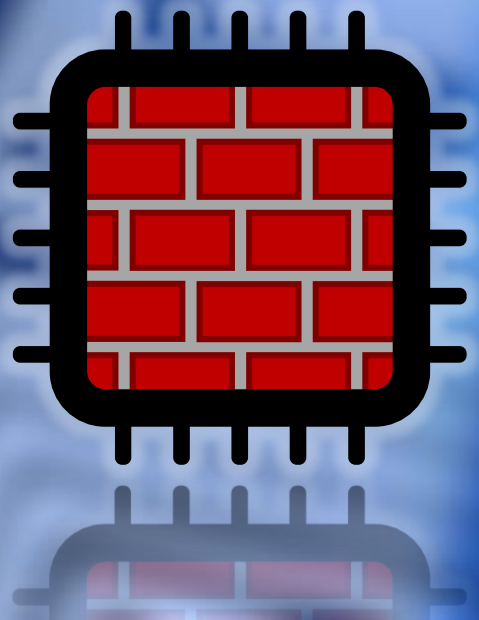
The Accelerator Wall:
Limits of Chip Specialization

Adi Fuchs and David Wentzlaff

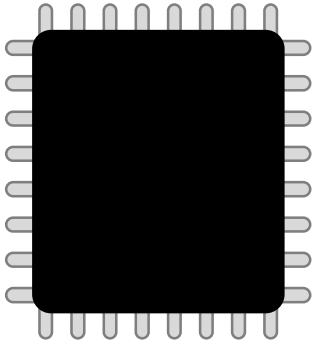
HPCA 2019 - Feb. 18, 2018



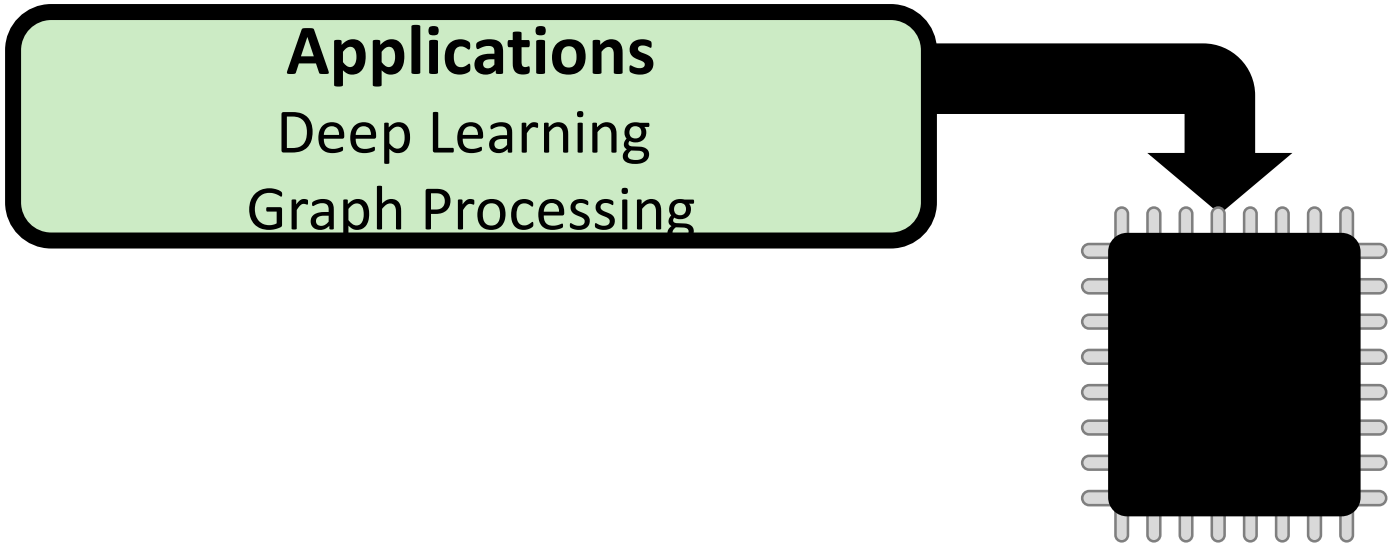
PRINCETON UNIVERSITY

Source: Dreamstime.com

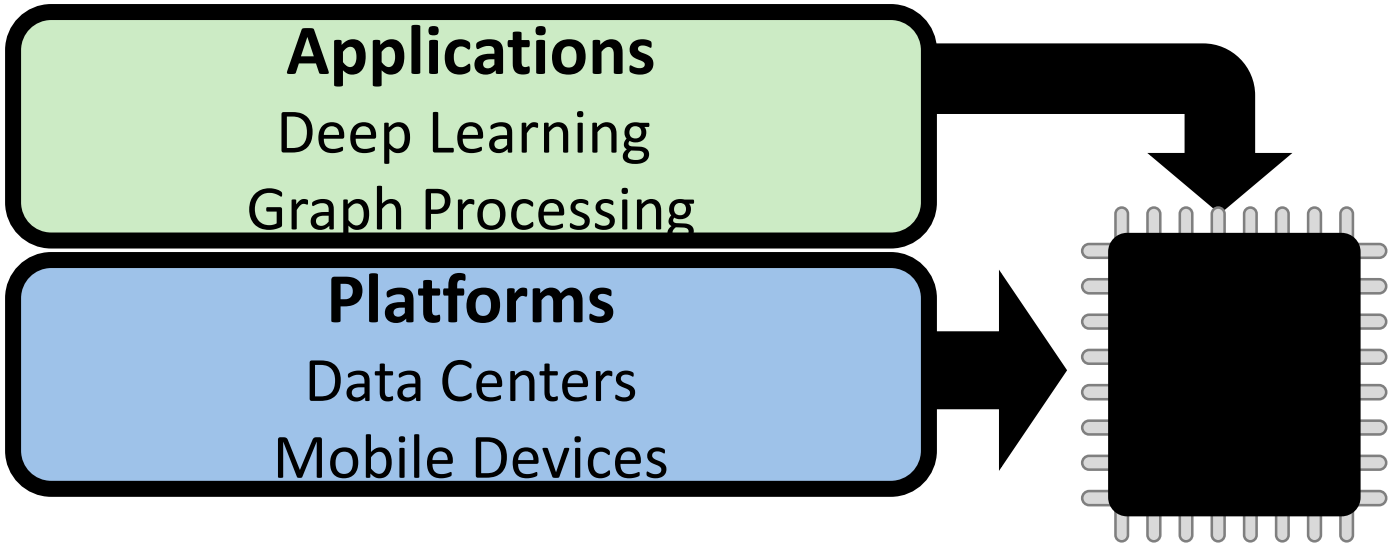
Background



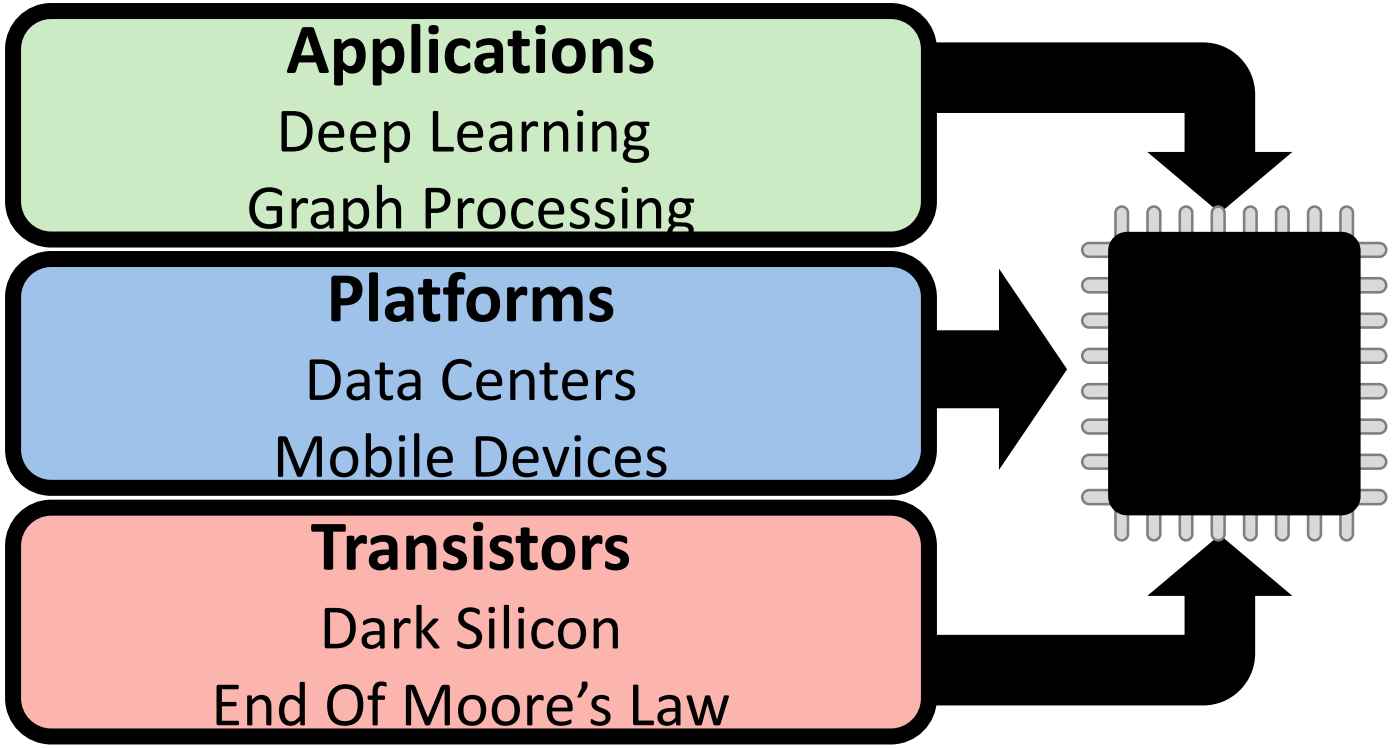
Background



Background


Applications
Deep Learning
Graph Processing
Platforms
Data Centers
Mobile Devices

Background

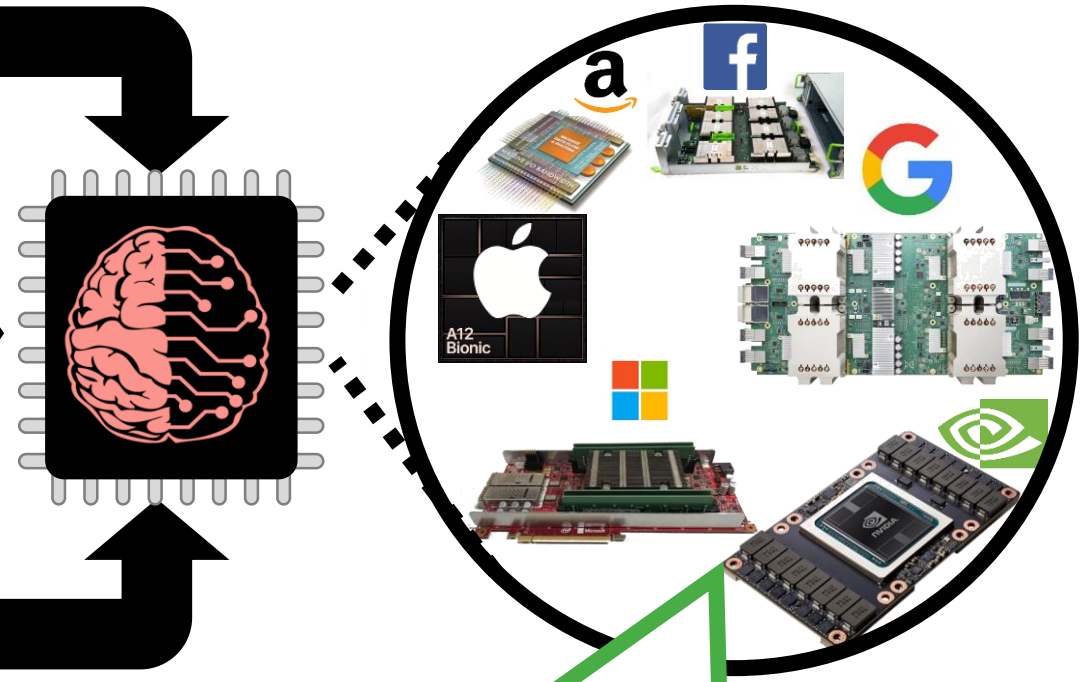

Applications
Deep Learning
Graph Processing
Platforms
Data Centers
Mobile Devices
Transistors
Dark Silicon
End Of Moore's Law

Background

Applications
Deep Learning
Graph Processing

Platforms
Data Centers
Mobile Devices

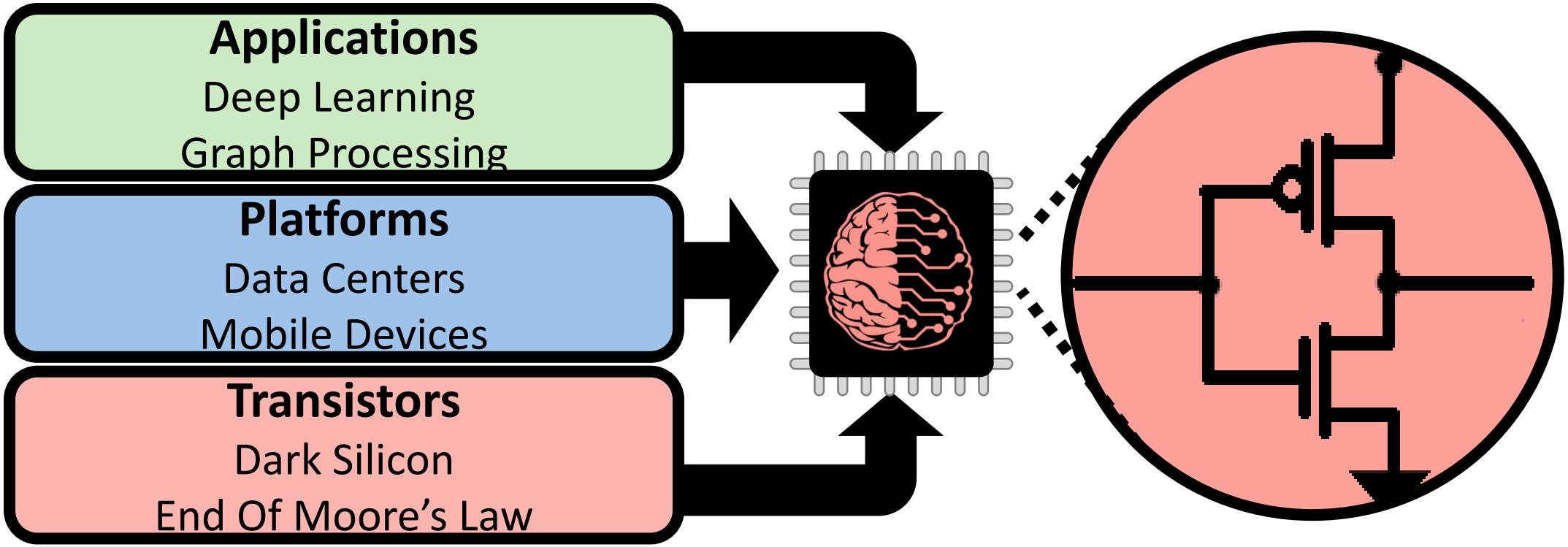
Transistors
Dark Silicon
End Of Moore's Law



Transistors Aren't Improving? We'll Use the Ones We Have Better!

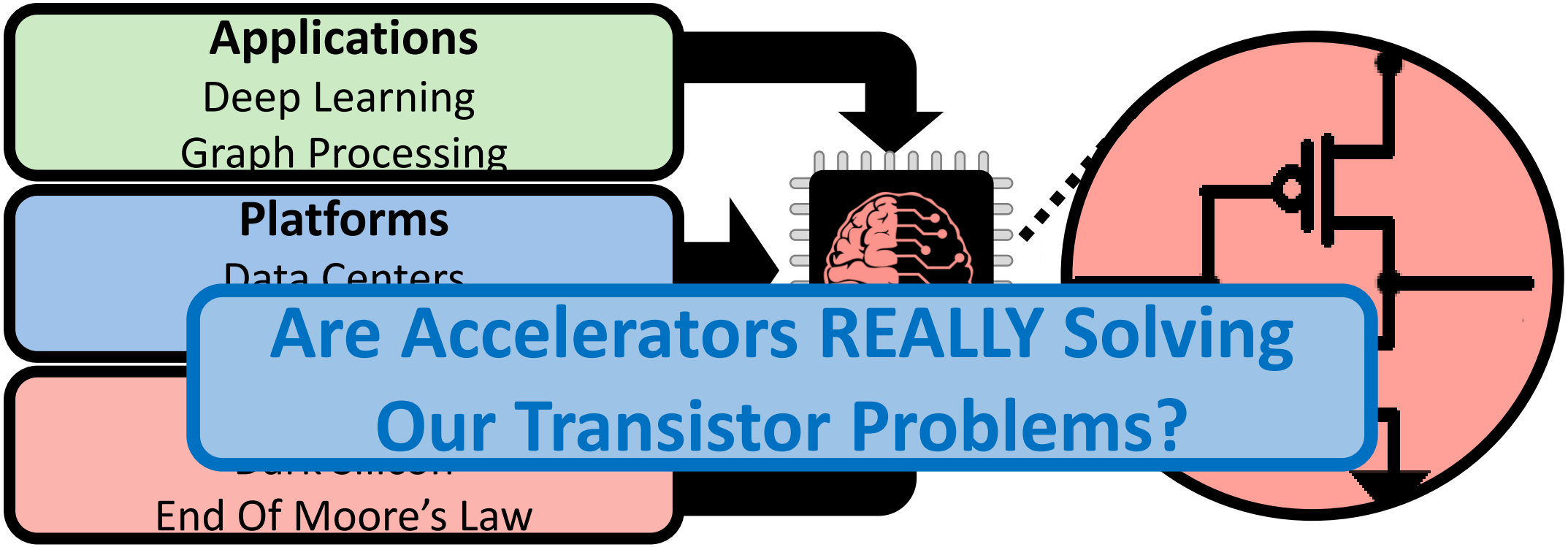
Sources:
"Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective", Hazelwood et al. HI
"Cloud TPU", Google
"FPGA Accelerated Computing Using AWS F1 Instances", David Pellerin, AWS summit 2017
"IPhone XS A12 Bionic", Apple https://www.apple.com/iphone-xs/a12-bionic/
"Microsoft unveils Project Brainwave for real-time AI", Doug Burger
"NVIDIA TESLA V100", NVIDIA

Background



Sources:
"Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective", Hazelwood et al. HPCA 2018
"Cloud TPU", Google
"FPGA Accelerated Computing Using AWS F1 Instances", David Pellerin, AWS summit 2017
"IPhone XS A12 Bionic", Apple https://www.apple.com/iphone-xs/a12-bionic/
"Microsoft unveils Project Brainwave for real-time AI", Doug Burger
"NVIDIA TESLA V100", NVIDIA

Background



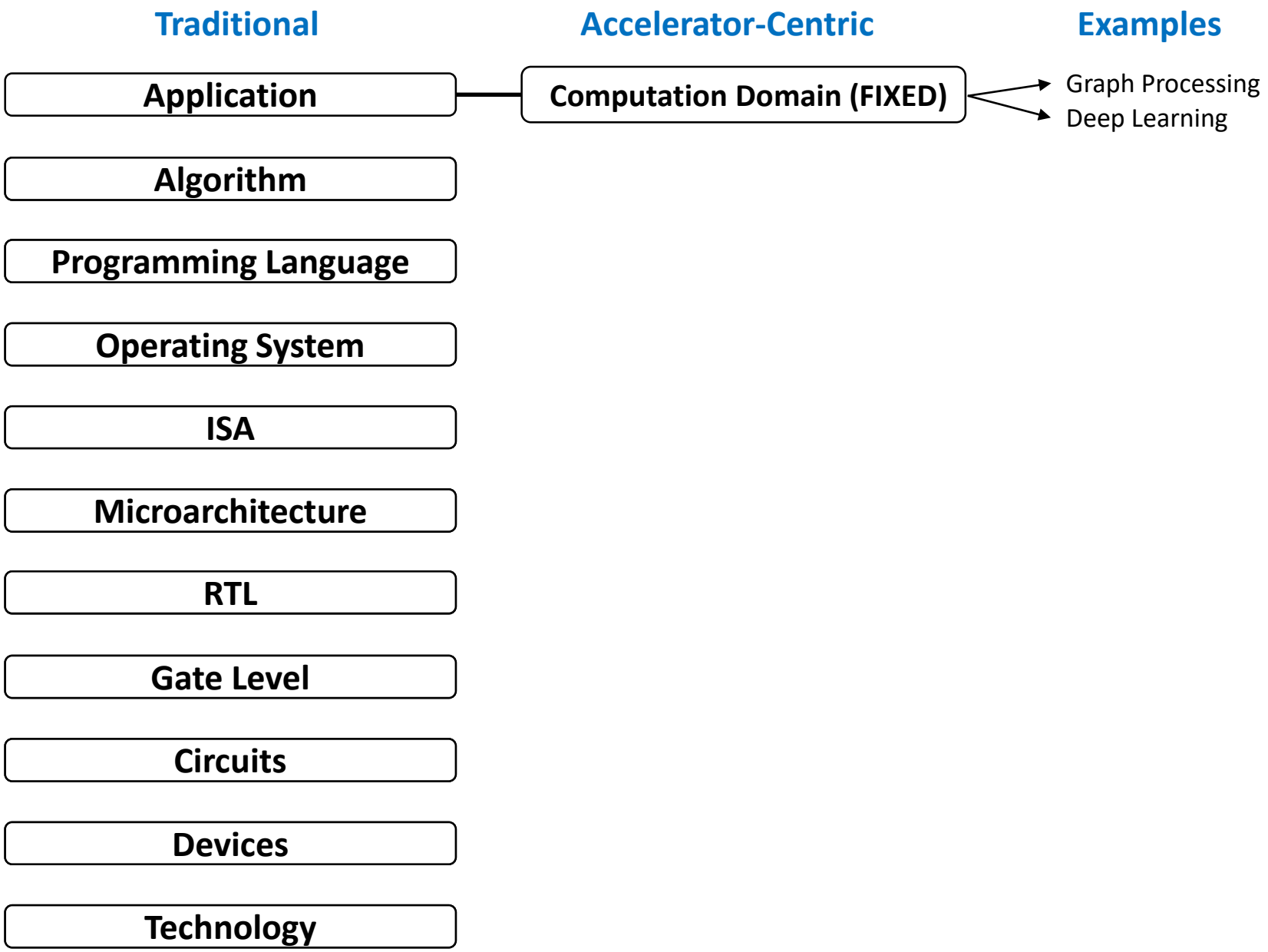
Sources:
"Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective", Hazelwood et al. HPCA 2018
"Cloud TPU", Google
"FPGA Accelerated Computing Using AWS F1 Instances", David Pellerin, AWS summit 2017
"IPhone XS A12 Bionic", Apple https://www.apple.com/iphone-xs/a12-bionic/
"Microsoft unveils Project Brainwave for real-time AI", Doug Burger
"NVIDIA TESLA V100", NVIDIA

Abstraction Layers: The Accelerator Viewpoint

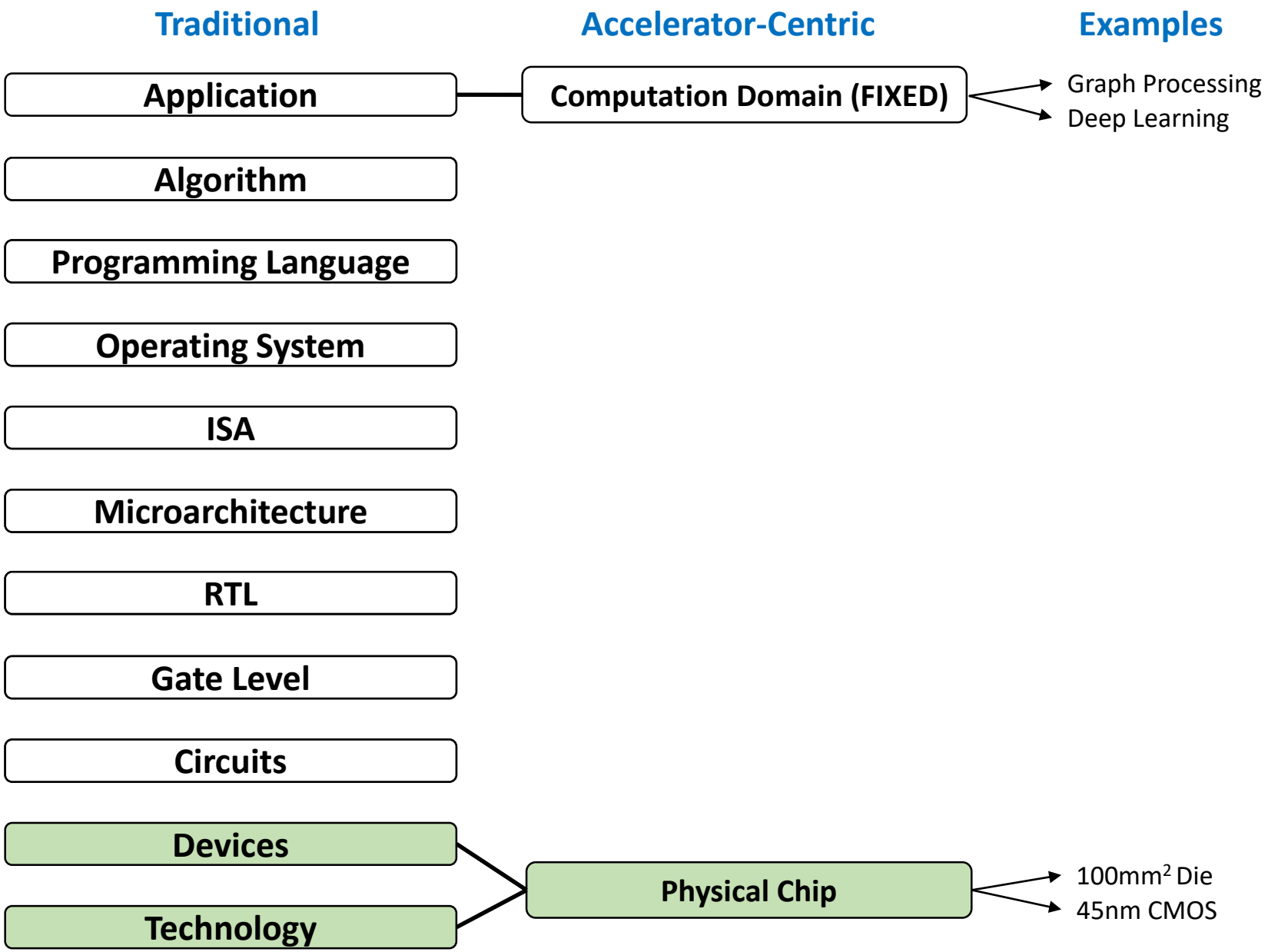
Traditional

- Application
- Algorithm
- Programming Language
- Operating System
- ISA
- Microarchitecture
- RTL
- Gate Level
- Circuits
- Devices
- Technology

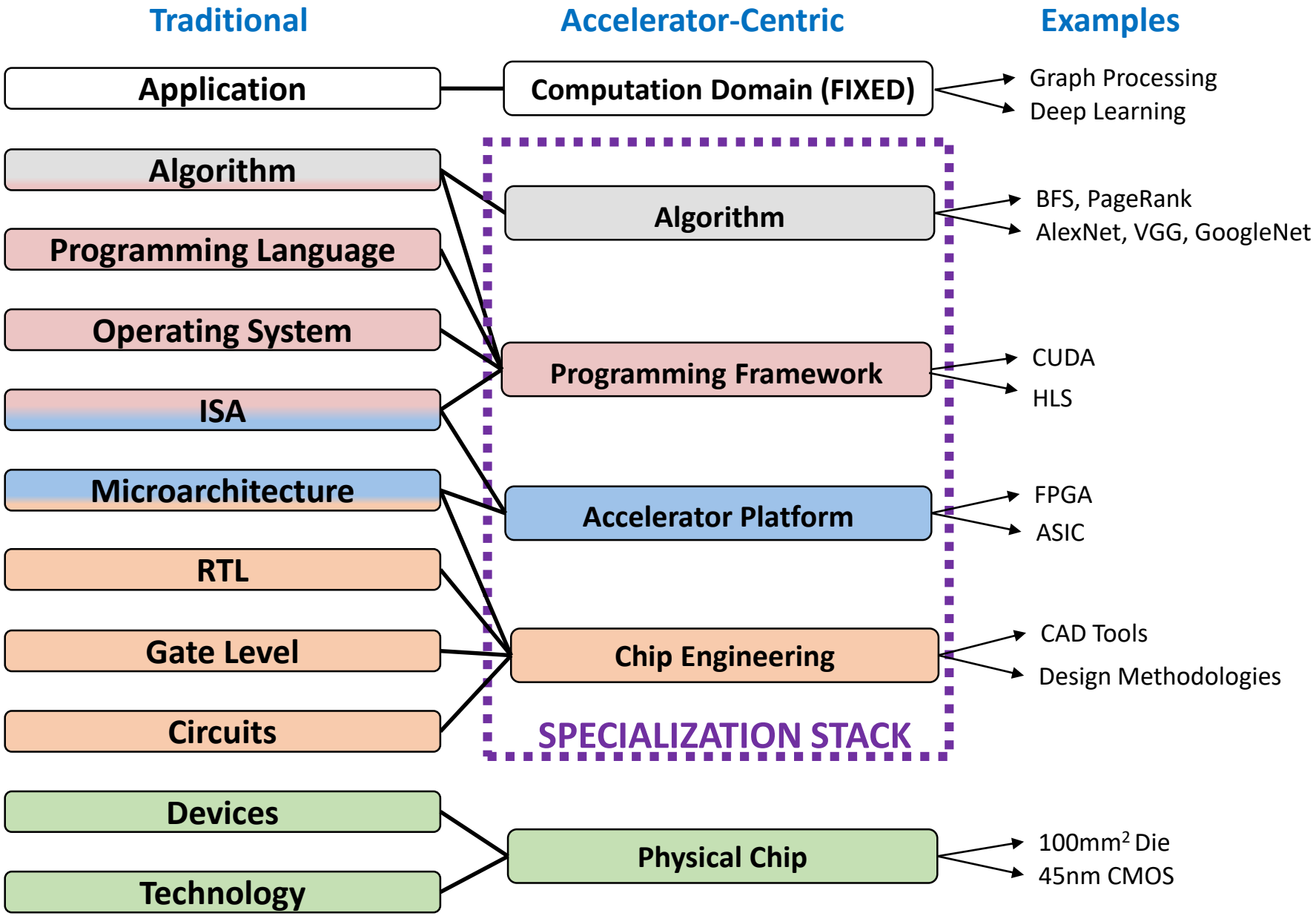
Abstraction Layers: The Accelerator Viewpoint



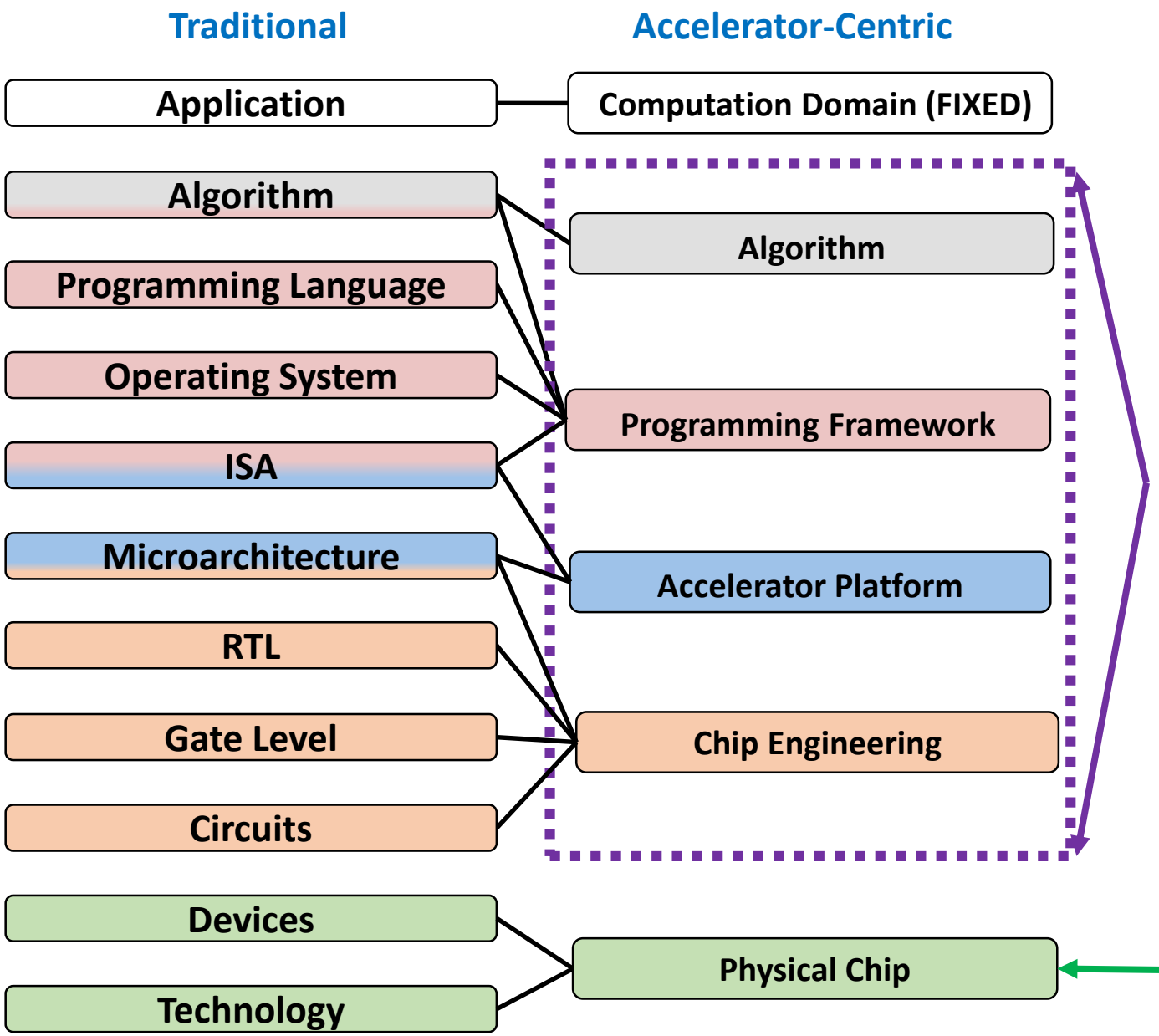
Abstraction Layers: The Accelerator Viewpoint



Abstraction Layers: The Accelerator Viewpoint



Abstraction Layers: The Accelerator Viewpoint



Are Accelerators Driven by:

SPECIALIZATION?

OR

TRANSISTORS?

Accelerator Evaluation

- **Motivation (Toy Example):**

	Accelerator A		Accelerator B
Application Throughput	10^6 op/s	Better →	2×10^6 op/s

Accelerator Evaluation

- **Motivation (Toy Example):**

	Accelerator A		Accelerator B
Application Throughput	10^6 op/s	Better →	2×10^6 op/s
Chip Transistors	1 Million		4 Million

Accelerator Evaluation

- **Motivation (Toy Example):**

	Accelerator A		Accelerator B
Application Throughput	10^6 op/s	Better →	2×10^6 op/s
Chip Transistors	1 Million		4 Million
Throughput "Per Silicon"	1	← Better	0.5

Accelerator Evaluation

- **Motivation (Toy Example):**

	Accelerator A		Accelerator B
Application Throughput	10^6 op/s	Better →	2×10^6 op/s
Chip Transistors	1 Million		4 Million
Throughput "Per Silicon"	1	← Better	0.5

VS.

Sources:
lamborghini.com

Accelerator Evaluation

- **Motivation (Toy Example):**

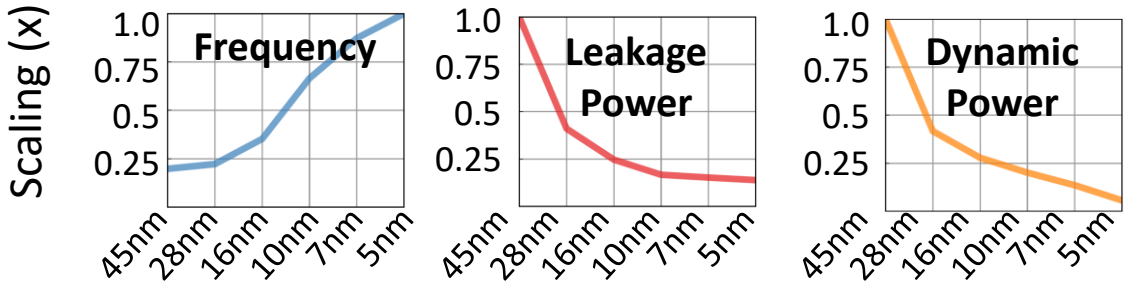
	Accelerator A		Accelerator B
Application Throughput	10^6 op/s	Better	2×10^6 op/s
Chip Transistors	1 Million		4 Million
Th			

IDEA: Build a Model for CMOS POTENTIAL To Account for CMOS Scaling Contributions

VS.

Sources:
lamborghini.com

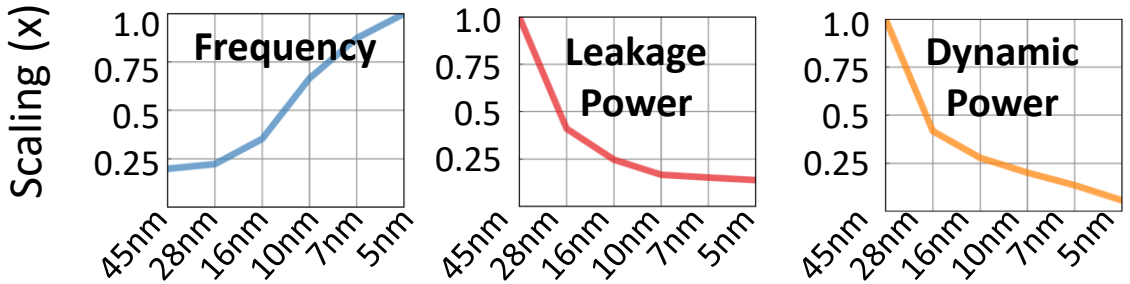
CMOS Potential Model



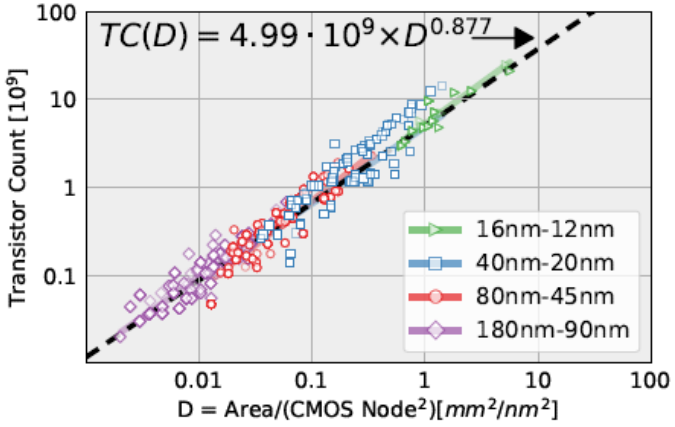
Sources:
"Scaling equations for the accurate prediction of CMOS device performance from 180 nm to 7 nm", Stillmaker and Baas, VLSI Journal 2017
"International Technology Roadmap For Semiconductors (ITRS) 2015 Edition", ITRS 2015 "International Roadmap For devices and systems (IRDS) 2017 Edition", IRDS 2017
www.techpowerup.com/cpudb, www.techpowerup.com/gpudb, cpudb.stanford.edu

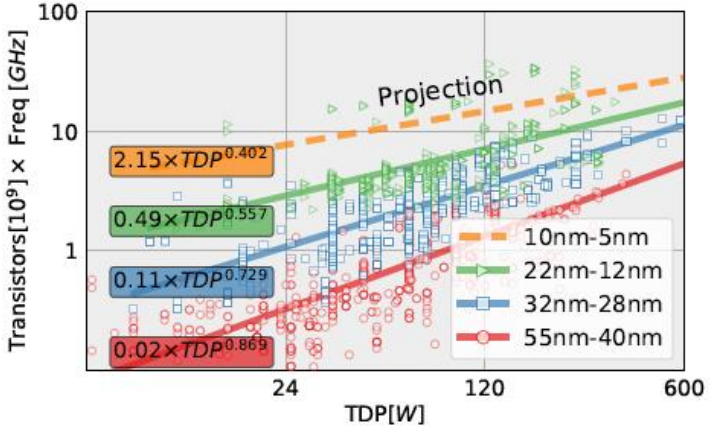
CMOS Potential Model

- Device-Level Scaling: Using CMOS Scaling study + ITRS/IRDS Projections



- Chip-Transistor Budget: Using Datasheets of **Thousands of Commercial Processors**
  - Model How Many Fit a **Chip Die** (Given Area + CMOS Node) and **Power Envelope** (Given TDP, Frequency etc.)

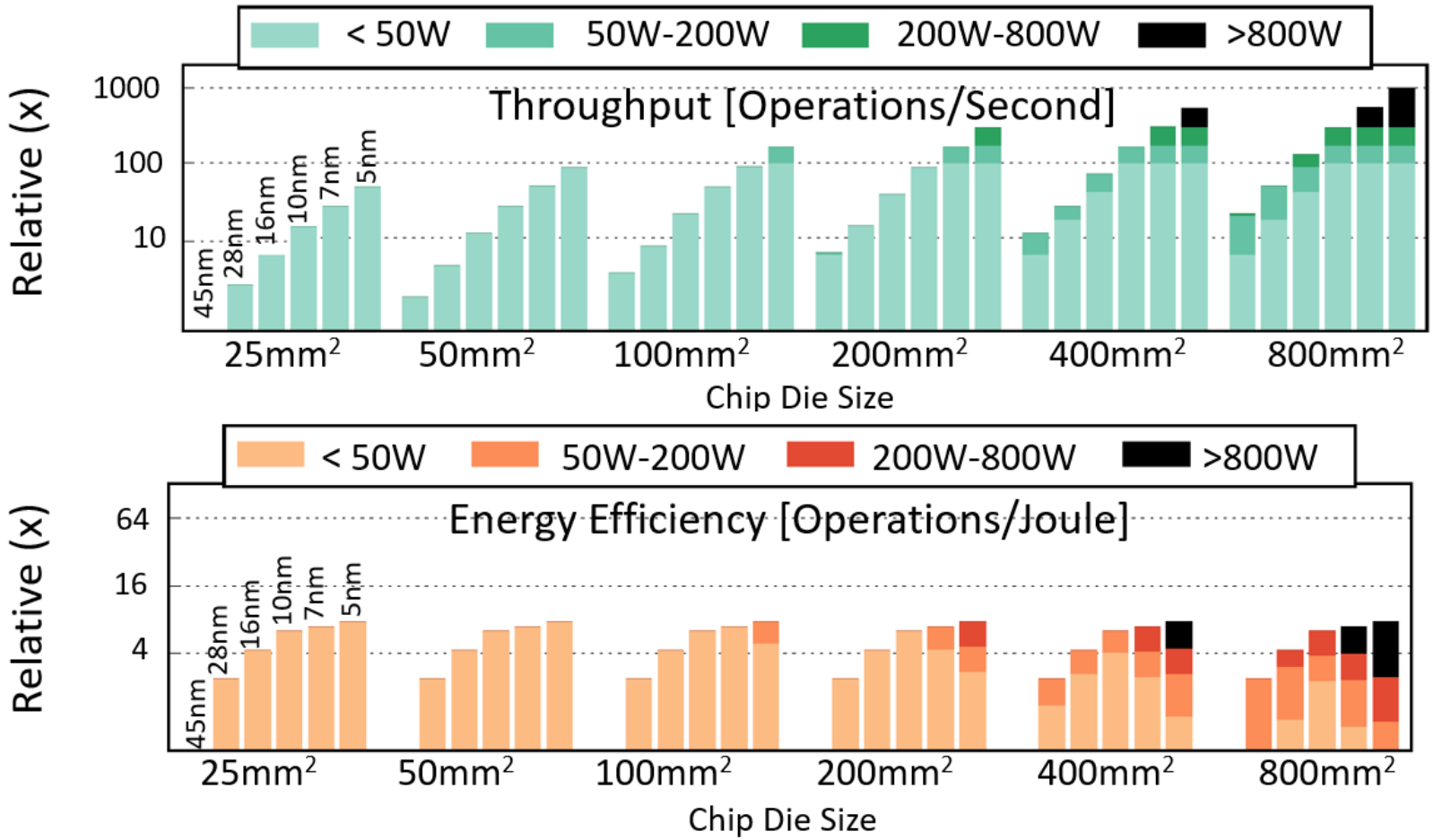




Sources:
"Scaling equations for the accurate prediction of CMOS device performance from 180 nm to 7 nm", Stillmaker and Baas, VLSI Journal 2017
"International Technology Roadmap For Semiconductors (ITRS) 2015 Edition", ITRS 2015 "International Roadmap For devices and systems (IRDS) 2017 Edition", IRDS 2017
www.techpowerup.com/cpudb, www.techpowerup.com/gpudb, cpudb.stanford.edu

CMOS Potential Model

- Integrate Device Scaling + Chip Budget Models To Build **CMOS Potential Functions.** For Example:



CMOS-Level Throughput and Energy Efficiency. Normalized to 45nm CMOS 25mm² chips

Accelerator Evaluation TAKE #2

- **Formal Definition:**

$$Gain = \frac{Gain}{CMOS\ Potential} \times CMOS\ Potential$$

Accelerator Evaluation TAKE #2

- **Formal Definition:**

$$Gain = \boxed{\frac{Gain}{CMOS\ Potential}} \times CMOS\ Potential$$

Chip Specialization Return (CSR)

Accelerator Evaluation TAKE #2

- **Formal Definition:**

$$Gain = \boxed{\frac{Gain}{CMOS\ Potential}} \times CMOS\ Potential$$

Chip Specialization Return (CSR)

- **Comparing Accelerators:**

$$\frac{Gain_{ACCELERATOR\ B}}{Gain_{ACCELERATOR\ A}} = \frac{CSR_B}{CSR_A} \times \frac{CMOS\ Potential_B}{CMOS\ Potential_A}$$

Accelerator Evaluation TAKE #2

- **Formal Definition:**

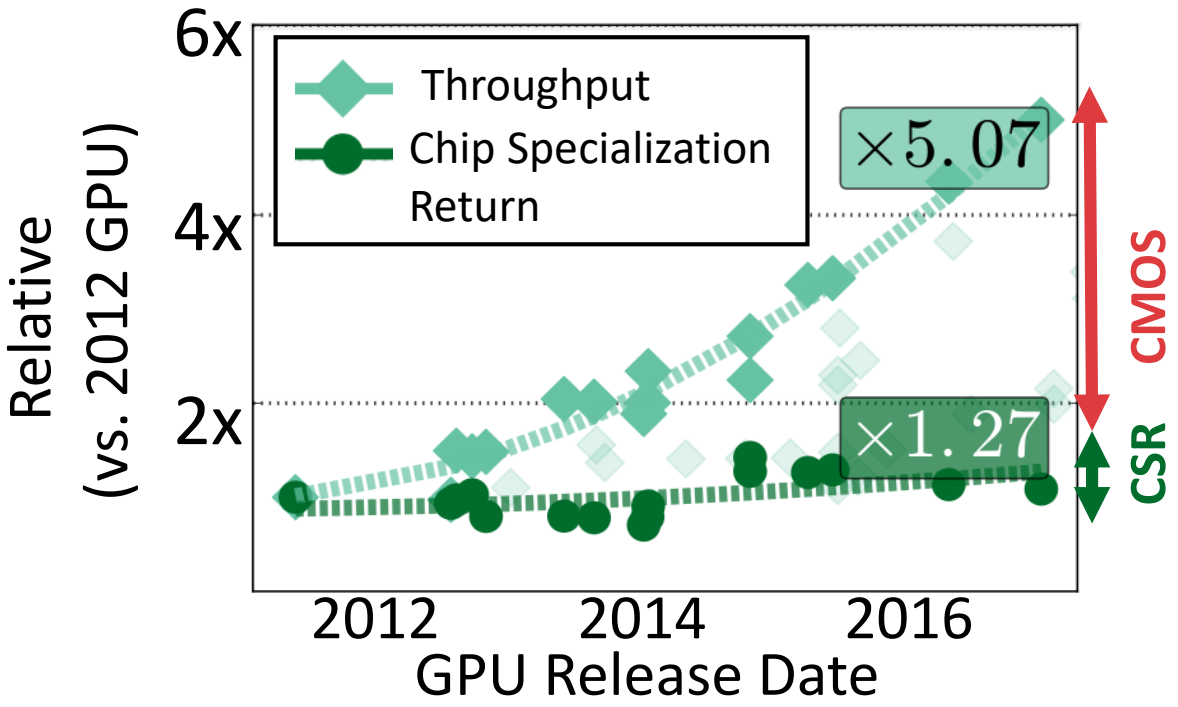
$$Gain = \boxed{\frac{Gain}{CMOS\ Potential}} \times CMOS\ Potential$$

Chip Specialization Return (CSR)

- **Comparing Accelerators:**

$$\frac{Gain_{ACCELERATOR\ B}}{Gain_{ACCELERATOR\ A}} = \frac{CSR_B}{CSR_A} \times \frac{CMOS\ Potential_B}{CMOS\ Potential_A}$$

- **Example:** Gaming Throughput on GPUs
  - **Throughput (Gain) Improvement: 5.07x**
  - **CMOS Scaling Contribution: 4x**
  - **Specialization Contribution: ONLY 1.27x**



Data Sources: www.techpowerup.com/gpudb
https://www.anandtech.com/bench/GPU18

Accelerator Evaluation TAKE #2

- **Formal Definition:**

$$Gain = \boxed{\frac{Gain}{CMOS\ Potential}} \times CMOS\ Potential$$

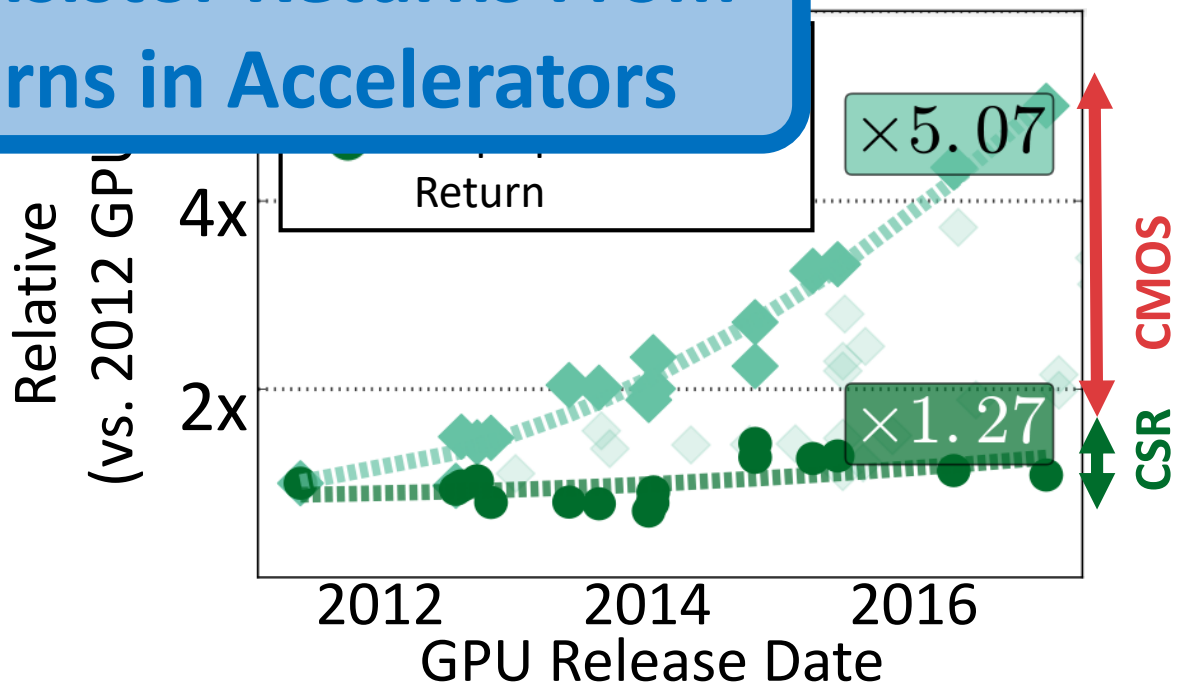
Chip Specialization Return (CSR)

- **Comparing Accelerators:**

$$\frac{Gain_{ACCELERATOR\ B}}{Gain_{ACCELERATOR\ A}} = \frac{CSR_B}{CSR_A} \times \frac{CMOS\ Potential_B}{CMOS\ Potential_A}$$

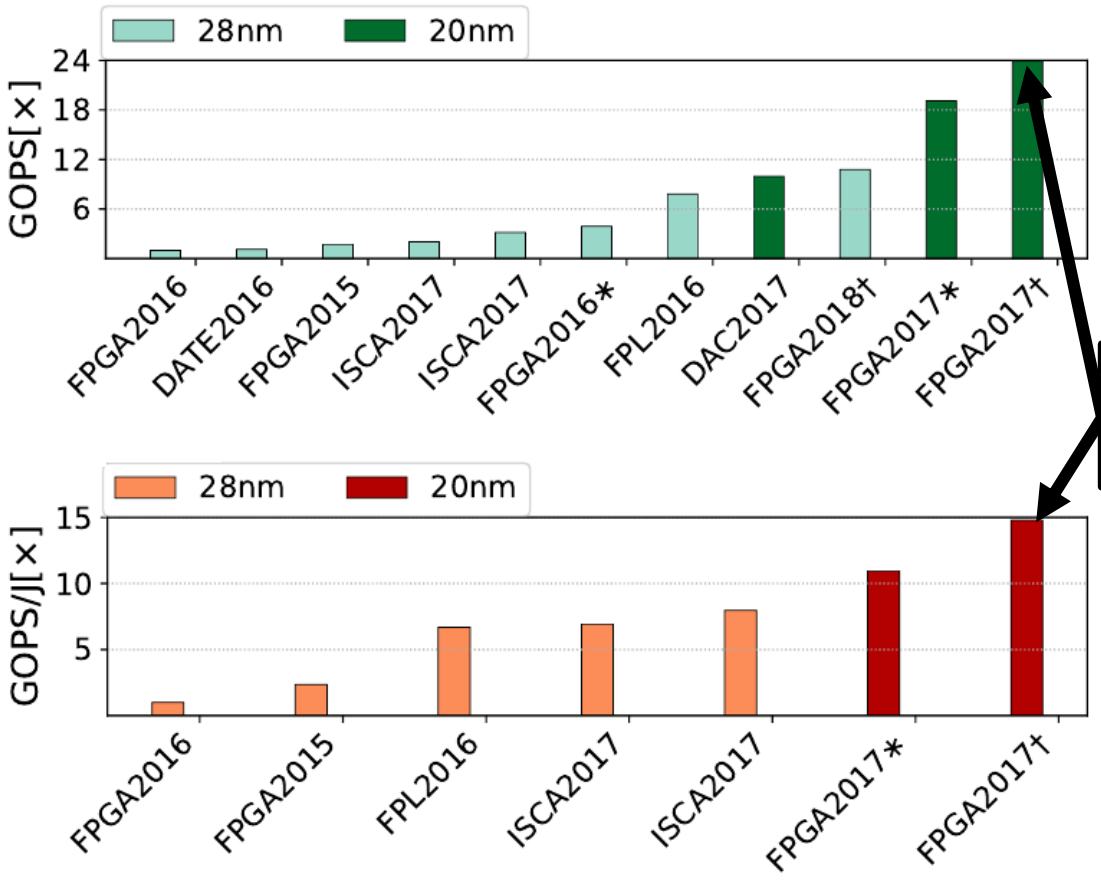
GOAL: Decouple Transistor Returns From Specialization Returns in Accelerators

- **Example:** Gaming Throughput on GPUs
  - **Throughput (Gain) Improvement: 5.07x**
  - **CMOS Scaling Contribution: 4x**
  - **Specialization Contribution: ONLY 1.27x**



Data Sources: www.techpowerup.com/gpudb
https://www.anandtech.com/bench/GPU18

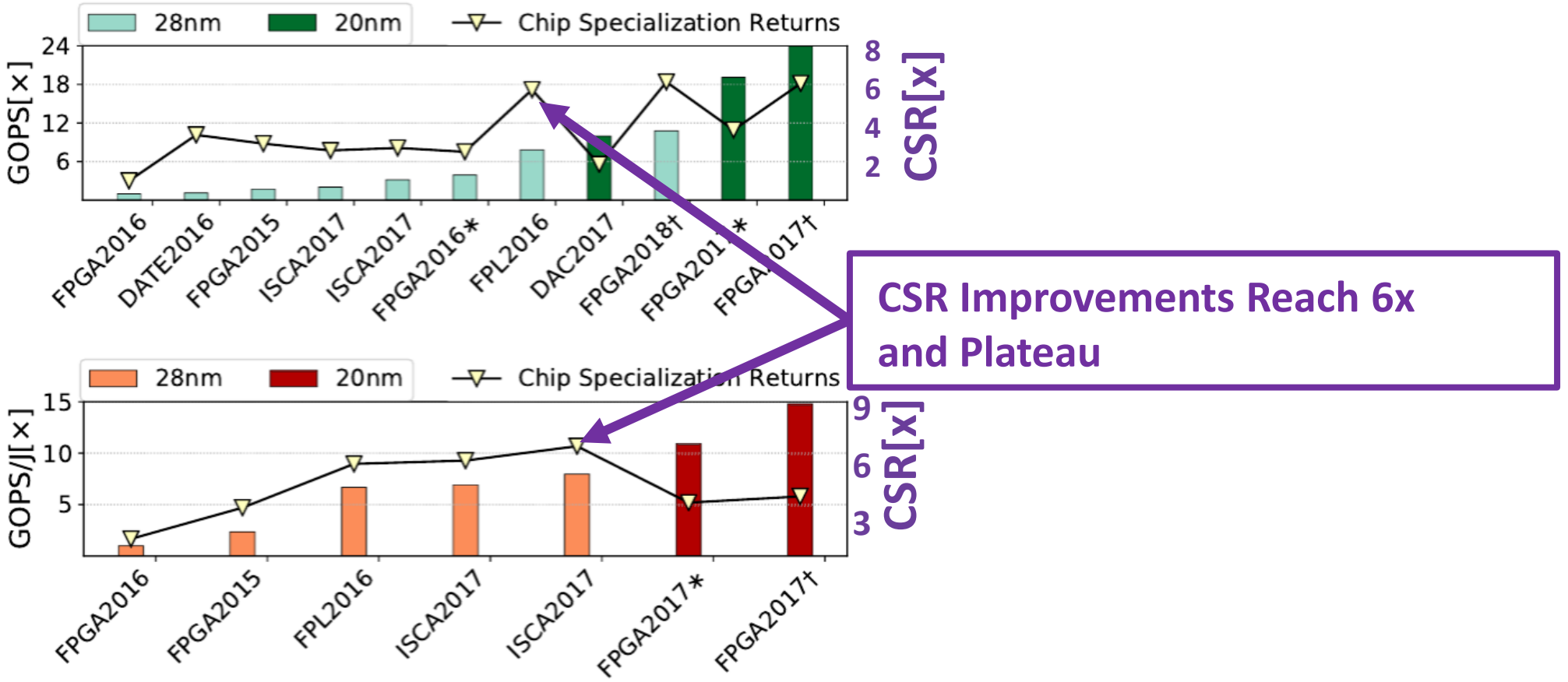
Case Study 1: Deep Learning on FPGAs



Throughput Improved by 24x
Energy Efficiency Improved by 15x

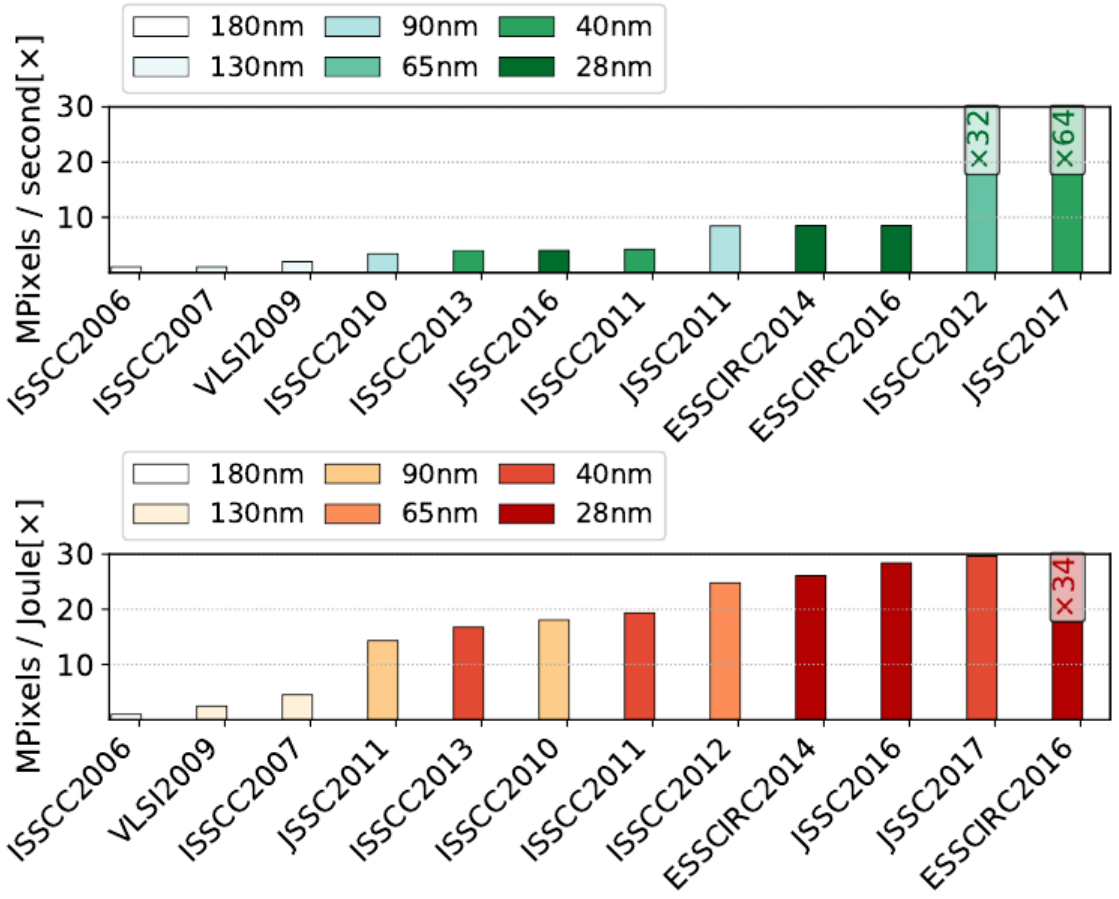
Case Study 1: Deep Learning on FPGAs

- 11 Implementations of The AlexNet CNN Architecture on FPGAs

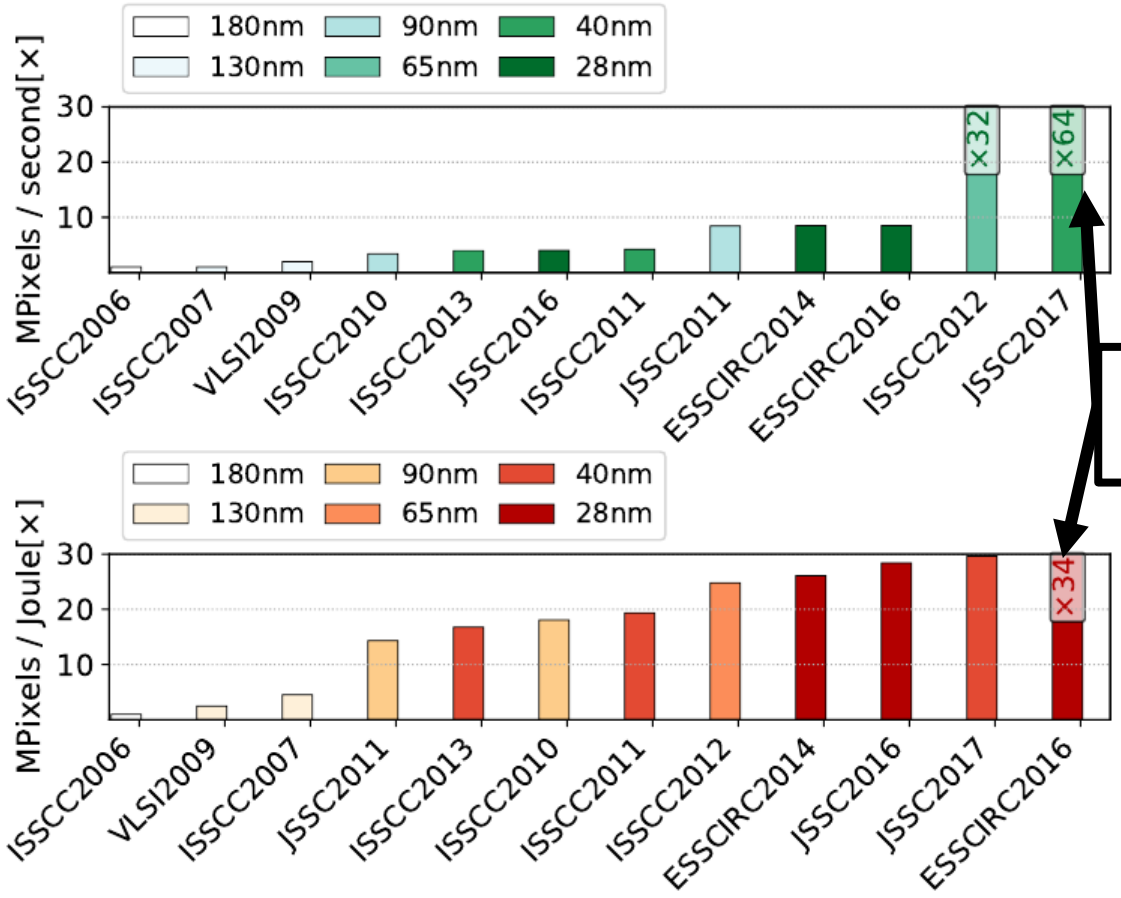


- **Why:** Recent Efforts Increased FPGA Utilization (More Transistors).
  Deep Learning is an **Emerging Domain**. **Still Hope for Better Returns**!

Case Study 2: Video Decoding on ASICs



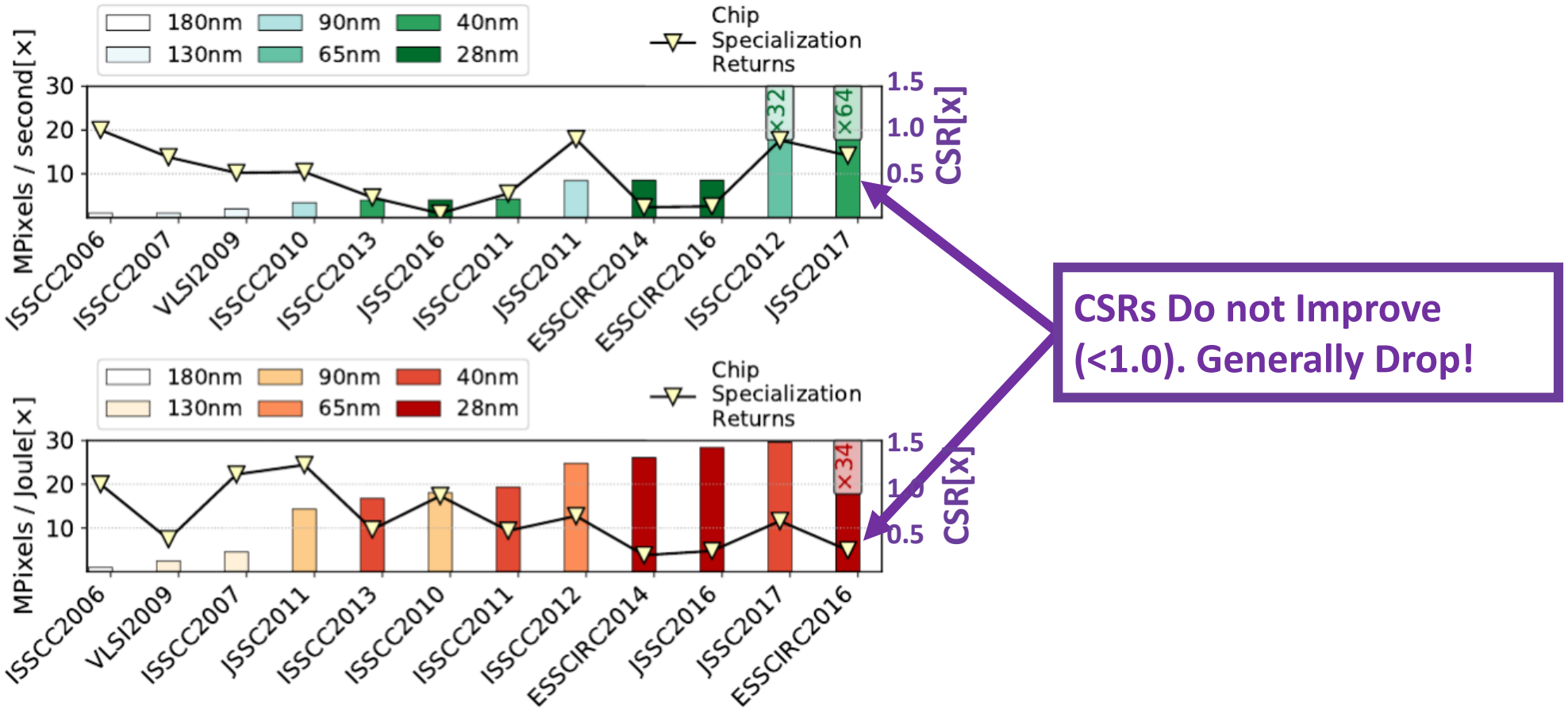
Case Study 2: Video Decoding on ASICs



Decoding Throughput Improved by 64x
Energy Efficiency Improved by 34x

Case Study 2: Video Decoding on ASICs

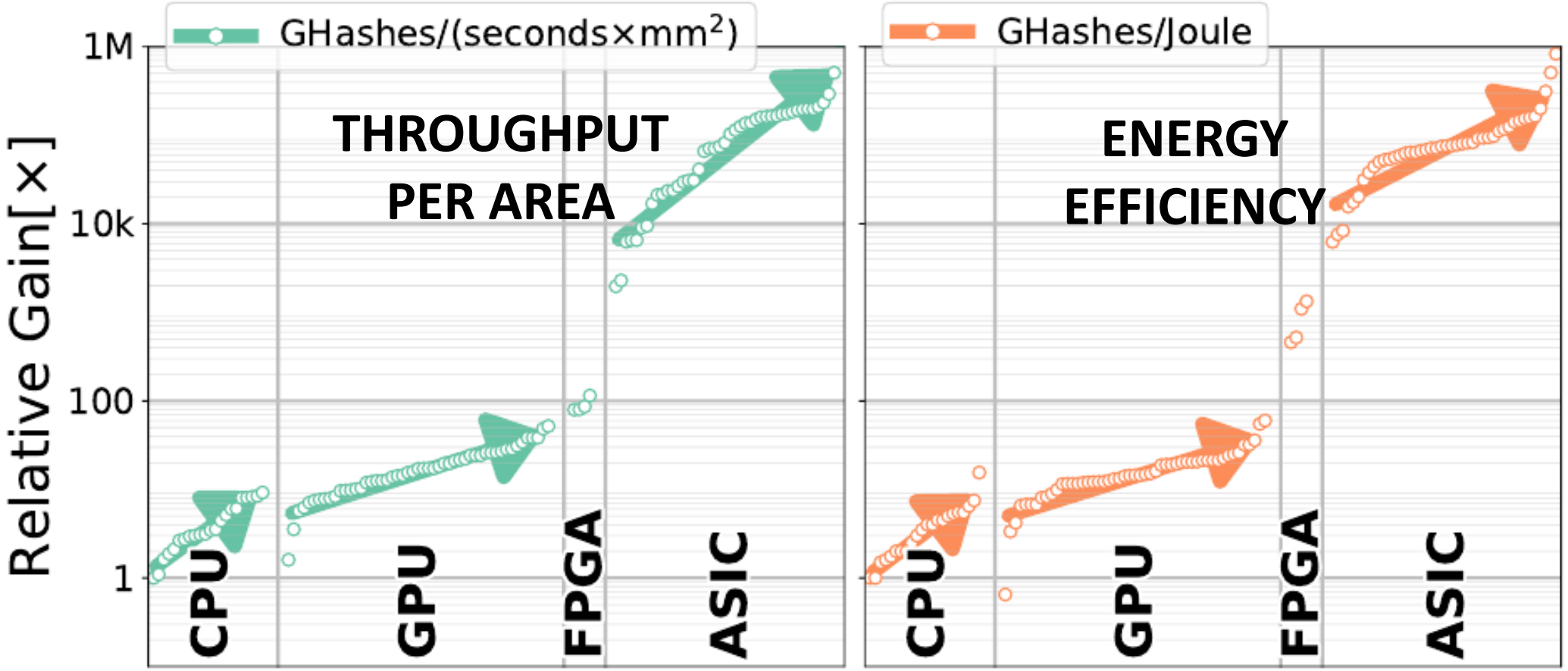
- 12 Video Decoding ASIC Chips, Taped-out Over an 11 Years Period.



- **<u>Why:</u> Domain Maturity.**
  - Designs Convergence After Decades of Specialization Efforts.

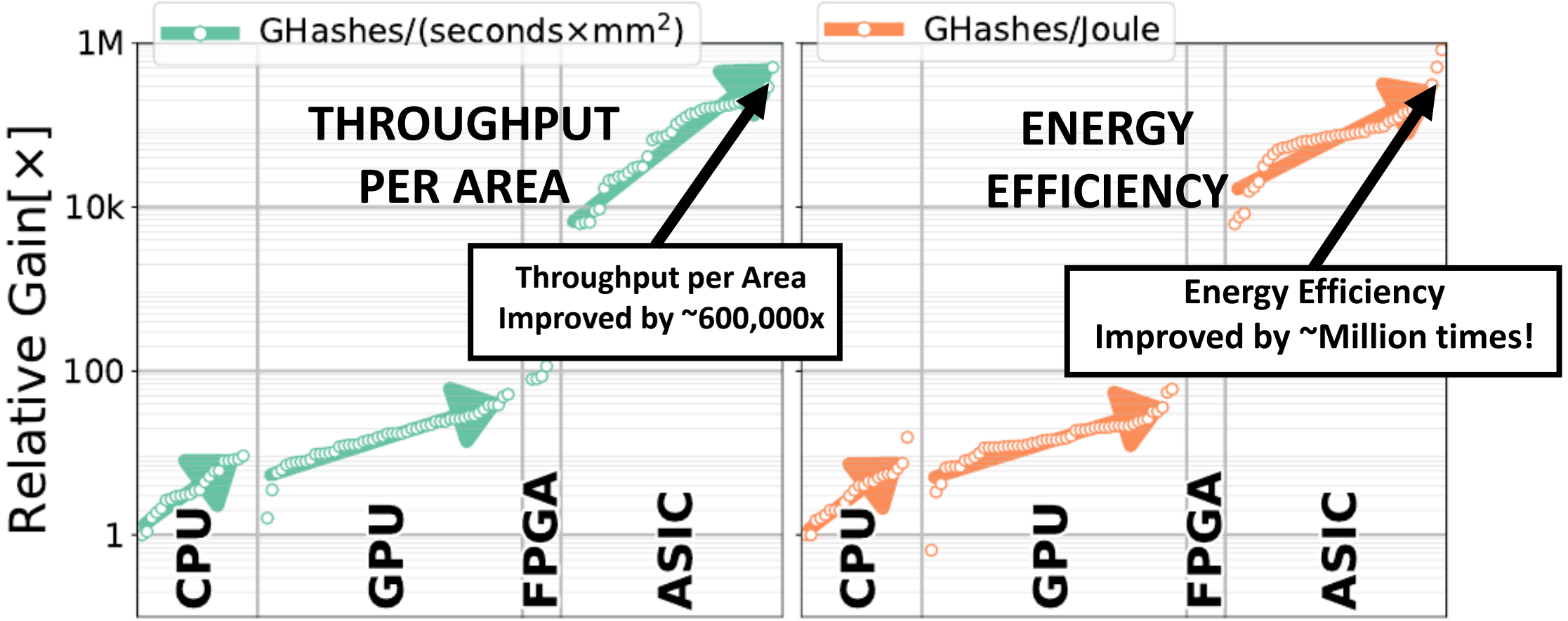
Case Study 3: Bitcoin Mining on CPU/GPU/FPGA/ASIC

- Analyze the **Cross-Platform Evolution:** From CPUs, to GPUs, FPGAs, and ASICs.



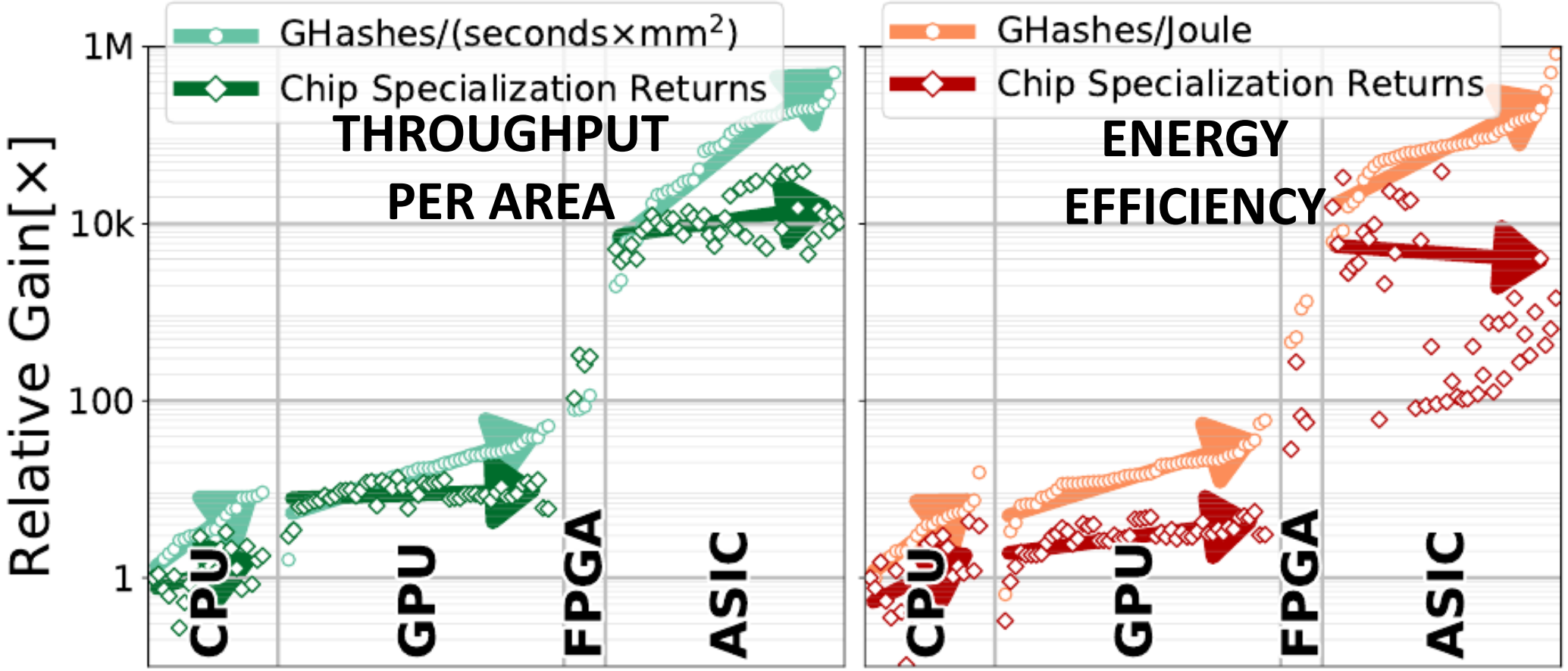
Case Study 3: Bitcoin Mining on CPU/GPU/FPGA/ASIC

- Analyze the **Cross-Platform Evolution:** From CPUs, to GPUs, FPGAs, and ASICs.



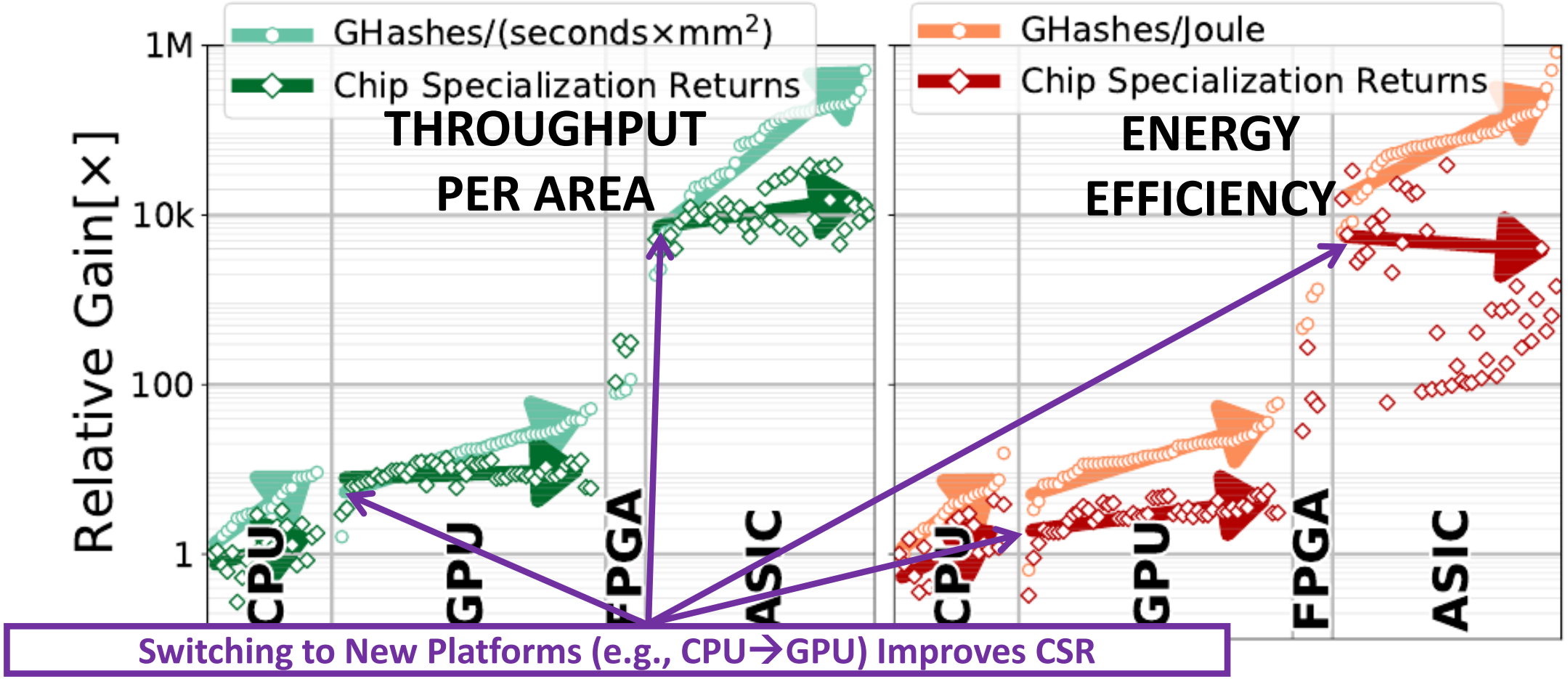
Case Study 3: Bitcoin Mining on CPU/GPU/FPGA/ASIC

- Analyze the **Cross-Platform Evolution:** From CPUs, to GPUs, FPGAs, and ASICs.



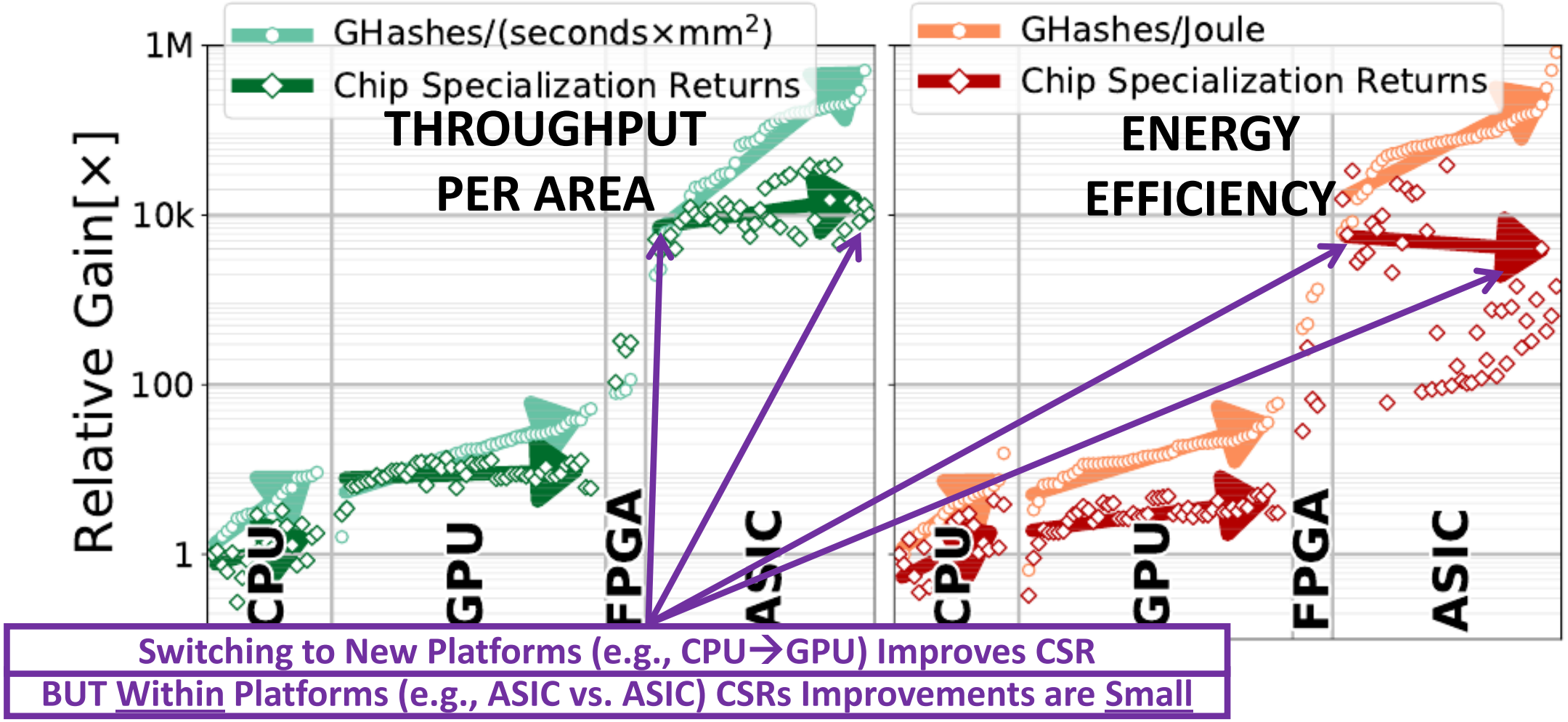
Case Study 3: Bitcoin Mining on CPU/GPU/FPGA/ASIC

- Analyze the **Cross-Platform Evolution:** From CPUs, to GPUs, FPGAs, and ASICs.



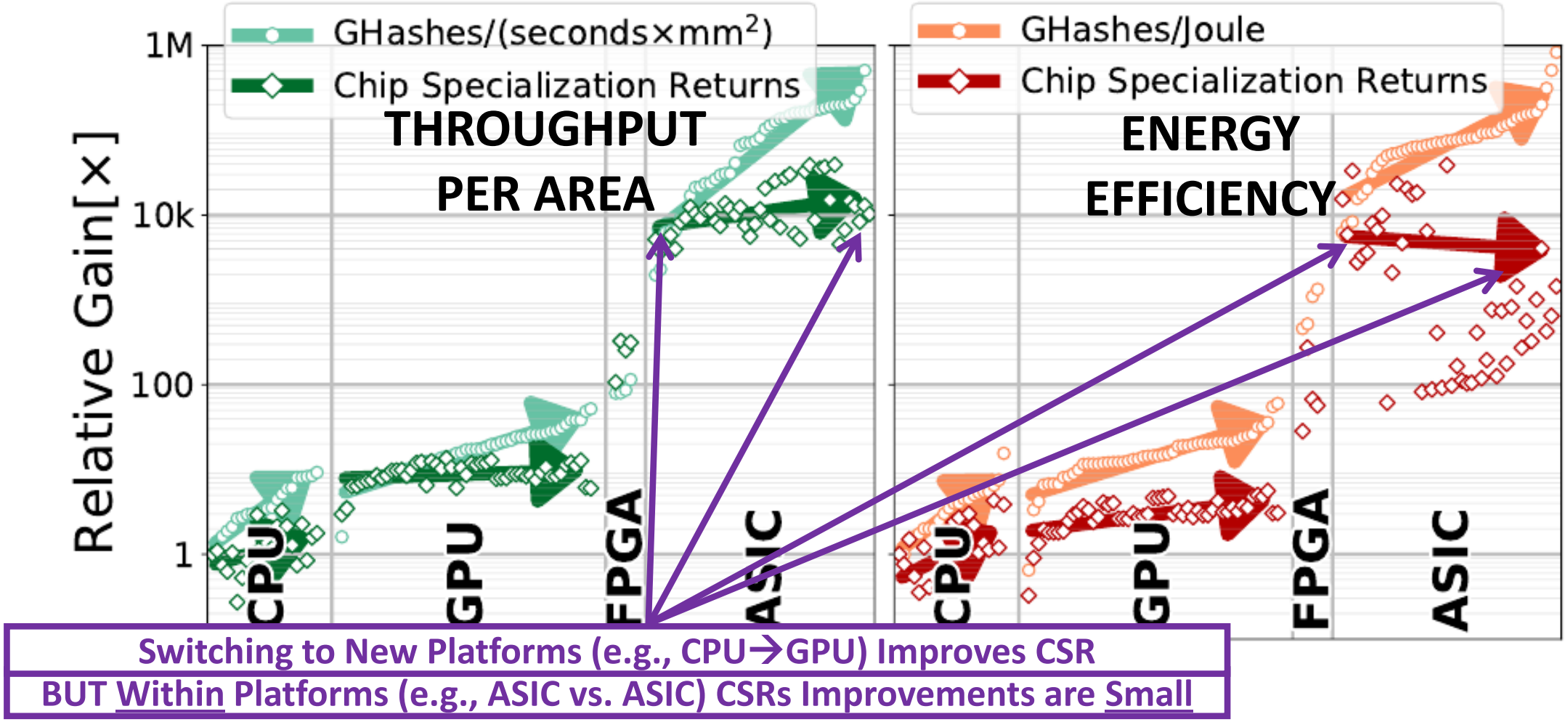
Case Study 3: Bitcoin Mining on CPU/GPU/FPGA/ASIC

- Analyze the **Cross-Platform Evolution:** From CPUs, to GPUs, FPGAs, and ASICs.



Case Study 3: Bitcoin Mining on CPU/GPU/FPGA/ASIC

- Analyze the **Cross-Platform Evolution:** From CPUs, to GPUs, FPGAs, and ASICs.



- **Why: Confined Computation** (Brute-force SHA256) + High-Pace CMOS Adoption

Chip Specialization Pitfalls

- **Massive Parallelism (e.g., GPU Gaming):**

  More Transistors → More Cores → More Parallelism

Chip Specialization Pitfalls

- **Massive Parallelism (e.g., GPU Gaming):**

  NO More Transistors → **NO** More Cores → **NO** More Parallelism

Chip Specialization Pitfalls

- **Massive Parallelism (e.g., GPU Gaming):**

  NO More Transistors → **NO** More Cores → **NO** More Parallelism

- **Confined Domains (e.g., Bitcoin Mining):**

  Limited Optimization Knobs

Chip Specialization Pitfalls

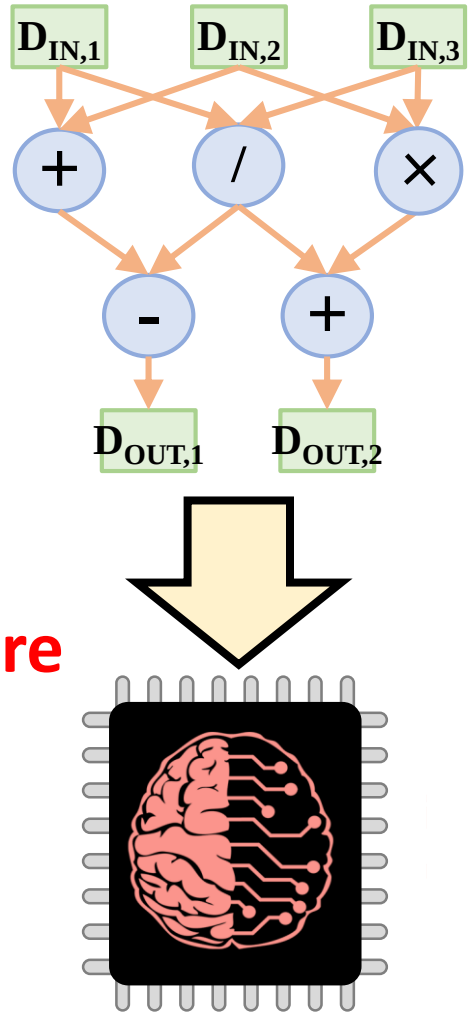
- **Massive Parallelism (e.g., GPU Gaming):**
  NO More Transistors → **NO** More Cores → **NO** More Parallelism

- **Confined Domains (e.g., Bitcoin Mining):**
  Limited Optimization Knobs

- **Mature Domains (e.g., Video Decoding):**
  Hard to Further Innovate Once Problem is Well-Studied

Chip Specialization Pitfalls

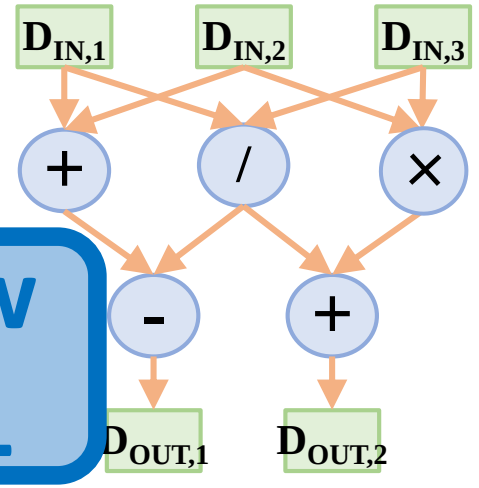
- **Massive Parallelism (e.g., GPU Gaming):**
  NO More Transistors → **NO** More Cores → **NO** More Parallelism
- **Confined Domains (e.g., Bitcoin Mining):**
  Limited Optimization Knobs
- **Mature Domains (e.g., Video Decoding):**
  Hard to Further Innovate Once Problem is Well-Studied
- **Finite Ways to Map a Computation Problem to Fixed-Hardware**
  Fixed Optimization Problems Have a Finite Solution Space!

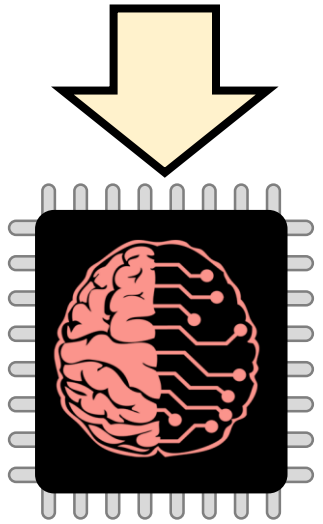


Chip Specialization Pitfalls

- **Massive Parallelism (e.g., GPU Gaming):**
  NO More Transistors → **NO** More Cores → **NO** More Parallelism
- **Confined Domains (e.g., Bitcoin Mining):**
  Limited Optimization Knobs
- **Mature Dor**
  Hard to F[illegible]

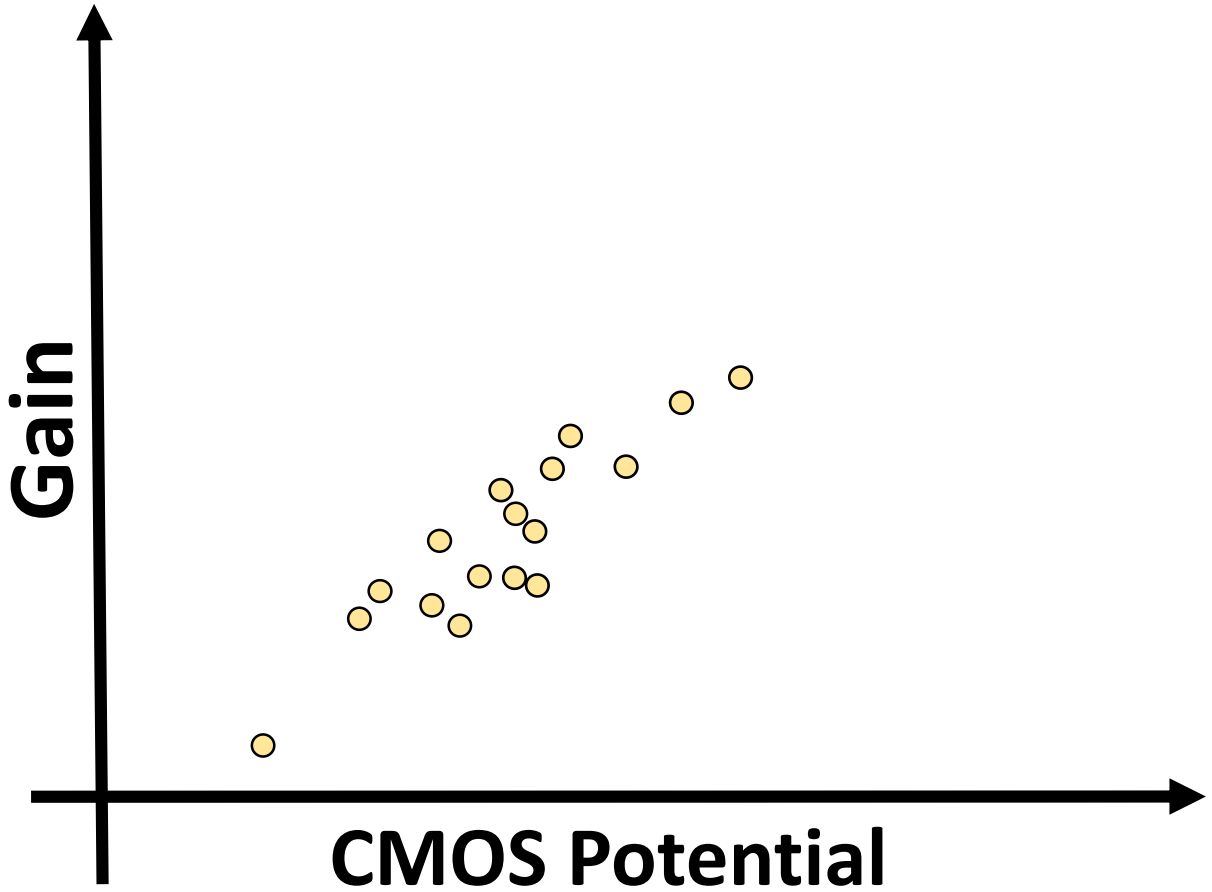
Following the End of Moore's Law
ACCELERATORS WILL HIT A WALL



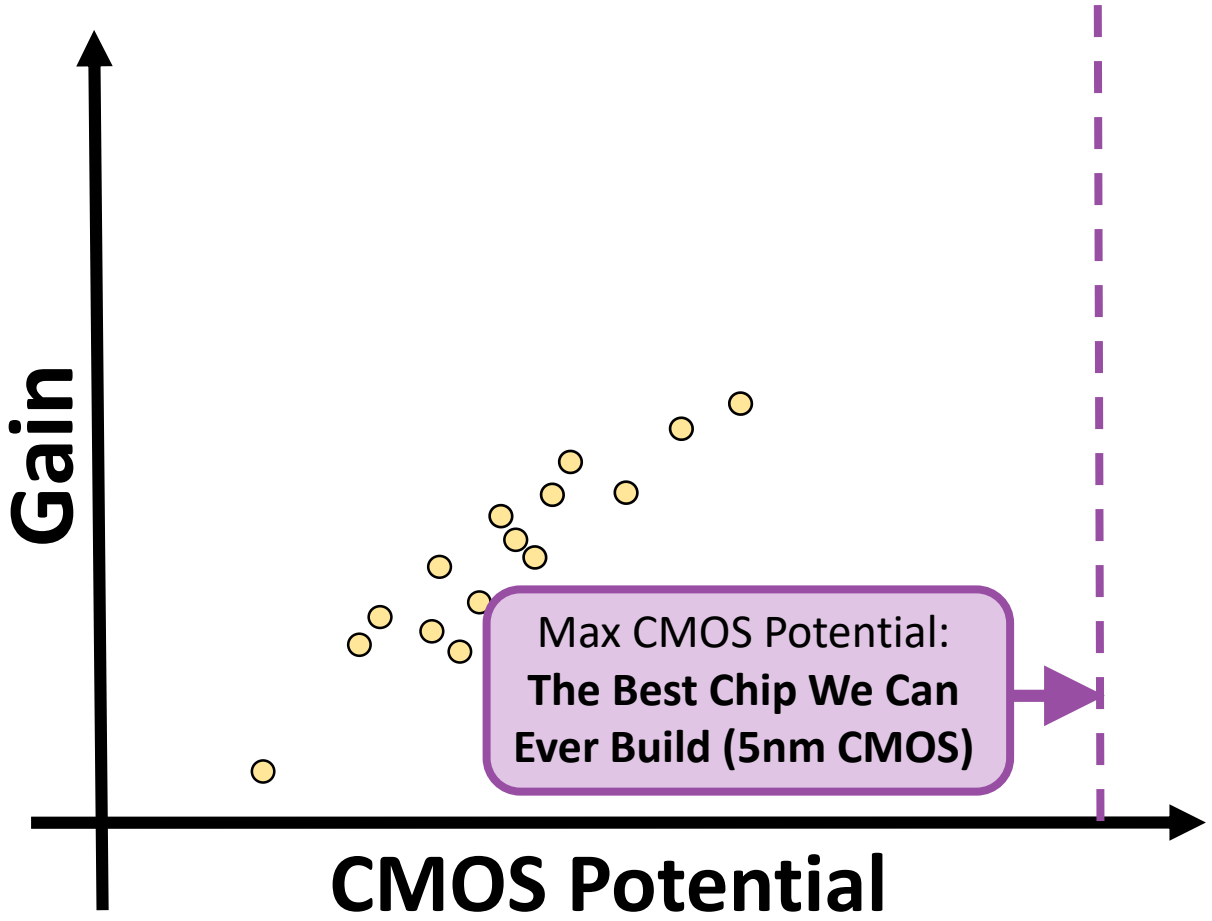


- **Finite Ways to Map a Computation Problem to Fixed-Hardware**
  Fixed Optimization Problems Have a Finite Solution Space!

The Accelerator Wall

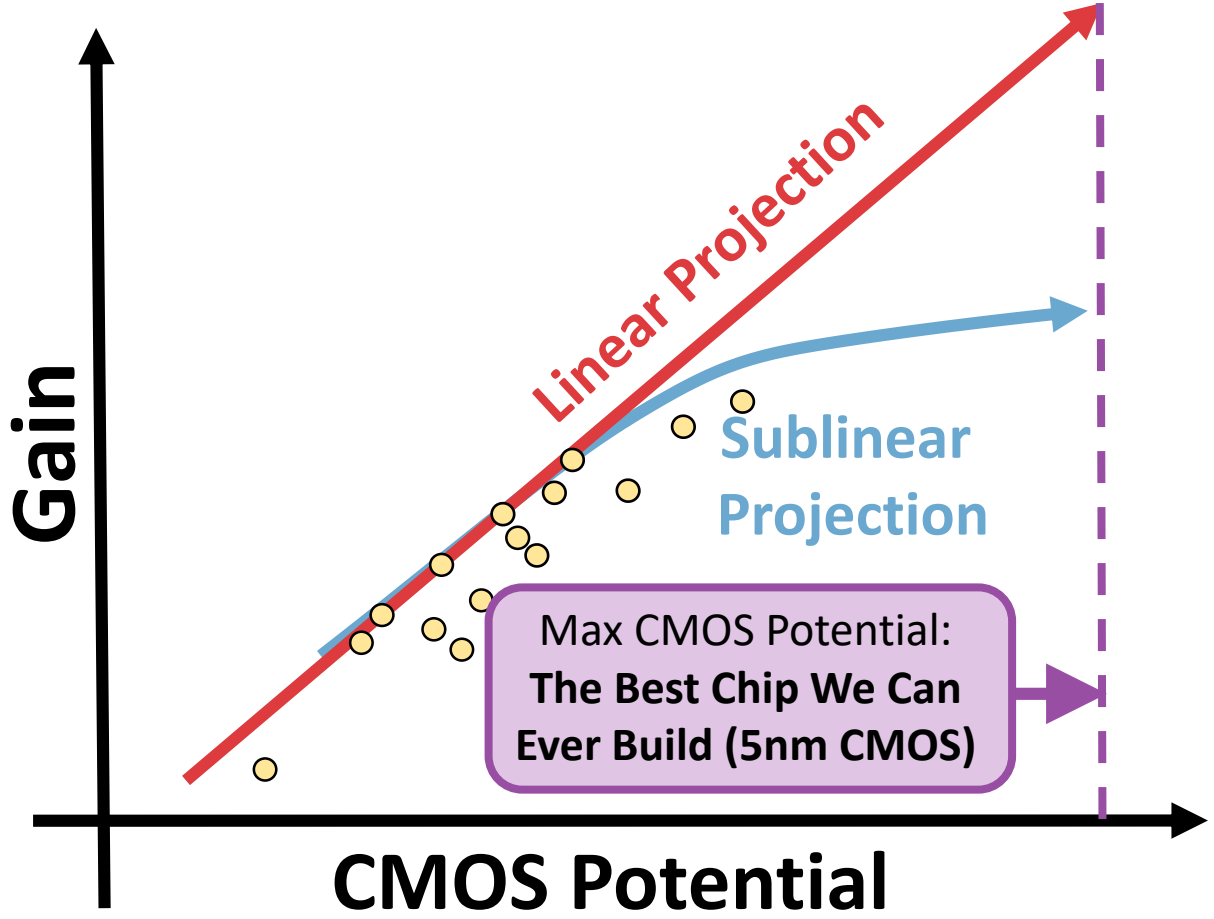


The Accelerator Wall



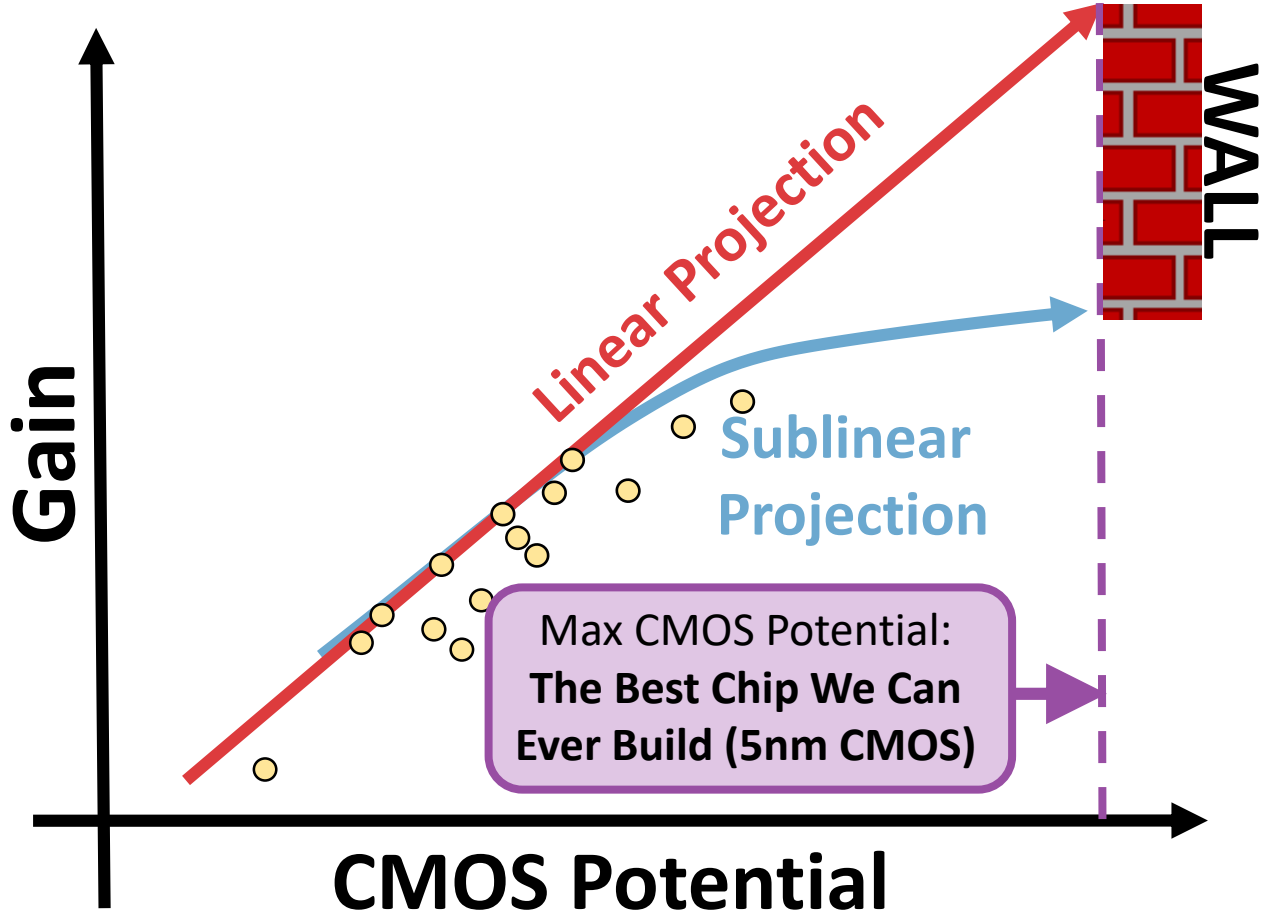
The Accelerator Wall

- Given a Group of Accelerators, Plot Corresponding Points in a “Gain vs. CMOS” Space.
- Use Projections to Predict **the Accelerator Wall** at the End of Moore’s Law (5nm CMOS)**:**



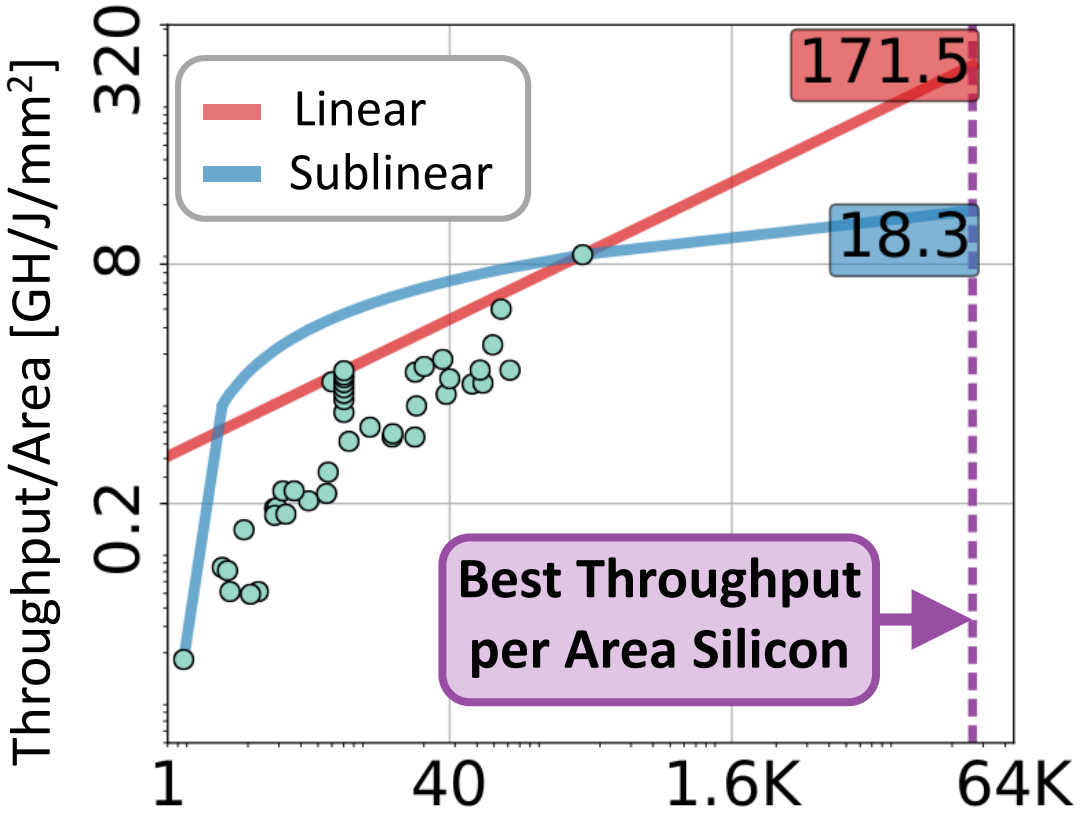
The Accelerator Wall

- Given a Group of Accelerators, Plot Corresponding Points in a “Gain vs. CMOS” Space.
- Use Projections to Predict **the Accelerator Wall** at the End of Moore’s Law (5nm CMOS)**:**



The Accelerator Wall

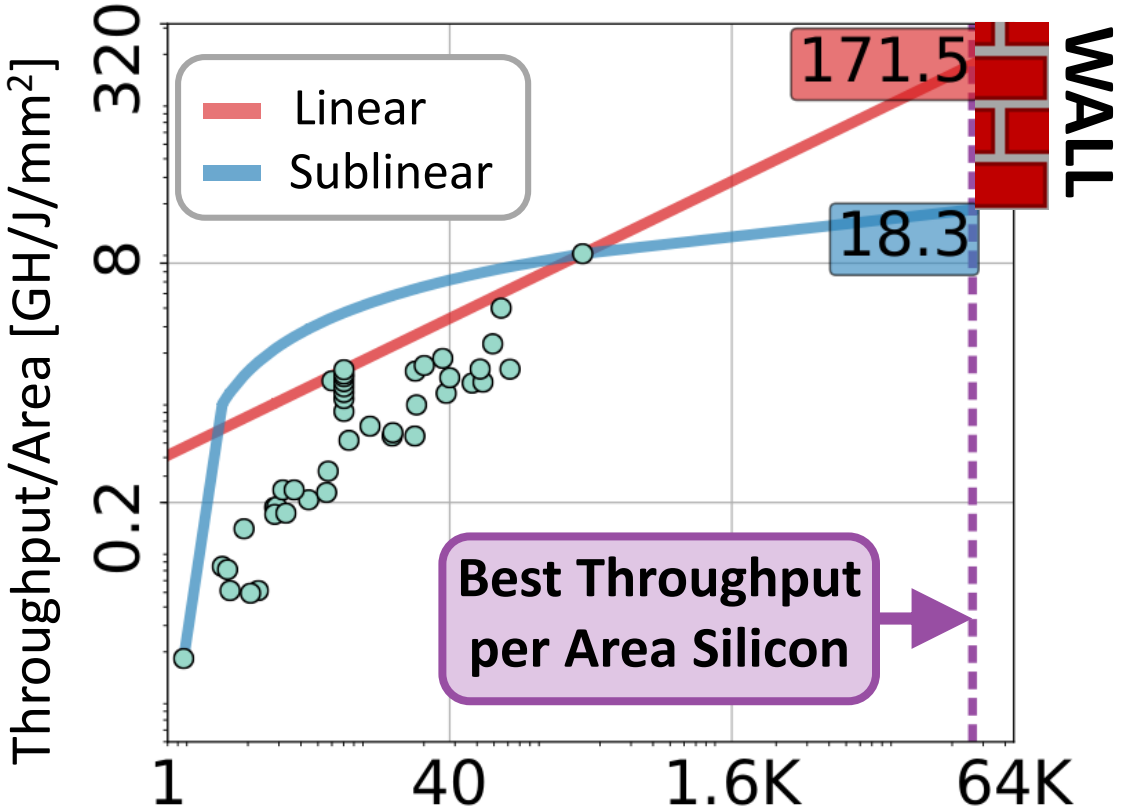
- **Example: ASIC Bitcoin Miners**



Throughput/Area Improved by 600,000x (vs. CPUs)

The Accelerator Wall

- **Example: ASIC Bitcoin Miners**

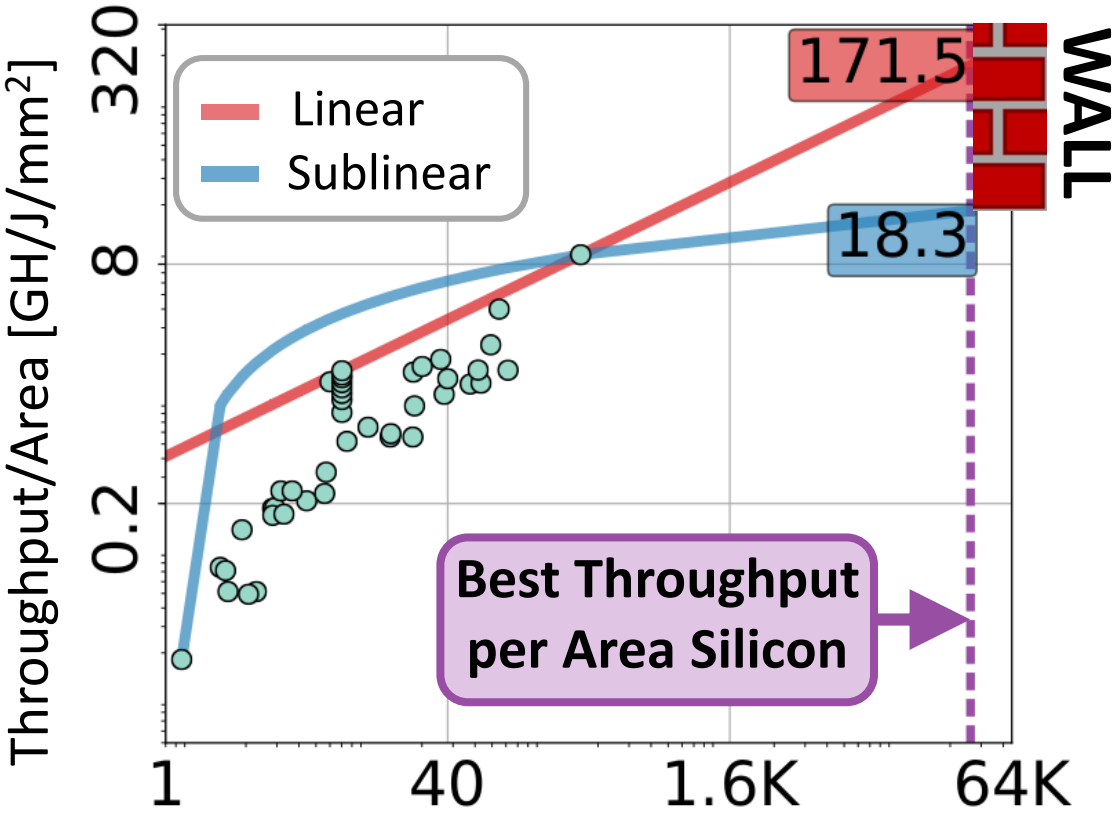


Throughput/Area Improved by 600,000x (vs. CPUs)

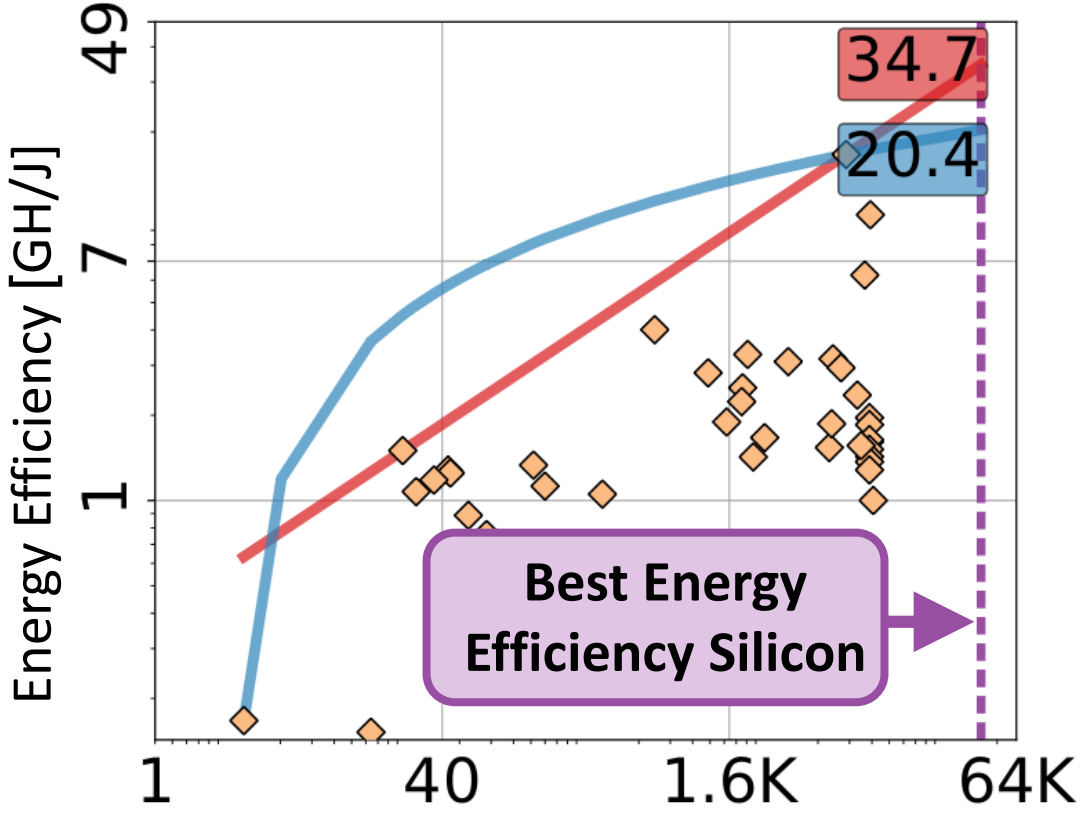
WALL follows an Additional 2-18.5x.

The Accelerator Wall

- **Example: ASIC Bitcoin Miners**



Throughput/Area Improved by 600,000x (vs. CPUs)

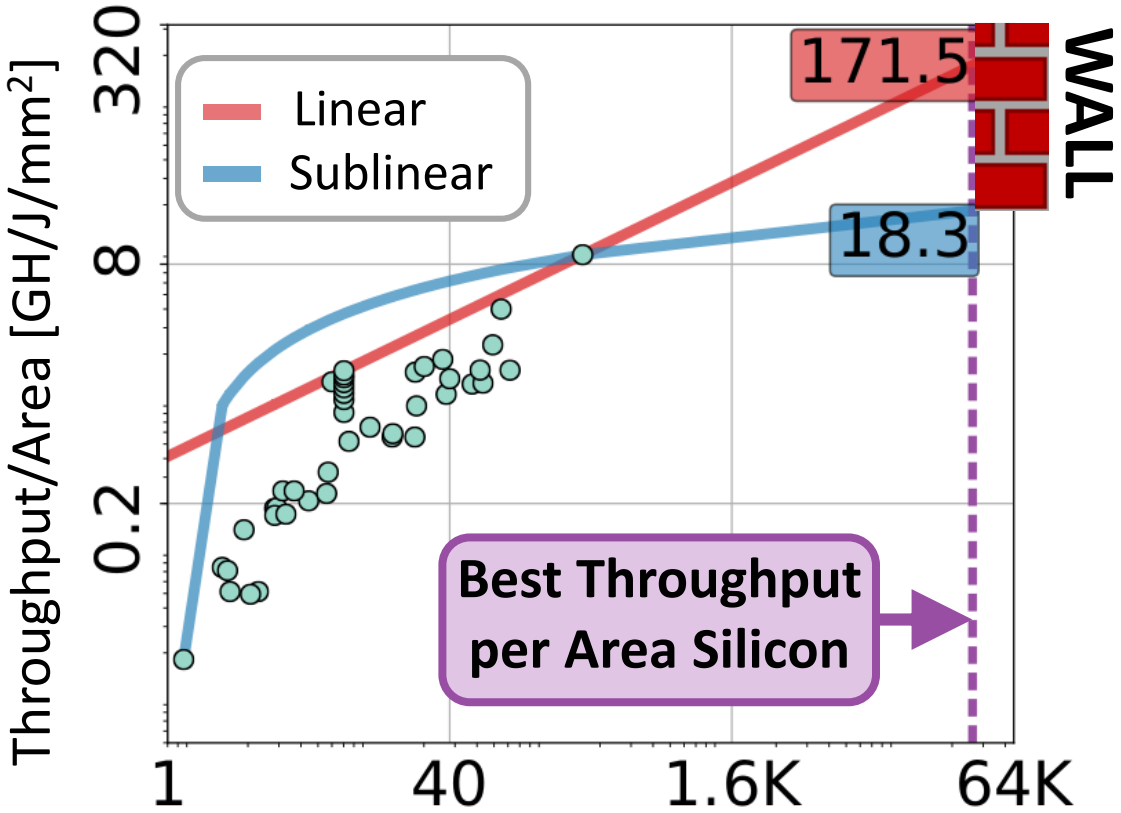


Energy Efficiency Improved by 1,000,000x (vs. CPUs)

WALL follows an Additional 2-18.5x.

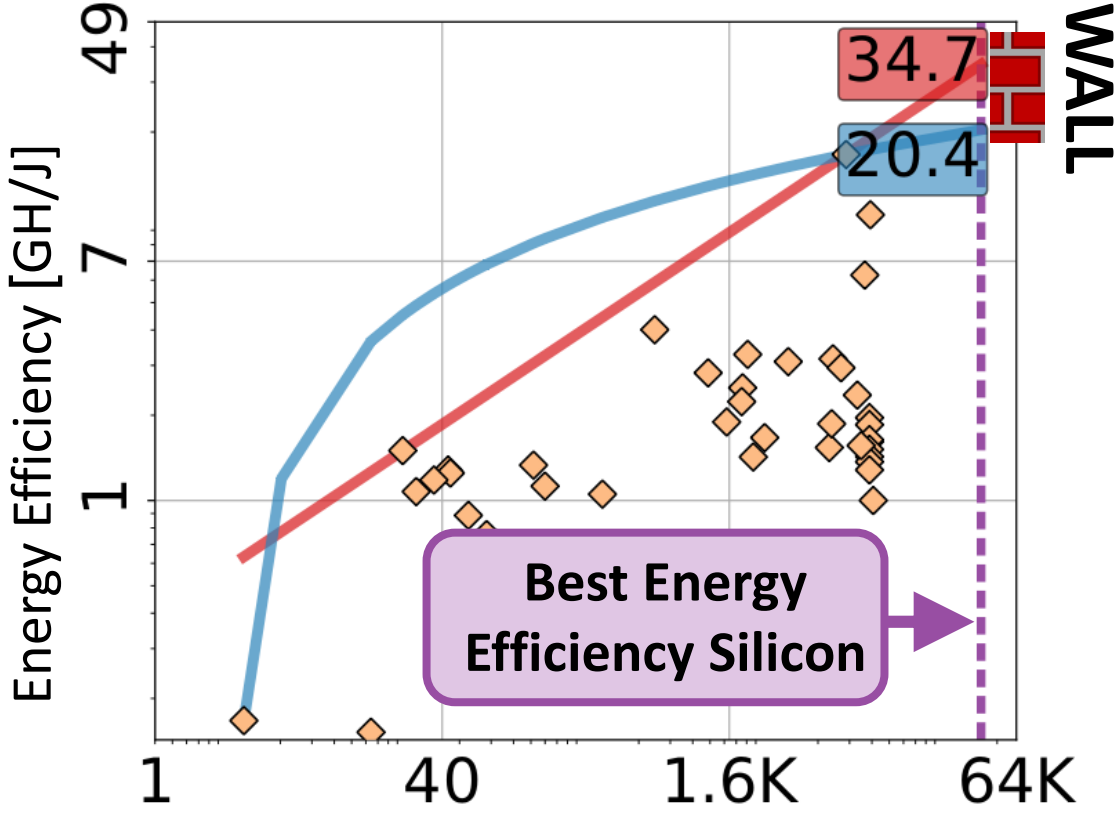
The Accelerator Wall

- **Example: ASIC Bitcoin Miners**



Throughput/Area Improved by 600,000x (vs. CPUs)

WALL follows an Additional 2-18.5x.



Energy Efficiency Improved by 1,000,000x (vs. CPUs)

WALL follows an Additional 1.3-2.1x.

Rankine: A CMOS Potential Modeling Tools

- Named After William Rankine That Termed: “Potential Energy”

Source: Wikipedia

- Based on Datasheets of **Thousands of Commercial Processors**
- Calculates CMOS Potential Functions Based on Physical Chip Properties
  - e.g., CMOS Process, Die Size/Number of Transistors, TDP, etc.

✓ **Quickly Explore CMOS Costs of Design Alternatives as You Build Accelerators.**

GitHub Repo: https://github.com/PrincetonUniversity/accelerator-wall

Accelerator Zoo: a Database of Accelerator Statistics

- A Database of Popular Specialized Applications and Accelerator Statistics
  - Deep Learning on FPGAs.
  - Video Decoding ASICs.
  - Bitcoin Mining ASICs.

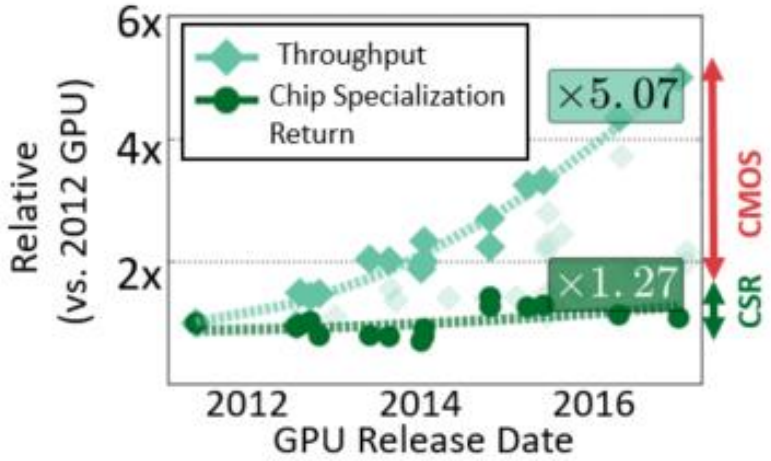
Performance Evolution of ASIC Bitcoin Miners

✓ **Evaluate Your Accelerator vs. Existing Accelerators in a Given Domain.**
✓ **Contribute Your Accelerator Data, Help us Understand New Domains.**

GitHub Repo: https://github.com/PrincetonUniversity/accelerator-wall

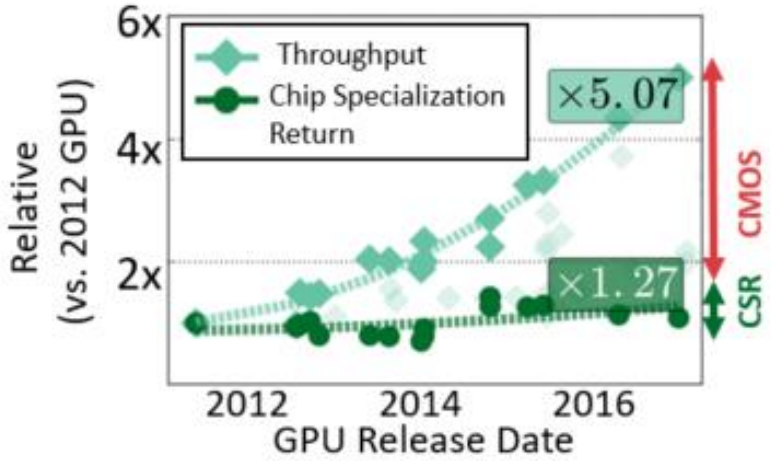
Conclusions

- **Be Mindful About Specialization vs. CMOS Returns in Accelerators**
  - Demystifying Shows Gains are Mostly CMOS-Dominated.
  - HW/SW Optimizations Play a Secondary Role.

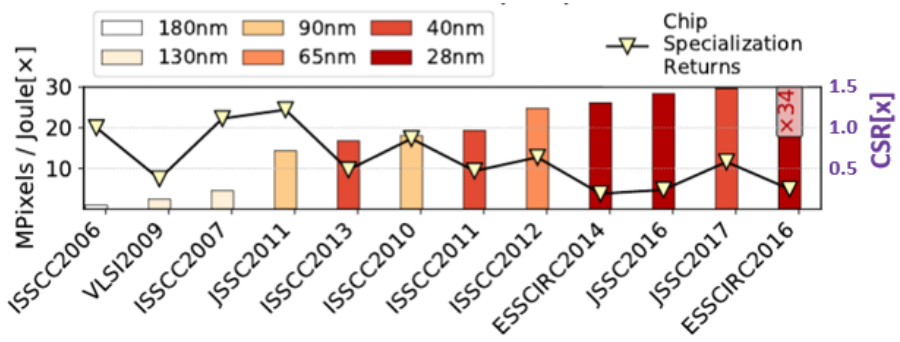


Conclusions

- **Be Mindful About Specialization vs. CMOS Returns in Accelerators**
  - Demystifying Shows Gains are Mostly CMOS-Dominated.
  - HW/SW Optimizations Play a Secondary Role.

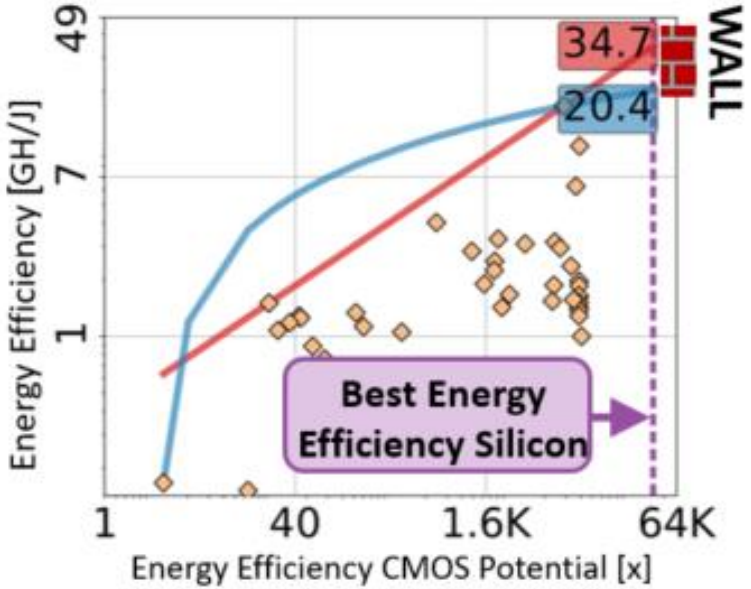


- **Be Mindful About Chip Specialization Pitfalls and Diminishing Returns**
  - **Parallelism** Dies With CMOS Scaling: No More Transistors = No More Cores.
  - **Confined** and **Mature** Domains Have Limited Improvement Opportunities.



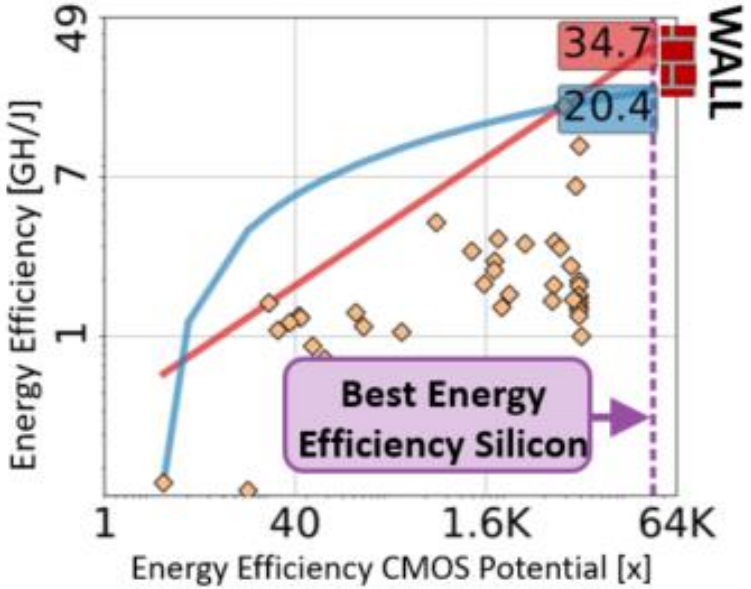
Conclusions

- **Chip Specialization is Not a Long-Term Remedy for The End of Moore's Law.**
  - **All Popular Domains Will Mature.** Diminishing Optimization Returns Will Follow.
  - After the End of Moore's Law, Accelerators Will Run Out of Steam **FASTER THAN EXPECTED**

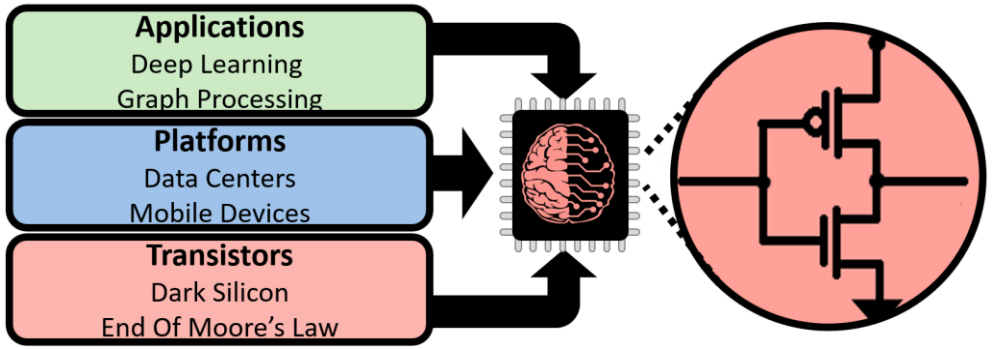


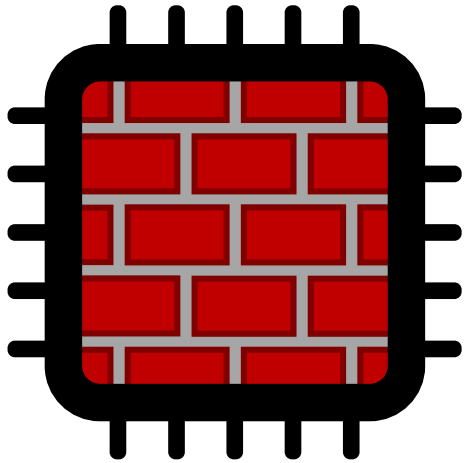
Conclusions

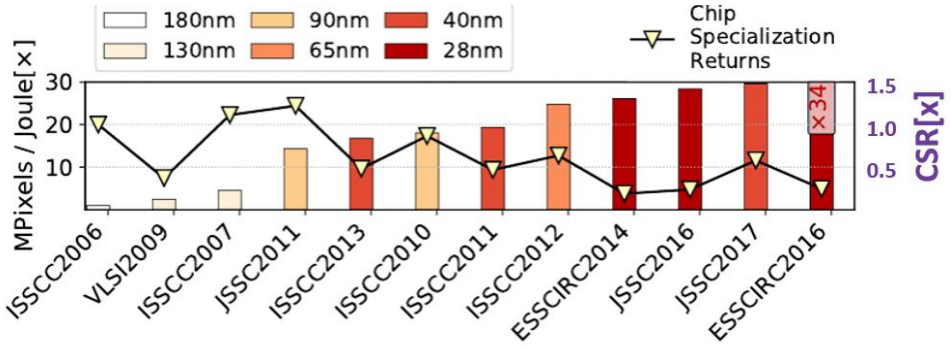
- **Chip Specialization is Not a Long-Term Remedy for The End of Moore's Law.**
  - **All Popular Domains Will Mature.** Diminishing Optimization Returns Will Follow.
  - After the End of Moore's Law, Accelerators Will Run Out of Steam **FASTER THAN EXPECTED**

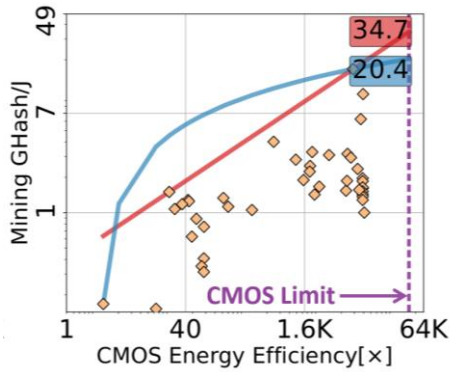


- **We Must Explore Other Forms of Optimization, That Are NOT CMOS Driven.**

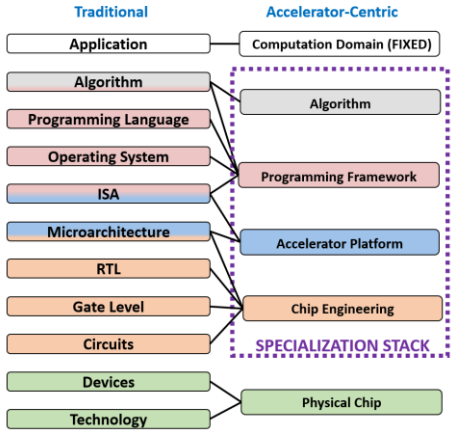








THANK YOU



Adi Fuchs and David Wentzlaff

{adif,wentzlaf}@princeton.edu

PRINCETON UNIVERSITY

GitHub Repo: https://github.com/PrincetonUniversity/accelerator-wall