Report from the Arch2030 Visioning Workshop: Where are Computer Architects headed and what does it mean for GreenMetrics?

Thomas F. Wenisch

(Special acknowledgements to Luis Ceze, Mark Hill, & 40+ members of the architecture community)

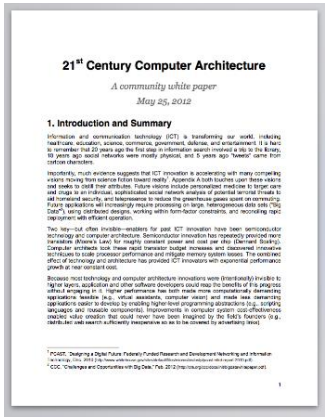
Arch2030 was supported by the Computing Community Consortium (CRA).

The (quick) backstory… (1/2)

Moore's Law is ending. For real this time.

In 2011, Architects (via the Computing Community Consortium) sent a white paper to NSF on "21st Century Computer Architecture"

Contributed to launch of NSF eXploiting Parallelism and Scalability program



21st Century Computer Architecture

A community white paper

May 25, 2012

1. Introduction and Summary

Information and communication technology (ICT) is transforming our world, reducing healthcare, education, science, commerce, government, defense, and entertainment. It is hard to remember that 20 years ago the first step in information search involved a trip to the library, 10 years ago social networks were mostly physical, and 5 years ago "tweets" came from cartoon characters.

Incontinently, much evidence suggests that ICT innovation is accelerating with many compelling visions moving from science fiction toward reality. Appendix A both touches upon these visions and seeks to distill their attributes. Future visions include personalized medicine to target care and drugs to an individual, sophisticated social network analysis of potential terrorist threats to aid homeland security, and telepresence to reduce the greenhouse gases spent on commuting. Future applications will increasingly require processing on large, heterogeneous data sets ("Big Data"), using distributed designs, working within form-factor constraints, and reconciling rapid deployment with efficient operation.

Two key—but often invisible—enablers for past ICT innovation have been semiconductor technology and computer architecture. Semiconductor innovation has repeatedly presented new iterations (Moore's Law) for roughly constant power and cost per chip (Dennard Scaling). Computer architects took these rapid transistor budget increases and discovered innovative techniques to scale processor performance and mitigate memory system losses. The combined effect of technology and architecture has provided ICT innovators with exponential performance growth at near-constant cost.

Because most technology and computer architecture innovations were (intentionally) invisible to higher layers, application and other software developers could reap the benefits of this progress without engaging in it. Higher performance has both made more computationally demanding applications feasible (e.g., virtual assistants, computer vision) and made less demanding applications easier to develop by enabling higher-level programming abstractions (e.g., scripting languages and reusable components). Improvements in computer system cost-effectiveness enables value creation that could never have been imagined by the field's founders (e.g., distributed web search sufficiently inexpensive so as to be covered by advertising fees).

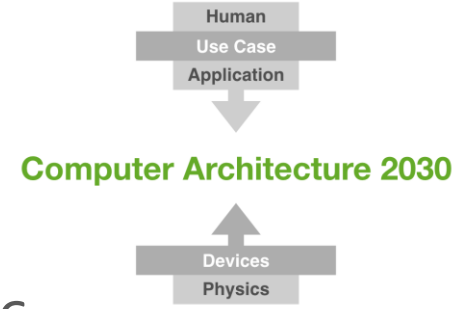
[1] PCAST, "Designing a Digital Future: Federally Funded Research and Development Networking and Information Technology," Dec. 2010 (http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf).

[2] CCC, "Challenges and Opportunities with Big Data," Feb. 2012 (http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf).

But, the world has changed in 5 years;

the community did not foresee some critical trends…

The (quick) backstory… (2/2)



To update the architecture research vision,

CCC sponsored **Arch2030** workshop with ISCA 2016.

This talk:

1) Context: What did architects say in 2011, and what is different now?
2) Summarize the technical story & recommendations of these whitepapers
3) Opine on where it intersects with GreenMetrics

20th century ICT set up:

Information & Communication Technology (ICT) has changed our world

<long list omitted>

Required innovations in algorithms, applications, programming languages, … ,
& system software

Key (invisible) enablers to (cost-)performance gains:

Semiconductor technology (“Moore’s Law”)

Computer architecture (~80x per Danowitz et al.)

21st century ICT promises more

Data-centric personalized health care

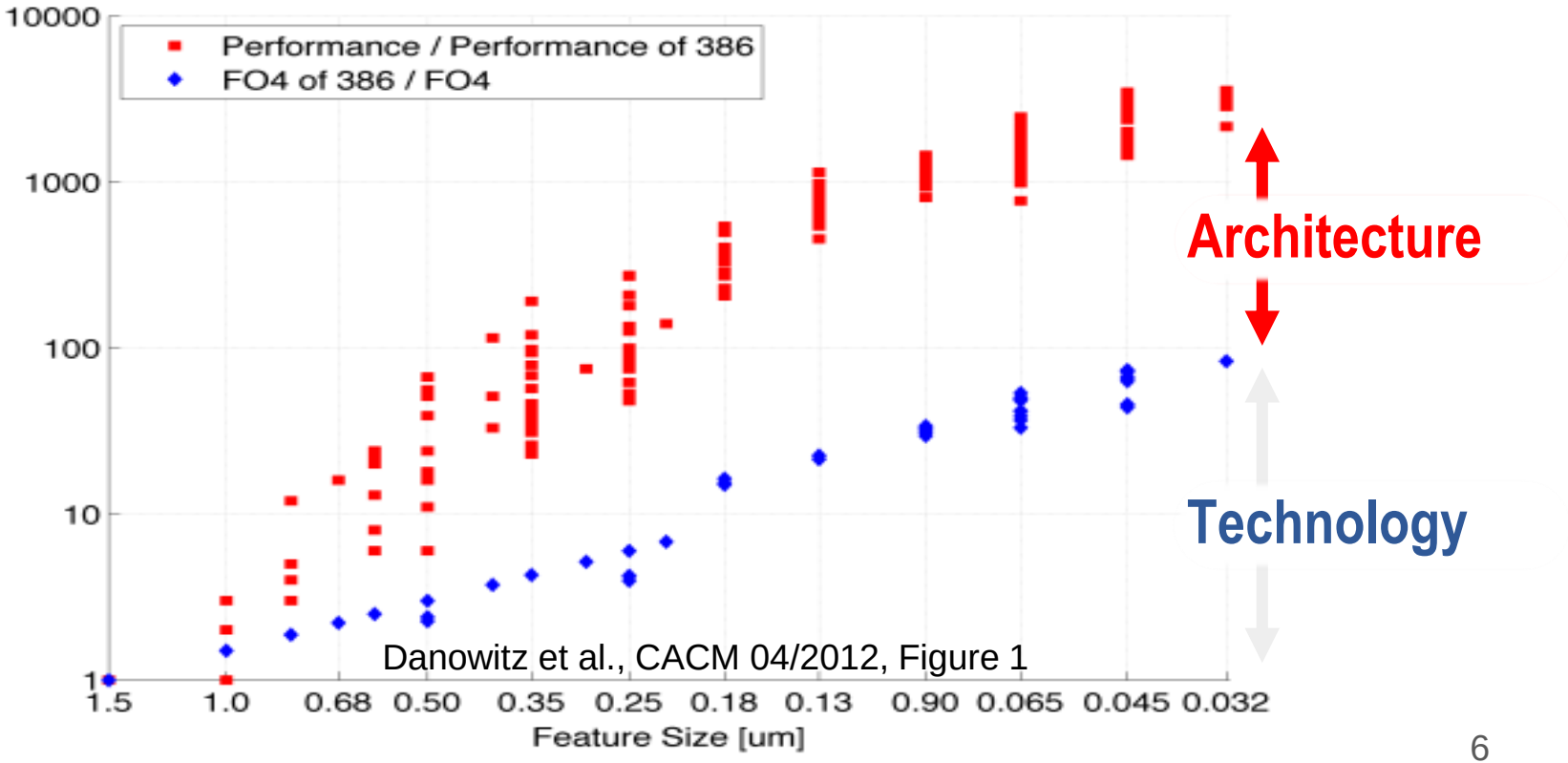
Computation-driven scientific discovery

"You never call, and the federal government will back me up on that."

Human network analysis

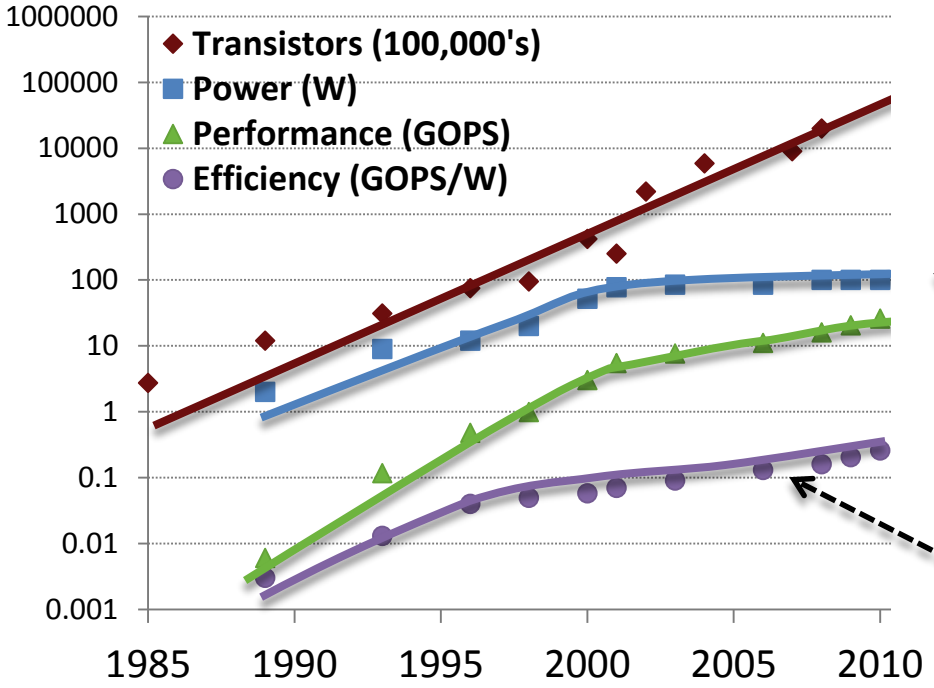
Much more: known & unknown

Enablers: Technology + Architecture



Danowitz et al., CACM 04/2012, Figure 1

Technology: a paradigm shift in the 2000s…

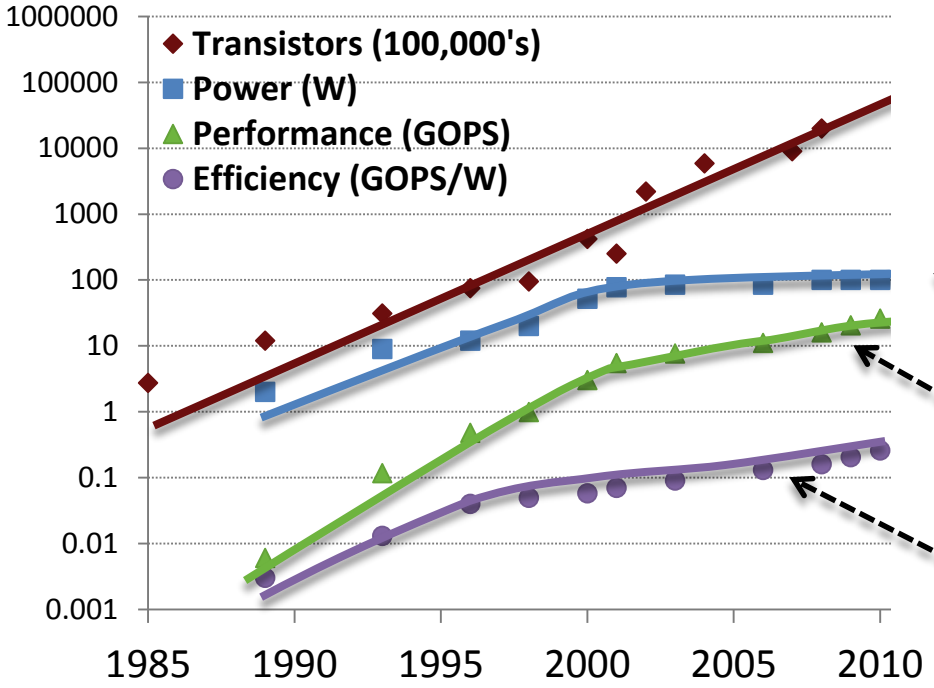


IEEE Computer—April 2001
T. Mudge

Limits on heat extraction

Limits on energy-efficiency of operations

Technology: a paradigm shift in the 2000s…



IEEE Computer—April 2001
T. Mudge

Limits on heat extraction

Stagnates performance growth

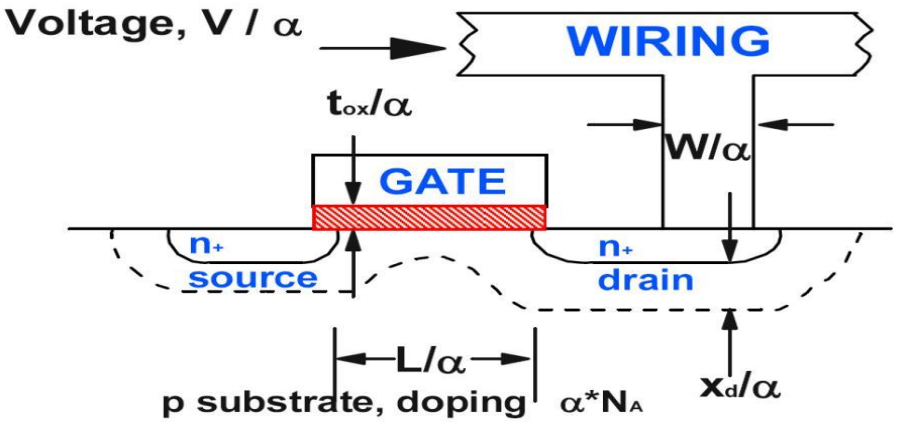
Limits on energy-efficiency of operations

Era of Delay-Constrained Computing

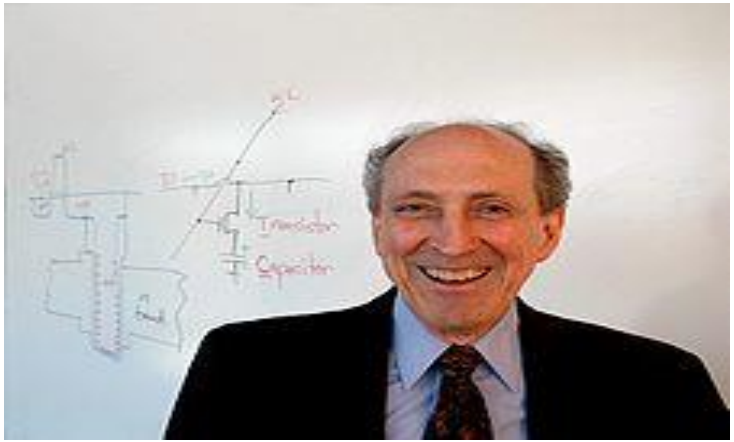
Era of Power-Constrained Computing

c. 2000

The magic underlying technology gains: Dennard Scaling



Dennard et. al., 1974



Robert H. Dennard, picture from Wikipedia

Dennard scaling in a nutshell:

Power = Capacitance × Voltage^2 × frequency

Scale transistor by α → density grows by α^2 → power nominally grows by α^2

To compensate: lower voltage by α → more transistors; constant power!

So what happened? Leakage killed Dennard Scaling

To distinguish zeros and ones, supply voltage must be about triple the transistor switching (threshold): $V_{dd}/V_{th} > 3$

So, scaling down supply requires scaling down threshold

But, transistor leakage power is exponential in V_{th}

→ V_{dd} can't go down anymore

No more free lunch… (2011 edition)

Dark Silicon: can't use all transistors all the time

Need system-level approaches to…

…turn increasing transistor counts into customer value

…without exceeding thermal limits

Energy efficiency is the new performance

21st Century Arch. – Key Challenges

Late 20th Century	The New Reality
Moore's Law — 2× transistors/chip	**Transistor count still 2× BUT…**
Dennard Scaling —~constant power/chip	**Gone.** Can't repeatedly double power/chip
Modest (hidden) transistor unreliability	**Increasing transistor unreliability** can't be hidden
Focus on computation over communication	**Communication (energy)** more expensive than computation
1-time costs amortized via mass market	**One-time cost** much worse & want **specialized** platforms

How should architects step up as technology falters?

Recommendations from 2011

20th Century	21st Century	
Single-chip in stand-alone computer	**Architecture as Infrastructure:** Spanning sensors to clouds Performance + security, privacy, availability, programmability, …	**Cross-Cutting:** Break current layers with new interfaces
Performance via invisible instr.-level parallelism	**Energy First** • Parallelism • Specialization • Cross-layer design	
Predictable technologies: CMOS, DRAM, & disks	**New technologies** (non-volatile memory, near-threshold, 3D, photonics, …) Rethink: memory & storage, reliability, communication	

Six years elapse…
… and some new realities emerge

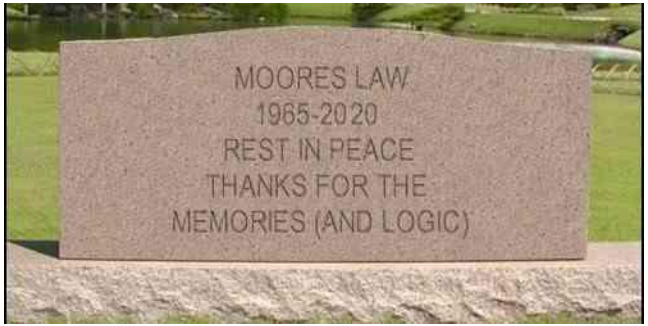
What changed?

Machine learning is a key workload

Specialization already happening at scale

Cloud is truly ubiquitous

Wide acceptance Moore's Law is really ending



Arch2030 Visioning Workshop: Process

Reached out to prior efforts

21st Century CA, Rebooting Computing

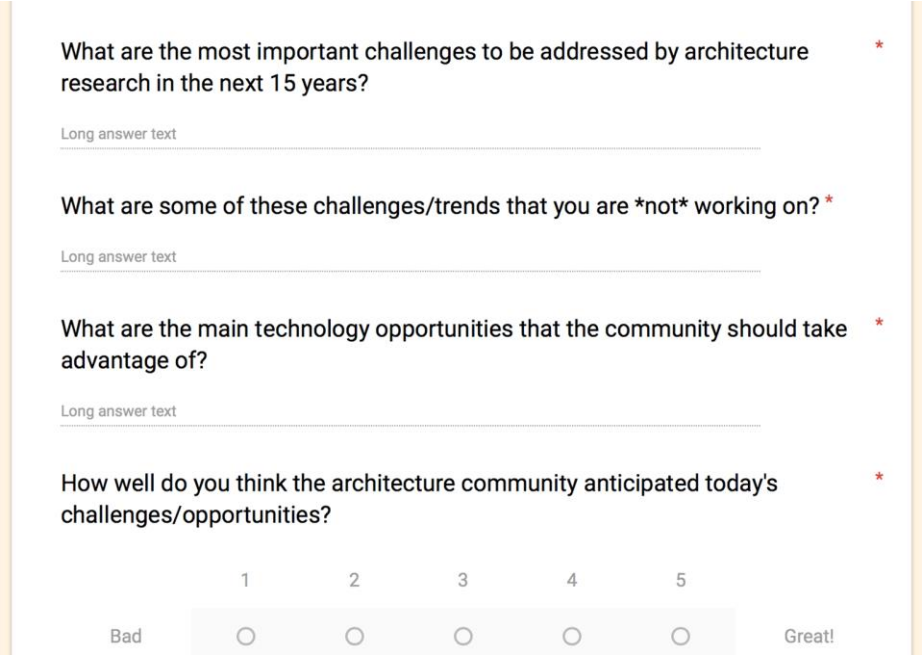
Reached out to community for input

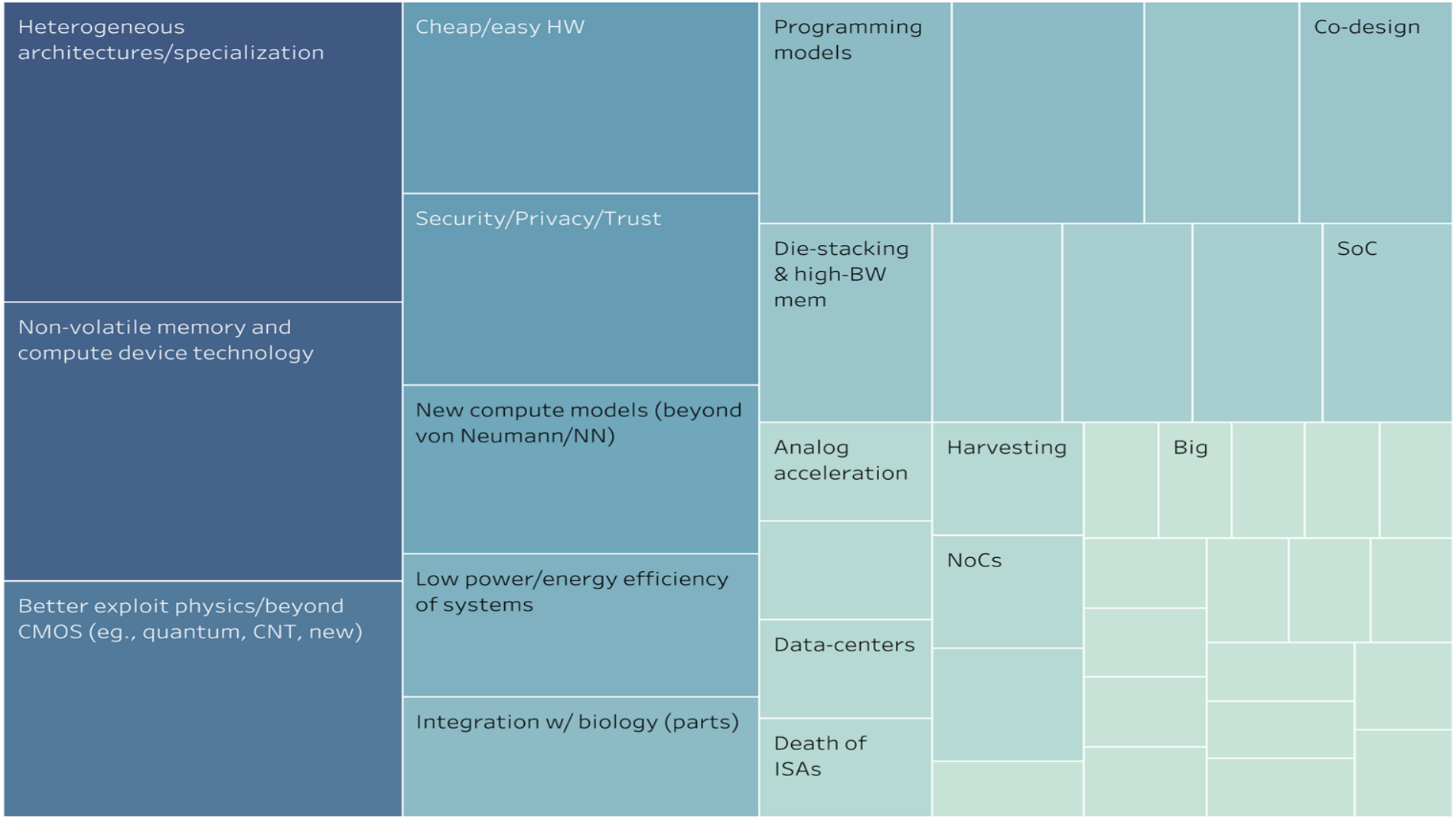
Invited experts (devices, applications)

Held with ISCA: 120+ participants

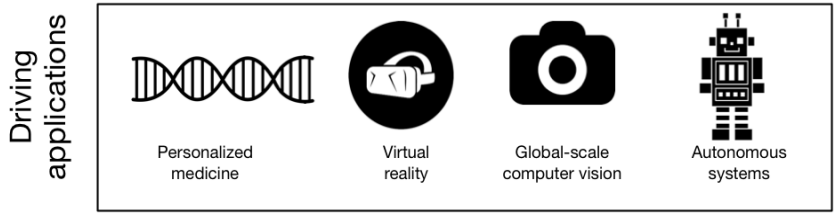
Sent out report for comments

40+ endorsers




Heterogeneous architectures/specialization
Non-volatile memory and compute device technology
Better exploit physics/beyond CMOS (eg., quantum, CNT, new)
Cheap/easy HW
Security/Privacy/Trust
New compute models (beyond von Neumann/NN)
Low power/energy efficiency of systems
Integration w/ biology (parts)
Programming models
Co-design
Die-stacking & high-BW mem
SoC
Analog acceleration
Harvesting
Big
NoCs
Data-centers
Death of ISAs

Arch2030: The upshot



Observation	***Implications for next 15 year***
1. Specialization gap	Democratize HW design: tools and open source designs
2. Ubiquitous cloud: innovation abstraction	Cloud model provides practical deployment path for new architectures
3. 3D stacking is real	Opportunities for new architectures and integration models
4. Getting “closer to physics”	Need for more adventurous architectures
5.Machine learning as key app. component	New architectures are enablers: need real collaboration with core ML community

1. Specialization Gap: Democratizing HW Design

Performance gap: many applications aren't possible without specialization

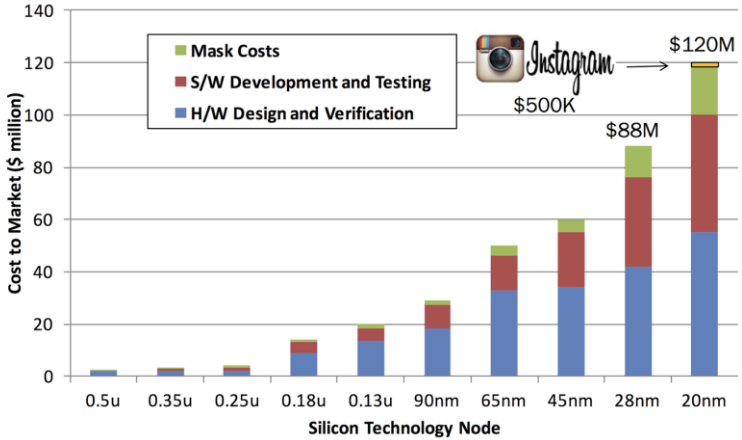
- AR/VR, autonomous vehicles, large-scale AI/ML
- General purpose processors aren't efficient enough

Design cost/effort gap:

- HW design costs growing too fast
- Need: better models, tools, open-source design

Can create new business/innovation forces

- Emerging "HW" companies: fitbit, Oculus, Pebble, Dropcam, …
- Open source can create agility for ASIC-based startups



Source: International Business Strategies, T. Austin

Developing specialized hardware must become as easy, inexpensive, and agile as developing software

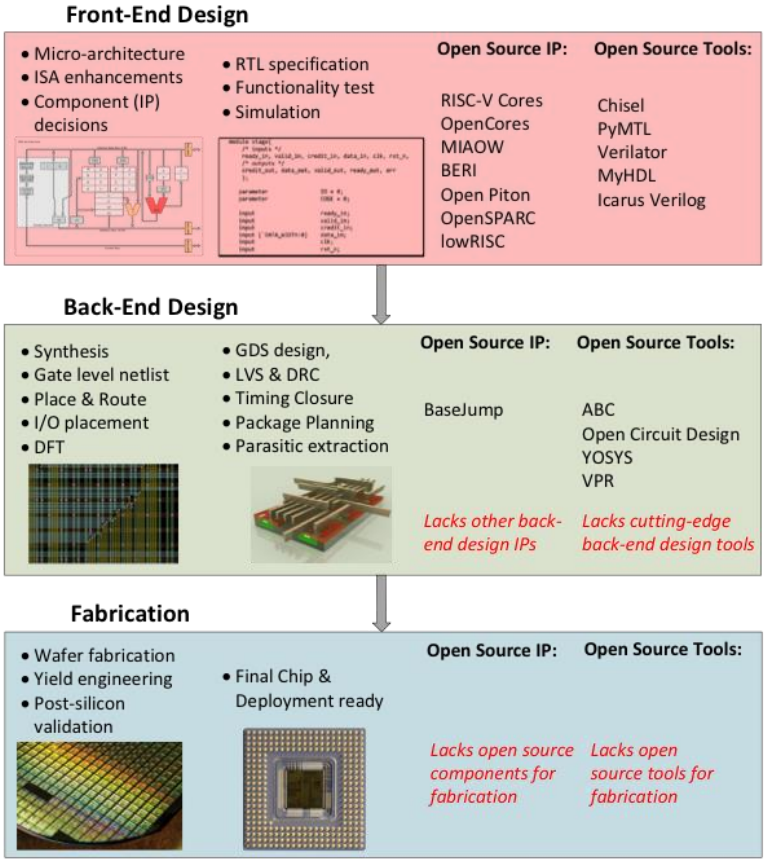
Opportunity:
Open-source hardware

Need infrastructure to reduce barrier-to-entry for custom ASICs

Faster impact via tightly integrated FPGAs

Need open/reusable IP cores and tools

Investigate “chiplet” / post-fab integration



Sankaralingam et al.

2. Cloud as Abstraction for Architectural Innovation

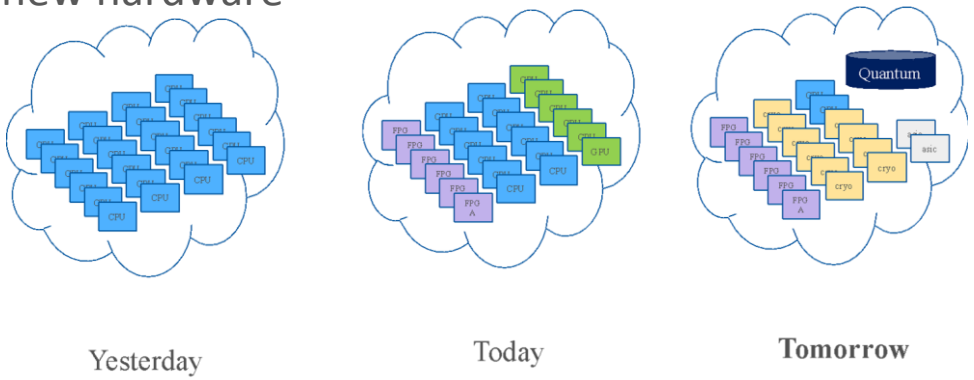
Ubiquitous public cloud infrastructure (Microsoft, Google, Amazon)

More than just software - entry point for new hardware

Clean service/microservice interfaces

Can hide exotic HW/devices

ASICs, FPGAs, quantum computers?



[Doug Carmean, ISCA'16 Keynote]

Through scale and virtualization, clouds can offer deep HW innovations transparently and at low cost

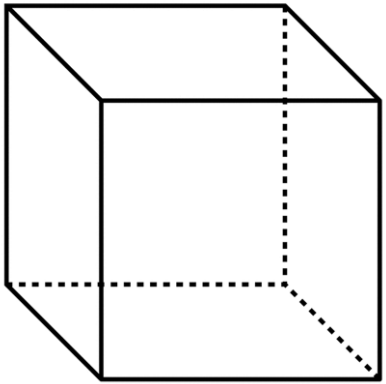
3. Going Vertical with 3D Integration

Denser memories, higher bandwidth

Capacity/bandwidth grows

Fundamental need for processing+memory integration

Integration of “chiplets” in 3D substrate a promising design/business model



$$\text{capacity} \propto L^3$$

$$\text{bandwith} \propto L^2$$

3D integration provides a new dimension of scalability

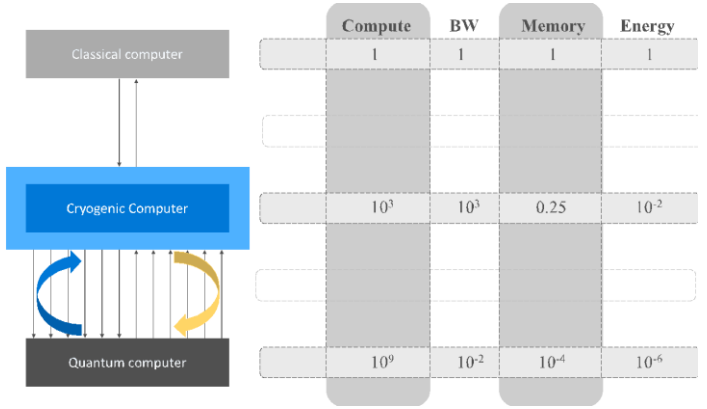
4. Getting Closer to Physics

New memories and devices

Carbon nanotubes

Quantum computing and superconducting logic

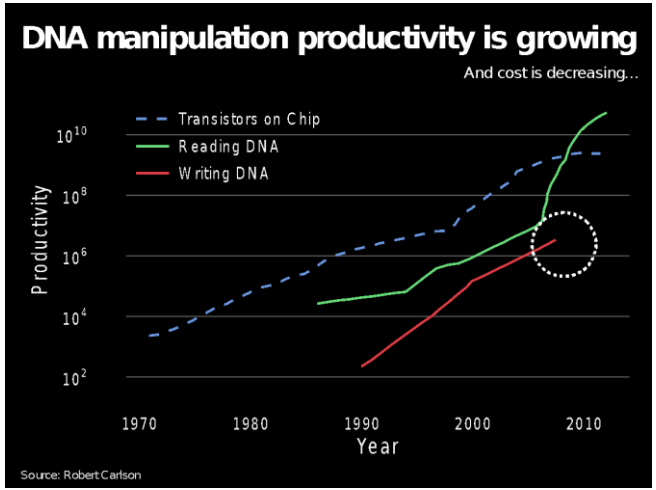
Borrowing from biology



[Doug Carmean, ISCA'16 Keynote]

	Access Time	Capacity	Durability
Flash	μs-ms	TBs	~5 yrs
HDD	10s ms	100s TBs	~5 yrs
Tape	minutes	PBs	~10s yrs
DNA-based Archival	hours	ZBs	~1000s yrs

[Bornholt et al.]



5. Machine Learning as a Key Workload

Training: HPC-like systems, turn-around time matters to evolution

Inference: Low latency, low power

Strong driver for architecture and systems innovation

Tensor flow, TPUs, MS CNTK, …

Google's Tensor Processing Unit

Hardware advancement enables machine learning over "bigger data"

The Future: Architecture + *X*

Application and technology driven

It's clear we are beyond a processor + memory centric world

Examples: sensor/compute fusion, intelligent networks, intelligent storage systems

Critical to reach out to other CS areas and fields

What does it mean for GreenMetrics?

What does it mean for GreenMetrics? (1 of 2)

All of computing needs to think about “Democratizing HW Design”

- What does source hardware mean for sustainability?
- Can we foster a “Github movement” for HW?
- What does the tech transfer pipeline from idea to system look like?

“The Cloud” does not mean commodity servers anymore

- Tensor Processing Unit
- Project Catapult at Microsoft
- FPGA instances at Amazon

What does it mean for GreenMetrics? (1 of 2)

Machine Learning is everywhere

How do we enable sustainable training & inference?

Unsustainable to do all this compute in centralized data centers;
it needs to be pushed to the client/edge

Exotic hardware is coming

3D transistors? Memristors? Carbon Nanotubes?

Quantum? Biology inspired? DNA storage?