MINIMIZATION OF CHIP SIZE AND POWER CONSUMPTION OF HIGH-SPEED VLSI BUFFERS

*D.Zhou** *X.Y.Liu*

The Department of Electrical Engineering
UNCC, Charlotte, NC 28223
Email: zhou@uncc.edu and xliu@uncc.edu
phone: (704) 547-4323
X.L. Wang
The Department of Physics Xian Jiaotong University
Xian, P.R. China

ABSTRACT

In this paper, we study optimal buffer design in high-performance VLSI systems. Specifically, we design a buffer for a given load such that chip area and power dissipation are minimal while circuit delay is no greater than a given upper bound. The explored direction, i.e., to minimize chip area and power consumption with circuit speed as a constraint, is a more realistic setting in practical VLSI design than conventional design objectives, where minimal circuit delay is usually sought. In fact, an optimal design must achieve an expected circuit speed with minimal system resources: chip area and power consumption. By solving the formulated constrained optimization problem, significant improvements in chip area and power consumption are achieved.

1. INTRODUCTION

Optimal buffer design is a fundamental and classical problem in VLSI design. To achieve the minimal delay, Mead and Conway in 1980 derived an "optimal" design where the buffer uses a circuit structure of cascaded inverters and the size-ratio between consecutive inverters takes the base value of the natural logarithm e [1]. The e size-ratio has been considered as "optimal" since then, and is widely cited in literature [4, 5]. In practical industrial design, the size-ratio is sometimes increased to a value between 3 and 5 in order to reduce the overall buffer size, a consequence of the reduced inverter stages [2]. A typical size-ratio 3.6 was proposed for an area efficient buffer design [16, 18]. Recently, the variable size-ratio was proposed to further improve the efficiency of buffers [15]. An example of using cascaded inverters to drive a capacitive load is shown in Figure 1, where the capacitor is used to model VLSI interconnects.

With a buffer consisting of cascaded inverters, most known design approaches use a constant size-ratio between consecutive inverters, and the size-ratio does not depend on the load [1, 16]. After the size-ratio is determined, the number of inverter stages is then related to the load. The main drawback of such approaches is that the circuit parameters are optimized in a separate way while they actually need to be considered together. Also, as noted in [15], the size-ratio should change with the stages to cope with the transistor's nonlinear property.

For high performance VLSI systems, not only minimal delay, but also minimal power dissipation and chip area are important design criteria. This paper studies the general relationships among these three objectives and optimizes them together. To properly formulate the optimal buffer design problem, we will first derive an analytic formula relating to circuit speed, buffer power dissipation, and chip size based on the transistor's nonlinear I-V characteristics. Using this formula, we further develop an optimization scheme to minimize the buffer size and power consumption with the circuit speed as a constraint. As a result, an optimal buffer is designed which runs faster than the e or 3.6 size-ratio buffers [16], consumes much less power and chip area. Figure 2 shows a practical design example where our optimal buffer not only has a shorter delay than that of the 3.6 and e size-ratio buffers, but also uses much smaller, less than one seventh, power and chip area.

This paper is organized as follows: Section 2. formulates the optimal buffer design problem. An analytic formula for the calculation of circuit delay is introduced, where both size-ratio and the number of inverter stages are variables. The closed form expressions for the buffer size and power consumption are also derived in this section. Based on the derived formulas, Section 3. studies the minimization of chip area and power consumption for a specified delay bound. Finally, Section 4. presents several design examples and Section 5. comments on the results and future researches.

2. CALCULATION OF DELAY, BUFFER SIZE AND POWER CONSUMPTION

Like the previous research [1, 2], this paper models a VLSI interconnect by a capacitor [1] and considers the buffer structure as cascaded inverters (Figure 1). CMOS technology is

* This research is supported in part by National Science Foundation NYI Award MIP-945402 and AirForce Office of Scientific Research grant F49520-96-1-0341.

[1] The result obtained based on the capacitive load model can be directly extended to the cases where the interconnect is modeled by the lumped RC or distributed RLC circuits, as briefly discussed in Section 5..

assumed in this paper [2]. The transistor size is measured by its gate area $W \times L$, where W and L are the channel width and length respectively. As a convention of IC design, all transistors have the same channel length L which is considered as a constant of the transistor's size optimization. Figure 3 shows a layout of four cascaded inverters using a constant e size-ratio between consecutive inverters.

For the circuit in Figure 1, let C_{g_i} be the gate capacitance, R_i be the output resistance, and τ_i be the delay of the ith inverter respectively. Let C_L be the load capacitance. With the simplest circuit model, the delay of the ith stage inverter is calculated by the product of its input capacitance C_{g_i} and the output resistance R_{i-1} of the previous stage inverter (Figure 4). We assume that the buffer always starts with a minimal sized inverter labeled as the 0th stage and has an output resistance R_s. The input to the 0th stage inverter is assumed to be a step function. The total delay of the buffer therefore is the sum of delays over all stages. That is,

$$\tau_{total} = \sum_{i=0}^{n-1} \tau_i = \sum_{i=0}^{n-1} R_i C_{g_{i+1}}, \tag{1}$$

where n is the number of the inverter stages, and $C_{g_n} = C_L$ is the load.

It is known that C_{g_i} and R_i are respectively proportional and inversely proportional to the transistor's size of the ith stage inverter. Usually, the linear region resistance of the transistor is used for R_i and the gate capacitance is used for C_{g_i} [4]. Under such an assumption, the R_i and C_i are written as

$$R_i = \frac{R_s}{W_i} \tag{2}$$

and

$$C_i = C_\square W_i \times L, \tag{3}$$

where W_i and L are respectively the channel width and length of the transistors in the ith stage inverter, and $C_\square$ is the unit square capacitance of the gate. In the rest of this paper, W_i represents the size of the n-channel transistor and the p-channel one is assumed to be properly sized to ensure symmetric signal rising and falling waveforms.

Let the variable size-ratio between the $(i+1)$th and ith stages be (Figure 1)

$$\frac{S_{i+1}}{S_i} = f(1+\alpha)^i \tag{4}$$

where both f and α are parameters to be determined by the optimization procedure. Following the same approach used in [15] the formula for calculating the circuit delay is obtained.

$$\begin{aligned} \tau_{total} &= \frac{f\tau_0((1+\alpha)^{n-1}-1)}{\alpha} \\ &+ \frac{R_s C_L}{f^{n-1}(1+\alpha)^{\frac{n^2-3n}{2}+1}} \end{aligned} \tag{5}$$

where τ_0 is the delay of the 0th stage minimal-size inverter, and n is the number of the inverter stages. Once the load capacitance C_L is given a minimal delay buffer can be designed using Eq. 5 by optimizing the parameters f, α and n [15]. Note that in Eq. 5 the size-ratio, number of inverter stages, and load capacitance all appear in one single equation calculating the delay, instead of being considered separately as in [1, 2, 5, 16]. It was shown in [15] that the optimization based on Eq. 5 leads to a much improved buffer design in terms of circuit speed, chip size and power consumption.

While all existing formulations of the optimal buffer design try to inimize the circuit delay, the resulted buffer size or power consumption have been generally ignored. As will be shown in the next section, a significant improvement can be made on the buffer size (or power) when delay is formulated as a constraint and the size (or power) as the optimization objective. To calculate the buffer size, we use Eq. 4 and have

$$\begin{aligned} S_{total} &= \sum_{i=0}^{n-1} S_i \\ &= \sum_{i=0}^{n-1} f^{i-1}(1+\alpha)^{\frac{i^2-i}{2}} S_o \\ &= \frac{\sqrt{\pi/2}(A-B)S_o}{\sqrt{\ln(1+\alpha)}\exp\left(\frac{\ln^2(f^2(1+\alpha))}{8\ln(1+\alpha)}\right)} \end{aligned} \tag{6}$$

where

$$A = \mathrm{Erfi}\left(\frac{\ln(f^2(1+\alpha)^{2n-1})}{2\sqrt{2\ln(1+\alpha)}}\right), \tag{7}$$

$$B = \mathrm{Erfi}\left(\frac{\ln(f^2(1+\alpha))}{2\sqrt{2\ln(1+\alpha}}\right), \tag{8}$$

and $Erfi(x)$ is an imaginary error function.

The energy used to drive a capacitor is the energy used in the charging/discharging process [3]. A transistor of size S_i has the input capacitance $C_{g_i} = C_\square S_i$. Therefore, the energy needed to charge C_{g_i} is

$$P_i = \frac{1}{2} C_{g_i} V_{dd}^2 = \frac{1}{2} V_{dd}^2 C_\square S_i \tag{9}$$

where V_{dd} is the supply voltage. The total energy is

$$P_{total} = \sum_{i=0}^{n-1} P_i = \frac{1}{2} V_{dd}^2 C_\square \sum_{i=0}^{n-1} S_i = \frac{1}{2} V_{dd}^2 C_\square S_{total} \tag{10}$$

From Eq. 10 it is easily seen that the difference between P_{total} and S_{total} is just a constant factor. Therefore, the optimization of the buffer size and its power consumption are equivalent. Without loss of generality, we only consider the former case in the following discussions.

[2] The other technologies can be discussed similarly.

[3] The inverter's transient power dissipation is not considered here to simplify the discussion. To verify this simplification, a SPICE simulation result is shown in Section 4..

3. OPTIMAL BUFFER DESIGN

We now study the problem defined as follows: *For a given load C_L and a specified upper bound on delay, determine the transistor's size of each inverter and the number of inverter stages such that the chip area S_{total} or the total power dissipation P_{total} of the buffer is minimal.*

Let τ_u be the upper bound on circuit delay. The minimal size buffer design with the specified upper bound on delay can thus be formulated as

$$\begin{aligned} \min \quad & S_{total}(\alpha, f, n) \\ s.t. \quad & \frac{f\tau_0((1+\alpha)^{n-1}-1)}{\alpha} + \frac{R_s C_L}{f^{n-1}(1+\alpha)^{\frac{n^2-3n}{2}+1}} \\ = \quad & \tau_u \end{aligned} \tag{11}$$

Considering Eq. 11 as an optimization constraint, an augmented Lagrangian function can be introduced and it can be solved by using a standard method [17]. Specifically, an unconstrained objective function is introduced

$$\begin{aligned} S_L &= S_{total}(\alpha, f, n) + \lambda \frac{f\tau_0((1+\alpha)^{n-1}-1)}{\alpha} \\ &+ \lambda\left(\frac{R_s C_L}{f^{n-1}(1+\alpha)^{\frac{n^2-3n}{2}+1}} - \tau_u\right) \end{aligned} \tag{12}$$

where λ is a Lagrangian parameter. Taking the derivatives of S_L with respect to the parameters f, α, n and λ, and setting each of them to zero we can determine their optimal values.

$$\begin{aligned} \frac{\partial S_L}{\partial \lambda} &= \frac{f\tau_0((1+\alpha)^{n-1}-1)}{\alpha} + \frac{R_s C_L}{f^{n-1}(1+\alpha)^{\frac{n^2-3n}{2}+1}} - \tau_u \\ &= 0 \end{aligned} \tag{13}$$

$$\begin{aligned} \frac{\partial S_L}{\partial n} &= \exp\left(\frac{\ln(f^4(1+\alpha)^{2n+1})\ln(1+\alpha)^{2n-3}}{8ln(1+\alpha)}\right) \\ &+ \lambda\left(\frac{f\tau_0(1+\alpha)^{n-1}\ln(1+\alpha)}{\alpha}\right. \\ &- \left.\frac{R_s C_L((n-3/2)\ln(1+\alpha)+\ln f)}{f^{n-1}(1+\alpha)^{\frac{n^2-3n}{2}+1}}\right) \\ &= 0 \end{aligned} \tag{14}$$

$$\begin{aligned} \frac{\partial S_L}{\partial f} &= \frac{\exp(\frac{\ln^2(f^2(1+\alpha)^{2n-1})}{8ln(1+\alpha)}) - \exp(\frac{\ln^2(f^2(1+\alpha))}{8ln(1+\alpha)})}{fln(1+\alpha)\exp(\frac{\ln^2(f^2(1+\alpha))}{8ln(1+\alpha)})} \\ &+ \frac{\sqrt{\pi/2}\ln(f^2(1+\alpha)^2)(C-D)}{2f\ln^{3/2}(1+\alpha)\exp(\frac{\ln^2(f^2(1+\alpha)^2)}{8ln(1+\alpha)})} \\ &+ \lambda\left(\frac{\tau_0((1+\alpha)^{n-1}-1)}{\alpha} - \frac{(n-1)R_s C_L}{f^n(1+\alpha)^{\frac{n^2-3n}{2}+1}}\right) \\ &= 0 \end{aligned} \tag{15}$$

where

$$C = Erfi\left(\frac{\ln(f^2(1+\alpha))}{\sqrt{8\ln(1+\alpha)}}\right), \tag{16}$$

$$D = Erfi\left(\frac{\ln(f^2(1+\alpha)^{2n-1})}{\sqrt{8\ln(1+\alpha)}}\right), \tag{17}$$

and

$$\begin{aligned} \frac{\partial S_L}{\partial \alpha} &= \lambda\left(\frac{f\tau_0(n-1)(1+\alpha)^{n-2}}{\alpha}\right. \\ &- \frac{f\tau_0((1+\alpha)^{n-1}-1)}{\alpha^2} \\ &- \left.\frac{(n^2-3n+2)R_s C_L}{2f^{n-1}(1+\alpha)^{\frac{(n^2-3n)}{2}+2}}\right) \\ &+ \frac{\sqrt{\pi/2}(E-F)}{2(1+\alpha)\ln^{3/2}(1+\alpha)\exp(\frac{\ln^2(f^2(1+\alpha)^2)}{8ln(1+\alpha)})} \\ &+ \frac{\sqrt{\pi/2}\ln(f^2(1+\alpha)^2)\ln(\frac{f^2}{(1+\alpha)^2})(G-H)}{8(1+\alpha)\ln^{5/2}(1+\alpha)\exp(\frac{\ln^2(f^2(1+\alpha)^2)}{8ln(1+\alpha)})} \\ &+ \frac{\sqrt{\pi/2}\exp(\frac{\ln(f^2(1+\alpha)^2)}{8ln(1+\alpha)})\ln(\frac{f^2}{1+\alpha})}{\sqrt{8\pi}(1+\alpha)ln^2(1+\alpha)\exp(\frac{\ln^2(f^2(1+\alpha)^2)}{8ln(1+\alpha)})} \\ &+ \frac{\sqrt{\pi/2}\exp(\frac{\ln(f^2(1+\alpha)^{2n-1})}{8ln(1+\alpha)})\ln(\frac{(1+\alpha)^{2n-1}}{f^2})}{\sqrt{8\pi}(1+\alpha)ln^2(1+\alpha)\exp(\frac{\ln^2(f^2(1+\alpha)^2)}{8ln(1+\alpha)})} \\ &= 0 \end{aligned} \tag{18}$$

where

$$E = Erfi\left(\frac{\ln(f^2(1+\alpha))}{\sqrt{8\ln(1+\alpha)}}\right), \tag{19}$$

$$F = Erfi\left(\frac{\ln(f^2(1+\alpha)^{2n-1})}{\sqrt{8\ln(1+\alpha)}}\right), \tag{20}$$

$$G = Erfi\left(\frac{\ln(f^2(1+\alpha)^{2n-1})}{\sqrt{8\ln(1+\alpha)}}\right), \tag{21}$$

$$H = Erfi\left(\frac{\ln(f^2(1+\alpha))}{\sqrt{8\ln(1+\alpha)}}\right), \tag{22}$$

Using the optimal parameter values we finally have the minimal buffer size for the given delay bound.

4. DESIGN EXAMPLES

In this section, we design several optimal buffers with a minimal chip area (power) for the given delay bound, including the minimal delay case. A wide range of loads, $\frac{C_L}{C_\square} = 300$, $\frac{C_L}{C_\square} = 10^3$ and $\frac{C_L}{C_\square} = 2 * 10^4$, are considered to model the short distance on-chip, medium distance on-chip, and long distance (inter-chip or pad to packaging) interconnects, respectively. Four different design approaches are compared: the traditional e size-ratio, modified 3.6 size-ratio [16], minimal delay with variable size-ratio [15], and minimal size (power) with the constrained delay bound. While the buffer is designed with one of these four methods, the actual delay

data in the figures and table are obtained from the SPICE simulation.

We first minimize the buffer size (power) with a given upper bound on delay. The results are shown in Figures 2, 5 and 6 respectively, where for each delay we designed a buffer and calculated its size. Over the whole range of loads, our design offers a much smaller chip area (power) than the other existing methods. The advantage of the proposed design method stands out sharply when driving heavy load (long distance interconnects $\frac{C_L}{C_\square} = 2 * 10^4$). An average about one order saving in area is achieved in this case.

In the previous section, we assumed a linear relationship between the power consumption and the buffer size. Based on this assumption the optimization of the buffer size and the power consumption becomes equivalent. To verify the validity of the assumption, we plotted power consumption obtained from SPICE simulation against the buffer size in Figure 7. It can be seen that the assumption holds well.

In many cases, the minimal delay is sought in buffer design. In Figure 8 (Table I), we compare the design methods of minimal delay with variable size-ratio, and minimal size (power) with delay constraints. [4]. For various loads, the minimal delay buffers are designed to test the performance of two methods at the extreme case. As can be seen from the data, the method proposed in this paper produces a better result over the minimal delay approach proposed in [15]. The data demonstrate that one can always set the circuit speed as a constraint and optimize the other design parameters. Such a formulation generally offers a better design result.

Figure 9 plots the buffer size against delay, where for each given delay bound, two buffers are designed using the optimal and the e size-ratio methods, respectively. Note that the curve generated from the optimal design always runs below that from the e size-ratio one. The difference between two methods is getting bigger as approaching the minimal delay. The curves rise sharply at the neighborhood of the minimal delay. The chip size and power consumption are very sensitive in this region and therefore, their optimization can be very effective.

Finally, we use our optimal buffer to drive a clock net in a real signal processing chip. The chip contains 160k transistors and its layout is shown in Figure 10. Again, for this real example, our design out-performed the existing methods (Figure 11).

5. DISCUSSION AND FUTURE RESEARCH

The minimal size (power) buffer with a constrained delay bound was formulated in this paper. Lagrangian method was used to find the optimal design. Real design examples demonstrated the correctness and effectiveness of the proposed method. An average of almost one order saving in chip area (power consumption) was achieved for driving large loads, mainly due to the new formulation of the problem.

[4] The e and 3.6 size-ratio buffer are not compared here for they are much more inferior to those listed here.

REFERENCES

[1] Mead, C. and Conway, L., Introduction to VLSI systems, Reading, MA: Addison-Wesley, 1980.

[2] H.B. Bakoglu, Circuits, Interconnections and packaging for VLSI, Addison-Wesley, 1990.

[3] D. Zhou, F.P. Preparata, and S.M. Kang, "Interconnection delay in Very High-Speed VLSI," IEEE Trans. Circuits Systems, VOL. 38, 1991.

[4] Douglas A. Pucknell, Kamran Eshraghian, Basic VLSI design, systems and circuits, 1989

[5] N. H. E. Weste, Kamran Eshraghian, Principles of CMOS VLSI design, A systems perspective, Addison-Wesley, 1988.

[6] D. Zhou, F. Tsui, D.S. Gao and J.S. Cong, "A distributed-RLC model for MCM layout," Proc. IEEE Multichip Module Conf., 1993.

[7] D. Zhou, S.Su and F.Tsui, D.S. Gao, J.S. Cong, "A simplified synthesis of transmission lines with a tree structure," Analog Integrated Circuits and Signal Processing," An International Hournal,pp15-30, 1993

[8] Chen-Ping Yuan, "Modelling and extraction of interconnection P parameters in VLSI", 1983

[9] J. Cong, S.Su and C.-K. Koh, "Simultaneous buffer and wire sizing for performance and power optimization," IEEE Trans. Very Large Scale Integrated Systems, Vol. 2, No. 4, pp 408-425, Dec. 1994

[10] F.N. Najm, "A survey of power estimation techniques in VLSI circuits," IEEE Trans. Very Large Scale Integrated Systems, Vol. 2, No. 4, PP.446-455, Dec. 1994

[11] A.P. Chandrakasan, S. Sheng, and R. W. Brodersn, "Low-power CMOS digital design," IEEE J. Solid-State Circ., Vol. 27, No. 4, PP. 473-484, April 1992

[12] S. Devadas, K. Keutzer, and J. White, "Estimation of power dissipation in CMOS combinational circuits using Boolean function manipulation," IEEE Trans. Computer-Aided Design, Vol. 11, No. 3, PP. 373-383, March 1992

[13] U. Jagau, "SIMCURRENT – An efficient program for the estimation of the current flow of complex CMOS circuits," IEEE Int. Conf. Computer-Aided Design, Santa Clara, CA, Nov. 11-15, 1990, pp. 396-399

[14] M. A. Cirit, "Estimaing dynamic power consumption of CMOS circuits," IEEE Int. Conf. Computer-Aided Design, Nov. 9-12, pp. 534-537, 1987

[15] D. Zhou and X.Y. Liu "On the Optimal Drivers of High-Speed Low Power ICs", to appear in International Journal of High-speed Electronics and Systems, Vol. 7, No. 2, June 1996.

[16] Jan. M. Rabaey, Digital Integrated Circuits, A Design Perspective, Prentice Hall, 1996.

[17] Philip E. Gill and et. al., "Practical Optimization", Academic Press, 1981.

[18] N. Hedenstiern et al., "CMOS Circuit Speed and Buffer Optimization", IEEE Trans. on Computer-Aided Design, Vol. CAD-6, No. 2, pp.270-281, March 1987.

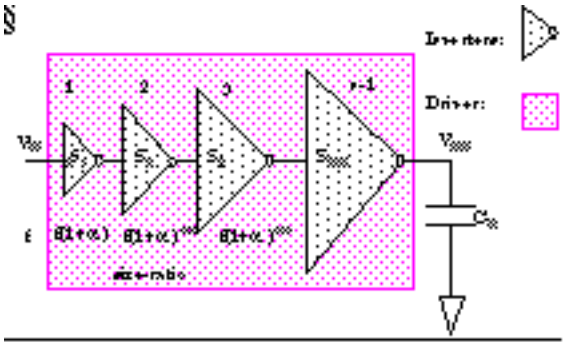

Figure 1: A buffer consists of cascaded inverters

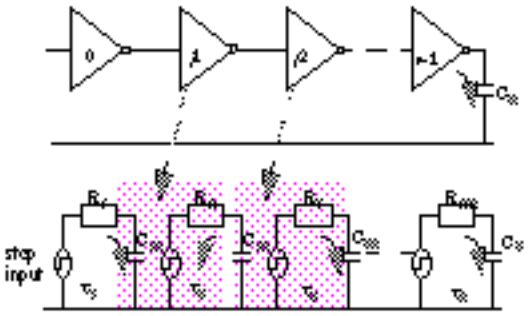

Figure 4: Cascaded inverters and their simple circuit model

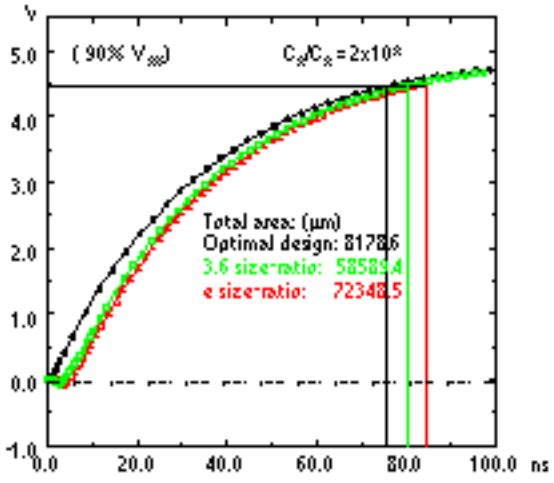

Figure 2: Comparison of different buffer design methods for long distance interconnect

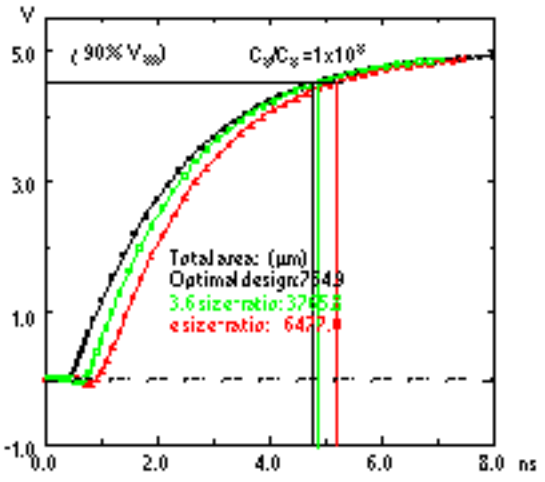

Figure 5: Comparison of different buffer design methods for medium distance interconnect

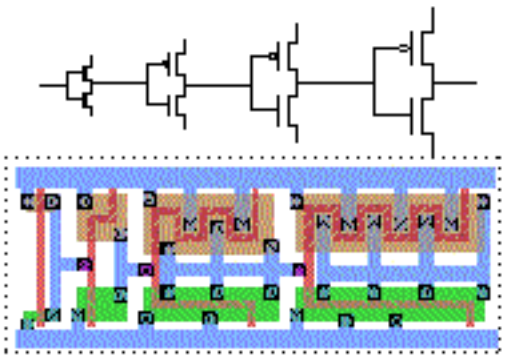

Figure 3: CMOS inverters and their layout

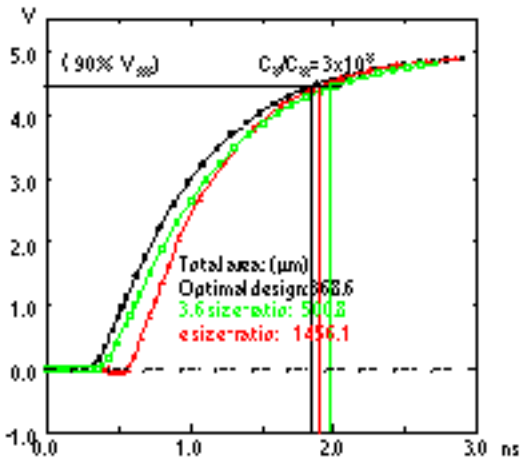

Figure 6 : Comparison of different buffer design methods for short distance interconnect

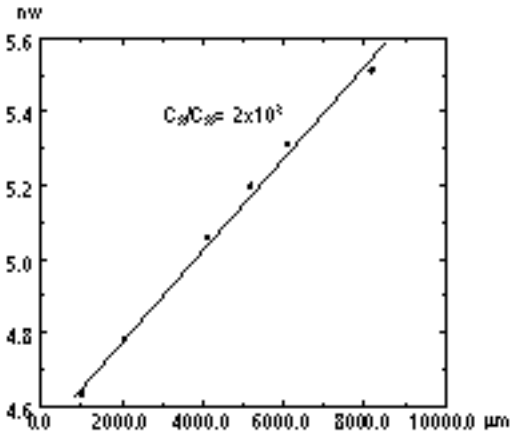

Figure 7: Power consumption vs buffer size

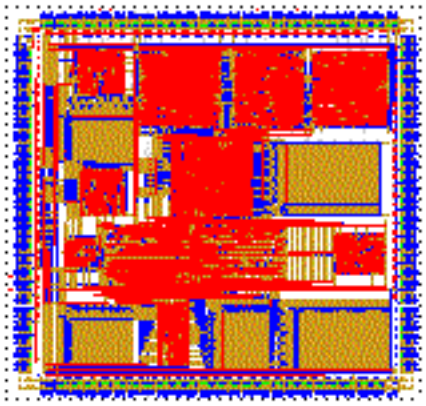

Figure 10: The layout of a real signal processing chip

Table I

C_g/C_{gs}	τ_{opt}(ns)	S_{opt}(μm)	S_{opt}(μm)	S_{opt}
2x10^3	78.0	11913.3	8178.63	31.1%
5x10^3	20.5	1542.3	1224.0	20.6%
3x10^3	1.9	492.3	368.6	25.0%

Figure 8:Minimal size vs minimal delay buffer design methods

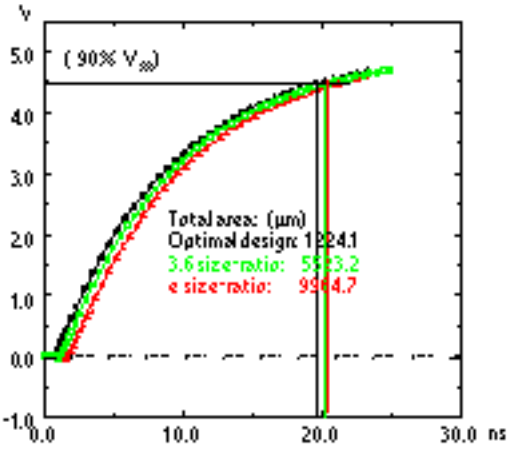

Figure 11: The signal response of the clock network driven by different buffers

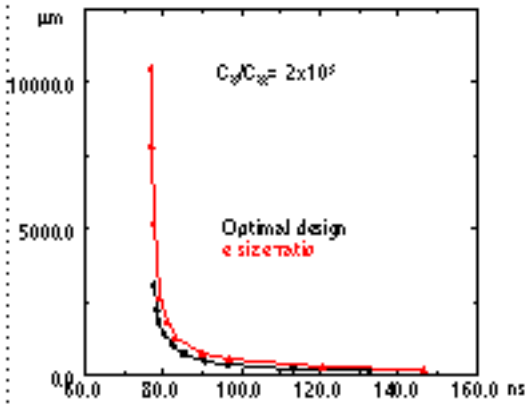

Figure 9: Minimal size buffer design with delay constraints