Ending the Tyranny of Amdahl's Law

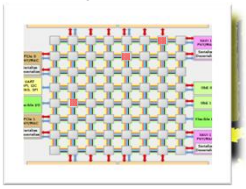
Todd Austin
University of Michigan

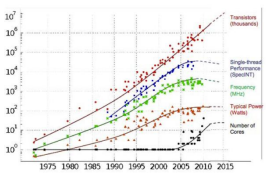
Perspectives on Scaling

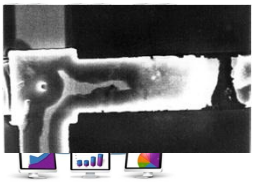
- **C-FAR**: Center for Future Architectures Research
  - Focused on scaling in 2020-2030 silicon
  - Performance, power and cost
  - 27 faculty at 14 universities, 82 students

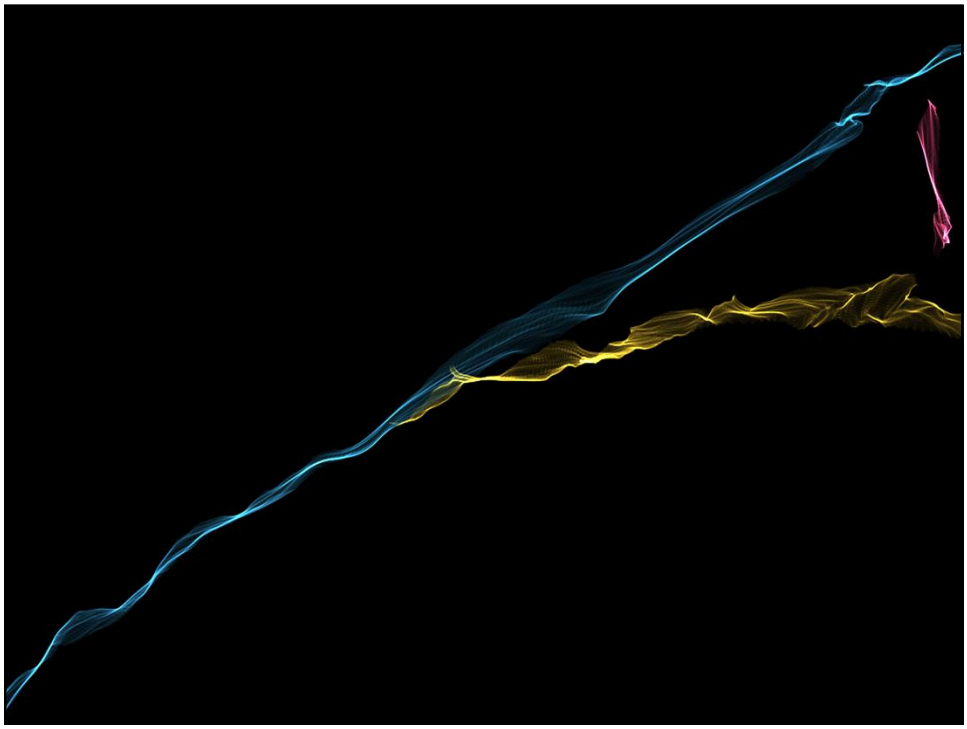
- **Why is C-FAR's mission important?**
  - The promise… tomorrow's applications need powerful systems
- **Why is C-FAR's mission challenging?**
  - The threats… slowing innovation and degrading silicon

Many-Core Vision Computer Vision End of Dennard Scaling Machine Learning Big Data Analytics Data Analytics

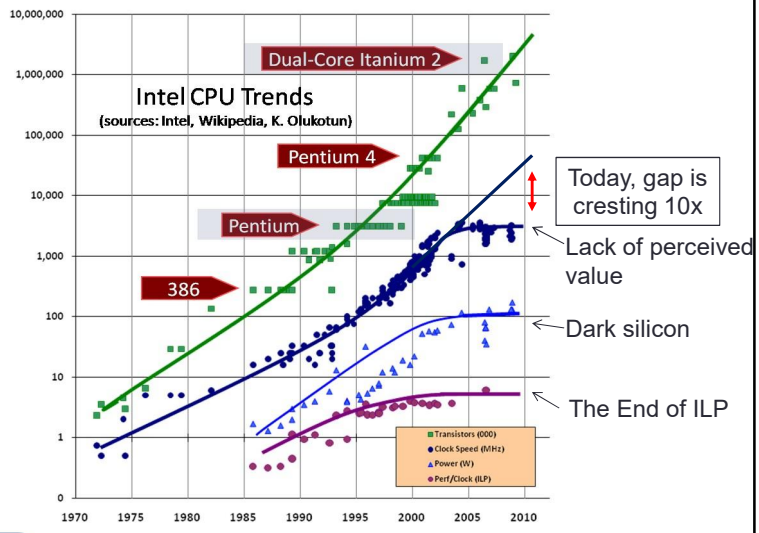






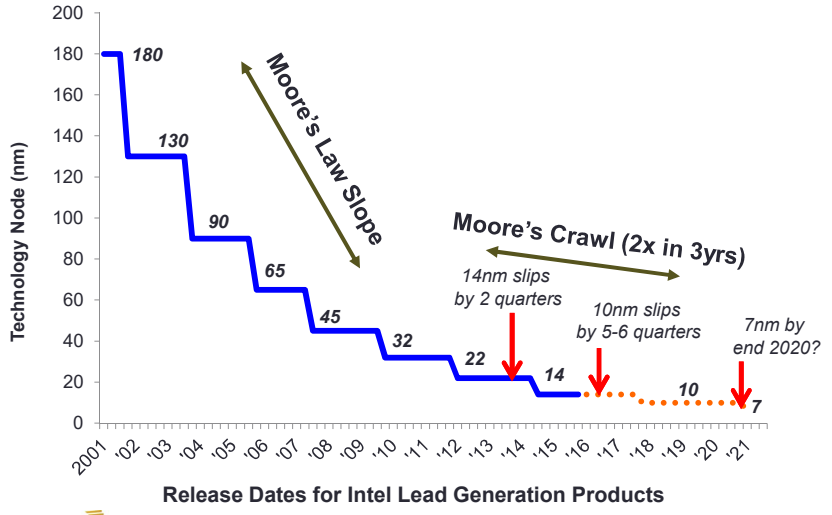


Moore's Law Performance Gap



4

Is Density Still Scaling?



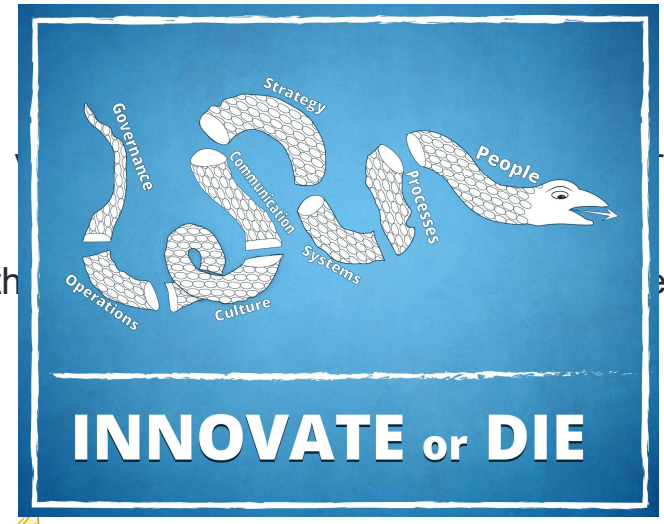
Courtesy David Brooks @ Harvard

5

What Does This All Mean to Architects?

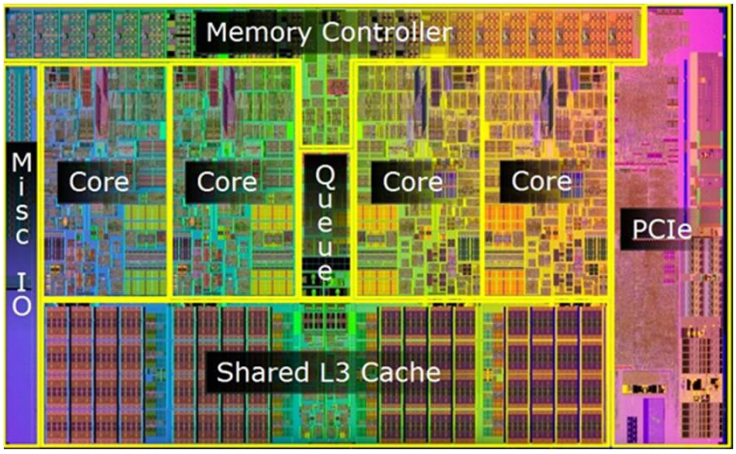
Today, v , cost).

But, th ft us.

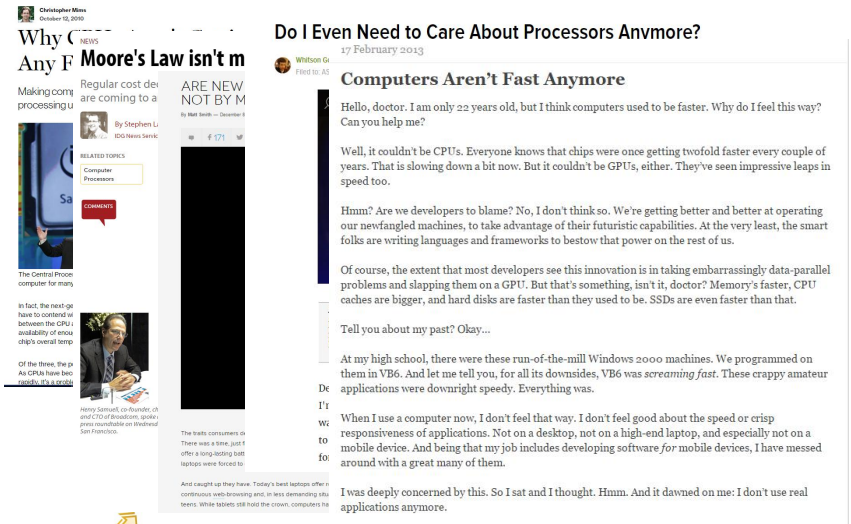


6

Attempt #1: Chip Multiprocessors



Underwhelming Results



Moore's Law isn't m

Regular cost de
are coming to a

ARE NEW
NOT BY M

Do I Even Need to Care About Processors Anvmore?

17 February 2013

Computers Aren't Fast Anymore

Hello, doctor. I am only 22 years old, but I think computers used to be faster. Why do I feel this way? Can you help me?

Well, it couldn't be CPUs. Everyone knows that chips were once getting twofold faster every couple of years. That is slowing down a bit now. But it couldn't be GPUs, either. They've seen impressive leaps in speed too.

Hmm? Are we developers to blame? No, I don't think so. We're getting better and better at operating our newfangled machines, to take advantage of their futuristic capabilities. At the very least, the smart folks are writing languages and frameworks to bestow that power on the rest of us.

Of course, the extent that most developers see this innovation is in taking embarrassingly data-parallel problems and slapping them on a GPU. But that's something, isn't it, doctor? Memory's faster, CPU caches are bigger, and hard disks are faster than they used to be. SSDs are even faster than that.

Tell you about my past? Okay...

At my high school, there were these run-of-the-mill Windows 2000 machines. We programmed on them in VB6. And let me tell you, for all its downsides, VB6 was *screaming fast*. These crappy amateur applications were downright speedy. Everything was.

When I use a computer now, I don't feel that way. I don't feel good about the speed or crisp responsiveness of applications. Not on a desktop, not on a high-end laptop, and especially not on a mobile device. And being that my job includes developing software *for* mobile devices, I have messed around with a great many of them.

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

9

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

10

We Investigate: Who's to Blame?

Educators

Programmers

?

CS Education in Ethiopia

- I have been working with Addis Ababa Institute of Technology to develop CS and IT coursework since 2009
- Special focus on building infrastructure and developing active learning
- Nearly 600 students in the CS program
- 2nd most popular major in the university
  - With many job opportunities
  - The first?

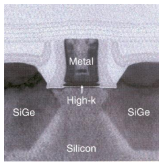
We Investigate: Who's to Blame?

Educators

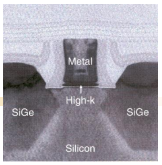
Programmers

?

The Transistor



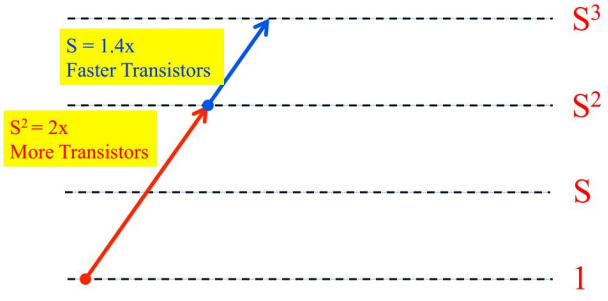
The Dark Silicon Dilemma



Advanced Scaling:

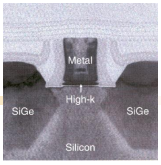
Dennard: "Computing Capabilities Scale by S^3 = 2.8x"

If S=1.4x …

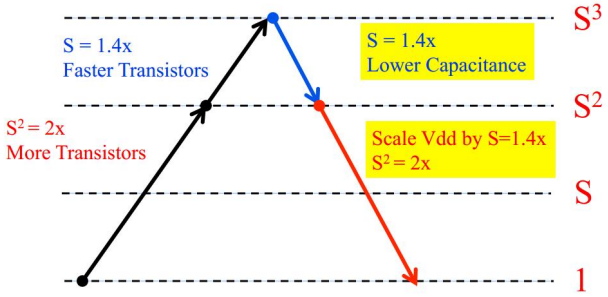


Courtesy Michael Taylor @ UCSD

The Dark Silicon Dilemma

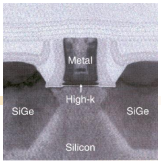


Dennard:
"We can keep power consumption constant"

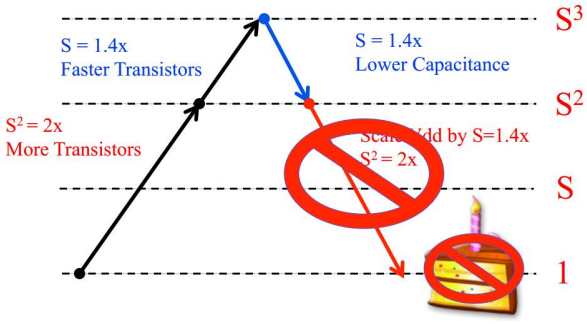


Courtesy Michael Taylor @ UCSD

The Dark Silicon Dilemma



Fast forward to 2005:
Threshold Scaling Problems due to Leakage Prevents Us From Scaling Voltage

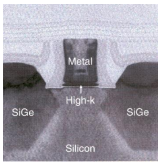


Courtesy Michael Taylor @ UCSD

We Investigate: Who's to Blame?

Educators

The Transistor



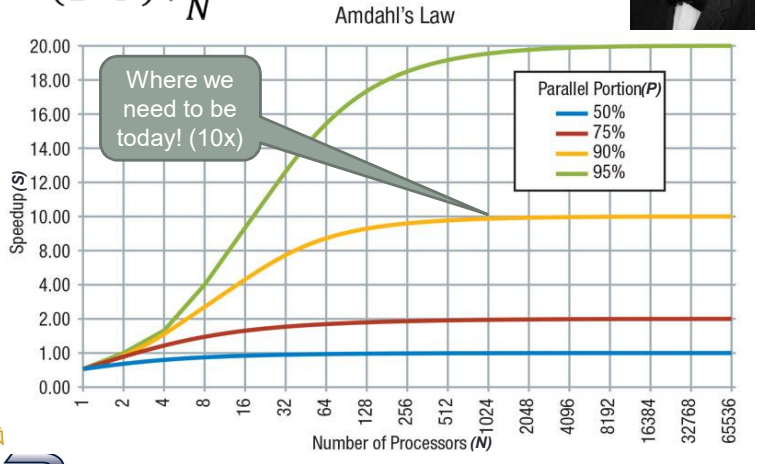
?

Programmers

Architects

The Tyranny of Amdahl's Law

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

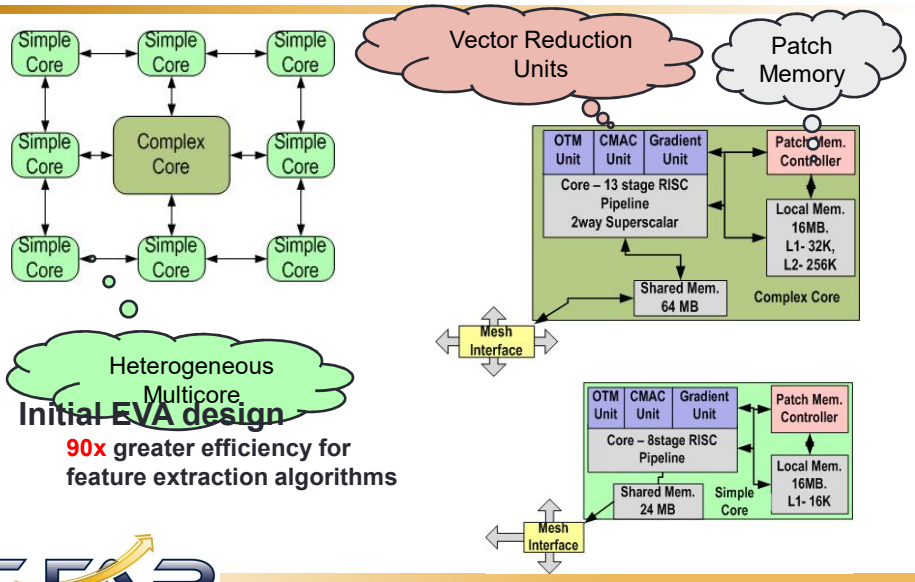


A Story about Jason and His Two Dads



19

EVA Embedded Vision Architecture

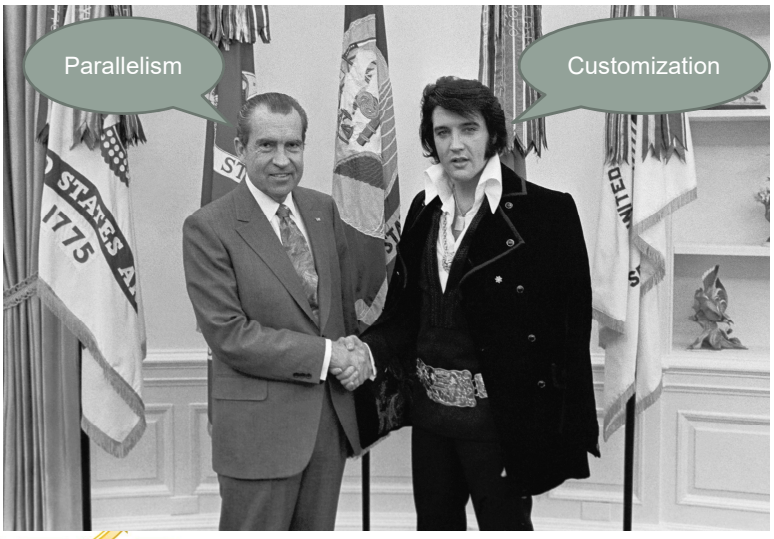


Initial EVA design

90x greater efficiency for feature extraction algorithms

20

Where We Need to Focus: Heterogeneous Parallel Systems



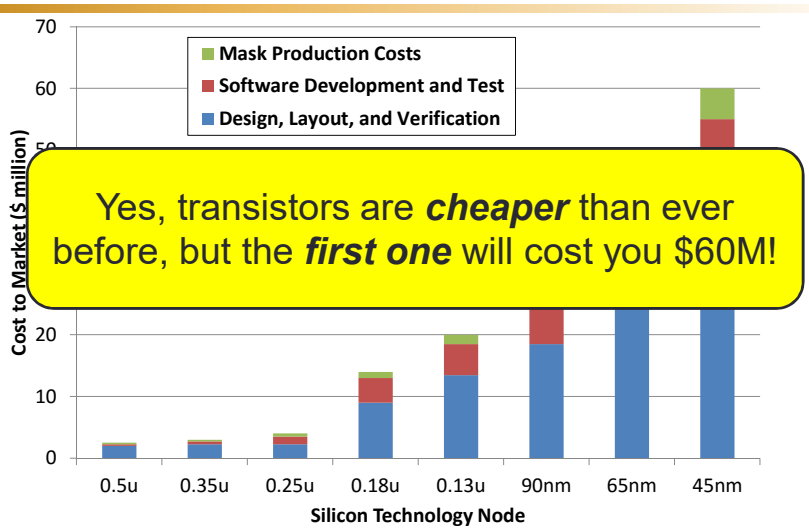
Silicon Today: The Good, the Bad and the Ugly

- ***The Good***: Heterogeneous parallel systems have the potential to close the Moore's Law performance gap
  - It's an old idea – it really works…
- ***The Bad***: Dennard scaling has all but stopped, Moore's Law is losing steam fast, leaving a growing performance/power scaling gap
  - All trends are bad…
- ***The Ugly***: The heterogeneous parallel designs needed to close the gap will be ***too expensive to afford***
  - Skyrocketing NREs will necessitate broadly applicable (vanilla and slow) H/W designs

What I Want You to Remember

- Successfully bridging the Moore's Law performance gap is less about "***How***" to do it, but more about "***How Much***" does it cost!
- ***My claim:*** *if we can effect a 100x reduction in the cost of bringing a design to market, scaling challenges will eventually solve themselves as the market flourishes with orders of magnitude more designs, some of which will be the big wins of tomorrow.*

Design Costs Are Skyrocketing

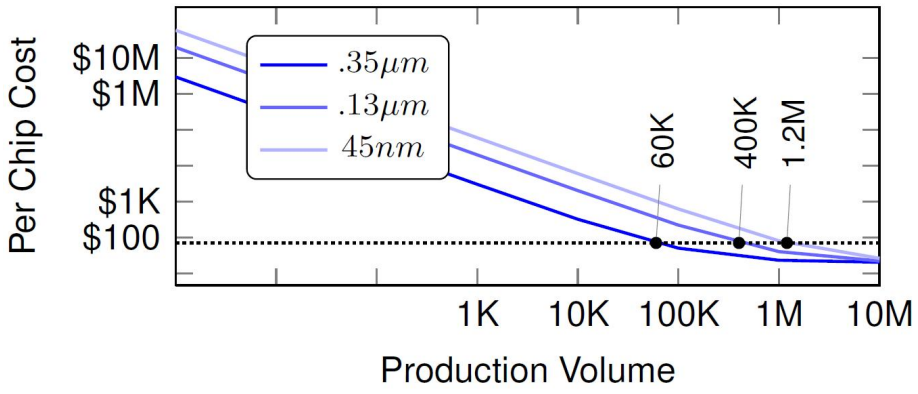


Yes, transistors are ***cheaper*** than ever before, but the ***first one*** will cost you $60M!

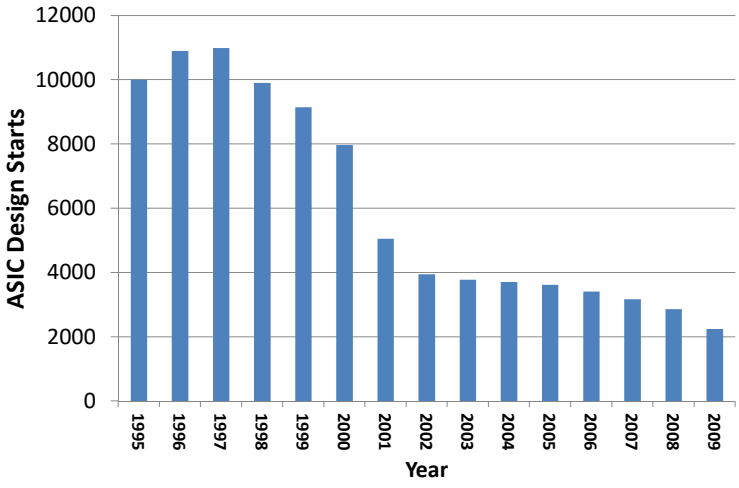
Source: International Business Strategies

High Costs Kill Customization

- Heterogeneous designs serve smaller markets



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 via Lower Design Cost

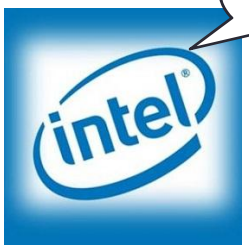
- Ultimate goal: make customized design sufficiently inexpensive that ***anyone can do it anywhere***
  - Address all NRE factors: market size, design costs, build costs
  - Take inspiration from Web 2.0, and subsequent innovation explosion

- Approach #1: Raise your expectations for scaling innovation
  - Abandon former metrics for those that can start closing the gap
- Approach #2: Reduce the cost to design custom hardware
  - With ***better tools*** that understand and leverage the benefits of customization
  - By embracing ***open-source hardware*** design solutions
- Approach #3: Widen the applicability of custom hardware
  - Increasing market applicability with ***composable customization*** mitigates potentially higher NREs
- Approach #4: Reduce the cost of manufacturing hardware
  - Utilize ***assembly-time customization*** to slash the cost of customization

1) Raise your expectations for scaling innovation

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

"I need 1% speedup for 1% area"

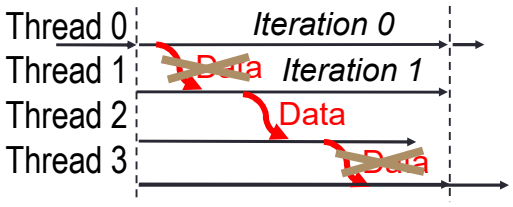


"Your idea needs to **deliver 2x or more**, or someone else should fund it"

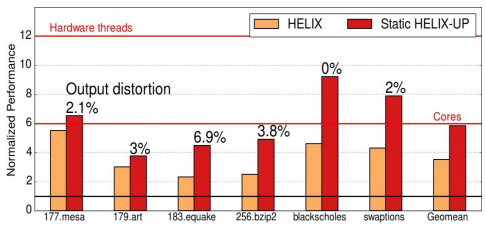
HELIX-UP Unleashed Parallelization

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



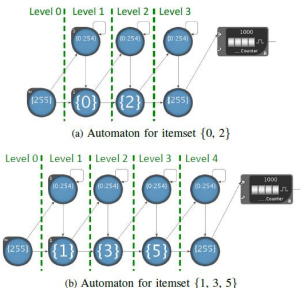
Nehalem 6 cores, 2 threads per core



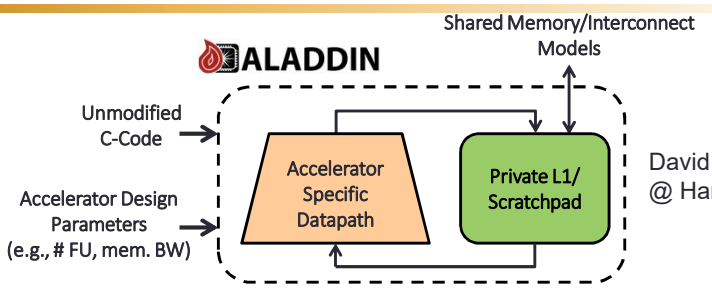
Association Rule Mining with the Automata Processor

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



2) Reduce the cost to design custom hardware

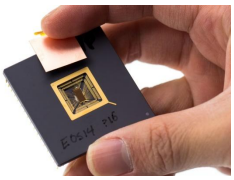


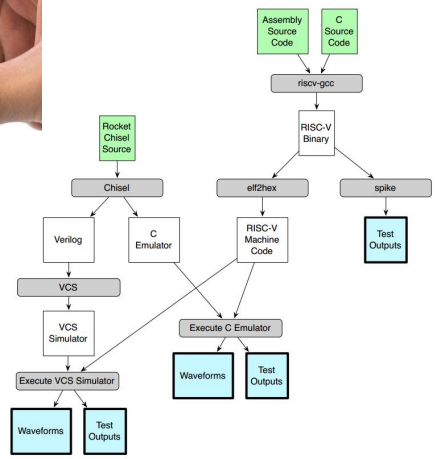
David Brooks @ Harvard

- Better tools and infrastructure
  - Scalable accelerator synthesis and compilation, ***generate code and H/W for highly reusable accelerators***
  - Composable design space exploration, e***nables efficient exploration of highly complex design spaces***
  - Well put-together benchmark suites to drive development efforts
- Embrace open-source concepts
  - Example: Berkeley's RISC-V architecture

Berkeley's RISC V Open-Source ISA

Krste Asanovic @ UC-Berkeley

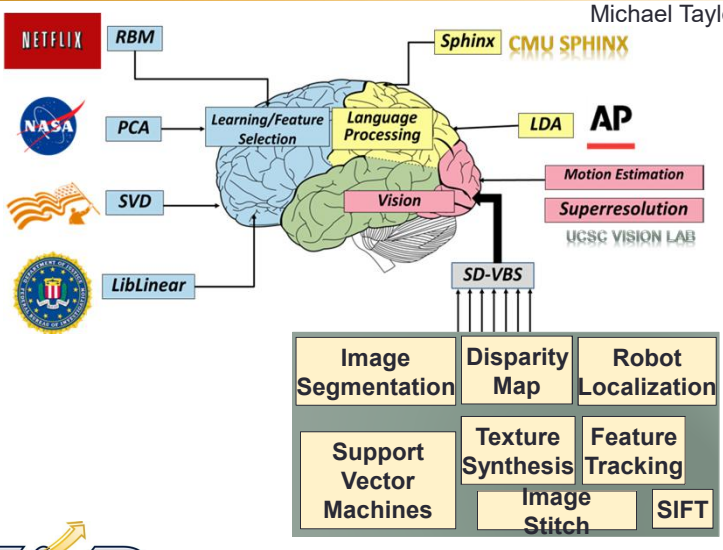




33

CortexSuite: A Synthetic Brain Benchmark Suite

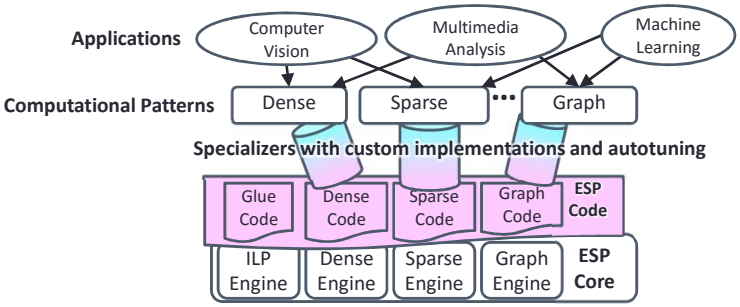
Michael Taylor @ UCSD



34

3) Widen the Applicability of Customized H/W

Krste Asanovic @ UC-Berkeley

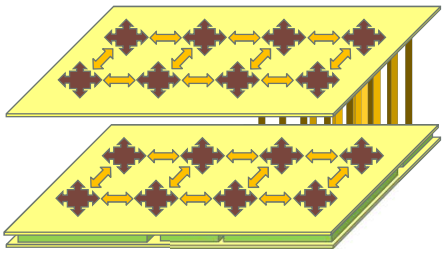


- ESP: Ensembles of Specialized Processors
  - Ensembles are algorithmic-specific processors optimized for code "patterns"
  - Patterns capture common operations across many applications, each with unique communication and computation structure
  - ***Approach has the promise of custom accelerator speed and efficiency that is widely applicable to general purpose programs***

4) Reduce the cost of manufacturing customized H/W

Martha Kim @ Columbia

- ***Brick-and-mortar silicon*** explores ***assembly-time customization***, i.e., 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

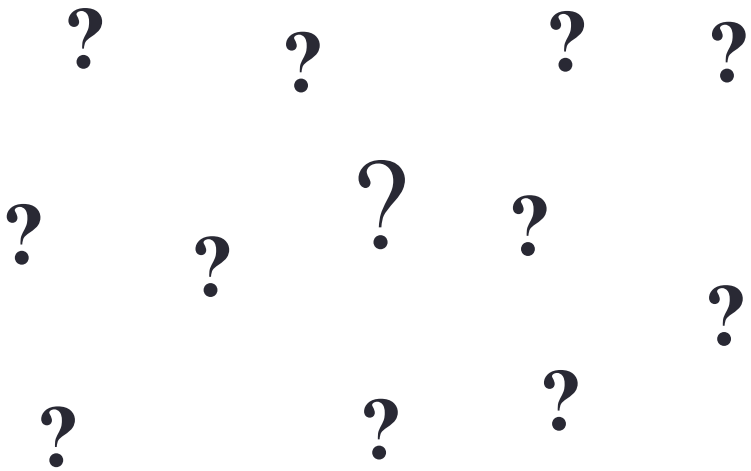
- Effective and affordable customization won't happen without our help
  1. Raise your expectations for scaling innovation
  2. Reduce the cost to design customized design
  3. Widen the applicability of customization
  4. Reduce the cost of custom manufacturing

- Increasing "nanodiversity" is a good thing
  - Better perf., power, cost, capability
  - More jobs, companies, and students
  - More competition and ***scalable innovation***

CFAR

37

Questions



CFAR