UC San Diego
UC San Diego Electronic Theses and Dissertations

Title
Design and optimization of high-performance low-power CMOS VLSI interconnects

Permalink
https://escholarship.org/uc/item/6wm241cm

Author
Zhang, Yulei

Publication Date
2011

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Design and Optimization of High-Performance Low-Power CMOS VLSI Interconnects

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in
Electrical Engineering (Electronic Circuits & Systems)

by

Yulei Zhang

Committee in charge:

Professor James Buckwalter, Chair
Professor Chung-Kuan Cheng
Professor Bill Lin
Professor Yuan Taur
Professor Michael Taylor

2011

Copyright
Yulei Zhang, 2011
All rights reserved.

The dissertation of Yulei Zhang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____

____

____

Chair

University of California, San Diego

2011

Dedication

To my parents and my lovely fiancee.

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Table of Contents v

List of Figures vii

List of Tables x

Acknowledgements xi

Vita xiii

Abstract of the Dissertation xv

I Introduction 1
1. Roadmap on VLSI interconnects 1
2. Current research efforts 5
3. Dissertation outline 7

II Background of On-Chip Transmission-Line 9
1. Basic theory 9
2. Modeling and simulation 11
3. Worst-case eye-diagram prediction 15

III Comparison of High-Performance On-Chip Global Interconnections . . 19
1. On-chip global interconnects 19
A. Overview 19
B. Interconnect schemes 21
C. Global wire modeling 24
D. Performance analysis 29
2. Design methodology 36
A. Single-ended T-lines 36
B. Differential T-lines 38
3. Performance prediction and comparison 39
A. Experimental settings 39
B. Latency 40
C. Other metrics 41
D. Critical length 43
4. Signal integrity 45
A. Single-ended T-lines 45

B. Differential T-lines 46
5. Summary and discussion 48
A. Discussion 48
B. Summary 49

IV Pipelined Global Interconnects with Voltage Scaling 50
1. Pipelined global interconnects 50
A. Overview 51
B. Glossary 52
C. Assumptions and modeling 54
2. Design objectives and metrics 59
A. Design objectives 59
B. Performance metrics 60
3. Performance evaluation flow 61
4. Experimental results 63
A. Experimental settings 63
B. Pipelining effect 64
C. Voltage scaling effect 66
D. Technology scaling 68
E. Design example 70
5. Summary 72

V Energy-Efficient Equalized Global Interconnects 74
1. Equalized on-chip global link 74
2. Driver design for on-chip transmission-line 76
A. Design guideline for tapered CML chain 77
B. Driver design example 78
3. Continuous-time linear equalizer design 81
A. CTLE modeling 81
B. CTLE design example 83
4. Equalized global link analysis 84
A. CTLE eye-opening analysis 84
B. System-level analysis 87
5. Driver-receiver co-design for low energy-per-bit 90
A. Driver-receiver co-optimization flow 93
B. Design space exploration 95
C. Full global link performance sign-off 98
6. Summary 102

VI Conclusion 104
1. Summary of contributions 104
2. Future works 105

A Performance analysis of ideal pipelined repeated RC wires 106

Bibliography 108

LIST OF FIGURES

Figure I.1: Cross-section of hierarchical scaling for MPU devices [60]. 2
Figure I.2: Scaling trends of interconnect and gate delay [80]. 3
Figure I.3: Scaling trends of interconnect and gate energy [80]. 4

Figure II.1: 3D parameter extraction model of on-chip T-lines considering adjacent and sub-adjacent layers. 11
Figure II.2: 2D inductance extraction model of on-chip T-lines considering sub-adjacent layers. 12
Figure II.3: One segment of three-line network using compact circuit model introduced in [17]. 13
Figure II.4: A n-pole foster RL filter representation of frequency-dependent impedance. 14
Figure II.5: A typical eye-diagram observed at the end of on-chip T-line. 16

Figure III.1: The organization of on-chip global interconnect structures. 20
Figure III.2: The multi-dimensional design tradeoffs of different global interconnect structures. 21
Figure III.3: One stage of pipelined repeated RC wire (P-RC structure). 22
Figure III.4: Single-ended T-line schemes for on-chip global interconnect. 23
Figure III.5: Differential T-line schemes for on-chip global interconnect. 23
Figure III.6: The wire configurations for on-chip T-line schemes. 26
Figure III.7: The design framework for on-chip global T-line structures (UT-TL/T-TL/UE-TL/PE-TL). 37
Figure III.8: The normalized delay of different global interconnection structures under min-d objective. 40
Figure III.9: The normalized energy per bit of different global interconnection structures under min-d objective. 41
Figure III.10: The throughput density of different global interconnection structures under min-d objective. 42
Figure III.11: Chip areas consumed by different global interconnection structures under min-d objective. 42
Figure III.12: Critical length of several chosen interconnect structure pairs in terms of different performance metrics under min-d objective. 44
Figure III.13: The wire configurations and worst case switching patterns of T-line structures for testing crosstalk effects. (a) single-ended; (b) differential. 45
Figure III.14: The influence of crosstalk effects on the eye-height of UE-TL and PE-TL structures. 47

Figure IV.1: Structure of pipelined global interconnect studied in this work. 51

Figure IV.2: Voltage scaled models built by using HSPICE simulation and curve regression. . . . 57
Figure IV.3: Impact of the number of pipelining stages on the performance of pipelined global interconnects using 45 nm CMOS process under different design objectives. . . . 65
Figure IV.4: Impact of supply voltage scaling on the performance of pipelined global interconnects using 45 nm CMOS process under different design objectives. . . . 67
Figure IV.5: Impact of technology scaling on the performance of pipelined global interconnects under different design objectives. . . . 69

Figure V.1: The overall structure of equalized on-chip global link studied in this work. . . . 75
Figure V.2: The CML buffer schematic and DC transfer characteristic. 77
Figure V.3: The cross-section of differential on-chip T-line. . . . 79
Figure V.4: The schematic (a) [23] and equivalent small-signal circuit (b) for Continuous-Time Linear Equalizer (CTLE). . . . 81
Figure V.5: The predicted and simulated eye-opening vs. CTLE power consumption for different modeling approaches. . . . 83
Figure V.6: The CTLE eye-opening for different p_2/p_1 ratios. The following parameters are assumed to generate the figure: α=1.2, T_C=50 ps, τ_I=250 ps, V_{DD}=1 V, v_{od}=100 mV, G_I=0.3 V. . . . 86
Figure V.7: The global link model used for system-level analysis. . . . 87
Figure V.8: 3D map and 2D contour of global link eye-opening (after CTLE) based on derived analytical model. This figure is generated using following parameter values: T_C=50 ps, V_{DD}=1 V, C_L=5 fF, R_L=1 kΩ, V_{sw}=300 mV, R_w=150 Ω, C_w=1 pF, α=1.2, v_{od}=100 mV. . . . 89
Figure V.9: Simulated eye-diagrams at different locations of proposed equalized on-chip global link using conventional design methodology. The design parameters: R_S=47 Ω, R_T=94 Ω, R_L=440 Ω, R_D=110 Ω, C_D=680 fF, V_{od}=60 mV. Simulated power consumption (driver+receiver w/o SA-latch) is 8.1 mW. . . . 91
Figure V.10: Simulated eye-diagrams at different locations of proposed equalized on-chip global link using low-power design methodology.The design parameters: R_S=148 Ω, R_T=1100 Ω, R_L=890 Ω, R_D=1430 Ω, C_D=150 fF, V_{od}=58 mV.Simulated power consumption (driver+receiver w/o SA-latch) is 3.8 mW. . . . 92
Figure V.11: The driver-receiver co-optimization flow. . . . 93
Figure V.12: 3D map of power dissipation and figure of merit (FoM) for equalized global link in the explored design space. Solid lines indicate the single-supply design, whereas dash lines indicate the split-supply design. . . . 96

Figure V.13: The schematic of Sense-Amplifier Based Latch (SA-latch) used in this work. 99
Figure V.14: Histograms of global link power and delay distribution under process variations for single-supply and split-supply design based on 500-run Monte Carlo simulations in HSPICE. 101

LIST OF TABLES

Table I.1: Delay and power scaling data of global wires [60]. 2

Table III.1: Design parameters for global *R-RC* and *P-RC* wires based on ITRS Roadmap 2007 and SPICE simulation 25
Table III.2: Parameters of on-chip global T-line used in UT-TL/T-TL schemes 27
Table III.3: Design parameters for UE-TL/PE-TL schemes (Wire Length = 5 mm) 27
Table III.4: Modeling performance metrics (normalized delay, normalized energy, normalized throughput, area) of six global interconnection structures using technology-defined parameters. 30
Table III.5: Crosstalk effects on the T-TL structure 46
Table III.6: Maximum crosstalk peak noise (mV) of differential T-line structures 46

Table IV.1: Design parameters for global pipelined interconnect based on ITRS Roadmap 2008 and predictive SPICE models. 52
Table IV.2: Symbols used for variables and parameters of pipelined global interconnects. 53
Table IV.3: Performance comparison of Nominal V_{dd} Design (Min-Latency) and Voltage Scaling Design (Max-TPEA) using 45 nm CMOS process. 71

Table V.1: Impact of T-line width/spacing on the received eye quality for a given I_{SS}=6 mA. Width + Spacing=2.0 μm 80
Table V.2: Impact of T-line pitch on the received eye quality for a given I_{SS}=6 mA. Width=Spacing=1/2*pitch 80
Table V.3: A 20 Gbps CML driver design example for 10 mm on-chip T-line. 81
Table V.4: A 20 Gbps CTLE design example for 10 mm on-chip T-line. 84
Table V.5: Performance sign-off and comparison of on-chip equalized global link for single-supply and split-supply design. 99

Table A.1: Design parameters used in performance analysis of *P-RC* structure 106

ACKNOWLEDGEMENTS

When four years ago I came to UCSD to pursue my PhD degree, I did not imagine that I would choose VLSI CAD as my research topics and also link my future career with this interesting and challenging field. Lots of help, support, and encouragement from all sides make this PhD dissertation come true in such a short four-year period. I would like to express my sincere gratitude here to all the people who make this period wonderful.

First of all, I want to thank my PhD advisor, Professor Chung-Kuan Cheng, who introduced me to the VLSI CAD field and inspired me to perform such exciting researches. During the days working with him, I do learn a lot from his wisdom, hard-working, patience, and passion for pursuing the truth. He also influenced me with his outstanding personality, by guiding me on how to communicate with other researchers and how to broaden my academic connections. All the things I learned from him become invaluable wealth for my entire life.

Secondly, Professor James Buckwalter, Professor Yuan Taur, Professor Bill Lin, and Professor Michael Taylor, who serve as my PhD committee members, provided a lot of insightful suggestions and discussions regarding my research work and PhD dissertation. It is a pleasure to thank all of them. I owe my special thanks to Professor James Buckwalter, who also served as my co-advisor. His deep knowledge and experience on circuit design help me to improve my research work in more practical way and to develop my designer's perspective.

Furthermore, I am grateful to thank every graduate student in VLSI CAD lab. He Peng, Rui Shi, Yi Zhu, Wanping Zhang, Renshen Wang, Xiang Hu, Peng Du, Shih-Hung Weng, Amirali Shayan, and many other students or post-doc researchers, shared the interesting research ideas and collaborated with me on different projects. It is joyful to work with all you guys! Special thanks go to Ling Zhang, who provided so many valuable suggestions on both my research works and my daily life.

Finally, I owe my deepest gratitude to my family. Without their everlasting support and understanding, I cannot finish this dissertation. My parents always encourage me to do what I choose since my childhood and give me support wherever and whenever I meet difficulties. My lovely fiancee Yunfei took care of me patiently with bearing lots of my complaint during my dissertation composition and gave me the confidence when I hesitate or feel frustrated. This dissertation is dedicated to them.

Chapter III includes the content of one accepted journal paper, "Prediction and Comparison of High-Performance On-Chip Global Interconnection", by Y. Zhang, X. Hu, A. Deutsch, A. E. Engin, J. F. Buckwalter, C. K. Cheng, which will appear in *IEEE Transaction on VLSI Systems*. The dissertation author was the primary investigator and author of the paper.

Chapter IV includes the content of one published conference paper, "Performance Prediction of Throughput-Centric Pipelined Global Interconnects with Voltage Scaling", by Y. Zhang, J. F. Buckwalter, C. K. Cheng, in Proceedings of *2010 IEEE International Workshop on System Level Interconnect Prediction*. The dissertation author was the primary investigator and author of the paper.

Chapter V includes the content of one submitted journal paper, "Energy Efficiency Optimization through Co-Design of the Transmitter and Receiver in High-Speed On-Chip Interconnects", by Y. Zhang, J. F. Buckwalter, C. K. Cheng, which is submitted to *IEEE Transaction on VLSI Systems*. The dissertation author was the primary investigator and author of the paper.

VITA

2007	B.Eng. in Electronic Engineering Tsinghua University, Beijing, P.R.China
2009	M.S. in Electrical Engineering (Electronic Circuits & Systems) University of California, San Diego
2011	Ph.D. in Electrical Engineering (Electronic Circuits & Systems) University of California, San Diego

PUBLICATIONS

Y. Zhang, J. F. Buckwalter, C. K. Cheng, "Energy Efficiency Optimization through Co-Design of the Transmitter and Receiver in High-Speed On-Chip Interconnects", *IEEE Trans. on VLSI Systems*, submitted.

Y. Zhang, X. Hu, A. Deutsch, A. E. Engin, J. F. Buckwalter, C. K. Cheng, "Prediction and Comparison of High-Performance On-Chip Global Interconnection", *IEEE Trans. on VLSI Systems*, accepted.

Y. Zhang, J. F. Buckwalter, C. K. Cheng, "High-Speed Low-Power On-Chip Global Link Design using Continuous-Time Linear Equalizer", *Proceedings of IEEE Electrical Performance of Electronic Packaging and Systems (EPEPS 2010)*

Y. Zhang, J. F. Buckwalter, C. K. Cheng, "Performance Prediction of Throughput-Centric Pipelined Global Interconnects with Voltage Scaling", *Proceedings of IEEE Int. Workshop on System Level Interconnect Prediction (SLIP 2010)*

Y. Zhang, J. F. Buckwalter, C. K. Cheng, "On-Chip Global Clock Distribution using Directional Rotary Traveling-Wave Oscillator", *Proceedings of IEEE Electrical Performance of Electronic Packaging and Systems (EPEPS 2009)*

Y. Zhang, X. Hu, A. Deutsch, A. E. Engin, J. F. Buckwalter, C. K. Cheng, "Prediction of High-Performance On-Chip Global Interconnection", *Proceedings of IEEE Int. Workshop on System Level Interconnect Prediction (SLIP 2009)*

Y. Zhang, L. Zhang, A. Deutsch, G. A. Katopis, D. M. Dreps, E. S. Kuh, C. K. Cheng, "Design Methodology of High Performance On-Chip Global Interconnect Using Terminated Transmission-Line", *Proceedings of IEEE Int. Symp. Quality of Electronic Design (ISQED 2009)*

Y. Zhang, L. Zhang, A. Deutsch, G. A. Katopis, D. M. Dreps, J. F. Buckwalter, E. S. Kuh, C. K. Cheng, "On-Chip Bus Signaling Using Passive Compensation", *Proceedings of IEEE Electrical Performance of Electronic Packaging and Systems (EPEPS 2008)*

Y. Zhang, L. Zhang, A. Tsuchiya, M. Hashimoto, C. K. Cheng, "On-chip High Performance Signaling Using Passive Compensation", *Proceedings of IEEE International Conference of Computer Design (ICCD 2008)*

L. Zhang, Y. Zhang, H. Cheng, B. Yao, K. Hamilton, C. K. Cheng, "On-Chip Interconnect Analysis of Performance and Energy Metrics under Different Design Goals", *IEEE Trans. on VLSI Systems*, March, 2011.

L. Zhang, W. Yu, Y. Zhang, R. Wang, A. Deutsch, G. A. Katopis, D. M. Dreps, J. F. Buckwalter, E. S. Kuh, C. K. Cheng, "Analysis and Optimization of Low Power Passive Equalizers for CPU-Memory Links", *IEEE Trans on Advance Packaging*, accepted.

R. Wang, Y. Zhang, N. C. Chou, E.F.Y. Young, C. K. Cheng, R. Graham, "Bus Matrix Synthesis based on Steiner Graphs for Power Efficient System-on-Chip Communications", *IEEE Trans. on CAD*, Feb., 2011.

L. Zhang, Y. Zhang, A. Tsuchiya, M. Hashimoto, E. S. Kuh, C. K. Cheng, "High Performance On-Chip Differential Signaling Using Passive Compensation for Global Communication", *Proceedings of IEEE Asian and South Pacific Design Automation Conference (ASP-DAC 2009)*

L. Zhang, W. Yu, Y. Zhang, R. Wang, A. Deutsch, G. A. Katopis, D. M. Dreps, J. F. Buckwalter, E. S. Kuh, C. K. Cheng, "Low Power Passive Equalization Design for Computer Memory Links", *Proceedings of IEEE Symposium on High-Performance Interconnects (HOTI 2008)*

X. Hu, W. Zhao, P. Du, Y. Zhang, A. Shayan, C. Pan, A. E. Engin, C. K. Cheng, "On the Bound of Time-Domain Power Supply Noise Based of Frequency-Domain Target Impedance", *Proceedings of IEEE Int. Workshop on System Level Interconnect Prediction (SLIP 2009)*

FIELDS OF STUDY

Major Field: Electrical Engineering (Electronic Circuits & Systems)
Studies in VLSI CAD
Professor Chung-Kuan Cheng

ABSTRACT OF THE DISSERTATION

Design and Optimization of High-Performance Low-Power
CMOS VLSI Interconnects

by

Yulei Zhang

Doctor of Philosophy in Electrical Engineering (Electronic Circuits & Systems)

University of California, San Diego, 2011

Professor James Buckwalter, Chair

As semiconductor technology advances in the ultra deep sub-micron era, on-chip global interconnections have been an ever-greater barrier to achieving high-performance and low-power for the increasingly larger system-on-chip (SoC) designs. Various on-chip interconnection schemes are proposed to tackle the scaling issue of global wires by manipulating the wire operation regions, changing signaling methods, and applying different equalization techniques. Optimization frameworks are also proposed to aid the transmitter-wire-receiver co-design based on user-defined constraints.

For the six representative global interconnection schemes, we investigate their performance metrics with technology scaling by performing optimizations using the proposed SQP-based framework. A set of simple models is also developed to enable early-stage system-level analysis. Performance of different interconnection schemes are predicted and compared over several technology nodes.

We further perform studies on the pipelined RC interconnection by exploring its performance metrics with voltage and technology scaling based on different design objectives. A performance evaluation flow is developed to generate the optimal designs for given objectives. Also, impacts of pipelining depth, voltage and technology scaling are illustrated.

Finally, we propose an energy-efficient high-speed on-chip global interconnection by employing continuous-time active equalization. Modeling and design of transmitter and receiver circuits are discussed. Analytical formula of received eye-opening is derived for system-level design planning. We further perform transmitter-receiver co-design through an optimization framework and explore the design space to generate design based on best energy-throughput tradeoff.

I

Introduction

I.1 Roadmap on VLSI interconnects

Interconnects on the VLSI chips are used to distribute the clock and other signals to the functional blocks of whole system as well as provide the power supply and ground connections. Modern logic chips (MPUs and ASICs) adopt a hierarchical scaling methodology to build the on-chip interconnect structure, as show in Fig.I.1.

With the technology scaling, the dimensions of local wires (Metal 1 and intermediate layers), also scale down, resulting in the increasing RC product[1]. But, considering the scaling length of local wires, the RC delay for such wires actually shows only a moderate increase. On the other hand, global interconnects do not enjoy the benefit of reduced wire length. Typically, the interconnect length in global wiring levels is fixed, determined by the chip size. As a result, the increasing RC delay for global interconnects becomes a critical issue for high performance VLSI chips. Interconnects also consume more power as the technology scaling, and gradually become dominant in the total power consumption, due

[1]More precisely, the resistance per unit length is increasing quadratically and capacitance per unit length is assumed to be constant if not considering the changes of resistivity and dielectric constant.

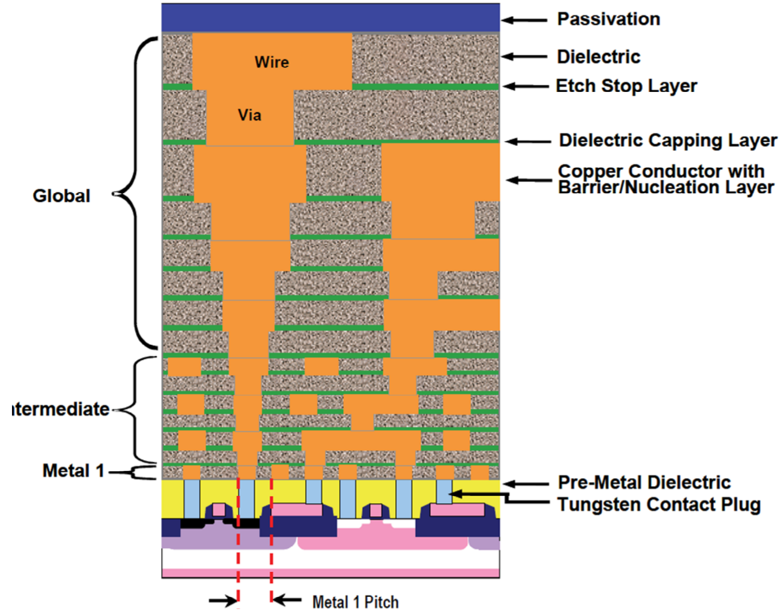


Figure I.1: Cross-section of hierarchical scaling for MPU devices [60].

Table I.1: Delay and power scaling data of global wires [60].

Year of Production	2010	2011	2012	2013	2014
Conductor effective resistivity ($\mu\Omega$-cm)	3.10	3.22	3.34	3.52	3.73
Capacitance per unit length (pF/cm)	2.0	2.0	2.0	2.0	2.0
RC delay (ps) for 1 mm global wire	542	713	846	1129	1562
Power index (W/GHz-cm^2)	1.8	2.0	1.8	2.0	2.3

to the increasing wire coupling capacitance and chip opearting frequency (supply voltage drops slowly in the prediction). To take a close look at the scaling trend of global interconnects from a quantitative perspective, we list some scaling data based on ITRS roadmap [60] in Table I.1.

It is noted that shrinking of wire width and thickness results in larger effective resistivity due to the scattering effect. Even with the improvement of dielectric constant, an increasing per-unit-length RC delay is observed as technology scales shown in the forth row of Table I.1. In the year 2014, the 1 mm wire delay could be 1562 ps, which is about 3 times of the number in the year 2010. Considering the logic gate delay, the 10 level fanout of 4 (FO4) delay of minimum

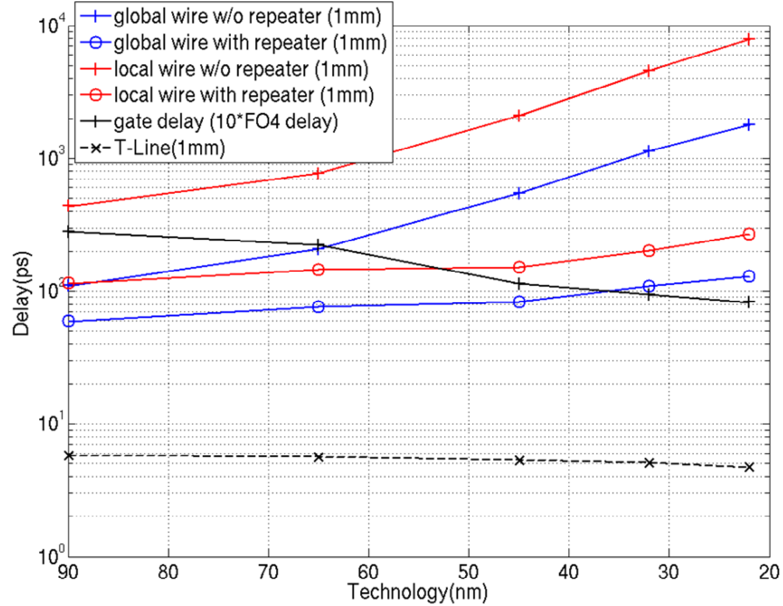


Figure I.2: Scaling trends of interconnect and gate delay [80].

sized inverter is only 145 ps at the 45 nm node (year 2010), whereas the RC delay of 1 mm-long, minimum pitch global wire is 542 ps at the same node. Given the fact that length of global wires keeps constant, the global wire delay will dominate and directly determine the system performance as technology advances. The delay scaling trends of interconnects and logic gates are also illustrated in Fig.I.2. It is seen that performance gap between global wire and logic gates will continue to increase as technology scales even with repeater inserted. To push the system for higher operating frequency, the low-latency global interconnects and corresponding design strategies remain a challenge in the near future.

In terms of power consumption, global wires also make a considerable contribution to the total dynamic power. Based on the power index numbers shown in Table I.1, the power index will reach up to 2.3 W/GHz-cm^2 in the year 2014. Suppose chip local clock frequency will be 7.5 GHz in 2014 (whereas the value is about 5 GHz nowadays), and chip area will remain 1 cm^2, then the power consumption of global wires could be 17.25 W, which is about two times of the value nowadays. Considering the high performance microprocessor today, which

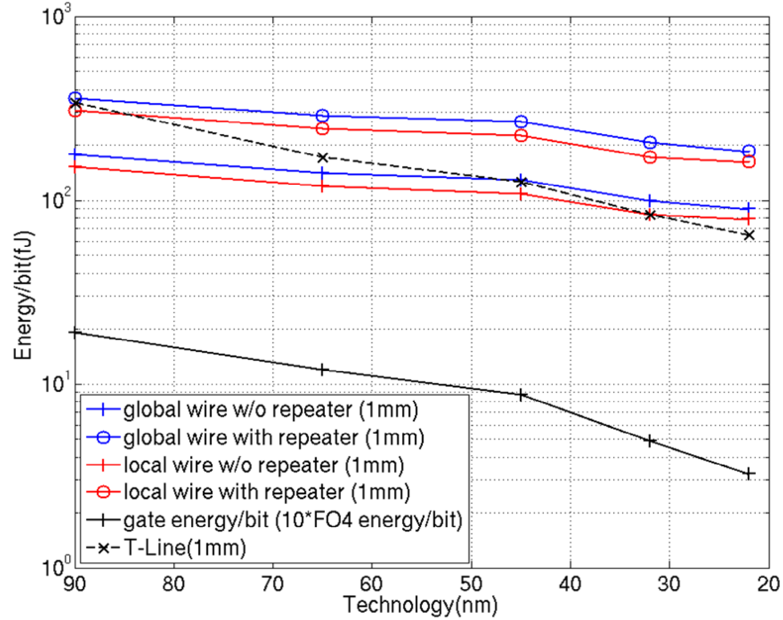


Figure I.3: Scaling trends of interconnect and gate energy [80].

consumes about 100 W, it is clear to see that global wires will become an important factor to determine the total power consumption. Some recent research works [47] [67] [76] also reveal the same observations. Fig.I.3 shows the predicted scaling trends of energy consumption for both interconnects and gate [80]. There exists about one order of magnitude difference between energy consumed by global wires and logic gates. In [47], it is shown that interconnect power accounted for half the total dynamic power in a 0.13 μm microprocessor designed for power efficiency. Therefore, power reduction on the global interconnects should become a key issue in the following generations.

In summary, VLSI interconnects, especially the global ones, have formed an invisible "interconnect wall" that blocks chip designers from achieving the performance requirements of ever increasingly larger system-on-chip (SoC) designs.

I.2 Current research efforts

To break the “interconnect wall” caused by the scaling of global wires, many approaches have been proposed to hasten on-chip global communication from different perspectives. Repeater insertion reduces the delay of global wire but brings in more power overhead. Transmission-line effects of global wires are utilized to achieve low-power signaling with consuming more chip area. To further improve interconnect throughput, either pipelining or equalization techniques can be applied based on different wire types and signaling methods.

The repeater insertion method introduced in [4] is firstly proposed to reduce the delay of global wires by breaking the long wire into segments and adding buffers. The repeater insertion method reduces total wire delay at the cost of additional power overhead. To reduce the power overhead, some works [6] [12] focus on power optimization of repeater insertion, with satisfying certain delay or bandwidth constraints. Furthermore, accurate global wire modeling approaches, such as considering inductance effects [32] [5] [72], have been taken into account when deriving the closed-form expressions [54] for various repeater insertion criteria. Tons of repeater insertion algorithms [1] [14] [45] are also proposed to foster the quick insertion of buffer chain in chip implementation with the minimum additional area or power cost. Noise, thermal, and other possible reliability issues during buffer insertion are also considered in some recent works [41] [11]. To better understand the trade-offs among different design metrics of repeated RC interconnect, [80] performs a set of numerical experiments to optimize buffer insertion with wire re-sizing for different metrics, and also summarizes the impacts of technology scaling on those metrics. In [50], buffer insertion methodology is described and reviewed from RTL to physical design to demonstrate how it is considered within the design flow of VLSI products.

Most conventional repeater insertion works focus on delay or power optimization, which cannot satisfy the stringent throughput requirement from in-

creasing capacity demand in emerging parallel computing architectures. As a s result, **throughput-centric** interconnect designs [61], like Networks-on-Chip (NoC) [34], have become necessary and hold more research attention. As the basic building block of NoCs, longer wires are pipelined through flip-flop insertion to meet the required clock period. Concurrent flip-flop and repeater insertion methods [46] [13] [44] are proposed to optimize the placement of flip-flops and repeaters to satisfy both latency and throughput requirements. [28] and [27] perform a power analysis on on-chip network and point out that deeper pipelining not only improves the overall throughput but also saves power by providing additional timing slack. As another pipelining method, wave-pipelined interconnects have been proposed [74] [81] [75] to implement the pipelining only using repeaters. However, special data synchronization needs to be handled in such structures as discussed in [16]. Noise and other reliability issues also need to be resolved [68] before wave-pipelining can be utilized in NoCs.

Recently, as another alternative approach, uninterrupted wire configurations [58] such as on-chip transmission-lines (T-lines) have been intensively studied to tackle the global communication issues. Compared with the repeated global wires, T-lines deliver signals with speed of light in the medium and consume much less power as well since the wave propagation eliminates the full-swing charge and discharge on the wire and gate capacitance. To better understand the performance and scaling trends of on-chip T-lines, some works [25] [24] study the throughput of different wire lengths for scaling technologies and compare the data with repeated RC wires. Latency and energy advantage of T-lines are also shown in [33]. However, the inter-symbol interference (ISI) could be the barrier for achieving high data rate for T-lines. [3] and [22] propose to add termination resistance at the wire-end to reduce the signal distortion and derive the formula of optimal resistance value. Similarly, signal distortion reduction by adding shunt resistance between differential T-lines is also demonstrated in [30] [10] [26].

Borrowed from the off-chip IO interconnect design, equalization tech-

niques have been used for on-chip T-lines to reduce ISI and improve throughput. [63] [42] [43] use various passive networks to build equalizer to compensate off-chip/on-chip T-line loss, and show the very low power consumption. However, larger chip area, possible static current, and sensitivity to process variations, limit the usage of passive equalization. On the other hand, active filters [38] [79] [51] have been adopted to build equalized interconnect schemes for high-throughput on-chip global communications. For the state-of-the-art active-equalized on-chip interconnect design using 90 nm CMOS process [39], silicon measurement results show about 2 Gbps/μm throughput density while consuming 0.7 pJ energy per single bit transmission over a 10 mm uninterrupted on-chip global wire. As the throughput demand increases, lots of equalizer schemes used in off-chip IO may find their applications for on-chip communication. As one category of off-chip equalizers, continuous-time transmitter or receiver structures, such as current-mode logic (CML) driver [69] [29] and continuous-time linear equalizer (CTLE) [82] [7], have gradually been used for on-chip global signaling due to the reduced cost of clocking and synchronization.

I.3 Dissertation outline

Chapter II reviews background of on-chip transmission-line which includes basic theory, T-line modeling and simulation approaches. Also, we briefly introduce a worst-case eye-diagram prediction algorithm, which is adopted all through this dissertation to estimate T-line eye-opening during optimization.

Chapter III gives an overview on six current on-chip global interconnect structures and develop simple performance models which can be used for early-stage system-level analysis. A general design framework is also proposed to design and optimize one category of global interconnects based on on-chip T-line technology. A group of experiments is performed to study and compare latency, energy per bit, throughput, area, and signal integrity of six global interconnect structures

over several technology nodes.

Chapter IV extends the pipelined *RC* structure discussed in Chapter III, and explores the performance of flip-flop-based pipelined global interconnects with considering voltage and technology scaling for different design applications. We propose a general evaluation flow to study the impacts of pipelining depth, voltage scaling, and different processes on the performance of pipelined interconnects and show the potential overall performance improvement through voltage scaling.

In Chapter V, we propose an equalized global link architecture for energy-efficient high-speed on-chip communication by utilizing current-mode logic (CML) driver and continuous-time linear equalizer (CTLE). Performance of the proposed global link is modeled and analyzed through linear system method, and the closed-form expression for receiver eye-opening is derived to provide high-level design guidelines. We also propose an energy-efficient driver-receiver co-design flow by adopting Sequential Quadratic Programming (SQP) non-linear optimization and apply it to explore the design space of equalized global link.

Chapter VI concludes the dissertation by summarizing the main contributions and discussing several future research directions.

II

Background of On-Chip Transmission-Line

In this chapter, we review background knowledge of on-chip transmission-line, including electromagnetic analysis, parameter extraction, and SPICE simulation using a synthesized compact circuit model. Also, a worst-case eye-diagram prediction algorithm, which is adopted all through this dissertation for estimating the T-line signal integrity, is briefly introduced.

II.1 Basic theory

On-chip T-Line is very lossy due to the miniaturization of the wire cross section. Given different operating frequencies and wire dimensions, the wire can operate in either RC or LC region [35].

In RC region, the frequency is low, which makes $\omega L \ll R$. Generally $G \approx 0$ for on-chip wires, then the propagation constant can be written as

$$\gamma = \sqrt{\frac{\omega RC}{2}} + j\sqrt{\frac{\omega RC}{2}}. \tag{II.1}$$

It is noted that both the attenuation and phase velocity depend on frequency in RC region.

The condition of $\omega L \ll R$ is usually satisfied up to 10 GHz for on-chip global wires. If the frequency increases such that $\omega L \gg R$, the wire will operate in LC region and the propagation constant becomes

$$\gamma = \frac{R}{2\sqrt{L/C}} + j\omega\sqrt{LC}. \tag{II.2}$$

Therefore the attenuation constant is

$$\alpha = \frac{R}{2\sqrt{L/C}} = \frac{R}{2Z_0} \tag{II.3}$$

where Z_0 is the characteristic impedance of T-line, and the phase velocity becomes

$$v = \frac{\omega}{\beta} = \frac{1}{\sqrt{LC}} = \frac{c_0}{\sqrt{\epsilon_r}} \tag{II.4}$$

where c_0 is the speed of light in free space and ϵ_r is the dielectric constant. In LC region, both the attenuation and phase velocity are independent of frequency. Therefore, ideally the signal will travel over the T-line at the speed of light with the certain constant attenuation.

Two parameters need to be considered when deciding which region the wires are operating in. The critical wire length distinguishes lumped-element region and distributive-element region, which can be computed as follows [35]:

$$L_{critical} = \left| \frac{0.25}{\sqrt{(R + j\omega L)(j\omega C)}} \right|. \tag{II.5}$$

The other is the corner frequency f_{LC} between RC region and LC region, which is defined as:

$$f_{LC} = \frac{1}{2\pi}\frac{R_{DC}}{L} \tag{II.6}$$

where R_{DC} is the DC resistance of the wire.

Based on (II.5) and (II.6), on-chip wires need to be modeled as distributive T-lines when the wire length is larger than critical length and the signal frequency is higher than corner frequency. For the high-speed global signaling ($L \geq 1$ mm and $f \geq 10$ GHz) studied in this dissertation, normally the above conditions are met.

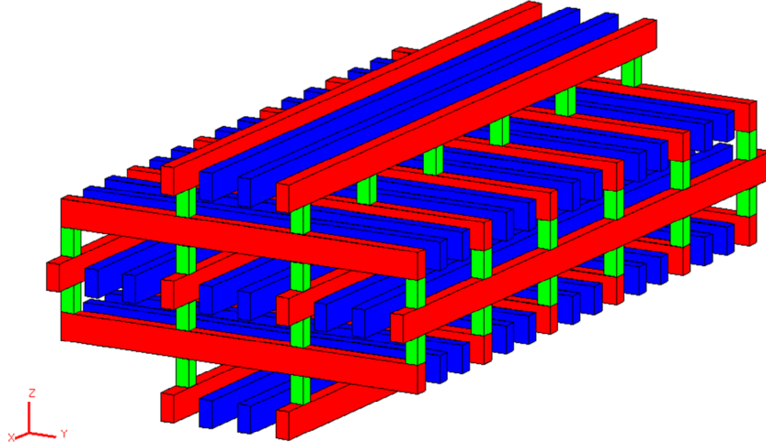


Figure II.1: 3D parameter extraction model of on-chip T-lines considering adjacent and sub-adjacent layers.

II.2 Modeling and simulation

In order to perform transient simulations of on-chip T-line in SPICE, parameter extraction and macro modeling approach are developed to generate SPICE-compatible T-line model. Proper templates need to be used during extraction to cover the full-spectrum frequency characteristics of on-chip T-line. Furthermore, extracted data are synthesized into high-quality macro models which can accurately mimic frequency response of T-line with certain physical meanings.

As shown in Fig.II.1, a complete 3D $R(f)L(f)C$ extraction[1] model needs to include vias (green wires), adjacent and sub-adjacent layers. To build well-controlled on-chip T-line structure, power and ground shielding (red wires) are inserted around signal wires (blue lines) to serve as the current return-paths for controlling the T-line inductance. The adjacent orthogonal loadings only affect the capacitance extraction and will introduce large run-time due to the non-uniform structure. In [20], it is shown that 2D capacitance extraction by replacing orthogonal layers with ground plane only slightly overestimates the value but is still

[1]Resistance and inductance of on-chip T-line are frequency-dependent due to the skin effect and proximity effect when frequency goes high. Because of the high resistivity of on-chip T-lines, dielectric loss is normally neglected and capacitance is nearly frequency independent.

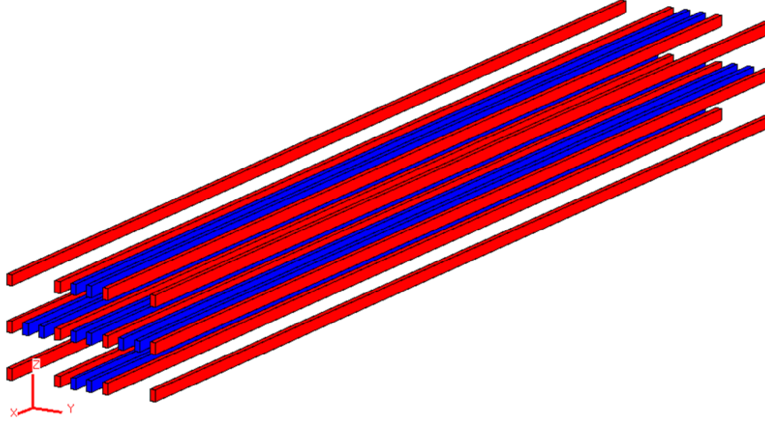


Figure II.2: 2D inductance extraction model of on-chip T-lines considering sub-adjacent layers.

acceptable when loading density and wire-to-wire coupling are high. As a result, 2D $R(f)L(f)C$ extraction is typically performed to study the characteristics of on-chip T-line [17].

For 2D extraction, capacitance is extracted based on a template which includes the studied signal lines, power/ground shielding located on the same layer, and ground reference planes above or below, representing the heavy-loaded adjacent orthogonal layers. For inductance extraction, more power/ground shieldings located on the same layer or sub-adjacent layers (in the parallel direction) need to be considered in order to capture the wide-band characteristic of wire inductance [20]. A typical 2D inductance extraction template is shown in Fig.II.2.

Based on the proper extraction templates, field solvers can be applied to extract the RLC parameters of on-chip T-line for given list of sample frequencies. The sample points of extraction frequency should be large enough to capture all the wide-band effects. Typically, extraction frequency ranges from DC to the point which is ten times larger than the signal frequency, and 3-5 points per decade are evenly chosen along the logarithm scale. In this dissertation, we use CZ2D of EIP tool suite from IBM [31] to perform 2D $R(f)L(f)C$ extraction. After extraction, frequency-dependent tabular models (such as W-element) are generated, and can

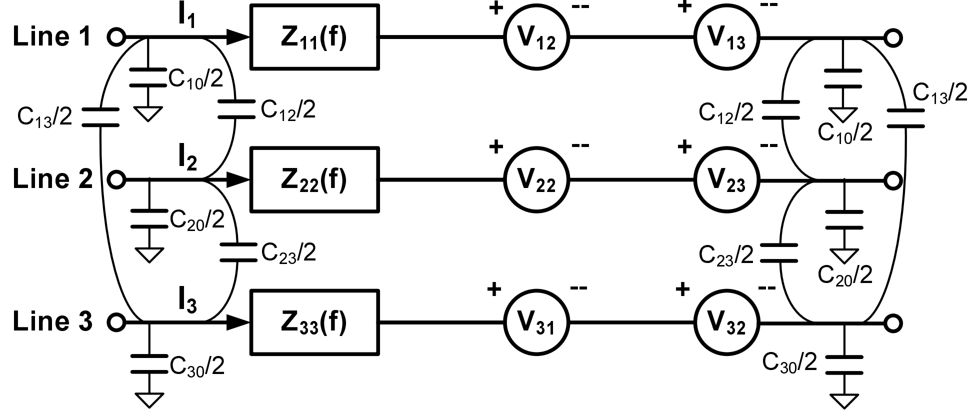


Figure II.3: One segment of three-line network using compact circuit model introduced in [17].

be directly used for SPICE simulation. However, due to the stability and accuracy issues of tabular models, we actually adopt a stable compact circuit model [17] synthesized from extracted tabular data to simulate on-chip T-line in this work.

Basically, the compact circuit model is a multi-line network which models each signal line as a frequency-dependent impedance $Z(f)$ in series with a bunch of current-controlled voltage sources (CCVSs) used for modeling the mutual terms of T-line impedance. To better illustrate this model, we take one segment of the three-line network shown in Fig.II.3 as an example. $Z_{ii}(f)$ indicates the self impedance of each signal line i and V_{ij} indicates the current-controlled voltage source, which can be written as $Z_{ij}I_j$ where Z_{ij} indicates the mutual impedance term and I_j is the current flowing into signal line j. Then the voltage drop across each line becomes

$$V_i = \sum_{j=1}^{3} Z_{ij} I_j. \tag{II.7}$$

Frequency-independent self and mutual capacitance are also added in the network according to extracted C_{i0} and C_{ij} values. To synthesize the frequency-dependent impedance Z_{ii} and Z_{ij} into a finite RL network, we adopt the approach in [40],

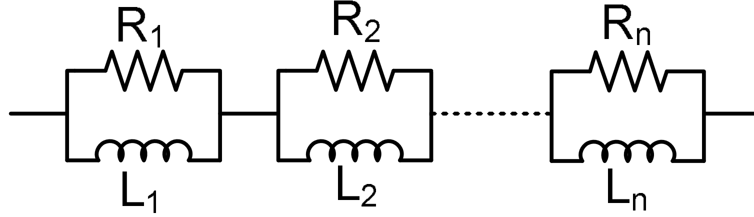


Figure II.4: A n-pole foster RL filter representation of frequency-dependent impedance.

and write this series impedance (self or mutual) as

$$Z_s(\omega) = R(\omega) + j\omega L(\omega) = R_{dc} + j\omega L_\infty + R_s(\omega) + j\omega L_s(\omega). \quad \text{(II.8)}$$

DC resistance R_{dc} and high-frequency inductance L_∞ are subtracted from approximation, and a so-called foster filter topology is used to synthesize the frequency-dependent part $R_s(\omega) + j\omega L_s(\omega)$ as described below.

Fig.II.4 shows the structure of foster filter composed of a series connection of parallel RL elements. The impedance of this foster form network is

$$Z(s) = \sum_{i=1}^{n} \frac{R_i s}{s + p_i} \quad \text{(II.9)}$$

where $p_i = R_i/L_i$ is the pole contributed by each parallel RL segment. Therefore, it is clear to visualize the contribution of various elements to the total frequency-dependent impedance in the foster filter. Now we need to solve the poles (p_i) and residues (R_i) to achieve a good fit to the given impedance $Z_s(\omega)$. A general fitting approach is to follow the way of model order reduction (MoR) by re-writing (II.9) as a rational polynomial representation

$$Z(s) = \frac{s\sum_{i=0}^{n-1} a_i s^i}{s^n + \sum_{i=0}^{n-1} b_i s^i} \quad \text{(II.10)}$$

where a_i and b_i are unknown coefficients replacing R_i and p_i in (II.9). By multiplying the denominator and substituting known impedance values (from extracted

tabular model) at sampled frequency point (ω_k), one can get the following set of linear equations

$$\left[(j\omega_k)^n + \sum_{i=0}^{n-1} b_i (j\omega_k)^i\right] [R(\omega_k) + j\omega_k L(\omega_k)] = (j\omega_k) \sum_{i=0}^{n-1} a_i (j\omega_k)^i. \qquad \text{(II.11)}$$

The above equations can be solved for the coefficients a_i and b_i, then poles p_i can be determined by calculating the roots of denominator polynomial in (II.10) and residues R_i can be found by partial fraction expansion or an additional linear least squares fit as discussed in [40]. The details of algorithm are omitted here due to page limit. In practical application, the above foster filter synthesis has been implemented in AQUAIA [21], which is part of EIP tool suite and integrated with the field solver to generate fitted T-line model right after extraction. To model the typical on-chip global wires ($\geq$1 mm), the number of poles $n \leq$10 and number of segments (number of cascaded multi-line networks shown in Fig.II.3) $N \leq$20 are enough to provide sufficient accuracy. In this work, we adjust n between 5 and 8 and use 2 segments per mm length for all the T-line modeling. Also, a perl script is developed to convert the compact circuit model generated by AQUAIA to a sub-circuit definition which can be directly used for HSPICE simulation.

II.3 Worst-case eye-diagram prediction

Eye-diagram is widely used to study the signal integrity of high-speed signaling system. A typical eye-diagram observed at the end of on-chip T-line is shown in Fig.II.5. It is created by overlapping the received time domain signal at a time window of certain multiple bit periods (Two bit periods in Fig.II.5). The diagram is so called because the pattern looks like a series of eyes between a pair of rails. To quantify the received signal quality, some metrics are defined based on eye-diagram. **Eye-opening**, which is the vertical distance between the lowest voltage level of received logic high signal (bit ‘1’) and the highest voltage level of received logic low signal (bit ‘0’), is used to measure the impact of noise

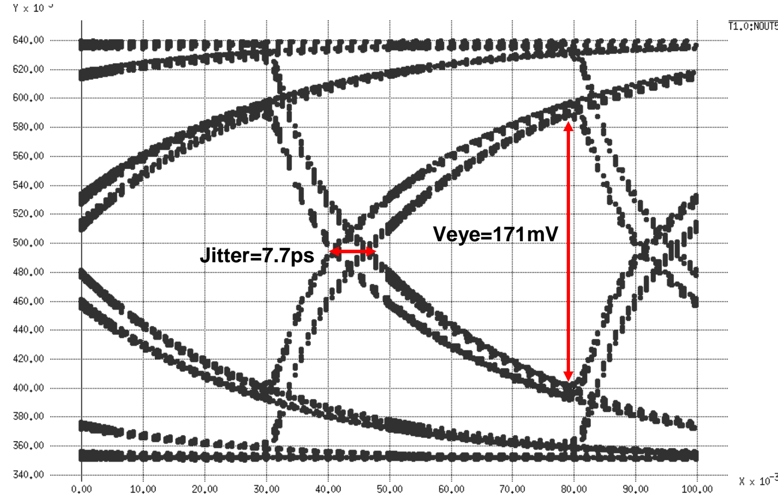


Figure II.5: A typical eye-diagram observed at the end of on-chip T-line.

or other interference on the received signal amplitude. **Timing Jitter**, which is the horizontal distance between the fastest and slowest time at which the received signal crosses the threshold voltage (normally defined as mid-point of high and low voltage level), is used to measure the impact of noise on the timing synchronization. For instance, in the eye-diagram shown in Fig.II.5, eye-opening and timing jitter are 171 mV and 7.7 ps, respectively.

Conventionally, to simulate the eye-diagram of on-chip T-line scheme, a pseudo-random bit sequence (PRBS) is adopted as input stimulus in the SPICE simulation. Usually the run contains thousands of clock cycles, which is very time-consuming. Also, due to the limited length of input sequence, worst-case scenario is not guaranteed to be captured in the simulation. Some research works [9] [2] are done to provide more efficient eye-diagram analysis methods which can predict accurate eye-opening and timing jitter values without PRBS simulation. In this dissertation, we adopt the prediction algorithm proposed in [62], which estimates the worst-case eye-diagram efficiently based on step-response of T-line.

The algorithm first assumes the transmitted signal can be treated as a

binary stream of zeros and ones

$$x(t) = \sum_{i=1}^{N} x_r(t - k_i^r T) - \sum_{i=1}^{N} x_f(t - k_i^f T) \tag{II.12}$$

where $x_r(t)$ indicates the zero-one transition for given rise time and $x_f(t)$ indicates the one-zero transition for given fall time. T is the time interval of each bit. Coefficients k_i^r and k_i^f are the slot numbers that the i-th zero-one and one-zero transition happen. If the whole signaling system is linear time-invariant (LTI), the output signal $y(t)$ can be modeled as the superposition of step responses

$$y(t) = \sum_{i=1}^{N} s_r(t - k_i^r T) - \sum_{i=1}^{N} s_f(t - k_i^f T) \tag{II.13}$$

where $s_r(t)$ is the step response of $x_r(t)$ and $s_f(t)$ is the step response of $x_f(t)$.

To analyze the worst-case eye-diagram, 8 pivot points are defined for given observed time point t_0 $(0 \leq t_0 \leq T)$ to characterize the eye-diagram. They represent the maximum and minimum voltage bounds for 4 different transition situations including zero-one, hold-one, one-zero, and hold-zero. It is shown that each voltage bound can be expressed as the maximum or minimum of a combination of $s_r(t)$ and $s_f(t)$ similar to (II.13). To calculate all these voltage bounds, the problem is formulated as

Given: *two arrays A and B*

$A = \{s_r(t_0 + T), s_r(t_0 + 2T), ..., s_r(t_0 + k_m T)\}$

$B = \{s_f(t_0 + T), s_f(t_0 + 2T), ..., s_f(t_0 + k_m T)\}$

where at time $(t_0 + k_m T)$ the step response becomes saturated.

Objective:

min or max $\sum_i A[i] - \sum_i B[i]$

Constraints:

The starting transition must be selected from A;

The transitions must be selected from A and B alternatively.

The above problem is solved by dynamic programming in [62] and proved to be the worst-case. After all the 8 voltage bounds are solved, the worst-case eye-opening

can be readily calculated based on its definition

$$V_{eye} = min\left(e_{lower01}(T), e_{lower11}(T)\right) - max\left(e_{upper10}(T), e_{upper00}(T)\right) \qquad \text{(II.14)}$$

where $e_{lower01}(T)$ indicates the voltage lower bound of transition zero-one at observed time T, similar denotation adopted for $e_{lower11}(T)$, $e_{upper10}(T)$, $e_{upper00}(T)$. For timing jitter, according to the definition, it can be written as

$$T_{jitter} = T_{jright} - T_{jleft} = max(T_{jr1}, T_{jr2}) - min(T_{jl1}, T_{jl2}) \qquad \text{(II.15)}$$

where T_{jright} indicates the time at which the latest zero-one transition (T_{jr1}) or one-zero transition (T_{jr2}) crosses the threshold voltage, and vice versa for T_{jleft}. In the algorithm implementation, simple binary search could be applied to find the T_{jr1}, T_{jr2}, T_{jl1}, and T_{jl2}.

The worst-case eye-diagram predication algorithm has been implemented in MATLAB and C. Over 20% prediction accuracy and 2000x run-time improvement have been demonstrated [62] using this method compared with the traditional 10000-bit PRBS SPICE simulation. Therefore, in this dissertation, we widely use this approach to estimate the signal integrity of on-chip T-line interconnections and incorporate it with our optimization flow to improve the accuracy and efficiency.

III

Comparison of High-Performance On-Chip Global Interconnections

In this chapter, we review six current on-chip global interconnect structures and develop simple models to analyze their architecture-level performance. We propose a general framework to design and optimize a new category of global interconnects based on on-chip transmission line (T-line). We perform a group of experiments to predict and compare the performance of different interconnections in terms of latency, energy per bit, throughput, area, and signal integrity over several technology nodes.

III.1 On-chip global interconnects

III.1.A Overview

On-chip global interconnect schemes can be divided into categories based on the operating region of wires, the signaling method, and other factors, as shown in Figure III.1. The widely-used repeated *RC* wire approach (referred as ***R-RC*** in this paper) belongs to the first category, which uses *RC*-mode dominant wires. To improve the bandwidth of repeated *RC* wires, the *R-RC* structure may be

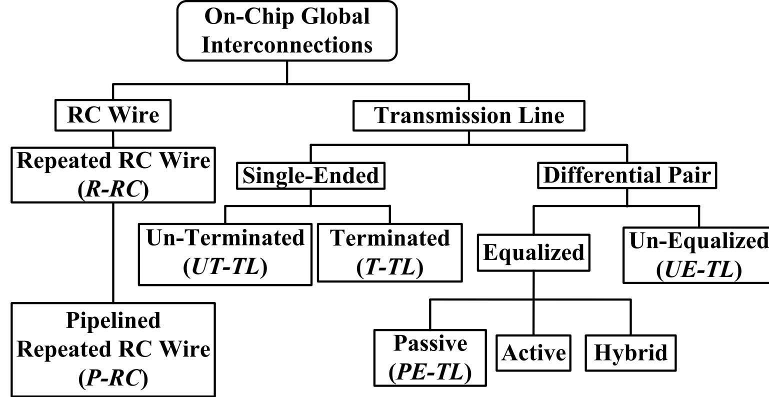


Figure III.1: The organization of on-chip global interconnect structures.

pipelined by breaking optimized *R-RC* wire into segments and inserting flip-flops. This pipelined *RC* wire strategy is subsequently denoted ***P-RC***. The other main category utilizing T-line effects of on-chip wires is comprised of two configurations, namely single-ended and differential pair, based on their respective signaling methods. For the single-ended configuration, capacitive or resistive loading (un-terminated or terminated T-line, referred as ***UT-TL*** or ***T-TL***) can be used at the wire end depending upon the throughput requirement [83]. For the differential pair configuration, conventional design mainly focuses on the optimization of T-line transceivers without adopting any equalization (referred as ***UE-TL***). Passive networks [78] [84] are used in some recent research to equalize on-chip T-line (referred as ***PE-TL***) whereas other on-chip equalization implementations using active circuits or even hybrid structures could be potential future research directions.

Multi-dimensional design tradeoffs, which are normally related to the latency, energy dissipation, throughput, area/cost, and reliability (noise), should be considered while designing an on-chip global interconnect scheme. For the six different structures mentioned above, we use a 45 nm CMOS process as an example to illustrate the tradeoff relations along multiple performance dimensions in Figure III.2. By observing this figure, designers can easily identify complex design tradeoffs and make determinations based on given specific applications. By

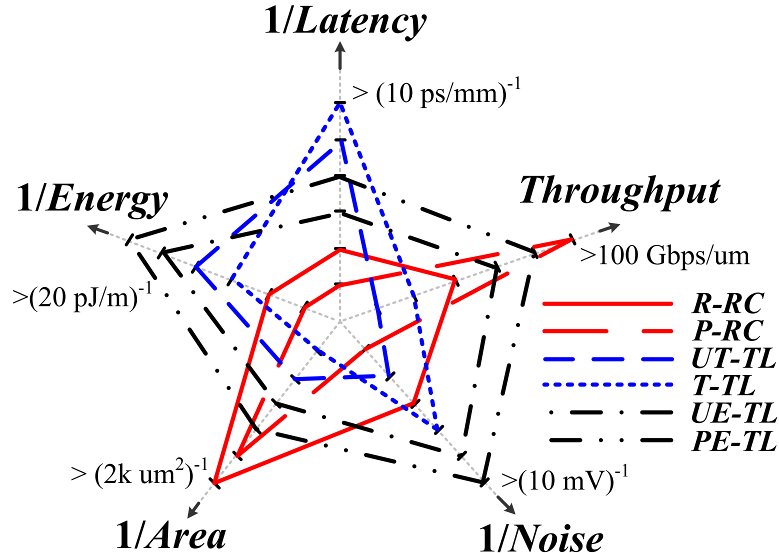


Figure III.2: The multi-dimensional design tradeoffs of different global interconnect structures.

using 45 nm CMOS, RC wire has advantages in throughput density (using P-RC) and area/cost (for both R-RC and P-RC) because of their small wire dimensions. On the other hand, single-ended T-lines (UT-TL and T-TL) could be used for low-latency application by utilizing wave propagation. In terms of low energy and noise, differential T-lines (UE-TL and PE-TL) should be better candidates due to their larger wire input impedance, low-power transceiver circuitry, and differential configuration. In order to identify these complex tradeoff relations at the early-stage and also from the architecture-level, we have developed simple performance models to help designers to do approximate but trend-following estimation, which will be discussed later in Subsection III.1.D.

III.1.B Interconnect schemes

We show the detailed structure for each global interconnection scheme mentioned above and briefly introduce the features of these schemes as follows. RC wires are composed of RC segments and inserted buffers, and flip-flops can

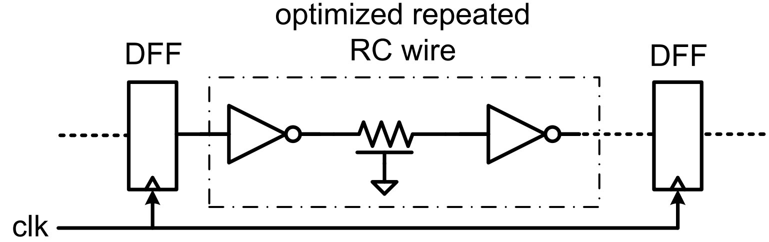


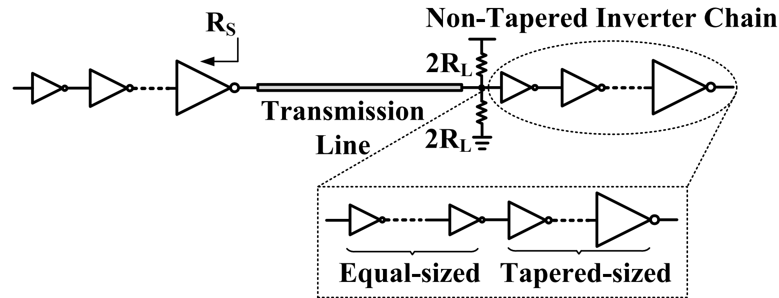
Figure III.3: One stage of pipelined repeated RC wire (P-RC structure).

be further added along the RC wires to improve the bandwidth. Inverter chains are utilized in single-ended T-lines to transmit and receive full-swing signals. For differential T-lines, low-swing signals are generated by differential driver, and recovered back to full-swing using sense-amplifier at the end.

For repeated RC wires (R-RC), the long wire is divided by repeaters into several RC segments. The strength of repeater (size of inverter) and length of wire segments could be optimized according to different design objectives for a given wire geometry. To further improve the bandwidth, P-RC is proposed as shown in Figure III.3. Assuming the R-RC wire between two flip-flops is **already optimized** based on one specific objective (minimum latency in this study), the only variable for P-RC wire optimization is the number of flip-flops inserted (a.k.a. pipelining depth). By utilizing pipeline, bandwidth of the R-RC wire is improved with overhead of energy and latency, therefore, the best pipelining depth can be decided in terms of the lowest energy over bandwidth ratio (conceptually similar to the energy-delay product, refer to the Appendix A for the mathematical derivation). In practice, there is an upper-bound for the maximum pipelining depth, so in the following experiments, P-RC is optimized based on **the lowest energy/bandwidth with the maximum pipelining depth constraint**.

For the single-ended T-line schemes (UT-TL and T-TL) which are shown in Figure III.4, tapered or non-tapered inverter chain is adopted as the driver and receiver, depending on different termination scenarios. Compared with unterminated scheme (as shown in Figure III.4(a)), resistive termination improves

(a) Un-Terminated T-Line scheme.



(b) Terminated T-Line scheme.

Figure III.4: Single-ended T-line schemes for on-chip global interconnect.

the bandwidth by alleviating the ISI, but lowers the swing of output signal and burns extra power on the termination. As a result, a non-tapered inverter chain (as shown in Figure III.4(b)) is devised to amplify the received signal and recover it back to digital level. In this kind of single-ended schemes, driver impedance (and terminated resistance, if any), first inverter size, and number of stages in the inverter chain are the variables to be optimized during the design.

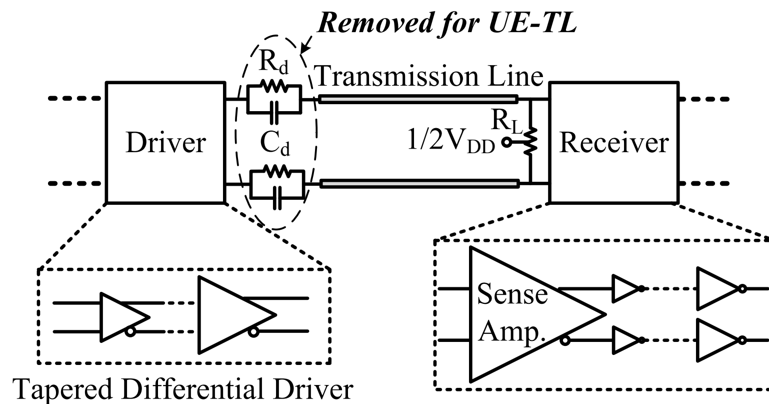


Figure III.5: Differential T-line schemes for on-chip global interconnect.

For the differential T-line schemes (*UE-TL* and *PE-TL*, shown in Figure III.5), the tapered differential drivers[1] could be used to provide the low driver impedance, whereas at the receiver side, a sense-amplifier (SA) [57] is adopted to amplify the attenuated signal. The following inverter chains further increase the slew rate to improve signal quality. Circuit design of T-line receiver has been discussed in the previous work [84]. In this work, we improve the design there to facilitate the SA bandwidth by using more sophisticated transistor-sizing strategy, which improves the bandwidth by about 2x compared with the one in [84]. Since the receiver is designed and optimized for each given technology individually, noise and sensitivity performance (capability of recovering 50 mV voltage difference) of receiver is guaranteed even for smaller technology node by automated transistor-sizing. As a result, real performance change of receiver circuit as technology scaling is modeled and considered during the study of T-line schemes in this work. For the equalization approach, the passive-equalized scheme adopts a parallel *RC* network at the driver side to flatten the overall frequency response by utilizing the high-pass transfer characteristic. In this scheme, driver impedance, resistance and capacitance value in the passive equalizer, and terminated resistance are optimized with the constraint that enough eye-opening should be observed at the end of T-line in order to be safely captured by the receiver.

III.1.C Global wire modeling

We model on-chip global wires using different approaches based on operating frequency and wire geometry. Global *RC* wires in scheme *R-RC* are normally represented by distributed Π model composed of wire resistance and capacitance. Following [76], 2D closed-from equations in [65] are utilized to calculate wire capacitance. The wire geometry and other design parameters for *R-RC* structure

[1]Drivers could be CML or other types, but in following optimization and experiments, differential T-line drivers are assumed to be voltage sources with output resistance R_s to simplify the analysis and optimization.

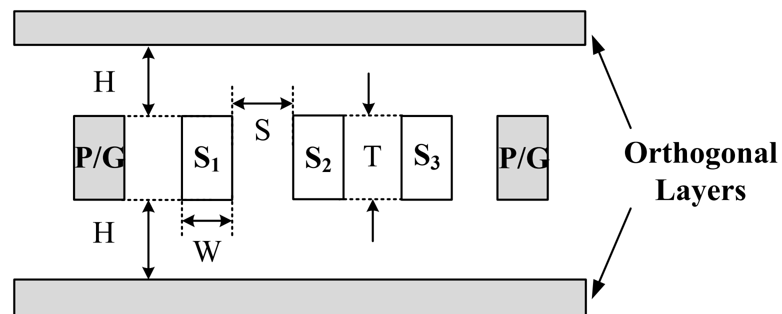
Table III.1: Design parameters for global *R-RC* and *P-RC* wires based on ITRS Roadmap 2007 and SPICE simulation

Tech Node/nm	90	65	45	32	22
supply voltage/(V)	1.2	1.1	1.0	0.9	0.8
dielectric constant ϵ_r	3.1	2.9	2.6	2.4	2.1
$\rho_{Cu}/(\mu\Omega\cdot\text{cm})$	2.53	2.73	3.10	3.52	3.93
min global pitch/(nm)	300	210	135	96	75
aspect ratio (A/R)	2.2	2.3	2.4	2.5	2.6
min-inv FO4 delay*/(ps)	11.9	8.4	4.6	2.8	1.6
flip-flop T^*_{c-q}/(ps)	34	24	13	8	4
flip-flop T^*_{setup}/(ps)	22	15	10	6	3
flip-flop C^*_{DFF}/(fF)	34.3	23.8	17.0	9.0	4.3

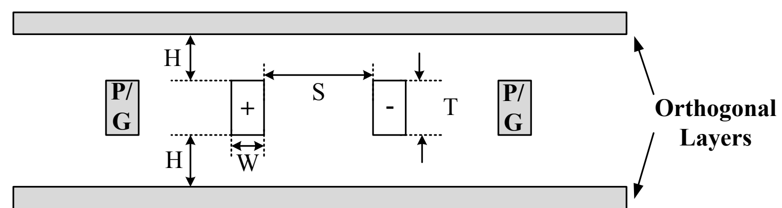
*Data are obtained by simulation using a predictive model [70].

are listed in Table III.1, based on the predictions of the 2007 ITRS roadmap [59]. For *P-RC* structure, flip-flop parameters including the clock-to-q time, setup time, and effective capacitance are derived by SPICE simulation using a predictive device model, as listed in the last three rows of Table III.1.

For other T-line schemes, we adopt single-ended or differential strip-line configurations to model global wires, as shown in Figure III.6. For single-ended scenarios, we insert power/ground (P/G) lines every three wires (shown in Figure III.6(a), following the typical wiring and power arrangement for global wide data bus [66]) to provide current return paths in order to form well-controlled on-chip T-line structures. The adjacent orthogonal layers could be replaced by the ground planes if performing 2D capacitance extraction. Considering orthogonal loading, the capacitance obtained by a 2D extraction is slightly overestimated compared with the 3D value [20], but still acceptable with the assumption that on-chip loading density and lateral wire-to-wire coupling are very high. The dimensions of this single-ended T-line structure are listed in Table III.2, following the settings in [83]. Fat and unscaled wires implemented on the top-layer are utilized to reduce the resistive loss in this scenario, first proposed in [56], to alleviate the increasing



(a) The cross section of single-ended stripline used in *UT-TL* and *T-TL* schemes.



(b) The cross section of differential stripline used in *UE-TL* and *PE-TL* schemes.

Figure III.6: The wire configurations for on-chip T-line schemes.

Table III.2: Parameters of on-chip global T-line used in UT-TL/T-TL schemes

W (μm)	S (μm)	H (μm)	T (μm)	R (Ω/mm)	Z_0 (Ω)	f^*_{LC} (GHz)
1.6	1.6	3.2	2.4	6.0	40	1.79

$^*f_{LC}$ indicates the corner frequency of RC and LC region.

Table III.3: Design parameters for UE-TL/PE-TL schemes (Wire Length = 5 mm)

Tech Node/nm	90	65	45	32	22
SA bandwidth/(GHz)	12.5	20.0	40.0	66.7	125
UE-TL wire width/(μm)	0.350	0.350	0.400	0.448	0.504
PE-TL wire width/(μm)	0.300	0.315	0.375	0.400	0.444

H=T=S=2W for T-lines used in UE-TL and PE-TL.

RC delay of scaled on-chip wires. With the improvement of device speed, the transmission-line effect does kick in and cannot be neglected while modeling such fat global wires; on the other hand, it has been verified by previous research works [17] [83] [20] that on-chip bus configuration comprising a low-impedance driver and uninterrupted fat wire outperforms repeated wire structures in high-performance applications (e.g., high-end processors [66]) due to the T-line effect. As a result, we adopt the configuration comprising a low-impedance driver and uninterrupted fat wire and assume the utilized wire geometry shown in Table III.2 maintains as technology scales. As shown in the last column of Table III.2, LC-mode behavior is dominant for this wire geometry, which speeds up the signal transmission through wave propagation.

We also devise a similar coplanar configuration for differential T-lines as shown in Figure III.6(b). Here, only one pair of wires is placed in a P/G bay in order to reduce the crosstalk noise. Wire dimensions of such configuration are determined by the resistive loss at given signal frequency, which is changing with the technology. Considering the differential T-line schemes discussed in this work, the overall signal bandwidth is limited by the SA in transceiver, as listed in Table

III.3 at each technology node. We derive the minimum wire widths of differential T-lines that satisfy the eye-opening constraint by binary search and SPICE simulations, as shown in the third and forth row of Table III.3.[2] By comparison, it can be seen that equalization helps to improve the data density by supporting narrower wires at the same bit rate.

The modeling and simulation of on-chip T-lines generally incorporate two steps. First, we extract the frequency-dependent $RLGC$ parameters for the given wire structure using field solvers. For on-chip wires, since dielectric loss can be ignored and wire capacitance is basically frequency independent, $R(f)L(f)C$ extraction is generally performed [17]. The frequency-dependent impedance extraction requires a group of P/G wires located in signal layer and sub-adjacent layers (parallel to the signal layer) to serve as return paths in order to capture the wide-band characteristic of wire inductance [20]. As a result, parameter extraction can generate the tabular model or other kinds of SPICE-compatible macro-model. In the second step, SPICE simulations can be performed to study transient characteristics and signal integrity. In this work, we evaluate performance metrics of all the T-line schemes based on the tabular model generated by 2D-$R(f)L(f)C$ extraction. As a more practical modeling approach, a stable compact circuit model [40] synthesized from 2D-$R(f)L(f)C$ extraction is used to study the signal integrity of global T-line schemes, as discussed in Section III.4.

The above discussed T-line structures can be implemented using CMOS process. For single-ended schemes, T-lines are implemented on the top-layers of copper stacking with well designed power/ground arrangement to control the T-line effect, as shown in the bus design for high-end processors [66]. Differential T-lines also have been implemented recently for global clock distribution [73]. Similar configuration can be borrowed here to implement global differential data bus [77].

[2]The wire width values in this table are different from the previous work [85] because of the improved wire width optimization subroutine in this work. It is shown that narrower T-lines can be utilized to satisfy the eye-opening constraint, resulting in higher throughput density in the following results and also affecting other metrics.

There is no any further specific layout style required for such T-line configuration, however, to improve the signal integrity (e.g., crosstalk), twisted structure might be used for the real chip designs [73] [58].

III.1.D Performance analysis

An approximate high-level analysis is performed here to reveal how architecture level performance metrics of different global interconnects behave with technology scaling under the **min-d** (minimum delay) objective[3]. As a result, we derive simple linear models, which can be used for designers to approximately estimate the performance of different interconnect structures at the early stage. In the following analysis, basic technology-determined parameters, including supply voltage V_{DD}, dielectric constant ϵ_r, min-sized inverter FO4 delay τ, as well as the total wire length L, are chosen to be the variables to build such models.

[3] *P-RC* scheme is optimized based on min-energy/bandwidth w/ maximum pipelining depth constraint, as discussed in Section III.1.B.

Table III.4: Modeling performance metrics (normalized delay, normalized energy, normalized throughput, area) of six global interconnection structures using technology-defined parameters.

Structures	$R\text{-}RC/P\text{-}RC$	$UT\text{-}TL/T\text{-}TL$	$UE\text{-}TL/PE\text{-}TL$
Normalized Delay	$K_1\sqrt{\epsilon_r/\tau} + K_2\tau/L$	$K_1\sqrt{\epsilon_r} + K_2\tau/L$	
Normalized Energy	$K_1\epsilon_r V_{DD}^2 + K_2\tau V_{DD}^2/L$	$K_1 + K_2\tau V_{DD}^2/L$	$K_1\tau V_{DD}^2 + K_2\tau V_{DD}^2/L$
Normalized Throughput	$1/(K_1\tau^2 + K_2\sqrt{\epsilon_r\tau L})$	$K_1 1/\tau + K_2 1/L$	$1/(K_1 L\sqrt{\tau} + K_2\tau\sqrt{L})$
Area	$K_1\tau L + K_2\tau^2$	$K_1 L + K_2\tau^2$	$K_1 L^{3/2} + K_2 L^2/\sqrt{\tau}$

Latency

For the latency evaluation, we define the **normalized delay** as:

$$delay_n = \frac{\text{propagation delay}}{\text{wire length}} \tag{III.1}$$

where the propagation delay includes both the wire delay and the gate delay (repeaters and flip-flops in *RC* wires or transceivers in other T-line schemes).

For *R-RC* structure, it can be shown that [76]

$$delay_n^{R-RC} \propto \sqrt{r_0 c_{nmos} r_w c_w} \propto \sqrt{\epsilon_r/\tau} \tag{III.2}$$

with the assumption that the output resistance of a min-sized inverter r_0 is roughly constant across different technologies, and FO4 delay τ reduces with the same scaling factor as the feature size. For *P-RC* structure, additional delay is introduced by the inserted flip-flops and is linearly proportional to the pipelining depth and FO4 delay τ. For long global *RC* wires ($\geq$5 mm) and advanced technologies (beyond 45 nm), the experimental results show that the maximum pipelining depth is chosen to reduce the overall energy over bandwidth ratio. Therefore, in our simple models, latency overhead of *P-RC* wire is assumed to be a linear function of FO4 delay τ only.[4]

For other T-line schemes, total latency can be expressed as the sum of wire delay and transceiver delay. For *LC*-mode dominant T-lines, normalized wire delay is proportional to $\sqrt{\epsilon_r}$. The transceiver delay could be simply represented by the FO4 delay τ linearly.

Considering the total wire length in our delay models, the final results are shown in the second row of Table III.4, where coefficients K_1 and K_2 reflect the different process technologies. It can be seen that if $\sqrt{\epsilon_r/\tau}$ item is dominant in the normalized delay expression for the *R-RC* and *P-RC* structures, the trend is an

[4] This assumption also holds for the following analysis of *P-RC* wire. Regarding the performance analysis of ideal *P-RC* structure without maximum pipelining depth constraint, refer to Appendix A.

increase in the normalized delay of RC wires as technology scales. The same table shows the opposite trend for the T-line schemes, where normalized delay decreases with the reduction of dielectric constant and scaling of transistors.

The most significant sources of error in our proposed delay model come from the simple modeling of transistor gate capacitance and the approximation of the *P-RC* scheme in the short wire-length range. As technology scales, the gate capacitance per unit width actually reduces instead of being constant[5], resulting in the decreasing c_{nmos} in (III.2), which partly cancels out the RC-wire slowing caused by the wire scaling. Modeling error of the *P-RC* scheme in the short wire-length range ($L \leq 3$ mm) is due to the neglect of optimal pipelining depth changing in the delay model. As discussed in Appendix A, the optimal pipelining depth in terms of lowest energy-bandwidth ratio reduces as the wire length decreases, and is proportional to the wire length. Therefore, for the cases with short wire length, the τ/L item in the delay model of *P-RC* structure approaches τ, causing relatively larger errors. This type of error source also applies to energy and throughput modeling of *P-RC* structure, as shown below.

Energy per Bit

The **normalized energy per bit** is used to evaluate the energy dissipation of global interconnect, which is defined as follows:

$$energy_n = \frac{\text{energy per bit}}{\text{wire length}} = \frac{\text{power}}{\text{bit rate} \times \text{wire length}}. \qquad \text{(III.3)}$$

The bit rate for RC wires is assumed to be the inverse of propagation delay over the total wire length for *R-RC* structure (not pipelined), or the inverse of delay between two flip-flops for *P-RC* structure (pipelined).

[5]CMOS gate capacitance per unit width (C_g) equals to $\epsilon_{ox} L/t_{ox}$, where L is the channel length, and t_{ox} indicates the oxide thickness. For long channel devices, L/t_{ox} is roughly constant due to the same scaling factor along different dimensions of transistor, whereas for short channel devices, oxide thickness cannot scale as fast as transistor width and length due to leakage and process considerations. Therefore, gate capacitance per unit width decreases as technology scales. In our study, C_g is 1.5 fF/μm at 90 nm node and 0.8 fF/μm at 22 nm node.

As discussed in [76], the normalized energy per bit for *R-RC* structure satisfies that

$$energy_n^{R-RC} \propto c_w V_{DD}^2 \propto \epsilon_r V_{DD}^2. \qquad \text{(III.4)}$$

For the *P-RC* structure, additional energy consumed by inserted flip-flops is represented by $C_{FF}V_{DD}^2$, which is approximately proportional to τV_{DD}^2. Similar to the delay modeling shown above, linear models are built by combining energy consumed on wire and gate together, shown in the third row of Table III.4.

For T-line schemes, we consider the power dissipation on the wire and transceiver individually. The power consumed on T-lines is basically proportional to V_{DD}^2 if assuming wire input impedance remains constant across technologies. Transceiver dynamic power is linearly proportional to fCV_{DD}^2, where f is the clock frequency and C represents the total gate capacitance of the transceiver. Combining these two factors together,

$$energy_n^{TL} \propto T_C V_{DD}^2/L \propto \tau V_{DD}^2/L \qquad \text{(III.5)}$$

with the assumption that cycle time T_C and gate capacitance C scale by the same rate as τ. Linear models based on the analysis are shown in Table III.4. As shown in (III.4), compared with *RC* wires, if $\epsilon_r V_{DD}^2$ item is dominant in the total normalized energy expression, T-lines will consume less energy as technology scales since τ shrinks more rapidly than ϵ_r does.

For energy modeling, the largest errors in *R-RC*/*P-RC* modeling come from the same sources as discussed in previous subsection. Errors in the modeling of T-lines may relate to the neglect of wire input impedance variations across different technologies.

Throughput

The **normalized throughput** (or throughput density) is defined as:

$$throughput_n = \frac{\text{bit rate}}{\text{wire pitch}} \qquad \text{(III.6)}$$

which is adopted to compare the amount of data can be transmitted for a given cross area in a given time interval.

From (III.2) and assuming that wire pitch also scales down as FO4 delay τ, for an *R-RC* structure,

$$throughput_n^{R-RC} \propto 1/(\sqrt{\epsilon_r}\tau L). \tag{III.7}$$

Regarding a *P-RC* structure, normalized throughput can be derived from the normalized delay expression. The flip-flop delay should account for the τ^2 item in the denominator of normalized throughput expression. The general throughput model for *RC* wires is summarized in the third row of Table III.4.

Unlike the *RC*-dominated structures, the bit rate of T-line schemes is usually limited by the bandwidth of transceivers (except for *UT-TL*, which is determined by the wire itself). As a result, assuming the transceiver bandwidth is inversely proportional to τ, the bit rate of *T-TL/UE-TL/PE-TL* structures is inversely proportional to τ. For *UT-TL*, the bit rate is constant as technology scales, but is approximately inversely proportional to wire length L. Therefore, a general model of bit rate of T-lines can be represented by,

$$\text{Bit Rate} = C_1 1/\tau + C_2 1/L \tag{III.8}$$

where C_1, C_2 are fitting coefficients. Considering wire pitch, for single-ended T-lines, wire pitch does not change with technology and wire length. However, for differential T-lines, larger wire pitch is required as technology scales or wire becomes longer. An approximate relation is shown below based on simple modeling of wire resistance considering DC and AC components separately,[6]

$$\text{Wire Pitch} = C_1\sqrt{L} + C_2 L/\sqrt{\tau} \tag{III.9}$$

[6] In (III.9), the $\sqrt{L}$ value comes from the DC component of wire resistance, whereas $L/\sqrt{\tau}$ item comes from the AC component of wire resistance, caused by the skin effect.

where C_1, C_2 are fitting coefficients. In deriving this equation, we neglect the minor change of supply voltage V_{DD} as technology scales, and assume the most important frequency component of T-line skin effect for each technology is also proportional to $1/\tau$. Combining (III.8) and (III.9), we derive the throughput models for each T-line scheme in Table III.4.

For most transceiver-limited T-line structures, even when considering the increasing wire pitch as technology scales, the throughput density will still exceed that of an *R-RC* structure due to the rapid improvement of transceiver bandwidth.

The proposed model may have a larger error for *T-TL* scheme at the most advanced technology node (22 nm) as wire length increases ($\geq$7 mm) because the bandwidth of the overall structure becomes wire-limited and does not improve as technology scales.

Area

Chip area consumed by different interconnect structures comprises two parts: wire area and circuit area. Wire area is the wire pitch multiplied by the total wire length L. The pitch scaling trend has been discussed in the previous section. Circuit area will reduce quadratically as technology scales, approximately proportionally to τ^2. Based on the analysis in previous subsections, the area model for each interconnection scheme is shown in the forth row of Table III.4.

Typically, wire area dominates total chip area. As a result, *RC* wires consume less area compared to T-line structures, and *RC* wire area decreases more quickly as technology scales compared to T-line structures. The area of differential T-lines actually increases as technology scales due to increasing wire pitch.

Area model of *RC* wires may show a substantial error due to the reasons cited in the subsection of delay modeling.

III.2 Design methodology

In this section, methodologies to design and optimize the six global interconnection structures are discussed. As previously mentioned, for the *R-RC* scheme, we adopt the optimization framework in [76], which is based on analytical formulae and numerical experiments to study the performance metrics under different design goals across multiple technology nodes. In terms of the *P-RC* structure, we develop a simple MATLAB flow to optimize the pipeline depth based on the lowest energy/bandwidth ratio, with the assumption that *R-RC* wire between flip-flops are optimized and the maximum pipeline depth is given. A detailed flow description is omitted here for the sake of brevity. For more information regarding performance analysis of ideal pipelined repeated *RC* wires (w/o maximum pipelining depth limit), the basis of our pipelined *RC* wire flow, please refer to the Appendix A.

This section focuses on the design of on-chip T-line schemes. Here, we propose a general framework by modeling on-chip T-line and transceiver circuitry separately and utilizing well-behaved optimization routines to generate the optimal design for given design specification, as illustrated in Figure III.7. We will introduce the application of this design framework on the single-ended T-line schemes (including *UT-TL*/*T-TL*) and differential T-line schemes (including *UE-TL*/*PE-TL*) respectively in the following.

III.2.A Single-ended T-lines

The methodology to optimize single-ended T-line structures is proposed and discussed in [83]. We summarize this methodology according to the general design framework shown in Figure III.7. Corresponding to the proposed framework, on-chip wire is modeled using the frequency dependent tabular model generated by the field solver, and the characteristic of the transceiver circuit is obtained by SPICE simulation. Also, we use SPICE to evaluate the performance metrics of

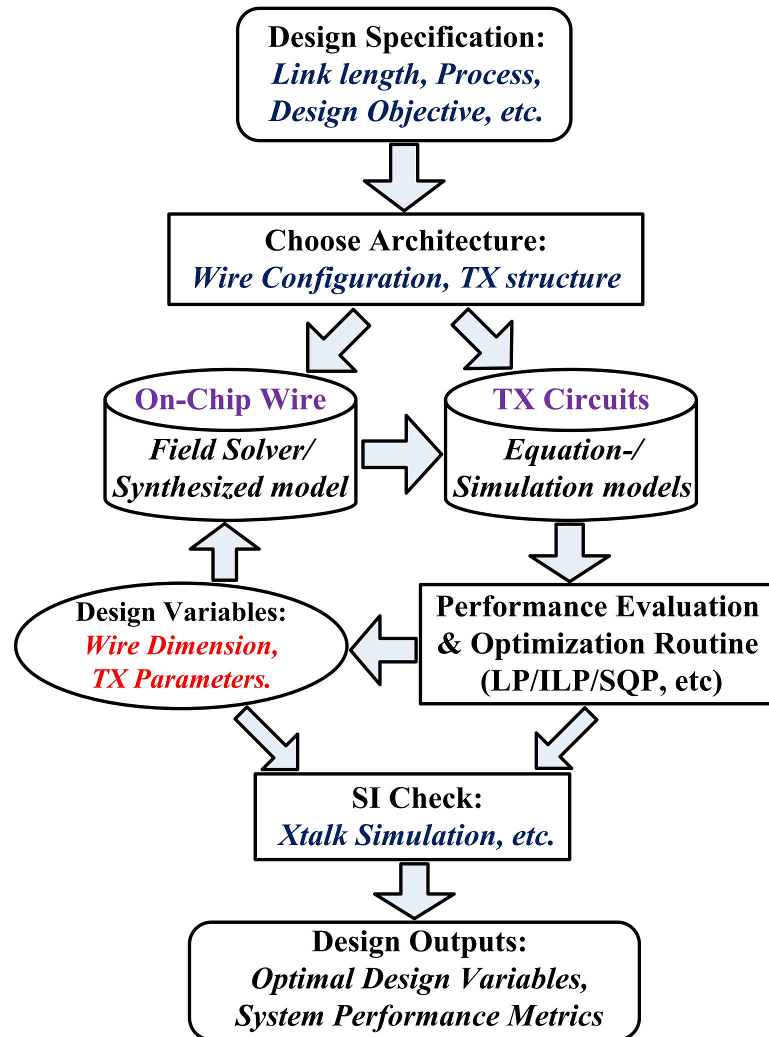


Figure III.7: The design framework for on-chip global T-line structures (UT-TL/T-TL/UE-TL/PE-TL).

whole structure. Since wire dimensions are well defined, design variables of interest relate only to the transceiver circuit (the inverter chain), including the first inverter size S_1 and number of stages N. For the terminated scheme (T-TL), termination resistance R_L also needs to be optimized for achieving high throughput. The optimization routine for this kind of scheme is comprised of two phases, namely determining the optimal clock rate and choosing the optimal variables in terms of the given objective. Finally, signal integrity is studied by SPICE simulation and the framework outputs the optimal design variables and corresponding performance metrics.

III.2.B Differential T-lines

For differential T-line schemes, the adopted methodology is based on the constrained non-linear programming formulation [78] and sequential quadratic programming (SQP) approach [8]. The flow begins with modeling of wires and transceivers using different means. For on-chip wires, 2D-$R(f)L(f)C$ tabular model is still utilized. For the transceiver circuit, though, we adopt a closed-form equation-based model, which is generated by fitting SPICE simulation data.[7] To evaluate the performance metrics of the whole structure, we combine the models of wire and transceiver together, and then utilize the approach in [62] to estimate the wire-end eye-opening. The optimization routine for differential schemes initially tries to find the smallest wire dimension that satisfies the eye-opening constraint by using binary search (which generates the data in Table III.3), and then calls the SQP subroutine to optimize the design variables for the given design objective. The design variables include driver impedance R_s, passive equalizer parameters R_d, C_d (for PE-TL structure), and termination resistance R_l. The key element in formulating the differential schemes is the eye-opening constraint. In this work,

[7]The details of these models can be referred to [78]. In this work, we adopt the same closed-form equations but re-calculate the coefficients based on the newer receiver circuit design generated. As shown in [78], these equation-based models can achieve less than 2% and 5% relative error for the delay and power fitting, respectively.

we choose the method used in [78] to consider this constraint by adding an exponential item to the cost function. After optimization, we check the signal integrity and finally output the system performance metrics.

III.3 Performance prediction and comparison

Applying the design methodologies discussed in Section III.2, we perform the experiments in this section to study the performance metrics of six different global interconnect structures under the **min-d** design objective across technology nodes, from 90 nm down to 22 nm.

III.3.A Experimental settings

For parameter extraction of on-chip lossy T-lines, we use the 2D field solver CZ2D of the EIP tool suite from IBM [31] to build T-line structures shown in Figure III.6, and extract the frequency-dependent tabular model for SPICE simulation. For circuit design and modeling, we adopt a predictive transistor model [70], which is a Synopsys level3 MOSFET model with the parameters tuned following the ITRS roadmap.

For system simulation and optimization, HSPICE is used to simulate the transient response of wires, evaluate the performance of circuit and the entire interconnection structure. Linear and non-linear regression methods and SQP optimization routine implemented in MATLAB are adopted to build circuit models and perform optimization.

In our study, we set the maximum pipeline depth to 20, and choose 5 mm as the wire length (which represents typical critical length for on-chip global interconnect) to evaluate and compare the performance metrics of different structures. We also extend each experiment to different wire lengths (0.5 mm, 1 mm, 3 mm, 7 mm, 9 mm) to study the wire length crossing points for some representative

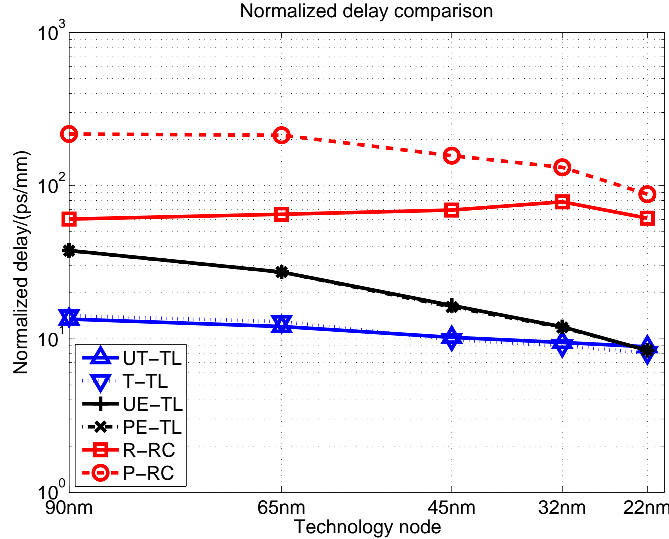


Figure III.8: The normalized delay of different global interconnection structures under min-d objective.

structure pairs in terms of different performance metrics, as shown in Subsection III.3.D.

III.3.B Latency

A comparison of the normalized delays of various global interconnect structures under the min-d objective is shown in Figure III.8. The trends of normalized latency with technology scaling verify our previous analysis in Subsection III.1.D. Normalized delay of *R-RC* structure increases due to the dominant effect of $\sqrt{\epsilon_r/\tau}$ on the total latency, whereas latency of *P-RC* decreases as the flip-flops dominate the total delay. Therefore, the latency penalty of pipelining *RC* wires is alleviated as technology scales.

On the other hand, due to the opposite scaling trend, all T-line structures outperform *R-RC* in terms of latency beyond the 90 nm node. The single-ended T-lines achieve lowest delay across all the five technology nodes. At 22 nm node, all the T-line structures show a similar delay around 8 ps/mm, whereas this number

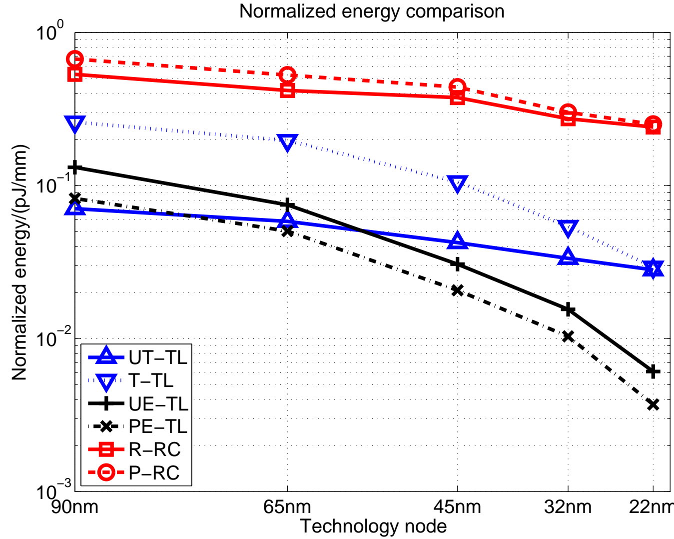


Figure III.9: The normalized energy per bit of different global interconnection structures under min-d objective.

is 60 ps/mm for *R-RC* and around 90 ps/mm for *P-RC*. Therefore, a delay reduction of at least 87% could be obtained by replacing global *RC* wires with T-line structures in this scenario.

III.3.C Other metrics

Under the min-d objective, every interconnection scheme shows a decreasing trend in energy dissipation as technology scales (Figure III.9), verifying our previous analysis in Subsection III.1.D. *RC* wires consume the largest energy among all six interconnection structures. Pipelining increases the energy of *R-RC* further due to the additional energy consumed by flip-flops, but this overhead decreases as technology scales because of the scaling of flip-flop capacitance. On the other hand, T-line structures consume less energy at each technology node. Beyond the 65 nm node, differential T-lines (*UE-TL*/*PE-TL*) consume the least energy due to power efficient SA-based receivers and the higher bit rate achieved by reducing signal swing at the wire-end. Further, the energy per bit could be

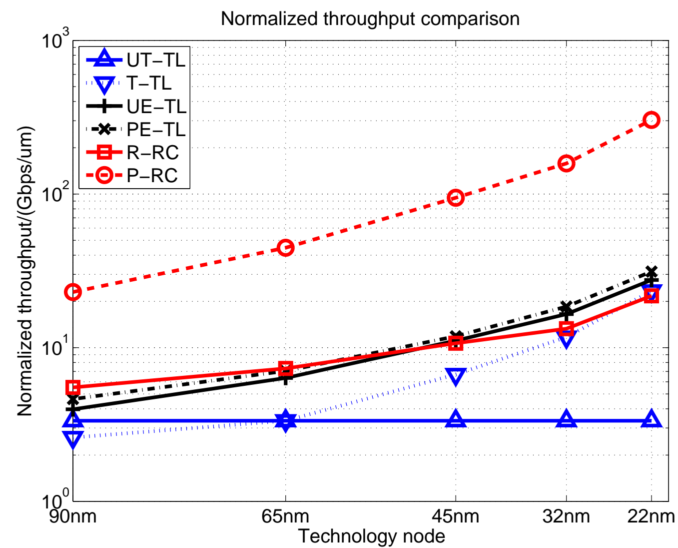


Figure III.10: The throughput density of different global interconnection structures under min-d objective.

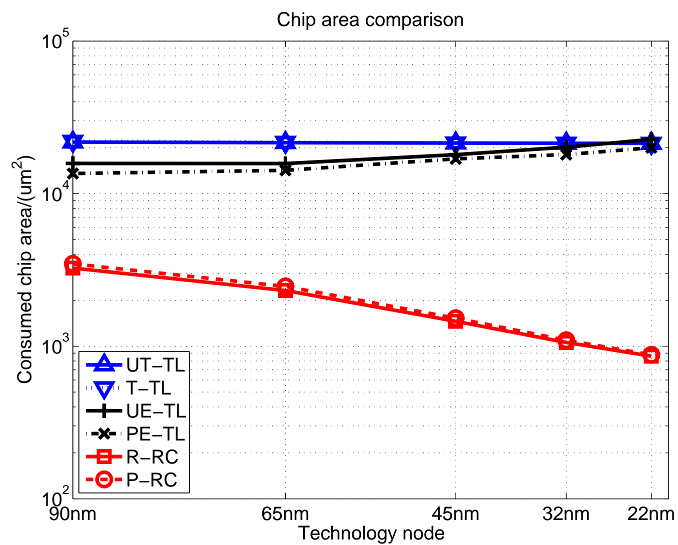


Figure III.11: Chip areas consumed by different global interconnection structures under min-d objective.

reduced by nearly 40% using a passive equalizer. At the 22 nm node, differential T-line schemes will reduce the energy per bit by two orders of magnitude compared with *RC* wires.

The throughput density of different schemes under the min-d objective is shown in Figure III.10. As discussed in Subsection III.1.D, this metric is improved for all the schemes as technology scales (except for *UT-TL*, which throughput density is constant, limited by the wire itself). *P-RC* achieves the highest throughput density across all the technologies by increasing the *R-RC* bandwidth using the smallest wire pitch. For T-line schemes, differential T-lines have larger throughput density compared with single-ended ones because of the higher achievable bit rate by utilizing SA-based receiver. Furthermore, the introduction of passive equalization makes the utilization of narrower wires possible, increasing the density even further. Beyond 45 nm node, differential T-lines will finally outperform *R-RC* in terms of throughput density.

The chip areas consumed by the various interconnect structures are compared in Figure III.11. According to the analysis performed in Subsection III.1.D, assuming wire area is dominant in total area consumption, *RC* wire area will decrease exponentially as technology scales, whereas area of other T-line schemes will remain the same (single-ended T-lines) or even increase (differential T-lines), as shown in Figure III.11.

III.3.D Critical length

A critical length study is also performed by running the optimization flow in several different wire lengths, from 0.5 mm to 9 mm. The results are summarized in Figure III.12. In this figure, a dashed line and dotted line located on the upper and lower sides indicate the upper-bound and lower-bound of wire length for on-chip global interconnects, corresponding to 10 mm and 0.5 mm, respectively. We chose eight representative interconnect structure pairs, and show the scaling trend

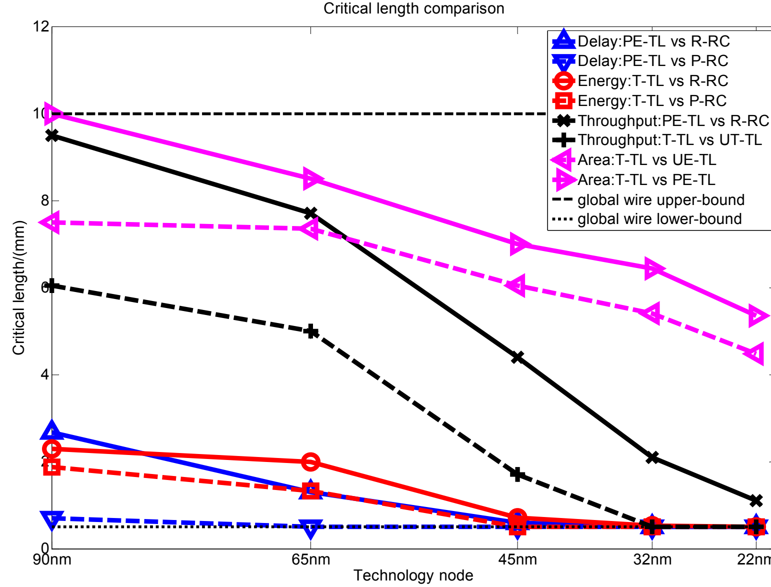


Figure III.12: Critical length of several chosen interconnect structure pairs in terms of different performance metrics under min-d objective.

of their critical lengths in terms of four different performance metrics. As an illustration, for "Delay:PE-TL vs R-RC" case, which corresponds to the solid line with upper triangle marker, the critical length at 90 nm node is about 2.5 mm, which means that when the wire length is larger than 2.5 mm at this node, PE-TL will outperform R-RC in terms of normalized delay. Based on this illustration to understand Figure III.12, we make the following general observations:

1) As technology scales (beyond the 45 nm node), T-line schemes will outperform RC wires in terms of normalized delay and energy within the entire length range of on-chip global wires;

2) In terms of throughput, at the 22 nm node, PE-TL will outperform R-RC while wire length is larger than 1 mm, and UT-TL will be replaced by T-TL within the entire length range; and

3) Single-ended T-lines will consume less chip area compared with differential counterparts for longer wire lengths. At the 22 nm node, T-TL occupies less area than PE-TL and UE-TL when the wire length is longer than 5.4 mm and

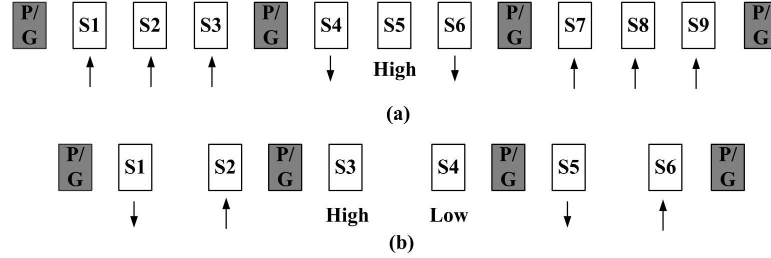


Figure III.13: The wire configurations and worst case switching patterns of T-line structures for testing crosstalk effects. (a) single-ended; (b) differential.

4.5 mm, respectively.

III.4 Signal integrity

In this section, we discuss the signal integrity issues of different interconnection structures, with the focus on the T-line schemes. Basically, we will study signal integrity by simulating the maximum crosstalk noise at the wire-end of quiet lines and the eye-height with and without crosstalk effects. For the maximum noise simulation, based on the previous work [19] and SPICE simulations, the worst case switching patterns of single-ended and differential T-lines are given in Figure III.13. In terms of eye-height simulation, HSPICE transient simulations for 500 cycle times are performed using one or several different PRBS input patterns. All the experimental results are summarized as follows.

III.4.A Single-ended T-lines

Single-Ended structures tend to be more sensitive to noise. For the un-terminated scheme ($UT\text{-}TL$), simulation shows that the maximum peak noise will be 380 mV at 45 nm node (1 V supply voltage)[8], and this situation could be more

[8] Here we follow the crosstalk simulation method in [18] and focus on the far-end noise (FEN). [18] also provides a more comprehensive study on the frequency-dependent crosstalk effects of on-chip single-ended un-terminated data bus.

Table III.5: Crosstalk effects on the *T-TL* structure

Tech Node/nm	90	65	45	32	22
Cycle Time/(ps)	90	70	40	25	15
Max Xtalk Noise/(mV)	212	188	170	137	78
Eye Height w/o Xtalk/(mV)	1002	851	769	574	405
Eye Height w/ Xtalk/(mV)	750	740	706	539	383
Supply Voltage/V	1.2	1.1	1.0	0.9	0.8

Table III.6: Maximum crosstalk peak noise (mV) of differential T-line structures

Tech Node/nm	90	65	45	32	22
UE-TL	7	8	10	11	13
PE-TL	2	6	8	10	13

severe as the technology scales since the supply voltage drops. Therefore, considering the crosstalk, full-swing signals cannot be guaranteed at the wire-end, which makes this conventional on-chip bus structure less reliable at advanced technology nodes in spite of its high-performance.

In comparison, *T-TL* provides improved noise performance as well as higher bandwidth. Since the cycle time of this structure changes as technology scales, we perform the simulation at different nodes and summarize results in Table III.5. The peak crosstalk noise reduces with technology scaling due to the reduced termination resistance and supply voltage (can be derived based on the formula presented in [71]). At the 45 nm node, the noise is only 170 mV, less than half that of *UT-TL*. Eye-heights also reduce because of the increasing bit rate. However, an eye around 380 mV could still be achieved at the 22 nm node even with the impact of crosstalk noise.

III.4.B Differential T-lines

Differential T-lines enjoy greater immunity to crosstalk due to the termination resistance and the impact of common-mode noise rejection [48]. Similar

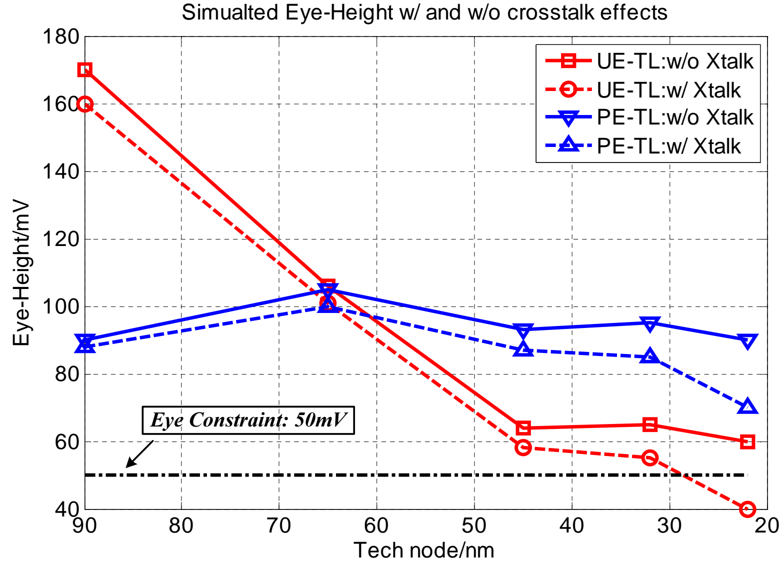


Figure III.14: The influence of crosstalk effects on the eye-height of $UE\text{-}TL$ and $PE\text{-}TL$ structures.

crosstalk peak noise simulations are performed using the switching pattern described in Figure III.13(b) for $UE\text{-}TL$ and $PE\text{-}TL$ structures, and the results are listed in Table III.6. The table shows that peak noise is far lower in differential T-lines than that of single-ended T-lines. Even with the higher inductive coupling as the bit rate increases, the peak noise in the differential T-line is only around the 10 mV range.

Eye-heights with and without crosstalk effects for two differential T-line structures are simulated and illustrated in Figure III.14. For $UE\text{-}TL$ structure, the optimal eye-height reduces as technology scales due to the increased bit rate. Considering the crosstalk, it will be harder for this scheme to meet the 50 mV eye constraint at advanced technology nodes (see the violation at 22 nm node). In comparison, by using passive equalization, $PE\text{-}TL$ can achieve larger than a 70 mV eye across all technologies even in the presence of crosstalk. Therefore, equalization improves signal integrity by boosting the eye-heights at higher bit rates.

III.5 Summary and discussion

III.5.A Discussion

In this work, **latency** was chosen to be the design objective for different interconnect schemes specifically designed for global wires (e.g., wide bus) in conventional high-end processors. To meet the increasing demand for computing capacity as process technology scales, throughput-centric interconnect design has become a hot research topic. New computing architectures have appeared, such as multi-cores and Networks-on-Chips (NoCs) [34]. New design metrics have also been proposed to balance throughput, energy, and chip area during the interconnect planning stage for different applications, as shown in [15]. Conventional repeater insertion with min-d optimization cannot satisfy the increasing bandwidth requirement for global interconnect. To enhance signal bandwidth, pipelining and other similar concepts (e.g., wave-pipelining [74]) are utilized to compensate for this performance gap. Since the purpose of our paper is to explore the potential of different interconnect options in **high-performance** applications, we did not include much optimization freedom during the preliminary study of the *P-RC* scheme. By adopting voltage scaling, buffer and wire sizing, the performance of *P-RC* scheme can be more fully studied, and the energy gap between *RC* wire and T-lines could perhaps be reduced. This would be further discussed in Chapter IV.

Some research has been done recently regarding chip-level CMOS implementation of novel global interconnect (e.g., uninterrupted *RC* wire, equalized on-chip interconnect, etc) [58] [39]. For the state-of-the-art equalized on-chip interconnect design using a 90 nm process [39], measured throughput density is similar to our prediction for differential T-line scheme (about 2 Gbps/μm), and energy per bit is about 1/2 of the passive equalized T-line scheme (about 700 fJ/b). Another possible option for global interconnect is low-swing signaling on *RC* wire, according to a recent study [52]. Although the energy dissipated using such a scheme can be

very low (similar or even lower than T-line schemes based on 0.13 μm simulation results), its latency is very large (2-4x of repeated *RC* wires). Therefore, we did not include reduced-swing *RC* signaling in this study.

III.5.B Summary

In this chapter, we compare six different global interconnect structures in terms of latency, energy per bit, throughput, chip area, and signal integrity, across technology nodes ranging from 90 nm down to 22 nm. A set of simple linear models is provided to link the architecture-friendly performance metrics of these interconnect structures with technology-defined parameters, and is verified by experimental results. A general design framework is introduced to optimize and evaluate the performance metrics of on-chip T-line interconnects.

Several observations based on the performance trends observed with technology scaling are summarized as follows: 1) T-line structures have the potential to replace *RC* wires at future technology nodes due to improved delay, energy per bit, throughput density (compared with *R-RC*), and reliability (crosstalk noise), but such schemes consume greater chip area; 2) Differential T-lines are better for high-throughput, low-power, and low-noise application compared with single-ended counterparts; and 3) Equalization approaches (such as passive equalization) can be utilized for on-chip global interconnects to improve throughput density and reduce energy dissipation.

Chapter III includes the content of one accepted journal paper, "Prediction and Comparison of High-Performance On-Chip Global Interconnection", by Y. Zhang, X. Hu, A. Deutsch, A. E. Engin, J. F. Buckwalter, C. K. Cheng, which will appear in *IEEE Transaction on VLSI Systems*. The dissertation author was the primary investigator and author of the paper.

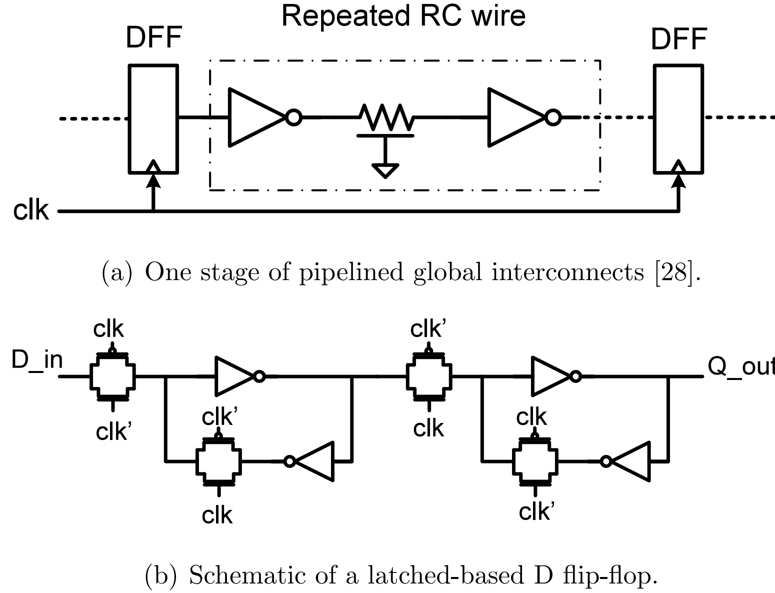
IV

Pipelined Global Interconnects with Voltage Scaling

In this chapter, we explore the performance of flip-flop-based pipelined global interconnects with more design freedoms under voltage and technology scaling for different applications in this chapter. Using the derived accurate voltage-scaled models of pipelined interconnects, we propose a general evaluation flow using numerical experiments to study how the pipelining depth, voltage scaling, and different processes affect the performance of pipelined interconnects under four different design objectives.

IV.1 Pipelined global interconnects

In this section, we firstly give an overview of pipelined on-chip global interconnects on their structure and configuration. Then variables and parameters related to such design are presented and summarized to make the following analysis clear. Finally, we talk about the performance modeling of pipelined interconnects based on the listed assumptions.



(a) One stage of pipelined global interconnects [28].

(b) Schematic of a latched-based D flip-flop.

Figure IV.1: Structure of pipelined global interconnect studied in this work.

IV.1.A Overview

Following work of Heo *et al.* [28], we adopted the pipelined global interconnect structure as shown in Fig.IV.1.[1] The long global wire is divided into multiple pipelined stages by flip-flops to meet the throughput requirement. For each stage (shown in Fig.IV.1(a)), repeaters are inserted to break the wire into several segments to reduce the wire latency. The size and interval of repeaters are optimized for given wire geometries (wire width, wire pitch) based on specific design objectives.

In terms of flip-flop, we adopt the common-used two-stage latch-based D flip-flop (shown in Fig.IV.1(b)) to implement the pipelining. To simplify the pipelined interconnect design, we assume a fix size for flip-flop (which is optimized

[1]As another pipelining method, wave-pipelined interconnects have been proposed [74] to implement the pipelining only using repeaters. However, data synchronization becomes an issue in such structures. Some special clock routing or local clock generation may be needed to latch the data in the receiver side [16], which makes the wave-pipelined not compatible with common SoC/NoC designs. To make the analysis here more general, we still choose the flip-flop based pipelining implementation in this work.

Table IV.1: Design parameters for global pipelined interconnect based on ITRS Roadmap 2008 and predictive SPICE models.

Year	2007	2010	2013	2016
Technology node (nm)	65	45	32	22
Target clock freq. (GHz)	5.06	5.88	7.34	9.18
Supply voltage (V)	1.1	1.0	0.9	0.8
Interlayer dielectric constant	2.9	2.6	2.4	2.1
Copper resistivity[1] ($\mu\Omega$·cm)	2.73	3.10	3.52	3.93
Min global pitch (nm)	210	135	96	75
Aspect ratio (A/R)	2.3	2.4	2.5	2.6
Resistance r_0 of min-repeater[2] (kΩ)	19.3	16.2	23.6	37.5
Leakage current at 100°C[2,3] (nA/nm)	0.22	0.085	0.18	0.38
Flip-flop capacitance[2] (fF)	16.4	10.2	6.94	4.78
Flip-flop delay[2,3] (ps)	90.3	63.2	58.4	57.3

[1] The copper resistivity includes scattering and barrier effect.

[2] Data are obtained by simulation using predictive models [70].

[3] Data are measured under nominal supply voltages.

for a 50x repeater loading), and use such configuration to characterize the timing and power dissipation of flip-flop in SPICE simulation as described in the following.

Generally speaking, for the pipelined global interconnects studied in this work, we manipulate the pipelining depth and scalable supply voltage on top of the optimization of wire geometries and repeater insertion to explore the best configuration (with target clock period met) in terms of different applications.

IV.1.B Glossary

To clarify the following analysis, we define all the parameters, variables, and expressions used in the pipelined interconnect study here. Most of them follow the work [76]. Overall pipelined interconnect design should meet target clock frequency (f_{clock}) for given wire length (l). Two knobs, pipelining depth (N) and supply voltage (V_{dd}), can be tuned according to different objectives. For

Table IV.2: Symbols used for variables and parameters of pipelined global interconnects.

f_{clock}	Target clock frequency [60]
l	Total wire length
N	Number of pipelined stages
V_{dd}	Supply voltage
w	Wire width
$pitch$	Wire pitch
s_{inv}	The scaled size of the repeater
l_{inv}	The repeater interval
t	Wire thickness
h	Dielectric height
ρ	The copper resistivity
r_w	Wire resistance per unit length
c_w	Wire capacitance per unit length
r_0	Output resistance of a min-sized repeater
c_{nmos}	Min-sized NMOS gate capacitance
I_{leak}	The leakage current for one min-sized repeater
η_{leak}	The ratio between leakage and dynamic power
C_{FF}	Effective capacitance of a flip-flop
d_{FF}	Delay of a flip-flop at nominal V_{dd}
g=1.34	P/N ratio of transistor width
f	The diffusion to gate capacitance ratio
a=0.4,b=0.7	Constants related to transistor switching model [4]
d_{seg}	The delay of each repeater-wire segment
e_{seg}	The energy dissipation of each repeater-wire segment

each configuration (N and V_{dd}), physical variables of interconnect (wire width w and pitch $pitch$) and repeaters (repeater size s_{inv} and interval l_{inv}) are optimized accordingly.

Physical characteristics of interconnects and performance of CMOS circuits (repeaters and flip-flops) are defined as parameters, which can be obtained from technology data or based on SPICE characterization. Wire resistance (r_w) and capacitance (c_w), which are utilized for building distributed Π model for global interconnects, are calculated using 2D closed-from equations in [65]. Some other parameters used for characterizing pipelined global interconnects are listed in Table IV.1, based on the prediction of ITRS roadmap 2008 [60] and SPICE simulation using predictive technology models [70]. Resistance (r_0) and capacitance (c_{nmos}) of repeaters, capacitance (C_{FF}) and delay (d_{FF})[2] of flip-flops, are also derived by SPICE timing and power characterizations. In this work, leakage power is considered in the whole performance evaluation by introducing a ratio (η_{leak}) of leakage power over dynamic power. Some design constants (a,b,f,g) are pre-defined, and basically follow the ones in [76].

Finally, we define d_{seg} and e_{seg} as the delay and energy dissipation of one repeater-wire segment, and will derive their expressions in the following text. All the parameters and variables discussed above are summarized and symbolized in Table IV.2.

IV.1.C Assumptions and modeling

To simplify our analysis in this work, we use the following assumptions:

- While searching optimal s_{inv}, l_{inv}, wire pitch varies from the minimum global wire pitch to the user-defined upper-bound, and wire width varies from the feature size[3] to the corresponding wire pitch.

[2] Delay of a flip-flop is defined as the sum of setup time and clock-to-q time.
[3] Feature size is defined as the metal 1 half pitch according to the ITRS roadmap.

- All the inserted repeaters or flip-flops have the same size, and number of total repeaters is even to avoid the signal inversion at the wire-end.
- Flip-flops are inserted at equal intervals into wires, and repeaters are inserted evenly between flip-flops.
- The size of flip-flop is fixed and flip-flop delay is characterized by loading an average-sized repeater.

Based on above assumptions, we derive the models of delay and power of pipelined global interconnects, as follows.

Delay and power modeling

Elmore delay model is used to derive the delay of repeated wire between flip-flops, as discussed in [36]. The wire delay for one repeater-wire segment (d_{seg}) is expressed as:

$$\begin{aligned} d_{seg} &= b(1+f)(1+g)r_0c_{nmos} + ar_wc_wl_{inv}^2 \\ &+br_0c_wl_{inv}/s_{inv} + b(1+g)r_wc_{nmos}l_{inv}s_{inv}. \end{aligned} \tag{IV.1}$$

Power consumed by repeated wires is expressed as the sum of dynamic power (including capacitance charging/dis-charging and short-circuit power) and static power. Following the approach in [76] to model each component, the energy dissipation of a repeater-wire segment (e_{seg}) becomes:

$$e_{seg} = \alpha_{sw}C_{eff}V_{dd}^2 \tag{IV.2}$$

where α_{sw} is the switching factor reflecting the data activity, and effective capacitance C_{eff} is:

$$C_{eff} = c_wl_{inv} + (1.1+\eta_{leak})(1+f)(1+g)c_{nmos}s_{inv}. \tag{IV.3}$$

Considering the pipelining effect, delay and power overhead of inserted flip-flops should be included in the performance modeling. Introducing the effective

capacitance C_{FF} and based on the previous assumptions, total delay and energy dissipation of pipelined global interconnects become:

$$d_{total} = (l/l_{inv})d_{seg} + Nd_{FF}, \tag{IV.4}$$

$$e_{total} = (l/l_{inv})e_{seg} + N\alpha_{sw}C_{FF}V_{dd}^2. \tag{IV.5}$$

As a result, throughput of such pipelined interconnects is:

$$f_{bw} = N/d_{total} = \frac{1}{(l/N)(d_{seg}/l_{inv}) + d_{FF}}. \tag{IV.6}$$

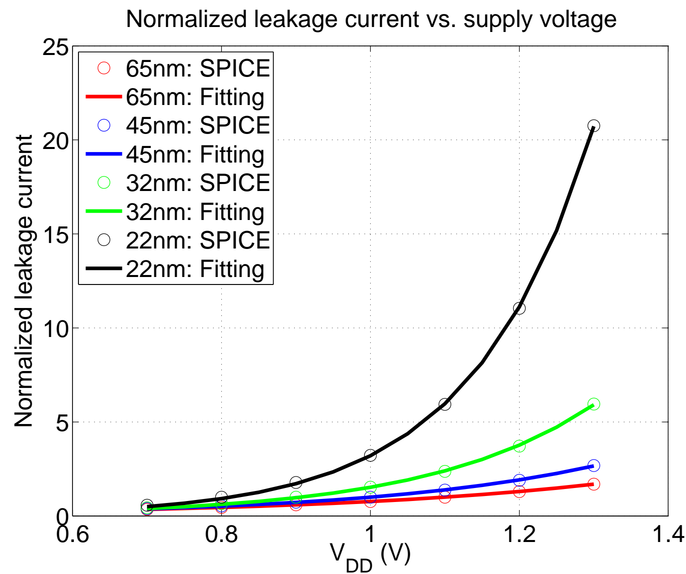
Eq.(IV.6) shows that, the throughput of pipelined interconnects can be improved by gradually adding more flip-flops to reduce the wire latency in each pipelined stage. However, more pipelined stages also bring extra delay and energy overhead based on (IV.4) and (IV.5) generally. Therefore, pipelining is not a free technique that can be used without any constraint. The good thing is, from practical point of view, required target clock frequency limits the number of pipelined stages, and by utilizing voltage scaling or other tunable knobs, the extra cost of adding flip-flops is minimized.

Voltage scaling modeling

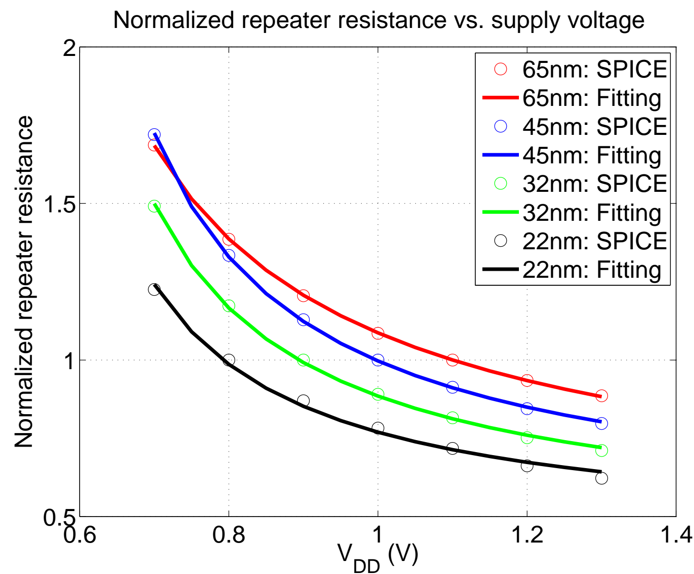
In order to incorporate the voltage scaling in the pipelined interconnect study, we model the voltage scaling as below. In terms of repeater models, output resistance r_0 and leakage factor η_{leak} need to be remodeled as a function of supply voltage V_{dd} to capture the delay and energy change due to the voltage scaling. In the voltage range [0.7 V, 1.3 V] considered in this work, we use SPICE to characterize the leakage current and output resistance of CMOS repeater, and fit the data using specific curve regressions, as shown in Fig.IV.2.

To fit the leakage current I_{leak} of a repeater, we use an exponential function [55]:

$$I_{leak}(V_{dd}) = K_1 e^{K_2 V_{dd}}. \tag{IV.7}$$



(a) Modeling leakage current with voltage scaling.



(b) Modeling repeater resistance with voltage scaling.

Figure IV.2: Voltage scaled models built by using HSPICE simulation and curve regression.

For output resistance r_0 of repeater, based on the α-power current law, following function is adopted [55]:

$$r_0(V_{dd}) = K_1 \frac{V_{dd}}{(V_{dd} - V_{th})^{K_2}} \tag{IV.8}$$

where V_{th} is the threshold voltages (around 0.4-0.5 V for predictive models). K_1,K_2 in (IV.7) and (IV.8) are the fitting coefficients, and are solved using Matlab Curve Fitting Toolbox [49]. Fitted coefficient K_2 (α in the α-power current law) in (IV.8) is around 1-1.2, which is reasonable for short-channel models. In this work, we define normalized I_{leak} and r_0 by normalizing to the values measured at the nominal V_{dd} of each process. Fig.IV.2(a) and Fig.IV.2(b) show the accuracy of the proposed models versus SPICE simulations for normalized I_{leak} and r_0. Generally, less than 4% and 2% maximum relative errors can be achieved for resistance and leakage modeling, respectively. Based on the normalized I_{leak}, normalized leakage factor (η_{leak}) can be computed as:

$$\eta^n_{leak}(V_{dd}) = \frac{\eta_{leak}(V_{dd})}{\eta_{leak}(V^{nom}_{dd})} = \frac{V^{nom}_{dd}}{V_{dd}} I^n_{leak}(V_{dd}) \tag{IV.9}$$

where η^n_{leak} and I^n_{leak} indicate the normalized η_{leak} and I_{leak}, and V^{nom}_{dd} indicates the nominal V_{dd} for given process. The following evaluation flow will take function η^n_{leak} and r^n_0 (normalized r_0) to calculate η_{leak} and r_0 at each V_{dd} point, using:

$$\eta_{leak}(V_{dd}) = \eta^n_{leak}(V_{dd}) \times \eta_{leak}(V^{nom}_{dd}), \tag{IV.10}$$

and

$$r_0(V_{dd}) = r^n_0(V_{dd}) \times r_0(V^{nom}_{dd}). \tag{IV.11}$$

The obtained results are later fed into (IV.1)-(IV.2) to evaluate the performance of repeated-wire at given V_{dd}.

For voltage-scaled modeling of flip-flops, following the same idea as modeling repeater resistance, we fit the flip-flop delay d_{FF} using (IV.8), and replace parameter d_{FF} with a scaled function $d_{FF}(V_{dd})$ in the flow after modeling.

Finally, after taking all the voltage-scaled models into account, performance of pipelined global interconnects, including $d_{total}(N, V_{dd})$, $e_{total}(N, V_{dd})$, and $f_{bw}(N, V_{dd})$ are remodeled using (IV.4)-(IV.6) and treated as functions of pipelining depth N and supply voltage V_{dd}.

IV.2 Design objectives and metrics

Different objectives are studied for the pipelined global interconnects here in order to deal with different applications, also several metrics are defined to evaluate the overall performance. We briefly discuss and summarize all the objectives and metrics used in this work as follows.

IV.2.A Design objectives

Conventional repeated global wire design tries to optimize the total latency of interconnects (like in the traditional high-end processor design), and if based on this objective to optimize pipelined global wire, we will end up with adding few flip-flops but larger repeaters and wider wire, to bring much more energy overhead. We label this design objective as **Min-Latency** in this work.

For the new design paradigm that considers throughput as the dominant performance measure of on-chip interconnects, with taking the energy and area overhead into account, several new design objectives are proposed to optimize pipelined global interconnects, following the work [15].

Max-TPE is introduced to optimize the global interconnects for low-power applications. Metric *TPE* is defined as the *throughput-per-bit-energy* of a single pipelined wire, which is maximized to reduce the total energy of parallel wires under the total throughput constraint [15].

Max-TPA is utilized for optimizing global interconnects in high performance applications, which only focus on the area reduction with satisfying the

throughput constraint. Therefore, *TPA* is defined as the *throughput-per-area* for a single pipelined wire, and in terms of area, we consider both interconnect area consumed by wires and silicon area by the repeaters/flip-flops, but in most cases, wire area will be dominant. To make this metric more physically clear, we use the *effective wire pitch* (total area over the wire length) to represent the area in TPA definition. In this scenario, maximizing TPA correlates to the throughput density increase.

To balance the throughput, energy dissipation, and also area, like in [15], metric *TPEA* is defined and optimized during the design of pipelined global interconnects for some moderate-performance/moderate-cost applications. Generally, TPEA is defined as the *throughput-per-energy-area* of a single pipelined wire. The metric is maximized to reduce the total power-delay product of a set of parallel wires [15] for a given total throughput constraint. Therefore, **Max-TPEA** is the objective that balances the tradeoffs of all the metrics we are looking into.

In this work, we will study the pipelined global interconnects under these four objectives, but with more attention focusing on the Max-TPEA design, which provides a better balanced point in the design space.

IV.2.B Performance metrics

Metrics used for evaluating pipelined global interconnects are defined as follows:

1. *Throughput.* Throughput is defined as the maximum clock frequency allowed for pipelined interconnects.

2. *Latency.* We define the normalized latency (unit: ps/mm) as

$$latency_n = \frac{\text{total latency}}{\text{wire length}} \tag{IV.12}$$

   to evaluate the delay performance of pipelined interconnects.

3. *Energy per Bit*. We define the normalized energy per bit (unit: pJ/mm) as

$$energy_n = \frac{\text{energy per bit}}{\text{wire length}} = \frac{\text{total power}}{\text{throughput} \times \text{wire length}} \quad \text{(IV.13)}$$

to evaluate the energy dissipation of pipelined interconnects.

4. *TPEA*. The TPEA (Throughput-per-energy-area, unit: Gbps/μm/pJ) is defined as

$$TPEA = \frac{\text{throughput}}{\text{energy per bit} \times \text{effective pitch}} \quad \text{(IV.14)}$$

to evaluate the overall performance of pipelined interconnects.

In the following text, if there is no specific clarification, the throughput, latency, and energy per bit metrics follow the definitions here.

IV.3 Performance evaluation flow

We develop the evaluation flow to derive the optimal pipelined interconnect designs that satisfy given target clock frequency constraint under different design objectives as discussed above. The flow can also dump out the performance metrics defined in Section IV.2.B based on the optimized designs as a reference. Due to the complexity of multi-variable optimization like this problem, we adopt the idea in [53] to build the flow. Instead of using complex nonlinear programming method, we simplify the problem by limiting the range of wire geometry (w, $pitch$) and discretize the variables by defining the minimum resolution. The general top-level algorithm (Algorithm 1) for optimizing pipelined global interconnects under given technology and objective is described below.

At the beginning of algorithm, some necessary global parameters, including the technology data, are defined for later usage. Design objective (*objective*) is defined to guide the following optimization. In the most outer for-loop, V_{dd} is swept from minimum to maximum in the predefined range. For each V_{dd} point, delay and energy parameters (η_{leak}, r_0, d_{FF}) of repeaters/flip-flops are computed

Algorithm 1 Pipelined Wire Optimization Algorithm

1: Define global and technology parameters
2: Define design *objective*
3: **for** $V_{dd} = V_{dd}^{min}$ to V_{dd}^{max} **do**
4: Compute η_{leak}, r_0, d_{FF}
5: $N \leftarrow 1$
6: **repeat**
7: **for** $pitch = pitch_{min}$ to $pitch_{max}$ **do**
8: **for** $w = w_{min}$ to $pitch$ **do**
9: Compute $r_w(pitch,w)$, $c_w(pitch,w)$
10: s_{inv}, l_{inv}= fminsearch($objective$, r_w, c_w)
11: Compute cost function f
12: **end for**
13: **end for**
14: Search minimum cost $f(N)$
15: Estimate throughput(N), delay(N), and energy(N)
16: $N \leftarrow N + 1$
17: **until** Throughput reaches the target frequency
18: **end for**
19: **return** Optimal design variables: $pitch$, w, s_{inv}, l_{inv}
performance: $f(V_{dd}, N)$, delay(V_{dd}, N), energy(V_{dd}, N)

based on scaled models derived in Section IV.1.B. To satisfy the target clock frequency with the minimum extra cost, we try to change the number of flip-flops N incrementally until the throughput constraint is met. For any given N, we build a matrix, which takes wire geometries $pitch$, w as the indices, to store the estimated performance metrics and cost function f. Repeater design parameters s_{inv}, l_{inv} are optimized based on wire geometries and design objective, by using the MATLAB built-in optimization engine $fminsearch$. After the execution of the two inner for-loops, we are able to compute the cost function f for pipelining depth N by searching the minimum (or maximum) element in the $f(pitch, w)$ matrix. With the obtained indices $(pitch, w)$, performance metrics for pipelining depth N are also derived based on corresponding matrix. If the throughput does not meet the requirement, the loop will continue by adding one flip-flop. Finally, after optimization for each V_{dd} point, the algorithm dumps out the optimal design variables $pitch$, w, s_{inv}, l_{inv}, and performance metrics in the matrix form which uses V_{dd} and N as indices. With these information, it is easy to study how the pipelining and voltage scaling affect the performance of pipelined interconnects, as shown in the following section.

IV.4 Experimental results

Applying the evaluation flow developed in Section IV.3, we perform the experiments in this section to study the performance scaling of pipelined global interconnects under different design objectives across technology nodes from 65 nm down to 22 nm.

IV.4.A Experimental settings

For the transistor models, we use the predictive technology models from Arizona State University [70], and the models are Level 54 BSIM3v3 models, which

are used broadly to predict performance scaling of CMOS circuits. HSPICE is utilized to characterize the delay and energy of repeaters and flip-flops to build the scaled models used in the optimization. We use MATLAB to do curve regression and implement the performance evaluation flow as discussed in Section IV.3.

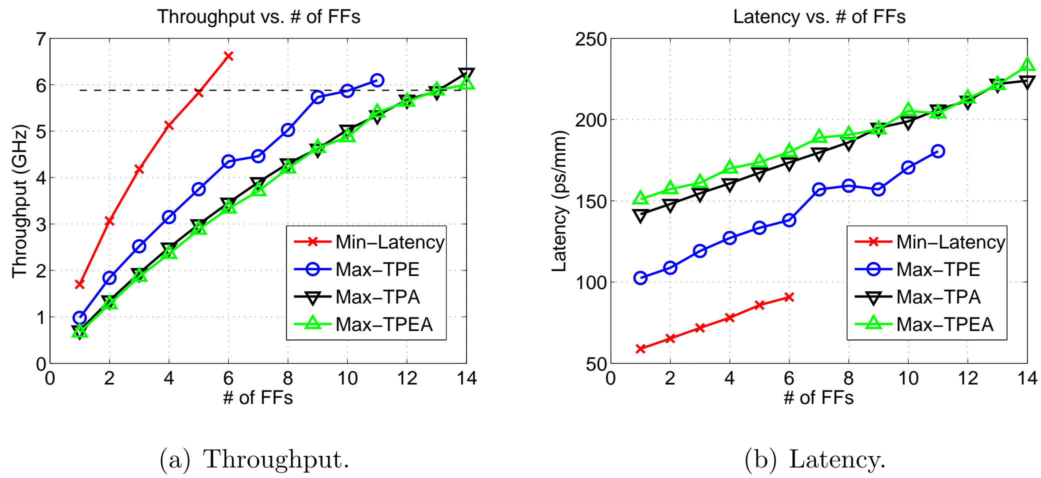
In our following study, we set the wire length l=10 mm, and switching factor α_{sw}=0.2, to represent the typical on-chip global wire scenario. During the performance evaluation, the upper limit of wire pitch is set to be 1 μm for all technologies. In terms of voltage scaling, we sweep V_{dd} from 0.7 V to 1.3 V incrementing by 50 mV. As the current technology node, we choose 45 nm process to study the impact of pipelining and voltage scaling on the performance of pipelined interconnects. We repeat the experiments for 65 nm, 32 nm, and 22 nm to discuss the technology scaling.

IV.4.B Pipelining effect

We study the impact of pipelining on the interconnect performance using 45 nm process under the nominal V_{dd} (1 V) voltage, as shown in Fig.IV.3.

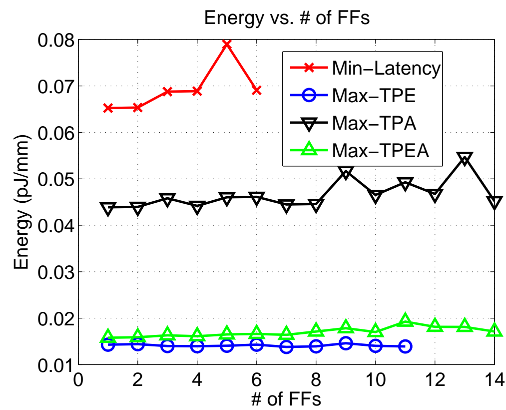
Fig.IV.3(a) shows that overall throughput is improved as pipelining goes deeper, and as described in Section IV.3, our flow stops adding flip-flops when the throughput reaches the target clock frequency, which is 5.88 GHz indicated by the dash line. Compared with Min-Latency, other objectives (Max-TPE/TPA/TPEA), which try to take area/energy into account, actually tend to use more flip-flops, to alleviate the timing slack and enforce more power/area reduction of repeated wires. As a result, Max-TPEA will need two times of number of flip-flops (14 vs. 6) compared with Min-Latency to reach the target clock frequency.

Fig.IV.3(b) and Fig.IV.3(c) show that latency and energy per bit increase with the pipelining depth N due to the delay and energy overhead brought by the flip-flops. For latency comparison, Min-Latency achieves the lowest latency compared with other objectives, because of the latency-target optimization and



(a) Throughput.

(b) Latency.



(c) Energy per bit.

Figure IV.3: Impact of the number of pipelining stages on the performance of pipelined global interconnects using 45 nm CMOS process under different design objectives.

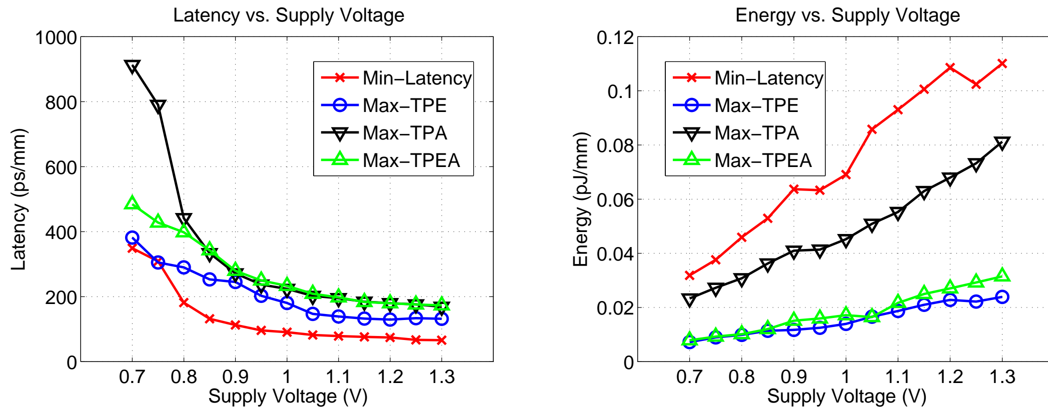
smaller N. Latency increases as objective is switched from Max-TPE, Max-TPA, to Max-TPEA. The latency overhead due to pipelining is around 50% for all the objectives. To reach the same clock frequency, 2.6x increase on the latency of Max-TPEA design is observed compared with the Min-Latency design. On the other hand, Min-Latency design does consume much more energy compared with other options (4.1x energy overhead compared with Max-TPEA design). It is noted that the energy overhead due to pipelining is very small, which is only 10% increase for all the objectives, mainly due to the dominance of wire and repeater capacitance.

IV.4.C Voltage scaling effect

Effects of voltage scaling on the pipelined global interconnects are explored in Fig.IV.4. Still, we use 45 nm process and assume pipelining depth N is chosen accordingly to satisfy the throughput constraint for different objectives.

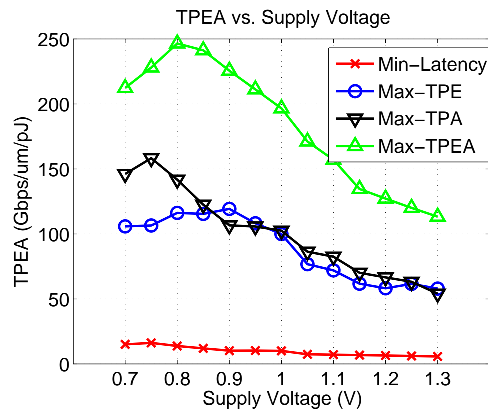
Fig.IV.4(a) and Fig.IV.4(b) show the scaling trends of latency and energy per bit within given supply voltage range. Latency decreases as supply voltage V_{dd} increases, but tends to be saturated as V_{dd} is larger than nominal value (1 V for 45 nm process). As V_{dd} drops to very low values ($\leq$ 0.8 V), the latency performance deteriorates very quickly for Min-Latency/Max-TPA designs, which exclude energy in the repeated wire optimization. This example helps to illustrate the stability of Max-TPE/TPEA designs under the voltage scaling. Also, we can verify this conjecture from the energy perspective. From Fig.IV.4(b), it is clear to see, as V_{dd} increases, the energy per bit of Min-Latency/Max-TPA designs increase much more quickly than the Max-TPE/TPEA designs. As a result, in the studied V_{dd} range, the difference of energy overhead due to voltage scaling for Max-TPEA and Min-Latency design is large (0.024 pJ/mm vs. 0.078 pJ/mm).

To evaluate how the overall performance of pipelined interconnects behaves with voltage scaling, we compare the TPEA metric for different objectives in Fig.IV.4(c). It is obvious that Max-TPEA design achieves the maximum TPEA



(a) Latency.

(b) Energy per bit.



(c) TPEA metric.

Figure IV.4: Impact of supply voltage scaling on the performance of pipelined global interconnects using 45 nm CMOS process under different design objectives.

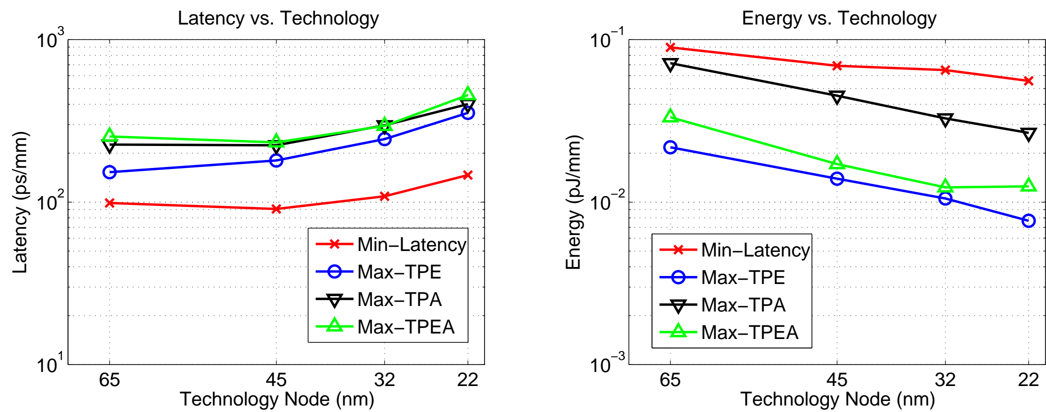
values compared with other options. TPEA of Min-Latency design is about one order of magnitude lower than that of Max-TPEA one. Furthermore, due to the throughput-energy tradeoff, there exists an optimal V_{dd} point in terms of maximum TPEA values for throughput-centric pipelined interconnect designs. For this case, the optimal V_{dd} values for Max-TPE/TPA/TPEA design are 0.9 V, 0.75 V, and 0.8 V, respectively. All these optimal V_{dd} values are below the nominal V_{dd} (1 V), showing the benefit of reducing V_{dd} to improve overall performance of pipelined global interconnects. Compared with nominal V_{dd} designs, TPEA values can be improved as much as 20%, 54%, and 25% using optimal V_{dd} for Max-TPE/TPA/TPEA, respectively.

IV.4.D Technology scaling

Performance scaling trends of pipelined global interconnects are studied using four different processes, and nominal V_{dd} with target throughput constraint is assumed in this part. Results are shown in Fig.IV.5.

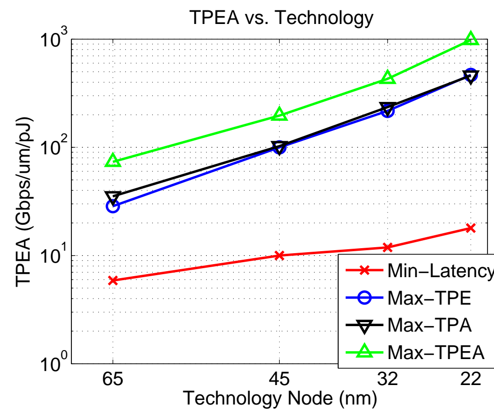
As technology scales, latency of pipelined interconnects increases approximately exponentially, around 1.2-1.4x per generation (Fig.IV.5(a)). The latency drop while scaling from 65 nm to 45 nm is mainly due to the improvement of CMOS process, that high-k/metal-gate and stress effect are incorporated in the newly released 45-22 nm PTM models [70]. Without such process breakthroughs, latency increases gradually because of the larger wire resistance, higher wire coupling, and especially deeper pipelining in the smaller feature size. Energy per bit, by contrast, decreases as technology scales, by about 30% per generation (Fig.IV.5(b)). The energy reduction is mainly due to the decrease of supply voltage and circuit gate capacitance. At 22 nm node, if choosing Max-TPEA design, latency can reach 456 ps/mm, whereas energy per bit is about 0.0125 pJ/mm.

TPEA is also improved as technology scales, as shown in Fig.IV.5(c). For throughput-centric designs, TPEA metric is improved exponentially, about



(a) Latency.

(b) Energy per bit.



(c) TPEA metric.

Figure IV.5: Impact of technology scaling on the performance of pipelined global interconnects under different design objectives.

2.4x per generation, whereas for Min-Latency design, the increasing slope is much lower, only 1.5x per generation. As a result, at 22 nm node, there will be a huge performance gap in terms of TPEA between Max-TPEA and Min-latency designs (982 Gbps/μm/pJ vs. 18 Gbps/μm/pJ). The benefit of switching to throughput-centric design becomes more promising in the future nodes.

IV.4.E Design example

Table IV.3: Performance comparison of Nominal V_{dd} Design (Min-Latency) and Voltage Scaling Design (Max-TPEA) using 45 nm CMOS process.

Design Variables and Performance Metrics	Nominal V_{dd} Design (Min-Latency)	Voltage Scaling Design (Max-TPEA)
Supply voltage (V_{dd}: V) / # of Flip-Flops (N)	1.0 / 6	0.8 / 22
Wire pitch ($pitch$: μm) / Wire width (w: μm)	0.957 / 0.735	0.222 / 0.055
Repeater size (s_{inv}) / Repeater interval (l_{inv}: mm)	260x / 0.417	26x / 0.455
Latency (ps/mm)	90.7 (1x)	397.9 (4.4x)
Energy per Bit (pJ/mm)	0.069 (1x)	0.010 (0.14x)
Throughput Density (Gbps/μm)	6.91 (1x)	24.96 (3.6x)
TPEA (Gbps/μm/pJ)	10.01 (1x)	246.52 (24.7x)

Finally, we give two 45 nm design examples of pipelined global interconnects that are optimized using the proposed flow, and compare design variables and performance metrics of those two designs. All the results are summarized in Table IV.3. To make the comparison more meaningful, we choose the Min-Latency using nominal V_{dd} as the reference design, and choose Max-TPEA with optimal V_{dd} to be compared to explore these two different design styles.

It is seen that instead of using very wide wire and strong repeaters to improve the latency in Min-Latency design, Max-TPEA tends to choose narrower wire (93% width reduction) and weak repeaters (90% size reduction) to reduce the energy and also area. In order to satisfy the throughput constraint, deeper pipelining is adopted in Max-TPEA design. Freedom of scaling V_{dd} also helps to reduce the energy further and improve the overall performance. As a result, in terms of performance metrics, Max-TPEA design does scarify some latency (4.4x increase), however, it can achieve lower energy (86% reduction), higher throughput density (3.6x increase), and finally have up to 25x improvement on the TPEA metric over the Min-Latency design.

IV.5 Summary

In this chapter, we study the performance scaling trends of pipelined global interconnects with voltage scaling as technology nodes advances from 65 nm to 22 nm. Simple but accurate voltage-scaled models are derived to capture the performance/energy degradation of repeaters/flip-flops, and used in the proposed optimization flow. Three design objectives (Max-TPE/Max-TPA/Max-TPEA) are introduced for throughput-centric pipelined global interconnect design and compared against conventional Min-Latency design. Some observations regarding the scaling of performance metrics of different objectives are summarized as follows: 1) deeper pipelining is utilized for throughput-centric designs to alleviate timing slack and reduce energy/area, 2) 20% to 50% overall performance improvement can be

obtained by reducing the nominal V_{dd} to optimal values, and 3) Max-TPEA design with voltage scaling can improve performance of Min-Latency design by 25x with only 4x latency overhead, and also shows more promising trends as technology scales.

Chapter IV includes the content of one published conference paper, "Performance Prediction of Throughput-Centric Pipelined Global Interconnects with Voltage Scaling", by Y. Zhang, J. F. Buckwalter, C. K. Cheng, in Proceedings of *2010 IEEE International Workshop on System Level Interconnect Prediction.* The dissertation author was the primary investigator and author of the paper.

V

Energy-Efficient Equalized Global Interconnects

In this chapter, we propose an equalized global link architecture for ultra-high-speed and low-energy on-chip communication by utilizing active continuous-time linear equalizer (CTLE). Accurate modeling of the CTLE is introduced to greatly improve the correlation between eye-opening prediction and SPICE simulation. Also, the proposed global link is analyzed using a linear system method, and the formula of CTLE eye-opening is derived to provide high-level design guidelines and insights. To reduce the energy-per-bit of the proposed link, a driver-receiver co-design flow is introduced by adopting Sequential Quadratic Programming (SQP) non-linear optimization and applied for design space exploration.

V.1 Equalized on-chip global link

Figure V.1 shows the overall structure of equalized on-chip global link proposed in this work. The whole system is composed of a chain of tapered CML buffers as driver, differential on-chip T-line with terminated resistance, CTLE and sense-amplifier based latch as receiver. The basic working principle is introduced

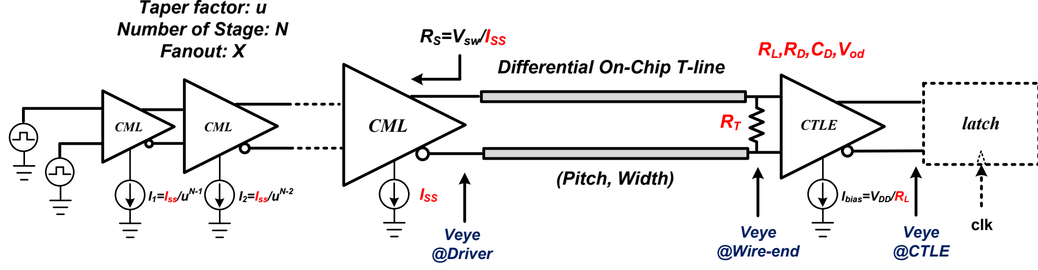


Figure V.1: The overall structure of equalized on-chip global link studied in this work.

as follows.

The transmitted high-speed digital signals (such as random bit patterns) first go through the chain of tapered CML buffers, to convert to low-swing differential signals to drive the following on-chip T-line. Similar to the delay optimization of CMOS inverter or buffer chain [4], tapered factor u and number of stages N can be decided based on the total fan-out X accordingly [29], which will be discussed later in Section V.2. For given specific driver output swing V_{sw}, bias current I_{SS} of final CML stage, can be optimized to trade-off the driver power consumption and eye-opening at wire's end. In this work, we treat driver swing V_{sw} as a design parameter of equalized global link, and define bias current I_{SS} as one of the design variables that can be optimized in the overall flow.

In terms of on-chip global wire, we model it as on-chip T-line by building uninterrupted differential wire surrounded by power and ground shielding on top of reference ground plane, which could be a high-density lower-level metal layer. The 2-D EM Field solver [31] and a synthesized compact circuit model [40] is adopted to model and simulate the transient response of such on-chip T-line structure. Geometries ($pitch$, $width$) of T-line are design parameters which can be tuned to adjust the characteristic impedance Z_0 and wire DC resistance to trade-off the signal attenuation with the wire area. We also add termination resistance R_T at the far-end of T-line to help improve the eye-quality after T-line [83]. The value of R_T is a design variable to be optimized in the flow.

At the receiver-side, one stage of continuous-time linear equalizer (CTLE) is used to recover the transmitted signal by boosting the eye-opening. CTLE parameters, including load resistance R_L, source degeneration resistance R_D and capacitance C_D, and over-drive voltage V_{od}, are optimized to improve the received eye quality as well as reduce receiver power consumption. To convert received signals back to digital level, a dedicated synchronous sense-amplifier based latch (SA-latch) is added after CTLE. In this work, we assume SA-latch is a pre-designed macro[1] using 45 nm predictive CMOS technology [70] which can recover the minimum input eye-opening V_{min}=50 mV to 1 V digital level at 20 Gbps data rate[2].

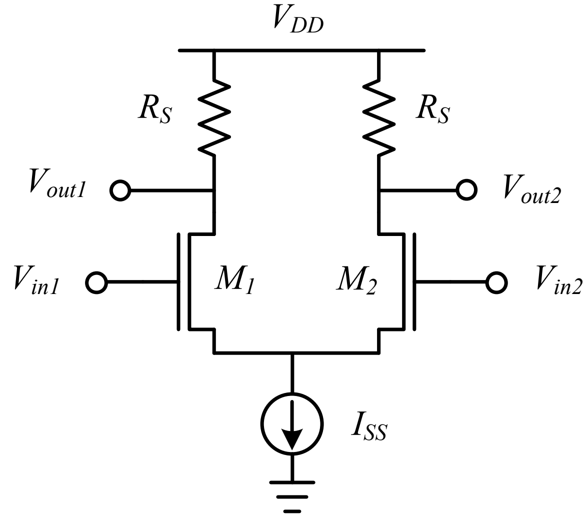
In the following sections, each building block is discussed in detail and the driver-receiver co-design methodology is also introduced to determine the best set of design variables [I_{SS},R_T,R_L,R_D,C_D,V_{od}] that achieves the lowest energy-per-bit for the proposed equalized global link.

V.2 Driver design for on-chip transmission-line

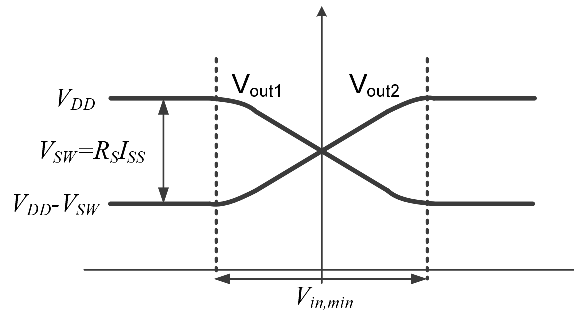
Tapered CML buffers are used as on-chip T-line driver in this work for achieving high-speed signal communication. The basic working principle has been introduced in previous works [69] [29] and here we summarize the design guideline for such driver configuration and then present a driver design example which combines CML driver design with T-line geometry optimization by considering the signal integrity after T-line.

[1]We adopt the sense amplifier design introduced in [57] and convert it to SA-latch by adding SR-latch at the output. SPICE simulation indicates that power dissipation of SA-latch slightly increases as input eye-opening decreases (5% increases as eye-opening varies from 250 mV to 50 mV), therefore we assume the power dissipation of SA-latch is constant value and do not consider it in the link power optimization.

[2]Based on the SPICE simulation, SA performance degrades (larger delay and slower output slew-rate) significantly when input signal amplitude is smaller than 50 mV for the design used in this work. As a result, we set the required minimum input eye-opening V_{min}=50 mV.



(a) Circuit schematic.



(b) DC transfer characteristic.

Figure V.2: The CML buffer schematic and DC transfer characteristic.

V.2.A Design guideline for tapered CML chain

The schematic and DC transfer characteristic of a CML buffer is shown in Fig.V.2. The design parameters of such CML buffer include load resistance R_S, size of input transistor W, and tail current I_{SS}. The output swing of such CML buffer is R_SI_{SS}. To guarantee the transistors operated in the saturation region,

$$V_{sw} = R_S I_{SS} \leq V_{th} \tag{V.1}$$

where V_{th} is the threshold voltage of CMOS transistor.

For given driver power budget I_{SS} or output swing requirement V_{sw}, driver resistance R_S can be optimized accordingly with T-line termination resistance R_T for the best eye-quality at the wire-end. Conventional chip-to-chip interconnect design methodology chooses driver resistance equal to characteristic impedance Z_0 of T-line to minimize the reflection. However, for on-chip T-line, we have the freedom to optimize R_S for better power consumption or signal integrity due to the large resistive attenuation of on-chip wires. As long as R_S and I_{SS} are determined, in order to make sure the output swing is larger than the minimum input voltage [69] such that CML buffer can be cascaded into a chain, size of CML input transistor becomes (assuming long-channel model),

$$W \geq \frac{2LI_{SS}}{\mu C_{ox} {V_{sw}}^2} \tag{V.2}$$

where L is the channel length, μ is the mobility, and C_{ox} is the per-unit-area oxide capacitance.

In terms of tapered CML chain design, after designing the final stage, total fan-out X can be calculated assuming the first-stage size is fixed. According to [29], to minimize the CML chain delay, taper factor u should be the base of natural logarithm e, which is around 2.7. Therefore, the number of stages becomes,

$$N = \lceil ln(X) + 1 \rceil. \tag{V.3}$$

Finally, each CML stage in the chain can be designed backward by scaling all the circuit parameters of next stage with the factor u.

V.2.B Driver design example

Adopting the methodology described above, we design the CML driver with T-line geometry optimization together. The T-line configuration used in this work is shown in Fig.V.3. We assume 45 nm 1P11M CMOS process[3] is used

[3]Predictive Transistor Models from [70] are used for transistor-level SPICE simulation.

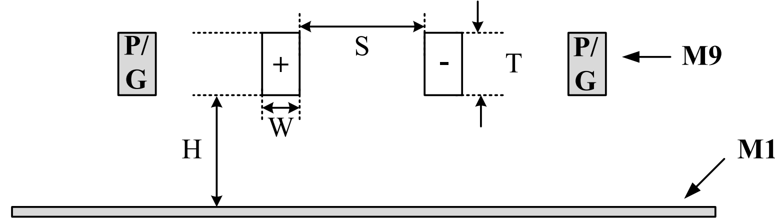


Figure V.3: The cross-section of differential on-chip T-line.

and differential T-line is built at Metal 9 where the power or ground shielding is implemented on the same layer. Metal thickness is 1.2 μm and height to the reference ground plane (assuming Metal 1 in this work) is 3.5 μm[4]. The *width* and *spacing* of T-line can be adjusted as studied in the following. We use EIP [31] to perform 2D EM extraction and adopt synthesized model [40] to simulate transient response in HSPICE. The algorithm in [62] is used to estimate the worst-case eye-opening at the wire-end based on simulated T-line step response. For this work, we target 20 Gbps global signaling over 10 mm on-chip T-line. To optimize the CML driver with T-line termination for the best eye-opening, we implemented a Sequential-Quadratic Programming [8] (SQP)-based non-linear optimization flow in MATLAB and apply it to find the best set of [R_S,R_T] in terms of highest wire-end eye-opening for given T-line geometries.

First we study how to choose the *width* and *spacing* of on-chip T-line for fixed *pitch* (=*width*+*spacing*). The results are summarized in Table V.1. For the fixed T-line *pitch*, it is seen that the optimal wire-end eye-opening is achieved when *width* and *spacing* becomes equal, which is the result of reducing T-line attenuation by balancing wire DC resistance and T-line characteristic impedance Z_0. Also, the optimal driver resistance R_S is slightly smaller than Z_0 for increasing

[4]Reducing the height to ground plane will increase capacitance and decrease inductance for T-line. As a result, signal loss becomes worse but the impact is not significant. SPICE simulation shows that, by reducing height of T-line from 3.5 μm to 1.2 μm, CTLE eye-opening loss is around 15 mV, which is equivalent to 15% power increase for the same driver-receiver design and eye-opening constraint. In real implementation, higher-level metal layer can be chosen as ground plane to save routing resources, but all the following analysis and design space exploration are still valid and can provide similar guidelines for smaller height scenarios.

Table V.1: Impact of T-line width/spacing on the received eye quality for a given I_{SS}=6 mA. Width + Spacing=2.0 μm

width (μm)	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3
Z_0 (Ω)	63.7	60.1	56.6	53.1	49.6	46.1	42.5	38.7
R_S (Ω)	42.5	44.3	44.0	42.5	40.3	37.3	33.8	22.5
R_T (Ω)	84.3	87.6	87.8	85.2	80.7	74.8	67.7	69.7
V_{eye} (mV)	10.9	14.9	17.5	19.0	19.3	18.4	16.6	13.8

Table V.2: Impact of T-line pitch on the received eye quality for a given I_{SS}=6 mA. Width=Spacing=1/2*pitch

pitch (μm)	1.0	1.2	1.4	1.6	1.8	2.0	2.2	2.4
Z_0 (Ω)	34.9	38.8	42.2	45.0	47.5	49.6	51.5	53.0
R_S (Ω)	N/A	N/A	14.7	24.5	32.6	40.3	47.0	53.1
R_T (Ω)	N/A	N/A	29.2	49.6	65.6	80.7	94.1	105
V_{eye} (mV)	closed	closed	1.0	4.6	10.8	19.3	29.0	39.5

the incident wave amplitude to boost the eye-opening [17].

By assuming the equal *width-spacing* T-line configuration, we sweep the *pitch* and observe the eye-quality change in Table V.2. It is shown that eye-quality is improved as T-line *pitch* increases because of larger Z_0. However, larger *pitch* introduces more area overhead as well as reduces the throughput density. Therefore, the minimum *pitch* which satisfies the wire-end eye-opening requirement (decided by noise/crosstalk or other possible signal-integrity considerations) should be chosen during the design of CML driver and loaded T-line.

Based on the above observations, a driver design example is derived to satisfy the constraint that T-line eye-opening is larger than 20 mV. The design parameters and SPICE simulated performance data are summarized in Table V.3. It is noted that the CML buffer chain shows shorter delay compared with conventional CMOS buffer chain, but tends to consume more power because of the static tail current. We should consider reducing CML driver power dissipation from the system-level and optimizing the driver design with CTLE receiver together to

Table V.3: A 20 Gbps CML driver design example for 10 mm on-chip T-line.

Design parameters	Simulated performance
$pitch = 2.2\ \mu$m	Driver delay: 30 ps
$R_S = 47\ \Omega\ R_T = 94\ \Omega$	T-line delay: 80 ps (8 ps/mm)
$I_{SS} = 6$ mA, $V_{swing} = 282$ mV	Driver power: 6.5 mA
final stage $W/L = 22.5\ \mu$m/45nm	Eye-opening at driver output: 232 mV
$u = 2.5$, $X = 100$, $N = 6$.	Eye-opening at wire-end: 22 mV

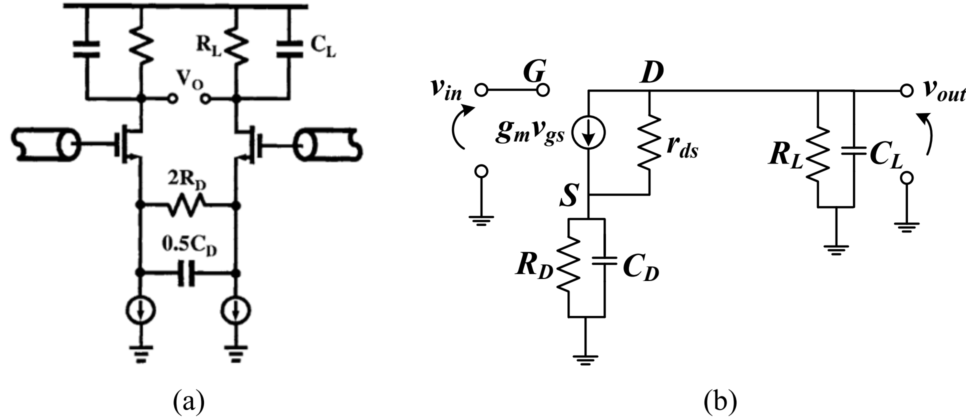


Figure V.4: The schematic (a) [23] and equivalent small-signal circuit (b) for Continuous-Time Linear Equalizer (CTLE).

reduce the power overhead.

V.3 Continuous-time linear equalizer design

The modeling, design, and optimization of CTLE are discussed in this section and one CTLE design example based on pre-designed CML driver and T-line is also shown to demonstrate the effectiveness of equalization approach for on-chip global link.

V.3.A CTLE modeling

Figure V.4 shows the schematic and equivalent small-signal circuit of CTLE studied in this work. Different from previous models [7], we try to consider

transistor output resistance r_{ds} as well as the parasitic capacitance C_D^{para} and C_S^{para}[5] to reduce the modeling error and get better correlation with SPICE simulation results. To calculate the small-signal parameters,

$$g_m = \alpha I_{bias}/V_{od}, r_{ds} = 1/(\lambda I_{bias}) \tag{V.4}$$

where I_{bias}=$V_{dd}/(2R_L)$ is the bias current for one branch of the CTLE. Coefficient α and λ are functions of over-drive voltage V_{od}. We perform the SPICE simulations and curve fitting to calculate these coefficients. In terms of parasitic modeling, we set C_D^{para}=C_S^{para}=1.5 fF/μm, and calculate the total effective capacitance by adding the external and parasitic one together.

Based on the small-signal analysis, the transfer function of CTLE becomes,

$$H(s) = G_{DC}\frac{1 + sR_DC_D}{1 + as + bs^2} \tag{V.5}$$

where

$$\begin{aligned} G_{DC} &= \frac{g_m r_{ds} R_L}{(g_m r_{ds}+1)R_D + r_{ds} + R_L}, \\ a &= \frac{(r_{ds}(R_LC_L + R_DC_D) + (g_m r_{ds}+1)R_DR_LC_L + R_DC_DR_L)}{(g_m r_{ds}+1)R_D + r_{ds} + R_L}, \\ b &= \frac{r_{ds}R_DC_DR_LC_L}{(g_m r_{ds}+1)R_D + r_{ds} + R_L}. \end{aligned} \tag{V.6}$$

To verify accuracy of the proposed CTLE modeling approach, we perform CTLE optimization on a test case (10 mm T-line, 20 Gpbs signal, 16 mV eye-opening at the wire-end) using the similar SQP-based flow as for driver design. CTLE parameters $[R_L, R_D, C_D, V_{od}]$ are optimized during the flow for the best eye-opening at CTLE output. To study the relation of eye-opening and CTLE power consumption, we set different power constraints and re-run the experiments. The results are visualized in Fig.V.5. It can be seen that eye-opening increases with relaxed power constraints, but tends to become saturated afterwards, which means

[5] C_D^{para} and C_S^{para} indicate the parasitic capacitances on the drain and source terminal of CMOS transistor, respectively.

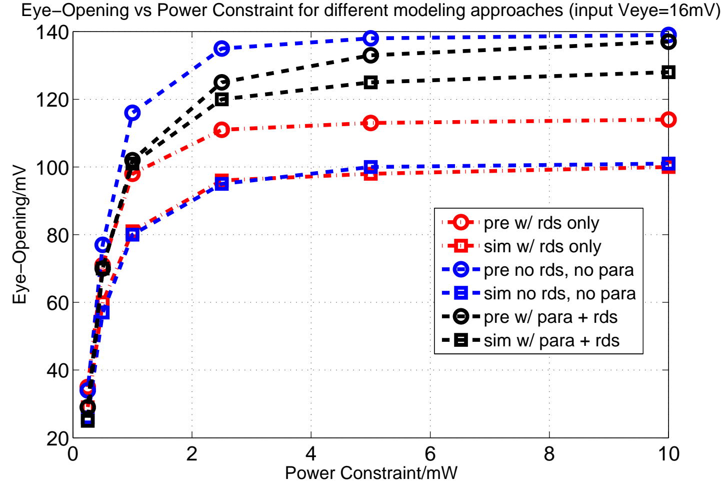


Figure V.5: The predicted and simulated eye-opening vs. CTLE power consumption for different modeling approaches.

higher power consumption may not always help to improve the eye-quality[6]. In terms of different modeling approaches, previous simple modeling (not considering r_{ds} and parasitic, blue lines) introduces the largest errors between the prediction (using algorithm in [62]) and HSPICE simulation results. Considering r_{ds} (red lines) can improve the correlation. With parasitic effects added (black lines), the modeling errors are greatly reduced. In the above test case, the relative error for eye-opening prediction can be reduced to around 5% using the proposed accurate CTLE model. More importantly, better optimization results (higher CTLE eye-opening) can be generated because of the tighter correlation between prediction and SPICE simulation.

V.3.B CTLE design example

Applying the proposed CTLE model with similar SQP-based optimization flow, we design and optimize CTLE by using pre-designed CML driver and T-line in Section V.2.B as driver stage. The cost function is to optimize eye-opening at

[6]This phenomenon will be explained later in Section V.4 after CTLE eye-opening formula is derived.

Table V.4: A 20 Gbps CTLE design example for 10 mm on-chip T-line.

Design parameters	Simulated performance
$R_L = 440\ \Omega$, $R_D = 110\ \Omega$, $C_D = 680$ fF	Receiver eye-opening: 87 mV
$I_{bias} = 1.14$ mA, W/L = 13.7 μm/45nm	Receiver power: 1.6 mA

the CTLE output. The CTLE design parameters and performance metrics are summarized in Table V.4. It is seen that after CTLE stage eye-opening is boosted to 87 mV compared with 20 mV at the wire-end. Therefore SA-latch can be added to capture the signal and convert it back to digital level.

V.4 Equalized global link analysis

Before introducing the driver-receiver co-optimization flow and showing the optimization results, we analyze the proposed equalized global link by simplifying the problem and applying linear system analysis approach. The derivations shown below also shed light on the results of design space exploration performed in Section V.5.B and provide design insights or guidelines to help designers understand the trade-offs of equalized global link design.

V.4.A CTLE eye-opening analysis

We first derive the expressions of eye-opening at the wire-end and CTLE output. To simplify the analysis, we treat on-chip global wire as *one-pole dominant linear system*, which means that the inductance effect (T-line effect) of global wire is neglected in the following analysis. This assumption is approximately valid for on-chip narrow wires because signal attenuation is very large due to the limited dimensions [67]. As a result, the step response of studied on-chip wire will have slower RC-type rise edge which can be simply modeled by a single time-constant in the exponential function [17].

Based on the above assumption, the transfer function of on-chip inter-

connect becomes,

$$H_I(s) = G_I \frac{1}{1 + s/p_I} \tag{V.7}$$

where G_I is the DC gain of on-chip interconnect and p_I indicates the assumed dominant pole. By applying inverse-*Laplace* transformation, we can derive the wire step-response and the eye-opening at the wire-end can be further written as,

$$V_{eye}^{wire} = G_I \left(1 - 2e^{\frac{-T_C}{\tau_I}}\right) \tag{V.8}$$

where T_C is the clock period (time interval of one bit data), and τ_I indicates the time constant of interconnect, which is the time that interconnect step-response reaches $0.63G_I$.

Firstly we consider CTLE as *1-zero, 1-pole linear system* (ignoring the 2nd non-dominant pole for now), and neglect the transistor output resistance r_{ds} and parasitics in the following analysis. Based on Eq.(V.5), eye-opening after CTLE can be derived as,

$$V_{eye}^{CTLE} = G_I G_{DC} \left(1 - 2e^{\frac{-T_C}{\tau_I}\frac{G_{HF}}{G_{DC}}}\right) \tag{V.9}$$

where G_{HF}=$g_m R_L$ is the high-frequency gain of CTLE, and G_{DC}=$g_m R_L/(1 + g_m R_D)$ is the DC gain of CTLE. Since G_{HF}/G_{DC} >1, Eq. (V.9) shows that CTLE boosts the signal eye-opening by reducing the time-constant of step-response.

Typically high-frequency gain of CTLE is limited by either power dissipation or circuit area. Suppose for certain G_{HF}, we want to optimize CTLE for given interconnect (τ_I) and data rate (T_C) that achieves the highest eye-opening. As a result, the optimal pole/zero ratio of CTLE can be derived as

$$\left.\frac{p}{z}\right|_{opt} = \left.\frac{G_{HF}}{G_{DC}}\right|_{opt} = 1.68\frac{\tau_I}{T_C}. \tag{V.10}$$

Based on Eq.(V.10), for more lossy on-chip interconnects (narrower or thinner wire) and higher data rate, pole of CTLE needs to be designed further away from the zero in order to boost the signal eye-opening higher. Plugging Eq.(V.10) back to Eq.(V.9), the highest achievable eye-opening of CTLE becomes

$$\left.V_{eye}^{CTLE}\right|_{opt} = 0.63 G_I G_{DC} = 0.1875\alpha \left(\frac{T_C}{\tau_I}\frac{V_{DD}}{v_{od}}\right) G_I. \tag{V.11}$$

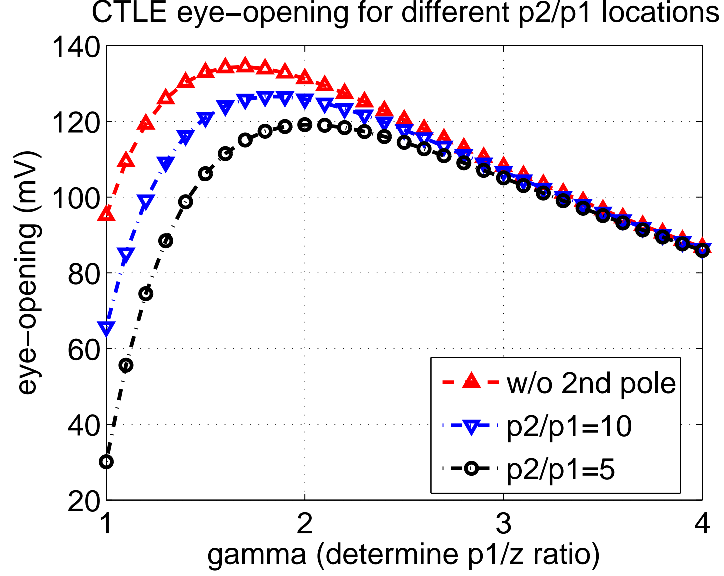


Figure V.6: The CTLE eye-opening for different p_2/p_1 ratios. The following parameters are assumed to generate the figure: α=1.2, T_C=50 ps, τ_I=250 ps, V_{DD}=1 V, v_{od}=100 mV, G_I=0.3 V.

Based on Eq.(V.11), the highest CTLE eye-opening for given interconnect and data rate is only affected by the over-drive voltage v_{od}, which is independent of CTLE power dissipation. In other words, the optimal eye-opening will not be improved by increasing the CTLE power dissipation continuously. This explains why eye-opening curve becomes saturated in Fig.V.5. On the other hand, Fig.V.5 also shows that CTLE eye-opening is improved with its power dissipation when the value is small, which is due to the effect of 2nd non-dominant pole of CTLE as discussed below.

In order to consider 2nd pole in following CTLE analysis, we assume,

$$\frac{p_2}{p_1} = \frac{R_D C_D/(1+g_m R_D)}{R_L C_L} > 1 \tag{V.12}$$

where p_1, p_2 indicate the $1^{st}, 2^{nd}$ pole of CTLE. Considering the 2nd pole in CTLE transfer function, we re-derive the CTLE eye-opening expression as follows,

$$V_{eye}^{CTLE} = \frac{K_2}{\gamma}\left(1 - \frac{2K_1}{K_1-\gamma}e^{-\gamma} + \frac{2\gamma}{K_1-\gamma}e^{-K_1}\right) \tag{V.13}$$

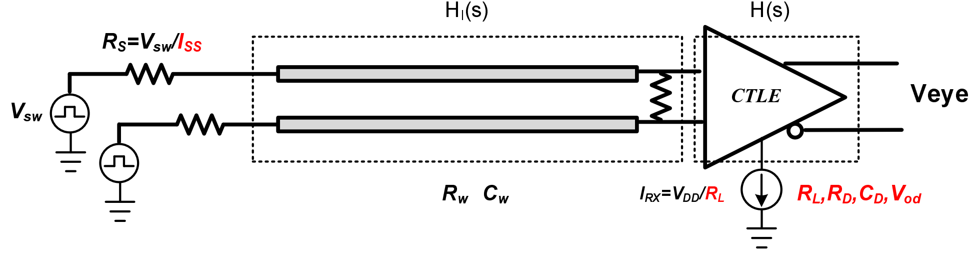


Figure V.7: The global link model used for system-level analysis.

where

$$K_1 = \frac{T_C}{R_L C_L}, \tag{V.14}$$

$$K_2 = \frac{G_I g_m R_L T_C}{\tau_I} = 0.5\alpha \frac{T_C}{\tau_I} \frac{V_{DD}}{v_{od}} G_I, \tag{V.15}$$

$$\gamma = \frac{G_{HF}}{G_{DC}} \frac{T_C}{\tau_I} = (1 + g_m R_D) \frac{T_C}{\tau_I}. \tag{V.16}$$

Based on Eq.(V.13)-(V.16), if 2nd pole is larger enough to be neglected ($K_1 \gg 1$), Eq.(V.13) will convert back to Eq.(V.9). In this scenario, the optimal γ=1.68 as derived in Eq.(V.10). On the other hand, when 2nd pole is approaching 1st pole (increasing load capacitance or decreasing CTLE bias current), parameter K_1 will kick in to affect CTLE eye-opening, which explains why eye-opening drops as CTLE current decreases in Fig.V.5. The optimal γ also becomes a function of 2nd pole location, and will increase slightly as the 2nd pole approaches 1st pole. To illustrate this 2nd pole effect, we plot CTLE eye-opening for different p_2/p_1 ratios in Fig.V.6. In this studied case, the optimal eye-opening will decrease by 11% if p_2/p_1 ratio decreases to 5. Accordingly, the optimal γ will increase slightly from 1.7 to 2.0, which means the 1st pole needs to be designed further away from the zero to compensate the performance degradation due to 2nd pole.

V.4.B System-level analysis

By modeling CML driver as an ideal voltage source that generates V_{sw} voltage with output resistance R_S as shown in Fig.V.7, we can re-write interconnect

DC gain and time constant as below,

$$G_I = \frac{R_T}{R_T + 2R_w + 2R_S} 2V_{sw}, \tag{V.17}$$

$$\tau_I = \left(\frac{1}{2}R_w + R_S\right) C_w \tag{V.18}$$

where R_w, C_w are total resistance and capacitance for on-chip interconnect. Plugging Eq.(V.17)-(V.18) back to Eq.(V.13) and assuming γ is given (such as constant value between 1.7 and 2.0), we can write the optimal eye-opening of global link as a function of driver and receiver current (I_{TX}, I_{RX}),

$$\left. V_{eye}^{CTLE} \right|_{opt} = f\left(K_1, K_2\right) = f\left(I_{TX}, I_{RX}\right) \tag{V.19}$$

where

$$K_1 = K_1\left(I_{RX}\right) = \frac{T_C}{V_{DD}C_L} I_{RX}, \tag{V.20}$$

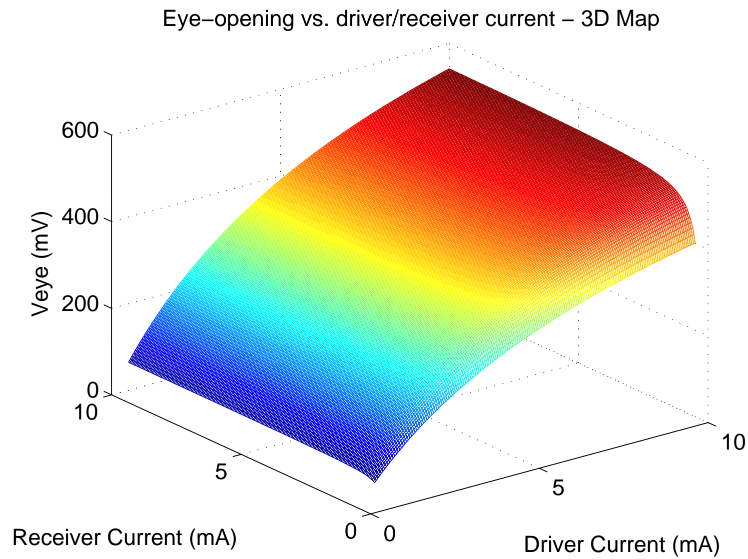
$$K_2 = K_2\left(I_{TX}\right) \simeq \alpha V_{sw} \left(\frac{V_{DD}}{v_{od}}\right) \frac{T_C}{\left(\frac{1}{2}R_w + 1.58\frac{V_{sw}}{I_{TX}}\right) C_w}. \tag{V.21}$$

Based on Eq.(V.19)-(V.21)[7], it is noted that[8]: 1) Global link eye-opening increases as driver or receiver current increases; 2) Eye-opening will become saturated as driver or receiver current is larger enough; 3) How to distribute the total current between driver and receiver needs to be investigated in order to reduce the total power of global link.

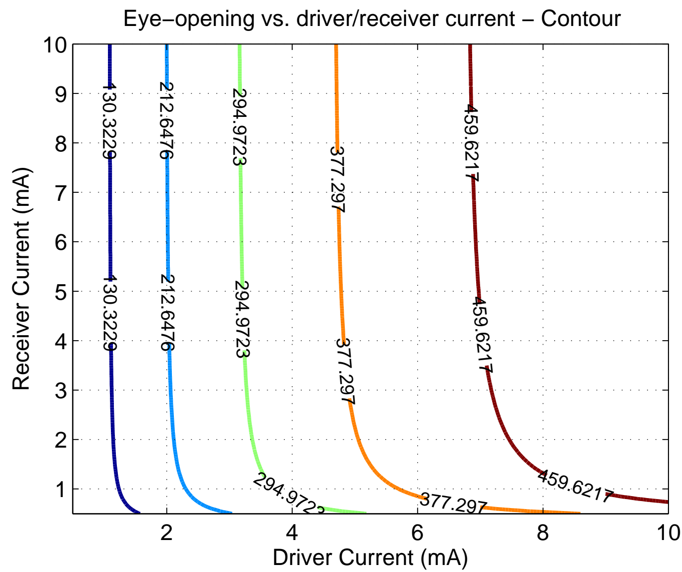
To study how to design driver and receiver current to reduce the total power dissipation for given global link eye-opening constraint, we plot the 3D map (Fig.V.8(a)) and 2D contour (Fig.V.8(b)) of eye-opening using the proposed analytical model for a test case in Fig.V.8. The 3D map shows that eye-opening increases as driver or receiver current increases, but becomes saturated quickly with

[7]To simplify Eq.(V.21), we assume T-line termination resistance R_T is larger enough to let the interconnect DC gain approximately equal to driver swing V_{sw}, which is often the case in order to boost the CTLE eye-opening.

[8]The following observations are valid when function f in Eq.(V.19) is increasing monotonically with variables K_1, K_2, which is easy to be proved by calculating the partial derivatives.



(a) 3D map for eye-opening of global link.



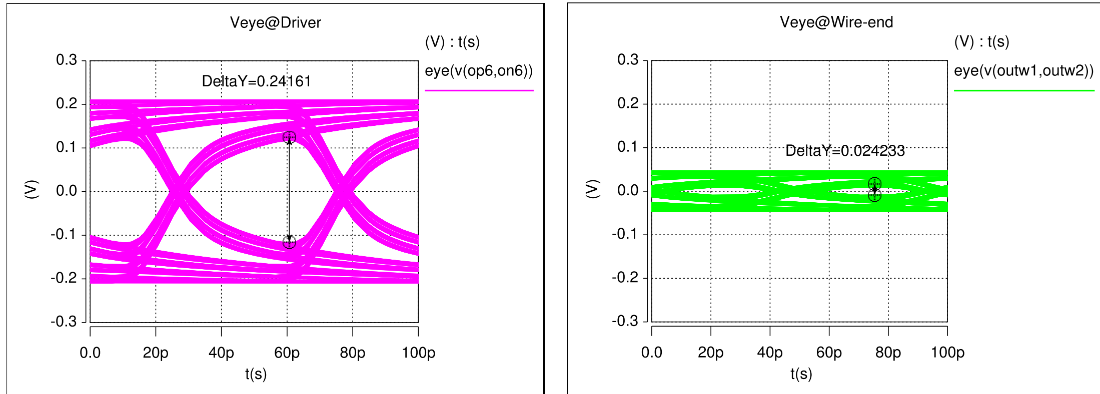
(b) 2D contour for eye-opening of global link.

Figure V.8: 3D map and 2D contour of global link eye-opening (after CTLE) based on derived analytical model. This figure is generated using following parameter values: T_C=50 ps, V_{DD}=1 V, C_L=5 fF, R_L=1 kΩ, V_{sw}=300 mV, R_w=150 Ω, C_w=1 pF, α=1.2, v_{od}=100 mV.

the receiver current, verifying previous observations. For given link eye-opening constraint, contour map can be used to determine the best current distribution for driver and receiver by finding the 45 degree tangent point for certain contour line, which is corresponding to certain constant eye-opening. For example, base on the case shown in Fig.V.8(b), to satisfy 130 mV eye constraint, driver and receiver current can be chosen as 1.35 mA and 0.7 mA respectively (tangent point located at the left-bottom corner of contour map), corresponding to the lowest total power dissipation 2.05 mA. The contour map proposed here is very useful for the designers to estimate global link eye-opening at the early stage, allocate the driver and receiver power dissipation from the system-level, and manipulate multi-dimensional design trade-offs (eye-opening, power, area, etc.) to satisfy design requirements.

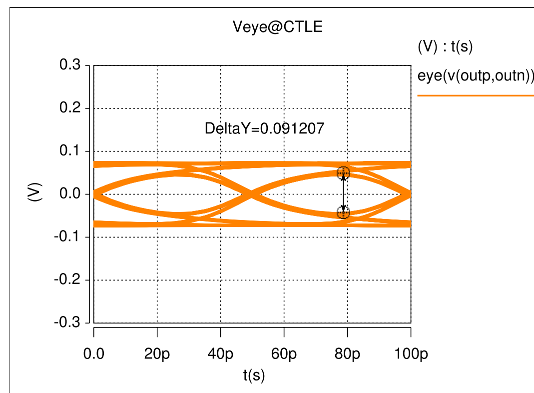
V.5 Driver-receiver co-design for low energy-per-bit

Combining each building block together using designs derived in previous sections, the whole system shown in Fig.V.1 can be built and we show the simulated eye-diagrams at different locations (CML driver output, wire-end, CTLE receiver output) as shown in Fig.V.9. This result is based on the design methodology which optimizes driver and receiver independently with specific eye-quality consideration, therefore eye-opening remains reasonable value at each observation node. However, this methodology may bring in unnecessary power overhead due to the over-design. In order to reduce the energy-per-bit for whole system, we explore the driver-receiver co-design by performing the non-linear optimization using the conventional design in Fig.V.9 as an initial solution. The co-optimization flow is developed using Sequential-Quadratic Programming method [8] and described in the following subsections. Using the proposed co-optimization flow, we perform a



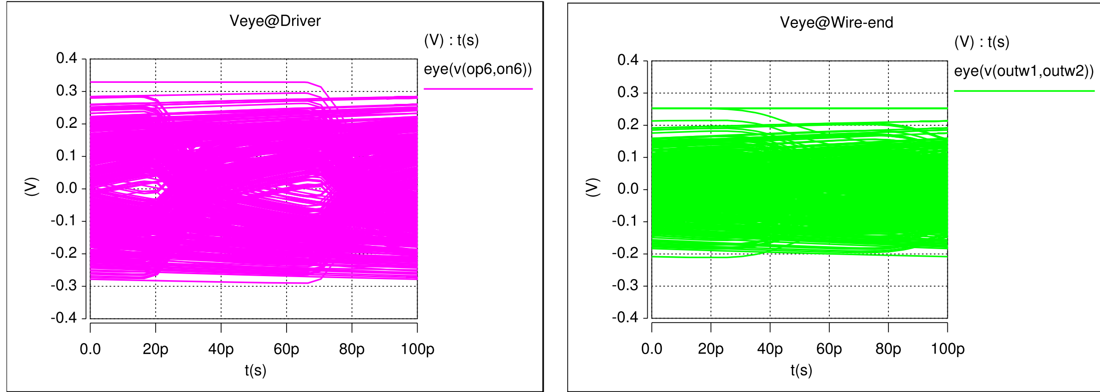
(a) Eye at driver output.

(b) Eye at wire-end.



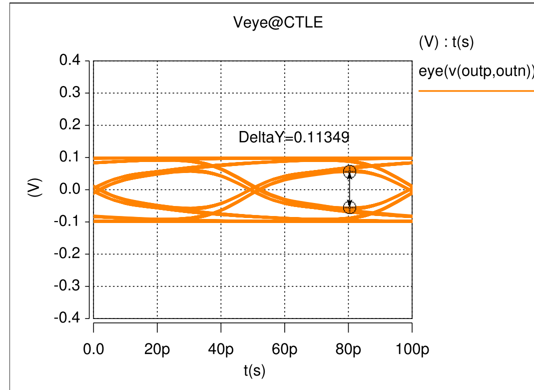
(c) Eye at CTLE output.

Figure V.9: Simulated eye-diagrams at different locations of proposed equalized on-chip global link using conventional design methodology. The design parameters: R_S=47 Ω, R_T=94 Ω, R_L=440 Ω, R_D=110 Ω, C_D=680 fF, V_{od}=60 mV. Simulated power consumption (driver+receiver w/o SA-latch) is 8.1 mW.



(a) Eye at driver output.

(b) Eye at wire-end.



(c) Eye at CTLE output.

Figure V.10: Simulated eye-diagrams at different locations of proposed equalized on-chip global link using low-power design methodology.The design parameters: R_S=148 Ω, R_T=1100 Ω, R_L=890 Ω, R_D=1430 Ω, C_D=150 fF, V_{od}=58 mV.Simulated power consumption (driver+receiver w/o SA-latch) is 3.8 mW.

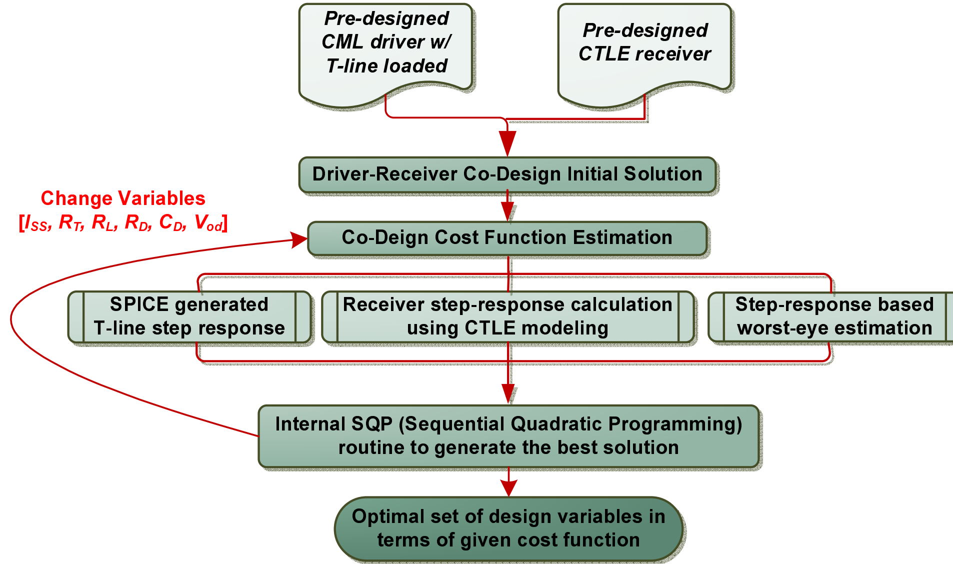


Figure V.11: The driver-receiver co-optimization flow.

set of experiments to study the performance tradeoffs of equalized on-chip global link in terms of different *T-line pitches* and *driver swings*. A full-link SPICE simulation with sense-amplifier latch loaded is also performed to show the performance metrics (including sensitivity to the process variations) of the proposed global link.

V.5.A Driver-receiver co-optimization flow

The proposed driver-receiver co-optimization flow is illustrated in Fig.V.11. In this flow, pre-designed CML driver and CTLE receiver are combined together as the initial solution. Co-design cost function is then estimated for certain specific solution. This stage is decomposed into three steps in the flow. Firstly, we use HSPICE to simulate the T-line step response for specified driver resistance (I_{SS}) and termination resistance (R_T). Secondly, step response after CTLE is calculated in MATLAB by adopting the proposed modeling approach (Eq.(V.5)). Finally, worst-case eye-opening after CTLE can be estimated using the algorithm in [62]. The co-design cost function, which will be minimized in the optimization, is defined

as,

$$f_1 = \frac{Power}{V_{eye}}, \tag{V.22}$$

or

$$f_2 = Power + c_1 e^{c_2(V_{min} - V_{eye})} \tag{V.23}$$

where c_1,c_2 are constant coefficients, and V_{min} is the user-defined minimal eye-opening constraint. Cost function f_1 indicates the power efficiency, which is minimized to reduce the power dissipated for boosting unit eye-opening. Cost function f_2 is used to minimize the total power dissipation of global link that satisfies the minimal eye-opening constraint V_{min}. In the following experiments, we first use cost function f_1 to demonstrate the effectiveness of co-design methodology, then switch to cost function f_2 to explore the design space in order to further reduce the power dissipation. After evaluating the cost function, a non-linear optimization routine which uses the internal Sequential Quadratic Programming (SQP) algorithm implemented in MATLAB [49] is called to permute the initial solution and guide the optimization iteration. In the end, the flow will generate the best solution, which is the optimal set of design variables of global link $[I_{SS}, R_T, R_L, R_D, C_D, v_{od}]$ in terms of user-defined cost function.

We use the proposed flow to optimize the initial solution in Fig.V.9 with cost function f_1, and present the eye-diagrams after optimization in Fig.V.10. It can be seen that eyes at the driver/T-line output are nearly closed after low-power optimization, but final eye after CTLE is actually improved from 91 mV to 113 mV, and total power dissipation drops from 8.1 mW to 3.8 mW as well. In terms of design parameters, driver resistance R_S and CTLE load resistance R_L increase dramatically to reduce the power dissipation, also other variables are adjusted accordingly to compensate the eye-opening loss due to larger resistance. The co-design solution also supports the analysis performed in Section V.4.B, which indicates that **total link power is minimized by reducing the extra driver/receiver power burned for optimizing eye-opening at internal nodes**.

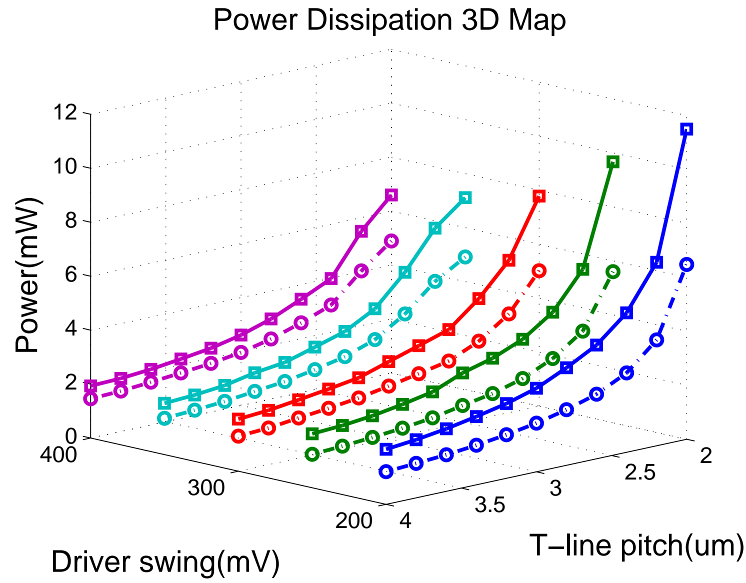
V.5.B Design space exploration

We explore the design space of equalized global link using the proposed driver-receiver co-optimization flow in this section. During the experiments, we sweep design parameters, which include driver output swing V_{sw} (swept from 200 mV to 400 mV with the interval 50 mV) and T-line *pitch*[9] (swept from 2 μm to 4 μm with the interval 0.2 μm), and run co-optimization flow to find the optimal global link design for each corresponding [V_{sw}, *pitch*]. The cost function f_2 (shown in Eq.(V.23)) is adopted in the optimization, and the minimal eye-opening constraint after CTLE is set to be 100 mV. The design target is set to be: **20 Gbps signaling over 10 mm on-chip wire in 45 nm predictive CMOS technology**.

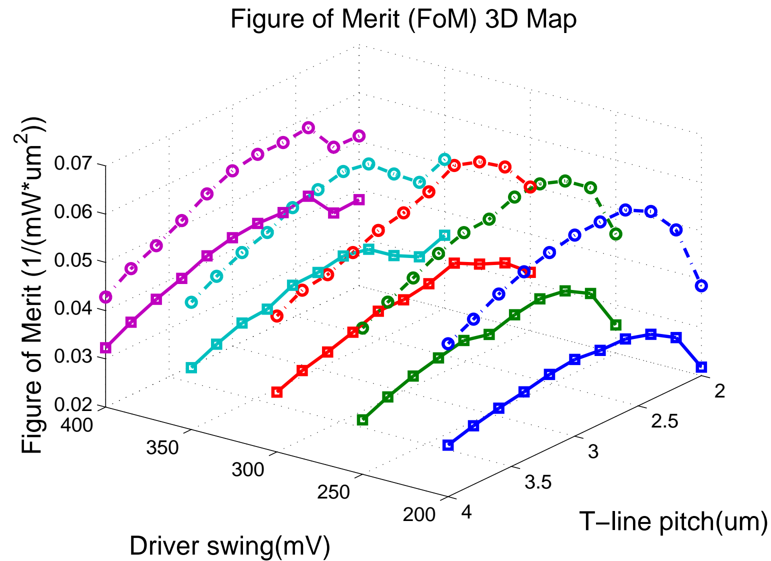
Based on the above settings, we run through the global link optimization flow and present the results in Fig.V.12 (**solid lines**). The average run time of generating the optimal solution for each [V_{sw}, *pitch*] set is about 20 min, which contains about 1000 iterations. The cost function evaluation for each design therefore takes about 1.2 sec using the CTLE modeling whereas the 1000-cycle PRBS HSPICE simulation normally takes 90 sec on the same machine. As a result, the proposed flow is demonstrated to achieve over 75x run time improvement and also guaranteed to catch the worst-case eye-opening scenario according to [62].

The power dissipation (CML driver and CTLE only, without including SA-latch yet) map in the explored [V_{sw}, *pitch*] space is shown in Fig.V.12(a). It is noted that global link power drops quickly as T-line pitch increases but tends to saturate when T-line pitch is larger than 3 μm. Based on Fig.V.12(a), power saving is between 60% and 70% when T-line pitch increases from 2 μm to 3 μm, whereas this number drops to about 25% to 35% when pitch increases from 3 μm to 4 μm. In terms of driver voltage swing, in order to boost the final eye-opening to satisfy the eye constraint, link power dissipation will increase as swing drops, however, this power increase becomes smaller and even can be ignored (less than

[9] T-line *width* is set to be $\frac{1}{2}$*pitch* as discussed in Section V.2.B.



(a) Power dissipation of global link versus different driver swings and T-line pitches.



(b) Figure of Merit (FoM) of global link versus different driver swings and T-line pitches.

Figure V.12: 3D map of power dissipation and figure of merit (FoM) for equalized global link in the explored design space. Solid lines indicate the single-supply design, whereas dash lines indicate the split-supply design.

15%) as T-line pitch is larger than 3 μm. In summary, total power dissipation of global link decreases for larger T-line pitch and higher driver voltage swing, but the power reduction will become very limited as pitch falls into [3 μm, 4 μm] region. The lowest achievable total power excluding SA-latch is about 1.9 mW (0.095 pJ/b for 20 Gbps data rate) in the explored design space.

To measure the overall performance of the proposed global link, we introduce and define the Figure of Merit (FoM) as follows,

$$FoM = \frac{1}{power \cdot pitch^2} \propto \frac{\text{Throughput Density}}{\text{Power} \times \text{Area}} \tag{V.24}$$

where throughput density is the amount of data (bandwidth) can be transferred per unit chip area (T-line pitch). Basically, larger FoM means more data can be transferred over the global link using less power and chip area. In Fig.V.12(b), we plot the map of FoM to show how it is affected by driver swing and T-line pitch. It is shown that there is an optimal T-line pitch value in terms of best FoM due to the trade-off of power and area. The optimal pitch slightly varies for different driver swings, but is in the range of [2.4 μm, 2.8 μm] based on Fig.V.12(b). FoM is also improved by increasing the driver swing but the gain is very limited, which is less than 20% in the explored space. As a result, **[400 mV, 2.4 μm]** is the optimal [V_{sw}, *pitch*] set which corresponds to the best FoM in the space. The power dissipation for this design is 3.98 mW (0.20 pJ/b for 20 Gpbs data rate) and the throughput density is about 2.78 Gpbs/μm.

In order to further reduce the total power of equalized global link, one solution is to provide split power supply for driver and receiver separately. Since the CML driver stage only needs to provide V_{sw} ($<V_{DD}$) driver swing, it is possible to reduce the driver supply to reduce driver power dissipation without affecting quality of transmitted signals. In order to make sure all the transistors in CML driver operated in saturation region, the lower bound of driver supply becomes,

$$V_{DD}^{CML} \geq V_{sw} + 2V_{ov} \tag{V.25}$$

where V_{ov} is the over-drive voltage of CML input and bias transistor (which provides the tail current). Assuming V_{ov}=150 mV in our design, we redraw the 3D map of power dissipation and FoM for split-supply design and overlay them with curves of single-supply design in Fig.V.12 by annotating with **dash lines**. It is shown that up to 45% power reduction (corresponding to 1.8x FoM improvement) can be achieved by adopting the split-supply approach. Also, since lower driver swing brings in more power saving according to Eq.(V.25), the optimal driver swing for split-supply design is lower than that of single-supply case. In the explored space, the optimal [V_{sw}, *pitch*] set is [**300 mV, 2.6 μm**] for split-supply design in terms of best FoM, which corresponds to 2.39 mW power (0.12 pJ/b for 20 Gbps data rate) and 2.56 Gpbs/μm throughput density. We will use this solution as well as the optimal single-supply solution to run the full link SPICE simulation (including SA-latch stage) to sign-off the overall performance and compare the tradeoffs of single-supply and split-supply approach in the following section.

V.5.C Full global link performance sign-off

To sign-off the performance of global link, SA-latch stage is included in the full link HSPICE simulation as shown in Fig.V.13. In this work, we adopt a double-tail latch-type sense-amplifier design [57] to build up the SA-latch, and add back-to-back *NOR* gate (the SR-latch) at the end to convert SA output signal from RZ (return-to-zero) to NRZ (non-return-to-zero) code same as the input PRBS patterns. The SA-latch is tuned based on predictive 45 nm CMOS process and can recover as low as 50 mV input differential signal back to 1 V full-swing. The power dissipation of SA-latch slightly changes with input signal level, and is around 1.5 mW for a 20 Gbps PRBS pattern with signal amplitude ranging from 50 mV to 250 mV.

Using the SA-latch shown in Fig.V.13 with the optimal single-supply and split-supply design found by the flow in Section V.5.B, we perform the full-

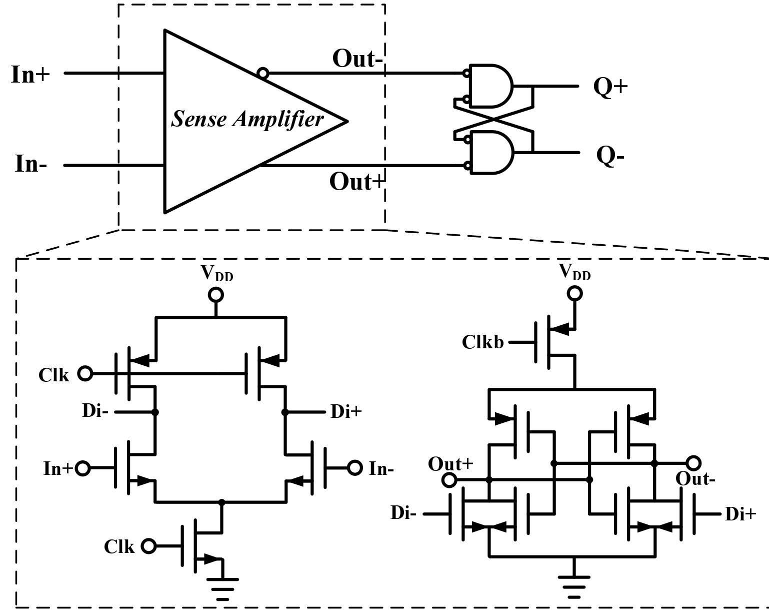


Figure V.13: The schematic of Sense-Amplifier Based Latch (SA-latch) used in this work.

Table V.5: Performance sign-off and comparison of on-chip equalized global link for single-supply and split-supply design.

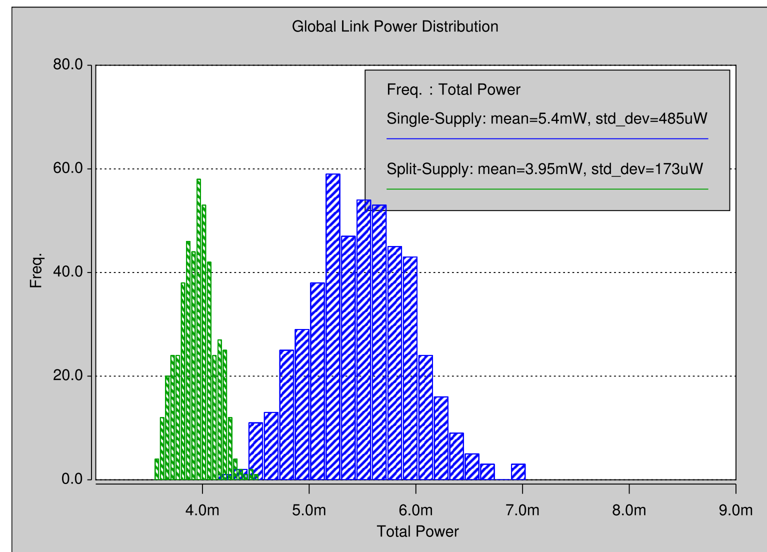
Performance Metrics	Single-Supply	Split-Supply
Total Power (mW)	5.54	3.91
Driver Power (mW)	3.30	1.88
Receiver Power (mW)	2.24	2.03
Total Delay (ps)	160	155
Driver Delay (ps)	45	40
T-line Delay (ps)	80	80
Receiver Delay (ps)	35	35
Energy per Bit (pJ/b)	0.277	0.196
Normalized Delay (ps/mm)	16	15.5
Throughput Density (Gbps/μm)	2.78	2.56
Yield* (%)	96	92.6

*Yield numbers are based on 500-run Monte Carlo simulations.

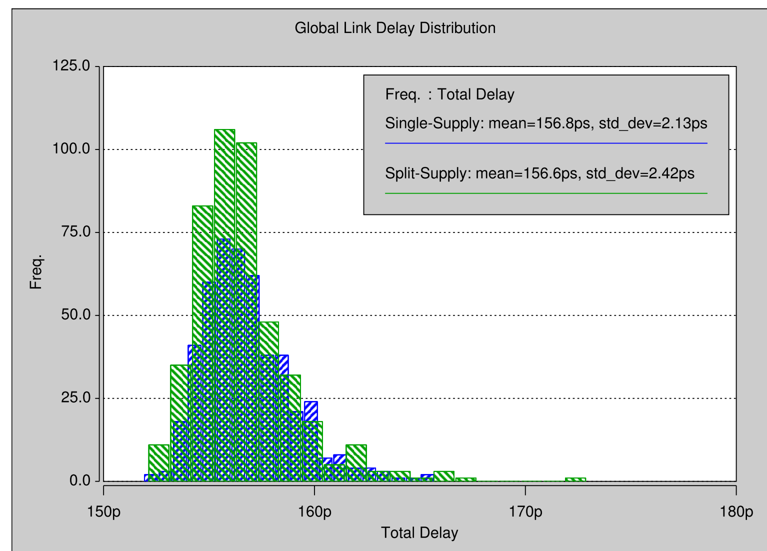
link HSPICE simulation and summarize the performance metrics in Table V.5. In the simulation, we employ 2^{10}-1 PRBS patterns to evaluate the signal quality and measure power and delay of the link. As shown in Table V.5, total power dissipation of global link has about 1.5 mW overhead compared with the values in Fig.V.12(a) due to the added SA-latch stage. The optimal split-supply design consumes less 30% power than single-supply one because of the great amount of power reduction (over 40% reduction) on the driver stage. In terms of total delay, both single-supply and split-supply design achieve very similar value, and the delay consumed on the driver and receiver is almost the same as time of flight of T-line. Split-supply design shows slightly smaller delay number because lower driver swing (300 mV) is used in the design. By using the normalized performance metrics to evaluate the global link, it is seen that split-supply design could achieve 20 Gbps signaling over 10 mm global wire with 0.196 pJ/b energy per bit, 15.5 ps/mm latency, and 2.56 Gpbs/μm throughput density. Split-supply design dominates single-supply design in terms of delay and power dissipation, but may increase chip area or introduce some yield loss due to more sensitivity to process variations, which will be studied in following Monte Carlo simulations.

In order to study the impact of process variations on the proposed global link, we perform 500-run Monte Carlo simulations in HSPICE to show the yield (percentage of runs can be recovered by SA-latch back to 1 V digital level), total delay and power distributions. To setup our Monte Carlo simulation, we assume all the transistors have 40 mV 3-σ deviation, and all the passive elements (resistance and capacitance) have 3-σ deviation equal to 10% of the nominal value except for R_D and C_D, which are assumed to have only 2%-nominal-value 3-σ deviation[10]. In addition, each supply voltage (two separate supplies in split-supply design)

[10]Because R_D and C_D are critical parameters affecting equalization quality of CTLE, during chip implementation, RC tuning approach [37] [64] can be used to control the variation of resistance and capacitance, especially RC product (R_DC_D), in certain limited range. As a result, smaller variation is assumed for R_D and C_D to reflect this tuning effort in the Monte Carlo simulation.



(a) Power dissipation distribution of global link.



(b) Delay distribution of global link.

Figure V.14: Histograms of global link power and delay distribution under process variations for single-supply and split-supply design based on 500-run Monte Carlo simulations in HSPICE.

is assumed to have 5%-nominal-value 3-σ deviation to model the power supply variation.

Applying the above settings, 500-run Monte Carlo simulation shows 96% and 92.6% yield number for single-supply and split-supply respectively, as shown in the last row of Table V.5. The possible reasons for yield loss can be: 1) poor eye-quality after CTLE due to either driver or CTLE receiver variations; 2) increased SA mismatch due to variations. The split-supply design has additional 3.4% yield loss compared with single-supply design, which can be attributed to additional supply variation added by a dedicated lower driver supply.

Excluding the outliers, which fail yield analysis, from the Monte Carlo simulation results, we draw histograms to illustrate the distribution of total delay and power dissipation of global link, and measure the approximate mean μ and standard deviation σ in Fig.V.14. Distribution of link power dissipation is shown in Fig.V.14(a). Blue bar and green bar indicate the single-supply and split-supply design, respectively. The single-supply design shows wider power distribution ($3\sigma/\mu$=26.9%), whereas the power variation of split-supply is much less ($3\sigma/\mu$=13.1%) due to the reduced driver power percentage in the total power. In terms of total delay, the equalized global link shows a very small variation. Single-supply and split-supply design have a nearly identical and narrow distribution as shown in Fig.V.14(b). The delay variation for single-supply $3\sigma/\mu$=4.0%, whereas the one for split-supply $3\sigma/\mu$=4.6%.

V.6 Summary

In this chapter, we propose an equalized on-chip global link structure by employing tapered CML driver, CTLE equalizer, and sense-amplifier based latch. Design guidelines to optimize CML driver with tuning the T-line dimensions are presented. Also, an accurate CTLE modeling is developed to improve the correlation of eye-opening prediction and SPICE simulation within 5%. Furthermore,

a set of analysis is performed to derive the theoretical formula for CTLE eye-opening, therefore eye-opening contour map can be generated to provide design insights from system-level. By adopting the SQP non-linear optimization method, we develop a driver-receiver co-optimization flow and apply it to explore design space of the proposed global link for lowest energy-per-bit.

The final optimal solution we found adopts separate supplies (600 mV for driver and 1 V for receiver) and 300 mV low-swing signaling, which can achieve 20 Gbps data rate over 10 mm, 2.6 μm-pitch on-chip T-line with 15.5 ps/mm latency and 0.196 pJ/b energy using predictive 45 nm CMOS process. The $3\sigma/\mu$ variations for power and delay are 13.1% and 4.6% respectively, based on 500-run Monte-Carlo simulation.

Chapter V includes the content of one submitted journal paper, "Energy Efficiency Optimization through Co-Design of the Transmitter and Receiver in High-Speed On-Chip Interconnects", by Y. Zhang, J. F. Buckwalter, C. K. Cheng, which is submitted to *IEEE Transaction on VLSI Systems.* The dissertation author was the primary investigator and author of the paper.

VI

Conclusion

VI.1 Summary of contributions

In this dissertation, we study the design and optimization of various on-chip interconnection schemes for high-speed and low-power global communication. Different wire operating modes (RC wire and transmission-line), signaling methods (single-ended and differential-pair; repeated, pipelined, and uninterrupted wiring), and equalization approaches (passive and active) are considered in the work, and the corresponding performance metrics are predicted and compared across several technology nodes. A general framework based on SQP non-linear programming, which is used to optimize the on-chip interconnection according to certain user-defined constraints, is also proposed and shown to be very effective in improving the energy efficiency of one active-equalized global interconnect design. The main contributions of each chapter are summarized as follows.

Chapter III reviews and compares six current global interconnection schemes in terms of latency, energy per bit, throughput, area, and signal integrity, and studies their scaling trends with the technology. A set of simple performance models for all these interconnections is proposed and used for early-stage estimation. An efficient general framework is also developed for optimizing on-chip T-line

interconnections.

Chapter IV explores the performance of pipelined *RC* interconnection with considering voltage and technology scaling for different applications. The impacts of pipelining depth, voltage and technology scaling are studied through numerical experiments by using a general evaluation flow. Overall performance improvement for adopting throughput-centric design methodology is demonstrated compared with traditional latency-aware optimization.

Chapter V proposes an active-equalized global interconnection for ultra-high-speed and low-power global communication. Accurate modeling and performance analysis are performed by adopting linear system method, and analytical formula for received eye-opening is derived to provide high-level design guidelines. A transmitter-receiver co-design methodology is developed to greatly improve energy-efficiency of the proposed interconnection.

VI.2 Future works

One potential future direction is to study the potential performance of on-chip global interconnection by employing emerging technologies, such as 3D interconnect using Through-Silicon Vias (TSVs), optical interconnect, carbon-nanotubes or nanowires, and so on. With the proper equivalent electronic modeling, the current optimization framework can be extended to optimize these new interconnection schemes and explore the design trade-offs compared with conventional copper interconnect.

The other research topic is to implement the proposed energy-efficient equalized interconnection with standard CMOS process and demonstrate the effectiveness through silicon measurement. The idea of energy reduction through transmitter and receiver co-design needs to be verified in real silicon. The design of on-chip T-line and accurate passive elements might be the main challenges need to be resolved during the implementation.

Appendix A

Performance analysis of ideal pipelined repeated *RC* wires

We analyze the performance metrics of pipelined repeated *RC* wire without maximum pipelining depth limit in the following, and define some parameters shown in Table A.1. Using defined parameters and assuming the energy and delay are evenly distributed within each pipelining stage the same as in previous long *R-RC* wires without flip-flop insertion, formulae for performance estimation can be derived as follows:

Table A.1: Design parameters used in performance analysis of *P-RC* structure

symbol	description
N	Pipelining depth (# of flip-flops inserted)
L	Total wire length
T_{FF}	Total latency of flip-flop (sum of T_{c-q} and T_{setup})
C_{FF}	Effective capacitance of flip-flop (obtained by power simulation)
t_{RC}	Normalized delay of optimized *R-RC* wire
e_{RC}	Normalized energy of optimized *R-RC* wire

$$\begin{aligned}
\text{Latency} &= NT_{FF} + t_{RC}L, && \text{(A.1)}\\
\text{Power} &= \frac{NC_{FF}V_{DD}^2 + e_{RC}L}{T_{FF} + t_{RC}(L/N)}, && \text{(A.2)}\\
\text{Bandwidth} &= \frac{1}{T_{FF} + t_{RC}(L/N)}, && \text{(A.3)}\\
\text{Energy} &= NC_{FF}V_{DD}^2 + e_{RC}L. && \text{(A.4)}
\end{aligned}$$

To derive the optimal pipelining depth N, we take the derivative of **Energy/Bandwidth**, and let it equal zero. The optimal N is shown to be,

$$N_{opt} = \sqrt{\frac{e_{RC}t_{RC}}{C_{FF}V_{DD}^2T_{FF}}}L \quad \text{(A.5)}$$

which is proportional to the wire length and shows an increasing trend with technology scaling. If there is no limit on the upper-bound of pipelining depth (ideal *P-RC* case), the performance metrics of *P-RC* in terms of **min-Energy/Bandwidth** can be obtained by plugging in (A.5) back to (A.1)-(A.4),

$$\begin{aligned}
\text{Latency} &= \left(\sqrt{\frac{e_{RC}t_{RC}T_{FF}}{C_{FF}V_{DD}^2}} + t_{RC}\right)L, && \text{(A.6)}\\
\text{Power} &= \left(\frac{e_{RC}}{T_{FF}}\right)L, && \text{(A.7)}\\
\text{Bandwidth} &= \frac{1}{T_{FF} + t_{RC}\sqrt{\frac{C_{FF}V_{DD}^2T_{FF}}{e_{RC}t_{RC}}}}, && \text{(A.8)}\\
\text{Energy} &= \left(\sqrt{\frac{e_{RC}t_{RC}C_{FF}V_{DD}^2}{T_{FF}}} + e_{RC}\right)L. && \text{(A.9)}
\end{aligned}$$

It can be seen that most metrics (latency, power, energy) of ideal *P-RC* structure are linearly proportional to L, except for the bandwidth, which is independent of the wire length L. Also, the bandwidth increases nearly exponentially as the technology scales, similar to the trend of transistor performance scaling.

Bibliography

[1] V. Adler and E.G. Friedman. Uniform repeater insertion in RC trees. *IEEE Trans. on Circuits and Systems I*, 47(10):1515–1523, Oct 2000.

[2] B. Analui, J. Buckwalter, and A. Hajimiri. Data-dependent jitter in serial communications. *IEEE Trans. Microwave Theory Tech.*, 53(11):1841–1844, 2005.

[3] A.Tsuchiya, M. Hashimoto, and H. Onadera. Design guidline for resistive termination of on-chip high-speed interconnects. In *Proc. IEEE Custom Integrated Circuits Conference*, pages 613–616, September 2005.

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

[5] K. Banerjee and A. Mehrotra. Accurate analysis of on-chip inductance effects and implications for optimal repeater insertion and technology scaling. In *Dig. of Tech. Papers Symp. on VLSI Circuits*, pages 195–198, Kyoto,Japan, Jun 2001.

[6] K. Banerjee and A. Mehrotra. A power-optimal repeater insertion methodology for global interconnects in nanometer designs. *IEEE Trans. on Electron Devices*, 49(11):2001–2007, Nov 2002.

[7] W. T. Beyene. The Design of Continous-Time Linear Equalizer Using Model Order Reduction Techniques. In *Proc. Dig. Elect. Perform. Electron. Packag.*, pages 187–190, San Jose, CA, Oct 2008.

[8] M. C. Biggs. Constrained Minimization Using Recursive Quadratic Programming: some alternative subproblem formulations. In *L.C.W. Dixon and G.P. Szego, eds.,Towards global optimization*, pages 341–349. North-Holland, Amsterdam, 1975.

[9] B.K. Casper, M. Haycock, and R. Mooney. An accurate and efficient analysis method for multi-Gb/s chip-to-chip signaling schemes. In *Proc. IEEE Symposium on VLSI Circuits*, pages 54–57, 2002.

[10] C.C.Liu, H.Zhu, and C.K.Cheng. Passive compensation for high performance inter-chip communication. In *Proc. of IEEE International Conference on Computer Design*, pages 547–552, October 2007.

[11] Chung-Ping Chen and Noel Menezes. Noise-Aware Repeater Insertion and Wire Sizing for On-Chip Interconnect Using Hierarchical Moment-Matching. In *Proc. of Design Automation Conf.*, pages 502–506, Jun 1999.

[12] Guoqing Chen and E.G. Friedman. Low-power repeaters driving RC and RLC interconnects with delay and bandwidth constraints. *IEEE Trans. on VLSI Systems*, 14(2):161–172, Feb 2006.

[13] Pasquale Cocchini. Concurrent flip-flop and repeater insertion for high performance integrated circuits. In *Proc. of Int. Conf. on CAD*, pages 268–273, 2002.

[14] J. Culetu, C. Amir, and J. MacDonald. A practical repeater insertion method in high speed VLSI circuits. In *Proc. of Design Automation Conf.*, pages 392–395, Jun 1998.

[15] V. V. Deodhar and J. A. Davis. Optimal Voltage Scaling, Repeater Insertion, and Wire Sizing for Wave-Pipelined Global Interconnects. *IEEE Trans. on Circuits and Systems I*, 55(4):1023–1030, May 2008.

[16] Vinita Deodhar. Throughput-Centric Wave-Pipelined Interconnect Circuits for Gigascale Integration. In *PhD Thesis*, Georgia Institute of Technology, 2005.

[17] A. Deutsch, P. W. Coteus, G. V. Kopcsay, H. H. Smith, C. W. Surovic, B. L. Krauter, D. C. Edelstein, and P. L. Restle. On-Chip Wiring Design Challenges for Gigahertz Operation. *Proc. IEEE*, 89(4):529–555, April 2001.

[18] A. Deutsch, H.H. Smith, C.W. Surovic, G.V. Kopcsay, D.A. Webber, P.W. Coteus, G.A. Katopis, W.D. Becker, A.H. Dansky, G.A. Sai-Halasz, and P.J. Restle. Frequency-dependent crosstalk simulation for on-chip interconnections. *IEEE Trans. Adv. Packaging*, 22(3):292–308, Aug 1999.

[19] I. M. Elfadel, A. Deutsch, G. V. Kopcsay, B. J. Rubin, and H. H. Smith. A CAD Methodology and Tool for the Characterization of Wide On-Chip Buses. *IEEE Trans. Adv. Packag.*, 28(1):63–70, Feb 2005.

[20] I. M. Elfadel, A. Deutsch, H. H. Smith, B. J. Rubin, and G. V. Kopcsay. A Multiconductor Transmission Line Methodology for Global On-Chip Interconnect Modeling and Analysis. *IEEE Trans. Adv. Packag.*, 27(1):71–78, Feb 2004.

[21] I.M. Elfadel, M.B. Anand, A. Deutsch, O. Adekanmbi, M. Angyal, H. Smith, B. Rubin, and G. Kopcsay. AQUAIA: a CAD tool for on-chip interconnect modeling, analysis, and optimization. In *Proc. Dig. Elect. Perform. Electron. Packag.*, pages 337–340, Oct 2002.

[22] M.P. Flynn and J.J. Kang. Global signaling over lossy transmission lines. In *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pages 985–992, November 2005.

[23] P. K. Hanumolu, G.-Y. Wei, and U.-K. Moon. Equalizers for High-Speed Serial Links. *International Journal of High Speed Electronics and Systems*, 15(2):429–458, 2005.

[24] M. Hashimoto, A. Tsuchiya, and H. Onodera. On-Chip Global Signaling by Wave Pipelining. In *IEEE. Electrical Performance of Electronic Packaging*, pages 311–314, 2004.

[25] M. Hashimoto, A. Tsuchiya, A Shinmyo, and H. Onodera. Performance Prediction of On-chip High-throughput Global Signaling. In *IEEE. Electrical Performance of Electronic Packaging*, pages 79–82, 2005.

[26] H.Chen, R.Shi, and C.K.Cheng. Surfliner: A distortionless electrical signaling scheme for speed-of-light on-chip communication. In *Proc. of IEEE International Conference on Computer Design*, pages 497–502, October 2005.

[27] Seongmoo Heo and Krste Asanovic. Power-Optimal Pipelining in Deep Submicron Technology. In *Proc. of Int. Symp. on Low Power Electronics and Design*, pages 218–223, 2004.

[28] Seongmoo Heo and Krste Asanovic. Replacing Global Wires with an On-Chip Network: A Power Analysis. In *Proc. IEEE Intl. Symp. on Low Power Elec. Design*, pages 369–374, Aug 2005.

[29] P. Heydari and R. Mohanavelu. Design of Ultrahigh-Speed Low-Voltage CMOS CML Buffers and Latches. *IEEE Trans. on VLSI Systems*, 12(10):1081–1093, 2004.

[30] H.Zhu, R.Shi, C.K.Cheng, and H.Chen. Approaching speed-of-light distortionless communication for on-chip interconnect. In *Proc. Asia and South Pacific Design Automation Conference*, pages 684–689, January 2007.

[31] IBM. IBM electromagnetic field solver suite of tools. In *http://www.alphaworks.ibm.com/tech/eip.*

[32] Y.I. Ismail and E.G. Friedman. Effects of inductance on the propagation delay and repeater insertion in VLSI circuits. *IEEE Trans. on VLSI Systems*, 8(2):195–206, Apr 2000.

[33] H. Ito, J. Inoue, S. Gomi, H Sugita, K. Okada, and K. Masu. On-Chip Transmission Line for Long Global Interconnects. In *IEEE. Int. Electron Device Meeting*, pages 677–680, 2004.

[34] A. Jantsch and H. Tenhunen, editors. *Networks on Chip*. Kluwer Academic Publishers, 2003.

[35] Howard Johnson and Martin Graham. *High-speed signal propagation*. Prentics Hall, 2003.

[36] P. Kapur, G. Chandra, and K.C. Saraswat. Power Estimation in Global Interconnects and Its Reduction Using a Novel Repeater Optimization Methodology. In *Proc. IEEE/ACM Design Automation Conf.*, pages 461–466, 2002.

[37] H. Khorramabadi and P. R. Gray. High-Frequency CMOS Continuous-Time Filters. *IEEE J. Solid-State Circuits*, SC-19(6):939–948, Dec 1984.

[38] B. Kim and V. Stojanovic. Equalized Interconnects for On-chip Networks: Modeling and Optimization Framework. In *Proc. IEEE Int. Conf. Computer Aided Design*, pages 552–559, San Jose, CA, Nov 2007.

[39] B. Kim and V. Stojanovic. A 4Gb/s/ch 356fJ/b 10mm Equalized On-Chip Interconnect with Nonlinear Charge-Injecting Transmit Filter and Transimpedance Receiver in 90nm CMOS. In *Proc. IEEE Int. Solid-State Circuits Conf.*, pages 66–68, Feb 2009.

[40] G. V. Kopcsay, B. Krauter, D. Widiger, A. Deutsch, B. J. Rubin, and H. H. Smith. A Comprehensive 2-D Inductance Modeling Approach for VLSI Interconnects: Frequency-Dependent Extraction and Compact Circuit Model Synthesis. *IEEE Trans. VLSI Systems*, 10(6):695–711, Dec 2002.

[41] Ja Chun Ku and Y. Ismail. Thermal-aware methodology for repeater insertion in low-power VLSI circuits. In *Proc. of Int. Symp. on Low Power Elec. and Design*, pages 86–91, 2007.

[42] L. Zhang and W. Yu and Haikun Zhu and Alina Deutsch and George A. Katopis and Daniel M. Dreps and Earnest Kuh and Chung-Kuan Cheng. Low Power Passive Equalizer Optimization Using Tritonic Step Response. In *Proc. IEEE/ACM Design Automation Conference*, 2008.

[43] L. Zhang and W. Yu and Y. Zhang and R. Wang and A. Deutsch and G. A. Katopis and D. M. Dreps and J. Buckwalter and E. Kuh and C.K Cheng. Low Power Passive Equalizer Design for Computer Memory Links. In *Proc. IEEE Symposium on High Performance Interconnects*.

[44] Weiping Liao and Lei He. Full-Chip Interconnect Power Estimation and Simulation Considering Concurrent Repeater and Flip-Flop Insertion. In *Proc. of Int. Conf. on CAD*, page 574, 2003.

[45] Xun Liu, Yuantao Peng, and M.C. Papaefthymiou. Practical repeater insertion for low power: what repeater library do we need? *IEEE Trans. on CAD*, 25(5):917–924, May 2006.

[46] Ruibing Lu, Guoan Zhong, Cheng-Kok Koh, and Kai-Yuan Chao. Flip-flop and repeater insertion for early interconnect planning. In *Proc. of Design Automation and Test in Europe Conf.*, pages 690–695, Paris,France, Mar 2002.

[47] N. Magen, A. Kolodny, U. Weiser, and N. Shamir. Interconnect-power dissipation in a microprocessor. In *Proc. Int. Workshop on System Level Interconnect Prediction*, pages 7–13, Paris, France, Feb 2004.

[48] Yehia Massoud, Jamil Kawa, Don MacMillen, and Jacob White. Modeling and analysis of differential signaling for minimizing inductive cross-talk. In *Proc. IEEE/ACM Design Automation Conf.*, pages 804–809, Las Vegas, NV, Jun 2001.

[49] MATLAB. In *The MathWorks, www.mathworks.com*, R2007a.

[50] R. Mclnerner, K. Leeper, T. Hill, H. Chan, B. Basaran, and L. McQuiddy. Methodology for Repeater Insertion Management in the RTL, Layout, Floorplan and Fullchip Timing Databases of the Itanium Microprocessor. In *Proc. of Int. Symp on Physical Design*, pages 99–104, 2000.

[51] E. Mensink, D. Schinkel, E. Klumperink, E. van Tuijl, and B. Nauta. A 0.28pJ/b 2Gb/s/ch Transceiver in 90nm CMOS for 10mm On-Chip interconnects. In *Dig. of Tech. IEEE. Int. Solid-State Circuits Conf.*, pages 414–612, 2007.

[52] J. C. Montesdeoca, J. A. Montiel-Nelson, and S. Nooshabadi. CMOS Driver-Receiver Pair for Low-Swing Signaling for Low Energy On-Chip Interconnects. *IEEE Trans. VLSI Systems*, 17(2):311–316, Feb 2009.

[53] R. Nagpal, A. Madan, A. Bhardwaj, and Y. N. Srikant. INTACTE: An Interconnect Area, Delay, and Energy Estimation Tool for Microarchitectural Explorations. In *Proc. ACM CASES' 07*, pages 238–247, 2007.

[54] A. Nalamalpu and W. Burleson. Repeater insertion in deep sub-micron CMOS: ramp-based analytical model and placement sensitivity analysis. In *IEEE Symp. on Circuits and Systems*, pages 766–769, Geneva,Switzerland, May 2000.

[55] Jan Rabaey. *Low Power Design Essentials*. Springer, 2009.

[56] G. A. Sai-Halasz. Performance Trends in High-End Processors. *Proc. IEEE*, 83(1):20–36, Jan 1995.

[57] D. Schinkel, E. Mensink, E. Klumperink, E.van Tuiji, and B. Nauta. A Double-Tail Latch-Type Voltage Sense Amplifier with 18ps Setup+Hold Time. In *Proc. IEEE Int. Solid-State Circuits Conf.*, pages 314–316, San Francisco, CA, Feb 2007.

[58] D. Schinkel, E. Mensink, E. A. Klumperink, E. van Tuijl, and B. Nauta. A 3-Gb/s/ch Transceiver for 10-mm Uninterrupted RC-limited Global On-Chip Interconnects. *IEEE Jour. Solid State Circuits*, 41(1):297–306, Jan 2006.

[59] Semiconductor Industry Association. International Technology Roadmap for Semiconductors. In *http://www.itrs.net*, 2004,2006,2007.

[60] Semiconductor Industry Association. International Technology Roadmap for Semiconductors. In *http://www.itrs.net*, 2007,2008 Update.

[61] Shah, Harshit and Shiu, Pun and Bell, Brian and Aldredge, Mamie and Sopory, Namarata and Davis, Jeff. Repeater insertion and wire sizing optimization for throughput-centric VLSI global interconnects. In *Proc. IEEE/ACM International Conference on Computer-Aided Design*, pages 280–284, 2002.

[62] R. Shi, W. Yu, Y. Zhu, E. S. Kuh, and C-K. Cheng. Efficient and Accurate Eye Diagram Prediction for High Speed Signaling. In *Proc. IEEE Int. Conf. Computer Aided Design*, pages 655–661, San Jose, CA, Nov 2008.

[63] Jaemin Shin and Kemal Aygun. On-Package Continuous-Time Linear Equalizer using Embedded Passive Components. In *Proc. IEEE Electrical Performance of Electronic Packaging*, pages 147–150, October 2007.

[64] J. Silva-Martinez, M. Steyaert, and W. Sansen. A 10.7-MHz 68-dB SNR CMOS Continuous-Time Filter with On-Chip Automatic Tuning. *IEEE J. Solid-State Circuits*, 27(12):1843–1853, Dec 1992.

[65] S. Sim, S. Krishnan, D. Petranovic, and N. Arora. A Unified RLC Model for High-Speed On-Chip Interconnects. *IEEE Trans. Electron Devices*, 50(6):1501–1510, June 2003.

[66] H. Smith, A. Deutsch, S. Mehrotra, D. Widiger, M. Bowen, A. Dansky, G. V. Kopcsay, and B. Krauter. R(f)L(f)C Coupled Noise Evaluation of an S/390 Microprocessor Chip. In *Proc. IEEE Custom Integrated Circuits Conf.*, pages 237–240, San Diego, CA, May 2001.

[67] D. Sylvester and K. Keutzer. A Global Wiring Paradigm for Deep Submicron Design. *IEEE Trans. on Computer Aided Design of Integrated Circuits and Systems*, 19(2):242–252, Feb 2000.

[68] Paul Teehan, Guy G. F. Lemieux, and Mark R. Greenstreet. Estimating reliability and throughput of source-synchronous wave-pipelined interconnect. In *Proc. of Int. Symp. on NoCs*, pages 234–243, 2009.

[69] A. Tsuchiya, T. Kuboki, and H. Onodera. Low-Power Design of CML Drivers for On-Chip Transmission-Lines. In *Proc. Workshop on Synthesis and System Integration of Mixed Information Tech.*, Apr 2006.

[70] S. Uemura, A. Tsuchiya, and H. Onodera. A Predictive Transistor Model Based on ITRS Roadmap. In *Proc. General Conf. IEICE*, page 81, Mar 2006.

[71] R. Venkatesan, J. A. Davis, and J. D. Meindl. Compact Distributed RLC Interconnect Models-Part IV:Unified Models for Time Delay, Crosstalk, and Repeater Insertion. *IEEE Trans. Electron Devices*, 50(4):1094–1102, Apr 2003.

[72] R. Venkatesan, J.A. Davis, and J.D. Meindl. Compact distributed RLC interconnect models - part IV: unified models for time delay, crosstalk, and repeater insertion. *IEEE Trans. on Electron Devices*, 50(4):1094–1102, Apr 2003.

[73] J. Wood, T. C. Edwards, and S. Lipa. Rotary Traveling-Wave Oscillator Arrays: A New Clock Technology. *IEEE Journal of Solid-State Circuits*, 36(11):1654–1665, Nov 2001.

[74] J. Xu and W. Wolf. A wave-pipelined on-chip interconnect structure for networks-on-chips. In *Proc. IEEE Symp. on High Performance Interconnects*, pages 10–14, Aug 2003.

[75] Jiang Xu and Wayne Wolf. Wave pipelining for application-specific networks-on-chips. In *Proc. of Int. Conf. on Compilers, Architecture, and Synthesis for Embedded Systems*, pages 198–201, 2002.

[76] L. Zhang, H. Chen, B. Yao, K. Hamilton, and C-K. Cheng. Repeated On-Chip Interconnect Analysis and Evaluation of Delay, Power, and Bandwidth Metrics under Different Design Goals. In *Proc. IEEE Int. Symp. Quality Electron. Design*, pages 251–256, San Jose, CA, Mar 2007.

[77] L. Zhang, J. Wilson, R. Bashirullah, L. Luo, J. Xu, and P. Franzon. Driver Pre-emphasis Techniques for On-Chip Global Buses. In *Proc. IEEE Int. Symp. Low Power Electronic Design*, pages 186–191, Aug 2005.

[78] L. Zhang, Y. Zhang, A. Tsuchiya, M. Hashimoto, E. S. Kuh, and C-K. Cheng. High Performance On-Chip Differential Signaling Using Passive Compensation for Global Communication. In *Proc. Asia and South Pacific Design Automat. Conf.*, pages 385–390, Yokohama, Japan, Jan 2009.

[79] Liang Zhang, John Wilson, Rizwan Bashirullah, Lei Luo, Jian Xu, and Paul Franzon. Driver pre-emphasis techniques for on-chip global buses. In *Prof. Int. Symp. on Low Power Elec. and Design*, pages 186–191, 2005.

[80] Ling Zhang, Yulei Zhang, Hongyu Chen, Bo Yao, K. Hamilton, and Chung Kuan Cheng. On-Chip Interconnect Analysis of Performance and Energy Metrics Under Different Design Goals. *IEEE Trans. on VLSI Systems*, 19(3):520–524, Mar 2011.

[81] Lizheng Zhang, Yuhen Hu, and Charlie Chung-Ping Chen. Wave-pipelined on-chip global interconnect. In *Proc. of Asia and South Pacific DAC*, pages 127–132, 2005.

[82] Y. Zhang, J. F. Buckwalter, and C. K. Cheng. High-speed and low-power on-chip global link using continuous-time linear equalizer. In *Proc. Dig. Elect. Perform. Electron. Packag.*, pages 5–8, Austin, TX, Oct 2010.

[83] Y. Zhang, L. Zhang, A. Deutsch, G. A. Katopis, D. M. Dreps, J. F. Buckwalter, E. S. Kuh, and C-K. Cheng. Design Methodology of High Performance On-Chip Global Interconnect Using Terminated Transmission-Line. In *Proc. IEEE Int. Symp. Quality Electron. Design*, pages 451–458, San Jose, CA, Mar 2009.

[84] Y. Zhang, L. Zhang, A. Tsuchiya, M. Hashimoto, and C-K. Cheng. On-Chip High Performance Signaling Using Passive Compensation. In *Proc. IEEE Int. Conf. Computer Design*, pages 182–187, Lake Tahoe, CA, Oct 2008.

[85] Yulei Zhang, Xiang Hu, Alina Deutsch, A. Ege Engin, James F. Buckwalter, and Chung-Kuan Cheng. Prediction of High-Performance On-Chip Global Interconnection. In *Proc. of the Intl. Workshop on System Level Interconnect Prediction*, pages 61–68, San Francisco, CA, Jul 2009.