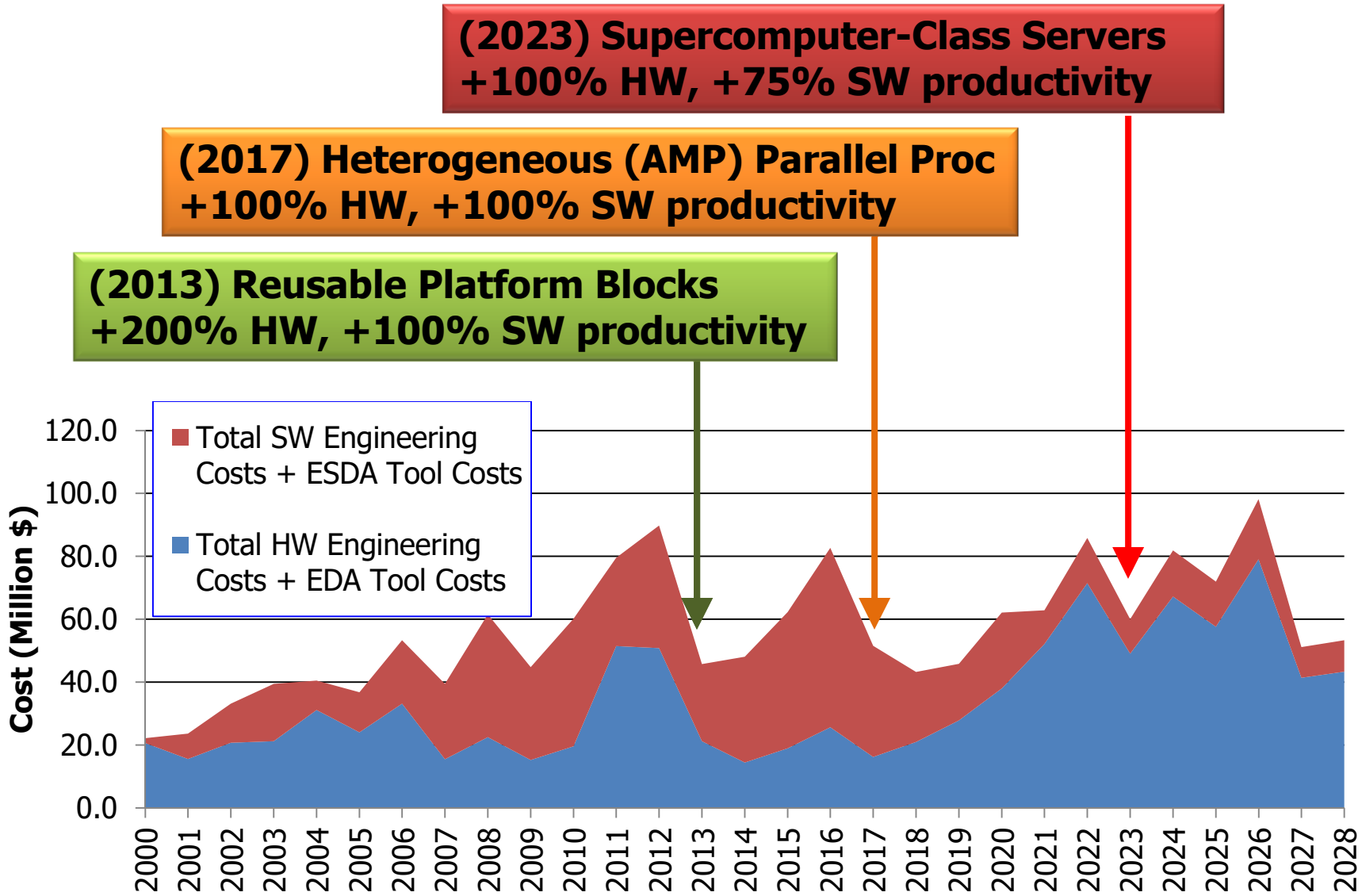
DA Perspectives and Futures: An Update

Andrew B. Kahng
CSE and ECE Department
UC San Diego
http://vlsicad.ucsd.edu/~abk/
abk@ucsd.edu

8:30AM: Wake Up, DA is Great!

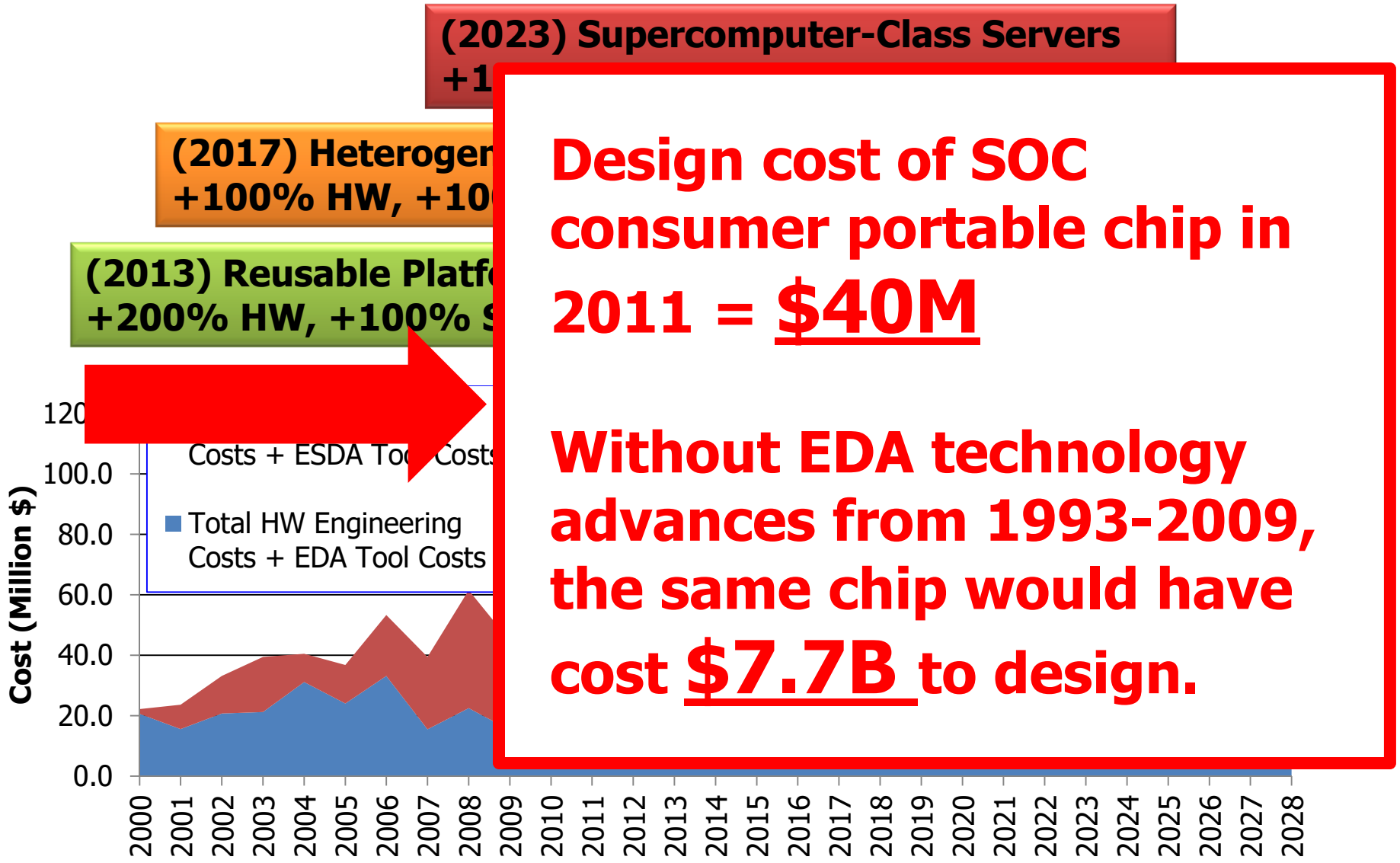
ITRS 2011, DAC13



A. B. Kahng IWLS 2017 keynote, 170618

8:30AM: Wake Up, DA is Great!

ITRS 2011, DAC13



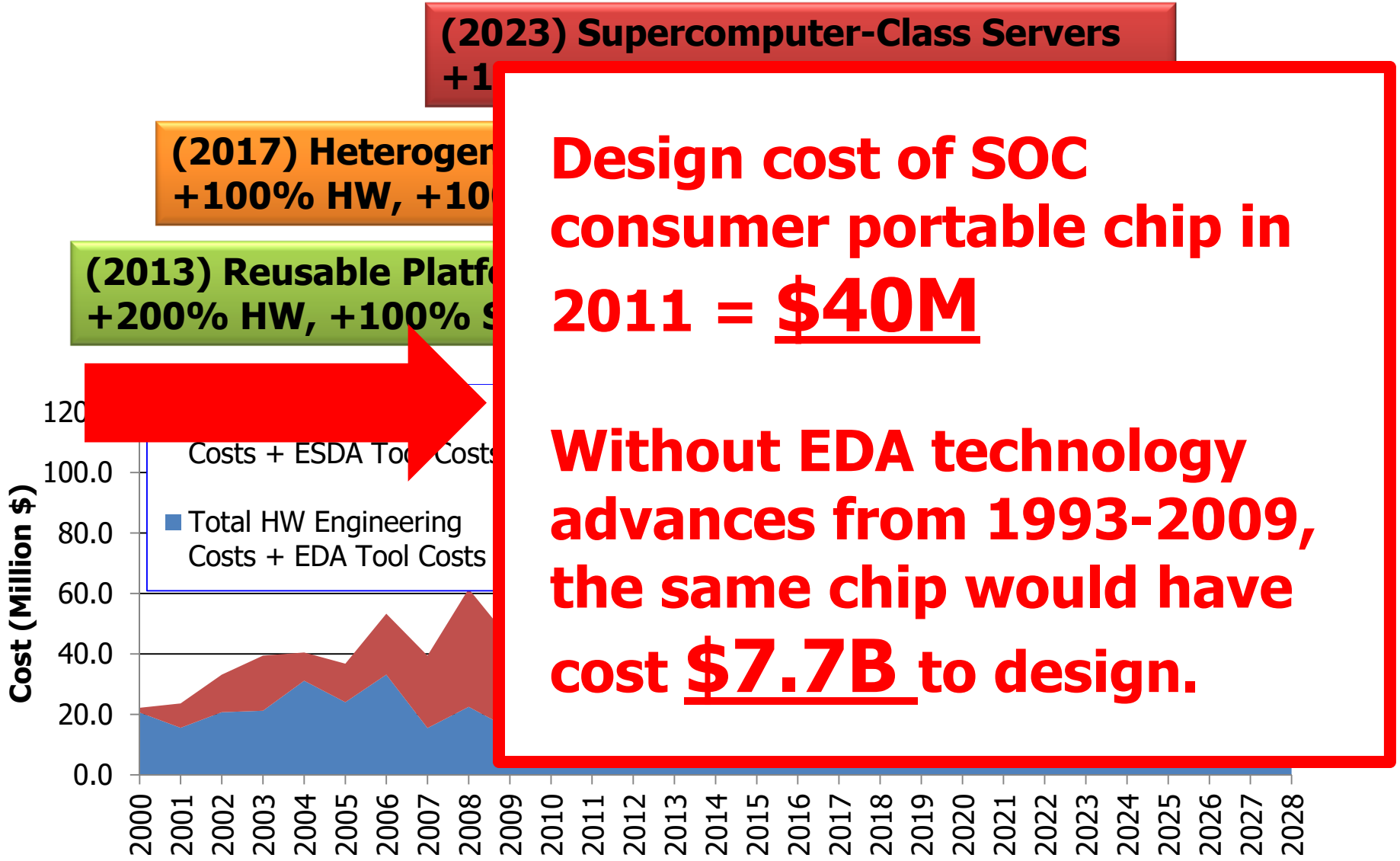
Design cost of SOC consumer portable chip in 2011 = $40M

Without EDA technology advances from 1993-2009, the same chip would have cost $7.7B to design.

A. B. Kahng IWLS 2017 keynote, 170618

EDA Impact on IC Design Cost

ITRS 2011, DAC13



Design cost of SOC consumer portable chip in 2011 = $40M

Without EDA technology advances from 1993-2009, the same chip would have cost $7.7B to design.

A. B. Kahng IWLS 2017 keynote, 170618

Outline

- Background: Evolving EDA Beyond its E-Roots
- First CEDA “DA Futures” Workshop (October 2016)
- Food for Thought

Outline

- Prelude: Looking Backward
- Background: Evolving EDA Beyond its E-Roots
- First CEDA "DA Futures" Workshop (October 2016)
- Food for Thought

"Those who cannot remember the past are condemned to repeat it." ☺

Looking Backward: EDA and the Roadmap

DAC13

- **2001: "Cost of design is the greatest threat to continuation of the semiconductor roadmap"**
- **Design technology innovations must keep on schedule to contain design costs, power**

A. B. Kahng IWLS 2017 keynote, 170618

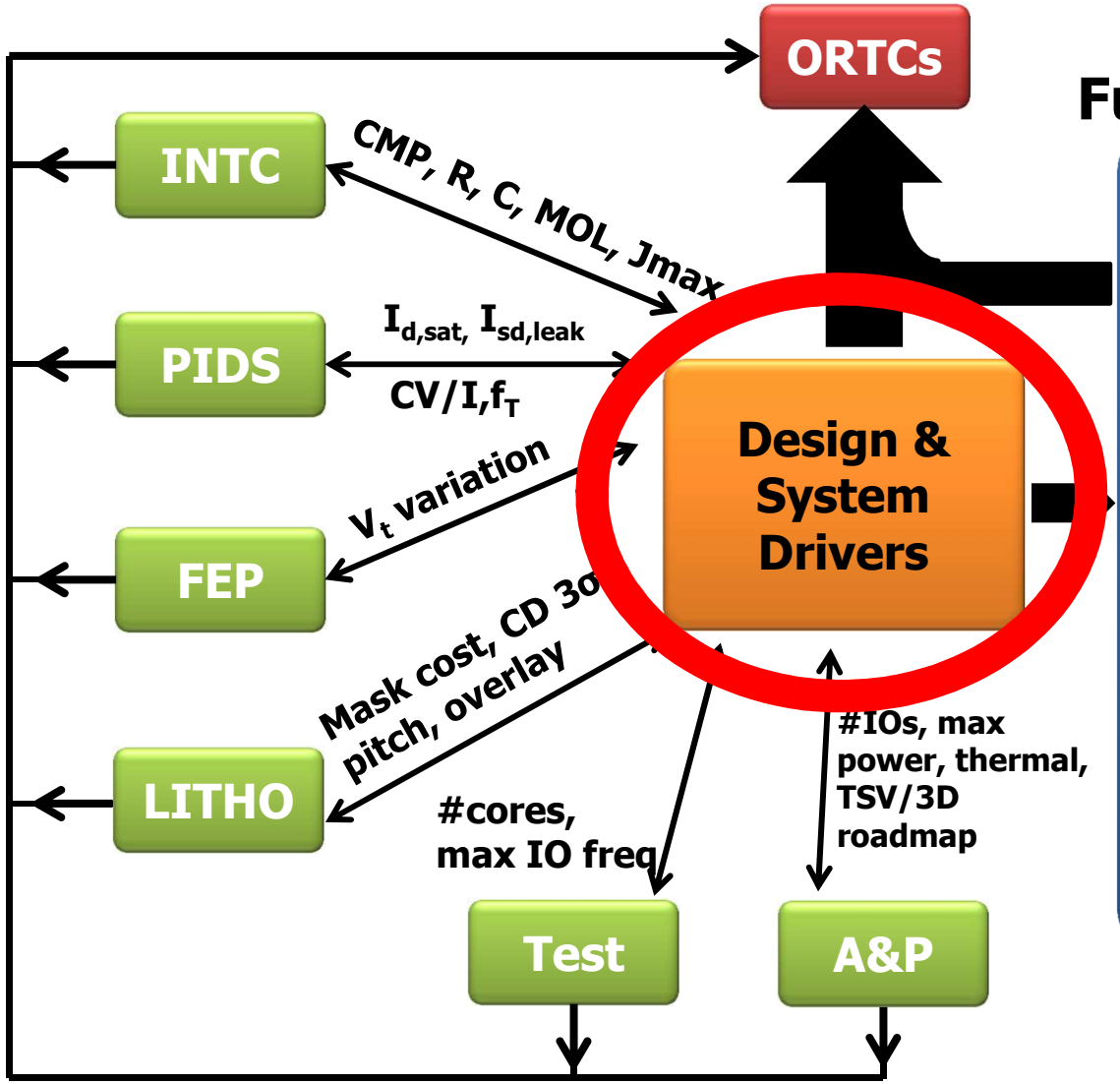
Looking Backward: EDA and the Roadmap

DAC13

- 2001: "Cost of design is the greatest threat to continuation of the semiconductor roadmap"
- Design technology innovations must keep on schedule to contain design costs, power
- 2007: System-level techniques are ultimately crucial to managing power
- 2009: Software and system-level design productivity are critical challenges
- 2009: Design-based equivalent scaling is essential to continuation of Moore's Law

EDA and Design are Central !

DAC13



Fundamental Models

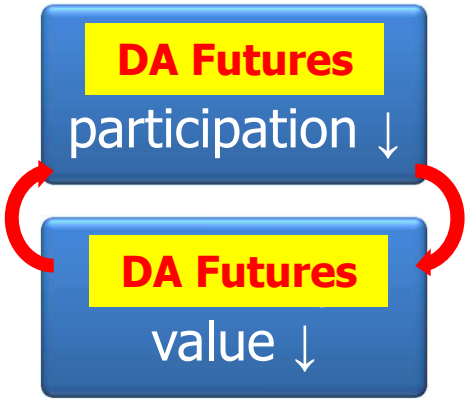
- max chip power
- layout density
- defect density
- transistor count
- chip size
- #distinct cores
- #cores
- IO speed
- max on-chip freq
- 3D TSV parasitics
- product/market drivers

A. B. Kahng IWLS 2017 keynote, 170618

Today: No (ITRS) Roadmap (Is "Futures" ~ "Roadmap"?)

DAC13

- **Fewer resources, wider scope**
  - Roadmap is not anyone's day job…
  - New scope: MEMS, More Than Moore, 3DIC, …
- **Risk of a "vicious cycle"**
  - Oligopolistic EDA industry
  - Disaggregation and consolidation in industry
  - Unwillingness to share "competitive" data



- **Explosion of post-CMOS, post-optical technology options**
  - Need better communication, synergy across
    - supplier industries
    - design-manufacturing
    - academia-industry

Outline

- Background: Evolving EDA Beyond its E-Roots
- First CEDA “DA Futures” Workshop (October 2016)
- Food for Thought

A Starting Point

- Electronic design automation (EDA)
  - Engineering success story
  - One of the first truly interdisciplinary fields
  - Has arguably focused on supporting “More Moore”
  - Has matured as semiconductor industry has also matured
  - Has seen diminished interest among ECE and CS students

A. B. Kahng IWLS 2017 keynote, 170618

Strongly Related (and a Plug ☺)

TUESDAY June 20, 4:30pm - 6:00pm | 12AB

TRACK: EDA, DESIGN

TOPIC AREA: GENERAL INTEREST, DIGITAL DESIGN, TEST & VERIFICATION

EVENT TYPE: PANEL

SESSION 30

COWS, COWBIRDS AND THE CONUNDRUM OF EDA RESEARCH FUNDING

Moderator: Andrew Kahng - *Univ. of California, San Diego, La Jolla, CA*

A perennial lament is that EDA companies don't fund academic EDA research at levels seen in other semiconductor supply chain industries. Perhaps, as the saying goes, "If you get free milk, why buy the cow?" But are EDA cows (research programs and their trained graduates) now scarcer? Starving? Migrating to greener, "DA Futures" fields? Today's graduates in EDA are hired by fewer EDA companies, who serve fewer customers, who have fewer internal EDA organizations. Research activity is a function of the number of faculty in the field, along with the amount of research funding. Research itself is slowed by limited access to tools, data, and problem formulations – while funding is limited by the years (from "DAC paper" to "production tool flow") needed to pay off the underlying investment. This panel brings together senior executives to discuss questions at the heart of these concerns: (1) Can the future bring a renaissance of EDA research funding? (2) For what kinds of research and results? (3) How must attitudes and actions change across researchers, industry, and government / consortia? (4) Who will ultimately contribute to such a renaissance, and how?

Panelists:
Mamta Bansal - *Qualcomm Technologies, Inc., San Diego, CA*
Tim Cheng - *Hong Kong Univ. of Science and Technology, Hong Kong*
Leon Stok - *IBM Systems and Technology Group, Yorktown Heights, NY*
Andreas Olofsson - *Defense Advanced Research Projects Agency, Arlington, VA*
Shishpal Rawat - *Council on EDA, Folsom, CA*

A Starting Point

- Electronic design automation (EDA)
  - Engineering success story
  - One of the first truly interdisciplinary fields
  - Has arguably focused on supporting “More Moore”
  - Has matured as semiconductor industry has also matured
  - Has seen diminished interest among ECE and CS students

- Urgent to revisit how the EDA field will evolve, grow
- Healthy growth requires exciting new directions as well as a steady supply of new experts trained at the graduate level

“Failing to Plan is Planning to Fail” ☺

"DA Perspectives and Futures"

- **In ~2013, IEEE Council on Electronic Design Automation (CEDA) began an initiative …**

- How can EDA paradigms and methodologies be leveraged for DA in other, emerging domains to solve impactful, real-world problems?
- EDA researchers actively contribute to DA in other fields …
- … but evolution and growth as a community requires a more systematic, coherent effort – as well as vision
- **Driving Question: How can EDA systematically "move beyond its E-roots"?**

CEDA's Initiative

- Two seeds
  - CCC workshops on extreme-scale design automation
  - CANDE (Computer-Aided Network DEsign) committee (1972-)
- Redirection of CANDE, focusing on new initiatives
- **Group I: Systemization of prior efforts**
  - F. Koushanfar, G. Qu, Z. Zhang. Summary/analysis of recent roadmaps; NSF EDA Expedition grants; SRC focused research centers
- **Group II: Metrics of DA research impact**
  - A. B. Kahng, G.-J. Nam, D. Pan. Analysis of research outputs, leveraged funding, industry and publication data with modern text mining
- **Group III: Vision for the EDA field.**
  - D. Chen, P. Kalla, S. Mitra, S. Levitan, M. Potkonjak. New opportunities inspired by technologies and applications, and real-world problems
- DAC-2015 DA Perspective Challenge
- ICCAD-2015 Evolving EDA Beyond Its E-Roots session, papers
- 2016 DA Futures Workshop

DA Perspective Challenge 2015 (1/2)

- What is the long-term problem?
- Why is the problem important and challenging?
- What is the state-of-the-art?
- What is the problem's relevance to existing DA tools and methods? How can DA help in addressing the challenges?
- What knowledge, skills, and/or tools are needed to address the problem?
- Is the problem interdisciplinary, requiring expertise other than DA?
- What are the broader impacts?

http://vlsicad.ucsd.edu/DAPerspectiveChallenge/

DA Perspective Challenge 2015 (1.5 / 2)

Food For Thought:

What would you expect from such a challenge?

- **Adjacencies: "DA for X" == "XDA"**
- **What else?**

DA Perspective Challenge 2015 (2/2)

- 30 submissions, 13 invited to workshop @DAC'15
- 3-minute talks to panel of 6 academic/industry judges
- 1st prize: Huang/Cheng (UCSB), **DA of flexible electronics**
- 2nd prize: Chen/Li (Penn State), **DA for neuromorphic**
- 3rd prize: Chang (KAIST), **Energy optimization for EVs**
- Audience favorite: (U of Calgary) **DA of energy systems**
- Honorable mentions: (IBM Research) **DA for networks of autonomous vehicles**; (MIT) **DA for trusted hardware**
- **Did we miss anything?** ☺

ICCAD-2015: Evolving Beyond E-Roots

MONDAY, NOVEMBER 2

Special Session 3D - From EDA to DA: Can We Evolve Beyond Our E-Roots?

Time: 4:00pm - 5:30pm | Room: Austin

Moderator:

Deming Chen - *Univ. of Illinois at Urbana-Champaign*

Organizers:

Andrew B. Kahng - *Univ. of California at San Diego*

Farinaz Koushanfar - *Rice Univ.*

As we approach a 2020 "wall" of silicon, patterning, device, interconnect and cost limits, and as EDA technologies and the industry itself remain in their present "mature" state, it is now more urgent to revisit question of how the field will evolve and grow. In particular, how can the paradigms and research methodologies of EDA be leveraged for design automation (DA) in other, emerging domains? Some researchers in our community are actively contributing to DA for a variety of other fields (e.g., emerging nanotechnologies, biomedical and security). But, evolution and growth as a community requires a much more systematic, coherent effort – as well as forward-looking vision to steer by. This special session is built upon and follows up on the seeds planted by the CCC Extreme Scale Design Automation workshop series. It focuses on recent efforts to systematize knowledge, trends, metrics, and visions that can help enable the EDA community to move beyond its E-roots. The audience will learn about a new IEEE CEDA technical activity group dedicated to this subject. The session includes four 15-minute talks, covering the overview and study teams within the technical activity group, as well as a 30-minute panel summarizing the recent DA challenge at DAC'15.

A. B. Kahng IWLS 2017 keynote, 170618

Research gap analysis (1/4)

- SRC studies to determine magnitude of research needs and world-wide "research gap"
  - Research needs derived from ITRS technology requirements in 2001 and 2003 studies
  - Massive funding gaps identified

Estimated Worldwide annual research investment to support 2008-14 needs

Science Area	Number of Tasks		Needs ($M)	
	2001	2003	2001	2003
Front-End Processing	22	19	150	513
Proc Integr, Dev & Struct	15	14	280	410
Patterning	25	22	245	338
Interconnects	13	13	153	244
Design CAD	26	31	280	398
Circuit Des & Sys Arch	20	40	168	359
ES&H	5	5	30	30
Assembly & Packaging	–	13	–	224
Factory Integration	5	5	100	155
Total	131	152	**1406**	**2671**

A. B. Kahng IWLS 2017 keynote, 170618

Research gap analysis (2/4)

- Circuit design needs (2003 study by ICSS)
  - Notable increase in circuit design needs

Year 2003		
Task #	Task Description	Quanta
1	Circuits that comprehend, manage and possibly leverage the effects of large sub-threshold source-drain leakage currents and large gate leakage currents	2
2	Exploit RF potential for aggressively scaled digital CMOS	2
3	Low-cost circuit and process techniques for improving soft-error immunity in sub-100 nm Leff technology generation	2
4	Novel device structures that permit complex circuit operations at the device level	3
5	Circuits that handle the increasing statistical variability of smaller numbers of dopant atoms and patterning at the diffraction limits of light	4
6	Innovative methods of reshaping state-of-the-art front-end analog processing for improving the robustness of subsequent digital processing	3
7	Circuit techniques that comprehend upgradability and reconfigurability	4
8	Novel signaling for low cost, low power, high bandwidth chip-to-chip communication	2
9	CMOS RF (1-10GHz) building blocks	1
10	Methods to include the package and PCB in the intrinsic design (5-20GHz)	2
11	Advanced power distribution and noise abatement techniques	2
12	Alternatives to aggressive voltage scaling to minimize power (e.g., noise-tolerant, ultra-low-swing logic schemes, adiabatic circuit styles, and on-chip magnetics)	2
13	Electrostatic Discharge (ESD) protection, device modeling and design	1
14	Circuit research that is closely coupled to architectural innovations for developing low power scalable digital systems	1.5
15	Circuits on non-silicon based technologies	2
16	Analog/mixed signal/RF built-in self test	2
17	High performance, low power data converters	2
	Total Research Needs	37.5

A. B. Kahng IWLS 2017 keynote, 170618

Research gap analysis (3/4)

- System architecture needs (2003 study)
  - Massive new tasks to address

Year 2003		
Task #	Task Description	Quanta
1	Application analysis for architectural requirements	3.6
2	Application mapping for nuclear spin state and/or Tunneling Phase Logic (TPL) devices	2.4
3	Efficient space-time processing communication algorithms and radio architectures	1.2
4	Energy efficient communications with MIMO	0.6
5	Energy efficient computing in leakage dominated regimes	0.9
6	On-chip multiprocessing architectures	2.4
7	Memory-based computing architectures; support for molecular computing	3
8	Defect tolerant architectures; support for CNT, nano-wires; dependability, availability and fault tolerance to unreliable devices	6
9	Architectural support for 3D integration	1.2
10	Biologically inspired computing architectures	1.2
11	Architectural advances to incorporate intelligence in silicon systems	3.6
12	Semantic preservation and architectural interface in compilation tools	1.2
13	Novel OS/RT architectures that enable incremental composition	4.5
14	Validation in compilation (proof-carrying compilation and assume guarantee frameworks)	1.2
15	Novel OS services for energy, precision/fidelity, and distribution	1.5
16	Application specific OS services	1.5
17	Heterogeneity: hybrid systems model and analysis	1.2
18	Quality of service models and optimizations	1.2
19	Correctness guarantees: composability, compositionality in SIP	2.4
20	Miniaturization of novel sensing techniques	3.6
21	Miniaturization of biological/chemical processing	3.6
22	Microfluidic processing integration (simulation, verification)	3
23	Bio/chemical processing in microelectronic systems: handling of biological samples and processing	1.2
	Total Research Needs	52.2

Research gap analysis (4/4)

- CAD and Test needs (2003 study)
  - Dominating factors in research needs

Year 2003		
Task #	Task Description	Quanta
1	Electronic system-level design methodology	2
2	Language-level modeling advances and composition interfaces	3
3	Architectural and microarchitectural design exploration	3
4	Design methods for embedded software	6
5	CAD infrastructure for rapid prototyping and evaluation of design tools (like MOSIS)	1
6	Interoperability standards for design technology	3
7	Collaborative, geographically distributed design methodologies	2
8	Assess/solve quality and yield impact due to test equipment limits (e.g., utility of at-speed test)	1.5
9	Fault modeling for defects/noise in nanometer technologies	2
10	Diagnosis and failure analysis for analog/mixed-signal parts	3
11	Self-test solutions and test methods for (heterogeneous) core-based SOC (including RF, analog, MEMs, etc.)	4
12	Intelligent testbench	1.5
13	Energy/power management at system, architecture, and microarchitecure levels	5
14	Power reduction/management at logic/physical/timing level	5
15	Power management addressing circuit-level issues and manufacturing variability	5
16	Unification of design-mask and design-silicon flows (RET, mask data prep/write/inspect, prep for future advanced patterning solutions)	4
17	Design under high manufacturing variability and parameter uncertainty	5
18	Cost-driven design that is aware of manufacturability/yield	4
19	Core-based design verification	2
20	Verification of heterogeneous systems	4
21	Verification for novel devices and high-frequency novel interconnects	2
22	Equivalence checking between design levels	2
23	Convergent, predictable physical implementation methodologies (e.g., one-pass design with incremental/partial specification)	5
24	Design with partial or probabilistic information	3
25	Core optimization algorithms and techniques	3
26	Reliability with permanent faults (reconfigurability, graceful degradation)	4
27	Design techniques to address SEU (soft errors)	1
28	Design with novel (future) devices	4
29	Analysis/optimization for system-in-package and 3D integration	3
30	Global (multi-disciplinary) solution of ITRS red bricks (Design-LITHO, Design-PIDS, Design-A&P, etc.)	2
31	Modeling and simulation tools for new devices	2.5
	Total Research Needs	99.5

Outline

- Background: Evolving EDA Beyond its E-Roots
- First CEDA “DA Futures” Workshop (October 2016)
- Food for Thought

CEDA's DA Futures Workshop Vision

- **Mission**
  - **Develop and maintain the premier inventory of future application areas for design automation theory and technologies, along with their current status**
- **Goals**
  - **Enable long-term, sustained focus on evolving DA field beyond its "classical" topics and scope**
  - **Address possible structural inefficiencies in**
    - Interactions among academia, industry and governmental entities
    - Interactions between design automation and the applications – systems – integration – technology stack, that waste time and resources today
  - **Bring together a critical mass of world-wide thought leaders**

The First DA Futures Workshop (Oct. 2016)

- **"DA for emerging memory, memory-based computing, and neuromorphic computing systems"**
- **Goal #1: Understand future DA (research) challenges**
- **Goal #2: Understand "research enablement" to enable timely solutions to challenges**
- **Archival workshop report** (4 Ph.D. student scribes), post-workshop survey, (breakout sessions, dinner 'exercise') …
  - Er… 20+ pages of notes still require final editing

DA Futures Workshop Program

- *Keynote I: TrueNorth Ecosystem for Brain-inspired Computing, Pallab Datta, IBM*
- *Session I: Memory-Driven Neuromorphic Computing*
- **Keynote II: Problems in Biology for which EDA Infrastructure is Potentially Useful, Lou Scheffer, HHMI**
- **Session II: Processor-Memory Integration for Energy-Intermittent Processing**
- *Keynote III: Memory-Driven Computing Architecture, Shivani Raghav, HPE*
- **Session III: New Computation Models**
- **Keynote IV: Design Automation Challenges of Google's TPU, Richard Ho, Google**
- **Session IV: EDA Infrastructure**

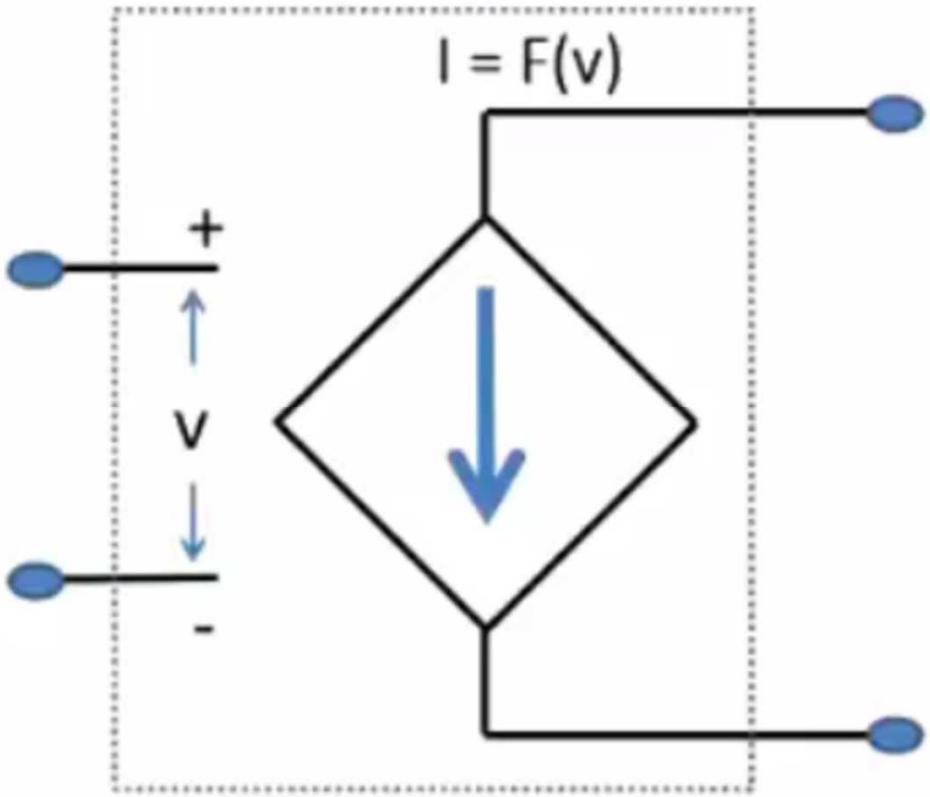
KEYNOTE II:
"PROBLEMS IN BIOLOGY FOR WHICH EDA INFRASTRUCTURE IS POTENTIALLY USEFUL"

LOU SCHEFFER, HHMI

[Scheffer]
A. B. Kahng IWLS 2017 keynote, 170618

Motivating Observations

- Many Biology problems involve chemical and electrical reaction
- E.g., neuron operation modeled essentially as voltage-controlled current source
  - Positive feedback non-linearity – “action potential” or spike



A. B. Kahng IWLS 2017 keynote, 170618

EDA and Biology Simulation: Commonalities

- IC simulation is similar to Biology simulation
  - Integration methods
  - Timestamp control
  - Accuracy control
- Biology simulation is still in time domain
- Many popular analyses in IC are missing
  - Sensitivity, frequency response, stability, noise analysis
- Connection with other simulators

Mixed-Level Simulation Requirement

- Multi-mode simulation
  - Simulation environment can be water, air, etc
- Multi-object simulation
  - Cooperate with simulators for muscle, hormones, cellular networks, etc.
- Languages for describing networks and interaction
  - Biology has no Verilog, VHDL, etc.
- APIs for allowing simulations to communicate

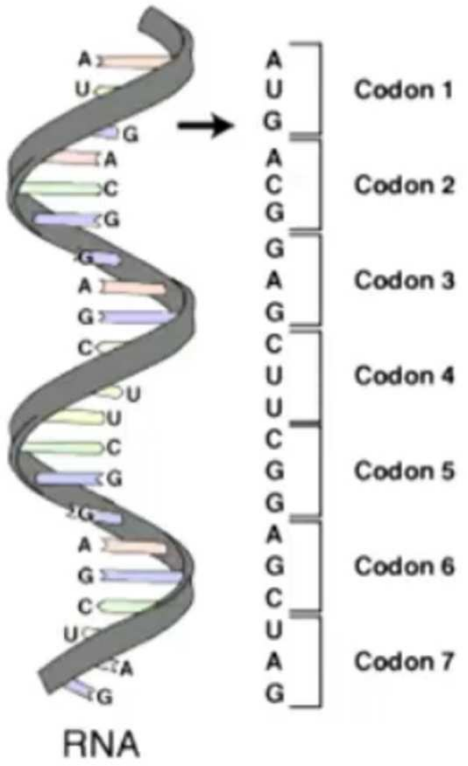
Temperature Corner Analogy

- IC functionality, in most cases, does not change over a temperature range
  - But, power and performance vary over temp range
- Similarly, cold-blooded animals work over a range of 10-40°C with largely unchanged behavior
  - But, reaction rates can change considerably over the temperature range
- Not well understood how this works !

A. B. Kahng IWLS 2017 keynote, 170618

Codons and Logic Synthesis

- Proteins are constructed based on base triples
  - Each has 4 possibilities (A, C, G, U)
  - 64 possible combinations
  - Only need to specify 20 different amino acids
- Each amino acid can be represented by one or more codons
- Like logic synthesis
  - Many ways to generate the same result



Takeaways From Keynote II (Lou Scheffer)

- EDA has lots of experience optimizing large systems with complex rules and conflicting constraints
- Similarly there are three kinds of biological network
  - Metabolic, Gene regulation and neural
- One known set of constructions, with rules (DNA encoding)

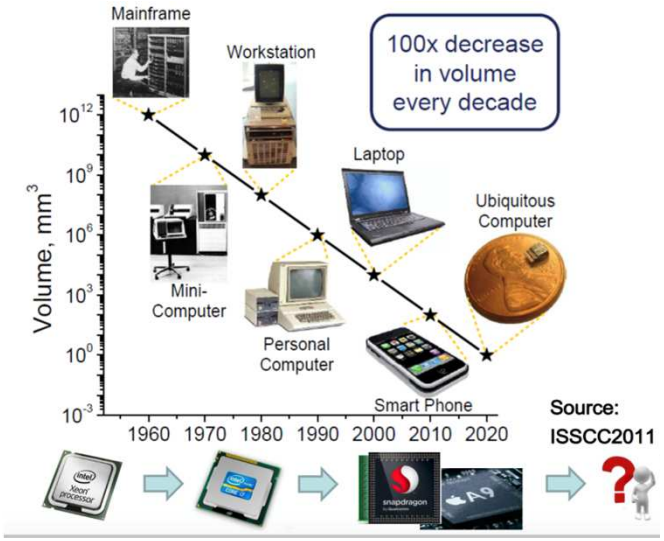
- Many possibilities of transfer from EDA (!!!)
- (Number of potential Ph.D. theses: Large ☺)

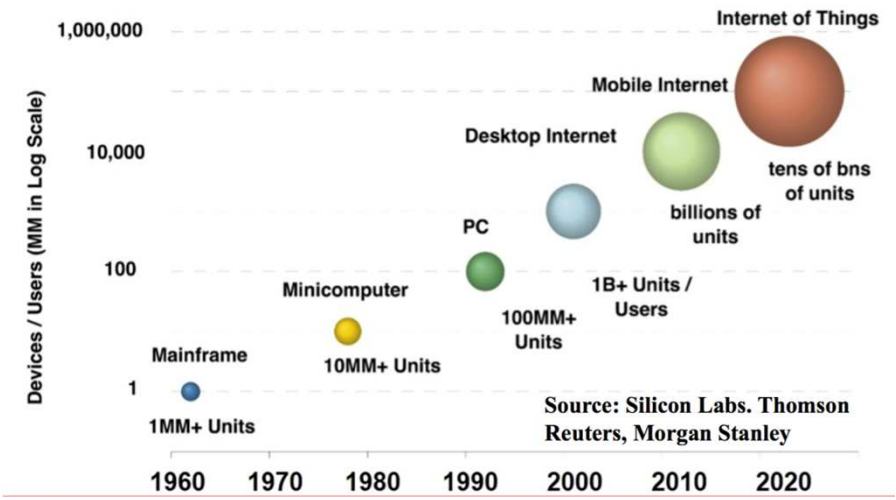
Session II: Processor-Memory Integration for Energy-Intermittent Processing

- [Zhang] Spintronics for Low-power Memory and Computing
  - New devices and library abstractions for emerging technologies enable low-power storage and computing
- [Liu] Design Automation for Ambient Energy Harvesting Nonvolatile Processor
- [Narayanan] Design Modeling & Architectures Using Novel Memory Devices
  - Challenge with new devices in emerging technologies (NCFET, PhaseFET, FeFET): theoretical models are not universal (depends on fabrication) – even as modeling and verification stacks are essential to evaluation

Design Automation for Ambient Energy Harvesting Nonvolatile Processor

- Electronic devices are becoming smaller and smaller
- Number of devices is increasing exponentially
- Big gap between battery capacity and power consumption
- Solution
  - Energy Harvesting
    - 2-3 Magnitudes Gap Generator and Consumer
    - Unstable Power Sources
    - Hard to Predict the future harvested energy
  - Low-power circuits



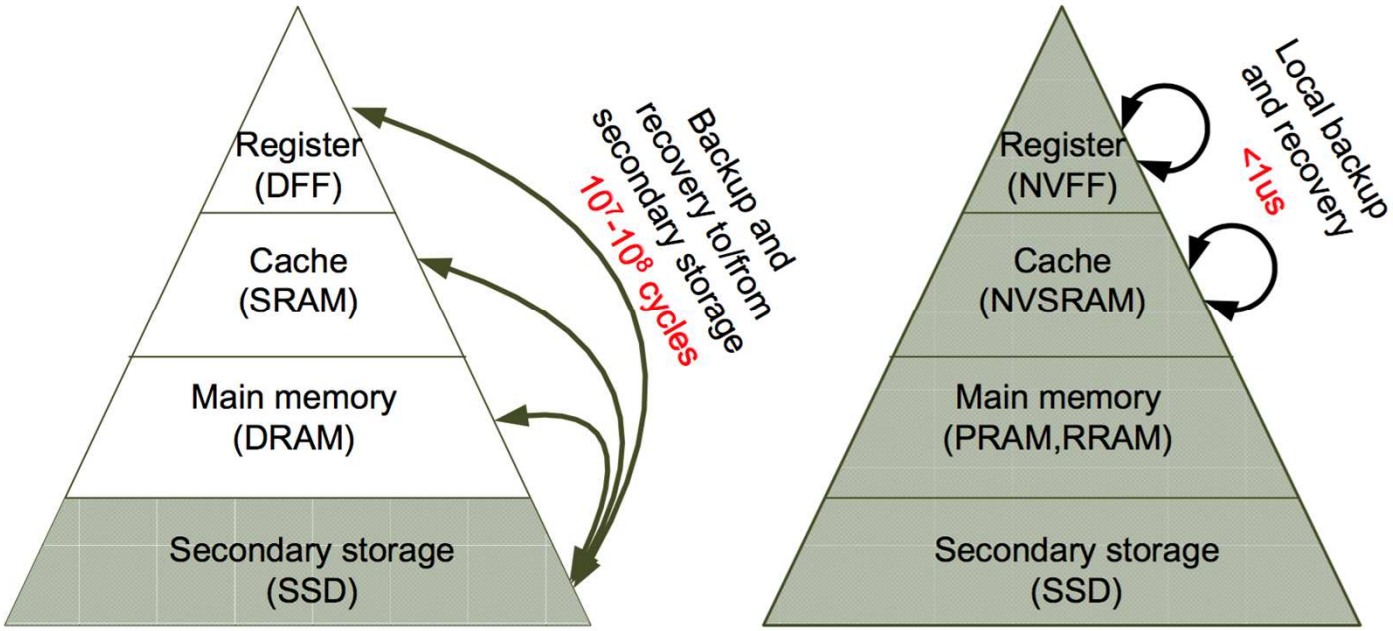


[Liu]

A. B. Kahng IWLS 2017 keynote, 170618

Design Automation for Ambient Energy Harvesting Nonvolatile Processor

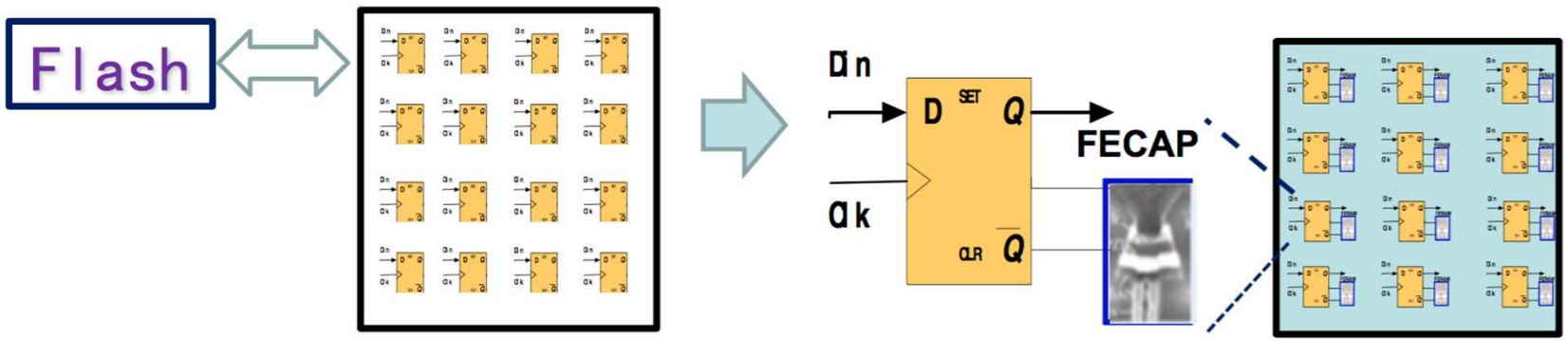
- Nonvolatile Processor, a reliable solution?
  - In place backup by adopting nonvolatile memory techniques
  - Much faster (<1us) than conventional backup and recovery (10^7-10^8 cycles)



A. B. Kahng IWLS 2017 keynote, 170618

Design Automation for Ambient Energy Harvesting Nonvolatile Processor

- Conventional Processor
  - Global Data Movement
  - Limited Memory Bandwidth
- Nonvolatile Processor
  - Local Data Movement
  - Flip-flop Level Parallelism



A. B. Kahng IWLS 2017 keynote, 170618

Takeaways: Yongpan Liu, Tsinghua Univ.

- Deploying energy harvesting is challenging
  - Limited power output
  - Unstable power
  - Hard to Predict
- Nonvolatile processor design
  - Circuit level (NVFF, NVSRAM)
    - Ferroelectric FF (FeFF)
    - NVSRAM
  - Architectural level

Design Modeling & Architectures Using Novel Memory Devices

- Need good design automation mechanisms to have a “smooth shift”: good models, validation technologies, and benchmarking
- Important to know where we are in an emerging technology shift
- Challenge with emerging technology: models work in theory, not necessarily for two different device fabricators
  - Fabrication affects model
- Must consider memory not only as storage element but also as computing element

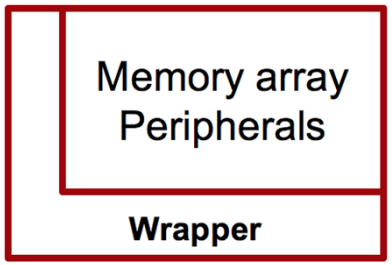
[Narayanan]
A. B. Kahng IWLS 2017 keynote, 170618

Session III: New Computation Models

- [Olshausen] Neuromorphic Architectures for Visual Perception and Scene Analysis
  - Encoding and storage of analog information in emerging memory
- [Gaillardon] Majority-based Synthesis for RRAM-based In-memory Computing
- [Datta] Future Emerging Devices & Computing Models
  - Phase transition materials with nonlinear threshold response and built-in memory can enable new classes of circuits and systems
- [Rowen] Cognitive Computing & Impact on Applications and Silicon
- [Salahuddin] Towards Energy-efficient Learning Machines (ENIGMA)
  - Efficient mapping of sparse matrix information enables synthesis of large-scale networks

Majority-based Synthesis for RRAM-based in-Memory Computing

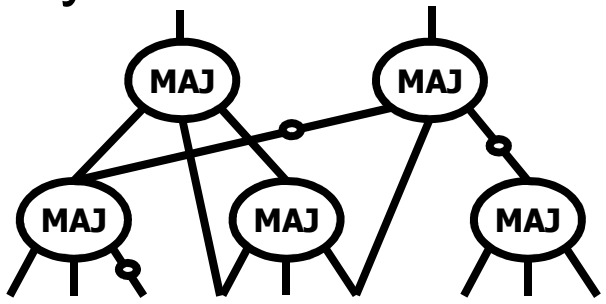
- **RRAM devices act as MAJ operators**
- **Open questions:**
  - How to deploy the complex signals involved with MAJ operation?
  - How to implement a dual-memory select with common ground?
  - How to schedule operations?
- **Programmable Logic-in-Memory (PLiM) paradigm: How to provide additional computing capabilities to regular memory arrays?**
  - Minimize controller overhead at cost of lower performance
    → *Architectural aspects*
  - Control of operations and scheduling
    → *CAD aspects*



[Gaillardon]
A. B. Kahng IWLS 2017 keynote, 170618

Majority-based Synthesis for RRAM-based in-Memory Computing

- **Computing with RRAMs goes beyond Logic Implication**
- MAJ operation done with a unique memory and simpler control signals
- The *PLiM* approach allows for:
  - Large memory arrays to natively perform small footprint processing
  - Memory manufacturers to add this functionality at very low area overhead
- The *PLiM* architecture is a holistic approach that considers:
  - Operations natively made within RRAMs
  - Control logic
  - Mapping and synthesis
- Low-cost distributed memory systems capable of operating on their own content may be the basis for new computing paradigms



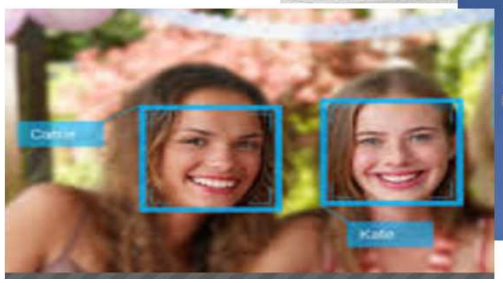
A. B. Kahng IWLS 2017 keynote, 170618

Takeaways: Pierre-Emmanuel Gaillardon, Utah

- Majority-oriented logic can be used to build any logic function
- Can be implemented by using binary nonvolatile memory, including cross-bar memories
- The combination of memory and logic is a key to breaking past current bottlenecks in computing
- Logic Synthesis may evolve towards more expressive primitives (MAJ-based Logic) and be more “technology-aware” or “-sensitive”
  - New DA techniques
  - New computing paradigms

Cognitive Computing and the Impact Upward on Applications and Downward on Silicon

- New HW & SW systems using neural-inspired methods to do human-like recognition and reasoning tasks
- Deep Learning Challenges
  - Need too much memory bandwidth
  - Need too much compute
  - Not accurate enough
  - Too much latency

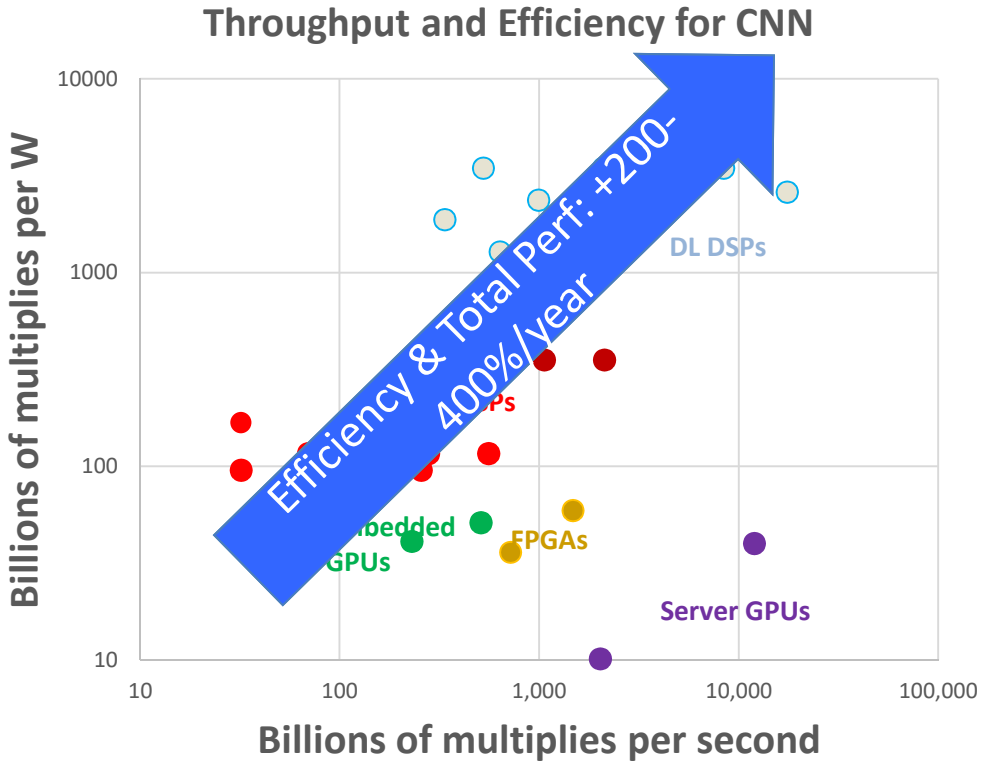


- AlexNet Example
  - ~60M model parameters (FP32: 240MB)
  - ~725M multiply accumulates (MACs) per image
  - At 1000 images/sec: 240GB/s DDR bandwidth (FP32)

[Rowen]
A. B. Kahng IWLS 2017 keynote, 170618

Cognitive Computing and the Impact Upward on Applications and Downward on Silicon

- **Silicon Implications of Deep Learning**
  - Applications demand extreme compute at low energy
  - Embedded implementation evolving rapidly towards specialized deep learning engines: up to 100x energy gain
  - Network structure synthesis to gain additional 5-10x in compute efficiency
  - Deep Learning Units become strategic building blocks – peers to CPU, GPU, media processors and memory



Source: Cadence Design Systems

A. B. Kahng IWLS 2017 keynote, 170618

Takeaways: Chris Rowen, Cognite Ventures

- Cognitive computing application is “right now”, and will be increasingly popular for handling recognition and reasoning tasks
- Cognitive computing brings challenging requirements for compute efficiency and performance
- ASICs will keep delivering increasing compute power as well as compute efficiency

A. B. Kahng IWLS 2017 keynote, 170618

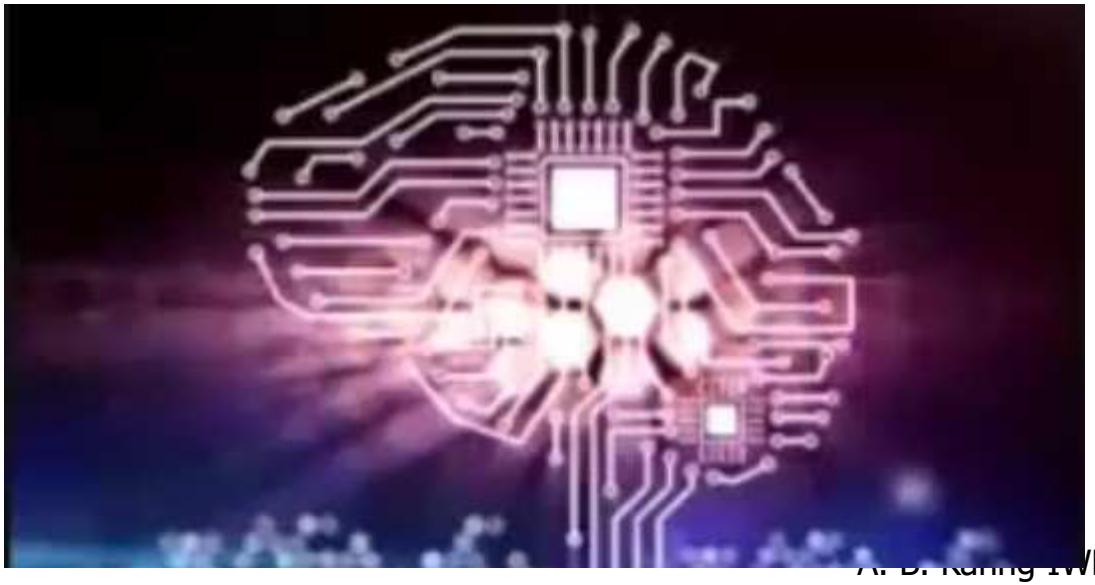
KEYNOTE IV: DESIGN AUTOMATION CHALLENGES OF GOOGLE'S TENSOR PROCESSING UNIT

RICHARD HO, GOOGLE

[Ho]
A. B. Kahng IWLS 2017 keynote, 170618

Overview and Motivation

- Exploding needs for Neural Networks
  - Growing uses of deep neural networks in many applications
  - Growing number of applications moving to DNNs
  - Growing popularity of certain apps:
- Slowing of Moore's Law
  - Custom accelerators can provide multiple factors improvement in performance/watt and operations/sec in equivalent silicon area



A. B. Kahng IWLS 2017 keynote, 170618

Google's Tensor Processing Unit

- Custom ASIC accelerator for machine learning in Google data centers since 2015
- 10X (7 years) of Moore's Law improvement for machine learning workloads
- Very fast design cycle for ASIC and system
- Rapid bringup and deployment

A. B. Kahng IWLS 2017 keynote, 170618

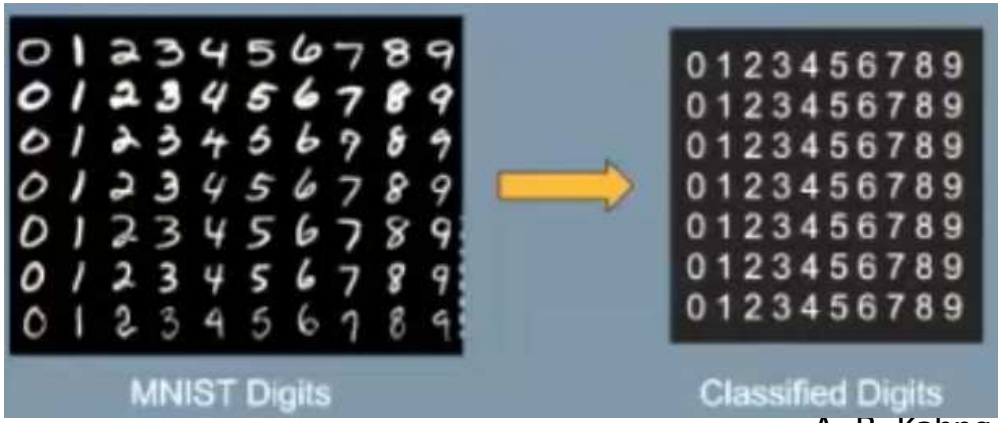
HW/SW Co-Design

- Accelerator hardware must take advantage of workload knowledge
- ASIC development needs greater agility
  - Adapting to changes in specification
  - Adapting to inner loop optimizations
  - Enhancements for performance bottlenecks
- Co-design and tradeoffs made across the full stack
  - API layer (e.g., TensorFlow)
  - System/Drive software
  - Firmware
  - Circuits

A. B. Kahng IWLS 2017 keynote, 170618

Unit Tests / Verification

- ASIC development needs to borrow this discipline from software development
- Unit tests guide implementation
  - Test-driven development
- Verification starts with unit tests
- End-to-end functional correctness
  - Integration early
  - Simulation early



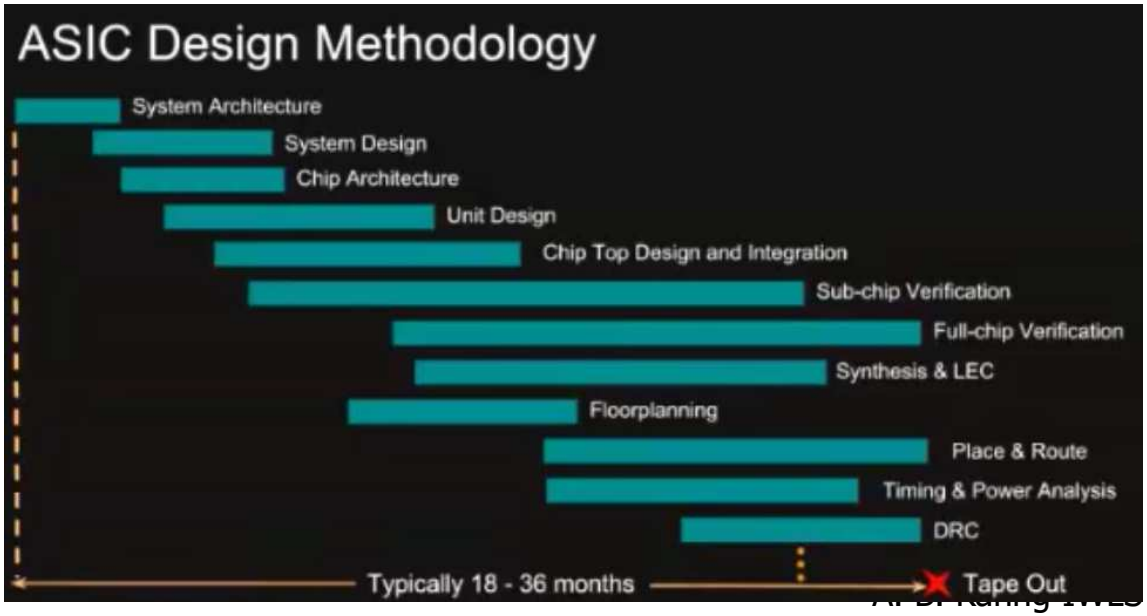
A. B. Kahng IWLS 2017 keynote, 170618

Development Flow Performance

- Time to complete a process = fn (iterations / day)
- Emulation and FPGA prototyping with system software development and verification
- Flexible infrastructure
  - Elastic compute, memory and storage resources
- Flexible EDA tools and licensing model
- Fast and frequent physical design loops
  - Estimation, measurement, refinement
- Balancing critical path

Takeaways From Keynote IV (Richard Ho)

- More agile design methodology
  - Power, performance, area estimates early and often
  - Physical-aware architecture and microarchitecture design
  - Parallel work on functional verification and physical design
  - Faster design and verification through aggressive reuse and unit testing
  - Balanced critical paths to reach aggressive schedule goals



2017 keynote, 170618

Session IV: EDA Infrastructure

- [Ghosh] EDA in Post Von-Neumann Era
- [Mitra] The N3XT 1,000X
- [Narayanan] A Multi-Layer Perspective on Design Challenges for Non-Volatile Memory-based Neuromorphic Systems
  - Potentially 25X speedup and 2580X lower power with better devices, scalable modeling and reconfigurable interconnect
- [Groeneveld] Future EDA "Infrastructure" Industrial Perspective

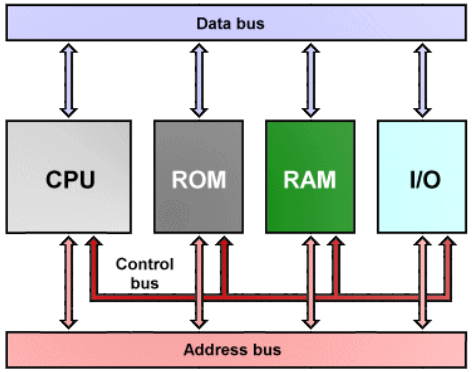
EDA in Post-Von Neumann Era

- **Neuromorphic ICs are in their early stage**
  - Notable difference compared to today's ICs
  - No mature design flow exists
- **Valuable EDA design paradigms today**
  - Separation of design creation and design verification
  - Hierarchical abstraction
  - Design for testing
  - Separation between design and manufacturing
- **Big question: What can be migrated from today's EDA to neuromorphic IC era?**

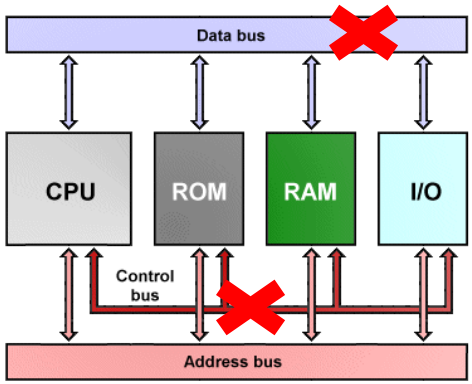
[Ghosh]
A. B. Kahng IWLS 2017 keynote, 170618

Von Neumann vs. Post Von Neumann

- Von Neumann Architecture: Turing machine
  - Turning Machine – A mathematical model of a hypothetical computing machine that can use a predefined set of rules to determine a result from a set of input variables



Von-Neumann Architecture



Post Von-Neumann Architecture

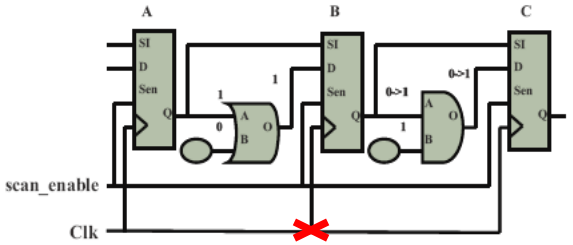
- Few characteristics of post von-Neumann Turing Architecture (early neuromorphic)
  - Asynchronous – helps in power saving as no clock
  - Compute and data are intermingled – no bus
  - Interconnect re-configurability – to emulate synapse

Source : http://www.technologyuk.net/computing/computer-systems/architecture.shtml

A. B. Kahng IWLS 2017 keynote, 170618

Impacts of No Clock to DFT and EDA Design

- How will DFT be done without clock?
  - What will be the equivalent of scan FF to keep state?
    - How will that impact the DFT methodology?
- Impact to Verification Flow
  - How will the chip be simulated at RTL level?
  - What will be equivalent of Gate simulation and Static Timing Analysis?



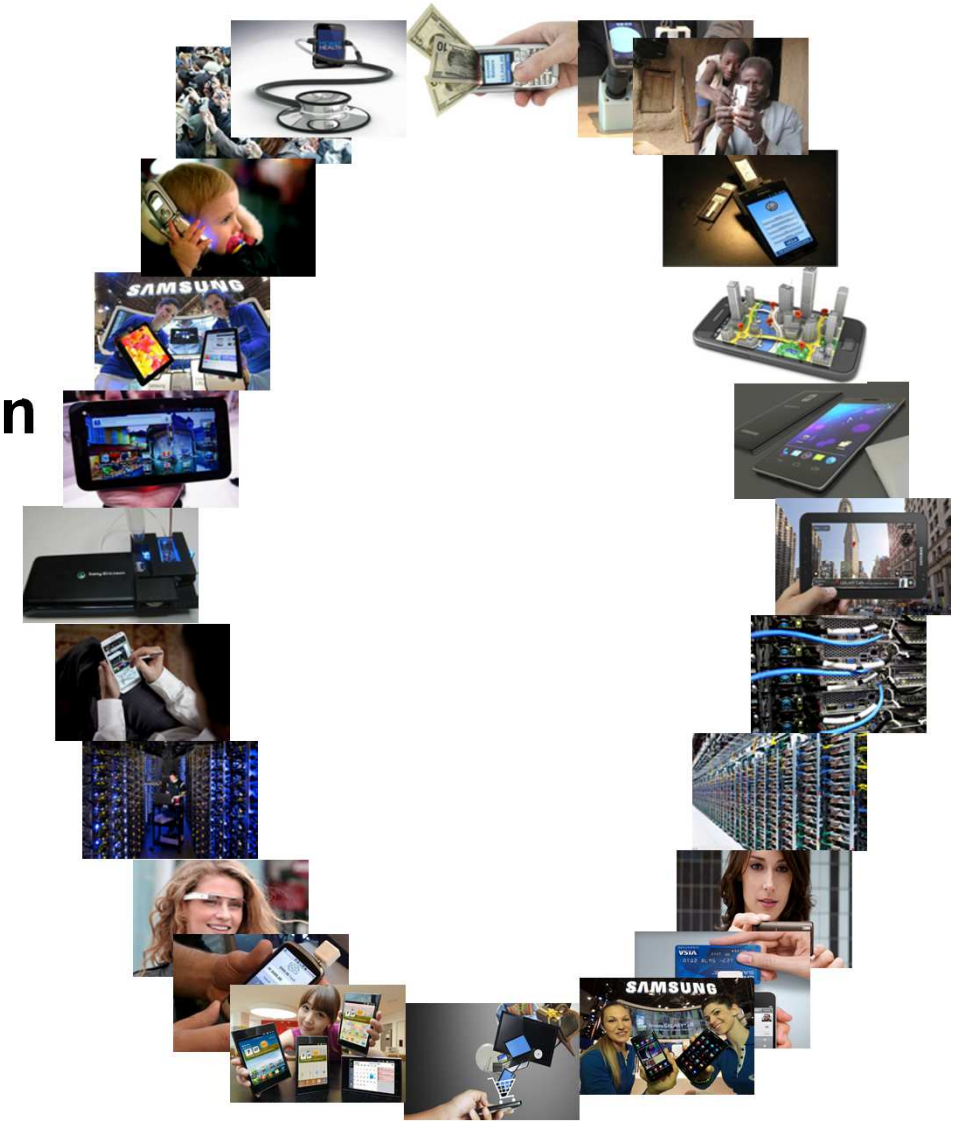
- Impact of asynchronous computing to separation of design and manufacturing paradigm
  - What is the value of extraction check?
    - Will asynchronous computing require more precise RC extraction?
    - Alternatively, will training-based programing lead to less precision of RCX?
  - What are impacts on retargeting?
    - Is the electrical impact of the process side retargeting more visible due to lack of clock-based design?
    - Do we need to make process side retargeting visible in PDK?

Takeaways: Pradiptya Ghosh, Mentor

- Build MCNC-like post-Von Neumann benchmarks to drive DA research
- Study whether DFT paradigm will need re-evaluation
- Develop design automation flows to exploit asynchronous computing on these chips
- Develop tools that fit into existing DA paradigms, flows / methodologies

Robust Nanosystems From Today to the N3XT 1,000X

- **World relies on electronics**
  - **Ensure robust operation**
  - **Enhance user experience**
  - **Expand application horizon**
- **Many walls, simultaneously**
  - **Power wall**
  - **Memory wall**
  - **Complexity wall**
  - **Resilience wall**
  - **Interconnect wall**
  - **Cooling wall**

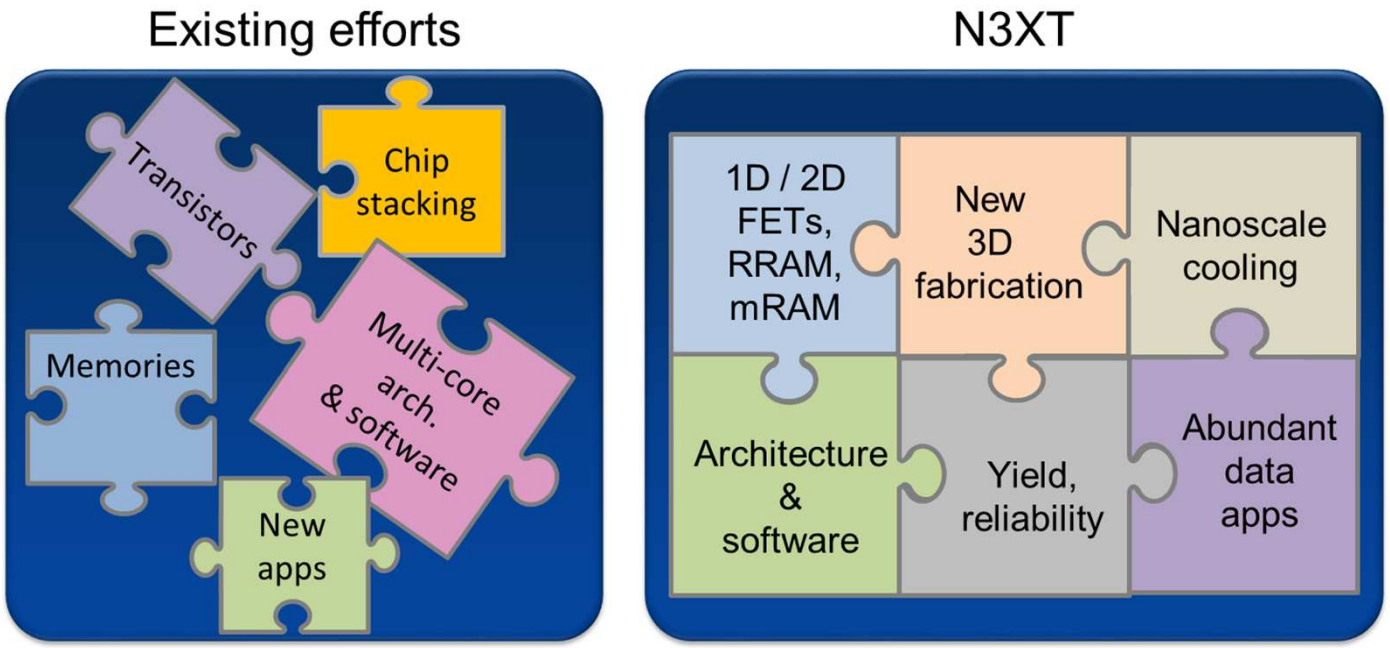


[Mitra]

A. B. Kahng IWLS 2017 keynote, 170618

Robust Nanosystems Today

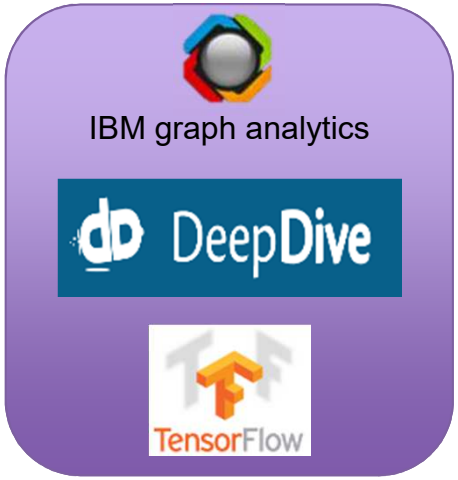
- New solutions – Elegantly simple and effective
  - Carbon Nanotube FET (CNFET)
    - ~10X Energy Delay Product (EDP) benefit
  - 3D Integration
    - From TSV to Nano-scale ILV → denser vertical interconnect
- Isolated improvements inadequate



A. B. Kahng IWLS 2017 keynote, 170618

Robust Nanosystems Tomorrow

- N3XT Framework
  - Heterogeneous nanotechnologies
  - Architecture design space
  - Physical design
  - Thermal, yield, variability



- ~1,000X benefits, software programmable

Energy	Execution time	Energy × Execution time
37x	**23x**	**851x**

A. B. Kahng IWLS 2017 keynote, 170618

Takeaways: Subhasish Mitra, Stanford

- Many unexploited opportunities to apply:
  - Specialization
  - Neuro-inspired
  - Technology innovations
- To achieve next 1,000X improvement:
  - Densely interwoven compute + memory + …
  - New FETs, RRAM/MRAM, monolithic 3DV, nano-cooling
  - And, we already have these building blocks (!)
  - *[Implicit: Realizing the 1000X is a good target for Future DA]*

Future EDA "Infrastructure": Industrial Perspective

- **Algorithmic Synthesis/Verification methodology**
  - Main challenge: managing scale and complexity
  - Strong abstraction levels with hierarchy
  - Design-agnostic
- **Iterative design facilitation**
  - Analysis tools for Analog design, Mechanical design
- **Physics**
  - TCAD Devices, device models, mask, optical
- **Design / IP**
  - Low-level building blocks: standard cells, memories
  - High-level: processors, USB, Neuromorphic, Cognitive computing

[Groeneveld]
A. B. Kahng IWLS 2017 keynote, 170618

Takeaways: Patrick Groeneveld, Synopsys

- **Moore's law with CMOS to stay for 10+ years**
  - Incremental adaptations of EDA method will do
- **New technologies and challenges to EDA**
  - Neuromorphic computing
  - Carbon nanotube based design
- **Perpetual "tough nuts" to crack**
  - **Parallel processing (incompatible with synthesis)**
  - **Automation of un-partitionable designs type (e.g., analog)**
  - **Automation of heterogeneous design (algorithms cannot handle)**
  - **Automation at architectural level (very application specific synthesis)**

A. B. Kahng IWLS 2017 keynote, 170618

ABK addendum: “Even Simple SP&R…”

- At what corner should synthesis be run so that the final post-P&R result has best {P, P, A} ?
- Why does MCMM synthesis lead to final post-P&R results that are worse than single-corner synthesis followed by “close the other corner”?
- Which corner should be the synthesis corner, and which corner should be “the other corner”?
- How can I predict final utilization from a post-synthesis netlist?
- How can I predict post-route DRVs from global/trial route?
- **These are questions that very strong chip implementation teams cannot find answers for !!!**

Outline

- Background: Evolving EDA Beyond its E-Roots
- First CEDA “DA Futures” Workshop (October 2016)
- Food for Thought

Recap …

- Growing EDA beyond its “E-roots” requires systematic, community-wide, strategic thinking and vision
- CEDA initiative: three study groups
  - Research gap analysis
  - DA Perspective Challenge 2015
  - ICCAD 2015
  - First DA Futures Workshop: emerging memory, memory-based computing, neuromorphic computing

Recap … And so?

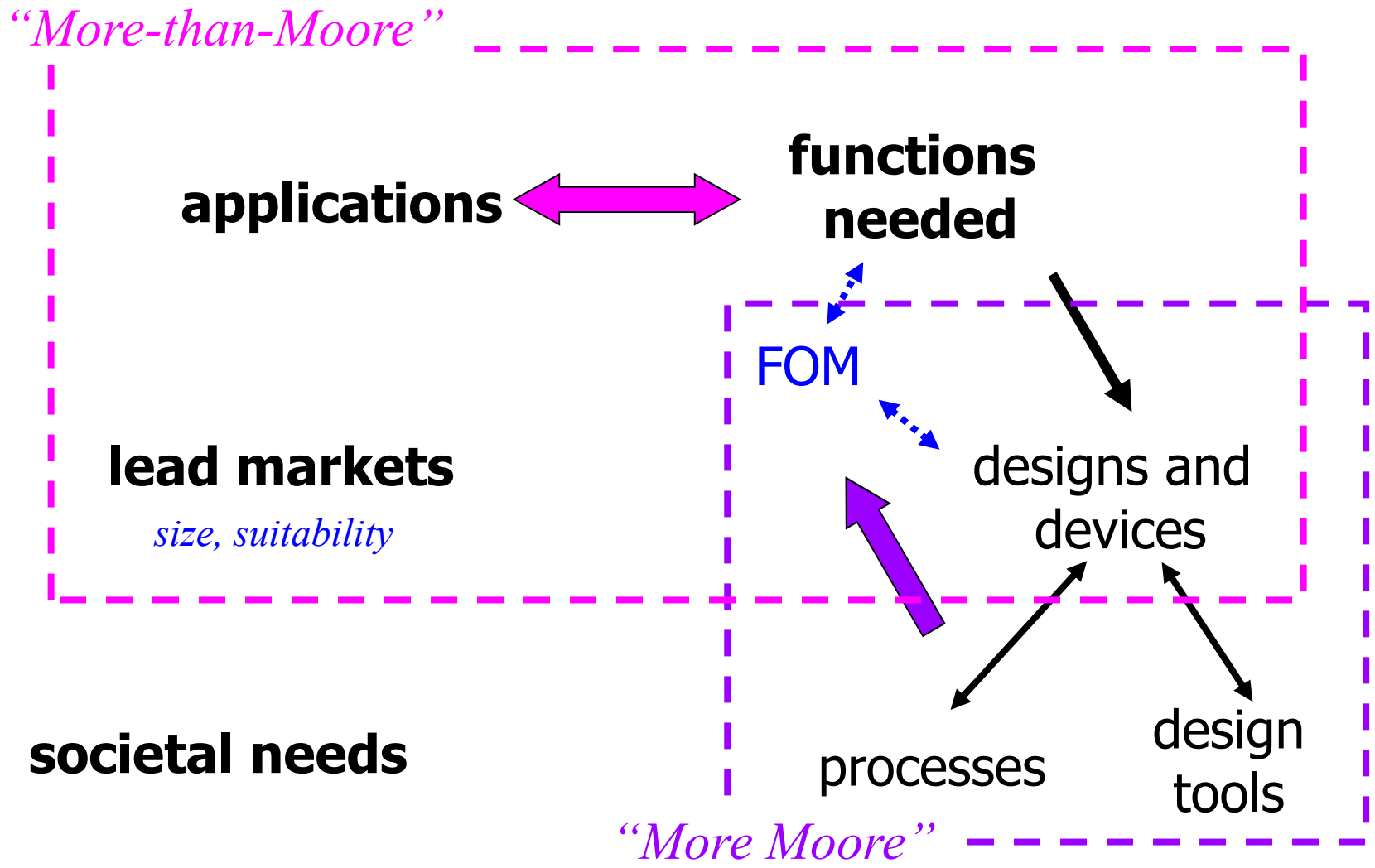
- Growing EDA beyond its "E-roots" requires systematic, community-wide, strategic thinking and vision
- CEDA initiative: three study groups
  - Research gap analysis
  - DA Perspective Challenge 2015
  - ICCAD 2015
  - First DA Futures Workshop: emerging memory, memory-based computing, neuromorphic computing
- **Second DA Futures Workshop = ? Third = ?**
- **The But-For question?**
- **How to learn from the past?**

Types of Futures for DA

- **Adjacencies: DA for X ("XDA")**
  - Deployment, system connection of new devices and tech
  - **X = technologies** (ML, new memory architectures, ANN ICs)
  - **X = application domains** (EVs, flexible electronics)
  - "DA for X": similar goals to DA for IC (power, TAT, …)
- **Parallel Universes: X =** "biological systems", … (movie scripts? presidential campaigns?)

ITRS ~2010

App–Func–Tech Interplay = Long Known



A. B. Kahng IWLS 2017 keynote, 170618

"Futures" ~ "Roadmap" ?

- **Beware of** casual, "let it come" approaches to identifying futures for your field !
- At some point, "futures" must have trajectories

"Roads? Where we're going, we don't need … roads!" ☺

A. B. Kahng IWLS 2017 keynote, 170618

Types of Futures for DA

- **Adjacencies: DA for X ("XDA")**
  - Deployment, system connection of new devices and tech
  - **X = technologies** (ML, new memory architectures, ANN ICs)
  - **X = application domains** (EVs, flexible electronics)
  - "DA for X": similar goals to DA for IC (power, TAT, …)
- **Parallel Universes: X =** "biological systems", … (movie scripts? systems of government? …)
- **Introspections: "DA for DA" ← let's not forget this!**
  - Predictability? QOR? Cloud / massive parallelism? Design- and technology-specific DA? Flows/methodologies?
    - ("First paper" vs. "Last paper" …)
  - Open-source / free tools? Learning-based tools and flows? New optimization foundations?
  - **Plenty left for "Design Automation" in the IC space, and for "Foundations of Design Automation" !!!**

More “Us” and “We” Needed

“We cannot predict the future, but we can invent it.” ☺

A. B. Kahng IWLS 2017 keynote, 170618

References

- A. B. Kahng and F. Koushanfar, "Evolving EDA Beyond its E-Roots: An Overview", (.pdf), (.pptx), *ICCAD*, 2015.
- A. B. Kahng, M. Luo, G.-J. Nam, S. Nath, D. Z. Pan and G. Robins, "Toward Metrics of Design Automation Research Impact", (.pdf), (.pptx), *ICCAD*, 2015.
- A. B. Kahng, "The ITRS Design Technology and System Drivers Roadmap: Process and Status", (.pdf), (.pptx), *DAC*, 2013
- DA-METRICS: http://vlsicad.ucsd.edu/DA-METRICS/
- EDA Roadmap Workshop, (link), Anaheim, CA, 2010
- DA Perspective Challenge, (link), San Francisco, CA, 2015
- A. B. Kahng, "IT and Society: How Far Into the Future Can We See?", (.pdf), Workshop on IT and Future Society, Jeju Island, Korea, 2011
- Design Automation Futures Workshop, (link), Fremont, CA, 2016

THANK YOU !

Acknowledgments: Many thanks to Mohammad Riazi, Lutong Wang, Bangqi Xu and Wei-Ting Jonas Chan for helping prepare slides for this talk. Thanks also to Farinaz Koushanfar, Shishpal Rawat and the IEEE CEDA leadership for their collaboration and support for the “DA Perspectives / Futures” efforts.