On the Efficacy of NBTI Mitigation Techniques

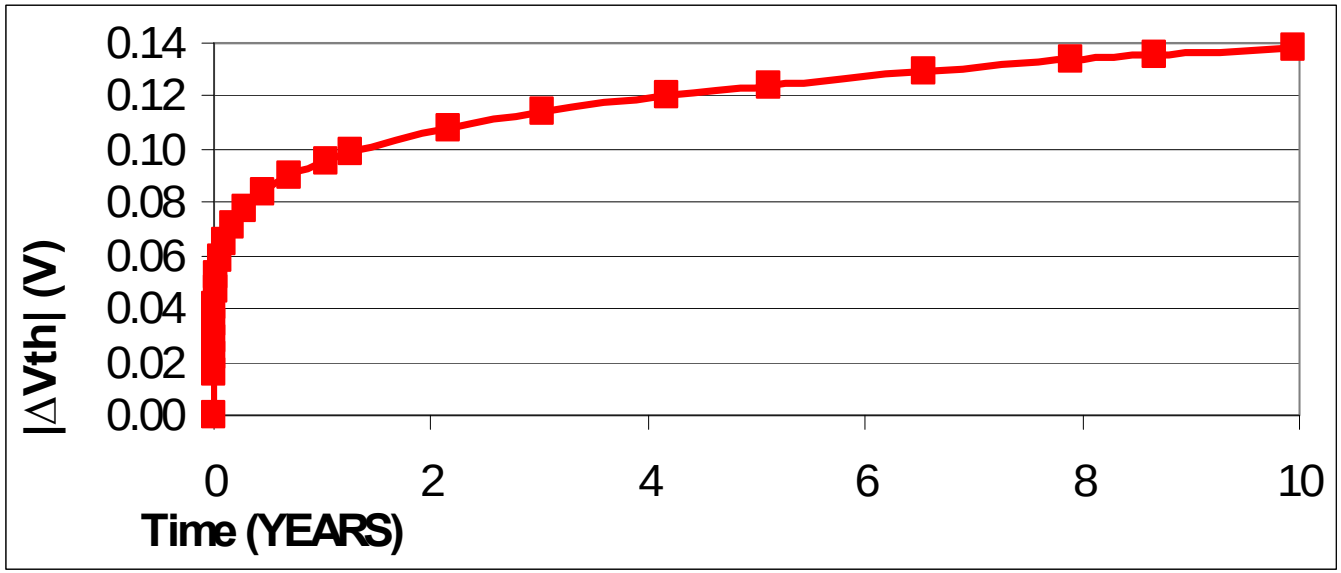
John Sartori (Illinois)
joint work with
Tuck-Boon Chan (UCLA)
Puneet Gupta (UCLA)
Rakesh Kumar (Illinois)

NBTI Background

- |Vth| increase for negatively biased PMOS
  - |Vth| increase causes **delay increase**
  - Delay increase can cause **timing failures**
- Degradation depends on **stress time** and **Vdd**

$$\Delta V_{th} \propto f(V_{dd}) \cdot t_{stress}^{time\_exponent}$$

- NBTI degradation is **front-loaded**

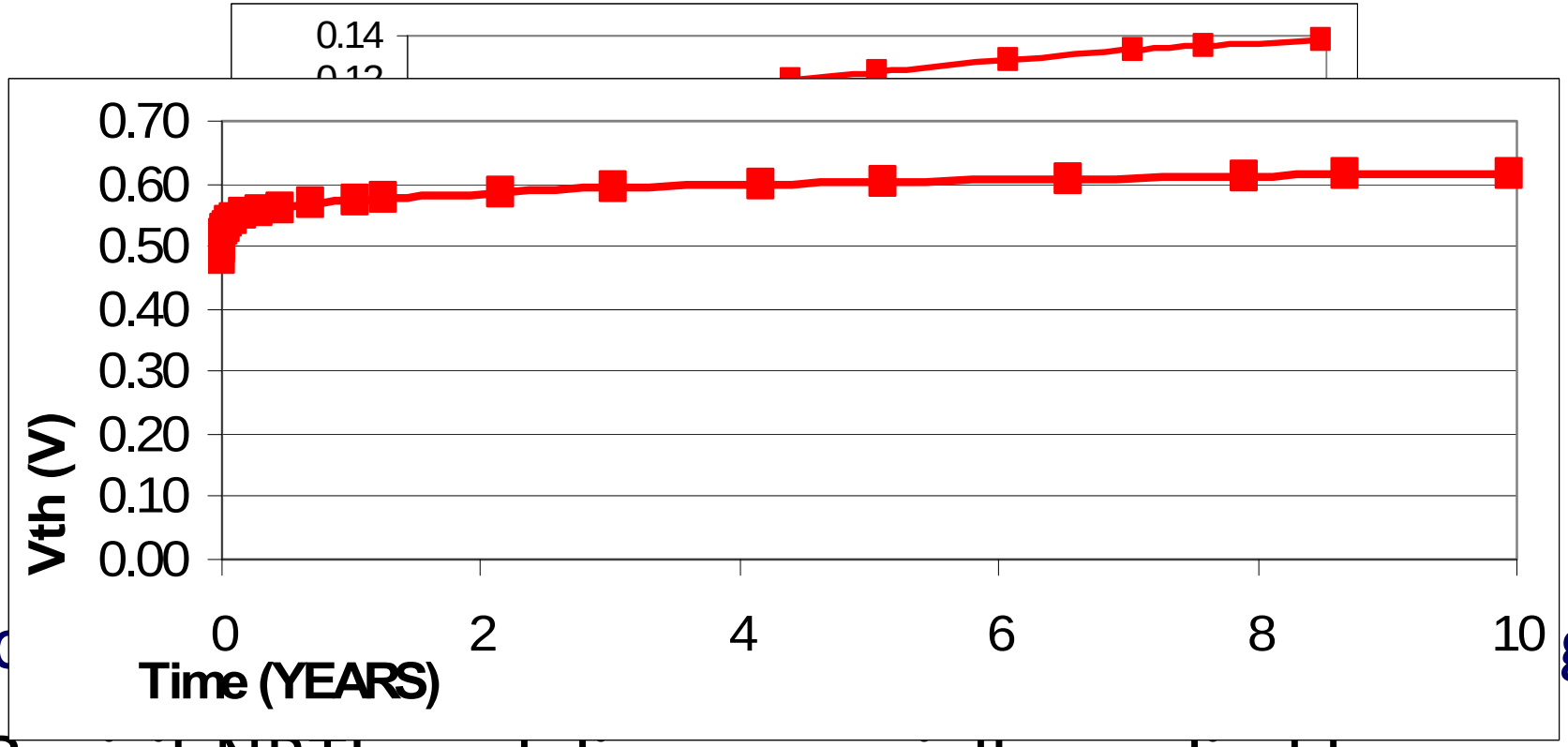


NBTI Mitigation Techniques

- **Guardbanding** is traditional way to deal with NBTI
  - Increase voltage / Reduce Frequency / Increase Area
- Many works propose techniques to reduce the cost of provisioning for NBTI
  - **Dynamic Voltage Scaling**
    - Always use lowest possible supply voltage
  - **Activity Management**
    - Attempt to put PMOS in idle state
  - **Power Gating**
    - Relax all nodes by turning power off

Motivation

- Benefits quoted by mitigation techniques seem to be at odds with front-loaded aging behavior
  - 50% of lifetime degradation occurs within 1.6 months



arc g?

- Revisit NBTI modeling, especially applied to architecture-level NBTI mitigation techniques

Outline

- Background and Motivation
- NBTI Modeling
- Application of NBTI model to architecture-level mitigation techniques
- Proposed NBTI model
- Methodology
- Revisiting architecture-level NBTI mitigation

NBTI Reaction-Diffusion Model

- Holes interact with H-passivated Si atoms
- Holes break Si—H bonds at Si/SiO_2 interface, generating traps and freeing H atoms
- H atoms anneal a trap or diffuse through SiO_2
- |Vth| increase proportional to number of traps

$$\Delta V_{th} = \frac{qN_{it}}{C_{ox}}$$

- Diffusion can also drive H atoms back toward interface when stress is relaxed → **Recovery**

NBTI Reaction-Diffusion Model

Reaction at surface

$$\frac{\partial N_{it}(t)}{\partial t} = k_f [N_0 - N_{it}(t)] - k_r N_{it}(t) C_H(x=0,t),$$

$$\frac{\partial N_{it}(t)}{\partial t} = -D \frac{\partial C_H(x,t)}{\partial x}\Big|_{x=0} + \frac{\delta}{2} \frac{\partial C_H(x,t)}{\partial t},$$

Diffusion in silicon oxide or poly

$$D \frac{\partial^2 C_H(x,t)}{\partial x^2} = \frac{\partial C_H(x,t)}{\partial t}$$

N_{it} – Interface Traps / Area

K_f – Si—H dissociation rate

K_r – Si—H annealing rate

D – Diffusion coefficient

N_0 – Number of initial bonds

C_H – H atoms / Area

δ – Interface thickness

Device-level Analytical Model

- Architecture-level techniques have been based on device-level analytical models

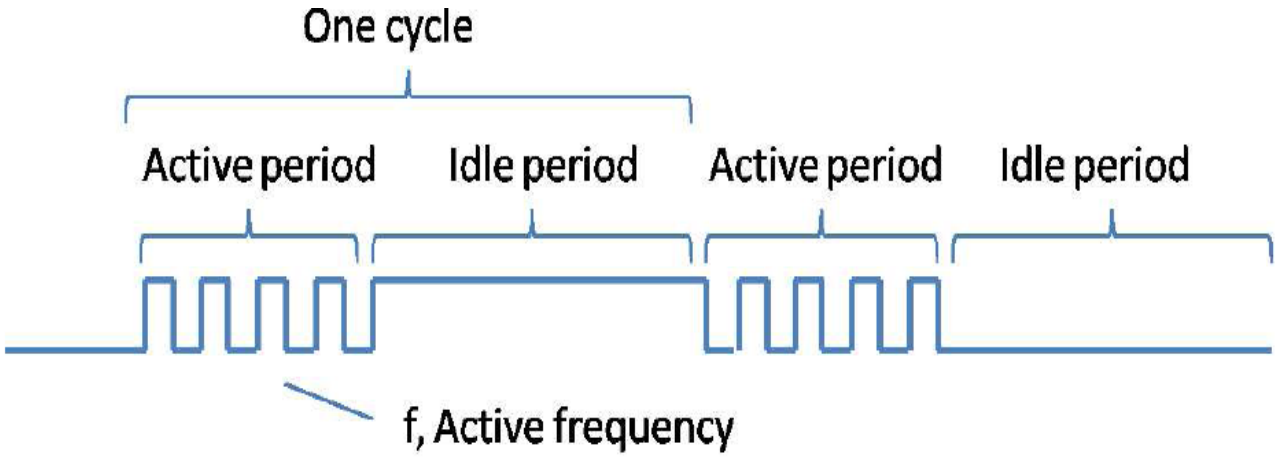
$$\Delta V_{th} = A_{NBTI} \cdot \tau_{ox} \cdot \sqrt{C_{ox}(V_{dd} - V_{th})} \cdot e^{\frac{V_{dd} - V_{th}}{\tau_{ox} E_0} - \frac{E_a}{kT}} \cdot t_{stress}^{0.25}$$

- Model fine for device-level analysis, **but**:
  - Assumes constant supply voltage
  - Assumes fixed, periodic signal with 100% activity
  - Does not model interactions over paths or circuits

Device-level analytical models not suitable to model the impact of dynamic, architecture-level techniques

Flexible, Numerical Aging Model for NBTI

- Solve Reaction-Diffusion equations **numerically**
- Same underlying NBTI model, but now we can account for the impact of architecture techniques
  - Supply voltage can be substituted at any time step
  - Arbitrary activity patterns can be simulated
  - Waveform is adapted to model path and circuit effects



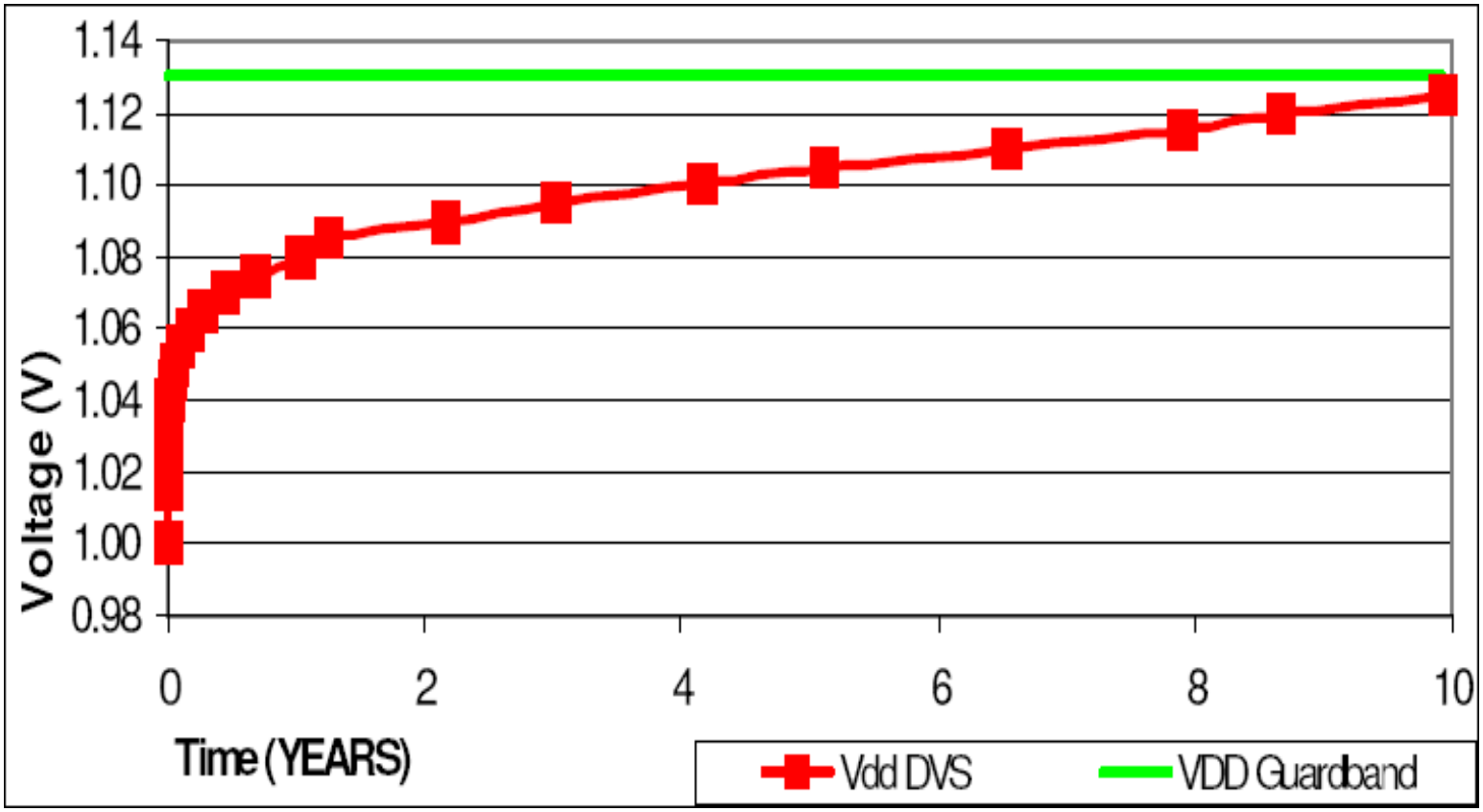
Outline

- Background and Motivation
- NBTI Modeling
- Application of NBTI model to architecture-level mitigation techniques
- Proposed NBTI model
- Methodology
- Revisiting architecture-level NBTI mitigation

Methodology

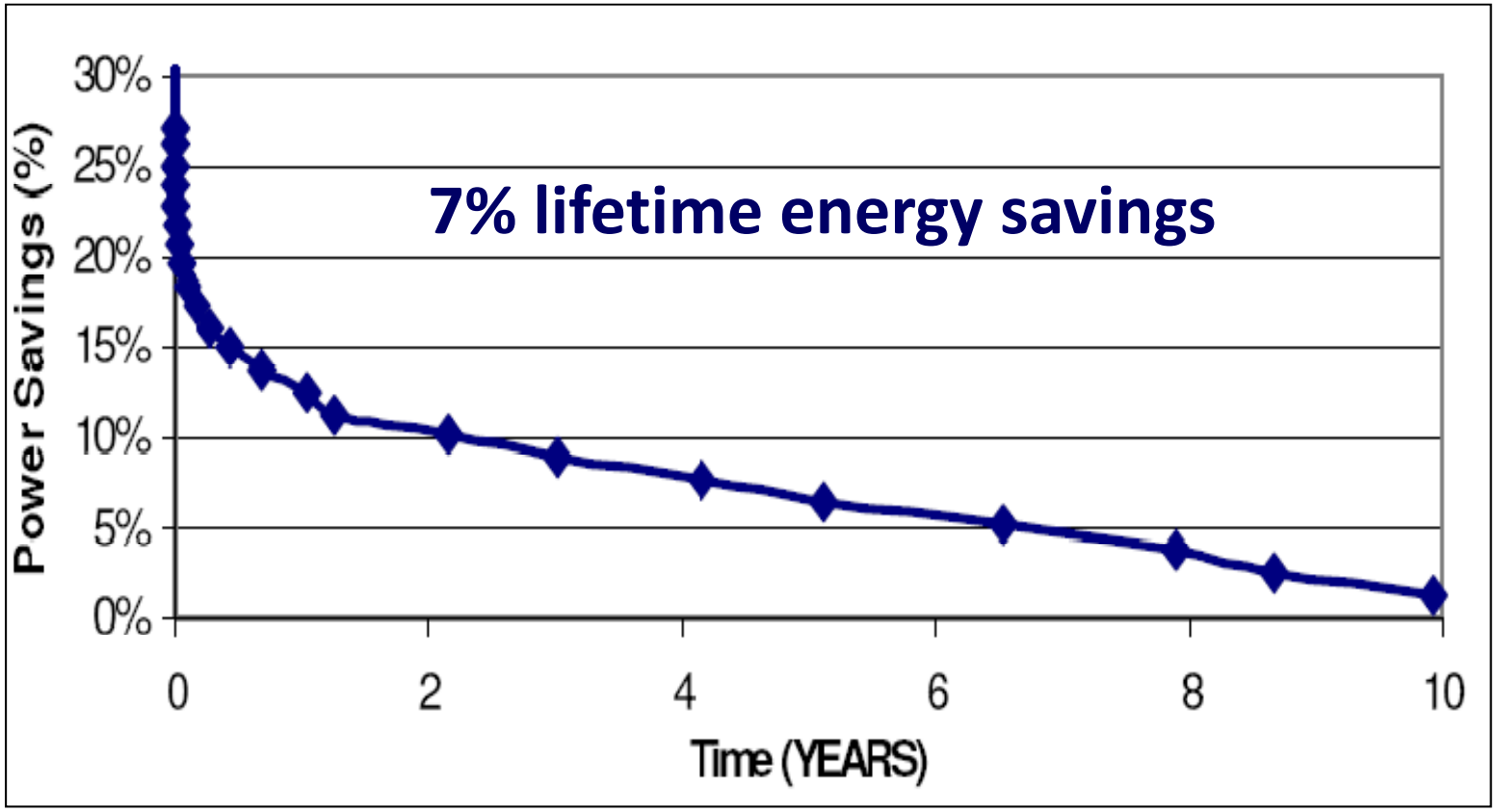
- SP&R OpenSPARC T1, characterize critical paths
- Numerical simulation framework calculates Vth degradation
- SPICE models degradation vs delay relationship for critical paths
- SMTSIM+SPEC characterize processor throughput and activity

Results – Dynamic Voltage Scaling



Supply voltage approaches guardband quickly in early lifetime. Power savings limited afterward.

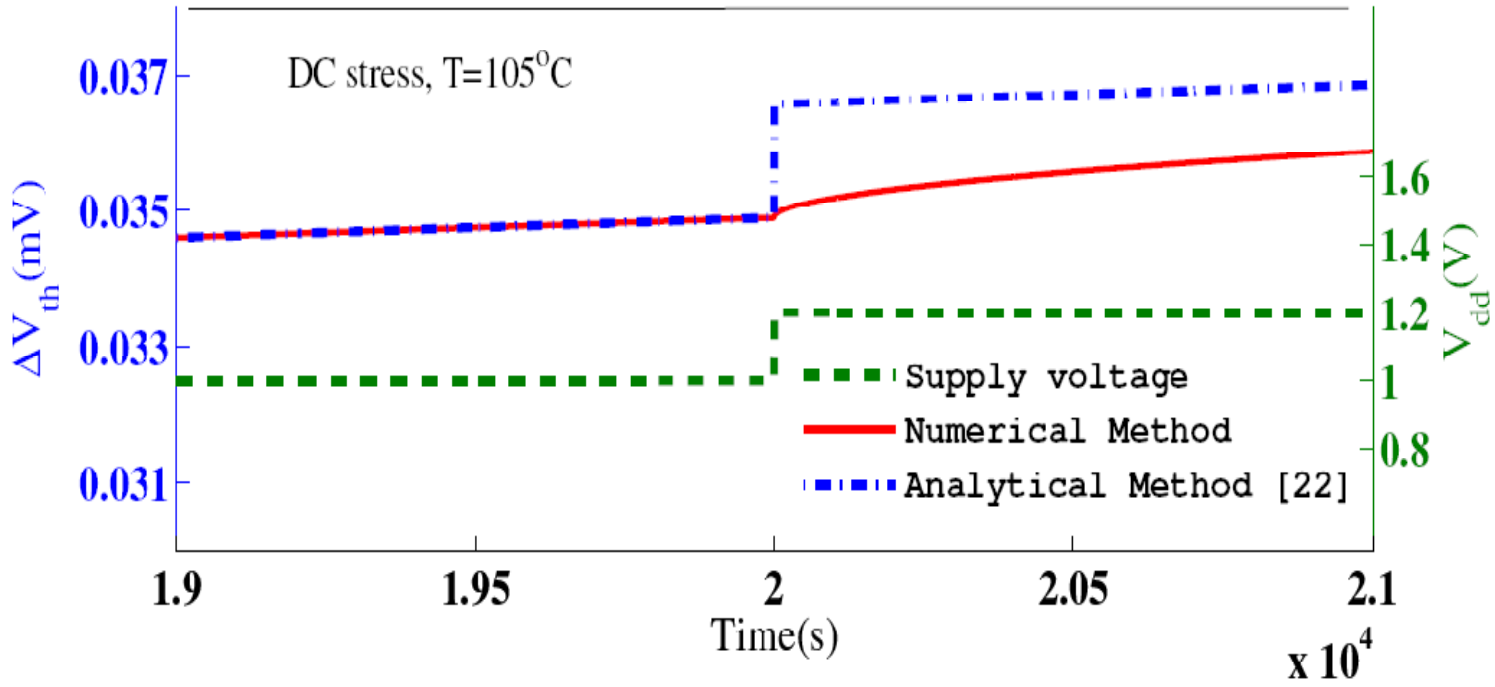
Results – Dynamic Voltage Scaling



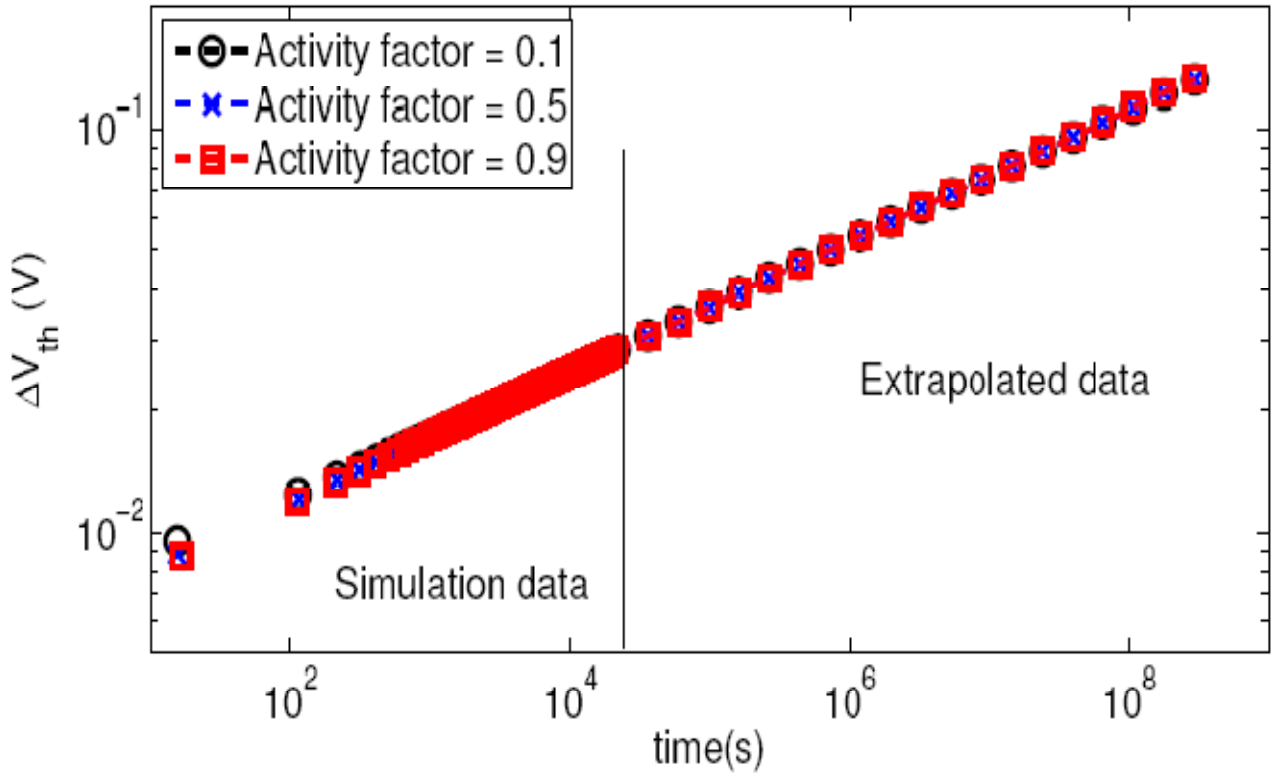
Significant power savings early on, limited later.
Benefits degrade for realistic DVS.

Limitations of Device-level Models

- Analytical equation does not model physical degradation phenomenon
- Changing voltage in analytical equation is like instantaneously changing internal device state



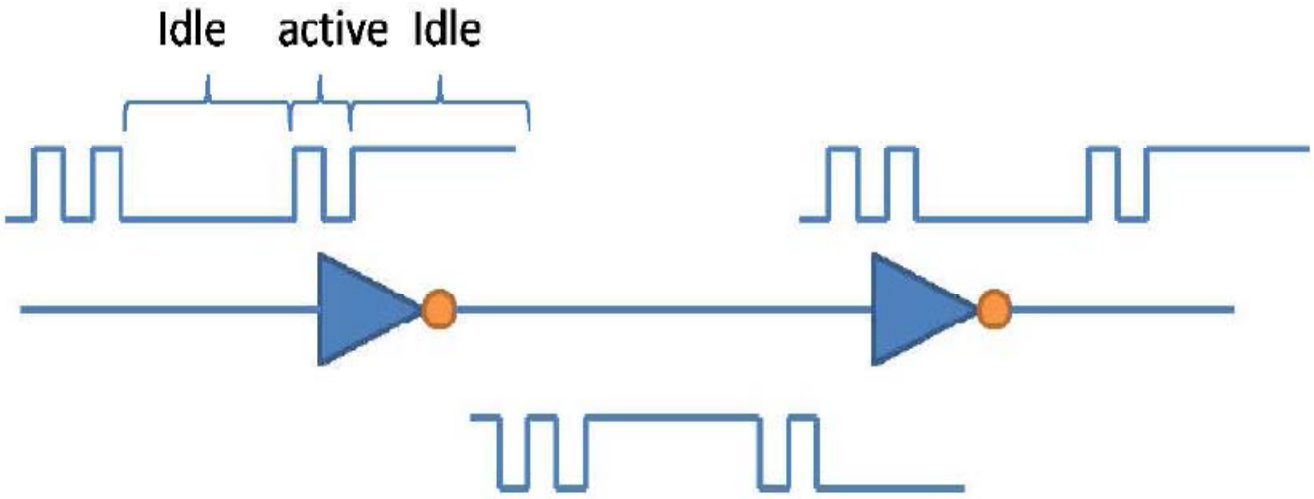
Results – Activity Management



Due to complementary nature of logic, idling and signal biasing techniques are not effective.

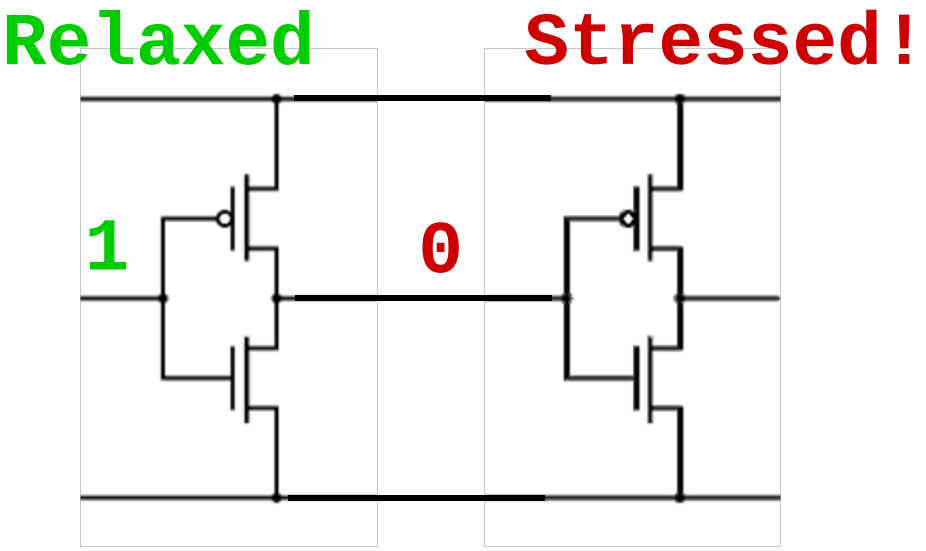
Limitations of Device-level Models

- AC signals used previously do not resemble typical digital signals in CMOS circuits
  - They assume 100% activity
  - They assume all PMOS behave the same time
    - When one is relaxed, all are relaxed



Limitations of Device-level Models

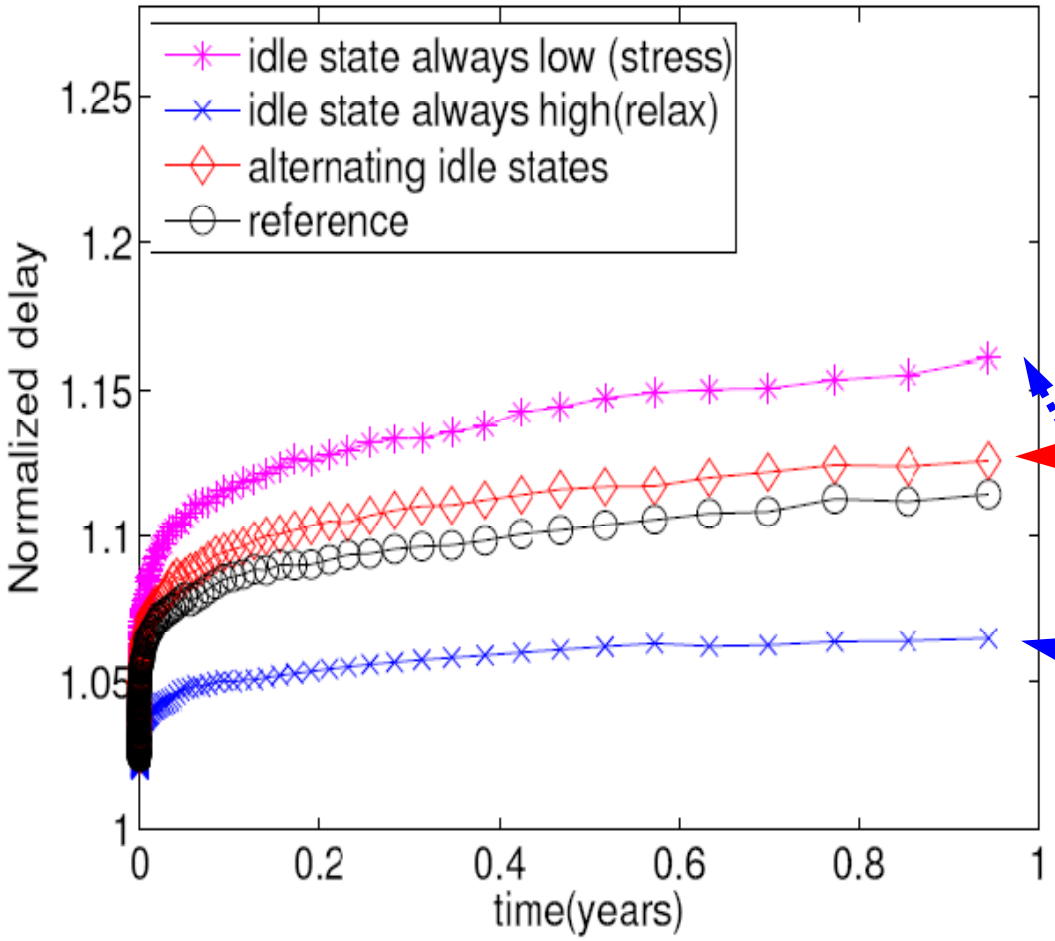
- CMOS stands for **Complementary** MOS
- Relaxation state at node implies stress state at next node

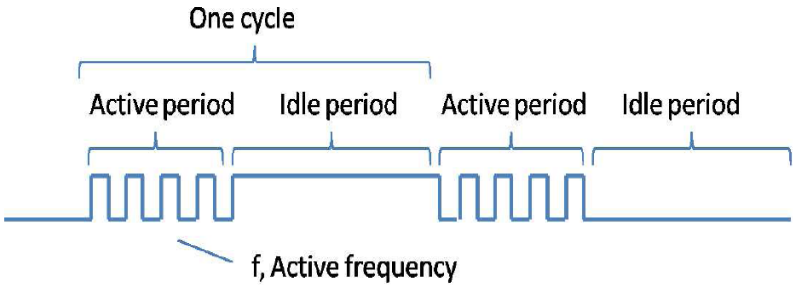


Device-level model ignores circuit-specific implications like path and circuit effects.

Limitations of Device-level Models

- Alternating values in idle state models averaging effect of degradation across logic path

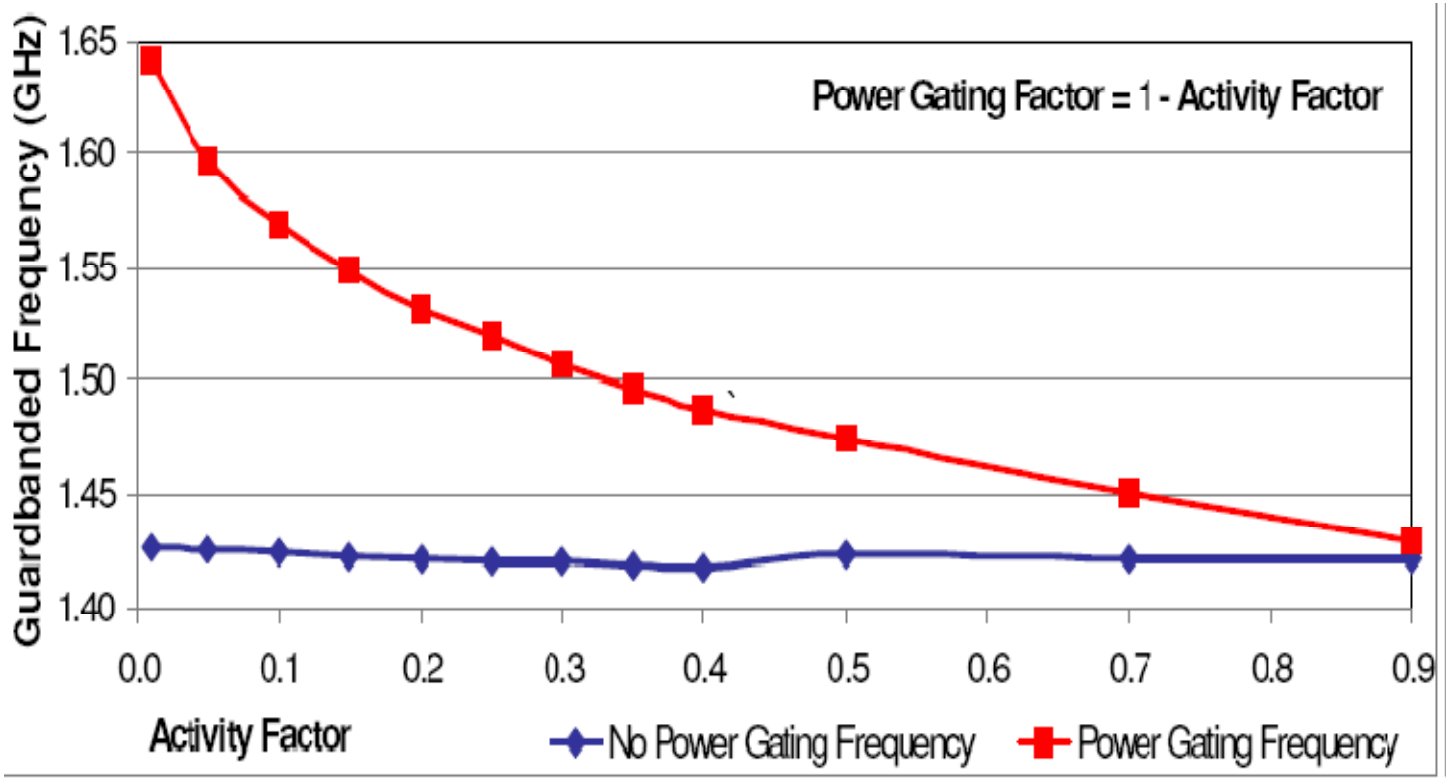




< 1% estimation error

±5% delay estimation error

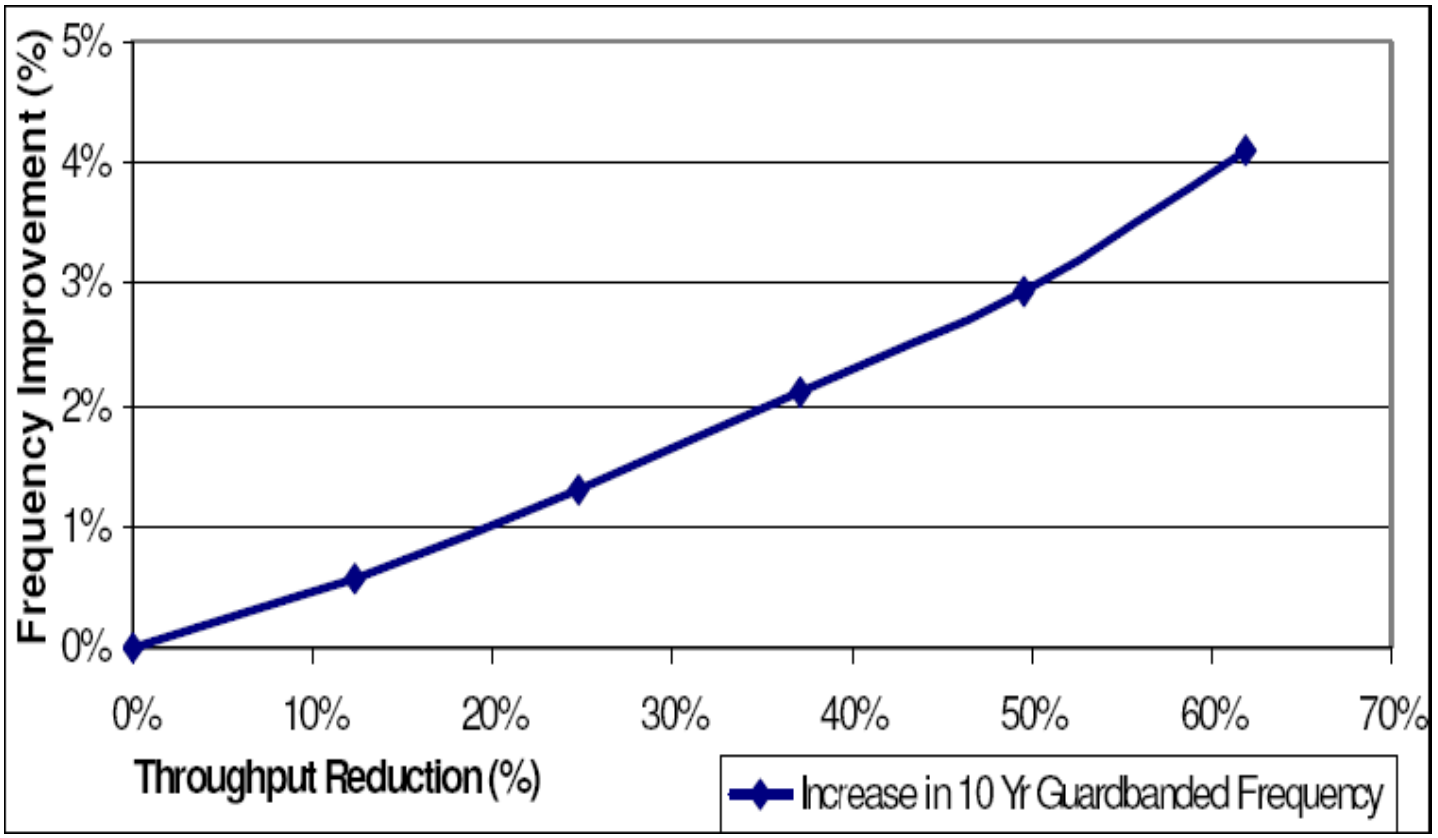
Results – Power Gating



Up to 15% improvement in guardbanded frequency for 9.9 yrs spent power gating. (10 yr lifetime)

6 yrs power gating buy 5% frequency improvement.

Results – Activity Management + Power Gating



Reduce activity, more power gating, less degradation.

60% throughput loss for 4% degradation reduction.

Summary and Conclusions

- Front-loaded nature of NBTI impacts the efficiency of architecture-level NBTI mitigation techniques
  - Reported benefits were inconsistent with device-level NBTI behavior
- Applied flexible numerical simulation approach to model impact of architecture-level techniques
- Results from evaluations using the proposed model consistent with device-level behavior
  - Guardbanding almost as good as ALL previously proposed techniques
- Numerical aging model available for download at

http://nanocad.ee.ucla.edu/Main/DownloadForm

Acknowledgments

- We thank NSF expedition for sponsoring this research

www.variability.org

BONUS SLIDES!

NBTI Reaction-Diffusion Model

Reaction at surface

$$\frac{\partial N_{it}(t)}{\partial t} = k_f [N_0 - N_{it}(t)] - k_r N_{it}(t) C_H(x=0,t),$$

$$\frac{\partial N_{it}(t)}{\partial t} = -D \frac{\partial C_H(x,t)}{\partial x}\Big|_{x=0} + \frac{\delta}{2} \frac{\partial C_H(x,t)}{\partial t},$$

Diffusion in silicon oxide or poly

$$D \frac{\partial^2 C_H(x,t)}{\partial x^2} = \frac{\partial C_H(x,t)}{\partial t}$$

N_{it} – Interface Traps / Area

K_f – Si—H dissociation rate

K_r – Si—H annealing rate

D – Diffusion coefficient

N_0 – Number of initial bonds

C_H – H atoms / Area

δ – Interface thickness

Numerical Model Details

- Trap generate rate slow compared to dissociation and annealing rates

$$\frac{\partial N_{it}(t)}{\partial t} \approx 0, \text{and}$$
$$N_{it(t)} << N_0,$$

- Simplified

$$C_H(x=0,t)N_{it}(t) \approx \frac{k_f}{k_r} N_0,$$
$$D\frac{\partial^2 C_H(x,t)}{\partial x^2} = \frac{\partial C_H(x,t)}{\partial t}$$

Numerical Model Details

$$\alpha = \frac{D\Delta t}{\Delta x^2}$$

$$C_H(x_0, t_i) = \begin{cases} [\frac{k_f N_0}{k_r \cdot N_{it}(t_i)}]^S & \text{if device under stress} \\ 0 & \text{if device under relaxation} \end{cases}$$

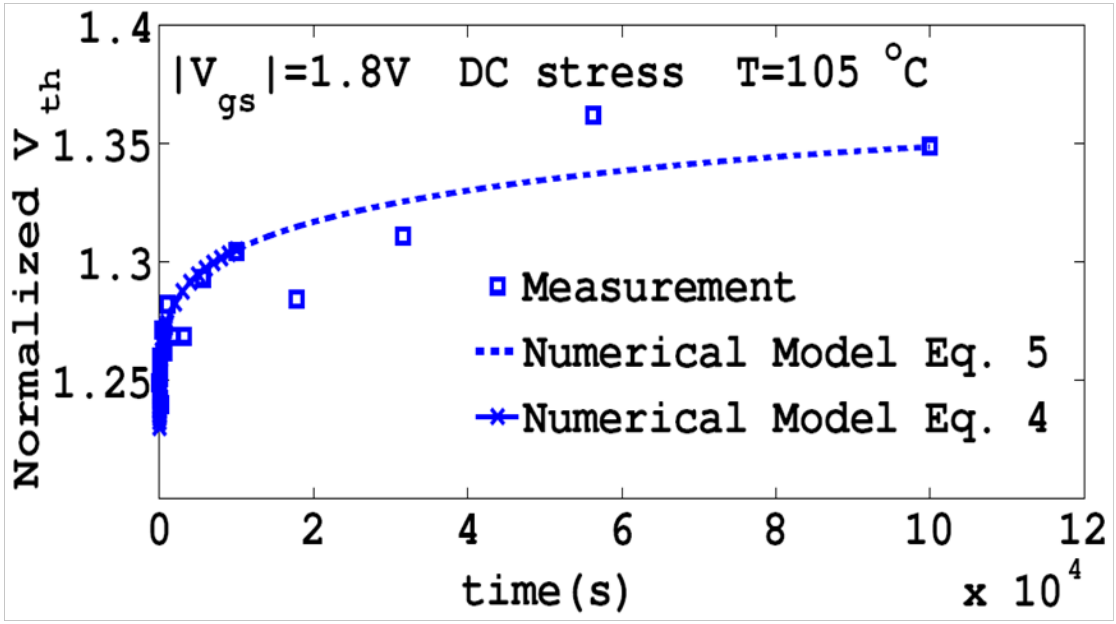
$$\mathbf{C}_H(t_{i+1}) = \mathbf{W}\mathbf{C}_H(t_i),$$

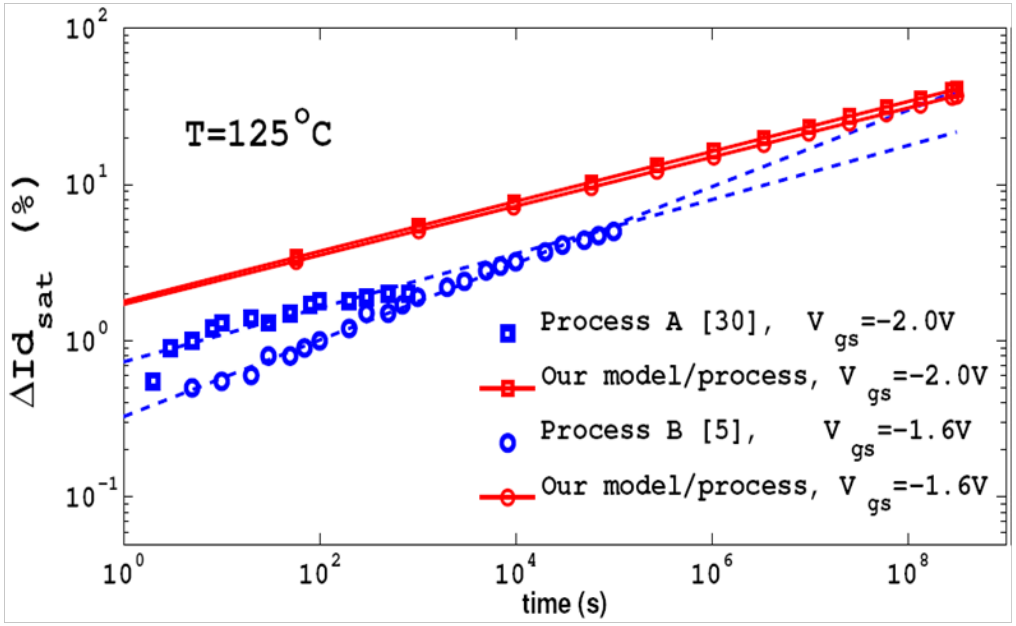
$$N_{it}(t_{i+1}) = S[1, 1, \ldots, 1]\mathbf{C}_H(t_i + 1),$$

$$\mathbf{C}_H(t) = \begin{bmatrix} C_H(x_0, t) \\ \vdots \\ C_H(x_n, t) \end{bmatrix},$$

$$\mathbf{W} = \begin{bmatrix} 1-\alpha & 1 & 0 & 0 & 0 & \ldots \\ 1 & 1-2\alpha & 1 & 0 & 0 & \ldots \\ 0 & 1 & 1-2\alpha & 1 & 0 & \ldots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}$$

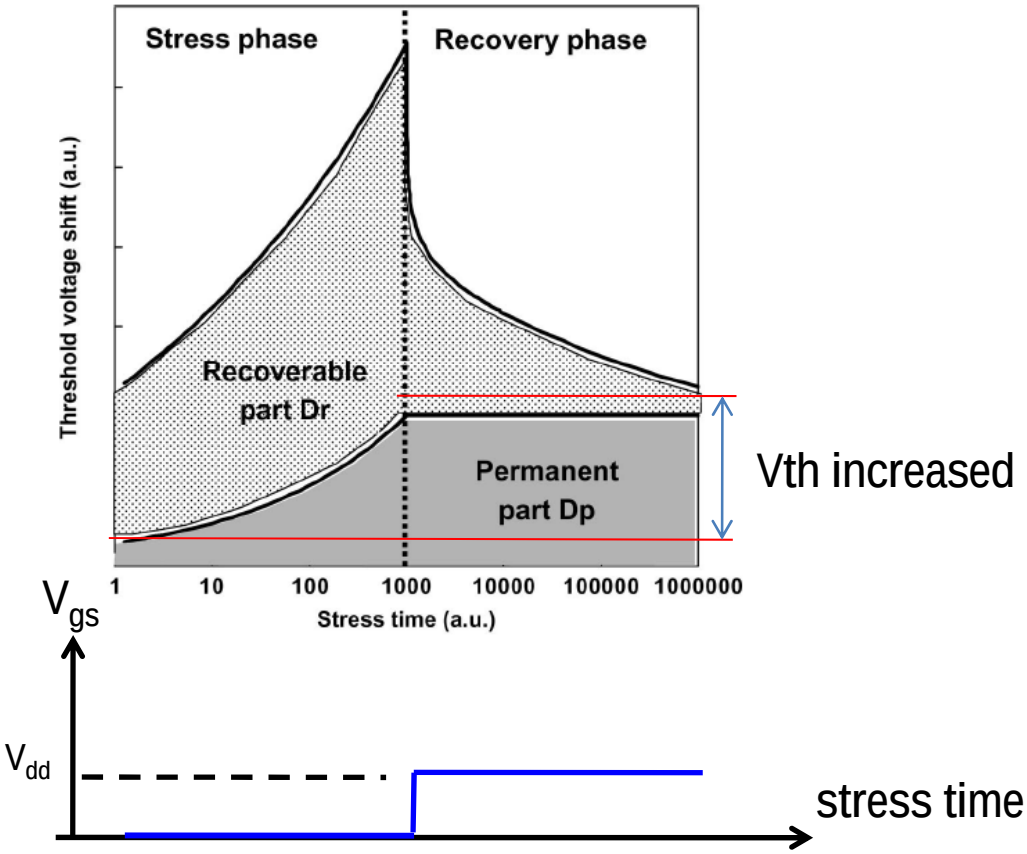
Numerical Model Details – Validation





NBTI Background: Stress and Recovery

- NBTI affects PMOS (PBTI affects NMOS)
- Two important phases of NBTI
  - **Stress:** |Vth| increases when a PMOS is on
  - **Recovery:** Part of the |ΔVth| degradation is recovered when PMOS is off



Activity Management

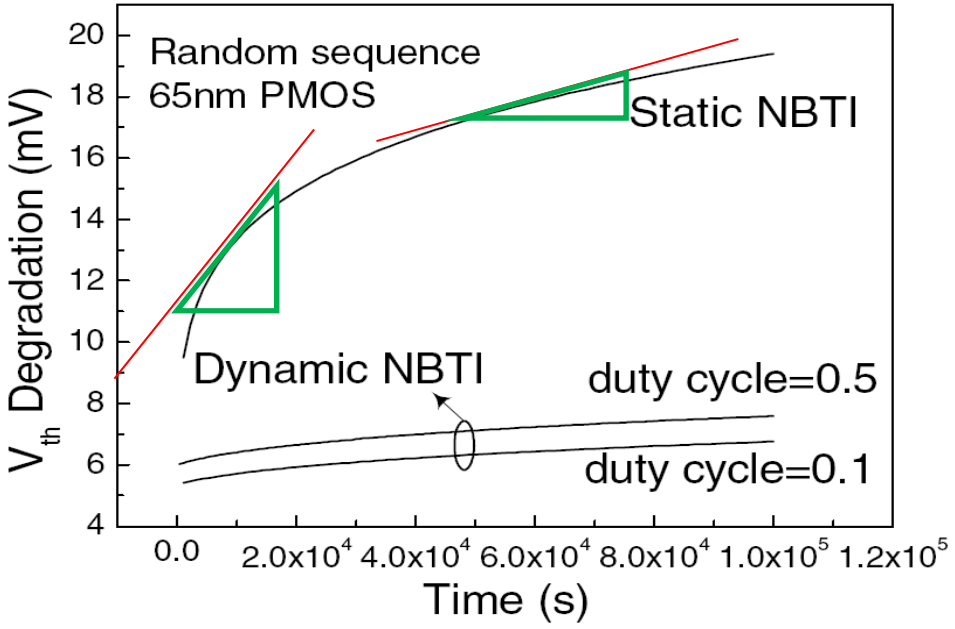
- Manage activity factor to control stress / relaxation ratio
- Bias signal probabilities to relax PMOS
- Throttle processor activity to enable power gating

Issues:

- CMOS (Complementary MOS) stresses roughly half of nodes even in idle mode
- Front-loaded degradation curves converge quickly after short active time

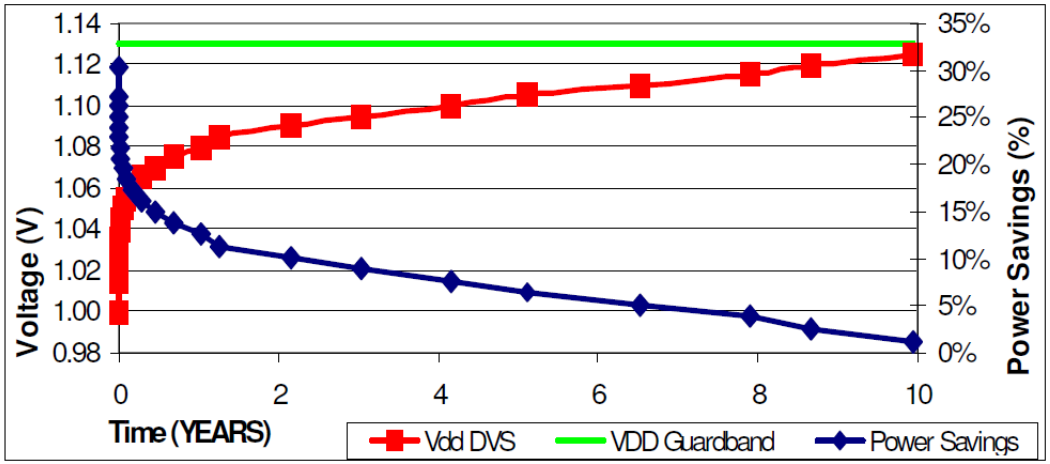
Background

- Degradation rate is fast initially, slows down after a PMOS is stressed for extended period of time
  - Similarly, recovery rate is fast initially and slows down after a short period → there is unrecoverable Δ|Vth|
- Static NBTI vs. dynamic NBTI
  - No recovery phase for static NBTI →large Δ|Vth|
- NBTI degradation increases when
  - Electric field across gate oxide increases (I.e., V_{gs} increases)
  - Operating temperature increases



NBTI Mitigation Techniques

- Dynamic Voltage Scaling
  - NBTI degradation happens very fast at the beginning
    - Rapid Supply voltage adjustments happen only at early lifetime
    - Power saving is not significant after early lifetime →efficiency of DVS reduces



  - Power saving will be less if overhead of implementing DVS is included
  - DVS has less peak power (happens at time ≈0) compared to simple guardbanding (use a larger V_{dd})

NBTI Mitigation Techniques

- Dynamic Instruction Scaling (DIS)
  - Changes instructions to control/limit circuit activity
    - Assume circuit only degrades when it is switching or active
  - But CMOS always has inverting signal → a PMOS is under recovery →PMOS at the next node is under stress
- Examine efficiency of DIS
  - Δ|Vth| is not sensitive to activity factor

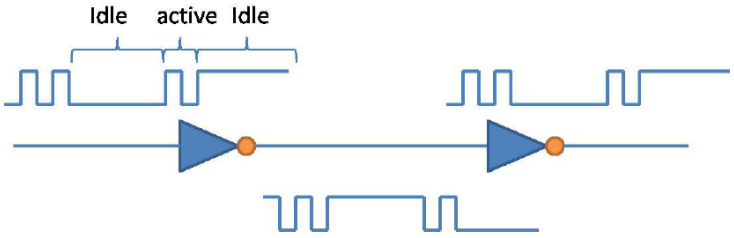
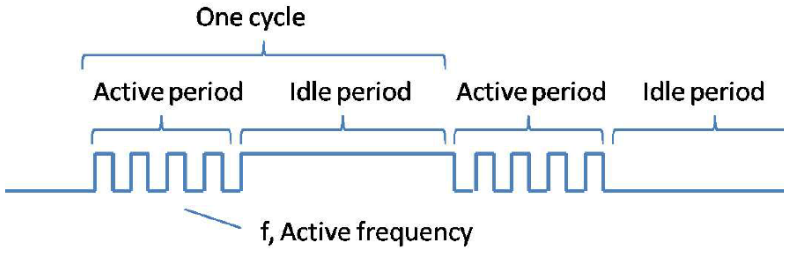
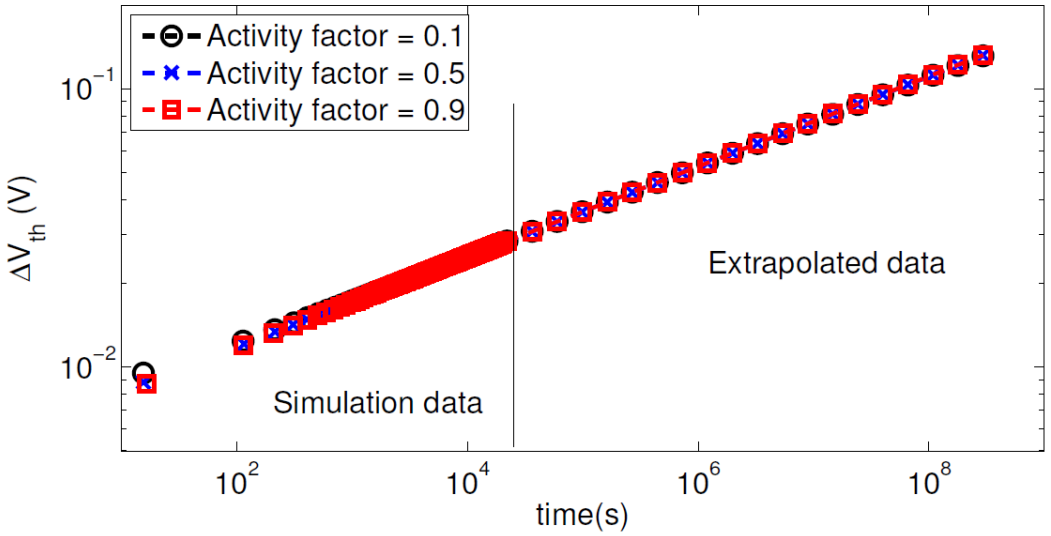


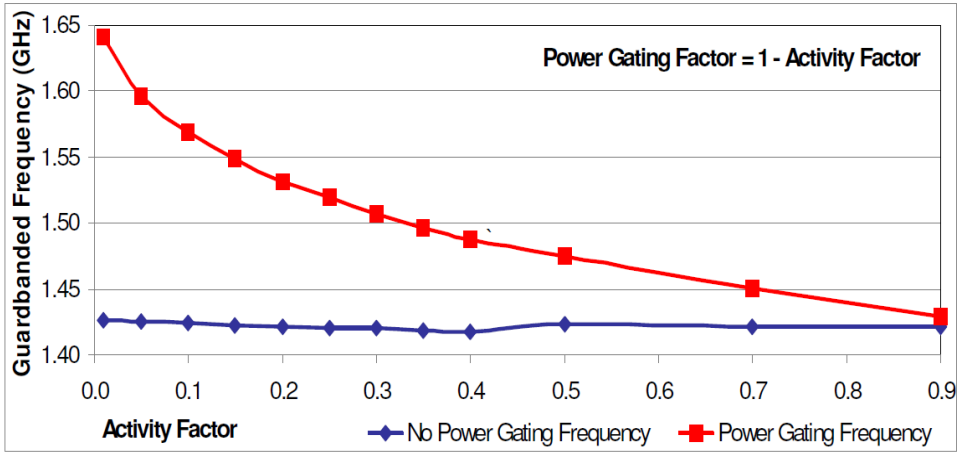
Fig. 3: Signal in a typical digital system.

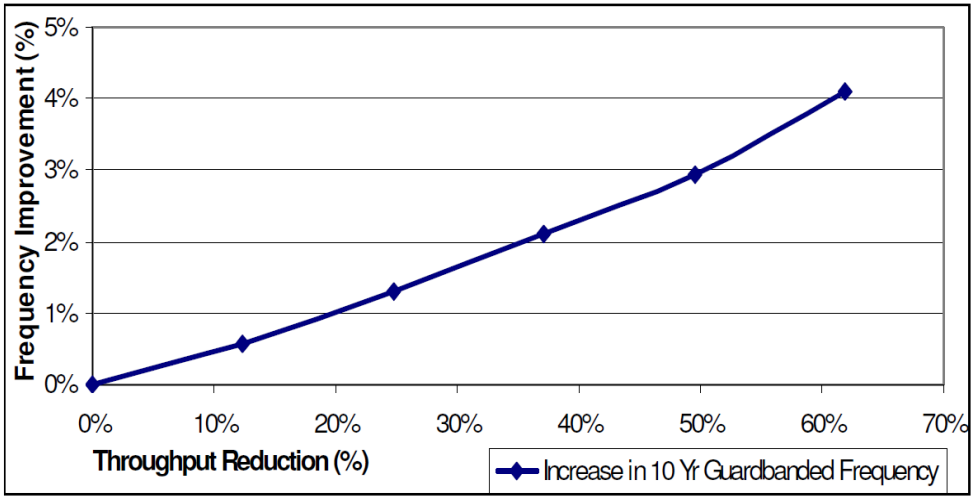




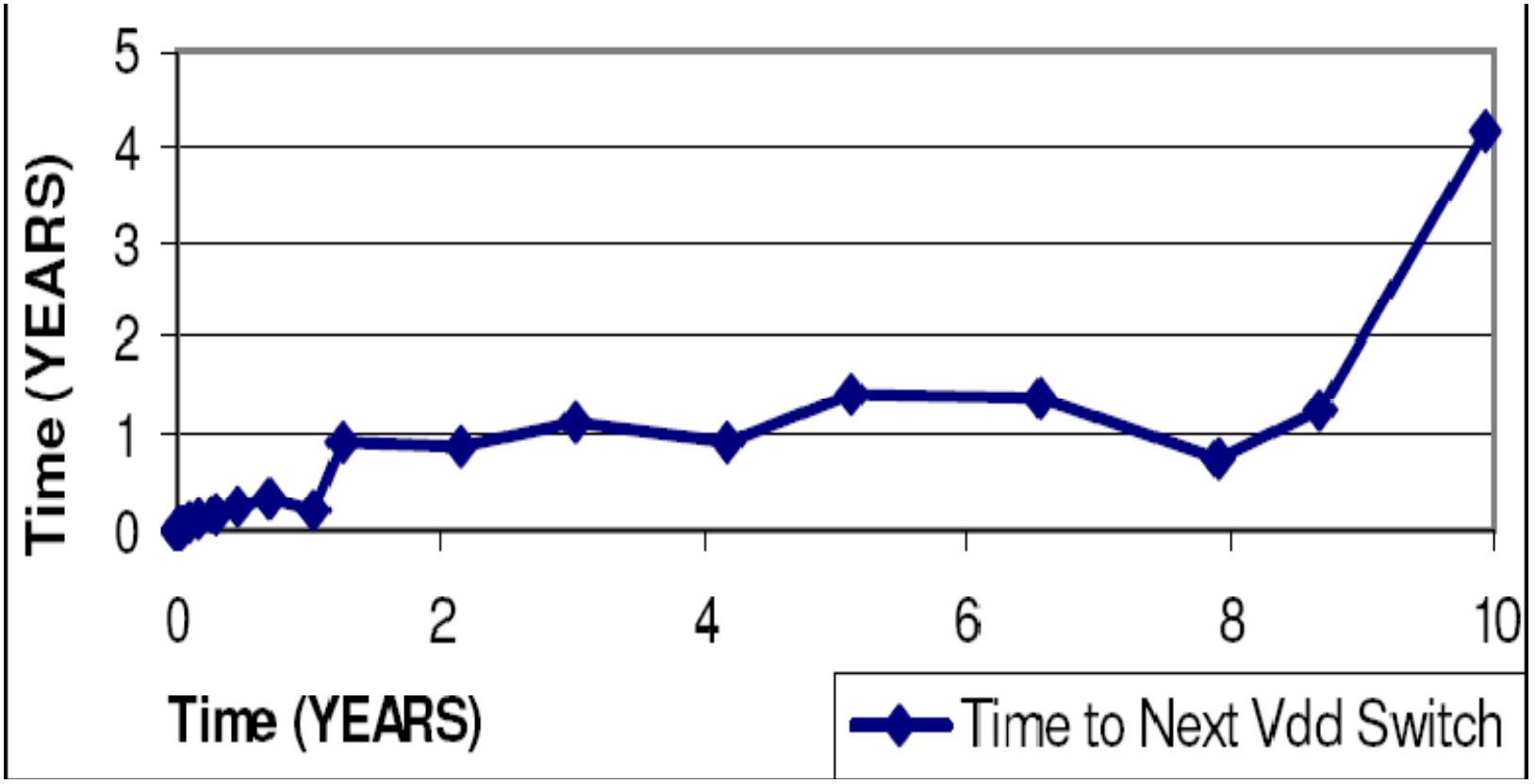
NBTI Mitigation Techniques

- Power-gating circuit to reduce degradation
  - PMOS recovers during power gating
- But NBTI degradation happens very quickly
  - Benefit of power gating is only significant for circuit with very low activity
- Adapt processor configuration to reduce activity
  - Throughput penalty is high





Results – DVS



Voltage switches frequent first few days/weeks.
Afterward, time between switches ~years.

Lifetime-Aware Adaptation

- Monitor MTTF and adapt processor to meet lifetime target

Issues:

- Average failures over lifetime
- Linear degradation curve rather than logarithmic

Motivation for NBTI Mitigation

- Guardbanding introduces a power / performance cost over the entire lifetime of the processor

If aging doesn't fully accumulate
until end of lifetime,
why pay full price for entire lifetime?

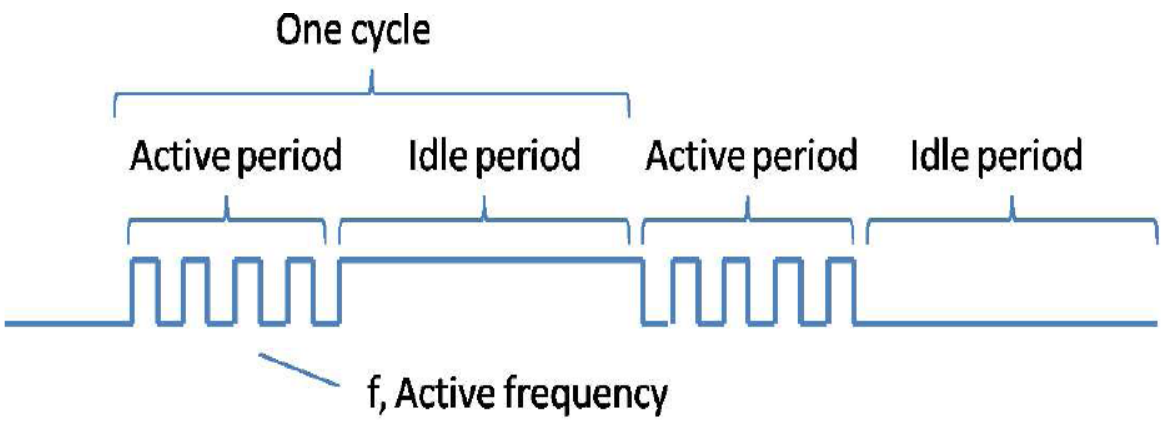
Revisiting NBTI Analysis

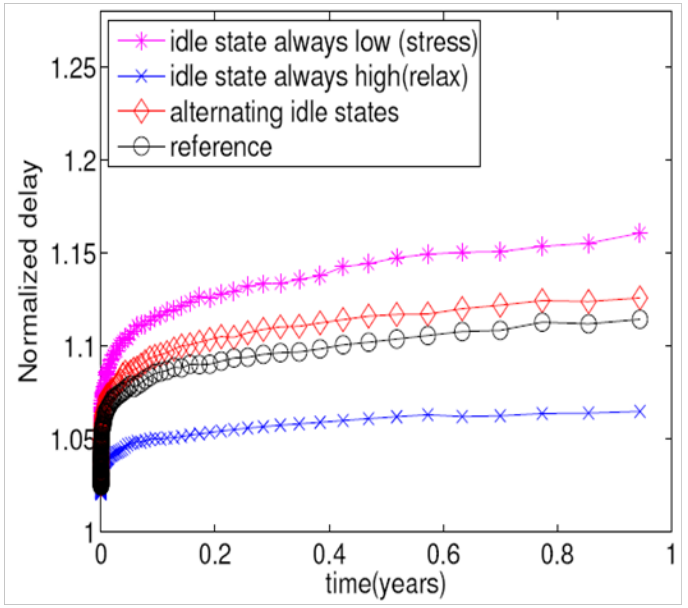
- Device-level models used to motivate and analyze architecture-level mitigation techniques
- Results and conclusions should be revisited with a capable model

We present a flexible numerical model for NBTI degradation that can model the impact of architecture-level NBTI mitigation techniques.

Flexible Numerical Model

- Dynamic voltage scaling
  - Numerical solution allows direct voltage substitution
- Signal modeling
  - Simulator can parse digital signal waveforms
- Inverting nature of CMOS
  - Use of alternating values in idle state





Conclusions

- Recent works propose architecture-level NBTI mitigation
- Architecture-level techniques based on inadequate device-level models
- We present a flexible numerical simulator that can model the impact of architecture-level techniques
- Re-evaluation of previous techniques shows that benefits may be less than suggested
- Guardbanding may still be the best approach
- Numerical aging model available for download and use in aging-related research

Download the Simulator – http://nanocad.ee.ucla.edu/Main/DownloadForm