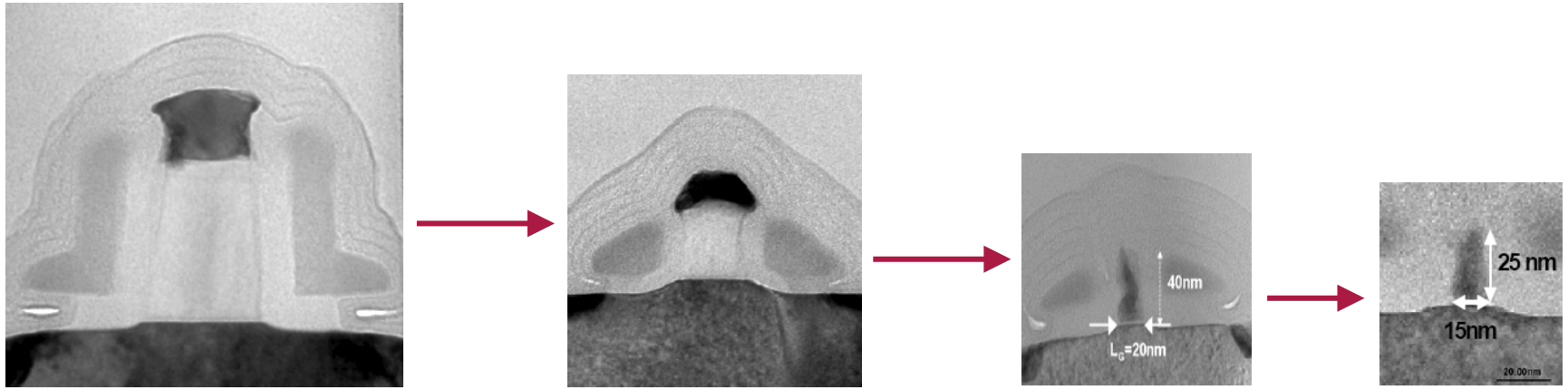
Dependable Design in Nanoscale CMOS Technologies: Challenges and Solutions

Vikas Chandra

ARM R&D

Reliability challenges

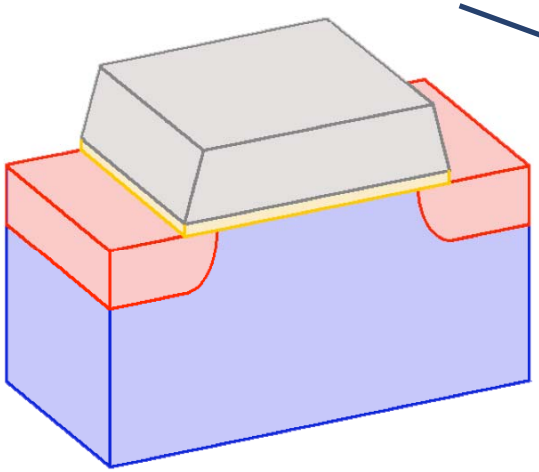


50nm 30nm 20nm 15nm

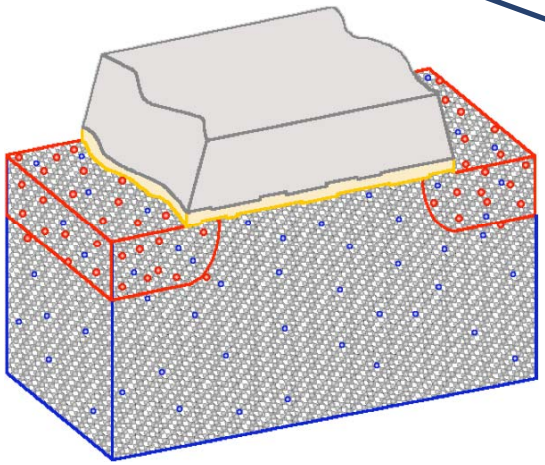
Source: M. Bohr, Intel, IRPS 2003

- Reasons of unreliable transistors
  - Random manufacturing defects
  - Significant increase in variability
  - Increasing electric field
  - Thin gate oxides
  - Voltage, Temperature variations
  - …

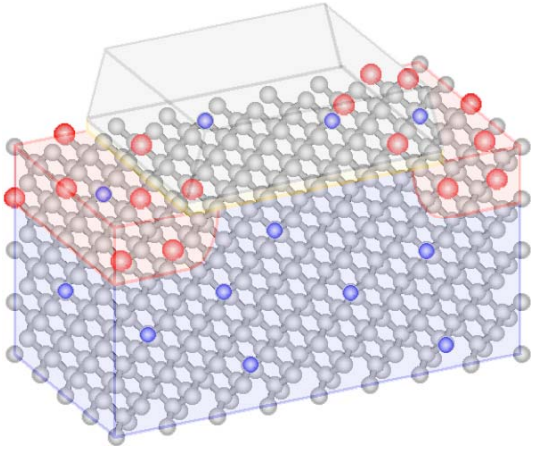
Atomistic scale devices



The simulation
Paradigm now



A 22 nm MOSFET
In production 2010



A 4.2 nm MOSFET
In production 2023

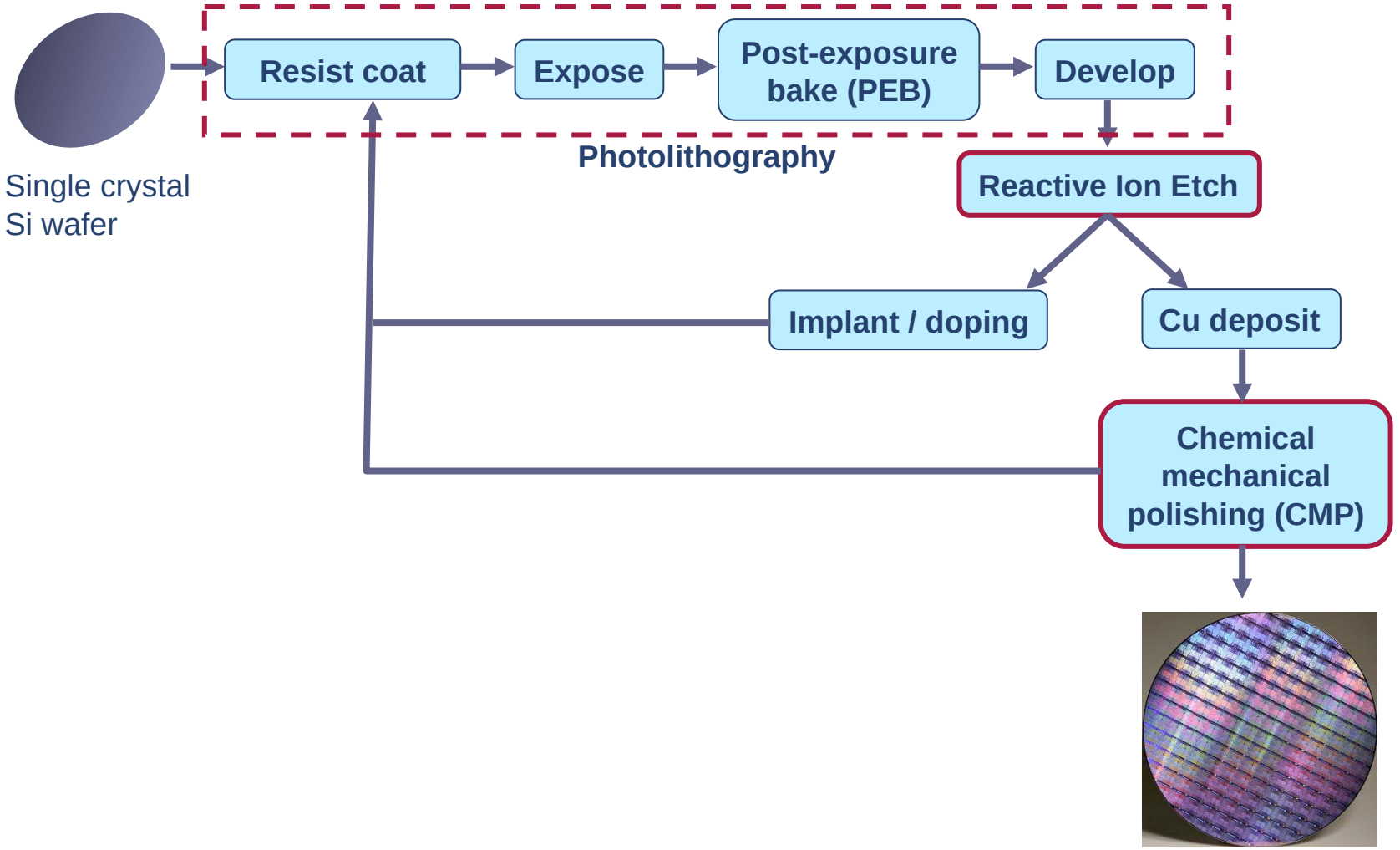
Source: A. Asenov

Types of variability

- Spatial
  - Variations due to the manufacturing process
  - Systematic, process and apparatus induced variations
  - Random variations
- Temporal
  - Mainly due to aging and wearout
  - NBTI
  - Gate oxide degradation
- Dynamic
  - Workload dependent
  - Voltage fluctuation
  - Temperature variation

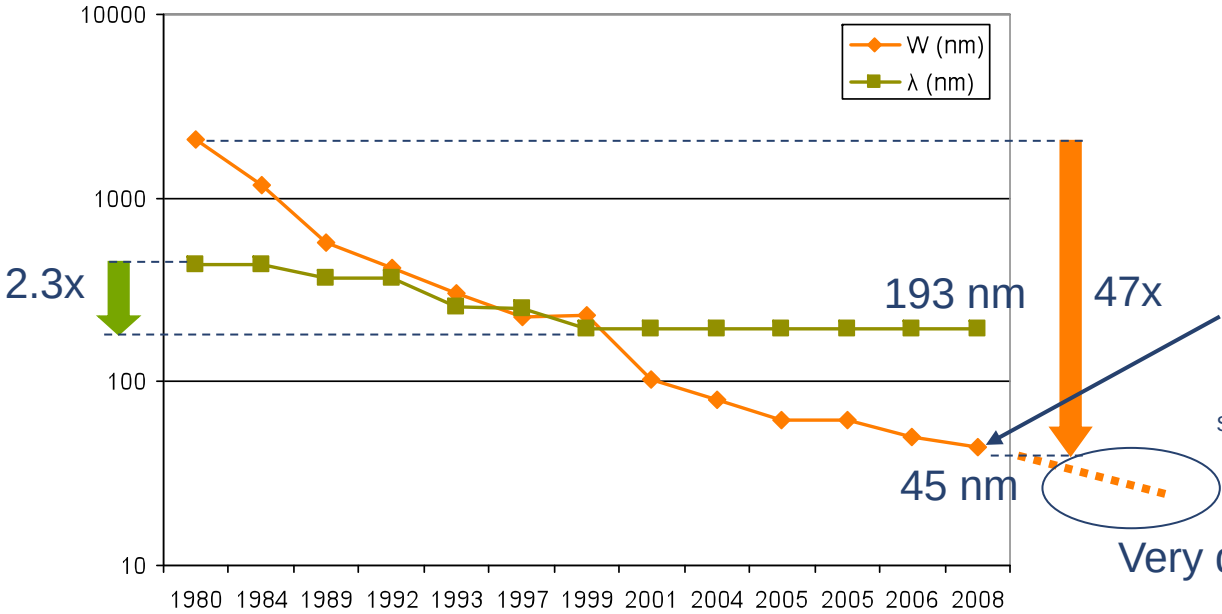
Spatial variations

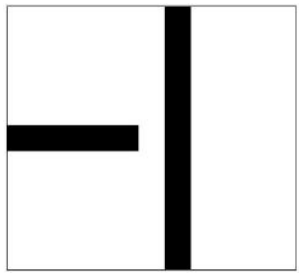
- Simplified Manufacturing Process Steps

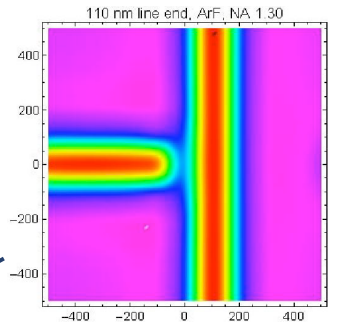


The Lithography Challenge: Reducing Feature Size

- Wavelength scaling has stopped!
  - Glass does not transmit
  - Source not bright enough
  - Reticle/mask too expensive to manufacture
- Deep sub-wavelength lithography
  - Finer lines than the point of a pen!



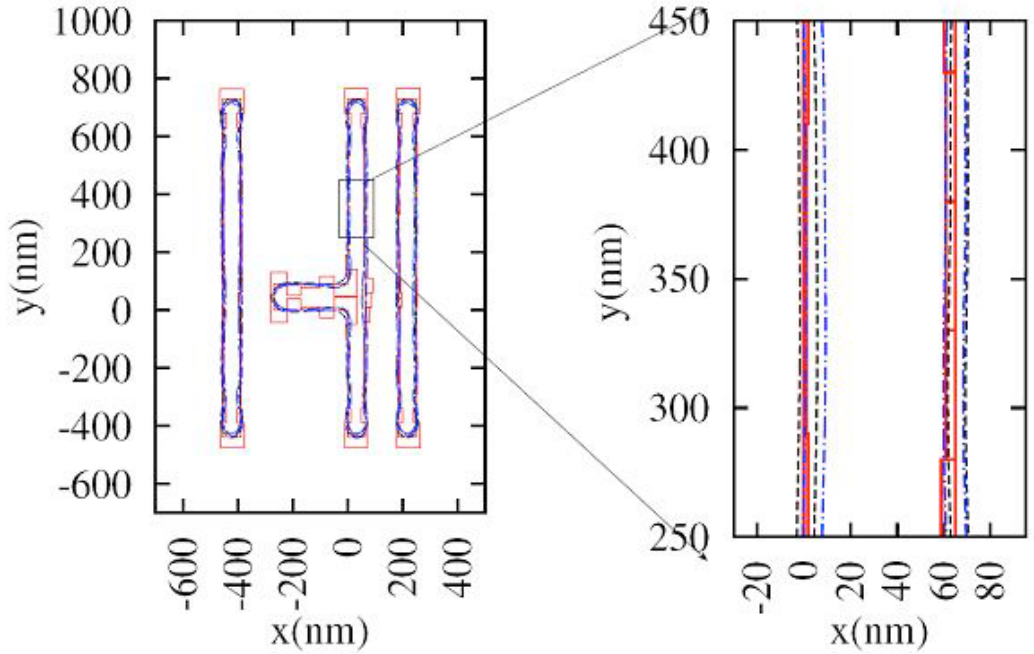




Source: Stephen Renwick, Nikon

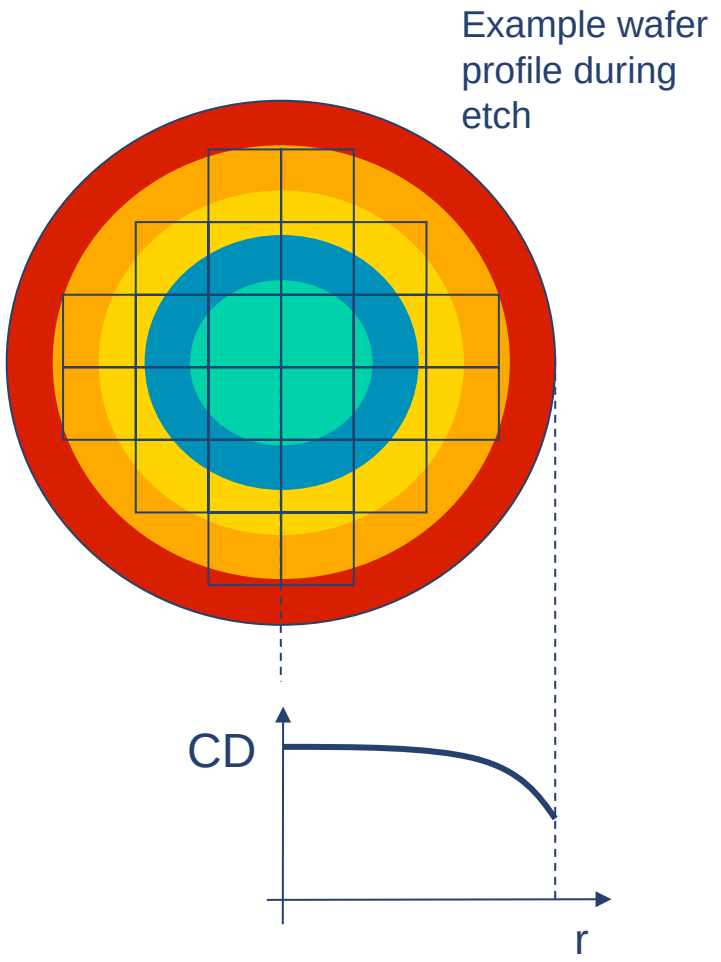
Lithography Variability

- Several sources of variation in lithography
  - Defocus variation
  - Exposure dose (intensity) variation
  - Mask errors
  - Overlay/mask alignment variation



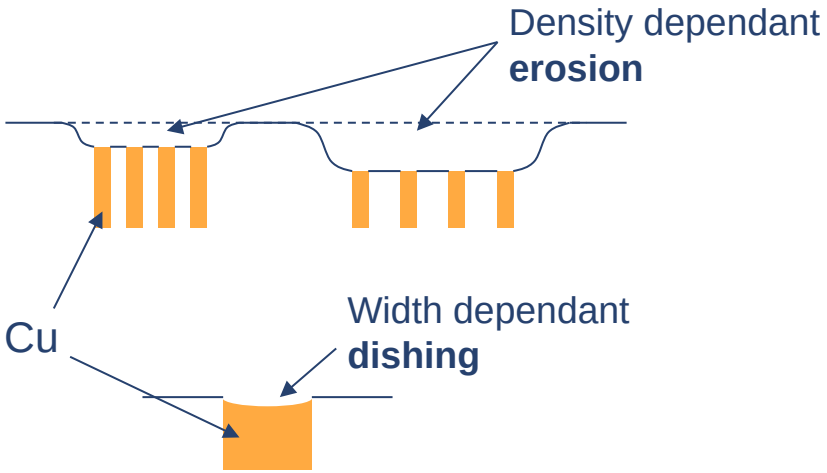
Etch Variability

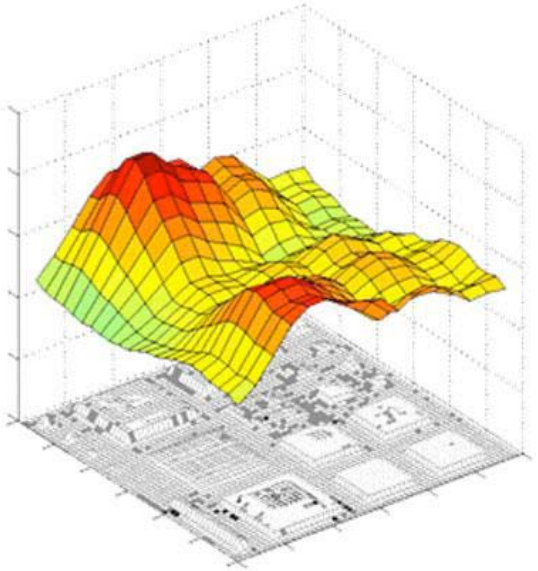
- Etching process has randomness
  - Poisson process for ions hitting the resist
  - Plasma gas flow can have turbulence
- Etch chuck temperature profile is radial – etch rate profile is radial
- Typically CD (linewidth) droops near wafer edge



Source, A. Singhee, IBM

CMP Variability





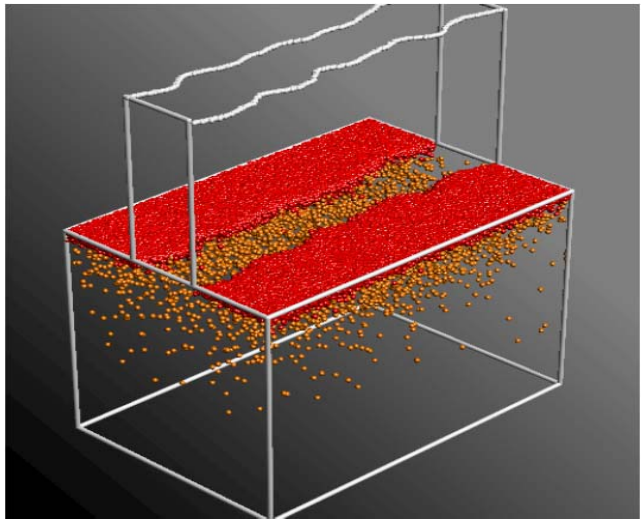
Source: Cadence Design Systems, Inc.

- Material removal depends on wire density and width
- Surface topography changes across the die with Copper density

- Wire resistance and capacitance variation
- Focus error for upper metal layers – wire width errors

Random Dopant Fluctuation

- Doping/implant is a random process
- Number of dopants in channel ~100



M. Hane, et. al., SISPAD 2003

- Dopant count is not repeatable
- Dopant position is not repeatable

- Large variations in threshold voltage

$$\sigma_{V_t} = \left(\frac{\sqrt[4]{4q^3 \varepsilon_{Si} \phi_B}}{2} \right) \cdot \frac{T_{ox}}{\varepsilon_{ox}} \cdot \frac{\sqrt[4]{N}}{\sqrt{W_{eff} L_{eff}}} \propto \frac{1}{\sqrt{W_{eff} L_{eff}}}$$

- ~10-15% $\sigma(V_t)$ at 45 nm and increasing
  - Typical ±3σ tolerance range >= ±30%!

Variability Challenges For Design: ITRS 2007

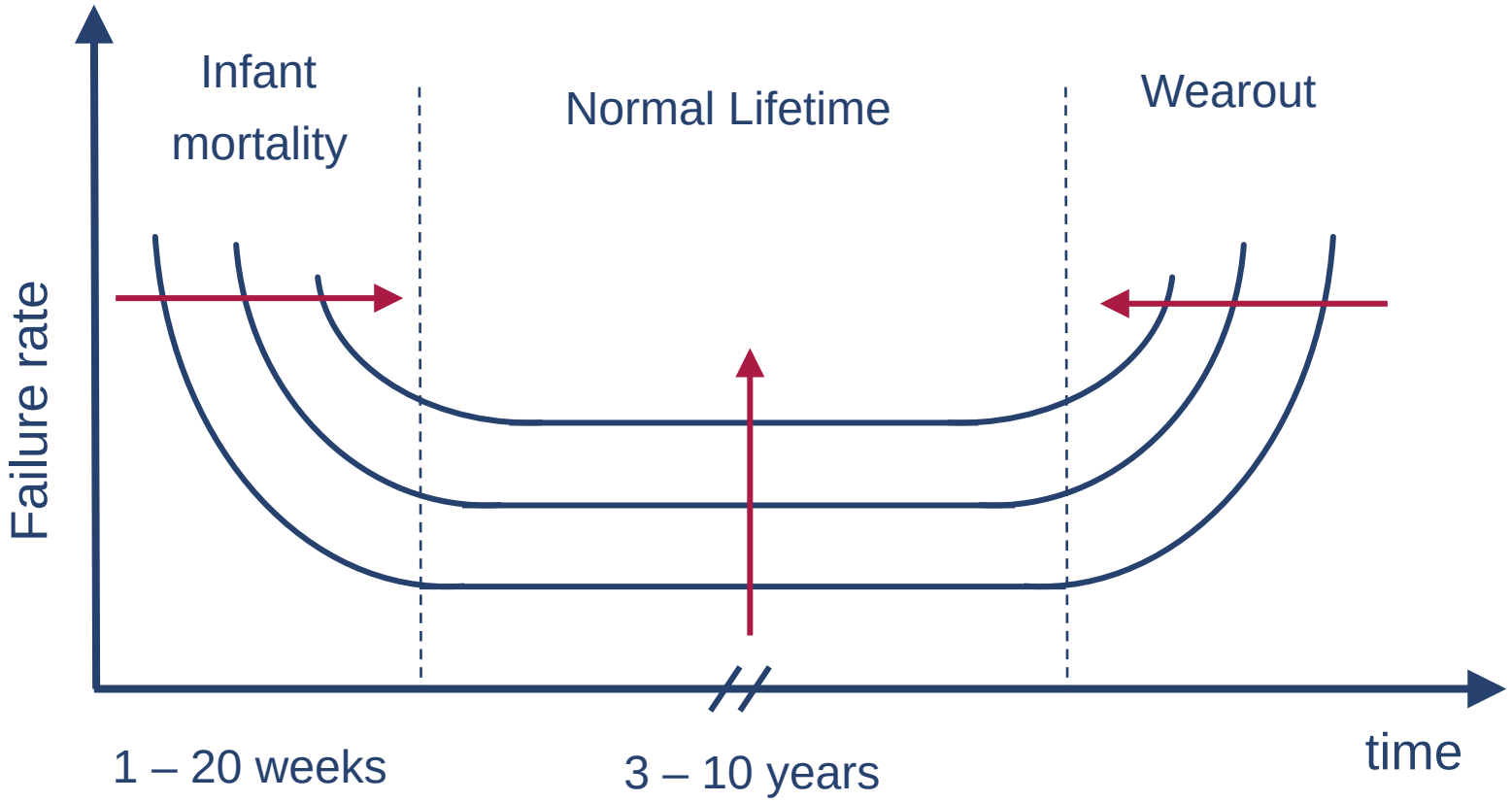
- Lots of RED ahead
- Economics of purely process solution are infeasible
  - Mask cost today up to $100,000
  - Litho tool cost today ~$50,000,000

Table DESN9a Design for Manufacturability Technology Requirements—Near-term Years

Year of Production	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*
Normalized mask cost from public and IDM data	1.0	1.3	1.7	2.3	3.0	3.9	5.1	6.6	8.7
% V_{dd} variability: % variability seen in on-chip circuits	10%	10%	10%	10%	10%	10%	10%	10%	10%
% V_{th} variability: doping variability impact on V_{th}, (minimum size devices, memory)	31%	35%	40%	40%	40%	58%	58%	81%	81%
% V_{th} variability: includes all sources	33%	37%	42%	42%	42%	58%	58%	81%	81%
% V_{th} variability: typical size logic devices, all sources	16%	18%	20%	20%	20%	26%	26%	36%	36%
% CD variability	12%	12%	12%	12%	12%	12%	12%	12%	12%
% circuit performance variability circuit comprising gates and wires	46%	48%	49%	51%	60%	63%	63%	63%	63%
% circuit total power variability circuit comprising gates and wires	56%	57%	63%	68%	72%	76%	80%	84%	88%
% circuit leakage power variability circuit comprising gates and wires	124%	143%	186%	229%	255%	281%	287%	294%	331%

- Need more process and variability-aware design

Temporal variations



- Infant mortality: *Increasing manufacturing defects*
- Normal lifetime: *Increasing transient errors*
- Wearout: *Acceleration of aging phenomena*

Temporal unreliability

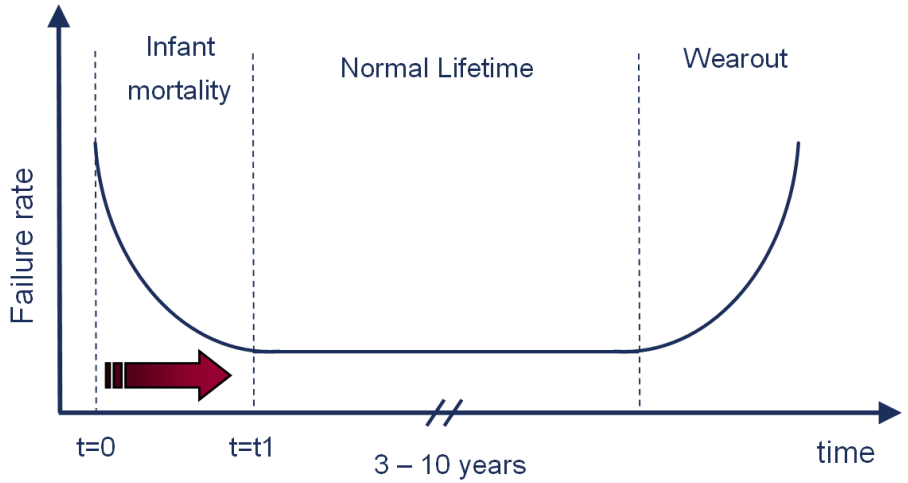
- Infant mortality
  - Marginal parts due to random manufacturing defects
  - Gate-to-source shorts
  - Small opens, poor vias & contacts
  - *Mitigated by Burn-in*
- Normal Lifetime
  - Soft errors in memory and logic
  - *Mitigated by design, architecture and ECC*
- Wearout
  - Transistor degradation (NBTI)
  - Gate oxide breakdown (GBD)
  - *Mitigated by circuit, architecture techniques and overdesign*

Infant mortality

- Also known as Early Life Failures (ELF)
  - Do not affect the circuit initially, but they get worse over time
- Due to manufacturing defects that are random in nature
  - Particles in interlevel oxide creating shorts between metal layers
  - Insulator cracks
  - Thin oxide defects
  - Metallization problems
  - Via defects
  - …
- ELFs follow log-normal failure distribution
  - Short mean lifetime and high sigma
  - Failure rate decreases over time

Burn-in testing

- Burn-in is stress testing for weeding out ELF defects
  - "Age" the circuits just beyond the infant mortality period
  - Weak (defective) parts break due to accelerated aging
  - Employs voltage and temperature to accelerate device aging



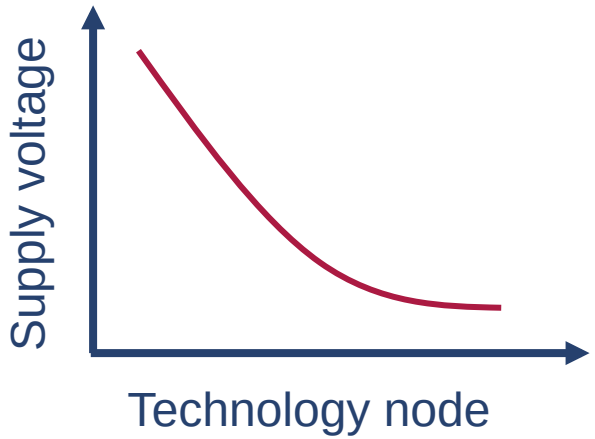
- Stress conditions
  - Voltage stress: Typically 30-40% over nominal Vdd
  - Temperature stress: Typically >120^{o} C
  - Stress time: Typically 10's of hours
    - Decreases as failure rate decreases

Temperature and Voltage stress

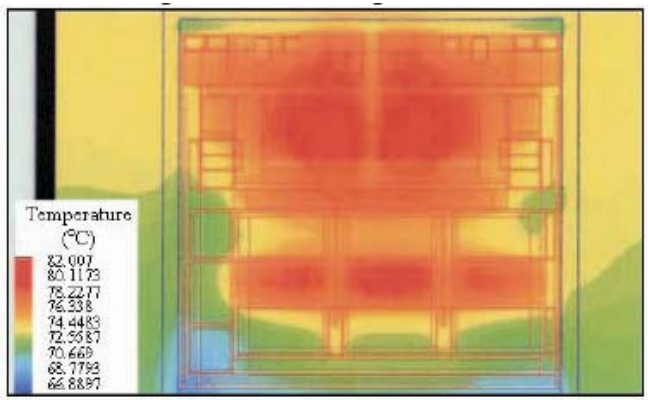
- Temperature acceleration factor $$TAF = e^{\frac{E_a}{K}\left(\frac{1}{T_{stress}} - \frac{1}{T_{use}}\right)}$$
- Voltage acceleration factor $$VAF = He^{\gamma\left(V_{stress} - V_{use}\right)}$$
- TAF targets: electromigration, metallization problems, contact/ via defects etc
- VAF targets: gate oxide defects

VAF and TAF trends

- Supply voltage is saturated
- $\Delta V = V_{stress} - V_{use}$
  - 40% of 3.3V → 1.32V
  - 40% of 1V → 0.4V
- VAF goes down exponentially



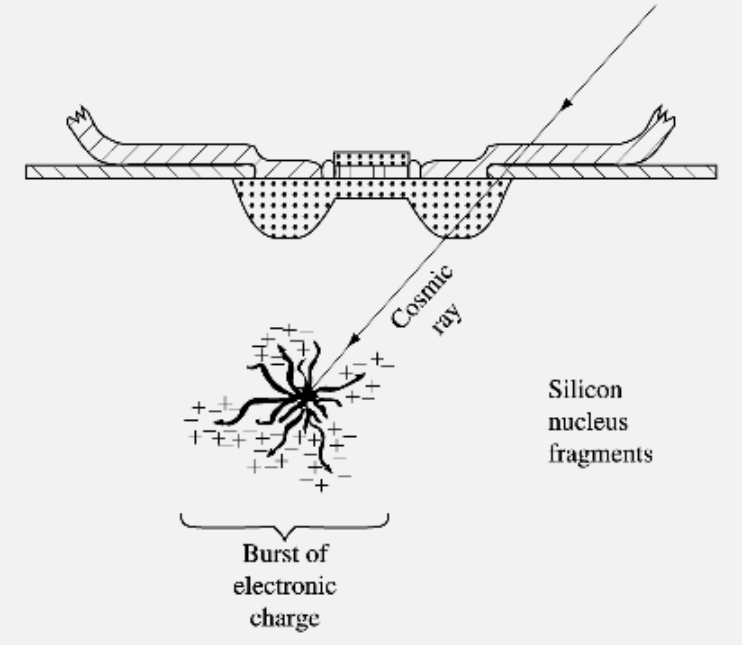
- On chip temperature is going up
- TAF goes down exponentially



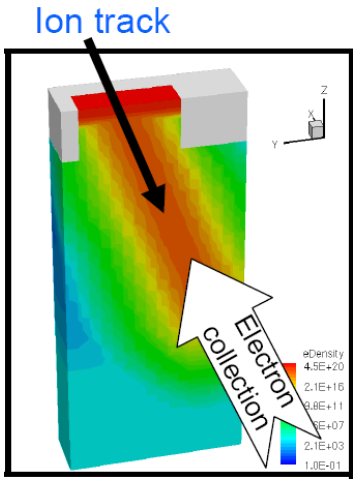
- **Burn-in testing running out of steam?**

Normal lifetime unreliability (Soft errors)

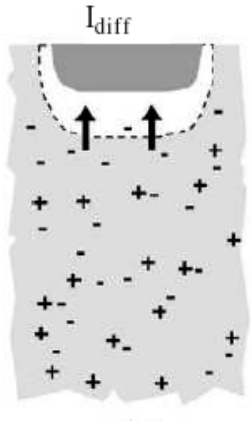
- Mechanism of soft errors due to high energy particles

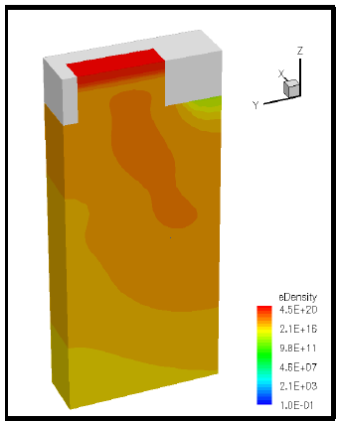


Particle strike creates hole electron pairs



Drift collection



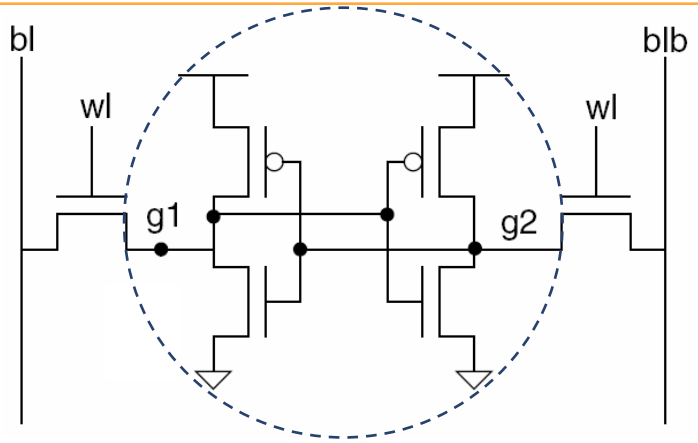


Diffusion collection

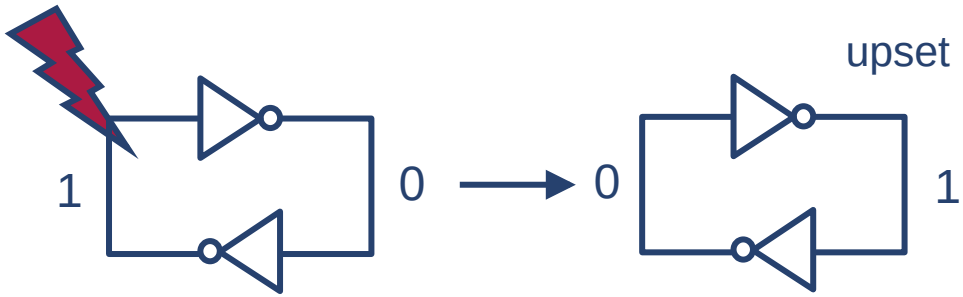
Source: Ziegler, et al., IBM J. of R&D, 1996
Source R. Baumann, *IEEE TDMR*, 2001

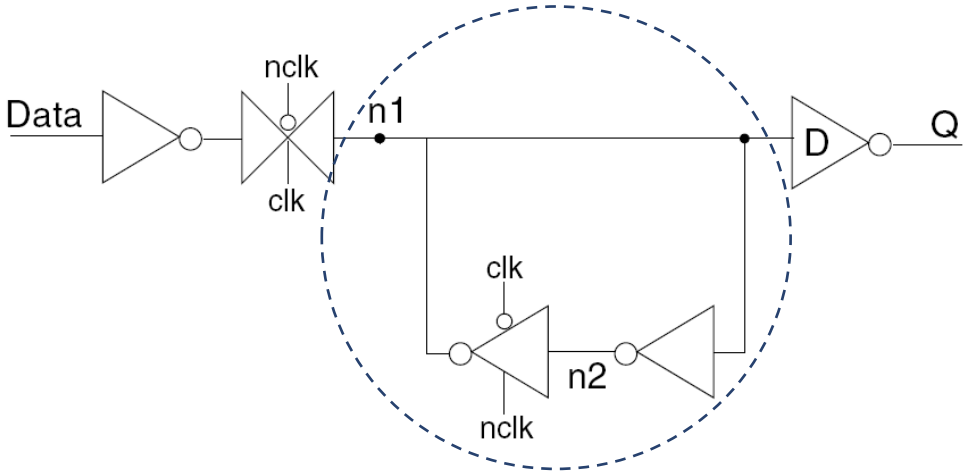
Source: P. Roche, ST, IRPS 2006

Impact on storage logic



6T bit cell



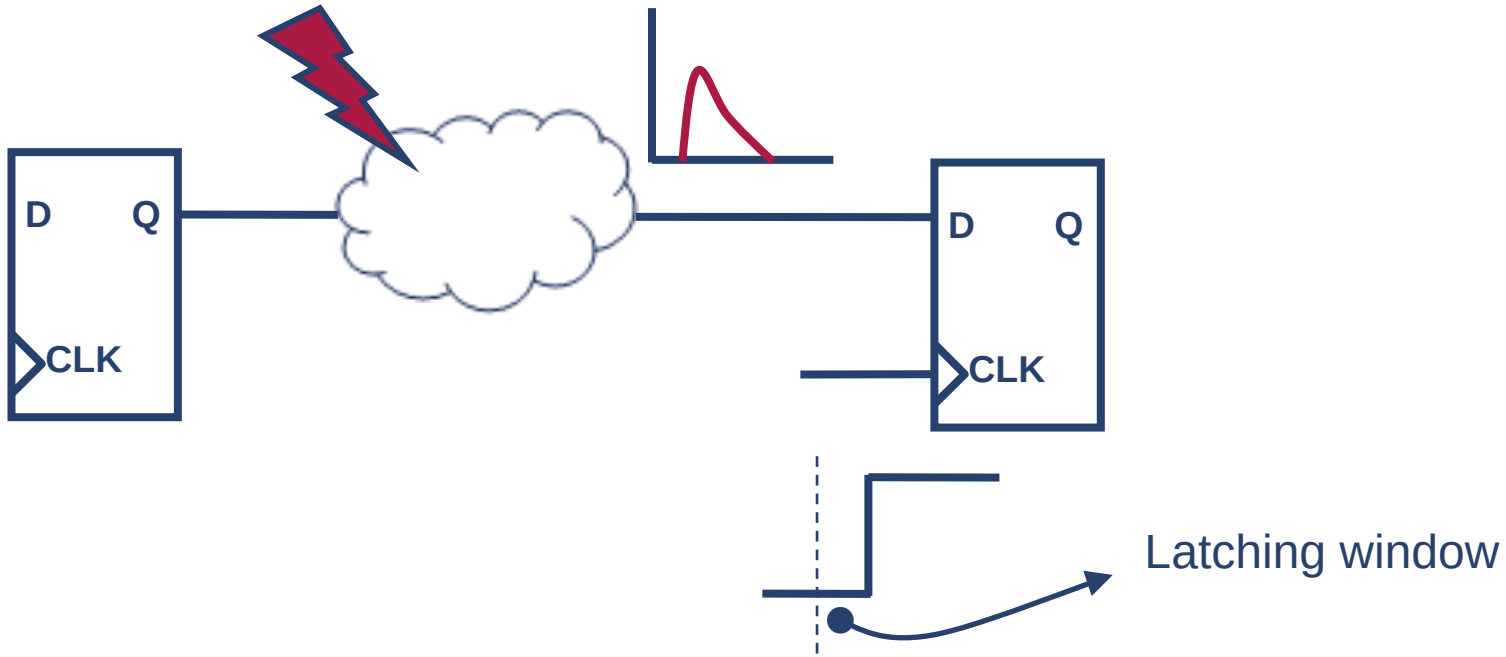


Latch

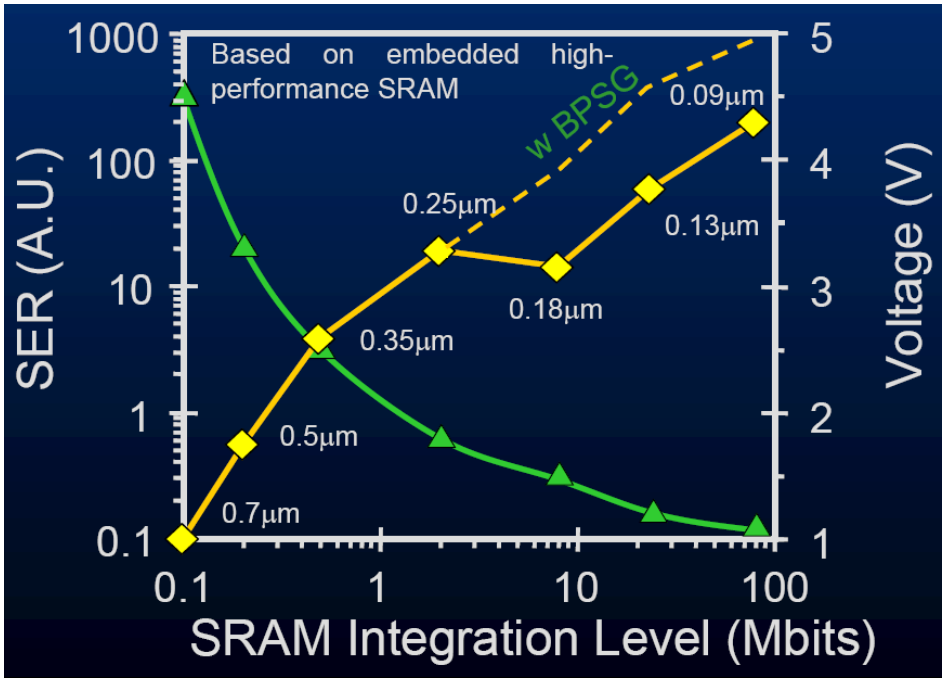
- Particle strike flips the stored value
- The flipped value stays due to regenerative feedback
- Corrupts the state of the system

Impact on combinational logic

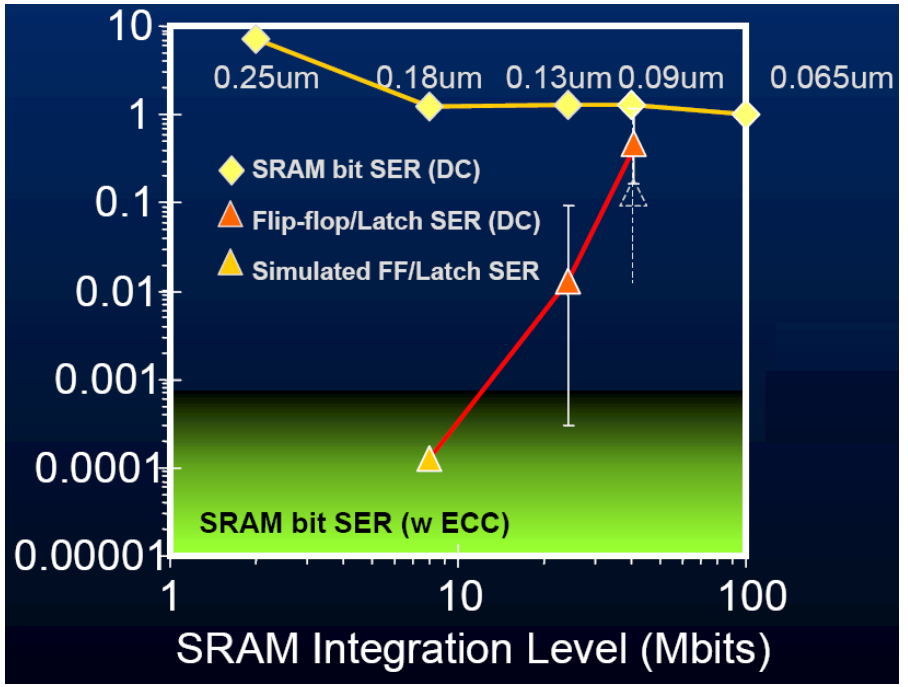
- Causes glitch at gate outputs
- Can be latched if transition happens during latching window
  - Can result in timing failure
  - Errors can be masked by electrical and logical masking
- Decreasing cycle time exacerbates this problem



Soft error trends



SRAM Trends



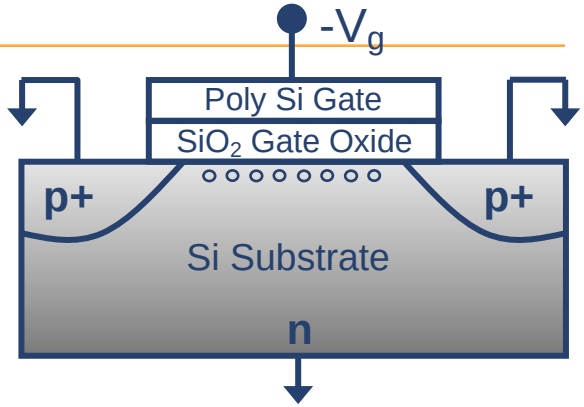
Latch Trends

- Substantial increase in soft error susceptibility with technology scaling!

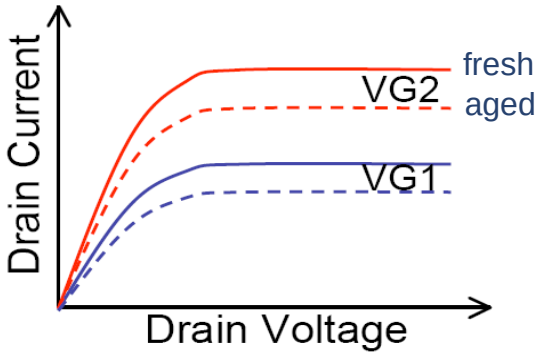
Source: R. Baumann, TI, SemaTech 2004

Wearout - NBTI basics

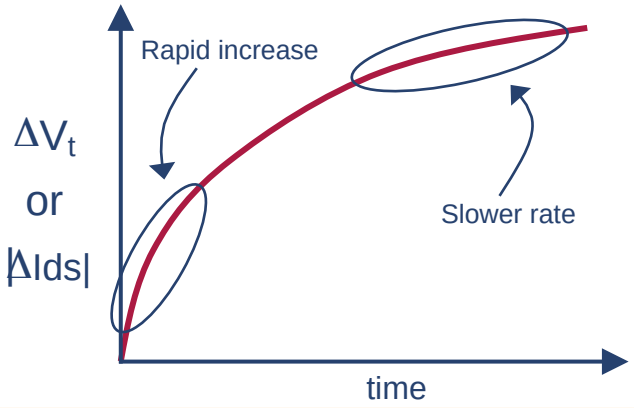
- NBTI stands for Negative Bias Temperature Instability
  - Degradation in PMOS performance over device lifetime
  - Due to traps at Si-SiO_2 interface
  - Instability refers to gradual shift in transistor parameters with time



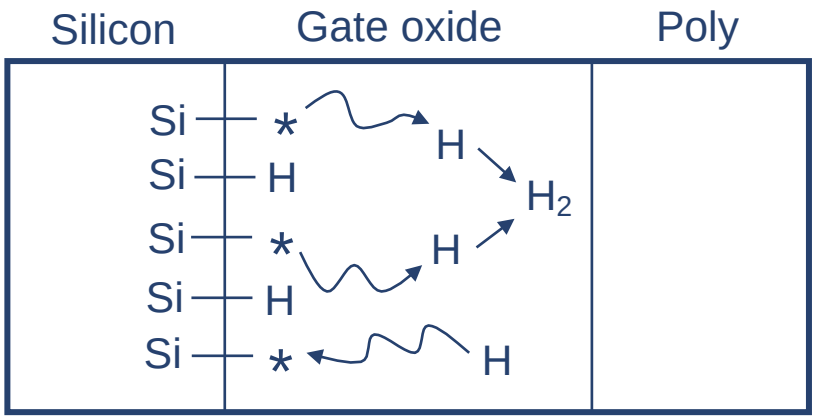
- Impact on transistor performance
  - V_t ↑
  - I_{ds}, g_m, I_{off} ↓



- Temporal behavior of NBTI induced aging

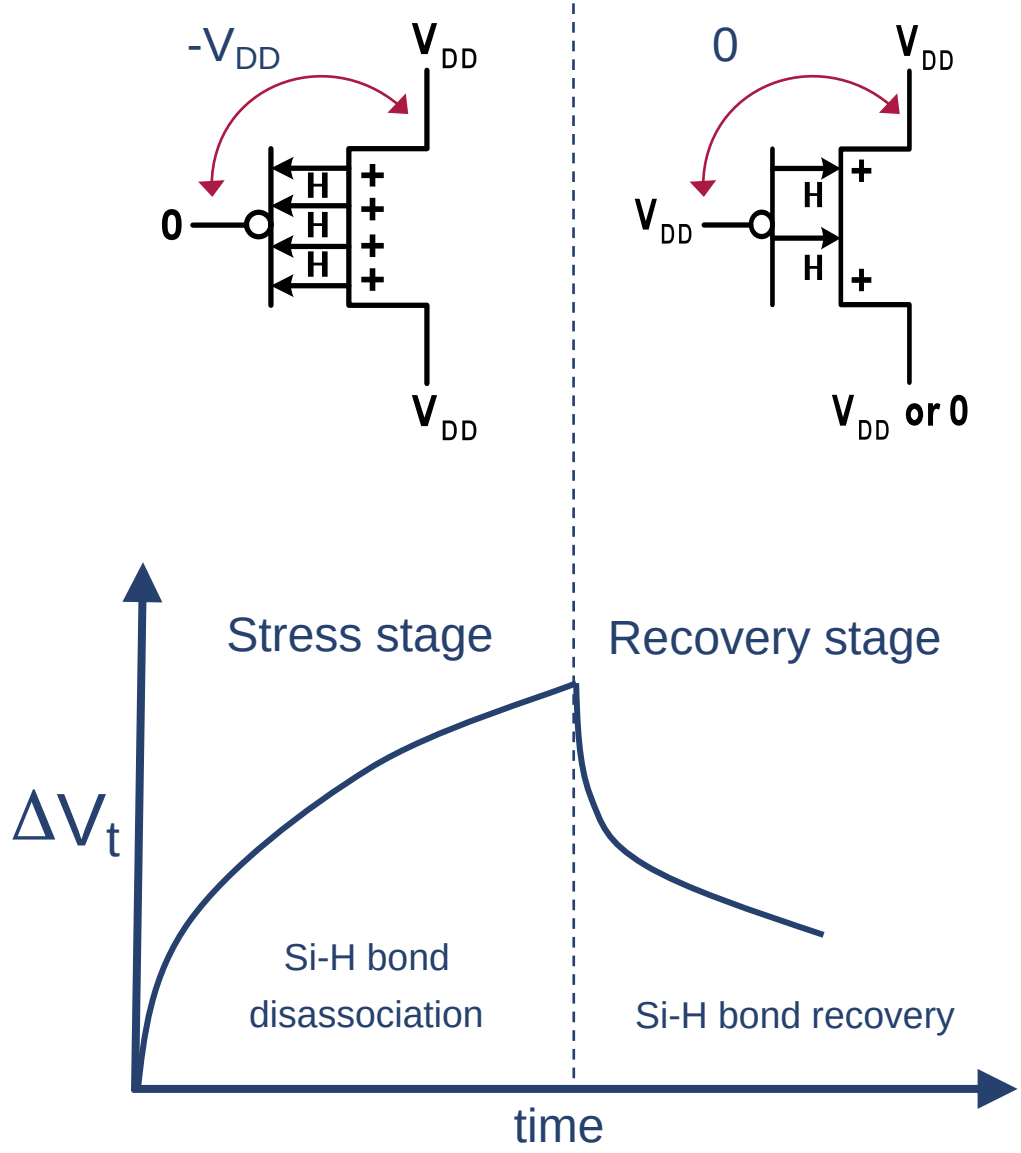


NBTI : Degradation – Recovery



Negative Bias: Si-H bond disassociation

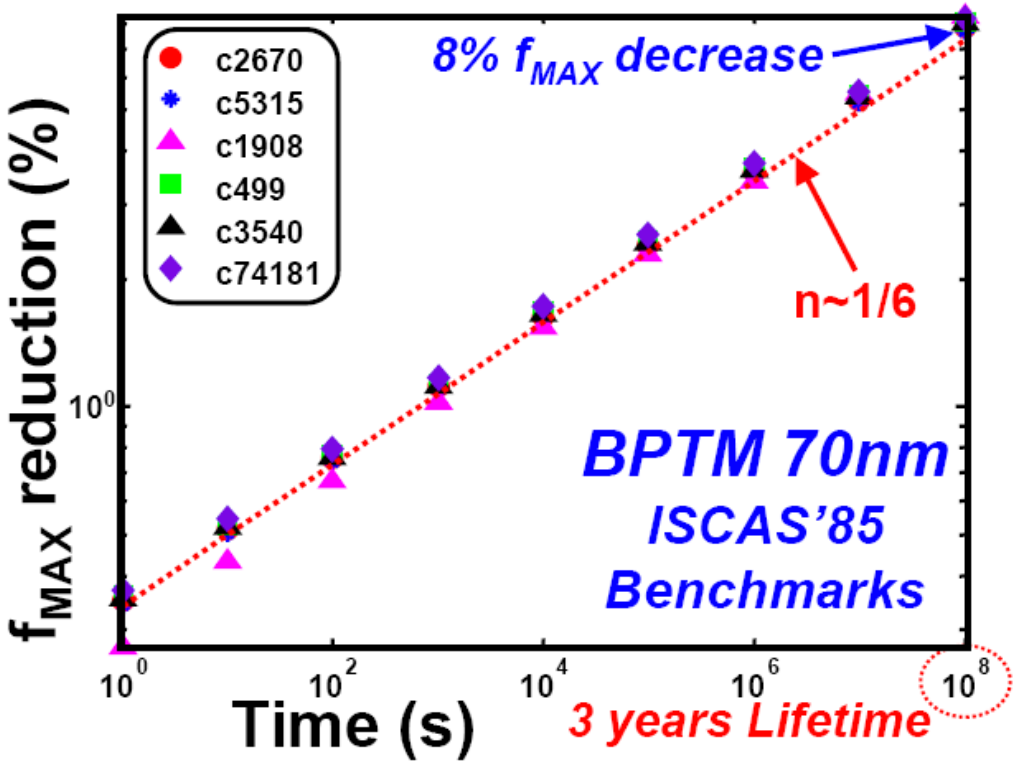
Zero Bias: Si-H bond recovery



Impact on logic circuits

- Temporal V_t shift in PMOS affects critical performance metrics
- Combinational circuits
  - F_{max} decreases ↓
  - Timing failure as circuits age
- Storage cells (SRAM, latch)
  - Static Noise Margin ↓
  - Read and write stability ↓
  - Parametric yield loss

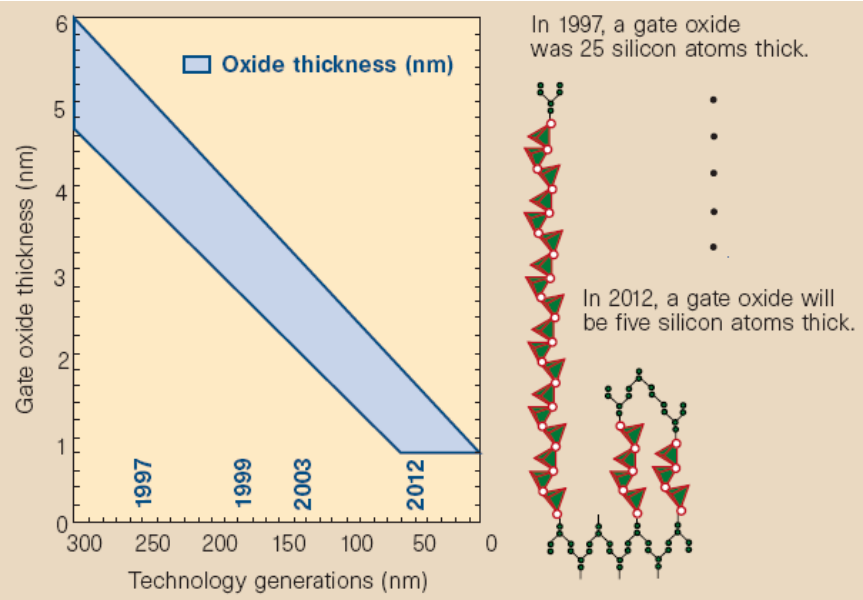
Circuit degradation



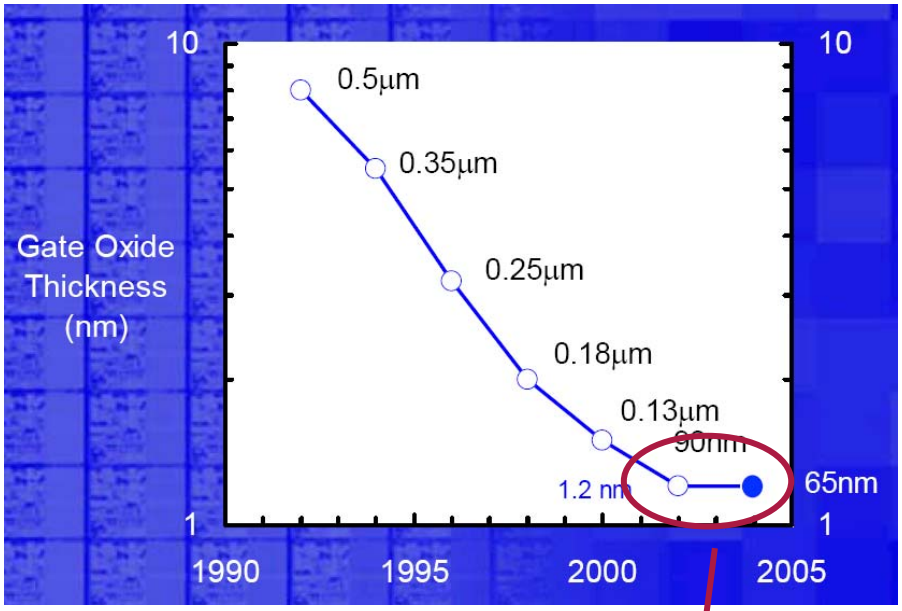
Source: K. Kang, IRPS, 2007

- Average degradation of ~8% in 3 years
- Degradation more dominant for PMOS dominated designs
- Complex circuits seem to degrade less

Gate oxide scaling trend

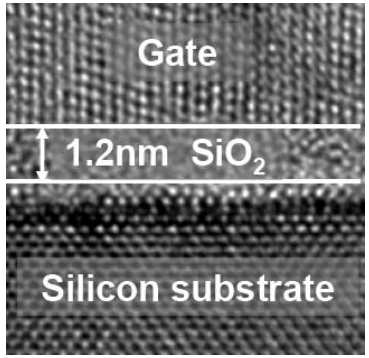


Source: Nature, June 1999

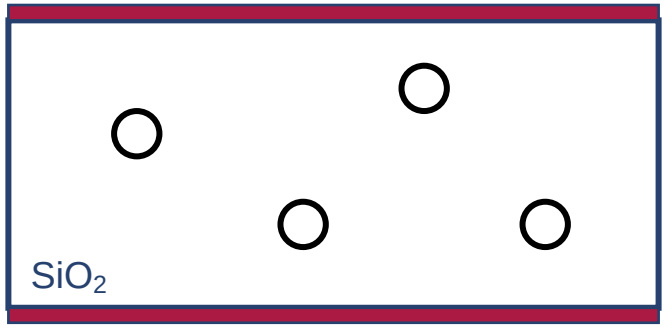


Source: Intel, 2005

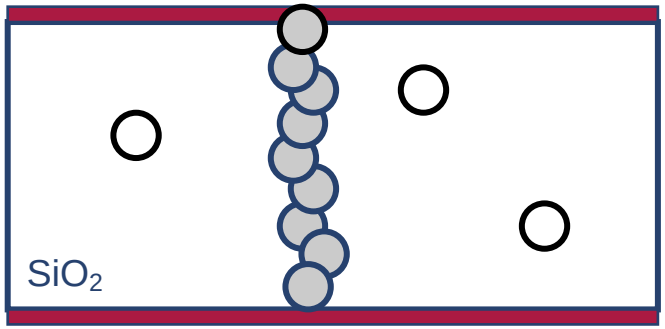
- To reduce power, Vdd is scaled
  - t_{ox} is reduced to reduce V_t
  - Performance increases, as well as leakage
- t_{ox} scaling has hit a plateau
  - Leakage, reliability…



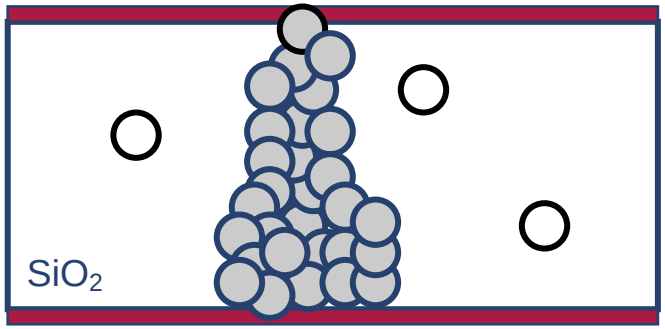
Gate oxide degradation



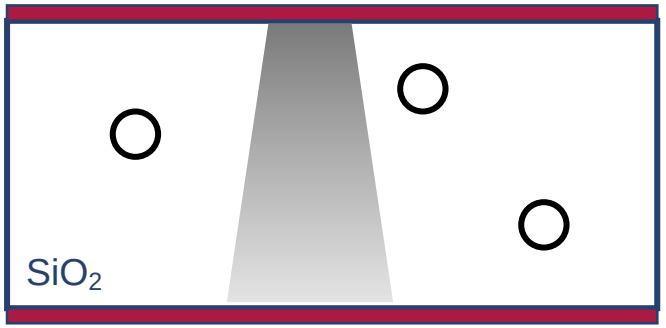
- Traps start to form in the Gate Oxide
  - Non overlapping
  - Do not conduct



- As more and more traps are created
  - Traps start to overlap
  - Conduction Path is created
  - Soft breakdown (SBD)

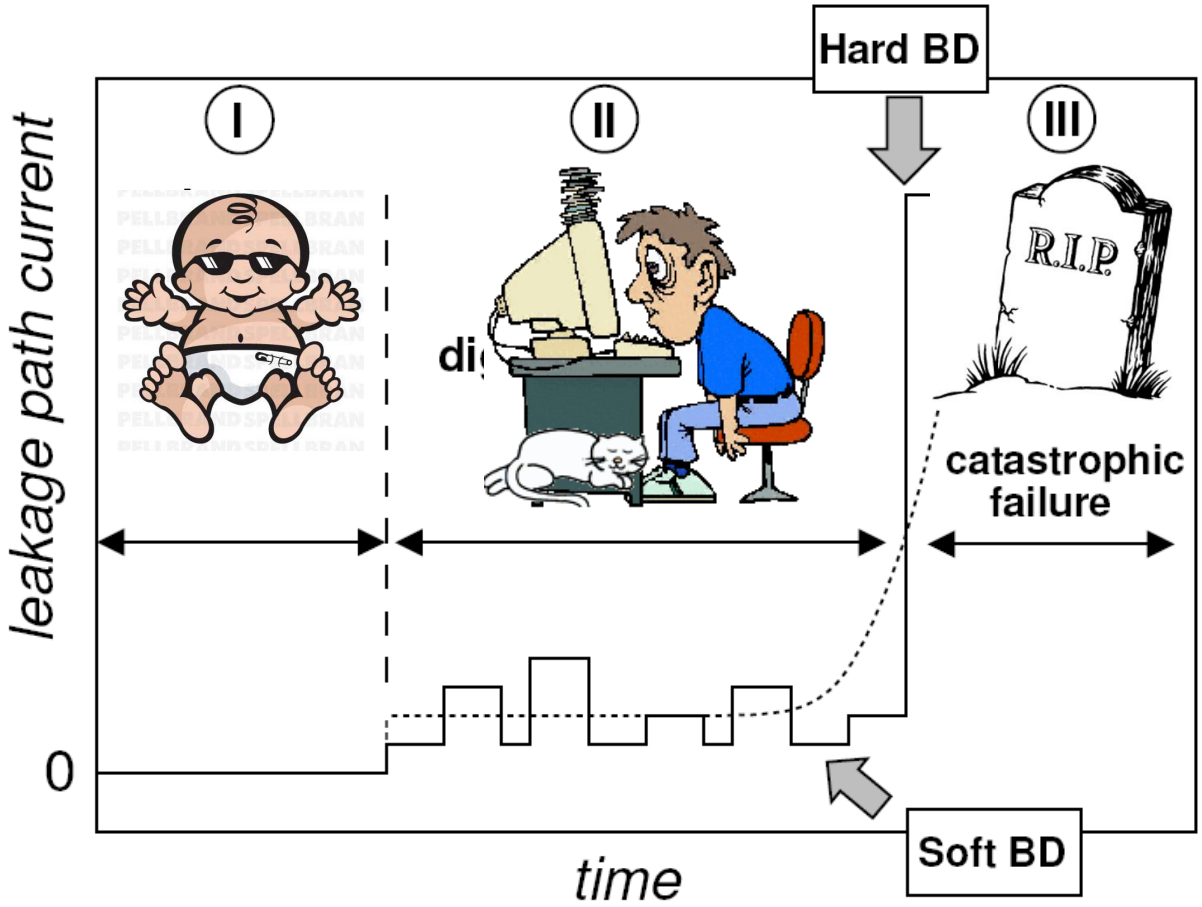


- **Thermal Damage**
  - Conduction leads to heat
  - Heat leads to thermal damage
  - Thermal Damage leads to Traps



- **Hard Breakdown**
  - Silicon in the breakdown spots melts
  - Oxygen is released
  - Silicon Filament is formed from Gate to Substrate (Hard Breakdown)

Temporal oxide degradation

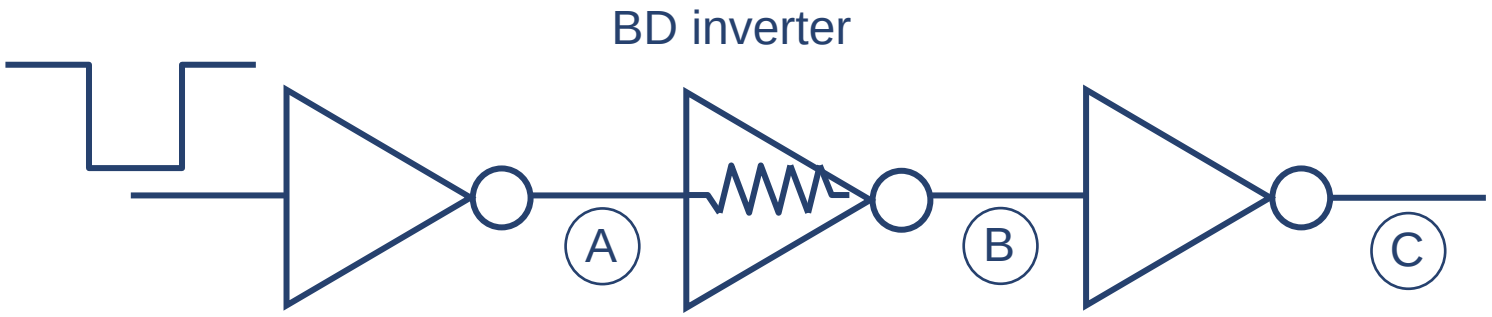


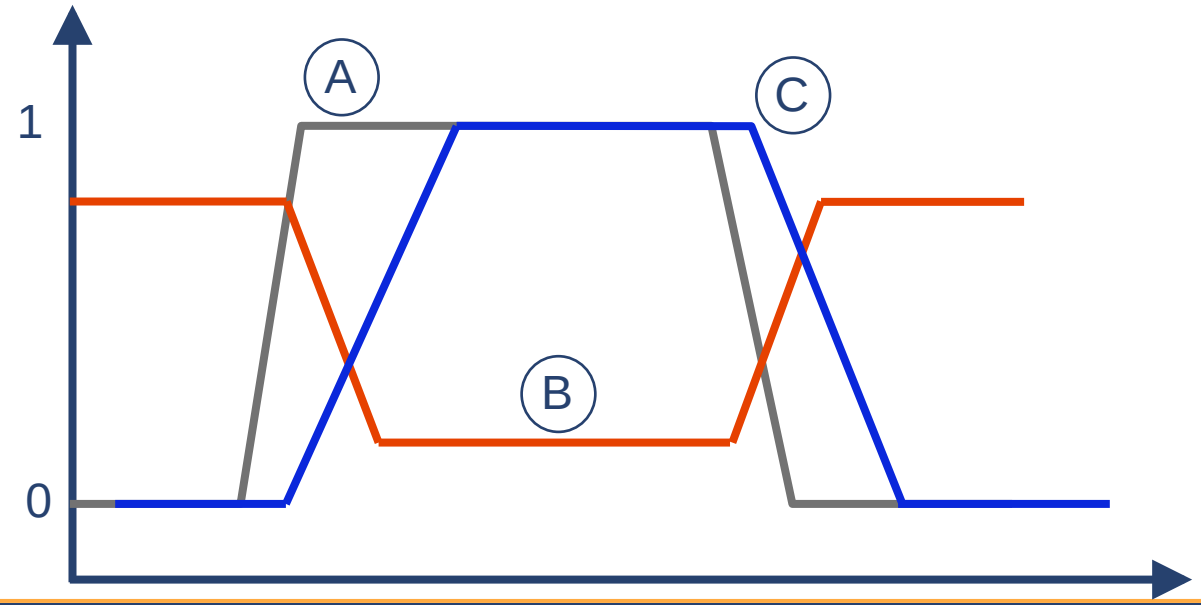
- Gate leakage fluctuates as the gate oxide degrades

Source: H. Wang et al, IEEE TDMR, 2007

Design Characteristic – Digital logic

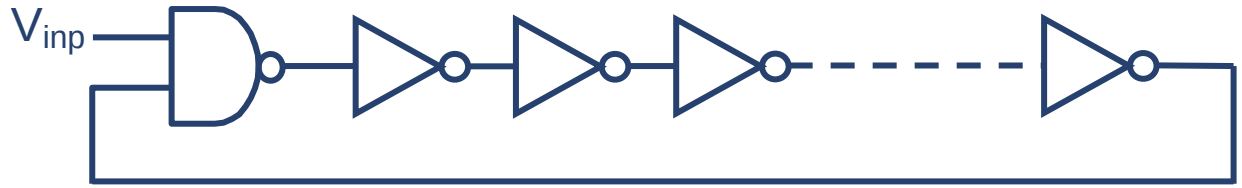
- CMOS logic inherently acts as noise rejecter

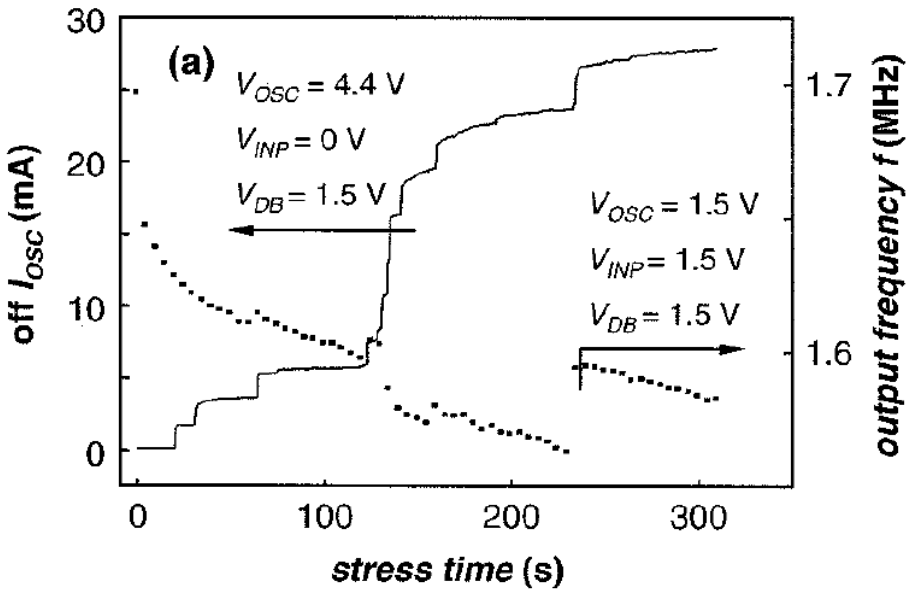


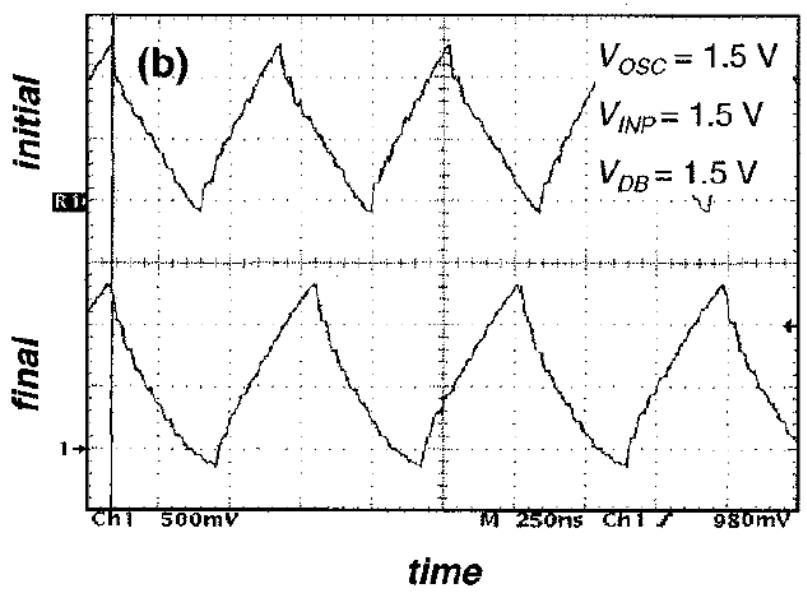


Design Characteristic – Digital logic

- Ring oscillators





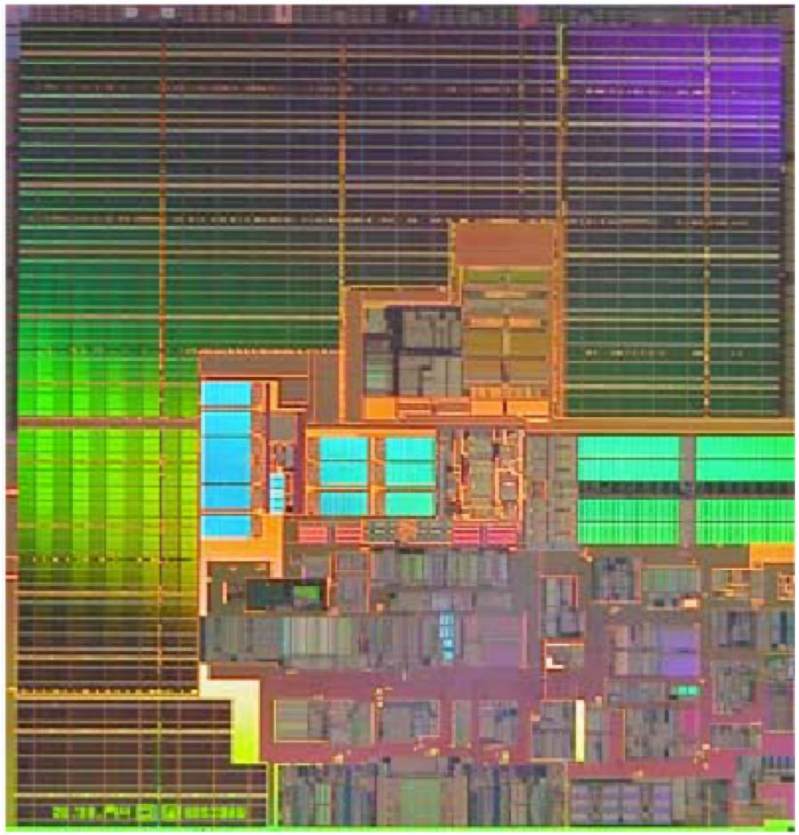


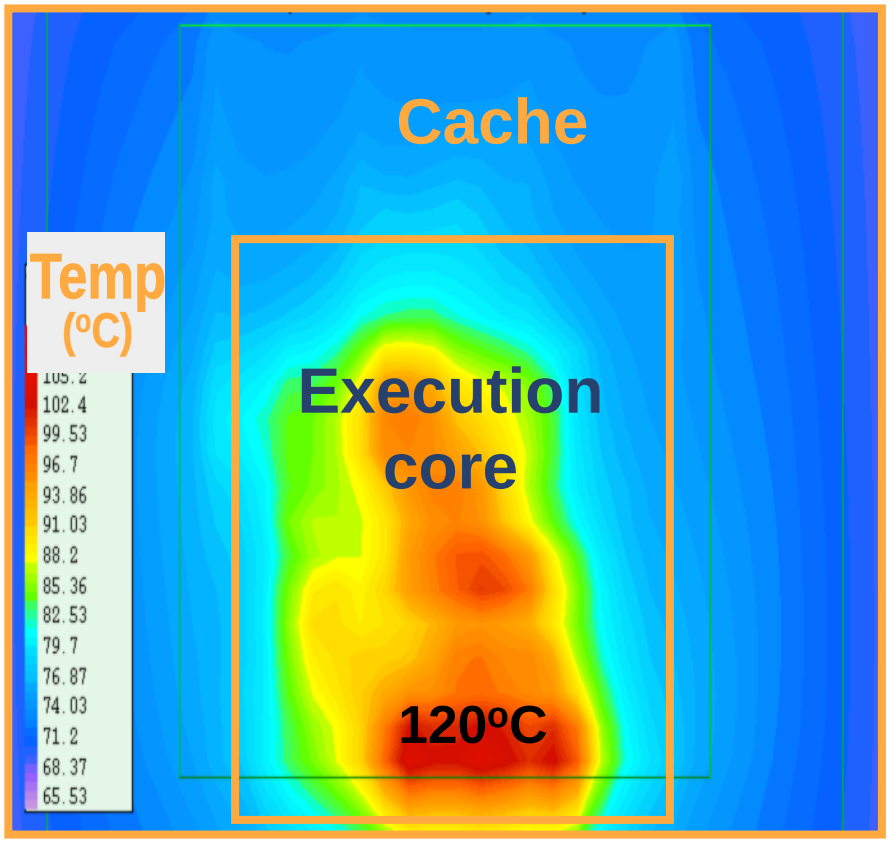
- 41 stage ring oscillator
  - Leakage current goes up after successive breakdowns
  - Still functional after multiple breakdowns
  - Oscillation frequency slows down

Source: B. Kaczer, Trans on Electron Devices, Mar 2002

Dynamic variations: Temperature

- Thermal map – 1.5 GHz Itanium map

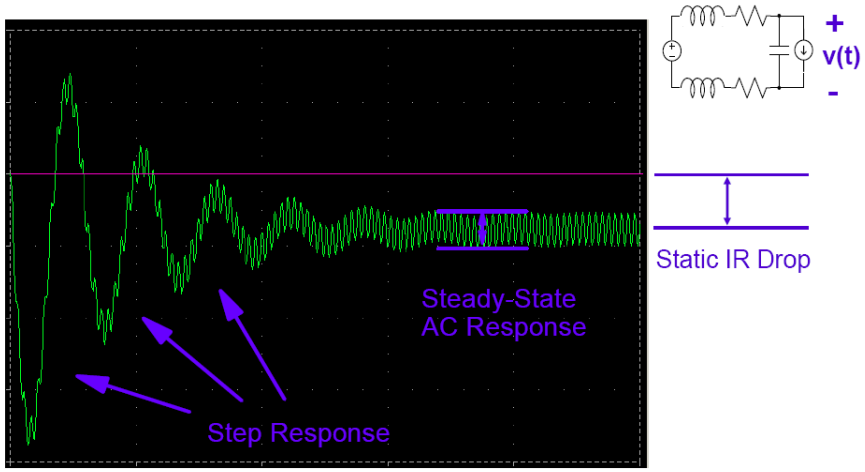




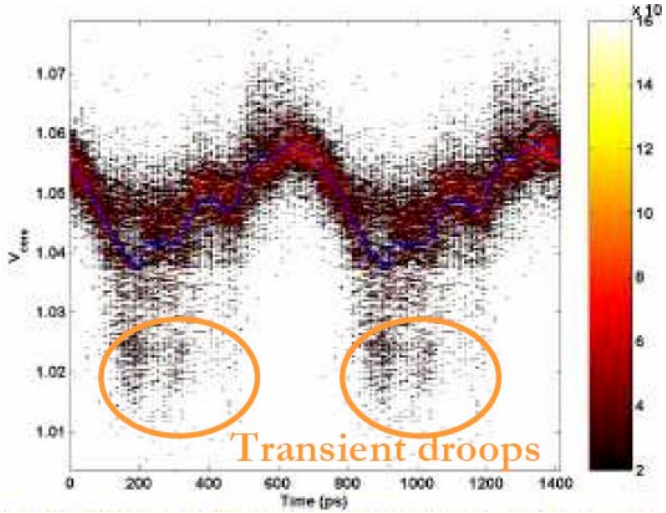
[Source: Intel Corporation and Prof. V. Oklobdzija]

Dynamic variations: Voltage, Power

Voltage variations

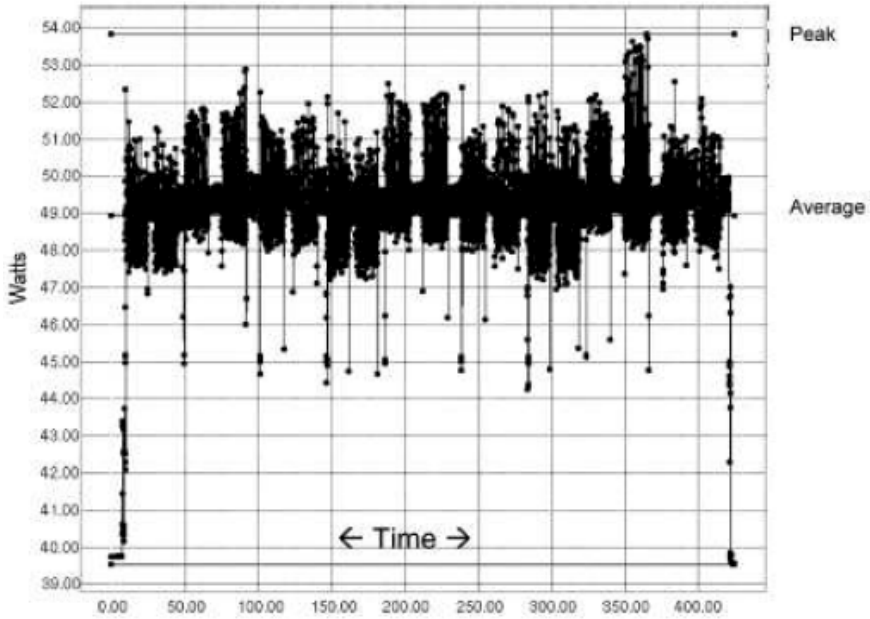


Source: D. Hathaway, SLIP 2005



Source: Naffziger *et al*, JSSC 2006

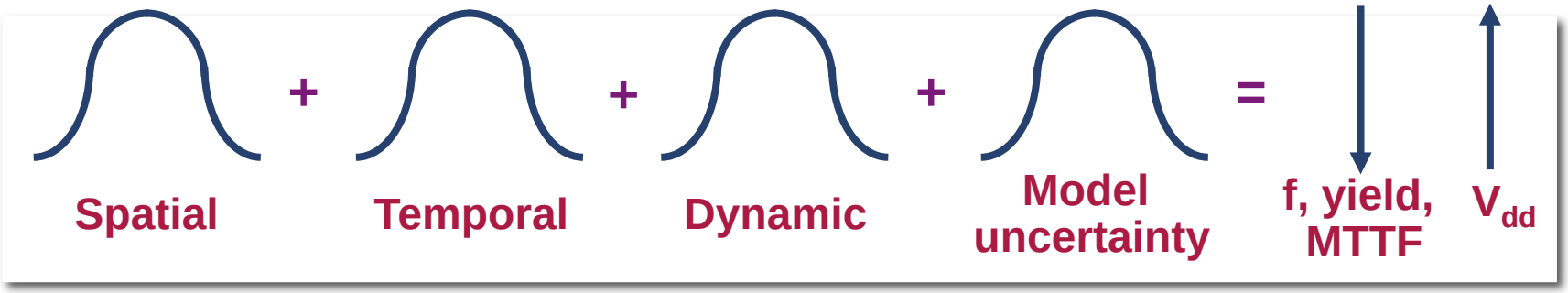
Power variations



Source: Naffziger *et al*, JSSC 2006

Design with margins

- Variability leads to margins



- Uncertainly leads to overheads in performance and power
  - Increasing intra- and inter-chip variation with process scaling
  - Sources: lithography, manufacturing (dopant fluctuation, pattern density effects), crosstalk noise, temperature variation, aging…
- Worst-case scenarios are highly improbable
  - Significant gain for circuits optimized for the common case

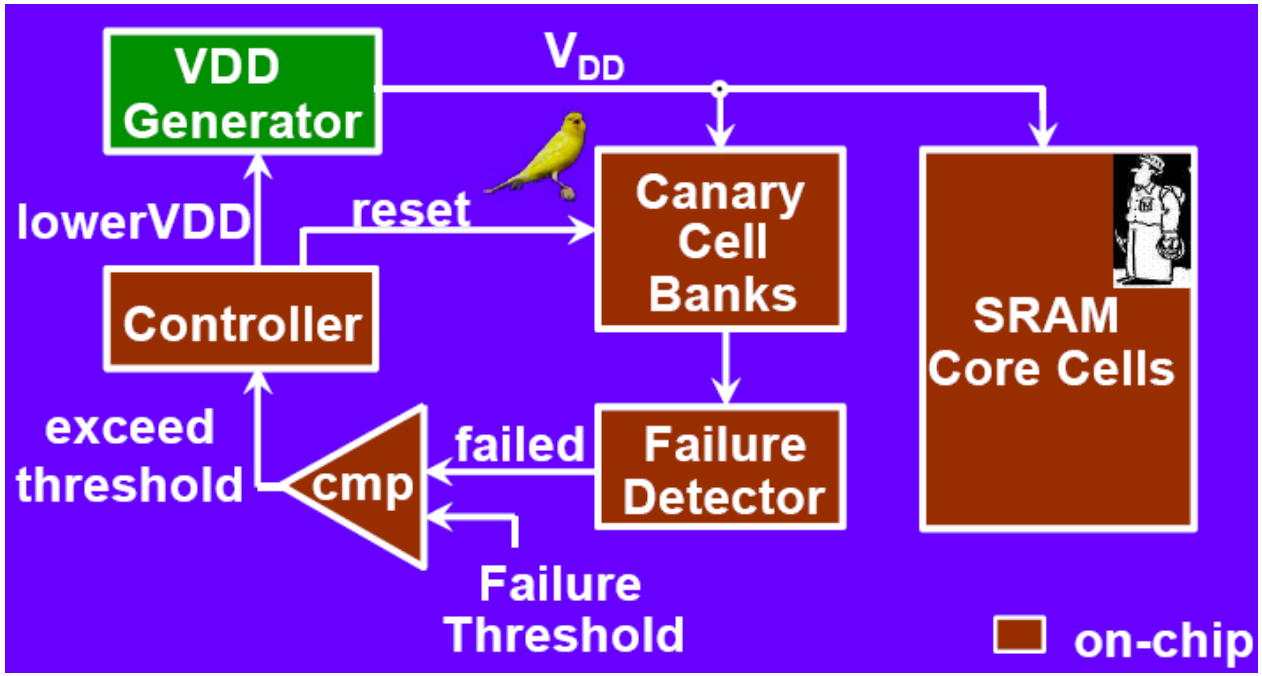
Adaptive designs

- Reduce guardbands due to variations
  - Spatial, temporal and dynamic
- Respond to variations by dynamic adaptation
- Three components required for adaptability
  - Failure prediction
  - Failure detection
  - Failure recovery

Failure prediction

- Predict the errors before they affect design functionality
  - More applicable to slow changing variations
- Adapt by changing frequency and/or voltage
- Possible ways to detects errors
  - Canary circuits: These circuits fail before the actual design fails
  - Pre-sampling: Sample the same data at different points in time
  - Aging monitor: Detect a transition in a guardband period

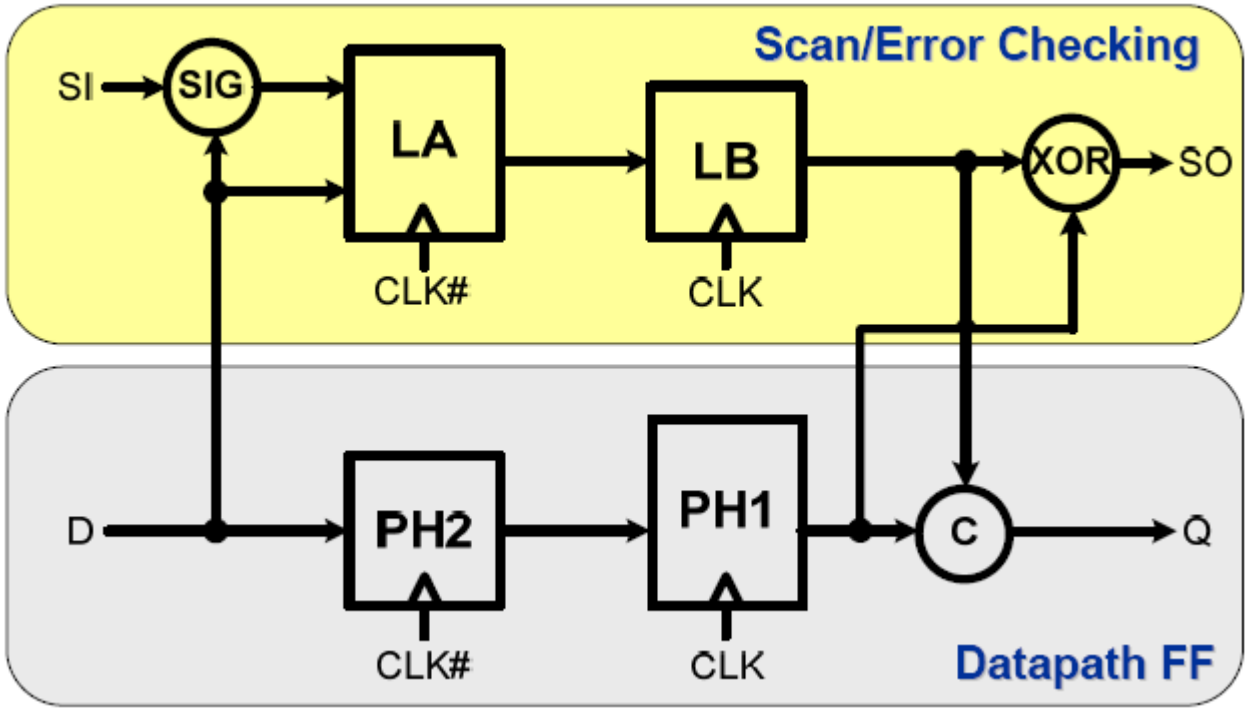
Failure prediction: Canary circuits



- SRAM example for choosing minimum Data Retention Voltage (DRV)
  - Use replica bitcells (canary bitcells) inspired by canary birds
  - Use Canary bitcells in closed-loop VDD scaling

Source: J. Wang et al, CICC 2007

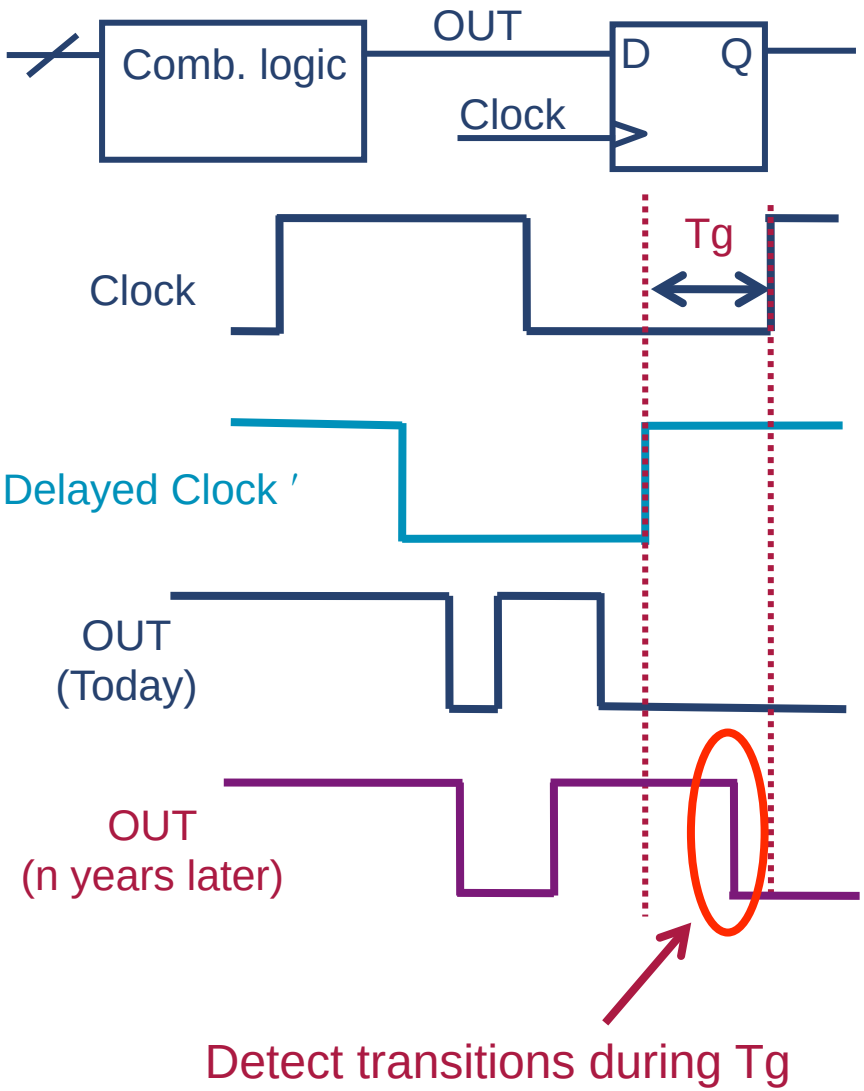
Failure prediction: Pre-sampling

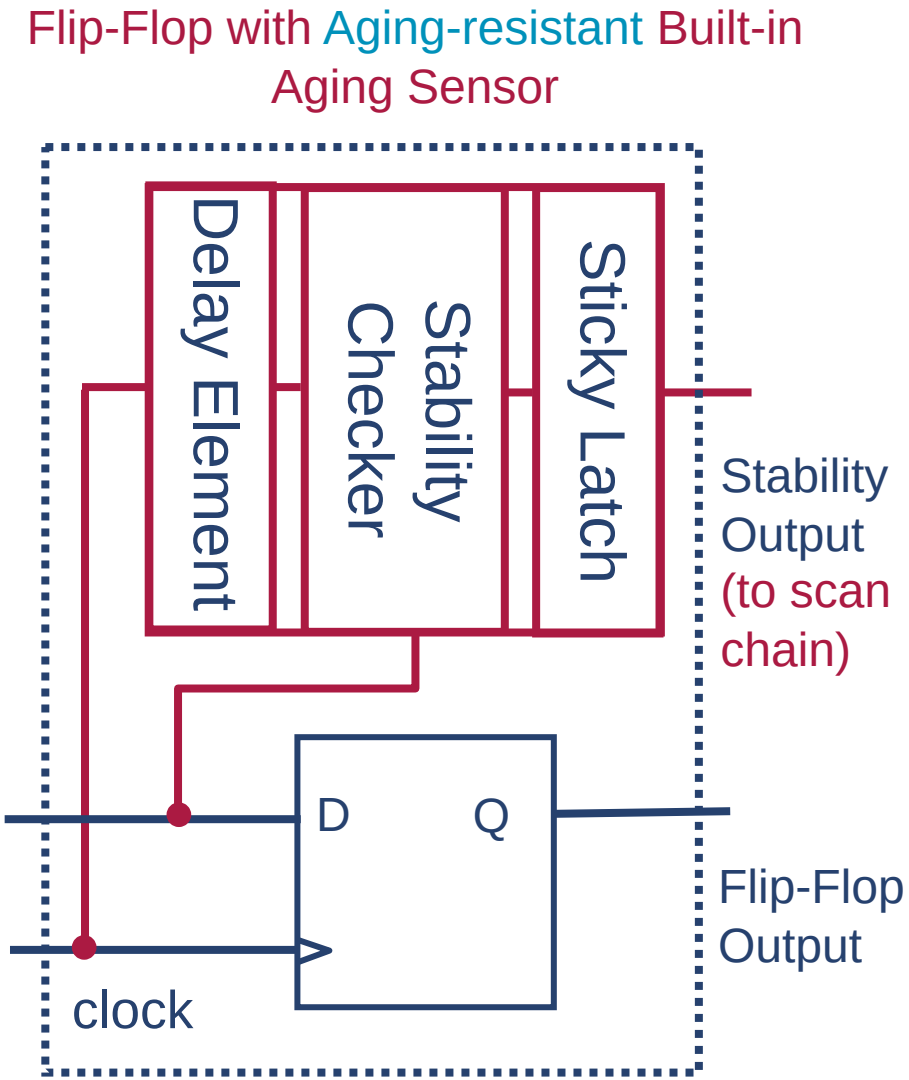


- Key features of AVERA cell
  - Scan circuit re-used for error checking and analysis
  - Circuit timing degradation detected by pre-sampling LA-LB
  - C-element for error correction

Source: M. Zhang, IOLTS '07

Failure prediction: Aging detector





Source: Agarwal, Mitra et al, VTS '07

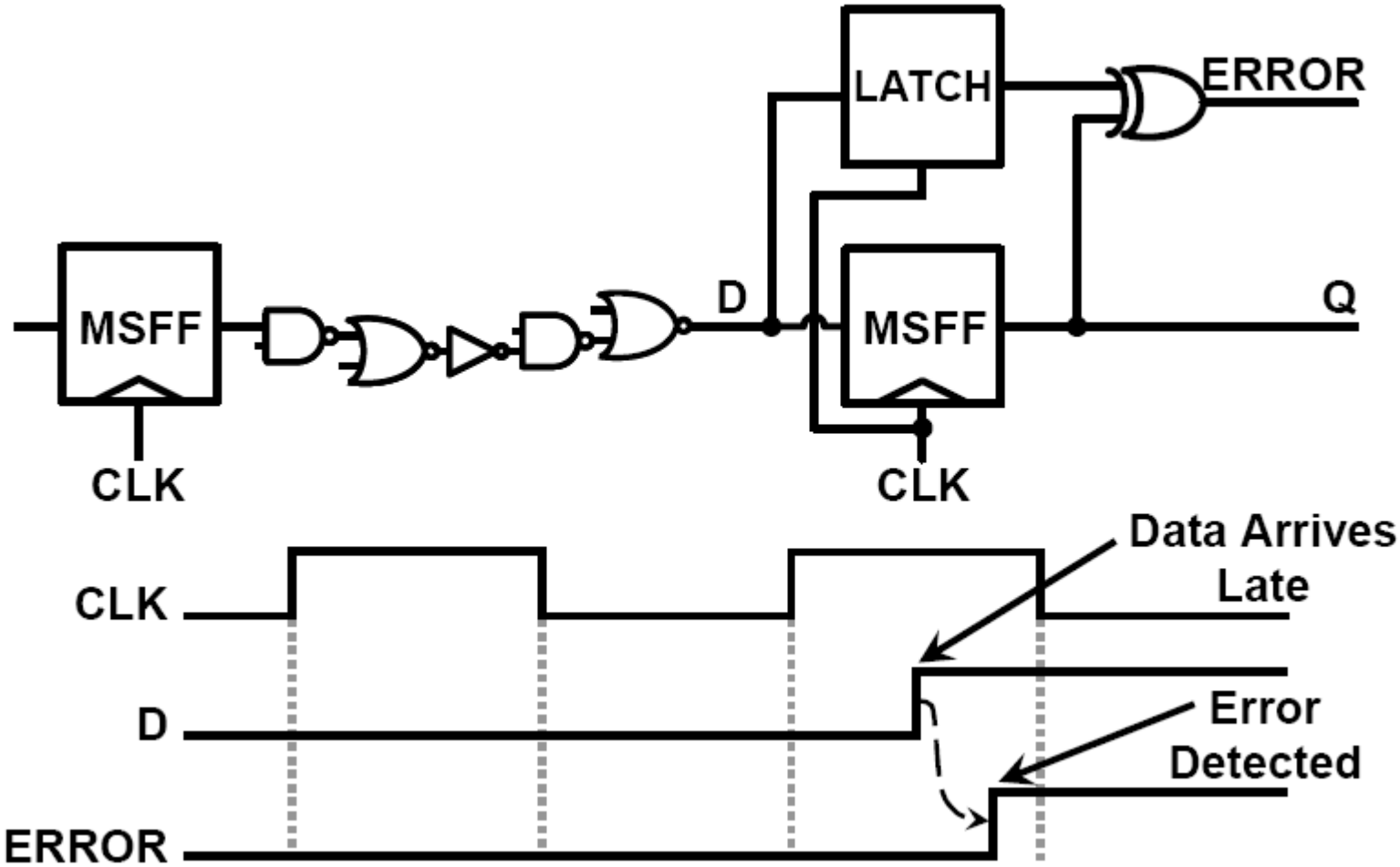
Failure detection

- Detect errors which affect functionality
  - Fast changing errors
    - Soft errors, transient errors due to voltage glitch etc.
  - Slow changing errors
    - Aging induce timing errors
    - Temperature induce timing errors

- Faliure detection methods
  - Software
  - Redundancy
  - Coding
  - Path-level delay fault detection
  - …

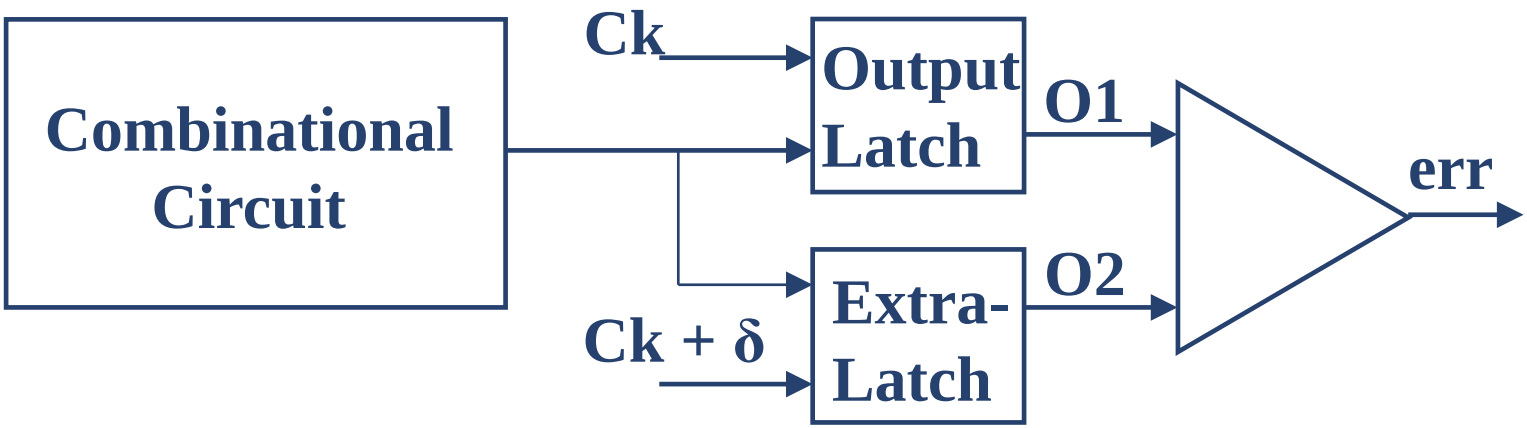
Failure detection

- Error detection by double sampling



Source: D. Ernst et al, Micro, 2003

Error-detection techniques for transient fault detection

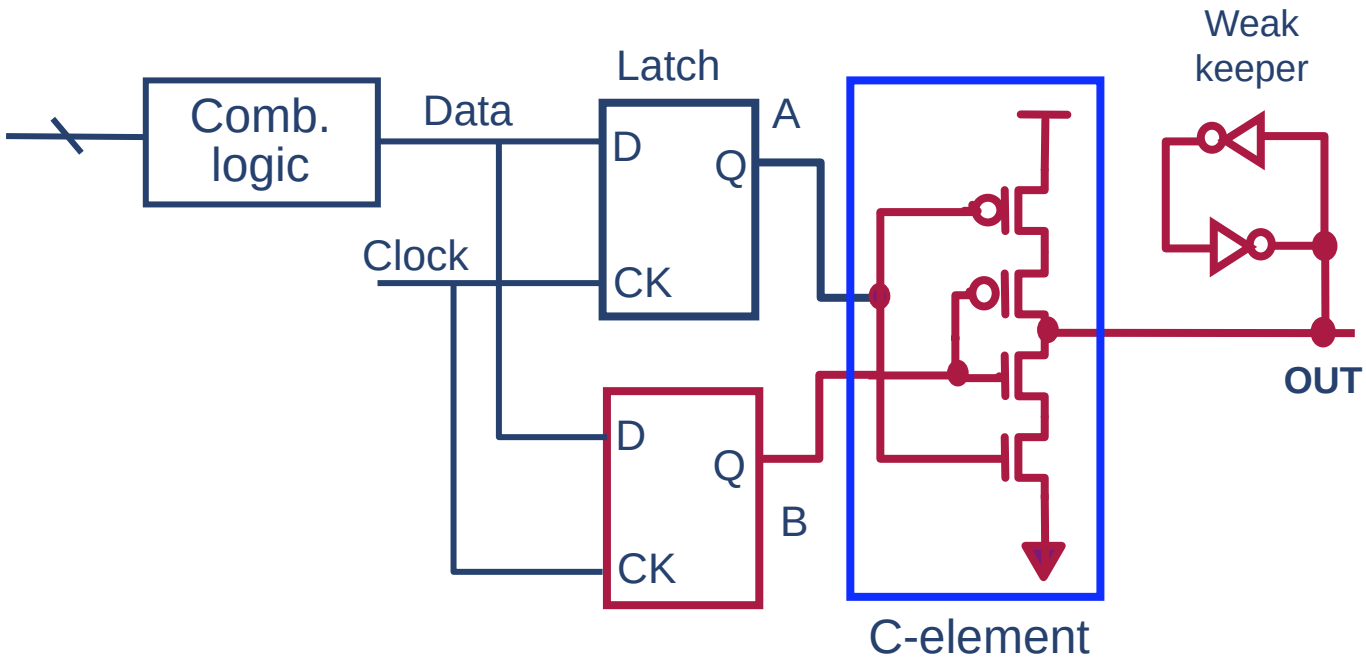


- Transient faults such as SEU manifest themselves as voltage pulses
- Temporal redundancy (sampling at 2 points in time) detects such an event
  - Error is flagged when the delayed sample does not agree with the first sample
- The error signal can be used for recovery

Source: Anghel & Nicolaidis '01

Transient error mitigation

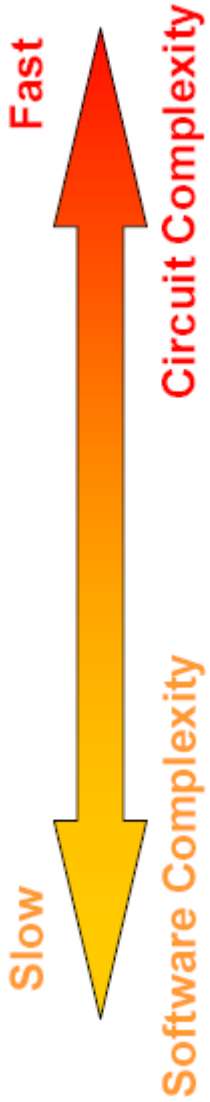
- Add redundancy to detect and correct transient errors (e.g. BISER FF)



A B	00	11	01	10
C-element (A, B)	1	0	Previous value retained	Previous value retained

Source: S. Mitra, Stanford

Failure recovery

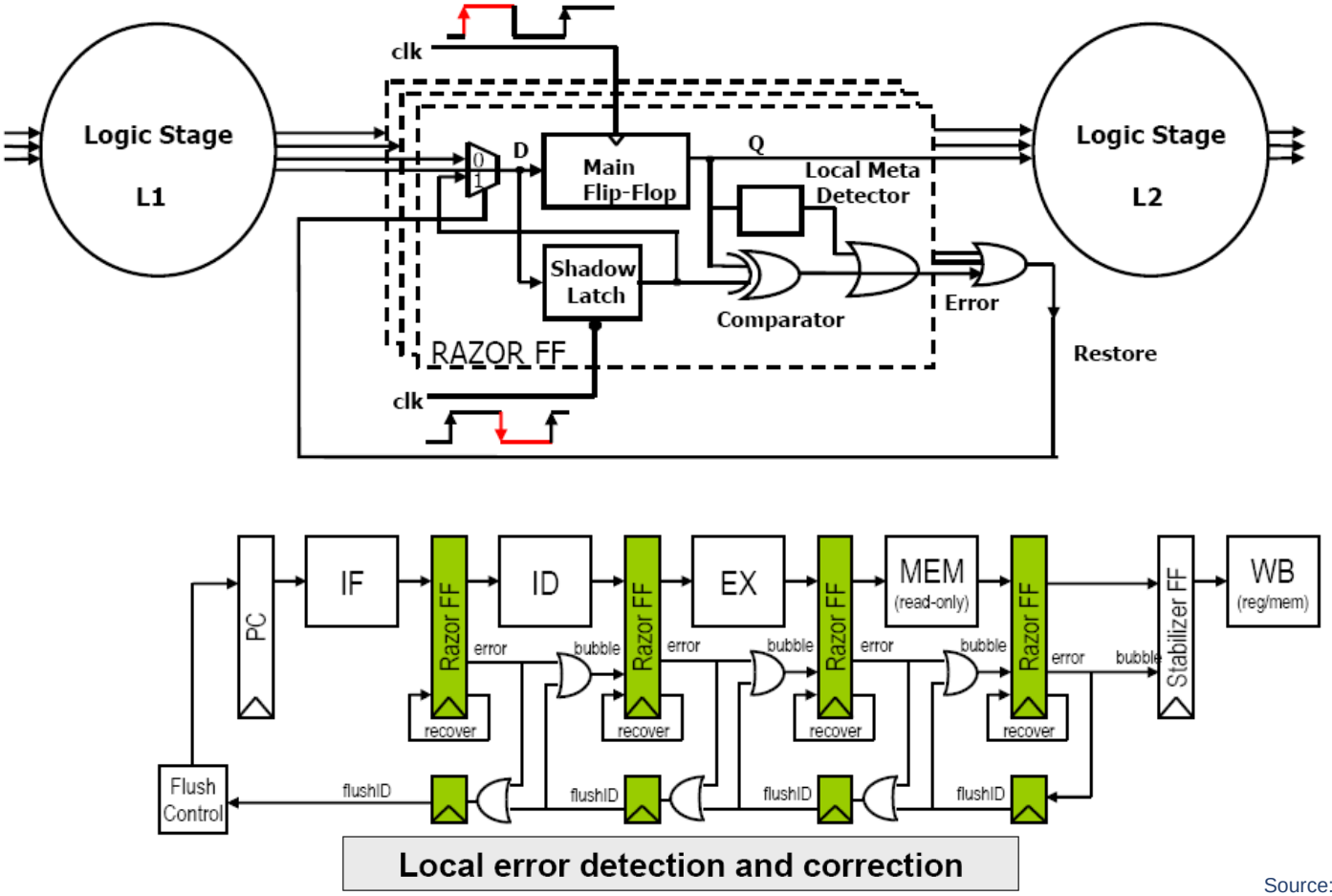


- Local recovery
  - Inject correct value into pipeline
  - Stall for one cycle and continue
- Instruction replay
  - Invalidate instructions in pipeline
  - Re-execute from failing instruction
- Checkpointing with roll-back
  - Periodically, save system state in memory
  - On error, roll back to last saved state

Source: Jim Tschanz, Intel

Failure recovery

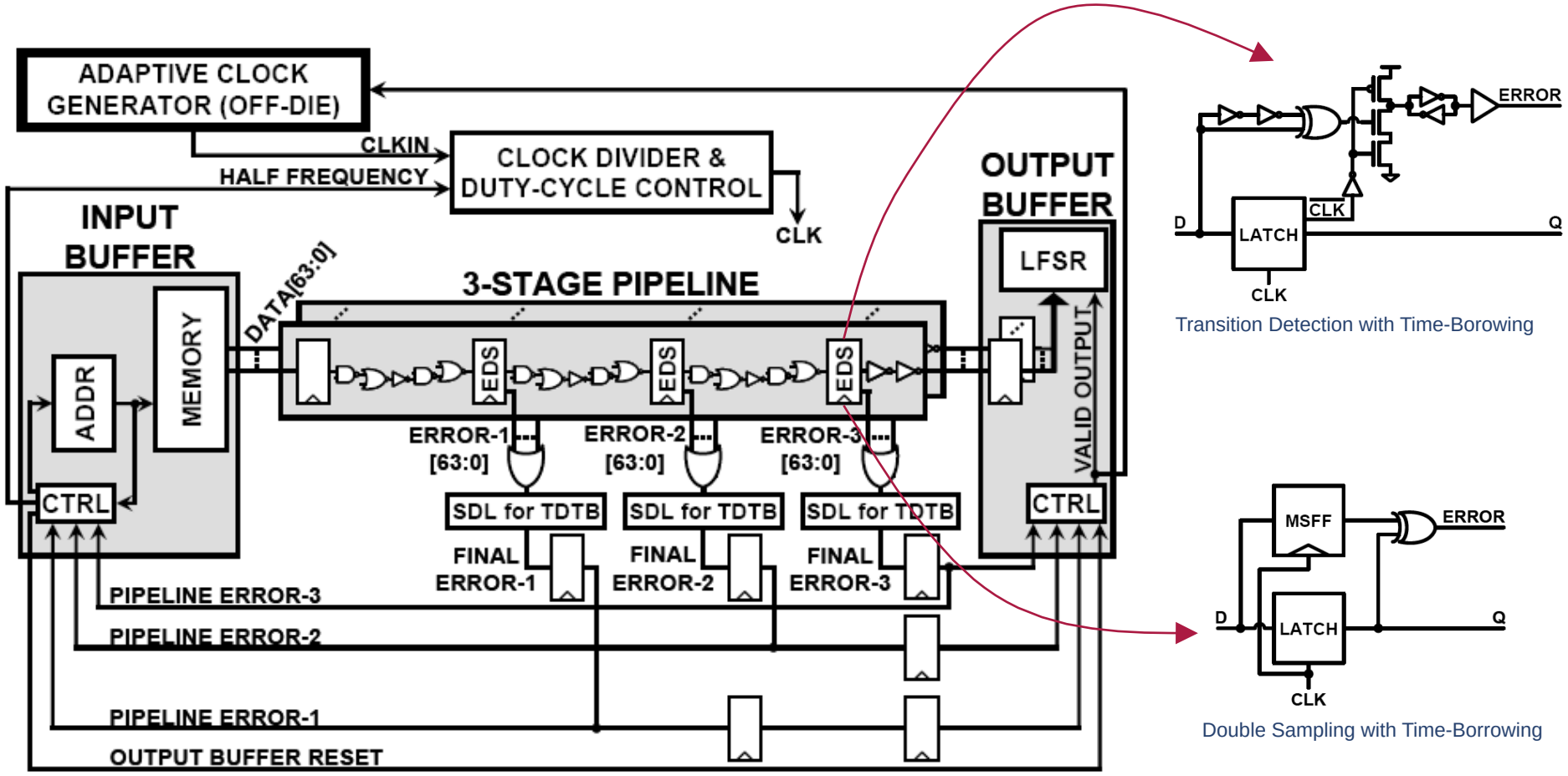
- Razor: Local error detection and correction on the fly
  - Upon failure: Overwrite main flip-flop with correct data from the shadow latch
  - Ensure that the shadow latch is always correct by conventional design



Source: S. Das et al, JSSC 2006

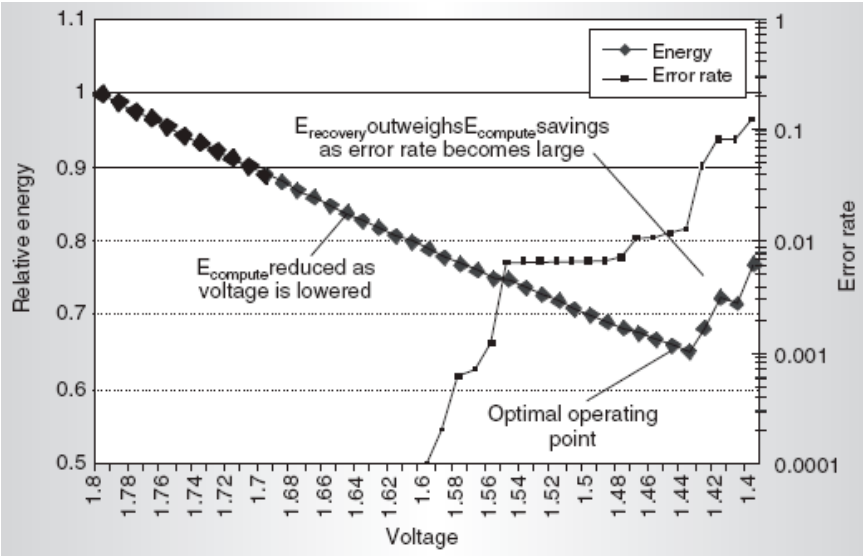
Failure recovery

- Error correction by instruction replay

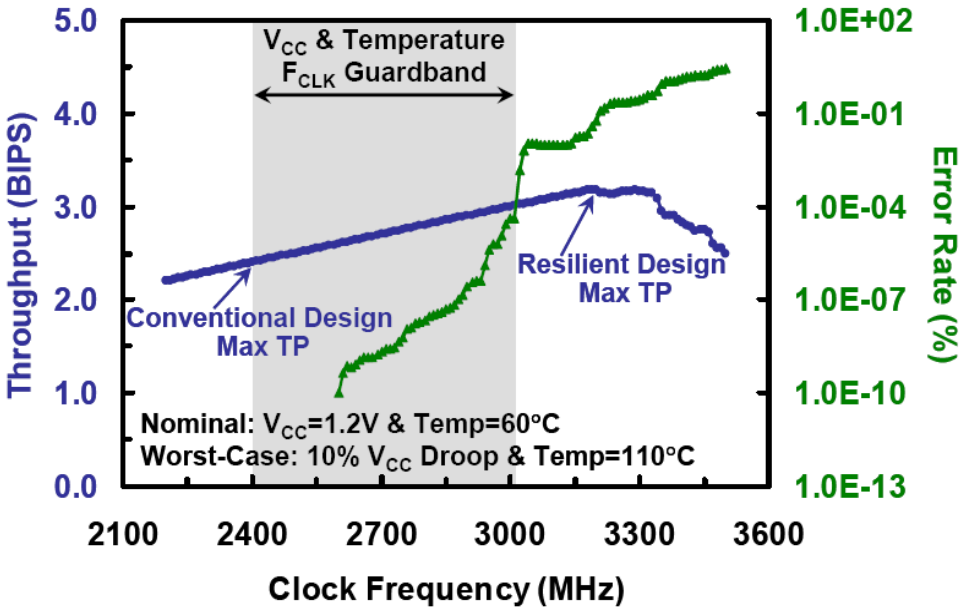


Source: K. Bowman, ISSCC 2008

Energy-error tradeoff



D. Ernst et al, IEEE Computers 2004



Source: K. Bowman, ISSCC 2008

- Adaptive designs have much lower V_{opt} than worse case designs
- Or alternatively, adaptive designs can run much faster at the same voltage

Conclusions

- Variations are becoming dominant with technology scaling
  - Spatial variations
  - Temporal variations
  - Dynamic variations
- Designing with margins is not a sustainable proposition
  - Too much power, performance overhead
- Resilient designs are needed which can adapt to variations
  - Three components required for adaptability
    - Failure prediction
    - Failure detection
    - Failure recovery

Fin