Interconnect impact on circuit performances in Deep Sub-Micron technologies

Cristiano Forzan

Bologna, June 2003
Outline

- CMOS device and interconnect scaling overview
- Interconnect resistance effects
- Coupling capacitance impact on performances
- Interconnect inductance: just around the corner
- References
CMOS technology scaling overview

1μm technology (1990) 0.13μm technology (2001)
Ideal CMOS scaling

<table>
<thead>
<tr>
<th>Scaled Parameter</th>
<th>Ideal Scaling Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>W, L, tox</td>
<td>S</td>
</tr>
<tr>
<td>NA</td>
<td>1/S</td>
</tr>
<tr>
<td>Vdd, Vth</td>
<td>S</td>
</tr>
</tbody>
</table>
# Interconnect reverse scaling

<table>
<thead>
<tr>
<th></th>
<th>Local Wiring</th>
<th>Global Wiring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linewidth &amp; Spacing</td>
<td>$S$</td>
<td>$S$</td>
</tr>
<tr>
<td>Wire Thickness</td>
<td>$S$</td>
<td>$S$</td>
</tr>
<tr>
<td>ILD Thickness</td>
<td>$S$</td>
<td>$S$</td>
</tr>
<tr>
<td>WireLength</td>
<td>$S$</td>
<td>$1/\sqrt{S}$</td>
</tr>
<tr>
<td>Resistance</td>
<td>$1/S^2$</td>
<td>$1/S^2$</td>
</tr>
<tr>
<td>Capacitance</td>
<td>$1$</td>
<td>$1$</td>
</tr>
<tr>
<td>RC delay</td>
<td>$1$</td>
<td>$1/S^3$</td>
</tr>
<tr>
<td>Current density</td>
<td>$1/S$</td>
<td>$1/S$</td>
</tr>
</tbody>
</table>
DSM scaling impact on delay

- Gate delay scales
- Local interconnect delay scales
- Global interconnect delay dominates
  - Global interconnects do not scale with feature sizes

(from ITRS 99)
Keeping wire resistance low

- Increase the wire aspect ratio \((AR = \text{height} / \text{width})\)
- Use copper instead of aluminium
Interconnect capacitances

Metal 2

Metal 1

Substrate

CROSSOVER

COUPLING

FRINGE

AREA

CROSSOVER

COUPLING

FRINGE

AREA
Interconnect capacitance trends

- Interconnect average length and routing density increase
- Wire aspect ratio increases
- Coupling capacitance dominates over ground capacitance
- Interconnect capacitance dominates gate capacitance
Traditional top-down design flow

- Any top-down flow requires some predictive capabilities

- Traditional top-down flows are effective when performances can be predicted with the wireload model
The evolutionary response to the DSM problem has been a stronger integration between the synthesis and the place&route process.

This is what Physical Synthesis does today.

Revolutionary breakthrough?
Outline

- CMOS and interconnect scaling overview
- Interconnect resistance effects
- Coupling capacitance impact on performances
- Interconnect inductance: just around the corner
- References
Interconnect resistance impact

- For technologies up to 0.35\(\mu m\) delay estimation based on the wireload model can be accurate
  - Few iterations for timing closure

- For DSM technologies (0.25\(\mu m\) and below) RC distributed effects become dominant
  - Many iterations and no guarantee to achieve timing closure
Wire resistance impact on delay

\[ T_{\text{stage}(A-R)} = T_{\text{gate}(A-Z)} + T_{\text{wire}(Z-R)} \]

\[ f(T_{\text{IN}}, C_Z) \]
Effective capacitance

- The propagation time across the gate decreases due to the shielding resistance effect.
- The effective capacitance concept is introduced.
Interconnect delay estimation

- The driving cell can be represented by a Thevenin equivalent model.
- The output waveform can be obtained via convolution of the driving point voltage waveform with the transfer function $H_{ZR}$.
- SPICE: generally too much expensive.
Elmore delay

\[ T_D = \int_0^\infty t \ h(t) \ dt \]

It is an upper bound on the delay of an RC tree response: \( T_D \geq \tau \)
Elmore delay calculation

\[ T_{D_i} = \sum_{k=1}^{N} R_{ki} C_k \]

\[ T_{D5} = R_1 C_1 + \]
\[ (R_1 + R_2)C_2 + \]
\[ (R_1 + R_2 + R_3)C_3 + \]
\[ (R_1 + R_2 + R_3 + R_4)C_4 + \]
\[ (R_1 + R_2 + R_3 + R_4 + R_5)C_5 + \]
\[ R_1 C_6 + \]
\[ R_1 C_7 \]

Elmore delay can be calculated very efficiently
- Linear time w.r.t the RC tree size
Model order reduction

Linear, time-invariant circuit

The transient response is generally dominated by a small number of poles

Set of dominant poles is obtained by using moment-matching techniques (AWE, PRIMA, Arnoldi …)

\[ H(s) = \frac{a_0 + a_1 s + a_2 s^2 + \cdots + a_n s^n}{1 + b_1 s + b_2 s^2 + \cdots + b_m s^m} \]
Moment matching technique (I)

- The goal of model order reduction techniques is to obtain a set of $q$ approximate dominant poles efficiently:

$$H^*(s) = \frac{k_1}{s-p_1} + \frac{k_2}{s-p_2} + \cdots + \frac{k_q}{s-p_q}, \quad q << m$$

$$h^*(t) = k_1 e^{p_1 t} + k_2 e^{p_2 t} + \cdots + k_q e^{p_q t}$$

- The $i$-th moment of the function $f(t)$ is defined as:

$$m_i = \frac{(-1)^i}{i!} \int_0^\infty t^i f(t) dt$$

$$m_i^* = \frac{k_1}{p_1^{i+1}} + \frac{k_2}{p_2^{i+2}} + \cdots + \frac{k_q}{p_q^{i+q}}$$
Moment matching technique (II)

The moments of $H^*(S)$ are forced to match the actual circuit moments:

\[
\frac{k_1}{p_1} + \frac{k_2}{p_2} + \cdots + \frac{k_q}{p_q} = m_0 \\
\frac{k_1}{p_1^2} + \frac{k_2}{p_2^2} + \cdots + \frac{k_q}{p_q^2} = m_1 \\
\vdots \\
\frac{k_1}{p_1^{2q-1}} + \frac{k_2}{p_2^{2q-1}} + \cdots + \frac{k_q}{p_q^{2q-1}} = m_{2q-1}
\]
Once the reduced order model is obtained, the output waveform is calculated via convolution with the input signal.

- Example: saturated ramp

\[
\begin{align*}
\text{For time interval } & \quad t_0 \leq t \leq t_1 \\
\vz(t) &= \frac{V_{DD}}{t_1 - t_0} \sum_{k=1}^{q} \frac{r_k}{p_k^2} \left( e^{p_k(t-t_0)} - 1 - p_k(t-t_0) \right)
\end{align*}
\]

\[
\begin{align*}
\text{For time interval } & \quad t \geq t_1 \\
\vz(t) &= \frac{V_{DD}}{t_1 - t_0} \sum_{k=1}^{q} \frac{r_k}{p_k^2} \left( e^{p_k(t-t_0)} - e^{p_k(t-t_1)} - p_k(t_1-t_0) \right)
\end{align*}
\]
Asymptotic Waveform Evaluation

A linear time-invariant network can be represented by the state equation:

\[ \dot{x} = Ax + Bu \]

In the frequency domain:

\[ sX(s) - x(0) = AX(s) + BU(s) \]

The network transfer function is expressed by:

\[ H(s) = \frac{X(s)}{U(s)} = (sI - A)^{-1}B \]
Asymptotic Waveform Evaluation (II)

- The transfer function is the Laplace transform of the impulse response:

\[ H(s) = \int_0^\infty e^{-st} h(t) \, dt \]

- The coefficients of the Maclaurin series expansion are the moments of the impulse response

\[ H(s) = \sum_{k=0}^{\infty} s^k \frac{(-1)^k}{k!} \int_0^\infty t^k h(t) \, dt = \sum_{k=0}^{\infty} s^k m_k \]

- From the previous representation:

\[ H(s) = (sI - A)^{-1} B = -A^{-1} (I + A^{-1} s + A^{-2} s^2 + \cdots) B \]
Moment determination

- The moments are computed recursively

\[ m_0 = A^{-1} B \]

\[ m_1 = A^{-2} B = A^{-1} m_0 \]

\[ m_q = A^{-(q+1)} B = A^{-1} m_{q-1} \]

- State matrix is inverted only once
Moment computation for RC trees

☐ Nodal Analysis instead of State Analysis
  - Based on KCL and KVL
  - It is not a general methodology but it is easier than State Equation

☐ Nodal Analysis formulation:

\[ \dot{v} = C^{-1} G v + C^{-1} B u \]

- \( u \) is a unit step input voltage
- \( C \) is the diagonal capacitance matrix
- \( G \) is the conductance matrix
First moments generation
- Replace input drivers with their final value
- Perform dc analysis
- Moments are the voltages across the capacitors

Successive moments
- Set input drivers to zero
- Replace each capacitor with a current source $I=m_{k-1}C$
- Perform dc analysis
- Moments are the voltages across the current sources
Outline

- CMOS and interconnect scaling overview
- Interconnect resistance effects
- Coupling capacitance impact on performances
- Interconnect inductance: just around the corner
- References
Coupling capacitance impact

- The ratio of the lateral vs. the vertical capacitance increases
- Coupling capacitance becomes a critical portion of the wire capacitance
Crosstalk induced timing effects

- Capacitive interaction between signals may cause:
  - Noise injection
  - Signal timing deviation: speed-up and slow-down
A crosstalk approach: switch factor

- Based on Miller’s theorem
- Switching in opposite directions: $C_1 = C_2 = 2Cc$
- Switching in same directions: $C_1 = C_2 = 0$
- Usually too much pessimistic. In some cases, optimistic!
A complex modeling problem

- Propagation delay on a *victim* depends on the *activity* of the adjacent *aggressors*
  - Signal *temporal positions*
  - Wires and drivers *physical position*
  - Driver *dynamic behaviour*
  - Victim/aggressors:
    - Coupling Capacitance
    - Ground Capacitance
    - Wire resistance
    - Driver resistance (Thevenin equivalent)
    - Signals transition times
Chip level modeling approach

- Circuit simulation (SPICE)
  - Not applicable to VLSI full-chip analysis

- Order reduction (AWE / PRIMA / PVL / ARNOLDI)
  - Much faster (1-2 orders of magnitude) and accurate, but still not extensively applicable to large VLSI ICs

- Worst-case models + heuristics
  - Conservative, but applicable to VLSI full-chip analysis
Complexity of the problem

- Design example:
  - 2.7 M gates
  - more than 1 M nets
  - 11K top level nets
    - Size of the detailed parasitics file 2.1Gbytes
    - More than 2M resistors
    - About 12M capacitors
    - About 20k gates
Two-step approach

- Preliminary screening of potentially crosstalk prone nets
  - Simple and fast model of the crosstalk effects is needed to identify signal integrity problems
- Refine estimation for critical nets (paths)
  - Accurate moment-matching based macromodel is used

This approach has been implemented in a tool (CAT) and successfully integrated in SYNOPSYS PrimeTime (PrimeTime-SI)
PrimeTime-SI implementation

Read parasitics and perform a coupling capacitance filtering process

Delay calculation:
- accounting arrival windows
- accounting signal correlations
- using less conservative model

Reselection phase: critical nets, if any, are reselected for the next iteration

Done!
Fast Xtalk Macromodel (FXM)
FXM – Injected noise waveform

- The transfer function between A and OV is modeled via a two-pole approximation by using the moment-matching technique:

\[
H_{A-OV}^*(s) = \frac{k_1}{s - p_1} + \frac{k_2}{s - p_2}
\]

- The output to a saturated ramp applied to A can be expressed as:

\[
v_{OV}(t) = \frac{V_{DD} A_n}{T_r (p_1 - p_2)} \left[ p_1 (1 - e^{-p_2 t}) - p_2 (1 - e^{-p_1 t}) \right], \quad t \leq T_r
\]

\[
v_{OV}(t) = \frac{V_{DD} A_n}{T_r (p_1 - p_2)} \left[ p_2 (1 - e^{p_1 T_r}) e^{-p_1 t} - p_1 (1 - e^{p_2 T_r}) e^{-p_2 t} \right], \quad t > T_r
\]
FXM - Injected noise peak evaluation

- Noise area
  \[ A_n = (R_V + R_{EV})(C_5 + C_6) + R_{WV}C_6 \]

- Time to peak
  \[ T_M = \frac{1}{p_1 - p_2} \ln \left( \frac{1 - e^{p_1T_r}}{1 - e^{p_2T_r}} \right) \]

- Peak amplitude
  \[ V_P = v_{OV}(T_M) = \frac{V_{DD}A_n}{T_r} \left( e^{p_2T_r} - 1 \right) e^{-p_2T_M} \]
FXM- Victim output estimation

\[ v_{OV}(t) = \frac{V_{DD}}{T_{IN}} \left[ \frac{1}{p} \left( e^{-pt} - 1 \right) + t \right], \quad t \leq T_{IN} \]

\[ v_{OV}(t) = \frac{V_{DD}}{T_{IN}} \left[ \frac{1}{p} \left( 1 - e^{pT_{IN}} \right) e^{-pt} + T_{IN} \right], \quad t > T_{IN} \]
The victim delay $T_D$ and crosstalk extra-delay $T_{XD}$ are computed via Newton-Raphson method.

$$V_{OV}(t_{50}) = \frac{V_{DD}}{2}$$

$$V_{OV}(t_x) = \frac{V_{DD}}{2} - V_P$$
FXM – Cluster of aggressors scenario
Accounting for the wire resistance

\[
V_p = \frac{V_{DD}}{1 + \frac{C_2}{C_C} + \frac{R_1}{R_2} \left( 1 + \frac{C_1}{C_C} \right)}
\]
Re determination

Re can be determined by

- Using a single pole approximation for the transfer function of the unbalanced resistive section of the interconnects
- Using Elmore delay to approximate the dominant time constant of the RC tree

\[ R_E = \frac{T_D(i)}{C_{TOT}} = \frac{\sum_j R_{ji}C_j}{C_{TOT}} \]
Example
Fixing crosstalk problems

- Use extra spacing or shielding
- Increase the strength of the driver
- Re-order the routes
- Buffer insertion
- Staggered inverters
Outline

- CMOS and interconnect scaling overview
- Interconnect resistance effects
- Impact of coupling capacitance
- Interconnect inductance: just around the corner
- References
Interconnect inductance role

- Higher clock frequencies
- Faster rise time \( f_s = \frac{1}{\pi t_r} \)
- Lower resistive materials

\[ \omega L \cong R_{\text{drive}} + R_{\text{wire}} \]
Inductance noise

\[
\begin{align*}
V_{\text{in}} & \quad V_{\text{RC}(Z)} \\
V_{\text{RLC}(Z)} &
\end{align*}
\]
Quantifying inductance effects

- **Dumping factor bound:**
  \[ \xi = \frac{Rl}{2\sqrt{LC}} > 1 \]

- **Transition time bound:**
  \[ t_r > 2l\sqrt{LC} \]

- **Inductance is important if:**
  \[ \frac{t_r}{2\sqrt{LC}} < l < \frac{2}{R\sqrt{LC}} \]

- **Inductance may be neglected if:**
  \[ t_r > 4\frac{L}{R} \]
Is inductance important?

- 0.13\( \mu m \) CMOS technology
- Metal 6 (minimum pitch)

Results:
- 32x: 70\( \mu m < l < 2000\mu m \)
- 16x: 160\( \mu m < l < 2100\mu m \)
Outline

- CMOS and interconnect scaling overview
- Interconnect resistance effects
- Impact of coupling capacitance
- Interconnect inductance
- References
References (I)

References (II)

Our works on this field (I)


- D. Pandini, P. Scandolara, C. Guardiani “Reduced order macromodel of coupled interconnects for timing and functional verification of sub half-micron IC designs”, Asia and Pacific DAC 1997
Our works on this field (II)


- B. Franzini, C. Forzan “Crosstalk Aware Static Timing Analysis Environment”, ESNUG 2001
On going activity: noise analysis

- With technology scaling:
  - coupling capacitances increase
  - operating frequencies increase
  - power supply decreases

- Coupling Noise may cause functional failures
Injected and propagated noise

Injected noise

Propagated noise

Failure!
Static Noise Analysis

Injected Noise
- Macromodel, Reduce Order Model (AWE, Prima, Arnoldi, …)
- Cell output steady state impedance characterization

Propagated Noise
- Need cell noise characterization

Criterion for deciding when noise is dangerous
- DC noise margins, noise rejection curves