Today's lecture: Scaling Laws for Transistors

I. Resistors, Capacitors, and Transistors
II. Delays of gates and wires
III. Scaling

Announcements:

Midterm on Feb. 28: Papers assigned for midterm:

- Architectural and Organizational Tradeoffs in the Design of the MultiTitan CPU. Norman P. Jouppi.
- A 0.18-μm CMOS IA-32 Processor With a 4-GHz Integer Execution Unit. Glenn Hinton, Michael Upton, *et al.*

Note that the IA-32 paper was previously assigned on the reading list but was never covered in lecture. I plan to post practice questions within a week.

1 Resistors, Capacitors, and Transistors

1.1 Resistors

A resistor restricts the flow of electrical current. For our water analogies, think of a narrow pipe – the narrower or longer the pipe, the more pressure is required to achieve the same flow.

The simplest model for an electrical resistor is to assume a linear relationship between voltage ($\sim$pressure) and current (i.e. flow). This is known as Ohm's law:

$$\begin{array}{rrcl} & I & = & \frac{V}{R} \\ \text{or, equivalently} & V & = & IR \end{array} \tag{1}$$

To keep the units straight, voltage is measured in "volts", where one volt is one joule per coulomb. A joule is a unit of energy (one kilogram$*$meter^2/second^2), and a coulomb is a unit of electrical charge ($-6.24 * 10^18$ electrons). Current is measured in "amperes ' where one ampere is one coulomb per second. Resistance is measured in ohms, where one ohm is one volt per ampere.

Now consider a rectangular bar of some resistive material (you can sketch it in in figure 1). Let ℓ be the length of the bar, and h and w be the height and width respectively. We expect the resistance to be proportional to the length, ℓ, and inversely proportional to the cross-sectional area, wh. The constant of proportionality is called the restivity of the material and is typically written with the Greek letter ρ. We have

$$R = \frac{\ell}{wh}\rho \tag{2}$$

Note that resistivity is measured in units of ohms $*$ meters.

On an integrated circuit, the thickness of a conducting layer is typically determined by the manufacturing process – the designer can't change it. Thus, we can divide the resistivity by the thickness of the layer to get the sheet resisitivity:

$$\rho_\square = \frac{\rho}{h} \tag{3}$$

Sheet resisitivity is in units of ohms, typically spoken as "ohms per square".

Figure 1: A resistor as a rectangular bar of metal

Figure 2: Resistors in series and parallel

Figure 3: A capacitor as a two parallel conductors

It's handy to know formulas for resistors in series and resistors in parallel. They are:

$$\begin{array}{rcl} R_{\text{series}} & = & R_1 + R_2 \\ 1/R_{\text{parallel}} & = & (1/R_1) + (1/R_2) \end{array} \tag{4}$$

1.2 Capacitors

A capacitor stores charge. A capacitor is typically formed by having two conductors separated by an insulator (you can draw the sketch in figure 3). To continue our water examples, a capacitor is like a water tank. You can pump water into the tank (from the bottom), but the more water that you pump in, the more pressure you need to exert to add more water. If we assume a linear relationship, we get

$$Q = CV \tag{5}$$

where Q is the charge stored in the capacitor in coulombs, and C is the "capacitance" of the capaictor. Capacitance is measured in "farads" where one farad is one coulomb per volt.

The capacitance of a water tank is proportional to its cross-sectional area. Likewise, if a capacitor consists of two parallel plates, each a rectangle that is w by ℓ, then the capacitance is proportional to $w\ell$. To figure out the constant of proportionality, we need to look a little more closely at our water tank analogy. In a water tank, pressure builds as the tank is filled because of the weight of the water. Electrons don't weigh very much. Instead, the pressure develops because electrons are repelled from one another. Charge builds up on the plates of a capacitor because the electrons on one plate are attracted to the positive charges on the other plate, even though they can't get there through the insulator. We can think of the insulator as a flexible barrier. Although the electrons can't get through, they can deform it and make room for more electrons. The thicker the insulator is, the harder it will be to deform. Thus, the capacitance is inversely proportional to d, the distance between the two plates, and we write ϵ to indicate the constant of proportionality. We get:

$$C = \frac{wl}{d}\epsilon \tag{6}$$

where ϵ is the "dielectric constant" in units of farads per meter. It is common to measure the "dielectric constant" as a multiple of that for a vacuum:

$$\epsilon = \epsilon_0 \epsilon_R \tag{7}$$

where ϵ_0 is the dielectric constant for a vacuum ($\epsilon_0 = 8.854 * 10^{-12}$ F/m), and ϵ_R is the relative dielectric constant. For glass (the common insulator between layers in a chip), the relative dielectric constant is about 4. For silicon nitride (the insulator between the gate and the channel, the relative dielectric constant is about 7.5. About the lowest dielectric constant for a solid is for teflon which is around 2.

Figure 4: A transistor

We can take our formulas above, and derive more useful formulas for understanding capacitors. First, we can calculate how much energy it takes to charge a capacitor from 0 to V volts. Note that a volt times a coulomb is a joule. For each coulomb that we shove into the capacitor, we can calculate what pressure is applied. This gives us the formula:

$$\begin{aligned} E &= \int_0^u Cu\,du \\ &= \tfrac{1}{2}CV^2 \end{aligned} \tag{8}$$

Note that $C\,du$ is the incremental charge, and u is the pressure that must be exerted to move that charge onto the capacitor.

Next, we can differentiate equation 5 with respect to time and get

$$\begin{aligned} \tfrac{d}{dt}Q &= C\tfrac{d}{dt} \\ I &= C\tfrac{d}{dt} \end{aligned} \tag{9}$$

where we've used the relationship that current is the rate of change of charge. In all of these formulas, we've assumed that C is constant. When capacitors are formed by semiconductors (i.e. around the transistors), the actual capacitance depends on the voltages of the various semiconductors. You can think of this as a water tank with varying cross-sectional area (curvy walls). For the analysis presented in this class, we'll make the simplifying approximation that capacitances don't vary. This obliterates many details of real circuit behaviour, but will be adequate for us to understand the basic scaling and asymptotic properties.

It's handy to know formula for capacitors in parallel (we won't be worrying about capacitors in series in this class). The formula is:

$$C_{\text{parallel}} = C_1 + C_2 \tag{10}$$

1.3 Transistors

As you can draw in figure 4, a transistor can be modeled as a switch that has some resistance when it is on. Likewise, it has capacitances from its gate, source, and drain, to ground. We write ℓ for the "length" of the transistor, this is the

Figure 5: A simple RC circuit

distance from the source to the drain. We write w for the "width" of the transistor, this is the parallel extend of the source and drain along the channel. Typically, transistors are much wider than they are long, but these are this is the standard terminology.

When a n-channel transistor is conducting, the gate has attracted a thin layer of electrons to the top of the channel. We can think of this as a sheet of resistive material, and we'll write $\rho_{n\square}$ to denote the sheet resisitivity of this material. From this, we have that the on-resistance of a n-channel device is

$$R_{\text{on,n}}(w,\ell) \quad = \quad \frac{\ell}{w}\rho n\square \tag{11}$$

Likewise, we can write $\rho_{p\square}$ fo the sheet resistivity of a p-channel transistor when it is conducting to get

$$R_{\text{on,p}}(w,\ell) \quad = \quad \frac{\ell}{w}\rho p\square \tag{12}$$

The capacitance of the gate and the channel form a parallel plate capacitor, and we conclude:

$$C_{\text{gate}}(w,\ell) \quad = \quad \frac{w\ell}{d_{Si_3N_4}}\epsilon_{Si3N_4} \tag{13}$$

Finally, the source and drain capacitances are proportional to the width of the transistor. The source capacitor is formed by the source-depletionLayer-substrate sandwich and likewise for the drain.

2 Delays

2.1 RC circuits

The critical thing to understand is that the product of a resistance and a capacitance is a time:

$$\begin{array}{rcl} 1\text{ohm} * 1\text{farads} & = & ((1\text{volt})/(1\text{ampere})) * ((1\text{coulomb})/(1\text{volt})) \\ & = & (1\text{coulomb})/(1\text{ampere}) \\ & = & (1\text{coulomb})/((1\text{coulomb})/(1\text{second})) \\ & = & 1\text{second} \end{array} \tag{14}$$

Consider the circuit you can draw in figure 5. Let the switch set the input to 0 volts for all time up to time 0. Then, the voltage on the capacitor will be 0 volts at time 0. At time 0 flip the switch to set the input voltage to U volts. We

Figure 6: A two-inverter chain

have:

$$\begin{array}{rcll} I_R(t) & = & (V_{in}(t) - V_{out}(t))/R, & \text{current through the resistor} \\ I_C(t) & = & C\frac{d}{dt}V_{out}, & \text{current through the capactor} \\ I_R(t) & = & I_C(t), & \text{Kirchoff's current law} \\ \frac{d}{dt}V_{out}(t) & = & \frac{V_{in}(t)-V_{out}(t)}{RC}, \text{a little algebra} & \\ V_{out}(0) & = & 0, \text{assumed} & \\ V_{in}(t) & = & U, \text{assumed, for } t > 0 & \\ V_{out}(t) & = & (1 - e^{-t/RC})U & \end{array} \tag{15}$$

Thus, RC is the time for the signal to transition to $(1 - e^{-1}) \approx 0.63$ of its final value. Since we're not worried about little constants here and there in this presentation, we'll consider this to be the transition time of the circuit.

3 Circuits with transistors

Typically, we want our circuits to go fast while using as little energy as possible. Note that making a transistor longer increases both its capacitance and its resistance. Neither helps us with our speed or energy goals. Thus, we'll assume that all transistors are designed to their minimum allowed width. In real circuits, there may be reasons to occasionally violate this assumption, but this rule is good enough for us to figure out the big picture trends. Let ℓ_0 be this minimum allowed transistor length. We can simplify our earlier formulas by defining:

$$\begin{array}{rcl} r_n & = & \ell_0 * \rho_{n,\square} \\ r_p & = & \ell_0 * \rho_{p,\square} \\ c_g & = & \frac{\ell_0}{d_{Si_3N_4}}\epsilon_{Si_3N_4} \end{array} \tag{16}$$

Typically, $r_p \approx 2r_n$. Let $\alpha = r_p/r_n$. We'll make the additional assumption that gates are designed so that the resistance of the pull-up and pull-down networks are equal. For example, for an inverter, this means that the width of the p-channel transistor will be α times that of the n-channel one. This isn't necessarily optimal, but it's close enough, and it simplifies the analysis.

First, we'll consider an inverter that drives one other inverter that is the same size as itself (draw it in figure 6). For the delay, we get:

3.1 Scaling transistor sizes

λ

Let λ be a scaling factor of a chip. For example, λ could be the minimum length for a transistor. We use λ to compare different manufacturing processes. For simplicity, we'll assume that all dimensions on a chip scale at the same rate. Thus, if we reduce the minimum transistor length be a factor of two, the minimum wire width, wire spacing, wire thickness, gate-oxide thickness, etc., are all reduced by a factor of 2 as well.

Number of logic gates on a chip: λ^{-2}

Assuming that the size of the chip remains constant, then the number of gates scales as $1/\lambda^2$. Thus, if we reduce the minimum transistor length by a factor of two and scale everything accordingly, the number of logic gates increases by a factor of four.

Power supply voltage: λ

As transistors are made smaller, the operating voltage must be reduced. Otherwise, the thin oxide layer between the gate and the substrate of the transistors would break down. If λ is the minimum transistor length, a good rule-of-thumb is that the power supply voltage, V_{dd}, is roughly $\lambda * \frac{10\text{volts}}{\mu}$, where 1μ is one micron (i.e. 10^{-6} meters). See also the notes on voltage scaling below.

Transistor resistance: 1

Scaling the gate-oxide thickness and the power supply voltage by λ leaves the strength of the electric field (volts/meter) unchanged. This makes sense, we were scaling voltage to prevent a breakdown of the gate-oxide.

With a constant strength for the field, the concentration of electrons under in the channel under the gate of a n-channel transistor with the gate high is constant under scaling (and likewise for holes with a p-channel transistor). This means that the sheet resistance of the channel for a conducting transistor remains unchanged under scaling. The scaling preserves the aspect ratio (width/length) of the transistor. Thus, the resistance of an "on" transistor is unchanged by scaling. A good rule-of-thumb is that the resistance for a n-channel transistor is $20 * 10^3 \Omega/\square$ for a n-channel transistor, and twice that for a p-channel device.

Transistor capacitance: λ

Recall that capacitance is given by $\frac{w\ell}{d}\epsilon$. Scaling w, ℓ, and d all by λ scales the capacitance by $\lambda^2/\lambda = \lambda$.

Note that this means that a transistor of fixed width has the same capacitance under process scaling. A good rule of thumb is that gate capacitance is $2fF/\mu$, where $1fF = 10^{-15}F$. Drain capacitance is rougly 0.7 to 1.0 times the gate capacitance. With good careful layout, the drain capacitance can often be reduced to half of this value.

Gate delay: λ

Recall that delay is resistance times capacitance. For circuits where the delay is dominated by the logic gates (i.e. there are no "long" wires), this scales with the product of the transistor resistance and the transistor capacitance. Thus, gate delay scales as λ: if the minimum transistor length is reduced by a factor of two, the logic circuits will be twice as fast, and the clock frequency can be twice as high.

A common measure of "gate delay" is the delay for a simple inverter driving four inverters of the same size. This is called a "fanout-of-four inverter delay" and abbreviated $FO4$. A good rule of thumb is

$$FO4 \quad = \quad \lambda * \frac{0.5ns}{\mu} \tag{17}$$

It's worth noting that in manufacturing processes that have been optimized for high performance (i.e., processes for general purpose CPUs), the manufacturing people do some tricks that make the effective length of the transistor about half what is drawn. Simply put, they get the source and drain regions to spread out under the gate by a carefully controlled amount. Because of this, the *FO4* delay for these processes (e.g. the manufacturing process for a Pentium-4) is about half what you would expect from the stated feature size.

Wire Resistance: λ^{-1}

Recall that the resistance of a rectangular bar is $(\ell/(wh))\rho$. Scaling w, ℓ, and d all by λ scales the resistance by $\lambda/(\lambda^2) = \lambda^{-1}$.

Note that the transition from aluminum wiring to copper provided a one-time, reduction of wire resistance by about 30much opportunity to go further in this direction (the resistivity of silver is only slighly lower than that of copper).

Wire Capacitance: λ

Same reasoning as for transistor capacitance.

Short Wire Delay: 1

Just multiply wire resistance by wire capacitance. Note that this is for wires whose length scale with everything else. This is what is meant by "Short Wire".

Long Wire Delay (unbuffered): λ^{-2}

Real designs have some fraction of their wires that cross the entire chip. Thus, w and h scale with λ, but ℓ remains fixed. We now get a capacitance of $((w+h)r/d)\epsilon$ where r is the distance across the chip, w is the wire width, h is wire height, and d is wire spacing. We have that w, h, and d scale as λ, and r and ϵ are constant. Thus, long wire capacitance scales as 1.

The resistance of a long wire is $r/(wh))\rho$ which scales as λ^{-2} thus long wire delay scales as λ^{-2}.

Buffer spacing: $\lambda^{1.5}$

As noted earlier, the clock period for logic scales as λ. Wire delay can be reduced by inserting buffers. Wire delay is minimized by choosing the separation between wires such that the wire segment delay equals the buffer delay. We've shown that buffer delay (a special case of gate delay) scales as λ. Thus, we need to choose our wire length, x such that wire delay scales as λ as well. The delay for a wire segment of length ℓ is:

$$\delta(\ell) \quad = \quad \frac{\ell}{wh}\rho * \frac{(w+h)\ell}{d}\epsilon$$

Noting that w, h, and d scale as λ, and ρ and ϵ are constant, we get that $\delta(\ell)$ scales as ℓ^2/λ^2. Thus, if the wire delay is to scale as λ, then ℓ^2 must scale as λ^3. This means that ℓ must scale as $\lambda^{1.5}$.

Thus, if we reduce the transistor length by a factor of two, the number of (smaller) gates that a wire can cross per *FO4* delay goes down by a factor of $\sqrt{2}$. If the clock period remains a fixed number of *F04* delays, then the number of optimally buffered segments that can be traversed in a clock period remains constant under scaling. However, these segments traverse $\sqrt{\lambda}$ fewer gates. Thus, the number of gates within one clock period of another goes down by a factor of λ. In other words, the region of synchronous design shrinks.

Power consumption (first analysis): 1 Power is, to a rough approximation, $\frac{\alpha}{2}nCV_{dd}^2 f$ where n is the number of logic gates; C is the capacitance per logic gate; V_{dd} is the power supply voltage; f is the clock frequency; and α is the fraction of clock cycles that each node changes. If we just scale the design, α remains constant; n goes as λ^{-2}; C goes as λ; V_{dd} goes as λ, and f goes as λ^{-1}. We multiply it all together and see that power consumption remains constant. This seems like wondeful news.

3.2 Power consumption (what went wrong?)

Voltage scaling down slower than predicted

Made possible by better materials and manufacturing.

Improves performance.

Necessitated by leakage currents.

Wires are taller than predicted by simple scaling. Taller wires reduce resistance, improves performance.

More trade-offs in wiring.

3.3 Voltage scaling