Abstract

ZHANG, LIANG LEON. Driver Pre-emphasis Signaling for On-Chip Global Interconnects (Under the direction of Professor Paul D. Franzon).

Signaling design for high performance VLSI systems has become an increasingly difficult task due to the delay/noise limitation for on-chip global interconnects. Repeater insertion techniques are widely used to improve the signal bandwidth of interconnect channels and to meet the delay goal of cross-chip communication, but even with a suboptimal delay approach, repeaters still consume a significant amount of power and area. They also increase the complexity of chip layout. As technologies continue to scale and operating frequencies continue to increase, the number of repeaters required increases exponentially. The intrinsic delay latency from repeaters themselves undermines total signal delay improvement.

The techniques proposed to avoid or minimize repeaters, as well as the challenge of on-chip global interconnects, are reviewed. A simplified delay design guideline is derived to determine whether inductive effects are important for the long on-chip interconnects used in this work (i.e. whether distributed RC or RLC model should be chosen).

Equalization techniques are verified as a capable solution to replace repeater insertion in achieving lower latency and higher data throughput for on-chip communication. A circuit for driver pre-emphasis is proposed by combining equalization techniques with a traditional

voltage-mode on-chip bus driver. It is demonstrated in 0.18μm CMOS technology for 10mm long interconnects at 2Gb/s. When compared to conventional repeater insertion techniques, driver pre-emphasis decreases repeater layout complexity and reduces power consumption by 12%-39% for data activity factors above 0.1.

To further improve the bandwidth and noise performance of on-chip interconnect channel, the combination of driver pre-emphasis and current-mode differential signaling is also explored in this work. A 32Gb/s 16-bit bus is demonstrated in 0.25μm CMOS technology. It reduces power by 15.0% at data activity factor of 0.1 and decreases peak current by 70%. The design is significantly less sensitive to crosstalk and delay variation, and occupies routing area comparable with conventional single-ended voltage-mode static buses.

Driver Pre-emphasis Signaling for On-Chip Global Interconnects

By

Liang Leon Zhang

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

ELECTRIAL ENGINEERING

Raleigh

2005

APPROVED BY:

PAUL D. FRANZON
Chair of Advisor Committee

JOHN M. WILSON
Co-chair

W. RHETT DAVIS

GIANLUCA LAZZI

DOUGLAS S. REEVES

Biography

Born in the quiet suburb of China's largest city, Shanghai, Liang Leon Zhang received the B.S. degree from Fudan University, Shanghai in 1997 and the M.E. degree from the National University of Singapore in 2001, both in electrical engineering. He is currently pursuing his Ph.D. degree in electrical and computer engineering at North Carolina State University. From 1997 to 1999, he was with Shanghai Belling Design Center, working on EEPROM circuits as an IC design engineer. His research interests include on-chip global communication and high-speed/mixed-signal circuit design.

Acknowledgements

As important as the achievements of defining and solving an original research problem, was the experience of knowing the people I worked with at NC State. Without them, this dissertation would not have been possible to complete and this program would not have been so enjoyable.

First, I would like to thank my advisor, Dr. Paul D. Franzon, who makes a perfect advisor for any self-motivated student. He has always been supportive and considerate through my studies. His enthusiasm and sharpness on research and his ability to keep generating creative ideas and convey them to industry have always amazed me.

I also would like to thank my co-advisor, Dr. John M. Wilson, who impresses me with his intelligence and diligence and treats me as both a student and a friend. I will certainly miss those days and nights of taking measurement with him in EGRC 301 and 410. I hope I can fulfill my promise someday by guiding him a month-long trip in my beautiful and historic home country.

I am grateful to Dr. Rizwan Bashirullah, who was my senior student then and is a professor in University of Florida at Gainesville now. His guidance on my research has always been valuable and he showed me how important commitment was to a PhD student, even to a person as smart as him.

I am also very thankful to my committee members, Dr. W. Rhett Davis, Dr. Gianluca Lazzi, Dr. Douglas S. Reeves, and Dr. Paul K. Bergey for always having time to advise me.

To my friends, Lei Luo and Jian Xu, I want to say it is always enjoyable working with smart and hardworking guys and I really like our brand name “John’s Chinese Gang”. I also would like to thank Dr. Stephen Mick for discussions on tape-out, Dr. Steve Lipa for help on wire-bonding, Fei Gao for helping me break the Alpha 21264 processor simulator, Yuantao Peng for teaching me the optimal repeater insertion algorithm, and Hao Hua for showing me the timing closure approach.

To my colleagues and friends, Evan Erickson, Karthik Chandrasekar, Julie Oh, Zhiping Feng, Ambrish Varma, Yongjin Choi, Mustafa Dagtekin, and Jayesh Nath, I enjoy the time working and hanging out with you guys. To Michele Joyner and Claire Sideri, thank you for helping me with all the paper work.

Finally, I thank my parents for their unconditional love and my Master advisor, Dr. Yongfeng Lu, for being my mentor in success, positive thinking, and motivation.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1 Motivation

Power consumption, delay and noise of global interconnects have become the major factors in deciding how long CMOS can serve the world's need for intelligent devices and communication [1], [2]. Due to the scaling nature of silicon technologies, it is no longer area, but global signaling and power dissipation that have become the limitations in integrating more functionality on a chip. Unlike local or intermediate interconnects, global interconnects do not scale down in length, since they communicate signals across a chip [3]. Together with a lack of new processes and materials based solutions for long interconnects, signaling design on global interconnects has become an increasingly important issue for circuit and architecture designers.

Conventional repeater insertion techniques have been effective at achieving lower latency and higher data throughput for on-chip RC dominated interconnects [4], [5]. However, the insertion of repeaters causes layout placement blockages that interrupt interconnect lines and circuits beneath. More importantly, the number of required repeaters increases as optimal repeater insertion spacing decreases with each technology node [7]. The power dissipation and delay latency associated with repeater themselves start to undermine the power/delay performance of global interconnects.

Because delay, throughput, power, area, and noise are all important performance metrics to be considered in on-chip signaling methodology, this work explores the possibilities of applying communication techniques to on-chip signaling by trading-off the various metrics. The idea is similar to an SRAM design [8]. Delay or power performance is improved by trading off some noise margin or signal swing, while a degradation of these properties is allowable within a confined domain of global buses, where noise levels are tightly controlled by circuit techniques and bus structures.

Transmitter equalization techniques [10] are implemented in a proposed driver pre-emphasis architecture. High frequency signal components are pre-emphasized at the driver side to improve interconnect channel bandwidth and obtain higher data rates. Current sensing and differential signaling techniques are also explored to understand bus bandwidth and robustness to crosstalk and delay variation.

1.2 Thesis Overview

In chapter 2, the various challenges of on-chip global interconnect design caused by technology scaling, delay, power, crosstalk, and inductive effects, are discussed. The techniques of on-chip global signaling, repeater insertion, low-swing signaling, current-mode signaling, transmission line behavior, and differential signaling, are reviewed based on the design performance metrics.

Chapter 3 proposes a simplified delay design guideline to suit distributed RC or RLC line models for long on-chip global interconnects. It is determined that resistive effects are still dominant for a wide range of wire parameters. For inductive lines, a first-incident-switching

delay model is derived and has accuracy within 10% of SPICE. An analysis of inductive effects on differential lines is also covered in this chapter.

In chapter 4, driver pre-emphasis is implemented using a traditional voltage-mode circuit to improve interconnect channel bandwidth. The attenuation of signal low-frequency components induced by the proposed pre-emphasis technique reduces inter-symbol-interference and reduces the power dissipation on interconnects.

Chapter 5 combines driver pre-emphasis with current-sensing differential signaling. Current-mode signaling is used to further improve the interconnect channel bandwidth and reduce dynamic power consumption, while differential signaling is used to reduce static power consumption and susceptibility to crosstalk. A novel circuit architecture and different bus structures are also explored.

Chapter 6 concludes this dissertation.

Chapter 2. On-Chip Global Interconnect

2.1 Introduction

Board and system designers have been dealing with interconnects for decades. Board-level interconnects are never an ideal communication medium. They have delay, reflections, crosstalk and require routing area. Today chip designers find themselves facing the same problems for on-chip interconnects. In this chapter, we are going to review the challenges of on-chip global interconnects and the techniques used and proposed for on-chip signaling.

2.2 Challenges of On-Chip Global Interconnects

Signaling over deep submicron global interconnects represents a major bottleneck in high performance VLSI systems due to the dominant limitation of signal propagation delays. Hence, on-chip global communication is first a delay problem and then a trade-off problem among the performance metrics as shown in Figure 2.1. For example, larger drivers and repeaters trade power, area, and noise for delay; wider wires trade area for delay; smaller pitch trade noise for area. Just like a typical analog circuit design, a high-performance on-chip global interconnect design can also treated as a multi-dimensional optimization problem.

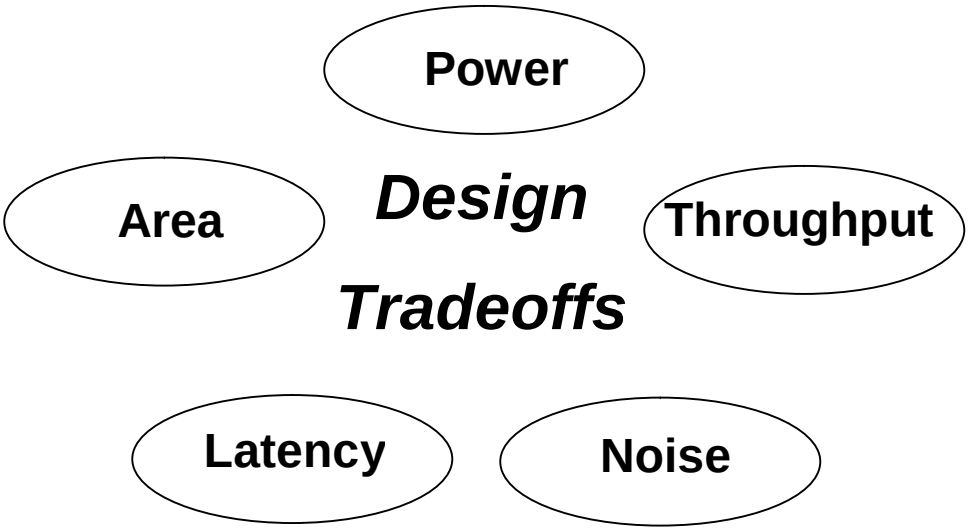


Figure 2.1 Global on-chip interconnect design tradeoffs

2.2.1 Delay latency

The implementation of copper and low-k dielectric technology mitigates the effect of scaling on signal delay in local and intermediate interconnects. However, the benefits of new material may not prove sufficient for global interconnects. Figure 2.2 shows the relative gate and wire delay scaling predicted by the International Technology Roadmap for Semiconductors (ITRS) [1]. Gate delay scales down by almost five times from technology node 250nm to 32nm. Metal-1 (local) wire delay almost tracks this trend because device gate dimensions/separations, which defines local interconnects, also scales. Global interconnects do not scale in length since they communicate signals across a chip and result in delay several hundred times worse than gate delay. Even with help from repeaters, global interconnect delay is still 10 times worse gate delay.

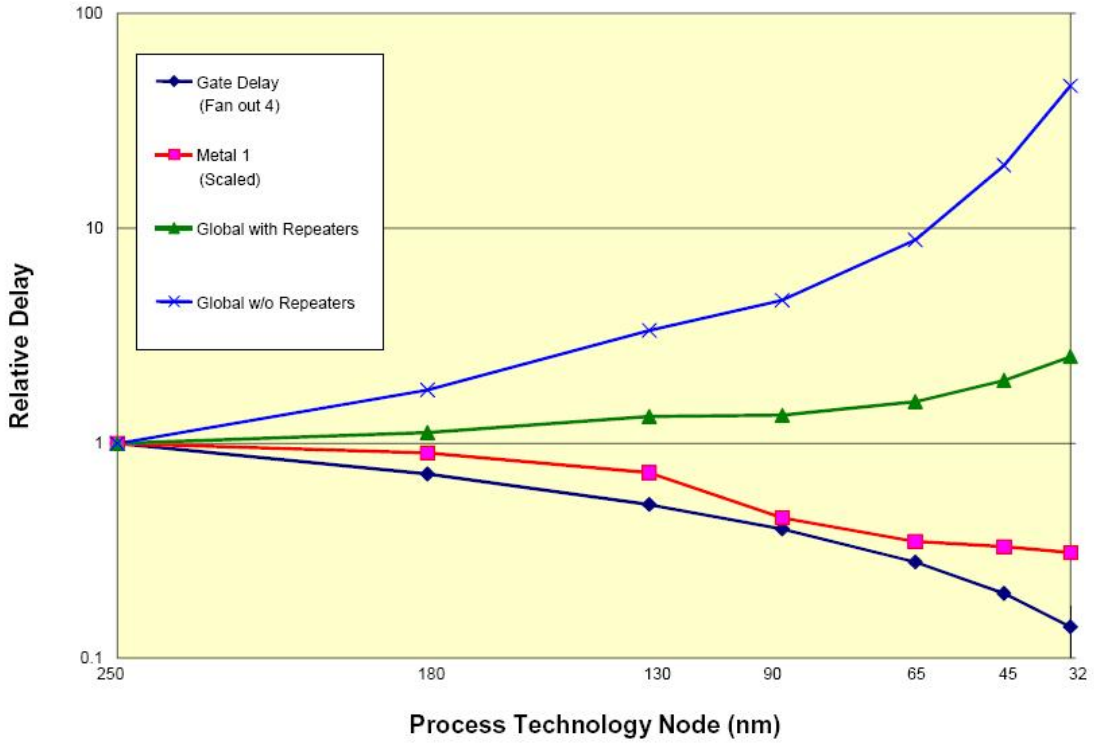


Figure 2.2 Relative delay for metal-1 and global wiring versus feature size (ITRS 2003)

Also predicted by the ITRS in Table 2.1 [1], designers are going to have to deal with a 12GHz clock frequency and 139ps global interconnect propagation delay within 5 years. Figure 2.3 shows the die picture of an Intel Itanium 2 microprocessor with chip size of about 20x20mm^2 [28] and the distance of a signal can travel during one clock cycle on such a chip, according to Table 2.1. Even if we assume chip size does not increase in the near future due to package or delay concerns, it still takes more than 30 clock cycles for a signal to travel across a chip without any advanced signaling strategies. Hence, on-chip global signaling is first a delay problem. The power and noise problems are often associated with the delay

problem or caused by the solutions to reduce delay. This is exactly the situation when we talk about the repeater insertion technique.

Table 2.1 Chip performance and MPU interconnect technology requirement – near-term (ITRS 2003)

Year of Production	2003	2004	2005	2006	2007	2008	2009
Technology Node	100	90	80	70	65	57	50
On-chip local clock (MHz)	2,976	4,171	5,204	6,783	9,285	10,972	12,369
Interconnect RC Delay (ps) for 1mm global line at minimum pitch	42	55	69	87	92	112	139

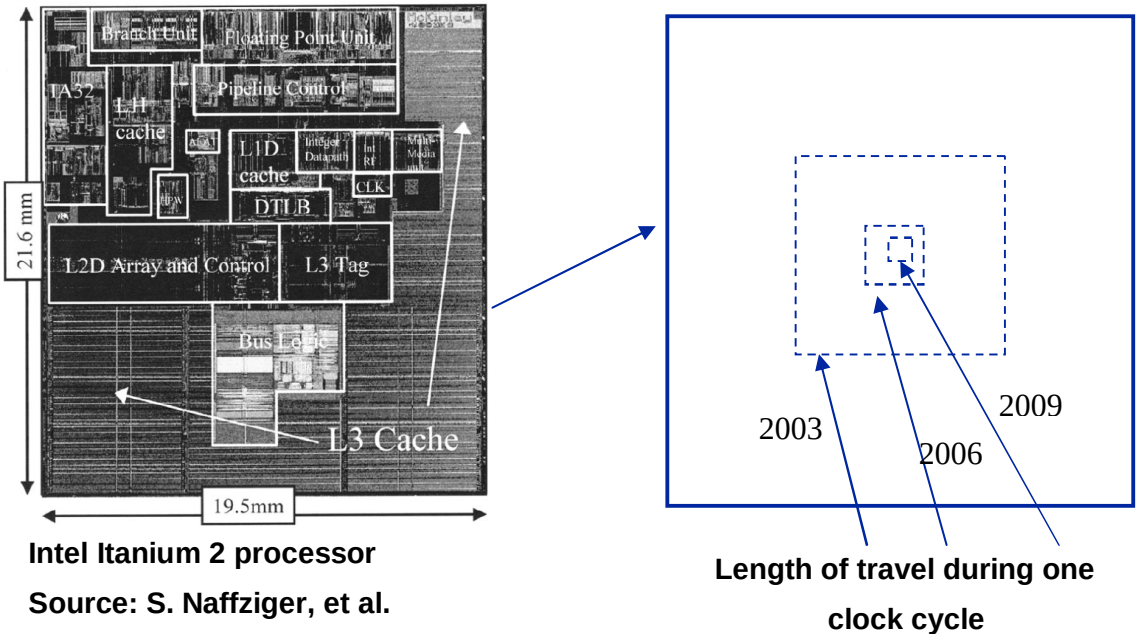


Figure 2.3 Signal travel length during one clock cycle

2.2.2 Repeaters

Repeater insertion techniques effectively improve the data rate for long on-chip interconnects by changing the quadratic relationship between line delay and line length to a linear relationship (Figure 2.4) [4]. Up to 80% of on-chip global interconnects in high performance ICs require repeaters to meet the delay goal [5]. Repeaters divide a long line into several shorter segments. This makes the coupling distance of long parallel lines shorter and prevents inductive effects. Significant work has been done on delay, power, or noise optimal repeater insertion techniques [18], [7], [3], [13], [14]. However, even with a suboptimal design, repeaters are still large devices and therefore require a significant amount of power and area. They also cause layout placement blockages to interrupt a line with repeaters, and complicate the placement of circuits beneath the line. More importantly, the number of required repeaters increases as optimal repeater insertion spacing decreases with each technology node [7]. The power dissipation and delay latency associated with repeater themselves start to undermine the power/delay performance of global interconnects.

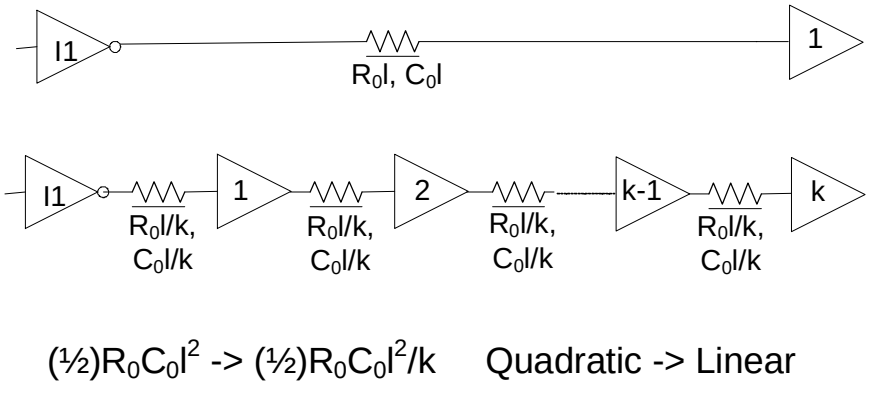


Figure 2.4 Repeater insertion techniques.

2.2.3 Crosstalk

Figure 2.5 shows a typical chip cross section [1]. Up to 14 layers could be needed to meet on-chip communication requirement. To prevent interconnects resistance from increasing too fast, meal thickness scales much slower than width. This explains the changing aspect ratio as technology scales and accounts for the large fraction of coupling capacitance in total line capacitance (Figure 2.6) [17]. CCM is the coupling capacitance multiplier factor between two lines (CCM is 0 for transitions in the same direction, 1 when there is no transition and 2 for transitions in opposite directions).

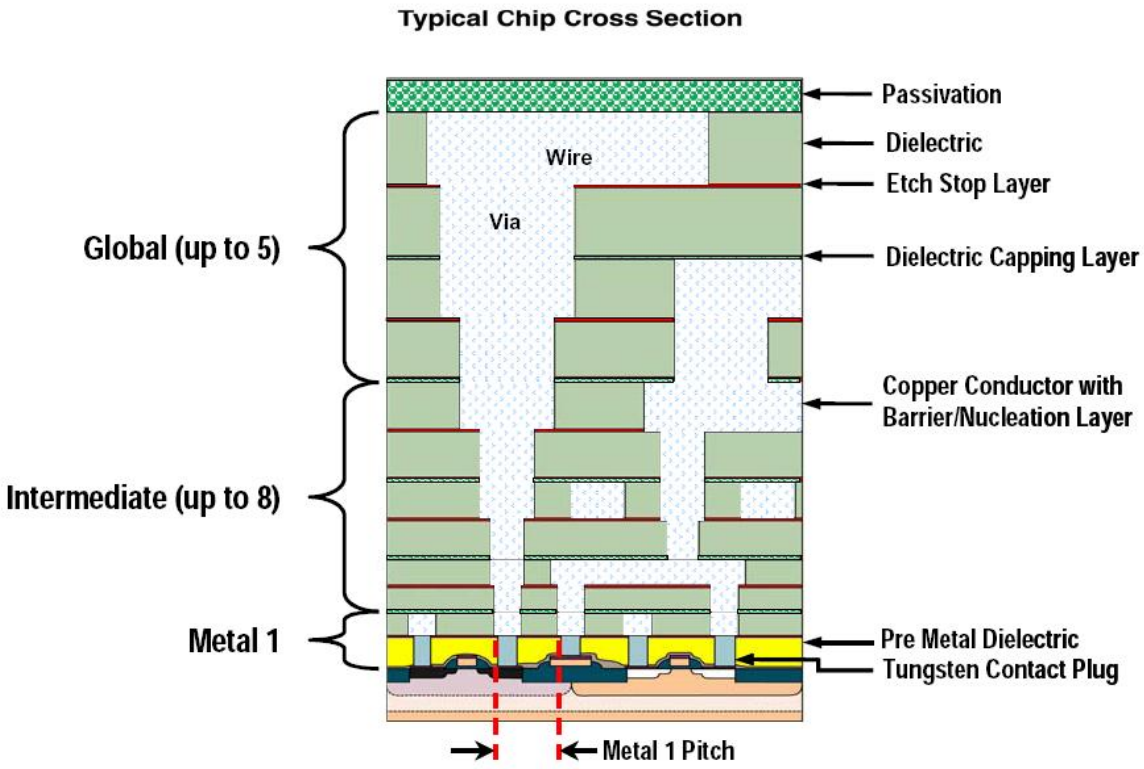


Figure 2.5 Cross-section of hierarchical scaling (ITRS 2003).

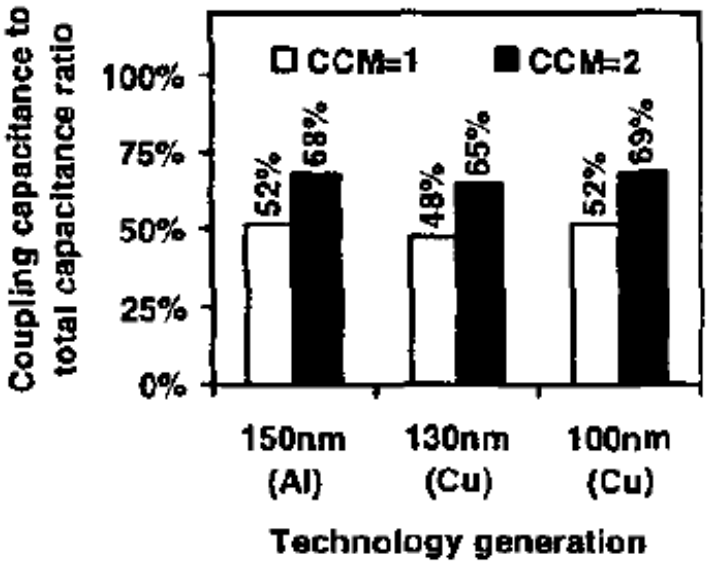


Figure 2.6 Technology scaling trends of metal-4 interconnect coupling capacitance [17].

A wiring solution with extensive spacing and shielding is simple and robust, but it is not optimal for a high-volume product [14]. A more aggressive wiring solution is likely to be chosen, but more risk comes from the increased susceptibility to crosstalk noise and data-dependant delay induced by such an aggressive solution. Interconnect noise-immune designs have become one of the major design issues in on-chip signaling.

2.2.4 Inductive effects

Global wires are typically routed in top-level metal layer with large cross section to reduce resistance. When increases in signaling frequency and signal edge rate are combined with lower resistance, inductance and related current return path issues, are quickly becoming an important consideration for on-chip signal integrity for global wiring [11]. While capacitance extraction can be restricted in a region around conductors of interest without

losing accuracy [19], no adequate extraction models exist for on-chip wire inductance due to the long-range effect of inductance. Partial-element-equivalent-circuit (PEEC) [20] or other full-wave solutions are accurate for small problem sizes. These techniques are extremely computationally expensive and are not feasible for a whole chip problem. Moreover, closed-form RC delay models have proven to be able to accurately estimate the delay for RC interconnects [12], [21], while RLC delay models are still immature and inadequate. Given this, the uncertainty of inductance effects causes severe signaling reliability problems.

2.3 On-Chip Signaling Techniques

Many signaling techniques have been proposed to cope with the challenges of global interconnects. They are trade-offs among the on-chip signaling performance metrics: delay, throughput, power, noise, and area, for different capacitive, resistive and inductive parasitics of the interconnect wires.

2.3.1 Low-swing signaling techniques

It is very effective to reduce the power consumption of global communication by reducing signal swing. The equation for dynamic power as a function of voltage swing on interconnect is,

$$P_{dyn} = afC_L V^2 \qquad (2\text{-}1)$$

Where α is the data activity, f is the working frequency, C_L is the total of wire and load capacitance, and V is the signal swing voltage. For the low-swing interconnect circuit shown in Figure 2.7, the driver circuit decreases the typical voltage swing to a reduced value, while

the receiver detects the signal and restores it to the normal swing value. Figure 2.8 shows the possible driver and receiver circuit. The signal is reduced to a lower voltage, VDD_L, and converted back to full-sing VDD by a cross-coupled structure at the receiver side. Significant power savings up to a factor of six have been observed [22].

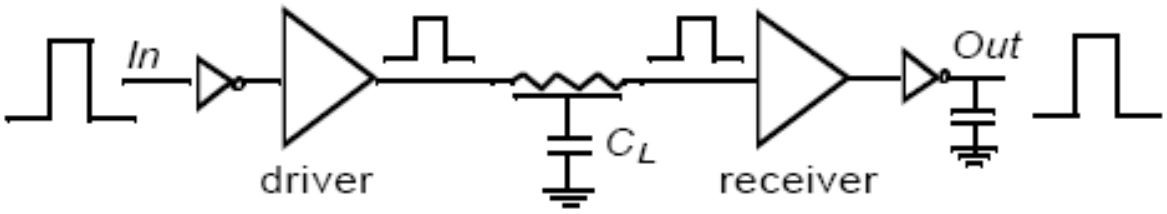


Figure 2.7 Low-swing interconnect circuit.

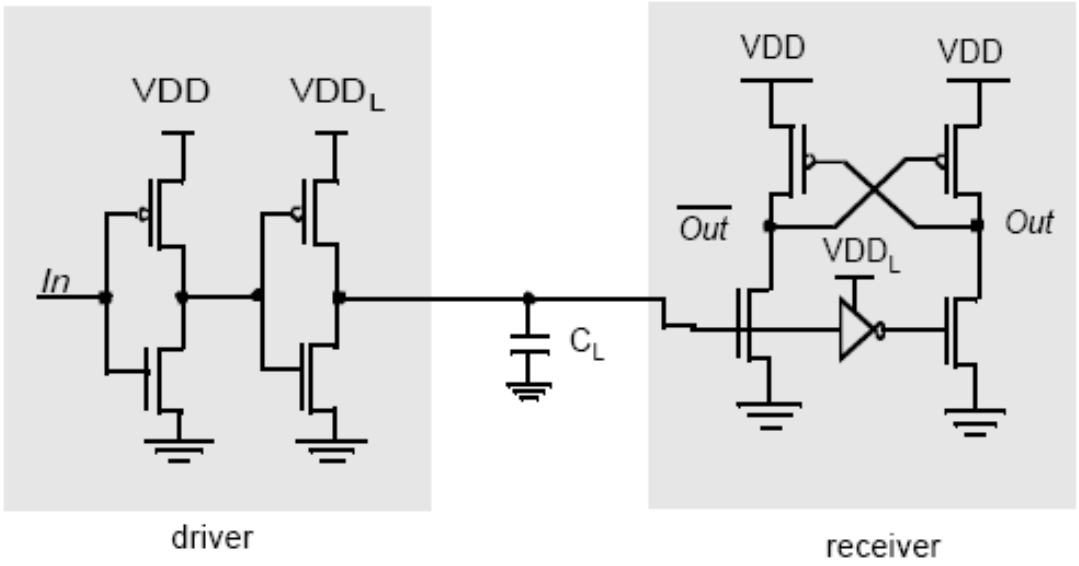


Figure 2.8 Simple low-swing driver and receiver circuit [22].

In addition to the short-circuit power of the low-swing receivers, an amplifier was often needed to restore the swing at the receiver side and it has a power consumption related to its gain. Although the swing reduction on wire dominates the power saving, there was an optimum voltage swing for minimum power [23]. The low-swing schemes have longer delays than the full-swing scheme. Isolation was also important between low-swing and full-swing signals to handle crosstalk noise. Therefore, low-swing signaling techniques generally sacrifice both noise-margin and bandwidth for power dissipation.

2.3.2 Low-swing differential bus

A low-swing differential interconnect architecture with distributed line equalization was proposed in [24]. As shown in Figure 2.9, the low swing differential signals on the long wires were equalized to a middle-level voltage for every clock cycle by evenly inserted N-transistors. This reduced the RC rise time and eliminated inter-symbol interference (ISI). The low-swing signal was recovered and restored to full-swing at the receiver side by a regenerative sense-amplifier with S/R latch (Figure 2.10).

This architecture used a clock signal to control the N-transistor equalizers. It increased the load of the clock distribution network, which is already another design challenge. In addition, just like repeater insertion technique, it added layout blockages by inserting transistors along the long wires.

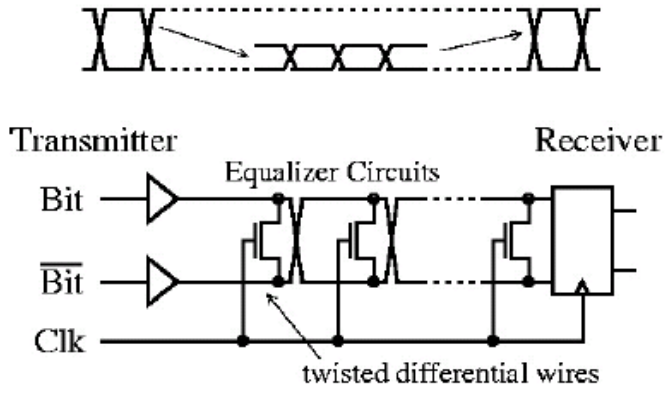


Figure 2.9 A low-swing differential interconnect architecture with distributed line equalization [24].

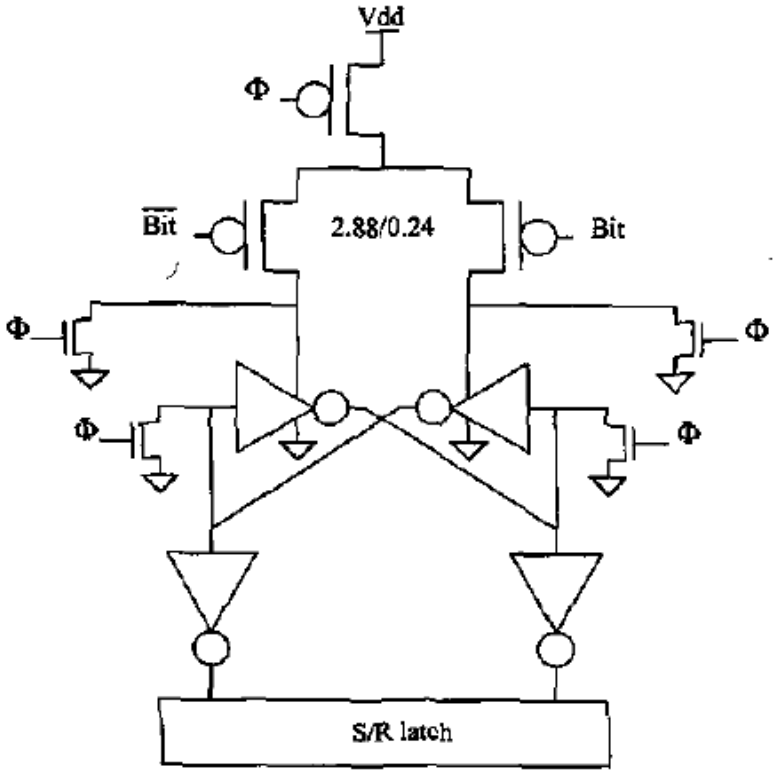


Figure 2.10 A regenerative sense-amplifier wth S/R latch.

2.3.3 Current-mode technique

Speed improvements on long interconnect lines are possible by using current-mode (CM) signaling rather than conventional voltage-mode (VM) techniques [29]. For a distributed RC interconnect model with source and load termination, as shown in Figure 2.11, its delay based on continuous ramp signals can be derived as,

$$dt = \frac{R_T C_T}{2} \cdot \left(\frac{R_B + \frac{R_T}{3} + R_L}{R_B + R_T + R_L} \right) + R_B C_T \cdot \left(\frac{R_L}{R_B + R_T + R_L} \right) \tag{2-2}$$

Where R_T and C_T are the total line resistance and capacitance, respectively, R_B is the driver output resistance, and R_L is the line termination resistance. R_L goes to infinity for a VM signaling case.

Noticed from (2-2), the delay can be minimized by either making R_B or R_L small. In VM signaling, the only option for designers is to make big drivers to have small R_B, while in CM singling, designers can play with both approaches.

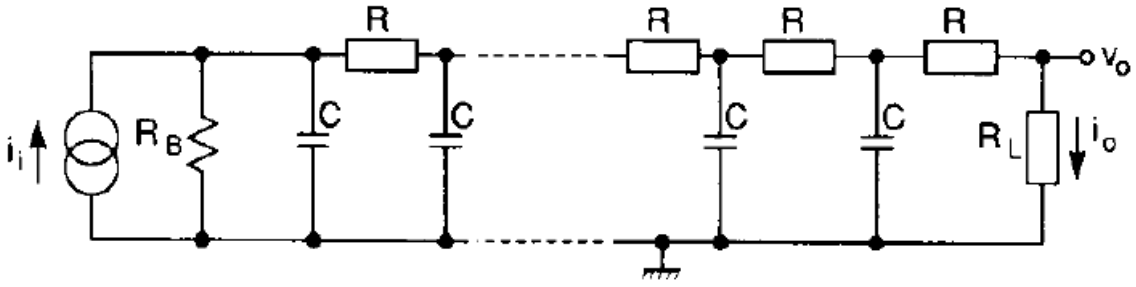


Figure 2.11 Distributed RC interconnect model with source and load termination.

Figure 2.12 shows a current sense-amplifier circuit for an SRAM cell [29]. The cross-coupled structure keeps both bit-lines at the same voltage V1+V2 by sizing and transitioning in saturation region. Hence, the bit-line load currents I+i are equal, as well as the bit-line capacitor currents iC. If I is the current the cell draws when accessed, it is then passed to the differential data-lines.

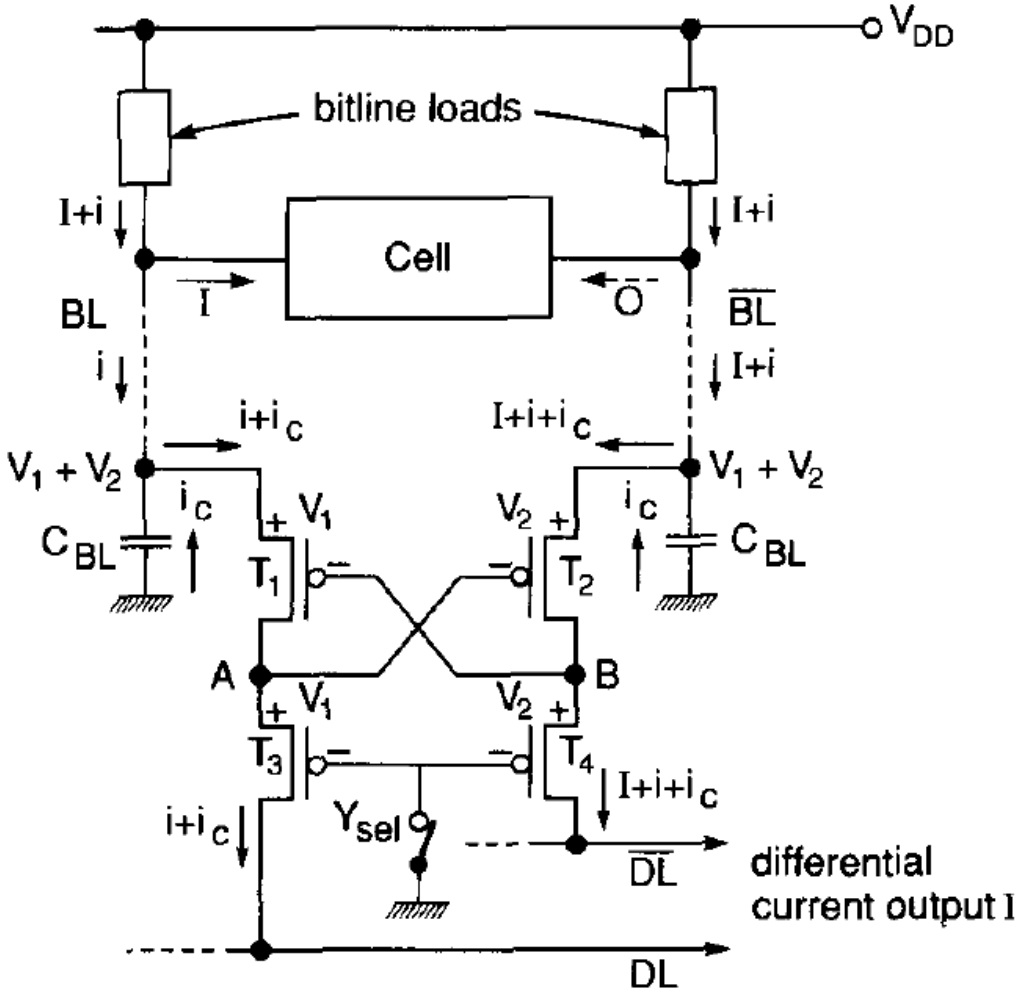


Figure 2.12 A current sense-amplifier circuit for SRAM [29].

2.3.4 Hybrid current/voltage mode bus

An ingenious adaptive bandwidth bus scheme based on hybrid CM/VM repeaters was reported in [25]. It switched to current mode at high data activity to take the advantage of the higher bandwidth of CM signaling and switched to voltage mode at low data activity to save the CM static power dissipation (Figure 2.13). As shown in Figure 2.14, this function can be simply implemented by changing the line termination resistance R_L. Rise time is long when R_L is large, but it has negligible static current consumption, while at small R_L, rise time is short at the cost of large static current consumption.

This proposed architecture has its disadvantages. It requires pipeline latency to accommodate its computational data-paths which determine the data activity before sending out the data to the adaptive bandwidth bus. Moreover, its power saving was not significant for low data activity bus.

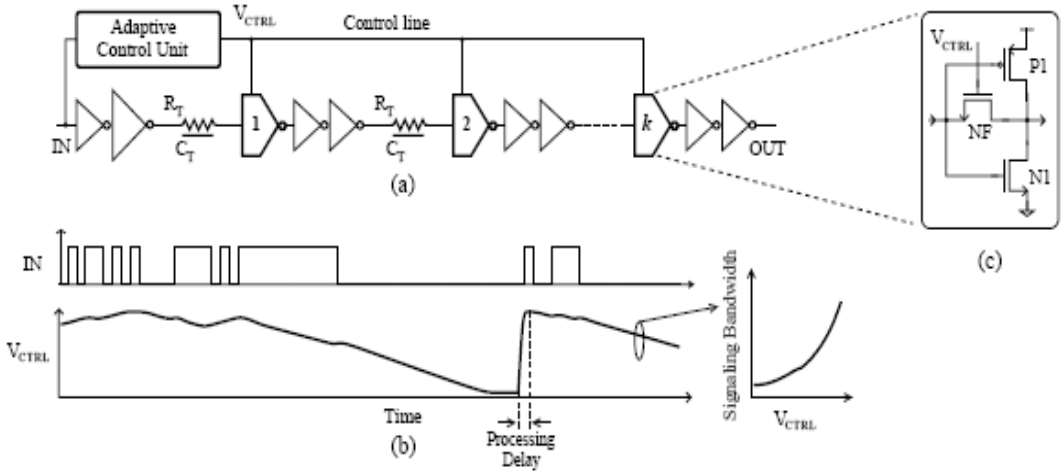


Figure 2.13 A Hybrid current/voltage mode bus [25].

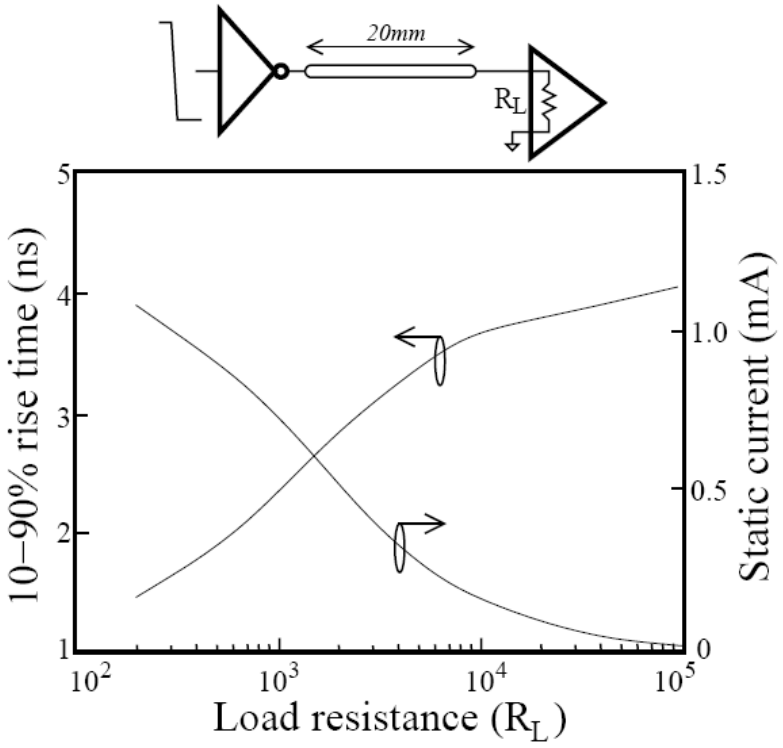


Figure 2.14 Effect of line termination resistance on rise time and static current [25].

2.3.5 Pulse-width pre-emphasis technique

A similar current sensing technique used in [25] was also presented in [26] (Figure 2.15). A twisted differential bus structure was used to alleviate intra-bus crosstalk. Figure 2.16 shows its driver circuit with pulse-width pre-emphasis. It always uses the second part of the symbol time to compensate for the remaining line charge and reduce ISI. However, this differential current-sensing bus consumes even more power than [25]. Its power dissipation performance is even worse than that of the tradition VM single-ended bus for data activity factors below 0.5.

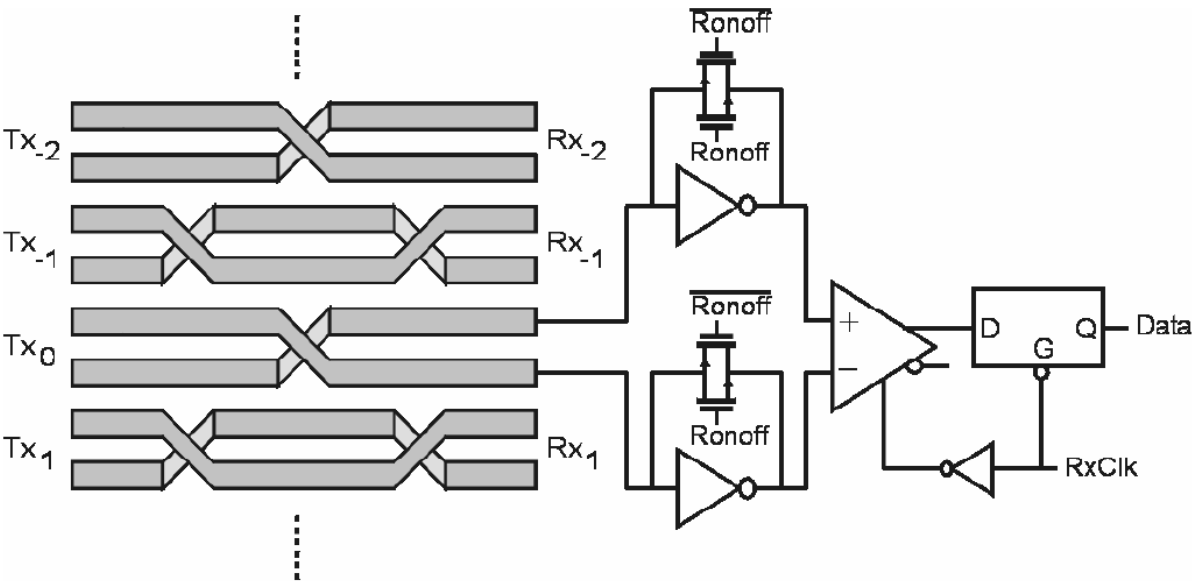


Figure 2.15 Current sensing receiver with twisted differential bus [26].

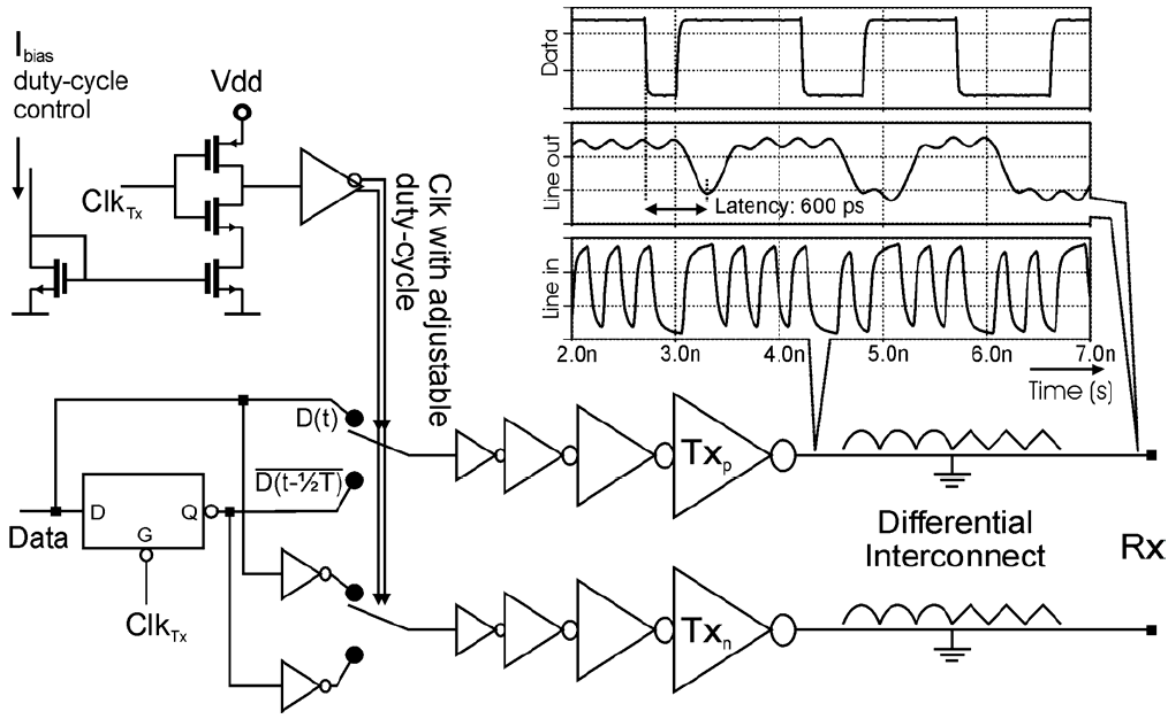


Figure 2.16 Driver with pulse-width pre-emphasis [26].

2.3.6 Inductance dominated interconnects

Instead of limiting and suppressing the inductance effects on interconnect, it has been proposed to emphasize the parasitic inductance and reduce the resistance [27]. As shown in Figure 2.17, this takes the advantage of the inductance-dominated high-frequency region of on-chip interconnects. Signal propagation near the speed of light in the dielectric surrounding interconnect was achieved. Figure 2.18 shows the modulation circuit to push the signal spectral component to the high-frequency region and to eliminate the low-frequency component that lags behind and contributes to ISI. Very wide interconnects were required and special care was needed for crosstalk prevention in this architecture.

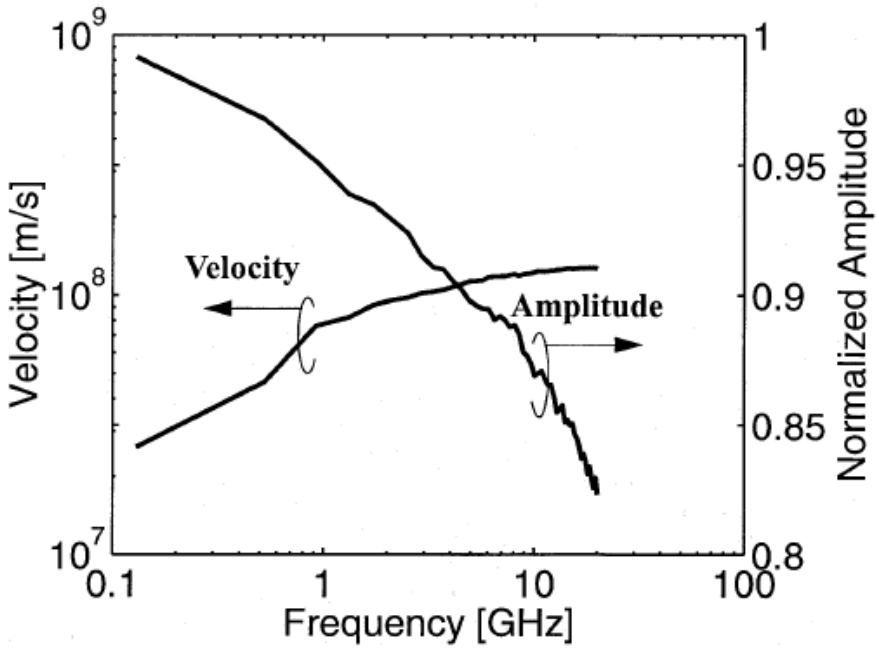


Figure 2.17 Frequency effect on velocity and attenuation for a 1-mm long coplanar waveguide [27].

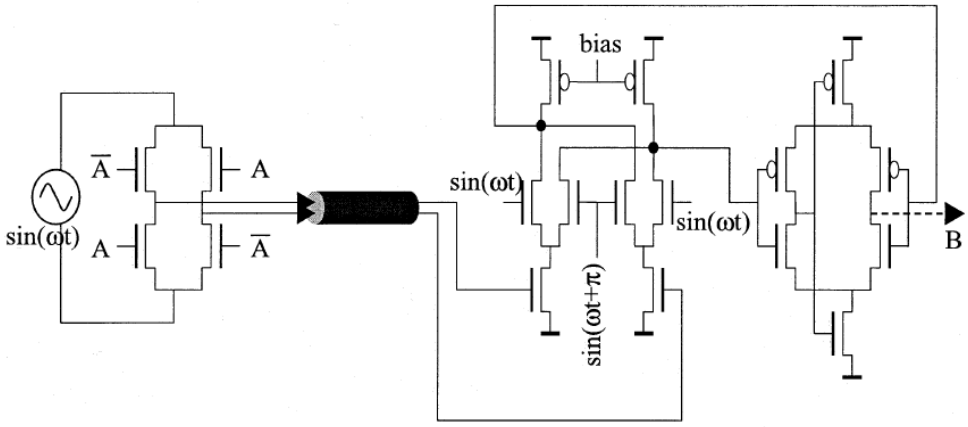


Figure 2.18 Simplified circuit schematic of an inductance-dominated interconnects system [27].

2.4 Summary

The on-chip signaling performance metrics: delay, throughput, power, noise, and area, are some of the challenges that designers are facing today, but they are also trade-offs that designers can manipulate for on-chip signaling strategies. The chart in Figure 2.19 summarizes this chapter and profiles the following parts of this thesis.

Inductive effects are becoming important, but thanks to the technology scaling, on chip interconnect is so resistive that it limits inductive effects to only wide wires in clock distribution networks. The distributed RC model is still accurate and sufficient for other applications in on-chip global communication. On-chip interconnects are discussed in Chapter 3.

The repeater insertion technique is still the most effective way to reduce delay on long wires by separating lines with invertors, buffers, or flip-flops. It is mainly a trade-off between

power and delay. It also helps reduce inductive effects and intra-bus crosstalk by breaking long lines. However, it increases simultaneous switching noise (SSN) and complicates layout.

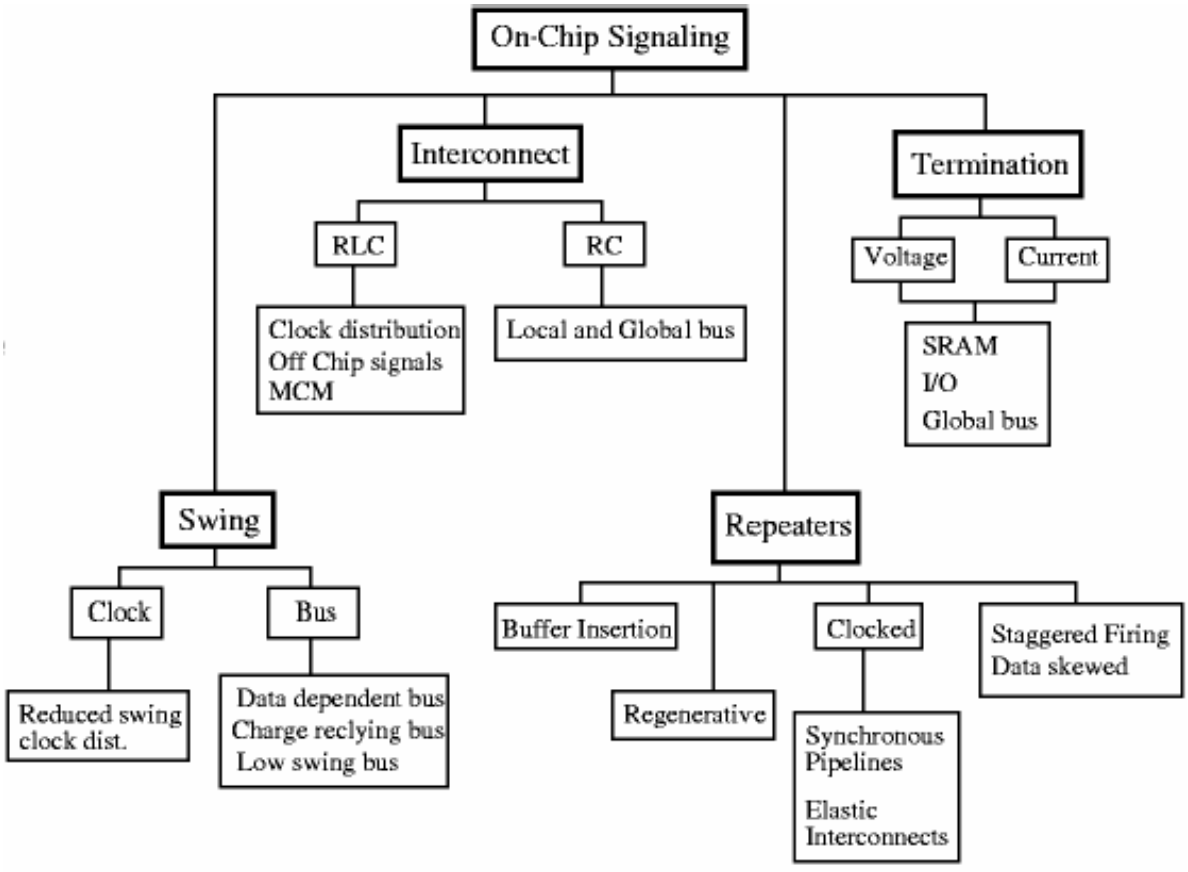


Figure 2.19 On-chip signaling techniques.

On-chip global communication could contribute up to 45% of the total power in a microprocessor design [56]. This makes low-swing signaling techniques attractive by trading off noise-margin and bandwidth for power dissipation. Chapter 4 and 5 will illustrate how to gain both power and bandwidth improvement by using low-swing signaling.

VM termination is still the dominant on-chip signaling technique. On-chip application of CM termination is limited to SRAM and I/O designs because of the concerns of its static power consumption and noise suceptability to full-swing signal. How to apply CM signaling to on-chip interconnects with acceptable static power overhead and noise robustness will be discussed in Chapter 5.

Chapter 3. On-Chip Inductive Effects

3.1 Introduction

In this chapter, a simplified delay design guideline is derived to determine whether inductive effects are significant for long on-chip global interconnects and to determine whether the distributed RC wire model is still accurate and sufficient for on-chip global signaling.

It is meaningful to analyze on-chip inductive effects because inductance causes severe signal reliability problem. It is generally an unpredictability problem, as a result of the extremely computationally expensive extraction of on-chip inductance, the increased coupling noise induced by on-chip inductance, and the inaccuracy of on-chip RLC interconnects delay formulas.

Modern technologies optimize their metal layers for three different tasks, lowest level metals for local interconnections; middle level metals typically for functional unit connection; top layer metals for global signaling and power distribution. Hence, on-chip interconnects can be divided into three categories, short, medium and long lines, as shown in Table 3.1 [11]. Although it is possible for short or medium lines to behave inductively if they are very wide and driven by very large drivers, the very high integration density desired on a chip limits the wiring dimensions and driver sizes for these lines. Therefore, short and medium lines are still resistance dominated, and the inductance problem is limited to long global lines.

Table 3.1 The categories of on-chip interconnect based on different tasks [11].

	Short Line	Medium Lines	Long Lines
Deutsch's three categories	R > 1000 Ω/cm, l < 500μm, Z_{drv} > 7Zo	R ~ 300-500 Ω/cm, l ~ 1-3 mm, Z_{drv} < 3Zo	R < 100 Ω/cm, l ~ 0.5-1cm, Z_{drv} ~ Zo

3.2 Simplified Delay Design Guideline

A closed-form interconnect delay formula is desirable in a high-speed circuit design to predict and control global wire delay. As shown in Figure 3.1, an analytical delay formula for interconnects can be used to avoid accurate but computationally expensive SPICE simulation and save the iteration time in a performance-driven synthesis or a global routing topology.

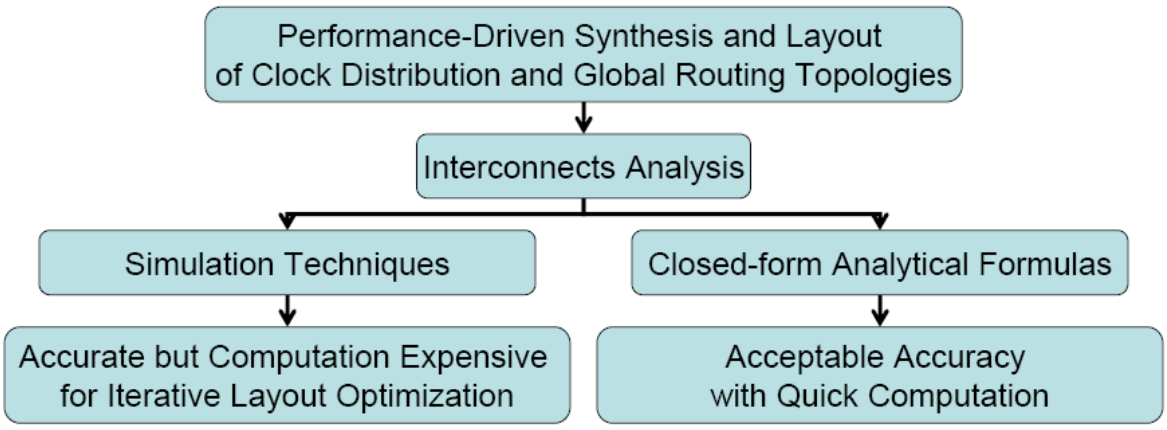


Figure 3.1 High level interconnects design flow.

For resistance-dominant interconnects, closed-form RC delay formulas ([35], [12]) have been proven acceptable and sufficient. Therefore, the delay problem of long global interconnects becomes a problem of if the interconnects are still RC dominant, or if not, how accurate a RLC delay formula can be.

The conventional RLC interconnect criteria to determine when inductive effects should be taken into account ([11], [30], [31]) mainly look at whether the near-end rise time of signal is much faster than the signal propagation velocity on the wire or whether the damping factor $Z_0/2R_{wire}$ is greater than one, where Z_0 is the wire characteristic impedance and R_{wire} is the distributed wire resistance. These rules actually limit the effective range of RC delay formula by ignoring source and load termination effects. The resistance and capacitance associated with the termination actually make an interconnect line more RC dominant.

Figure 3.2 elaborates the approach used in this work to find the design guideline determining whether inductive effects are important. An interconnect line with source and load termination is first described by a general transmission line transfer function [34]. Second, this function is approximated by a two-pole simplification [33]. Third, the effective damping factor is derived [6]. At last, a RC model or a transmission line model can be applied respectively according to the damping factor.

For the on-chip interconnect line with source and load termination shown in Figure 3.3, its transfer function can be written as [34],

$$H(S)=\frac{V_2(S)}{V_1(S)}=\frac{1}{\left[\cosh(qh)+\frac{Z_S}{Z_o}\sinh(qh)\right]+\frac{1}{Z_L}\left[Z_0\sinh(qh)+Z_S\cosh(qh)\right]} \quad (3\text{-}1)$$

Where $Z_S = R_S$ and $Z_L = \frac{R_L}{1 + R_L C_L S}$ are the source and load impedance, h is the line length, $q = \sqrt{(R_0 + SL_0)SC_0}$ is the propagation constant, $Z_0 = \sqrt{(R_0 + SL_0)/SC_0}$ is the characteristic impedance, and R_0, L_0, and C_0 are the distributed line parameters.

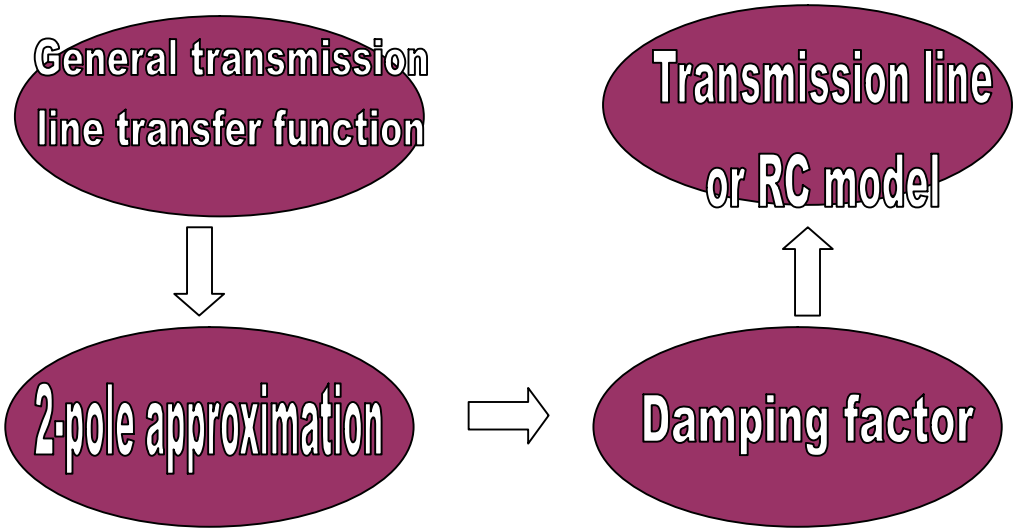


Figure 3.2 Approach to the simplified delay design guideline.

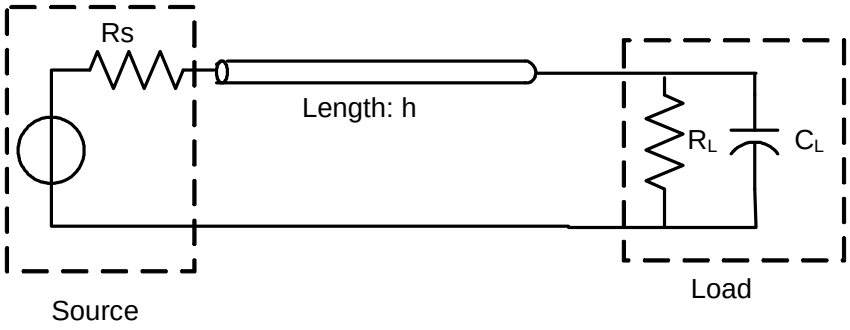


Figure 3.3 An on-chip interconnect line with source and load termination.

This transfer function can be approximated as a two-pole system by expanding cosh and sinh,

$$H(S) \approx \frac{m_0}{1 + sb_1 + s^2 b_2} \tag{3-2}$$

Where $b_1 = a_1 m_0$, $b_2 = a_2 m_0$, and m_0 serves as the zero moment and indicates the system DC gain,

$$m_0 = \frac{R_L}{R_S + R_T + R_L} \tag{3-3}$$

$$a_1 = R_S (C_T + C_L) + \frac{R_T C_T}{2} + R_T C_L + \frac{R_S R_T C_T + 2L_T}{2R_L} \tag{3-4}$$

$$a_2 = \frac{1}{6} R_S R_T C_T^2 + \frac{1}{2} R_S R_T C_T C_L + \frac{(R_T C_T)^2}{24} + \frac{1}{6} R_T^2 C_T C_L + \frac{L_T C_T}{2} + L_T C_L + \frac{\frac{1}{3} R_T L_T C_T + \frac{1}{120} R_T^3 C_T^2 + \frac{1}{24} R_S R_T^2 C_T^2 + \frac{1}{2} R_S L_T C_T}{R_L} \tag{3-5}$$

Where $R_T = R_0 h$, $C_T = C_0 h$, and $L_T = L_0 h$ are the line resistance, capacitance and inductance.

The transfer function of a two-pole system can also be expressed as,

$$H(S) = \frac{m_0}{\frac{S^2}{w_n^2} + \frac{2x_{eff} S}{w_n} + 1} \tag{3-6}$$

From (3-2) and (3-6),

$$w_n = \frac{1}{\sqrt{b_2}} \tag{3-7}$$

$$x_{eff} = 2\frac{b_1}{\sqrt{b_2}} \tag{3-8}$$

By analogy with the line attenuation constant $x = \frac{Z_0}{2R_0 h}$, ξ_{eff} is called the effective attenuation constant and used as the criterion to determine whether inductance effects are significant. By doing this, the interconnect line and the termination impedance are treated as a whole system. $\xi_{eff} \geq 1$ indicates a RC region and RC delay formula are still accurate for RLC interconnects, while $\xi_{eff} < 1$ indicates a RLC region and RLC delay formula is required.

3.3 Guideline Verification

3.3.1 RC delay formula

Closed-form RC delay formulas (3-9) and (3-10) ([35], [12]) are used to verify the proposed delay design guideline. In (3-9) it was proposed for voltage-mode signaling and in (3-10) it was proposed for both voltage-mode and current-mode signaling.

$$\frac{t_v}{R_T C_T} = 0.1 + \ln\left(\frac{1}{1-v}\right)\left(R_{ST}C_{LT} + R_{ST} + C_{LT} + 0.4\right) \tag{3-9}$$

$$\frac{t_v}{R_T C_T} = \frac{\ln\left(\dfrac{1.2 + 0.5\left(m_0 / R_{LT}\right)}{1-v}\right)}{1 + 1.0058\ln\left(1.2 + 0.5\left(\dfrac{m_0}{R_{LT}}\right)\right)} * \left(\frac{m_0}{2}\left(1 + \frac{1}{3R_{LT}}\right) + m_0 R_{ST}\left(1 + \frac{1}{2R_{LT}}\right) + m_0 C_{LT}\left(1 + R_{ST}\right)\right) \tag{3-10}$$

Where $R_{ST} = R_S / R_T$ $R_{LT} = R_L / R_T$ $C_{LT} = C_L / C_T$ and delay t_v is defined as the time from (t=0) to the time when the normalized voltage reaches threshold v at the receiving end of the line.

Table 3.2 compares the results from HSPICE simulation and the RC delay formula for a line with designated parameters. These parameters (R_0 = 150Ω/cm, L_0 = 2.46nH/cm, C_0 = 1.76pF/cm, Rs = 250Ω, $R_L = \infty$, C_L = 250fF, and v_{th} = 0.5) are chosen to show the worst case of inductive effects in an on-chip bus environment using aluminum. A typical global bus line is more RC dominant and therefore less likely to behave inductively. Although the traditional wire damping fact ξ indicates the wire is inductive, both of the RC delay formulas maintain good accuracy along a large range of line lengths. Figure 3.4 also shows the convergence between SPICE simulation and the RC delay formulas.

The limitation of wire damping fact is caused by its ignorance of termination. An inductive line with adequate source and load impedance still behaves like a RC line. The effective damping factor gives the effective range of a RC domain by treating a line and its termination as a whole system.

Table 3.2 Comparison between HSPICE simulation and the RC delay formula (R_0 = 150Ω/cm, L_0 = 2.46nH/cm, C_0 = 1.76pF/cm, Rs = 250Ω, $R_L = \infty$, C_L = 250fF, and v_{th} = 0.5).

Length (um)	1	10	100	1000	10000	20000
ξ	2.0x10-4	2.0x10-3	2.0x10-2	0.20	2.0	4.0
ξ_{eff}	126	39.8	12.5	4.08	1.80	1.52
RC delay (ps)	46.27	46.59	49.78	82.54	497.2	1143
HSPICE (ps)	43.27	43.59	46.65	78.17	482.7	1132
Error	6.9%	6.9%	6.7%	5.6%	3.0%	1.0%

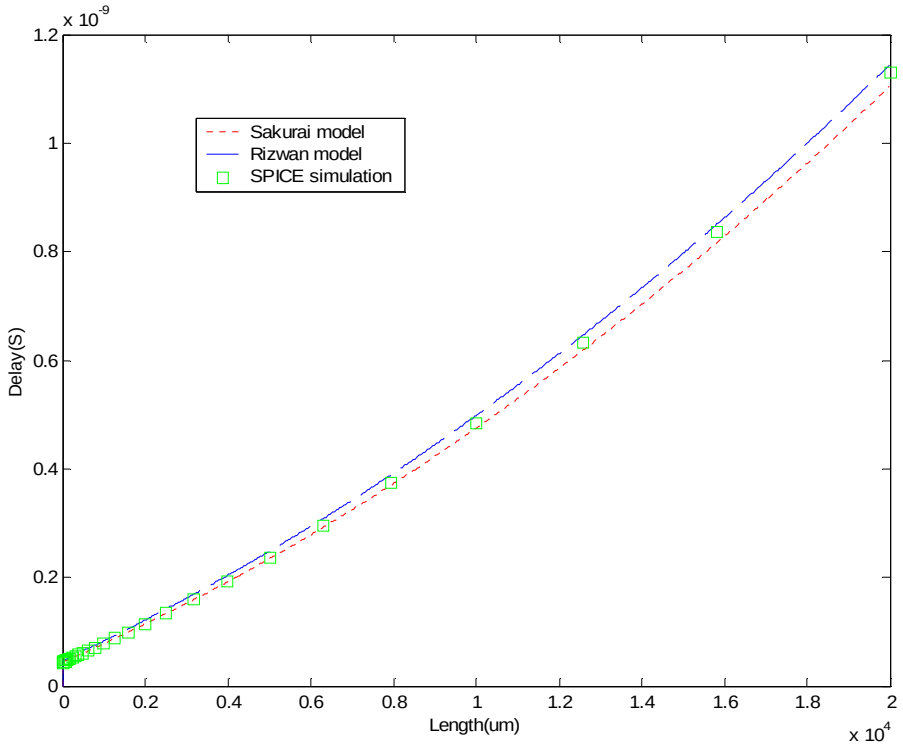


Figure 3.4 Comparison between SPICE simulation and the RC delay formula (R_0 = 150Ω/cm, L_0 = 2.46nH/cm, C_0 = 1.76pF/cm, Rs = 250Ω, R_L = ∞, C_L = 250fF, and v_{th} = 0.5).

3.3.2 Existing RLC delay formulas

Although aluminum global bus lines were just proven RC dominant, inductive effects can be dominant in a copper process and have been reported important on wide lines of a clock distribution network [40]. It is therefore meaningful to find an approximated closed-form analytical formula for RLC interconnects.

A formula based on first and second moments [33] or a formula based on the damping factor of a two-pole system [6] generally imply the same order of moment information,

because the first 2q-1 moments can approximate the response of a lower q-pole model [32]. These formulas were able to model some non-monotonic behavior; given this, they were more accurate than the first-moment-based Elmore Delay [36]. However, much higher orders of moments are required to model a transfer system with significant inductive effects (i.e. transmission line effects). Therefore, these formulas without moment order higher than two could only be accurate within a very limited range of parameters.

Another analytical formula modeled RLC interconnects as both RC distributive line and lossy transmission line and chose the maximum value as the delay [9]. Unfortunately, this maximum criterion was not true all the time and it failed to clarify the grey region which could not be modeled by either RC or lossy transmission line model [4].

In the next two sections, we propose to use design guidelines to rout interconnects in either the RC or the RLC region and derive a delay formula based on first incident switching for RLC region.

3.3.3 First incident switching delay formula

In the RLC region, lossy transmission line models become more accurate than distributed RC line models. Because most transmission line analysis is performed in the frequency domain, an inverse Laplace transform would be computationally expensive when a transmission line analysis has to span a large time interval. In this work, instead of deriving a lossy transmission line delay model in the frequency domain, a first incident switching delay model is directly obtained in the time domain.

The smallest propagation delay on a transmission line is obtained if the first signal arriving at the end of the transmission line has sufficient magnitude to switch the gate. Otherwise, at least one extra round trip time interval will be needed. For this reason, the closed-form formula in the RLC regions is derived based on first incident switching to achieve minimum delay.

A transmission line can be modeled at various points along the line by using circuit models [4]. Similarly, we model a lossy transmission line at points I, X and L (Figure 3.5). For modeling point I,

$$V_i(0) = \frac{Z_0}{Z_0 + R_S} V_S \tag{3-11}$$

Where V_S is the normalized step input voltage and $Z_0 \approx \sqrt{L_0 / C_0}$. For modeling point X, the voltage at point X + dh would be,

$$V_X(X + dh) = \frac{Z_0}{2Z_0 + R_0 dh} 2V_X(X) \tag{3-12}$$

When the signal reaches the end of the line, the voltage value can be integrated as,

$$V_X(h) = e^{-R_0 h / 2Z_0} V_i(0) \tag{3-13}$$

Considering the load capacitance C_L, we assume the initial voltage for first incident switching at modeling point L is zero and the final voltage is the division of two resistors,

$$V_L = \frac{R_L}{R_L + Z_0} 2V_X(h) \tag{3-14}$$

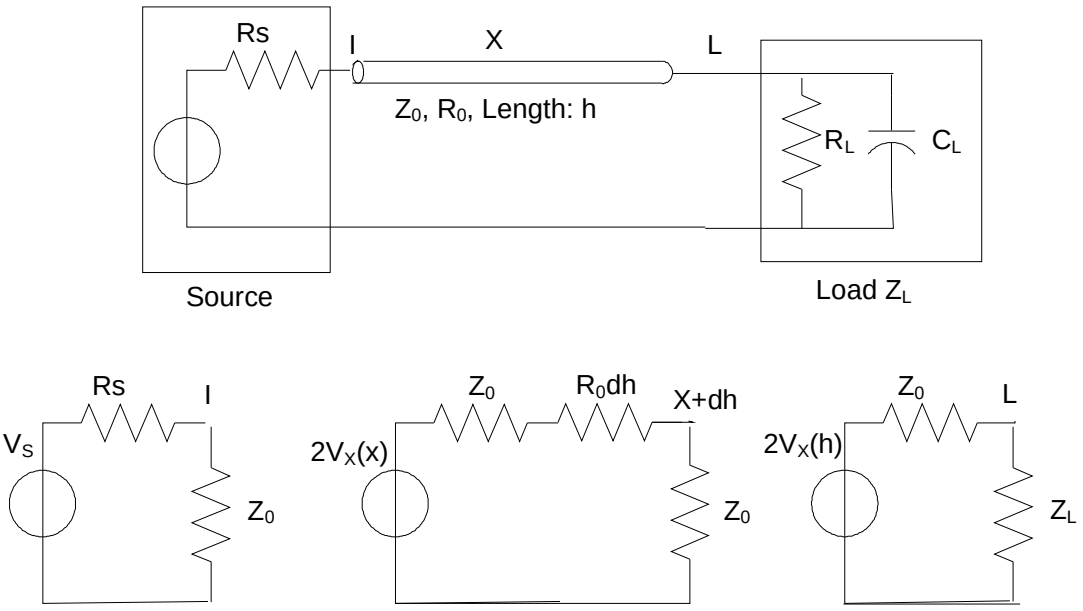


Figure 3.5 Circuit model for a lossy transmission line.

Hence, the signal waveform at the load end for first incident switching can be approximated as,

$$V_{out}(t) = V_L(1 - e^{-t/t}) \tag{3-15}$$

$$t = \frac{Z_0 R_L}{Z_0 + R_L} C_L \tag{3-16}$$

The rise time for the first incident switching waveform is then derived as,

$$t_{FIS} = -\frac{Z_0 R_L}{(Z_0 + R_L)C_L} \ln\left(1 - \frac{vm_0}{\frac{Z_0}{Z_0 + R_S}\frac{R_L}{R_L + Z_0} 2e^{-R_0 h/2Z_0}}\right) \tag{3-17}$$

If $t_f = \sqrt{L_0 C_0 h^2}$ is defined as the line flight time, the final closed-form delay formula is obtained as,

$$t_v = t_f + t_{FIS} \tag{3-18}$$

Figure 3.6(a) compares this first incident switching delay formula with SPICE simulation. The results converge very well at different threshold ν along a wide range of line lengths, while the Sakurai RC delay model [12] deviates largely from the SPICE simulation. Notice that unlike RC lines, the propagation delay of RLC lines does not change rapidly against the threshold. This inductance-induced sharp edge rate is bad for crosstalk and simultaneous switching noise (SSN), but good for reducing skew and short circuit current.

Figure 3.6(b) shows the delay change versus load termination R_L and C_L. It is within 10% accuracy of SPICE simulation for a large range of termination variation. Because R_L is also a variable, this proposed formula can be applied to both voltage and current-mode signaling designs [37].

3.3.4 Maximum length guideline

Due to the slow RC rise time from 90% to 100% for the final voltage settling, it greatly increases the propagation delay if the desired output voltage is designed to be in this region. Therefore, we have

$$V_{out} \le 0.9 V_L \tag{3-19}$$

Combined with (3-12) – (3-14),

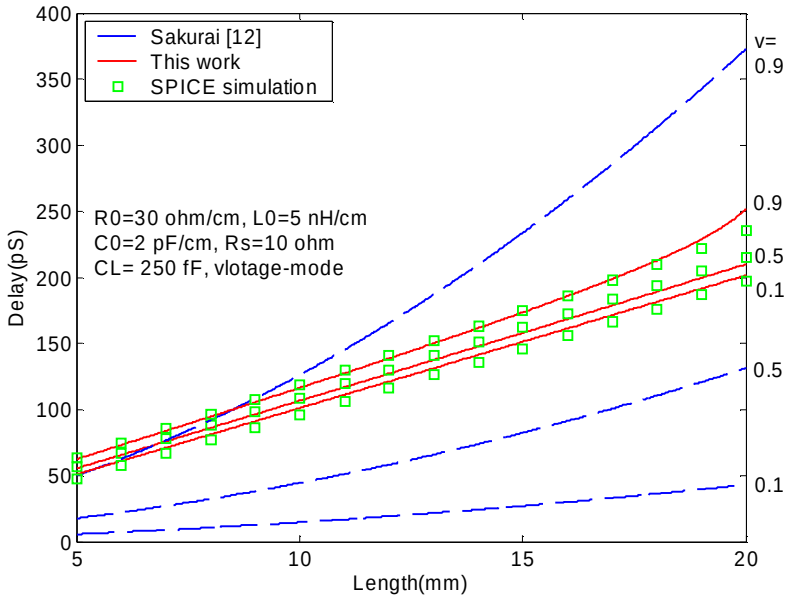
$$h \le \frac{2Z_0}{R_0} \ln \frac{1.8R_L Z_0}{vm_0 (R_L + Z_0)(R_S + Z_0)} \tag{3-20}$$

This equation gives the guideline of the maximum usable length for a lossy transmission line if first incident switching is targeted. It consequently avoids the grey region which cannot be modeled by either RC or lossy transmission line model.

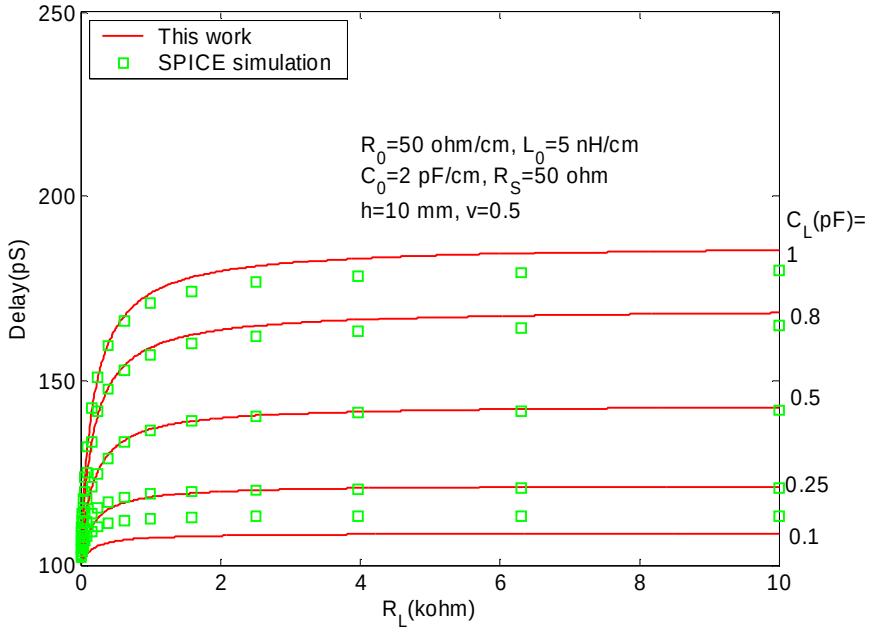
The effective attenuation constant guideline (3-8) and the maximum length guideline (3-20) define the RLC and grey regions in Figure 3.7(a), and the RLC region, grey region, and the RC region in Figure 3.7(b). The proposed first incident switching model (3-18) can be used in the RLC region and the RC model [12] in the RC region for delay analysis. The results indicate convergence with SPICE simulation in both regions.

To verify the guidelines proposed in (3-8) and (3-20), the delay formulas and SPICE simulation are compared in parameter ranges of $5\text{mm} \le h \le 20\text{mm}$, $0\Omega \le R_L \le \infty$, $0 \le C_L \le 1\text{pF}$, $0 \le R_S \le Z_0$, $0 \le R_0 \le 100\Omega$ and accuracy within 10% is achieved. Line parameters can be either extracted by using a field solver simulator or characterized by using the method in Appendix B.

The delay formula in grey region is theoretically derivable by using inverse Laplace transform or convolution method, but the maximum length guideline constrains the failing point of first incident switching in the grey region and the failing point determines if one extra round trip time is needed. Therefore, interconnect design in the grey region is undesirable due to both the delay concern and timing unpredictability concern. By using the simplified delay design guidelines proposed in this work, the grey region can be avoided by either changes in line parameters or termination.

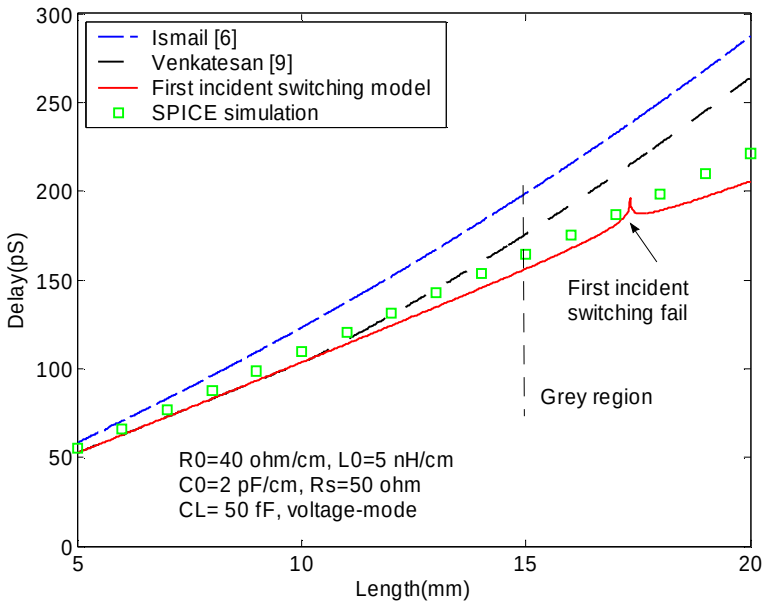


(a)

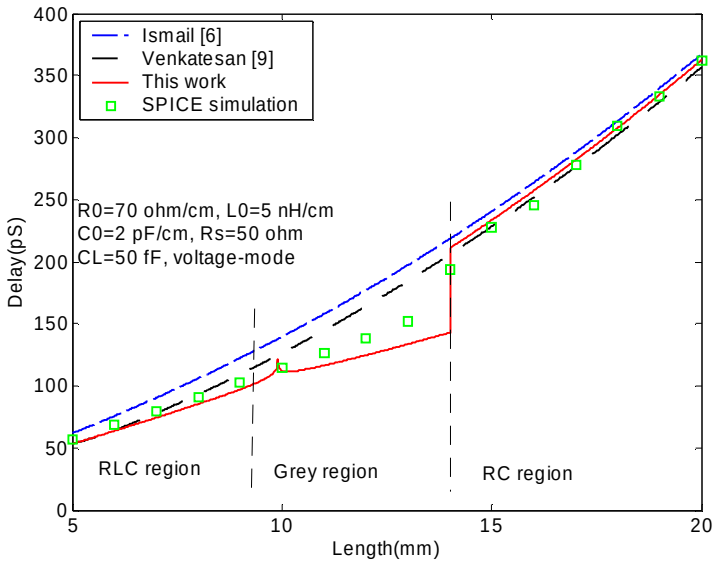


(b)

Figure 3.6 First incident switching delay formula compared to SPICE, (a) delay vs. length at different ν, (b) delay vs. R_L at different C_L.



(a)



(b)

Figure 3.7 Design guidelines compared to SPICE, (a) RLC, grey and RC regions, (b) RLC and grey regions.

3.4 Inductance Model for Differential Signaling

Inductance modeling for a differential pair of buses is analyzed in this section to prepare and justify the application of the driver pre-emphasis technique to differential signaling in chapter 5.

The magnetic interaction model of [40], [41] is used in the inductance analysis of one pair of repeater-free differential interconnects. Figure 3.8 shows current I_a flowing through the interconnect $line_a$. The relationship between the time-derivative of I_a and the induced voltage V_{ind} on $line_b$ is,

$$V_{ind} = L_{ba} \frac{dI_a}{dt} \tag{3-21}$$

Where L_{ba} is the mutual inductance of $line_a$ upon $line_b$. V_{ind} results from the integration of the induced electric field E_{ind} and E_{ind} is created by the time varying magnetic flux Φ,

$$V_{ind} = -\int_b \vec{E}_{ind} \cdot d\vec{x} = \frac{d\Phi}{dt} \tag{3-22}$$

If all the current in $line_a$ is assumed to be condensed to its axis, it generates a magnetic field $B_0 = \frac{m_0}{2pPitch} I_a$ at the center of $line_b$. μ_0 is the permeability of free space. If the magnetic field along the cross section of $line_b$ is approximated as B_0, we have,

$$\Phi = \int_{Area} \vec{B} \cdot d\vec{S} = WidthLength \frac{m_0}{2pPitch} I_a \tag{3-23}$$

Where S is the surface of $line_b$ on XY plane. By combining (3-21) - (3-23),

$$V_{ind} = \frac{WidthLength}{Pitch} \frac{m_0}{2p} \frac{dI_a}{dt} \tag{3-24}$$

From (3-21) and (3-24),

$$L_{ba} = \frac{WidthLength}{Pitch}\frac{m_0}{2p} \tag{3-25}$$

Hence, this simple closed-form calculation can be used for the inductance extraction of differential interconnects.

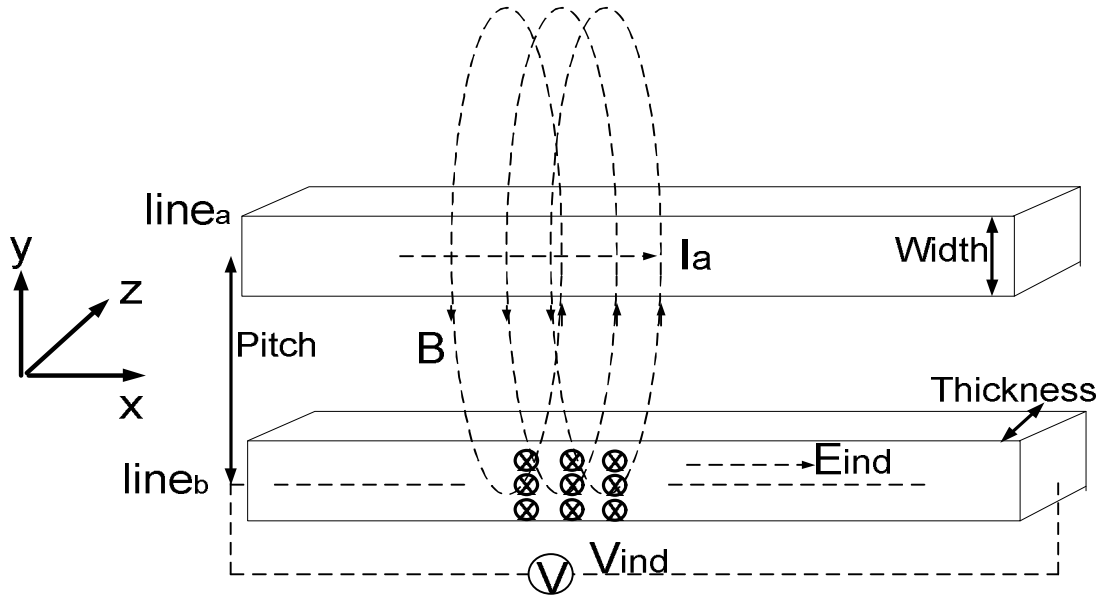


Figure 3.8 Magnetic field created by the time-variant current in $line_a$ induces voltage in $line_b$.

The skin effect and proximity effect are ignored in this analysis because both the skin depth and proximity depth are larger than the line width used in our interconnect scheme. For a 50pS rise time tr, the characteristic frequency can be defined as $f = \frac{0.35}{t_r} = 7GHz$.

For an aluminum conductivity σ=3.8x108$(\Omega m)^{-1}$, skin depth δ is given by [41],

$$d = (pfm_0 s)^{-0.5} = 0.31mm \tag{3-26}$$

Line proximity effect can be modeled as [42],

$$R_{prox}(f) = R_{DC}\left[1 + \frac{1}{2}\left(\frac{u_0 Width^2}{R_{sheet} Pitch} f\right)^2\right] \tag{3-27}$$

We define the proximity depth as the width where proximity effect changes the resistance by 5% of R_{DC}. For R_{sheet}=0.076Ω/square, it is roughly 5μm. Driver pre-emphasis and current-mode signaling allow us to use interconnects as narrow as 0.4μm for signal transmission. It is smaller than both 2δ and the proximity depth. Therefore, the skin effect and proximity effect are ignorable in this case.

3.6 Summary

An effective damping factor was derived as the rule to decide whether an on-chip wire system was underdamped or overdamped, i.e. whether inductance effects were important or not. For aluminum on-chip global buses, RC delay models were proven still valid even if the attenuation constant, or the wire damping factor, indicates inductive behavior. For wide wires in a clock distribution network or copper wires where inductive effects could be verified important, a first-incident-switching delay model was derived to treat the interconnect line as a lossy transmission line. Accuracy within 10% of SPICE simulation was obtained. Inductance models for differential pairs were also derived to verify that differential signaling had higher computation and simulation efficiency with respect to inductance analysis.

Chapter 4.Single-Ended Voltage-Mode Signaling with Pre-emphasis

4.1 Introduction

Low-swing signaling techniques are effective in reducing power consumption by trading off noise-margin and bandwidth [22]. However, bandwidth is most likely the major design goal for an on-chip signaling design and cannot be traded. This chapter explores the possibilities of applying equalization technique to on-chip signaling by trading-off noise margin for both bandwidth and power. The noise levels in the domain of global buses are tightly controlled by circuit techniques and bus structures to compensate for the loss of noise margin.

4.2 Transmitter Equalization (Driver Pre-emphasis)

Equalization techniques improve interconnect channel bandwidth and reduce delay latency by compensating for the high-frequency component loss in a low-pass channel [43]. Figure 4.1 shows a lossy transmission line in an off-chip application. Without any signaling strategy, the signal transmitted through this channel has severe inter-symbol-interference (ISI) at the receiver input for single-pulse-one and single-pulse-zero. An equalizer compensates the channel high-frequency loss by either emphasizing high-frequency signal components or de-emphasizing low-frequency components to transmit an equalized signal to the receiver

input. A flatter channel-frequency-response is achieved by combining the equalizer with the channel. Notice that different vertical scales are used in the figure to illustrate the frequency compensation.

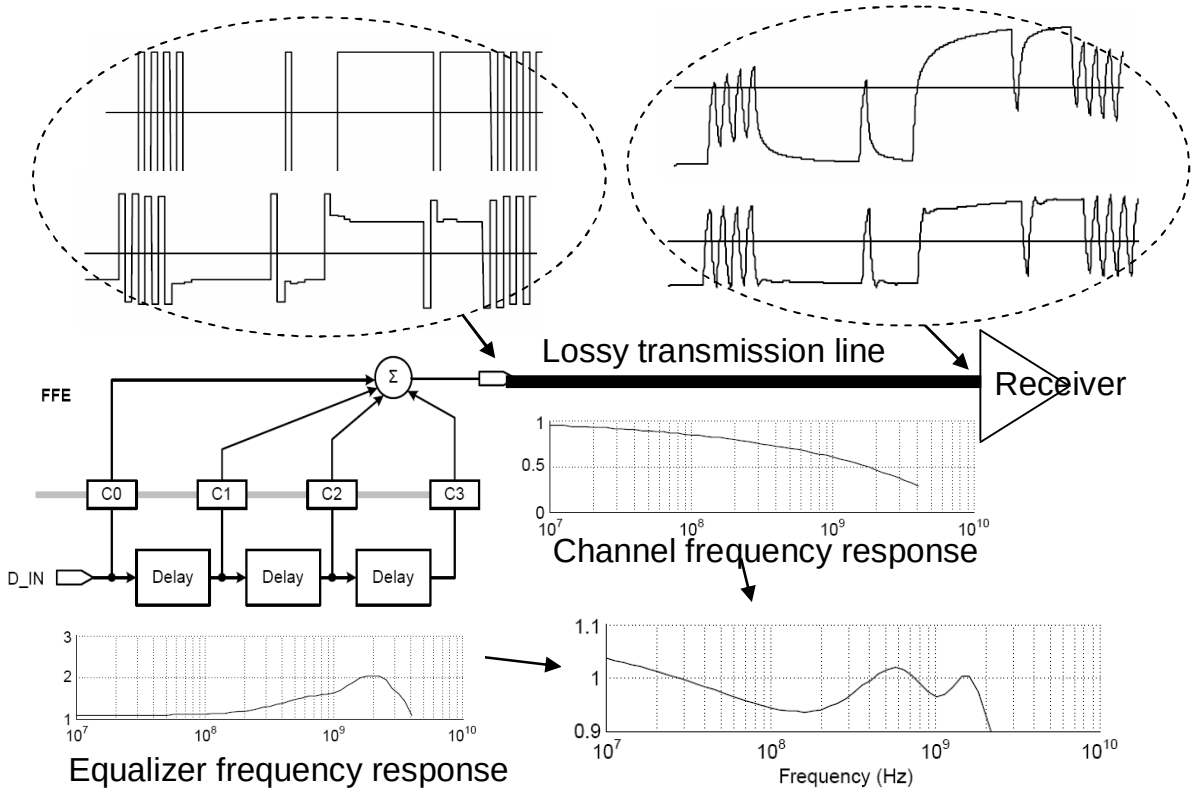


Figure 4.1 Off-chip transmitter equalization.

Equalization techniques were traditionally limited in off-chip applications because on-chip signals were more controllable than off-chip signals. Therefore, it was not necessary to consume the extra power and area overhead and extra delay latency associated with equalization for on-chip applications. As technology scaling keeps producing faster logic but slower global interconnects, a high performance VLSI design is reaching the point where on-

chip signal bit error rate (BER) must be considered. Off-chip communication techniques are required for on-chip communication to accommodate this trend [44].

Two main equalization architectures, Feed Forward Equalization (FFE) and Decision Feedback Equalization (DFE), are shown in Figure 4.2. DFE is only used at the receiver side Equalization at the driver side is easier to implement for non-variant channels like on-chip interconnects, an FFE driver pre-emphasis architecture shown in Figure 4.3 is used in this work. It has only one tap on the equalization path to reduce power overhead and its output can be either singled-ended or differential.

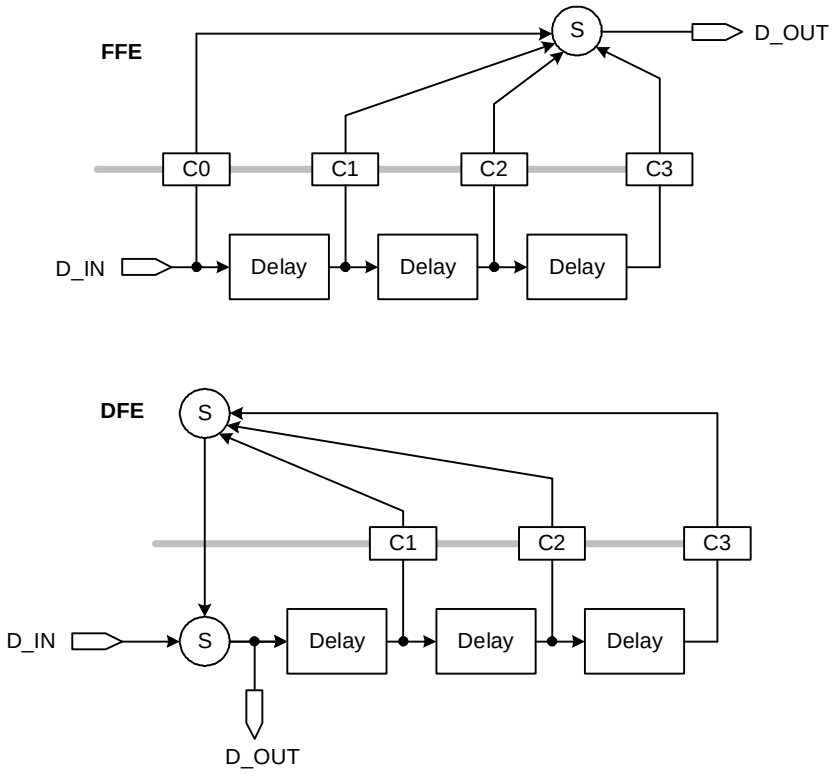


Figure 4.2 Two main equalization architectures.

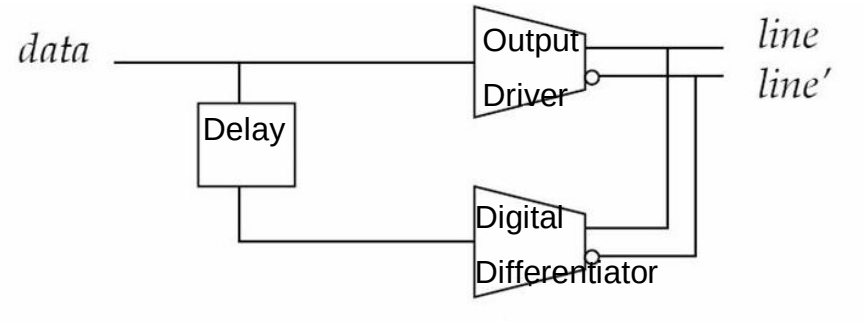


Figure 4.3 Driver architecture with one-tap pre-emphasis.

The metal-4 layer in the TSMC 0.18μm six-level aluminum technology is used to demonstrate this proposed driver pre-emphasis architecture. Based on the interconnect analysis in Chapter 3, aluminum wires in on-chip global bus applications are still RC dominant. To verify that driver pre-emphasis technique can also compensate the high-frequency loss of a RC interconnect channel in addition to a RLC channel, Figure 4.4 shows the frequency responses of a RC interconnect, a pre-emphasis equalizer with co-efficient 0.2, and their combination. The interconnect is modeled as a distributed RC channel with resistance R_0=240Ω/cm, parasitic capacitance C_0=2.5pF/cm, driver resistance R_S=250Ω, and load capacitance C_L=20fF. These parameters are later tested in a chip experiment. Driver pre-emphasis improves the system -3dB frequency from 0.5GHz to 1GHz. The bandwidth of the RC channel is doubled and sufficient for transferring 2Gb/s NRZ data.

In the time domain, MATLAB models based on the Sakurai RC delay formula [12] are used to compare driver pre-emphasis technique with repeater insertion for a wide range of interconnect lengths (Figure 4.5). If delay t_v is defined as the time from (t=0) to the time when the normalized voltage reaches threshold ν at the receiving end of the line, $t_{0.5}$ is used

for repeater insertion and ($t_{0.5}$-$t_{0.2}$) is used for driver pre-emphasis assuming the equalization co-efficient is 0.2. A repeater line is divided by (k-1) uniformly spaced repeaters, which represents a line with k segments. For equally sized drivers and repeaters, the pre-emphasis technique results in lower delay latencies than repeater insertion. The latency is 411ps for a line length of 10mm, a 26% and 19% improvement over lines with one repeater and four repeaters, respectively. Given this, it is again verified that driver pre-emphasis techniques can be used to replace repeater insertion technique to achieve 2Gb/s data rates. In addition, because a driver with pre-emphasis only requires about three times of the area of a traditional driver, it saves active layout area when two or more repeaters are required for a given target data rate.

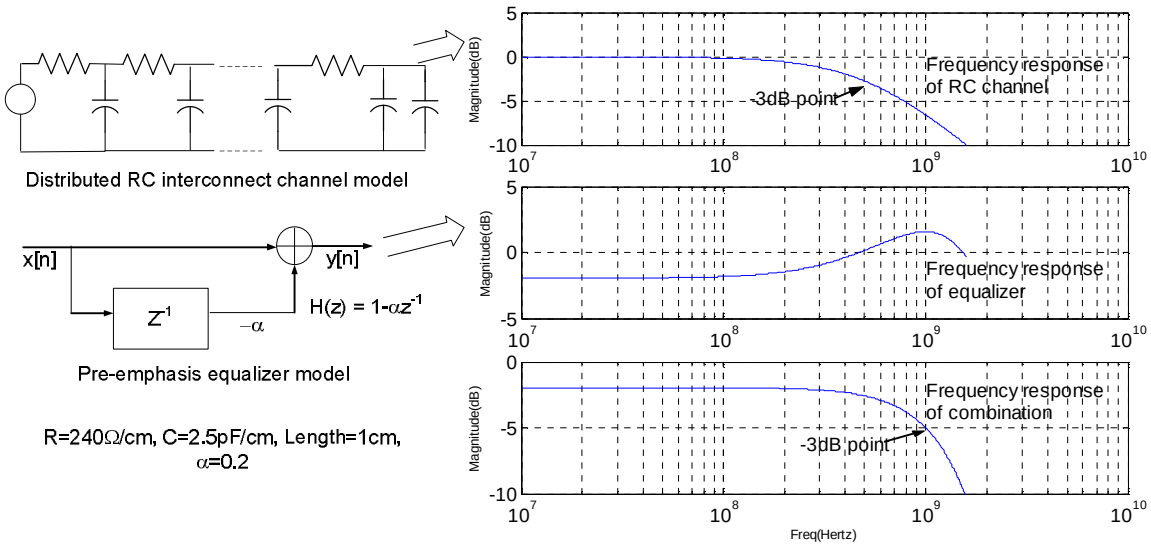


Figure 4.4 Frequency responses of a distributed RC interconnect channel, pre-emphasis equalizer, and their combination.

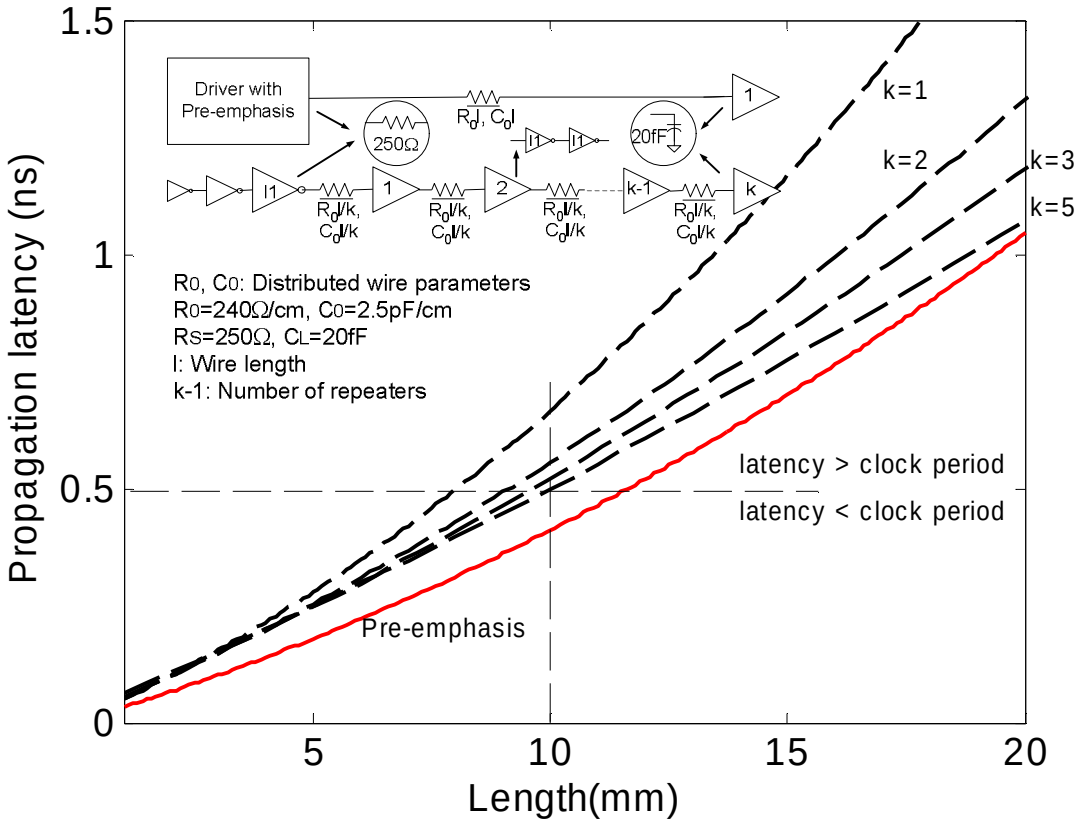


Figure 4.5 Propagation latency comparison of driver pre-emphasis and repeater insertion.

4.3 Demonstration

4.3.1 Circuit implementation

To transmit 2Gb/s signal on the previous analyzed interconnect, a driver pre-emphasis circuit is designed based on the architecture proposed in Figure 4.3. The circuit in Figure 4.6 consists of a one-tap FIR filter and a simple DAC. The FIR filter controls the two tri-state gates in the DAC. Transistor P1 or N1 is only turned on when there is a “0” to “1” or “1” to “0” transition and provides full signal swing at Dout. The transistors are sized to produce a

swing from (V_{dd}-V_{th}) to V_{th} at the receiver input, R_{in}. Transistors N2/N3 or P2/P3 are turned on for consecutive "1"s or "0"s and are only needed to maintain the voltage swing. Therefore, a smaller transistor size, 2.5x minimum, is used. A DFF is used to save every previous-sent bit in the FIR filter. A static DFF is chosen because of its low power consumption. Inverters are used as receivers to amplify the attenuated signal back to full-swing signal. The noise margin loss is later discussed in the performance evaluation section. The extra power overhead of the pre-emphasis driver is mainly from the logic cells and DFF in the FIR filter. It is less than 0.2pJ/bit and is likely to further scale with technology.

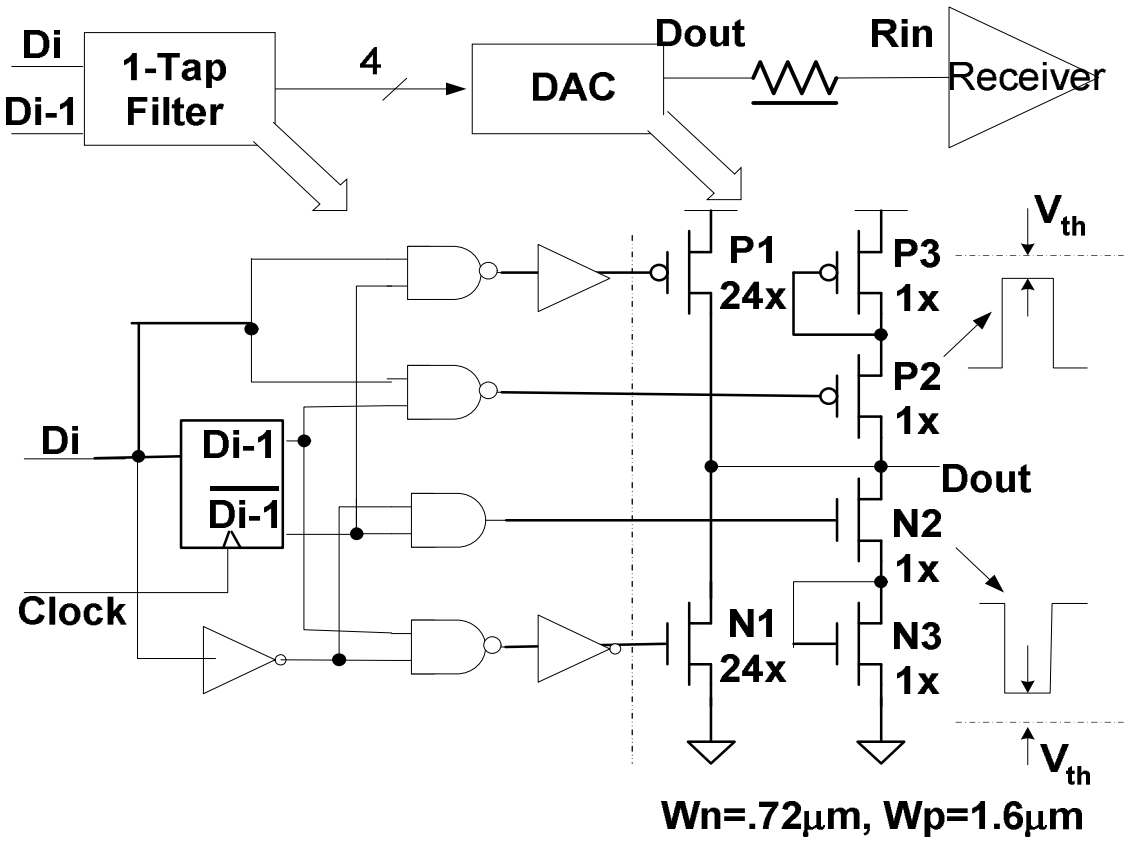


Figure 4.6 Circuit design for driver pre-emphasis.

Figure 4.7 shows the simulation results of the waveforms at the driver output and receiver input. An equalized signal is achieved at the receiver input. The illustrative timing diagram is shown in Figure 4.8. Pre-emphasis is determined by every previous-sent-bit. Data sequence does not need to be pipelined or delayed as in [25] before appearing at the driver output. Therefore, it does not introduce any extra clock-period of latency into the timing. All consecutive "1"s or "0"s are attenuated by one threshold voltage (V_{th}) at Dout and "0" to "1" or "1" to "0" transition are emphasized. The signal swing attenuation reduces power consumption and the overdrive increases signaling speed by providing a larger signal than required at the receiver input [24]. Both the attenuation and overdrive are the by-products of driver pre-emphasis, which de-emphasizes the low frequency component of signal to reduce inter-symbol interference (ISI) and improve bandwidth.

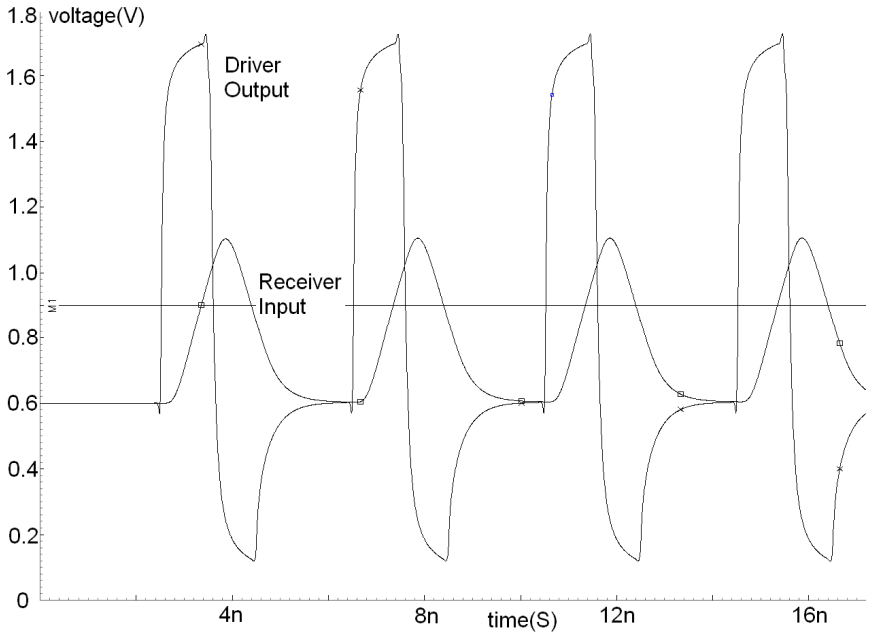


Figure 4.7 Driver output and receiver input waveforms of pre-emphasis bus.

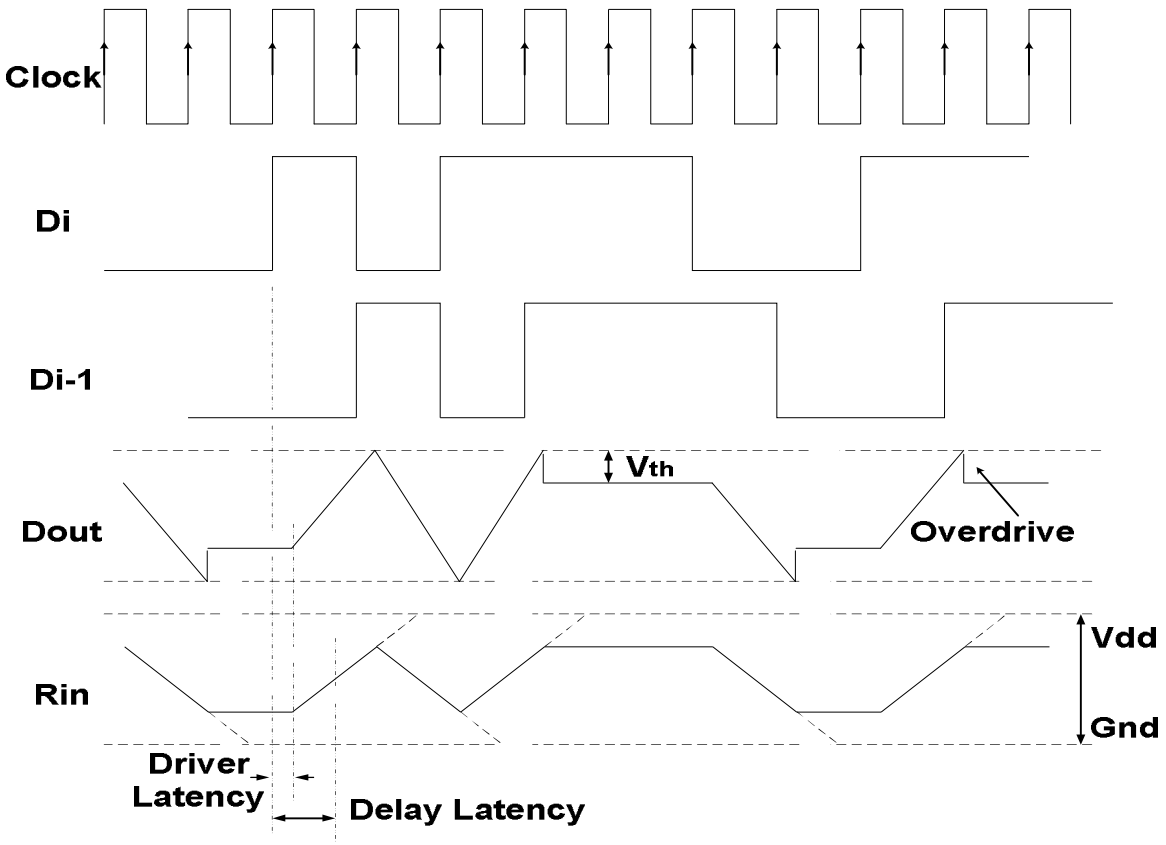


Figure 4.8 Illustrative timing diagram.

4.3.2 Silicon implementation

Figure 4.9 shows the test chip demonstrating the proposed driver pre-emphasis technique in TSMC 0.18µm CMOS technology. Meandered metal-4 lines with length of 10mm and width of 4.5µm are used. Three test cases, driver pre-emphasis interconnect, simple interconnect, and repeater interconnect, are implemented. For fair comparison, all of the drivers and repeaters in the test chip use the same transistor size as P1 and N1 in Figure 4.6.

The measurement setup for probing is shown in Figure 4.10. Because of inductance consideration, bare test chips instead of packaged ones were ordered from MOSIS and were wire-bonded on PCB board. Signals were probed in by using GGB model 40A GSSG probes and probed out by using GGB model 35A high impedance probes.

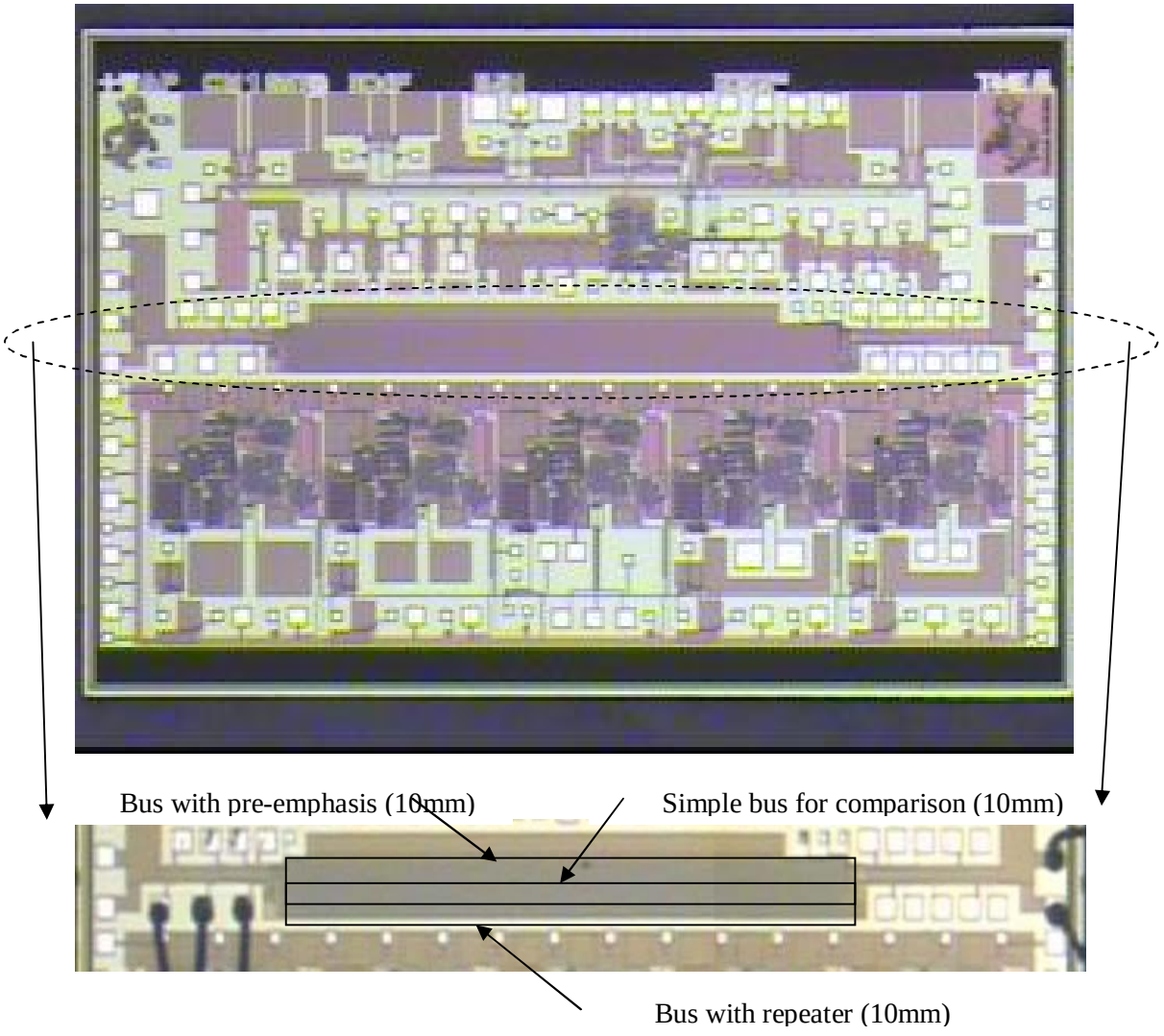


Figure 4.9 Die photograph.

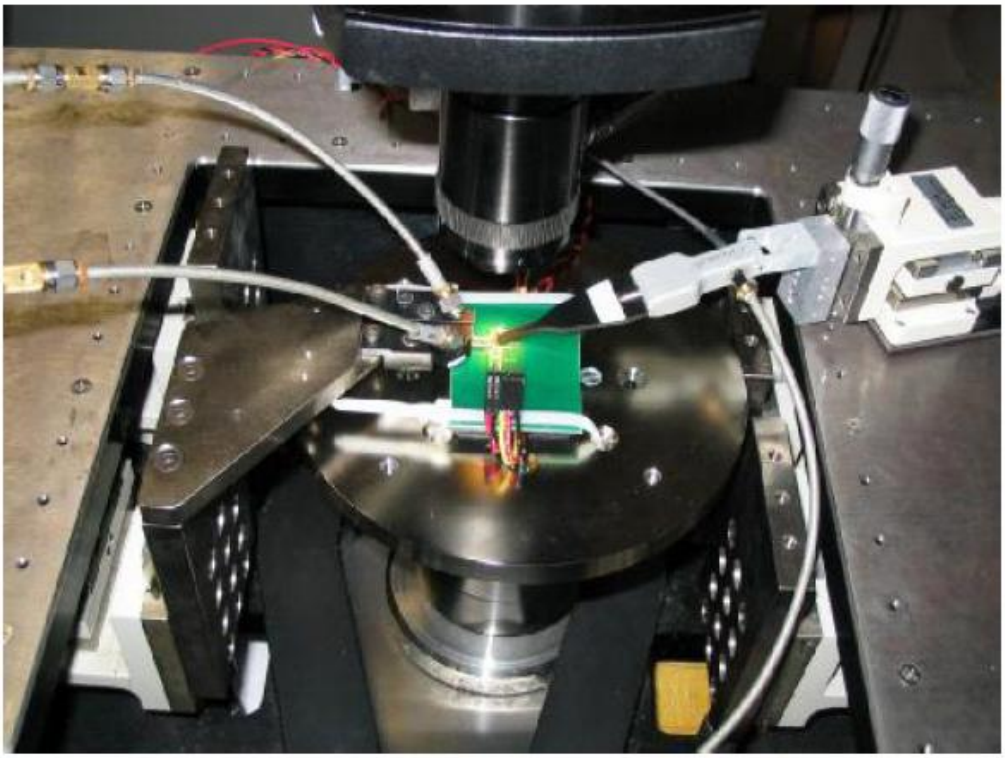


Figure 4.10 Probing measurement setup.

4.3.3 Measurement results

A 127-bit Pseudo-Random Binary Sequence (PRBS) input and data pattern profiles with different signal activity factors are probed in from an Agilent 81134A source. The measurement results at the receiver input are shown in Figure 4.11, simple interconnect with no repeater (left), interconnect with one repeater (middle), and interconnect with driver pre-emphasis (right). A data pattern is used for the waveform measurement to show inter-symbol interference (ISI) and the PRBS input is used for the eye diagram measurement. At 2Gb/s, the simple interconnect has severe ISI, resulting in eye closure. The interconnect containing repeater reduces ISI by boosting the whole signal, while the interconnect using driver pre-

emphasis does this by attenuating the low-frequency signal components. Both approaches increase bandwidth, but driver pre-emphasis saves power. An interconnect delay latency of 420ps is measured and matches the simulation results.

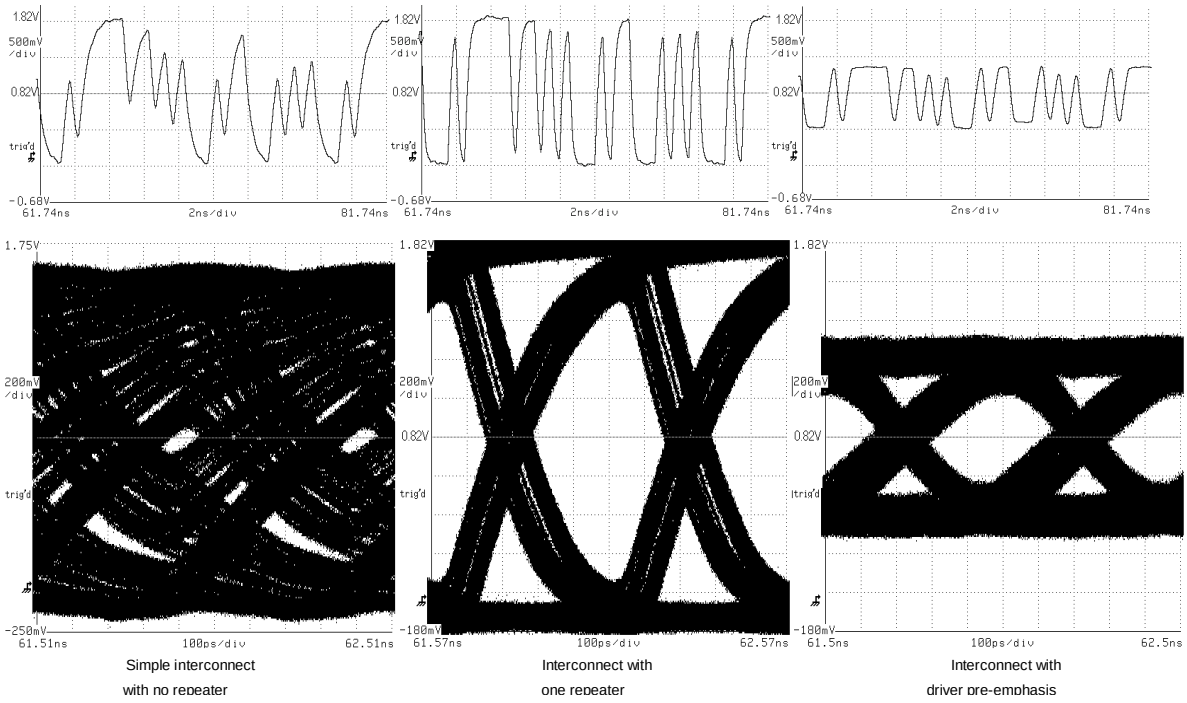


Figure 4.11 Measurement results at the receiver input with a data pattern input and a 127-bit PRBS input.

In Figure 4.11, low-swing signal can be observed on the interconnect with driver pre-emphasis. Unlike the low-swing schemes in [22], which generally sacrifice both noise-margin and bandwidth for power dissipation, this proposed pre-emphasis technique improves bandwidth while only trading off noise-margin due to the reduction in voltage swing. With an eye opening of 400mV, a simple inverter is used as a receiver with negligible increase in

static power. V_{th} variation has an impact on noise margin. The DC points at both the driver output and the receiver input are dependent on V_{th}. If V_{th} variation between the driver and receiver track each other, the DC points will also track and cause no noise margin penalty. Only slow N and fast P at one side and fast N and slow P at the other side degrade noise margin. In this case, sense amplifiers instead of simple inverters are needed as receivers for power optimum voltage swings [23]. As long as there is still signal noise margin to be traded off for power and bandwidth as Vdd and Vth scale, this driver pre-emphasis technology scales.

Figure 4.12 shows the power dissipation measurement for PRBS data at different frequencies. The simple interconnect does not work above 1Gb/s. The pre-emphasis interconnect decreases power consumption by up to 40% when compared to using repeaters.

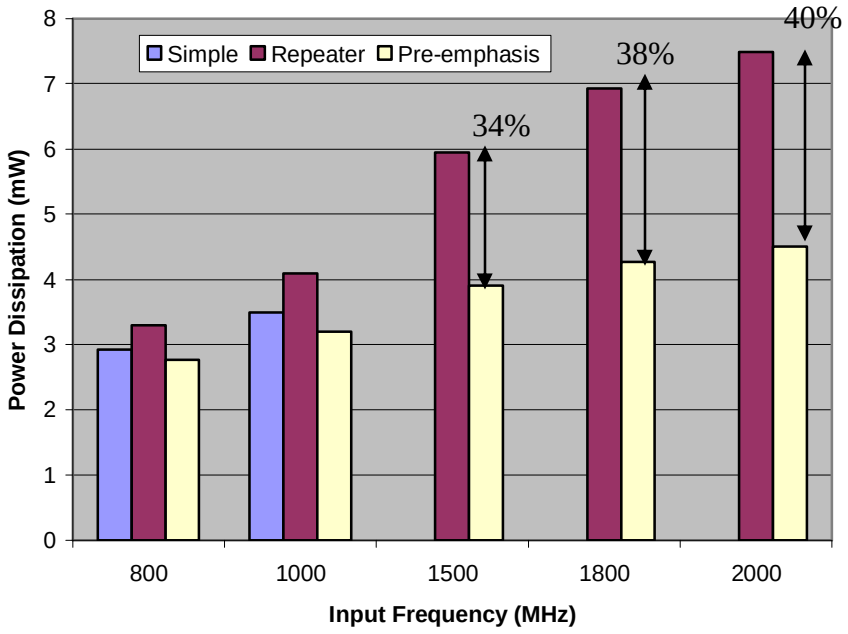


Figure 4.12 Power dissipation measurement at different frequencies with PRBS input.

Figure 4.13 shows the power dissipation measurement for 2Gb/s data patterns with different data activity factors. For activity factors above 0.1, the use of driver pre-emphasis reduces power by 12% to 39% when compared to the traditional repeater insertion technique. With a typical on-chip bus data activity factor of 0.15 [48], the power saving is 25%.

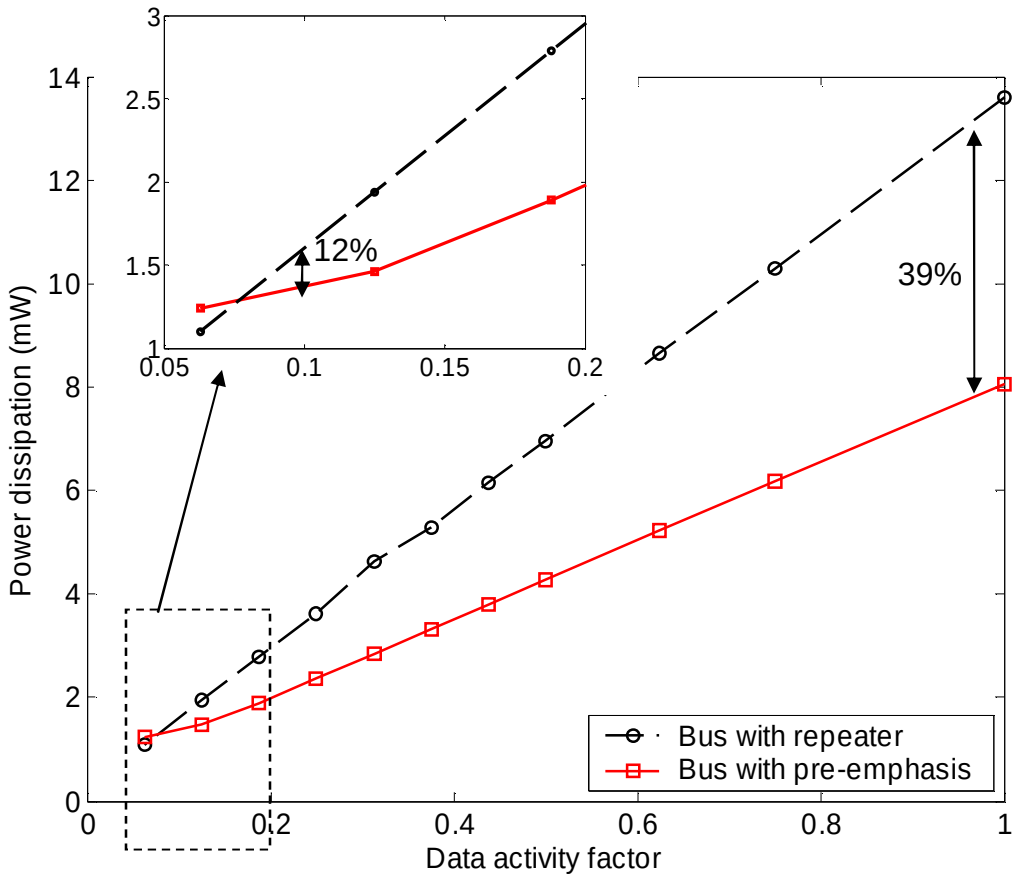


Figure 4.13 Power dissipation measurement at different data activity factors with 2Gb/s data pattern input.

4.4 Performance Evaluation

4.4.1 Bus structures

The previous results are mainly power performance comparison of driver pre-emphasis technique and repeater insertion technique based on the same target date rate and bus routing area. The section explores a thorough comparison of the performance metrics in on-chip signaling design: delay, noise, power, and area.

Figure 4.14(a) illustrates a 16-bit repeater bus structure with the same wire width and spacing. One Vdd/Gnd line is inserted for every 4-bit signal lines to provide a current return path. The distributed RC model of 10mm long and 0.45μm wide metal-4 lines in TSMC 0.18μm technology can be extracted as R_0=1.73kΩ/cm and C_0=3.48pF/cm [45]. In Appendix A, Lagrange relaxation method is performed to find that the optimal segment number k=5 and the optimal repeater size $W=36xW_{min}$ for most power-optimal repeater insertion for a wire target delay of 1ns.

Figure 4.14(b) shows a 16-bit pre-emphasis bus structure using the same wire pitch as the repeater bus in Figure 4.14(a). For the pre-emphasis driver proposed in Figure 4.6, the driving ability of the tri-state gate to build up the voltage swing on interconnects is 24 times stronger than the tri-state gate to maintain the swing. Therefore, it causes severe crosstalk problem when the neighboring line of a quiet line switches. To quantify the worst-case area penalty for pre-emphasis bus to achieve the same or better noise performance as full-swing repeater bus, each signal line of the 16-bit pre-emphasis bus is shielded by one Gnd line with four distributed connections to top power metals. Therefore, if bus routing area is the major concern, this pre-emphasis bus does take 57.1% more area when compared to the repeater

bus at the same wire pitch of 0.9µm, but it has 16.6% less delay latency, 34.3% less power consumption, and ignorable crosstalk and data-dependant delay variation (Table 4.1).

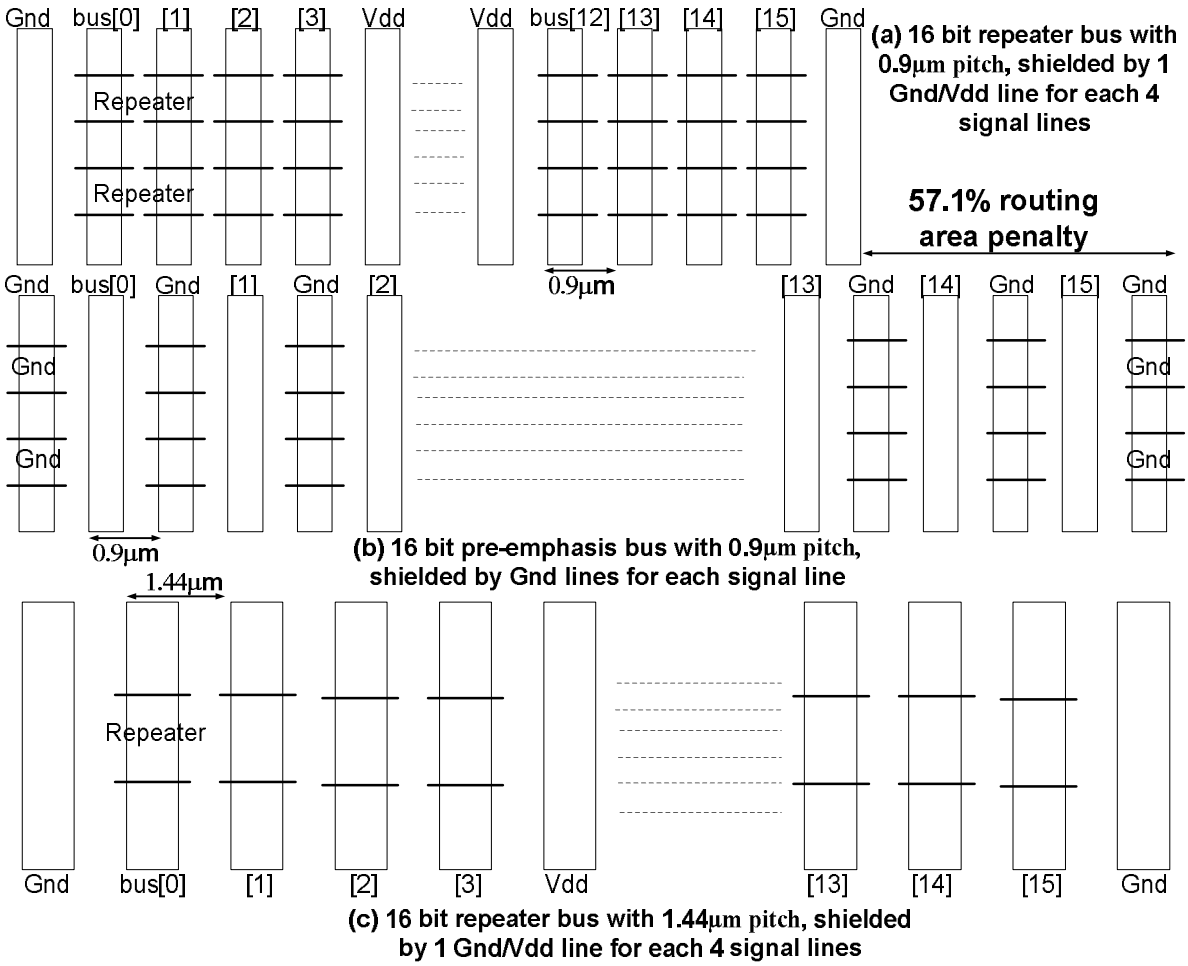


Figure 4.14 16-bit repeater and pre-emphasis bus structures, meanders and dummy underlying metal layers not shown.

Figure 4.14(c) shows a 16-bit repeater bus structures using the same wire routing area as the pre-emphasis bus in Figure 4.14(b). With 1.44µm pitch, it is extracted that R0=1.08kΩ/cm and C0=2.78pF/cm ([45]). Appendix A derives that the optimal segment

number k=3 and the optimal repeater size W=20xWmin for most power-optimal repeater insertion for a wire target delay of 1ns. Table 4.1 shows that at the same routing area this repeater bus consumes 22.3% more power and has significantly more crosstalk and data-dependant delay variation than the pre-emphasis bus. In addition, the pre-emphasis bus saves all of the layout blockage and active area of repeaters.

Table 4.1 Performance comparison.

		Repeater Bus with 0.9µm Pitch		Repeater Bus with 1.44µm Pitch		Pre-emphasis Bus with 0.9µm Pitch	
			Δ		Δ		Δ
Delay	-+- (worst)	1.25	+26%	1.03	+27%	.829	+0.4%
(ns)	0+0	0.99	_	0.81	_	.826	_
	+++ (best)	0.50	-49%	0.45	-44%	.823	-0.4%
Crosstalk (mV)		550		360		43	
Power (mW) (act=0.15)		1.02		0.82		0.67	
Width of routing area (µm)		18.9		30.2		29.7	
R(kΩ/cm)		1.73		1.08		1.73	
Cbottom(pF/cm)		.388		.608		.388	
Ccoup(pF/cm)		.774		.543		.774	

The power consumption in Table 4.1 is compared at a data activity factor of 0.15 [48]. A thorough bus data activity analysis based on a typical microprocessor is later shown in Chapter 5. If the data-dependent delay on an interconnect line is defined as the delay dependence of the data patterns on neighboring lines, the highest data-dependent delay happens when the two neighboring lines switch in the opposite direction (-+-), the lowest data-dependent delay happens when the two neighboring lines switch in the same direction (+++), and the typical data-dependent delay happens when the two neighboring lines are quiet (0+0). The repeater buses have 26-49% of data-dependent delay variation while the pre-emphasis bus has negligible variation and its intra-bus crosstalk noise is only one tenth of the noise on the repeater buses.

4.4.2 Full-swing to low-swing crosstalk

Although intra-bus crosstalk is the most important noise source for on-chip bus design, just like in any low-swing bus design, it is still important to analyze the crosstalk on the proposed pre-emphasis bus from full-swing signals. Because a global bus structure defines a confined domain in one metal layer, crosstalk from the full-swing signals in the same layer is tightly controlled by shielding. Figure 4.15 shows a 16-bit full-swing bus orthogonally crossing beneath the 16-bit pre-emphasis bus at receiver side. The noise is negligible due to the small coupling capacitance of 1fF between the two different layers. Similar measurement result about full-swing to low-swing crosstalk can be found in Chapter 5.

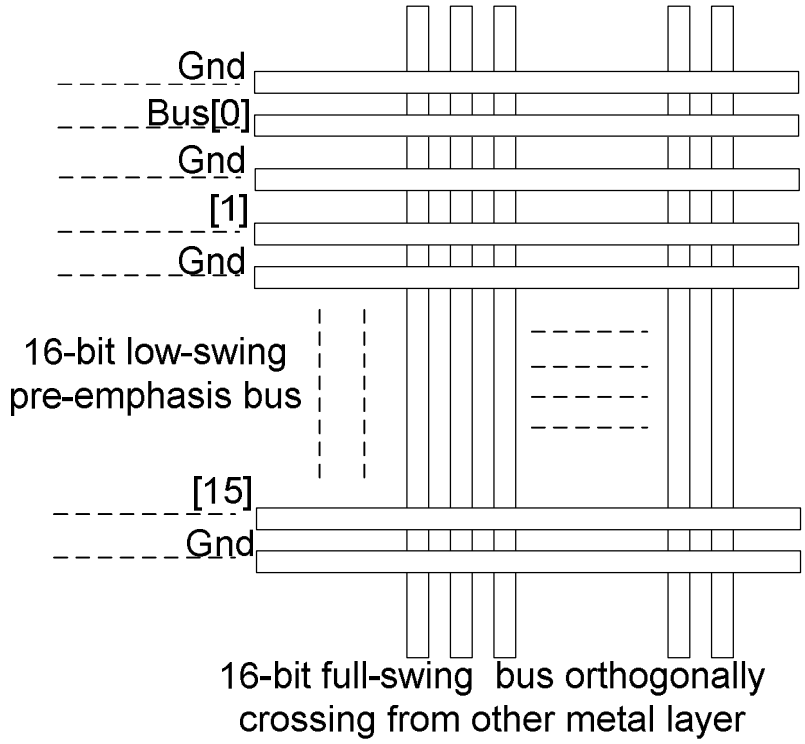


Figure 4.15 Analysis of crosstalk from full-swing signal.

4.4.3 Improved driver pre-emphasis circuit

Figure 4.16 shows an improved driver pre-emphasis circuit. The attenuated voltage reference, V_{pre} and G_{pre}, can be provided by either a series-resistor structure or diode-connected transistors. Diode-connected transistors do not require any extra static power consumption, but they have noise margin disadvantage due to V_{th} variation. The static power consumed on series resistors can be shared by the 16-bit bus. It is only 0.02pJ/bit at 1Gb/s data rate and is negligible. As technology scales, multiple-Vdd and multiple-Vth algorithms have been investigated extensively to reduce power without drastically degrading circuit

performance or increasing leakage current. Driver pre-emphasis technique can be invested to take advantage of these algorithms.

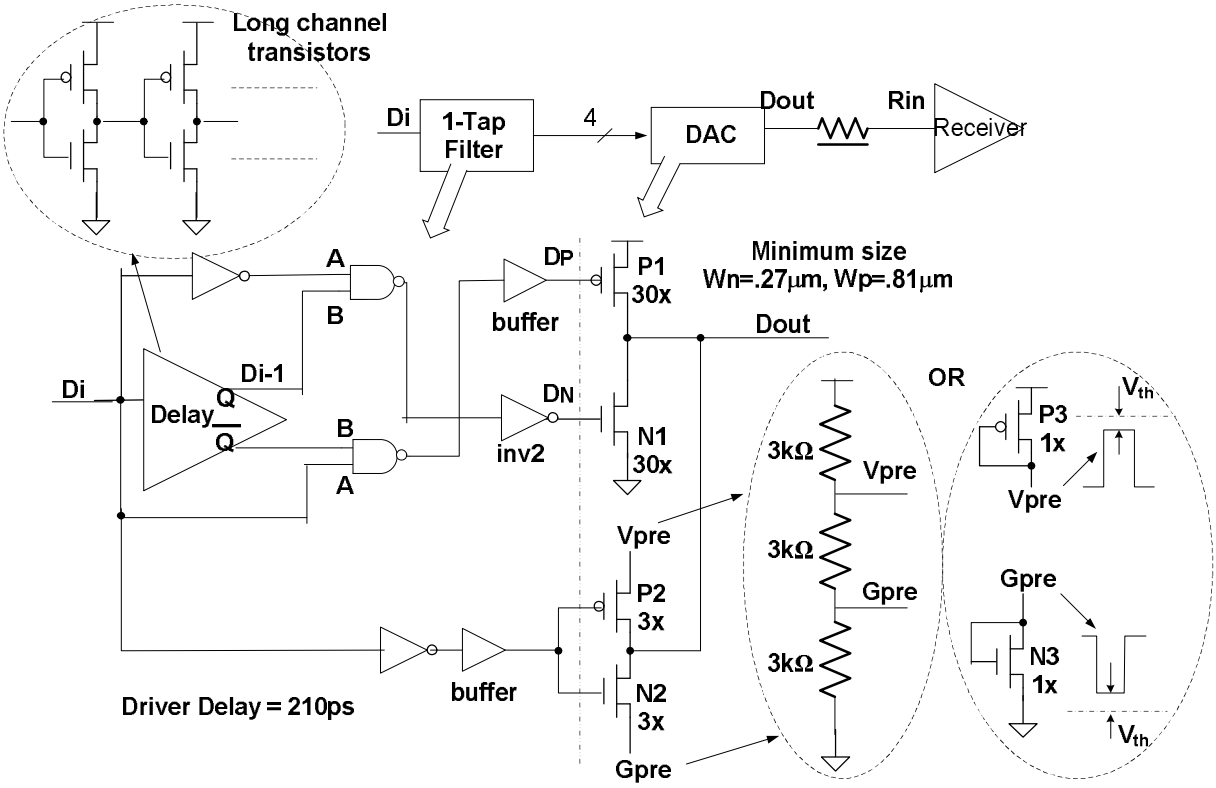


Figure 4.16 Improved driver pre-emphasis circuit.

A delay cell with long channel transistors is used to replace the DFF to detect a “0” to “1” or “1” to “0” transition and compensate transistor process variation. At slow process corners, the delay is larger and the signal pulses at nodes D_P and D_N are wider. This produces more pre-emphasis on driver output signal to compensate the slow driver. While at fast corners, the output signal needs less pre-emphasis and the D_P/D_N pulses are narrower.

4.5 Summary

A single-ended voltage-mode driver pre-emphasis architecture for on-chip global bus was used to minimize the number of repeaters required to meet the goal of signal latency and throughput. For 10mm metal-4 interconnects at 2Gb/s in TSMC 0.18μm technology, it had no extra clock latency and obtains 12%-39% power saving. A thorough comparison between driver pre-emphasis technique and repeater insertion technique was also developed based on the on-chip signaling design metrics.

Chapter 5.Current-Mode Differential Bus with Pre-emphasis

5.1 Introduction

As discussed in the previous chapter, the voltage-mode (VM) driver pre-emphasis technique requires a boosted voltage reference to emphasize high-frequency signal component or an attenuated voltage reference to de-emphasize low-frequency component. The weak driving abilities of these references make the pre-emphasis bus vulnerable to crosstalk noise. Current-mode (CM) signaling is easier to implement with driver pre-emphasis technique because its logic states are determined by current values instead of voltage levels.

The Digital Alpha 21264 design demonstrates that low-swing differential signaling is feasible for on-chip global communication [38]. Low-swing differential signaling creates less noise and is more immune to inductive noise than its single-ended full-swing counterpart [39]. This chapter explores the possibilities of applying driver pre-emphasis techniques to a current-mode differential bus.

5.2 Architecture Implementation

5.2.1 Current-mode sensing

Current-mode signaling can be used to provide higher interconnect bandwidth when compared to traditional full-swing voltage-mode signaling, at the expense of increased DC power dissipation [29]. For the current-sensing circuit architecture shown in Figure 5.1(a), a static current path always exists between the driver and receiver stages even if there is no data activity on the interconnect.

To compensate for this static current, we propose to use a pair of differential interconnects with a bridge resistor termination R_B (Figure 5.1(b)). The static current is reduced by at least 50% due to the resistance increase on the current path. Because a virtual ground is set up in the middle of R_B with a voltage of $V_{dd}/2$, the system RC time constant is the same as that of a single line system.

This architecture requires less CM static current and has all the advantages of differential signaling. The current return path is well-defined in a differential structure and it reduces the impact of inductive effects [51]. Besides, the combination of driver pre-emphasis, current-mode sensing, and differential signaling increase interconnect channel bandwidth and allow for narrow and resistive interconnects. It therefore dominates inductive effects. The area concern is discussed later in section 5.2.3. It shows that for a 16-bit bus this technique takes only 7.9% more bus routing area than the single-ended bus and requires none of the repeater area.

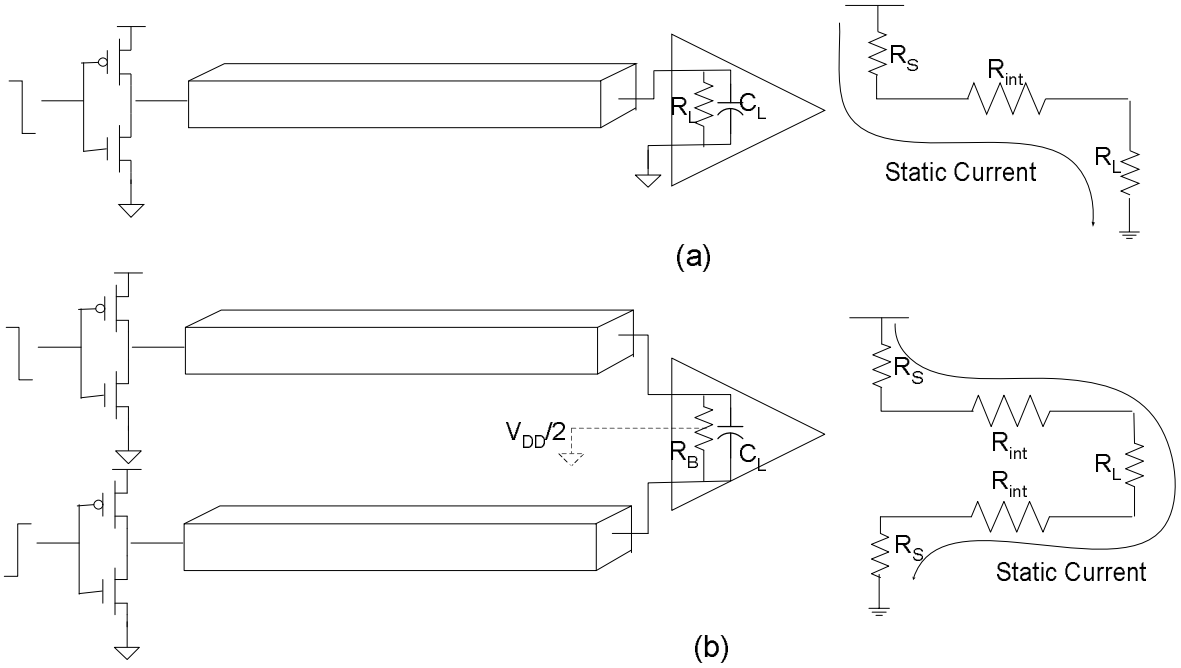


Figure 5.1 CM static current for (a) single-ended bus and (b) differential bus with bridge resistor termination.

5.2.2 Circuit design

Figure 5.2 shows the circuit for driver pre-emphasis CM differential bus. It follows the same design procedure for the discussed pre-emphasis VM single-ended signaling. The interconnects are first analyzed and their parameters are extracted. The signal swing on the bridge resistor is limited by the sensibility of the sense-amplifier receiver. For input offset tolerance, a 100mV signal swing (200mV differential) is established in the design. The driver size is decided by the targeted data rate the static current overhead, 2Gb/s and 1.6mA in this case.

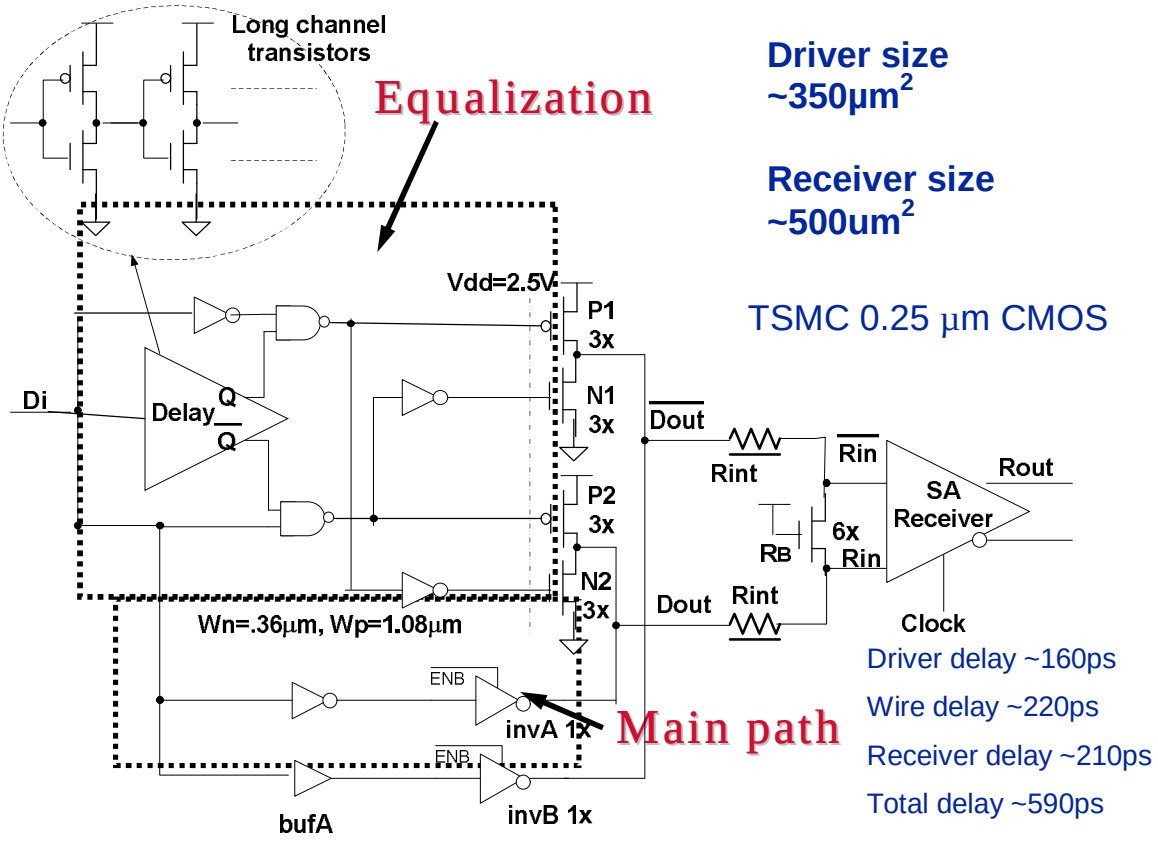


Figure 5.2 Circuit for driver pre-emphasis CM differential bus.

The proposed circuit has an equalization path and a main path. The equalization path has a single-ended to differential conversion circuit, a one-tap FIR filter, and a simple DAC. The main path uses minimum-size inverters for “invA” and “invB” to reduce static current and maintain the 100mV signal swing at the receiver input for consecutive “1”s or “0”s.

Transistors P1/N1 and P2/N2 form two tri-state gates and are only turned on when there is a “0-1” or “1-0” transition. They are only 3x minimum size transistors. The benefits of the small drivers are small peak current and thereby reduce power supply noise. The peak current reduction is illustrated and discussed in Figure 5.19.

Buffers “bufA” and “bufB” are placed to compensate for the data skew between their following inverter drivers and the tri-state gates. The data sequence does not need to be pipelined or delayed as in [54] before appearing at the bus input. Pre-emphasis is determined by every previous sent bit. Therefore, it does not introduce any extra clock-period of latency into the timing. At the receiver side, an nmos transistor is used as the resistive termination. A flip-flop sense-amplifier amplifies the 100mv signal swing and converts it to a full-swing single-ended output. Longer channel transistors and dummy layout cells are used in the receiver to compensate for input offset voltage.

Long channel transistors are also used in the delay cell to detect a “0” to “1” or “1” to “0” transition. For the case of slow process corners, longer delays result in additional pre-emphasis on the driver output to compensate for process variation, however, with fast corners, the output requires less pre-emphasis and the overall delay is shorter. Fast N slow P or slow N fast P corner could induce a 120mV common-mode bias change at the receiver side, which is small compared to the Vdd/2 bias at receiver inputs Rin and Rin_. The transistor, R_B, is always kept in the linear region by the Vdd/2 bias and has a large (Vgs-Vth) value. Its resistance deviation caused by Vth variation is smaller than 8%, indicating a ±8mV change for a 100mV signal swing. Because each differential pair lines are evenly routed, it is reasonable to assume the ±20% metal variation deviates in the same direction. Therefore, this variation does not change the common-mode bias at receiver side, but only changes the total resistance of the current path by less than ±1.5%.

A flip-flop sense-amplifier is used for the receiver in Figure 5.3 [55]. Unlike a PMOS input receiver in [24], the Vdd/2 bias at the receiver input in this circuit allows an NMOS

selection, which operates faster and has smaller latency. Besides, the Vdd/2 bias helps build a large (Vgs-Vth) value on both inputs of the differential sense amplifier to make it less sensitive to transistor mismatch. Special considerations in layout, large input transistors and dummy cells, are ued to compensate input offset. The driver size is around 350μm^2 and the receiver size is around 500μm^2.

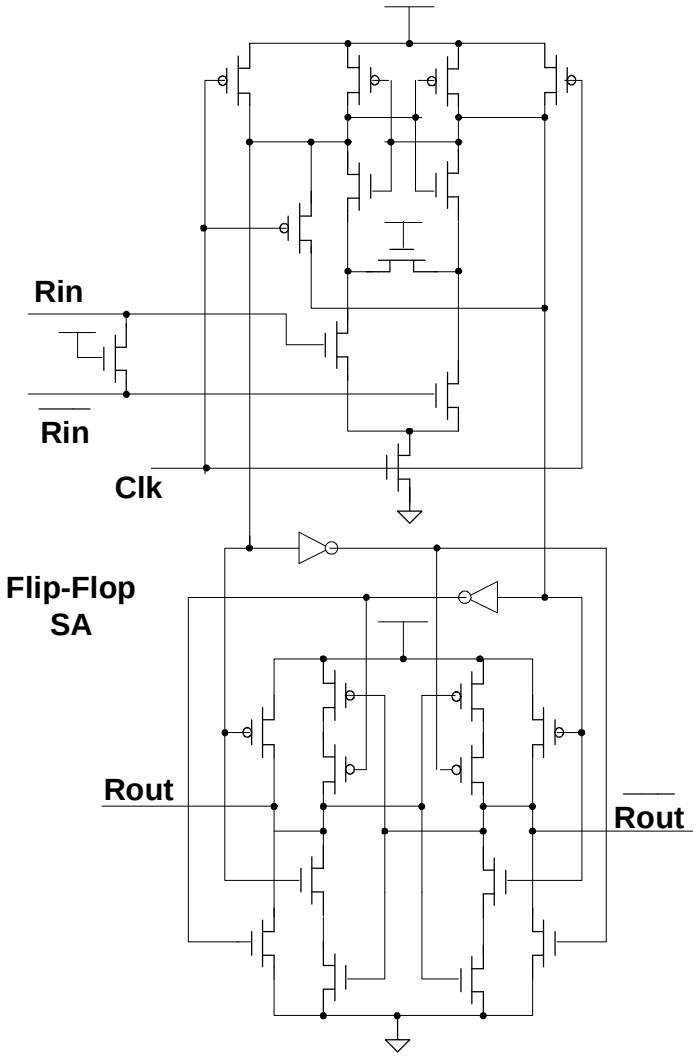


Figure 5.3 Flip-flop sense-amplifier.

The flip-flop sense-amplifier does add the equivalent load of five minimum size inverters to the clock distribution network, but it saves one latch when compared to a simple inverter receiver. Notice that the clock load of a typical static DFF is already equal to four minimum size inverters and a dynamic flip-flop with more clock load has to be used at the high speed, as in our design cases.

5.2.3 Bus structures

Bus layout for the 16-bit differential and single-ended buses are shown in Figure 5.4. Metal-4 with 0.8μm pitch-minimum (Pmin) in TSMC 0.25μm technology is used for signal lines. Every differential pair is drawn at minimum pitch with 0.4μm width and 0.4μm spacing. The pairs have a spacing of 2μm and therefore a pitch of 3.2μm, or 2xPmin per line. The lines are 10mm long with three meanders. Dummy layers of underlying metal-3 to metal-1 with 50% coverage are used to emulate a realistic chip environment. For clarity, neither the meanders nor the dummy layers are shown in this figure. One ground line at each side of the 16-bit bus is used to shield the low-swing signal. To run the single-ended full-swing bus at the same speed wider wires with 3xPmin are used and one Vdd/Gnd shielding line is inserted for each 4-bit bus to provide signal return path. Because each differential pair is driven by a pair of 3x tri-state gates and 1x inverters, an 8x driver is used for each bit of single-ended bus for fair comparison. Two repeaters, with equally sized drivers, need to be inserted into each 10mm long line. The proposed differential bus uses only 7.9% more bus routing area than the single-ended bus and it requires none of the active area needed for repeaters.

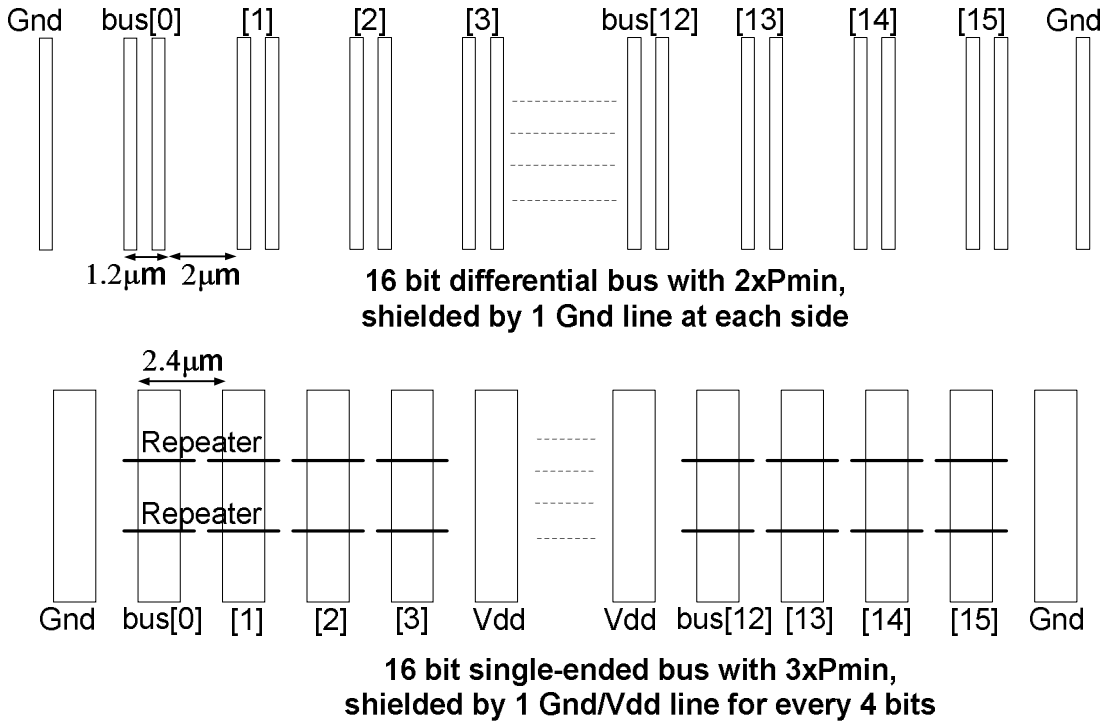


Figure 5.4 Differential and single-ended 16-bit bus structures, meanders and dummy underlying metal layers not shown.

In the reference, the 3xPmin buses with two 16x repeaters are not optimized for power [18], but in this test case the total repeater capacitance is only 5% of the total line capacitance. Additional power optimization will not yield significant power improvement to challenge the validity of the power comparison results. A 2xPmin or 1xPmin buses can be used to save the routing area for the reference but that requires many more repeaters to meet the delay goal. Moreover, a smaller pitch can also be used in the proposed differential bus architecture by inserting one or two repeater with pre-emphasis. The proposed architecture always requires less repeaters than the reference. The purpose of this work is to compare delay, power and noise performance based on similar bus routing area.

METAL™ from OEA [45] is used to extract the parasitic interconnect capacitance (Table 5.1). For the differential bus, the total capacitance per line is,

$$C_{tot} = C_a + C_{f1} + C_{f2} + 2 \times C_{diff} + CCM \times C_c \quad (5\text{-}1)$$

Where Ca=.145pF/cm is the area capacitance to bottom layers, C_{f1}=.270pF/cm and C_{f2}=.094pF/cm are the two fringe capacitances, C_{diff}=.806pF/cm is the coupling capacitance between one differential pair, (the multiplier of C_{diff} is fixed at 2 for differential lines so that C_{diff} is not counted as coupling capacitance,) Cc=.179pF/cm is the coupling capacitance from the neighbor differential pair lines, and CCM is the coupling capacitance multiplier factor, (CCM is 0 for transitions in the same direction, 1 when there is no transition, and 2 for transitions in opposite directions,). The coupling capacitance to total capacitance (Cc/Ctot) ratios are 7.8% and 14.4% for CCM=1 and 2, respectively. This is a significant improvement from a coupling capacitance ratio of 50% in deep sub-micro technologies [17] and allows for more noise rejection and less data-dependent delay.

The Cc/Ctot reduction is the result of both the low-swing differential signaling [15] and the width/spacing configuration used in this work. If a similar configuration is used for the VM single-ended bus in the reference to achieve the same Cc/Ctot ratio, the reference will require significantly more repeaters and be non-competitive in delay and power. In addition, smaller spacing can be used in the proposed architecture to save more bus routing area with a reasonable increase in total capacitance and noise.

For single-ended bus, the total capacitance per line is,

$$C_{tot} = C_a + 2 \times C_f + CCM \times C_c \quad (5\text{-}2)$$

Where Ca=.435pF/cm, C_f=.283pF/cm, Cc=.393pF/cm, and CCM is 0, 1, 2, 3 or 4 because the two neighboring lines can transition in any direction. The worst case of coupling capacitance to total capacitance ratio is 61.2%, a huge degradation.

Table 5.1 Parasitic capacitance for one interconnect line.

	Differential	**Single-ended**
Ca (pF/cm)	.145	.435
C_{f1} (pF/cm)	.270	.283
C_{f2} (pF/cm)	.094	.283
C_{diff} (pF/cm)	.806	/
Worst CCM	2	4
Cc (pF/cm)	.179	.393
C_{tot} (pF/cm)	2.48	2.57
Coupling ratio	14.4%	61.2%

Figure 5.5 shows the signal waveforms at the receiver input for the CM differential bus with driver pre-emphasis. All consecutive “1”s and “0”s are equalized by the pre-emphasis and a 200mV differential signal swing is achieved. Crosstalk is shown by transitioning the two neighbor pairs in various directions. The waveform is clean when the two neighboring lines are quiet (top).

Due to the 14.4% of coupling capacitance to total capacitance ratio, this bus structure has very good differential mode noise rejection. When the two neighboring lines switch (middle

two), the crosstalk on the differential signal swing is controlled under 20% of total swing. 80mV common mode noise is observed on the bottom waveform while the two neighboring pairs couple the differential lines to the same direction. From 1V – 1.5V the common mode rejection ratio (CMRR) of the differential sense amplifier is 50 and is able to reject this 80mV noise.

5.2.3 Alternative bus structures

Alternative bus architectures can be used to save even more bus routing area (Figure 5.6). The main impact of reducing dimensions is the intra-bus cross-coupling increase, which could become very high when two lines run in parallel for a long distance. Extra ground shielding lines inserted between every differential pair reduce the pair pitch from 3.2µm from 2.4µm. For a 16-bit bus, extra shielding saves 25% of area at the cost of 11% total line capacitance and 6.3% coupling capacitance ratio increase. To prevent long on-chip interconnects from behaving like floating lines, shielding lines need to be interrupted for better grounding.

Another more aggressive architecture is to use a transposed structure. Theoretically, the coupling from a neighboring differential pair can be cancelled with enough twists. Hence, the minimum spacing can be used between pairs and 50% of bus routing area can be saved. Just like the previous extra shielding architecture, this area saving is also at the cost of 11% total line capacitance increase. In addition, it adds more layout complexity and via resistance.

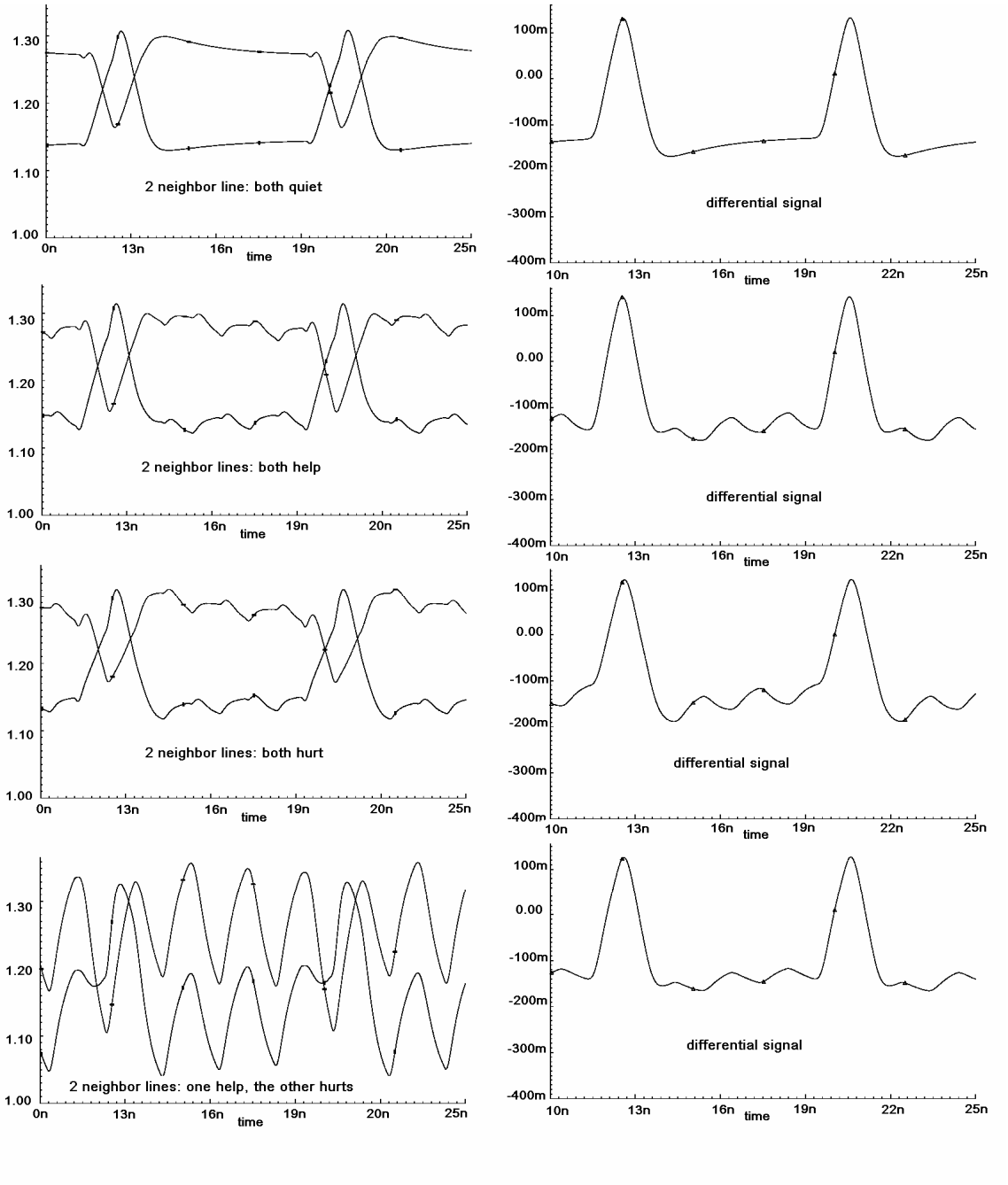


Figure 5.5 Signal waveforms at the receiver input with two neighboring pairs transitioning in various directions.

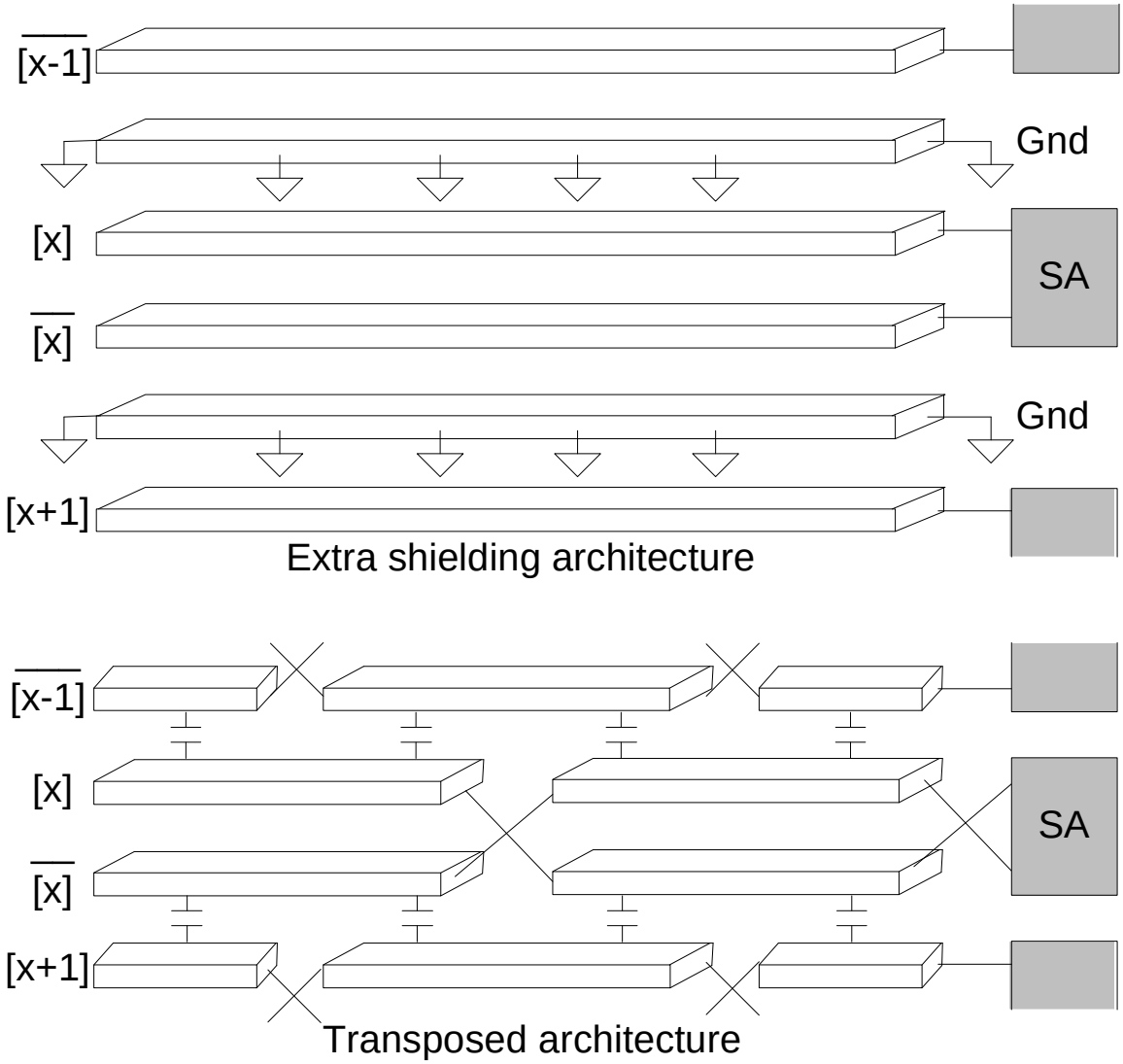
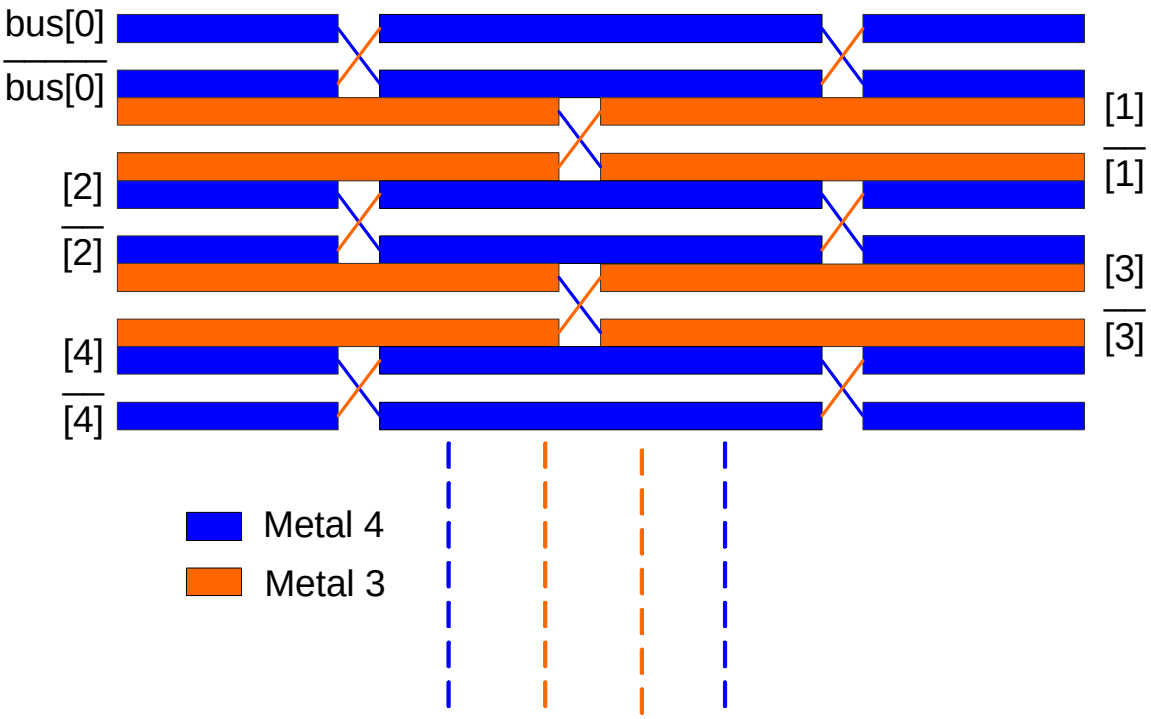
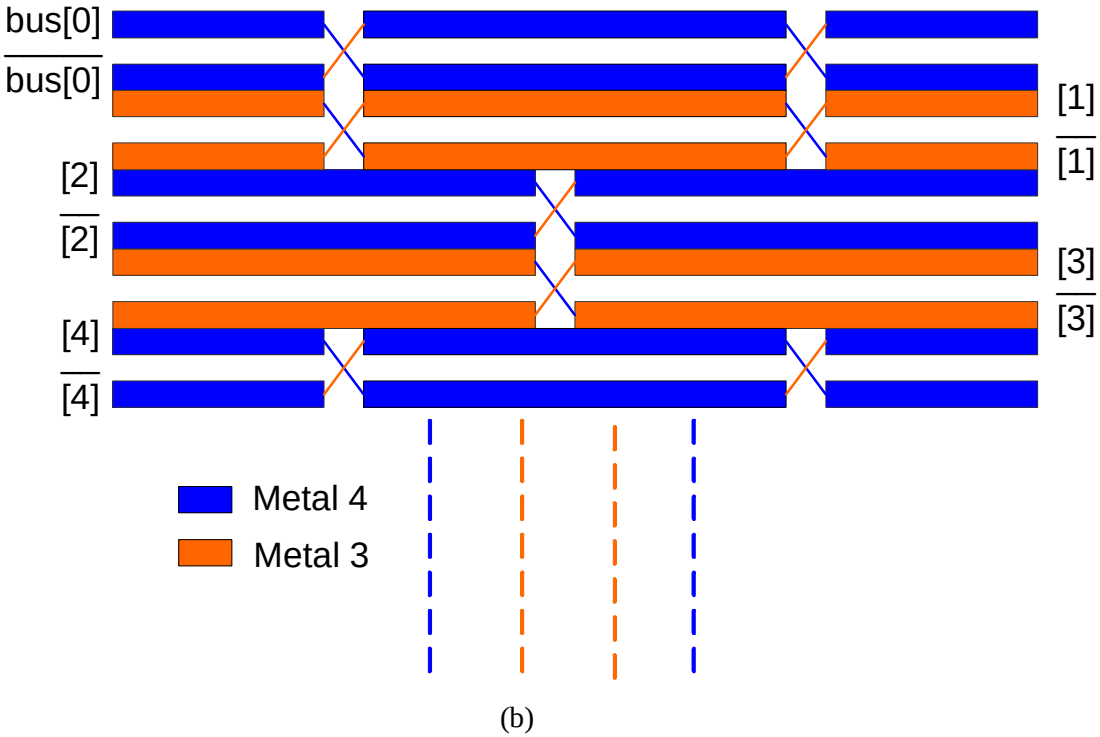


Figure 5.6 Alternative single-layer bus structure Implementation.

An ingenious way to reduce coupling noise and routing area is to use multi metal layers. Because the spacing between metal-3 and metal-4 is even larger than minimum spacing in a single layer, it causes more crosstalk if the two lines of one differential pair are put on different layers. Our approach is to put alternate pairs on two different layers. Figure 5.7 shows two different structures with transposing wires in two different ways.



(a)



(b)

Figure 5.7 Multi-Layer Bus Structure Implementation.

The bus structure as shown in Figure 5.8 is used as the input for Q3D simulation [46] to analyze the crosstalk performance of the bus structures in Figure 5.7 (assuming straight lines with no twisting). Based on the simulation results listed in Table 5.2, the differential capacitance is still dominant due to the minimum pitch used. The main crosstalk in the same metal layer is between the neighboring differential pairs and the main crosstalk between the different layers is between the two closest lines in the two layers.

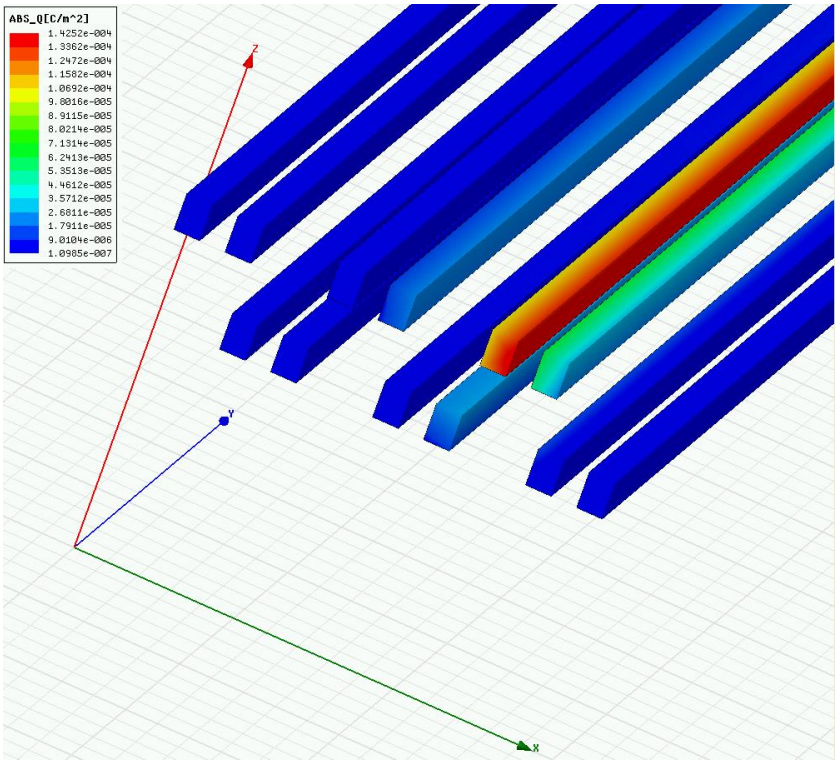


Figure 5.8 Electric charge surface density of multi-layer bus

In Figure 5.7(a), the diagonal Cc between bus[2]_ and bus[3] is minimized by twisting and the parallel Cc between bus[2]_ and bus[4] is about 24.0% of Ctot. This small Cc is the result from the 3x minimum spacing in the same layer. In Figure 5.7(b), the parallel Cc between bus[2]_ and bus[4] is minimized by twisting and the diagonal Cc between bus[2]_ and bus[3] is about 23.6% of Ctot. This small Cc is the result from the diagonal structure in the different layers. Both of the bus structures yield about 25% in area saving.

Table 5.2 Single-layer and multi-layer coupling capacitance from Q3D simulation, assuming straight lines with no twisting.

Cc(fF/cm)	[2]	[2]_	[3]	[3]_	[4]
[2]	5	**830**	74	8	39
[2]_	**830**	5	**257**	74	**262**
[3]	74	**257**	36	830	80
[3]_	8	74	830	28	259
[4]	39	**262**	80	259	5

5.3 Demonstration

5.3.1 Test chip

Figure 5.8 shows the demonstrated 16-bit current-mode differential driver pre-emphasis (CDP) bus. On-chip pseudo-random bit sequence (PRBS) generator and bit error rate

analyzer (BER) are implemented based on an 8-bit data generator structure. Semi-dynamic flip-flop (SDFF) is chosen for the data generator to take advantage of its negative setup time.

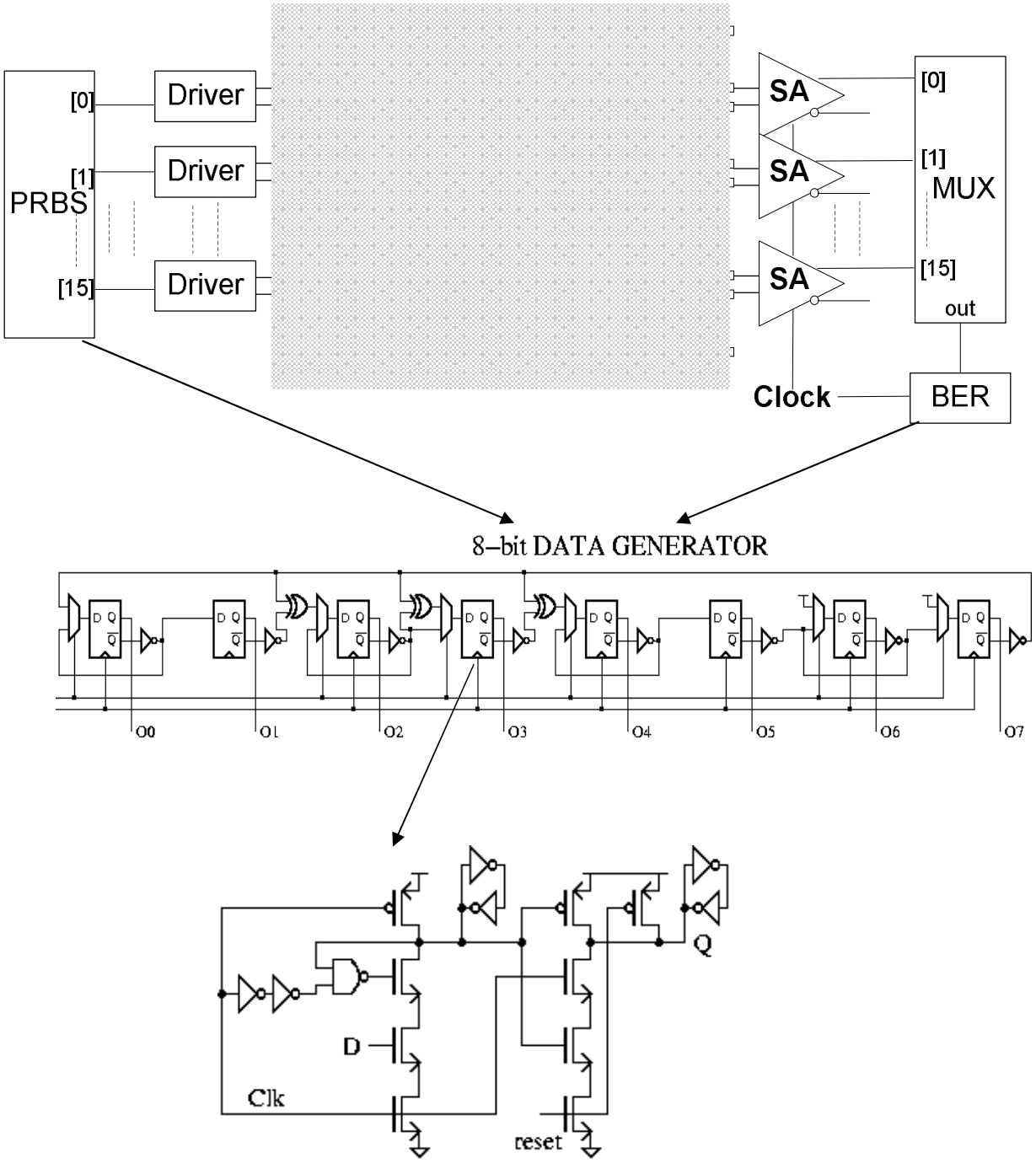


Figure 5.9 Demonstrated 16-bit pre-emphasis bus architecture.

The on-chip BER switches when an error is detected. At data rate as high as 2Gb/s, output is also probed out to an Agilent Error Performance Analyzer to prevent misdetection of even number of error bits.

The die picture of the test chip in TSMC 0.25μm CMOS technology is shown in Figure 5.10. It has test cases of a 16-bit 32Gb/s CDP 5mm bus, an 8-bit 16Gb/s CDP 10mm bus, and a 16-bit 32Gb/s VM 5mm bus as benchmark. All of the buses are meandered due to cost concern. The chip is wire-bonded in the lab. High speed signals are probed in through DC probes and probed out through high-impedance probes.

5.3.2 Measurement results

Figure 5.11 shows the measured eye diagram at the receiver output of CDP. The GGB model 35A high impedance probe attenuates the waveform by ten times and the power slitter, which connects the probe output to both the Tektronix TDS 8000B digital sampling oscilloscope and the Agilent 863130A 3.6Gb/s error performance analyzer (BERT), attenuates the waveform by two times. Both on-chip BER tester and BERT report immeasurable BER ($<10^{-12}$) with 15 minute tests. A 230ps clock offset margin is found at BER 10^{-12} (Figure 5.12) by adjusting the receiver sampling clock.

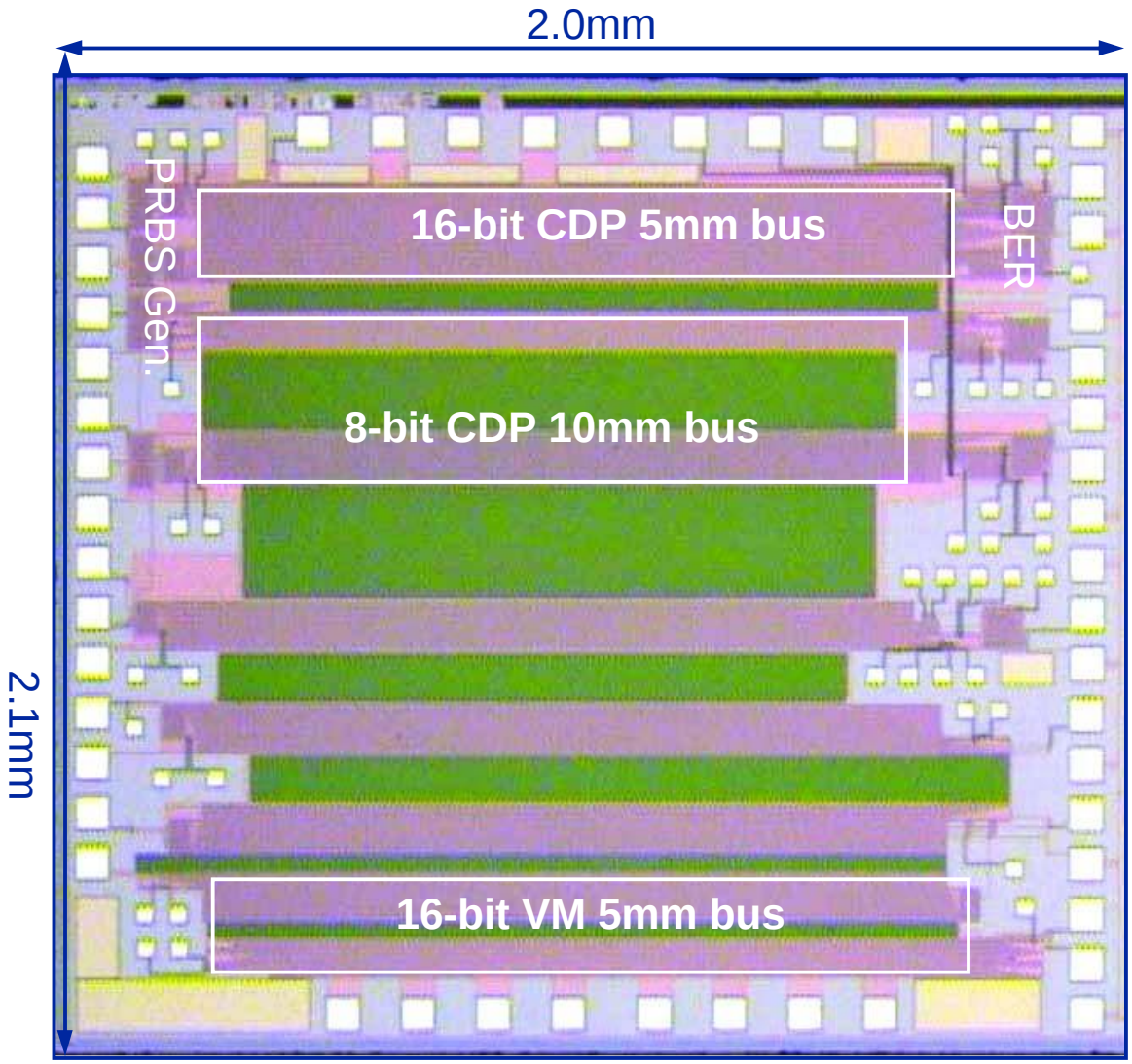


Figure 5.10 Die picture in TSMC 0.25μm CMOS technology.

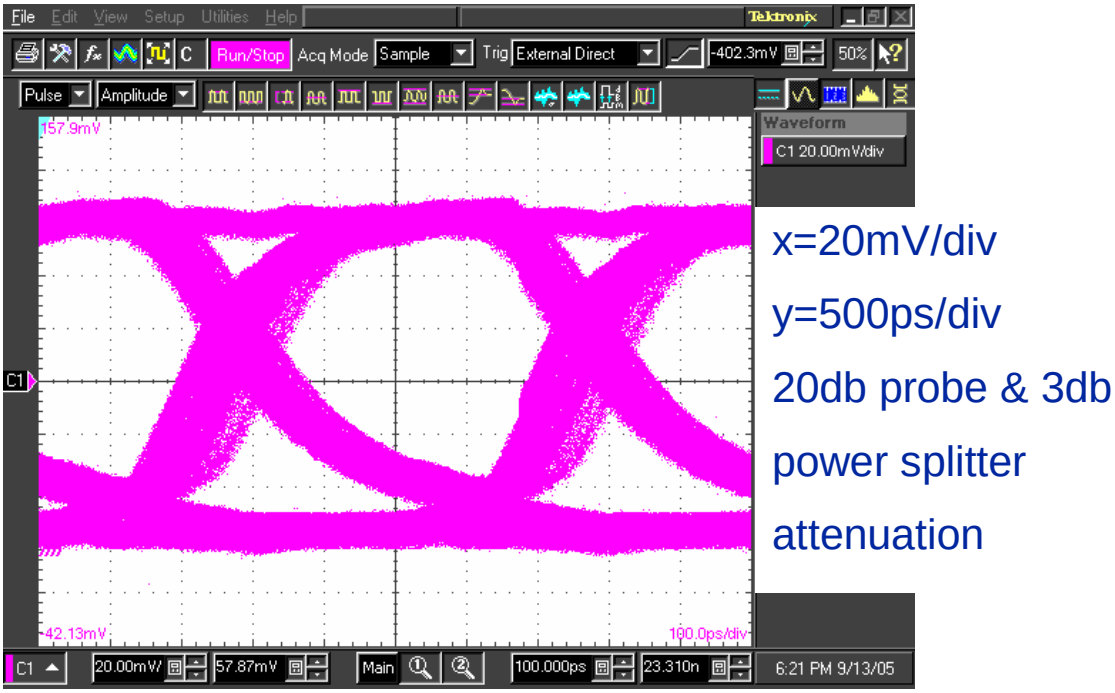


Figure 5.11 Eye diagram at receiver input.

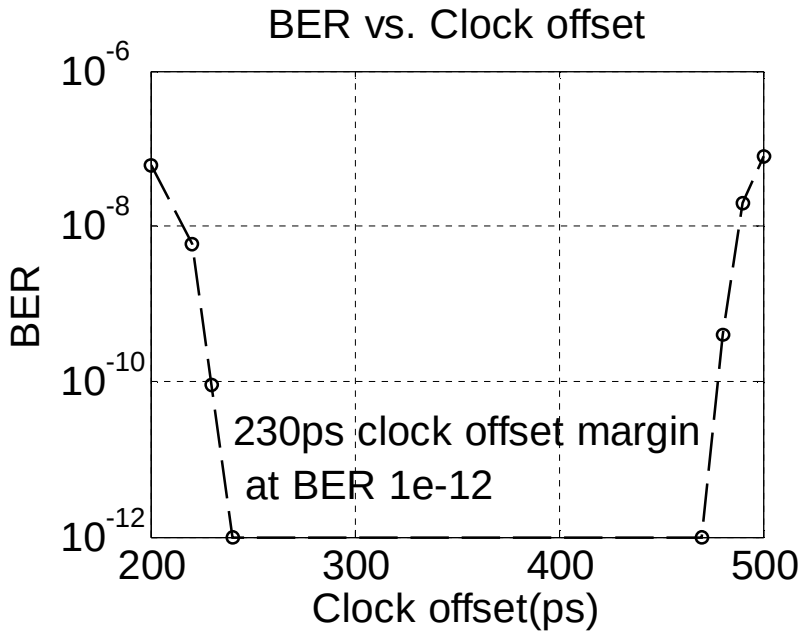


Figure 5.12 Bit error rate performance at different sampling clock offset.

The intra-bus crosstalk performance of the CDP is shown in Figure 5.13. Waveforms at the receiver input are measured when the victim pairs are active and quiet and then converted to differential signal. The crosstalk between adjacent lines mainly behaves as common-mode noise. The differential signal on a pair of quiet lines has only 36mV of noise swing, which is 14.4% of the 250mV measured signal swing. Figure 5.14 shows the eye diagrams of the differential signal at the receiver input and the single-ended signals of the two inputs. A 250mV differential signal swing and 200mV eye opening are observed when all of the 16-bit are switching randomly.

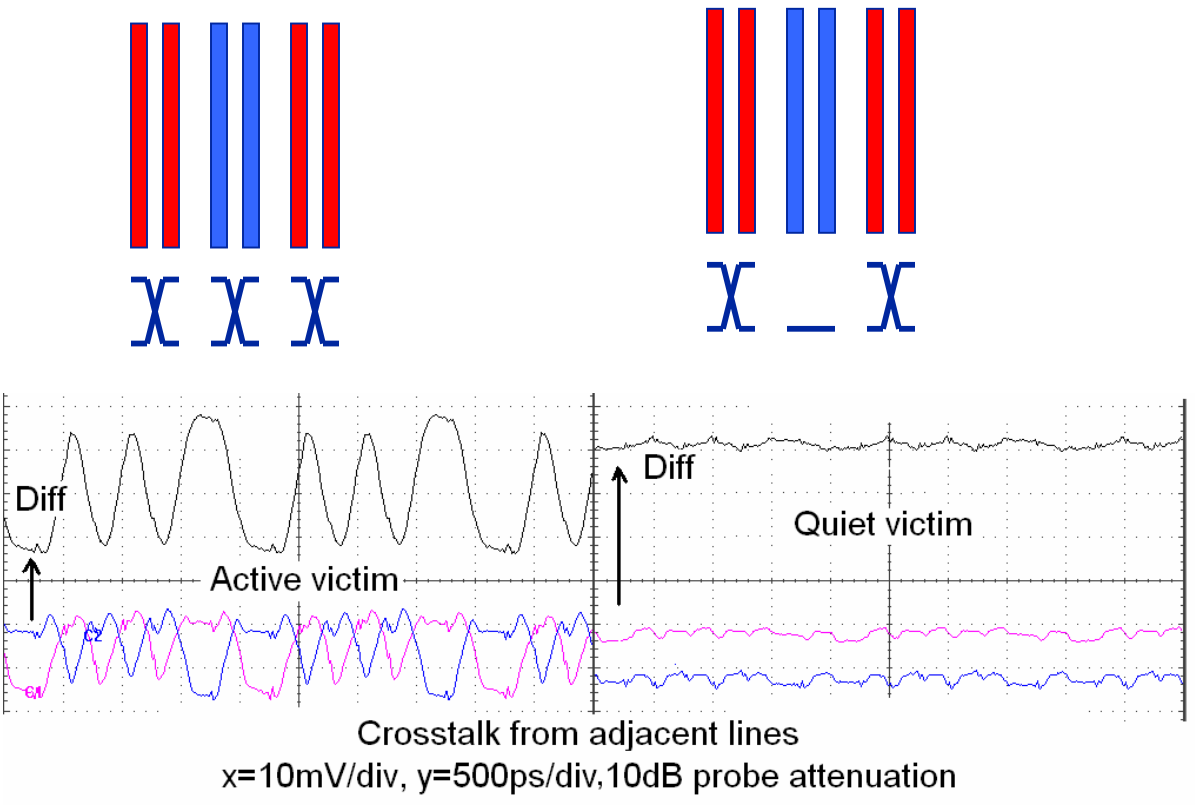


Figure 5.13 Intra-bus crosstalk.

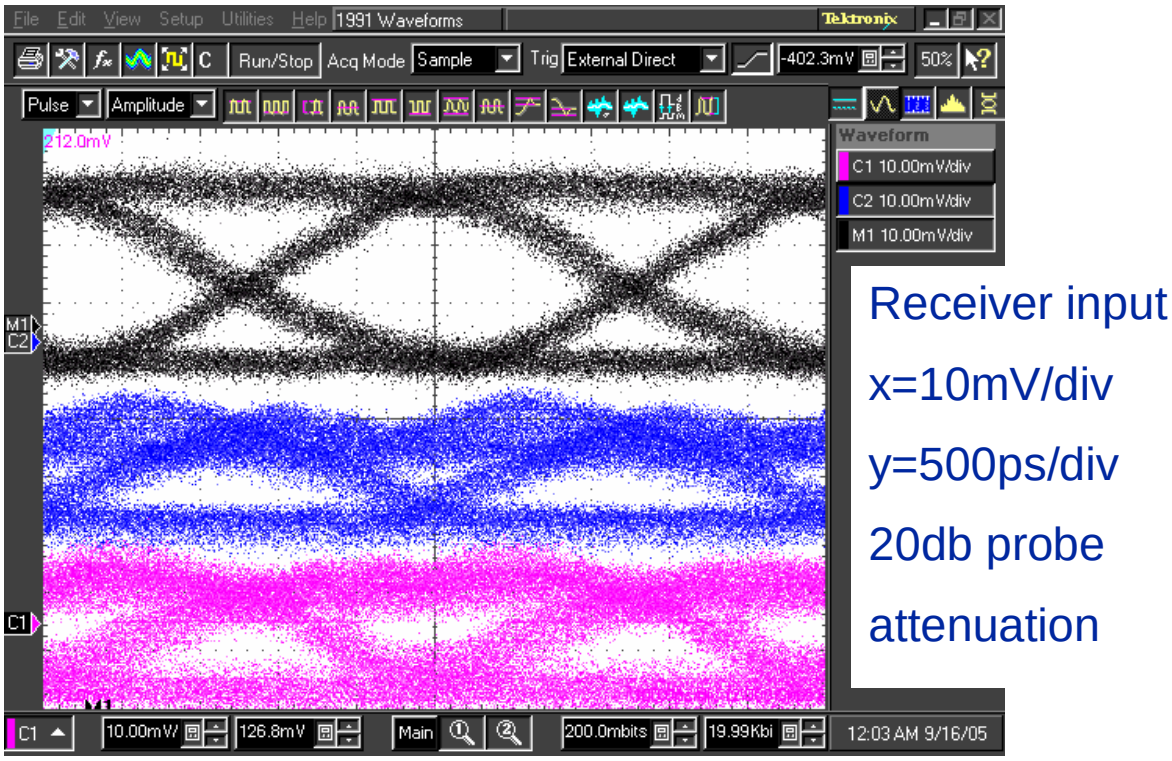


Figure 5.14 Receiver input eye diagrams with crosstalk, differential signal (top) and two single-ended signals (bottom two).

To analyze the crosstalk on the low-swing bus from full-swing bus, a test structure as shown in Figure 5.15 is used, a full-swing 8-bit VM bus crossing orthogonally beneath the 16-bit low-swing bus at the receiver side. The worst case happens when signals switch in the same direction at the same time. Figure 5.16 shows the noise is still mainly common-mode and ignorable due to the small coupling capacitance between different layers. The 10mm bus test case shows the similar BER and noise performance.

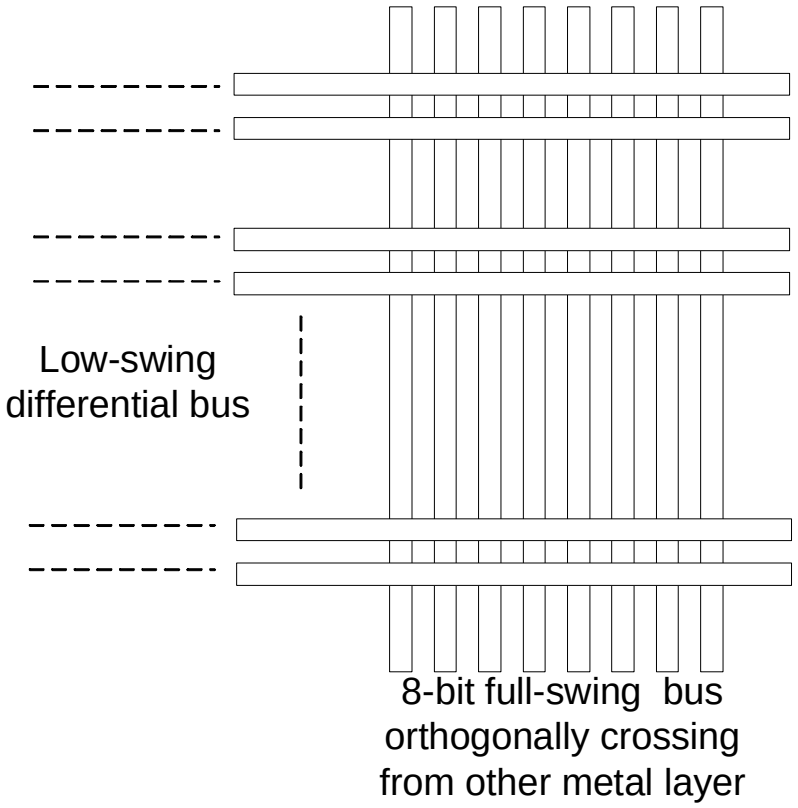


Figure 5.15 8-bit full-swing bus orthogonally crosses 16-bit low-swing bus.

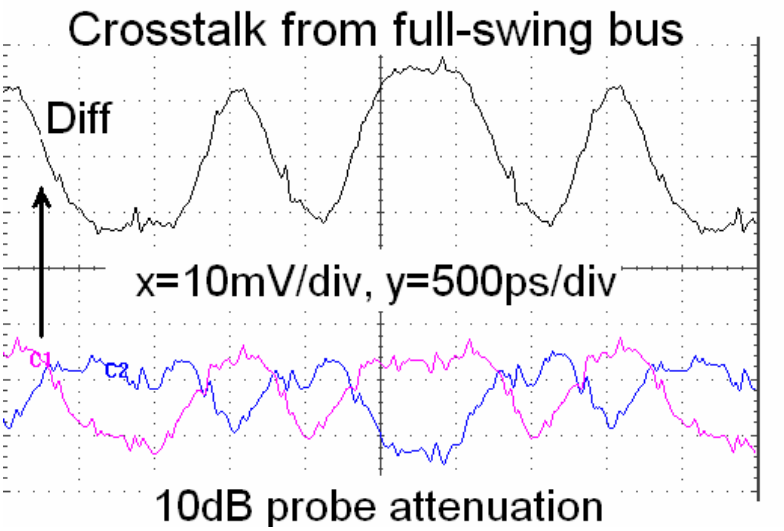


Figure 5.16 Crosstalk from full-swing bus.

Figure 5.17 shows the power dissipation measurement at different data activity factors. For activity factors above 0.1, CDP bus reduces power by 15.0%-67.5% comparing to traditional VM repeater bus. The power performance of CDP bus dose not become worse until the activity factor is as low as 0.07. Notice that tradition current-mode signaling requires activity factor to be above 0.5 (random data) to achieve better power performance.

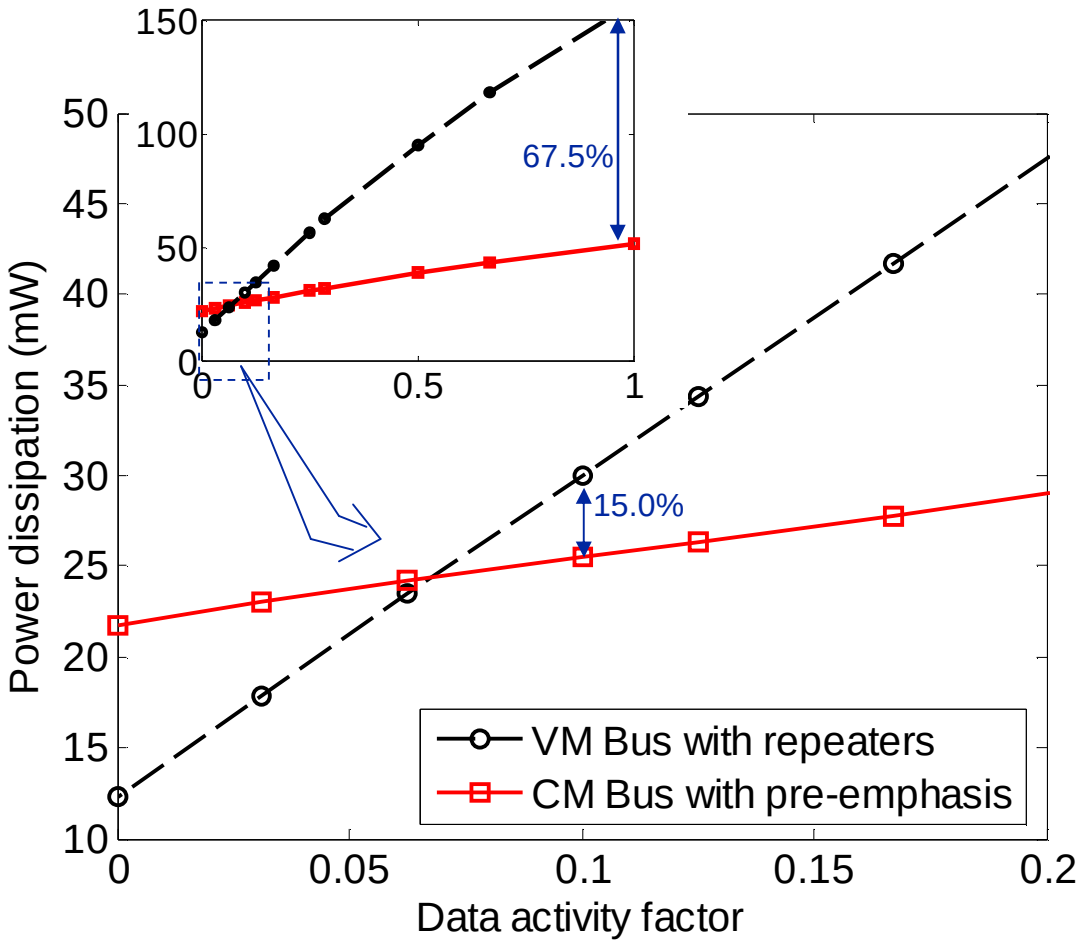


Figure 5.17 Power dissipation comparison.

5.3.3 Bus data activity

To analyze the power performance of the proposed CDP bus in a real application, a time-based Alpha 21264 processor simulator program [49] was modified to extract data activity profiles on instruction and data (i.e. load/store) streams. A total of 100 million 32-bit instructions and another total of 100 million 32-bit data were collected for benchmarks from the SPECint2000 test suite.

Figure 5.18 shows the accumulated data activity profiles of instruction address (a), instruction (b), data address (c), and data (d) patterns from the GCC benchmark (i.e. C Programming Language Compiler). It can be observed that instruction, data address, and data buses exhibit a more uniform activity distributions within the bus lines than the instruction address bus does. The application of CDP can save 52.1% of power on the instruction bus, 13.2% on the data address bus, and 20.4% on the data bus.

CDP only saves 1.4% of power on the instruction address bus, but the instruction address bus exhibits a high correlation of switching activity for the lower order bits, which indicates a higher spatial locality amongst the address streams since instructions are usually stored in adjacent locations of memory. A bus scheme with CDP on lower order bits and traditional VM bus on higher order bits can be proposed to take advantage of this high spatial locality.

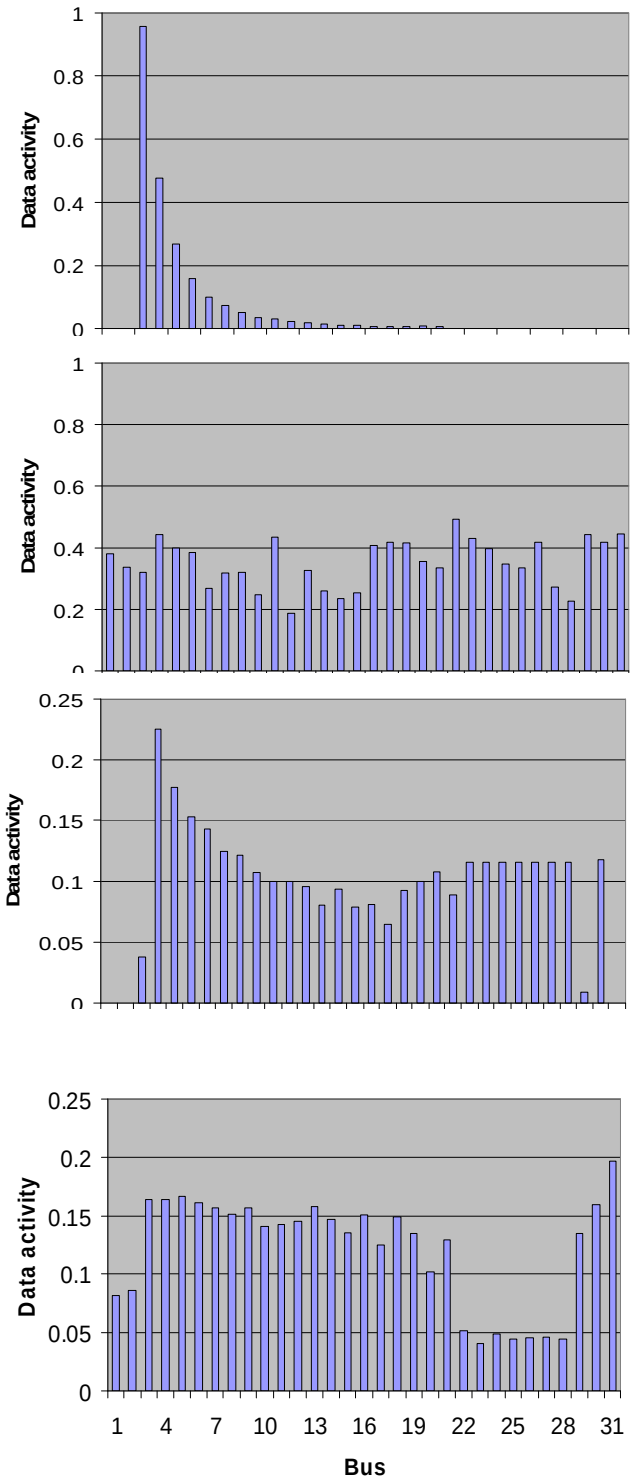


Figure 5.18 Power savings on buses of an Alpha 21264 microprocessor.

5.3.4 Simulation results

Performance improvement on peak current and process variation cannot be obtained through measurement but is analyzed through HSPICE simulation.

The peak current consumption in a single test channel for the two bus architectures is compared in Figure 5.19. Due to its small drivers and small signal swing, CDP reduces the peak current by 70.0% over the single-ended full-swing VM bus. This is a significant for simultaneous switch noise (SSN) reduction. The static current is 0.126mA, only 0.158pJ/bit at 2Gb/s.

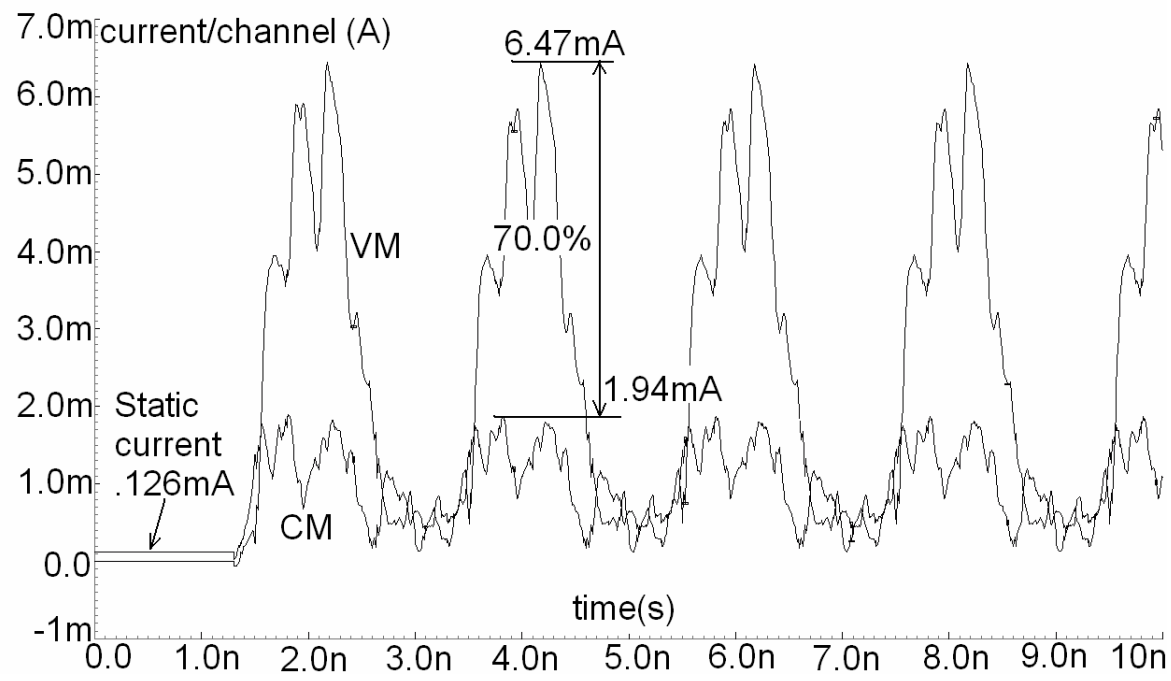


Figure 5.19 Peak current comparison of VM bus with repeaters and CM bus with pre-emphasis.

Table 5.3 shows the impact of transistor process variation on driver and delay (Figure 5.20). Traditional VM signaling has ±28% variation for FF and SS corners, while CDP bus improves it to ±18%. The variation of SF and FS corners falls between FF and SS corners.

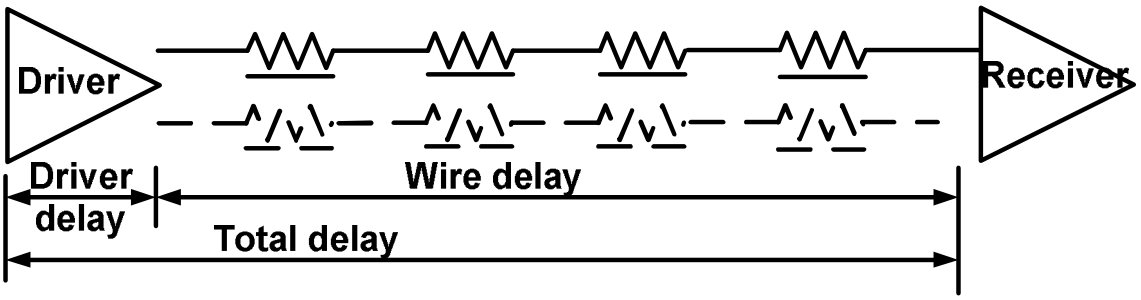


Figure 5.20 Delay illustration.

Table 5.3 Transistor process variation.

Delay (ps)	VM			
	Typ	FF	SS	Variation
Driver	313	257	399	±28%
Wire	541	478	697	±28%
Total	854	735	1096	±28%
	CDP			
Driver	167	125	222	±33%
Wire	228	213	244	±7.0%
Total	395	338	466	±18%

5.4 Potential Architectures

Because the low-swing nature of CDP bus, it is more suitable for application with mono-direction buses. For a bi-direction application shown in Figure 5.21, extra shielding between bus lines is required to deal with near-end and far-end crosstalk.

For a multi-drop bus application, differential sense-amplifiers can be simply connected to the current loop (Figure 5.22). The differential swing across the current loop is from 250mV at receiver side to 500mV at receiver side.

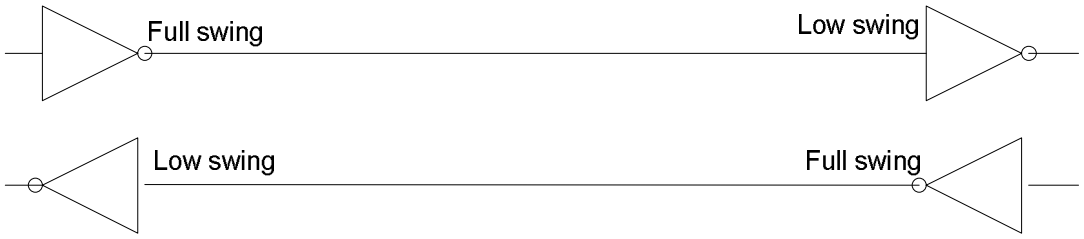


Figure 5.21 Bi-direction bus.

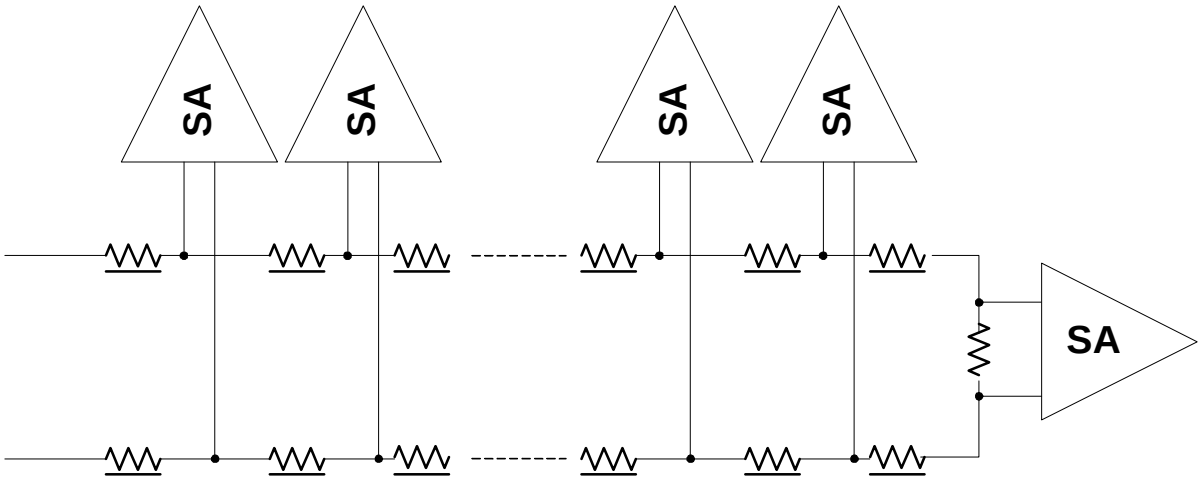


Figure 5.22 Multi-drop bus architecture.

Figure 5.23 shows a high-level repeater estimation flow for Intel Itanium microprocessors [5]. Driver pre-emphasis technique can easily fit into this estimation flow by just adding one-step of high-level estimation.

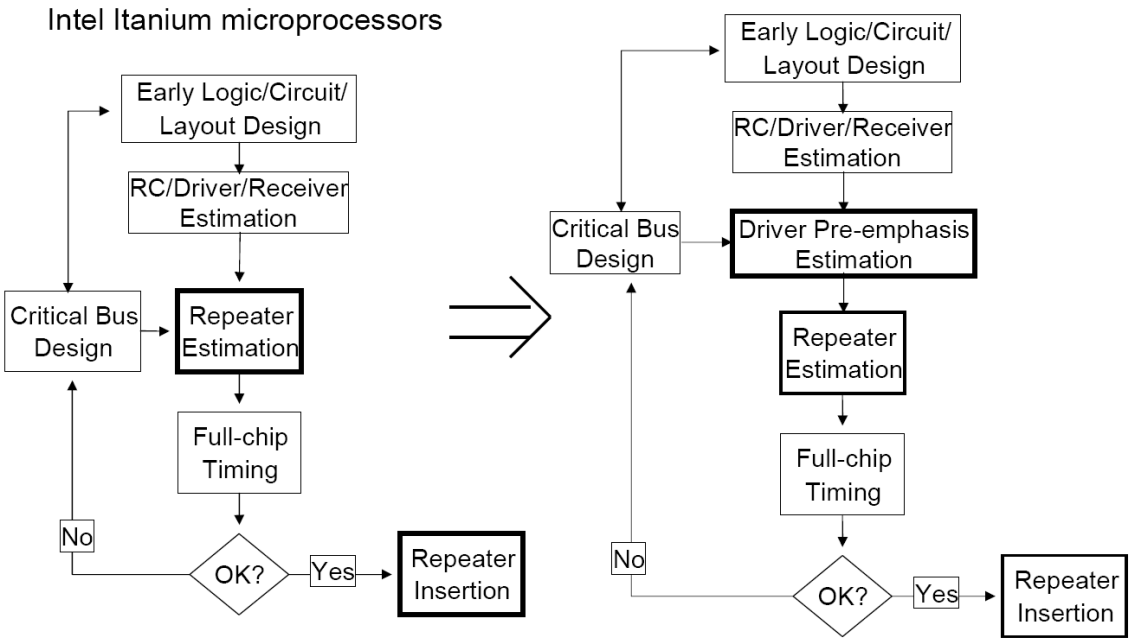


Figure 5.23 High-level repeater estimation flow change.

5.5 Summary

Table 5.4 shows the performance summary of the traditional VM repeater bus and the CDP bus. The CDP bus obtains better delay latency performance, better power performance, and better peak current (SSN) performance. Its crosstalk degradation can be controlled in a confined global bus domain. Its bus routing area overhead can be reduced by trading off noise or bandwidth in different bus structures. It does have 64.4% area penalty of driver and

receiver circuit, but it is only 2% of the total bus routing area. In addition, the CDP bus had negligible data-dependent delay variation.

As long as a stack of four transistors can still be implemented in a differential sense amplifier, the CDP circuit should scale as technology scales. The 125mV signal swing at the receiver input may not decrease significantly, but as Vdd scales, the static power will decrease and it makes CDP more promising for power-saving.

Table 5.4 Performance summary.

	VM repeater	CDP	Improvement
Latency (ns)	.828	.594	28.3%
Crosstalk	.252V/2.5V	36mV/250mV	-4.3%
Peak current (mA)	6.47	1.94	70.0%
Power (mW) (act=0.10)	30.0	25.5	15.0%
Width of routing area (µm)	50.4	54.4	-7.9%
Circuit area(μm^2)	8030	13200	-64.4%
R(kΩ/cm) Ccoup(pF/cm) Ctotal(pF/cm)	633 .393 2.57	1.9k .358 2.48	N/A

Chapter 6. Conclusions

6.1 Summary

This work dealt with trade-offs among the on-chip signaling design metrics, delay, throughput, power, noise, and area, by applying different communication techniques, equalization, current-mode, and differential signaling techniques to on-chip global communication.

The basic and novel idea behind was first to achieve bandwidth and power improvement by trading off noise margin and signal swing. The degradation of noise margin and signal swing is then limited within a confined domain of global buses, where noise levels are tightly controlled by circuit techniques and bus structures. Finally, power is traded back for signal swing at receiver side. Because the power saving on interconnects dominates the extra receiver power overhead, both power and bandwidth are improved.

This work also elaborated that it was under designers' control to trade off between noise and area. Intra-bus crosstalk could be alleviated by using shielding or transposing structure to reduce bus routing area.

The demonstration results showed the number of repeaters required was minimized by improving interconnect channel bandwidth with the proposed signaling techniques. It helped reduce delay latency and power consumption and save repeater layout area and blockages.

Peak current, which causes simultaneous switching noise (SSN), was largely reduced because interconnect channel bandwidth was no longer improved by large drivers and repeaters. Data-dependent delay variation was improved due to a small coupling capacitance ratio Process-dependent delay variation was improved due to both the process-compensation scheme in pre-emphasis driver and the robustness of current-mode signaling against process variation.

The proposed differential structure produced a well-defined current return path for long bus lines. It also reduced the unpredictability of inductive effects. In addition, equalization and current-mode signaling increased bandwidth and allowed for narrow and resistive interconnects. This further decreased inductive effect

A simplified delay design guideline was derived and verified to show that the long interconnects used in this work were still RC dominated.

6.2 Future work

On-chip signaling is a multi-dimension design of delay, throughput, power, noise, and area. It is also a cross-field design of architecture, circuit, and process.

A benchmark to verify performance improvement is critical when so many design perspectives are involved. Instead of building a voltage-mode repeater case for comparison, the idea is to put this proposed driver pre-emphasis current-sensing differential bus architecture in a real memory or DSP application. This can verify how well noise levels can be controlled in this low-swing bus region and make the improvement statements more convincing.

The major power consumption of global communication is from clock distribution. It is worthy to explore the possibilities of using current-mode differential signaling to reduce power consumption on clock distribution network and using even fractional equalization to improve clock slew rate.

The scalability of this work is determined by the offset of sense-amplifier receivers. Although the receiver half-Vdd bias built by the current loop alleviate the threshold voltage deviation of input transistors, it is still important to do a thorough research on offset effects.

Appendix A. Power-Optimal Repeater Insertion

Figure A.1 shows an interconnect line with length l, distributed resistance R_0 and capacitance C_0 and evenly separated into k segments by repeaters. Same size of driver, repeaters and receiver, W times of minimum transistors, is assumed If only the repeater output resistance Rs/W and input capacitance $C_L W$ are considered and the repeater intrinsic delay is ignored. The signal propagation delay on one segment of interconnect can be derived as [12],

$$\frac{t_{05}}{(R/k)(C/k)} = 0.377 + 0.693(R_T C_T + R_T + C_T) \qquad \text{(A-1)}$$

Where $R=R_0 l$, $C=C_0 l$ and $R_T=(Rs/W)/R$, $C_T=(C_L W)/C$.

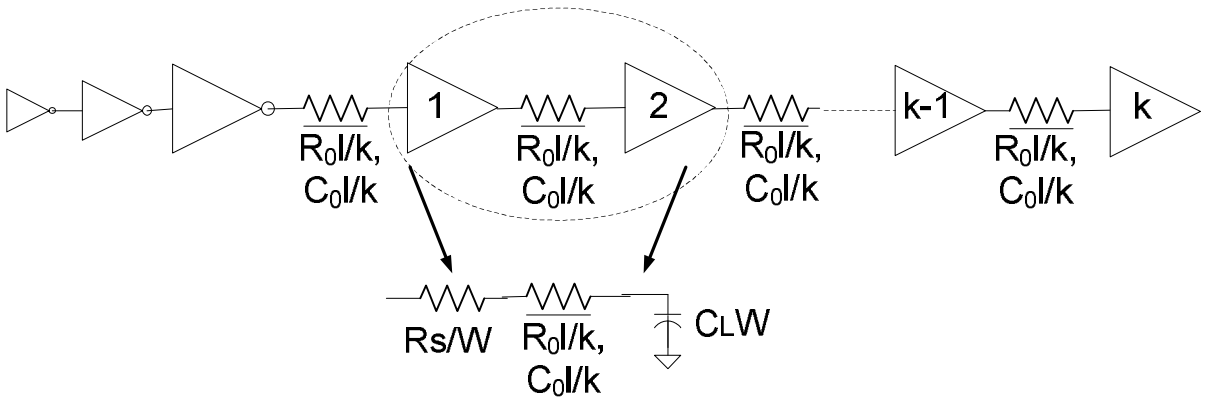


Figure A.1 Circuit model for interconnect with repeaters.

$$t_{05} = 0.377\frac{RC}{k^2} + 0.693(R_S C_0 + \frac{R_S}{W}\frac{C}{k} + \frac{R}{k}C_0 W) \quad \text{(A-2)}$$

The total delay is,

$$kt_{05} = 0.377\frac{RC}{k} + 0.693(R_S C_0 k + \frac{R_S}{W}C + RC_0 W) \quad \text{(A-3)}$$

To find the most power-optimal repeater insertion at a designated timing target d_0, Lagrange relaxation method is used to solve the function [47], [18],

$$f = \min Power = akW \quad \text{(A-4)}$$

$$\begin{aligned} d_0 = kt_{05} &= 0.693R_S C_0 k + 0.377\frac{RC}{k} + 0.693RC_0 W + 0.693\frac{R_S}{W}C \\ &= Ak + \frac{B}{k} + CW + \frac{D}{W} \end{aligned} \quad \text{(A-5)}$$

Where $A=0.693R_SC_0$, $B=0.377RC$, $C=0.693RC_0$, and $D=0.693R_SC$.

$$f = akW + l(d - d_0) = akW + l(Ak + \frac{B}{k}) + l(CW + \frac{D}{W}) - ld_0 \quad \text{(A-6)}$$

$$aW + l(A - \frac{B}{k^2}) = 0 \Rightarrow W + \frac{l}{a}(A - \frac{B}{k^2}) = 0 \quad \text{(A-7)}$$

$$ak + l(C - \frac{D}{W^2}) = 0 \Rightarrow k + \frac{l}{a}(C - \frac{D}{k^2}) = 0 \quad \text{(A-8)}$$

Where λ is the Lagrange coefficient.

From (A-7) and (A-8),

$$\frac{W}{k} = \frac{A - \frac{B}{k^2}}{C - \frac{D}{w^2}} \tag{A-9}$$

From (A-5),

$$CW + \frac{D}{W} = d_0 - Ak - \frac{B}{k} \tag{A-10}$$

From (A-9) and (A-10),

$$\frac{Ad_0}{2}k^2 + (DC - AB - \frac{d^2}{4})k + \frac{B}{2}d_0 = 0 \tag{A-11}$$

$$2CW = d_0 - 2\frac{B}{k} \tag{A-12}$$

Power-optimal segment number and repeater size can be obtained by solving (A-11) and (A-12). For 10mm long and 0.45μm wide metal-4 lines in TSMC 0.18μm technology, R_0=1.73kΩ/cm and C_0=3.48pF/cm [45]. Same Rs and C_L value are used as in [18]. The optimal segment number is k=5 and the optimal repeater size is W=36xW_{min} for most power-optimal repeater insertion for a wire target delay of 1ns. With 1.44μm pitch, R0=1.08kΩ/cm and C0=2.78pF/cm, the optimal segment number is k=3 and the optimal repeater size is W=20xWmin for the same wire target delay.

Appendix B. On-Chip Interconnect Characterization

A 1.8mm long interconnect line was fabricated in the UMC .18µm copper process for characterization. As shown in Figure B.1, 1.44µm wide metal-5 lines are used for both the signal line in the middle and the two return paths on the sides. Metal-4 grid is used to reduce inductive effects by providing shorter current return paths. Both time domain reflectometry (TDR) analysis and S-parameter analysis are considered to characterize the line parameters.

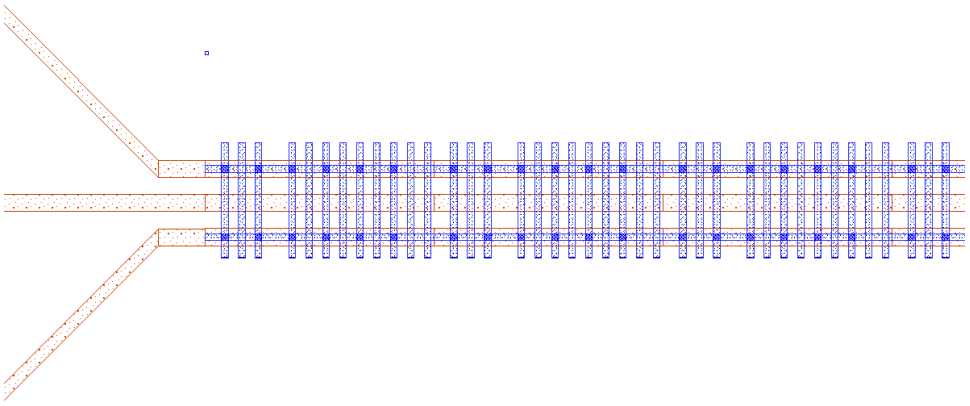


Figure B.1 Layout of Cu interconnect lines.

To make TDR measurement feasible, the line flight time should be at least 10 times longer than the 17.5ps rise time, which is the fastest that the sampling head can offer. The line length needs to be at least,

$$length = t_{flight} \cdot C \Big/ \sqrt{e} = \frac{175ps \cdot 3x10^{8}\, m/s}{\sqrt{4}} \approx 2.6cm \qquad \text{(B-1)}$$

Where C is the speed of light and ε=4 is the approximated silicon dielectric. With the 1.8mm line length, the TDR measurement shown in

Figure B.2 is typically a RC response as expected.

Therefore, S-parameter analysis is the only option to characterize the line distributive parameters, R, L and C. At least two different lengths of interconnects are needed to characterize S-parameters in one frequency range of interest. Moreover, open and short de-embedding structures are necessary to calibrate measurement results.

The transfer function of an interconnect line can be derived by using ABCD parameters,

$$H(S) = \frac{V_2(S)}{V_1(S)} = \frac{1}{[A + Z_S C] + \frac{1}{Z_L}[B + Z_S D]}$$

$$= \frac{1}{\left[\cosh(qh) + \frac{Z_S}{Z_o}\sinh(qh)\right] + \frac{1}{Z_L}[Z_0 \sinh(qh) + Z_S \cosh(qh)]} \qquad \text{(B-2)}$$

Where Z_S and Z_L are the source and load impedance, h is the line length, $q = \sqrt{(R_0 + SL_0)SC_0}$ is the propagation constant, and $Z_0 = \sqrt{(R_0 + SL_0)/SC_0}$ is the characteristic impedance with R_0, L_0, and C_0 as the line distributed parameters.

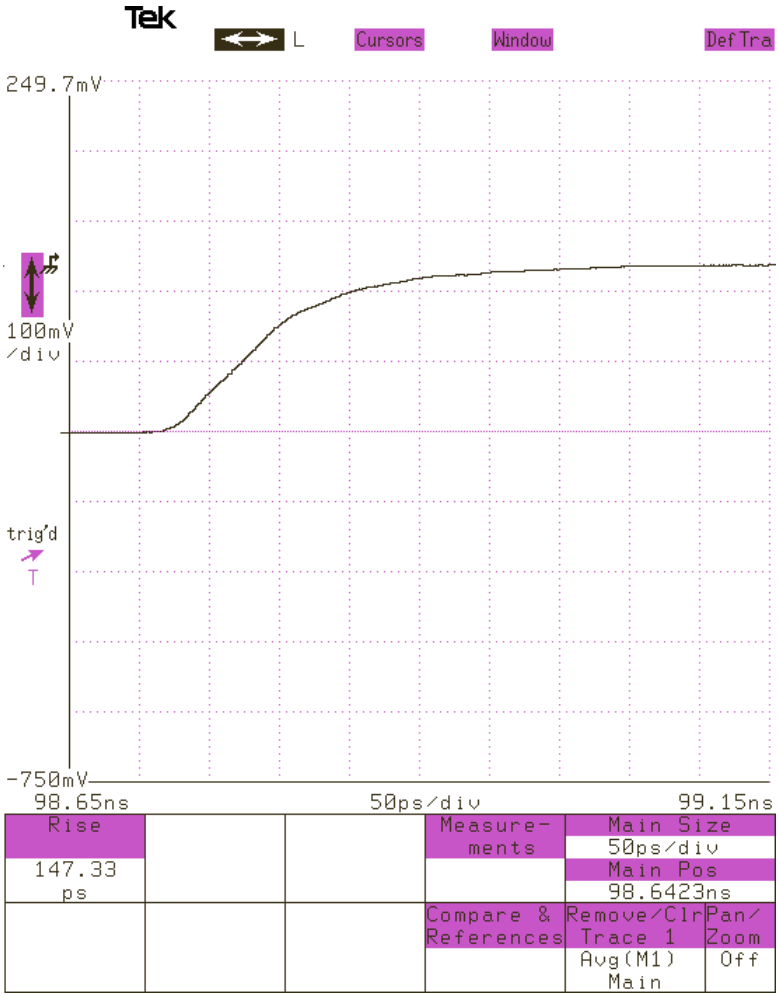


Figure B.2 TDR measurement result.

The conversion between S and ABCD parameters,

$$A=\cosh(\gamma h)=\frac{(1+S_{11})(1-S_{22})+S_{12}S_{21}}{2S_{21}}$$

$$B=Z_0\sinh(\gamma h)=Z_0\frac{(1+S_{11})(1+S_{22})-S_{12}S_{21}}{2S_{21}}$$

$$C=\frac{1}{Z_0}\sinh(\gamma h)=\frac{1}{Z_0}\frac{(1-S_{11})(1-S_{22})-S_{12}S_{21}}{2S_{21}} \quad \text{(B-3)}$$

$$D=\cosh(\gamma h)=\frac{(1-S_{11})(1+S_{22})+S_{12}S_{21}}{2S_{21}}$$

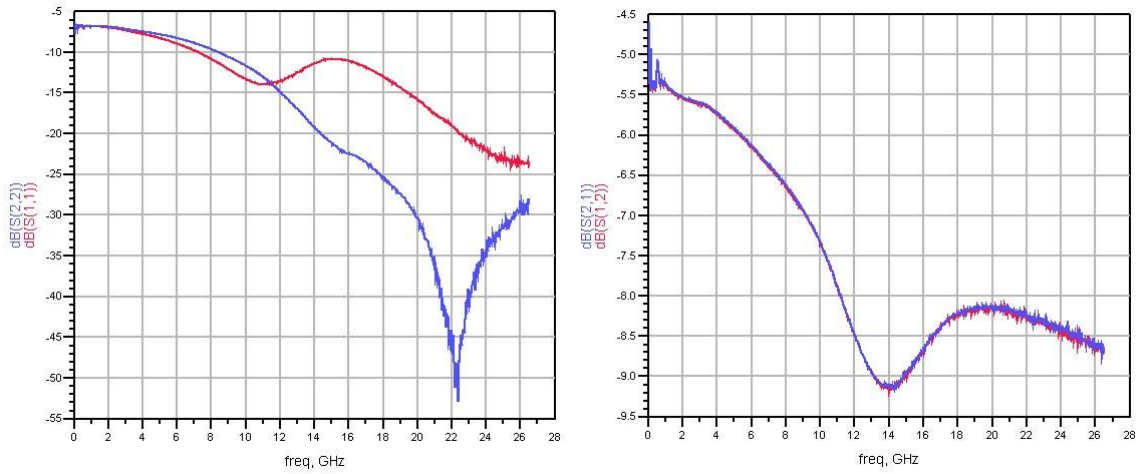
θ(f) can be obtained from (B-3). Combined with $q = \sqrt{(R_0 + SL_0)SC_0}$, we have,

$$\begin{aligned} \mathrm{Re}(q^2) &= -w^2 C_0(f) L_0(f) \\ \mathrm{Im}(q^2) &= -wR_0(f) C_0(f) \end{aligned} \tag{B-4}$$

For sub-gigahertz frequency, $R_0(f) = R_{DC}$, we can therefore obtain $C_0(f) = \frac{\mathrm{Im}(q^2)}{wR_{DC}}$. Unlike R_0 and L_0, C_0 has very small frequency dependence, finally we have,

$$\begin{aligned} C &= \frac{\mathrm{Im}(q^2)}{wR_{DC}} \\ R(f) &= \frac{\mathrm{Im}(q^2)}{wC} \\ L(f) &= \frac{\mathrm{Re}(q^2)}{-w^2 C} \end{aligned} \tag{B-5}$$

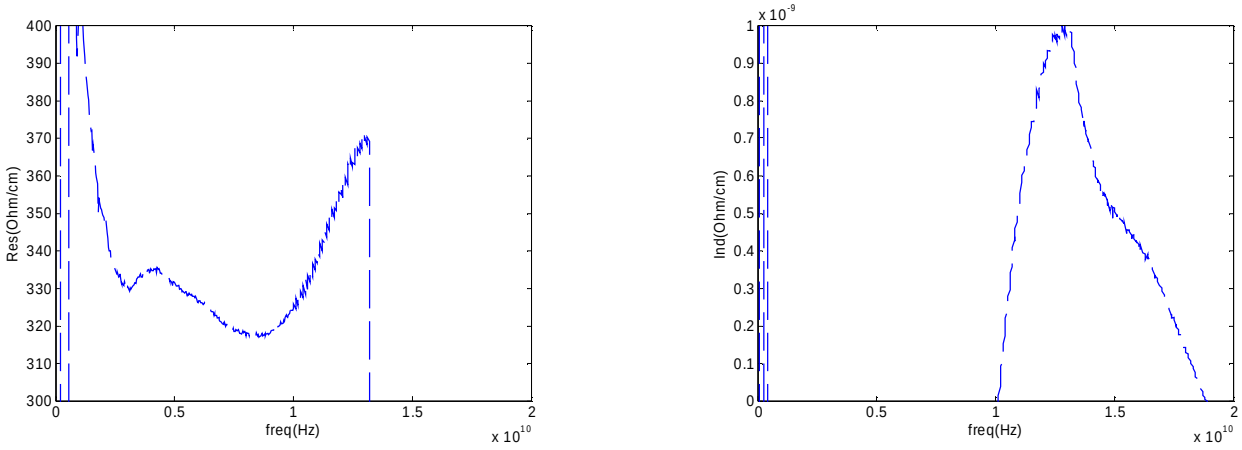
Figure B.3 shows the magnitude of the measured S-parameter. The difference between S11 and S22 might be caused by parasites.



(a) **(b)**

Figure B.3 Magnitude of S parameters (a) S11, S22, (b) S12, S21.

Network analyzer gives R_{DC}=60.9Ω. $R_{0DC} = 60.9/0.18cm = 338\Omega/cm$. Using (B-5), we have $C = 4pF/cm$. $R_0(f)$ and $L_0(f)$ are shown in Figure B.4. R_0 increases from 330Ω/cm to 370Ω/cm at frequency 10-13GHz. L_0 decreases from 1nH/cm at 13GHz. The resistance and inductance value is in the expected range.



(a) **(b)**

Figure B.4 Frequency-dependent line (a) resistance and (b) inductance.

Bibliography

[1] *International Technology Roadmap for Semiconductors (ITRS),* Semiconductor Industry Association, 2003.

[2] C. Hu, "CMOS for one more century?" Custom Integrated Circuits Conference, Keynote Speech, Oct 2004.

[3] R. Ho, K. W. Mai, and M. A. Horowitz, "The future of wires," Proc. IEEE, vol. 89, no. 4, pp. 490-504, Apr 2001.

[4] H. Bakoglu, *Circuits, Interconnections and Packaging for VLSI,* Addison-Wesley, 1990.

[5] R. McInerney, et al., "Methodology for repeater insertion management in the RTL, layout, floorplan and fullchip timing databases of the Itanium microprocessor," ISPD Proc., pp. 99-104, Apr 2000.

[6] Y. I. Ismail and E. G. Friedman, "Effects of inductance on the propagation delay and repeater insertion in VLSI circuits," IEEE Trans. VLSI Syst., vol. 8, no. 2, pp. 195-206, Apr 2000.

[7] J. Cong, "An interconnect-centric design flow for nanometer technologies," Proc. of the IEEE, vol. 89, no. 4, pp. 505-528, Apr 2001.

[8] J. M. Rabaey, *Degital Integrated Circuits: A Design Perspective*, Prentice Hall, 1996.

[9] R. Venkatesan, J. A. Davis, and J. D. Meindl, "Compact distributed RLC interconnect models – part IV: unified modles for time delay, crosstalk, and repeater insertion," IEEE trans. Electron Devices, vol. 50, pp. 1094-1102, Apr 2003.

[10] W. Dally and J. Poulton, *Digital Systems Engineering*, Cambridge Univ. Press, Cambridge, UK, 1997.

[11] A. Deutsch, et al., "On-chip wiring design challenges for gigahertz operation," Proc. of the IEEE, vol. 89, no. 4, pp. 529-555, Apr 2001.

[12] T. Sakurai, "Closed-form expressions for interconnection delay, coupling, and crosstalk in VLSI's," IEEE Trans. on Elec. Dev., vol. 40, pp.118-124, Jan 1993.

[13] M. L. Mui, K. Banerjee, and A. Mehrotra, "A global interconnect optimization scheme for nanometer scale VLSI with implications for latency, bandwidth, and power dissipation," IEEE Trans. on Electron Devices, vol. 51, no.2, pp. 195-203, Feb 2004.

[14] R. Kumar, “Interconnect and noice immunity design for the Pentium 4 processor,” DAC, pp. 938-943, Jun 2003.

[15] D. Sylvester and H. Kaul, “Power-driven challenges in nanometer design,” IEEE Design & Test of Computers, vol. 18, issue. 6, pp. 12-21, Nov 2001.

[16] D. Sylvester and K. Keutzer, “A global wiring paradigm for deep submicron design,” IEEE Trans. on CAD of ICs and Systems, vol. 19, no. 2, pp. 242-252, Feb 2000.

[17] M. Khellah, J. Tschanz, Y. Ye, S. Narendra, and V. De, “Static pulsed bus for on-chip interconnects,” Symp. VLSI Circuits, pp. 78-79, Jun 2002.

[18] K. Banerjee and A. Mehrotra, “A power-optimal repeater insertion methodology for global interconnects in nanometer designs,” IEEE Trans. Electron Devices, vol. 49, no. 11, pp. 2001-2007, Nov 2002.

[19] V. Veremey and R. Mittra, “Efficient computation of interconnect capacitances using the domain decomposition approach,” 7th Topical Meeting Electrical Performance of Electronic Packaging, proc. dig. Of IEEE, pp. 277-280, Oct 1998.

[20] A. E. Ruehli, “Inductance calculation in a complex integrated circuit environment,” IBM J. Res. Develop., pp. 16470-16485, 1972.

[21] R. Bashirullah, W. Liu, and R. K. Cavin, “Delay and power model for current-mode signaling in deep submicron global interconnects,” CICC, proc. of IEEE, pp. 513-516, May 2002.

[22] H. Zhang, V. George, and M. Rabaey, “Low-swing on-chip signaling techniques: effectiveness and robustness,” IEEE Trans. VLSI, vol. 8, no. 3, pp. 264-272, Jun 2000.

[23] C. Svensson, “Optimum voltage swing on on-chip and off-chip interconnect,” IEEE J. Solid-State Circuits, vol. 36, no. 7, pp. 1108-1112, Jul 2001.

[24] R. Ho, K, Mai, and M. Horowitz, “Efficient on-chip global interconnects,” Symp. VLSI Circuits, pp. 271-274, Jun 2003.

[25] R. Bashirullah, et al., “A 16Gb/s adaptive banwidth on-chip bus based on hybrid current/voltage mode signaling,” Symp. VLSI Circuits, pp. 392-393, Jun 2004.

[26] D. Schinkel, et al., “A 3Gb/s/ch transceiver for RC-limited on-chip interconnects,” ISSCC, pp. 386-387, Feb 2005.

[27] R. Chang, N. Talwalkar, C. Yue, and S. Wong, “Near speed-of-ling signaling over on-chip electrical interconnects,” IEEE J. Solid-State Circuits, vol. 38, no. 5, pp. 834-838, May 2003.

[28] S. D. Naffziger, et al., "The implementation of the Intanium 2 microprocessor," IEEE J. Solid-State Circuits, vol. 37, no. 11, pp. 1448-1460, Nov 2002.

[29] E. Seevinck, P. van Beers, and H. Ontrop, "Current-mode techniques for high-speed VLSI circuits with application to current sense amplifier for CMOS SRAM's," IEEE J. Solid-State Circuits, vol. 26, no. 4, pp. 525-536, April 1991.

[30] K. Agarwal, D. Sylvester, and D. Blaauw, "An effective capacitance based driver output model for on-chip RLC interconnects," Design Automation Conference, pp. 376-381, Jun 2003.

[31] Y. I. Ismail, E. G. Friedman, and J. L. Neves, "Figures of merit to characterize the importance of on-chip inductance," IEEE Trans. VLSI Syst., vol. 7, no. 4, pp. 442-449, Dec 1999.

[32] L. T. Pillage and R. A. Rohrer, "Asymptotic waveform evaluation for timing analysis," IEEE trans. Computer-Aided Design, vol. 9, pp. 352-366, Apr 1990.

[33] A. B. Kahng and S Muddu, "An analytical delay model for RLC interconnects," IEEE trans. Computer-Aided Design, vol. 16, pp. 1507-1514, Dec 1997.

[34] L. N. Dworsky, *Modern Transmission Line Theory and Applications*, New York: Wiley, 1979.

[35] R. Bashirullah, W. Liu, and R. K. Cavin, "Current-mode signaling in deep submicrometer global interconnects," TVLSI, vol. 11, pp. 406-417, Jun 2003.

[36] W. C. Elmore, "The transient response of damped linear networks with particular regard wideband amplifiers," J, Appl. Phys., vol. 19, pp. 55-63, Jan 1948.

[37] L. Zhang, et al., "Simplified delay design guidelines for on-chip global interconnects," GLVLSI, pp. 29-32, Apr 2004.

[38] M. K. Gowan, et al., "Power considerations in the design of the Alpha 21264 microprocessor," DAC, pp. 804-809, 2001.

[39] Y. Massoud et al., "Modeling and analysis of differential signaling for minimizing inductive crosstalk," DAC, pp 804-809, 2001.

[40] X. Huang, et al., "Loop-based interconnect modeling and optimization approach for multigigahertz clock network design," IEEE J. Solid-State Circuits, vol. 38, no. 3, Mar 2003.

[41] M. Beattie and L. Pileggi, "Inductance 101: modeling and extraction," DAC, pp. 323-328, Jun 2001.

[42] J. Jackson, *Classical Electrodynamics, 2nd Edition*, J. Wiley, New York, 1984.

[43] W. Dally and J. Poulton, “Transmitter equalization for 4-Gbps signaling,” IEEE Micro, pp. 47-56, Jan/Feb 1997.

[44] H. Man, “Ambient intelligence: gigascale and nanoscale realities,” ISSCC, pp. 29-35, Feb 2005.

[45] http://www.oea.com/document/metal.pdf

[46] www.ansoft.com/products/si/q3d_extractor

[47] X. Liu, Y. Peng, and M.C. Papaefthymiou, “Practical repeater insertion for low power: What repeater library do we need?” DAC, pp. 30-35, Jun 2004.

[48] A. P. Chandrakasan and R. W. Brodersen, *Low Power Digital CMOS Design*, Kluwer Academic Publishers, Norwell, MA, 1995.

[49] D. Burger and T. M. Austin, “The SimpleScalar tool set, version 2.0,” University of Wisconsin, Madison, Technical Report CS-TR-97-1342, June 1997.

[50] L. Zhang, et al., “Driver pre-emphasis techniques for on-chip global buses,” ISLPED, pp. 186-191, Aug 2005.

[51] L. Zhang, J. Wilson, R. Bashirullah, and P. Franzon, “Differential current-mode signaling for robust and power efficient on-chip global interconnects,” EPEP, pp. 315-318, Oct 2005.

[52] L. Luo, J. Wilson, S. Mick, J. Xu, L. Zhang, and P. Franzon, "3Gb/s AC coupled chip-to-chip Communication using a low swing pulse receiver", Accepted by JSSC.

[53] L. Luo, J. Wilson, S. Mick, J. Xu, L. Zhang, and P. Franzon, "3Gb/s AC-coupled chip-to-chip communication using a low-swing pulse receiver," ISSCC, pp. 522-523, Feb 2005.

[54] R. Bashirullah, W. Liu, R. Cavin, and D. Edwards, “A hybrid current/voltage mode on-chip signaling scheme with adaptive bandwidth capability,” TVLSI, vol. 12, pp. 876-880, Aug 2004.

[55] B. Nikolic, et al., “Improved sense-amplifier-based flip-flog design and measurements,” JSSC, vol. 35, pp. 876-884, Jun 2000.

[56] H. Kawaguchi and T. Sakurai, “A reduced clock-swing flip-flop (RCSFF) for 63% power reduction,” JSSC, vol. 33, pp. 807-811, May 1998.