MIT Reversible Computing Project

Memo #M9

Optimal Voltages for CMOS Circuits

WORKING DRAFT MEMO

Revision: 1.3

Michael P. Frank
MIT AI Lab, Rm. 747
545 Technology Sq.
Cambridge, MA 02139
http://www.ai.mit.edu/~mpf

Started December 11, 1997.
Revision Date: 1997/12/11 18:01:30 GMT
Formatted December 11, 1997.

A current version is available online at
http://www.ai.mit.edu/~mpf/rc/memos/M09_voltage/voltage.html

Abstract

As CMOS logic circuits become smaller and faster, power dissipation starts to become a limiting factor on performance. By operating a circuit at lower voltages, we can reduce its power dissipation, but at the cost of increasing the effective resistance of its transistors, leading to larger RC delays. Thus, for any given CMOS technology, there is an optimal operating voltage, at which speed is equally limited by both RC delay and power dissipation. This optimal voltage depends only on the cooling capacity per logic gate, and on the effective transconductance of the gate's logic network, and interestingly is independent of load capacitance.

For an example $0.5\,\mu$m technology and a densely packed design that minimizes price/performance, we estimate the optimal voltage would be about $0.7\,$V, given a maximum heat flux of $100\,\mathrm{W/cm^2}$. At this voltage, the RC delay and dissipation time are both around $43\,$ps, suggesting a multi-gigaHertz maximum clock speed which is 20 times faster than the speed limit due to power dissipation when running at the $3.3\,$V voltage level for which the technology was designed. Furthermore, as device lengths decrease, the optimal voltage scales down proportionately, and the speed when running at this voltage increases proportionately as well.

However, below about $0.05\,\mu$m length scales, we will be unable to lower the power dissipation of room-temperture devices enough to continue these speed improvements, since sub-threshold conduction will then dominate the power consumption, and prevent further increases in device density. Soon after this point, speeds will be limited to a constant, determined by the power dissipation per device, regardless of any further decreases in line widths. Any improvements in price/performance beyond this point will require either sophisticated cooling systems, low-temperature operation, or some new, non-MOSFET device technology.

1 Analysis

One figure of merit that we generally wish to minimize in a computer circuit is its cost/performance ratio. For parallelizable tasks, performance can often be improved by replication of the circuit, but if doing this takes more silicon surface area, it may not by iteself be a cost-effective solution to the problem at hand. So generally we would like to improve cost/performance, that is, improve the performance (number of operations per second) per given unit of area.

Formally speaking, if A is the average silicon area per logic device, and n is the number of devices in an IC component, then the cost \$ of the component is generally proportinal to nA, that is, $\$ \sim nA$. The performance p, on the other hand, can be considered proportional to the number of devices times the frequency of operation f of each device, that is $p \sim nf$. So the price/performance ratio $\$/p \sim nA/nf \sim A/f$. The frequency f can be represented as the inverse of the time t taken by an individual device operation, $f = 1/t$. Therefore, $\$/p \sim At$. So to minimize price/performance, we must minimize the area-time product At of our devices.

Supposing the device area A to be fixed, this tells us that we want to minimize the time t per device operation. We will now proceed to do this. However, we shall see that it is important to keep in mind that what we really want to minimize is the area-time product At, since otherwise, when heat removal is the limiting factor on speed, we might be misled into thinking we should just spread our devices out over larger areas. Recalling the cost factor A prevents us from making this mistake.

Now, we examine constraints on the time t between device operations. First, this time is constrained to be greater than the characteristic time RC required to bring the voltage on the logic gate's output capacitance C a significant fraction of the way to the desired voltage during switching, given the effective resistance R of the path through which the output is charged. In other words, $t > RC$. Depending on our threshold voltages and noise margins, we may actually wish to allow several RCs before considering the output valid, and permitting a new input to be processed. Let us call the number of RCs that constitute a sufficient waiting period our *delay factor* d, and assume for now that d is a constant independent of all other parameters. A value of around $d = 3$ should be sufficient, in an aggressive design. Thus we have:

$$t \geq dRC. \tag{1}$$

The second constraint on t is that it must be a large enough time for the heat dissipated during the device's operation to diffuse out of the device and be removed by whatever cooling mechanism is available. We assume that any cooling system will have some upper limit F on the heat flux that can be removed per unit time per unit of die surface area. Thus, the maximum power dissipation per device is $P = FA$. The minimum energy that is dissipated when a logic gate's output capacitance C is switched through a voltage V is well-known to be $\frac{1}{2}CV^2$. Generally in switching a CMOS device there is energy dissipation due to transient short-circuit currents, and to leakage as well. However, in fast circuits, the dissipation is usually dominated by the $\frac{1}{2}CV^2$ switching energy, so for now we will ignore the other contributions to energy.

In any case, we know that the time t must be large enough for the energy $\frac{1}{2}CV^2$ to be removed by the cooling power P, that is,

$$t > \frac{\frac{1}{2}CV^2}{P}. \tag{2}$$

One way to decrease this bound is to increase the allowable dissipation $P = FA$. If this is done by increasing the heat-removal flux F per unit area, that's useful, but if we try to increase P by increasing the device area A, we have no net effect on the area-time product At, which is ultimately what we would really like to minimize. (If we only wanted to increase performance without regard for cost, we could often just replicate the whole component and increase parallelism, rather than improving device performance. One reason we don't do this is that cost is important too.)

Another way to decrease this power bound, when it is the limiting factor on t, is to decrease V. This will work for a while, but it will tend to increase the resistance R of our logic gates, by lowering the number of charge carriers in the transistor channels, which will increase the RC bound on time. Thus, at some particular value of V, the time t will be minimal; increasing V below this point will increase t due to the the greater power dissipation, and decreasing V will increase t due to the higher resitance through the transistors.

Let us now derive an approximate formula for the value of V at which the delay t is minimized. This occurs when the bounds on time given by equations (1) and (2) are equal:

$$dRC = \frac{CV^2}{2P}. \tag{3}$$

Before we solve this for V we must break down the left-hand side in terms of V. The relationship between V and R is rather complex, since the effective resistance of a MOSFET is not constant during charging; it depends on the output voltage, on short-channel effects, and on the device's threshold voltage. Moreover, the resistance through the transistor network of a multi-input logic gate is generally data-dependent.

However, we will ignore all these nit-picky factors and observe that the resistance through a network of MOSFETs is, roughly speaking,

$$R \approx \left(\frac{L}{W}\right)\frac{1}{k'V},$$

where L and W are the average effective channel length and width, respectively, and k' is the transconductance parameter for transistors in the given technology, with units of current per square voltage.

Furthermore, in a good high-speed design, logic gate fan-in is kept small, and devices are sized so that the typical effective L/W is kept about the same as that for a single minimum-sized MOSFET, where the channel width is about the same as its length. We have to be a little careful here, since widening our devices also affects the area factor A in the area-time metric, and also the increases the load capactance C on upstream gates. But for now we will assume these effects are relatively small. In any case they would introduce constant factors that should not radically affect the overall form of our scaling results significantly. Anyway, for now we will take the effective L/W to be about 1, so that:

$$R \approx \frac{1}{k'V}. \tag{4}$$

Substituting this into eq. (3), we get

$$\frac{dC}{k'V} = \frac{CV^2}{2P}.$$

which can be readily solved for V:

$$V = \left(\frac{2dP}{k'}\right)^{1/3}. \tag{5}$$

Note that the Cs cancel out, indicating that the optimal voltage point is independent of the load capacitances in the circuit.

How does this voltage scale as the device length scale ℓ decreases? Well, $P \sim A$, and $A \sim \ell^2$, so $P \sim \ell^2$. We know from device physics that $k' = \mu C_{\mathrm{ox}}$, where μ is a mobility constant and C_{ox} is the per-area gate capacitance, $C_{\mathrm{ox}} = \epsilon_{\mathrm{ox}}/t_{\mathrm{ox}}$; ϵ_{ox} is constant and t_{ox} is the gate oxide thickness. With all lengths scaling equally, $t_{\mathrm{ox}} \sim \ell$, so $C_{\mathrm{ox}} \sim 1/\ell$ and $k' \sim 1/\ell$. So $V \sim \left(\frac{\ell^2}{1/\ell}\right)^{1/3} = \ell$. The optimal voltage scales down with device length. This is consistent with other important constraints on voltage: if voltages were to get too high compared to gate oxide thicknesses, the gate oxides will short out due to the high field strengths.

Let us now derive a formula for the time per operation at the optimal voltage. Substituting eq. (5) into eq. (4), we have that the effective resistance is

$$R = \frac{1}{k'}\left(\frac{k'}{2dP}\right)^{1/3}$$

which we can reexpress as

$$R = \frac{1}{(2dPk'^2)^{1/3}}. \tag{6}$$

Since $P \sim \ell$ and $k' \sim 1/\ell$, this R stays relatively constant as length scales change.

Continuing, the minimum time from eq. (1),

$$t = dRC,$$

becomes

$$t = \frac{dC}{(2dPk'^2)^{1/3}}.$$

Since $C \sim C_{\mathrm{ox}}WL \sim (1/\ell)\ell^2 = \ell$, the time per operation at this optimum voltage should scale down roughly in proportion to length. This is as expected.

2 Example

Using equation (5), we calculated the optimum voltage for an example device technology, namely the $0.5\,\mu$m technology HP14 supplied by Hewlett-Packard. The above analysis assumes that the threshold voltages of the technology are adjusted to be enough below whatever optimal voltage we settle on is still enough for the devices to still work.

The fundamental device parameters of the technology are shown in table 1.

In order to estimate the allowable power per device P, we assumed an overall cooling capacity of $100\,\mathrm{W/cm^2}$, probably about the highest value attainable with cooling techniques in common use today, and an average logic gate area

Parameter	Value	Meaning
V_{dd}	$3.3\,\mathrm{V}$	Suggested operating voltage
L_{min}	$0.5\,\mu\mathrm{m}$	Minimum channel length
t_{ox}	$9\,\mathrm{nm}$	Gate oxide thickness
μ_{n}	$978\,\mathrm{cm}^2/\mathrm{Vs}$	Electron mobility (n-channel)
C_{ox}	$3.9\,\mathrm{fF}/\mu\mathrm{m}^2$	Gate capacitance per area, $\epsilon_{\mathrm{ox}}/t_{\mathrm{ox}}$
k'	$380\,\mu\mathrm{A}/\mathrm{V}^2$	Transconductance, $\mu_{\mathrm{n}} C_{\mathrm{ox}}$

Table 1: Parameters of an example technology (HP14).

of $25\,\mu\mathrm{m}^2$ corresponding to a very densely-packed design. This imples a maximum power dissipation per gate of $25\,\mu\mathrm{W}$. We assumed the minimum tolerable number of RC delays to be $d = 3$.

The optimum voltage as derived from equation (5) then becomes $V = 0.734\,\mathrm{V}$, much smaller than the value of $3.3\,\mathrm{V}$ for which the technology was designed. In fact, this value of V is comparable to the threshold voltage of the technology, which means, among other things, that eq. (4) would actually be a very bad approximation for the resistance in this regime. In order for this low a supply voltage to be useful, we would require a significantly lower threshold voltage for the technology, perhaps on the order of 100–200 mV. However a more detailed analysis would be required to determine the exact optimal threshold.

The effective minimum-size device resitance at our optimum voltage, from eq. (6), turns out to be about $3600\,\Omega$. The exact speed depends on the load capacitance. The gate capacitance of minimum-sized devices is $0.975\,\mathrm{fF}$; assuming the average fanout to be four times this leads to $C = 3.9\,\mathrm{fF}$. The delay $t = dRC$ is then about 42 ps. This is 20 times smaller than the minimum delay of 850 ps that is required if running at the nominal voltage of 3.3 V, given our constraint on power dissipation.

This indicates that the HP14 technology could have permitted a much higher level of performance per unit area if it had permitted much lower threshold voltages, to support this low-voltage mode of operation.

3 Discussion of Scaling Limits

Note that as line lengths become even smaller, the optimal voltage would decrease even further. For an order of magnitude smaller $0.05\,\mu\mathrm{m}$ technology, the optimal voltage of $\sim$73 mV would only be about three times the room-temperature thermal voltage of $k_{\mathrm{B}}T/q = 26\,mV$, leading to a difference between "on" and "off" transistor currents of only approximately $exp(3) \approx 20$. At this point the leakage power starts to become comparable to the switching power, and so further decreases in voltage do not significantly decrease the power consumption. So device density stops improving because it is limited by the fixed leakage power per device. This takes away the ℓ^2 contribution that length scaling would otherwise make to the area-time product. However, we can still continue to improve At by decreasing the speed of individual devices by decreasing their load capacitances. In fact, with the oxide thickness constant, gate capacitances decrease with the square of the length scale. But this just means that soon, wire capacitances will dominate the capacitance, since the wires are not getting any shorter after the device density has reached its limit. We can make the wires thinner to keep their capacitance going down, but this increases their resitance, so that soon the wires dominate the total resistance, and thus dominate the total RC delay in the circuit, and the RC of a fixed-length wire cannot be reduced by making it thinner. So at this point there are no further speed increases, and we have hit a brick wall. Another way to see this fixed-speed limit coming is to note that once devices stop getting closer together, they cannot get faster indefinitely, because eventually the speed-of-light delay between them will dominate the time for them to charge up the wire leading to their neighbor. (Although conceivably this latency could be hidden to some degree by allowing a sequence of many signals at once to be carried by a wire behaving as a transmission line, or through some other medium supporting wave propagation or other near-ballistic means of information transfer.)

One way out of this impasse is to start scaling down to lower temperatures, which scales down the thermal voltage $k_{\mathrm{B}}T/q$ and allows everything else to scale down as before. Alternatively, if we can scale up the heat flux that we can remove with our cooling system, then devices can continue getting closer together for a while longer. Still, either approach will eventually hit some technologically or economically insurmountable barrier, and then we will need to abandon MOSFETs and jump to a completely different technology. It is not yet clear what that would be.

Symbol	Meaning
A	Area per logic gate
n	Number of gates in component
\$	Cost of component
p	Performance of component
f	Frequency of device operation
t	Time for device operation
R	Effective resistance of charging path through gate
C	Effective load capacitance on gate output
d	Number of RC delays to wait between inputs
F	Maximum rate of heat flux per unit area in cooling system
P	Maximum power dissipation permitted per device
V	Voltage difference between low and high logic values
L	Channel length
W	Channel width
k'	Transconductance parameter $\mu_\mathrm{n} C_\mathrm{ox}$
ℓ	Length scale
t_ox	Gate oxide thickness
μ	Electron mobility
C_ox	Gate capacitance per area, $\epsilon_\mathrm{ox}/t_\mathrm{ox}$
k_B	Boltzmann's constant (natural entropy unit), 1.38×10^{-23} J/K
ϵ_ox	Permittivity of SiO_2, 3.5×10^{-13} F/cm
T	Operating temperature, assumed room temperature or 300 K
q	Electron charge, 1.6×10^{-19} C
$\sim$	Relation: Scales proportionally to

Table 2: Notations used in this document.