System Design for Machine Intelligence

CSE 240D

Spring 2026

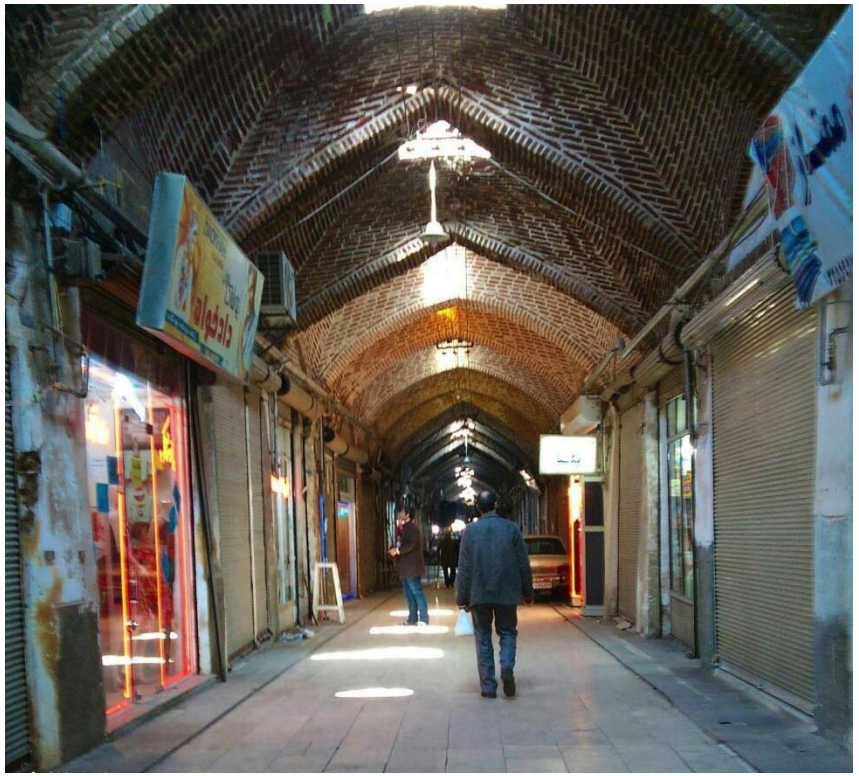
Hadi Esmaeilzadeh

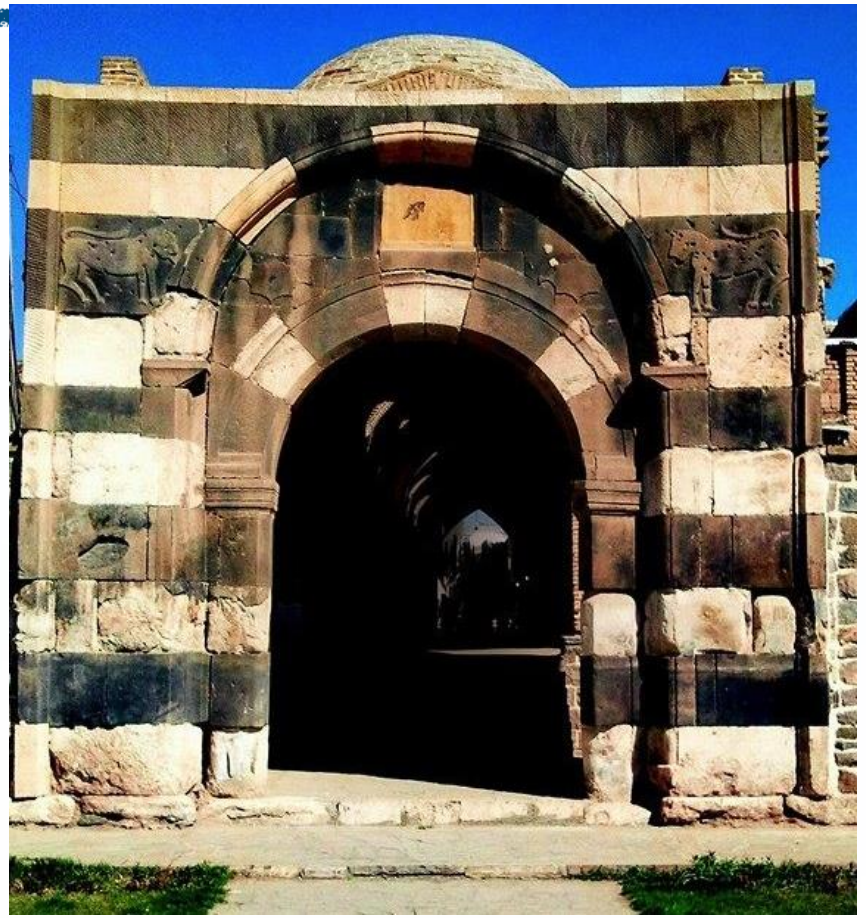
hadi@ucsd.edu

University of California, San Diego

Hadi Esmaeilzadeh

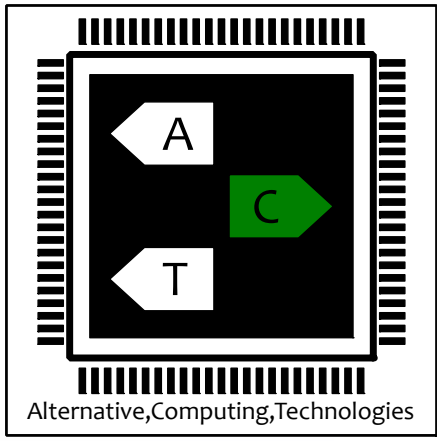
From Khoy, Iran





Research: ACT Lab

Alternative Computing Technologies



- System design for immersive machine intelligence
- System design for robotics
- System design for health

- Analog computing
- General-purpose approximate computing
- Bridging neuromorphic and von Neumann models of computing

Agenda

1. Who is Hadi
2. **Course organization**
3. Why CSE 240D System Design for Deep Learning
    1. How we became and industry of new capabilities
    2. Why we might become an industry of replacement
    3. Specialization and accelerators

Objective

- Explore cutting-edge research on new paradigms of deep learning
- Empower you with higher-order critical thinking
- Improve your technical writing and speaking
- Innovate in advanced microarchitecture

Format

- Seminar course
  - Reading papers
  - Critiquing and discussing the papers
  - Brainstorming about new ideas
  - Developing new technologies
  - Class Website: https://hadianeh.github.io/hadiclass.github.io/cse240d-sp26/
    - Everything you need to know is there, please check it regularly
- Half you presentation and half my presentations

Grading rubric

Component	Fraction
Class Presentation	20%
Class Participation	10%
Critiques	30%
Final Project	40%

Class presentation

- Objective: Communicate and analyze ideas

- **4 points: Clearly presenting the key ideas**
- 1 points: Clear, well-organized slides
- **5 points: Stimulating interesting discussion**
- 1 point bonus

Class participation

- You have to say something interesting!

- At least two comments/questions within your critique
  - Your new ideas
  - Critical questions about methodologies and conclusions
  - Why will the paper be cited
  - What you learned
  - Main insights from the papers

Critiques

- Objective: developing high-order critical thinking

- Summary (quarter a page)
- Strengths (1-3 sentences)
- Weaknesses (1-3 sentences)
- Analysis I (1 paragraph)
- Analysis II (1 paragraph)

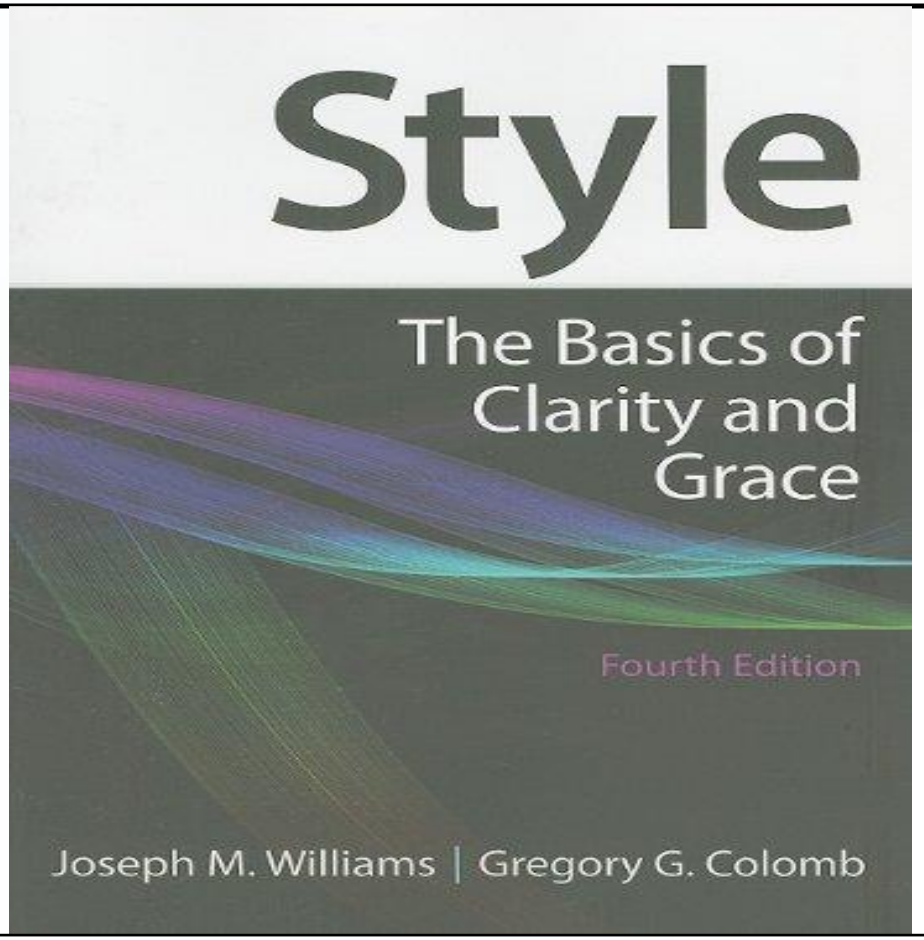
- Please read the “The task of the referee by Alan Jay Smith”
- ***Submit two days before class on Gradescope***

Reading material for writing critiques

The task of the referee

Allen Jay Smith

Style: The Basics of Clarity and Grace
Joseph M. Williams



Final project

- Groups of two (Tentative)
- Options
  - Implementing a new idea
  - Extending an existing paper
  - Re-implement a paper
  - Survey at least ten papers
- Evaluation
  - Implementation
  - Writing
  - Oral presentation

Prerequisites

- Understand a subset of
  - VLSI Circuits
  - Computer architecture
  - Programming Languages
  - Machine learning
- Do
  - Programming

Agenda

1. Who is Hadi
2. Course organization
3. **Why CSE 240D System Design for Deep Learning**
    1. How we became and industry of new capabilities
    2. Why we might become an industry of replacement
    3. Specialization and accelerators

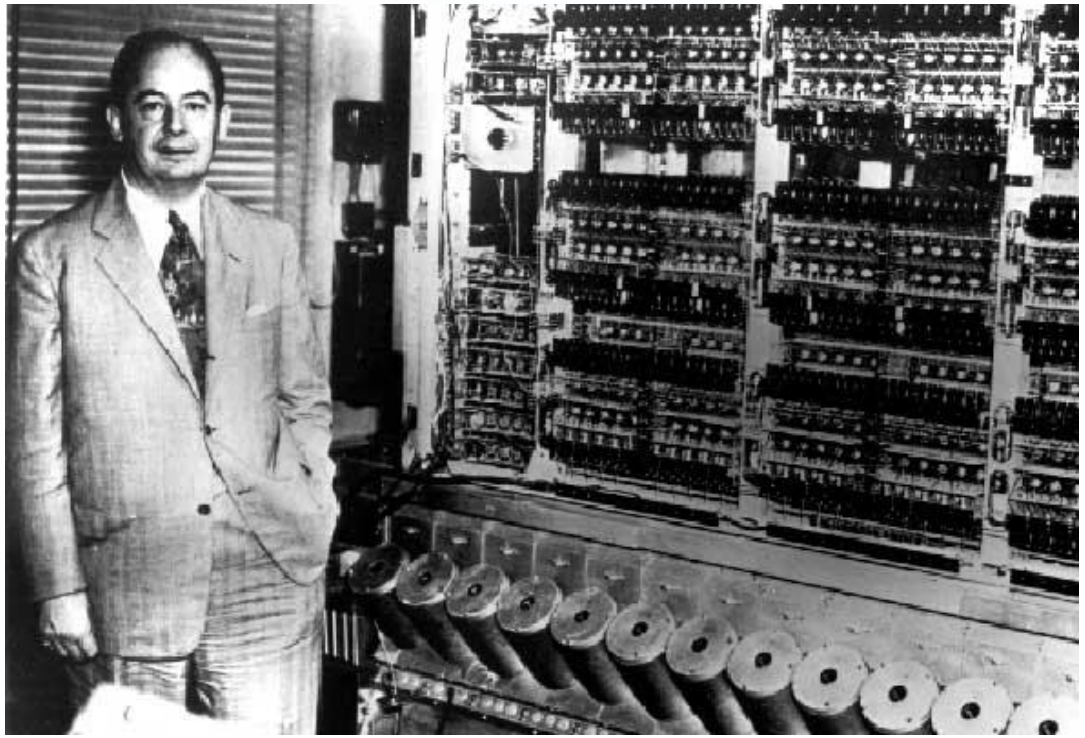
What has made computing pervasive? What is the backbone of computing industry?

Programmability

```
public class TcpClientSample
{
    public static void Main()
    {
        byte[] data = new byte[1024]; string input, stringData;
        TcpClient server;
        try{
            server = new TcpClient(" . . . . ", port);
        }catch (SocketException){
            Console.WriteLine("Unable to connect to server");
            return;
        }
        NetworkStream ns = server.GetStream();
        int recv = ns.Read(data, 0, data.Length);
        stringData  = Encoding.
          ASCII.GetString(data, 0, recv);
        Console.WriteLine(stringData);
        while(true){
            input = Console.ReadLine();
            if (input == "exit") break;
                newchild.Properties["ou"].Add
                ("Auditing Department");
                newchild.CommitChanges();
                newchild.Close();
```

Networking

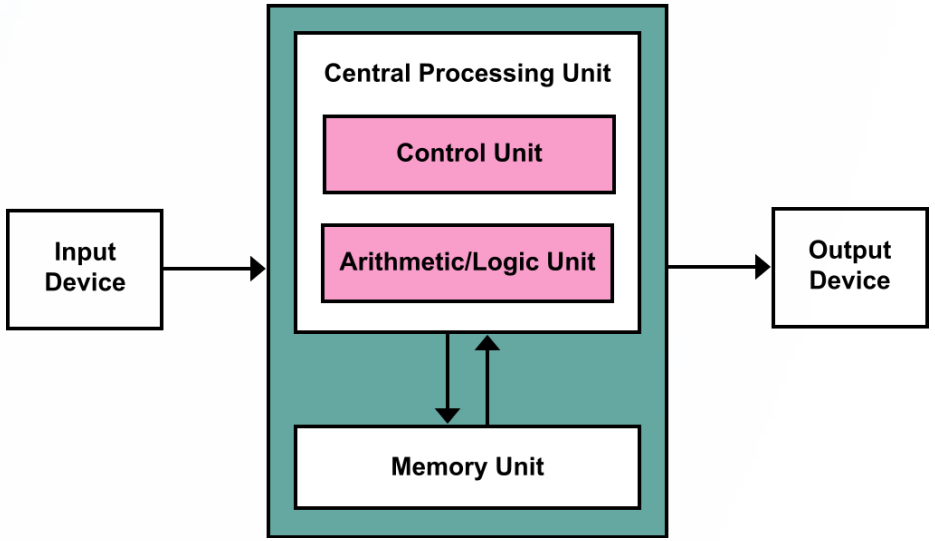
What makes computers programmable?



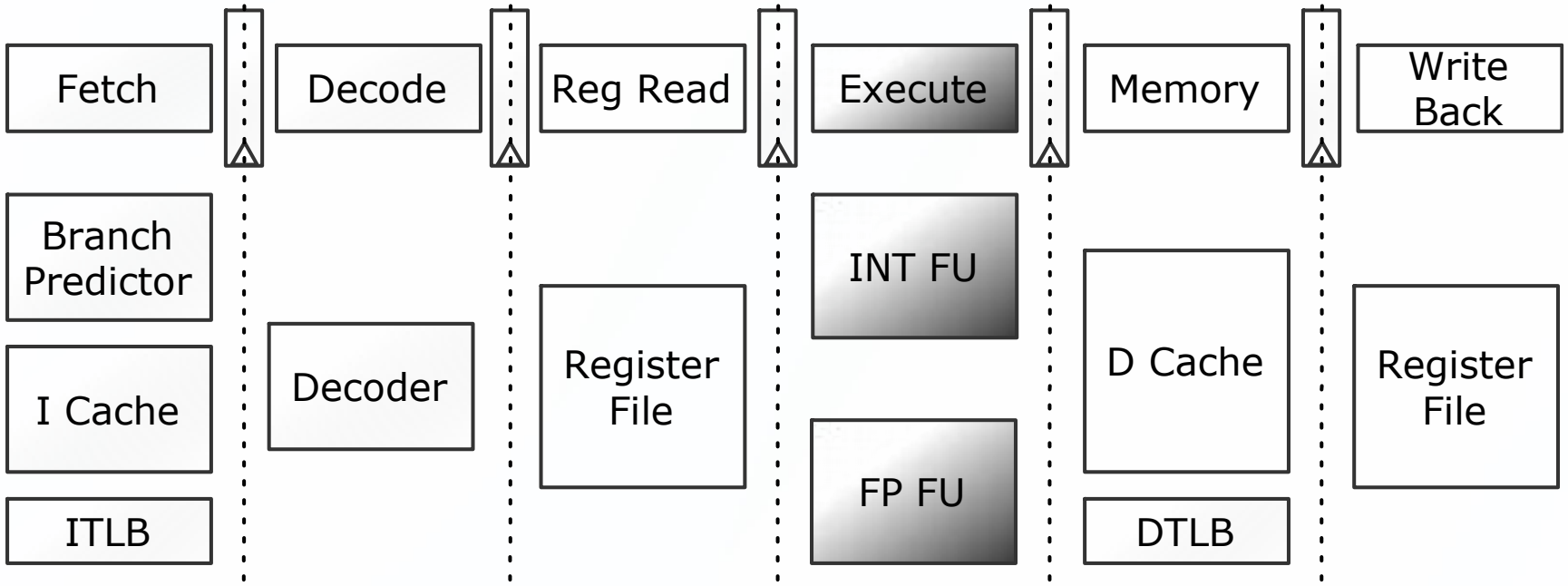
von Neumann architecture

General-purpose processors

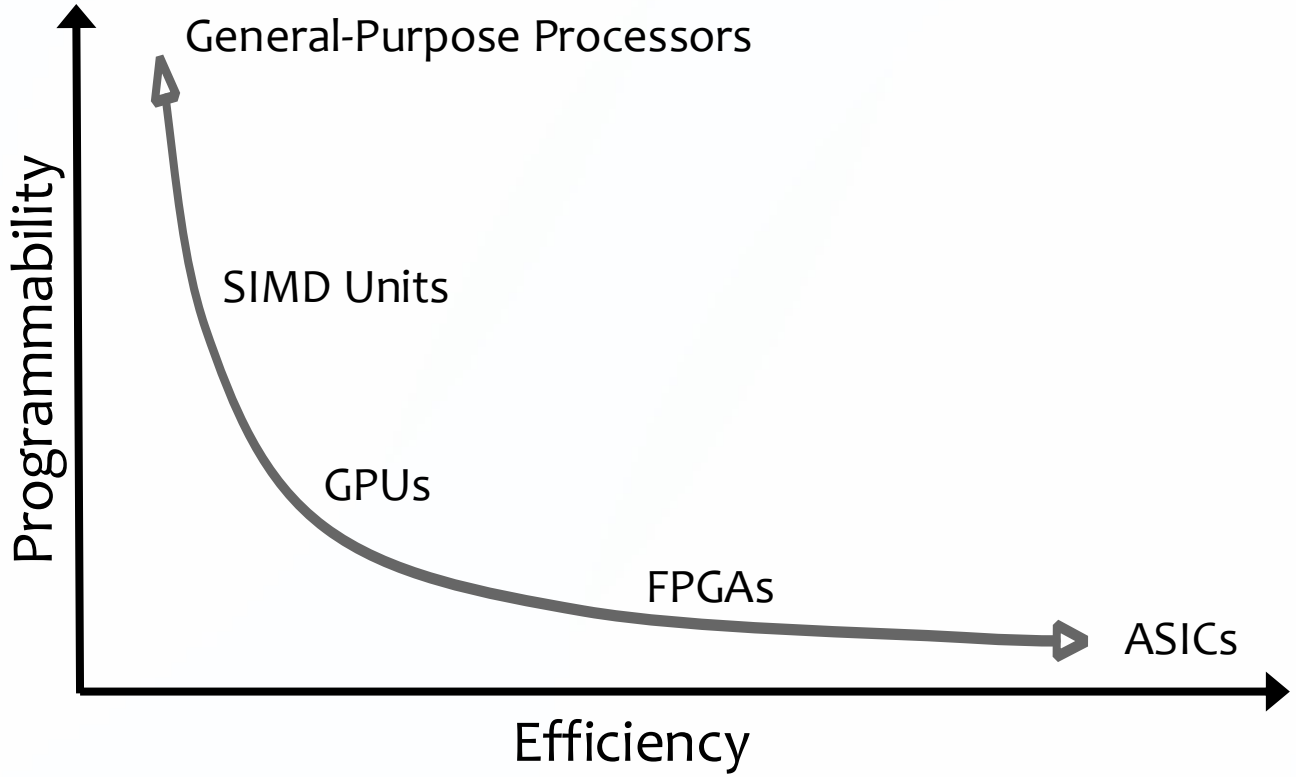
- Components
  - Memory (RAM)
  - Central processing unit (CPU)
    - Control unit
    - Arithmetic logic unit (ALU)
  - Input/output system
- Memory stores program and data
- Program instructions execute sequentially



Programmability versus Efficiency



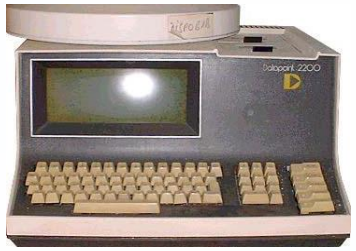
Programmability versus Efficiency



What is the difference between the computing industry and the toothpaste industry?

Industry of Replacement

1971 2024




?

Industry of New Capabilities

Can we continue being an industry of new possibilities?

Personalized healthcare

Virtual reality

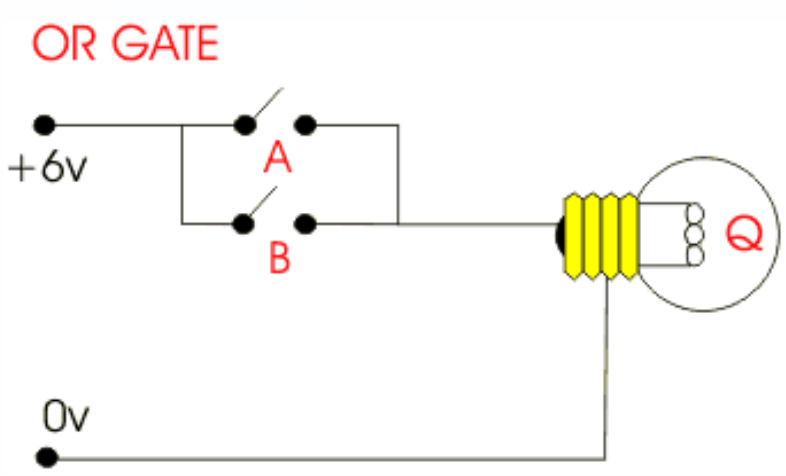
LLMs

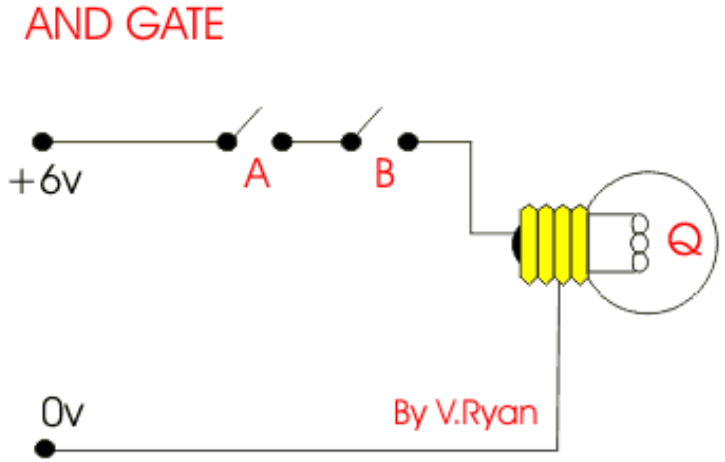
Agenda

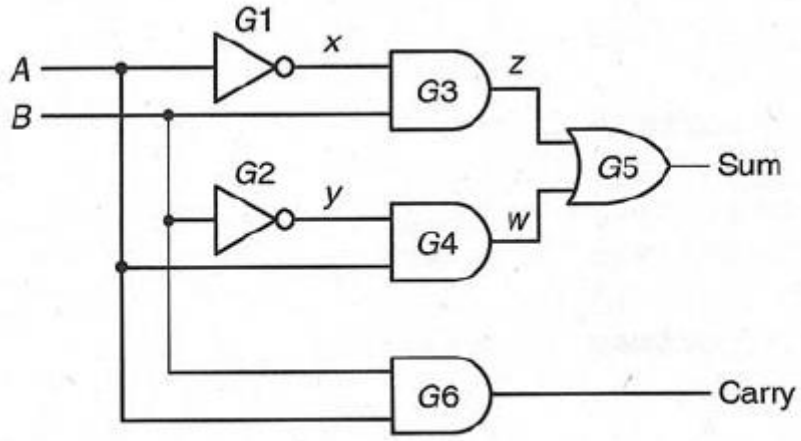
1. Who is Hadi
2. Course organization
3. Why CSE 240C Advanced Microarchitecture
    1. **How we became and industry of new capabilities**
    2. Why we might become an industry of replacement
    3. Specialization and accelerators

Transistors/switches

Building blocks of computing







Logic diagram

Moore's Law

Or, how we became an industry of new capabilities

Every 2 Years

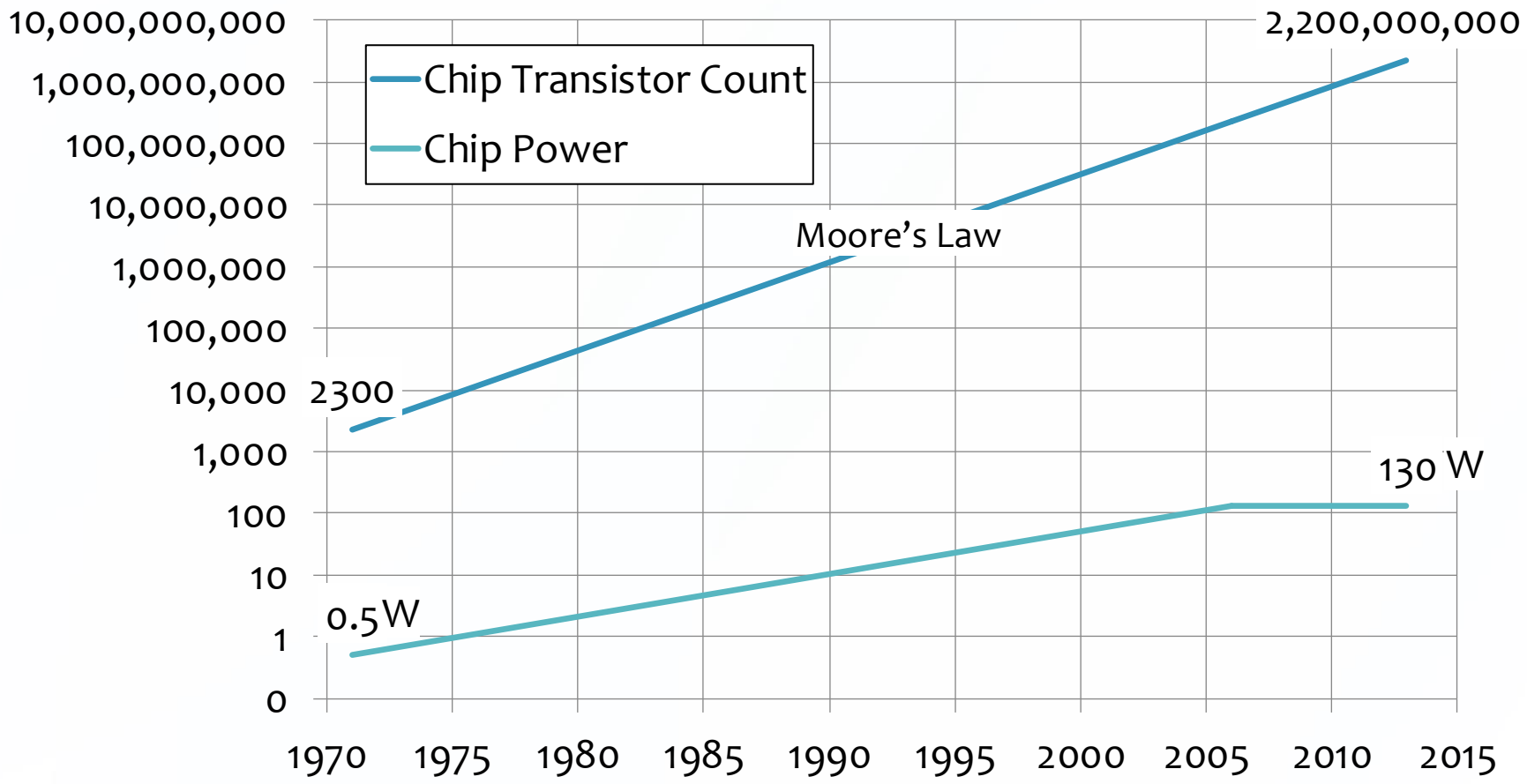
Double the number of transistors

Build higher performance general-purpose processors

- Make the transistors available to masses
- Increase performance (1.8×↑)
- Lower the cost of computing (1.8×↓)

What is the catch?

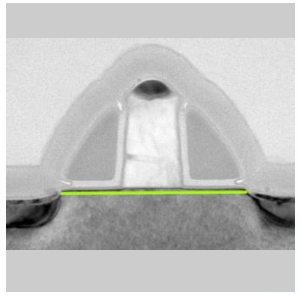
Powering the transistors without melting the chip



Dennard scaling:

Doubling the transistors; scale their power down

Transistor: 2D Voltage-Controlled Switch

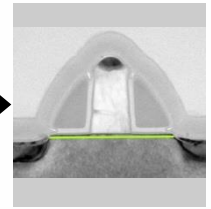


Dimensions

Voltage

Doping Concentrations

×0.7



Area — 0.5×↓

Capacitance — 0.7×↓

Frequency — 1.4×↑

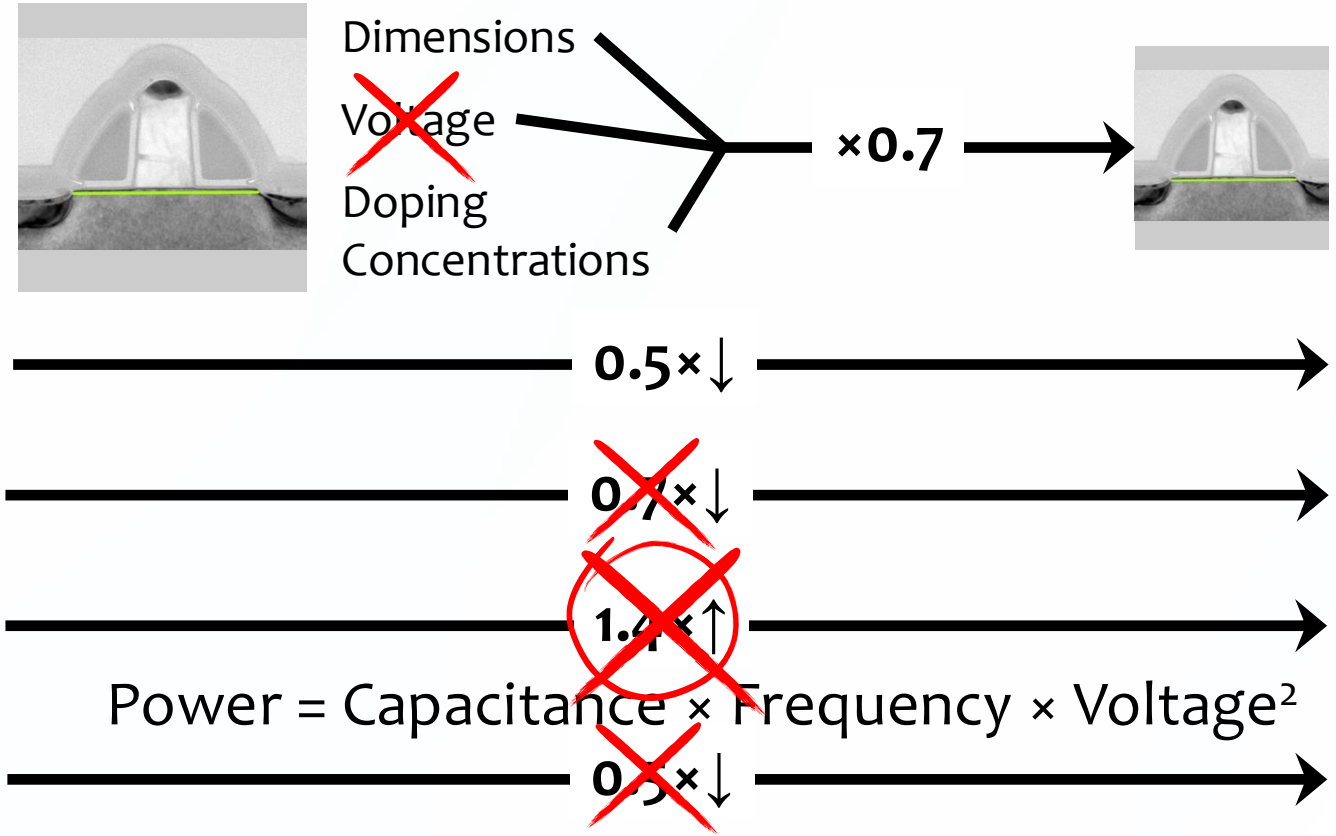
Power = Capacitance × Frequency × Voltage²

Power — 0.5×↓

Dennard Scaling Broke:

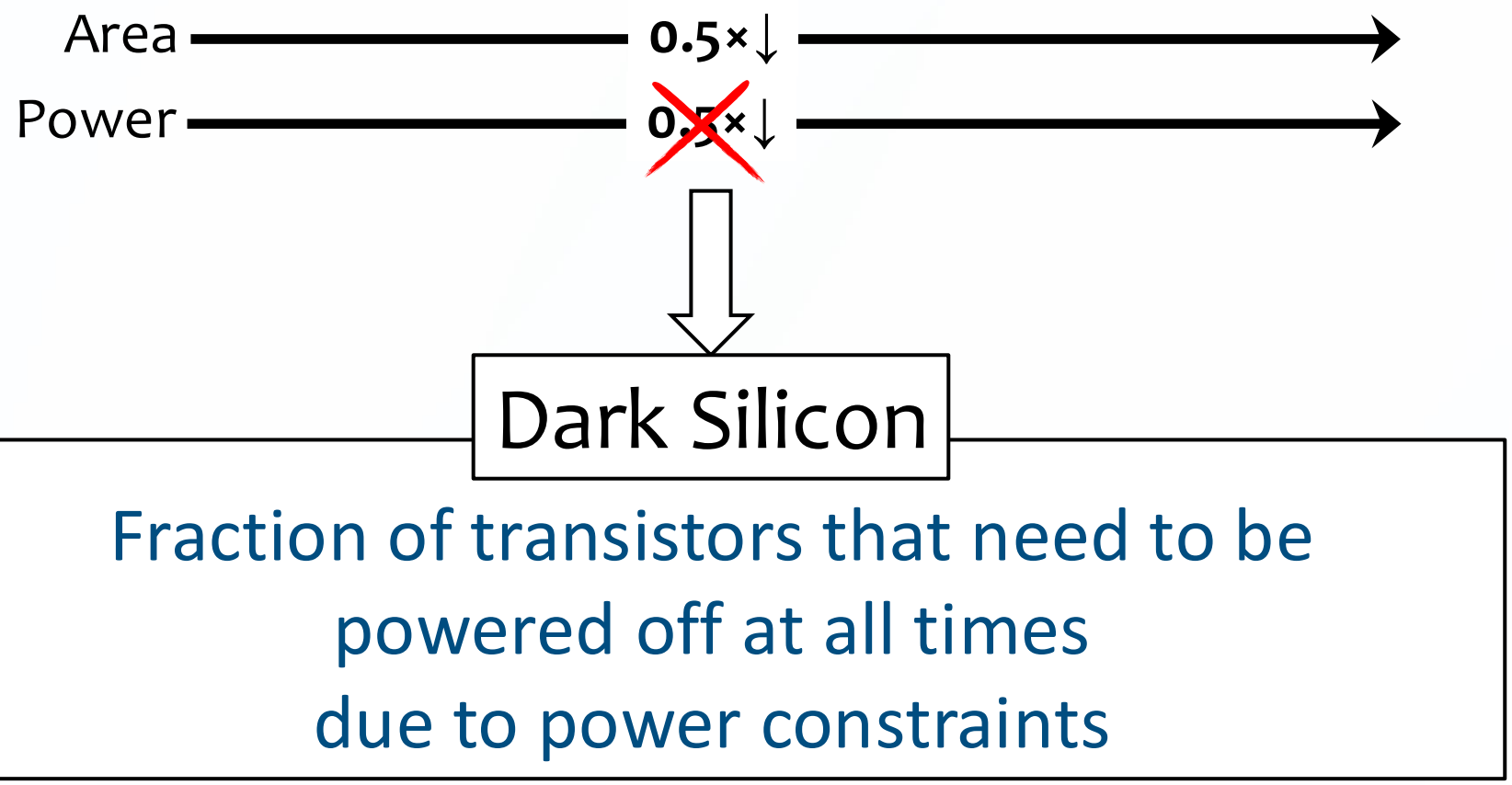
~~Double the transistors; still scale their power down~~

Transistor: 2D Voltage-Controlled Switch

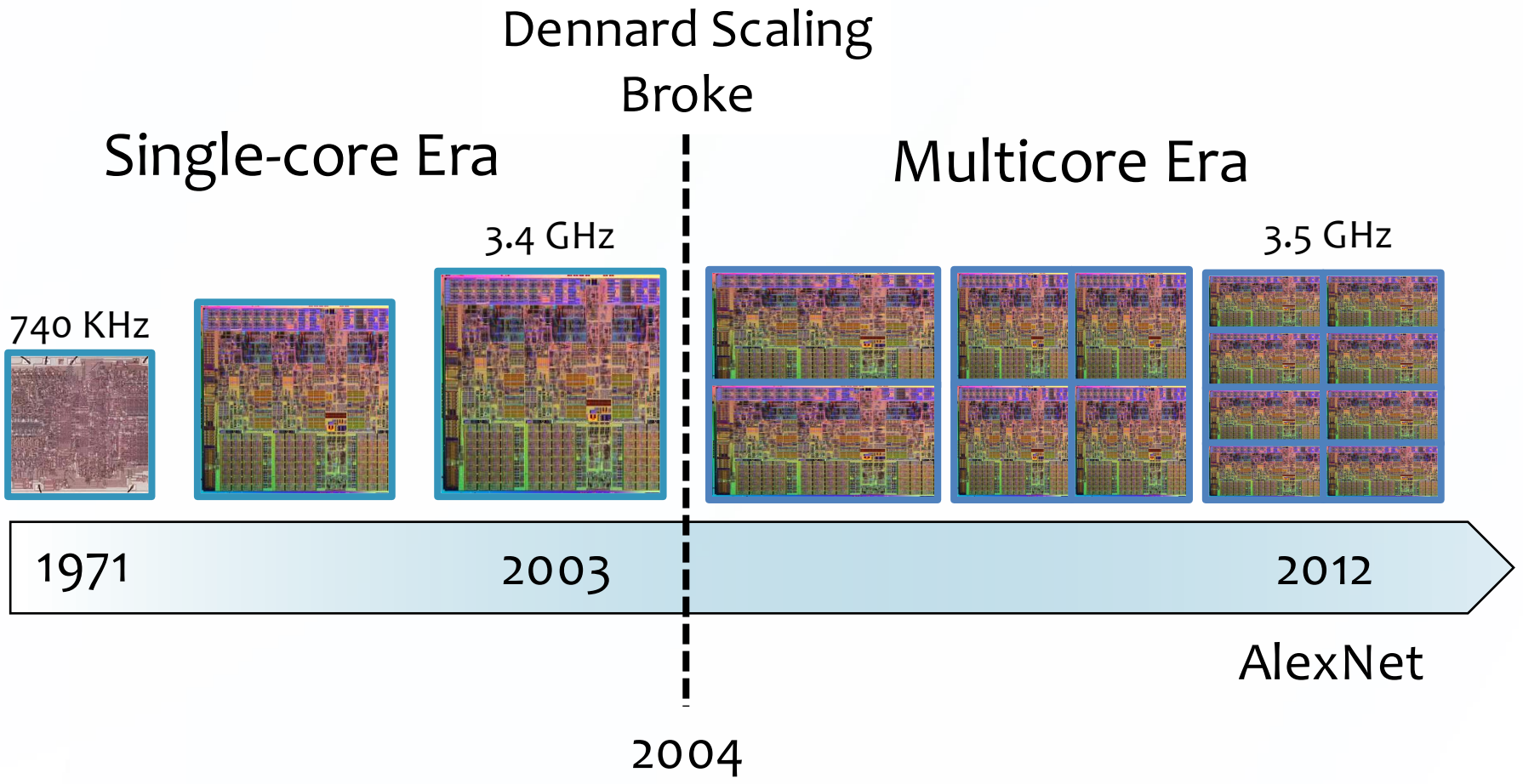


Dark Silicon

If you cannot power them, why bother making them?



Looking Back
Evolution of processors



Are multicores a long-term solution or just a stopgap?

Agenda

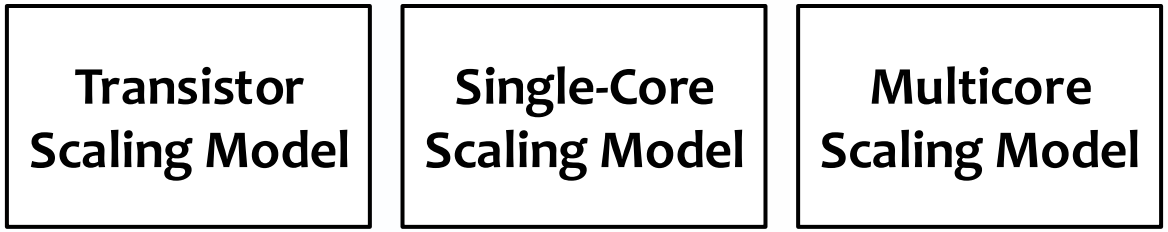
1. Who is Hadi
2. Course organization
3. Why CSE 240C Advanced Microarchitecture
    1. How we became and industry of new capabilities
    2. **Why we might become an industry of replacement**
    3. Specialization and accelerators

Modeling future multicores

Quantify the severity of the problem

Predict the performance of best-case multicores

- From 45 nm to 8 nm
- Parallel benchmarks
- Fixed power and area budget



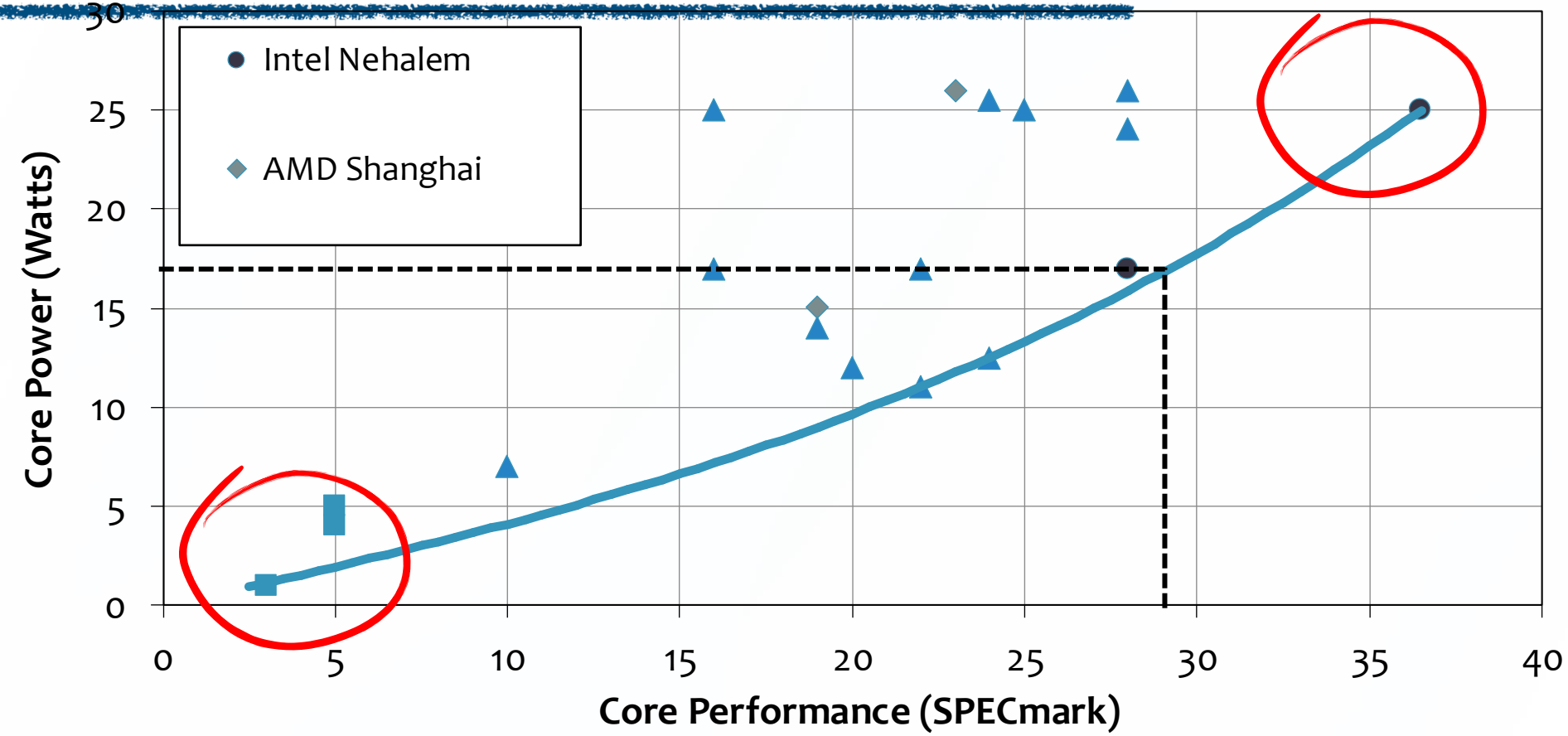
Esmaeilzadeh, Belem, St. Amant, Sankaralingam, Burger, "Dark Silicon and the End of Multicore Scaling," ISCA 2011

Transistor scaling model

From 45 nm to 8 nm

	[Dennard, 1974]	[ITRS, 2010]	[VLSI-DAT, 2010]
	Historical Scaling	**Optimistic Scaling Model**	**Conservative Scaling Model**
Area	32× ↓	32× ↓	32× ↓
Power	32× ↓	8.3× ↓	4.5× ↓
Speed	5.7× ↑	3.9× ↑	1.3× ↑

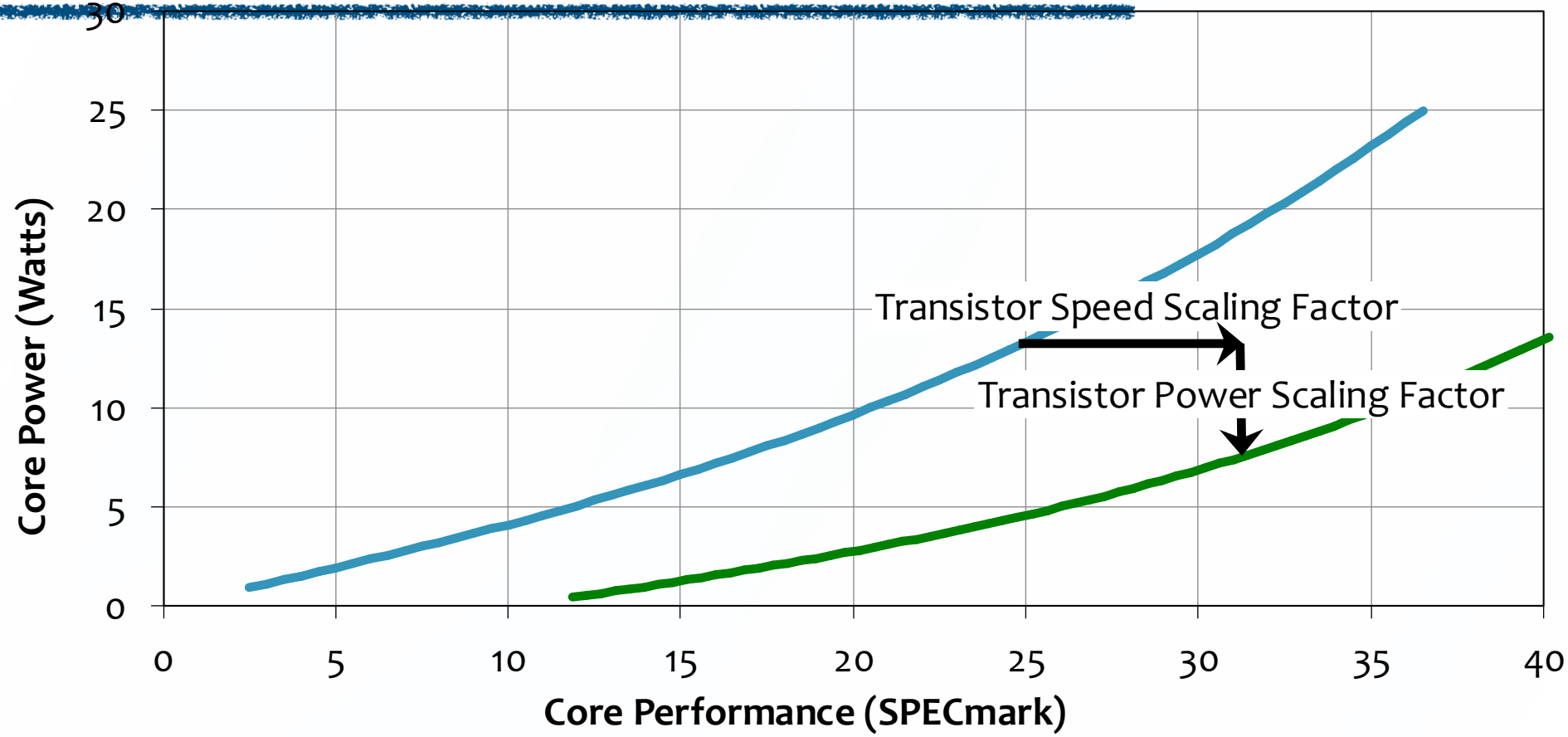
Single-core model (45 nm)



Power-Performance and Area-Performance Pareto Optimal Frontiers

Single-core scaling model

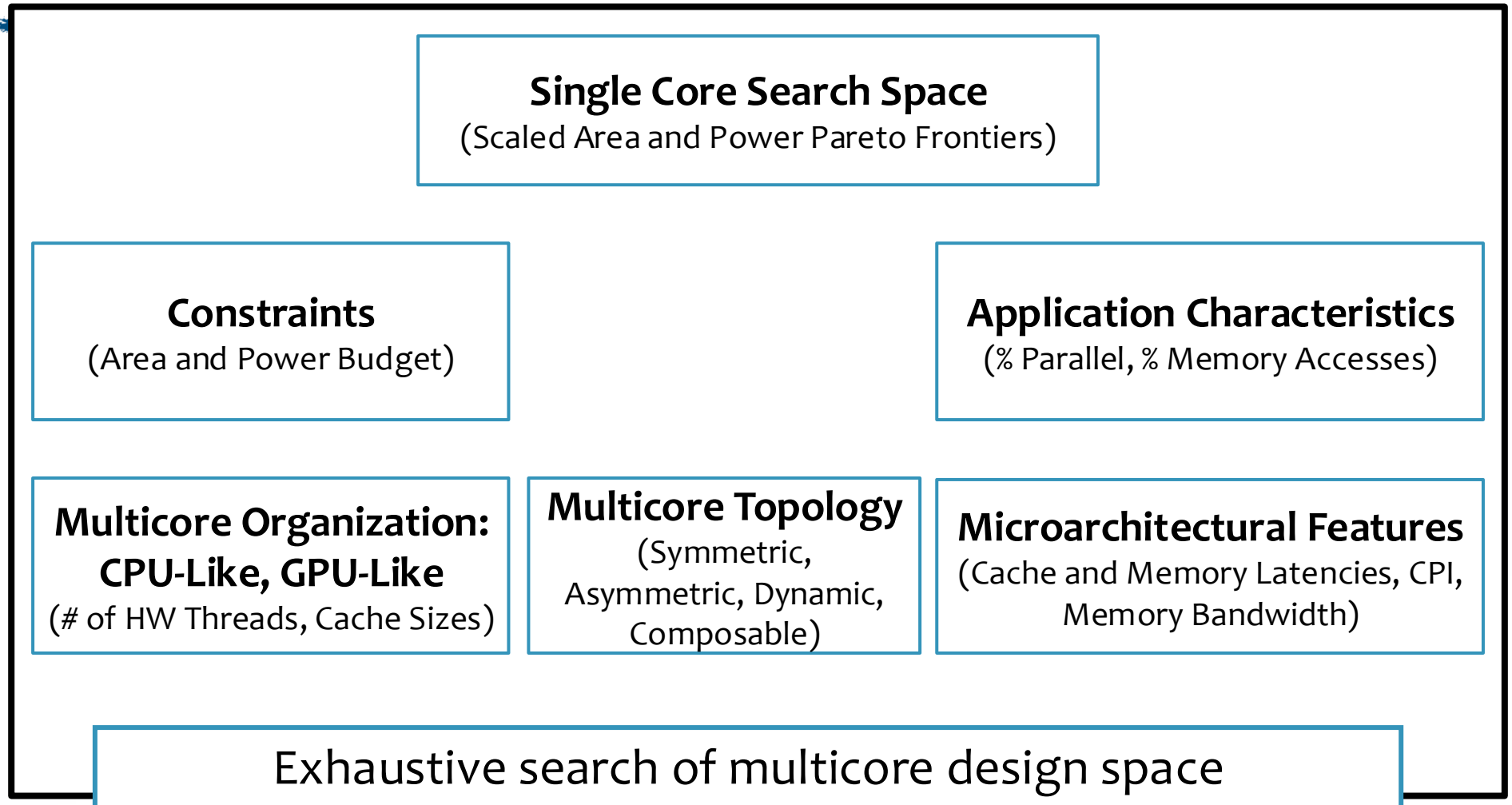
From 45 nm to 8 nm



Single-core Scaling Model:
Single-core Model × Transistor Scaling Model

Multicore scaling model

From 45 nm to 8 nm



Multicore model (Amdahl's Law)

$$\text{Speedup} = \frac{1}{\frac{1-f_{Parallel}}{\text{Serial Speedup}} + \frac{f_{Parallel}}{\text{Parallel Speedup}}}$$

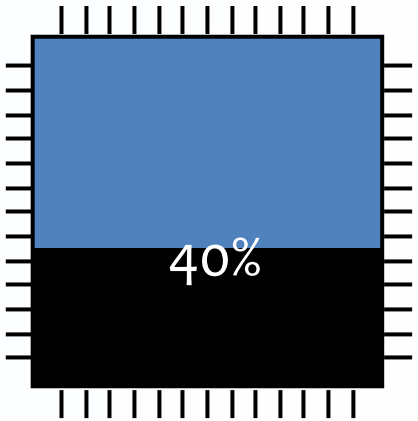
$$\text{Serial Speedup} = 1 \times \text{Core Performance}$$

$$\text{Parallel Speedup} = \text{N} \times \text{Core Performance}$$

Dark silicon

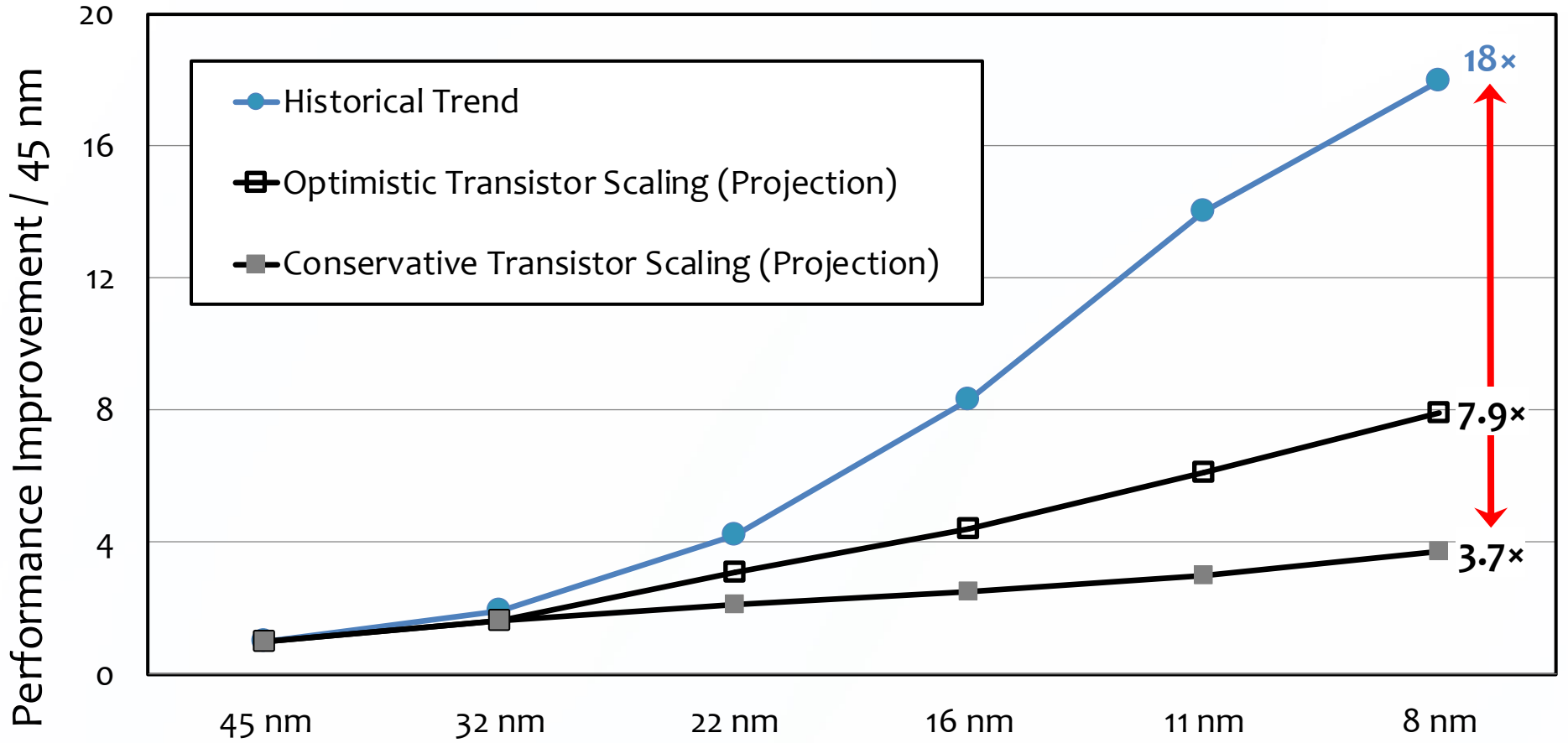
$$N_{Core} = min(\frac{Area\ Budget}{Area_{Core}}, \frac{Power\ Budget}{Power_{Core}})$$

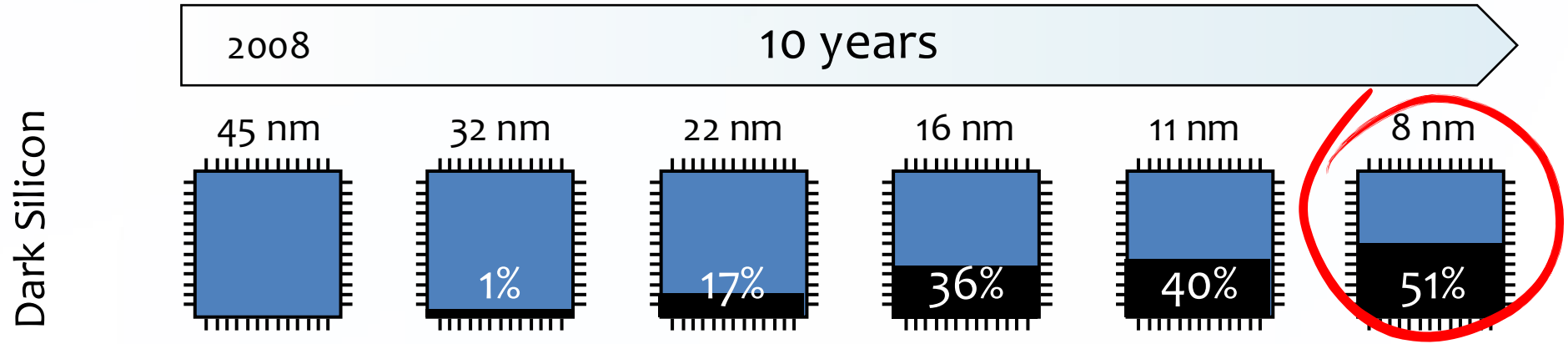
$$Dark\ Silicon = 1 - \frac{N_{Core} \times Area_{Core}}{AreaBudget}$$



Evaluation Setup

- Applications:
  - 12 PARSEC Parallel Benchmarks
- Baseline:
  - The best multicore design available at 45 nm
- Constraints:
  - Driven from the best multicore design at 45 nm
    - Fixed Power Budget: 125 W
    - Fixed Area Budget: 111 mm^2


Performance Improvement / 45 nm
20
16
12
8
4
0
Historical Trend
Optimistic Transistor Scaling (Projection)
Conservative Transistor Scaling (Projection)
18×
7.9×
3.7×
45 nm
32 nm
22 nm
16 nm
11 nm
8 nm


Dark Silicon
2008
10 years
45 nm
32 nm
22 nm
16 nm
11 nm
8 nm
1%
17%
36%
40%
51%

The Shift Towards Domains-Specific Accelerators

Esmaeilzadeh et al. “Dark Silicon and the End of Multi-Core Scaling,” ISCA 2011

CACM Research Highlight

IEEE Micro Top Picks

The New York Times

Progress Hits Snag: Tiny Chips Use Outsize Power

Share full article 78

By John Markoff

July 31, 2011

For decades, the power of computers has grown at a staggering rate as designers have managed to squeeze ever more and ever tinier transistors onto a silicon chip — doubling the number every two years, on average, and leading the way to increasingly powerful and inexpensive personal computers, laptops and smartphones.

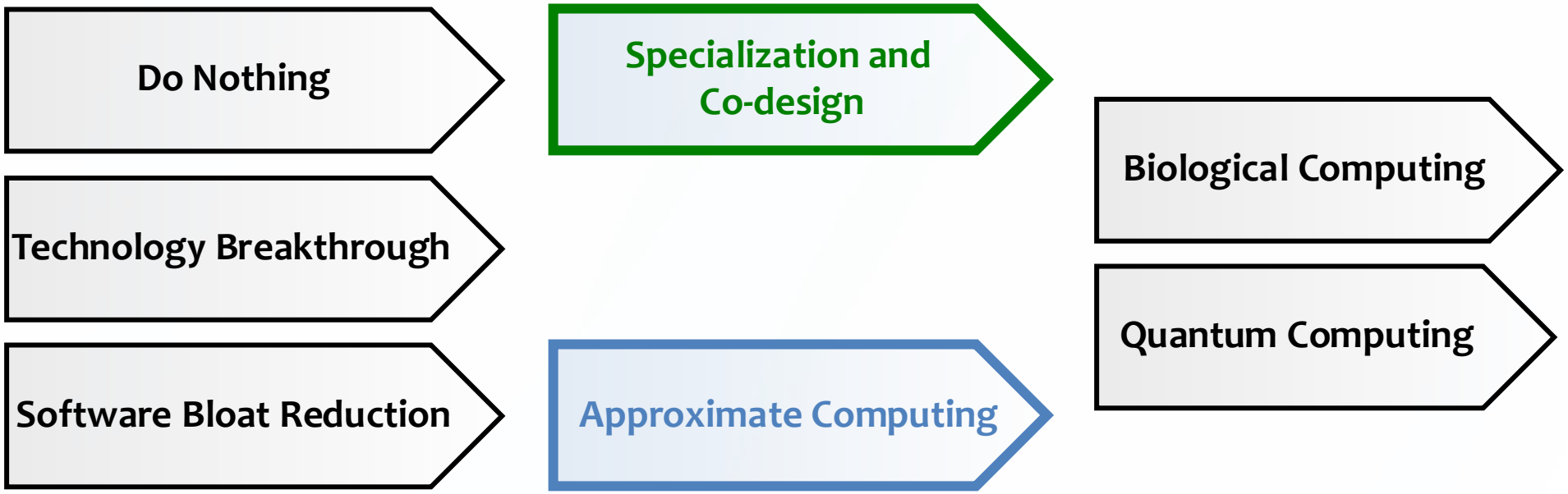
Industry of replacement?

- Multicores are likely to be a stopgap
  - Not likely to continue the historical trends
  - Do not overcome the transistor scaling trends
  - The performance gap is significantly large
- Radical departures from conventional approaches are necessary
  - Extract more performance and efficiency from silicon while preserving programmability
  - Explore other sources of computing

Agenda

1. Who is Hadi
2. Course organization
3. Why CSE 240C Advanced Microarchitecture
    1. How we became and industry of new capabilities
    2. Why we might become an industry of replacement
    3. **Specialization and accelerators**

Possible paths forward

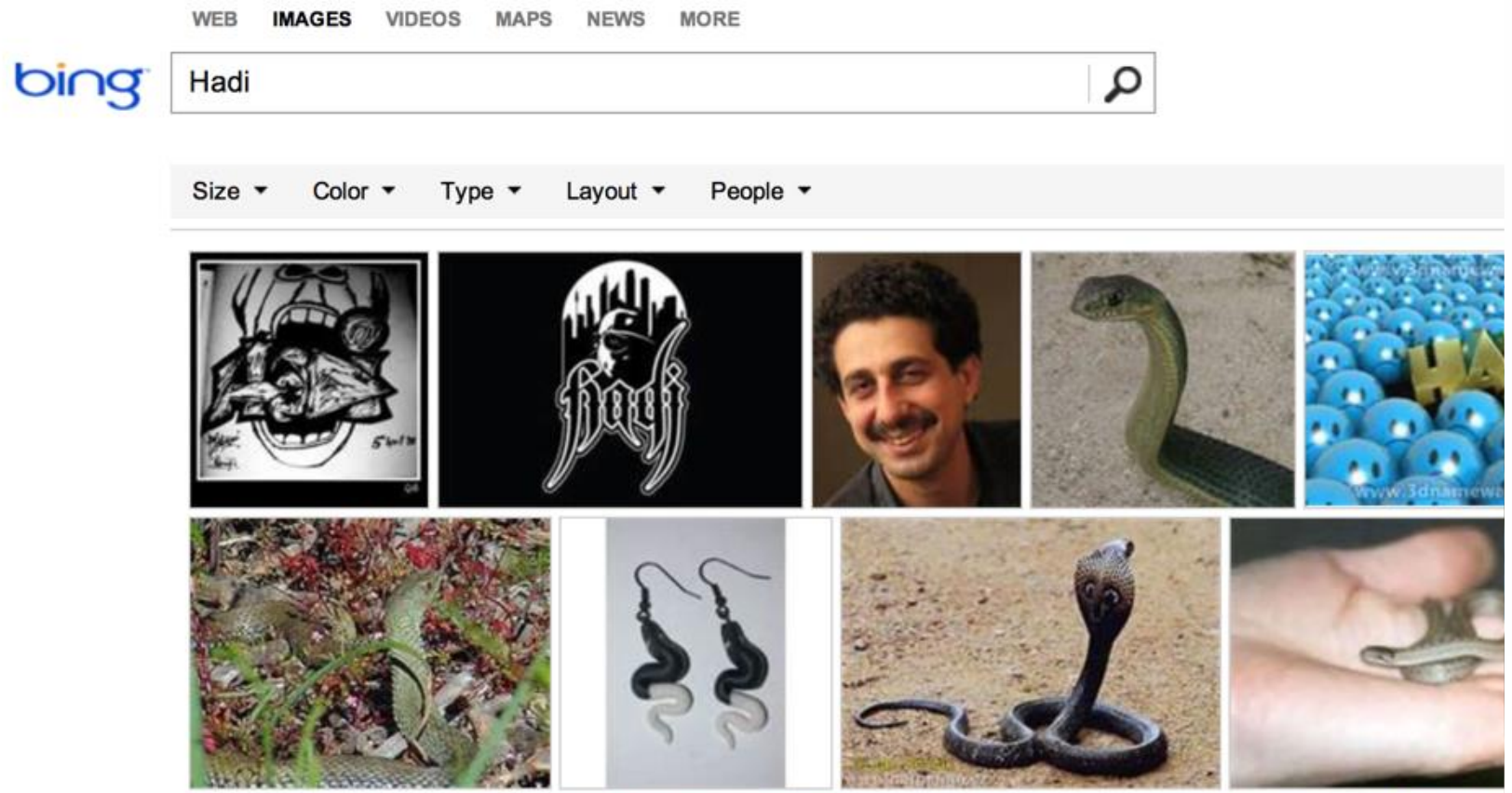


Approximate computing
Embracing error

- Relax the abstraction of near-perfect accuracy in general-purpose computing

- Allow errors to happen in the computation
  - Run faster
  - Run more efficiently


Insta


WEB
IMAGES
VIDEOS
MAPS
NEWS
MORE
bing
Hadi
Size
Color
Type
Layout
People

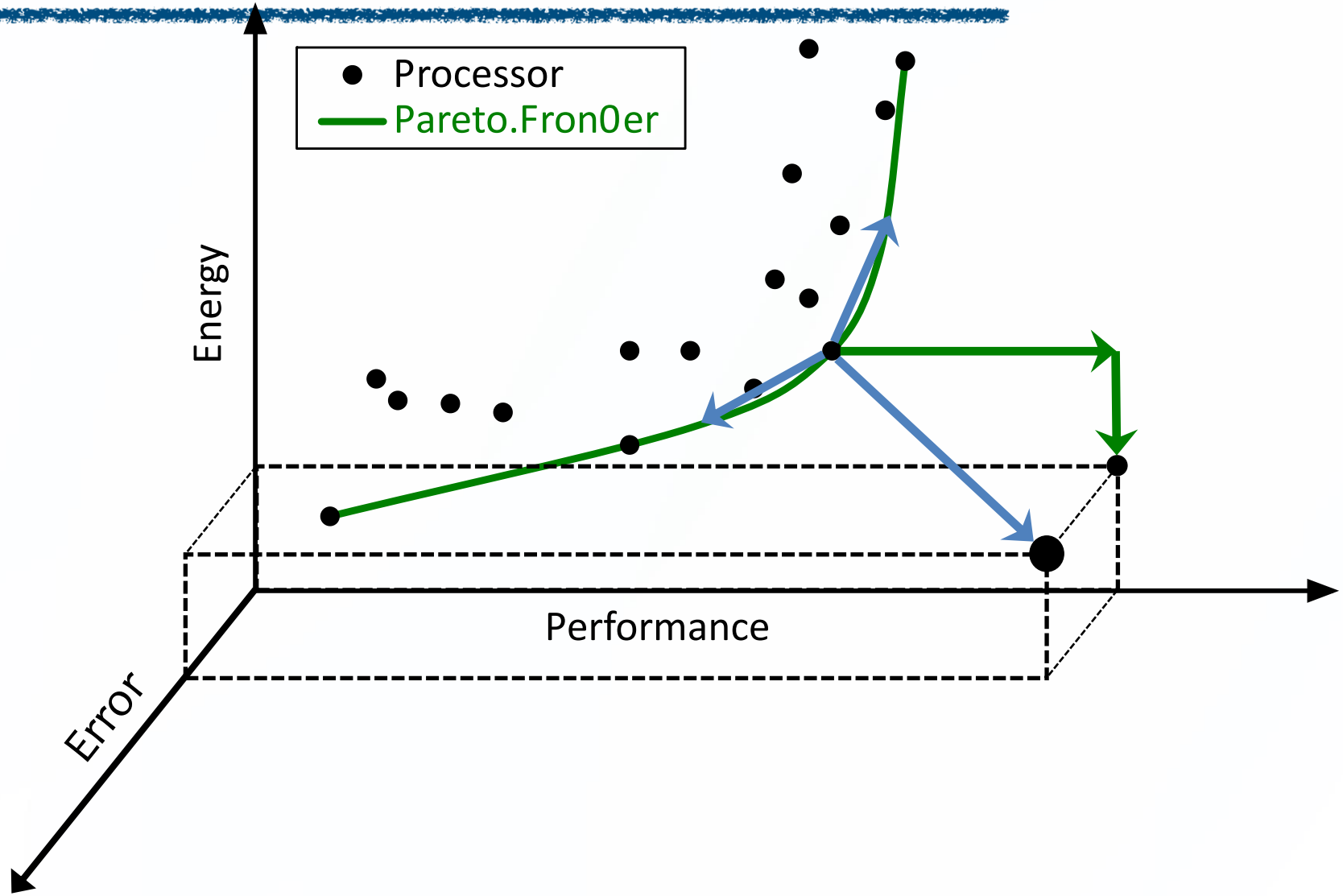
New landscape of computing

Personalized and targeted computing

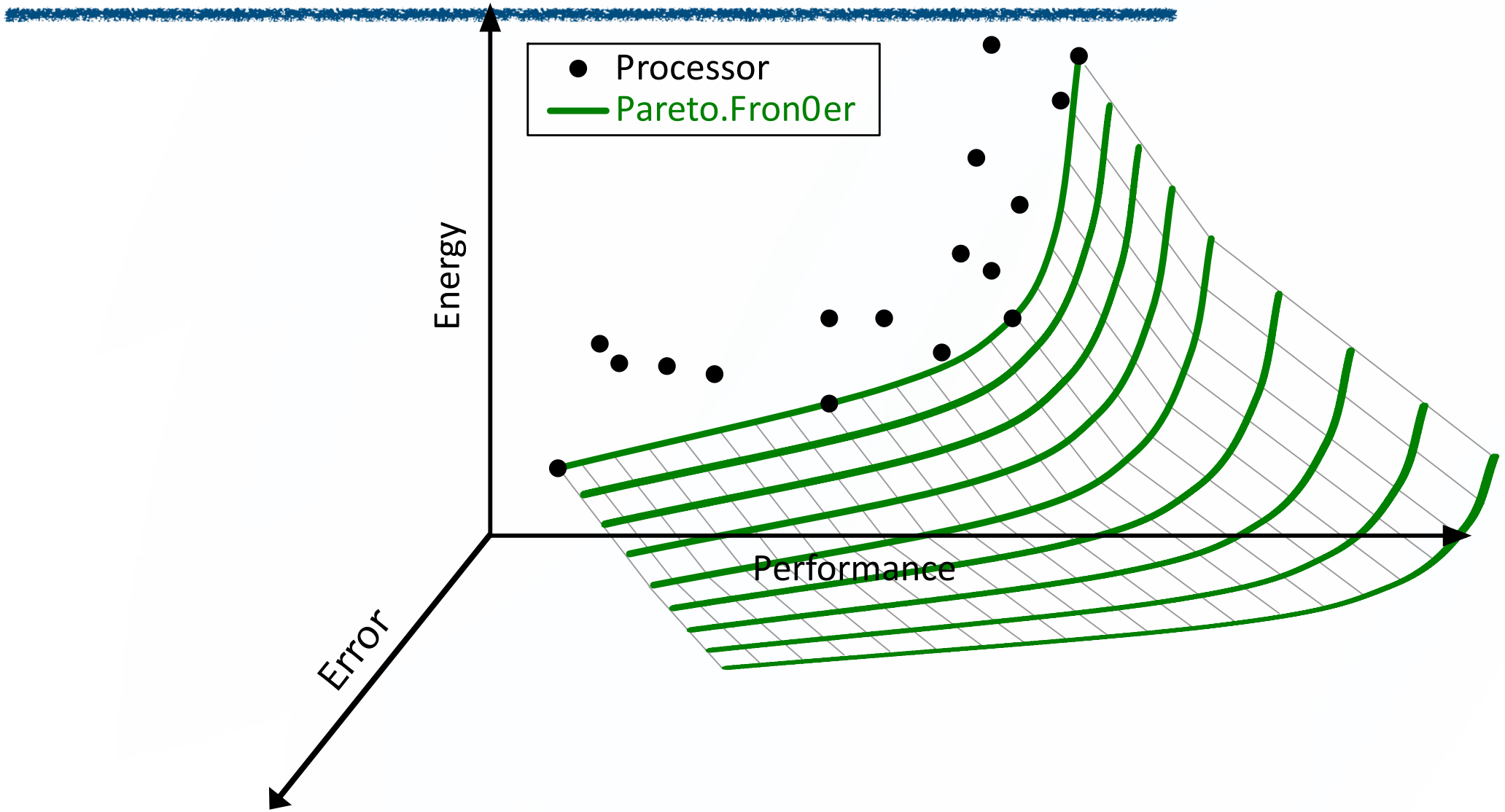
Classes of approximate applications

- Programs with analog inputs
  - Sensors, scene reconstruction
- Programs with analog outputs
  - Multimedia
- Programs with multiple possible answers
  - Web search, machine learning
- Convergent programs
  - Gradient descent, big data analytics

Adding a third dimension
Embracing Error



A fertile ground for innovation

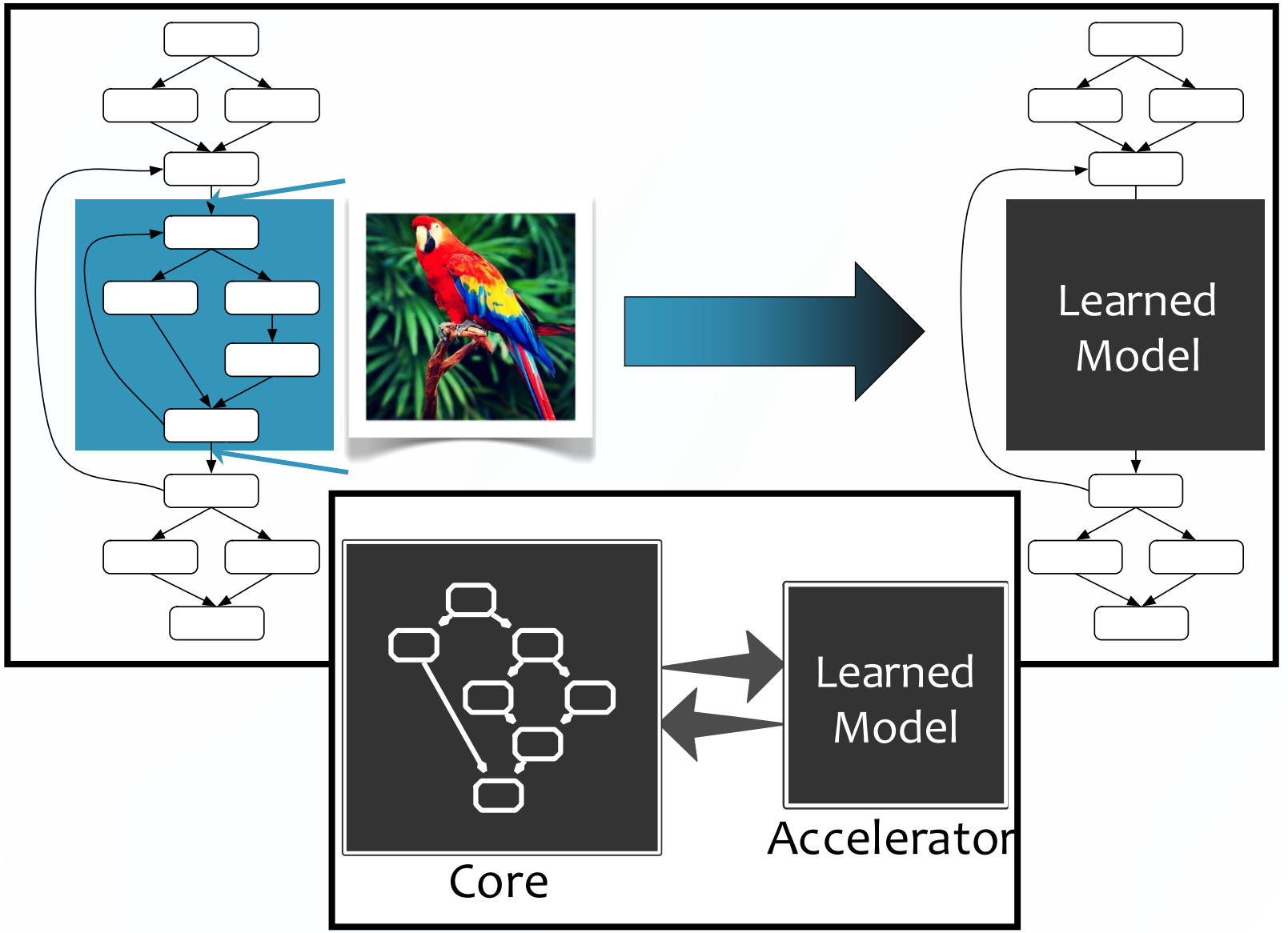


Approximate computing techniques

Same Model	From Model to Model

• Sampling
  • Loop perforation (MIT)
• Compression
  • Sage (Michigan)
• Early termination
  • Green (MSR)
• Replacement
  • Green (MSR)
• Lower voltage
  • Truffle (Rice, UW)
	• von Neumann to Neural
	  • NPUs (UW, GaTech, UCSD)

Parrot Algorithmic Transformation



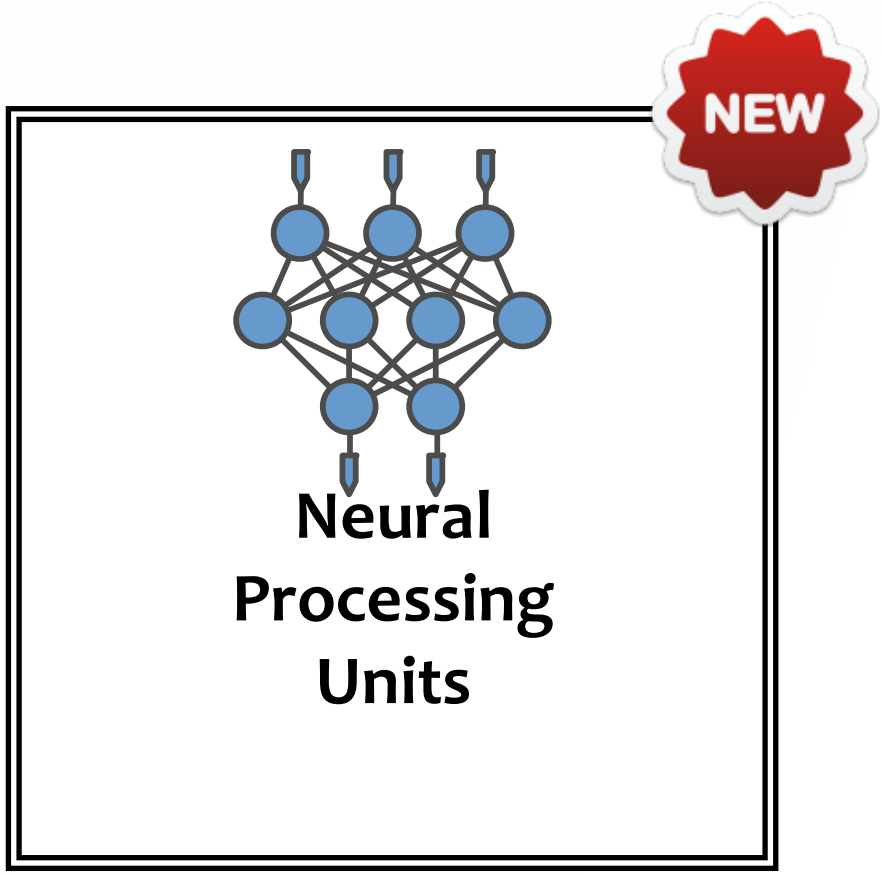
Neural Networks for Code Approximation

Powerful prediction tools

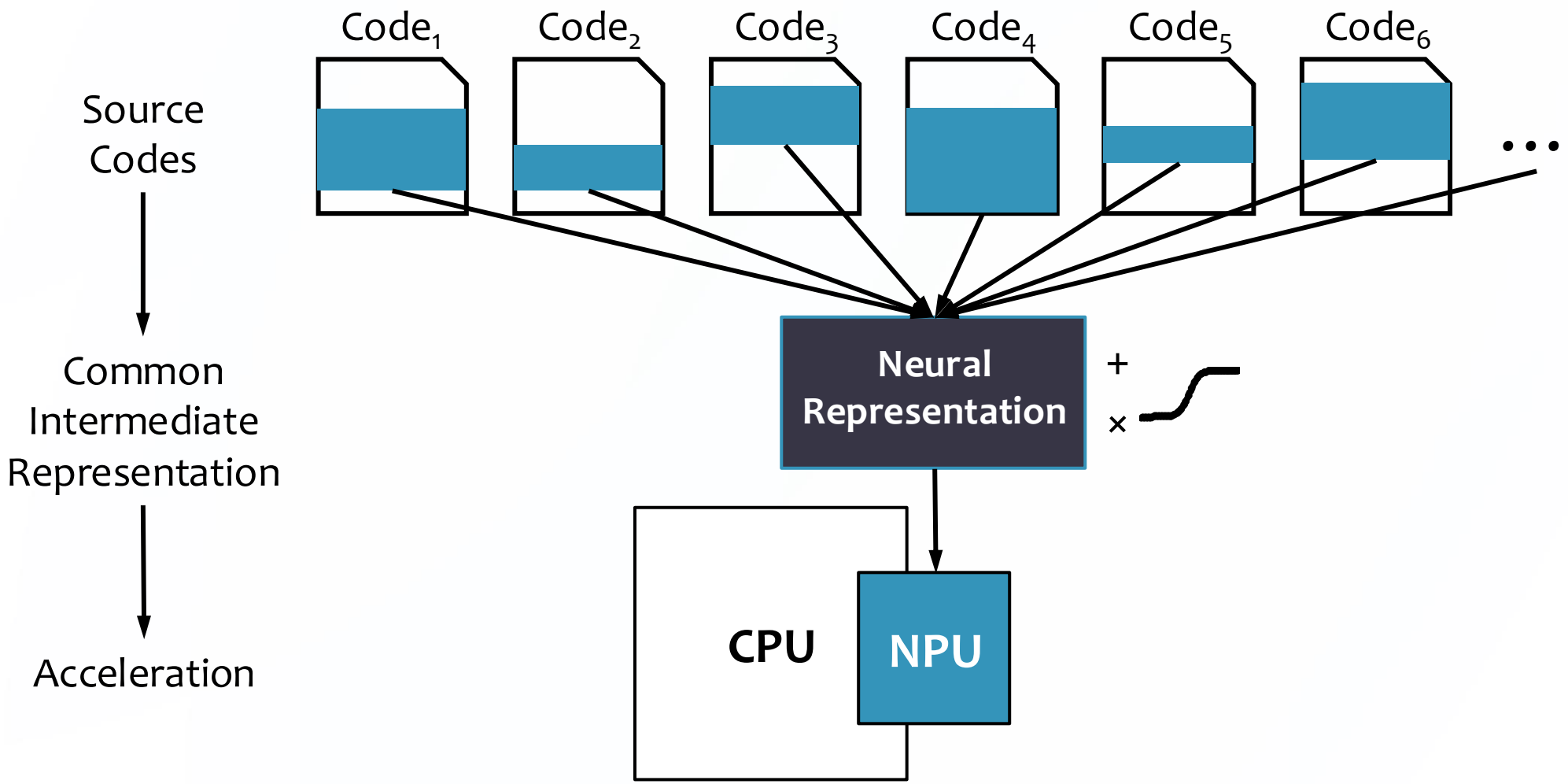
Highly parallel

Efficiently implementable with both digital and analog hardware

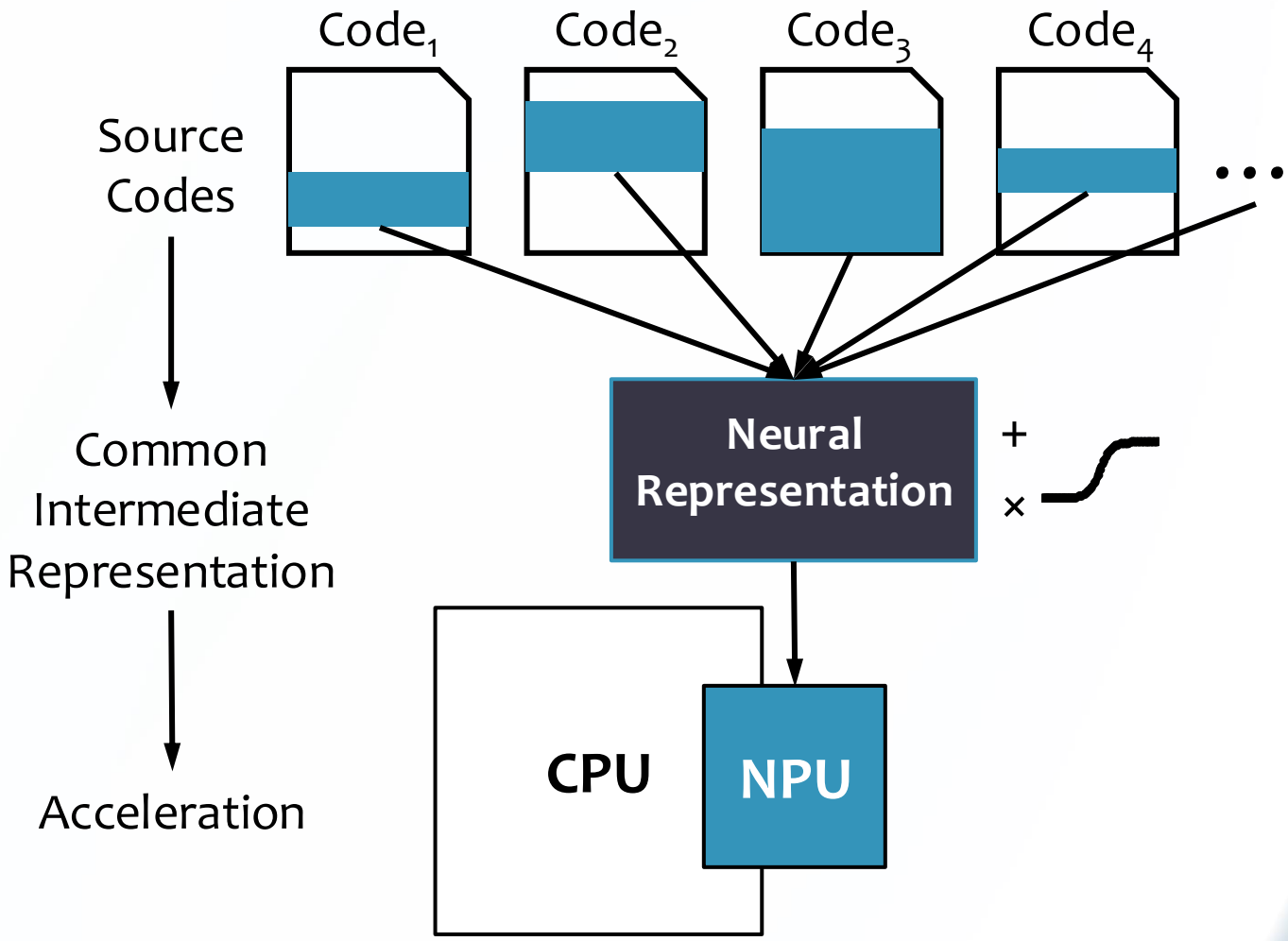
Fault tolerant



NPU Acceleration



Neural Processing Units (NPUs)

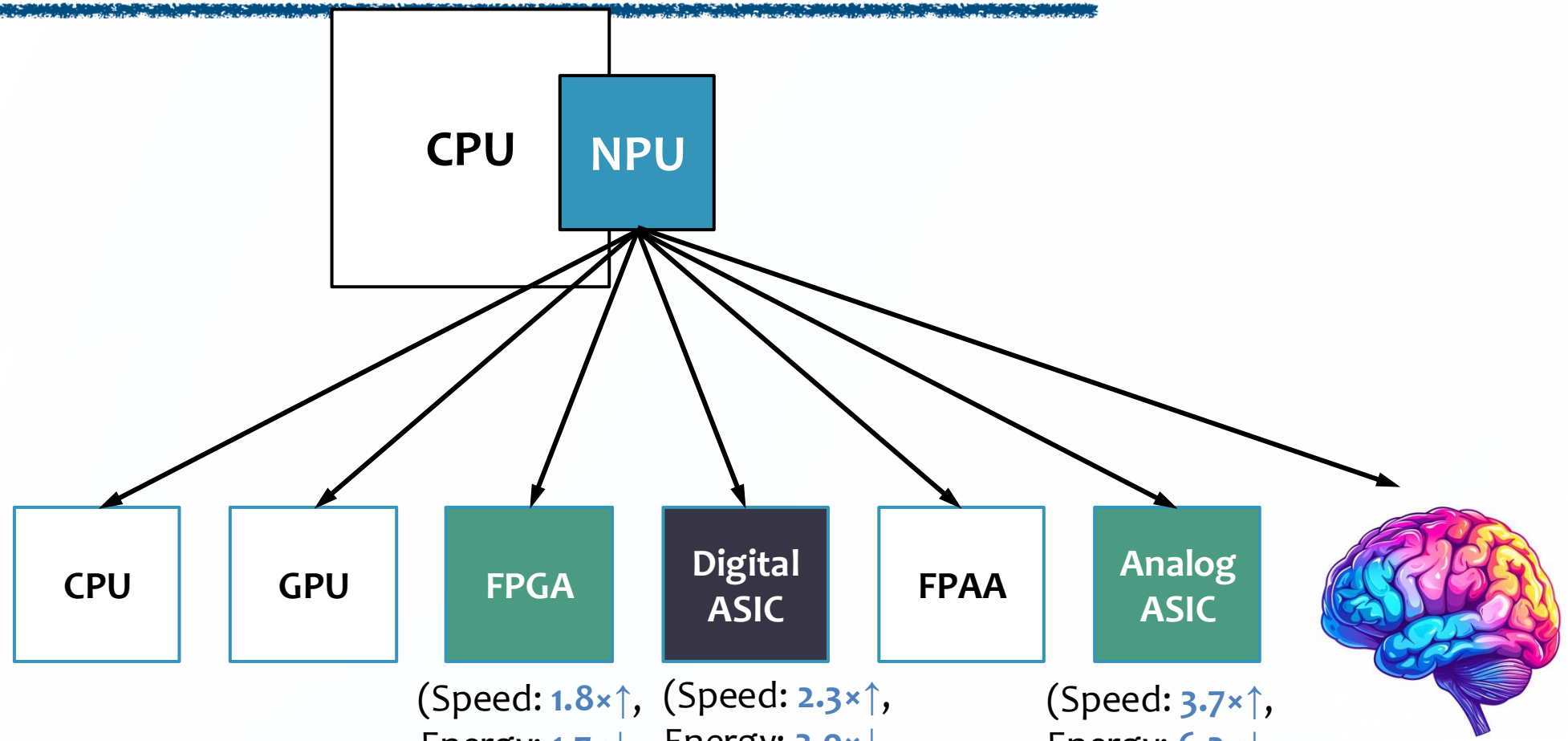


Esmaeilzadeh et al. "Neural Acceleration for General-Purpose Approximate Programs," Micro 2012

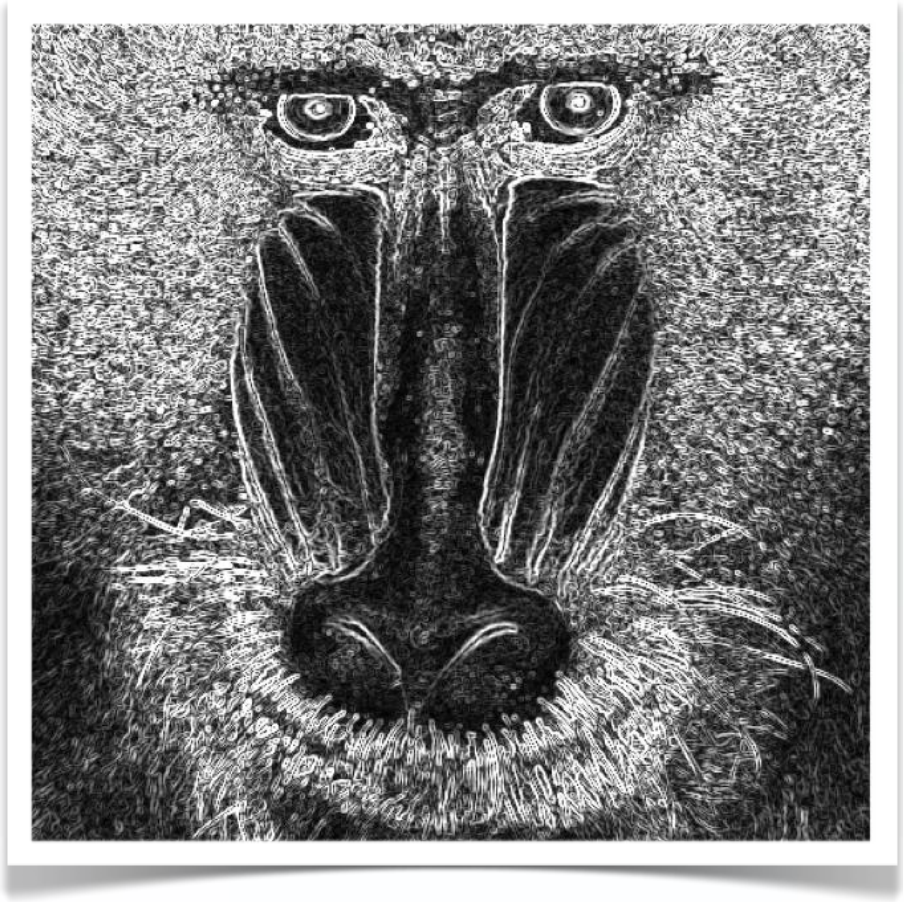
CACM Research Highlights

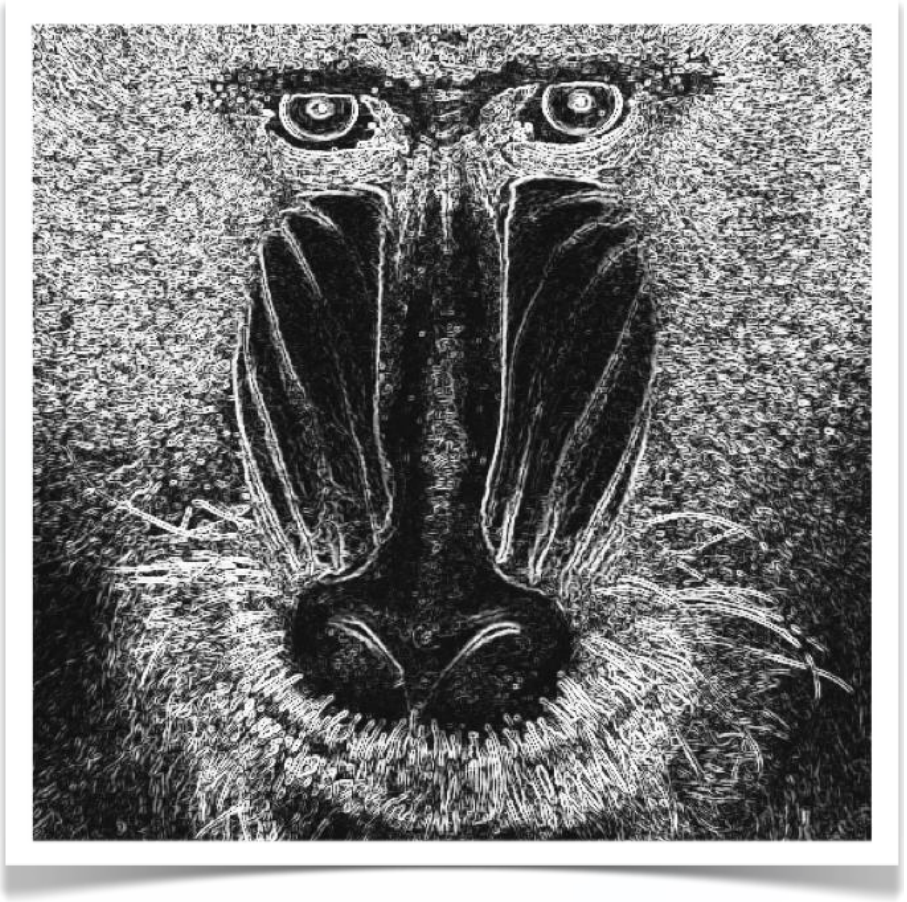
IEEE Micro Top Picks

NPU design alternatives



Approximate Computing versus Conventional Computing

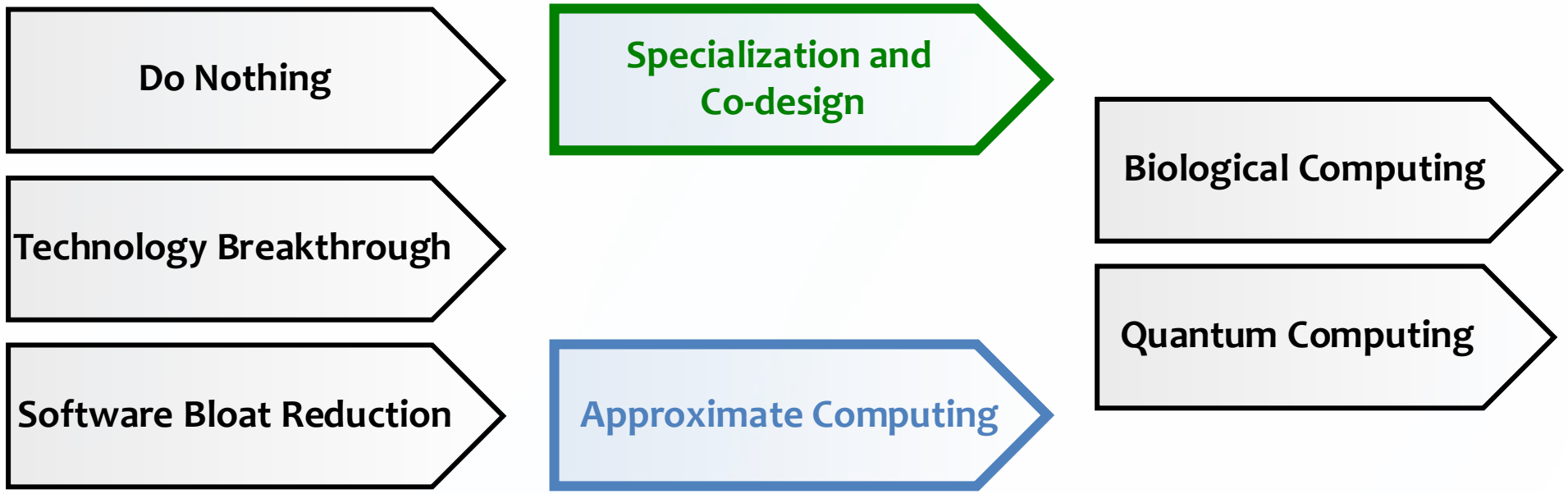




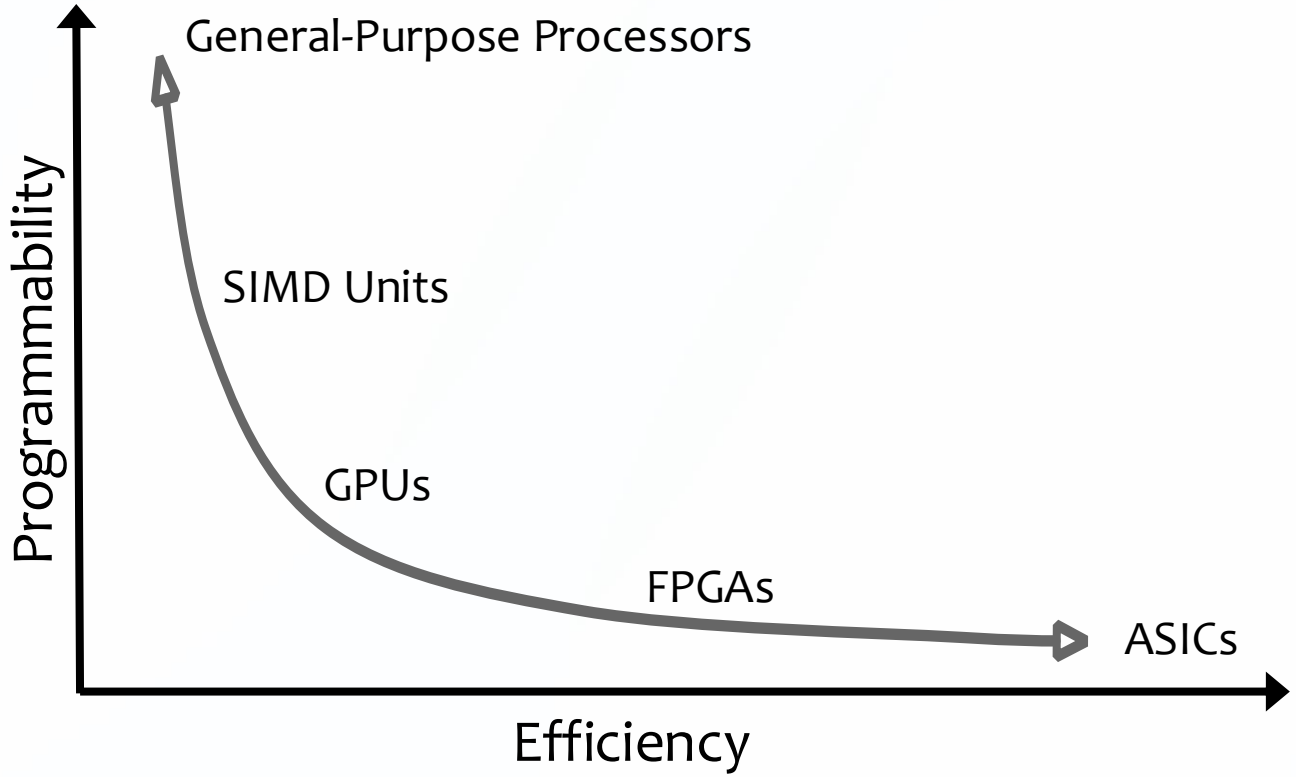
Analog Computing: Computing with Physics

http://youtu.be/dAyDi1aa40E

Possible paths forward



Programmability versus Efficiency




IMAGES
VIDEOS
MAPS
NEWS
SEARCH HISTORY
MORE
MSN
OUTLOOK.COM
Make Bing my homepage
2
250
Sign in
bing
Large-Scale Reconfigurable Computing in a Microsoft Datacenter
Hot Chips '26 – Aug '24
© Image Credits
© 2014 Microsoft | Privacy and Cookies | Legal | Advertise | About our ads | Help | Feedback

Microsoft Cloud Services

Windows Azure

bing

Microsoft SQL Server

XBOX LIVE

COSMOS

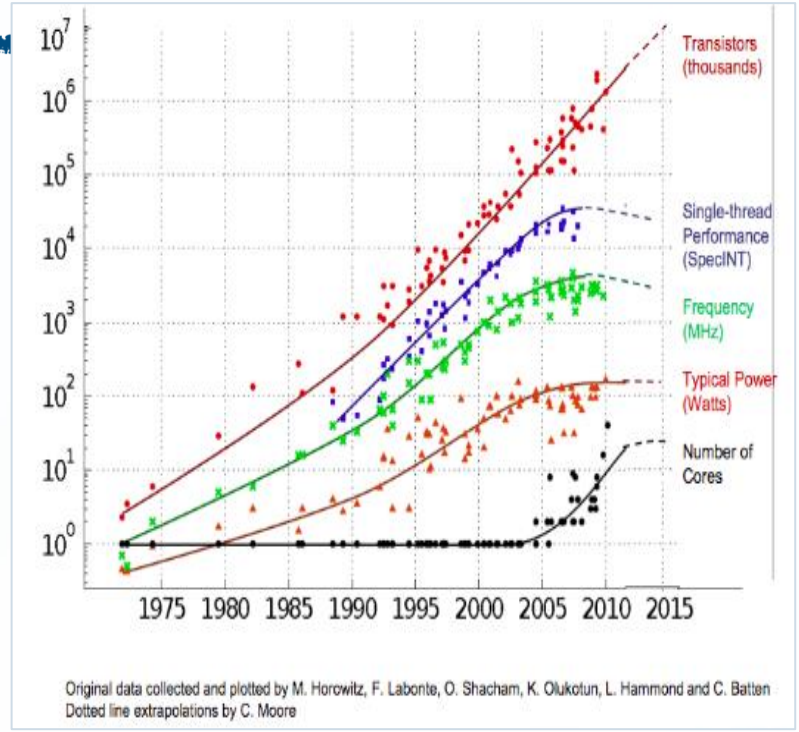
skype

msn

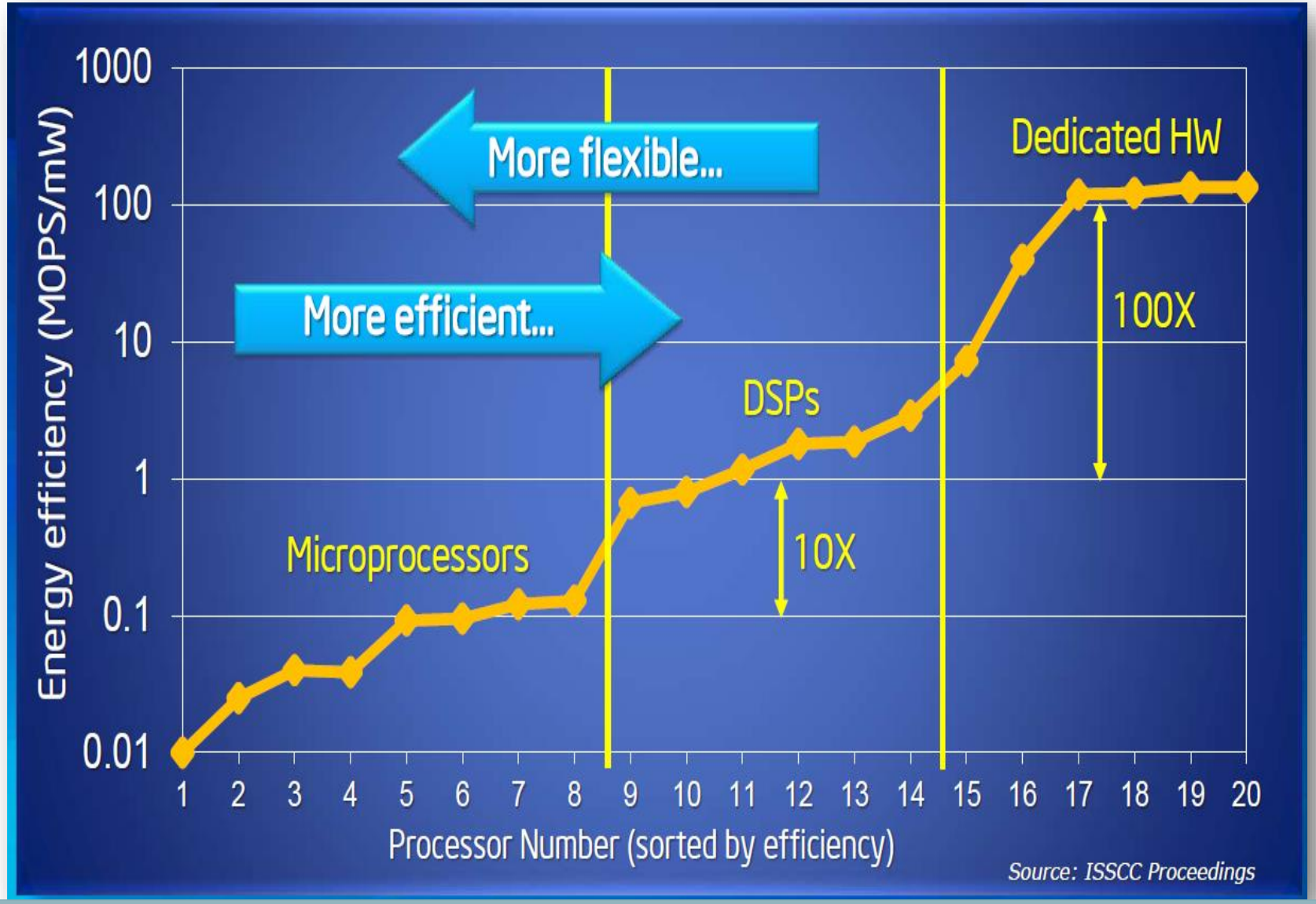
OneDrive

Outlook.com

Microsoft Office 365

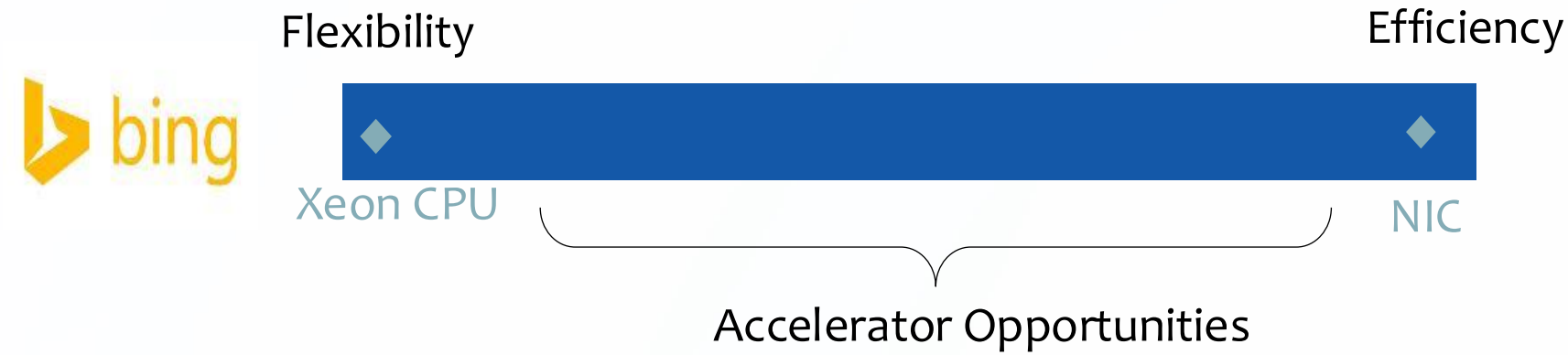


Capabilities, Costs

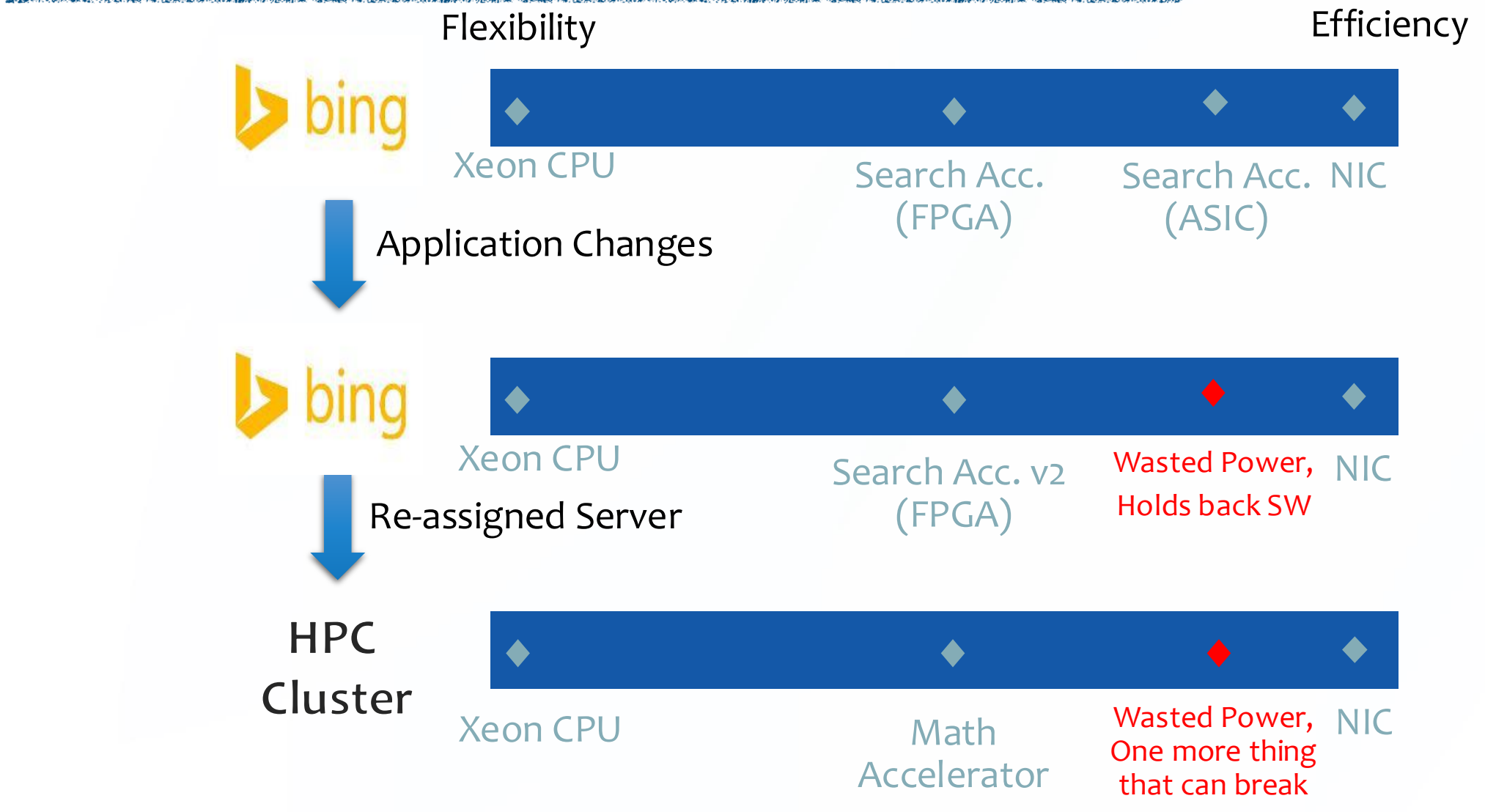


Increase Efficiency with Hardware Specialization

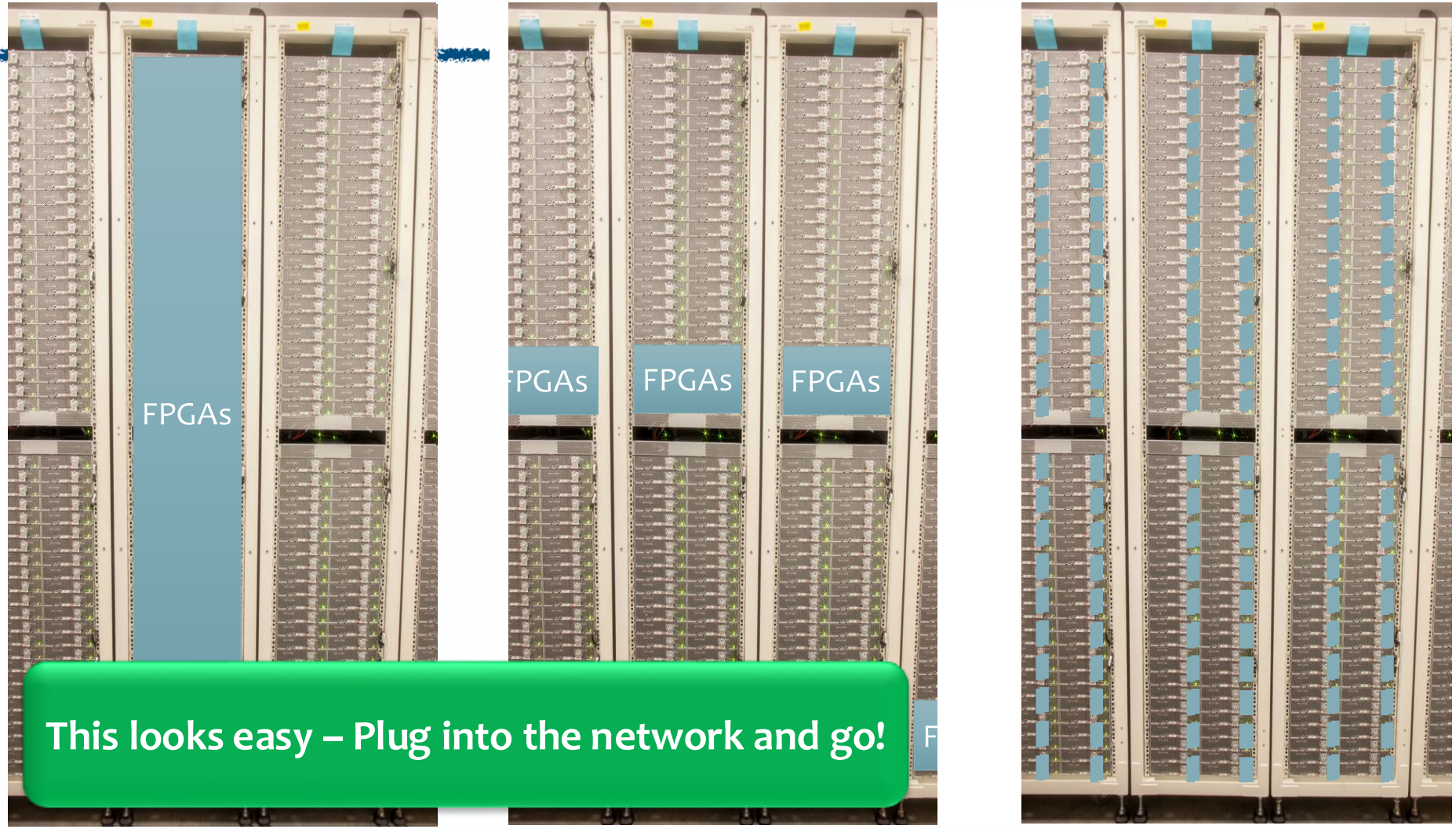
One Application's Accelerator



One Application's Accelerator



Integrating FPGAs into the Datacenter



Centralized **Distributed**

Microsoft Open Compute Server

Air flow

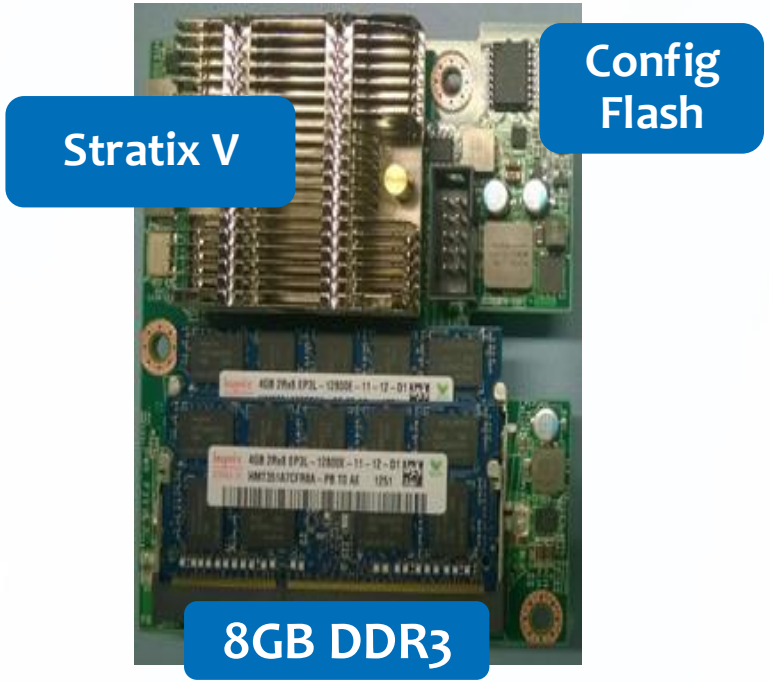
200 LFM

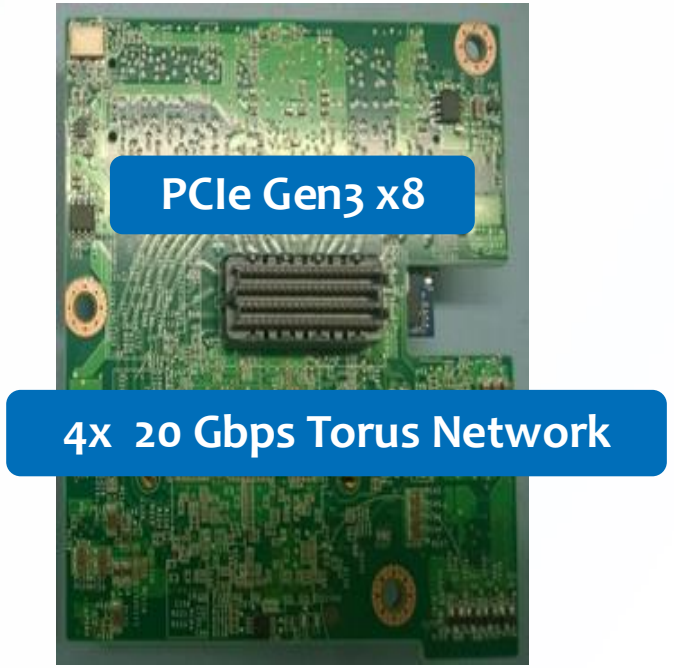
68 °C Inlet

Two 8-core Xeon 2.1 GHz CPUs
64 GB DRAM
4 HDDs @ 2 TB, 2 SSDs @ 512 GB
10 Gb Ethernet
No cable attachments to server

Catapult FPGA Accelerator Card

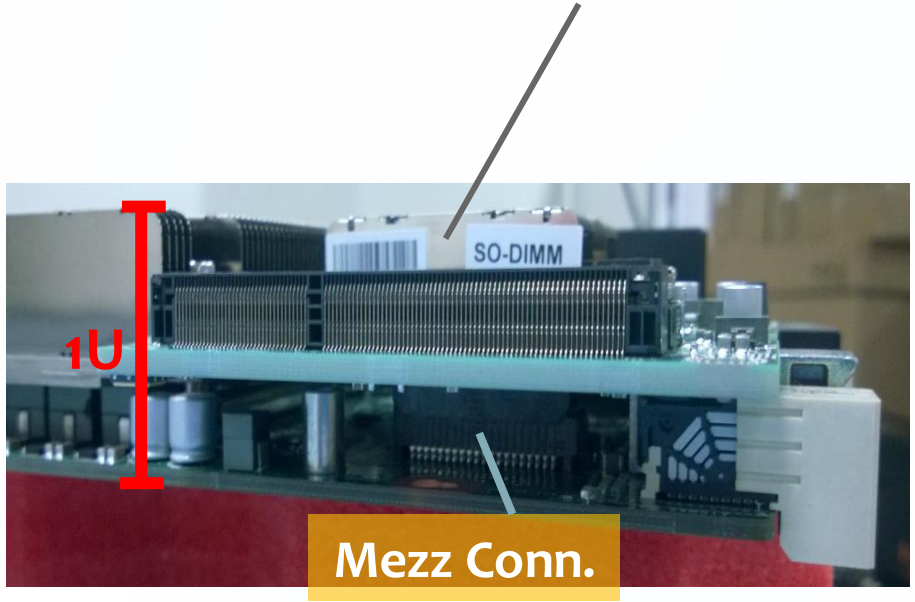
- Altera Stratix V GS D5
  - 172k ALMs, 2,014 M20Ks, 1,590 DSPs
- 8GB DDR3-1333
- 32 MB Configuration Flash
- PCIe Gen 3 x8
- 8 lanes to Mini-SAS SFF-8088 connectors
- Powered by PCIe slot

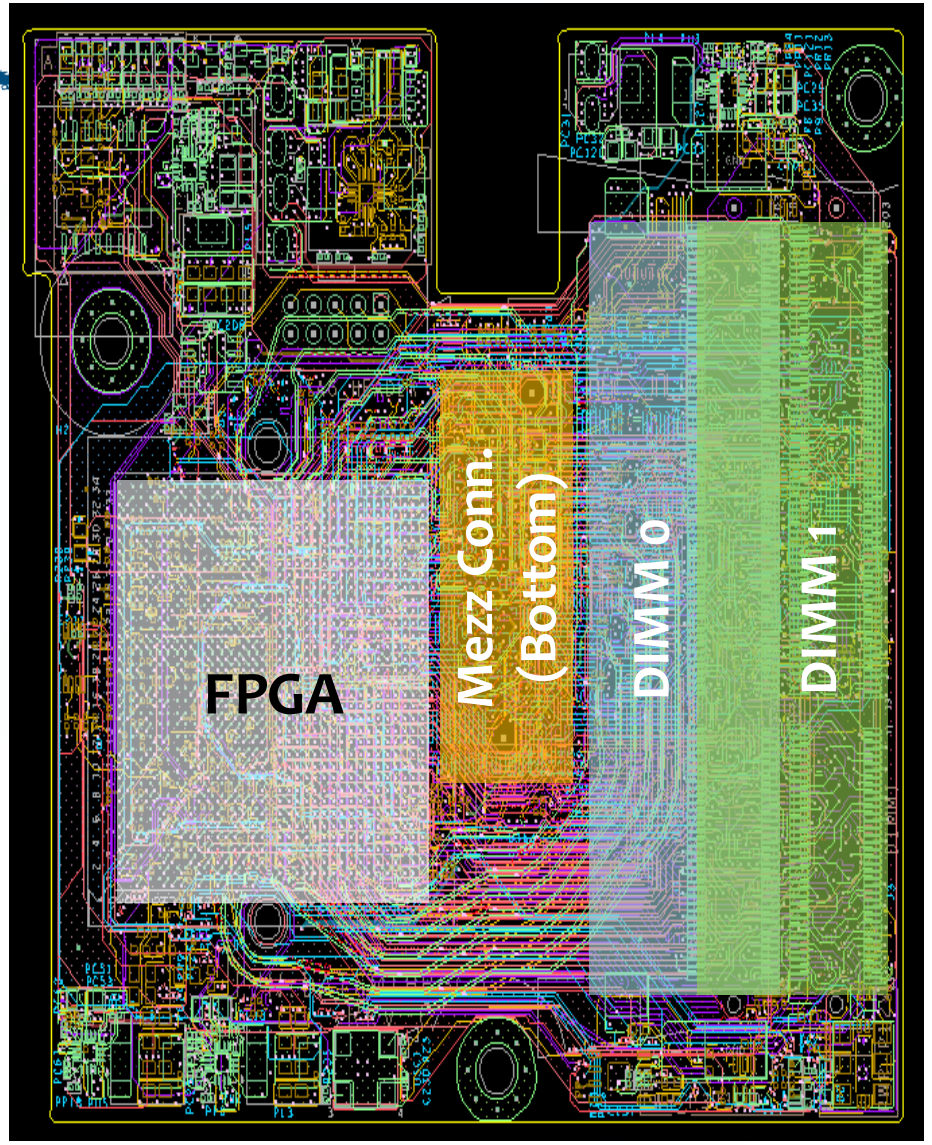




Board Details

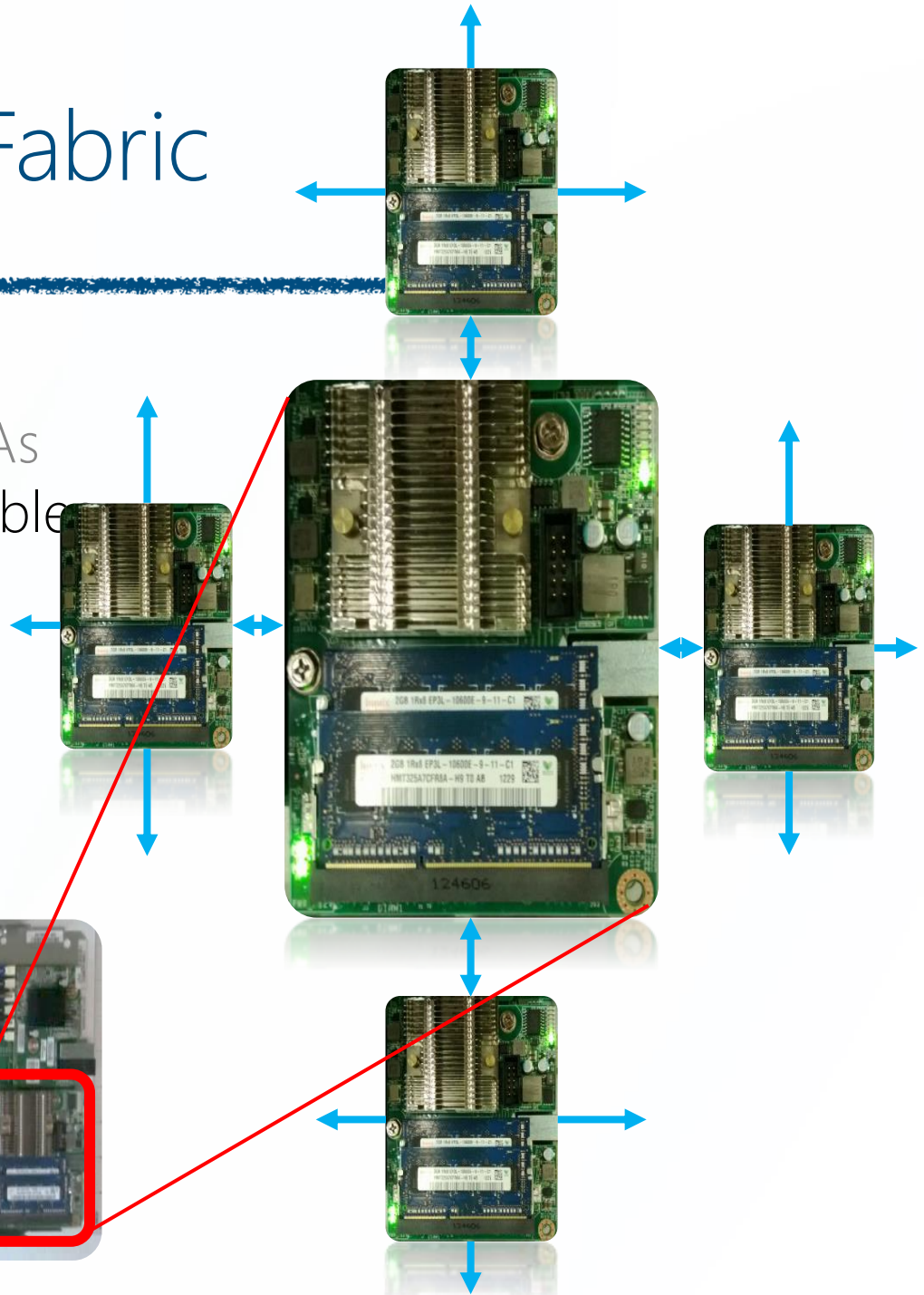
16 Layer, FR408
9.5cm x 8.8cm x 115.8 mil
35mm x 35mm FPGA
14.2mm high heatsink





Scalable Reconfigurable Fabric

1 FPGA board per Server
48 Servers per ½ Rack
6x8 Torus Network among FPGAs
20 Gb over SAS SFF-8088 cable



Data Center Server (1U, ½ width)

Microsoft Research

Microsoft Research
@MSFTResearch

Follow

Catapult propels datacenter services into the future @Bing @MSFTResearch @dcburger #FPGA bit.ly/1lzp10f

Reply Retweet Favorite More

RETWEETS 56 FAVORITES 30

3:00 PM - 16 Jun 2014

Flag media

Reply to @MSFTResearch @bing @dcburger

Intel's $16.7 Billion Altera Deal Is Fueled by Data Centers

by Ian King

June 1, 2015 – 8:34 AM EDT *Updated on* June 1, 2015 – 4:13 PM EDT

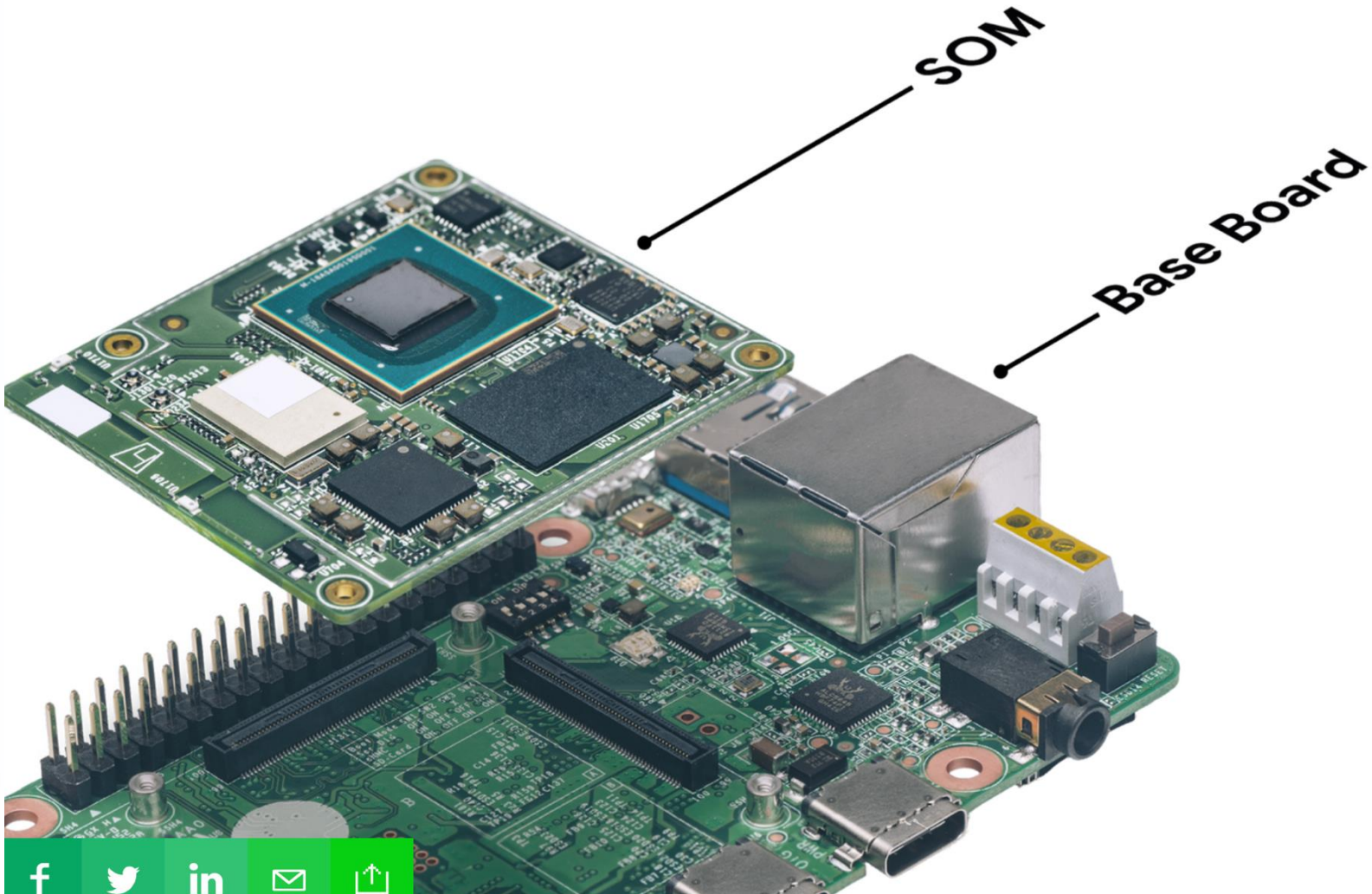
Intel Acquiring Altera in $16.7B Chipmaker Combination

Intel Corp. agreed to buy Altera Corp. for $16.7 billion to defend its presence in data centers, forging a deal that will add to a record year for industry consolidation.

Google is making a fast specialized TPU chip for edge devices and a suite of services to support it

Matthew Lynley @mattlynley / 2 months ago

Comment

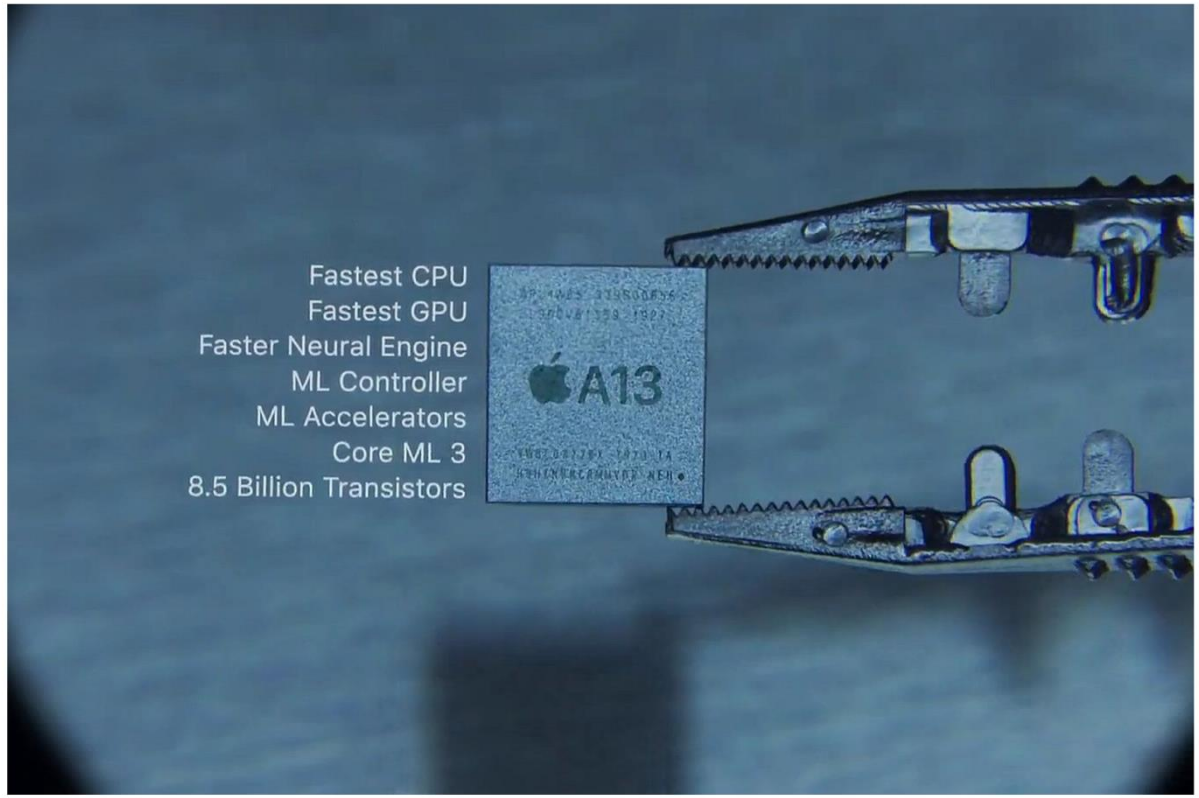


Apple says its new A13 Bionic chip brings hours of extra battery life to new iPhones

Plus a 20 percent performance boost across the board

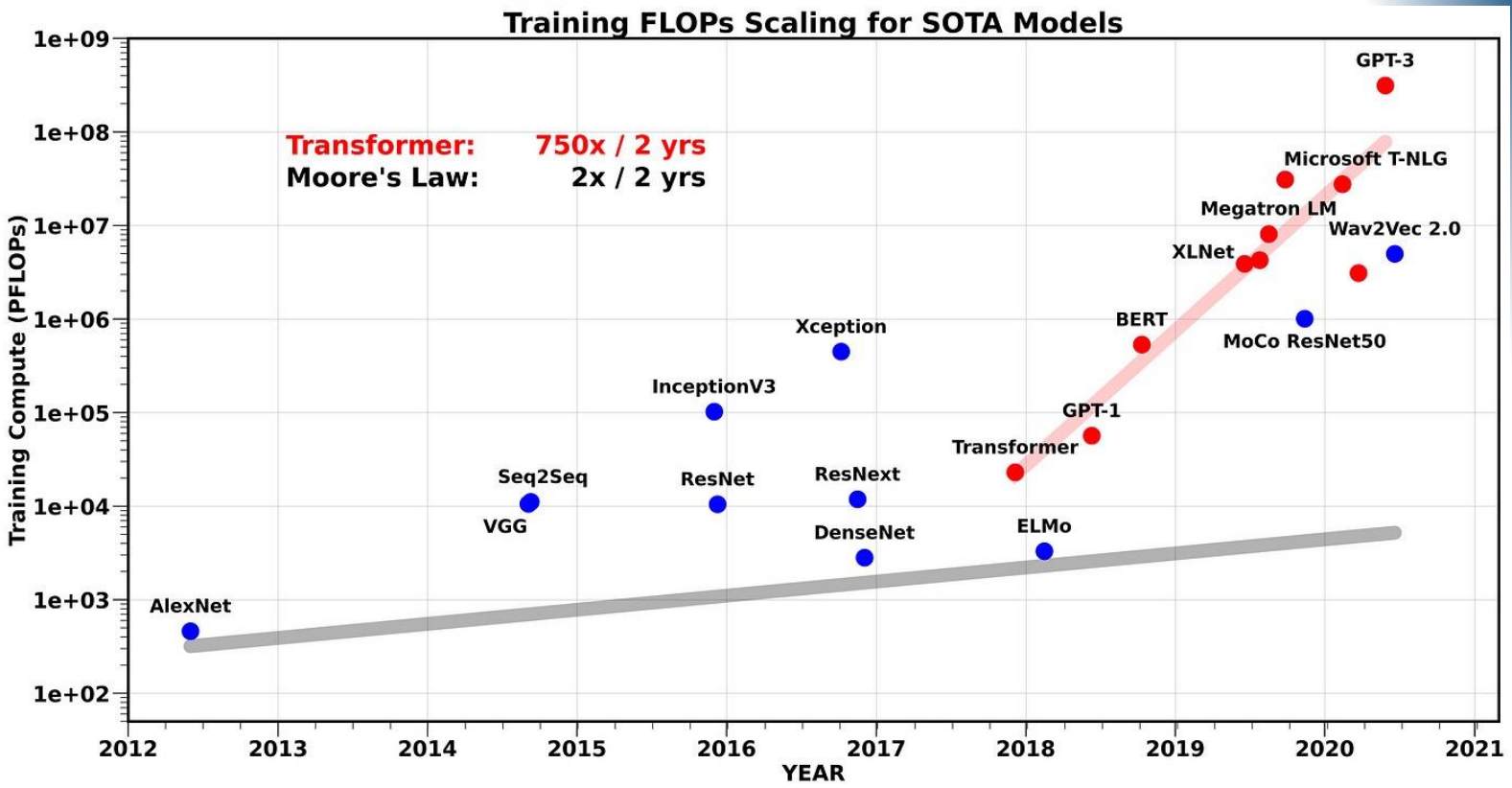
By Sean Hollister | @StarFire2258 | Sep 10, 2019, 2:02pm EDT

SHARE



Apple has revealed the chip that will power its new 2019 iPhones: the A13 Bionic. And as you'd expect, the company is wasting no time in explaining that it's the most powerful silicon ever to grace the inside of a smartphone — just as it has every year for the past three years. But if you care about battery life, you'll want to pay attention.

The Era of Neural Networks



Article

Conscious Empathic AI in Service

Journal of Service Research
2022, Vol. 25(4) 549–564

DOI: 10.1177/10946705221103531
journals.sagepub.com/home/jsr
SAGE

Hadi Esmaeilzadeh[1,*] and Reza Vaezi[2,*]

Abstract

Recent advances in artificial intelligence (AI) have achieved human-scale speed and accuracy for classification tasks. Current systems do not need to be conscious to recognize patterns and classify them. However, for AI to advance to the next level, it needs to develop capabilities such as metathinking, creativity, and empathy. We contend that such a paradigm shift is possible through a fundamental change in the state of artificial intelligence toward consciousness, similar to what took place for humans through the process of natural selection and evolution. To that end, we propose that consciousness in AI is an emergent phenomenon that primordially appears when two machines cocreate their own language through which they can recall and communicate their internal state of time-varying symbol manipulation. Because, in our view, consciousness arises from the communication of inner states, it leads to empathy. We then provide a link between the empathic quality of machines and better service outcomes associated with empathic human agents that can also lead to accountability in AI services.

The Fifth Day of Creation …

By Maestro Mahmoud Farshchian