ModSim-2025

ModSim Challenges in Secure and Resilient AI (SARA) System Design

Pradip Bose

Distinguished Research Scientist and

Manager of Efficient and Resilient Systems

IBM Research

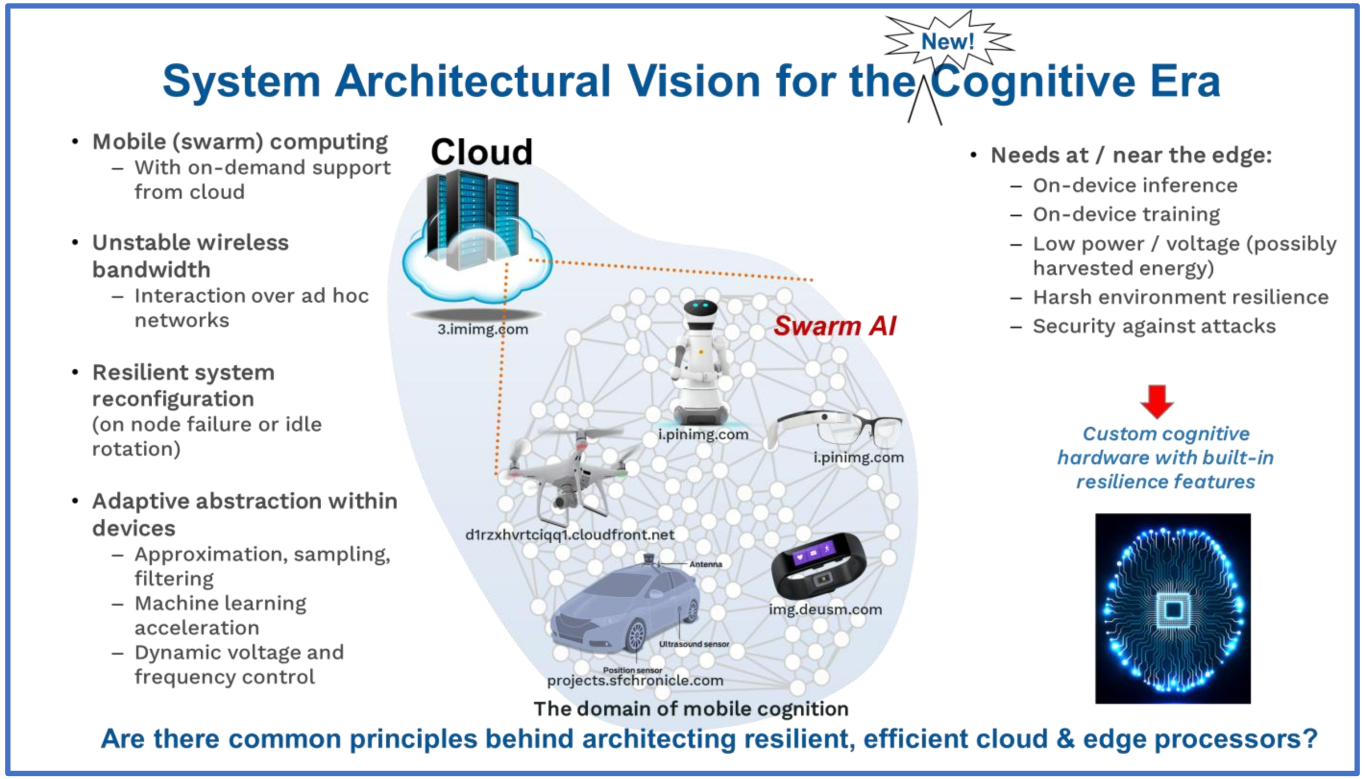
pbose@us.ibm.com

This research was developed in part with funding from the Defense Advanced Research Projects Agency (DARPA) and later from the DoD/RAMP-C program. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

DARPA-hard Challenges:

a good way of pushing the envelope in systems R&D

2011



New!

System Architectural Vision for the Cognitive Era

- **Mobile (swarm) computing**
  - With on-demand support from cloud
- **Unstable wireless bandwidth**
  - Interaction over ad hoc networks
- **Resilient system reconfiguration** (on node failure or idle rotation)
- **Adaptive abstraction within devices**
  - Approximation, sampling, filtering
  - Machine learning acceleration
  - Dynamic voltage and frequency control

The domain of mobile cognition

- Needs at / near the edge:
  - On-device inference
  - On-device training
  - Low power / voltage (possibly harvested energy)
  - Harsh environment resilience
  - Security against attacks

Custom cognitive hardware with built-in resilience features

Are there common principles behind architecting resilient, efficient cloud & edge processors?

Meanwhile, the modern age of AI had begun in 2011

- IBM Watson (Deep Q&A, Jeopardy champion)
- Siri (iPhone/Apple) edge NLP

- Agile SoC
- Programmability

- Data security
- Privacy

PERFECT [Bob Colwell, Joe Cross, …] →

2013 – 2018

Power Efficiency Revolution for Embedded Computing Technologies

1 GF/W → 75 GF/W

IBM + Stanford, Harvard, U of Virginia

DSSoC [Tom Rondeau] →

2018 → 2023/ongoing

Domain-Specific System on Chip

Power-perf, programmability, productivity metrics

IBM + Columbia, Harvard, UIUC

[Tom Rondeau]

DPRIVE

2021 → ongoing

(IBM was not part of DPRIVE; but we pursued the same goal, 2022-2025 w/support from DoD/RAMP-C), *IBM + Columbia*

Executive Summary of ModSim Challenges Faced

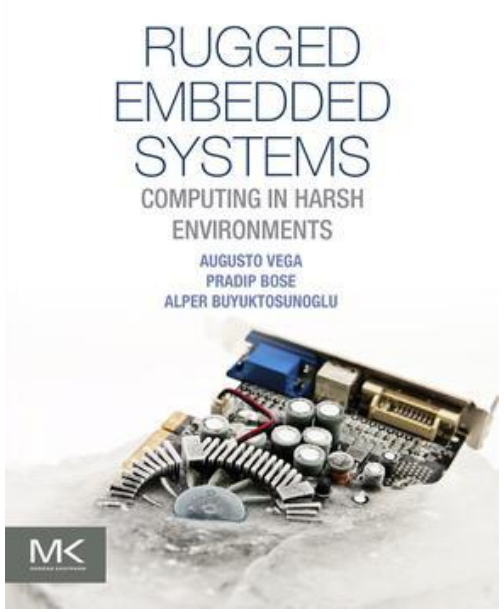
(across the three govt-sponsored R&D projects)

1. Design Verification (and Test!)
   - Architects woefully lack tools and metrics to gauge verification complexity in pre-silicon modeling
   - *Agile SoC design* claims avoid factoring in verification time
2. Robust Power Management
   - On-chip, workload-driven power management architectures have become increasingly more advanced and sophisticated
   - But…ModSim-driven reliability & security guarantees are lacking
3. Security Metrics and Pre-Silicon Modeling
   - Largely absent! (Urgent need)

Deficiencies above cause shortfalls in system resilience and inhibit product quality deployment of devised solutions

RESILIENCE

In machine terms, it roughly means *reliable operation under error-prone or harsh environments*

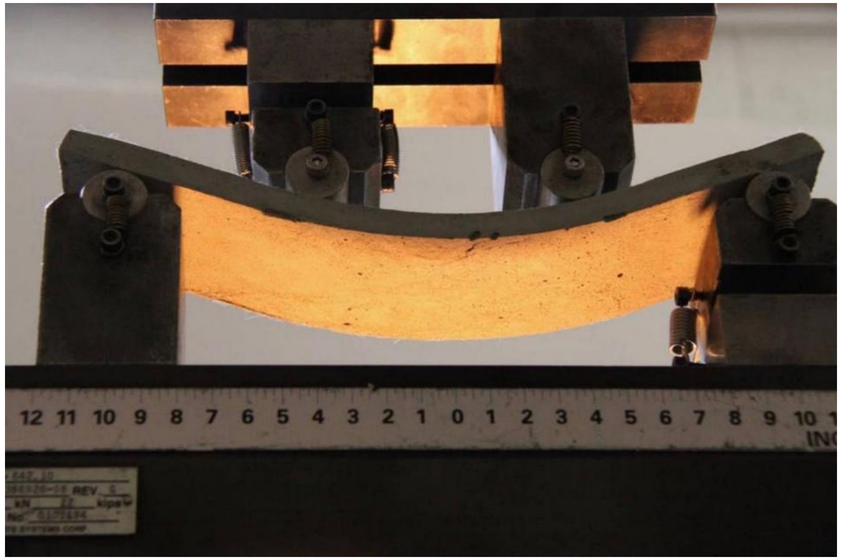


RUGGED EMBEDDED SYSTEMS

COMPUTING IN HARSH ENVIRONMENTS

AUGUSTO VEGA
PRADIP BOSE
ALPER BUYUKTOSUNOGLU

MK



In human (and perhaps AI?) terms, on the other hand….

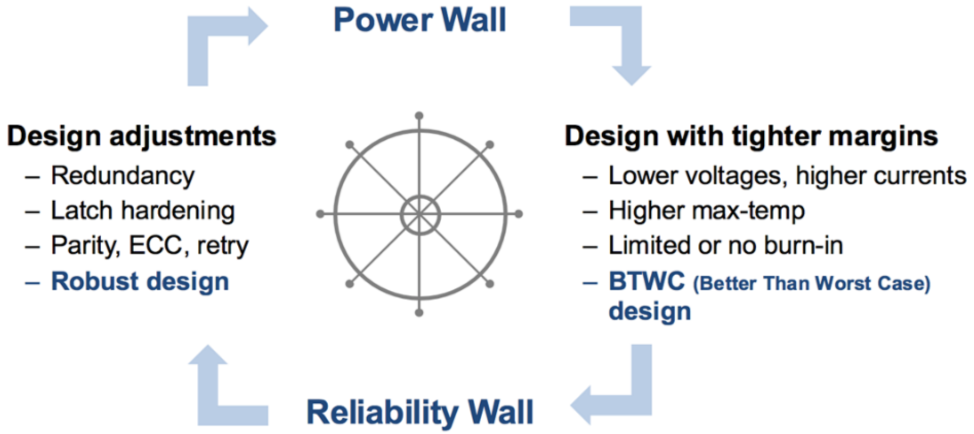
Resilience, a key component of emotional intelligence, is essentially the ability to “bounce back” from stressful experiences.

https://www.psychologytoday.com/us/blog/comfort-cravings/201308/getting-back-emotional-intelligence-and-resilience

What is ***Efficient* Resilience**?

- System design approach to improve **efficiency** with “guarantees” of operational **correctness or quality** for a given application workload (even under hostile circumstances)

PERFECT: achieve 75 GF/W without giving up resilience

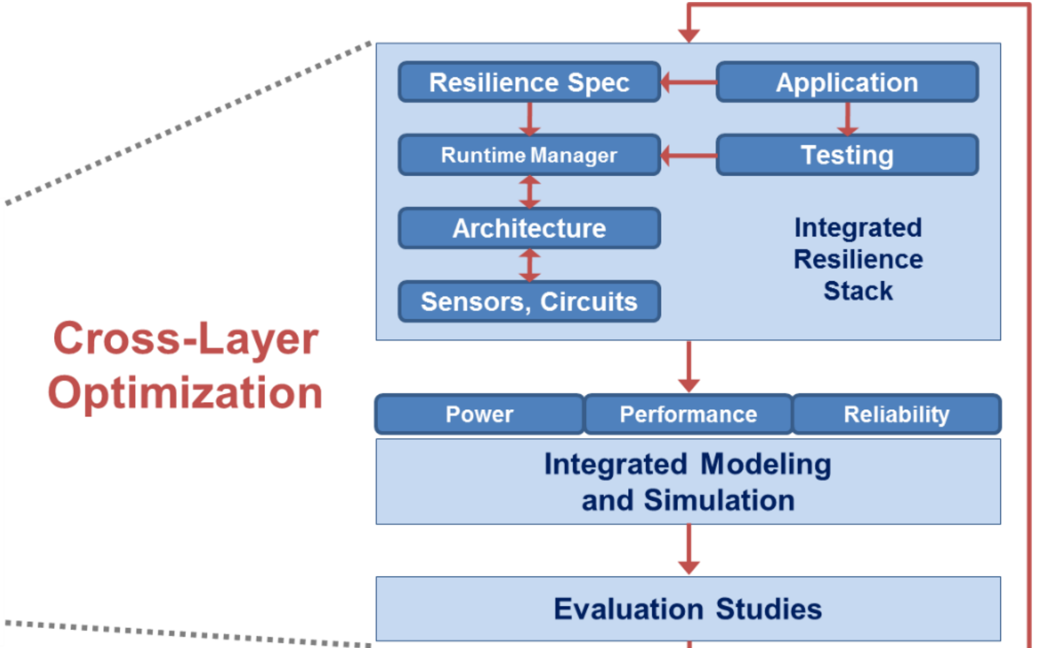


Power Wall

Design adjustments
- Redundancy
- Latch hardening
- Parity, ECC, retry
- **Robust design**

Design with tighter margins
- Lower voltages, higher currents
- Higher max-temp
- Limited or no burn-in
- **BTWC** (Better Than Worst Case) **design**

Reliability Wall



Cross-Layer Optimization

Resilience Spec, Application, Runtime Manager, Testing, Architecture, Sensors, Circuits

Integrated Resilience Stack

Power, Performance, Reliability

Integrated Modeling and Simulation

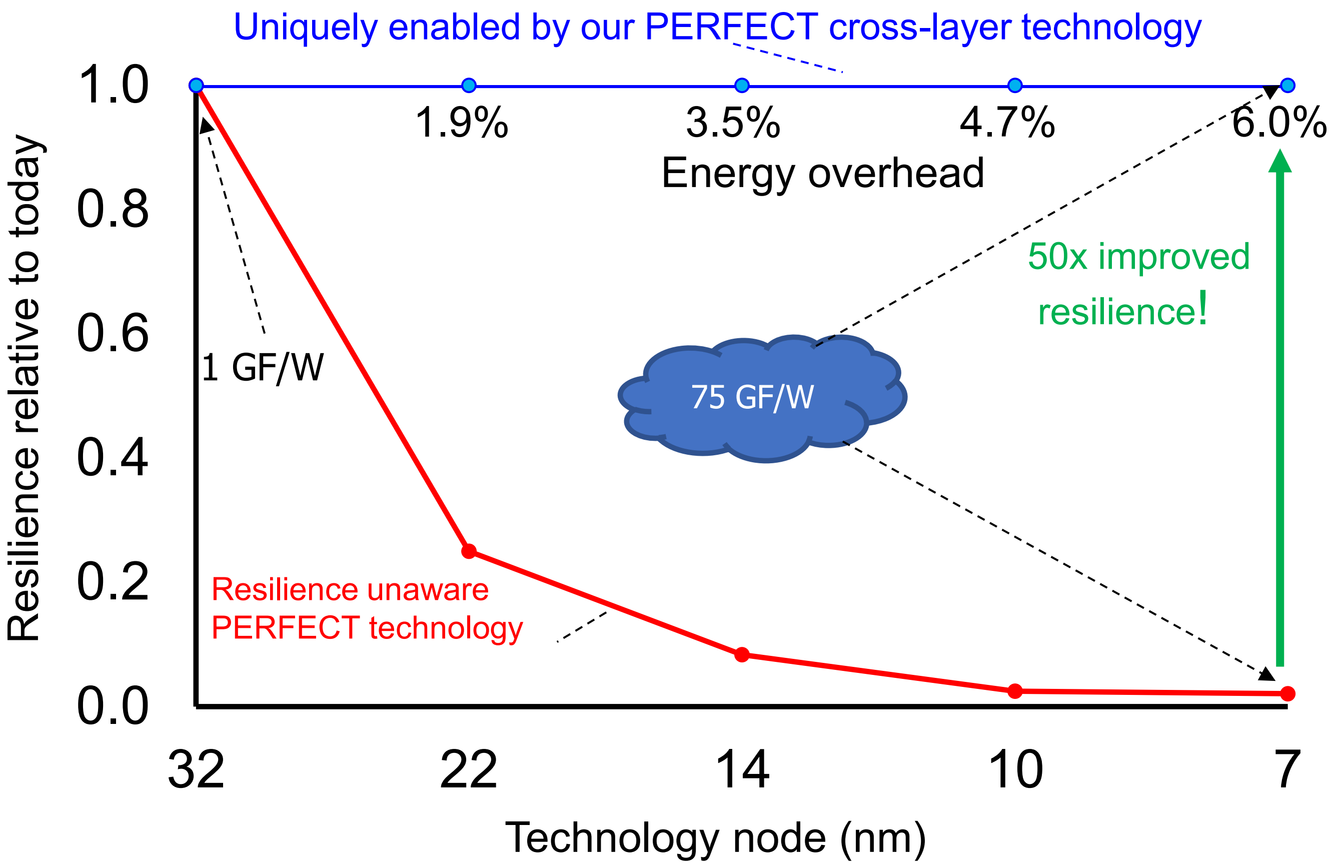
Evaluation Studies

- Security is another aspect of system resilience that has similar overhead-related concerns
- Resilience is part of basic functionality that customers expect to get for free!

Key enabler for mobile real-time cognitive computing

Applies to future server/mainframe, and supercomputing systems as well

The ModSim-Driven PERFECTion



- Experimental set-up used: a full-stack software-hardware system consisting of an FPGA implementation of an open-source processor (LEON3-OpenSparc) with matrix multiplication application
- Resilience improvement for current system, with our cross-layer technology was evaluated using **fault injection at the latch level**
- Cross-layer knobs used: Selective latch hardening (circuit), parity (logic/microarch), control/dataflow checking (microarch), algorithm based fault tolerance, ABFT (software)

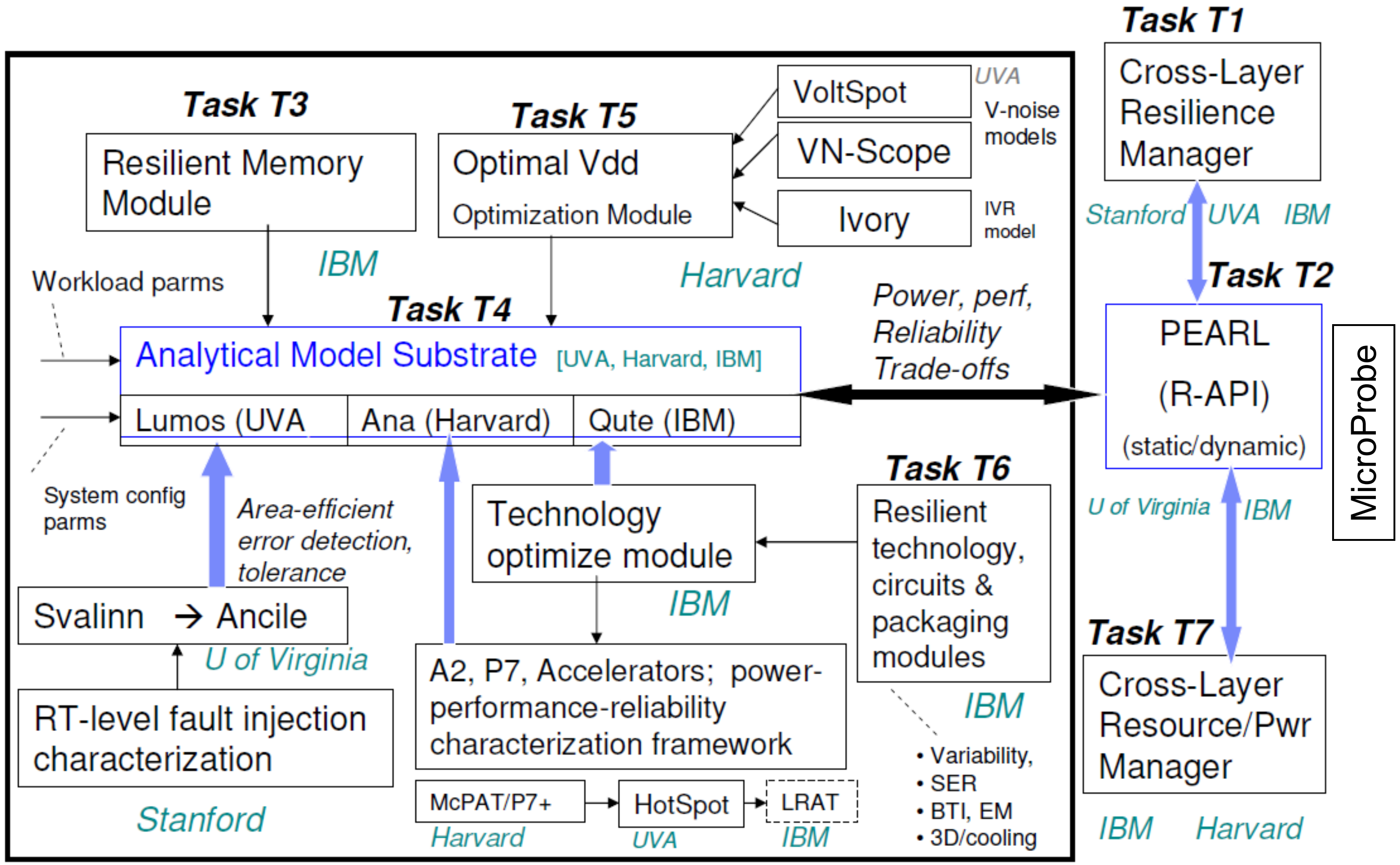
Calculation Assumptions

Node	Supply	FIT (shrink)	FIT (voltage)	FIT (total)
32 nm	1.00 v	1x	1x	1x
22 nm	0.85 v	2x	2x	4x
14 nm	0.65 v	4x	8x	12x
10 nm	0.50 v	8x	32x	40x
7 nm	0.50 v	16x	32x	48x

- FIT = unit of failure rate; 1 FIT = 1 failure in a billion hours; system mean time to failure, MTTF ~ 1/FITs
- System FITs will increase with technology node (bad!)
  - Two effects considered here: (a) device size shrinkage per Moore's Law: 2x component count increase per generation; and (b) increase of transient error rates (SER, voltage noise) with voltage reductions required to meet end target of 75 GF/W
- Note: FITs are additive; so last column = sum of the prior two

PERFECT: Overall System Modeling Framework

(Delivered in Phase-1; analytical models, open-source software toolset)



Task T3

Resilient Memory Module

IBM

Task T5

Optimal Vdd Optimization Module

Harvard

VoltSpot

VN-Scope

Ivory

UVA

V-noise models

IVR model

Workload parms

System config parms

Task T4

Analytical Model Substrate [UVA, Harvard, IBM]

Lumos (UVA | Ana (Harvard) | Qute (IBM)

Power, perf, Reliability Trade-offs

Area-efficient error detection, tolerance

Svalinn → Ancile

U of Virginia

RT-level fault injection characterization

Stanford

Technology optimize module

IBM

A2, P7, Accelerators; power-performance-reliability characterization framework

McPAT/P7+ → HotSpot → LRAT

Harvard | UVA | IBM

Task T6

Resilient technology, circuits & packaging modules

IBM

- Variability,
- SER
- BTI, EM
- 3D/cooling

Task T1

Cross-Layer Resilience Manager

Stanford | UVA | IBM

Task T2

PEARL

(R-API)

(static/dynamic)

MicroProbe

U of Virginia | IBM

Task T7

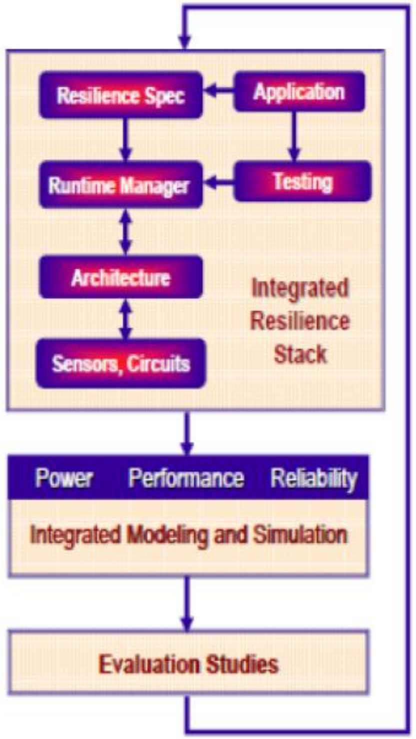
Cross-Layer Resource/Pwr Manager

IBM | Harvard

SHIVA-1 Framework

SHIVA-2 delivered in Phase-2 includes cycle-accurate processor core and accelerator elements

Cross-layer
Efficient Reslience
Technologies



Resilience Spec

Application

Runtime Manager

Testing

Architecture

Sensors, Circuits

Integrated Resilience Stack

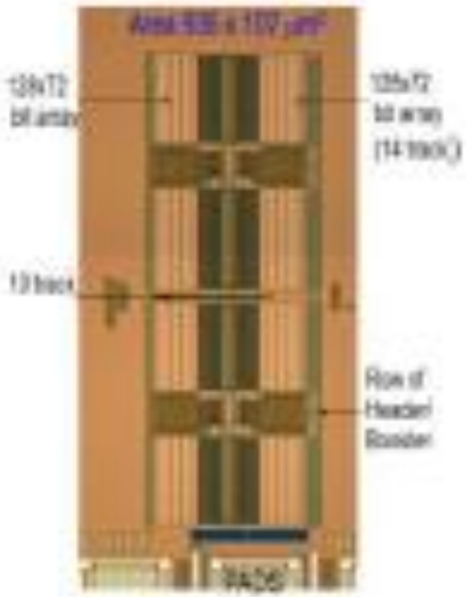
Power | Performance | Reliability

Integrated Modeling and Simulation

Evaluation Studies

Latch-accurate SHIVA-3 model in Phase-3 will be fully *design-ready,* with key FPGA component prototype implementations

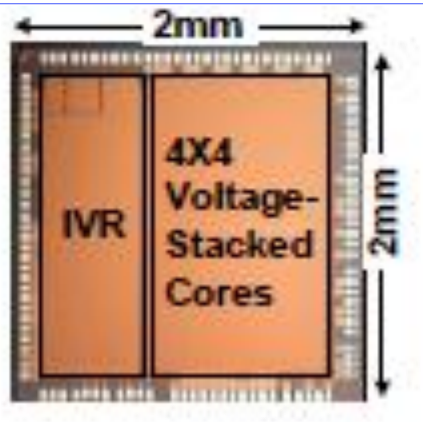
Test Chips to Validate Modeled PERFECT Innovations in Efficient Resilience; three accepted papers at VLSI Tech. & Circuits Symposia (Kyoto)



Ultra low-Vmin SRAM is a major technology breakthrough – in the quest for 75 GF/W embedded systems

14nm FinFET Based Supply Voltage Boosting Techniques for Extreme Low Vmin Operation

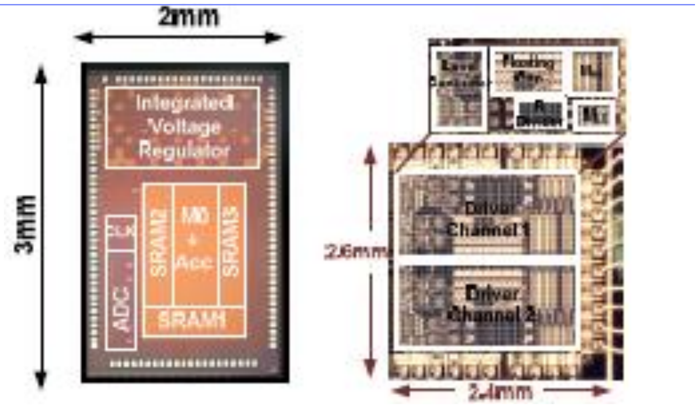
R. V. Joshi, M. Ziegler, H. Wetter, C. Wandel, H. Ainspan, **IBM**



IVR model calibration/& proof of voltage-stacking efficacy is a key new advance in exploring optimal Vdd settings for targeted embedded systems

A 16-core voltage-stacking system with an integrated switched-capacitor DC-DC converter

S. K. Lee, T. Tong, X. Zhang, D. Brooks, G-Y. Wei, **Harvard University**



Robo-bees brain SoC chip tests provide validation insights about ultra low power cognitive acceleration

A Multi-Chip System Optimized for Insect-Scale Flapping-Wing Robots

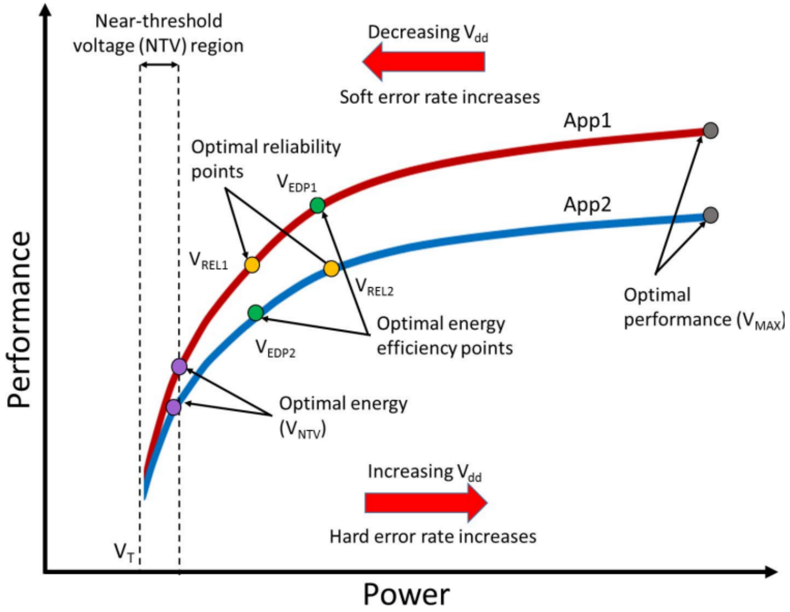
X. Zhang, M. Lok, T. Tong, S. Chaput, S. K. Lee, B. Reagan, H. Lee, D. Brooks, G-Y. Wei, **Harvard University**

A Couple of Key ModSim-Relevant Papers from our PERFECT Project

CLEAR Cross-Layer Resilience: A Retrospective. IEEE Des. Test 42(3): 74-85 (2025); Eric Cheng et al. (Stanford-led work)

*A key ModSim takeaway: architectural abstractions in fault-injection simulation are hazardous, the conclusions can be grossly misleading! **Up to 45x inaccuracy***

BRAVO: Balanced Reliability-Aware Voltage Optimization. HPCA 2017: 97-108
Karthik Swaminathan et al. (IBM work)



ModSim-driven discourse on how to optimize the voltage-frequency operating point to achieve highest performance without violating power and reliability constraints

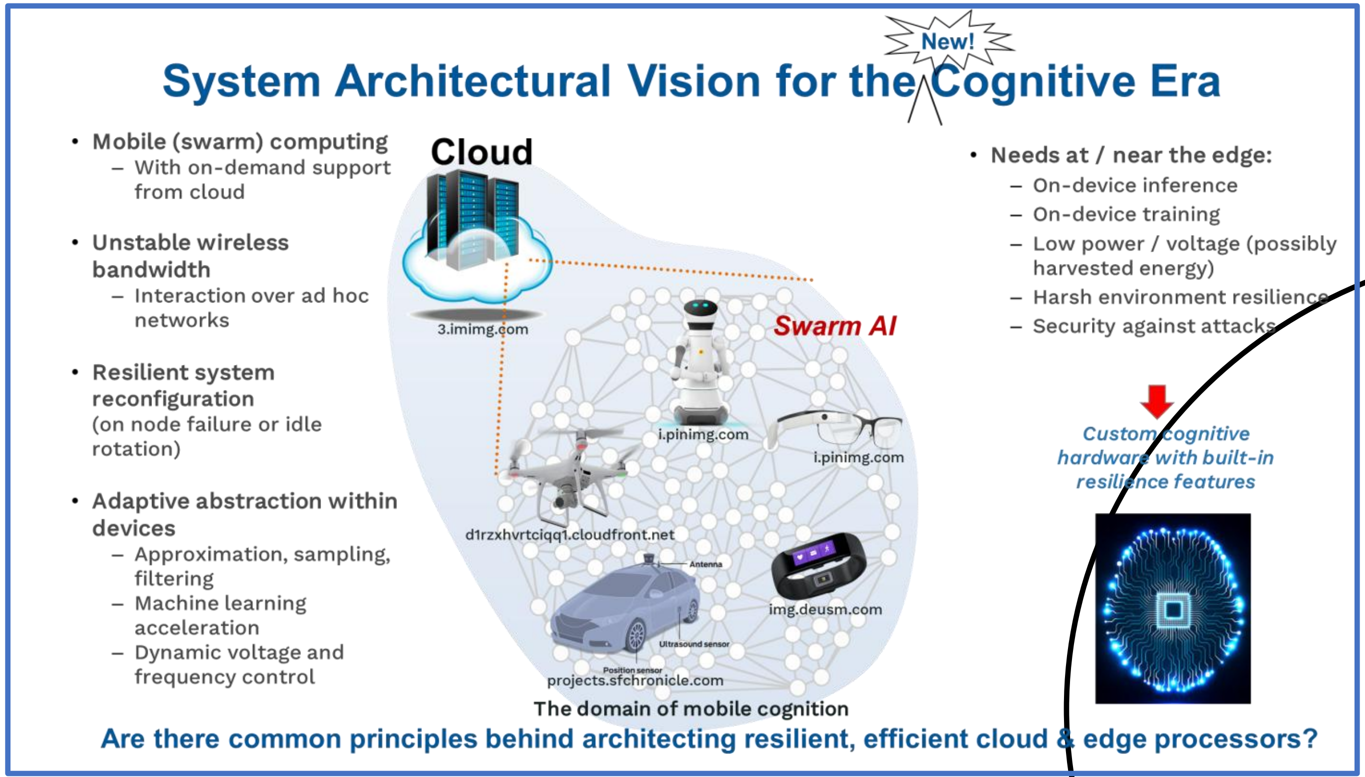
PERFECT-regime Real-Time Experimental Demo

https://www.youtube.com/watch?v=YvbHXz3lccc

That was 10 years ago!

DARPA-hard Challenges: ⇨ Onward to DSSoC

a good way of pushing the envelope in systems R&D


New!
Cloud
Swarm AI
3.imimg.com
i.pinimg.com
i.pinimg.com
d1rzxhvrtciqq1.cloudfront.net
Antenna
img.deusm.com
Ultrasound sensor
Position sensor
projects.sfchronicle.com

System Architectural Vision for the Cognitive Era

- Mobile (swarm) computing
  - With on-demand support from cloud
- Unstable wireless bandwidth
  - Interaction over ad hoc networks
- Resilient system reconfiguration (on node failure or idle rotation)
- Adaptive abstraction within devices
  - Approximation, sampling, filtering
  - Machine learning acceleration
  - Dynamic voltage and frequency control

The domain of mobile cognition

- Needs at / near the edge:
  - On-device inference
  - On-device training
  - Low power / voltage (possibly harvested energy)
  - Harsh environment resilience
  - Security against attacks

Custom cognitive hardware with built-in resilience features

Are there common principles behind architecting resilient, efficient cloud & edge processors?

- Agile SoC
- Programmability

PERFECT [Bob Colwell, Joe Cross, …] →

2013 – 2018

Power Efficiency Revolution for Embedded Computing Technologies

1 GF/W → 75 GF/W

IBM + Stanford, Harvard, U of Virginia

DSSoC [Tom Rondeau]

2018 → 2023/ongoing

Domain-Specific System on Chip

Power-perf, programmability, productivity metrics

IBM + Columbia, Harvard, UIUC

Our recently-completed DARPA (DSSoC) sponsored project:

EPOCHS: Efficient Programmability of Cognitive Heterogeneous Systems

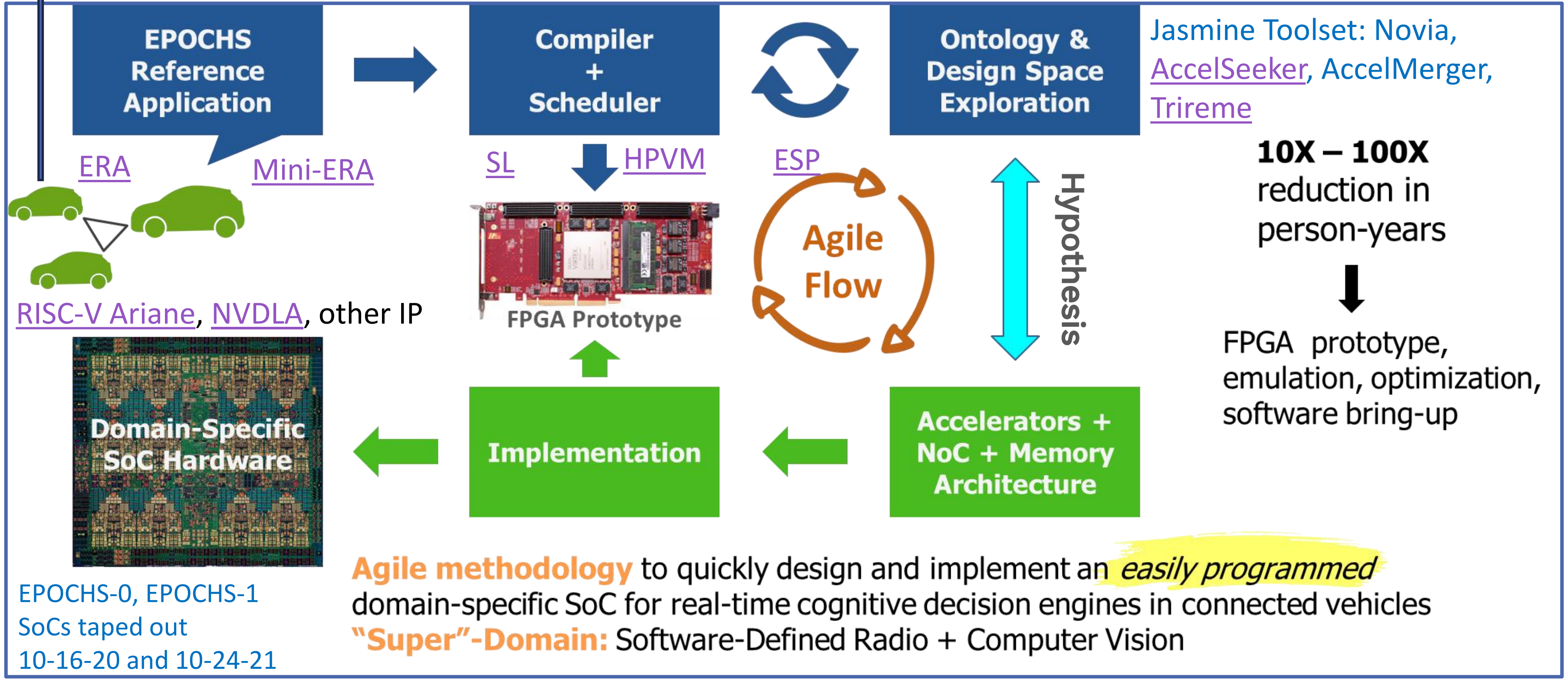
connected autonomous vehicles (CAVs)**

- Agile design of heterogeneous DSSoCs with programmability as a primary consideration
- Open-source software and hardware
- Technology transition: within IBM and outside, including DoD entities

one example

https://mas400.com

Technical Approach



EPOCHS Reference Application
Compiler + Scheduler
Ontology & Design Space Exploration
ERA
Mini-ERA
SL
HPVM
ESP
FPGA Prototype
Agile Flow
Hypothesis
RISC-V Ariane, NVDLA, other IP
Domain-Specific SoC Hardware
Implementation
Accelerators + NoC + Memory Architecture

Jasmine Toolset: Novia, AccelSeeker, AccelMerger, Trireme

10X – 100X reduction in person-years

FPGA prototype, emulation, optimization, software bring-up

EPOCHS-0, EPOCHS-1 SoCs taped out 10-16-20 and 10-24-21

Agile methodology to quickly design and implement an *easily programmed* domain-specific SoC for real-time cognitive decision engines in connected vehicles

"Super"-Domain: Software-Defined Radio + Computer Vision

Tightly knit collaborative team: IBM + UIUC, Harvard and Columbia

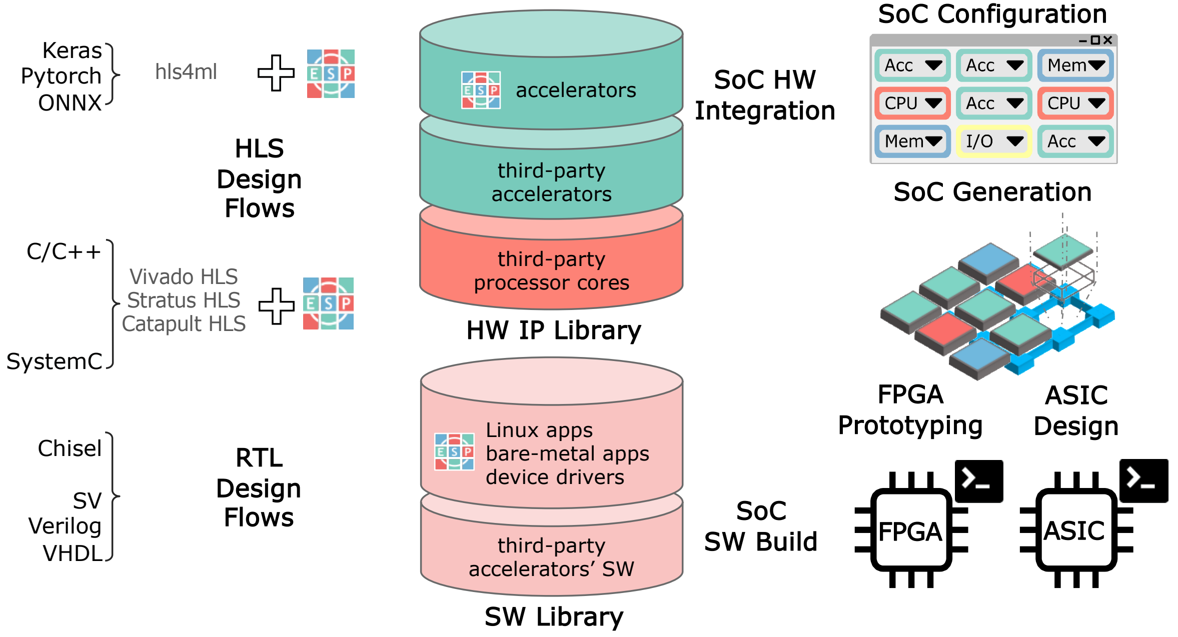
Targeted impact on AI hardware roadmap:
energy reduction, without giving up inferential accuracy

** an embodiment of "swarm intelligence" (bio-inspired AI application)

Agile SoC Design Flow: the Heart of EPOCHS ModSim

ESP: the open-source agile flow for system-on-chip (SoC) design

- *Seamless* integration of SoC components: NoC & platform services
- *Push-button* generation of SoC RTL for ASIC physical design
- **Two EPOCHS full SoC ASIC chip tapeouts**
- Rapid FPGA prototyping → early application development



Keras
Pytorch
ONNX
hls4ml

HLS Design Flows

C/C++
SystemC
Vivado HLS
Stratus HLS
Catapult HLS

Chisel
SV
Verilog
VHDL

RTL Design Flows

accelerators

third-party accelerators

third-party processor cores

HW IP Library

Linux apps
bare-metal apps
device drivers

third-party accelerators' SW

SW Library

SoC HW Integration

SoC SW Build

SoC Configuration

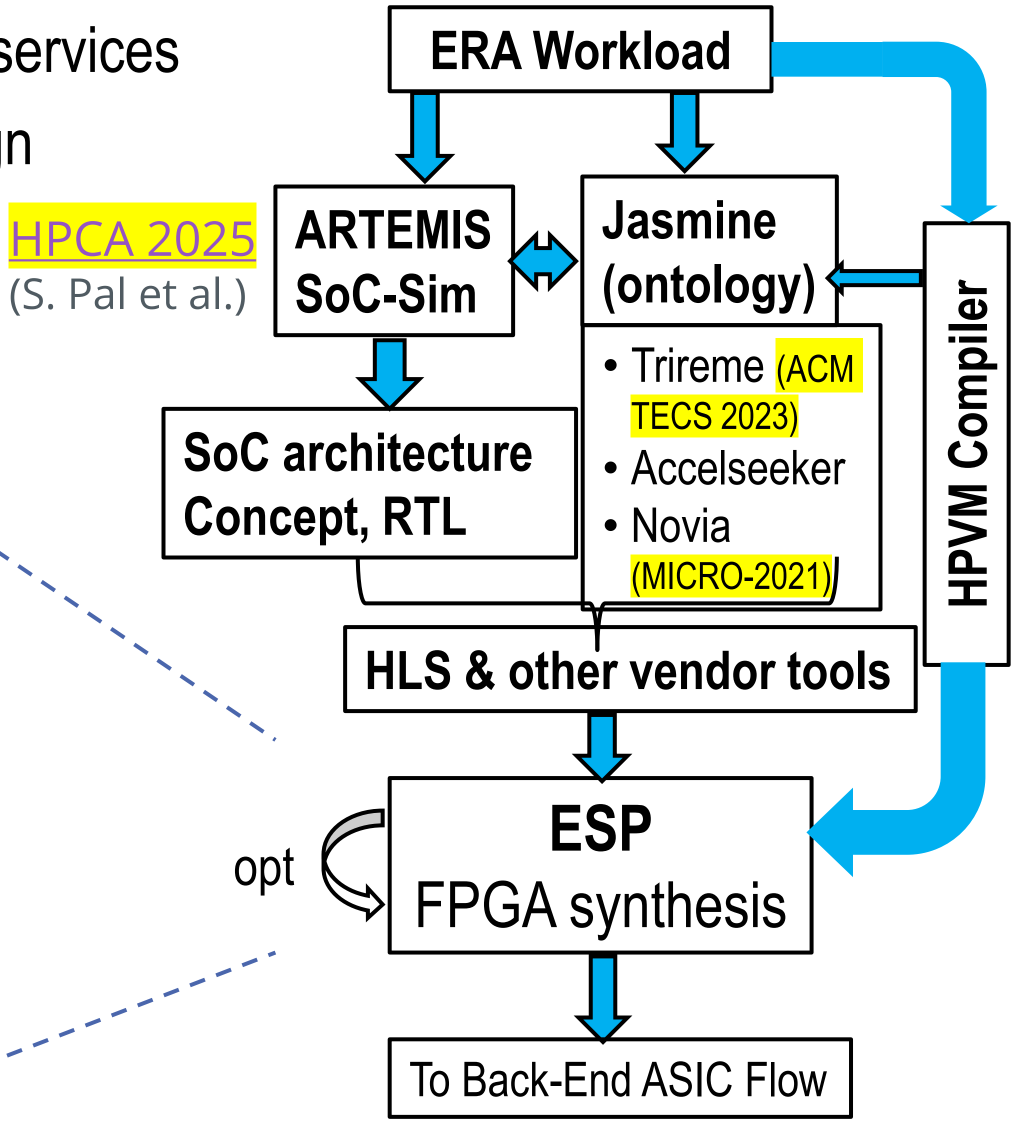
SoC Generation

FPGA Prototyping

ASIC Design

www.esp.cs.columbia.edu

open-source toolset



ERA Workload

HPCA 2025
(S. Pal et al.)

ARTEMIS SoC-Sim

Jasmine (ontology)

- Trireme (ACM TECS 2023)
- Accelseeker
- Novia (MICRO-2021)

HPVM Compiler

SoC architecture Concept, RTL

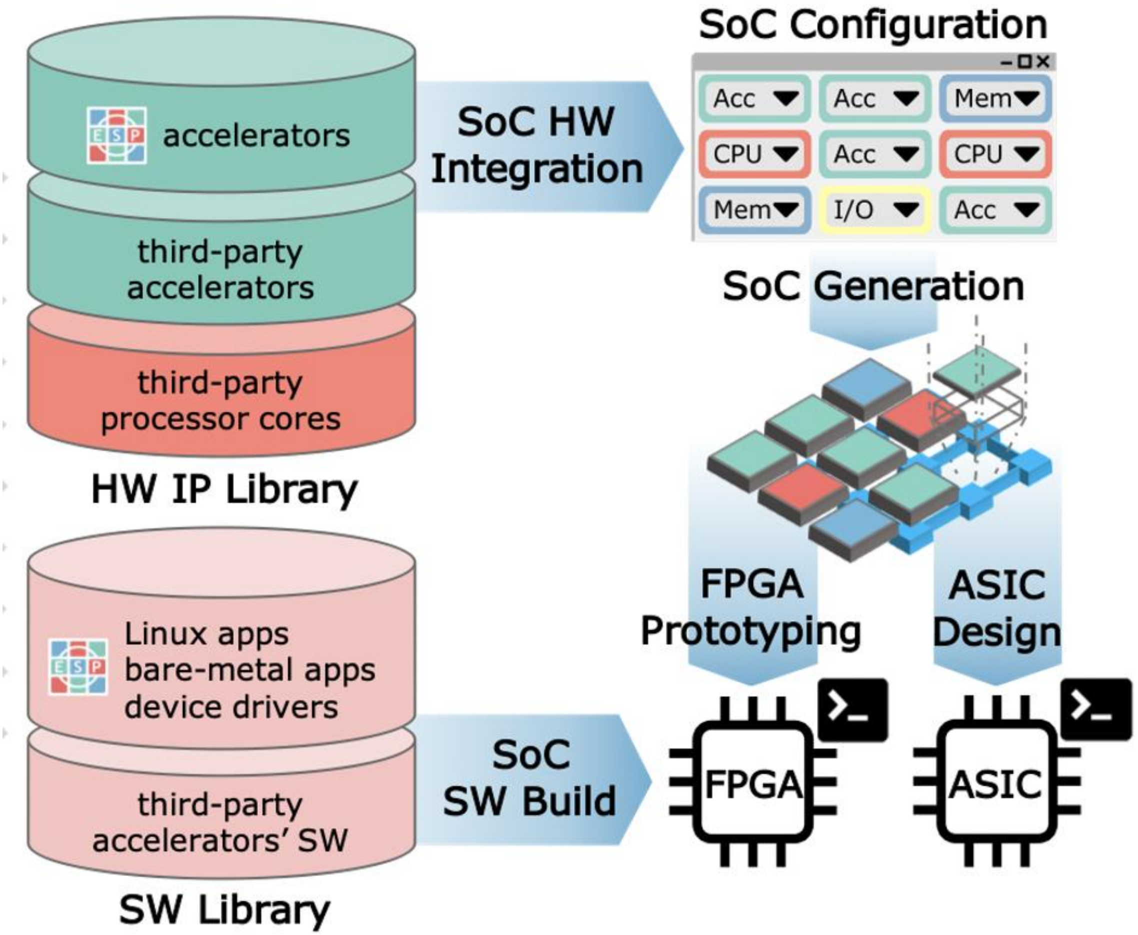
HLS & other vendor tools

opt

ESP
FPGA synthesis

To Back-End ASIC Flow

ESP SoC Flow



accelerators

third-party accelerators

third-party processor cores

HW IP Library

Linux apps
bare-metal apps
device drivers

third-party accelerators' SW

SW Library

SoC HW Integration

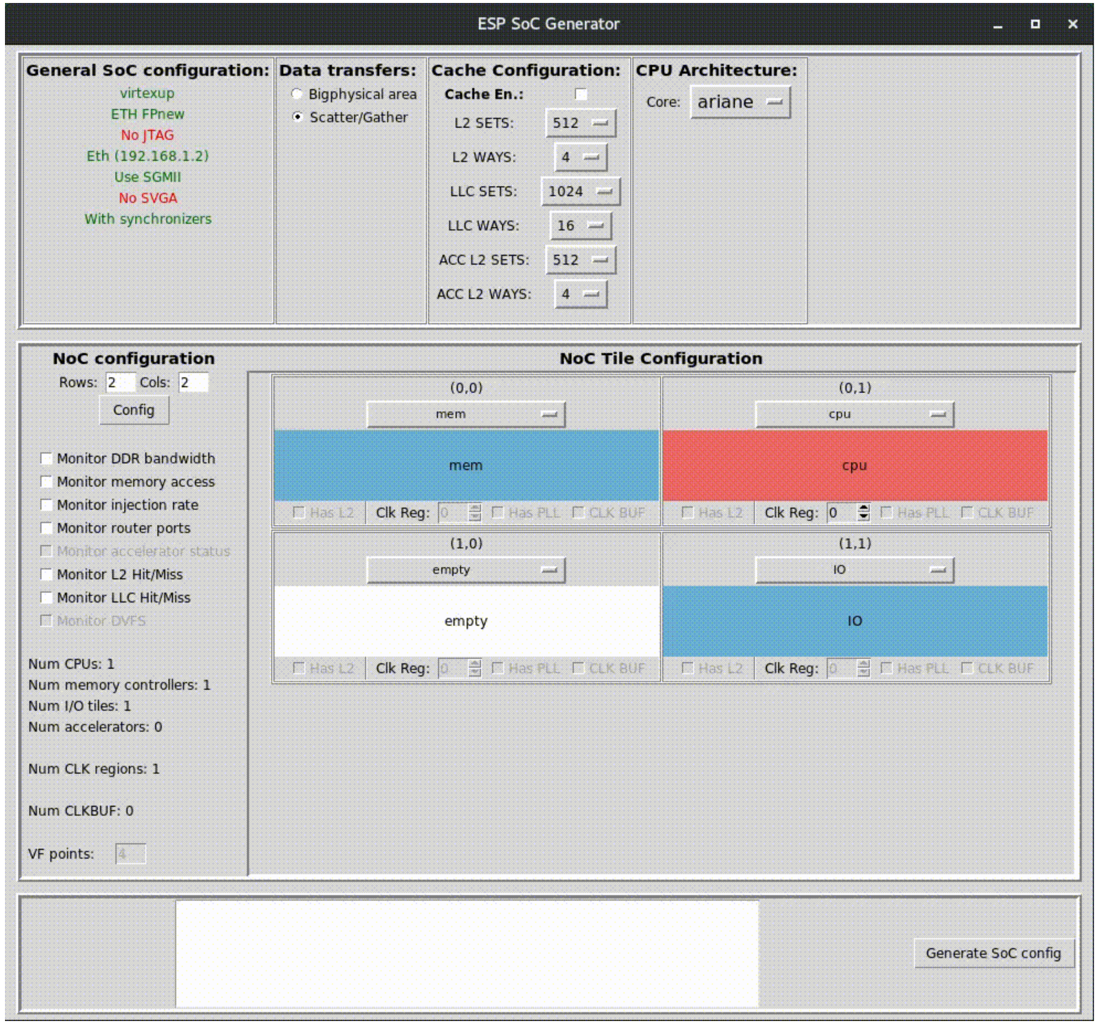
SoC Configuration

SoC Generation

FPGA Prototyping

ASIC Design

SoC SW Build



ESP SoC Generator

General SoC configuration:

virtexup
ETH FPnew
No JTAG
Eth (192.168.1.2)
Use SGMII
No SVGA
With synchronizers

Data transfers:

○ Bigphysical area
● Scatter/Gather

Cache Configuration:

Cache En.: □
L2 SETS: 512
L2 WAYS: 4
LLC SETS: 1024
LLC WAYS: 16
ACC L2 SETS: 512
ACC L2 WAYS: 4

CPU Architecture:

Core: ariane

NoC configuration

Rows: 2 Cols: 2

Config

□ Monitor DDR bandwidth
□ Monitor memory access
□ Monitor injection rate
□ Monitor router ports
□ Monitor accelerator status
□ Monitor L2 Hit/Miss
□ Monitor LLC Hit/Miss
□ Monitor DVFS

Num CPUs: 1
Num memory controllers: 1
Num I/O tiles: 1
Num accelerators: 0

Num CLK regions: 1

Num CLKBUF: 0

VF points: 4

NoC Tile Configuration

(0,0)
mem
mem
□ Has L2 Clk Reg: 0 □ Has PLL □ CLK BUF

(0,1)
cpu
cpu
□ Has L2 Clk Reg: 0 □ Has PLL □ CLK BUF

(1,0)
empty
empty
□ Has L2 Clk Reg: 0 □ Has PLL □ CLK BUF

(1,1)
IO
IO
□ Has L2 Clk Reg: 0 □ Has PLL □ CLK BUF

Generate SoC config

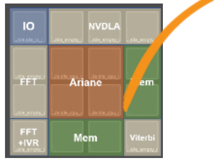
Slide courtesy: Joseph Zuckerman, Luca Carloni et al., Columbia University

EPOCHS/DSSoC: Accomplishments Summary

IBM

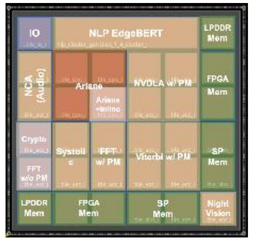
EPOCHS-0 SoC tapeout

– 4×4 SoC fabricated



Scaled-out EPOCHS-1 SoC tapeout

– 6×6 SoC with new accelerators



Chip back from fab + packaging (July 2022)
Respin: Nov 2023

Significant design cost mitigation

– 10×–100× reduction in person-years

Hardware-agnostic programming of heterogeneous SoCs

– HPVM compiler, smart scheduler…

Click!

DARPA ERI Summit Demo, Aug. 2023

SCAN ME

Open-source ecosystem for collaboration

ERA: github.com/IBM/era

HPVM: gitlab.engr.illinois.edu/llvm/hpvm-release

Mini-ERA: github.com/IBM/mini-era

STOMP: github.com/IBM/stomp

ESP: www.esp.cs.columbia.edu

Scheduler: github.com/IBM/scheduler-library

Spandex: github.com/sld-columbia/esp/tree/master/rtl/caches

ESSCIRC-2022 paper

Simultaneous apps

4 (goal: ≥ 2)

Integration time for new accelerators

2 weeks average (goal: ≤ 3 months)

Power

NoC: 7.2% of chip (goal: ≤ 40% of chip)

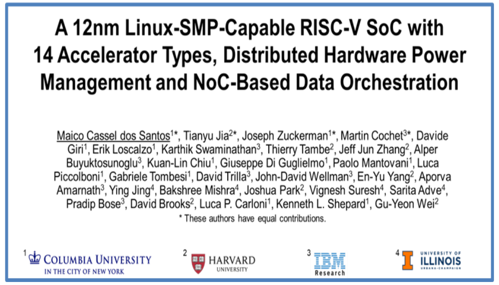
Chip: 240mW – 1.83W (op. range: 0.5V – 1.0V)

Peak frequency at 1.0V: 1.52 GHz

Benefits of acceleration

	FFT	Viterbi
Performance	71×	20×
Energy	233×	56×

Even more amazing results!



A 12nm Linux-SMP-Capable RISC-V SoC with 14 Accelerator Types, Distributed Hardware Power Management and NoC-Based Data Orchestration

Maico Cassel dos Santos[1]*, Tianyu Jia[2]*, Joseph Zuckerman[1]*, Martin Cochet[3]*, Davide Giri[1], Erik Loscalzo[1], Karthik Swaminathan[3], Thierry Tambe[2], Jeff Jun Zhang[2], Alper Buyuktosunoglu[3], Kuan-Lin Chiu[1], Giuseppe Di Guglielmo[1], Paolo Mantovani[1], Luca Piccolboni[1], Gabriele Tombesi[1], David Trilla[3], John-David Wellman[3], En-Yu Yang[2], Aporva Amarnath[3], Ying Jing[4], Bakshree Mishra[4], Joshua Park[2], Vignesh Suresh[4], Sarita Adve[4], Pradip Bose[3], David Brooks[2], Luca P. Carloni[1], Kenneth L. Shepard[1], Gu-Yeon Wei[2]

* These authors have equal contributions.

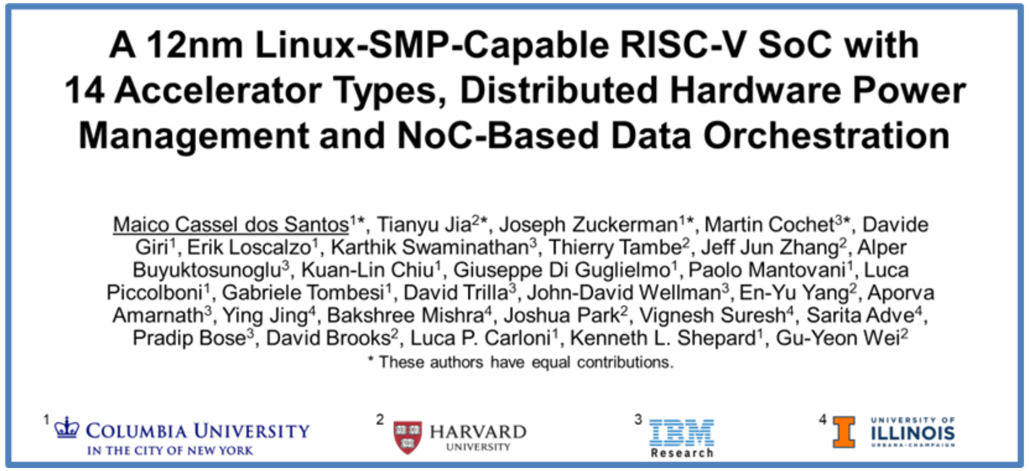
[1] Columbia University in the City of New York [2] Harvard University [3] IBM Research [4] University of Illinois Urbana-Champaign

ISSCC-2024

ISCA-2024

VLSI Symp.2024

EPOCHS-1 SoC Highlights



A 12nm Linux-SMP-Capable RISC-V SoC with 14 Accelerator Types, Distributed Hardware Power Management and NoC-Based Data Orchestration

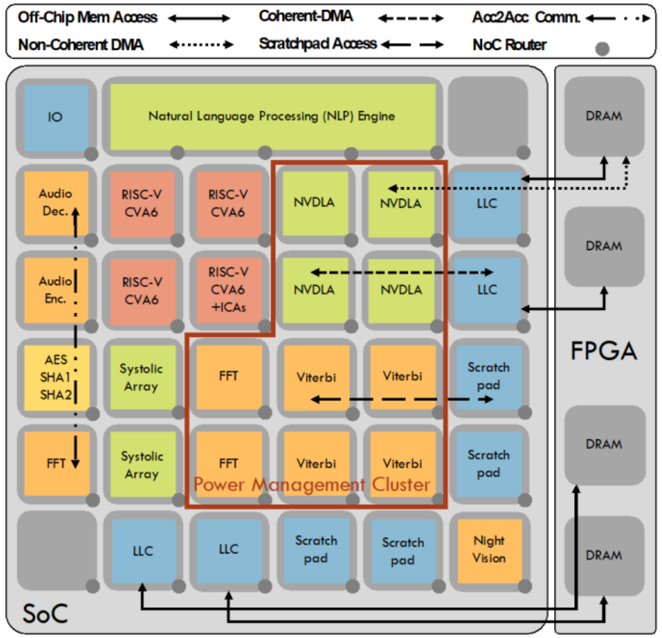
Maico Cassel dos Santos[1]*, Tianyu Jia[2]*, Joseph Zuckerman[1]*, Martin Cochet[3]*, Davide Giri[1], Erik Loscalzo[1], Karthik Swaminathan[3], Thierry Tambe[2], Jeff Jun Zhang[2], Alper Buyuktosunoglu[3], Kuan-Lin Chiu[1], Giuseppe Di Guglielmo[1], Paolo Mantovani[1], Luca Piccolboni[1], Gabriele Tombesi[1], David Trilla[3], John-David Wellman[3], En-Yu Yang[2], Aporva Amarnath[3], Ying Jing[4], Bakshree Mishra[4], Joshua Park[2], Vignesh Suresh[4], Sarita Adve[4], Pradip Bose[3], David Brooks[2], Luca P. Carloni[1], Kenneth L. Shepard[1], Gu-Yeon Wei[2]

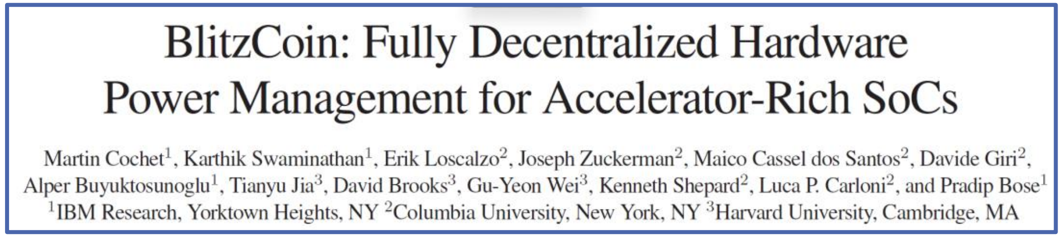
* These authors have equal contributions.

[1] Columbia University in the City of New York [2] Harvard University [3] IBM Research [4] University of Illinois Urbana-Champaign

ISSCC-2024 Paper

- **64** mm² SoC designed in **12** nm FinFET
- **35** clock domains; **23** power domains
- **8.4** MB on-chip SRAM memory
- Tile-based SoC architecture
- **34** tiles connected by a **6**-plane 2-D mesh NoC
- The **74** Tbps NoC provides flexible orchestration of data
- **23** accelerators of **14** different types
- **10** accelerators compose a cluster demonstrating a novel distributed hardware power management scheme
- Designed by a small team of PhD students, postdocs, and industry researchers in **3** months with **ESP**, an open-source platform for agile SoC design



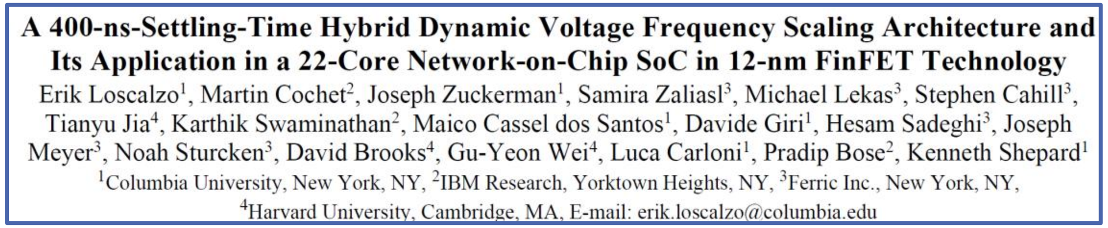


BlitzCoin: Fully Decentralized Hardware Power Management for Accelerator-Rich SoCs

Martin Cochet[1], Karthik Swaminathan[1], Erik Loscalzo[2], Joseph Zuckerman[2], Maico Cassel dos Santos[2], Davide Giri[2], Alper Buyuktosunoglu[1], Tianyu Jia[3], David Brooks[3], Gu-Yeon Wei[3], Kenneth Shepard[2], Luca P. Carloni[2], and Pradip Bose[1]

[1]IBM Research, Yorktown Heights, NY [2]Columbia University, New York, NY [3]Harvard University, Cambridge, MA

ISCA-2024 paper

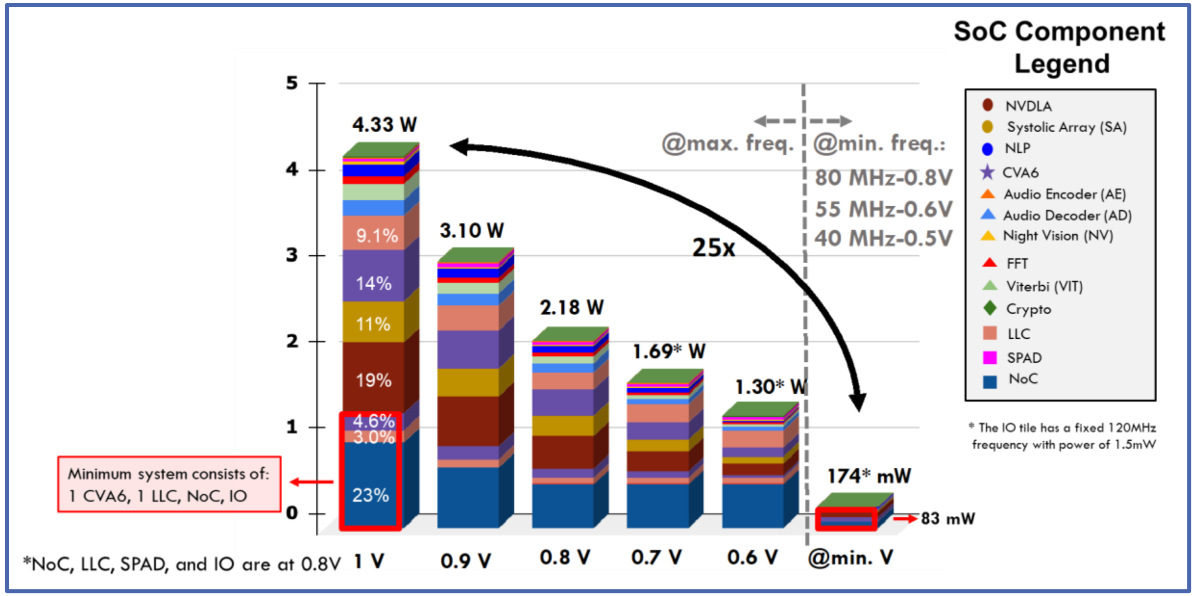


A 400-ns-Settling-Time Hybrid Dynamic Voltage Frequency Scaling Architecture and Its Application in a 22-Core Network-on-Chip SoC in 12-nm FinFET Technology

Erik Loscalzo[1], Martin Cochet[2], Joseph Zuckerman[1], Samira Zaliasl[3], Michael Lekas[3], Stephen Cahill[3], Tianyu Jia[4], Karthik Swaminathan[2], Maico Cassel dos Santos[1], Davide Giri[1], Hesam Sadeghi[3], Joseph Meyer[3], Noah Sturcken[3], David Brooks[4], Gu-Yeon Wei[4], Luca Carloni[1], Pradip Bose[2], Kenneth Shepard[1]

[1]Columbia University, New York, NY, [2]IBM Research, Yorktown Heights, NY, [3]Ferric Inc., New York, NY, [4]Harvard University, Cambridge, MA, E-mail: erik.loscalzo@columbia.edu

VLSI Symp. 2024 paper



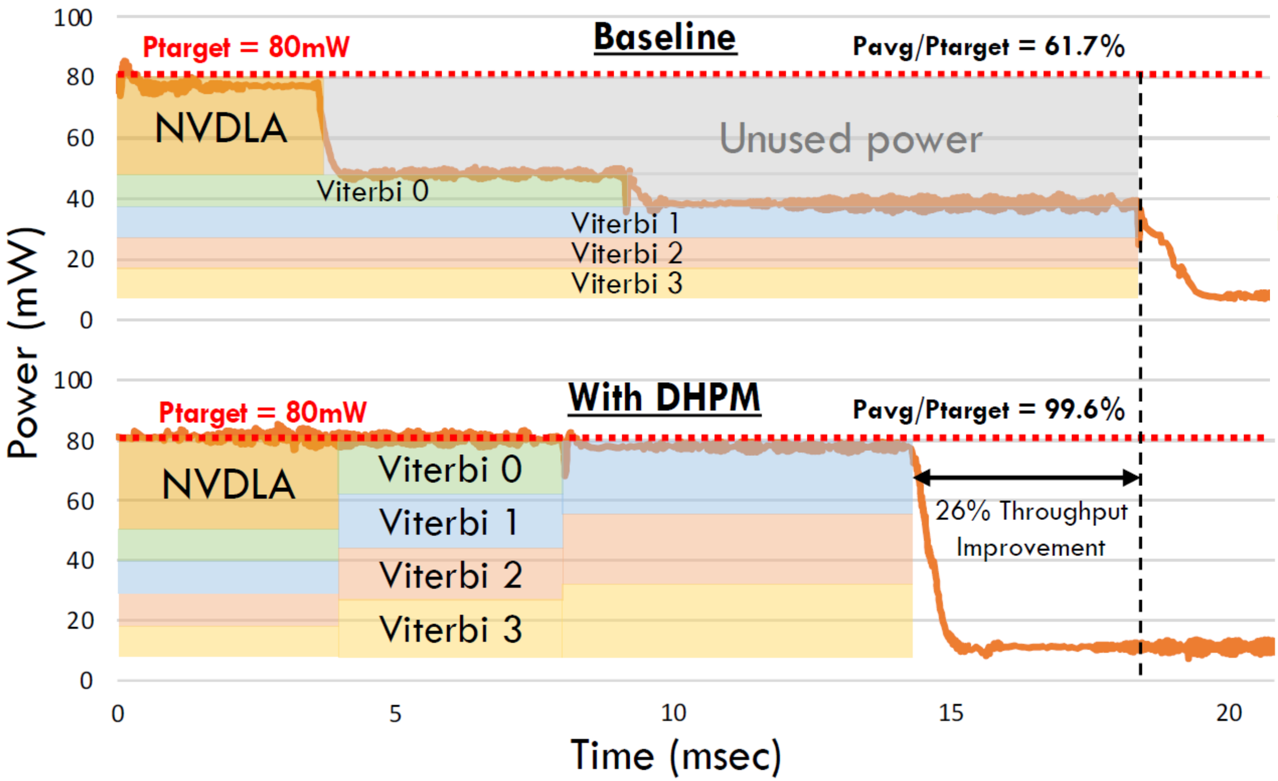
Minimum system consists of: 1 CVA6, 1 LLC, NoC, IO

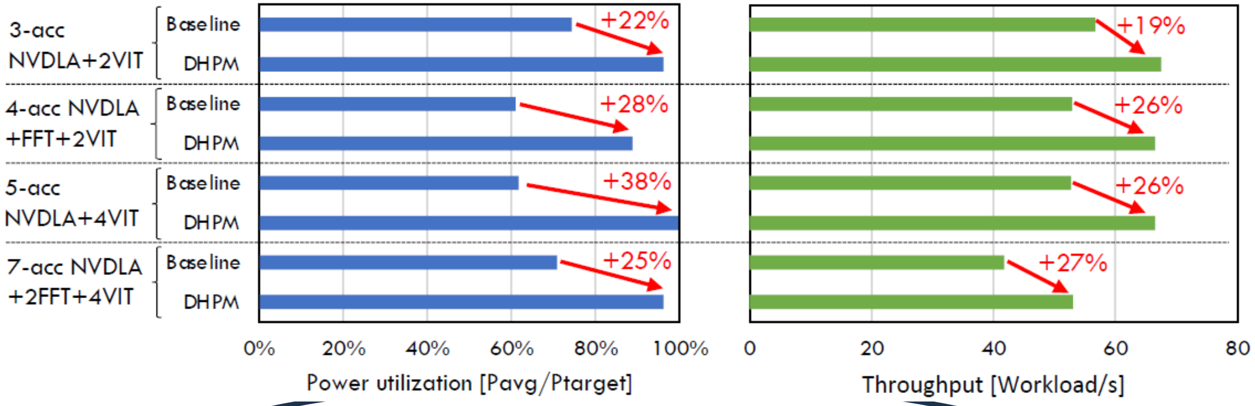
* The IO tile has a fixed 120MHz frequency with power of 1.5mW

*NoC, LLC, SPAD, and IO are at 0.8V

Distributed Hardware Power Management

- Concurrent execution of 5 accelerators under fixed 80mW power cap
- Without DHPM (baseline), each tile is allocated a fixed power
- With DHPM, power is dynamically reallocated among tiles





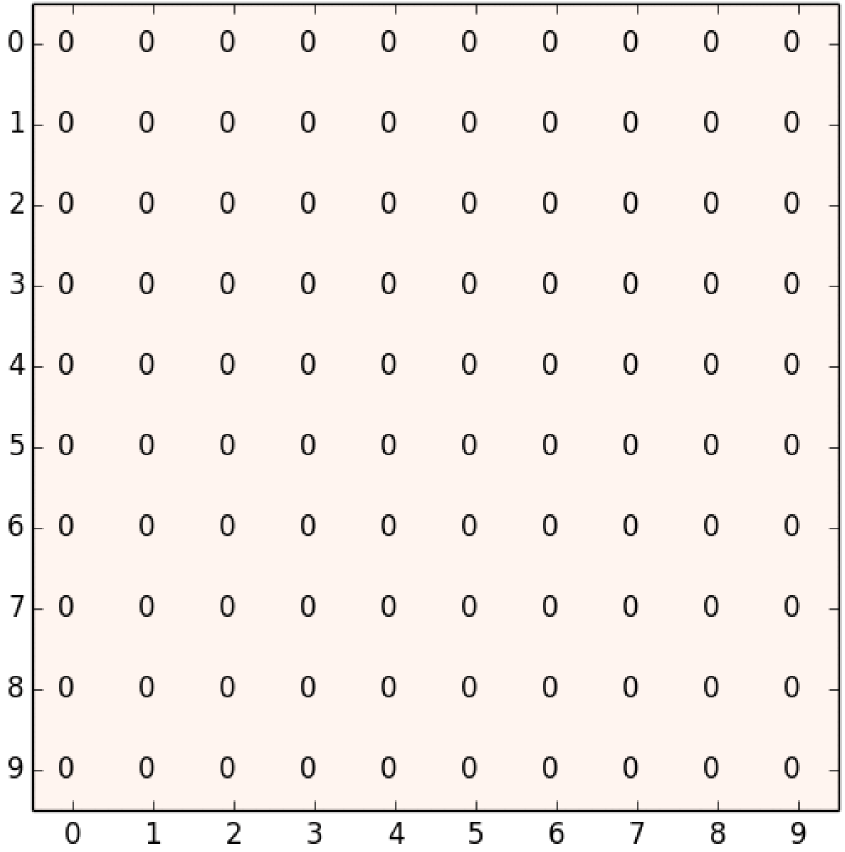
→ 22-38% power utilization improvement translating to 19-27% throughput improvement with full-hardware scalable implementation

Token Shortfall Situations

(early-stage concept ModSim)

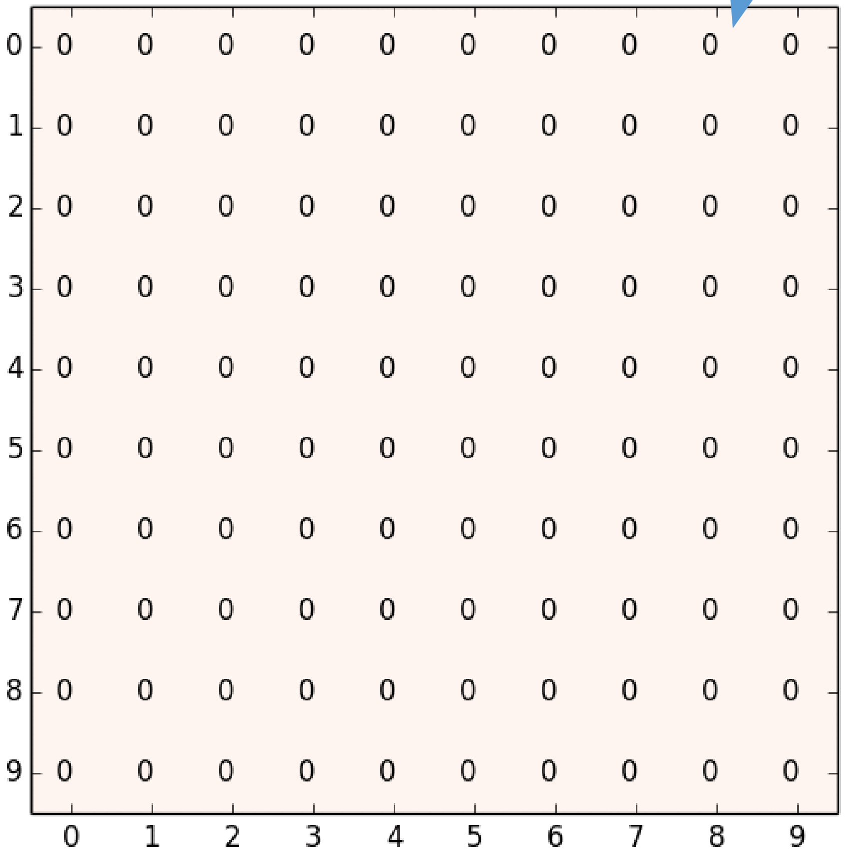
Numbers in each core represent the deficit of tokens (the lower the better)

Disabled*



	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0

Enabled v4*



	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0

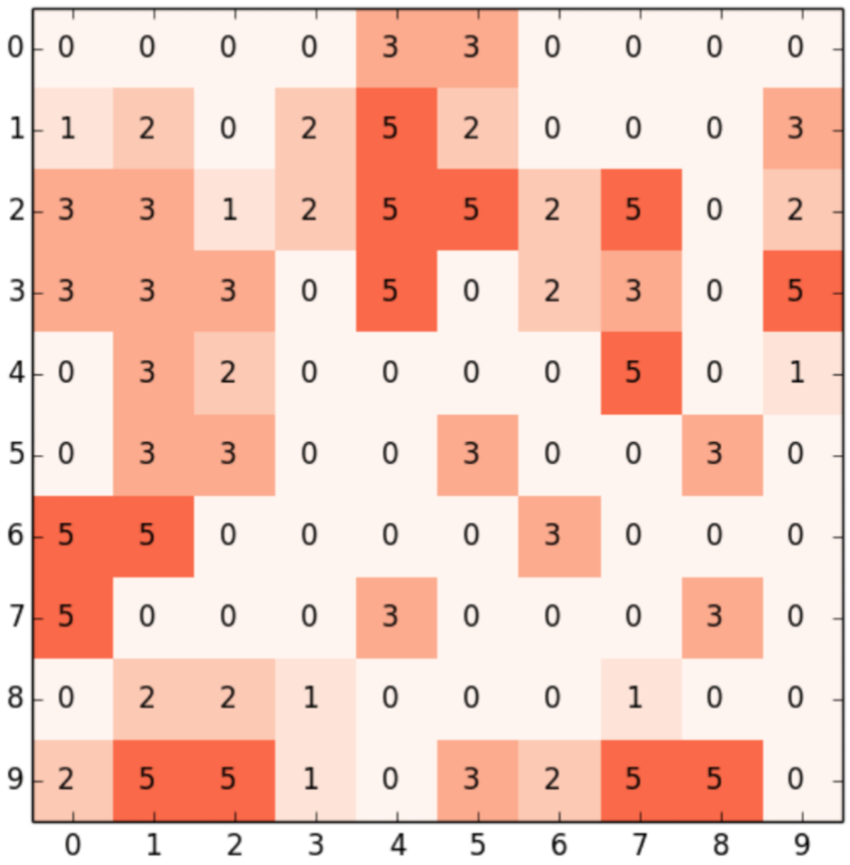
** Animation frames taken every 100 simulation iterations* (animations won't show up in pdf, sorry!)

Token Shortfall Situations

(early-stage concept ModSim)

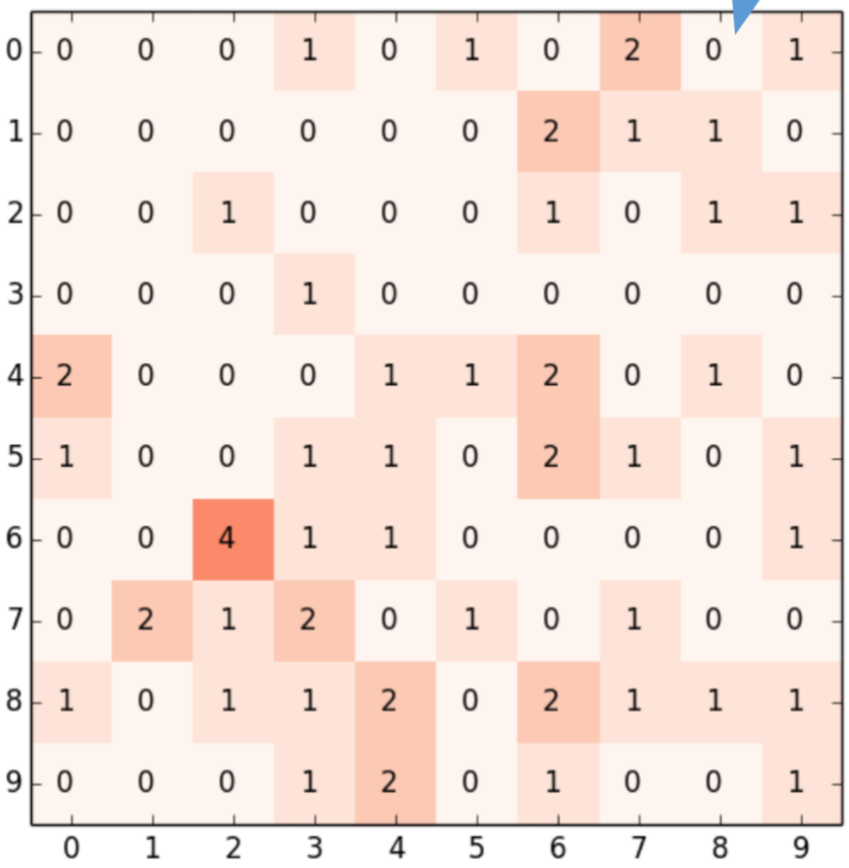
Numbers in each core represent the deficit of tokens (the lower the better)

Disabled*



	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	3	3	0	0	0	0
1	1	2	0	2	5	2	0	0	0	3
2	3	3	1	2	5	5	2	5	0	2
3	3	3	3	0	5	0	2	3	0	5
4	0	3	2	0	0	0	0	5	0	1
5	0	3	3	0	0	3	0	0	3	0
6	5	5	0	0	0	0	3	0	0	0
7	5	0	0	0	3	0	0	0	3	0
8	0	2	2	1	0	0	0	1	0	0
9	2	5	5	1	0	3	2	5	5	0

Enabled v4*



	0	1	2	3	4	5	6	7	8	9
0	0	0	0	1	0	1	0	2	0	1
1	0	0	0	0	0	0	2	1	1	0
2	0	0	1	0	0	0	1	0	1	1
3	0	0	0	1	0	0	0	0	0	0
4	2	0	0	0	1	1	2	0	1	0
5	1	0	0	1	1	0	2	1	0	1
6	0	0	4	1	1	0	0	0	0	1
7	0	2	1	2	0	1	0	1	0	0
8	1	0	1	1	2	0	2	1	1	1
9	0	0	0	1	2	0	1	0	0	1

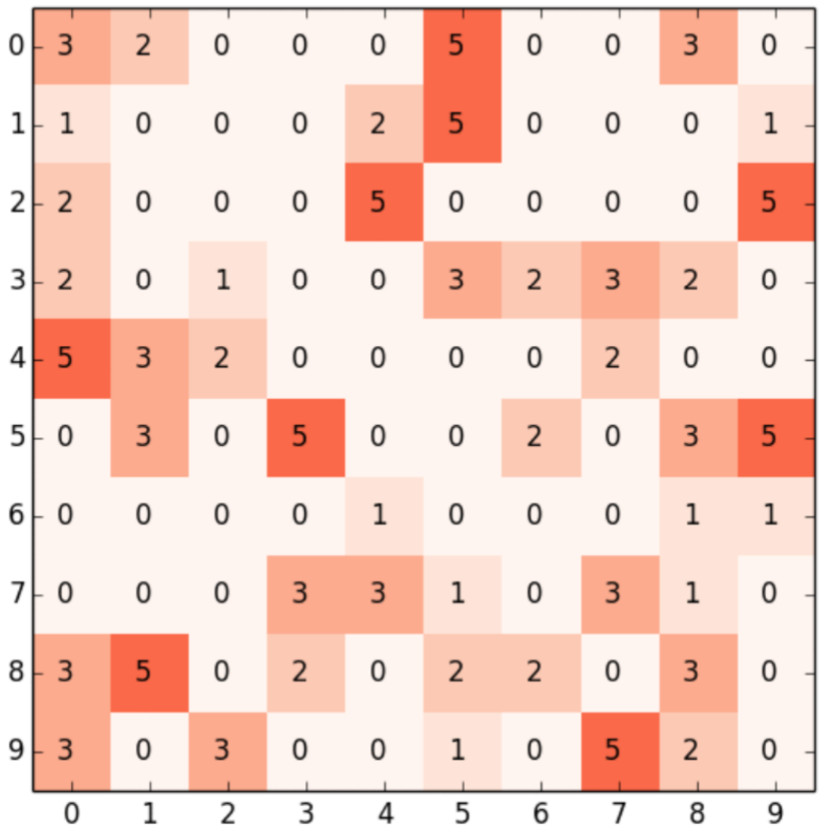
* *Animation frames taken every 100 simulation iterations* (animations won't show up in pdf, sorry!)

Token Shortfall Situations

(early-stage concept ModSim)

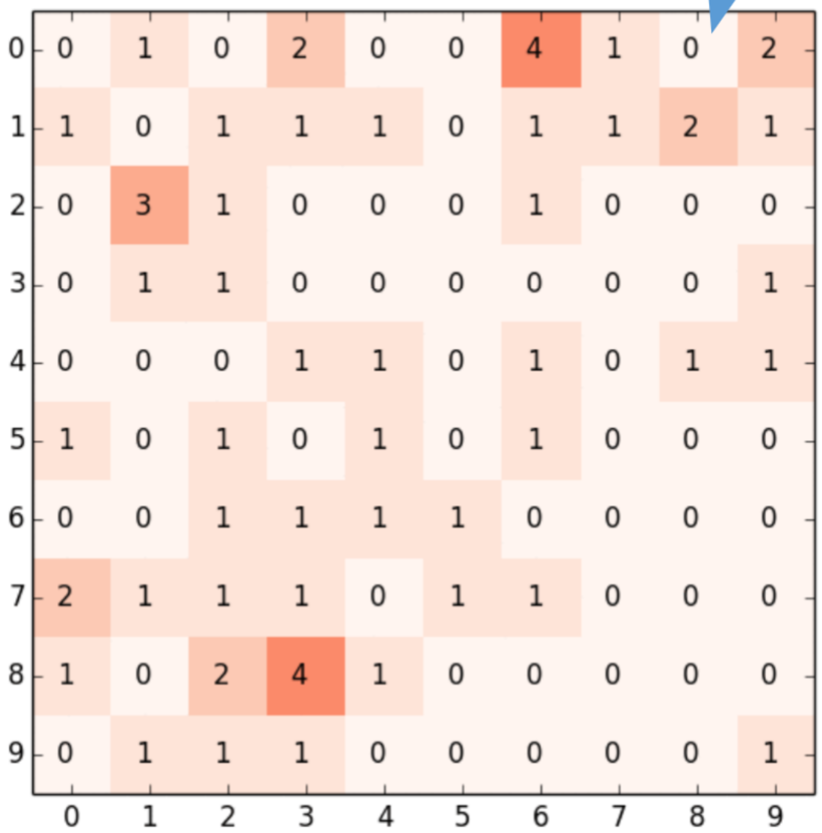
Numbers in each core represent the deficit of tokens (the lower the better)

Disabled*



	0	1	2	3	4	5	6	7	8	9
0	3	2	0	0	0	5	0	0	3	0
1	1	0	0	0	2	5	0	0	0	1
2	2	0	0	0	5	0	0	0	0	5
3	2	0	1	0	0	3	2	3	2	0
4	5	3	2	0	0	0	0	2	0	0
5	0	3	0	5	0	0	2	0	3	5
6	0	0	0	0	1	0	0	0	1	1
7	0	0	0	3	3	1	0	3	1	0
8	3	5	0	2	0	2	2	0	3	0
9	3	0	3	0	0	1	0	5	2	0

Enabled v4*



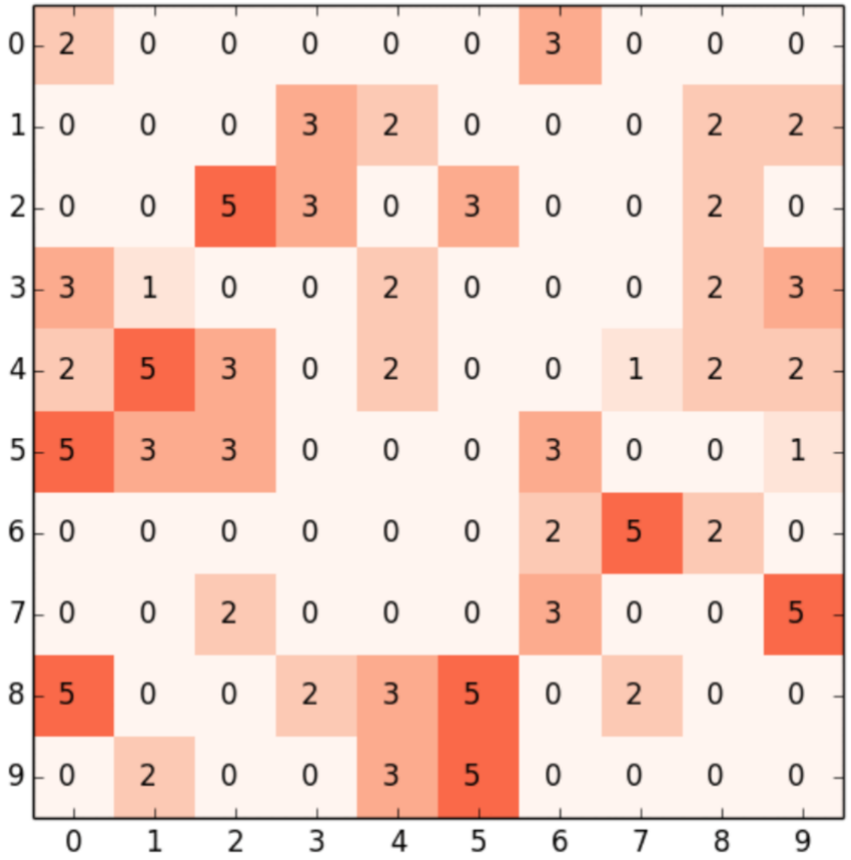
	0	1	2	3	4	5	6	7	8	9
0	0	1	0	2	0	0	4	1	0	2
1	1	0	1	1	1	0	1	1	2	1
2	0	3	1	0	0	0	1	0	0	0
3	0	1	1	0	0	0	0	0	0	1
4	0	0	0	1	1	0	1	0	1	1
5	1	0	1	0	1	0	1	0	0	0
6	0	0	1	1	1	1	0	0	0	0
7	2	1	1	1	0	1	1	0	0	0
8	1	0	2	4	1	0	0	0	0	0
9	0	1	1	1	0	0	0	0	0	1

* *Animation frames taken every 100 simulation iterations* (animations won't show up in pdf, sorry!)

Token Shortfall Situations

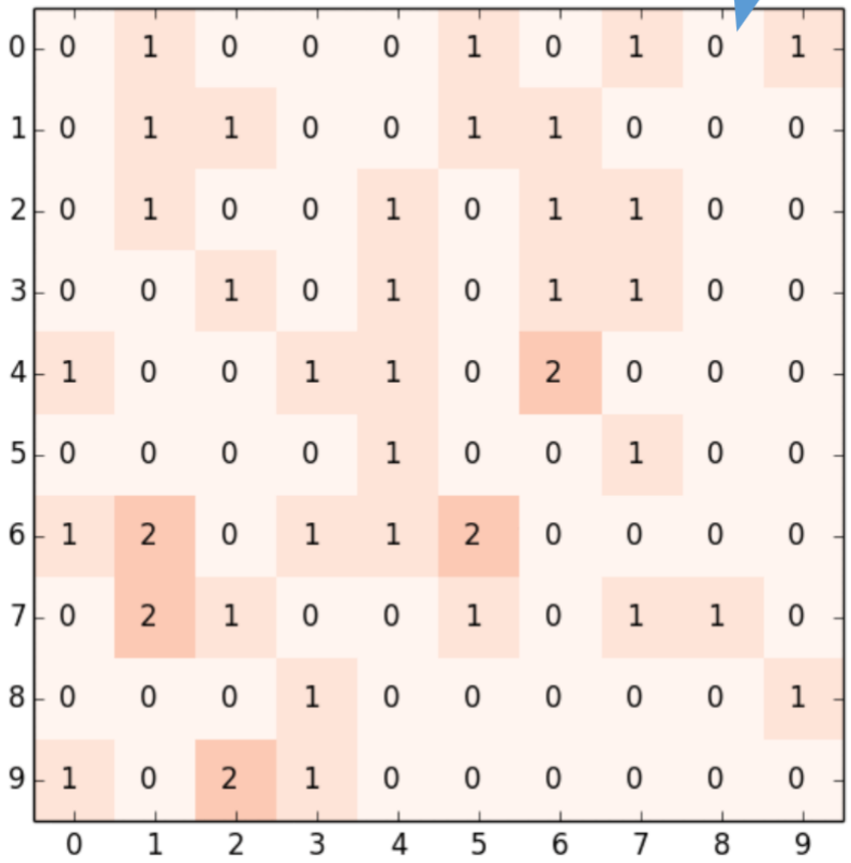
(early-stage concept ModSim)

Disabled*



	0	1	2	3	4	5	6	7	8	9
0	2	0	0	0	0	0	3	0	0	0
1	0	0	0	3	2	0	0	0	2	2
2	0	0	5	3	0	3	0	0	2	0
3	3	1	0	0	2	0	0	0	2	3
4	2	5	3	0	2	0	0	1	2	2
5	5	3	3	0	0	0	3	0	0	1
6	0	0	0	0	0	0	2	5	2	0
7	0	0	2	0	0	0	3	0	0	5
8	5	0	0	2	3	5	0	2	0	0
9	0	2	0	0	3	5	0	0	0	0

Enabled v4*



	0	1	2	3	4	5	6	7	8	9
0	0	1	0	0	0	1	0	1	0	1
1	0	1	1	0	0	1	1	0	0	0
2	0	1	0	0	1	0	1	1	0	0
3	0	0	1	0	1	0	1	1	0	0
4	1	0	0	1	1	0	2	0	0	0
5	0	0	0	0	1	0	0	1	0	0
6	1	2	0	1	1	2	0	0	0	0
7	0	2	1	0	0	1	0	1	1	0
8	0	0	0	1	0	0	0	0	0	1
9	1	0	2	1	0	0	0	0	0	0

Numbers in each core represent the deficit of tokens (the lower the better)

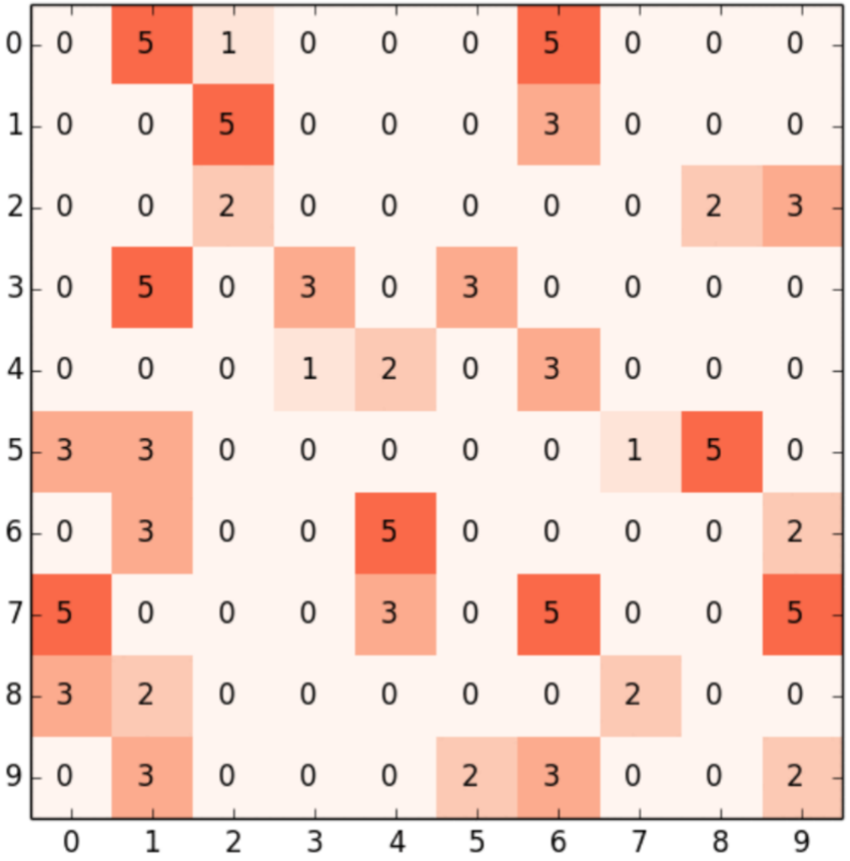
* *Animation frames taken every 100 simulation iterations* (animations won't show up in pdf, sorry!)

Token Shortfall Situations

(early-stage concept ModSim)

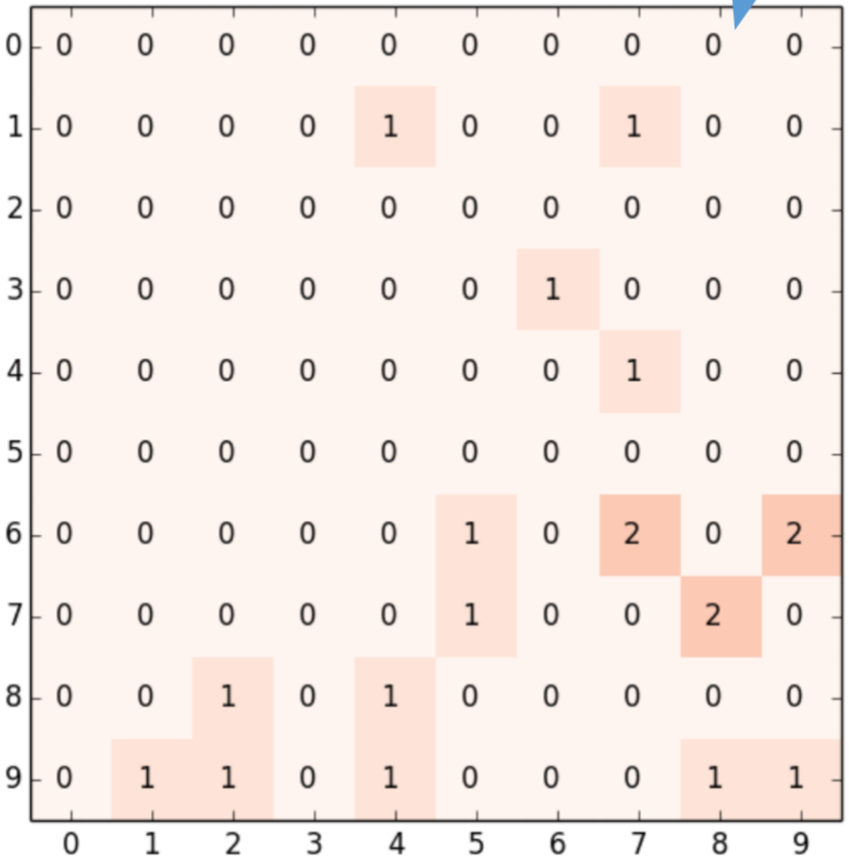
Numbers in each core represent the deficit of tokens (the lower the better)

Disabled*



	0	1	2	3	4	5	6	7	8	9
0	0	5	1	0	0	0	5	0	0	0
1	0	0	5	0	0	0	3	0	0	0
2	0	0	2	0	0	0	0	0	2	3
3	0	5	0	3	0	3	0	0	0	0
4	0	0	0	1	2	0	3	0	0	0
5	3	3	0	0	0	0	0	1	5	0
6	0	3	0	0	5	0	0	0	0	2
7	5	0	0	0	3	0	5	0	0	5
8	3	2	0	0	0	0	0	2	0	0
9	0	3	0	0	0	2	3	0	0	2

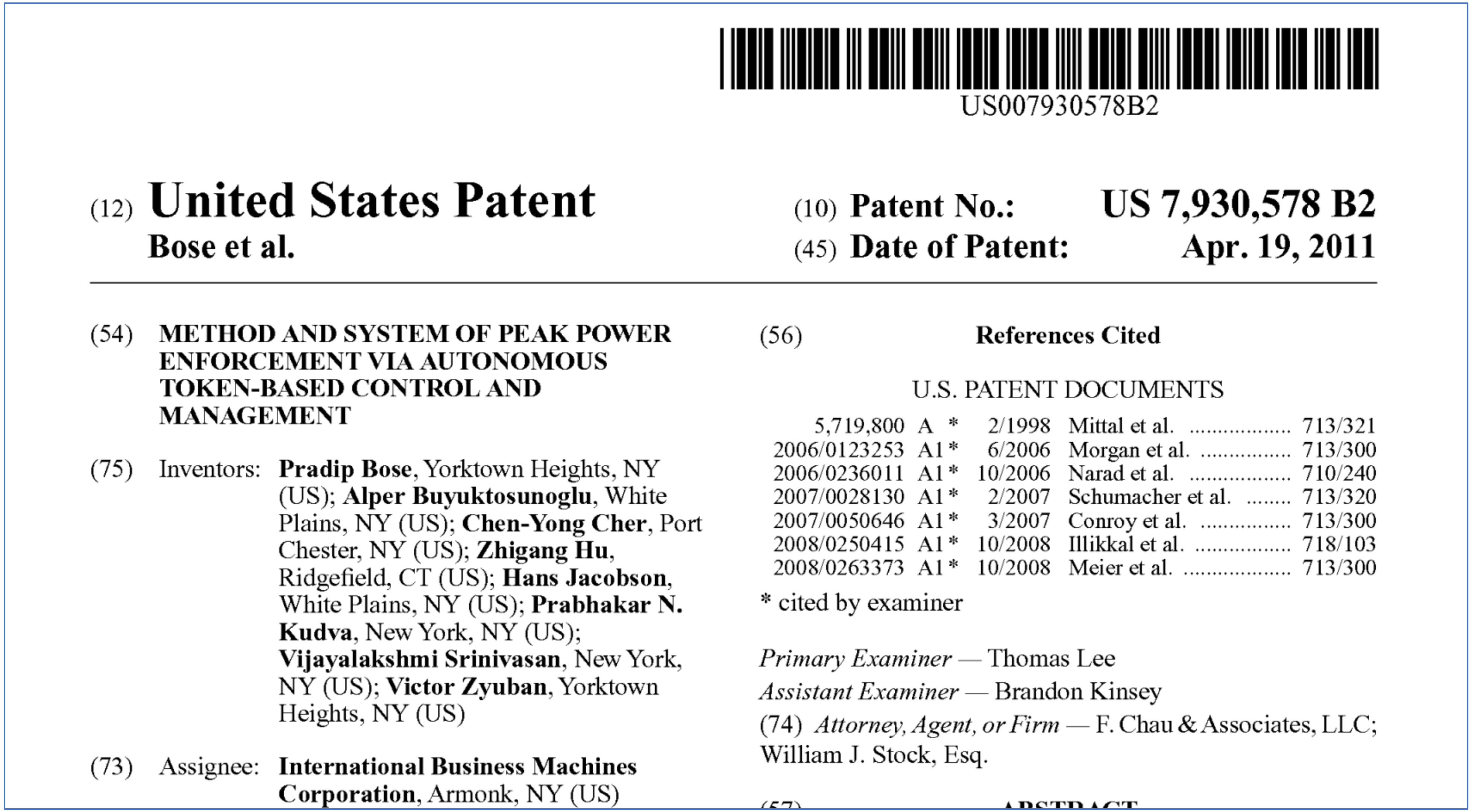
Enabled v4*



	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	1	0	0	1	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	1	0	0	0
4	0	0	0	0	0	0	0	1	0	0
5	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	0	2	0	2
7	0	0	0	0	0	1	0	0	2	0
8	0	0	1	0	1	0	0	0	0	0
9	0	1	1	0	1	0	0	0	1	1

** Animation frames taken every 100 simulation iterations* (animations won't show up in pdf, sorry!)

Early Interest in Token-Based Power Management



US007930578B2

(12) **United States Patent**
Bose et al.

(10) **Patent No.:** **US 7,930,578 B2**
(45) **Date of Patent:** **Apr. 19, 2011**

(54) **METHOD AND SYSTEM OF PEAK POWER ENFORCEMENT VIA AUTONOMOUS TOKEN-BASED CONTROL AND MANAGEMENT**

(75) Inventors: **Pradip Bose**, Yorktown Heights, NY (US); **Alper Buyuktosunoglu**, White Plains, NY (US); **Chen-Yong Cher**, Port Chester, NY (US); **Zhigang Hu**, Ridgefield, CT (US); **Hans Jacobson**, White Plains, NY (US); **Prabhakar N. Kudva**, New York, NY (US); **Vijayalakshmi Srinivasan**, New York, NY (US); **Victor Zyuban**, Yorktown Heights, NY (US)

(73) Assignee: **International Business Machines Corporation**, Armonk, NY (US)

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,719,800	A *	2/1998	Mittal et al.	713/321
2006/0123253	A1 *	6/2006	Morgan et al.	713/300
2006/0236011	A1 *	10/2006	Narad et al.	710/240
2007/0028130	A1 *	2/2007	Schumacher et al.	713/320
2007/0050646	A1 *	3/2007	Conroy et al.	713/300
2008/0250415	A1 *	10/2008	Illikkal et al.	718/103
2008/0263373	A1 *	10/2008	Meier et al.	713/300

* cited by examiner

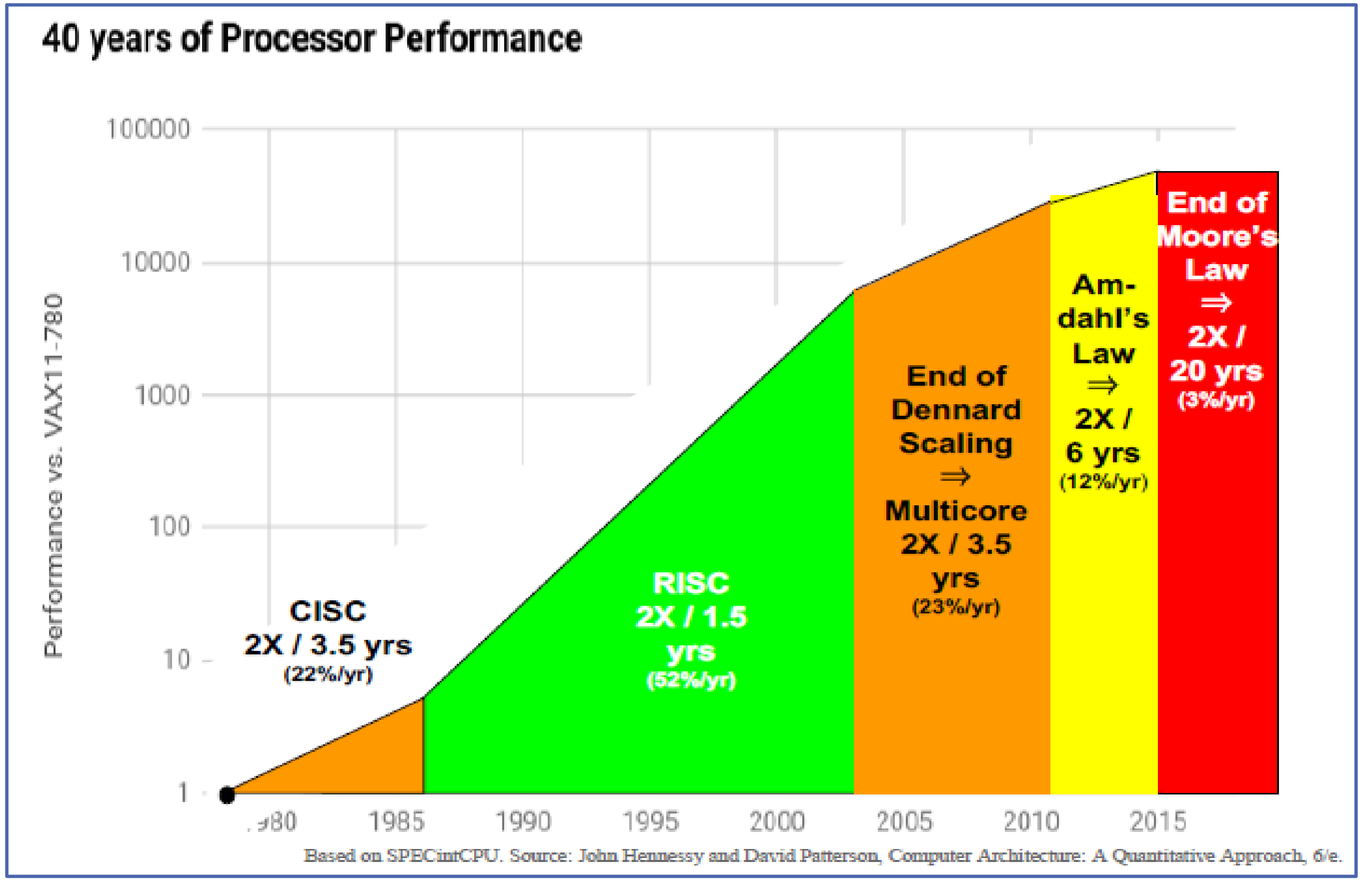
Primary Examiner — Thomas Lee
Assistant Examiner — Brandon Kinsey
(74) *Attorney, Agent, or Firm* — F. Chau & Associates, LLC; William J. Stock, Esq.

(57) ABSTRACT

But what about security?

DSSoC was not just an edge vision or strategy – it applied to server/cloud as well!

In the late CMOS era, domain specific accelerators will dominate



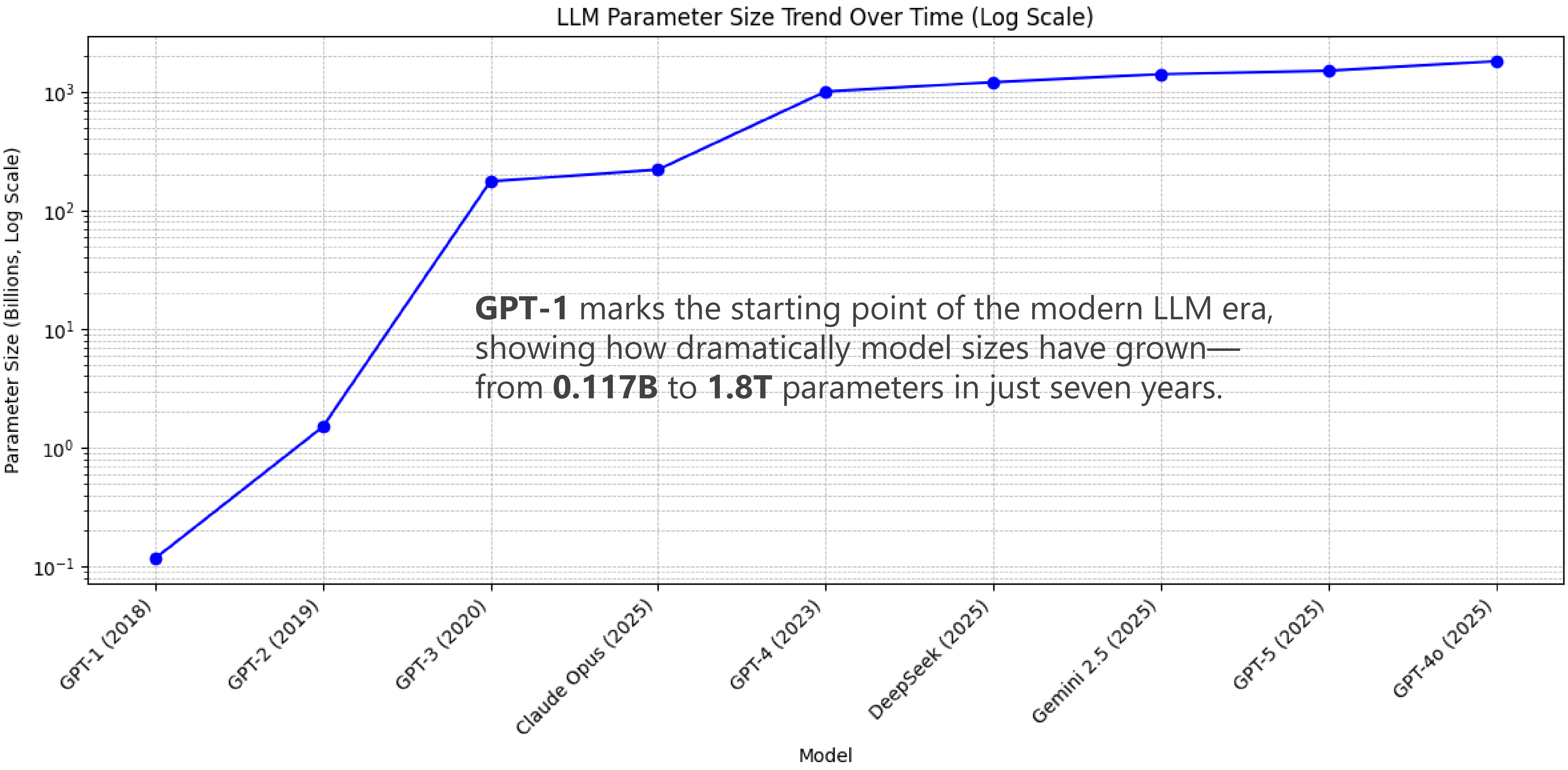
Based on SPECintCPU. Source: John Hennessy and David Patterson, Computer Architecture: A Quantitative Approach, 6/e.

- Primary server refresh at data center may be progressively delayed
- Differentiation (feature, performance) via domain-specific accelerators
- AI as a domain – changes (scales up) at an astounding rate → see next slide!

Agile hardware-software accelerator system synthesis is key to retaining customer base

- Learn customer workloads
- Design plug-in accelerator offerings; refresh choices every 6 mos.
- Highly automated design flow ➔ **small team**

Case in Point: Large Language Model (LLM) Parameter Growth Over Time*



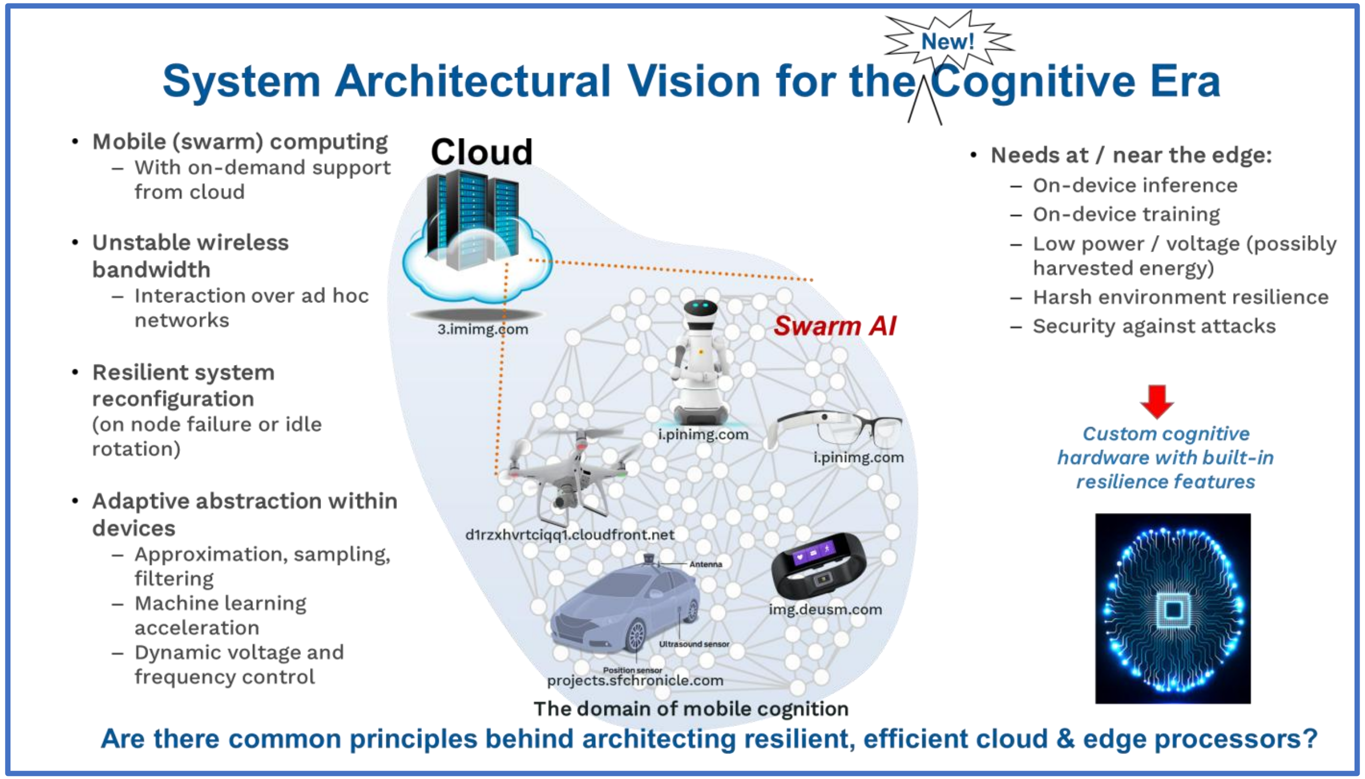
GPT-1 marks the starting point of the modern LLM era, showing how dramatically model sizes have grown—from **0.117B** to **1.8T** parameters in just seven years.

* Plot generated by Microsoft Co-Pilot

DARPA-hard Challenges:

a good way of pushing the envelope in systems R&D

⟹ Onward to Data Security



System Architectural Vision for the Cognitive Era

- Mobile (swarm) computing
  - With on-demand support from cloud
- Unstable wireless bandwidth
  - Interaction over ad hoc networks
- Resilient system reconfiguration (on node failure or idle rotation)
- Adaptive abstraction within devices
  - Approximation, sampling, filtering
  - Machine learning acceleration
  - Dynamic voltage and frequency control

The domain of mobile cognition

- Needs at / near the edge:
  - On-device inference
  - On-device training
  - Low power / voltage (possibly harvested energy)
  - Harsh environment resilience
  - Security against attacks

Custom cognitive hardware with built-in resilience features

Are there common principles behind architecting resilient, efficient cloud & edge processors?

- Agile SoC
- Programmability

- Data security
- Privacy

PERFECT [Bob Colwell, Joe Cross, …] →

2013 – 2018

Power Efficiency Revolution for Embedded Computing Technologies

1 GF/W → 75 GF/W

IBM + Stanford, Harvard, U of Virginia

DSSoC [Tom Rondeau] →

2018 → 2023/ongoing

Domain-Specific System on Chip

Power-perf, programmability, productivity metrics

IBM + Columbia, Harvard, UIUC

[Tom Rondeau]

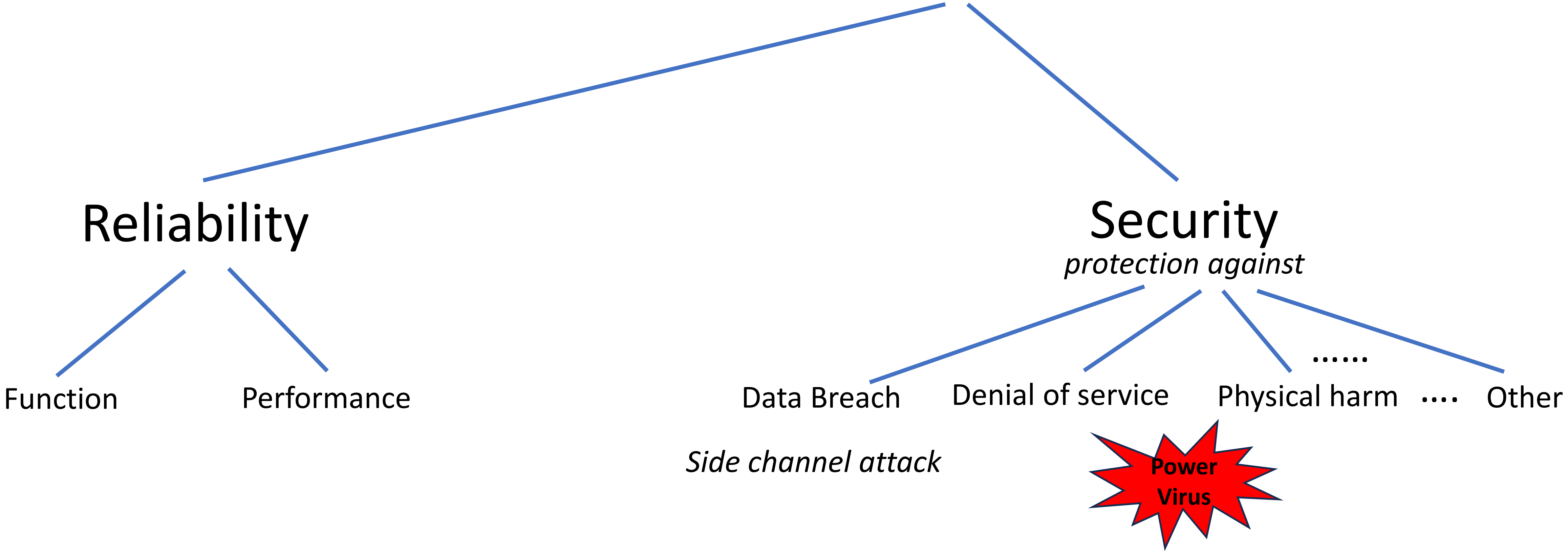
DPRIVE

2021 → ongoing

(IBM was not part of DPRIVE; but we pursued the same goal, 2022-2025 w/support from DoD/RAMP-C), *IBM + Columbia*

Beyond the DARPA DSSoC program .....

Next phase of R&D: worrying about DSSoC resilience at affordable power cost



Reliability

Function

Performance

Security

protection against

Data Breach

Denial of service

......

Physical harm

....

Other

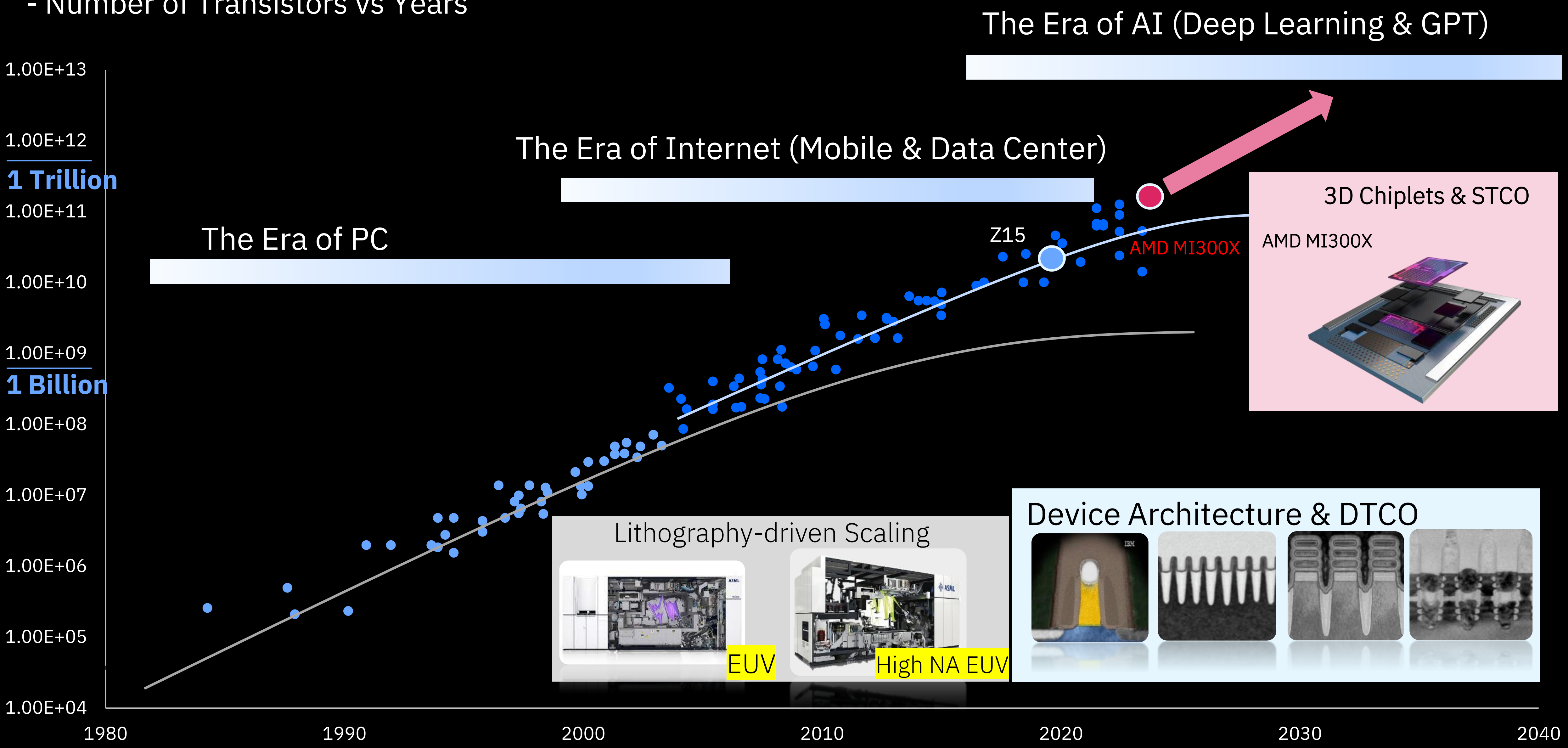
Side channel attack

Power Virus

... in the context of emerging trends in semiconductor and packaging technology

Technology Path to 1 Trillion Transistors

- Number of Transistors vs Years



The Era of AI (Deep Learning & GPT)

The Era of Internet (Mobile & Data Center)

The Era of PC

3D Chiplets & STCO

AMD MI300X

Device Architecture & DTCO

Lithography-driven Scaling

EUV

High NA EUV

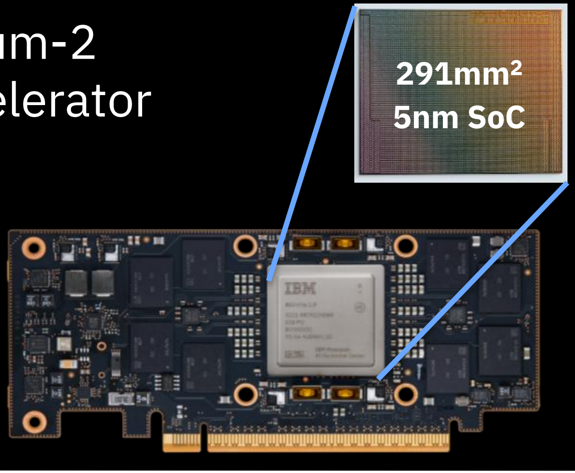
IBM AIU: Roadmap of Foundation Model AI accelerators

Key technology and enablement needs:

- State-of-the-art foundry CMOS
- State-of-the-art silicon-verified IP blocks for support functions (memory controllers, I/O interfaces)
- Chiplets and 3D stacking

AIU 1.0[1]

IBM z System Telum-1, Telum-2 announcements; Spyre accelerator at Hot Chips 2022, 2023



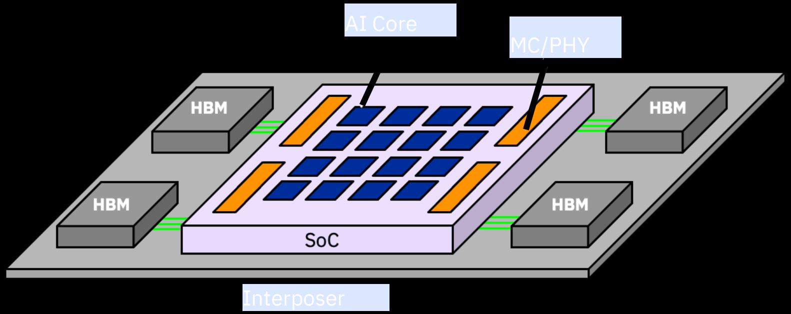
Optimized for FM Inference

[1] - Announced October 2022

AIU 1.0+

Optimized for FM Inference and Fine-Tuning, + Training

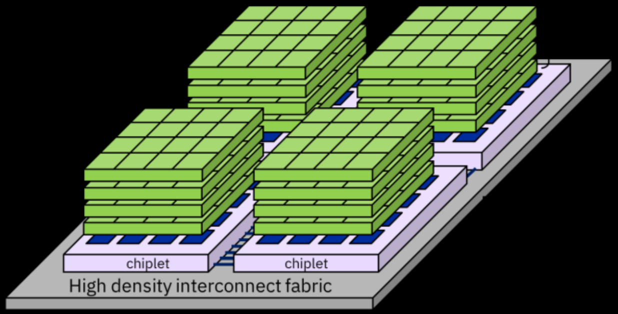
Leverage HBM



AIU Next

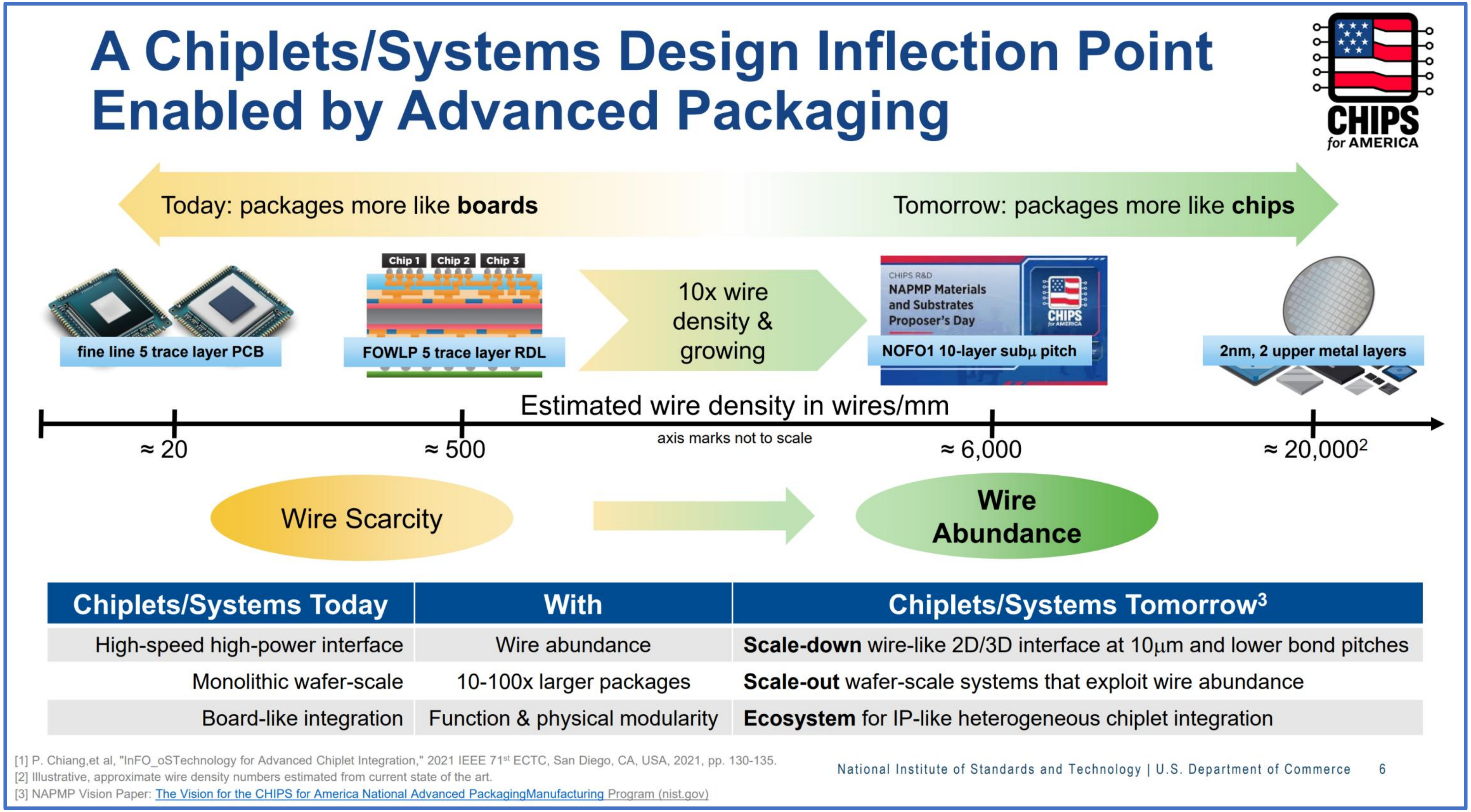
Optimized for future very large FM Inference + Fine-Tuning + Training

Leverage 3D-stacked memory + chiplet technologies



Huiming Bu, IBM IEEE CAS/EDS ***AI Compute Symposium 2023***

CHIPS-Act Linked NAPMP Funding Opportunity



A Chiplets/Systems Design Inflection Point Enabled by Advanced Packaging

CHIPS for AMERICA

Today: packages more like **boards**

Tomorrow: packages more like **chips**

fine line 5 trace layer PCB

Chip 1 Chip 2 Chip 3

FOWLP 5 trace layer RDL

10x wire density & growing

CHIPS R&D NAPMP Materials and Substrates Proposer's Day

NOFO1 10-layer subμ pitch

2nm, 2 upper metal layers

Estimated wire density in wires/mm

axis marks not to scale

≈ 20 ≈ 500 ≈ 6,000 ≈ 20,000[2]

Wire Scarcity

Wire Abundance

Chiplets/Systems Today	With	Chiplets/Systems Tomorrow[3]
High-speed high-power interface	Wire abundance	**Scale-down** wire-like 2D/3D interface at 10μm and lower bond pitches
Monolithic wafer-scale	10-100x larger packages	**Scale-out** wafer-scale systems that exploit wire abundance
Board-like integration	Function & physical modularity	**Ecosystem** for IP-like heterogeneous chiplet integration

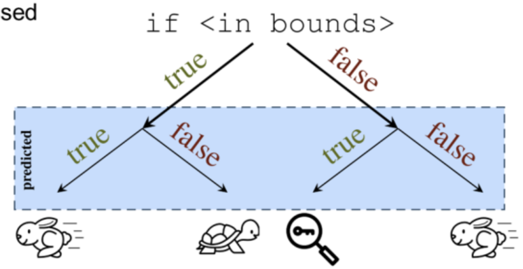
[1] P. Chiang,et al, "InFO_oSTechnology for Advanced Chiplet Integration," 2021 IEEE 71st ECTC, San Diego, CA, USA, 2021, pp. 130-135.
[2] Illustrative, approximate wire density numbers estimated from current state of the art.
[3] NAPMP Vision Paper: The Vision for the CHIPS for America National Advanced PackagingManufacturing Program (nist.gov)

National Institute of Standards and Technology | U.S. Department of Commerce 6

From the Proposers' Day slideset

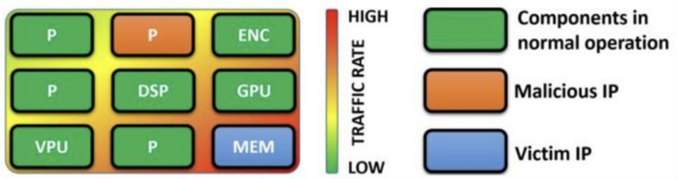
Threats to AI hardware

- Side channel attacks
  - Extract sensitive information (e.g., data, model parameters) using hardware side channels (timing, power, etc.)
  - E.g., cache-based side channels like Spectre [Kocher et al., S&P 2019] and Meltdown [Lipp et al., USENIX 2018] can be used to extract data regularly accessed by a model



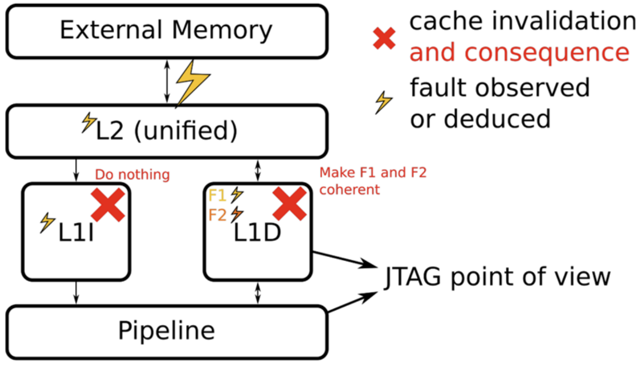
[Kocher et al., S&P 2019]

- Hardware trojans
  - Insert malicious hardware at design time to impact AI functionality
  - E.g., hardware trojan hidden in a unit of an SoC can launch a denial-of-service attack when triggered that prevents AI model from continuing computations



[Charles et al., DATE 2019]

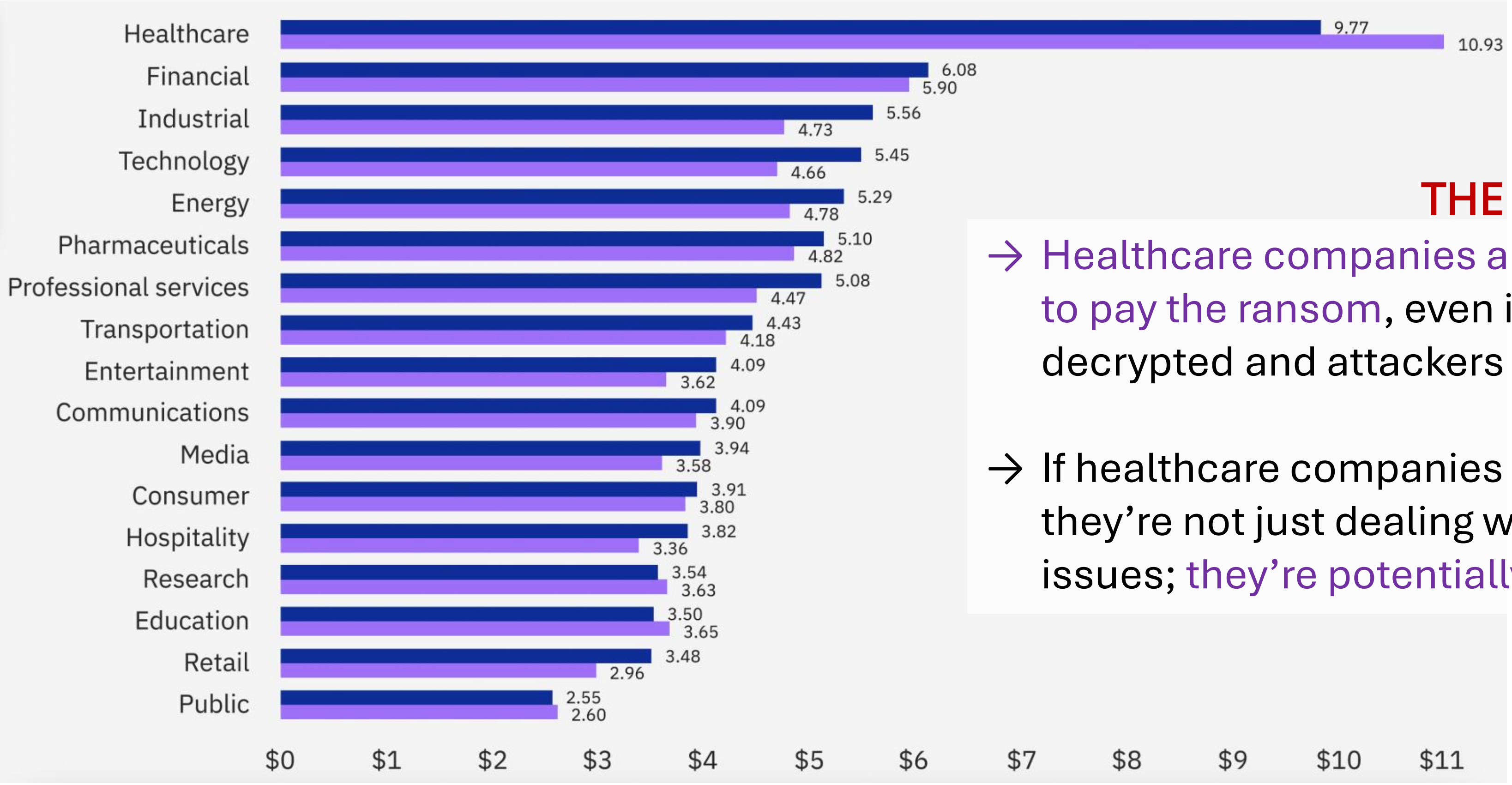
- Physical attacks
  - E.g., laser or voltage manipulation to alter system behavior and functionality
  - [Trouchkine et al., CoRR 2019] showed electromagnetic fault injection attacks can be used to target individual subsystems within an SoC



[Trouchkine et al., CoRR 2019]

The Need for Privacy-Aware Computing

Cost of a data breach by industry (in USD millions)



■ 2024 ■ 2023

Source: https://www.ibm.com/think/insights/healthcare-industry-attack-trends-2024

THE RANSOMWARE CASE

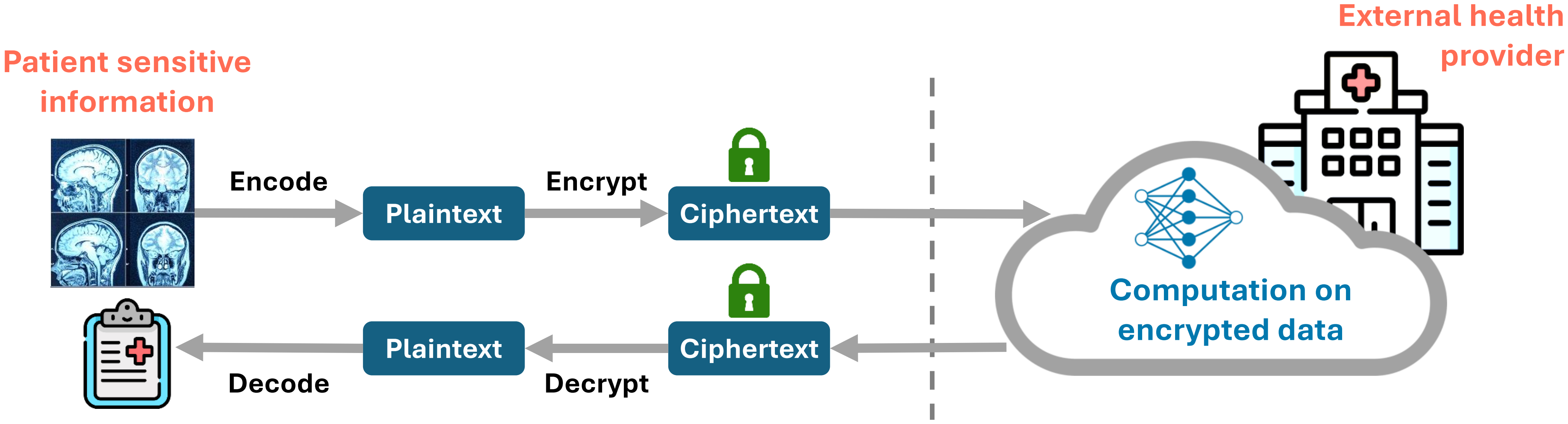
→ Healthcare companies are more likely than other industries to pay the ransom, even if there's no guarantee data will be decrypted and attackers won't try again

→ If healthcare companies hold the line and refuse to pay, they're not just dealing with financial and operational issues; they're potentially putting patients at risk

What is Homomorphic Encryption (HE)?

- Cryptographic technique that enables **processing and manipulation of encrypted data**
- Traditional crypto algorithms require data to be unencrypted for processing

HOMOMORPHIC ENCRYPTION PIPELINE

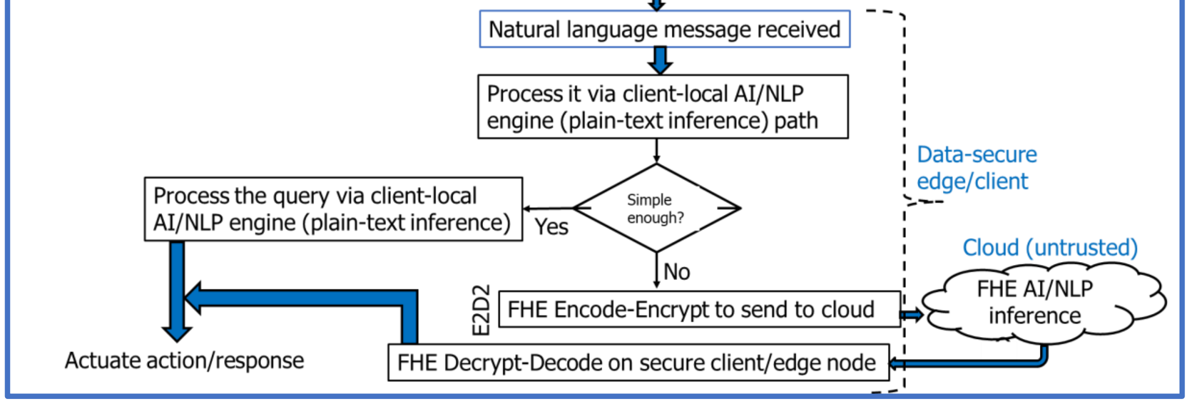


Real-life Business Use Cases

Real-life business use case (1)

1) Call center message handling automation: customer privacy-protected call center voice/text/data traffic handling

- Process voice/email queries: automated handling where possible
- Protect data privacy for cloud-hosted inferencing, analytics if applicable



Fetch next message (group/batch) from queue

Natural language message received

Process it via client-local AI/NLP engine (plain-text inference) path

Process the query via client-local AI/NLP engine (plain-text inference)

Data-secure edge/client

Cloud (untrusted)

FHE Encode-Encrypt to send to cloud

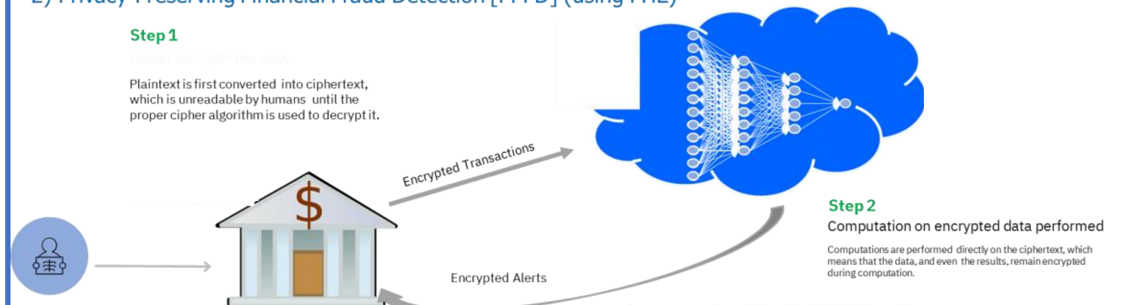
FHE AI/NLP inference

Actuate action/response

FHE Decrypt-Decode on secure client/edge node

Real-life business case (2)

2) Privacy-Preserving Financial Fraud Detection [PPFD] (using FHE)



Step 1

Owner encrypts the data

Plaintext is first converted into ciphertext, which is unreadable by humans until the proper cipher algorithm is used to decrypt it.

Step 2

Computation on encrypted data performed

Computations are performed directly on the ciphertext, which means that the data, and even the results, remain encrypted during computation.

Step 3

Owner unlocks data and results

The encrypted results are then sent back to the data owner. Data and results can only be decrypted by that data owner, so security is never compromised.

Latest FHE version		128-bit security			256-bit security		
		24-cpu	88-cpu	GPU	24-cpu	88-cpu	GPU
predict (server)		792.02	509.3	26.59	1177.6	1106.1	42.49
one-time setup	context	733	806	1797	1881	1416	3688
	model	510	138	124	248	236	262
Encrypt (client)		50.56	20.22	34.12	38.78	42.33	72.43
Decrypt (client)		3.47	2.19	1	5.22	4.67	2.1
RAM used (MB)		4452	947	966	1623	1656	1575
Input ciphertext (MB)		8.39			16.78		
Output ciphertext (MB)		0.26			0.52		

F1=~0.68

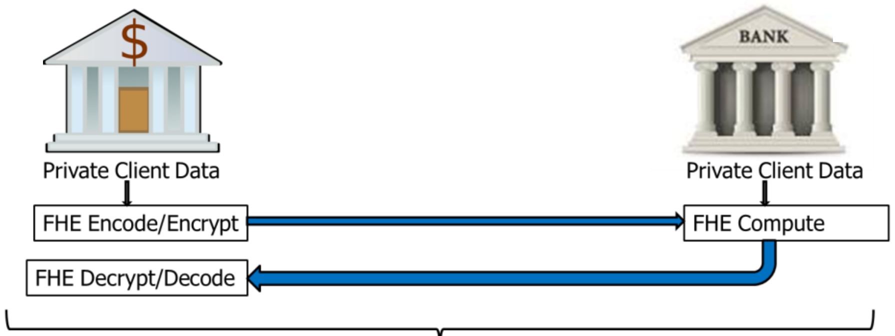
The measured results show that the E2D2 segment of PPFD on the on-prem server system in the bank is in itself a key bottleneck that is amenable to H/W acceleration

Early FHE version		128-bit security			256-bit security		
		24-cpu	88-cpu	GPU	24-cpu	88-cpu	GPU
predict (server)		255.84	243.91	18.31	531.37	498.52	30.56
one-time setup	context	615	529	1385	1451	1112	2405
	model	103	100	81	223	181	163
Encrypt (client)		16.55	17.04	28.09	31.7	33.52	57.63
Decrypt (client)		2.98	2.37	1	5.99	5.13	2.04
RAM used (MB)		596	599	647	1021	1018	1015
Input ciphertext (MB)		6.29			12.58		
Output ciphertext (MB)		0.26			0.52		

F1=~0.50

Real-life business case (3)

3) z2z (server to server without Cloud involvement)



Private Client Data

Private Client Data

FHE Encode/Encrypt

FHE Compute

FHE Decrypt/Decode

And, vice versa, in terms of cooperative computing with privacy-protected data sharing

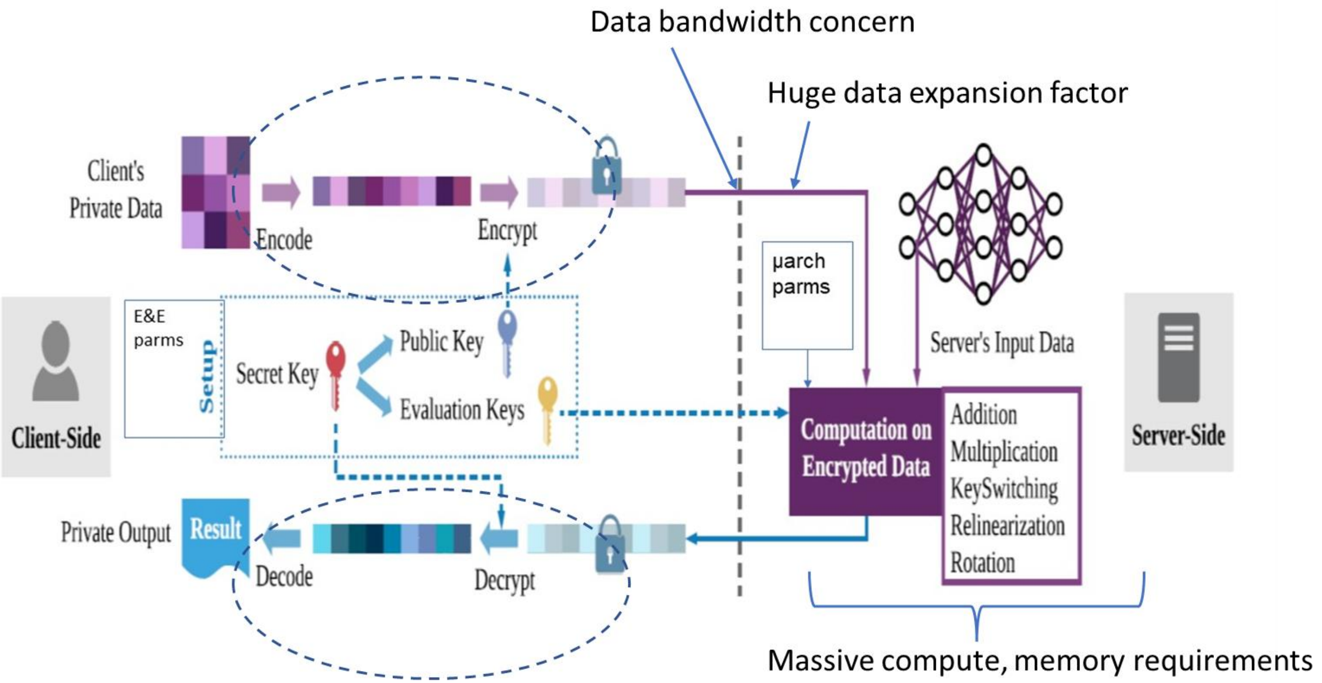
Both banks in z2z need E2D2 and FHE Compute capability

- There are many other use cases of course.
- The ones mentioned are representative of IBM's core mainframe business in the financial sector
- The connected autonomous vehicle (CAV) edge sector remains a major area of interest as well

AI/FHE Motivation: DARPA-hard Challenge (hardware acceleration)

The larger context

Data-Secure Computing: the End-to-End Picture



The Encode-Encrypt, Decrypt-Decode (E2D2) client-side task is also important!

- Poster child application to utilize the emerging trends in semiconductor and packaging technology (e.g. chiplets/3DHI)
- Aligned with semiconductor business strategy; linked also to the government’s CHIPS ACT related thrusts
- 1000x – 500000x acceleration needed to meet performance (e.g. real-time) needs, as addressed in the DARPA DPRIVE program

We started up this project in January 2022 with the above challenge and business opportunity in mind

- The initial (primary) focus is on AI-embedded transactional workloads like financial fraud detection (FFD).
- But there are many other edge-cloud application workloads (with privacy-protection needs) that map into this space (e.g. the cloud-backed connected autonomous vehicular space just mentioned).

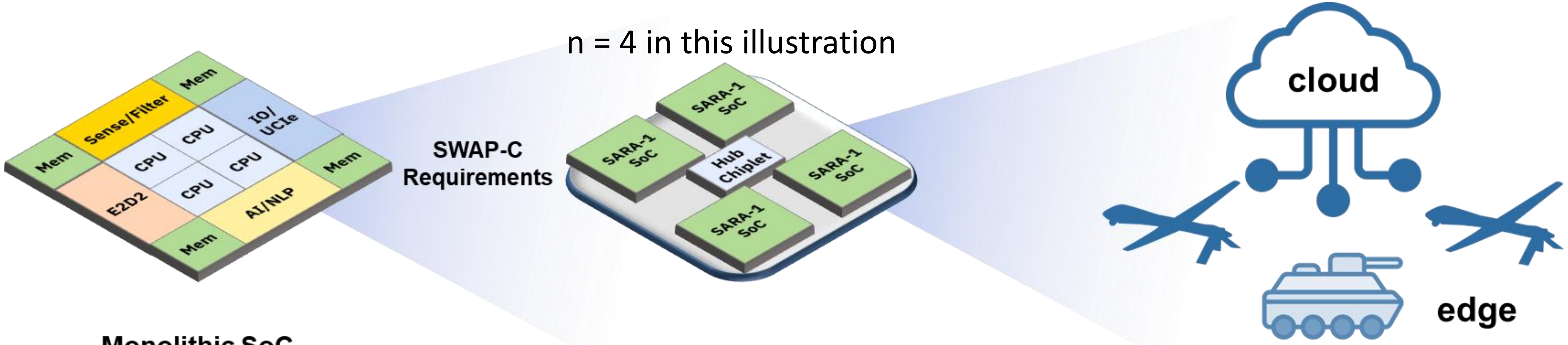
AI/FHE Hardware Acceleration: Enabling Privacy-Protected Edge-to-Cloud AI Computation

Our design strategy and long-term research vision:

1. Leverage our prior agile SoC design methodology ***(EPOCHS)*** to implement a AI/FHE SoC (chiplet) in order to demonstrate basic viability for a class of AI-centric inference workloads with an n-chiplet SiP solution (where n = 1, 2 or 4 in the first generation)
   - ✓ Individual chiplet size (area) is determined by yield (cost) constraints for a new technology node
   - ✓ Integrated UCIe interface allows scaled-up system solution with multiple chiplets and on-package memory modules (DDR or HBM)

2. *Scalable solution:* start with an edge E2D2 capability, scale up to a cloud AI/FHE compute capability
   - ▪ Leverage 3DHI and chiplet technology to meet memory capacity and bandwidth requirements

n = 4 in this illustration

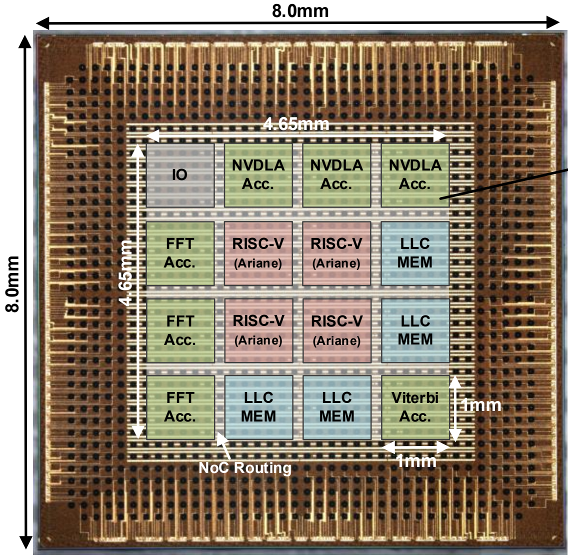


Monolithic SoC (SARA-1)

Chiplet Design (SARA-2) (2.5D Packaging)

Hub chiplet provides connectivity to host via PCIe and memory via HBM and/or DDR controllers

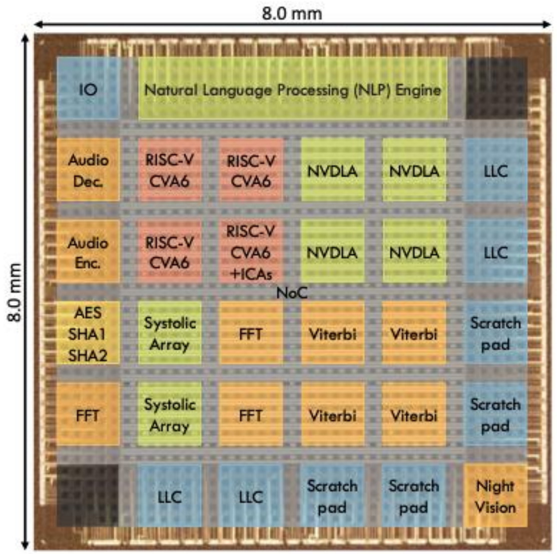
Circling back to emphasize the benefit of our agile SoC design methodology driven by Columbia's ESP: Impressive Productivity Gain



EPOCHS-0

Oct. 2020

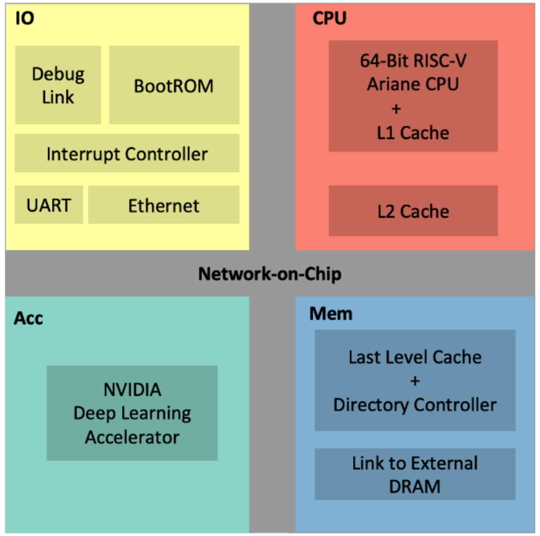
2 RTL/Verif. Engineers
6 PD Engineers



EPOCHS-1

Nov. 2022

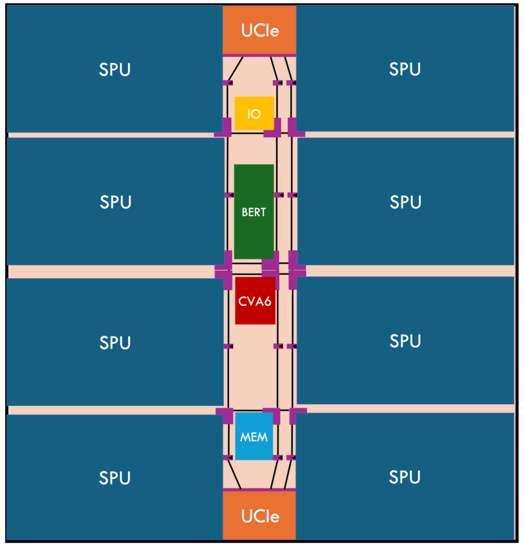
3 RTL/Verif. Engineers
6 PD Engineers


IO
Debug Link
BootROM
Interrupt Controller
UART
Ethernet
CPU
64-Bit RISC-V Ariane CPU + L1 Cache
L2 Cache
Network-on-Chip
Acc
NVIDIA Deep Learning Accelerator
Mem
Last Level Cache + Directory Controller
Link to External DRAM

Mini SoC

Jan. 2024

1 RTL/Verif. Engineer
6 PD Engineers



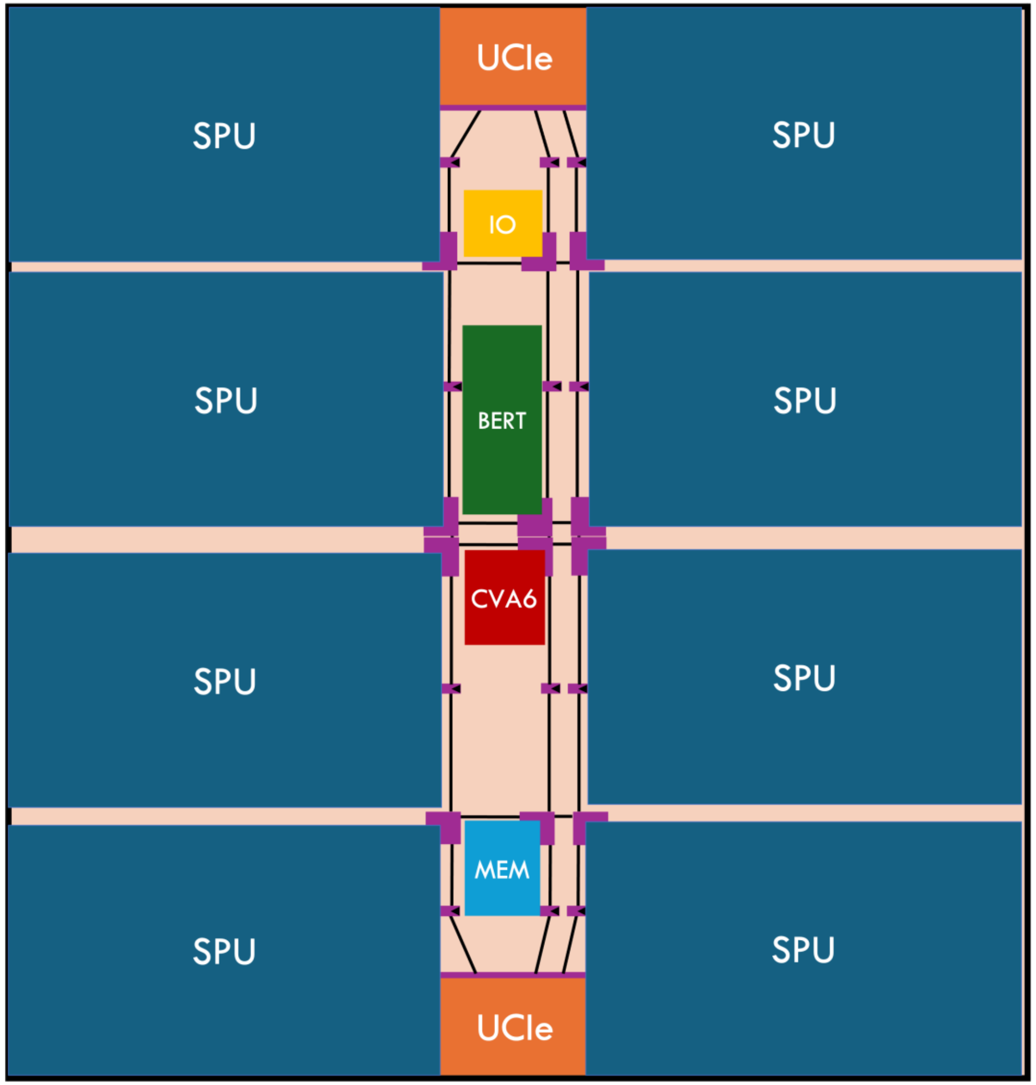
SARA-1

Sep. 2025

9 RTL/Verif. Engineers
6 PD Engineers

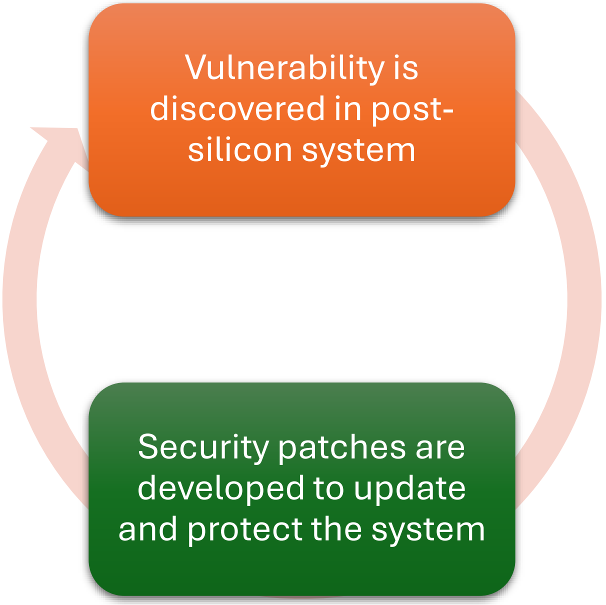
SARA-1

- Same advanced technology node
- 8 copies of the SARA Processing Unit (SPU)
  - Programmable accelerator
  - Large (>10mm^2)
- Application has complex data-dependency patterns
  - One-to-many communication
- Chiplet-based w/ UCIe
- Heterogeneous tile sizes complicate physical design
- Design completes in September 2025



Motivation for pre-silicon security analysis

Typical security cycle


Vulnerability is discovered in post-silicon system
Security patches are developed to update and protect the system

- Can we prioritize security during early design cycle stages?
  - Power, area, and performance are prioritized but can we include security in these early design considerations as well?
  - How can vulnerabilities be found in pre-silicon stages?
  - How can security be measured?
  - Can security evaluation be automated early in the design cycle to ease designer effort?

However, it can be expensive and less effective to implement mitigation solutions in post-silicon stages of design or after production.

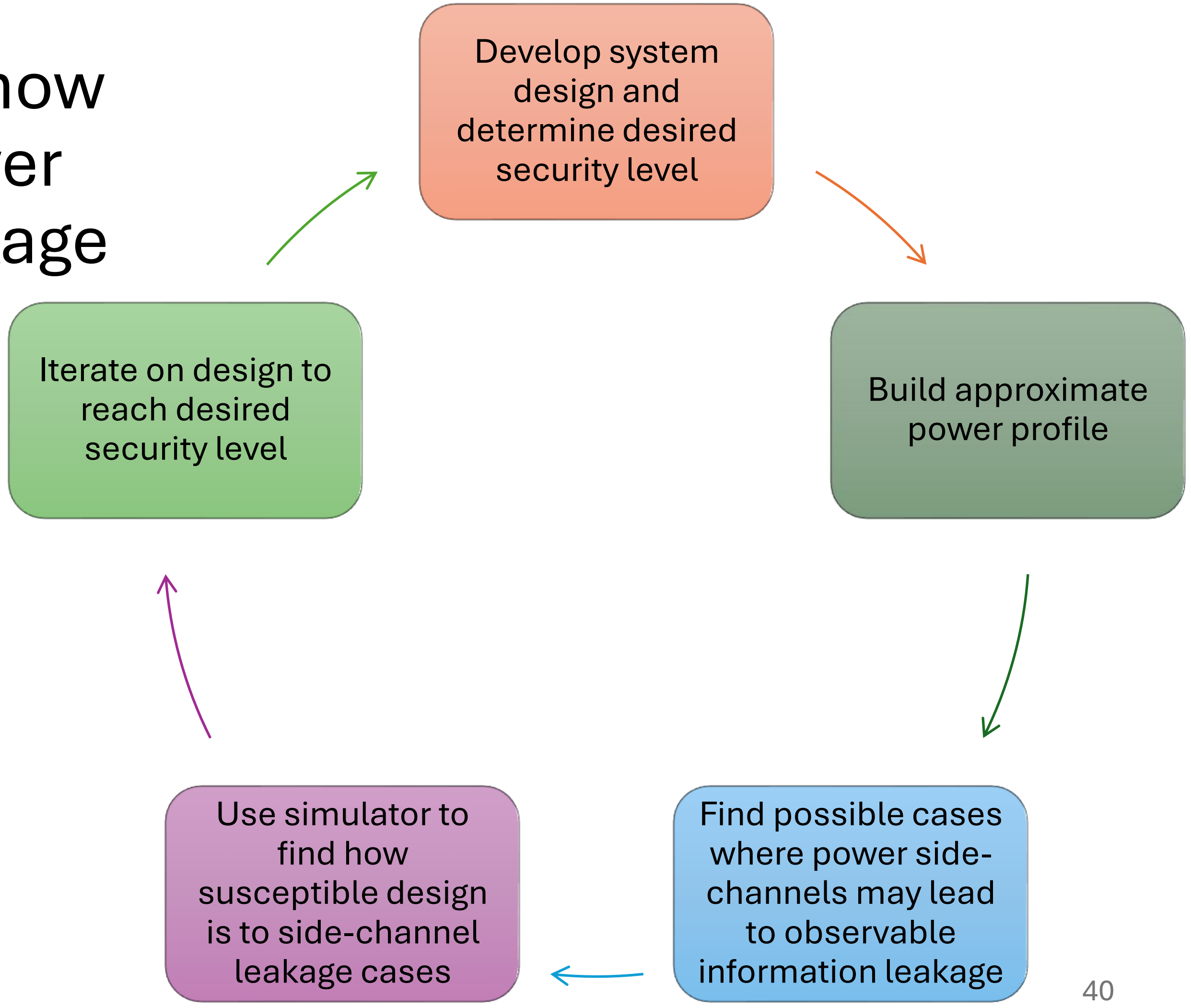
Focus on side channel analysis from early design stages

Pre-silicon: design cycle stages prior to taping out chip
Post-silicon: taped out chip stages

Naorin Hossain et al., ICMC-2024

Our approach to designing systems with power side channel leakage possibility in mind

- Develop a **metric** to quantify how susceptible a design is to power side-channel information leakage
  - E.g., 0 to 1 values
    - 0 = no information leakage
    - 1 = fully vulnerable



Develop system design and determine desired security level

Build approximate power profile

Find possible cases where power side-channels may lead to observable information leakage

Use simulator to find how susceptible design is to side-channel leakage cases

Iterate on design to reach desired security level

Naorin Hossain et al., ICMC-2024

Benefits of our approach

Our ideas can be applied based on the needs of the system

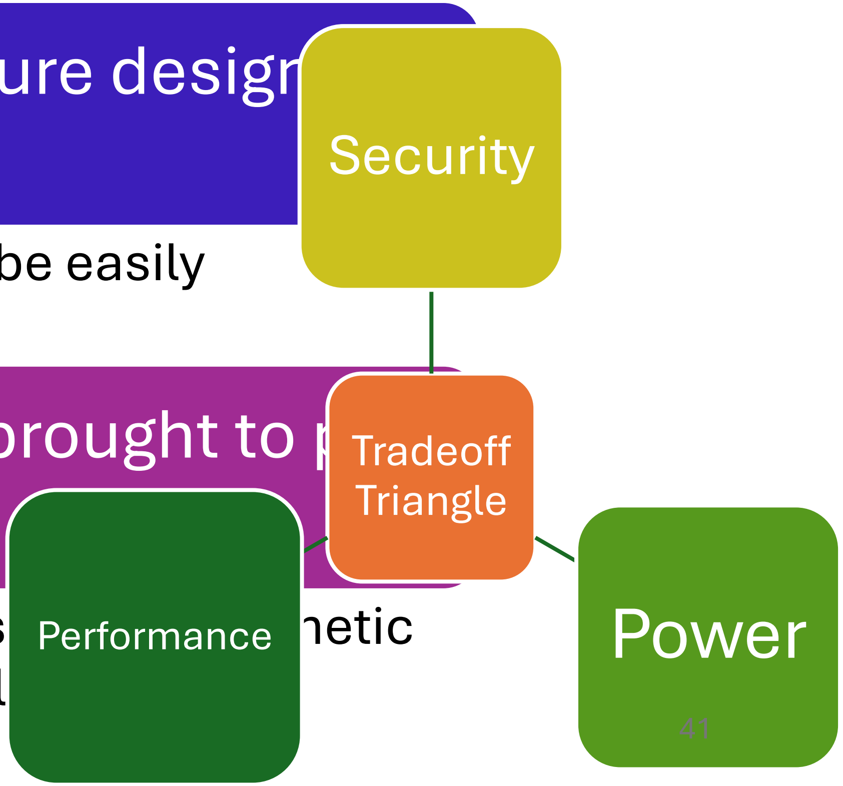
- E.g., systems handling sensitive data may prioritize secure design over performance

Our approach improves and eases secure desig efforts

- E.g., side-channel vulnerability evaluation can be easily automated

Other security considerations may be brought to silicon design stages

- E.g., other physical design phenomena such as netic emanations may be similarly modeled and eval



Security

Tradeoff Triangle

Performance

Power

Naorin Hossain et al., ICMC-2024

Executive Summary of ModSim Challenges Faced

(across the three govt-sponsored R&D projects)

Don’t forget the SDC scare!

1. Design Verification (and Test!)
   - Architects woefully lack tools and metrics to gauge verification complexity in pre-silicon modeling
   - *Agile SoC design* claims avoid factoring in verification time
2. Robust Power Management
   - On-chip, workload-driven power management architectures have become increasingly more advanced and sophisticated
   - But…ModSim-driven reliability & security guarantees are lacking
3. Security Metrics and Pre-Silicon Modeling
   - Largely absent! (Urgent need)

DFV/DFT and ModSim thereof

Deficiencies above cause shortfalls in system resilience and inhibit product quality deployment of devised solutions

PERFECT: Efficient Resilience
In Embedded Computing

SARA: Secure & Resilient AI

Thank You!

Pradip Bose, Augusto Vega, IBM T. J. Watson Research Center
Sarita Adve, Vikram Adve, Sasa Misailovic, University of Illinois at Urbana-Champaign
Luca Carloni, Ken Shepard, Columbia University
David Brooks, Gu-Yeon Wei, Vijay Janapa Reddi, Harvard University
Kevin Skadron, Mircea Stan, University of Virginia