Design of Variable Input Delay Gates for Low Dynamic Power Circuits

Tezaswi Raja[1], Vishwani Agrawal[2], and Michael Bushnell[3]

[1] Transmeta Corp., Santa Clara, CA. traja@transmeta.com
[2] Auburn University, Aubrun, AL. vagrawal@eng.auburn.edu
[3] Rutgers University, Piscataway, NJ. bushnell@caip.rutgers.edu

Abstract. *The time taken for a CMOS logic gate output to change after one or more inputs have changed is called the* output delay *of the gate. A conventional multi-input CMOS gate is designed to have the same input to output delay irrespective of which input caused the output to change. A gate which can offer different delays for different input-output paths through it, is known as a variable input delay(VID) gate and the maximum difference in delay between any two paths through the same gate is known as "u_b". These gates can be used for minimizing the active power of a digital CMOS circuit using a previosuly described technique called variable input delay(VID) logic. This previous publication proposed three different designs for implementating the VID gate.*
In this paper, we describe a technique for transistor sizing of these three flavors of the VID gate for a given delay requirement. We also describe techniques for calculating the u_b of each flavor. We outline an algorithm for quick determination of the transistor sizes for a gate for a given load capacitance.

1 Introduction

We first describe the prior work and motivation for this work in this section. We then describe the sizing procedures and algorithms in the following sections followed by contributions and conclusion.

1.1 Prior Work

Dynamic power consumed in the normal operation of a circuit consists of essential power and also *glitch power*. Glitches are spurious transitions caused by imbalance in arrival times of signals at the input of a gate. Techniques such as *delay balancing, hazard filtering, transistor sizing, gate sizing* and *linear programming* have been proposed for eliminating glitches [1–16]. For further reference the reader is directed to recent books and articles [17–24]. Our focus in this paper is a recent technique known as *variable input delay logic* [12, 13]. Raja *et al.* described a technique for reducing glitches using special gates are known as *variable input delay(VID) gates* where the delay through any input-output path can be manipulated without affecting the delays of the other paths upto

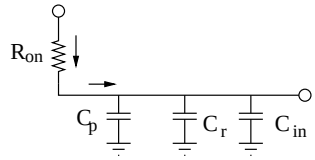

Fig. 1. The RC components along the charging path.

a certain limit. This limit is known as the *differential delay upper bound* or u_b. This u_b is determined by the technology in which the circuit is implemented and is needed for finding the optimal solution to the linear program.

1.2 Motivation

Raja *et al.* describe three new ways of implementing the VID gate *viz. Capacitance manipulation, nMOS transistor insertion* and *CMOS transistor insertion* [12,13]. Each of these gate designs can be used for efficient manipulation of input delay without altering the output delay of the gate. However, the paper has the following shortcomings.

- How are the transistor sizes determined from the delay assignment?
- How is the u_b calculated for every gate type?
- What is the algorithm for finding the right sizes and what are the trade-offs?

These are the questions we try to answer in this paper.

1.3 Components of RC Gate delay

Gate Delay is the time taken for the output signal at the output of the gate to reach 50% of *Vdd* after the signal at the input of the gate has reached 50% of *Vdd* [25,26].

Consider the path shown in Figure 1. The delay of a gate is a function of the on resistance R_{on}(ignoring saturation effects) and the load capacitance C_L. The load capacitance is given by:

$$C_L = C_p \ + \ C_r \ + \ C_{in} \tag{1}$$

where C_p is the parasitic capacitance due to the on transistor, C_r is the routing capacitance of the path and C_{in} is the input capacitance of the fanout transistors. C_{in} is the major component of C_L. C_r and C_p are non-controllable and hence, we ignore them in the current discussion. The delay of the path during a signal transition is given by:

$$Delay = R_{on} \times C_L \tag{2}$$

The delay can be manipulated by changing the C_L or the R_{on} by sizing the transistor accordingly. This alters the gate delay along all paths equally. This is called *conventional gate sizing*.

For VID logic, we describe the gate delay as the sum of *input delay* and *output delay* through the gate. *Output delay* is the common delay component of the gate no matter which input has caused the transition. *Input delay* is the delay component on input 1 that is present only on the input-output path through input 1 of the gate. Both input and output delays should be independent. Clearly, conventional gate sizing cannot be used for designing a VID gate. In this paper, we describe the *variable input delay gate sizing* for the VID gates proposed by Raja *et al.* [12, 13].

2 Gate Design by Input Capacitance Manipulation

The overall gate delay is given by Equation 2. In the new gate design we need to manipulate the input delay of the gate without affecting the output delay too much. Substituting Equation 1 into Equation 2, we get:

$$Delay = R_{on} \times (C_p \ + \ C_r) \ + \ R_{on} \times C_{in} \tag{3}$$

$$= Output\ Delay \ + \ Input\ Delay \tag{4}$$

From the above analysis we separate the input and output delays of the gate. The output delay depends on C_p and C_r, which are unalterable. The input delay is a function of R_{on} and C_{in} of the transistor pair. Thus, *input delay* of an input X can be changed by increasing the C_{in} offered by the transistor pair connected to X. Note that this does not alter the input delays of the other inputs of the gate(this is not always true as shown in Sec 2.2).

2.1 Calculation of u_b

The delay of the transistor pair can be calculated by using Equation 3. The input capacitance of a transistor pair is given by:

$$C_{in} = W \times L \times C_{ox} \tag{5}$$

where W is the transistor width, L is the transistor length and C_{ox} is the oxide capacitance per unit area, which is technology and process dependant. The range of manipulation for C_{in} is limited by the range of W and L of the transistors allowed. The range of dimensions for digital design, in any technology, is governed by second-order effects, such as *channel length modulation*, *threshold voltage variation*, *standard cell height* etc [25, 26]. We have chosen the limit of the transistor length for 0.25μ technology as $3\mu m$, which is determined by the standard cell height. The minimum gate length in the same technology is $0.3\mu m$. Hence, the maximum difference in input capacitance is $2.7 \times C_{ox}$. The maximum

differential delay d_{dif} and the minimum differential delay d_{min} obtainable in the technology can thus be:

$$\textit{Maximum Differential Delay } d_{dif} = R_{on} \times 2.7 \times C_{ox}$$
$$\textit{Minimum Gate Delay } d_{min} = R_{on} \times 0.3 \times C_{ox}$$

Thus, the gate differential delay upper bound u_b is given by:

$$u_b = \frac{d_{dif}}{d_{min}} = \frac{R_{on} \times 2.7 \times C_{ox}}{R_{on} \times 0.3 \times C_{ox}} = 9$$

Thus, the u_b of the technology can be calculated by using the bounds on the dimensions of the transistors in the particular technology. There are several design issues in this gate design as described below.

2.2 Design Issues

The gate design proposed in the previous section has several drawbacks.

- In this gate design output and input delays are not independent for both falling and rising transitions. For example, the NAND gate consists of two pMOS transistors in parallel and two nMOS transistors is series. The gate has different rising delays along both inputs if pMOS transistors are sized differently. But the same is not true for a falling delay. Altering the size of one of the nMOS transistors affects the R_{on} of the output discharging path and, thus, the output delay. This dependancy makes the sizing, for a given delay, a non-linear problem which can be difficult to converge.
- The parasitic capacitance C_p is assumed to be constant and independent of the transistor sizes. But in reality, C_p is a function of the transistor sizes. Altering the sizes of one transistor can affect C_p and the output gate delay.
- When the transistors are connected in series to one other, some of them are ON and some are OFF. This causes the threshold voltages of the transistors to change drastically due to *body effect* [25,26]. This makes the output delay of the gate, input pattern dependant. This is a problem as the LP gives a single delay for every gate output [10,11].

3 Gate Design with nMOS Pass Transistors

In the design proposed in Sec. 2, the main problem was the inter-dependence of output and input delays. In this second design, we propose to leave the input capacitance unaltered, and increase the resistance of the path.

3.1 Effects of Increasing Resistance and Input slope

Consider the charging path shown in Figure 1. Energy is drawn from the supply to charge the C_L through R_{on}. The energy consumed by a signal transition is

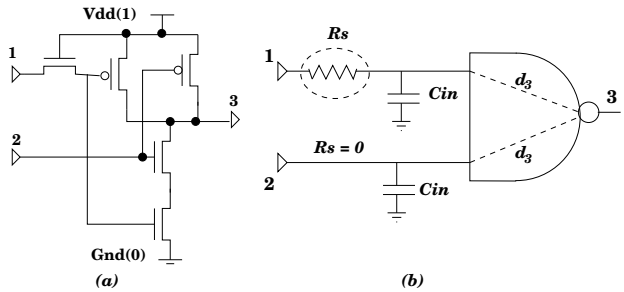

Fig. 2. The proposed single added nMOSFET VID NAND gate. (a) Transistor Level showing the nMOS transistor added and (b) charging path for transitions along the different paths through the gate.

given by $0.5C_L{V_{dd}}^2$, where C_L is the load capacitance and V_{dd} is the supply voltage. Note that the energy expression does not include resistance R_{on} in it. The resistance governs the switching time but the overall energy per transition remains the same. Hence, *increasing the resistance of the path does not alter the energy consumed per transition.* Increasing resistance of the slope however, degrades the slew of the input waveform. This increase in input slope affects gate delay and needs to be acounted for.

$$Gate\ Delay = t_{step}\ +\ t_{slew}$$

where t_{step} is the gate delay when the input is a step waveform and t_{slew} is the gate delay due to the input slope or *slew*. Thus, by increasing R_{on} we manipulate t_{slew} part of the gate delay. But increasing the input slew decreases the *robustness* and noise immunity of the circuit [25]. A large input slope means that the circuit is *in transition* for a longer period of time and is more susceptible to noise and short-circuit power. The input slope is *restored* or improved by using *regenerative gates*. The CMOS logic gates are regenerative as they improve the slope of the waveform while passing the signal transition from the input to the output. In our new VID gate design by inserting resistance, we use this regenerative property of the CMOS gates in the output for restoring the slope. However, the slope restoration also has limits and hence, there is a practical limit to degrading the input slope. This is one of the major factors that influence the practical value of u_b of a given technology.

3.2 Proposed Gate Design

We insert a single nMOS transistor that is always ON, with resistance R_s, in the series charging path. A modified NAND gate is shown in Figure 2. The delays of the gate along both I/O paths are given by:

$$d_{2\rightarrow 3} = R_{on} \times C_L \tag{6}$$

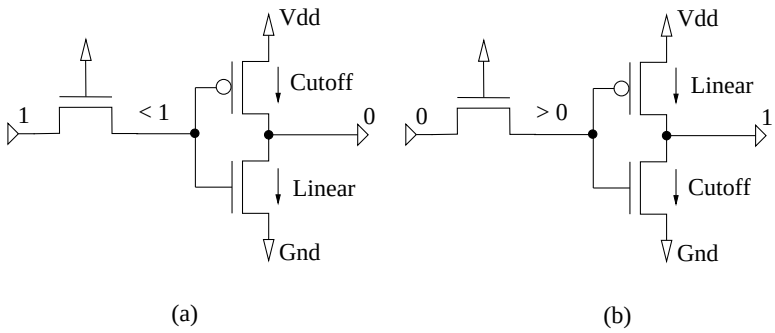

Fig. 3. The logic degradation of the single nMOS transistor addition (a) When logic 1 is passed through and (b) When logic 0 is passed through the gate.

$$d_{1\rightarrow 3} = R_{on} \times C_L \ + \ R_s \times C_L \tag{7}$$

$$= Ouput\ Delay \ + \ Input\ Delay \tag{8}$$

Thus, the input and output delays are separated completely from each other. The output delay can be controlled by sizing the gate transistors and the input delay can be controlled through R_s. $d_{2\rightarrow 3}$ is not affected by altering $d_{1\rightarrow 3}$. This concept can be extended to a n-input gate. The differential delay of path x with respect to the other $n-1$ paths, can be controlled by inserting $n-1$ transistors in series with the inputs. These paths can be independently controlled by sizing the $n-1$ transistors. Thus, we have a VID gate design that is extendible to all multi-input gate types.

3.3 Calculation of u_b

As seen from Equation 8, the input delay can be controlled independently by altering the size of the nMOS transistor. The nMOS transistor passes logic 0 effectively but degrades the signal when passing logic 1. Let us assume that there is a degradation of voltage λ when a logic 1 is passed through the transistor [25, 26]. When the transistor is acting as a resistor, there is an IR voltage drop also across the capacitor. The drop can be significant for two reasons:

- If the drop is too large, then the transistors in the fanout will not switch OFF completely. This increases short circuit dissipation of the fanout gate.
- The leakage power of the transistors is a function of the *gate to source voltage* (V_{gs}). Hence, larger drop would increase leakage current of fanout gate.

The circuit in Figure 3(a) shows a single transistor pair at the output of the nMOS. The operating regions for the transistors are as shown. The critical condition in this configuration is the pMOS transistor remaining in cutoff. If this condition is not met, the pMOS transistor is also ON and, hence, there is a direct path from the supply to the ground. This increases the short circuit dissipation. To meet the condition, we need to make sure that $V_g > V_{dd} \ - \ V_{tp}$,

where V_{tp} is the threshold voltage of the pMOS transistor. There are two factors that control the input voltage V_g is this case, (1)$I_{ds}R_s$, where I_{ds} is the drain to source stand-by current through the series transistor and (2)the signal degradation λ [25].

$$V_{dd} - \lambda \ - \ I_{ds}R_s > V_{dd} \ - \ V_{tp} \text{ or } R_s < \frac{V_{tp} \ - \ \lambda}{I_{ds}} \tag{9}$$

Consider the input configuration in Figure 3(b). The nMOS transistor passes a logic 0 without any degradation($\lambda = 0$). The critical condition here is the nMOS transistor in cutoff. By using a similar analysis as above, the condition is given by:

$$I_{ds}R_s < V_{tn} \text{ or } R_s < \frac{V_{tn}}{I_{ds}} \tag{10}$$

Equations 9 and 10 give the upper bound on R_s. This limits the amount of resistance that can be added to the charging path. Thus, the amount of input delay that can be added is also limited by this condition.

$$u_b = \frac{d_{diff}}{d_{min}} = \frac{R_{max} \times C_L}{R_{on} \times C_L} = \frac{R_{max}}{R_{on}} \tag{11}$$

where R_{max} is the maximum resistance that can be added and C_L is the load capacitance of the gate. This is the *theoretical limit* of u_b but the practical limit is governed by signal integrity issues as explained in Sec. 3.1.

3.4 Design Issues

This new VID gate design, although an improvement over the design in Sec. 2 has the following issues.

- Theoretical u_b can be further reduced by dimension limits on the series nMOS transistors.
- The short circuit dissipation is a function of the ratio of the input and output waveform slopes [25]. By inserting resistance we are increasing the input waveform slope thereby increasing the short circuit dissipation.
- The leakage power is a function of the gate to source voltage (V_{gs}). Since $\lambda > 0$ when passing a 1, the leakage power of the fanout transistors increases. This drawback is alleviated in the design discussed in the next section.
- This design has an area overhead due to extra transistors added.

4 Gate Design with CMOS Pass Transistors

In the gate design described in Sec. 3, the single nMOS transistor degrades logic 1, thereby increasing leakage power. This disadvantage can be alleviated by adding a CMOS pass transistor instead. The CMOS pass transistor consists of an nMOS and a pMOS transistor connected in parallel. Both transistors are kept always ON and $\lambda = 0$ while passing either logic 1 or logic 0.

4.1 Calculation of u_b

The u_b calculation is similar to the single nMOS added design but with $\lambda = 0$. Note that the resistance R_s is the effective parallel resistance of both the transistors together.

4.2 Design Issues

The design issues involved in this gate design are:

- R_s is the effective series/parallel resistance of both the nMOS and the pMOS transistors. Hence, effective resistance per unit length reduces and the transistors have to be longer to achieve the same resistance as a single nMOS transistor.
- Larger area overhead than the design in Sec. 3.

5 Technology Mapping

The process of designing gates that implement a given delay by altering the dimensions of the transistors is called *technology mapping* or *transistor sizing*. In this section we describe the transistor sizing of VID gates. From Eqn. 2, gate delay is dependant on C_L of the gate, which is dependant on the dimensions of the fanout gate size. Hence, to obtain a valid transistor sizing for delay at a gate G, the sizes of the gates in the fanout of G have to be decided. Therefore, to design an entire circuit, we use a *reverse breadth first search* methodology and first design the gates connected to the primary outputs and work towards the inputs of the circuit.

The objective is to design a gate with a load capacitance C_L in a particular instance, in order to have a required delay d_{req}. The procedure involves searching for the appropriate sizes for all of the transistors in the gate. The dimensions for the search space of an n-input gate are load capacitance, $2n$ transistor widths and $2n$ lengths for a total of $4n + 1$ dimensions. This can be a time consuming process to do for large circuits. So we propose to do this in two stages. The first stage is to generate a look-up table of sizes by simulation, for different d_{req} and C_L. For every gate type, we simulated the gate with the smallest sizes to find rising delay d_{rise} and falling delay d_{fall}. The objective function is to minimize $\epsilon = \frac{|d_{req}-d_{rise}| + |d_{req}-d_{fall}|}{d_{req}}$. The d_{rise} and d_{fall} can be increased by increasing the length of the transistors and decreased by increasing the width. Thus, by an iterative process an implementation for the given d_{req} and C_L can be achieved(to within acceptable values of error ϵ) and noted in the look-up table. Thus, the look-up table has size assignments for all different gate types and some values of C_L. This look-up table can be used for all circuits.

When a particular circuit is being optimized, the look-up table may not have the exact C_L. In such cases, we go to the second stage of fine tuning the sizes. We start with the closest entry in the look-up table. Each dimension is perturbed

by one unit(since dimensions are discrete in a technology) and the sensitivty is calculated where:

$$Sensitivity = \frac{d_{current}}{|d_{req} - d_{rise}| \ + \ |d_{req} - d_{fall}|}$$

where $d_{current}$ is the present measured gate delay, and d_{rise} and d_{fall} are the rise and fall delays after a perturbation in the dimension. There can be 8 perturbations, two for each of the dimensions. The perturbation with the highest sensitivity is incorporated and the gate is simulated again. The objective function is to minimize ϵ given earlier. This procedure is called the *steepest descent* method as the objective function is minimized by driving the dimensions based on sensitivities. The complexity is greatly reduced by using the lookup table as the search is limited to the neighborhood of the solution. Hence, local minima will not be a problem. The procedure can also be tuned for including the area of the cell in the objective function.

6 Summary

In this paper, we explained why conventional CMOS gates cannot be used as VID gates. We presented three new implementations of the VID gate. We presented an analysis of each of the gates and listed their shortcomings. Then we proposed a two-step approach for fixing the transistor sizes of every instance in the circuit. The main idea of this paper is to present the transistor level implementation details of the *variable input delay logic*. The advantages of the technique, its power reduction results and comparisons with other techniques are the same as presented in earlier publications and are not duplicated here [11–13].

References

1. Agrawal, V.D.: Low Power Design by Hazard Filtering. In: Proc. of the International Conference on VLSI Design. (1997) 193–197
2. Agrawal, V.D., Bushnell, M.L., Parthasarathy, G., Ramadoss, R.: Digital Circuit Design for Minimum Transient Energy and Linear Programming Method. In: Proc. of the International Conference on VLSI Design. (1999) 434–439
3. Berkelaar, M., Jacobs, E.: Using Gate Sizing to Reduce Glitch Power. In: Proc. of the ProRISC Workshop on Circuits, Systems and Signal Processing, Mierlo, The Netherlands (1996) 183–188
4. Berkelaar, M., Buurman, P., Jess, J.: Computing Entire Area/Power Consumption Versus Delay Trade-off Curve for Gate Sizing Using a Piecewise Linear Simulator. IEEE Transactions on Circuits and Systems **15** (1996) 1424–1434
5. Berkelaar, M., Jess, J.A.G.: Transistor Sizing in MOS Digital Circuits with Linear Programming. In: Proc. of the European Design Automation Conference, Mierlo, The Netherlands (1990) 217–221
6. Berkelaar, M., Jacobs, E.T.A.F.: Gate Sizing Using a Statistical Delay Model. In: Proc. of the Design Automation and Test in Europe Conference, Paris, France (2000) 283–290

7. Sathyamurthy, H., Sapatnekar, S.S., Fishburn, J.P.: Speeding up pipelined circuits through a combination of gate sizing and clock skew optimization. In: Proc. of the International Conference on Computer-Aided Design. (1995) 467–470
8. Benini, L., DeMicheli, G., Macii, A., Macii, E., Poncino, M., Scarsi, R.: Glitch power minimization by gate freezing. In: Proc. of the Design Automation and Test in Europe Conference. (1999) 36
9. Kim, S., Kim, J., Hwang, S.Y.: New Path Balancing Algorithm for Glitch Power Reduction. IEE Proceedings: Circuits, Devices and Systems **148** (2001) 151–156
10. Raja, T., Agrawal, V.D., Bushnell, M.L.: Minimum Dynamic Power CMOS Circuit Design by a Reduced Constraint Set Linear Program. In: Proc. of the International Conference on VLSI Design. (2003) 527–532
11. Raja, T., Agrawal, V.D., Bushnell, M.L.: CMOS Circuit design for Minimum Dynamic Power and Highest Speed. In: Proc. of the International Conference on VLSI Design. (2004) 1035–1040
12. Raja, T., Agrawal, V.D., Bushnell, M.L.: Variable Input Delay Logic and Its Application to Low Power Design. In: Proc. of the International Conference on VLSI Design. (2005) 598–605
13. Raja, T.: Minimum Dynamic Power Design with Variable Input Delay Logic. PhD thesis, Rutgers University, Dept. of ECE, Piscataway, New Jersey (2004)
14. Datta, S., Nag, S., Roy, K.: ASAP: A Transistor Sizing Tool for Area, Delay and Power Optimization of CMOS Circuits. In: Proc. of the IEEE International Symposium on Circuits and Systems. (1994) 61–64
15. Musoll, E., Cortadella, J.: Optimizing cmos circuits for low power using transistor reordering. In: Proc. of the European Design Automation Conference. (1995) 219–223
16. Hashimoto, M., Onodera, H., Tamaru, K.: A Practical Gate Resizing Technique Considering Glitch Reduction for Low Power Design. In: Proc. of the Design Automation Conference. (1999) 446–451
17. Chandrakasan, A.P., Brodersen, R.W.: Low Power Digital CMOS Design. Kluwer Academic Publishers, Boston (1995)
18. Chandrakasan, A.P., Sheng, S., Brodersen, R.W.: Low Power CMOS Digital Design. IEEE Journal of Solid-State Circuits **27** (1992) 473–484
19. Chandrakasan, A., Brodersen, R., eds.: Low-Power CMOS Design. IEEE Press, New York (1998)
20. Nebel, W., Mermet, J.: Low Power Design in Deep Submicron Electronics. Kluwer Academic Publishers, Boston (1997)
21. Rabaey, J.M., Pedram, M., eds.: Low Power Design Methodologies. Kluwer Academic Publishers, Boston (1996)
22. Rabaey, J.M., Pedram, M.: Low Power Design Methodologies. Kluwer Academic Publishers, Boston (1995)
23. Roy, K., Prasad, S.C.: Low-Power CMOS VLSI Circuit Design. Wiley Interscience Publications, New York (2000)
24. Yeap, G.: Practical Low Power Digital VLSI Design. Kluwer Academic Publishers, Boston (1998)
25. Rabaey, J., Chandrakasan, A., Nikolic, B.: Digital Integrated Circuits: A Design Perspective. Prentice Hall, Upper Saddle River, NJ (2003)
26. Weste, N., Eshraghian, K.: Principles of CMOS VLSI Design: A Systems Approach. Addison Wesley Publications, Reading, MA (1985)