Power Efficient Inter-Module Communication for Digit-Serial DSP Architectures In Deep-Submicron Technology

Imed Ben Dhaou  Elena Dubrova  Hannu Tenhunen
Electronic System Design Lab
Department of Electronics
Royal Institute of Technology
S-164 40 Kista, Sweden
{imed,elena,hannu}@ele.kth.se

Abstract

This paper investigates the use of quaternary current mode signaling to minimize the power dissipation associated with inter-module communication. We formulate a condition specifying when the insertion of the encoder-decoder pair between the two modules results in a reduction of the overall power consumption of the system. An algorithm LIBCOM is developed which utilizes this condition to insert the encoder-decoder pair between the two modules only if it is advantageous. The HSPICE results obtained for 0.35 μm CMOS process show that LIBCOM can reduce the power consumption by 15%. As technology scales down, the power saved by our algorithm can be several orders of magnitude higher.

1. Introduction

It is widely accepted that in deep-submicron CMOS technology interconnect is a limiting factor for achieving higher integration. One problem is that, as technology scales down, power and delay associated with interconnect becomes several orders of magnitude higher than those attributed to unloaded gates [7]. Another problem is that shrinking of device size and power supply levels, as well as increase in operating speed, result in reduction of noise margins. This makes the circuit increasingly sensitive to transient faults caused by single event upsets like atmospheric neutrons and alpha particles [2] as well as to delay faults [1]. In order to maintain acceptable level of reliability, it is becoming mandatory to design ICs that are capable of avoiding these faults [14].

Several technique for reducing the energy dissipation associated with communication have been proposed, including low swing signaling, charge re-cycling, data coding [3]-[6]. A scheme combining low swing signaling with buffer insertion to avoid noise and delay caused by interconnect has been studied in [13]. An alternative technique for reduction of the cost of on-chip communication, using current-mode multiple-valued signaling, has been examined in [9]. It was shown that the power associated with interconnect between the two modules can be reduced by inserting a binary-to-quaternary encoder and a quaternary-to-binary decoder and allowing the modules to communicate via a quaternary current-mode link.

However, [9] left open an important question, namely, in which situation we should use the quaternary link and in which we shouldn’t. Clearly, a reduction of interconnect between the two modules comes for the expense of addition of the extra hardware (encoder and decoder) which contributes to the area, delay and power of the whole system. Sometimes, the cost of this additional hardware may exceed the theoretical advantage it brings.

In this paper, we address this open problem. We formulate a condition specifying when the insertion of the encoder and the decoder between the two modules results in a reduction of the overall power consumption of the system. Our target application area is digit-serial DSP architectures [11] and on-chip busses [12]. We design an algorithm which utilizes this condition to insert the encoder-decoder pair between the two modules only if it is advantageous. The HSPICE results are presented to justify our theoretical analysis.

The rest of the paper is organized as follows. In Section 2 the condition for insertion of the encoder and the decoder is formulated and proved. In Section 3 the algorithm LIBCOM is designed. Section 4 describe the CMOS MVL circuits for encoder and decoder which we use in our experiments. Section 5 presents the simulation results. Finally, Section 6 concludes this work.
2. Minimizing power consumption associated with interconnect

Let two blocks, \( A_1 \) and \( A_2 \), be connected using \( n \) parallel wires of length \( d \) each (see Figure 1). Such a model of on-chip communication is usual for digit-serial DSP architectures [11] or on-chip busses [12]. The total power consumption of such a system is given by

\[
P = f \cdot V_{dd}^2 \cdot (\alpha_1(C_{out1} + d \cdot n \cdot C_w + C_{in2}) + \alpha_2 C_{out2})
\]

where \( C_{out1} \) and \( C_{out2} \) are the intrinsic output capacitance of \( A_1 \) and \( A_2 \), respectively, \( C_{in2} \) is the input capacitance of \( A_2 \), \( C_w \) is the per-unit-length value of the capacitance of a wire, \( f \) is the clock frequency, and \( \alpha_1 \) and \( \alpha_2 \) are the switching probabilities of \( A_1 \) and \( A_2 \), respectively [8].

![Figure 1. A model of the on-chip communication between \( A_1 \) and \( A_2 \)](image)

Suppose we insert between \( A_1 \) and \( A_2 \) a binary-to-ternary encoder and a ternary-to-binary decoder and allow the blocks \( A_1 \) and \( A_2 \) to communicate via a ternary current-mode link. Since \( n/2 \) wires carrying ternary signals are needed to encode \( n \) wires carrying binary signals, such a scheme reduces the number of wires between \( A_1 \) and \( A_2 \) by half. Following the standard procedure [8], we can show that the total power consumption such a system is given by:

\[
P_{2 \rightarrow 4} = f \cdot V_{dd}^2 \cdot (\alpha_1(C_{out1} + C_{in1}) + \alpha_6(C_{out4} + d \cdot n_{ed} \cdot C_w + C_{in2}) + \alpha_3 C_{out4})
\]

where \( n_{ed} \) is the number of wires between the encoder and the decoder, \( C_{in1}, C_{in2}, C_{out1}, \) and \( C_{out2} \) are the input and output capacitances of the encoder and decoder \( D_1 \), respectively, and \( \alpha_6 \) and \( \alpha_3 \) are the switching probabilities of the encoder and the decoder, respectively.

Clearly, a reduction of the wires between \( A_1 \) and \( A_2 \) comes for the expense of addition of the extra hardware, encoder-decoder pair, which contributes to the area, delay and power of the whole system. We have to make sure that the cost of this additional hardware does not exceed the improvement it brings. The following Lemma formulates a condition showing when the insertion of the encoder-decoder pair results in a reduction of the overall power consumption of the system. The proof is done on the base of the assumption that the power consumed by glitches is zero (zero delay model). Later, in Section 5, we will analyze the consequences of this assumption.

**Lemma 1** The power consumption \( P_{2 \rightarrow 4} \) of two modules with an encoder-decoder pair is smaller than the power consumption \( P \) of two modules without an encoder-decoder pair if

\[
d > n \frac{c_{codec}}{C_w}
\]

where \( d \) is the length of the wires between the modules, \( n \) is the number of wires, \( c_{codec} \) is the intrinsic capacitance of the encoder-decoder pair and \( C_w \) is the per-unit-length value of the capacitance of a wire.

**Proof:** Assume that the power consumed by glitches is zero. Also assume that all the blocks, \( A_2 \), \( A_2 \), encoder and decoder have the same switching activity. Then \( P_{2 \rightarrow 4} \) is less than \( P \) if and only if the following condition is satisfied:

\[
C_{in1} + C_{out1} + C_{out2} < d C_w (n - n_{ed}) .
\]

Since \( n_{ed} < n \), there exists a real number, \( 0 < \beta < 1 \), such that \( n_{ed} = \beta n \). If we denote by \( C_{codec} \) the sum \( C_{in1} + C_{out1} + C_{out2} \), then eq.1 becomes

\[
C_{codec} < C_w d (1 - \beta) .
\]

The power saving factor \( \eta \) is given by

\[
\eta = 1 - \frac{C_{codec}}{C_w d (1 - \beta)} .
\]

Let \( c_{codec} \) be the equivalent capacitance of the binary-to-ternary encoder-decoder pair. Then, the total capacitance of the system is given by

\[
C_{codec} = n (1 - \beta) c_{codec} .
\]

Substituting \( C_{codec} \) in eq.2, we get

\[
\eta = 1 - \frac{n \cdot c_{codec}}{C_w d} .
\]

Clearly, if \( d > \frac{n \cdot c_{codec}}{C_w} \), then \( \eta > 1 \) and thus \( P_{2 \rightarrow 4} \) is less than \( P \).
From Lemma 1 we can conclude that the minimum wire length between the modules at which the addition of the encoder-decoder pair brings the improvement of the total power consumption of the systems is:

\[ d_{\text{min}} = n \cdot \frac{c_{\text{codec}}}{C_w}. \] \[(5)\]

We use this information in the algorithm, developed in the next section, to guide the choice of insertion of the encoder-decoder pair between the modules.

**Example:** Consider the case of a system with the per-unit-length value of the capacitance \( C_w = 0.3pF/mm \), the number of wires \( n = 8 \) and the intrinsic capacitance of the encoder-decoder pair \( c_{\text{codec}} = 30fF \). Using equation (5), we compute \( d_{\text{min}} = 8 \cdot \frac{30}{300} = 0.8mm \). If the wire length between the modules is \( d = 10mm \), then \( P_{2 \rightarrow 4} \) is 92% smaller than \( P \). If \( d = 0.01mm \), then \( P_{2 \rightarrow 4} \) is 79% larger than \( P \).

A potential application area of our technique is low-power implementation of DSP algorithms using digit-serial style [11].

3. LIBCOM algorithm

The name LIBCOM stands for Low-power Inter-Block Communication, and it is used to decide where in the system to insert an encoder-decoder pair between two modules so that the overall power consumption is minimized. The pseudo-code for LIBCOM is given in Figure 2.

LIBCOM takes as its input a data flow graphs (DFG). DFG is a directed graph, \( G = \{ V, E \} \) that has \( N_V \) vertices and \( N_E \) edges. Each vertex of a DFG, \( V_{in_1},...,\in N_V \), is either a computational or storage unit i.e., latches, multipliers, adders, mixers and switches etc, with \( N_{I}^V \) inputs and \( N_{O}^V \) outputs. The vertex \( V_j \) receives data from the vertex \( V_i \) through an edge \( e_{i,j} \).

At the layout level, the edge \( e_{i,j} \) is corresponds to a wire. A wire has an electrical parasitic such as resistance, capacitance and inductance. These values as usually expressed as the Per-Unit Value (PUV). The PUV of L, R and C are computed by solving the multi-dimensional Maxwell equation which is done with the help of a field solver[18]. For larger circuits, the use field solver is almost impossible due to the high computation cost. In order to get around this problem, interconnect library has been widely used, where the idea is to have a predetermined layout structures (such as polygons, squares), then give the layout, its parasitic is computed by matching, in the least square sense, the actual layout geometry to the one given by the interconnect library. This method is referred to as statistical intercon-

**Figure 2. Pseudo code for LIBCOM algorithm**

**LIBCOM\((V, E, C_w, c_{\text{codec}})\)**

**Input:** a set of vertices \( V = \{ V_1, V_2, ..., V_{N_V} \} \), a set of edges \( E = \{ e_{1,1}, e_{1,2}, ..., e_{N_E} \} \), the capacitance of wires \( C_w \), the capacitance of the encoder-decoder pair \( c_{\text{codec}} \).

**Output:** locations \( X \) of the encoder-decoder pair

\[ X = \emptyset; \]
For each edge \( e_{i,j} \in E \):

- estimate the length \( d \) of \( e_{i,j} \);
- determine the number of wires, \( n \), between the vertex \( V_i \in V \) incident with \( e_{i,j} \);
- if \( \frac{n \cdot c_{\text{codec}}}{C_w} < d \) then \( X = X \cup e_{i,j} \)

return \( X \);

nect modeling[17]. However, in this paper the PUV for R, L and C are assumed to be known.

Due to interconnect parasitics, the signal transmitted by \( V_i \) arrives at the vertex \( V_j \) after a delay \( \tau_{i,j} \). The delay, \( \tau_{i,j} \), is caused by the wiring parasitics and it depends on the square of the wire-length (length of the edge \( e_{i,j} \)). In order to overcome this problem, the interconnect has to be partitioned into smaller sections. Each section is then driven by an inverter [10]. Consequently, the power consumption due to signaling between \( V_i \) and \( V_j \) is an increasing function on the wire-length.

Given a DFG, which is represents by a set of connected vertices, for each vertex \( V_i \) incident with the edge \( e_{i,j} \), LIBCOM places an encoder-decoder pair if the length of \( e_{i,j} \) and the number \( n \) of outputs of \( V_i \) satisfy Lemma 1. Consequently, accurate interconnect length estimation is needed in order to effectively use LIBCOM, this suggests that LIBCOM can only be implemented after detailed routing. However, if this information is available in higher levels, e.g. logic level, then LIBCOM can be used during logic synthesis. This problem is however beyond the scope of this paper, we thus assume that interconnect length is a priori known.

The complexity and the correctness of LIBCOM is given by the following Theorem. The proof is omitted due to space limitation (available from the authors on request).

**Theorem 1** For a given DFG, Algorithm 2 inserts the encoder-decoder pair in an optimal way, minimizing the overall power dissipation of the system represented by the DFG. The complexity of the algorithm is linear in the number of edges of the DFG, \( O(|E|) \).
Table 1. Encoding of levels for output current $I_M$

<table>
<thead>
<tr>
<th>$V_0$</th>
<th>$V_1$</th>
<th>$I_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$V_{dd}$</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>$K \cdot I$</td>
</tr>
<tr>
<td></td>
<td>$V_{dd}$</td>
<td>$(K + 1) \cdot I$</td>
</tr>
</tbody>
</table>

Table 2. Decoding of $I_M$

<table>
<thead>
<tr>
<th>$I_M$</th>
<th>$V_0$</th>
<th>$V_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0 \leq I_M &lt; K_0 \cdot I$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$K_0 \cdot I \leq I_M &lt; K_1 \cdot I$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$K_1 \cdot I \leq I_M &lt; K_2 \cdot I$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$I_M &gt; K_2 \cdot I$</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

4. Encode-decoder pair

A number of different implementations of current-mode CMOS encoder and decoder circuits have been proposed (see [15] for an overview). For our experiments, we have adopted the encoder and decoder described in [16].

The circuit diagram of the encoder is shown in Fig.3. It takes as input two voltages, $V_0$ and $V_1$, and produces current signal, $I_M$, which can assume four logic levels. The transistors $M_1, M_2, M_5$ and $M_6$ are used to establish the reference current $I$. The current in $M_4$ is $K$ times larger than the current in $M_3$, where $K$ is a positive integer greater than one. The two currents $I$ and $K \cdot I$ are composed together to form the output current, $I_M$. The level of $I_M$ is controlled by the voltages, $V_0$ and $V_1$, through the pass transistors $M_5$ and $M_6$.

Table 1 shows the values of the output current $I_M$ for different assignments of $V_0$ and $V_1$. In our experiments, we use $K = 2$, i.e., the four levels of $I_M$ are 0, $I$, $2I$ and $3I$. The circuit diagram of the decoder is shown in Fig.4. The four-valued current, $I_M$, is passed through three current comparators formed by the transistors $CP_1 = (M_7, M_8, M_{14}, M_{15}, M_9), CP_2 = (M_7, M_{10}, M_{14}, M_{15}, M_{11})$ and $CP_3 = (M_7, M_{13}, M_{14}, M_{15}, M_{12})$. Each of the current comparators, $CP_{1,2,3}$ produces a LOW voltage if and only if $I_M$ is larger than its drain current. For example, let $I_M$ takes values from $\{0, 1, 2I, 3I\}$, if $I_M = 0.5$, then $CP_1$ produces a logical LOW voltage whereas the others comparators produce a logic HIGH voltage. In order to restore the encoded voltages, circuits consisting of the transistors $M_{16} - M_{23}$ generates the output voltages $V_0$ and $V_1$ which are determined by $CP_1, CP_2$ and $CP_3$ (see Table 2).

5. Experimental results

We have implemented the encoder-decoder pair using 0.35μm 3.3V CMOS process. In order to verify how our assumption that the power consumed by glitches is zero influences the correctness of our results, we have simulated (using HSPICE) the encoder-decoder pairs of different numbers of bits. The results are summarized in Table 3. The row $I_{est}$ shows the estimated current of the binary-to-ternary encoder-decoder pair, computed using the formula $I_{est} = C_{codec} \cdot Vdd \cdot f$, with $C_{codec} = n(1 - \beta)c_{codec}$ (eq.3 from the proof of Lemma 1). The parameter $\beta$ is assumed to be $\beta = 0.5$. The row $C_{codec,est}$ gives the current of the encoder-decoder pair obtained from HSPICE simulator. The row Error shows the error, computed as $Error = (I_{est} - I)/I$. The average error is less than 9%, which...
is an acceptable result. The maximum error is 17% because our power prediction does not account for glitch power consumption.

<table>
<thead>
<tr>
<th>Size</th>
<th>4x4</th>
<th>6x6</th>
<th>7x7</th>
<th>8x8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_{\text{est}}$ (mA)</td>
<td>1.45</td>
<td>2.25</td>
<td>2.99</td>
<td>3.54</td>
</tr>
<tr>
<td>$I$ (mA)</td>
<td>1.7</td>
<td>2.55</td>
<td>3.98</td>
<td>4.34</td>
</tr>
<tr>
<td>Error</td>
<td>17%</td>
<td>13%</td>
<td>-0.26%</td>
<td>-3.72%</td>
</tr>
</tbody>
</table>

Table 3. Simulated versus estimated capacitance of the encoder-decoder pair.

In order to evaluate the power-saving capabilities of our algorithm, we have simulated five different DFGs (diagrams are shown in Figure 5) using HSPICE. The interconnect has been modeled as a lumped capacitance with the $C_{ui} = 0.3pF/mm$. The power-saving factor is computed as $\eta = (I - I_{2+4})/I$, where $I$ is the current consumed by the original DFG and $I_{2+4}$ is the current consumed by the DFG after adding the encoder-decoder pair. Table 4 summarizes the results. DFGs from Figure 5 are numbered as DFG1,...,5. We can see that, for DFG1 and for DFG2, the power-saving factor $\eta$ increases as the length of the wires $d$ between the modules increases. For example, for DFG1, $\eta > 1$ for $d \geq 4cm$. With $d$ changing from 4 cm to 10 cm, $\eta$ is increasing from 3% to 13%, i.e. by 77%. However, for some DFGs, like DFG4, the proposed scheme does not save power. This is because DFG4 contains two multipliers and for a multiplier the glitch power consumption represents an important fraction of the overall power consumption. By adding an encoder-decoder pair to DFG4, the power consumed by glitch increases even further. This problem can be avoided by pipelining DFG4.

<table>
<thead>
<tr>
<th>$d$ (mm)</th>
<th>6</th>
<th>12</th>
<th>20</th>
<th>40</th>
<th>80</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$DFF1$</td>
<td>$I_{\text{est}}$ (mA)</td>
<td>2.45</td>
<td>3.33</td>
<td>3.65</td>
<td>3.96</td>
<td>4.13</td>
</tr>
<tr>
<td>$I_{\text{est}}$ (mA)</td>
<td>1.7</td>
<td>2.55</td>
<td>3.98</td>
<td>4.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\eta$</td>
<td>47%</td>
<td>-17</td>
<td>-17</td>
<td>-17</td>
<td>-17</td>
<td>-17</td>
</tr>
<tr>
<td>$DFF2$</td>
<td>$I_{\text{est}}$ (mA)</td>
<td>3.51</td>
<td>4.40</td>
<td>4.66</td>
<td>4.80</td>
<td>4.86</td>
</tr>
<tr>
<td>$I_{\text{est}}$ (mA)</td>
<td>1.7</td>
<td>2.55</td>
<td>3.98</td>
<td>4.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\eta$</td>
<td>5%</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
</tr>
<tr>
<td>$DFF3$</td>
<td>$I_{\text{est}}$ (mA)</td>
<td>3.51</td>
<td>4.40</td>
<td>4.66</td>
<td>4.80</td>
<td>4.86</td>
</tr>
<tr>
<td>$I_{\text{est}}$ (mA)</td>
<td>1.7</td>
<td>2.55</td>
<td>3.98</td>
<td>4.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\eta$</td>
<td>5%</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
</tr>
<tr>
<td>$DFF4$</td>
<td>$I_{\text{est}}$ (mA)</td>
<td>1.47</td>
<td>2.07</td>
<td>2.07</td>
<td>2.07</td>
<td>2.07</td>
</tr>
<tr>
<td>$I_{\text{est}}$ (mA)</td>
<td>1.47</td>
<td>2.07</td>
<td>2.07</td>
<td>2.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\eta$</td>
<td>40%</td>
<td>-40</td>
<td>-40</td>
<td>-40</td>
<td>-40</td>
<td>-40</td>
</tr>
<tr>
<td>$DFF5$</td>
<td>$I_{\text{est}}$ (mA)</td>
<td>2.78</td>
<td>3.38</td>
<td>3.38</td>
<td>3.38</td>
<td>3.38</td>
</tr>
<tr>
<td>$I_{\text{est}}$ (mA)</td>
<td>2.78</td>
<td>3.38</td>
<td>3.38</td>
<td>3.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\eta$</td>
<td>2%</td>
<td>-2%</td>
<td>-2%</td>
<td>-2%</td>
<td>-2%</td>
<td>-2%</td>
</tr>
</tbody>
</table>

Table 4. Power saving

Overall, the results presented Table 4 may look pessimistic (max. 15% for the case of DFG2). This is due to the fact that the experiments were performed for 0.35μm technology, where the intrinsic capacitance of the encoder-decoder pair is still high compared to the capacitance of interconnect used in this experiments. However, as technology scales down below 0.25μm, the delay and power-consumption of the encoder-decoder pair decreases. For example, the current consumption of the encoder-decoder pair implemented in 0.25μm is 0.24mA whereas for the case of 0.35μm, it is 0.82mA. In ultra deep-submicron technology, the intrinsic capacitance of the encoder-decoder pair becomes much smaller than the capacitance of interconnect. Therefore, the power-saving factor $\eta$ achievable by our algorithm can be several orders of magnitude higher than the one achieved using 0.35μm.

6. Conclusion

In this paper, we investigated the use of quaternary current mode signaling to minimize the power dissipation associated with inter-module communication in a given dataflow graph. We formulated a condition specifying when insertion of the encode-decoder pair between the two modules results in a reduction of the overall power consumption of the system. An algorithm LIBCOM is developed which utilizes this condition to insert the encoder-decoder pair between the two modules only if it is advantageous.

The HSPICE results obtained for 0.35μm CMOS process have showed that LIBCOM can reduce the power by 15% for an interconnect of length 10cm. This result may look modest, but it is due to the fact that in 0.35μm technology the intrinsic capacitance of the encoder-decoder pair...
is still high compared to the capacitance of interconnect used in these experiments. As technology shrinks down, the power-saving achievable by LIBCOM can be several orders of magnitude higher than the one achieved using 0.35μm.

References


