Energy Efficient Signaling in Deep Submicron CMOS Technology

Imed Ben Dhaou, Vijay Sundararajan, Hannu Tenhunen* and Keshab K. Parhi!

*Electronics System Design Laboratory, Dept. of Electronics
KTH-Electrum, Electrum-229
Royal Institute of Technology,
SE-16440 Kista, Sweden
E-mail: imed@ele.kth.se

!Dept. of ECE, University of Minnesota,
Minneapolis, 55455
E-mail: parhi@ece.umn.edu

Abstract

In this paper we propose an efficient technique for energy savings in DSM technology. The core of this method is based on low-voltage signaling over long on-chip interconnect with repeaters insertion to tolerate DSM noise and to achieve an acceptable delay. We elaborate a heuristic algorithm, called VIJIM, for repeaters insertion. VIJIM algorithm has been implemented to design a robust inverter chain for on-chip signaling using 0.25μm, 2.5V, 6 metal layers CMOS process. An average of 70% of energy-saving has been achieved by reducing the supply voltage from 2.5V down to 1.5V.

1 Introduction

With the advance of technology, gate length of a few tens of a nanometer for CMOS has become a distinct possibility, also the amount of transistors in one single die is increasing steadily over time towards ULSI or GSI integration level. This increase in integration has created a lot of problems e.g., packaging noise and power dissipation. Thus power savings and noise tolerance have emerged as important factors to consider at all design abstraction levels. Power savings have been a major concern within the circuit designer community in submicron CMOS technology. So far, the most effective method of lowering power dissipation has been scaling down of the gate supply voltage [4]. This results in a speed degradation of the gates. In order to restore the speed, techniques like retiming, pipelining and parallel processing have been widely deployed with the main focus on gate level parameters (capacitance, delay, and size). However, in DSM technology, it has been reported in many papers [2] that the power dissipated due to interconnects (delay as well) has become several orders of magnitude larger than that caused by unloaded gates. Therefore, energy saving techniques focusing only on gate power optimization are not adequate in DSM technology.

In order to tackle these problems i.e. noise, delay and power in DSM, our approach to design a power-efficient, robust signaling techniques that allow for reliable communication between cells. This will be referred to as energy-efficient on-chip signaling.

The rest of the paper is organized in the following way. In Section 2 we formulate the problem and we put our work into perspective with existing approaches. In Section 3 we derive VIJIM heuristic algorithm for buffer insertion. In Section 4 we present the experimental results. Finally in Section 5 we summarize our paper.

2 Power savings in DSM technology

2.1 Problem formulation

Let a CMOS gate, referred to as a transmitter, sending signals to a receiver situated at a distance \( d \) from the transmitter. The receiver is a CMOS gate. The medium used for signaling between the transmitter and the receiver is the On-Chip Interconnect (OCI)\(^1\). In order to define the efficiency of the communication between the transmitter and the receiver, we use the following definition.

\(^1\)OCI is sometimes referred to as a net
Definition 1 Let us denote by \( s_t(t) \) the transmitted signal, \( s_r(t) \) the received signal and by \( \delta \) the propagation delay of the signal \( t \) sent over a distance \( d \), where \( d \) is the distance between the receiver and the transmitter. The communication between the transmitter and the receiver, is said to be perfect if and only if the received signal, \( s_r(t) \), at the instant of time \( t \) sent by the transmitter at \( t - \delta \) is within the noise margin of the receiver satisfying the following inequalities: 
\[
|s_t(t - \delta) - s_r(t)| \leq NM_H \text{ if } s_t(t - \delta) \text{ is a logic high or}
\]
\[
|s_t(t - \delta) - s_r(t)| \leq NM_L \text{ if } s_t(t - \delta) \text{ is a logic low [1]}
\]

In order to achieve a perfect communication, a priori knowledge and characterization of the OCI is required. The characterization of the net is a challenging task that requires the use of a multi-dimensional field solver which is computationally appealing. In order to get around this problem, OCI, in submicron technology, has been approximated by lumped RC elements. However, as in DSM technology, it was found that the inductance, \( L \), started to be a non-negligible value, due to the fast fall and rise time of the signal[8]. As a result, OCI needs to be modeled as a lumped RLC circuits.

As the technology scales down (towards DSM and Ultra DSM), the transistor feature size is decreasing (width, gate-oxide, channel length) and the number of metalization levels is increasing. As a result, the spacing between metal layers is decreasing, and thus, the parasitics elements \( R, L, \) and \( C \) are increasing.

Consequently, power dissipated to send the signal over a distance \( d \) is becoming larger than the power consumed by the signaling gate. In addition, the delay caused by the OCI is becoming dominant compared to the delay caused by the input capacitance of the receiving gate. In most cases, the delay, denoted by \( \Delta \), is a design parameter that needs to be satisfied. If we denote by \( \xi(d, \Delta) \) the required energy, to send the data over a distance \( d \) under a specific delay \( \Delta \) then power saving in DSM can be summarized into the following optimization problem.

\[
\text{OPT-SIG} \\
\begin{align*}
\text{minimize } & \xi(d, \Delta) \\
\text{subject to,} & \\
& \delta \leq \Delta \\
& \text{a perfect communication}
\end{align*}
\]

Figure 1. Problem formulation for energy-efficient signaling in DSM

2.2 Related work

Traditionally, buffer\(^2\) insertion has mainly been done to distribute the clock network at a tolerable skew or to minimize the delay when a gate is driving a high load capacitance [4] using full-swing signals. Recently, many published reports have considered the problem of low-swing signaling with repeater insertion. In [3], algorithm for clock network distribution using low-voltage signaling combined with repeater insertion has been reported. However, in his model, the OCI has been modeled as a lumped RC circuit. In addition, the switching noise has been totally ignored. It has been found in [5] that up to 40% improvement in delay is achievable if OCI is modeled as lumped RLC compared to the case if it is modeled as lumped RC. However, buffer resizing for optimal delay has not been considered. Moreover, the location of the buffer has been computed without taking into consideration DSM noise. Also, power saving has not been addressed. In order to tackle DSM noise for buffer insertion, in [9] an algorithm for buffer insertion for delay and noise optimization has been reported. However power saving has not been addressed. Furthermore, buffer resizing has not been considered, instead the algorithm requires a library of buffers.

In order to address all of the problems neglected in the mentioned work, we describe a heuristic algorithm, VIJIM, that estimates the location of the buffer and resizes it in order to solve the optimization problem given in Fig.1. In order to minimize \( \xi(d, \Delta) \), our approach is based on low-swing signaling. VIJIM algorithm is used to locate the position, number and sizes of the buffers. This approach can be seen as a generalization to the algorithm proposed in [7]. However, as we will see in Section 3, the algorithm, used to locate the buffers, partitions OCI into \( n + 1 \) equidistant nodes, as a result, it inherently assumes that the noise at each node segment is within the noise margin of the receiving gate. In addition, to the best of our knowledge, there is no efficient algorithm that estimates \( n \). As an example, Fig.2 shows two coupled OCI before and after VIJIM algorithm. The figure contains a quiet node (a) and an active node (b) of length \( d = 0.5\mu m \). The crosstalk noise depends on the ratio of crosstalk capacitance to the self capacitance and the interconnect length (\( d \)). In full custom design, crosstalk capacitance depends on the routing algorithm (cf. see e.g. [13]). In Fig.2, we have assumed that the crosstalk capacitance \( (C_v) \) equals 0.624pf/mm and the self capacitance \( (C_e) \) equals 1.56pf/mm. These values were obtained using LINPAR [14]. The waveforms for \( v(ar) \) and \( v(arb) \) are depicted in Fig.3. The buffers have been implemented using 0.25\( \mu \)m, 2.5V, 6 metal layers CMOS process. Fig.3 shows that thanks to an inserted buffer, the amplitude of the coupled noise (\( v(ar) \)) has been reduced from 1.22V...
Figure 2. A noise tolerance scheme using buffers.

Figure 3. Waveforms of v(ar) and v(arb) to 0.74V. In both cases (before and after buffer insertion), the delay of v(b) is 0.77 nanosecond and 0.841 nanosecond, respectively. That means if the desired delay of the signal v(b) is 1 nanosecond then buffer insertion algorithm for optimum delay (e.g. MZhou) sees no reason for buffer insertion. However, from noise point of view, a buffer must be inserted in order to guarantee a perfect communication. Furthermore, crosstalk noise causes false switching of the gates connected to the net a. Consequently, the power consumption increases due to unnecessarily switching.

The example, depicted in Fig.3, clearly shows that buffer insertion for delay optimization does not guarantee a perfect communication. This case can be generalized in the following Lemma.

Lemma 1 A net that has been optimally buffered to minimize delay alone may be susceptible to noise variation.

Proof 1 The proof is given in [9]. □

3 VIJIM algorithm

The concept of geometric ratio sizing is a well known design technique for choosing buffer sizes. In [4] a technique for choosing the number and sizes of inverters, to drive an interconnect wire at an optimal delay, has been demonstrated. Using simplified inverter charging/discharging models it was concluded in [4] that successive inverters driving the interconnect line should be sized up in a geometric progression. According to this technique, assuming an inverter of size $S$ is to drive a load of capacitance $C_L$, then the first driver/inverter of size say $(S)$ is followed by uniformly spaced inverters of size $eS, e^2 S, e^3 S, ..., e^n S$. Where $n = \log_2 \left( \frac{C_L}{C_g} \right)$, $C_g$ is the gate capacitance of a minimum sized inverter. It was later proved in [7] that this technique cannot achieve an optimal delay. By taking into account the slope of the input signal as well as the RC interconnect models and more exact behavior of inverter charging/discharging equations, D.Zhou et al. have found that in order to achieve an optimal delay, the inverters need to be sized with a pseudo-fixed ratio. The pseudo ratio should be of the form $a \times (1 + \epsilon)^i$ for sizing stage $i$. This implies that the inverter sizes will vary as, $S, a(1 + \epsilon)S, a^2(1 + \epsilon)S, a^3(1 + \epsilon)S, ... , a^n(1 + \epsilon)^{n(n+1)/2}S$. If the length of the interconnect is $d$ then the inverters need to be located with uniform spacing $\frac{d}{n}$. In [7] a detailed technique was then developed for minimal powered chained-driver configuration for achieving a specified delay. While this is our objective in this paper, we note that the inductive effects and noise for buffer insertion have been totally ignored in [7]. This has motivated us to present VIJIM algorithm for developing a robust chain of inverters scheme for energy-efficient signaling with interconnects that exhibit RLC behavior under DSM noise constraints. Before we introduce VIJIM algorithm we will first present an algorithm for low power signaling over interconnects exhibiting RLC behavior and meeting a specific delay requirement. This technique ignores the noise characteristics of the interconnect.

The expression for the power consumed by the interconnect and the load between the $i^{th}$ and $(i+1)^{th}$ driver is denoted by $pow(i)$, and it is a function of several parameters such as interconnect length, driver size, load size etc. We assume that the first driver is of size $S$ and the ratio of the size of the $i^{th}$ driver to the size of the $i-1^{th}$ driver is $\alpha(1 + \epsilon)^i$. The exact value of $\alpha$ and $\epsilon$ will be determined by the optimization heuristic that will be referred to as the Modified ZHOU (MZHOU) heuristic, given in Fig.4. The MZHOU heuristic inherently assumes that an interconnect...
of length $d$ is divided in $n$ equi-distant sections, where $n$ is a parameter to be determined by the above optimization problem. MZHOU algorithm assumes that the noise, at each node situated at distances $\frac{d}{n}$, is within the noise margin of the circuit. Thus, MZhou doesn’t guarantee a perfect communication (cf. see the constraints given in Fig.1). A robust algorithm for buffer insertion in DSM should solve the optimization problem given in Fig.5. One of the possible solution of VIJIM-OPT is achievable by VIJIM algorithm. Its pseudo code is given in Fig.6. VIJIM algorithm relies on a well characterized OCI, an accurate delay estimation and an accurate cross-talk noise estimation algorithms. These functions are given as an argument to VIJIM algorithm: $f_{\text{compute}}(\text{delay},\text{OCI},a,b)$ and $f_{\text{compute}}(\text{noise},\text{OCI},b)$. An efficient algorithm for crosstalk estimation is given in [10].

VIJIM algorithm works in the following way. At the $i^{th}$ iteration, VIJIM algorithm uses a fast searching technique to locate the ”best” position of the $i^{th}$ buffer such that the received signal from the $(i-1)$ buffer $V_{i-1}$ satisfies the perfect communication constraints (cf. see Definition1). The location of the buffer is determined within a given precision($\sigma$) which can be set at the onset of VIJIM. The convergence speed4 of VIJIM depends on the value of $\gamma$.

In other words, for a small value of steps VIJIM estimates the location of the $i^{th}$ buffer with a faster convergence time provided the location of $b_{opt}$ is closer to the initial position of $b$ than to $a$.

4 Experimental results

VIJIM algorithm has been implemented to design a robust chain of inverters implemented using 0.25$\mu$m, 6 metal layers, 2.5V CMOS process. The OCI model has been simplified to a first order RLC circuit. The transfer function, $H(s)$, of an OCI of length $d$ is given in the following equation.

$$H(s) = \frac{V_{out}(s)}{V_{in}(s)} = \frac{1}{d^2 s^2 LC + s RC + 1},$$  \hspace{1cm} (1)$$

where $V_{in}(s) = L(v_{in}(t))$ is the input voltage in s-domain, $V_{out}(s) = L(v_{out}(t))$ is the output voltage in s-domain, $s = j2\pi f$ is Laplacian variable, and $f$ is the operating frequency. From the above equation it is clear that as the wire-length increases ($d$ increases), the transfer function decreases, and thus the output-voltage decreases. This has motivated us to use the transfer function to measure the delay of the OCI. The algorithm used to verify if an intercon-
nect satisfies the delay requirements is given in the following Lemma.

**Lemma 2** Given an interconnect approximated as a first order RLC circuit that has a transfer function \( H(s) \) and a delay \( \tau_{oci} \). There exist a parameter \( \beta \) such that \( |H(\frac{2\pi}{\tau_{oci}})| = \beta \). The delay is then given by

\[
\tau_{oci} = \frac{2\pi}{\sqrt{\max\left(\frac{-b^2 - 2a^2\gamma}{2\pi^2}, \frac{-b^2 - 2a^2}{2\pi}\right)}}
\]  

(2)

where \( \gamma = \sqrt{(b^2 - 2a)^2 - 4a^2(1 - 1/\beta)} \), \( b = d^2 LC \), and \( a = d^2 RC \). Let a given \( \tau_{specified} \) close to \( \tau_{oci} \). If \( |H(\frac{2\pi}{\tau_{specified}})| \geq \beta \) then \( \tau_{specified} \leq \tau_{oci} \).

**Proof 2** The proof is given in Appendix-A.

The comparison between \( \tau_{oci} \) given in Eq.2 and HSPICE is depicted in Fig.7. OCI parameters are: \( L = 0.32\Omega\text{H/mm}, R = 1.59\Omega\text{/mm}, \) and \( C = 0.16\text{pF/mm} \). These parameters, obtained using a commercial software called LINPAR[6], are reasonably close to the measured parameters published in [11]. In VIJIM algorithm, the value of \( \beta \) is equal to 0.8. The comparison of VIJIM implemented using the simplified delay model and Sakurai’s delay model [12] (referred to as SVIJM), is shown in Tab.1.

<table>
<thead>
<tr>
<th>Param.</th>
<th>Wire length (d) in mm</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d [\text{mm}] )</td>
<td>1</td>
</tr>
<tr>
<td>( \tau [\text{ns}] )</td>
<td>2.82</td>
</tr>
<tr>
<td>( P [\text{mW}] )</td>
<td>0.27</td>
</tr>
<tr>
<td>( \xi_{SV}[\text{pJ}] )</td>
<td>0.16</td>
</tr>
<tr>
<td>( \xi_{VI}[\text{pJ}] )</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table 1. Comparison of VIJIM and SVIJIM for the case of \( f = 600 MHz \)**

VIJIM algorithm has been also compared to MZHou algorithm. The results are summarized in Tab.2. The results show that VIJIM algorithm has better energy-saving factor compared to MZHou algorithm.

By using linear regression techniques, depicted in Fig.8, the relation between \( d \) and \( \xi \) is given in Eq.3.

\[
\xi_V = 54e^{-4} \cdot d^3 - 0.07 \cdot d^2 + 22.81e^{-2} \cdot d^2 + 0.89 \cdot d - 0.38
\]

(3)

where \( \xi_V \) is the energy, in pJ, needed to transmit a signal over a distance \( d \) in mm. As expected, Eq.3 shows that the energy-consumption for on-chip signaling, \( \xi(d, \Delta) \) (cf. see Fig.1), is proportional to the wire-length. In order to minimize \( \xi(d, \Delta) \), our approach is to select the optimum supply voltage from a given set of possible supply voltages. In our experiments, we are given three supply voltages: 2.5V (full-swing), 1.8V, and 1.5V. The results, given in Tab.3, show that 1.5V has the lowest \( \xi(d, \Delta) \). This observation rises the following question: why don’t we just use the lowest supply voltage that guarantees a perfect communication?. In order to see this, we have conducted the same experiments using supply voltage equal to 1V. The results, tabulated in Tab.3, show that \( \xi_{15} \) is very close to \( \xi_1 \). This is because in...
order to compensate for the delay caused by over scaling the supply voltage then more buffers are needed if the supply voltage equals 1V compared to the case if the supply voltage is 1.5V.

\[
\begin{array}{|c|c|c|c|c|c|c|}
\hline
\text{Param.} & \text{Wire length (in mm)} \\
\hline
\xi_1[pJ] & 0.76 & 1.84 & 5.75 & 10.6 & 13.8 & 29.3 \\
\xi_2[pJ] & 0.37 & 1.03 & 3.89 & 5.66 & 9.84 & 15.6 \\
\xi_3[pJ] & 0.25 & 0.75 & 2.96 & 3.7 & 4.94 & 10.3 \\
\xi_4[pJ] & 0.42 & 0.63 & 1.84 & 3.83 & 4.26 & 10.1 \\
\xi_5 & -51 & -44 & -32 & -46 & -28 & -46 \\
\xi_6 & -67 & -60 & -94 & -65 & -64 & -64 \\
\hline
\end{array}
\]

Table 3. Energy-saving based on reduced swing signaling

5 Conclusion

In this paper we proposed a novel technique for energy saving in DSM technology. Our power saving scheme is based on low-voltage signaling over interconnects with buffer insertion and resizing. A buffer insertion algorithm, VIJIM, that takes into consideration both delay and DSM noise has been proposed. Experimental results using 0.25\textmu m 2.5V CMOS process have showed that VIJIM algorithm has better energy saving factor compared to MZhou algorithm. For on-chip signaling using 0.25\textmu m CMOS process, over 60% of energy has been achieved by reducing the supply voltage from 2.5V down to 1.5V.

Appendix-A: Proof of Lemma 2

The proof is done by contradiction. Firstly, the existence of \( \beta \) will be proved, secondly, the relation between \( \tau_{\text{specified}} \) and \( \beta \) will be elaborated.

It is well known that the transfer function of a first order RLC circuit is a bell shaped function of \( f \). If we denote by \( f_{\text{max}} \) the frequency such that \( |H(j2\pi f)| \) is at the maximum then in the interval \( I = [0, f_{\text{max}}] \), \( |H(j2\pi f)| \) is a monotonically increasing function. In Laplacian domain, we know that \( |V_{\text{out}}(s)| = |V_{\text{in}}(s)||H(s)| \). Let the rise-time of \( V_{\text{in}}(s) \) be a negligible value then the curve of \( |V_{\text{out}}(s)| \) follows exactly the curve of \( |H(s)| \). That means \( |V_{\text{out}}(s)| \) reaches its maximum when \( f = f_{\text{max}} \). Since \( |V_{\text{out}}(s)| \) reaches its maximum when \( f = f_{\text{max}} \) then \( \tau_{\text{oci}} < t_{\text{max}} \), where \( t_{\text{max}} = \frac{1}{f_{\text{max}}} \). Because \( H(s) \) is a continuous function then there exist a unique value \( \beta = |H(j2\pi/\tau_{\text{oci}})| \). In time domain, if \( t < \tau_{\text{oci}} \) then \( v_{\text{out}}(t) < v_{\text{out}}(\tau_{\text{oci}}) \). By using standard techniques, Eq.2 can be easily obtained.

Let us assume for a given \( \tau_{\text{specified}} \) close enough to the measured delay of the OCI, \( \tau_{\text{oci}} \), satisfying the following inequality

\[ |H(\frac{j2\pi}{\tau_{\text{specified}}})| \geq \beta. \]

Suppose that \( \tau_{\text{specified}} \geq \tau_{\text{oci}} \) then from Eq.4 and the previous discussion, we can easily see that in the interval \( I \), if \( |H(\frac{j2\pi}{\tau_{\text{specified}}})| > |H(\frac{j2\pi}{\tau_{\text{oci}}})| \) then \( \tau_{\text{specified}} < \tau_{\text{oci}} \). Thus our hypothesis is absurd. □

References