Chapter 9

Estimating circuit speed

9.1 Counting gate delays

The simplest method for estimating the speed of a VLSI circuit is to count the number of VLSI logic gates that the input signals must propagate through to become output signals. This is a fairly straightforward method, except for the fact that a circuit symbol may actually be composed of multiple gates.

For instance, in Figure 9.1, the $\mathbf{S}_0$ is computed from the $\mathbf{X}_0$, $\mathbf{Y}_0$ and Carry in in 3 gate delays instead of 2, because the full adder is composed of two VLSI logic gates. One of the gates computes the carry from the three inputs, while the other computes the sum from the three inputs and the carry output. This is why the carry from the first full adder to the second full adder is computed in only two gate delays.

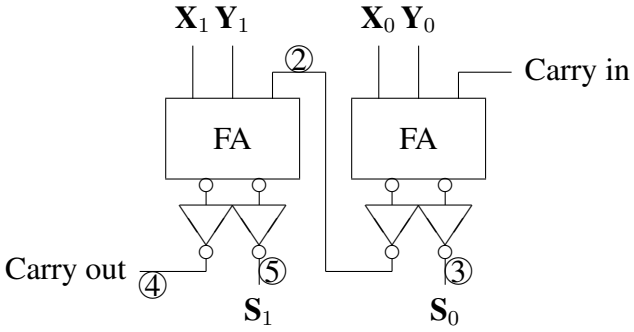

Figure 9.1: Two bit adder with computed gate delay.

Other symbols such as the AND and OR logic gates also hide multiple VLSI logic gates. Since the AND logic gate cannot be created directly, the logic gate must be constructed from a NAND logic gate and an inverter or from two inverters and a NOR gate. Similarly, the OR logic gate must be constructed from a NOR logic gate and an inverter or from two inverters and a NAND gate.

If the adder is optimized as described in Chapter 7 to eliminate inverters from the carry chain, then the gate delay changes as well, as shown in Figure 9.2. Note that in the Figure 9.2(a), the carry from the first full adder reaches the second full adder with same gate delay as the inverted inputs. In contrast, in Figure 9.2(b), the carry from the first full adder reaches the second full adder

with two gate delays, so that the sum takes longer to compute (although it will still be faster then the unoptimized adder of Figure 9.1 when extended to three or more bits).

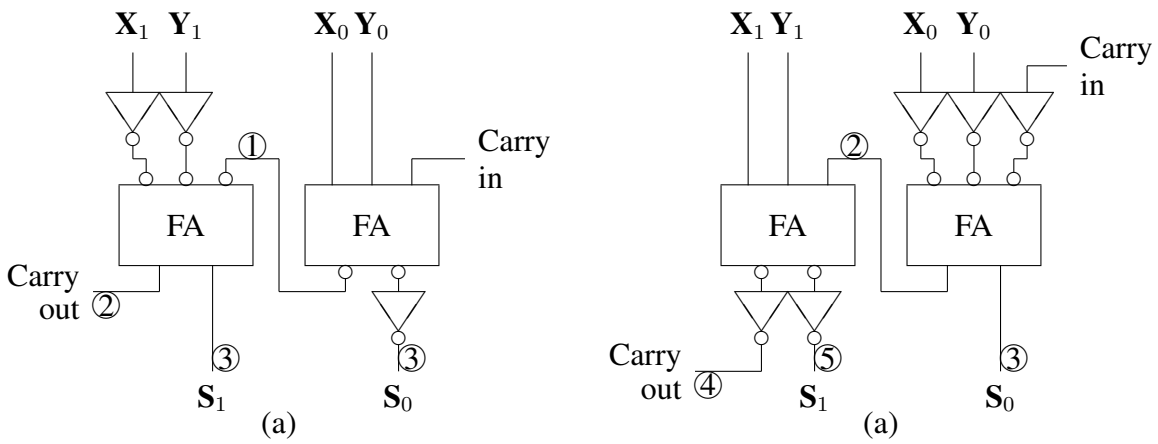

Figure 9.2: Optimized two bit adder.

9.2 Computing RC delays

Counting gate delays is a poor method for estimating circuit speed. It doesn't take into account the speed reduction that occurs when a VLSI logic gate drives many other VLSI logic gates, or the difference between series and parallel transistor structures within a VLSI logic gate. It also doesn't account for the effect of the transistor sizes on the VLSI logic gate's speed.

A better method is to model each transistor as a capacitor and a resistor. In this method, each transistor is modeled as having a capacitor between its gate terminal and ground, and a resistor and an ideal switch in series between the source and drain terminals as shown in Figure 9.3. The capacitor models the effect of driving multiple VLSI logic gates, and the resistor models the reduced speed of the circuit when the transistors are connected in series. As before the switch is open or closed depending on the logic level at the gate terminal.

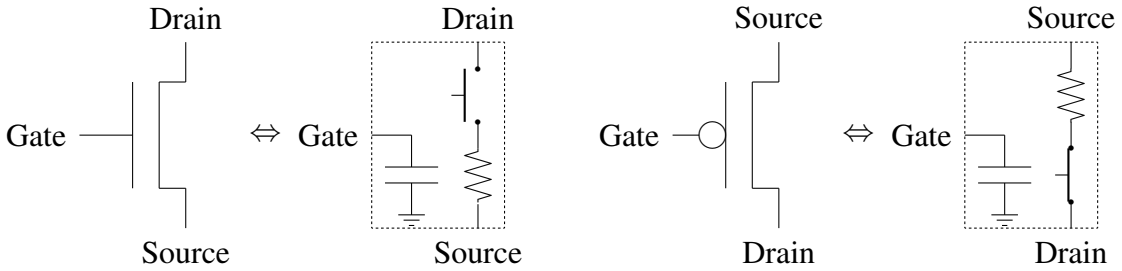

Figure 9.3: RC transistor model.

With this new transistor model, we can model the interaction between two logic gates. For instance, the model converts the two inverters shown in Figure 9.4(a) first into the circuit shown in Figure 9.4(b), then into the circuit shown in Figure 9.4(c).

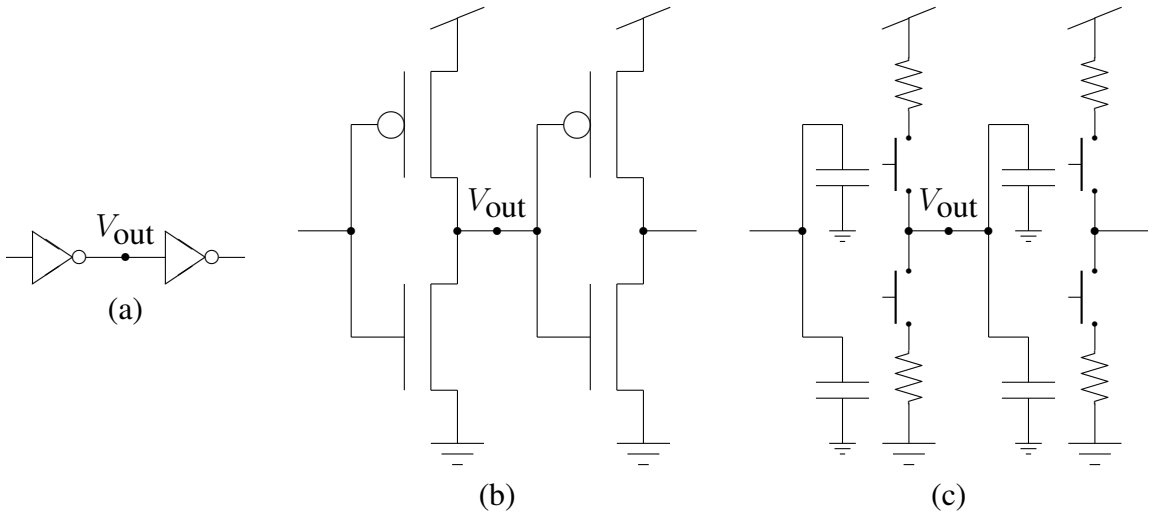

Figure 9.4: Modeling two inverters.

When the output of the first inverter transitions from 1 to 0, the resistor originating from the NFET in the first inverter will discharge the capacitance originating from the transistors in the second inverter. Hence the output will follow the exponential decay described by:

$$V_{\text{out}} = \text{Vdd} \cdot e^{\frac{-t}{R_{\text{N}}C}}$$

where t is time, R_{N} is the NFET resistance and C is the total capacitance at the input of the second inverter (that is, the sum of the capacitance C_{N} originating from the NFET and the capacitance C_{P} originating from the PFET). This exponential decay is sketched in Figure 9.5(a), and the product $R_{\text{N}}C$ is referred to as the *time constant*.

When the output of the first inverter transitions from 0 to 1, we also get an exponential decay, but one that approaches Vdd asymptotically instead of Gnd. This exponential decay is described by:

$$V_{\text{out}} = \text{Vdd} \cdot (1 - e^{\frac{-t}{R_{\text{P}}C}})$$

and is sketched in Figure 9.5(b). The resistance R_{P} in this equation is the resistance originating from the PFET, not the NFET, and may be quite different, so that the time constant $R_{\text{P}}C$ may also be quite different.

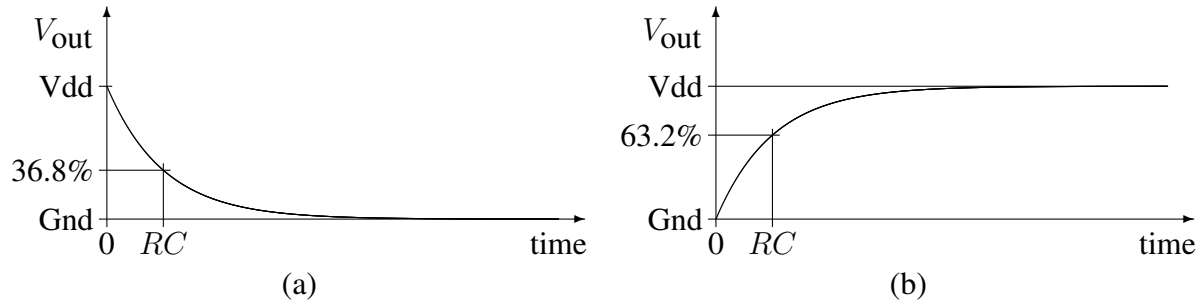

Figure 9.5: Exponential decays. (a) Decay to Gnd. (b) Decay to Vdd.

9.3 Effects of device sizing

Consider the two transistors shown in Figure 9.6(a). We can draw these two transistors in a layout by drawing a vertical diffusion (or active) rectangle with two separate horizontal polysilicon rectangles as shown in Figure 9.6(b). However, since the gates of both transistors are connected together, we can move the polysilicon rectangles towards one another until they touch as shown in Figure 9.6(c). This results in a single transistor of length L which is twice the length of the individual original transistors. Since the resistance of two series resistors adds and the capacitance of two parallel capacitors also adds, we can conclude that:

$$R_{\mathrm{N}} \propto \mathrm{L} \;\text{ and }\; C_{\mathrm{N}} \propto \mathrm{L}$$

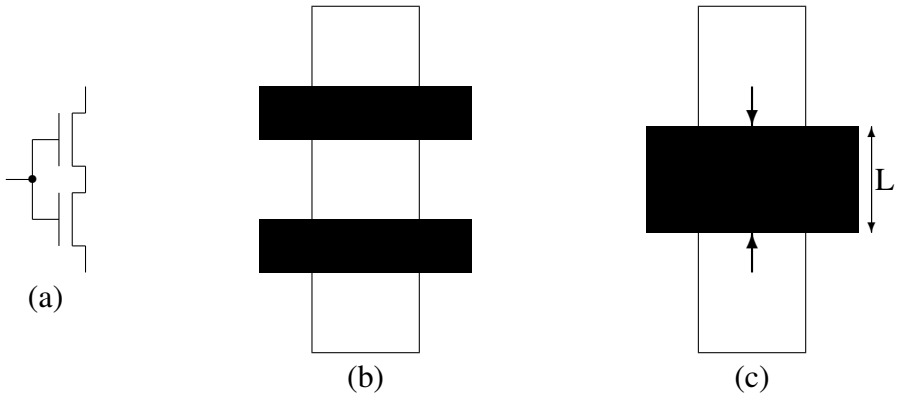

Figure 9.6: Series transistors with common gate. (a) Schematic. (b) Layout with separate polysilicon rectangles. (c) Layout with merged polysilicon rectangles.

We can perform the same exercise with parallel transistors as shown in Figure 9.7 to observe the effect of the width W. Since the conductance (reciprocal of resistance) of two parallel resistors adds and the capacitance of two parallel capacitors also adds, we conclude that:

$$R_{\mathrm{N}} \propto \frac{1}{\mathrm{W}} \;\text{ and }\; C_{\mathrm{N}} \propto \mathrm{W}$$

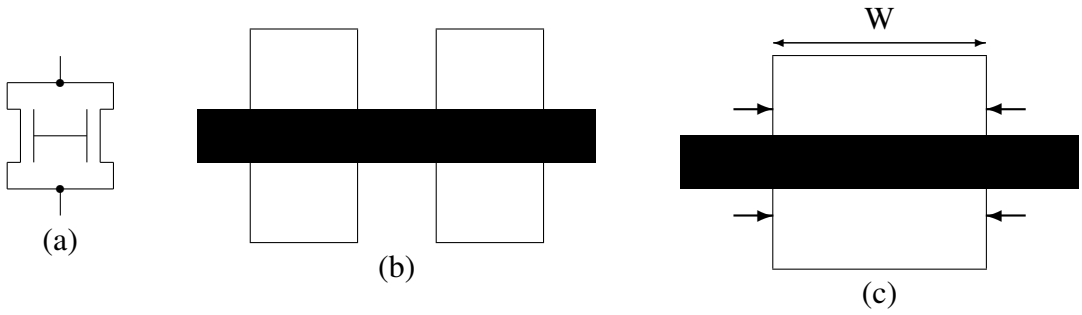

Figure 9.7: Parallel transistors with common gate. (a) Schematic. (b) Layout with separate diffusion rectangles. (c) Layout with merged diffusion rectangles.

Similar conclusions hold for PFETs, so that in summary, we have:

$$R_{\rm N}, R_{\rm P} \propto \frac{\rm L}{\rm W} \;\text{ and }\; C_{\rm N}, C_{\rm P} \propto {\rm L} \cdot {\rm W}$$

9.4 Adjusting device width and length

In section 9.2 we found that the shape of the transient waveform for a high to low transition is an exponential decay. This means that the *fall-time* (the time from when the output starts to fall to when it reaches the input threshold of the next gate) is proportional to the time constant. This is because if Vinv is the input threshold for an inverter, then the fall-time t_f is given by:

$$\text{Vinv} = \text{Vdd} \cdot e^{\frac{-t_f}{R_{\rm N}C}} \;\Leftrightarrow\; t_f = R_{\rm N}C \ln\left(\frac{\text{Vdd}}{\text{Vinv}}\right) \;\Rightarrow\; t_f \propto R_{\rm N} \cdot (C_{\rm N} + C_{\rm P})$$

Similarly, the *rise-time* t_r is also proportional to the time constant:

$$\text{Vinv} = \text{Vdd} \cdot (1 - e^{\frac{-t_r}{R_{\rm P}C}}) \;\Leftrightarrow\; t_r = R_{\rm P}C \ln\left(\frac{\text{Vdd}}{\text{Vdd} - \text{Vinv}}\right) \;\Rightarrow\; t_r \propto R_{\rm P} \cdot (C_{\rm N} + C_{\rm P})$$

If $\text{Vinv} = \frac{\text{Vdd}}{2}$ then the average propagation delay is proportional to the average time constant $t_{\rm RC}$:

$$t_{\rm pd} = \frac{t_f + t_r}{2} \propto \frac{R_{\rm N} + R_{\rm P}}{2} \cdot (C_{\rm N} + C_{\rm P}) \;\Rightarrow\; t_{\rm pd} \propto t_{\rm RC}$$

where

$$t_{\rm RC} = RC \;\text{ and }\; R = \frac{R_{\rm N} + R_{\rm P}}{2} \;,\; C = C_{\rm N} + C_{\rm P}$$

It is important to note that the resistance in these expressions is the output resistance of the transistors in the driving logic gate (such as the leftmost inverter in Figure 9.4) while the capacitance is the input capacitance of the transistors in the driven logic gate (such as the rightmost inverter in Figure 9.4).

Obviously, if all transistors have the same width and length the distinction is superfluous. However, suppose that we make the transistors in the driving logic gate α_w wider than the minimum size that we can draw, and the transistors in the driven logic gate β_w wider than the minimum size that we can draw. Then the conductance of the driving logic gate increases by a factor of α_w, and the capacitance of the driven logic gate increases by a factor of β_w. The effect on the propagation delay is:

$$t_{\rm RC} = \frac{R_{\rm N} + R_{\rm P}}{2\alpha_w} \cdot \beta_w \, (C_{\rm N} + C_{\rm P}) = \frac{\beta_w}{\alpha_w} RC$$

Hence, to improve speed (*i.e.* reduce the propagation delay), we generally want to make the driving transistors wider and the driven transistors narrower. (Note that this relationship is somewhat approximate, because changing the device widths also affect the Vinv.)

If on the other hand, if we make the transistors in the driving logic gate α_l longer than the minimum size that we can draw, and the transistors in the driven logic gate β_l longer than the minimum size that we can draw, then the resistance of the driving logic gate increases by a factor

of α_l, and the capacitance of the driven logic gate increases by a factor of β_l. The effect on the propagation delay is:

$$t_{\text{RC}} = \alpha_l \frac{R_{\text{N}} + R_{\text{P}}}{2} \cdot \beta_l \left(C_{\text{N}} + C_{\text{P}}\right) = \beta_l \cdot \alpha_l \cdot RC$$

Hence, to improve speed we generally want to make all transistors shorter.

9.5 Optimizing device width and length

Since increasing the width of the transistors can both improve and degrade the performance of a circuit, it is necessary to find the optimal widths on a case by case basis. Suppose we have three inverters as shown in Figure 9.8.

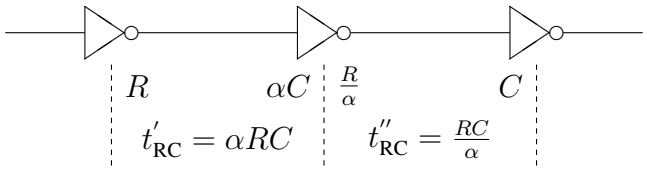

Figure 9.8: Time constant calculation for three inverters.

If we increase the width of the transistors in the middle inverter by α, then the overall time constant becomes:

$$t_{\text{RC}}(\alpha) = t'_{\text{RC}} + t''_{\text{RC}} = \left(\alpha + \frac{1}{\alpha}\right) RC$$

To optimize the value of α, we set the first derivative of $t_{\text{RC}}(\alpha)$ to zero as follows:

$$\frac{\mathrm{d}t_{\text{RC}}(\alpha)}{\mathrm{d}\alpha} = \left(1 - \frac{1}{\alpha^2}\right) RC = 0 \quad \Rightarrow \quad \alpha = 1$$

We should not conclude from this exercise that adjusting the transistor widths is pointless. Indeed, suppose that the middle inverter drives multiple inverters as shown in Figure 9.9.

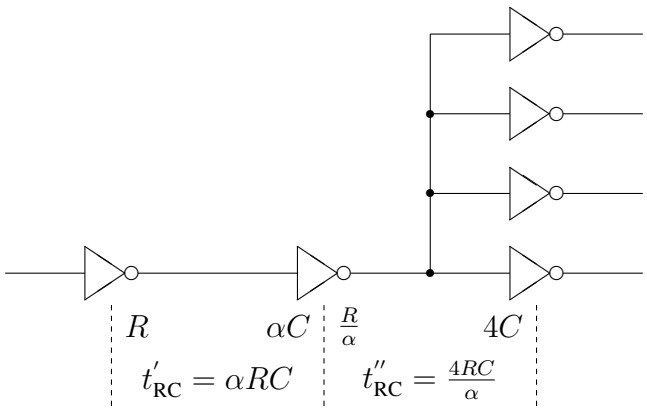

Figure 9.9: Time constant calculation for three inverters with a fan-out of 4.

The same optimization now yields:

$$\frac{\mathrm{d}t_{\mathrm{RC}}(\alpha)}{\mathrm{d}\alpha} = \left(1 - \frac{4}{\alpha^2}\right) RC = 0 \quad \Rightarrow \quad \alpha = 2$$

9.6 General optimization guidelines

Suppose we have n+1 inverters as shown in Figure 9.10, labeled from 0 to n, where the width of the transistors in inverter i is α_{i} times wider than the minimum (and $\alpha_0 = 1$).

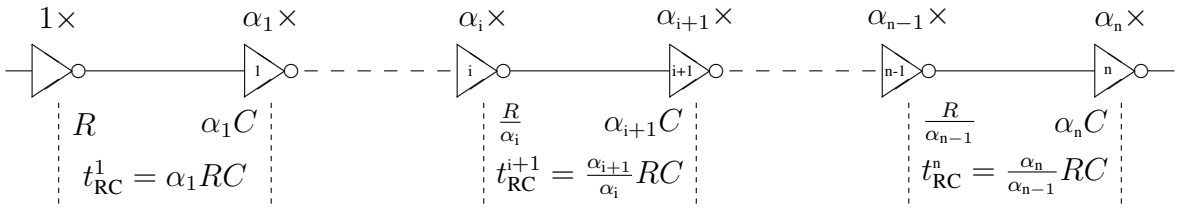

Figure 9.10: Time constant calculation for many inverters.

The overall time constant is the sum of all the individual time constants:

$$t_{\mathrm{RC}}(\alpha_1, \ldots, \alpha_{\mathrm{n}}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} t_{\mathrm{RC}}^{\mathrm{i}} = RC \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\alpha_{\mathrm{i}}}{\alpha_{\mathrm{i}-1}}$$

Since $t_{\mathrm{RC}}()$ is now a function of multiple variables, optimization require setting all the partial derivatives to zero as follows:

$$\frac{\delta t_{\mathrm{RC}}(\alpha_1, \ldots, \alpha_{\mathrm{n}})}{\delta \alpha_{\mathrm{i}}} = 0 \ \text{ for } \ \mathrm{i} = 1 \ldots \mathrm{n}$$

For every α_{i}, there are only two terms in the summation that contain α_{i}: $RC\frac{\alpha_{\mathrm{i}}}{\alpha_{\mathrm{i}-1}}$ and $RC\frac{\alpha_{\mathrm{i}+1}}{\alpha_{\mathrm{i}}}$. Hence:

$$\frac{\delta t_{\mathrm{RC}}(\alpha_1, \ldots, \alpha_{\mathrm{n}})}{\delta \alpha_{\mathrm{i}}} = \left(\frac{1}{\alpha_{\mathrm{i}-1}} - \frac{\alpha_{\mathrm{i}+1}}{\alpha_{\mathrm{i}}^2}\right) RC = 0 \quad \Rightarrow \quad \alpha_{\mathrm{i}}^2 = \alpha_{\mathrm{i}-1} \cdot \alpha_{\mathrm{i}+1}$$

This gives us the recurring relation:

$$\alpha_{\mathrm{i}+1} = \frac{\alpha_{\mathrm{i}}^2}{\alpha_{\mathrm{i}-1}} \ \text{ and } \ \alpha_0 = 1$$

for which the solution is (proved by induction):

$$\alpha_{\mathrm{i}} = \alpha^{\mathrm{i}} \ \text{ and } \ \alpha = \sqrt[\mathrm{n}]{\mathrm{N}}$$

where $\mathrm{N} = \alpha_{\mathrm{n}}$. This shows that the optimal transistor sizing yields a geometric progression, where each inverter is α times bigger than the previous one.

The overall time constant is now:

$$t_{\mathrm{RC}} = RC \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\alpha_{\mathrm{i}}}{\alpha_{\mathrm{i}-1}} = RC \sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha = \mathrm{n}RC\alpha = \mathrm{n}RC\sqrt[\mathrm{n}]{\mathrm{N}}$$

If we allow n to vary while keeping the size of the last inverter constant, we find that we can also optimize the number of inverters to minimize the overall time constant. To facilitate the minimization, we will instead minimize the time constant as a function of α. Since:

$$\alpha = \sqrt[\mathrm{n}]{\mathrm{N}} \;\Leftrightarrow\; \alpha^{\mathrm{n}} = \mathrm{N} \;\Leftrightarrow\; \mathrm{n}\ln\alpha = \ln\mathrm{N} \;\Leftrightarrow\; \mathrm{n} = \frac{\ln\mathrm{N}}{\ln\alpha}$$

we can express the overall time constant's dependence on α as:

$$t_{\mathrm{RC}}(\alpha) = \frac{\ln\mathrm{N}}{\ln\alpha} RC\alpha = RC\ln\mathrm{N}\frac{\alpha}{\ln\alpha}$$

If we plot $\frac{\alpha}{\ln\alpha}$ as a function of α, we get the curve shown in Figure 9.11.

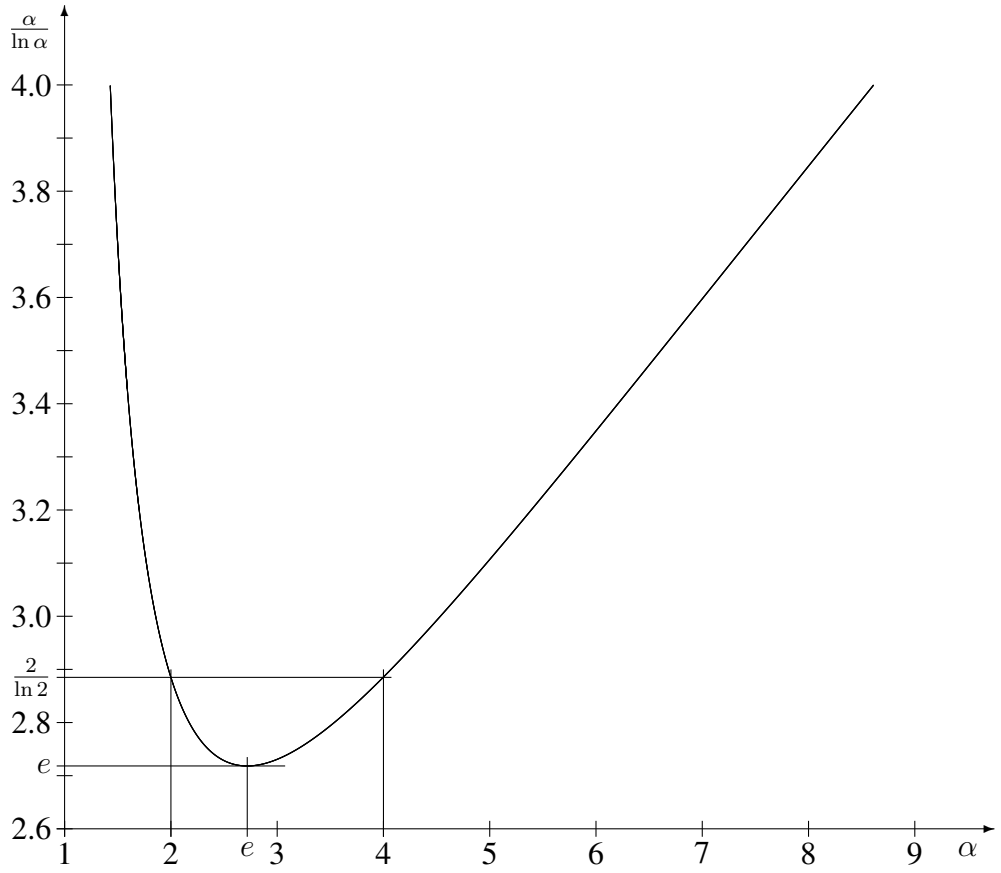

Figure 9.11: $\frac{\alpha}{\ln\alpha}$ as a function of α.

Setting the derivative of $t_{\mathrm{RC}}(\alpha)$ to zero we get:

$$\frac{\mathrm{d}t_{\mathrm{RC}}(\alpha)}{\mathrm{d}\alpha} = RC\ln\mathrm{N}\frac{\ln\alpha - 1}{(\ln\alpha)^2} = 0 \;\Rightarrow\; \ln\alpha = 1$$

The optimal value for α and n is therefore:

$$\alpha = e \;\text{ and }\; \mathrm{n} = \ln\mathrm{N}$$