Delay and Energy Efficient Design of an On-Chip Bus with Repeaters Using a New Spatial and Temporal Encoding Technique

Zhang Qingli†, Wang Jinxiang, Yu Mingyan, and Ye Yizheng

(*Microelectronics Center, Harbin Institute of Technology, Harbin* 150001, *China*)

Abstract: On-chip global buses in deep sub-micron designs consume significant amounts of energy and have large propagation delays. Thus, minimizing energy dissipation and propagation delay is an important design objective. In this paper, we propose a new spatial and temporal encoding approach for generic on-chip global buses with repeaters that enables higher performance while reducing peak energy and average energy. The proposed encoding approach exploits the benefits of a temporal encoding circuit and spatial bus-invert coding techniques to simultaneously eliminate opposite transitions on adjacent wires and reduce the number of self-transitions and coupling-transitions. In the design process of applying encoding techniques for reduced bus delay and energy, we present a repeater insertion design methodology to determine the repeater size and inter-repeater bus length, which minimizes the total bus energy dissipation while satisfying target delay and slew-rate constraints. This methodology is employed to obtain optimal energy versus delay trade-offs under slew-rate constraints for various encoding techniques.

Key words: on-chip buses; delay; energy; encoding; repeaters
EEACC: 2570
CLC number: TN47 **Document code**: A **Article ID**: 0253-4177(2008)04-0724-09

1 Introduction

Low-power and high-performance are necessary for all components in deep submicron (DSM) microprocessors and system-on-chip (SoC) designs. This is especially true for global buses, which suffer from higher power consumption and larger propagation delay with technology scaling[1]. According to the International Technology Roadmap for Semiconductors (ITRS)[2], gate delay and local wire delay decrease with technology scaling, while global wire delay tends to increase. Therefore, the propagation delay through global buses will act as a major limiting factor in many high-performance SoC designs. Furthermore, increasing silicon integration levels has resulted in larger chip areas and higher connectivity demands between functional units on a chip. Consequently, on-chip interconnection networks will consume a significant portion of the total system power in many large SoCs through increased wire lengths and number of wires. This is especially prominent in high performance designs, where interconnects are heavily buffered to improve degraded intrinsic RC wire delays and signal slew rates while the heavy use of buffers increases the power dissipation of the interconnect[3,4].

In DSM technology, due to the increased aspect ratio of the metal wires, the coupling capacitances between adjacent wires become significantly larger than the parasitic capacitances between a wire and the substrate (the self capacitance)[1]. Hence, inter-wire capacitive coupling during a switching occurrence becomes a major source of the aforementioned issues. Hereafter, self-transitions can be defined as transitions on the self-capacitance, whereas coupling-transitions can be defined as transitions on the coupling capacitance.

From a power perspective, average power consumption determines battery life, whereas the peak power consumption dictates the packaging and thermal regulation mechanisms that determine the reliability of high performance chips. For a particular wiring pitch and length, the total wire capacitance and resistance are fixed. The factors that determine the peak energy of a buffered bus include voltage supply, the size and number of the repeaters in the bus, and the maximum number of self-transitions and coupling-transitions on the bus in any given clock cycle. In contrast, the average energy per bus cycle is determined by the average number of self-transitions and coupling-transitions per bus cycle[5].

From a performance perspective, the crosstalk between bus signals, caused by increased capacitive coupling, is considered one of the major factors that

† Corresponding author. Email: qinglee@hit.edu.cn
Received 15 June 2007, revised manuscript received 2 December 2007

affect the worst-case delay of bus signals[6]. For a particular wiring pitch and length, the worst-case delay of a buffered bus is determined by the size and number of the repeaters, and the worst-case coupling-transitions in any given clock. Therefore, elimination of the worst-case coupling-transitions can reduce the worst-case delay.

Many encoding techniques using spatial redundancy have been presented to alleviate power or delay problems. For example, the bus-invert method[7] can limit the maximum number of self-transitions and coupling-transitions to 50% and result in potential gains of 50% in peak energy. Various encoding schemes[5,8] are applied to minimize both average self-transition and coupling-transition activity for bus average power reduction; and crosstalk avoidance codes[9] simultaneously provide power efficiency and eliminate the worst-case crosstalk delay. These encoding techniques have a large routing area overhead due to the need for additional bus wires. We refer to such an encoding scheme as spatial encoding. There are also some encoding techniques[10,11] using temporal redundancy to minimize both delay and power. We refer to such an encoding scheme as temporal encoding.

However, all the aforementioned works focus on the effects of encoding schemes on bus delay and power consumption, but do not consider the effects of repeater insertion (the number and size of the repeaters inserted in the bus). Some previous work can be found in the literature that attempts to address the issue of optimizing the repeater insertion in the design process of applying bus encoding techniques for reduced delay and power[11,12]. However, these papers do not provide any closed-form expression for their repeater optimization methods. Instead, they performed the optimization by exhaustively sweeping the number and size of repeaters at the expense of considerable SPICE simulation time. Therefore, their design methods are not suitable for integration in a CAD tool flow.

In our previous work, we presented an energy-efficient temporal encoding circuit[13]. In this paper, we propose a spatial and temporal encoding technique to further reduce energy by combining spatial bus-invert coding[7] with the temporal encoding circuit technique presented in our previous work. In addition, we present a repeater insertion design methodology to determine the repeater size and inter-repeater bus length that minimizes the total bus energy dissipation while satisfying the target delay and slew-rate constraints. This methodology is employed to obtain energy versus. delay trade-offs under slew-rate constraints for the proposed encoding technique.

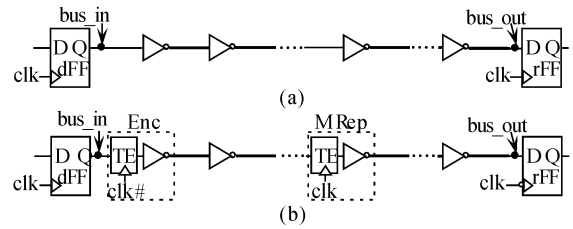

Fig. 1　Topology of original uncoded (a) and temporally encoded bus with repeaters (b)

2　Spatial and temporal encoding method

In this section, we first describe the abstract of the temporal encoding circuit technique proposed in our previous work[13]. Then, we propose a new spatial and temporal encoding method based on the temporal encoding circuit.

2.1　Temporal encoding circuit technique

The topologies of the original uncoded and temporally encoded bus with repeaters are shown in Fig. 1. We are interested in the optimal energy-delay trade-off for transmitting data from the node bus_in to the node bus_out. For the uncoded (UNC) bus, the worst-case coupling-transitions between adjacent wires are '↑ ↓' and '↓ ↑'. The temporal encoding (TE) technique, which eliminates the worst-case coupling-transitions without the cost of additional wire area, is based on the following property: for any given n-bit input data stream, if n-bit shield signals (e.g. all 0' or all 1's) are inserted in the data stream every other data value, it is observed that there exist no '↑ ↓' and '↓ ↑' transitions in the newly generated data stream. The key idea of the technique is that the TE circuit (as shown in Fig. 2) can dynamically build shield signals depending on the results of the logic AND operation of the current and previous state of input data signals instead of inserting fixed shield signals (e.g. all 0's). Thus, on the one hand, the TE-coded bus not only achieves bus switching activity dependent on input switching behavior, but also switches only once in a cycle when static input data switches. Thus, the TE technique has the self transition profile of an uncoded bus. On the other hand, since the TE technique completely eliminates opposite transitions

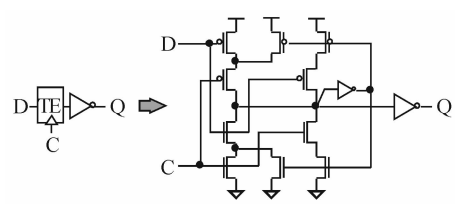

Fig. 2　Temporal encoding circuit

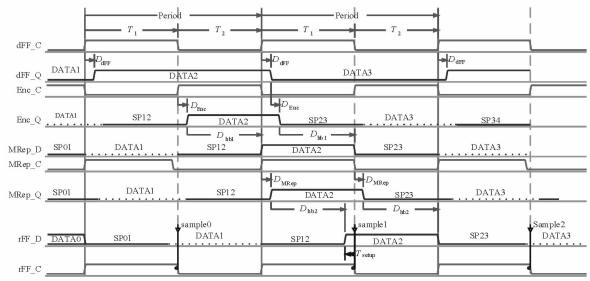

Fig. 3 Timing plan of the TE-coded bus

on adjacent wires, this technique reduces both the maximum number of coupling-transitions in any given clock cycle and the average number of coupling-transitions per bus cycle.

The modified repeater (MRep) near the middle of the TE bus topology has the data latching property, which allows the data signal to traverse the second half of the bus while the first half of the bus transmits shield signal. This ensures full throughput of useful data transmission without the danger of data pulse evaporation. This is illustrated by the timing plan of the TE-coded bus in Fig. 3. From the timing plan, we observe that the T_1 and T_2 phase of clock are determined by the data signal arrival instead of the shield signal arrival; that is, $T_1 \geqslant D_{\mathrm{MRep}}^{\mathrm{TE}} + D_{\mathrm{hb2}} + T_{\mathrm{setup}}$ and $T_2 \geqslant D_{\mathrm{Enc}}^{\mathrm{TE}} + D_{\mathrm{hb1}}$, respectively. Therefore, the bus delay of the TE technique is given by

$$\begin{aligned} T_{\mathrm{d}}^{\mathrm{TE}} &= T_1 + T_2 - D_{\mathrm{dFF}} - T_{\mathrm{setup}} \\ &= D_{\mathrm{hb1}} + D_{\mathrm{hb2}} + D_{\mathrm{codec}}^{\mathrm{TE}} - D_{\mathrm{dFF}} \end{aligned} \quad (1)$$

where D_{hb1} and D_{hb2} are the delays of the first and second half of the bus, respectively, $D_{\mathrm{codec}}^{\mathrm{TE}}$ is the sum of $D_{\mathrm{Enc}}^{\mathrm{TE}}$ and $D_{\mathrm{MRep}}^{\mathrm{TE}}$ which are the delays of the TE circuits (Enc and MRep, respectively), D_{dFF} is the clock-to-out delay of the driver flip-flop (dFF), and T_{setup} is the setup time of the receiver flip-flop (rFF).

2.2 Combination of TE and bus-invert coding

In order to further improve the power efficiency, we combine a spatial bus-invert (BI) coding with the TE technique because the BI coding is a simple but effective low-power coding scheme through self transition activity reduction. Since the TE technique has the self transition profile of an uncoded bus, the BI encoder can be followed by TE encoding without destroying the effectiveness of the self transition activity reduction of the BI coding. The bus topology of the joint coding based on the combination of TE and BI coding is shown in Fig. 4. The original input data value is first encoded through the BI encoder, which computes the Hamming distance H_{d} between the next data value and the present bus value (including the invert bit). The data value is inverted for transmission and the invert bit is set to the high level if $H_{\mathrm{d}} > n/2$ for the n-bit bus; otherwise, the data value is unchanged and the inverted bit is set to the low level. Then, the newly generated data and invert bit are encoded through the TE technique to eliminate opposite transitions on adjacent wires. In summary, the joint coding exploits the benefits of both coding techniques. However, the effectiveness of BI coding decreases as the bus width increases. Therefore, for wide buses, the whole bus is partitioned into several sub-buses each with its own inverted bit to improve switching activity reduction[14]. We refer to the joint coding as the BI(g)TE method, where g is the number of sub-buses. By modifying Eq. (1), we obtain the bus delay of BI(g)TE as follows

$$T_d^{\mathrm{BITE}} = D_{\mathrm{hb1}} + D_{\mathrm{hb2}} + D_{\mathrm{codec}}^{\mathrm{TE}} + D_{\mathrm{dec}}^{\mathrm{BI}} - D_{\mathrm{dFF}} \quad (2)$$

where $D_{\mathrm{dec}}^{\mathrm{BI}}$ is the delay of BI decoder (XOR gate).

3 Delay and energy models of bus with repeaters

3.1 Delay and transition time model of uncoded bus with repeaters

As shown in Fig. 5, a global bus of length L, resistance r per unit length, self capacitance c_{s} per unit length, and coupling capacitance c_{c} per unit length is evenly divided into k segments of length l by identical repeaters. For a repeater of size s, the total output resistance $R_{\mathrm{tr}} = r_{\mathrm{s}}/s$, the total output parasitic capacitance $C_{\mathrm{p}} = sc_{\mathrm{p}}$, and the total input capacitance is $C_{\mathrm{g}} = sc_{\mathrm{g}}$, where r_{s}, c_{p}, and c_{g} are the output resistance, output capacitance, and input capacitance, respectively, of a minimum-sized repeater. The delay t_{ds} and transition time t_{rs} of a wire segment in the bus are obtained by modifying the expression in Ref. [15] as follows

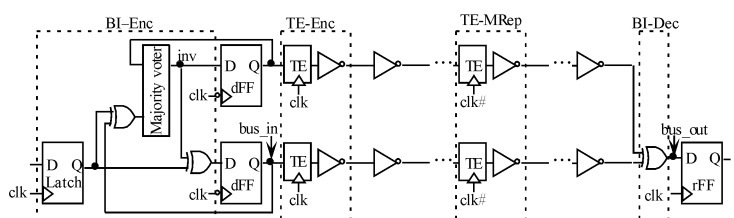

Fig. 4 Bus topology of the BI(g)TE joint coding

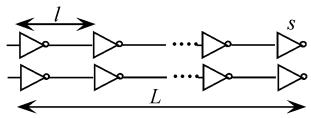

Fig.5　Repeater insertion in a global bus

$$t_{ds}(p,t_{rin},C_L)=\ln 2\left[\frac{r'_s}{s}(sc_p+(c_s+pc_c)l+C_L)+rlC_L\right]+(0.1+0.4\ln 2)r(c_s+pc_c)l^2+\gamma t_{rin} \quad (3)$$

$$t_{rs}(p,C_L)=2.5\ln 3\left[\frac{r''_s}{s}(sc_p+(c_s+pc_c)l+C_L)+rlC_L+0.4r(c_s+pc_c)l^2\right] \quad (4)$$

where p is the coupling factor, which takes values $p=0,1,2,3,$ and 4 depending on the transitions occurring on adjacent wires[1]; t_{rin} is the input transition time of the driving repeater of the wire segment; C_L is the load at the end of the segment, and $C_L=C_g$ if the load is the driving repeater of the next segment; r'_s and r''_s are the different versions of the output resistances r_s used to calculate the delay t_{ds} and transition time t_{rs}, respectively; γ is the coefficient representing the contribution of t_{rin} to the repeater delay, which is determined as[15]

$$\gamma=\frac{1}{2}\left(1-\frac{1-\upsilon_{tn}}{1+\alpha_n}-\frac{1-\upsilon_{tp}}{1+\alpha_p}\right) \quad (5)$$

where $\upsilon_{tn}=V_{tn}/V_{dd}$, $\upsilon_{tp}=V_{tp}/V_{dd}$, and α is the velocity saturation index. The total delay of a wire in the bus is given by

$$T_{wire}(p,L,t_{rdi},C_L)=\left(\frac{L}{l}-1\right)t_{ds}(p,t_{rri},C_g)+t_{ds}(p,t_{rdi},C_g)+t_{ds}(p,t_{rri},C_L) \quad (6)$$

where t_{rdi} is the input transition time of the driver of the wire, t_{rri} is the input transition time of the repeaters in the middle of the wire, thereby, $t_{rri}=t_{rs}(p,C_g)$.

When coding is not employed, p takes the worst-case value of 4. Therefore, the worst-case delay and transition time of an uncoded bus with repeaters are, respectively, given by

$$T_d^{UNC}=T_{wire}(4,L,t_{rdi},C_{in}^{dff}) \quad (7)$$

$$T_r^{UNC}=t_{rs}(4,\max(C_g,C_{in}^{dff})) \quad (8)$$

where C_{in}^{dff} is the input capacitance of the slave flip-flop.

3.2　Energy dissipation model of uncoded bus with repeaters

The energy dissipation of an uncoded bus with repeaters is composed of dynamic, short-circuit, and leakage energy dissipation in the repeaters and dynamic energy dissipation in the interconnects. We assume that the number of self-transitions and coupling-transitions on the bus are N_s and N_c in a given clock cycle, respectively. Therefore, the dynamic energy E_{ds}, short-circuit energy E_{ss}, and leakage energy E_{ls} dissipated in a bus segment during that clock cycle are obtained by modifying the expressions in Ref.[16] as follows

$$E_{ds}(N_s,N_c,C_L)=\frac{1}{2}V_{dd}^2(N_s(lc_s+sc_p+C_L)+N_c lc_c) \quad (9)$$

$$E_{ss}(N_s,t_{rin})=\frac{N_s}{5\ln 3}\ln\left(\frac{1+\upsilon_{tp}}{\upsilon_{tn}}\right)t_{rin}V_{dd}sW_{n_{min}}I_{sc} \quad (10)$$

$$E_{ls}(B)=B\times\frac{1}{2}s(I_{off_n}W_{n_{min}}+I_{off_p}W_{p_{min}})\frac{V_{dd}}{f_{clk}} \quad (11)$$

where V_{dd} is the power supply, f_{clk} is the clock frequency, I_{off_n} (I_{off_p}) is the leakage current for nMOS (pMOS) per unit width, $W_{n_{min}}$ ($W_{p_{min}}$) is the width of the nMOS (pMOS) transistor in minimized sized repeater, I_{sc} is the per unit width short-circuit current, and B is the bit width of bus. Therefore, the total energy dissipation in the whole bus is given by

$$E_{bus}(N_s,N_c,B,L,t_{rdi},C_L)=E_{ds}(N_s,N_c,C_L)+E_{ss}(N_s,t_{rdi})+\left(\frac{L}{l}-1\right)(E_{ds}(N_s,N_c,C_g)+E_{ss}(N_s,t_{rri}))+\frac{L}{l}E_{ls}(B) \quad (12)$$

For the number of self-transitions and coupling-transitions, we have the following computational principle. Both charging and discharging transitions on the self capacitance are counted as self-transitions for energy dissipation. For the transitions on the coupling capacitance, there are three possible cases: charging, discharging, and toggling. Toggling is defined as the case where adjacent lines switch simultaneously in the opposite direction. A toggling case has four times more energy dissipation than the charging or discharging case[5]. Thus, if the number of coupling-transitions is 1 for the charging or discharging case, a toggling event is equivalent to 4 coupling-transitions. Assume that $P_s(X_i)$ is the probability of self capacitance i (of line i) with X_i self-transitions per bus cycle and $P_c(X_i)$ is the probability of coupling capacitance i (between lines i and $i+1$) with X_i coupling-transitions per bus cycle. Then, the expected number of self-transition $\widetilde{N}_s$ and coupling-transitions $\widetilde{N}_c$ per bus cycle on B-bit bus with both outer bus wires having a grounded (shield) wire as a neighbor are, respectively, given by

$$\widetilde{N}_s=\sum_{i=1}^{B}P_s(X_i=1) \quad (13)$$

$$\widetilde{N}_c=\sum_{i=1}^{B-1}\sum_{m=1}^{4}mP_c(X_i=m)+P_s(X_1=1)+P_s(X_n=1) \quad (14)$$

Hereafter, we assume that the original uncoded input data are spatially and temporally independent and a uniformly distributed random n-bit pattern

stream. Then, for an uncoded bus (i.e. $B^{\mathrm{UNC}} = n$), we have $P_s(X_i = 1) = 0.5$, $P_c(X_i = 1) = 0.5$, $P_c(X_i = 2) = P_c(X_i = 3) = 0$, and $P_c(X_i = 4) = 0.125$ for any i[14]. Consequently, the expected number of self-transitions $\tilde{N}_s^{\mathrm{UNC}}$ and coupling-transitions $\tilde{N}_c^{\mathrm{UNC}}$ per bus cycle on an uncoded bus are

$$\begin{cases} \tilde{N}_s^{\mathrm{UNC}} = n/2 \\ \tilde{N}_c^{\mathrm{UNC}} = n \end{cases} \tag{15}$$

The maximum number of self-transitions $\hat{N}_s^{\mathrm{UNC}}$ and coupling-transitions $\hat{N}_c^{\mathrm{UNC}}$ in any given cycle are

$$\begin{cases} \hat{N}_s^{\mathrm{UNC}} = n \\ \hat{N}_c^{\mathrm{UNC}} = 4n - 2 \end{cases} \tag{16}$$

Therefore, the peak energy and average energy of the uncoded bus with repeaters are, respectively, given by

$$E_{\mathrm{peak}}^{\mathrm{UNC}} = E_{\mathrm{bus}}(\hat{N}_s^{\mathrm{UNC}}, \hat{N}_c^{\mathrm{UNC}}, B^{\mathrm{UNC}}, L, t_{\mathrm{rdi}}, C_{\mathrm{in}}^{\mathrm{dff}}) \tag{17}$$

$$E_{\mathrm{avg}}^{\mathrm{UNC}} = E_{\mathrm{bus}}(\tilde{N}_s^{\mathrm{UNC}}, \tilde{N}_c^{\mathrm{UNC}}, B^{\mathrm{UNC}}, L, t_{\mathrm{rdi}}, C_{\mathrm{in}}^{\mathrm{dff}}) \tag{18}$$

3.3 Effects of bus encoding techniques on delay and transition time

As described in the section 2, the TE and BI(g) TE bus encoding techniques eliminate opposite transitions on adjacent wires, reducing the maximum coupling factor from $p = 4$ to $p = 2$, thereby, reducing the worst-case delay of the bus. However, the delay overhead of encoding circuits has a negative effect on the bus delay reduction. From Eqs. (1), (2), and (6), the worst-case delays of the TE and BI(g)TE coded buses with repeaters are given by equations (B1) and (B2) shown below, respectively, where $C_{\mathrm{in}}^{\mathrm{mrep}}$ is the input capacitance of the M-Repeater, $C_{\mathrm{in}}^{\mathrm{dec}}$ is the input capacitance of the BI decoder (XOR gate), $t_{r_{\mathrm{out}}}^{\mathrm{enc}}$ and $t_{r_{\mathrm{out}}}^{\mathrm{mrep}}$ are the output transition times of TE circuits (Encoder and M-Repeater, respectively), and L_{hb1} and L_{hb2} are the lengths of the first and second half of the bus, respectively. The maximum transition times of the both buses are, respectively, given by

$$T_r^{\mathrm{TE}} = t_{\mathrm{rs}}(2, \max(C_g, C_{\mathrm{in}}^{\mathrm{mrep}}, C_{\mathrm{in}}^{\mathrm{dff}})) \tag{19}$$

$$T_r^{\mathrm{BITE}} = t_{\mathrm{rs}}(2, \max(C_g, C_{\mathrm{in}}^{\mathrm{mrep}}, C_{\mathrm{in}}^{\mathrm{dec}})) \tag{20}$$

3.4 Effects of bus encoding techniques on energy

For the self-transitions on each line, since the TE and BI(g)TE bus techniques have the switching characteristics of an uncoded bus and a BI-coded bus, respectively, results from Ref. [14] can be applied here.

For TE coding:

$$P_s(X_i = 1) = 0.5 \tag{21}$$

For BI(g)TE coding:

$$P_s(X_i = 1) = \frac{1}{2} - 2^{-(n+1)} C\left(n, \frac{n}{2}\right) \tag{22}$$

For the coupling-transitions between adjacent wires, the TE and BI(g) TE techniques transform a toggling event into a discharging followed by a charging event (or vice-versa), reducing the number of coupling-transitions from 4 to 2. Thus, in various switching scenarios on a coupling capacitance, there are only three possible values: 0, 1, and 2 for the number of coupling-transitions. Similar to the Markov-based approach employed in Ref. [14], by modeling the TE and BI(1) TE coding processes as Markov chains, we obtain $P_c(X_i = 1)$ and $P_c(X_i = 2)$ for TE and BI(1)TE techniques, respectively, as follows

For TE coding:

$$\begin{cases} P_c(X_i = 1) = 0.5 \\ P_c(X_i = 2) = 0.125 \end{cases} \tag{23}$$

For BI(1)TE coding:

$$P_c(X_i = 1) = 2\sum_{h=0}^{n/2-1} C(n-1, h) 2^{-n} = 0.5 \tag{24}$$

$$P_c(X_i = 2) = 2^{-3} - 2^{-n-1} C\left(n-1, \frac{n}{2} - 1\right) \tag{25}$$

Furthermore, $B^{\mathrm{TE}} = n$ and $B^{\mathrm{BI(1)TE}} = n + 1$. Consequently, the expected number of self-transitions and coupling-transition per bus cycle for both coding techniques, are, respectively, given by

$$\begin{cases} \tilde{N}_s^{\mathrm{TE}} = \dfrac{n}{2} \\ \tilde{N}_c^{\mathrm{TE}} = \dfrac{3(n-1)}{4} \end{cases} \tag{26}$$

$$\tilde{N}_s^{\mathrm{BI(1)TE}} = (n+1)\left(\frac{1}{2} - 2^{-(n+1)} C\left(n, \frac{n}{2}\right)\right) \tag{27}$$

$$\tilde{N}_c^{\mathrm{BI(1)TE}}(n) = 1 + \frac{3}{4}n - (n+2)2^{-n} C\left(n-1, \frac{n}{2} - 1\right) \tag{28}$$

Now, we consider the coupling-transitions for BI(g)TE buses with $g \geqslant 2$. Suppose the bus lines are partitioned into g equal-sized groups, each of which has $m = n/g$ lines, excluding invert lines. Then, we have $B^{\mathrm{BI}(g)\mathrm{TE}} = n + g$. Using a similar approach to Ref. [14], we obtain the expected number $\tilde{N}_c^{\mathrm{p}}$ of coupling-transitions between the invert line of group j and the first bus line of group $j+1$ as follows

$$T_d^{\mathrm{TE}} = T_{\mathrm{wire}}(2, L_{\mathrm{hb1}}, t_{r_{\mathrm{out}}}^{\mathrm{enc}}, C_{\mathrm{in}}^{\mathrm{mrep}}) + T_{\mathrm{wire}}(2, L_{\mathrm{hb2}}, t_{r_{\mathrm{out}}}^{\mathrm{mrep}}, C_{\mathrm{in}}^{\mathrm{dff}}) + D_{\mathrm{enc}}^{\mathrm{TE}} + D_{\mathrm{mrep}}^{\mathrm{TE}} - D_{\mathrm{dff}}$$

$$= \left(\frac{L}{l} - 2\right) t_{\mathrm{ds}}(2, t_{\mathrm{rri}}, C_g) + t_{\mathrm{ds}}(2, t_{r_{\mathrm{out}}}^{\mathrm{enc}}, C_g) + t_{\mathrm{ds}}(2, t_{r_{\mathrm{out}}}^{\mathrm{mrep}}, C_g) + t_{\mathrm{ds}}(2, t_{\mathrm{rri}}, C_{\mathrm{in}}^{\mathrm{mrep}}) + t_{\mathrm{ds}}(2, t_{\mathrm{rri}}, C_{\mathrm{in}}^{\mathrm{dff}}) + D_{\mathrm{codec}}^{\mathrm{TE}} - D_{\mathrm{dff}} \tag{B1}$$

$$T_d^{\mathrm{BITE}} = \left(\frac{L}{l} - 2\right) t_{\mathrm{ds}}(2, t_{\mathrm{rri}}, C_g) + t_{\mathrm{ds}}(2, t_{r_{\mathrm{out}}}^{\mathrm{enc}}, C_g) + t_{\mathrm{ds}}(2, t_{r_{\mathrm{out}}}^{\mathrm{mrep}}, C_g) + t_{\mathrm{ds}}(2, t_{\mathrm{rri}}, C_{\mathrm{in}}^{\mathrm{mrep}}) +$$

$$t_{\mathrm{ds}}(2, t_{\mathrm{rri}}, C_{\mathrm{in}}^{\mathrm{dec}}) + D_{\mathrm{codec}}^{\mathrm{TE}} + D_{\mathrm{dec}}^{\mathrm{BI}} - D_{\mathrm{dff}} \tag{B2}$$

$$E_{\text{peak}}^{\text{TE}} = E_{\text{bus}}(\hat{N}_{\text{s}}^{\text{TE}}, \hat{N}_{\text{c}}^{\text{TE}}, B^{\text{TE}}, L_{\text{hb1}}, t_{\text{r}_{\text{out}}}^{\text{enc}}, C_{\text{in}}^{\text{mrep}}) + E_{\text{bus}}(\hat{N}_{\text{s}}^{\text{TE}}, \hat{N}_{\text{c}}^{\text{TE}}, B^{\text{TE}}, L_{\text{hb2}}, t_{\text{r}_{\text{out}}}^{\text{mrep}}, C_{\text{in}}^{\text{dff}}) + \hat{E}_{\text{codec}}^{\text{TE}}$$
$$= \left(\frac{L}{l} - 2\right)(E_{\text{ds}}(\hat{N}_{\text{s}}^{\text{TE}}, \hat{N}_{\text{c}}^{\text{TE}}, C_{\text{g}}) + E_{\text{ss}}(\hat{N}_{\text{s}}^{\text{TE}}, t_{\text{rri}})) + \frac{L}{l}E_{\text{ls}}(B^{\text{TE}}) +$$
$$E_{\text{ds}}(\hat{N}_{\text{s}}^{\text{TE}}, \hat{N}_{\text{c}}^{\text{TE}}, C_{\text{in}}^{\text{mrep}}) + E_{\text{ds}}(\hat{N}_{\text{s}}^{\text{TE}}, \hat{N}_{\text{c}}^{\text{TE}}, C_{\text{in}}^{\text{dff}}) + E_{\text{ss}}(\hat{N}_{\text{s}}^{\text{TE}}, t_{\text{r}_{\text{out}}}^{\text{enc}}) + E_{\text{ss}}(\hat{N}_{\text{s}}^{\text{TE}}, t_{\text{r}_{\text{out}}}^{\text{mrep}}) + \hat{E}_{\text{codec}}^{\text{TE}} \quad \text{(T1)}$$

$$E_{\text{peak}}^{\text{BITE}} = E_{\text{bus}}(\hat{N}_{\text{s}}^{\text{BITE}}, \hat{N}_{\text{c}}^{\text{BITE}}, B^{\text{BITE}}, L_{\text{hb1}}, t_{\text{r}_{\text{out}}}^{\text{enc}}, C_{\text{in}}^{\text{mrep}}) + E_{\text{bus}}(\hat{N}_{\text{s}}^{\text{BITE}}, \hat{N}_{\text{c}}^{\text{BITE}}, B^{\text{BITE}}, L_{\text{hb2}}, t_{\text{r}_{\text{out}}}^{\text{mrep}}, C_{\text{in}}^{\text{dff}}) + \hat{E}_{\text{codec}}^{\text{BI}} + \hat{E}_{\text{codec}}^{\text{TE}}$$
$$= \left(\frac{L}{l} - 2\right)(E_{\text{ds}}(\hat{N}_{\text{s}}^{\text{BITE}}, \hat{N}_{\text{c}}^{\text{BITE}}, C_{\text{g}}) + E_{\text{ss}}(\hat{N}_{\text{s}}^{\text{BITE}}, t_{\text{rri}})) + \frac{L}{l}E_{\text{ls}}(B^{\text{BITE}}) +$$
$$E_{\text{ds}}(\hat{N}_{\text{s}}^{\text{BITE}}, \hat{N}_{\text{c}}^{\text{BITE}}, C_{\text{in}}^{\text{mrep}}) + E_{\text{ds}}(\hat{N}_{\text{s}}^{\text{BITE}}, \hat{N}_{\text{c}}^{\text{BITE}}, C_{\text{in}}^{\text{dff}}) + E_{\text{ss}}(\hat{N}_{\text{s}}^{\text{BITE}}, t_{\text{r}_{\text{out}}}^{\text{enc}}) + E_{\text{ss}}(\hat{N}_{\text{s}}^{\text{BITE}}, t_{\text{r}_{\text{out}}}^{\text{mrep}}) + \hat{E}_{\text{codec}}^{\text{BI}} + \hat{E}_{\text{codec}}^{\text{TE}} \quad \text{(T2)}$$

$$\tilde{N}_{\text{c}}^{\text{p}} = P_{\text{s}}(X_i = 1)(2 - P_{\text{s}}(X_i = 1))$$
$$= \left(\frac{1}{2} - 2^{-(n+1)}C\left(n, \frac{n}{2}\right)\right)\left(\frac{3}{2} + 2^{-(n+1)}C\left(n, \frac{n}{2}\right)\right) \quad (29)$$

Therefore, the expected number of coupling-transitions per bus cycle on an BI(g)TE bus is

$$\tilde{N}_{\text{c}}^{\text{BI}(g)\text{TE}} = g\,\tilde{N}_{\text{c}}^{\text{BI}(1)\text{TE}}(m) + (g-1)\,\tilde{N}_{\text{c}}^{\text{p}} \quad (30)$$

The maximum number of self-transitions and coupling-transitions for the both coding techniques in any given cycle are

$$\begin{cases} \hat{N}_{\text{s}}^{\text{TE}} = n \\ \hat{N}_{\text{c}}^{\text{TE}} = 2n \end{cases} \quad (31)$$

$$\begin{cases} \hat{N}_{\text{s}}^{\text{BI}(g)\text{TE}} = n/2 \\ \hat{N}_{\text{c}}^{\text{BI}(g)\text{TE}} = n \end{cases} \quad (32)$$

From Eqs. (12), (31), and (32), the peak energy dissipation of the TE and BI(g)TE buses with repeaters are given by equations (T1) and (T2) above, respectively, where $\hat{E}_{\text{codec}}^{\text{TE}}$ is the peak energy dissipation of the TE codec circuits and $\hat{E}_{\text{codec}}^{\text{TE}}$ is the peak energy dissipation of the BI codec circuits. The average energy dissipation per bus cycle of both buses can be obtained in a similar way, but they are not shown here due to limited space.

4 Repeater insertion optimization method

As described in Section 3, the worst-case delay, maximum transition time, and energy dissipation are function of repeater size s and segment length l (or the number of segments $k = L/l$). There exits the delay optimal point in the design space of l and s. However, this delay-minimal repeater design methodology is not necessarily an appropriate strategy in practical circuits. The delay is not sensitive to the size of the repeaters near the optimal point. Therefore, significant power and area are wasted to achieve only a small improvement in speed when approaching the optimal point for minimum delay. So, in many instances, such as non-critical global buses, a target delay is desired rather than a minimal delay to reduce energy dissipation. Here, we present a repeater insertion design methodology for achieving the minimum bus energy dissipation at each target delay point while satisfying a maximum slew-rate constraint in the design process of applying bus encoding techniques for reduced delay and power.

Here, we investigate the repeaters insertion of an uncoded bus to illustrate the optimization method. Assume that the desired delay is $T_{d_{\text{target}}}$ and the maximum transition time (slew-rate) constraint is $T_{r_{\max}}$, then we set

$$T_d^{\text{UNC}} \leqslant T_{d_{\text{target}}} \quad (33)$$

$$T_r^{\text{UNC}} \leqslant T_{r_{\max}} \quad (34)$$

If $T_{d_{\text{target}}}$ is greater than the optimal delay, then there exist many combinations of l and s that satisfy Eq. (33). For each s, an optimal l exists to achieve the minimum peak energy Eq. (17) or average energy Eq. (18). If l is too large, the signal transition time will be large and the short-circuit energy becomes larger. There are many combinations of l and s that satisfy Eq. (34). The minimum energy dissipation satisfying the slew-rate constraint can be achieved with minimum-sized repeaters. For minimum-sized repeaters, the corresponding l and bus delay, however, are impractically large. In order to produce an effective repeater insertion, the delay and slew-rate constraint should be considered simultaneously. We apply a genetic algorithm (GA) and sequential quadratic programming (SQP) method based solvers in the Matlab toolbox to solve the non-linear constrained optimization problem with the objective function being Eq. (17) or (18), and the constrained functions being Eqs. (33) and (34). First, we use the genetic algorithm to find a good starting point (l, s) for the global solution. Next, since $k(=L/l)$ is usually not an integer, the nearest two integers are used to determine the minimum energy dissipation while the SQP based solver is applied to further refine the value of s.

5 Experimental results

The methodology described in Section 4 is employed to obtain energy versus delay trade-offs under

Table 1 Wire RC parasitic for different bus pitches

Bus pitch	r/(Ω/mm)	c_s/(fF/mm)	c_c/(fF/mm)
1× min. pitch ($W/S=0.2\mu$m/0.2μm)	303	76.29	91.55
1.125× min. pitch ($W/S=0.2\mu$m/0.25μm)	303	84.06	72.65
1.25× min. pitch ($W/S=0.2\mu$m/0.3μm)	303	91.59	59.30

slew-rate constraints for uncoded and encoded buses with repeaters. In all simulations, we use industrial 0.13μm CMOS technology. We consider a 16-bit bus in the metal-6 layer when both outer bus wires have a shield wire as a neighbor. If all uncoded and encoded buses are assumed to be routed at minimum pitch, there is an area penalty for the buses coded with BI (g)TE (i.e. a 12.5% increase in routing area for $g=2$, and 25% increase in routing area for $g=4$). For a fair comparison, the wire width and spacing of the uncoded and TE-coded buses are re-optimized for minimum energy within the increased routing area. This re-optimization results in increased spacing since the coupling capacitance decreases rapidly, reducing both energy and delay of the bus. The capacitances per unit length for the different bus pitches used in the design were extracted using Synopsys's Raphael and are listed in Table 1. Device parameters were extracted using SPICE simulation similar to Refs. [15, 16]. The relevant technology parameters are shown in Table 2. The minimized-size repeater is defined to have $W_n=2L_{drawn}=0.26\mu$m and $W_p=2.4W_n$. The codec circuits are sized so as to operate at the "knee" of their respective energy-delay curves. The knee points typically result in energy that is 10%～20% higher than the minimum energy of codec circuits and yield nearly constant delay over a range of load capacitances. Hence, for simplicity, we assume that the codec circuits have fixed configurations, delay, and energy overheads for all delay targets. These values, measured by SPICE simulation, are shown in Table 3. In this paper, buses of length 9mm are optimized and the maximum transition time constraint at every point on the bus is set to 240ps (～3× the transition time at the output of an inverter driving a fanout-of-4 load).

Table 2 Device parameters for an industrial 0.13μm CMOS technology

Parameter	Value	Parameter	Value
r_s'	7.44kΩ	r_s''	3.54kΩ
c_g	1.35fF	c_p	2.59fF
V_{tn}	0.42V	V_{tp}	−0.36V
α_n	1.23	α_p	1.45
I_{offn}	4.15mA/m	I_{offp}	6.97 mA/m
I_{sc}	50μA/μm	V_{dd}	1.2V

Table 3 Overheads of codec circuits and other related design parameters

Overheads	Value	Parameter	Value
D_{enc}^{TE}	53ps	D_{dff}	120ps
D_{rep}^{TE}	67ps	t_{rdi}	100ps
D_{dec}^{BI}	15ps	$t_{r_{out}}^{enc}$	240ps
$\hat{E}_{codec}^{TE}$	1.93～1.6 6pJ	$t_{r_{out}}^{mrep}$	240ps
$\bar{E}_{codec}^{TE}$	1.43～1.55pJ	C_{in}^{mrep}	16fF
$\hat{E}_{codec}^{BI}$	0.79pJ	C_{in}^{dec}	4fF
$\bar{E}_{codec}^{BI}$	0.65pJ	C_{in}^{dff}	4fF

The peak and average energy versus worst-case delay trade-off curves for the uncoded and encoded buses with repeaters are shown in Fig. 6 and Fig. 7. The left most point of each curve represents the delay-optimized solution and, hence, consumes the highest peak (average) energy. Meanwhile, the right most point of each curve represents the peak (average) energy-optimized solution due to the slew rate constraints. For the routing area constraint of 1.125× minimum pitch (Fig. 6(a) and Fig. 7(a)), TE and BI (2) TE achieve peak energy gains of 59.2% and 74.4%, respectively, and average energy gains of 55.1% and 52.4% over the uncoded bus, respectively, at the minimum achievable delay points of the uncoded bus.

Furthermore TE and BI (2) TE allow more aggressive delay targets to be met (23.6% and 17.2% faster, respectively), while still dissipating less peak

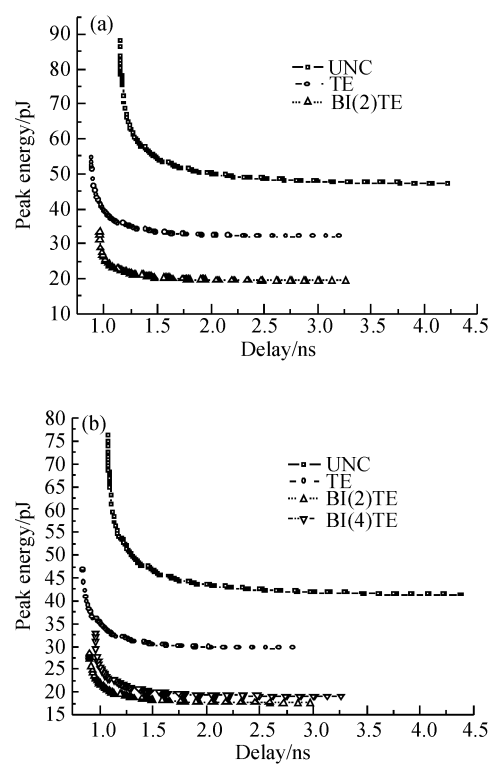

Fig. 6 Peak energy versus delay curves for a routing area constraint of 1.125× (a) and 1.25× (b) minimum pitch

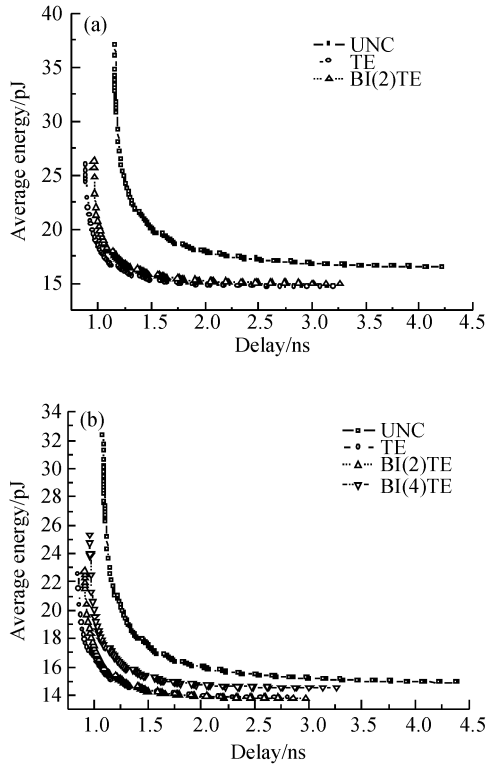

Fig. 7　Average energy versus delay curves for a routing area constraint of 1.125×(a) and 1.25×(b) minimum

energy (38% and 62.1%, respectively) and average energy (29.7% and 28.9%, respectively). In many instances, such as non-critical global buses, a target delay is larger than the delay number corresponding to the minimum peak(average) energy point, and reducing energy is the primary goal. In such cases, TE and BI(2)TE can result in 32% and 58.9% reductions in peak energy, respectively, and 10.8% and 9.1% reductions in average energy, respectively. Thus, TE and BI(2)TE can provide peak(average) energy savings over the uncoded bus at all target delay values. Similar analyses were carried out for the routing area constraint of 1.25× minimum pitch (Fig. 6(b) and Fig. 7(b)). The relevant data points for all cases have been tabulated in Table 4. The number of segments and size of repeaters required to minimize peak energy while meeting the delay and slew rate constraints for a bus with 1.125× minimum pitch are plotted in Fig. 8. Due to space limitations, we have not shown optimal repeater configurations for the 1.25× minimum pitch and average energy cases.

Table 4 shows that the proposed BI(g)TE technique yields much better results in peak energy reduction than the TE technique, but trivial (or even negative) improvements in average energy reduction. This is because the energy gains that result from a weak reduction in the average number of self- and coupling-

Table 4　Gains achieved by TE, BI(2)TE, and BI(4)TE over uncoded bus with repeaters for 9mm bus

Metrics	Coding	E_{peak}		E_{avg}		Delay	
		×min. Pich		×min. Pich		×min. Pich	
		1.125	1.25	1.125	1.25	1.125	1.25
Max. gain /%	TE	59.2	56.3	55.1	51.5	23.6	21.7
	BI(2)TE	74.4	73.3	52.4	50.5	17.2	16.4
	BI(4)TE	–	69.8	–	45.8	–	11.6
Gains at min. energy point/%	TE	32	28.4	10.8	7.3	24.2	36
	BI(2)TE	58.9	57.4	9.1	7.5	22.9	32.2
	BI(4)TE	–	53.7	–	2.5	–	25.7

transitions are counteracted by the fact that the TE-coded bus energy can also be reduced further under the increased area penalty incurred by BI(g)TE. With respect to the efficiency of BI(g)TE with various g values to reduce peak/average energy, BI(2)TE results in larger gains in peak and average energy than BI(4)TE at all target delay values (though BI(2)TE should ideally result in the same peak energy gains with BI(4)TE and less average energy gains), since, under the same area constraint, the bus pitch for BI(2)TE is more relaxed than BI(4)TE due to fewer control signals.

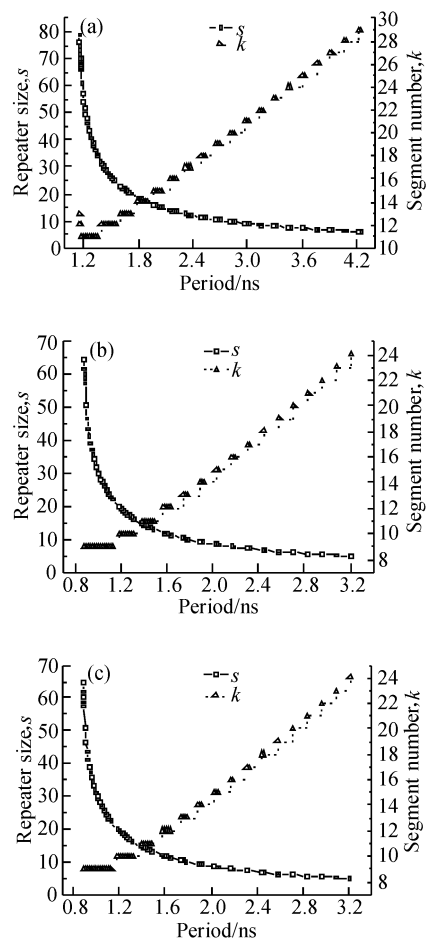

Fig. 8　Optimal repeater configurations for UNC (a), TEC (b) and BI(2) TE (c) at 1.125× minimum pitch

6 Conclusion

A new spatial and temporal encoding approach has been proposed for global on-chip buses in DSM SoC design to allow higher performance than a uncoded repeater tuning strategy while reducing peak and average energy. Compared with the temporal encoding technique presented in our previous work, the proposed spatial and temporal encoding approach further improves peak energy reduction. In the design process of applying bus encoding techniques for reduced bus delay and energy, we presented a repeater insertion design methodology for achieving the minimum bus energy with delay and slew-rate constraints, which can be employed to obtain energy versus. delay trade-off curves under slew-rate constraints for buses with repeaters, thereby, providing convenience for comparisons of the efficiency of the various encoding techniques.

References

[1] Sotiriadis P P. Interconnect modeling and optimization in deep submicron technologies. PhD Dissertation, Massachusetts Inst Technol, Cambridge, May 2002

[2] International Technology Roadmap for Semiconductors. Semiconductor Industry Association, 2003

[3] Sylvester D, Keutzer K. Getting to the bottom of deep submicron II: the global wiring paradigm. Proc ISPD, 1999

[4] Liu X Y, Chen S M. A low-latency low-power schme for on-chip global interconnects. Chinese Journal of Semiconductors, 2005, 26(9): 1854(in Chinese)[刘祥远, 陈书明. 一种低延迟低功耗的片上全局互连方法. 半导体学报, 2005, 26(9): 1854]

[5] Zhang Y, Lach J, Skadron K, et al. Odd/Even bus invert with two-phase transfer for buses with coupling. Proc ISLPED, 2002: 80

[6] Victor B, Keutzer K. Bus encoding to prevent crosstalk delay. Proc ICCAD, 2001: 57

[7] Stan M R, Burleson W P. Bus-invert coding for low-power I/O. IEEE Trans VLSI System, 1995, 3(1): 49

[8] Sotiriadis P P, Chandrakasan A. Low power bus coding techniques considering inter-wire capacitances. Proc CICC, 2000: 507

[9] Sridhara S R, Ahmed A, Shanbhag N R. Area and energy- efficient crosstalk avoidance codes for on-chip buses. Proc ICCD, 2004: 12

[10] Mutyam M, Eze M, Vijaykrishnan N, et al. Delay and energy efficient data transmission for on-chip buses. Proc ISVLSI, 2006: 6

[11] Anders M, Rai N, Krishnamurthy R, et al. A transition-encoded dynamic bus technique for high performance interconnection. IEEE J Solid-State Circuit, 2003, 38(5): 709

[12] Kaul H, Sylvester D, Anders M A, et al. Design and analysis of spatial encoding circuits for peak power reduction in on-chip buses. IEEE Trans VLSI Systems, 2005, 13(11): 1225

[13] Zhang Q L, Wang J X, Ye Y Z. An energy-efficient temporal encoding circuit technique for on-chip high performance buses. Proc GLSVLSI, 2006: 273

[14] Lin R B. Coupling reduction analysis of bus-invert coding. Proc ISCAS, 2005: 5862

[15] Chen G Q, Friedman E G. Low-power repeaters driving rc and rlc interconnects with delay and bandwidth constraints. IEEE Trans VLSI System, 2006, 14(2): 161

[16] Banerjee K, Mehrotra A. A power-optimal repeater insertion methodology for global interconnects in nanometer designs. IEEE Trans Electron Devices, 2002, 49(11): 2001

基于一种新的时空编码技术的片上总线的低延迟低能耗设计

张庆利† 王进祥 喻明艳 叶以正

(哈尔滨工业大学微电子中心, 哈尔滨 150001)

摘要: 在深亚微米设计中,降低能耗和传播延迟是片上全局总线所面对的两个最主要设计目标.本文提出了一种用于片上全局总线的时空编码方案,它既提高了性能又降低了峰值能耗和平均能耗.该编码方案利用空间总线倒相编码和时间编码电路技术的优点,在消除相邻连线上反相翻转的同时,减少了自翻转数和耦合翻转数.在应用该总线编码技术降低总线延时和能耗的设计中,给出了一种总线上插入中继驱动器的设计方法,以确定它们合适的尺寸和插入位置,使得在满足目标延时和翻转斜率要求的同时总线总的能耗最小.该方法可用来为各种编码技术获得翻转斜率约束下的总线能耗与延时的优化折中.

关键词: 片上总线; 延时; 能量有效; 编码; 中继驱动器

EEACC: 2570

中图分类号: TN47 **文献标识码**: A **文章编号**: 0253-4177(2008)04-0724-09

† 通信作者. Email: qinglee@hit.edu.cn

2007-06-15 收到, 2007-12-02 定稿