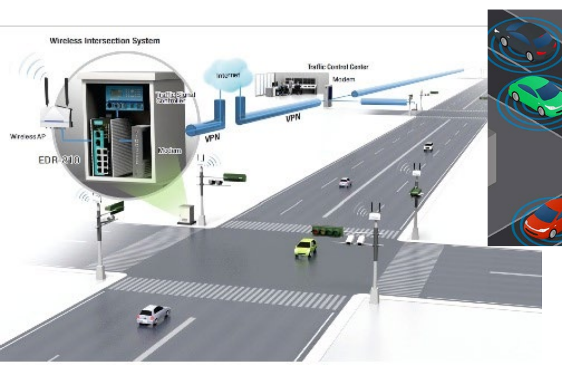
From Cross-Layer Resilience for On-Chip Systems to Robust Machine Learning

M. Shafique

Smart Cyber Physical Systems & Internet-of-Things



Smart Automobiles
http://www.it5g.com/latest-software-enhancements-in-the-auto-industry/

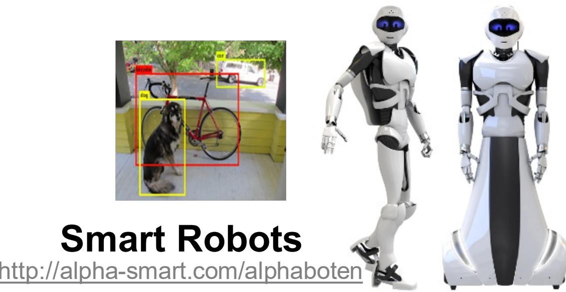
Smart Traffic Control
https://www.emaze.com/@ACIOWOWR/IMSA-Slide-Show

Smart Transport Systems
https://www.automotiveworld.com/analysis/automotive-cyber-physical-systems-next-computing-revolution/

Host Device
Body Gateway
PAN
WAN
Care Server
Internet
ADSL
Off-body sensors
User IP Box
Doctor/User PC

Smart Health Care

Applications



Smart Robots
http://alpha-smart.com/alphaboten

Industry 4.0:
Smart Industrial Automation
https://vimeo.com/145877805

Smart Houses
https://www.linkedin.com/pulse/smart-homes-private-secure-future-intelligent-home-tripti-jha

Smart Grids
http://solutions.3m.com/wps/portal/3M/en_EU/SmartGrid/EU-Smart-Grid/

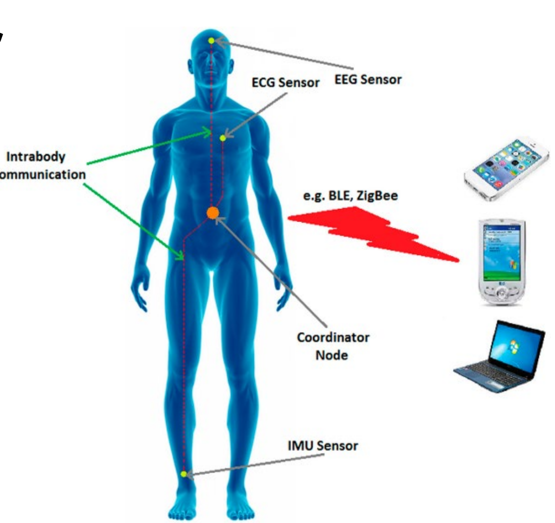
Smart CPS & IoT => The Big Data Processing Challenge!

… should consider

- ❑ **Performance**
  - ❑ Throughput
  - ❑ Latency

- ❑ **Robustness**
  - ❑ Reliability
  - ❑ Security

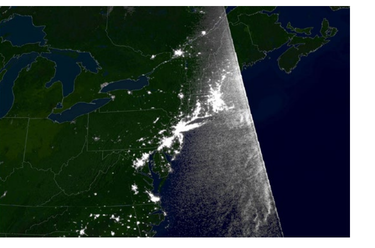
- ❑ **Others**
  - ❑ Adaptability
  - ❑ Safety
  - ❑ Privacy
  - ❑ Interoperability



Smart Healthcare
(Energy and time constraints)

Norwegian C-130 crash (2012)
https://en.wikipedia.org/wiki/2012_Norwegian_C-130_crash

Failure of F-22 Raptor (2007)
http://www.dailytech.com/Lockheeds+F22+Raptor+Gets+Zapped+by+International+Date+Line/article6225.htm



Satellite imagery of the Northeastern United-States taken before and during the blackout

Toronto, on the evening of August 14, 2003

Northeast blackout of 2003
https://en.wikipedia.org/wiki/Northeast_blackout_of_2003

Hacking Jeep Cherokee 4x4 (2015)

Sent the instructions through Entertainment systems

- Change the in-car temperature
- Control the steering
- Control the braking system

https://www.ophtek.com/4-real-life-examples-iot-hacked/
https://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/

Smart CPS & IoT => The Big Data Processing Challenge!

But,

Processing, Cybersecurity, and Reliability requires Power / Energy!

Minimal power footprint => Market Winner

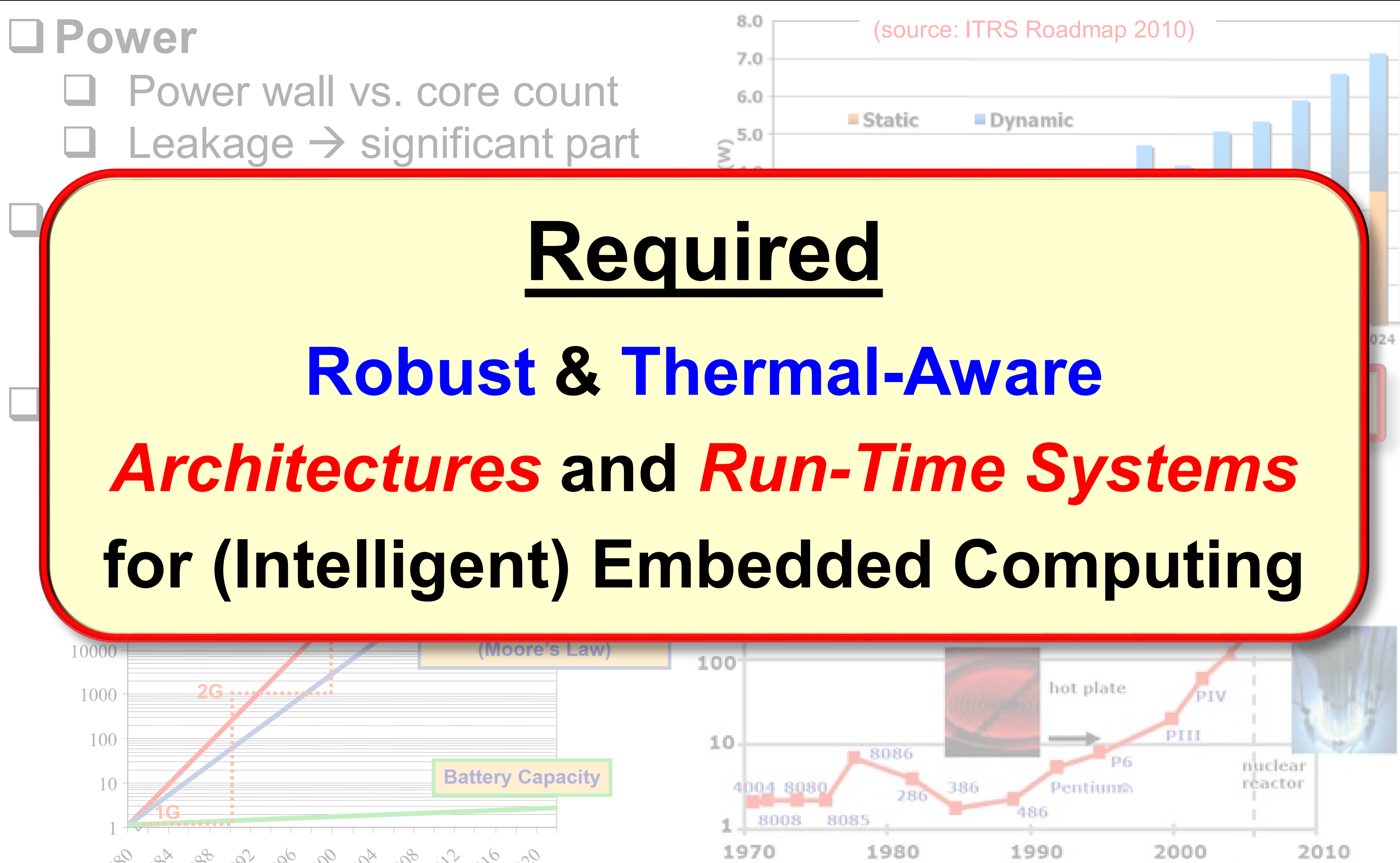
❑ Reliability

Challenging Question

How to process such huge amount of data in a robust yet energy-efficient way?

Why to care about Low-Power Computing?

Power is the Limiting Factor for Technology Scaling

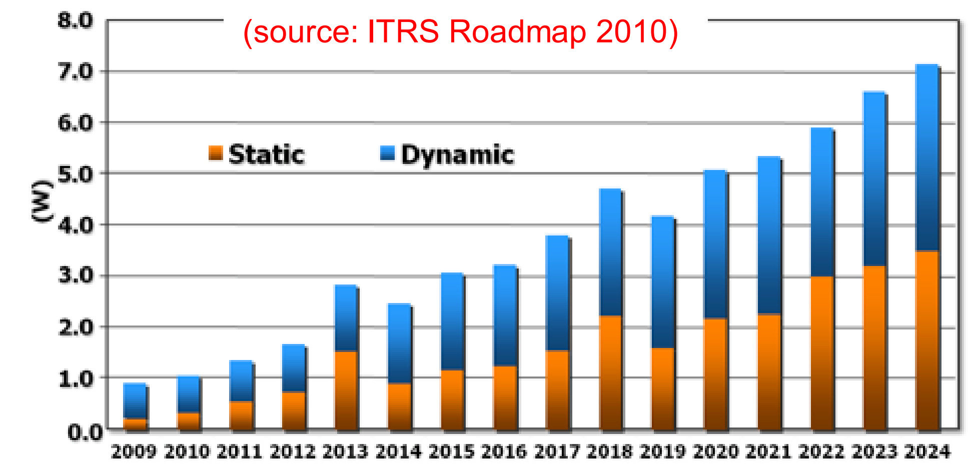


(src: A. Cuomo, ST Micro, Stockholm, Sept.8, 2004)

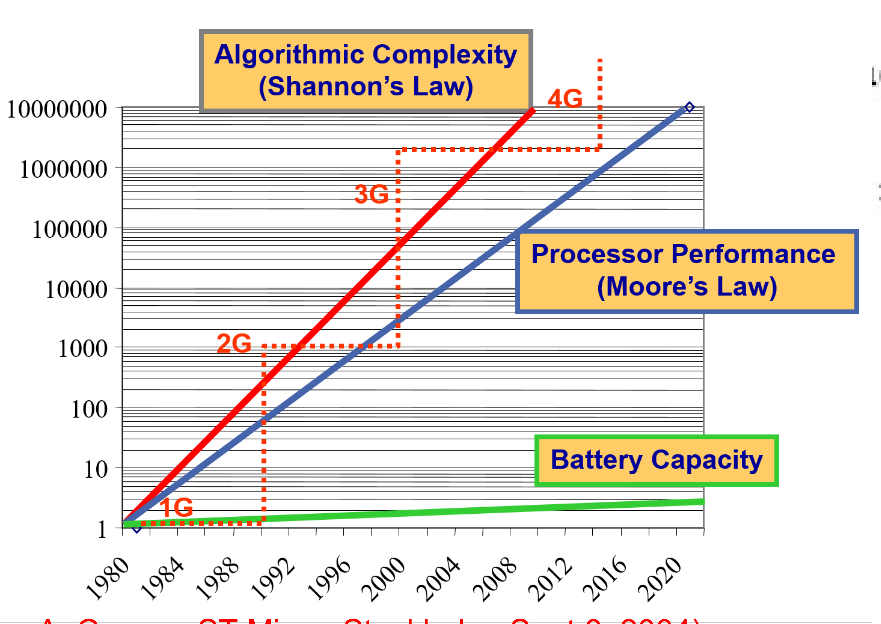
The growing power density (measured in W/cm²) of Intel's microchip processor families. (Source: Intel)

Why to care about Low-Power Computing?
Power is the Limiting Factor for Technology Scaling

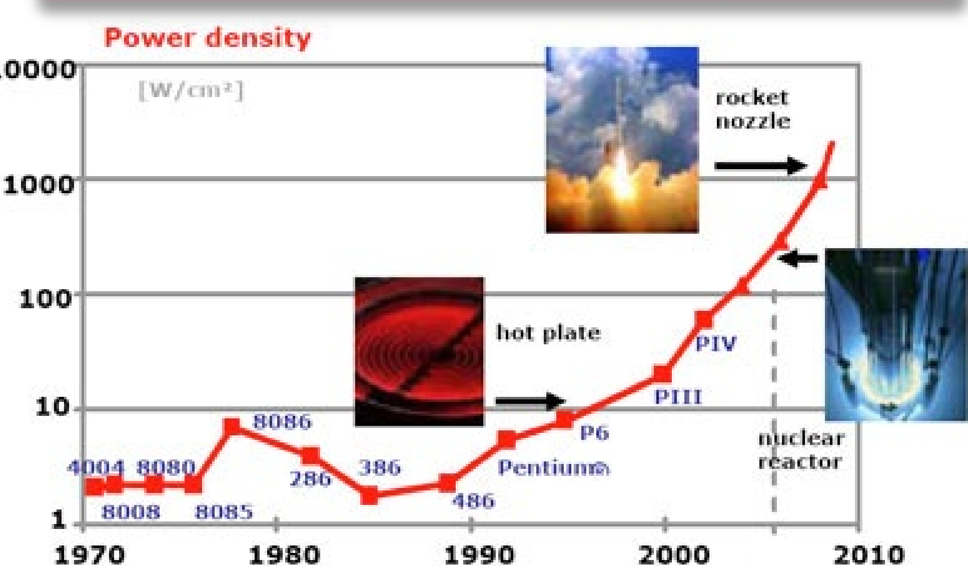
- ❑ **Power**
  - ❑ Power wall vs. core count
  - ❑ Leakage → significant part
- ❑ **Leads to high Temperature**
  - ❑ Dark Silicon Issue
  - ❑ Aggravates Reliability
- ❑ **Low Battery Lifetime**



Power density continues to get worse



(src: A. Cuomo, ST Micro, Stockholm, Sept.8, 2004)



The growing power density (measured in W/cm²) of Intel's microchip processor families. (Source: Intel)

Outline

- Different Types of Reliability Threats
- **Cross-Layer Resilience**
  - Modeling => Bridging the Gap between HW and SW
  - Optimization => Engage Multiple Layers of the System Stack
  - A Self-Healing Framework for Building Resilient CPS
  - Power / Temperature Considerations for Resilience
- **Robust Machine Learning**
- Conclusion

Technology Scaling

Transistor Count
10^{10}
10^9
10^8
10^7
10^6
10^5
10^4
10^3

Gordon E. Moore
(co-founded Intel in 1968)

GM 200 (Nvidia)
62 core Xeon Phi
Apple A8
8 core Xeon(Nehalem)
Power 6
Core M
AMD K10
Core i7 (quad)
Itanium 2
Core 2 Duo
AMD K8
Cell
Pentium IV
Barton
AMD K7
Atom
AMD K6
Pentium III
Pentium
Pentium II
i486
AMD K5
i386
80286
8086
8080
4004

Transistor count doubling every 18 months

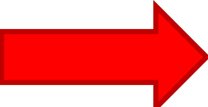
Src: Moore@Electronics'65, Intel, Nvidia, Chipworks

1970 1975 1980 1985 1990 1995 2000 2005 2010 2015
Year

[src: 3DTV-LX9500, LG, Jan Madsen DTU , Mercedes, KIT]

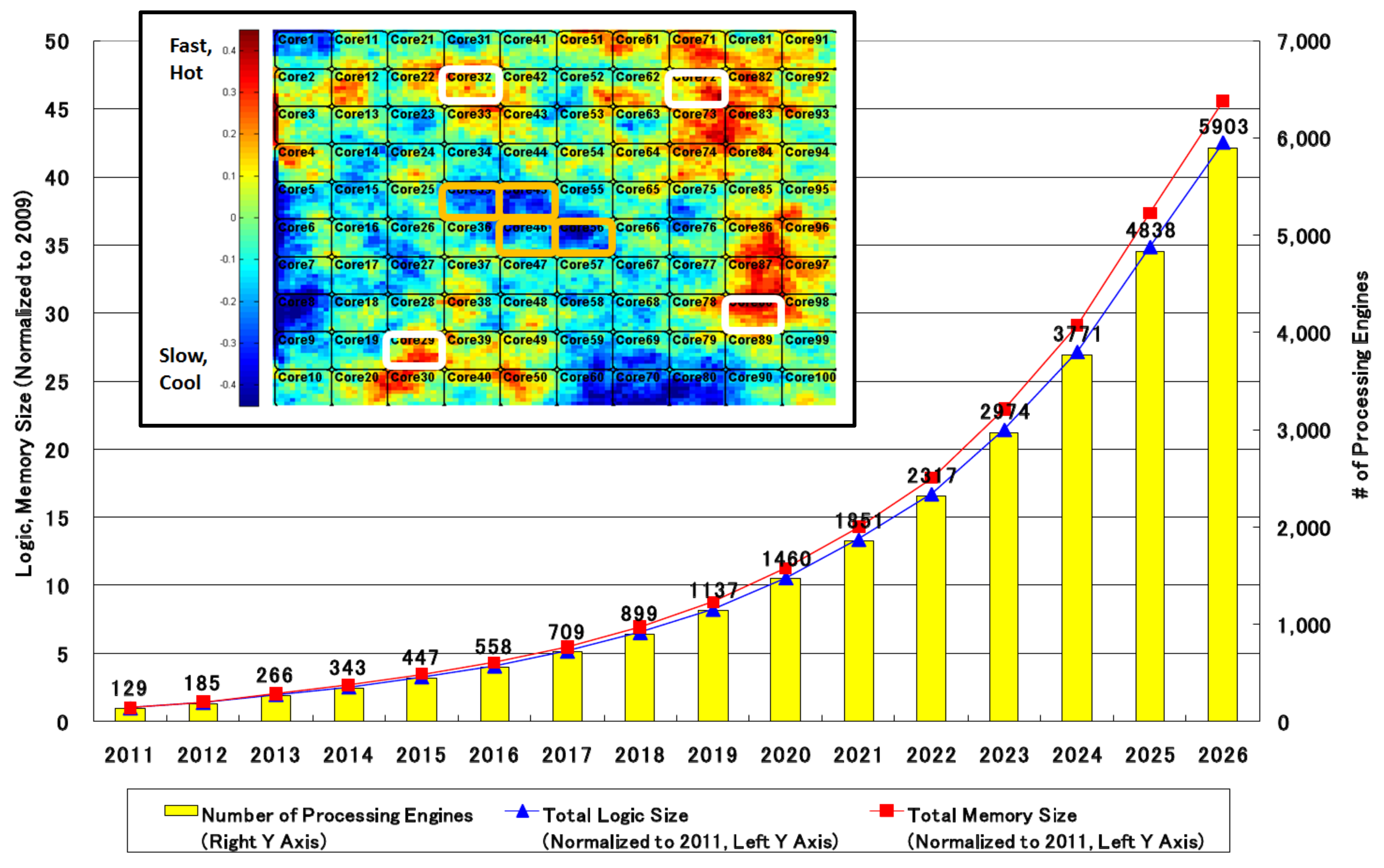
❑ Moore's Law provided a win-win situation:

- ❑ Smaller feature size
- ❑ Higher integration density
- ❑ More functionality
- ❑ Higher speed
- ❑ Lower power consumption
- ❑ Lesser cost per transistor



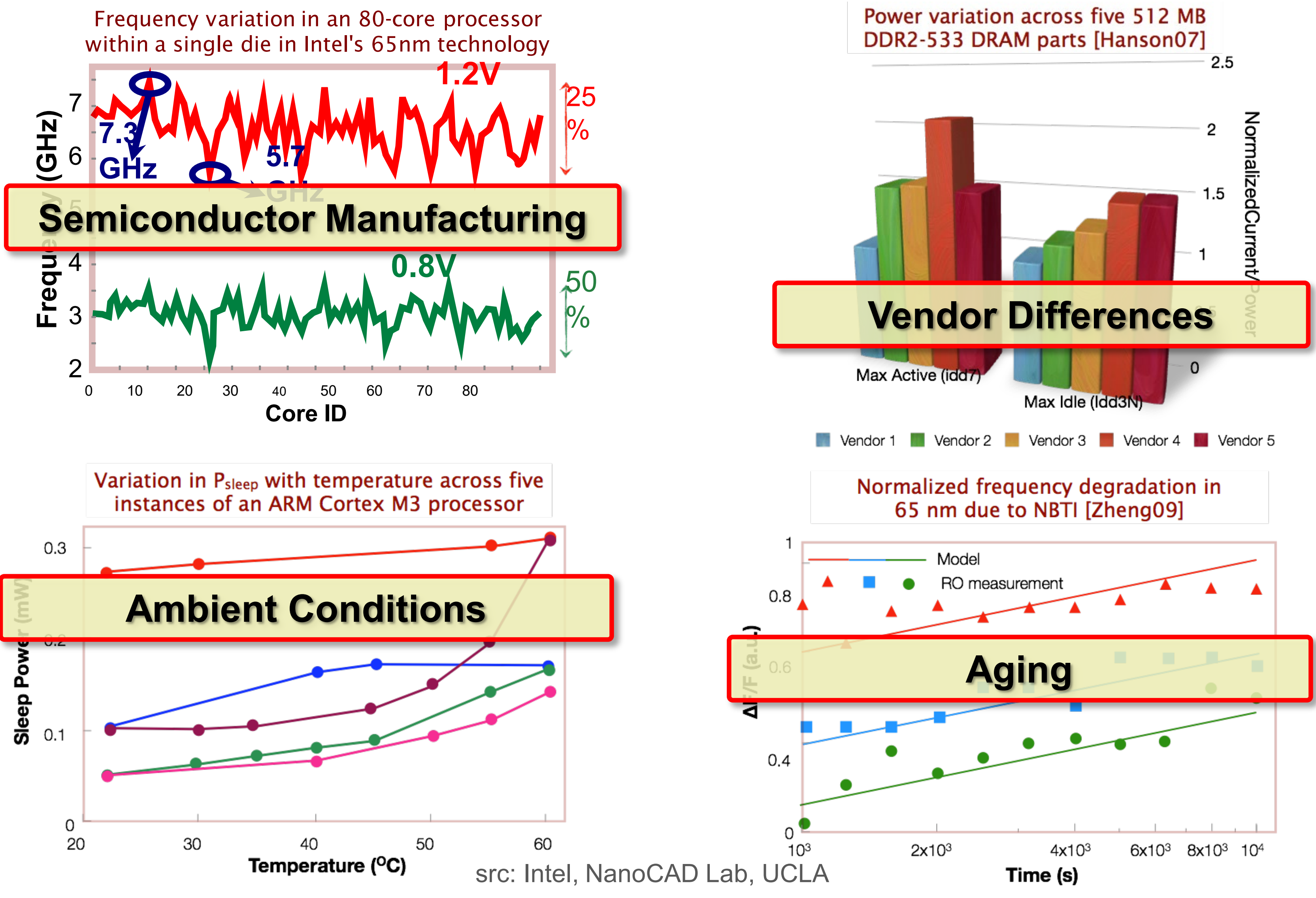
Reliability?

Future Many-Core Systems and Process Variations



Source: ITRS 2011 (ITRS:International Technology Roadmap for Semiconductors until 2017)
Now called IRDS: International Roadmap for Devices and Systems

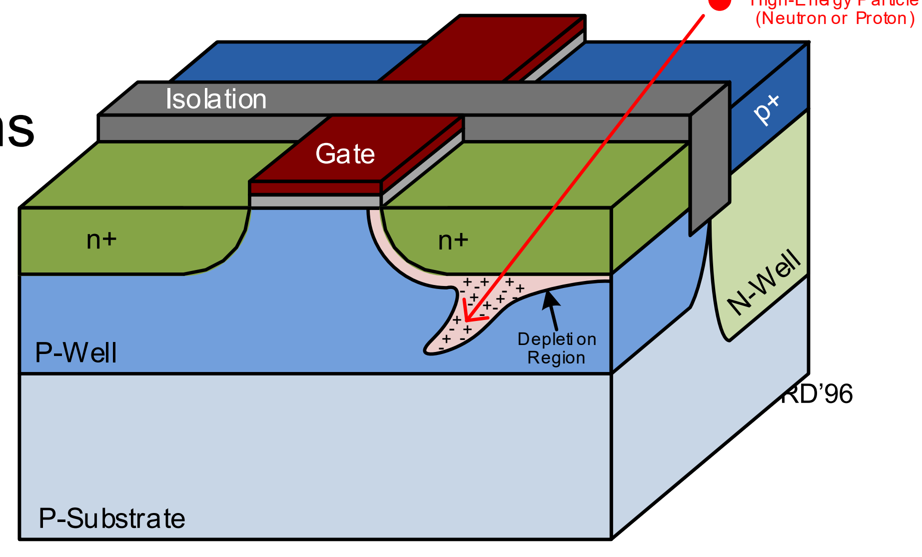
Key Reliability Threats: Variability Sources



src: Intel, NanoCAD Lab, UCLA

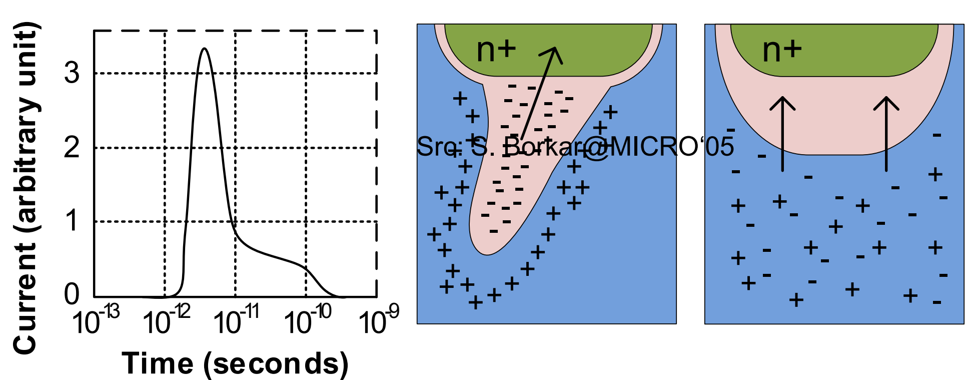
Key Reliability Threats: Soft Errors

- Radiation effects on Hardware → **Bit Flips in Software**
  - Alpha particles
  - Low-energy neutrons
  - High-energy neutrons/protons
- Radiation event
  - Ion track formation
  - Ion drift
  - Ion diffusion



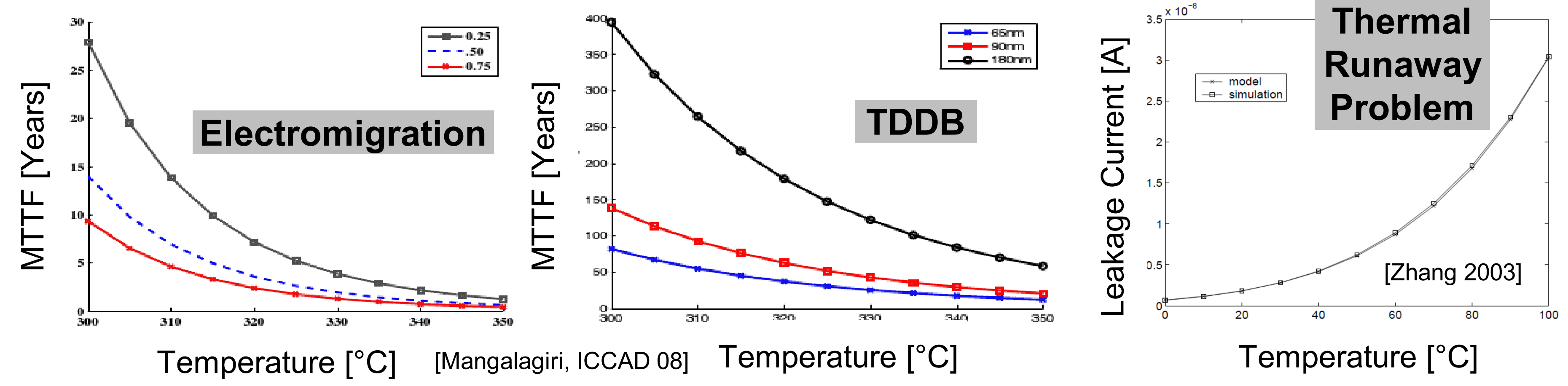
Src: Baumann, TI@Design&Test'05, Ziegler, IBM@IBM JRD'96

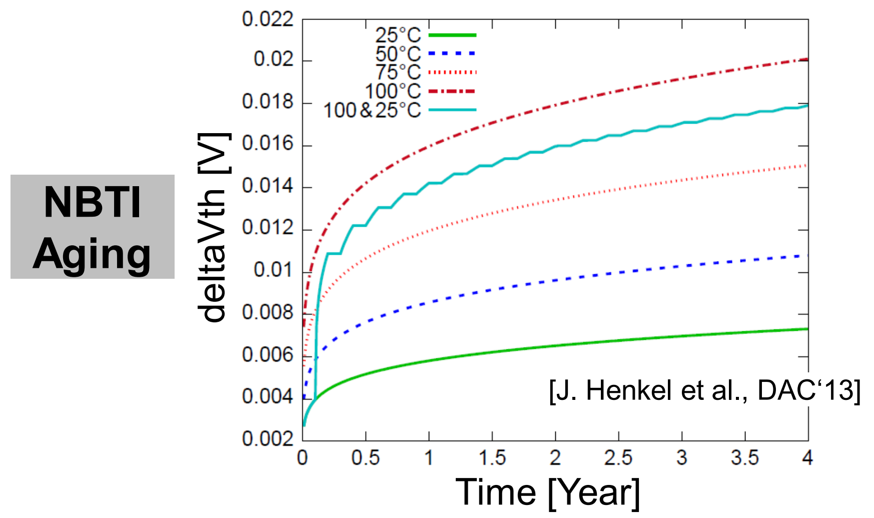
- Accelerated through technology advancements
  - Representation of bits through
  - smaller and smaller charges
  - More transistor per area => System SER ↑

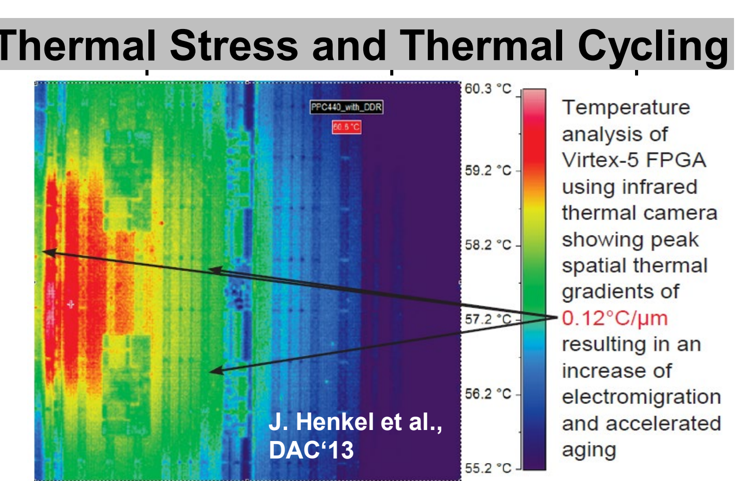


High Temperature => Aggravates Reliability Threats

- **Increased** aging and soft errors => reduced MTTF and lifetime

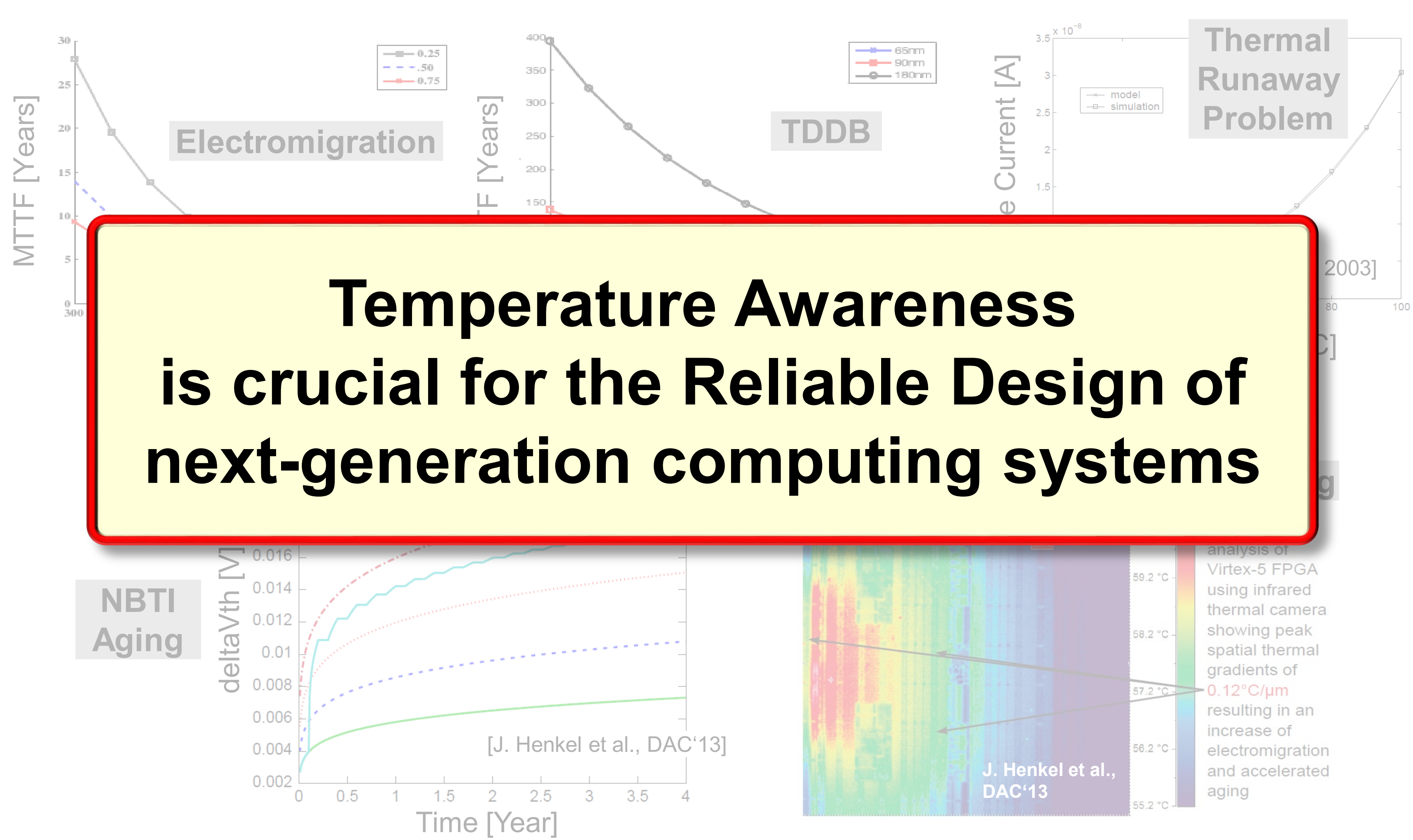






High Temperature => Aggravates Reliability Threats

❑ **Increased** aging and soft errors => reduced MTTF and lifetime



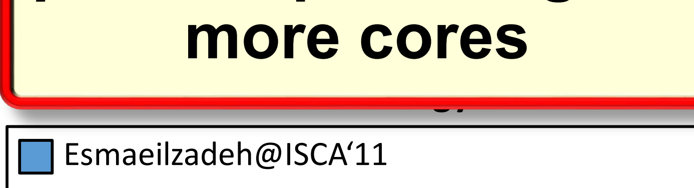
Temperature leads to the Dark Silicon Problem

22nm

11nm and Beyond

344.9
342.8
340.7
338.6
336.5
334.4
332.3
331.0

High power density and temperature *prohibit* powering-on more cores

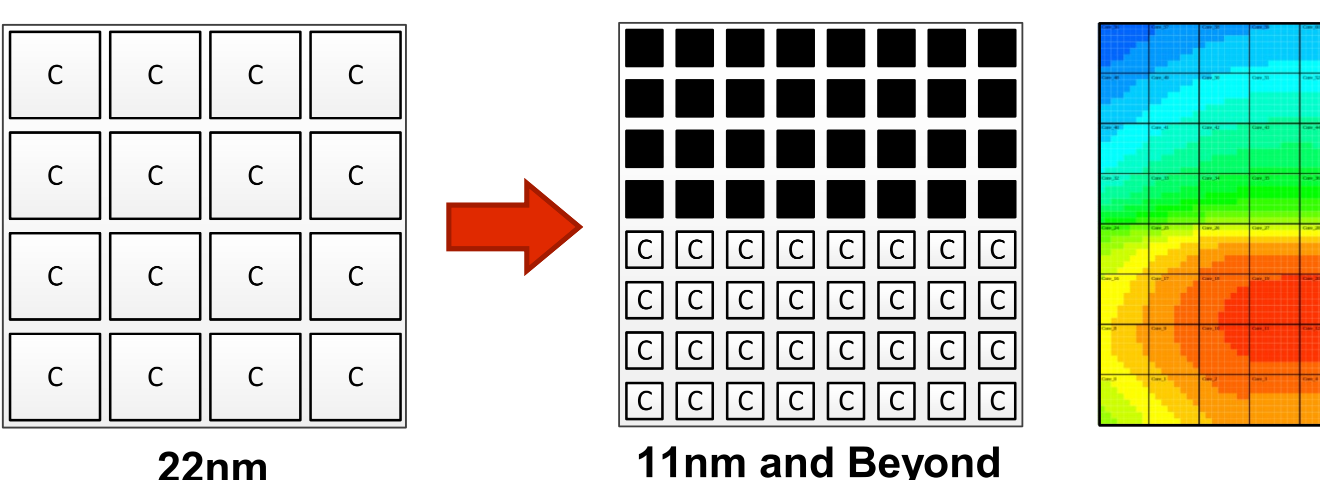


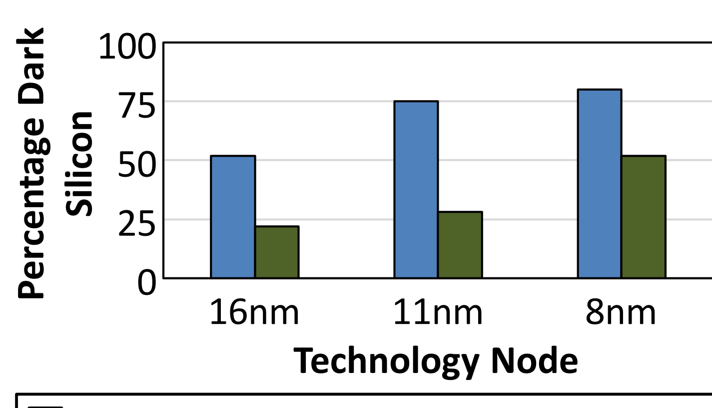
Classical Scaling

Device count	S^2
Device frequency	S
Device power (cap)	1/S
Device power (V_{dd})	$1/S^2$
Power Density	**1**

(Src: "Dennard Scaling")

Temperature leads to the Dark Silicon Problem





Classical Scaling

Device count	S^2
Device frequency	S
Device power (cap)	1/S
Device power (V_{dd})	$1/S^2$
Power Density	**1**

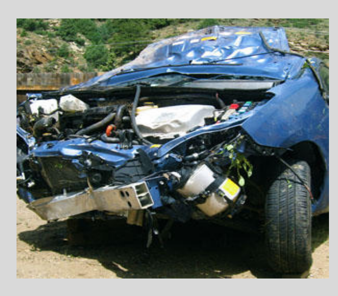
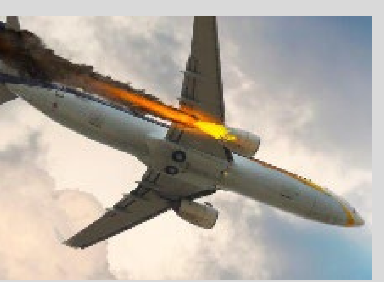
Limited Scaling

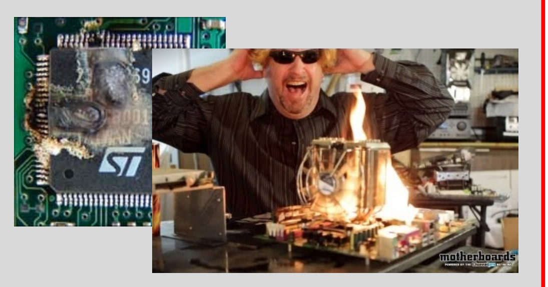
Device count	S^2
Device frequency	S
Device power (cap)	1/S
Device power (V_{dd})	**~1**
Power Density	**S^2**

(Src: "Dennard Scaling")

Summary of Reliability Problems

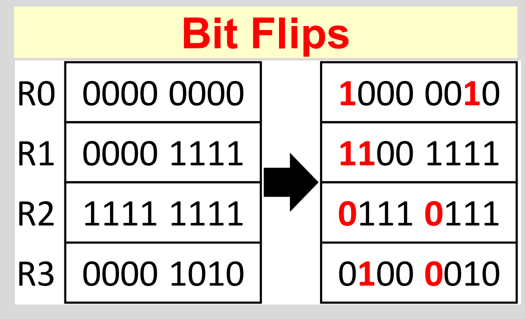
Impact

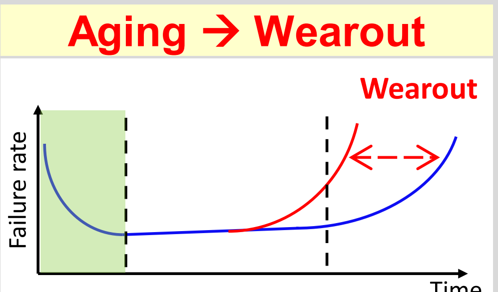


Mani-festation

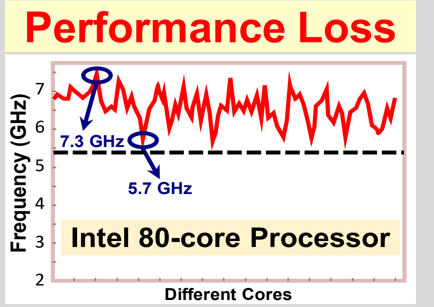
Bit Flips			
R0	0000 0000	→	1000 0010
R1	0000 1111		1100 1111
R2	1111 1111		0111 0111
R3	0000 1010		0100 0010



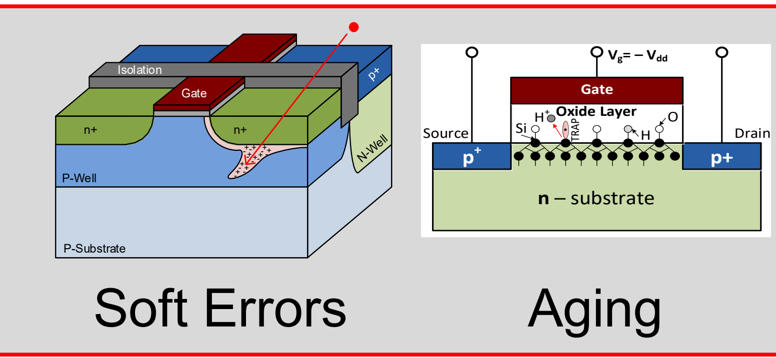
Aging → Wearout

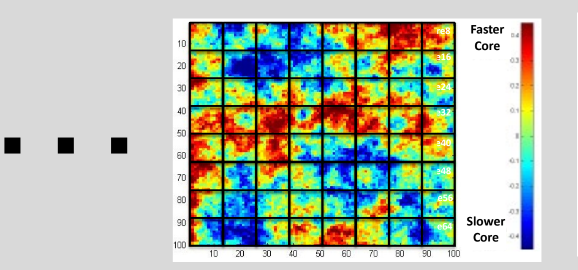


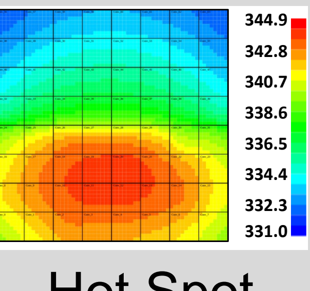
Performance Loss



Physical Level







Soft Errors   Aging   . . .   Process Variation   Hot Spot

Summary of Reliability Problems

Reliability-Related Questions

1. How to manage reliable program execution under process variations, transient and permanent faults?
2. How to develop efficient and accurate cross-layer reliability models?
3. How to handle both functional and timing errors?
4. Do we have to do something special towards Robust Machine Learning and Embedded AI?

Mani-

Physical Le

P-Substrate

n – substrate

Slower Core

334.4
332.3
331.0

Soft Errors Aging Process Variation Hot Spot

State-of-the-Art

- **Mitigating aging and variations**
  - **Timing guardbands:** Considering frequency drop [Agarwal et al.@VTS'07][Kang et al.@ASP-DAC'08][Tiwari et al.@Micro'08]
    - Performance loss
  - **Workload management:** Control usage of cores [Masrur et al.@RTCSA'12][Allred et al.@ICCD'13]
    - Soft-error and variation unaware
- **Mitigating soft errors**
  - **Hardware-level:** spatial/temporal redundancy [Mukherjee et al.@ ISCA'02][Hu et al.@DSN'06]
  - **Software-level:** instruction/data duplication [Oh et al.@TR'02]
    - May aggravate aging

State-of-the-Art

- **Mitigating aging and variations**
  - **Timing guardbands:** Considering frequency drop [Agarwal et al.@VTS'07][Kang et al.@ASP-DAC'08][Tiwari et al.@Micro'08]

Required

A Cross-Layer Reliability Framework

that leverages the knowledge from multiple system layers to address different reliability threats

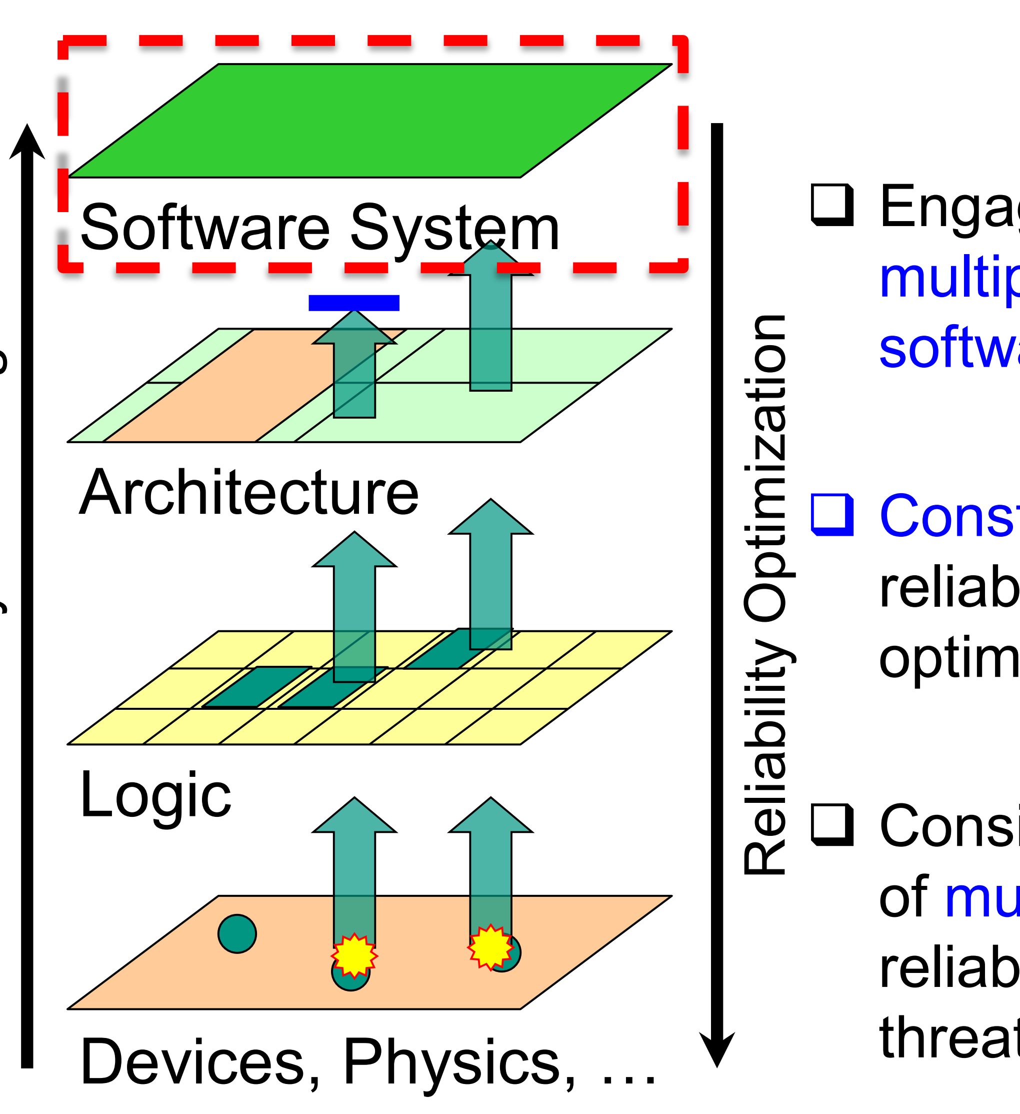
- May aggravate aging

Outline

- Different Types of Reliability Threats
- **Cross-Layer Resilience**
  - Modeling => Bridging the Gap between HW and SW
  - Optimization => Engage Multiple Layers of the System Stack
  - A Self-Healing Framework for Building Resilient CPS
  - Power / Temperature Considerations for Resilience
- **Robust Machine Learning**
- Conclusion

Cross-Layer Approach: Goals Beyond State-of-the-Art

- ❑ Bridging the gap between the hardware and software
- ❑ Estimating software reliability at different levels of granularity



- ❑ Engaging multiple software layers
- ❑ Constrained reliability optimization
- ❑ Consideration of multiple reliability threats

… was also ino the key scope of the DFG's SPP1500 "Dependable Embedded Computing" and NSF's Variability Expedition Inttiatives

GetSURE: A Cross-Layer Reliability Framework with Hardware- and Software-Level Techniques

NBTI-Aging Estimates for Gates and SRAM Cells

Fault Models [Transient, Permanent, #bit-flips, SEUs, SETs]

Reliability and Resilience Modeling

- Program Reliability Estimation
- IVI, FVI Distribution
- Application Resilience Modeling
- TRF Analysis
- Full-Processor Aging Modeling and Estimation

Software Techniques

Reliability-Driven Compilation
1) Resilience-Aware Performance Overhead Budgeting
2) Reliability-Driven Software Transformations

Multiple Compiled Versions with Reliability – Perf. Tradeoff

Reliability-Driven Offline System
1) Reliability-Aware Function Scheduling
2) Reliability-Aware Task Prioritization

Program-level Soft Error Masking

Set of Applications

Hardware Techniques

Heterogeneous Fault-Tolerance Modes
1) Reliable Multi-level Caches
2) Embedded Processors
3) Superscalar Processors
4) Dynamic Voltage & Freq. Scaling

Process Variation (PV)

Faster Core

Slower Core

C_4 C_2 C_3 C_1

µ-Proc. Config.

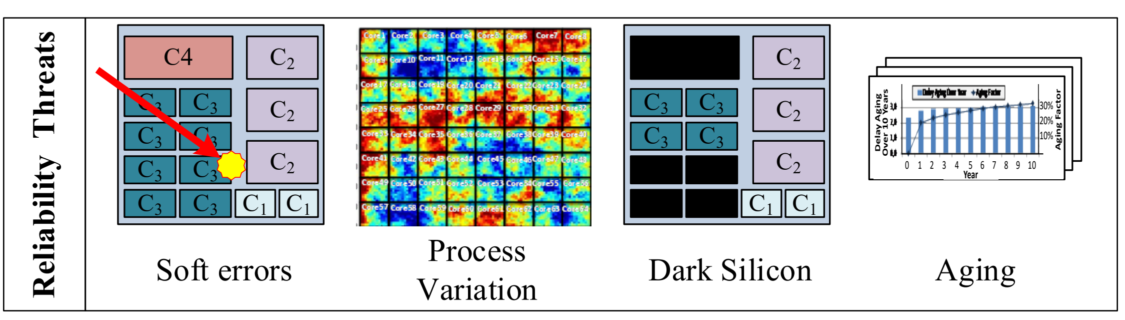
Run-time System

Dynamic Application Version Selection
1) Aging-aware
2) DVFS

Scenario - I

Scenario - II

Core and Application Version Selection
1) Variation-aware
2) Aging-aware
3) Perf.-aware
4) Dark-silicon



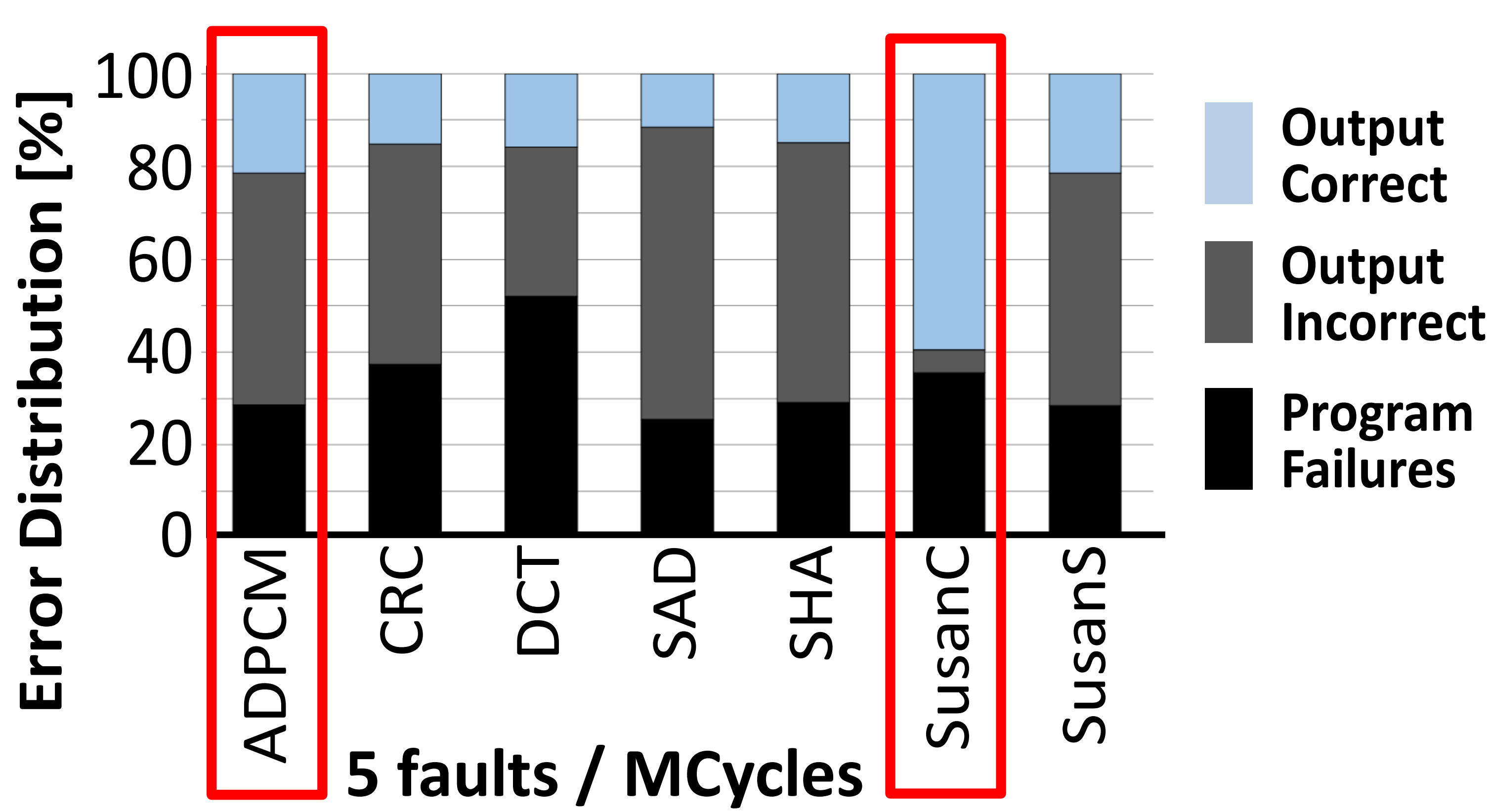
DAC (2013, 14, 15, 16, 17), **DATE** (2013, 14, 15, 16, 17), **Codes+ISSS** (2011, 14, 15), **TC** (2016, 17), **TVLSI** 2016, **TCAD** 2014; see more in Refs.

Fault Injection Campaigns: Different Parameters

Parameter	Description	Properties/Values
Distribution	Distribution models for fault generation	random
Bit Flips	Min/Max number of bits flipped	1/1, 1/2, …
Fault Probability	Probability that strike becomes a fault	10%-100%
Fault Location	List of target processor components	Register file, PC, IW, IEU, Cache contl., etc.
Processor Layout/Area	Size of the complete target device	in gate equivalents or mm^2
Component Area	Area of different processor components given as percentage of processor area	0%-100%
Place and latitude	City and altitude at which the device is used to determine the flux rate (N_{Flux})	Karlsruhe, Germany, 1- 20km
Frequency	Operating frequency of the processor	50, 100 MHz

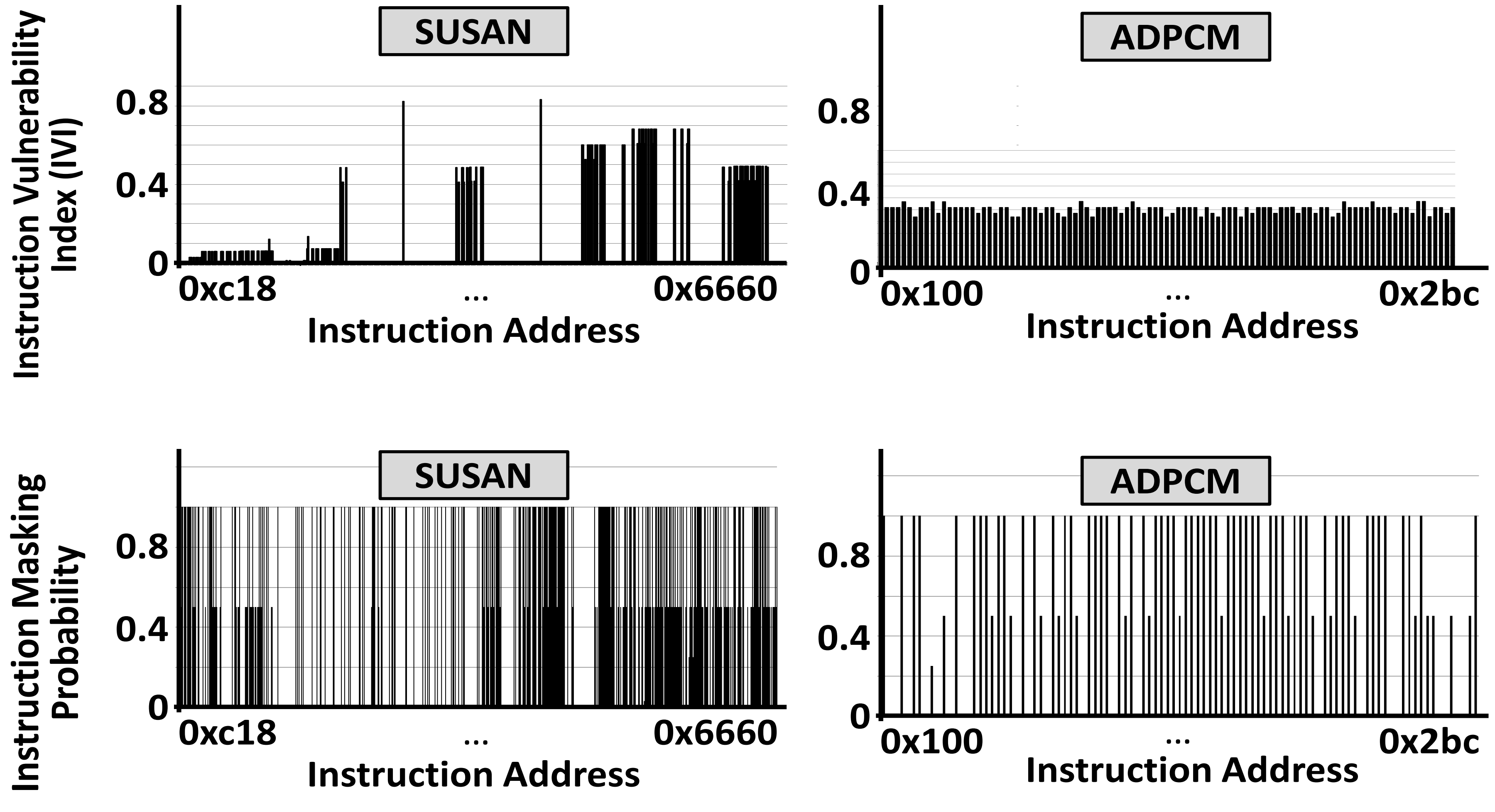
Application Resilience vs. Error Distribution

- ❑ **Different applications have distinct resilience properties**
  - ❑ Error distribution
  - ❑ Soft error masking potential



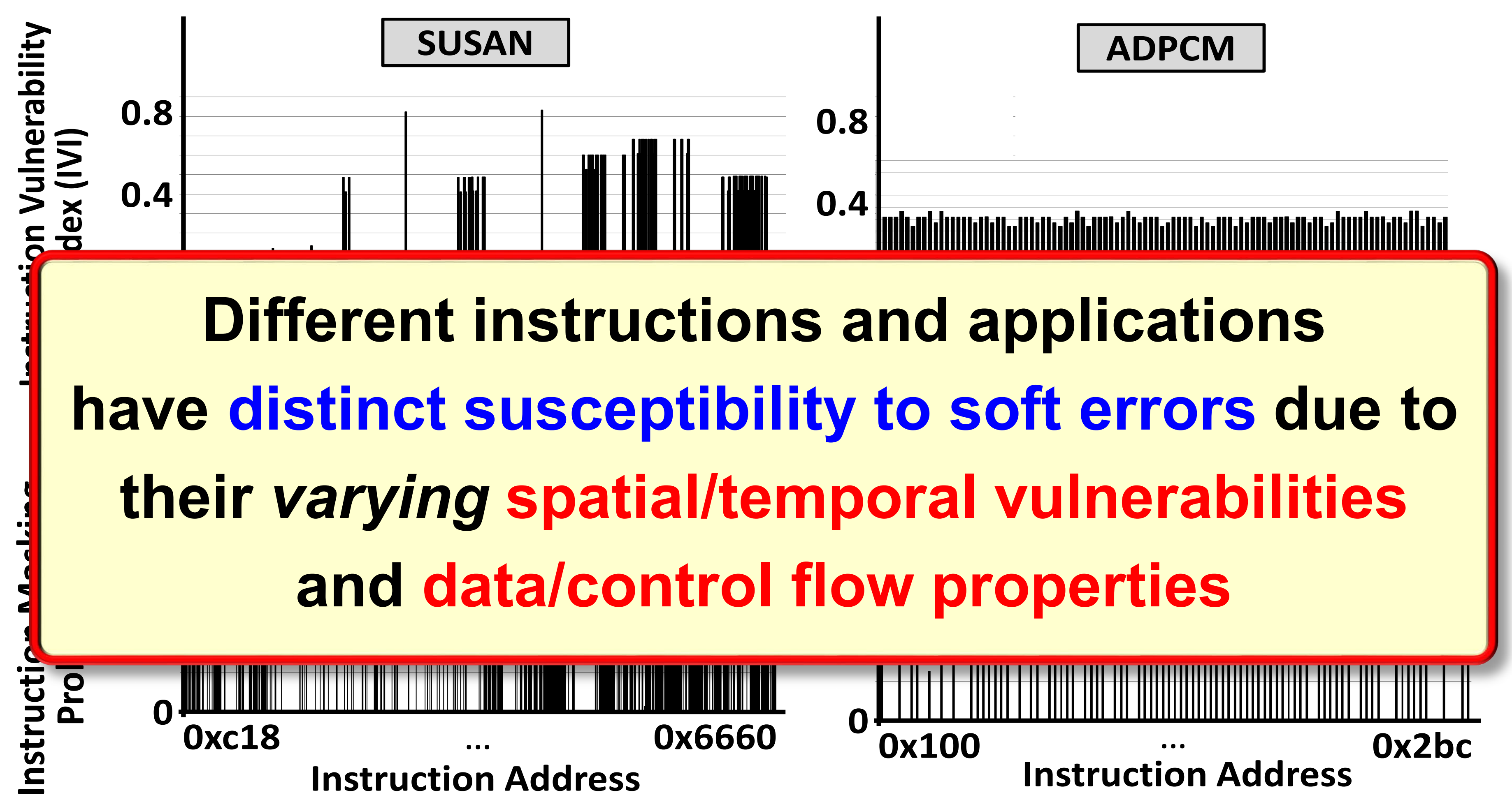
Src: Rehman et al.@CODES+ISSS'11

Instruction-Level Analysis: Vulnerabilities and Masking Probabilities



Src: Shafique et al. DAC 2013; Rehman et al.@CODES+ISSS'11

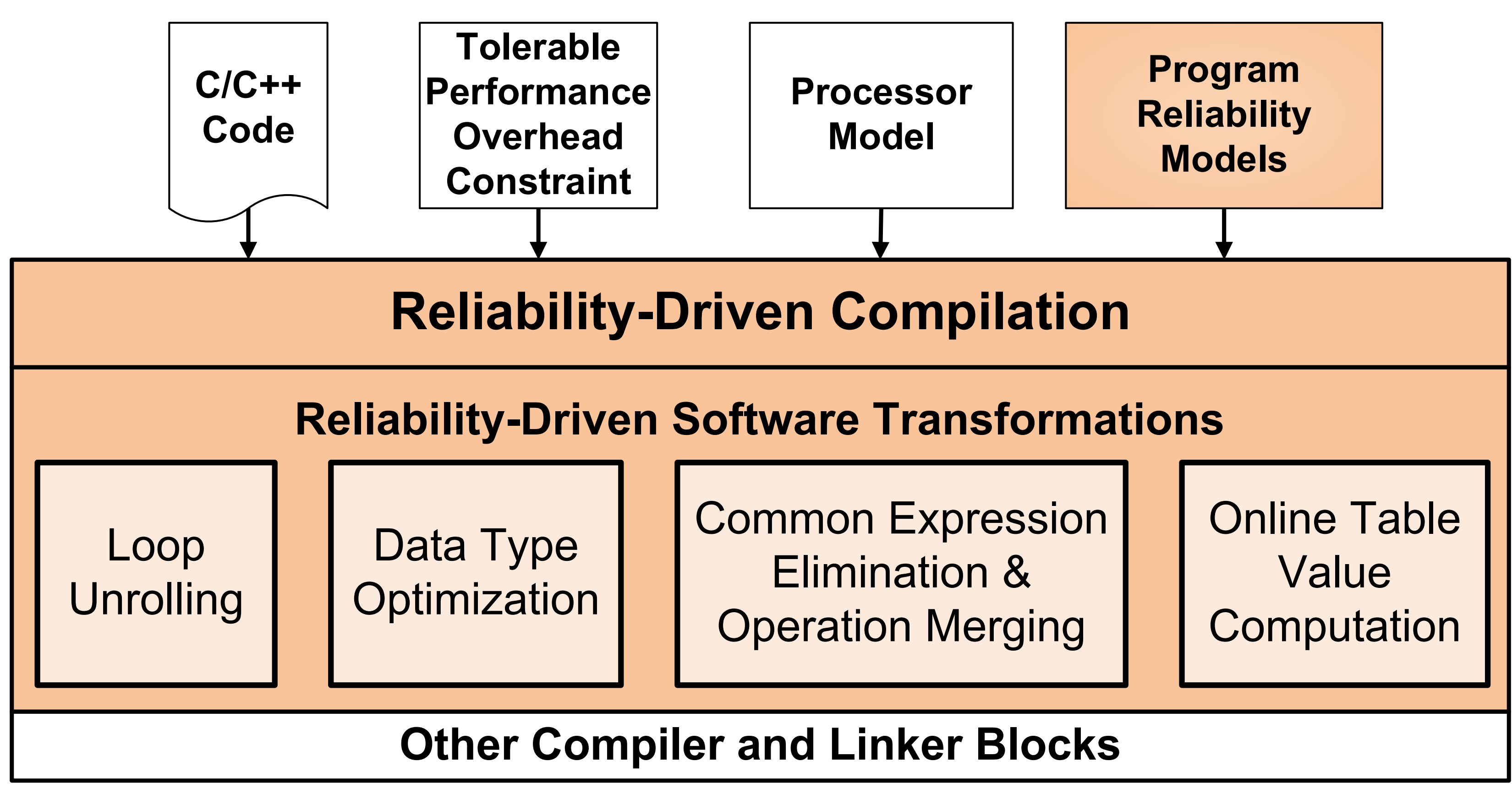
Instruction-Level Analysis: Vulnerabilities and Masking Probabilities



Different instructions and applications have distinct susceptibility to soft errors due to their *varying* spatial/temporal vulnerabilities and data/control flow properties

Src: Shafique et al. DAC 2013; Rehman et al.@CODES+ISSS'11

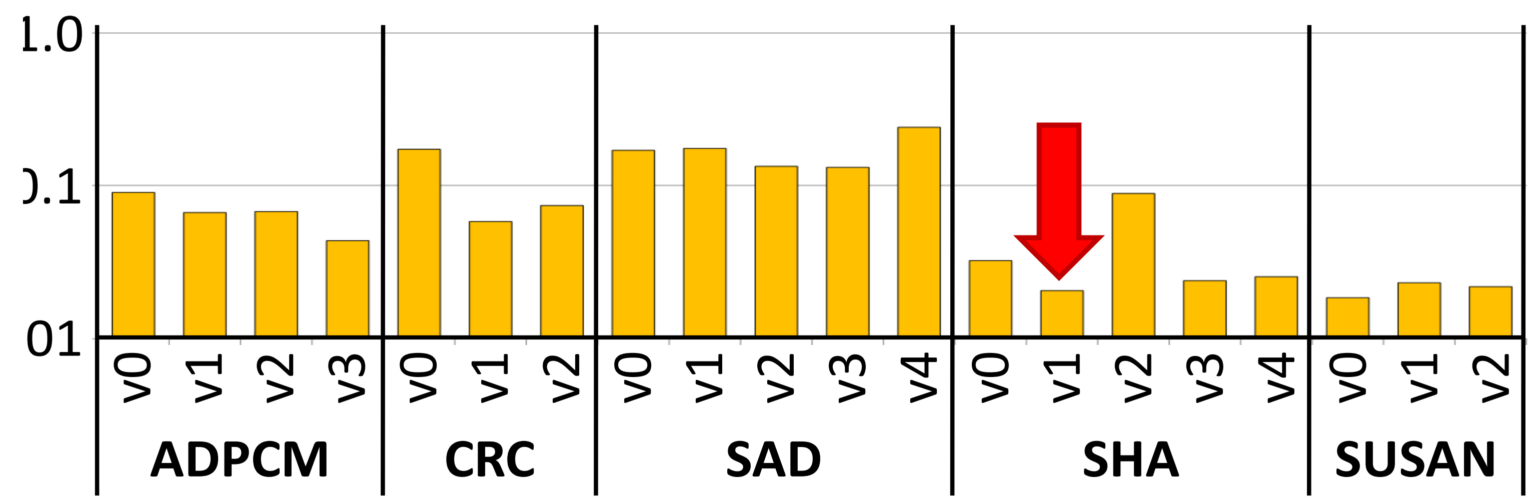
Reliability-Driven Code Generation

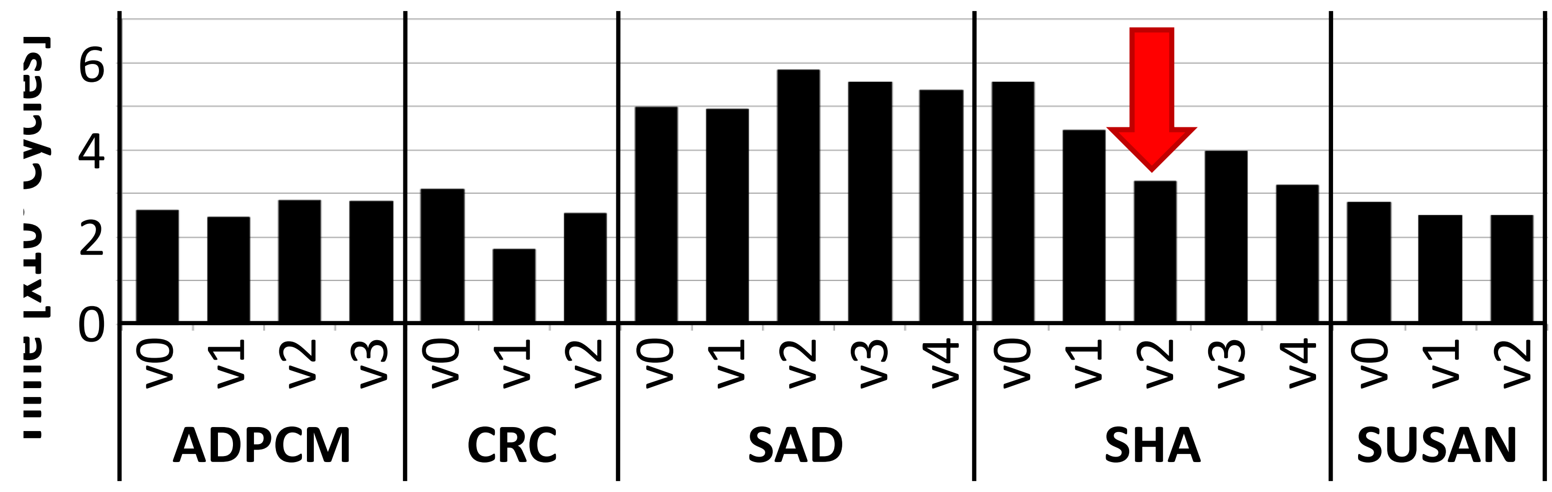


- Reducing error probability by minimizing the spatial and temporal vulnerabilities and critical instructions executions

Src: Rehman et al.@CODES+ISSS‘11 , TCAD’14, RTAS’13, TC‘16

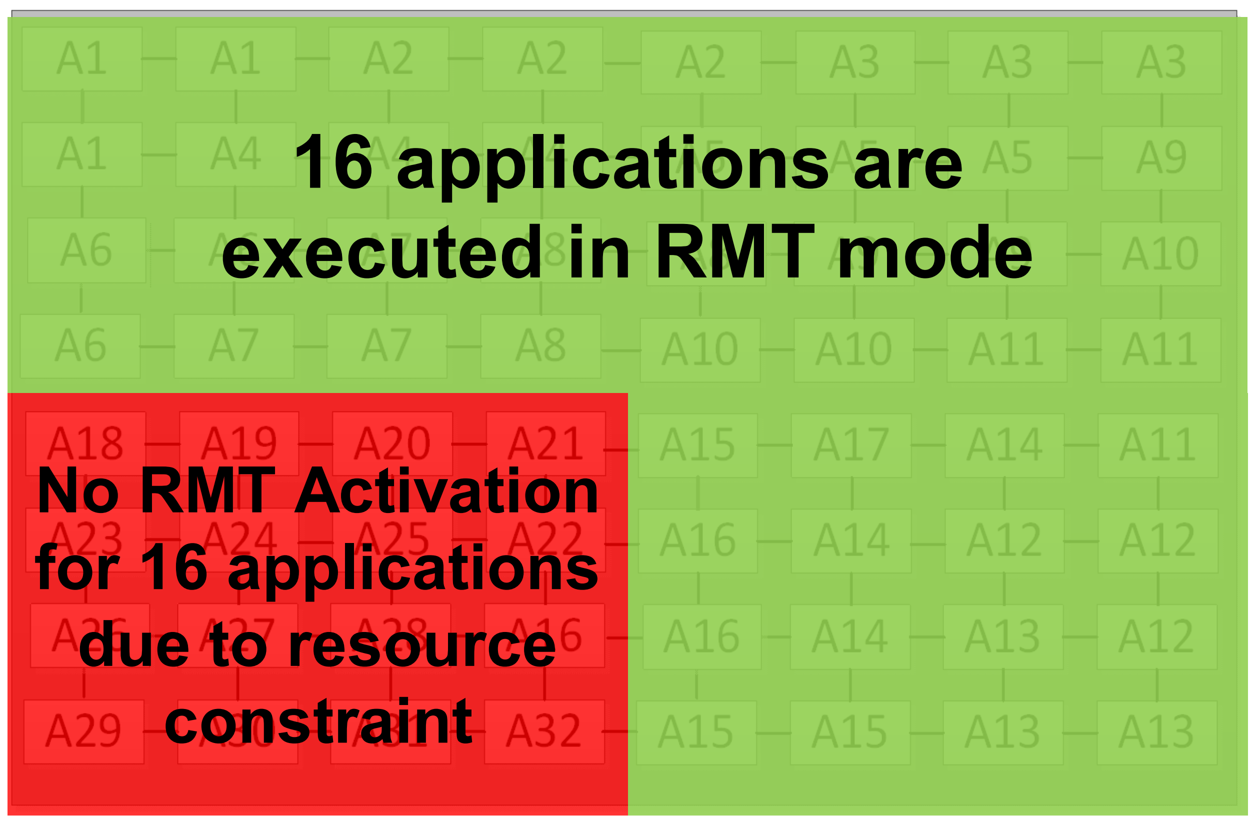
Reliability-Driven Code Generation: Results for Multiple Reliable Code Versions





Src: Rehman et al.@CODES+ISSS'11 , TCAD'14, RTAS'13, TC'16

Redudant Multi-Threading:
Core Allocation and RMT (De)activation Decision



- 64 cores
- 32 applications

Question
Which application should run in RMT mode?

Not all applications can be served with RMT due to resource constraint

Src: Rehman et al.@DAC'14

Aging Impact of RMT

- Multi-core subjected to process variation for different years

Variation Map

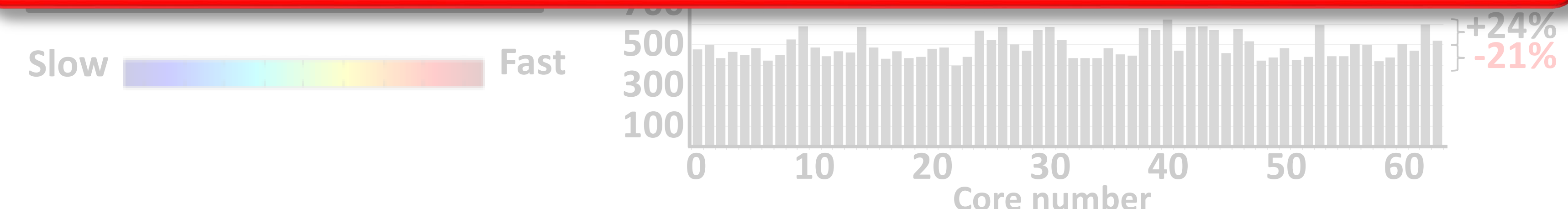
Frequency Variations *before* Aging @Year 0

What is Required?

Aging-minimizing RMT technique, that

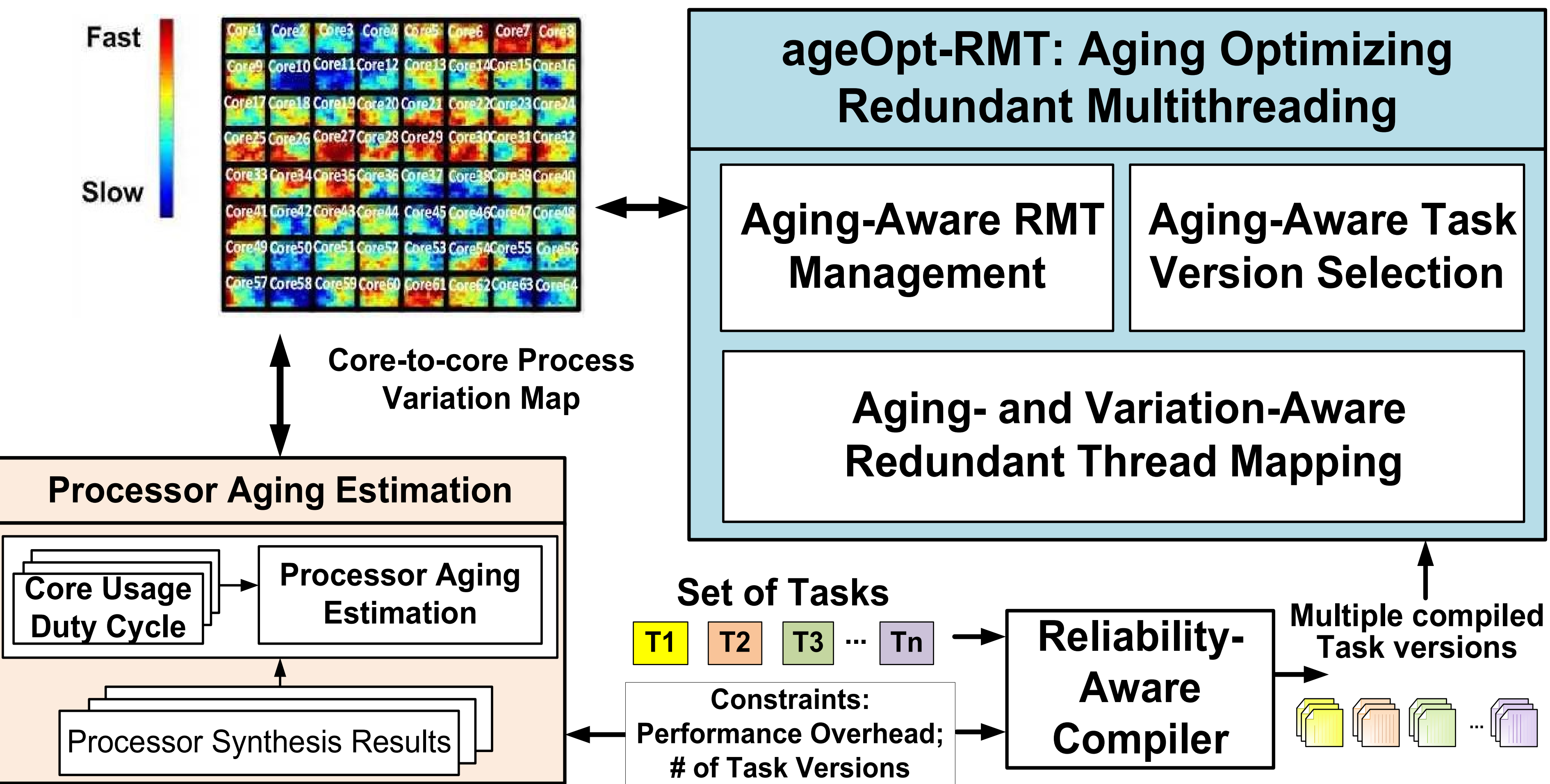
(1) Reduces the aging of slower cores

(2) Achieves a balanced aging profile



Src: Kriebel et al.@DAC‘16

ageOpt-RMT: Aging & Variation-Aware RMT



F. Kriebel, S. Rehman, M. Shafique, J. Henkel, "ageOpt-RMT: Compiler-Driven Variation-Aware Aging Optimization for Redundant Multithreading", **DAC**, 2016.

Reliability Improvement Results Under Process Variation and Aging

- Considering soft errors, aging and process variations
- Reliability savings compared to different state-of-the-art solutions

6x6 Cores

With Aging @ Year=0

With Aging @ Year=10

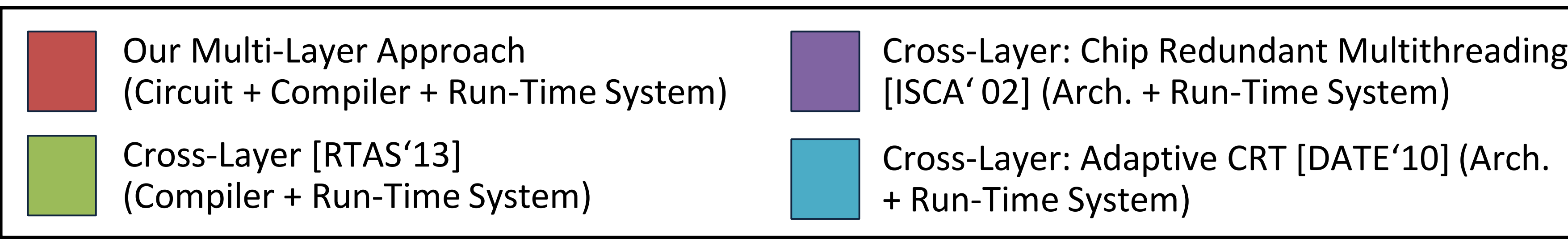
Percentable Reliability Improvement

100
80
60
40
20
0
-20

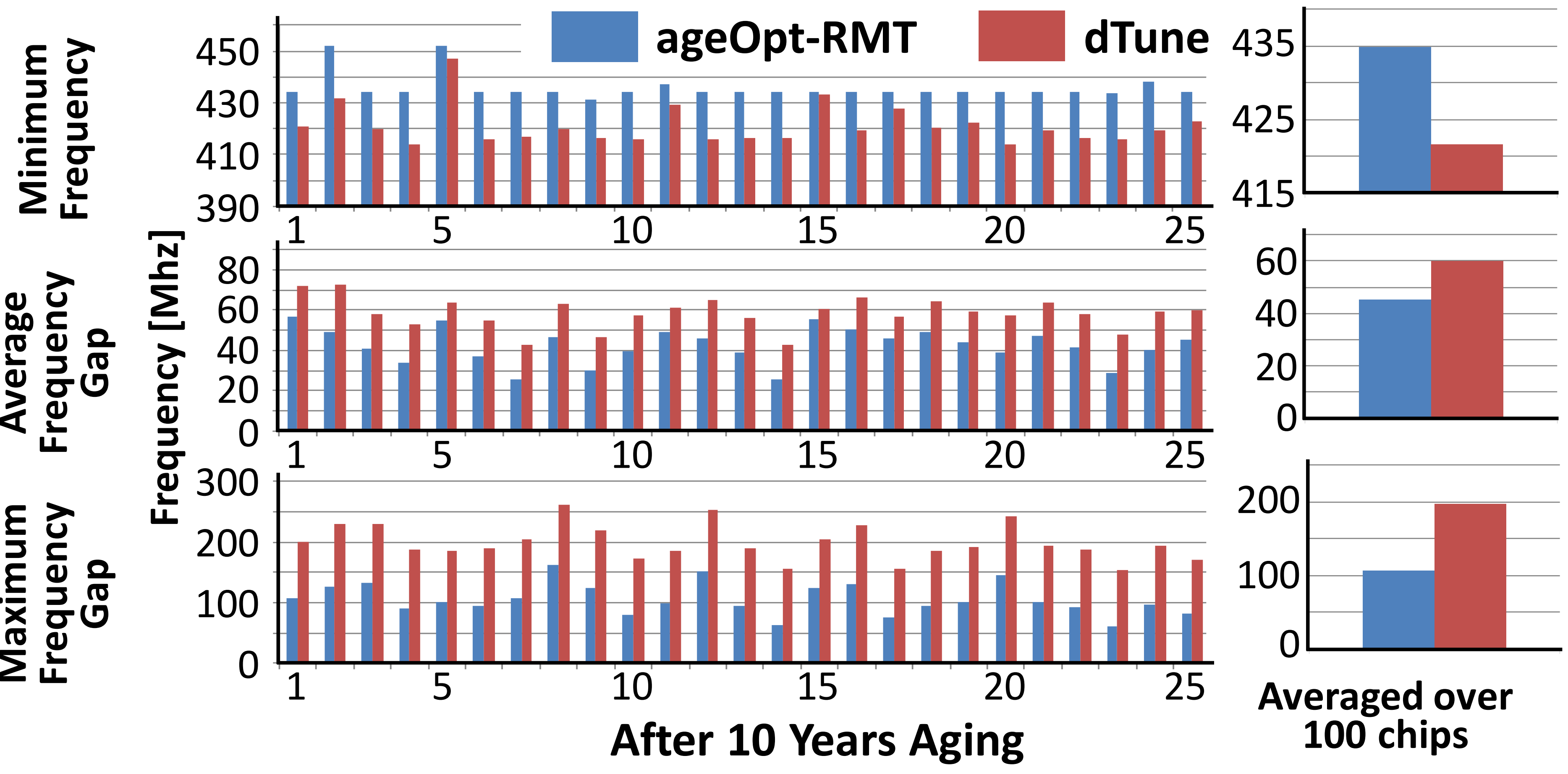
5 10 15 20 25 30

5 10 15 20 25 30

Number of Concurrently Executing Tasks



Aging Imapact of RMT: Avg. over Different Chips



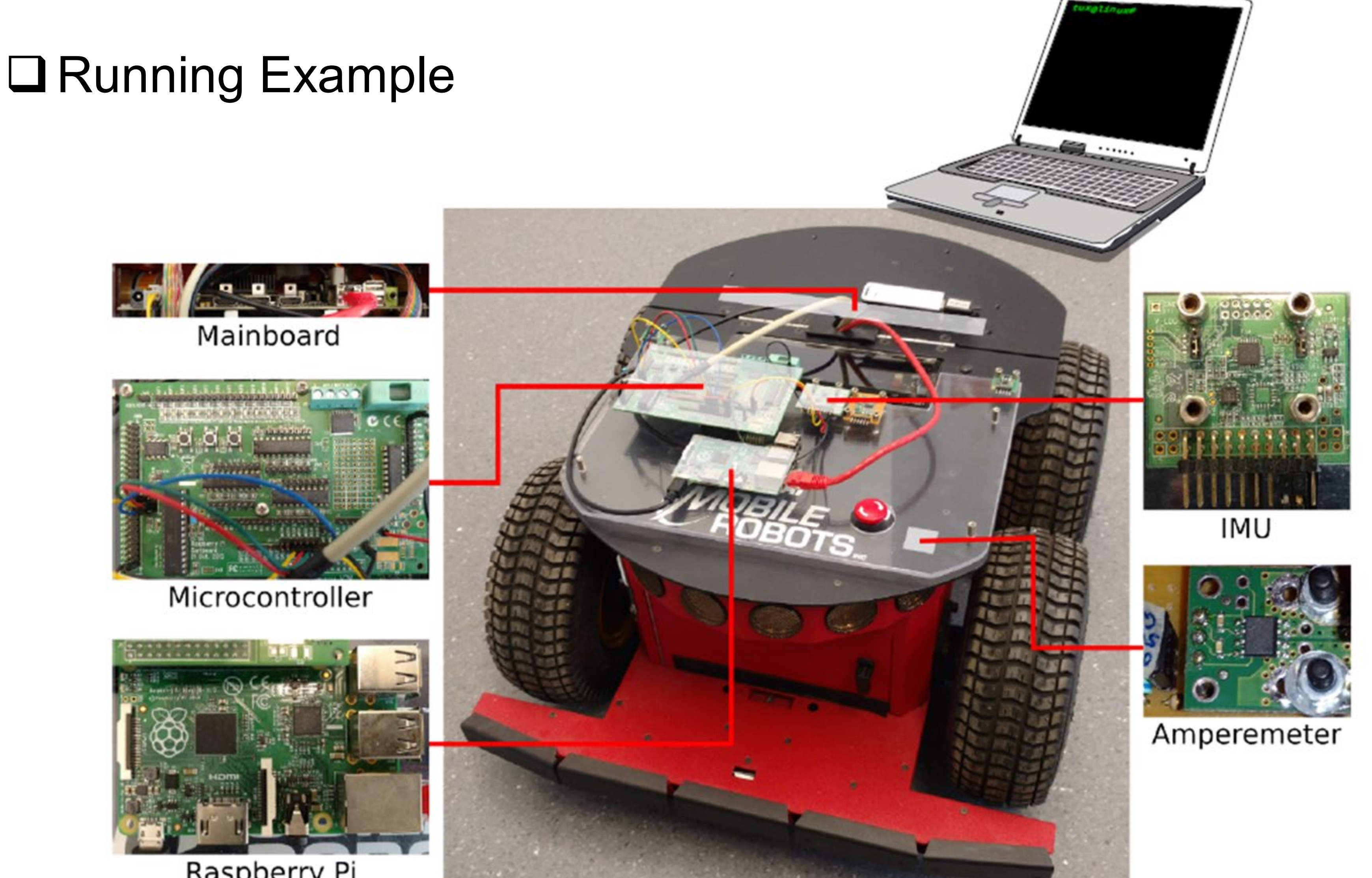
Src: Kriebel et al.@DAC'16

Outline

- Different Types of Reliability Threats
- **Cross-Layer Resilience**
  - Modeling => Bridging the Gap between HW and SW
  - Optimization => Engage Multiple Layers of the System Stack
  - A Self-Healing Framework for Building Resilient CPS
  - Power / Temperature Considerations for Resilience
- **Robust Machine Learning**
- Conclusion

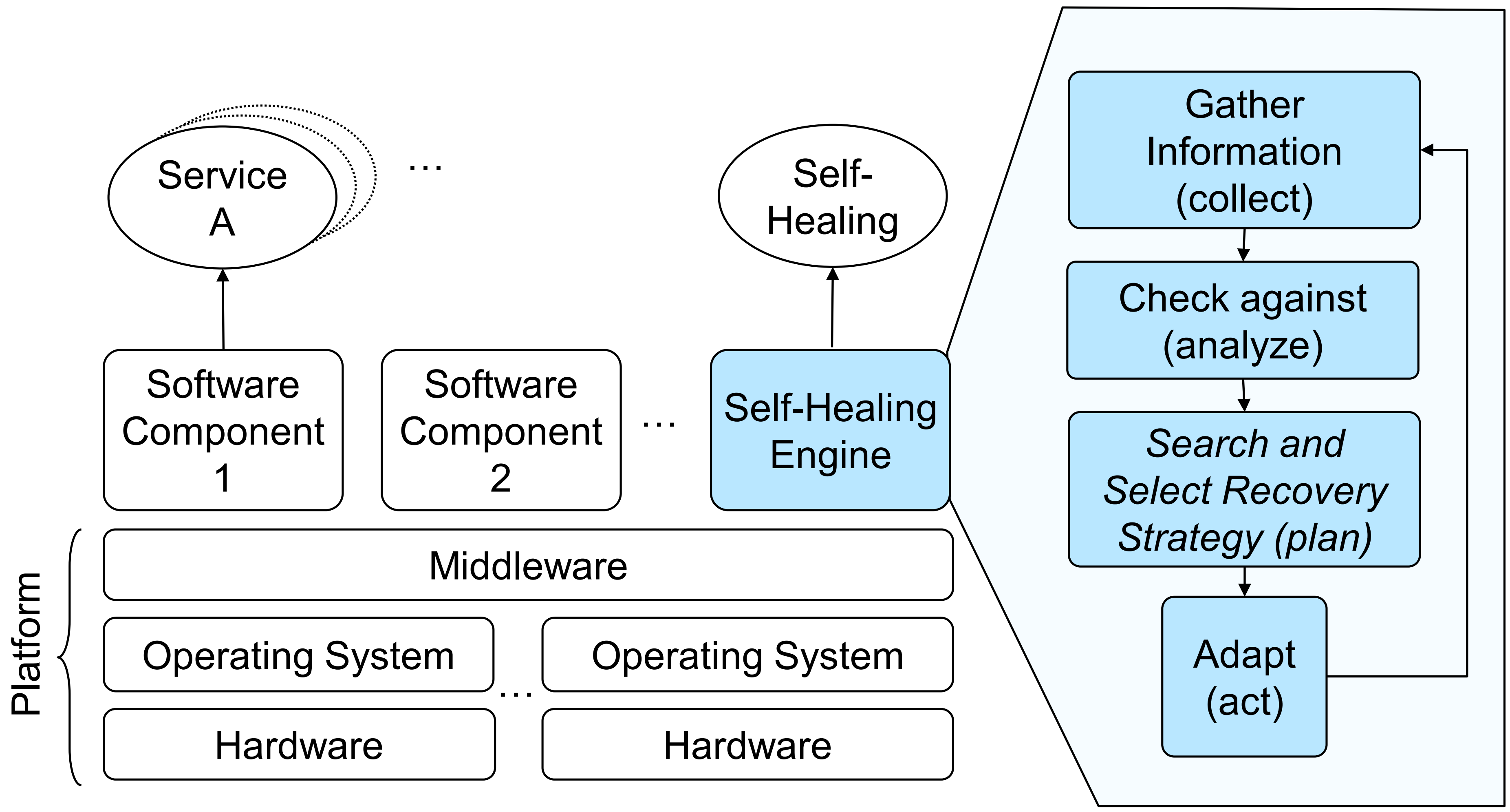
Demonstrating Self-Healing in Automotive CPS

❑ Running Example



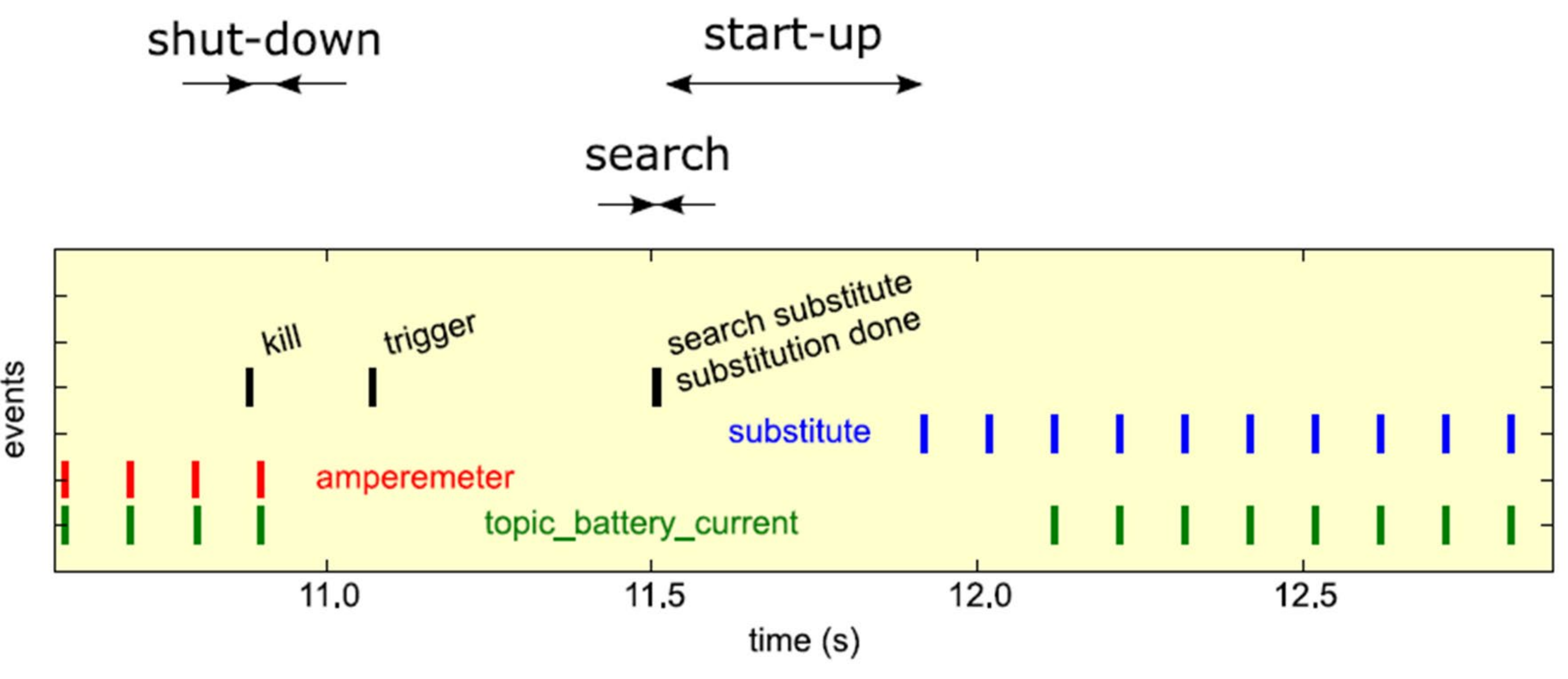
D. Ratasich, O. Höftberger, H. Isakovic, M. Shafique, R. Grosu, "A Self-Healing Framework for Building Resilient Cyber-Physical Systems", 19th IEEE International Symposium on Real-Time Computing (ISORC), May 2017

Self-Healing: System Model and Overview



D. Ratasich, O. Höftberger, H. Isakovic, M. Shafique, R. Grosu, "A Self-Healing Framework for Building Resilient Cyber-Physical Systems", 19th IEEE International Symposium on Real-Time Computing (ISORC), May 2017

Experiment on a Real-Wolrd CPS Prototype: A Run-Time Scenario



D. Ratasich, O. Höftberger, H. Isakovic, M. Shafique, R. Grosu, "A Self-Healing Framework for Building Resilient Cyber-Physical Systems", 19th IEEE International Symposium on Real-Time Computing (ISORC), May 2017

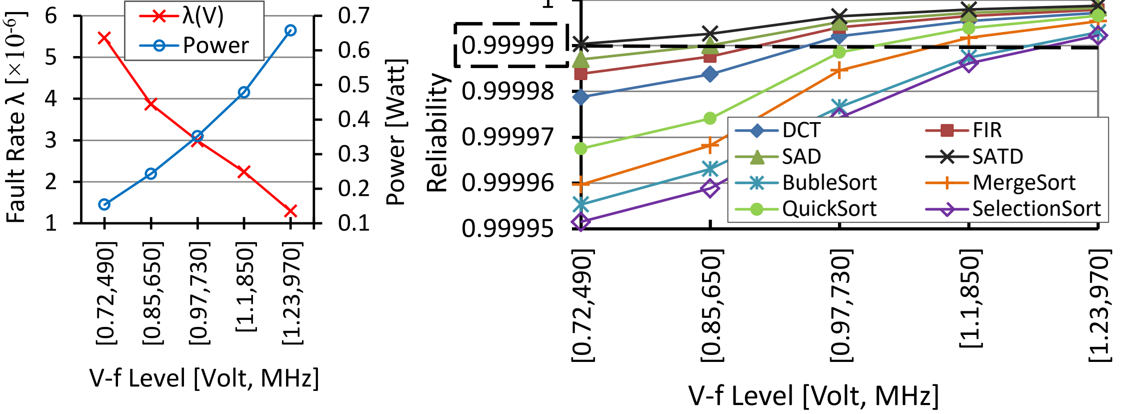
Outline

- Different Types of Reliability Threats
- **Cross-Layer Resilience**
  - Modeling => Bridging the Gap between HW and SW
  - Optimization => Engage Multiple Layers of the System Stack
  - A Self-Healing Framework for Building Resilient CPS
  - Power / Temperature Considerations for Resilience
- **Robust Machine Learning**
- Conclusion

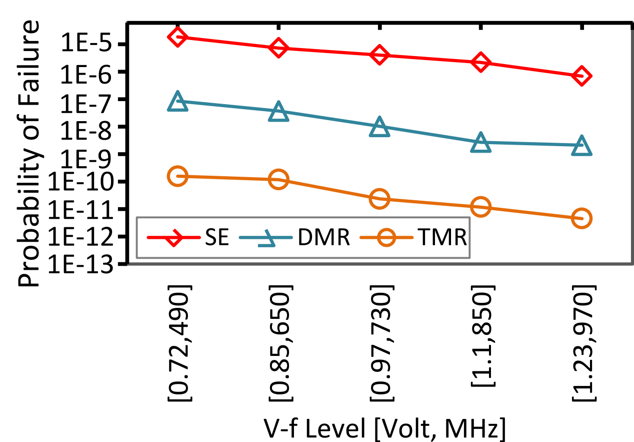
Power-Aware Software Reliability

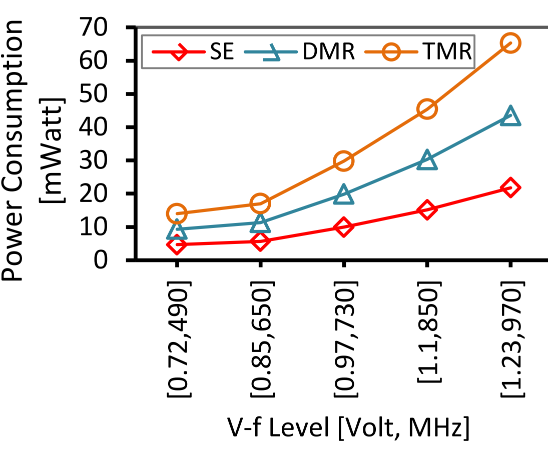
$$system\text{ - }wide\ reliabiliy \propto e^{-(fault_rate \times software_vulnerability \times execution_time)}$$

- **HW-Level Tradeoffs**
  - Fault rate $\lambda(V) \propto 10^{(V_{\max} - V)}$
  - Power $P(V, f) \propto V^2 f$
- **SW-Level Tradeoffs**
  - Reliability variations at different V-f levels



- ***Tradeoffs for Different Redundancy Modes***

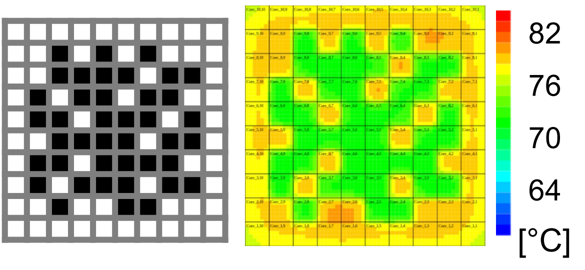




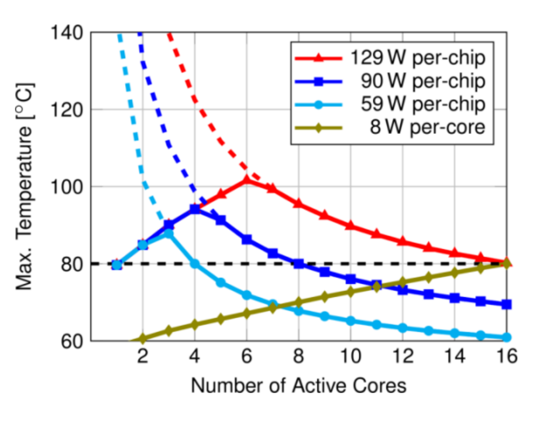
See Salehi et al. **ISLPED'15, Codes+ISSS'15**

Mitigating the Power Density and Dark Silicon:
Techniques for *High Performance* and *Reliability*

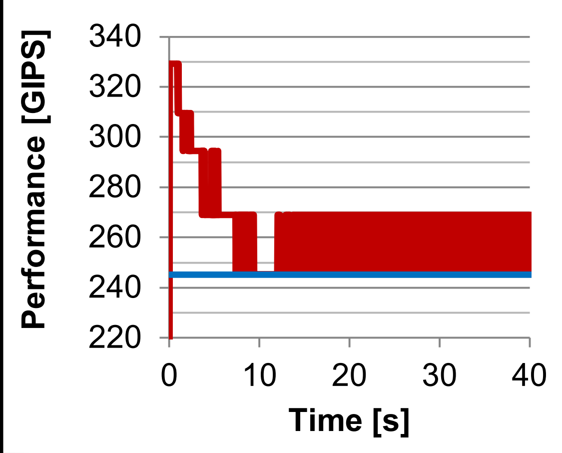
Dark Silicon Management
(Patterning and Resource Management)



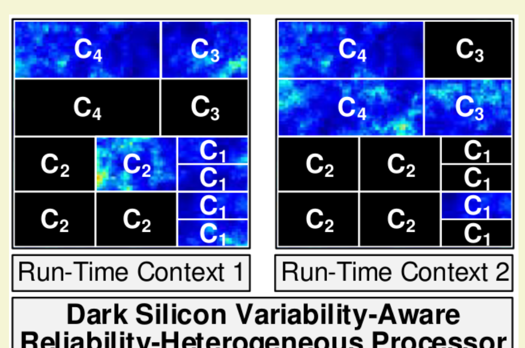
Thermal Safe Power (TSP)
(Abstract from temperature using efficient power budgets)



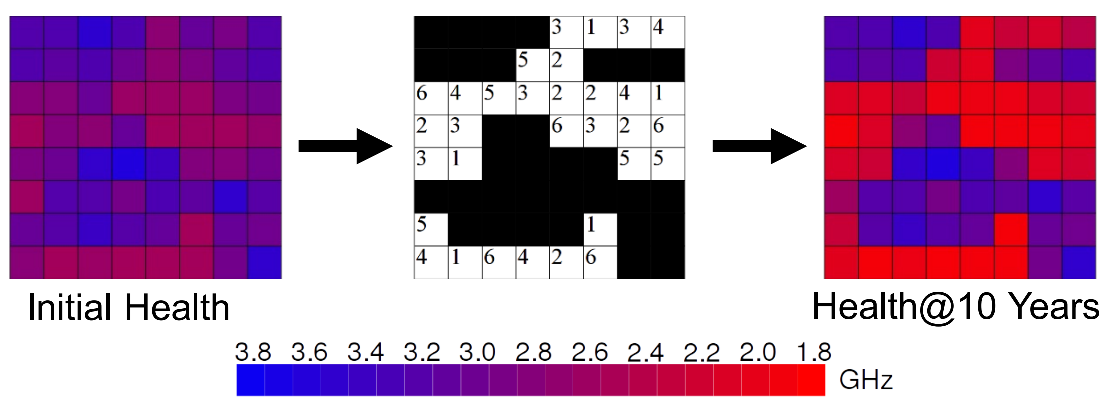
STC / NTC vs. Boosting
(Constant frequency vs. control-loop based boosting)



Dark Silicon-Aware Soft Error Tolerance



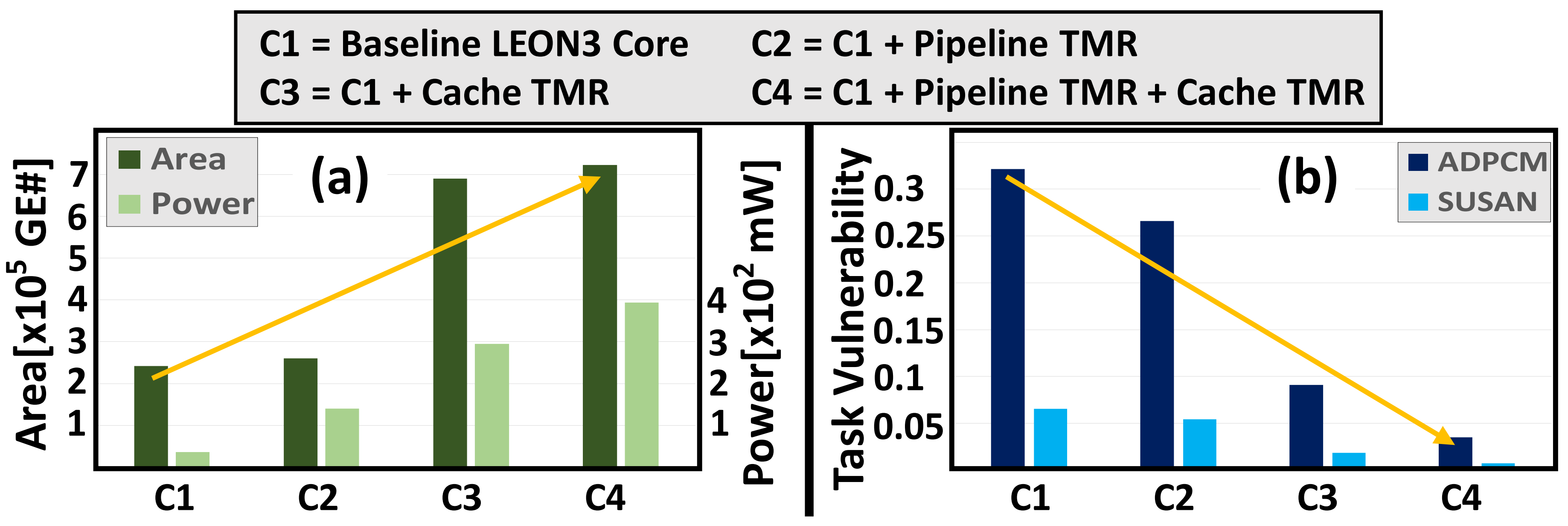
Dark Silicon-Aware Aging Optimization



Shafique et al., @ DAC (2014, 15), DATE (2015, 16), CODES+ISSS (2014, 15), TC (2017, 18), D&T (2016)

Adaptive Soft Error Resilience: Turning Dark Silicon Problem into a Solution

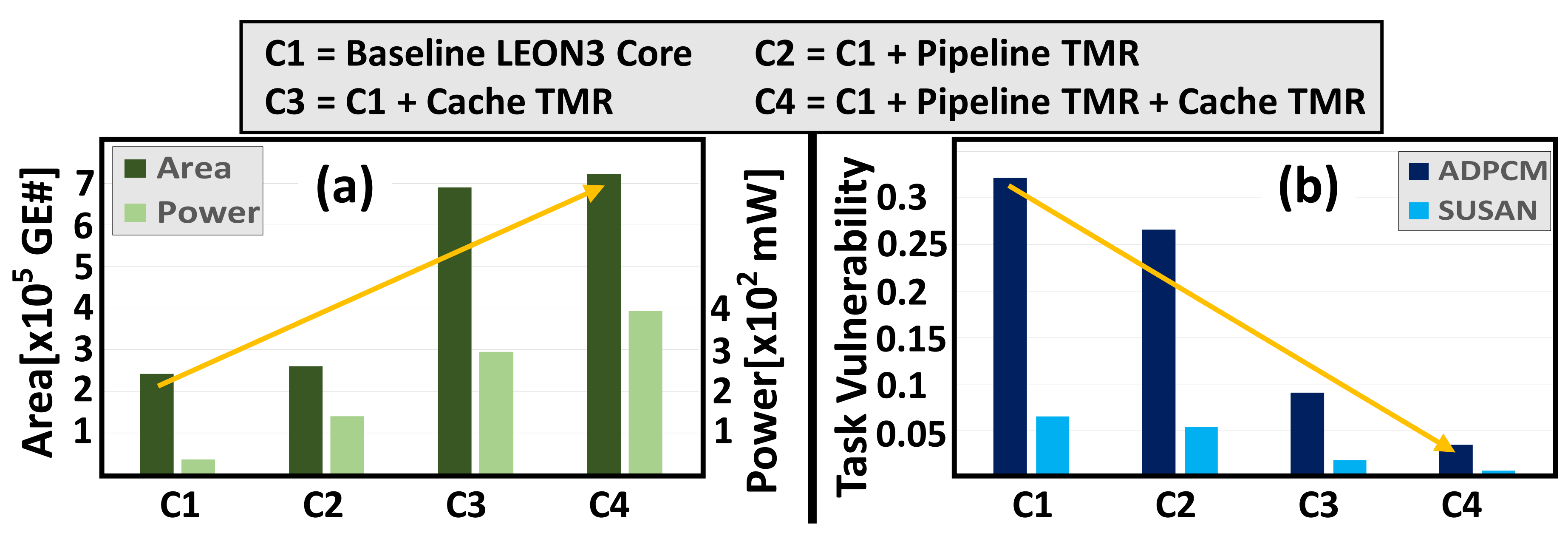
- Leverage available dark silicon chip with reliability-wise specialized cores offering a distinct degree of reliability, i.e., protection against soft errors
  - Multiple “iso-ISA reliability-heterogeneous cores”
  - Higher protection against soft error => more power and area



Src: Kriebel et al.; ASER, @DAC‘14

Adaptive Soft Error Resilience: Turning Dark Silicon Problem into a Solution

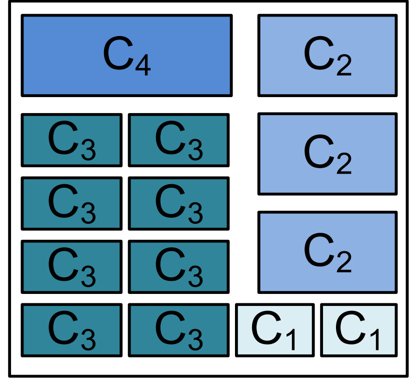
Applications' varying error vulnerability and masking probabilities show different reliability requirements



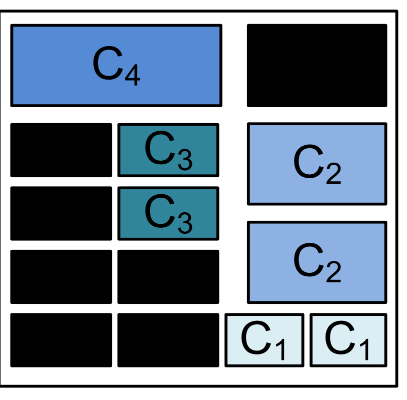
Src: Kriebel et al.; ASER, @DAC'14

ASER – Adaptive Soft Error Resilience: Patterning and Mapping for Reliability Boosting

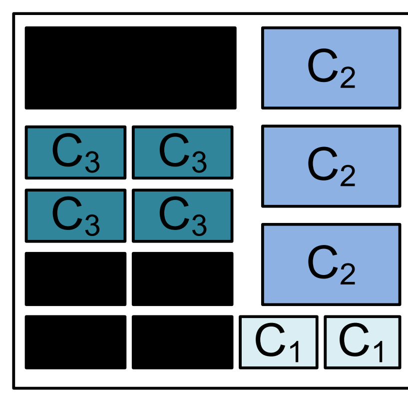
- Within the chip's **TDP constraint**, *only a subset of cores can be powered-on at run-time* and remaining cores stay dark
- A run-time system to manage reliability under thermal constraints.



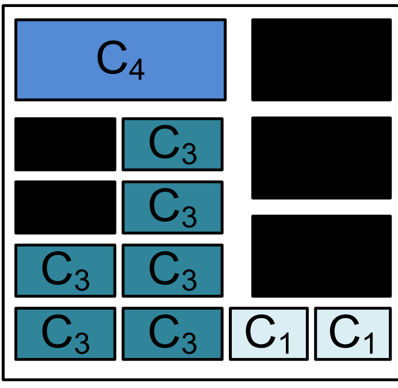
darkRHP @Design-Time



Run-Time Context-1



Run-Time Context-2



Run-Time Context-3

Src: Kriebel et al.; ASER, @DAC'14

Reliability Heterogenous Cores: Synthesis Results

	frequency = 250 MHz				frequency = 1 GHz			
	area[# of gate Eq.*10^5]	Power[mW]			area[# of gate Eq.*10^5]	Power[mW]		
		leakage	dynamic	total		leakage	dynamic	total
C1	**4.83**	**4.36**	**78.19**	**82.55**	**5.00**	**4.76**	**269.58**	**274.34**
C2	4.99	4.50	79.75	84.25	5.20	4.98	274.91	279.89
C3	13.62	12.26	223.39	235.65	14.26	13.83	767.73	781.56
C4	5.41	4.88	86.46	91.35	5.58	5.23	298.21	303.44
C5	13.77	12.40	224.96	237.36	14.46	14.08	773.52	787.60
C6	5.56	5.03	88.02	93.05	5.77	5.52	303.45	308.98
C7	14.19	12.79	231.73	244.51	14.94	14.35	796.26	810.61
C8	**14.35**	**12.93**	**233.30**	**246.23**	**15.02**	**14.56**	**801.79**	**816.34**

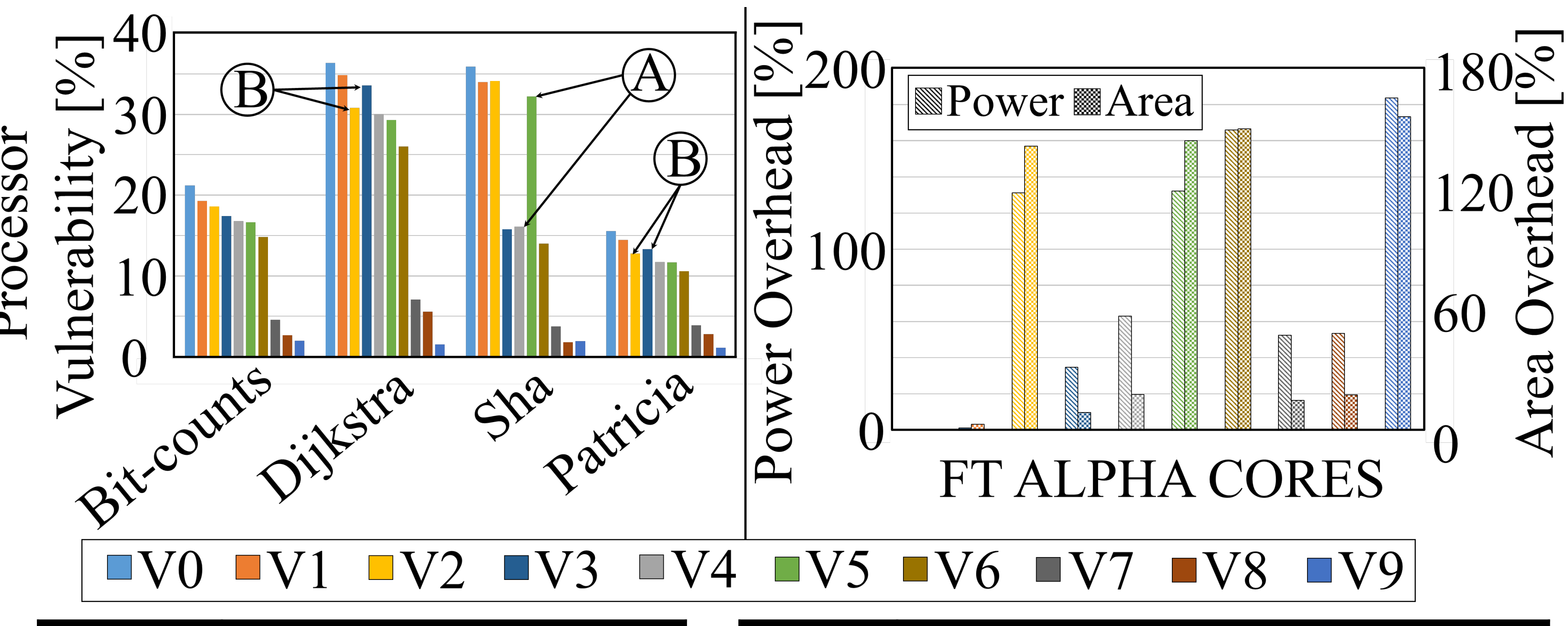
- **TSMC 45nm technology library**
- **Different process corners & frequencies**

Reliability Savings are 20%-60% compared to state-of-the-art

C1	**Baseline core**
C2	Pipeline TMR
C3	Cache TMR
C4	Register File TMR
C5	Pipeline TMR + Cache TMR
C6	Pipeline TMR + Register File TMR
C7	Cache TMR + Register File TMR
C8	Pipeline TMR + Cache TMR + Register File TMR

Src: Kriebel et al.; ASER, @DAC'14

Superscalar Processors: Fault-Tolerant Alpha Cores



Variant	Hardened Components
V0	Unprotected
V1	Int. RF, FP RF
V2	IQ, LQ, SQ
V3	Int. RF, FP RF, RM
V4	IQ, RM

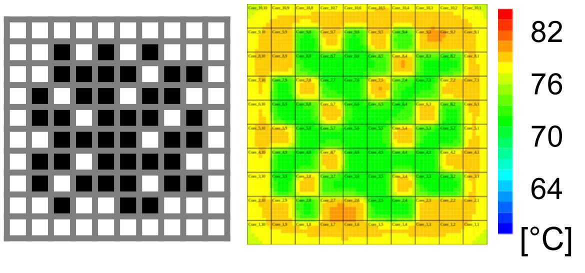
Variant	Hardened Components
V5	Int. RF, FP RF, IQ, LQ, SQ
V6	Int. RF, FP RF, IQ, LQ, SQ, RM
V7	RM, ROB
V8	Int. RF, FP RF, RM, ROB
V9	IQ, LQ, SQ, RM, ROB

Src: Kriebel & Bharath et al.; @IEEE Access 2019

Mitigating the Power Density and Dark Silicon:
Techniques for *High Performance* and *Reliability*

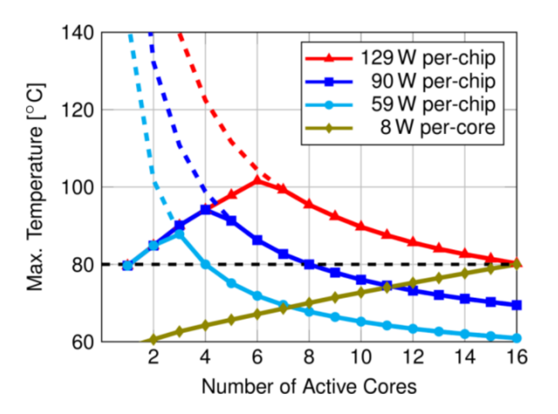
Dark Silicon Management

(Patterning and Resource Management)



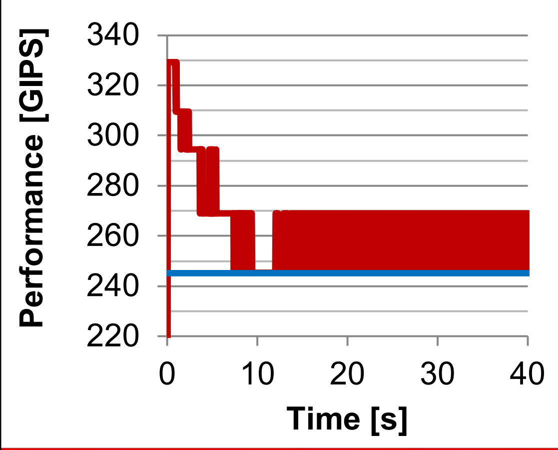
Thermal Safe Power (TSP)

(Abstract from temperature using efficient power budgets)

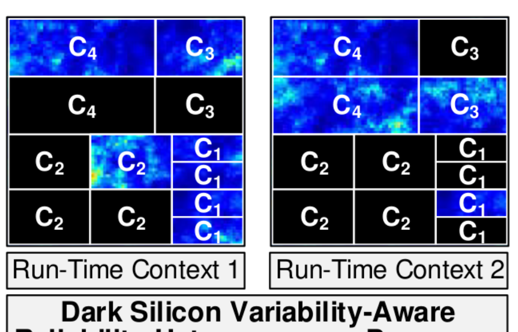


STC / NTC vs. Boosting

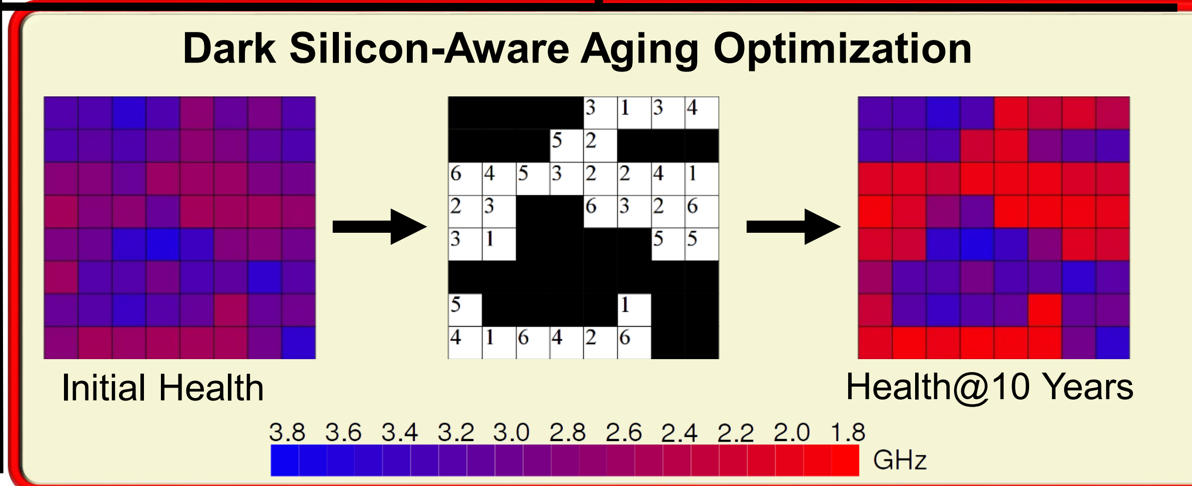
(Constant frequency vs. control-loop based boosting)



Dark Silicon-Aware Soft Error Tolerance



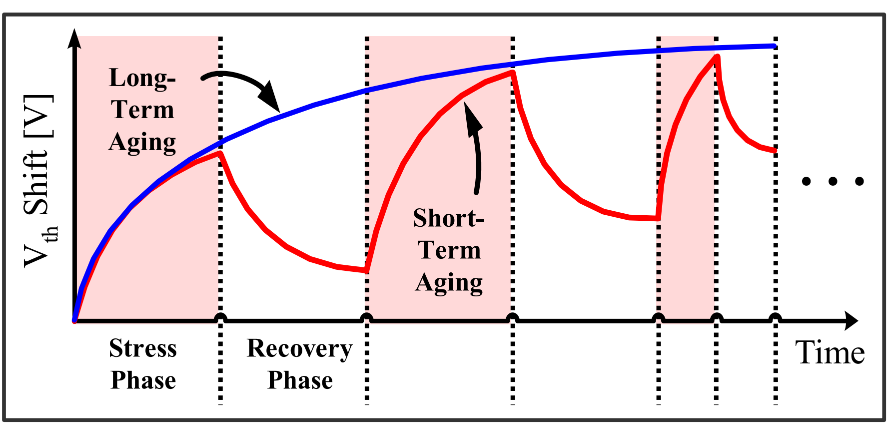
Dark Silicon-Aware Aging Optimization



Shafique et al., @ DAC (2014, 15), DATE (2015, 16), CODES+ISSS (2014, 15), TC (2017, 18), D&T (2016)

High Temperature => Aggravates Reliability Threats: NBTI-induced Aging

Negative Bias Temperature Instability (affects mostly PMOS)

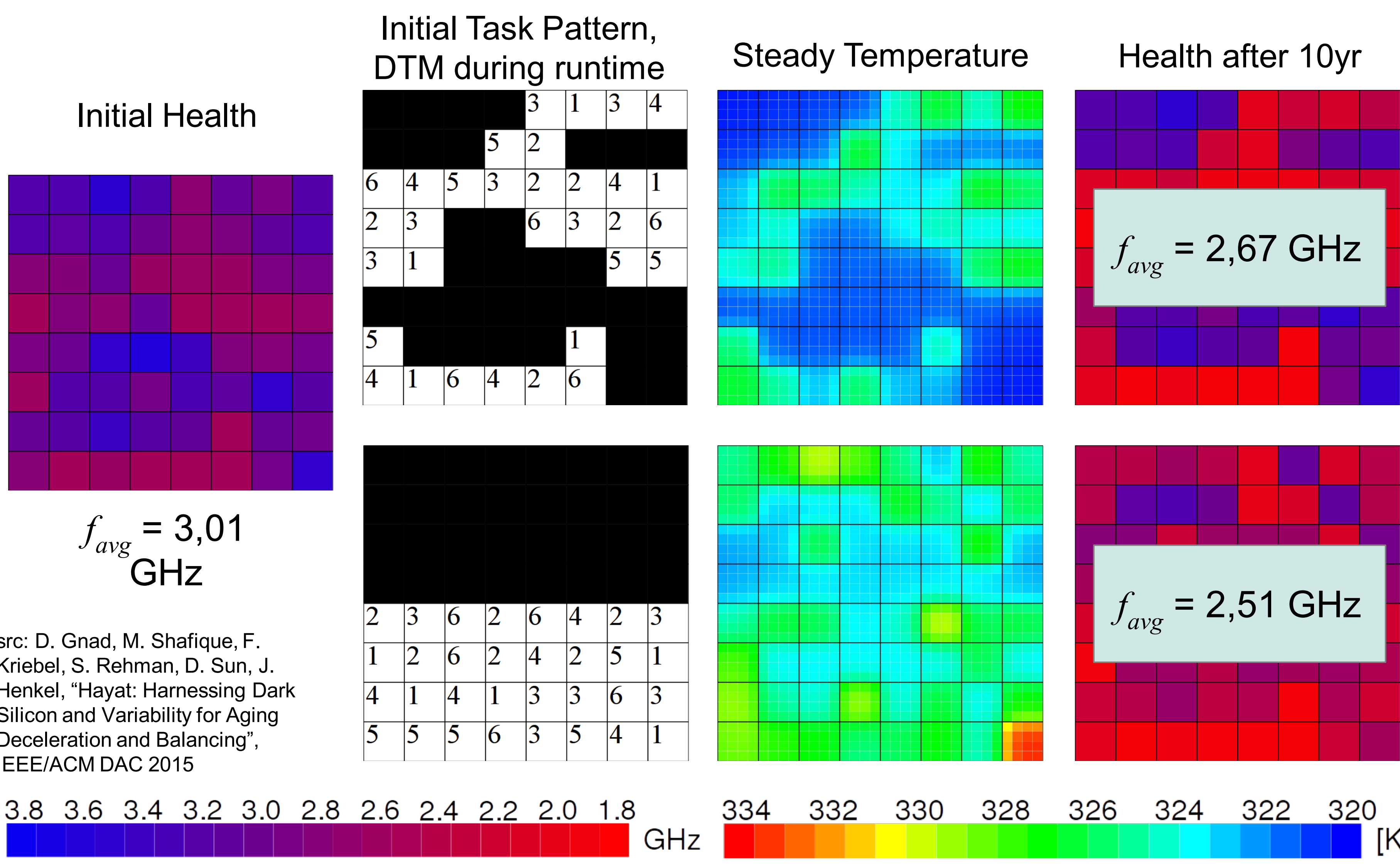


early-aging phase (usage/duty-cycle critical)

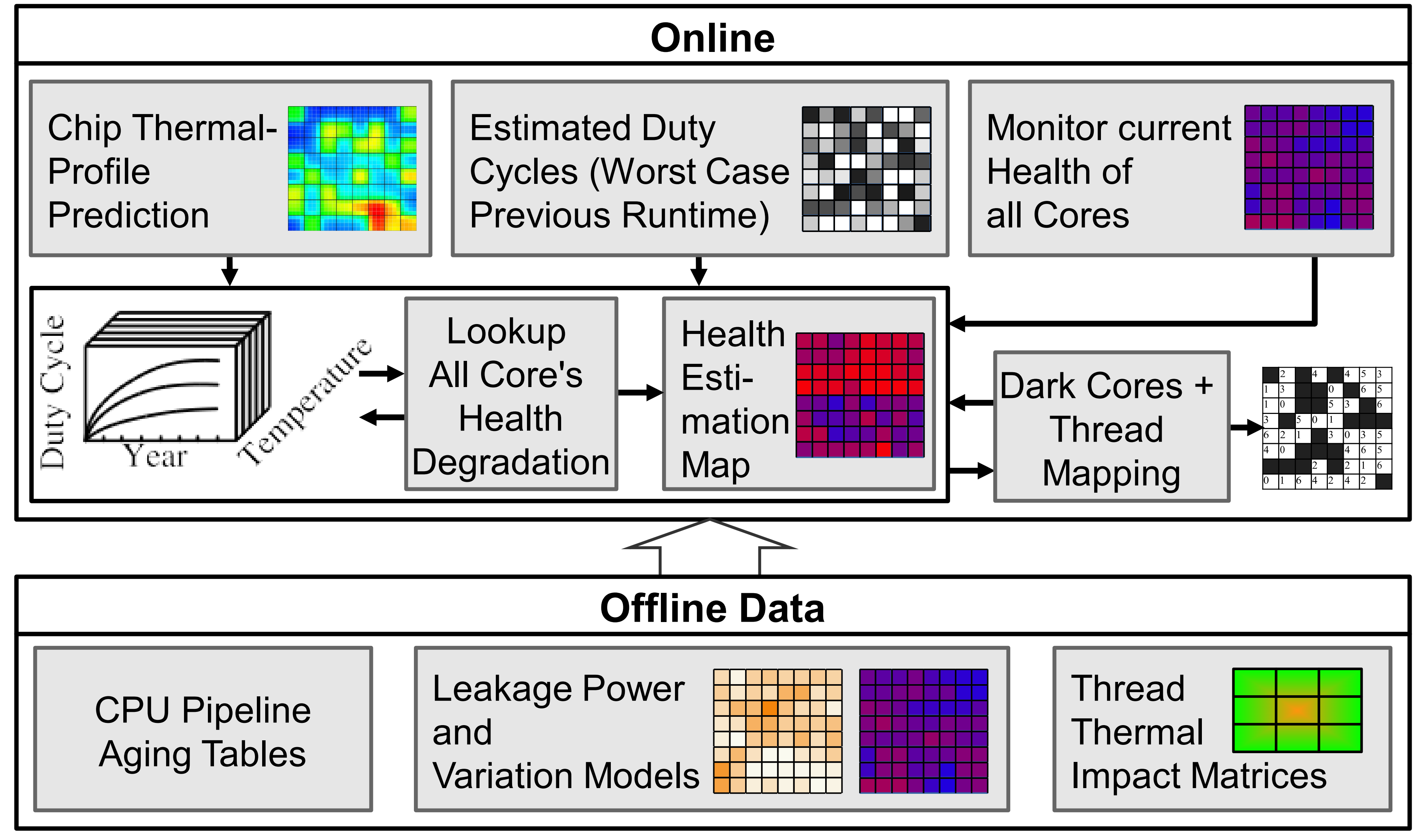
late-aging phase (temperature critical)

Src: Gnad and Shafique et al.; Hayat, @DAC'15

Impact of Dark Silicon on the Chip Aging Profile

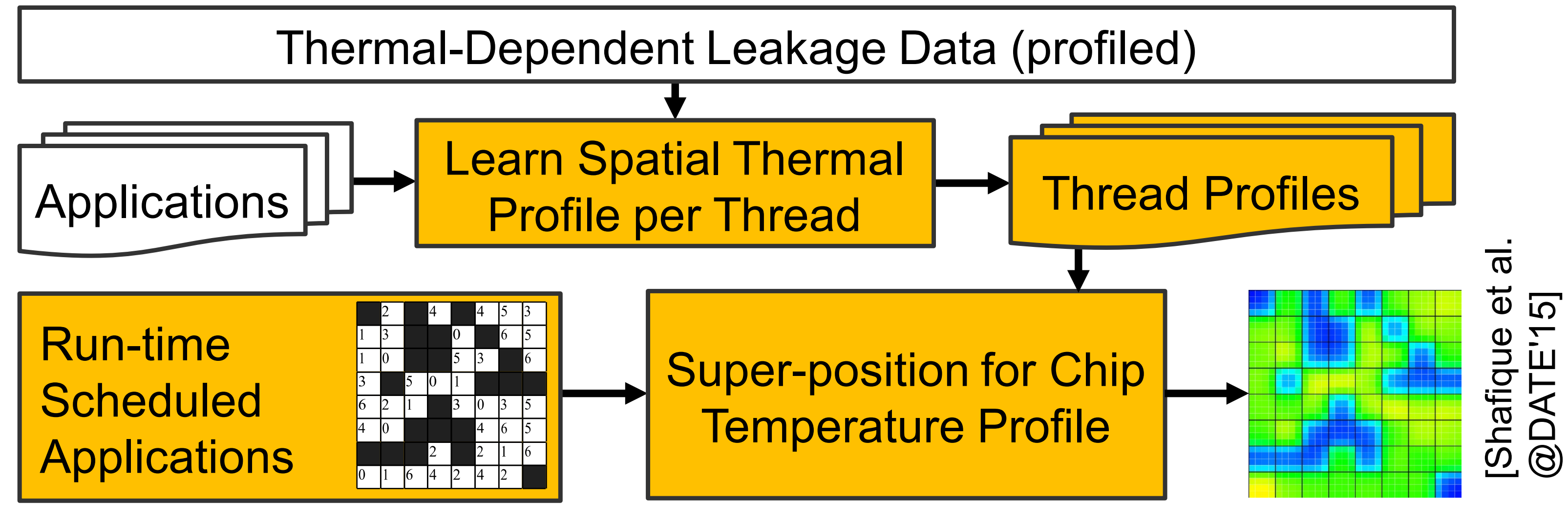


Hayat: Dark Silicon-Aware Aging Optimization

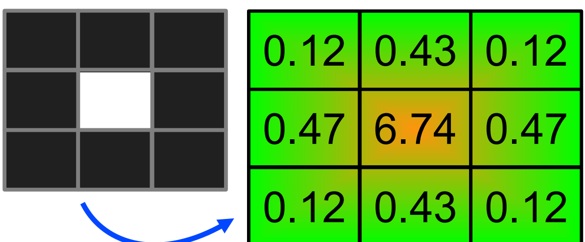


Src: Gnad and Shafique et al.; Hayat, @DAC'15

Variability-Aware Dark Silicon Management: Online Temperature Prediction



1. **Offline-profile** linear offset of temperature influence, e.g., 3x3



0.12	0.43	0.12
0.47	6.74	0.47
0.12	0.43	0.12

2. Use profiles for **lightweight Online Temperature Prediction**

	0.13	0.45	0.13
0.13	0.92	6.94	0.49
0.49	6.94	0.92	0.13
0.13	0.45	0.13	

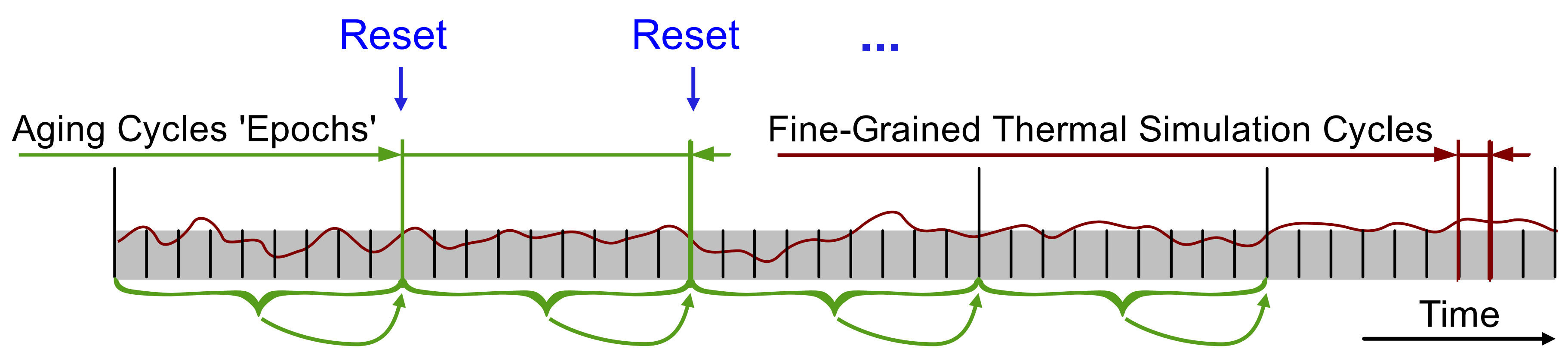
…

Refine for thermal-dependent leakage

Src: Gnad and Shafique et al.; @DATE'15 and DAC'15

Hayat: Dark Silicon-Aware Aging Optimization
Evaluation with Epochs

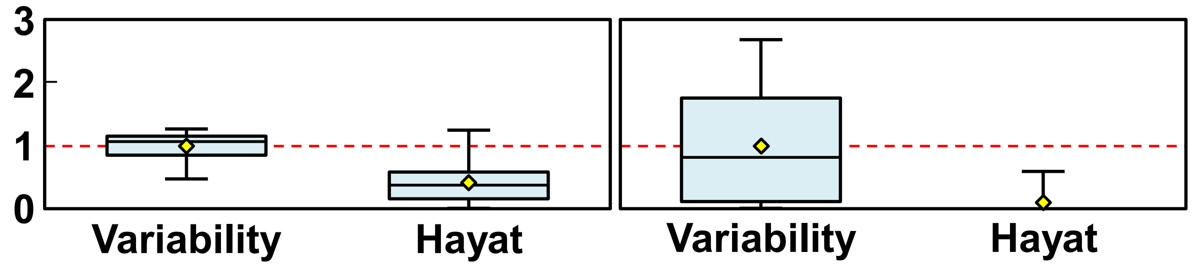
- Issues evaluating aging for longer times (complete system):
  - Longterm-Aging → Transient simulation would take *years*
    → Run *fine-grained* thermal simulation
    → Run *coarse-grained* aging cycles ("Epochs")
  - Evaluating the system capabilities at different aging states
    → Simulation of the same scenario for each epoch



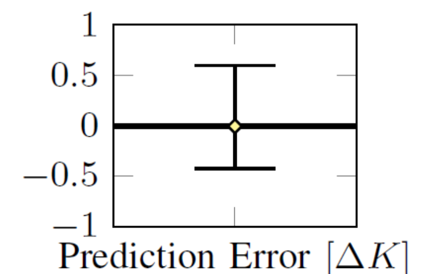
Src: Gnad and Shafique et al.; Hayat, @DAC'15

Results: Aging and DTM Events

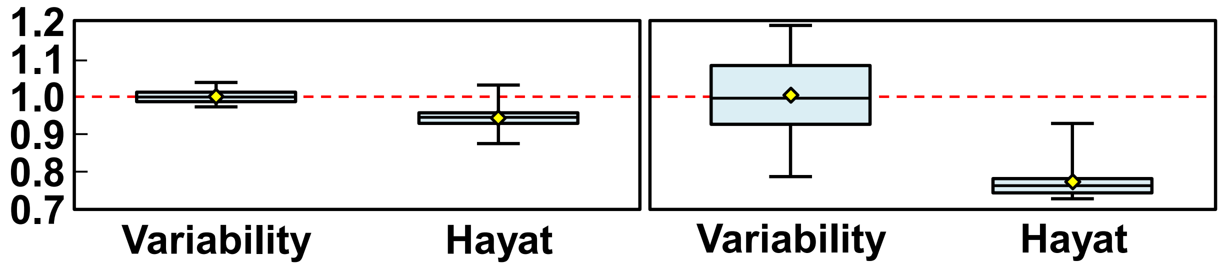
Aging Rate of Chip-Level f_{max} (Δf / 10 year)



Temperature Accuracy

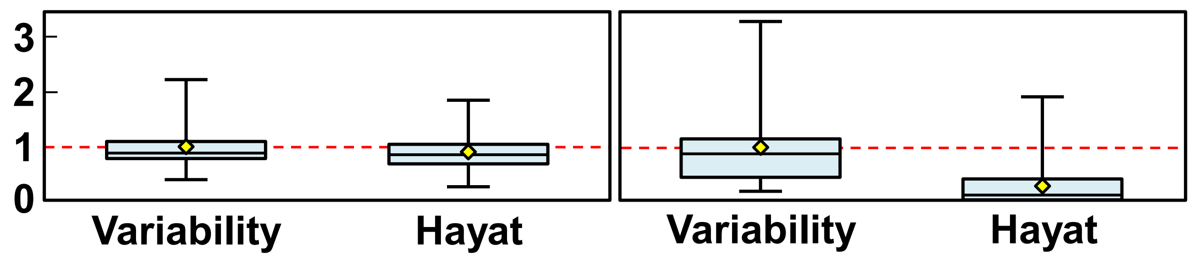


Aging Rate of Chip-Average f (Δf / 10 year)



- Left: Min. 25% Dark Silicon
- Right: Min. 50% Dark Silicon

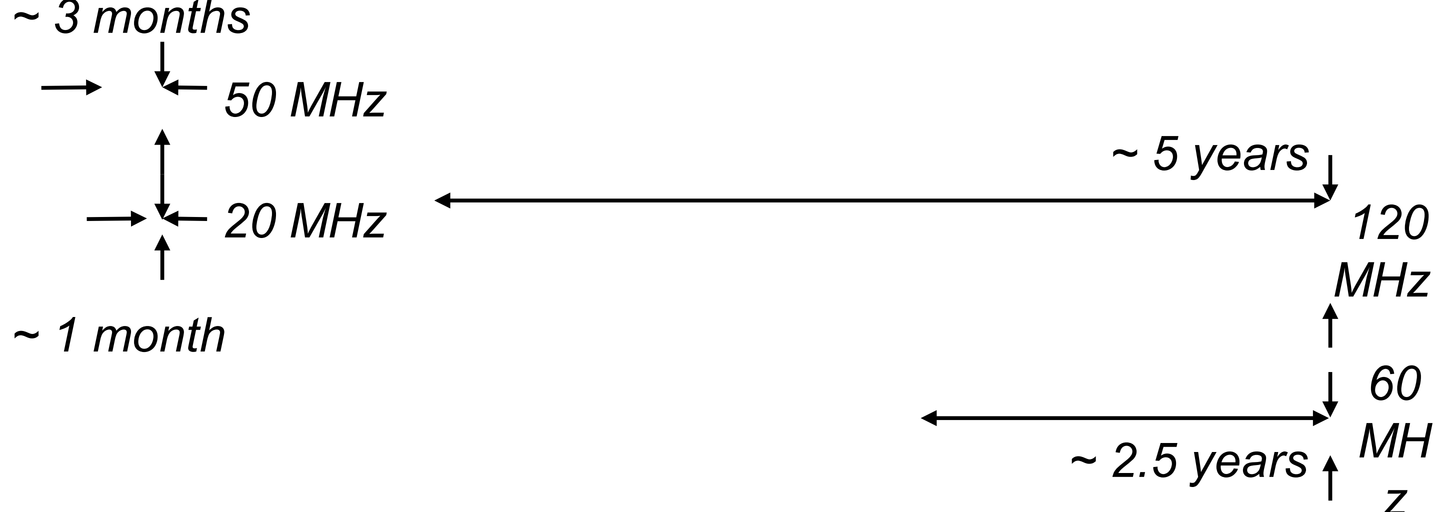
DTM Events



Src: Gnad and Shafique et al.; Hayat, @DAC'15

Hayat: Dark Silicon-Aware Aging Optimization: Results

Average Max. Frequency over all Chips for 25% and 50% Dark Silicon



Src: Gnad and Shafique et al.; Hayat, @DAC'15

Outline

- Different Types of Reliability Threats
- **Cross-Layer Resilience**
  - Modeling => Bridging the Gap between HW and SW
  - Optimization => Engage Multiple Layers of the System Stack
  - A Self-Healing Framework for Building Resilient CPS
  - Power / Temperature Considerations for Resilience
- **Robust Machine Learning**
- Conclusion

Robustness for Machine Learning: News Feed

Beware: Galaxy S10's Facial Recognition Easily Fooled with a Photo

Jesus Diaz • Freelance Writer
Updated Mar 11, 2019

Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian

Tempe police said car was in autonomous mode at the time of the crash and that the vehicle hit a woman who later died at a hospital

Hackers trick a Tesla into veering into the wrong lane

https://www.youtube.com/watch?v=a7L51u23YoM

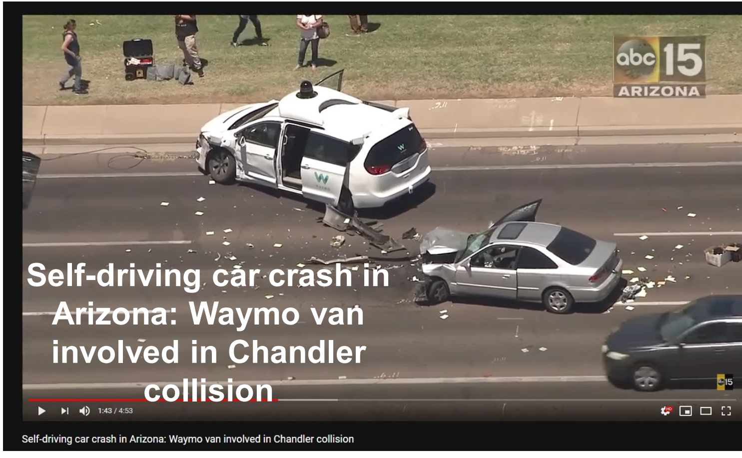
BBC

Tesla Model 3: Autopilot engaged during fatal crash

17 May 2019

Tesla driver dies in first fatal crash while using autopilot mode

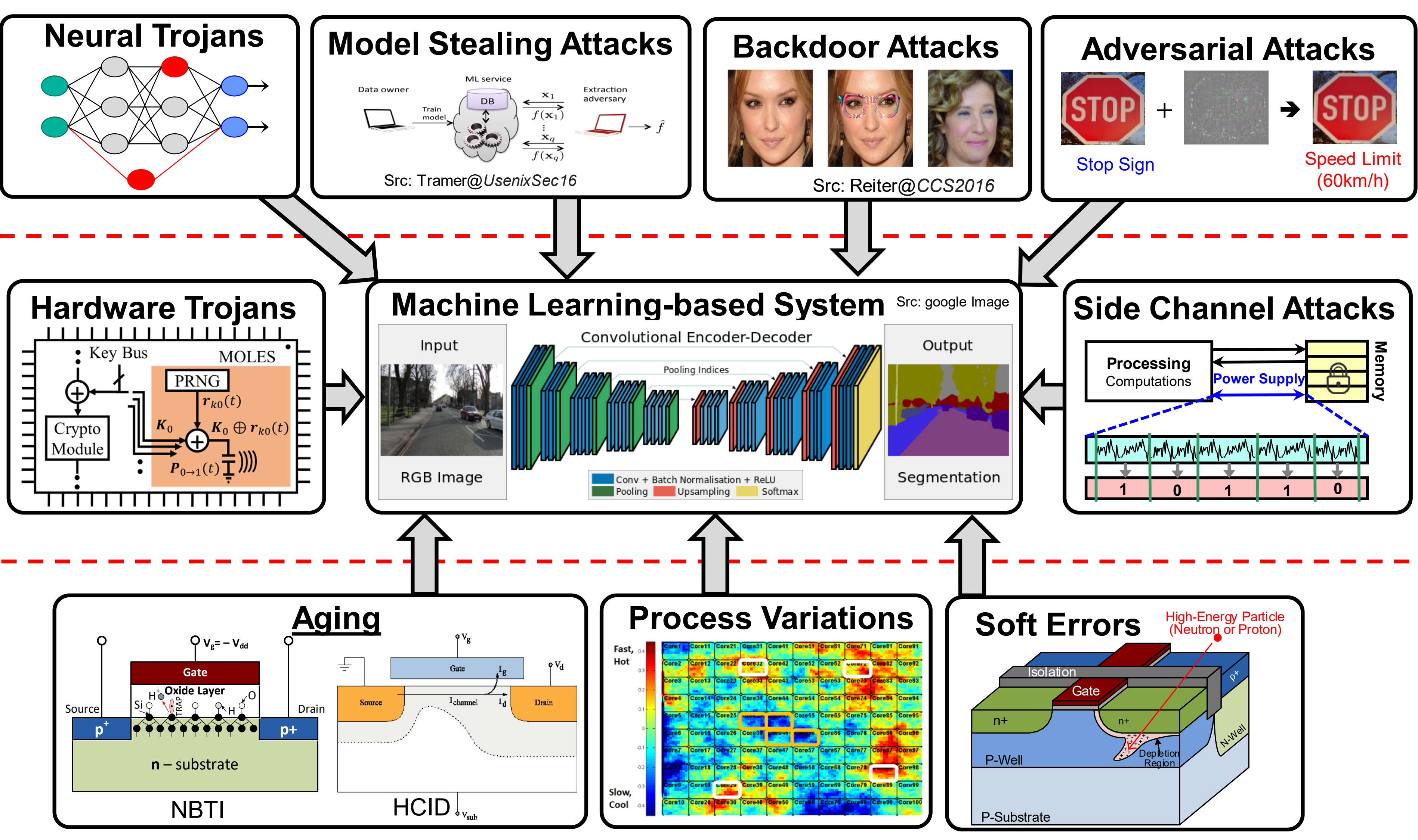
The autopilot sensors on the Model S failed to distinguish a white tractor-trailer crossing the highway against a bright sky



Self-driving car crash in Arizona: Waymo van involved in Chandler collision

https://www.technologyreview.com/f/613254/hackers-trick-teslas-autopilot-into-veering-towards-oncoming-traffic/

Reliability and Security for Machine Learning Systems

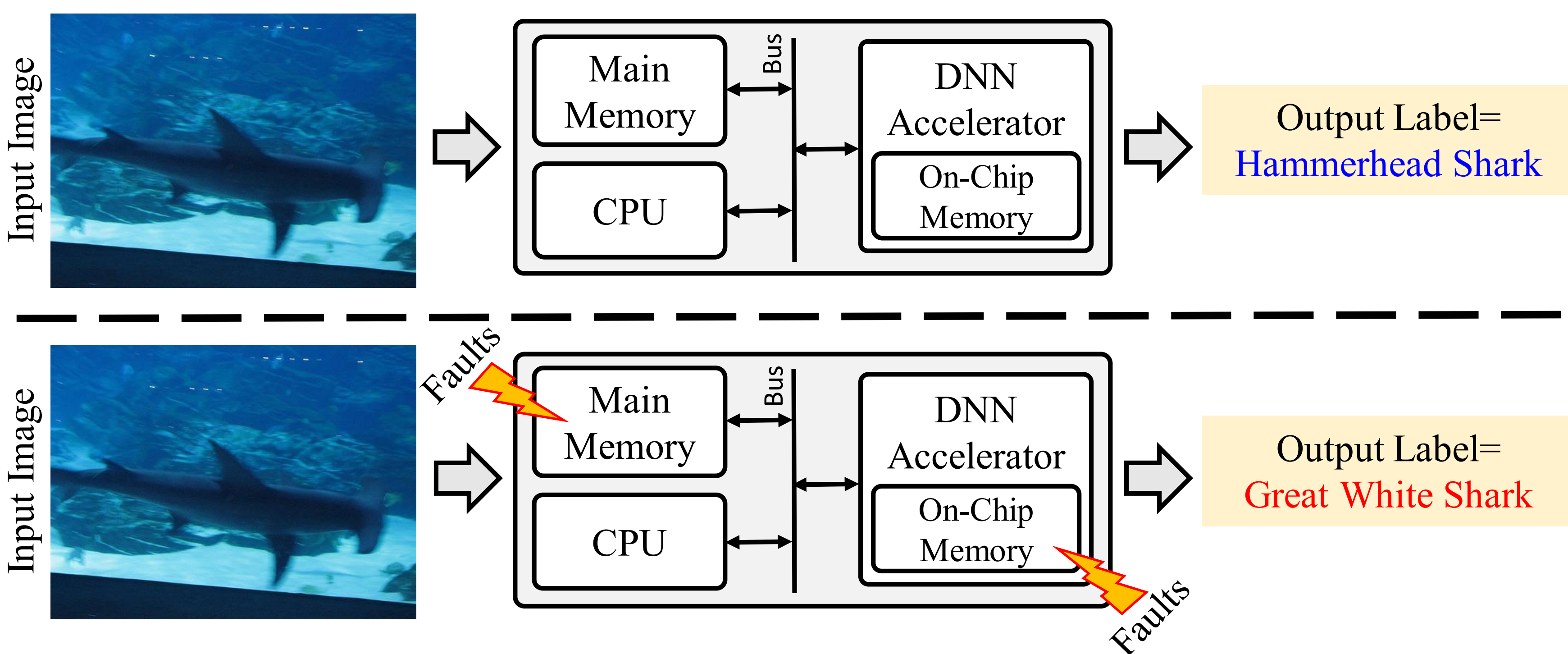


- M. A. Hanif, F. Khalid, R. V. W. Putra, S. Rehman, M. Shafique, "Robust Machine Learning Systems: Reliability and Security for Deep Neural Networks", in IOLTS-2018, Platja d'Aro, Spain, pp. 257 - 260.
- F. Kriebel, S. Rehman, M. A. Hanif, F. Khalid, M. Shafique, "Robustness for Smart Cyber-Physical Systems and Internet-of-Things: From Adaptive Robustness Methods to Reliability and Security for Machine Learning", ISVLSI-2018, Hong Kong, China, pp. 581-586.

Motivation for Reliability in Machine Learning

- Faults injection in weight memory of a DNN

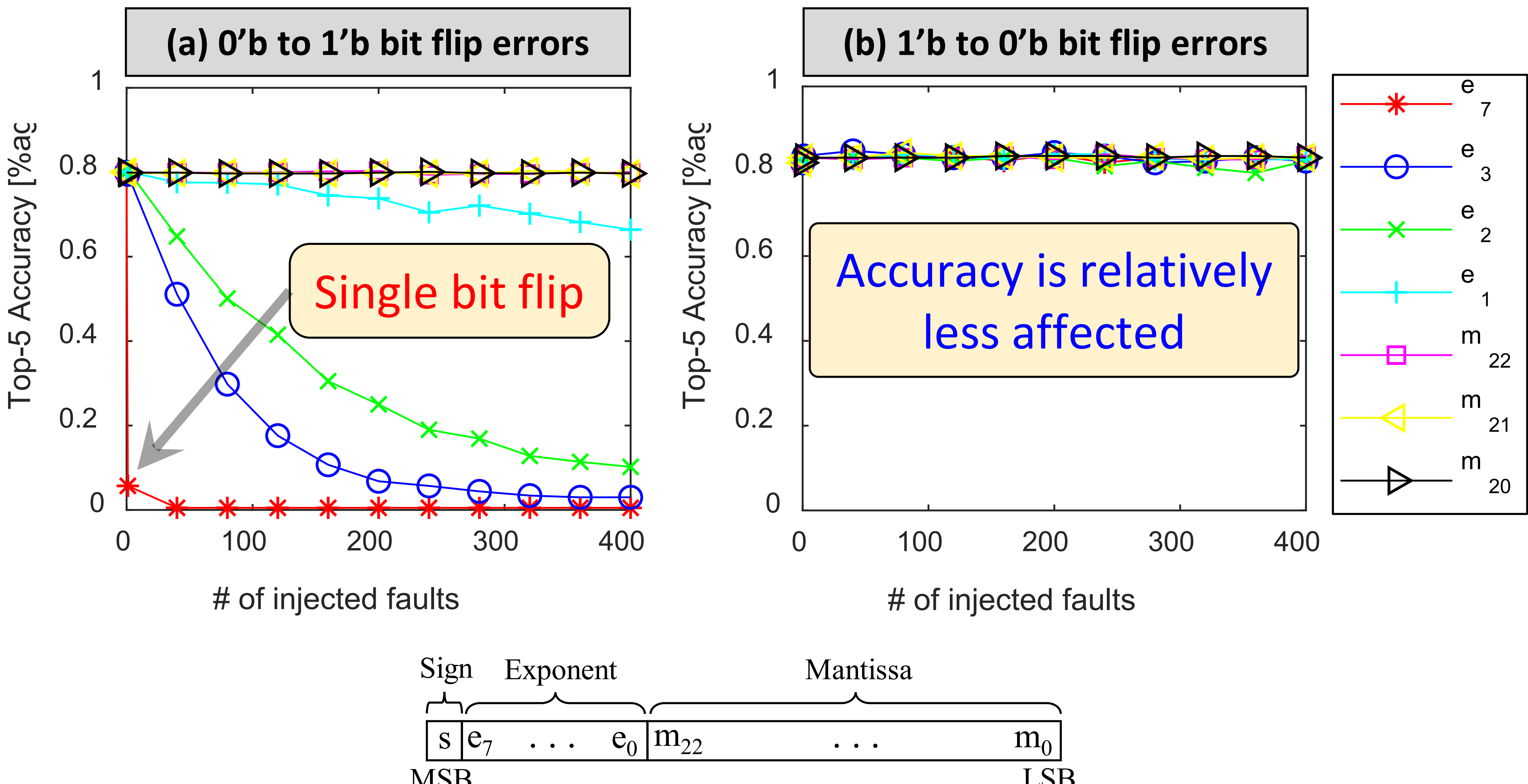
- Assuming single precision floating point weights and activations



Hanif, Khalid, Putra, Rehman, Shafique, *"Robust Machine Learning Systems: Reliability and Security for Deep Neural Networks"*, IOLTS 2018

Impact of Memory Bit-Flip Errors on Accuracy

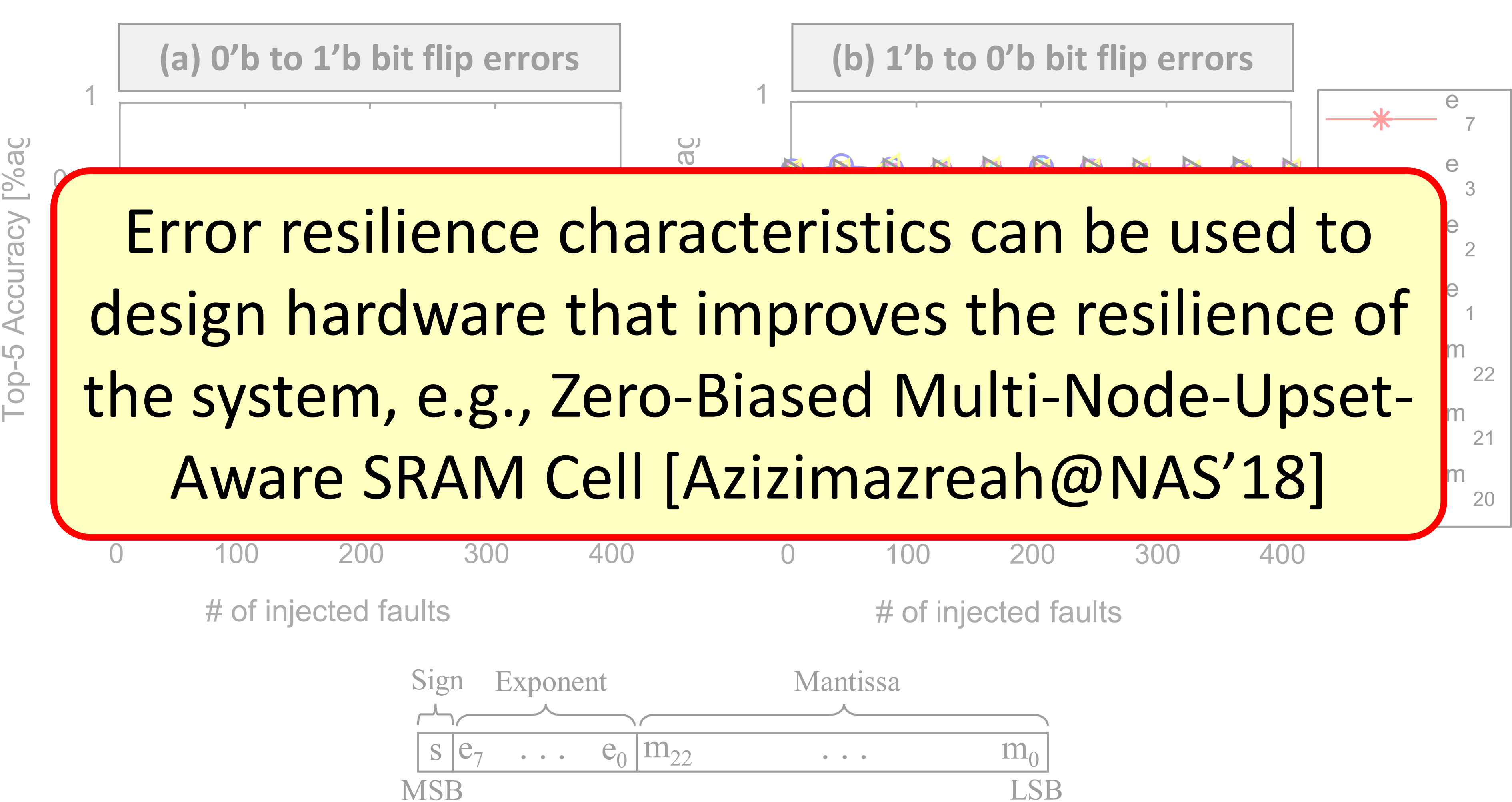
- Impact of memory bit-flip errors on accuracy



Hanif, Khalid, Putra, Rehman, Shafique, *"Robust Machine Learning Systems: Reliability and Security for Deep Neural Networks"*, IOLTS 2018

Impact of Memory Bit-Flip Errors on Accuracy

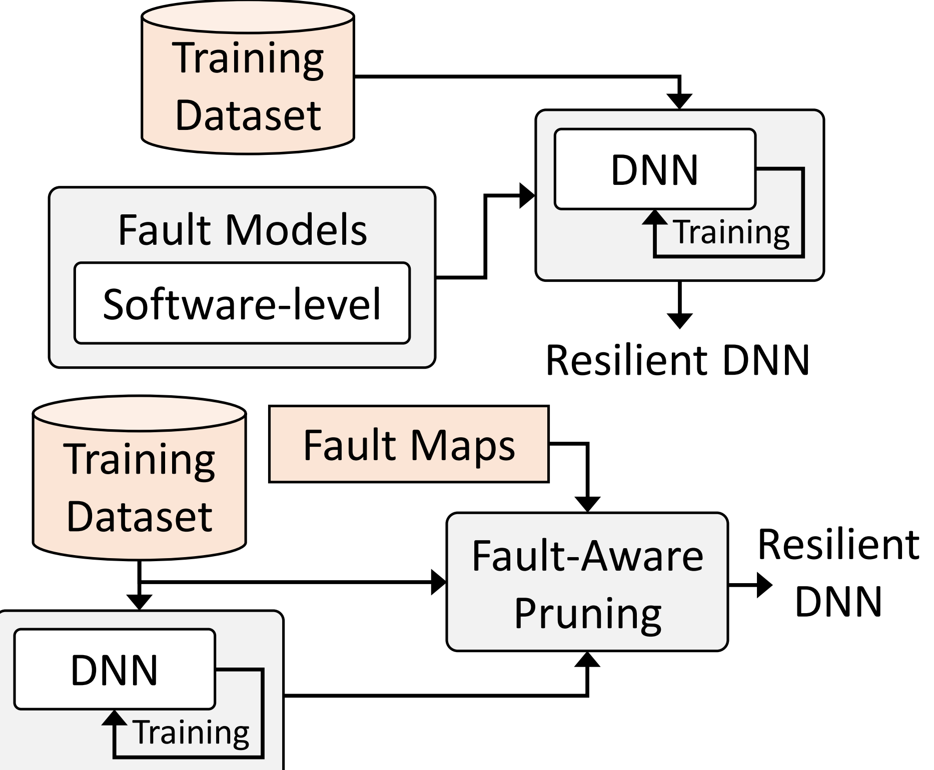
- Impact of memory bit-flip errors on accuracy



Hanif, Khalid, Putra, Rehman, Shafique, *"Robust Machine Learning Systems: Reliability and Security for Deep Neural Networks"*, IOLTS 2018

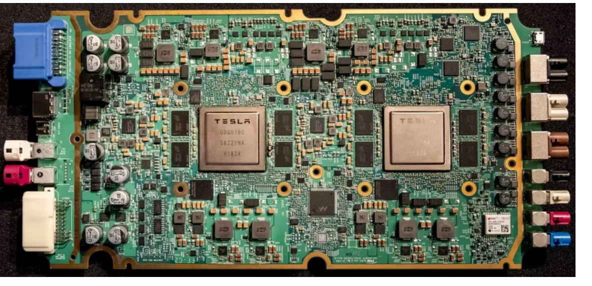
Techniques for Fault Mitigation

- ❑ **Fault-aware Training**
  - ❑ Make the training process aware of the faults
  - ❑ DNNs adapt as per the error characteristics of the system



- ❑ **Fault-aware Optimizations**
  - ❑ Make the DNN optimization process aware of the faults

- ❑ **Redundancy**
  - ❑ Use redundant components to verify the decisions made by the system

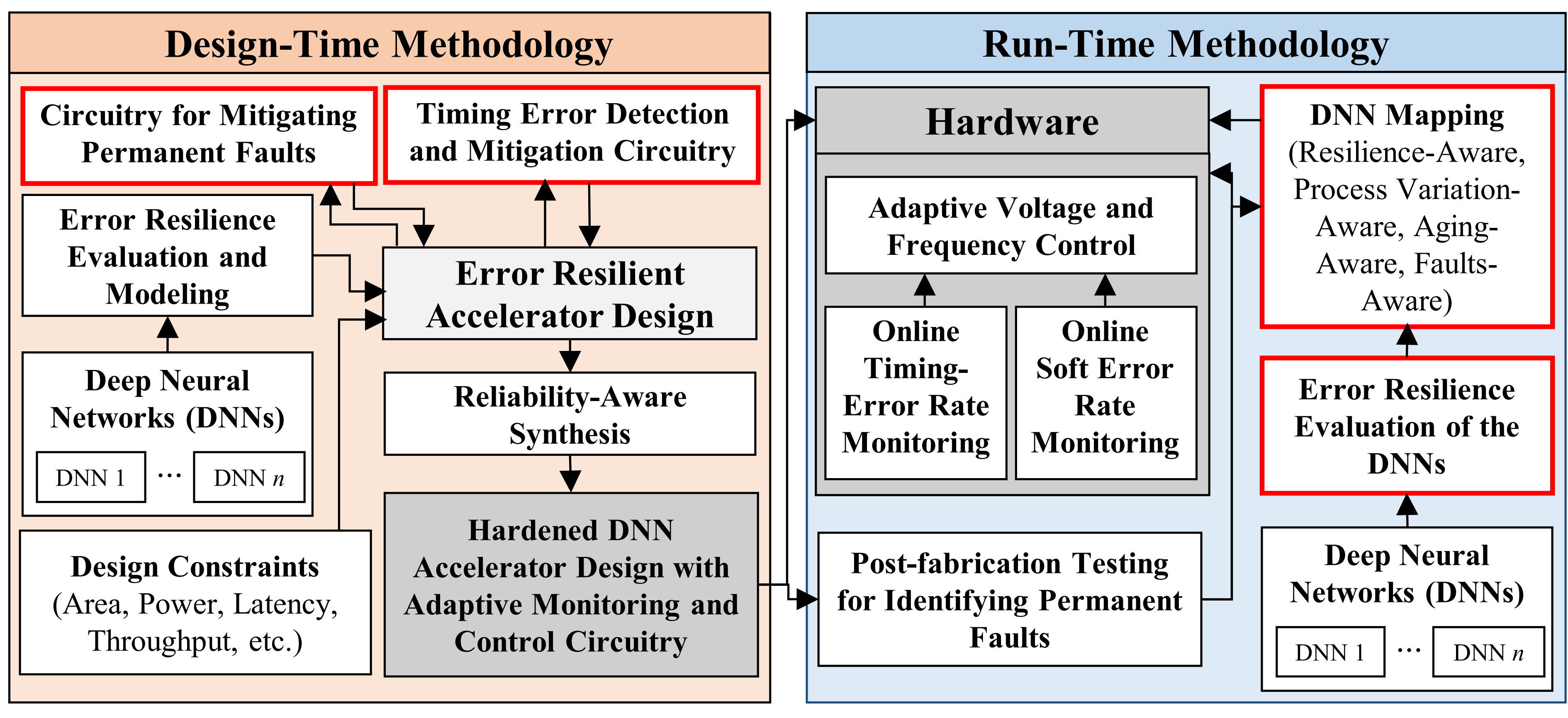


Tesla's Two AI processors hardware for Self Driving Cars

Src: Siddharth, Theo and Shafique et al.; @DAC'19

Reliable Systems for ML Applications

- Our methodology for designing reliable systems for ML-based applications



Hanif, Khalid, Putra, Rehman, Shafique, *"Robust Machine Learning Systems: Reliability and Security for Deep Neural Networks"*, IOLTS 2018

Details on Secure ML: Refer to Our Works

- **Tutorial** on *"Adversarial ML and Vehicular Networks: Strategies for Attack and Defense"* at IEEE 89th Vehicular Technology Conference (VTC).
- **Selectd Research Papers**
  - **DAC'19:** "Building Robust Machine Learning Systems: Current Progress, Research Challenges, and Opportunities".
  - **ICML'19** Workshop on Robust ML: "CapsAttacks: Robust and Imperceptible Adversarial Attacks on Capsule Networks"
  - **DATE'19:** "FAdeML: Understanding the Impact of Pre-Processing Noise Filtering on Adversarial Machine Learning"
  - **IOLTS'19:** "TrISec: Training Data-Unaware Imperceptible Security Attacks on Deep Neural Networks".
  - **IOLTS'19:** "QuSecNets: Quantization-based Defense Mechanism for Securing Deep Neural Network against Adversarial Attacks".

Summary of Cross-Layer Resilience

- Each new technology node introduces new reliability problems or makes existing ones worse
- **Cross-Layer is the Key:**
  - Engage multiple HW/SW Layers
    => Bridge the gap between hardware and software
  - Heterogeneity can be used for improved fault tolerance,
    => turning dark silicon from a problem to a solution
  - Multi-objective optimization based on applications' requirements under power/thermal constraints
- **Robust ML** needs to account for both reliability and security vulnerabilities at design and run time.
  => Traditional solutions would be too costly for ML-based Systems
- **Open-Sourcing** is important to progress as a community

Tools and Open-Source Releases

- Reliability-driven Instruction Set Simulator with Fault Injection and Reliable compiled binaries
- Gate-Level Aging and Masking Estimator (GAME)
- C# based Memory Aging Analysis Tool
- Multiple Libraries of Approximate Adders and Multipliers
- A Lightweight Multithreaded HEVC Software

Selected GetSURE Publications on Robust Computing

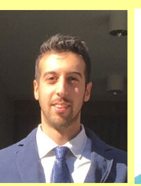
- S. Rehman, M. Shafique, J. Henkel, "Reliable Software for Unreliable Hardware – A Cross Layer Perspective", **Book: Springer** Science+Business Media, LLC, 2016.
- S. Rehman, K.-H. Chen, F. Kriebel, A. Toma, M. Shafique, J.-J. Chen, J. Henkel, "Cross-Layer Software Dependability on Unreliable Hardware", in IEEE Transactions on Computers (**TC**), vol. 65, no. 1, pp. 80-94, 2016.
- D. Gnad, M. Shafique, F. Kriebel, S. Rehman, D. Sun, J. Henkel, "Hayat: Harnessing Dark Silicon and Variability for Aging Deceleration and Balancing", **DAC**, 2015.
- S. Rehman, F. Kriebel, D. Sun, M. Shafique, J. Henkel, "dTune: Leveraging Reliable Code Generation for Adaptive Dependability Tuning under Process Variation and Aging-Induced Effects", **DAC**, 2014.
- F. Kriebel, S. Rehman, D. Sun, M. Shafique, J. Henkel, "ASER: Adaptive Soft Error Resilience for Reliability-Heterogeneous Processors in the Dark Silicon Era", **DAC**, 2014.
- M. Shafique, S. Rehman, P. V. Aceituno, J. Henkel, "Exploiting Program-Level Masking and Error Propagation for Constrained Reliability Optimization", **DAC**, 2013.
- S. Rehman, A. Toma, F. Kriebel, M. Shafique, J.-J. Chen, J. Henkel, "Reliable Code Generation and Execution on Unreliable Hardware under Joint Functional and Timing Reliability Considerations", **RTAS**, 2013.
- S. Rehman, M. Shafique, J. Henkel, "Instruction Scheduling for Reliability-Driven Compilation", **DAC**, pp. 1288-1296, 2012.
- S. Rehman, M. Shafique, F. Kriebel, J. Henkel, "Reliable Software for Unreliable Hardware: Embedded Code Generation aiming at Reliability", **CODES+ISSS**, 2011, **Best Paper Award**.

Selected GetSURE Publications on Robust Computing

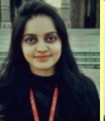
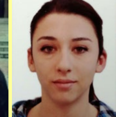
- F. Kriebel, S. Rehman, M. Shafique, S. Garg, J. Henkel, "Variability- and Reliability-Awareness in the Age of Dark Silicon", IEEE Design and Test (**D&T**), vol. 32, no. 4, 2016.
- M. Salehi, M. K. Tavana, S. Rehman, M. Shafique, A. Ejlali, J. Henkel, "Two-State Checkpointing for Energy-Efficient Fault-Tolerance in Hard Real-Time Systems", IEEE Transactions on Very Large Scale Integration Systems (**TVLSI**), 2016.
- K.-H. Chen, J.-J. Chen, F. Kriebel, S. Rehman, M. Shafique, J. Henkel, "Task Mapping for Redundant Multithreading in Multi-Cores with Reliability and Performance Heterogeneity", in IEEE Transactions on Computers (**TC**), 2016.
- M. Shafique, M. U. K. Khan, J. Henkel, "Content-Aware Low-Power Configurable Aging Mitigation for SRAM Memories", in IEEE Transactions on Computers (**TC**), 2016.
- F. Kriebel, A. Subramaniyan, S. Rehman, S. J. B. Ahandagbe, M. Shafique, J. Henkel, "R2Cache: Reliability-Aware Reconfigurable Last-Level Cache Architecture for Multi-Cores", **CODES+ISSS**, 2015. **Best Paper Award**.
- F. Kriebel, S. Rehman, M. Shafique, J. Henkel, "ageOpt-RMT: Compiler-Driven Variation-Aware Aging Optimization for Redundant Multithreading", **DAC**, 2016.
- M. Salehi, M. Shafique, F. Kriebel, S. Rehman, M. K. Tavana, A. Ejlali, J. Henkel, "dsReliM: Power-Constrained Reliability Management in Dark-Silicon Many-Core Chips under Process Variations", **CODES+ISSS**, 2015.
- M. Salehi, M. K. Tavana, S. Rehman, F. Kriebel, M. Shafique, A. Ejlali, J. Henkel, "DRVS: Power-Efficient Reliability Management through Dynamic Redundancy and Voltage Scaling under Variations", **ISLPED**, 2015.

My Research Team and Collaborators

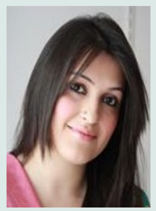
Post-Docs and PhDs



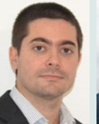
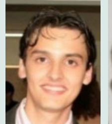
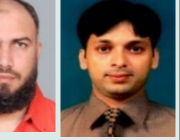
MS/BS Students

Collaborators



Previous Students

CARE Tech Computer Architecture and Robust, Energy-Efficient Technologies Group

Thank You!
Questions?