Abstract

BASHIRULLAH, RIZWAN. Current-mode Band-limited Signaling for Global On-Chip Interconnects (Under the direction of Professor Wentai Liu).

Global on-chip interconnects are a limiting factor in modern high performance VLSI systems due to cross-talk noise, signal delay and wire bandwidth constraints. This dissertation addresses these limitations with a fundamental change in signaling technique – the use of *current-mode band-limited* pulses. This work is intended to establish the theoretical basis for the proposed current-mode and band-limited signaling scheme while formulating its impact on signal delay, bandwidth and cross-talk noise both analytically and experimentally.

A simple yet accurate closed-form delay expression for inverter driven on-chip interconnects with arbitrary receive-end termination is presented. The solution can be used for both resistive and capacitive termination to adequately model current and voltage mode sensing schemes. The model is extended to consider fast input slope and input-to-output capacitance effects of a CMOS inverter. A test chip fabricated in AMI 1.6μm bulk CMOS technology is used to experimentally verify the proposed model. Further analysis shows that the model can be used for sub-micrometer process to accurately estimate delay and bandwidth performance of long on-chip interconnects. Additionally, a new power dissipation model for current-mode signaling is developed to understand the design tradeoffs between current and voltage sensing. Based on the results and derived formulations, a comparison between voltage and current mode repeater insertion for long global deep sub-micron interconnects is presented.

The performance of band-limited pulse signaling for cross-talk noise reduction in high-density on-chip interconnects is addressed. The proposed signaling scheme uses reduced edge-rate pulses to minimize the effect of high frequency components on cross-talk noise. A simple low-overhead adjustable slew-rate driver/repeater is designed to generate the band-limited pulses. In order to assess the various design tradeoffs imposed by the proposed signaling scheme, a comparative analysis is presented to evaluate performance in cross-talk noise, driver/receiver power dissipation and propagation delay.

A test chip with on-chip cross-coupled interconnect test patterns is fabricated in AMI 1.6μm for experimental verification. It is shown that for a coupling-to-signal capacitance ratio of 0.6 and data-rate of 50Mb/s, band-limited signaling yields approximately 72% reduction in peak noise voltage.

This dissertation also explores a novel on-chip bus architecture based on hybrid current/voltage mode repeaters to address signal latency and throughput while minimizing the static power dissipation associated with low-impedance current sensing. The adaptive bandwidth bus (ABB) presented herein is designed to automatically increase or decrease the interconnection bandwidth to track the input data activity envelope, thereby minimizing the overall power dissipation without a penalty in maximum attainable data rate. To demonstrate the feasibility of the proposed adaptive bandwidth bus (ABB) scheme, a preliminary "analog" architecture has been fabricated in AMI 1.6μm bulk CMOS technology. The results and limitations of this design are addressed, and an improved all-digital ABB architecture along with a power dissipation modeling approach based on circuit-level and statistical analysis of microprocessor data streams is presented to evaluate the performance of the proposed bus. Attaining a maximum aggregate bandwidth of 16Gb/s (i.e. 1Gb/s per line) across lossy on-chip interconnects spanning 1.75cm in length, the digital bus core fabricated in TSMC 0.35μm CMOS technology dissipates approximately 93mW with a supply of 2.5V and signal activity of 0.5. Experimental results indicate a reduction in power of 50% over current-mode (CM) sensing, and an improvement in interconnection delay and signaling bandwidth of 35%-70% and 66% over voltage-mode (VM) sensing, respectively.

CURRENT-MODE BAND-LIMITED SIGNALING FOR GLOBAL ON-CHIP INTERCONNECTS

by
RIZWAN BASHIRULLAH

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

ELECTRICAL ENGINEERING

Raleigh, NC

2004

APPROVED BY:

Chair of Advisory Committee

Biography

Born in the coastal town of Cumaná, Venezuela, Rizwan Bashirullah received his B.S.E.E. degree from the University of Central Florida, Orlando, in December of 1997, the M.S.E.E. degree from North Carolina State University (NCSU), Raleigh, in August of 1999 and is his PhD degree in electrical engineering at NCSU in March of 2004. Since 1998, he has served as a Research Assistant in the Department of Electrical and Computer Engineering, NCSU. His research interests include circuit design techniques for high-performance analog/digital VLSI, clock/data recovery circuits and RF bio-telemetry systems.

Acknowledgements

My accomplishments as a graduate student has been to a great extent determined by those of whom I have had the honor and pleasure to work with. My thesis is simply irrelevant when you consider the quality and depth of individuals who have helped shape and train my academic and technical growth. I cannot imagine a better environment to learn, exchange ideas and exercise creativity in its purest of forms. I have always believed that the people behind any undertaking are important; in this case, it has not only proven to be true but also exceeded all possible expectations.

I sincerely thank my PhD advisor, Dr. Wentai Liu, who has always supported me and allowed me to explore new ideas. His guidance, patience and encouragement throughout my graduate studies are a testament to his character as an advisor and a teacher. He has an amazing ability to tailor his inter-personal relationships with his students to maximize productivity. At a personal level, he never once hesitated to help me and for that I will always be indebted to him. I am also grateful to my MS advisor, Dr. Amir Mortazawi, who has taught me valuable lessons in research and problem solving. I am deeply honored and thankful to them for allowing me to be a part of their research groups.

To my committee members, I thank them for being a part of my graduate study experience. Gracefulness comes to my mind when I think of Dr. Ralph Cavin – he has been an integral part of my research. I wish to extend my gratitude to Dr. Paul Franzon, for always being accommodating and understating. To Dr. Gianluca Lazzi, possibly one of the nicest professors I have met, I am thankful for our conversations and his guidance regarding academia.

My brief association with Semiconductor Research Corporation (SRC) has been one of the most fulfilling professional experiences. My involvement with Techon and numerous design review meetings has allowed me to present my work and exchange ideas on a one-to-one basis with top researchers from other institutions. I am grateful to Dale Edwards and Dr. Cavin for being a part of this research and always allocating time to review our work. Dale has been the best of industry liaisons one can hope for – he has introduced me to numerous industry managers and professors and always been accommodating in conferences and meetings.

To my colleagues at EGRC, I am indebted to you for making my experience an enjoyable one. To Steve Lipa, who has been unconditionally supportive of my work, I am grateful to him for helping me in uncountable occasions to accomplish my work. His patience and experience during laboratory testing was invaluable. I am grateful to Kasin Vichienchom, a mentor in research and the way of all things with patience, for our numerous conversations regarding circuit design. I am appreciative of Mark Clements, the system level guy, who has taught me to approach problems from an architectural perspective. John Wilson, who has been supportive and helpful in many ways, has always been willing to discuss and share ideas. To the rest of the gang, Musto, Jeff, Rajeev, Chris, Liang, Mehmet, Alper, Guoxing, Ying, Toby, Alan, David, David, Jayesh, Praveen, Stephen, John; thanks for a great experience.

I would like to extend my gratitude to Sandra Bronson and Michele Joyner for putting up with me and helping me with all the paper work.

My fiancée, Ayşegül, whom I met while pursuing my graduate studies, has made my years in school all worthwhile the wait. I always knew I would meet her while pursuing my degree, which explains why I took my sweet time to finish.

To my family, I thank them for their unconditional love. I thank my father, Dr. Bashirullah, who has set the highest of standards for me to achieve. I am thankful to my brother, Dr. Bashirullah, for having someone to look up to. To my mother, the business woman and the wealthiest of us all, who never understood the point of working so hard; I thank her for keeping my feet grounded and reminding me that health and happiness are the only important things in life to strive for.

This work was supported in part by the National Science Foundation and by Semiconductor Research Corporation via Advanced Micro Devices (AMD) under award 983.001.

Table of Contents

List of Figures

List of Tables

Chapter I

Introduction

1.1. Motivation

Long global on-chip interconnects has been perceived as a major bottleneck in high-performance VLSI systems due to an increase in signal propagation delays relative to gate delays, increasing effects of cross-talk noise due to wire coupling capacitance, and signaling bandwidth limitations due to diffusive and dispersive effects of signals traveling over RC dominated wires. Therefore, as CMOS continues to scale in the deep sub-micron regime, increasing performance measures such as higher clock rates will become more difficult to attain. Assuming an approximate 30% reduction in feature size and 0.7X scaling of interconnects for each technology node, the approximate increase in wire RC delay is 70% per technology node [1]. Materials and process approaches to minimizing relative delay and cross-talk are yielding smaller incremental performance improvements with successive technology generations of the ITRS (The International Roadmap for Semiconductors) [2]. Moreover, the increase in interconnect aspect ratio and reduction of metal pitch with each technology generation increases the ratio of coupling-to-total capacitance, progressively exacerbating the signal integrity problem.

Process scaling also allows for a higher level of system complexity and integration. Raising the level of integration on a chip brings about a larger number of on-chip interconnects. Of particular interest are the global interconnects used to connect individual processing units, which can traverse the chip edge and thus be in the centimeter range lengths. Although the probability of occurrence of global interconnects is much lower than local interconnects [3], the increased propagation delays between processing units is likely to limit the throughput performance of high-end processors.

The most popular solution to mitigate the increase in delay and cross-talk and improve signaling bandwidth has been the systematic insertion of repeaters in long global interconnects [4]. The underlying assumption of repeater insertion is that a chip is typically wire limited, resulting in some unutilized silicon area that can be used to accommodate the repeaters [5]. Often, however, optimal repeater insertion cannot be achieved due to underlying critical processing units, which can result in increased signal delays. Furthermore, as the optimal repeater insertion distance decreases with each technology node due to increased interconnect resistive effects, the overall improvement in delay can be undermined by the exponential increase in the number of repeaters on-die and associated driver/repeater power dissipation [6].

This dissertation explores alternative solutions to address the propagation delay, signaling bandwidth and cross-talk noise limitations in deep sub-micrometer global interconnect by introducing a fundamental change in the signaling approach: the use of *current-mode* and *band-limited signaling*. This work is intended to establish the theoretical basis for the proposed current-mode and band-limited signaling scheme and analytically formulate its impact on signal delay, bandwidth and cross-talk noise reduction. Study of circuit design issues is followed by prototype fabrication for conceptual and experimental verification.

1.2. Thesis Overview

A literature review of global interconnects technology trends and associated signaling solutions are presented in chapter II. Based on projections made by the ITRS, future global interconnect trends are described while focusing on signal delay propagation, power dissipation and cross-talk noise. The effect of copper and low k dielectrics, delay degradation and coupling capacitance due to adjacent wires, die size model and cycles time across chip is discussed in more detail. Signaling schemes and existing solutions to problems associated with global interconnects are examined.

In Chapter III, the theoretical basis for the proposed *current-mode signaling* is presented in detail. Propagation delay and power dissipation are addressed for current/voltage mode signaling in deep sub-micrometer global interconnects. Based on the effective lumped element resistance and capacitance approximation of distributed RC lines, simple yet accurate closed-form expressions of delay and power dissipation are presented. The usefulness of these solutions is that a resistive load termination for current mode signaling is accurately modeled. Measurement results are provided to verify the model accuracy. The test setup and precautions for the implementation of fast edge-rates are also discussed. Based on the results and derived formulations, a comparison between voltage and current mode repeater insertion for long global deep sub-micron interconnects is presented.

In chapter IV, a *band-limited signaling* scheme to minimize signal cross-talk noise for global on-chip interconnects is presented. The fundamental difference between this technique and other solutions is that band-limited signaling minimizes the source of cross-talk noise by eliminating fast edge rates pulses within critical interconnections links. The effect of signal edge rate reduction on delay and power dissipation is analyzed to address the potential disadvantages of the proposed technique. Theoretical expressions are derived to evaluate the cross-talk noise waveforms due to an aggressor line operated in both current and voltage modes. Circuit implementation issues of a simple band-limited repeater are discussed and experimentally verified to evaluate the effect of band-limited pulses on cross-talk noise, delay and power.

In chapter V, a novel on-chip *adaptive bandwidth bus* architecture based on hybrid current/voltage mode repeaters is explored to address signal latency and throughput while minimizing the static power dissipation associated with low-impedance current sensing. The adaptive bandwidth bus presented in chapter V is designed to automatically increase or decrease the interconnection bandwidth to track the input data activity envelope, thereby minimizing the overall power dissipation without a penalty in maximum attainable data rate. In addition, a power dissipation modeling approach based on circuit-level and statistical analysis of data streams is presented to evaluate the performance of the proposed bus. Circuit implementation details and measured comparative results of an analog and digital implementation of the adaptive bandwidth technique are presented with supporting experimental results.

Concluding remarks of the proposed techniques and related areas of future work are presented in chapter VI.

1.3. References

[1] A.K. Stamper, M. Fuselier, X. Tian, "Advanced Wiring RC Delay Issues for sub-0.25-micron Generation CMOS," International Interconnect Technology Conference, pp 62-64, 1998.

[2] Semiconductor Industry Association, *The International Technology Roadmap for Semiconductors*, 2001 Edition.

[3] W. E. Donath, "Wire-length distribution for placements on computer logic," IBM J. Research and Development, vol. 2, no. 3, pp. 152-155, May 1981.

[4] H.B. Bakoglu, Circuits, Interconnections, and Packaging for VLSI. Reading, MA: Addison-Wesley, 1990.

[5] R. Venkatesan, J.A. Davis, K. Bowman, J. Meindl, "Optimal repeater insertion for n-tier multilevel interconnect architectures," International Interconnect Technology Conference, pp. 132-134, 2000.

[6] R. Krishnamurthy, A. Alvandpour, V. De, S. Borkar "High-performance and Low Power Challenges for Sub-70nm Microprocessor Circuits," Custom Integrated Circuits Conference, pp. 125-128, 2002.

Chapter II

Global Interconnect Trends and On-chip Signaling Techniques

2.1. Introduction

An increasing attention is placed on the design of on-chip interconnects due to the dominant limitation of global interconnect signal delays, power dissipation and wire coupling cross-talk noise on overall system performance. Therefore it is imperative that future on-chip interconnect designs meet the increasing transmission speeds, reliability and signal integrity needs despite the aggressive scaling trend of CMOS technology.

It is clear that global interconnect performance required for future generations of integrated circuit technology will need to improve in the long-term via fundamental changes in material technology. The use of copper and low-*k* interconnects has helped mitigate some of the problems associated with signal distribution, however further changes will be required. Architectural changes and new design approaches to high-performance systems such as microprocessors will also be required. For instance, to overcome the bandwidth limitations in the deep submicron regime, new approaches to high-speed signaling are needed to meet the increasing clock frequency projected by the ITRS [1].

In this chapter, a review of global interconnects scaling trends and associated signaling solutions are presented. The literature review on scaling trends and the various signaling schemes proposed to date will focus on delay, throughput, power and signal-integrity issues instead of process fabrication related aspects and challenges. A performance comparison amongst several signaling approaches will be presented in an effort to identify design tradeoffs and associated circuit complexity.

2.2. Global Interconnect trends

As the complexity of high-performance systems continues to increase, the need for substantial planning prior to the circuit implementation becomes evermore critical. In particular, the process of design floor-planning to identify critical interconnect paths during the micro-architectural definition is crucial to meeting performance goals. Critical interconnection paths are often determined by global interconnects which can span lengths that transverse the entire chip edge, connecting circuit blocks within the chip. At a system-level the number of such global paths constitute only a fraction of the overall wire-length distribution [2]-[4], as observed from the bi-modal wire-length behavior depicted in Figure 2.1. The occurrence of these global wires, however, often limit the maximum operating frequency and throughput due to delay latency and bandwidth bottlenecks.

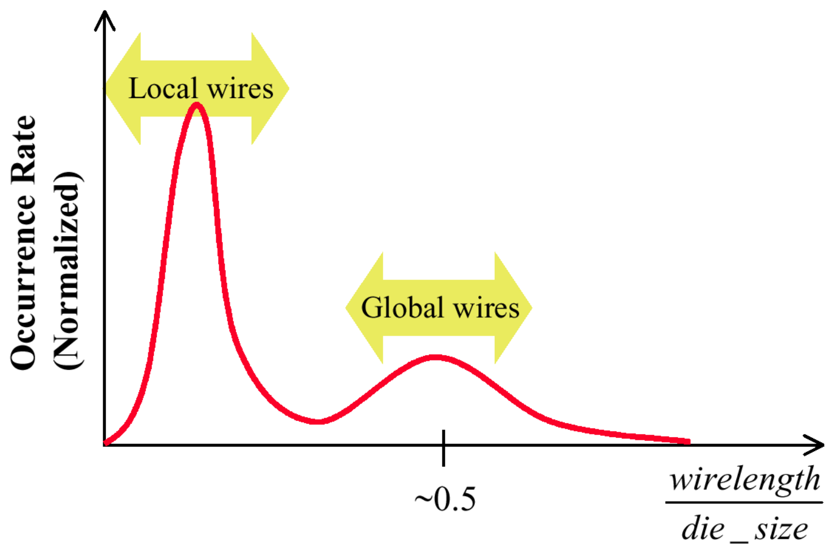


Figure 2. 1. Typical bi-modal wire length distribution. The first peak represents the occurrence rate of local interconnects used for wiring of internal cells, whereas the second peak represents the occurrence rate of global wiring.

To adequately describe the overall impact of global interconnects on the performance of future technology nodes, it is important to consider the scaling scenario of global interconnects as well as projections of system level and intrinsic device parameters. A summary of scaling projections for short-term (2001-2007) and long-term (2010-2016) goals is extracted from the ITRS 2001 [1], as shown in Tables 2.1, 2.2 and 2.3.

Table 2.1 summarizes the scaling trends of global interconnects which determine physical as well as electrical properties of global wires. Table 2.2 highlights the projected trends of system-level scaling such as operating frequency, power dissipation and number I/O pins. Table 2.3 summarizes the intrinsic device properties such as saturation current, leakage current, parasitic resistance and capacitance and device time constants. The discussions and projected results are thus based on parameters extracted from these tables.

Table 2. 1. Global Interconnect parameters.

Year	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2010**	**2013**	**2016**
Technology Node (nm)	130	115	100	90	80	70	65	45	32	22
Chip Size (mm^2)	310	310	310	310	310	310	310	310	310	310
Number of Metal Levels	8	8	8	9	10	10	10	10	11	11
Jmax (A/cm_2) –wire	5.8E5	7.1E5	8.0E5	9.6E5	1.1E6	1.3E6	1.4E6	2.1E6	3.7E6	4.6E6
Imax (mA) – via	0.36	0.36	0.33	0.32	0.29	0.27	0.24	0.18	0.16	0.11
Minimum Pitch (nm)	670	565	475	460	360	320	290	205	140	100
Wire aspect ratio(copper)	2	2	2.1	2.1	2.2	2.2	2.2	2.3	2.4	2.5
Via aspect ratio (copper)	1.8	1.8	1.9	1.9	2	2	2	2.1	2.2	2.3
Copper Effective Resistivity (μm-Ω)	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2
ILD dielectric (εr)	3-3.6	3-3.6	3-3.6	2.6-3.1	2.6-3.1	2.6-3.1	2.3-2.7	2.1	1.9	1.8

Table 2. 2. System level parameters

Year	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2010**	**2013**	**2016**
Generation (um)=DRAM 1/2pitch	130	115	100	90	80	70	65	45	32	22
Devices (M/cm^2) (including SRAM)	89	112	142	178	225	283	357	714	1427	2854
Chip Size (mm^2)	310	310	310	310	310	310	310	310	310	310
Signals (from TEST section)	1024	1024	1024	1024	1024	1024	1024	1280	1408	1472
Pins	1700	1870	2057	2263	2489	2738	3012	4009	5335	7100
Clock (MHz) = on-chip clock (max)	1700	2320	3090	3990	5170	5630	6740	12000	19000	29000
IO (MHz) = Chip-to-Board Peripheral Buses	1700	1870	2057	2262	2488	2737	3011	4009	5339	7100
Wiring Levels	8	8	8	9	10	10	10	10	11	11
Vdd (V)	1.1	1	1	1	0.9	0.9	0.7	0.6	0.5	0.4
Power (W)	130	140	150	160	170	180	190	218	250.7	288.3

Table 2. 3. Intrinsic device parameters.

Year	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2010**	**2013**	**2016**
Generation (um)=DRAM 1/2pitch	130	115	100	90	80	70	65	45	32	22
T_{ox} Electrical equivalent (nm)	2.3	2.1	2	2	1.9	1.9	1.4	1.2	1	0.9
Nominal V_{dd} (V)	1.2	1.1	1	1	0.9	0.9	0.7	0.6	0.5	0.4
Nom. High Performance (HP) NMOS sub-threshold leakage current, I_{sdleak} (at 25C) (μA/μm)	0.01	0.03	0.07	0.1	0.3	0.7	1	3	7	10
Nominal HP sat. current I_{dd} (at 25C) (μA/μm)	900	900	900	900	900	900	900	1200	1500	1500

Table 2.3. (continued)

Year	2001	2002	2003	2004	2005	2006	2007	2010	2013	2016
Parasitic source+drain resistance (R_{SD}) (μΩ-μm)	190	180	180	180	180	170	140	110	90	80
Parasitic source+drain resistance (R_{SD}), percent of ideal channel resistance (V_{dd}/I_{dd})	16%	16%	17%	18%	19%	19%	20%	25%	30%	35%
HP NMOS device $\tau = C_{gate}(V_{dd}/I_{dd})$(ps) C_{gate} includes overlap/fringing capacitances	1.6	1.3	1.1	0.99	0.83	0.76	0.68	0.39	0.22	0.15
Parasitic capacitance; % of ideal gate cap. (This capacitance is included in τ calculation)	19%	22%	24%	27%	29%	32%	27%	31%	36%	42%

2.2.1. Delay and Cycles Across Chip

As shown in Table 2.1, it is clear that the implementation of copper and low *k* materials will minimize the impact of scaling on wiring delay. An approximate delay formula based on cross-sectional resistance and parallel plate capacitance formulations can be derived in terms of parameters from Table 2.1, expressed as,

$$\frac{RC}{l^2} = \frac{(\varepsilon_o \rho)\varepsilon_r}{\left(\frac{A}{R}\right)_{wire}\left(\frac{A}{R}\right)_{via}\left(\frac{P_m}{2}\right)^2}\left(1 + n\frac{\left(\frac{A}{R}\right)_{wire}\left(\frac{A}{R}\right)_{via}\left(\frac{P_m}{2}\right)^2}{w \cdot s}\right) \tag{2-1}$$

where n is the number of adjacent wires at a metal spacing s. From Eq. (2-1), a projected 67% decrease in wiring delay due the combined effect of copper/low-k materials is predicted, as shown in Figure 2.2a. However, the overall increase in propagation delay over future technology nodes is unavoidable. For instance, the wiring delay of a 1-cm long global unrepeated interconnect is projected to increase from 160-ps in 2001 to over 2.4-ns in 2016, a 1500% increase[1]. Notice that the increase in delay is independent of the interconnection width, at least to a first order approximation.

Signal delay performance will also worsen due to the increasing loading of adjacent interconnect lines, as shown in Figure 2.2b. Assuming that the metal width and spacing are equal, an approximate 10x increase in metal width is required to minimize the delay

[1] These projections are based on global interconnect parameters of Table 2.1, extracted from the 2001 ITRS roadmap. A simple parallel plate capacitance formulation is used to derive the capacitance of the line and the effect of driver resistance and far-end loading has been neglected.

due to the projected limits of a single wire shown in Figure 2.2a. The use of wide lines or ‘fat’ wires, can minimize the effect of adjacent interconnects on delay at the expense of a significant area penalty.

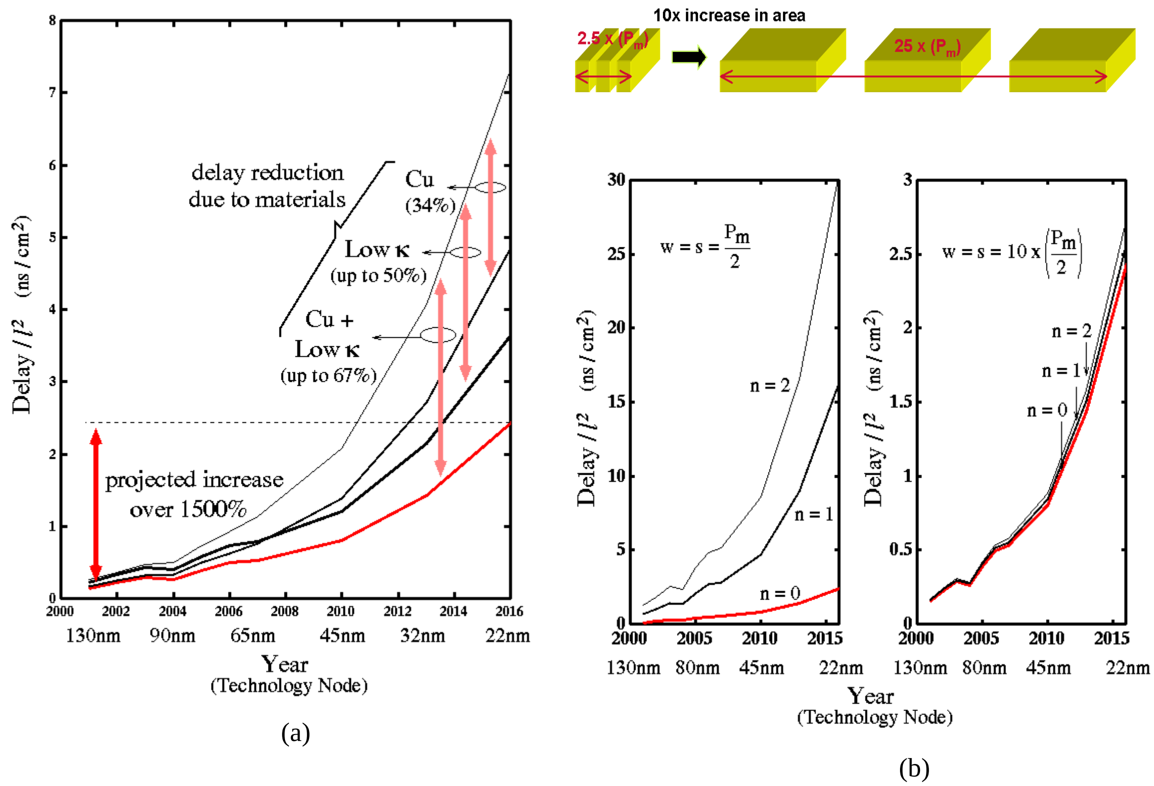


Figure 2. 2. Projected delay performance over future technology nodes (a) combine effect of copper and low k on delay performance. (b) Effect of adjacent wires on delay.

An important performance metric that is closely related to the increase in global propagation delays is the number of clock cycles required to transverse the chip edge, which depends on the projected clock frequency and chip size model. Interestingly, the die size of high performance processors is projected to saturate at 310mm^2 due to the increase in cost and power dissipation. According to the 2001 ITRS, additional logic content will not be efficiently usable due to package power dissipation limitations, and additional memory content (i.e. larger caches, more levels of memory integrated on-chip) will not be cost-effective beyond a certain point. Based on a constant chip size assumption, the number of clock cycles to transverse a chip edge for an unrepeated wire is projected to increase from 2 cycles in 2001 to 220 cycles in 2016, as illustrated in Figure 2.3a (this assumes only interconnect RC delay). A more realistic perspective of chip size features and cycles across chip is shown for the Alpha Microprocessors in Figure 2.3b and 2.3c [5]. Notice that the chip size of the EV8 is greater than $400mm^2$, well above the $310mm^2$ constant chip size model projected by the ITRS 2001 [1]. In addition, the cycles across chip have steadily increased, with 7 cycles for the latest EV8 design operating at nearly 1.5GHz.

It is evident that the need for repeater insertion and other high-speed on-chip signaling techniques will be fundamental to cope with the continued scaling trends. However the increase in the number of repeaters on-die and associated driver/repeater power dissipation and placement constraints due to underlying logic will pose even greater challenges. A more detail discussion on interconnect performance with repeaters is provided in section 2.2.3.

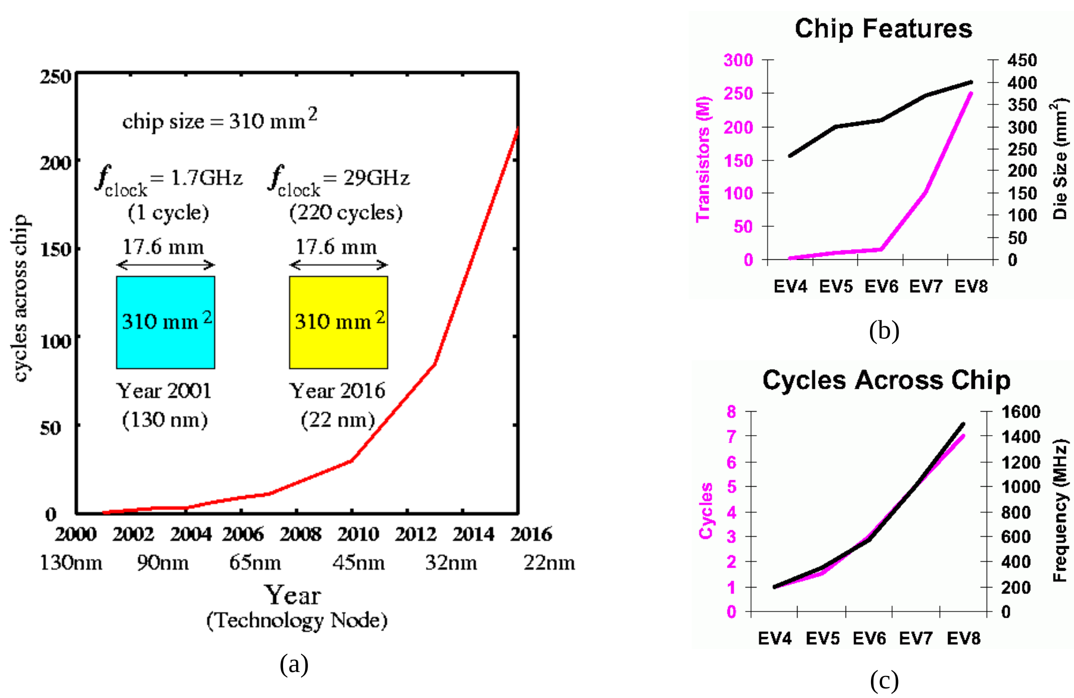


Figure 2. 3. Chip size model and cycles across chip. (a) Projected cycles across chip for future technology nodes for a constant chip-size model. (b) Alpha MPU die sizes and (c) clock frequency and cycles across chip [5].

2.2.2. Cross-talk Noise

Coupling capacitance between adjacent metal interconnects actually scales down with technology node despite the aggressive decrease in metal pitches, an effect due to the reduction in inter-layer dielectric (IMD) constant. However, the slower scaling in metal thickness to maintain the line resistance from increasing rapidly, results in higher aspect ratios (A/R), which combined with increasing global interconnect lengths ultimately causes cross-talk noise to increase. Cross-talk noise on a victim line from adjacent interconnects can cause fault detection errors as well an increase in propagation delay. The increase in delay due to simultaneous opposite transitions of adjacent wires causes an increase in effective capacitance seen by the victim line. This effect has been reported to degrade the delay performance by 30% [6].

To a first order approximation, the ratio of coupling capacitance (C_C) to capacitance to ground (C_V) can be approximated as $C_C/C_V=(A/R)_{wire}(A/R)_{via}$ – where it has been assumed that the metal $w=s=P_m/2$; w=width, s=spacing, P_m=minimum pitch. This simple relationship indicates a 60% increase in C_C/C_V over future technology generations, as listed in Table 2.1. This trend, combined with faster switching edge-rates and reduced noise margins due to threshold voltage scaling will undoubtedly pose a significant design challenge. An example of the coupled noise effect is reported in [7] for the IBM S/390 G4 300-400MHz microprocessor. The G4 CPU is 17.35x17.3 mm^2 in size and was designed in a 0.2μm L_{eff} CMOS 2.5V process with a total of 7.8 million transistors [8]. A histogram of the noise voltage for 20095 relevant nets is shown in Figure 2.4a. The allocated noise margin is in the order of 500mV-600mV. Figure 2.4a shows a long tail of nets that exceed the noise margin, with 476 and 157 nets exceeding 0.8 and 1V, respectively. Figure 2.4b shows a histogram of the ratio of C_C/C_V. Notice that the majority of the nets have a C_C/C_V ratio under 0.5, which is clearly much smaller that the current projected trends for future technology nodes. Thus, coupling noise is likely to pose bigger challenges in future technology generations.

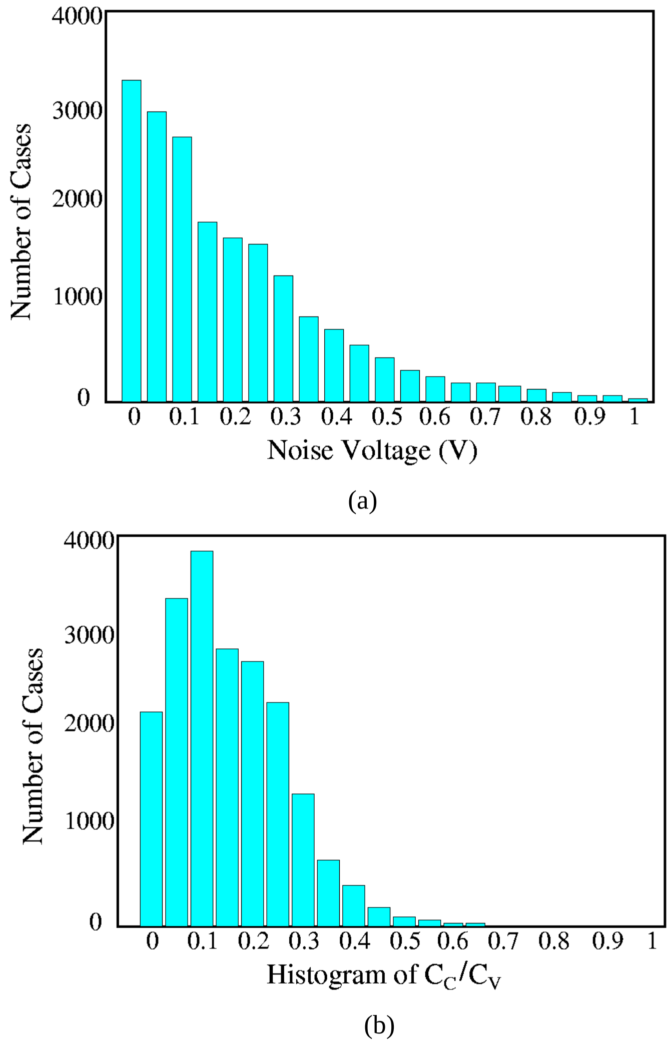
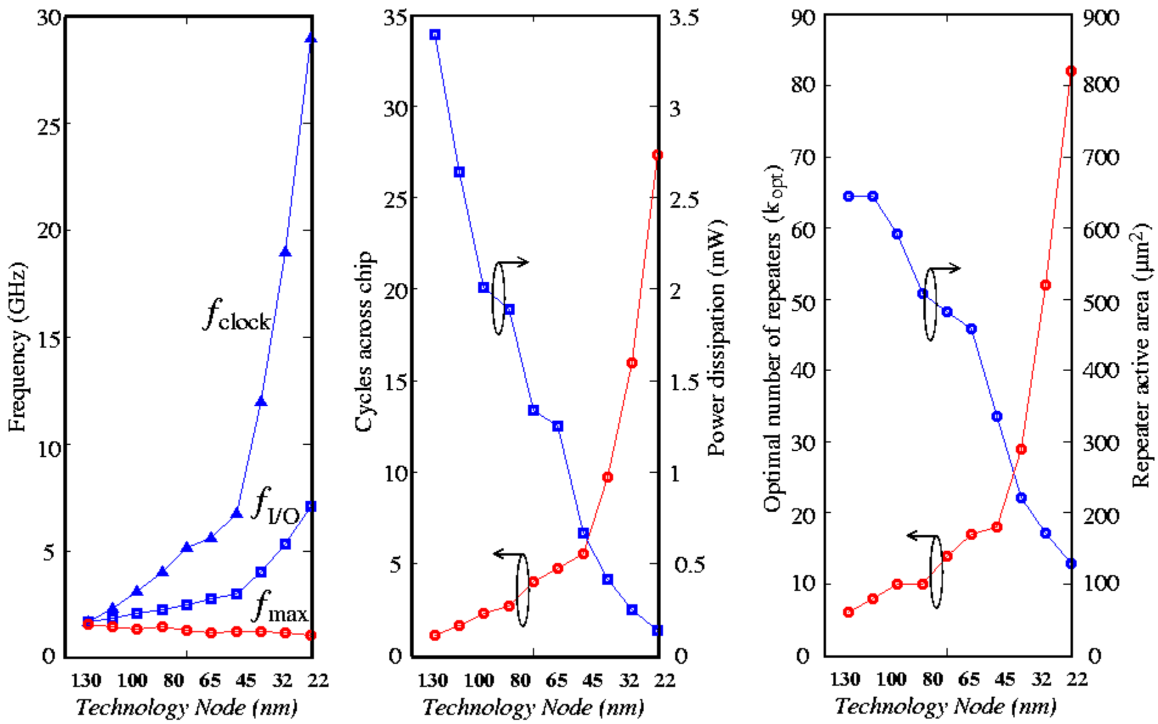


Figure 2. 4. Histogram for noise cross-talk analysis for the IBM S/390 Microprocessor (a) Noise voltage (b) ratio of coupling-to-ground capacitance (C_C/C_V).

2.3.3. Interconnect performance with optimal repeaters

It is estimated that the number of repeaters required for the 70nm node will exceed 1.6million [9], [10]. Although projections of this nature are important, estimating the repeater count does not reveal some of the fundamental limiting trends for signaling across long global interconnections. Historically, optimal repeater insertion techniques have been used to minimize delay [11], often a determining factor for the system cycle time. However, as the projected operating frequencies continue to increase, attaining global bus speeds that are a mere fraction of the system frequency will become increasingly difficult.



(a) (b) (c)

Figure 2. 5. Global interconnect with optimal repeater insertion for future technology nodes. (a) Maximum frequency of an optimally repeated interconnect, (b) cycles across chip and power dissipation and (c) optimal number of repeaters and associated active area. All projections assume w=s=P_m, two adjacent lines, l_{max}=$(310)^{1/2}$ mm. These projections are for a single global interconnect with scaling parameters extracted from Table 2.1-2.3.

To better illustrate the speed and propagation delay limitations of on-chip global interconnects, projected trends using optimal repeater insertion for global signaling is shown in Figure 2.5. The results are based on device, interconnect and system parameters from Table 2.1, 2.2 and 2.3. It is assumed that the length of the global line spans one chip edge of the constant chip size model of 310mm^2 (i.e. l_{max} =$310^{1/2}$mm or 1.76cm). In addition, the metal width and spacing from adjacent lines are defined as w=s=2P_m (where P_m is the minimum pitch from Table 2.1). In Figure 2.5a, the maximum data frequency of the global line (f_{max}) constitutes a small fraction of the projected clock and I/O frequencies. It should be noted that increasing the metal spacing and the width of the line only gives small incremental results as long as the line remains RC dominated. In Figure 2.5b, the total number of cycles is shown to increase across generations, whereas the power dissipation decreases. The later is due to decreasing capacitive loading of repeaters and interconnect segments. It is to be understood that the overall interconnect power dissipation is likely to increase in future generations due an increasing transistor, repeater and interconnect count. In Figure 2.5c, the optimal number of repeaters required for minimizing delay is shown to increase with each generation, as expected. The total active area of the optimally sized repeaters decreases due to the reduction of feature size.

As depicted in Figure 2.5, attaining higher performance global signaling will require a combination of the following solutions; 1) an above optimal number of repeaters to maximize speed, 2) better performance on-chip global tracks, 3) high-speed on-chip signaling schemes. All of the above mentioned requirements are ultimately limited by the power, heat removal, placement blockage and the ubiquitous interconnection limited chip constraints.

2.3. On-chip Signaling Techniques

On-chip signaling can be defined as the method to transmit information from one location to the other within the chip [12]. It is important to differentiate on-chip signaling from off-chip signaling techniques used in chip-to-chip, cable or optical links. When communicating within a chip, interconnects typically do not present controlled characteristic impedances and reflections are often not a problem because the wires are RC dominated. In the RC regime, the signal response is damped due to slow diffusive propagation.

There are instances, however, in which the line impedances of on-chip interconnects are optimized for RLC transmission line behavior. Signal propagation in the high-frequency RLC regime can allow for electromagnetic wave transmission as supposed to the slow diffusive signal propagation. For instance, clock distribution within the chip requires well-controlled transmission lines to minimize frequency dependent attenuation or dispersion and prevent unwanted reflections to control the maximum clock edge-rate and skew throughout the entire chip [13], [14]. On the other hand, the interconnection wiring of a wide bus that communicates to and amongst the I-Cache and D-Cache, for example, is more likely to exhibit dominant RC behavior.

The discussions presented herein are valid for any type of transmission line interconnections. However, it is important to keep in mind that long on-chip interconnects are typically not optimized for electromagnetic signal propagation, mainly due to area overhead of optimized transmission line structures (see section 2.3.6). On-chip interconnects, for the most part, are predominantly resistive and capacitive and exhibit a quadratic relationship with line length.

2.3.1. Performance Metrics

The solutions to signaling over on-chip interconnects described in the literature focuses on achieving certain performance metrics which can be summarized as follows:

1) **Delay Latency:** this refers to the propagation delay, often defined as the 50% crossover delay of the input/output waveforms. Delay latency is crucial in microprocessor design because it determines the critical paths of the design, often resulting in stall cycles. If pipelining is not possible due to area or added system complexity, delay latency can also limit the maximum operating frequency.
2) **Power:** power dissipation of on-chip interconnects has been reported to be 20% or more of the total chip power consumption in CMOS VLSI's [15], reaching tens of watts.
3) **Data throughput:** unlike delay latency, data throughput refers to the maximum data rate that can be supported by on-chip interconnects. The internal bus of a microprocessor often operates at a much lower data rate then the system clock due to the RC delays of on-chip interconnects.
4) **Noise:** refers to any deviation of the desired signal from its intended value [12]. In this work, the noise induced in on-chip interconnects will be limited to cross-talk noise from adjacent lines.
5) **Area:** the physical implementation of on-chip signaling solutions may result in added interconnection area (i.e. differential interconnects as supposed to single-end lines) and/or overhead active area due to associated design complexity of circuits. In addition, placement blockages due to underlying circuits can play an important factor in determining the feasibility of physical implementation.

2.3.2. Repeater Insertion

Most high-performance microprocessors require the insertion of repeaters at a global wiring level to mitigate interconnection delays, an increasingly challenging task given a 0.7X reverse-interconnect scaling trend per generation, a 14% increase in die size, and doubling of clock operating frequency per technology node [16]. The main advantage of inserting uniformly spaced and identical repeaters, dividing a long interconnect line into smaller subsections, is the reduction in the quadratic relationship of RC delay on line length to a less severe linear dependence [11], [17]. As depicted in Figure 2.6, inserting k uniformly spaced repeaters can reduce the Elmore delay approximation [18], [19] of an interconnection with total resistance (capacitance) of R_T (C_T) from $0.5R_TC_T$ to $0.5R_TC_T/k$. The additional delay due to the buffers can be taken into account to a first order approximation by [11]:

$$t_{50\%} = k\left[0.7R_o\left(\frac{C_T}{k} + C_o\right) + \frac{R_T}{k}\left(0.4\frac{C_T}{k} + 0.7C_o\right)\right] \qquad (2\text{-}2)$$

where R_o and C_o are the input capacitance and output resistance of the repeater. The effectiveness of repeater insertion in reducing the interconnection RC delays has proved invaluable in modern day process technology and remains the most popular technique used in high performance processors [20], [21]. As such, delay modeling for repeater optimization in on-chip interconnects has been extensively addressed in the literature [22]-[25].

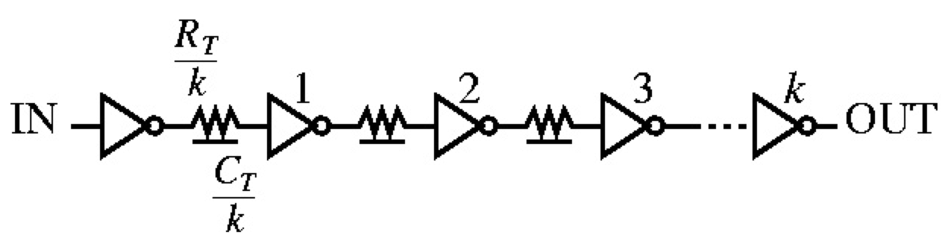


Figure 2. 6. Uniform repeater insertion methodology.

Repeaters can also be used to increase the maximal data rate of on-chip interconnects [26]. In [27]-[28], a methodology for increasing the data transfer rate on long interconnection wires is proposed. It consists of inserting latched repeaters to form synchronous pipelines, similar to a FIFO (first-in first-out). Figure 2.7a shows a block diagram of the repeater. In the pipeline, data propagates from one repeater stage to the other as long as the next storage repeater stage is available. If the last repeater nearest to the receiver-end cannot deliver its data, the current data is buffered and appropriate control signal is sent to the preceding stage indicating that it will stop receiving data at the next cycle [26]. Thus, the congestion and decongestion operations ripple from one stage to the other each cycle. A similar idea coined as "elastic interconnects" was also reported in [29]. The elastic interconnect circuitry is shown in Figure 2.7b. Its main feature is the use of tri-state repeaters capable of sampling and maintaining data line voltage levels. A dedicated congestion control line is used to hold data if the pipeline is full or decongest the line otherwise. It should be pointed out that the implementation in [29] aims to relieve the latency in multi-processor tile systems, instead of maximizing the data throughput as in [27]-[28].

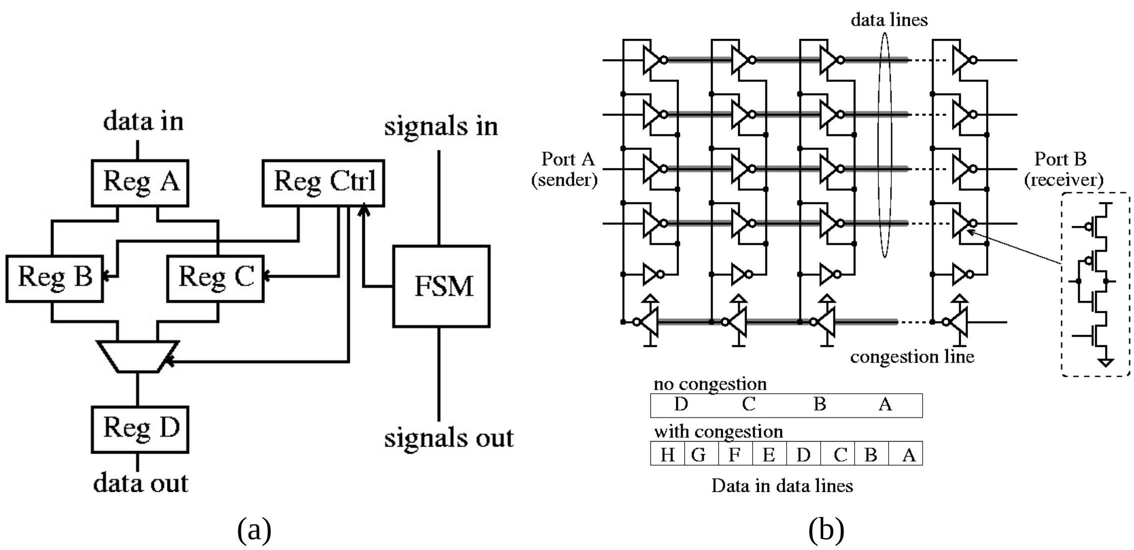


Figure 2. 7. Repeater insertion with data storage capabilities (a) Synchronous pipeline based on repeater insertion [27], [28]. (b) Elastic interconnects for multiple tile processor units [29].

Another repeater insertion technique used to reduce propagation delay is based on regenerative repeaters [30]-[33]. Unlike typical repeater insertion techniques where an interconnect line is divided into smaller sections and buffered, the parallel regeneration scheme uses direction-independent buffering repeaters by sensing the beginning of a transition and subsequently enforcing it. Thus, this type of repeaters is ideal for bi-directional buses. Regenerative repeaters can be classified into dynamic or static, as shown in Figure 2.8 [32]. In the dynamic regenerative repeaters (Figure 2.8a-2.8c), a gated clock is used to pre-charge the line to the supply voltage. After pre-charge, in the event of a falling transition due to a logical high at the input of the line driver, which can be implemented with a single NMOS pull-down driver, the regenerative repeaters sense the beginning of the transition and locally provides a pull-down path. In the static implementation (Figure 2.8d), the pre-charge transistor is removed and replaced by the complement of the pull-down circuitry.

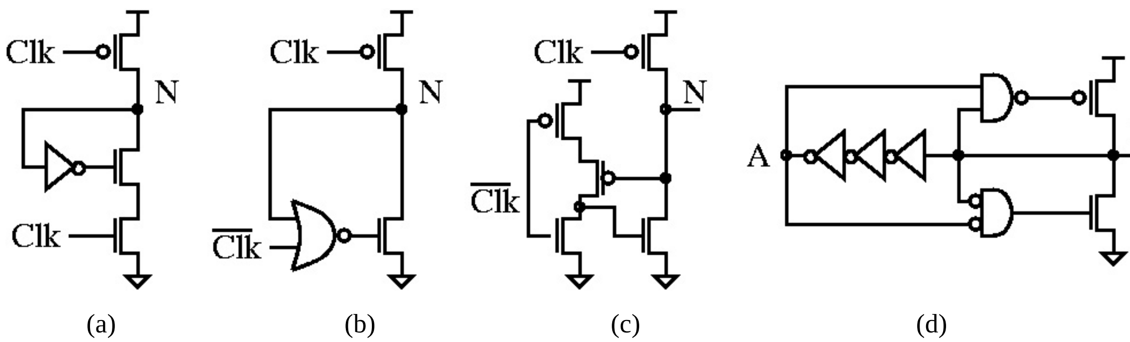


Figure 2. 8. Dynamic and static regenerative repeaters [32] (a)-(c) dynamic (d) static.

Another static implementation of a regenerative amplifier is shown in Figure 2.9 [33]. The transient sensitive trigger (TST) block senses early transitions on the line activating the acceleration transistors to locally speed up the transition. In steady state, both pull-up and pull-down paths are turned off so that the line is not actively driven. The transient sensitive accelerator in [33] also achieves higher cross-talk immunity due to the clamp transistors.

In summary, various repeater insertion techniques were presented. Repeaters can be optimized for minimal delay and/or maximum data throughput. Typically, higher throughput is achieved at the expense of larger latencies due to the intrinsic delay contribution of an increasing number of repeaters. Repeaters for delay optimization can be implemented either by segmenting the interconnect line and inserting buffers or by using a parallel regeneration scheme. Since regenerative repeaters are not inserted in the signal path, it can be used in bi-directional signaling schemes.

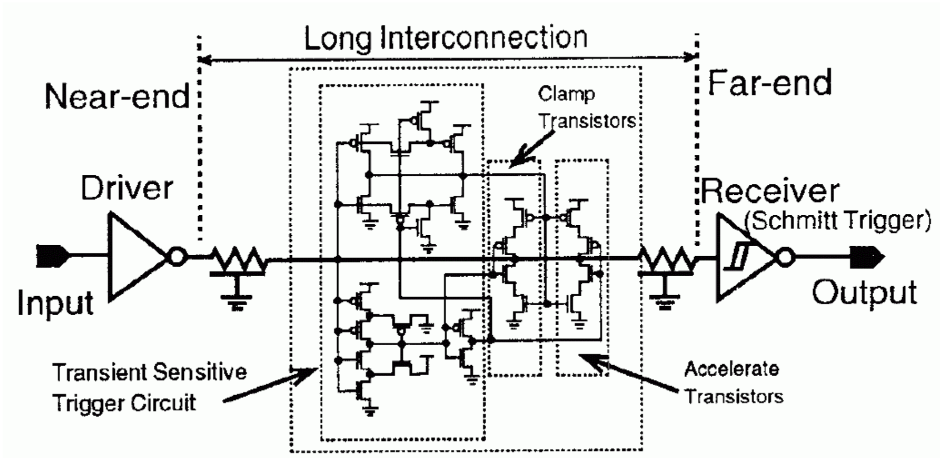


Figure 2. 9. Transient sensitive accelerator regenerative repeater [33].

2.3.3. Low swing bus

One of the most efficient ways of decreasing the power consumption in on-chip interconnects is to reduce the operating voltage [34]. However, unless the devices scale proportionally, simply reducing the supply voltage will cause speed degradation and can offset the benefits in power savings. One possible way to achieving low operating voltage while minimizing its effect on propagation delay is to reduce the signal swing only in the internal bus [35], [36]. The proposed architecture for internal bus swing reduction in [35] is shown in Figure 2.10a. The output of the bus driver is a reduced swing of V_{CC}/n, which is restored back to V_{CC} by the receiver. An internal supply generator is used to create voltages V_{CL} and V_{SL}, which are fed to the source-offset bus driver circuit, as shown in Figure 2.10b and 10c. Since low-Vt devices are used for both MOSFET's in the bus driver, and the input is operated with a signal swing of V_{CC}, the overdrive voltage for both on and off state remains the same as the conventional CMOS driver.

The receiver circuit consists of a symmetric level converter, which restores the output voltage to full swing, as shown in Figure 10c. However, due to the reduced output current drive of the level converter, the bus receiver exhibits added conversion delay. This internal reduced swing bus architecture has been reported to achieve power savings of approximately 1/3 for a 100MHz operating frequency [35], [36]. This implementation of low swing bus requires two extra supplies (i.e. V_{CL} and V_{SL}) to be routed across chip and uses special low-Vt devices to compensate for the current drive loss.

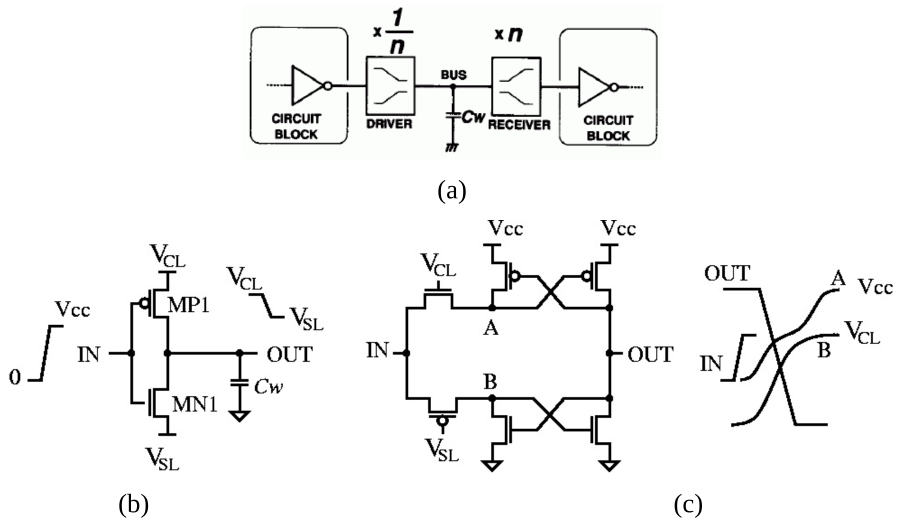
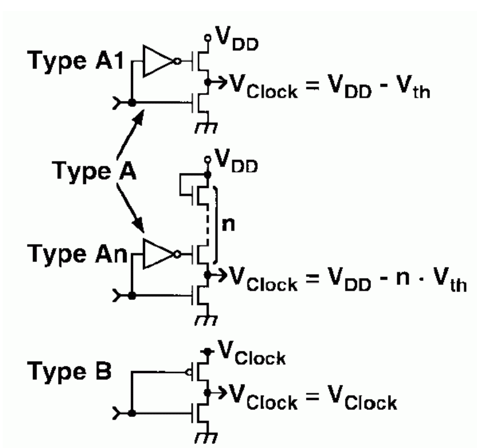
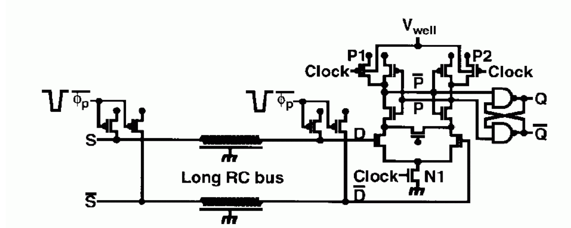


Figure 2. 10. Reduced swing signaling (a) Low swing bus architecture in [13] with (b) reduced swing line



	Driver	V_{Clock} [V]	Power	Delay	Area
Conventional		3.3	100%	100%	100%
RCSFF	Type A1	2.2	59%	82%	83%
V_{well} = 6V	Type A2	1.3	48%	123%	83%
W_{Clock} = 10μm	Type B	2.2	48%	82%	83%
f_{Clock} = 100MHZ	Type B	1.3	37%	123%	83%



(a) (b)

Figure 2. 11. Low swing signaling for clock distribution (a) Several reduced swing driver implementations [15] and (b) associated performance for signaling in global interconnects.

Another implementation of reduced voltage swing is reported in [37], [38]. Although the reported application in [37] targets the reduction of clock power, the implementation is applicable to an internal bus. A reduced clock swing of 0 to V_{Clock} (i.e. $V_{Clock} < V_{DD}$) output is created using one of the driver topologies shown in Figure 2.11a. The receiver consists of a sense amplifier flip-flop [39] with the exception of the pre-charge transistors P1 and P2. As shown in Figure 2.11b, the well of P1 and P2 are tied to a high voltage supply (i.e. V_{WELL}=6V) thereby increasing the threshold voltage V_{TP} and allowing the low clock swing to switch off P1 and P2 completely during the evaluation period. A performance comparison is shown in Figure 2.11b for power, Clk-to-Q delay and area. Various other schemes for reducing the clock swing are reported in [40]-[42].

In [43], a low swing interconnect system with distributed line equalization is explored. As shown in Figure 2.12, the system consists of three parts: a low swing transmitter with overdrive and distributed equalization scheme; differential interconnects, twisted to minimize noise effects; and a receiver using a clock regenerative amplifier. Before each bit transmission, the differential lines are pre-equalized to recycle the charge. The pre-equalization transistors are driven by global clock signal potentially increasing the clock load and clock distribution slightly more complicated. The driver, which consists of N-fet pull-ups and pulldowns, are operated from an additional reduced voltage supply.

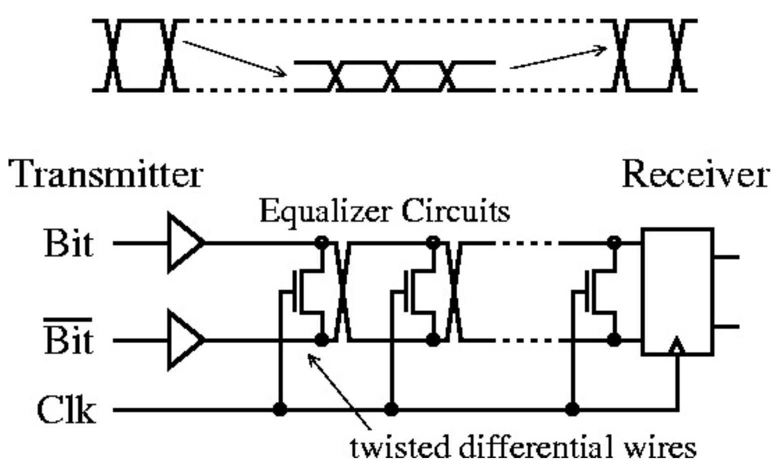
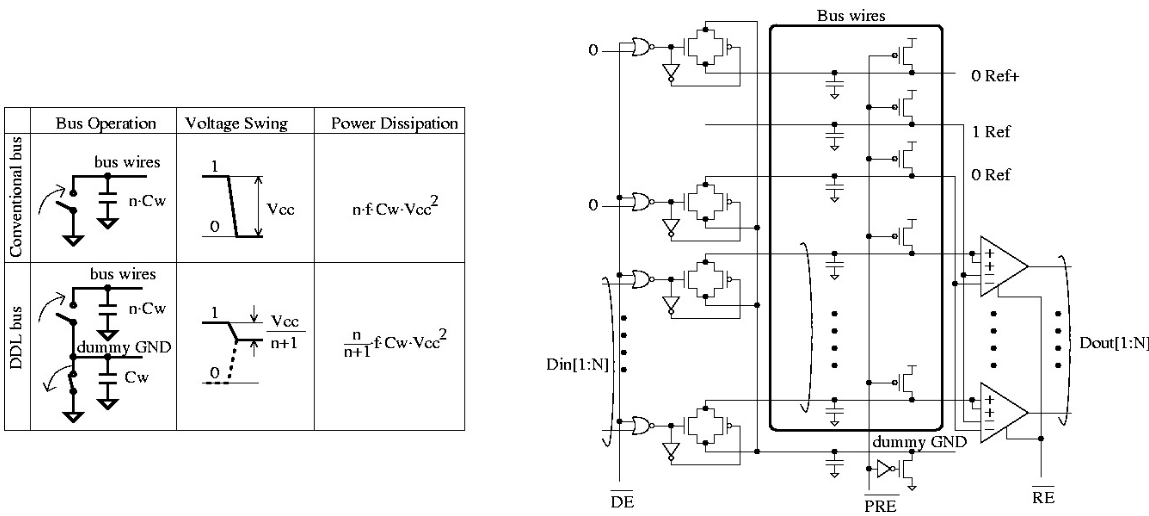


Figure 2. 12. Low swing interconnect system [43].

A charge inter-shared bus (CISB) and charge-recycling (CRB) implementation of a low swing bus has been proposed in [44], [45] and [46]-[48], respectively. In a conventional bus, the individual bus lines are either charged to the supply voltage or discharged to ground depending on the input data. Alternatively, in a charge inter-shared and recycling schemes, the capacitive bus lines are shared with a dummy ground [44] or amongst each other via switches [46], which result in a reduction of voltage swing by a factor of n due to capacitive charge-sharing between multiple data bit lines of a bus (where n is the number of bits). The CISB scheme uses single-ended lines with extra bit lines for references and the CRB uses fully differential signaling.



(a) (b)

Figure 2. 13. Charge recycling bus technique (a) Concept of data dependent charge sharing bus, (b) circuit implementation.

To illustrate the concept of a charge sharing bus, the operation and architecture of CISB in [44] is shown in Figure 2.13. In Figure 2.13a, the bus operation in each cycle consists of pre-charging the data bit lines and pre-discharging the dummy ground. In Figure 2.13b, when the bus drivers are enabled after pre-charging, the charge in the bus lines is either discharged to the dummy ground or stay at the supply voltage. At the receiver, an intermediate voltage reference is extracted from the reference lines “0” Ref and “1” Ref, which is compared to the reduced swing data bit line voltages and converted into CMOS-level outputs. The operation of the CRB in [46] uses the same concept of charge sharing with the exception that the charge is recycled or reused between the individual bit lines. The implementation of these bus schemes require multiple timing signals, well capacitively matched bus lines and are susceptive to cross-talk noise to the floating nature of the interconnects [49].

To summarize, several reduced swing bus architectures have been presented [34]-[49]. In general, most of the low swing schemes reported operate at low clock frequencies (between 50-100MHz) and are susceptible to cross-talk noise due to reduced noise margins. Differential signaling schemes can be used to improve signaling bandwidth and cross-talk at the expense of increased wiring area. Delay overhead due to low-swing to full-swing level conversion is typically unavoidable. Power savings of approximately 50% is typically achieved, with larger reported savings for the charge-sharing schemes. In addition, most schemes require extra control signals, added power supply wire distribution and low-Vt devices.

2.3.4. Noise-aware Bus

The combined effect of decreasing metal pitches, increasing interconnect aspect-ratios and faster switching edge rates will progressively exacerbate the signal integrity problem due to capacitive cross-talk noise from adjacent interconnects. The simplest and perhaps the most effective way to decrease the effect of cross-talk noise due to adjacent interconnects is to increase the spacing of the metal wires and/or use extensive shielding. However, the continuing demands for high-level of system integration often does not permit for this. The systematic insertion of repeaters can mitigate this problem in long bus wires [20], [50], which often run adjacent to each other. Since repeaters divide long interconnects into smaller segments, the cross-coupling capacitance can be decreased resulting in smaller cross-talk noise [51].

Cross-talk noise on a victim wire from adjacent bus lines can cause the receiver logic to trigger erroneously. Furthermore, simultaneous opposite transitions of adjacent wires causes bus delay increase. One way of mitigating this effect is to carefully interleave busses to suppress the simultaneous opposite transitions between neighboring wires. In [6], a staggered firing bus (SFB) scheme was used to reduce the delay degradation due to cross-talk noise, as shown in Figure 2.14. The basic idea is to drive interconnects at different timing by applying additional delay on adjacent lines. Similar schemes have been proposed in [52], [53]. The regenerative repeater insertion technique described in

[33] is also used to minimize the effect cross-talk noise via clamp transistors that maintain the line voltages stable in the presence of noise.

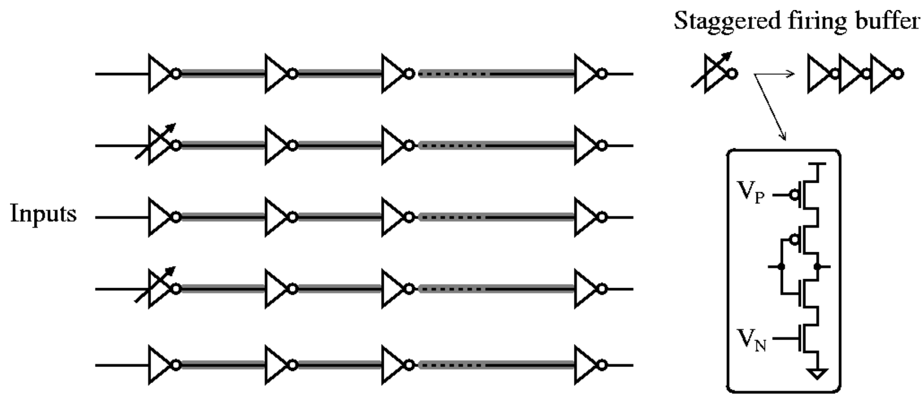


Figure 2. 14. Stagger firing repeaters for reduced cross-talk noise.

In [54], a transition encoded dynamic bus scheme is proposed to reduce the effect of coupling capacitance on delay. The worst-case delay occurs when neighboring lines transition simultaneously in opposite direction to the signal line, resulting in a 2x increase of effective coupling capacitance. The worst case coupling capacitance can be reduced to 1x with a dynamic bus that evaluates conditionally depending on input data. However, now that the lines are pre-charged every cycle regardless of input data activity, unlike static CMOS busses, there is substantial power loss for low input switching activity signals. To compensate for the loss of power, encoder and decoder circuits are inserted at the input and output of the bus lines, respectively. The encoder translates input data activity into a logic state. A ‘0’ output indicates no transition occurred on the input and a ‘1’ output indicates the input transitioned from a ‘0’ to ‘1’ or ‘1’ to ‘0’. The decoder uses the encoded signal circuit to reconstruct the original input to the encoder, and uses the previously stored state to distinguish between the two transitions. The results in [54] indicate a 10%-35% performance improvement with this technique.

Another technique proposed in [55], uses a static pulsed bus (SPB) driver, repeater and receiver technique to reduce the worst-case effective coupling capacitance to 1x. As shown in Figure 2.15, a pulse generator circuit produces a narrow pulse for every data transition. The pulse propagates through the chain of repeaters and interconnect RC segments, and drives the input of a toggle flip-flop (TFF), as shown in Figure 2.15. The TFF converts the received pulse back into a data transition at the receiver output. At the driver end, transmitted pulses are monotonic since the data is transition encoded. In addition, the repeaters are skewed in time to provide additional immunity from signals in neighboring lines. The delay improvement of the SPB technique is therefore due to the effective coupling capacitance reduction that results from transition ending the data and repeater skewing.

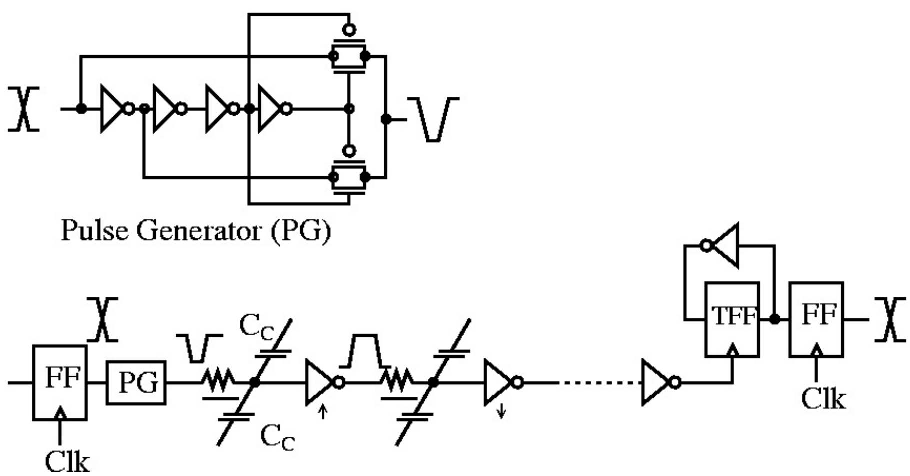


Figure 2. 15. Static Pulsed Bus technique.

2.3.5. Current-mode bus

Sensing current instead of voltage with a low impedance receiver can dramatically increase the bandwidth of long global interconnects and reduce the overall propagation delay. However, despite the frequent use of current sensing techniques in high-speed SRAM circuits [56]-[58], current-mode sensing has received limited attention for global bus on-chip signaling [59]. Therefore, the discussion presented herein will focus on current sensing techniques used in memory circuits. However, the circuits presented can also be applied for global on-chip signaling.

The speed performance of current sensing versus voltage sensing in SRAM data path was analyzed in [60], [61]. The proposed current sensing data path circuit is shown in Figure 2.16a, which consists of a positive feedback current sense amplifier followed by a differential to single-end output section for generating CMOS signal levels. The circuit responds to a small change in current (about 150μA of read current) and exhibits a sensing delay of 2ns for a bit line capacitance of 2pF. A slightly different circuit configuration was demonstrated in [62], [63]. The circuit schematic of the clamped bit-line sense amplifier for 1T DRAM in [62] is shown in Figure 2.16b. Although the circuit operation occurs in three phases (pre-charge, sense and restoration), the main feature of this circuit is that devices M5 and M6 (clamp transistors) provide a low impedance termination to the bit lines thus enhancing the speed performance. The sensing of the bit line currents is implemented via the positive feedback provide by transistors M1 and M2.

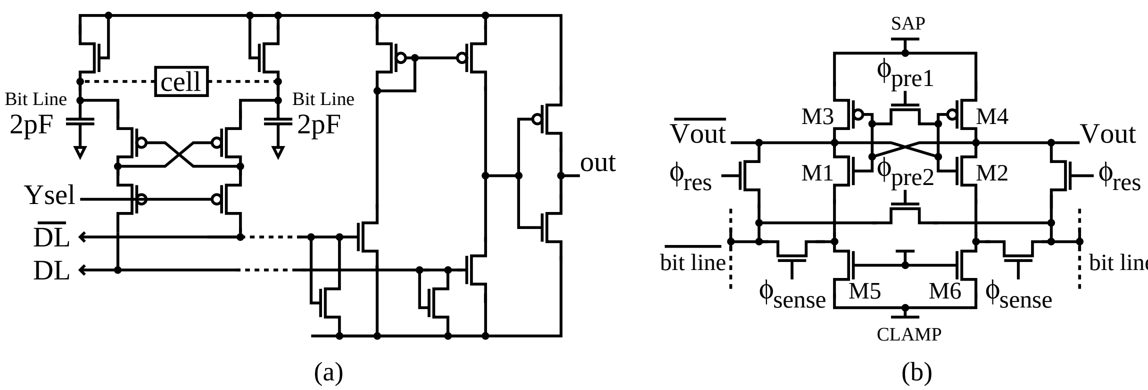


Figure 2. 16. Current mode sense amplifiers for memory circuits (a) Current sense amplifier in [60], (b) clamped current sense amplifier in [62].

A single-ended current direction sense circuit for multiport SRAM's was proposed in [64], as shown in Figure 2.17a. The M5 and M6 transistors operate in the linear region to provide a low-impedance termination to the bit line (denoted as INT). A change in current direction creates a voltage change in the drain of M5 (i.e. INT), which is sensed by the positive feedback loop of M3 and M4. The difference in voltage V1 and V2 is amplified into CMOS level by the second amplifier. Another implementation of a current sense amplifier is presented in [65]. The input of the sense amplifier consists of a current-conveyor type circuit, which provides a very low impedance termination to the bit lines via the feedback amplifiers A1 and A2, as shown in Figure 2.17b. Although this implementation of a current sense amplifier is limited in speed by the response time of the amplifiers A1 and A2, it is capable of detecting very small changes in bit line voltages (~3mV). The transistors M1 and M2 are intentionally made with large gate lengths such that a small change in bit line voltage creates a detectable change in V1 and V2, which is subsequently sensed and amplifier into CMOS level by an amplifier (not shown).

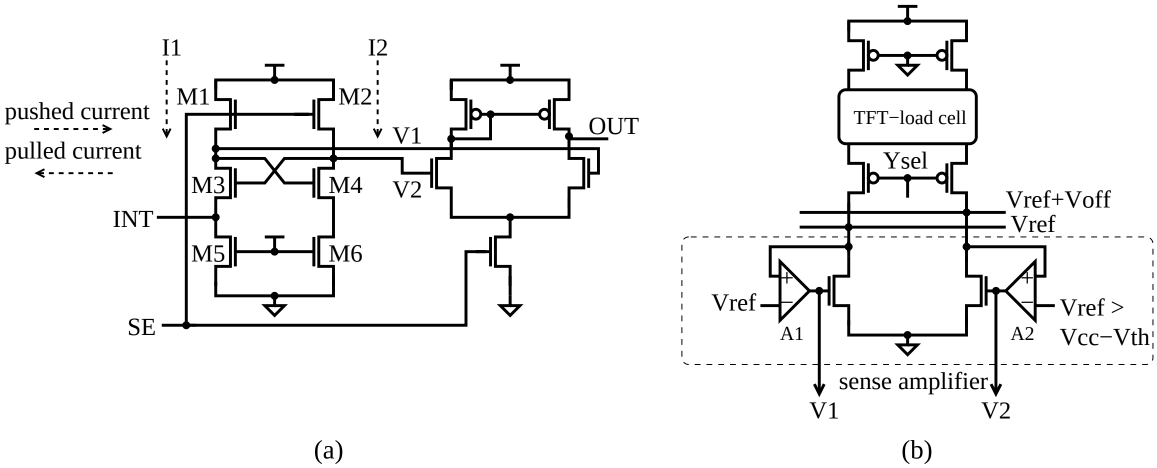


Figure 2. 17. Current-mode sense circuits (a) Single-ended current sense amplifier for multi-port SRAM [64], (b) current conveyor implementation for differential SRAM [65].

In summary, the key to high-speed current sensing schemes is the low input impedance of the sense amplifier, which is typically implemented by terminating the line with a diode connected transistor [60] or a transistor biased in the linear region of operation [62]-[64]. The voltage induced by the change in current at the input of the sense amplifier is commonly amplified by a regenerative positive feedback loop and converted to CMOS level by a second amplifier. One drawback of the high input sensitivity of the sense amplifier is the relatively low noise margins, which is undesirable due to cross-talk noise from adjacent interconnects. Other current sensing schemes based on non-regenerative amplification are commonly used to drive I/O at high speeds [66]-[68], although typically result in higher power dissipation.

2.3.6. On-chip transmission lines

Electromagnetic wave propagation in on-chip transmission line structures is desirable because the peak phase velocity is determined by the speed of light in the dielectric surrounding the interconnect (i.e. $3\text{x}10^8/\sqrt{\varepsilon}$ m/s). Unlike distributed RC lines, where signals travel slowly by diffusion, the LC nature of an optimized transmission line structure allows for faster signal transmission [69]. Signal transmission in the higher frequency regime ($R \ll \omega L$) is explored in [69]. The technique uses a 7.5Ghz local oscillator to up-convert and transmit 1Ghz of data bandwidth across an optimized on-chip transmission line spanning 20mm in length. Although improvement in delay (i.e. flight time) is achieved over standard repeater insertion techniques, there is a penalty in interconnection area and local oscillator signal distribution. Moreover, signal transmission at the higher frequencies required for electromagnetic wave propagation results in losses that depend not only on the well-controlled series resistance of the line, but also on frequency dependent losses such as skin effect, proximity effect and substrate coupling effects [70].

In an effort to understand on-chip interconnects, a significant amount of work has been dedicated to the modeling and characterization of on-chip transmission lines [71]-[76]. Interconnect line structures have been designed for improved performance in [77]-[79]. In [77]-[78], grooves are placed underneath the clock signal lines and in [79] patterned ground shields are used to minimize the frequency dependent losses. The basic idea behind segmenting or patterning the ground plane is to reduce the effective capacitance seen by the signal line, which increases the characteristic impedance Z_o and thus l_{max} (the length that can be achieved with low loss transmission line behavior). Although cross-talk noise coupling from lower metal layers can become an issue, this technique results in improved performance without special processing requirements.

The use of off-chip transmission lines as a shared medium for global communications has been proposed in [80]-[82]. In [80], a coplanar-waveguide (CPW) is used as an “off-chip” but “in-package” transmission medium shared by multiple I/Os. The interconnect system is intended to be used as a miniature LAN with multiple users and capacitive couplers as near-field antennas. Capacitive coupling as well as inductive coupling is used in [81] to achieve high I/O density and signal bandwidth for multi-chip modules (MCM) via ball-grid array flip-chip technology. In [82], a chip carrier with thin film wiring (TFW) is used to create a “DiePack” which contains several smaller ICs connected via solder bump technology. The TFW of the chip carrier can thus be used for intra-chip as well as inter-chip interconnections. These signaling schemes based on off-chip transmission line carriers are more suitable for off-chip communications due the associated power and circuit overhead. Furthermore, historically on-chip interconnects have always been most economical when compared MCM type solutions.

2.4. Summary

Signaling across long global interconnects poses significant design challenges in present and future technology nodes. The primary focus to date has been to minimize the propagation delay, although wire cross-talk noise is progressively receiving much deserved attention. In addition, the drive for low power systems has spawned significant activity to decrease the power dissipation of on-chip interconnects by reducing the voltage swings. As technology continues to scale, delay, noise and power performance of on-chip interconnects will become increasingly important and is likely to limit overall system performance. An added complication will result from the signaling bandwidth limitations of on-chip global interconnect due to the increase in wire resistance. Evidently, the designer will have to trade-off among these performance metrics to best meet the desired target goals.

A summary of design techniques for on-chip signaling is presented in Table 2.4. The summary of published work in Table 2.4 clearly shows that minimizing delay and power dissipation have been target goals in both high performance MPUs and DSP cores. A look at current signaling solutions and future technology trends suggest that minimizing power dissipation will remain at the center of target performance metrics. Although delay latency is crucial in high-performance systems, it is likely to be unavoidable even with optimal repeater insertion due to the rapidly increasing clock frequencies.

Solutions to the cross-talk noise problem mainly consist of systematically inserting buffers and/or regenerative repeaters whereas data throughput is addressed via synchronous and elastic repeaters. It seems, however, that cross-talk will become an increasingly difficult problem to solve due to higher wiring aspect ratios, faster edge-rates and smaller metal pitches. In addition, longer global interconnects due to increasing chip size and resistive effects will limit the maximum bandwidth of global signals.

Table 2. 4. Summary of signaling techniques.

Technique	Ref.	Objective	Comments	Application
Buffer Insertion	[11], [17]-[25]	Minimize propagation delay.	Reduces cross-talk	MPU
Regenerative Amplification	[30]-[33]	Minimize propagation delay.	Reduces cross-talk	MPU
Synchronous Pipelines	[26]-[28]	Maximize throughput	Significant area overhead. Suitable for WSI	WSI
Elastic Interconnects	[29]	Maximize throughput and data contention/congestion	Sensitive to noise.	SOC, Multiple MPUs
Low swing internal bus	[35], [36]	Minimize power dissipation	Requires internal supply conversion, multiple supply voltages and low V_t devices	DSP cores
Low swing equalized bus	[42]	Minimize power, improve cross-talk	Differential signaling, Clocking overhead	MPU
Reduced clock swing	[37],[38], [40]-[41]	Minimize power dissipation in clock lines	Requires high-voltage or multiple supplies. Delay penalty. Req. low swing latches/FF. Low V_t devices.	DSP cores, low end MPU
Data dependent Bus/ Charge-recycling Bus	[44],[45], [46]-[48]	Minimize bus power dissipation.	Sensitive to noise. Implementation complexity Low Speed: 50-100MHz	DSP cores
Staggered Firing Repeaters	[6]	Minimize delay due to cross-talk noise	Possible repeater placement constraints	MPU, DSP
Intentional data skewing in repeater chains	[52], [53]	Minimize delay due to cross-talk noise	Additional digital logic required.	MPU, DSP
Transition Encoded dynamic bus	[54]	Minimize delay due to coupling capacitance	Worst-case 1xCc, clocking overhead	MPU
Transition encoded static bus	[55]	Minimize delay due to coupling capacitance	Worst-case 1xCc, Repeater staggering, small circuit overhead,	MPU
Current-Sense amplifiers	[56]-[58], [60]-[63], [65]	Minimize delay in SRAMs	Very small signal swings. Voltage swing conversion delay penalty.	SRAM, DRAM
Current direction sense technique	[64]	Minimize delay in single-ended SRAMs	Sensitive to noise	SRAM, DRAM

Table 2.4. (continued)

Technique	Ref.	Objective	Comments	Application
Current-mode signaling for I/O	[66]-[68]	High-speed I/O	Added power dissipation. Off-chip signaling	I/O, MPU
Current-mode sensing for global bus	[59]	Minimize on-chip delay	Increased static power dissipation.	MPU
On-chip transmission lines	[69]	Minimize on-chip delay	Local oscillator overhead, wide optimized lines	PU
On-chip transmission lines	[77]-[79]	Minimize transmission delay. Patterned ground planes to increase TL Z_o.	Controlled TL are suitable for clock distribution..	MPU
Shared off-chip transmission line mediums	[80]-[82]	Increase bandwidth and I/O density	Off-chip TL carriers. Added power dissipation. Suitable for MCMs.	SoC, MCM.

2.5. References

[1] Semiconductor Industry Association, *The International Technology Roadmap for Semiconductors*, 2001 Edition.

[2] W. E. Donath, “Wire-length distribution for placements on computer logic,” IBM J. Research and Development, vol. 2, no. 3, pp. 152-155, May 1981.

[3] B.S. Landman and R.L. Russo, “On a pin versus block relationship for partiotions of logic graphs,” IEEE Trans. Comput., vol. C-20, pp. 1469-1479, Dec. 1971.

[4] P. Zarkesh-Ha, J.A. Davis, J.D. Meindl, “Prediction of Net-Length Distribution for Global Interconnects in a Heterogeneous System-on-a-Chip,” IEEE Trans. VLSI, vol. 8, no. 6, Dec. 2000.

[5] B. Herrick, “Design Challenges in Multi-GHz Processors,” Slide presentation, Alpha Microprocessor Development.

[6] K. Nose, T. Sakurai, “Two schemes to reduce interconnect delay in bi-directional and uni-directional buses,” Symposium on VLSI Circuits, pp. 193-194, 2001.

[7] A. Dansky, H. Smith, P. Williams, “On-chip Coupled Noise Analysis of a High Performance S/390 Microprocessor,” Elec. Comp. Tech. Conf., pp. 817-825, 1997.

[8] C. Webb et al., “A 400-MHz S/390 Microprocessor,” IEEE J. Solid-State Circuits, vol. 32, no. 11, pp. 1665-1675, April 1997.

[9] J. Cong, “An interconnect-centric design flow for nanometer technologies,” Proc. of the IEEE, vol. 89, no. 4, pp. 505-528, April 2001.

[10] K. Banerjee, S. Souri, P. Kapur, K. Saraswat, “3-D ICs: a novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration,” Proc. of the IEEE, vol. 89, no. 5, May 2001.

[11] H.B. Bakoglu, Circuits, Interconnections, and Packaging for VLSI. Reading, MA: Addison-Wesley, 1990.

[12] W.J. Dally, J.W. Poulton, Digital Systems Engineering. Cambridge, UK: Cambridge Univ. Press, 1998.

[13] P.J. Restle, et. al., “A Clock Distribution Network for Microprocessors,” IEEE J. Solid-State Circuits, vol. 36, no. 5, May 2001.

[14] T. Xanthopoulos, et. al., “The Design and Analysis of the Clock Distribution Network for a 1.2GHz Alpha Microprocessor,” ISSCC Digest of Technical Papers, Feb. 2001.

[15] D. Liu, C. Svensson, “Power Consumption Estimation in CMOS VLSI Chips,” IEEE J. Solid-State Circuits, vol. 29, no. 6, June 1994.

[16] A.K. Stamper, M. Fuselier, X. Tian, “Advanced Wiring RC Delay Issues for sub-0.25-micron Generation CMOS,” International Interconnect Technology Conference, pp 62-64, 1998.

[17] H.B. Bakoglu, J.D. Meindl, "Scaling and Performance Aspects of Technology," ISSCC Digest of Technical Papers, pp. 164-165, 1984.

[18] W.C. Elmore, "The transient response of damped linear networks with particular regard to wide-band amplifiers," J. Applied Physics, vol. 19, no. 1, pp. 55-63, Jan. 1948

[19] J. Rubinstein, P. Penfield, M. Horowitz, "Signal Delay in RC Tree Networks," IEEE Trans. Computer-Aided Design, vol. 2, no. 3, July 1993, pp. 202-210.

[20] R. McInerney et al., "Methodology for Repeater Insertion Management in the RTL, Layout, Floorplan and Fullchip Timing Databases of the Itanium™ Microprocessor," ISPD, pp. 99-104, 2000.

[21] N. Kojima et al., "Repeater Insertion Method and its Application to a 300MHz 128-bit 2-way Superscalar Microprocessor," Design Autamation Conf., pp. 641-646, 2000.

[22] C. Wu, M. Shiau, "Delay models and Speed Improvement Techniques for RC Tree Interconnections Among Small-Geometry CMOS Inverters," IEEE J. Solid-State Circuits, vol. 25, no. 5, pp. 1247-1256, Oct. 1990.

[23] T. Sakurai, "A unified Theory for Mixed CMOS/BiCMOS Buffer Optimization," IEEE J. Solid-State Circuits, vol. 27, no. 7, pp. 1014-1019, Oct. 1992.

[24] V. Adler, E.G. Friedman, "Uniform Repeater Insertion in RC Trees," IEEE Trans. VLSI Systems, vol. 47, no. 10, pp. 1515-1523, Oct. 2000.

[25] Y. Ismail, E.G. Friedman, "Effects of Inductance on the Propagation Delay and Repeater Insertion in VLSI Circuits," IEEE Trans. VLSI Systems, vol. 8, no. 2, pp. 195-206, April 2000.

[26] P.L. Edefors, "Investigation on maximal Throughput of a CMOS Repeater Chain," IEEE Trans. Cir. Systems I, vol. 47, no. 4, pp. 602-606, April 2000.

[27] D. Audet, Y. Savaria, N. Arel, "Pipelining Communications in Large VLSI/ULSI Systems," IEEE Trans. VLSI Systems, vol. 2, no. 1, pp. 1-10, Mar. 1994.

[28] D. Audet, Y. Savaria, "An Architectural Approach for Increasing Clock Frequency and Communication Speed in Monolithic WSI Systems," IEEE Trans. Comp., Packg., Manuf. Tech., vol. 17, no. 3, pp. 362-368, Aug. 1994.

[29] M. Mizuno, W.J. Dally, H. Onishi, "Elastic Interconnects: Repeater-Inserted Long Wiring Capable of Compressing and Decompressing Data," ISSCC Digest of Technical Papers, pp. 346-347, 2001.

[30] L. Glasser, D. Dobberpuhl, The design and analysis of VLSI circuits. Reading MA: Addison-Wesley Publishing Company, 1985.

[31] M. Nekili, Y. Savaria, "Parallel regeneration of interconnections in VLSI and ULSI circuits," IEEE Symp. Circuits Systems, pp. 2023-2026, May 1993.

[32] I. Dobbelaere, M. Horowitz, A. El Gamal, "Regenerative Feedback Repeaters for Programmable Interconnections," IEEE J. Solid-State Circuits, vol. 30, no. 11, pp. 1246-1253, Nov. 1995.

[33] T. Iima, M. Mizuno, T. Horiuchi, M. Yamashina, “Capacitive Coupling Immune, Transient Sensitive Accelerator for Resistive Interconnect Signals of Subquarter Micron ULSI,” IEEE J. Solid-State Circuits, vol. 31, no. 4, pp. 531-536, April 1996.

[34] C. Svensson, “Optimum Voltage Swing on On-chip and Off-Chip Interconnect,” IEEE J. Solid-State Circuits, vol. 36, no. 7, pp. 1108-1112, July. 2001.

[35] Y. Nakagome, K. Itoh, M. Isoda, K. Takeuchi, M. Aoki, “Sub-1-V Swing Architecture for Future Low-Power ULSI’s,” Symposium VLSI Circuits Dig. Tech. Papers, pp. 82-83, 1992.

[36] Y. Nakagome, K. Itoh, M. Isoda, K. Takeuchi, M. Aoki, “Sub-1-V Swing Internal Bus Architecture for Future Low-Power ULSI’s,” IEEE J. Solid-State Circuits, vol. 28, no. 4, pp. 414-419, April 1993.

[37] H. Kawaguchi, T. Sakurai, “A Reduced Clock-Swing Flip-Flop (RCSFF) for 63% Clock Power Reduction,” Symposium on VLSI Circuits, pp. 97-98, 1997.

[38] H. Kawaguchi, T. Sakurai, “A Reduced Clock-Swing Flip-Flop (RCSFF) for 63% Power Reduction,” IEEE J. Solid-State Circuits, vol. 33, no. 5, pp. 807-811, May 1998.

[39] M. Matsui, et al., “A 200 MHz 13mm^2 2-D DCT macrocell using sense amplifying pipeline flip-flop scheme,” IEEE J. Solid-State Circuits, vol. 29, no. 12, pp. 1482-1490, Dec. 1994.

[40] H. Kojima, S. Tanaka, K. Sasaki, “Half-Swing Clocking Scheme for 75% Power Saving in Clocking Circuitry,” IEEE J. Solid-State Circuits, vol. 30, no. 4, pp. 432-435, April 1995.

[41] C. Kim, S. Kang, “A Low-Swing Clock Double-Edge Triggered Flip-Flop,” IEEE J. Solid-State Circuits, vol. 37, no. 5, pp. 648-652, May 2002.

[42] J. Kim, S. Lee, H. Park, “A High-Speed 50% Power-Saving Half-Swing Clocking Sceme for Flip-Flop with Complementary Gate and Source Drive,” Int. Conf. VLSI and CAD, pp. 574-577, 1999.

[43] R. Ho, K. Mai, M. Horowitz, “Efficient On-Chip Global Interconnects,” Symposium on VLSI Circuits, pp. 271-274, 2003.

[44] M. Hiraki et al., “Data-dependent logic swing internal bus architecture for ultra-low-power LSI’s,” Symposium on VLSI Circuits Dig. Tech. Papers, pp. 29-30, June 1994.

[45] M. Hiraki et al., “Data-dependent Logic Swing Internal Bus Architecture for Ultralow-Power LSI’s,” IEEE J. Solid-State Circuits, vol. 30, no. 4, pp. 397-402, April 1995.

[46] H. Yamauchi et al., “A low power complete charge-recycling bus architecture for ultra-high data rate ULSI’s,” Symposium on VLSI Circuits Dig. Tech. Papers, pp. 21-22, June 1994.

[47] H. Yamauchi, H. Akamatsu, T. Fujita, “An Asymptotically Zero Power Charge-Reclying Bus Architecture for Battery-Operated Untrahigh Data Rate ULSI’s,” IEEE J. Solid-State Circuits, vol. 30, no. 4, pp. 423-431, April 1995.

[48] H. Yamauchi, H. Matsuzawa, “A Signal-Swing Suppressing Strategy for Power and Layout Area Savings Using Time-Multiplexed Differential Data-Transfer Scheme,” IEEE J. Solid-State Circuits, vol. 31, no. 9, pp. 1285-1294, Sept. 1996.

[49] H. Zhang, V. George, J.M. Rabaey, “Low-Swing On-chip Signaling Techniques: Effectiveness and Robustness,” IEEE Trans. VLSI, vol. 8, no. 3, pp. 264-272, June 2000.

[50] R. Kumar, “Interconnect and Noise Immunity Design for the Pentium 4 Processor,” Intel Technology Journal Q1, pp. 1-12, 2001.

[51] T. Sakurai, “Closed-form expression for interconnection delay, coupling, and cross-talk in VLSI,’s” IEEE Trans. Electron Devices, vol. 40, pp. 118-124, Jan. 1993.

[52] J. Yim, C. Kyung “Reducing Cross-coupling among Interconnect Wires in Deep-Submicron Datapath Design,” Design Automation Conf. pp. 485-490, 1999.

[53] K. Hirose, h. Yasuura, “A Bus Delay Reduction Technique Considering Crosstalk,” Design Automation Test Conf. pp. 441-445, 2000.

[54] M. Anders, N. Rai, R. Krishnamurthy, S. Borkar, “A Transition-Encoded Dynamic Bus Technique for High-Performance Interconnects,” Symposium on VLSI Circuits Dig. Tech. Papers, pp. 16-17, June 2002.

[55] M. Khellah, J. Tschantz, Y. Ye, S. Narendra, V. De, “Static Pulsed Bus for On-Chip Interconnects,” Symposium on VLSI Circuits Dig. Tech. Papers, pp. 78-79, June 2002.

[56] M. Matsumiya et al., “A 15ns 16Mb CMOS SRAM with Reduced Voltage Amplitude Bus,” ISSCC Dig. Tech. Papers, pp. 214-215, 1992.

[57] K. Sasaki et al., “A 7-ns 140-mW 1-Mb CMOS SRAM with Current Sense Amplifier,” IEEE J. Solid-State Circuits, vol. 27, no. 11, pp. 1511-1518, Nov. 1992.

[58] N. Shibata et al., “A 2-V 300-MHz 1-Mb Current-Sensed Double-Density SRAM for Low-Power 0.3-μm CMOS/SIMOX ASICs,” IEEE J. Solid-State Circuits, vol. 36, no. 10, pp. 1524-1537, Oct. 2001.

[59] A. Maheshwari, W. Burleson, “Current Sensing Techniques for Global Interconnects in Very Deep Submicron (VDSM) CMOS,” IEEE Comp. Soc. Work. VLSI, pp. 66-70, 2001.

[60] E. Seevinck, P. van Beers, 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.

[61] E. Seevinck, “A current sense-amplifier for fast CMOS SRAMs,” Symp. VLSI Circuits, pp. 71-72, 1990.

[62] T. Blalock, R. Jaeger, “A high-speed sensing scheme for 1T dynamic RAM’s utilizing the clamped bit-line sense amplifier,” IEEE J. Solid-State Circuits, vol. 27, no. 4, pp. 618-625, April 1992.

[63] T. Blalock, R. Jaeger, “An experimental 2T cell RAM with 7ns access time at low temperature,” Symp. VLSI Circuits, pp. 13-14, 1990.

[64] M. Izumikawa, M. Yamashina, “A current direction sense technique for multiport SRAM’s,” IEEE J. Solid-State Circuits, vol. 31, no. 4, pp. 546-551, April 1996.

[65] K. Ishibashi et al., “A 6-ns 4-Mb CMOS SRAM with Offset-Voltage-Insensitive Current Sense Amplifiers,” IEEE J. Solid-State Circuits, vol. 30, no. 4, pp. 480-486, April 1995.

[66] M. Ishibe et al., “High-Speed CMOS I/O Buffer Circuits,” IEEE J. Solid-State Circuits, vol. 27, no. 4, pp. 671-673, April 1992.

[67] B. Chappel et al., “Fast CMOS ECL Receivers With 100-mV Worst-Case Sensitivity,” IEEE J. Solid-State Circuits, vol. 23, no. 1, pp. 59-67, Feb. 1988.

[68] S. Long, J. Zhang, “Low Power GaAs Current-Mode 1.2 Gb/s Interchip Interconnections,” IEEE J. Solid-State Circuits, vol. 32, no. 6, pp. 890-897, June 1997.

[69] R.T. Chang, N. Talwalkar, P.Yue, S. Wang, “Near Speed-of-Light Signaling Over On-Chip Electrical Interconnects,” IEEE J. Solid-State Circuits, vol. 38, no. 5, pp. 834-838, May 2003.

[70] B. Kleveland, C.H. Diaz, D. Vook, L. Madden, T.H. Lee, S. Wang, “Exploting CMOS Reverse Interconnect Scaling in Multigigahertz Amplifier and Oscillator Desing,” IEEE J. Solid-State Circuits, vol. 36, no. 10, pp. 1480-1488, Oct. 2001.

[71] A. Deutsch et al. “On-Chip Wiring Design Challeneges for Gigahertz Operation,” Proc. of the IEEE, vol. 89, no. 4, pp. 529-555, April 2000.

[72] A. Deutsch et al. “Modeling and characterization of long on-chip interconnections for high-performance microprocessors,” J. Res. Develop., vol. 39, no. 5, pp. 547-567, Sept. 1995.

[73] D.C. Edelstein, G.A. Sai-Halasz, Y. Mii, “VLSI on-chip interconnection performance simulations and measurements,” J. Res. Develop., vol. 39, no. 5, pp. 383-401, July 1995.

[74] K. Soumyanath, S. Borkar, C. Zhou, B. Bloechel, “Accurate On-Chip Interconnect Evaluation: A Time-Domain Technique,” IEEE J. Solid-State Circuits, vol. 34, no. 5, pp. 623-631, May 1999.

[75] P.J. Restle, K.A. Jenkins, A. Deutsch, P.W. Cook, “Measurement and Modeling of On-Chip Transmission Line Effects in a 400 MHz Microprocessor,” IEEE J. Solid-State Circuits, vol. 33, no. 4, pp. 662-665, April 1998.

[76] X. Qi et al., “On-chip Inductance Modeling of VLSI Interconnects,” ISSCC Digest of Technical Papers, pp. 172-173, 2000.

[77] M. Mizuno et al., “On-chip Multi-GHz Clocking with Transmission Lines,” ISSCC Digest of Technical Papers, pp. 366-367, 2000.

[78] M. Mizuno et al., “Clock Distribution Networks with On-Chip Transmission Lines,” Int. Interconnect Tech. Conf, pp. 3-5, 2000.

[79] R. Lowther, S. Lee, “On-Chip Interconnect Lines with Patterned Ground Shields,” Microwave Guided Lett., vol. 10, no. 2, Feb. 2000.

[80] F. Chang et al., “RF/Wireless Interconnect for Inter- and Intra-Chip Communications,” Proc. of the IEEE, vol. 89, no. 4, pp. 456-466, April 2000.

[81] S. Mick, J. Wilson, P. Franzon, “4 Gbps High-Density AC Coupled Interconnection,” IEEE Cust. Int. Cir. Syst., pp. 133-140, 2002.

[82] E. Davidson, B. McCredie, W. Vikelis, “Long Lossy Lines (L^3) and their Impact Upon Large Chip Performance,” IEEE Trans. Comp. Packg. Manuf. Tech., vol. 20, no. 4, pp. 361-375, Nov. 1997.

Chapter III

Current-mode Signaling for Global On-chip Interconnects

3.1. Introduction

It appears, with the increasing speed requirements in VLSI circuits [1], that current mode signal transporting techniques may provide an attractive solution to some of the challenges caused by aggressive interconnect scaling. Current mode signaling has already been shown to provide drastic speed enhancements in SRAM CMOS circuits [2]-[4]. Nonetheless, in order to apply current mode techniques to on-chip global interconnect signaling, new models will be required to accurately and efficiently estimate propagation delay and power dissipation during the early design stages of high performance VLSI systems. Various techniques based on simulations and/or analytical closed-form formulations have been proposed to model delay in interconnects. However, the bulk of the work dedicated to this area targets capacitively terminated lines for voltage mode signaling [5]-[7].

A useful closed-form delay analysis of current mode signaling for RC interconnects was presented in [2]. However, the analysis assumes the output response is a linear-ramp signal, and does not include transient effects due to a step or fast edge rate inputs. In this work, in section 3.3, we present an analytical model based on the closed-form formulation of the effective resistance and capacitance of a driven distributed RC line with arbitrary termination. In section 3.4, based on this model and by satisfying certain boundary conditions, a new closed-form single exponent approximation of interconnection delay under step input excitation is derived for arbitrary load terminations. It is shown that the accuracy of this work is the same as Sakurai's formulation for voltage mode design [8], extended herein to accommodate current-mode type circuits. Thereafter, the delay model is extended to take into account finite rise time ramp input signals.

In addition to delay formulation, power dissipation in long global interconnects must be accurately estimated. Power dissipation in on-chip interconnects constitutes a significant fraction of the total power dissipation of modern day microprocessor designs. To accurately model both dynamic and static power dissipation components of current-mode signaling, a new closed-form power model is developed in section 3.5. In section 3.6, the optimum line width that minimizes the total delay for current mode circuits is discussed. This analysis is extended in section 3.7, using the derived delay and power models, a comparative study of current and voltage repeaters driving long global interconnects is presented. In section 3.8, the delay model is extended to take into account the effect of miller capacitance and is experimentally verified in section 3.9. Finally, concluding remarks are presented in section 3.10.

3.2. Current-mode vs. voltage-mode

The terms “current-mode” and “voltage-mode” are frequently used to describe circuit topologies in which the “mode” of operation determines whether current or voltage signals are processed. It is apparent, however, that signals do not stay exclusively in current or voltage domain, and hence the “mode” of operation is a loosely defined term, which does not qualify circuits in a strict sense [9]. However, there seems to be a widely excepted premise that current-mode circuits exhibit higher bandwidth performance than its voltage-mode counterpart. Much of this bandwidth enhancement is owed to sensing signals with low impedance nodes that result in smaller signal swings and hence the parasitic capacitances do not need to be charged or discharged as the signal is processed [10][11].

Perhaps implicitly defined, the key to current-mode signal transporting is the extension of signaling bandwidth and reduction of the system time constants that result from sensing signals with low impedance nodes. Hence, from hereon after, for the purpose of signaling in on-chip interconnects, *current-mode or current sensing refers to sensing a signal with a low impedance termination at the receive-end which results in a* *shift or extension in dominant pole position thereby increasing the bandwidth of the line*. Figure 3.1 shows an inverter driven interconnect line terminated by a receiver. Assuming all things are equal, the parallel termination R_L determines the impedance of the receiver and hence the current or voltage mode operation of the line.

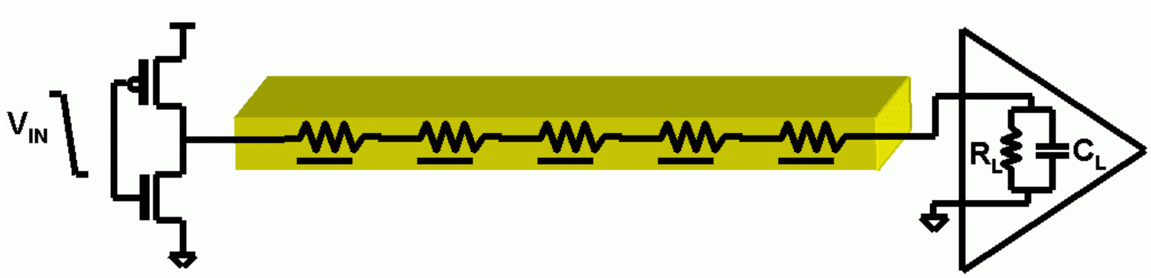


Figure 3. 1. Inverter driven interconnect terminated by an arbitrary impedance.

3.3. Analysis model of voltage and current mode signals

3.3.1. Generalized Analysis Model

Long global interconnects can be modeled by distributed RC transmission lines as long as the overall line resistance dominates the response (i.e. R»jωL). These RC interconnect lines can be driven either by voltage-mode or current-mode signals. Figure 3.2 shows typical voltage and current mode driver and receiver circuits. Current-mode drivers are similar to voltage mode drivers in that a low output impedance node drives the transmission line. Therefore, from a signaling point of view, both voltage and current mode drivers can be approximated by a voltage source and a linear resistance. Current-mode receivers, on the other hand, provide a low impedance node at the receiver whereas voltage-mode receivers present a high impedance capacitive termination.

The key to current-mode signal transporting is the low impedance termination at the receiver which results in reduced signal swings and increased bandwidth performance [2]. It can be shown, that the bandwidth performance of current-mode signaling is a direct tradeoff with the signal amplitude across R_L at the receive-end, similar to the gain-bandwidth tradeoff concept of analog amplifiers [12].

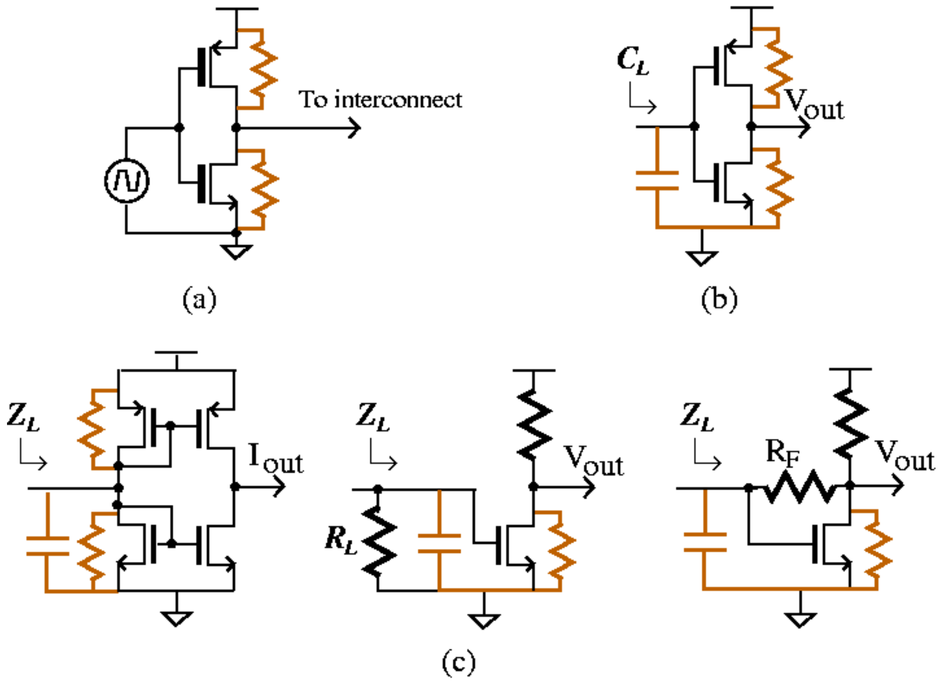


Figure 3. 2. Typical circuits for voltage and current sensing; (a) single-ended transmitter (b) voltage-mode receiver (c) current-mode receivers.

A generalized model for a driven distributed RC line is shown in Figure 3.3a. The driver is modeled as a voltage source with output resistance R_S and capacitance C_S. Note that previously derived models do not explicitly consider the effect of driver output capacitance C_S. For sake of generality the output of the line is terminated with a resistor R_L and load capacitance C_L. For voltage-mode signaling, the termination resistance R_L is infinite and the output voltage is seen across C_L. In current-mode signaling, the terminating resistance R_L is finite.

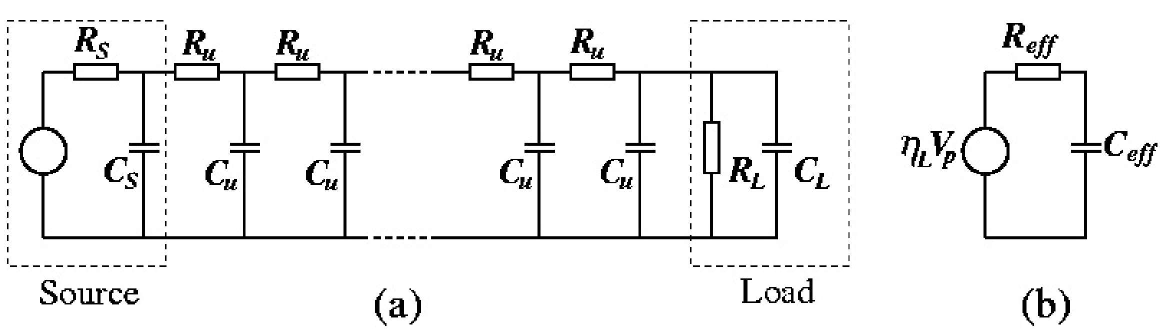


Figure 3. 3. R_L and C_L are determined from the receiver circuit topologies shown in Fig 3.1b and 3.1c; (a) generalized distributed RC model (b) approximate effective lumped element model.

3.3.2. Effective Resistance and Capacitance

A simple lumped element model with effective resistance and capacitance, to account for the distributed transmission line behavior, can be used to provide analogous results to the Elmore Delay [5]. Since Elmore's formulation is basically a first moment (i.e. mean) approximation of the signal delay time, Moment-Matching Methods [13]-[14] can be used to derive a first-order RC network with effective lumped element parameters for voltage and current mode signaling.

The effective resistance and capacitance of Figure 3.3b is derived in Appendix A, which can be expressed as,

$$R_{eff} = \eta_L \frac{R_T}{\sqrt{2}}\left(\frac{N+1}{N}\right)\cdot\left(1+\frac{R_T}{3R_L}\left(\frac{N-1}{N}\right)\right)+2\eta_L \frac{R_S}{\sqrt{2}}\left(1+\frac{R_T}{2R_L}\left(\frac{N-1}{N}\right)\right) \tag{3-1a}$$

$$C_{eff} = \frac{C_T}{\sqrt{2}} + C_S \eta_S \frac{R_L + R_T}{R_{eff}} + C_L \eta_L \frac{R_S + R_T}{R_{eff}} \tag{3-1b}$$

and

$$\eta_L = \frac{R_L}{R_L + R_S + R_T} \tag{3-1c}$$

$$\eta_S = \frac{R_S}{R_L + R_S + R_T} \tag{3-1d}$$

In Eq. (3-1), R_T and C_T are the total resistance and capacitance calculated from the per unit-section components R_u=R_T/N and C_u=C_T/N; N is the number of distributed sections; R_S (R_L) and C_S (C_L) are the source(load) resistance and capacitance, respectively. η_L and η_S are defined as voltage loss factors of the load and source, respectively. In the ideal case for voltage mode signaling (R_S =0, $R_L = \infty$, $C_S = C_L$ =0), the total delay is:

$$T_{DELAY} = \frac{R_T C_T}{2}\left(\frac{N+1}{N}\right) = R_u C_u \frac{N(N+1)}{2} \tag{3-2}$$

which is the same result obtained in Elmore's formulation. Thus, the lumped element resistance and capacitance in Eq. (3-1) are the effective components that model the distributed transmission line. The generalized effective lumped element model is shown in Figure 3-3b. For current mode signals the source voltage (Vp) is scaled by η_L. The usefulness of this model is that an equivalent lumped element model can predict the approximate step response of an interconnect line for both current and voltage mode signals. Furthermore, due to its relative simplicity, the expression in Eq. (3-1) can be used in performance driven layout tools to optimize interconnect speed.

3.4. Delay formulation

3.4.1. Delay under step input excitation

The step voltage response for a distributed line with 1000 segments and the lumped element model with effective resistance and capacitance is shown in Figure 3.4. Since Eq. (3-1) is derived by reducing a multi-pole system to a single dominant pole, the effective model voltage response of Eq. (3-1) will cross-over the distributed output response. This cross-over point occurs at approximately 63% of the normalized output voltage (i.e. when the delay is $R_{eff}C_{eff}$). An improved analytical approximation of the delay can be obtained by satisfying the boundary conditions at v_{063} and v_x with a single exponent, as shown in Figure 3.4. The new approximation to the normalized output response is given by:

$$v(t) = 1 - (1 + v_x) e^{-\frac{t}{RC}} \tag{3-3}$$

where RC is the time constant of the new approximation single exponent function. By forcing the boundary condition at $v(t{=}t_{063})$, the following equality is obtained:

$$0.63 = 1 - e^{-\frac{t_{063}}{R_{eff} C_{eff}}} = 1 - (1 + v_x) e^{-\frac{t_{063}}{RC}} \tag{3-4}$$

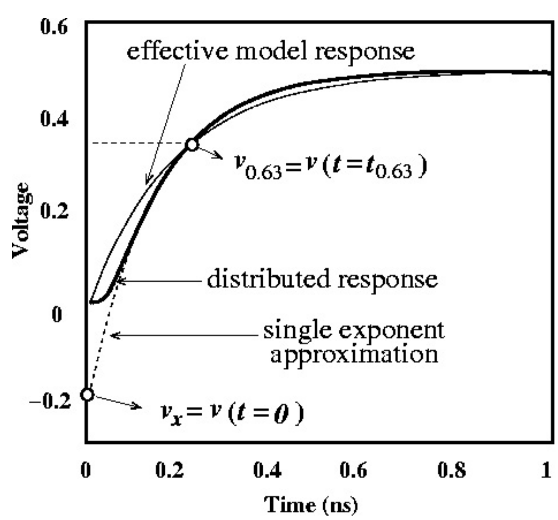


Figure 3. 4. Single exponential function approximation for the step response of a distributed interconnect line.

The delay (t_v) defined as the time from (t=0) to the time when the normalized voltage reaches v at the end of the line is given as:

$$t_v = \frac{\ln\left(\frac{1+v_x}{1-v}\right)}{1-\frac{\ln(1+v_x)}{\ln(0.37)}} R_{eff} C_{eff} \tag{3-5}$$

In Eq. (3-5), R_{eff} and C_{eff} are given by Eq. (3-1). An optimal value of v_x was found iteratively and is approximated by:

$$v_x = v(t=0) = 0.5\eta_L \frac{R_T}{R_L} + 0.2 \tag{3-6}$$

From equations (3-1), (3-5) and (3-6), the final expression of delay for voltage and current mode signals given an input step excitation is derived as:

$$\frac{t_v}{R_T C_T} = \frac{\ln\left[\frac{1.2+0.5(\eta_L/R_{LT})}{1-v}\right]}{1+1.0058\ln\left[1.2+0.5\left(\frac{\eta_L}{R_{LT}}\right)\right]} \times \left\{\frac{\eta_L}{2}\left(1+\frac{1}{3R_{LT}}\right)+\eta_L R_{ST}\left(1+\frac{1}{2R_{LT}}\right)+\eta_S C_{ST}(1+R_{LT})+\eta_L C_{LT}(1+R_{ST})\right\} \tag{3-7}$$

where $R_{LT}=R_L/R_T$, $R_{ST}=R_S/R_T$, $C_{ST}=C_S/C_T$ and $C_{LT}=C_L/C_T$. R_T and C_T are the total interconnect resistance and capacitance, respectively. The term inside the brackets depicts the contribution of the normalized source and load terminations (i.e. R_{ST}, R_{LT}, C_{ST} and C_{LT}), exhibiting great similarity to the equations derived by Sakurai [8]. However, Eq. (3-7) can be used to predict the delay of current mode signals ($R_L < \infty$) in addition to voltage mode signals ($R_L = \infty$). The accuracy of this formula is shown in Figure 3.5. In Figure 3.5a, the delay formula is compared with the distributed RC response with 1000 segments (i.e. N=1000) and Sakurai's voltage mode model. The analytical formula in Eq. (3-7) matches well with Sakurai's model, which is determined to have an error less than 3.5% for a large range of parameters [8]. This accuracy is maintained for current mode signals across a wide range of values, as shown in Figure 3.5b-3.5d.

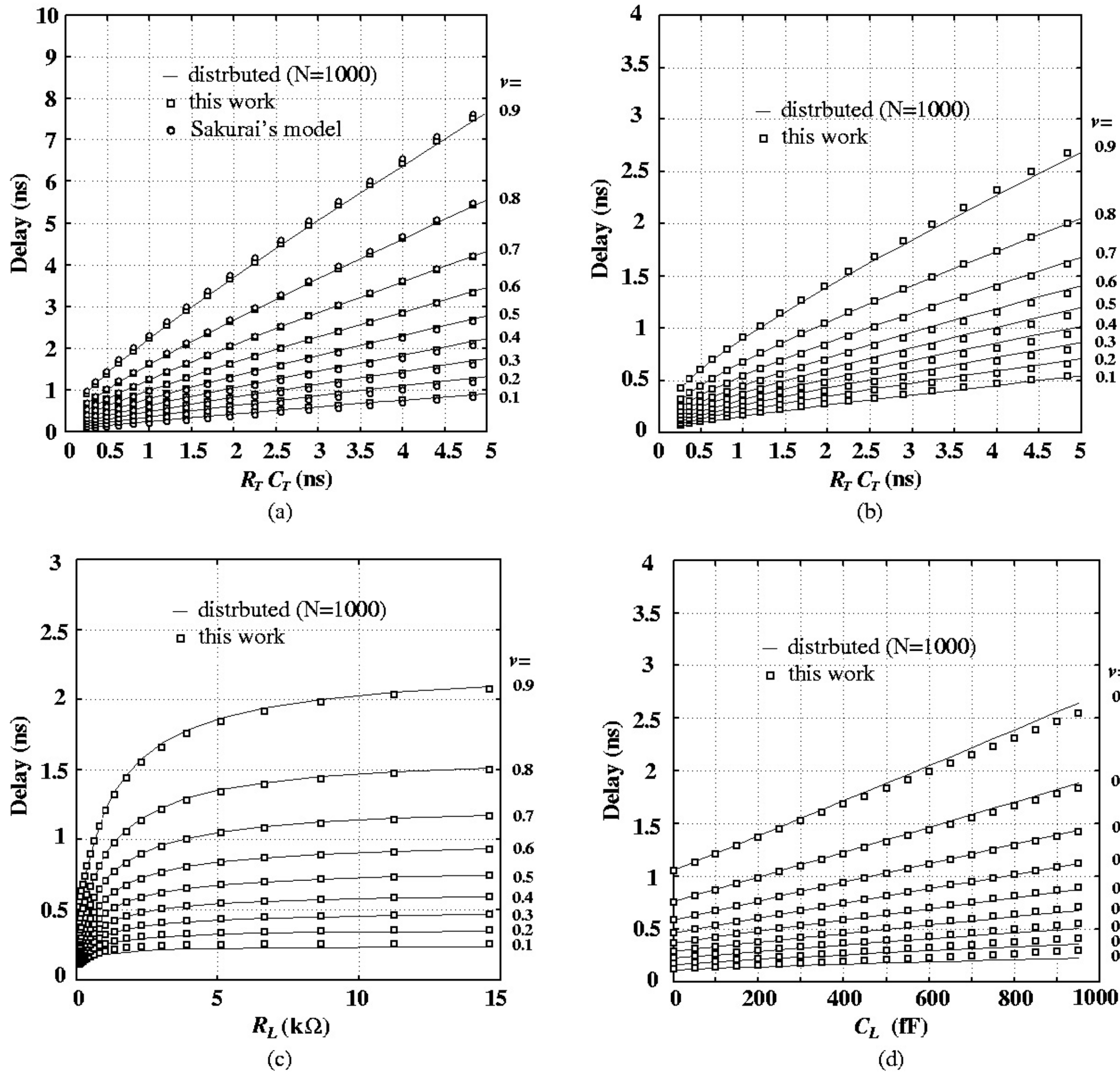


Figure 3. 5. Comparison between SPICE simulations for a distributed line with 1000 segments and Eq. (3-7). Parameter v represents the normalized threshold voltage at which the delay is calculated. (a) Voltage-mode (b) current-mode, R_L=500Ω. In (a) and (b), R_T and C_T are the total interconnect resistance and capacitance, respectively. Similarly, the delay is shown versus (c) R_L and (d) C_L. Unless, otherwise specified, R_S=250Ω and C_L=250fF.

3.4.2. Delay for slow ramp inputs

Propagation delay under linear ramp input excitation is typically desired since it closely approximates the real input conditions of digital CMOS circuits. Under steady-state conditions, the output response to a linear ramp input can be approximated by a delayed and scaled ramp signal waveform. This approximation is valid only when the rise time (t_R) of the ramp input is greater than the total steady-state time delay (t_{ST}) of the step response obtained in Eq. (3-7), that is, when $t_R > t_{ST}$ (the approximate t_{ST} value is obtained in section 3.4.3). In other words, the approximation is valid when the propagation delay is independent of the input rise time. Under these conditions, Eq. (3-7) can be approximated by the $R_{eff}C_{eff}$ product, which is expressed as:

$$\frac{t_{SR}}{R_T C_T} = \frac{\eta_L}{2}\left(1+\frac{1}{3R_{LT}}\right)+\eta_L R_{ST}\left(1+\frac{1}{2R_{LT}}\right)+\eta_S C_{ST}\left(1+R_{LT}\right)+\eta_L C_{LT}\left(1+R_{ST}\right) \quad (3\text{-}8)$$

It is interesting to note that when the effect of the source and load capacitance is not considered (i.e. $C_{ST} = C_{LT} = 0$), the result in Eq. (3-8) can be shown to be identical to [2].

3.4.3. Generalized Delay formulation for ramp inputs

A generalized delay approximation for ramp inputs with arbitrary rise time (t_R) can be extrapolated from equations (3-7) and (3-8) using a piece-wise linear fitting function. First, the boundary conditions for equations (3-7) and (3-8) are defined as $t_R = 0$ and $t_R > t_{ST}$, respectively, where t_{ST} was previously defined as the steady-state delay of the step response. It follows that the desired delay formulation for arbitrary rise time ramp signals falls in the shaded region depicted in Figure 3.6a, bounded by the propagation delays of the step and slow ramp output response waveforms. To obtain a generalized delay equation for ramp inputs, a piece-wise linear fitting function is used to weight equations (3-7) and (3-8) according to the rise time of the input ramp, as shown in Figure 3.6b. The dependence of the 50% ramp delay (t_{50R}) on rise time (t_R) are depicted in terms of the normalized formulations (t_{50R} / t_{50}) and (t_R / t_{50}), respectively.

It follows that in region I ($0 \le t_R / t_{50} < t_{ST} / t_{50}$), t_{50R} / t_{50} can be expressed as,

$$\frac{t_{50R}}{t_{50}} = 1 + \left(\frac{\frac{t_{SR}}{t_{50}} - 1}{\frac{t_{ST}}{t_{50}}}\right)\frac{t_R}{t_{50}} \quad (3\text{-}9)$$

From equations (3-7) and (3-8), equation (3-9) can be approximated by,

$$\frac{t_{50R}}{t_{50}} = 1 + \frac{0.3068 \cdot t_R}{t_{ST} \ln\left(2.4 + \frac{\eta_L}{R_{LT}}\right)} \quad (3\text{-}10)$$

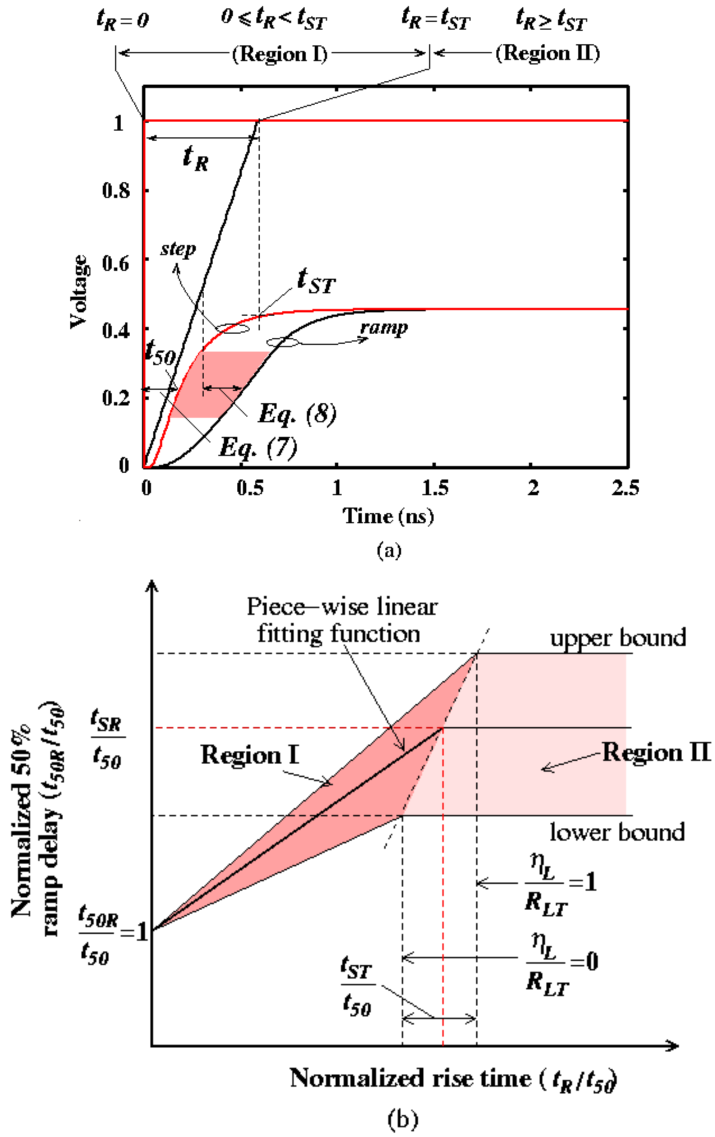


Figure 3. 6. Generalized delay formulation for ramp inputs with arbitrary rise time; (a) step and slow ramp response. The shaded area corresponds to the desired delay formulation. Regions of validity are defined to extrapolate the approximate 50% delay (t_{50R}) formula for ramp inputs. (b) A piece-wise linear approximation fitting function is used to model the dependence of t_{50R} on rise time t_R.

As shown in Figure 3.6b, the term t_{ST} determines the slope of the fitting function in region I, and hence the accuracy of Eq. (3-10). A good approximation to the steady state delay term (t_{ST}) was found via SPICE simulations to be $t_{ST} = t_{98}$ (i.e. recall that t_{98} is the time delay when the normalized output waveform reaches 98 % of its maximum value). Then, Eq. (3-10) can be rewritten in general form to satisfy both region I and II as follows,

$$
t_{50R} = \begin{cases} t_{50} + \dfrac{0.3068}{\ln\left(60 + 25\dfrac{\eta_L}{R_{LT}}\right)} \cdot t_R & \begin{matrix} 0 \le t_R < t_{98} \\ \text{(Reg. I)} \end{matrix} \\ t_{SR} & \begin{matrix} t_R \ge t_{98} \\ \text{(Reg. II)} \end{matrix} \end{cases} \tag{3-11}
$$

Since $0 \le \eta_L/R_{LT} \le 1$ holds true (i.e. η_L/R_{LT}=0 when RL=∞ and η_L/R_{LT}=1 for R_L=R_S=0), the range of t_{98} in Eq. (3-11) for which the boundary between region I and II are satisfied, can be expressed as $3.63t_{50} \le t_{98} \le 4.67t_{50}$. Notice that this finite range approximation to the steady-state delay time is simply used to provide a good fit of the proposed formula Eq. (3-11) to the entire continuous range of t_{50R} across a wide range of simulation parameters. The validity of the above mentioned fitting method can be verified in Figure 3.7, where each sample represents one out of a total of eighteen hundred SPICE simulations of a distributed line with 1000 subsections and various load and source terminations. The ranges of parameters simulated are 20fF ≤ C_L ≤ 1pF, 20fF ≤ C_S ≤ 250fF, 200 ≤ R_T ≤ 1kΩ, 1pF ≤ C_T ≤ 4pF, 100Ω ≤ R_S ≤ 500Ω, 100Ω ≤ R_L ≤ ∞. The calculation results by the proposed formula agree well with the SPICE simulation results across both regions I and II, as shown in the error scatter-plot in Figure 3.7b. A histogram of the overall error distribution is shown in Figure 3.7c. The mean error of Eq. (3-11) was found to be approximately 1.9% with a standard deviation of 1.5%.

Notice that the result in Eq. (3-11) does not require an iterative solution since all the variables can be derived entirely in closed form. This result is an improvement over existing RC delay models since it accurately models the distributed nature of transmission lines while taking into account the effect of both resistive and capacitive loads (R_L and C_L) with a complex driving source impedance (R_S and C_S) for arbitrary rise time (t_R) ramp inputs.

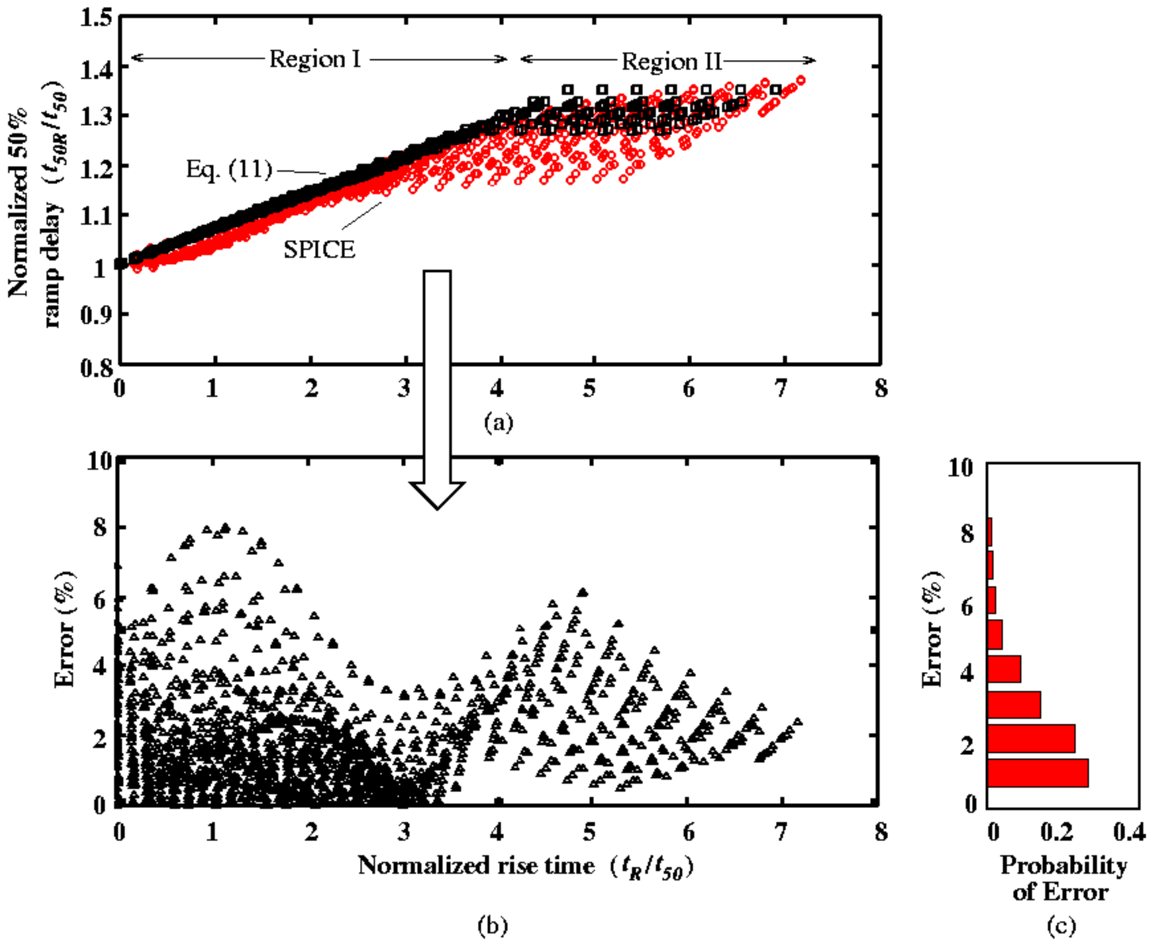


Figure 3. 7. Probability of error distribution of Eq. 3-11; (a) comparison between delay approximation using Eq. 3-11 and SPICE simulations of a distributed RC line with N=1000. A total of 1800 individual cases were simulated for a wide range of circuit parameters. (b) Error distribution of calculated and simulated delay for each individual data point and, (c) probability of error histogram. The mean error is determined to be 1.9% with a standard deviation of 1.5%.

3.5. Power dissipation model for current-mode circuits

There are three primary sources of power dissipation in current-mode circuits: static, dynamic and short-circuit power dissipation. Typically, the major component is the static power dissipation that arises from the constant current path from V_{dd} to ground via the resistive termination R_L, as shown in Figure 3.8. The dynamic power is dissipated when the capacitive components are charged through the PMOS device and discharged via the NMOS device. The third source of power dissipation arises from the finite input signal edge rates that result in short-circuit current. Generally, careful control of input edge rates can minimize the short-circuit current component to within 20% of the total dynamic power dissipation [15].

The dynamic power dissipation of current-mode circuits cannot be modeled by the well-known equation $P_{dyn}=V_{dd}^2C_T f$, since it assumes that all the capacitive components of the distributed RC line are charged to V_{dd}. As illustrated in Figure 3.8, the voltage at any point of a resistively terminated line will be less than V_{dd}, resulting in a smaller dynamic power dissipation component. To accurately model this effect, we find the steady state voltage at each node of an N-segment distributed RC line,

$$V_i = V_{dd}\eta_i \quad i = 0,1,2\ldots N \tag{3-12}$$

where the voltage drop factor η_i is calculated as

$$\eta_i = \frac{R_i}{R_L + R_S + R_T} \tag{3-13}$$

In Eq. (3-13), R_i is the resistance seen to the right of the ith node, expressed as

$$R_i = R_T + R_L - iR_u \tag{3-14}$$

where R_u is the unit segment resistance and R_T=NR_u. The dynamic power component at each node can be written as

$$P_i = (V_{dd}\eta_i)^2 C_i f \cdot act \tag{3-15}$$

C_i is the capacitance at each node and *act* is the switching activity factor. Hence, the total dynamic power dissipation can be expressed as

$$P_{dynamic} = \sum_{i=0}^{N} P_i = \sum_{i=0}^{N} (V_{dd}\eta_i)^2 C_i f \cdot act \tag{3-16}$$

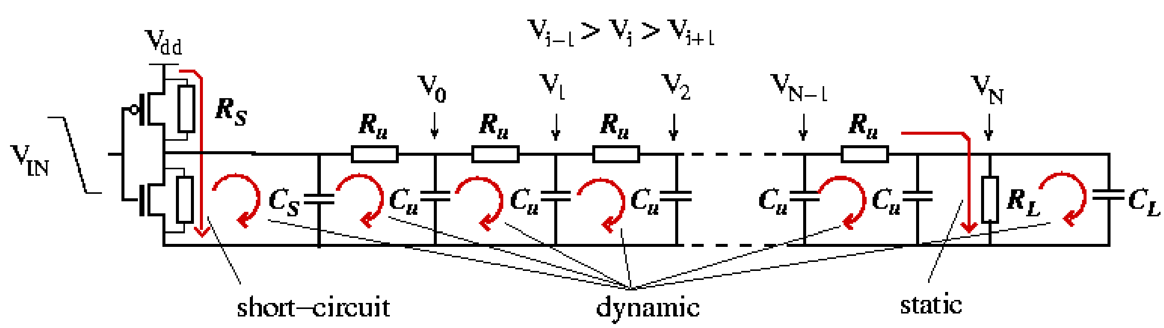


Figure 3. 8. Power dissipation breakdown of current-mode interconnect signaling.

Evaluating the summation in closed-form and letting N approach infinity results in the final expression for total dynamic power dissipation for current-mode circuits

$$P_{dynamic} = (\eta_L V_{dd})^2 f \cdot act \cdot \left\{ C_L + C_S \left[1 + \frac{2R_T}{R_L} + \left(\frac{R_T}{R_L} \right)^2 \right] + C_T \left[1 + \frac{R_T}{R_L} + \frac{1}{3} \left(\frac{R_T}{R_L} \right)^2 \right] \right\} \tag{3-17}$$

where η_L is given by Eq. (3-1c). Note that as R_L approaches infinity (i.e. voltage-mode), Eq. (3-17) reduces to the more familiar dynamic power dissipation formulation of voltage-mode circuits.

The static power dissipation component of current-mode circuits can easily be expressed as

$$P_{static} = \frac{\eta_L V_{dd}^2}{R_L} \tag{3-18}$$

In Eq. (3-18), as R_L approaches infinity, the static component reduces to zero. The short-circuit power dissipation component can be expressed as a fraction (χ) of the total dynamic power dissipation in Eq. (3-17)

$$P_{short-circuit} = \chi \cdot P_{dynamic} \tag{3-19}$$

where χ typically ranges from 0.1 to 0.2 [15].

3.6. Optimum line width

Let us consider the relationship between total delay and interconnection line width, where the former is determined by the sum of driver (R_S, C_S), interconnection (R_T, C_T) and load termination (R_L, C_L) delays. It can be easily shown that for a given interconnection line, there exists an optimal line width that minimizes the total delay. For the special case of voltage-mode signaling (i.e. $R_L = \infty$), the total delay t_{50} can be rewritten from Eq. (3-7) as,

$$\frac{t_{50}}{R_T C_T} = 0.37 + 0.74 \cdot (R_{ST} C_{ST} + R_{ST} C_{LT} + R_{ST} + C_{LT}) \tag{3-20}$$

Next, assume that the total line resistance and capacitance is given by,

$$R_T = R_{sh} \frac{l}{w} \tag{3-21b}$$

$$C_T = C_{area} wl + C_{fringe} l \tag{3-21b}$$

where R_{sh}, C_{area} and C_{fringe} are specific to process technology and are defined as the sheet resistance (Ω/sq), parallel plate area capacitance (F/m^2) and fringe capacitance (F/m) of the interconnection metal layer, respectively; l and w are the interconnect line length and width, respectively. Then, by minimizing Eq. (3-20) with respect to interconnection line width of the R_T and C_T terms, we have

$$w_{opt} = \sqrt{\frac{R_{sh}}{C_{area}}\left(\frac{C_{fringe}l + 2C_L}{2R_S}\right)} \tag{3-22}$$

Figure 3.9a shows the 50% delay for voltage mode sensing given by Eq. (3-20) versus line width for several loading capacitances. As predicted in Eq. (3-22), the optimal line width increases as the loading capacitance increases. Notice that the effect of fringe capacitance on the optimal line width (w_{opt}) can be neglected when C_L » $C_{fringe}\ l$.

A slightly more complicated yet generalized formulation for the optimal line width can be similarly derived for current-mode signaling by minimizing Eq. (3-7). For simplicity, only the terms inside the curly brackets {} in Eq. (3-7) are considered. Then, differentiating in terms of w and equating to zero, the following relationship is obtained,

$$6R_{Sl}R_{Ll}\left(R_{Sl}+R_{Ll}\right)w^4 + 12R_{Sl}R_{Ll}w^3 + \left[2\left(R_{Sl}+R_{Ll}\right) - 3R_{Sl}^2\left(C_{fa}+2C_{Sl}\right) - 3R_{Ll}^2\left(C_{fa}+2C_{Ll}\right)\right]w^2$$
$$-2C_{fa}\left(R_{Sl}+R_{Ll}\right)w - C_{fa} = 0 \tag{3-23a}$$

where R_{Sl}=R_S/($R_{sh}l$), R_{Ll}=R_L/($R_{sh}l$), C_{Ll}=C_L/($C_{area}l$), C_{Sl}=C_S/($C_{area}l$), and C_{fa}=C_{fringe}/C_{area}. Eq. (3-23a) can be easily solved numerically to yield the optimal line width (w_{opt}). A further simplification of Eq. (3-23a) is possible if the fringe capacitance is neglected. Then the following expression holds:

$$w_{opt} = \frac{1}{R_{Sl}+R_{Ll}} \cdot \left\{\sqrt{1 - \frac{\left(R_{Sl}+R_{Ll}\right)^2}{3R_{Sl}R_{Ll}} + \left(\frac{R_{Sl}+R_{Ll}}{R_{Sl}R_{Ll}}\right)\left(R_{Ll}^2C_{Ll} + R_{Sl}^2C_{Sl}\right)} - 1\right\} \tag{3-23b}$$

Although the fringe capacitance in any modern process technology should not be neglected, the approximation in Eq. (3-23b) exhibits an asymptotically improved accuracy when C_{Ll}, C_{Sl} » C_{fa} and R_{Ll}, R_{Sl} »0. The dependence of w_{opt} on R_S, R_L and C_L is shown in Figure 3.9b and 3.9c using Eq. (3-23a) and Eq. (3-23b) for ζ=0, 2 and 4, where ζ is defined as ζ= C_{fringe}/(C_{area} μm). Notice that for small values of R_L, Eq. (3-23a) should be used to accurately model the effect of fringe capacitance. Additionally, since no constraints are placed on the range of w_{opt}, it is to be understood that w_{opt} in Eq. (3-23) is valid only for values larger or equal to the minimum interconnect width (w_{min}) specific to the metal layer and process technology. Also, it is easy to verify that in the limit of $R_L = \infty$, Eq. (3-23) reduces to Eq. (3-22), as to be expected.

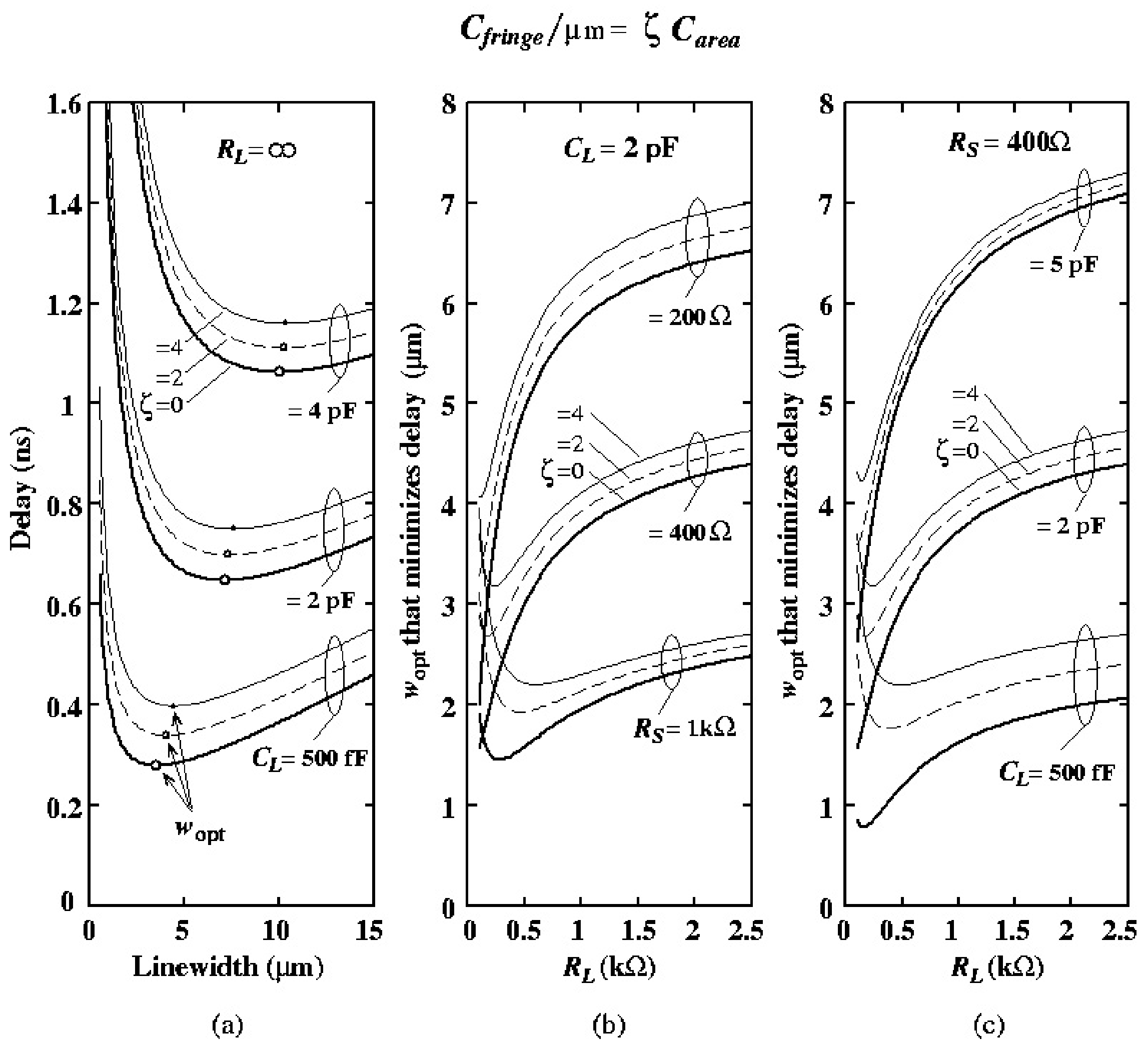


Figure 3. 9. Shown in (a) is the total delay versus line width, indicating that there exist an optimal width (w_{opt}) that minimizes delay. The optimum line width that minimizes delay versus load termination R_L is shown for several (b) R_S values and (c) C_L loading. Results are based on 0.18-μm copper process with M5 signal line and M3 shielding for a 10-mm long line (R_{sh}=0.07Ω/sq, C_{area}=14 aF/ μm2, and $C_{fringe}=\zeta$ C_{area} μm). Unless otherwise specified, R_S=200Ω and C_S=150fF.

One important result of Figure 3.9b is the reduction in optimum width (w_{opt}) for decreasing values of R_L, indicating that higher interconnection densities can be achieved with current-mode signaling. At a glance, this result may not be as intuitive and an attempt is made to explain this behavior. Lets first assume that $R_L = \infty$ and that Eq. (3-20) holds. A closer look at Eq (3-20) reveals that the only terms that depend on the interconnection width w are the product terms $R_S C_T$ and $R_T(C_L+C_{fringe}l)$ – i.e., these terms are obtained by expanding and multiplying out by $R_T C_T$. Thus the optimal width that minimizes delay can be approximately expressed as a function of the ratio $R_T(C_L+C_{fringe}l)/(R_S C_T)$, as described in Eq. (3-22). Next, as R_L is allowed to decrease and referring to Eq. (3-7), this ratio can be roughly approximated by $[R_T(C_L+C_{fringe}l)]/[R_S C_T(1+0.5R_T/R_L)]$, which decreases as R_L is reduced. Consequently, the reduction of R_L effectively increases the weighting factor on line capacitance C_T and as a result, the optimum width decreases to minimize the line capacitance.

The exact dependence of w_{opt} on delay parameters is more complex and can be approximated by Eq. (3-23). Notice that even the interconnection length l has an effect on the optimum line width. To summarize, as R_L is reduced the optimum line width generally decreases to minimize the capacitance of the line. This reduction in w_{opt} is more pronounced for small values of R_S. In Figure 3.9c, the optimum width is shown versus R_L for various C_L values. Notice that the reduction in optimum width is also prominent for large C_L values, indicating that current-mode signaling can be desirable for heavily loaded lines such as clock distribution networks.

3.7. Repeater insertion with current and voltage mode signaling

The most popular speed improvement technique used in global interconnect buses is the insertion of repeaters with full-swing voltage mode signaling. It is a well-known fact that repeaters improve the quadratic dependency of delay on interconnect line length to a less severe linear relationship [7]. Thus, most high performance systems such as microprocessors require repeating signals at the global level to some extent. For instance, the Intel Itanium™ microprocessor [16] fabricated in a 0.18-µm 6-metal layer process and with a total of 25.4 million transistors, requires approximately 85% of its global nets to be repeated. Of about 13,000 total nets at the top level approximately 11,000 require insertion of repeaters. Interestingly, the average number of repeater drop points per net is approximately 1.5 [16]. That is, only a sub-optimal number of repeaters were used in the physical design [16], [17].

Since optimal repeater insertion can seldom be achieved in an actual design due to the complexity that arises with its physical implementation (i.e. placement, area, power, noise), a comparison of sub-optimal/optimal repeaters with voltage and current mode signaling is made. Figure 3.10a shows the comparison benchmark used for the analysis, where k uniformly spaced drivers divide a long interconnect with total resistance R_T and total capacitance C_T. A possible implementation of the voltage and current mode repeaters and receiver are shown in Figure 3.10b and 3.10c, respectively. The generalized model in Figure 3.10d is used to approximate both voltage and current mode repeater circuits. Notice that a conversion delay (t_{cd}) is added to account for the low to high swing conversion delay penalty of the current-mode repeater.

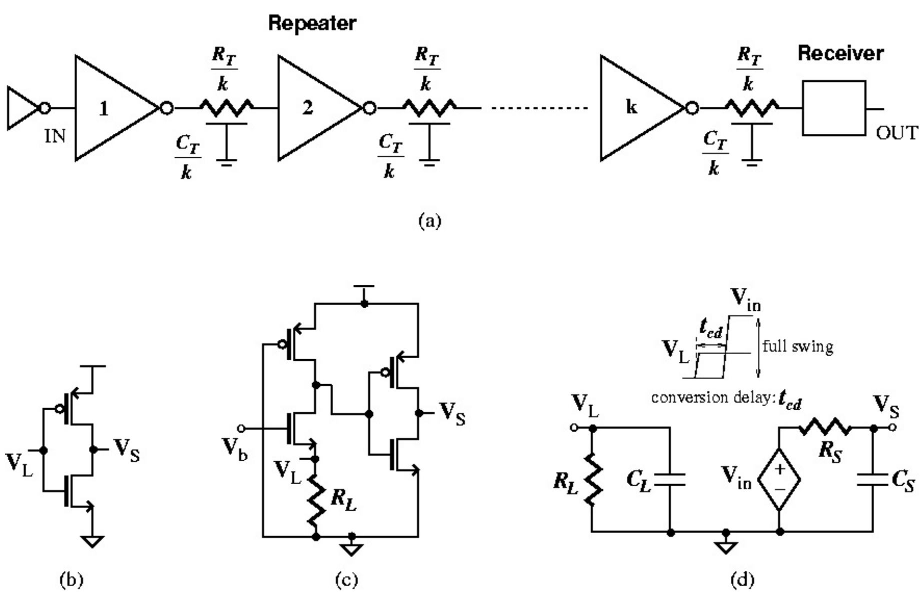


Figure 3. 10. Analysis benchmark for voltage and current mode repeater insertion (a), with repeaters/receiver circuit topologies for (b) voltage and (c) current sensing. (d) Simplified repeater circuit model with conversion delay (t_{cd}).

Figure 3.11 illustrates the delay latency, power dissipation and maximum throughput projections of voltage and current sensing repeaters. The results are based on equations (3-7) and (3-17)-(3-19) for a 0.18-μm process with interconnect resistance (R_T) and capacitance (C_T) varied from 50-1000Ω and 100fF-2pF, respectively. Other pertinent parameters are defined as follows: V_{dd}=1.8V, Wp/Wn=3, activity factor act=0.5, short-circuit dissipation factor χ=0.1, h_{opt}=50(Wp/Wn) and R_L=100Ω. h_{opt} is the optimal driver size and Wp(Wn) is the minimum size width of the pMOS(nMOS) transistor.

In Figure 3.11a, the delay latency of an unrepeated interconnect line with a single current-mode (CM) receiver exhibits comparable performance to a repeated line with k uniformly spaced voltage-mode (VM) drivers (i.e. k-1 repeaters). The optimal number of drivers (k_{opt}), is derived by minimizing Eq. (3-20) with respect to k, which yields,

$$k_{opt} = \sqrt{\frac{R_T C_T}{2(C_L + C_S)R_S}} \tag{3-24}$$

where R_T and C_T are given by Eq. (3-21). Notice that the inclusion of the source capacitance (C_S) in the delay formulations presented in this paper yields a lower k_{opt} than [7]. Figure 3.11a shows that the conversion delay (t_{cd}) overhead due to low-to-high swing conversion of the CM repeaters/receiver can have a noticeable impact on the overall latency of a repeated line.

Figure 3.11b shows the total power dissipation (P_T) (i.e. dynamic + static + short-circuit) dependence on interconnect delay R_TC_T for several k values and a target data frequency of 2GHz (i.e. 4Gb/s NRZ data rate). In the CM sensing case, for short interconnect lines (i.e. R_TC_T < 0.5ns), a large static power is dissipated due to the current path from the power supply to ground via R_L, resulting in large overall power dissipation (P_T). This indicates that CM signaling should be used for long global resistive interconnects to minimize the static power dissipation. However, the minimum power dissipation (P_{Tmin}) that is required to achieve a target data frequency of 2GHz is smaller when CM repeaters are used. For simplicity, in the CM sensing projections, the power dissipated in the driver, interconnect and receiver is assumed to be larger than the power dissipation of the level conversion circuit. Any additional power dissipated in the CM level conversion circuits can be added to the projected trends. In general, the result of Figure 11(b) can be attributed to the reduction in dynamic power dissipation due to the lower voltage swing and a fewer number of CM repeaters. As shown in Figure 3.11b, CM signaling can achieve the target data rates with nearly 1/3 the number of VM repeaters. This trend is illustrated in Figure 3.11c, where the improvement in maximum throughput (f_{max}) of the CM sensing scheme is apparent. For instance, at R_TC_T=1ns, 3 CM repeaters achieves nearly 4.8Gb/s more NRZ bandwidth than 3 VM repeaters, and exhibits the same data rate performance as 9 VM repeaters. This comparison is extended in Figure 3.11d to illustrate the dependence of the power and data rate ratio (P_T/f_{max}) on the number of repeaters (k-1). Again, the advantage of CM mode sensing over VM signaling is more evident for long interconnections (i.e. R_TC_T =2ns).

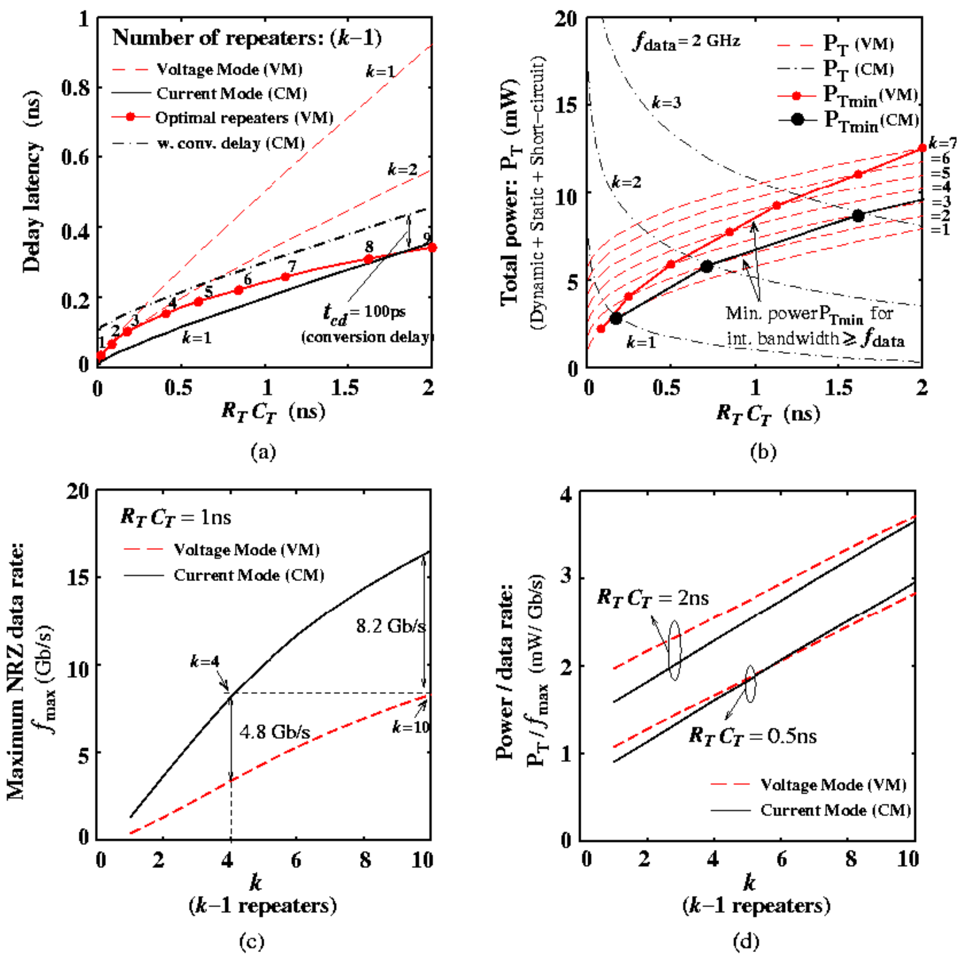


Figure 3. 11. Delay, power and maximum throughput comparison of current (CM) and voltage (VM) mode signaling. (a) Delay latency vs. R_TC_T for sub-optimal/optimal voltage mode repeaters and a single current mode driver (k=1). (b) Total power dissipation for various k values assuming a 2GHz data frequency (f_{data}). To determine the minimum power dissipation (P_{Tmin}) curves, the bandwidth of each repeated section must be $\geq f_{data}$. That is, the steady state delay (t_{ST}) defined in section 3.4.3, must be satisfied (i.e. $t_{ST} \leq 1/f_{data}$) for each of the interconnect line segments across all k values. (c) Maximum NRZ data throughput and (d) power data rate ratio vs. number of VM and CM repeaters.

3.8. Delay Overshoot due to finite edge rates

The analysis in the previous sections assumes a linear resistance (R_S) and capacitance (C_S) model for the interconnect line driver (i.e. inverter). However, when an inverter is driven by a sharp input edge, the input-to-output or miller capacitance (C_M) of the inverter causes a voltage overshoot at the output, which may add to the overall delay. This effect is shown in Figure 3.12, indicating that the recovery time (t_{miller}) of this overshoot voltage can contribute appreciably to the overall delay time [18]. An offset delay term (t_{offset}) can be used to account for this effect,

$$t_{offset} = \left\{ \frac{t_R}{2}\left(1+\frac{V_T}{V_{DD}}\right) + \frac{2C_M}{C_{eff}}\left[0.37 + 0.74\left(R_{ST}C_{ST} + R_{ST}C_{LT} + R_{ST} + C_{LT}\right)\right] \right\} - \alpha \cdot t_o \tag{3-25}$$

where C_{eff} is the effective load capacitance seen by the driver, given by Eq. (3-1b). The first term in {} brackets of Eq. (3-25) represents the recovery time (t_{miller}) and t_0 (i.e. t_v when v=0) is the zero crossing delay time of Eq. (3-7). The term α is used as a pseudo-empirical fitting factor that can be extracted from simulations, typically ranging $0 \leq \alpha \leq 2$. The overall delay (t_{dv}) can be approximated from equations (3-7) and (3-25) as,

$$t_{dv} = t_v + t_{offset} - v \cdot t_R \qquad t_R < t_{miller} \tag{3-26}$$

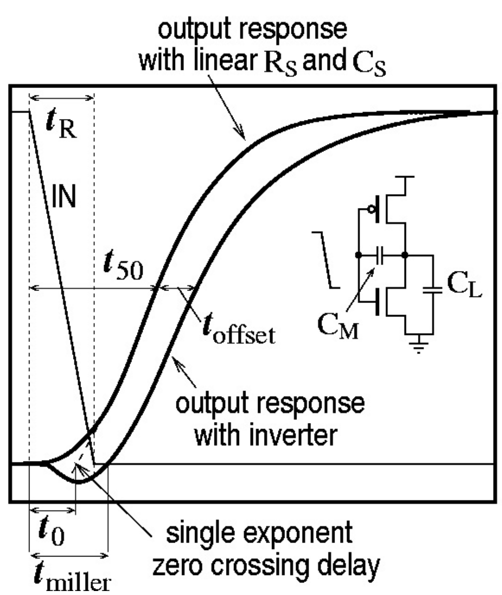


Figure 3. 12. The miller effect due to a sharp transition at the input causes add recovery time (t_{miller}).

3.9. Experimental Results

The test chip fabricated in AMI 1.6μm bulk CMOS technology consists of six parallel-terminated test lines to experimentally verify the proposed model. Figure 3-13 shows the 2x2 mm^2 die micrograph with model 10 and 34A GGB probes. Due to the small die area, poly interconnects were used to model the effect of long metal RC interconnects (l >10mm). Parameter extraction of resistance and capacitance was performed using a 4275A LCR meter, with measured values given as follows: R_T=500Ω, C_T=1.63pF, C_L=771fF, R_L=4.44kΩ, 2.22k Ω, 1.11k Ω, 555Ω, 277Ω, and an open ended termination for voltage mode signal. The driver is an inverter with the PMOS and NMOS sized as Wp/Lp = 5x(48μm/3.2μm) and Wn/Ln = 5x(14.8μm/3.2μm), respectively. The extracted linear driver resistance and source and miller capacitance is R_S=250Ω, C_S=1.3pF and C_M=350fF, respectively.

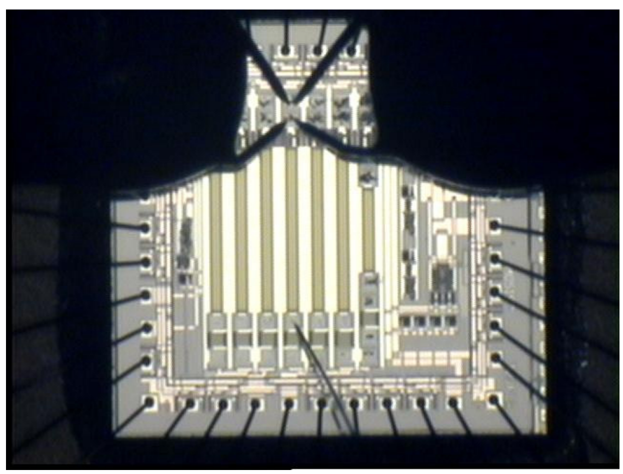


Figure 3. 13. Chip micrograph for delay formula verification.

The test setup used to create a fast input edge (i.e. at least 5V for AMI 1.6μm devices) and measure the input/output waveforms is shown in Figure 3.14. A fast input edge is created using a PSPL 4015B output driver, a 3dB attenuator and a 5541A bias-T. The PSPL 4015B driver creates a 30ps –9V output edge, which is attenuated and DC biased to create a sharp falling input edge (i.e. approximately 5V to –1V). The input edge is measured to be approximately 300ps with a high-impedance model 34A GGB probe. In order to de-embed the measured waveforms from the probing and testing equipment, a single probe is used to measure both input and output waveforms. The measurements vs. SPICE simulation and model results are shown in Figure 3.15. Measurements of the input and output waveforms were made using a TEK 11801A oscilloscope with an SD-24 sampling head. The measured result is seen to agree well with the proposed model. The accuracy of the measurements is within 100ps, limited mainly by the bandwidth of the model 34A probes.

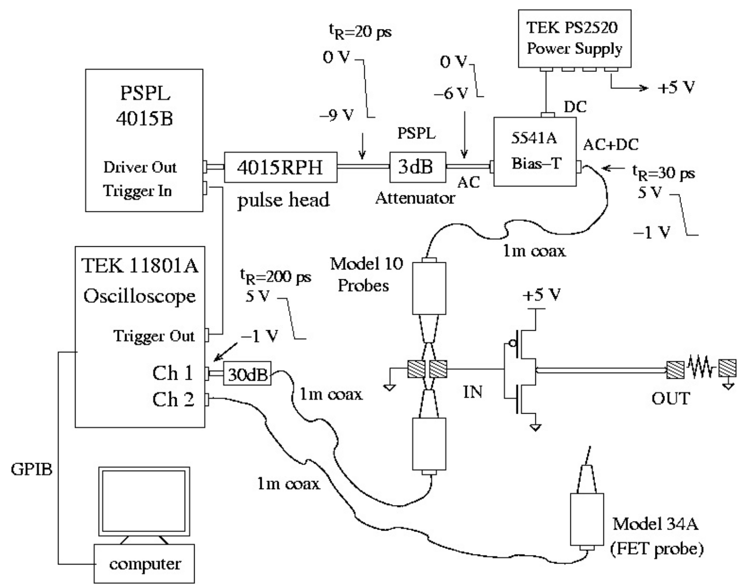


Figure 3. 14. Experimental setup for delay measurement.

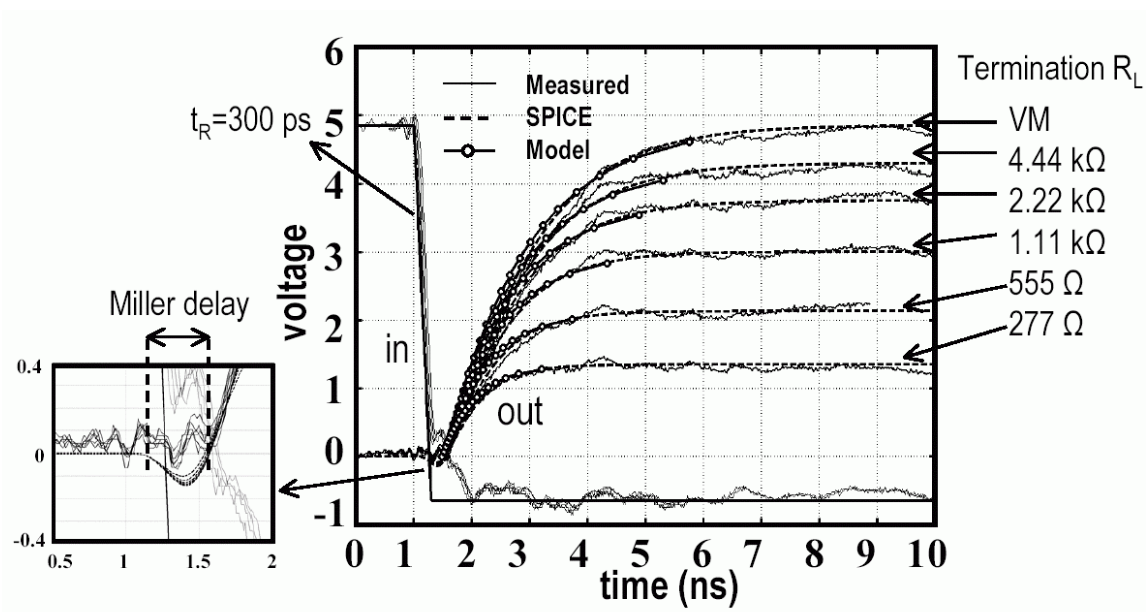


Figure 3. 15. Measured and theoretical results of delay analysis.

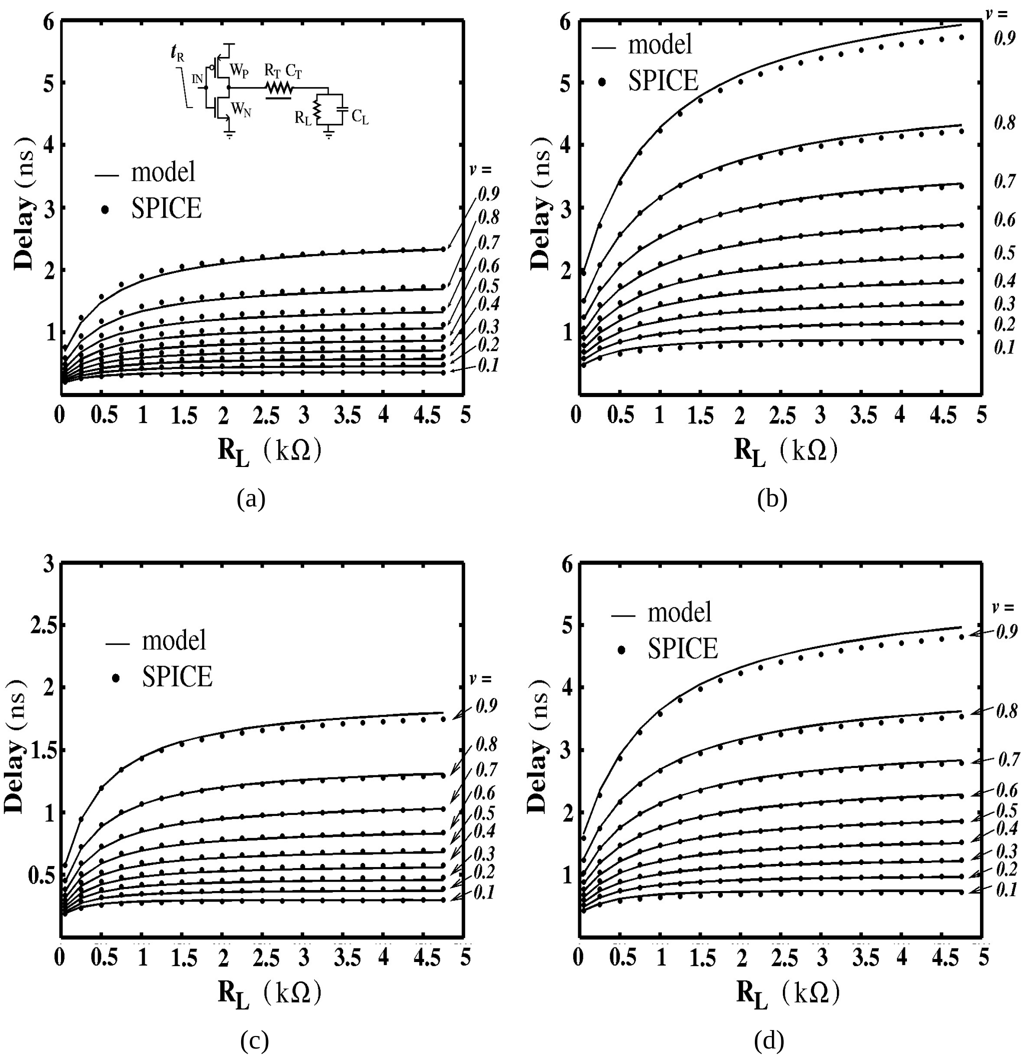


Figure 3. 16. SPICE simulations vs. model results for TSMC 0.35μm Al process with M3 line between M4 and M2; w=2um, Lp=Ln=0.4μm and Wp/Wn=k (8μm/3μm). The input edge is t_R=100ps. (a) l=10 mm, k=5 (b) l=20 mm, k=5 (c) l=10 mm, k=10 and (d) l=20 mm, k=10.

Further model comparisons are made in sub-micrometer process via SPICE simulations. Figure 3.16 shows simulation vs. model comparison results in TSMC 0.35μm Al process for 2μm wide metal-3 lines that are 10mm and 20mm in length, sandwiched between metal-4 and metal-2 planes. The accuracy of this model is over 90% for the simulated range of parameters. Notice that as the resistive load termination (R_L) increases, the delay asymptotically approaches that of a driven interconnect with a purely capacitive load (C_L) termination (i.e. voltage mode). To achieve reasonable delay reduction, R_L should be chosen such that $R_L \ll R_S+R_T$. An estimate of the maximum NRZ data throughput in TSMC 0.35 μm is shown in Fig 3.17. For simplicity, it is assumed that the data throughput is equivalent to $1/t_{90}$ (i.e. the reciprocal of 0-90% delay). The results indicate over a 2x increase in data bandwidth as R_L is decreased.

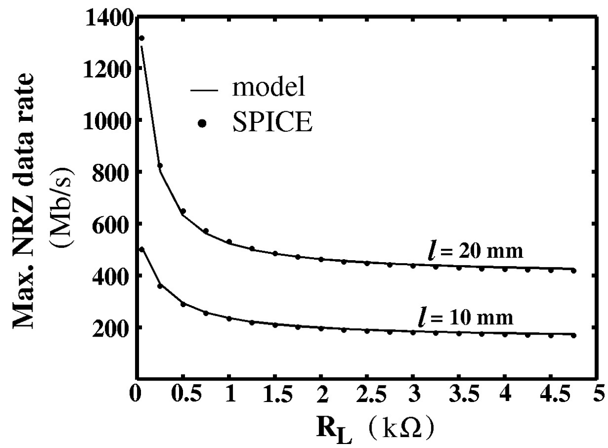


Figure 3. 17. Maximum NRZ data rate vs. R_L. Driver size is Wp/Wn=5 x (8μm/3μm).

3.10. Summary

A simple closed-form expression for the effective resistance and capacitance of a driven distributed RC line with resistive termination is presented. Based on this model, an analytical closed-form delay expression under step input excitation is derived. The formulation is valid for both voltage and current mode signaling and exhibits an error less than 5% for a large set of simulated circuit parameters. An extension to this formulation for ramp inputs with arbitrary rise time exhibits a mean error of approximately 1.9% and standard deviation of 1.5%. To model the power dissipation in current-mode sensing schemes, a new closed-form expression for dynamic power dissipation is presented to accurately predict the reduction in dynamic power due to the voltage drop along the resistively terminated interconnect line. In addition, the optimum interconnect line width that minimizes the total delay is formulated. The optimum line width for a current mode bus is shown to be typically much smaller than that of voltage mode signaling. This reduction in optimum line width is found to be more prominent when the driver source resistance (R_S) is small and the load capacitance (C_L) is large.

The effect of fast input edge and input-to-output capacitance of a CMOS inverter is considered. Analysis shows less then 10% error in model uncertainty for a wide range of parameters. Results from experimental measurements of an AMI 1.6μm test chip is shown to agree with the proposed model. The proposed model can be used to estimate delay and bandwidth performance of both current and voltage mode signaling schemes.

The key to current-mode signal transporting is the shift in pole position and reduction of the system time constants that result from sensing signals with low impedance nodes. The analysis presented herein shows that current-mode signal transporting techniques can be used to enhance the overall delay and bandwidth performance in long global on-chip interconnects.

Using the delay and power expressions, a comparative study of delay latency, power dissipation and maximum throughput is presented for voltage and current mode repeaters. It is shown that a single current-mode driver achieves comparable performance to the optimally spaced voltage mode repeaters. An increased margin of interconnection delay latency can be used to accommodate the low-to-high swing delay conversion of the current-mode repeater/receiver, while maximizing the data throughput. It is found that current mode signaling is best suited for long resistive interconnects to minimize the static power dissipation, which can be significant for short interconnects. It has been shown that the increase in NRZ data throughput for current sensing repeater insertion schemes can be significantly larger than the voltage mode signaling schemes. In addition, current-mode signaling can be used to meet the increasingly difficult target data frequencies with a fewer number of repeaters, minimizing the complexity associated with inserting a large number of repeaters in critical delay paths. Perhaps, a hybrid repeater insertion approach, with voltage mode repeaters for intermediate interconnects and current mode repeaters for long critical interconnections can be envisioned in future deep sub-micron systems.

3.11. References

[1] Semiconductor Industry Association, *The International Technology Roadmap for Semiconductors*, 2001 Edition.

[2] E. Seevinck, P. van Beers, 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.

[3] M. Izumikawa, M. Yamashina, “A current direction sense technique for multiport SRAM’s,” *IEEE J. Solid-State Circuits*, vol. 31, no. 4, pp. 546-551, April 1996.

[4] T. Blalock, R. Jaeger, “A high-speed sensing scheme for 1T dynamic RAM’s utilizing the clamped bit-line sense amplifier,” *IEEE J. Solid-State Circuits,* vol. 27, no. 4, pp. 618-625, April 1992.

[5] W.C. Elmore, “The transient response of damped linear networks with particular regard to wide-band amplifiers,” *J. Applied Physics*, vol. 19, no. 1, pp. 55-63, Jan. 1948

[6] J. Rubinstein, P. Penfield, M. Horowitz, “Signal Delay in RC Tree Networks,” *IEEE Trans. Computer-Aided Design,* vol. 2, no. 3, July 1993, pp. 202-210.

[7] H.B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*. Reading, MA: Addison-Wesley, 1990.

[8] T. Sakurai, “Closed-form expression for interconnection delay, coupling, and cross-talk in VLSI,’s” *IEEE Trans. Electron Devices,* vol. 40, pp. 118-124, Jan. 1993.

[9] H. Schmid, “Why the terms ‘current-mode’ and ‘voltage-mode’ neither divide nor qualify circuits,” IEEE International Symp. On Circuits and Systems, 2002, pp. 29-32.

[10] Y. Cong, R.L. Geiger, “Performance association and discrimination between current-mode and voltage mode operation of monolithic linear circuits,” Southwest Symposium on Mixed-Signal Design, 2001.

[11] C. Tomazou, F.J.Lidgey, D.G. Haigh, Analogue IC design: the current-mode approach, vol 2 of IEE Circuit and Systems Series 2, Peter Perigrinus Ltd. 1990.

[12] W.J. Dally, J.W. Poulton, Digital Systems Engineering. Cambridge, UK: Cambridge Univ. Press, 1998.

[13] L.T. Pillage, R.A. Rohrer, “Asymptotic waveform evaluation for timing analysis,” IEEE Trans.Computer Aided Design, vol. 9, no. 4, pp. 352-366, April 1990.

[14] R. Achar, M.S. Nakhla, “Simulation of high-speed interconnects,” Proc. of the IEEE, vol. 89, no. 5, pp. 693-728, May 2001.

[15] M.K. Gowan, L.L. Biro, D.B. Jackson, “Power considerations in the design of the Alpha 21264 microprocessor,” Proc. Design Automation Conference, pp. 726 – 731, 1998.

[16] R. McInerney, et al. “Methodology for repeater insertion management in RTL, layout, floorplan and fullchip timing databases of the ItaniumTM microprocessor,” Proc. Int. Symp. Physical Design, pp. 99-104, 2000.

[17] S. Rusu, G. Singer, “The first IA-64 microprocessor,” IEEE J. Solid-State Circuits, vol. 35, no. 11, pp. 1539-1544, Nov. 2000.

[18] Daga J.M, Auvergene D., “A comprehensive Delay macro Modeling for Submicrometer CMOS Logics,” IEEE JSSC, V34, N. 1, January 1999.

Chapter IV

Band-limited Signaling for Cross-talk Reduction

4.1. Introduction

The current interconnect scaling trend in CMOS technology suggest that mitigating the effect of increasing wiring resistance on propagation delay, exemplified by high metal aspect ratios, is of significant relevance in high performance VLSI [1]. However, higher aspect ratios combined with decreasing metal pitches with each technology generation will progressively exacerbate the signal integrity problem due to an increase in coupling capacitance between adjacent conductors. Therefore, its is likely that the cross-talk noise due to coupling effects will become increasingly important and will eventually become a dominant problem over intermediate and global propagation delays. The potential problem due to cross-talk is augmented given the reduced noise margins in modern high performance non-static logic circuits (i.e. dynamic circuits with multiple phase clocks and pre-charge logics), as well as the decrease in device threshold voltages.

To reduce cross-talk noise due to capacitive and inductive effects, we propose a signaling scheme based on band-limited basis waveforms to partially replace the fast edge-rate pulses in critical high-density interconnect communication links within the system. The fundamental difference between this technique and other solutions is that band-limited signaling for critical paths results in the minimization of the cross-talk noise source itself. If we operate under the assumption that signaling across global interconnect tracks result in several clock cycles of latency [2], or that the processing margin is available, the increase in propagation delay due to band-limited signaling can be acceptable. The rationale for this is that in highly integrated large-scale systems, the latency in global propagation delay is unavoidable and will continue to worsen with future technology nodes [1]. In contrast, the penalty in fault detection errors due to cross-talk noise coupling effects is usually unacceptable.

In order for band-limited signaling to be efficiently deployed on silicon, its implementation should have minimal circuit overhead and must be compatible with repeater insertion methodologies used in high-performance designs. Moreover, the band-limited approach can be designed to support current-mode schemes to minimize cross-talk noise and maximize signaling bandwidth.

In this chapter, we investigate the effect of band-limited signaling on cross-talk noise and address relevant issues such as delay and power dissipation related to the proposed signaling scheme. In section 4.2, the band-limited signaling scheme is defined and a cross-talk noise model for current/voltage mode coupled on-chip interconnects is analyzed in both frequency and time domain. Theoretical expressions are derived to evaluate the cross-talk noise waveforms due to an aggressor line. In section 4.3, circuit implementation issues of a band-limited repeater are discussed. A test chip fabricated in AMI 1.6μm bulk CMOS technology is used to experimentally verify the effect of band-limited pulses on cross-talk noise. Concluding remarks are presented in section 4.4.

4.2. Band-Limited Signaling

4.2.1. Definition

If we recall, data transfer in digital communication systems involves the construction of orthogonal basis functions and mapping these functions into a vector space to form a signal constellation [3]. The vector space of a baseband pulse amplitude modulation (PAM) signal can be represented as,

$$x(t) = \sum_{n} I_n \cdot g\left(t - nT_b\right) \tag{4-1}$$

In Eq. (4-1), the I_n*'s* are the transmitted symbols, $g(\cdot)$ is the basis function and T_b is the bit time. From a circuit implementation perspective, due to its simplicity in generation and moderate complexity in detection, $g(\cdot)$ is typically chosen to be a constant throughout the bit time (i.e. square pulses). In band-limited signaling, the basis function $g(\cdot)$ is modified to improve its spectral characteristics for transmission in long global interconnects. Since typical implementations of pulse-shaping filters for digital communication systems [4] incur in large power dissipation and speed limitations, an alternative approach is required. In this work, the basis waveform $g(\cdot)$ is generated instead by approximating a Raised Cosine pulse at 100% excess bandwidth with a reduced edge-rate signal. For modeling purposes, the reduced edge-rate signal transitions can be represented by a half-cycle of a sinusoid to take into account for the soft turn on/off transitions of MOS transistors. An example of the waveform approximation is depicted in Figure 4.1, referred to as Raised Cosine Approximation (RCA) from hereon.

The RCA pulse $g_{RCA}(\cdot)$ is defined as,

$$g_{RCA}(t) = \frac{A}{2}\left(1 + \cos\left(\frac{\pi}{T_b}\left(t - T_b\right)\right)\right) \quad 0 \le t < 2T_b \tag{4-2}$$

Its Fourier transform is easily derived and expressed as,

$$G_{RCA}(f) = \frac{A}{2}\left(\frac{\sin\left(2\pi f T_b\right)}{\pi f\left(1 - 4f^2 T_b^2\right)}\right) \cdot e^{-j2\pi f T_b} \tag{4-3}$$

It can easily be shown that the RCA pulse is the frequency *dual pair* of the RC filter at 100% excess bandwidth. The spectrum of the RCA pulse contains tails or side-lobes that decay inversely as f^{6}, whereas the spectrum of a rectangular pulse decays at a rate inversely proportional to f^{2} [3].

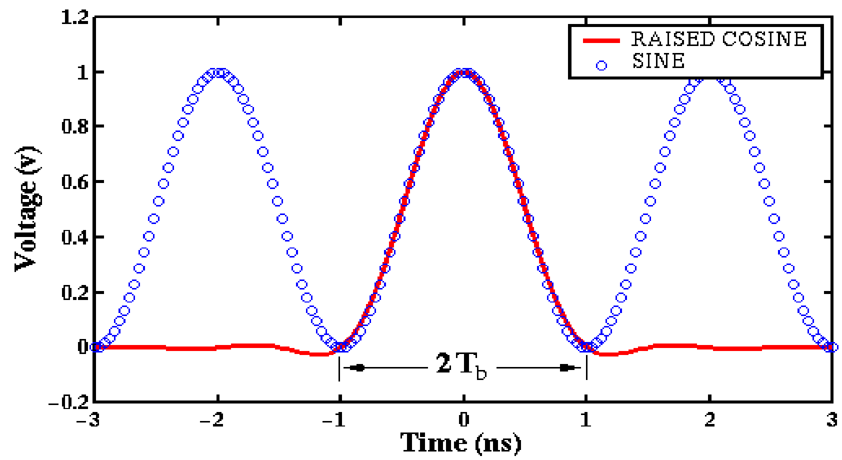


Figure 4. 1. Raised Cosine Approximation (RCA) waveform.

4.2.2. Cross-talk Noise Model

4.2.2.A. Frequency domain Analysis

Using the previously derived effective resistance and capacitance in Eq. (3-1), a simple model can be obtained for two coupled interconnect lines of length l, as shown in Figure 4.2a. The approximating lumped element model is shown in Figure 4.2b, where C_C and C_{Ei} (i=1,2) are the effective coupling and vertical capacitances of interconnects, respectively, and R_{Ei} is the effective resistance defined in Eq. (3-1a). Results obtained from extensive HSPICE simulations suggest that a good approximation to the effective coupling capacitance is $C_C= 0.85C_{cu}\, l$, where C_{cu} is the distributed unit-length coupling capacitance. The driver and receiver circuits are model with linear resistance and capacitance values previously defined as R_S, C_S, R_L and C_L. Note that the input loads of the signal and quiescent line receivers may be resistive or capacitive depending on the signal transporting technique (i.e. voltage or current mode).

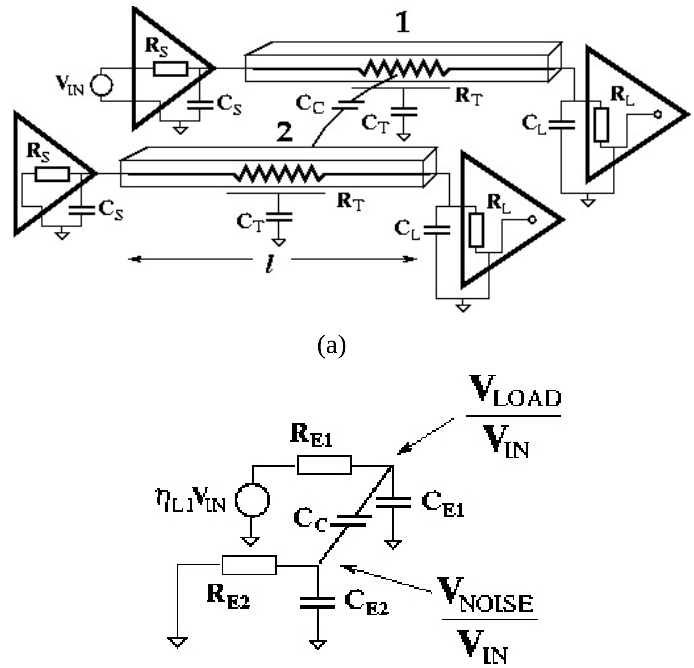


Figure 4. 2. Equivalent cross-talk model for two adjacent drivers; (a) driver/interconnect models (b) equivalent circuit model.

Since the worst-case cross-talk for long on-chip interconnects occurs at the receiver end [5], the signal and cross-talk noise transfer functions given by $V_{NOISE}(s)/V_{IN}(s)$ and $V_{LOAD}(s)/V_{IN}(s)$, respectively, are derived,

$$\frac{V_{LOAD}(s)}{V_{IN}(s)} = \frac{\left[1 + s \cdot \eta_{L2} R_{E2}\left(C_{E2} + C_C\right)\right] \cdot \eta_{L1}}{\left\{\begin{matrix} s^2\left[\eta_{L1}\eta_{L2}R_{E1}R_{E2}\left(C_{E1}C_{E2} + C_{E1}C_C + C_{E2}C_C\right)\right] + \\ s\left[\eta_{L1}R_{E1}\left(C_{E1} + C_C\right) + \eta_{L2}R_{E2}\left(C_{E2} + C_C\right)\right] + 1 \end{matrix}\right\}} \tag{4-4a}$$

$$\frac{V_{NOISE}(s)}{V_{IN}(s)} = \frac{\left(\eta_{L1}\eta_{L2}R_{E2}C_C\right) \cdot s}{\left\{\begin{matrix} s^2\left[\eta_{L1}\eta_{L2}R_{E1}R_{E2}\left(C_{E1}C_{E2} + C_{E1}C_C + C_{E2}C_C\right)\right] + \\ s\left[\eta_{L1}R_{E1}\left(C_{E1} + C_C\right) + \eta_{L2}R_{E2}\left(C_{E2} + C_C\right)\right] + 1 \end{matrix}\right\}} \tag{4-4b}$$

The resulting transfer functions $V_{LOAD}(s)/V_{IN}(s)$ and $V_{NOISE}(s)/V_{IN}(s)$ are depicted in Figure 4.3. Note that $V_{LOAD}(s)/V_{IN}(s)$ can be viewed as the sum of two separate transfer functions composed of a *low-pass* (i.e. $V_{LP}(s)/V_{IN}(s)$) and a *band-pass* (i.e. $V_{BP}(s)/V_{IN}(s)$) component (shown in dashed lines), such that,

$$\frac{V_{LOAD}(s)}{V_{IN}(s)} = \frac{V_{LP}(s)}{V_{IN}(s)} + \frac{V_{BP}(s)}{V_{IN}(s)} \tag{4-5}$$

The objective of the following frequency domain analysis is to gain some understanding into the frequency sensitivity of a quiescent line to cross-talk noise from an adjacent aggressor. To model the frequency and cross-talk sensitivity of the quiescent line, we introduce parameters,

$$\alpha_f = \frac{f_o - f_{3dB}}{f_o} \tag{4-6a}$$

$$\beta_c = \frac{C_C}{C_C + C_T} \tag{4-6b}$$

where α_f and β_c are defined as the *frequency offset* and *cross-talk* factors, respectively. In (4-6), f_o and f_{3dB} are the maximum gain and 3dB frequencies of $V_{NOISE}(s)/V_{IN}(s)$ and $V_{LOAD}(s)/V_{IN}(s)$ transfer functions, respectively. The parameter α_f is simply a measure of the frequency deviation between the peak noise frequency (i.e. f_o for $V_{NOISE}(s)/V_{IN}(s)$) and 3dB signal frequency (i.e. f_{3dB} for $V_{LOAD}(s)/V_{IN}(s)$), whereas β_c represents a loss factor determined by the aspect ratio of the parallel interconnects.

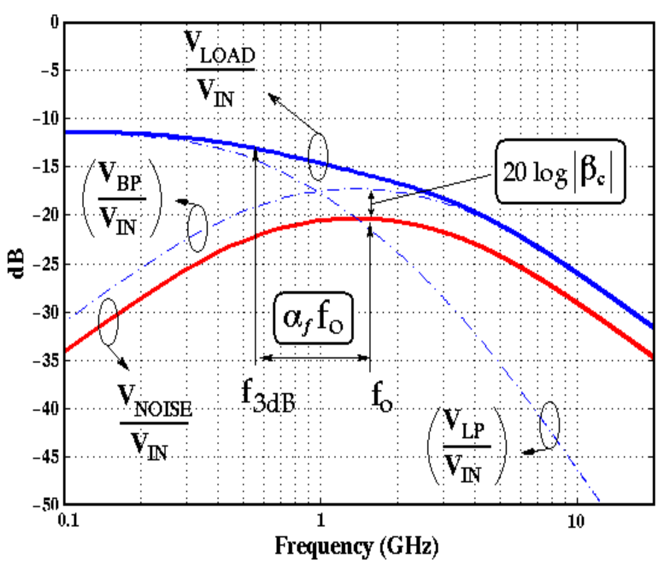


Figure 4. 3. Frequency response of aggressor and victim lines for coupled lines in Al 0.18-um process with ground planes above and below (W=1-um, spacing=0.28-um).

For the special cases when both interconnect lines are equal, the 3-dB frequency (f_{3dB}) can be expressed in terms of f_o and β_c in the following form,

$$f_{3dB} = f_o \left[\sqrt{\left(\frac{\beta_c^2}{1-\beta_c^2}\right)^2 + 1} - \left(\frac{\beta_c^2}{1-\beta_c^2}\right) \right]^{\frac{1}{2}} \tag{4-7}$$

From (4-7), f_{3dB} can also be written as,

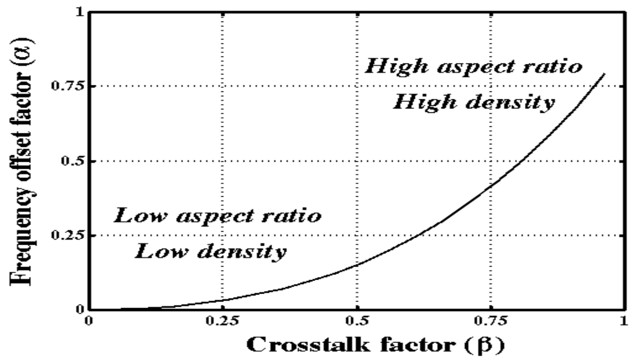
$$f_{3dB} = f_o \left(1 - \alpha_f\right) \tag{4-8}$$

which results in the following expression for α_f in terms of β_c:

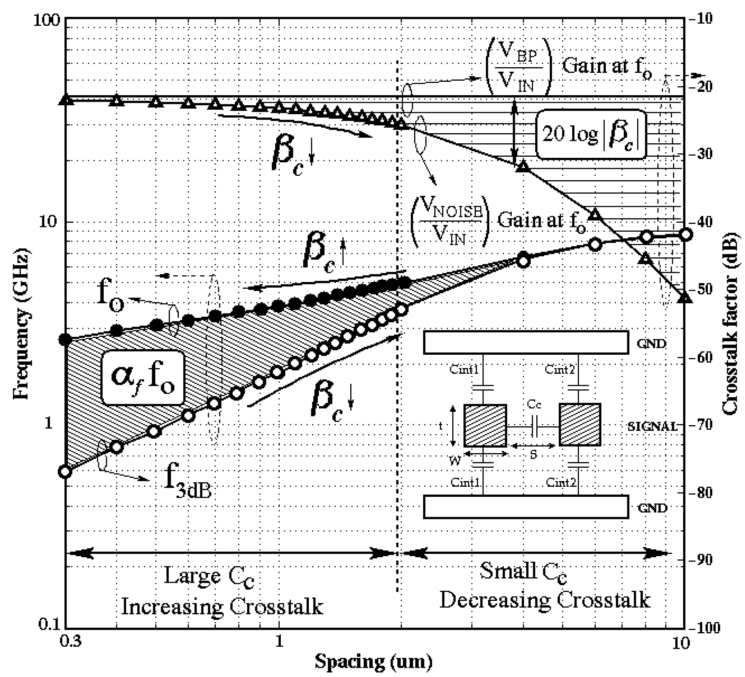
$$\alpha_f = 1 - \left[\sqrt{\left(\frac{\beta_c^2}{1-\beta_c^2}\right)^2 + 1} - \left(\frac{\beta_c^2}{1-\beta_c^2}\right) \right]^{\frac{1}{2}} \tag{4-9}$$

Equations (4-7) and (4-9) imply that the relative position of f_o and f_{3dB} can be controlled with interconnect design parameters. A plot of α_f versus β_c is shown in Figure 4.4a. A low β_c represents small coupling capacitance C_C whereas a large β_c indicates large C_C. Notice that as β_c is increased, α_f increases exponentially indicating that the difference f_o-f_{3dB} increases rapidly. A similar result is obtained in Figure 4.4b, where the dependence of α_f and β_c with interconnect spacing are shown in shaded regions. As to be expected, at large spacing s (i.e., β_c <<1) the cross-talk between the aggressor and victim lines is minimized, whereas the opposite is true for small s.

In general, *the increase in β_c implies that the victim line is more susceptible to the input excitation of the aggressor line, and the increase in α_f implies that the susceptibility due to higher frequency components tends to be larger*. This suggests, that in a high cross-talk factor β_c interconnect environment, the cross-talk noise can be minimized if band-limited signaling such as RCA pulses is used instead of square pulses.



(a)



(b)

Figure 4. 4. Frequency analysis of cross-talk noise (a) αf versus βc (b) Cross-talk frequency offset and loss factors vs. spacing. The analysis is based on the UMC 0.18-μm Al process with interconnects in Metal 4 (W=1-μm, T=0.57-μm and H=0.8-μm, L=10000-μm.

To gain a better understanding of the potential impact of the band-limited signaling scheme on interconnect wires with large aspect ratios or equivalently high β_c, we first formulate the cross-talk noise reduction in terms of the SQ and RCA pulse spectrums. The cross-talk noise power spectral density with input source pulse shapes $G_{SQ}(f)$ and $G_{RCA}(f)$ can be defined by:

$$V_{no,SQ}^2(f) = \left| \frac{V_{LOAD}(f)}{V_{IN}(f)} \right|^2 G_{SQ}^2(f) \tag{4-10a}$$

$$V_{no,RCA}^2(f) = \left| \frac{V_{LOAD}(f)}{V_{IN}(f)} \right|^2 G_{RCA}^2(f) \tag{4-10b}$$

where $G_{RCA}(f)$ is defined in Eq. (4-3) and $G_{SQ}(f)$ is given by the ubiquitous:

$$G_{SQ}(f) = AT_b \left[\frac{\sin(\pi f T_b)}{\pi f T_b} \right] \cdot e^{-j\pi f T_b} \tag{4-11}$$

Next, using equations (4-3), (4-10) and (4-11) we define $\gamma(f)$ as the *frequency envelop* of the ratio of the two cross-talk noise power spectral densities:

$$|\gamma(f)| = \begin{cases} \left| \dfrac{\cos(\pi \cdot fT_b)}{1-4(fT_b)^2} \right| & 0 \le fT_b < 1 \\ \left| \dfrac{1}{1-4(fT_b)^2} \right| & 1 \le fT_b < \infty \end{cases} \tag{4-12}$$

Equation (4-12) gives the frequency dependent envelope of the best-case cross-talk noise improvement of RCA over SQ signaling versus normalized bit time (fT_b), as shown Figure 4.5. The relative improvement at the first (i.e. fT_b=1) and second (i.e. fT_b=3) components of the data frequency is 9.54dB and 30.88dB, respectively. Notice that assuming a linear time invariant interconnect channel, the improvement in cross-talk noise over SQ signaling is independent of the noise transfer function, as to be expected.

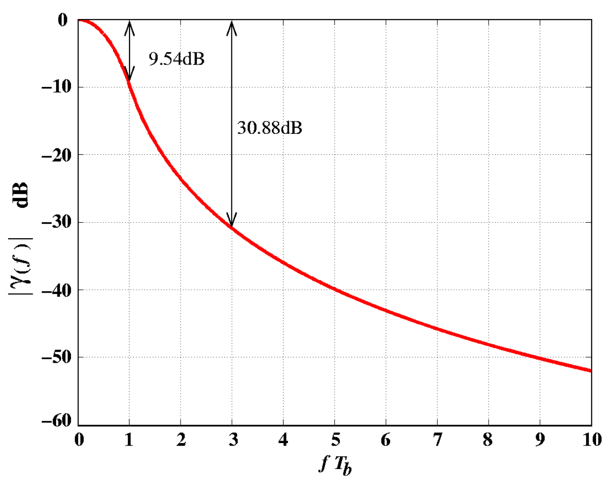


Figure 4. 5. Relative performance improvement of RCA signaling over SQ signaling.

4.2.2.B. Time domain Analysis

Unlike frequency domain analysis of cross-talk noise, time domain expressions can provide information on the instantaneous changes of cross-talk voltage in the quiescent line due to an aggressor. Thus, the peak noise voltage can be found to determine whether it exceeds the noise margin of the receiver. In order to determine the cross-talk voltage waveform due to a band-limited step input at the aggressor, the input waveform is first defined as the superposition of two sinusoidal terms, which can be expressed as,

$$V_{IN}(t) = \left(\frac{A}{2}\right) \cdot \left[\left(1 - \cos\left(\frac{\pi \cdot t}{t_R}\right)\right) \cdot u(t) + \left(1 - \cos\left(\frac{\pi \cdot t}{t_R}\right)\right) \cdot u(t - t_R) \right] \qquad (4\text{-}13)$$

where u(t) denotes the unit step function, A is the amplitude of the waveform and t_R is the input rise time. Notice that the expression in Eq. (4-13) is general and can be used to represent a SQ input by letting t_R approach zero, whereas an RCA waveform can be represented when $t_R = T_b/2$ (i.e. t_R is half the bit time). Figure 4.6 shows the input waveform composed of two sinusoidal terms.

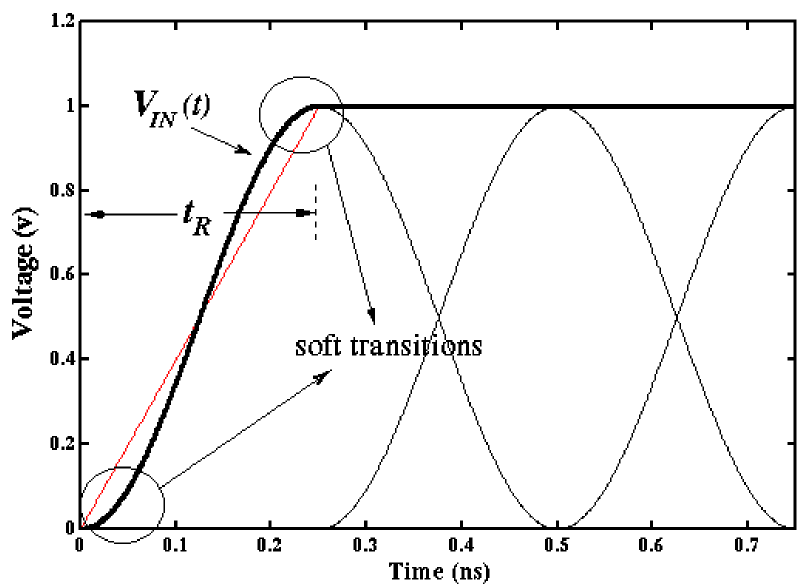


Figure 4. 6. Representation of a reduced edge-rate step with rise time tR. The waveform is composed of two sinusoidal terms.

For completeness, the output response of both the aggressor and victim lines at the receiver-end are derived. The aggressor line output response given the effective lumped element model of Figure 4.2b with input waveform as defined in Eq. (4-13) can be derived from the transform domain multiplication of Eq. (4-4a) and the Laplace of the input waveform. The Laplace of the input waveform can be expressed as,

$$V_{IN}(s) = \frac{A}{2}\left(\frac{1}{s} - \frac{s}{s^2+\omega_b^2}\right)\cdot\left(1+e^{-t_R s}\right) \tag{4-14}$$

The product of Eq. (4-4a) and Eq. (4-14) yields the following expression:

$$V_{LOAD}(s) = \frac{A}{2}\cdot\left(\frac{a_1 s}{s^2+b_1 s+b_0} + \frac{a_0}{s^2+b_1 s+b_0}\right)\left(\frac{1}{s} - \frac{s}{s^2+\omega_b^2}\right)\left(1+e^{-t_R s}\right) \tag{4-15}$$

where a_1, a_0, b_1 and b_0 are the coefficients of the transfer function defined in Eq. (4-4a). The output time-domain response can be expressed as:

$$V1_{OUT}(t) = V_{LOAD}(t) + V_{LOAD}(t-t_R) \tag{4-16}$$

In Eq. (4-16), $V_{LOAD}(t)$ is the kernel of the output response. Assuming that the coupled interconnects are symmetric (i.e. $R_{E1}=R_{E2}=R_E$ and $C_{E1}=C_{E2}=C_E$), the kernel can be written after some algebraic manipulation as:

$$V_{LOAD}(t) = \frac{A\eta_L}{2}\left\{1 - 0.5\cdot\left[\frac{\left(\frac{1}{2t_R f_o}\right)^2\left(\frac{1+\beta_c}{1-\beta_c}\right)}{1+\left(\frac{1}{2t_R f_o}\right)^2\left(\frac{1+\beta_c}{1-\beta_c}\right)}e^{-\omega_o\sqrt{\frac{1-\beta_c}{1+\beta_c}}\cdot t} - \frac{\left(\frac{1}{2t_R f_o}\right)^2\left(\frac{1-\beta_c}{1+\beta_c}\right)}{1+\left(\frac{1}{2t_R f_o}\right)^2\left(\frac{1-\beta_c}{1+\beta_c}\right)}e^{-\omega_o\sqrt{\frac{1+\beta_c}{1-\beta_c}}\cdot t}\right] - \frac{\sqrt{1-\beta_c^2}\cos\left(\pi\frac{t}{t_R} - |\phi|\right) - \left(\frac{1}{2t_R f_o}\right)\sin\left(\pi\frac{t}{t_R} - |\phi|\right)}{\sqrt{\left(1-\left(\frac{1}{2t_R f_o}\right)^2\right)^2\left(1-\beta_c^2\right) + 4\left(\frac{1}{2t_R f_o}\right)^2}}\right\} \tag{4-17}$$

where f_o and φ are defined as,

$$f_o = \frac{1}{2\pi\eta_L C_C R_E}\sqrt{\frac{\beta_c}{1+\beta_c}} \tag{4-18}$$

$$\phi = \frac{2\left(\frac{1}{2t_R f_o}\right)\sqrt{\frac{1}{1-\beta_c}}}{1-\left(\frac{1}{2t_R f_o}\right)} \tag{4-19}$$

In similar manner, the transform domain of the noise transfer function is given by the product of Eq. (4-4b) an the input waveform in (4-14), which can be written as,

$$V_{NOISE}(s) = \frac{A}{2}\cdot\left(\frac{a_o s}{s^2 + b_1 s + b_o}\right)\cdot\left(\frac{1}{s} - \frac{s}{s^2+\omega_b^2}\right)\cdot\left(1+e^{-t_R s}\right) \tag{4-20}$$

where a_0, b_1 and b_0 are the coefficients of the transfer function defined in (4-4.b). The output time-domain response of the quiescent line due to a finite rise time input waveform can be expressed as:

$$V2_{OUT}(t) = V_{NOISE}(t) + V_{NOISE}(t - t_R) \tag{4-21}$$

The kernel $V_{NOISE}(t)$ in Eq. (4-21), can be written as:

$$V_{NOISE}(t) = \frac{A\eta}{2}\left\{0.5\cdot\left[\frac{\left(\frac{1}{2t_R f_o}\right)^2\left(\frac{1+\beta_c}{1-\beta_c}\right)}{1+\left(\frac{1}{2t_R f_o}\right)^2\left(\frac{1+\beta_c}{1-\beta_c}\right)}e^{-\omega_o\sqrt{\frac{1-\beta_c}{1+\beta_c}}\cdot t} - \frac{\left(\frac{1}{2t_R f_o}\right)^2\left(\frac{1-\beta_c}{1+\beta_c}\right)}{1+\left(\frac{1}{2t_R f_o}\right)^2\left(\frac{1-\beta_c}{1+\beta_c}\right)}e^{-\omega_o\sqrt{\frac{1+\beta_c}{1-\beta_c}}\cdot t}\right] + \right.$$
$$\left. \frac{\beta_c\cdot\left(\frac{1}{2t_R f_o}\right)\cdot\sin\left(\pi\frac{t}{t_R} - |\phi|\right)}{\sqrt{\left(1-\left(\frac{1}{2t_R f_o}\right)^2\right)^2\left(1-\beta_c^2\right)+4\left(\frac{1}{2t_R f_o}\right)^2}}\right\} \tag{4-22}$$

where the initial phase φ is given by Eq. (4-19). The expression in (4-21) is used to find the time-domain response to input step with finite rise-time. For an arbitrary input pulse $p_T(t)$ of finite duration T_b, the noise voltage can be written as,

$$p_T(t) = V2_{OUT}(t) - V2_{OUT}(t - k\cdot t_R) \tag{4-23}$$

In Eq. (4-23), the pulse width is defined as $T_b = kt_R$, where k is any positive integer. For the special case when the pulse width is t_R (i.e. k=1), the peak cross-talk noise can be expressed in closed form by assuming that the peak noise occurs at the maximum of the sinusoidal term. Solving for $t = t_{peak}$ yields,

$$t_{peak} = t_R \left(0.5 + \frac{|\phi|}{\pi} \right) \tag{4-24}$$

Thus, an approximate closed form expression to the peak cross-talk noise is given by $V_{peak} = V_{NOISE}\,(t = t_{peak})$. In general, the peak cross-talk voltage can be obtained from Eq. (4-23), where less than 10 iterations of simple back substitution are typically sufficient. The aggressor load and victim cross-talk voltage are shown in Figure 4.7. Note that a resistive load termination (R_L) results in decreased noise only due to the reduced voltage swings of current mode signaling. That is, for the same voltage and current mode line response, the ratio of the peak noise to the voltage swing at the far-end of the aggressor line remains fairly constant for different R_L.

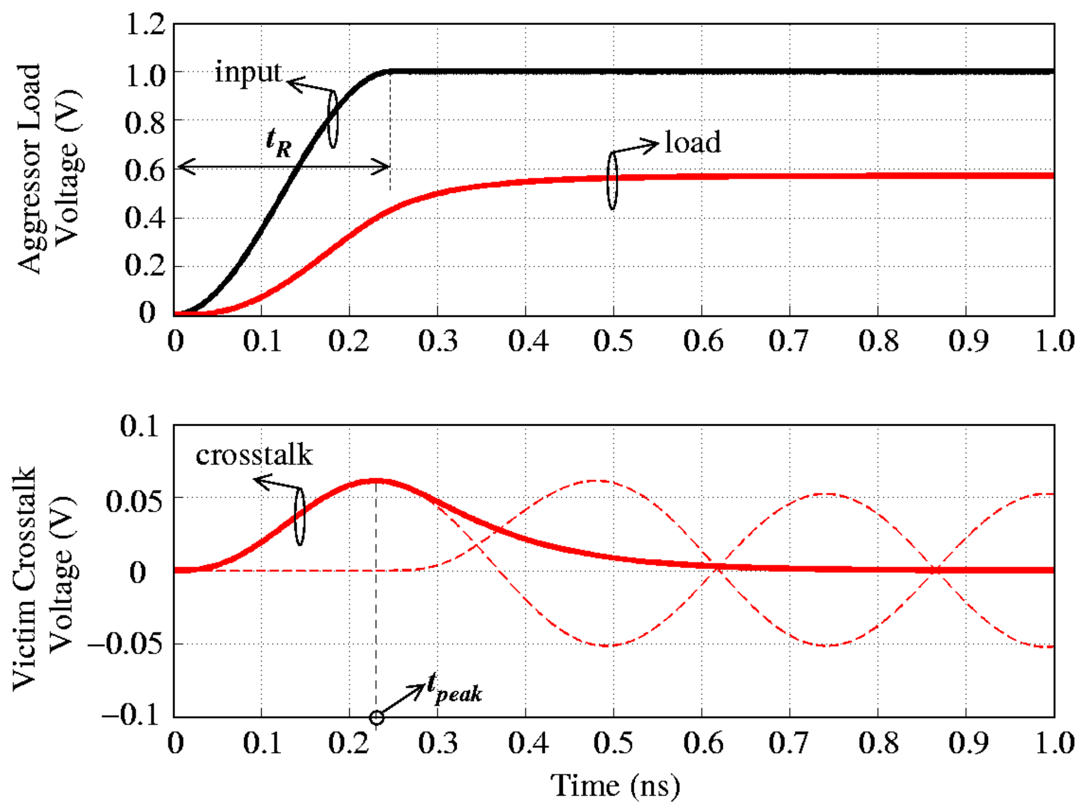


Figure 4. 7. Load and cross-talk voltage waveforms. The following parameters are used: R_L=1kΩ, C_L=500fF, C_T=2pF, R_T=500, R_S=250, C_S=100fF and t_R=2.5n.

4.3. Circuit Implementation

4.3.1. Adjustable Slew-rate Repeater

Repeaters, in general, are uniformly inserted in long global on-chip interconnects to minimize delay and improve data signaling bandwidth [5], [6]. In this work, repeaters are also used to limit the bandwidth of the output pulse. To realize a wide range of output edge-rates, a simple adjustable slew-rate repeater for on-chip signaling is proposed in Figure 4.8. Unlike typical implementation of slew-rate control drivers [7], [8], based on weighted current switches that are activated sequentially with signals derived from delay lines or counters, the proposed circuit has small overhead. This is particularly important given the increasing number of on-chip repeaters used in modern high performance VLSI designs [9].

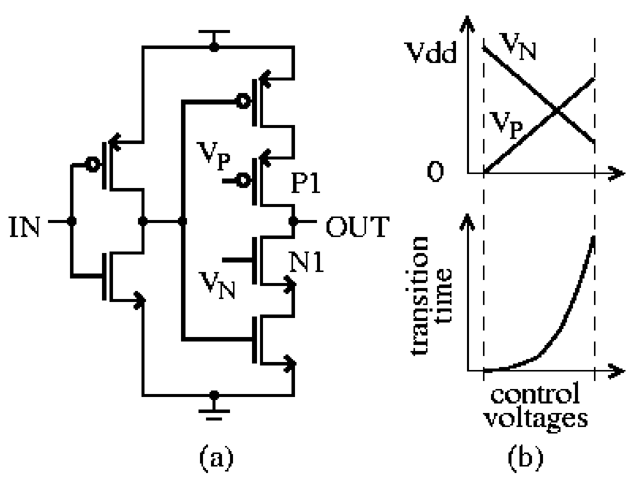


Figure 4. 8. Adjustable slew-rate repeater (a) circuit (b) transition time and control voltages and (b) fall time.

The non-inverting adjustable slew-rate repeater shown in Figure 4.8a operates as a typical buffer with the exception of voltage controlled MOS resistors P1 and N1. The dependence of output transition time on bias voltages V_P and V_N, designed to be symmetric about Vdd/2, is shown in Figure 4.8b. Notice that by setting V_N to Vdd and V_P to ground, fast transition times can also be obtained. Measured output transition waveforms from a test chip fabricated in AMI 1.6μm bulk CMOS technology are superimposed for several bias voltages, as shown in Figure 4.9.

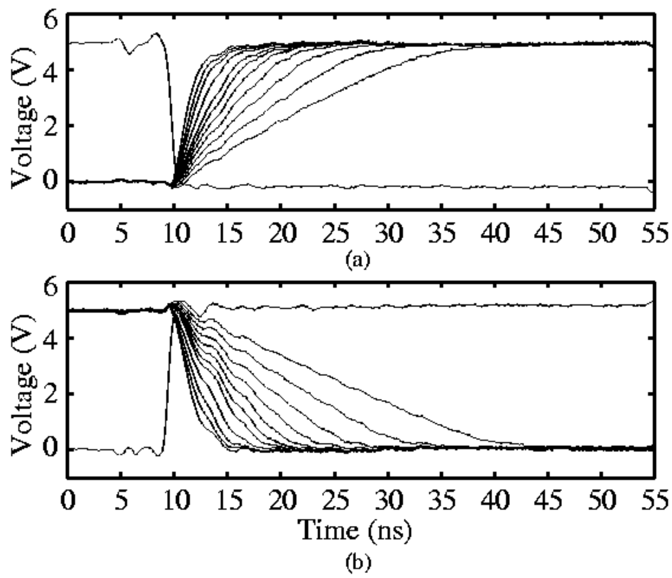


Figure 4. 9. Measured output driver response (a) rise time and (b) fall time.

Figure 4.10a shows the measured rise and fall transition times versus bias voltages V_P and V_N, respectively. Due to device mismatch, an offset voltage is required to equalize the transition times, as shown in Figure 4.10b. The mean offset voltage is approximately 140mV.

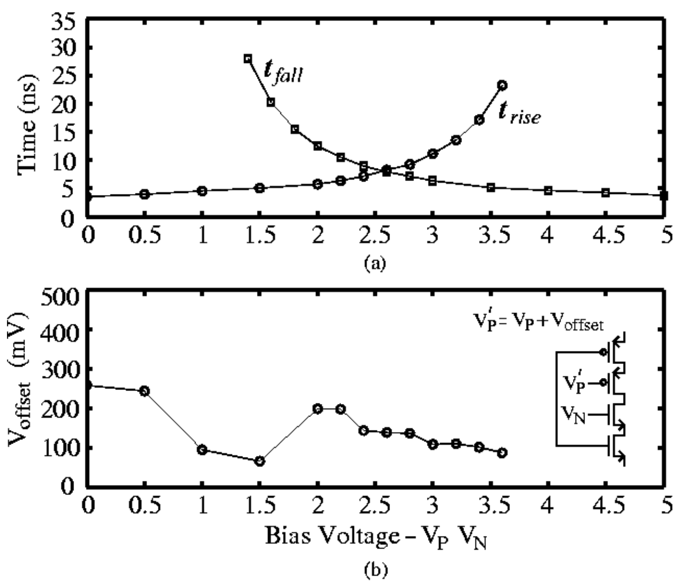
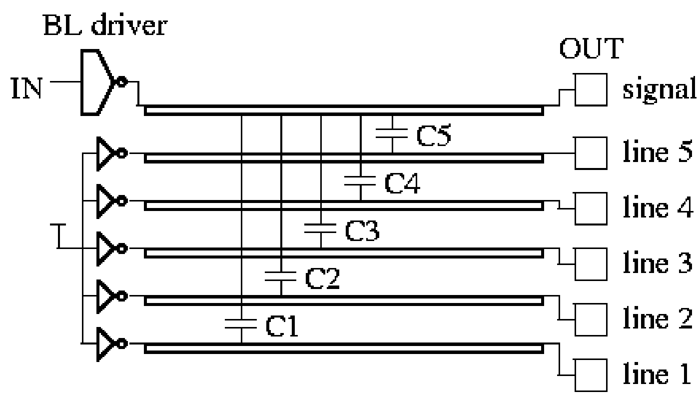


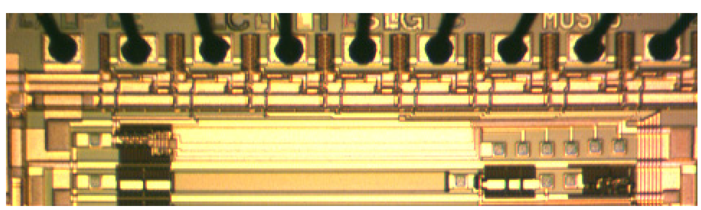
Figure 4. 10. Measured rise and fall time (a) vs. bias voltages VP and VN. (b) Offset voltage required to equalize rise and fall transition times.

4.3.2. Cross-talk Noise

In order to experimentally evaluate the effect of band-limited pulses on cross-talk noise, measurements were performed with cross-coupled test patterns, as shown in Figure 4.11a. The test chip fabricated in AMI 1.6μm, is shown in Figure 4.11b. To model the lateral wiring capacitance exhibited in deep sub-micrometer process technology, the cross-coupling capacitance between aggressor and victim lines were implemented using the vertical capacitance of metal 1 and metal 2 layers instead. Table 4.1 summarizes the signal and coupling capacitance values.



(a)



(b)

Figure 4. 11. Experimental test setup (a) test cases (b) die-section microphotograph.

Table 4. 1. Summary of capacitance values.

	Signal	**C1**	**C2**	**C3**	**C4**	**C5**
Capacitance	1.85pF	95fF	190fF	380fF	571fF	1.14pF
$C_{COUPLING}/C_{SIGNAL}$	-	0.05	0.1	0.206	0.31	0.62

Measured signal and noise waveforms for output rise and fall transitions are shown in Figure 4.12. Notice that the peak noise voltages in the victim lines, except lines 1 and 2 on the falling output transitions (i.e. C1 and C2), exceed the device threshold voltage for this process – V_{TN}=0.56V and V_{TP}= -0.97V. This occurs even for small coupling-to-signal capacitance ratios of 0.05 (see Table 4.1). Thus, in deep sub-micrometer process, where the device threshold voltages are low and the aspect ratios of minimum pitch global wires are projected to exceed 2.5 [1], the effect of noise cross-talk will become a dominant problem.

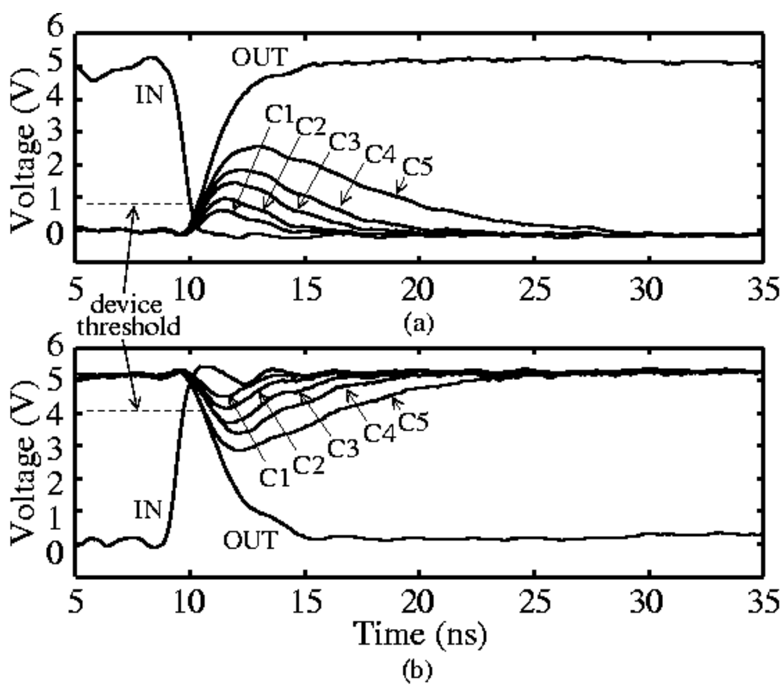


Figure 4. 12. Measured signal and noise waveforms due to (a) rising edge and (b) falling edge.

The measured peak noise voltage in the victim line versus output signal rise time is shown in Figure 4.13. The effect of limiting the output pulse bandwidth on cross-talk noise is more prominent at faster edge-rates, indicating that small changes in rise time can yield significant improvement in noise performance. For instance, in the case of line 5 (i.e. C5), decreasing the rise time by 2.18ns from an initial 3.5ns yields a reduction in peak noise of 500mV, whereas decreasing it by 10.7ns from 23ns yields a reduction of only 193mV. Notice, however, that the results in Figure 4.13 indicate that a relative large change in signal rise time is required to guarantee that the peak noise does not exceed the device threshold voltages.

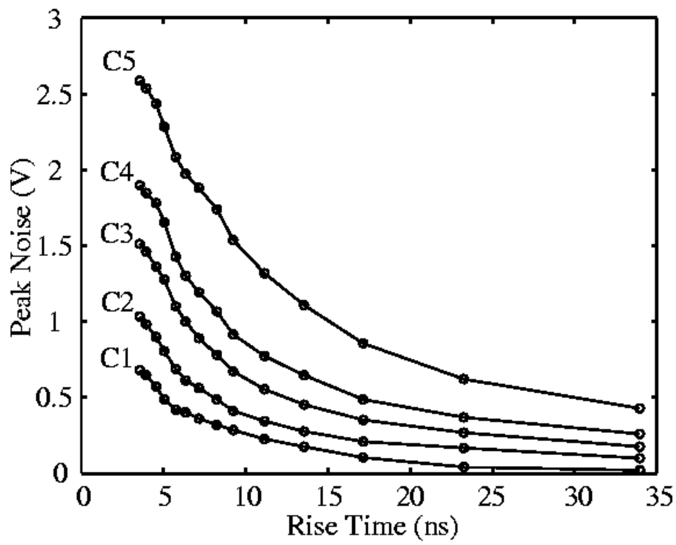


Figure 4. 13. Measured peak noise voltage vs. rise time.

4.3.3. Performance Evaluation

It is clear that the bandwidth of an output data stream cannot be limited or chosen arbitrarily. Thus, it is useful to formalize the definition of band-limited signaling for the purpose of digital on-chip signal transmission as it pertains to this work. Based on the pulse basis function definition in Eq. (4-2), the reduced edge-rate signal can be rewritten as,

$$g(t) = \begin{cases} A & 0 \le |t| \le (1-\alpha_R)\cdot\dfrac{T_b}{2} \\ \dfrac{A}{2}\left\{1+\cos\left[\dfrac{\pi\alpha}{T_b}\left(|t|-(1-\alpha_R)\cdot\dfrac{T_b}{2}\right)\right]\right\} & (1-\alpha_R)\cdot\dfrac{T_b}{2} \le |t| \le (1+\alpha_R)\cdot\dfrac{T_b}{2} \\ 0 & |t| > (1+\alpha_R)\cdot\dfrac{T_b}{2} \end{cases} \tag{4-25}$$

In Eq. (4-25), A is the signal amplitude and α_R is defined as the ratio of signal transition time (t_R) and bit time (T_b), or

$$\alpha_R = \frac{t_R}{T_b} \qquad 0 \le \alpha_R \le 1 \tag{4-26}$$

The definition of α_R in Eq. (4-26) indicates that the range of output transition time is ultimately limited by the target data-rate. This also implies, given fixed driver/receiver and interconnect parameters, that the maximum reduction in cross-talk noise occurs when α_R=1 (i.e. t_R=T_b). Figure 4.14 illustrates measured waveforms for an arbitrary 50Mb/s data stream with α_R=0.18 (Figure 4.14b) and α_R=1 (Figure 4.14c). The induced noise waveforms on C5 of the test setup in Figure 4.11a indicate a reduction in peak cross-talk voltage of 72%.

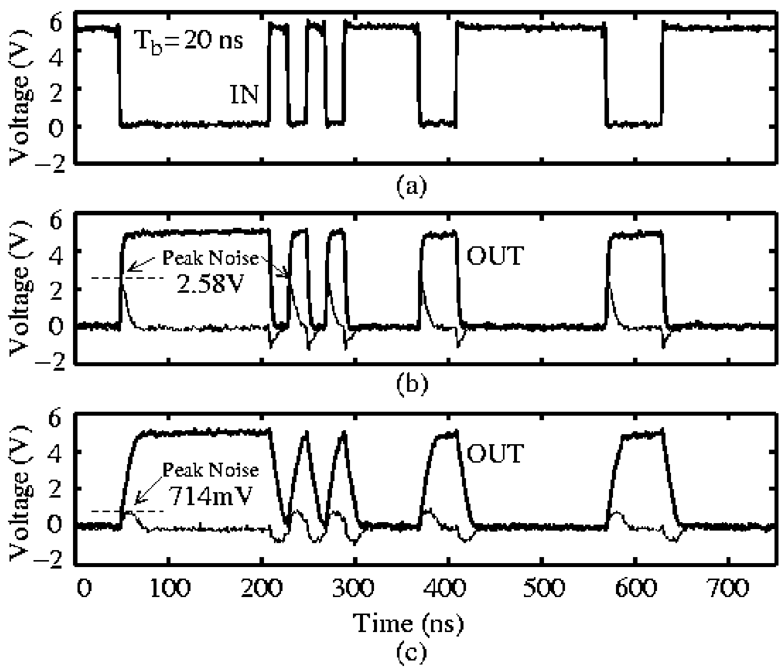


Figure 4. 14. Measured signal and noise waveforms for a 50Mb/s data stream (a) input; signal and noise waveforms for (b) α_R=0.18 and (c) α_R =1.

When using band-limited signaling for an on-chip communication link, it is important to consider two additional factors - power and delay. To evaluate the power and delay performance due to increased output transition time, simulations were carried out using the test model depicted in Figure 4.15a. Indicated in Figure 4.15b is the power dissipation versus rise time (t_R) for the aggressor and victim receivers, denoted as P_{REC} and P_{NOISE}, respectively. The receivers consists of two cascaded inverters with Wp/Lp=12μm/1.6μm and Wn/Ln=4μm/1.6μm.

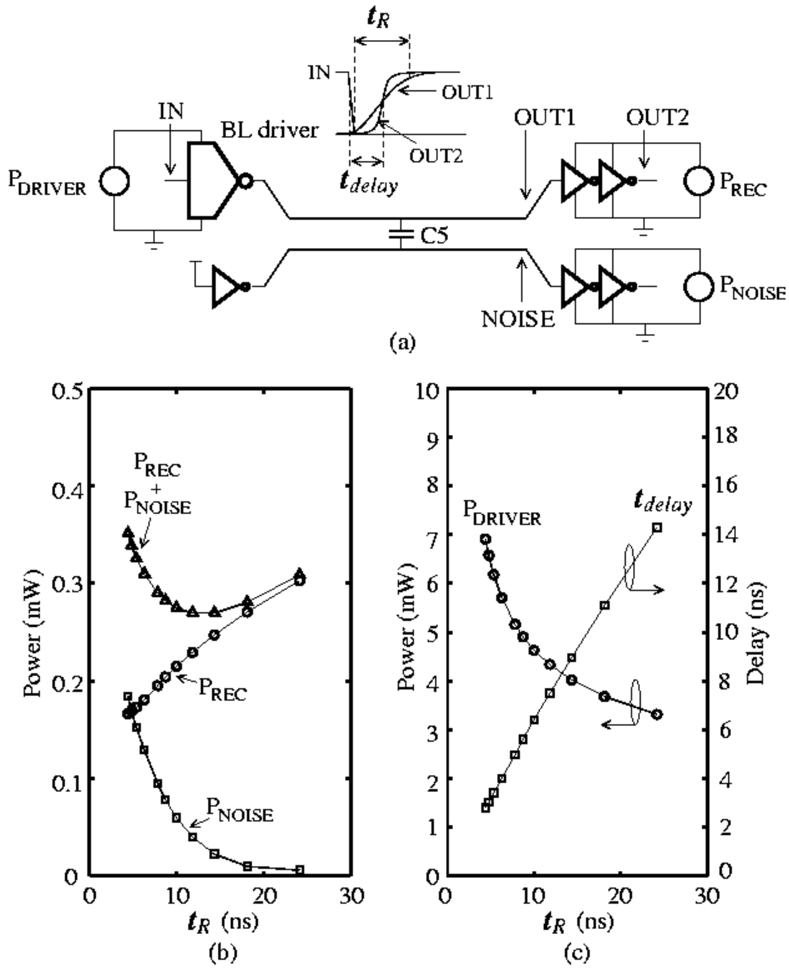


Figure 4. 15. Simulated power and delay tradeoffs due to band-limited pulse signaling, (a) test model, (b) signal and victim receiver power dissipation and (c) driver and delay performance.

The results shown in Figure 4.15b can be explained as follows. An increase in t_R at the input of the aggressor receiver (i.e. OUT1) causes larger short-circuit power to be dissipated, hence increasing P_{REC}. However, the noise voltage induced in the victim line (i.e. NOISE) decreases with increasing t_R, causing less current to conduct and P_{NOISE} to decrease. Thus, when the peak noise voltage in the victim line is large enough to disturb the internal node of the victim receiver, there exists an optimum t_R that minimizes power.

Figure 4.15c shows the driver power (P_{DRIVER}) and overall delay (t_{delay}) performance of the aggressor – t_{delay} is defined as the 50% point between IN and OUT2, as depicted in Figure 4.15a. P_{DRIVER} decreases significantly as t_R increases because a larger fraction of time is allocated to charge the capacitances associated with the interconnect line and driver/receiver circuits (i.e. CV^2/t_R). However, t_{delay} increases more or less linearly with trise, indicating that careful attention should be given so as not to violate allocated timing margins. A summary of performance tradeoffs versus α_R given a target signaling data-rate of 50Mb/s is shown in Table 4.2.

Table 4. 2. Band-limited signaling performance tradeoffs for Tb=20ns (data-rate=50Mb/s).

α_R (t_R / T_b)	Peak Noise Voltage (V)	Power Dissipation (mW)		Delay/ bit-time (t_{delay} / T_b)
		$P_{REC} + P_{NOISE}$	P_{DRIVER}	
0.18	2.58	0.351	6.91	0.14
0.25	2.25	0.325	6.18	0.17
0.5	1.4	0.274	4.63	0.32
0.75	1.0	0.269	4	0.448
1	0.714	0.29	3.5	0.575

4.4. Summary

A band-limited signaling scheme has been proposed to reduce cross-talk noise in global on-chip interconnects. Unlike typical implementations of cross-talk noise reduction techniques, the proposed signaling scheme aims to reduce the noise source itself by decreasing the signal edge-rates. Theoretical formulations of cross-talk noise are presented to understand the effect of band-limited signaling on interconnect coupling noise. Based on the effective resistance and capacitance extraction of coupled interconnects, closed form time domain equations have been derived for the noise and signal voltages. Further analysis of cross-talk noise is carried out in frequency domain to understand the performance improvement of band-limited signaling over square wave signals.

Design tradeoffs for band-limited pulse signaling in on-chip interconnects are presented and experimentally verified. It is shown that reducing the edge-rate of data streams in on-chip high-density interconnects yield significant improvement in cross-talk noise – up to 72% reduction in peak noise voltage for coupling-to-signal capacitance of 0.6 and data-rate of 50Mb/s. Although the propagation delay increases nearly linearly with transition time, the power dissipation associated with driver, interconnect and receiver circuits decrease significantly. It is observed that, if the coupled peak noise voltage is large enough to induce a change in victim receiver output voltage, there exists an optimum transition time that minimizes the power dissipation of both aggressor and victim receivers.

More importantly, since materials and process approaches to minimizing relative crosstalk are yielding smaller incremental performance improvements with successive technology generations of the ITRS, the use band-limited signaling – or the option to do so with adjustable slew-rate repeaters – may serve integral in meeting performance goals.

4.5. References

[1] Semiconductor Industry Association, *The International Technology Roadmap for Semiconductors*, 2001 Edition.

[2] B. Herrick, "Design Challenges in Multi-GHz Processors," Slide presentation, Alpha Microprocessor Development.

[3] J.B. Anderson, "Digital Transmission Engineering," IEEE Press, 1999.

[4] M. Bank, J. Gavan, "Practical realization of a raised-cosine filter," Elect. Letters, vol. 32, no. 5, 438-440, Feb. 96.

[5] H.B. Bakoglu, Circuits, Interconnections, and Packaging for VLSI. Reading, MA: Addison-Wesley, 1990.

[6] P.L. Edefors, "Investigation on maximal Throughput of a CMOS Repeater Chain," IEEE Trans. Cir. Systems I, vol. 47, no. 4, pp. 602-606, April 2000.

[7] T. Gabara, W. Fischer, J. Harrington, W. Troutman, "Forming Damped LRC Parasitic Circuits in Simultaneously Switched CMOS Output Buffers," IEEE J. Solid-State Circuits, vol. 32, no. 3, pp. 407-418, Mar. 1997.

[8] H. Wilson, M. Haycock, "A six-port 30-Gb/s Nonblocking Router Component Using Point-to-Point Simultaneous Bidirectional Signaling for High-Bandwidth Interconnects," IEEE J. Solid-State Circuits, vol. 36, no. 12, pp. 1954-1963, Dec. 2001.

[9] R. McInerney, et al. "Methodology for repeater insertion management in RTL, layout, floorplan and fullchip timing databases of the ItaniumTM microprocessor," Proc. Int. Symp. Physical Design, pp. 99-104, 2000.

Chapter V

Adaptive Bandwidth Bus (ABB)

5.1. Introduction

Repeaters have been traditionally used for optimizing wire delays [1]. However, as operating frequencies continue to increase, repeaters will be required to improve the signaling bandwidth of interconnects for the purpose of achieving high-speed on-chip data transmission [2]. A dominant limitation of interconnects at higher frequencies is the signal dispersive effects due to lossy conductors. One possible approach to extend the bandwidth of lossy interconnects is via an alternative repeater insertion technique based on current-mode signaling. Current-mode signaling uses low-impedance receive-end signal sensing to fundamentally enhance the interconnection bandwidth and hence increase the maximum attainable data rates in addition to reducing the wire delays.

The advantages of current-mode signaling can be summarized as follows:

1) An increase in signaling bandwidth (i.e. higher data rates) and the reduction in interconnection latency.
2) Reduction in the number of repeaters for a given target data rate.
3) Decrease in dynamic power dissipation via the reduction of signal voltage swings.

To compensate for the increase in static power dissipation associated with current-mode signaling, the proposed adaptive bandwidth signaling scheme allocates the required interconnection bandwidth only when input data transitions are sensed, otherwise remaining in low-bandwidth voltage-mode to eliminate the static power consumption altogether.

To demonstrate the feasibility of the proposed adaptive bandwidth bus (ABB) scheme, a preliminary "analog" architecture has been fabricated in AMI 1.6μm bulk CMOS technology, as described in section 5.2. The limitations of this design are addressed, and an improved all-digital ABB architecture is presented section 5.3. In section 5.4, a power dissipation modeling approach based on circuit-level and statistical analysis of data streams is presented to evaluate the performance of the proposed bus. Circuit implementation details and measured comparative results of an experimental adaptive bus fabricated in TSMC 0.35μm CMOS technology, which operates at a sustained bandwidth of 16Gb/s (i.e. 1Gb/s per line), are presented in section 5.5. Concluding remarks are presented in section 5.6.

5.2. Preliminary Study - Analog ABB

The key to current-mode signal transporting is in the reduction of RC time constants that results from sensing signals with low impedance nodes [3]. In addition, it should be noted here that although line inductance is assumed to be negligible because of high conductor attenuation (i.e. R››ωL), current sensing techniques have been shown to be effective for driving long interconnects in the presence of on-chip inductive effects [4]. As shown in Figure 5.1, simulations carried out in HSPICE based on 1.8V-0.18μm bulk CMOS technology parameters for a fixed driver size with metal-6 interconnection length of 20mm, minimum-width of 0.5μm and minimum adjacent metal spacing, the signal rise-time at the receiver decreases as the termination load resistance R_L is reduced. This improvement, however, is at the expense of an increase in power dissipation due to the static current path between driver and receiver circuits, which is not present for voltage-mode signals (i.e. $R_L \sim \infty$)

In this section, we explore an analog adaptive bandwidth bus scheme that uses current-mode sensing to allocate interconnection bandwidth when input data transitions are sensed, and otherwise remains in low-bandwidth voltage-mode to compensate for the increase in static power dissipation associated with current-mode signaling.

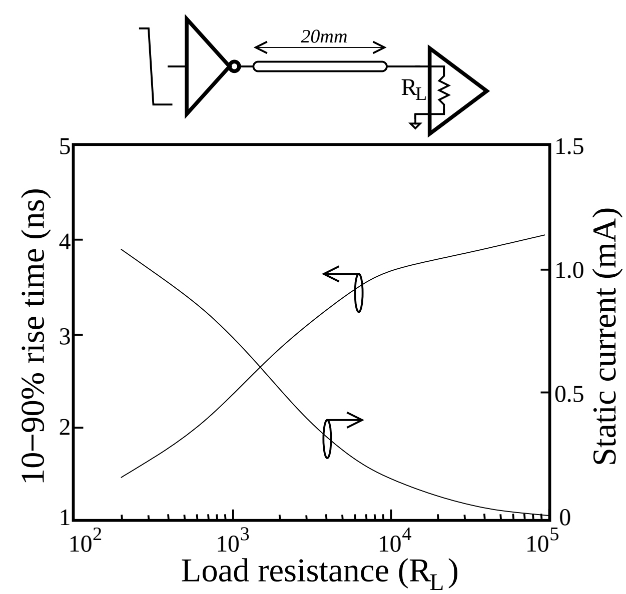


Figure 5. 1. Effect of receiver termination on signal rise-time and static power dissipation.

5.2.1. Principle of Operation

Figure 5.2a shows an adaptive bandwidth line divided into smaller segments of resistance R_T and capacitance C_T with appropriately spaced repeaters. Each repeater is used to terminate the line, amplify the signal and drive the following interconnect segment. An adaptive control unit is placed at the input of the bus and connected to a control line, which is distributed to all the repeater stages.

The operation of the bus can be described as follows. The control unit senses input data transitions generating a voltage (V_{CTRL}) that is proportional to the input data activity of the bus (Figure 5.2b). A special type of hybrid current/voltage mode repeater senses changes in the control line voltage V_{CTRL} and adjusts the termination resistance of each interconnect segment, effectively changing the signaling bandwidth of the lines. Therefore, the adaptive bus exploits the higher signaling bandwidth of current-mode operation (i.e. low impedance termination) only when data transitions are sensed.

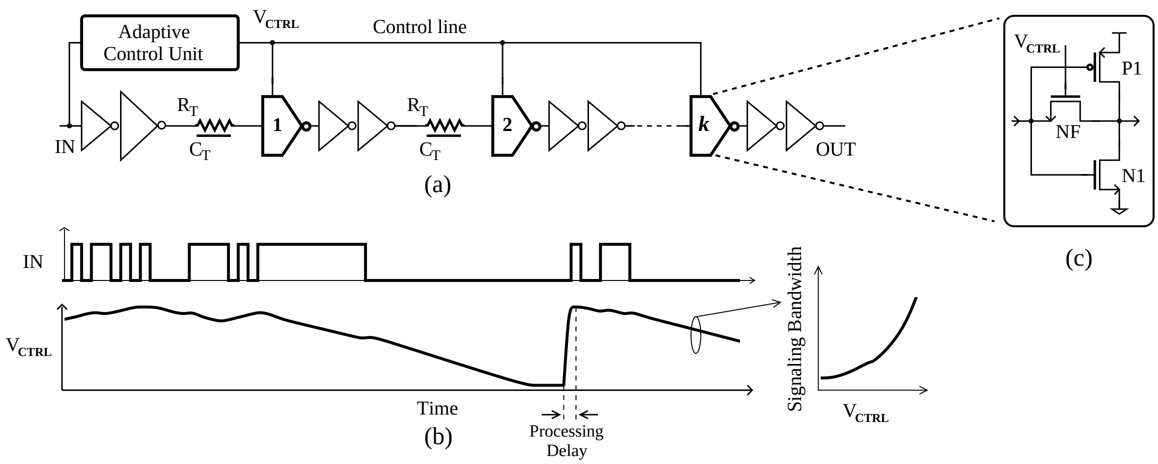


Figure 5. 2. Adaptive bandwidth line (a) architecture (b) input data activity and control line waveforms (c) hybrid current/voltage mode repeater.

Notice that in the absence of data transitions, V_{CTRL} is discharged to minimize power dissipation (Figure 5.2b). Therefore, the adaptive bus must ensure that V_{CTRL} is set high before a pulse is launched in the repeater chain so as to allocate the required interconnect bandwidth. The time required to charge the control line is defined as the setup time or processing delay of the bus. The setup time of an adaptive bus imposes that input data transitions are sensed early in the datapath. To further reduce this processing delay of the ABB, an improved design architecture has been proposed in section 5.3.

5.2.2. Hybrid Current/Voltage Mode Repeater

Figure 5.2c shows the schematic of the input stage of the hybrid current/voltage mode repeater (HCVR), which consists of an inverter with a voltage controlled feedback transistor. The HCVR operates as an inverter when the control voltage (V_{CTRL}) is less that the feedback transistor threshold voltage. Because of the high input impedance of the inverter, the line operates in full-swing voltage mode. As V_{CTRL} is increased, however, the feedback device turns on causing the repeater to operate as a self-biased inverter with low input impedance for current sensing operation.

The operation of the HCVR was validated via a test chip fabricated in AMI 1.6μm bulk CMOS technology. The test setup consisted of two cascaded inverters (sized at W_P/W_N=240μm/80μm) driving an interconnect line with resistance R_T=375Ω and capacitance C_T=1.43pF, and the HCVR at the receiving-end of the line. The HCVR devices were sized at W_{P1}=192μm, W_{N1}=64μm and W_{NF}=40μm (see Figure 5.2c). A minimum L=1.6μm was used for all devices.

Shown in Figure 5.3 are the measured voltage waveforms of the HCVR. When V_{CTRL}=0V (i.e. voltage-mode), the line voltage rise time at the input of the HCVR is 2.96ns and the overall delay is 2.45ns, as shown in Figure 5.3a. When V_{CTRL}=5V (i.e. current-mode), the rise time and overall delay are 1.35ns and 2.04ns, respectively (Figure 5.3b). Notice that this includes the CMOS level conversion latency. The results indicate a reduction in rise time of 55% and in delay of 17%.

The maximum bandwidth or NRZ data rate of the HCVR for several V_{CTRL} voltages is depicted in Figure 5.4. In order to characterize the bandwidth performance of the interconnection and low-to-high swing conversion circuit, measurements were taken at the input (i.e. INh) and output (i.e. OUT) nodes of the HCVR, respectively. The maximum bandwidth of the interconnection, obtained via 1/rise-time measurements at INh, indicate an increase of nearly 400Mb/s over the V_{CTRL} range. However, the maximum NRZ data rate measured at the output node (i.e. OUT) was limited to 660Mb/s, above which the pulses began to distort due to the speed limitation of the low-to-high swing circuit. Although in the present design the low-to-high swing conversion consists of a couple of cascaded inverters, a more advanced process with faster devices is likely to remove this limitation.

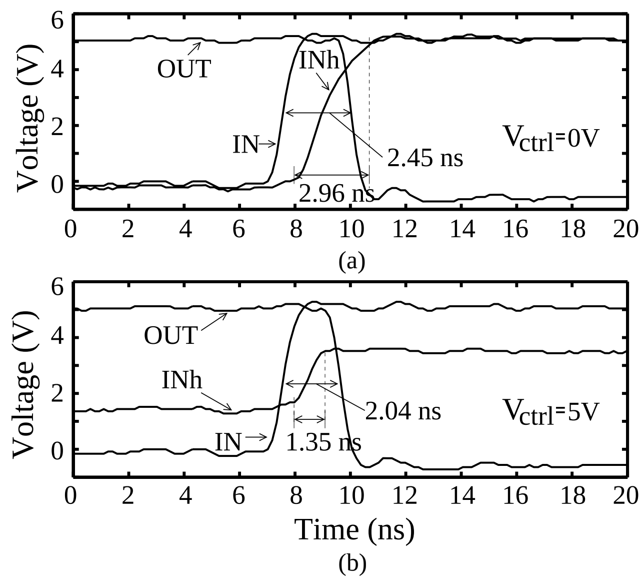


Figure 5. 3. Measured waveforms of the HCVR (a) voltage-mode (b) current-mode.

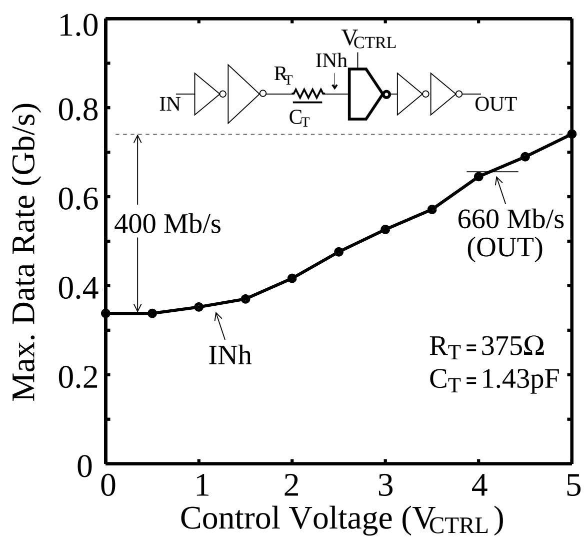


Figure 5. 4. Measured data rate performance of HCVR.

5.2.3. System Implementation

Figure 5.5 shows the system implemented to evaluate the performance of the adaptive lines. The control unit, which generates a voltage (V_{CTRL}) that is proportional to input data activity, consists of transition detector (TD), a charge pump (CP) and a lone pulse detector (LPD).

For every input data transition, the transition detector generates short pulses that are three inverter delays wide. These short pulses are used to increase V_{CTRL} via the charge pump circuit shown in Figure 5.5. A current source biased at V_{down} simultaneously discharges V_{CTRL} at a fixed rate. Thus, in the absence of data transitions, the control line is discharged to bias the HCVR in voltage mode

A lone pulse detector circuit, shown in Figure 5.5, is used in parallel with the charge pump to ensure correct data transmission given the onset of a lone input pulse – i.e. a narrow pulse preceded by a long period of inactivity. This circuit is activated only when V_{CTRL} is less that V_{TH} and an input data transition is detected. When both conditions are satisfied, the LPD forces V_{CTRL} high via large PMOS transistors to update the HCVR operation to current-mode and allocate the required interconnection bandwidth.

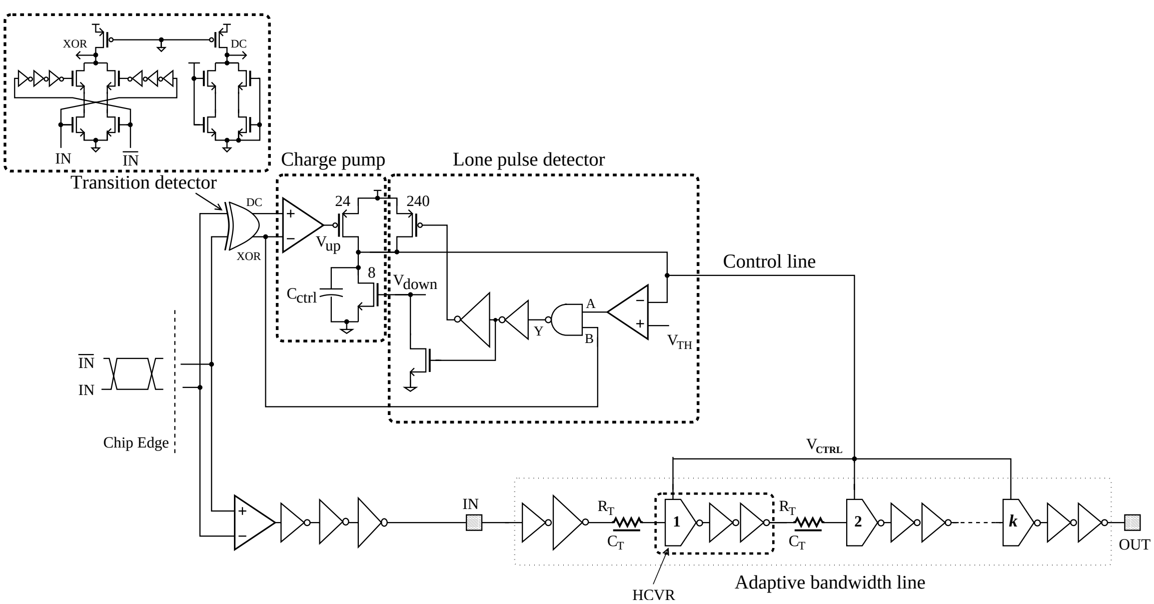


Figure 5. 5. Test chip implementation

Notice that when the LPD is activated, a fast V_{CTRL} rising transition may cause undesirable signal coupling effects from the control line to an adjacent adaptive bus line. Therefore, the routing of the control signal must take into consideration cross-talk effects via conventional techniques, such as increased metal spacing and/or partial shielding. Other possible coupling effects due to gate-to-source/gate-to-drain capacitance of the HCVR input amplifier stage are attenuated since the internal node is statically clamped (i.e. resistive path) to the supply rails.

In Figure 5.5, the bias voltages V_{TH} and V_{down} also serve to externally force V_{CTRL} high or low for current or voltage mode operation, respectively. For instance, by setting V_{TH}=Vdd and V_{down}=GND, V_{CTRL} will remain high regardless of input data transitions – a condition used to test the repeater chain under current-mode (CM) operation. Conversely, reversing the bias voltages forces voltage-mode (VM) operation (i.e. V_{TH}=GND and V_{down}=Vdd).

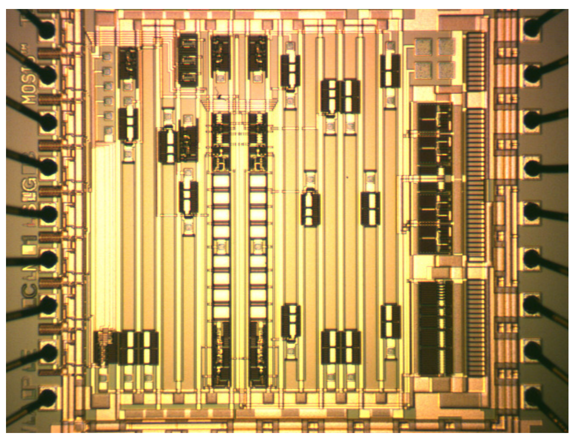


Figure 5. 6. Test chip photograph.

5.2.4. Performance Evaluation

5.2.4.A. Experimental Results

Shown in Figure 5.6 is a photograph of the test chip fabricated in AMI 1.6µm. Due to the small die area, poly interconnects were used to model the effect of long RC wires. Measured waveforms shown in Figure 5.7 depict the operation of the ABB. Figure 5.7a shows an arbitrary input data stream (IN) and control voltage (V_{CTRL}), which reacts to the input signal activity. In the absence of transitions, the charge-pump discharges V_{CTRL} low within approximately 36ns, whereas the LPD updates V_{CTRL} high on the onset of a pulse within 2ns. Notice that the fast update time is necessary to allocate bus-signaling bandwidth before data is launched. Figure 5.7b shows the waveforms at the output node (OUT) indicating correct data transmission.

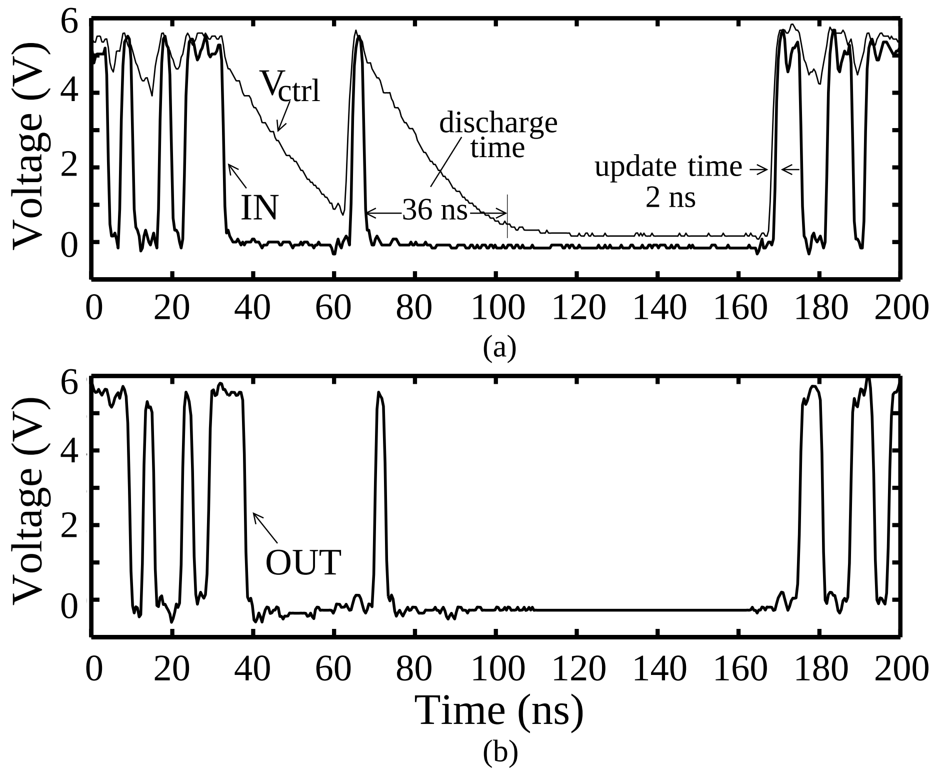


Figure 5. 7. Measured waveforms of the ABB operation (a) input and control voltage (b) output.

Shown in Figure 5.8 is a comparison of the power dissipation versus data rate performance for current-mode (CM), voltage-mode (VM) and adaptive-mode (ABB) signaling schemes. The measured results are based on a test line with four uniformly spaced repeaters and interconnect line segments of R_T=436Ω and C_T=1.65pF each. At Vdd=5V (Figure 5.8a), the ABB scheme achieves the same maximum data rate as the CM scheme (approximately 40% higher than VM signaling), while minimizing the power dissipation by nearly 62% over the CM bus when the activity rate is low. Reducing the supply voltage to Vdd=3V (Figure 5.8b) yields similar improvements in relative performance for the ABB scheme – approximately 50% increase in signaling rate over the VM bus and 63% reduction in power over the CM bus. Furthermore, a comparison between the ABB scheme at Vdd=3V and the VM bus at Vdd=5V indicates nearly 2.5x reduction in power dissipation with only 4.5% penalty in maximum data rate (i.e. 420Mb/s over 440Mb/s).

It is worthwhile noting that the discharge time of V_{CTRL} determines the data rate for which the 'Adaptive' and 'CM' curves in Figure 5.8 merge, and should be considered to further improve the power dissipation performance of the ABB. Since V_{CTRL} sets the operation of the bus in current or voltage mode, the slower it discharges the longer it remains in current-mode dissipating static power. Thus, reducing the V_{CTRL} discharge time yields a reduction in power dissipation at higher data rates, effectively shifting the 'Adaptive' curve in Figure 5.8 to the right.

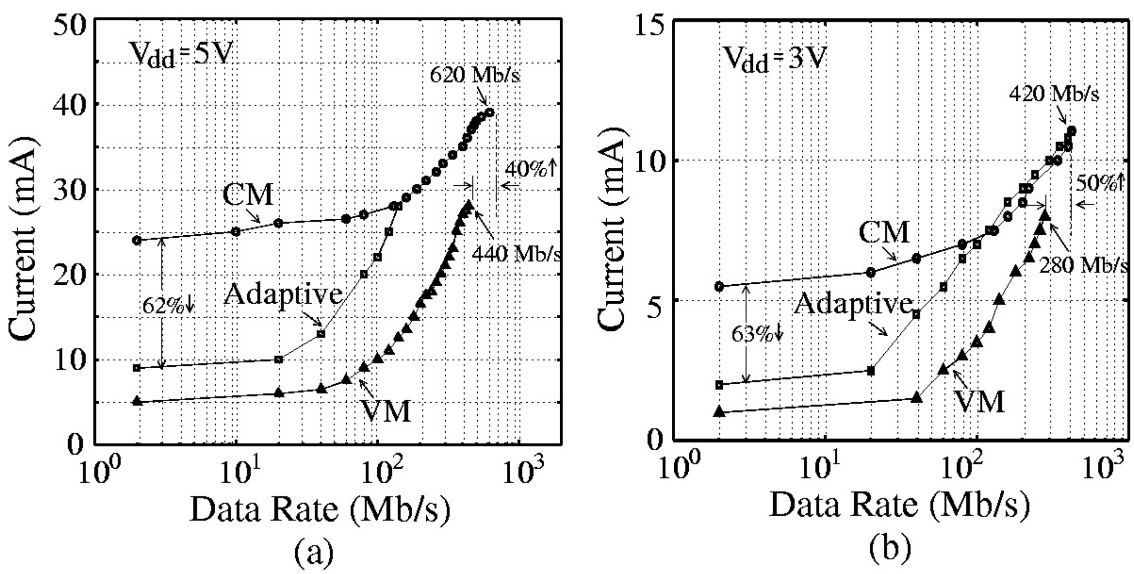


Figure 5. 8. Measured power dissipation vs. data-rate (a) Vdd=5V, (b) Vdd=3V.

In the present design, a capacitance (C_{CTRL}=1.5pF) was added to the output of the CP (Figure 5.5) to regulate the control voltage (V_{CTRL}). However, experimental results indicated that this is not necessary. In fact, the addition of C_{CTRL} causes the discharge time to increase, which limits the improvement in power dissipation of the ABB scheme at higher data rates.

In general, an important advantage of CM sensing over VM signaling is that its dynamic power dissipation component can be significantly reduced as a result of smaller voltage swings in the interconnect. This can be inferred from Figure 5.8 by noting that the slope at which the power increases with increasing data rates is smaller for the CM bus. For instance, at Vdd=5V, the measured change in current is 24μA/Mb/s for CM signaling and 55μA/Mb/s for VM signaling – a reduction of 56%. Thus, at higher operating frequencies, the power dissipation of full-swing VM signaling is likely to become more dominant.

Figure 5.9 shows the overall delay performance of four uniformly spaced repeaters given an input step for both CM and VM signaling schemes. The 50% delay is measured to be 6.2ns for CM sensing and 8.8ns for VM signaling, indicating nearly 30% reduction in delay.

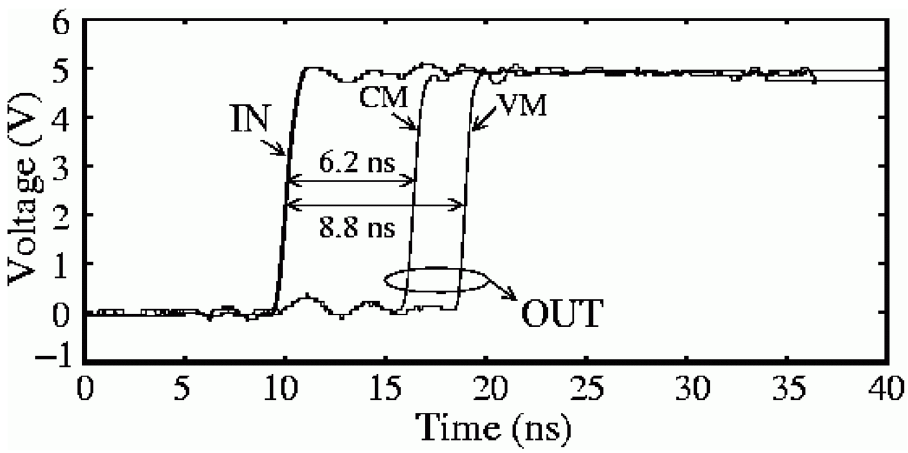


Figure 5. 9. Measured delay of CM and VM signaling schemes. Test line consists of four uniformly spaced repeaters (each interconnect segment has R_T=436Ω and C_T=1.65pF.

5.2.4.B. ABB Circuit Overhead

Several factors contribute to the overall area of the ABB scheme, such as 1) the number of repeaters per bus line, 2) the number of bus lines sharing a single control unit and 3) the length of the interconnect segments between repeaters. A summary of the active circuit area of the ABB, which include its control unit and HCVR buffer, are shown in Table 5.1. Area projections for sub-micrometer process, assuming a scaling factor of S=0.7 per technology node [5], are also provided for comparison.

The scaling projections suggest that the active area of the ABB scheme can be reduced to less than 1.5% if 0.18μm process is used. Moreover, the equivalent metal bus wiring area in 0.18μm technology that is required to match the resistance and capacitance of poly-silicon interconnects used in the present design, represents a significant fraction of the total bus area. For instance, the poly-silicon wire segments of R_T=436Ω and C==1.62pF roughly correspond to a Metal 5 wire over Metal 4 ground plane which is 1.8μm wide and 9.8mm in length – based on 0.18μm parameters R_{SQ}=0.08Ω/□, C_{AREA}=38aF/μm^2 and C_{FRINGE}=55aF/μm. The ratio of repeater to interconnect area - $A_{HCVR}/(w_{INT} \times l_{INT})$, is roughly 1%.

Table 5.2 shows the percentage overhead area of the ABB control unit parameterized by the number of repeaters and bus lines. As the number of repeaters per bus line is increased, the relative overhead of the control unit decreases. A more significant reduction in overhead results from the sharing of the control unit with multiple bus lines – less than 12% overhead for shaded areas in Table 5.2.

Table 5. 1. Actual and projected area of ABB circuits. [1]Calculated from AMI 1.6μm design.

Feature Size(μm)	1.6[1]	0.35	0.18
Control Unit Area - A_{CU} (μm^2)	49.5k	2.85k	685
HCVR Area – A_{HCVR} (μm^2)	13.5k	778	186

Table 5. 2. ABB circuit overhead parameterized by the number of repeaters (k) and bus lines (N).

Number of Repeaters (k)	ABB % of active area overhead vs. N-bus lines $\frac{A_{CU}}{A_{HCVR}}\left(\frac{1}{k \cdot N}\right)\times 100$			
k	**N=1**	**N=4**	**N=8**	**N=16**
2	183%	46%	22.9%	11.4%
4	91.6%	22.9%	11.4%	5.2%
8	46%	11.4%	5.7%	2.8%
12	30.5%	7.6%	3.8%	1.9%
16	22.9%	5.7%	2.8%	1.4%

5.3. Digital Adaptive Bandwidth Bus (ABB)

5.3.1. Analog vs. Digital ABB

One of the drawbacks in the analog implementation of the ABB is that the bus input-to-output propagation delay varies with V_{CTRL} – as the bandwidth of the line is proportional to V_{CTRL}. This uncertainty in propagation delay and data arrival time with respect to the global clock boundaries can cause detection errors at the receiving end of the bus. In order to ensure correct data detection, the digital ABB implementation uses a binary or digital control line to set the bandwidth of the bus lines. Thus, in the event of data transitions, V_{CTRL} is set to logical level HI; and in the absence of data transitions V_{CTRL} is set to logical level LO. Since data transitions are sent across the bus lines only when the control voltage is set to the appropriate voltage level, the propagation delay remains constant.

5.3.2. Concept and Architecture

Figure 5.10 shows the adaptive bandwidth bus (ABB) consisting of bus lines divided by appropriately spaced repeaters that are capable of terminating the interconnection segments with either a high impedance capacitive load or a low impedance resistive load. The bus uses a separate control line, which is shared among multiple bus lines, to indicate whether the repeaters should operate in voltage-mode (i.e. capacitive load) or current-mode (i.e. resistive load). The control signal used to indicate the operation mode of the bus is generated by a set of transition detectors placed at a relevant location in the pipeline stage of each bus line. These pipeline stages are logical-function independent to the operation of the bus, but are required to provide the additional timing margin for the ABB, as explained in section 5.3.3. The transition detectors can be implemented by a sampled XNOR function embedded in flip-flops to isolate the pipeline data changes from the control line.

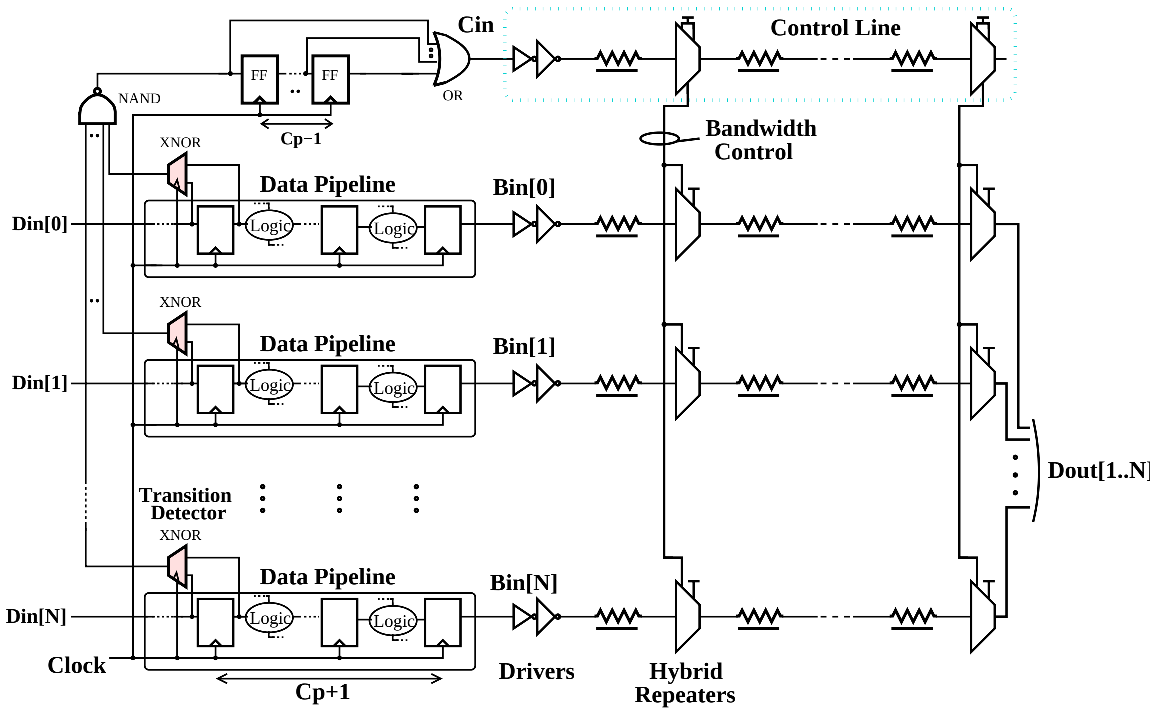


Figure 5. 10. Architecture of the adaptive bandwidth bus (ABB).

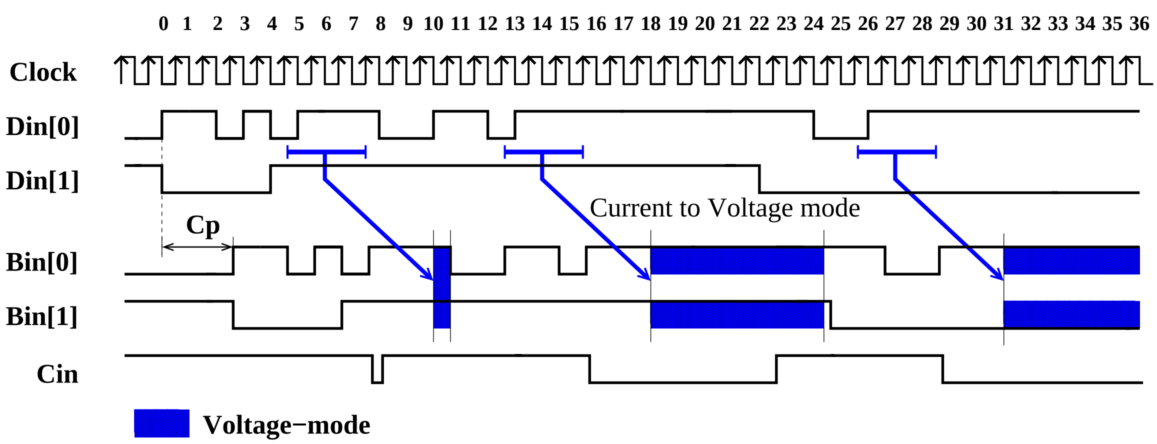


Figure 5. 11. Cycle-by-cycle operation of the ABB.

Figure 5.11 shows the cycle-by-cycle operation of a 2-bit wide bus with a single shared control line Cin. Data sequences Din[0] and Din[1] are pipelined and delayed by C_P cycles before appearing at the bus inputs denoted as Bin[0] and Bin[1], respectively. For simplicity, it is assumed that the logical levels of the pipelined data are not modified. In Figure 5.11, the bus initially operates in CM, as indicated by a logical level “1” in the control line. In the absence of data transitions for C_P+1 cycles, the control signal Cin switches to “0” setting the bus lines to VM operation. In the event that data transitions are detected early in the pipeline stages, the transition detectors activate the control line to revert the bus lines to CM operation, allocating the signaling bandwidth required to reliably transmit data. Therefore, the ABB dynamically switches from CM to VM and vice-versa to eliminate the static power dissipation, indicated by the shaded regions in Figure 5.11.

5.3.3. Control Signal Generation and Timing Constraints

The mechanism of generating the control signal for the ABB can be synchronous, as indicated in Figure 5.10, or asynchronous, via an interlocking handshaking operation with valid and acknowledge signals [6]. However, in either case, the timing margins (i.e. setup and hold time) must be satisfied for proper operation. Since the interconnection bandwidth must be allocated prior to launching data transitions into the bus lines, timing margin violations may result in 1) loss of data due to lack of signaling bandwidth (i.e. narrow pulses), or 2) data-dependent delay uncertainty due to signal latency variations causing increased jitter at the bus outputs.

Figure 5.12 shows the setup and hold time constraints of the ABB. The setup time is crucial to the ABB and can be defined as the time prior to a data transition (Bin[0:N]) that is required to update the bus operation from VM to CM. It consists of the worst-case delay determined by the signal path of the transition detectors, control logic and control line. For long global busses, the control line signal path is dominant due to driver, interconnection, control repeater and buffering delays to the bus line repeaters. However, if the control line is assumed to be identical to the bus lines and continuously operated in CM, only the first column of bus line repeater stages are to be updated before the input data sequences (Bin[0:N]) are launched. As the first wave of data transitions propagate in the first interconnection segments, the following stages are subsequently updated to CM, similar to a falling domino effect. As a result, the setup time can be significantly reduced. Therefore, the worst-case bus line repeater spacing distance must also be considered to determine the minimum data pipeline depth in order to satisfy the setup time constraints.

The hold time constraint of the ABB can be defined as the time after the last data transition (Bin[0:N]), prior to shifting the bus lines from CM to VM, that is required to guarantee reliable data transmission. Since the control line is identical to the bus lines, the ABB has an inherent positive hold time – the "OR" gate and buffering to the bus line repeaters provide additional delay. As shown in Figure 5.12, an optional delay element such as a flip-flop (i.e. shaded) can be inserted in the control line signal path to guarantee hold time constraints. Notice that increasing the hold time does not limit the operation of the bus, however it retains the bus in CM operation for additional time dissipating static power.

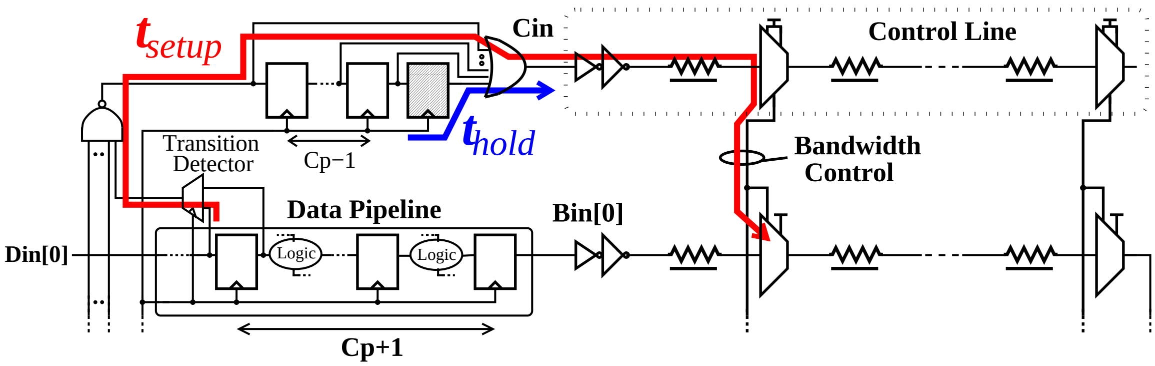


Figure 5. 12. Setup and hold time constraints of the ABB.

5.3.4. Design Considerations

As in any on-chip bus design, it would be desirable to meet the target data rate and latency constraints with the fewest possible number of repeaters with the goal of achieving lower power designs. In order to maximize the spacing between repeaters and lower the bus power dissipation, several factors must be taken into consideration.

1) Data pipeline latency. Unlike conventional busses, the ABB uses the latency built into pipeline stages of existing designs to meet its timing constraints. This implies that the maximum setup time allocated to the bus, and hence the maximum length of interconnection segments between repeaters, is determined by the available latency in the data pipeline – assuming that the signaling data rate and noise constraints are satisfied. Additionally, if the data pipeline is deep enough, the control lines can be operated using the slower VM instead of CM signaling techniques, in order to further reduce the power dissipation of the ABB.

2) Cross-talk noise. Since the voltage swings of the bus lines are reduced, induced noise from adjacent aggressor lines can deteriorate the noise margins of the repeater stages. This is particularly important if adjacent lines use full-swing VM signaling, although it's effect on cross-talk noise is not linearly proportional to the voltage swings. That is, since the bandwidth of an adjacent voltage-mode line is smaller than its current-mode counterpart, the edge-rates of the aggressor signals are also smaller, and hence partially but not entirely canceling the effect of larger voltage swings. In addition, the static low-impedance nodes at the receiving-end of the repeater stages provide better immunity to noise when compared to dynamic or capacitively terminated lines. However, the overall increased sensitivity to cross-talk noise compared to standard full-swing static busses should be taken into account via conventional techniques, such as increased metal spacing and/or partial shielding.

3) Number of control lines. A wide bus (i.e. 32 or 64 bits) can be divided into several groups of lines, each one group with its own dedicated control line. Since the ABB operates by allocating signaling bandwidth when transitions are detected in the bus lines, dividing the bus into smaller groups has the potential advantage of isolating the transition activity within the groups from each other. That is, the transition activity of the lower order bits, for instance, can be isolated from the rest of the bus lines and vice-versa. As a result, by selecting the appropriate number of control lines, it is be possible to optimize the power dissipation of the ABB for a given type of bus activity profile (i.e. instruction address bus, data bus). The disadvantage, however, is that as the number of control lines increases, the associated power dissipation increases as well. The tradeoff between the number of control lines and power dissipation are dealt in more detail in section 5.4.

5.4. ABB Power Dissipation Analysis

5.4.2. Bus Transition Statistics

The power dissipated in an adaptive bus line depends not only on the number of data transitions, contrary to conventional busses, but also on the number of cycles between each transition. The rationale for this is that the number of cycles between transitions determines the current or voltage mode operation of the ABB, and hence its power dissipation. Therefore, the probability that an n-bit wide data pattern remains unchanged for *i* consecutive clock cycles can be defined as,

$$pr_{n,i} = i \cdot \frac{N_i}{N_c} \tag{5-1}$$

where N_i is the number of occurrences in which any n-bit data pattern remains unchanged for i consecutive clock cycles before at least one transition occurs; and N_c is the data stream length in clock cycles.

To obtain the probabilities of sequential bit patterns as defined in Eq. (5-1), a time-based Alpha 21264 machine simulator program [7] was modified to extract statistics on instruction and data (i.e. load/store) address streams. A total of 100 million 32-bit instruction and data addresses were collected for benchmarks from the SPECINT2000 test suite. The addresses were grouped into half-bytes (4-bits) and the number of clock cycles before each 4-bit pattern change was accumulated.

Figure 5.13 shows the accumulated transition statistics of instruction and data address patterns extracted from the GCC benchmark (i.e. C Programming Language Compiler). The probability distribution of a 4-bit data address pattern transitioning in *i*-clock cycles, as shown in Figure 5.13a, provides useful information as to the correlation between switching activity and the number of cycles prior to transitions. In similar manner, the activity profile of instruction and data address streams shown in Figure 5.13b and 5.13c, respectively, illustrates the correlation between sequential bit patterns and bus lines (i.e. [0:3], [4:7], etc). In Figure 5.13b and 5.13c, each bar is divided into 1, 2, 3, 4, 5 and greater than 5 clock cycle bins. For instance, "1" refers to the percentage of total simulated clock cycles in which a 4-bit pattern remains unchanged for 1 cycle; "2" refers to the percentage of total simulated clock cycles in which a 4-bit pattern remains unchanged for 2 cycles; and so forth. The results indicate that the instruction addresses exhibit 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. Conversely, data addresses exhibit a more uniform switching activity distribution within the bus lines – representative of a lower probability of in-sequence address streams.

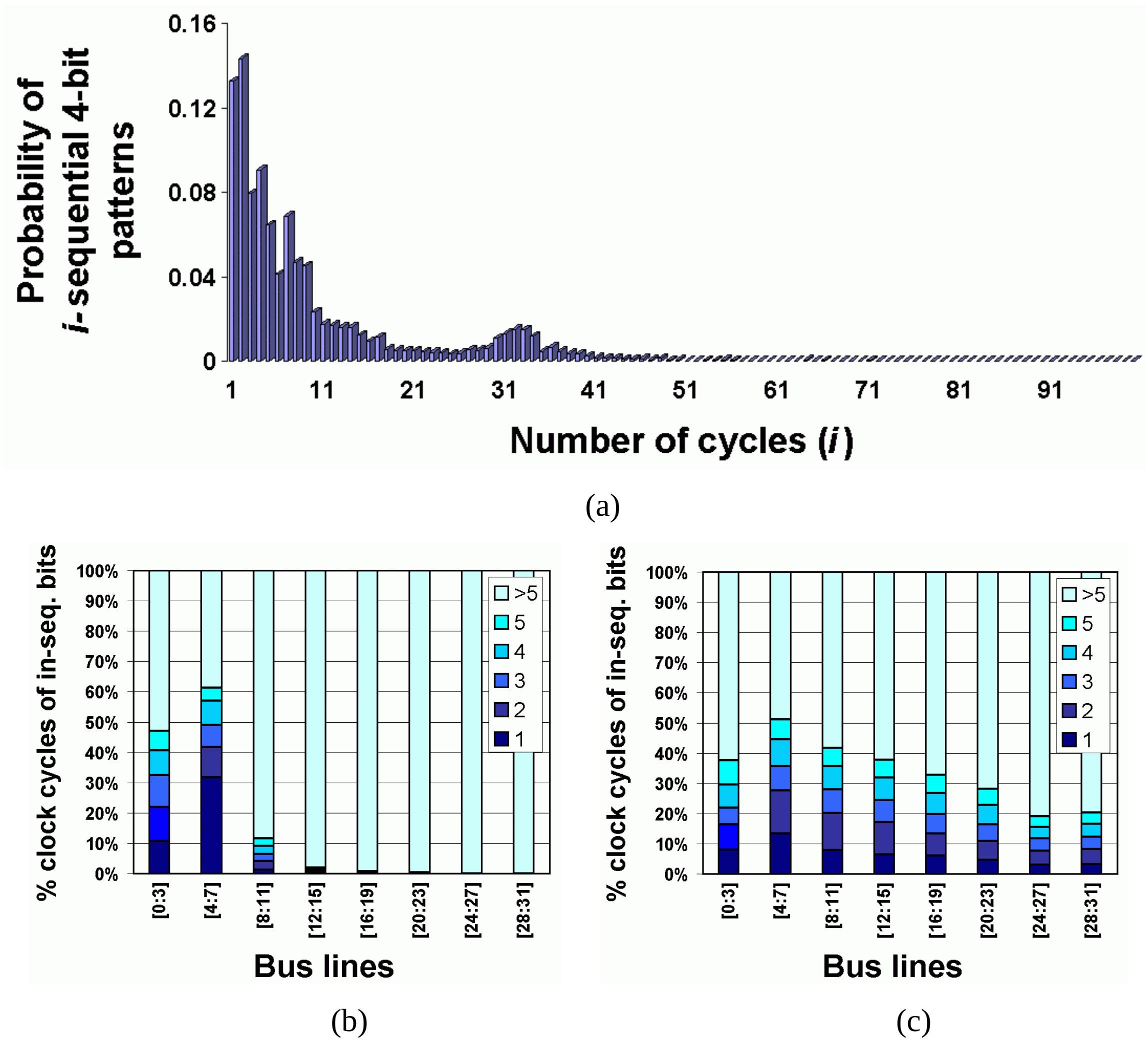


Figure 5. 13. Probability that 4-bit wide data address streams remains unchanged for *i*-consecutive clock cycles, based on the GCC program benchmark: (a) probability distribution $pr_{4,i}$ vs. number of clock cycles. Sequential activity profile as a percentage of clock cycles for (b) instruction and (c) data (i.e. load/store) address streams.

5.4.2. Analytical-Based Power Dissipation Model

The analytical procedure used to determine the power dissipation of a bus line for an arbitrary data input pattern is summarized in Figure 5.14. Using superposition, the average power P_i dissipated in a bus line is evaluated for periodic waveforms as shown in Figure 5.14 (i.e. i is the number of cycles before each transition), multiplied with the probability of sequential bits defined in Eq. (5-1), and added to obtain the total power, P_T. Referring to Figure 5.14 and evaluating the summation, we obtain

$$P_T = \sum_{i=1}^{N_c} P_i \cdot pr_{1,i} = \sum_{i=1}^{N_c} P_i \cdot i \frac{N_i}{N_c} \tag{5-2}$$

For the special case when P_i=Vdd^{2}C/iT; where C is the line capacitance and T is the clock period, Eq. (5-2) reduces to P_T=P_D=(Vdd^{2}C/T) $\sum(N_i/N_c)$, where $\sum(N_i/N_c)$ is the activity factor and P_D is the dynamic power dissipation of the bus line. Therefore, formulating the power dissipation using Eq. (5-2) yields higher accuracy since it includes all sources of power dissipation (i.e. static, short-circuit and dynamic) related to a given bus line topology. Notice that P_i can be easily extracted and tabulated from SPICE simulations.

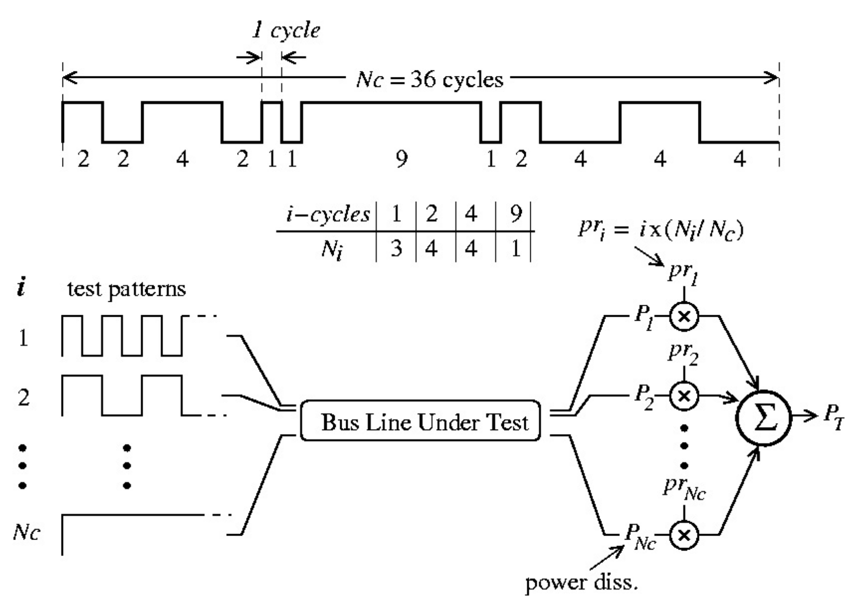


Figure 5. 14. Power dissipation model.

If we formulate the upper bound to the power dissipation of n-bus lines as $P_{n,i}=nP_i$, which is the worst case power dissipation since it assumes that all n-bus lines transition simultaneously, the worst case power dissipation in n-adaptive lines can be approximated as,

$$P_{bus,n} = \sum_{i=1}^{\mathrm{Cp}} pr_{n,i} P_{n,i} \quad + \quad P_{n,Cp} \sum_{i=Cp+1}^{Nc} \frac{\mathrm{Cp}}{i} pr_{n,i} \tag{5-3}$$

In Eq. (5-3), C_P is the ABB setup time in clock cycles and $P_{n,Cp}$ is the power dissipation of n-bus lines given that the data pattern remains unchanged for C_P cycles. The first term in Eq. (5-3) is the power dissipated while the bus operates in CM whereas the second term represents the power dissipation after switching to VM. The power dissipation in VM operation is due to the finite setup time of C_P cycles, which causes the bus to remain in CM for at least C_P cycles before switching to VM.

In similar manner, recalling that the control line will transition only if the data patterns remain unchanged for i>C_P cycles, the average power P_{ctrl} dissipated in the control line can be approximated as,

$$P_{ctrl} = \left(\sum_{i=1}^{\mathrm{Cp}} pr_{n,i} \right) P_{n,Nc} \quad + \quad \sum_{i=Cp+1}^{Nc} \left(pr_{n,i} \cdot P_{n,i-Cp} \right) \tag{5-4}$$

Equations (5-3) and (5-4) can be used to approximate the power dissipation of n adaptive bus lines.

5.4.3. Simulation Analysis

5.4.3.A. Circuit test benchmark

The overall power dissipation performance of current and voltage mode signaling for on-chip interconnects is evaluated using a circuit-level test benchmark designed in 3V-0.35μm CMOS technology with 4-Al metal layers, as shown in Figure 5.15. A 1-cm long metal-3 interconnect over metal-2 ground plane is modeled with a 1000 segment distributed RC line. The resulting total resistance (R_T) and capacitance (C_T), including fringing capacitance due to adjacent interconnects, is given by 175Ω and 2.56pF, respectively. To fairly compare the power dissipation performance of both schemes, we deliberately add inverters “I1” after the current-mode receiver interface circuit. The inverters are sized with Wp=2x10μm and Wn=2x3μm and minimum drawn length of L=0.4μm. The target maximum data rate was set at 1Gb/s (i.e. bit time T_b=1ns), which requires at least two VM repeaters, whereas no repeaters were required for CM signaling.

Figure 5.16 shows the overall power dissipation performance of the test benchmark for several *i*, where *i* represents the number of cycles in bit times (T_b) for which the logical level remains unchanged. At relatively large *i*, the VM line in Figure 5.15b exhibits lower overall power dissipation than the CM line in Figure 5.15a. This is due to the static power dissipation inherent to parallel resistive termination of CM signaling. However, as *i* is decreased, the dynamic power dissipation of full-swing VM signaling dominates. For this example, the crossover point occurs at approximately *i*=2.5 or T_b=2.5ns, which is equivalent to a bus frequency of 200Mhz (i.e. 1/5ns) – relatively small compared to current GHz processors.

Notice that the slope at which the power dissipation increases is smaller for the CM signaling case, a result due to the reduced voltage swing in the interconnect line. It should be pointed out, that unlike low-swing VM signaling schemes, CM signaling reduces the voltage swing while enhancing the bandwidth of the line. The results depicted in Figure 5.16 suggest that CM signaling is beneficial at higher signaling data-rates.

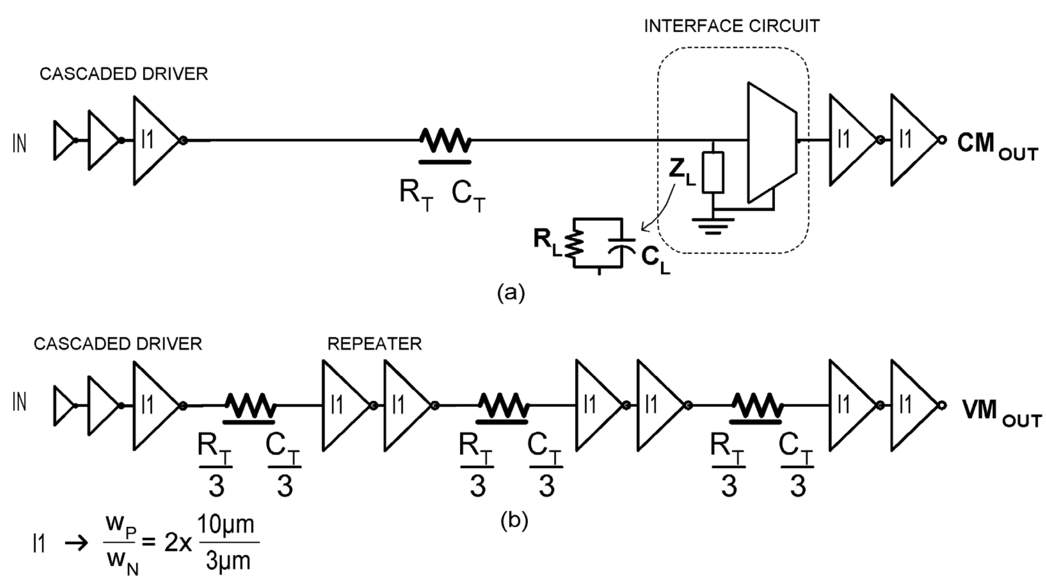


Figure 5. 15. SPICE simulation benchmark for power analysis, (a) current and (b) voltage mode. The design was based on TSMC 0.35μm parameters.

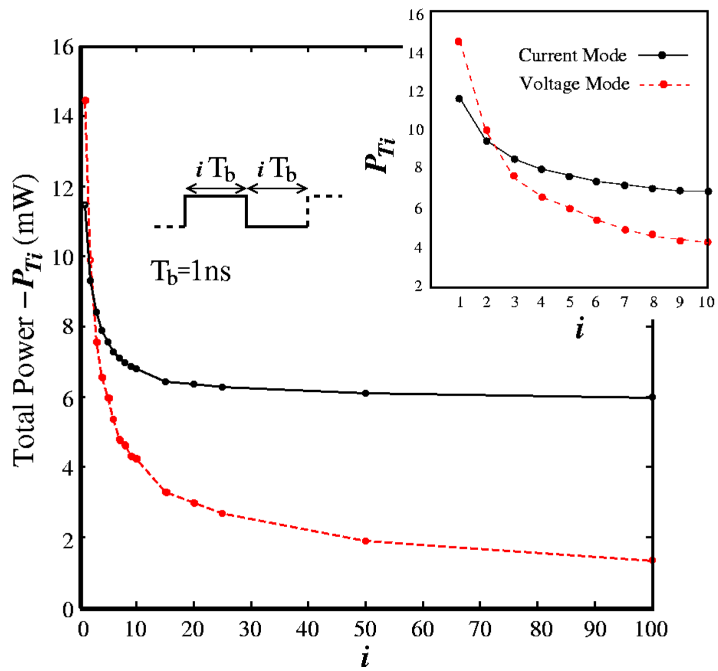
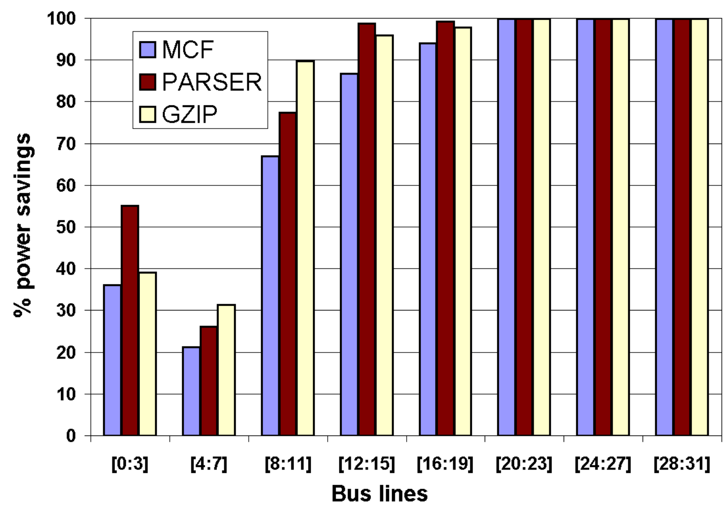


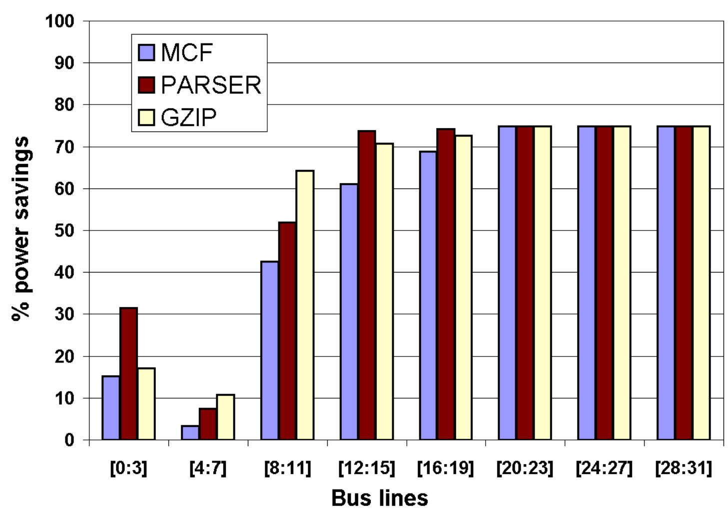
Figure 5. 16. Power dissipation comparison of current and voltage mode benchmarks depicted in Figure 15.

5.4.3.B. Power Savings

Based on the proposed power dissipation model described in section 5.4.2, instruction address streams were extracted for three benchmarks from the SPECINT2000 test suite - MCF (Combinatorial Optimization), PARSER (word-processor) and GZIP (compression) – to verify the savings in power dissipation of the adaptive bus technique over a current-mode bus. In this example, the adaptive bus uses one control line to update the state of four bus lines (i.e. total of 8 control lines for 32 bus lines). The control lines operate in current-mode and are assumed to be identical to the bus lines, with and update time latency of 3 cycles (i.e. C_P=3). Figure 5.17a and 5.17b shows the power savings without and with the added power of the control lines, respectively, indicating that higher performance gains could be obtained by minimizing the total number of control lines. The mean power savings of all three simulated benchmarks including control lines is over 50%.



(a)



(b)

Figure 5. 17. Percent reduction in power dissipation of the adaptive hybrid current/voltage mode bus technique over current-mode bus. The adaptive bus uses 1 control line per 4 bus lines, (a) performance without power dissipation of control lines, (b) with control lines included.

5.4.3.C. Power dissipation dependence on data profile and control line design

The results depicted in Figure 5.17 clearly indicate that the static power dissipation inherent to current sensing techniques - most dominant in bus lines with low switching activity - can be significantly minimized with the proposed adaptive bus. However, address busses may also exhibit low probability of in-sequence address streams, as in the case of data addresses (i.e. load/stores) for the GCC benchmark shown in Figure 5.13.

To examine the effect of varying switching activity distribution within bus lines on power dissipation, the performance of the adaptive bus is compared against both VM and CM signaling schemes. In this example, two designs for control lines are also compared, as shown in Figure 5.18. The type-I adaptive bus consists of 8 control lines, each one used to update the signaling state (i.e. CM or VM) of 4 bus lines; whereas the type-II adaptive bus uses 2 control lines, each one updating the state of 16 bus lines. The main difference between the two control line design approaches, apart from the obvious reduction in the number of control lines, is that a type-II bus will shift from CM to VM only when all 16 bus lines remain inactive for more than Cp clock cycles, whereas in a type-I bus only 4 bus lines need to be inactive. Thus, the probability that a type-II adaptive bus will remain in CM operation for a longer fraction of total simulated clock cycles is likely to be higher than the type-I bus.

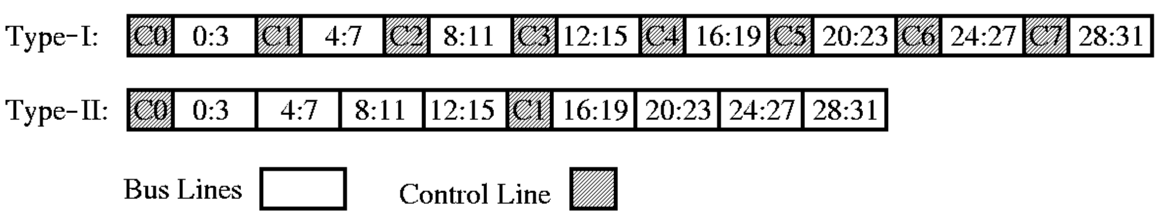


Figure 5. 18. Control line design for 32-bit adaptive bus. Type-I uses 8 control lines (1 per 4 bus lines) and Type-II uses 2 control lines (1 per 16 bus lines).

The overall power dissipation performance of a 32-bit wide bus for simulated statistics of the GCC benchmark is shown in Figure 5.19. The following observations can be inferred from these results:

1) The CM bus exhibits the highest power dissipation; nearly 2.5 and 1.4 times higher than the VM bus for instruction and data addresses, respectively – this is due to static power dissipation of CM signaling. However, the relative change in power dissipation for instruction and data address streams is only 10% for the CM bus whereas the VM bus changes by 94%. This indicates that CM signaling is more suitable for increasing switching activity, an effect due to the reduction in voltage swings.

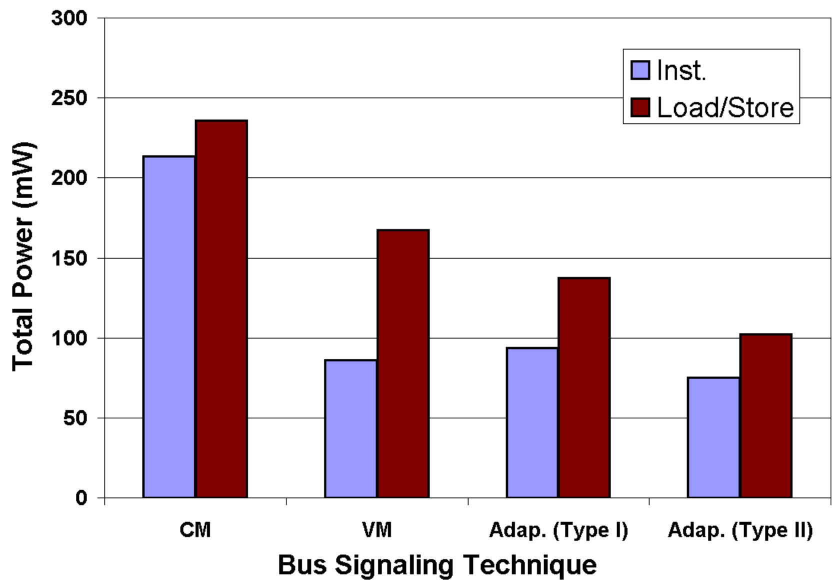


Figure 5. 19. Total power dissipation comparison for GCC benchmark.

2) The type-II adaptive bus outperforms the type-I bus for both instruction and data address streams. In fact, the type-II bus remains in CM operation for a longer percentage of total simulated clock cycles – because the probability of all 16 bus lines remaining inactive is likely to be lower than 4 bus lines remaining inactive. However, there is an increase in power dissipation due to the additional control lines of the type-I bus, making the type-II bus more suitable.

3) The type-II bus exhibits nearly 13% and 40% improvement over the VM bus for both instruction and data address streams, respectively, and up to 65% power savings over the CM bus.

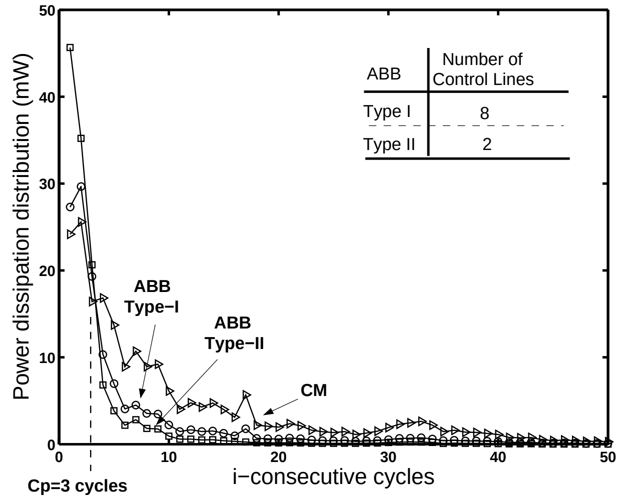
In addition to the power savings of the adaptive bus technique, an important result that stems from using CM signaling is the reduction in the number of repeaters. As shown in Table 5.3, the 32-bit type-II adaptive bus can achieve the target data rate of 1Gb/s across a 1-cm long wire with 34 instead of 96 repeaters/receivers required for the VM bus.

Table 5. 3. Total number of repeaters and receivers for several bus signaling schemes.

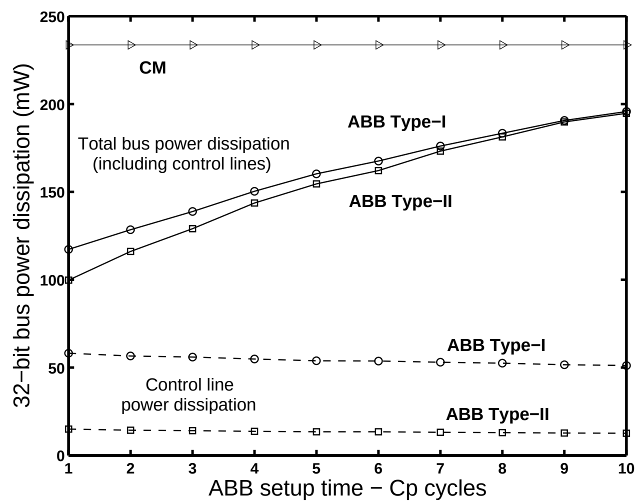
	CM	VM	Adaptive (Type-I)	Adaptive (Type-II)
Number of Repeaters + Receivers	32	96	32+8=40	32+2=34

5.4.3.D. Power dissipation dependence on sequential data (*i*) and setup time (C_P)

Figure 5.20a shows the power dissipation distribution versus the number of sequential address streams expressed in clock cycles of a 32-bit 1Gb/s/line adaptive and current-mode bus for the data activity profile in Figure 5.13c. The ABB minimum pipeline latency or setup time is assumed to be C_P=3 clock cycles. In addition, both type-I and type-II adaptive bus designs are considered.



(a)



(b)

Figure 5. 20. ABB power dissipation analysis for GCC data address streams shown in Figure 5.13c. (a) Power dissipation distribution of adaptive and CM busses versus sequential data patterns and, (b) ABB power dissipation dependence on setup time (C_P).

Since the probability that a type-II adaptive bus will remain in CM operation for a longer fraction of total simulated clock cycles is likely to be higher than the type-I bus, the power dissipation is higher for sequential patterns under C_P cycles, as shown in Figure 5.20a. However, as the number of sequential patterns increase above C_P cycles, the type II bus dissipates less power due the increased control line overhead of the type I bus. This overhead is evident in Figure 5.20b, which illustrates the overall power dissipation of the bus as the setup time C_P is increased. As C_P increases, the ABB remains in CM operation for a larger number of cycles causing an overall increase in power dissipation. Notice that although the power dissipation distribution of the ABB is unique to each data activity profile, the adaptive behavior remains the same. That is, above C_P cycles of sequential data streams, the ABB dynamically minimizes the power dissipation.

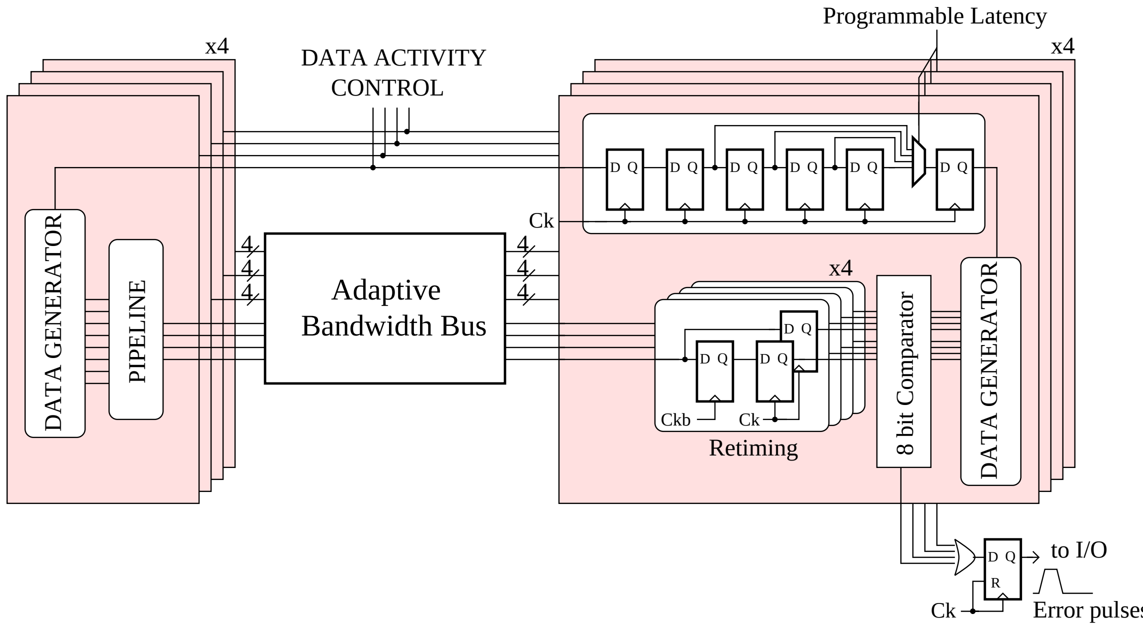
5.5. ABB Circuits and Performance Evaluation

5.5.1. Experimental ABB Test Chip

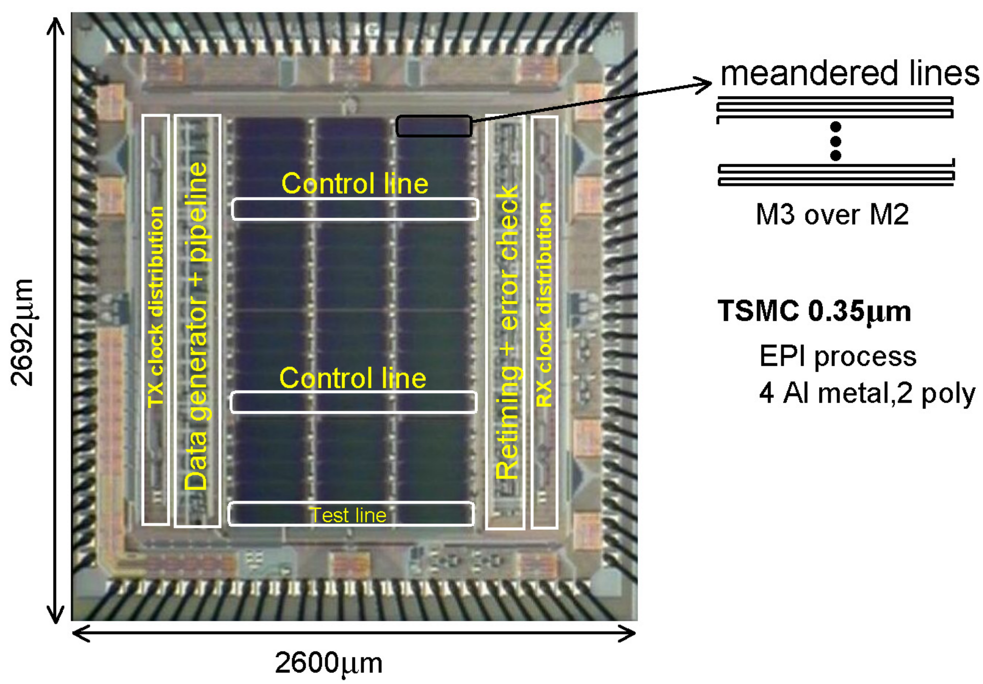
As described in section 5.3.2, the circuits required to implement the ABB consists of transition detectors, flip-flops and hybrid current/voltage mode repeaters. Although drop-in replacement of static bus implementations is possible, additional design is required to ensure setup time constraints of the ABB. Moreover, since the bus is designed for pipeline compatible circuits, the overhead of the transition detectors and the small FIFO that precedes the control line (see Figure 5.10) must be taken into account.

The architecture of the experimental test chip used to evaluate the performance of the ABB is shown in Figure 5.21a. The ABB consists of 16 bus lines operating over on-chip interconnects spanning 1.75cm in length at a maximum of 1Gb/s each, with two additional and identical control lines (i.e. 1 control line for every 8 bus lines). A 32-bit pseudorandom input data sequence profile is generated with four independent and externally controlled 8-bit data generators, pipelined with C_P=2.5 cycles of latency to model a data pipeline, and 2X sampled with both edges of a 500MHz clock to generate a maximum of 16Gb/s of total aggregate NRZ data bandwidth. For experimental comparisons, the bus can be operated in current, voltage and adaptive modes. In addition, the output resistance of the hybrid repeater stages in each bus line can be varied to characterize the bus input-to-output latency. At the receiving-end of the bus, the output is retimed to generate 32-bit parallel data sequences. To ensure proper bus operation and retiming during testing, a programmable latency controller and an error checking circuit are built into the chip to generate error pulses, which can be viewed in an oscilloscope.

The chip microphotograph of the experimental bus is shown in Figure 5.21b. All bus lines in the 7mm^2 test chip are fabricated using metal-3 over metal-2 interconnects and meandered due to limited die area. The chip was fabricated in 3.3V-0.35µm EPI CMOS technology and wire bonded to a pin-grid array (PGA) package, which is attached to a PGA socket mounted on a 4-layer custom test board. Due to the high inductance per pin (i.e. an average of 12nH [8]) and multiple signal path discontinuities, all high-speed I/O circuits were implemented with differential topologies, terminated with on-chip resistors and limited to 300-500mV I/O swings (see section 5.5.5).



(a)



(b)

Figure 5. 21. Test Chip (a) architecture (b) chip microphotograph.

5.5.2. Pseudo-random Data Profile Generator

A 16-bit high-speed data activity profile is generated from four independently controlled 8-bit pseudo-random maximum length linear feedback shift registers (LFSR). Each LFSR generates an 8-bit parallel data sequence and is sampled with both clock edges to generate 4-bit data at twice the speed, as shown in Figure 5.22a. Therefore, the core digital logic can be operated at 500MHz while generating 1Gb/s NRZ data sequences.

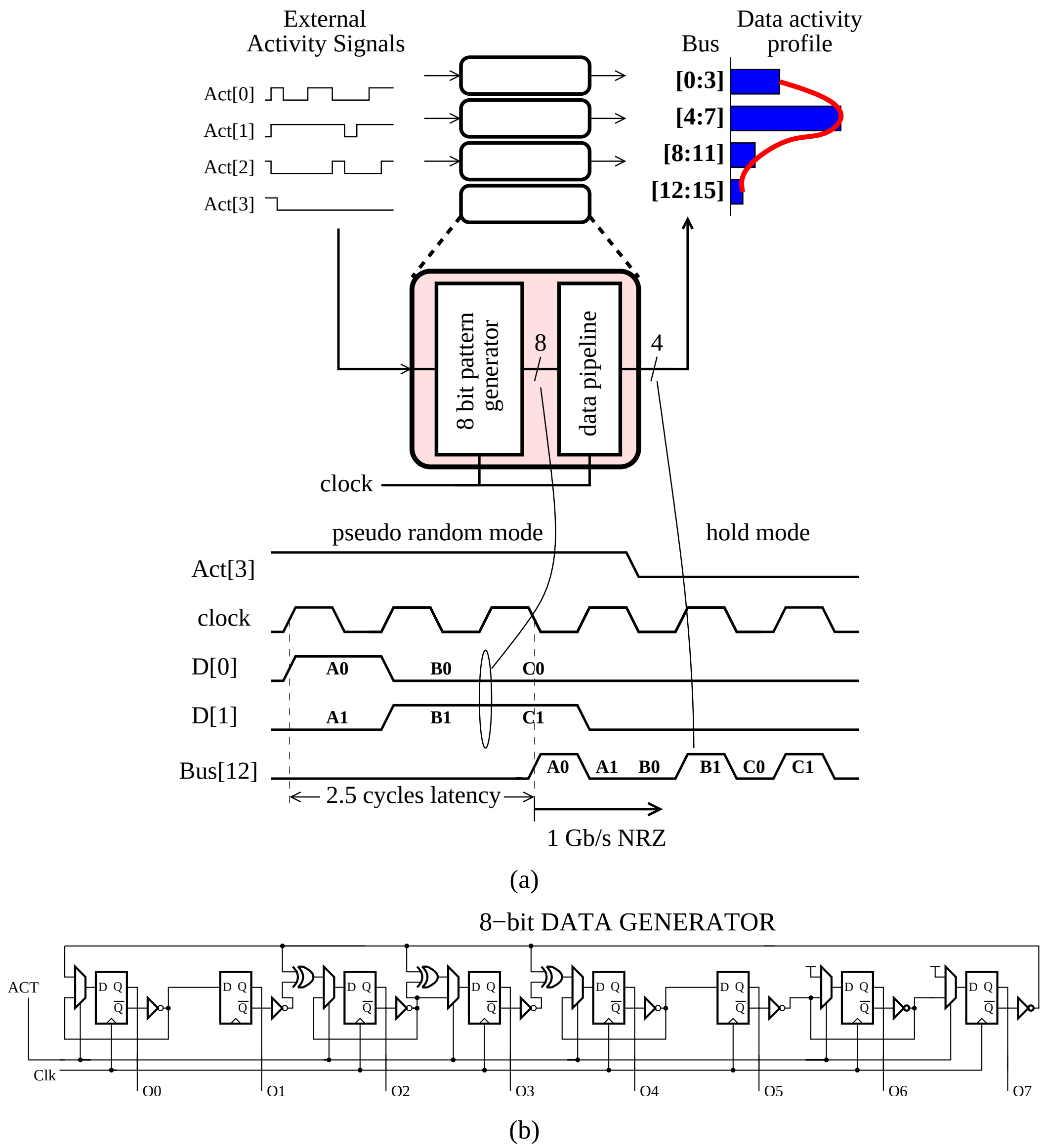


Figure 5. 22. High-speed data profile generator (a) block diagram and timing (b) 8-bit data generator topology.

The LFSR is modified to enable two independent modes of operation via an externally controlled signal “Act” from a High Frequency Stimulus system (HFS 9000). The modified LFSR is shown in Figure 5.22b. A high logical level in “Act” indicates pseudo-random operation, whereas a logical low level holds the previous data pattern. Therefore, the external “Act” signals can be programmed to generate an arbitrary data profile. This feature is used to emulate realistic data activity profiles similar to those extracted in section 5.4.

5.5.3. On-chip Clock and Power Supply Distribution

Figure 5.23 shows the 3-level H-tree clock distribution scheme implemented on both transmit (TX) and receive (RX) sides of the ABB test chip. The clock tree consists of differential buffers, each one use to drive four additional buffers. The outputs of the last 16-stages are shorted together to minimize the effect of localized device variations on clock skew. To ensure minimum clock skew between the differential clocks, each clock signal is generated from both true and complementary clocks from the preceding stages and followed by two cascaded inverters to reduce the effect of PMOS and NMOS device variations. Metal-3 and metal-4 layers were used to distribute the clock signals and the lengths at each stage of the clock tree are made equal to reduce the effect of wire loading.

A differential on-chip clock distribution scheme was also employed to minimize the effect of current spikes from clock buffers on supply and ground voltage. Although these noise peaks occupy a wide-band frequency spectrum, it is dominant at multiples of the clock frequency and is amplified by the inductance in the supply network [9]. Multiple supply and grounds pins (i.e. 6-Vdd and 6-Gnd) were used to reduce the di/dt noise from the package inductance.

To provide a stable power supply, wide metal-1 and metal-2 lines with low resistance and decoupling capacitance cells are distributed on-chip. Fig 5.23 shows the cross-section of the power and ground lines. On-chip decoupling cells are placed under the power and ground lines, which are routed alternately to minimize the parasitic resistance of the substrate contacts. The decoupling capacitors in each cell are made of large N-fet transistors with the gate tied to the supply, whereas the source, drain and bulk contacts are tied to ground, as shown in Figure 5.23. Notice that the gate-to-substrate capacitance of a 4-terminal MOS transistor biased in strong inversion does not vary significantly at higher frequencies because the inversion layer charge is removed or provided by the heavily doped source/drain regions. The frequency dependent gate-to-substrate capacitance behavior is present in 2-terminal devices, in which the inversion layer is in a sense isolated from the outside world and the electron concentrations is changed only by slow mechanisms of thermal generation and recombination [10].

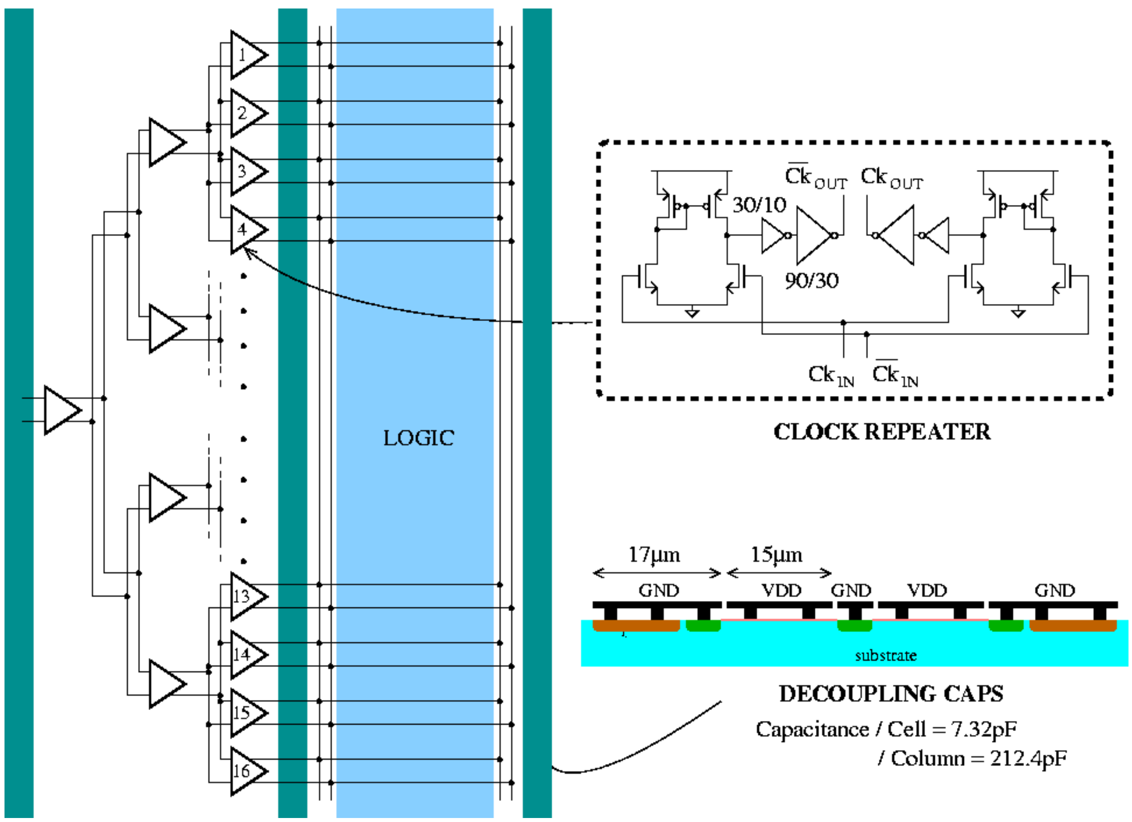


Figure 5. 23. On-chip clock distribution and power/ground lines with decoupling capacitor cells.

Localized power supply grids are distributed to clock buffers and logic circuits from three columns of global supply/ground rails, and placed in close proximity to the circuits to minimize substrate noise. The close proximity of decoupling capacitors is required since the resistance between two substrate contacts in an epitaxial process with a heavily doped p-type bulk is independent of distance once a seperation of approximately four times the EPI layer thickness has been reached [11]. That is, the substrate can be regarded as a single node and any noise injected to the substrate via the epitaxial layer will spread throughout the chip. Since the EPI layer thickness of a standard 0.35μm CMOS technology is approximately 4μm, the edges of decoupling capacitor cells were placed as close as 1-2μm from noisy clock buffers and digital circuits.

The total decoupling capacitance allocated to the core logic and clock distribution network on each side of the test chip is approximately 637pF. This is 70 times the combined load seen by the clock buffers on the transmit side of the bus, which is approximately 9.2pF - i.e. CK/CKb loads are 4.36pF/2.98pF, respectively. On the receive side of the chip, the total clock load is slightly smaller at approximately 7.3pF – CK/CKb loads are 3.8pF/1.56pF, respectively. It should be pointed that the interconnection loading of both CK and CKb networks on each side of the chip is approximately 1.9pF, and therefore cannot be neglected during the design stage.

The clock loads is mainly due to the 230 flip-flops that make up the bulk of the digital core logic. The circuit topology of the flip-flops was chosen according to the clock load/FF, the speed and the setup/hold time constraints. In order to exploit the advantages of each FF, two circuit topologies were chosen for different datapaths. In Figure 5.24a, the Power PC FF [12] has the advantage of low internal power dissipation and the positive setup time that is characteristic to master/slave flip-flops [13]. Therefore, it was used in the data pipelines and FIFOs stages to ensure hold-time constraints – i.e. assuming that the skew between CK and CKb is less than one inverter delay, the positive setup-time virtually ensures hold time constraints.

For the high-speed datapaths where a logic depth of at least 6-7 gate delays was required, as in the LFSR circuits, the semi-dynamic flip-flop (SDFF) circuit topology shown in Figure 5.24b was used [14]. This circuit has the advantage of negative setup time with respect to the rising clock edge and conditional evaluation that reduces the delay dependence on process variation. The negative setup time, in particular, is useful as the operating frequency is increased since it allows the data edges to arrive slightly after the clock rising edge. Notice, however, that a negative setup time can cause minimum path delay errors due to hold time constraints. The clock load in gate-microns of the SDFF was approximately 24μm, slightly lower than the combined load of the PowerPC FF at 34.4μm (i.e. 17.2μm/clock signal).

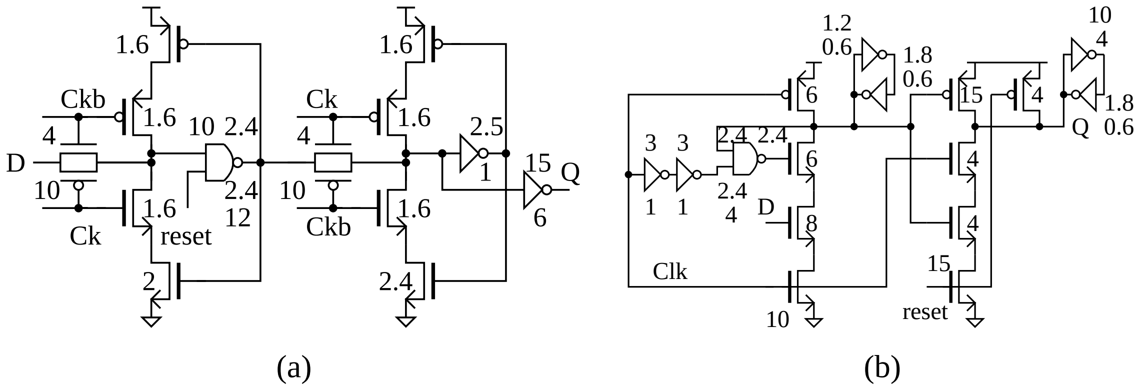


Figure 5. 24. Flip-flop circuits used (a) PowerPC [12] and (b) semi-dynamic FF [14].

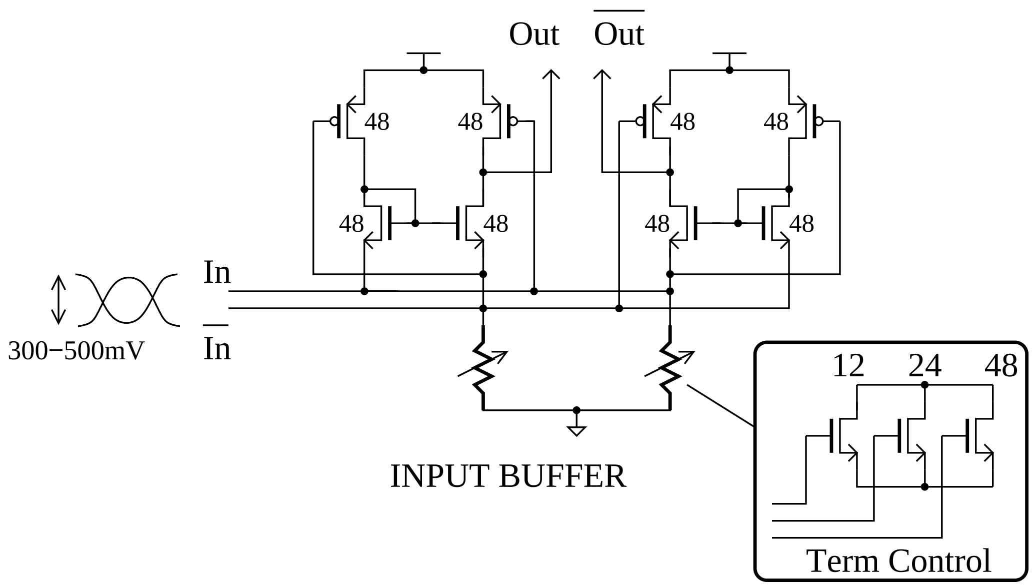
5.5.4. I/O Interfacing Considerations

5.5.4.A. I/O circuits

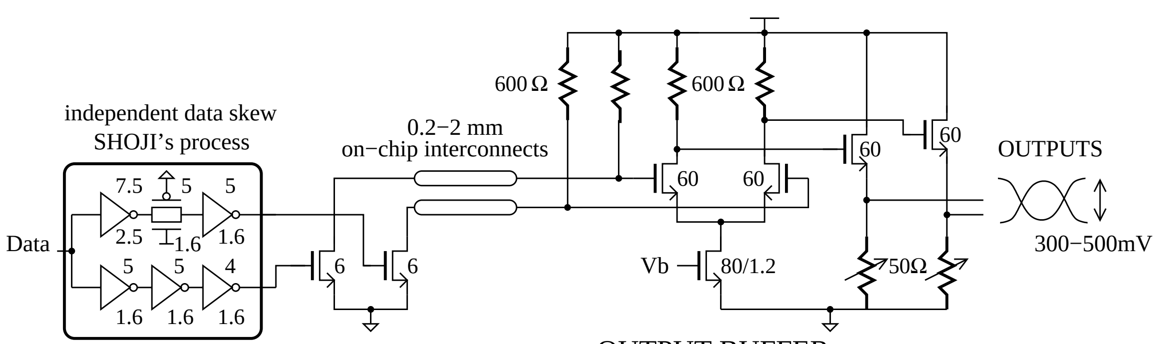
The ABB test chip, when fully operational, requires two 500MHz input clocks (i.e. TX and RX) and four 500Mb/s activity signals, while generating up to four 1Gb/s NRZ output scan signals. Since the performance of the ABB test chip depends to a large extent on the quality of the I/O signals, the I/O buffer design was carefully addressed to guarantee proper operation.

In order to achieve high data-transfer rate in low-performance packages, differential I/O buffers with on-chip terminations are used. On-chip termination is preferable over off-chip termination because of unwanted reflections that are generally introduced due to package parasitics. In addition, since electro-static discharge (ESD) protection is not used in order to maximize the I/O bandwidth, the on-chip terminations provide a discharge path for electro-static events. The buffers were designed to be operational with 300-500mV swings, which are referenced to ground to maintain compatibility with testing equipments.

Figure 5.25a shows the circuit topology of the input buffer. It consists of two cross-coupled direct-drive Chappell receivers [15] for pseudo-differential signal amplification. The signals are applied directly to the source nodes of the n-channel devices. The high gain common-gate stages and the additional overdrive from the diode-connected devices provide fast amplification of very low signal amplitudes (i.e. 300mV). The output signals, which are amplified to nearly CMOS level, are distributed differentially across the chip and converted to single-ended with a grounded source-coupled stage. The termination resistance consists of three parallel n-channel devices, which are externally controlled to compensate for process variations. Due to the small input signal swings referenced near the chip ground, the resistance value of the termination devices biased in triode region remains fairly constant.



(a)



(b)

Figure 5. 25. I/O circuits (a) input buffer (b) output buffer.

To ensure good signal integrity for the high-speed output signals, the differential scheme shown in Figure 5.25b is used. Complementary signals are generated from single-ended internal scan signals using a process independent data skew approach [16]. This technique is based on independently matching the sum of the pull-down delays of one logic chain to the sum of pull-down delays of the second logic chain. Similarly, the sum of pull-up delays of both logic chains is also forced to be equal. Since the pull-up and pull-down delays are independently matched, process variation in n-channels or p-channel devices affects both chains equally, and process dependent skew can be eliminated to a large extent.

These complementary signals are distributed across chip using open-drain source-coupled n-channel devices with differential interconnects to minimize the effect of cross-talk, as shown in Figure 5.25b. The interconnect length varies from 200-2000μm, and are terminated with 600Ω pull-up resistors in the output buffer cells. The output drivers consists of differential amplifiers followed by source follower buffers, which are terminated on-chip with n-channel devices operated in triode region. The buffers are designed for 300-500mV signal swings.

To a large extent, the differential I/O and signal distribution scheme was chosen to minimize ground bounce and supply noise due to all ten I/O buffers switching simultaneously. However, even in differential signaling, a small skew between the two signals can generate large current spikes that are easily coupled to the low resistance bulk affecting the entire chip. Ensuring equal board trace lengths can minimize the skew of the high-speed differential input signals. In addition, common-centroid layout is implemented in all I/O circuits to minimize the effect of device mismatch.

To further reduce supply noise, dedicated power supplies are used for the input and output buffers. In addition, on-chip decoupling capacitance is used for each I/O cell. Every input and output buffer has 50pF and 10pF of decoupling capacitance, respectively. The smaller decoupling capacitance of the output buffers is due to area limitation, and has the negative effect of shifting the power supply resonant frequency closer to the nominal output buffer operating frequency. To avoid noise accumulation due to the possibility of oscillations in the output buffer supply network, the effective series resistance (ESR) of the on-chip decoupling capacitors is intentionally increased.

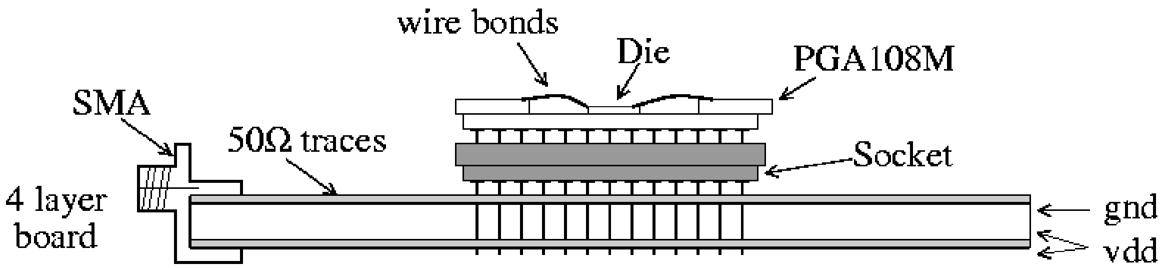
Although this results in a slight increase in resistive voltage drop, the damping factor of the supply network is also increased, thus minimizing the risk of noise accumulation due to power supply resonance [17].

P-type guard rings are extensively used in close proximity to all I/O circuits to minimize substrate noise. Dedicated ground pins are used to bias the substrate contacts in order to decouple the I/O ground noise from the bulk and improve the effectiveness of the guard rings [11]. Therefore, the source connections of all n-channel devices used in the I/O circuits and I/O decoupling capacitors are separated from the bulk connection.

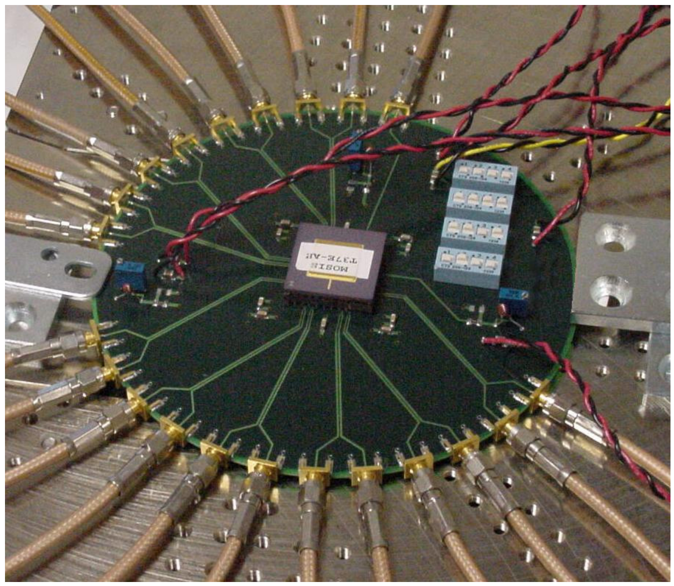
5.5.4.B. Packaging and Board design

To accommodate the large number of differential high-speed signals in a 50-Ω environment, a four-layer board with top and bottom pre-peg stack-up thickness of 7.5 mils is used. With a dielectric constant of approximately 4.5, the trace widths with ½ oz. Copper are roughly 24mils. The traces are tapered to 7mils near the package due to routing density constraints. 50-Ω differential microstrip lines are placed on the top layer with a ground plane underneath. The multiple power supply planes and biasing signals are routed in the third and bottom layers, as shown in Figure 5.26a. To reduce supply noise, the board's power supplies are tightly decoupled with discrete capacitors, ranging from 0.1uF to 4.7uF. Since the smaller capacitors have lower parasitic inductance and higher self-resonant frequencies, a wide range of noise frequencies is bypassed. In addition, the internal board capacitance between the second ground and third supply layers provide additional noise decoupling.

A medium sized cavity 108 pin-grid array (PGA) package provided by MOSIS is used to electrically connect the chip die to a socket, as shown in Figure 5.26a. The parasitic inductance of the I/Os due to wire bonds (1nH/mm), PGA traces (4-7nH) and pins, and socket pins can be significant, ranging from 7nH – 15nH [18]. The socket is soldered to the center of the test board, as shown in Figure 5.26b. A circular test board of approximately 6-inches in diameter is fabricated in order to minimize the difference in flight time in all high-speed I/O signals. The 50-Ω traces, which are mounted at the board edges using SMA connectors, span approximately 14cm in length.



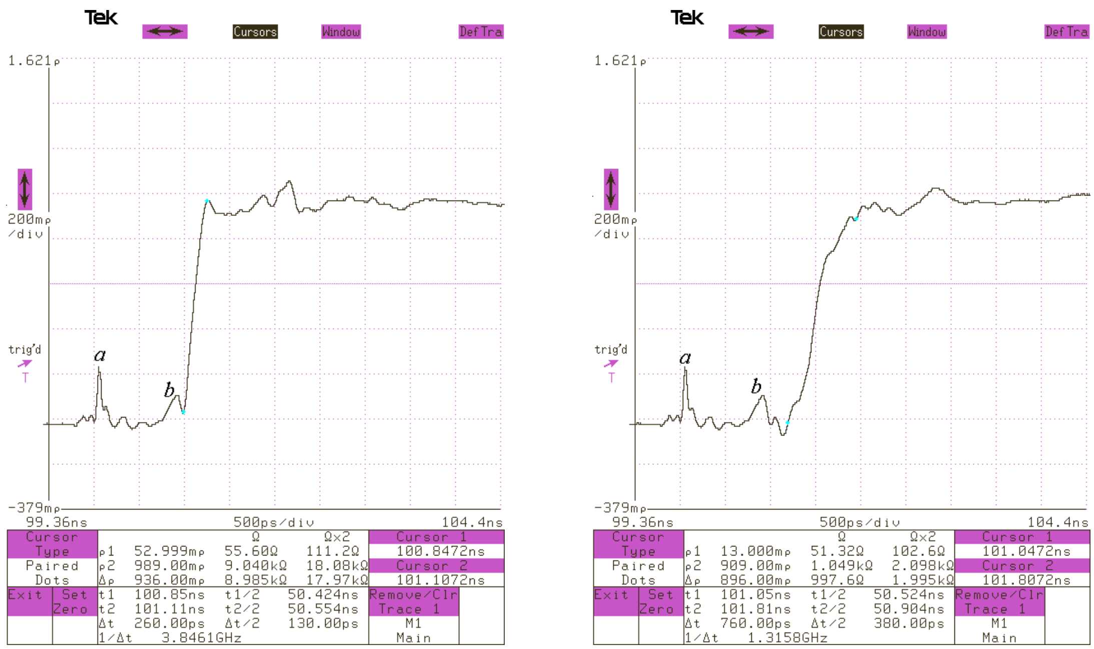
(a)



(b)

Figure 5. 26. Test board setup (a) board and package cross-section (b) experiment test board.

Figure 5.27 shows the time domain reflectometry (TDR) of the high-speed differential signal path. In Figure 5.27a, the PGA package is removed from the test setup to study the effect of the SMA edge launcher, signal traces and socket. The peak at point *a* is due to the SMA edge launcher which appears to represent an inductive effect that arises from a failure to bring the board's reference planes all the way to the edge of the board [19]. Following this peak, the trace impedance is approximately 50-Ω as designed. At point *b*, the impedance increases slightly and is followed by a dip due to the socket and via pin capacitance. The increase in characteristic impedance is due to tapering the differential microstrip lines near the socket fixture. The true edge rate of the signal path at the socket end is approximately 260ps.



(a) (b)

Figure 5. 27. TDR measurements of high-speed I/O traces with (a) socket and (b) PGA package inserted into socket (without on-chip termination).

The signal path degrades significantly as the PGA package is inserted into the socket, as shown in Figure 5.27b. The TDR shows a bigger dip that follows point *b* due to the package pin capacitance. Notice that the measured TDR after inserting the PGA package into the socket shows a slower true edge-rate of 760ps – this does not include on-chip resistive termination. The degradation in edge rate is partly due to the multiple reflections as a result of impedance discontinuities between the board, socket and PGA package. Therefore, the on-chip termination resistors are expected to improve the signal edge rates of the high-speed differential lines (see section 5.5.6).

5.5.5. Hybrid Current/Voltage Repeater

The hybrid repeater circuit, shown in Figure 5.28a, consists of a variable input resistance amplifier followed by a variable output resistance driver, the later of which is used to characterize the effect of driver resistance on line delay. The input amplifier consists of a self-biased inverter stage with a feedback transmission gate, isolating the input from the internal node ‘A’ when the control voltage (V_{CTRL}) is set to ‘0’ (i.e. voltage-mode), and otherwise providing a low resistive termination when V_{CTRL} transitions to ‘1’ (i.e. current-mode). The additional leg of transistors on the internal node is used to adjust the duty cycle, the input of which is typically biased at Vdd/2. When operating in CM, the low swing voltage at the input node labeled ‘IN’ is amplified in two stages to CMOS level at node ‘B’, with approximately 300ps of conversion latency.

As shown in Figure 5.28, the bus lines are divided into 3 equally spaced segments spanning 5830μm in length with a line width of 2μm – with approximately R_T=204Ω, C_T=1.07pF per segment. Figure 5.28b shows the effect of driver resistance and supply voltage variation on interconnection delay, measured from the input-to-output 50% waveforms points. The worst-case measured delay variation was 8.56ns and 1.54ns for VM and CM sensing, respectively, an improvement of 5.5x. At the nominal supply voltage of 2.5V, the reduction in CM delay is in the range of 35% to 70%.

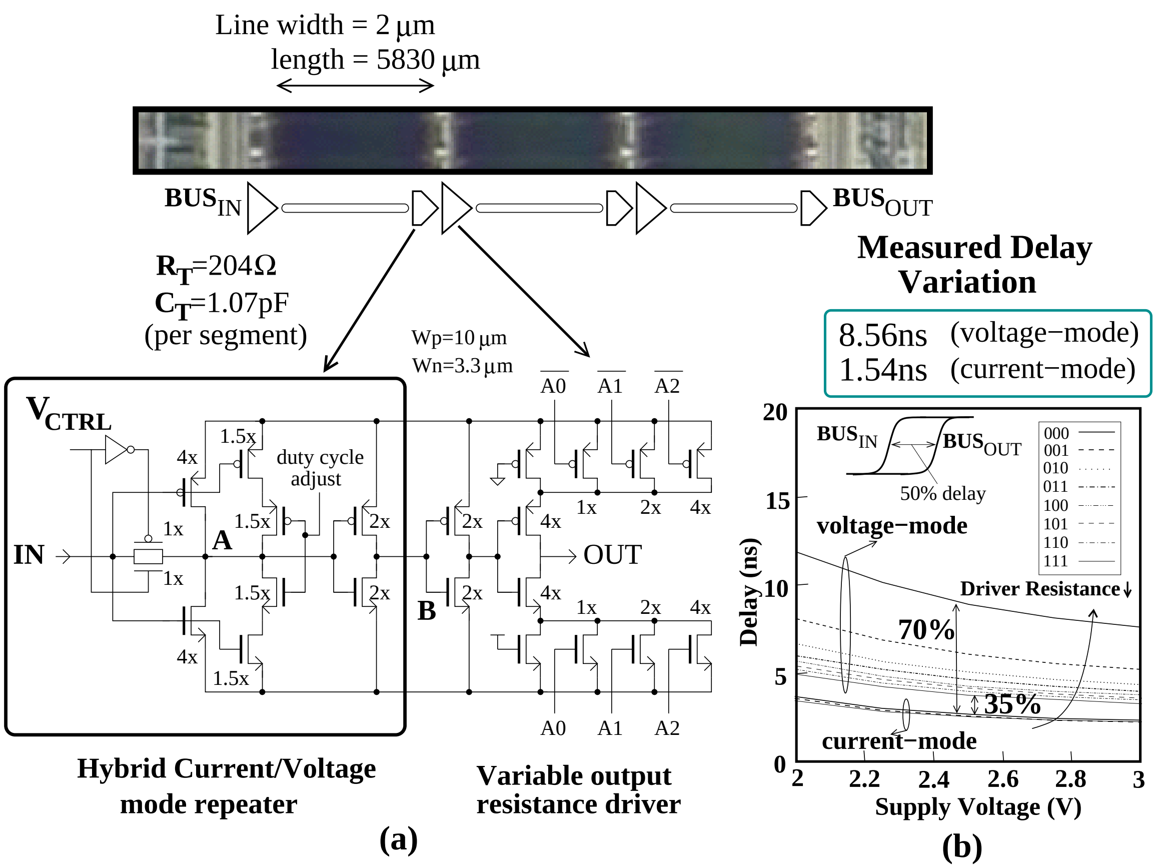


Figure 5. 28. Repeater HCVR (a) circuit (b) measured delay versus supply voltage and driver resistance.

5.5.6. ABB Waveforms

Figure 5.29 shows relevant ABB waveforms corresponding to an adaptive bus line (i.e. bus[16]). The waveforms illustrated in Figure 5.29 indicate that as input data transitions are sensed, the control signal switches to set the ABB in CM operation, and otherwise reverts to VM to eliminate the DC power. This adaptive operation is shown in Figure 5.30, where the dynamic change in voltage swing at the input of the repeater is apparent. In the absence of data transitions, the internal nodes of the hybrid repeaters are ideally isolated from the bus line causing the line voltage to reach the supply rails. When data transitions are sensed, the input resistance to the repeater is decreased to allocate the required interconnection bandwidth and reduce the signal voltage swing. Although, this reduction in voltage swing results in higher sensitivity to cross-talk compared to standard full-swing busses, it provides better performance compared to low-swing VM techniques due to the static low-impedance nodes at the receiving-end of the repeater stages.

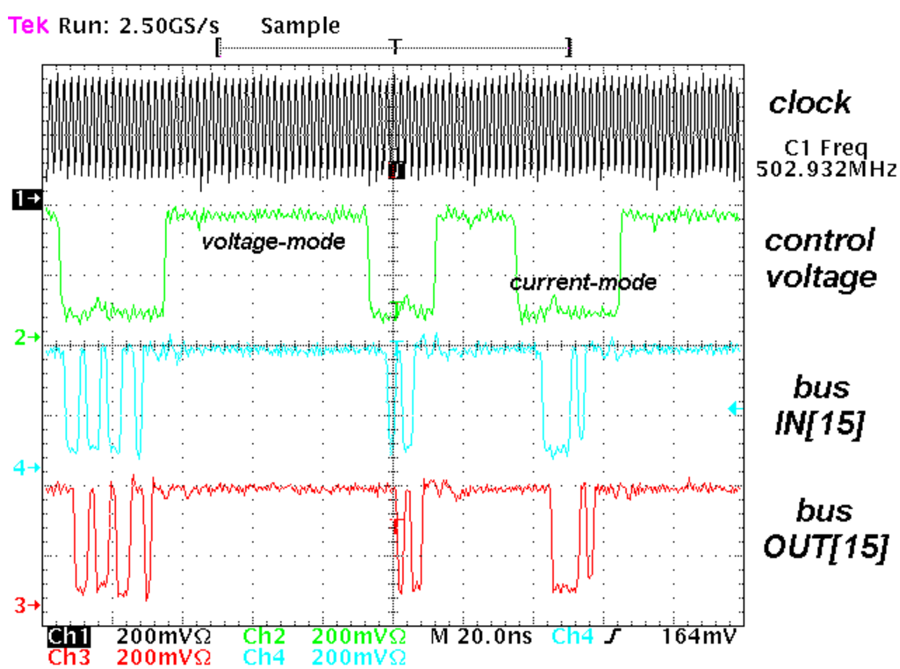


Figure 5. 29. Measured I/O waveforms of an adaptive line (clock frequency ~ 500MHz, dual-data rate).

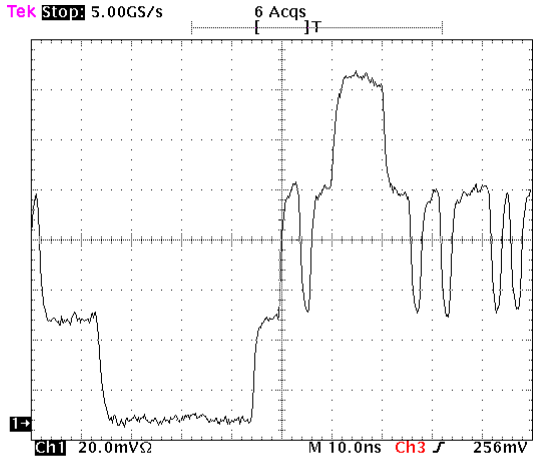
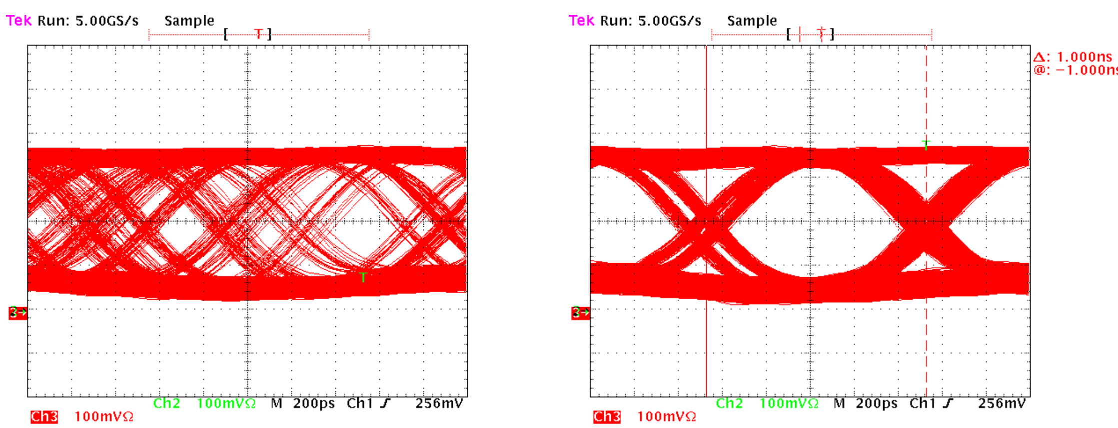


Figure 5. 30. Probe measurements of internal repeater waveform indicating adaptive bandwidth operation.

Figure 5.31 shows measured output eye diagrams at 1Gb/s operation. In Figure 5.31a, the bus is operated in voltage-mode and the output eye is completely closed due to ISI. Figure 5.31b shows the output eye when the adaptive bus is enabled resulting in an eye-width of approximately 650ps. The output eye-width opening at several data-rates for the adaptive bus is indicated in Table. 5.4.

Table 5. 4. Eye-width opening versus line data rate for the adaptive bus.

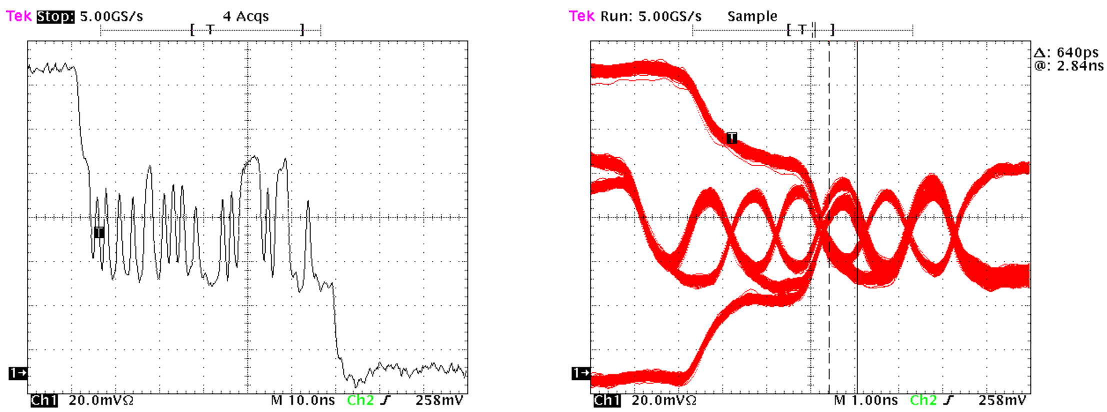
Data rate/bus line	1Gb/s	0.8Gb/s	0.6Gb/s
Eye-width	650ps	930ps	1.4ns



(a) (b)

Figure 5. 31. Bus line output eye-diagrams at 1Gb/s (a) voltage-mode (b) adaptive-mode.

Measurements using high-impedance probes show the adaptive operation at the input of the last repeater stage for a pseudo-random pattern at 1Gb/s, as shown in Figure 32a, and its corresponding eye diagram in Figure 5.32b. The worst-case eye-width of 640ps occurs when the line switches from VM to CM and is subsequently followed by a 1-0 or 0-1 pattern. Notice that the ABB setup time (i.e. the time required to update the line from VM to CM operation) is approximately 2.5ns (i.e. C_P=2.5), as shown in Figure 5.32b.



(a) (b)

Figure 5. 32. Measurements using probes (i.e. 20x signal attenuation) at the input of the last hybrid repeater stage given a pseudo-random data pattern at 1Gb/s. (a) Internal adaptive line operation and (b) eye-diagram.

5.5.7. Substrate Noise

A simple common-gate amplifier is used to sense the substrate noise of the test chip in order to evaluate the noise injected by the ABB. The amplifier is placed away from noisy digital circuits and surrounded with guard rings with a dedicated pin to avoid noise modulation via the amplifier's bulk node. Dedicated pins are also used for the supply and gate bias voltage, as shown in Figure 5.33. Substrate noise coupling capacitors are placed underneath the bus lines and are implemented with finger-structured MOS capacitors. The substrate contacts are connected to drain/source nodes. Each capacitor consists of eight fingers with W/L=80/1.6. A total of 19 capacitor cells are placed along the center of all the bus lines, which adds to approximately 87.5pF. The substrate noise is coupled to the capacitor's gate node, and distributed with appropriate noise shielding to the source node of the common-gate amplifier.

Figure 5.33 shows an arbitrary data pattern (Bus[16]) and the substrate noise voltage (V_N) due to input data activity in all 16 bus lines. The maximum measured peak-to-peak noise voltage is 370mV - the simulated gain of the amplifier is 2, so the actual substrate noise is 185mV. Similarly, when 8 bus lines are simultaneously activated the measured peak-to-peak voltage is 286mV (or actual 143mV). Notice that the worst-case peak noise occurs when the bus lines adaptively switch from current to voltage mode. The peak noise, which is mainly due to the high parasitic inductance of the package, is well within acceptable margin of operation.

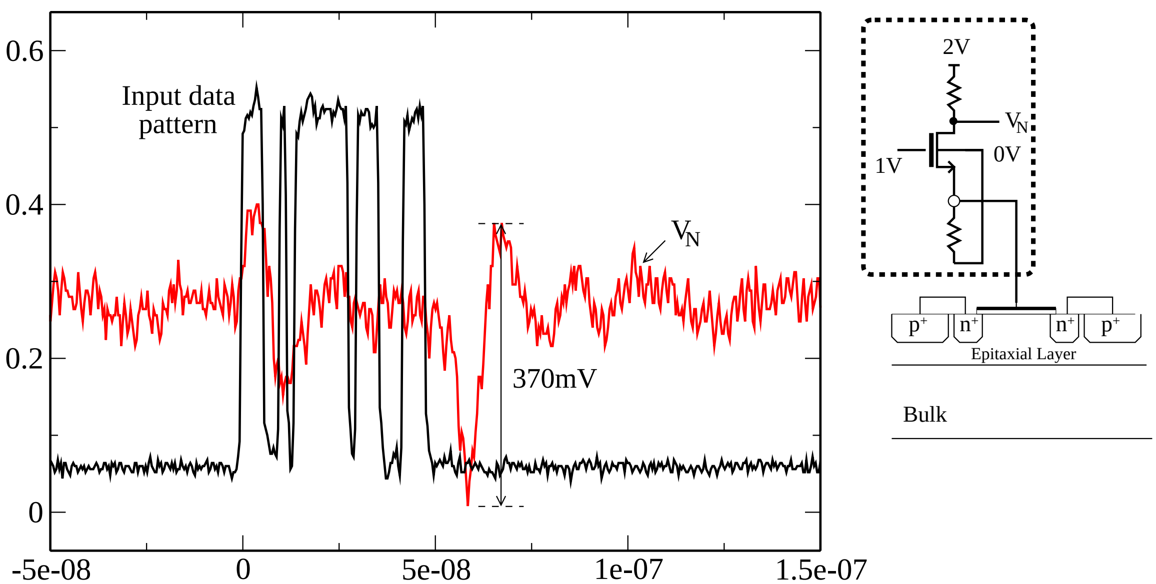


Figure 5. 33. Substrate noise due to adaptive bus operation of 16 bus lines.

5.5.8. Power Dissipation Comparisons

For comparison, the overall power dissipation of the bus in voltage, current and adaptive mode of operation for varying data rates and data activity profiles is shown in Figure 5.34. The data activity profiles are representative of MPU instruction and data address streams. These are generated by externally controlling the operation mode of four independent 8-bit pattern generators - in hold-mode, it retains the previous data pattern and in active-mode, it generates pseudo-random data sequences.

As shown in Figure 5.34, the ABB exhibits 50%-60% reduction in power over the CM bus and 66% improvement in data rate over the VM bus. Although the VM bus is shown to dissipate less power at 600Mb/s/line, at least 2-3 additional repeaters per line (i.e. 32-48 additional repeaters for 16 lines) are required to meet 1Gb/s/line causing an increase in overall power dissipation. At 16Gb/s, the ABB dissipates 93mW@2.5V, or equivalently energy consumption of 5.81pJ/bit for 1.75cm lines.

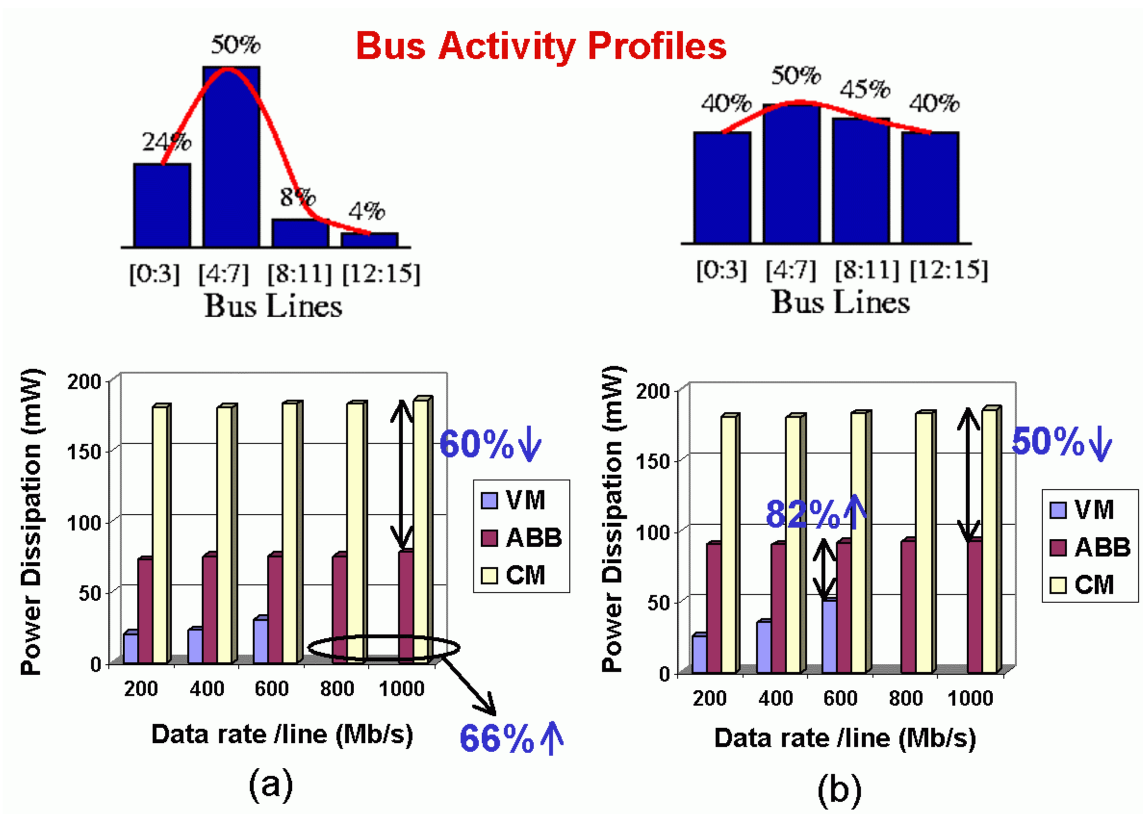


Figure 5. 34. Measured power dissipation versus data-rate and data activity profile (a) instruction (b) data address streams.

5.6. Summary

An adaptive bandwidth bus (ABB) architecture based on hybrid current/voltage mode repeaters for signaling across long global interconnects has been explored to minimize the static power dissipation inherent to current-mode sensing while achieving high-speed data transmission. This technique allows for minimal repeater insertion while maximizing the operating bandwidth. The attractiveness of current-mode signaling stems from the fact that relatively high data-rates can be attained despite the continuing reverse interconnect scaling trends. To compensate for the increase in static power dissipation inherent to current sensing, the proposed bus technique adaptively changes the mode of operation from current to voltage when the signal activity is low and from voltage to current mode otherwise. Thus, the bus energy expenditure can be minimized to support the required bus signal activity only.

A preliminary analog ABB fabricated in AMI 1.6μm bulk CMOS technology has been implemented to evaluate key aspects of the adaptive bus. Measurement results indicate a 62% improvement in static power dissipation over CM sensing techniques while achieving 40% increase in maximum data rate over VM signaling. A comparison of the ABB and VM signaling schemes operating at Vdd=3V and Vdd=5V, respectively, shows that the ABB dissipates nearly 2.5x less power with only 4.5% penalty in maximum data rate.

A digital ABB for pipelined systems eliminates the data-dependent delay associated with the analog architecture. Results from realistic MPU data streams extracted from an Alpha 21264 time-based microprocessor simulator show the potential benefits of the ABB. In addition, the power dissipation dependence on data activity profiles and design variables such as number of control lines and pipeline depth are also addressed.

An experimental ABB test chip fabricated in TSMC 0.35μm CMOS technology attains 16Gb/s of aggregate NRZ bandwidth with 16 bus lines operating at a maximum of 1Gb/s with an eye-width opening of 640ps across lossy on-chip interconnects spanning 1.75cm in length. Results indicate a 35%-70% reduction in delay and 66% improvement in maximum data rate over a voltage-mode bus, and 50%-60% reduction in power dissipation over the current-mode bus. The energy consumption of the bus is 5.81pJ/bit for 1.75cm lines.

5.7. References

[1] H.B. Bakoglu, Circuits, Interconnections, and Packaging for VLSI. Reading, MA: Addison-Wesley, 1990.

[2] P.L. Edefors, “Investigation on maximal throughput of a CMOS repeater chain,” IEEE Trans. Cir. Systems I, vol. 47, no. 4, pp. 602-606, April 2000.

[3] E. Seevinck, P. van Beers, 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.

[4] S.V. Morton, “Techniques for driving interconnect,” Design of High Performance Microprocessor Circuits, IEEE Press, pp. 352-376, 2001.

[5] Semiconductor Industry Association, *The International Technology Roadmap for Semiconductors*, 2001 Edition.

[6] S.E. Schuster and P.W. Cook, “Low-power synchronous-to-asynchronous- to-synchronous interlocked pipelined CMOS circuits operating at 3.3-4.5 GHz,” IEEE Journal of Solid-State Circuits, vol. 38, no. 4, pp. 622-630, April 2003.

[7] 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.

[8] M. Badaroglu, M. van Heijningen, V. Gravot, J. Compiet, S. Donnay, G.E. Gielen, H.J. De Man, “Methodology and Experimental Verification for Substrate Noise Reduction in CMOS Mixed-Signal ICs With Synchronous Digital Circuits,” IEEE Journal of Solid-State Circuits, vol. 37, no. 11, pp. 1383-1395, Nov. 2002.

[9] M. Heijningen, J. Compiet, P. Wambacq, S. Donnay, M.G. Engels, I. Bolsens “Analysis and Experimental Verification of Digital Substrate Noise Generation for Epi-Type Substrate,” IEEE Journal of Solid State Circuits, vol. 35, no. 7, 1002-1008 July 2000.

[10] Y. Tsividis, *Operation and Modeling of the MOS transistor*, Second Edition, McGraw-Hill, 1999.

[11] D.K. Su, M.J. Loinaz, S. Masui, B. Wooley, “Experimental Results and Modeling Techniques for Substrate Noise in Mixed-Signal Integrated Circuits,” IEEE Journal of Solid State Circuits, vol. 28, no. 4, pp. 420-430, April 1993.

[12] G. Gerosa, S. Gary, C. Dietz, P. Dac, K. Hoover, J. Alvarez, H. Sanchez, P. Ippolito, N. Tai, S. Litch, J. Eno, J. Golab, N. Vanderschaaf, and J. Kahle, “A 2.2W, 80MHz Superscalar RISC microprocessor,” IEEE Journal of Solid State Circuits, vol. 29, no. 12, pp. 1440-1452, Dec. 1994.

[13] V. Stojanovic, V.G. Oklobdzija, “Comparative Analysis of Master-Slave Latches and Flip-Flops for High-Performance and Low-Power Systems,” IEEE Journal of Solid State Circuits, vol. 34, no. 4, pp. 536-548, April 1999.

[14] F. Klass, “Semi-Dynamic Flip-Flops with Embedded Logic,” IEEE Symposium on VLSI Circuits, pp. 108-109, 1998.

[15] B.A. Chappell, T.I. Chappell, S.E. Schuster, H.M. Segmuller, J.W. Allan, R.L. Franch, P.J. Restle, “Fast CMOS ECL Receivers With 100-mV Worst-Case Sensitivity,” IEEE Journal of Solid State Circuits, vol. 23, no. 1, pp. 59-67, Feb. 1988.

[16] M. Shoji, “Elimination of Process-Dependent Skew in CMOS VLSI,” IEEE Journal of Solid State Circuits, vol. 21, no. 5, pp. 875-880, Oct. 1986.

[17] P. Larsson, “Resonance and Damping in CMOS Circuits with On-Chip Decoupling Capacitance,” IEEE Transactions on Circuits and Systems I, vol. 45, no. 8, pp. 849-858, Aug. 1998.

[18] K. Vichienchom, “A Multi-Gigabit CMOS Transceiver with 2x Oversampling Linear Phase Detector,” PhD. Dissertation, NCSU, 2003.

[19] Dally W.J., Poulton J.W., “Digital System Engineering,” Cambridge University Press, 1998.

Chapter VI

Conclusions

6.1. Summary and Contributions

This dissertation explores the use of *current-mode* and *band-limited* pulse signaling to minimize delay, reduce cross-talk noise and improve transmission bandwidth in deep sub-micron global on-chip interconnects. The contributions of this work can be summarized as follows:

1) The use of band-limited pulse signaling for on-chip data transmission as a means to reduce the source of cross-talk noise. Silicon implementation of a low overhead reduced edge-rate driver compatible with repeater insertion methodologies for full-chip system implementation. Theoretical formulations of cross-talk noise for capacitive and resistive terminations.
2) Closed-form delay and bandwidth formulations for current and voltage mode signaling techniques based on single-exponent approximation to distributed RC line system response. Power dissipation model for current and voltage mode receive-end sensing. Experimental validation of proposed model with silicon test chip.
3) Adaptive bandwidth bus based on hybrid current/voltage mode signaling used to maximize signaling bandwidth and reduce delay while minimizing the static power dissipation associated with current signal sensing. Silicon implementation of analog and digital adaptive bus architectures in AMI 1.6μm bulk CMOS and TSMC 0.35μm CMOS technologies, respectively.

4) Statistical analysis of realistic microprocessor data streams and power dissipation modeling techniques to estimate power associated with current, voltage and adaptive mode busses.

6.2. Future Work

The challenges associated with on-chip interconnects will continue to drive system architects and circuit designers to implement new interconnect concentric solutions to address the cross-talk noise, delay latency, bandwidth and power limitations. These limitations tend to be system application specific. However, as operating frequencies continue to increase, the noise and bandwidth limitation of on-chip interconnects are likely to determine the direction of future research in this area. Delay latency, to a large extent, will be unavoidable and must be taken into account during the early stages of architectural planning. Several possible directions of future research are outlined in the following sections.

6.2.1. Noise Aware Designs

The side-wall capacitive and signal transitions of an aggressor line can cause detection errors on a quiescent victim line and degrade the delay performance of adjacent wires. The continuing increase in wire aspect ratios to mitigate resistive effects only helps exacerbate this signal integrity problem. Currently, the cross-talk noise problem seems to be a critical obstacle in interconnect performance. Therefore, interconnect designs must take into account the coupling capacitance effects. Possible areas of research are:

1) *Differential interconnects*: signals transmitted over differential wires theoretically outperform single-ended approaches due to its inherent common-mode rejection. Improved rejection from adjacent differential or single ended wires can be improved by ‘twisting’ the wires. Low overhead singled-ended to differential and differential to single-ended circuits are required to minimize the power, latency and area. Low swing differential swing techniques can be used to reduce the additional wire related power dissipation.

2) *Transition encoded busses*: encoding signal transitions reduces the capacitive multiplier factor due to adjacent switching interconnects (i.e. the effective capacitance due to side-wall capacitance). Low overhead encoding and decoding circuits as well as dynamic pre-charge techniques can be investigated.

3) *Reduced edge rate interconnects*: full-chip implementation of repeaters with pulse shaping capabilities can further investigated to evaluate the overall impact in noise, power and delay performance.

6.2.2. Signaling Bandwidth

The signaling bandwidth of future interconnects are likely to become increasingly important. Typically, minimizing global delays has been of critical importance. However, as the average global interconnect length and the clock frequency continues to increase, achieving high-speed on-chip data transmission will become equally important. Therefore, techniques for maximizing data throughput for global interconnects within on-die processing units are required. Some possible areas of research are outlined:

1) *High-speed current-mode signaling*: although much of the work presented in this dissertation addresses current-mode signaling, the static power dissipation associated with this technique can be further reduced. Repeater stages with dynamic resistive terminations can be investigated to potentially eliminate the static current when the line is quiescent. Conditional evaluation of clocked current-mode receivers can be used to minimize the static power to a fraction of the clock period.

2) *Line equalization techniques*: In order to improve the ISI due to resistive losses, equalization techniques can be implemented to maximize the repeater insertion distance. Noise effects and circuit area overhead must be carefully addressed to ensure the feasibility of this technique.

6.2.3. Adaptive Interconnects

Adaptive, self-compensating circuits are often used to minimize local device and interconnection timing variations, leakage currents and overall power dissipation in large-scale systems. In order to achieve high computational efficiency, the compensating variable is usually a voltage, such as a supply voltage used to track operating frequency or body bias voltage used to minimize leakage currents. Other approaches may implement scheduling algorithms to make efficient use of functional units or maintain peak power dissipations within a specified level.

The research proposed as part of this dissertation is aimed to reduce power dissipation by adaptively changing the signaling bandwidth of a subsystem in order to track the computational load. Since signaling bandwidth is inversely proportional to power, adaptively changing the bandwidth reduces the system power dissipation. Possible areas of future research are new techniques to apply this concept to computational units. By identifying critical nodes in the data-path, it possible to implement circuit topologies that compensate for the speed limitations. Hybrid current and voltage mode techniques can be used to adaptively change the system time constants to allocate signaling bandwidth as necessary.

Appendix A

A.1. Effective Resistance and Capacitance Derivation

A lumped, linear, time-invariant circuit such as that of a generalized distributed RC line shown in Figure 3.3 (i.e. in chapter III), can be conveniently expressed in terms of state equations by using the modified nodal matrix (MNA) representation [1]. The generalized output equation can be expressed in the Laplace domain as:

$$[G + sC]\cdot[X(s)] = b(s) \tag{A-1}$$

where G and C are the nodal conductance and capacitance matrix, respectively; $X(s)$ is the vector of node voltages; and $b(s)$ is the input source excitation. An NxN nodal conductance matrix G for the circuit topology shown in Figure 3.3b can be written as:

$$[G] = \begin{bmatrix} G_S + G_u & -G_u & 0 & \cdots & & 0 \\ -G_u & 2G_u & -G_u & 0 & & \\ 0 & -G_u & 2G_u & -G_u & 0 & \\ \vdots & & & \cdots & & \vdots \\ & 0 & -G_u & 2G_u & -G_u & 0 \\ & & 0 & -G_u & 2G_u & -G_u \\ 0 & & \cdots & 0 & -G_u & G_L + G_u \end{bmatrix} \tag{A-2}$$

G_u is the segment conductance of the distributed transmission line and G_L is the load conductance. In Eq. (A-2), $G_S=1/(R_{S1}+R_u)$; where R_{S1} is the source resistance.

As described in [2]-[3], the node voltage vector $X(s)$ is expanded using the Taylor series and coefficients of similar powers of s are equated to obtain the following expressions:

$$\begin{aligned} GM_0 &= b \\ GM_q &= -CM_{q-1} \quad q > 0 \end{aligned} \tag{A-3}$$

where M_q represents the moment vector. A general closed-form expression for the q^{th} moment and the k^{th} node voltage of $X(s)$ is given by:

$$m_{q+1}^{k} = \sum_{j=1}^{N} -G_{inv}(k,j)\cdot C(j)\cdot m_q^{j} \tag{A-4}$$

where N is the number of distributed segments and G_{inv} is the inverse matrix (G-1) which can be expressed as:

$$G_{inv}(k,j)=\begin{cases}\dfrac{[G_u+(N-k)\cdot G_L]\cdot[G_u+(j-1)\cdot G_S]}{G_u\cdot[G_uG_S+(N-1)\cdot G_LG_S+G_LG_u]} & j\le k\\ \dfrac{[G_u+(N-j)\cdot G_L]\cdot[G_u+(k-1)\cdot G_S]}{G_u\cdot[G_uG_S+(N-1)\cdot G_LG_S+G_LG_u]} & j>k\end{cases} \tag{A-5}$$

Similarly, we can express the 0th moment for the k^{th} nodal voltage as:

$$m_0^k=\frac{G_S[G_u+(N-k)\cdot G_L]}{[G_uG_S+(N-1)\cdot G_LG_S+G_LG_u]} \tag{A-6}$$

Given equations (A-4), (A-5) and (A-6), all higher order moments can be derived.

To obtain the effective resistance and capacitance of the first order AWE approximation, we express the reduced single order rational transfer function $\hat{H}(s)$ in terms of the polynomial coefficients as:

$$\hat{H}(s)=\frac{\hat{k}}{s+\hat{p}}=\frac{\frac{a_0}{b_1}}{s+\frac{1}{b_1}} \tag{A-7}$$

where a_0 and b_1 are calculated from the 0th and 1th moments from (A-4) and (A-6), given as:

$$a_0=m_0 \qquad b_1=-\frac{m_1}{m_0} \tag{A-8}$$

Since the pole of $\hat{H}(s)$ is $1/(R_{eff}C_{eff})$, the effective resistance and capacitance can be derived from Eq. (A-8), which can be written in closed-form as:

$$R_{eff}=\eta_L\frac{R_T}{\sqrt{2}}\left(\frac{N+1}{N}\right)\cdot\left(1+\frac{R_T}{3R_L}\left(\frac{N-1}{N}\right)\right)+2\eta_L\frac{R_S}{\sqrt{2}}\left(1+\frac{R_T}{2R_L}\left(\frac{N-1}{N}\right)\right) \tag{A-9a}$$

$$C_{eff}=\frac{C_T}{\sqrt{2}}+C_S\eta_S\frac{R_L+R_T}{R_{eff}}+C_L\eta_L\frac{R_S+R_T}{R_{eff}} \tag{A-9b}$$

and

$$\eta_L=\frac{R_L}{R_L+R_S+R_T} \tag{A-9c}$$

$$\eta_S=\frac{R_S}{R_L+R_S+R_T} \tag{A-9d}$$

A.2 References

[1] C.W. Ho, A.E. Ruehli, P.A. Brennan, “The modified nodal approach to network analysis,” IEEE Trans. Circuits and Systems, vol. CAS-22, pp. 504-509, June 1975.

[2] L.T. Pillage, R.A. Rohrer, “Asymptotic waveform evaluation for timing analysis,” IEEE Trans.Computer Aided Design, vol. 9, no. 4, pp. 352-366, April 1990.

[3] R. Achar, M.S. Nakhla, “Simulation of high-speed interconnects,” Proc. of the IEEE, vol. 89, no. 5, pp. 693-728, May 2001.