ETH zürich

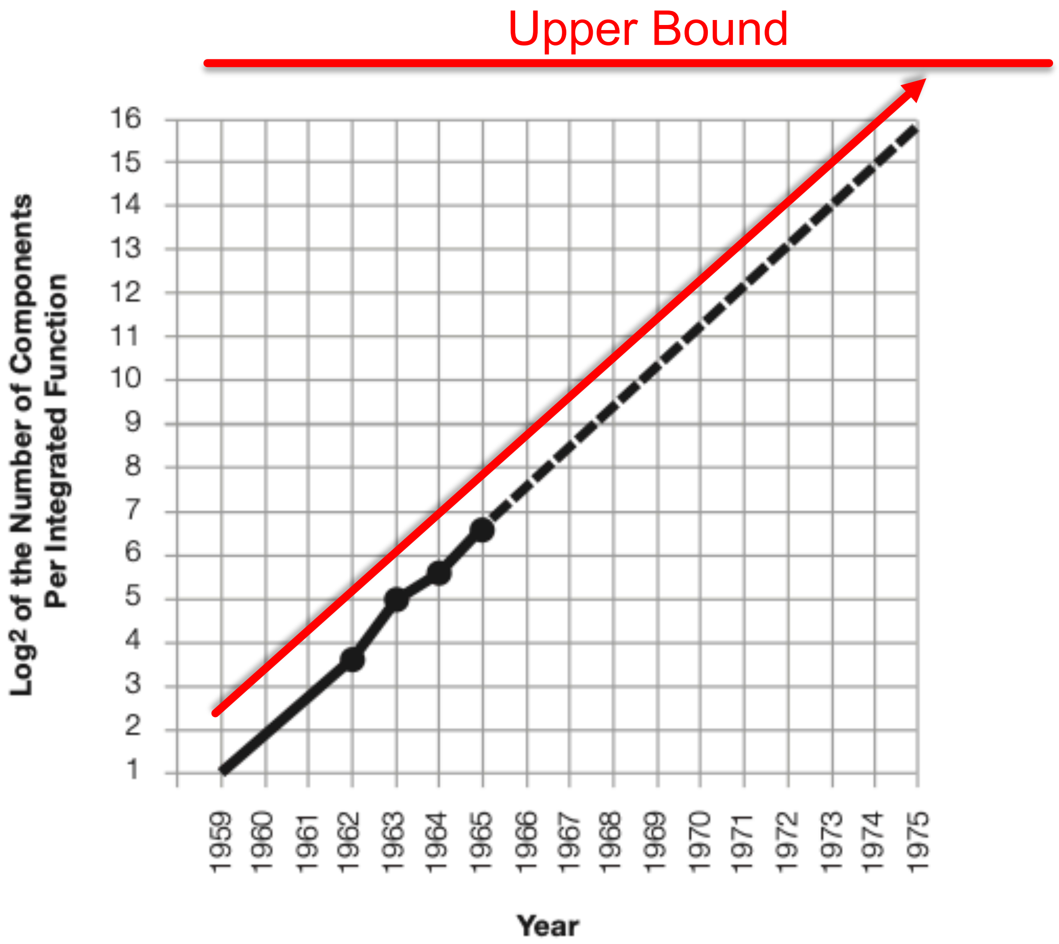
The Accelerator Wall: Limits of Chip Specialization (HPCA19)

Adi Fuchs, David Wentzlaff - *Princeton University (Department of Electrical Engineering)*

Presented by Manuel Burger, ETH Zürich

Moore‘s Law

- Cramming more components onto integrated circuits – Gordon E. Moore,1965
- Final Transistor Size 5nm



https://newsroom.intel.com/wp-content/uploads/sites/11/2018/05/moores-law-electronics.pdf

Accelerators to the rescue

- **Sacrifice flexibility** and target **specific application** domains
  - Performance (Throughput)
  - Energy Efficiency (Throughput / Watt)
- Again **limited by transistor size** at some point

Outline

- Problem & Goal
- Multi Phase Analysis
  - New Metric CSR (Chip Specialization Return)
  - Physical Transistor Model
  - Case Studies on specific domains and accelerators
  - Accelerator Wall
- Strengths
- Weaknesses
- Discussion

Problem

- CMOS scaling is ending
- Throughput and Efficiency through accelerators
- Gains (partly / entirely) rely on CMOS scaling
- Slowdown and eventual breakdown in performance scaling

- **Accelerator Wall**

Goal

- **Analyzing the limits of chip specialization**

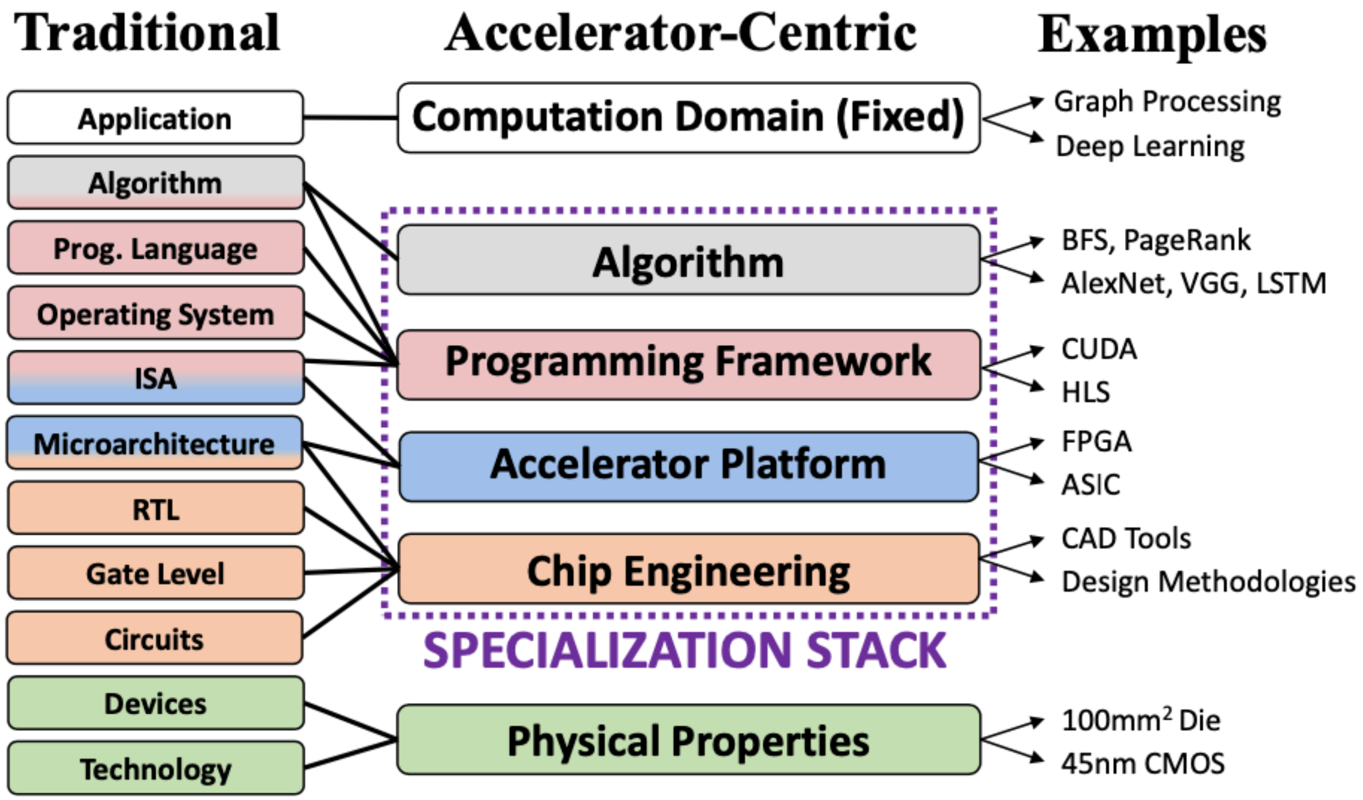
- **Analyze application scaling** behavior on various accelerator architectures
- **Build projection models** for performance metrics for fixed application domain
- **Predict upper performance limit** for fixed application domain
- **Understand the origins** of these imposed upper bounds

Outline

- Problem & Goal
- Multi Phase Analysis
  - **New Metric CSR (Chip Specialization Return)**
  - Physical Transistor Model
  - Case Studies on specific domains and accelerators
  - Accelerator Wall
- Strengths
- Weaknesses
- Discussion

Analyzing the software stack

- **Are accelerators driven by Specialization or Transistors?**
- Objective functions:
  - Throughput
  - Energy Efficiency
- **Goal: isolate** contribution of **pyhsical layer**



CSR (Chip Specialization Return)

- *CSR:* „How much did the chip‘s compute capabilities improve under a fixed physical budget?“

$$CSR(Alg, Fwk, Plt, Eng) = \frac{Gain(Alg, Fwk, Plt, Eng, Phy)}{Gain(Phy)}$$

- *Gain(Alg, Fwk, Plt, Eng, Phy)*
  - Effective Gain (measured by execution)
  - Coupled to objective function
- *Gain(Phy)*
  - **Theoretical** gain by physical scaling
  - **CMOS Potential**

CSR – Gain Metric

- Objective: Abstract chip specialization
- Relative metric
- **CSR > 1**: Overall gains rely on specialization stack optimizations
- **CSR < 1**: Gains rely on CMOS
- **Higher CSR**: advances in Alg, Fwk, Plt, Eng
- How to compute?
  - Gain(Alg,Fwk,Plt,Eng,Phy) → simple
  - Gain(Phy) → ???

$$CSR(Alg, Fwk, Plt, Eng) = \frac{Gain(Alg, Fwk, Plt, Eng, Phy)}{Gain(Phy)}$$

Computing Gain(Phy)

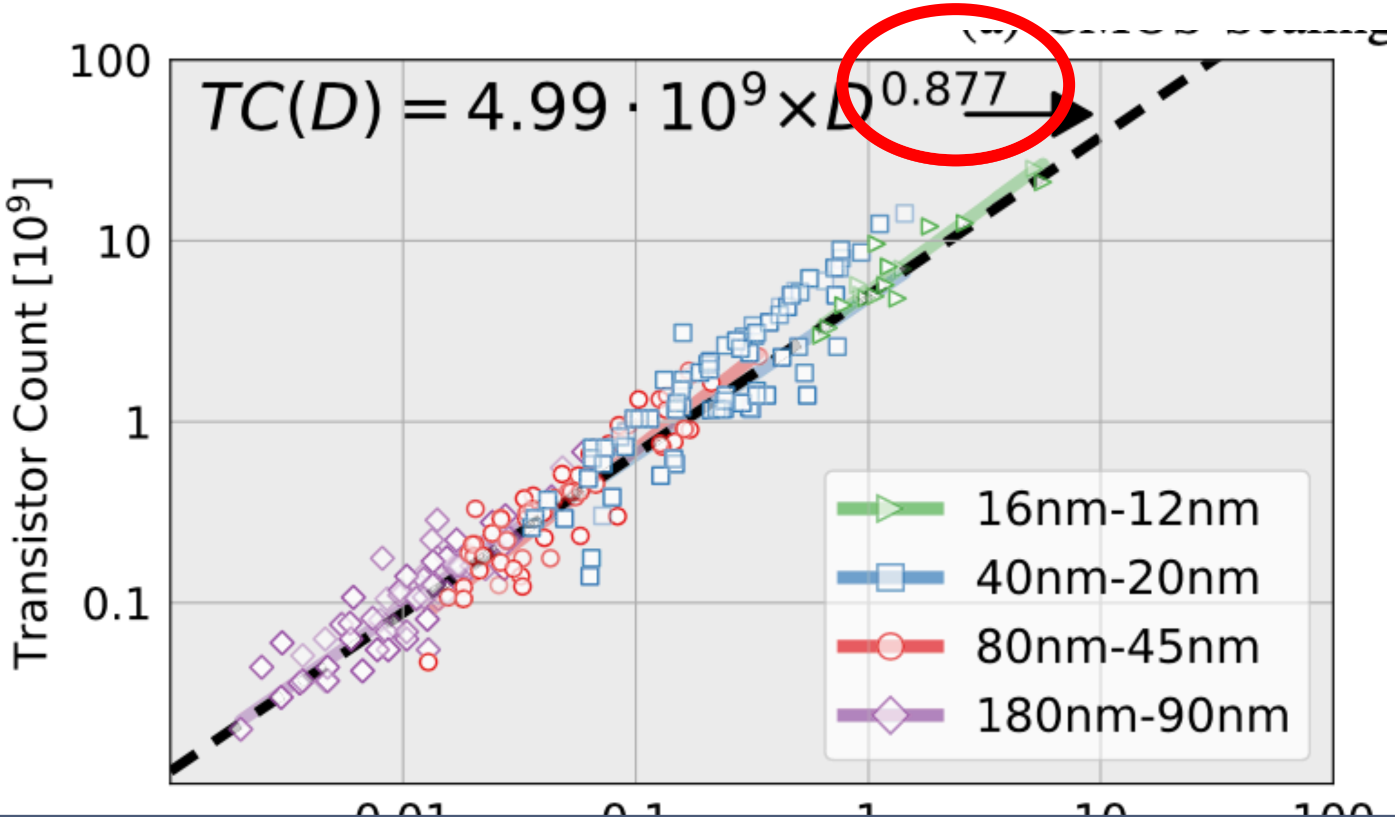
- Extract or **predict # of active transistors** on chip on new CMOS technology
- Assume (almost) **perfect scaling** of application
- **Scaling factor** of # active trans. yields **Gain(Phy)**

- Why?
  - Fixed application domain
  - Accelerators used for applications with high levels of parallelism
  - Active transistors
  - Worst case analysis

Outline

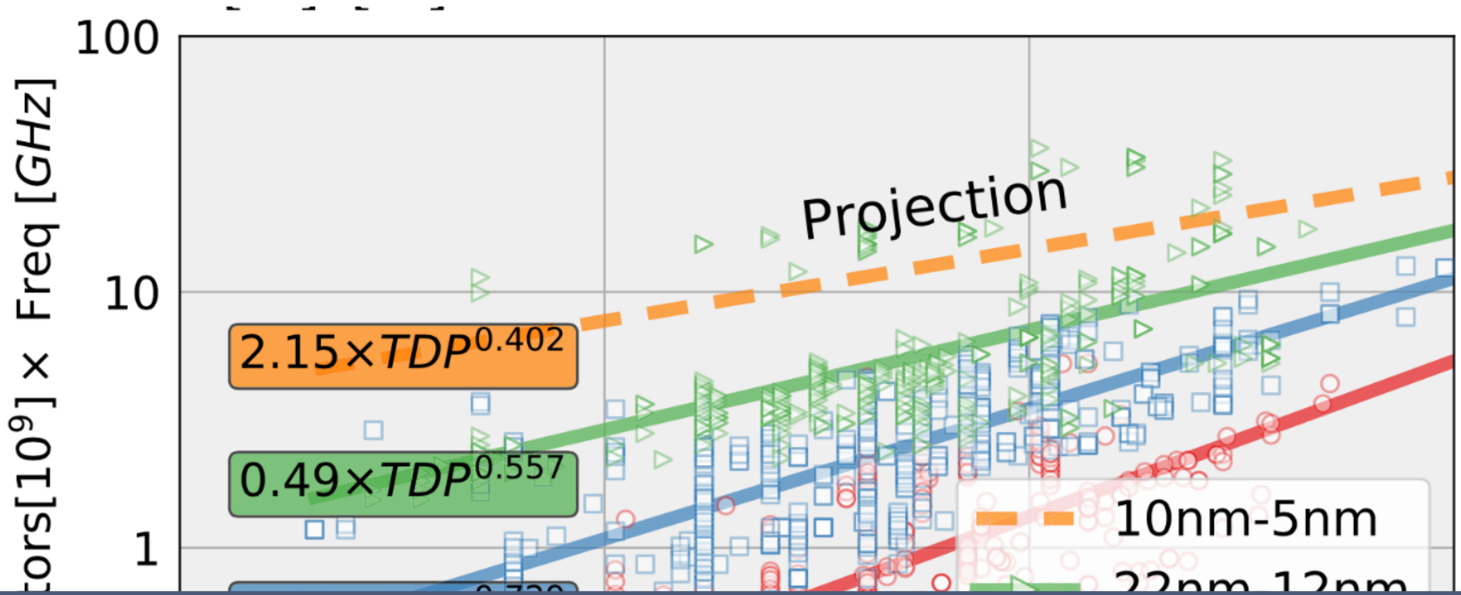
- Problem & Goal
- Multi Phase Analysis
  - New Metric CSR (Chip Specialization Return)
  - **Physical Transistor Model**
  - Case Studies on specific domains and accelerators
  - Accelerator Wall
- Strengths
- Weaknesses
- Discussion

Physical Transistor Model – Transistor Budget



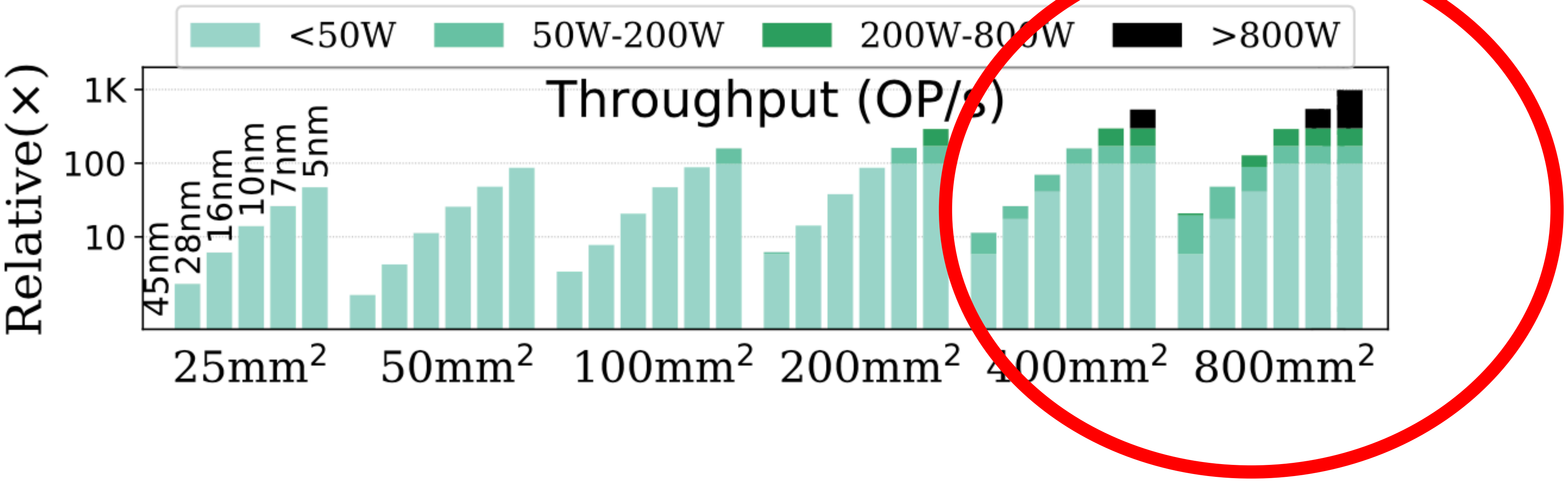
Sub-linear scaling due to dead silicon

Physical Transistor Model – Active Transistor Budget



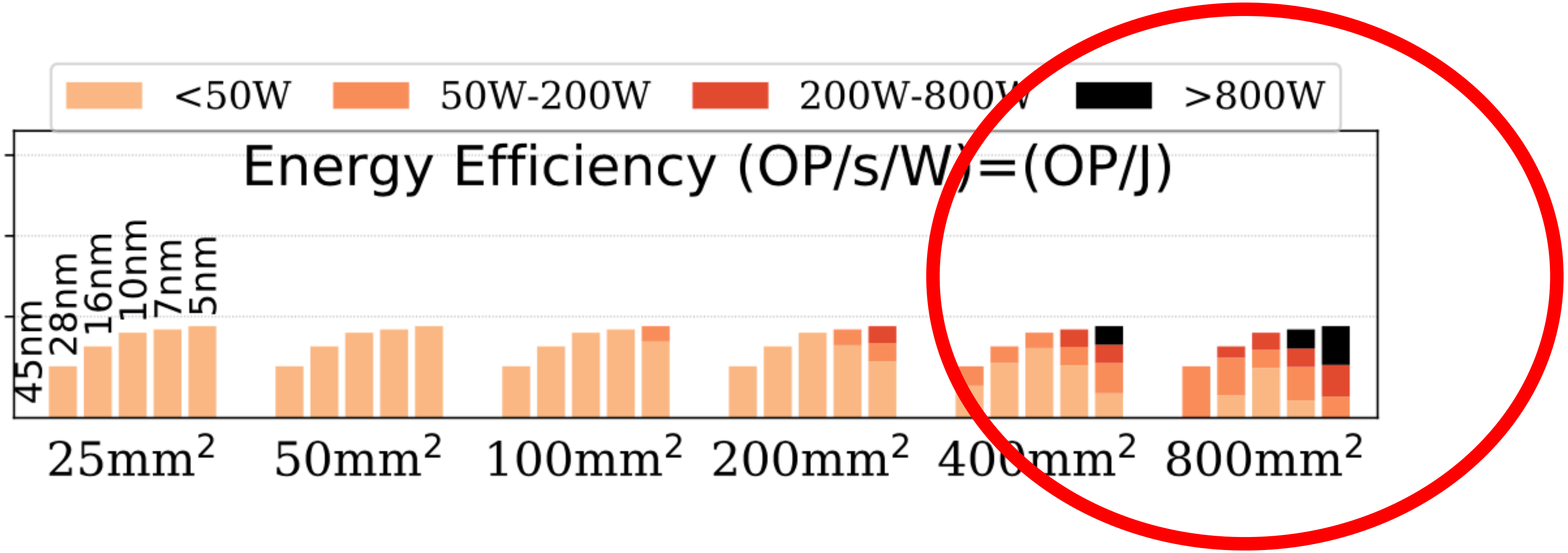
- TDP limits active transistor count
- Smaller nodes more affected by TDP constraints

Physical Transistor Model – CMOS Potential



Limiting TDP caps (especially for larger chips)

Physical Transistor Model – CMOS Potential



Energy efficiency favours smaller chips (due to static power consumption)

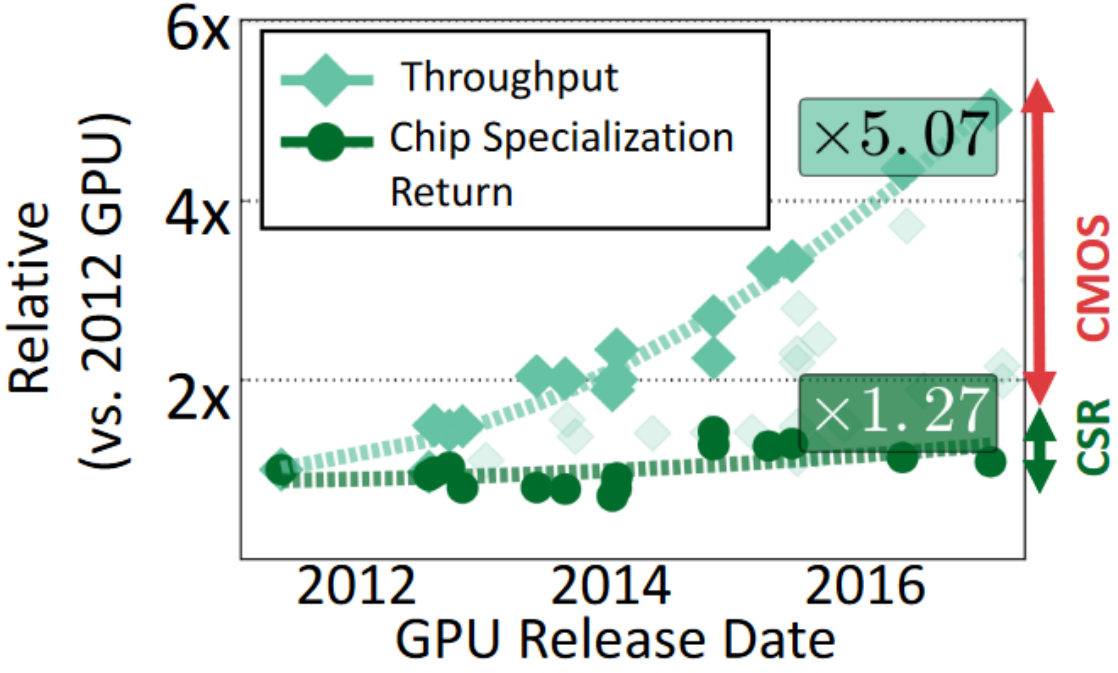
Outline

- Problem & Goal
- Multi Phase Analysis
  - New Metric CSR (Chip Specialization Return)
  - Physical Transistor Model
  - **Case Studies on specific domains and accelerators**
  - Accelerator Wall
- Strengths
- Weaknesses
- Discussion

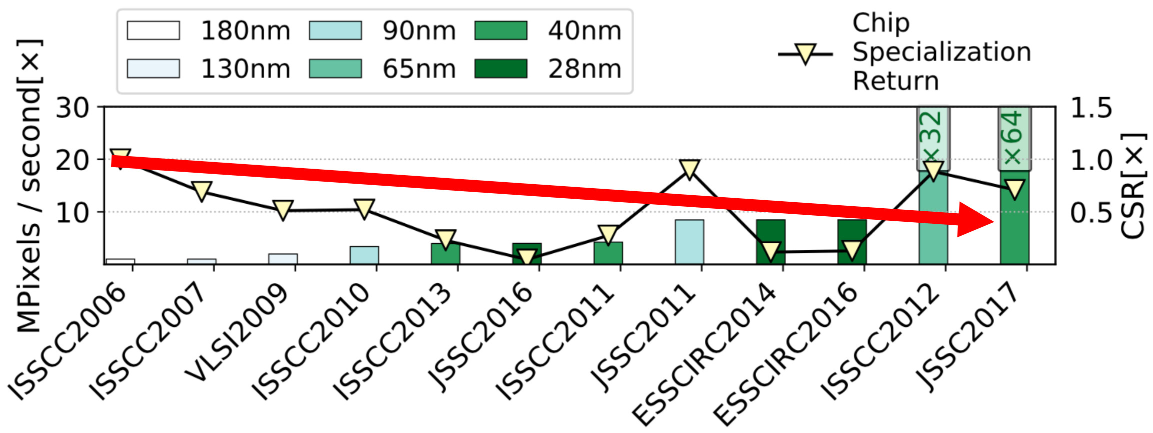
Case Study 1 – GPUs for graphics

- **Throughput (Gain) Improved: 5.07x**
- **Specialization Contribution: 1.27x**
- **CMOS Scaling Contribution: 4x**
- Similarly for **energy efficiency**

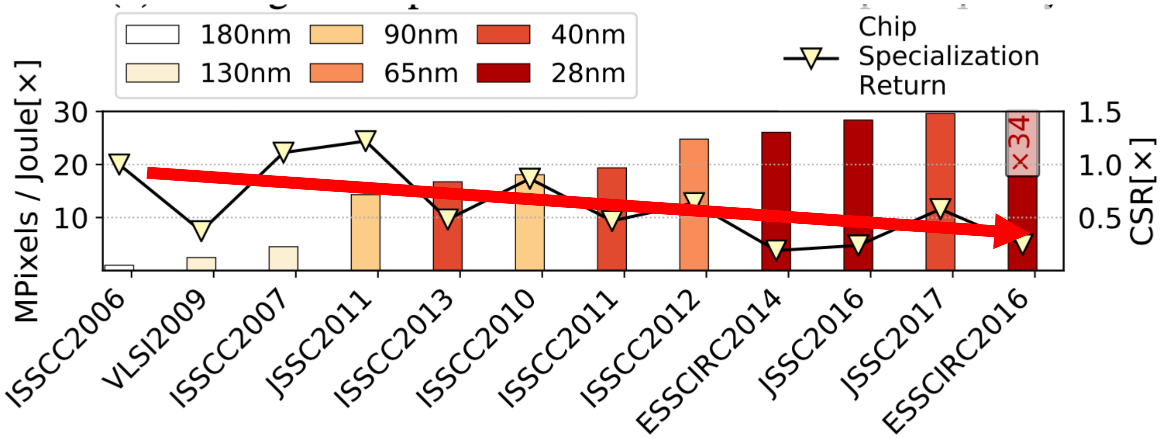
GTA V FHD



Case Study 2 – ASIC video decoders



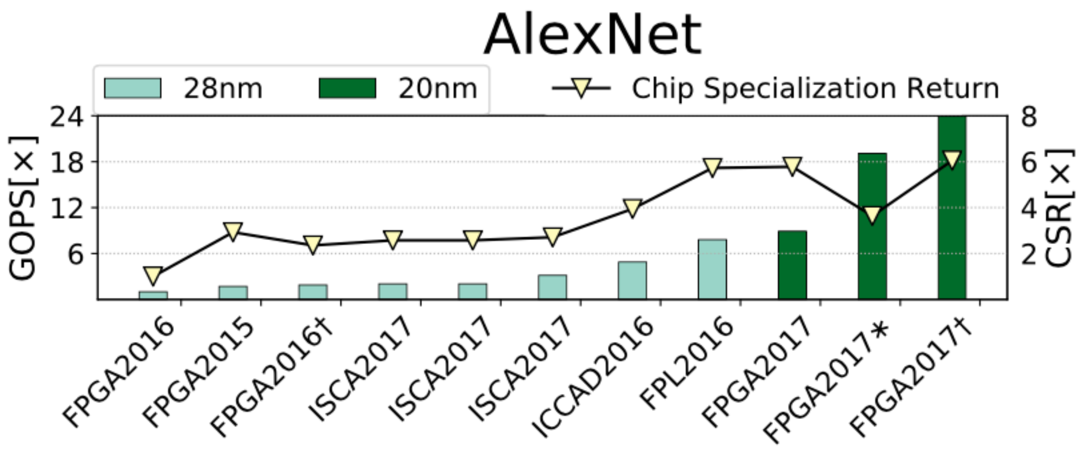
(a) Scaling of Performance and Chip Specialization Return

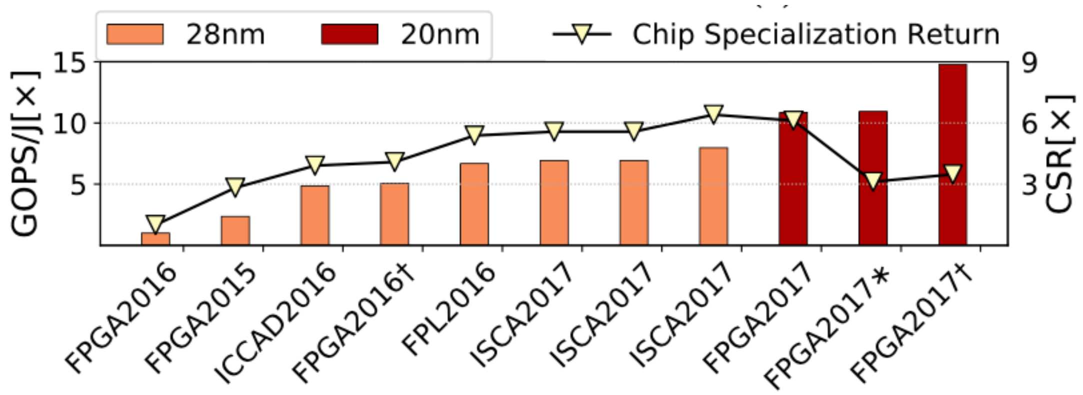


(c) Scaling of Energy Efficiency and Chip Specialization Return

- Diminishing CSR
- Gain relying on CMOS potential and scaling

Case Study 3 – FPGA for conv. neural nets

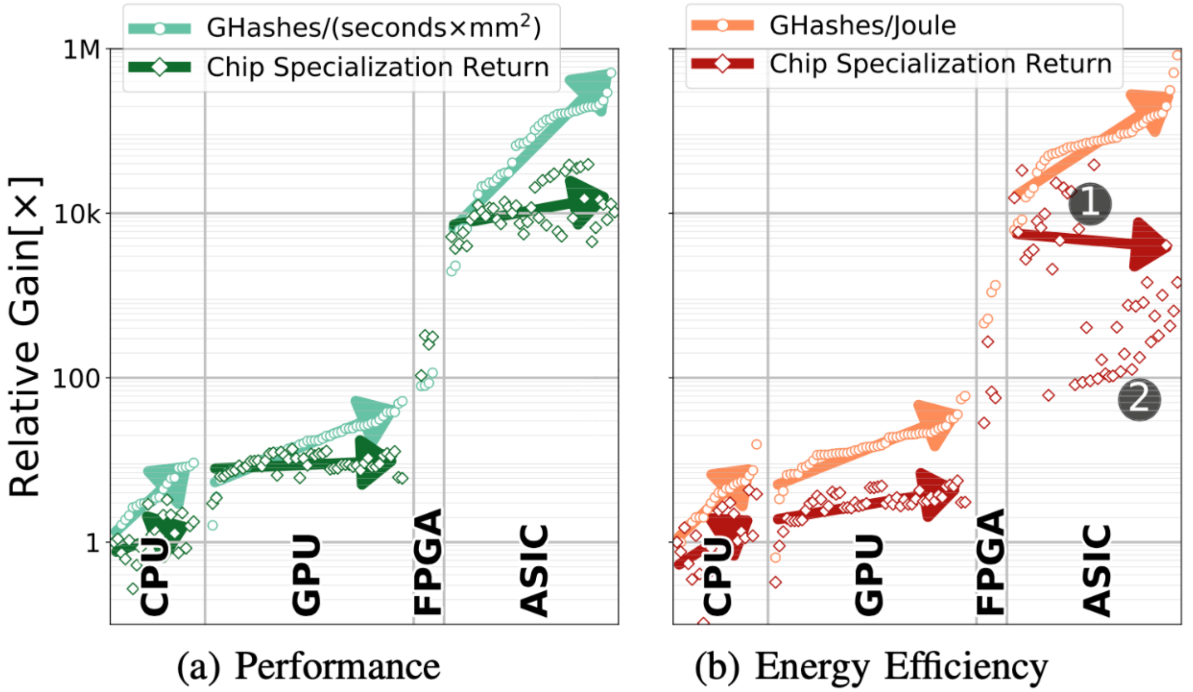




- Newer domains show better CSR values
- Stagnating CSR

Case Study 4 – Bitcoin mining across platforms

- High **CSR boost on platform change**
- Almost **constant CSR within platform**
- **Decline in CSR on ASICs** shows heavy reliance on CMOS scaling
- **Extremely specific computation**, small optimization space



(a) Performance (b) Energy Efficiency

Figure 9: Bitcoin Mining Capabilities of CPU, GPU, FPGA and ASIC chips (vs. AMD Athlon 64 CPU Miner).

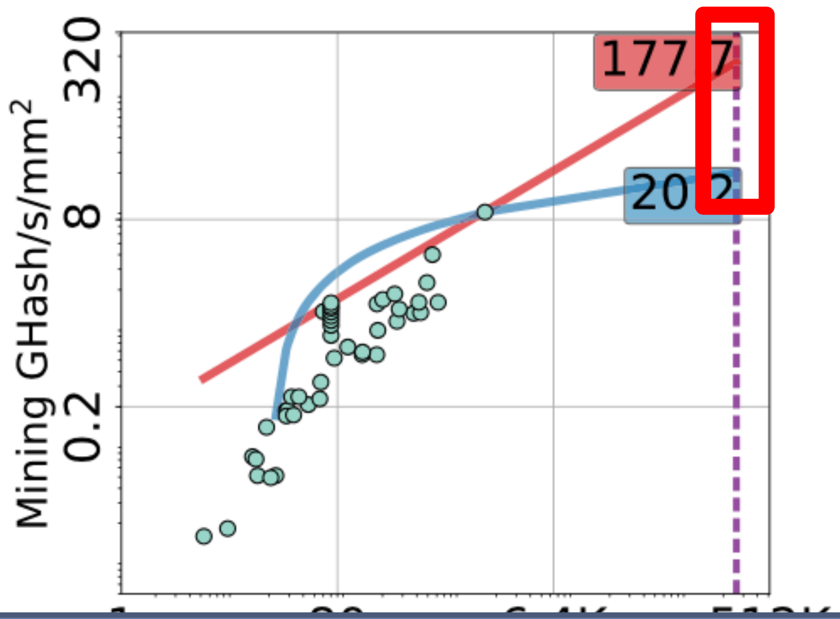
Case Studies - Summary

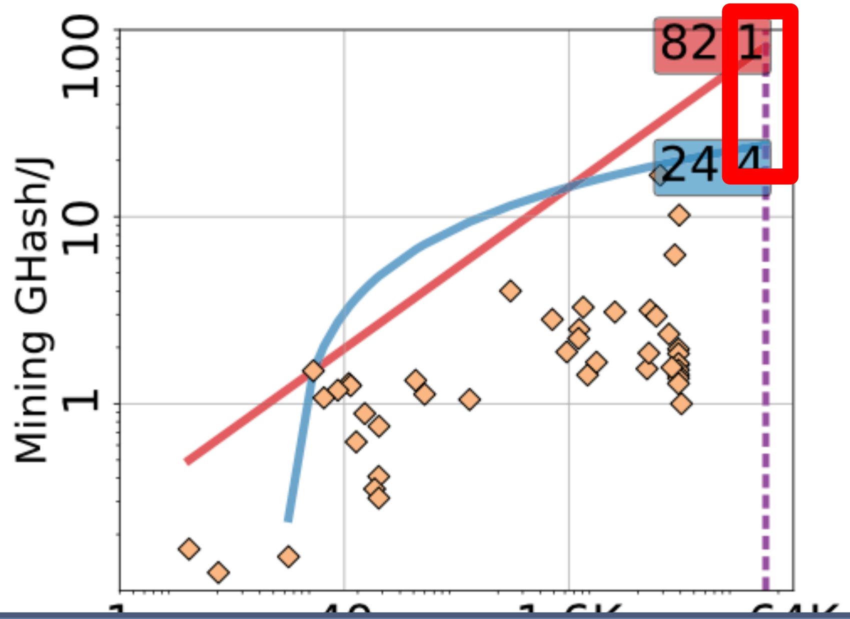
- Specialization returns and computation maturity
- Introduction of a new specialization platform
- Confined computations
- Dependence on CMOS scaling

Outline

- Problem & Goal
- Multi Phase Analysis
  - New Metric CSR (Chip Specialization Return)
  - Physical Transistor Model
  - Chip Specialization Upper Bounds
  - Case Studies on specific domains and accelerators
  - **Accelerator Wall**
- Strengths
- Weaknesses
- Discussion

The Accelerator Wall





The Accelerator Wall

The Accelerator Wall

- Performance scaling linear with CMOS
- Energy efficiency to scale sub-linear (logarithmic)
- Predictions easier for more mature domains (algorithmically stable)

Conclusion

- Developped metric for analysis
- Modelled potential physical gains by CMOS scaling
- Characterize influence of CMOS scaling on well-known application domains
- Show the accelerator wall based on the developed models and concepts

- So the goals have been achieved…..have they?

Questions?

Outline

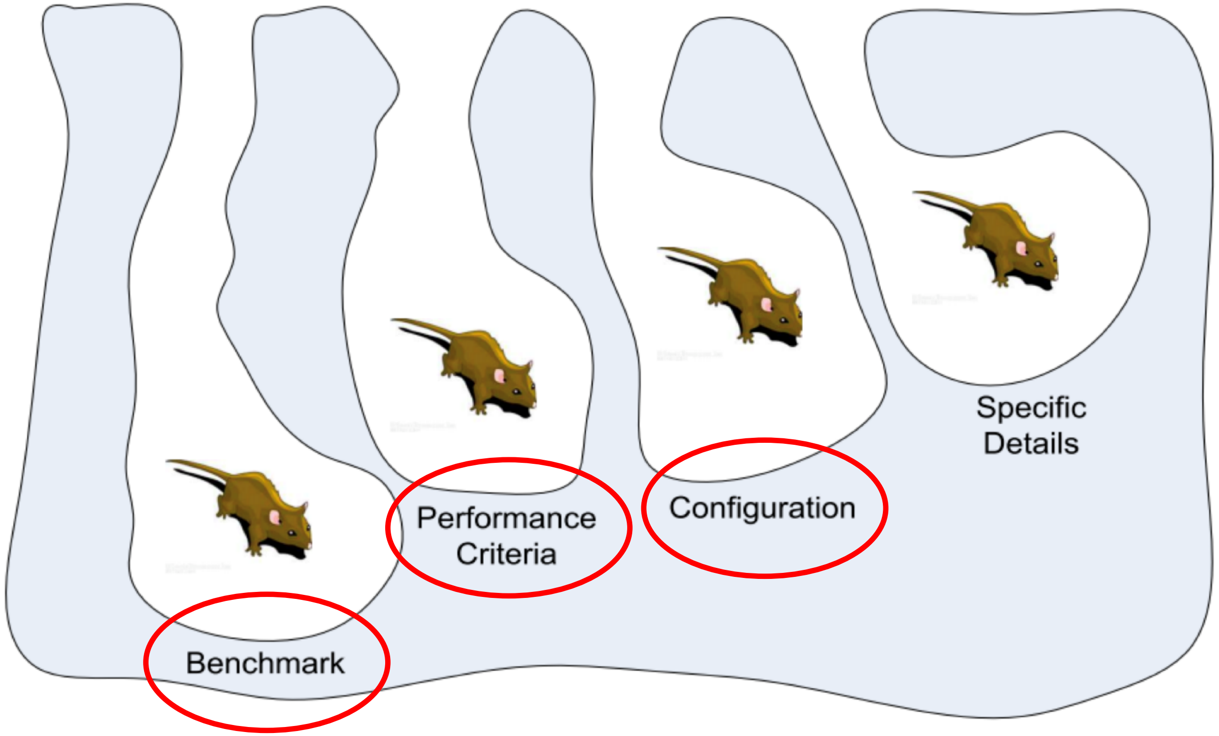
- Problem & Goal
- Multi Phase Analysis
    - New Metric CSR (Chip Specialization Return)
    - Physical Transistor Model
    - Chip Specialization Upper Bounds
    - Case Studies on specific domains and accelerators
    - Accelerator Wall
- **Strengths**
- **Weaknesses**
- **Discussion**

Strengths

- High level of abstraction
  - CSR metric
- Analysis across a wide dataset incorporating many different use cases, maturities and platforms
- Developped general procedure and tools, which could be applied to many other application domains
- Insights into accelerator development over time

Weaknesses

Performance Analysis Rat Holes

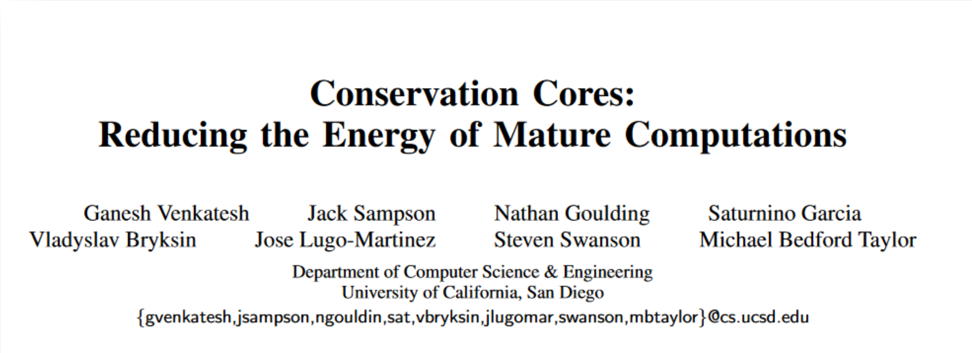


Source: P. Jarupunphol, “Using Buddhist Insights to Analyse the Cause of System Project Failures,” Ph.D. Thesis, 2013

Weaknesses

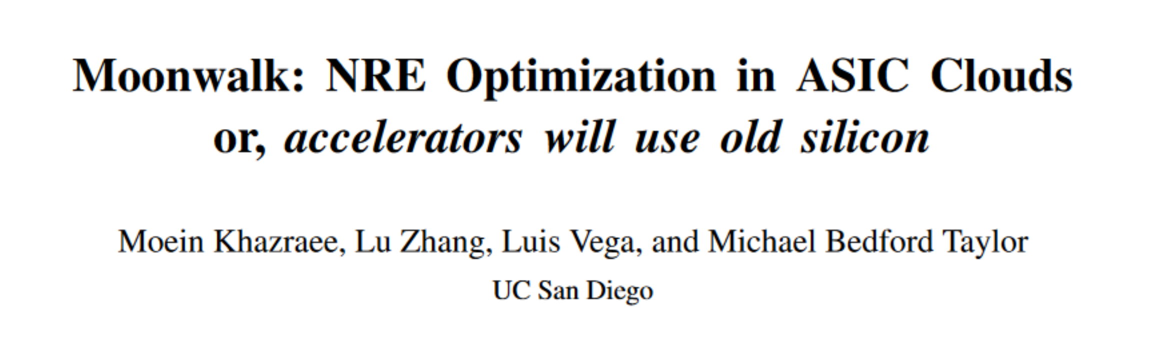
- Transistor model based on CPU/GPU data
- Unreliable data sources
- Evaluation not too focused
  - Many domains
  - Many configurations
- High dependency on fitting curves (many implicit assumptions)
- Assuming perfect scaling: Amdahl‘s law
- Difficult to start reading
  - Introduction of many new and own concepts

Related Work



Conservation Cores:
Reducing the Energy of Mature Computations

Ganesh Venkatesh, Jack Sampson, Nathan Goulding, Saturnino Garcia
Vladyslav Bryksin, Jose Lugo-Martinez, Steven Swanson, Michael Bedford Taylor
Department of Computer Science & Engineering
University of California, San Diego
{gvenkatesh,jsampson,ngouldin,sat,vbryksin,jlugomar,swanson,mbtaylor}@cs.ucsd.edu

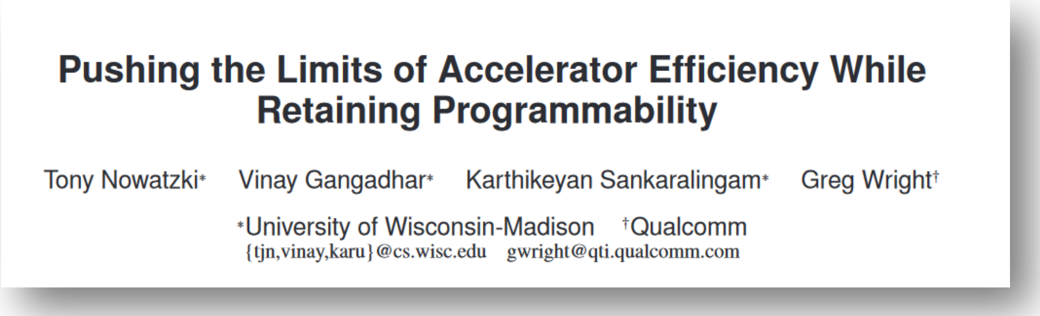


Moonwalk: NRE Optimization in ASIC Clouds
or, *accelerators will use old silicon*

Moein Khazraee, Lu Zhang, Luis Vega, and Michael Bedford Taylor
UC San Diego

- ASPLOS (2010, 2017)
- **Dark Silicon** limits number of usable transistors

Related Works

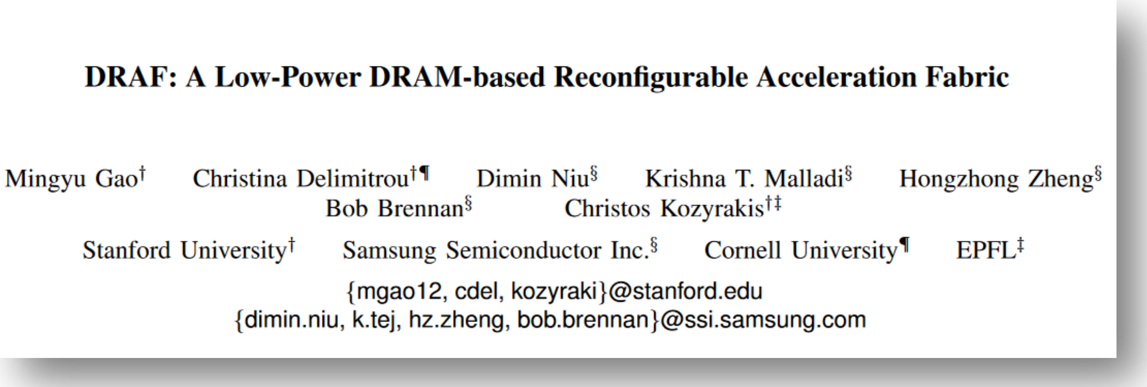


Pushing the Limits of Accelerator Efficiency While Retaining Programmability

Tony Nowatzki[*] Vinay Gangadhar[*] Karthikeyan Sankaralingam[*] Greg Wright[†]

[*]University of Wisconsin-Madison [†]Qualcomm

{tjn,vinay,karu}@cs.wisc.edu gwright@qti.qualcomm.com



DRAF: A Low-Power DRAM-based Reconfigurable Acceleration Fabric

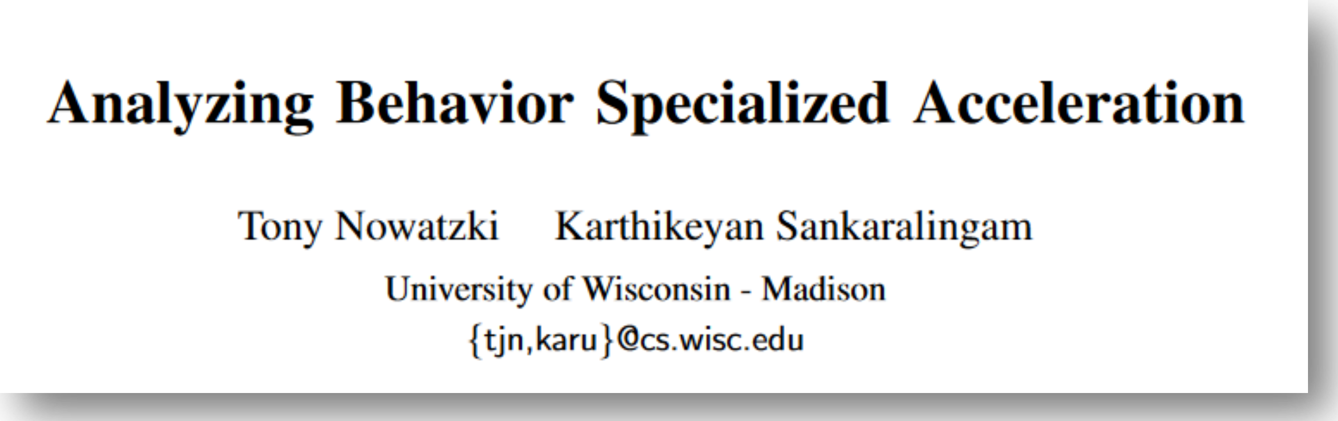
Mingyu Gao[†] Christina Delimitrou[†¶] Dimin Niu[§] Krishna T. Malladi[§] Hongzhong Zheng[§]
Bob Brennan[§] Christos Kozyrakis[†‡]

Stanford University[†] Samsung Semiconductor Inc.[§] Cornell University[¶] EPFL[‡]

{mgao12, cdel, kozyraki}@stanford.edu
{dimin.niu, k.tej, hz.zheng, bob.brennan}@ssi.samsung.com

- HPCA and ISCA 2016
- Various studies about **improving accelerator reusability and optimization** techniques

Related Work

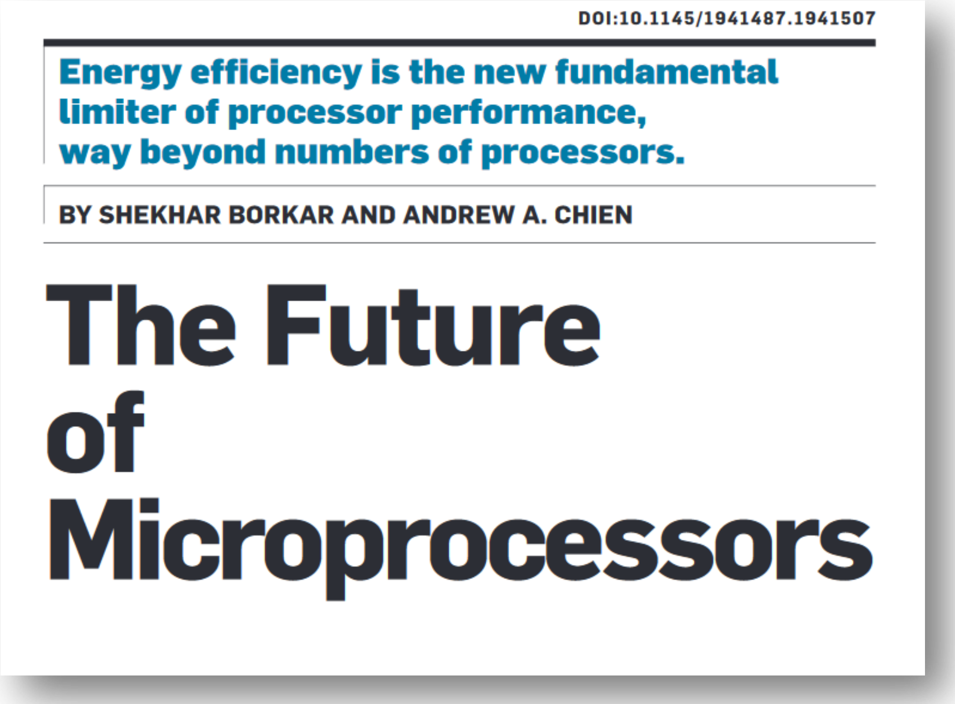


Analyzing Behavior Specialized Acceleration

Tony Nowatzki Karthikeyan Sankaralingam
University of Wisconsin - Madison
{tjn,karu}@cs.wisc.edu

- ASPLOS 2016
- Accelerator **modelling using dependence graphs**

Related Work



DOI:10.1145/1941487.1941507

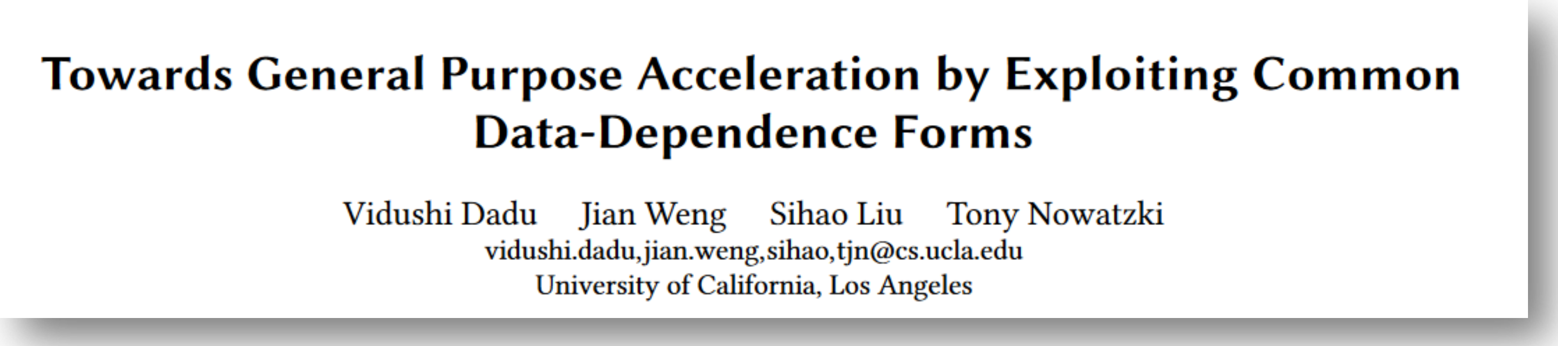
Energy efficiency is the new fundamental limiter of processor performance, way beyond numbers of processors.

BY SHEKHAR BORKAR AND ANDREW A. CHIEN

The Future of Microprocessors

- Article in: *Communications of the ACM, 2011*
- **Decouple chip and application performance** to estimate impact of microarchitecture on general-purpose microprocessors

Related Work – cited by



Towards General Purpose Acceleration by Exploiting Common Data-Dependence Forms

Vidushi Dadu Jian Weng Sihao Liu Tony Nowatzki
vidushi.dadu,jian.weng,sihao,tjn@cs.ucla.edu
University of California, Los Angeles

- MICRO 2019
- Increase performance by **accelerating common data dependency patterns**

Discussion

- What‘s your impression on the CSR metric?
  - Do you think it is a useful and sensible abstraction?
  - Can you think of a better way to abstract gains of different optimization layers
- What‘s your impression on the active transistor count model and physical gain model?
  - Realistic, what is missing?
  - Can we just assume perfect scaling for abstraction purposes?
  - Can we use machine learning to predict performance metrics and transistor counts?
- Do you think we will hit an accelerator wall?

Discussion

- How far will the use of accelerators go in the future?
  - Will GP-CPUs go away?
  - What implications for system architecture does high accelerator usage bring?
- The paper shows an accelerator wall for a few specific application domains, what other important domains can you think of?
  - Do the paper‘s assumptions hold there as well?
- Large ML chip in introductory lecture, what‘s the paper‘s answer to chip size scaling?
- What about completely new physical technologies
  - Compound Semiconductors
  - Quantum Computing
  - Graphene and Carbon Nanotubes

Thank you for your attention!

Moodle: http://bit.ly/accelerator-wall

Related Work

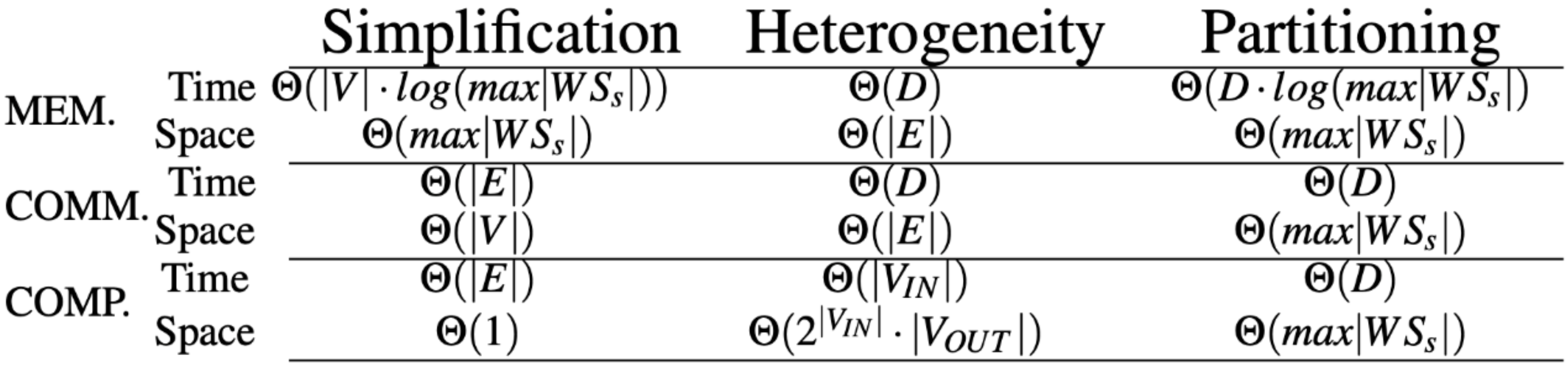
- D. W. Wall, “Limits of instruction-level parallelism,” in *Intl. Conf. on Arch. Support for Programming Languages & Operating Systems (ASPLOS)*, pp. 176–188, ACM, 1991.

- Limits of exploiting instruction level parallelism

Related Work

- A. Arunkumar, E. Bolotin, B. Cho, U. Milic, E. Ebrahimi, O. Villa, *et al.*, "**MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability**," in *Intl. Symp. on Computer Architecture (ISCA)*, pp. 320–332, ACM, 2017
- Scale beyond monolithic GPUs performance by putting several GPU cores modules on a single die

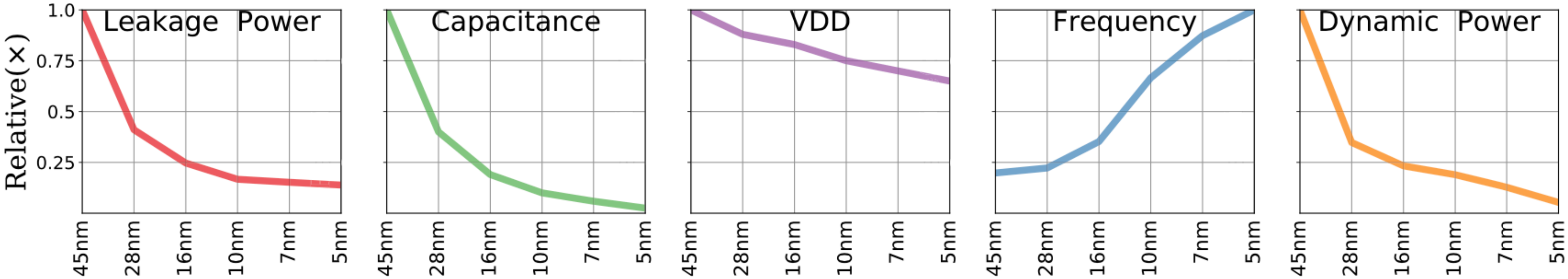
Chip Specialization – Dataflow and Tight Bounds



		Simplification	Heterogeneity	Partitioning
MEM.	Time	$\Theta(\|V\| \cdot log(max\|WS_s\|))$	$\Theta(D)$	$\Theta(D \cdot log(max\|WS_s\|)$
	Space	$\Theta(max\|WS_s\|)$	$\Theta(\|E\|)$	$\Theta(max\|WS_s\|)$
COMM.	Time	$\Theta(\|E\|)$	$\Theta(D)$	$\Theta(D)$
	Space	$\Theta(\|V\|)$	$\Theta(\|E\|)$	$\Theta(max\|WS_s\|)$
COMP.	Time	$\Theta(\|E\|)$	$\Theta(\|V_{IN}\|)$	$\Theta(D)$
	Space	$\Theta(1)$	$\Theta(2^{\|V_{IN}\|} \cdot \|V_{OUT}\|)$	$\Theta(max\|WS_s\|)$

Table II: Summary of Time and Space Complexity Limits for Chip Specialization Concepts, in Terms of DFG Definitions.

Physical Transistor Model – Device Scaling

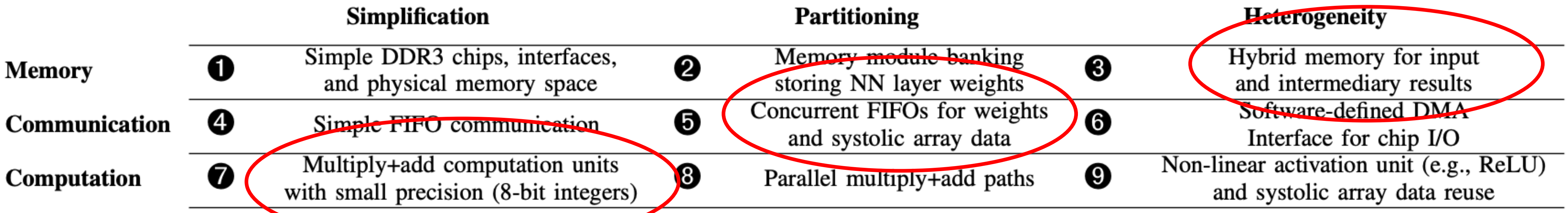


- Device Scaling Models from [20-22]

Chip Specialization - Limitations

- CMOS scaling **ends at 5nm**
- **Fixed # active transistors**
- But we can still be smart right?! (Alg, Fwk, Plt, Eng)
  - Alg, Fwk: Fixed application domain
    - Limited solution space
    - Often already quite exhausted
  - Plt, Eng: Limited ways to map problems to silicon
    - Upper bounds given by abstracted dataflow

Chip Specialization – Concepts



	Simplification	Partitioning	Heterogeneity
Memory	❶ Simple DDR3 chips, interfaces, and physical memory space	❷ Memory module banking storing NN layer weights	❸ Hybrid memory for input and intermediary results
Communication	❹ Simple FIFO communication	❺ Concurrent FIFOs for weights and systolic array data	❻ Software-defined DMA Interface for chip I/O
Computation	❼ Multiply+add computation units with small precision (8-bit integers)	❽ Parallel multiply+add paths	❾ Non-linear activation unit (e.g., ReLU) and systolic array data reuse

Table I: Chip Specialization Concepts. Examples From a TPU ASIC Chip.

- **Simplification:** reduce functionality and simplify datapaths
- **Partitioning:** exploit parallelism
- **Heterogenity:** tailor to application patterns

Chip Specialization – Dataflow and Tight Bounds

- **Simplification:** reduce functionality and simplify datapaths
- **Partitioning:** exploit parallelism
- **Heterogenity:** tailor to application patterns

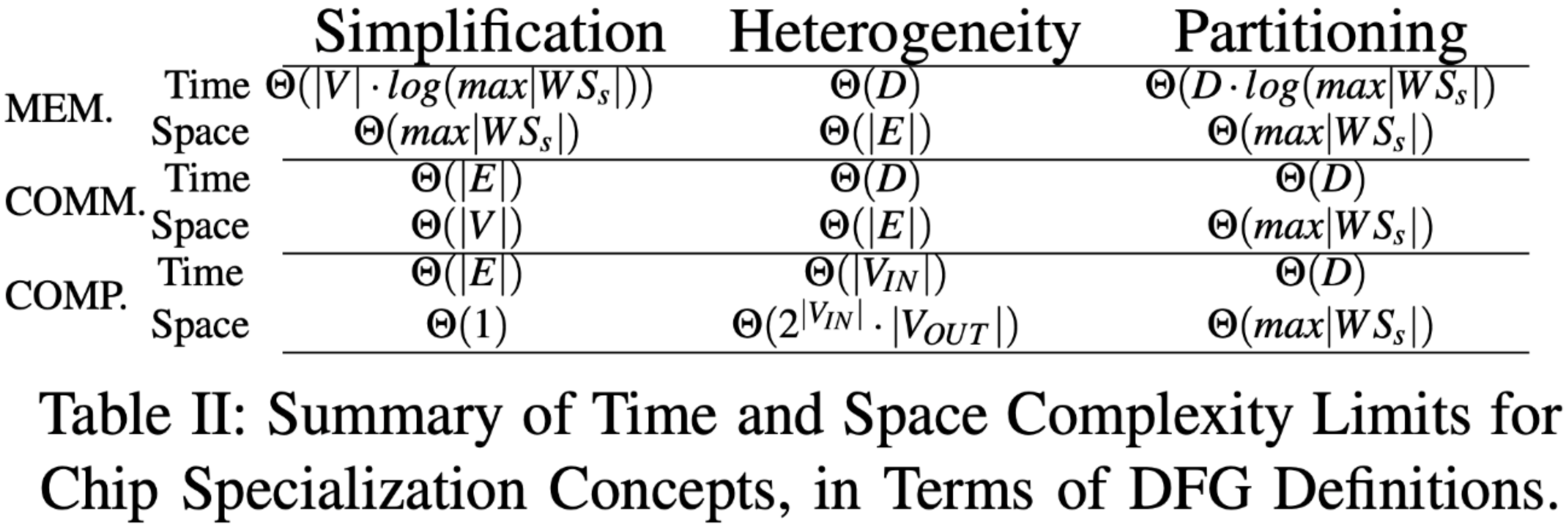
- **Model application** in dataflow graph
- **Couple # transistors** and dataflow graph

Chip Specialization – Dataflow and Tight Bounds

- Space, Simplified:
  - $\Theta(1)$ all mathematical ops, constant number of gates
- Time, Simplified:
  - $\Theta(E)$ computation limited by number of edges in dataflow
- Space, Heterogenity:
  - $\Theta(2^{|Vin|} \text{ x } |V_{out}|)$ lookup table
- Time, Heterogenity:
  - $\Theta(|V_{in}|)$ read in input

- **Chip specialization is coupled to the # transistors**

Chip Specialization – Dataflow and Tight Bounds



		Simplification	Heterogeneity	Partitioning
MEM.	Time	$\Theta(\vert V\vert \cdot log(max\vert WS_s\vert))$	$\Theta(D)$	$\Theta(D \cdot log(max\vert WS_s\vert)$
	Space	$\Theta(max\vert WS_s\vert)$	$\Theta(\vert E\vert)$	$\Theta(max\vert WS_s\vert)$
COMM.	Time	$\Theta(\vert E\vert)$	$\Theta(D)$	$\Theta(D)$
	Space	$\Theta(\vert V\vert)$	$\Theta(\vert E\vert)$	$\Theta(max\vert WS_s\vert)$
COMP.	Time	$\Theta(\vert E\vert)$	$\Theta(\vert V_{IN}\vert)$	$\Theta(D)$
	Space	$\Theta(1)$	$\Theta(2^{\vert V_{IN}\vert} \cdot \vert V_{OUT}\vert)$	$\Theta(max\vert WS_s\vert)$

Table II: Summary of Time and Space Complexity Limits for Chip Specialization Concepts, in Terms of DFG Definitions.

Chip Specialization – Accelerator Gains Bounds

- Aladdin: modelling tool for accelerator design
- Runtime vs. Power Efficiency
- Importance of CMOS technology

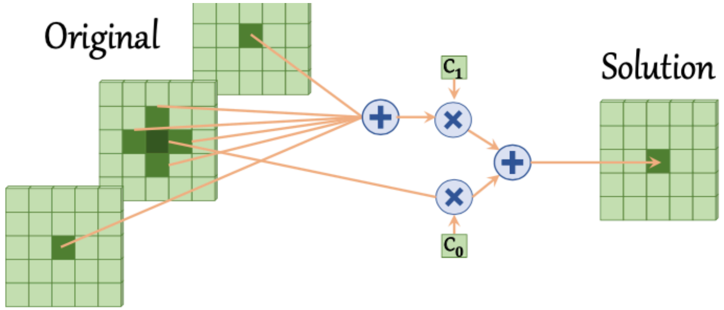


Figure 12: Visualization of a 3D Stencil Computation

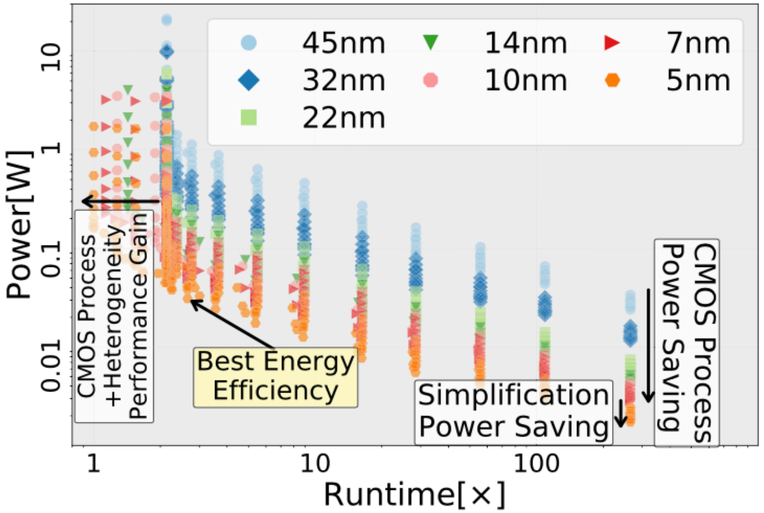


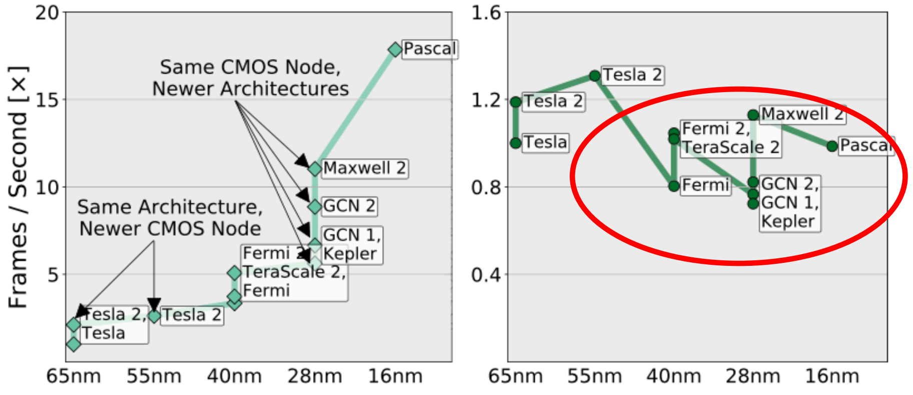
Figure 13: 3D Stencil Power, Timing, and CMOS Sweep. Arrows Highlight Optimal Point and Gain Sources.

Chip Specialization – Verdict

- Chip specialization performance gains are eventually coupled to CMOS scaling
  - Dataflow abstraction
  - Common optimization techniques
- Couples Gain(Plt) and Gain(Eng) to CMOS scaling

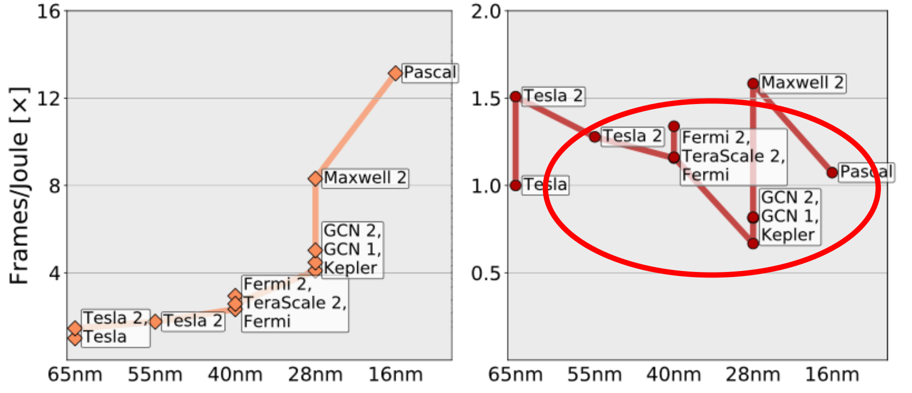
Case Studies – GPUs for graphics

- Bad CSR on architecture changes
- Better CSR within same architecture



(a) Absolute (b) Chip Specialization Return

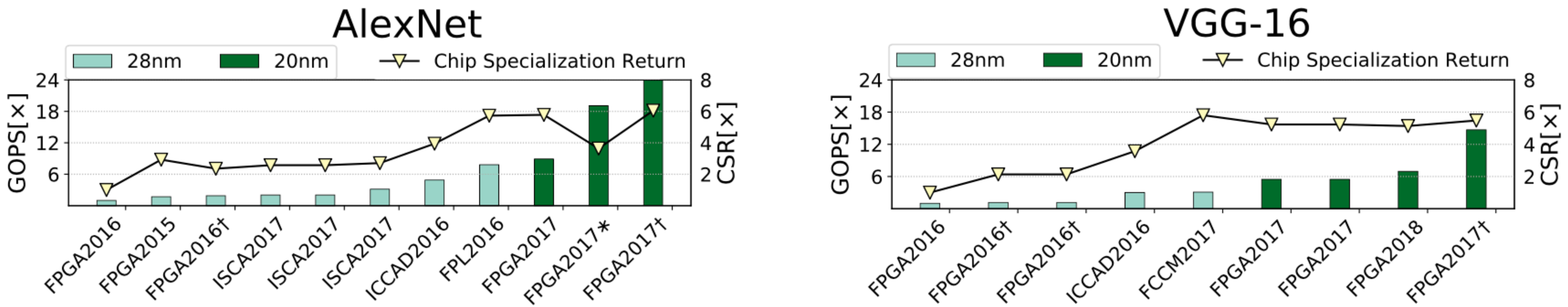
Figure 6: Architecture + CMOS Scaling: Throughput



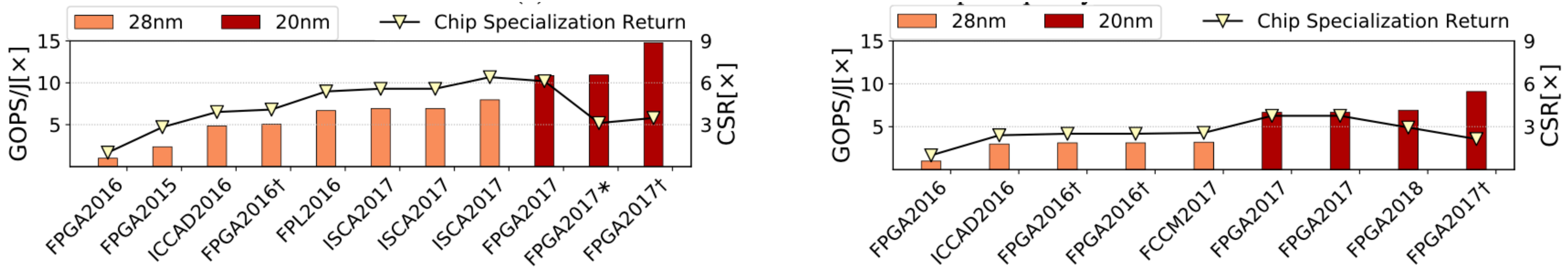
(a) Absolute (b) Chip Specialization Return

Figure 7: Architecture + CMOS Scaling: Energy Efficiency

Case Studies – FPGA for conv. neural nets



(a) Performance Scaling of FPGAs and Respective CMOS Nodes: Absolute and Chip Specialization Return

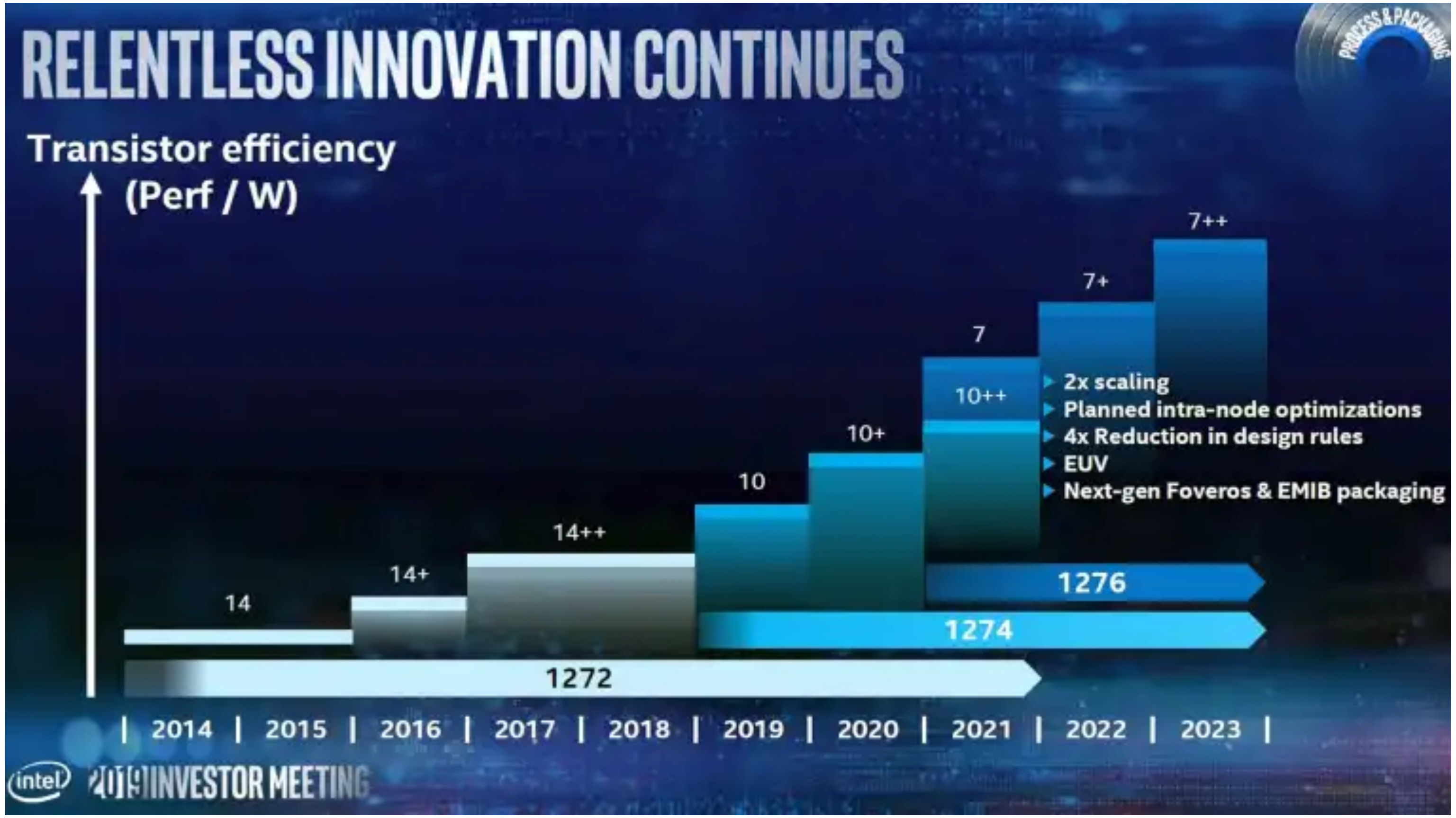


(c) Energy Efficiency Scaling of FPGAs and Respective CMOS Nodes: Absolute and Chip Specialization Return

Related Works

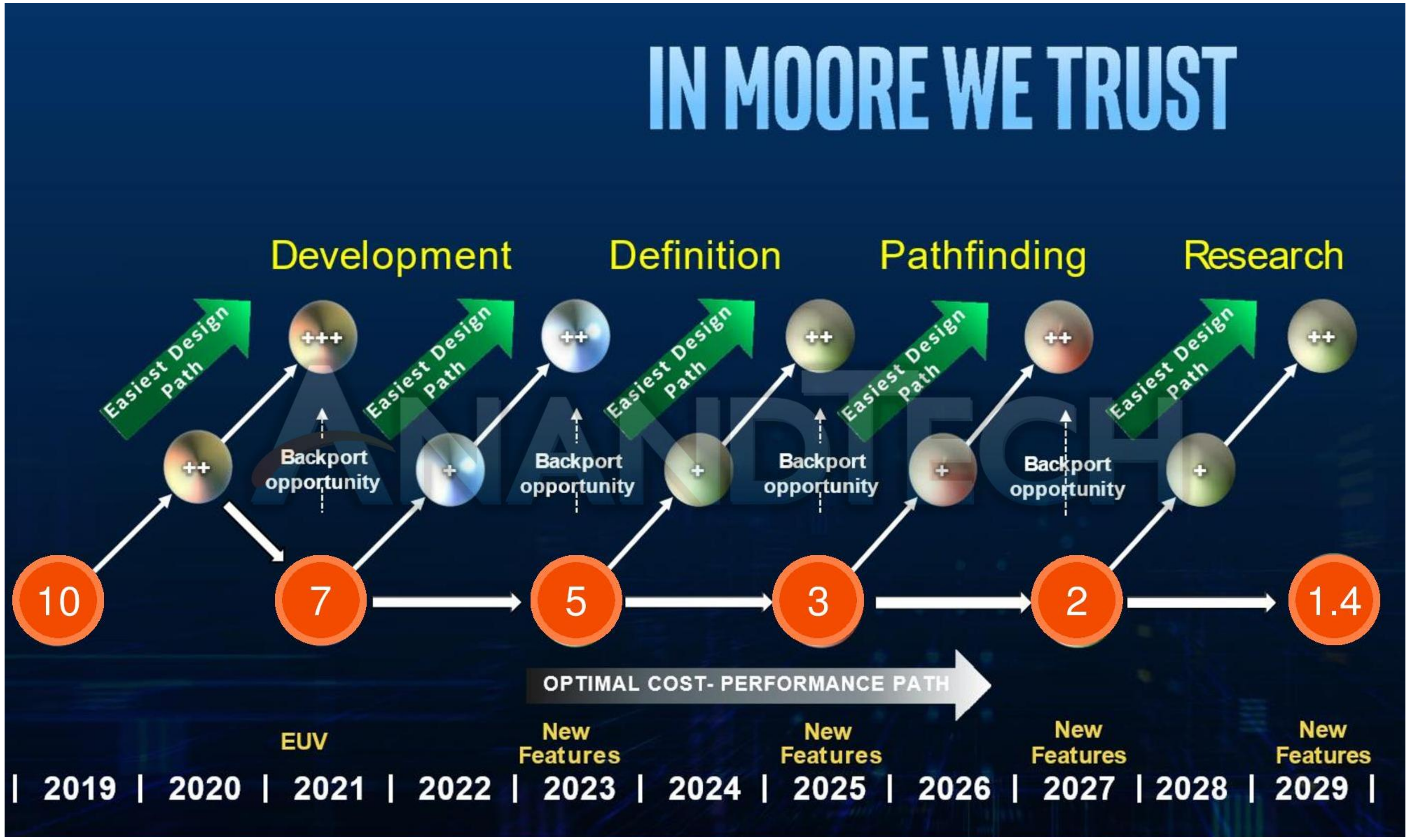
- A. Fuchs and D. Wentzlaff, “**Scaling datacenter accelerators with compute-reuse architectures**,” in *Intl. Symp. on Computer Architecture (ISCA)*, pp. 353–366, 2018.
- Intensive use of pre-computed results with new low energy non-volatile memory solutions in **accelerators (memoization)**

Intel Roadmap for CMOS architectures



https://www.heise.de/newsticker/meldung/Intel-plant-7-nm-Chips-ab-2021-4418708.html

Intel Roadmap for CMOS architectures



https://www.anandtech.com/show/15217/intels-manufacturing-roadmap-from-2019-to-2029