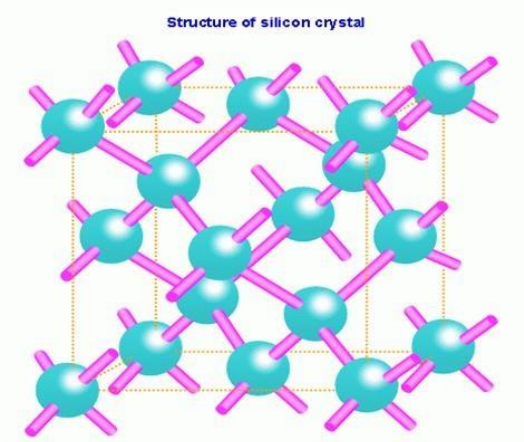
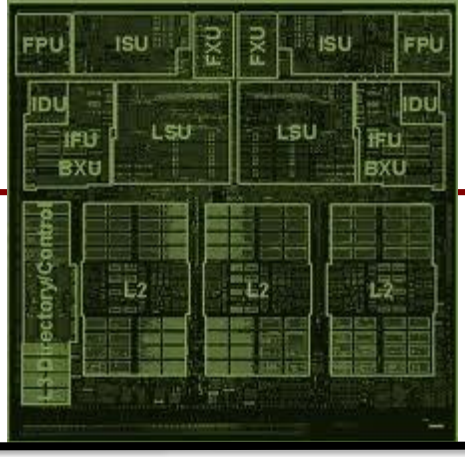
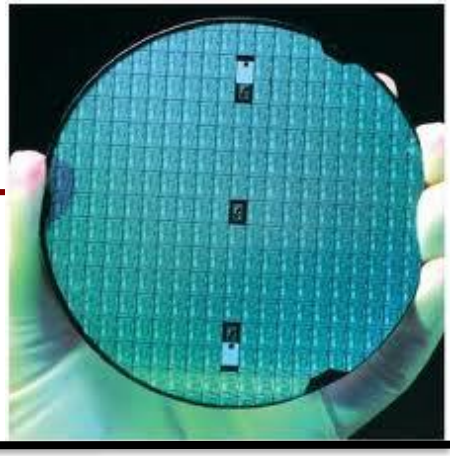






ECE 225

Lecture 11 & 12

Interconnect Design under Thermal, Power, Reliability & Variability Constraints

Prof. Kaustav Banerjee

Electrical and Computer Engineering

University of California, Santa Barbara

Voltage Drop Effects in Power Distribution Networks

A. H. Ajami, K. Banerjee and M. Pedram, International Journal of Analog Integrated Circuits and Signal Processing, Vol. 42, No. 3, pp. 277-290, Springer, 2005.

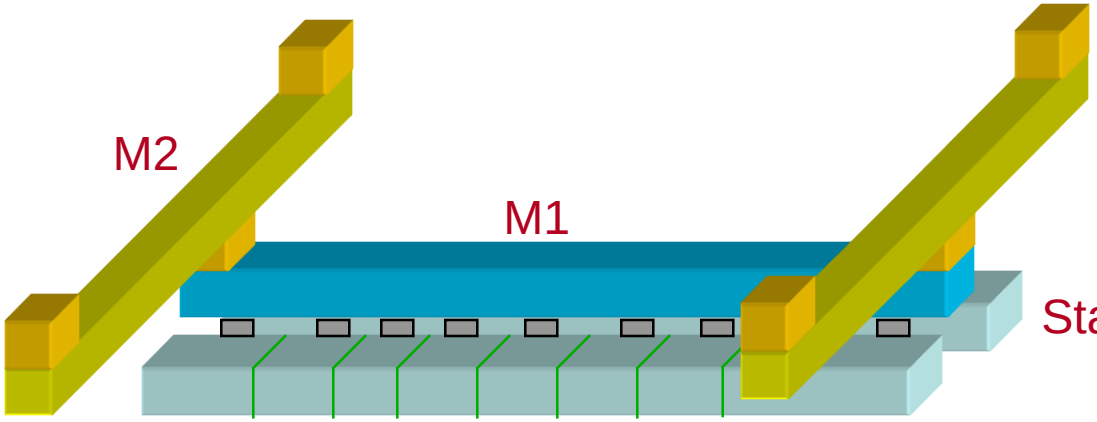
Voltage Drop Effects

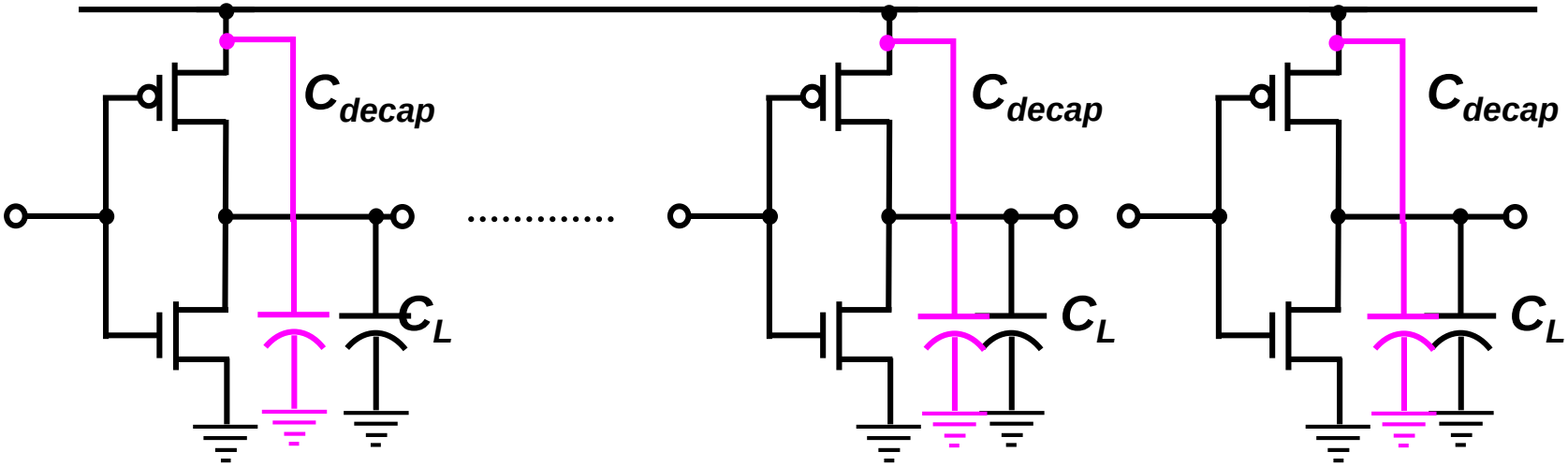
- Voltage-drop is an inseparable aspect of DSM power distribution network (PDN)
  - resistive voltage drop ΔV (on-chip resistance)
  - switching noise, Ldi/dt (off-chip inductance)
- Degrades device switching speed and *DC* noise margins
- Can cause functional failure in dynamic logic and timing violation in static logic
- 10% voltage-drop → 8% increase in device propagation delay (0.18μm tech.)

Challenges in P/G Network Design

- Satisfying the electromigration (EM) rules for each power routing segment
- Minimizing the area consumed by PDN
- Limiting the worst-case voltage-drop (~10% of *Vdd* for current technology)
  - Power network wire-sizing (↑ routing area)
  - Decoupling-cap insertion (↑ substrate area)

Local Power Distribution



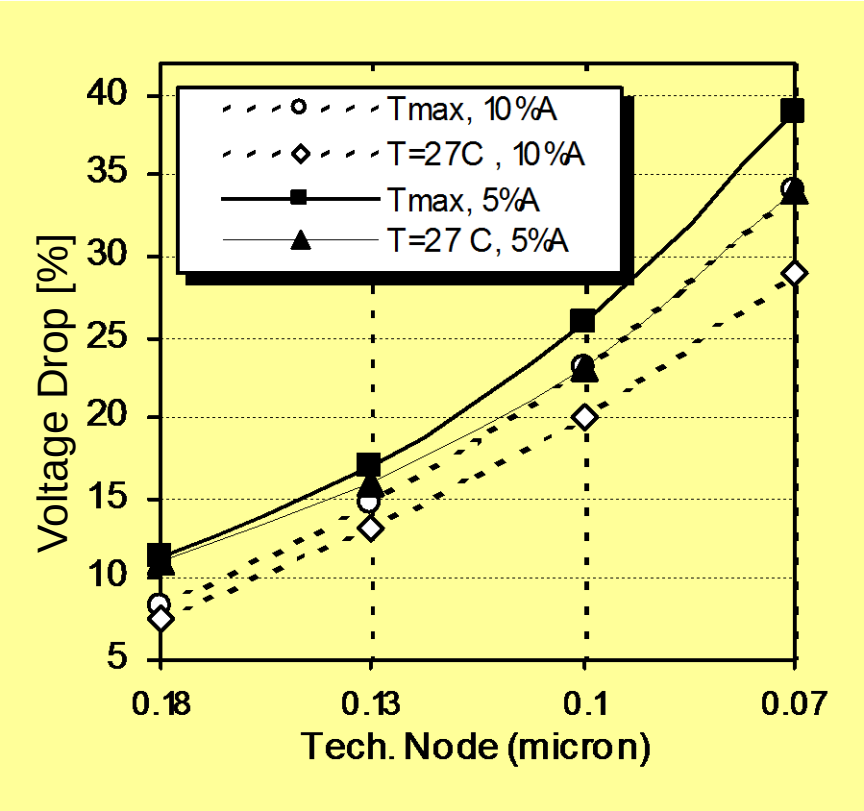


Technology Scaling Impacts

- Scaling of line dimension (*R, L, C*)
- Scaling of chip frequency, power, and # of power pads
- Interconnect thermal effects
  - Electromigration rules
  - R
- Thin-film scattering and barrier thickness effects in DSM interconnects
  - R

Global/Semi-global PDN IR-Drop

Allocation of 5% and 10% of the routing area for wire sizing to minimize the voltage-drop:

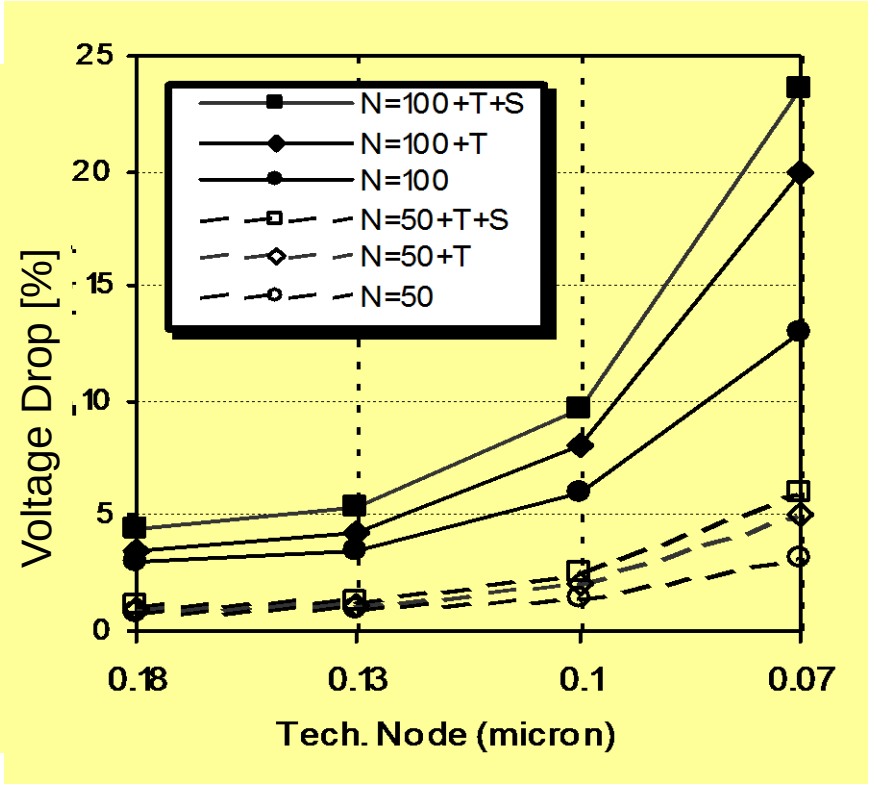


Voltage-drop increases rapidly with technology scaling….

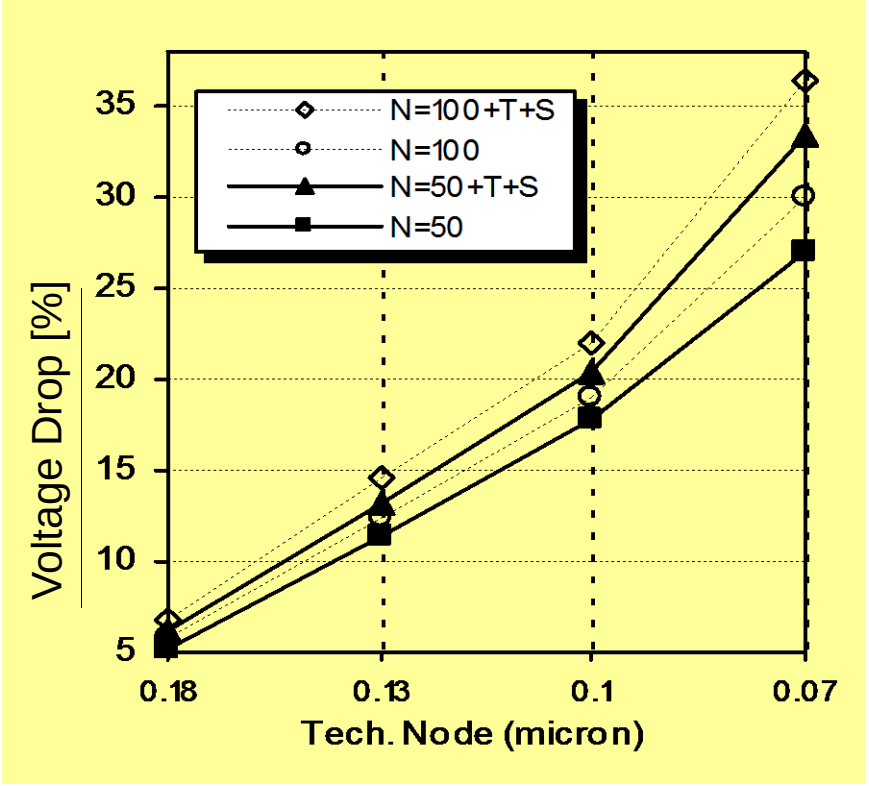
Interconnect temperature has a major impact on the voltage-drop of global segments…..

Full Chip IR-Drop

Local worst-case IR-Drop
for 50 and 100 switching cells

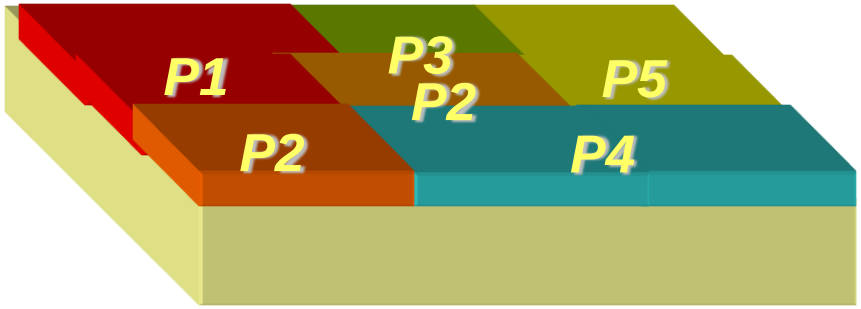


Total chip worst-case IR-Drop
10% routing, 5% chip area



T– considering temperature effect **S—considering thin-film effects**

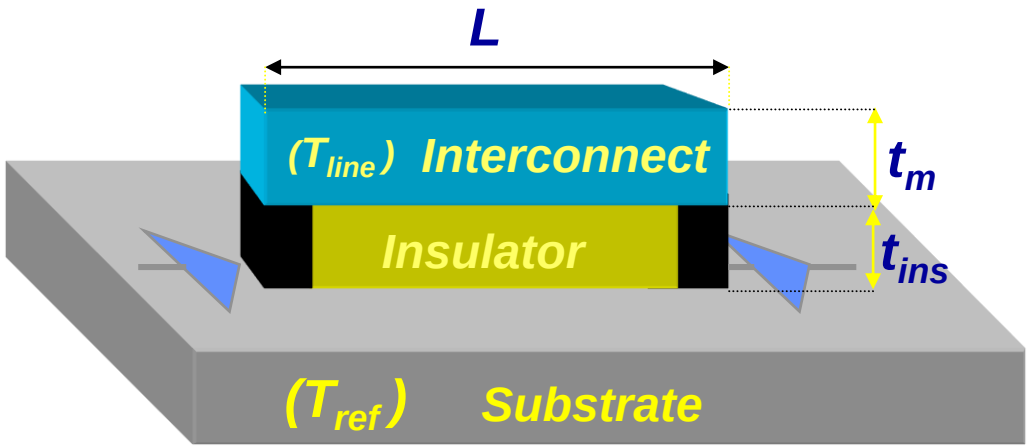
Within-Die Temperature Variations

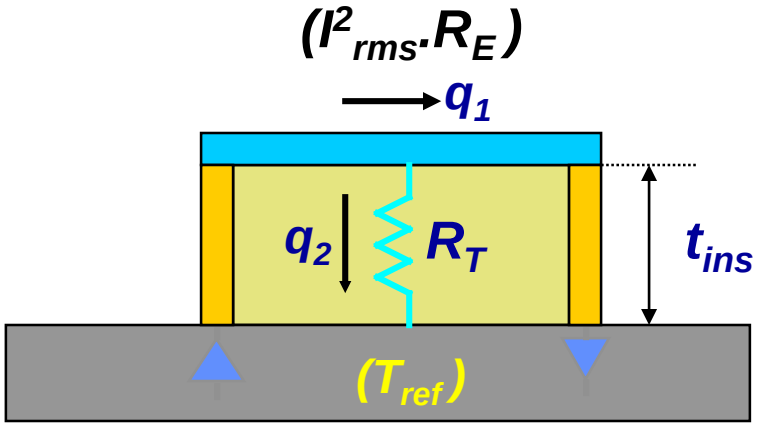


- Substrate power generation distribution is generally non-uniform
  - Functional block clock gating
  - System-level power management
  - Non-uniform distribution of gate sizing and switching activities in different blocks
- Substrate thermal profile is non-uniform
  - Thermal time constant is of the order of *ms*
  - Switching activities at the block level are more important
  - Introduces non-uniformity in the global interconnect thermal profile

Interconnect Thermal Profile

A. Ajami et al., DAC 2001





$$\frac{d^2T_{line}}{dx^2} = -\frac{Q}{k_m}$$

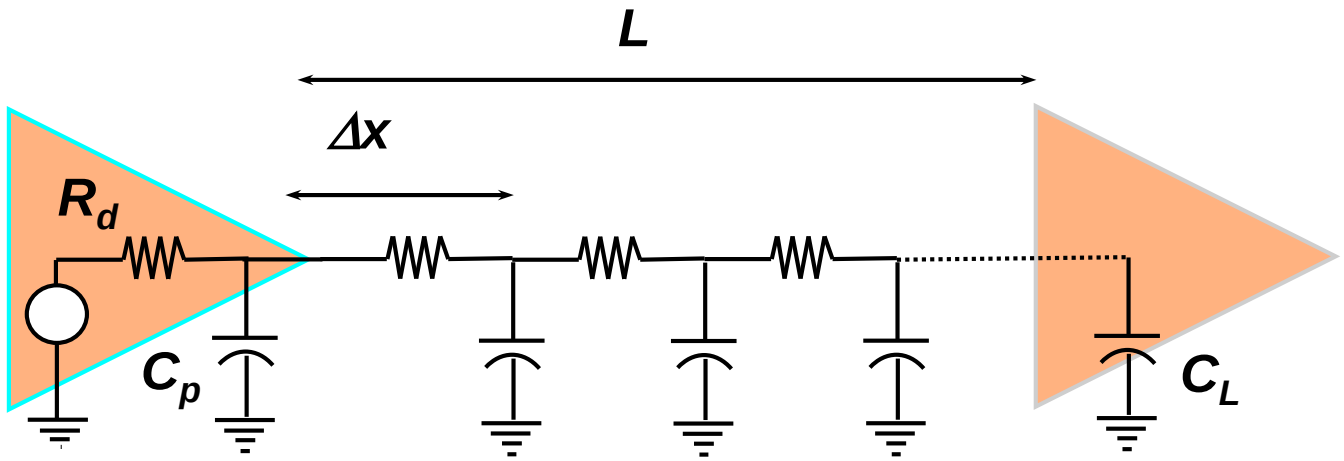
$$Q = q_1 - q_2$$

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

λ and θ are constants

$$f(L, t_m, k_m, t_{ins}, k_{ins}, I_{rms}, R_E)$$

Non-Uniform Temperature-Dependent Delay



$$D = R_d\left(C_P + C_L + \int_0^L c_0(x)dx\right) + \int_0^L r_0(x)\left(\int_x^L c_0(\eta)d\eta + C_L\right)dx$$

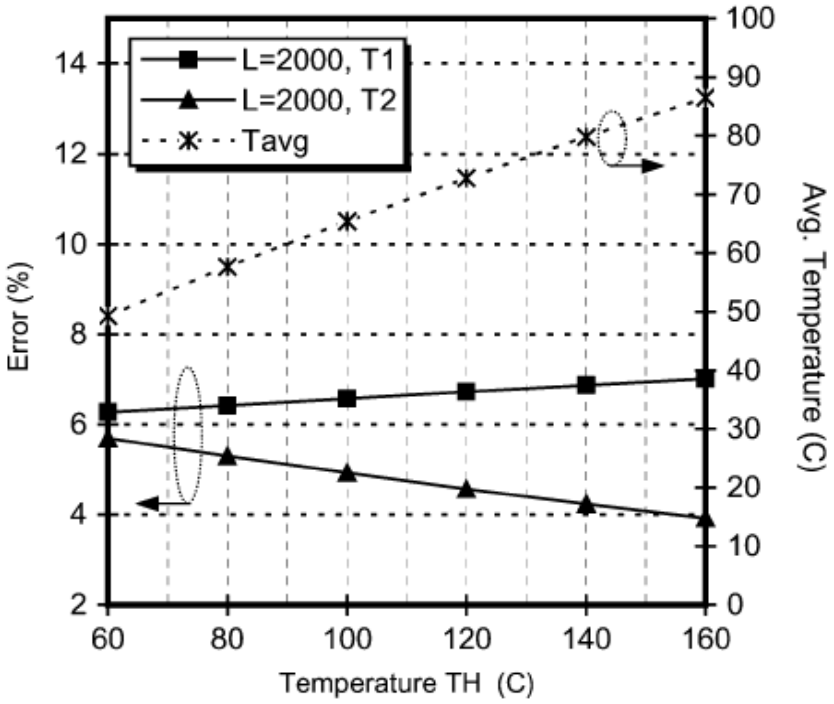
$$D = D_0 + (c_0 L + C_L)\rho_0\beta\int_0^L T(x)dx - c_0\rho_0\beta\int_0^L xT(x)dx$$

D_0 is the Elmore delay model at reference temp.

Implications for Delay and IR-Drop

Impact on delay estimation…..

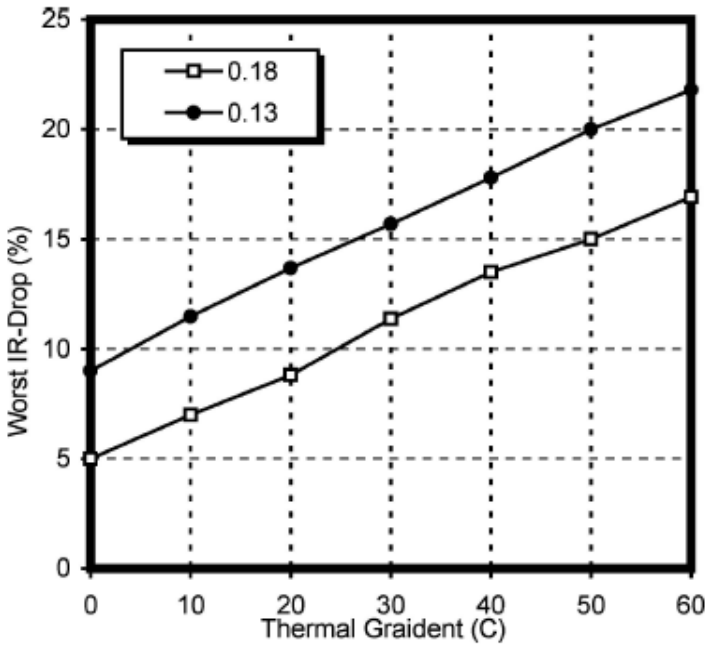
T1: positive exponential gradient
T2: negative exponential gradient
For a fixed T_Low



Ajami et al., TCAD 2005

Impact on IR-drop analysis…..

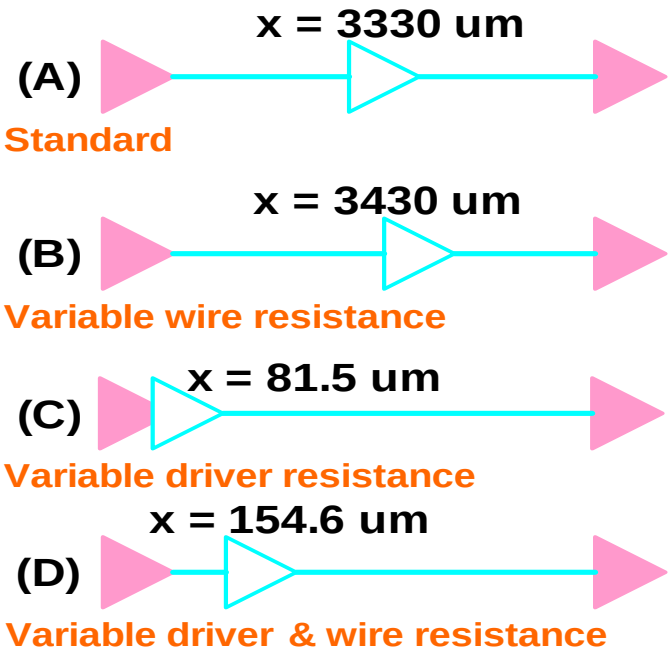
Worst-case voltage-drop (V_{IR}/V_{dd}) increases in the presence of thermal gradients



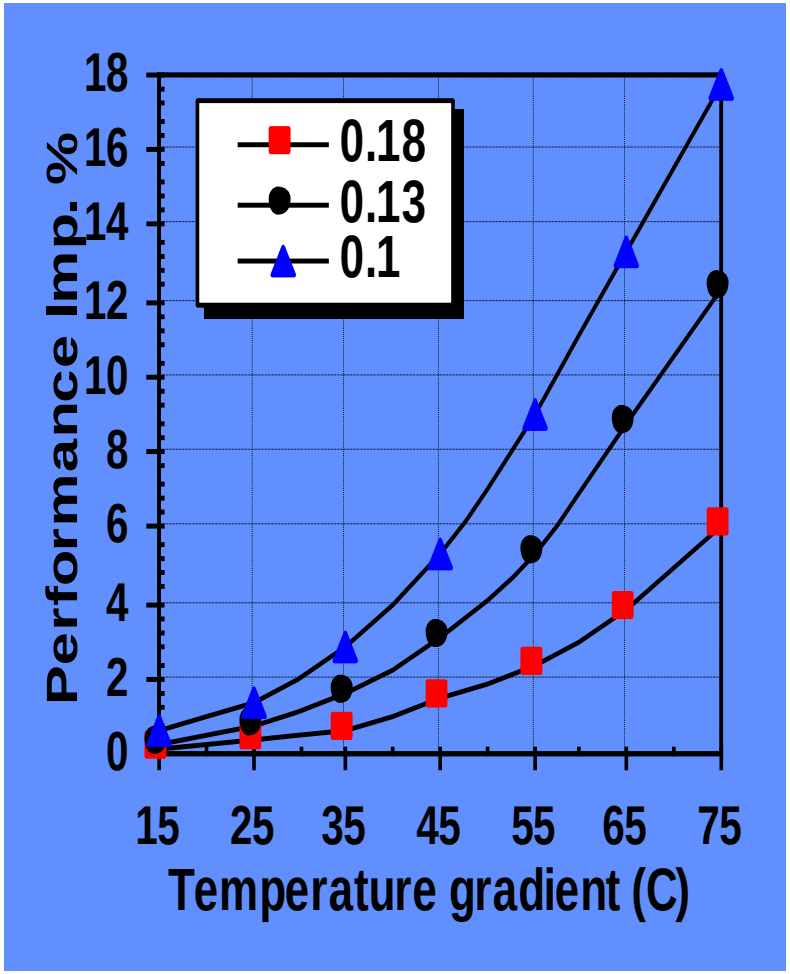
Ajami et al., JAICSP, 2005

Impact on Buffer Insertion

Buffer movement in a 6660 um line (180 nm node)



Delay improvement after thermally-aware buffer insertion



Ajami et al., ICCAD 2001

Interconnect Thermal Effects and Reliability

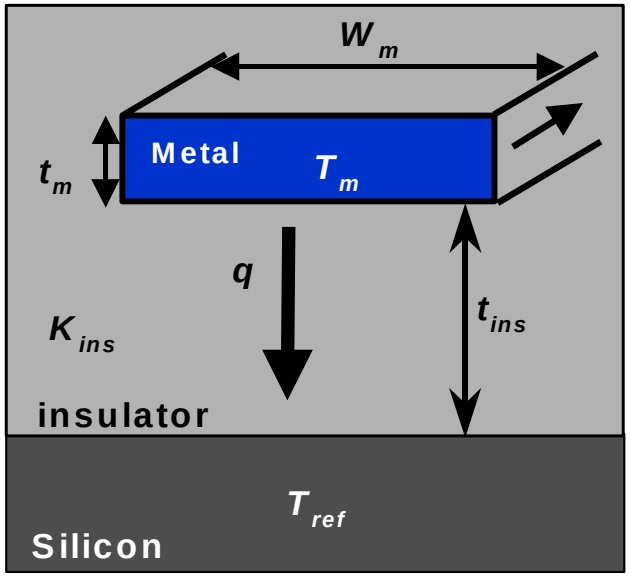
K. Banerjee and A. Mehrotra, IEEE Circuits and Devices Magazine, pp. 16-32, Sept. 2001.

Self-Heating or Joule Heating

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

q_j is the effective thermal impedance:

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

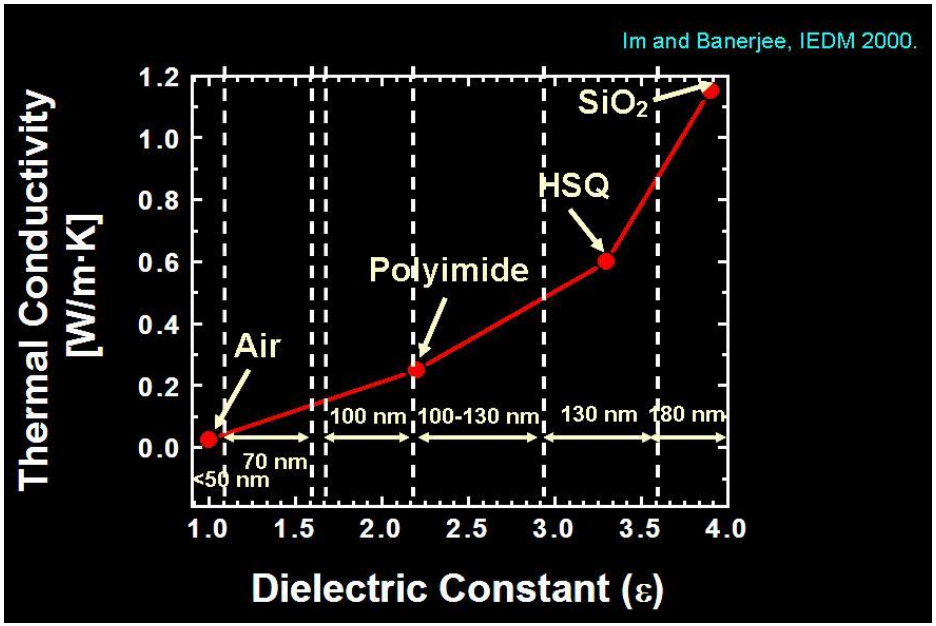


W_{eff} is the effective metal width to account for quasi-2D heat conduction.

Interconnect Scaling Trends

Implications of Technology Scaling on Thermal Effects

- Low-k materials have lower thermal conductivity than silicon dioxide.
- Scaling trends that cause increasing thermal effects are:
  - increasing current density
  - increasing interconnect levels
  - low-k dielectrics
  - increasing thermal coupling

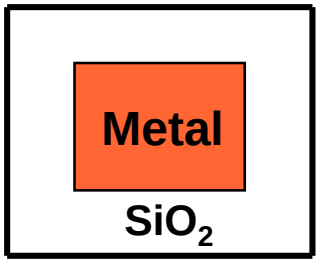


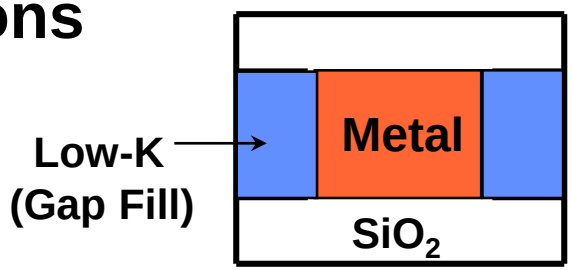
Interconnect Scaling Trends

Impact of Line Width Scaling Using Low-k

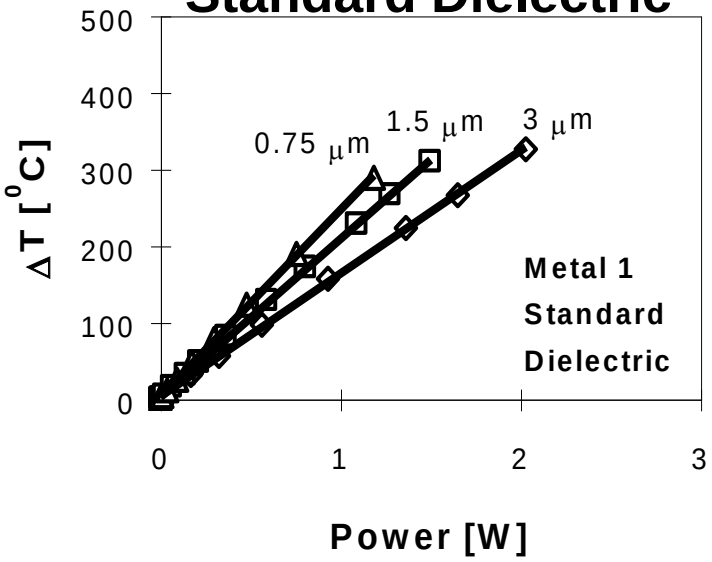
Banerjee et al., IEDM, 1996

DC Conditions

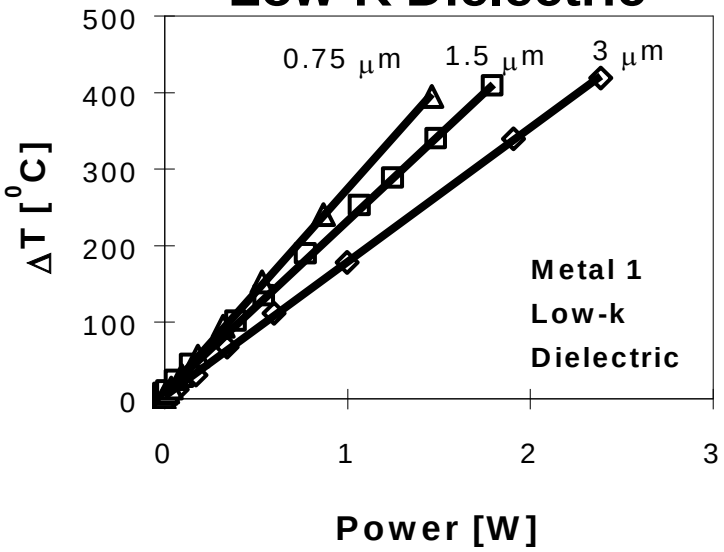




Standard Dielectric



Low-K Dielectric



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

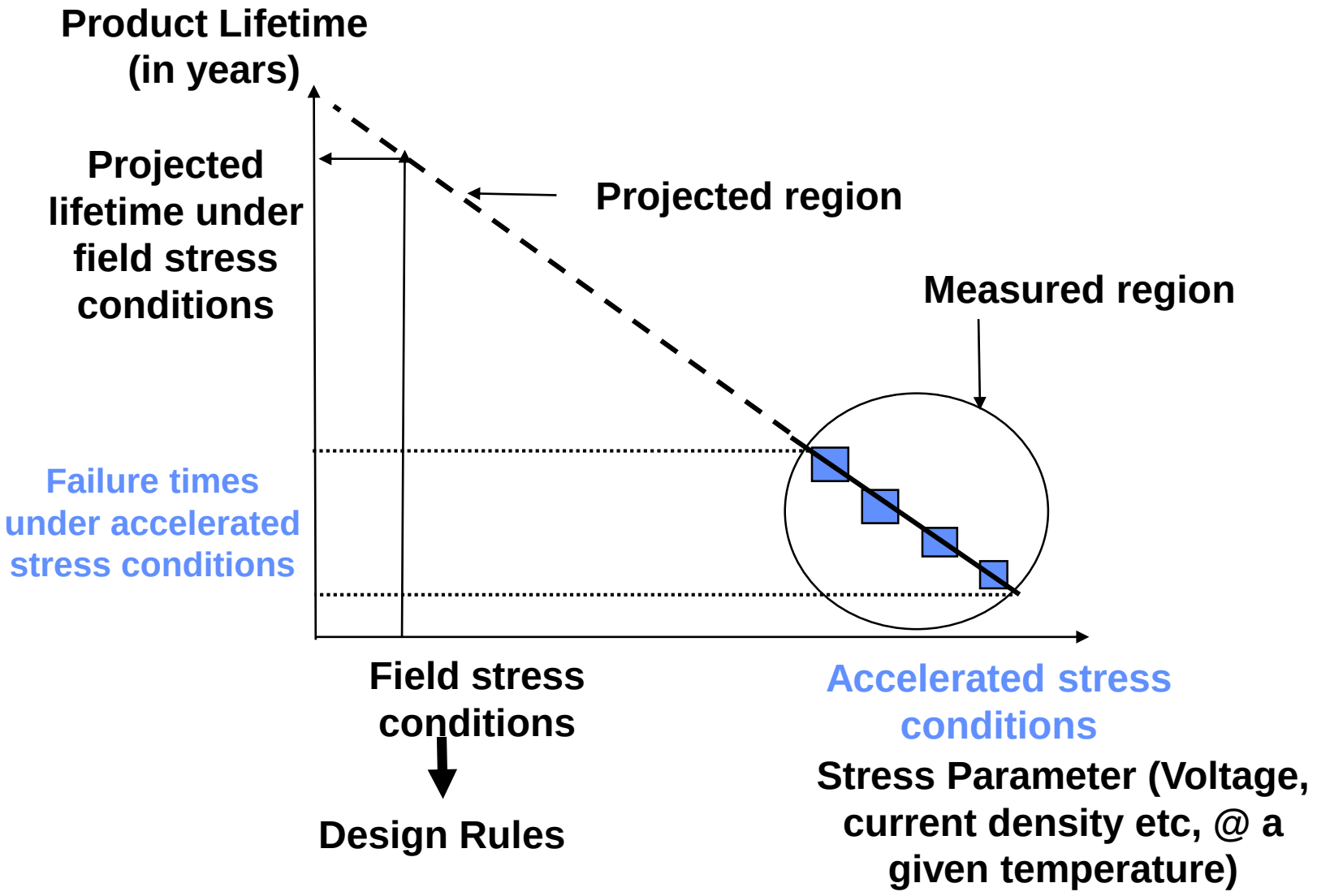
Electromigration Reliability

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

$$TTF = A\, j_{avg}^{-2}\, exp\left(\frac{Q}{k_B T_m}\right)$$

- EM stress data gives a technology limit of the current density (j_{avg}) - - *for a 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**

VLSI Reliability: Lifetime Estimation and Design Rule Generation



Self-Consistent Design

Typical EM Analysis

- Accelerated EM stress data yields *A, Q*, and *n* in Black's equation, and a value of log-normal s_{LN}.
- 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).
- Typical goal: achieve a 10 year lifetime.
- The j_{avg} limit does not comprehend self heating, and is therefore not self-consistent.

Self-Consistent Design

Current Density Definitions

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

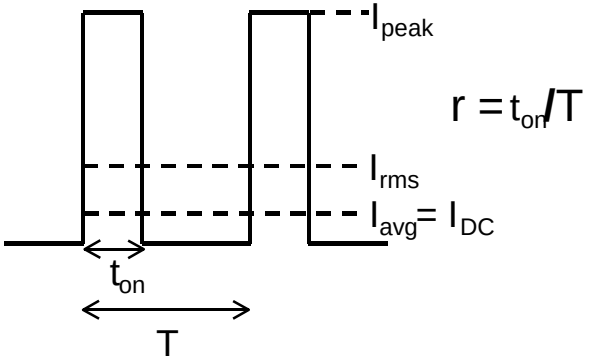
$$j_{peak} = \frac{I_{peak}}{A} \qquad j_{avg} = \frac{1}{T}\int_0^T j(t)\,dt \qquad j_{rms} = \sqrt{\frac{1}{T}\int_0^T j^2(t)\,dt}$$

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}.**

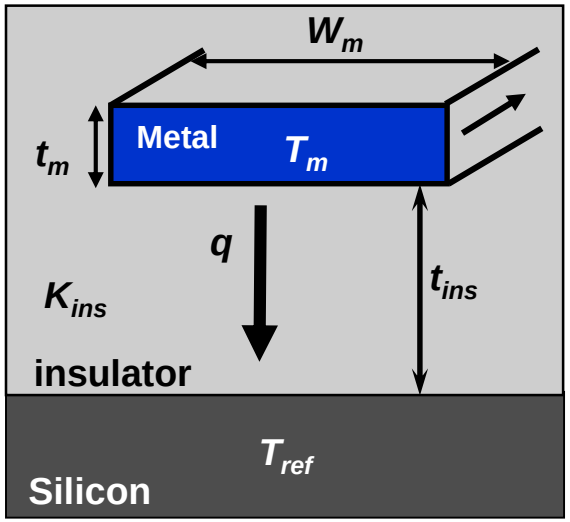
Self-Consistent Design

Impact of Self-Heating on EM

- EM lifetime given by: $TTF = A\, j^{-n} \exp\left(\frac{Q}{k_B T_m}\right)$
- Due to self-heating: $T_m = T_{ref} + DT_{self\text{-}heating}$

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

$$j_{rms}^2 = \frac{(T_m - T_{ref}) K_{ins} W_{eff}}{t_{ins} t_m W_m \rho_m (T_m)} \quad (1)$$



R_q 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.

Self-Consistent Design

What is Self-Consistent Design?

- Typically, design rules specify j_{avg} from EM and j_{rms} from self-heating separately.
- Self-consistent approach: comprehends EM and self-heating simultaneously.
- In order to achieve an EM reliability lifetime goal mentioned earlier, 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).

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

Self-Consistent Design

What is Self-Consistent Design?

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

$$\frac{j_{avg}^2}{j_{rms}^2} = r \qquad (3)$$

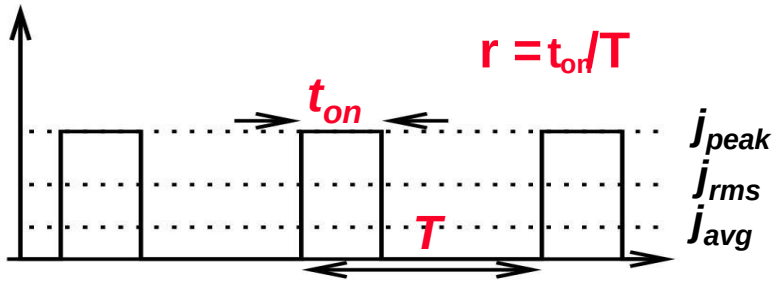
- Incorporating the j^2_{rms} and j^2_{avg} values from (1) and (2) in (3) yields the self-consistent equation,

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

This is a single equation in the single unknown temperature T_m. Once the self-consistent T_m is obtained, the corresponding j_{rms} and j_{peak} and j_{avg} values can be calculated.

Coupled Analysis

Unipolar Case:



- **Coupled equation:**

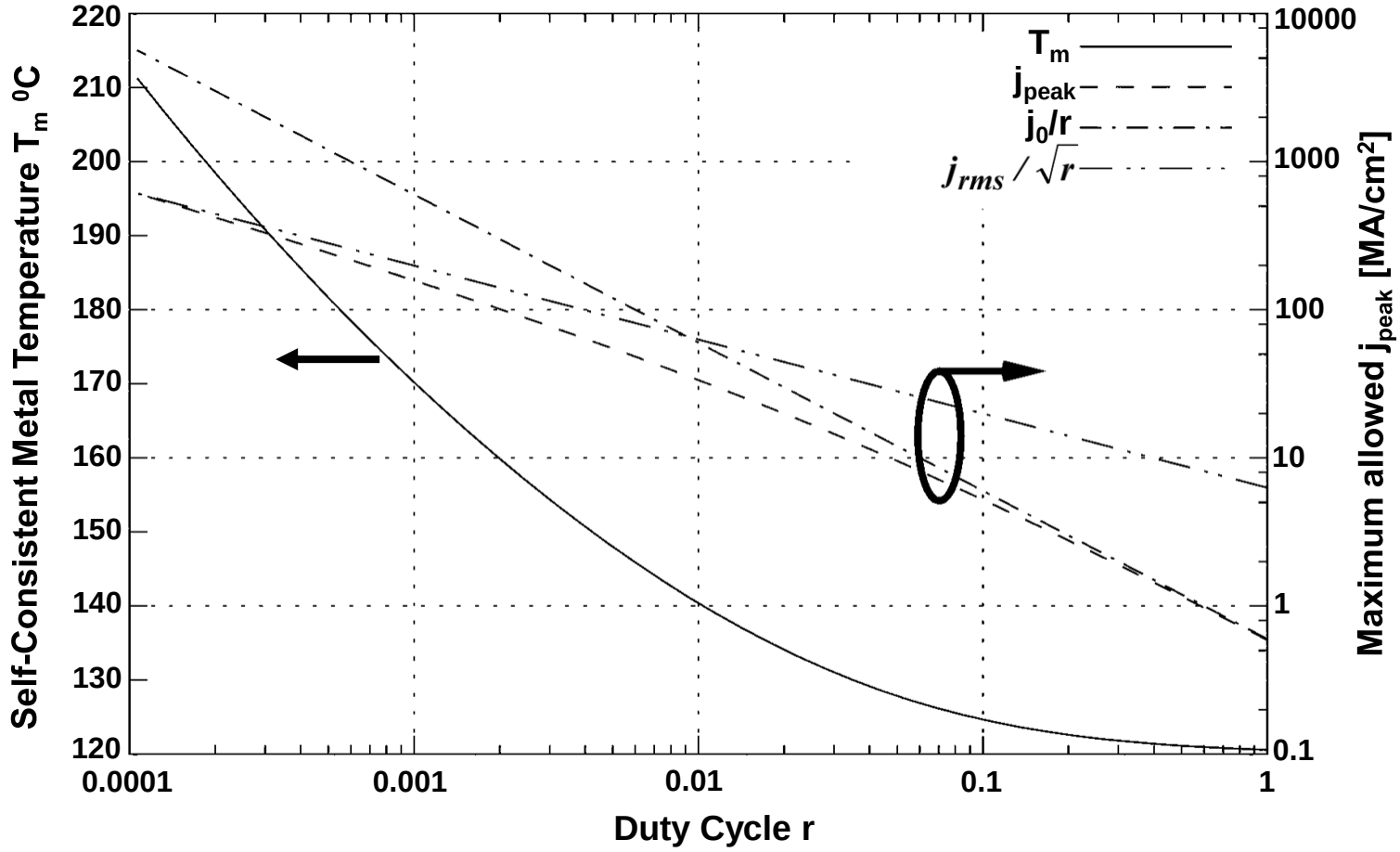
$$\frac{j_{avg}^2}{j_{rms}^2} = r$$

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

Once T_m is calculated, j_{rms} and j_{peak} can be obtained

Coupled Analysis

Unipolar Case: Metal 6, 180 nm node, j_0 = 0.6 MA/cm^2



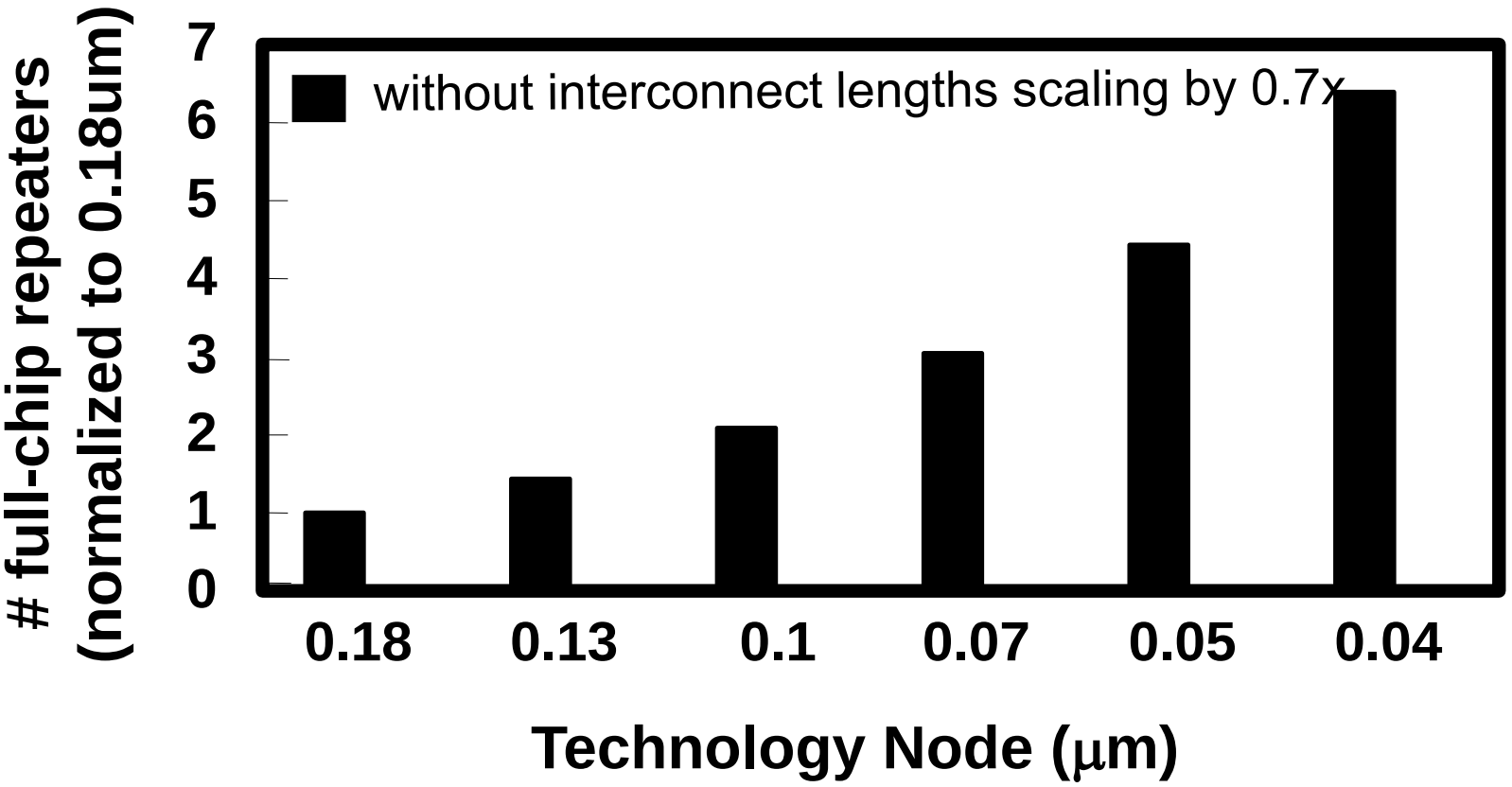
Interconnect Delay and Power Dissipation

K. Banerjee and A. Mehrotra, IEEE Transactions on Electron Devices, Vol. 49, No. 11, pp. 2001-2007, November 2002.

Performance Optimization: Global Wires

- **Increasing chip complexity and scaling**
  - **number and (electrical) length of global lines & latency increase**
- **Repeater insertion used for lowering the signal delays**
- **Delay of optimally repeatered line is linear in its length**
- **High-performance designs: large number of repeaters (> 10^6 for sub-100 nm designs)**

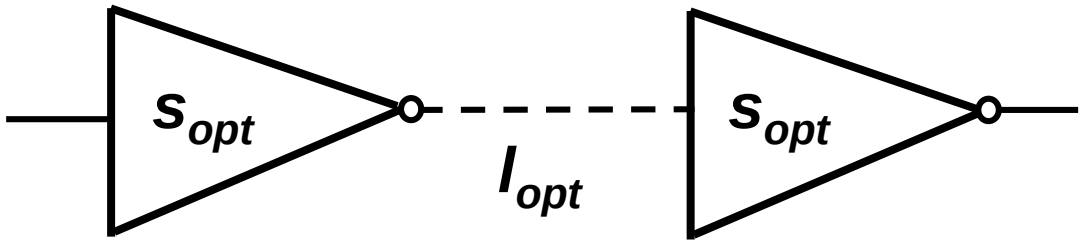
Increasing Number of Repeaters



Source: Intel Corporation

Optimal Repeater Insertion

- Long interconnects can be optimally buffered

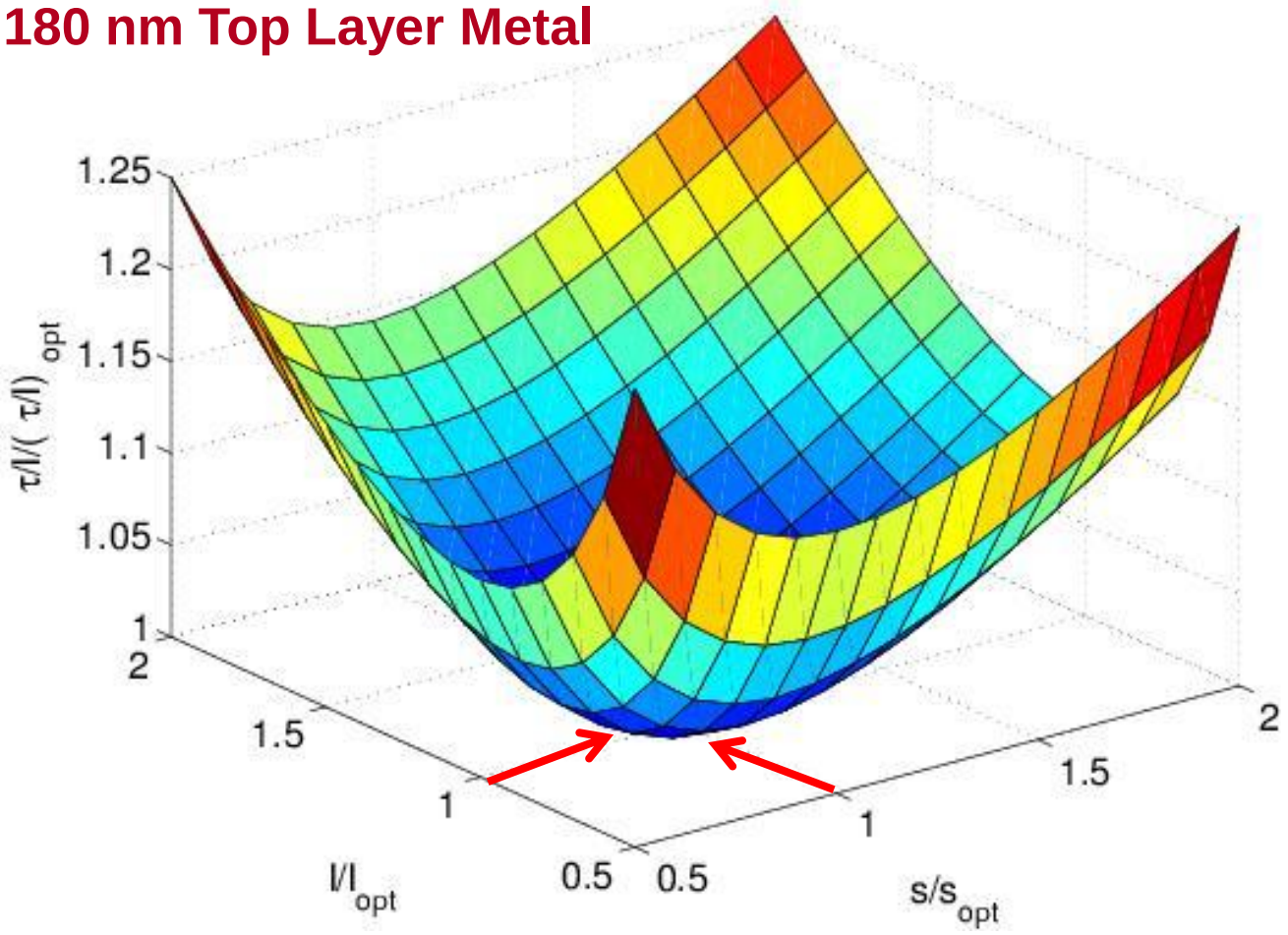


- Large size: could be ~ 450 times the minimum sized inverters available in a relevant technology (K. Banerjee et al., DAC 1999)
- Can contribute to significant on-chip power dissipation!

Delay Vs Power Tradeoff for Global Wires

- All global wires are not on the critical path
- Optimal buffering is not always necessary
- Some delay penalty can be tolerated on non-critical wires
- Potential for large power savings by using smaller repeaters and larger inter-repeater wire length

Interconnect Delay Optimization



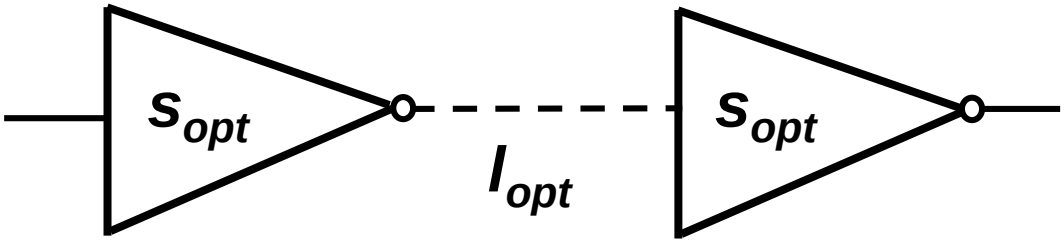
For $s = s_{opt}/2$ and $l = 2\ l_{opt}$; Delay Penalty is only 25%

Power-Optimal Repeater Insertion

Banerjee and Mehrotra, TED 2002.

- A methodology to estimate repeater size and inter-repeater wire length which minimizes the total interconnect power dissipation for a given delay penalty
- Estimate power-optimal buffering schemes for various ITRS technology nodes
- Analyze relative importance of various components of power dissipation as a function of technology scaling

Methodology



delay per unit length:

$$\frac{\tau}{\ell} = \frac{1}{\ell} r_s \left(c_0 + c_p\right) + \frac{r_s}{s} c + r s c_0 + \frac{1}{2} r c \ell$$

For $\ell = \ell_{opt}$, $s = s_{opt}$:

$$\left(\tau / \ell\right)_{opt} = 2 \sqrt{r_s c_0 r c} \quad \{1 + \sqrt{0.5 \left(1 + \frac{c_p}{c_0}\right)}\}$$

Methodology

$$P_{repeater} = P_{switching} + P_{leakage} + P_{short\ circuit}$$

$$P_{switching} = \alpha \left(s \left(c_p + c_0 \right) + l\, c \right) V_{DD}^2\ f_{clk}$$

$$P_{leakage} = V_{DD}\ I_{leakage} = V_{DD}\ I_{off}\ W_{n_{min}}\ s$$

$$P_{short\ circuit} = \alpha\, t_r\, V_{DD}\ I_{peak}\ f_{clk} = \alpha\ t_r\, V_{DD}\ W_{n_{min}}\ s\ I_{short\ circuit}\ f_{clk}$$

- α is the activity factor (=0.15)
- t_r = rise time of input voltage to change from V_{tn} to V_{DD} - V_{tp}

Methodology

If the delay penalty is f, then: $\frac{\tau}{l} = (1+f)(\tau / l)_{opt}$

$$\frac{P_{repeater}}{l} = k_1 \left(s / l \left(c_p + c_0 \right) + c \right) + k_2 \frac{s}{l} + k_3 s$$

switching leakage short-circuit

$$k_1 = \alpha V_{DD}^2 f_{clk}$$

$$k_2 = \frac{3}{2} V_{DD} I_{off\,n} W_{n\,min}$$

$$k_3 = \alpha V_{DD} W_{n\,min} I_{short\ \ circuit\ \ } f_{clk} \log_e 3 (1+f)(\tau / l)_{opt}$$

Methodology

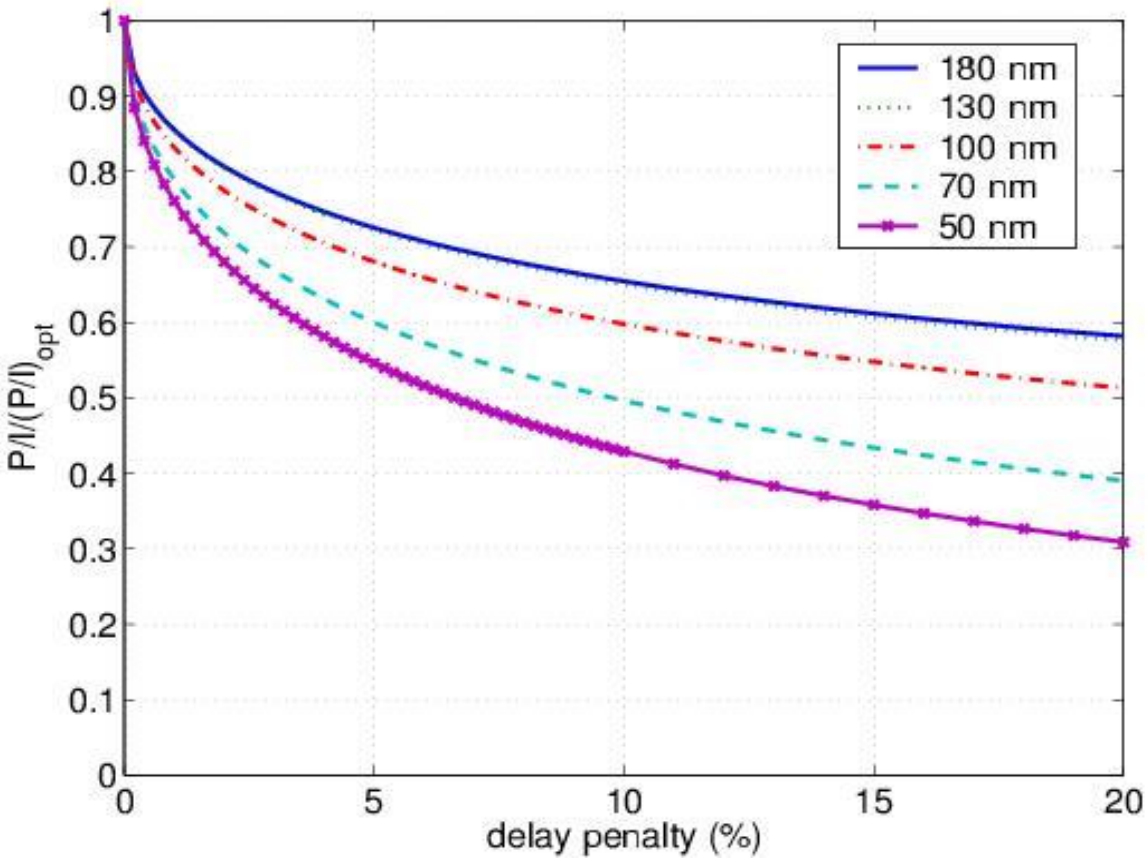
Minimize $P_{repeater}/l$ w.r. t. s:

$$\frac{P_{repeater}}{l} = k_1 \left(s / l \left(c_p + c_0 \right) + c \right) + k_2 \frac{s}{l} + k_3 s$$

switching *leakage* *short-circuit*

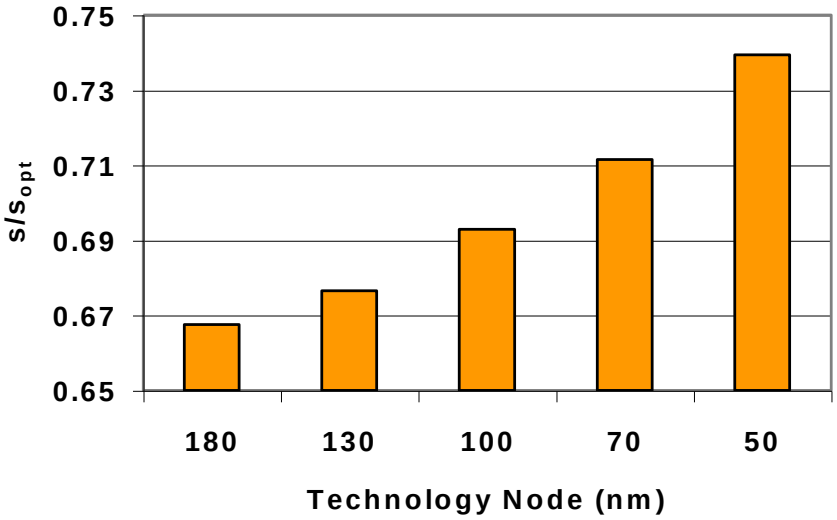
- **Involves three non-linear equations with three unknowns: *l*, *s*, and *dl/ds***
- **Used Newton-Raphson to solve numerically**

Results: Power Vs Delay Penalty

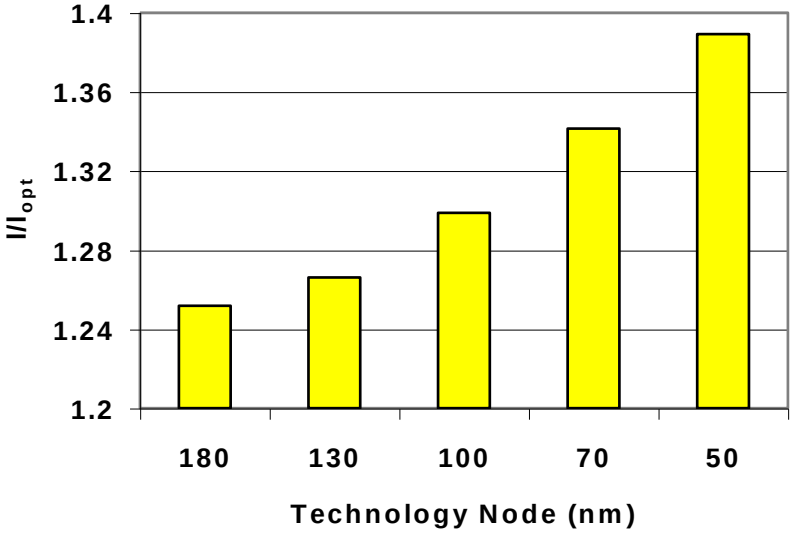


- Normalized power per unit length reduces as delay penalty increases
- Incremental reduction in P/I is high for small values of the delay penalty
- 180 nm and 130 nm results are almost identical
- However, normalized P/I decreases with technology scaling: leakage power

Optimization Results: 5% delay penalty



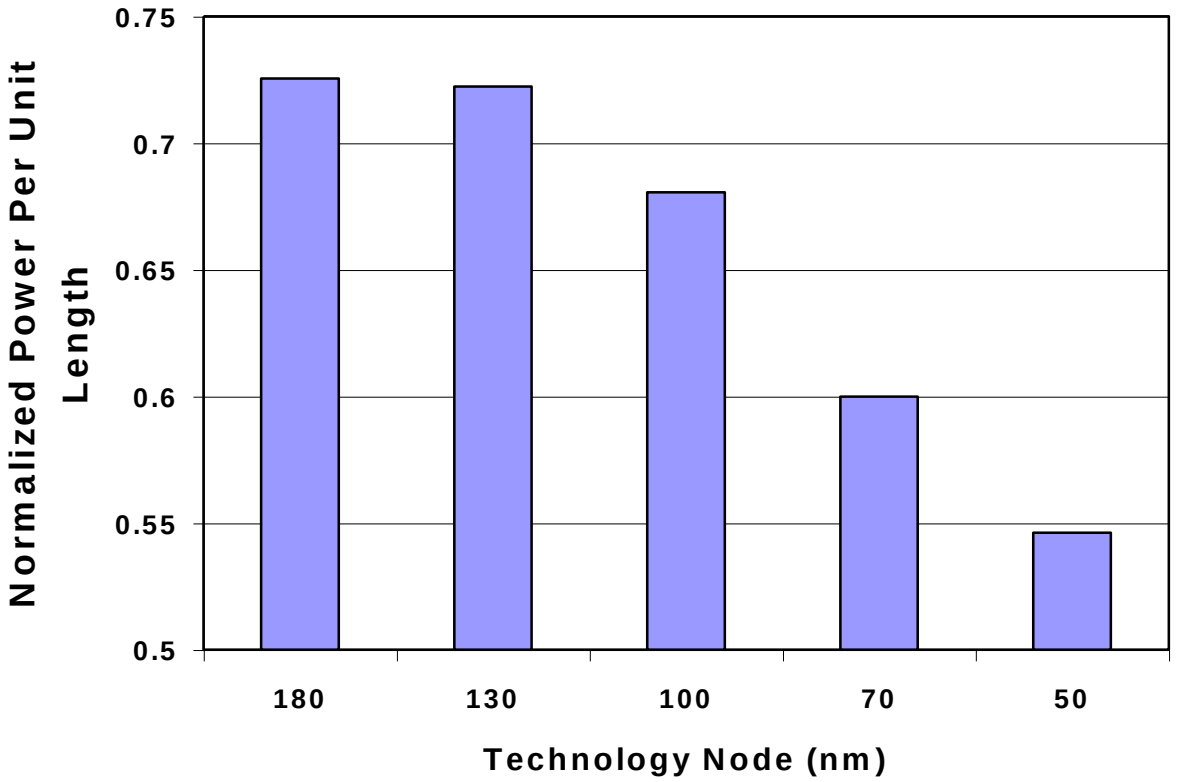
$S < S_{opt}$



$l > l_{opt}$

For optimal power dissipation: repeater size needs to be reduced and inter-buffer wire length needs to be increased

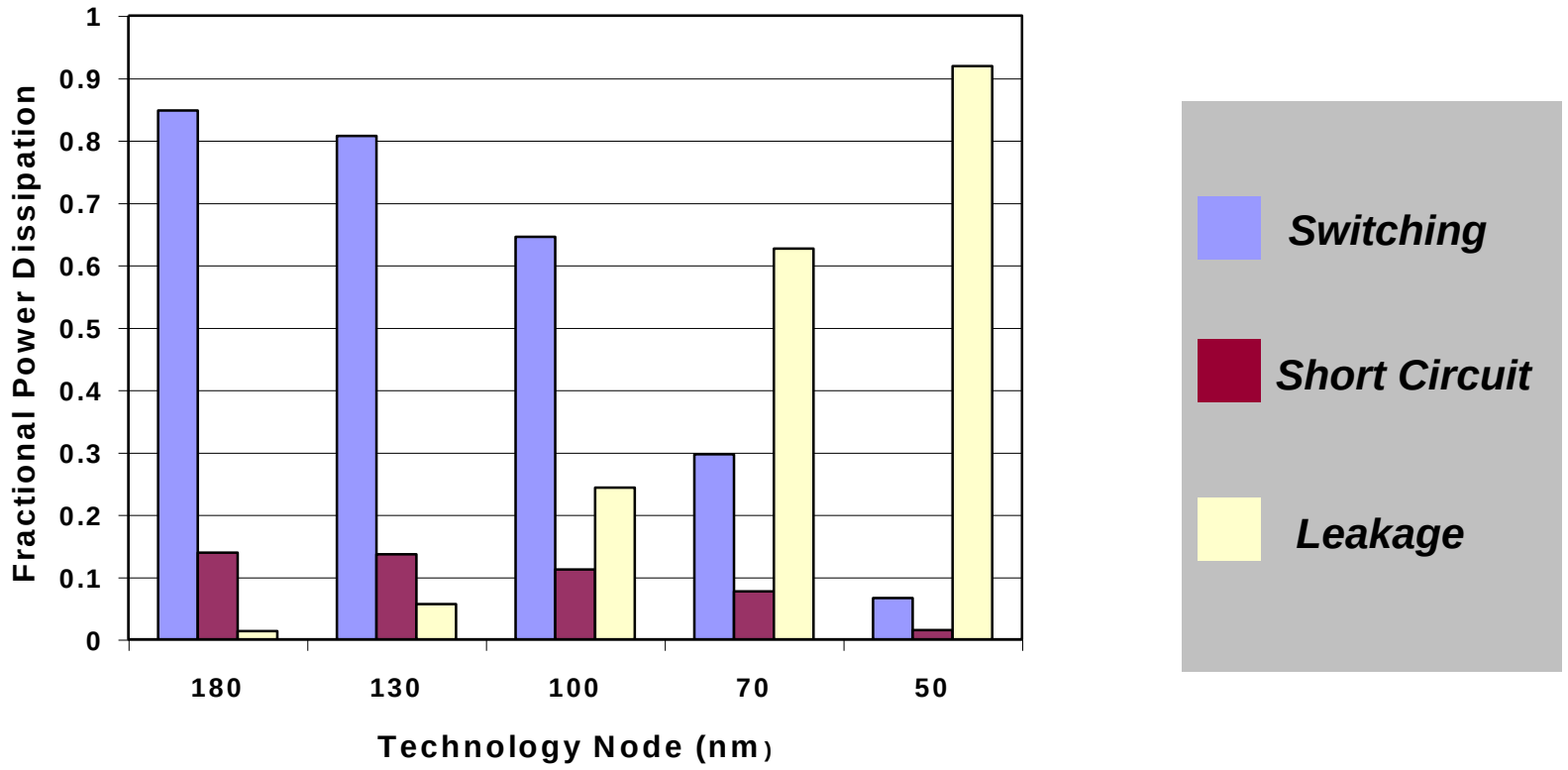
Optimization Results: 5% delay penalty



Normalized power per unit length decreases rapidly with scaling: due to substantial increase in leakage current

Optimization Results: 5% delay penalty

Relative contributions of the three components of PD

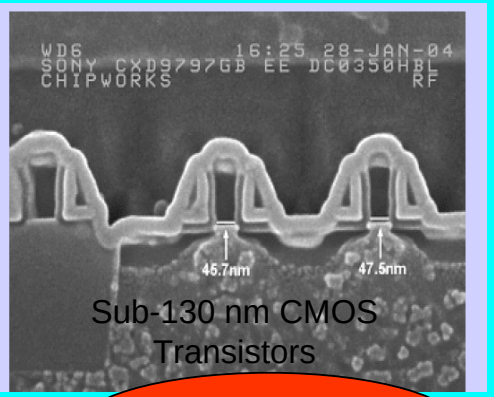


- Leakage power is the dominant component for advanced nodes
- Short circuit power is also non-trivial across all nodes

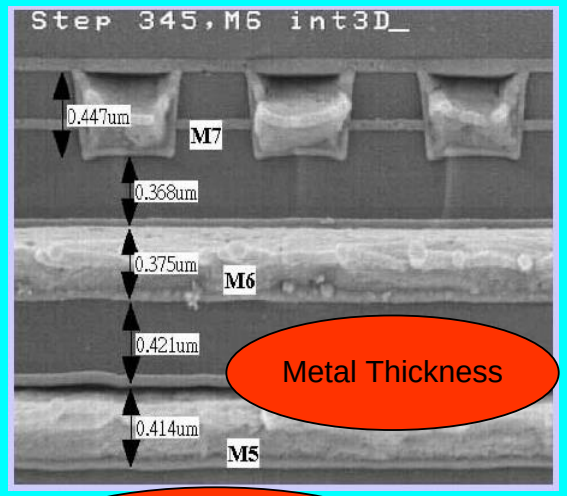
Implications of Parameter Variations

V. Wason and K. Banerjee, IEEE International Symposium on Low Power Electronic Design (ISLPED), 2005, pp. 131-136.

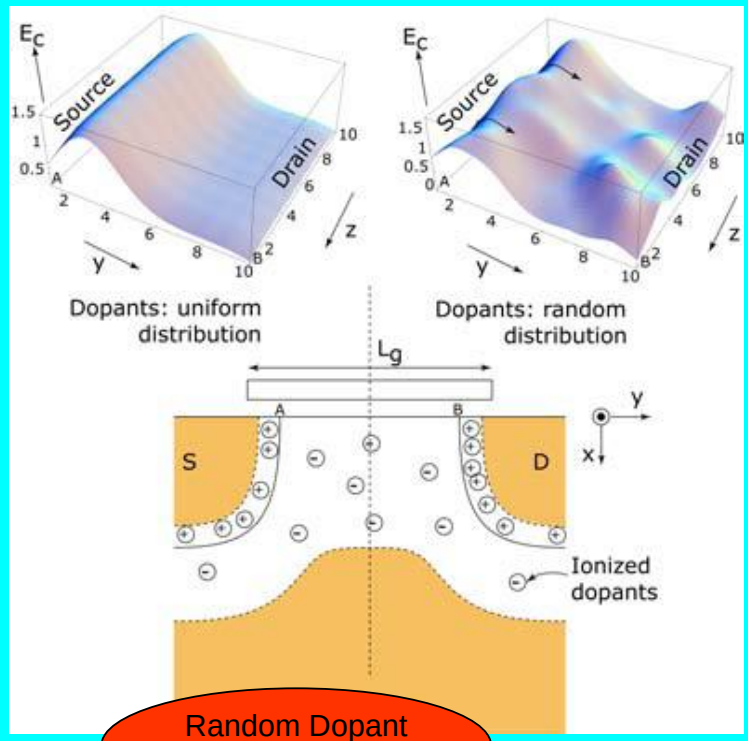
Parameter (P-V-T) Variations



Transistor channel length

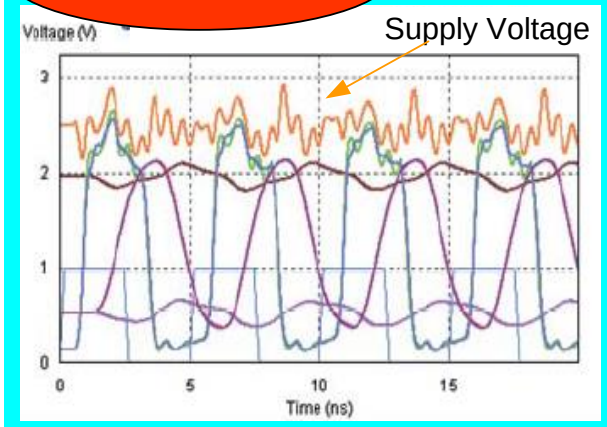


Metal Thickness

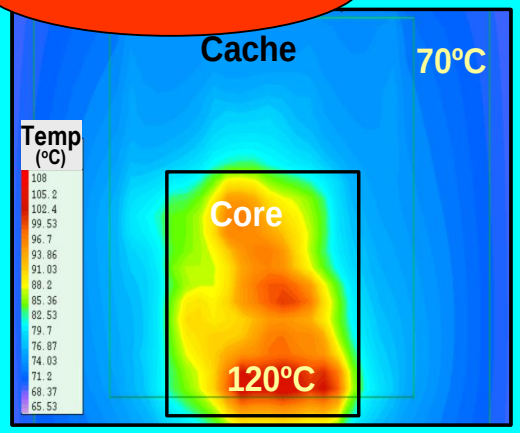


Random Dopant Fluctuations

Power Supply Voltage

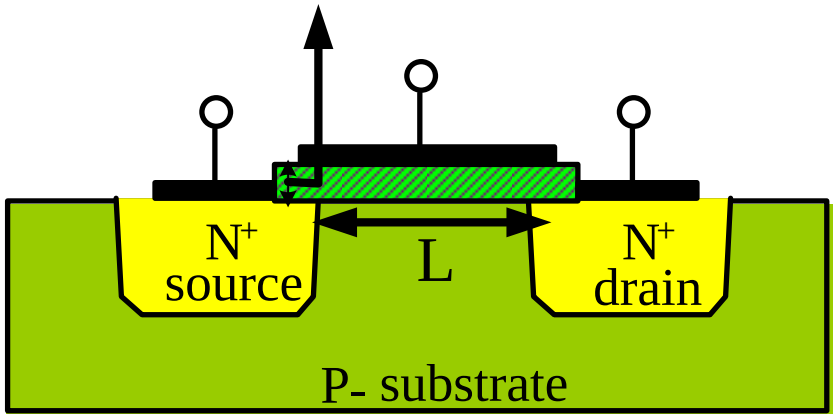


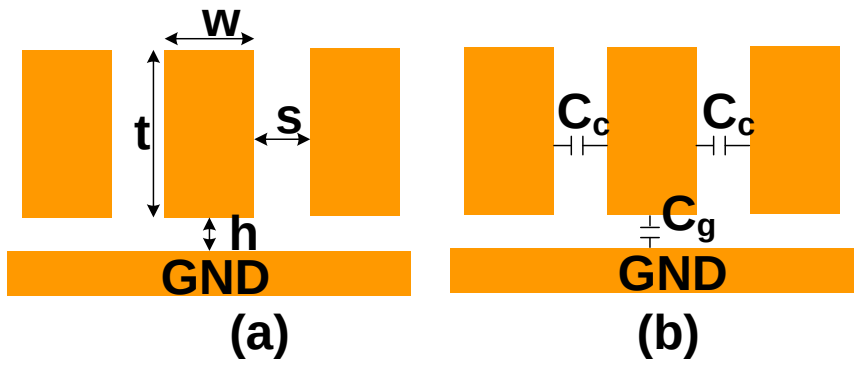
Chip Temperature



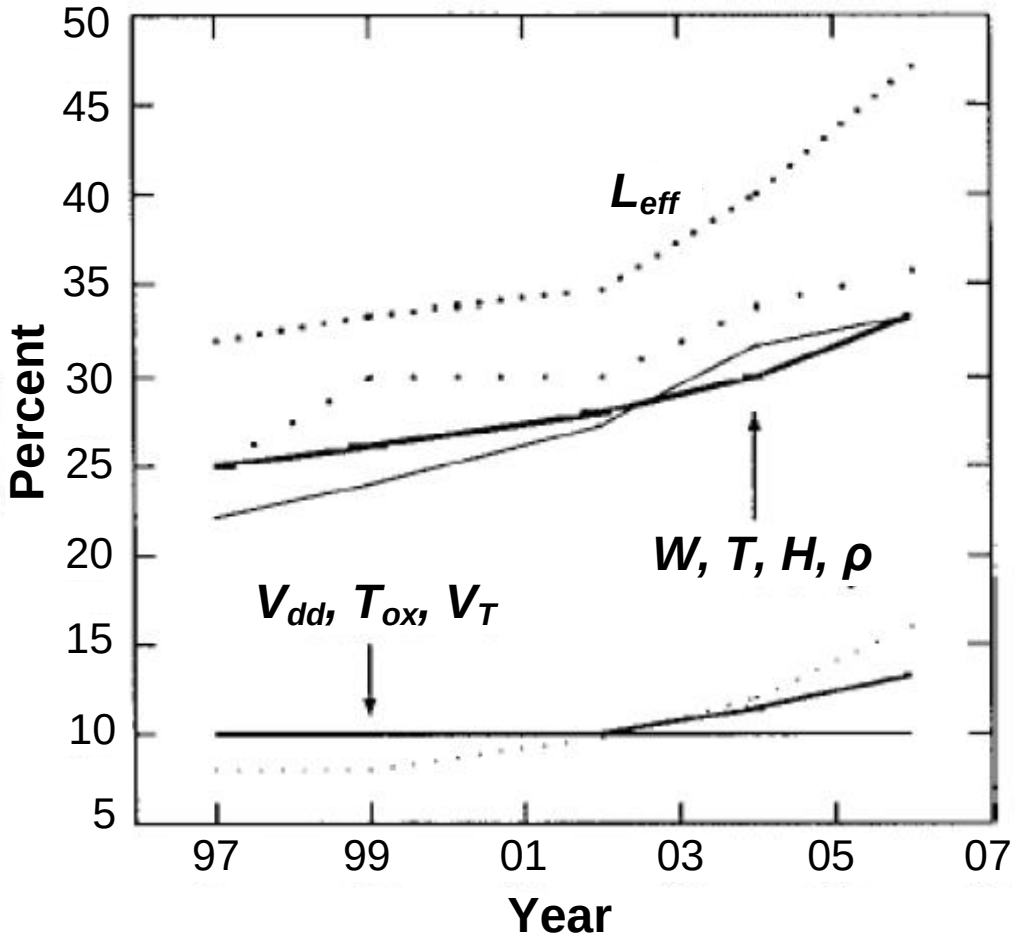
Process Variations

- Device Variations
  - Channel Length (L)
  - Oxide Thickness (t_{ox})
  - Dopant Density (N_a)
- Interconnect Variations
  - Line Width (w)
  - Spacing (s)
  - Metal Thickness (t)
  - Dielectric Thickness (h)



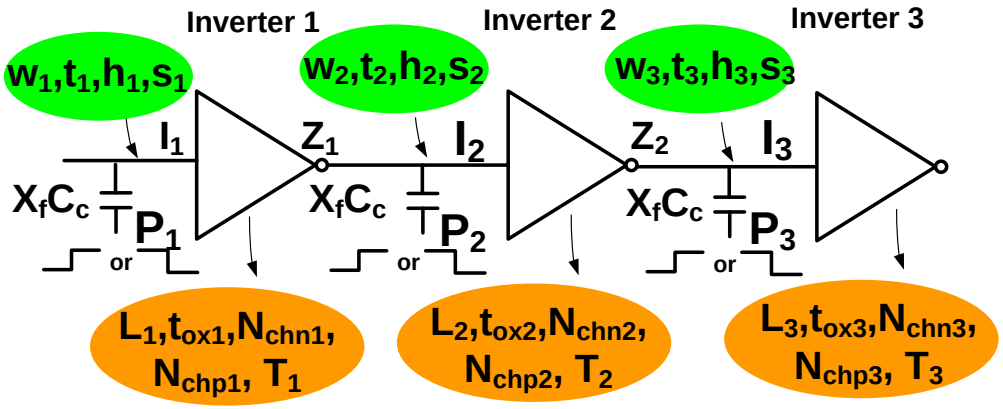


Device vs. Interconnect Variations



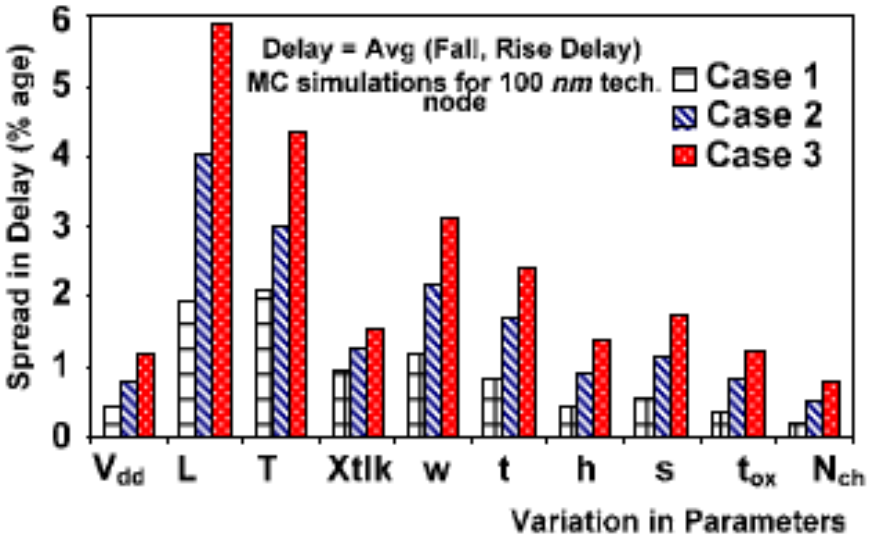
S. Nassif (IBM)

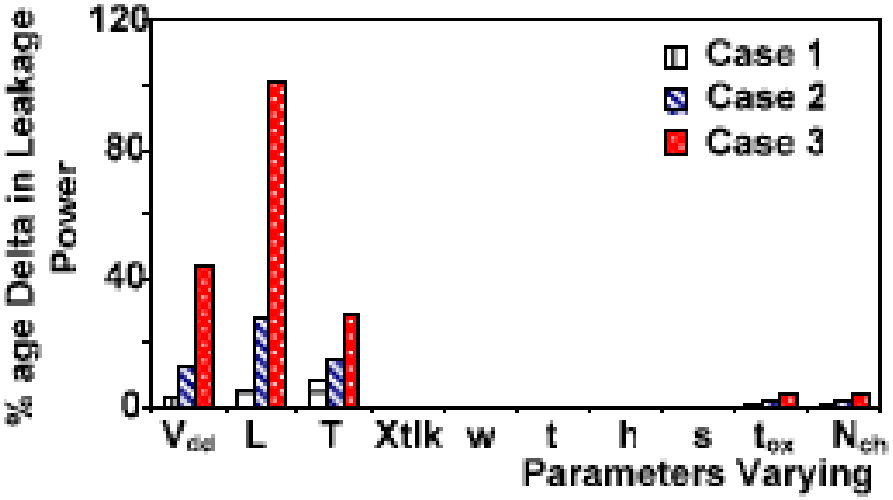
Sensitivity Analysis of Delay and Power



3 cases considered;
% variations increase from case 1 to case 3

Wason and Banerjee, ISLPED 2005

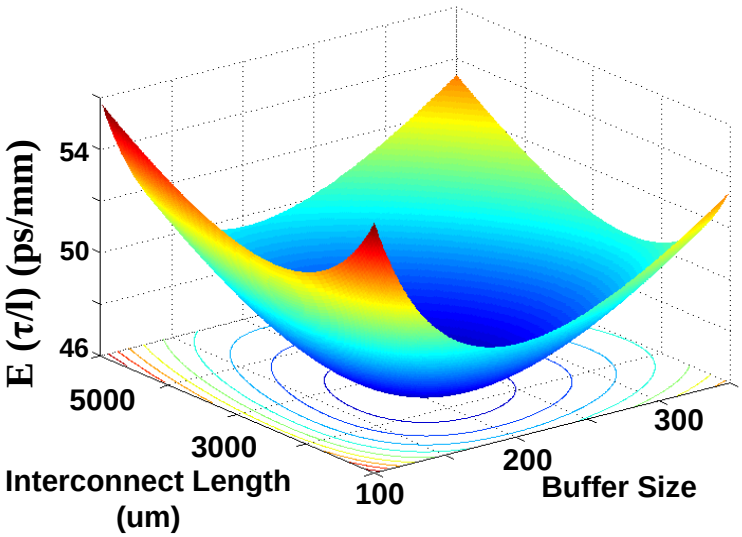


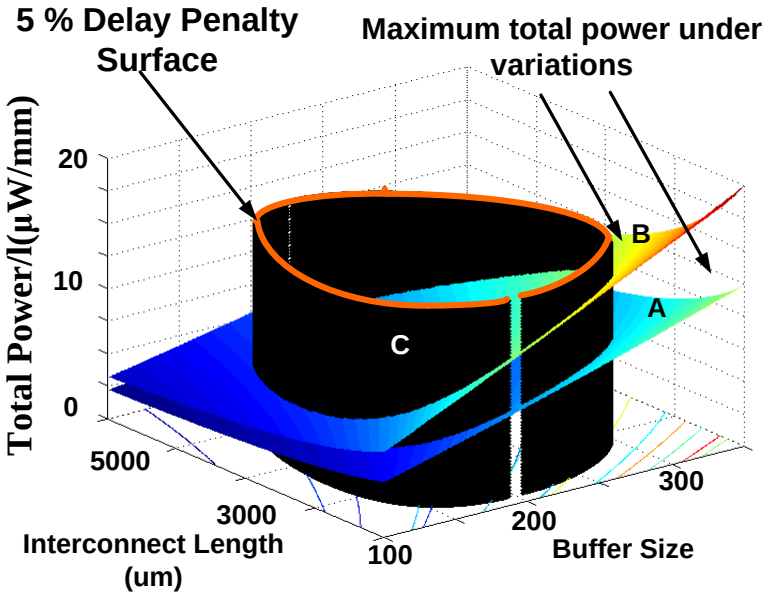


- Variations have significant impact on leakage power and delay

Interconnect Design Considering Variations

- Significant Implications for Interconnect Design
- Power-Optimal Repeater Insertion- tradeoff delay with power

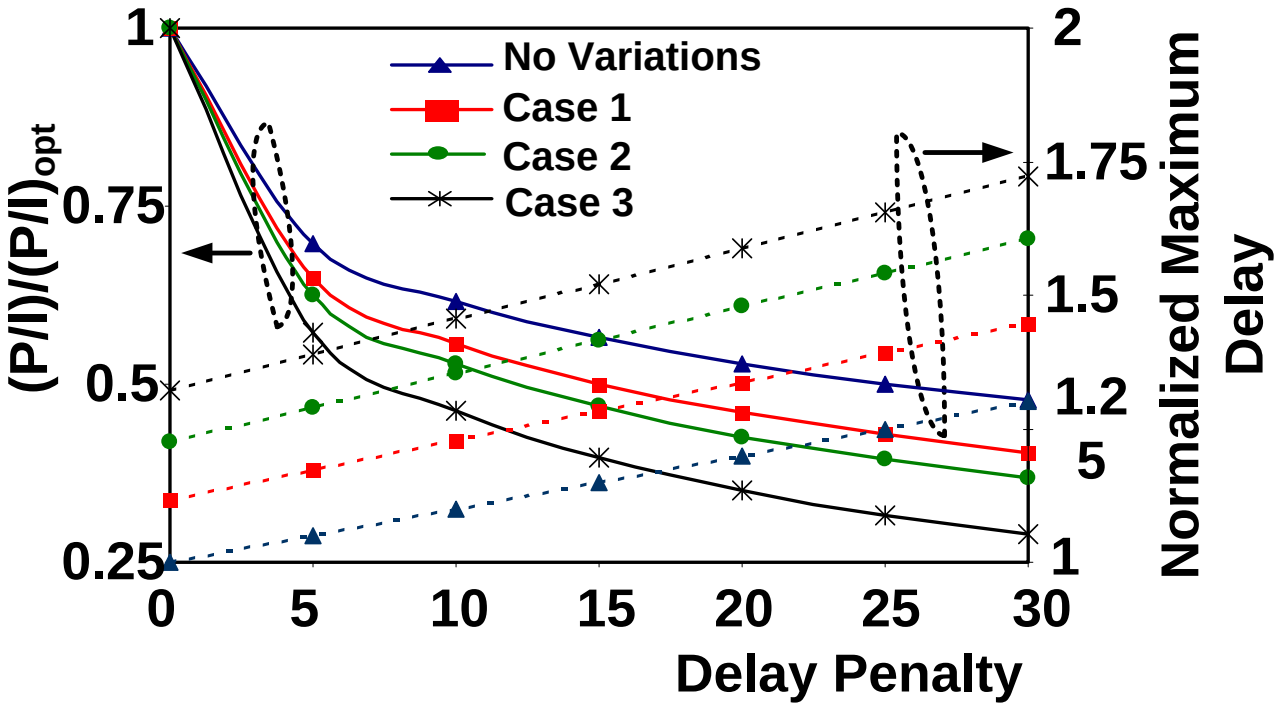




Wason and Banerjee, ISLPED 2005

Variations can significantly impact power-optimal repeater insertion

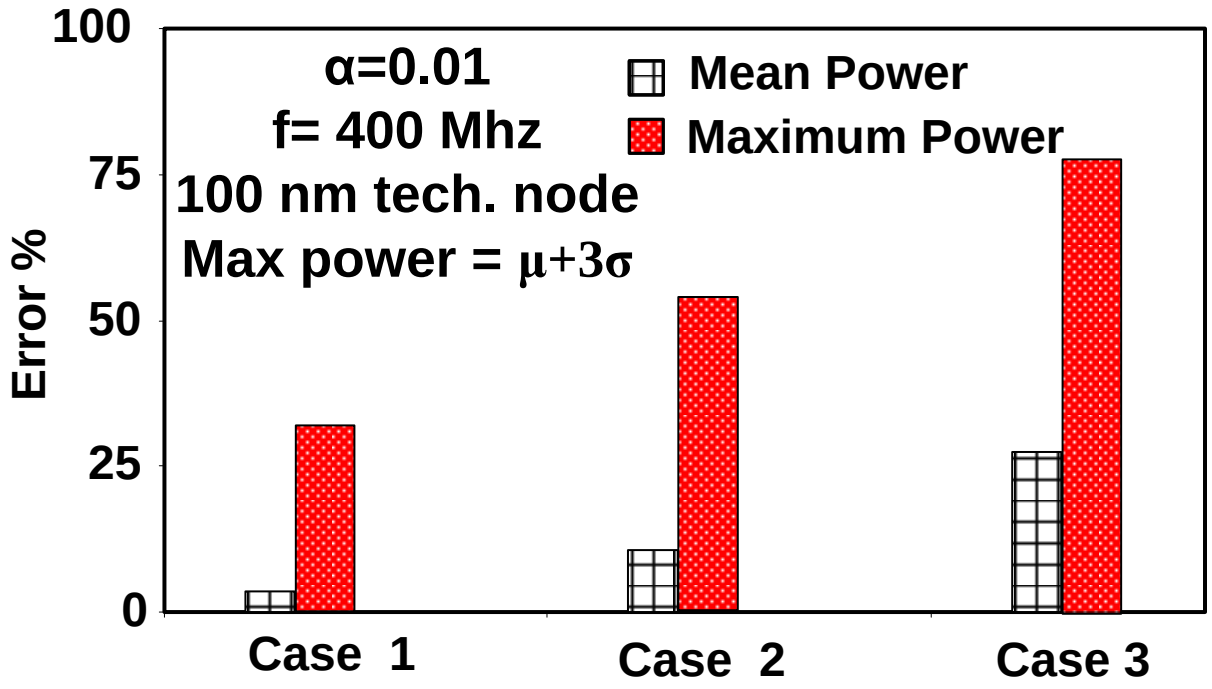
Optimization under Parameter Variations



Wason and Banerjee, ISLPED 2005

- The normalized delay increases under variations
- Power savings are higher for the same penalty in delay, as variations increase
- Importance of power-optimal repeater insertion increases

Implications for Power Estimations



- Error in power estimation can be significant if variations are neglected