Analysis and Optimization of Thermal Issues in High Performance VLSI

Kaustav Banerjee
Stanford University

Massoud Pedram and Amir H. Ajami
University of Southern California

Presentation Outline

- Introduction
- Thermal Effects and Reliability
- Interconnect Performance Optimization
- High-Current Effects: ESD
- Analysis of Non-uniform Chip Temperature
- Non- Uniform Temperature Dependent Delay
- Circuit Optimization: Clock Skew
- Summary

K. Banerjee, M. Pedram and A. H. Ajami

Introduction

- Sources of chip power dissipation
- Chip temperature model
- Thermal effects in interconnects
- Scaling trends and implications

K. Banerjee, M. Pedram and A. H. Ajami

Sources of Chip Power Dissipation

Devices: Close to Heat Sink

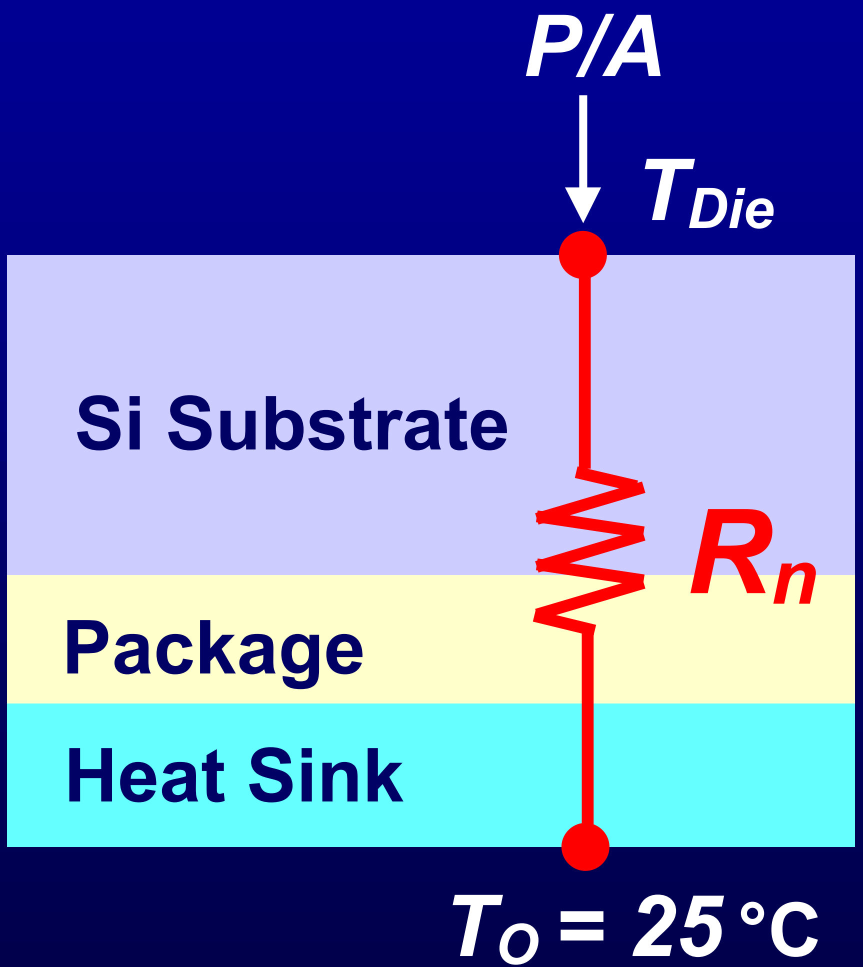
- Dynamic Power: $\propto CV^2f$ ⟶ most significant
- Leakage Power: increasing with scaling
- C dominated by interconnects
- Affects interconnect temperature

Interconnects: Away from Heat Sink

- Joule Heating: I^2R

K. Banerjee, M. Pedram and A. H. Ajami

Chip Temperature Model



- 1-D Heat Conduction

$$T_{Die} = T_O + R_n\left(\frac{P}{A}\right)$$

- T_{Die} = 120 °C (180 nm Node)
- R_n = 4.75 cm^2 °C/W

- Assuming same Packaging and Cooling Technologies (Same R_n)

→ T_{Die} at Other Technology Nodes Calculated

K. Banerjee, M. Pedram and A. H. Ajami

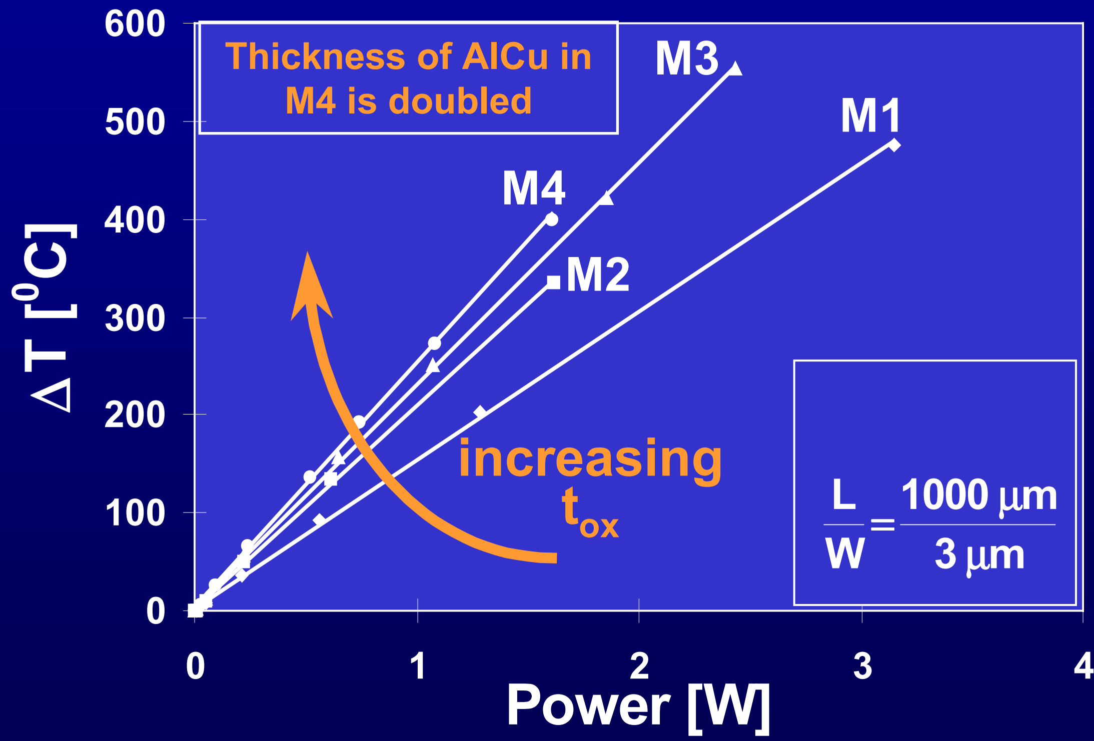
Thermal Effects in Interconnects

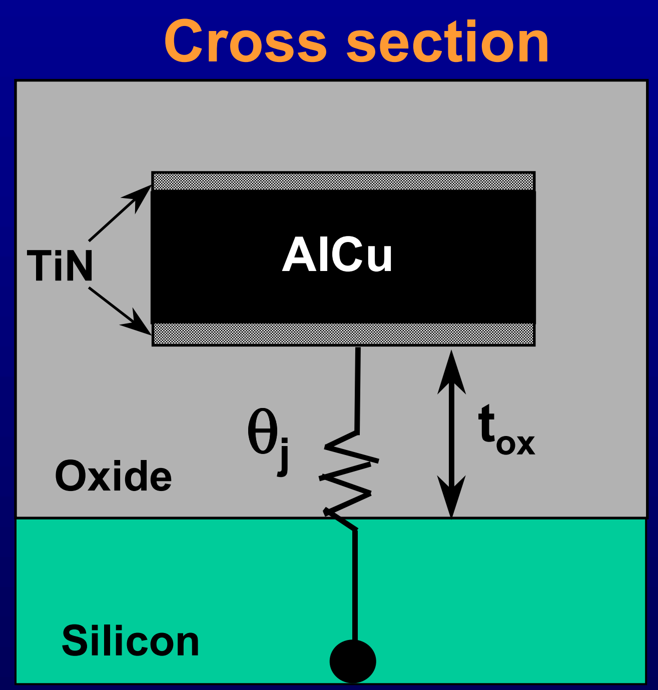
- An inseparable aspect of electrical power distribution and signal transmission through the interconnects
- Arise due to self-heating (or Joule heating) of interconnects caused by current flow
- Thermal effects impact interconnect electromigration reliability and design

K. Banerjee, M. Pedram and A. H. Ajami

Thermal Effects in Interconnects

Self Heating under DC Stress *(IRPS 96)*





Thermal impedance θ_j, defined by $\Delta T = P \times \theta_j$

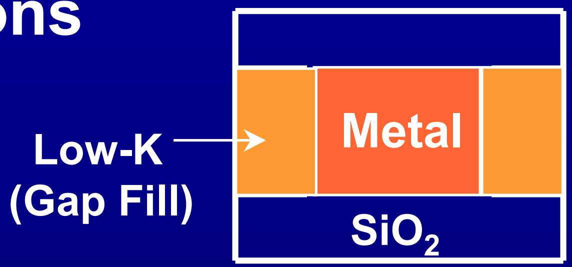
- ΔT increases with increasing t_{ox}

K. Banerjee, M. Pedram and A. H. Ajami

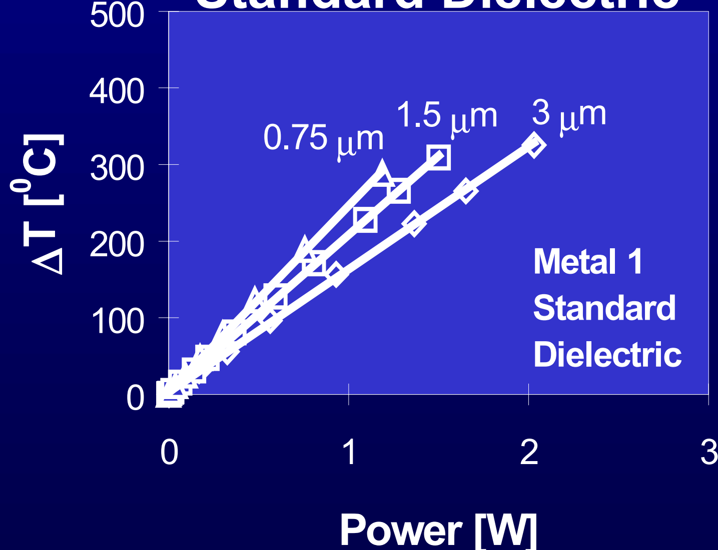
Thermal Effects in Interconnects

Impact of Scaling Using Low-k (*IEDM 96*)

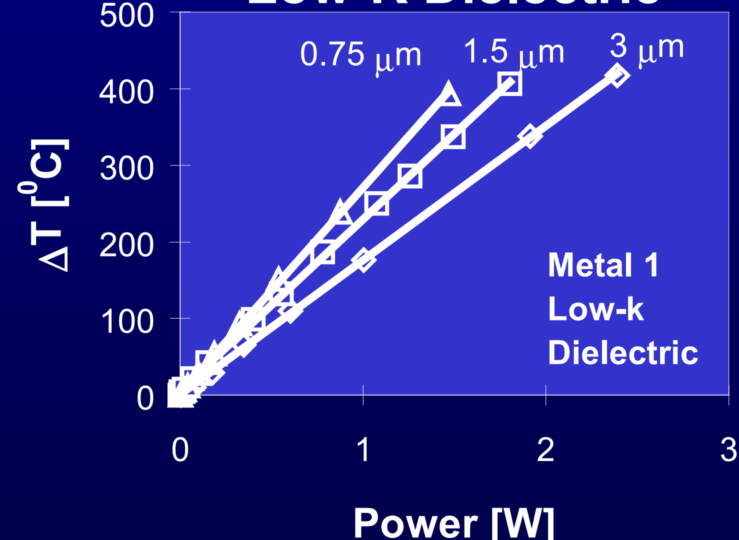
DC Conditions



Standard Dielectric



Low-K Dielectric



- As W decreases SH increases.
- Low-k increases SH by 10-15%

K. Banerjee, M. Pedram and A. H. Ajami

Scaling Trends and Implications

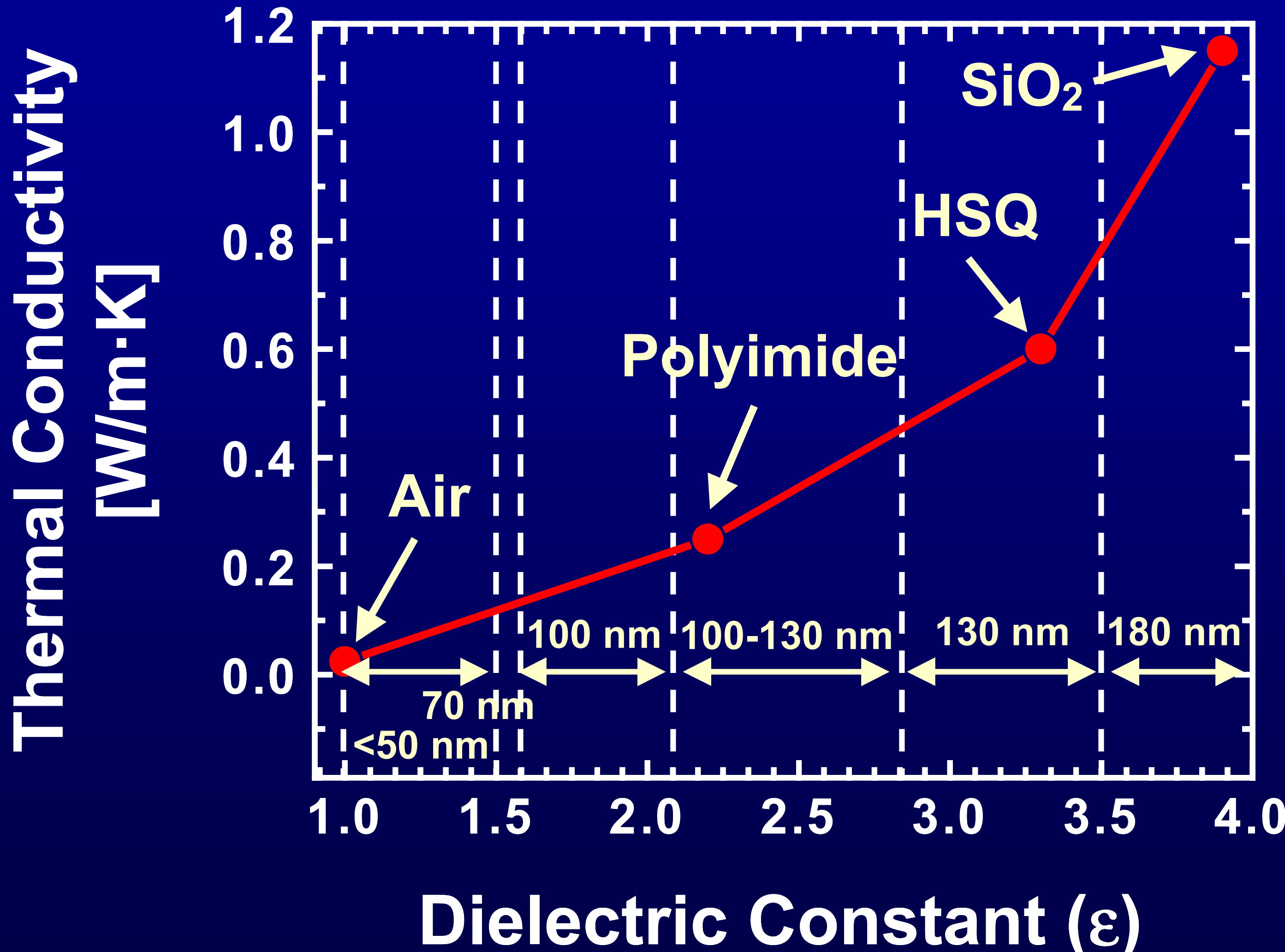
- **Scaling Effects (ITRS '99)**
  - Chip Power and Area increases
  - Negligible Change in Power Density
  - Current Density in Metal Lines Increases
  - Number of Metal Levels Increases

→ **Chip Temperature Distribution ?**

- **As Temperature Increases**
  - Electromigration (EM) Time to Failure Decreases
  - Increased ρ (T) → Wire Delay Increases

K. Banerjee, M. Pedram and A. H. Ajami

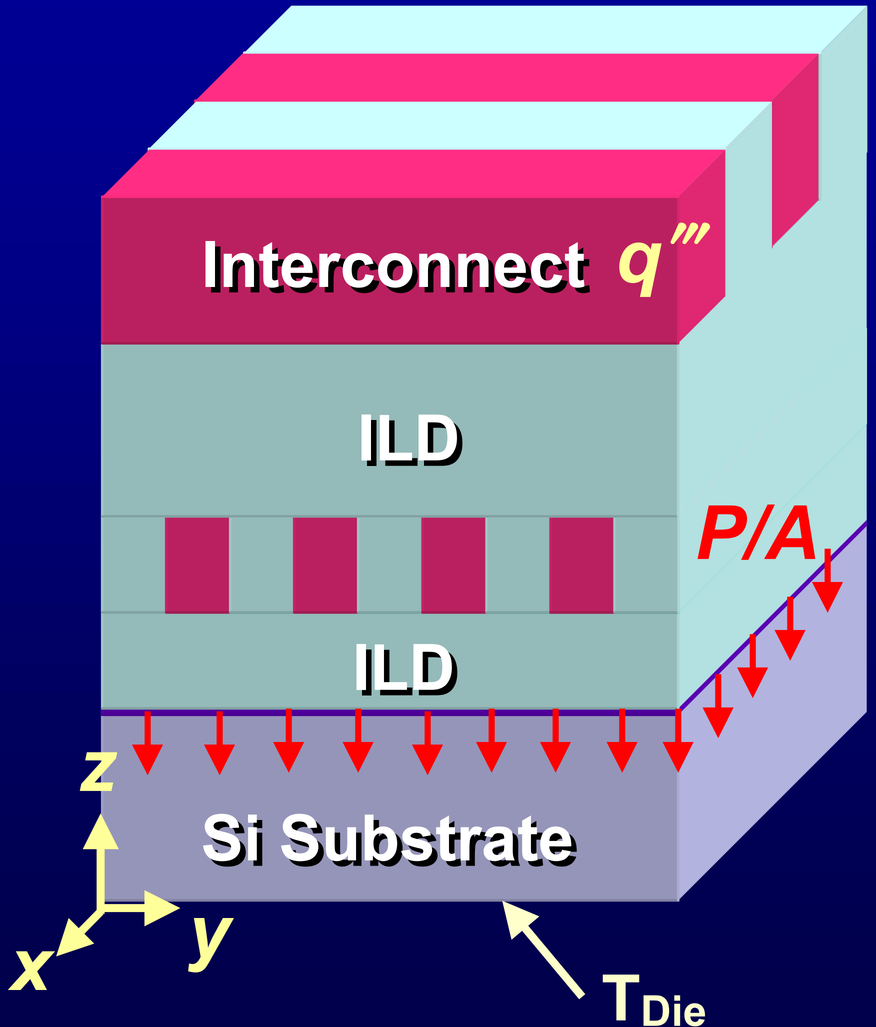
Scaling Effects (1) :
Thermal Conductivity of Dielectrics



(ITRS '99)

K. Banerjee, M. Pedram and A. H. Ajami

Full Chip Thermal Analysis



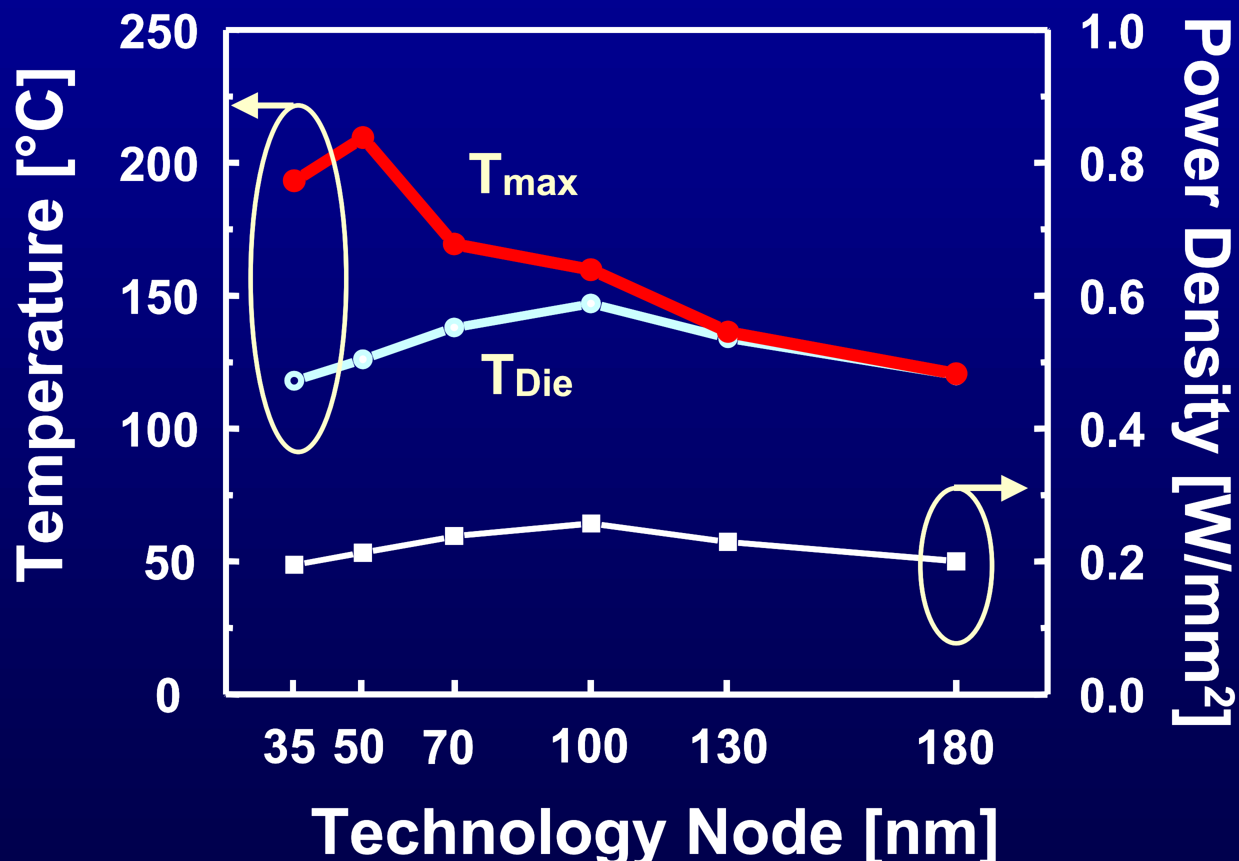
- **Three Dimensional Heat Conduction**
  - Steady State, Uniform Heat Generation (q'''), Constant Properties (k)

$$\nabla^2 T + \frac{q'''}{k} = 0, \qquad \nabla^2 T = 0$$

(Interconnect) (Others)

- **Worst Case Simulation**
  - Uniform j_{rms} for all Metal Lines *(ITRS '99)*

Scaling Effects (2) : Maximum Chip Temperature *(IEDM 2000)*

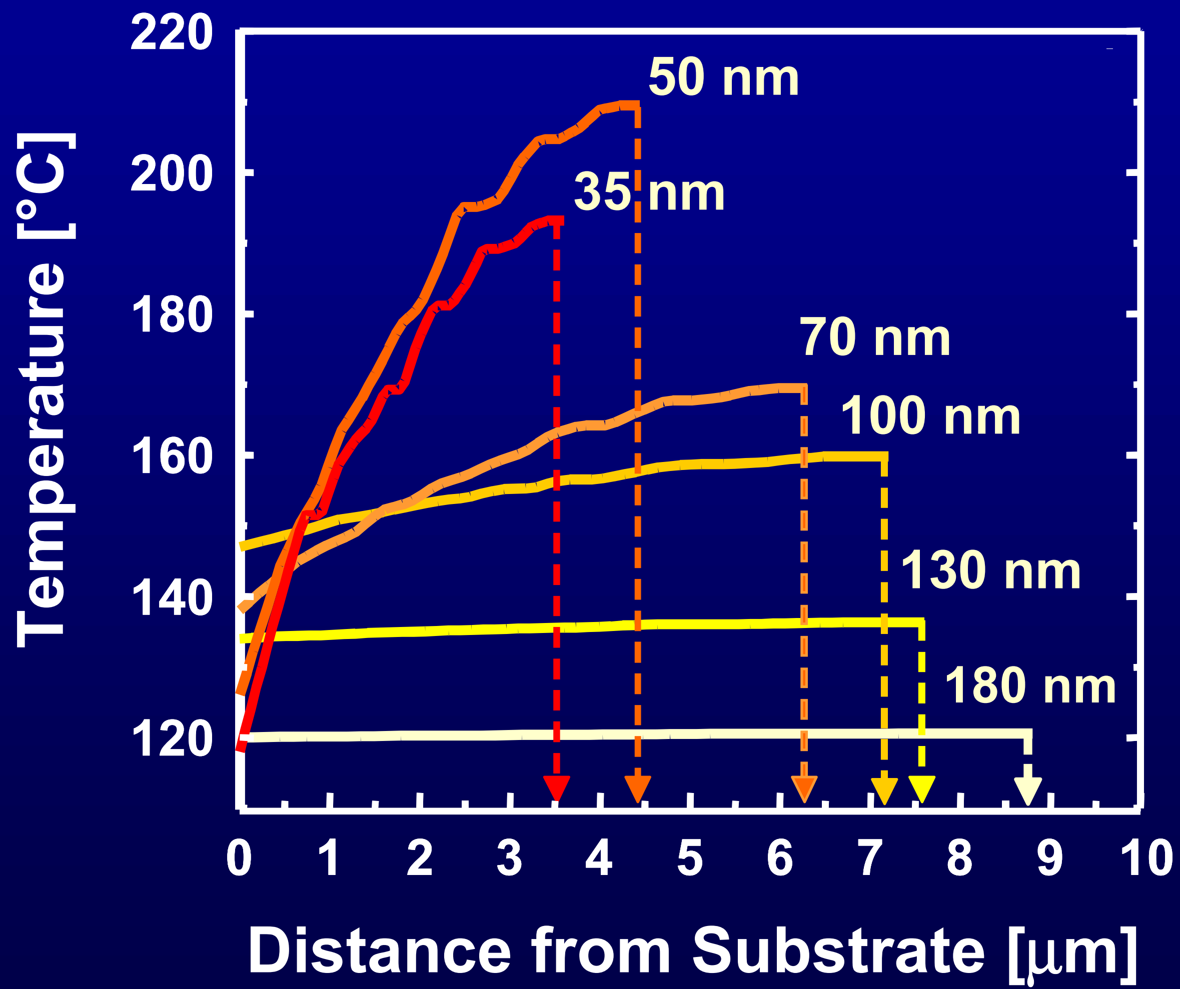


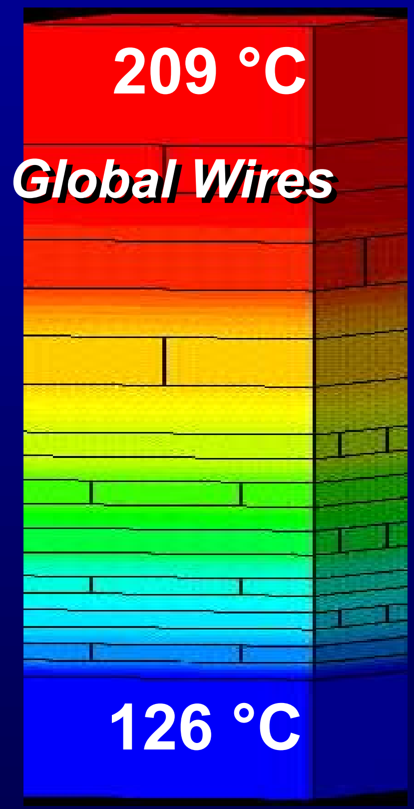
- **Negligible Change in Power Density**
  (ITRS '99)
- T_{Die} **= 133±15°C**
- **Increase in** T_{max} **Due to Joule Heating of Interconnects**

(FEM Simulation)

K. Banerjee, M. Pedram and A. H. Ajami

Scaling Effects (3) :
Temperature Distribution *(IEDM 2000)*

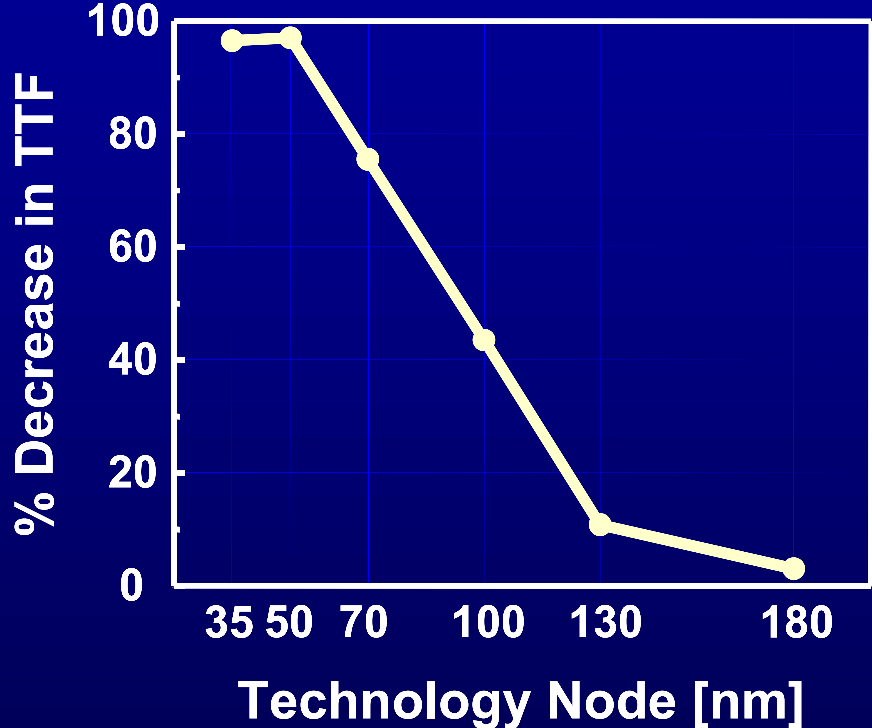


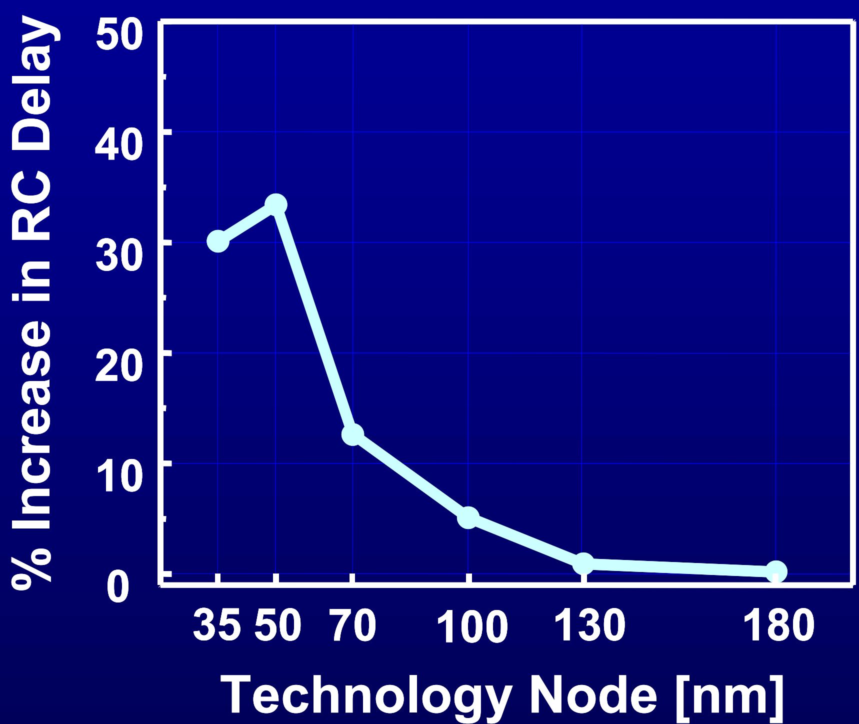


50 nm Node

K. Banerjee, M. Pedram and A. H. Ajami

Scaling Effects (4) : Effects on Reliability & Performance *(IEDM 2000)*





$$\%D\ \ in\ \ TTF = \left[1 - \frac{TTF(T_{max})}{TTF(T_{Die})}\right] \times 100$$

$$\%I\ \ in\ \ RC\ Delay = \left[\frac{\rho(T_{max})}{\rho(T_{Die})} - 1\right] \times 100$$

K. Banerjee, M. Pedram and A. H. Ajami

Scaling Trends and Implications (Summary)

- Scaling trends that cause increasing thermal effects:
  - increasing interconnect levels
  - increasing current density
  - low-k dielectrics
  - increasing thermal coupling

K. Banerjee, M. Pedram and A. H. Ajami

Presentation Outline

- Introduction
- Thermal Effects and Reliability
- Interconnect Performance Optimization
- High-Current Effects: ESD
- Analysis of Non-uniform Chip Temperature
- Temperature Dependent Performance
- Circuit Optimization: Clock Skew
- Summary

K. Banerjee, M. Pedram and A. H. Ajami

Reliability Implications

Electromigration (EM)

- Transport of mass in metal interconnects under an applied current density
- EM lifetime reliability modeled using Black's equation given by,

$$TTF = A\, j^{-n} \exp\left(\frac{Q}{k_B T_m}\right)$$

TTF is the time-to-fail

A is a constant that depends on line geometry and microstructure

j is the DC or average current density

Q is the activation energy for EM (~ 0.7 eV for AlCu)

T_m is the metal temperature

Reliability Implications

Typical EM Analysis

- Accelerated EM stress data yields *A, Q*, and *n* in Black's equation, and a value of log-normal σ_{LN}
- Typical goal: achieve a 10 year lifetime
- EM stress data + Black's equation gives a technology limit to the maximum allowed current density (j_{avg}) for the required failure rate and a desired lifetime at a reference temperature T_{ref} (~ 100 ^{0}C)
- The j_{avg} limit does not comprehend self heating

K. Banerjee, M. Pedram and A. H. Ajami

Reliability Implications

Current Density Definitions (*DAC 99)*

- **Peak, Average, and RMS current densities:**

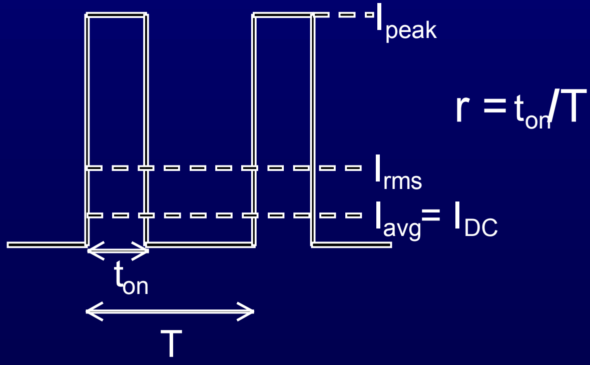
$$j_{peak} = \frac{I_{peak}}{A} \qquad j_{avg} = \frac{1}{T}\int_0^T j(t)\,dt \qquad j_{rms} = \sqrt{\frac{1}{T}\int_0^T j^2(t)\,dt}$$

A **is the cross sectional area of interconnect,** ***T*** **is the time period of the current waveform.**

- **For an unipolar waveform:**

$$j_{avg} = r\, j_{peak} \qquad j_{rms} = \sqrt{r}\, j_{peak}$$

r **is the duty factor**



- **EM is determined by j_{avg}, and self-heating by j_{rms}**

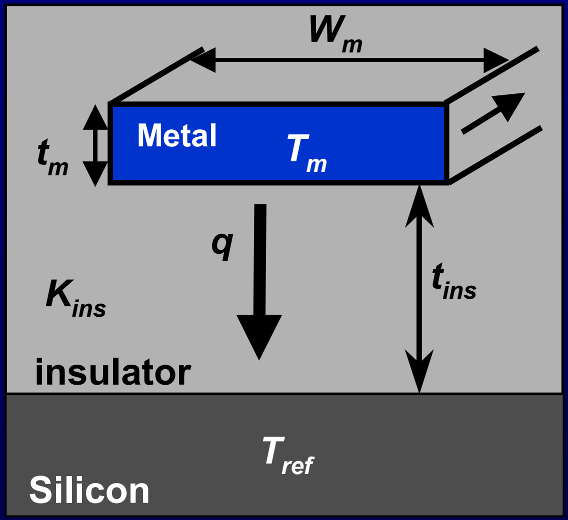
Reliability Implications

Impact of Self-Heating on EM *(DAC 99)*

- EM lifetime given by: $TTF = A\, j^{-n} \exp\left(\frac{Q}{k_B T_m}\right)$

- Due to self-heating: $T_m = T_{ref} + \Delta T_{self\text{-}heating}$

$$\Delta T_{self-heating} = (T_m - T_{ref}) = I_{rms}^2 R R_\theta$$



R_θ is the effective thermal impedance given by,

$$R_\theta = \frac{t_{ins}}{K_{ins} L W_{eff}}$$

W_{eff} is the effective metal width to account for quasi-2D heat conduction.

K. Banerjee, M. Pedram and A. H. Ajami

Reliability Implications

Self-Consistent Design *(Hunter 97)*

- Typically, design rules specify j_{avg} from EM and j_{rms} from self-heating separately.
- Self-consistent approach: comprehends EM and self-heating simultaneously.
- The lifetime at any j_{avg} and metal temperature T_m, should be equal to or greater than the lifetime value (e.g., 10 year) under the design rule current density (j_0).

$$\frac{exp\left(\frac{Q}{k_B T_m}\right)}{j_{avg}^2} \geq \frac{exp\left(\frac{Q}{k_B T_{ref}}\right)}{j_0^2}$$

K. Banerjee, M. Pedram and A. H. Ajami

Reliability Implications

Self-Consistent Equation

- Using the relationship between j_{avg}, j_{rms}, j_{peak} and r for an unipolar waveform described earlier, it can be shown that,

$$\frac{j_{avg}^2}{j_{rms}^2} = r$$

- Incorporating the j^2_{rms} and j^2_{avg} values from yields the self-consistent equation,

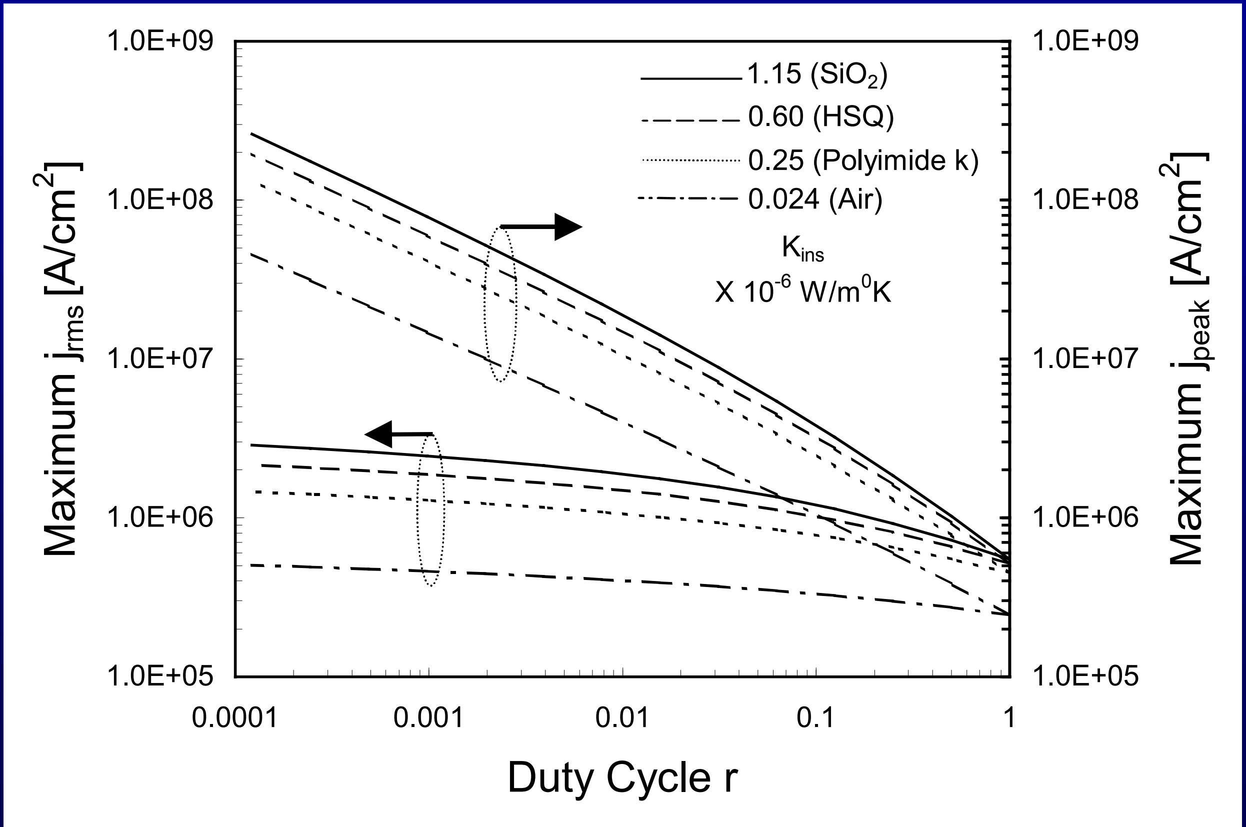
$$r = j_0^2 \frac{\exp\left(\frac{Q}{k_B T_m}\right)}{\exp\left(\frac{Q}{k_B T_{ref}}\right)} \frac{t_{ins}\, t_m\, W_m\, \rho_m(T_m)}{(T_m - T_{ref}) K_{ins}\, W_{eff}}$$

This is a single equation in the single unknown temperature T_m

K. Banerjee, M. Pedram and A. H. Ajami

Reliability Implications

Low-k/Cu: Implications for Current Density Limits

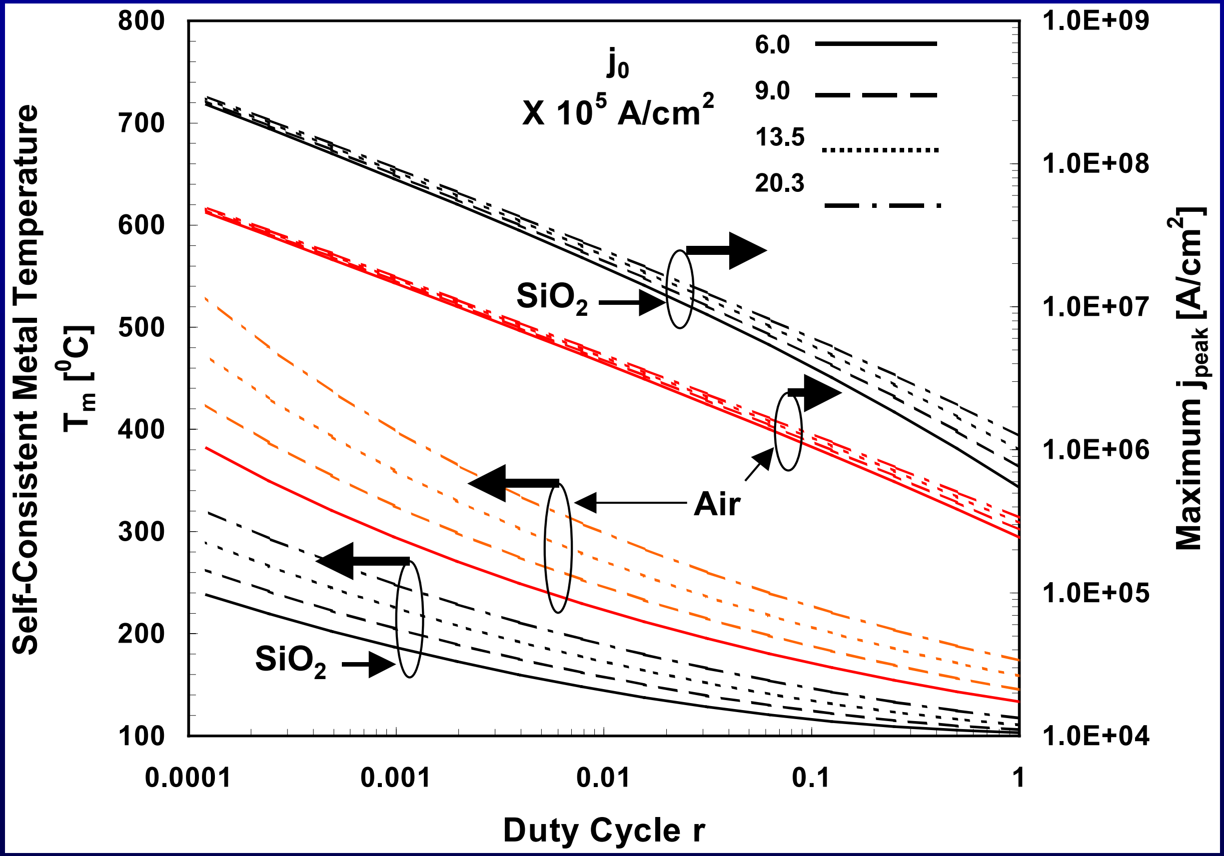


- Self-consistent j_{rms} and j_{peak} decrease significantly as low-k materials are introduced.

K. Banerjee, M. Pedram and A. H. Ajami

Reliability Implications

Implications for Interconnect Technology



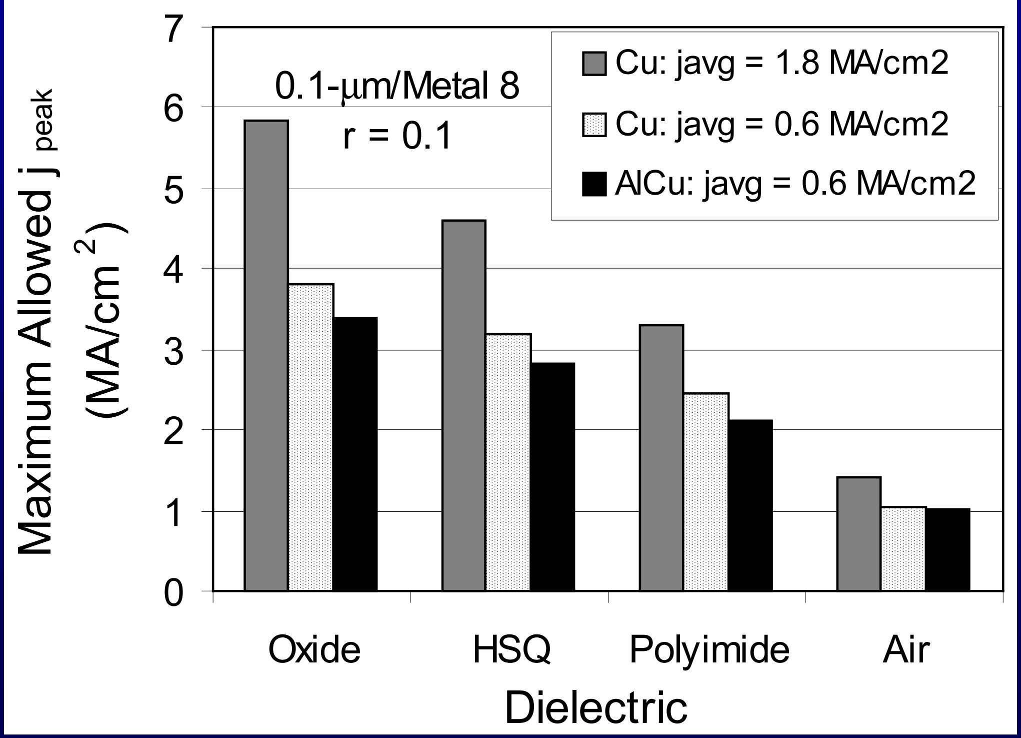
- As *r* decreases, material changes (increasing j_0) will become ineffective in increasing j_{peak}.

K. Banerjee, M. Pedram and A. H. Ajami

Reliability Implications

Implications for Current Density Limits

- Comparison with AlCu



- Thermal effects reduce the advantage of Cu as low-k materials are introduced

K. Banerjee, M. Pedram and A. H. Ajami

Presentation Outline

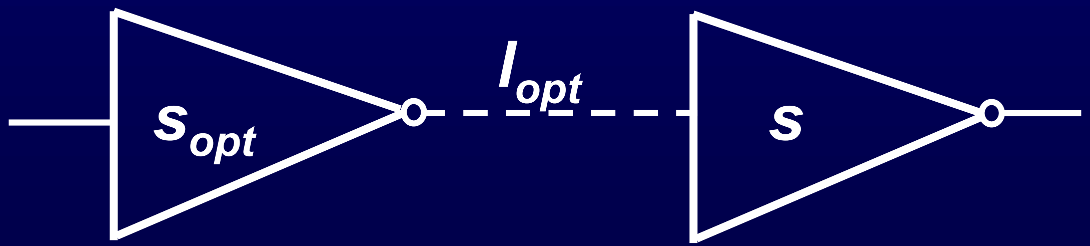
- Introduction
- Thermal Effects and Reliability
- Interconnect Performance Optimization
- High-Current Effects: ESD
- Analysis of Non-uniform Chip Temperature
- Temperature Dependent Performance
- Circuit Optimization: Clock Skew
- Summary

K. Banerjee, M. Pedram and A. H. Ajami

Performance Optimization

Semi-Global and Global Wires

- Thermal effects predominant in semi-global and global interconnects which are:
  - Away from the Si substrate
  - Long
  - Typically split into buffered segments
- Long interconnects can be optimally buffered.

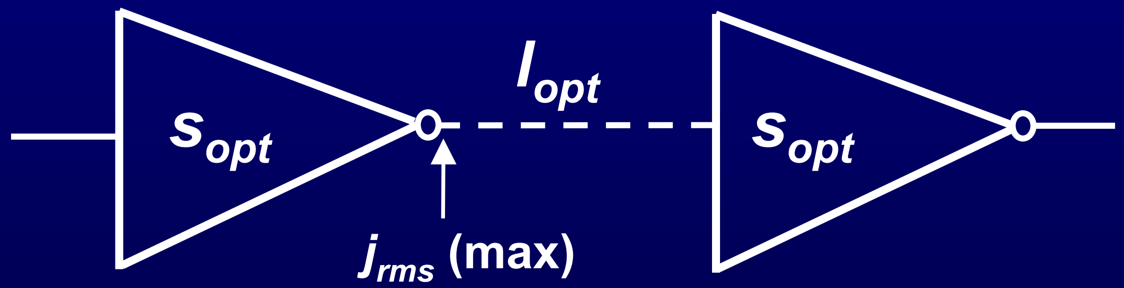


K. Banerjee, M. Pedram and A. H. Ajami

Performance Optimization

Performance Based Current Density (Signal Lines)

- 0.25 μm and 0.1 μm technology
- Full 3-D Interconnect capacitance extracted
- Accurate l_{opt} and s_{opt} values determined by SPICE simulations

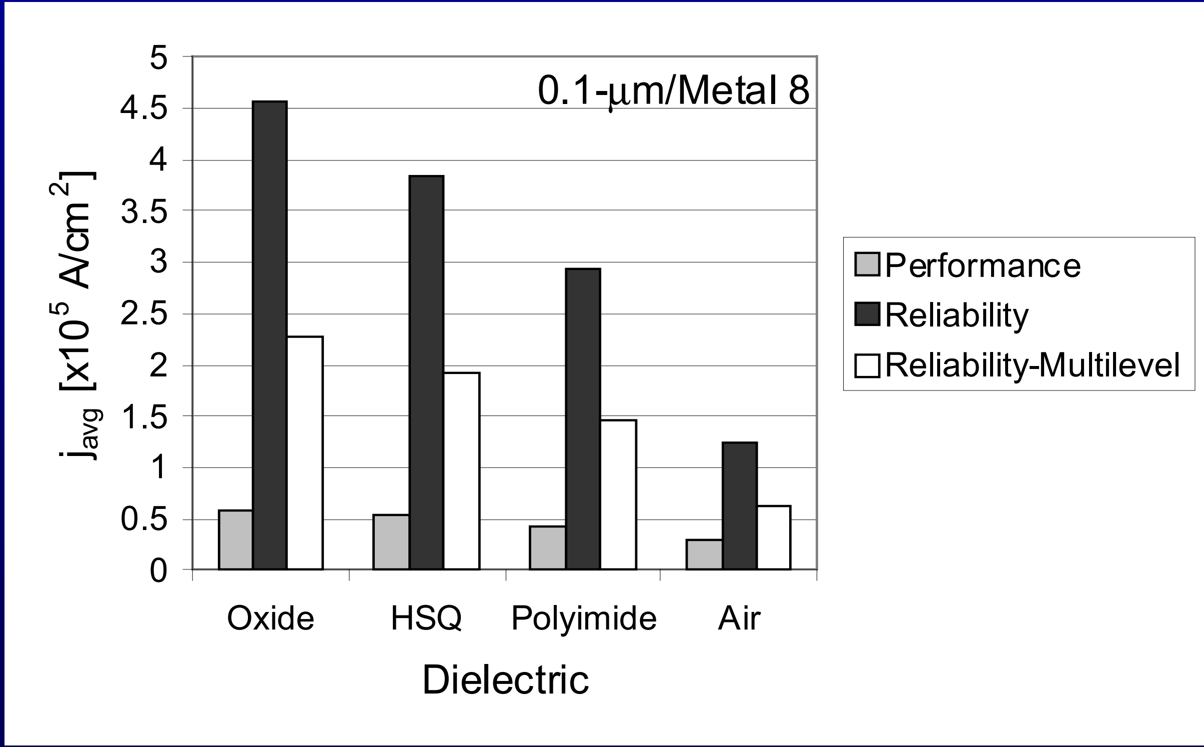


- j_{rms} (max) occurs close to the repeater output due to the distributed nature of the interconnect.

K. Banerjee, M. Pedram and A. H. Ajami

Performance vs Reliability

- Effect of Thermal coupling Included (NTRS Based)



Gets Worse for ITRS Data

- For point-to-point interconnects reliability design limits satisfied even after considering thermal coupling

Presentation Outline

- Introduction
- Thermal Effects and Reliability
- Interconnect Performance Optimization
- High-Current Effects: ESD
- Analysis of Non-uniform Chip Temperature
- Temperature Dependent Performance
- Circuit Optimization: Clock Skew
- Summary

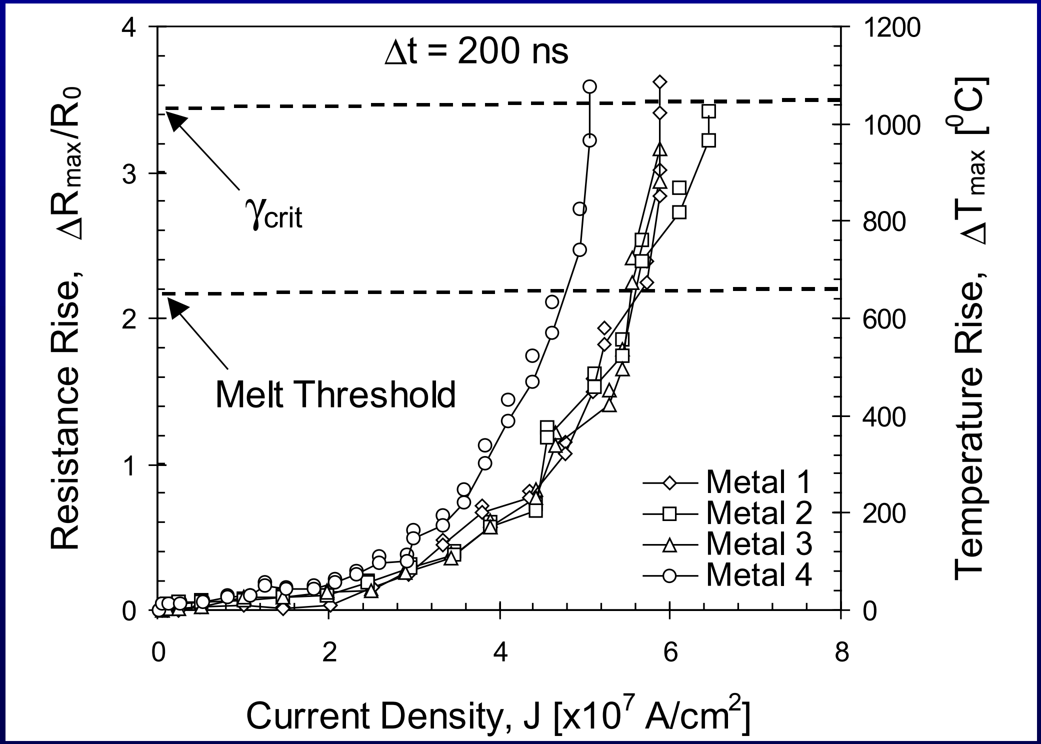
High-Current Effects

Non Steady-State Scenarios

- Electrostatic Discharge (ESD)
  - A short duration (< 200 ns), high current (> 1 A) event
  - Can cause open circuit failure of metals and latent damage that impact EM reliability

Non Steady-State Self-Heating

- Self-heating characteristics of AlCu lines under short-pulse stress conditions *(Electron Device Letters 97)*



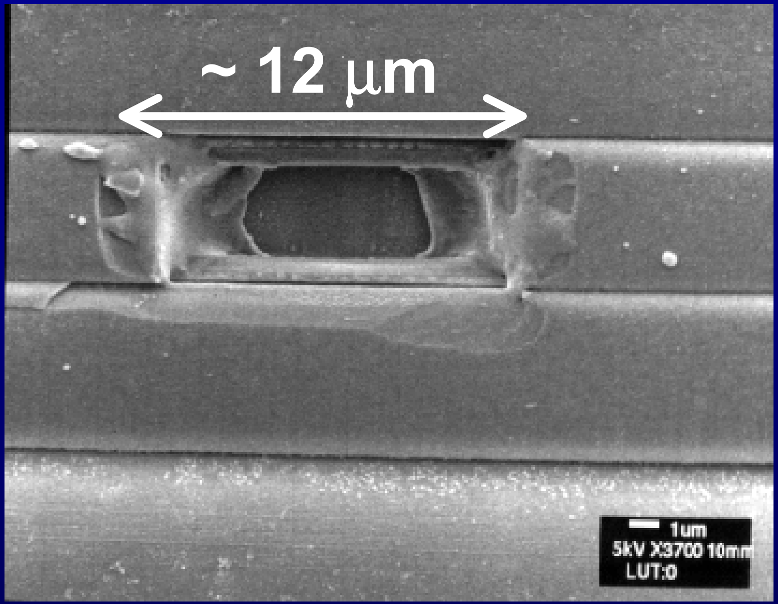
- Metal 1, 2, & 3 show **identical SH**
- Higher SH in Metal 4 is due to smaller surface area to volume ratio
- Interconnect failure temperature is **~ 1000 ^{0}C**

- Failure current densities are much higher than under normal circuit conditions

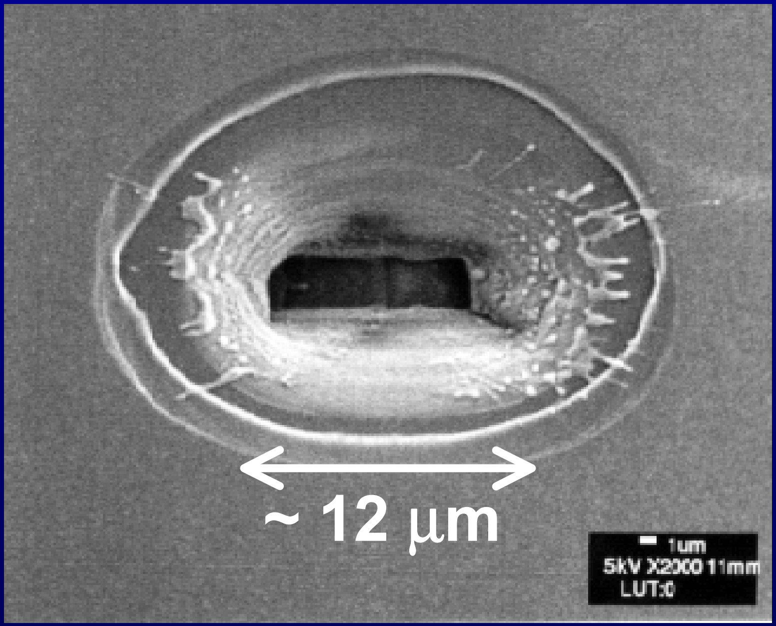
K. Banerjee, M. Pedram and A. H. Ajami

High-Current Effects

Open Circuit Failure *(IRPS 2000)*



Metal 4



Metal 1

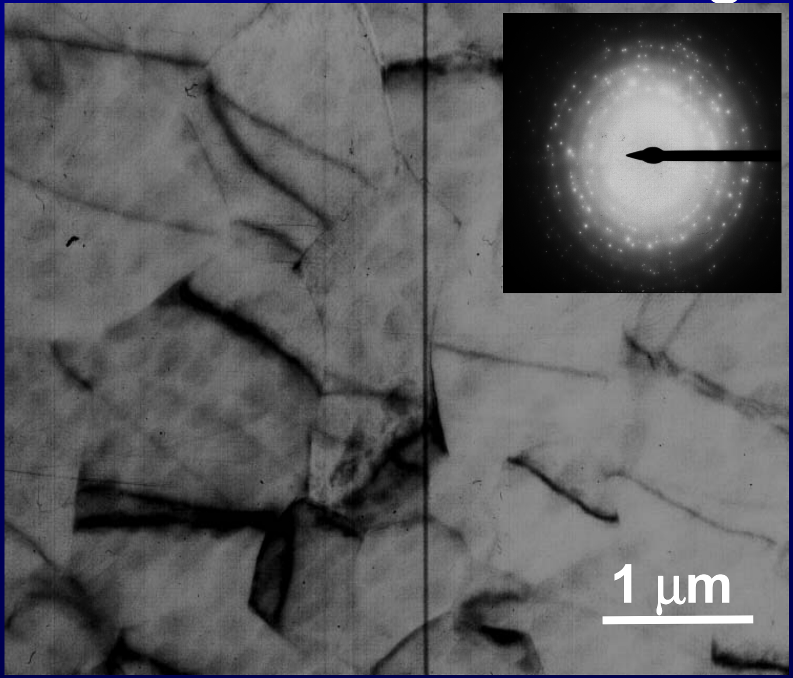
- Passivation fracture due to the expansion of critical volume of molten AlCu. (@ 1000 ^{0}C)
- Independent of overlying dielectric thickness.

K. Banerjee, M. Pedram and A. H. Ajami

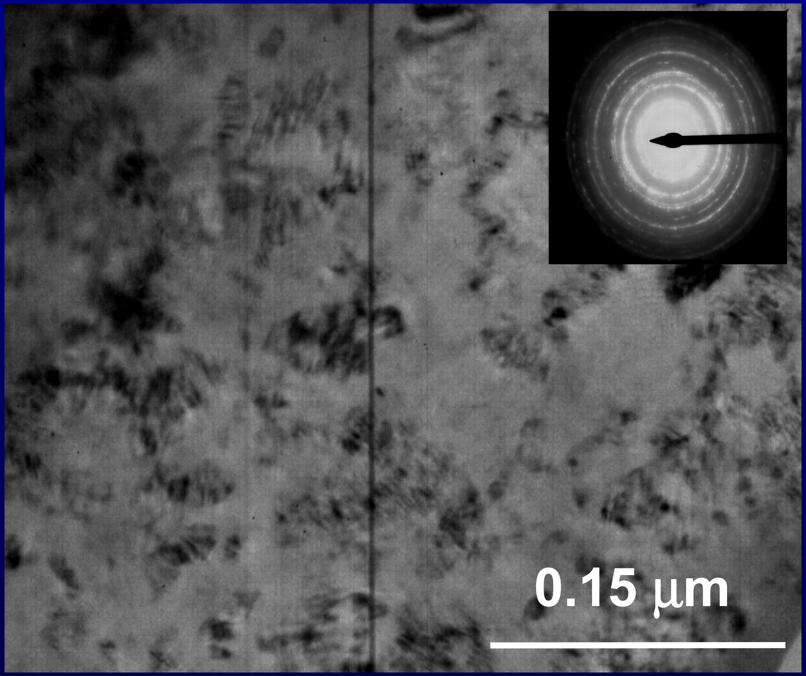
High-Current Effects

Latent Interconnect Damage *(IRPS 2000)*

Significant Electromigration Performance Degradation



Unstressed AlCu



Stressed AlCu

K. Banerjee, M. Pedram and A. H. Ajami

Summary (1)

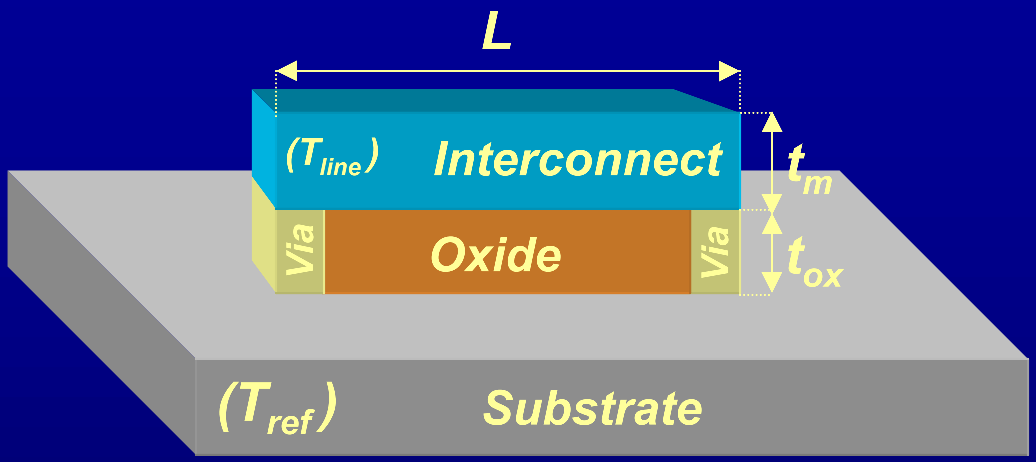
- Thermal Analysis including Interconnect Joule Heating based on ITRS '99
  - Peak Temperatures in ICs Increase with Technology Scaling in Spite of Constant Power Density
  - Significant Implications for Performance and Reliability
  - Advanced Chip Cooling Techniques may be Necessary
- Thermal Effects and Reliability
  - Thermal Effects Strongly Impacts EM
  - Self-Consistent Analysis: Thermal + EM
  - Point-to-Point Interconnects Optimized for Performance Meets Reliability Based Current Density Limits
  - High-Current Design Rules Must be Followed for I/O and ESD Protection Circuit Interconnects

K. Banerjee, M. Pedram and A. H. Ajami

Presentation Outline

- Introduction
- Thermal Effects and Reliability
- Interconnect Performance Optimization
- High-Current Effects: ESD
- Analysis of Non-uniform Chip Temperature
- Temperature Dependent Performance
- Circuit Optimization: Clock Skew
- Summary

Interconnect Temperature



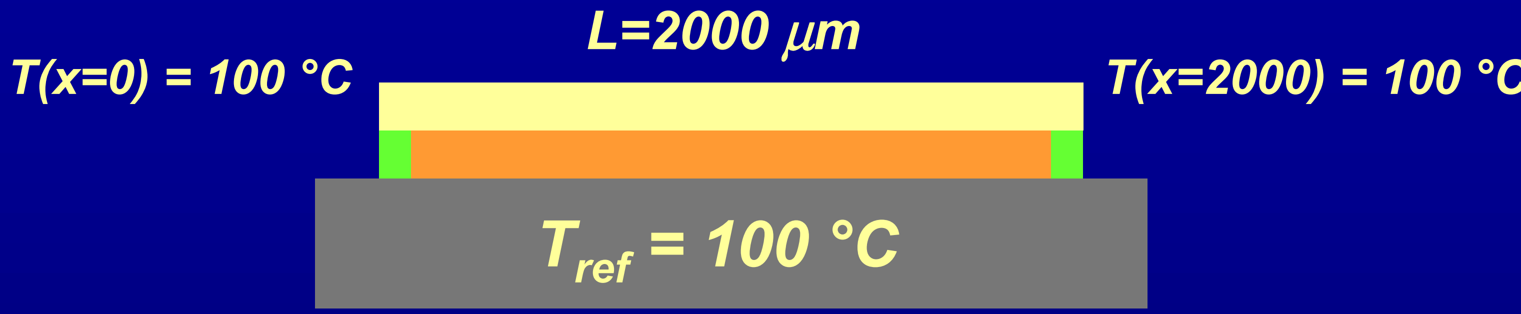
Heat equation in Interconnect (DAC 2001)

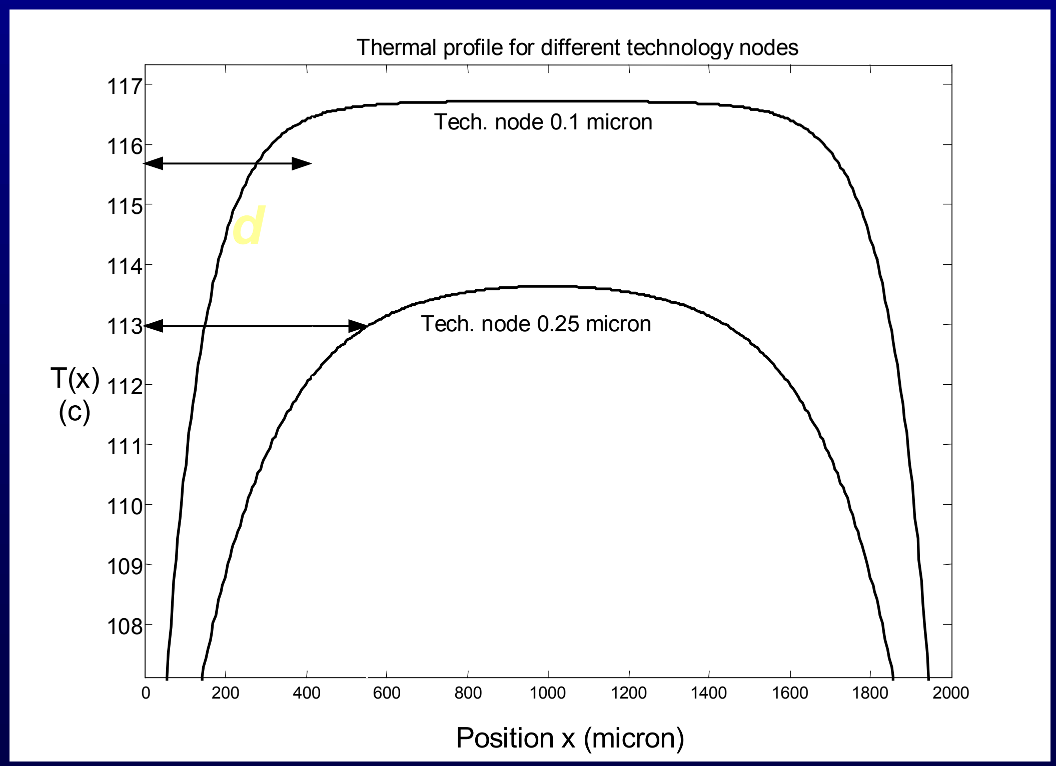
$$\frac{d^2 T_{line}}{dx^2} = -\frac{Q}{k_m} \Rightarrow \frac{d^2 T_{line}(x)}{dx^2} = \lambda^2 T_{line}(x) - \lambda^2 T_{ref}(x) - \theta$$

λ and θ are constants

K. Banerjee, M. Pedram and A. H. Ajami

Solution to Heat Equation

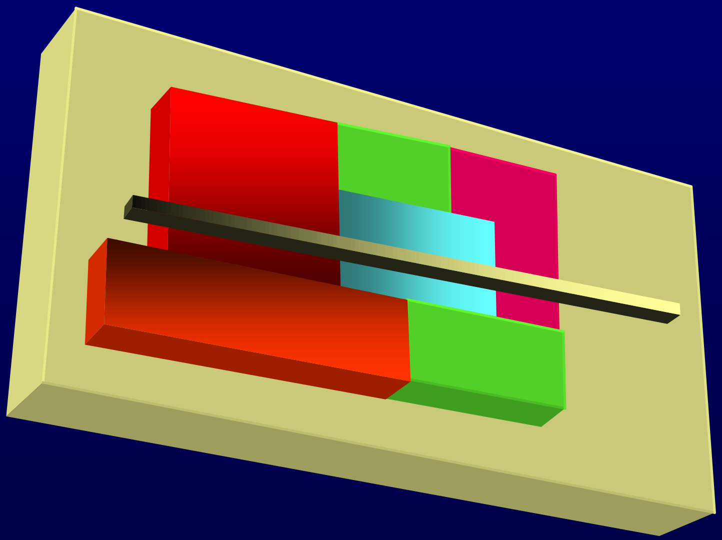




K. Banerjee, M. Pedram and A. H. Ajami

Non-Uniform Substrate Temperature

- Due to different switching activities, substrate temperature is generally non-uniform.
  - DPM, Functional block clock gating
  - Thermal time constant is much higher that signal propagation constant



Presentation Outline

- Introduction
- Thermal Effects and Reliability
- Interconnect Performance Optimization
- High-Current Effects: ESD
- Analysis of Non-uniform Chip Temperature
- Temperature Dependent Performance
- Circuit Optimization : Clock Skew
- Summary

Non-uniform Interconnect Thermal Profile

- Long global interconnects span large area
  - Experience substrate thermal non-uniformity with high probability
- Assuming a uniform substrate thermal profile results in delay estimation errors
  - Introduces error in wire-planning and optimization steps

Temperature Dependence of Resistance

- Resistance is dependent on Temeprature

$$r(x) = \rho_0 (1 + \beta \cdot T(x))$$

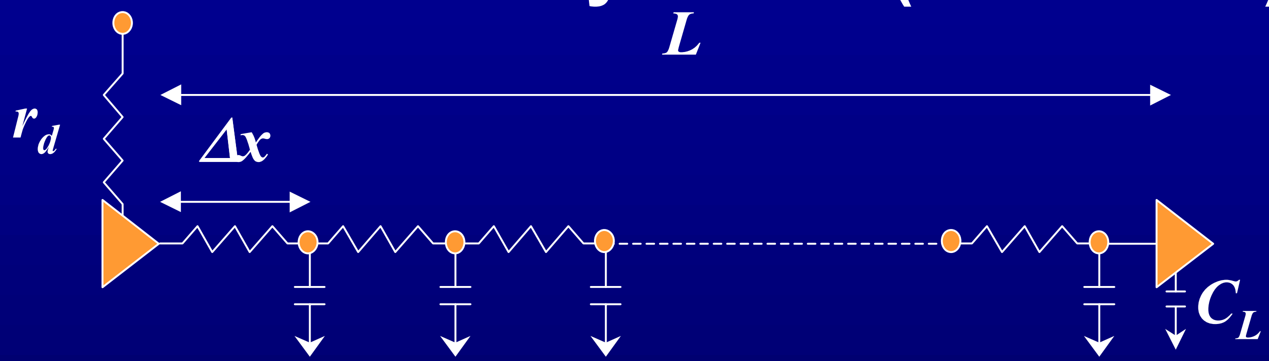
  - ρ_0 is the resistance per unit length at reference temperature
  - β is the temperature coefficient of resistance (1/°C)

- Non-Uniform line temperature ⇒ non-uniform resistance profile
  - Unit length capacitance is not affected

K. Banerjee, M. Pedram and A. H. Ajami

Non-Uniform Temperature Dependent Delay

- Distributed RC delay model (DAC 2001)



$$D = R_d(C_L + \int_0^L c_0(x)dx) + \int_0^L r_0(x)(\int_x^L c_0(\tau)d\tau + C_L)dx$$

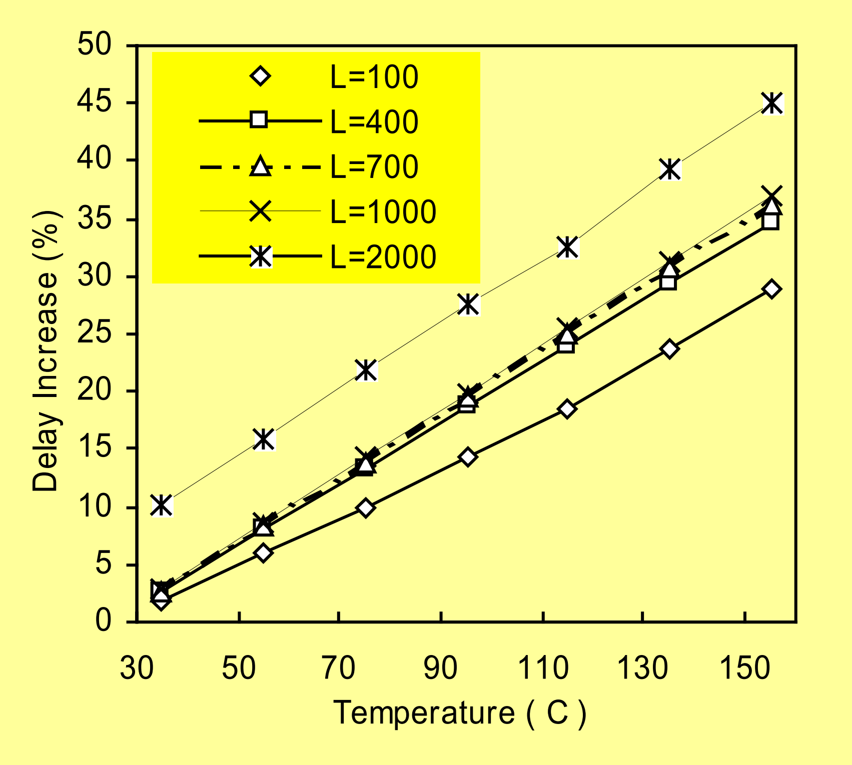
$$D = D_0 + (c_0L + C_L)\rho_0\beta\int_0^L T(x)dx - c_0\rho_0\beta\int_0^L xT(x)dx$$

D_0 is the Elmore delay model at 0 °C

K. Banerjee, M. Pedram and A. H. Ajami

Delay Degradation with Uniform Tref(x)

- With uniform thermal profile (0.25 μm):



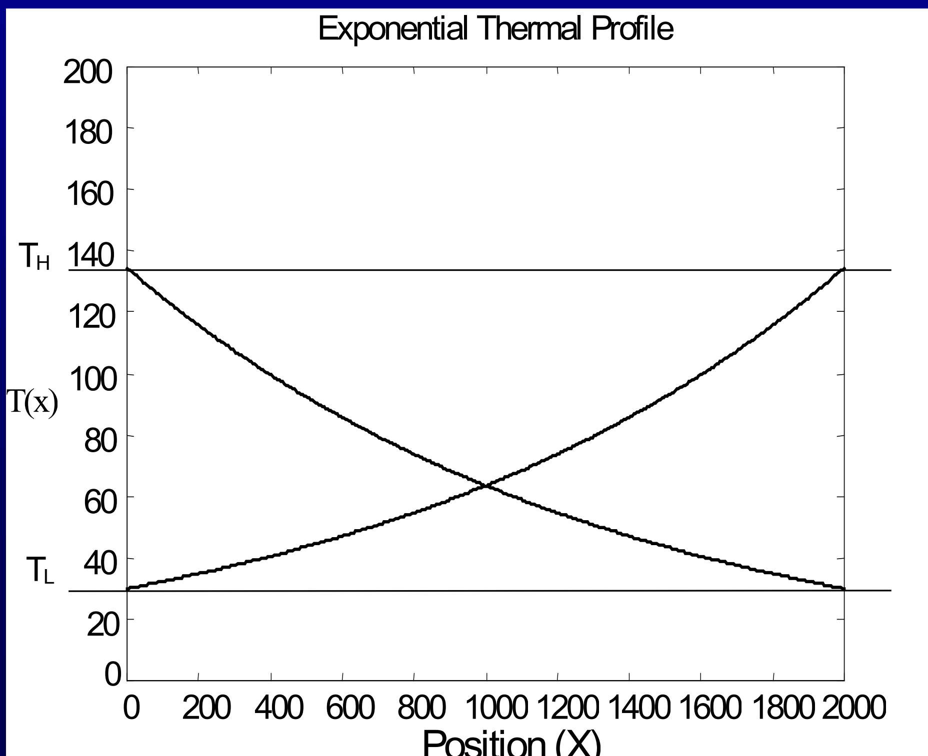
- 5-6% increase for each 20-degree increase in long global lines

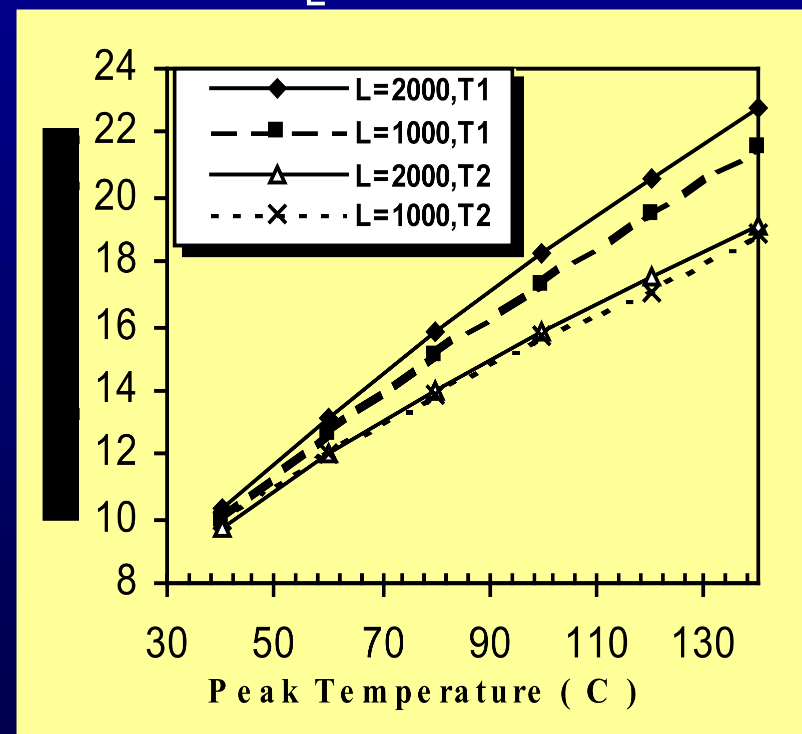
K. Banerjee, M. Pedram and A. H. Ajami

Delay Degradation with Non-uniform Tref(x)

- Effect of exponential thermal profiles:

T_L=30 °C



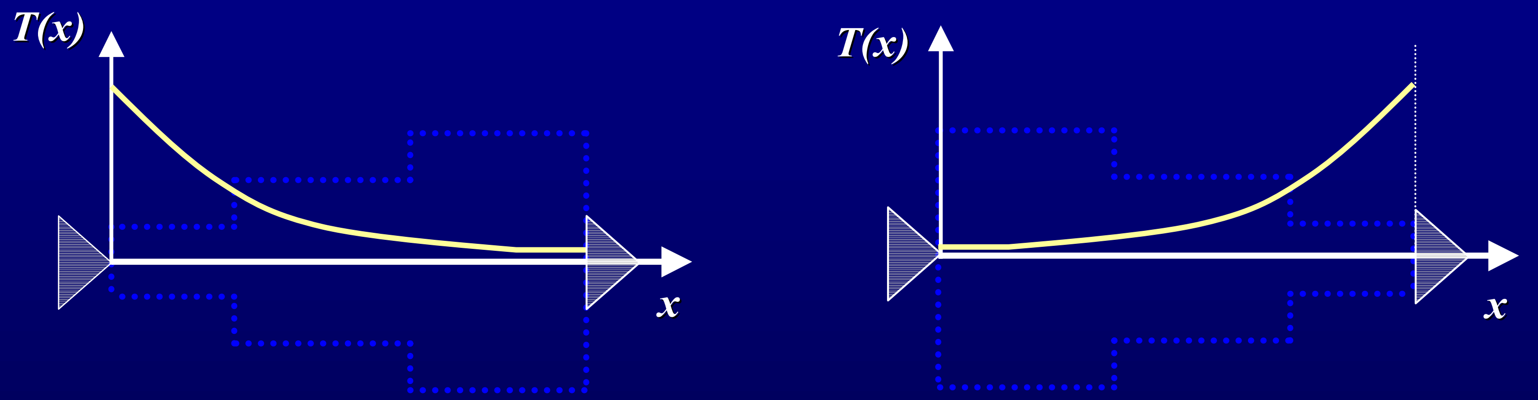


- Direction of Thermal Gradient is Important

K. Banerjee, M. Pedram and A. H. Ajami

Directional Thermal Profile

- Increasing (decreasing) thermal profile is equivalent to decreasing (increasing) sizing profile for uniform resistance wire



- Increasing thermal profile has better performance than that of decreasing thermal profile (optimal wire sizing)

K. Banerjee, M. Pedram and A. H. Ajami

Presentation Outline

- Introduction
- Thermal Effects and Reliability
- Interconnect Performance Optimization
- High-Current Effects: ESD
- Analysis of Non-uniform Chip Temperature
- Temperature Dependent Performance
- Circuit Optimization : Clock Skew
- Summary

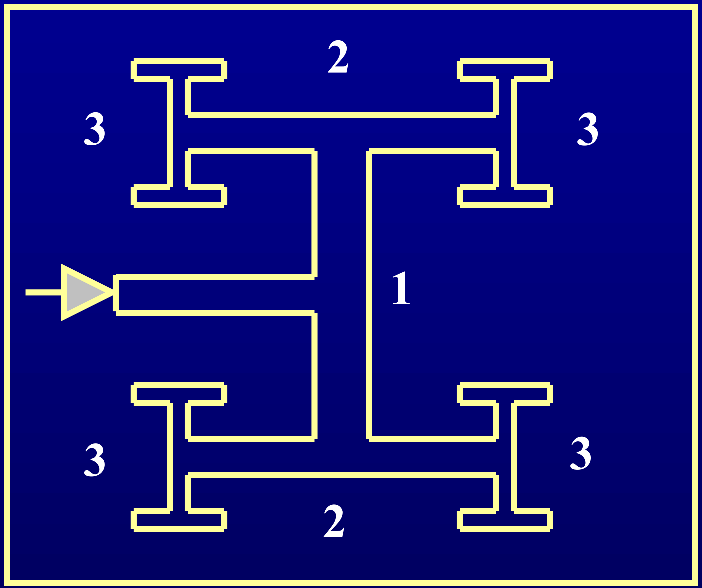
Clock Net Routing

- Clock is the most vulnerable signal to the underlying thermal non-uniformity
  - Have long global segments in the highest metal layers
  - delay variations affect skew

- Clock nets must have near-zero skew among their sinks to guarantee correct functionality of the circuits

H-Tree's

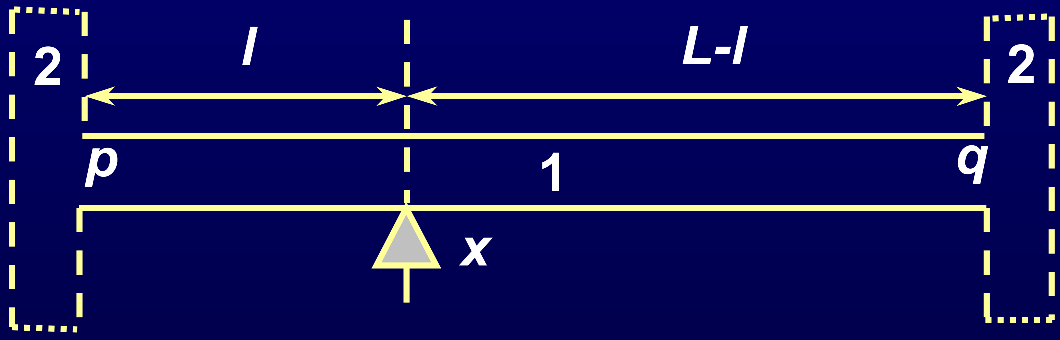
- H-Tree or bottom-up merging techniques



- Balancing loads seen at merging point in H-Tree to have zero-skew at two sides of each branch

Branching Point *(C/CC 2001)*

- Equal load at each sink: middle point is the branching point (*l*)
- With non-uniform thermal profile, branching point dependent on the profile



K. Banerjee, M. Pedram and A. H. Ajami

Branching Point cont'd

- Using thermally dependent delay, optimal branching location (l^*) is:

$$\beta \int_0^{l^*} T(x)\,dx + l^* - A = 0$$

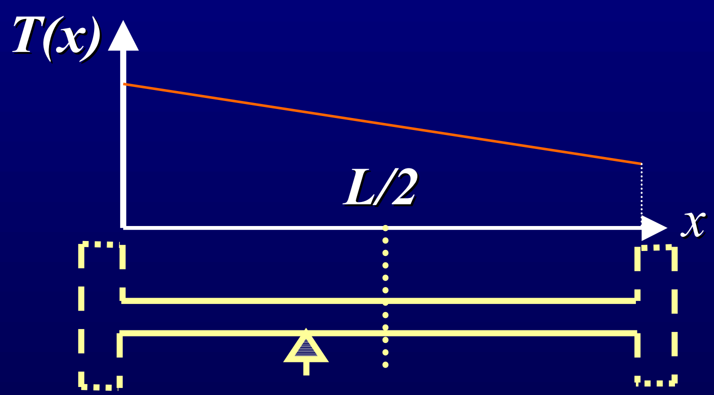
A is constant

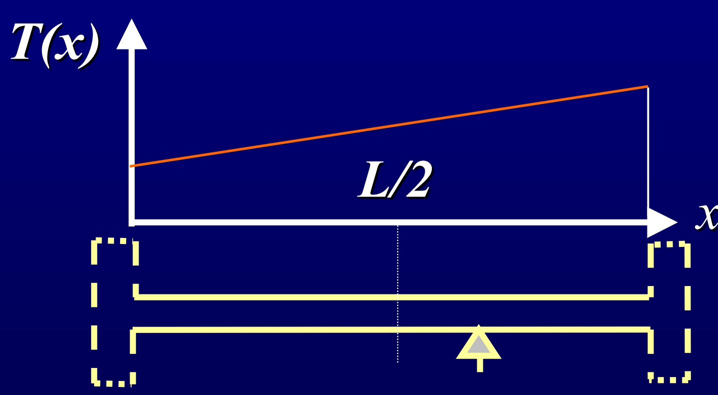
- With symmetric non-uniform thermal profile, the branching point is still at $l^*=L/2$

K. Banerjee, M. Pedram and A. H. Ajami

Movement of Branching Point

- In gradually decreasing (increasing) thermal profile, optimal length l^* has to be less than (greater than) $L/2$.





K. Banerjee, M. Pedram and A. H. Ajami

Thermally Dependent Merging

- Thermal non-uniformity can introduce a significant skew in the clock tree
- Thermally-dependent bottom-up merging must be used to minimize the skew

Results *(CICC 2001)*

$T_{line}(x)$	params	$l=l^*$	$l=L/2$ skew%
$T(x) = ax + b$, $a = \frac{T_H - T_L}{L}$, $b = T_L$	T_H=170, T_L=90	1042	5.42
	T_H=170, T_L=110	1032	3.98
	T_H=170, T_L=130	1021	2.65
$T(x) = a \cdot e^{-bx}$, $b = \frac{1}{L}\ln(\frac{T_H}{T_L})$, $a = T_H$	T_H=170, T_L=90	957.5	5.24
	T_H=170, T_L=110	968.66	3.63
	T_H=170, T_L=130	979.5	2.40
$T(x) = T_{max} \cdot e^{-\frac{-(x-\mu)^2}{2\sigma^2}}$	μ=2000, σ=1000	1210	7.78
	μ=1000, σ=400	1000	0.0
	μ=300, σ=700	911	9.57

K. Banerjee, M. Pedram and A. H. Ajami

Effects of Non-Uniform Temeprature on EDA Flow

- Interconnect non-uniform thermal profile can affect many EDA flow steps
  - Optimal layer assignment
  - Buffer insertion
  - Wire sizing
  - Gate sizing

Summary (2)

- Impact of Non-Uniform Substrate Temperature
  - Different switching activities in the substrate cause thermal gradients
  - Interconnect temperature is strongly dependent on substrate thermal profile
  - As technology scales, effect of substrate temperature becomes more important

Summary (2)

- Performance dependency
  - Delay model for non-uniform line temperature presented
  - Delay based on uniform worst case line temperature is not sufficient
  - Direction of thermal gradients is important
- Signal Integrity : Clock Skew
  - Non-uniform substrate temperature introduces skew in the clock tree
  - Bottom-up merging techniques must consider non-uniform interconnect thermal profile
  - Skew can be minimized by suitable merging