Analysis and Optimization of Thermal Issues in High Performance VLSI

Kaustav Banerjee
Stanford University

Massoud Pedram and Amir H. Ajami
University of Southern California

ISPD 2001

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**

ISPD 2001

Introduction

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

Sources of Chip Power Dissipation

Devices: Close to Heat Sink

- **Dynamic Power: $\propto CV^2f$ ⟶ most significant**

- **Leakage Power: increasing with scaling**

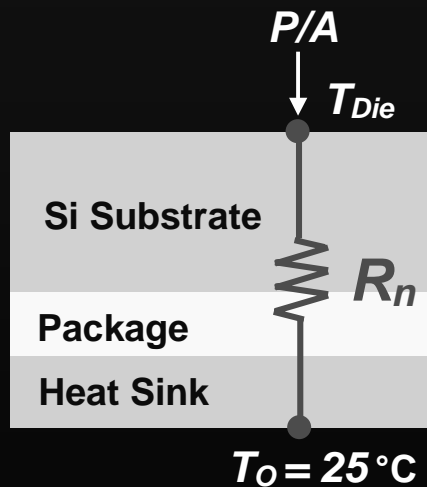- **C dominated by interconnects**

- **Affects interconnect temperature**

Interconnects: Away from Heat Sink

- **Joule Heating: I^2R**

Chip Temperature Model

P/A

1-D Heat Conduction

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

T_{Die}

Si Substrate

R_n

Package

Heat Sink

$T_O = 25\,°C$

- T_{Die} = 120 °C (180 nm Node)
- R_n = 4.75 cm² °C/W

- Assuming same Packaging and Cooling Technologies
  → (Same R_n)
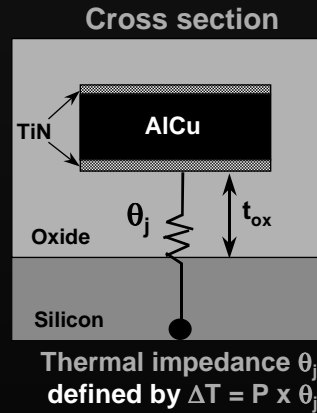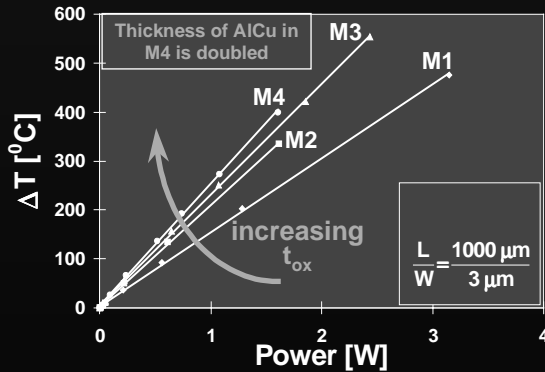 T_{Die} at Other Technology Nodes Calculated

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

Thermal Effects in Interconnects
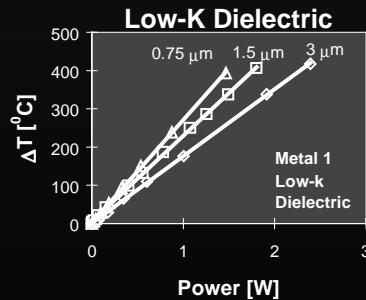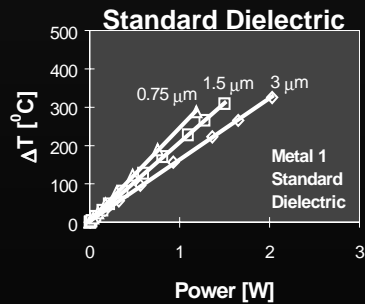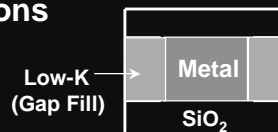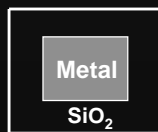
Self Heating under DC Stress *(IRPS 96)*

Thickness of AlCu in M4 is doubled

$\Delta T \, [^0C]$ — 0, 100, 200, 300, 400, 500, 600

M3, M1, M4, M2

increasing t_{ox}

$$\frac{L}{W} = \frac{1000 \, \mu m}{3 \, \mu m}$$

Power [W] — 0, 1, 2, 3, 4

Cross section

TiN — AlCu

θ_j — t_{ox}

Oxide

Silicon

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

- ΔT **increases** with increasing t_{ox}

Thermal Effects in Interconnects

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

DC Conditions

Metal

SiO_2

Low-K (Gap Fill) → Metal

SiO_2

Standard Dielectric

$\Delta T \, [^oC]$ — 0, 100, 200, 300, 400, 500

$0.75 \, \mu m$ $1.5 \, \mu m$ $3 \, \mu m$

Metal 1 Standard Dielectric

Power [W] — 0, 1, 2, 3

Low-K Dielectric

$\Delta T \, [^oC]$ — 0, 100, 200, 300, 400, 500

$0.75 \, \mu m$ $1.5 \, \mu m$ $3 \, \mu m$

Metal 1 Low-k Dielectric

Power [W] — 0, 1, 2, 3

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

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

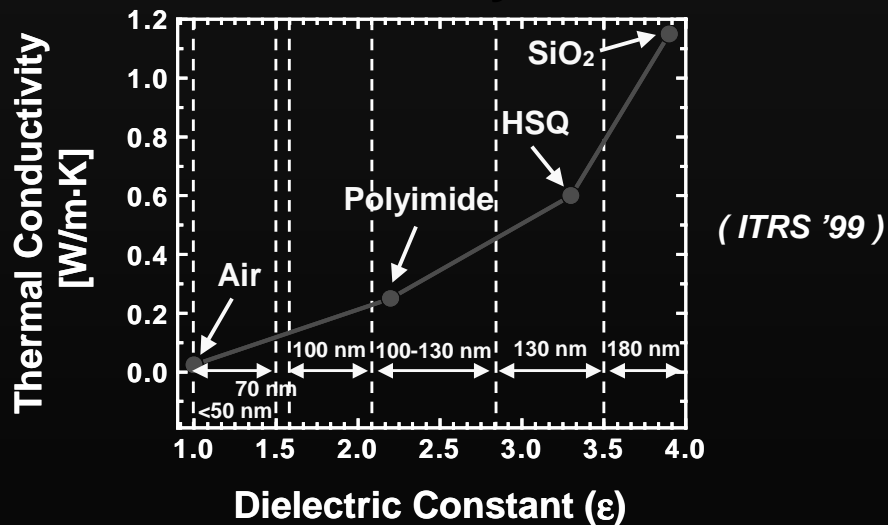   ⟶ **Chip Temperature Distribution ?**

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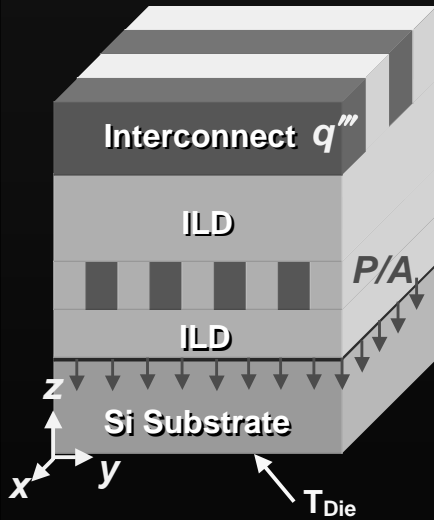 - Increased ρ (T) ⟶ Wire Delay Increases

ISPD 2001

Scaling Effects (1) :
Thermal Conductivity of Dielectrics

ISPD 2001

Full Chip Thermal Analysis

Interconnect q'''

ILD

ILD

Si Substrate

P/A

z

x y

T_{Die}

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

$$\nabla^2 T + \frac{q'''}{k} = 0, \qquad \nabla^2 T = 0$$
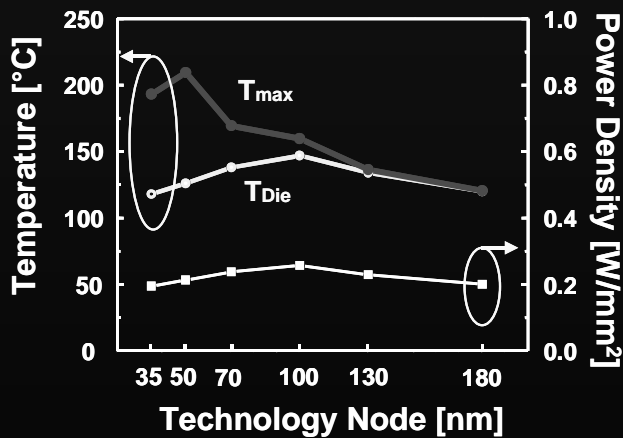**(Interconnect)   (Others)**

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

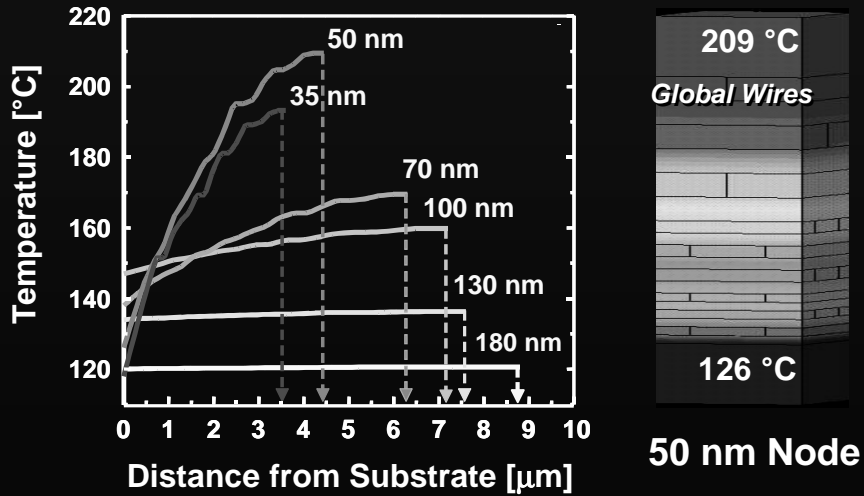# *Scaling Effects (2)* :
Maximum Chip Temperature *(IEDM 2000)*

Temperature [°C]

Power Density [W/mm²]

Technology Node [nm]

250 — 1.0

200 — 0.8

150 — 0.6

100 — 0.4

50 — 0.2

0 — 0.0

T$_{max}$

T$_{Die}$

35 50 70 100 130 180

- **Negligible Change in Power Density**
  (ITRS '99)

- T_{Die} **= 133±15°C**

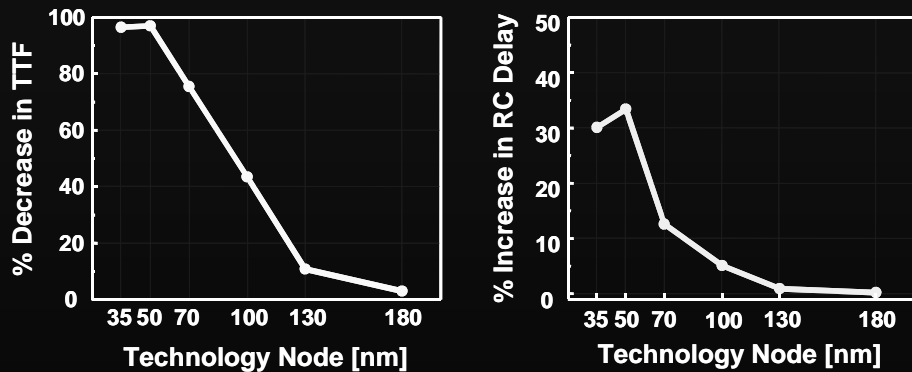- **Increase in T_{max} Due to Joule Heating of Interconnects**
  (FEM Simulation)

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

Temperature [°C] vs **Distance from Substrate [μm]**

Curves labeled: 50 nm, 35 nm, 70 nm, 100 nm, 130 nm, 180 nm

209 °C
Global Wires
126 °C

50 nm Node

ISPD 2001

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

% Decrease in TTF vs **Technology Node [nm]** (35 50 70 100 130 180)

% Increase in RC Delay vs **Technology Node [nm]** (35 50 70 100 130 180)

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

ISPD 2001

Scaling Trends and Implications (Summary)

- **Scaling trends that cause increasing thermal effects:**

  - **increasing interconnect levels**
  - **increasing current density**
  - **low-k dielectrics**
  - **increasing 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**

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,**

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

  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 ⁰C)**
- **The j_{avg} limit does not comprehend self heating**

Reliability Implications

Current Density Definitions (*DAC 99)*

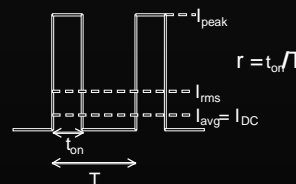- **Peak, Average, and RMS current densities:**

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

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

$$r = t_{on}/T$$

$$I_{peak}$$

$$I_{rms}$$

$$I_{avg} = I_{DC}$$

$$t_{on}$$

$$T$$

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

Reliability Implications

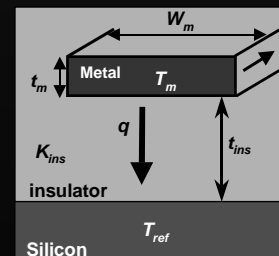## Impact of Self-Heating on EM *(DAC 99)*

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

- 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,

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

$$W_m$$

Metal T_m

$$t_m$$

$$q$$

$$K_{ins}$$

$$t_{ins}$$

insulator

$$T_{ref}$$

Silicon

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

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).**

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

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,

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

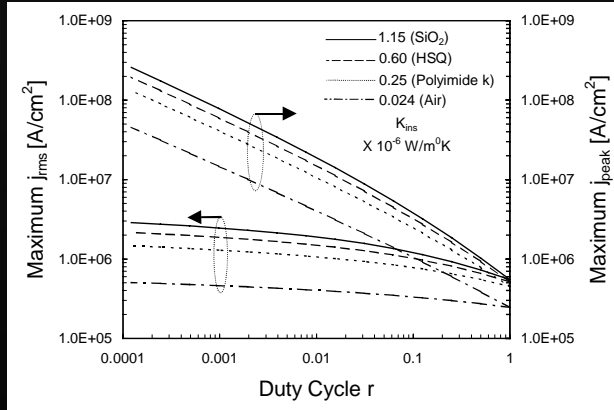- Incorporating the j^2_{rms} and j^2_{avg} values from yields the self-consistent equation,

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

This is a single equation in the single unknown temperature T_m

Reliability Implications

Low-k/Cu: Implications for Current Density Limits
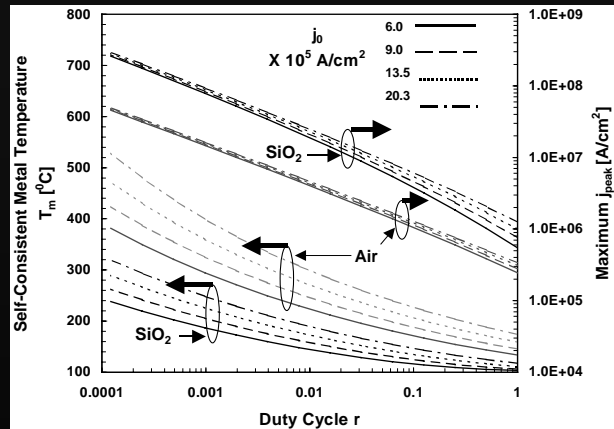

- Self-consistent j_{rms} and j_{peak} decrease significantly as low-k materials are introduced.

ISPD 2001

Reliability Implications

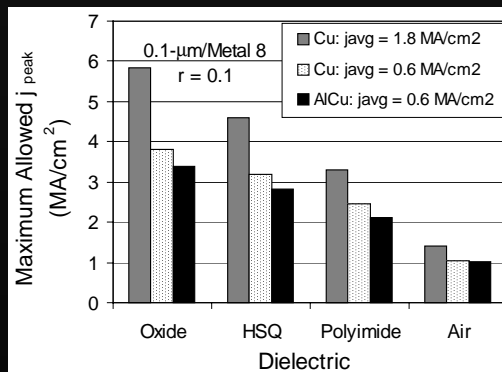## Implications for Interconnect Technology

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

ISPD 2001

Reliability Implications

Implications for Current Density Limits

- **Comparison with AlCu**

Chart:
- Y-axis: Maximum Allowed j_{peak} (MA/cm^2), scale 0 to 7
- X-axis: Dielectric — Oxide, HSQ, Polyimide, Air
- Legend:
  - Cu: javg = 1.8 MA/cm2
  - Cu: javg = 0.6 MA/cm2
  - AlCu: javg = 0.6 MA/cm2
- 0.1-µm/Metal 8, r = 0.1

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

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**

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
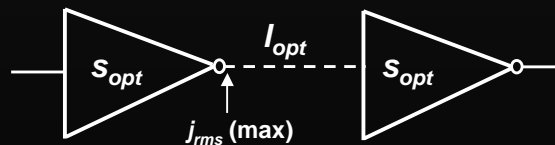- Long interconnects can be optimally buffered.
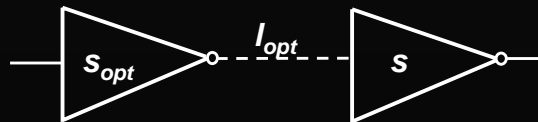
Performance Optimization
Performance Based Current Density
(Signal Lines)

- $0.25\ \mu m$ and $0.1\ \mu m$ technology
- Full 3-D Interconnect capacitance extracted
- Accurate I_{opt} and s_{opt} values determined by SPICE simulations

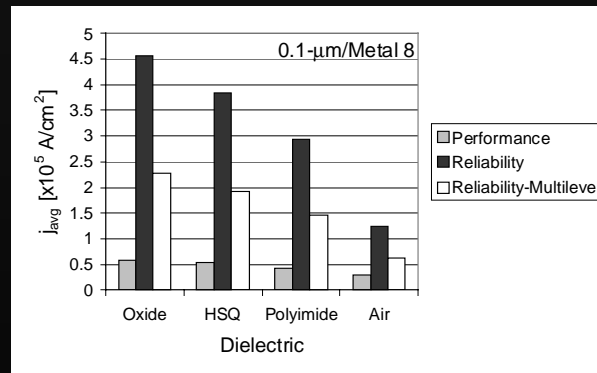- j_{rms} (max) occurs close to the repeater output due to the distributed nature of the interconnect.

Performance vs Reliability

● **Effect of Thermal coupling Included (NTRS Based)**

Chart: 0.1-μm/Metal 8, y-axis j_{avg} [×10^5 A/cm^2] from 0 to 5, x-axis Dielectric (Oxide, HSQ, Polyimide, Air). Legend: Performance, Reliability, Reliability-Multilevel.

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
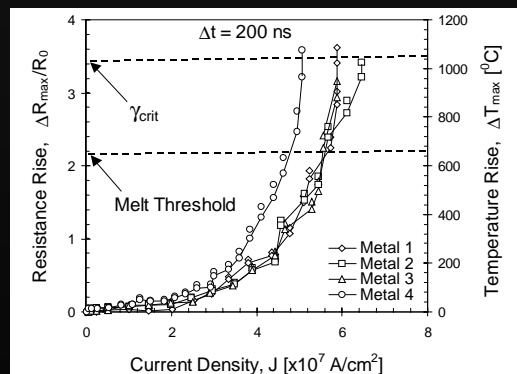
Non Steady-State Scenarios

- **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**

ISPD 2001

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**
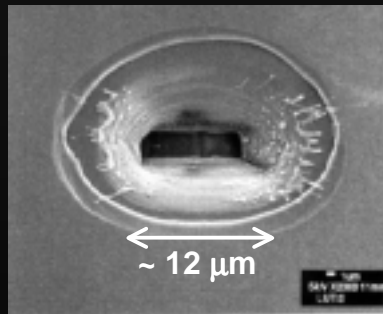
ISPD 2001

16

High-Current Effects

Open Circuit Failure *(IRPS 2000)*

~ 12 μm

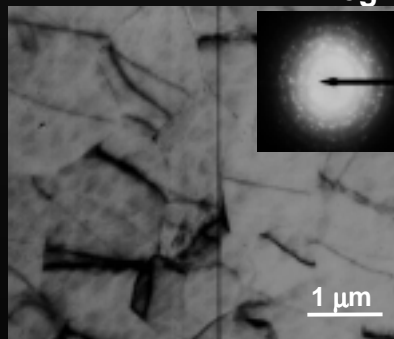**Metal 4**

~ 12 μm
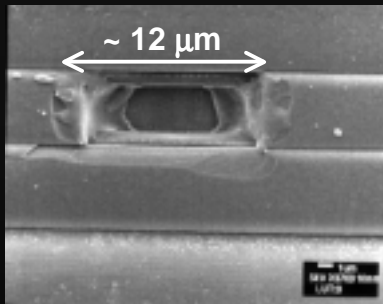
Metal 1

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

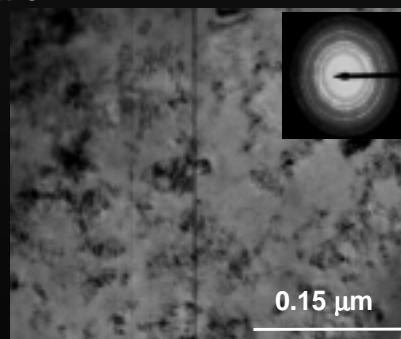- **Independent of overlying dielectric thickness.**

ISPD 2001

High-Current Effects

Latent Interconnect Damage *(IRPS 2000)*

Significant Electromigration Performance Degradation

1 μm

Unstressed AlCu

0.15 μm

Stressed AlCu

ISPD 2001

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**
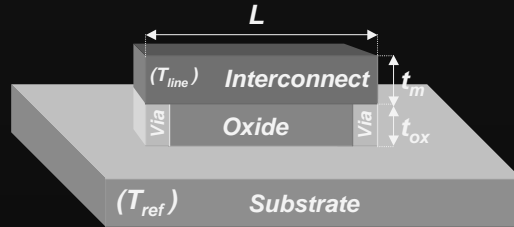 - **High-Current Design Rules Must be Followed for I/O and ESD Protection Circuit Interconnects**

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**
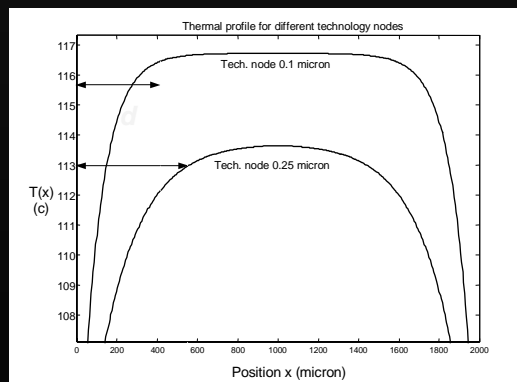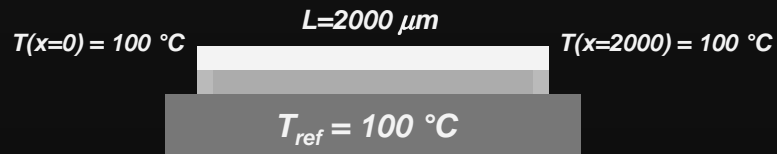- **Summary**

Interconnect Temperature

Heat equation in Interconnect (DAC 2001)

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

λ and θ are constants

Solution to Heat Equation

L=2000 μm

T(x=0) = 100 °C   *T(x=2000) = 100 °C*

T_{ref} = 100 °C

Thermal profile for different technology nodes

Tech. node 0.1 micron

Tech. node 0.25 micron

T(x) (c)

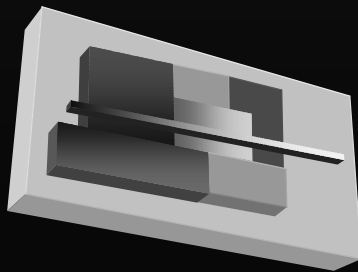Position x (micron)

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 $\Rightarrow$ non-uniform resistance profile**
  - **Unit length capacitance is not affected**

Non-Uniform Temperature Dependent Delay

- **Distributed RC delay model (DAC 2001)**

L

r_d

Δx

C_L
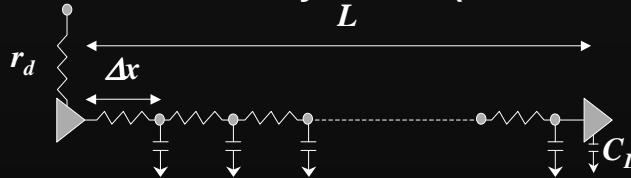
$$D = R_d(C_L + \int_0^L c_0(x)dx) + \int_0^L r_0(x)(\int_x^L c_0(\tau)d\tau + C_L)dx$$
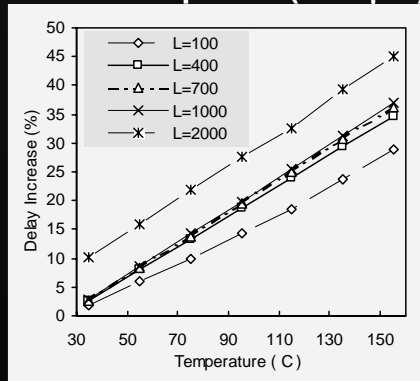
$$D = D_0 + (c_0L + C_L)\rho_0\beta\int_0^L T(x)dx - c_0\rho_0\beta\int_0^L xT(x)dx$$

D_0 is the Elmore delay model at 0 °C

ISPD 2001

Delay Degradation with Uniform Tref(x)

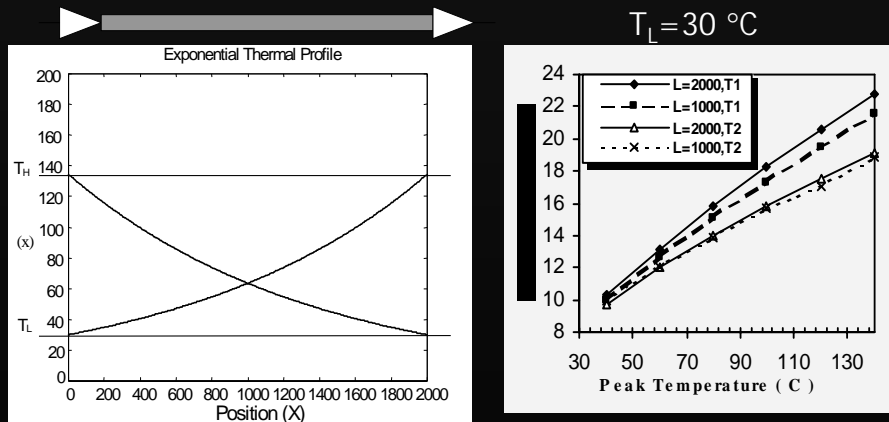- **With uniform thermal profile (0.25 μm):**

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

ISPD 2001

Delay Degradation with Non-uniform Tref(x)

- **Effect of exponential thermal profiles:**

$T_L = 30\ °C$

- **Direction of Thermal Gradient is Important**

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)**

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

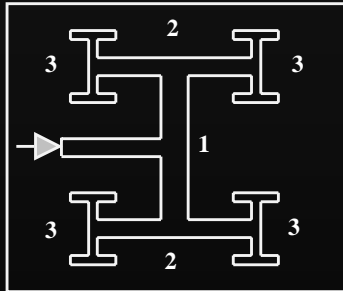- Clock nets must have near-zero skew among their sinks to guarantee correct functionality of the circuits
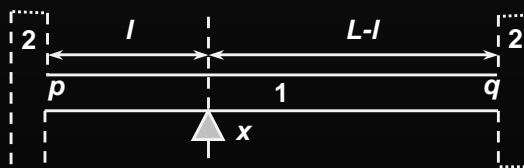
H-Tree's

- **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 *(CICC 2001)*

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

Branching Point cont'd

- **Using thermally dependent delay, optimal branching location (l^*) is:**
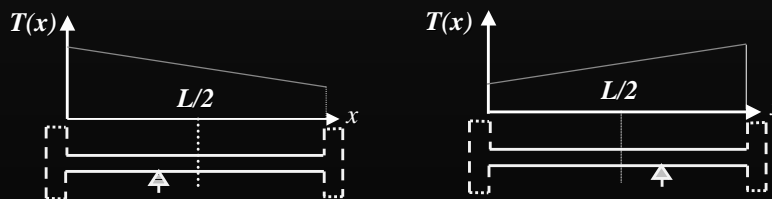
$$\beta \int_0^{l^*} T(x)\,dx + l^* - A = 0$$

A is constant

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

Movement of Branching Point

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

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%
$T(x) = ax + b$	T_H=170, T_L=90	1042	5.42
	T_H=170, T_L=110	1032	3.98
$a = \dfrac{T_H - T_L}{L} \quad b = T_L$	T_H=170, T_L=130	1021	2.65
$T(x) = a \cdot e^{-bx}$	T_H=170, T_L=90	957.5	5.24
	T_H=170, T_L=110	968.66	3.63
$b = \dfrac{1}{L}\ln(\dfrac{T_H}{T_L}) \quad a = T_H$	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

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**

ISPD 2001