Throughput Optimization for Area-Constrained Links With Crosstalk Avoidance Methods

Basel Halak and Alex Yakovlev

***Abstract*—The effect of crosstalk avoidance codes on the throughput of fixed width communication channels is studied. Closed form expressions of the throughput which incorporate the dimensions of the interconnects and the wiring overheads incurred by such techniques are derived for lines under different buffering conditions. These formulae are utilized to optimize the bandwidth of constrained-area parallel buses under different latency and power constraints. Our results are confirmed by the simulations we have performed in Spectre for a UMC CMOS 90-nm technology.**

***Index Terms*—Crosstalk, interconnect, performance.**

I. Introduction

As VLSI technology progresses toward integration densities that will allow for more than a billion active devices per chip, the cost of high-speed wire networks will become excessive. The economic demands to continue the exponential reduction in price per function will force the use of area efficient wiring methodologies that will require a shift from a low-density latency-centric global wire design to a high-density throughput-centric wire design [1]–[3]. The interconnect performance and power consumption in current deep sub-micrometer technologies is greatly affected by crosstalk noise due to the decreasing wire separation and increased wire aspect ratio [4]. This trend is anticipated to worsen in the future. Techniques to avoid the crosstalk delay have been proposed by many researchers [5]–[10]. These methods can generally be implemented on the physical or data link layers of the design or on both levels. Physical layer solutions include wire sizing optimization and buffer insertion [2], [9]. The techniques implemented on the data link layer consist of data encoding to avoid crosstalk [10]. Although the use of crosstalk avoidance techniques improves the wire delay, it incurs wiring overheads, which may reduce the link throughput.

This paper explores the design tradeoffs of global interconnect architectures. The key question that we try to answer is, given a fixed area in which to distribute interconnect, what is the best arrangement of wires to obtain the highest bandwidth and/or minimum energy dissipation? Is it to use all the wires to send data at a low signalling frequency, or to implement crosstalk avoidance techniques to operate at a higher signalling frequency but with less number of wires? What effects does repeater insertion have?

Opportunities for achieving high throughput and energy efficient links are revealed through the creation of new physical models for interconnect throughput. These new models incorporate the channel geometry (wires dimensions) and the power and wiring overheads incurred by crosstalk avoidance coding schemes (CACs).

To the best of our knowledge the effect of CACs on the throughput has not been addressed before.

The organization of this paper is as follows: Section II summarizes the derivation a closed-form analytical expression for communication throughput for fixed width links. In Section III we formulate three optimization problems which are based on the design constraints of the communication link, we have considered three types of buses: throughput-centric, latency-constrained and power-constrained channels. Algorithmic and analytical solutions for these problems are detailed in Sections IV–VI, respectively. Our results are verified using spectre for a standard UMC 90-nm technology. Finally, conclusions are drawn in Section VII.

Manuscript received June 04, 2008; revised September 05, 2008 and December 16, 2008; accepted March 01, 2009. First published August 04, 2009; current version published May 26, 2010.

The authors are with the School of Electrical, Electronic, and Computer Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, U.K.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2009.2017915

II. Bandwidth Model Derivation

Throughput in this paper refers to the number of data bits per second that is delivered over a physical link. It is the product of the number of data carrying wires and the clock frequency. The latter depends on the worst-case wire delay. The maximum frequency of the channel is given as follows:

$$f_{\max} = \frac{1}{\eta * D}. \tag{1}$$

η is a safety factor which depends on the application and on the variability of wire delays. η is chosen to be 1.5 in order to account for the 50% expected variability of wire parasitic in future technology as indicated in [4].

D is the worst case wire delay, which is widely accepted to be the 50% propagation delay of signals.

For uniformly buffered resistance-capacitance (*RC*) lines, it can be calculated as follows [2]:

$$D = (0.7R_d/H) * (c_g + p * c_c + K * H * c_d) + R * (((0.4 * c_g + 0.4 * p * c_c)/K) + 0.7H * c_d) \tag{2}$$

where p is the capacitive coupling factor, it is a function of the transition activities and can have one the following values $\{1,2,3,4\}$ for simultaneously switching signals [5]. R_d and C_d are the input capacitance and output resistance of the repeaters, H, K is the size and number of the repeaters. R, C_g and C_c are wire resistance, ground and inter-wire capacitances, respectively, they are functions of wire dimensions and metal and dielectric properties. The explicit equations of these parasitic elements are included in our previous publication [1].

For an N-wire communication channel, the bandwidth (throughput) can be calculated as follows:

$$BW = N * f_{\max}. \tag{3}$$

For communication links with a fixed width (cw) (see Fig. 1), the total number of wires is a function of the wire width ($w = w_{\min} * wn$) and spacing ($s = w_{\min} * sn$). This can be written as follows:

$$N = \frac{(cw/w_{\min}) - sn}{wn + sn} \tag{4}$$

where $w_{\min}$ is the minimum wire width in the considered technology, *wn* and *sn* are multiples of $w_{\min}$.

Crosstalk avoidance codes can be used to reduce the wire delay by decreasing p to 3, 2, or 1. This, however, will decrease the number data carrying wires, in which case the bandwidth is given as follows:

$$BW = \frac{Cr * N}{\eta * D} \tag{5}$$

where Cr is the coding rate for the CAC under consideration. The coding rate of CACs is a function of the total number of wire available in the channel (N) and its coupling factor. Using curve fitting

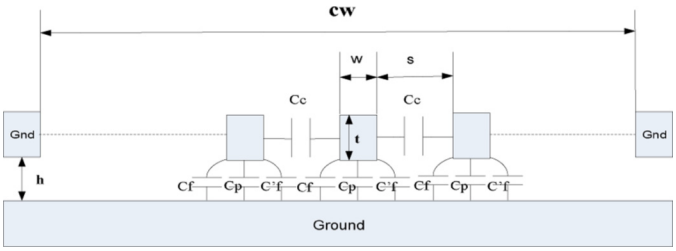

Fig. 1. Cross section of global interconnect ($h = 0.94$ um, $t = 0.81$ um, $w_{\min} = 0.56$ um).

techniques Cr has already been derived in our previous work, The accuracy of Cr formulae was shown to be more than 97% [1].

For linear CACs the coding rate on wide buses ($N > 8$) is given as follows:

$$Cr = 0.2355 * e^{(0.3586*p)}. \quad (6)$$

The coding rates of nonlinear CAC's have a strong dependency on the number of wires [6]. For the range of links ($16 < N < 64$), the coding rate is given as follows:

$$Cr = 0.4145 * \ln(p) + 0.4128. \quad (7)$$

III. Problem Formulation

Consider a communication channel with a width denoted *cw* (see Fig. 1). Let $w_{\min}$ be the minimum width of the wires, *wn* the width of wires and *sn* the spacing between adjacent wires. The rest of the parameters, i.e., wire thickness, dielectric height are usually specified by the technology, so they are not design parameters.

For this channel, there are several crosstalk avoidance methods that can be employed to reduce its worst case crosstalk capacitance from $4Cc$ to $p * Cc$ to reduce bus delay, these desired properties come at the expense of the number of data carrying wires, which may reduce the link bandwidth [see (5)]. We consider three types of buses, namely the following.

A. Throughput-Centric Buses

The throughput is the most crucial aspect of this type of communication link; this is the case for non-interactive or bulk traffic. All design parameters (*p, wn, sn*) should in this case be optimized to maximize the throughput.

B. Latency-Constrained Buses

A good example where low latency is necessary is at bottlenecks such as a microprocessor to cache connections. It is well known that high cache latency can dramatically reduce the amount of work that can be usefully done by a processor. In such cases the dimensions of the wires are predefined by delay requirements, we will investigate if the throughput can be improved by optimizing the capacitive coupling factor p.

C. Power-Constrained Buses

At present low power design is of great interest driven mainly by the need to extend battery life per unit weight in mobile application [7], [8], [11].

Both performance and energy are critical in this case, which means all design parameters (*p, wn, sn*) should be optimized to gain the maximum throughput for a given bit-transition-energy (*BW/E*).

TABLE I
Analytical Solutions for Throughput-Centric Optimizations of Fixed Width Links With Nonlinear CACs

K	H	Optimum Parameters			BW(GHz)		BW Gain (%)
		p	wn	sn	Max	No Optimization	
1	2	2	1	1	23.4	19.6	19
1	5	2	1	1	28.9	24.77	17
1	10	2	2	1	30.8	25.4	21
2	5	2	1	1	41.88	37.5	12
2	10	3	2	1	42.3	39	8

TABLE II
Analytical Solutions for Throughput-Centric Optimizations of Fixed Width Links With Linear CACs

K	H	Optimum Parameters			BW(GHz)		BW Gain (%)
		p	wn	sn	Max	No Optimization	
1	2	4	1	2	20.14	19.6	3
1	5	4	2	2	27.6	24.77	11
1	10	4	2	2	30.07	25.4	18
2	5	4	1	1	37.5	37.5	0
2	10	4	2	2	40.93	39	5

TABLE III
Optimum Coupling Factor for a $65 * \mathrm{wmin}$ Bus With Nonlinear CACs ($K = 1, H = 5$, wire length $= 10$ mm)

Wire Dimensions		Optimum p	
wn	sn	Calculated	Simulated
1	1	1.72	2
2	1	1.83	2
3	1	1.94	2
1	2	2.64	2
3	2	3	3

Based on the above mentioned classification of buses, three throughput-centric optimization problems can be formulated as follows.

- *Problem 1:* For a fixed width channel, find the wire width (*wn*), the wire spacing (*sn*) and the crosstalk avoidance method (p) that achieve the maximum bandwidth (*BW*).
- *Problem 2:* For a fixed width channel with specified geometry (i.e., w, s are given), find the crosstalk avoidance method (p) which achieves the maximum bandwidth.
- *Problem 3:* For a fixed width channels find (w, s, p) that maximize the bandwidth per bit transition energy (*BW/E*).

Analytical and algorithmic solutions are provided for these three problems in the following sections.

IV. Throughput-Centric Optimization for High Performance Interconnect

An exhaustive search algorithm was implemented to find the optimal design parameters (*wn, sn, p*) that maximize the throughput for a link with a width of ($65 * w_{\min}$) and a wire length of 10 mm. We chose metal 9 in 90-nm technology to be our medium of communication. The optimization was performed on links under different buffering conditions. The input capacitance (C_d) and output impedance (R_d) of the minimum size repeater used in this work were estimated to be (39 fF) and (400 Ω), respectively. For comparison reasons, the bandwidth of the link with no optimization was calculated in each case. The results are outlined in Tables I and II.

TABLE IV
THROUGHPUT PER BIT-TRANSITION ENERGY FOR A FIXED WIDTH LINK WITH LINEAR CACS

Buffering Conditions		BW/E (ZBit/J.S)											
H	K	Sn=1, Wn=1				Sn=2, Wn=1				Sn=2, Wn=3			
		Un-coded	HS	S	D&S	Un-coded	HS	S	D&S	Un-coded	HS	S	D&S
2	1	0.84	1.08	1.72	**2.82**	2.04	2.37	2.62	2.84	0.50	1.73	2.26	1.91
2	4	0.92	1.22	1.95	2.67	1.70	2.39	2.77	**3.06**	0.46	1.59	2.07	1.65
5	1	1.12	1.43	2.20	**3.27**	2.07	2.82	3.11	3.62	0.92	3.08	**3.88**	3.16
5	4	1.45	1.86	2.72	2.93	2.47	3.32	3.69	3.45	0.79	2.67	3.30	2.41

The results in Table I show that the use of nonlinear crosstalk codes can achieves significant improvement of the link throughput (up to 21% in some cases). A combination of both coding and wire sizing is required sometimes to achieve the maximum bandwidth. An example of such case is when $(K = 1, H = 10)$.

Note that the throughput gain obtained using these methods, is higher for link with less number of buffers. This is due the fact that repeater insertion reduces the impact of capacitive crosstalk on the overall delay of a buffered line: The higher the number and/or size of the repeaters; the more the delay sensitivity to repeater delay and the less the delay sensitivity to wire delay, and to capacitive crosstalk [see (2)].

Although aggressive repeater insertion achieves high bandwidth, it comes at a very high power and area prices, which makes coding a more attractive solution.

Table II shows that linear crosstalk avoidance methods are not very useful in the context of throughput optimization; wire sizing approach achieve better results in this case. To verify the calculations, we run simulations in spectre for a standard UMC 90-nm technology. The accuracy of results in Tables I & II ranges between 80% and 90% compared to the simulations. The bandwidth gain from the simulations ranges between 6% to 25%.

V. THROUGHPUT-CENTRIC OPTIMIZATION FOR LOW LATENCY INTERCONNECT

Closed form expressions of the optimum coupling for a given bus geometry have been obtained by finding the roots of (8).

$$\frac{\partial BW}{\partial p} = 0. \tag{8}$$

For buses with nonlinear CACs the coupling factor at which the bandwidth reaches a maximal point can be calculated using (9).

$$\begin{aligned} p = \exp(&lambertw(1/Cc/(7RdK + 4RH) \\ &* (7RdKCg + 7RdK^2HCd \\ &+ 4RHCg + 7RH^2CdK) \\ &* \exp(-17/4145)) + 17/4145) \end{aligned} \tag{9}$$

where *Lambertw* is the inverse function of $f(x) = xe^x$, where e^x is the natural exponential function and x is any complex number.

For links with linear CACs, it has been found that the bandwidth reaches a minimum point at the coupling factor given in (10)

$$\mathrm{p} = 2.788 - \left[\frac{(\frac{0.7RdCg}{H}) + 0.7R_dC_d + (\frac{0.4RCg}{K}) + 0.7RHCd)}{(\frac{0.7RdCc}{H}) + (\frac{0.4RCc}{K})}\right]. \tag{10}$$

These formulae indicate that the optimum coupling factor which maximizes the throughput of a physical link depends on the interconnect length, its structure (wire width, spacing, etc.) and on the strength (H) and number (K) of the inserted buffers. In order to verify the accuracy of these formulae, we run simulations in spectre to measure the delay for different wire geometries under all possible crosstalk cases (see Table III).

It can be seen that our formula predicts the optimum solution with good accuracy in most cases. The same experiment was performed on buses with linear crosstalk codes. Simulations showed that the bandwidth achieves its maximum at $p = 4$ and its minimum at $p = 1$ for all the considered wire sizes. Equation (10) successfully predicted the minimum point to be close to $(p = 1)$.

VI. THROUGHPUT-CENTRIC OPTIMIZATION FOR LOW DYNAMIC POWER INTERCONNECT

Energy efficient VLSI design is of great interest given the proliferation of mobile computing devices. Designing low-power high performance communication links is a challenging problem because it requires us to explore the energy-performance curve. It is not enough to reduce communication energy but to be able to achieve sufficient bandwidth. Therefore, it is important to maximise the throughput performance for a given bit-transition energy (*BW/E*) for such power-constrained applications. The average energy dissipation (E) consists of two parts, the energy consumed on the links $(E_{\mathrm{interconnect}})$, and the energy consumed in the coding/decoding circuitry (E_{codec}). The former depends on the transition activity on the bus and the latter on the complexity of the codec. It can generally be stated that the more the delay reduction (i.e., less p) a crosstalk avoidance coding method achieves, the more complex its coding circuitry becomes, hence less energy efficiency. However, to the best of our knowledge, there is no formula which relates energy dissipation of a certain code with the delay reduction it achieves. Therefore, In order to obtain the optimal design parameters (*p, wn, sn*) to optimize (BW/E), we needed to implement some practical codes. We refer to each code as (n, i, p) code, where i is the number of inputs of the encoder circuits, n the number of inputs of the decoder circuits, p is the maximum coupling factor on the coded link. We considered four representatives CACs that achieve different degrees of delay reduction [6]: Forbidden Overlap Condition (FOC) code (5, 4, 3); Forbidden Transition Condition (FTC) code(4, 3, 2); Forbidden Pattern Condition (FPC) (5, 4, 2)code and One Lambda Condition (OLC) code (8,4,1). All the above codes are nonlinear, we also consider some linear crosstalk avoidance methods, namely, half- shielding (HS) with $p = 3$, shielding (S) with $p = 2$ and duplication and shielding (D&S) with $p = 1$. A $(65 * w_{\min})$ width communication link was considered, for which we studied several wire sizing options (see Tables IV and V). The crosstalk avoidance codes described above were implemented in each case. In order to do that, we had to construct the link from sub-channels, the width of which depends on the code under consideration. For example for an FOC (5, 4, 3), each sub-channel consists of 5 wires. While combining the sub-channels we made sure that there is no forbidden pattern on the boundaries.

The average energy dissipation of the codecs has been estimated form a synthesized gate-level netlists obtained using a 90-nm standard cell library. The link dynamic energy has been calculated for each

TABLE V
THROUGHPUT PER-BIT TRANSITION ENERGY FOR A FIXED WIDTH LINK WITH NONLINEAR CROSSTALK AVOIDANCE CODES

Buffering Conditions		BW/E (ZBit/J.S)														
H	K	Sn=1, Wn=1					Sn=2, Wn=1					Sn=2, Wn=3				
		Un-coded	FOC	FTC	FPC	OLC	Un-coded	FOC	FTC	FPC	OLC	Un-coded	FOC	FTC	FPC	OLC
2	1	0.84	0.96	1.91	1.57	**3.21**	2.04	2.31	2.91	2.21	2.00	0.50	1.44	2.23	1.64	1.52
2	4	0.92	1.07	2.12	1.69	2.94	1.70	2.32	**2.96**	2.20	2.14	0.46	1.32	2.03	1.47	1.31
5	1	1.12	1.26	2.42	1.95	**3.66**	2.07	2.74	3.38	2.53	2.54	0.92	2.55	3.80	2.78	2.51
5	4	1.45	1.61	2.87	2.19	3.11	2.47	3.20	**3.76**	2.71	2.38	0.79	2.20	3.20	2.28	1.89

coding method using the formula provided in [12]. It is noteworthy that the average bus energy per bus transitions is a function of the statistical distribution of the data, which depends on the application. Here we assume that the data is spatially and temporally uncorrelated, with "0" and "1" having the same probability. To assess the relevance of these techniques for specific applications, the link energy has to be estimated for the real data patterns. The wire delays were estimated in spectre in the same manner described in Section V. The throughputs per bit-transition energy calculated are listed in Tables IV and V.

The results indicates that crosstalk avoidance codes can be very useful for reducing the amount of energy needed to send data on long on chip interconnect. A comparison between the results obtained using linear and non linear CACs reveals that they have similar performances; however linear methods have no coding overheads, hence more practical. Although the use of repeater insertion method improves (*BW/E*), it is a sub-optimal technique compared to coding. Further, aggressive repeater insertion was found to degrade the energy efficiency of the link. There are optimum size (H_{opt}) and number (K_{opt}) of repeaters that maximize (BW/E), for example $(H_{\mathrm{opt}}, K_{\mathrm{opt}})$ is equal to (5, 4) for the minimally sized-wire link with an FOC technique. The optimum buffer insertion solution differs depending on the crosstalk avoidance method; this can be observed in Tables IV and V. A final remark is that achieving the maximum energy efficiency for a required throughput can only be obtained through a combination of design methods. For example the highest throughputs-per-bit transition energy obtained in our experiment for nonlinear CACs by using FTC method combined with wire sizing $(wn = 2, sn = 1)$ and repeater insertion $(K = 4, H = 5)$.

VII. CONCLUSION

Interconnects are rapidly becoming a bottleneck for the performance and cost in high-speed VLSI circuits. This paper has explored the design tradeoffs of area-constrained links. New models of the throughput have been derived which incorporate the interconnect structures (wire width, thickness and spacing), its length and the wiring overheads incurred by crosstalk avoidance methods. These expressions were then used to investigate the best combination of wire sizing solutions and crosstalk avoidance techniques that yield the maximum throughput for given metal and energy resources. For throughput-centric buses, the nonlinear crosstalk avoidance codes were found to increase the bandwidth. On the contrary, linear crosstalk avoidance methods seem to have rather negative effect, this mainly is due to their low coding rates. For latency-centric buses, it has been found that there is a clear optimum of the maximum capacitive coupling factor on the channel at which the throughput is maximized; this optimum can be calculated using the closed form expressions we derived. The significance of this result comes from the fact that it indicates that crosstalk avoidance methods may improve the link throughput in addition to its latency, this depends on the interconnect physical structure and the number of inserted buffers and their sizes. Our derived formulae of the optimum coupling factor can be utilized by the designer as a quick tool to establish whether or not employing CACs help improve the bandwidth. For power-constrained buses, the results show that a significant improvement in the energy efficiency of constrained area-links can be obtained by employing crosstalk avoidance methods. However, in order to achieve the maximum bandwidth-per bit energy transition, a combination of physical layer solutions (e.g., wire sizing and repeater insertion) and data link layer methods (e.g., crosstalk avoidance codes) must be employed. The results we have presented in this article can conveniently be used to optimize on-chip buses.

REFERENCES

[1] B. Halak and A. Yakovlev, "Bandwidth-centric optimisation for area-constrained links with crosstalk avoidance methods," in *Proc. Des., Autom. Test Conf. Eur.*, 2008, pp. 438–443.
[2] D. Pamunuwa, L.-R. Zheng, and H. Tenhunen, "Maximizing throughput over parallel wire structures in the deep submicrometer regime," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 11, no. 2, pp. 224–243, Apr. 2003.
[3] V. V. Deodhar and J. A. Davis, "Optimization of throughput performance for low-power VLSI interconnects," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 13, no. 3, pp. 308–318, Mar. 2005.
[4] ITRS, "International Technology Roadmap for Semiconductors," 2007. [Online]. Available: www.itrs.net
[5] B. Victor and K. Keutzer, "Bus encoding to prevent crosstalk delay," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des.*, 2001, pp. 57–63.
[6] S. R. Sridhara and N. R. Shanbhag, "Coding for reliable on- chip buses: A class of fundamental bounds and practical codes," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 26, no. 5, pp. 977–982, May 2007.
[7] Q. Zhang, J. Wang, and Y. Ye, "Delay and energy efficient design of on-chip encoded bus with repeaters," in *Proc. Int. Conf. VLSI Des.*, 2008, pp. 377–382.
[8] S. R. Sridhara and N. R. Shanbhag, "Coding for system-on-chip networks: A unified framework," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 13, no. 6, pp. 655–667, Jun. 2005.
[9] H. Shah, P. Shiu, B. Bell, M. Aldredge, N. Sopory, and J. Davis, "Repeater insertion and wire sizing optimization for throughput-centric VLSI global interconnects," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des.*, 2002, pp. 280–284.
[10] B. Vaidyanathan and Y. Xie, "Crosstalk-aware energy efficient encoding for instruction bus through code compression," in *Proc. IEEE Int. SOC Conf.*, 2006, pp. 193–196.
[11] Q. Zhang, J. Wang, and Y. Ye, "Low-power crosstalk avoidance encoding for on-chip data buses," in *Proc. IEEE Asia Pac. Conf. Circuits Syst.*, 2006, pp. 1611–1614.
[12] P. P. Sotiriadis and A. P. Chandrakasan, "A bus energy model for deep submicron technology," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 10, no. 6, pp. 341–350, Jun. 2002.