Formal Sizing Rules of CMOS Circuits

Daniel Auvergne, Nadine Azemard, Vincent Bonzom, Denis Deschacht, Michel Robert

Formal Sizing Rules of CMOS Circuits

D. AUVERGNE, N. AZEMARD, V. BONZOM, D. DESCHACHT, M.ROBERT

LAMM : Laboratoire d'Automatique et de Microélectronique
de Montpellier (UA D03710 CNRS)
Université de Montpellier II : Pl. E. Bataillon, 34095 MONTPELLIER Cedex 5,
FRANCE (Telex: 490944F)

Abstract

In this paper we present a local strategy for sizing CMOS circuits. We show how the explicit definition of delays can be used to define delay / area optimal sizing rules. Examples are given for sizing irregular inverter arrays, NAND gates and adder cells, starting from an initial electrical netlist and ending with the fully automatically generated layout. Direct comparisons of speed / area performances are given for a linear matrix style layout implementation.

Introduction

High performance circuit design necessitates the mapping of functional specifications into a technology, following predefined performance objectives such as high speed, low power and minimum area. Transistor sizing is one basic technique used to produce designs which meet the performance goals. This is usually carried out using a circuit simulator and critical path analysis tools to adjust iterative transistor sizes to the specifications ; in this case, optimal solutions are not guaranteed [1].

Global mathematical optimization techniques [2,3] can be used to solve this sizing problem, but the extent of the optimization space and the accuracy of the models involved in evaluating delays quickly limit their effectiveness. However, combined with accurate initial solutions, they can be used efficiently at the final stage of any optimal sizing problem [4].

In fact, the lack of an accurate initial estimation of transistor sizes imposes an impractical global strategy giving rise to serious convergence problems in the treatment of the divergence branches of arrays with irregular loading. It has been generally recognized that for an implementation which is nearly optimal with respect to delays, the silicon area required for speed improvement quickly becomes intolerable. It is then of prime importance to define a sizing methodology allowing an efficient trade between delay and area. Moreover, the electrical power available in a switching device being proportional to the width of the corresponding transistor, any minimum area solution guarantees minimum power dissipation for fixed delay constraints. As a result, delays and transistor widths are the parameters to be determined in any optimal sizing problem.

We recently showed [5] that the accurate modelization of delays in CMOS structures can be obtained through closed form equations, allowing the explicit formulation of delays with clear evidence of technological, structural and environmental parameters. As an attempt to find an accurate initial estimation for the optimal sizing of a CMOS data path, we propose in this paper a local strategy allowing the backward processing of data paths using sizing criteria defined from the explicit formulation of delays.

Delay Modeling

It has been shown [6,7,8] that real delay evaluation of general ANDORI can be obtained from a linear combination of the driving (i–1) and the controlled (i) structure's step responses. For example,

$$t_{HL}(i) = \frac{A \cdot t_{LHs}(i-1) + t_{HLs}(i)}{1 + \alpha \cdot A \cdot \frac{t_{LHs}(i-1)}{t_{HLs}(i)}} \qquad (1)$$

$$t_{LH}(i) = \frac{B \cdot t_{HLs}(i-1) + t_{LHs}(i)}{1 + \alpha \cdot B \cdot \frac{t_{HLs}(i-1)}{t_{LHs}(i)}}$$

where A,B, are linearization coefficients, α is an input slope effect correcting term, t_{HLs}, t_{LHs} are the fall, rise step responses of general ANDORI, which can be written in reduced units as follows:

$$\frac{t_{HLs}}{\tau_{ST}} = \frac{(n-1).X_O + Y}{2X_N} + K'.\frac{X_O}{X_{TG}}.\frac{(n-1)+\frac{Y}{2X_O}}{(n-1)+\frac{Y}{X_O}}.\left(\frac{6.(n-1).Y}{X_O} + 2.(n-1)^2 + 1\right) \quad (2)$$

$$\frac{t_{LHs}}{\tau_{ST}} = \frac{\mu_N}{\mu_P}.\frac{(n-1).X_O + Y}{2X_P} + K'.\frac{X_O}{X_{TG}}.\frac{(n-1)+\frac{Y}{2X_O}}{(n-1)+\frac{Y}{X_O}}.\left(\frac{6.(n-1).Y}{X_O} + 2.(n-1)^2 + 1\right)$$

where: – **K'** is a slowly varying technological coefficient, and τ_{ST}, the elementary fall time characteristic of the technology as defined in [7],

– **Y, X_N, X_P, Xo** represent respectively the output load, the N and P transistor gate capacitances, the gate (X_{TG}) and the parasitic capacitances of serial array transistors, normalized with respect to a reference capacitance which can be the minimum capacitance available in the technology or any capacitance used as a reference for delay evaluation (note that with this definition X_N,X_P and Xo depend only on the width of the corresponding transistors),

– **n** represents the number of serial transistors in the array under consideration.

In these equations, the first term on the right hand side represents the step response associated with the switching device with the full load lumped to its output node; it directly gives the step response of the inverter for n = 1; the second term is a propagation term characteristic of the serial array structure, the complete equation allowing the direct evaluation of NAND and NOR gates.

As shown, it is only when real delay is considered (equ.1) that the convexity of the sizing problem of inverters is guaranteed, allowing direct optimization of the structure. As a result, a minimum optimization cell must always be considered in the individual sizing of gates. Such a cell, consisting of the gate under consideration and its driving inverter, is illustrated in fig.1; the size of the driving inverter defines the reference capacitance used in equation (2) as a unit of fan out measurement.

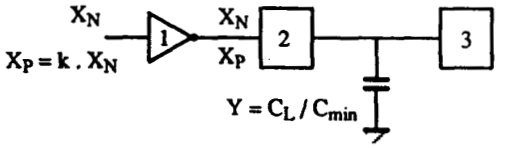


fig.1: minimum optimization cell
1/ is the driving inverter used as a reference
2/ is the gate under study
3/ represents the total active load included in Y.

Sizing Strategy

Using equations (1) and (2), it is then possible to evaluate delays on any data path. Depending on the initial data file two steps have to be considered:

– initial solution sizing. This can be done directly after the technology mapping of a logical synthesis solution; it allows a first estimation of the performance of the data path concerned and can be of great interest regarding performance driven synthesis.

– post layout sizing applied to an electrical description extracted from a layout implementation. This step allows a real evaluation of performances, taking into account layout style and full loading environment.

Initial solution sizing

Let us consider an irregular array of inverters, as shown in fig.2. Optimal sizing of the transistors in this array can be directly obtained by cancelling the derivative of the rise and fall delay equations, evaluated on the complete array with respect to the transistor sizes.

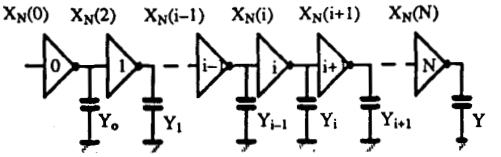


Fig.2: Ex. of irregular array of inverters (there is no correlation with the Yi passive load and $X_N(i+1)$ active ones).

This directly gives the internal configuration ratio of each inverter cell, which depends only on the dissymmetry of the N and P transistors. For symmetrical threshold voltages ($V_{tN} = V_{tp}$) we get: $X_P/X_N = \mu_N/\mu_P$ as generally recognized for the optimal internal sizing ratio, where μ_N and μ_P represent the mobility of N and P transistors respectively. Total delays t_{HL} and t_{LH} are equal for the value of this internal ratio.

The external configuration ratio which depends on the real load of each inverter is then obtained from:

$$\left(1+\frac{\mu_N}{\mu_P}\right)\frac{X_N(i)}{X_N(i-1)} = \frac{Y_i + \left(1+\frac{\mu_N}{\mu_P}\right).X_N(i+1)}{X_N(i)} \quad (3)$$

The solution of this equation defines the sizing of transistors, allowing absolute minimum delay for the structure under consideration. This equation also reflects the full extent of the complexity of the global sizing problem.

Equation (3) represents a system of dependent equations ($X_N(i) \sim X_N(i+1) . X_N(i-1)$) whose complexity is equal to the number of elements to be sized; it has to be solved through successive iterations from initial solutions [2]. In this case, major difficulties occur in solving divergence

branches, due to the capacitive loading effects of path interactions.

As an alternative, we propose a local solution of equation (3) allowing the backward processing of the array (including divergence branches) from the last inverter to the input. In order to do this, it must be noted that equation (3) defines the optimal sizing corresponding to the minimum propagation delay of the structure ; the reduced values $X_N(i) \sim X_N(i+1)$. $X_N(i-1)$, represent the area to be paid in order to reach this minimum. The optimization of each cell with its real load, but with respect to the minimum configuration of fig.1, gives:

$$X_N(i) = \frac{1}{\sqrt{1+\frac{\mu_N}{\mu_P}}} \cdot \sqrt{Y_i \cdot X_N(0)} \qquad (4)$$

where: $X_N(0)$ represents the reduced width of the driving inverter of fig.1 and

Yi is the real load evaluated on the array, including passive and active capacitances.

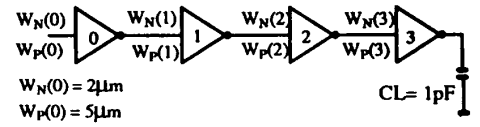
Note that equation (4) gives the sizing conditions of each of the array's inverters corresponding to the optimal delay defined through the minimum cell of fig.1. This constitutes a **local solution**, because each $X_N(i)$ is not determined with respect to the unknown $X_N(i-1)$ driving inverter but with respect to a reference one, $X_N(0)$, identical for all cells. So, if $X_N(0)$ is minimum, this local solution corresponds to a sizing for optimal delay and minimum area (or dynamic power), and allows an accurate initial estimation of the size of the array processed from the end point to the input. Sizing all branches in the same way solves the divergence problem encountered in the system of equation (3).

The resulting minimum delays t_{HL} and t_{LH} associated with each cell are easily obtained from equation (2) as follows:

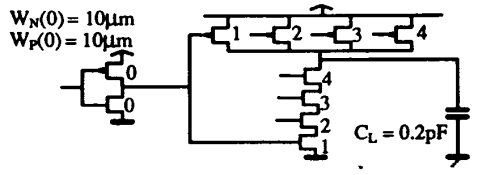
$$\frac{t_{HL}}{\tau_{ST}} = \frac{t_{LH}}{\tau_{ST}} = (1+A) \cdot \sqrt{1+\frac{\mu_N}{\mu_P}} \cdot \sqrt{\frac{Y}{X_N(0)}} \qquad (5)$$

The above equation clearly shows the influence of the size of the driving inverter in defining the real effective load.

As an example of the application of these results, the first table 1 compares the sizing solutions of an inverter array, obtained by MARPLE [3] using mathematical optimization methods, to the values we obtained directly by iterations from local solutions. As shown, equation (4) allows a fast sizing solution of inverter arrays with a power area implementation smaller than [3], at the expense of a weak delay penalty. The second table.1 illustrates the application of this sizing methodology to NAND gates; here too, a local sizing solution allows a minimum area/power implementation without any significant increase of delays. This illustrates the well known result: near global minimum, delay improvement is highly expensive in area. All these results are compared with the values of absolute minimum obtained from the Rosenbrock's minimization algorithm [14].



array of 4 inverters	Mathematical optimization [3] MARPLE case		with the formulation local case (equ.4)	with the formulation global case (equ.3)	absolute minimum [14]	
	$W_N(i)$ (μm)	$W_P(i)$ (μm)	$W_N(i)$ (μm)	$W_N(i)$ (μm)	$W_N(i)$ (μm)	$W_P(i)$ (μm)
inv1	5	7	3.4	6	7	11.5
inv2	12	15	6	18.5	23	26.5
inv3	34	48	18	55	54	89.5
T_{HL} form. T_{LH} form.	0.81 ns 0.96 ns		0.96 ns 0.96 ns	0.86 ns 0.86 ns	0.82 ns 0.82 ns	
T_{HL} ELDO T_{LH} ELDO	0.8 ns 0.95 ns		0.95 ns 0.95 ns	0.85 ns 0.85 ns	0.8ns 0.8ns	
$\sum_{W(i)}$ (μm)	125		105	285	220	



NAND $n_1 = 1$ $n_2 = 4$	Math. Opt. MARPLE case $W_N(i)$ (μm)	Math. Opt. MARPLE case $W_P(i)$ (μm)	with the formulation (6) $W_N(i)$ (μm)	with the formulation (6) $W_P(i)$ (μm)	absolute minimum [14] $W_N(i)$ (μm)	absolute minimum [14] $W_P(i)$ (μm)
1	22	9	14.5	15	15	16.5
2	24	9	14.5	15	15	16.5
4	22	9	14.5	15	15	16.5
4	20	9	14.5	15	15	16.5
T_{HL} form T_{LH} form.	0.4 ns 0.9 ns		0.52 ns 0.53 ns		0.51 ns 0.5 ns	
T_{HL} ELDO T_{LH} ELDO	0.55 ns 0.8 ns		0.6 ns 0.45 ns		0.6 ns 0.45 ns	
$\sum_{W(i)}$ (μm)	145		140		145	

Tables.1: Sizing solutions for an array of inverters and a 4 input NAND gate; comparisons are given between the values obtained from the MARPLE's mathematical optimization method, the explicit solution and the absolute minimum. For the inverter array we compare, local and global sizing methods, for illustration purposes. The technological minimum length is 2μm.

The generalization of these results to ANDORI allows the definition of general local sizing rules by :

$$X_P = \frac{\frac{W_N}{W_P}}{\sqrt{1+\frac{W_N}{W_P}}} \cdot \sqrt{\frac{Y.TH \cdot \left(1+0.4\,(n_1-1)\right)}{\left(1+0.6\,\sqrt[3]{n_2-1}\right)}}$$

$$X_N = \frac{1}{\sqrt{1+\frac{W_N}{W_P}}} \cdot \sqrt{\frac{Y.TH \cdot \left(1+0.7\,(n_2-1)\right)}{\left(1+0.9\,\sqrt{n_1-1}\right)}} \qquad (6)$$

where n_1, n_2 respectively represent the number of serial transistors in a P and N array.

For any structure, these equations allow a fast initial estimation of transistor size corresponding to a minimum area/power implementation for a nearly optimal delay.

Post layout sizing

This is achieved on an electrical file extracted from the layout implementation of the structure under study. This file includes not only the technological mapping and the interconnecting of the initial sizing file, but also the layout and interconnection parasitics, characteristic of the design style.

In order to do this, the inclusion of parasitic capacitances can easily be done on equations (6) by adding :

- the parasitic of the parallel network to the load Y,
- the reduced value of the serial array's parasitic capacitance characteristic to X_N, X_P,

as shown in figure 3 which illustrates the different terms involved in the post layout sizing.

Table 2 illustrates, for various loading conditions, the difference obtained in area and speed for a 6 input NAND gate using minimum size transitors, initial sizing defined directly from the electrical netlist using equations (6) and post layout sizing defined from the netlist extracted from the corresponding layout. Figure 3 illustrates the different terms involved in that calculation. Two delays values are given in the initial sizing solution :

- the delay t_i without parasitic capacitance calculated from the sizing solution obtained before layout,
- the delay t_p which takes the extracted parasitic capacitances into account.

For each solution we give the total area of the cell, the delays and the amount of parasitic capacitances. As expected, the table illustrates that minimum size implementation gives the minimum area cell at the expense of delays. The initial sizing solution improves (t_i). The increase in delays due to the inclusion of parasitic capacitances (t_p) highlights the necessity to perform post layout evaluation.

However it appears possible, using post layout sizing (real configuration evaluation), to minimize delays with a reasonable increase in area. Note the advantage in considering the t_i delay value as an ideal limiting value (no parasitic effect) for performance oriented implementation.

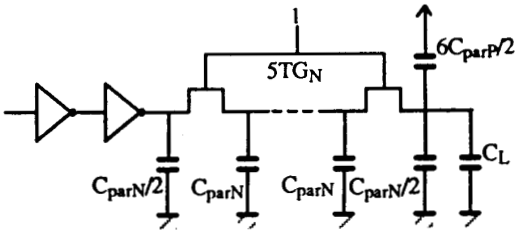


Fig.3 : extracted 6 input NAND gate representation.

load	C_L = 50fF							
cycles	with formulation initial sizing		post layout sizing		absolute min [14] initial sizing		post layout sizing		NAND min	
W_N (µm)	10.5		20.5		11		20.5		3	
W_P (µm)	9		17		9		17		3	
area with SOLO 2000	5050		6400		5050		6400		4300	
(ns)	t_i	t_p			t_i	t_p			t_i	t_p
T_{HL} ELDO	0.6	1.3	1.1		0.6	1.3	1.1		1.2	2.7
T_{LH} ELDO	0.5	0.9	0.9		0.5	0.9	0.9		0.6	1.4

table.2: Comparison of the different sizing solutions for area and delay of a 6 input NAND gate.
t_i : delay without parasitic capacitance
t_p : delay with the extracted parasitic capacitances

Application

For illustration purposes we applied this local sizing strategy to a static adder cell, as given by the electrical scheme and layout of fig. 4.
As expected, the correct sizing of transistors can improve the speed of the structure at the expense of a significant increase in power consumption (ΣW), even if the total area increase is much smaller. The important result is that if the initial inclusion of parasitic results in an important modification of the structure's performances, correct sizing ensures almost conservative speed characteristics. Table 3 compares for the adder cell of fig. 4 the difference in performance between minimum size implementation, initial sizing and post layout sizing.

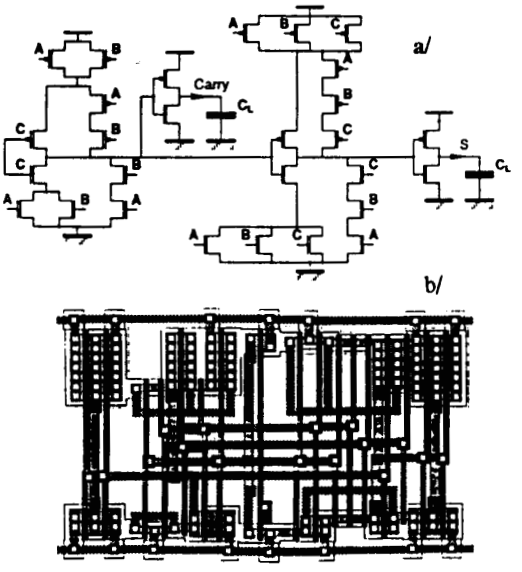


Fig.4 : a/ extracted one bit adder slice representation
b/ layout

load	CL = 1pF				
cycles	with formulation initial sizing	post layout sizing	absolute minimum initial sizing	post layout sizing	area min
	1	2	1	2	
area with SOLO 2000	18600	20500	30900	42500	13700
(ns)	t_i t_p		t_i t_p		t_i t_p
T_{HL} ELDO	1.6 3	2.8	1.4 2.7	2.7	3.4 7
T_{LH} ELDO	1.6 3	2.8	1.4 2.7	2.7	6.8 11

Table.3 : Comparison of the different sizing solutions for area and delay in sizing an one bit adder slice.
t_i : delay without parasitic capacitance
t_p : delay with the extracted parasitic capacitances

As explained in [13] the automatic generation of a layout in linear matrix style results in a compact cell with minimum parasitic capacitances. As a result, final sizing allows a reduction of delay by an important factor (5 times for this example), resulting only in 50% of increase of the cell size.

Conclusion

We have presented a new strategy allowing a nearly optimal time area solution through the local definition of explicit sizing rules. We have defined post layout sizing rules which allow a real characterization of structure performances as well as a direct evaluation of a layout style induced parasitic load. The generalization of these results can be applied to on-line definition of bufferization condition as well as logic resynthesis for delay, technology mapping under constraints and performance driven layout.

As part of a cell compiler, PRINT [10], under development is an automatic procedure (P.SIZE) allowing the direct sizing of transistors from a layout extracted electrical description file.

References

[1] M. HOFMANN, J.K. KIM: *"Delay Optimization of Combinational Static CMOS Logic"*, 24th ACM, IEEE Design Automation Conference, p 125–132, 1987.

[2] C. LEE, M. SOUKUP: *"An Algorithm for CMOS Timing and Area Optimization"*, IEEE J. of Solid States Circuits, vol. SC–19, n°5, p781–787, October 1984.

[3] D. MARPLE: *"Optimal Selection of Transistor Sizes in Digital VLSI Circuits"*, Advanced Research in VLSI, 1987 Standford Conference, p151–167, P.Loslebem ed.

[4] F.W. OBERMEIER, R. KATZ: *"An Electrical Optimizer that consider Physical Layout"*, 25th ACM[IEEE DAC, p453–459, 1988.

[5] D. DESCHACHT, M. ROBERT, D. AUVERGNE: *"Explicit Formulation of delays in CMOS data paths"*, IEEE Journal of Solid States Circuits, vol. 23, n°5, p1257–1264, October 1988.

[6] D. AUVERGNE, D. DESCHACHT, M. ROBERT: *"Explicit Formulation of Delays in CMOS VLSI"*, Electronics Letters, vol. 23, n°14, p741–742, July 1987.

[7] D. AUVERGNE, N. AZEMARD, D. DESCHACHT, M. ROBERT: *"Evaluation dynamique et optimisation des structures CMOS et VLSI"*, TSI, VOL.8, n°6, Decembre 1989.

[8] D. DESCHACHT, P. PINEDE, M. ROBERT, D. AUVERGNE: *"PATH–RUNNER: an accurate and fast timing analyser"*, EDAC, Scotland, 12–15 March 1990.

[9] J.M. SHYU, A. SANGIOVANNI–VINCENTELLI, J.P. FISHBURN, A.E. DUNLOP: *"Optimization based transistor sizing"*, IEEE J.of Solid State Circuits, Vol.23, No2, pp 400–408, April 1988.

[10] M. ROBERT, D. DESCHACHT, G. CATHEBRAS, S. PRAVOSSOUDOVITCH, D. AUVERGNE: *"PRINT methodology: a compilation approach for cell library generation"*, 'ISCAS'88, Vol.2, p965–968, FINLAND, 1988.

[11] N. AZEMARD, V. BONZOM: *"CMOS circuit speed optimization based on closed form equations"*, ISMEC 90, Yugoslavia, Zagreb, 21–24 May 1990.

[12] D. AUVERGNE, N. AZEMARD, D. DESCHACHT, M. ROBERT: *"Input waveform slope effects in CMOS delays"*, IEEE J. of solid state circuits, to be published.

[13] M. ROBERT, G. CATHEBRAS, J. TRAUCHESSEC, D. DESCHACHT, D. AUVERGNE: *"Linear matrix oriented optimal automatic layout of digital CMOS cells"*, EDAC. Holland, Amsterdam, February 1991.

[14] D.M. HIMMELBLAU, Applied Nonlinear programming, *"Rosenbrock's method"*, chp.4, Mc Graw–Hill Ed., 1972.

[15] B. HENNION, P. SENN: *"ELDO: a new third generation circuit simulation using the one step Relaxation Method"*, Proceeding of ISCAS, KYOTO, JAPAN, IEEE, 1985.