High Speed CMOS Circuit Design

Lecture 8: Interconnect RC

(c) 1997 David Harris

1.0 Introduction

Once upon a time, the delay of a circuit was well approximated as the sum of the delays of the gates. Wires could be viewed as equipotential nodes that instantly carried the voltage to all ends. As transistors have shrunk, the relative delay of gates has decreased. At the same time, the delay of wires per unit length has actually increased. Thus, wires now are a very important part of total delay.

Wire delay comes from two sources. One is the intrinsic, speed-of-light delay. The second is the lossy nature of on-chip wires; because the resistance of such wires is very high, the wires form RC circuits. Speed of light delay is proportional to the length of the wire, while RC delay increases with the square of the wire length. Thus, for long wires, RC delay dominates. For shorter wires, speed of light delay still dominates. However, such delay over a short wire is still relatively small compared to gate delay.

In addition to increasing circuit delay, wires are subject to coupling noise. According to Maxwell's equations, changes in voltages and currents create electric and magnetic fields which in turn cause changes in voltages and currents on nearby conductors. These effects are called capacitive and inductive crosstalk, and get worse as conductors move closer together and edge rates get sharper. They now can greatly exceed the noise margins of many circuits unless interconnect is carefully designed.

We will first examine the RC wire effects which are most important in current circuits. We'll see how to compute wire resistance and estimate wire capacitance, then use these two properties in wire delay models. We'll also look at capacitive crosstalk and circuit failures caused by IR drops, electromigration, and self-heating. Since delay and crosstalk are so severe, we will explore how to optimize wire geometries, add repeaters, and drive busses to improve performance.

In the next chapter, we'll consider how RC effects scale with decreasing feature size. Then we'll address inductance, an effect which becomes more significant for wire pitches in the one micron range. We'll look at inductive delay effects, di/dt noise, and inductive crosstalk and mutter some banal warnings since little concrete is known about the subject.

2.0 Wire Resistance

The resistance of a wire can easily be computed from its resistivity ρ, length L, width W, and thickness t:

$$R = \rho L/Wt \qquad \text{(EQ 1)}$$

Since the thickness is not under designer control, resistance is frequently rewritten as $R_{\square}$ L/W, where $R_{\square}$ is the sheet resistance having units of ohms / square. Typical values for various interconnect levels are:

TABLE 1. Typical Interconnect Sheet Resistances

Material	$R_{\square}$
diffusion	10
polysilicon	10
thin metal (m1, m2, ...)	0.07
thick metal (top level)	0.03

Without special treatment, diffusion and polysilicon would have much higher resistances that might degrade performance. Therefore, they are usually covered by a thin layer of a refractory metal such as tungsten which alloys with the silicon to form a “silicide” and reduce resistance to approximately 10 ohms/square. Even so, polysilicon should only be used for short interconnect within cells because it is still rather resistive. Diffusion should never be used for routing in high-performance circuits because it has tremendous parasitic capacitance as well as resistance.

Lower metal layers are usually thin, allowing very tight spacing but higher sheet resistance. Metal 1 can be especially thin and resistive. Higher metal layers intended for power distribution and critical signals are thicker and built on a wider pitch, reducing sheet resistance. Most wires are aluminum, but extensive research on other materials with slightly lower resistivity such as copper and gold are beginning to pay off in commercial processes. Such materials can have nasty effects if they contaminate transistors, so exotic schemes are used to prevent their infiltration of the substrate.

Every time a signal crosses metal layers, it must pass through a contact cut, adding further resistance. Such cuts generally have 1-10 ohms of resistance. Low-resistance connections therefore use an array of many cuts.

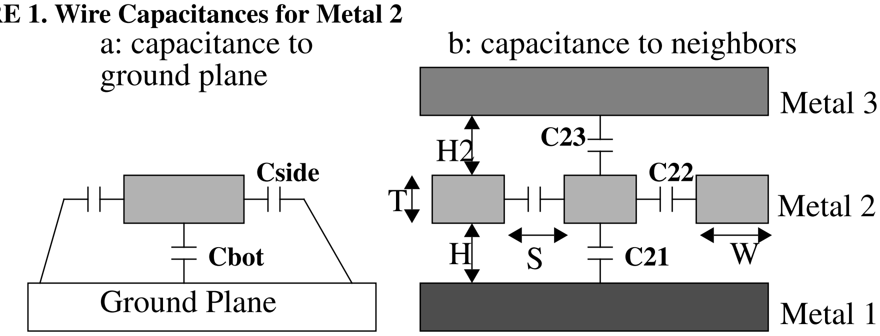
3.0 Wire Capacitance

Wires also have capacitance to the substrate and to other nearby wires. A handy rule of thumb is that most wires have about 0.2 fF/µm of total capacitance, about 1/10 that of a transistor gate. Therefore, when estimating sizes, one can treat 10 λ of interconnect as about 1 λ of gate loading.

In the good old days, wires were much wider than thick. They could therefore be approximated as having parallel plate capacitance to the grounded substrate. Now, wires often have comparable thickness and width and are closer to adjacent wires than to layers above or below. Moreover, the layers above and below are much closer than the substrate. Thus, a negligible part of a wire's capacitance is to ground; instead it is to nearby wires which may transition and couple noise into the hapless wire.

These capacitances are shown in Figure 1:

FIGURE 1. Wire Capacitances for Metal 2



If all the capacitance is from a conductor to a solid ground plane, it can be represented as the sum of a parallel plate capacitance on the bottom and fringing capacitance on the sides. There are many physical and empirical models for such computing such capacitances relatively accurately[1]. Unfortunately, this situation is rare in actual designs; instead, coupling occurs to neighbors on both sides and on the top and bottom. No closed-form equations are possible in general; instead, capacitances must be extracted with a 2 or 3-dimensional field solver using Poisson's equation. Many CAD tools avoid this time-consuming process by pre-extracting tables of capacitances for various geometries and interpolating between the tabulated results closest to the actual geometries of the circuit. Such a table for a 0.8 micron 4-level process can be found at:

`http://www-leland.stanford.edu/class/ee371/lib/wirecap`

Even in such a process, over 75% of the capacitance of a metal3 line over substrate can be to adjacent lines because the adjacent lines are much closer than the substrate. In more advanced processes, the coupling to adjacent lines gets even larger as the lines get closer. In a later section will analyze the amount of crosstalk such coupling can cause.

4.0 Wire Delay Modeling

Wires contribute delay from both the speed-of-light limit and the RC delay. For long wires, the RC term dominates because it increases quadratically with length while speed-of-light delay increases linearly. For short wires, speed-of-light would matter, except that

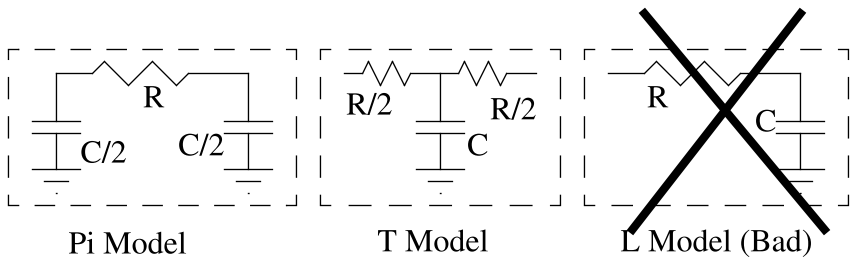
1. See E. Barke, "Line-to-Ground Capacitance Calculation for VLSI: A Comparison," IEEE Transactions on Computer-Aided Design, Vol. 7, No. 2, Feb. 1998, pp. 295-298.

the wires are short enough that the delay is usually assumed to be negligible. This assumption will change in the future as picoseconds of delay cease to be negligible.

4.1 Models

The distributed RC delay can be modeled by breaking up a wire into one or more segments and using a lumped model for each segment. If each segment has resistance R and capacitance C, popular lumped models include:

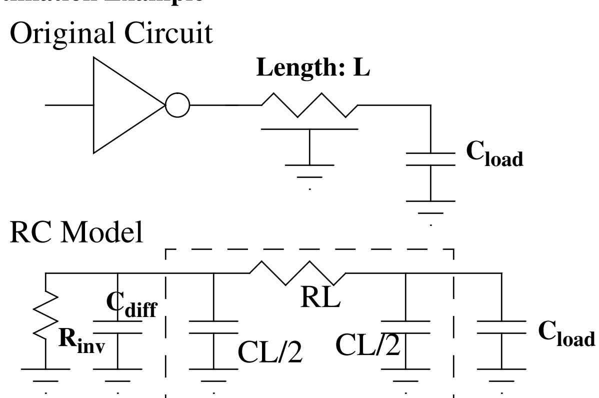
FIGURE 2. Lumped models of wires



Note that R and C depend on the length of the segment. The Pi model and T model, named for their shapes, are both good models. The Pi model can be slightly more convenient because it has fewer circuit nodes. The L model is much less accurate, requiring many more segments for the same accuracy of delay estimation. For example, the delay of a line can be modeled to within 5% accuracy with 4 Pi segments or 64 L segments! Thus, it is usually reasonable to model long wires with 4 Pi segments in simulation and with a single Pi segment for hand calculation.

The delay of the driver must also be included when estimating the delay of a wire. For example, we can compute the delay of an inverter driving an L mm wire with resistance R/mm and capacitance C/mm, inverter drive resistance R_{inv}, drain capacitance C_{diff}, and total load on the wire C_{load}:

FIGURE 3. Delay Estimation Example



The Elmore delay of the model is:

$$R_{inv}*(C_{diff} + CL/2) + (R_{inv} + RL)*(CL/2 + C_{load}) \quad \text{(EQ 2)}$$

This can be rewritten as:

$$R_{inv}*(C_{diff} + C_{load} + CL) + RCL^2/2 + RLC_{load} \quad \text{(EQ 3)}$$

This expression has three terms. The first represents the inverter driving it's own diffusion, the load capacitance, and the lumped capacitance of the wire. The second is the quadratic delay of the wire's self-loading. It is proportional to RC/2, rather than RC, because the capacitance is distributed along the wire rather than lumped at the end. The third term is the extra delay contributed by the wire resistance discharging the load capacitance.

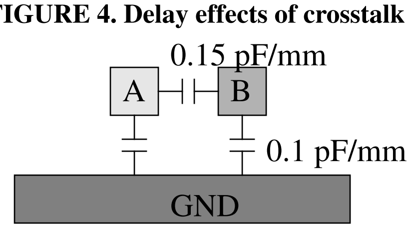
Remember that the units in this equation are somewhat curious. R_{inv} is Ω. R is Ω*µm. C_{diff} and C_{load} are fF, while C is fF/µm.

5.0 Capacitive Crosstalk

The RC models presented above assume that all capacitance is to ground, or some other node which is not transitioning. When the capacitance couples to another wire which might be switching, the crosstalk between wires causes noise and change in delay.

When two adjacent wires switch opposite directions, their delay is longer than if one switched while the other remained steady. This is because the capacitance between the wires sees a total of twice as much voltage change and thus requires twice as much current to be charged up. We can model this case for delay with a capacitor to ground twice as big as the capacitance between lines. For example, if a wire had 0.1 pF/mm capacitance to ground and 0.15 pF/mm capacitance/mm to an adjacent wire which might switch in the opposite direction, it can be modeled as a total of 0.4 pF/mm of capacitance to ground for delay purposes. Conversely, if the adjacent wire is switching in the same direction, the coupling capacitor will see zero voltage change and have no effect on delay. In the previous example, it could be modeled as 0.1 pF/mm of total capacitance to ground. At first, this seems good because the delay will be less. However, the delay of the wire becomes much less predictable because its effective capacitance changes by a factor of 4 depending on what the neighboring line does. If delays of two signals must be matched for some reason, the variation in delay caused by such coupling is very important.

FIGURE 4. Delay effects of crosstalk



A	B	Effective capacitance of A
0->1	constant	0.25 fF/mm
0->1	1->0	0.4 fF/mm
0->1	0->1	0.1 fF/mm

Worse yet, the capacitive crosstalk introduces noise that could corrupt a signal. We can conservatively model crosstalk noise with a capacitive voltage divider. Suppose victim signal A is supposed to be 0. Adjacent aggressor signal B rises. The coupling between B and A tends to cause A to rise as well. The amount A rises depends on the ratio of coupling capacitance C_{AB} to total capacitance $C_{A\text{-}GND} + C_{AB}$. It is also influenced by the driver of A, which will fight the crosstalk, trying to maintain A at a 0. If the crosstalk happens fast enough that the capacitive impedances are much lower than the driver resistance, the driver will not help much and the voltage noise on A is:

$$C_{AB} / (C_{AB} + C_{A\text{-}GND}) * V_{DD} \quad \text{(EQ 4)}$$

When $C_{AB} = C_{A\text{-}GND}$, the noise will be 50% and could cause a false transition on a gate at the end of the wire. Most static circuits can tolerate such noise glitches. Dynamic circuits, however, will be corrupted irrecoverably. Therefore, dynamic nodes are especially sensitive to coupling. Moreover, in a bus, a wire can receive coupling from both sides. In fact, it is even possible to receive coupling from orthogonal wires on layers above and below. Usually, this coupling is negligible. However, if a wide bus above or below switches entirely in one direction, the coupling can be significant. Most of the time, this coupling is ignored by verification tools. One day, chips may start to fail in a data-dependent way from such coupling on other layers!

When a node is actively driven, the small driver resistance does reduce the amount of coupling noise. Simulation may be the only way to determine exactly how much the resistance helps.

6.0 Other Wire Effects

In addition to delay and coupling noise, wires are susceptible to IR drops, electromigration, and self-heating.

6.1 IR Drops

Wires such as power busses which carry large DC currents droop in voltage due to the resistance of the wire multiplied by the current carried on the wire. This is bad because, for instance, it reduces the value of V_{DD} seen at the center of a chip when the power supply is driven by pads around the perimeter.

IR drops get to be a bigger problem as lower power supplies are used. For a fixed wattage, a lower V_{DD} means a higher I_{DD} flows through the power distribution network. For the same power network resistance, a higher IR drop is caused. Moreover, this drop is even a larger fraction of V_{DD}!

For example, a 100 watt chip operating at 2V requires 50 amps of supply current! If IR drop should be no more than 5% of VDD, this requires a power network with only 2 mΩ of resistance. To meet the same spec, a 1V chip requires 100 amps of current and can only tolerate 0.5 mΩ of resistance in the power distribution network! Remembering that a typical top-level metal layer has a resistance of about 30 mΩ / square, the power network can only have a small fraction of a square of size. Even a solid power plane may be insufficient; instead power may be distributed in a thick plane in the package and dropped all over the chip with C4 bumps to reduce the resistance.

6.2 Electromigration

When high currents flow through a wire, they create an "electron wind" which can apply force on metal atoms in the wire. The atoms may actually move slightly, in an effect known as electromigration. If electromigration continues for a long time, it may actually cause a break in a wire, resulting in an open circuit and chip failure. This is especially unpleasant because it may happen after years of operation.

Such failure is avoided by limiting the current flowing through a wire. DC current is worst because it pushes atoms in a single direction. AC currents are less of a problem for electromigration because they push the atoms back and forth without a net average movement. Thus, power and ground lines are most prone to electromigration failure. A rule of thumb for many processes is to allow no more than 1 mA of current per micron of wire width. Wider wires may tolerate somewhat higher current densities than this. Processes with very thin wires tolerate less current. Check with your vendor or be conservative with wire widths.

6.3 Self-heating

IR drops in wires cause the wires to heat up. Oxide is a good insulator, so the heat may build up in the wires, resulting in the wires being hotter than the chip. This can increase wire resistance and degrade performance. In extreme cases, it is possible to experience thermal runaway, in which wires get hotter, thus become more resistive, causing larger IR drops and more heat until the wire melts down!

Such self-heating sets minimum wire widths even for lines carrying only AC signals.

7.0 Optimizing Wire Geometry

We have seen that wire resistance and capacitance cause significant amounts of delay and noise. Thus, optimizing wires can be as important as optimizing gates.

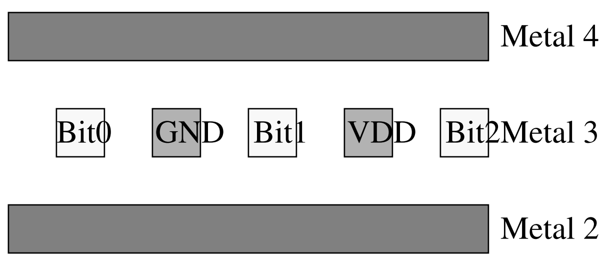
The best way to optimize wires is to keep them as short as possible. This requires careful floorplanning and good exploitation of datapath structure. Nevertheless, it is inevitable that a chip will have some long wires. What can we do to minimize delay and noise of these wires?

Wire design is fairly simple because there are only a few degrees of freedom. The designer can control the width of each wire, the spacing between wires, and the arrangement of which wires are adjacent to which other wires.

To reduce delay, the designer may adjust each of these factors. If the delay is dominated by the wire RC product, increasing wire width is often useful. This decreases wire resistance proportionally, while only slightly increasing total wire capacitance if the capacitance from the top and bottom of the wire is a small part of the whole. Increasing spacing between wires is also good because it decreases the total wire capacitance and the amount of coupling to adjacent wires. This is especially important when the adjacent wires might be switching and Miller-multiplying the coupling capacitance. Finally, interleaving busses such that the adjacent wires are not switching at the same time prevents Miller-multiplication of coupling capacitance. For example, a bus driven in phase 1 could be interleaved with a bus driven in phase 2 so no two adjacent wires switch at once.

The designer also may use these degrees of freedom to reduce crosstalk. The first tactic is to increase the spacing between lines, thus reducing coupling. At some point, this yields diminishing returns. The next option is to increase the width of the wire or to shield the wire with power/ground lines. Increasing the width of a wire is sometimes helpful because it increases the fraction of capacitance to the layers above and below rather than to an adjacent coupling wire. It also reduces the RC delay. In contrast, shielding the wire with a nearby minimum pitch power or ground line is extremely effective because it eliminates almost all capacitive coupling. It also increases the number of wires in the power and ground grids which may improve the power supply network. The drawback is that it may increase the total amount of capacitance on a wire and thus increase RC delay. Shielding every bit of a wide dynamic but also has horrible effects on area.

FIGURE 5. Shielding adjacent lines of a dynamic bus



As mentioned earlier, the capacitance to adjacent wires and to the upper and lower layers is a complex function of wire width and spacing. It may be necessary to use a table of capacitances to understand how width and spacing influence noise and delay. The designer faces a constant tradeoff between delay and area; there is no single pitch optimum for all problems.

8.0 Repeaters

Even when a wire is well optimized to minimize its resistance and capacitance per unit length, wire delay still increases quadratically with distance. At some point, it becomes better to break a long wire in half and insert a buffer, called a repeater, to drive the second half. Each half of the wire contributes 1/4 of the RC delay, thus reducing total RC delay by a factor of 2. The buffer, however, introduces extra delay. The best wire length for introducing a repeater therefore is when the repeater delay is less than the amount of wire delay saved. For very long lines, breaking the line into more than 2 pieces and adding multiple repeaters becomes beneficial.

By splitting a wire every time it gets too long and adding a repeater, delay can be made to be proportional to the length, instead of the square of length. This results in a "effective speed of light" in wires driven with optimal repeater placement. The effective speed scales with transistor performance because faster transistors mean faster repeaters and shorter optimal wire lengths between repeaters. In Appendix A, we prove that the effective speed of light is:

$$\sqrt{C_W R_W C_T R_T}\,(2 + \sqrt{2}) \quad \text{(EQ 5)}$$

where C_W, C_T, R_W, and R_T are the capacitances and resistances per unit length of wires and transistors.

For a 0.4 micron technology, this delay is around 50 ps/mm using optimal repeater placement of inverters with 23 micron NMOS devices every 2.5 mm of wire. The repeater placement gets more frequent as transistors get faster and wires get slower. This compares with about 6 ps/mm for the true speed of light on a chip. Increasing wire pitch or building faster transistors improves the RC delay.

When the final load is comparable to the capacitance of a repeater, it is interesting to see that each repeater is sized identically. The tapered buffer chain concept of inverters increasing exponentially in size along the chain is not appropriate because the goal is not to power up to drive a huge load, but rather to break the wire into lengths that can be driven quickly. In fact, raising the size of the repeaters will actually slow the circuit down as well as hurt area and power. Knowing the appropriate repeater sizing and placement for your process is important.

There are several tricky issues with actually using repeaters. Unless an even number of inversions are always used, the logic at the receiving end may get the wrong polarity of signal. Repeaters must also be placed in the layout; this may require extensive communication between designers so that the repeaters for cross-chip interconnect are placed in suitable locations. When long interconnect forks to multiple destinations, such as in critical stall signals, repeater placement to minimize delay while providing the right polarity becomes even trickier.

9.0 Advanced Signaling Techniques

There are many tricks to improve interconnect performance further, but each introduces risks. Precharging busses can be very fast and integrates well with dynamic circuits, but is susceptible to coupling noise. Small swing busses can be even faster and more noise-prone. Bi-directional busses may be beneficial for area, but make using repeaters difficult.

9.1 Precharged Busses

Precharged busses are fast for the same reasons precharged gates are fast. They can be driven by fast gates, improving repeater performance. The repeaters can also have switching thresholds adjusted to favor the critical edge, responding after a smaller signal swing.

Unfortunately, precharge busses are not actively driven while in the precharge state. Therefore, crosstalk noise is especially severe. Even if they have a small keeper, the keeper is too weak to cancel much of the crosstalk.

Hence, many designers limit use of precharged busses to within a unit so that the designer has complete control over the coupling sources. Good routing and verification tools are necessary to support global precharged busses.

9.2 Small Signal Swings

The delay of a heavily capacitive system is set by the time charging the capacitance:

$$\Delta t = C/I \, \Delta v \qquad \text{(EQ 6)}$$

Hence, reducing the Δv required to indicate a logic level change also reduces the time required to propagate the information. This trick is especially popular in SRAM arrays and on I/O pads. In RAM arrays, a differential sense amplifier detects a small change in voltage between the bitline and its complement. If the gain of the amplifier is sufficiently high, only enough voltage difference must develop to overcome offset voltages in the amplifier. Similarly, many low-voltage I/O techniques use swings of one volt or less.

Coupling is especially serious when small swing signals are located near full swing signals, because even a small amount of coupling from a full swing line onto a small signal line could upset the value on the victim line. I/O pads avoid this problem if all external I/O signals are small swing. SRAM bitlines are often twisted in such a way as to guarantee the same coupling onto the true and complementary bitlines so that the net differential noise is zero.

9.3 Bidirectional Busses

Sometimes area can be reduced by sharing a single bus among many seldom-used drivers. Each driver uses a tri-state circuit to only drive the bus when it is granted control.

Such a design has many drawbacks. Each driver must be large enough to drive the heavily loaded bus. The drivers which are not on therefore present large parasitic drain capacitances, slowing the active driver. Moreover, building repeaters on the bus is difficult because the driver might be located at either end of the wire. Therefore, the repeaters cannot be inverters which only drive one direction. Instead, they must be bidirectional circuits which sense a transition starting and pass it on. Such circuits are slower and more prone to noise than regular repeaters.

In general, bidirectional busses should be avoided in high performance systems. Point to point links should be used instead. When many seldom-used drivers must be connected to many receivers, the drivers can each drive point-to-point links to a central location where a multiplexor selects the appropriate source and drives it to all the receivers.

10.0 Wire Scaling

As feature size continues to scale, transistors get progressively faster. Interconnect, however, slows down because smaller wires packed more densely have higher resistance and capacitance. Thus, wire delay accounts for a significant and growing portion of overall path delay on many modern circuits. Understanding how such delays will scale with advances in process technology is important when considering chip designs.

In a process shrink, the x and y dimensions of a chip (transistor and wire lengths and widths) are scaled by a factor of α. A typical value of α is 0.7. Gate oxides are also shrunk by this factor. Conductor and dielectric thicknesses may be shrunk or may be kept constant. Table 1 lists how the resistance, capacitance, and delay of wires and transistors scales with such process shrinks.

TABLE 2. Process Scaling

Case	R	C	RC
Interconnect (shrink thicknesses)			
per μm	α^{-2}	1	α^{-2}
total	α^{-1}	α	1
Interconnect (constant thicknesses)			
per μm	α^{-1}	$1<C<\alpha^{-1}$	$\alpha^{-1}<RC<\alpha^{-2}$
total	1	$\alpha<C<1$	$\alpha<RC<1$
Transistors			
per μm	α	1	α
total	1	α	α

The entries per micron indicate how electrical characteristics change per unit length. For example, if a new chip is designed in a more advanced process, the delay of a wire traveling 10 mm will be much longer than the delay of a 10 mm wire in the older process. The "total" entries indicate how characteristics change when an old design is simply shrunk. Wire lengths will get shorter, compensating for the fact that wires get slower per unit length. However, wires will not improve as much as transistors, making wire-dominated paths more critical.

To derive the interconnect entries in the table, remember that wire resistances scale as L/Wt. W is always scaled by α. t scales by α in the shrink thickness case, or remains constant in the second case. L is normalized out in the per micron case, but scales by α in the total case. In contrast, C/micron scales as the ratio of the ratio of two distances. When all three dimensions are scaled uniformly, C/micron therefore remains constant because all ratios of distances remain constant. When thickness is held constant, capacitance might increase slightly. The capacitance from the bottom of a wire to the lower layer decreases slightly because the wire gets narrower but the distance to the next layer stays the same. The capacitance from the side wall of the wire to an adjacent wire will increase, however, because the wall remains as tall but the spacing gets smaller. Since spacings between wires are now smaller than separation between layers, the second effect usually dominates. In summary, capacitance per micron increases somewhat, but not as much as $1/\alpha$. Multiplying the resistance and capacitance, we see that delay per micron increases substantially. Several factors favor constant thickness scaling. The table indicates that wire delay is better, perhaps by as much as a factor of α if sidewall capacitance is not dominant. Wire resistance and current handling capability also is good, important for IR drops and electromigration. Unfortunately, constant thickness scaling results in increased coupling capacitance as a fraction of the whole capacitance. Constant thickness scaling has been most popular because of its advantages, but means that coupling capacitance is now a major challenge.

Transistor scaling can be done with constant voltage or constant fields. For years, constant voltage scaling was used, maintaining the 5 volt standard. As long as transistors did not become velocity saturated, their delay scaled as α^{-2}, resulting in great performance improvements. Transistors are now velocity saturated, so holding voltage constant has little benefit. Thus, constant field scaling is now popular. Voltage and feature size are reduced by about the same amount. Electric fields depend on the ratio of voltage to distance and remain roughly constant, giving the name constant field scaling. Resistance per micron depends on VDD/(L*tox) and thus scales by α. Capacitance per micron depends on ratios of distances. Since all distances scale equally, it remains nearly constant between generations, in the range of 1-2 fF/µm. When all widths are also scaled, resistance and capacitance scale inversely.

11.0 Introduction to Inductance

Inductance is a phenomenon which is poorly understood but which is beginning to worry many designers of high speed chips. It is well-known that I/O pads suffer problems from inductive Ldi/dt noise and engineers know how to use bypass capacitance and improved packaging to handle such problems, as we will see. On-chip inductance, in contrast, is poorly understood. Many designers see that the resistance of on-chip interconnect is so high that wire RC seems to swamp out the speed of light delay set by on-chip inductance. Simply raising the frequency of a particular chip won't impact the RC or LRC delays; so it is not obvious why inductive problems will become more serious in the future than it is now. Moreover, there are existence proofs from several high speed microprocessors on the market that 0.35 micron chips can be built without much attention to inductance and still work reliably. Why must we worry about it in the future? It turns out that for very high speed chips, on-chip inductance becomes an issue in four areas: increased wire resistance due to the skin effect, increased signal propagation delay, ground bounce, and inductive crosstalk. Before looking at each of these issues, let us review some electromagnetism theory necessary to understand inductance.

What is the physical origin of inductance? Ampere's law says that a current J in a wire produces a magnetic field H:

$$\oint_C \vec{H} \bullet d\vec{s} = \int_S \vec{J} \bullet d\vec{a} + \frac{d}{dt}\int_S \varepsilon_0 \vec{E} \bullet d\vec{a} \qquad \text{(EQ 7)}$$

Faraday's law says that a change in the magnetic field passing through a loop causes a voltage drop around the loop:

$$\oint_C \vec{E} \bullet d\vec{s} = -\frac{d}{dt}\int_S \mu_0 \vec{H} \bullet d\vec{A} \qquad \text{(EQ 8)}$$

Inductance relates these two effects such that the voltage drop around a loop is proportional to the rate of change of current causing the magnetic field: v = L di/dt. Therefore, the inductance of a circuit depends on the geometry of the current loops. Depending on loop shape, inductance may increase linearly or logarithmically with loop area. In any event, it is important to minimize the area of the loop to reduce inductance. This means that a return path for current must be located close to the signal of interest. Typical values of on-chip inductance with a reasonable power grid are on the order of 1 nH/mm. A nearby reference plane reduces this to around 0.5 nH/mm.

Current follows the path of least impedance. Since impedance Z = R + jωL, resistance dominates impedance at low frequencies and inductance dominates at high frequencies. Therefore, low frequency current will flow in the lowest resistance path. This path could be the substrate if doping levels are high, or the power grid and package power plane if the substrate is resistive. At high frequency, current will flow in the lowest inductance path. This is the path forming the smallest loop. Since the paths carrying the current are a function of frequency, the resistance and inductance seen by the circuit are also functions of frequency.

12.0 Propagation Delay

Inductance and capacitance determine the speed of light delay through a wire: t = $\sqrt{LC}$. The speed of light delay and RC delay are not additive; when RC is significantly larger, the speed of light delay barely impacts overall wire flight time. Let's look at these two delays and compare their magnitudes. In free space, light takes 3.3 ps to travel a millimeter. SiO_2 has a dielectric constant $\varepsilon/\varepsilon_0$ of about four; the speed of light is slowed by the square root of this ratio to 6.6 ps/mm. The resistance of a typical 1 μm wide M3 signal line is about 50 Ω/mm and the capacitance is about 0.2 pF/mm. Therefore, the RC product is 10 ps/mm^2. Since the RC line is distributed, delay is about half the RC product, but even so, the RC delay is comparable to speed of light delay for 1 mm lines and completely overwhelms speed of light delay for longer lines. Therefore, it could be argued, speed of light can be ignored unless errors of less than 10 ps are important enough to justify the effort.

Unfortunately, this argument has a fallacy, namely that the return path under the signal wire is an ideal ground plane. Charges redistribute rapidly on the surface of the substrate so the substrate looks like a good ground plane as far as capacitance calculations are concerned. Unfortunately, the substrate has sufficient resistance that return currents travel hundreds of microns down in the substrate and the substrate is not a good ground plane for inductive modeling. In some packages, the current actually returns in low-resistance metal planes in the package instead. Therefore, the size of the current loop is much larger than if current returned in a plane on the surface of the substrate. The inductance depends on the area of the current loop and therefore is much larger due to this substrate return path. Speed of light delay depends on the $\sqrt{LC}$ product and therefore gets worse. This effect is known as a "slow mode" of wave propagation and has been seen often by RF engineers.

The existence of slow waves on chips depends on the materials and geometries being used. If the substrate is lightly doped, its impedance is too high to carry much return current. Instead, the current returns in the power grid or package. If the substrate is more heavily doped, it may carry substantial current deep beneath the surface and cause more severe slow mode waves.

If slow mode waves do exist, the speed-of-light delay will increase. For lines with short RC delay, speed of light may become important.

Slow-mode waves are eliminated with a good power and ground plane located close to the signal lines.

13.0 Inductive Crosstalk

The most sinister threat of inductance is inductive crosstalk caused by magnetic coupling between lines. Current in each line creates a magnetic field circulating around the line. Changes in this current produce changes in the magnetic field, which cause a voltage drop in adjacent lines. This effect is known as mutual inductance. The mutual inductance between wires A and B is a function of the ratio of the distance between A and B to the distance between each wire and the return current path, usually the power grid. Larger ratios give higher mutual inductances. Thus, if the power grid is sparse or far away, a wire will be coupled not only to the two adjacent wires, but also to many other nearby wires in a wide bus. This is different than capacitive coupling, which tends to terminate on adjacent lines.

Another way of understanding inductive crosstalk is to think of a 64 bit bus and its return path in a power plane as a 64 turn solenoid. Suppose all the bits except one in the middle switch from low to high. The current flowing through each bit of the bus and back through the return plane creates a magnetic field. The fields add, just as they do in a tightly wound solenoid. Therefore, they can cause huge voltage noise on the signal that is not switching, just as a N:1 turn transformer can step up voltage by a factor of N.

Inductive crosstalk involves coupling from all lines that are as close to the victim as to their return current paths. Thus, the best way to avoid such crosstalk is to provide nearby return current paths, either with a power and ground plane or by interleaving power and ground lines with every few bits of the data bus.

14.0 Ground Bounce

Ground bounce occurs when the current through the GND busses rapidly changes. For example, if the floating point unit is switched from a low power sleep mode to a fully active mode in one cycle, the GND current will greatly increase. This produces a voltage $V = Ldi/dt$, making the GND line rise in potential relative to other GND busses on the chip.

15.0 Skin Effect

From Maxwell's equations, it can be shown that magnetic fields take finite time to diffuse into a conductor. Thus, very high frequency currents will produce fields that do not fully penetrate a thick conductor. The currents are carried along the outer edges of the current, a phenomenon known as the skin effect. This may increase the resistance of the conductor.

Currents penetrate to a skin depth δ which is a function of frequency, conductivity, and magnetic permeability:

$$\delta = \sqrt{\frac{2}{\omega\mu\sigma}} \tag{EQ 9}$$

If this depth is comparable to half the thickness of a wire, it may increase the effective wire resistance since current is only carried in a fraction of the wire. The frequency of interest is not the clock frequency, but rather the edge rate, which is substantially higher. For instance, signals with a 75 ps rise time correspond to frequencies of 4 GHz. For metal lines with a resistivity of around 4 $\mu\Omega$*cm, the skin depth is still 1.5 microns, thicker than a typical conductor. Therefore, skin effect is not yet a major problem for on-chip interconnect.

16.0 Solutions to On-chip Inductance

On chip-inductance is an especially difficult problem to manage because it is so poorly understood. The general approaches are to model it or to minimize it.

Minimizing inductance is easiest: dedicate at least two metal layers on the chip to power and ground planes, just as a PC board uses dedicated power and ground planes. In such a case, on-chip inductive effects will usually be negligible. DEC chose this approach for the Alpha 21264, dedicating the third and sixth level of metal to GND and VDD, respectively. DEC probably chose this expensive approach because it needed a very low resistance path for the power supply to limit IR drops. Lacking C4 packaging technology, dedicated planes may have been the only solution. The planes have a side benefit of reducing inductance and capacitive coupling.

Modeling inductance might result in chips that are less expensive to manufacture, but is a more difficult task. Merely extracting inductance is difficult, requiring 2D or sometimes 3D field solvers such as FastHENRY and a good understanding of return current paths. Moreover, understanding what nets must be simulated with extracted inductance is difficult. Finally, doing the simulation is difficult; most existing tools either to not handle inductance or converge poorly when it is introduced. For example, modeling a 64 bit bus with substantial mutual inductance between bits in HSPICE usually results in HSPICE not converging.

The circuits most impacted by inductance are the power network, low-resistance wide lines such as the clock, and wide busses. The power network should be checked for di/dt noise and designed to provide low-inductance return paths for other nets. Wide lines have low resistance and are thus most impacted by LC delay relative to RC delay. Wide busses switching simultaneously are subject to severe mutual inductive coupling unless the return path is very close, such as a solid ground plane or power/ground lines interleaved among the signals.

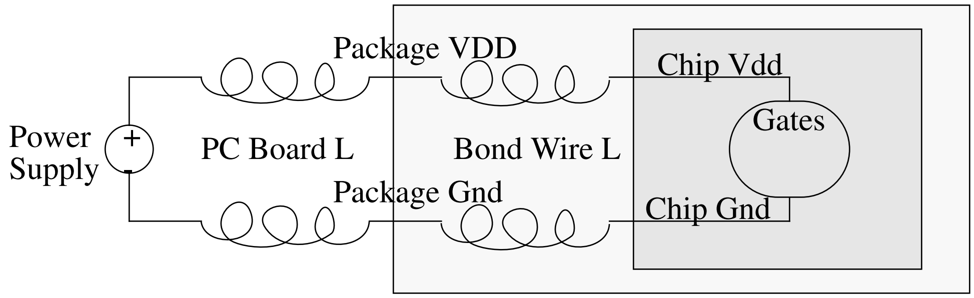
On-chip inductance is still in the research stage. Slow-mode waves and the skin effect will at worst cause increased wire delays. Ground bounce and inductive crosstalk can introduce enough noise to cause circuit failure in fast systems with poor return current paths. Designers who do not plan to minimize or model these effects may discover unexpected failures in their silicon.

17.0 I/O Pad Inductance

On-chip inductance is an emerging problem, but I/O pad inductance has been a problem for years. When a chip quickly changes current demands (as would happen when it wakes up from sleep mode), a large di/dt spike is produced. If this current is supplied from the off-chip power supply, the current must pass through traces on the PC board and through the bond wires onto the chip. Bond wires have an inductance of roughly 1 nH/mm, creating di/dt noise.

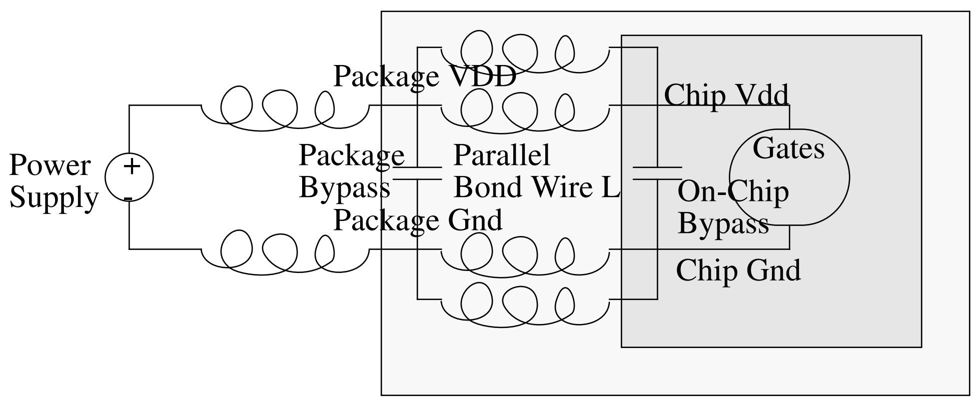
For example, consider the circuit in Figure 1. If all of the gates begin switching and raise the current from 0 to 10A in one nanosecond, di/dt = 10^{10} A/s. If there is a total of 4 nH of inductance from bond wires and the PC board, the power supply would be reduced by L di/dt = 40 volts. This is obviously impossible for a 2.5 volt chip; instead, as current begins ramping up, VDD and GND would start collapse toward each other until the chip sees only a very small potential difference between them. At such a point, transistors would barely turn on and less current would actually get drawn. In any event, the circuit would produce incorrect results.

FIGURE 6. External Inductance Sources



To reduce the inductance, designers may use hundreds of pins dedicated to power and ground to provide parallel current paths. Better packaging, such as C4 packages with low-inductance solder balls instead of bond wires, also helps. Even so, supplying enough current to a switching I/O pad from off chip is very difficult. The next step is to build bypass capacitors which supply the instantaneous current demand, then are recharged more slowly from the external power supply. Bypassing is used externally, in the package, and on chip. These techniques are shown in Figure 2:

FIGURE 7. Parallel bond wires and bypass capacitors



Fortunately, all the gate capacitance on a chip which is not switching serves as bypass capacitance for the switching gates. For example, SRAM arrays do an excellent job of decoupling the power supply in certain regions of a chip because they have a high capacitance per area. As di/dt demands increase, even the implicit bypass capacitance becomes insufficient and designers must add explicit capacitors under routing channels and in other unused portions of the chip. The capacitors must be close to the current demand to be effective. In some cases, chip area may actually be increased to provide sufficient bypass capacitance.

18.0 References

[1] H. Johnson, and M. Graham, *High Speed Digital Design*, Englewood Cliffs, NJ: Prentice Hall, 1993.

A good text on PC-board design issues. Some guidelines on how PC-board makers deal with inductance.

[2] D. Priore, "Inductance on Silicon for Sub-Micron CMOS VLSI," in *Proc. VLSI Symposium*, p. 17-18, June 1993.

A short and crisp paper illustrating the presence of inductive crosstalk for on-chip interconnect.

[3] A. Deutsch, et. al, "Modeling and characterization of long on-chip interconnections for high-performance microprocessors," IBM J. Res. Develop, Vol. 39, No. 5, pp. 547-567, Sept. 1995.

The most thorough study yet published of on-chip inductance measured by test structures.

[4] M. Shoji, *High-Speed Digital Circuits*, Reading, MA: Addison-Wesley, 1996.

A very unusual book attempting to build a theory of circuits involving inductance. Dense, difficult reading, but the first chapter contains a good introduction to inductance.

Appendix A Derivation of Repeater Sizing

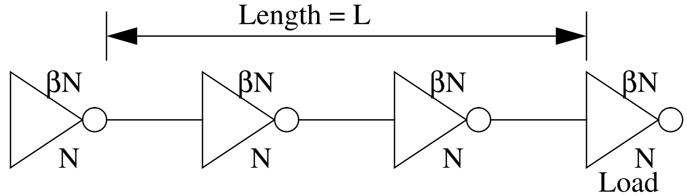
Delay of a single long wire in an integrated circuit is proportional to the square of the length because both the resistance and capacitance increase linearly with length. By breaking the wire into multiple stages and placing repeaters at the start of each stage, the delay can be made linear with length. This appendix explores the optimal sizing of buffers and number of repeaters to minimize the delay. It computes delay (ps/µm) as a function of process parameters at the optimal sizing.

We begin with several simplifying assumptions to make the analysis cleaner.

- Source / drain diffusion capacitance is negligible
- Inverters are sized for equal rise and fall times (faster results may be possible with unequal times)
- Wire pitch and spacing are preset
- Propagation speeds are low enough that transmission line effects may be neglected
- Load capacitance equals repeater capacitance

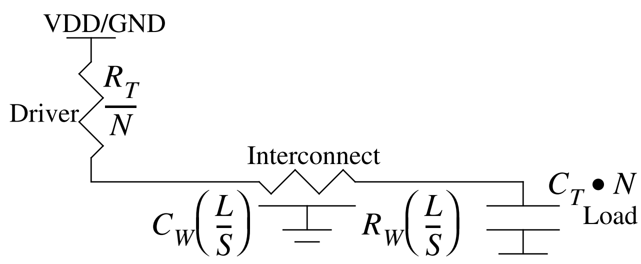
The interconnect of total length L is divided into S segments. Each inverter has a pull-down N microns wide and a pullup βN microns wide to achieve equal rise and fall times. The figure below illustrates a wire divided into three segments:

FIGURE 8. Interconnect divided into 3 segments



The path can be broken into S identical segments for analysis. A model of one segment is shown below. Resistances are in ohms / micron; capacitances are in pF / micron. Notice that C_T is the combined capacitance of the PMOS and NMOS transistors in an inverter sized at N=1.

FIGURE 9. Model of single segment



Observing that a distributed RC line has the same delay as a lumped RC element with half the resistance, we can derive the following expression for the delay through a segment. The delay through the entire interconnect is S times the delay through each segment:

$$t_{interconnect} = S\left[C_T N\left(\frac{R_T}{N} + R_W\left(\frac{L}{S}\right)\right) + C_W\left(\frac{L}{S}\right)\left(\frac{R_T}{N} + \frac{R_W}{2}\left(\frac{L}{S}\right)\right)\right] \qquad \text{(EQ 10)}$$

$$\text{simplifying, } t_{interconnect} = SC_T R_T + L\left(C_T R_W N + \frac{C_W R_T}{N}\right) + L^2\left(\frac{C_W R_W}{2S}\right) \qquad \text{(EQ 11)}$$

Take partial derivatives with respect to N and S to minimize total delay:

$$\frac{\partial t_{interconnect}}{\partial N} = C_T R_W L - \frac{C_W R_T L}{N^2} = 0 \Rightarrow N = \sqrt{\frac{C_W R_T}{C_T R_W}} \qquad \text{(EQ 12)}$$

$$\frac{\partial t_{interconnect}}{\partial S} = C_T R_T - \frac{C_W R_W L^2}{2S^2} = 0 \Rightarrow S = L\sqrt{\frac{C_W R_W}{2 C_T R_T}} \qquad \text{(EQ 13)}$$

Hence, the total delay is:

$$t_{interconnect} = L\sqrt{\frac{C_W R_W C_T R_T}{2}} + L\left(\sqrt{C_W R_W C_T R_T} + \sqrt{C_W R_W C_T R_T}\right) + L\sqrt{\frac{C_W R_W C_T R_T}{2}} \quad \text{(EQ 14)}$$

$$= L\sqrt{C_W R_W C_T R_T}\,(2 + \sqrt{2}) \quad \text{(EQ 15)}$$

and the optimal length of each segment is:

$$\sqrt{\frac{2C_T R_T}{C_W R_W}} \quad \text{(EQ 16)}$$