Analysis of Substrate Thermal Gradient Effects on Optimal Buffer Insertion

Amir H. Ajami and Massoud Pedram
University of Southern California

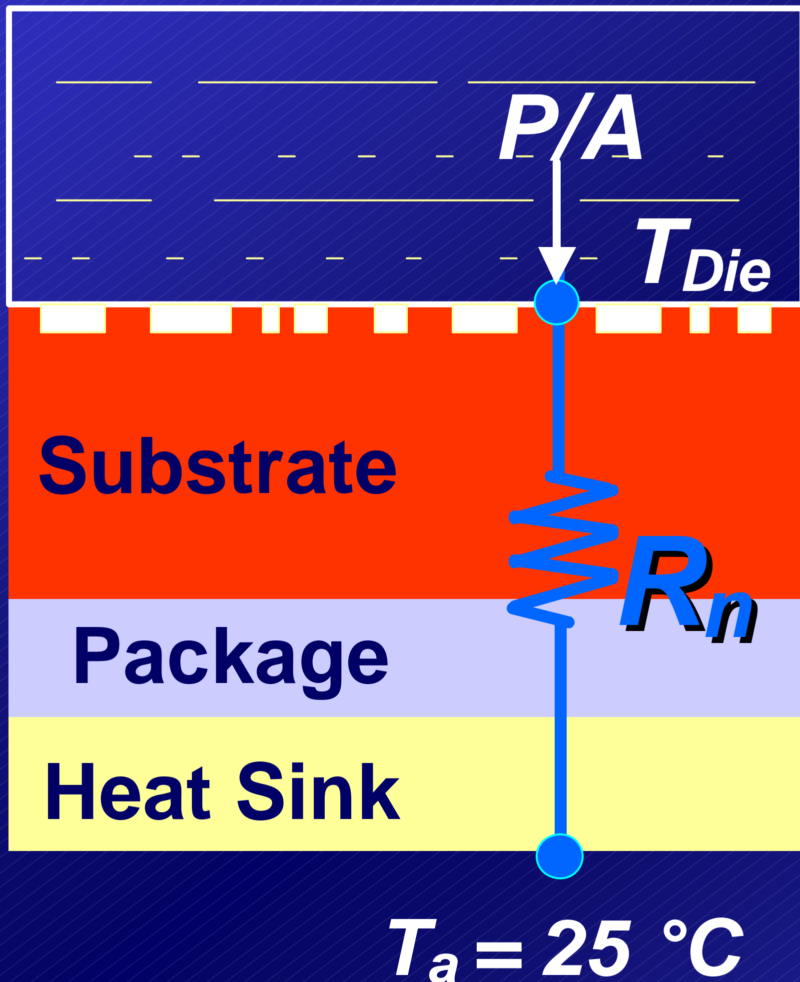
Kaustav Banerjee
Stanford University

Nov. 5th, 2001

Outline

- Introduction
- Non-Uniform Chip Temperature Profile
- Buffer Insertion Techniques
- Temperature-Dependent Buffer Insertion
- Summary

Average Chip Thermal Model

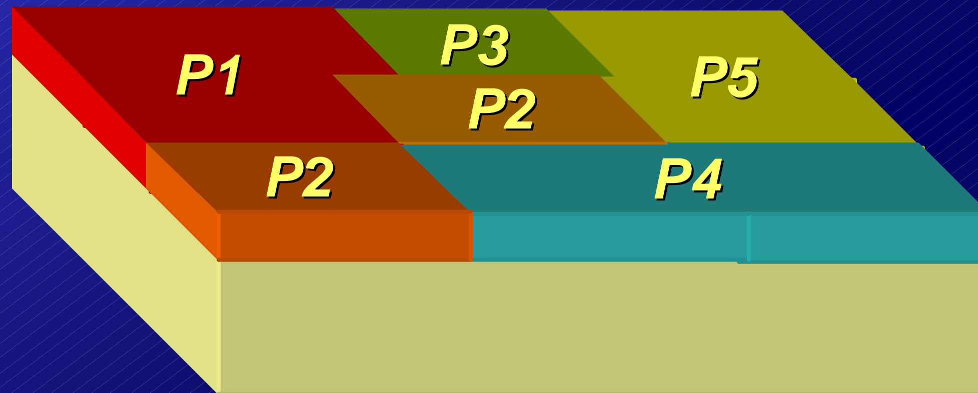


- 1-D heat conduction model

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

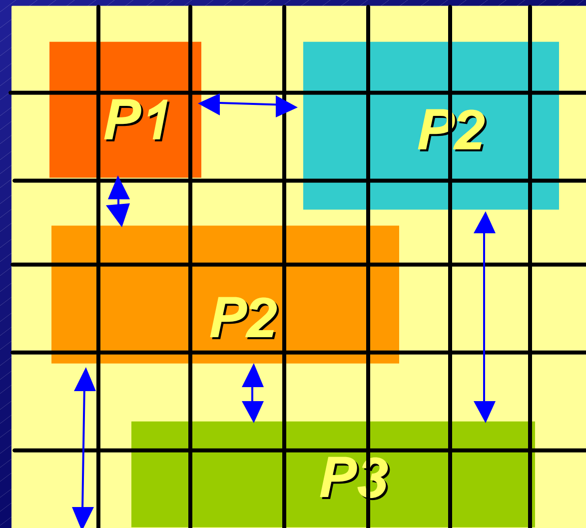
- Due to excessive Joule heating and the distance from the heat-sink, global interconnect lines are the hottest locations inside the chip

Non-Uniform Substrate Power Map



- 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

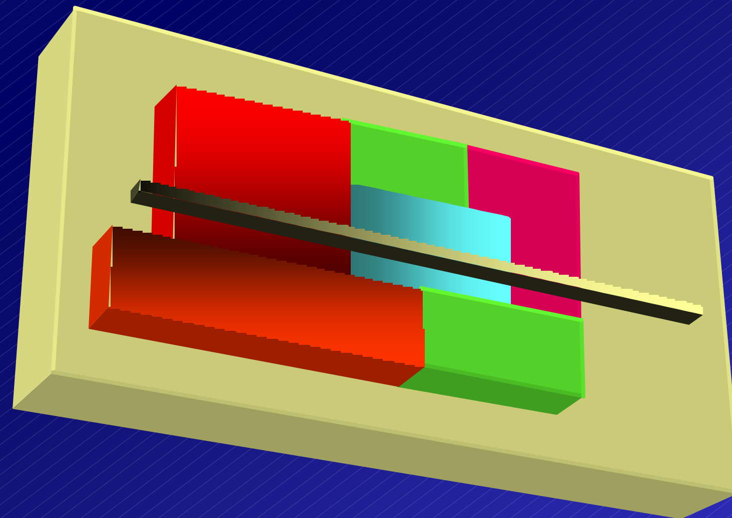
Substrate Temperature



- Consider 4 neighbors of each square and solve 1-D heat conduction model
- Generating matrix $\mathbb{R}_t \cdot \mathbb{P} = \mathbb{T}$ using 6 neighbors of each grid in 3-D space
- Using FST to speed up the computation (*Kang* et al.)

Non-Uniform Substrate Temperature

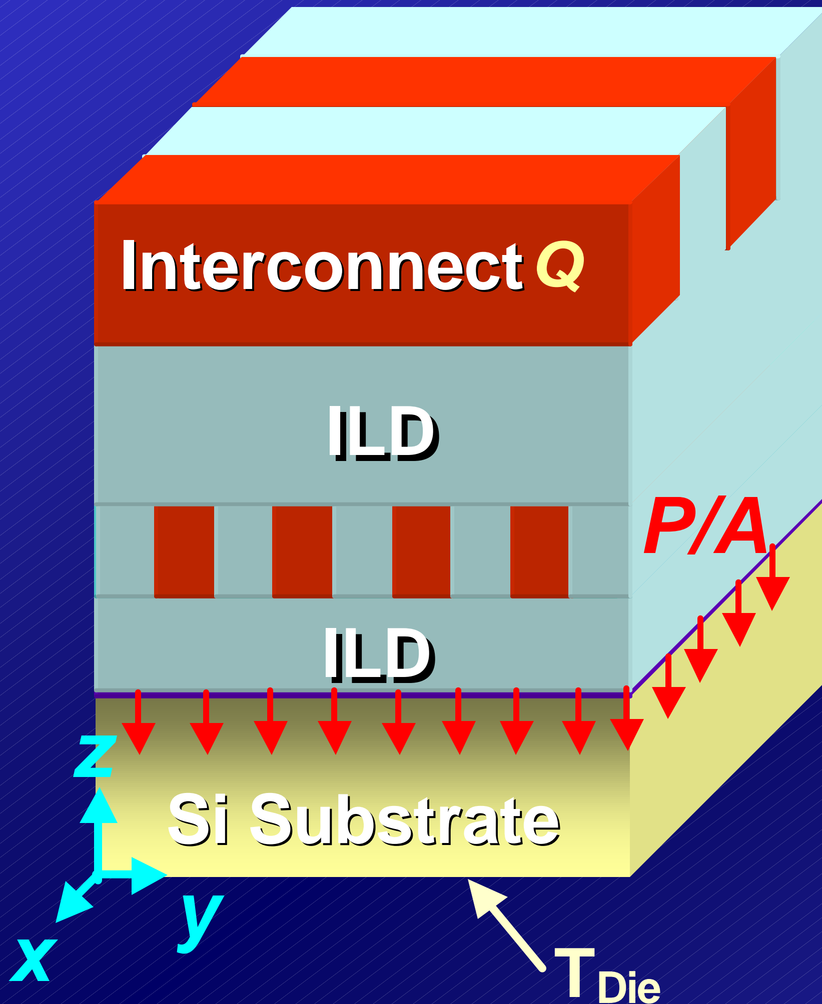
- Substrate thermal profile is non-uniform
  - Thermal time constant is of the order of *ms*
  - Switching activities in the block level are more important
  - Introduces non-uniformity in the global interconnect thermal profile



Outline

- Introduction
- Non-Uniform Chip Temperature Profile
- Buffer Insertion Techniques
- Temperature-Dependent Buffer Insertion
- Summary

Interconnect Thermal Profile



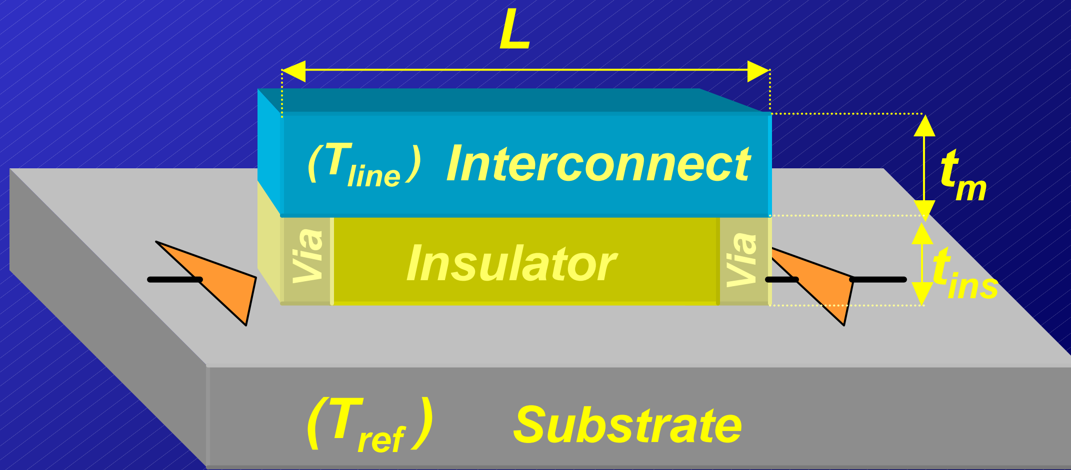
- Three dimensional heat conduction in steady state

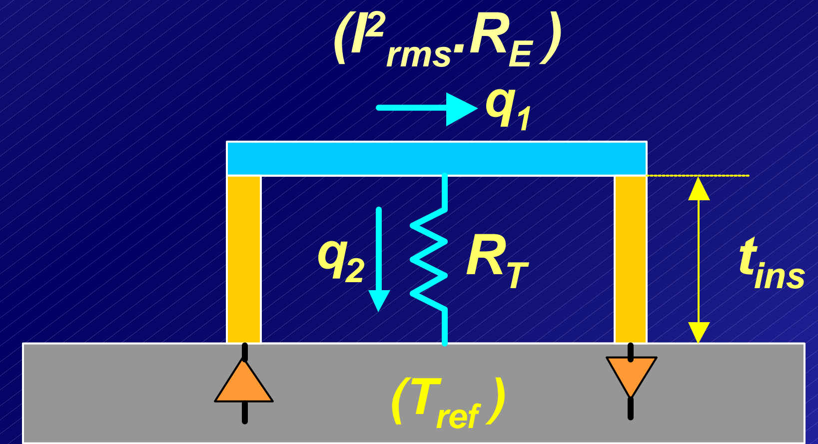
$$\tilde{N}^2 T = 0$$

- With an effective heat generation Q in the interconnect and a constant thermal conductivity k_m

$$\nabla^2 T + \frac{Q}{k_m} = 0$$

1-D Heat Equation for Interconnects





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

$Q = q_1 - q_2$

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

λ and θ are constants

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

Spatial Temperature Distribution

$T(x=0) = 100\ °C$

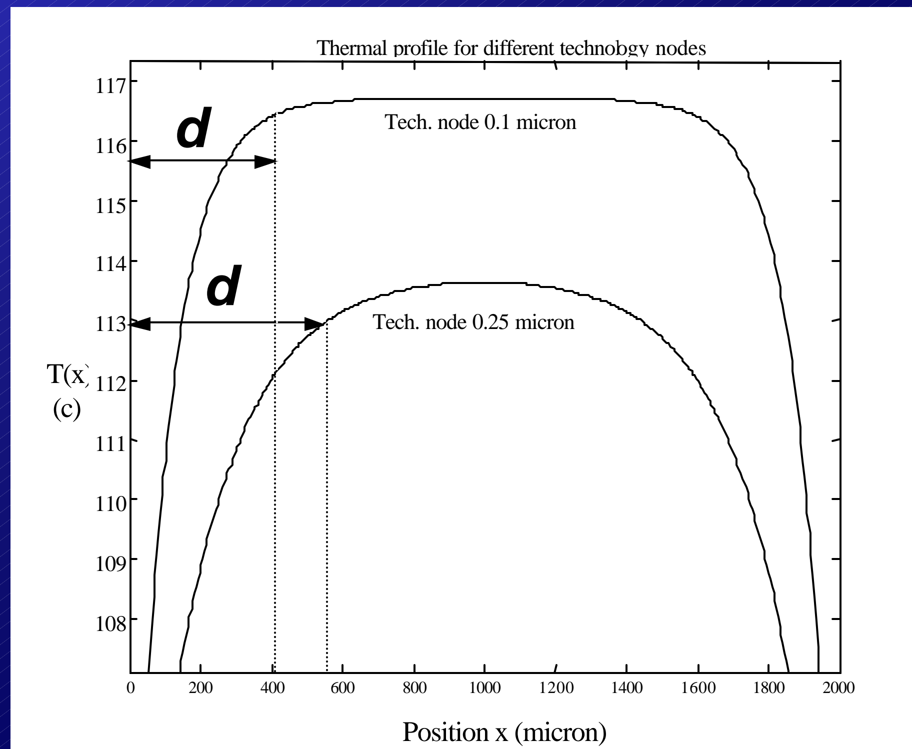
$T(x=2000) = 100\ °C$

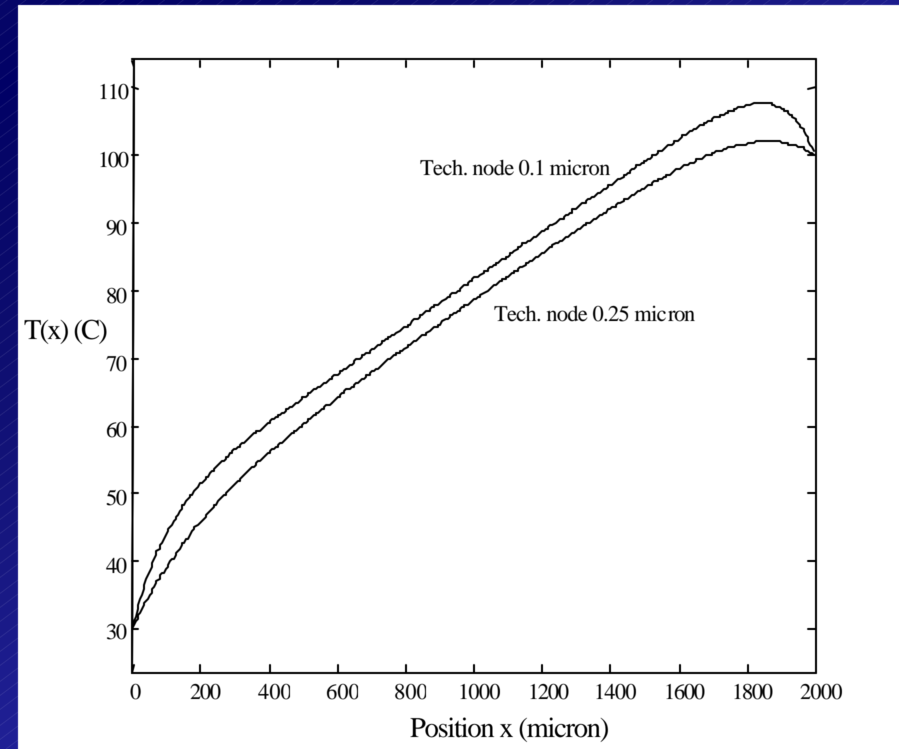
$L=2000\ \mu m$

$T(x=0) = 30\ °C$

$T(x=2000) = 100\ °C$

$T_{ref} = 100\ °C$





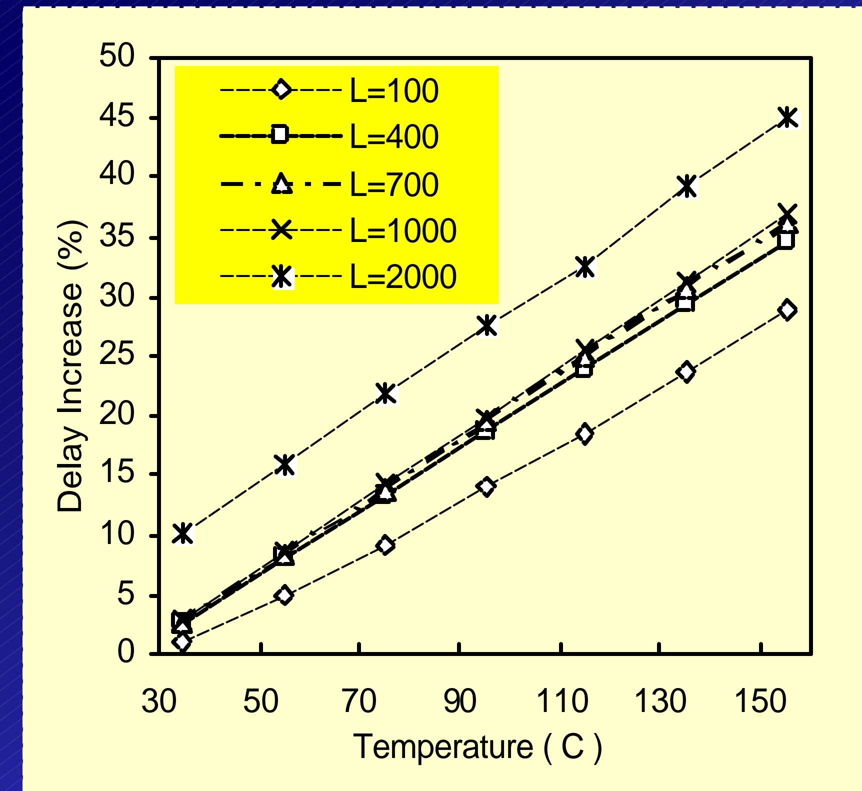
Temperature Dependency of Delay

- Interconnect delay dependent on *T* due to the *T* dependence of the resistance

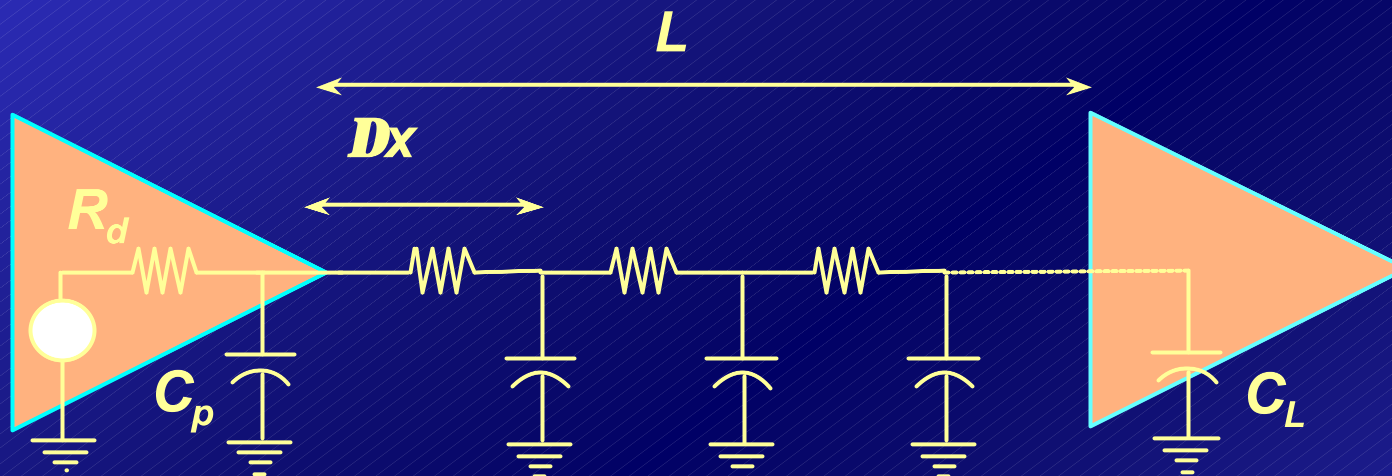
$$r(x) = ?_0(1 + ßT(x))$$

- r_0 : resistance per unit length at reference temperature
- b : temperature coefficient of resistance (1/°C)

T(x) = Constant



Non-Uniform Temperature-Dependent Delay



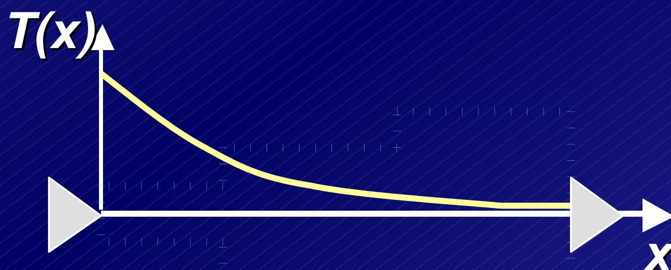
$$D = R_d\left(C_P + C_L + \int_0^L c_0(x)dx\right) + \int_0^L r_0(x)\left(\int_x^L c_0(h)dh + C_L\right)dx$$

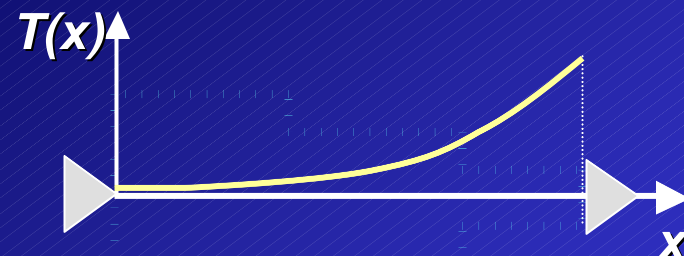
$$D = D_0 + (c_0 L + C_L)?_0 ß\int_0^L T(x)dx - c_0 ?_0 ß\int_0^L xT(x)dx$$

D_0 is the Elmore delay model at reference temp.

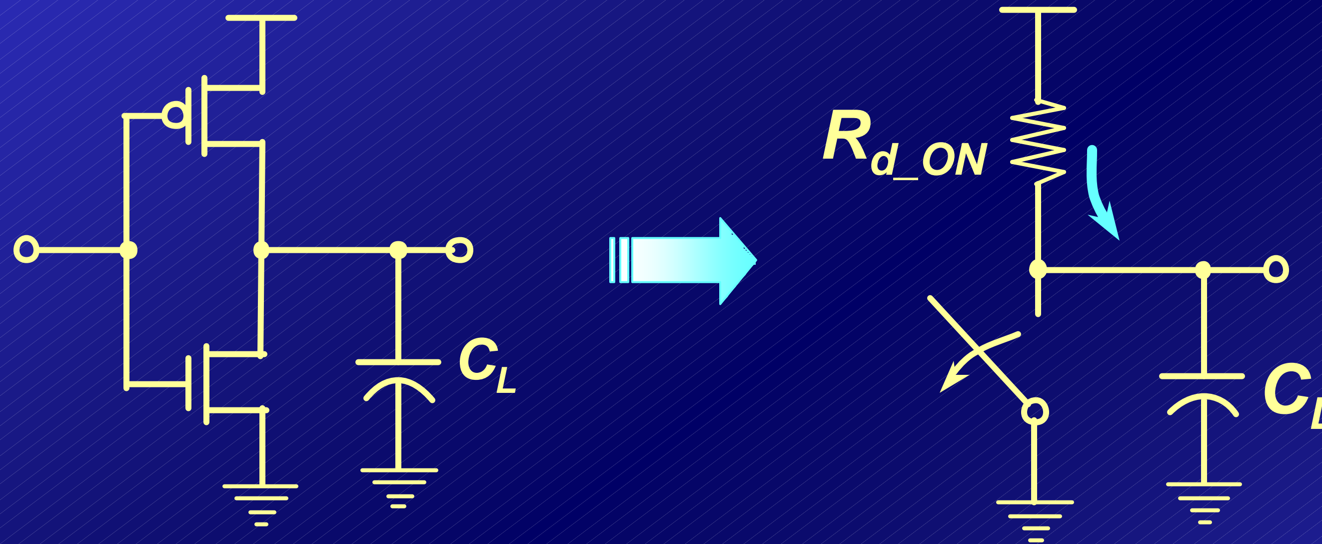
Direction of Thermal Profiles

- Decreasing (increasing) thermal profile is equivalent to increasing (decreasing) sizing profile for uniform resistance wire *(DAC'01)*
- Increasing thermal profile has better performance than that of decreasing thermal profile (optimal wire sizing)





Inverter ON-driving Resistance



$$R_{d_ON} \cong \frac{L_{eff}/w}{\mu C_{ox}(V_{DD} - V_T)}$$

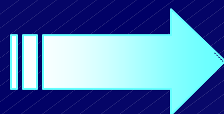
- V_T and *m* are dependent on the cell temperature

Temperature-dependent R_d

Q_B: Depletion region charge
C_{ox}: Gate oxide capacitance
E_g: Energy gap of Silicon
q: Electron charge
μ: Electron mobility
W: Gate width
L_{eff}: Channel width

$$V_T \cong 2j_f - \frac{Q_B}{C_{ox}} \cong E_g - \frac{Q_B}{C_{ox}}$$

$$\frac{\partial V_T}{\partial T} = \frac{E_g/q + V_T}{T}$$

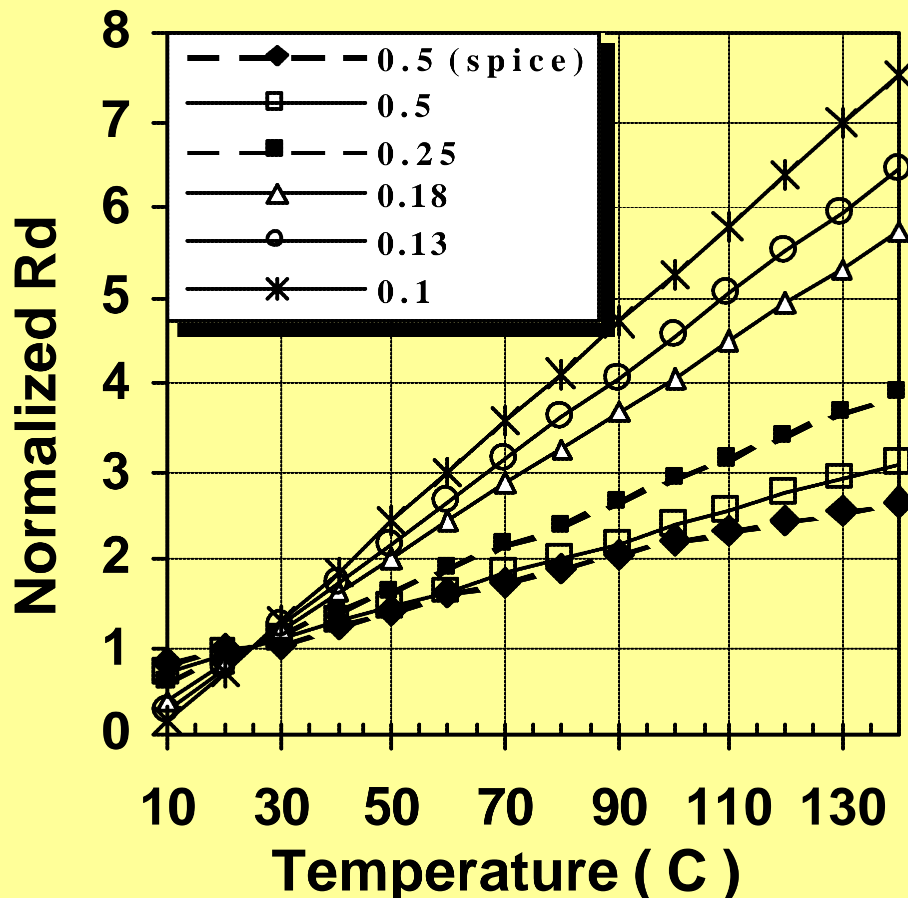


$$\frac{?R_d}{R_d} = \frac{E_g/q + V_T}{V_{DD} - V_T} \times \frac{?T}{T}$$

E_g/q » 1.12 V

ON-Resistance (R_d) Variations

Normalized to R_d at 25ºC



0.25 µm	V_T=0.6 V	V_{dd}=3.3 V
0.18 µm	V_T=0.36 V	V_{dd}=1.8 V
0.13 µm	V_T=0.3 V	V_{dd}=1.5 V
0.10 µm	V_T=0.24 V	V_{dd}=1.2V

$$R_d(x) = R_{d0}(1 + \beta_c T(x))$$

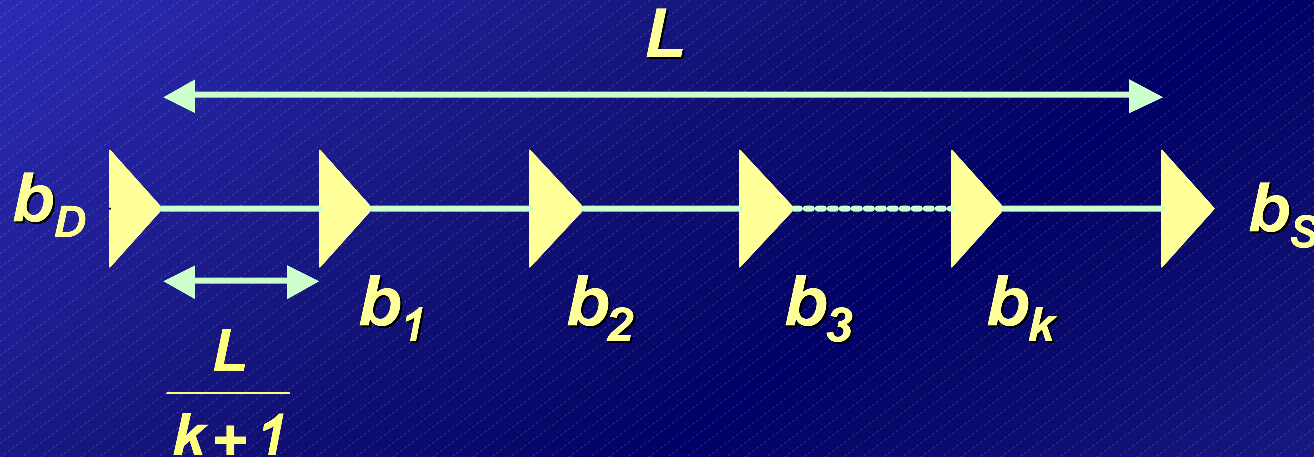
- **Thermal dependency of R_{d_ON} is much severe than that of R_{int}**

- Introduction
- Non-Uniform Chip Temperature Profile
- Buffer Insertion Techniques
- Temperature-Dependent Buffer Insertion
- Summary

Buffer Insertion

- Improving the performance in signal nets with high capacitive loads by inserting buffers
- Finding the number of inserted buffers, their sizes and locations along the the net in order to minimize the delay
- In a given technology, the critical length between each two buffers and optimal buffer sizes can be extracted from the technology parameters (*Otten* et al., *Alpert* et al.)

Methodology



$$l_{crit} = \sqrt{\frac{r_0 c_0 (1 + \frac{c_p}{c})}{rc}}$$

$$s_{opt} = \sqrt{\frac{r_0 c}{r c_0}}$$

r_0 : min. size transistor output *resist.*
c_0 : min. size transistor input *cap.*
c_p : min. size transistor parasitic *cap.*
r : unit length line *resistance*
c : unit length line *capacitance*

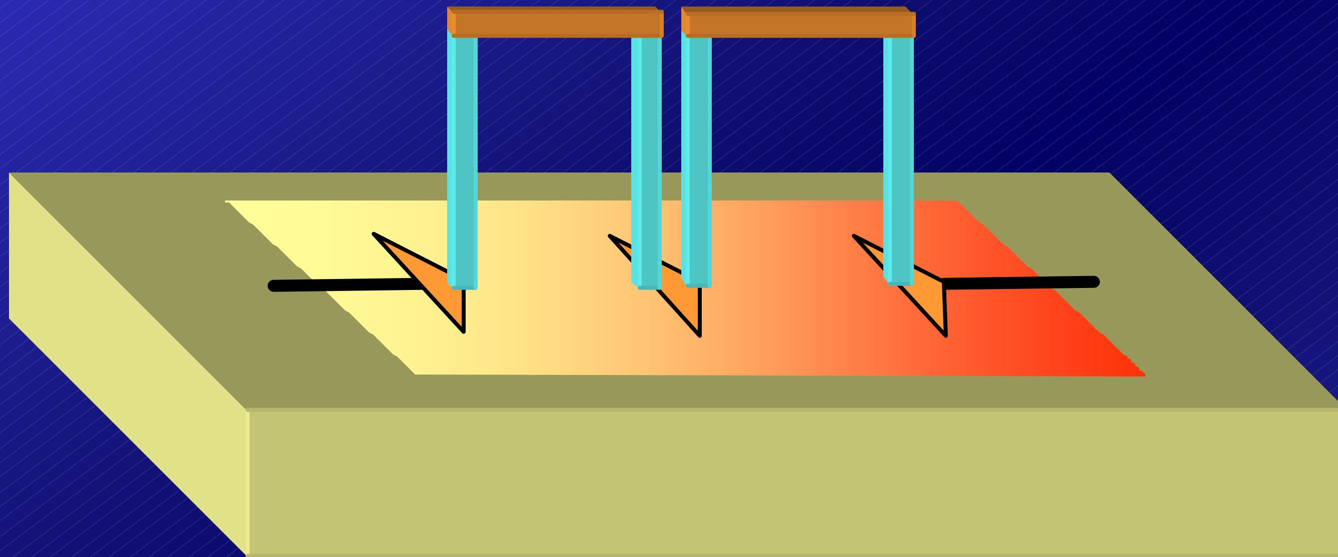
Methodology

- Equal distances between each two adjacent buffers while having identical source and sink buffers
- Uniform line resistance per unit length r and min size driving resistance r_0

$$k = \left\lfloor -0.5 + \sqrt{1 + \frac{2rcL^2}{R_d(C_L + C_p)}} \right\rfloor$$

$$R_d = \frac{r_0}{S_{opt}} \qquad C_L = c_0 . s_{opt} \qquad C_P = c_p . s_{opt}$$

Effects of Non-uniform Substrate Temp.

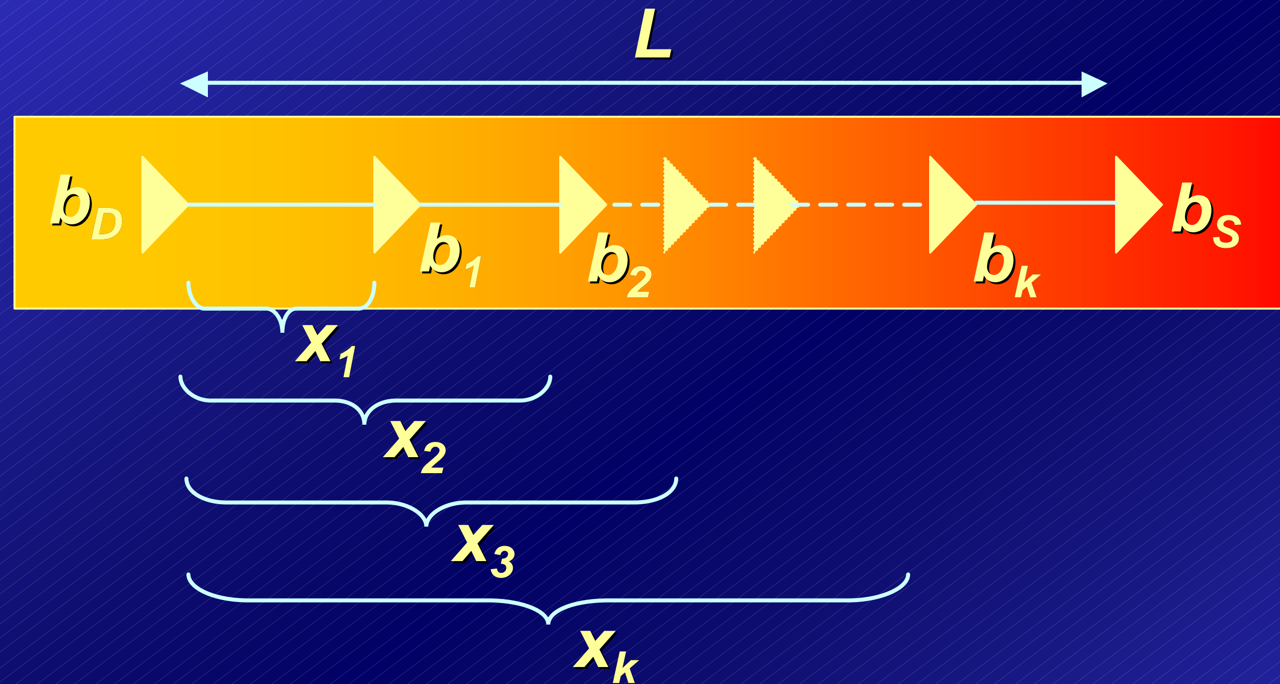


- Non-uniform substrate temperature causes:
  - Non-uniform interconnect resistance profile
  - Non-uniform ON-driving resistance profile for placed buffers

Outline

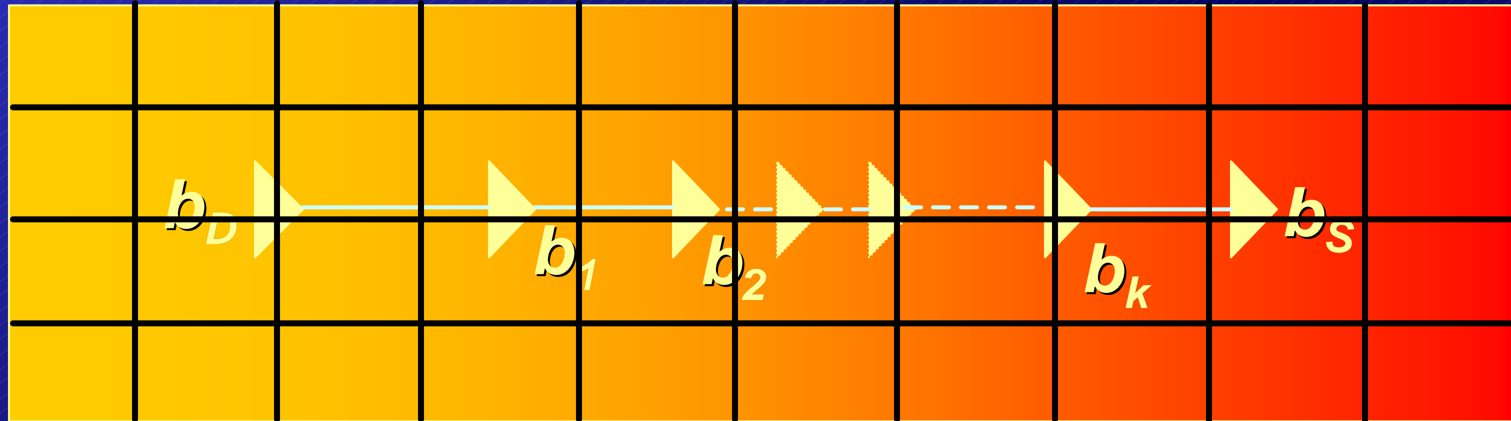
- Introduction
- Non-Uniform Chip Temperature Profile
- Buffer Insertion Techniques
- Temperature-Dependent Buffer Insertion
- Summary

Thermally-Dependent Buffer Insertion



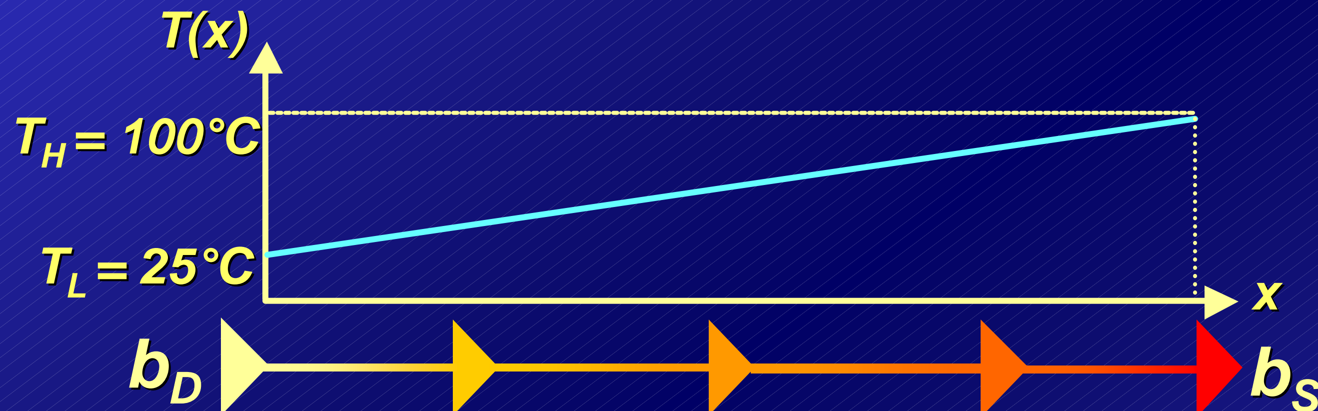
$$D = \sum_{i=1}^{k+1} \left(\int_{x_i}^{x_{i+1}} R(t)(c(x_i - t) + C_L) dt \right) + \sum_{i=1}^{k+1} R_d(x_{i-1})(cx_i - cx_{i-1} + C_L + C_p)$$

Assumption



- Temperature of each grid square is a function of the *total* power consumption of gates located in that square
- In steady state, each inserted gate reaches to the temperature of its surrounding area

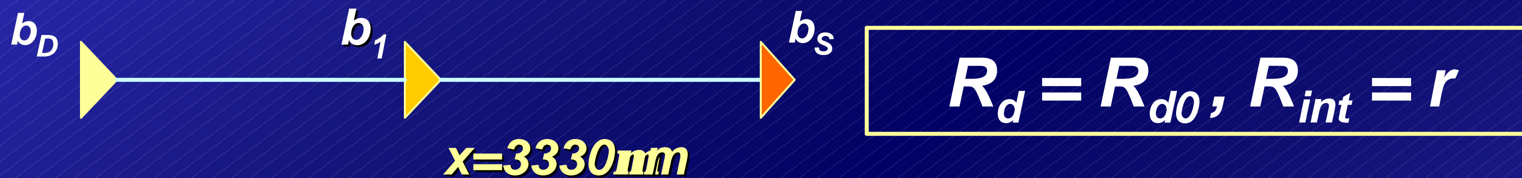
R_d vs. R_{int} Thermal Dependencies

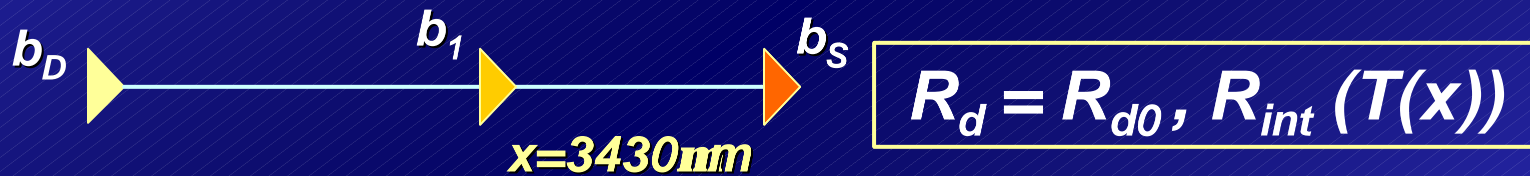


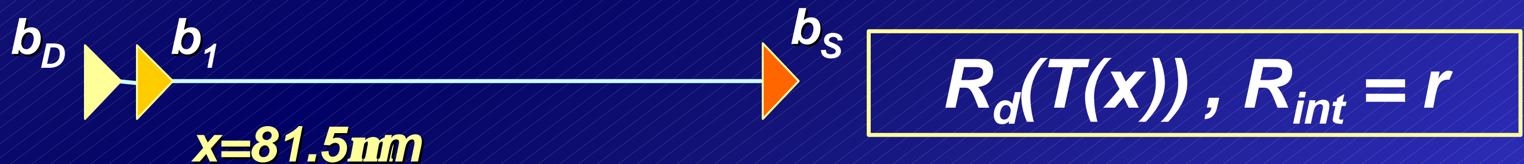
- Gradually increasing R_{int} pushes the inserted buffers toward the sink buffer
- Gradually increasing R_d pushes the buffers toward the source buffer

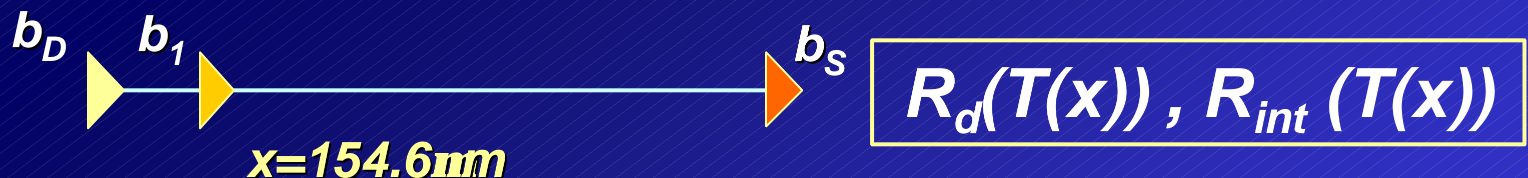
Buffer Movements

L=6660 μm (0.18μm)



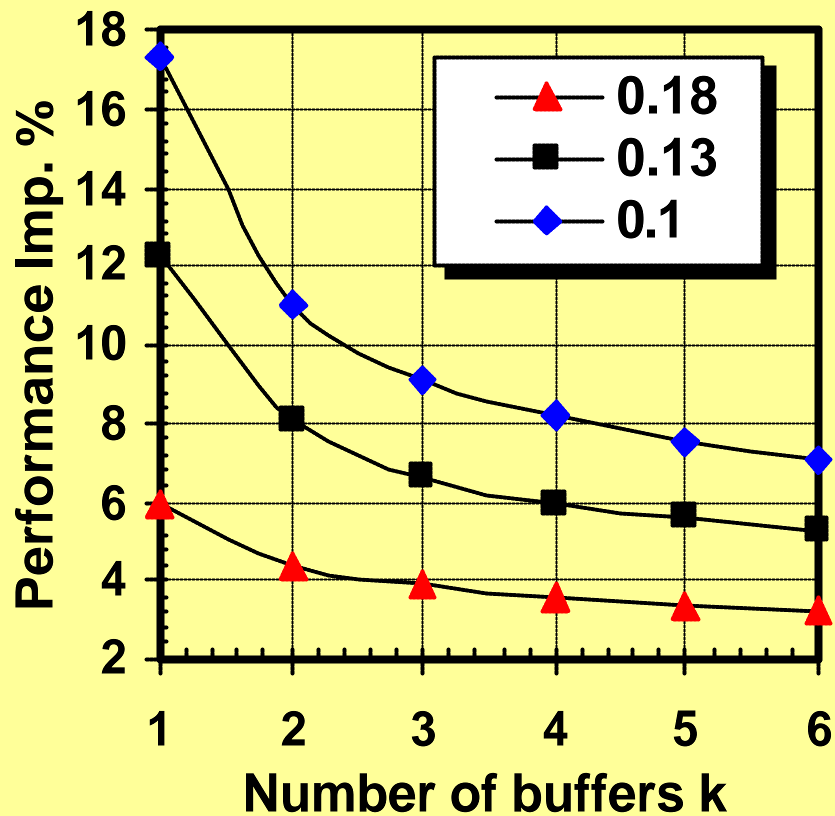




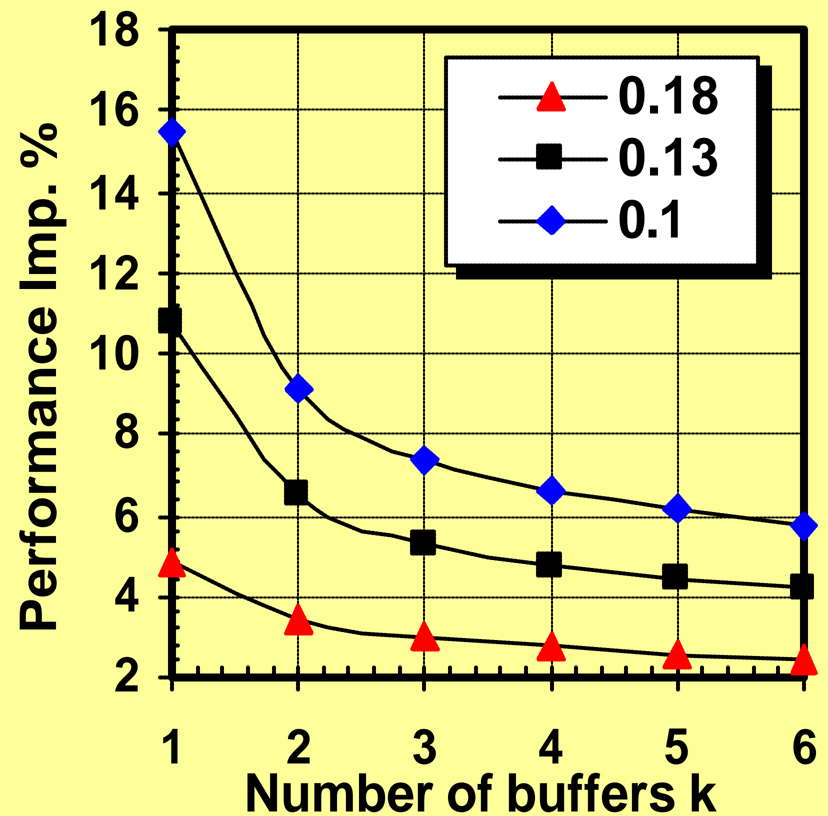


Performance Improvement

$R_d(T(x))$, $Rint = r$

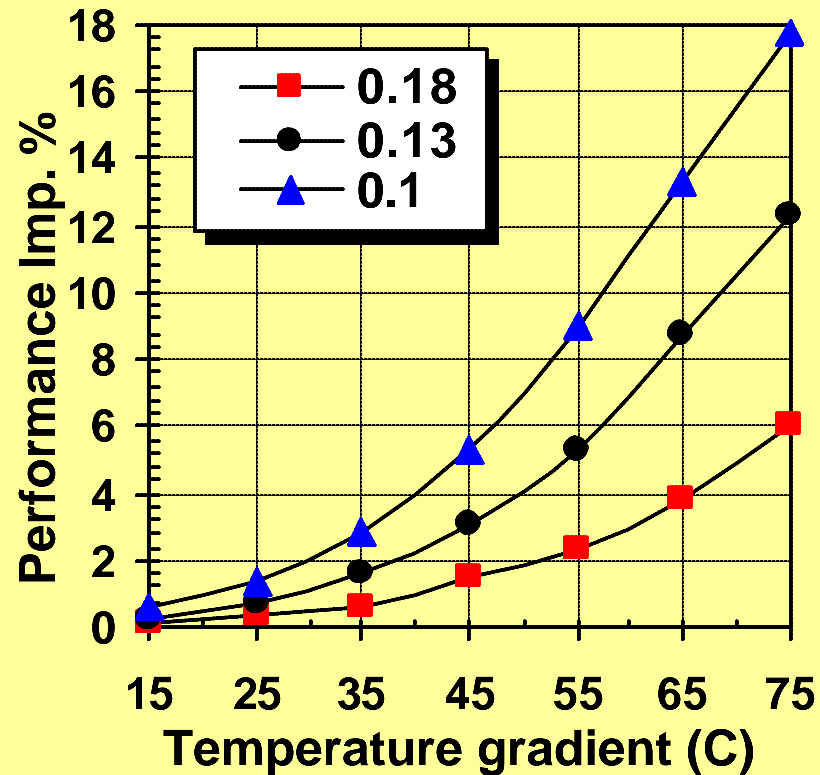


$R_d(T(x))$, $Rint(T(x))$



Effect of Thermal Gradient Magnitude

L=6660 μm (0.18μm)



- Due to different switching activities along with low power design policies, substrate & interconnect thermal maps are non-uniform
- Substrate thermal non-uniformities:
  - Affects the signal performance in interconnects
  - Severely impacts the device switching performance
  - Have serious effects on different EDA flow steps, specifically the buffer insertion routines
- Non-uniform substrate thermal profiles must be considered in the design flow of high-performance VLSI systems