Nanometer Scale Issues for On-Chip Interconnections

Kaustav Banerjee
University of California, Santa Barbara

UCSB
IC performance is being dominated by interconnects.....

![Graph showing optimal interconnect delay and typical gate delay vs. feature size.](ITRS_99.png)

Sub-100 nm Interconnect Issues....

RC Delay ↑

Circuit Solutions: Repeater Insertion

Large # of Repeaters: Power Dissipation

Power-Performance Tradeoffs? Need RLC Delay?

Materials Solutions: Cu/Low-k

Macro scale

Thermal Effects

Micro scale

- Non-uniform Die Temperature
- Effect of Leakage Power

- Reliability Implications
- Reliability-Performance Tradeoffs?
- Implications for IR-Drop

Kaustav Banerjee
Materials Solutions
Low-k Dielectrics

IBM Says Process Can Make Chips More Powerful

By Bruce Meyerson
Associated Press

NEW YORK — IBM Corp. has developed a method of insulating tiny circuits on a computer chip so they can be placed even closer together, making room for more computing power on each semiconductor.

The new process, which coats the circuits with a chemical polymer instead of silicon dioxide, can either improve a chip's processing speed by as much as 30 percent or halve the amount of power it uses, IBM said yesterday.

The insulation reduces magnetic interference, or "cross-talk," between the electrical pulses that travel along millions of copper circuits that a chip uses to crunch data.

Because the electrical pulses are better shielded from the magnetic fields generated by neighboring pulses, the distance between circuits can be reduced by more than a quarter to 0.13 microns -- a measurement 600 times thinner than a human hair.

Thanks to the space savings, chipmakers can double the number of transistors they cram onto a piece of silicon to about 400 million, said Bijan Davari, a vice president of IBM's semiconductor research and development center in East Fishkill, N.Y.

Chips made with the new insulation are expected to be available next year, IBM said.

The first application will be to produce powerful microprocessors for major business systems like those used to run Web sites and communications networks.

Old dielectric: SiO₂, K = 4.0
New dielectric: Silk, K = 2.5
Ultimate low-k: air-gap, K=1
Thermal Conductivity of Dielectrics

Im and Banerjee, IEDM 2000.

Kaustav Banerjee
IUMRS-ICAM 2003
Yokohama, Oct. 8-13
Low-k Increases Thermal Effects

K. Banerjee et al., IEDM 1996

Impact of Scaling Using Low-k

As $W$ decreases, $SH$ increases.

Low-k increases $SH$ by 10-15%
Cu Resistivity: Effect of Scaling

- **Effect of Cu diffusion Barrier**
  - Barriers have higher resistivity
  - Barriers can’t be scaled below a minimum thickness

- **Effect of Electron Scattering**
  - e scattering from the surface
  - increases effective resistivity

**Problem is worse than anticipated in the ITRS roadmap**
Maximum Chip Temperature

(FEM Simulation)

Temperature [°C]

Power Density [W/mm²]

Technology Node [nm]

Tmax

TDie

Global Wires

209 °C

126 °C

50 nm Node

Im and Banerjee, IEDM 2000.

Kaustav Banerjee

IUMRS-ICAM 2003

Yokohama, Oct. 8-13
Implications for Reliability
Electromigration (EM) 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 °C)

- The \( j_{avg} \) limit does not comprehend self-heating
Coupled Analysis

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

Self-consistent

Kaustav Banerjee

IUMRS-ICAM 2003

Yokohama, Oct. 8-13

K. Banerjee et al., IRPS 2000
Coupled Analysis

Unipolar vs Bipolar:
Metal 6, 180 nm node, $j_0 = 0.6 \text{ MA/cm}^2$, $R = 0.5$

![Graph showing self-consistent metal temperature $T_m$ and maximum allowed $j_{rms}$ vs duty cycle $r$.](image)
Current Density Limits in Signal Lines: Reliability Based

Impact of Low-k Materials

- Coupled $j_{rms}$ values decrease significantly as low-k materials are introduced.
Current Density Limits: Performance Based

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

\[ j_{rms}^{(\text{max})} \text{ occurs close to the repeater output due to the distributed nature of the interconnect} \]
Performance vs Reliability

For signal lines reliability design limits satisfied

Kaustav Banerjee
Buffers connected to global lines by a series of vias

Vias at the local and semi-global tiers are pitch-matched to the metal lines at these levels

Vias must carry the same current as the global lines

Even for the signal lines:

\[ J_{\text{via \ (performance)}} > J_{\text{via \ (Reliability)}} \]
### Local Via Temperature Rise: DC Case

**3-D FEM Analysis**

<table>
<thead>
<tr>
<th>Technology Node [nm]</th>
<th>(I_{\text{max}}) (Global Wire)</th>
<th>(I_{\text{max}}) (Semi-Global Wire)</th>
<th>(I_{\text{max}}) (Local Wire)</th>
<th>(I_{\text{max}}) (Via)-ITRS</th>
</tr>
</thead>
<tbody>
<tr>
<td>160</td>
<td>60</td>
<td>50</td>
<td>40</td>
<td>30</td>
</tr>
<tr>
<td>120</td>
<td>100</td>
<td>90</td>
<td>80</td>
<td>70</td>
</tr>
<tr>
<td>80</td>
<td>150</td>
<td>140</td>
<td>130</td>
<td>120</td>
</tr>
<tr>
<td>40</td>
<td>200</td>
<td>190</td>
<td>180</td>
<td>170</td>
</tr>
<tr>
<td>0</td>
<td>250</td>
<td>240</td>
<td>230</td>
<td>220</td>
</tr>
</tbody>
</table>

**ITRS ref. temp. (105 °C)**

- Average via temp. = 197 °C

- (45 nm node)
  - Avg. via temp. = 197 °C

*Im, Banerjee and Goodson, IEDM 2002*
IR-Drop
IR-Drop in P/G Networks

- Degrades device switching speed and DC noise margins
- Can cause functional failure in dynamic logic and timing violation in static logic
- 10% voltage-drop $\rightarrow$ 8% increase in device propagation delay (0.18\(\mu\)m tech.)
Technology Scaling Impacts

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

Full Chip Worst-Case IR-Drop

- Allocation of 10% of the routing area for wire-sizing to minimize the IR-drop
Circuit Solutions
Increasing Number of Repeaters

![Bar chart showing the increasing number of full-chip repeaters across different technology nodes. The chart includes data up to 0.04 μm technology node with interconnect lengths scaling by 0.7x normalized to 0.18 μm technology node. The chart is courtesy of R. Krishnamurthy, Intel Corporation.]

Courtesy of R. Krishnamurthy, 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!
Interconnect Delay Optimization

For \( s = s_{\text{opt}}/2 \) and \( l = 2 l_{\text{opt}} \); Delay Penalty is only 25%
Power-Optimal Repeater Insertion

- Developed a methodology to estimate repeater size and inter-repeater wire length which minimizes the total interconnect power dissipation for a given delay penalty

- Estimated power-optimal buffering schemes for various ITRS technology nodes

- Analyzed relative importance of various components of power dissipation as a function of technology scaling

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.

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

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
RLC Delay?
Motivation

- Increasing clock speeds and decreasing rise times cause “inductive” effects
- Low resistance wires more susceptible to inductive effects
  - global wires
  - Copper interconnects
- Line inductance depends on current return path
  - strongly dependent on input vectors
  - difficult to predict a-priori
  - large variations
Signal Delay

![Graph showing signal delay with time (ns) on the x-axis and \( \frac{v(t)}{V_0} \) on the y-axis. The graph compares RLC and RC circuits, with RLC showing a faster response.]
Equivalent RLC Circuit

Minimum Delay Per Unit Length Comparison

![Graph showing minimum delay per unit length comparison for different line inductances (nH/mm) and process technologies.]
Macro-Scale Thermal Effects
On-Chip Temperature Variations

- Activity & ambient change
- Dynamic: 100-1000us

High-Performance Microprocessor

Temp (°C)

108
105.2
102.4
99.53
96.7
93.86
91.03
88.2
85.36
82.53
79.7
76.87
74.03
71.2
69.37
65.53

Cache 70°C
Core 120°C

Courtesy of S. Borkar, Intel Corporation.
Non-uniform Interconnect Thermal Profile

\[ \frac{d^2 T_{line}}{dx^2} = \frac{Q}{k_m} \]

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

\[ Q = q_1 - q_2 \]

\[ \lambda \text{ and } \theta \text{ are constants} \]

\[ f(L, t_m, k_m, t_{ins}, k_{ins}, I_{rms}, R_E) \]
# Impact on Clock Skew

Ajami, Pedram and Banerjee, CICC 2001

<table>
<thead>
<tr>
<th>( T_{\text{line}}(x) )</th>
<th>( \text{params} )</th>
<th>( l = l^* )</th>
<th>( l = L/2 \skew% )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T(x) = ax + b )</td>
<td>( T_H=170, T_L=90 )</td>
<td>1042</td>
<td>5.42</td>
</tr>
<tr>
<td>( a = \frac{T_H - T_L}{L} ) ( b = T_L )</td>
<td>( T_H=170, T_L=110 )</td>
<td>1032</td>
<td>3.98</td>
</tr>
<tr>
<td>( a = \frac{T_H - T_L}{L} ) ( b = T_L )</td>
<td>( T_H=170, T_L=130 )</td>
<td>1021</td>
<td>2.65</td>
</tr>
<tr>
<td>( T(x) = a \cdot e^{-bx} )</td>
<td>( T_H=170, T_L=90 )</td>
<td>957.5</td>
<td>5.24</td>
</tr>
<tr>
<td>( b = \frac{1}{L} \ln\left(\frac{T_H}{T_L}\right) ) ( a = T_H )</td>
<td>( T_H=170, T_L=110 )</td>
<td>968.66</td>
<td>3.63</td>
</tr>
<tr>
<td>( b = \frac{1}{L} \ln\left(\frac{T_H}{T_L}\right) ) ( a = T_H )</td>
<td>( T_H=170, T_L=130 )</td>
<td>979.5</td>
<td>2.40</td>
</tr>
<tr>
<td>( T(x) = T_{\text{max}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} )</td>
<td>( \mu=2000, \sigma=1000 )</td>
<td>1210</td>
<td>7.78</td>
</tr>
<tr>
<td>( T(x) = T_{\text{max}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} )</td>
<td>( \mu=1000, \sigma=400 )</td>
<td>1000</td>
<td>0.0</td>
</tr>
<tr>
<td>( T(x) = T_{\text{max}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} )</td>
<td>( \mu=300, \sigma=700 )</td>
<td>911</td>
<td>9.57</td>
</tr>
</tbody>
</table>
Impact of Leakage Current on Junction Temperature

For leakage dominant technologies the chip power is strongly coupled to the junction temperature ....

Self-consistent junction temperature estimation is needed, which will impact interconnect temperature....

K. Banerjee et al., to appear in IEDM 2003
Conclusions

● Micro-scale thermal effects:
  ● Self-heating—increases with low-k, interface electron scattering, and junction temperature
  ● Coupled analysis of self-heating and EM is necessary—impacts interconnect design rules
  ● Current continuity, thermal coupling, and chip junction temperature rise affects via design and scaling
  ● IR-drop in the power/ground network—strongly affected by global interconnect thermal profile

● Macro-scale thermal effects:
  ● On-chip temperature variations: an increasingly important issue
  ● Non-uniform die temperature affects signal integrity, clock skew, buffer insertion
  ● For leakage dominant technologies, various electrothermal couplings need to be modeled for accurate junction and interconnect temperature estimation
Conclusions (Cont’d)

- A methodology has been developed to estimate repeater size and inter-repeater wire length which minimizes the total interconnect power dissipation for a given delay penalty.
- Relative power savings increase with scaling due to increasing leakage current with technology scaling.
- Inductive effects diminish for scaled on-chip lines.
Thank You!