A High Speed CMOS Buffer for Driving Large Capacitive Loads in Digital ASICs

Radu M. Secareanu and Eby G. Friedman
Department of Electrical Engineering
University of Rochester
Rochester, NY 14627-0231
radums@ee.rochester.edu, friedman@ee.rochester.edu

Abstract— **A High Speed High-Drive (HD) CMOS Buffer is described in this paper which is an alternative to the widely used CMOS tapered buffer. The paper introduces the principle of operation of the HD buffer and compares it with a tapered buffer. Depending upon the capacitive load, the HD buffer as compared to an equivalent tapered buffer can provide increased speed (up to 2.2x) or multiple speed/power/area trade-offs. Clock distribution networks, large data busses, and I/O buffers are some possible applications of this buffer structure.**

I. Introduction

Continuous technology scaling has changed many of the problems that modern ASIC designers face in designing high performance digital circuits. With deep submicrometer technologies [1], on-chip interconnect has become a fundamental issue. High interconnect resistance and capacitance have become an important factor in limiting performance. Driving large off-chip capacitive loads is also an important design issue. Techniques to efficiently drive these large loads have been developed, tapered buffers being the primary example [2–8]. Technological improvements have also contributed, with the development of low dielectric constant and low resistivity materials being prime examples.

The proposed high speed High-Drive (HD) buffer structure is a circuit capable of driving large capacitive loads at higher speeds than a tapered buffer. Depending on the capacitive load and on the desired speed/power/area trade-offs, the HD buffer can provide higher speed, lower power, significant area savings, and/or improved transition times. As compared to the FS buffer developed by Huang and Chu [9], the HD buffer improves the speed and noise susceptibility and minimizes the dissipated power, while decreasing the circuit complexity, reducing area, and simplifying the circuit design process.

A detailed description of the operation of the HD buffer circuit is presented in Section II. Some sizing considerations for an optimal HD buffer are summarized in Section III. Simulation results comparing the performance of both the tapered buffer and the HD buffer are described in Section IV. Finally, some conclusions are presented in Section V.

This research was supported in part by the National Science Foundation under Grant No. MIP-9423886 and Grant No. MIP-9610108, the Army Research Office under Grant No. DAAH04-G-0323, a grant from the New York State Science and Technology Foundation to the Center for Advanced Technology—Electronic Imaging Systems, and by grants from the Xerox Corporation, IBM Corporation, and Intel Corporation.

II. The High-Drive Buffer

A transistor level schematic of the proposed HD buffer is shown in Figure 1. A more simplified version, provided to enhance the explanation of the circuit behavior, is shown in Figure 2. There are three categories of transistors in this circuit: driving transistors (n and p) that source or sink each stage up to the load, nulling transistors (m and q), which reestablish the logic levels, and glue logic transistors (the input gates and the inverters), that provide different logic levels inside the HD buffer. The HD buffer example shown in Figure 1 is six stages long, the driving transistors being n1–n6 and p1–p6, each stage increasingly larger.

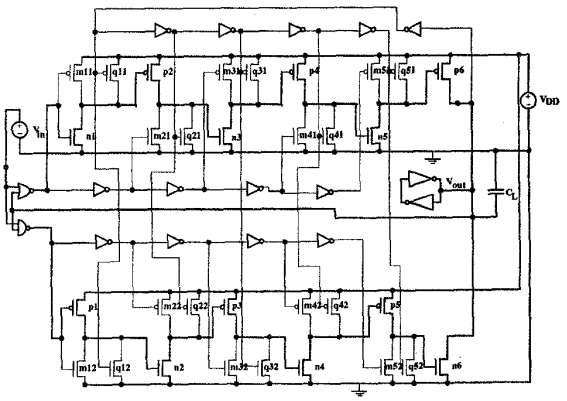

Fig. 1. The transistor level schematic of the HD buffer

To obtain the savings in speed, power, and area, the HD buffer uses two separate data paths, one for each logic state of the input signal (0 or 1). Each path is optimized to amplify a state and transmit the signal to the load with minimal energy consumption. The path that amplifies and transmits the 1 (0) state is called the "fast 1" ("fast 0") path. The fast 1 path starts with the smallest n transistor from stage 1, n1, and continues with p2, n3, p4, n5, and p6 as shown in the upper side of the HD circuit in Figure 1. The fast 0 path starts with p1, and continues with n2, p3, n4, p5, n6 as shown in the lower side of Figure 1. When the desired drive current is reached, these two paths are combined to drive the output load (the common drain node of p6 and n6).

As shown in Figure 1, the load for each driving transistor consists of the gate capacitance of the following driving transistor along the path and the drain capacitance of the corresponding nulling transistors. For a tapered buffer system [2–8], the corresponding load is much larger, consisting of the junction capacitance of the transistor pair plus the gate capacitance of both transistors of the following buffer stage. The gain in speed of the HD buffer is derived from the smaller capacitance at the internal stages, as well as from the following effect. In a tapered buffer, the N and P transistors of a stage are designed to have the same transconductance. During a transition, the delay of each inverter can substantially increase since the PMOS and NMOS transistors conduct simultaneously, creating a DC path between V_{DD} and GND. This effect adversely degrades the charge/discharge process of the load capacitance for each stage of the buffer, increasing the delay [10]. For the HD buffer, this effect does not exist, since, when a driving transistor is turned on, the nulling transistor pair is off (the nulling process is completed). The smaller capacitance at the internal stages also provides important savings in the dynamic power as well as higher stage to stage tapering coefficients which implies less stages are necessary to drive a particular load. Less stages permit the HD buffer to operate faster.

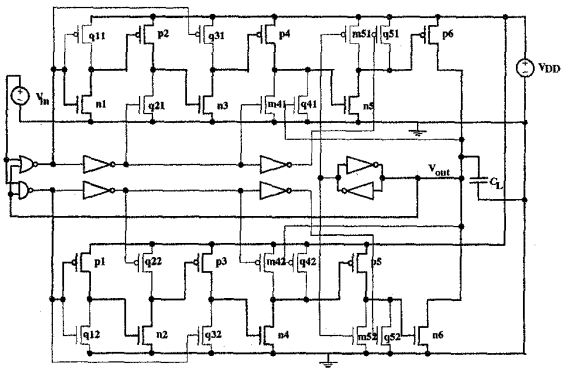

Fig. 2. A simplified transistor level schematic of the HD buffer.

In order to generate a fast output transition, only one of the final transistors must pull up or down at any one time. Several circuit techniques are employed to realize this objective.

1. The fast 1 path is driven by a NOR gate while the fast 0 path is driven by a NAND gate. The feedback path from the buffer output to the input of the two logic gates permits the fast 1 and the fast 0 paths to stay on for a short period of time (from the change of state at the data input until the output reaches the final state).

2. The m and q nulling transistors are separated from the data paths and do not affect the speed or the dynamic power dissipation of the data paths. By employing a synchronous driving technique, the short circuit or DC power dissipation is eliminated by the nulling transistors. Note that in Figure 1 each stage has double nulling transistors. The m transistors designate the maintenance nulling transistors (MNT) and the q transistors designate the fast nulling transistors (FNT). In between transitions, when no path is driven, each gate drives the MNT through the glue logic transistors. As shown in Figure 1, the FNT are driven by the output through the upper inverters.

Consider the case where the output of the HD buffer is in the 1 state, sourced by the p6 transistor. The next 0 output is generated by the fast 0 path through the n6 transistor. To provide a fast output signal, only the p6 or n6 should be on. Since a fast 1 to 0 output transition is desired, only the n6 transistor must be on during the transition. The p6 transistor is therefore turned off (nulled) in between transitions when no path is driven by slowly charging the gate of the p6 transistor through the corresponding nulling transistors. To efficiently turn off the p6 transistor, the n5, p4 n3, p2, and n1 transistors are each turned off by the corresponding nulling transistors of each stage by charging or discharging the gates of the driving P or N type transistors. This process prepares the driver for the impending fast transition. This nulling process is realized by the FNT fed back from the output signal.

Note that when the fast 1 path is nulled, only the FNT of the fast 1 path are driven while the FNT of the 0 path are off. The internal nodes of the 0 path can float for up to half a period of the input signal. A noise glitch or leakage present at any of these nodes could create a malfunction of the buffer, reflected in either an increase in delay, and/or an increase in the dissipated power. The solution implemented by the HD buffer is to use the MNT. The MNT are minimally sized, their only role is to minimize the leakage current and noise susceptibility of the internal nodes for up to half a period of the input signal.

To eliminate any short circuit and DC power for all stages of both the fast 1 and fast 0 paths, each group of nulling transistors are synchronously driven so that no DC path is created between a driving and nulling transistor pair. Each inverter in the nulling path introduces a delay equal to the time necessary to null one stage. Thus, when the nulling process of one stage is completed, the next stage is nulled. For example, when the nulling process of stage 1 is terminated, stage 2 begins to be nulled; p2 is off, eliminating any DC path. This technique substantially reduces the dissipated power of the HD buffer. An expression for the total dissipated power of the HD buffer is

$$P = (C_L + \Sigma C_i + \Sigma C_j) V_{DD}^2 f, \qquad (1)$$

where C_i are the capacitances at the internal nodes in the 0 path and 1 path and C_j are the node capacitances of the nulling transistors.

If the input signal frequency is low, the FNT can be driven by the input gates without affecting the speed and power dissipation of the HD buffer. Only single nulling transistors are necessary, since the FNT also behave as MNT, providing the noise and leakage immunity.

The HD buffer shown in Figure 2 uses single nulling transistors for the less significant stages, and double nulling transistors for the final stages that require larger nulling transistors. The example circuit shown in Figure 2 does not employ synchronous driving. However, to reduce short circuit and DC power, synchronous driving can be employed.

3. In between transitions, when no path is driven and the nulling process for the current path is completed, C_L floats and any leakage current could destroy the output state. This effect is prevented by the small latch at the output, which maintains the output signal on C_L. The effect of C_L floating can also be compensated for by noting that the nulling time is inversely proportional to the width of the nulling transistor. By making the nulling transistors small, the nulling process is delayed and C_L floats for a shorter amount of time. At the limit, the nulling time equals the time between transitions and C_L does not float at all and no latch is needed. However, a fast transition is achieved when only one final transistor pulls at any given time. Therefore, the nulling transistors must be properly sized so that the nulling process is completed before but close to the moment when the next transition occurs.

III. HD BUFFER SIZING CONSIDERATIONS

The sizing of an optimal HD buffer is presented in this section. The sizing process is based on charging (discharging) the capacitive load at each node of the circuit by a PMOS (NMOS) transistor.

The HD buffer driving transistors are sized based on the condition that each stage introduces the same delay, equal to the delay introduced by the final stage. The input gates are sized to be equivalent to a minimal sized inverter. The n1 transistor is either chosen to be minimal for a specific technology, or to be equivalent to a load having a tapering coefficient of $\beta = e' \approx 2.7$ (for minimum speed) or $\beta \approx 10$ (to optimize power) [6] for the input gate. The minimal delay is determined as a function of the number of stages of the HD buffer and of the n1 transistor size.

A sizing alternative for the driving transistors is to consider the final stage of the HD buffer to be the same size as for an equivalent tapered buffer. The HD buffer driving transistors of the remaining stages (the predriver) are sized so that each stage introduces the same delay (not equal to the delay introduced by the final stage). The minimal delay of the predriver is determined as a function of the number of stages of the predriver and of the size of the first stage. This alternative provides an improved speed/power/area trade-off than the previous sizing process.

The FNT are sized considering that if the input signal has a period T and a duty factor of 50%, then the nulling process must be performed in less than T/2 to ensure that the final driving transistors are not on at the same time. If the HD buffer has k stages, the time allocated to null each stage is $\approx T/2k$. Each nulling transistor is sized to perform the nulling process during the T/2k allocated time for the particular capacitive load.

The MNT are minimum sized and depend only on the leakage current characteristic and on the imposed noise immunity level. The chains of inverters driving the MNT are minimum sized. The chain of inverters driven by the output signal synchronously drives the FNT to minimize the short circuit and DC power. Each inverter in the chain introduces the same delay as the nulling process for one stage, which as shown is T/2k. Computing the capacitance at each node (consisting of the gate capacitances of an FNT and of the following inverter in the chain plus the junction capacitance present at that node), permits each of the inverters in the chain to be properly sized.

The output latch is sized based on the time during which C_L is floating (which is T/2 minus the total nulling time for all stages), and the leakage characteristics of the capacitive load. The latch is sized so that the leakage is minimized during the time that C_L floats.

The optimally sized HD buffer has less stages than a tapered buffer for a similar capacitive load. The larger the capacitive load, the larger the difference in the number of stages between the two buffer types, and the greater the speed and power savings of the HD buffer.

IV. SIMULATION RESULTS

Circuit simulations based on Cadence-Spectre and a 1.2 μm CMOS technology are described in this section. A minimum speed tapered buffer is compared in terms of delay, power, area, and transition time with an HD buffer sized as above, optimized for speed and power. The HD buffer optimized for power (called the same speed HD buffer, SHD) has the same delay as the equivalent tapered buffer. The HD buffer optimized for speed (called the predriver HD buffer, PHD) has the same area as the equivalent tapered buffer. A range of capacitive loads from 1 pF to 500 pF is considered in this discussion.

The delay as a function of the number of stages for an optimal speed tapered buffer and HD buffer driving a 500 pF load is shown in Figure 3. Note that for this load, the HD buffer is up to 2.2 times faster than the tapered buffer. Note also that the optimum size for the tapered buffer is ten stages, and as shown in Figure 3, between nine and thirteen stages for the HD buffer. Nine stages provides the least power and area.

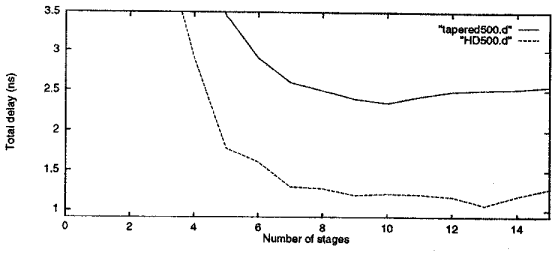

Fig. 3. Tapered buffer and HD buffer delay as a function of the number of stages. $C_L = 500\,pF$ and $n1 = 1.8\,\mu m$.

In Figs. 4 – 7, the delay, power, area, and transition time are compared. The performance of the SHD, PHD, and tapered buffers are ranked in Table I, from the point of view of these design criteria.

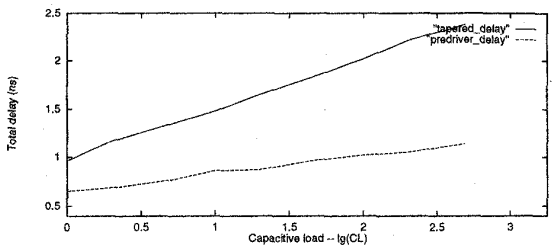

Fig. 4. Comparison of delay of tapered buffer and HD buffer. C_L=1 pF to 500 pF, plotted logarithmically.

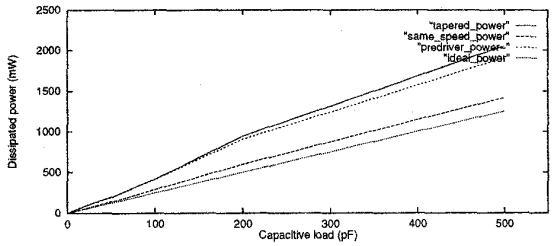

Fig. 5. Comparison of the power of the HD buffer and tapered buffer.

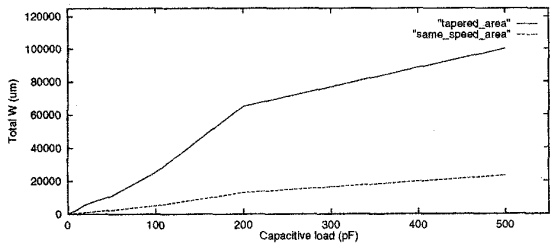

Fig. 6. Comparison of the area of the tapered buffer and HD buffer.

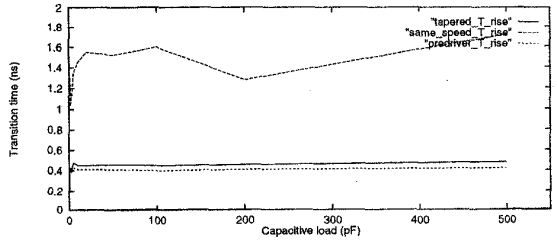

Fig. 7. Comparison of the output transition time of the tapered buffer and HD buffer.

Note that for the data presented in Figure 5, the ideal_power curve depicts the power dissipated exclusively on C_L. All other power curves include the power dissipated on C_L and within the buffer. Also note that for the SHD buffer, the total power dissipation is closest to the ideal_power. For a 500 pF load, the total power dissipated within the SHD buffer is $\approx 7\%$ of the power dissipated on C_L. Note also that the SHD buffer provides the same speed as a tapered buffer, and has the lowest power dissipation while saving up to 500% in area, but has the slowest output signal transition times. The PHD buffer offers up to a 2.2 times decreased delay, the same area, less than 10% power savings, and up to 10% faster transition times.

TABLE I
RANKING THE TAPERED, PHD, AND SHD BUFFERS IN TERMS OF SPEED, POWER, AREA, AND TRANSITION TIMES.

Buffer	Speed	Power	Area	T_{rise}
Tapered	2	3	2	2
PHD	1	2	2	1
SHD	2	1	1	3

V. Conclusions

A circuit structure called the High Drive (HD) buffer is proposed for driving large capacitive loads in high speed digital ASICs. If speed is the primary design objective, the PHD buffer is shown to provide up to a 2.2 times smaller delays, the same area, less than 10% power savings, and up to 10% faster transition times as compared to an equivalently tapered buffer. If power and area are the primary design objectives, the SHD buffer is shown to provide the same speed as an equivalently tapered buffer, the lowest power dissipation (close to the ideal $C_L V^2 f$ power), a savings of up to 500% in area, but with the slowest output signal transition times.

References

[1] T. Sakurai and A. R. Newton, "Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas," *IEEE Journal of Solid-State Circuits*, Vol. SC-25, No. 2, pp. 584–594, April 1990.

[2] H. C. Lin and L. W. Linholm, "An Optimized Output Stage for MOS Integrated Circuits," *IEEE Journal of Solid-State Circuits*, Vol. SC-10, No. 2, pp. 106–109, April 1975.

[3] N. Hedenstierna and K. O. Jeppson, "CMOS Circuit Speed and Buffer Optimization," *IEEE Transactions on Computer Aided Design*, Vol. CAD-6, pp. 270–281, March 1987.

[4] N. C. Li, G. L. Haviland, and A. A. Tuszynski, "CMOS Tapered Buffer," *IEEE Journal of Solid-State Circuits*, Vol. SC-25, pp. 1005–1008, August 1990.

[5] S. Dhar and M. A. Franklin, "Optimum Buffer Circuits for Driving Long Uniform Lines," *IEEE Journal of Solid-State Circuits*, Vol. SC-26, pp. 151–155, January 1991.

[6] B. S. Cherkauer and E. G. Friedman, "A Unified Design Methodology for CMOS Tapered Buffers," *IEEE Transactions on VLSI Systems*, Vol. VLSI-3, No. 1, pp. 99–111, March 1995.

[7] B. S. Cherkauer and E. G. Friedman, "Design of Tapered Buffers with Local Interconnect Capacitance," *IEEE Journal of Solid-State Circuits*, Vol. SC-30, No. 2, pp. 151–155, February 1995.

[8] J. M. Rabaey, *Digital Integrated Circuits, A Design Perspective*. Prentice-Hall, Inc, 1996.

[9] H.-Y. Huang and Y.-H. Chu, "Feedback controlled split-path CMOS buffer," *Proceedings of the IEEE International Symposium on Circuits and Systems*, Vol. 4, pp. 300–303, May 1996.

[10] D. Hodges and H. Jackson, *Analysis and Design of Digital Integrated Circuits*. McGraw-Hill, 1988.