SAFER:
Single Aging Factor Enhanced Rings for Data Annotation & Early Warning in Online Aging Monitor of Automotive SoCs

Tsung Huang and Jing Huang
National Changhua Univ. of Edu.
2020/06/16

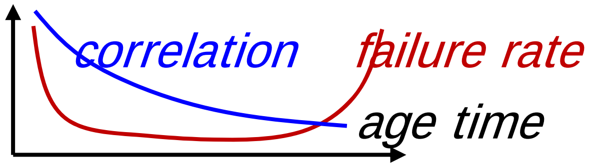
Outline

- **Introduction**
  - **Importance of Data Annotation & Early Warning**
  - **Basic concept**
- **Previous Work**
  - **AI Monitoring for Power/Thermal Management**
  - **On-Chip Accelerated Aging Trackers**
- **Proposed Online Aging Monitor**
  - **Architecture**
    - ✓ Stratified Sampling Detectors
    - ✓ Single Aging Factor Enhanced Rings
  - **Co-Learning Method**
    - ✓ 1st Stage: Data Annotation
    - ✓ 2nd Stage: Stress Adapting for SAFER Selection
- **Evaluations and Comparison**
  - **High-level profiling**
  - **Low-level parametric extraction**
- **Conclusions**

Motivation 1: Early Warning

Motivation 2: Demands on High Correlation

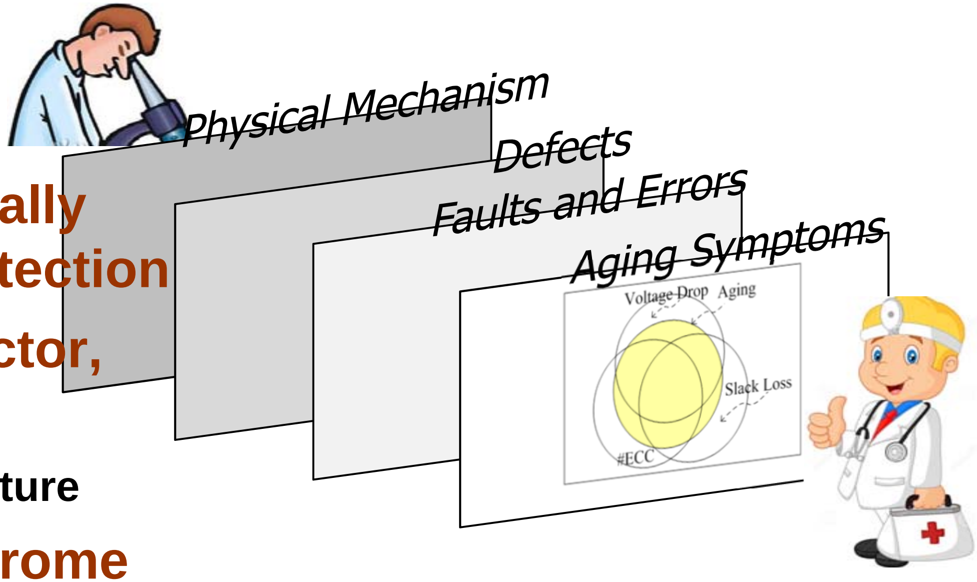
- **Basic Learnings:**
  - **Supervised:**
    - ✓ classification ➔ clustering ➔ data annotation (labeling) ➔ training
    - ✓ **with high correlation ➔ high accuracy and high safety**
  - **Reinforcement: Implied reward**
  - **Unsupervised: complicate, time-consuming, danger**
- **Correlation**

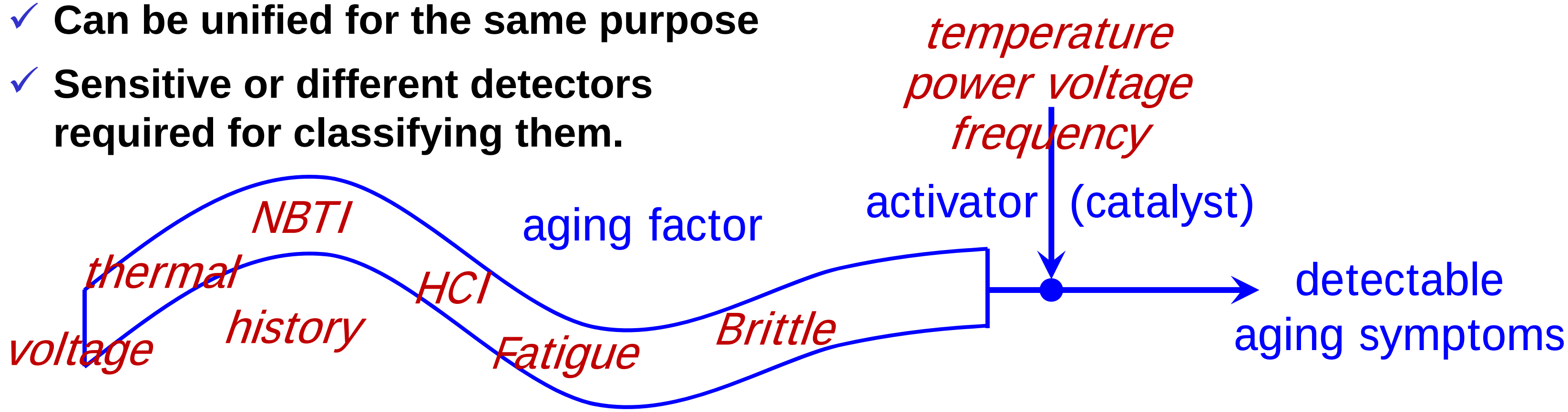


  - **Early Stage:**
    - ✓ **Inherent (Innate) Correlation:**
      - **Process Technology ➔ Lot ➔ Chip ➔ Cell Type ➔ Thermal & Backgnd.**
    - ✓ **Suitable Strategies:**
      - **Stratified Sampling**
      - **Data Annotated by the (accelerated) aged (old men ~ experienced men)**
  - **Steady Stage:**
    - ✓ **Acquired Correlation:**
      - **Required to be trained by labels**
  - **Divergence:**
    - ✓ Different operations ➔ ton duration & switching activity (sa) ➔ hot spots
    - ✓ Trained detectors

Concept of Aging Causality

- **Observation:**
  - **Diagnosis (classifier) is usually more challengeable than detection**
  - **Some parameters can be factor, catalyst and/or symptom**
    - ✓ **power/energy/thermal/temperature**
  - **Some symptoms are a syndrome due to more than two factors.**
    - ✓ **(NBTI, HCI) ➔ ΔVth ➔ slack loss**
    - ✓ **Can be unified for the same purpose**
    - ✓ **Sensitive or different detectors required for classifying them.**





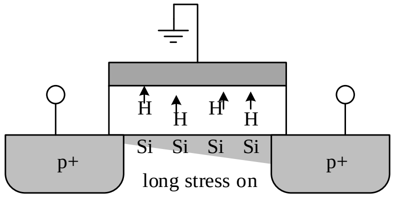
Failure Mechanisms of ICs

- Failure Mechanisms Related to the Wafer Process
  - Negative Bias Temperature Instability (NBTI) ← pMOS, ton, T
  - Hot Carrier Injection (HCI) ← nMOS, VDD, Switching, T, ta
  - Time Dependent Dielectric Breakdown (TDDB) ← tox, ta
  - Electro-migration ← Process, T, ta
  - Stress Migration
  - Soft Error
  - Reliability of Non-Volatile Memory
- Failure related to Packaging, Assembly & Use
  - Wire Bonding Reliability (Au-Al Joint Reliability) Ag Ion Migration
  - Al Sliding
  - Filler Whiskers
  - Moisture
  - Cracks
  - Electrostatic Breakdown and Electrical Overstress Breakdown
  - Latchup
  - Power MOS FET Damage

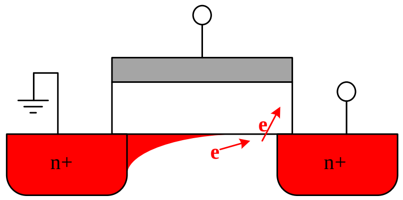
RENESAS Ltd. Semiconductor Reliability Handbook, 2017.

Major Aging Factors

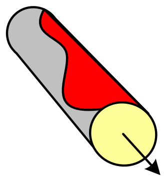
- **Negative-bias temperature instability (NBTI)**
  - **Typical α-Power Model**
    - ✓ $\Delta V_{th} = A \cdot t_{on}^{\alpha} \cdot V_s^{\beta} \cdot (1 - \eta^{0.5}) \cdot e^{-\frac{\gamma E_a}{kT}}$
    - ✓ "power" coefficients: 0.5 $< \alpha, \beta, \gamma <$ 1.0 roughly
    - ✓ Ea: activation energy. η: recovery coefficient



- **Hot Carrier Injection (HCI)**
  - **Typical α-power Model**
    - ✓ $\frac{\Delta P}{P} = A \left(\frac{I_D}{W}\right)^n \left(\frac{I_B}{I_D}\right)^{mn} t^n L^{-p} e^{-\left(\frac{1}{T} - \frac{1}{T_{ref}}\right) E_a/k}$



- **Fatigue Resistance**
  - **Basic α-power Model of thermal fatigue**
    - ✓ $\Delta \rho = \rho_o \left(\frac{T}{T_o}\right)^{\alpha} t^{\beta}$



NBTI vs. HCI

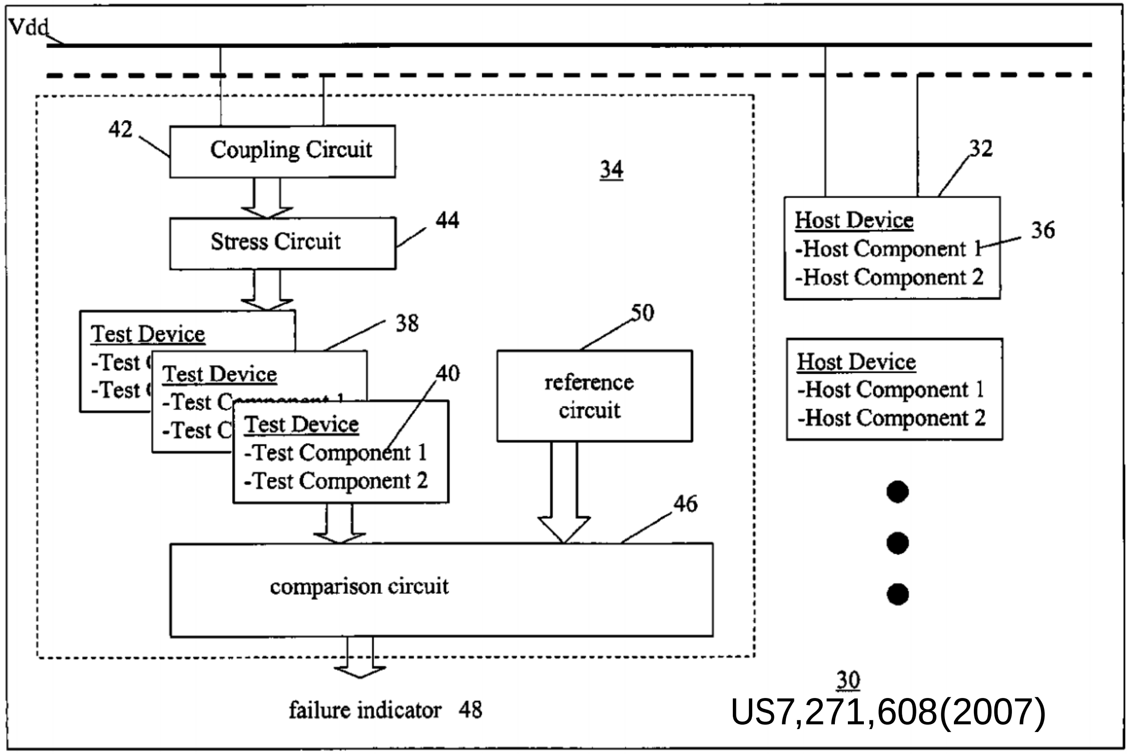
- **Hard to distinguish due to similar response**
  - **They has ever been unified for PM early.**
  - **whole life aging tracking ➔ Recurrent NN ?**

Aging	Factors		Catalysts (Activators)		Effect (Response)	
Comparison	NBTI	HCI	NBTI	HCI	NBTI	HCI
High Voltage	V	V				
High Temperature		**V**	**V**			
MOS Type Majored	pMOS	nMOS				
Frequency			**V**	**V**		
Stress State	ON	Switching				
$\Delta V_{th}/V_{th0}$					**V**	V
I_{DDQ}					**reduced**	reduced
delay					**V**	V

Y. Wang, *et al*. "A unified aging model of NBTI and HCI degradation towards lifetime reliability management for nanoscale MOSFET circuits," NanoArch2011.

Previous Work on Online Early Warning for ICs

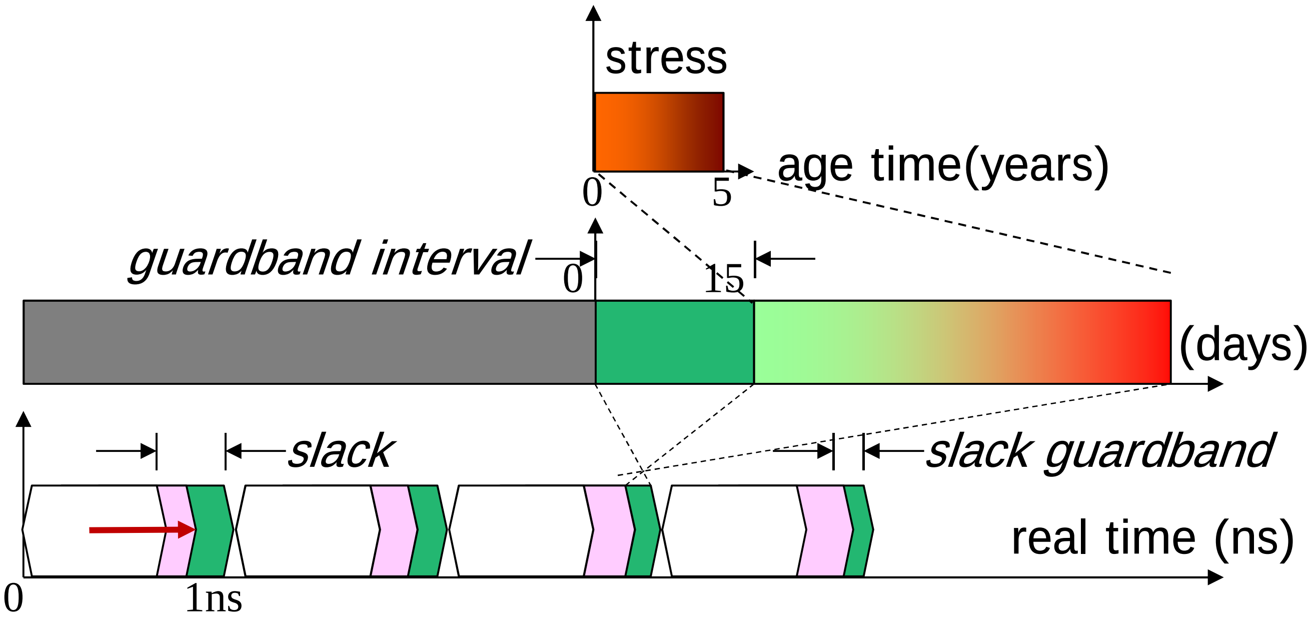
- Duplicate circuit stressed as aging tracker
  - With very high correlation
  - Compared with TMR:
    - ✓ Stressed one can be healthy, while the minority may be right.
  - High cost



Bert M. Vermeire and Harold G. Parks. Prognostic Cell for Predicting Failure of Integrated Circuits. US Patent 7,271,608 B1, Sep. 18, 2007.

Previous Work on Online Early Warning for ICs

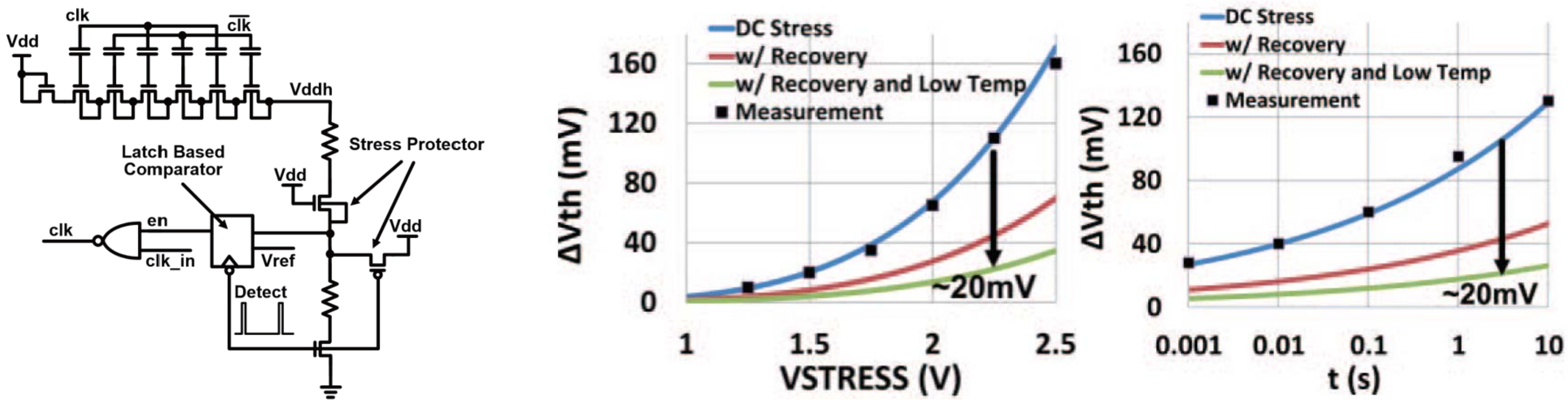
- Mitra [VTS07] proposed a concept on the worst case guardbanding.
  - Focused on NBTI-related aging
  - Taking a general circuit in 1GHz frequency is usually with 3~7 years of life to estimate the slack and guardband.
  - In an about (1 years/3 days)-aging acceleration technique, the guardband adjustment is estimated.



M. Agarwal, B. C. Paul, M. Zhang and S. Mitra, "Circuit Failure Prediction and Its Application to Transistor Aging,“ 25th IEEE VLSI Test Symposium (VTS'07), Berkeley, CA, 2007, pp. 277-286.

Previous Work on Online Early Warning for ICs

- You & Gu [ASPDAC17] shortened 15d to 12s-ton by +2.2V Stress for NBTI.
  - Proposed a charge pumper without stress infection on normal circuits by ground level shifting.
  - A set of ORs are stressed in each rebooting time for aging tracking and provide early warning.



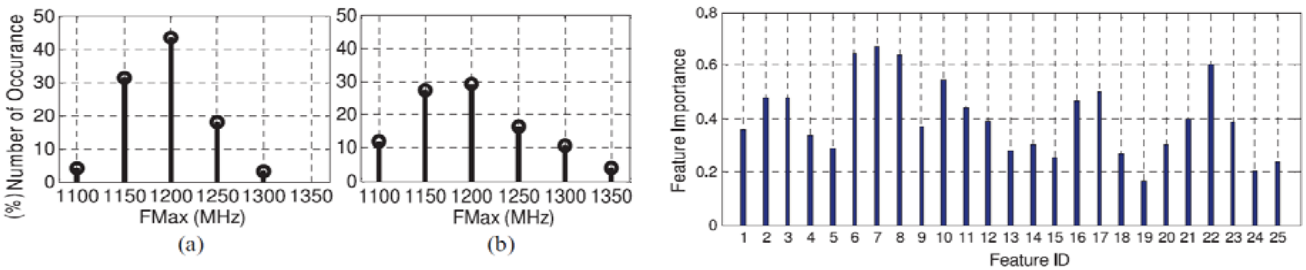
Y. You and J. Gu, "Exploiting accelerated aging effect for on-line configurability and hardware tracking," 22nd Asia & South Pacific Design Automation Conf., Chiba, 2017, pp. 348-353.

Previous Work on Online Early Warning for ICs

Major Previous Work related to On-Line Aging Monitoring Methods

References	Factor				Catalyst		Symptoms			On/Off-Line	Object	Target	Parameter	Monitoring	Meas/Sim	Learning
	NBTI	HCI	TDDB	Fatigue	None	Temp	Delay	Vdrop	f							
TCAD2014Lai							slack			online						NA
TCAD2015SYHuang							slack									NA
VTS2016Anghel									V	online						NA
JSSC2008									V							NA
TVLSI2012Wang							slack									NA
TDAES2015Firouzi							slack									NA
ASPDAC2017You	V	V				V	slack			NA		prech		Tracking	Monte Carlo	NA
ITC2013Firouzi							slack			online						NA
TCAD2017Sengupta							slack					prech			bound-derive	NA
TVLSI2017Tenentes	V							V		online						NA
TCAD2017Sadi							slack			online	TSV	Binning	speed			ML

- No work applies NN obviously except TCAD2017Sadi and PM work introduced later.
- Sadi's work is actually either an unsupervised classifier or implied by the scalar parameter – slack time.
- A famous unsupervised learning example tells about how a 1-year kid how to distinguish a DXX and a CXX. Its response is learned by **innate values** and finally it cannot tell the names due to no **labels**.
- TCAD2017Sadi's work can only classified into 2~25 "feature IDs" with considerable mismatches.



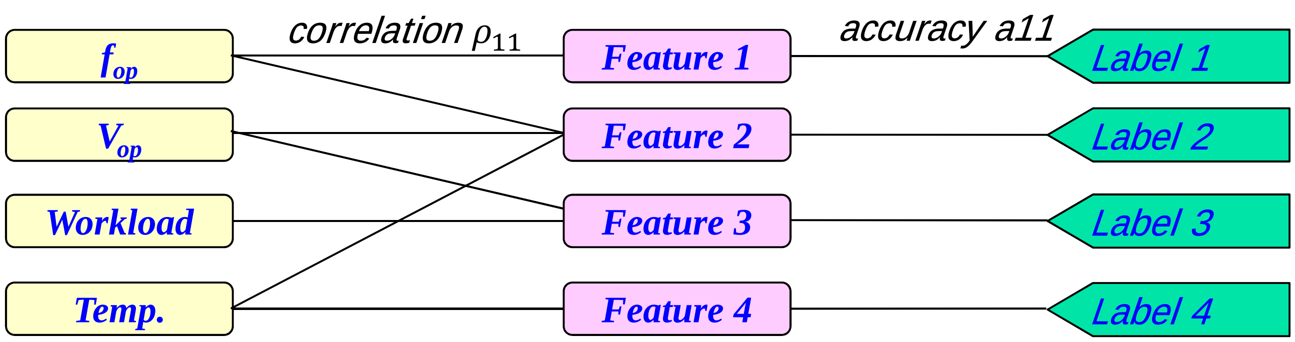
Previous Work on NNs for PM

Year	Work	Optimize (G) / Constraint (C)				Optimization Knobs			Architecture			Machine Learning Technique		
		Perf	***Power***	***Energy***	***Temp***	Task Alloc	DPM	DVFS	Single Core	Homogeneous Multicore	Heterogeneous Multicore	Supervised Learning	Unsupervised Learning	Reinforcement Learning
2016	[48]	C		G			✓			✓	✓			TD(A)-learning
2015	[15]	G	C					✓		✓				Q-learning
2015	[14]				G	✓				✓				Q-learning
2015	[16]	G	C		G/C	✓		V		✓	✓			Q-Learning
2015	[36]	C		G				✓		✓			Clustering	
2015	[24]	C		G		✓			✓					TD(A)-learning
2015	[18]	C		G/C				V		✓		Rigid linear regression		
2015	[37]	C		G/C				✓		✓				Q-learning
2014	[38]	C			G	V		V		✓				Q-Leaming
2014	[9]	C		G	C	V	✓			✓		Neural Network		Q-learning
2014	[25]	C		G			✓		✓					TD(A)-learning
2013	[26]	C		G	C		V	V	✓					Q-learning
2013	[6]	C	G				✓		✓			Bayes classifier		TD(1)-learning
2013 2011	[39] [40]	C		G	C			V		✓		Least squares regression		
2012	[7]	G		C			✓		✓					Q-learning
2012	[41]				G	.(				V		Genetic algorithm	k-means clustering	
2012	[27]	G/C		G/C	G/C			V	✓					Q-leaming
2011	[28]	G/C	G/C				✓		✓			Bayes classifier		TD(A)-learning
2011	[29]	G			C			V	V				k-means clustering	Q-learning
2011	[30]	C		G					✓	✓	Least squares	regression		
2011	[42]	G/C			G/C	V		V		✓				ad hoc
2010	[43]	G			C		V	V		✓		Least squares regression	k-means clustering	
2010	[8]	C		G				✓		✓		Bayes classifier		
2010	[44]			C	G/C			V		✓	Least squares	regression		
2010	[45]	C/G				V		V		V				Observe-decide act
2010	[31]	C	G					V	✓			Least mean square linear predictor		
2009	[32]	C	G				V		✓					Q-learning
2009	[33]	G		G			✓	V	✓					ad hoc
2008	[46]				G/C			V		✓		LWPR		
2008	[47]	G			G/C	✓	✓	V		✓				ad hoc
2005	[21]	C		G				V	✓		Least squares	regression		
2002	[34]	G/C	G				V		V					Markov Decision
1999	[35]			G			.(		V			Adaptive learning tree		

S. Pagani, P. D. S. Manoj, A. Jantsch and J. Henkel, "Machine Learning for Power, Energy, and Thermal Management on Multicore Processors: A Survey," Trans. CAD of IC & Sys. 39(1): 101-116, Jan. 2020.

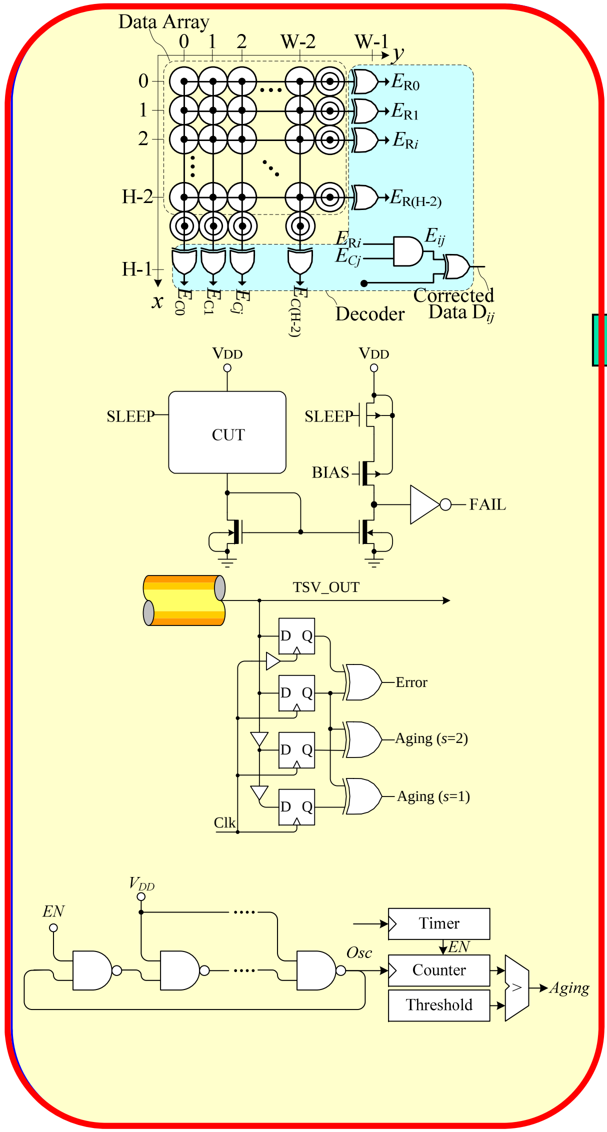
Previous Work on NN Data Annotation

- **10 best data annotation companies in web lionbridge.ai**
  - Data annotation is not only an issue but a significant and commercial task.
  - Usually processed automatically off-line also by AI methods
  - Finally decided by at least 2 experts in the field.
  - No literature works on data annotation related to early warning of IC aging.
- **Relation Graph of Correlations:**

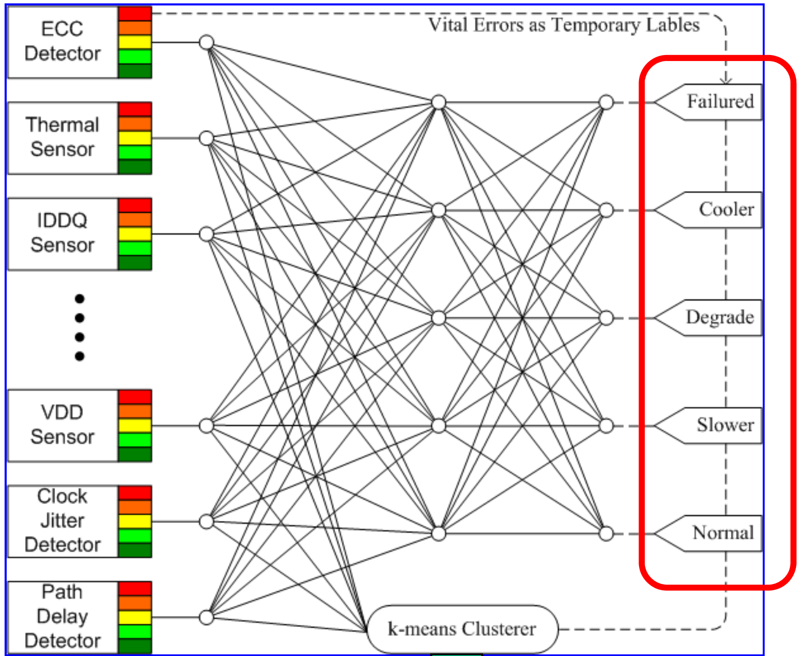


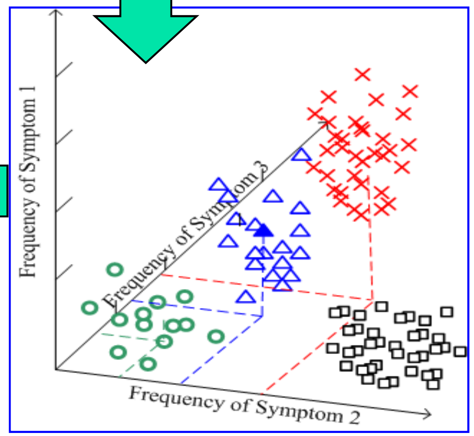
- **Supervised:** $p(\rho|Label) \cdot quality(Label) = p(\rho)$
- **Unsupervised but with single label (scalar thermal or slack with implied direction):** *p(single label) = 100%*

Subproject-4: Big-Data Driven Online Testing, Reconfiguration, and Reliability Enhancement for AI Hardware Accelerators



Statistical Sampling + Sensor Allocation





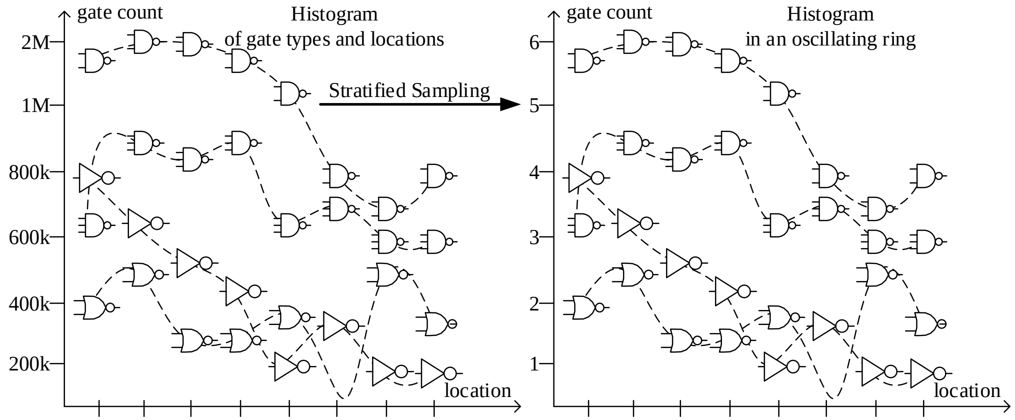
Contingency Policy

Cooling,
Degrading,
Slower,
Substituting, *etc*.

Stratified Sampling

Stratified Sampling

- **Concept: Examples with respective to both gate types and locations:**



- **High correlation to the average effect of the whole circuit under test**
- **Personalities of critical paths, TSVs, hot-spot are lost ➔ should be kept**

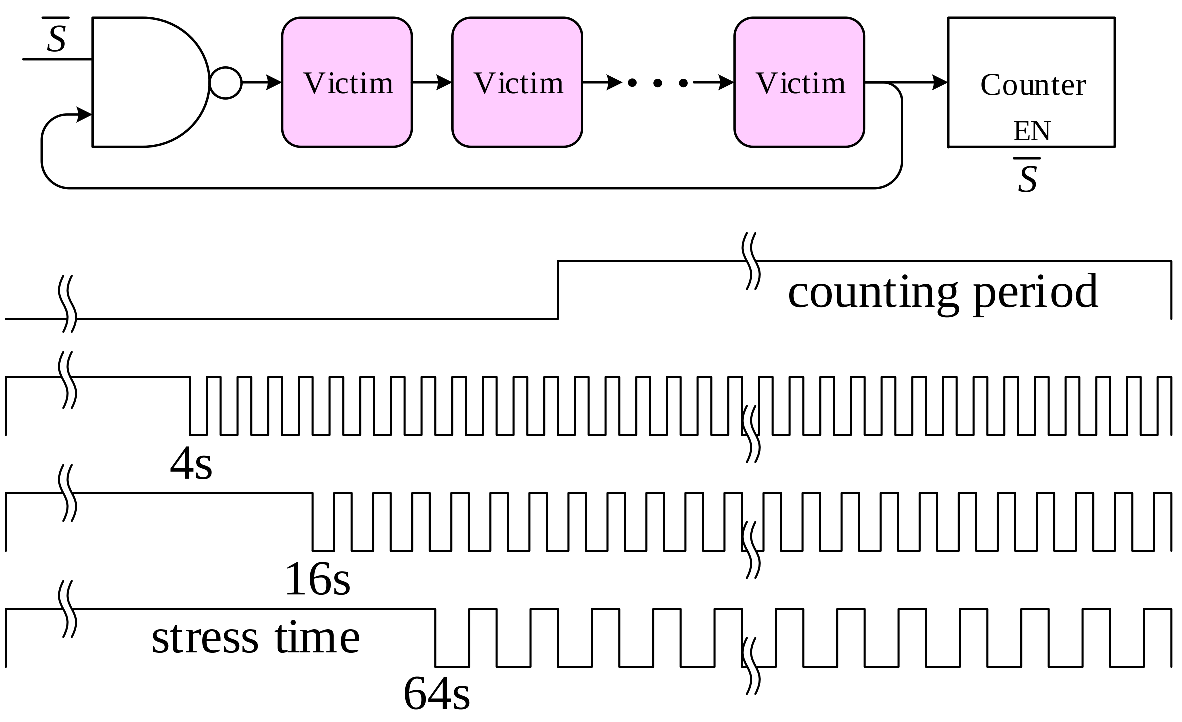
High-Risk Sampling

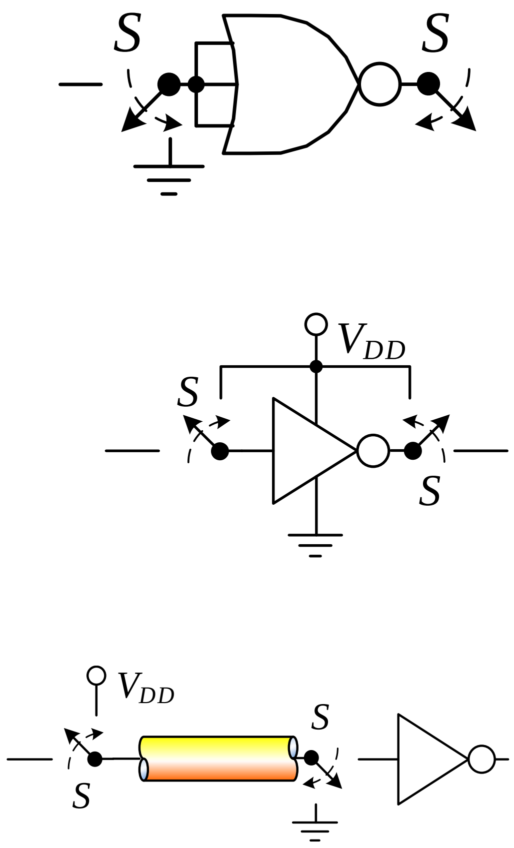
- **High risk in normal operations that cannot be replaced and stressed, but with high correlation to specific SAFER Indicators**
- **Hard to reflect the portion of the whole circuits under test**
- **Examples:**

Factors	High-Risk Patients
NBTI	Idle SRAM, FFs, Latches
HCI	Delay-lines, Flip-flops
Fatigue	Hot spots, TSVs, Contacts

SAFER Array

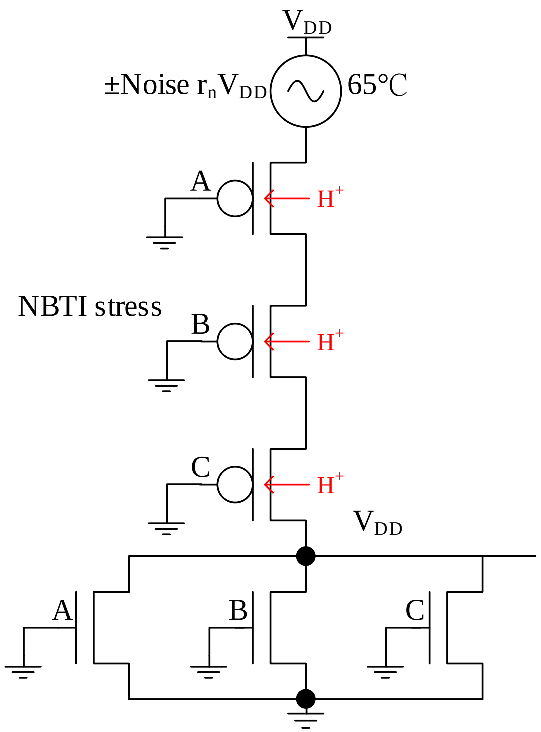
Single Aging-Factor Enhanced Rings

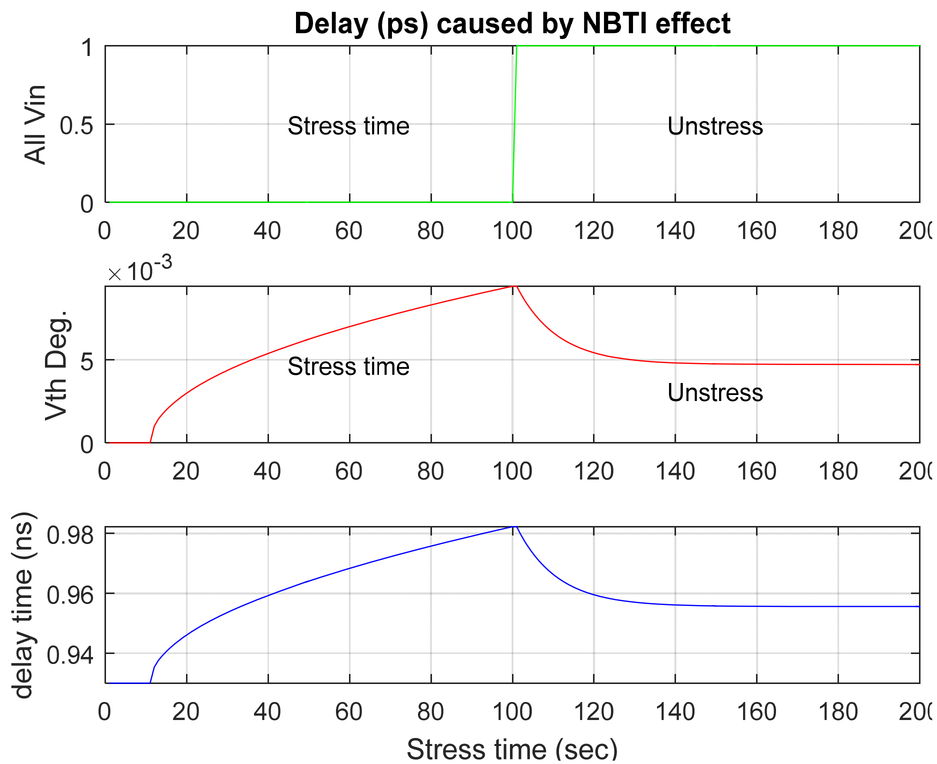




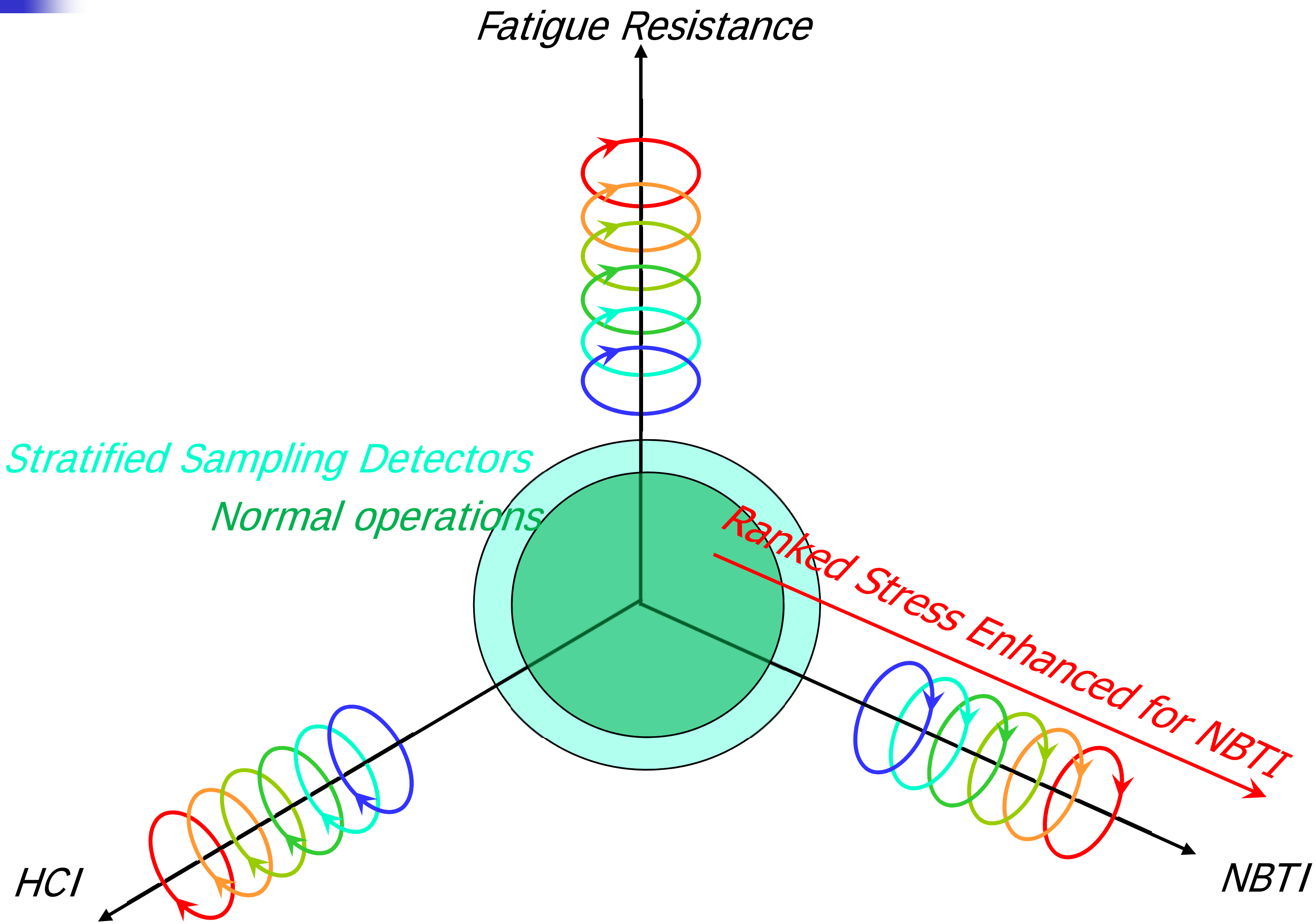
NBTI Enhanced Cell for Aging Acceleration

■ Mapping NBTI effect to circuit delay

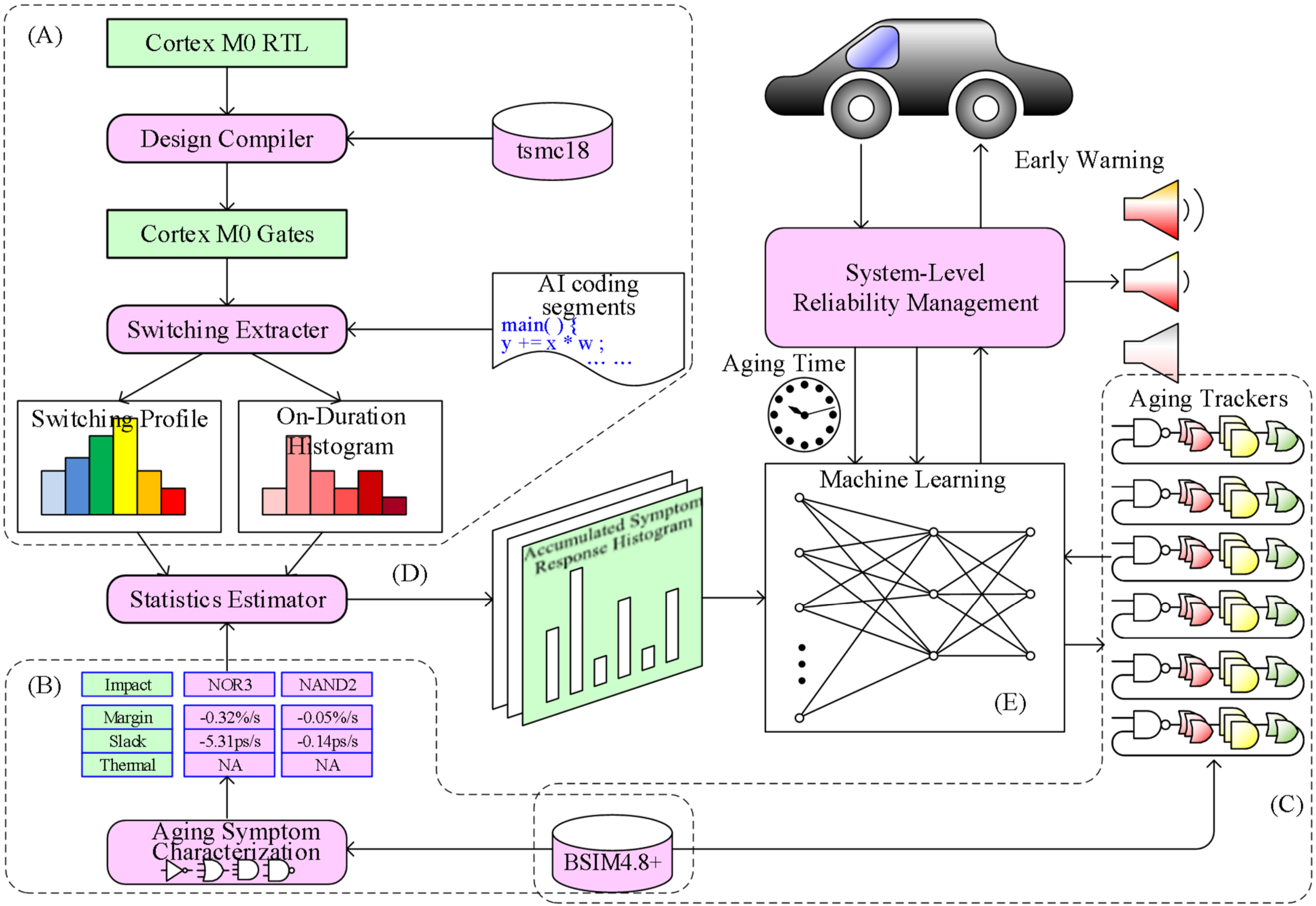




Basic Concept of SAFERs

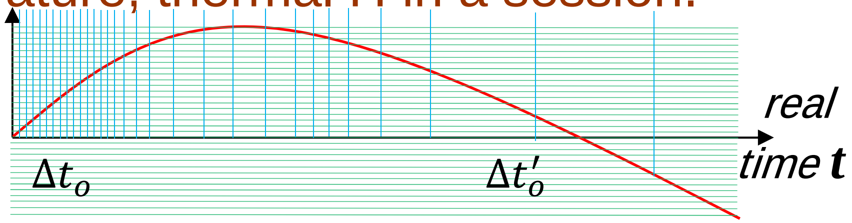


Proposed Online Aging Monitor

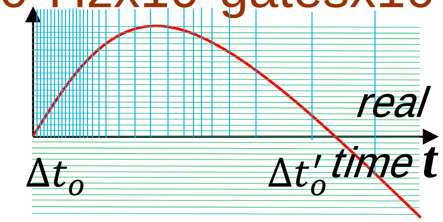


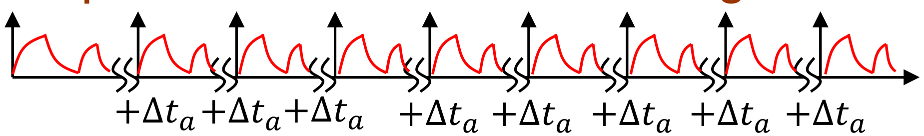
General Simulation Acceleration

- Fresh Simulation (*eg*. Conventional HSPICE)
  - Cannot change $agetime, $Temperature, thermal H in a session.
  - Δt can be accelerated to $\Delta t'$ under precision control.



- Aging Simulation in Real Time
  - Dynamic Range (Precision) Issue.
  - Extremely Terrible Time Complexity, *eg*. 10^9Hzx10^7gatesx10^6sec !
- Aging Simulation in Aging Time
  1. Aging acceleration: with an AgingTime Scaling Factor (ATSF)$\gg 500$
  2. Realtime-Sampling: Aging time space is split from real time with a large unit *ua* (s or h).





- Statistics-Base Simulations
  1. Probabilistic Simulation: without memory or history
  2. Stochastic Simulation: random process from distributions in previous state to distributions in next state.

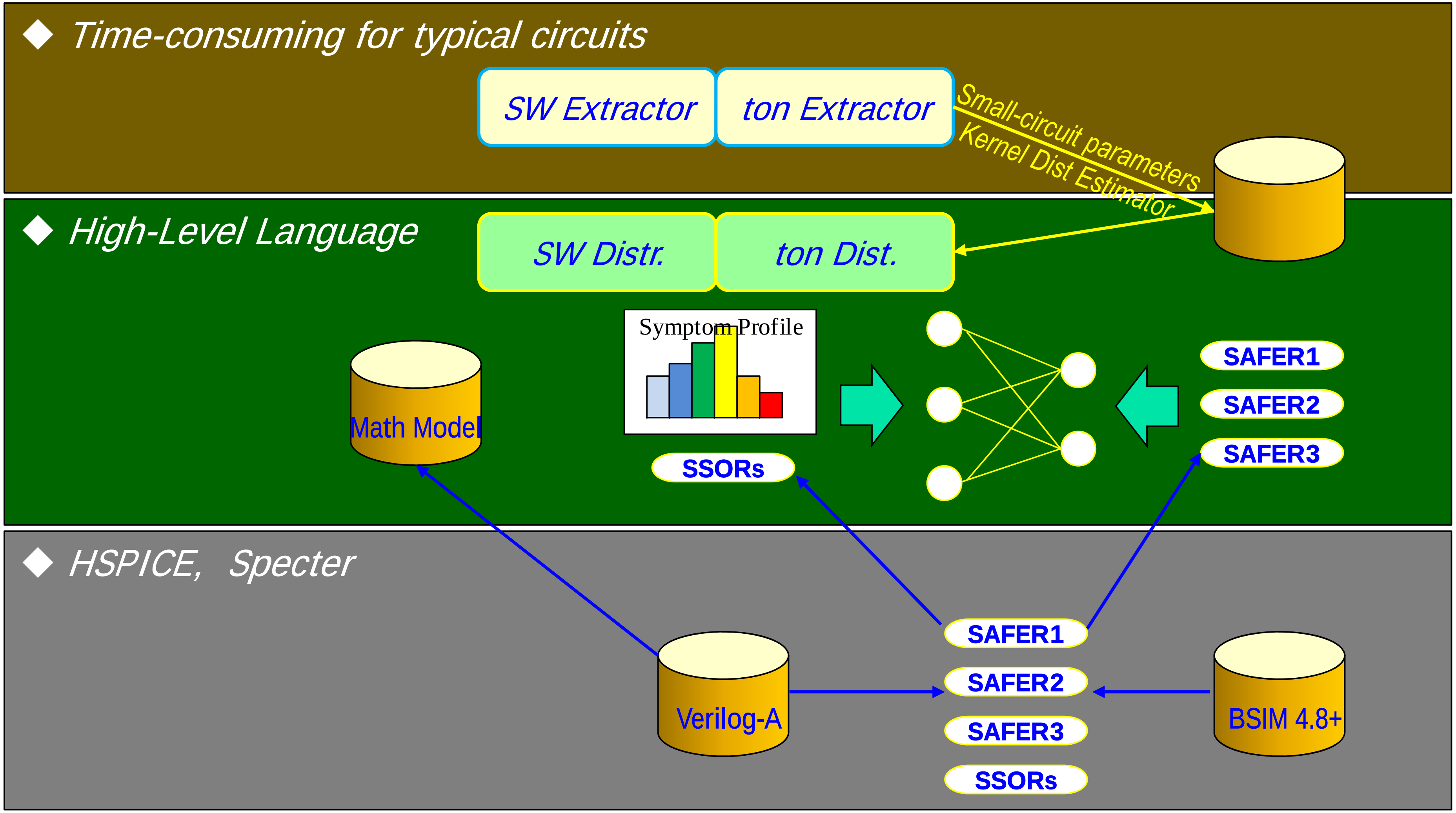
Introduction to Verilog-A

- Quick tutorial for HSPICE
  - ΔV_{th} sim
  - IR drop
  - IV thermal
  - Pattern-indep. stochastic sim can be done in hi-level languages, eg. python
  - So far it's impossible to sim pat-dep. M-gate G-Hz circuits with any effect.
  - Symptom profiles are usually assumed to be similar to that of a small circuit or during a short run.

```
`lib 'tsmc45.l' TT
`include "disciplines.vams“

.Model Diode D(BV=6.3)

module DVA(A,C);
    electrical A,C;
    branch (A,C) AC;
    parameter real is=1e-14 from [1e-30:inf);
    parameter real n=1 from [0:10];
    real vd, id;
    analog begin
        vd = V(AC);
        id = is * (limexp( vd / (n * $vt)) - 1);
        I(AC) <+ id; // accumulated
        // demo for aging simulation
        t = $temperature;
        agingtime = $abstime*atsf;
        ais = f(agingtime, atsf, …);
    end
endmodule

D1      1       2       Diode
X2      1       2       DVA
V1      1       0       PULSE(0 1 0 0 0 1n 2n)
R2      2       0       1K
.TRAN   10p     10n
.END
```

Difficulties and Considerations on Verification

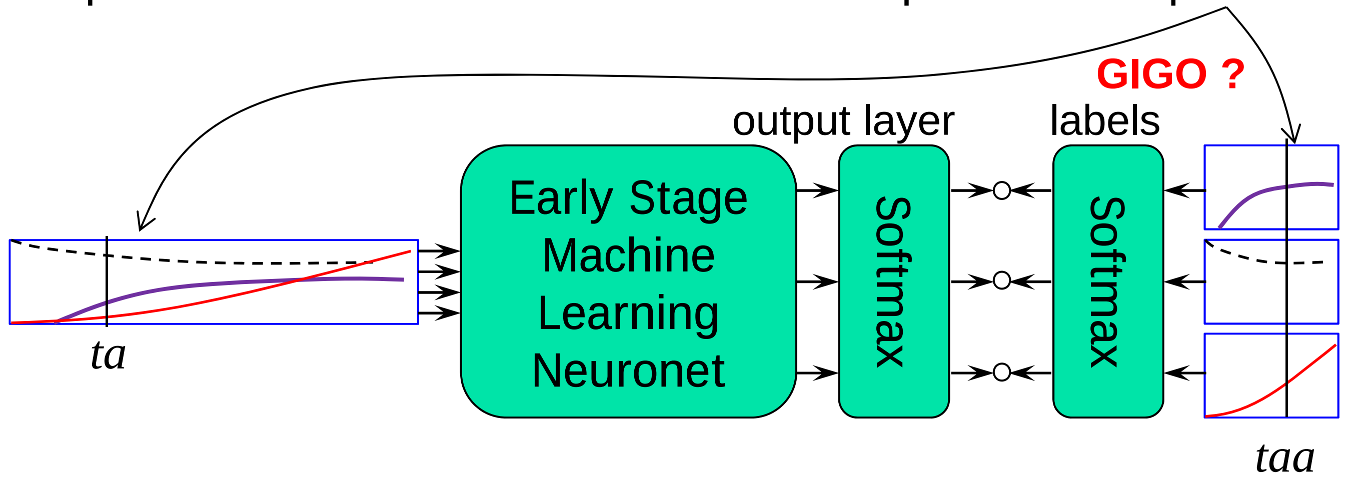
Two-Level Simulations



Stochastic Simulation

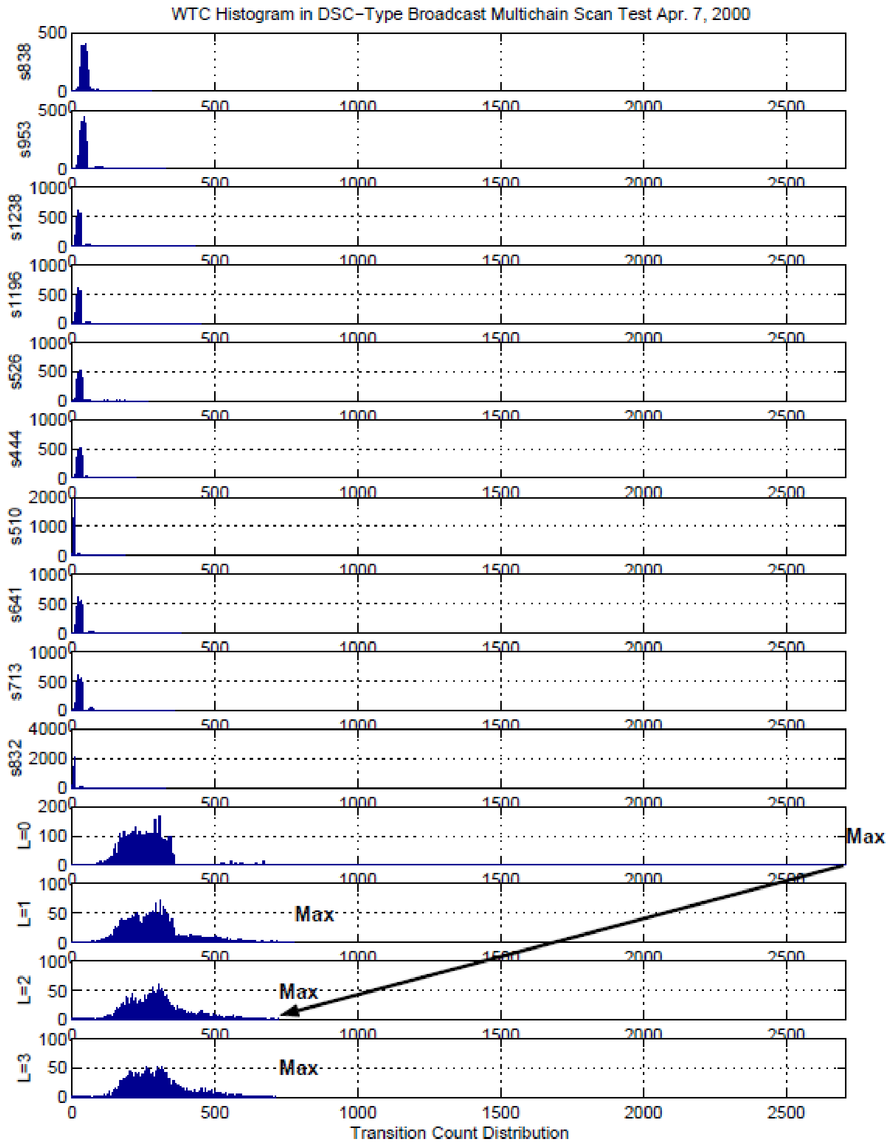
Simulation Flow

- Profile Extraction from Small Circuits in a Short Duration
  - Switching Activity (SA)
  - Turn-on Duration (t_{on})
- Symptom Mapping from Realistic Profiles ➔ High Complexity
  - V_{th} ➔ Delay (Slack loss)
  - In Adiabatic Models ➔ Temperature Variance
  - SA ➔ Voltage-Drop & Ground Bounce
- Kernel Distribution Estimation (KDE, *python/seaborn*) from Profiles
  - Estimate the distributions
  - Adjust statistic parameters $(\overline{X}, S)_k$
  - Extrapolation from stochastic distribution to probabilistic profiles



Profile Extraction from Small Circuits

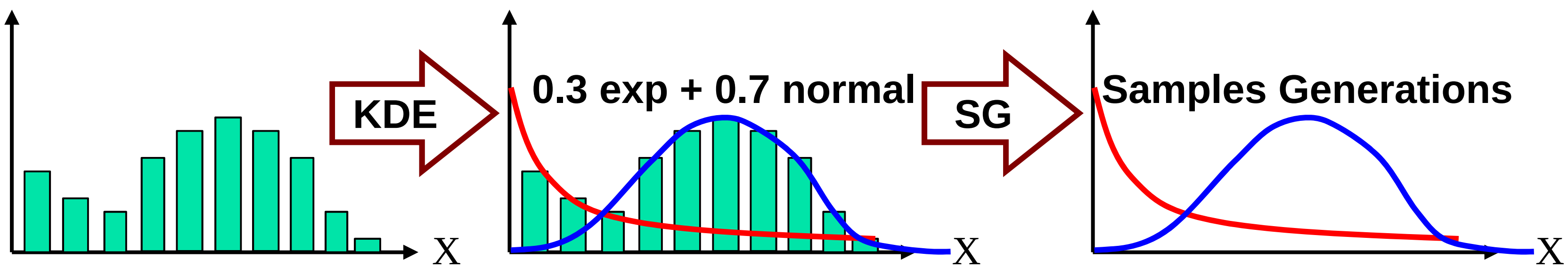
- **Profile Extractor**
  - Revised from our peak power estimator / Verilog parser
- **Extractable Parameters**
  - Peak Power
  - Average Power
  - IDDQ
  - Switching Activity (SA)
  - C0/C1 Values
  - Turn-on Duration (ton, or top)
- **Categories**
  - Whole CUT
  - Specific with Uniqueness
    - ✓ Delay line composed NOT gates
    - ✓ Register Files (Top & Bottom)
    - ✓ SRAM Cells
    - ✓ PLL, CG, Many-input gates



Kernel Distribution Estimation

Stochastic KDE Probabilistic Profiles

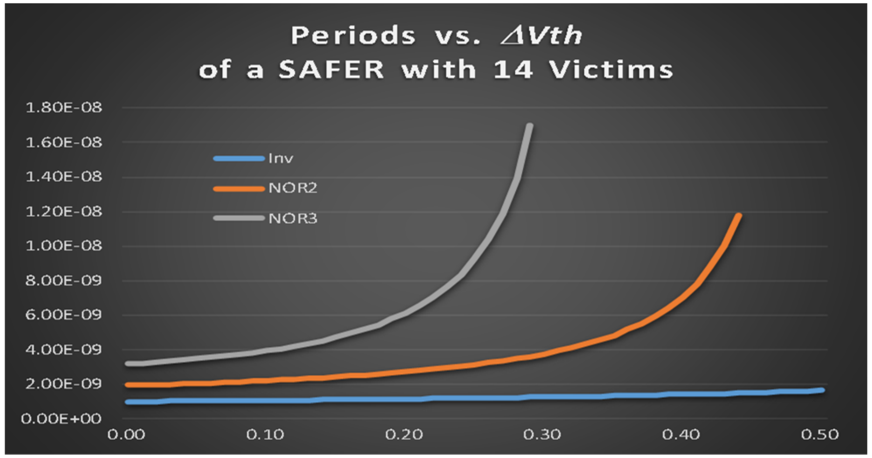
- Profile Extraction from Small Circuits in a Short Duration
  - ✓ Switching Activity (SA)
  - ✓ Turn-on Duration (t_{on}) (Vdd (±5%) & T (60-80℃) are set to uniform, so far)
- Symptom Mapping from Realistic Profiles ➔ High Complexity
  - ✓ V_{th} ➔ Delay (Slack loss)
  - ✓ In Adiabatic Models ➔ Temperature Variance
  - ✓ SA ➔ Voltage-Drop & Ground Bounce
- Kernel Distribution Estimation (KDE, *python/seaborn*) from Profiles
  - ✓ Estimate the distributions
  - ✓ Adjust statistic parameters $(\bar{X}, S)_k$
  - ✓ Extrapolation from stochastic distribution to probabilistic profiles



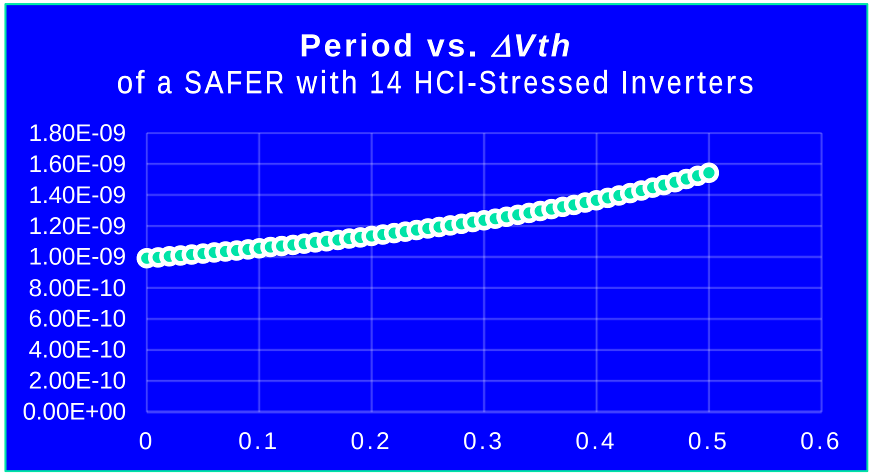
Experimental Results on Vth Degradation

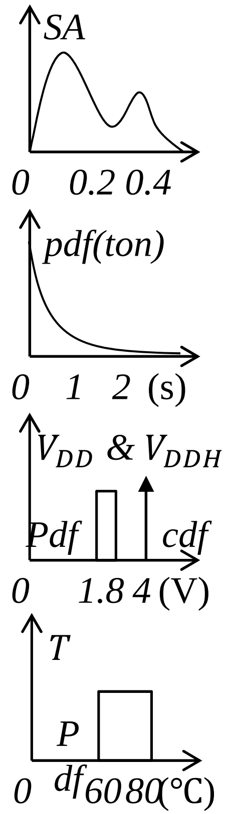
HSPICE Simulations

- TSMC18 Model transistors for other circuits
- Verilog-A Model for Victim transistors of the Surrogate Cells with n~0.5
- NBTI:



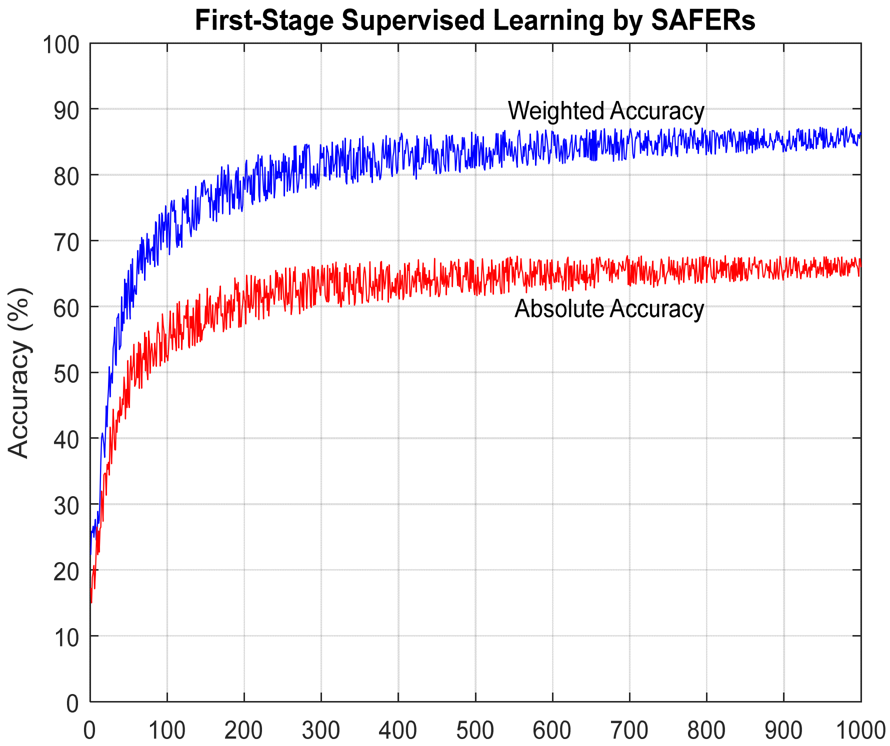
- HCI:





Simulation on Supervised Learning

- 3 symptoms (SW (instead of T), Delay, IR drop) x 3 SAFER Arrays
- Low absolute accuracy because the symptoms and response of SAFERs are still the combinations of multiple factors.
- If the Bayesian analyses are applied to calculate the weighted accuracy, the weighted accuracy can be pulled up to 84%
- Actually, the safety is better than 100% classifier without any correlation.



Comparison with Previous Work

Comparison on the aging tracker design

Comparison	ASPDAC2017Pagani	Our SAFER Array
On/Off Line	Off-line	On-line
Trigger	System Reboot	Next Guardband Interval
Aging Tracking	V	V
AI Network	None	Neural Network
Data Annotation	None	1st Stage
Intensity Adapting	None	2nd Stage
Early Warning	Too Conservative	V
Dimensions	m times x n samples	f factors x n intensities

Estimation of Cost

Device	Sensor	Count	#gates	Estimated Area (μm^2)
Import	System Timer	1	0	0
Integrated Detectors	Path delay	7	392	3,360
	IDDQ	0	0	0
	Thermal	1	1 (x10)	85
	Clk TSV	0	0	0
	Data TSV	1	44V+672G	9,531
Isolated	SSORs	4	52	445
Isolated SAFERs	NBTI	12	156	1,337
	HCI	4	52	445
	Fatigue	4	4V+8G	411
Total		34	48V+1352G	8.5% overhead

Conclusions & Future Work

Novelties & Contributions

1. **We propose a SAFER array suitable for**
   - ✓ Data Annotation to symptoms with high correlation
   - ✓ Classified and annotated symptoms taken to select proper SAFERs for early warning
2. **Two-Stage co-learning (self-annotation & self-selection) strategy**
   - ✓ Reasonable accuracy ($\gg 1/\#L$)

Future Work

1. **Medium-sized circuits aging profile extractors**
   - ✓ Making the KDE more trustable by Fmax tests.
   - ✓ Extracting more realistic profiles
2. **Developing more accuracy NN model including Bayesian analysis**
   - ✓ Improving the learning accuracy
   - ✓ Study the overlapped spectrum (Syndrome) to reason the inaccuracy
   - ✓ Multiple monitors for unsupervised intensity learning
3. **Guardband reduction by error redundancy**
   - ✓ Reliable Neural Network Accelerators
   - ✓ Taking error correctable capacity for data annotation
   - ✓ Reducing slack guardband and provide longer guardband intervals