Buffer Sizing for Minimum Energy-Delay Product by Using an Approximating Polynomial*

Chang Woo Kang, Soroush Abbaspour, and Massoud Pedram
University of Southern California/EE-systems
3740 McClintock Ave. EEB-314
Los Angeles, CA 90089, USA
{ckang, sabbaspo, pedram}@usc.edu

ABSTRACT
This paper first presents an accurate and efficient method of estimating the short circuit energy dissipation and the output transition time of CMOS buffers. Next, the paper describes a sizing method for tapered buffer chains. It is shown that the first-order sizing behavior, which considers only the capacitive energy dissipation, can be improved by considering the short-circuit dissipation as well, and that the second-order polynomial expressions for short-circuit energy improves the accuracy over linear expressions. These results are used to derive sizing rules for buffered chains, which optimize the overall energy-delay product.

Categories and Subject Descriptors
B.7.1 [**Integrated Circuits**]: Types and Design Styles – *VLSI (very large scale integration), advanced technologies.*

General Terms
Performance, Design.

Keywords
Buffer sizing, Short circuit energy, Polynomial approximation.

1. INTRODUCTION
Reduction of energy dissipation in CMOS digital circuits has become an important goal of the design optimization process. Optimization tools rely on accurate and efficient energy analysis and estimation techniques. These techniques, in turn, need to account for all the key components of the energy consumption in CMOS circuits. One such component, which is generally referred to as the short circuit or rush-through energy dissipation, is the energy consumed by flow of current from *Vdd* to *Gnd* through a direct current path that is temporarily established during an output transition. Short circuit energy dissipation is becoming an important factor as the number of buffers increase. Without considering the short circuit energy dissipation, sizing a multi-stage buffer to drive a large capacitive load may result in a poor solution in terms of the energy-delay product.

*This research was supported in part NSF under grant no. 9988441.

GLSVLSI'03, April 28-29, 2003, Washington, DC, USA.

The focus of this work is multi-stage buffer sizing for the minimum energy-delay product where the energy term accounts for both the capacitive and the short circuit components. The latter component is calculated by using an approximating polynomial. By having the input transition time, the size of a buffer, and the output capacitive load, the short circuit energy dissipation, $E_{sc,}$ and output transition time, τ_{out}, for buffer chains can be accurately evaluated by the proposed formula, and furthermore, the formula can be used to find optimal sizes of buffers for the minimum energy-delay product in buffer chains. This scheme is applicable for CAD tools requiring accuracy as well as fast computation time.

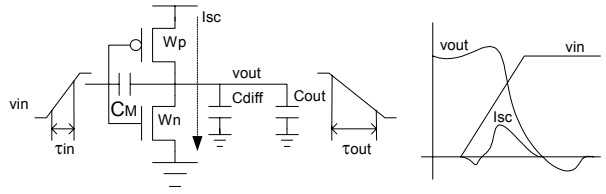

Figure 1: An inverter driving a capacitive load and electrical waveform showing short circuit current.

There has been much research done on developing closed-form expression [1][2][4][7][9]. Ko and Balsara proposed a gate-sizing technique for reducing overall power dissipation on non-critical paths [6]. Turgis et al. introduced the notion of a short-circuit capacitance to capture the short circuit power dissipation [7]. Short circuit energy dissipation occurs when NMOS and PMOS transistors establish a direct path from power supply to ground because they are simultaneously turned on during input transition as shown in Figure 1. When input, V_{in}, changes from low to high, the PMOS transistor enters the linear region. Before V_{in} reaches the threshold voltage of the NMOS transistor, current flows from the output load to the power supply because of the overshoot caused by gate-to-drain coupling capacitance C_M. When the NMOS transistor enters the saturation region after the threshold voltage is crossed, short circuit current, I_{sc}, starts flowing from the power supply to the ground. After that, the PMOS transistor enters the saturation region and then is turned off while the NMOS transistor enters the linear region [1][2][9]. E_{sc} and τ_{out} in a CMOS gate are dependent on the size of transistors, the input transition time, and the output load. E_{sc} can be measured by integrating positive I_{sc} at the PMOS transistor for the falling output transition. Discussion for the rising output transition is symmetric.

The size of inverter, W, represents the sum of the widths of NMOS and PMOS transistor of an inverter. According to [3][7], E_{sc} increases linearly with the input transition time, τ_{in}, because a long input transition time increases the time, during which, both transistors are on. E_{sc} is also a linear function of the inverse of the output capacitance. A large output capacitive load keeps the voltage between the drain and source of a transistor at a small value for a long period, resulting in small amount of short circuit current. E_{sc} increases linearly as the size of inverter increases. The gate width of a transistor is the key factor in limiting the amount of short circuit current. Similarly, there exist linear relationships between τ_{out} and the input transition time, the output load, and inverse of the size of a transistor [7]. As the input transition time becomes longer, τ_{out} increases linearly. τ_{out} is also a linear function of the output load, C_{out}. τ_{out} is the time duration for discharging the output load. The resistance of a transistor is proportional to the inverse of the transistor width. Therefore, τ_{out} is a linear function of the inverse of inverter size.

We propose first-order and second–order approximating polynomials for estimating short circuit energy dissipation and output transition time in Section 2. Approximation results and optimal sizing solutions for the minimum energy-delay product in a multi-stage buffer chain are provided in Section 3. Concluding remarks are given in Section 4.

2. POLYNOMIAL APPROXIMATION

As noted in Section 1, E_{sc} is a linear function of the transistor width, the input transition time, the inverse of the output load.[1] In comparison, τ_{out} is a linear function of the inverse of the transistor width, the input transition time, and the output load. Hspice simulations are used to determine E_{sc} and τ_{out} for all combinations of two inverter sizes, two input transition times, and two output loads. Next, we calculate E_{sc} and τ_{out} for a given input triplet by interpolating between these eight corner combinations.

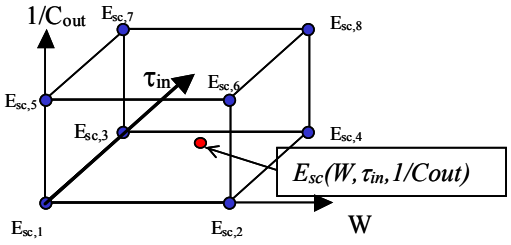

Figure 2: Three-dimension linear approximation.

Consider two distinct transistor widths W_i, two distinct input transitions $\tau_{in,i}$, and two distinct inverse load values $1/C_{out,i}$. Suppose we have obtained short circuit energy simulation results $E_{sc,1}$ through $E_{sc,8}$ for all eight combinations of $<W_i, \tau_{in,i}, 1/C_{out,i}>$ where i=0,1 as shown in Figure 2. To estimate E_{sc} of a new combination, $<W, \tau_{in}, 1/C_{out}>$, which may be inside or outside of the box in Figure 2, we use the following equations:

$$\begin{aligned} E_{sc} = {} & E_{sc,1}(1-X)(1-Y)\left(1-\hat{Z}\right) + E_{sc,2}X(1-Y)\left(1-\hat{Z}\right) \\ & + E_{sc,3}(1-X)Y\left(1-\hat{Z}\right) + E_{sc,4}XY\left(1-\hat{Z}\right) \\ & + E_{sc,5}(1-X)(1-Y)\hat{Z} + E_{sc,6}X(1-Y)\hat{Z} \\ & + E_{sc,7}(1-X)Y\hat{Z} + E_{sc,8}XY\hat{Z} \end{aligned} \quad (1)$$

$$X = \frac{W - W_1}{W_2 - W_1} \quad Y = \frac{\tau_{in} - \tau_{in,1}}{\tau_{in,2} - \tau_{in,1}} \quad \hat{Z} = \frac{1/C_{out} - 1/C_{out,1}}{1/C_{out,2} - 1/C_{out,1}}$$

where W is the size of an inverter, and τ_{in} is the 10-90% input transition time, and C_{out} is the output load. After simplifying the above equation, we have:

$$E_{sc}(W, \tau_{in}, 1/C_{out}) = \sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1} m_{ijk} \frac{W^i \tau_{in}^{\ j}}{C_{out}^{\ k}} V_{DD} \quad (2)$$

where m_{ijk} are constant coefficients that are dependent on the technology and W is sum of the widths of PMOS and NMOS transistors. Notice that i, j and k values in equation (2) are the exponents of the corresponding variables W, τ_{in} and C_{out}.

Similarly, if we compute τ_{out} for eight different $<1/W_i, \tau_{in,i}, C_{out,i}>$ combinations, then τ_{out} for each new triplet of parameters will be:

$$\tau_{out}(1/W, \tau_{in}, C_{out}) = \sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1} p_{ijk} \frac{C_{out}^{\ i} \tau_{in}^{\ j}}{W^k} \quad (3)$$

where p_{ijk} are constant coefficient numbers dependent on technology and again W is sum of the widths of PMOS and NMOS. Notice that the diffusion capacitance of an inverter is linearly proportional to the inverter size W, and therefore, its effect on E_{sc} and τ_{out} is captured through appropriate adjustments to the coefficients of W and $1/W$ in equations (2) and (3). From simulation results, we observe that although the prediction accuracy of the linear approximating function for τ_{out} is quite high, the prediction accuracy for E_{sc} as a function of the input transition time and the inverse of output load may be improved by using a second order approximating polynomial. Therefore, we modeled E_{sc} by using a linear equation for its dependency on the size of an inverter and second-order equations for its dependency on the input transition time and output load, resulting in equations (4).

$$E_{sc}(W, \tau_{in}, C_{out}) = \sum_{i=0}^{1}\sum_{j=0}^{2}\sum_{k=0}^{2} m_{ijk} \frac{W^i \tau_{in}^{\ j}}{C_{out}^{\ k}} V_{DD} \quad (4)$$

This requires more coefficients, but yields more accurate results. We compared results from equation (3) and (4) with Hspice simulation and the accuracy was 1.1% and 1.6%. Results are shown in 3.

3. MODEL ACCURACY AND THE MINIMUM ENERGY-DELAY PRODUCT

Figure 3(a)-(f) show the results of the second-order approximation and Hspice simulation for a falling output transition. Solid lines denote our energy model predictions. We performed 119 simulations. Using the first-order approximations, the average error for the short energy dissipation was 4.0% whereas the average error for output transition time was only 1.1%. When we used the second-order approximation for E_{sc}, the average error for E_{sc} reduced to 1.6%. In this case, the maximum errors for E_{sc} and τ_{out} are 6% and 3%, respectively. Finding inverter sizes to minimize the energy-delay product is essential to save energy and/or improve circuit speed.

[1] A function of multiple variables is (multi)-linear if it is linear with respect to each of its variables.

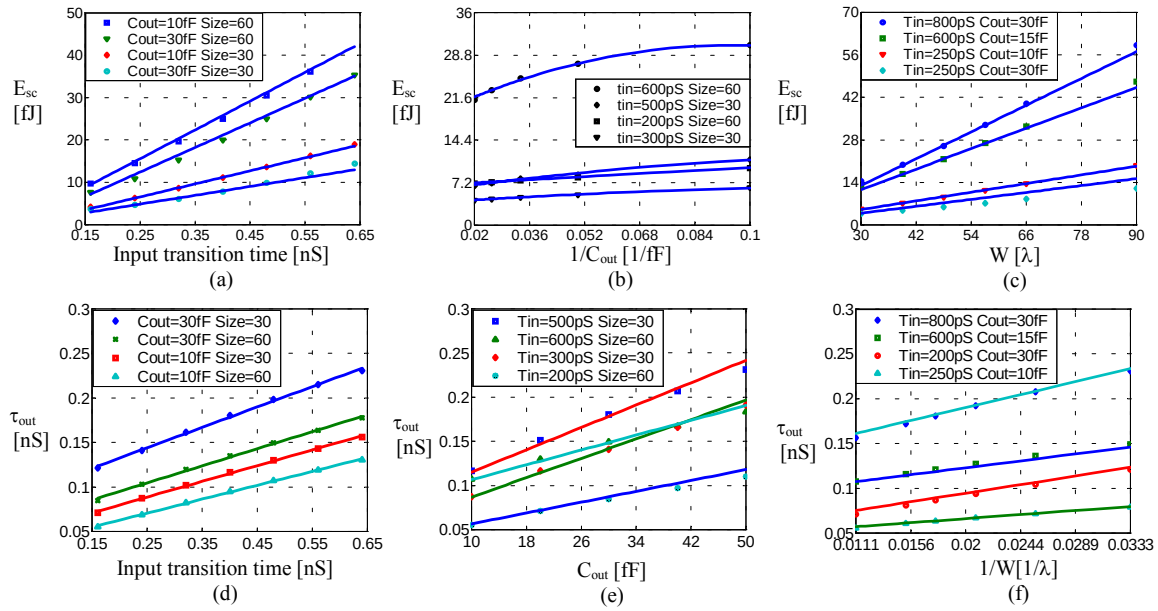

Figure 3: Short circuit energy dissipation and output transition time comparison between Hspice results (markers) and results from the second-order approximation (solid lines).

Suppose the input transition time for the first stage buffer, the size of the first buffer, and output load capacitances are given. Equipped with equations (3) and (4), and without direct Hspice simulation, we can estimate E_{sc} and τ_{out}. By using our formula we can also determine sizes of inverters in each stage in order to achieve the minimum energy-delay product. For this problem, we consider the gate capacitance and the diffusion capacitance as functions of the inverter size. In other words, $C_{g0}=\beta W_0$, $C_{d0} = \alpha W_0$, $C_{gx}=\beta W_x$, and $C_{dx} = \alpha W_x$ in Figure 4 here α and β are gate and diffusion capacitance of unit size. In addition, we assume that an inverter has equal rising and falling time. Therefore, the propagation delay will be proportional to either the rising or the falling transition time [10]. In this section, we present a methodology to find the optimal size of each buffer in stages using our formula for short circuit energy dissipation and output transition time. The buffer chain must show the minimum energy-delay product.

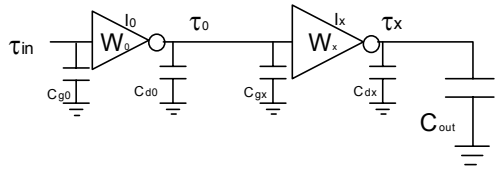

Figure 4: Finding optimal size W_x in a two-stage buffer chain for minimum energy-delay product.

Figure 4 shows a two-stage buffer sizing problem. By using our formula we can express delay and energy dissipation as follows:

$$E_{capacitive} = 1/2\left(C_{g0}+C_{d0}+C_{gx}+C_{dx}+C_{out}\right)V_{DD}^{\ 2}$$
$$= 1/2\left[(\alpha+\beta)\times(W_0+W_x)+C_{out}\right]V_{DD}^{\ 2}$$

$$E_{sc0} = E_{sc}\left(W_0,\tau_{in},1/C_{gx}\right) = E_{sc}\left(W_0,\tau_{in},1/\beta W_x\right)$$

$$\tau_0 = \tau_{out}\left(1/W_0,\tau_{in},C_{gx}\right) = \tau_{out}\left(1/W_0,\tau_{in},\beta W_x\right)$$

$$E_{scx} = E_{sc}\left(W_x,\tau_0,1/C_{out}\right)$$

$$\tau_x = \tau_{out}\left(1/W_x,\tau_0,C_{out}\right)$$

$$Delay \propto (\tau_0+\tau_x)$$

$$E_{sc} = E_{sc0}+E_{scx}$$

$$EDP \propto \left(E_{capacitive}+E_{sc}\right)\times Delay$$

These equations can be expressed as functions of W_x:

$$E_{capacitive} = a_0 + a_1 W_x$$

$$E_{sc0} = \sum_{i=0}^{1}\sum_{j=0}^{2}\sum_{k=0}^{2} m_{ijk}\frac{W_0^{\,i}\tau_{in}^{\,j}}{(\beta W_x)^k}V_{DD} = \frac{b_0}{W_x^2}+\frac{b_1}{W_x}+b_2$$

$$\tau_0 = \sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1} p_{ijk}\frac{(\beta W_x)^i\tau_{in}^{\,j}}{W_0^{\,k}} = c_0+c_1W_x$$

$$E_{scx} = \sum_{i=0}^{1}\sum_{j=0}^{2}\sum_{k=0}^{2} m_{ijk}\frac{W_x^{\,i}(c_0+c_1W_x)^j}{C_{out}^{\ k}}V_{DD} = d_0+d_1W_x+d_2W_x^2+d_3W_x^3$$

$$\tau_x = \sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1} p_{ijk}\frac{C_{out}^{\ i}(c_0+c_1W_x)^j}{W_x^{\,k}} = \frac{e_0}{W_x}+e_1+e_2W_x$$

$$EDP \propto \left(E_{capacitive}+E_{sc}\right)\times Delay = \sum_{i=-3}^{4} f_i\times W_x^{\ i}$$

To determine the optimum buffer sizes, we have:

$$\frac{d}{dW_x}EDP \propto \sum_{i=-4}^{3} i\times f_i\times W_x^{\ i-1} = 0$$

We solved this non-linear equation in MATLAB by using the *least squares method*. The results are depicted in Figure 5. Figure 5(a) shows optimal values of W_x for different combinations of W_0, C_{out}, and τ_{in}. Figure 5(b), (c), and (d) show energy dissipation, delay, and energy-delay product for the optimum W_x obtained from Figure 5(a) as a function of C_{out} and W_0 for τ_{in} = 300 ps. We make a few observations. From Figure 5(b), we can see that the capacitive energy dissipation increases linearly with W_0 and C_{out}, whereas the short circuit energy dissipation increases as W_0 increases and decreases as C_{out} increases. Furthermore, the percentage of the short circuit

energy dissipation is less than 20% of total energy dissipation. From Figure 5(c), the delay increases as C_{out} increases and W_0 decreases, which is the expected result. This increase in delay is, however, quite small (and hence negligible) in the region to the left of the line that is marked by *Z*. Therefore, from Figure 5(b) and (c), we conclude that for a given output load, we should use the minimum W_0 value that does not result in a significant delay increase (i.e., move us to the right of line Z), which, in turn, may cause a required arrival time constraint violation. Figure 5(d) shows clearly that the energy-delay product remains nearly constant over a large range of values for W_0. This is because the total energy dissipation and the delay change in opposite directions with respect to W_0. Therefore, given C_{out} and τ_{in} values (which is the typical scenario that we encounter during the circuit optimisation flow), one can easily trade off energy for delay or vise versa without changing the overall energy-delay product.

Similarly, we determined the optimal inverter sizes for a three-stage inverter chain. Notice that W_x and W_y are the sizes of inverters in the second and third stages, respectively. When τ_{in} is 300 *ps*, Figure 6(a) shows optimal sizes for buffers in a three-stage buffer chain with different size of the first stage buffer and output capacitive load. Figure 6(b) shows the optimum energy-delay product as a function of C_{out} and W_0 for τ_{in} = 300 ps (i.e, the energy-delay product for optimum values of W_x and W_y). We have generated optimum sizing results for four and five-stage inverter chains by using the same methodology. Results are similar and not included here due to space limitation.

4. CONCLUSION

This paper has presented an accurate and efficient method of estimating the short circuit energy dissipation and the output transition time of CMOS buffers by using first-order and second-order polynomial approximations and a methodology to find an optimal buffer sizing solution in terms of the energy-delay product where the energy term accounts for both the capacitive and the short circuit components. Simulation and optimal sizing results for a CMOS buffer chain in a 0.18 µm process technology have been presented.

5. References

[1] H. Veendrick, "Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits," *IEEE J. Solid-State Circuits,* vol. SC-19, pp. 468-473, 1984.

[2] T. Sakurai, A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter and other formulas," *IEEE J. Solid-State Circuits*, vol. 25, pp. 584-594, April 1990.

[3] S. Nikolaidis and A. Chatzigeorgiou, "Modeling the transistor chain operation in CMOS gates for short channel devices," *IEEE Transactions on Circuits and Systems,* vol. 46, no. 10, October 1999.

[4] P. Maurine, M. Rezzoug and D. Auvergne, "Output transition time modeling of CMOS structures", *IEEE International Symposium on Circuits and Systems,* vol. 5, pp. 363-366, 2001.

[5] A. Chatzigeorgiou and S. Nikolaidis, "Collapsing the CMOS transistor chain to an effective single equivalent transistor," in IEE *Proc. on Circuits, Devices and Systems,* vol. 145, no. 5, pp. 347-353, October 1998.

[6] U. Ko and P. T. Balsara, "Short-circuit power driven gate sizing technique for reducing power dissipation," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 3, no. 3, September 1995.

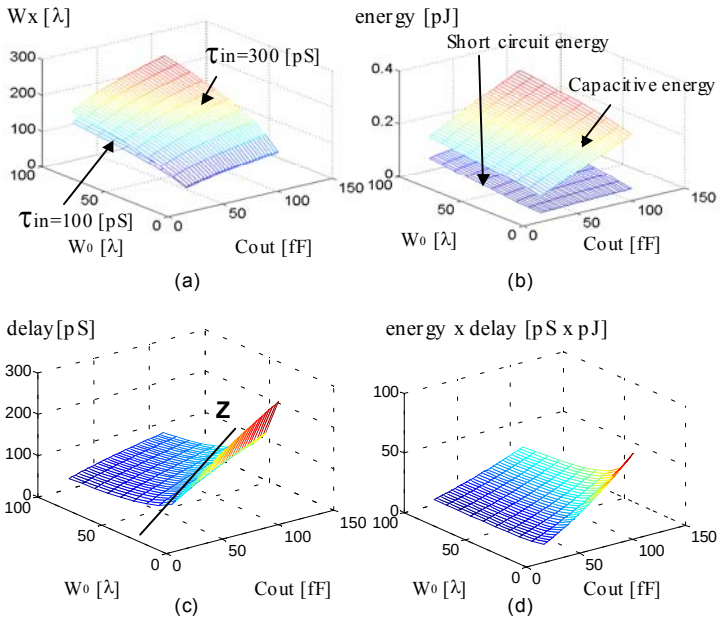

Figure 5:Two-stage buffer chain; (a) optimal size W_x for two different τ_{in} values; (b) energy dissipation, (c) delay, and (d) energy-delay product for τ_{in} = 300 ps.

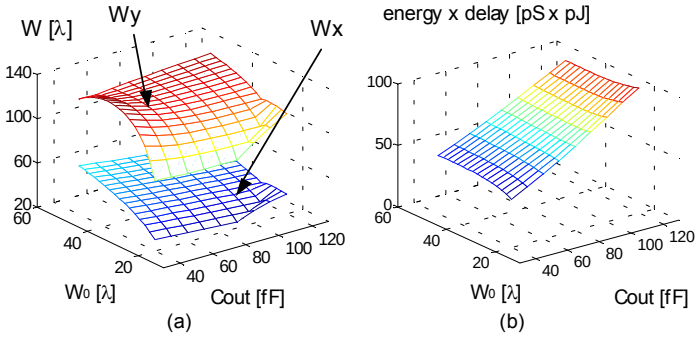

Figure 6: Three-stage buffer chain; (a) optimal size W_x and W_y; (b) energy-delay product for τ_{in} = 300 ps.

[7] S. Turgis, N. Azemard, and D. Auvergne, "Explicit evaluation of short-circuit power dissipation and its influence on propagating delay for static CMOS gates," *in Proc. IEEE Int. Symp.on Circuits and Systems*, vol. 4, pp. 751-754, May 1996.

[8] M. Borah, R. Michael Owens, and M. J. Irwin, "Transistor sizing for low power CMOS circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 15, no.6, June 1996.

[9] L. Bisdounis, O. Koufopavlou, and S. Nikolaidis, "Accurate evaluation of CMOS short-circuits power dissipation for short-channel devices," *in Proc. Int. Symp. Low Power Electronics Devices*, pp. 189-192, August 1996.

[10] J. M. Rabaey, Digital integrated circuits: a design perspective, Upper Saddle River, NJ: Prentice Hall, 1996