Distributed RC Delay Line Model and MOS PLA Timing Estimation

Chao-Lin Chiang and S. Lennart Johnsson
YALEU/DCS/TR-738
September 1989

Distributed RC Delay Line Model and MOS PLA Timing Estimation

Chao-Lin Chiang and S. Lennart Johnsson*

Abstract

An accurate RC delay line model is derived for wires in MOS technologies. PLA and ROM delay models are derived using the wire delay model. The PLA delay model expresses the worst case delay in terms of the number of inputs, minterms, and outputs. With information about bit patterns more accurate estimates can be obtained.

1 Introduction

The resistance and capacitance of wires in MOS technologies affect the propagation of signals along wires. The wire delay in state-of-the-art MOS circuits is often larger than the charge transit time of transistors. The wire delay does not scale well with the technology [7]. Uniform scaling of circuit geometry without any change in doping levels results in constant delay for wires whose length is reduced by the scaling factor. For a wire of constant length the delay increases [1]. For a long wire the increase is proportional to the square of the scaling factor [9]. In Table 1, some typical nMOS circuit parameters are listed and the effect of these parameters on the delays in an nMOS circuit is shown.

To estimate the delay in digital MOS circuits, transient analysis of a circuit that contains distributed RC lines is important. In [5] a computationally simple method to estimate the delay of nodes in an RC tree network is presented. This paper develops a delay model for wires with uniformly distributed resistance and capacitance. The delay estimates are more accurate than those obtained in [5], and can be applied to estimate the delay of MOS storage circuits such as PLA and ROM.

In the next section, a MOS wire is modeled as a two-port network. The distributed resistance and capacitance are approximated by lumped resistors and capacitors. A simple

---

1

Feature	$\lambda = 2\mu m$		$\lambda = 0.25\mu m$	
Object	Resistance $\Omega/\square$	Rel. Length of Equal R	Resistance $\Omega/\square$	Rel. Length of Equal R
Transistor	10^4	1	10^4	1
Diffusion	100	100	800	12.5
Poly	25	400	200	50
Metal	0.03	3×10^5	0.24	4×10^4
	Capacitance $10^{-4}pF/\mu m^2$	Rel. Area of Equal C	Capacitance $10^{-4}pF/\mu m^2$	Rel. area of Equal C
Gate	5.0	1	40	1
Diffusion	1.2	4	10	4
Poly	0.5	10	4	10
Metal	0.33	15	2.7	15

Table 1: The parasitic Parameters as a chip size scales down.

equation for estimating the delay of a wire with uniformly distributed resistance and capacitance is derived. The wire is assumed to be driven by a voltage source having an internal resistance. The delay in a static nMOS PLA is estimated in section 3. The effect of scaling the technology is analyzed.

2   The Distributed RC Delay Line Model

Wires in MOS circuits are often modeled as transmission lines as shown in Figure 1 [9, 1]. Signal propagation across the wire is described by the one-dimensional diffusion equation:

$$rc\frac{\partial V(x,t)}{\partial t} = \frac{\partial^2 V(x,t)}{\partial x^2} \tag{1}$$

where $V(x,t)$ is the voltage along the wire, t time, x the spatial coordinate, r the resistance and c the capacitance per unit length.

The solution of this diffusion equation is complex and it is unreasonable to solve it for every element in a VLSI circuit. However, with uniformly distributed resistance and capacitance along the wire, simple approximate solutions of good accuracy can be found. In the following a wire with uniformly distributed resistance and capacitance is referred to as a URC element [4, 5].

Two-port Analysis of a URC Element

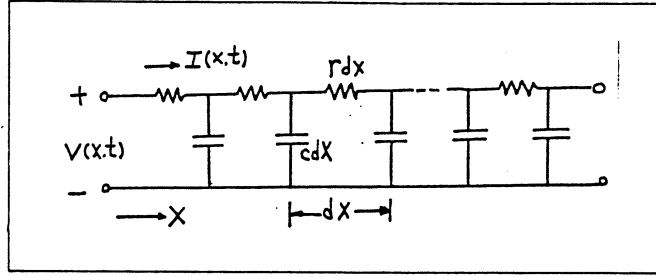A URC element has two parameters: total resistance R and total capacitance C. Let r

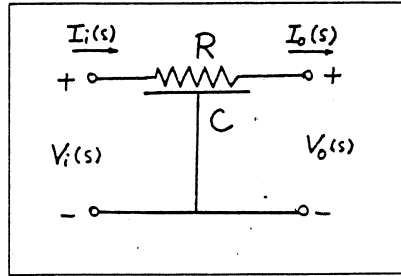Figure 1: A wire modeled as a transmission line.

Figure 2: A URC element modeled as a two-port network.

and c be the resistance and capacitance per unit length, respectively. Then for a wire of length L, $R = rL$ and $C = cL$.

A URC element can be considered as a two-port network [3, 4]. Its transmission matrix in the Laplace domain is:

$$\begin{bmatrix} \cosh\Gamma & \frac{\sinh\Gamma}{Z} \\ Z\ \sinh\Gamma & \cosh\Gamma \end{bmatrix}$$

where $\Gamma = \sqrt{sRC}$, and $Z = \sqrt{\frac{R}{sC}}$. That is,

$$\begin{bmatrix} V_i(s) \\ I_i(s) \end{bmatrix} = \begin{bmatrix} \cosh\Gamma & \frac{\sinh\Gamma}{Z} \\ Z\ \sinh\Gamma & \cosh\Gamma \end{bmatrix} \begin{bmatrix} V_o(s) \\ I_o(s) \end{bmatrix} \tag{2}$$

where $V_i(s)$, $V_o(s)$, $I_i(s)$, $I_o(s)$ are the Laplace transforms of the input voltage, output voltage, input current and output current, respectively, as shown in Figure 2.

Lumped Approximation of a URC element

A URC element can be approximated to first order by lumped resistors and capacitors. Expanding the entries of the transmission matrix in equation (2) in a Taylor series and keeping only first order terms results in,

$$\begin{bmatrix} \cosh\Gamma & \frac{\sinh\Gamma}{Z} \\ Z\ \sinh\Gamma & \cosh\Gamma \end{bmatrix} = \begin{bmatrix} 1 + \frac{sRC}{2} & R \\ sC & 1 + \frac{sRC}{2} \end{bmatrix} + \text{high order } s \text{ terms} \tag{3}$$
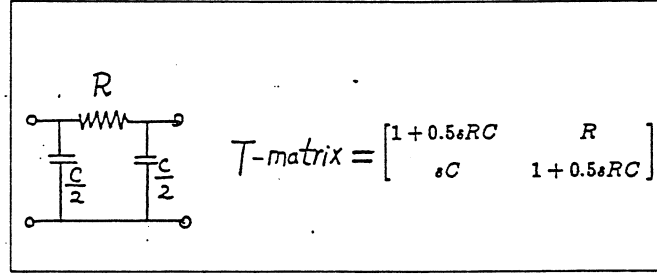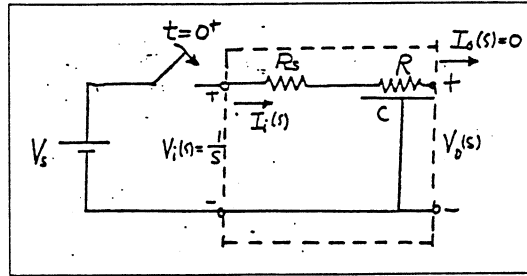
3

Figure 3: A lumped approximation of a URC element.

Figure 4: A URC element driven by a voltage source with resistance R_S.

which suggests a Π network with a resistor of value R on top and two capacitors of value $\frac{C}{2}$ on each side as a model of a URC element, Figure 3.

A URC element driven by a voltage source with source resistance R_S

A URC element with total resistance R and total capacitance C driven by a voltage source having source resistance R_S is shown in Figure 4. The source resistance can be considered as the equivalent channel resistance of the driving transistor.

The two-port network approximation is

$$\begin{bmatrix} V_i(s) \\ I_i(s) \end{bmatrix} = \begin{bmatrix} 1 & R_s \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cosh \Gamma & \frac{\sinh \Gamma}{Z} \\ Z \sinh \Gamma & \cosh \Gamma \end{bmatrix} \begin{bmatrix} V_o(s) \\ I_o(s) \end{bmatrix} \tag{4}$$

With open output and the input driven by a unit step voltage, i.e., $I_o(s) = 0$ and $V_i(s) = \frac{1}{s}$,

$$V_o(s) = \frac{1}{s \left(\cosh \Gamma + \frac{R_S}{Z} \sinh \Gamma \right)} \tag{5}$$

The time domain representation of the step response can be obtained by an inverse Laplace transform:

$$V_o(t) = \mathcal{L}^{-1} \left[\frac{1}{s \left(\cosh \Gamma + \frac{R_S}{Z} \sinh \Gamma \right)} \right] \tag{6}$$
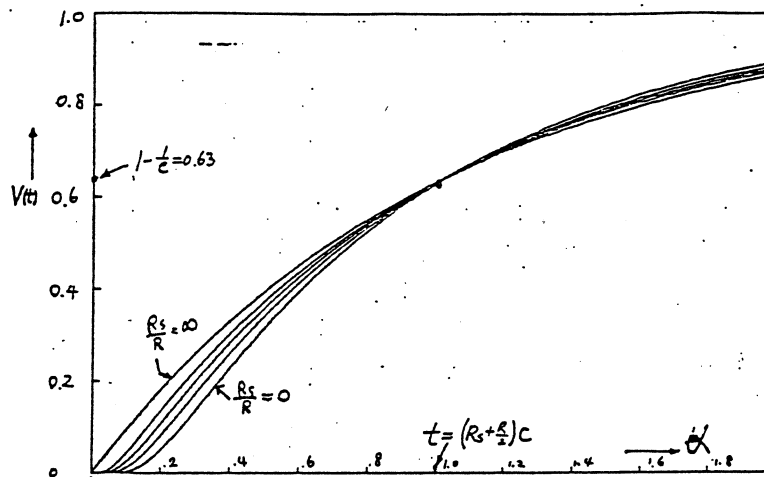
4

Figure 5: The unit step voltage response of a URC element driven by a voltage source with resistance R_S.

R_S/R	$V_o(\sigma = 1)$
∞	0.63213
10	0.63210
2	0.63180
1	0.63127
0.5	0.63044
0.1	0.62930
0	0.62922

Table 2: Unit step responses at $t = (R_S + R/2)C$.

However, there is no simple closed form for this transform. Some numerical results for different ratios of R_S/R are plotted in Figure 5. The time–axis is normalized by $(R_S + R/2)C$. The normalized time is denoted $\sigma = \frac{t}{(R_S+R/2)C}$.

For $R_S/R = \infty$, i.e., $R = 0$ the URC element is simply a pure capacitor and the unit step response is an exponential function with time constant $\tau = R_S C$. Hence at time $t = R_S C$, the output reaches $1 - \frac{1}{e} \approx 0.63$ of its final value.

As the ratio R_S/R decreases, the output response changes accordingly. Let $V_a(t)$ be the output voltage response of the circuit for a given ratio of R_S/R and $V_b(t)$ the response for a larger ratio. In the beginning ($t = 0^+$), $V_b(t)$ is rising slower than $V_a(t)$ and $V_b(t) < V_a(t)$. As time proceeds, the rising rate of $V_b(t)$ increases faster than the rising rate of $V_a(t)$, and eventualy exceeds that rate so that $V_a(t) = V_b(t)$ at some time t. From Figure 5 and Table 2, it is observed that the curves for different ratios R_S/R cross at about $((R_S + R/2)C , 0.63)$. The curves indeed intersect each other within such a small neighborhood of $((R_S + R/2)C , 0.63)$ that it with good acuracy can be considered a
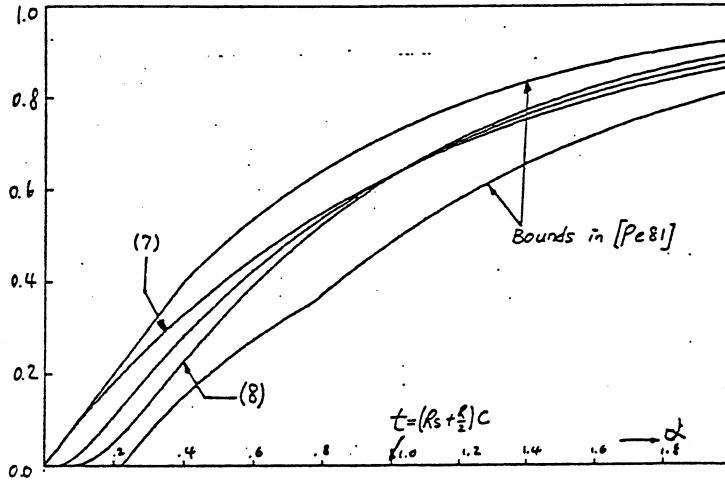
5

Figure 6: The bounds V^0 and V^∞ and the bounds by Penfield for unit step responses of a URC element driven by a voltage source with resistance R_S, $R_s/R = 1$.

point. Hence, the following proposition:

Proposition 1: The time $t = (R_S + R/2)C$ can be used as a time constant τ for the circuit in Figure 4.

Note that the curves in Figure 4 are bounded by two extreme cases: $R_S/R \to \infty$ and $R_S/R \to 0$. The unit step responses for these two cases are given by equations (7) and (8) [6].

$$V^0(t) = 1 - e^{-t/\tau} \tag{7}$$

$$V^\infty(t) = 1 + \frac{4}{\pi} \sum_{n=1}^{\infty} (-1)^n \frac{\exp \frac{(2n-1)^2 \pi^2 t}{4\tau}}{2n - 1} \tag{8}$$

While $t < 0.99(R_S + R/2)C$ the upper bound is $V^0(t)$ and the lower bound is V^∞. For $t < 1.01(R_S + R/2)C$ the upper bound is V^∞ and the lower bound is V^0. The bounds given by equations (7) and (8) are considerably tighter than those given in [5] as is indicated in Figure 6.

A URC Element Driven by a Voltage Source with Source Resistance and Source Capacitance

A URC element with total resistance R and total capacitance C driven by a voltage source with resistance R_S and capacitance C_S is shown in Figure 7.

An approximation of the output voltages can again be obtained by a two-port network analysis. The unit step responses depend only on two parameters: $K_r = R_S/R$ and $K_c = C_S/C$. However, over a large range of resistance and capacitance values the unit step responses are bounded in a way similar to those shown in Figures 5 and 6. Table 3 gives the unit step repsonses at time $t = R_S C_S (R_S + R/2)C$ for different ratios K_r and
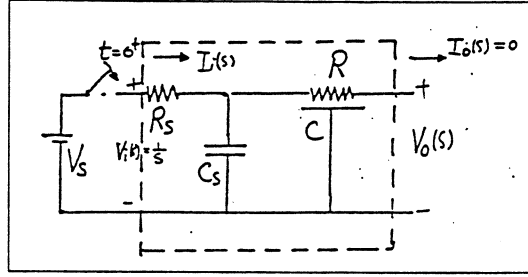
6

Figure 7: A URC element driven by a voltage source with resistance R_S and capacitance C_S.

K_c	K_r						
	01	02	05	1	2	5	10
0.1	0.629	0.629	0.627	0.623	0.615	0.607	0.615
0.2	0.629	0.629	0.626	0.621	0.616	0.620	0.630
0.5	0.630	0.630	0.628	0.623	0.626	0.630	0.631
1	0.631	0.631	0.630	0.630	0.631	0.632	0.632
2	0.632	0.632	0.632	0.632	0.632	0.632	0.632
5	0.632	0.632	0.632	0.632	0.632	0.632	0.632
10	0.632	0.632	0.632	0.632	0.632	0.632	0.632

Table 3: Unit step responses at $T = R_S C_S + (R_S + R/2)C$ of a URC element driven by a voltage source with resistance R_S and capacitance C_S.

K_c.

The fact that all values of $V_o(t)$ for $t = R_S C_S + (R_S + R/2)C$ are close to 0.63 suggests:

Proposition 2: The time $t = R_S C_S + (R_S + R/2)C$ can be used as a time constant τ for the circuit in Figure 7.

Hence, as in the case for a one-dimensional array of lumped circuits, the time constant for a composite circuit is to first order equal to the sum of the time constants of the circuit constituting the array. Next, a delay model for a nMOS PLA is developed.

3 nMOS PLA Delay Estimation

Cherry [2] has given a formula for the delay in a nMOS PLA ignoring wire delay. A formula that takes the wire delay into account is developed below. The wires in the PLA are modeled as URC elements. Consider a typical nMOS PLA as shown in Figure 8. The path that passes the points a,b,d,f,g is a critical path. Its equivalent circuit is also shown

Path section	Transition $0 \rightarrow 1$	Transition $1 \rightarrow 0$
$a \rightarrow b$	$R_t C_1$	$k R_t C_1$
$b \rightarrow d$	$(k R_t + R_2/2) C_2$	$(R_t + R_2/2) C_1$
$d \rightarrow e$	$k R_t (C_3 + C_4)$	$R_t (C_3 + C_4)$
$e \rightarrow f$	$R_4 C_4/2$	$R_4 C_4/2$
$f \rightarrow g$	$k R_t C_5$	$R_t C_5$

Table 4: Delays for the PLA path shown in Figure 8.

in Figure 8.

The poly wire starting at point b that runs in parallel with the wire cd has resistance R_1 and capacitance C_1. The poly wire between points c and d has resistance R_2 and capacitance C_2. The metal wire connecting d and e has capacitance C_3. Its resistance can be ignored. The poly wire between d and f has resistance R_4 and capacitance C_4. The metal wire connecting f and g has capacitance C_5. The output loading capacitance is C_{out}. Let R_t be the channel resistance of a pull-down transistor. The pull-up to pull-down transistor ratio is denoted k, and it is assumed to be 4 in the following calculations.

Based on propositions 1 and 2, the signal delay from the input a to the output h for $0 \rightarrow 1$ and $1 \rightarrow 0$ transitions are to first order accuracy given in Table 4. Wire resistances and capacitances are assumed to be uniformly distributed.

The magnitudes of $T_{0 \rightarrow 1}$ and $T_{1 \rightarrow 0}$ depend on the values of C_1, C_2, C_3, C_4, C_5 and C_{out}. As a rule of thumb, if the number of minterms is greater than twice the number of outputs (i.e., the AND-plane dominates the OR-plane), then $T_{0 \rightarrow 1} > T_{1 \rightarrow 0}$. Otherwise $T_{1 \rightarrow 0} > T_{0 \rightarrow 1}$. The delay from a to h for a 0 to 1 transition at point a is estimated as follows:

$$
\begin{aligned}
T_{0 \rightarrow 1} \approx \ & R_t C_1 + k R_t C_2 && \text{— input buffer delay} \\
& + R_2 C_2/2 && \text{— AND-plane wire delay} \\
& + k R_t (C_3 + C_4) && \text{— AND-plane switching delay} \\
& + R_4 C_4/2 && \text{— OR-plane wire delay} \\
& + R_t C_5 && \text{— OR-plane switching delay} \\
& + k R_t C_{out} && \text{— output buffer delay}
\end{aligned}
\tag{9}
$$

$T_{1 \rightarrow 0}$ can be computed in the same way. Note that $R_2 C_2/2$ and $R_3 C_3/2$ are wire delays due to the resistance of the poly wires.

Figure 9 shows the layout of a basic PLA cell. It is $7 \times 7\lambda^2$. The gates are minimum sized ($2\lambda \times 2\lambda$) and separated by 5λ poly wires. The width is 2λ for both poly and diffusion wires. Metal wires are 3λ wide. The pull-up to pull-down transistor ratio k is 4.
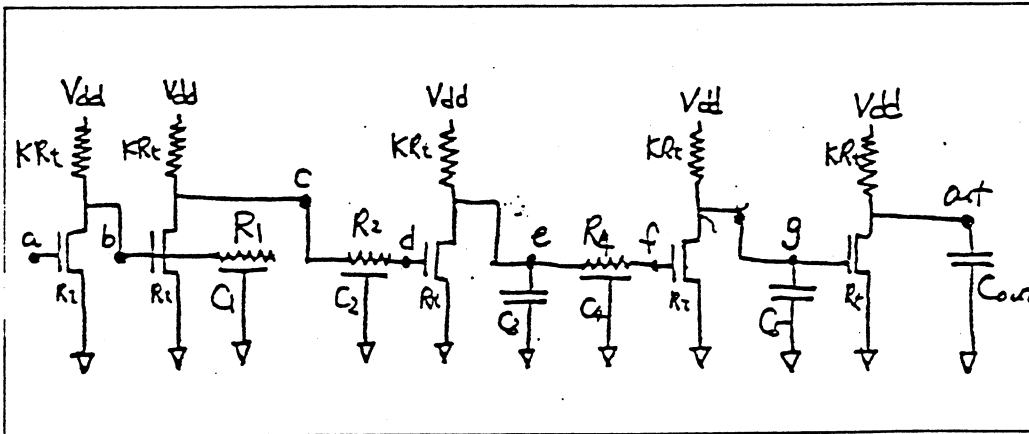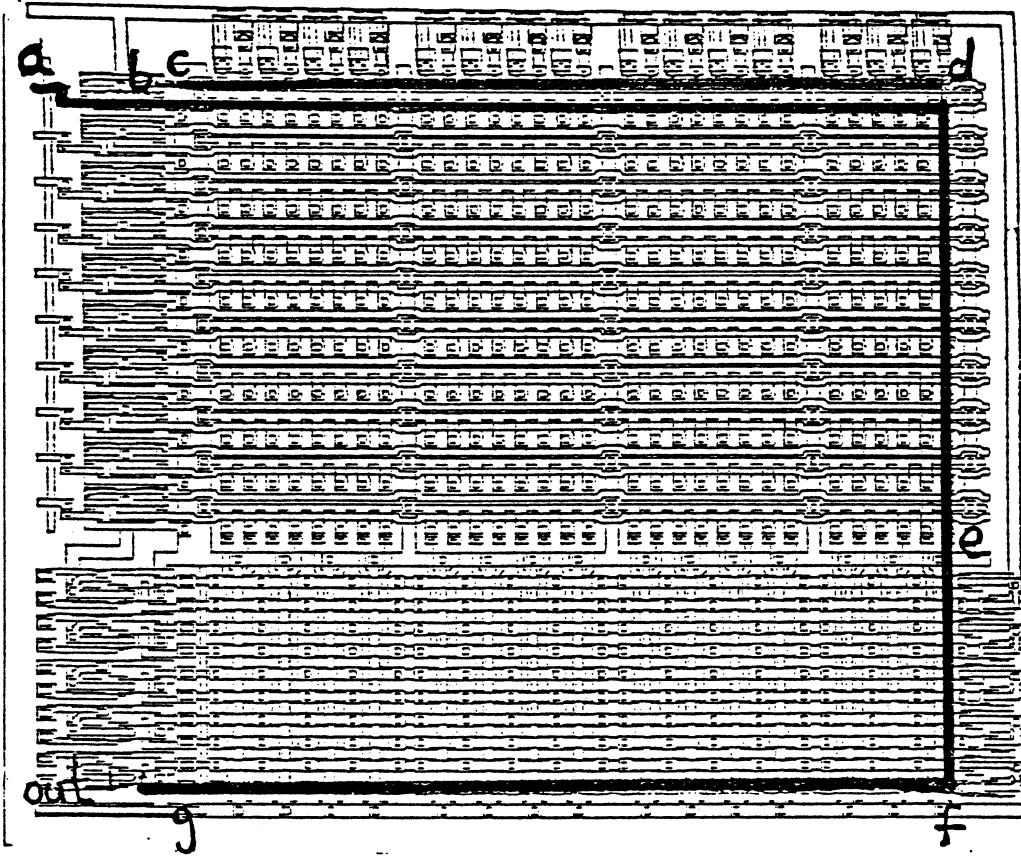
8

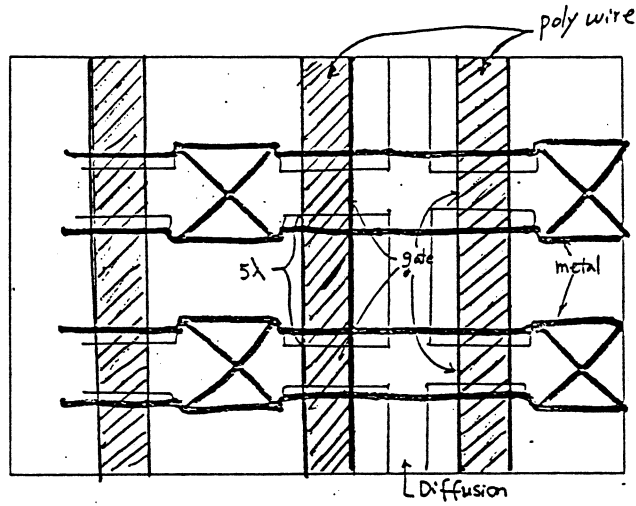Figure 8: A critical path in a typical nMOS PLA and its equivalent circuit.

Figure 9: The layout of a basic PLA cell.

Let r_p denotes the sheet resistance of a poly wire $(\Omega/\square)$ and C_g denotes the capacitance of a minimum gate $(pF/(2\lambda)^2)$. The poly wire capacitance is about $\frac{1}{10}C_g/(2\lambda)^2$, and the metal wire capacitance is about $\frac{1}{15}C_g/(2\lambda)^2$.

Worst Case Delay

Let

$$
\begin{array}{rcl}
P & = & \text{number of minterms} \\
I & = & \text{number of inputs} \\
O & = & \text{number of outputs} \\
f_{out} & = & C_{out}/C_g = \text{fan-out factor}
\end{array}
$$

Then the maximum value of the $C_i's$ and $R_i's$ can be obtained under the assumption that in every cell there is a gate present.

$$
\begin{array}{rcll}
C_1 = C_2 & = & \left(\left(1 + \frac{5}{2}\cdot\frac{1}{10}\right)C_g\right)P & = 1.25 C_g P \\[2mm]
C_3 & = & \left(\frac{7\times3}{2\times2}\cdot\frac{1}{15}C_g\right)2I & = 0.70 C_g I \\[2mm]
C_4 & = & \left(\left(1 + \frac{5}{2}\cdot\frac{1}{10}\right)C_g\right)2O & = 2.50 C_g O \\[2mm]
C_5 & = & \left(\frac{7\times3}{2\times2}\cdot\frac{1}{15}C_g\right)P & = 0.35 C_g P \\[2mm]
R_2 & = & \left(\frac{7}{2}r_p\right)P & = 3.5 r_p P \\[2mm]
R_4 & = & \left(\frac{7}{2}r_p\right)2O & = 7.0 r_p O
\end{array}
\tag{10}
$$

Substituting these values into equation (9) yields the maximum delay,

$$
T_{0\to1} \approx \left(2.2\left(\frac{r_p}{R_t}\right)(P^2 + 4O^2) + 6.5P + 2.8\,I + 10.0\,O + 4\,f_{out}\right)R_t C_g
\tag{11}
$$

10

The value of $R_t C_g$ is the time constant for a minimum size transistor to drive a minimum gate. It is equal to the charge transit time (τ) for a minimum size transistor [7]. The values of r_p and R_t depend on the IC process technology. At 4μm feature size, R_t is of the order $10^4 \Omega/\square$ and r_p is in the range from 10Ω to $100\Omega/\square$. Substituting $r_p=25\ \Omega/\square$ and $R_t = 10^4\ \Omega/\square$ into equation (11) yields

$$T_{0\to1} \approx \left(0.0055\ (P^2 + 4O^2) + 6.5P + 2.8\ I + 10.0\ O + 4\ f_{out}\right)\tau \qquad (12)$$

Equation (12) gives the worst case delay of a static nMOS PLA terms of the number of inputs I, outputs O and minterms P. For a particular design, the number of gates in each column and row is known.

More Accurate Estimation for Known Bit Patterns

Let

$$\rho_1 = \frac{\text{max(the number of gates each poly wire drives in the AND plane)}}{P}$$

$$\rho_2 = \frac{\text{max(the number of gates each poly wire drives in the OR plane)}}{2 \cdot O}$$

Then the relations between the actual wire capacitance (C_i') and resistance (R_i') and the maximum wire capacitance and resistance are:

$$
\begin{aligned}
C_1' &\leq (1.25\rho_1\ P + 0.035(1-\rho_1)\ P)C_g = (0.28 + 0.72\rho_1)C_1 \\
C_2' &\leq (1.25\rho_1\ P + 0.035(1-\rho_1)\ P)C_g = (0.28 + 0.72\rho_1)C_2 \\
C_3' &= C_3 \\
C_4' &\leq (1.25\rho_2\ 2O + 0.035(1-\rho_2)\ 2O)C_g = (0.28 + 0.72\rho_2)C_4 \qquad (13)\\
C_5' &= C_5 \\
R_2' &= R_2 \\
R_4' &= R_4
\end{aligned}
$$

$$(14)$$

where $C_1, C_2, C_3, C_4, C_5, R_2, R_4$ are the maximum wire resistances and capacitances given in equation (10). The approximated delay is

$$
\begin{aligned}
T_{0\to1}' \approx\ & (0.0055((0.28 + 0.72\rho_1)P^2 + 4(0.28 + 0.72\rho_2)O^2) \\
& + 6.5(0.28 + 0.72\rho_1)P + 2.8I + 10.0(0.28 + 0.72\rho_2)O + 4f_{out})\tau \qquad (15)
\end{aligned}
$$

The effects of scaling

The coefficients in equations (12) and (15) are based on the resistance and capacitance values typical at the current state of the technology ($\lambda = 2\mu$m, i.e., $r_p = 25\ \Omega/\square$ and $R_t =$

$10^4 \; \Omega/\square$). As feature sizes are reduced by α, the wire resistances/$\square$ increase by α, while the channel resistance of a transistor essentially remains constant. The capacitance/μm^2 increases by α [7, 8]. To summarize:

$$
\begin{aligned}
x &\rightarrow x/\alpha \\
R_t &\rightarrow R_t \\
C_g &\rightarrow C_g/\alpha \\
\tau &\rightarrow \tau/\alpha \\
r_p &\rightarrow r_p \, \alpha \\
\frac{r_p}{R_t} &\rightarrow \frac{r_p}{R_t}\alpha
\end{aligned}
\tag{16}
$$

Equations (12) and (15) become

$$
T_{0 \rightarrow 1} \approx (0.0055\alpha(P^2 + 4O^2) + 6.5P + 2.8 \, I + 10.0 \, O + 4 \, f_{out})\tau'
\tag{17}
$$

and

$$
\begin{aligned}
T'_{0 \rightarrow 1} \approx \; & (0.0055\alpha(0.28 + 0.72\rho_1)P^2 + 4(0.28 + 0.72\rho_2)O^2) \\
& + 6.5(0.28 + 0.72\rho_1)P + 2.8I + 10.0(0.28 + 0.72\rho_2)O + 4f_{out})\tau'
\end{aligned}
\tag{18}
$$

where $\tau' = \tau/\alpha$.

The wire delay is comparable with the switching delay when

$$
\begin{aligned}
0.0055\alpha \, P^2 &\approx 6.5 \, P &\rightarrow \quad P > 1181/\alpha \\
0.0055\alpha \, O^2 &\approx 10.0 \, O &\rightarrow \quad O > 454/\alpha
\end{aligned}
\tag{19}
$$

At $0.5\mu m$ feature size $\alpha \approx 8$ and equality holds at about 150×60 bits. Hence, unless, e.g., the sheet resistance is reduced, the delay of large nMOS storage structures ($> 10K$ bit) will essentially be proportional to the storage size (bits).

Performance enhancements

By inspecting the delay equations (17) and (18), the delay introduced by a PLA can be reduced in several ways:

1. Use superbuffers, or a sequence of exponentially sized drivers [7] to reduce the switching delay $6.5P + 2.8I + 10.0$.

2. Reorganize the bit patterns in order to minimize $P^2 + 4O^2$, such that the wire delay is equally divided between the AND-plane and the OR-plane.

3. Improve the process technology to reduce r_p and therefore the wire delay (e.g., the sheet resistance of silicide iss only one tenth of that of polysilicon).

4. Limit the driving needs on each clock phase by pipelining the operations (e.g., input buffer, AND-plane, OR-plane, output buffer).

5. Use hierarchical storage architectures as suggested in [7].

4   Conclusions

A simple lumped approximation to the one dimensional diffusion equation as a model for wires in MOS technology is derived. The model is accurate over a large range of resistance and capacitance values. The lumped model is used to derive estimates of delays in a nMOS PLA. The delay is computed in terms of inputs, outputs, and minterms.

The derivation of the lumped model can easily be adapted to other technologies such as CMOS. Moreover, the delay equations can be adopted to estimate the speed of storage structures. For large MOS storage structures the delay is essentially proportional to the storage size, i.e., the square of the chip dimension. The diffusion model is the proper delay model in this case.

References

[1] Gianfranco Bilardi, M. Pracchi, and Frnco P. Preparata. A critique and an appraisal of VLSI models of computation. In *VLSI Systems and Computations*, pages 81–88. Computer Sciences Press, 1981.

[2] Jim Cherry. Pla speed calculations. Technical report, MIT, 1979.

[3] Charles A. Desoer and Ernest S. Kuh. *Basic Circut Theory*. McGraw Hill, 1972.

[4] Mohammed S. Ghausi and John J Kelly. *Introduction to Distributed Parameter Networks with Application to Integrated Circuits*. Holt, Rinehart and Winston, 1968.

[5] Paul Penfield Jr. and Jorge Rubinstein. Signal delay in rc tree network. In *Proceedings of the Second Caltech Conference on VLSI*. Dept of Computer Science, California Institute of Technology, 1981.

[6] Kaufman. Tapered distributed filters. *IRE Trans. Circuit Theory*, CT-9:329–336, 1962.

[7] Carver A. Mead and Lynn Conway. *Introduction to VLSI Systems*. Addison-Wesley, 1980.

[8] Krishna C. Saraswar and Farrokh Mohammadi. The effect of scaling of interconnections on the time delay of VLSI circuits. *IEEE J. on Solid State Circuits*, SC-17:275–280, 1982.

[9] Charles L. Seitz. Self-timed VLSI systems. In *Proc. of the Caltech Conference on Very Large Scale Integration*, pages 345–355. Computer Science, California Institute of Technology, 1979.