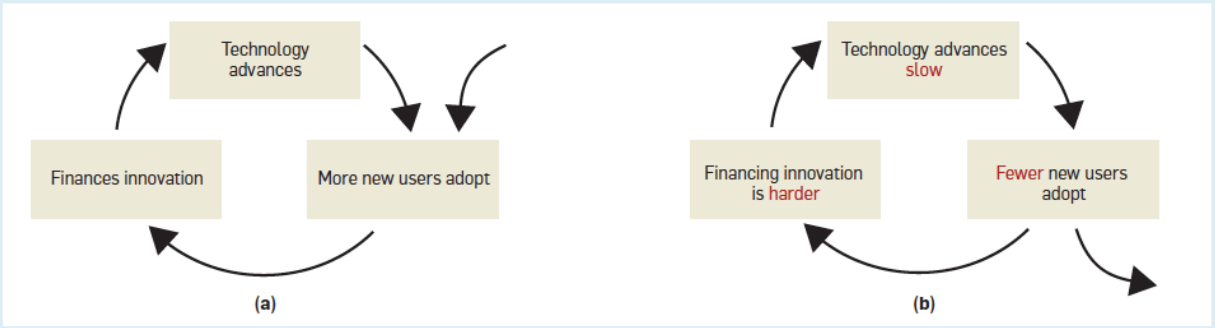


(b)

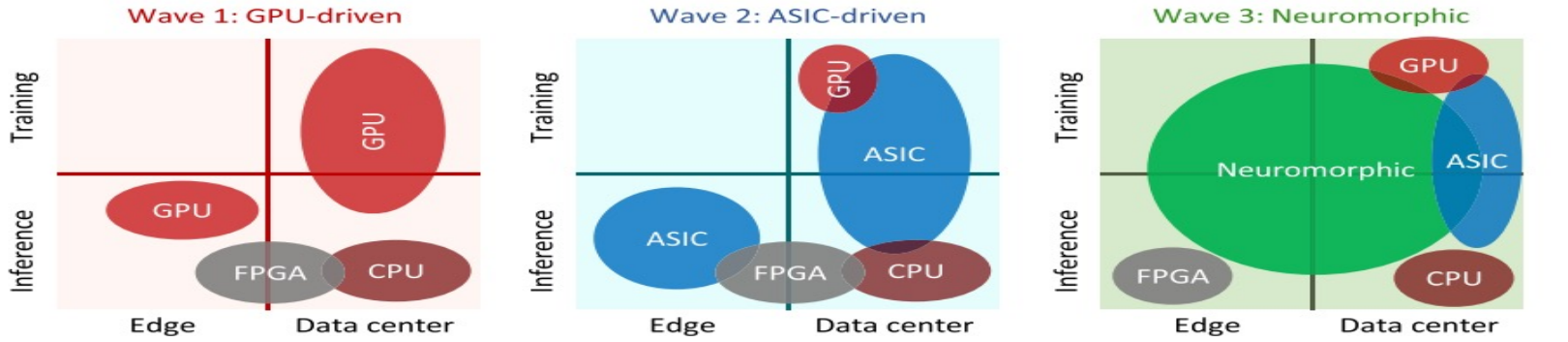
Number of manufacturers with leading-edge production capabilities by node size and year (based on data from Bloom et al.[2appx] and Hemsoth[12appx])

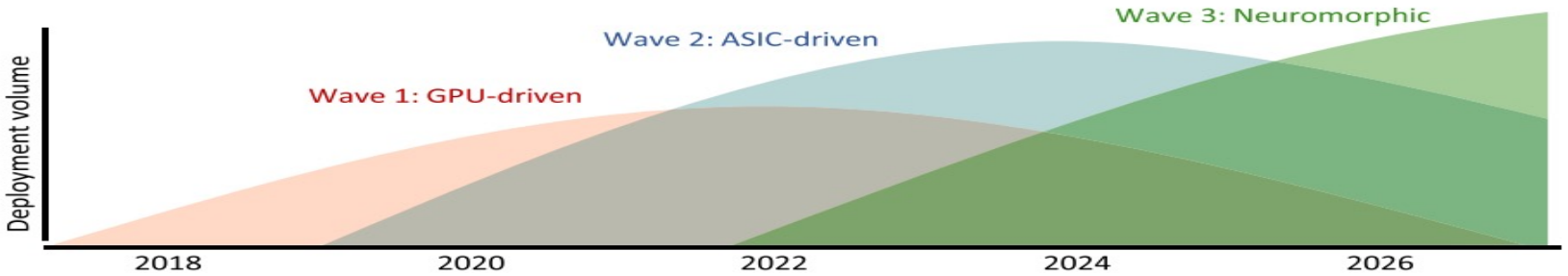
Dennard Scaling Ends and Moore's Law Slows Specialization Gains Market Share

Figure 1. The historical virtuous cycle of universal processers (a) is turning into a fragmentation cycle (b).



N. Thompson, S. Spanuth, Decline of Computers as a General Purpose Technology, Comm of the ACM March 2021





J. Kendall, S. Kumar, "The building Blocks of a Brain-Inspired-Computer" Applied Physics Reviews 2020

So no big deal right ? Specialization can fill the gap……..

Gain = CMOS-Driven Gains x Chip Specialization Gains

DSA's are Silicon thus Dependent on Transistor Scaling

CSR: Chip Specialization Return:

Take transistor scaling out

Evaluate gain only due to specialization

When Moore's Law ends CSR is all we are left with !

A. Fuchs, D. Wentzlaff "The Accelerator Wall: Limits of Chip Specialization" HPCA 2019

Nex Time: The Accelerator Wall

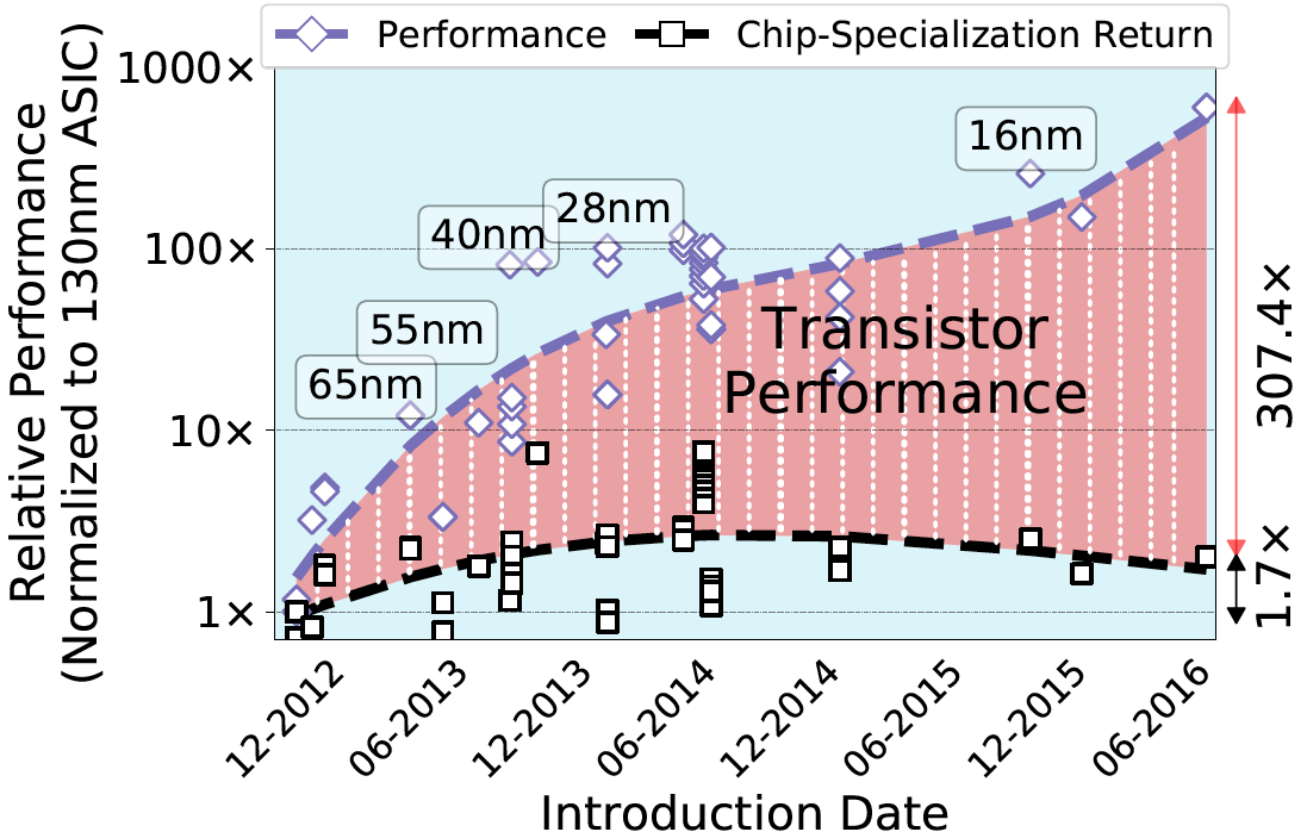


Figure 1: Evolution of Bitcoin Mining ASIC Chips. Performance metric is SHA256 Hashing Throughput per Chip Area (Hashes/Seconds/mm^2).

A. Fuchs, D. Wentzlaff "The Accelerator Wall: Limits of Chip Specialization" HPCA 2019

A.Fuchs, D. Wentzlaff "The Accelerator Wall: Limits of B.Chip Specialization" HPCA 2019

Key Insights:

1. Specialization gives nice initial boost but further increases from additional specialization tails off. 1st generation new Parallel architecture matching parallelism of new domain can Give impressive speedup. Once exploited performance boosts Of each iteration of "new hardware tricks" diminishes. Can also get increasingly expensive.

2. Majority of observed generational improvements are derived From device scaling (aka Moore's Law). Reinforces the importance of finding some new form of device scaling.

Cost and time for hardware customization may become critical

The Specialization Stack

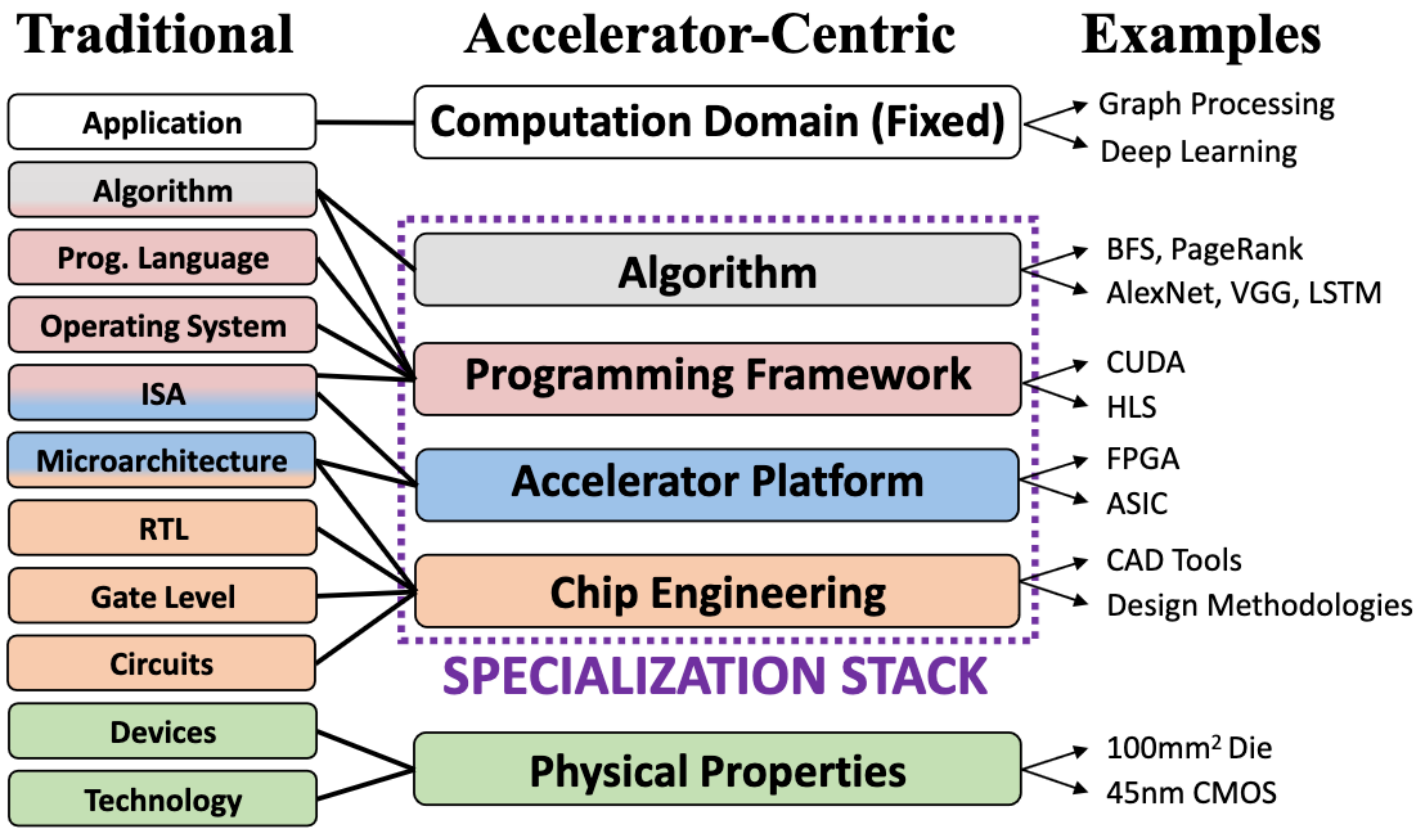
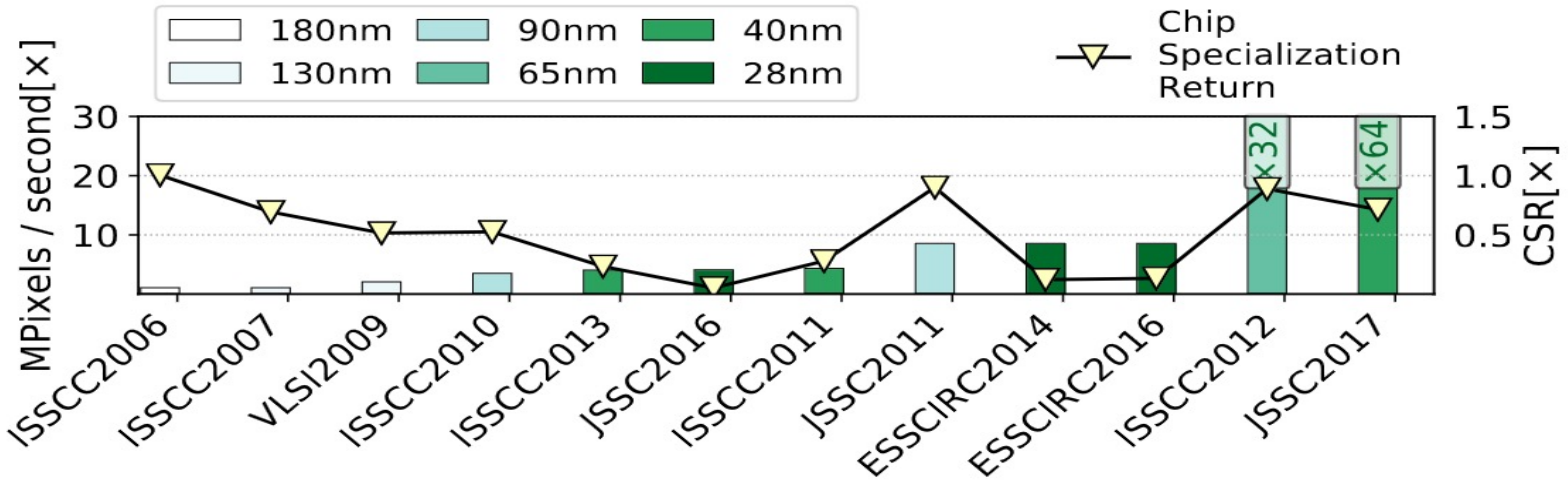
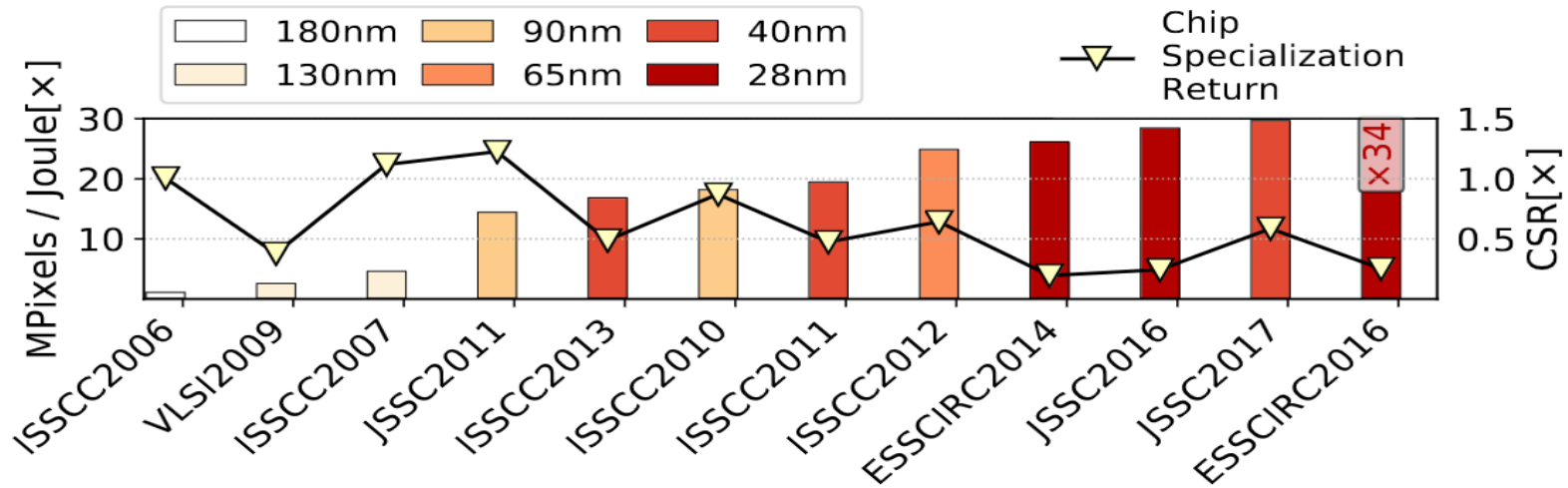


Figure 2: Abstraction Layers: Traditional and Accelerated Systems. Dashed Box Groups The Layers of Specialization.

Throughput due to scaling increases with technology node
Throughput due to specialization < 1.0



(a) Scaling of Performance and Chip Specialization Return



(c) Scaling of Energy Efficiency and Chip Specialization Return

The Law of Diminishing Returns

Computational Confinement: Domains with fixed hardware implementation have limited future specialization returns

Massive Parallelism: Domains like GPU graphics processing enabled via higher transistor counts: Relies on Moore's Law and limited by dark silicon

Domain Maturity: Only so many ways to skin a cat !

An Impending Accelerator Wall !

A. Fuchs, D. Wentzlaff "The Accelerator Wall: Limits of Chip Specialization" HPCA 2019

An Impending Accelerator Wall !

We will hit the Accelerator Wall after Moore's law finally dies, and transistor budgets plateau.

Chip Specialization is not a long-term remedy for the ending of Moore's Law

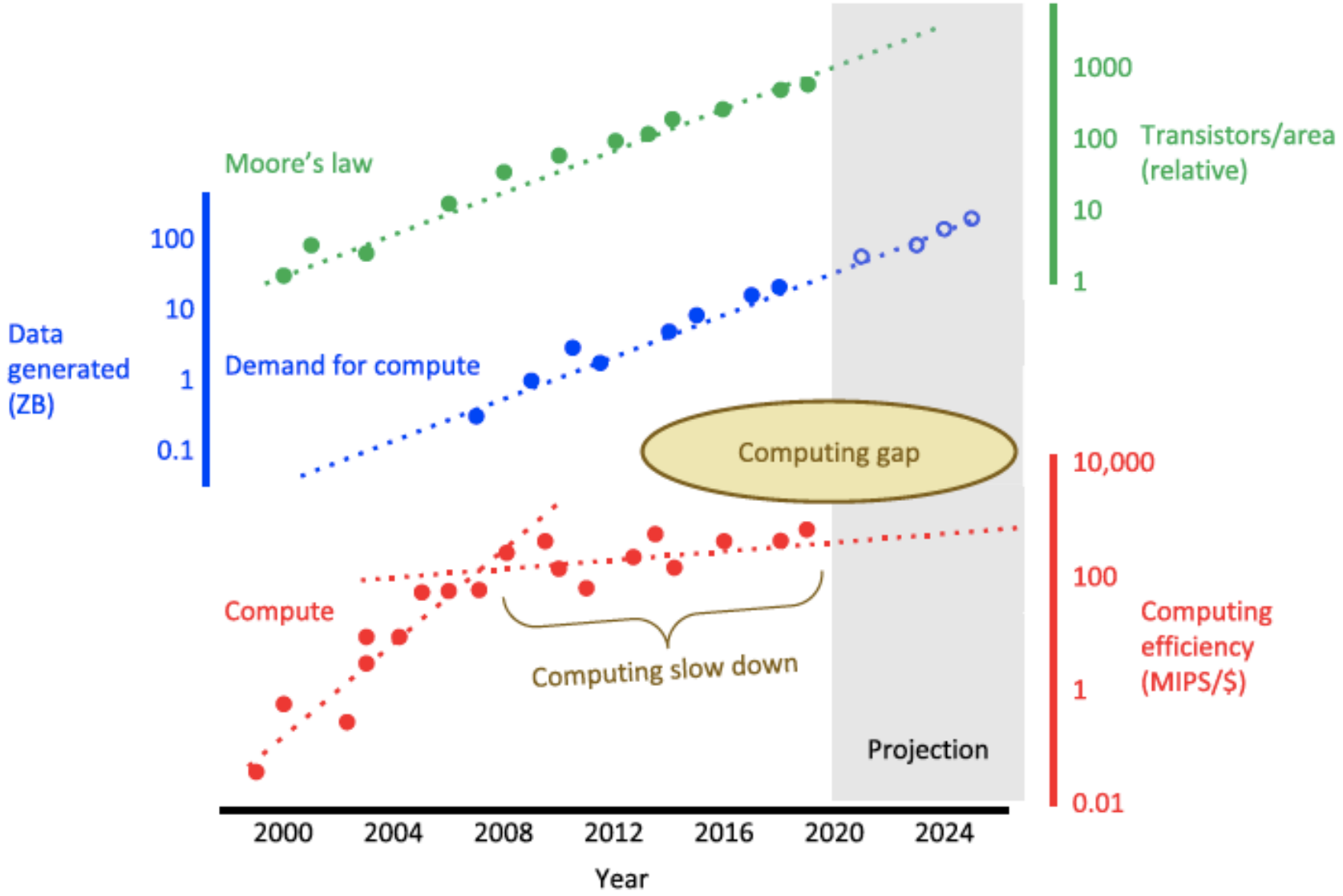
Yes, this applies to CPUs, GPUs, DSA's and FPGAs !

Can we re-establish Virtuous Cycle in a Post Moore's Law era ?

- A "New" General Purpose Technology
- Exponential Scaling of New Technologies

Save discussions are for a different class. ☺

Whats the big deal ?



J. Kendall, S. Kumar, The Building Blocks of a Brain-Inspired Computer, Applied Physics Reviews, Sept. 2019