High Speed CMOS VLSI Design

Lecture 5: Interconnect LRC

1.0 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 1. 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.

2.0 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:

$$\oint_C \vec{H} \bullet d\vec{s} = \int_S \vec{J} \bullet \; d\vec{a} + \frac{d}{dt}\int_S \varepsilon_0 \vec{E} \bullet \; d\vec{a} \qquad \text{(EQ 1)}$$

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

$$\oint_C \vec{E} \bullet d\vec{s} = -\frac{d}{dt}\int_S \mu_0 \vec{H} \bullet d\vec{A} \qquad \text{(EQ 2)}$$

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.

3.0 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 over-

whelms 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.

4.0 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.

5.0 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.

6.0 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:

$$\delta = \sqrt{\frac{2}{\omega\mu\sigma}} \qquad \text{(EQ 3)}$$

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.

7.0 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.

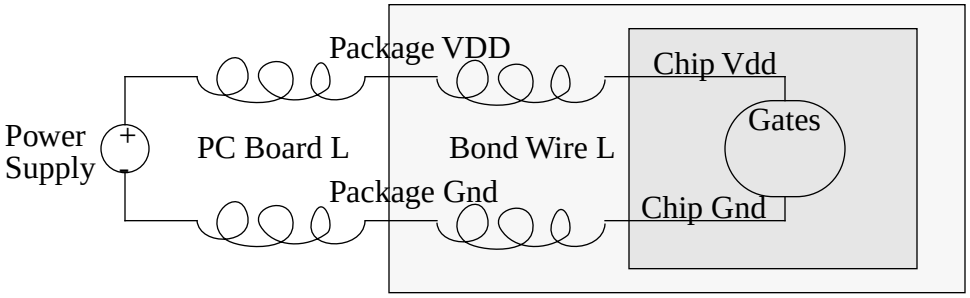
On-chip inductance is still in the research stage. Slow-mode waves and the skin effect will at worst cause increased wire delays. Ground bounce and inductive crosstalk can introduce enough noise to cause circuit failure in fast systems with poor return current paths. Designers who do not plan to minimize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

8.0 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.

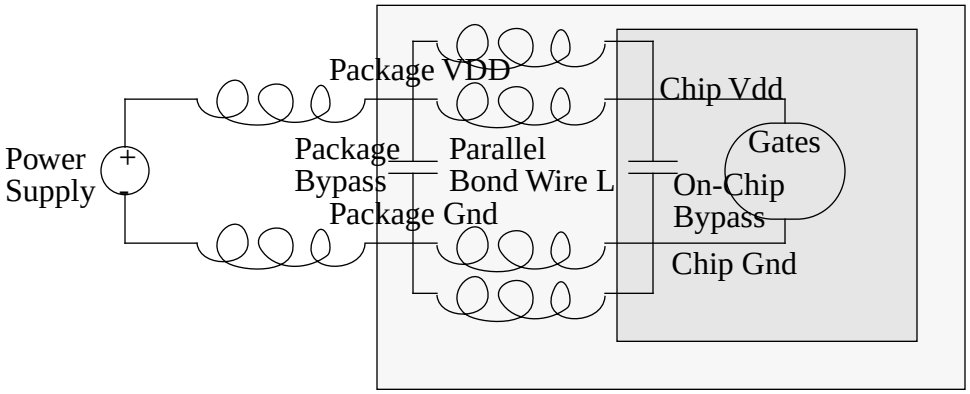
For example, consider the circuit in Figure 1. If all of the gates begin switching and raise the current from 0 to 10A in one nanosecond, $di/dt = 10^{10}$ A/s. If there is a total of 4 nH of inductance from bond wires and the PC board, the power supply would be reduced by $L\ di/dt = 40$ volts. This 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 1. 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 2. 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.

9.0 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.