Why Build Hardware Accelerators? (& Course Introduction)

Scott Beamer
sbeamer@ucsc.edu

1/7/19

University of California, Santa Cruz

CMPE 293: Programmable Hardware Accelerators

https://cmpe293-winter19-01.courses.soe.ucsc.edu

Efficiency is Critical at All Scales

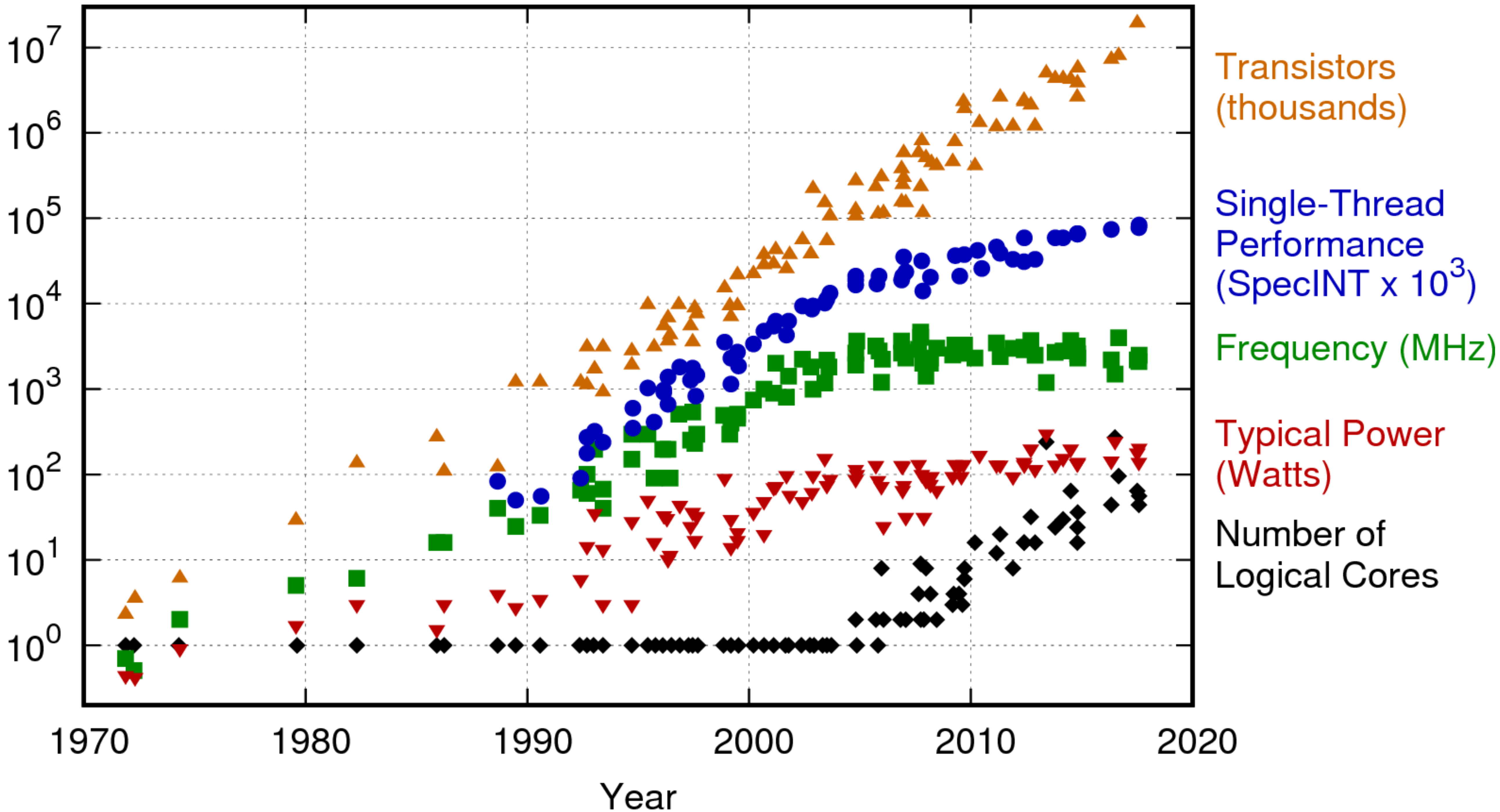
Mobile

Datacenter

Internet of Things

High-performance Embedded

Power Wall Limits Scaling



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 2010-2017 by K. Rupp

https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/

Demise of Dennard Scaling

- Way to scale transistor parameters (inc. voltage) to keep power density constant

Table 1
Scaling Results for Circuit Performance

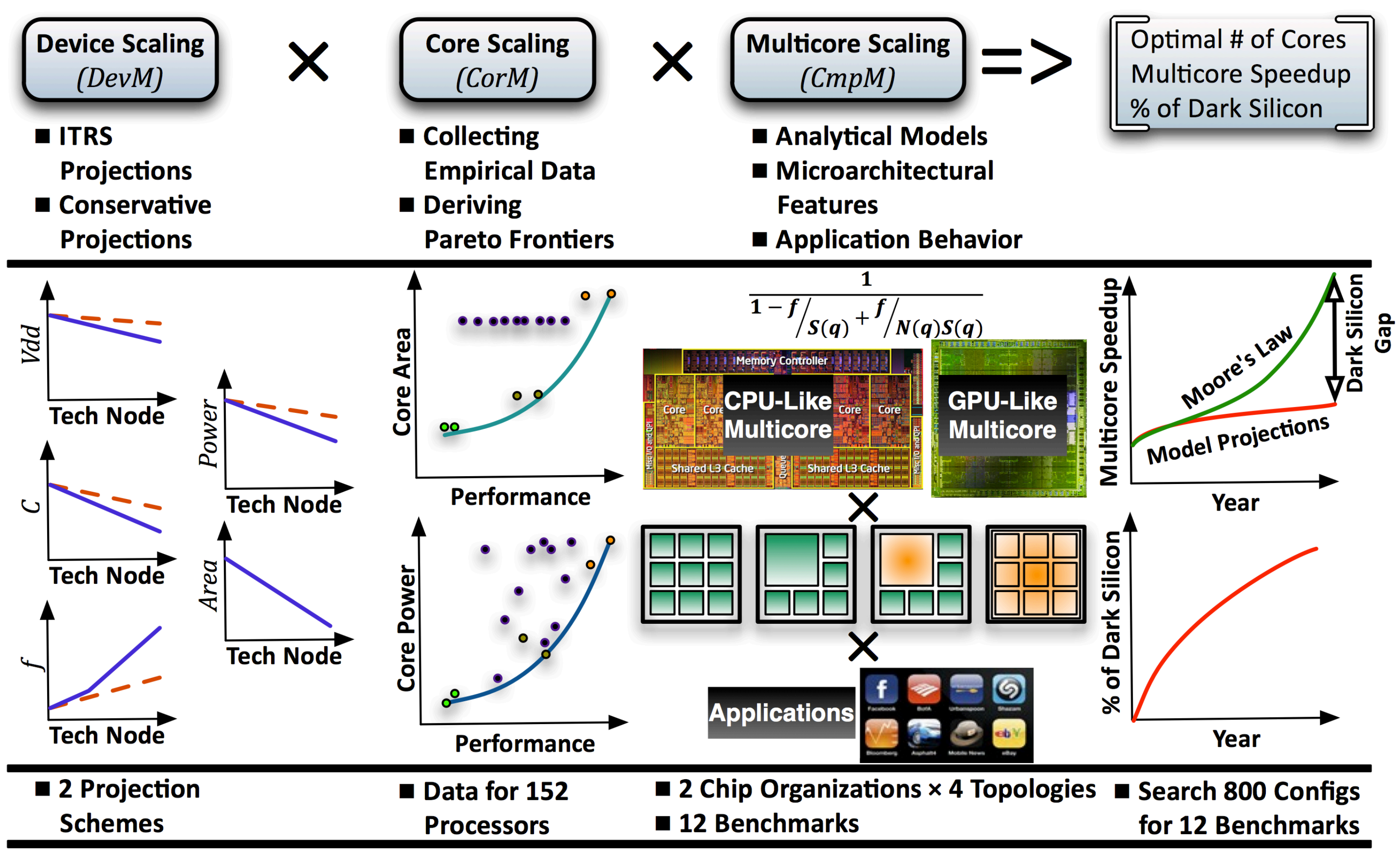
Device or Circuit Parameter	Scaling Factor
Device dimension t_{ox}, L, W	$1/\kappa$
Doping concentration N_a	κ
Voltage V	$1/\kappa$
Current I	$1/\kappa$
Capacitance $\epsilon A/t$	$1/\kappa$
Delay time/circuit VC/I	$1/\kappa$
Power dissipation/circuit VI	$1/\kappa^2$
Power density VI/A	1

"Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions" (JSSC 1974)

- Unfortunately, couldn't sustain so overall power increased until it hit cooling limits

Dark Silicon Argument

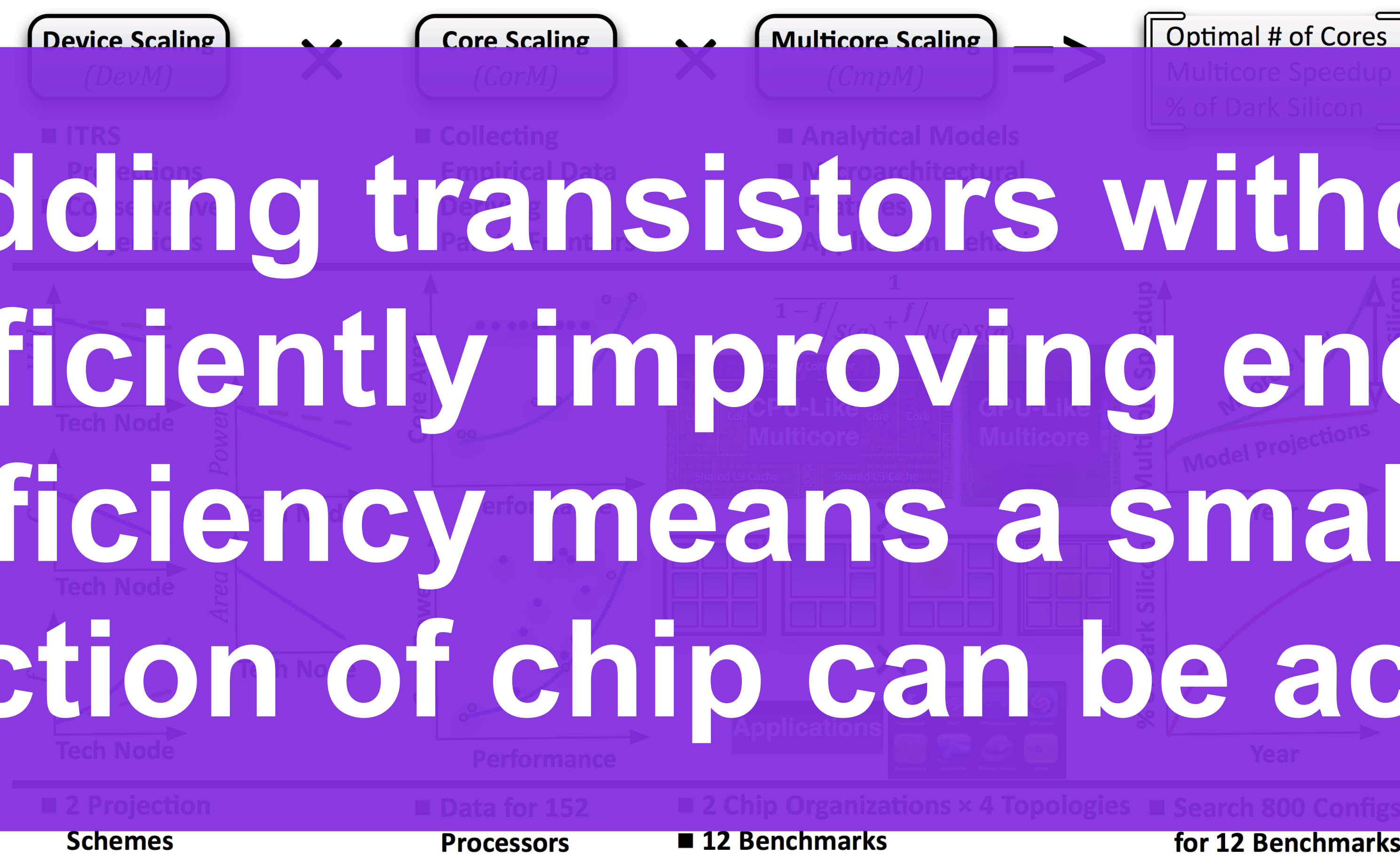
- If transistor energy efficiency is improving slower than the number of transistors implies for constant power, not all additional transistors may be active



"Dark Silicon and the End of Multicore Scaling" (ISCA 2011)

Dark Silicon Argument

- If transistor energy efficiency is improving slower than the number of transistors implies for constant power, not all additional transistors may be active

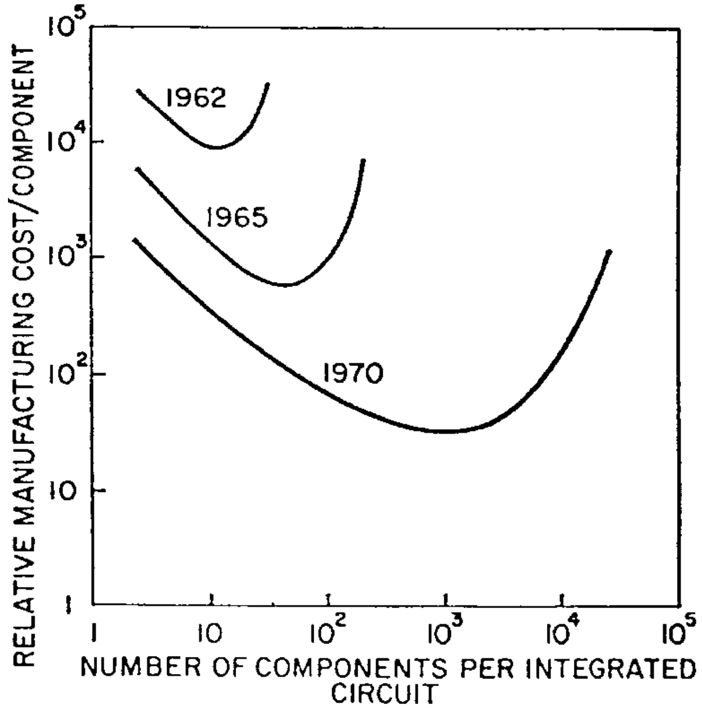


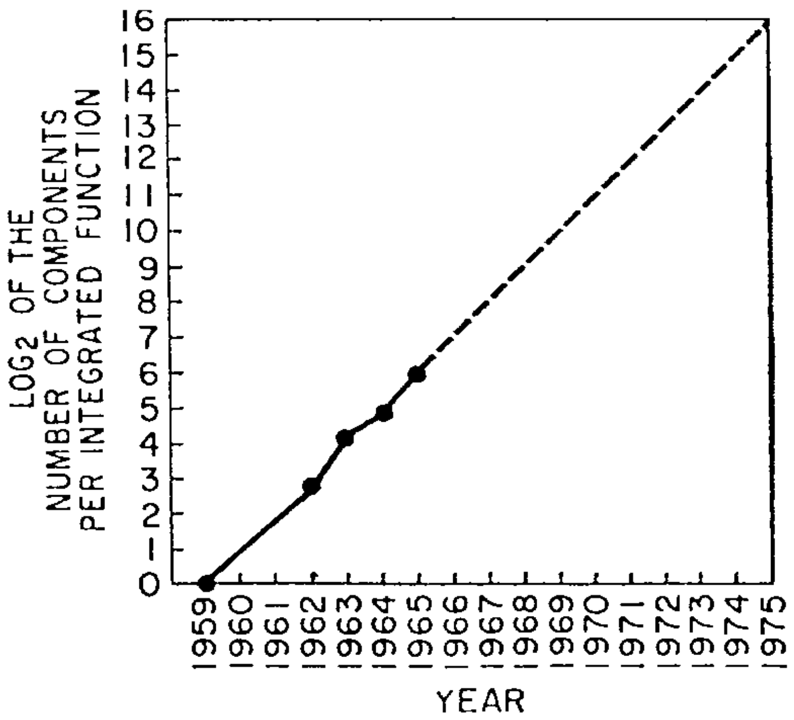
Adding transistors without sufficiently improving energy efficiency means a smaller fraction of chip can be active

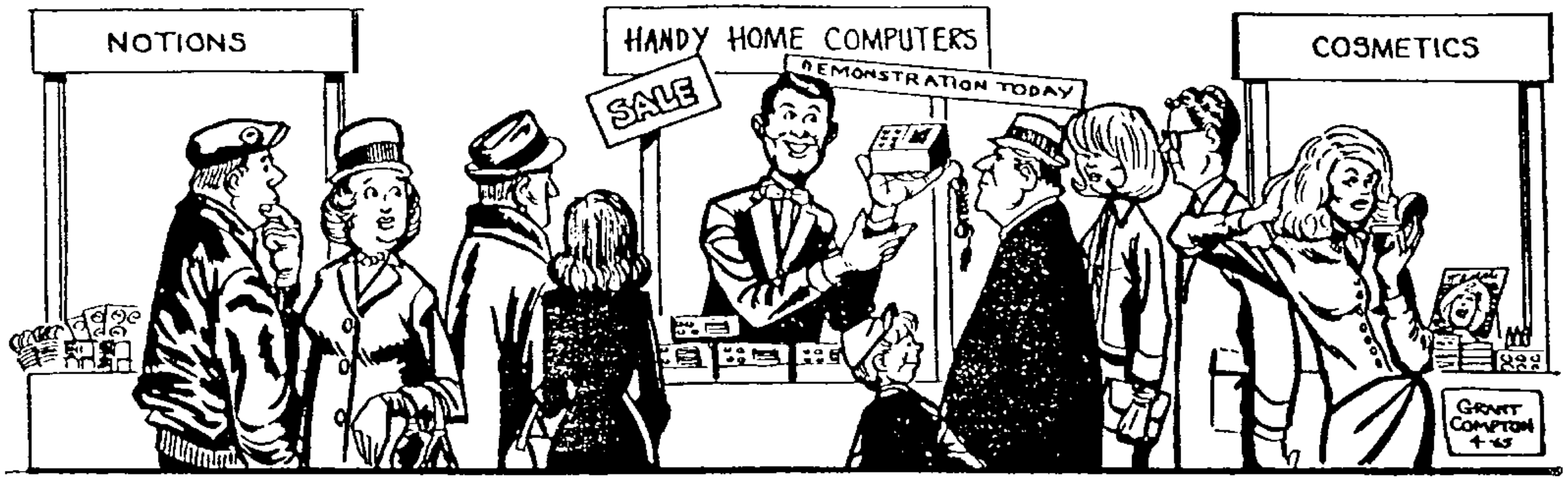
“Dark Silicon and the End of Multicore Scaling” (ISCA 2011)

Moore's Law

- Observation that # of transistors in an economical IC doubles every year







"Cramming More Components onto Integrated Circuits " (Electronics 1965)

Moore's Law is Slowing for Intel

- Even by own timeline Intel is behind schedule
- Competitors (TSMC & Samsung) are pushing ahead, but maybe only a few more steps left
- Time gap between new processes may grow

	Intel First Production
1999	180 nm
2001	130 nm
2003	90 nm
2005	65 nm
2007	45 nm
2009	32 nm
2011	22 nm
2014	14 nm
2016	~~10 nm~~
2017	~~10 nm~~
2018	10 nm?
2019	10 nm!

https://www.anandtech.com/show/12693/intel-delays-mass-production-of-10-nm-cpus-to-2019

Moore's Law is Slowing for Intel

- Even by own timeline Intel is behind schedule
- Competitors (TSMC & Samsung) are pushing ahead, but maybe only a few more steps left
- Time gap between new processes may grow

	Intel First Production
1999	180 nm
2001	130 nm
2003	90 nm
2005	65 nm
[illegible]	[illegible]
2009	32 nm
[illegible]	[illegible]
2014	14 nm
2016	~~10 nm~~
2017	~~10 nm~~
2018	10 nm?
2019	10 nm!

https://www.anandtech.com/show/12693/intel-delays-mass-production-of-10-nm-cpus-to-2019

May no longer be able to count on exponentially more transistors each year

Specialization is Necessary

- Dark Silicon + End of Moore's Law = Fewer additional active transistors
- Implies to do more, must get more benefit from each transistor
- *Specialization* - perform better at some tasks by sacrificing some flexibility

Example Specialization Exploits

- Hardwire control flow
- Increase parallelism
- Deeper pipelines
- Reduce arithmetic precision
- Custom-tailor memory hierarchy
- Hardwire data flow
- Systolic data exchanges

Opportunity for an Accelerator

- For an accelerator opportunity to be worthwhile, it should satisfy the following:

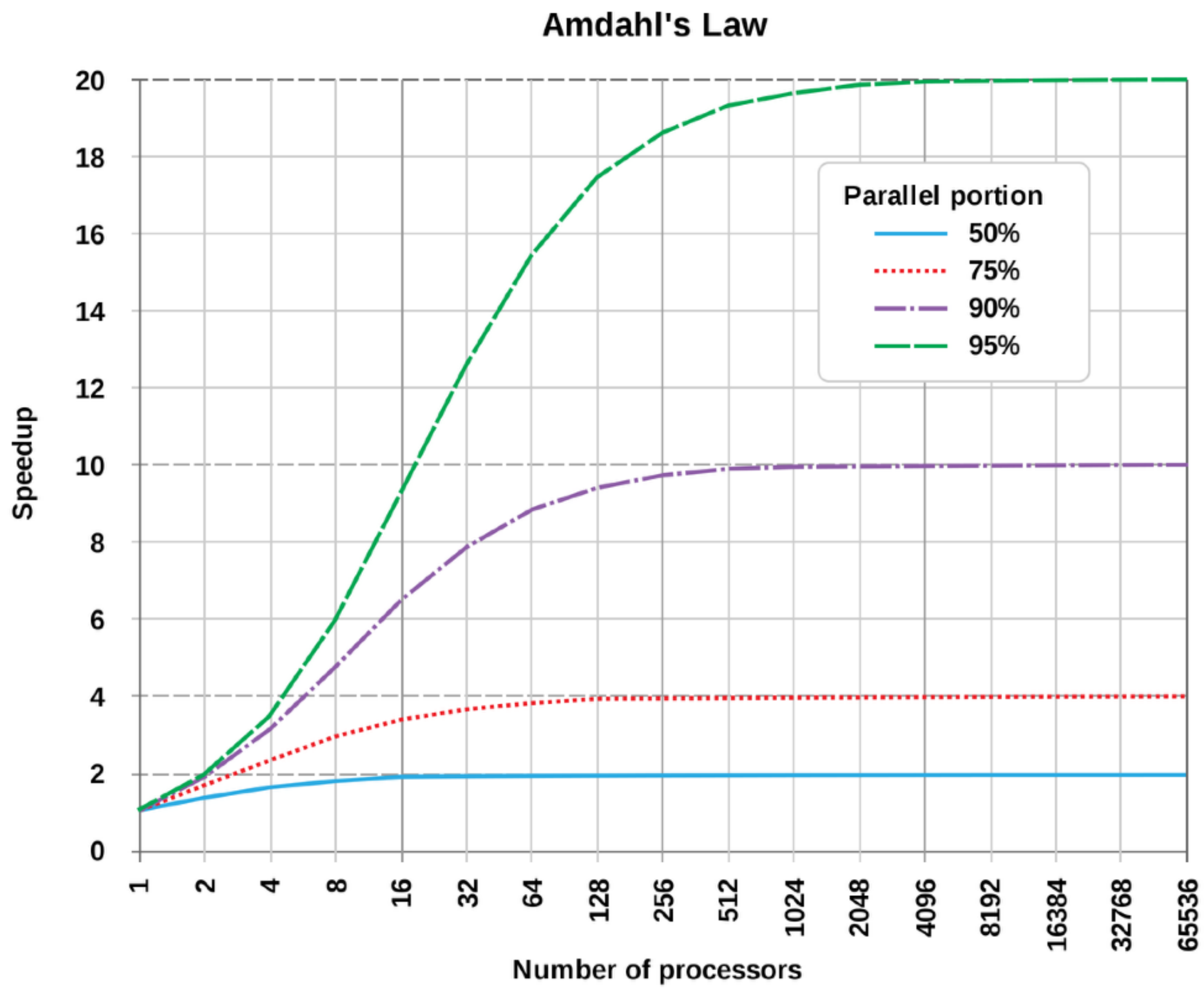
1. *Restricted* target workload - hard to improve on everything at once
2. Current platforms are *inefficient* at target workload - need room to improve
3. Potential *benefit* to target workload makes accelerator cost worthwhile

Amdahl's Law

$$\text{speedup} = \frac{1}{(1-f) + f/P}$$

f is workload fraction

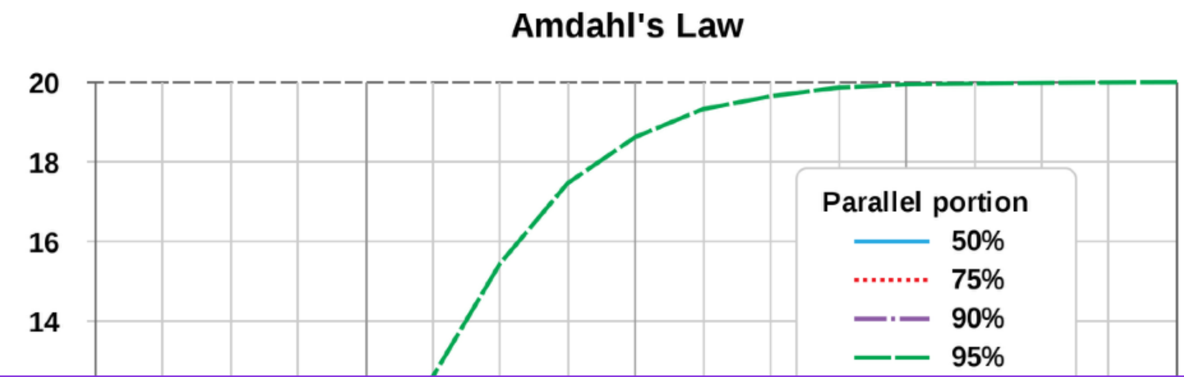
P is parallel speedup



- Potential speedup from parallelization is bounded by the fraction of the workload being parallelized (or specialized)

Amdahl's Law

$$\text{speedup} = \frac{1}{(1-f) + f/P}$$



Amdahl's Law implies need to target a significant portion of the workload to get overall speedup

- Potential speedup from parallelization is bounded by the fraction of the workload being parallelized (or specialized)

Accelerator Design Considerations

- Identify what workload features to exploit
  - This enables efficiency gains
  - Look for parallelism and locality
  - Implicitly, also picking what features to not support (or at least not efficiently)
- How will it be programmed?
- How will it interface with the rest of the system and workload?

Accelerator Programming Options

software support cost

- Expose low-level complicated/brittle interface
  - *BAD* in practice, but maybe ok for new research
- Hide behind library calls (e.g. HEVC decoder)
  - Great, but offers less programmability
- Develop a new general abstraction
  - *Example*: SIMT/CUDA for GPUs
  - Passes some complexity to programmer
- Domain-specific language w/ optimizing compiler
  - *Example*: Halide
  - Near-ideal, but largest software support cost

Accelerator Integration Options

- ISA extensions (e.g. FPU or SIMD)
  - Accelerator instructions intermixed with host CPU instructions and accelerator is essentially another functional unit
- Coprocessor (e.g. GPU)
  - CPU triggers operations on accelerator, but lets it operate asynchronously
- Fully independent system (e.g. Anton)
  - Requires no host CPU or access to shared memory

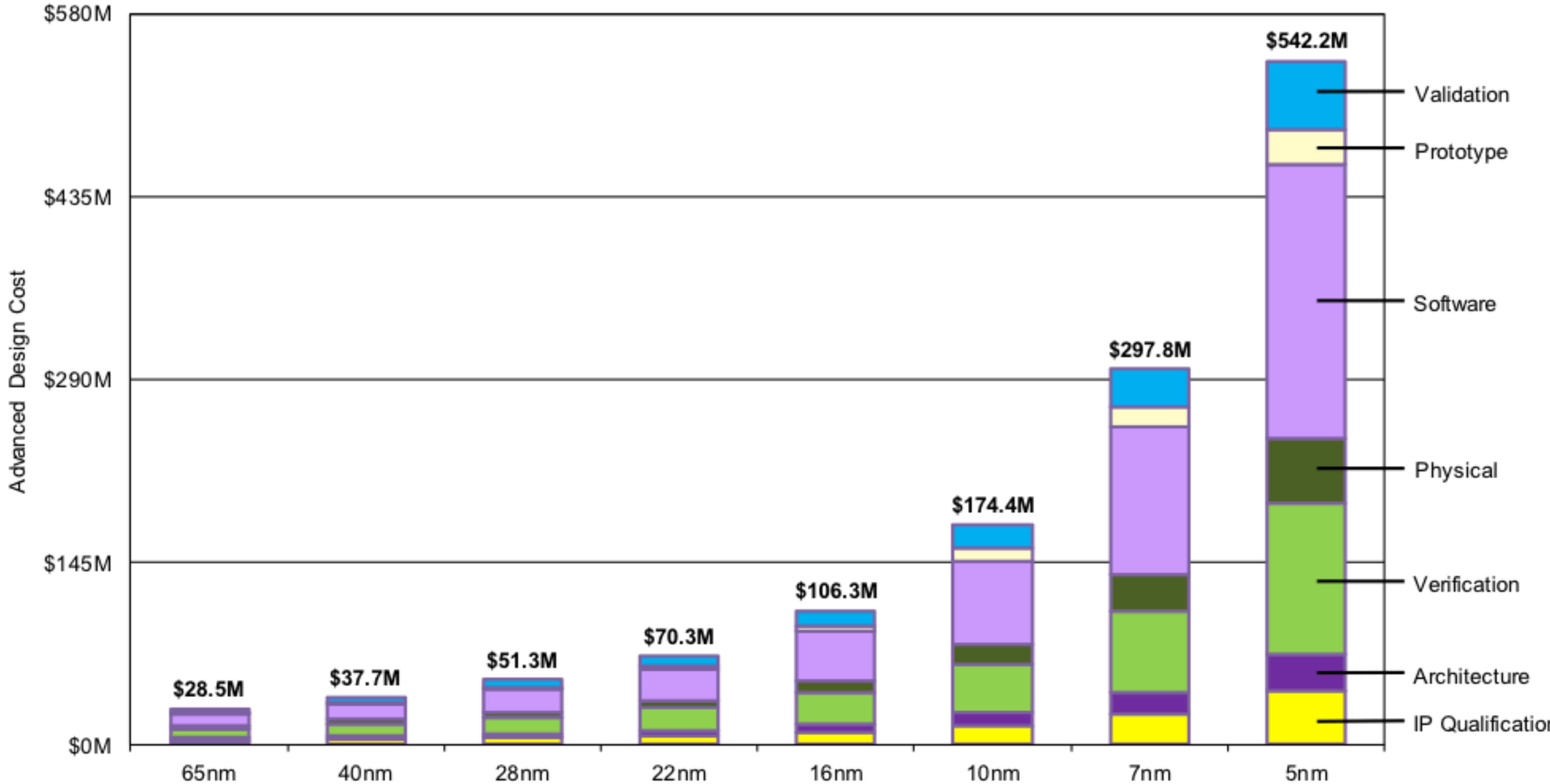
needs larger work chunks

needs more from host CPU

Design Challenges

- Flexibility vs specialization tradeoff
  - Specialization needed to gain efficiency
  - Increased flexibility can future-proof against algorithmic changes and even increase potential market
- Making benefit outweigh cost
  - Benefit is improvement and market size
  - Cost includes design (time and $$) as well as integration

Design Costs Rise for Leading Process



from International Business Strategies (IBS)
https://semiengineering.com/big-trouble-at-3nm/

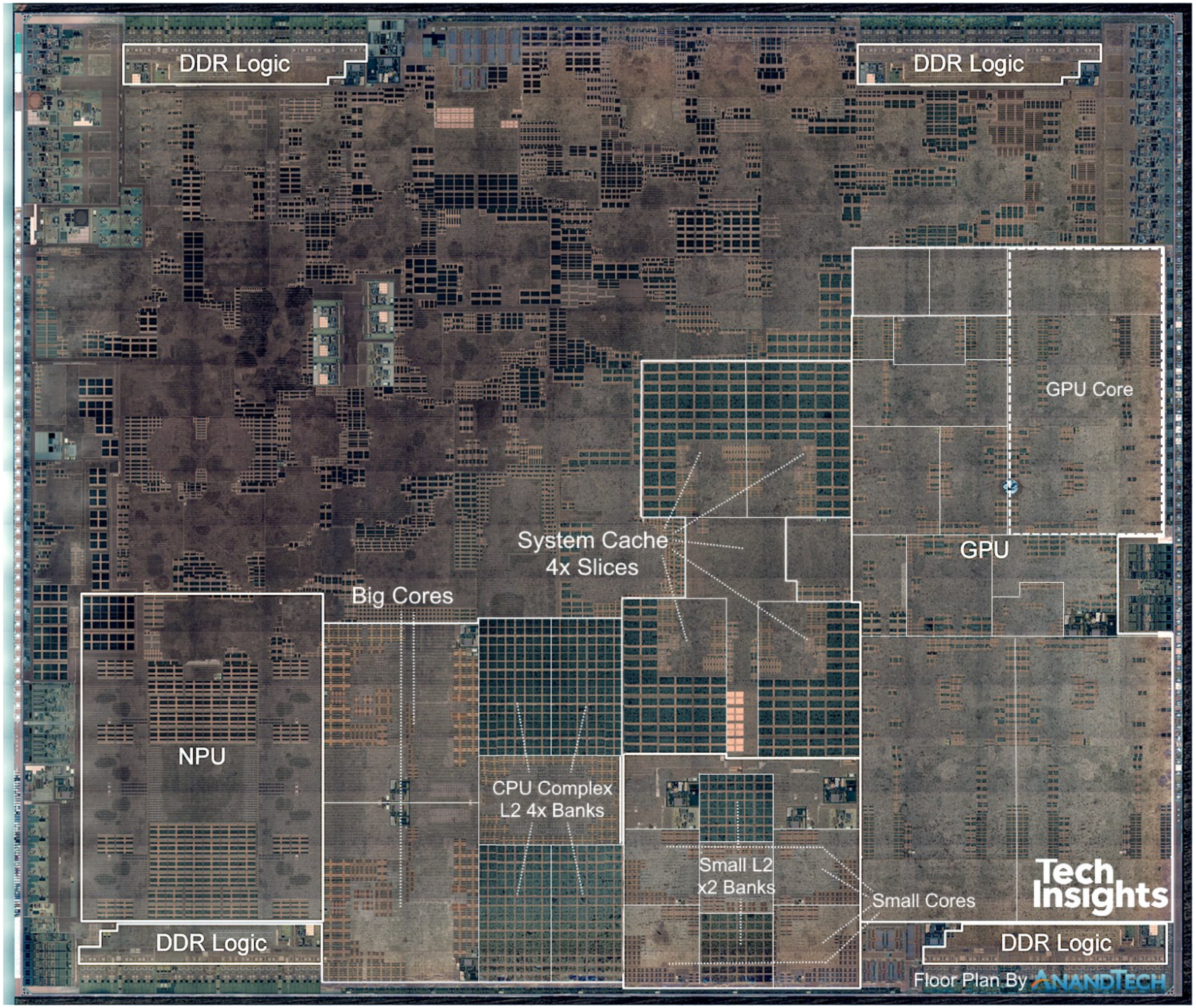
Ways to Reduce Design Costs

- Leverage new design methodologies
  - Design at a higher level and go agile
- Reuse open-source building blocks
- Use an older process - with specialization, may still provide benefit
- Consider using an FPGA
  - Trades design cost for efficiency

Are CGRAs The Solution?

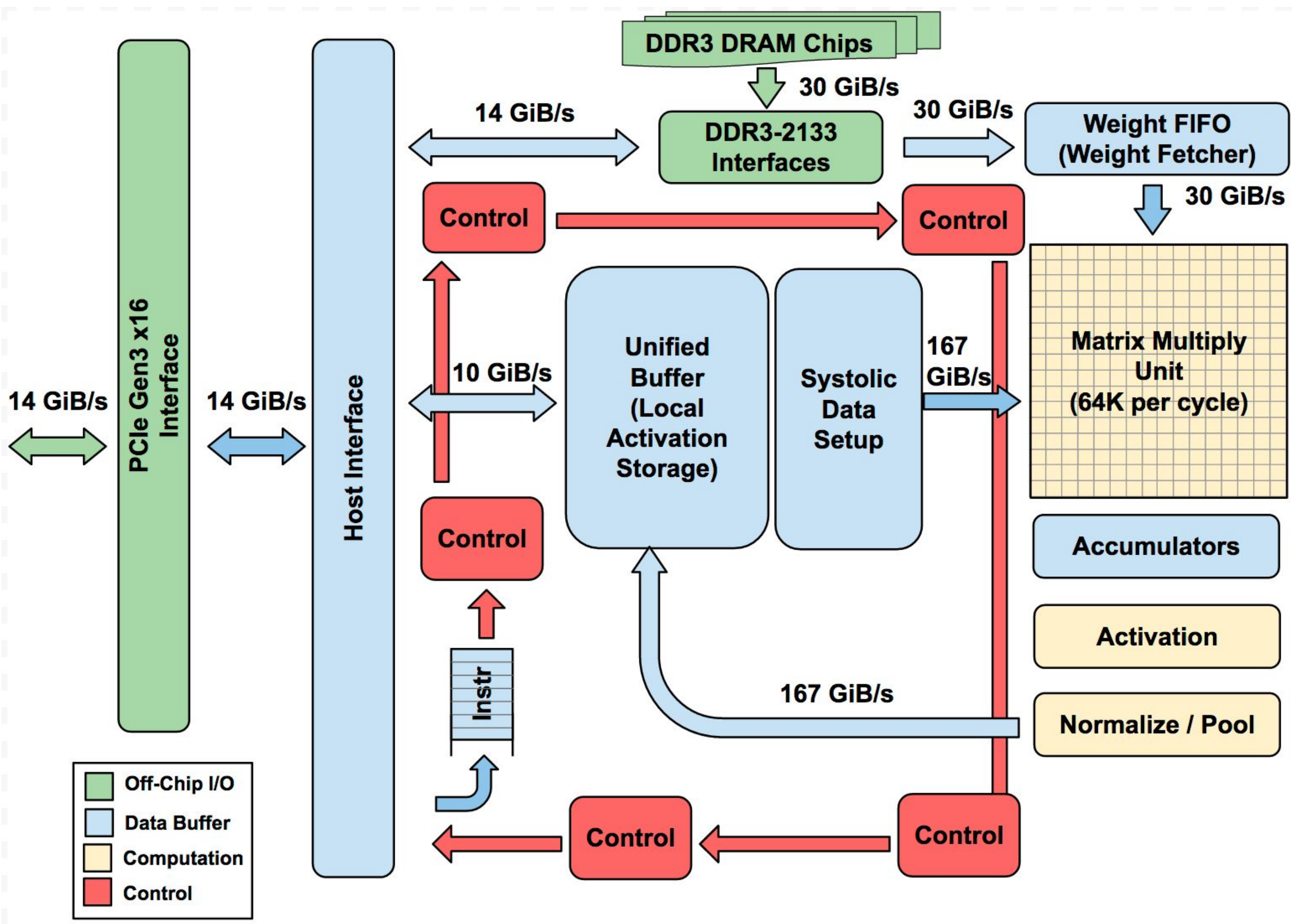
- FPGAs provide tremendous configurability (down to single-bit signals and logic gates)
  - Naturally, causes many resources to be dedicated to routing and configuration, reducing efficiency
- Coarse Grain Reconfigurable Array (CGRA)
  - Trades some flexibility for efficiency
  - Provide accelerator building blocks, let configuration put them together
  - Old research idea experiencing large revival

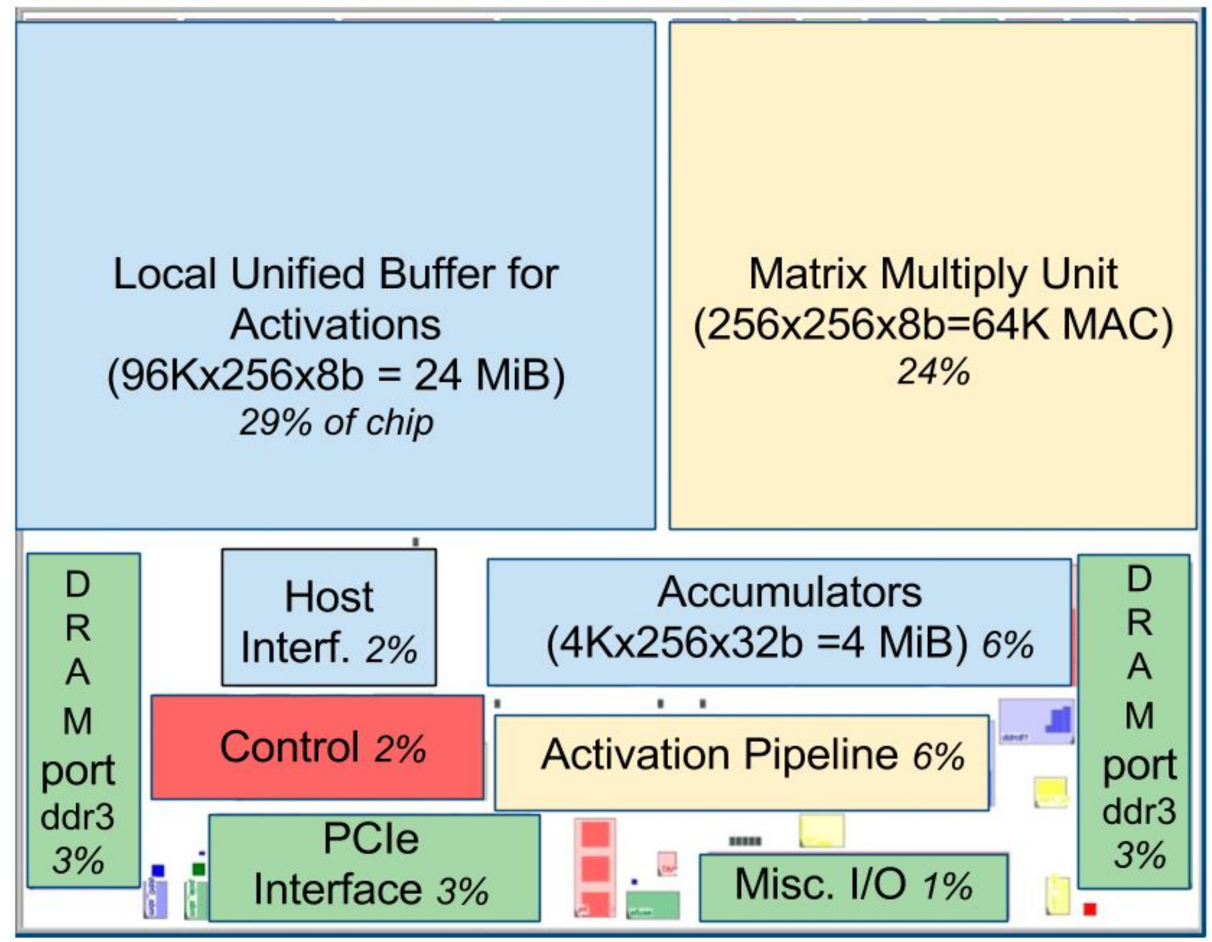
Peek at A12 SoC (2018 iPhone)



https://www.anandtech.com/show/13392/the-iphone-xs-xs-max-review-unveiling-the-silicon-secrets/2

Google's Tensor Processing Unit (TPU)





"In-Datacenter Performance Analysis of a Tensor Processing Unit" (ISCA 2017)

- Leverages on-chip memory and systolic data movement to communicate efficiently

Accelerator Summary

- Efficiency is crucial, and technology scaling trends make it even more important
- Accelerators specialize for target workload
  - Trade generality for efficiency
- Accelerator design is a vibrant research area with large industry impact

Introductions

- Please share:
  - Name
  - Major & Year
  - Research area (or interest)
  - Fun: If you had world-class athletic ability, which Olympic sport would you compete in?

Course Learning Outcomes

- After completing the course, the student will be able to ...
  - Characterize a workload
  - Understand how accelerators provide benefit
  - Suggest an accelerator for a target workload
- + improve skills as researcher

Course Activities

- In-class paper discussion
  - Read (& summarize) 1-2 papers (per class)
  - Lead discussion (once per quarter)
  - Scribe 1 discussion (once per quarter)
  - (Participation & Attendance)
- Course Project
  - Project proposal (1)
  - Project reports (2)
  - Peer Review (1)
  - Project Presentation (1)

Grade Breakdown

- 50% - Project
- 20% - Reading Summaries
- 10% - Presenting Paper & Leading Discussion
- 5% - Scribing Paper Discussion
- 15% - Participation & Attendance

Projects Steps

- Proposal - pitch topic (1 page)
- Part 1 - Characterize workload (4 pages)
  - identify key workload features
- Peer review one part 1 submission
- Part 2 - High-level design (8 pages)
  - analyze design + revised part 1
- Presentation

Departing Details

- Office hours MW 4-5pm E2-229
- Sign up for piazza
- Submit paper preferences on canvas (for discussion lead)
- Start reading 2 papers for Friday
  - Summary due 9 AM Friday