A novel low-swing interconnect optimization model with delay and bandwidth constraints*

Zhu Zhang-Ming(朱樟明)†, Hao Bao-Tian(郝报田),
Yang Yin-Tang(杨银堂), and Li Yue-Jin(李跃进)

Microelectronics School, Xidian University, Xi'an 710071, *China*

(Received 27 February 2010; revised manuscript received 9 April 2010)

Interconnect power and repeater area are important in the interconnect optimization of nanometer scale integrated circuits. Based on the RLC interconnect delay model, by wire sizing, wire spacing and adopting low-swing interconnect technology, this paper proposed a power-area optimization model considering delay and bandwidth constraints simultaneously. The optimized model is verified based on 65-nm and 90-nm complementary metal-oxide semiconductor (CMOS) interconnect parameters. The verified results show that averages of 36% of interconnect power and 26% of repeater area can be saved under 65-nm CMOS process. The proposed model is especially suitable for the computer-aided design of nanometer scale systems-on-chip.

Keywords: interconnect power, repeater area, low-swing circuit, time delay, bandwidth

PACC: 7850G

1. Introduction

Due to the increasing complexity of systems-on-chip (SOC) and continued development of networks-on-chip (NOC), more and more functionality modules are integrated on a chip. It leads to a growing chip size despite the device dimensions have been shrinking with the aggressive scaling down of very large scale intigration (VLSI) technology. The developing trend also leads to an increase in the use of global interconnects to carry signals across the chip, thus the performance of modern integrated circuits (ICs) is mainly determined by global interconnect performance. As communication mechanisms between subsystems (such as intellectual property blocks and routers) and microprocessors are realized using high-speed bus structures or point-to-point interconnects, on-chip communication is becoming a performance bottleneck and a significant consumer of power and area budgets. Therefore, it is meaningful to solve the performance, power, and area problems arising from global interconnect.

In order to meet performance requirements arising from increasing clock frequency and increasing complexity, many methods, such as optimal repeater insertion (ORI),[1,2] wire spacing and sizing,[3] low-swing circuit,[4,5] bus encoding,[6] and diverse combinations of them,[7,8] have been proposed. Uniform repeater insertion is an effective technique for driving long interconnects and the authors in Ref. [1] presented closed form expressions for optimum repeater insertion to minimize the delay of a propagating signal. A small delay penalty will save considerable amount of power. Based on distributed RLC interconnect model, the authors in Ref. [2] proposed a methodology to calculate repeater size and interconnect length which minimizes the sum of power dissipation and repeater area while satisfying a target delay constraint. Using a specified figure of merit (FOM) function, a new methodology was presented in Ref. [3] for determining the optimal wire sizing of buffered global interconnects and the impact of wire width and space on performance was investigated. Considering Vdd and Vth optimization, the authors in Ref. [4] presented an analytical repeater insertion method which optimizes power under delay constraint for a single net. Being different from the method of repeater insertion, an irredundant bus encoding scheme for on-chip buses was proposed in Ref. [6] to tackle thermal issue as well as power issue. Based on RC delay

*Project supported by the National Natural Science Foundation of China (Grant Nos. 60725415 and 60971066), the National High-Tech Program of China (Grant Nos. 2009AA01Z258 and 2009AA01Z260), and the National Science & Technology Important Project of China (Grant No. 2009ZX01034-002-001-005).

†Corresponding author. E-mail: zmyh@263.net

© 2010 Chinese Physical Society and IOP Publishing Ltd

model, crosstalk and delay are optimized simultaneously by wire sizing, wire spacing and repeater insertion in Ref. [8].

In the present paper, a new performance driven interconnect optimization methodology is presented. Combined with wire sizing, wire spacing, repeater insertion and low-swing circuit, the methodology is focused on interconnect power and repeater area optimization with both delay and bandwidth constraints. It is generally beneficial for any global interconnect design and is extremely suitable for interconnect bus design used for data communication between subsystems and many-core microprocessors. The paper investigates the effects of wire width, wire space and repeater supply voltage on interconnect power and repeater area first, and then proposes a power-area optimization model considering wire width, wire spacing and repeater supply voltage. Based on this optimization model, top level interconnects under 90-nm and 65-nm CMOS technology are optimized and optimization results are analysed theoretically. A simple conclusion is made in Section 5.

2. The interconnect optimization strategy

2.1. Overview of interconnect optimization

Consider a low-swing interconnect with total length of $L_{\rm w}$ buffered by repeaters, as shown in Fig. 1. A traditional circuit, conventional level converter (CLC) is used to convert a low swing signal back to a full swing one. Though simple, this circuit is very robust against noise.[4] The primary objective of uniform repeater insertion system is to minimize the time for a signal to propagate through a long interconnect. It is widely believed that interconnect delay is actually very slight with respect to both repeater size and separation close to the minimum point.[2] Large power saving can be achieved by a small delay penalty.

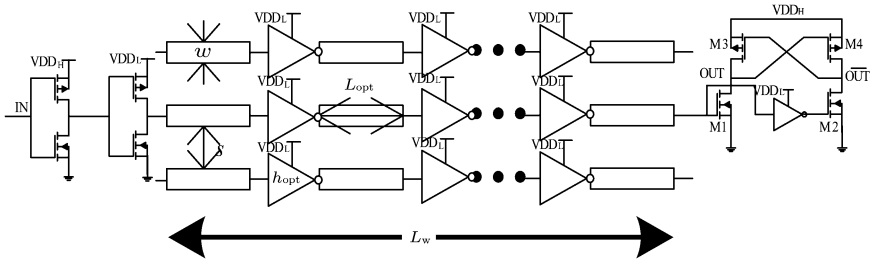


Fig. 1. Buffered interconnect bus structure.

Supposing pitch, the sum of wire width and wire space is 3 times that of typical wire width, effect of wire width on power dissipation and repeater area is illustrated in Fig. 2 with the same delay constraint. Obviously, adopting wire width equal to wire space is not the best choice, as it will substantially increase interconnect power and repeater area. The optimal power and optimal area do not appear at the same wire width. A compromise between power and area should be made on during optimization process. Supposing the pitch is 3 times that of typical wire width and the wire width equals the wire space, the effect of repeater supply voltage $\rm VDD_L$ on power dissipation and repeater area is illustrated in Fig. 3 with the same delay constraint. Clearly the interconnect power declines rapidly while the repeater area rises dramatically when voltage drops and thus a compromise should be made also.

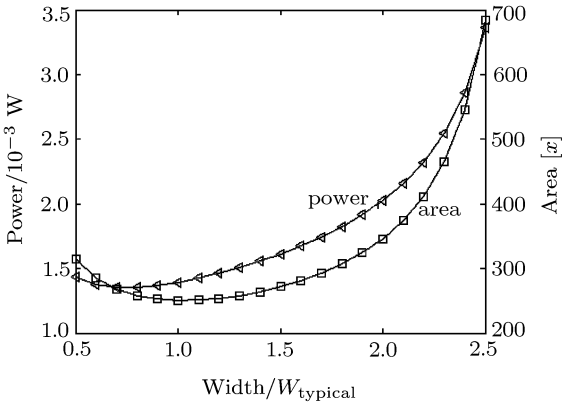


Fig. 2. Effect of interconnect width on power dissipation and repeater area.

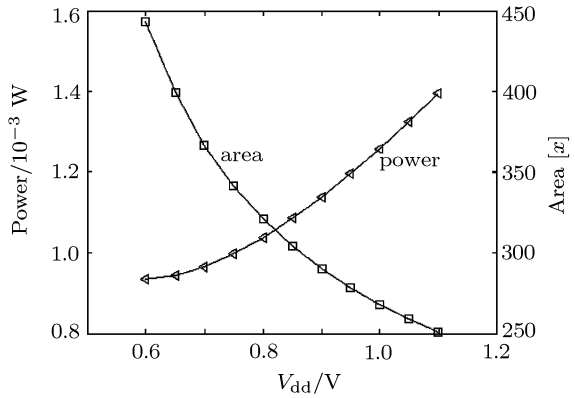


Fig. 3. Effect of repeater supply voltage on power dissipation and repeater area.

2.2. Interconnect optimization model

As communication mechanisms between subsystems (such as IP blocks and routers) and microprocessors are realized by using high-speed bus structures or point-to-point interconnects, bus performance plays a key role in today's ICs. In order to realize real time communication and guarantee considerable data throughput in a fixed area, it is necessary to introduce delay and bandwidth as constraint conditions during the optimization of power dissipation and repeater area.

Based on the above considerations, we propose a figure-of-merit (FOM) function to minimize power dissipation and repeater area, which is a function of repeater size h, optimal interconnect length L, interconnect width w, interconnect space s and repeater supply voltage $\mathrm{VDD_L}$. Its expression can be given as:

$$\mathrm{FOM} = \mathrm{Power}^{i} \cdot \mathrm{Area}^{j}, \tag{1}$$

where Power and Area denote interconnect power and repeater area, respectively, and i, j are the weight factors of the power and area, respectively. Then the optimization model can be finally expressed as:

$$\begin{aligned} &\min.\ \mathrm{FOM} \\ &\text{s.t.} \begin{cases} BW = BW_{\rm tar} \\ T_{\rm total} = T_{\rm tar} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} w \\ s \\ \mathrm{VDD_L} \end{bmatrix} \geq \begin{bmatrix} w_{\min} \\ s_{\min} \\ \mathrm{VDD_{L,min}} \end{bmatrix}, \end{cases} \end{aligned} \tag{2}$$

where $\mathrm{BW_{tar}}$ and $T_{\rm tar}$ denote the bandwidth target and time delay target, respectively. This model subjects to constraints consisting of two equations and three inequalities. The main purpose of this model is to find the optimal solution $\{h, L, w, s, \mathrm{VDD_L}\}_{\rm opt}$. A local optimal solution for the FOM exists for each set of target constraints. It is a minimum problem of a constrained nonlinear multivariable function and solution can be obtained numerically.

With optimization flow shown in Fig. 4, all designer's work is to input weight factors and target constraints. The computer will seek the optimal solution meeting these requirements automatically through numerical method and then output the optimal solution and corresponding interconnect power and repeater area as design reference.

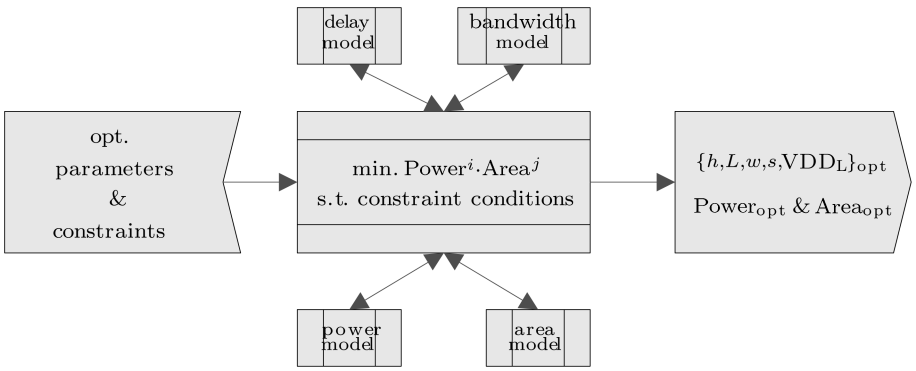


Fig. 4. Interconnect optimization flow with delay and bandwidth constraints.

3. Buffered interconnect model

Optimization flow depicted in Fig. 4 indicates requirement of accurate yet simple interconnect models. The required models include delay model, power model, repeater area model, bandwidth model and the most basal models of resistance, capacitance and inductance. Referring to related works,[1,2,9−12] this sec-

tion will introduce several models that can meet these requirements.

3.1. RLC parameter models for nanometer scale interconnect

With the fast development of VLSI technology, the size of top level interconnect decreases continually, which makes interconnect resistance increase and thus resistance plays a more important role in interconnect delay. At the same time, interconnect-to-ground capacitance is constantly decreasing, which highlights the position of coupling capacitance. In order to reduce delay and crosstalk, nanometer scale CMOS process introduced low-k dielectric material. Medium material's relative dielectric constant under 65-nm technology node has a value of 2.2, far less than that of silicon oxide, so parasitic interconnect inductance becomes increasingly important in time delay. Being the same as models used in Ref. [11], we adopt resistance, capacitance and inductance model provided by Berkeley Predictive Technology Model (BPTM).[10,11]

3.2. Delay model considering effect of inductance

Reconsider the global interconnect with total length of $L_{\rm w}$ depicted in Fig. 1. Total interconnect delay should consist of two parts: interconnect RLC delay and CLC gate delay. Only RLC delay is considered here, as the gate delay is so small compared with interconnect delay that it can be neglected.

Assume that the line and repeater dimensions have been optimized for some purposes and the optimal wire width and space are $w_{\rm opt}$ and $s_{\rm opt}$, respectively. The line is partitioned by repeaters into $k_{\rm opt}$ equal segments and each segment has a length of $L_{\rm opt}$. Then based on RLC parameter models mentioned above, the resistance per unit length r, capacitance per unit length c and inductance per unit length l can be denoted by these parameters. Suppose the repeater size is $h_{\rm opt}$ and then the input capacitance is $h_{\rm opt}c_{\rm p}$, the output capacitance is $h_{\rm opt}c_{\rm o}$ and the output resistance is $r_{\rm s}/h_{\rm opt}$, where c_0, $c_{\rm p}$ and $r_{\rm s}$ are the input capacitance, output capacitance and output resistance of repeater with minimum size, respectively.

The output resistance can be affected greatly by repeater supply voltage. Considering supply voltage $\mathrm{VDD_L}$, the output resistance $r_{\rm s}$ can be expressed as:[5]

$$r_{\rm s} = K\frac{\mathrm{VDD_L}}{(\mathrm{VDD_L} - V_{\rm th})^{\lambda}}, \tag{3}$$

where K is the device parameter, $V_{\rm th}$ the device threshold voltage and λ is about 1.25 for current technology. Assume $r_{\rm s0}$ is the output resistance of repeater supplied by normal voltage $\mathrm{VDD_H}$, then $r_{\rm s}$ can be rewritten as

$$r_{\rm s} = r_{\rm s0}\frac{\mathrm{VDD_L}}{\mathrm{VDD_H}}\left(\frac{\mathrm{VDD_H} - V_{\rm th}}{\mathrm{VDD_L} - V_{\rm th}}\right)^{\lambda}. \tag{4}$$

Referring to Ref. [1], time delay of each segment can be described as follows:

$$\begin{aligned} T_{\rm d} &= \left(\mathrm{e}^{-2.9\xi^{1.35}} + 1.48\xi\right)/\omega_n, \\ \xi &= \frac{rL_{\rm opt}}{2}\sqrt{\frac{c}{l}}\cdot\frac{r_{\rm s}/rL_{\rm opt}h_{\rm opt} + c_{\rm p}h_{\rm opt}/rL_{\rm opt} + r_{\rm s}c_{\rm p}/(rL_{\rm opt})^2 + 0.5}{\sqrt{1 + c_{\rm p}h_{\rm opt}/rL_{\rm opt}}}, \\ \omega_n &= 1/\sqrt{l\cdot L_{\rm opt}(cL_{\rm opt} + c_{\rm p}/h_{\rm opt})}. \end{aligned} \tag{5}$$

For an inductance interconnect partitioned into $k_{\rm opt}$ equal segments the total delay can be written as:

$$T_{\rm total} = T_{\rm d}\cdot\frac{L_{\rm w}}{L_{\rm opt}}. \tag{6}$$

3.3. Power dissipation model

The power dissipation of a buffered interconnect consists of switching power, leakage power and short-circuit power. Compared with the other two parts, the short-circuit power is so small that it can be neglected in the power model. For a long interconnect with length of $L_{\rm w}$, considering power dissipated by the CLC, the total power dissipation can be written as the following equation (7). And $h_{\rm L}$ expresses the size of the left-most inverter in Fig. 2, $h_{\rm R}$ denotes the size of transistors M1, M2, assuming that the size of M3 and M4 is equal and is two times that of $h_{\rm R}$. c_l is the load capacitance driven by the CLC circuit, $\mathrm{VDD_H}$ is the normal power supply voltage, $\mathrm{VDD_L}$ is repeater supply voltage, $f_{\rm clock}$ is clock frequency, α is the switch-

ing factor which is commonly taken as 0.15, $I_{\rm offn}$ is the leakage current per unit n-metal-oxide-semiconductor (NMOS) transistor width and $W_{n\,\rm min}$ is the width of the NMOS transistor in the minimum sized inverter. Thus

$$\begin{aligned}\text{Power} = {} & k_{1\rm L}\left(h_{\rm opt}\frac{L_{\rm w}}{L_{\rm opt}}\left(c_0+c_{\rm p}\right)+cL_{\rm w}+\frac{2}{3}h_{\rm L}c_{\rm p}\right)\\ & +k_{1H}\left(\left(h_{\rm L}+2h_{\rm R}\right)\left(c_0+c_{\rm p}\right)-\frac{2}{3}h_{\rm R}c_{\rm p}+c_l\right)\\ & +k_{2\rm L}h_{\rm opt}\frac{L_{\rm w}}{L_{\rm opt}}+k_{2\rm H}\left(h_{\rm L}+2h_{\rm R}\right),\end{aligned} \tag{7}$$

where

$$\begin{aligned}&k_{1\rm L}=\alpha{\rm VDD}_{\rm L}^2 f_{\rm clk},\quad k_{1H}=\alpha{\rm VDD}_{\rm H}^2 f_{\rm clk},\\ &k_{2\rm L}=\frac{3}{2}{\rm VDD}_{\rm L}I_{\rm offn}W_{n\,\rm min},\\ &k_{2\rm H}=\frac{3}{2}{\rm VDD}_{\rm H}I_{\rm offn}W_{n\,\rm min}.\end{aligned}$$

The first two parts on the right-hand side of Eq. (7) denote the switching power, which dominates total power dissipation. The left two parts denote the leakage power which increases rapidly with technology scaled down and becomes key part of total power dissipation.[13] Power dissipated by the inverter in the CLC is ignored as its size is small.

3.4. Repeater area model

Repeaters are always used to reduce the total delay by subdividing the line into short sections. However, the number of the repeaters can be prohibitively high ($> 10^6$ for sub-100-nm designs) for high-performance designs and the size of the repeaters can be quite large. A large number of repeaters occupying a significant fraction of silicon area affect chip structure and layout aggressively and have been a significant factor in interconnect optimization.

The repeater area for an optimal buffered global interconnect is proportionally to $h_{\rm opt}$ and inversely proportional to $L_{\rm opt}$. Assuming that the CMOS inverter adopts devices with minimum channel length, the expressions of repeater area can be written as

$$\text{Area} = h_{\rm opt}+h_{\rm L}+2h_{\rm R}, \tag{8}$$

where Area denotes the total silicon area occupied by repeaters and it has been normalized to the area occupied by repeater with minimum size.

3.5. Bandwidth model

Let BW denote the bandwidth, i.e., the rate at which bits can be transmitted across unit length of buffered interconnect in a given chip. It is inversely proportional to the delay per unit length. The number of such lines is inversely proportional to the metal pitch, which is the sum of wire width w and wire space s. Therefore BW can be expressed by:

$$\text{BW} = \frac{1}{(w+s)\,T_{\rm total}/L_{\rm w}}. \tag{9}$$

The aim of interconnect design is to have a large BW while maintaining a small delay per unit length. Larger bandwidth means bigger data throughput and small delay per unit length means higher performance.

4. Model verification and discussion

BPTM[9] and the international technology roadmap for semiconductor (ITRS)[10] provided typical integrated circuit parameters such as listed in Table 1. The minimum delay $T_{\rm min}$ and maximum bandwidth $\text{BW}_{\rm max}$ corresponding to each technology node can be obtained by the model provided in Ref. [1]. Taking $\text{BW}_{\rm max}$ and 2/3 time of $\text{BW}_{\rm max}$ respectively as target bandwidths, $T_{\rm min}$ as target time delay, based on optimization model shown in this paper, we carried out optimization of top level interconnect under 65-nm and 90-nm CMOS technology nodes separately.

Table 1. Physical and electric parameters for nanometer CMOS interconnect.

tech. node/nm	$w_{\rm typical}$/nm	$s_{\rm typical}$/nm	$t_{\rm m}$/nm	$t_{\rm ins}$/nm	$\varepsilon_{\rm r}$	$r_{\rm s}$/kΩ	$C_{\rm p}$/fF	C_0/fF	$\text{VDD}_{\rm H}$/V	$I_{\rm offn}$/μA·μm
65	450	450	1200	200	2.2	9.6	1.03	1.22	1.1	20
90	500	500	1200	300	2.8	9.04	1.1	2.04	1.1	3.56

During the optimization, the size of transistors in the CLC is chosen as $h_{\rm L} = 0.9h_{\rm opt}$ and $h_{\rm R} = 0.19h_{\rm opt}$, as experience has found that the optimal results can be achieved under this condition, and the load capacitance is chosen as 10 fF. Optimization results of a 10-mm long top level interconnect under 65-nm CMOS technology node are shown in Table 2, Fig. 5 and Fig. 6. The optimal wire width and optimal repeater supply voltage should be taken under each operating frequency and listed in Table 2 corresponding to different weight factors and bandwidth constraints. The 44600 GHz and 29700 GHz of bandwidth constraints correspond to pitches of two times and three times typical wire width $w_{\rm typical}$, respectively. From Table 2 we can see that the optimal wire width is not one-fold and 1.5-fold $w_{\rm typical}$, and designer taking wire width equal to wire space cannot obtain the most superior result. $\rm VDD_L$ has a clear declining trend along with the increase of operating frequency and power weight factor. It explains that repeater supply voltage will play the vital role in optimization of high performance ICs or in optimization preferentially focused on power.

Taking 1 time and 2/3 time of maximum bandwidth $\rm BW_{max}$ as bandwidth constraints respectively, applications of model in this paper and model in Ref. [1] obtain optimization results of power and repeater size. Figure 5 shows the power optimization results and figure 6 shows the area optimization results. One may see from the figures that, under the same condition the present model can save an average of 36% of power consumption and 26% of the repeater area compared with the model in Ref. [1] when power and area weight factors are taken as one. When relaxing the bandwidth constraint of 2/3 times $\rm BW_{max}$, the present model can save an average of 52% of the power consumption and 42% of the repeater area. When optimization is particularly focused on power consumption ($i = 2$), regarding to 1 time of $\rm BW_{max}$ bandwidth constraint, the power dissipation further drops 6.1% averagely, the repeater area averagely rises 31%, and regarding to 2/3 time of $\rm BW_{max}$ bandwidth constraint, the power further drops 4.5% averagely, the buffer area rises 28% averagely. This explained that more power reduction is obtained by sacrificing the repeater area. Using the same method, this paper also carried out optimization of a 10-mm long top level interconnect under 90-nm CMOS technology, and obtained similar results as shown in Figs. 7 and 8. Optimization results under 65-nm technology node are more encouraging compared with these under 90-nm technology node, indicating the optimization model is better suited to high performance ICs with advanced process technology.

Table 2. Optimization results of top level interconnect under 65-nm CMOS technology.

i	1			2			$\rm BW_{TAR}$/GHz	$T_{\rm TAR}$/ps
f/GHz	1	3	5	1	3	5		
w [$w_{\rm typical}$]	0.78	0.76	0.75	0.75	0.75	0.75	44600	249
$\rm VDD_L$/V	1.1	1.06	1.01	0.74	0.86	0.83		
w [$w_{\rm typical}$]	0.97	0.95	0.94	0.99	0.73	0.80	29740	
$\rm VDD_L$/V	1.08	0.95	0.91	0.80	0.64	0.74		

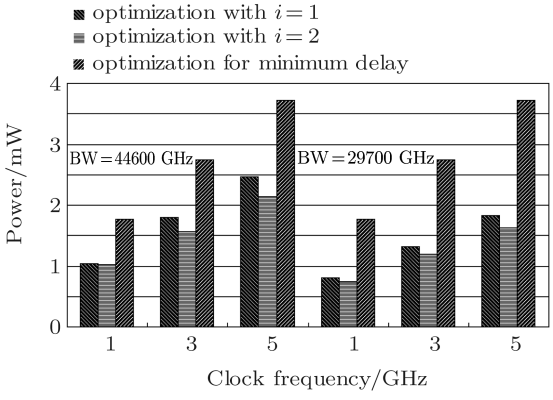


Fig. 5. Power dissipation of optimal top level interconnect under 65-nm CMOS technology.

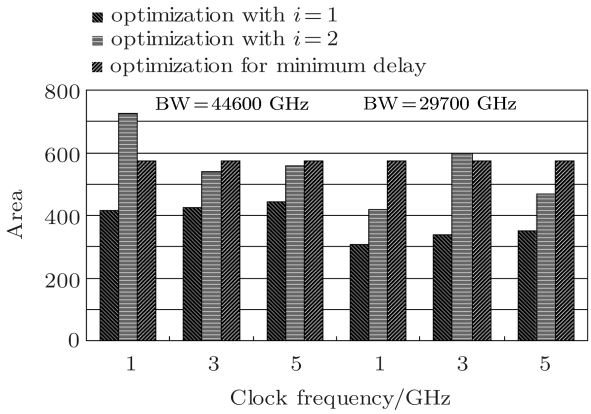


Fig. 6. Repeater area of optimal top level interconnect under 65-nm CMOS technology.

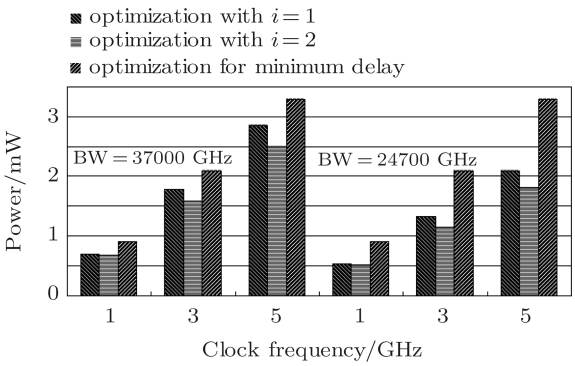


Fig. 7. Power dissipation of optimal top level interconnect under 90-nm CMOS technology.

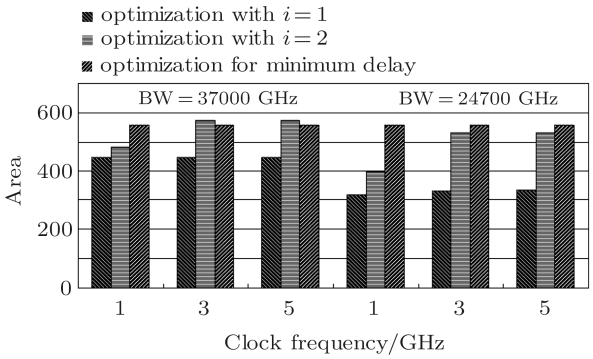


Fig. 8. Repeater area of optimal top level interconnect under 90-nm CMOS technology.

From the above analysis we can see that the achievement of our model is quite remarkable. Under the same conditions the present model can save an average of 36% of power consumption and 26% of repeater area, and when relaxing the bandwidth constraint of 2/3 times $\mathrm{BW}_{\max}$, this model can save an average of 52% of power consumption and 42% of repeater area. Better optimization results can be achieved for more advanced process. Moreover, the model performance does not degenerate along with the rise of system operating frequency. The proposed model is quite suitable for optimization of global interconnect in high performance ICs with advanced process technology.

5. Conclusion

The paper investigates the impacts of wire width, wire space and repeater supply voltage on interconnect power and repeater area first. Then, based on the RLC interconnect delay model, by combination of wire sizing, wire spacing and by optimization of size, number and supply voltage of low-swing repeater, we proposed a power-area optimization model considering delay and bandwidth constraints simultaneously. Verification based on 65-nm and 90-nm CMOS process technology confirms the validity of the proposed model. Under exactly the same conditions, the present model can save an average of 36% of power and 26% of repeater area compared with the model in Ref. [1] under 65-nm technology node, and it may play an guiding role in design and optimization of nanometer scale CMOS ICs.

References

[1] Lsmail Y I and Friedman E G 2000 *IEEE Trans. VLSI Systems* **8** 195

[2] Zhu Z M, Qian L B and Yang Y T 2009 *Chin. Phys.* B **18** 1188

[3] Tang M and Mao J F 2008 *Proceedings of International Conference on Microwave and Millimeter Wave Technology* Nanjing, China, April 21–24, 2008 p. 479

[4] Chang Y C, Tam K H and He L 2005 *ACM International Symposium on Low Power Electronics and Design* San Diego, USA, August 8–10, 2005 p. 137

[5] Zhang H and Rabaey J 1998 *ACM International Symposium on Low Power Electronics and Design* California, USA, August 10–12, 1998 p. 161

[6] Wang F, Xie Y and Vijaykrishnan N 2006 *IEEE Design Automation and Test in Europe* Munich, Germany, March 27–30, 2006 p. 1

[7] El-Moursy M A and Friedman E G 2005 *Integration the VLSI Journal* **38** 205

[8] Hasani F and Masoumi N 2008 3*th Int. Conf. on Design and Technology of Integrated Systems in Nanoscale Era* Cairo, Egypt, March 25–27, 2008 p. 1

[9] http://www.eas.asu.edu/∼ptm/

[10] Semiconductor Industry Association 2007 *International Technology Roadmap for Semiconductors*

[11] Zhu Z M, Qian L B and Yang Y T 2009 *Acta Phys. Sin.* **58** 2631 (in Chinese)

[12] Li X, Wang J M and Tang W Q 2009 *Acta Phys. Sin.* **58** 3603 (in Chinese)

[13] Banerjee K and Mehrotra A 2002 *IEEE Trans. Electron. Dev.* **49** 2001