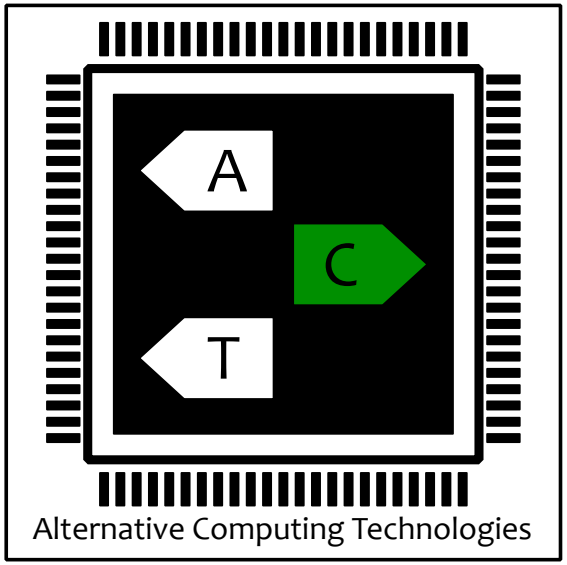
Introduction to Computer Architecture

CSE 141
Spring 2020

Hadi Esmaeilzadeh
Aka “Professor Hadi” or “Dr. Hadi”
hadi@eng.ucsd.edu
University of California, San Diego

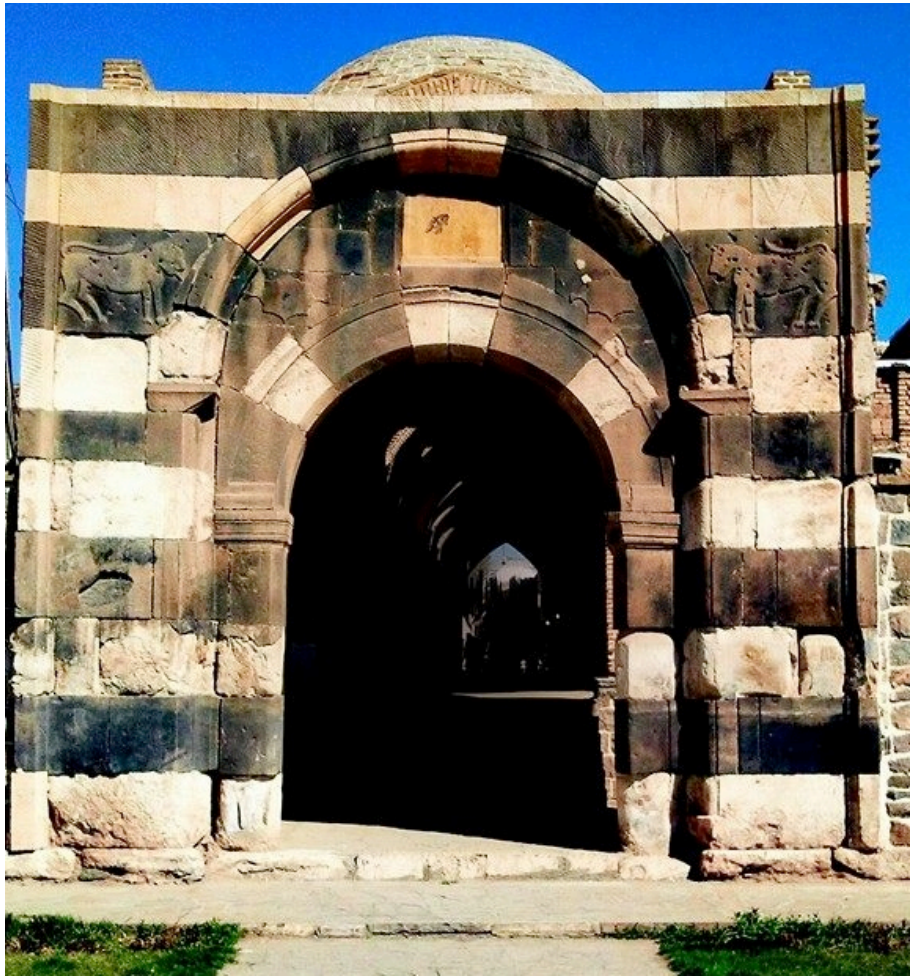


Reading Assignment for next class:
1.3, 2.1-2.3, 2.5

Optional Reading: https://cseweb.ucsd.edu/~hadi/doc/paper/2013-cacm-dark_silicon.pdf
https://www.nytimes.com/2011/08/01/science/01chips.html

Hadi Esmaeilzadeh

From Khoy, Iran



PhD in CSE, University of Washington

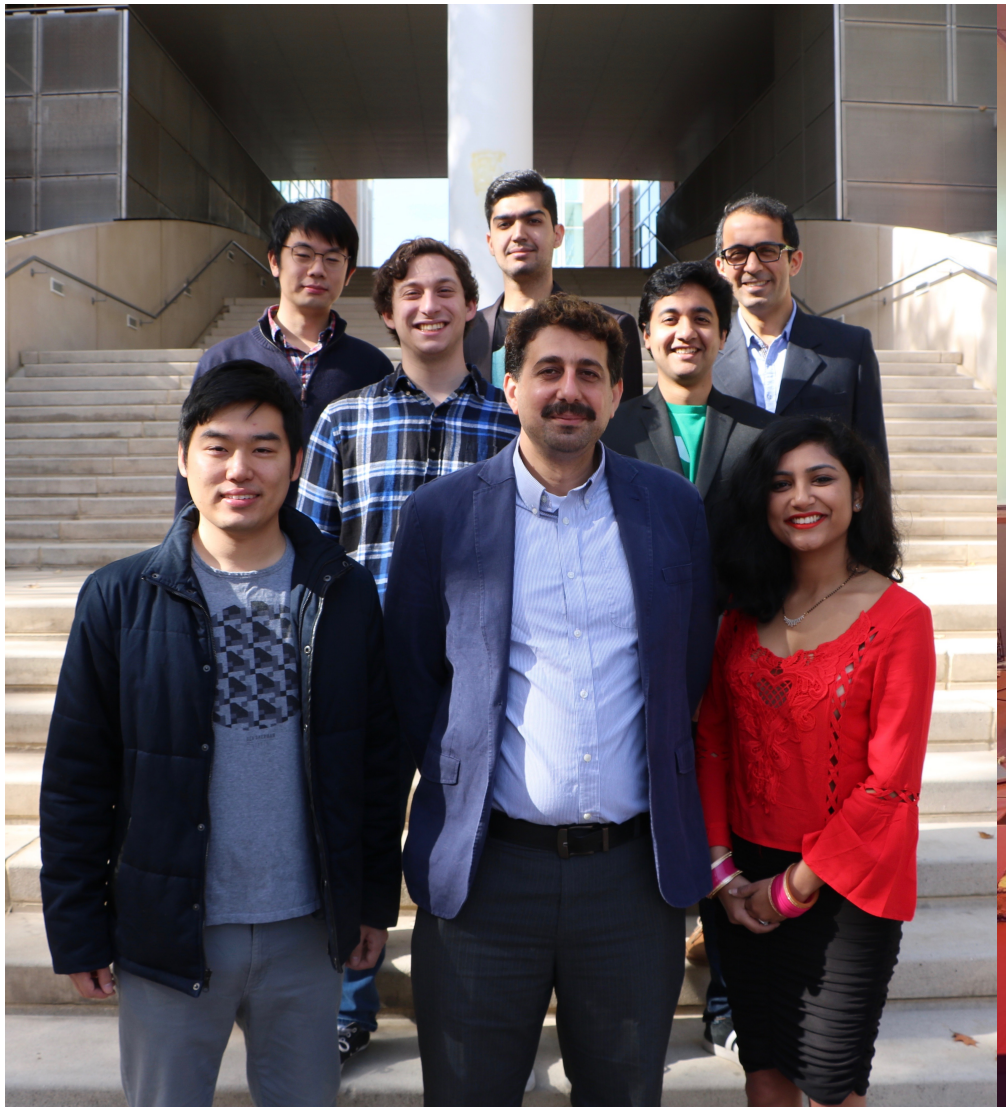
Doug Burger and Luis Ceze

2013 William Chan Memorial Best Dissertation Award

MSc in CS, The University of Texas at Austin

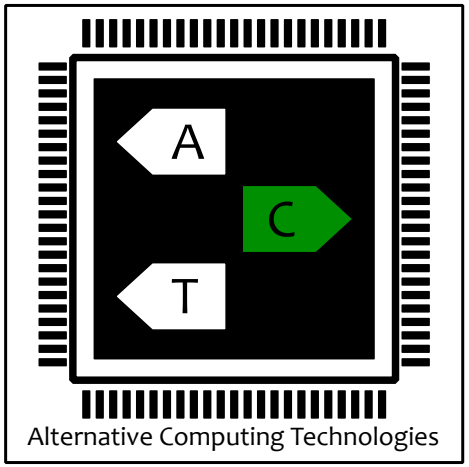
MSc and BSc in ECE, University of Tehran

Alternative Computing Technologies (ACT) Lab



Research: ACT Lab

Alternative Computing Technologies



- System design for machine learning
- Analog computing
- General-purpose approximate computing
- Privacy for artificial intelligence
- Compilation for machine learning
- Bridging neuromorphic and von Neumann computing

Agenda

1. Who is Hadi
2. **Course organization**
3. Why CS 141 - Introduction to Computer Architecture

Course Information and Communication

- Piazza
- Gradescope

- If you have any issues with these two websites please email Byung Hoon Ahn <bhahn221@eng.ucsd.edu>

Textbook

Patterson & Hennessy, 5th edition

"Computer Organization, the Hardware / Software Interface"

- Hennessy is former president of Stanford and chairman of Alphabet:
  - Recently stepped down as President of Stanford
  - Co-founded of MIPS Computer Systems
- Patterson is emeritus professor at Berkeley:
  - Lead RISC project (foundation of SPARC)
  - Lead RAID (redundant array of inexpensive disks)

Patterson and Hennessy just the 2017 Turing Award!

Basics – Your grade

- Bonus: 5%
  - In class questions and polls
- Reading and other quizzes: 3%
- Professionalism: 5%
- Homework: 21%
  - Must follow procedures for online submission
  - No late homework (but, drop your lowest score)
- Midterm: 31%
- Final: 40-45% Total: 105%-110%

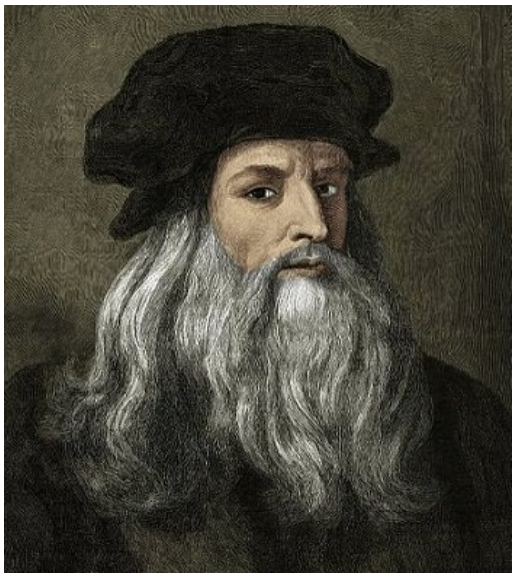
Objective:
To teach you the art of design and its beauty

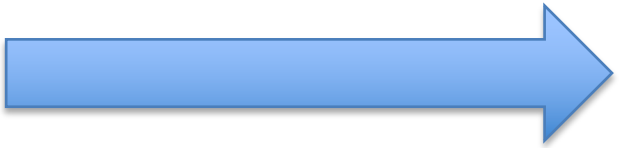
- Be the guide in **your** journey to understand how computers work and how to design them
- Create situations and post problems that set the scene for **your** exploration
- Answer **your** questions
- Not spend lecture reading the textbook to you with slightly different words

What do you do in class?

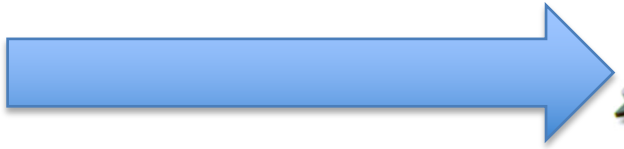
- Prepare to maximize in-class learning
  - Listen, read, and solve
- In class: engage with your colleagues and the class, engage with the ideas
  - Turn them upside down and sideways, think about what common errors or misconceptions might be
- Seek help and seek to help others
  - In class, piazza forums, office hours, discussion section
  - **I expect each class member to contribute to a positive environment of mutual aid and cooperation**

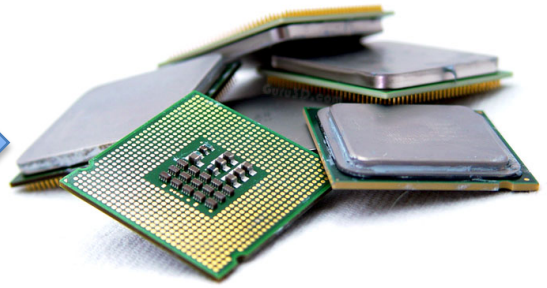
FAQ: "But professor, wouldn't it be more efficient if you just taught us with the right answer to begin with?"





You





Agenda

1. Who is Hadi
2. Course organization
3. **Why CSE 141**
    1. How we became and industry of new capabilities
    2. Why we might become an industry of replacement

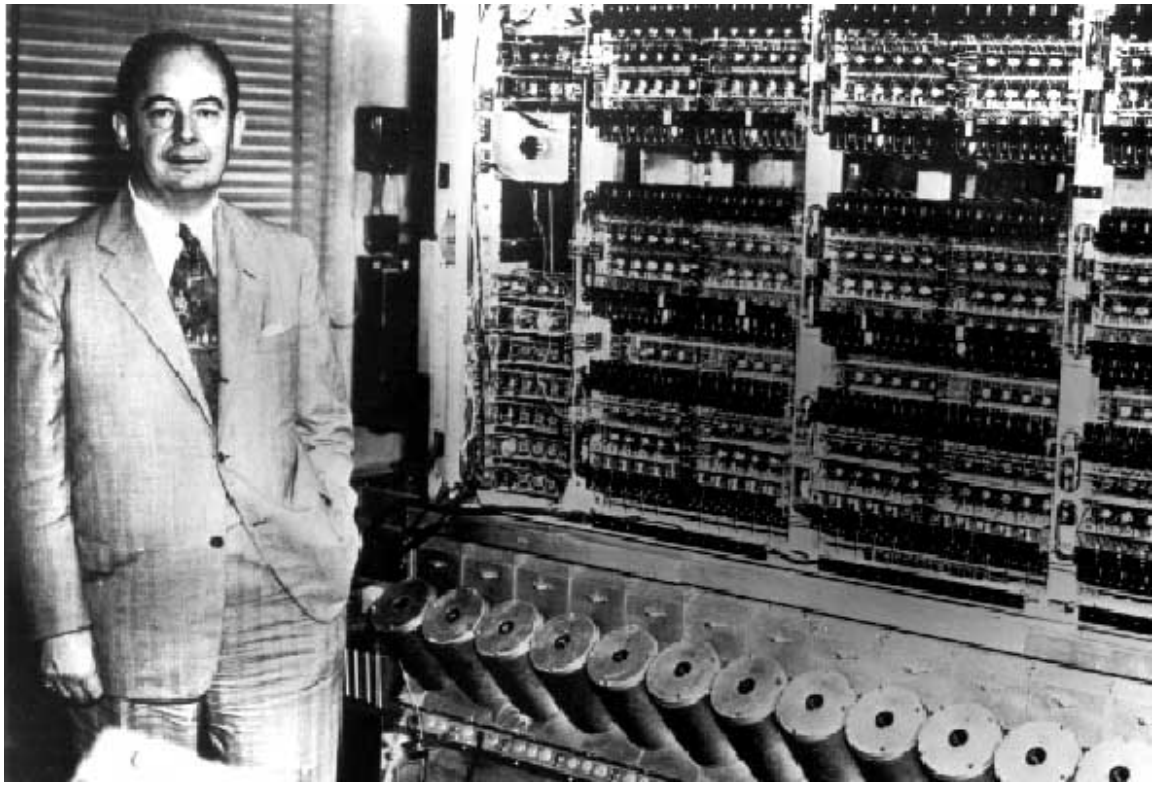
What has made computing pervasive? What is the backbone of the computing industry?

Programmability

Networking

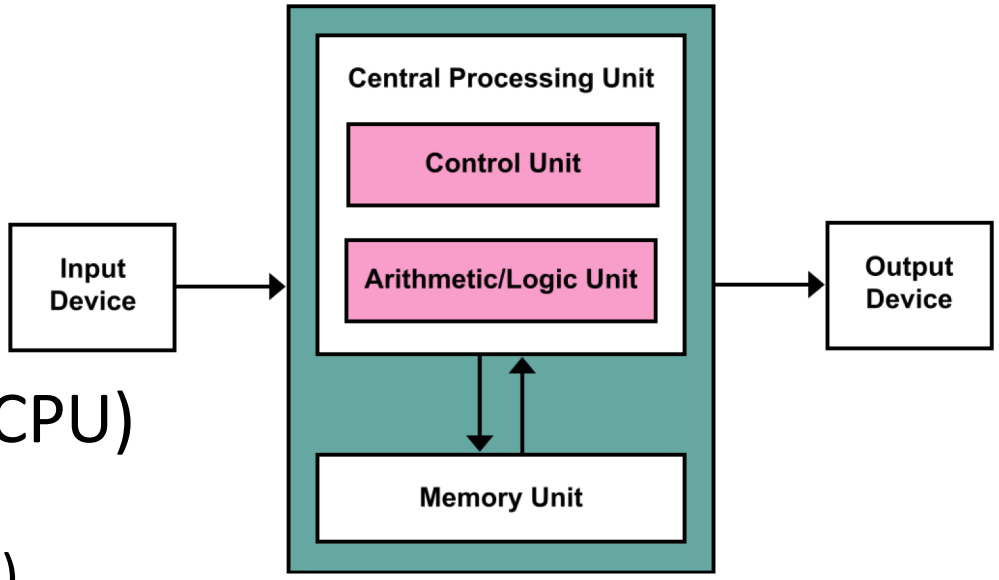
```
public class TcpClientSample
{
    public static void Main()
    {
        byte[] data = new byte[1024]; string input, stringData;
        TcpClient server;
        try{
            server = new TcpClient(" . . . . ", port);
        }catch (SocketException){
            Console.WriteLine("Unable 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();
```

What makes computers programmable?



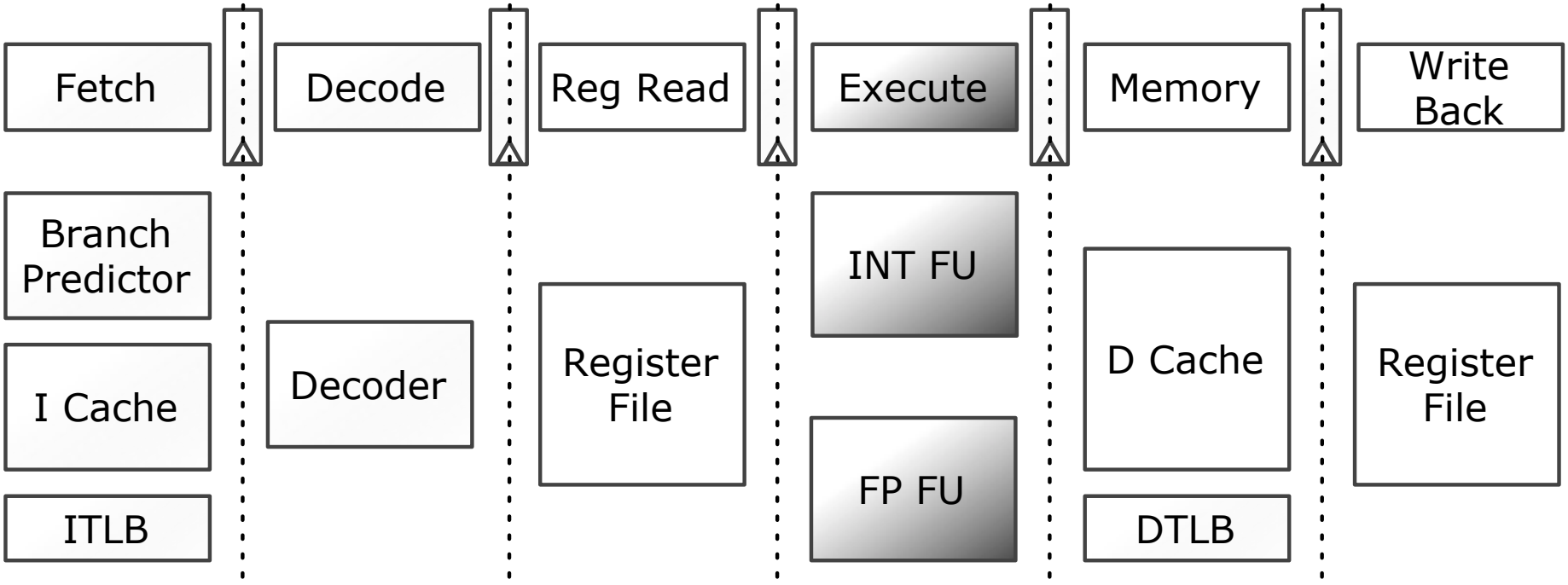
Von Neumann architecture

General-purpose processors

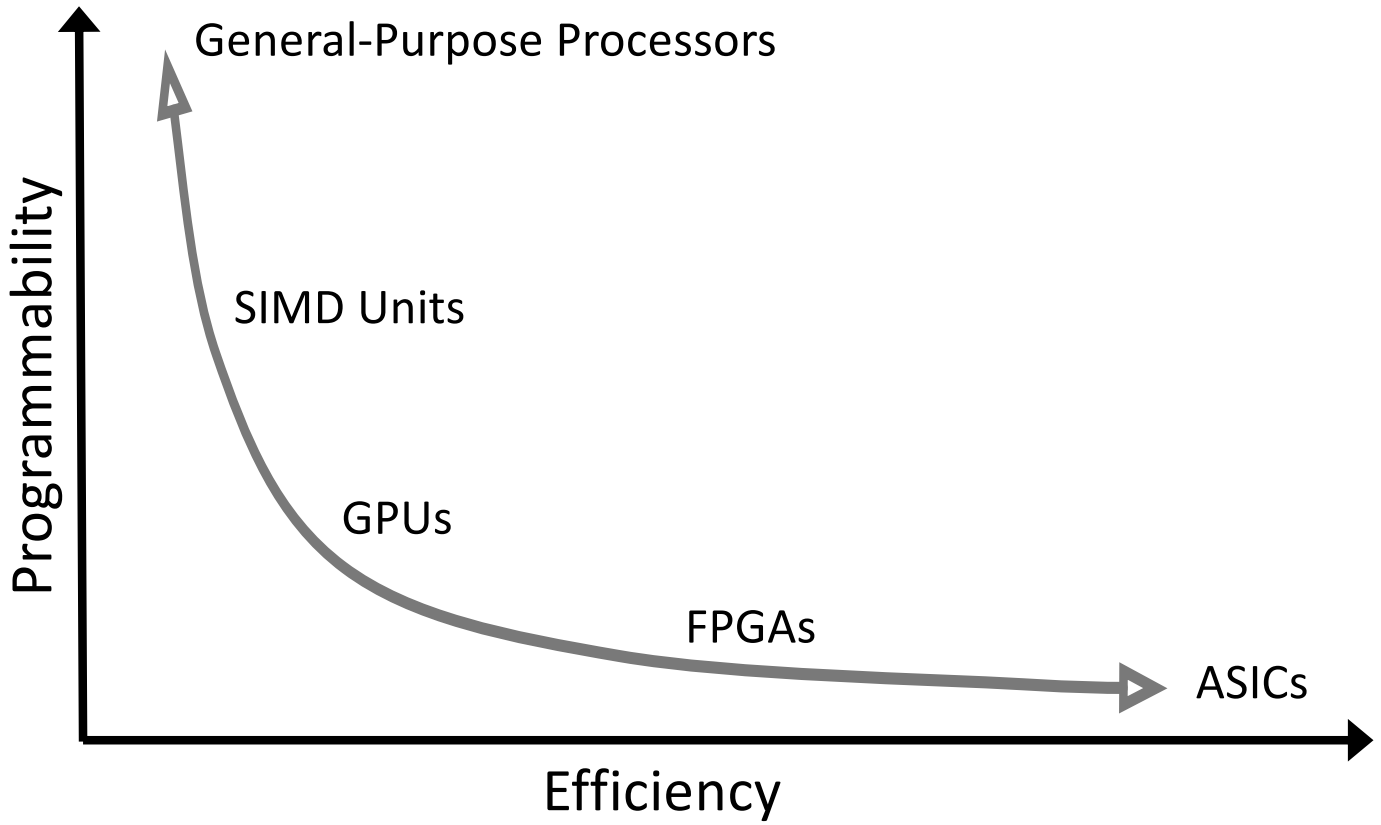


- Components
  - Memory (RAM)
  - Central processing unit (CPU)
    - Control unit
    - Arithmetic logic unit (ALU)
  - Input/output system
- Memory stores program and data
- Program instructions execute sequentially
  - Program Counter PC

Programmability versus Efficiency



Programmability versus Efficiency



What is the difference between the computing industry and the tissue paper industry?

Industry of replacement

1971 2014

?

Industry of new capabilities

Can we continue being an industry of new capabilities?

Personalized healthcare

Virtual reality

Real-time translators

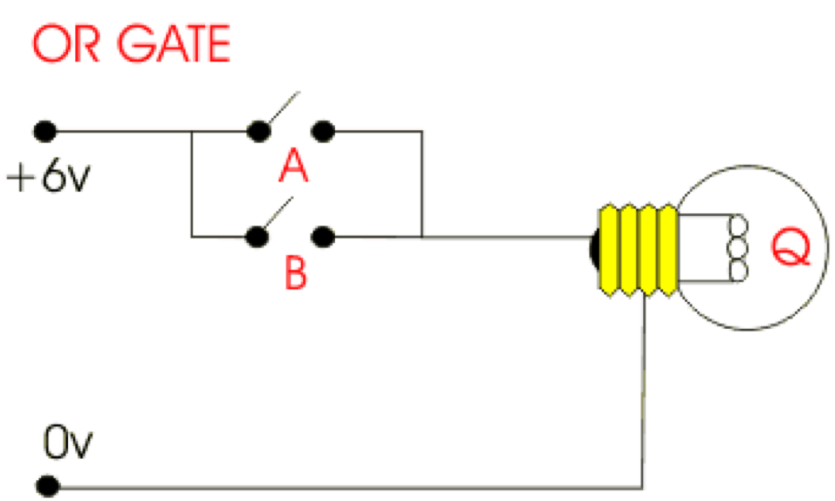
1) How we became an industry of new capabilities
   – Moore's Law and transistor scaling

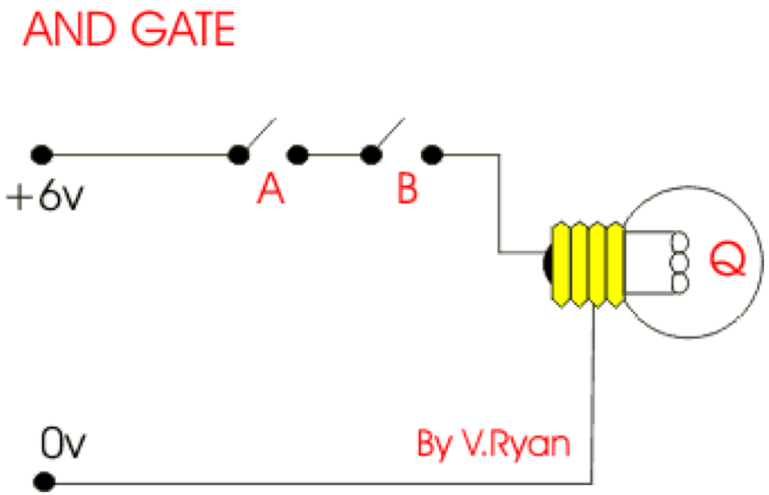
2) How we may become a replacement industry
   – Modeling future multicores

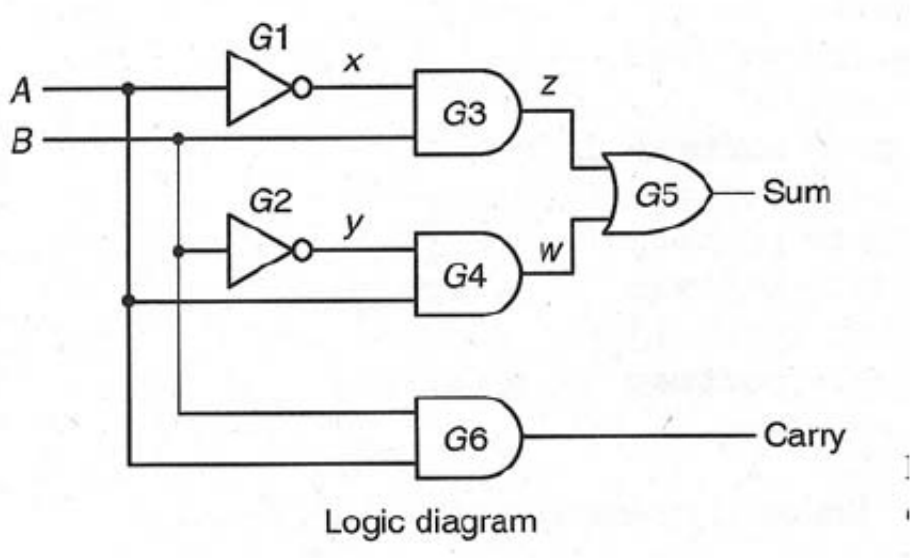
3) One new possible path forward
   – Specialization

Transistors/switches

Building blocks of computing







Logic diagram

Moore's Law

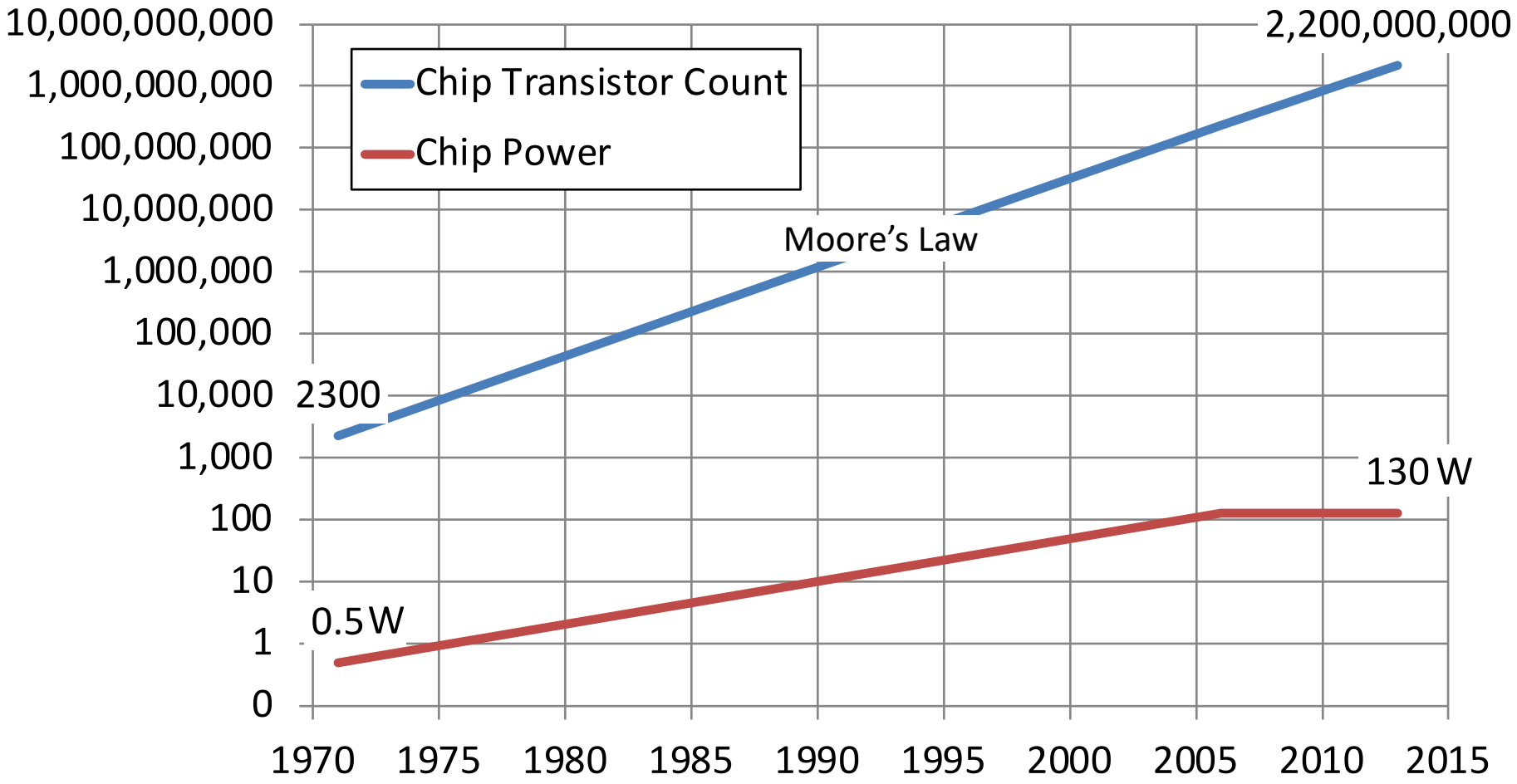
Or, how we became an industry of new possibilities

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



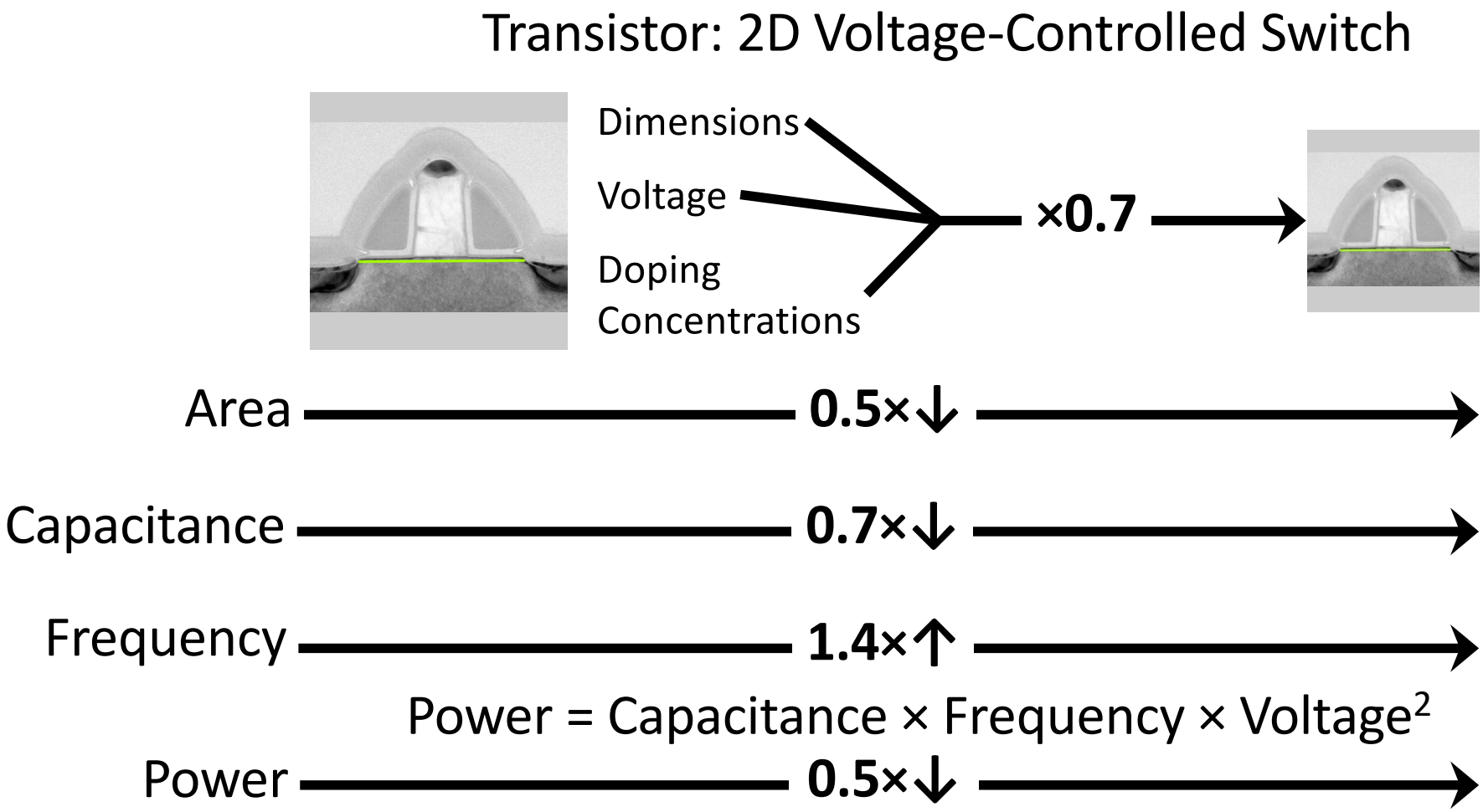
Double the transistors; scale their power down

Every 2 Years

- Historically
  - Reduce transistor power by half
  - While switching transistors 1.4× faster
- Currently
  - Slowed scaling in transistor power
  - Build multicores
  - Adding more cores is not scalable

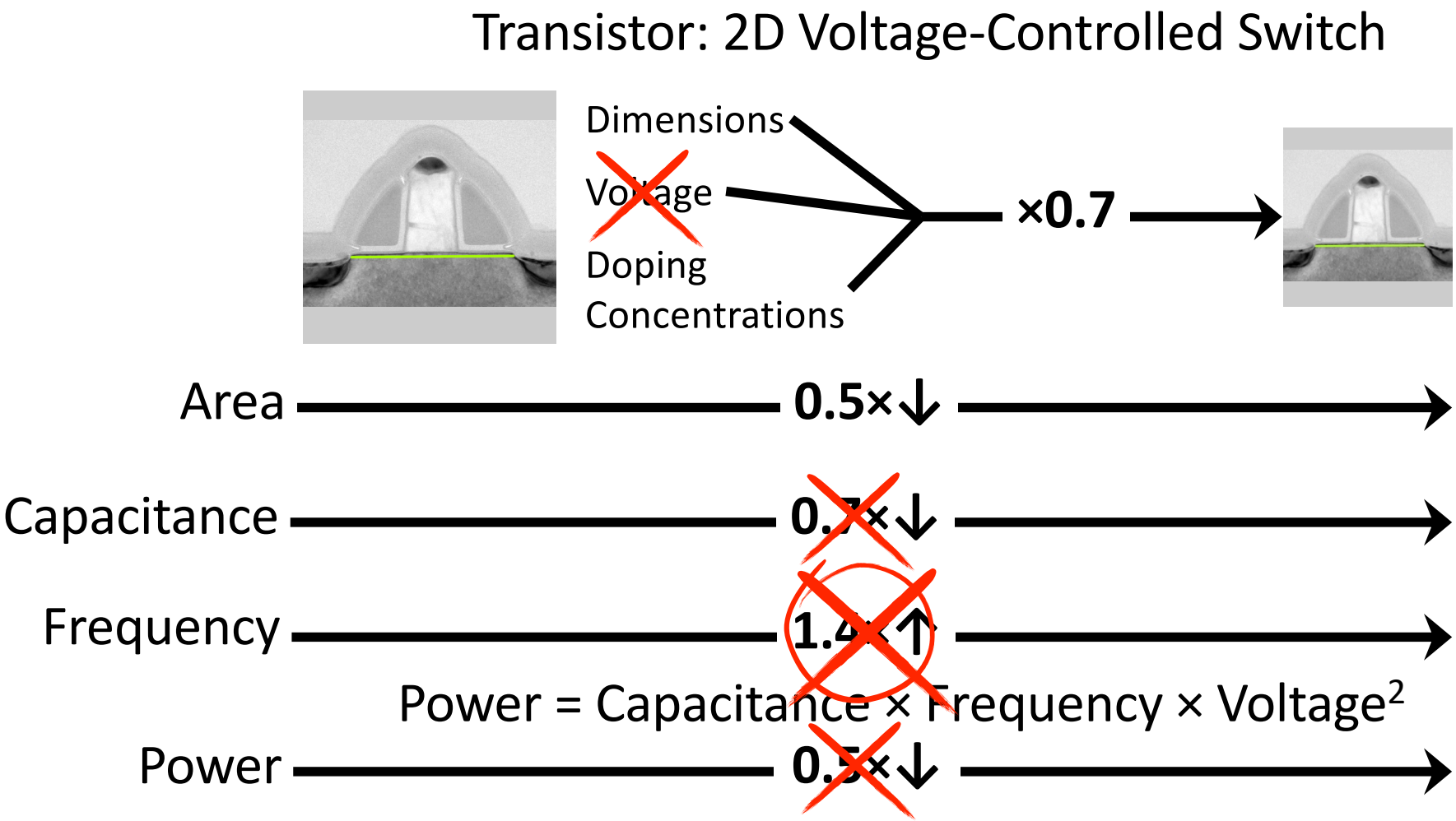
Dennard scaling:

Doubling the transistors; scale their power down

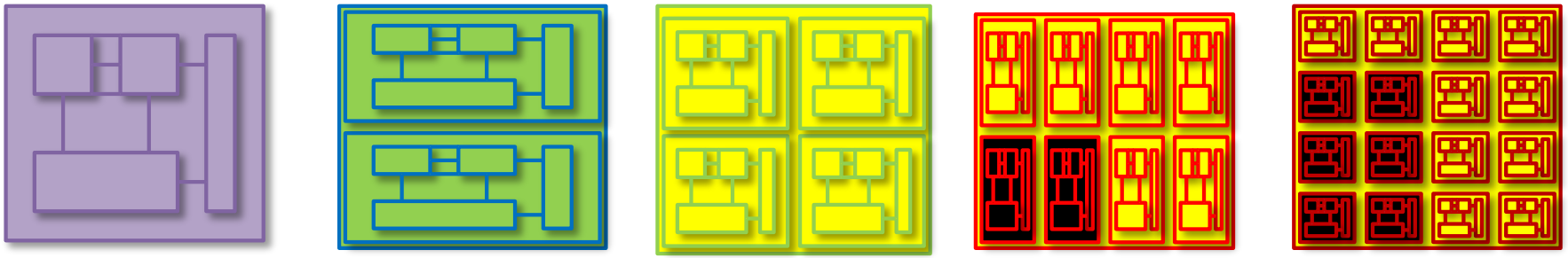


Dennard scaling broke:

Double the transistors; still scale their power down



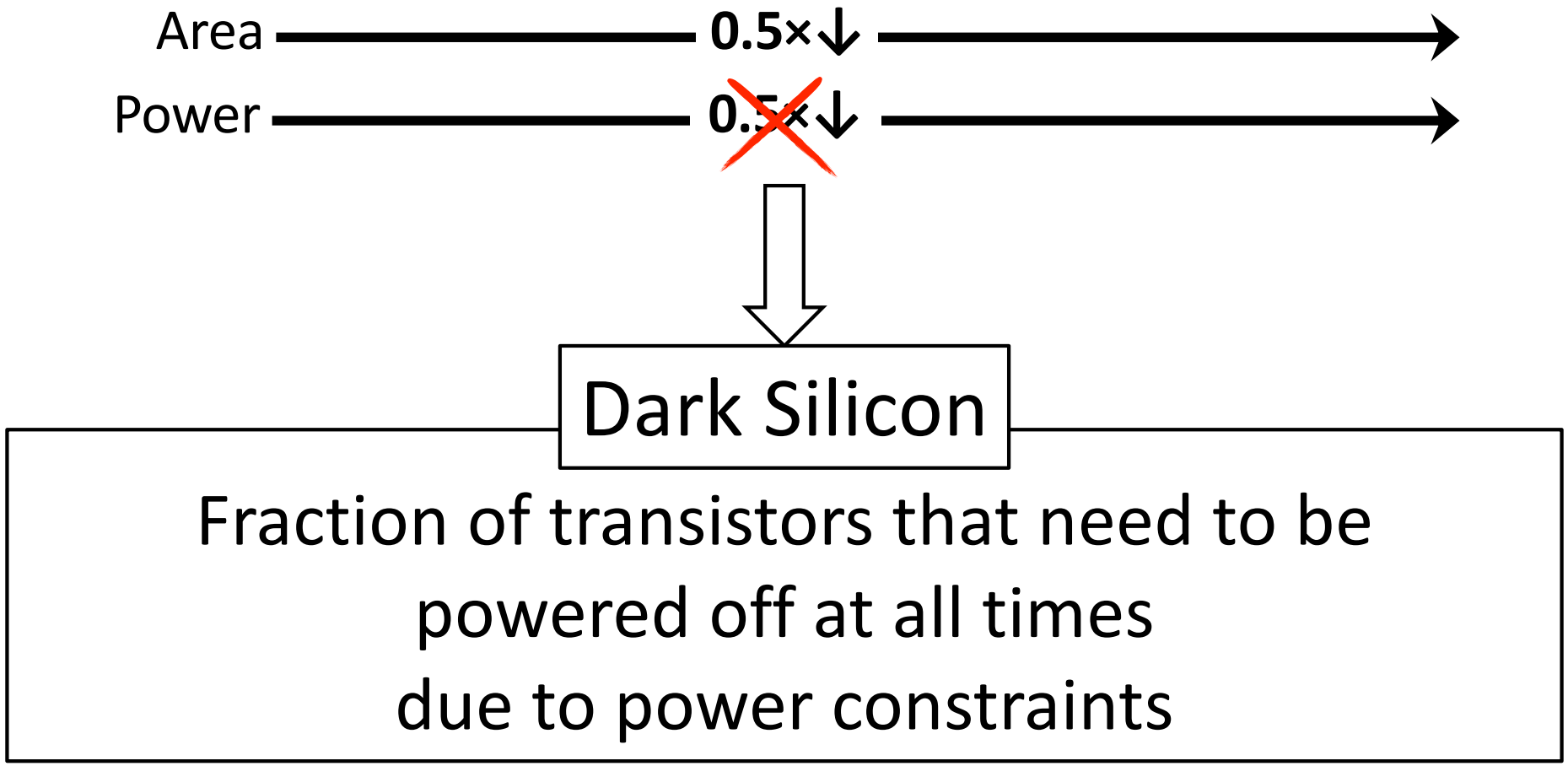
Why Diminishing Returns?



- Transistor area is still scaling
- Voltage and capacitance scaling have slowed
- Result: designs are power, not area, limited

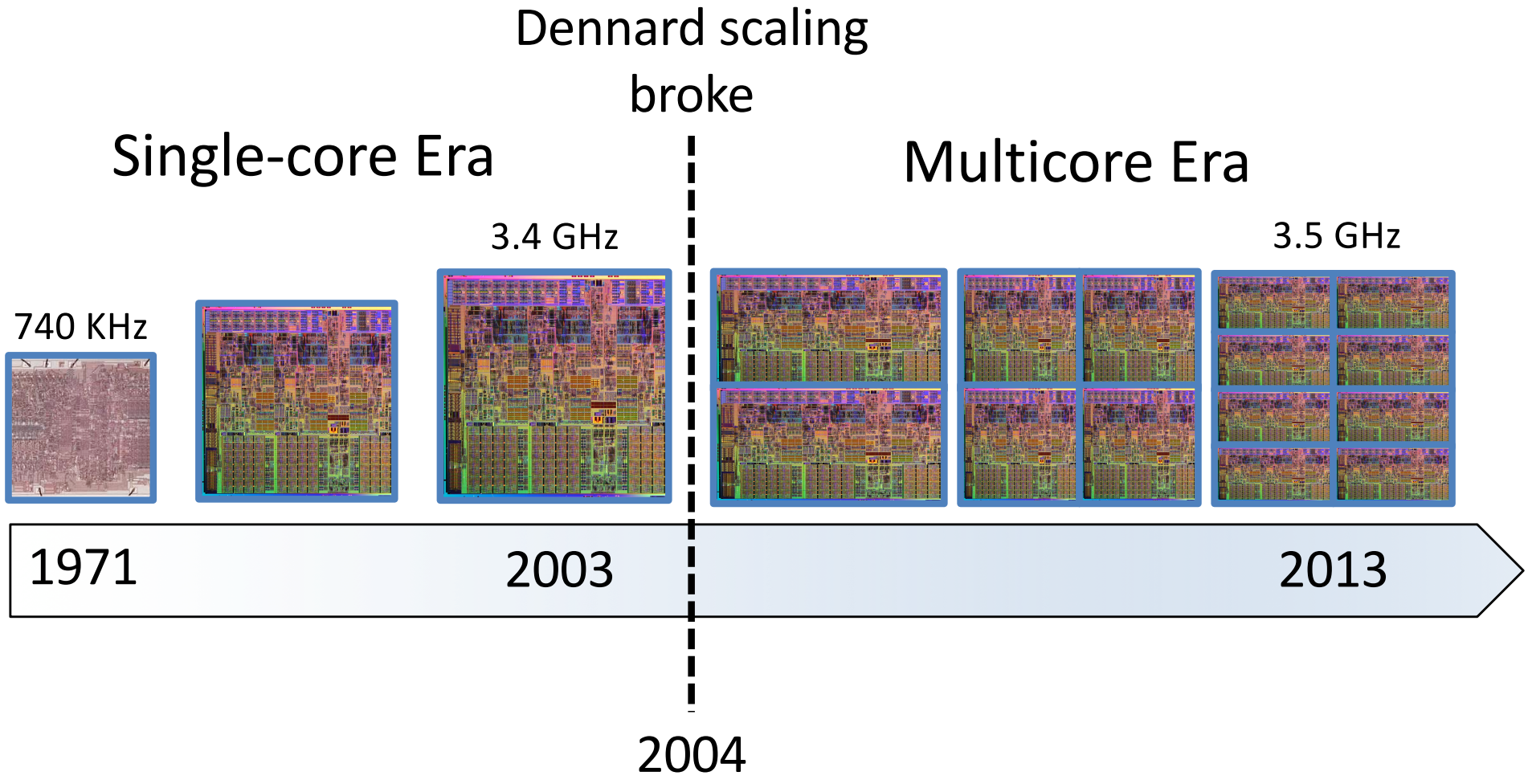
Dark silicon

If you cannot power them, why bother making them?



Looking back

Evolution of processors



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

1) How we became an industry of new possibilities
   – Moore's Law and transistor scaling

2) How we may become a replacement industry
   – Modeling future multicores

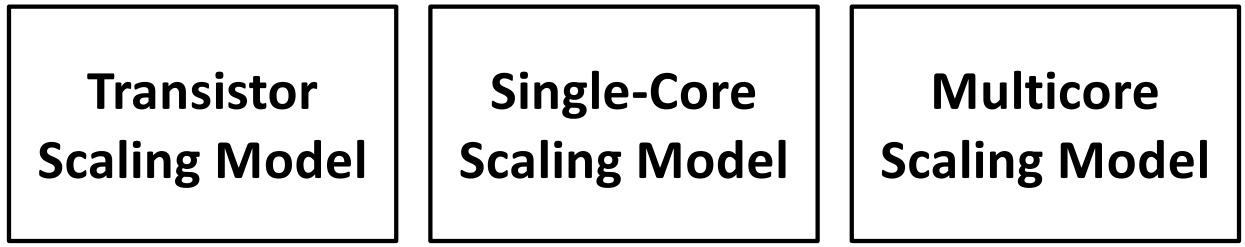
3) One new possible path forward
   – Specialization

Modeling future multicores

Quantify the severity of the problem

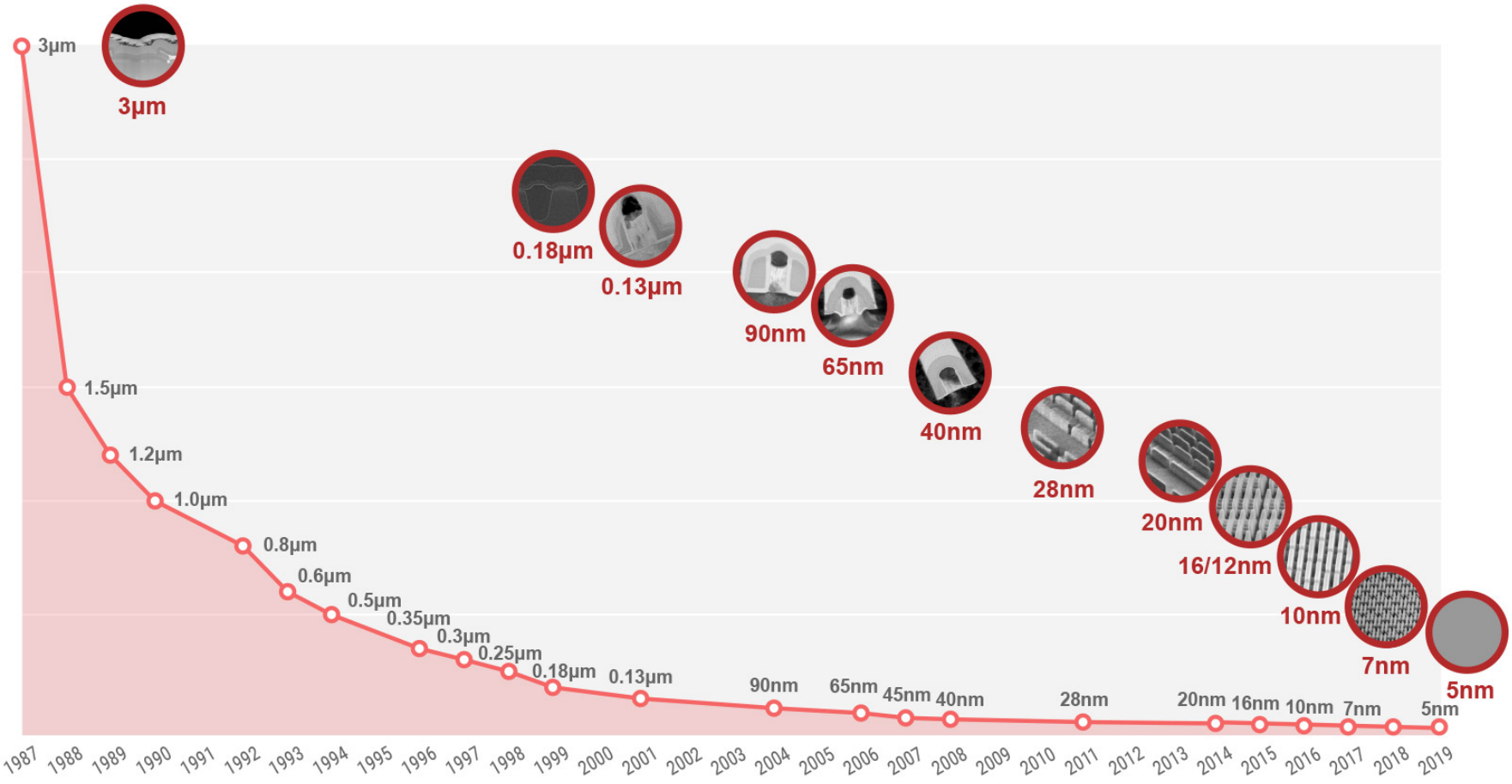
Predict the performance of best-case multicores

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

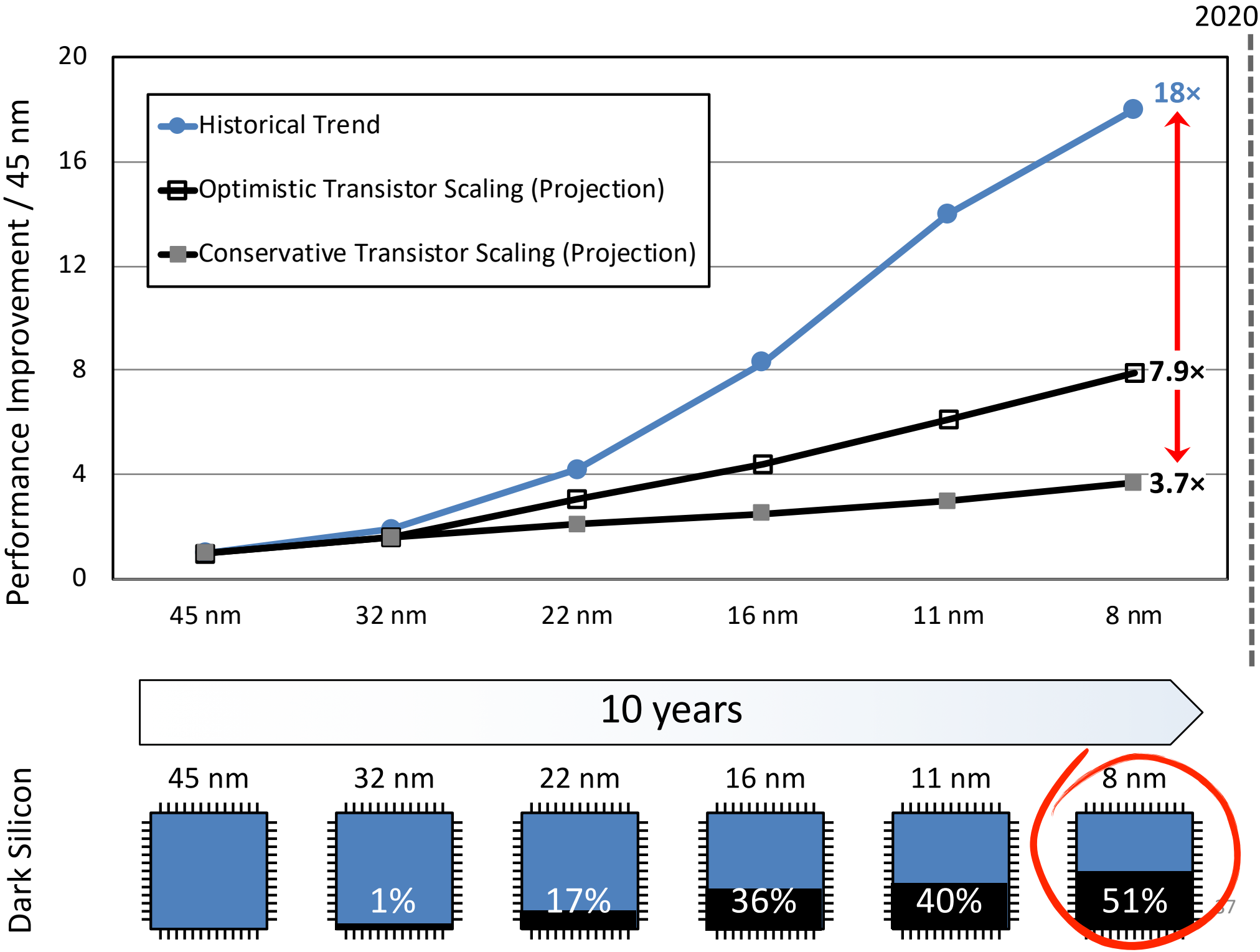


Esmaeilzadeh, Belem, St. Amant, Sankaralingam, Burger, "Dark Silicon and the End of Multicore Scaling," ISCA 2011

Taiwan Semiconductor Manufacturing Company (TSMC):
World's largest dedicated independent semiconductor foundry



Source: https://www.tsmc.com/english/dedicatedFoundry/technology/logic.htm


2020
20
16
12
8
4
0
Performance Improvement / 45 nm
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
10 years
Dark Silicon
45 nm
32 nm
22 nm
16 nm
11 nm
8 nm
1%
17%
36%
40%
51%

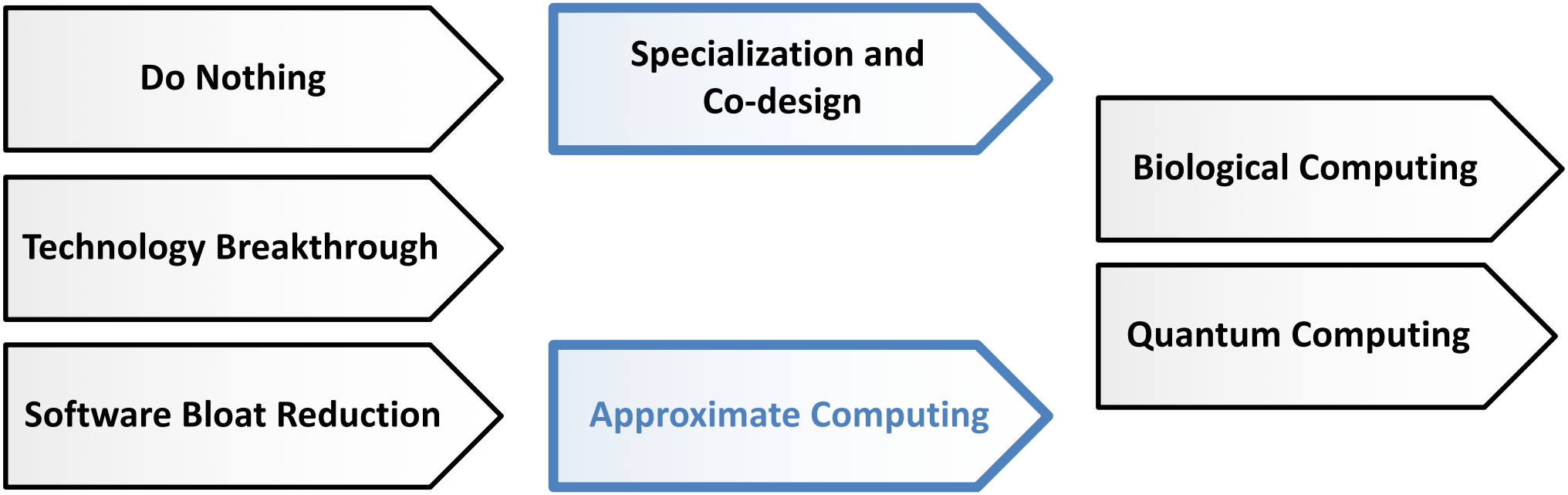
Industry of replacement?

- 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 models of computing

Agenda

1. Who is Hadi
2. Course organization
3. Why CSE 141
    1. How we became and industry of new capabilities
    2. Why we might become an industry of replacement
    3. **Specialization**

Possible paths forward



Easy for me!

My research!

Way long term!

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

WEB IMAGES VIDEOS MAPS NEWS MORE

bing Hadi

Size Color Type Layout People

New landscape of computing

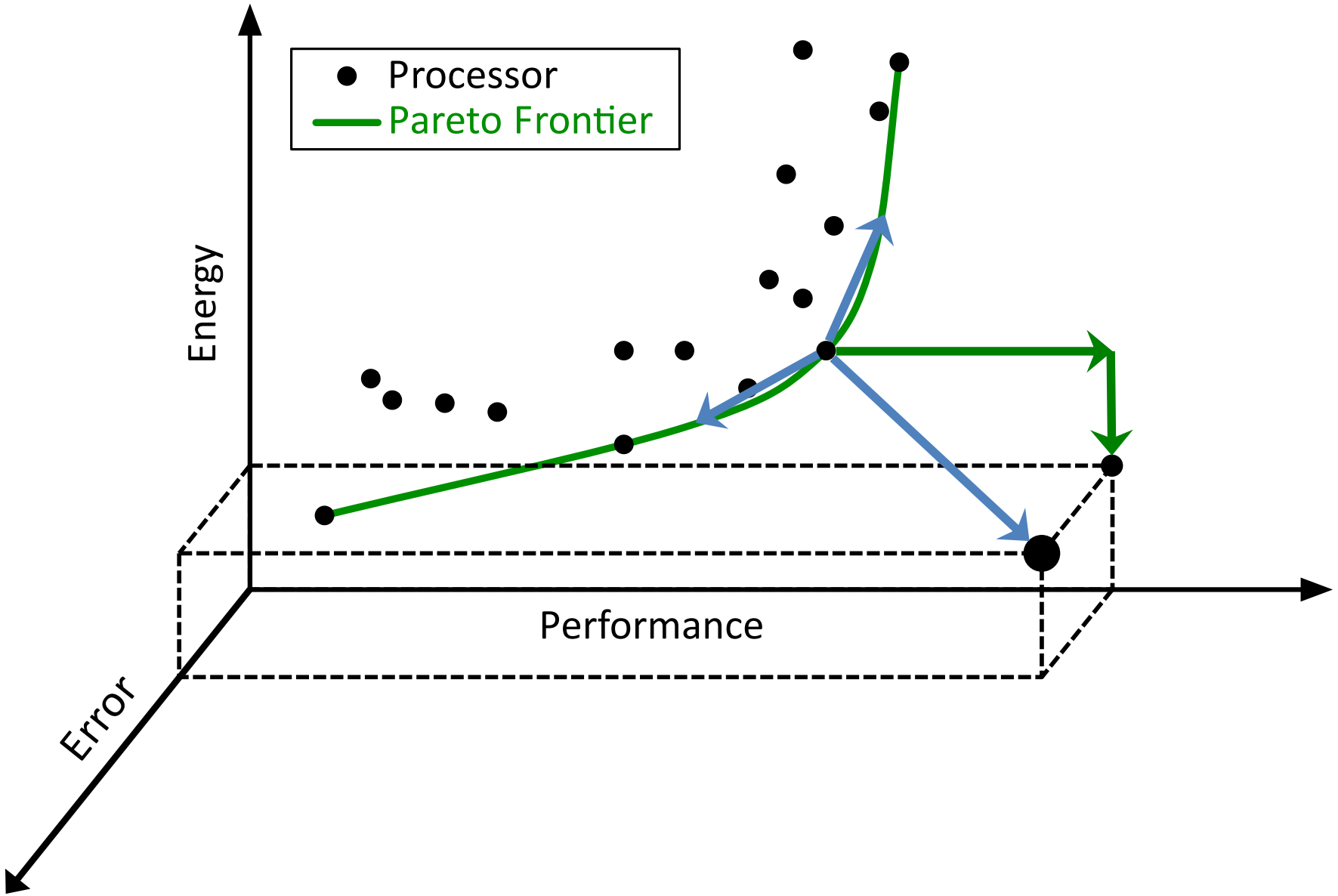
Personalized and targeted computing

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

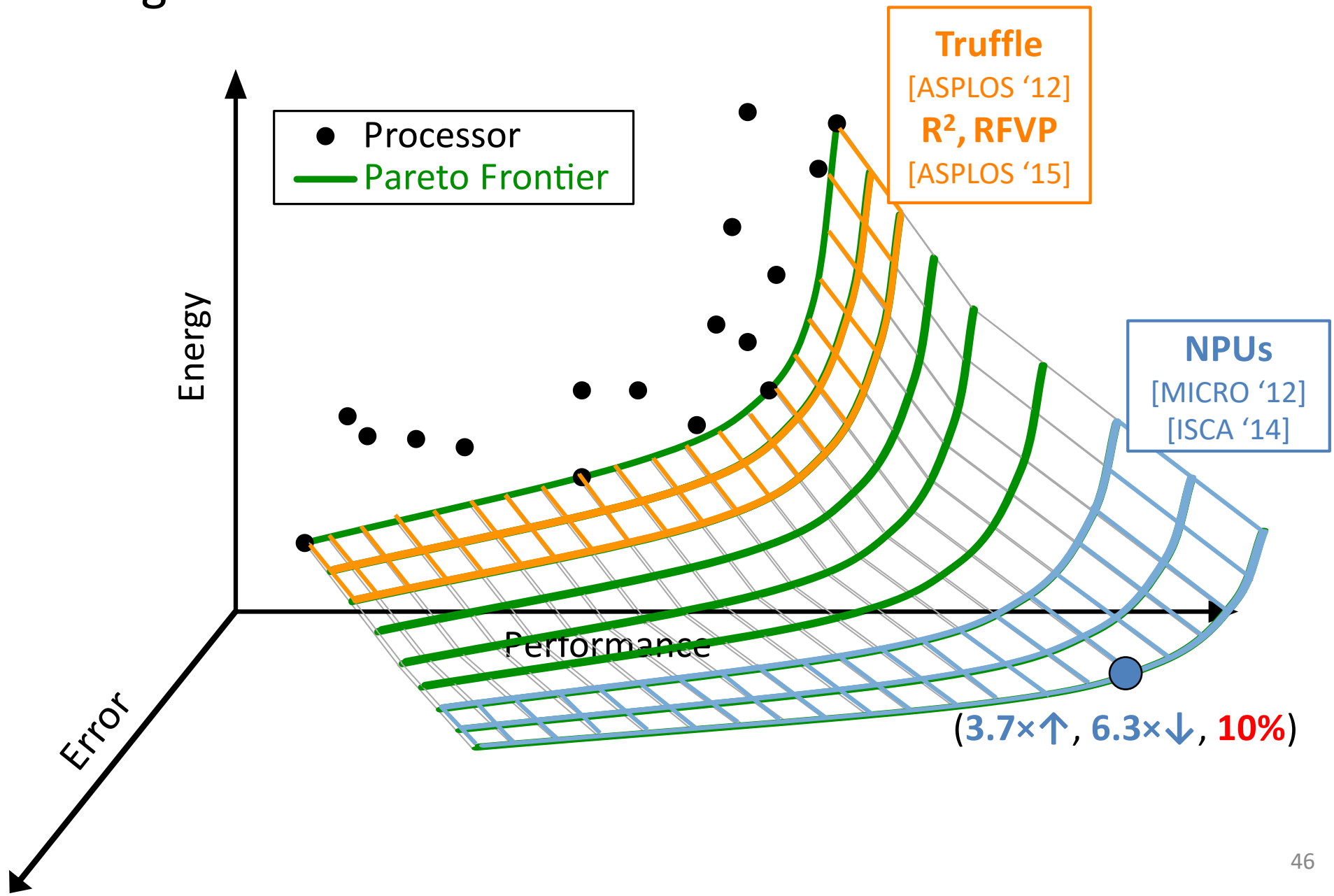
Adding a third dimension

Embracing Error

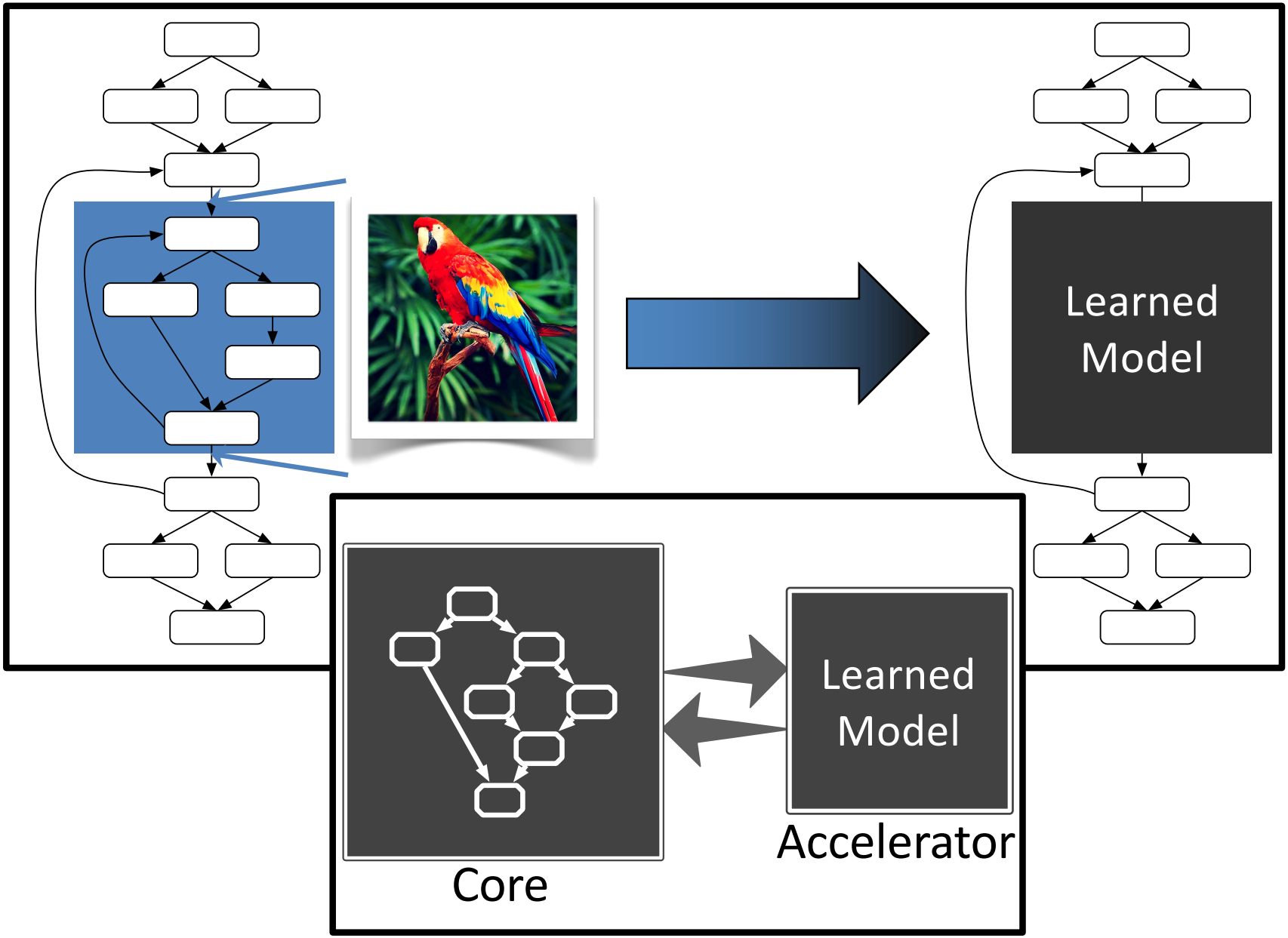


Adding the Dimension of Error

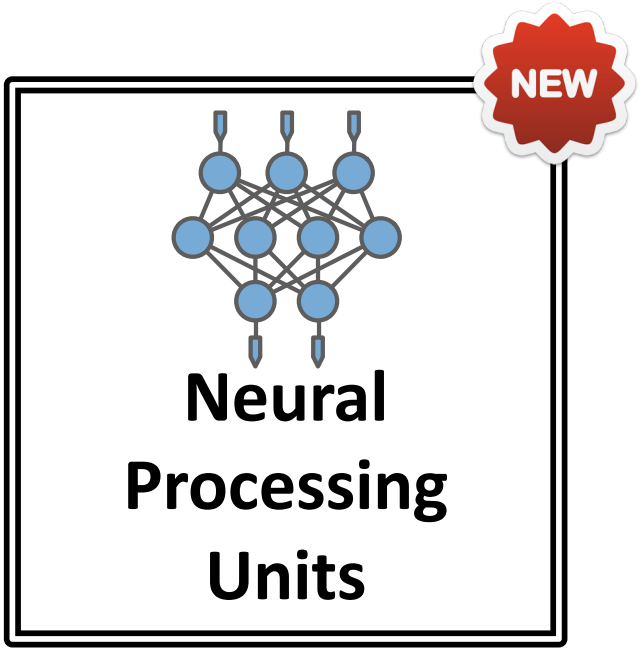
Finding the Pareto surface



Parrot algorithmic transformation

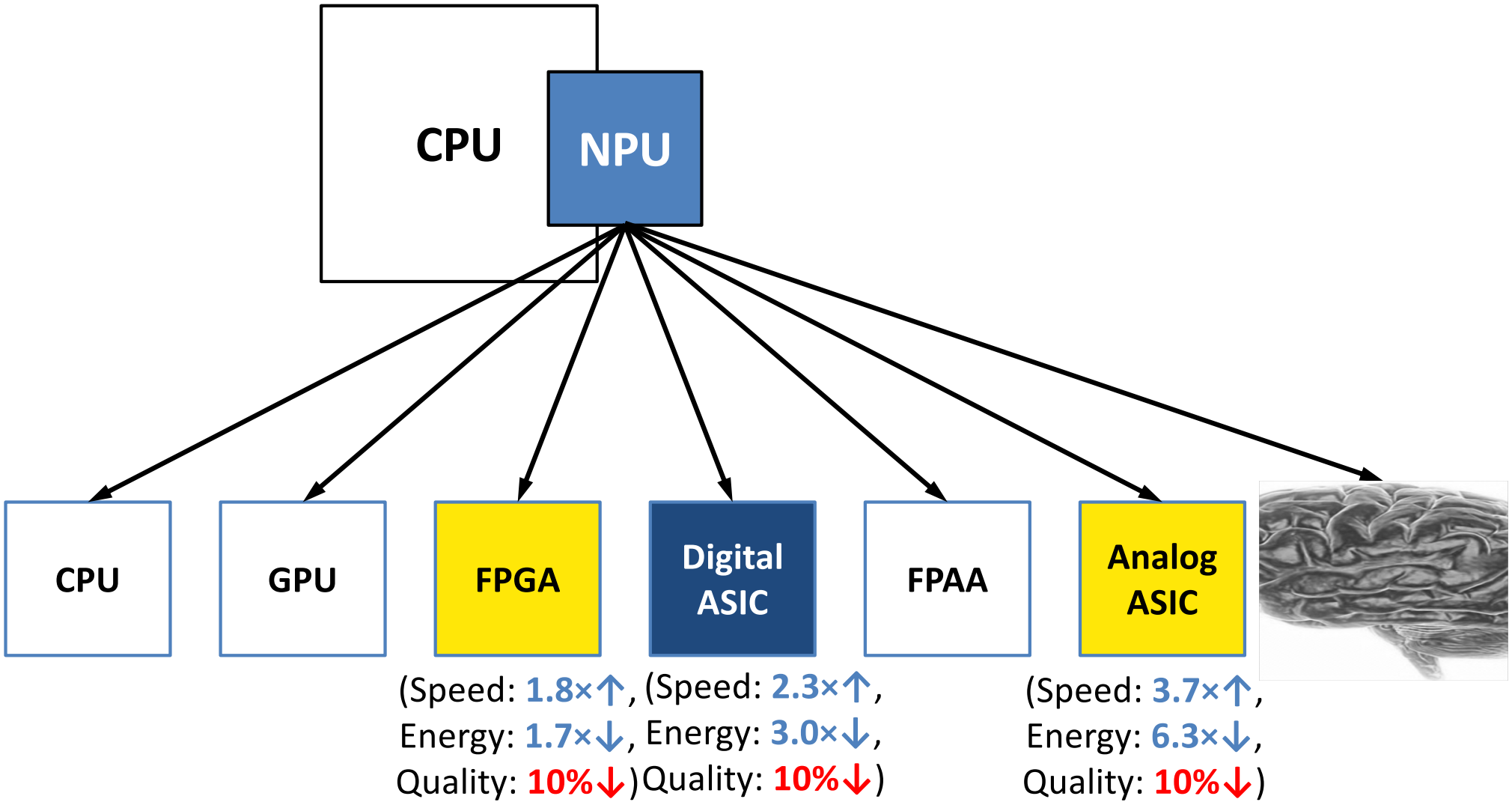


Neural networks for code approximation

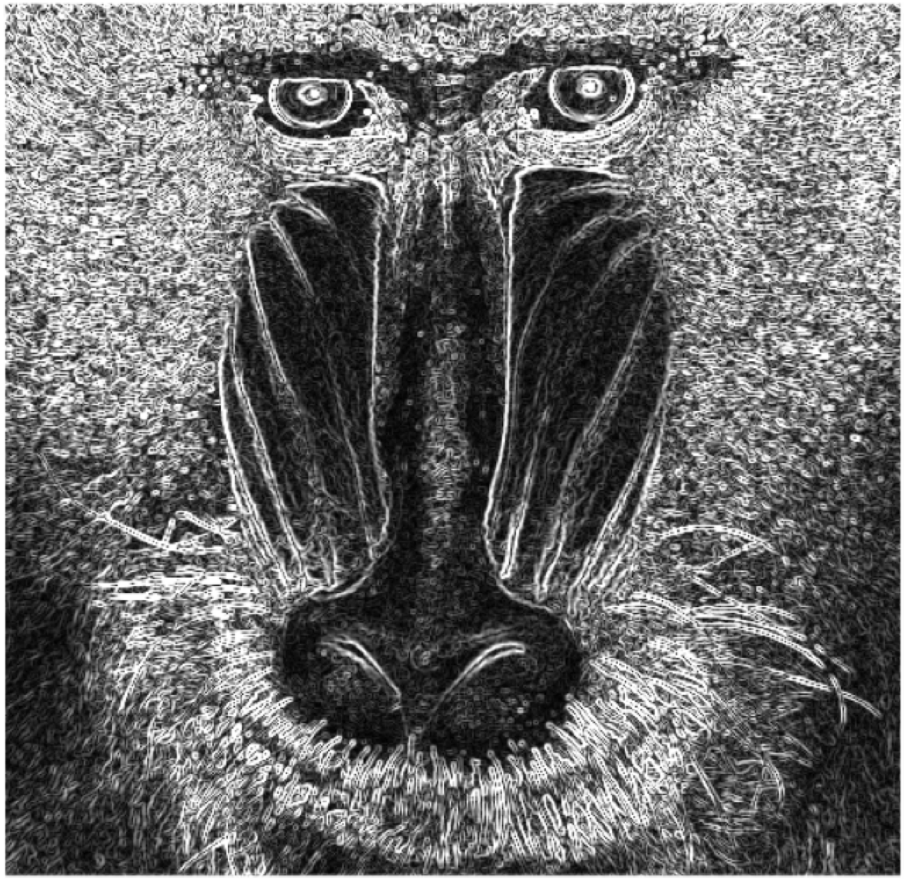


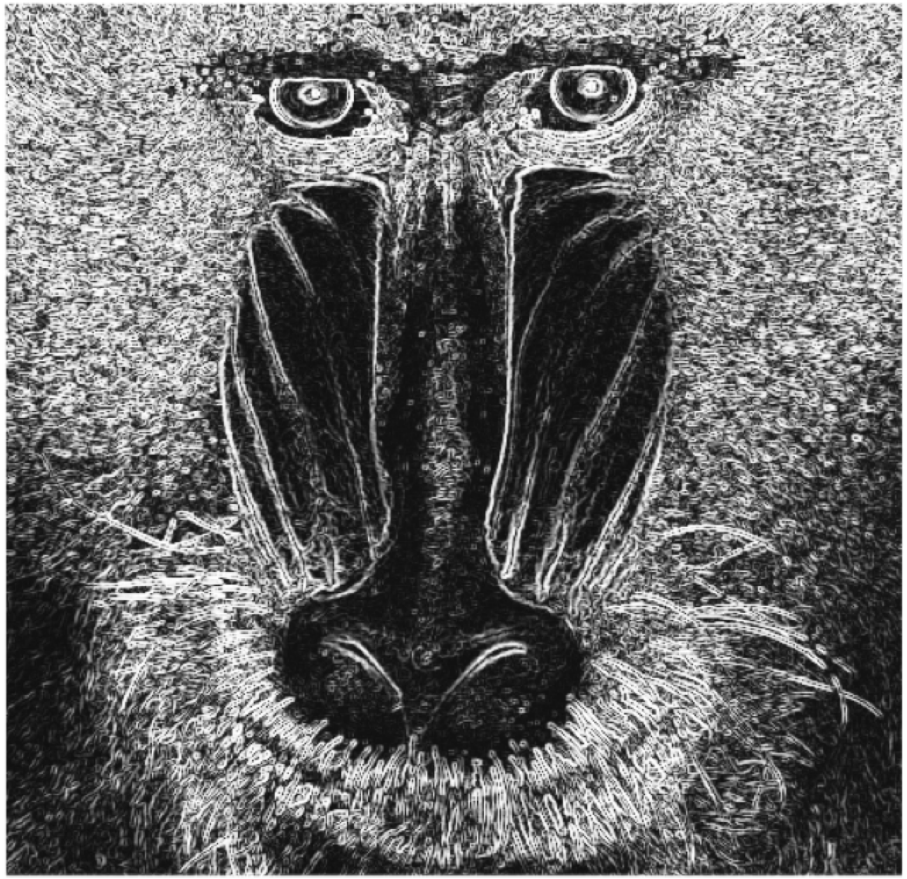
- Powerful prediction tools
- Highly parallel
- Efficiently implementable with hardware
  - Both digital and analog
- Fault tolerant

NPU design alternatives

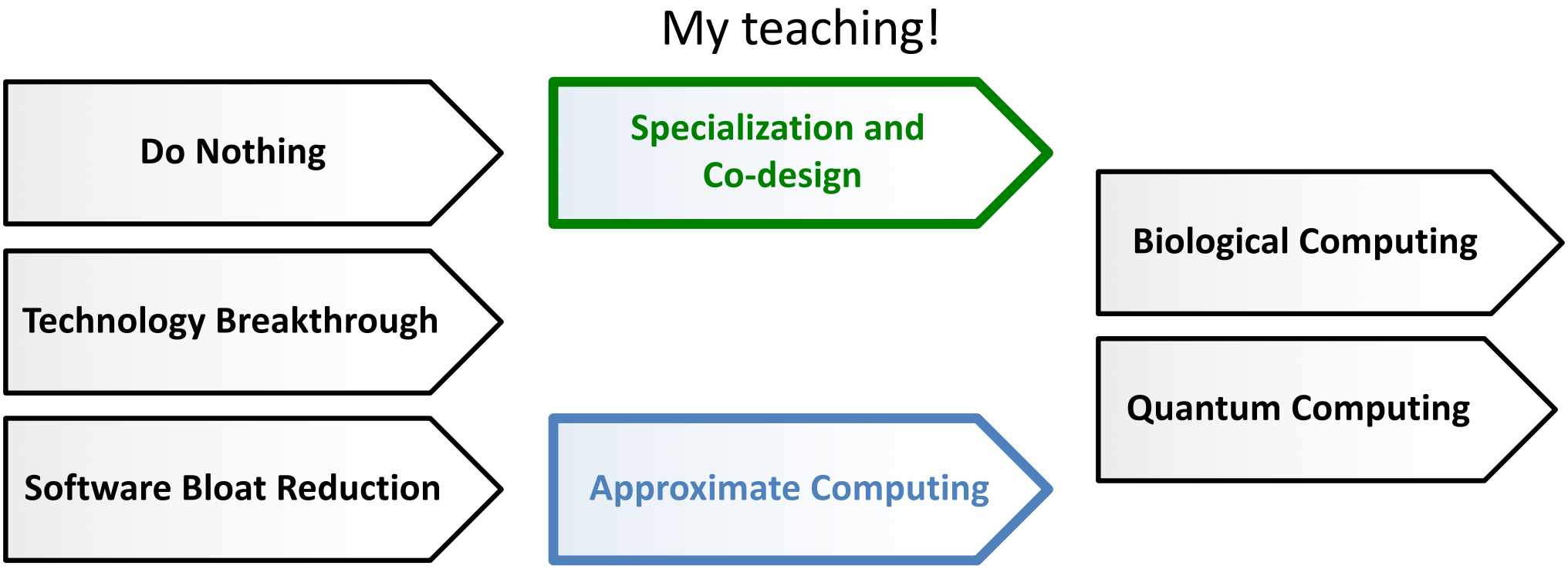


Approximate computing versus conventional computing

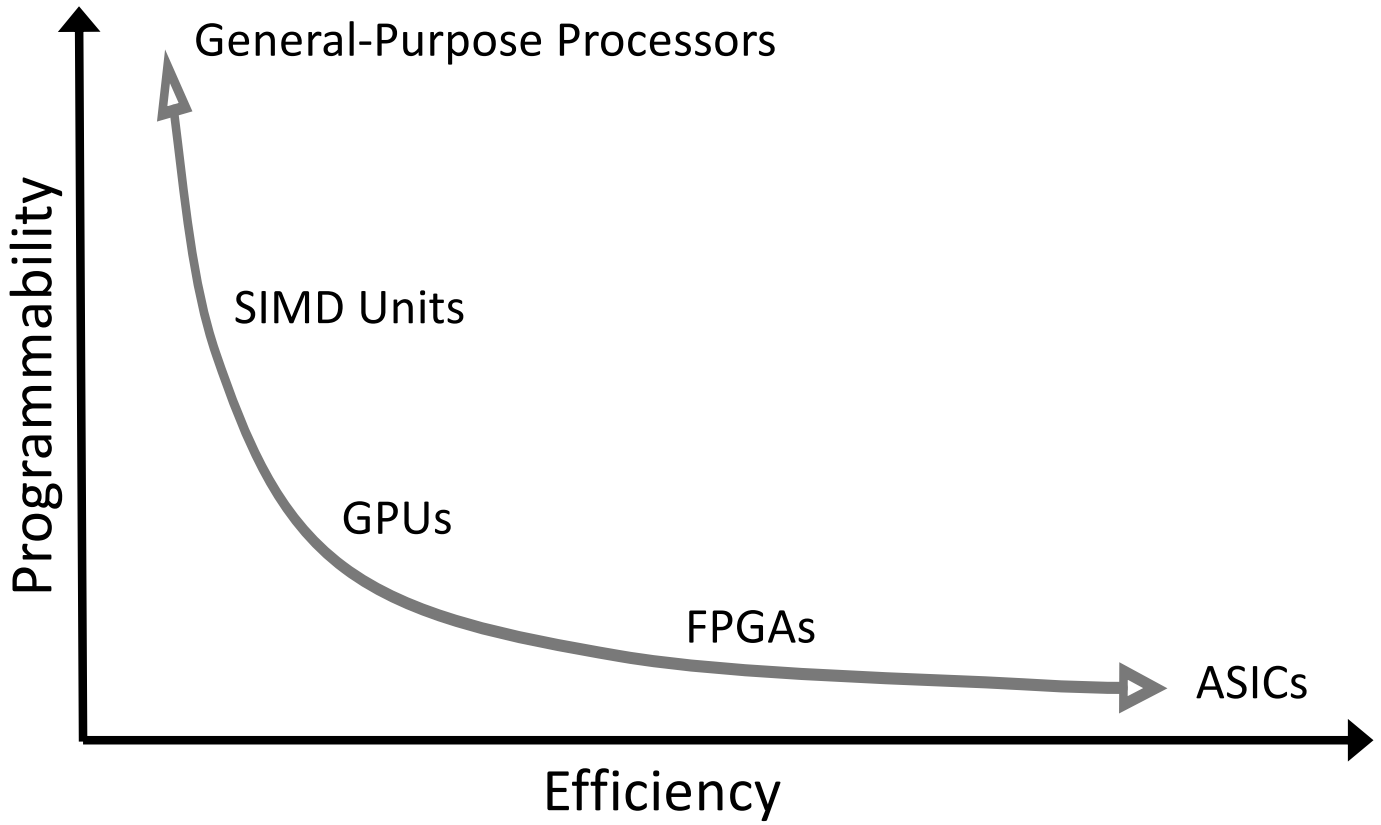




Possible paths forward



Programmability versus Efficiency




MAGES 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
Image Credits
© 2014 Microsoft | Privacy and Cookies | Legal | Advertise | About our ads | Help

Microsoft Cloud Services

Windows Azure

bing

Microsoft SQL Server

XBOX LIVE

COSMOS

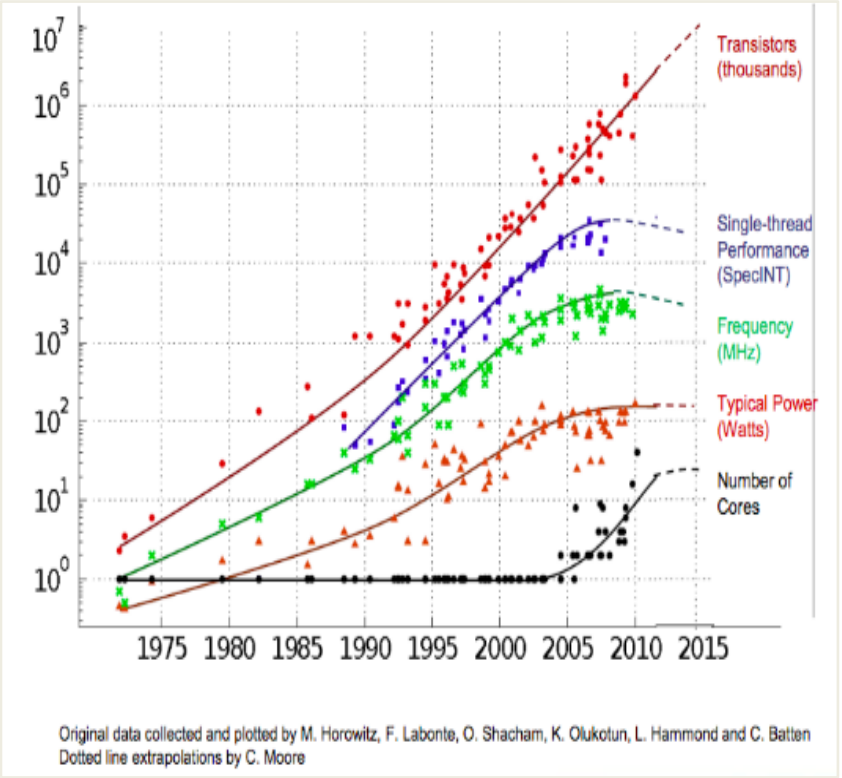
skype

msn

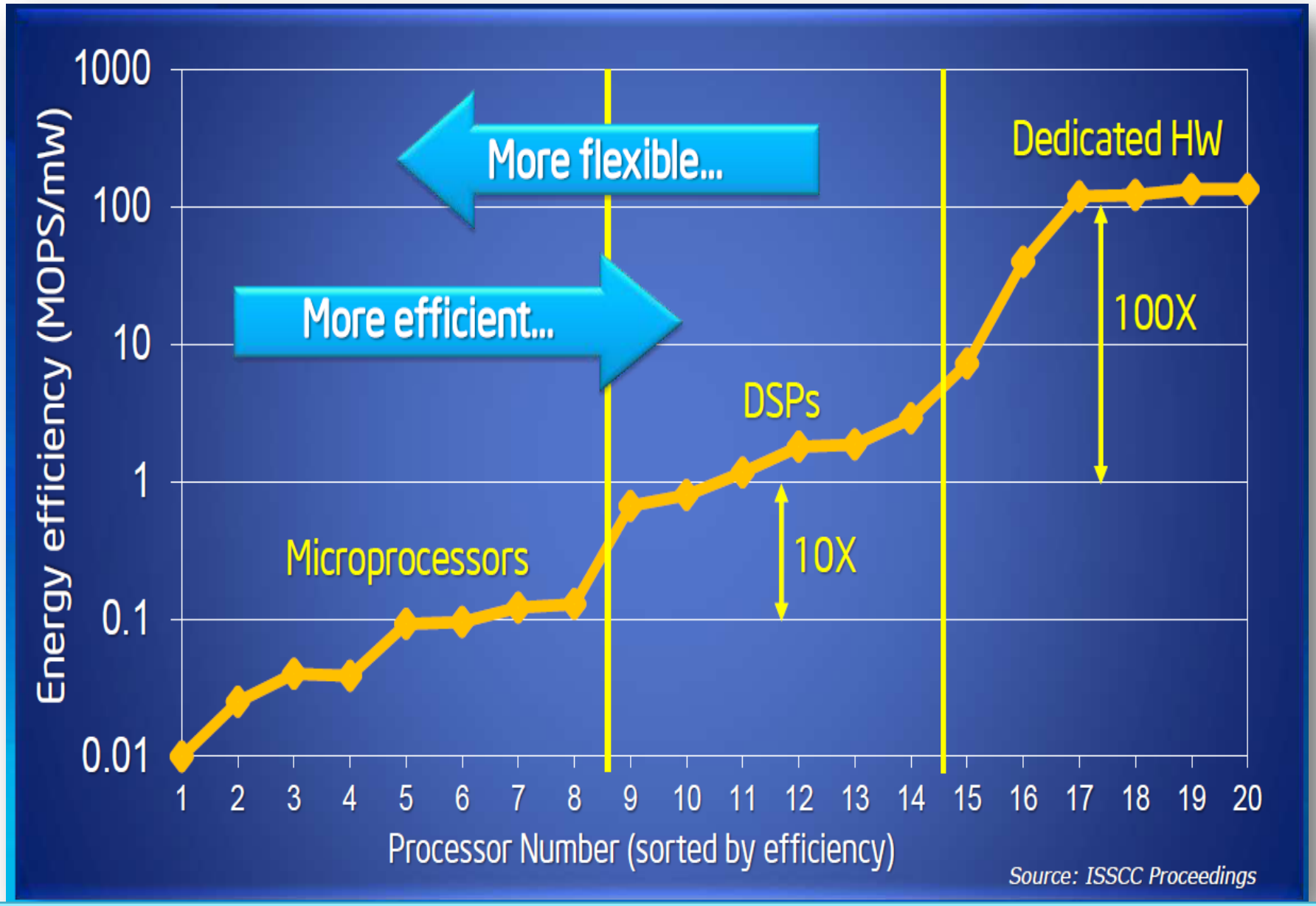
OneDrive

Outlook.com

Microsoft Office 365

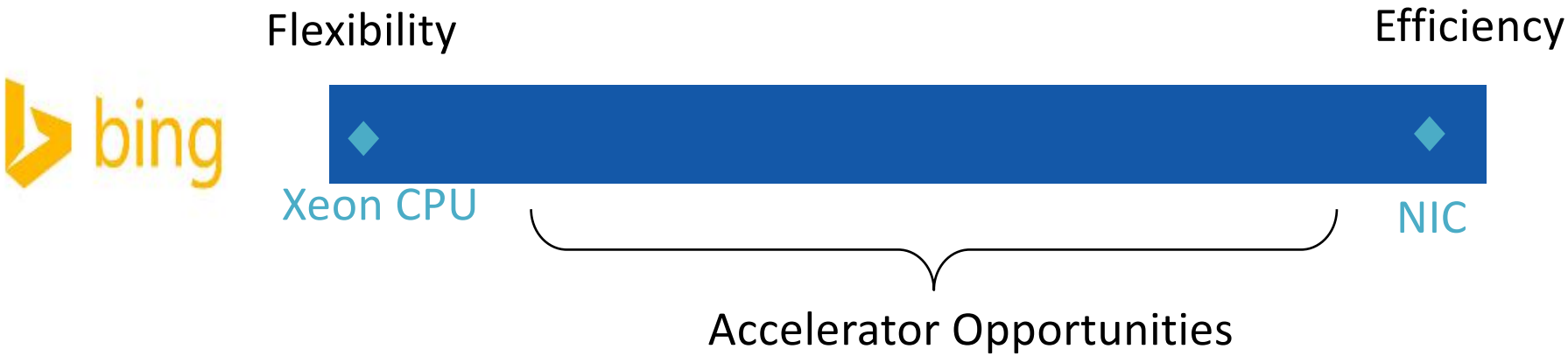


Capabilities, Costs



Increase Efficiency with Hardware Specialization

One Application's Accelerator

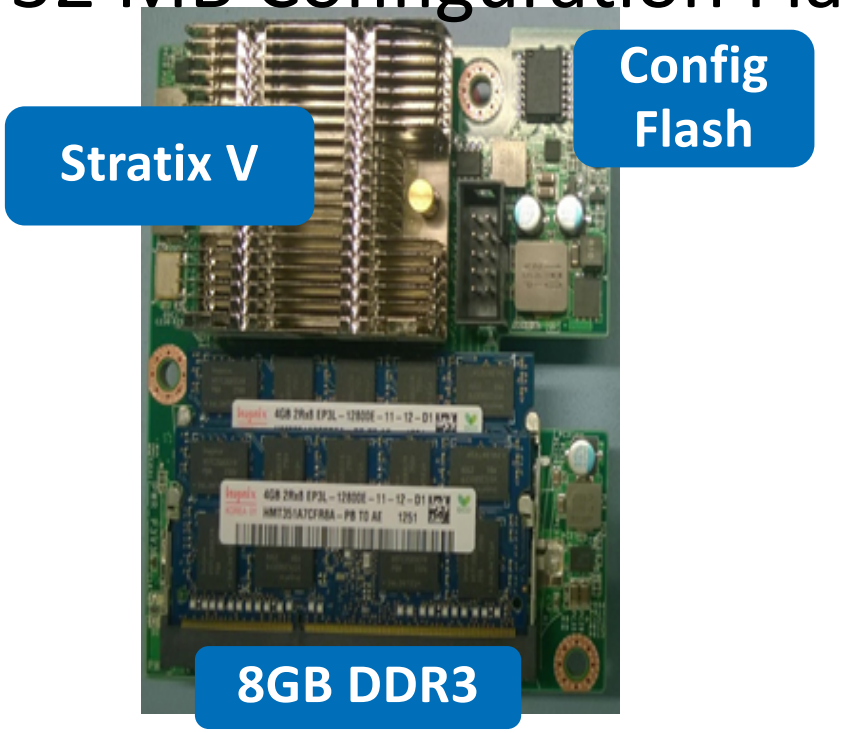


Catapult FPGA Accelerator Card

–Altera Stratix V GS D5
- 172k ALMs, 2,014 M20Ks, 1,590 DSPs

–8GB DDR3-1333

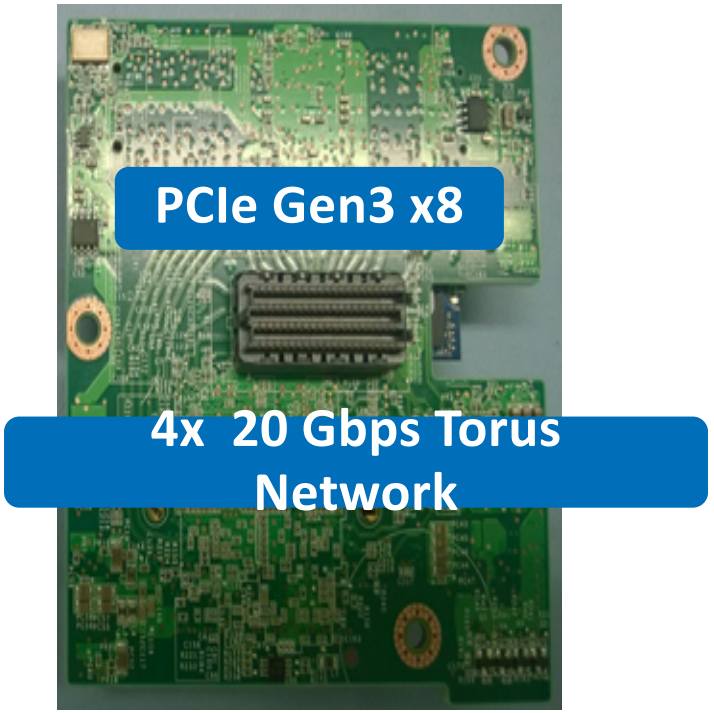
–32 MB Configuration Flash



–PCIe Gen 3 x8

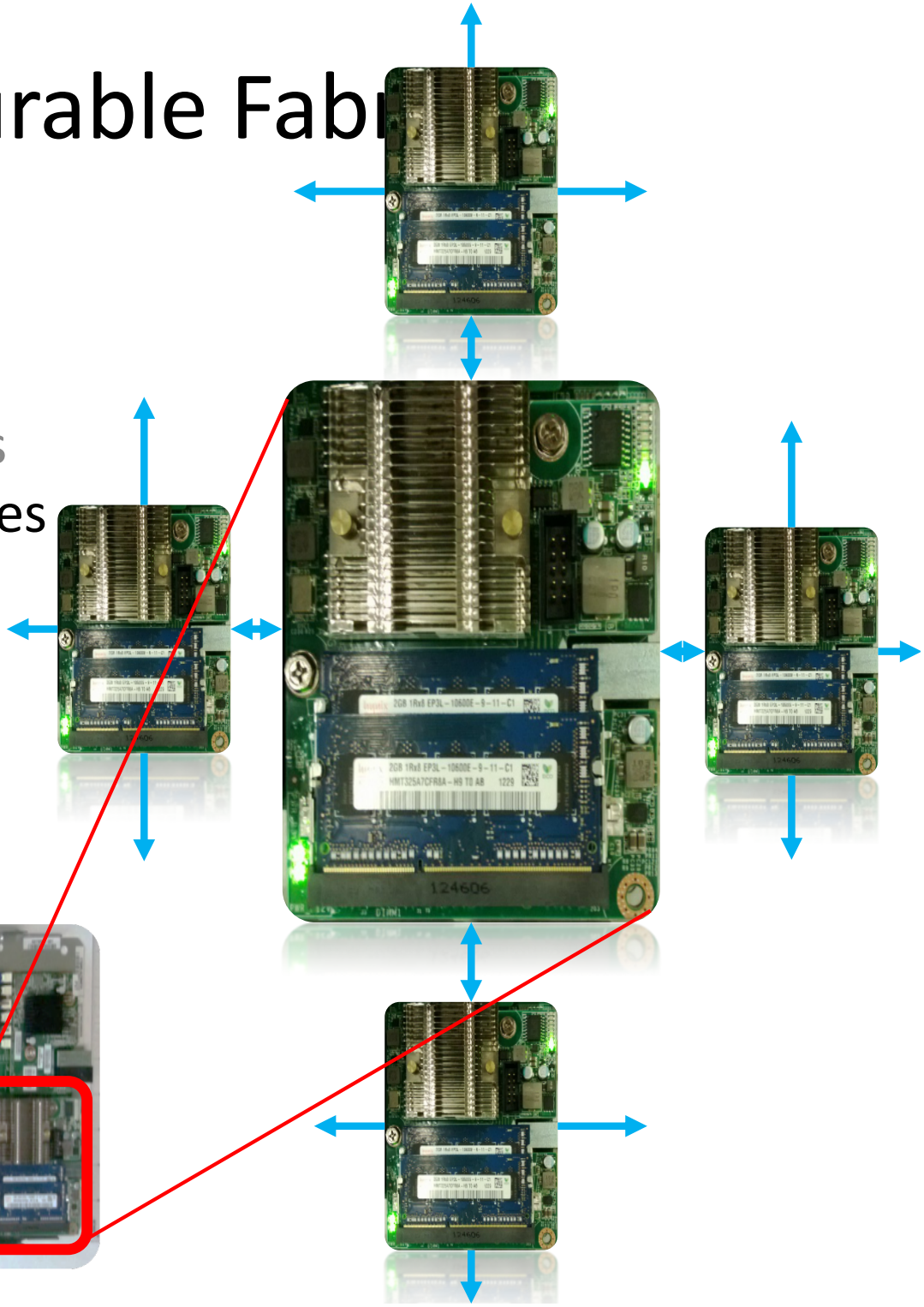
–8 lanes to Mini-SAS SFF-8088 connectors

–Powered by PCIe slot

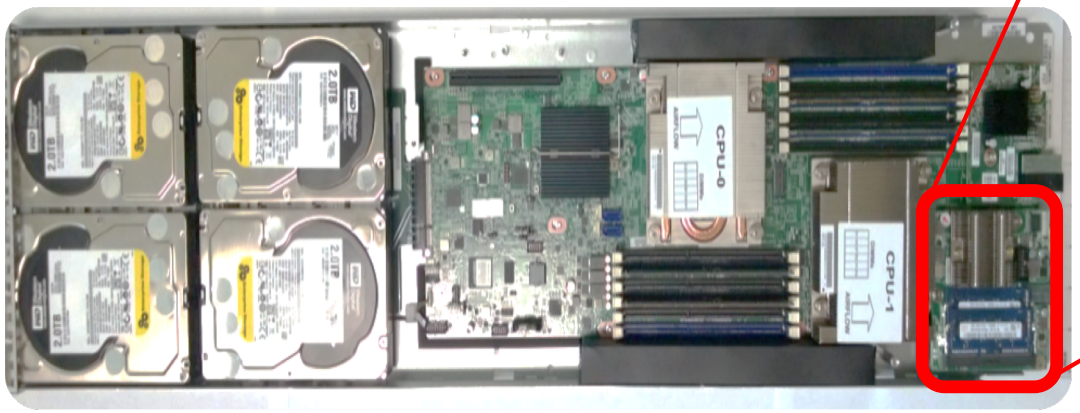


Scalable Reconfigurable Fabr

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

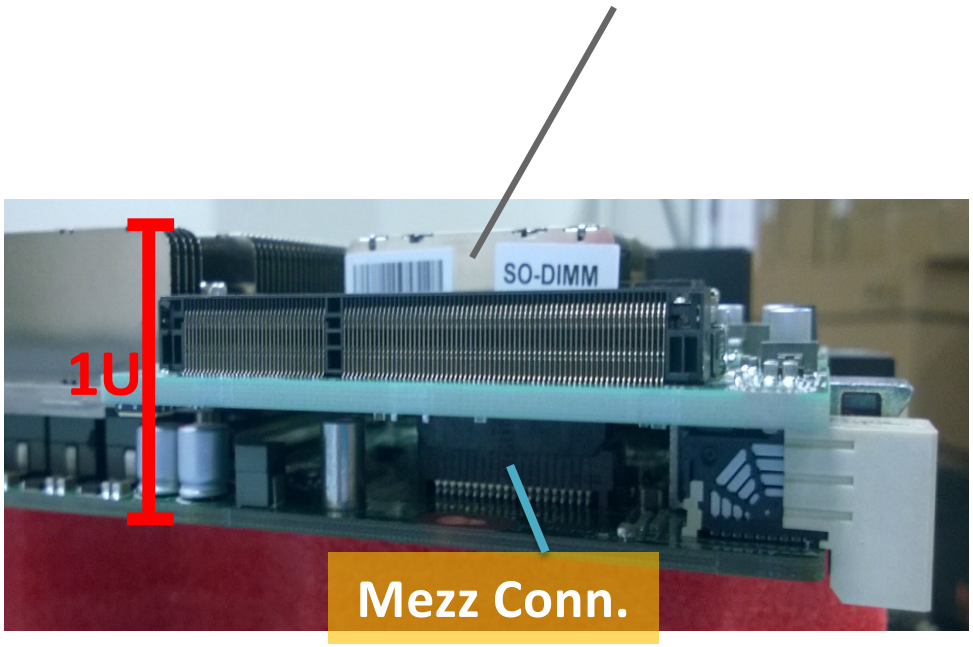


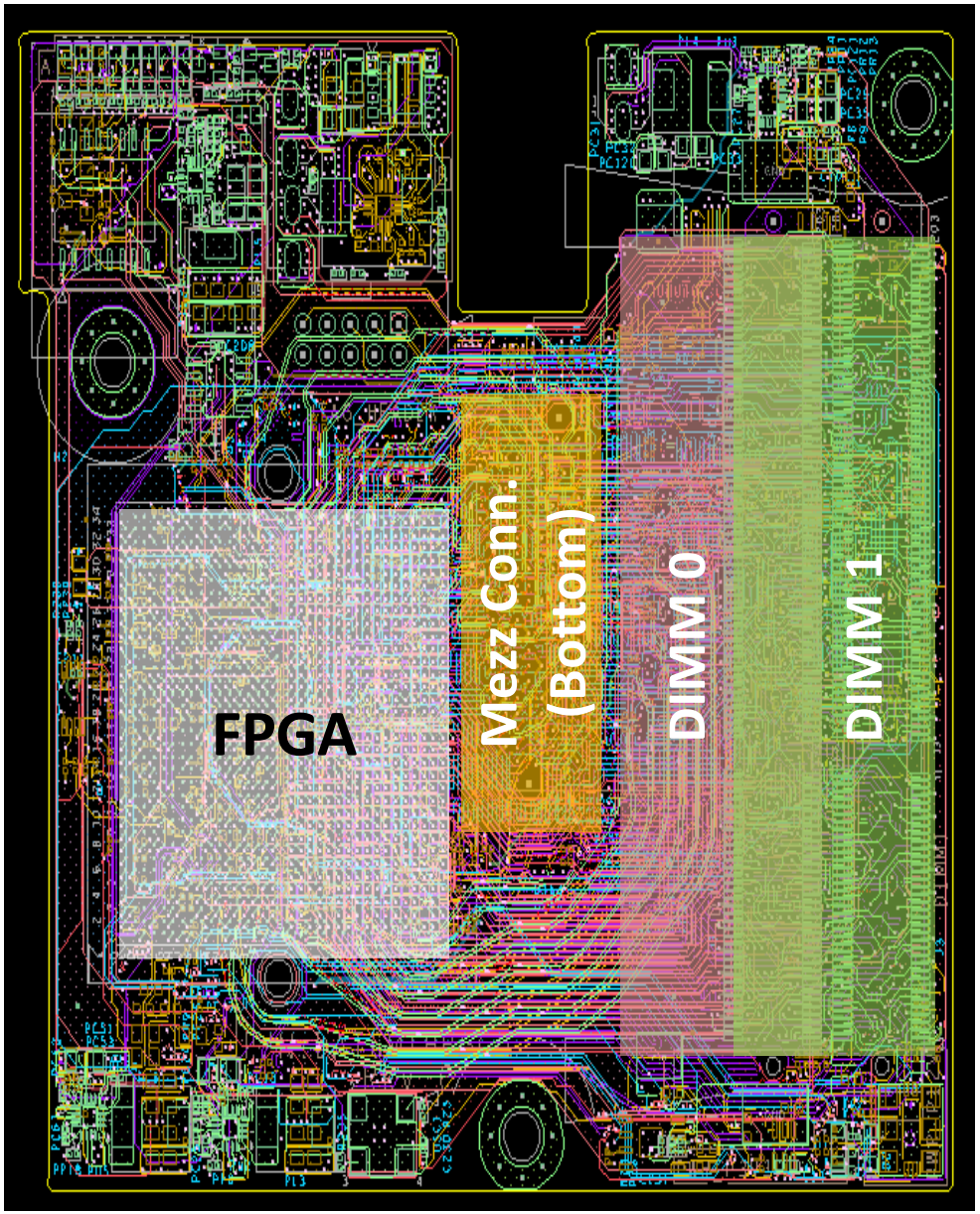
Data Center Server (1U, ½ width)



Board Details

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





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

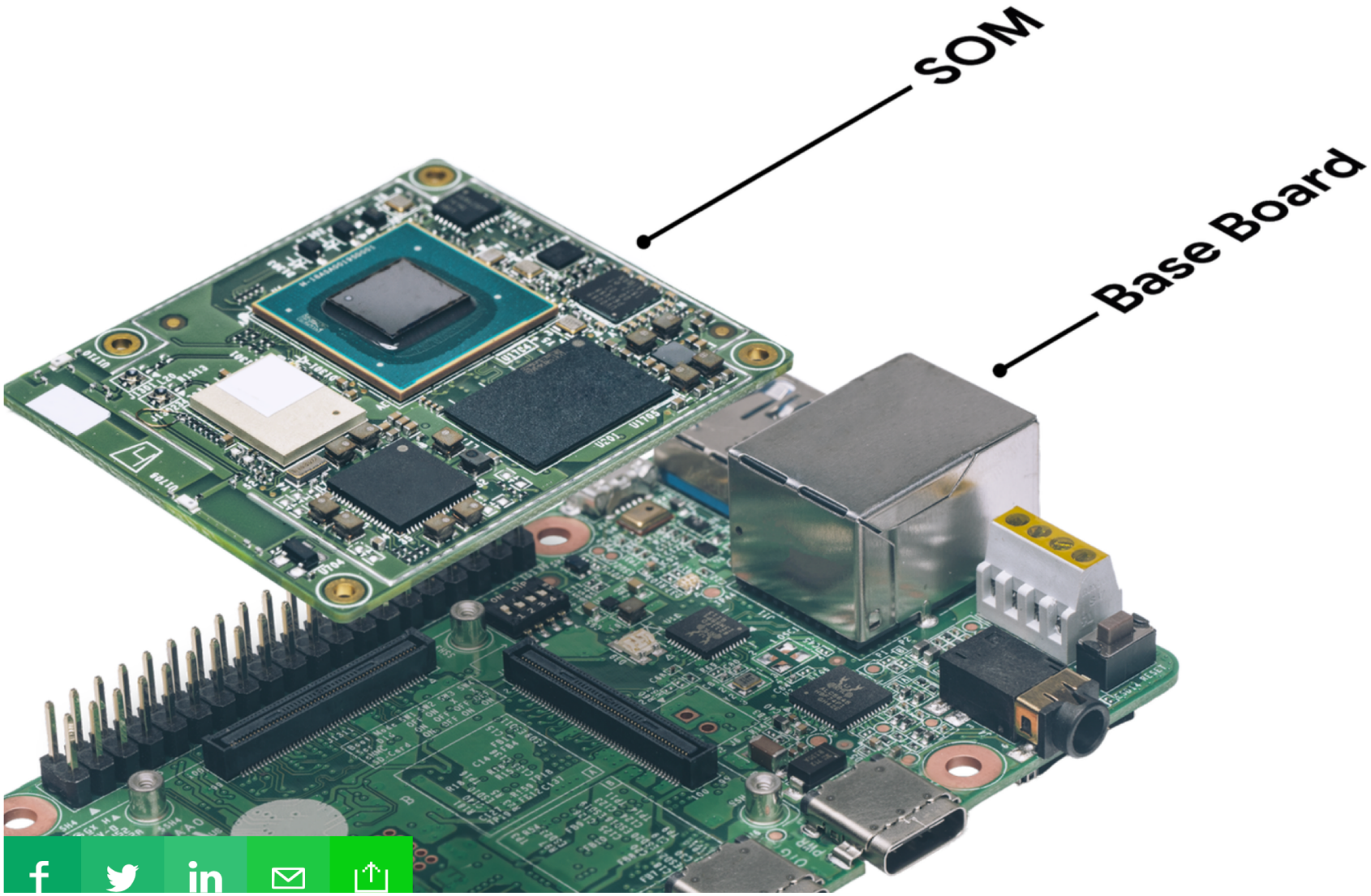
Intel Acquiring Altera in $16.7B Chipmaker Combination

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

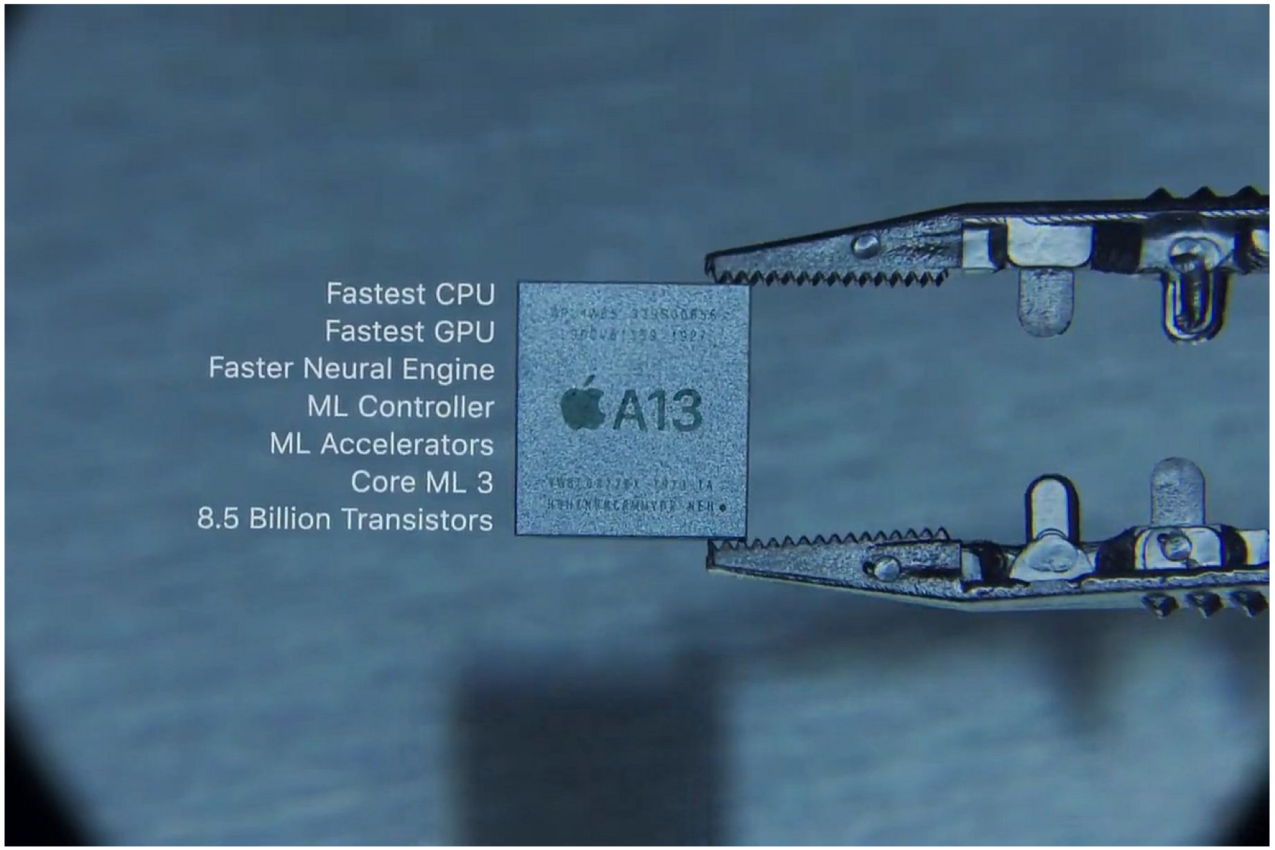


Apple says its new A13 Bionic chip brings hours of extra battery life to new iPhones

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.