SYSTEM (CHIP) LEVEL PHYSICAL DESIGN

FOR LARGE "SYSTEM ON A CHIP" DESIGNS, MUST CONSIDER ISSUES SUCH AS:

- **INTERCONNECT DELAY BETWEEN BLOCKS**
  - Especially high speed data paths
  - Can be as large as gate delays (or even larger)
- **CROSSTALK BETWEEN SIGNALS**
- **FLOORPLANNING OF BLOCKS & ROUTING BETWEEN THEM**
- **INPUT/OUTPUT PADS**
- **POWER DISTRIBUTION**
- **GROUND BOUNCE**

Many of these issues (e.g., interconnect delay) are "built in" to a typical top-down design flow.

*Figure 14.1* Design flow showing chip-level physical design issues

- Floorplanning, cell placement, wire routing, clock trees, etc. all impact timing analysis.
- Several iterations often required to meet timing requirements.
**MODELING OF INTERCONNECT DELAYS**

![Isolated interconnect line](image)

Figure 14.2 Isolated interconnect line

- **Often work with resistance and capacitance of lines on a “Per Unit Length” basis:**

\[
R_{\text{line}} = R_s \left( \frac{W}{L} \right) = (R_s L) \frac{L}{L} = R_{\text{line}}
\]

**WHERE:** \( Y = \frac{R_s}{W} \) (\( \text{m}^{-1} \text{cm} \) (or \( \text{m}^{-1} \text{mm} \))

→ **Note that \( Y \) depends on \( W \), so it is different for each value of \( W \)**

→ **Not provided with process info, just a typical “rule of thumb” used by designers**

\[
C_{\text{line}} = \left( \frac{\varepsilon_{\text{ox}}}{T_{\text{ox}}} \right) W = \left( \frac{\varepsilon_{\text{ox}} W}{T_{\text{ox}}} \right) L = C_{L F} = C_{\text{line}}
\]

**WHERE:** \( C = \left( \frac{\varepsilon_{\text{ox}}}{T_{\text{ox}}} \right) W \) (\( \text{F} \text{cm}^{-1} \) (or \( \text{F} \text{mm}^{-1} \))

→ **Cline depends on \( T_{\text{ox}} \), the thickness of the oxide between the metal and substrate**

→ **Higher metals have larger \( T_{\text{ox}} \), and therefore lower capacitance (e.g., \( C_{\text{M1}} > C_{\text{M2}} > C_{\text{M2}} \ldots \))**

→ **Cline = “Self-capacitance” of line**

(*Note: process info typically contains tables of resistance & capacitance for each metal layer*)
Figure 14.3 Electric field lines for an isolated interconnect

- NOW, THE SIMPLE EXPRESSION OF \( C_{\text{line}} = \left( \frac{\varepsilon_0}{t_{\text{ox}}} \right) \frac{W}{L} \) NEGLCTS THE EFFECTS OF FRINGING = ELECTRIC FIELD LINES FROM SIDES

A MORE ACCURATE EQUATION IS:

\[
C = \varepsilon_0 \left[ 1.15 \left( \frac{W}{t_{\text{ox}}} \right) + 2.8 \left( \frac{L}{t_{\text{ox}}} \right)^{0.223} \right] \text{ F/cm}
\]

\( (\text{PAPER REFERENCE IN TEXT}) \)

Example 14.1
Consider a first-level metal interconnect that has cross-sectional dimensions of \( w = 0.35 \mu m \) and \( l = 0.7 \mu m \) and runs over an oxide layer that has a thickness of \( t_{\text{ox}} = 0.3 \mu m \). The capacitance per unit length is

\[
c = (3.9)(8.854x10^{-14}) \left[ 1.15 \left( \frac{0.35}{0.9} \right) + 2.8 \left( \frac{0.7}{0.9} \right)^{0.223} \right] = 1.07 \text{ pF/cm}
\]

(14.7)

If the sheet resistance is \( R_s = 0.02 \Omega \), then the resistance per unit length is

\[
r = \frac{0.02}{0.35x10^{-4}} = 571 \ \Omega/cm
\]

(14.9)

An interconnect with a length of \( l = 40 \mu m \) is characterized by

\[
R_{\text{line}} = (571)(40x10^{-4}) = 2.29 \ \Omega
\]

\[
C_{\text{line}} = (1.07)(40x10^{-4}) = 4.28 \ \text{fF}
\]

(14.9)

Increasing the length to \( l = 225 \mu m \) gives

\[
R_{\text{line}} = (571)(225x10^{-4}) = 12.85 \ \Omega
\]

\[
C_{\text{line}} = (1.07)(225x10^{-4}) = 24.1 \ \text{fF}
\]

(14.10)

While the resistance remains relatively small (because \( R_s \) is small), the parasitic line capacitance is on the order of MOSFET values, making it important to the analysis.
Once \( R_{LINE} \) & \( C_{LINE} \) are found, we can calculate the time constant for the line:

\[ \tau = R_{LINE} C_{LINE} \]

Which leads to:

\[ V_L(t) = V_{PP} \left(1 - e^{-\frac{t}{\tau}}\right) \] for a 0→1 transition

or,

\[ V_L(t) = V_{PP} e^{-\frac{t}{\tau}} \] for a 1→0 transition

(assumes a step input to the line, an approximation)

---

**Figure 14.4** Single-rung ladder model

**Figure 14.9** Step response for the interconnect circuit
Now, the simple R-C model is actually pessimistic, because it lumps the total capacitance at the end of the line, when the real line has distributed resistance and capacitance.

![Physical model of an interconnect line](image)

*Figure 14.5* Physical model of an interconnect line

Note the voltage on the line is actually a function of both time & position, while this is really a transmission line, we can improve the accuracy of our lumped-element model by adding more sections!

**Using Elmore formula:**

For 2 rungs:

\[ \gamma_2 = C_2 (2R_2) + C_2 R_2 = 3R_2 C_2 \]

For 3 rungs:

\[ \gamma_3 = C_3 \left( 3R_3 \right) + C_2 (2R_2) + C_2 R_2 = 6R_2 C_2 \]

**Where:**

\[
\begin{align*}
R_m &= \text{Rishe-Cleve} \\
M &= \text{Number of rungs in ladder} \\
C_m &= \text{Cleve} \\
M &= \frac{C_m}{M}
\end{align*}
\]

In general:

\[ \gamma_m = \frac{M (M+1) R_m C_m}{2} \]

Or,

\[ \gamma_m = \left( \frac{M+1}{2} \right) \text{Rishe-Cleve} \]

As \( M \to \infty \), \( \gamma \to \frac{1}{\Delta \text{Rishe-Cleve}} \)

*Figure 14.6* Multiple-rung ladder circuits
• Since \( \lim_{M \to \infty} \gamma_M = \frac{1}{2} \text{RLINE} \text{CLINE} \), we can use a simple "\( \gamma \)-model" and get reasonable accuracy.

\[ R_{\text{line}} \quad \frac{1}{2} \text{C}_{\text{line}} \quad \frac{1}{2} \text{C}_{\text{line}} \quad R_{\text{line}} \]

Figure 14.7 Simple RC Interconnect model

Which yields \( \gamma = \frac{\text{RLINE} \left( \frac{\text{CLINE}}{2} \right)}{2} \), and for a load of \( C_{\text{IN}} \) due to another logic gate \( \gamma = \frac{\text{RLINE} \text{CL}_{\text{IN}}}{2} \), with \( \text{CL}_{\text{IN}} = \text{C}_{\text{IN}} + \frac{1}{2} \text{CLINE} \)

Interconnect

(a) Physical structure

(b) RC model

Figure 14.8 Using the Interconnect model to estimate signal delays
Finally, we know that \( \gamma = \frac{1}{2} R \text{line} C \text{line}, \) but how does \( T \) vary with \( L \)?

Use the simple equations for \( R \& C \) (neglect fringe):

\[
\begin{align*}
\Rightarrow & \quad R \text{line} = R_s \left( \frac{1}{T_{0x}} \right) \\
\Rightarrow & \quad C \text{line} = \left( \frac{60 \pi}{T_{0x}} \right) WL \\
\Rightarrow & \quad T \approx \frac{1}{2} R_s \left( \frac{1}{T_{0x}} \right) \left( \frac{60 \pi}{T_{0x}} \right) WL
\end{align*}
\]

\( \Rightarrow T \approx B L^2 \)

where \( B = \frac{1}{2} R_s \left( \frac{60 \pi}{T_{0x}} \right) \)

\(: T \text{ goes up with } L^2 !\)

Figure 14.11 Parabolic dependence of the time delay on line length

---

**Example 14.2**

Suppose that the signal delay on an interconnect of length 50 \( \mu \)m is known to be 0.13 ps. If the line is increased to 100 \( \mu \)m, the delay rises to a value of

\[
\tau = \left( \frac{0.13}{50} \right) 100^2 = 0.52 \text{ ps (14.36)}
\]

where we have used the given data to find \( B \) in the equation. A line that is 200 \( \mu \)m long has a delay of

\[
\tau = \left( \frac{0.13}{50} \right) 200^2 = 2.06 \text{ ps (14.37)}
\]

This shows that the relative lengths of interconnect wires become the important factor.
"Crosstalk" occurs when a signal on one line capacitively couples onto another line.

![Capacitive coupling between two lines](image)

To model this, assume the following structure:

![Geometry used for coupling capacitance calculation](image)

Which can be analyzed (see reference in text), to give the following equation for coupling capacitance:

$$ C_0 = \varepsilon_0 \left[ 0.03 \left( \frac{W}{\text{in}} \right) + 0.83 \left( \frac{H}{\text{in}} \right) - 0.07 \left( \frac{S}{\text{in}} \right)^{0.222} \right] \left( \frac{S}{\text{in}} \right)^{1.34} \, \text{pF/in} $$

$$ \text{C}_{\text{total}} = C_0 \cdot L_c $$

where, $L_c$ = the length over which the lines couple.
Crosstalk can be modeled by lumping the coupling capacitance into a single cap in the middle of the line, or using multiple caps!

(a) Physical structure

(b) RC model

Figure 14.14  Lumpd-element coupling circuit model

Figure 14.15  Alternate model for coupling circuit
Crosstalk can also occur between multiple lines.

Figure 14.16 Multiple-line coupling

Figure 14.17 Circuit model for 3-line coupling problem

or between lines on different metal layers.

Figure 14.19 Overlap capacitance
**Reducing Crosstalk**

- The best way to reduce crosstalk is to space the lines further apart, thereby reducing the coupling capacitance.

  \[ \text{Recall } C_c \propto \left( \frac{\varepsilon}{\varepsilon_{\text{ox}}} \right)^{-1.34} = \left( \frac{\varepsilon_{\text{ox}}}{\varepsilon} \right)^{1.34} \]

  \[ \Rightarrow \text{AS } S \uparrow \text{ Cc} \downarrow \]

  But, this increases the area required on chip for the interconnect lines.

- Another way to reduce crosstalk is to add "shield lines," so that \( C_c \) couples to VDD or VSS instead of another signal line.

![Figure 14.36 Isolation using VSS lines](image)

- Can also use shields between metal layers. E.g., use a M2 shield between signal lines on M1 and M3.
How Wide to Make Lines and Spaces?

1st, Look at RC Delay of Lines:

\[ \gamma \propto \frac{1}{2} \text{Rline Cline} \]

WHERE: \[ \text{Rline} = \frac{R_S}{C_{line}} \quad \text{Cline} = \frac{C_{0x}}{(\epsilon_{0x}/\text{TOX})} \]

\[ \Rightarrow \gamma \approx \frac{1}{2} R_S \left( \frac{C_{0x}}{\text{TOX}} \right) L^2 \]

This would lead us to believe that \( \gamma \) is independent of \( W \) (e.g., doubling \( W \) causes \( R \) to cut in half and \( C \) to double \( \Rightarrow \text{RC} = \text{constant} \))

BUT, this neglects the fringe cap from the sides of the line!

Including \( C_{\text{fringe}} \)

\[ \text{Cline} = \text{Carea WL} + C_{\text{fringe}} L \]

\[ \Rightarrow \gamma = \frac{1}{2} R_S \left( \frac{1}{W} \right) \text{Carea WL} + \frac{1}{2} R_S \left( \frac{1}{W} \right) C_{\text{fringe}} L \]

\[ \gamma = \frac{1}{2} R_S L^2 \left[ \text{Carea} + \frac{C_{\text{fringe}}}{W} \right] \]

\( \star \gamma \downarrow \text{as } W \uparrow \text{ until the fringe cap is } \ll \text{ area gap!} \)

For the space between adjacent lines, increasing this space reduces crosstalk until \( C_c \ll C_{\text{line}} \)

Rule of Thumb:

2µm wide lines \( \Rightarrow \) typically gives 2µm wide spaces \( \Rightarrow \) good results

\( \star \) Be sure to check this for new processes!
**Floorplanning**

- "Floorplanning" refers to the process of deciding where on the chip to place each block need to consider:
  - Which blocks talk to each other?
  - Where are high-speed lines needed?
  - Where are large buses needed?
  - Which blocks talk to input and/or output pads?
  - Size and shape of blocks?

*Each chip is different!*

![System block diagram](a)
![Initial floorplan](b)

**Figure 14.20** Using a block diagram for initial floorplanning

![Global](a)
![Detailed](b)

**Figure 14.24** Routing steps

- Typically leave room for "wiring channels" between major blocks.
- Wiring inside blocks is handled by place & route (P&R) tools.
**INPUT/OUTPUT PADS**

- "**PAPS**" ARE USED TO CONNECT CIRCUITS ON A CHIP TO EXTERNAL (OFF-CHIP) COMPONENTS.

**3 TYPES:**

- **INPUT PAPS**
- **OUTPUT PAPS**
- **I/O PAPS (BI-DIRECTIONAL)**

![Figure 14.26 Input/Output port types](image)

- IN ADDITION TO CONNECTING THE CHIP TO THE EXTERNAL WORLD, PAPS INCLUDE SPECIAL CIRCUITS TO PROTECT AGAINST ESD EVENTS (ELECTROSTATIC DISCHARGE).

→ ESD EVENTS CAN "ZAP" A CHIP WITH 100V OR EVEN 1000V OF VOLTS!

(SIMILAR TO THE SHOCK YOU GET TOUCHING A DOOR KNOB ON A COLD, DRY DAY)

**INPUT PAP!**

![Figure 14.29 Input ESD protection circuit](image)

- HERE THE DIODES TURN ON IF THE VOLTAGE ON THE PAD GETS TOO HIGH AND THE RESISTORS LIMIT CURRENT.

→ $V_{pad}$ NEGATIVE FORWARD BIASES DIODES
→ $V_{pad}$ POSITIVE REVERSE BIASES DIODES, WHICH TURN ON AT THEIR "ZENER VOLTAGE" (Seldom used)
A better input pad!

Figure 14.30 Input resistor-diode structure

- CAN SOMETIMES USE A DIFFUSION RESISTOR, TO CREATE A DISTRIBUTED "R- DIODE" STRUCTURE

Figure 14.31 Alternate input protection circuits

- THIS PAD USES DIODES TO BOTH VDD AND GND, TO LIMIT VPAD TO: \( VDD + V_{JPH} > VPAD > V_{GND} - V_{PH} \)

- ALSO USES AN FET BUILT USING FIELD OXIDE FOR ITS GATE, TO MAKE IT'S \( VT > VDD \)
  
  → FET TURNS ON IF VOLTAGE GETS TOO HIGH, DIVERTING CHARGE AWAY FROM CORE CMOS CIRCUITS
  
  → FIELD-OX FET OFTEN REPLACED WITH A "SNAP BACK" DEVICE

- RESISTORS LIMIT CURRENT INTO CORE, CAUSING IT TO FLOW THROUGH DIODES/FET INSTEAD
Schmitt Trigger

- Often used on input pads to prevent false switching due to noise.
- Positive feedback used to create hysteresis, which causes the trip points to be different for a 0 → 1 and a 1 → 0 transition.

(a) Symbol  
(b) Voltage transfer curve

Figure 14.32 An inverting Schmitt trigger

Figure 14.33 A mirror CMOS Schmitt trigger

Figure 14.34 A non-inverting Schmitt trigger circuit
**OUTPUT PAP**

- **OUTPUT PAP DRIVERS** often just look like big inverters, to drive large off-chip capacitances (10% of PF)

- **SOMETIMES DRIVE FINAL PMOS & NMOS SEPARATELY,** with "break-before-make" architecture to prevent large shoot-through currents

- **LARGE PN JUNCTION DIODES ON DRAINS OF FINAL INVERTER** are typically enough for ESD protection

- **To drive a large capacitance with minimal delay time,** requires a chain of scaled inverters

$$V_i \xrightarrow{1X} \xrightarrow{\alpha X} \xrightarrow{\alpha^2 X} \xrightarrow{\alpha^N X} \frac{V_0}{C_L}$$

$$N+1 \text{ INVERTERS} \quad \text{OPTIMAL } \alpha = e^{2.7} \quad (\text{OFTEN USE } \alpha = 3)$$

* ALSO USE SCALED INVERTER CHAINS TO DRIVE LARGE ON-CHIP CAPS (E.G., ACROSS CHIP)
BI-DIRECTIONAL I/O PADS

- Contain circuits for both input and output
- Output buffer is tri-stated (open circuit) when used as an input pad
- Can be used as a dedicated input pad, output pad, or switched between input & output on the fly as needed

Figure 14.36 A bi-directional I/O circuit
**Power Supply Distribution**

- Separate power supply grids needed for VDD and VSS.
- A grid structure is typically used to achieve high packing density while meeting current needs.
- Two major concerns: 1) Electromigration 2) Ohmic (IR) drops
- VDD/VSS rings often used around edges of chip to connect all pads to a common supply for ESD.

**Figure 14.37** Linewidth sizing for power distribution

**Figure 14.39** Power distribution scheme
GROUND BOUNCE & VDD BOUNCE (AKA "SIMULTANEOUS SWITCHING NOISE")

(a) Interconnect line

(b) Equivalent element

Figure 14.40 Origin of line inductance

\[ V(x) = L \frac{dI}{dt} \]

* INDUCTANCE COMES FROM!

- BOND WIRES
- MOST PACKAGE LEADS
- IMPORTANT SUPPLY GRID

* LARGE CURRENT SPIKES OCCUR WHEN LOGIC GATES SWITCH
  \[ \Rightarrow \text{BIG} \frac{dI}{dt}, \text{BOTH} +/− \]

Figure 14.41 Current flow in an inverter circuit

* MANY LOGIC GATES SWITCHING AT DIFFERENT TIMES
  \[ \Rightarrow \text{LOTS OF NOISE ON} \]
  \[ \text{VDD, GND} \]

* CAN BE 100s OF mV, UNLESS CARE IS TAKEN TO REDUCE INDUCTANCE
  (E.g., MULTIPLE VDD, GND TAPS)

Figure 14.42 Simultaneous switching noise example

Figure 14.43 Switching current in a random logic chain