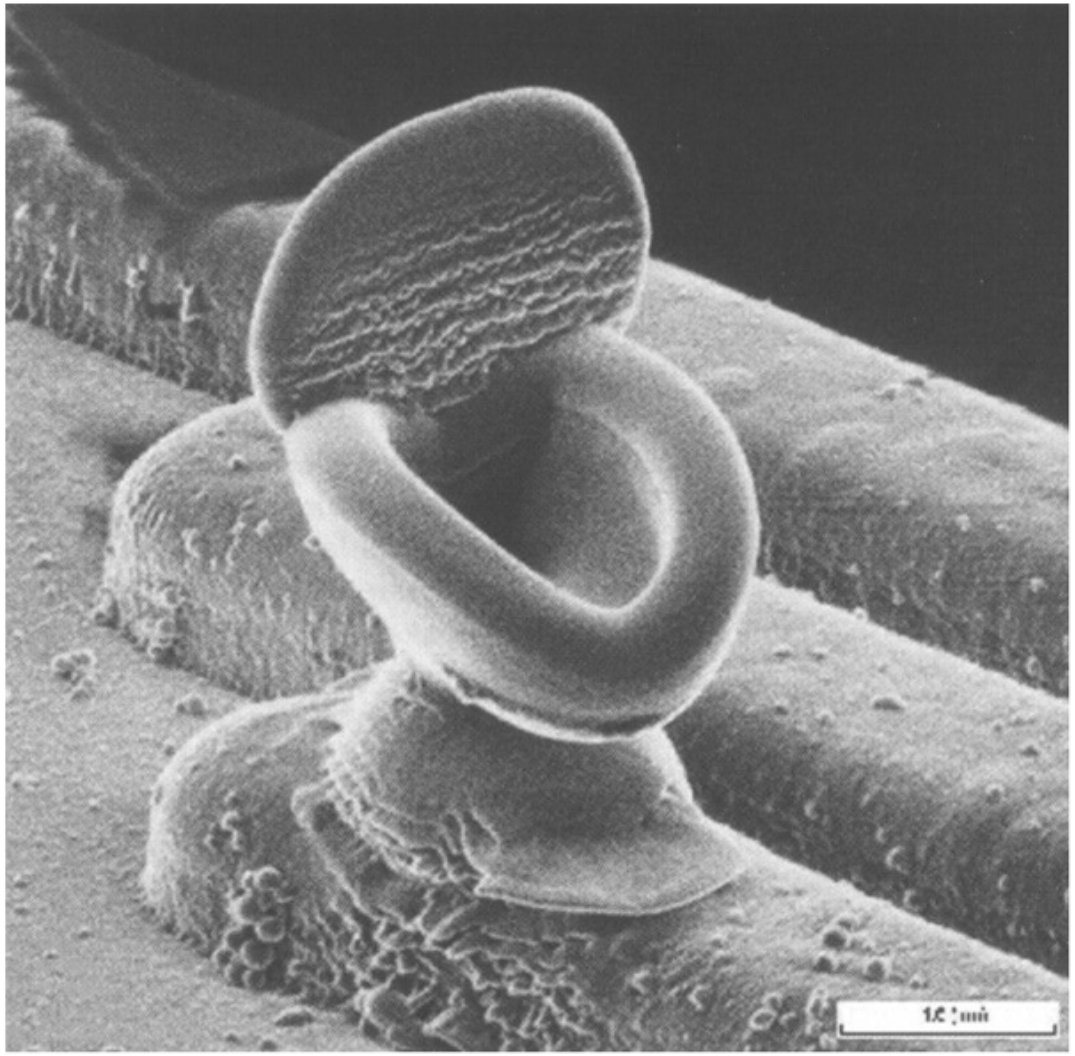
Building reliable chips in future technologies: Fact, fiction or an oxymoron?

Vikas Chandra
ARM Research
San Jose, CA

Acknowledgments

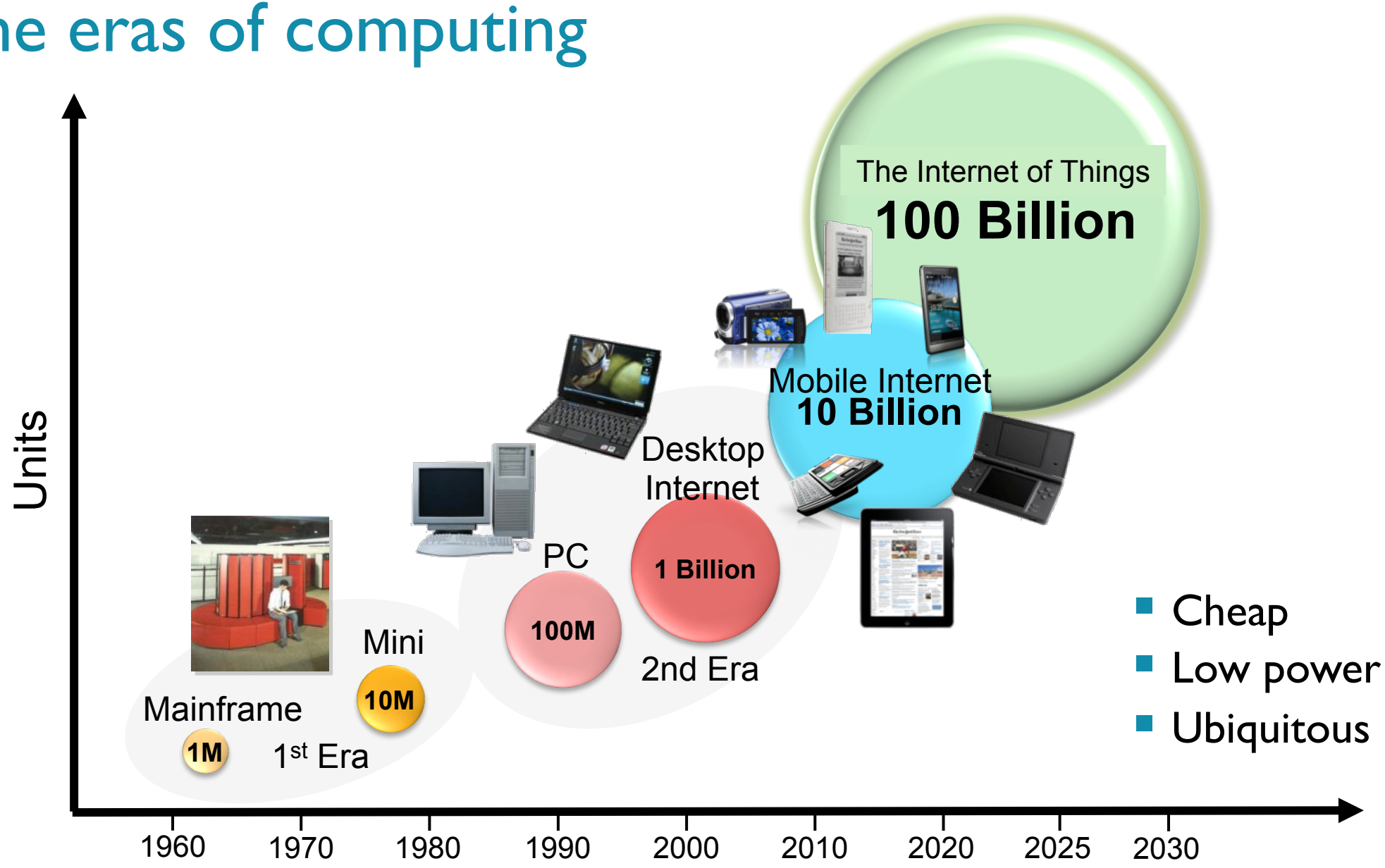
- Rob Aitken (ARM)
- Greg Yeric (ARM)
- Subhasish Mitra (Stanford)

Some things are beyond our control!



Courtesy Takahashi Kaito (SII Nanotechnology Inc.)

The eras of computing

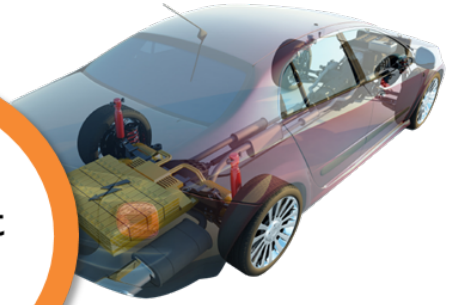


- Cheap
- Low power
- Ubiquitous

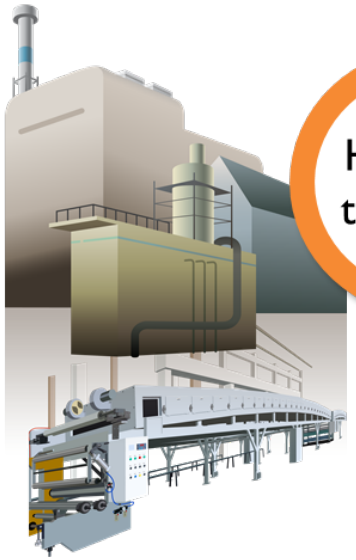
Need for reliability

High susceptibility

High impact of failure



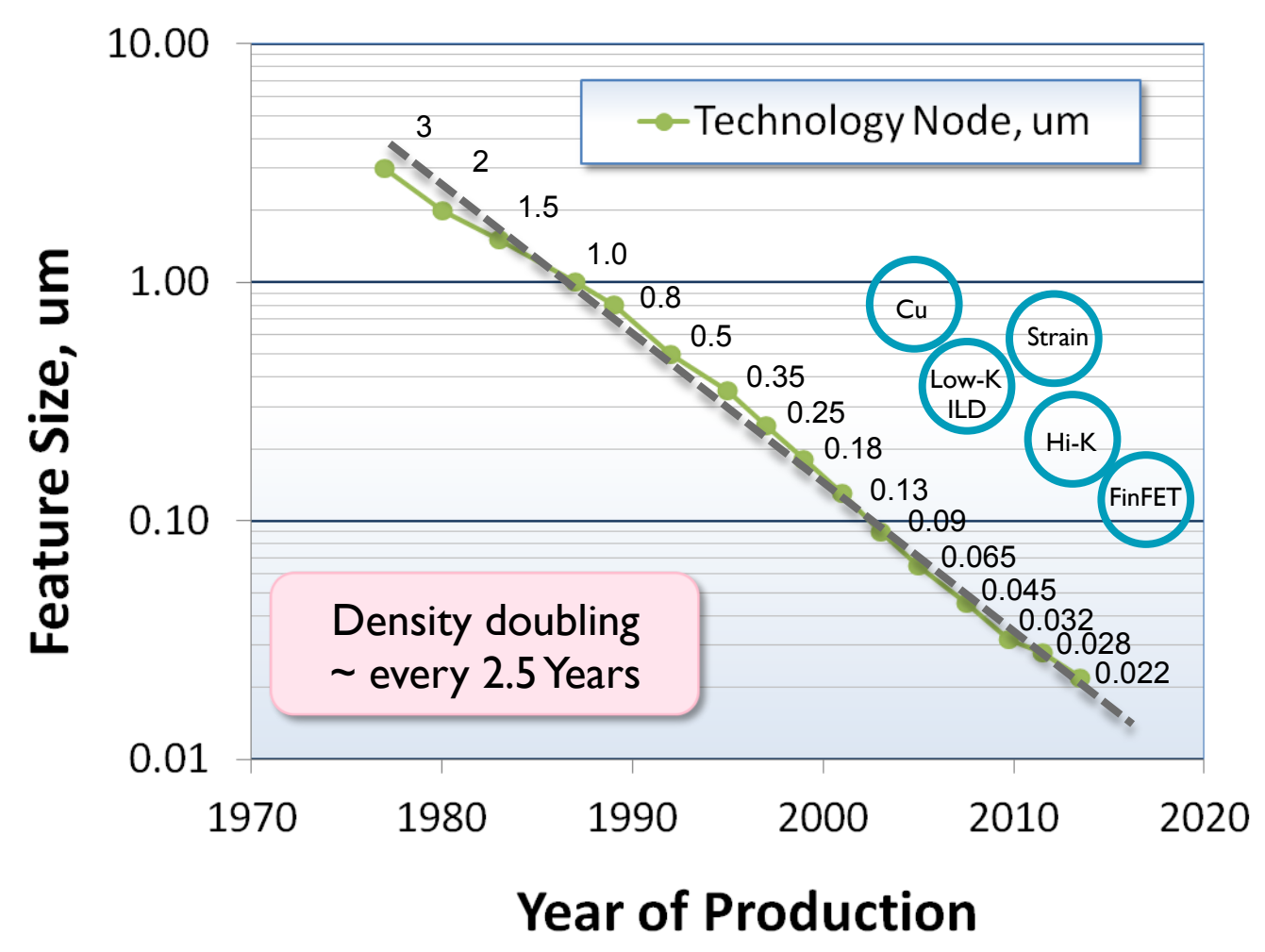
High intrinsic value



High cost to replace

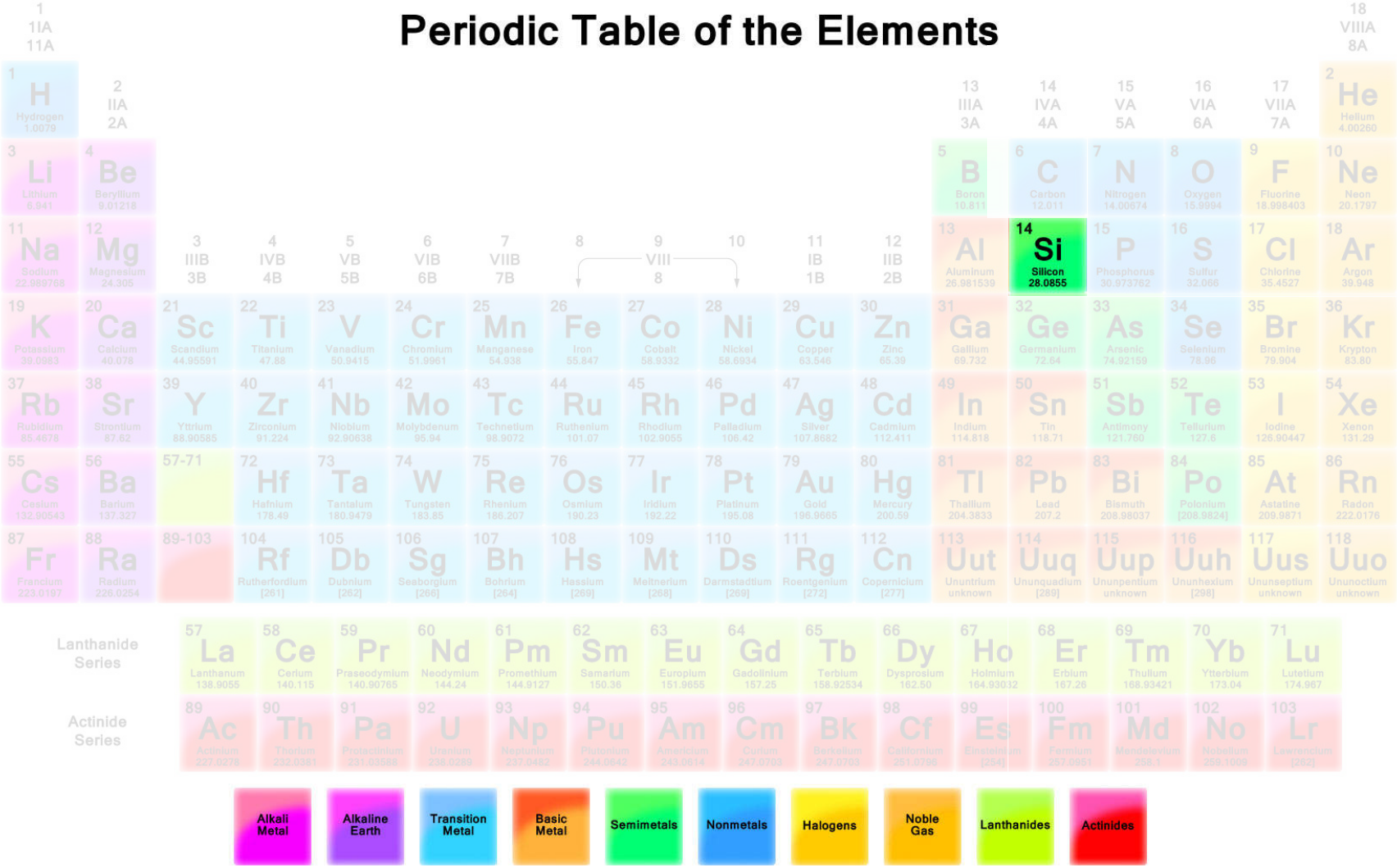
Lifetime profiles

Feature size scaling

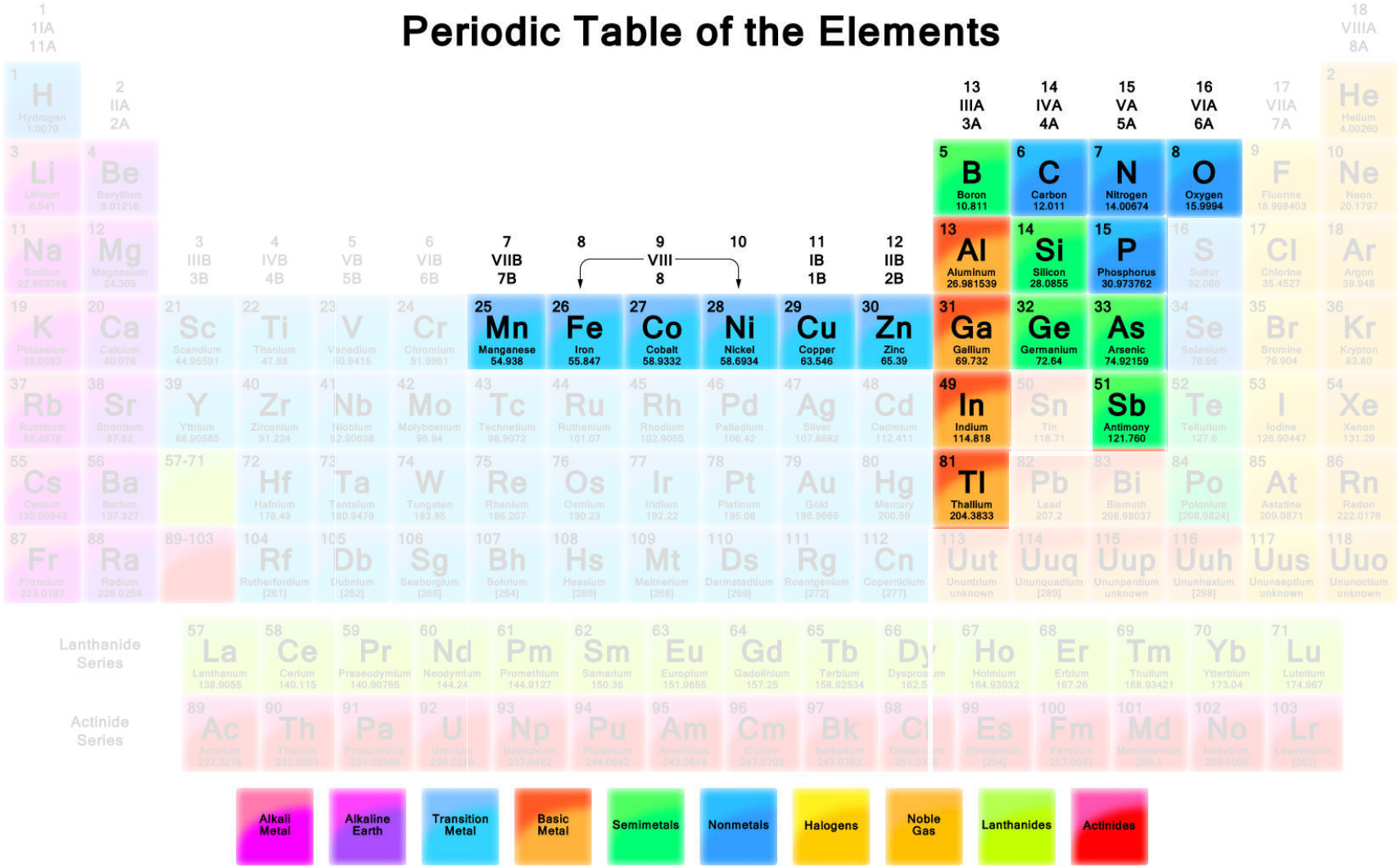


- Impact of feature size scaling
  - Charge discretization
  - Random manufacturing defects
  - Increasing electric field
  - Thin gate oxides
  - Interface defects at Si/SiO_N interface
  - Metal defects
  - Susceptibility to soft errors
- Beyond 14nm, scaling is more reliant on new materials
  - Will require more reliability analysis

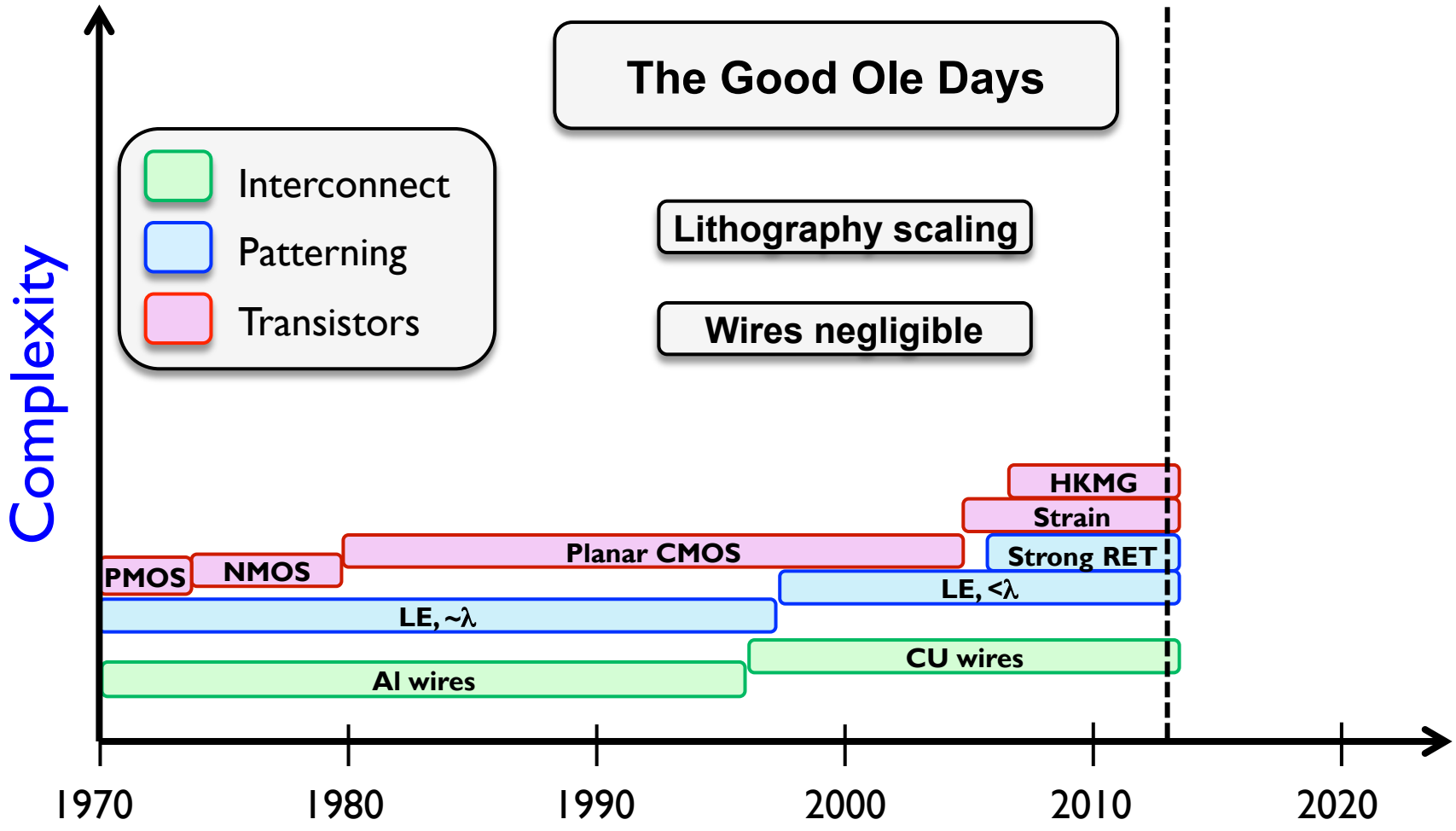
CMOS Switches: 19xx - 2015



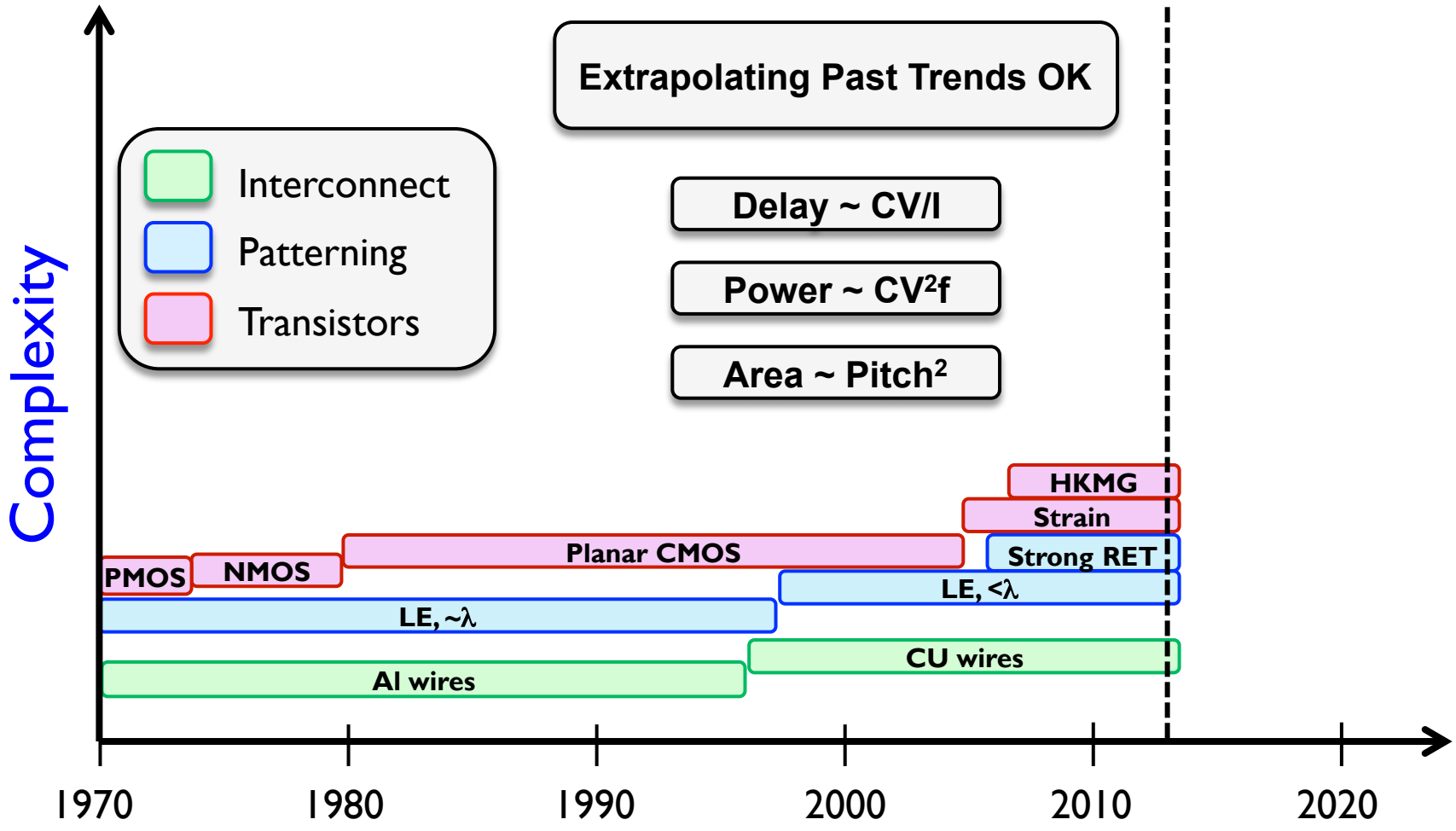
Devices 2015 – 2025



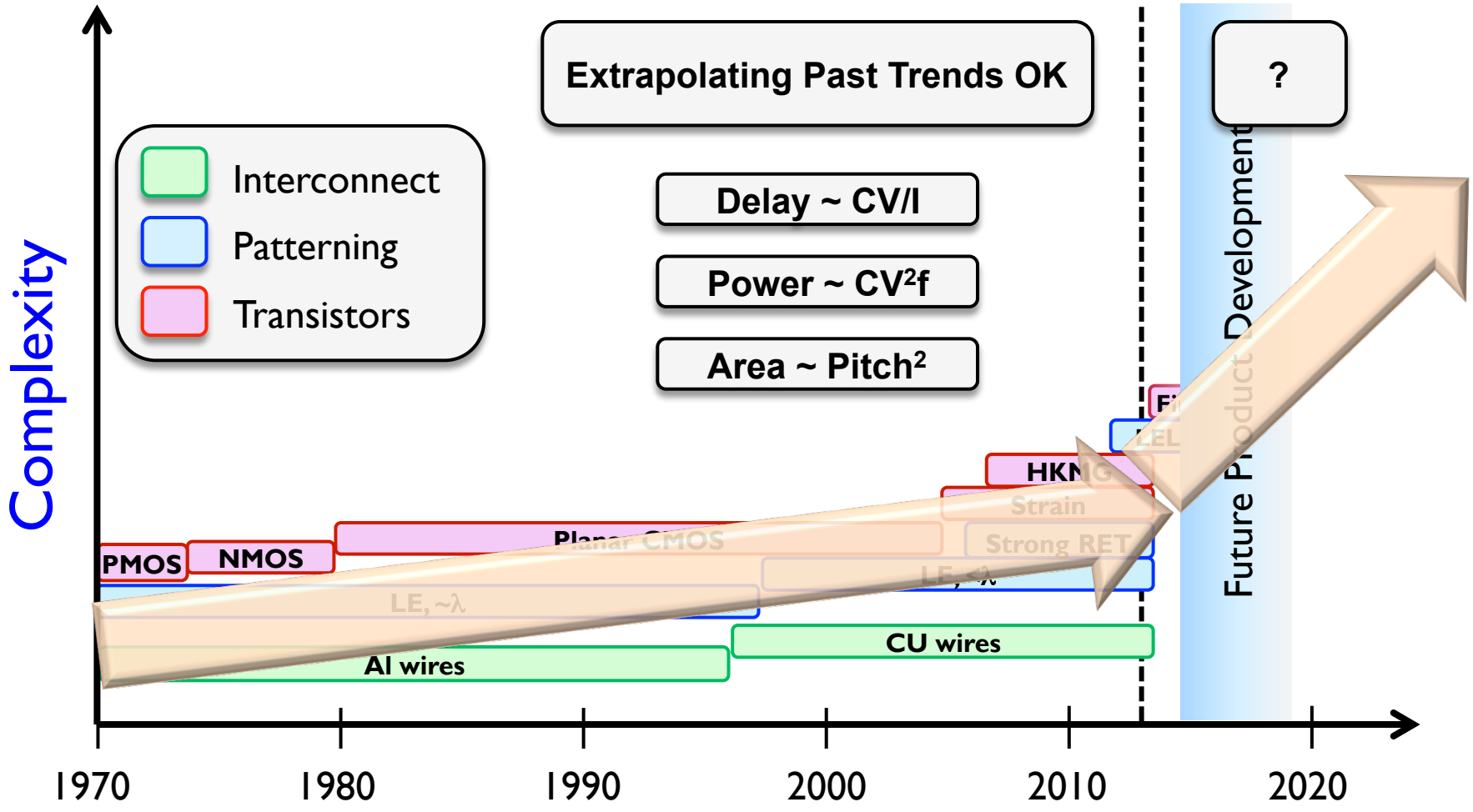
Technology scaling overview



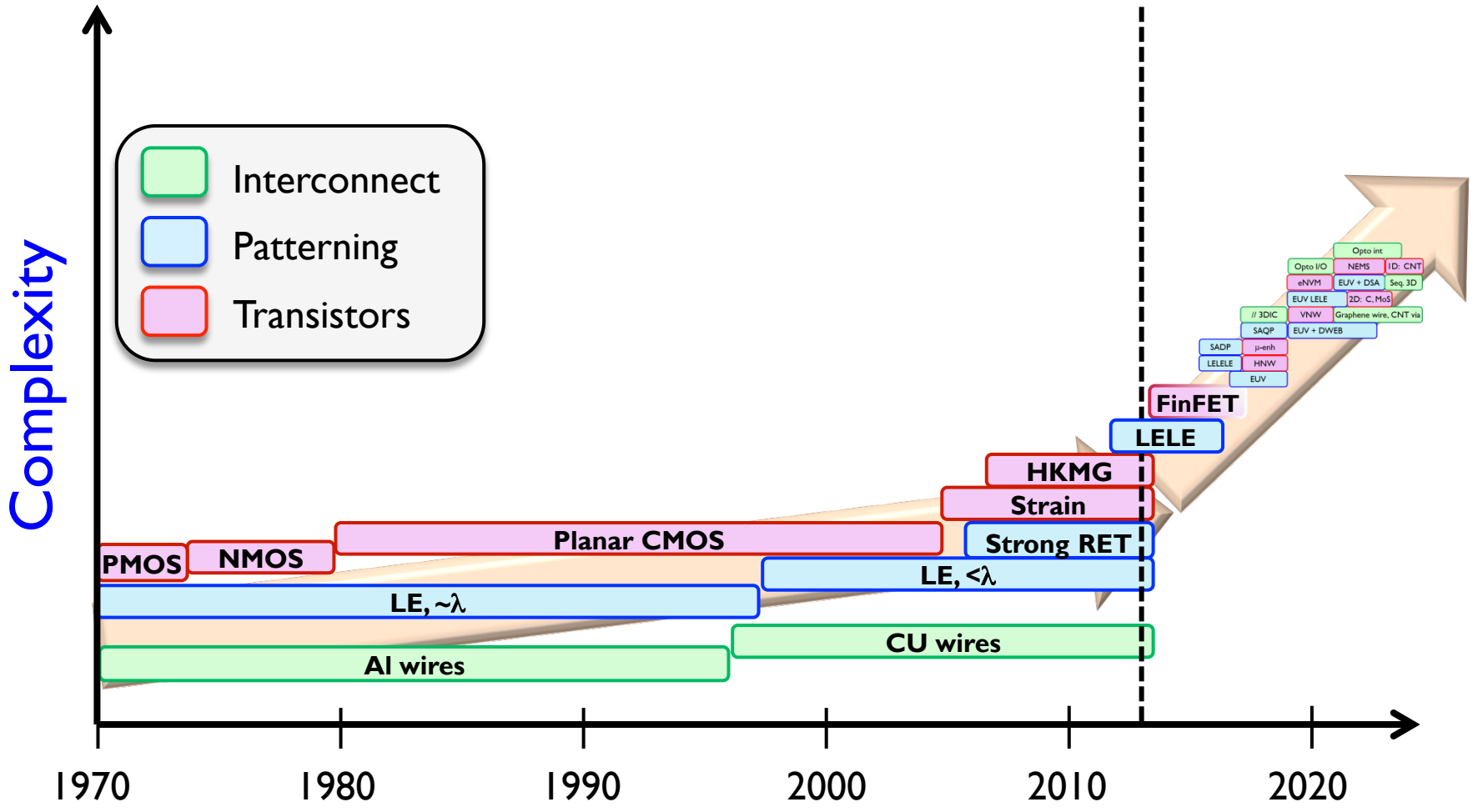
Technology scaling overview



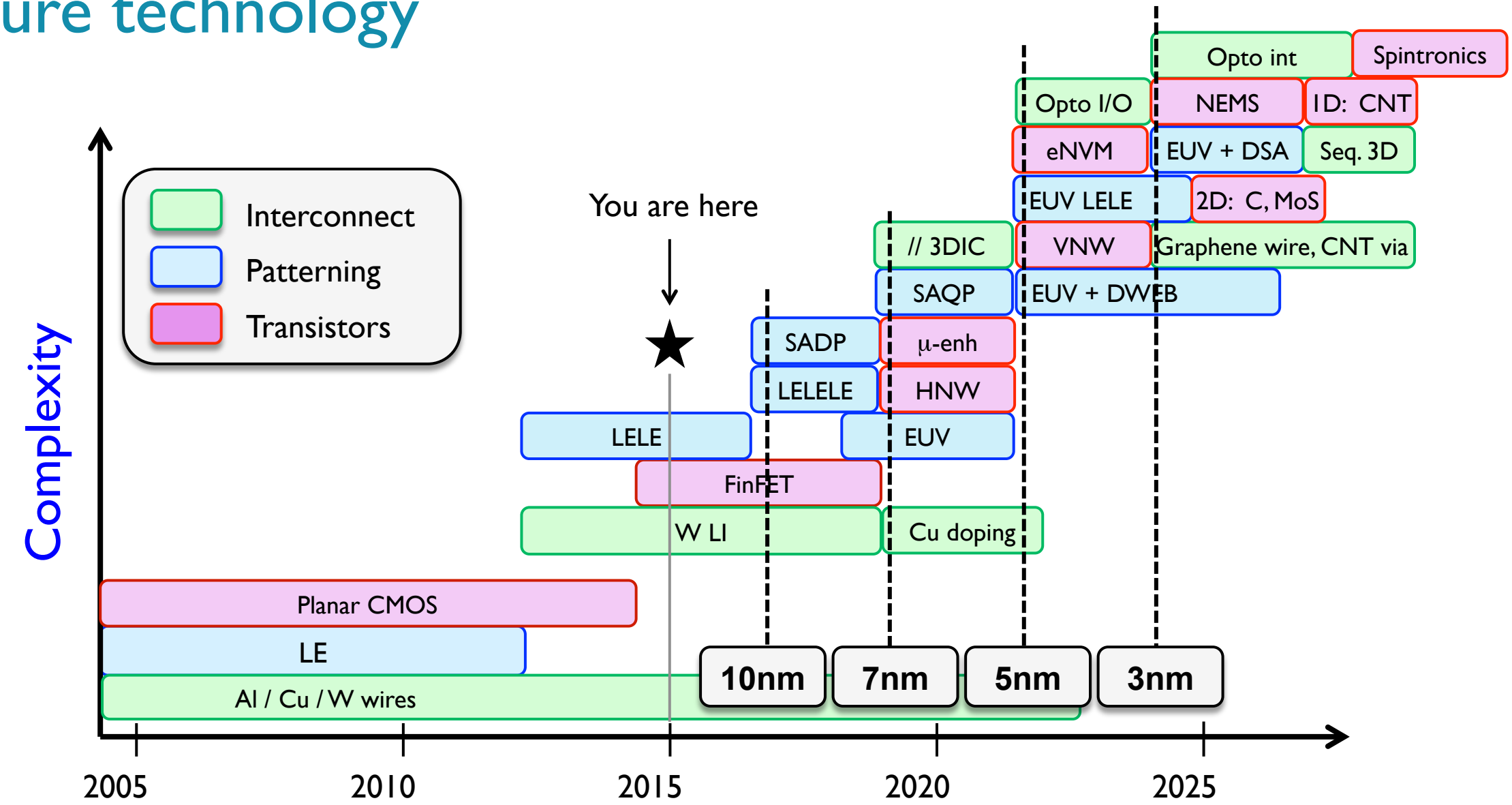
Technology complexity inflection point?



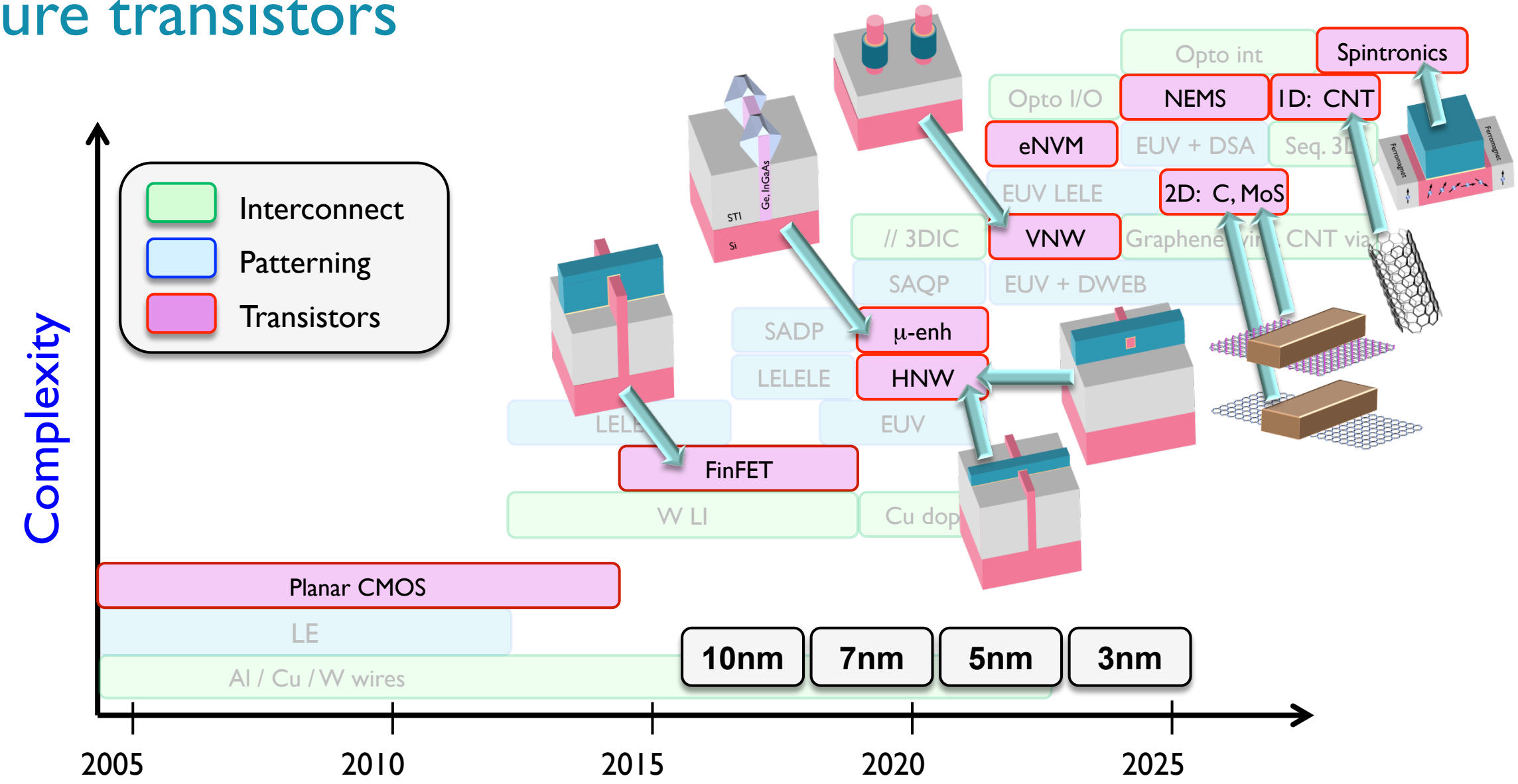
Technology complexity inflection point?



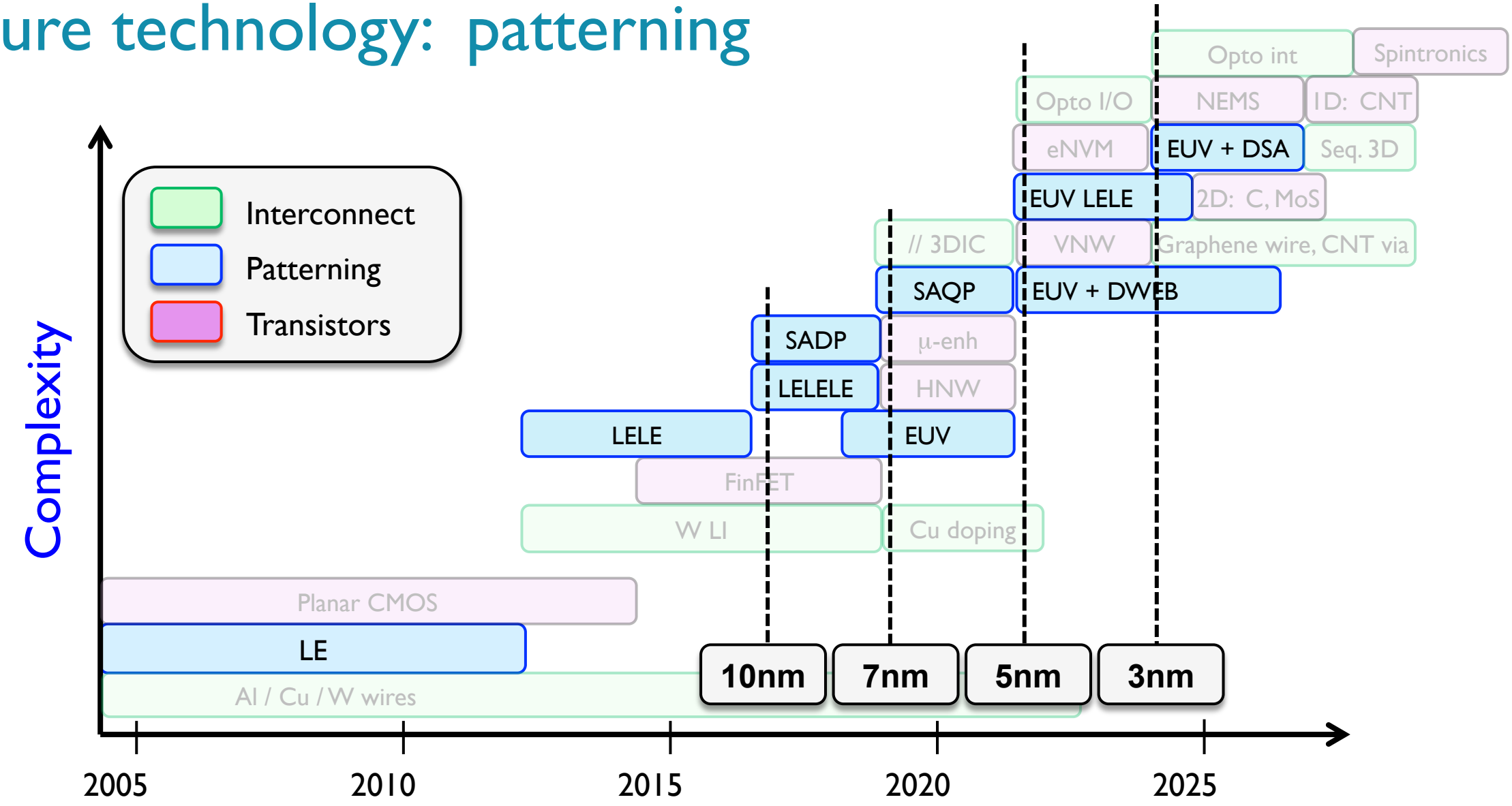
Future technology



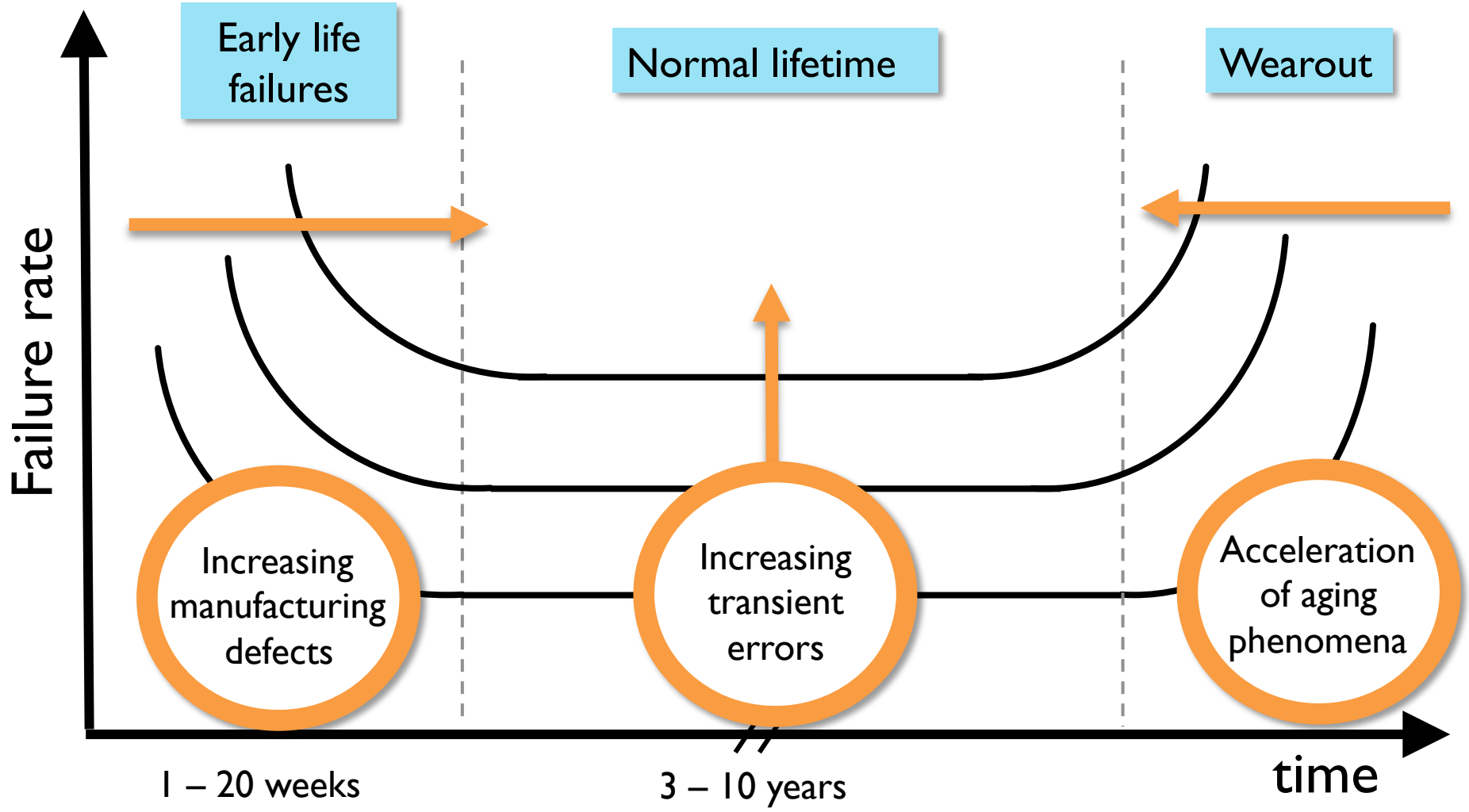
Future transistors



Future technology: patterning



Device lifetime and failure rate



Unreliable transistors – 3 phases

Early life failures (1)

Gate to source shorts
Insulator cracks
Thin oxide defects
Small opens
Poor vias/contacts

Normal lifetime (2)

Soft errors
- Memory/Flip-flops
- Combinational cells

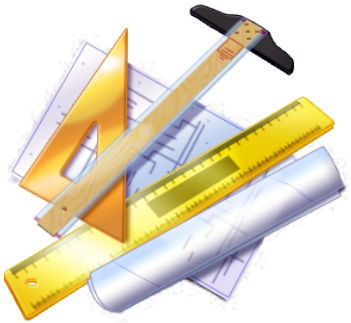
Noise
- Electrical
- RTN

Wearout/Aging (3)

Transistor wearout
- BTI, TDDB

Metal wearout
- Electromigration

ILD wearout



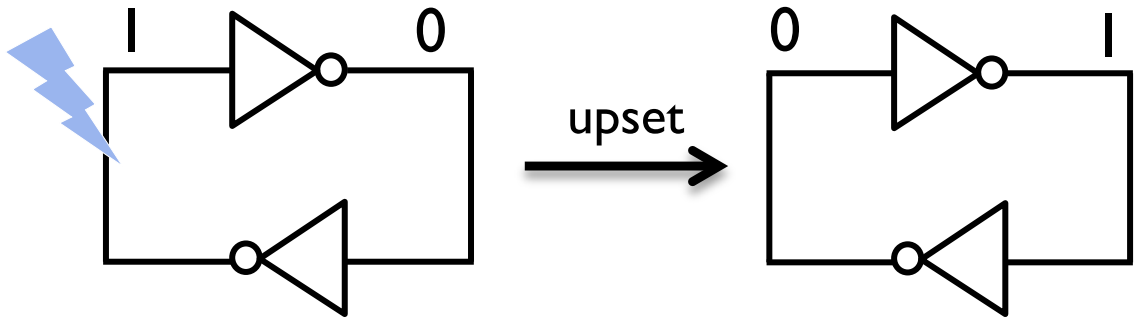
Burn-in

Design
Architecture
ECC

Design margins
Overdesign

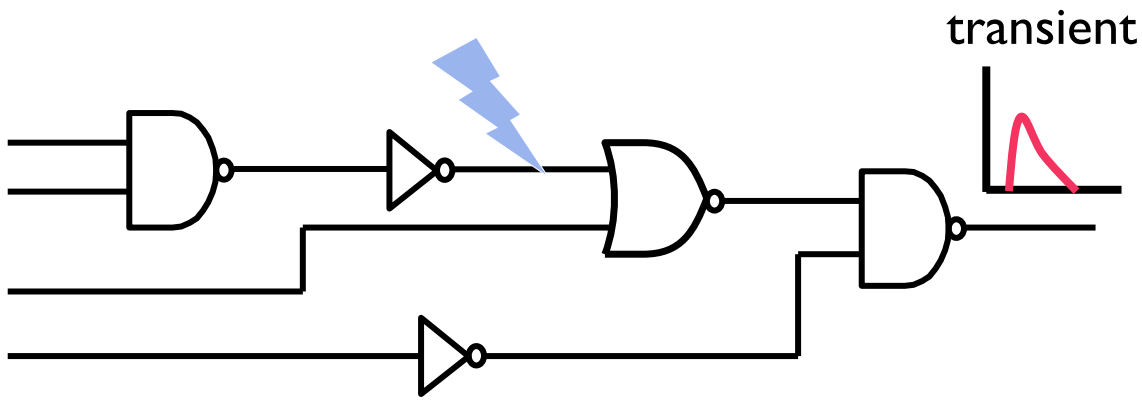
ARM®

Soft error: Impact on circuits

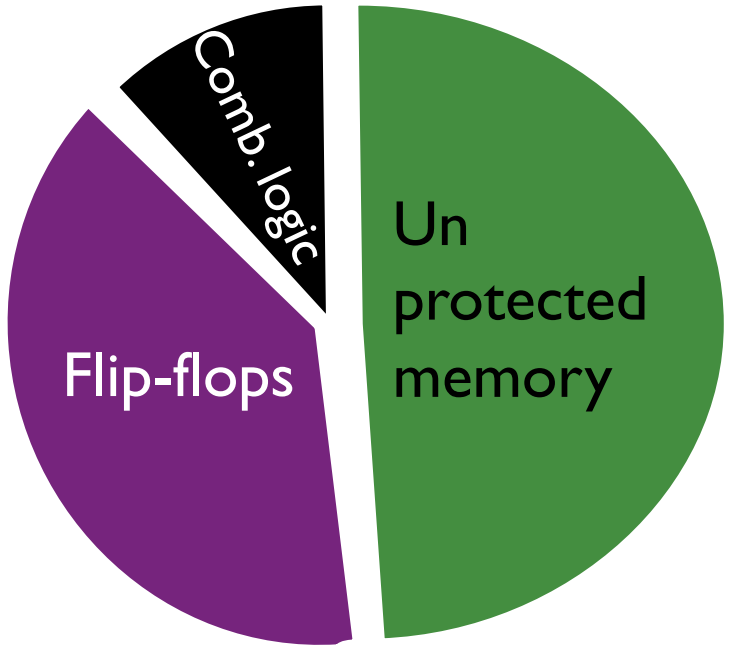


Storage cells

- SRAM bit cells, Flip-flops



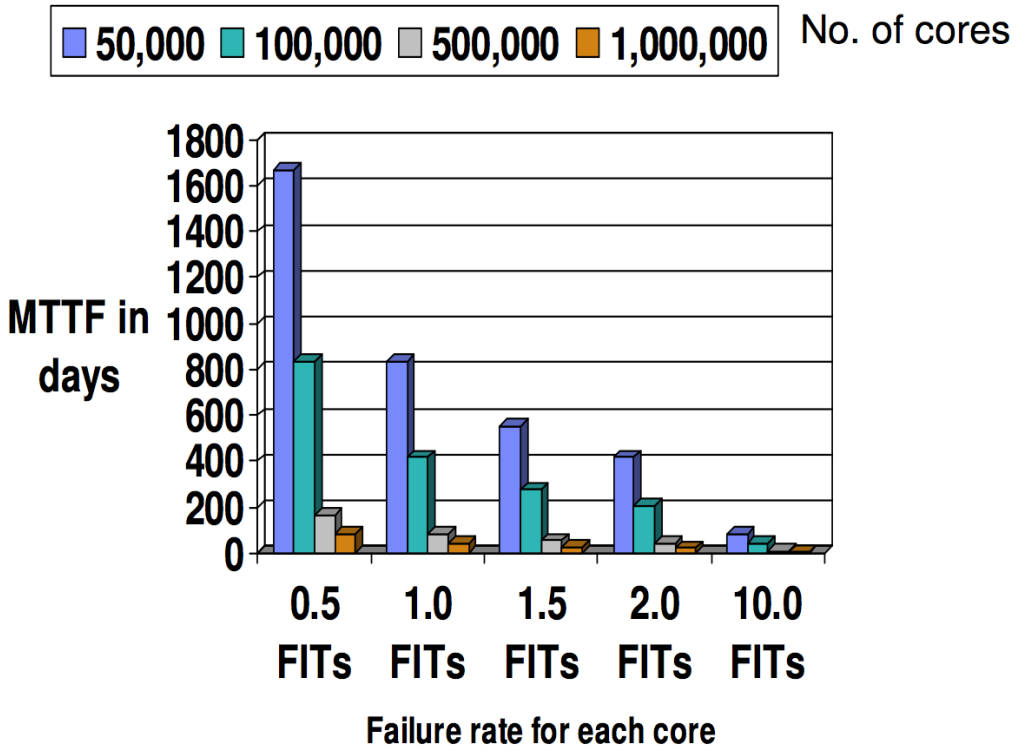
Combinational cells

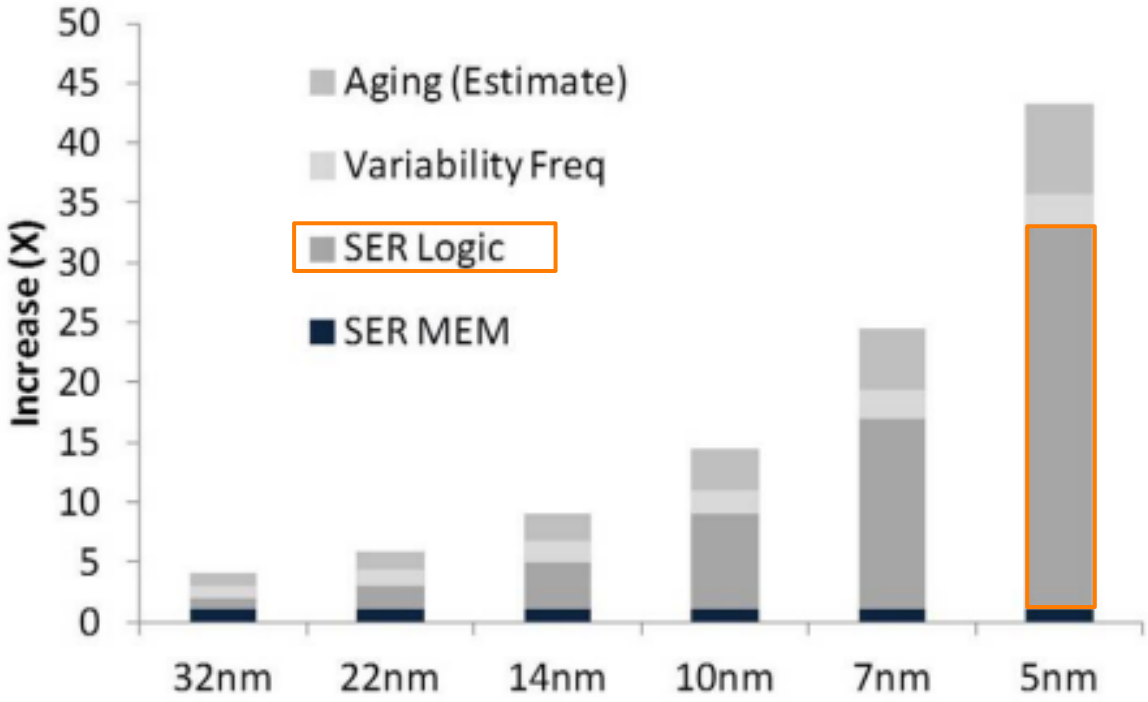


Soft error contribution rate

Big system reliability

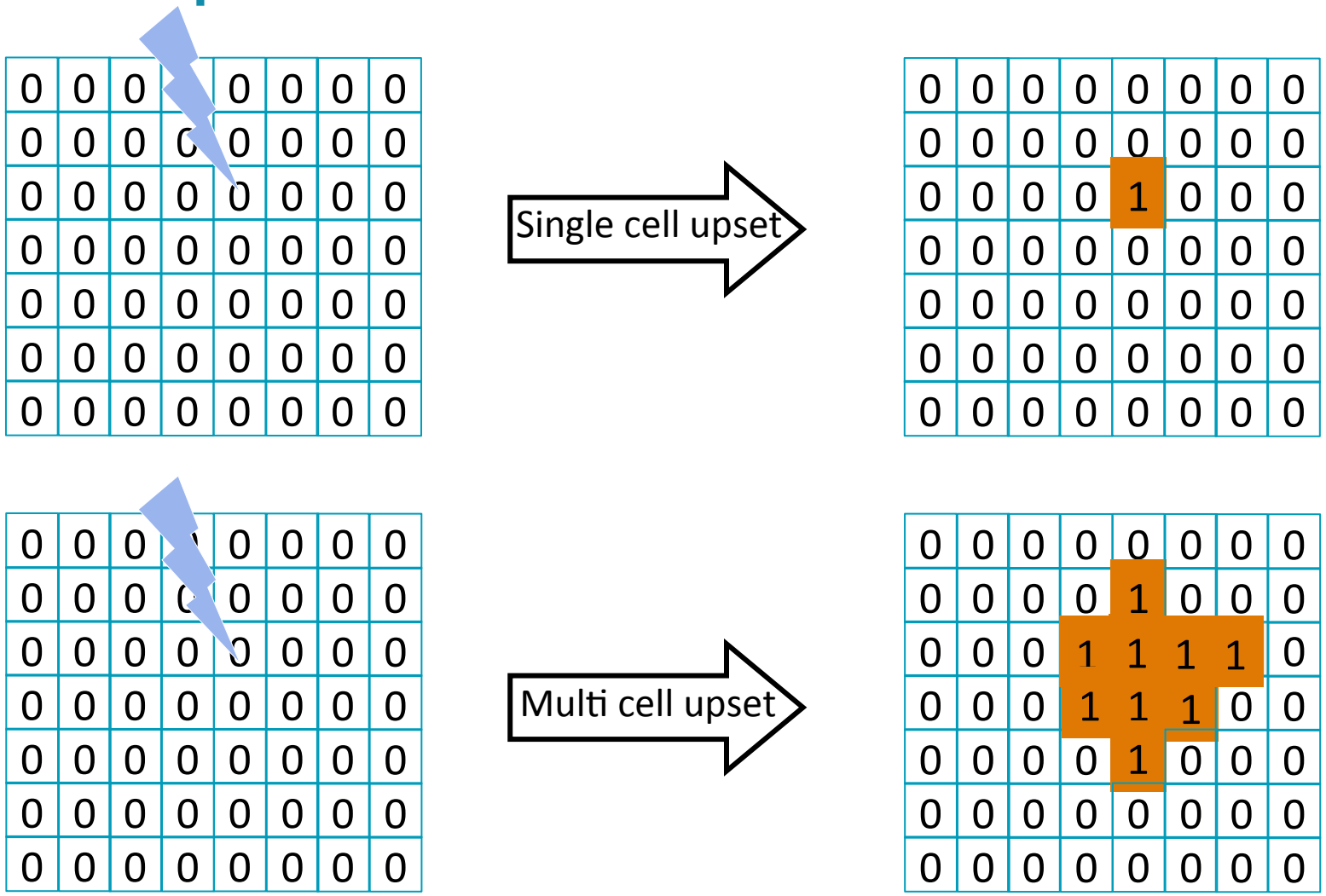
- "Always on" system are especially vulnerable to soft error
  - Vulnerability further increases substantially at low supply voltage
- As technology scales down, number of cores scales up
  - Rates of failures increases: From 2 weeks (current) to 1 hour





Source: Draft ICiS 2012 Reliability Workshop

SRAM upsets



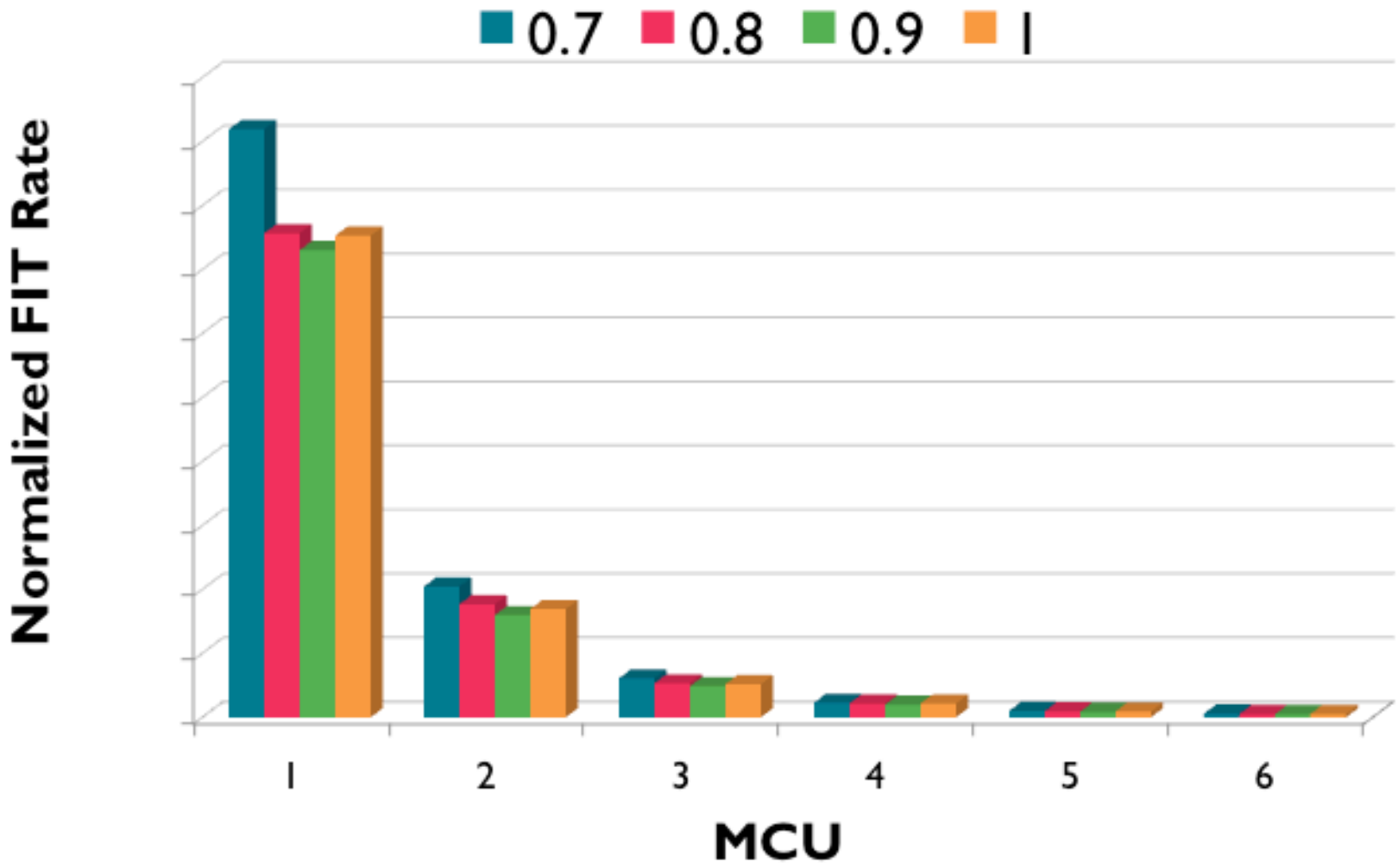
Word arrangement in a mux-4 arch

1	2	3	4	1	2	3	4

Feature size scaling

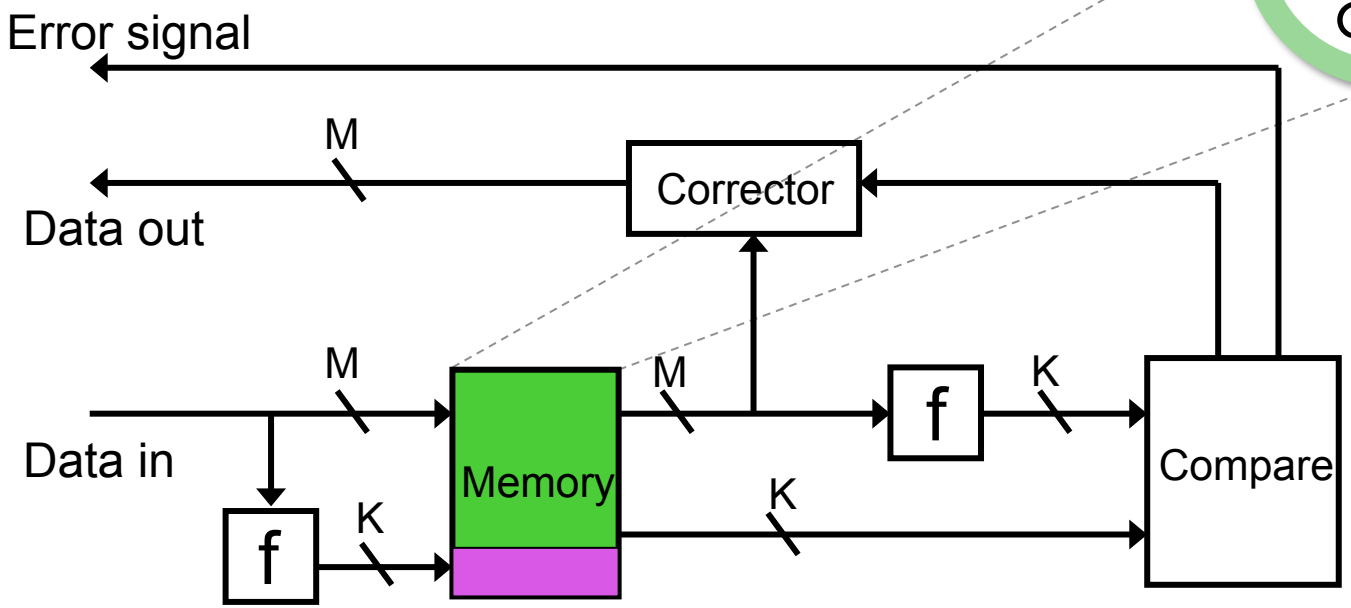
cell area ↓
cell capacitances ↓
Q_{coll} ↓

FIT rate in 28nm: SRAM

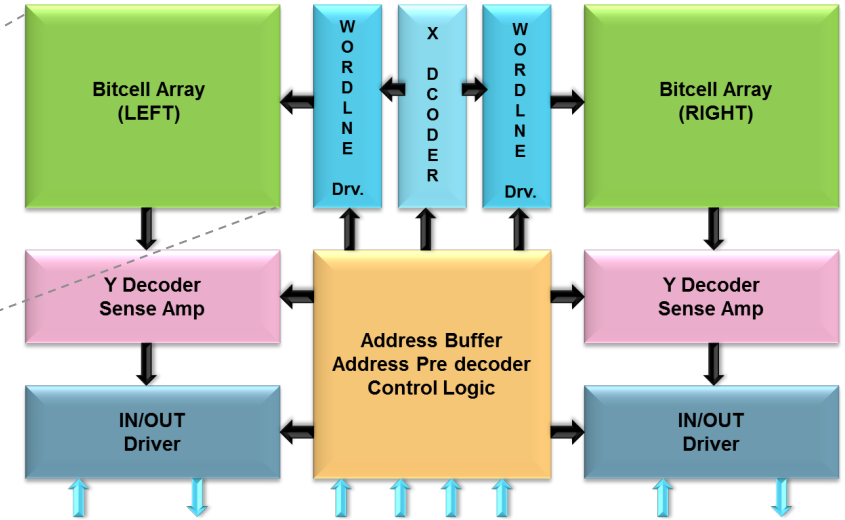


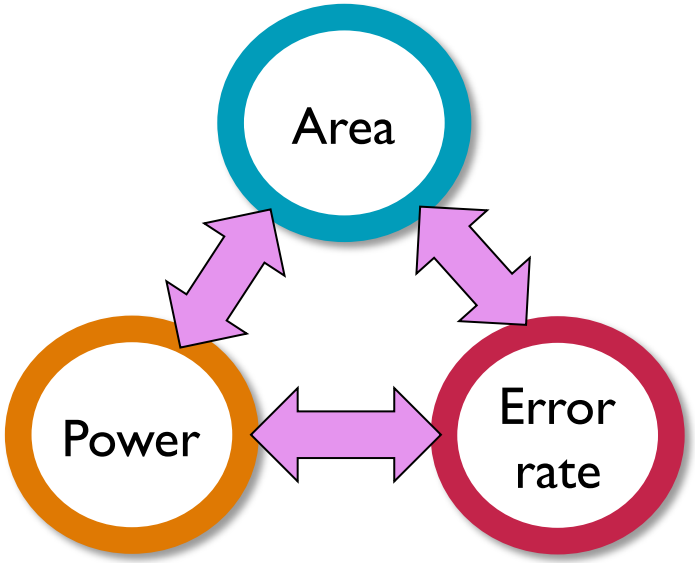
MCU probability = 0 for > 6 cells

Soft error mitigation - SRAM



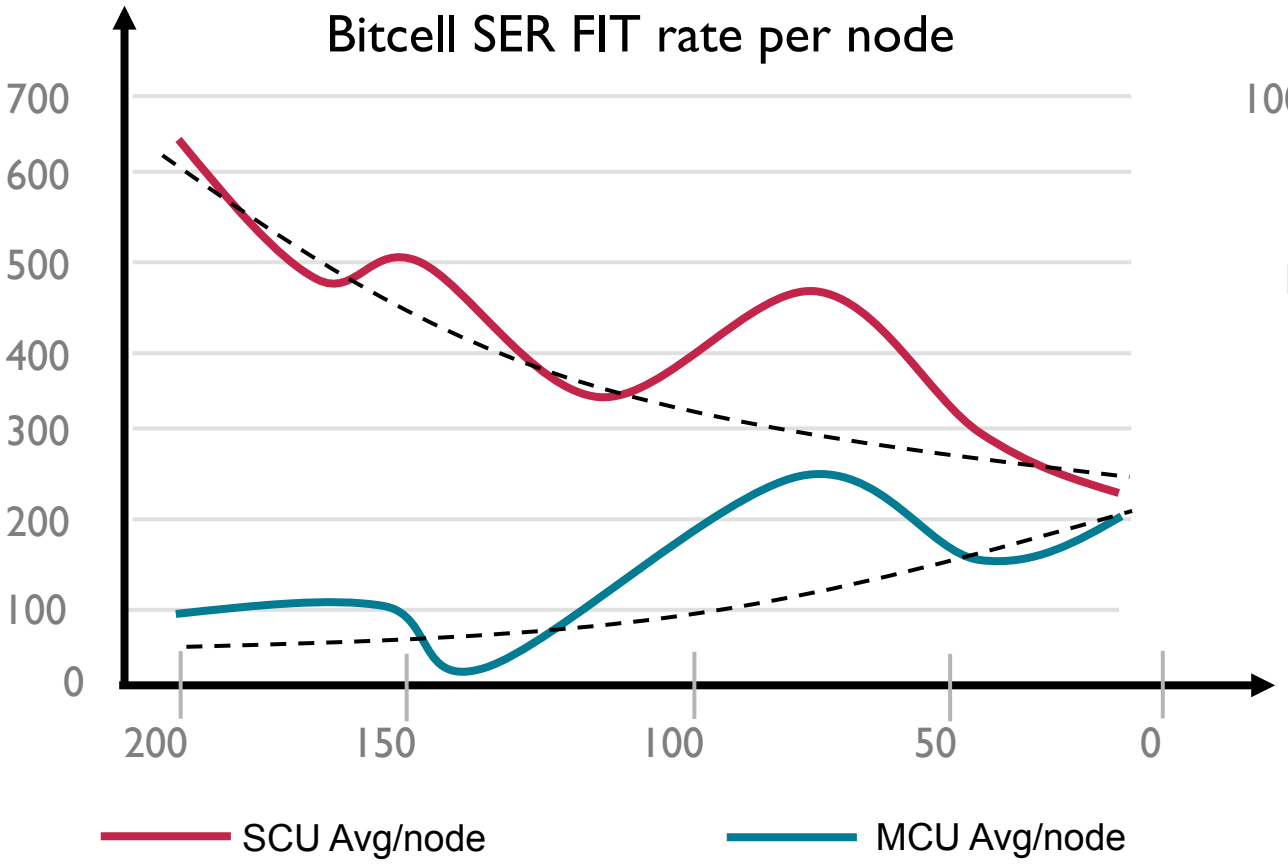
Error Correcting Code

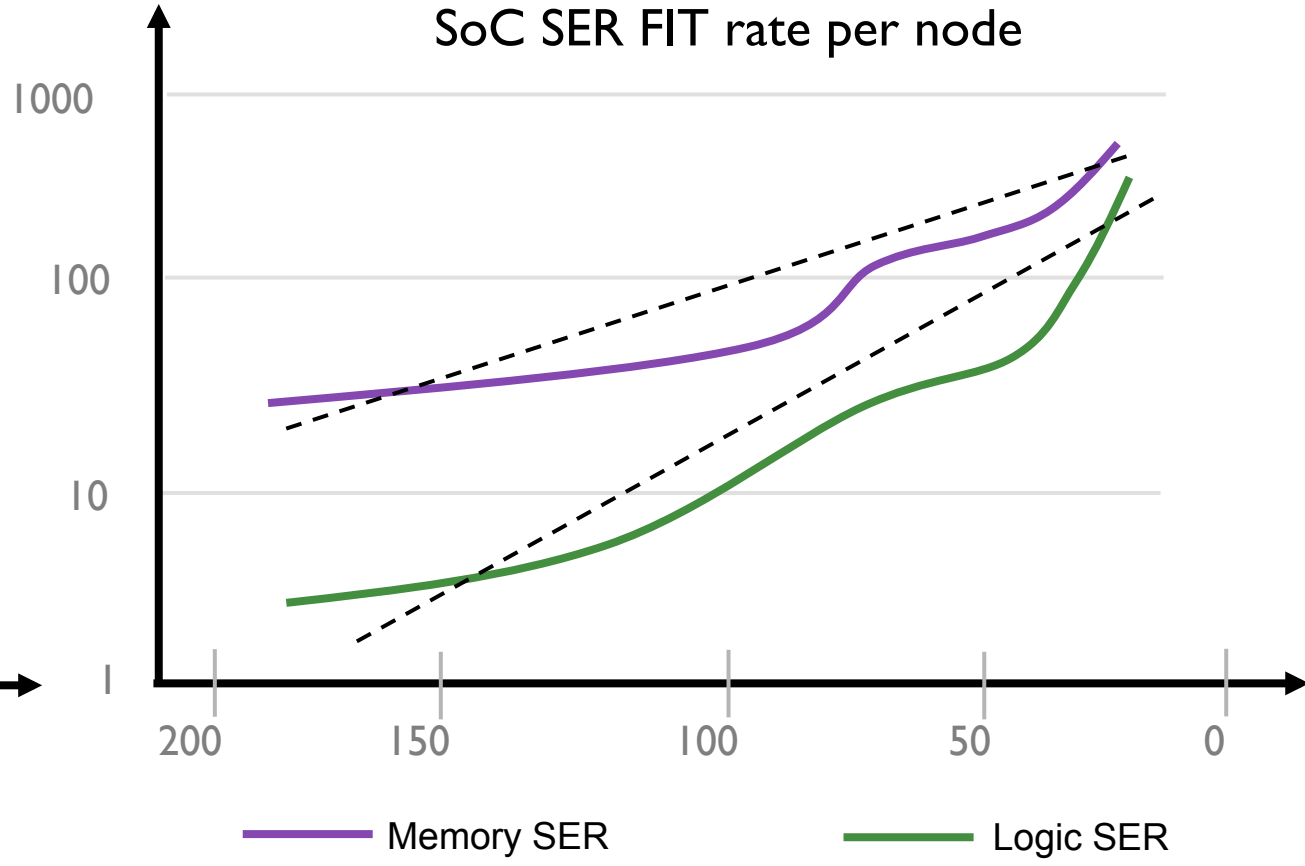




- ECC with physical interleaving reduces the FIT rate to ~0
- Physical interleaving: multi cell error → single bit error
- Temporal scrubbing to correct single bit errors

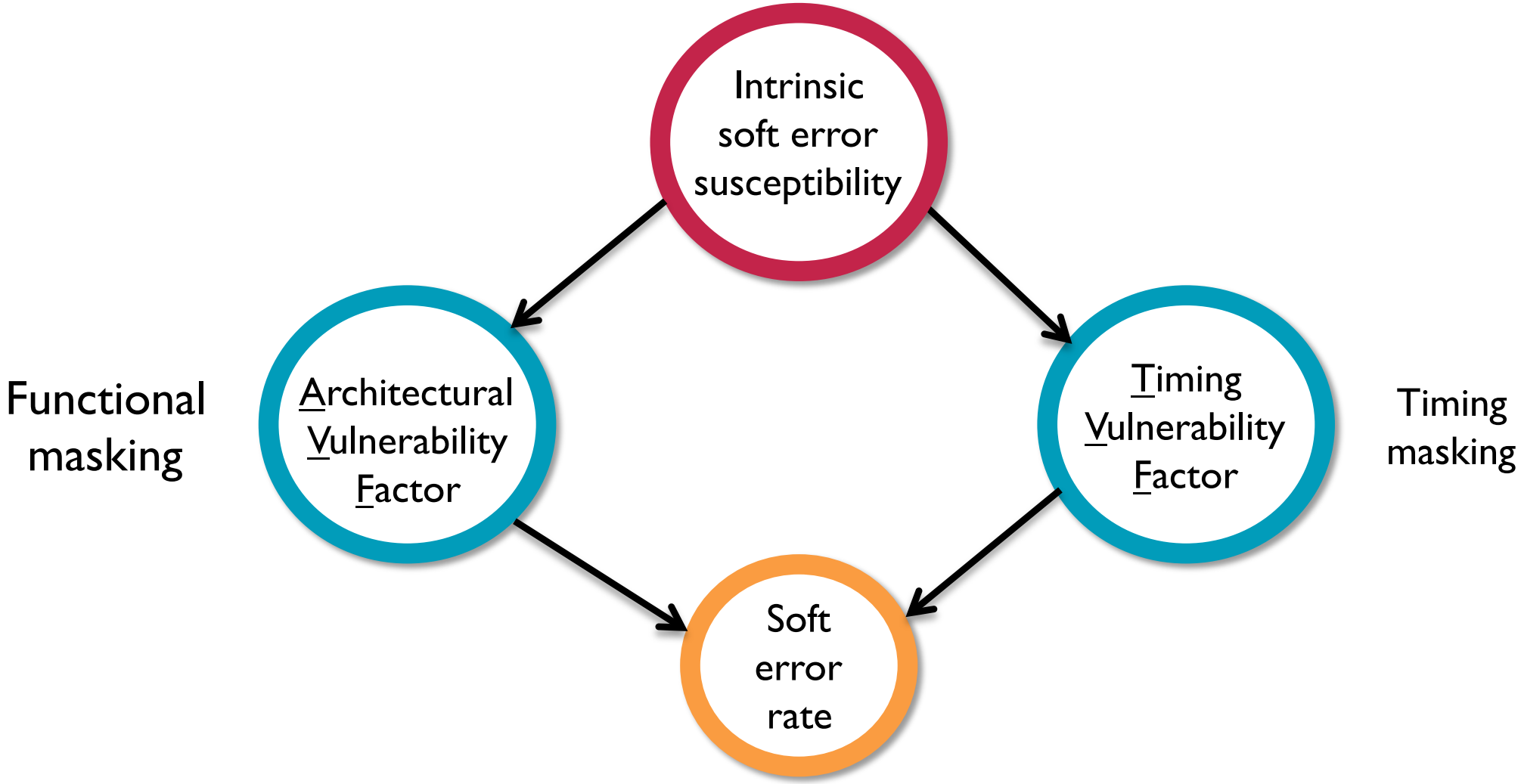
SoC soft error trends





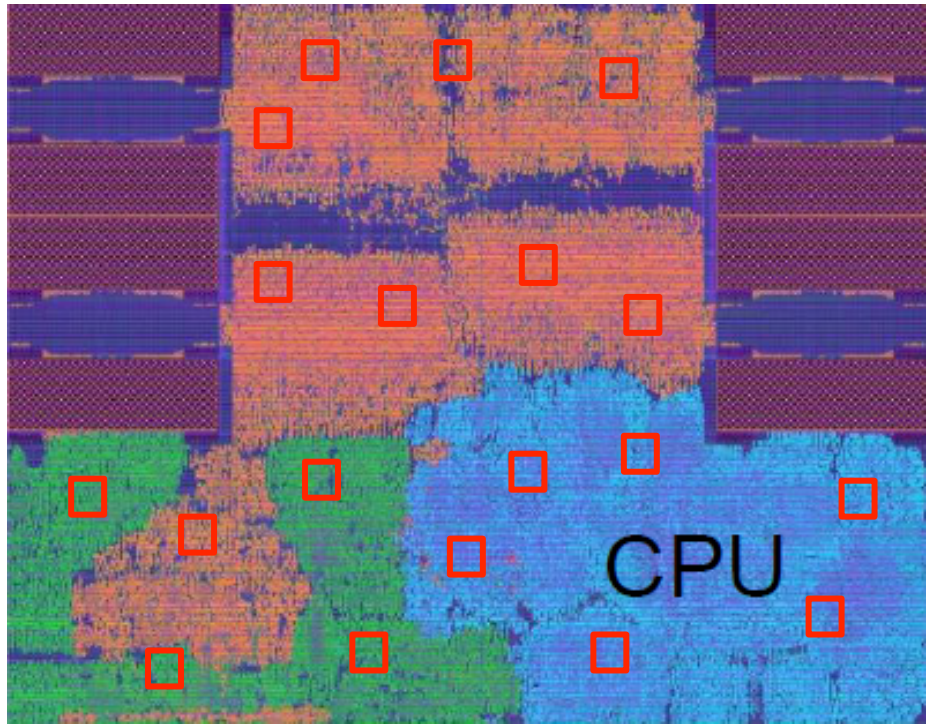
Even though per memory bitcell SER sensitivity is decreasing, overall FIT per SoC is increasing

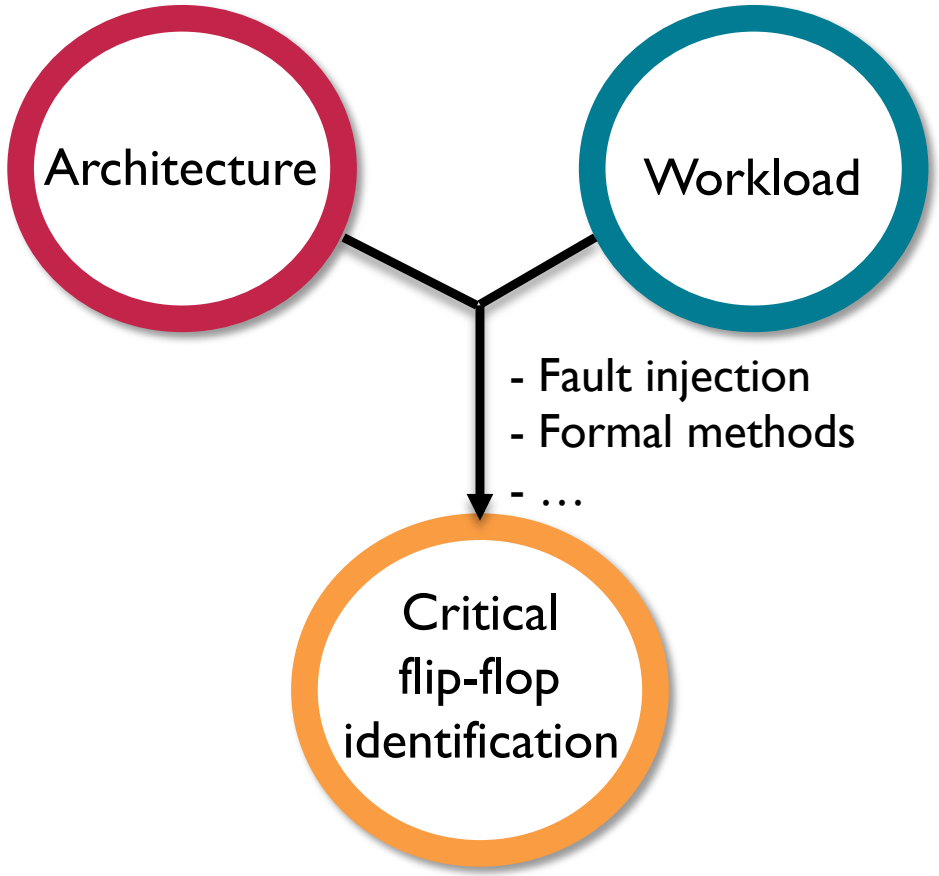
Are all bits equally vulnerable?



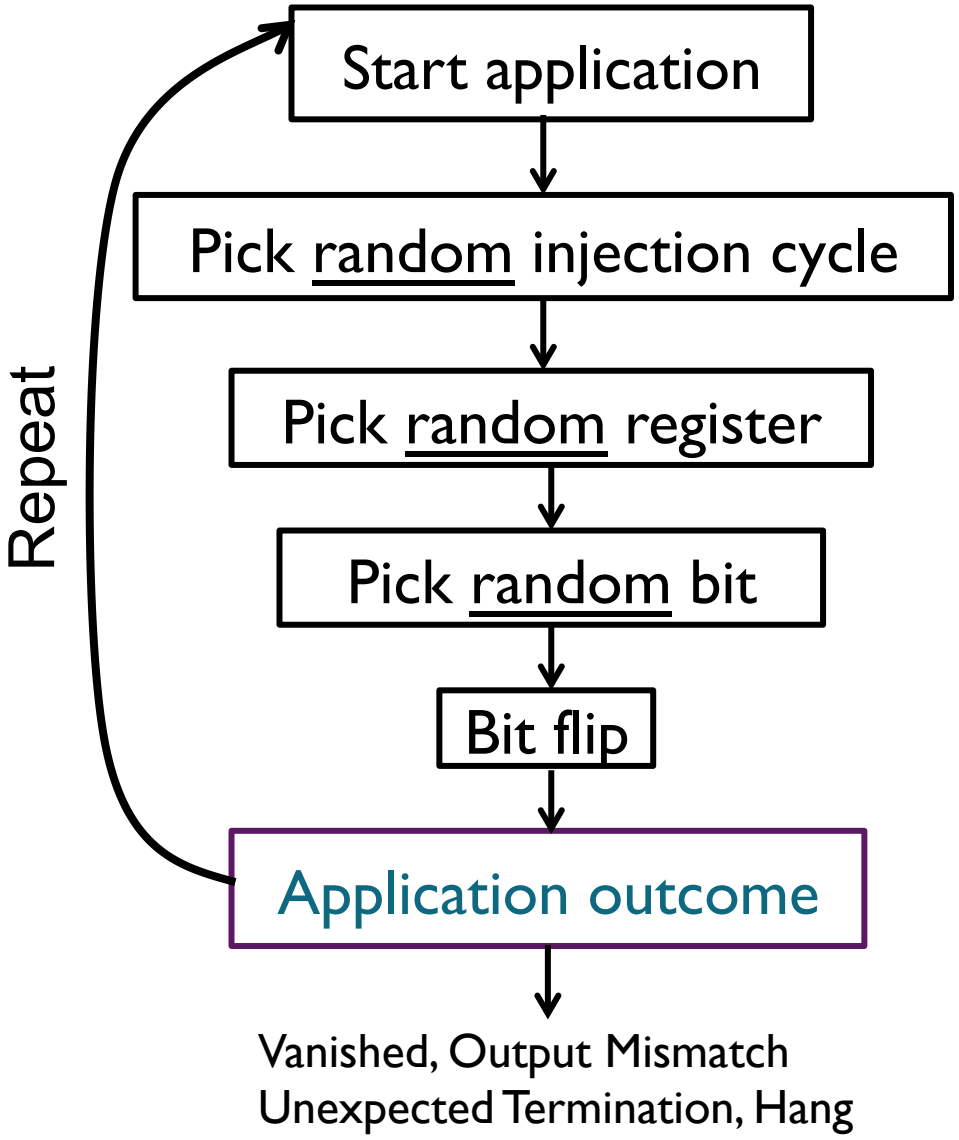
Soft error in logic

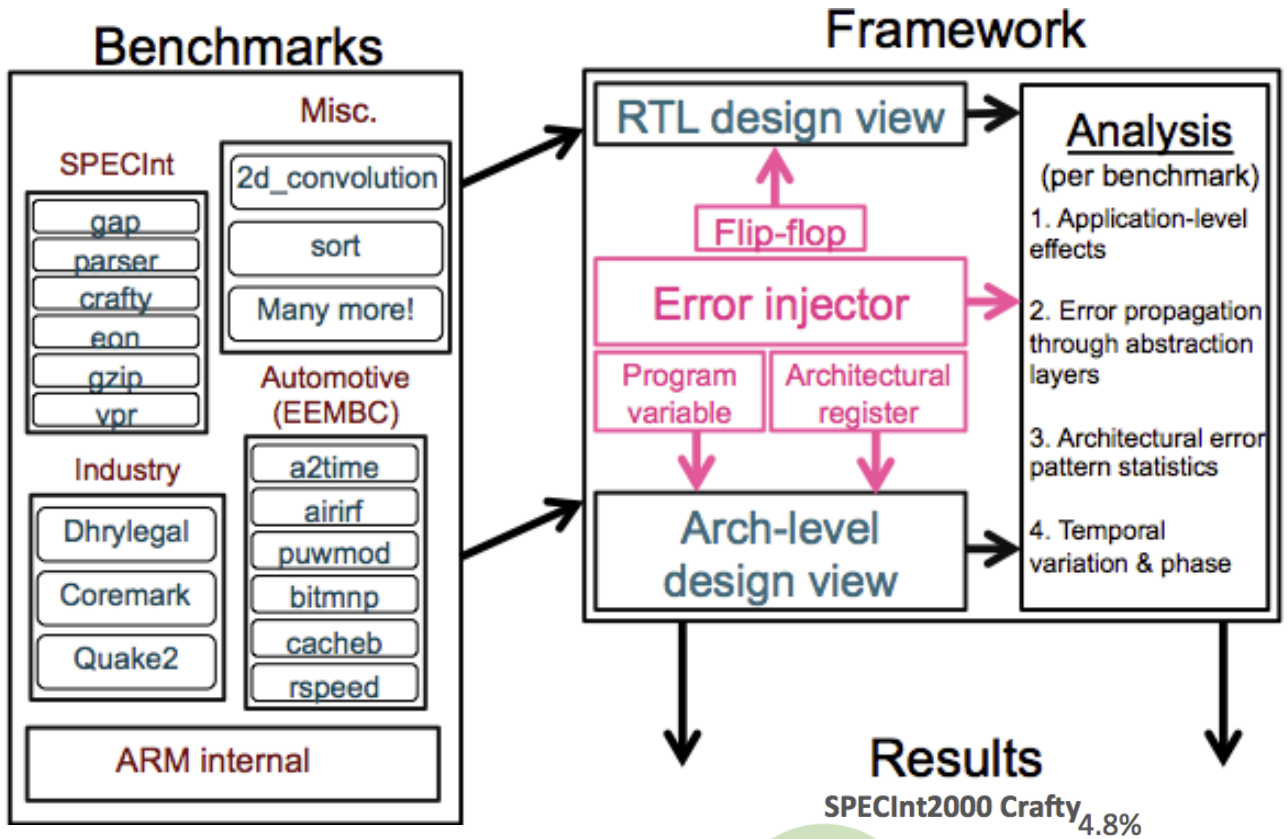
- No "cheap" way to protect - spatially distributed bits
- Vulnerability factor analysis helps to identify critical ones



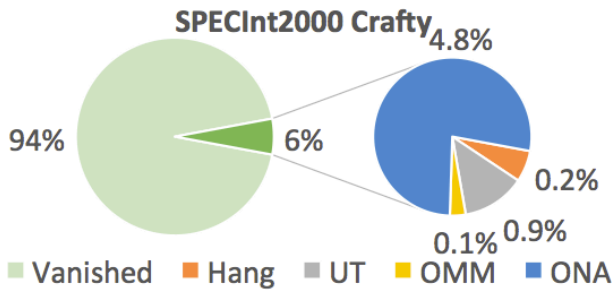


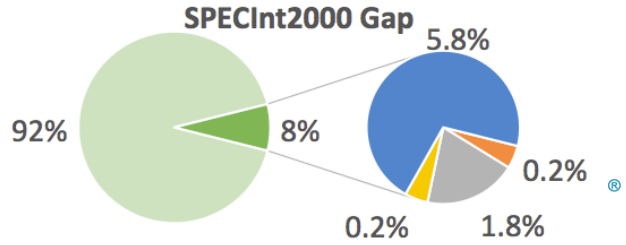
Error injection flow



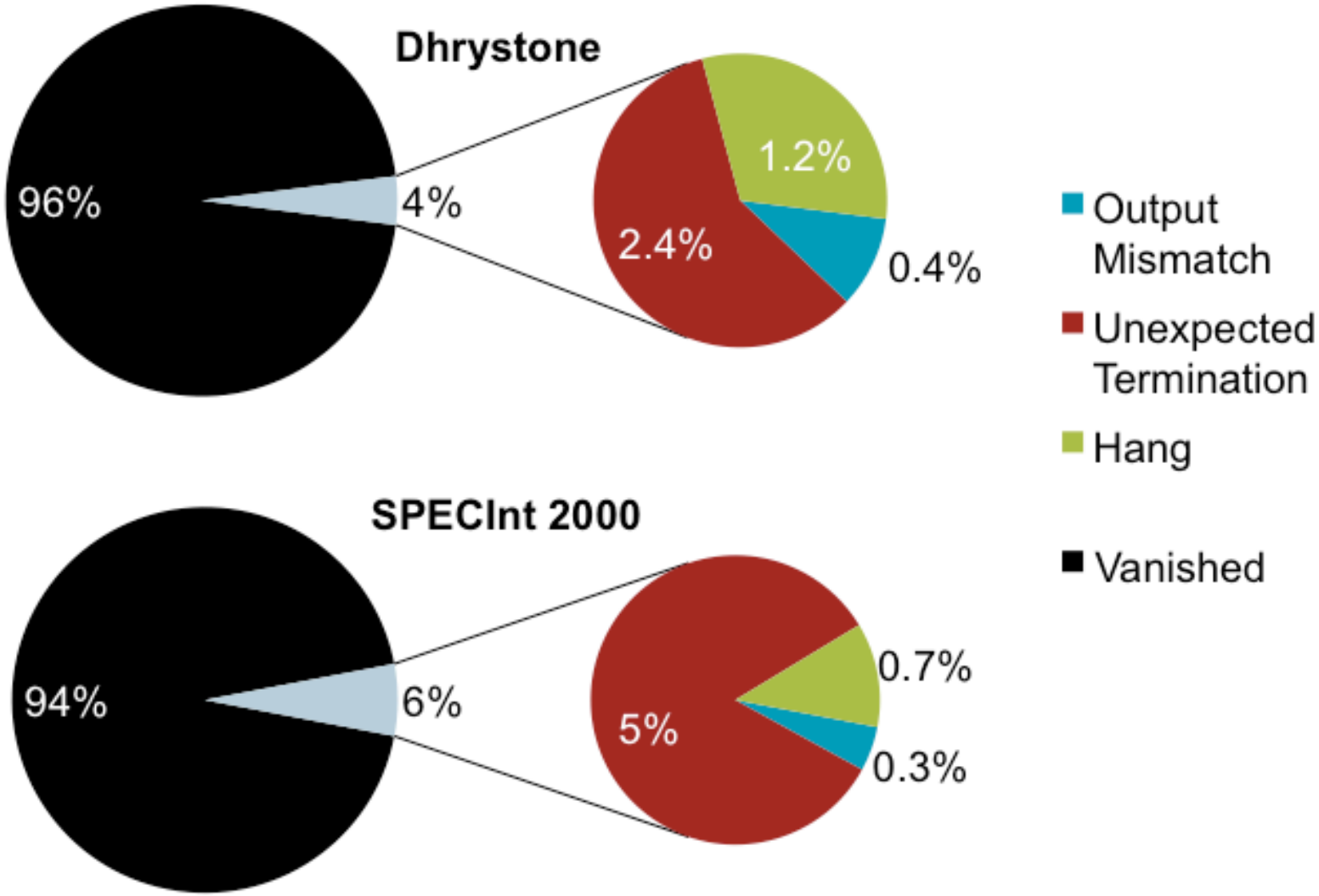


Results



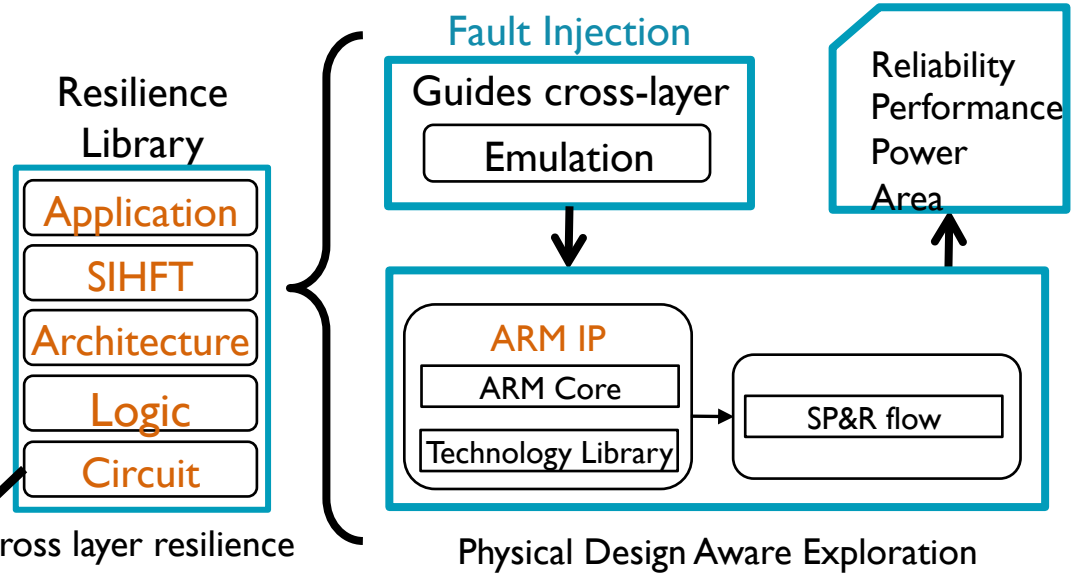


Fault injection results

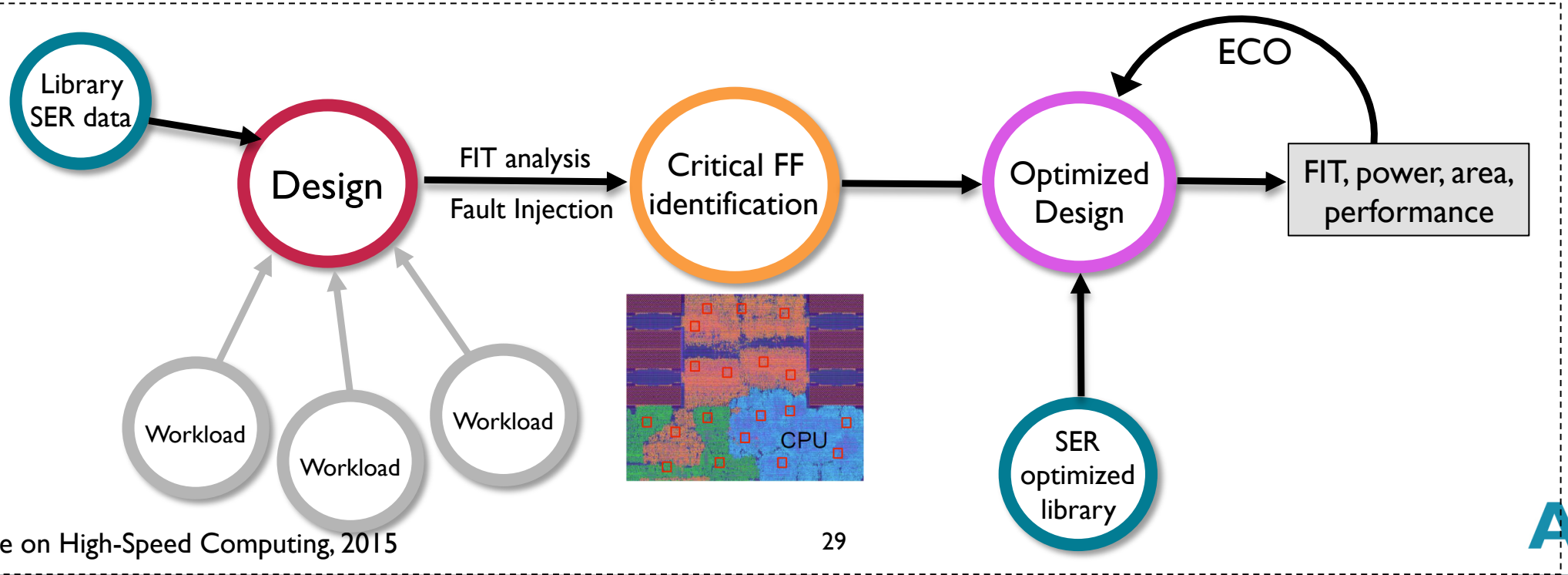


The Key message here is that even though most faults vanish, we still need to worry about the remaining 4-8% of faults.

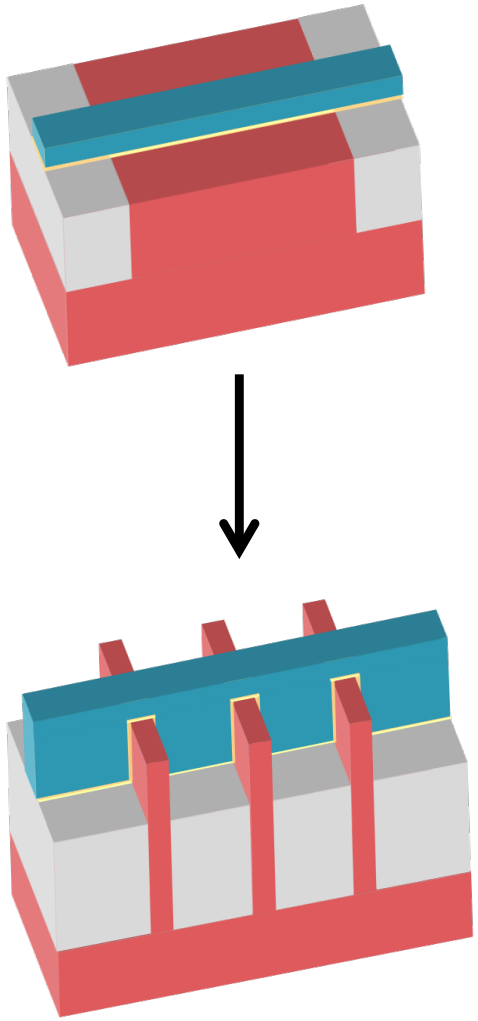
Cross layer resilience approach

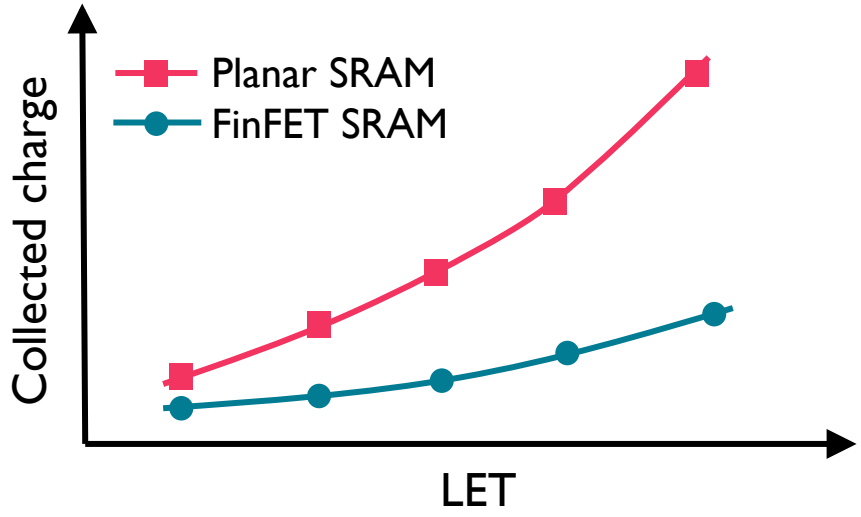


Circuit resilience example



FinFET and soft error rate

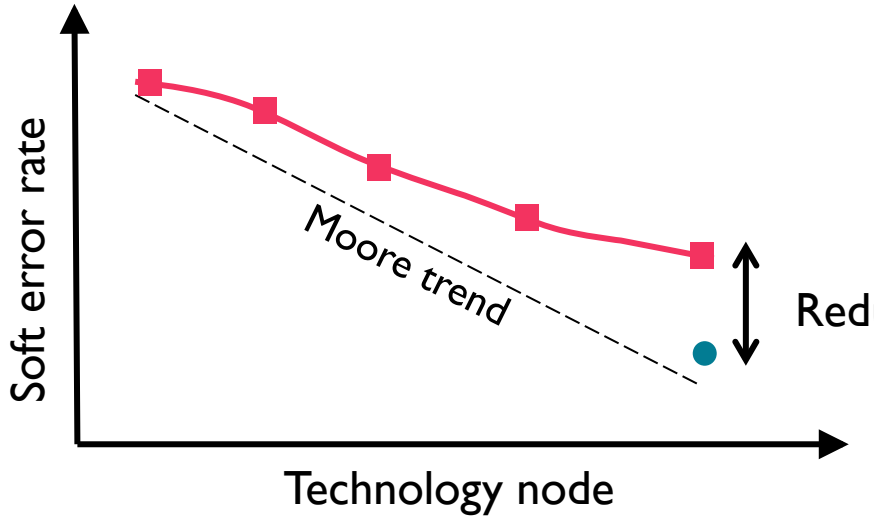




$$SER = A_{diff}\, e^{(-Qcrit/Qcoll)}$$

Qcoll ↓ leads to SER ↓

Source: Fang and Oates, TDMR 2011

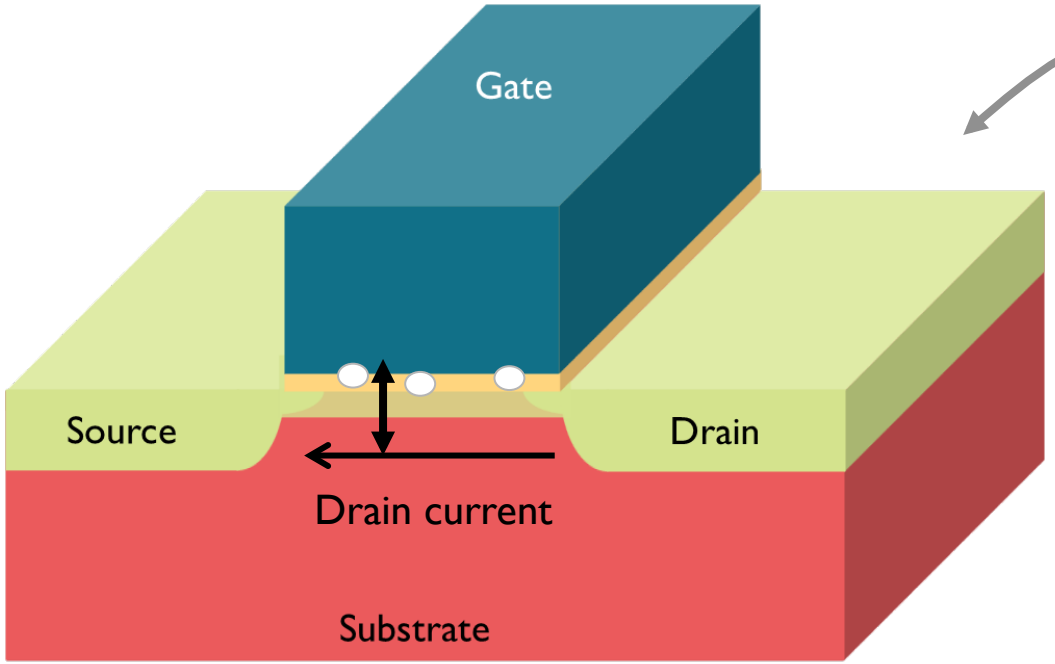


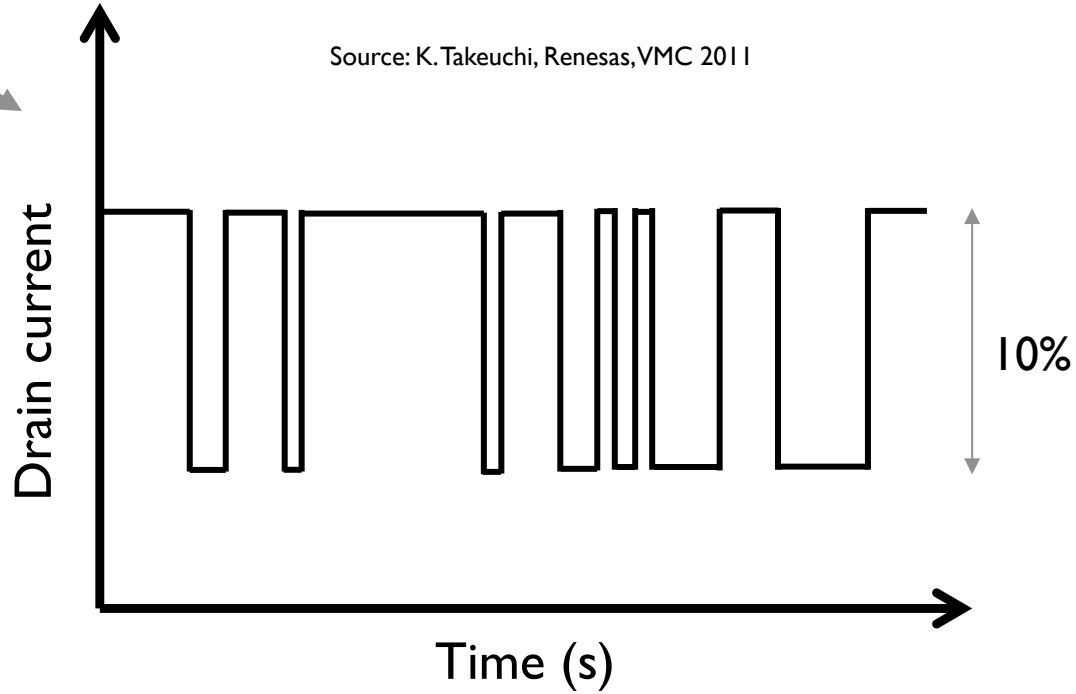
Reduction due to FinFET

Source: S. Ramey, *et al*, IRPS 2013

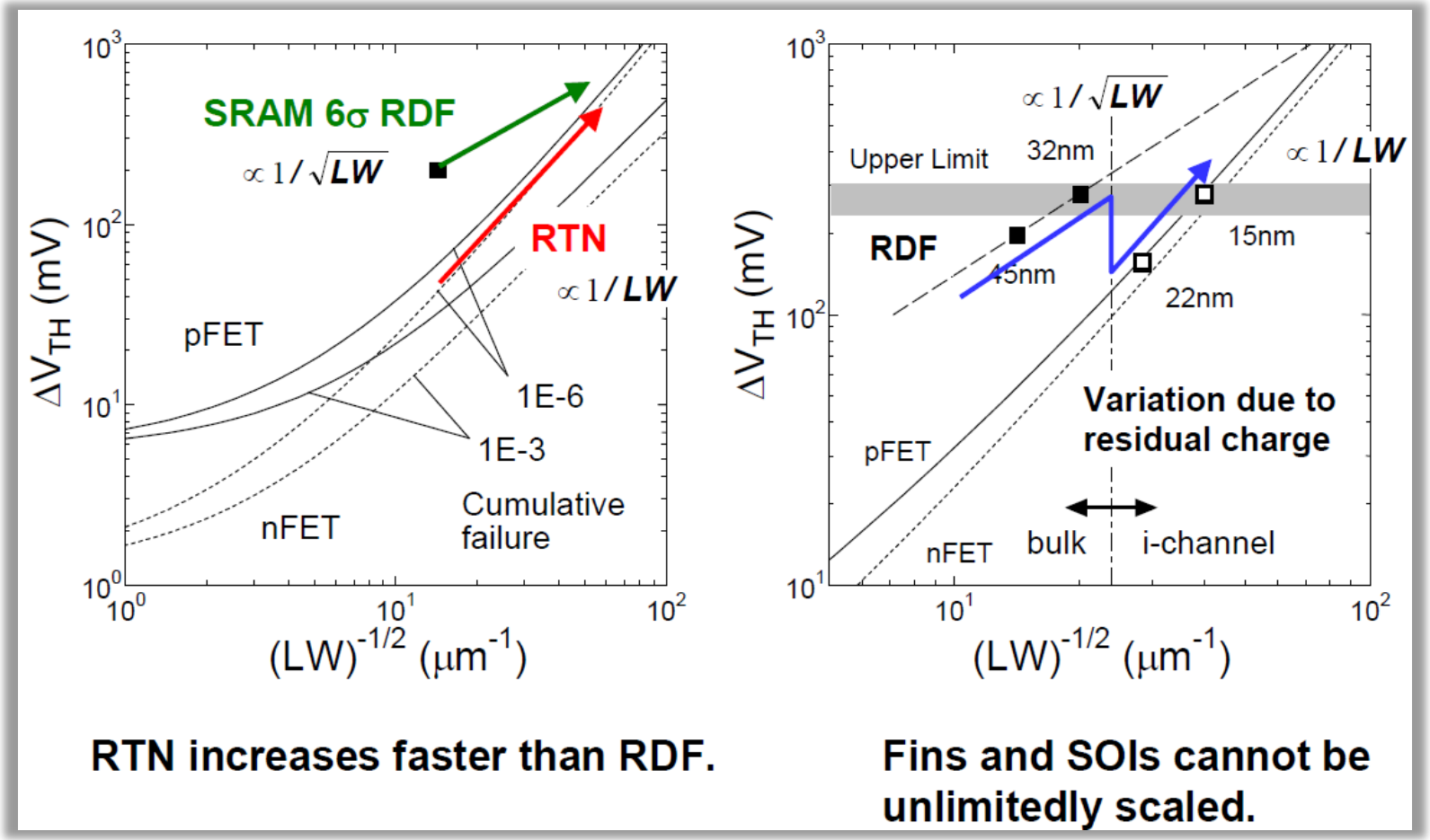
Random telegraph noise (RTN)

- Discrete level current fluctuation with time
- Caused by charge trapping/de-trapping in dielectric
- Behavior is results in 1/f noise for large FETs





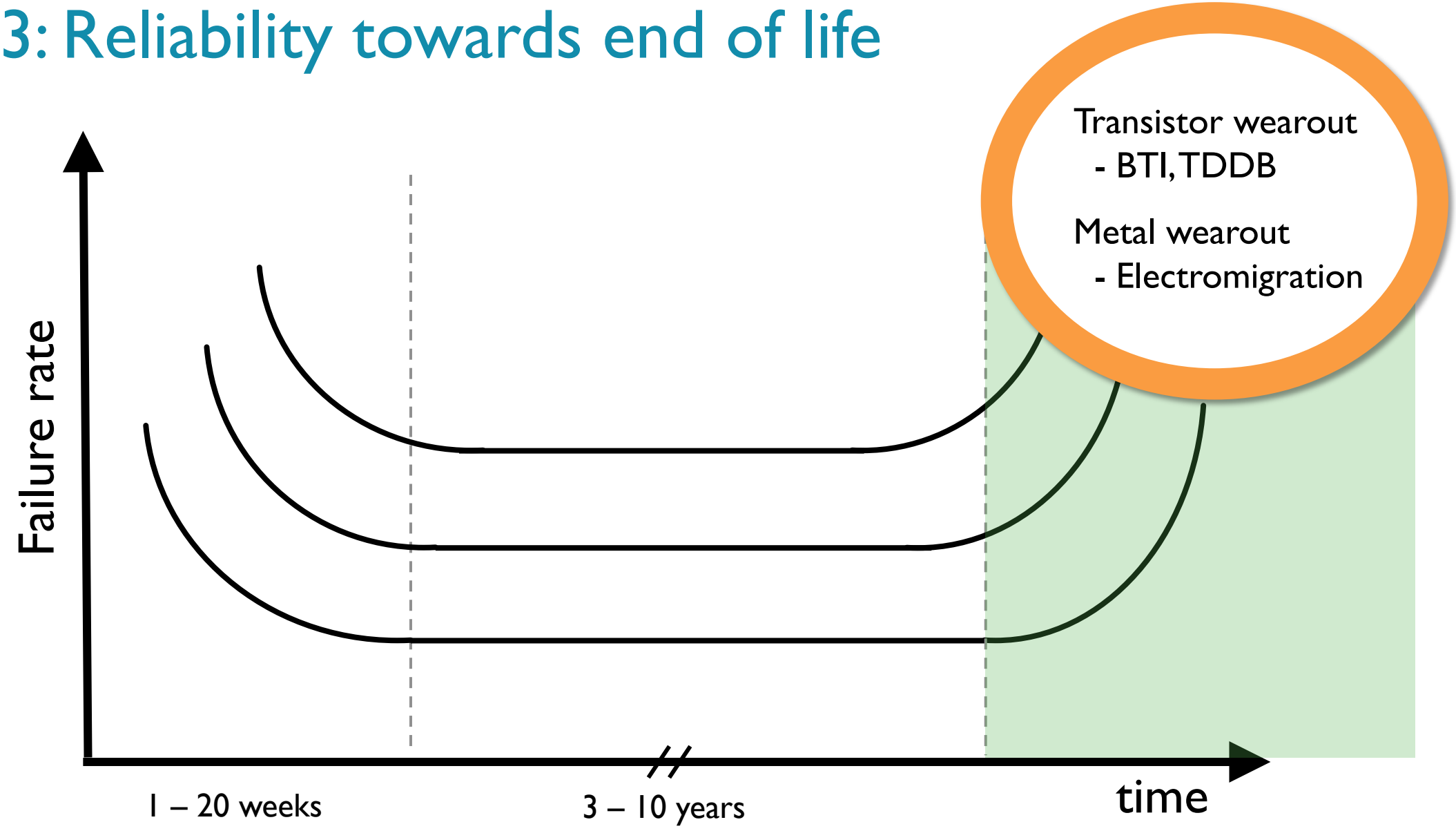
RTN scaling



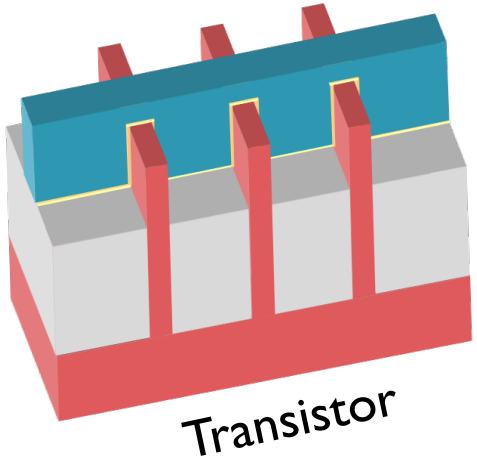
Source: K. Takeuchi, Renesas, VMC 2011

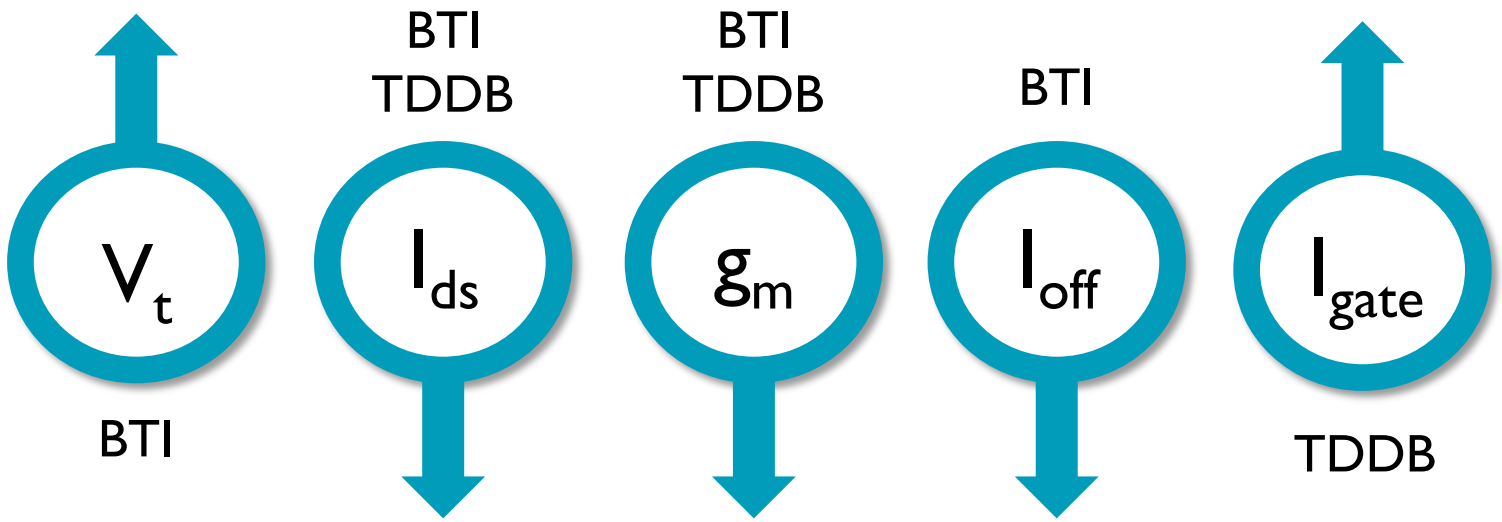
RTN scales faster (1/LW) as opposed to other variability phenomena which scale as $1/(LW)^{1/2}$

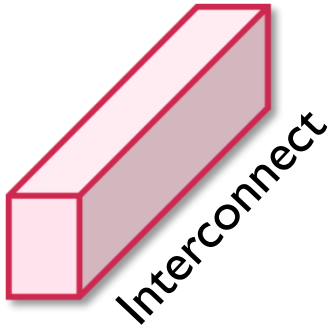
Phase 3: Reliability towards end of life

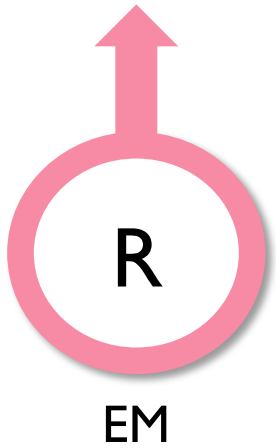


Wearout/aging









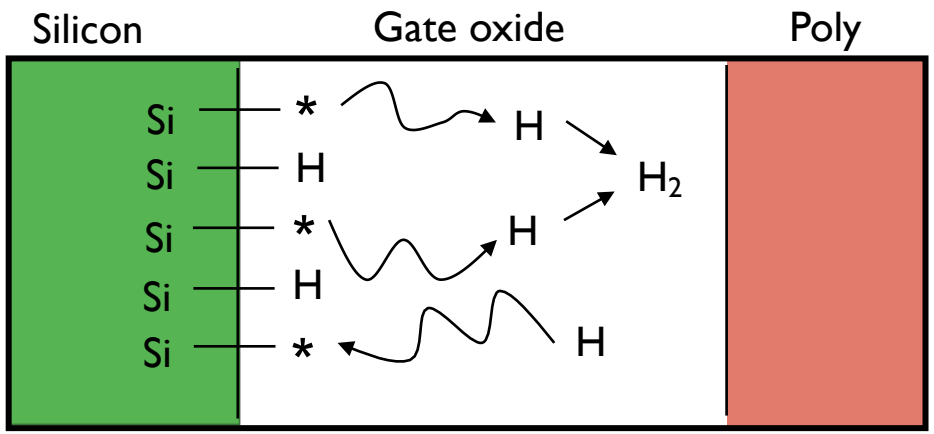
BTI → Bias Temperature Instability

TDDB → Time Dependent Dielectric Breakdown

EM → Electromigration

Bias Temperature Instability (BTI)

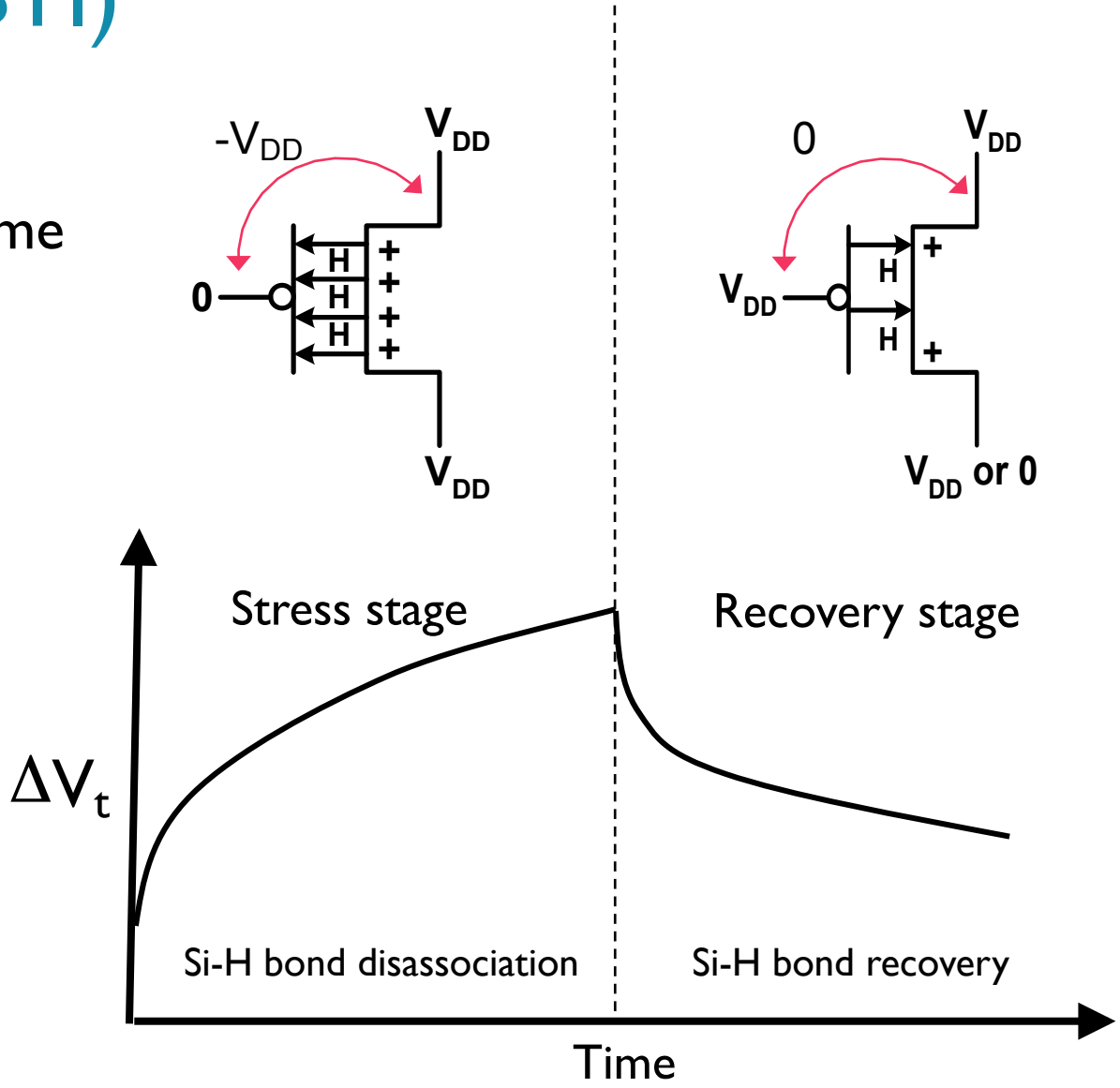
- Reliability concern at Si-SiO_N interface
- Gradual shift in transistor parameters with time



PMOS

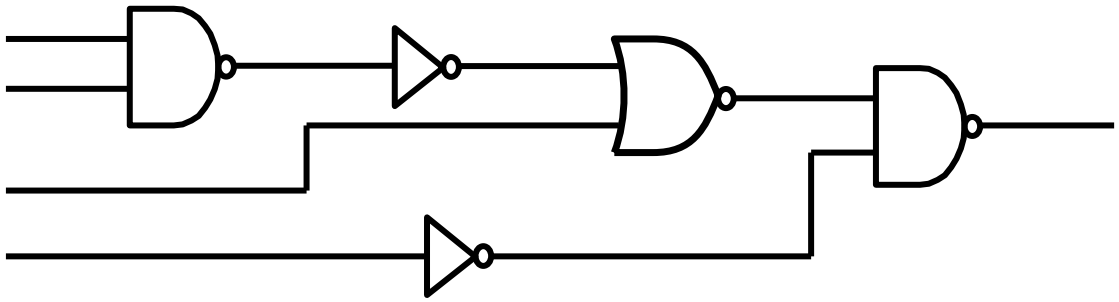
Negative Bias: Si-H bond disassociation

Zero Bias: Si-H bond recovery



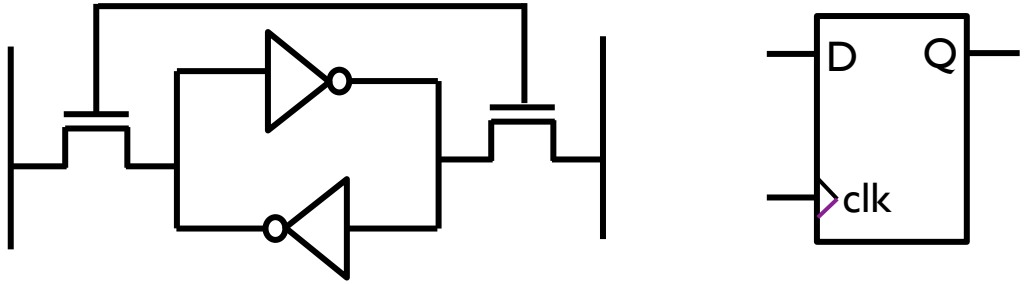
Wearout: Impact on circuits

Combinational



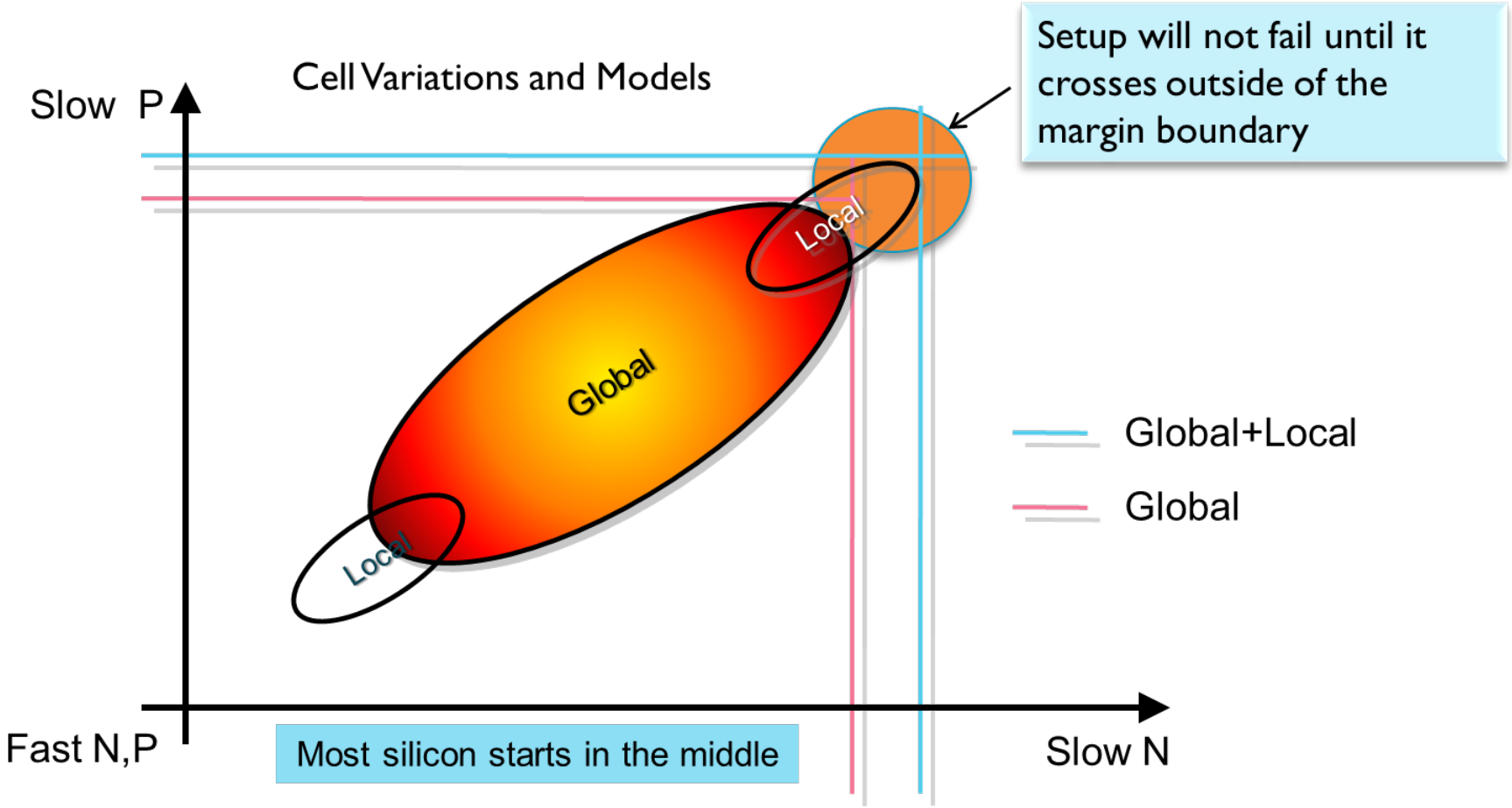
- F_{max} ↓
- Timing failure as circuits age

State-holding cells
(bit cells, flip-flops)



- Static Noise Margin ↓
- Read and write stability ↓
- Parametric yield loss

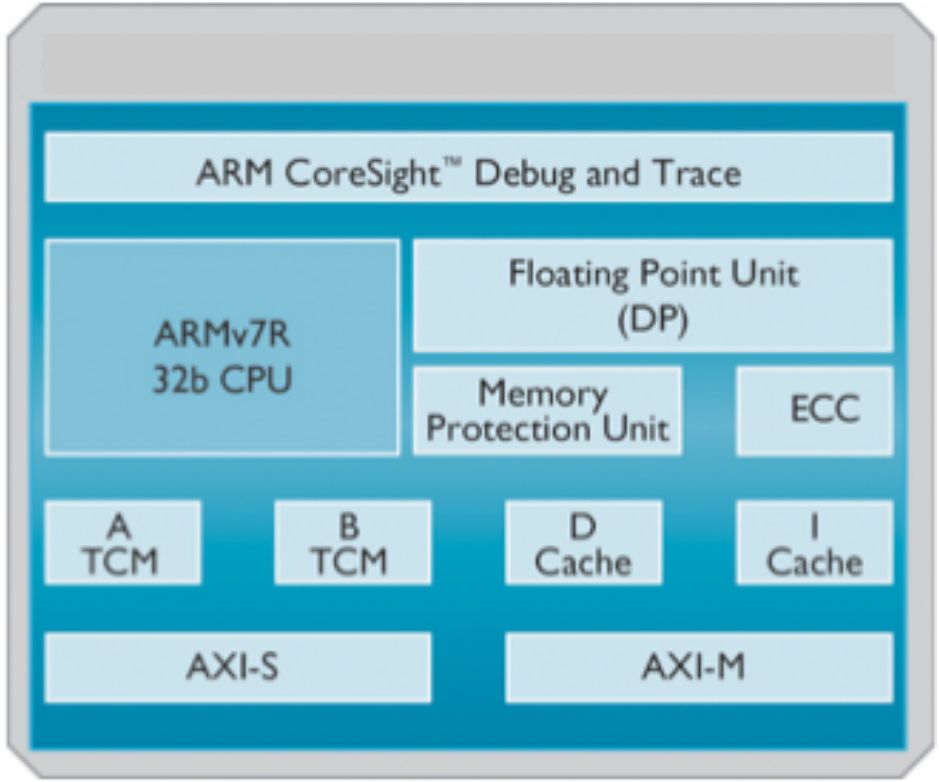
Delay Degradation ≠ Failure



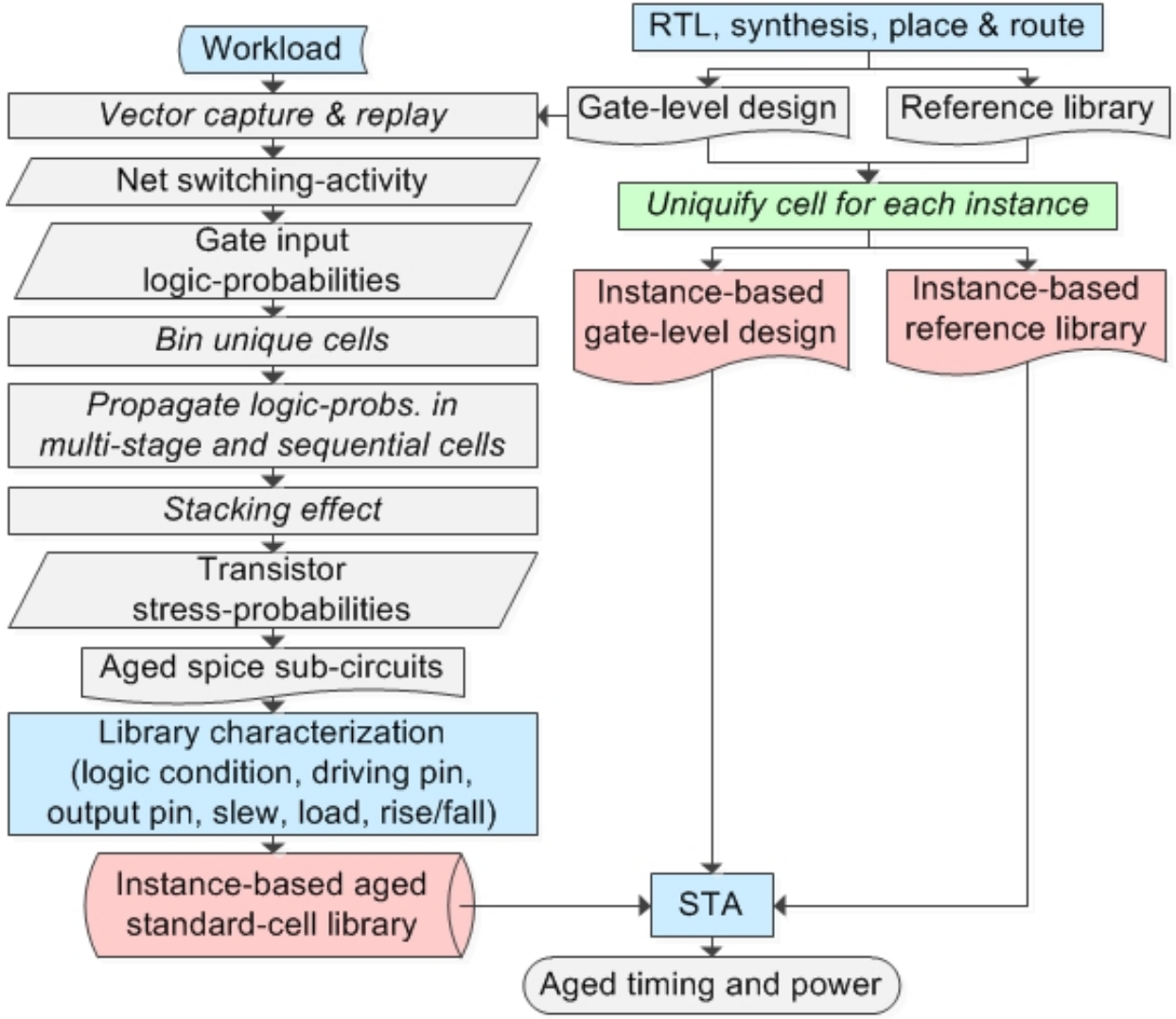
But there could be hold errors due to aging in clock path

CPU workload dependent aging: A case study

- Mid-size ARM CPU
  - > 100K instances, > 10K sequentials
  - Large enough to be interesting
  - Small enough for rapid turnaround time
- Simulation settings
  - Vdd = 0.9V, Temp = 105 °C, Lifetime = 3 years
  - CP 28LP standard-cell library
  - >1K cell-topologies, >10K timing-arcs
- NBTI and PBTI aging model
  - RD: Reaction-Diffusion [Gielen 11, Zheng 09]
  - TD: Trapping-Detrapping [Velamala 12]



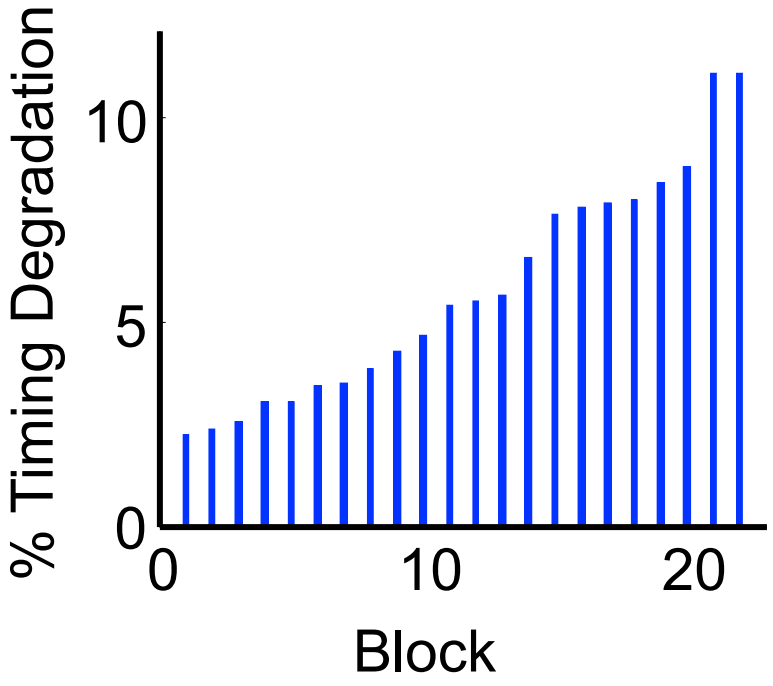
Instance based simulation flow

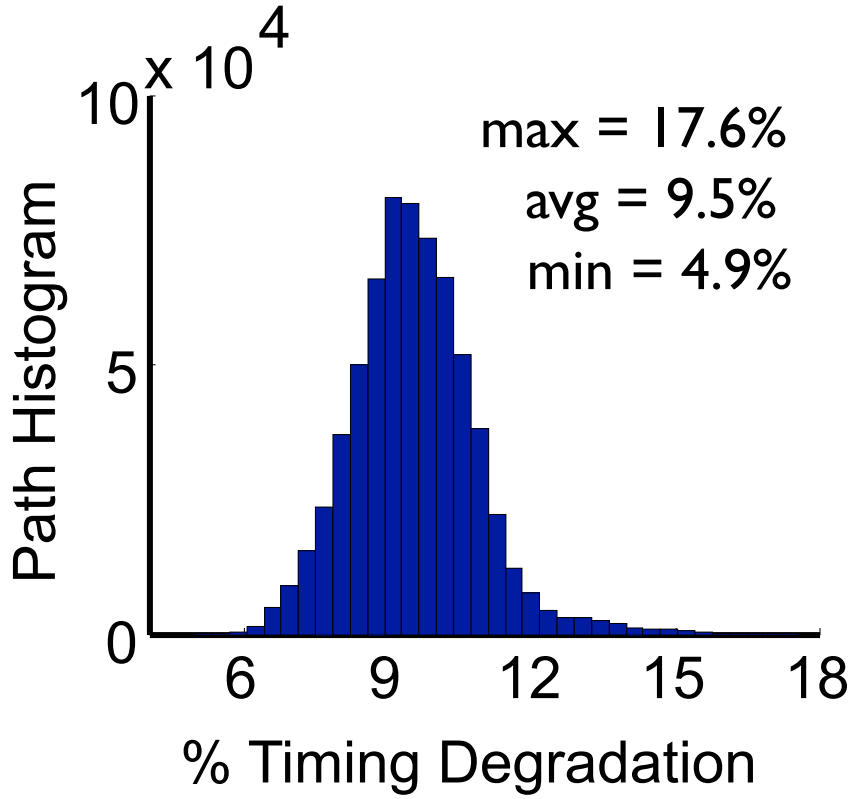


For more details:
Workload Dependent NBTI and PBTI Analysis for a sub-45nm Commercial Microprocessor, *Intl. Reliability Physics Symposium (IRPS)*, 2013.

Block and Path Timing Degradation

- Dhrystone workload





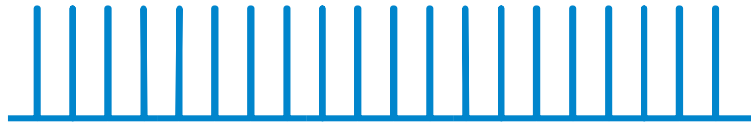
Path Rank Analysis

- Rank paths in fresh and aged design, sorted by slack
- Non-critical paths can become critical and vice versa

Path rank		% Timing Degradation
Fresh	Aged (Dhrystone)	
1	14084	7.64
2	9781	7.94
3	9329	8.02
4	12345	7.87
5	6220	8.31
6	36672	7.16
7	7771	8.19
8	11580	7.96
9	28975	7.40
10	20054	7.66

Path rank		% Timing Degradation
Aged (Dhrystone)	Fresh	
1	179394	15.61
2	145042	15.41
3	134419	15.18
4	1413427	17.57
5	272323	15.67
6	224034	15.46
7	331934	15.76
8	275422	15.56
9	481425	16.06
10	208561	15.24

Workload Power-State Trace

Workload	Power-State (Active / Sleep) vs. Time	Average Active Time
mp3		0.03
web-browse		0.07
3D rendering		0.24
Dhrystone		0.40
video H264		0.54

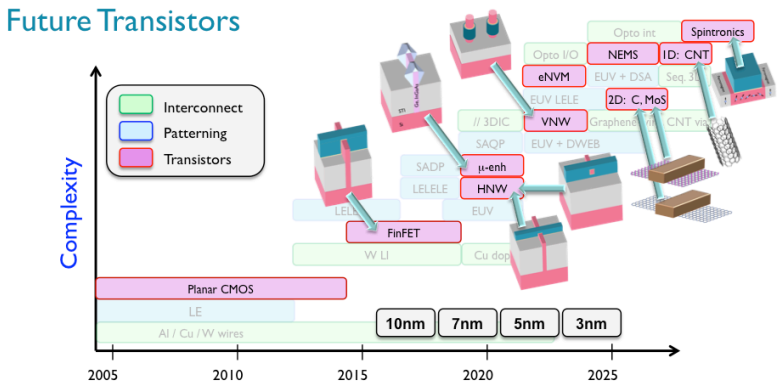
Workload-Dependent Processor Aging

Workload	% Timing Degradation					
	Switching-activity		Power-state		Switching-activity and power-state	
	RD	TD	RD	TD	RD	TD
mp3	10.0	6.3	3.6	2.5	2.3	1.6
web-browse	10.6	7.3	6.1	4.2	4.1	2.8
3D rendering	12.3	8.4	6.9	4.7	5.4	3.7
Dhrystone	11.2	7.7	10.1	6.9	7.3	5.0
video H264	12.5	8.5	11.4	7.8	9.1	6.2
worst-case	15.6	10.7	15.6	10.7	15.6	10.7

RD: Reaction-Diffusion model
TD: Trapping-Detrapping model

Conclusions

- Reliability is more critical than ever for RAS critical designs – Servers, HPC etc.
- Errors can be mitigated at devices, circuits, architecture or even software level
  - Lots of opportunities to design reliable systems from ground up
- As we scale, some things are becoming worse, some better
  - BTI, TDDB, EM ↑
  - SER ↓
- Future technology challenges
  - Need to keep an eye on the evolution of future devices
  - Quantification of wearout impact on CPU PPA scaling
  - RTN scaling in the era of new devices (nanowires, compound semiconductors, CNT)



Fin