ASPIRE

Algorithms and Specializers for Provably Optimal Implementations with Resiliency and Efficiency

Krste Asanović, Elad Alon,
Jonathan Bachrach, Jim Demmel, Armando Fox,
Kurt Keutzer, Borivoje Nikolić, David Patterson,
Koushik Sen, Vladimir Stojanović, John Wawrzynek

krste@berkeley.edu

http://aspire.eecs.berkeley.edu

ASPIRE End of Project Party
UC Berkeley
December 5, 2017

It all began back at end of 2011…

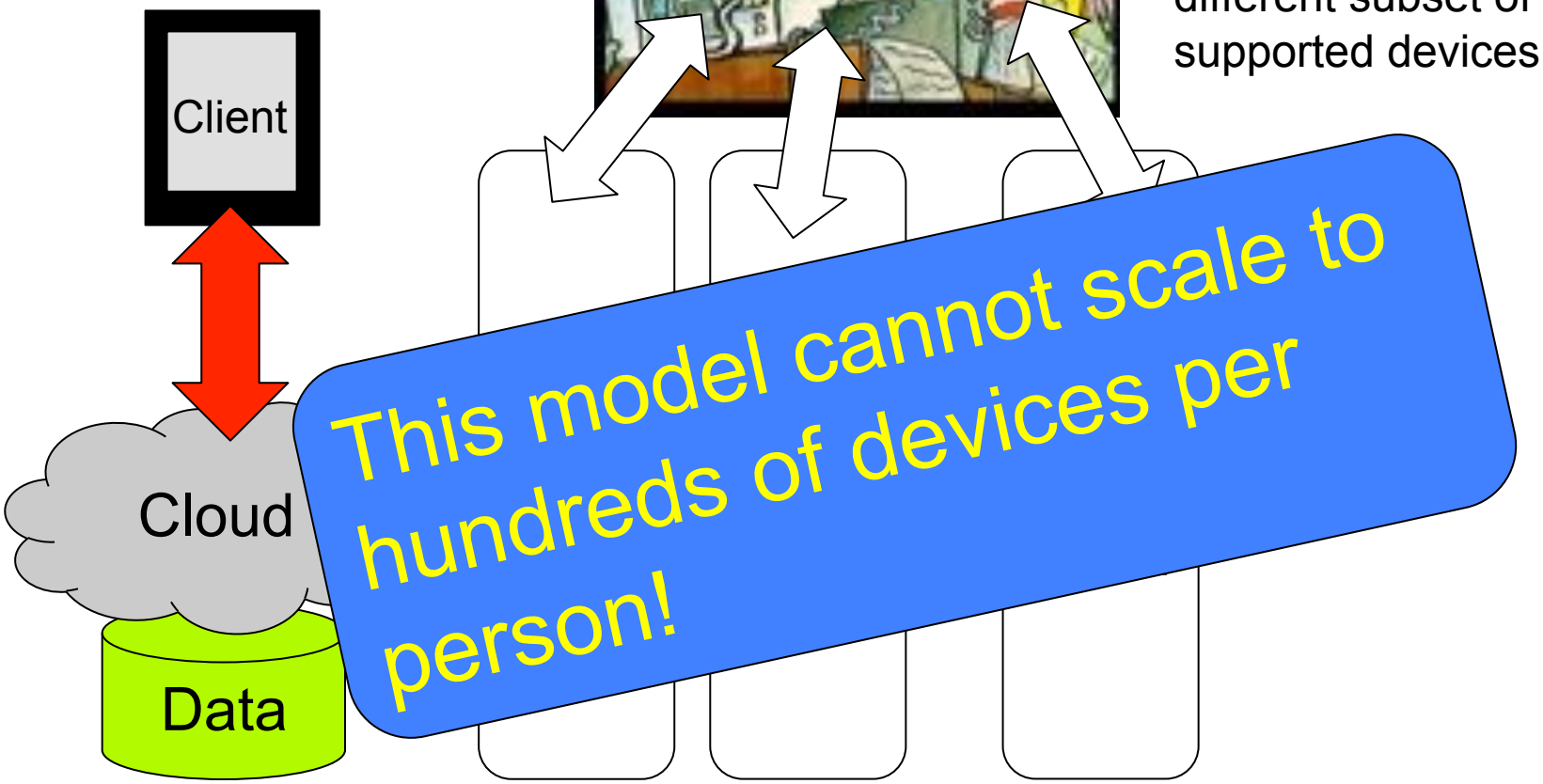
- We were searching for what would come after Par Lab (2008-2013)

Today's software paradigm: "Apps"

What's the problem: Functionality Encased in App Silos

Each task involves the manual invocation of a manually scheduled series of applications

Each application has a different: user interface, account to register, password, software updates, payment model, different subset of supported devices…



This model cannot scale to hundreds of devices per person!

What we want to do in future: Interact with Intelligent Personal Assistant

Like Ms. Potts in *Iron Man*

Or like Jeeves in P.G. Wodehouse's *Aunts Aren't Gentlemen*

- Have a dialogue with assistant who you can give high level goals vs. simple steps, who remembers past tasks and can find information for you

Post-ParLab Vision

- Goal: “Computing continuum”: users’ perception is always connected, always available
- From mobile client → sensors + mobile client + cloud
- From GUI → Virtual Assistant via Natural User Interfaces
- From a giant menu of applications → set of integrated services
- From local-client scheduling → intelligent Client+Cloud partitioning
- Enable productive development of an integrated set of services that meet end-user requirements of speed, privacy, and reliability through intelligent use of client and cloud computing.

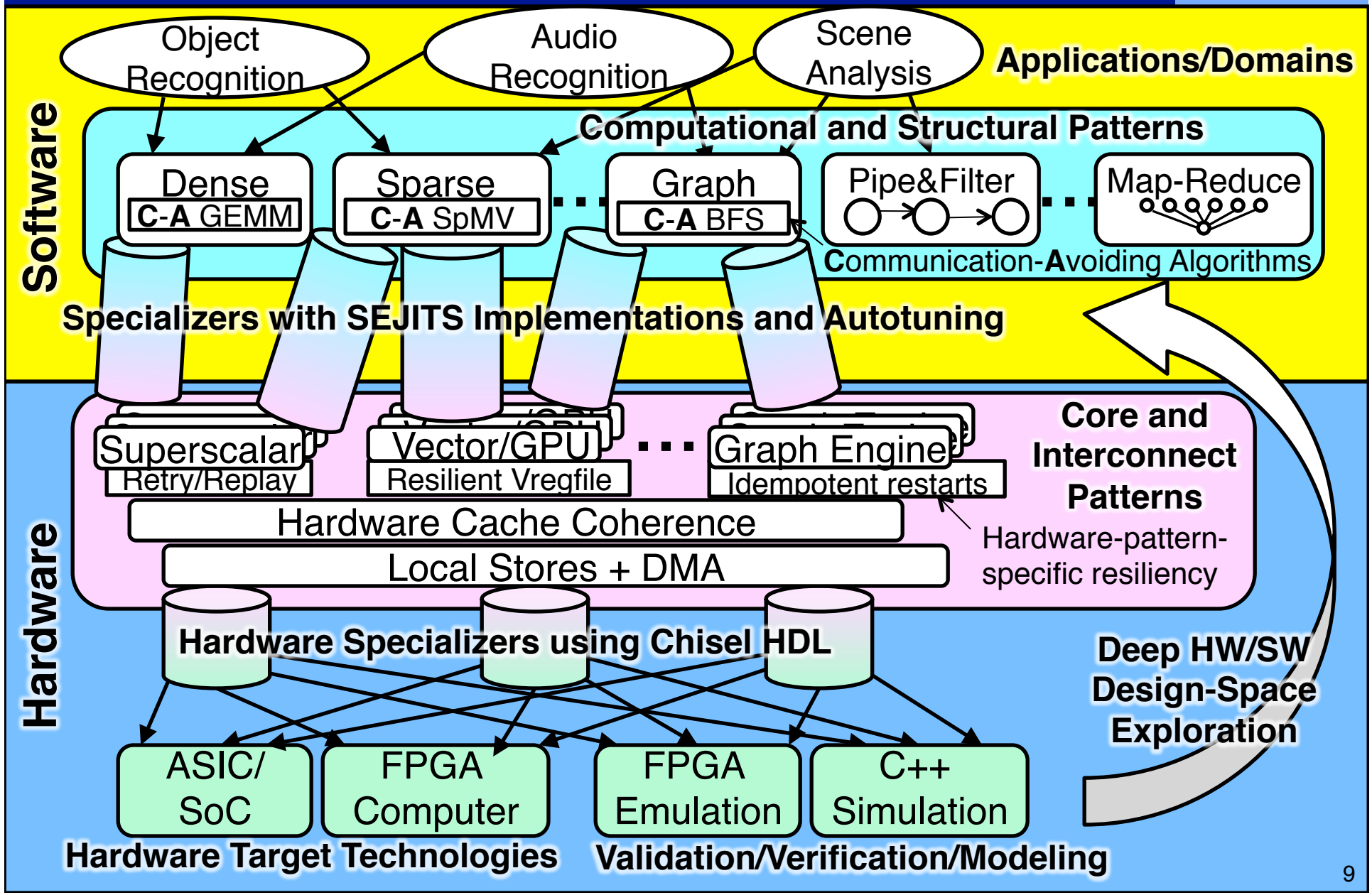
Sponsor's Feedback:

"Work on Energy-Efficient Computing"

Upheaval in Computer Design

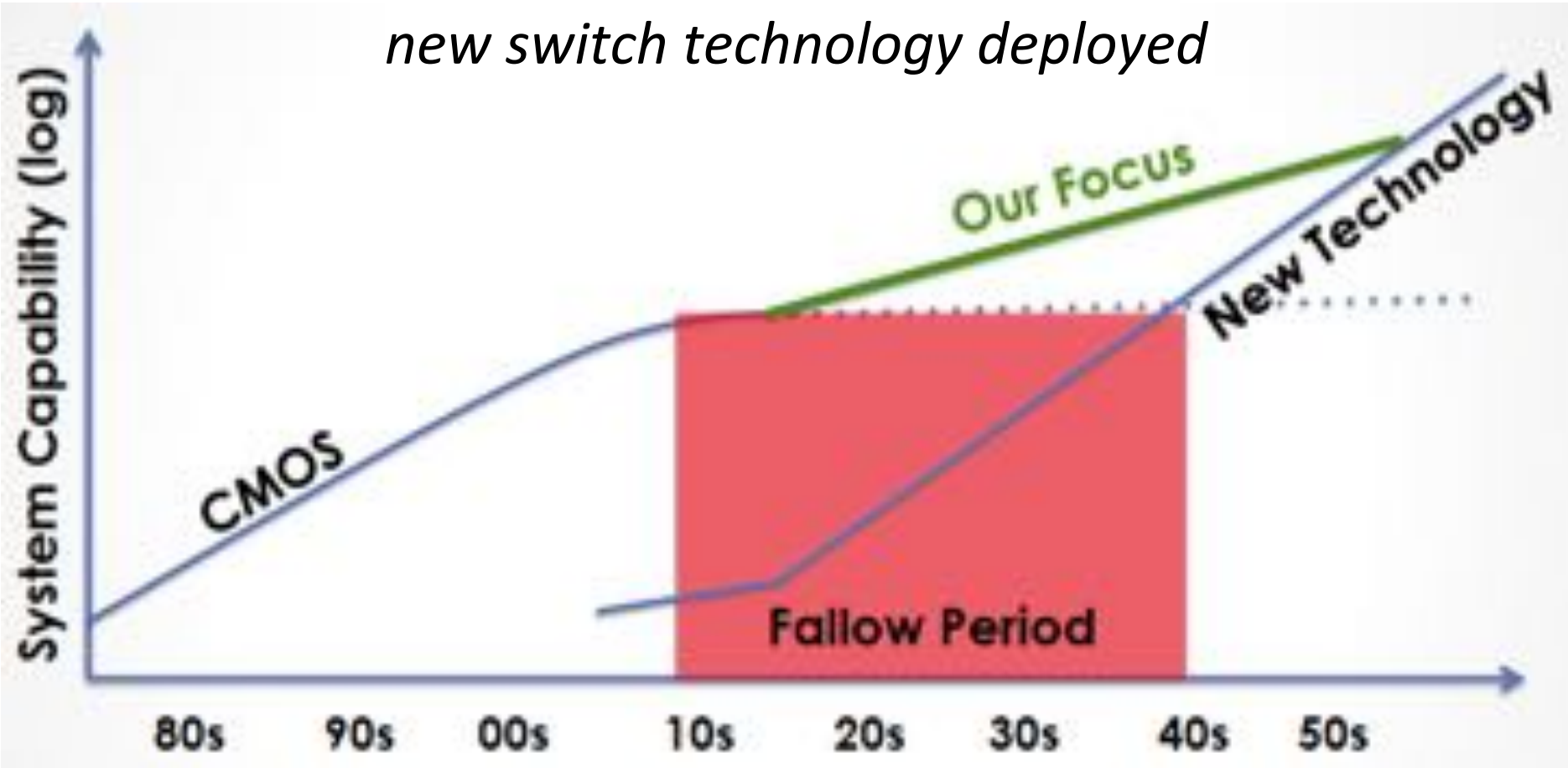
- Most of last 50 years, Moore's Law ruled
- Last decade, technology scaling slowed
  - Dennard scaling over (supply voltage ~fixed)
  - Moore's Law (cost/transistor) over?
  - No competitive replacement for CMOS anytime soon
- Users' imagination unlimited, creating new useful applications with endlessly growing compute demands
- Energy efficiency constrains everything
  - Mobile
  - High-performance embedded computing
  - Warehouse-scale computing
- Parallel computing going mainstream, but only one-time improvement in energy efficiency
- What next?

UCB HW/SW Specializer Stack



ASPIRE Goal

Keep computers' performance and energy efficiency improving past end of CMOS transistor scaling until new switch technology deployed

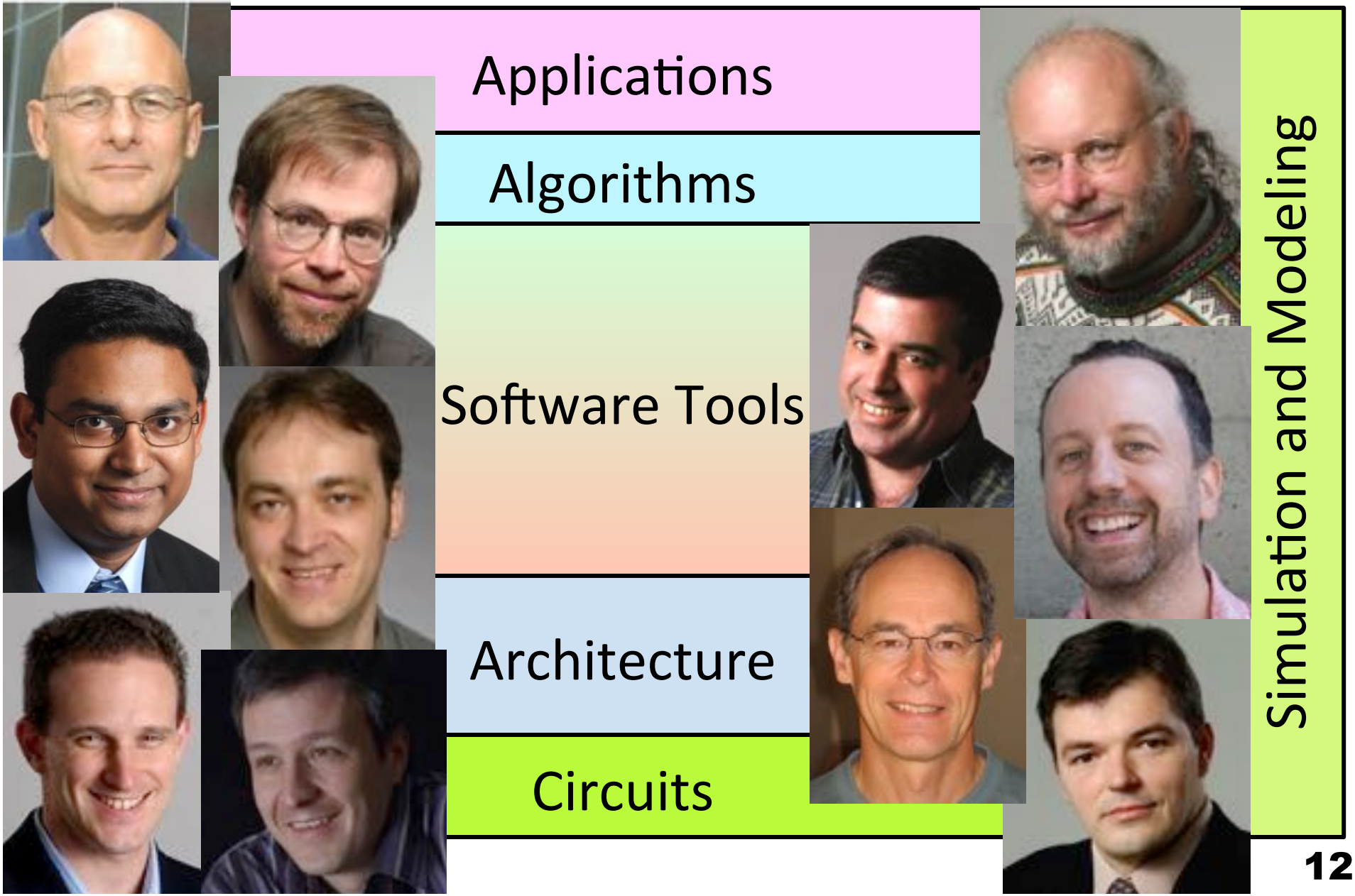


[Graph from "Advancing Computers without Technology Progress", Hill, Kozyrakis, et al., DARPA ISAT 2012]

ASPIRE Explained

- ***A***lgorithms with ***P***rovably minimal data movement
- ***S***pecialization of software and hardware based on recurring patterns of computation and communication
- Autotuning & design-space exploration of software & hardware ***I***mplementations for best ***E***fficiency
- ***R***esiliency from holistic whole-stack design

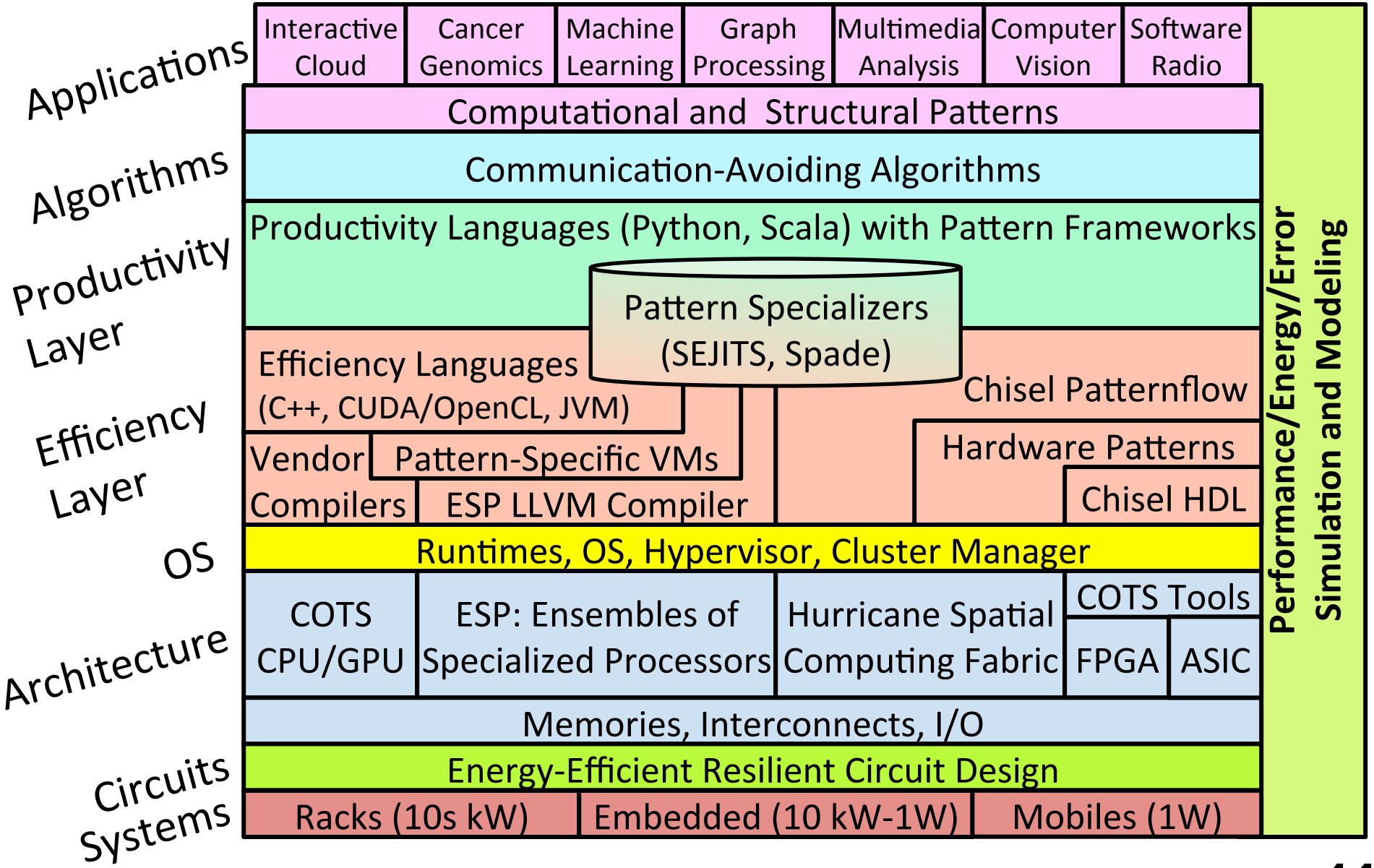
Algorithms and Specializers for Provably Optimal Implementations with Resiliency and Efficiency



ASPIRE System Targets

Systems	Racks (10s kW)	Embedded (10 kW-1W)	Mobiles (1W)

ASPIRE Lasagna



Driving Applications

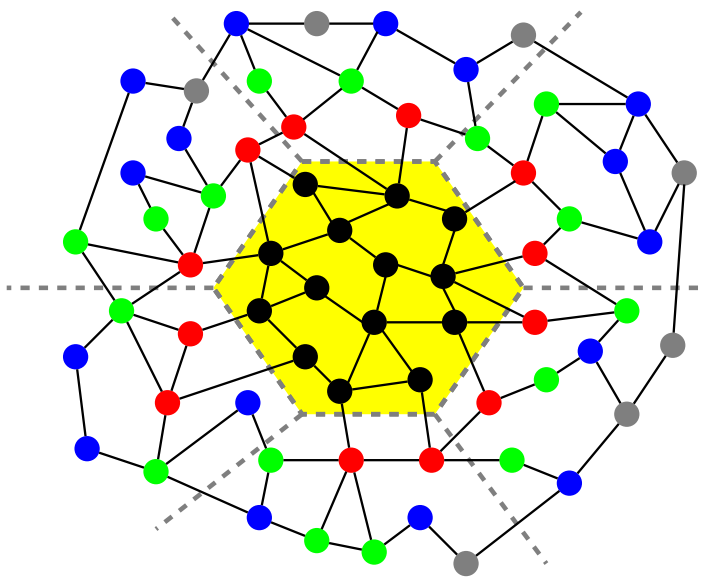
Applications

Interactive Cloud	Cancer Genomics	Machine Learning	Graph Processing	Multimedia Analysis	Computer Vision	Software Radio

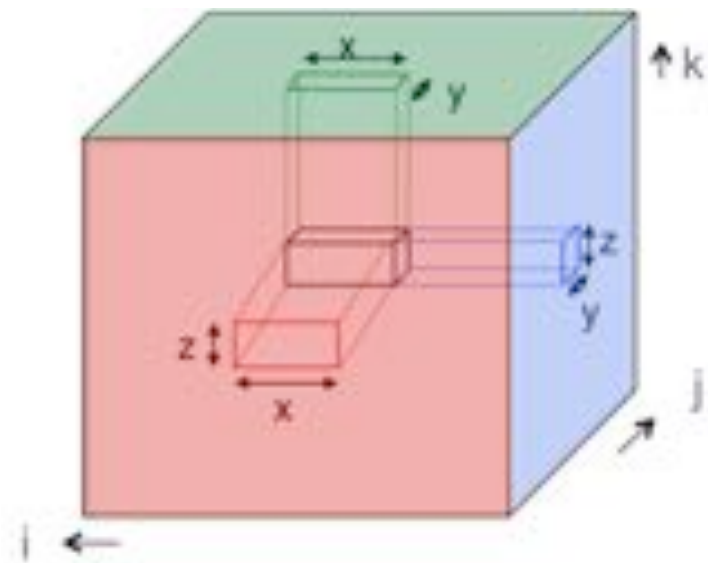
Squeezdet exceeded our DARPA Phase-III goals from 2012 proposal: 57 fps at 1.4 J/frame (>2000 improvement in efficiency)

Communication-Avoiding Algorithms

Applications	Interactive Cloud	Cancer Genomics	Machine Learning	Graph Processing	Multimedia Analysis	Computer Vision	Software Radio
	Computational and Structural Patterns						
Algorithms	**Communication-Avoiding Algorithms**						



Idea #1: read a piece of a sparse matrix (= graph) into fast memory and take



Idea #2: replicate data (including left-hand side arrays, as in C in C=A*B) and compute partial results, reduce later

Seemingly endless "Best Paper" and other awards for this work, due to very real speedups on well-studied kernels

Computing with Specialized Hardware?

Well known, for some applications, custom hardware 100-1000x performance and energy efficiency over general-purpose processors, but how to deploy?

- **Custom chip per application (ASIC)?**
  - NRE costs exploding, ~$50-100M for cutting-edge ASIC
  - Development time too long, months or years, high risk
  - Number of ASIC starts dropping over time
- **Custom "Chiplets"? (aka Lego, Bricks&Mortar, MoChi,…)**
  - i.e., small custom chips connect to standard parts as System-in-Package
  - No industry standards, packaging cost, yield
  - Data movement energy/performance costs
- **Standard programmable logic parts, FPGAs?**
  - Lower NRE and risk
  - Historically, poor energy efficiency and terrible programming productivity
- **Standard programmable Systems-on-Chip (SoC)?**
  - Common for both server and mobile platforms today
  - Specialized hardware as heterogeneous accelerators

Evaluating Hardware Choices

Applications

Interactive Cloud	Cancer Genomics	Machine Learning	Graph Processing	Multimedia Analysis	Computer Vision	Software Radio

What to target?

What are productivity/cost/performance tradeoffs?

Architecture

CPU	GPU	FPGA	ASIC
Memories, Interconnects, I/O			

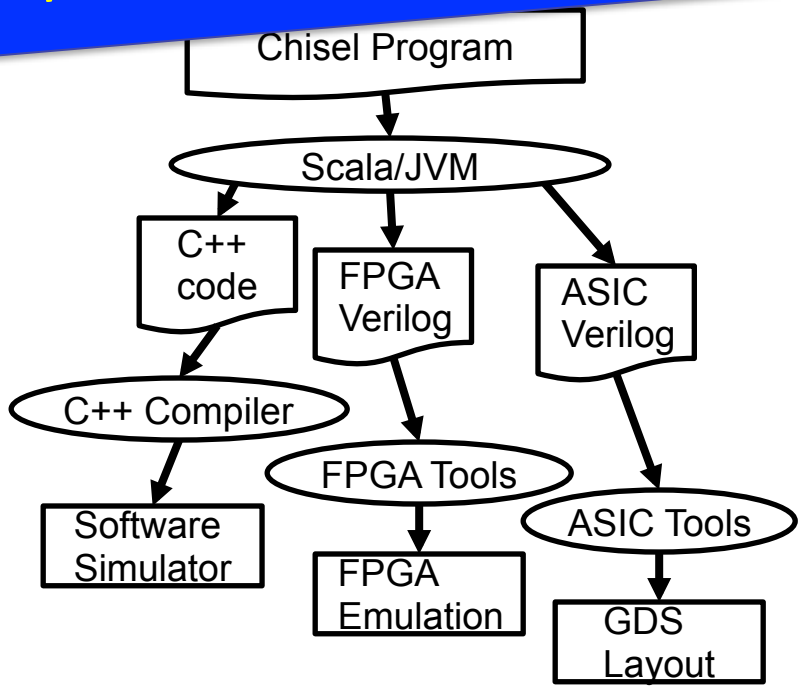
Systems

Racks (10s kW)	Embedded (10 kW-1W)	Mobiles (1W)

Chisel: Constructing Hardware In a Scala Embedded Language

- Embed hardware-description language in Scala, using Scala's extension facilities: Hardware module is just data structure in Scala
- Different output routines generate different types of output (C, FPGA-Verilog, ASIC-Verilog) from same hardware representation
- Full power of Scala for writing hardware

Chisel 3.0/ FIRRTL 1.0 released. External industry + academia uptake accelerating.



RISC-V ISA

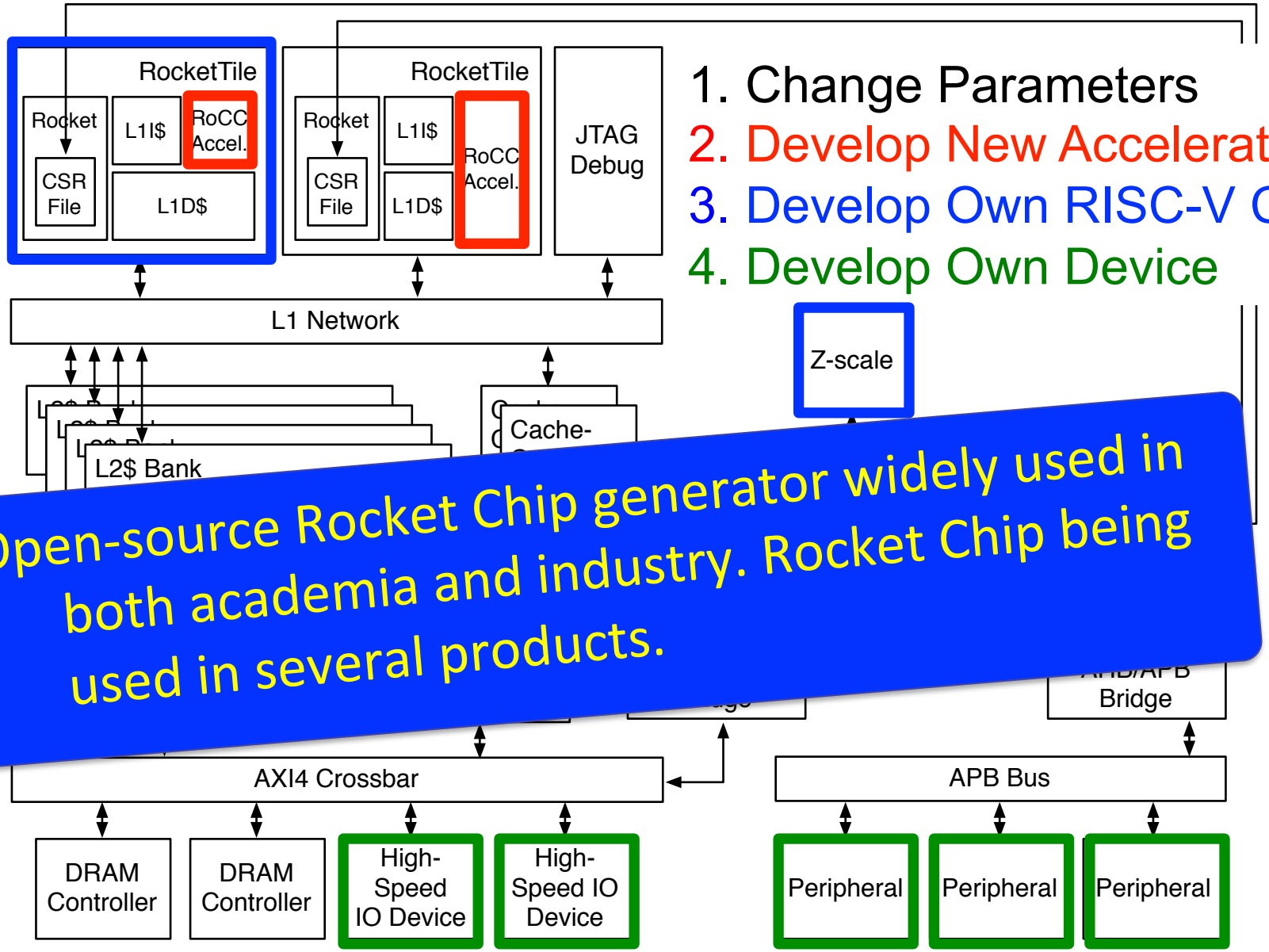
www.riscv.org

- A new completely free and open ISA
  - Already GCC, Linux, glibc, LLVM, upstreamed!
  - RV32, RV64, and RV128 variants for 32b, 64b, and 128b address spaces defined
- Base ISA only <50 integer instr

RISC-V experiencing strong momentum as legitimate alternative to other proprietary ISAs.
70+ companies are members of RISC-V Foundation
Nvidia & Western Digital committed to using RISC-V.
Linley group "Best Technology in 2016".
Becoming the standard ISA for education.

implemented at Berkeley so far (45nm, 28nm, 16nm)

Rocket Chip Generator

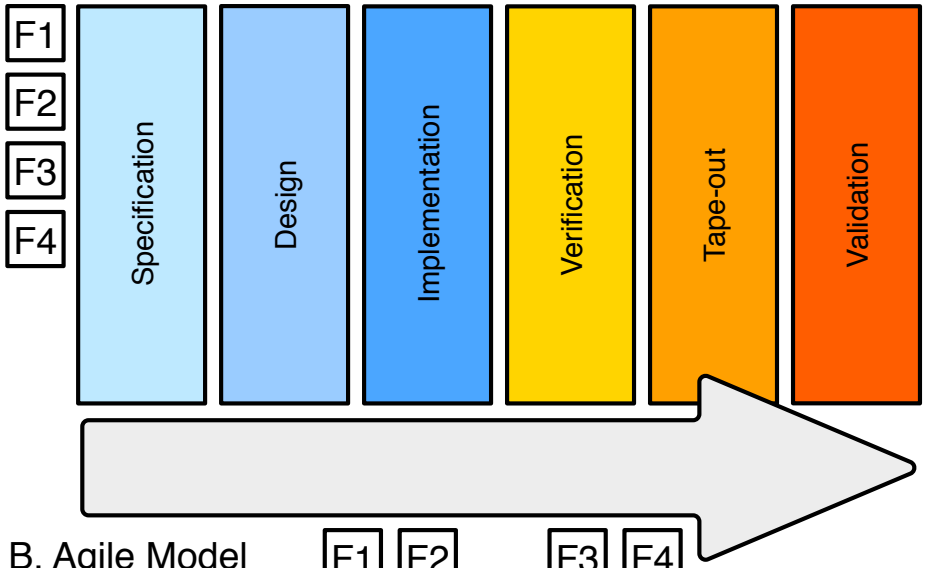


1. Change Parameters
2. Develop New Accelerators
3. Develop Own RISC-V Core
4. Develop Own Device

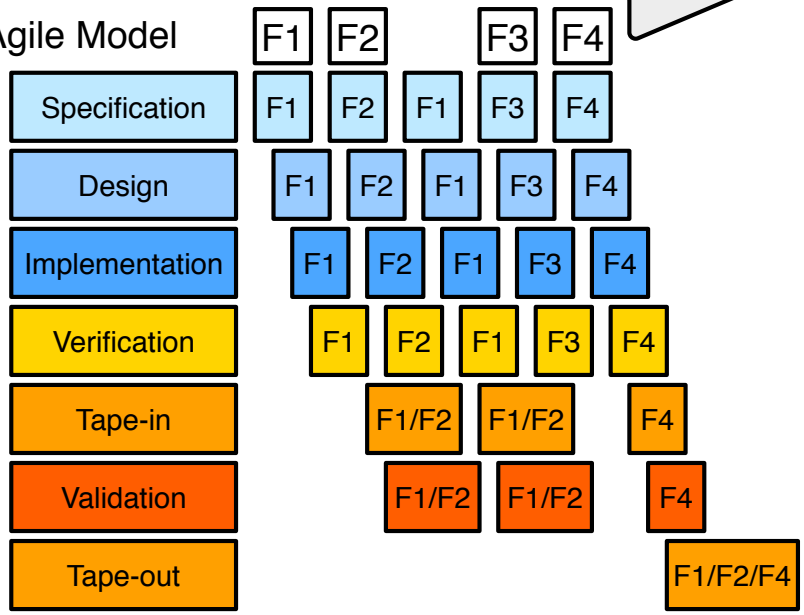
Open-source Rocket Chip generator widely used in both academia and industry. Rocket Chip being used in several products.

Agile Hardware Development

A. Waterfall Model

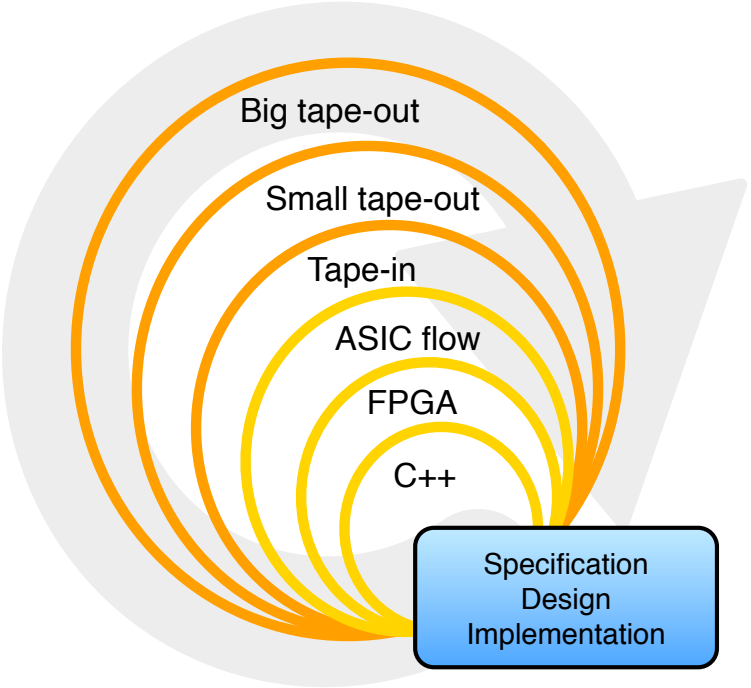


B. Agile Model

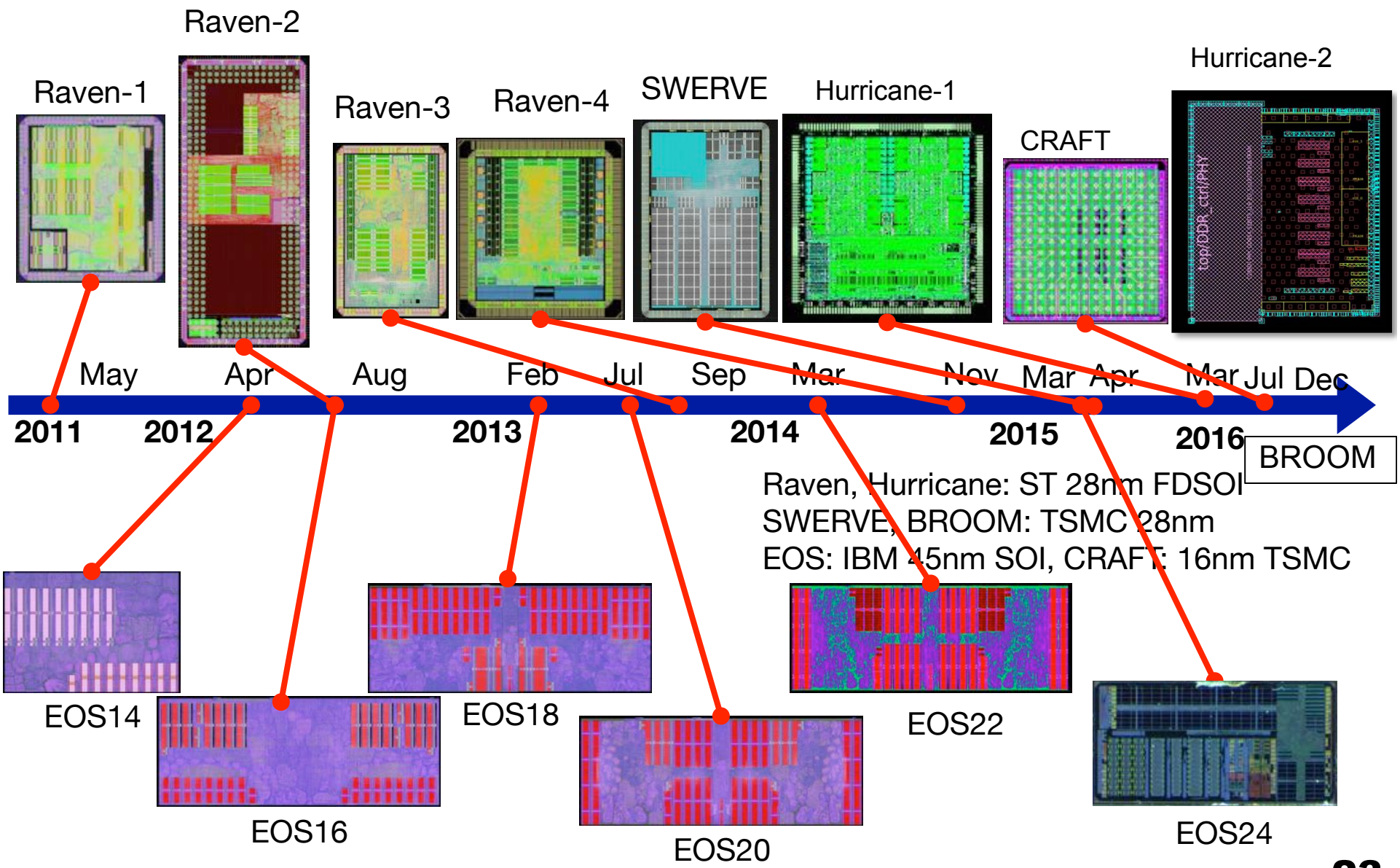


Always have a fabricatable working prototype, incrementally add features and push to "tapein" often

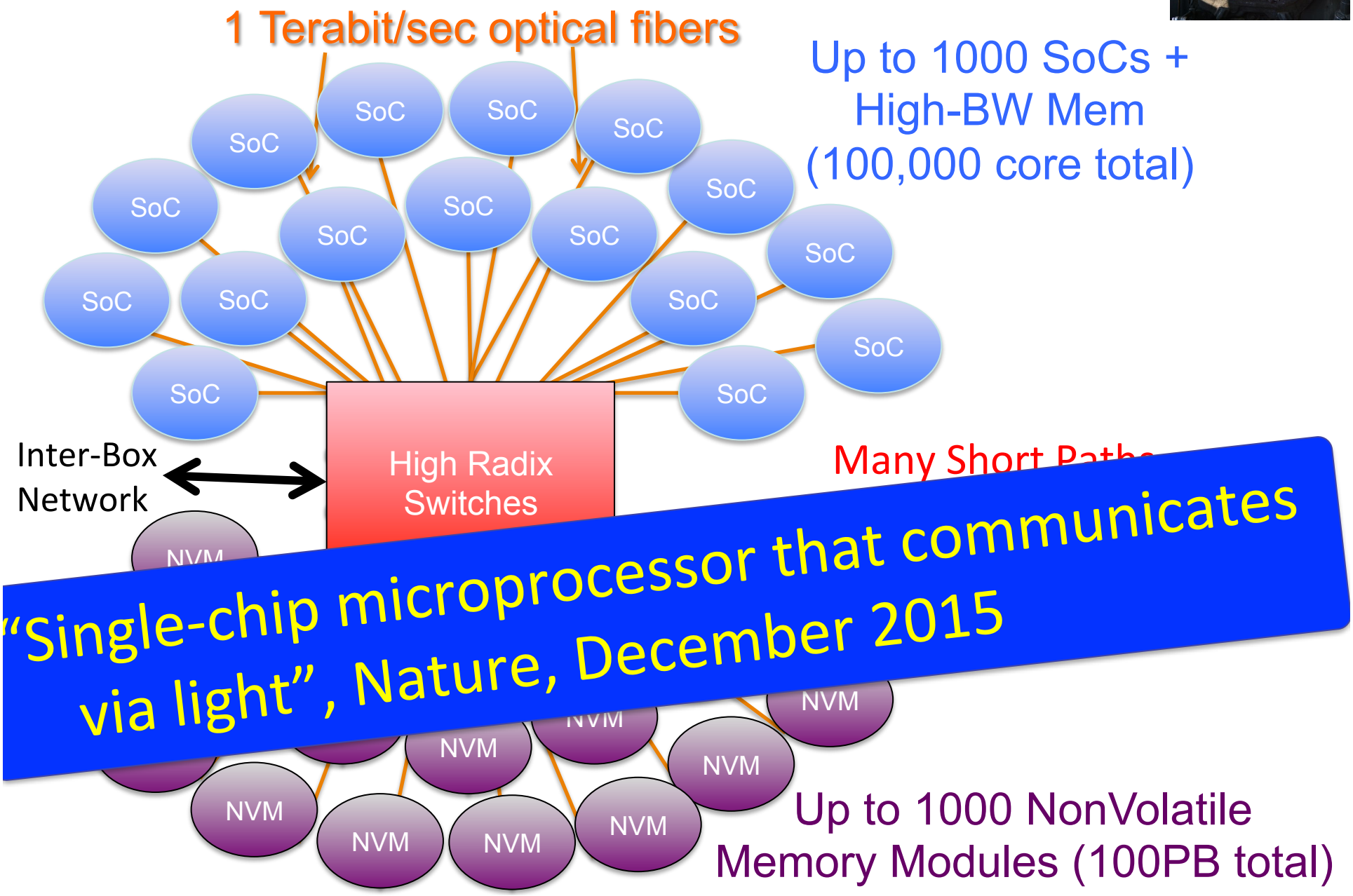
C. Validation Through Agile Iteration



RISC-V Chips Designed at Berkeley



FireBox Overview

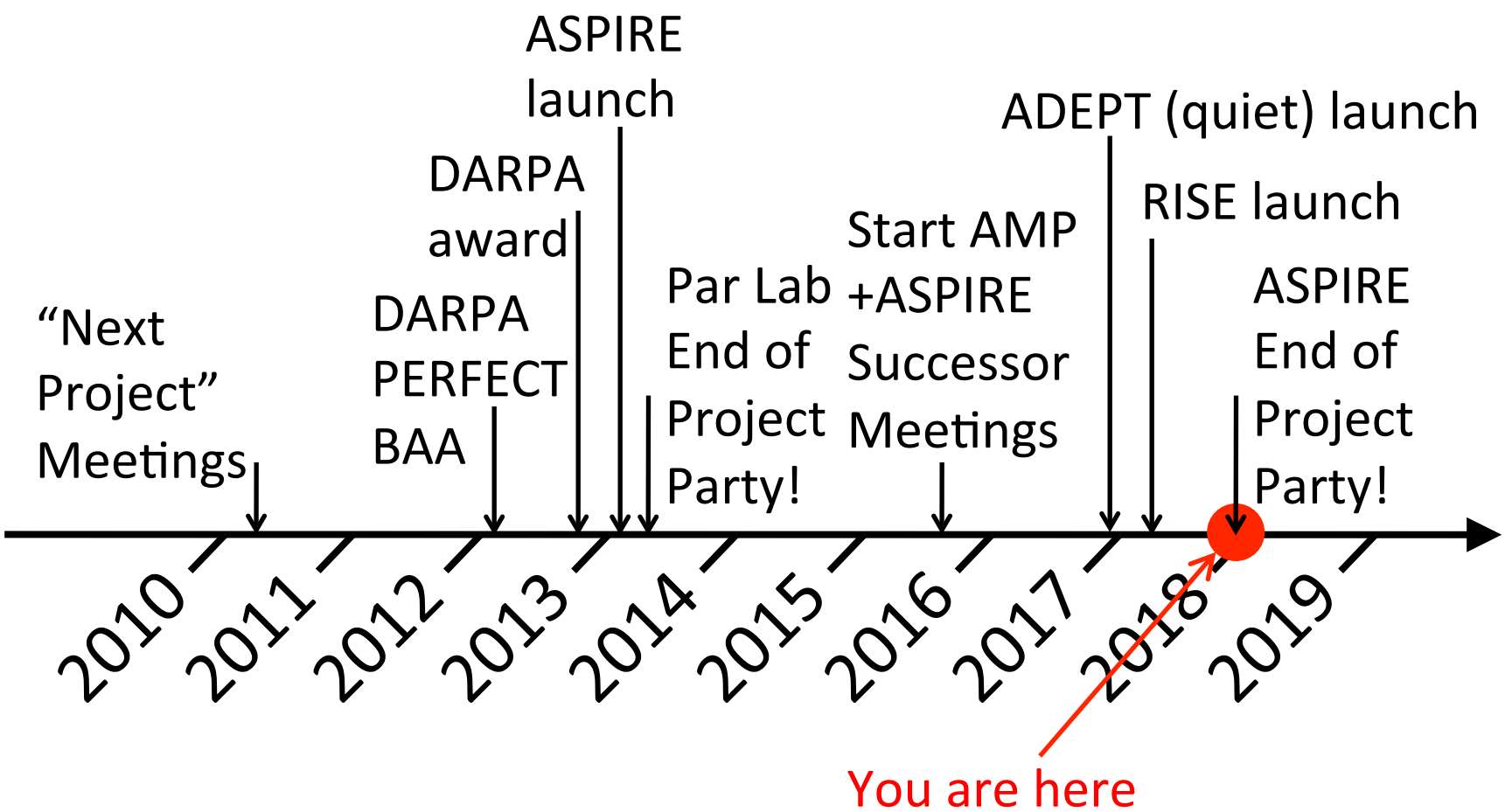


"Single-chip microprocessor that communicates via light", Nature, December 2015

ASPIRE Summary

- Architect and program whole ***applications*** using computational and structural patterns
- Develop ***algorithms*** that minimize data movement
- Pattern-specific ***compilers*** to specialize and autotune software for best hardware performance
- New "general-purpose specialized" ***architectures*** employing pattern-specific acceleration
- Parameterized ***generators*** and design-space exploration to optimize hardware µarchitecture
- Resilient ***circuit*** design, new tech. (photonics, NVM)
- Fast flexible accurate whole-system ***FPGA simulation***
- Real ***silicon prototypes*** to validate ideas

ASPIRE Timeline



ASPIRE Sponsors

- DARPA PERFECT program
- DARPA POEM program (Si photonics)
- DARPA CRAFT program (new)
- STARnet Center for Future Architectures (C-FAR)
- Lawrence Berkeley National Laboratory
- Industrial sponsors
  - Intel
- Industrial affiliates
  - Google
  - HPE
  - Huawei
  - LGE
  - NVIDIA
  - Oracle
  - Samsung

ASPIRE
UC Berkeley

riselab
UC Berkeley

riselab UC Berkeley

PEOPLE PROJECTS PUBLICATIONS Q

Making Intelligent Decisions on Big Data

In the RISELab, we are working on developing Secure Real-time Decision Stack, an open source platform, tools and algorithms for low-latency decisions on live data with strong security.

Agile Design of Efficient Processing Technologies

Krste Asanovic, Bora Nikolic, Elad Alon,
Jonathan Bachrach, Jonathan Ragan-Kelley,
Koushik Sen, Sanjit Seshia

ADEPT: New 5-Year Project

adept.eecs.berkeley.edu

- Moore's Law has effectively ended
  - But silicon manufacturing is already wonderful technology
- General-purpose computing stuck in plateau
  - Software's free ride is over
- Specialized silicon needed for new applications
  - Machine learning/inference, IoT sensing/processing, automotive,…
- But too expensive to design custom chips today!
  - semiconductor industry model broken for lots of specialized chips.
- Solutions: Reduce NRE!
  - Develop next-generation of chip-design automation
  - Back to the future: use *right* node instead of *next* node
  - New industry business models: share risks, increase rewards
  - Open-source hardware: RISC-V ISA is one element

CMOS at End of Moore's Law

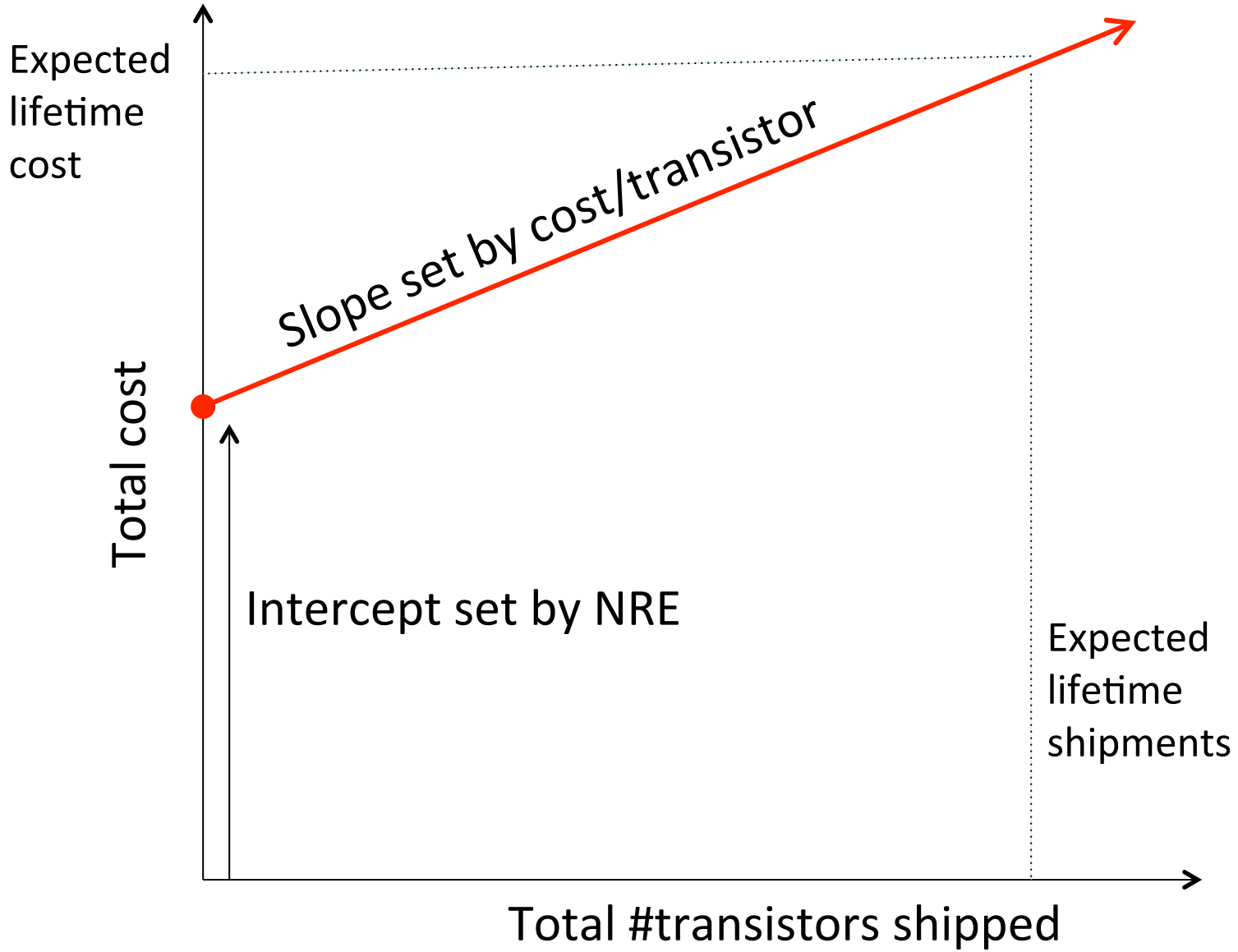
- Wonderful, almost magical, technology
- Fabricate billions of transistors
- Connect them with billions of wires
- Clock at a few GHz
- Dissipate <100W
- Near 100% yield, cost a few $/die

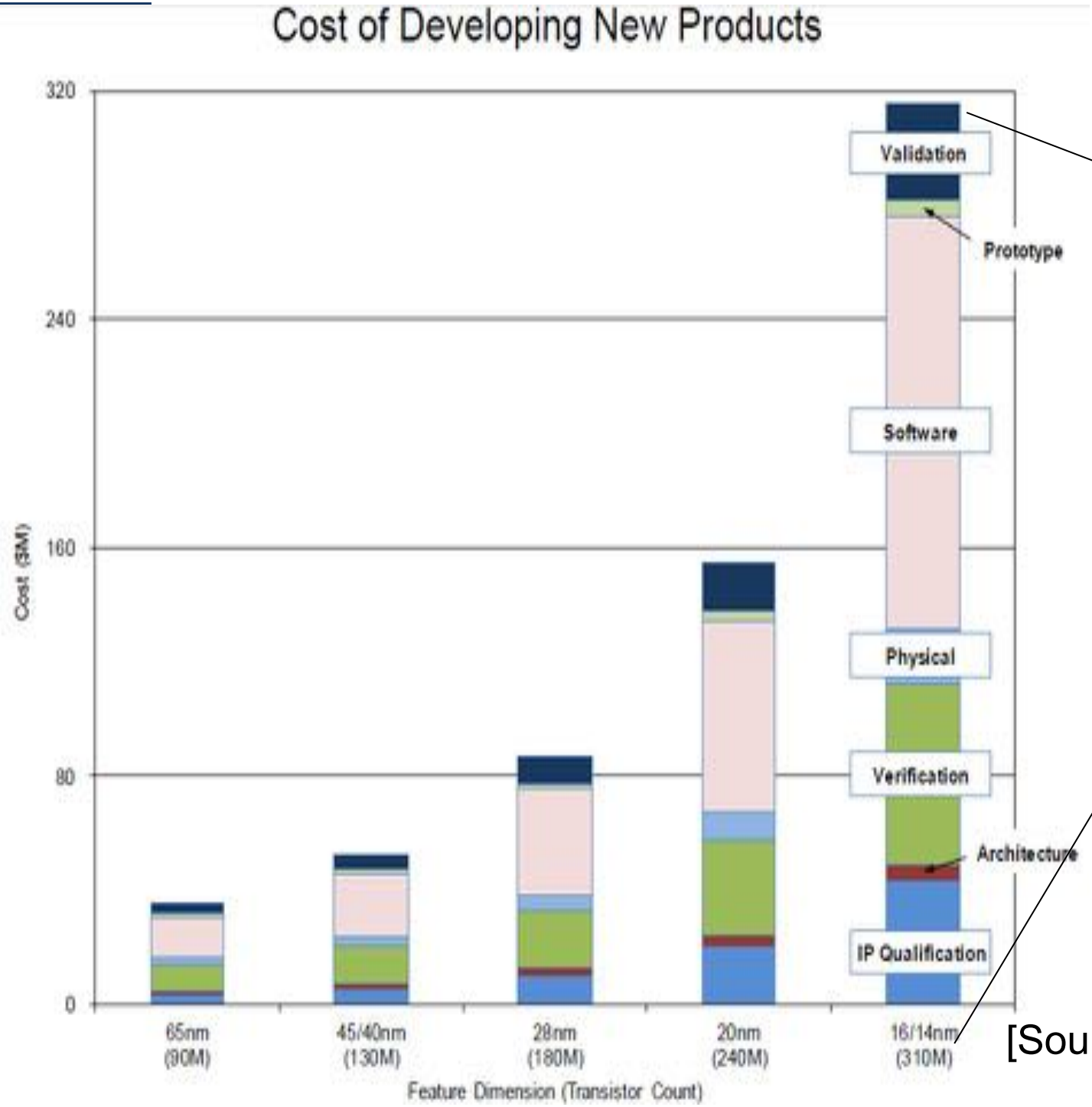
- Manufacturing is least of our problems
- Good, because nothing waiting in wings to replace anytime soon (20+ years to competitor?)

Custom Chip Costs

- NRE: Non-Recurring Engineering costs
  - Design + tooling for production
- Manufacturing cost
  - Cost of each chip made once in production
  - Silicon manufacture plus other costs including package & test, which get worse relatively with smaller nodes

Overall Cost/Project



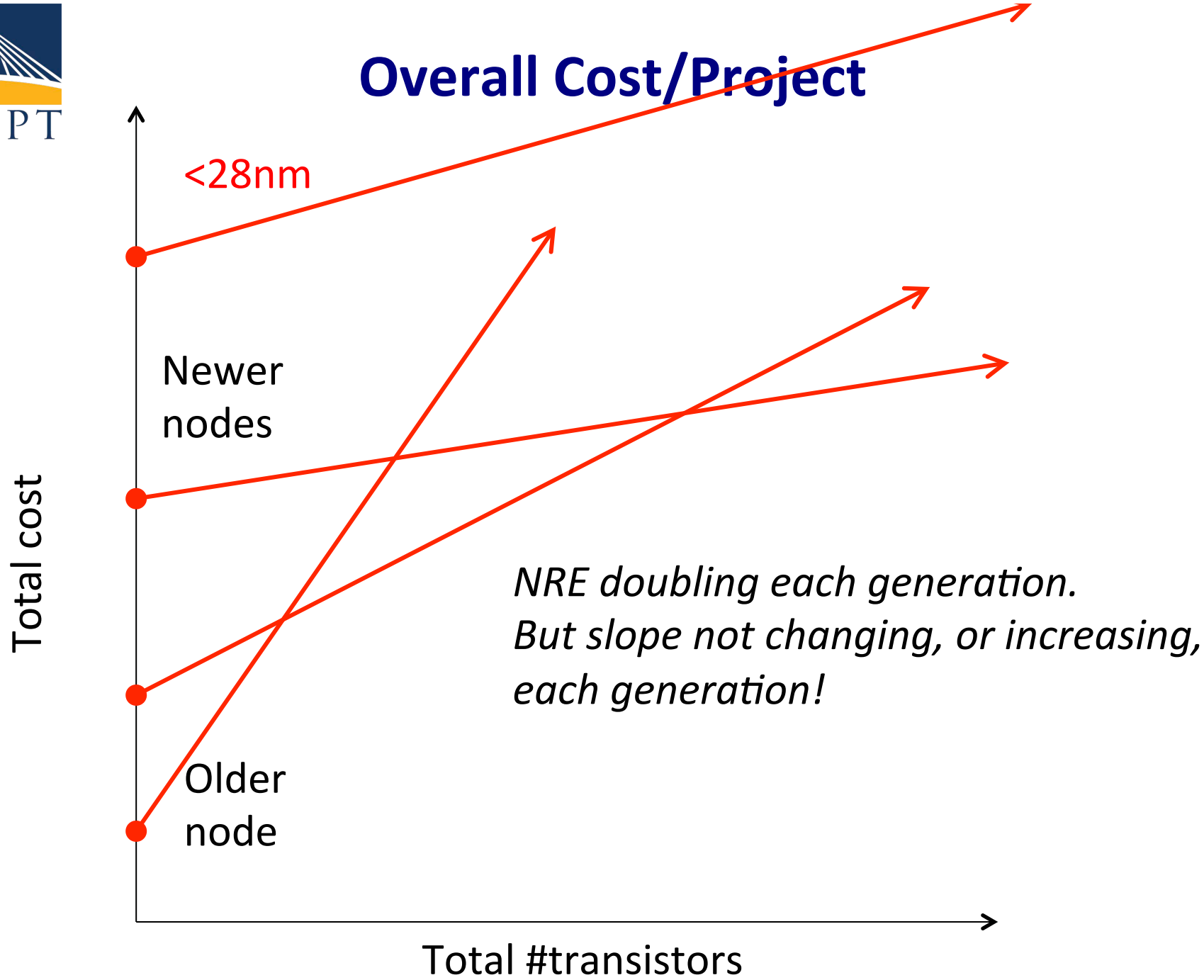


~$300M development cost

~$5/chip manufacturing cost

[Source: IBS]

Overall Cost/Project



Old Semiconductor Business Model

- Design standard part for the "killer socket"
  - Was PC, now smartphone
- Sell 100s millions parts
- Ideally, ~0 customers, ~infinite volume
- Not working so well now with high NREs, end of Moore's Law, and fragmenting markets

No good ideas, so merge to cut costs

Biggest Semiconductor Acquisition Agreements

Ranking	Acquisition-Buyer (Year Announced)	Price Tag ($B)
1	**NXP** by Qualcomm (2016)	$39.0
2	**Broadcom** by Avago (2015)	$37.0
3	**ARM** by SoftBank (2016)	$32.0
4	**SanDisk** by Western Digital (2015)	$19.0
5	**Freescale** by U.S Investment Companies (2006)	$17.6
6	**Altera** by Intel (2015)	$16.7
7	**Linear Technology** by Analog Devices (2016)	$14.8
8	**Freescale** by NXP (2015)	$11.8
9	**Burr Brown** by TI (2000)	$7.6
10	**LSI** by Avago (2013)	$6.6
11	**National Semiconductor** by TI (2011)	$6.5
12	**ATI** by AMD (2006)	$5.4
13	**Spansion** by Cypress (2014)	$5.0
14	**Agere** by LSI (2006)	$4.0
15	**Chartered** by GlobalFoundries (2009)	$3.9

7 of top 8 in last 2 years

Source: Companies, IC Insights (*2017 McClean Report*)

New Silicon Model emerging

Instead of chip company's standard product, chip customers want own differentiated designs:

- Apple, Samsung for phones
- Google, Amazon, Microsoft, for client/cloud
- Tesla?, Uber? for cars?
- IoT products, FitBit? maybe?
- End-system value/profit justifies chip NRE

But what about the non-huge firms that can't risk NRE, or don't have chip experience?

- Can we cut NRE to <<$100M, or <<$10M or <<$1M?

Attacking all the NRE components

Using data center (**FireBox**) and embedded vision chips (**Raven/Hurricane/Eagle**) as driving applications:

- Reduce software cost with high-level programming tools (**Halide,**serverless) and standard ISA (**RISC-V**)
- Reduce architecture design and verification with reusable open-source chip generators (**Chisel/FIRRTL**)
- Share costs of analog/mixed-signal IP development by sticking to a few nodes, analog generators (**BAG**), maybe open-source analog!
- Automate backend physical design, save on tool costs (**HAMMER/FICL**)
- Improve verification and validation, accelerate with cloud-hosted FPGAs

ADEPT Sponsors

- Cornerstone funding: **Intel Science and Technology Center on Agile Design** (*ADEPT: Agile Development of Efficient Processing Technologies*)
  - PIs: Asanovic, Nikolic, Bachrach, Ragan-Kelley, Seshia
  - Funding for 3 years initially, started in December 2016
- Also *AHDEPT: Agile Hardware Design in Extreme Process Technologies*
  - Part of **DARPA CRAFT** program, started in May 2016
  - Pis: Alon, Nikolic, Bachrach, Sen
  - Partners: Northrop-Grumman, Cadence
- Industrial sponsors: **Siemens** and **SK Hynix**
- Looking for more!
- First retreat: Monterey, January 8-10, 2018

ASPIRE End of Project

- It's been a great adventure
- Thank you all for feedback over the years!
- Many impactful results
  - SqueezeDet -> DeepScale startup
  - CA Algorithms -> Many awards
  - FireBox photonics -> Nature article on first photonic micro, Ayar Labs startup
  - RISC-V ISA -> widespread industry and academia adoption
  - Chisel + Agile Hardware Design -> SiFive, JITx startups
  - **https://aspire.eecs.berkeley.edu/aspire-best-of-papers/**
- Many amazing graduated/graduating students
- Strong foundation for the next 5-year ADEPT project