Definition of P/N Width Ratio for CMOS Standard Cell Library

A. Verle, P. Maurine, N. Azémard, D. Auvergne

LIRMM, UMR CNRS/Université de Montpellier II, (C5506),161 rue Ada, 34392 Montpellier, France
pmaurine, azemard, auvergne@lirmm.fr

Abstract-The efficiency of cell-based design synthesis of high performance circuit is strongly dependent on the content of the library. Great effort has been given in the design of libraries, to define the optimal selection of the logic gate drive strength. But few justifications are available to determine the P/N width ratio of each cell.

In this paper we use an extension of the logical effort model to characterize the dissymmetry of gate delay and define the best P/N width ratio allowing a path minimum area implementation under delay constraint. This delay model explicitly represents the sensitivity of delay to gate structure and P/N width ratio. Application is given on a 0.18µm process on different logic path implementations.

I. INTRODUCTION

The relative merits of different cell libraries can be evaluated in terms of area/power necessary in achieving a particular delay for implementing a specific circuit. For that important effort has been devoted to supply high-performance standard cell libraries. Work has been devoted to define the optimal content of the library [1] as well as for determining the best selection of drives [2,3]. A fluid cell approach is emerging [4], in which a cell generation tool is used to create a discrete library with 10 to 25 drive strengths and 1 to 4 different P/N width transistor ratios.

The question arise to know if it is possible to define an optimal value for the gate transistor ratio allowing to implement a CMOS logic circuit with the best delay/power trade-off.

Recent work has been published [5] on P/N transistor width selection, based on an asymmetric implementation of the gate rise and fall delays.

Considering that, for an array of inverters, the minimum delay can be obtained using asymmetric edges, [5] use a first order gate delay model to determine an optimal transistor width ratio for each gate. They minimize the average of the rising and falling delays to obtain an optimal solution in which they in over-size the Nand gates and under-size the NOR gates, with respect to that of an inverter.

In fact on a logical path a separate consideration of the falling and rising edges must be given. In this case it can be shown that, considering only the critical edge, the fastest solution is obtained for an inverter implementation with balanced fall and rise delays. Moreover, for gates great attention must be given to the modeling of the transistor serial array current. On a non critical path the minimum gate area solution can be obtained with unbalanced edges, resulting from identical equal N and P transistor sizes. We want to demonstrate, here, that on a critical path performance constraint satisfaction may result in transistor over-sizing and extra power consuming, if no care is given to the balancing of the gate rise and fall delays.

In this paper we propose a method for determining the P/N transistor width ratio for implementing high performance library cells.

The method is based on a delay model developed around the logical effort [6], but with an explicit consideration of the input ramp and Miller effects. This model is presented in part II. In part III, this model is used to determine the optimal value of the P/N transistor width ratio. Application to different logical paths is given in part IV, where we compare, at constant delay value, the area resulting from different sizing alternatives to that resulting from the proposed solution.

II. DELAY MODEL

The delay of a CMOS logic gate is load, gate size and input slew dependent. Following the value of the gate internal P/N width ratio, consideration of different rising and falling edges must be given.

Considering the I/0 coupling [7], the input slope effect can be introduced in the propagation delay as:

$$t_{HL}(i) = \frac{v_{TN}}{2}\tau_{INLH}(i-1) + \left(1+\frac{2C_M}{C_{LTOT}}\right)\frac{\tau_{outHL(i)}}{2}$$
$$t_{LH}(i) = \frac{v_{TP}}{2}\tau_{INHL}(i-1) + \left(1+\frac{2C_M}{C_{LTOT}}\right)\frac{\tau_{outLH}(i)}{2} \quad (1)$$

where for each element the delay is evaluated as the time interval between the input and output waveforms, evaluated at mid-supply voltage value.

$\tau_{INHL,LH}$ ($\tau_{out,HL,LH}$) is the transition time of the signal applied to the input (generated at the output). It is evaluated as the time duration of a linear approximation of the output waveform of the controlling (switching) structure. Indexes (i), (i-1) refers to the location of the cell in the array. $v_{TN,P}$ represent the reduced values of the N(P) transistor threshold voltage with respect to the supply voltage V_{DD}. $C_{LTOT}=C_L+C_M+C_{DIFF}$ includes the coupling capacitance C_M between the gate input and output nodes, the load C_L at the output node and the diffusion capacitance C_{DIFF}. As shown, for each edge, the delay expression is a linear combination of the output

transition time of the controlling and the switching gate.

Let us first consider a simple situation in which each switching gate is equivalent to a constant current generator. Following [8] the gate output transition time value can be obtained from

$$\tau_{outHL} = \tau \cdot (1+k) \cdot DW_{HL} \cdot \frac{C_{LTOT}}{C_{IN}}$$
$$\tau_{outLH} = \tau \cdot \frac{(1+k)}{k} \cdot R \cdot DW_{LH} \cdot \frac{C_{LTOT}}{C_{IN}} \quad (2)$$

where τ is a unit delay characterizing the process. The configuration ratio $k = C_P/C_N$, represents the P/N transistor width ratio. R is an equivalent mobility ratio between N and P transistors, that can directly be calibrated on an inverter.

As defined in [8], $DW_{HL,LH}$ is the current reduction factor in the gate series-connected transistor array, evaluated as the ratio of maximum current available in an inverter and a gate of identical size. $C_{IN} = C_N+C_P$, represents the active gate input capacitance. Considering that the input-to-output coupling capacitance and the parasitic capacitance are input gate size dependent, the effective loading factor can be defined as

$$\frac{C_{LTOT}}{C_{IN}} = A + \frac{C_L}{C_{IN}}. \quad (3)$$

Where A is gate input capacitance independent for a cell sizing at 2 or 3 times the minimum value allowed by the process.

Eq.2 can then be identified to the logical effort model delay expression [6]

$$\tau_{outHL,LH} = \tau \cdot (p_{HL,LH} + g_{HL,LH} \cdot h) \quad (4)$$

The comparison of eq.2 and 4 gives an easy way to explicit the different parameters of the logical effort model for a general gate configuration. Here **p** is the gate input independent parasitic delay contribution, which is edge and configuration ratio dependent.

h is the electrical effort defined by the ratio of output load to gate input capacitance and

$$g_{HL} = (1+k) \cdot DW_{HL}$$
$$g_{LH} = R \cdot \frac{1+k}{k} \cdot DW_{LH} \quad (5)$$

represent the detailed expression of the logical efforts. As shown, the value of this parameter is a direct indicator of the symmetry of the response and the efficiency of a logic structure.

Note here that (4) identifies a transition time and, in general implementations, is not accurate enough to represent a switching delay as given in (1) and almost for evaluating the delay on a logic path.

III. P/N CONFIGURATION RATIO DETERMINATION

Considering the logic path delay, the contribution of each element is due to the loading term including I/O coupling effect (second term on the right hand part of (1)) and the input slope effect on the next stage. The total stage contribution is thus given by

$$t_{HL}(G(i)) = \left(1 + \frac{2C_M}{C_{TOT}} + v_{TP}\right) \frac{\tau_{outHL}(i)}{2}$$
$$t_{LH}(G(i)) = \left(1 + \frac{2C_M}{C_{TOT}} + v_{TN}\right) \frac{\tau_{outLH}(i)}{2} \quad (6)$$

We represent in Fig.1 the variation of the delay of different cells (inverter, two input, Nand and Nor gate), versus the value of their internal configuration ratio value, $k = C_P/C_N$.

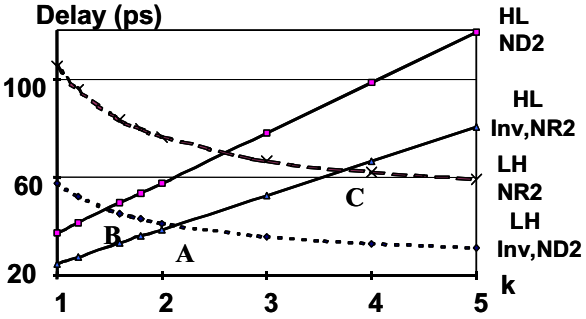

Fig.1. Variation of the gate delay (0.18µm) with the P/N width ratio value.

The intersect points A, B, C correspond to the value of k balancing the edges of the inverter, Nand and Nor gates, respectively. As shown, for each gate, the edge symmetry is obtained for a well defined value of k. Not fulfilling this condition increases the edge asymmetry. We can note that the delay average of both edges corresponds nearly to the value of delay obtained for the symmetrical case. This may explain the claim of [5] that asymmetric rise and fall delay will result in a minimum delay implementation on an array of inverters.

In fact in an array of gates the delay is given by the contribution of the successive edges. As given in (1) the delay of each gate is modified by the transition time of its preceding gate and in the same way modifies the delay (from its output transition time contribution) of its loading gates. In this case it is easy to understand from Fig.1 that the symmetric implementation will always have the shortest delay or under delay constraint, will result in the smallest area/power implementation.

Let us now determine the k value balancing the rise and fall delay gate contribution to a logic path. Developing eq.6 we obtain

$$t_{HL}(G(i)) = \left(1 + \frac{2\alpha_P C_P}{C_{LTOT}} + v_{TP}\right) \cdot \left(p_{HL}(i) + g_{HL}.h(i)\right)$$
$$t_{LH}(G(i)) = \left(1 + \frac{2\alpha_N C_N}{C_{LTOT}} + v_{TN}\right) \cdot \left(p_{LH}(i) + g_{LH}.h(i)\right) \quad (7)$$

where $\alpha_{P,N}=C_M/C_{P,N}$ for falling or rising edges respectively, are the Miller coefficients [9] (a #0.5 or may be calibrated on the process). Balancing the delay contribution of the gate (i), imposes equal contribution of each edge. In the usual case, the expression of the configuration ratio value satisfying

this condition is quite complicated and will be discussed later.

To get a first idea, let us consider the asymptotic value of (7), corresponding to a large value C_{LTOT}. In this case the I/O coupling contribution becomes negligible. Searching for the value of k that balances the falling and rising delays, (7) gives

$$k_{asympt.} = R.\frac{DW_{LH}}{DW_{HL}} \cdot \frac{1+v_{TN}}{1+v_{TP}} \quad (8)$$

where the value of R and $DW_{HL,LH}$, characterize each structure [10]. In Table I we give the resulting values of k_{asympt}, calculated from (8), for a library developed in a 0.18µm process (R =2.25, v_{TN}=0.39 v_{TP}=0.36). The DW values are directly calibrated on the process.

Table I

	DW_{HL}	DW_{LH}	k_{asympt}
INV	1	1	2.3
ND2	1.55	1	1.5
ND3	2.1	1	1.1
ND4	2.6	1	.88
NR2	1	1.9	4.4
NR3	1	2.9	6.6
NR4	1	3.6	8.3
AOI21	1.55	1.9	2.9
AOI31	2.1	1.9	2.1
OAI21	1.55	1.9	2.9
OAI31	1.55	2.9	4.4

The values given in this table are evaluated for the critical input. For Nand and Nor gates this has been shown to be the input controlling the transistor near the output node. For complex gates the most critical branch must be considered. In any way the edge symmetry will only be obtained for the critical switching condition, that must be optimized on a critical path.

Considering now the input–to-output coupling effect we obtain, from (7), the general solution as

$$k_{sym} = \frac{1}{2}\frac{k_{asymp.}-1}{1+\frac{\alpha_P}{F'_0}\cdot\frac{3+v_{TP}}{1+v_{TP}}} + \sqrt{\frac{1}{4}\left(\frac{k_{asymp.}-1}{1+\frac{\alpha_P}{F'_0}\cdot\frac{3+v_{TP}}{1+v_{TP}}}\right)^2 + k_{asymp.}\frac{1+\frac{\alpha_N}{F'_0}\cdot\frac{3+v_{TN}}{1+v_{TN}}}{1+\frac{\alpha_P}{F'_0}\cdot\frac{3+v_{TP}}{1+v_{TP}}}} \quad (9)$$

where $F'_0 = A+C_L/C_{IN}= A + F_0$.

We can easily verified that k_{asymp} is the asymptotic value of this expression, for large F'_0 value. As shown, in the general case and for small values of the loading factor, the P/N with ratio, balancing the rising and falling gate delays, is load dependent. This is a direct result of the input-to-output coupling effect.

The curves in Fig.2 represent, for the simple gates of Table I, the simulated (Spice level 49 on a 0.18µm process) variation with the loading factor F_0, of the k value balancing the rise and fall gate delays.

As shown, for the interval of F_0 values ranging between 2 to 6, which corresponds to nearly optimal design conditions for performance, the P/N width ratio value balancing the delay edges, is constant for Nand and inverters, but is load dependent for Nor gates.

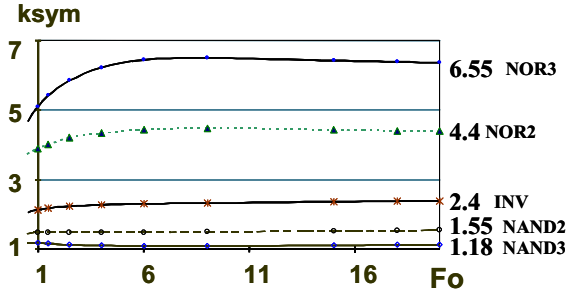

Fig.2: P/N width ratio sensitivity to the load.

In Table II we compare the values of k_{sym}, calculated from (9), to the simulated values (Fig.2). As shown the agreement between simulated and calculated values, obtained for different values of the loading factor, is quite satisfactory.

Table II

F_0	1	2	3	5
Simul. Inv	**2.1**	**2.2**	**2.27**	**2.3**
Calcul.	**1.9**	**2.1**	**2.14**	**2.21**
Simul. ND2	**1.48**	**1.48**	**1.48**	**1.48**
Calcul.	**1.37**	**1.42**	**1.44**	**1.46**
Simul. ND3	**1.18**	**1.15**	**1.13**	**1.11**
Calcul.	**1.01**	**1.11**	**1.11**	**1.12**
Simul. NR2	**3.89**	**4.13**	**4.28**	**4.42**
Calcul.	**3.26**	**3.68**	**3.9**	**4.1**
Simul. NR3	**5.11**	**5.7**	**6**	**6.36**
Calcul.	**4.54**	**5.3**	**5.6**	**6**

4. EXPERIMENTAL VALIDATION

We have determined in the preceding part the value k of the P/N width ratio for balancing the gate delay contribution on a logical path. As a result (Table II) it may appear that imposing a gate sizing for delay edge balancing results in a gate over-sizing. At first sight, an implementation with unbalanced edges, using k=1 or the value recommended in [5], appears to be less area consuming. We demonstrate, in this part, that under a delay constraint, a logic path implemented with the k values given in Table II, necessitates less area than an unbalanced edge implementation.

In Fig.3 we represent the simulated (Hspice) variation of the delay of a logic path with respect to the sum of

the gate transistor widths of each implementation, under the following protocol.

- We consider 3 P/N width ratio values, k=1, k from [5], corresponding to a gate implementation with asymmetric fall and rise delays, and the k value from (9) for symmetrical edge implementation.
- The gate input size value is obtained, for each implementation, by imposing a constant value to the derivative of the total delay equation with respect to each input gate capacitance. Varying this constant from a negative value to zero allows to explore the design space and to reach the minimum delay achievable on the path [11].

The variations corresponding to k = 1 and k from [5], correspond to the critical edge of the path delay (the delay imbalance is 23% and 37% respectively). The variation corresponding to k, from (9), represents the variation of the rise and fall path delays (5% of delay imbalance).

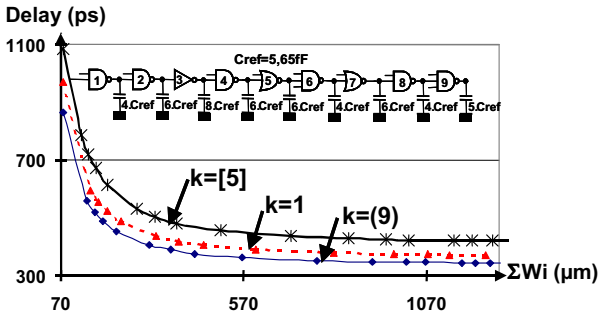

Fig. 3. Logic path delay sensitivity to the gate P/N width ratio value (0.18µm process).

As shown, the gate implementation with balanced rise and fall delays results in the less area consuming implementation. For a weak delay constraint the area difference between the implementations is not very important. However for a more severe constraint, not only the recommended P/N width ratio value (9) results in a reduced area implementation, but may give the possibility to obtain a smaller delay value than allowed by the other methods.

In Table III we detail the results of the comparison of the different sizing alternatives (k from [5], k=1, k from (9)) on benchmark circuits of various complexity. These values are just given for completeness. We compare the values of the minimum allowed delay, the sum of the transistor widths and the delay symmetry of the rising and falling edges of the different implementations. We also compare the area necessary to satisfy different delay constraints, defined with respect to the minimum delay obtained with the implementation obtained from (9). (xxxx) corresponds to a non satisfaction of the delay constraint, ie a configuration where the delay constraint value is smaller than the minimum delay allowed by the corresponding implementation.

The maximum delay value has been obtained by imposing the input capacitance of all the gates to be at the minimum allowed value, C_{REF} = 5.65fF.

The minimum delay value has been obtained in the same way than for Fig.3, by cancelling the derivative of the total delay equation of each path with respect to each gate input capacitance.

As shown, in all the considered situations, the P/N sizing method, we propose, allows a fastest and smaller area implementation. This is particularly verified for strong delay constraints.

Considering the very low sizing sensitivity of the delay near the minimum we have to note here that the area corresponding to the minimum delay is far to be of practical use.

V. Conclusion

Using an extension of the logical effort model, we have developed in this paper a method for determining the best P/N width ratio of gates in a standard cell library. We have defined the explicit expression of the P/N width ratio, which is shown to be loading factor and structure dependent.

Validations have been obtained, with respect to Spice simulations on a 0.18µm process, by comparing, on different benchmarks, simulated values of the delay using different P/N width ratio strategies. We have obtained clear evidence that imposing on a logic path equal rise and fall gate delay, results in a high performance implementation for the best area- delay trade-off.

References

[1] K. Scott & Al, "Improving cell libraries for synthesis", IEEE Custom Integrated Circuit Conference, pp. 7.2.1.-7.2.4, 1994.

[2] O. Coudert, "Gate Sizing: a General Purpose Optimization Approach", European Design & Test Conf., pp214-231, 1996.

[3] Y. Jiang, & Al, "Interleaving buffer insertion and transistor sizing into a single optimization" IEEE Trans. On VLSI Systems, vol.6, n°4, pp.625-633, 1998.

[4] G. Northrop & Al, "A semi-custom design flow in high performance microprocessor design", Proc. of Design Automation Conference, pp.426-431,2001.

[5] D.S. Kung & Al, "Optimal width ratio selection for standard cell libraries", IEEE conf. On Comp.-Aided Design, pp. 178-184, 1999.

[6] I. Sutherland, & Al, "Logical Effort: Designing Fast CMOS Circuits", Morgan Kaufmann Publishers, INC., San Francisco, California, 1999.

[7] K. O. Jeppson, "Modeling the influence of the transistor gain ratio and the input-to-output coupling capacitance on the CMOS inverter delay", IEEE J. Solid State Circuits, vol.29, pp.646-654, 1994.

[8] P. Maurine, M. Rezzoug, N. Azémard, D. Auvergne "Transition time modeling in deep submicron CMOS" IEEE trans. on Computer Aided Design, Vol.21, n°11, pp.1352-1363, nov. 2002.

[9] J. Meyer "Semiconductor Device Modeling for CAD" Ch. 5, Herskowitz and Schilling ed., Mc Graw Hill, 1972.

[10] Lasbouygues[1], J. Schindler[2], S. Engels[2], P. Maurine[3], N. Azémard[3], D. Auvergne[3], " Continuous representation of the performance of a CMOS library" pp. 595-598, ESSCIRC 03, Lisbon.

[11] A. Verle, X. Michel, P. Maurine, N. Azémard, D. Auvergne "Delay bound based CMOS gate sizing technique" ISCAS 04, pp.V-189-192, Vancouver, Canada may 2004.

Table III

								Delay constraint 1.3 t_{min}		Delay constraint 1.5 t_{min}		Delay constraint 2 t_{min}	
	P/N width ratio k	Delay Max. (ps)	ΣW (µm)	Edge Dissy. %	Delay Min. (ps)	ΣW (µm)	Edge Dissy. %	ΣW (µm)	Edge Dissy. %	ΣW (µm)	Edge Dissy. %	ΣW (µm)	Edge Dissy. %
9 gates	[5]	1083	76	38	421	1294	27	541	29	302	32	158	35
	1	970	76	30	371	1234	18	302	20	198	22	132	25
	(9)	863	76	3	345	1275	5	232	4	165	3	111	1.5
11 gates	[5]	1520	92	49	539	2482	43	562	48	329	54	197	43
	1	1442	92	47	495	2570	31	439	39	293	39	182	39
	(9)	757	92	7	472	1466	15	225	13	138	9	92	7
15 gates	[5]	2722	128	51	769	8988	36	xxx	xxx	1250	41	474	41
	1	2483	128	46	696	8336	27	1568	30	784	30	371	30
	(9)	2033	128	4	586	6405	4	690	4	485	4	287	2
21 gates	[5]	3555	188	51	979	16127	47	2380	47	1150	47	530	44
	1	3186	188	43	852	13000	28	1210	32	800	37	450	36
	(9)	2347	188	7	884	14776	15	1030	14	650	12	370	24
31 gates	[5]	5078	276	46	1222	42238	38	xxx	xxx	4100	40	1420	40
	1	4591	276	40	1078	42412	26	4690	26	2320	27	1040	28
	(9)	3903	276	5	941	37043	7	2280	5	1520	6	780	3