Preparing for a Post Moore's Law World

Todd Austin

University of Michigan

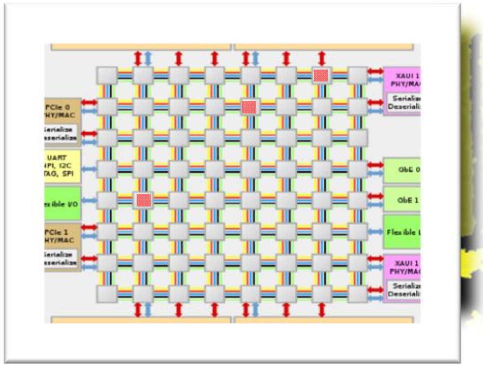
Perspectives on Scaling

- **C-FAR**: Center for Future Architectures Research
  - Focused on scaling in 2020-2030 silicon
  - Performance, power and cost
  - 27 faculty at 14 universities, 92 students

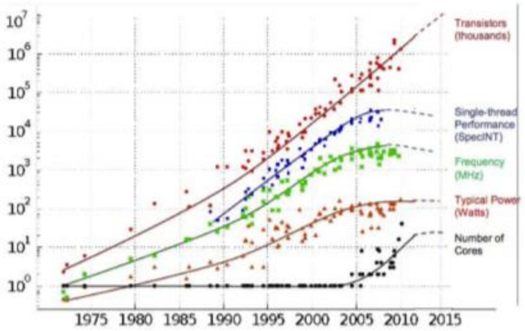
- **Why i**
  - The
- **Why i**
  - The threats… slowing innovation and degrading silicon

All of the work presented in this talk is that of C-FAR faculty.

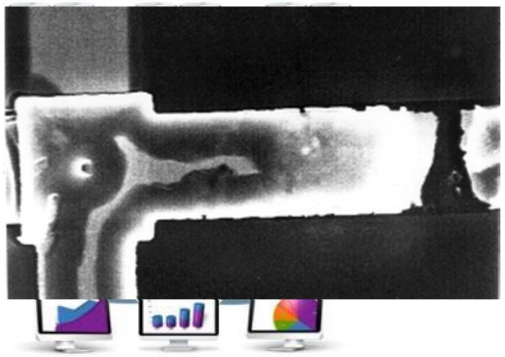
Many Cores
Computer Vision



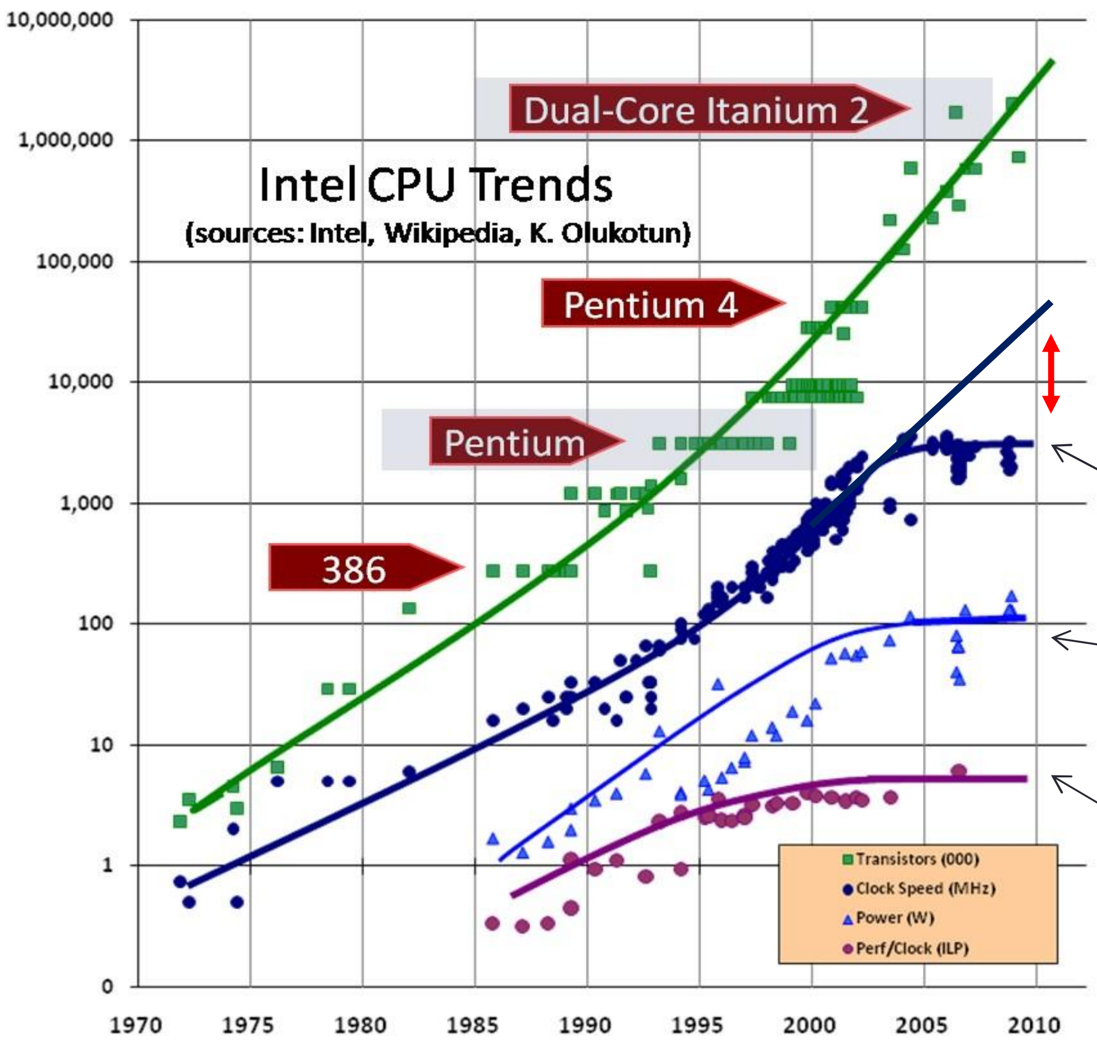
End of Dennard Scaling
Machine Learning



Big Data Analytics
Silicon Defects



Moore’s Law Performance Gap



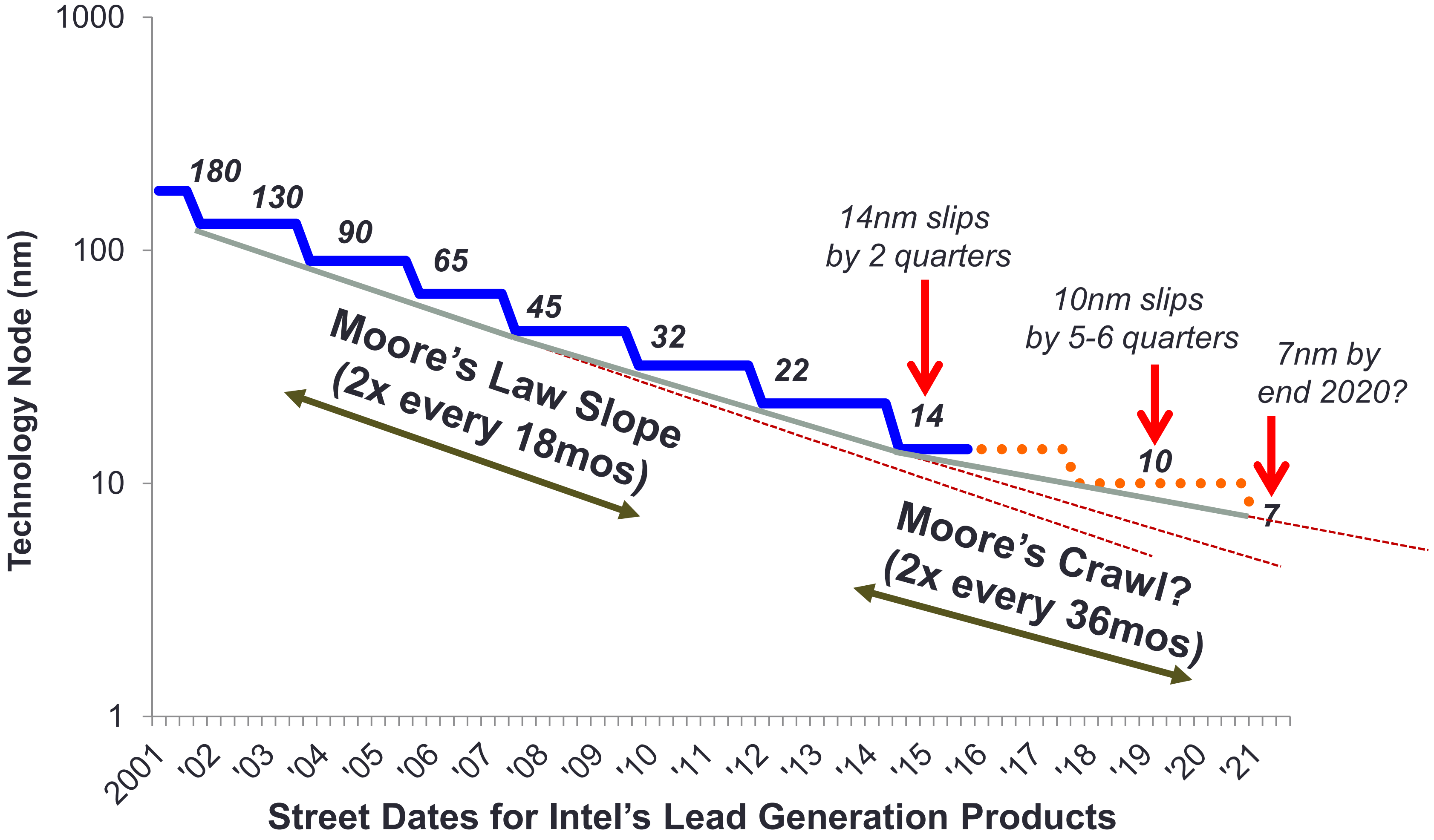
Today, gap is cresting 10x

Lack of perceived value

Dark silicon

Diminished ILP

Is Density Still Scaling?



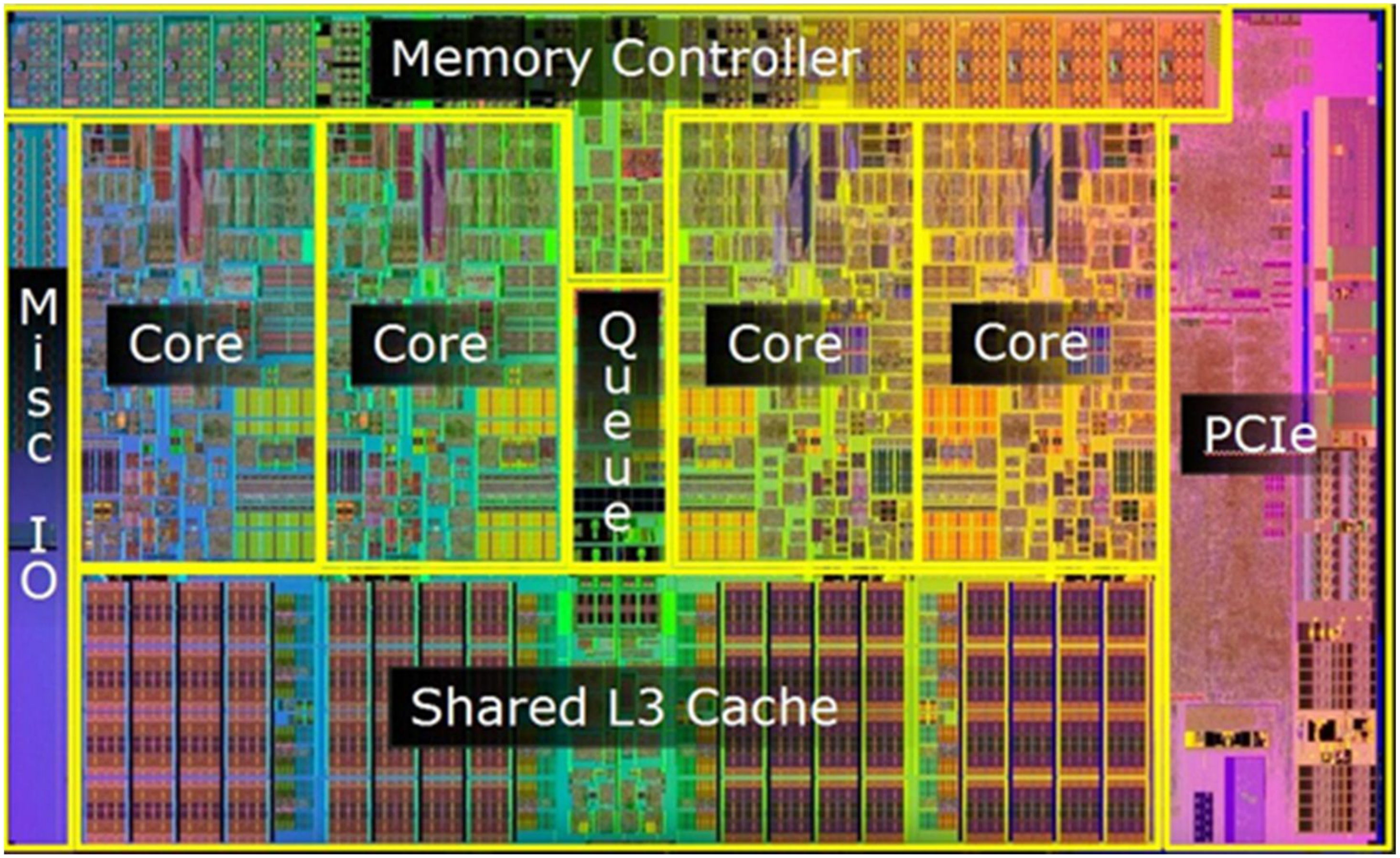
Courtesy David Brooks @ Harvard

What Does This All Mean to Architects?

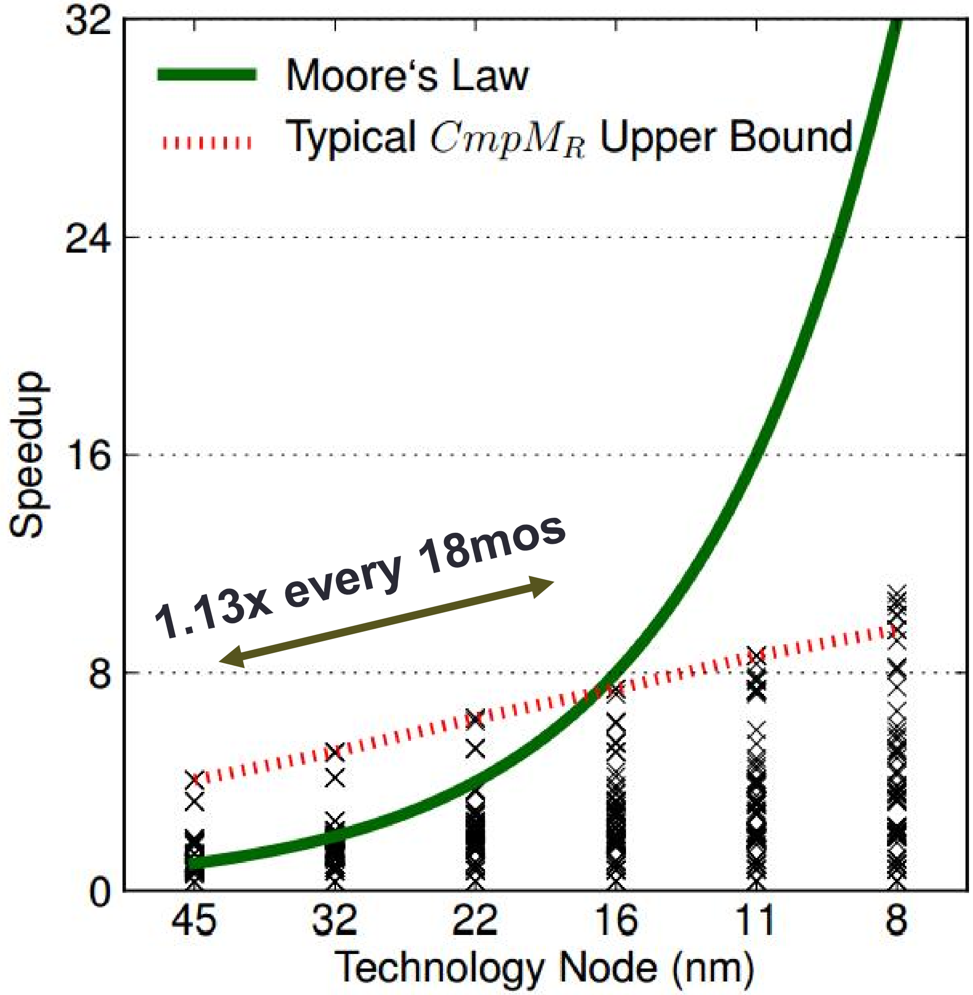
Today, value = scalability (performance, power, cost).

But, the technology scaling component has left us.

Remedy #1: Chip Multiprocessors

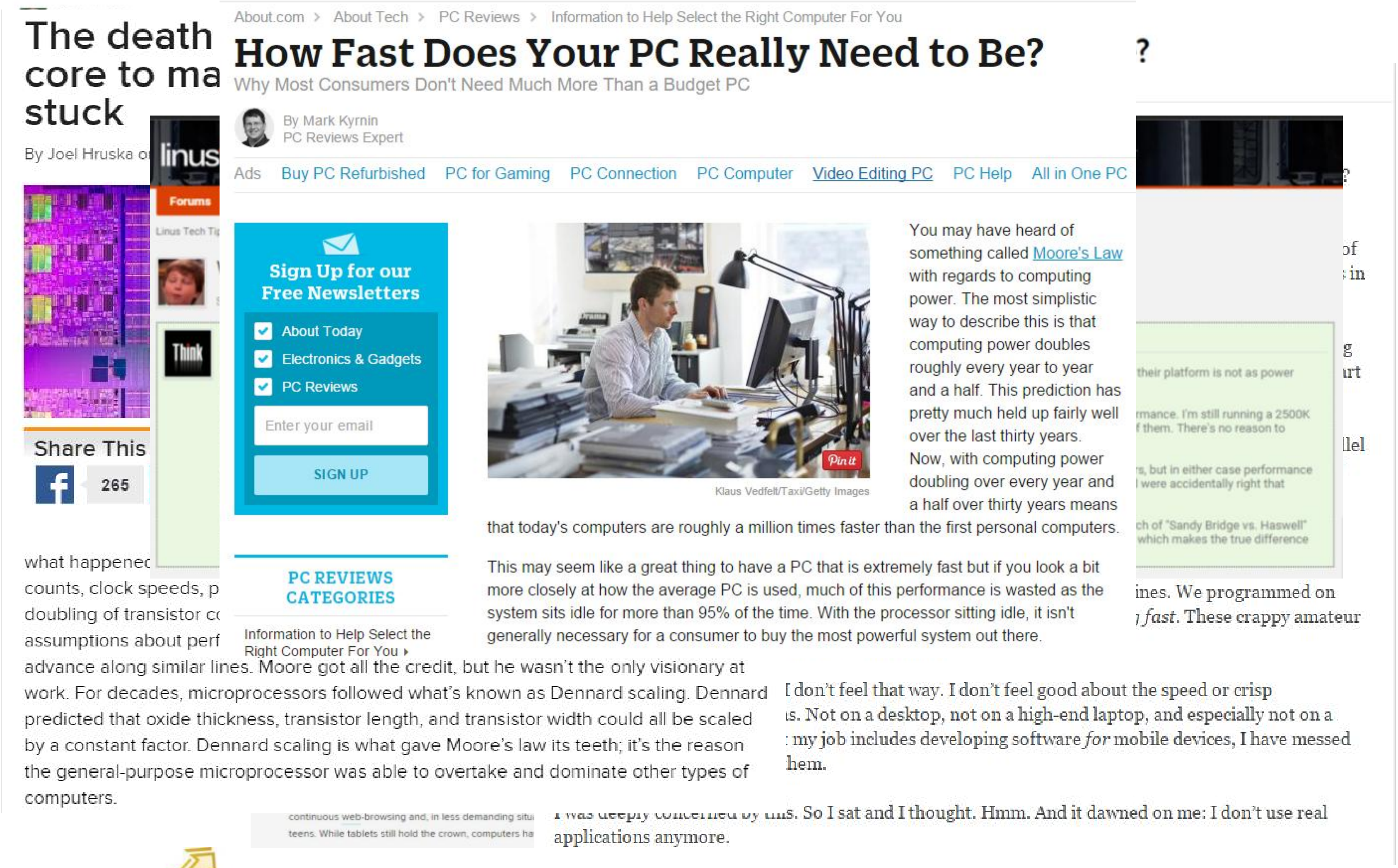


CMP Performance Scaling for the Highly Parallel PARSEC Benchmarks



From "Dark Silicon and the End of Multicore Scaling," by Esmaeilzadeh *et al*.

What Does the Press Think?



The death core to ma stuck

By Joel Hruska o

Share This

265

what happenec counts, clock speeds, p doubling of transistor c assumptions about perf advance along similar lines. Moore got all the credit, but he wasn't the only visionary at work. For decades, microprocessors followed what's known as Dennard scaling. Dennard predicted that oxide thickness, transistor length, and transistor width could all be scaled by a constant factor. Dennard scaling is what gave Moore's law its teeth; it's the reason the general-purpose microprocessor was able to overtake and dominate other types of computers.

linus

Forums

Linus Tech Ti

Think

their platform is not as power

rmance. I'm still running a 2500K f them. There's no reason to

s, but in either case performance were accidentally right that

ch of "Sandy Bridge vs. Haswell" which makes the true difference

continuous web-browsing and, in less demanding situ teens. While tablets still hold the crown, computers ha

About.com › About Tech › PC Reviews › Information to Help Select the Right Computer For You

How Fast Does Your PC Really Need to Be?

Why Most Consumers Don't Need Much More Than a Budget PC

By Mark Kyrnin
PC Reviews Expert

Ads Buy PC Refurbished PC for Gaming PC Connection PC Computer Video Editing PC PC Help All in One PC

Sign Up for our Free Newsletters

☑ About Today
☑ Electronics & Gadgets
☑ PC Reviews

Enter your email

SIGN UP

PC REVIEWS CATEGORIES

Information to Help Select the Right Computer For You ›

Pin it

Klaus Vedfelt/Taxi/Getty Images

You may have heard of something called Moore's Law with regards to computing power. The most simplistic way to describe this is that computing power doubles roughly every year to year and a half. This prediction has pretty much held up fairly well over the last thirty years. Now, with computing power doubling over every year and a half over thirty years means that today's computers are roughly a million times faster than the first personal computers.

This may seem like a great thing to have a PC that is extremely fast but if you look a bit more closely at how the average PC is used, much of this performance is wasted as the system sits idle for more than 95% of the time. With the processor sitting idle, it isn't generally necessary for a consumer to buy the most powerful system out there.

?

?

of

s in

g

rt

llel

ines. We programmed on

fast. These crappy amateur

I don't feel that way. I don't feel good about the speed or crisp

s. Not on a desktop, not on a high-end laptop, and especially not on a

my job includes developing software *for* mobile devices, I have messed

hem.

I was deeply concerned by this. So I sat and I thought. Hmm. And it dawned on me: I don't use real applications anymore.

We Investigate: Who’s to Blame?

?

Programmers

Largest NA Bitcoin Miner

- GPGPU-based system
- Fills 2000 sq.ft. warehouse
- Computes 1 petahash/s
- Reportedly generates $8M in Bitcoins per month

- Unfortunately soon to be obsolete as Bitcoin difficulty continues to scale

We Investigate: Who’s to Blame?

Educators

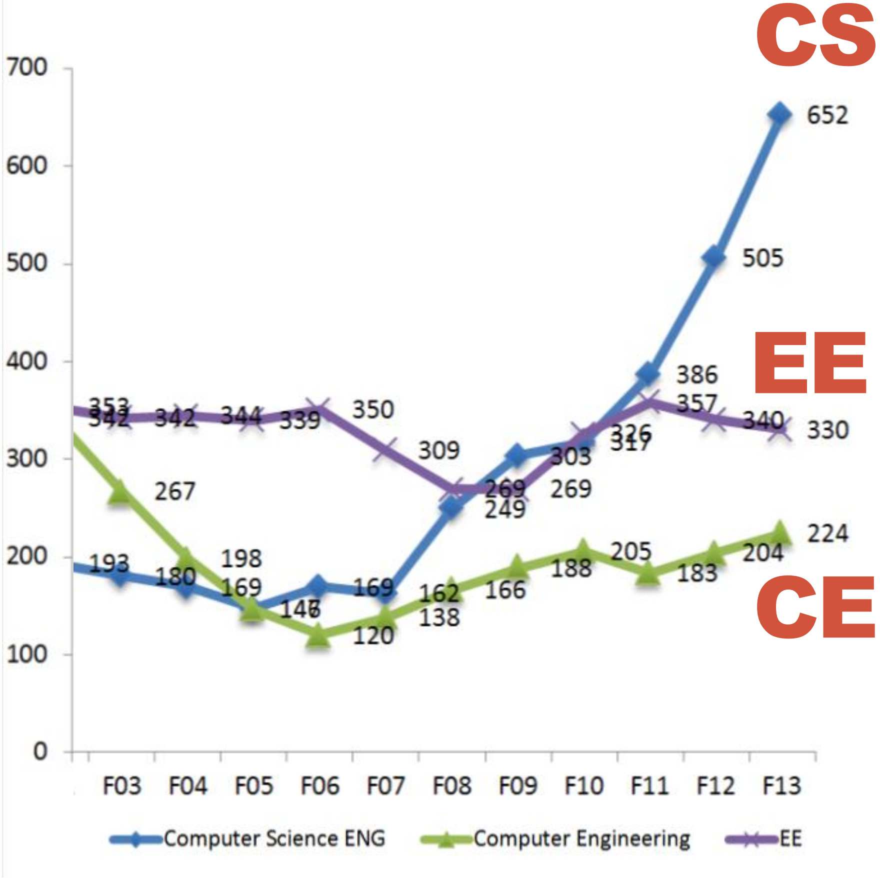
?

Programmers

CS Education is Booming

- CS enrollment on a fast-rising trajectory for a decade
- Parallel programming at UM
  - EECS 381, Object-Oriented and Advanced Programming
  - EECS 482, Operating Systems
  - EECS 570, Parallel Computer Architecture
  - EECS 587, Parallel Computing
  - EECS 591, Distributed Systems
  - EECS 598, Ubiquitous Parallelism
- I have been teaching and developing CS in Ethiopia
  - Nearly 600 students in the CS program
  - 2nd most popular major in the university

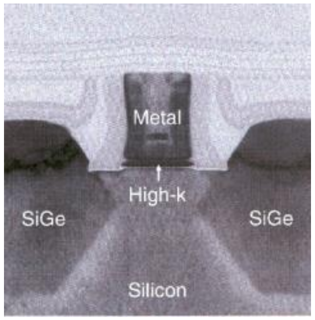
UM EECS Enrollment



We Investigate: Who’s to Blame?

Educators

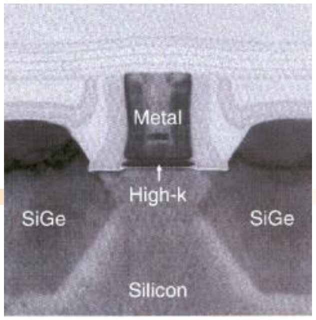
The Transistor



?

Programmers

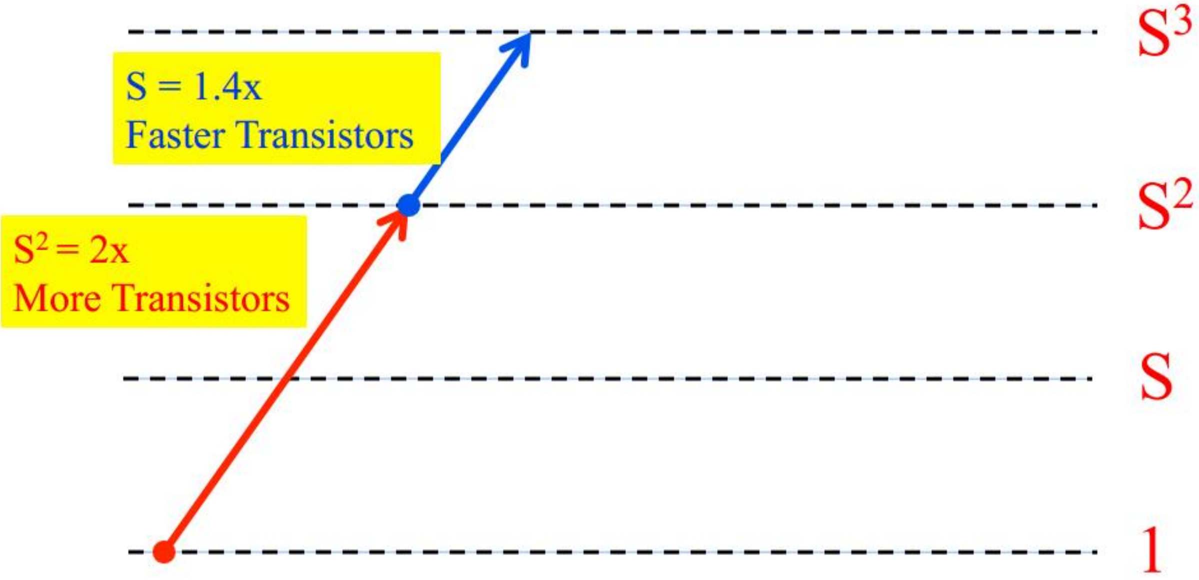
The Dark Silicon Dilemma



Advanced Scaling:

Dennard: "Computing Capabilities Scale by S^3 = 2.8x"

If S=1.4x …

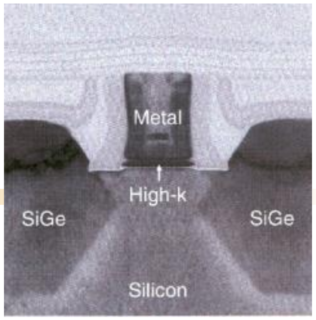


S = 1.4x
Faster Transistors

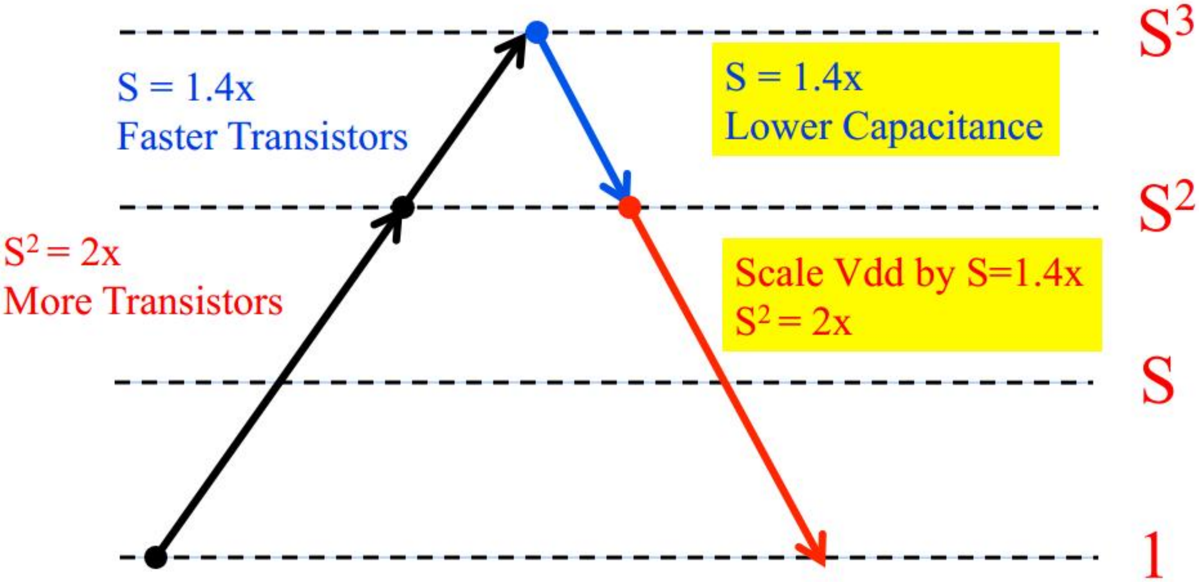
S^2 = 2x
More Transistors

Courtesy Michael Taylor @ UCSD

The Dark Silicon Dilemma



Dennard:
"We can keep power consumption constant"



S = 1.4x
Faster Transistors

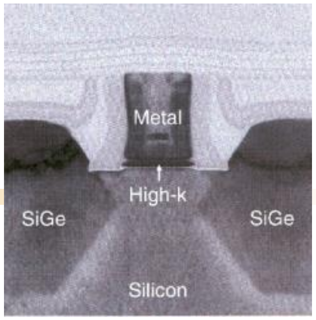
$S^2 = 2x$
More Transistors

S = 1.4x
Lower Capacitance

Scale Vdd by S=1.4x
$S^2 = 2x$

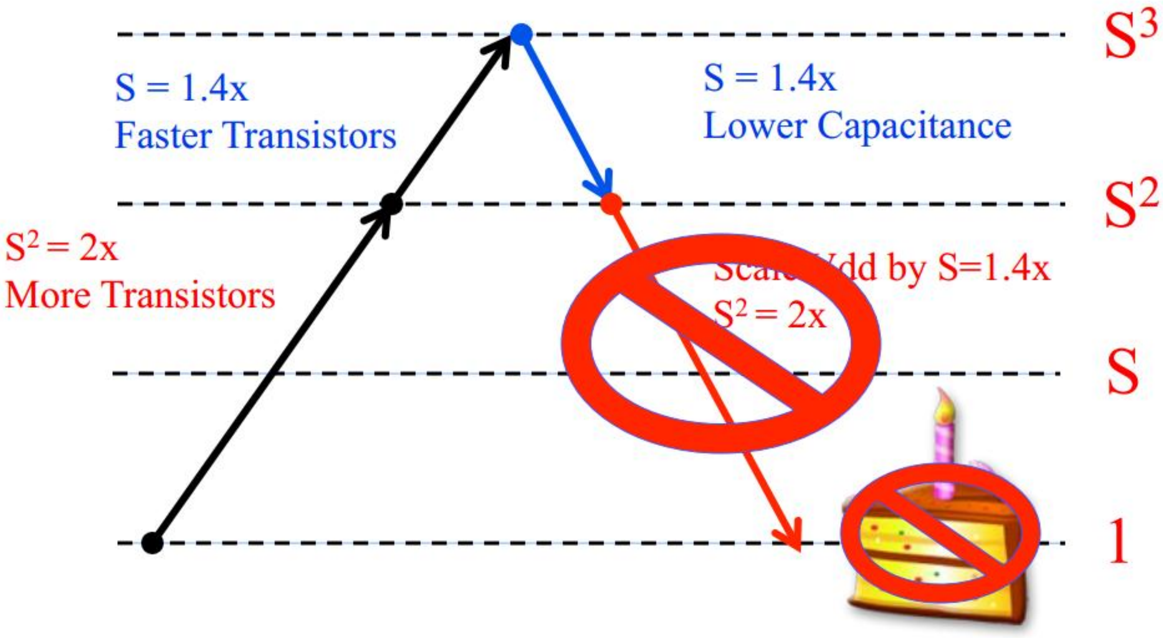
Courtesy Michael Taylor @ UCSD

The Dark Silicon Dilemma



Fast forward to 2005:

Threshold Scaling Problems due to Leakage Prevents Us From Scaling Voltage



S^3

S = 1.4x
Faster Transistors

S = 1.4x
Lower Capacitance

S^2

$S^2 = 2x$
More Transistors

Scale Vdd by S=1.4x
$S^2 = 2x$

S

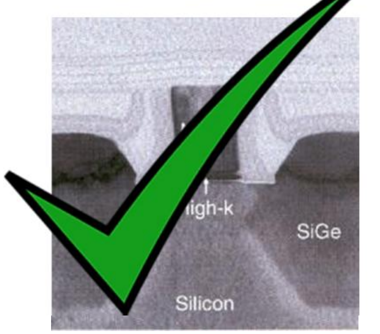
1

Courtesy Michael Taylor @ UCSD

We Investigate: Who’s to Blame?

Educators

The Transistor



?

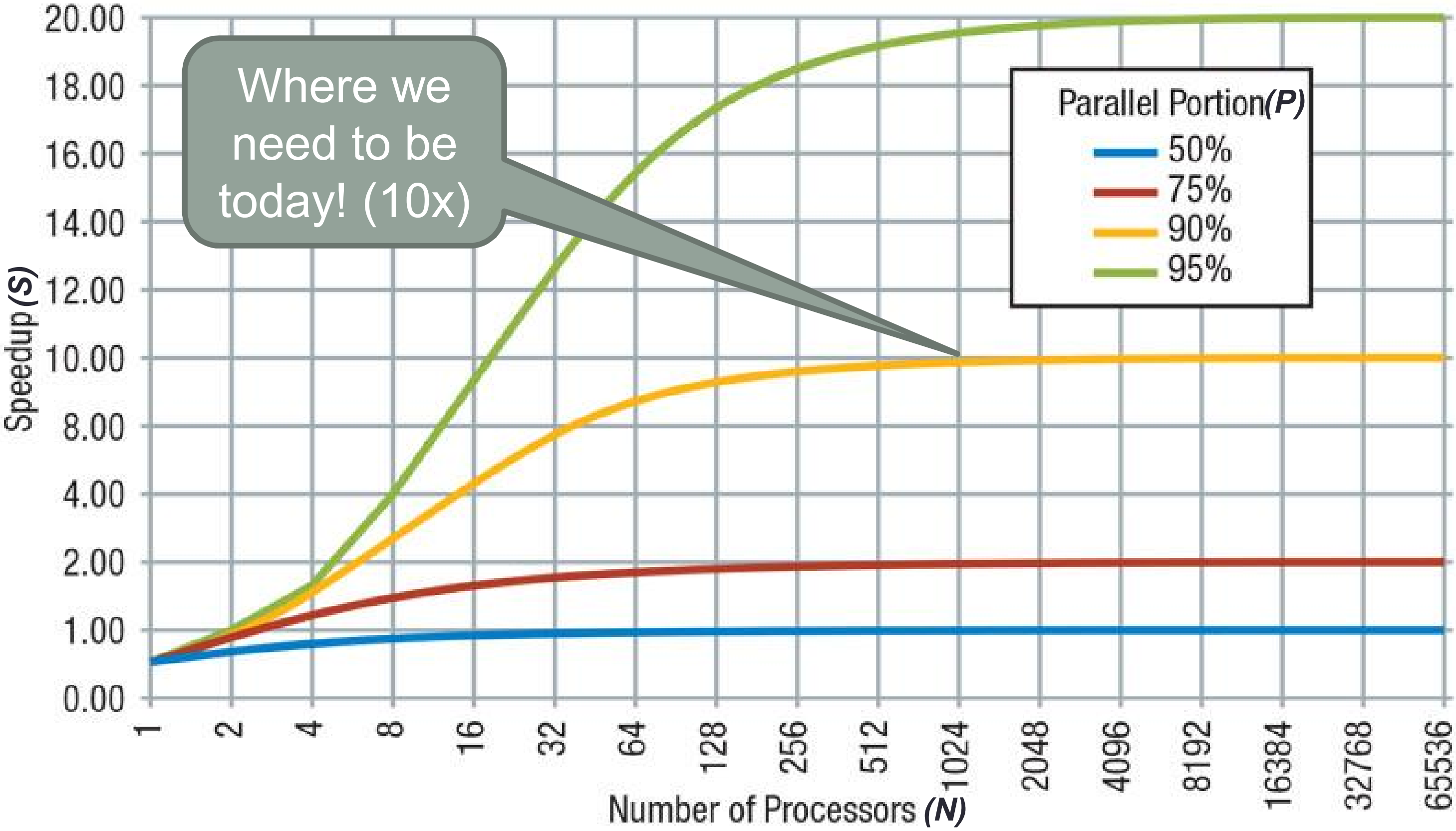
Programmers

Architects

The Tyranny of Amdahl’s Law

$$S(N) = \frac{1}{(1-P)+\frac{P}{N}}$$

Amdahl’s Law



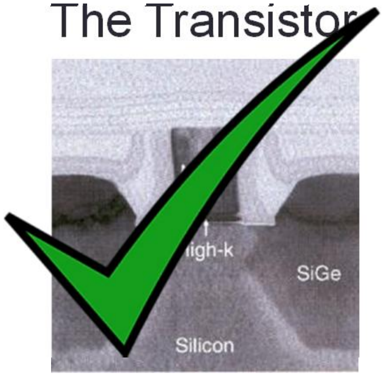
Where we need to be today! (10x)

We Investigate: Who’s to Blame?

Educators

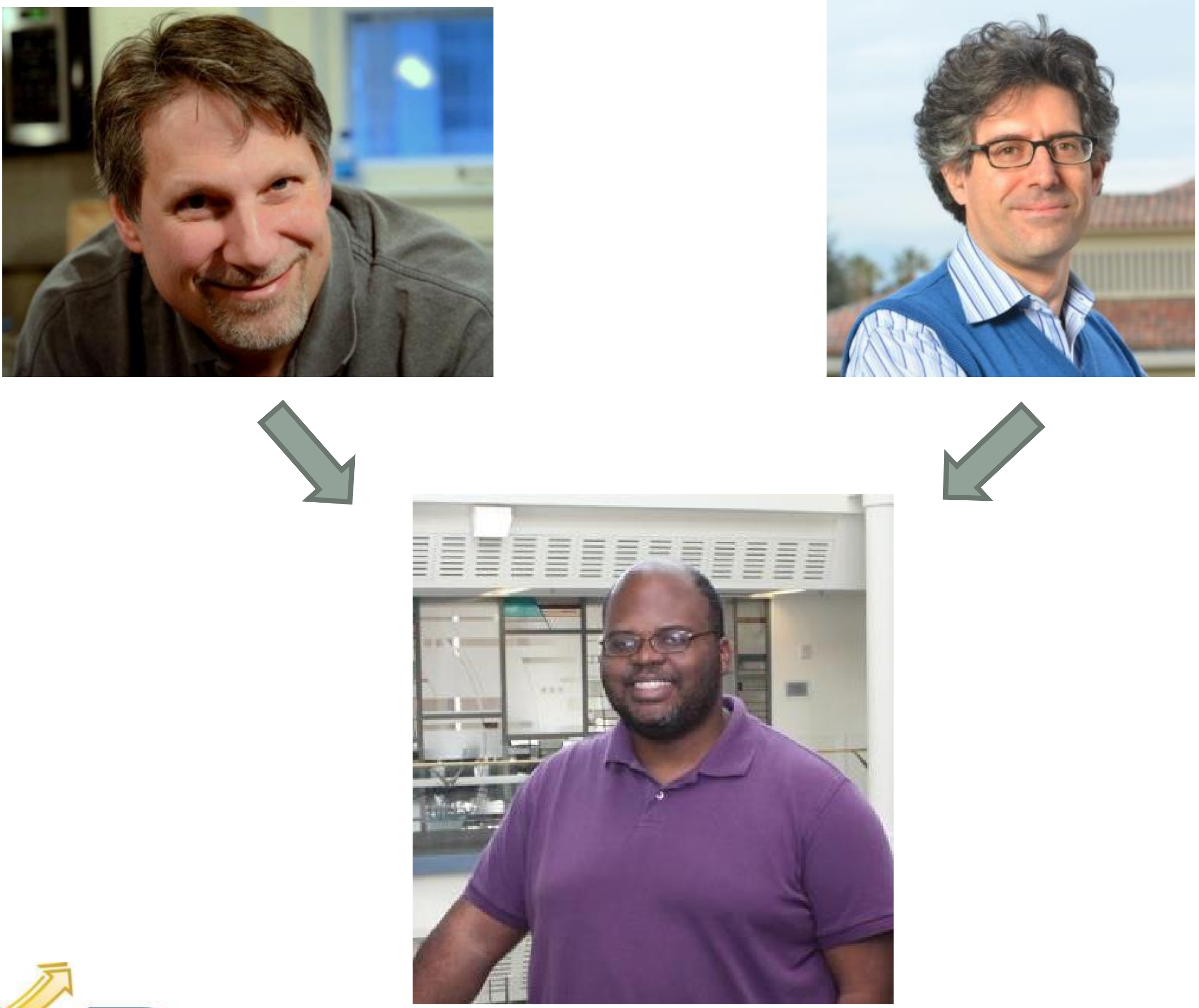
Programmers

The Transistor

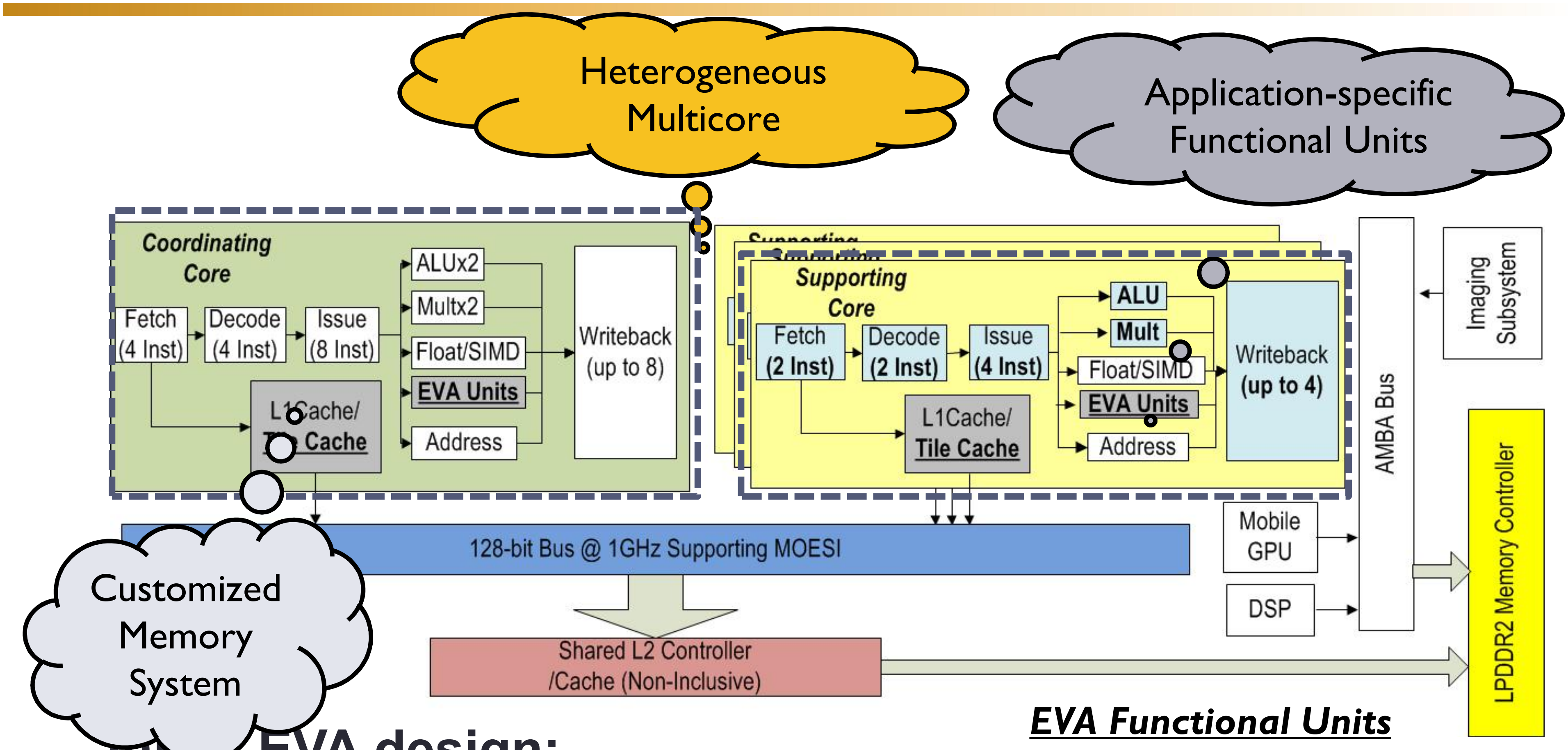


Architects

A Story about Jason and His Two Advisors



EVA: Embedded Vision Architecture



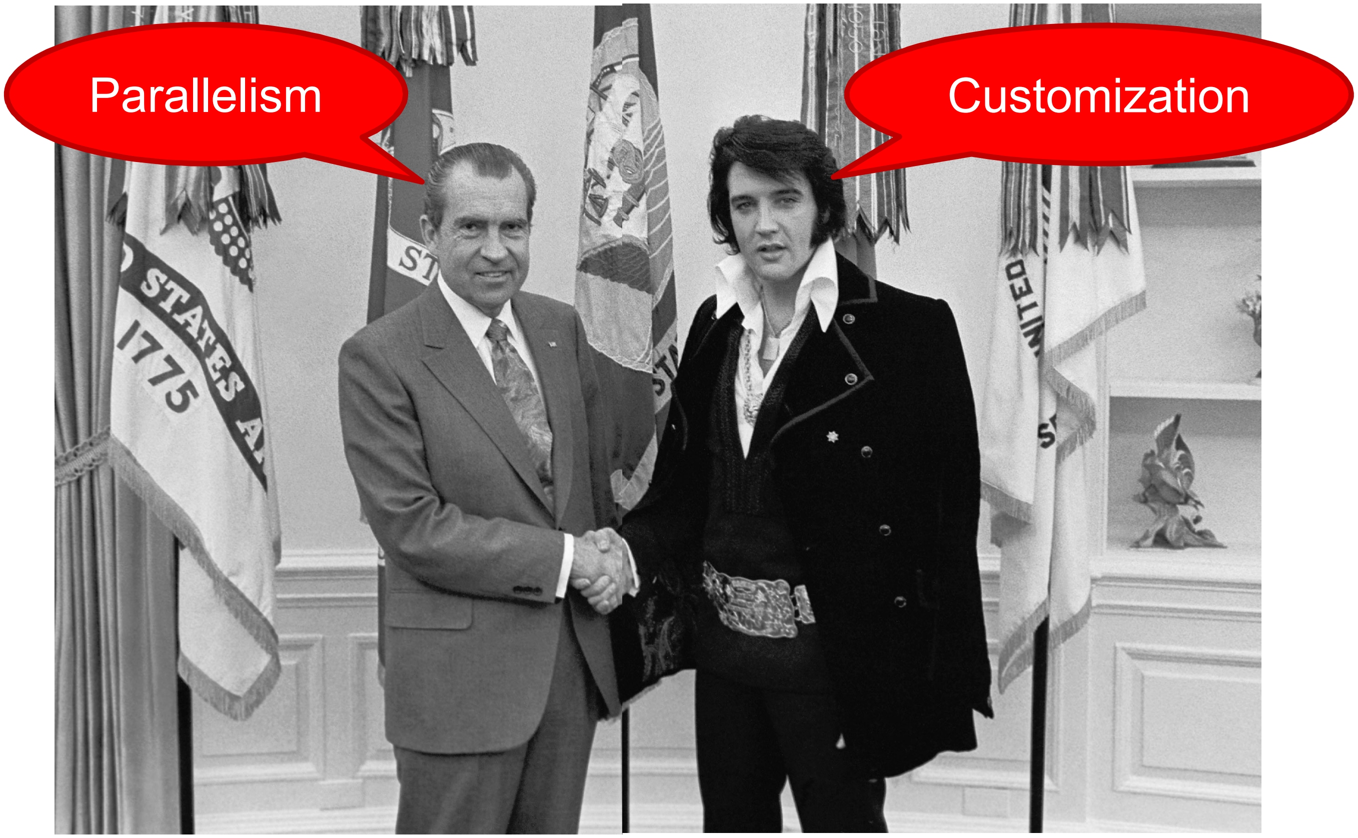
Initial EVA design:

90x greater efficiency for computer vision algorithms

EVA Functional Units

Monopoly Compare,
Dot Product Unit,
Vector Max,
Decision Tree Compare

Where We Need to Focus



Heterogeneous parallel systems

overcome ***dark silicon*** and the ***tyranny of Amdahl's Law.***

Why These Ideas Will Likely Fail, Unless We Make a Change…

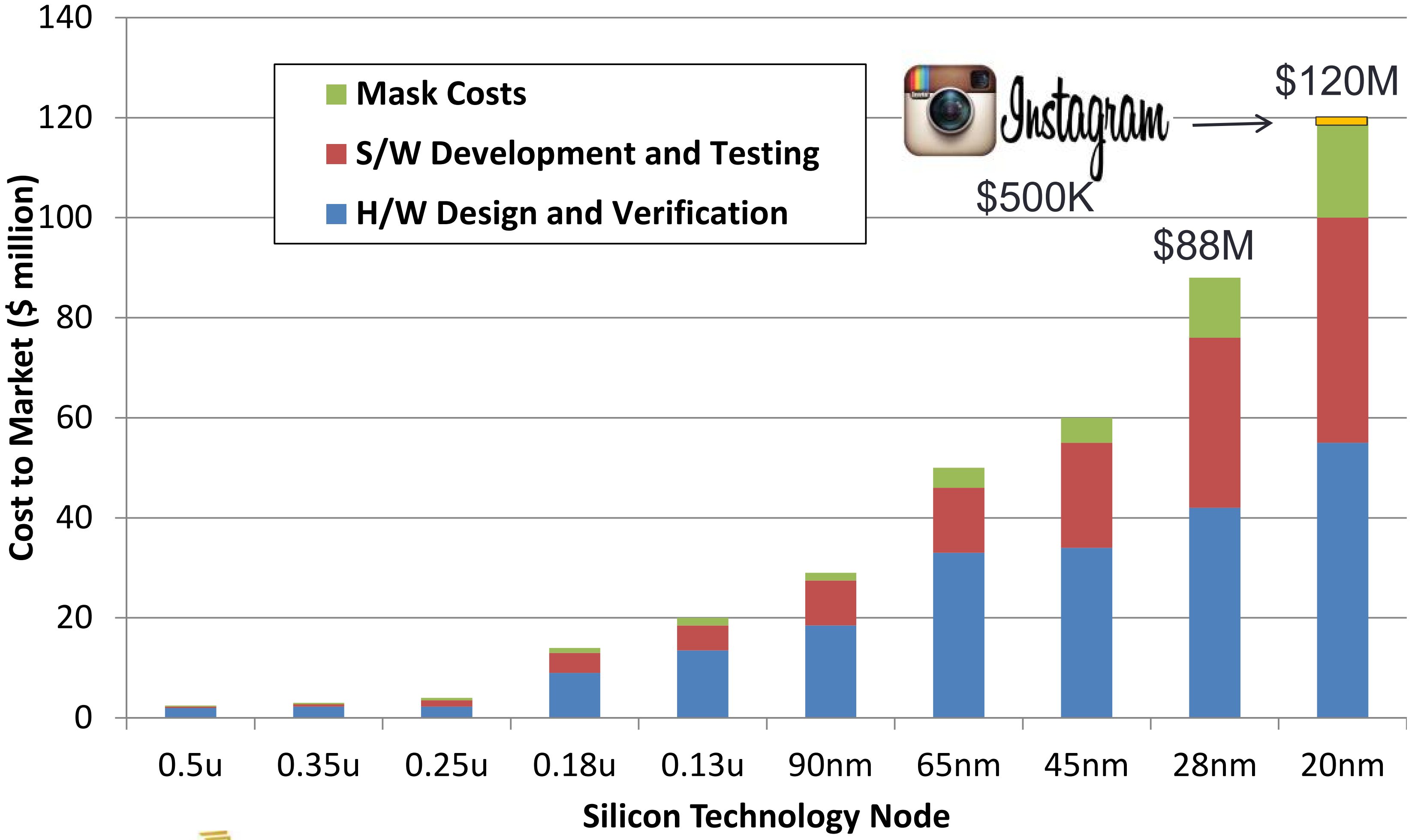
- ***The Good***: Hetero-parallel systems can close the Moore's Law gap
- ***The Bad***: Dennard scaling has stopped, Moore's Law is slowing, leaving a growing gap
- ***The Ugly***: Hetero-parallel designs needed to close the gap will be ***too expensive to afford***
    - We must make design much ***cheaper***!

What I Want You to Remember

- Successfully bridging the Moore’s Law performance gap is less about “***How***” to do it and more about “***How Much***” does it cost!

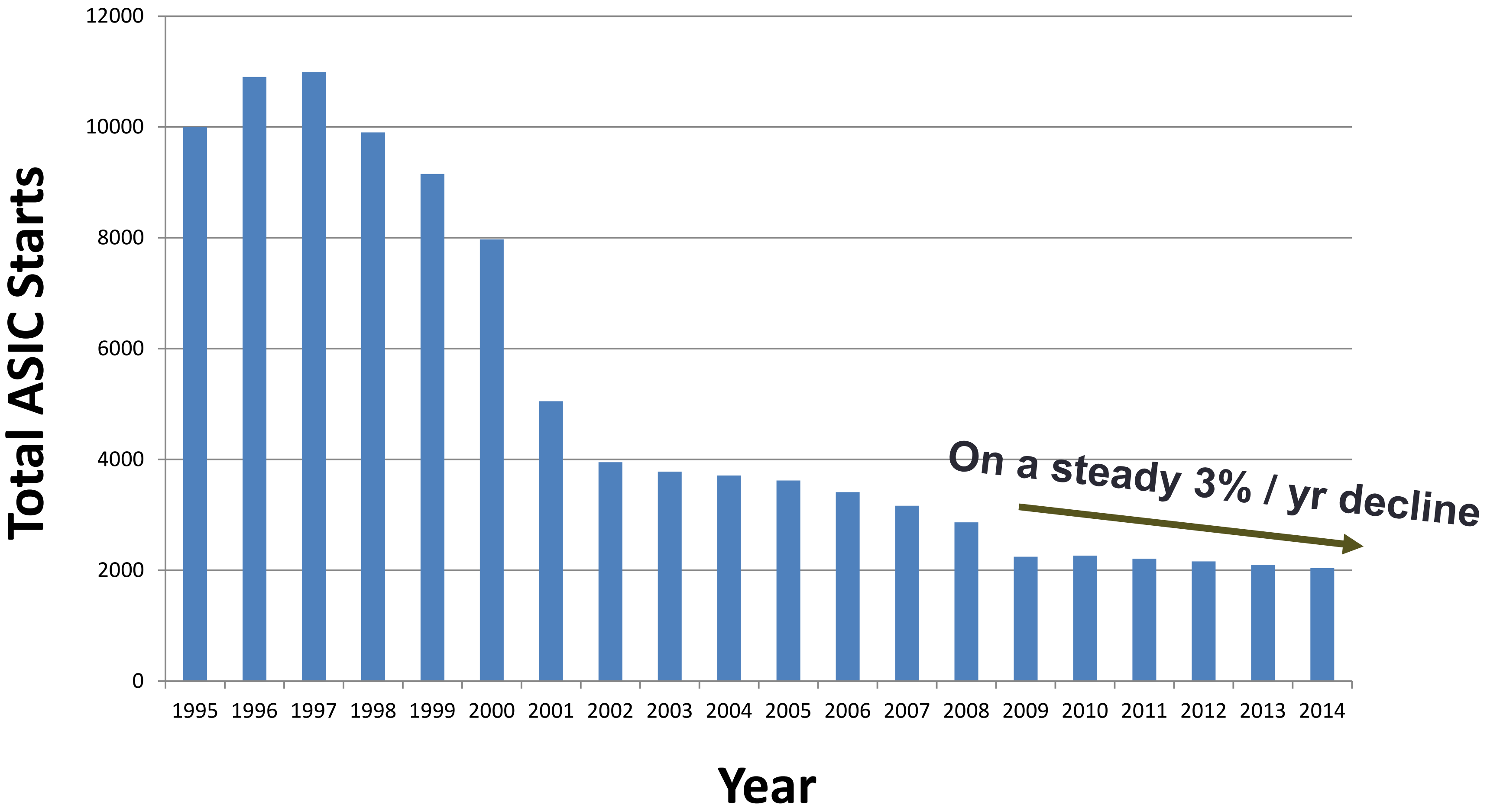
- **My claim:** if we can effect a ***100x reduction*** in the cost to bring a design to market, ***innovation will flourish*** and scaling challenges will be overcome.

Design Costs Are Skyrocketing



Source: International Business Strategies

Outcome: “Nanodiversity” is Dwindling



Source: Gartner Group

Inexpensive “Design” Promotes Innovation and Adaptation

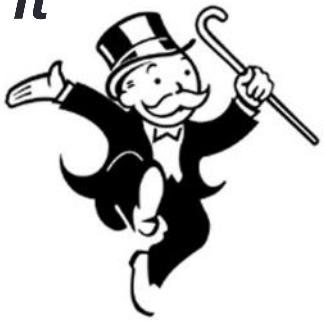
- Don’t Believe Me? Ask Mother Nature!
  - *r/K* selection theory is a biological mechanism that organisms use to better adapt to their environment

- In unstable environments, ***r-selection*** predominates as the ability to reproduce quickly is crucial

- In stable environments, ***K-selection*** predominates as the ability to compete successfully for limited resources is crucial

The Remedy: Scale Innovation

- Ultimate goal: ***accelerate system architecture innovation*** and make it sufficiently inexpensive that ***anyone can do it anywhere***



- Approach #1: Expect more from architectural innovation
- Approach #2: Reduce the cost to design custom hardware
- Approach #3: Embrace open-source concepts
- Approach #4: Widen the applicability of custom hardware
- Approach #5: Reduce the cost of manufacturing custom H/W

1) Expect more from architectural innovation

ISCA 2011

Proceedings of the 38th Annual

INTERNATIONAL SYMPOSIUM on COMPUTER ARCHITECTURE

Sponsored by
ACM SIGARCH
IEEE Computer Society TCCA

June 4-8, 2011
San Jose, CA, USA

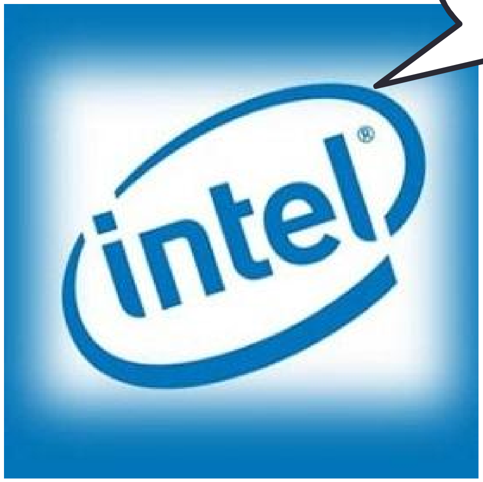
IEEE computer society

IEEE

Association for Computing Machinery

SIGARCH

"Give me 15% speedup and I'll accept your paper"



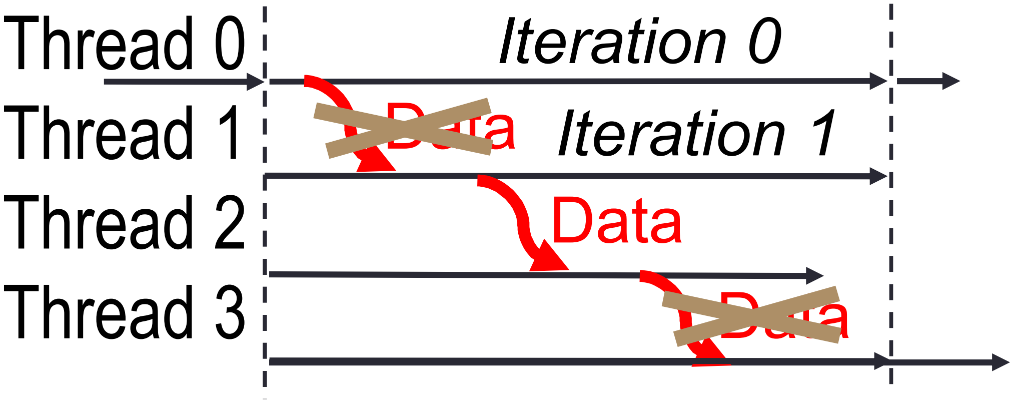
"I need 1% speedup for 1% area"

"Your idea needs to **deliver 2x or more**, or someone else should fund it"

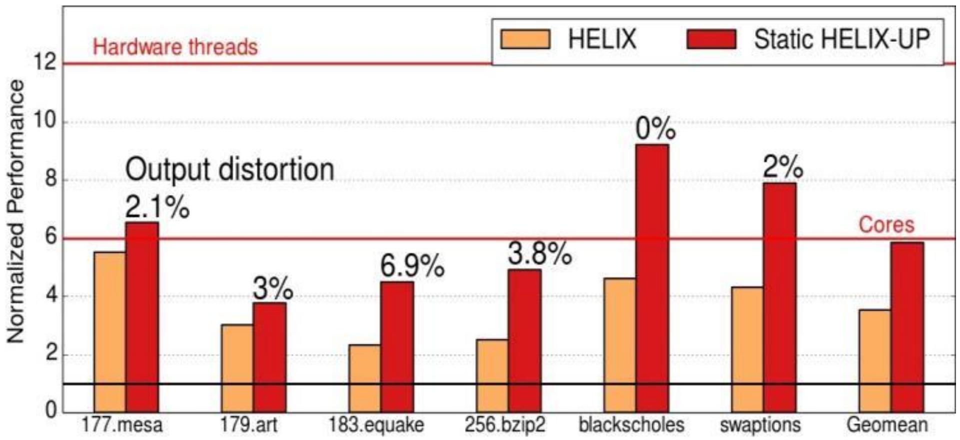
HELIX-UP Unleashed Parallelization

David Brooks @ Harvard

- Traditional parallelizing compilers must honor ***possible*** dependencies
- HELIX-UP manufactures parallelism by profiling which deps do not exist and ***which are not needed***
  - Based on user supplied ***output distortion function***
- Big step for parallelization
  - **2x speedup** over parallelizing compilers, 6x over serial, < 7% distortion



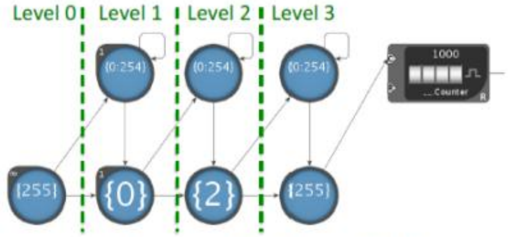
Nehalem 6 cores, 2 threads per core



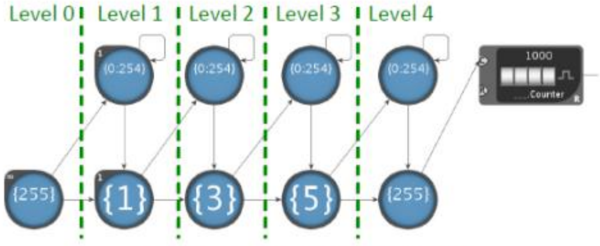
Association Rule Mining with the Automata Processor

Kevin Skadron @ UVA

- Micron’s Automata processor
  - Implements FSMs at memory
  - Massively parallel with accelerators
- Mapped data-mining ARM rules to memory-based FSMs
  - ARM algorithms identify relationships between data elements
  - Implementations are often memory bottlenecked
- Big-data sets had big speedups
  - 90x+ over single CPU performance
  - **2-9x+ speedups** over CMPs and GPUs
- Joint effort with UVA and Micron

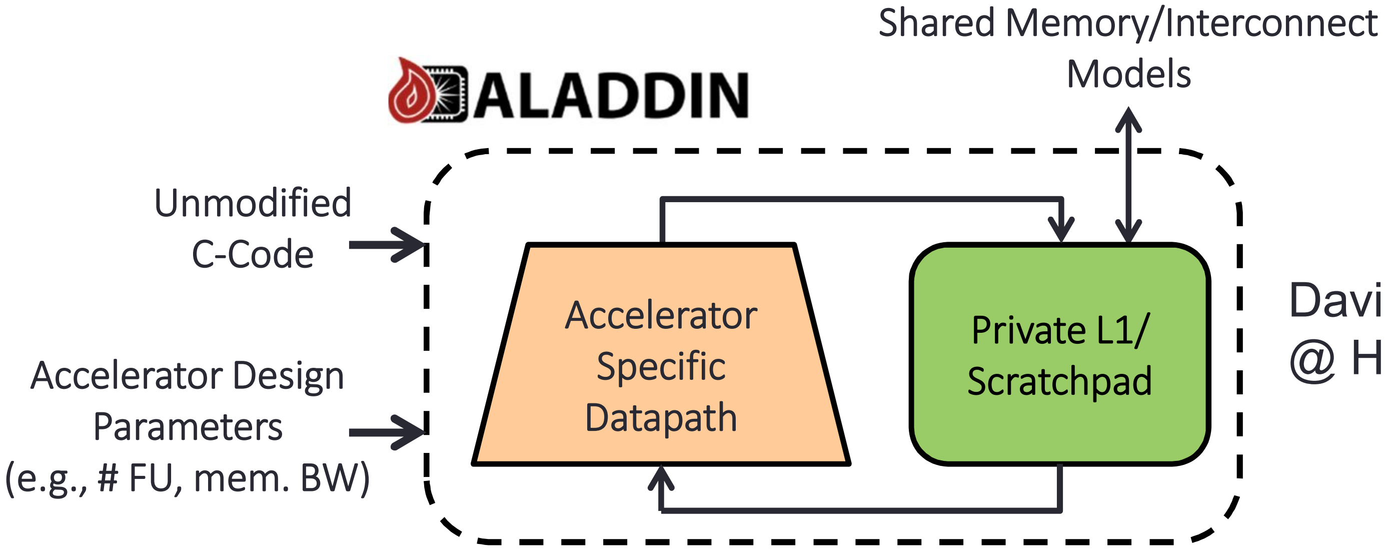


(a) Automaton for itemset {0, 2}



(b) Automaton for itemset {1, 3, 5}

2) Reduce the cost to design custom hardware


ALADDIN
Shared Memory/Interconnect Models
Unmodified C-Code
Accelerator Design Parameters (e.g., # FU, mem. BW)
Accelerator Specific Datapath
Private L1/ Scratchpad

David Brooks @ Harvard

- Better tools and infrastructure
  - Scalable accelerator synthesis and compilation, ***generate code and H/W for highly reusable accelerators***
  - Composable design space exploration, ***enables efficient exploration of highly complex design spaces***
  - Well put-together benchmark suites to drive development efforts

CortexSuite: A Synthetic Brain Benchmark Suite

Michael Taylor @ UCSD

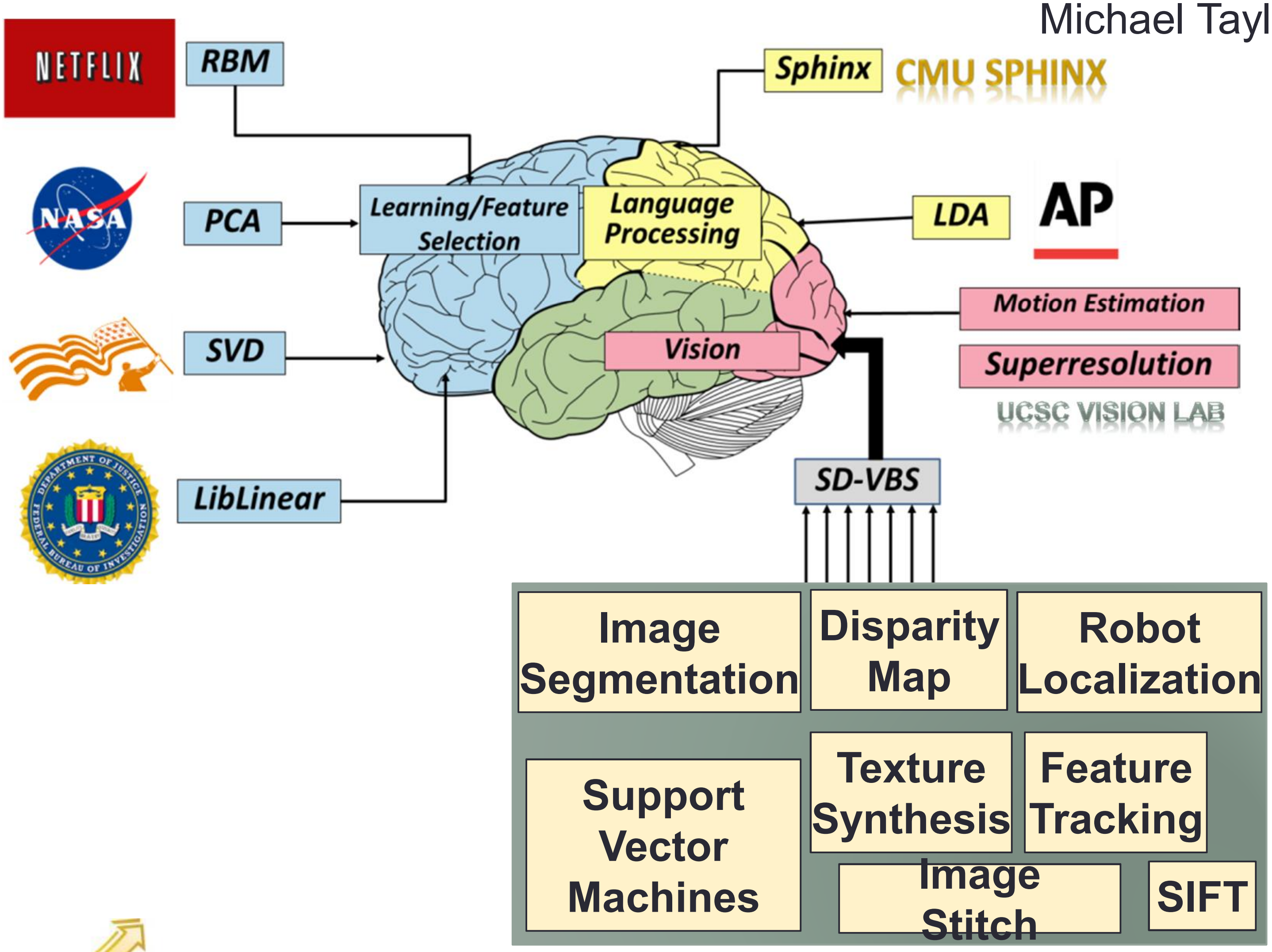


Image Segmentation

Disparity Map

Robot Localization

Support Vector Machines

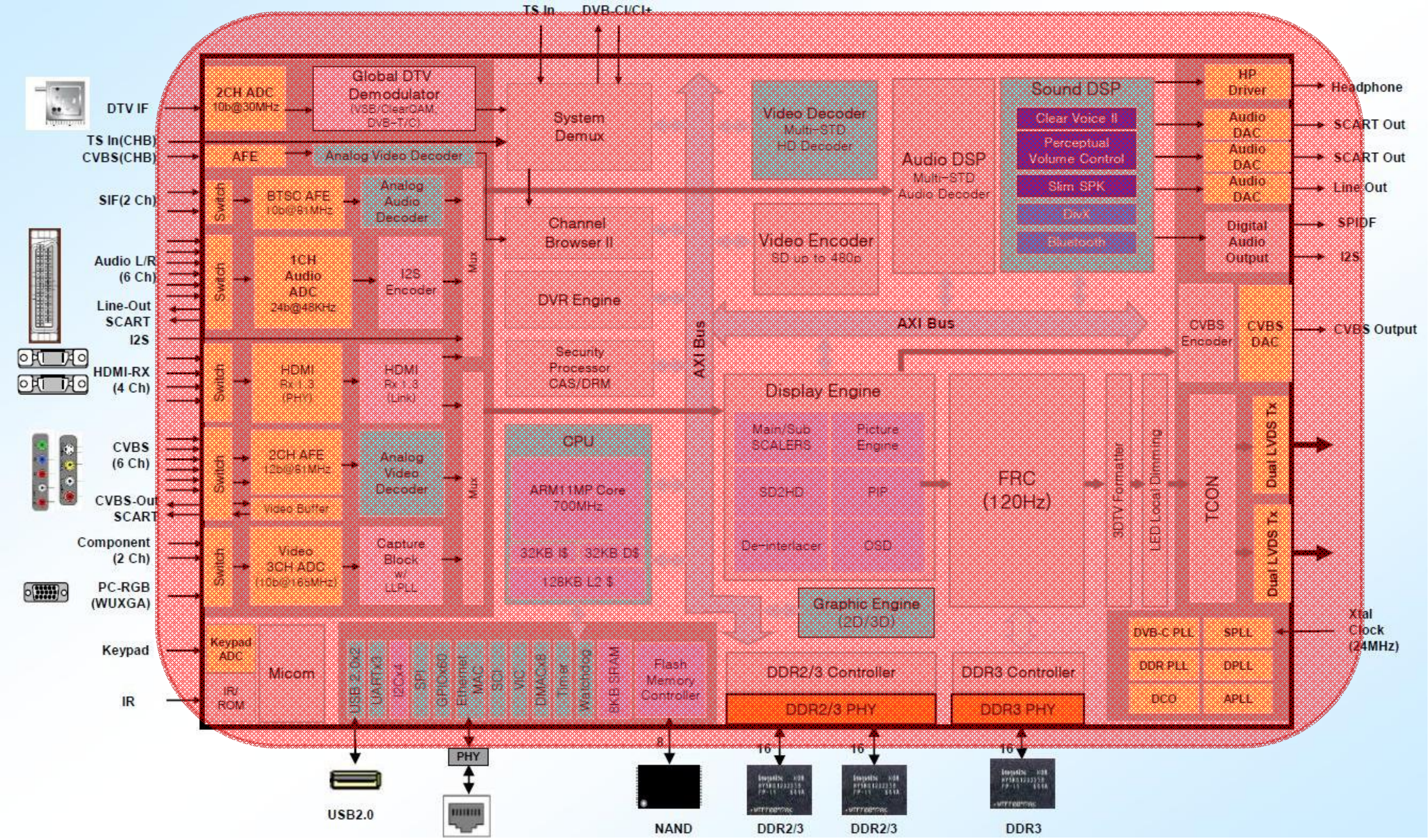
Texture Synthesis

Feature Tracking

Image Stitch

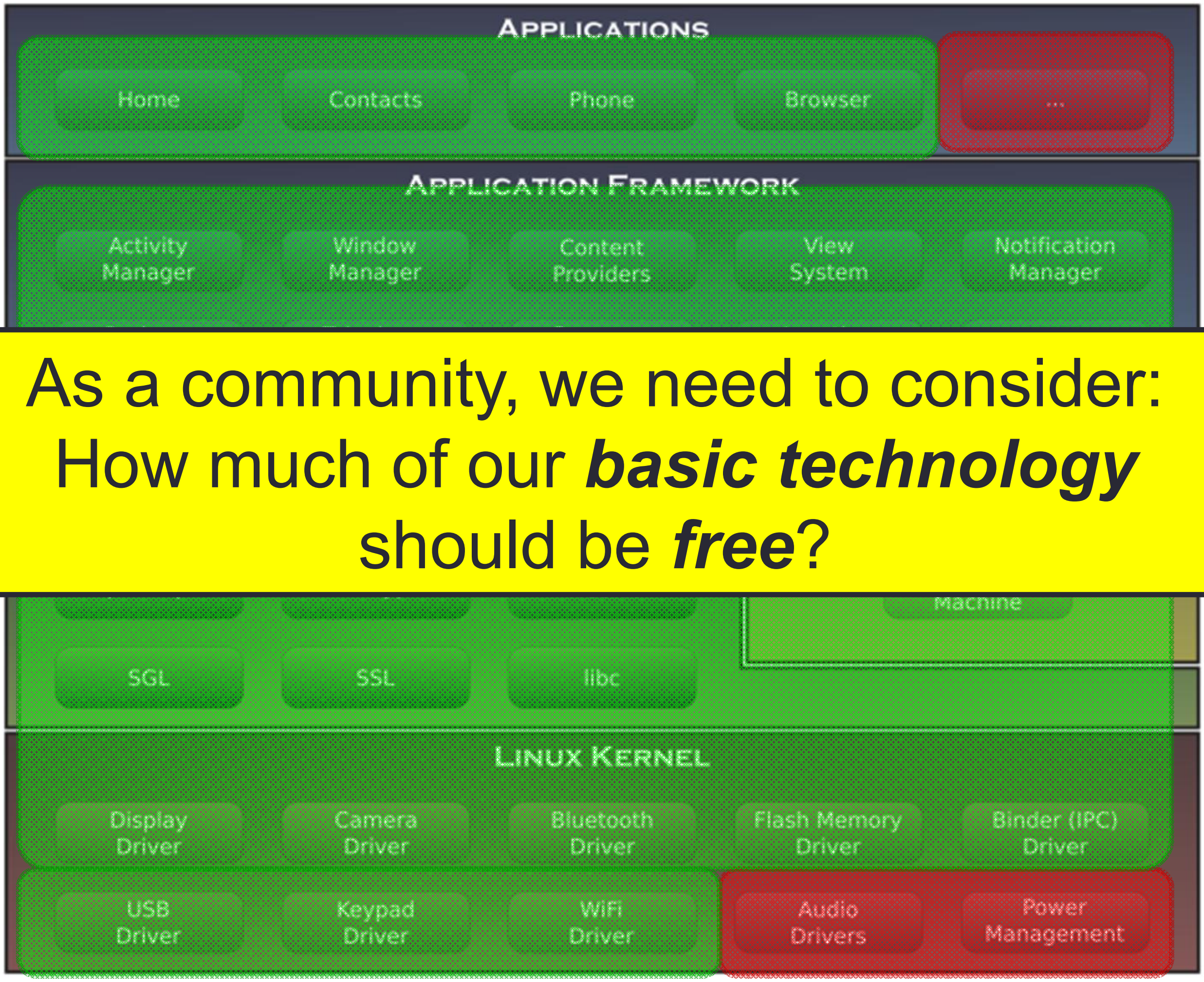
SIFT

3) Embrace Open-Source Concepts



Red = non-free IP, **Green** = free IP

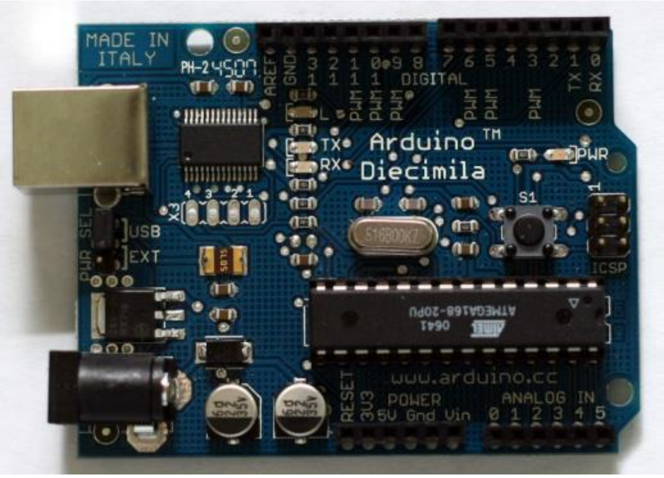
3) Embrace Open-Source Concepts



As a community, we need to consider:
How much of our ***basic technology***
should be ***free***?

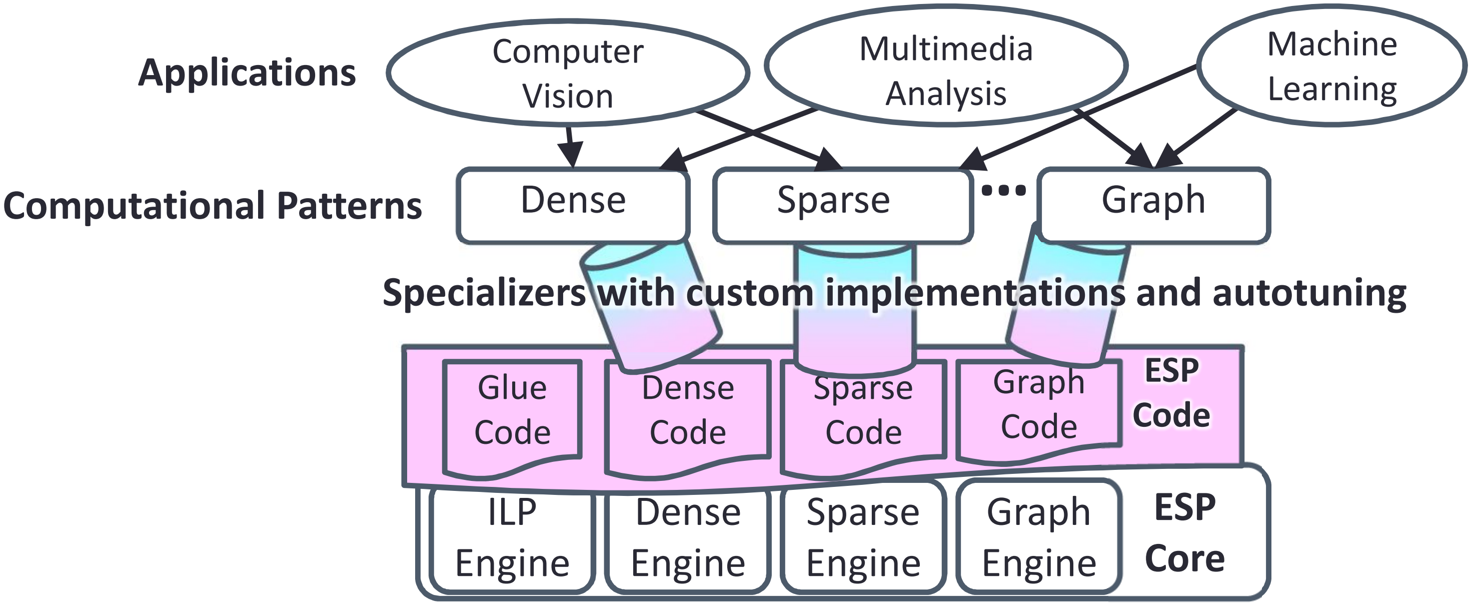
Red = non-free IP, **Green** = free IP

Open-Source H/W is Growing



4) Widen the Applicability of Customized H/W

Krste Asanovic @ UC-Berkeley

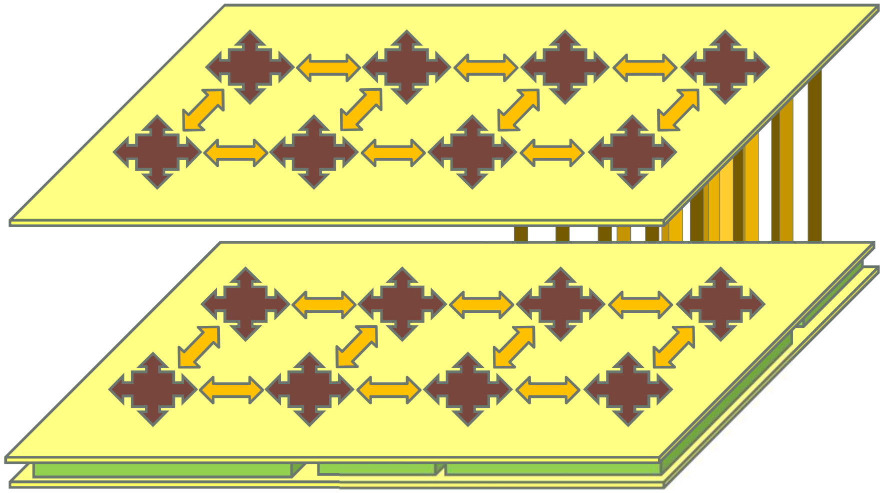


- ESP: Ensembles of Specialized Processors
  - Ensembles are algorithmic-specific processors optimized for code “patterns”
  - Approach uses ***composable customization*** to deliver speed and efficiency that is widely applicable to general purpose programs
  - Grand challenges remain: ***what are the components*** and ***how are they connected***?

5) Reduce the cost of manufacturing customized H/W

Martha Kim @ Columbia

- Another thought experiment: what if building a house were like fabricating a chip?
- ***Brick-and-mortar silicon***: ***assembly-time customization***, e.g. MCMs + 3D + FPGA interconnect



Brick-and-mortar silicon design flow:

1) Assemble brick layer
2) Connect with mortar layer
3) Package assembly
4) Deploy software

- Diversity via brick ecosystem & interconnect flexibility
- Brick design costs amortized across all designs
- Robust interconnect and custom bricks rival ASIC speeds

Conclusions

- Heterogeneous design could continue Moore's law perf. scaling via innovation alone
  - But, it requires a diverse hardware ecosystem with affordable customization

- Effective and affordable customization won't happen without our help
  1. Expect more from architectural innovation
  2. Reduce the cost to design customized design
  3. Embrace open-source concepts
  4. Widen the applicability of customization
  5. Reduce the cost of custom manufacturing

- Increasing "nanodiversity" is a good thing
  - More jobs, companies, and students
  - More competition and ***scalable innovation***

Questions

? ? ? ?

? ? ? ?

?

? ? ?