COMPARISON STUDY OF FUTURE ON-CHIP INTERCONNECTS FOR HIGH PERFORMANCE VLSI APPLICATIONS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Kyung-Hoae Koo
March 2011

2011 by Kyung Hoae Koo. All Rights Reserved.
Re-distributed by Stanford University under license with the author.

This work is licensed under a Creative Commons Attribution-Noncommercial 3.0 United States License.
http://creativecommons.org/licenses/by-nc/3.0/us/

This dissertation is online at: http://purl.stanford.edu/kv550rj8376

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Krishna Saraswat, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

James Harris

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Philip Wong

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

This page is intentionally left blank

Abstract

Moore's law has driven the scaling of digital electronic devices' dimensions and performances over the last 40 years. As a result, logic components in a microprocessor have shown dramatic performance improvement. On the other hand, an on-chip interconnect which was considered only as a parasitic load before 1990s became the real performance bottleneck due to its extremely reduced cross section dimension. Now, on-chip global interconnect with conventional Cu/low-k and delay optimized repeater scheme faces great challenges in the nanometer regime, imposing problems of slower delay, higher power dissipation and limited bandwidth. Carbon based materials such as carbon nanotubes and graphene nanoribbons, and optical interconnect have been proposed for the alternate solution for the future nodes due to their special physical characteristics.

This dissertation investigates the basic physical properties of novel materials for future interconnect, and describes the analytical and numerical models of local and global wire system based on new materials and novel signaling paradigms. This work also compares their basic performance metrics and circuit architectures to cope with

the interconnect performance bottlenecks. We quantify the performance of these novel interconnects and compare them with Cu/low-k wires for future high-performance ICs. We find that for a local wire, a CNT bundle exhibits a smaller latency than Cu for a given geometry. In addition, by leveraging the superior electromigration properties of CNT and optimizing its geometry, the latency advantage can be further amplified. For semiglobal and global wires we compare both optical and CNT options with Cu in terms of commonly used elementary metrics: latency and power dissipation. The above trends are studied with technology node. In addition, for a future technology node, we compare the relationship between system's metrics such as bandwidth density, power density and latency, thus alluding to the latency and power penalty to achieve a given bandwidth density for each type of wire. Optical wires have the lowest latency and the highest possible bandwidth density using wavelength division multiplexing. While, CNT has a lower latency than Cu. The power density comparison is highly switching activity dependent, with high switching activity being optics favorable. At low switching activity, optics is only found to power density favorably beyond a critical bandwidth density. We have also quantified the impact of improvement in optical and CNT device, material, and system parameters on the above comparisons. A small detector and modulator capacitances for optical interconnects (~10fF) yields superior, at least comparable, performance with CNTs (practical, electron mean free path of 0.9μm) and Cu for greater than 35% and 20% switching activity, respectively. However, improving mean free path of CNTs (~2.8μm) increases this crossover switching activity to 80%The above trends are studied with technology node and bandwidth density in terms of latency and power

dissipation. Optical wires have the lowest latency and power consumption, while a carbon nanotube (CNT) bundle has lower latency than Cu. The new circuit scheme, 'capacitively driven low-swing interconnect (CDLSI)', has the potential to effect a significant energy saving and latency reduction. We present an accurate analytical, optimization model for the CDLSI wire scheme. In addition, we quantify and compare the delay and energy expenditure for not only different interconnect circuit schemes, but also various future technologies such as Cu, carbon nanotube and optics. We find that the CDLSI circuit scheme outperforms the conventional interconnects in latency and energy per bit for lower bandwidth requirement, while these advantages degrade for higher bandwidth requirements. Lastly, we explore the impact of CNT bundle and CDLSI on a via blockage factor. CNT shows a significant reduction in via blockage while CDLSI does not help to alleviate it, though CDLSI results in a reduced number of repeaters, due to differential signaling scheme. There exists the uncertainty in previously published literatures regarding experimental work of m-SWCNTs both in DC and AC characterization. Therefore, we focused our attention to experimental AC characterization of m-SWCNTs. We also conducted DC measurement to verify no discrepancy of DC measurement with RF characterization. We find that the existence of a high kinetic inductance (~60nH/ μm) is clear. Subsequent DC (f=0Hz) resistance extraction from s-parameter measurements proves that the resistance of the m-SWCNT varies as applied input power varies.

Acknowledgements

For the last five and a half years, Stanford has offered me wonderful opportunities to intellectually reshape myself. Especially, Stanford department of electrical engineering has exposed me to an intellectually demanding environment with its top class faculties, staffs, and students. In addition, Stanford Nanofabrication Facility, being a world class research lab, has provided me with the tools and knowledge which are needed to perform my daily research. All these things together have relentlessly inspired and motivated me to move forward with my work in a consistent manner. There are so many people whom I want to thank.

First of all, I would like to express my deepest gratitude to my principal advisor, Prof. Krishna C. Saraswat. I would not have been able to finish my Ph.D. research without his support and care. My curiosity about on-chip interconnect area could be realized as practical output because he has given me right guidance and continuously offered me intellectual challenges. On the other hand, he has allowed a lot of degree of freedom in conducting my research. This has enabled me to experience not only system performance analysis of interconnect but also hands-on carbon nanotube interconnect fabrication at Stanford Nanofabrication Facility. His deepest understanding about interconnect and broad knowledge in overall micro-electronics have been one of key factors for me to reach this far.

I also want to thank to my associate advisor H.-S. Philip Wong. His inspirational feedback through discussion on carbon nanotube work including support from his group members has helped me tremendously to solve many difficult problems in my research. Without his collaboration, I would not have been able to finish my Ph.D. work.

Next, I would like to thank Prof. James Harris. I remember that he has been really enthusiastic about optics throughout all of his classes that I have taken. His overall insight on optics has inspired me to get involved in on-chip optical interconnect research. His advice and recommendation for completing this dissertation were a great help for me.

I also want to thank Dr. Pawan Kapur. He had been not only a great mentor for my research but also a wonderful co-worker. His expertise on on-chip interconnects and broad spectrum on general electronics from nano-electronics to optics had enabled me to gain insight on system analysis of interconnects. Whenever I had to halt my research for a moment to solve difficult problems, his practical advice has always been a crucial ingredient for breakthrough.

I also would like to thank to administrative associate Gail Chun-Creech for all of her help. I would not have been able to focus on my research without her professional administrative support.

Acknowledgements

I would like to thank to all of my lab mates Hoyeol Cho, Ali Kemal Okyay, Hoon Cho, Crystal Kenny, Donhyun Kim, Duygu Kuzum, Gunhan Ertosun, Hyung-Yong Yu, Jin-Hong Park, Sarves Verma, Shyam Raghunathan, Arunanshu Roy, WooShik Jung, Yeul Na, Aneesh Nainanee, Donguk Nam, Ashish Pal, Dave Sukhdeo, Jason Lin, Suyog Gupta, and Ju-Hyung Nam. Each of them was very smart researcher. It was my privilege to have a chance to work and have long discussions with them

I would like to thank many Korean friends in CIS. Most of the time, I had to spend long hours over midnight at CIS to make things work. I could laugh because I could chat with them whenever I was in good or bad mood. CIS become a brighter place because of them although there weren't many windows for sunlight in CIS.

I would like to thank to my parents in Korea. They have always been my best supporters and friends throughout entire of my life. I deeply appreciate their endless prayer for me and my family. Thank you so much.

Most of all, I thank to my lovely wife Jisun Ahn who has been my soul mate since spring 2002. Her support and sacrifice have always been my best motivation for all of the work that I have achieved. My little son, Ja-yoon Koo, has been the most precious gift to me for the last two years. I want to thank him for being a great joy for me in long unreachable tunnel of Stanford graduate school. Now, finally, I have made it because of my family.

Table of Contents

List of Tables

List of Figures

Chapter 3

Chapter 4

Chapter 1

Introduction

1.1 Motivation

The ever decreasing interconnect cross section dimensions give rise to increase in resistance. In addition, surface and grain-boundary scattering of electrons in Cu becomes a serious problem as the wire size becomes almost comparable to the grain size of Cu. eventually leading to higher resistivity than bulk Cu [1][2]. Putting all these together, degradation of the RC time constant of on-chip metal wires becomes more serious. As a result, the continuous performance degradation of on-chip Cu/low-k interconnects is one of the greatest challenges to keep Moore's law alive while the scaling of transistors' dimension has provided relentless delay improvement as shown in Fig. 1.1 [3]

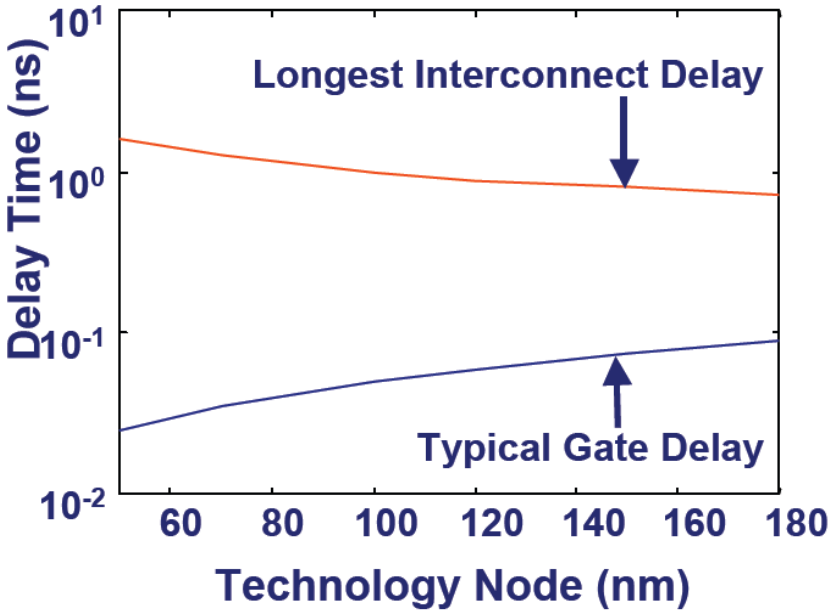

Fig. 1.1 Delay as a function of technology node both for global interconnect and typical CMOS gate

The scaling of wire dimension deteriorates not only delay time but also all related interconnect performance metrics, such as power dissipation, reliability, and bandwidth, for local, semi-global and global tiers. The on-chip power dissipation problem is coupled with an increasing number of repeaters to alleviate long RC time constant of Cu wire, switching activity factor, and increase of operating frequency. The reliability issue also becomes very important since future systems will require higher current density within the reduced wire cross section to maintain or boost the operating frequency. This is directly coupled with electromigration induced hillocks and voids in Cu as shown in fig 1.2. Both hillocks and voids are detrimental to on-chip signaling because they are responsible for shorts between adjacent interconnect lines and opens to single signal path which are main causes of functional failure of the system [4].

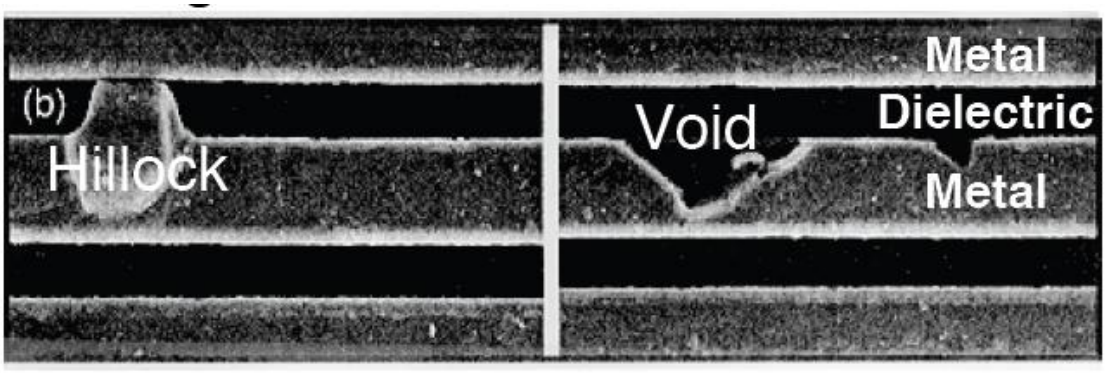

Fig. 1.2 Hillocks and voids induced by electromigration with high current density in Cu interconnect.

In addition, the increasing number of cores per processor, while the number of connections between cores is not, places a premium on high bandwidth density (bandwidth per unit cross sectional distance) (Φ_{BW}), and low latency links between cores as shown in Fig. 1.3. Furthermore, the budget of core peripheries will be more limited because the die size tends to remain unchanged. This will make Φ_{BW} even more critical [5].

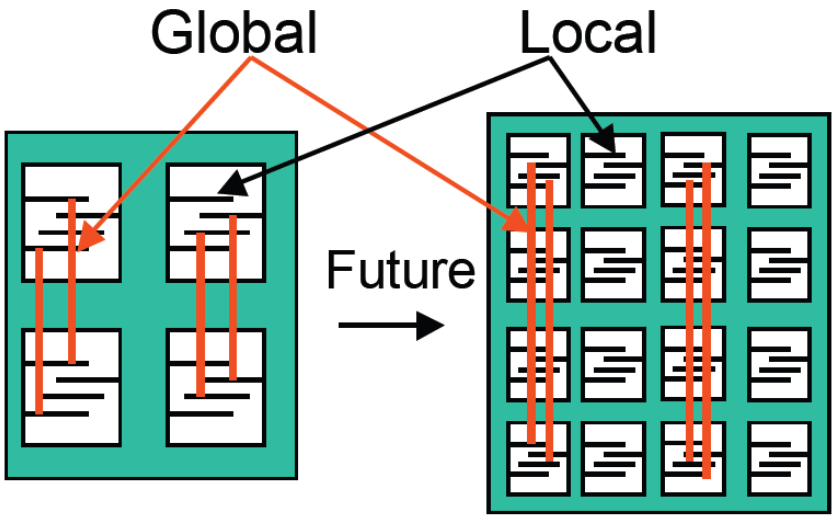

Fig. 1.3 Comparison between 4 cores and 16 cores and interconnect layout, illustrating that future many cores will be more bandwidth density hungry.

Therefore it is imperative to investigate novel interconnect technologies which can alleviate the aforementioned problems of Cu/low-k interconnects. Metallic carbon-based (carbon nanotube (CNT), graphene nanoribbons (GNR)) and optical interconnects are considered as two promising alternatives to cope with the problems [6][7]. CNTs exhibit performance advantages over Cu because the ballistic transport of electrons over distances of micrometer scale results in a much lower resistivity, and strong bonds between carbon atoms create a much higher electromigration tolerance [8]. For instance, a 1Ghz CNT-integrated oscillator has been demonstrated with multi-wall (MW)CNT interconnects, expediting the advent of high performance CNT-based interconnect fabric [9]. On the other hand, optical interconnects differ fundamentally from the electrical schemes (CNT and Cu). First, a large part of the latency and the entire power dissipation is in the end-devices instead of the waveguide. Second, the nature of power dissipation is mostly static rather than dynamic [7][10]. These differences, when coupled with favorable wire architectures, present new opportunities for optical interconnects. Although promising for most interconnect metrics, optics does suffer from the drawback of a relatively large transmission medium (waveguide) pitch (~0.6μm). The resulting Φ_{BW} limitation can be surmounted using the unique wavelength division multiplexing (WDM) option available for optical interconnects [11].

While these new interconnect technologies show promise for meeting future system interconnect requirements, they are currently impractical due to manufacturability limitations, although physics grants a possibility. On the other hand, a new low swing interconnect circuit scheme – “capacitively driven low swing

interconnects" (CDLSI) – is highly practical, while being equally promising, and hence warrants a detailed analysis [12][13]. The advantages of CDLSI over the conventional schemes are two fold: first, it can enormously reduce the energy per bit from a reduced voltage swing. Second, it can achieve a smaller delay from the pre-emphasis effect (explained in chapter 3). Thus it is also important to investigate this novel circuit scheme.

1.2 SPICE Model and Performance Metrics

Most important performance figures of merit of interconnect are speed, power, signal integrity, and bandwidth. In this section, we introduce a SPICE model for interconnects and define useful interconnect figures of merit.

1.2.1 SPICE Model

Basic physical properties of metal (Cu or Al) based interconnects consist of resistance (*R*), capacitance (*C*), and inductance (*L*). In an on-chip interconnect inductance value is mostly not taken into account because wire rise (fall) time is more dominant over the time of flight. Therefore on-chip interconnects can be approximated as a lossy RC network as shown in fig 1.4

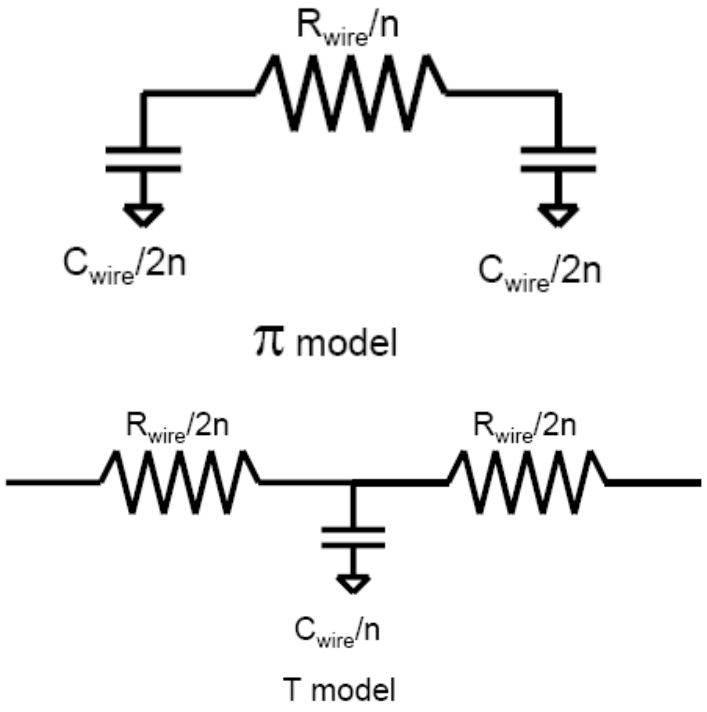

Fig. 1.4 One segment of distributed wire model in SPICE model a). is a π model and b) is T model.

In a SPICE simulation, interconnects are modeled as a distributed RC line with *n* number of segments. Ideally infinite *n* and infinitesimal segment length are needed for the most accurate delay estimation. This is because a simple RC network model for a long on-chip wire results in significant errors due to large resistance and capacitance. There are two types of lumped electrical wire models. One is the π model and the other is the T model as shown in Fig. 1.4. The accuracy of the models is determined by the number of segments. For example,a chain of more than three consecutive π stages gives an error of less than 3% [14].

1.2.2 Delay Model

Delay is one of the most important performance metrics. In most cases, building the equation of waveform in a time domain of the distributed RC network is very complex while setting up the s-domain equation can be simply done with differential equations. The Elmore delay equation, of which the mathematical meaning is the first moment of the impulse response, can help simplify the delay calculation of complex RC network. [15]. The Elmore delay is given by

$$\tau_{elmore,i} = \sum_{k=1}^{N} C_k R_{ik} \quad (1.1)$$

$$R_{ik} = \sum R_j \quad (1.2)$$

The Elmore delay is simply the sum of *RC* time constant of each node with common path resistance (R_{ik}) between node *k* and *i* where *k* is the index of each node and *i* is the node where delay need to be measured. Using this simple relationship, the wire delay of equivalent circuit in fig 1.5 can be simply given by

$$\tau_{wire} = \alpha \cdot R_{dr}C_p + \beta \cdot R_w C_w + \beta \cdot R_{dr}C_w + \beta \cdot (R_{dr} + R_w)C_L \quad (1.3)$$

where α and β are determined by the type of network and points of interest of input step response (summarized in table 1.1). R_{dr} is driver resistance. C_p and C_L are driver parasitic and transmitter load capacitance, respectively.

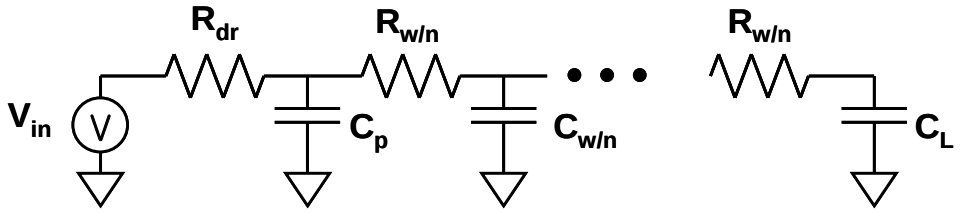

Fig. 1.5 Equivalent circuit of a distributed RC interconnect with step input function.

Voltage	α (Lumped RC)	β (distributed RC)
0 → 50%	0.69	0.38
0 → 60%	1	0.5
10% → 90%	2.2	0.9
0 → 90%	2.3	1

Table 1.1 α and β for lumped and distributed network for different points of interest

1.2.3 Power Dissipation Model

The power consumption of interconnects can be partitioned into three components, dynamic, static, and dynamic short circuit power. Dynamic power dissipation is due to charging and discharging of load capacitance (C_L). C_L includes wire capacitance, parasitic and input capacitance of repeaters. Each time the gate is switching, either charge is supplied from the power supply to C_L while PMOS transistors burn the

power or they are drawn to ground with CMOS burning the power. Dynamic power is given by

$$P_{dyn} = a \cdot C_L V_{swing} V_{DD} f_{0\to1} \quad (1.4)$$

where a is switching activity factor and $f_{0\to1}$ is the frequency of energy-consuming transitions. Static power consumption means the power dissipation without any switching activity. This includes gate leakage, source-drain leakage, and junction leakage in repeaters. Putting these all together, static power can be described as

$$P_{dyn} = I_{stat} V_{DD} \quad (1.5)$$

where I_{stat} is the current from V_{DD} to GND and there is no switching activity. Dynamic short circuit power represents the power dissipation due to the current flow when both NMOS and PMOS are in their saturation regions.

1.2.4 Bandwidth/Bandwidth Density

The bandwidth of interconnects represents their ability to send how many bits per second via wire. If the delay of the interconnect is τ, then ideally the inverse of τ is the number of bits that interconnect can handle within one second. If the interconnect is pipelined or repeated, then the throughput of the interconnect further increases. For example, if the delay of the pipeline segment is τ_{seg}, then the bandwidth of this system is $1/\tau_{seg}$ instead of inverse of total wire delay (τ). Currently, the bandwidth density (Φ_{BW}) becomes an even more important performance figure of merit rather than just the wire bandwidth. This is because a core in the on-chip die tends to have more limited

periphery with an advent of multi-core paradigm whereas interconnects should be laid out within its limit. This makes the chip more bandwidth hungry. Φ_{BW} is given by

$$\Phi_{BW} = \frac{f_{clk}}{W_{pitch}} \quad (1.6)$$

f_{clk} is the system clock defined by timing constraints of the system. W_{pitch} is the wire pitch.

1.2.5 Signal Integrity

Noise is one of major concerns in order to maintain correct functionality in a digital system. Noise in digital signaling can be categorized into the noise proportional to signal swing and that which is independent of signal.

$$V_N = K_N V_S + V_{NI} \quad (1.7)$$

where V_N and V_S are noise and signal voltages. K_N is a noise coefficient proportional to signal swing. V_{NI} is a noise source independent of signal swing. It is critical to minimize V_N to cope with errors in digital signaling.

Capacitive coupling cross talk is the main source of K_N. Fig 1 illustrates the example of the capacitive coupling between two wires.

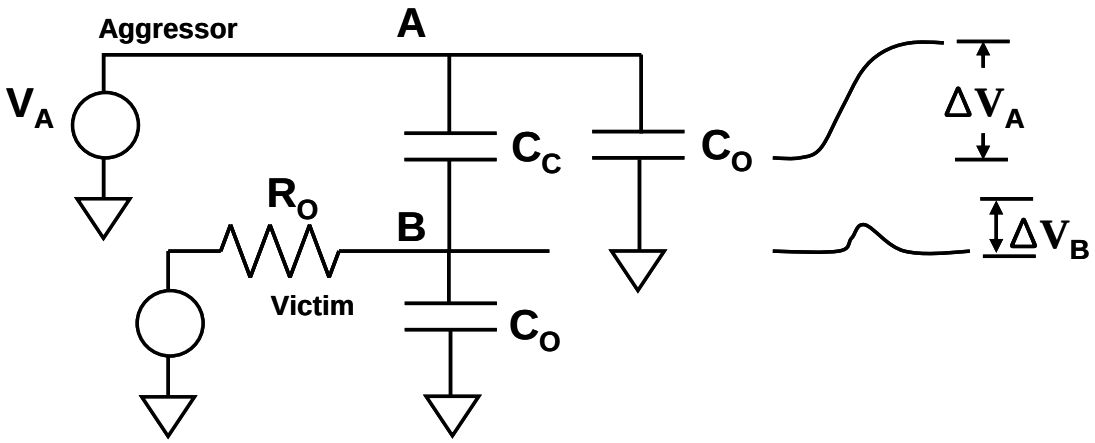

Fig. 1.6 Capacitive coupling cross talk caused by wire A in wire B.

A coupling capacitance between aggressor and victim lines forms a capacitive voltage divider creating unwanted voltage overshoot on the victim node. V_{NI} in (1.7) is mainly determined by transmitter or receiver offset voltages. Power supply noise is random noise due to non-ideal impedance of the power supply rail [16].

1.3 Dissertation Organization

The dissertation is organized as follows.

Chapter 1: Introduction

Chapter 1 - Introduction

Chapter 2 : Performance Comparison Study between Cu/low-k, m-SWCNT Bundle, and Optical Interconnects

Chapter 3: Capacitively Driven Low Swing Interconnect (*CDLSI*) and Its Analytical Modeling

Chapter 4: DC and AC Characterization of Metallic Single Wall Carbon Nanotube

Chapter 5: Conclusions and Future Recommendations

Bibliography

[1] W. Steinhogl, G. Schindler, and M. Engelhardt, “Size-dependent resistivity of metallic wires in the mesoscopic range,” Phys. Review B, vol. 66, 075414, 2002.

[2] P.Kapur, J. P. McVittie, and K. C. Saraswat, “Technology and reliability constrained future copper interconnects-part I: Resistance modeling,” Trans. On Electron Devices, vol. 49, No. 4, pp. 590-597, April 2002.

[3] R. H. Havemann, and J. A. Hutchby, “High Performance Interconnecs: An Integration Overview”, Proceedings of the IEEE, vol. 89, No. 5, May 2001

[4] C. Ryu , K.-W. Kwon, A.L.S.Loke, H. Lee, T. Nogami, V. M. Dubin, R.A. Kavari, G. W. Ray, S.S. Wong, “Microstructure and reliability of copper interconnects**”**, Trans. on Electron Devices, vol. 46, pp.1113-1120, June,1999.

[5] H. Cho, K-H Koo, P. Kapur, and K. C. Saraswat, “Performance Comparisons between Cu/Low-k Carbon-Nanotube, and Optics for Future On-Chip Interconnects”, IEEE Electron Device Letters, vol. 29, No. 1, Jan. 2008

[6] Naeemi, A.; Sarvari, R.; Meindl, J.D.; , "Performance comparison between carbon nanotube and copper interconnects for GSI," *Electron Devices Meeting, 2004. IEDM Technical Digest. IEEE International* , vol., no., pp. 699- 702, 13-15 Dec. 2004

[7] Miller, D.A.B.; , "Rationale and challenges for optical interconnects to electronic chips," *Proceedings of the IEEE* , vol.88, no.6, pp.728-749, Jun 2000

[8] C. T. White, and T. N. Todorov, "Carbon nanotube as long ballistic conductors," *Nature*, vol. 393, pp. 240-242, 1998.

[9] G. F. Close, S. Yasuda, B. Paul, S. Fujita, H.-S. P. Wong, "1-GHz integrated circuit with carbon nanotube interconnects and silicon transistors," *Nano Letters*, vol. 8, vo. 2, pp. 706 – 709, Feb. 13, 2008

[10] P. Kapur, and K. C. Saraswat, "Comparisons between electrical and optical interconnects for on-chip signaling," *Intl. Interconnect Technology Conference*, pp. 89-91, 2002.

[11] H. Cho, K. H. Koo, P Kapur, and K. C. Saraswat, "The delay, energy, and bandwidth comparisons between copper, carbon nanotube, and optical interconnects for local and global wiring application," *Intl. Interconnect Technology Conference*, pp. 135, Jun, 2007.

[12] R. Ho, T. Ono, R.-D. Hopkins, A. Chow, J. Schauer, F. Y. Liu, and R. Drost, "High speed and low energy capacitively driven on-chip wires," *J. Solid State Circuits*, vol. 43, no. 1, pp. 52-60, Jan., 2008

[13] E. Mensink, D. Schinkel, E. Klumperink, E. Tuijl, and B. Nauta, "A 0.28pJ/b 2Gb/s/ch transceiver in 90nm CMOS for 10mm on-chip interconnects," *Intl. Solid State Circuits Conference*, pp. 209, Sep., 2003

[14] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, "Digital Integrated Circuits", Prentice Hall, Jan., 2003

[15] Chung-Kuan Cheng, John Lillis, Shen Lin, and Norman Chang, "Interconnect Analysis and Synthesis" Wiley-Interscience, Nov., 1999

[16] W. J. Dolly, and J. W. Poulton, "Digital Systems Engineering" Cambridge University Press, April, 2008

Chapter 2

Performance Comparison Study between Cu/low-k, m-SWCNT Bundle, and Optical Interconnects

2.1 Circuit Parameters Modeling

2.1.1 Implication of Scaling for Modeling of Copper Interconnect

As wire cross-sectional dimension and grain size become comparable to the bulk mean free path of electrons in Cu, two major physical phenomena occur to the current-

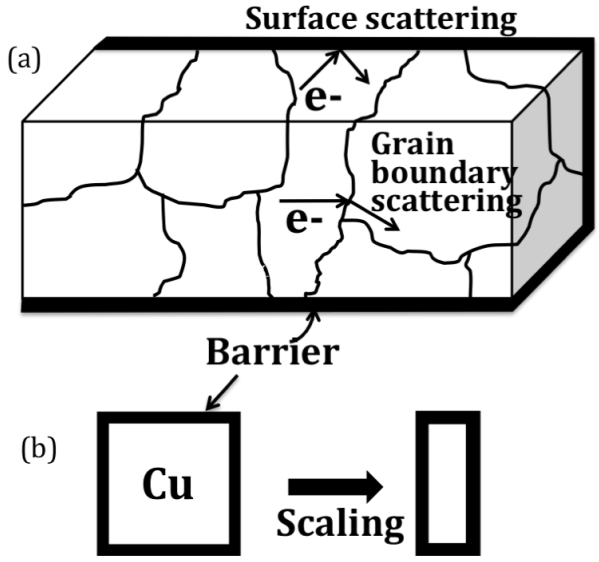

Fig. 2.1 : (a) Schematic illustration of the surface and grain boundary scatterings, and the barrier effect. (b) Impact of scaling on barrier effect. Cu can be scaled while barrier cannot.

carrying electrons in Cu in addition to bulk phonon as illustrated in fig. 2.1(a). One is interface scattering (Eq. (2.1) [1]) and the other is grain boundary scattering (Eq. (2.2) [2]). These increase Cu resistivity more than the ideal bulk resistivity (ρ_o=1.9 Ω·cm) [3]. The Fuchas-Sondheimer model and the theory of Mayadas and Shatzkes quantify these two effects. Eq. (2.1) and (2.2) describe Fuchas-Sondheimer (ρ_{FS}) model and Mayadas-Shatzkes (ρ_{MS}) effect, respectively.

$$\frac{\rho}{\rho_o} = 1 + \frac{3}{4}(1-p)\frac{l_o}{w} \quad (2.1)$$

$$\frac{\rho_o}{\rho_{MS}} = 3\left[\frac{1}{3} - \frac{\alpha}{2} + \alpha^2 - \alpha^3 \ln(1+\frac{1}{\alpha})\right] \quad (2.2)$$

$$\alpha = \frac{l_o}{d}\frac{R}{1-R}$$

Here, ρ_o is the bulk Cu resistivity and p is the electron fraction scattered specularly at the interface (assumed to be 0.6 [1]). p ranges from 0 to 1 and approaches 1 as electrons have more interface scattering. w is the wire width (from ITRS [4]), and l_o is the bulk mean free path of electrons. For Eq. (2.2), R is the reflectivity coefficient representing the electron fraction not scattered at grain boundaries (assumed to be 0.5 [1]). This coefficient also ranges from 0 to 1, approaching 0 as electrons have more grain boundary scatterings. d is the average grain size (d~w). Based on above models, at the 22nm node, Cu resistivity for minimum width wire increases to 5.8μΩ·cm (~3X that of bulk). On the other hand, Cu interconnects typically need a diffusion barrier, which could come in the form of Ta, Ru, and Mg based materials. Because the resistivity of these materials is much higher than that of Cu, in effect, the barrier reduces the useful interconnect cross-section. One way to capture this effect is to

define the problem in terms of an increased effective resistivity, which would be applicable to the original cross-sectional area. Fig. 2.1(b) shows that scaling exacerbates the barrier problem as barrier thickness does not scale proportionately to the aggressive interconnect cross-sectional scaling. This, in turn, results in an increase in the effective resistivity with the technology node. Three dashed curves in Fig. 2.2 quantify this effect. The effective resistivity here also captures both the surface and grain boundary scatterings. It is clear that with technology scaling, effective resistivity increases dramatically. Further, lowering the barrier thickness (ALD: 3nm vs. 1nm) has a big impact on effective resistivity [3].

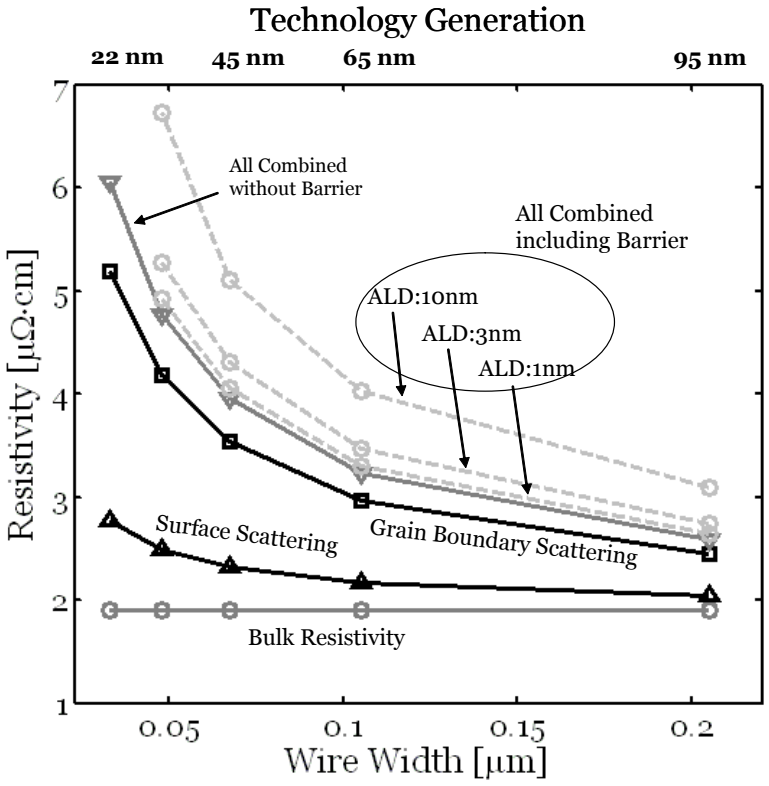

Fig. 2.2 : Cu resistivity in terms of wire width taking into account the surface and grain boundary scattering and barrier effect. The barrier layer is assumed to be uniformly deposited, e.g., using atomic layer deposition (ALD).

Fig. 2.3 shows the scaling trend of the electrical wire dimensions and its impact on the bandwidth and power budget of the wire.

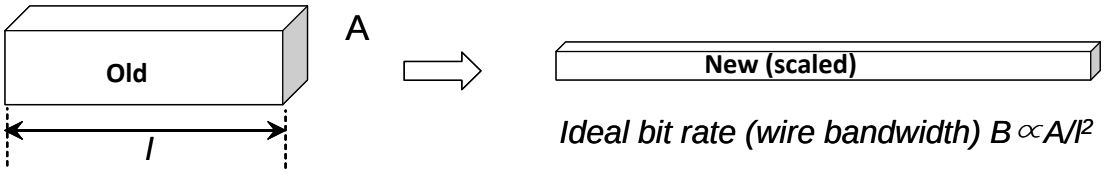

Fig. 2.3 : The impact of interconnect scaling. Scaled wire with lower A and longer l has higher resistance resulting in higher delay, increased power, and reduced bandwidth

The increase in wire length (*l*) in addition to the reduction in cross-section area (*A*) further exacerbates wire resistance, subsequently limiting the signal rise time and the bandwidth. This can be well understood from the simple relationship between the ideal bit rate (*B*), and the cross-sectional area and the wire length in Fig. 2.3. Typically, the wire buffering with multiple repeaters mitigates the bandwidth shortfall. However, it swallows a significant portion of the power budget (this will be explained more in detail in part *B*). Thus, for the electrical interconnect, it becomes more difficult to meet the bandwidth requirement and power budget simultaneously [5].

2.1.2 Conventional Interconnect Circuit and Its Performance Limit

In general, the interconnect performance has been optimized in terms of delay such that repeater distance (l) and size (w) are leveraged for equal delay from the repeater and wire [6][7].

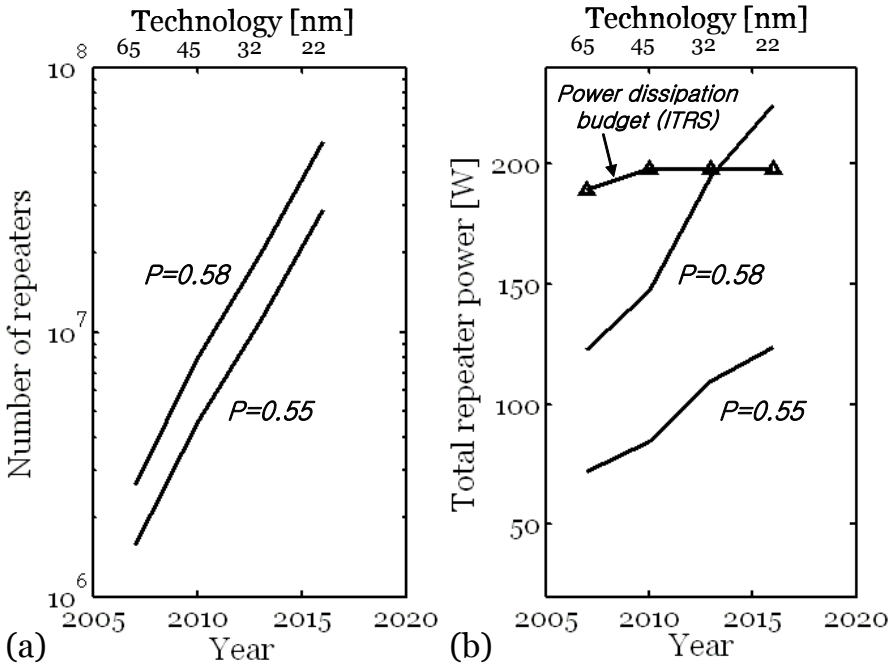

Fig. 2.4 (a) Number of global wire repeaters vs. wire dimensions, (b) total energy per bit (pJ) in global wire vs. wire dimension.

However, this approach suffers from undesirable problems. While this methodology can significantly reduce delay, it results in inordinate power consumption. This excessive power expenditure has a close relationship with increasing number of repeaters, which is the consequence of meeting the small delay requirement [8]. Fig. 2.4 shows this phenomenon, assuming a minimum power

dissipation budget defined by ITRS [4]. Furthermore, a massive number of repeaters leads to the via blockage problem, which severely reduces the wiring efficiency [9][10]. On the other hand, if the global wires do not scale at all, staying at a constant pitch to avoid these problems, it will cause a rapid increase in the number of metal layers [10].

2.1.3 Modeling Parameter for CNT bundles

CNTs have attracted great attention because of their interesting physical and electrical properties [11]. Their near one-dimensional shape supports ballistic transport, making them potentially useful in many applications, such as transistors, sensors, and interconnects. In addition, *CNTs* offer great mechanical strength due to their strong sigma (σ) bonds between neighboring carbon atoms. These excellent physical properties have been proven theoretically and experimentally by intensive research for more than a decade. *CNTs* can be categorized as semiconducting or metallic depending on their chiral configurations. Fig. 2.5 shows the band structures of armchair and zigzag *CNTs* corresponding to metallic and semiconducting nanotubes, respectively. For the interconnect application, we only consider metallic carbon nanotubes. Depending on the shape, *CNTs* can be categorized as *SWCNT* or multi-walled carbon nanotube (*MWCNT*). An *SWCNT* is constructed by wrapping a graphene layer into a cylindrical shape. Its diameter ranges from 0.4nm to 4nm. Its typical diameter is around 1nm. An MWCNT is formed of multiple layers of *SWCNTs*

with different diameters. Its diameter ranges from 10nm to 100nm. Fig. 2.6 shows the difference between *SWCNT* and *MWCNT*.

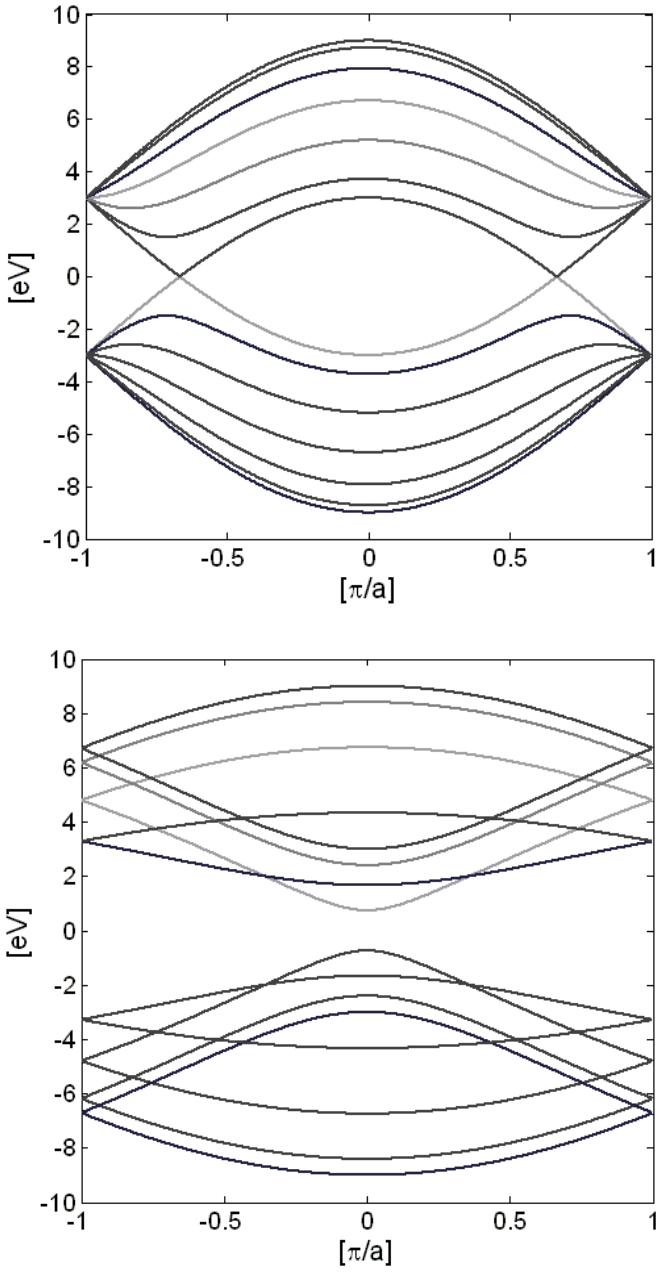

Fig. 2.5: E-k band diagram of *SWCNT* with chirality of (a) Armchair (6,6) SWCNT, (b) Zigzag (7, 0) SWCNT.

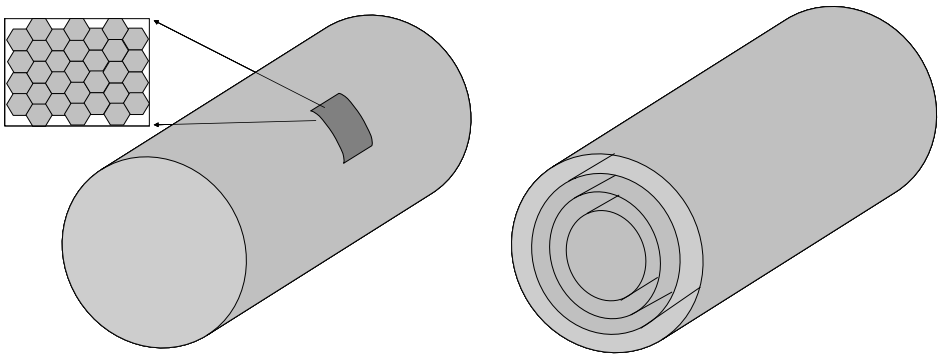

Fig. 2.6 : three dimensional illustration of (a) *SWCNT*, (b) *MWCNT*.

2.1.3.1 Resistance (Conductance) of m-SWCNT and Bundle

The resistance of a *CNT* bundle depends on the total wire cross section area and the fractional packing density (*PD*) of metallic *CNTs* within it [12]. Henceforth, *PD* is assumed to be 33% unless specified otherwise. This is because given an equal probability of the wrapping vectors, about 1/3rd and 2/3rd *SWCNTs* turn out to be metallic and semiconducting, respectively [13].

For an *SWCNT*, in the absence of any type of scattering, the maximum quantum conductance is limited as shown in Eq. (2.3)

$$G_{SWCNT} = 4q^2/\mathrm{h}\ G_{SWCNT} = \frac{2q^2}{\pi\eta} \quad (2.3)$$

where $\hbar$ is the Planck constant and q is a charge of one electron. The multiple four accounts for the two channels due to electron spin and another two channels due to sub-lattice degeneracy. Thus, the quantum resistance of an *SWCNT* is 6.45 kΩ. This

resistance is fairly large to allow use in interconnect applications. One way to get around this problem is to use a bundle of *SWCNTs*, as illustrated in Fig 2.7.

The number of *SWCNTs* in a bundle is given by

$$n_w = \frac{w - d_t}{x}, n_h = \frac{h - d_t}{\left(\sqrt{3}/2\right)x} + 1$$
$$n_{CNT} = n_w n_h - n_h/2 \quad \text{if } n_h \text{ is even} \qquad (2.4)$$
$$= n_w n_h - \frac{n_h - 1}{2} \quad \text{if } n_h \text{ is odd}$$

where n_w is the number of "columns," n_h the number of "rows," and n_{CNT} the number of *SWCNTs* in a bundle.

The quantum resistance of an SWCNT (R_Q) is 6.45KΩ, as explained earlier. However, in practice, electrons do get scattered by either defects or phonons, hence, possess a finite mean free path (l_o). Therefore, a linear dependence of resistance on length, based on recent results on high quality (long mean free path) *CNTs*, can be defined as [14]:

$$R_{SWCNT} = R_Q\left(1 + \frac{l}{l_o}\right) \text{ and } R_w = \frac{R_Q}{n_{CNT}}\left(1 + \frac{l}{l_o}\right) \qquad (2.5)$$

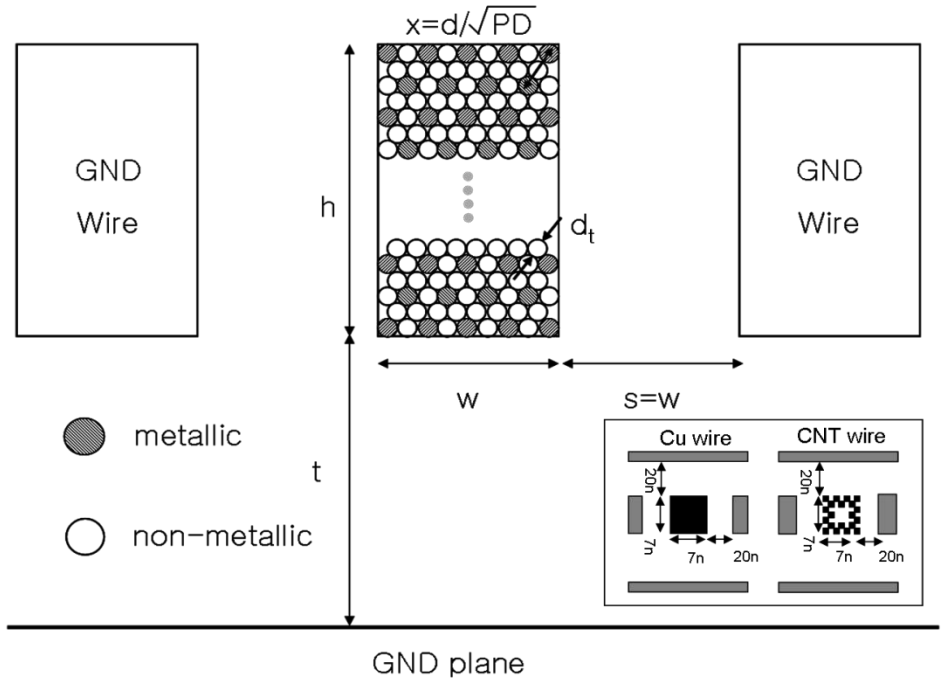

Fig. 2.7: Schematic of the interconnect geometry with *CNT* bundles, taking packing density (*PD*) into account. *x* is the distance between the closest metallic *SWCNTs*, d_t is the diameter of an *SWCNT*. (inset) *CNT* and *Cu* interconnect cross-section geometry for Fastcap simulation.

where R_{SWCNT} and R_w are the resistances of one *SWCNT* and a bundle respectively. *l* is the wire length and l_o is the electron mean free path. l_o is proportional to the *SWCNT* diameter with theoretically and experimentally derived proportionality constants of 2.8μm/nm (henceforth, ideal model) [15] and 0.9μm/nm (henceforth, practical model) [14], respectively. Thus, for a 1nm-diameter *SWCNT*, ideal and practical models yield an l_o of 2.8μm and 0.9μm, respectively. These values are valid at a small voltage drop across the *CNT* interconnect (< 0.16V) [16]. Operation at higher voltage drop can degrade performance. An additional resistance component in the form of contact resistance has recently been shown to be reduced down to a few kΩ per tube [16][17]. The total resistance of a *CNT* bundle is simply all the resistance components of a

single tube divided by the number of metallic tubes in the bundle (dictated by *PD*), as shown in Eq. (2.5).

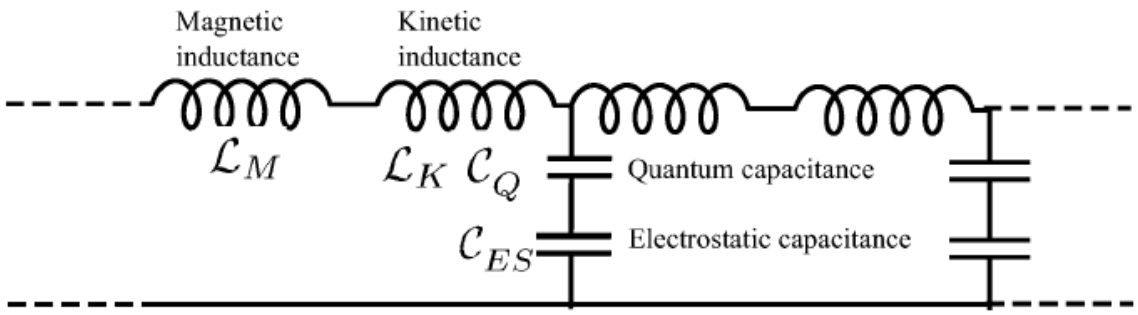

Fig. 2.8 transmission line LC components of *SWCNT*.

It is very important to investigate the transmission line components of SWCNT to understand its applicability for an interconnect. Figure 2.8 shows a transmission line *LC* equivalent circuit of *SWCNT*.

2.1.3.2 Capacitance of m-SWCNT

An *SWCNT* has two capacitance components. One is electrostatic capacitance (C_E) purely determined by *SWCNT's* geometric distance to the ground plane. The other component is quantum capacitance (C_Q) [18] due to a reduced 2-D density of states for electrons. To add an electron in an *SWCNT*, one must add it at an available quantum state above the Fermi energy (E_F) due to Pauli's exclusion principle. C_E and C_Q can be described by Eqs. (2.6) and (2.7), respectively.

$$C_E = \frac{2\pi\varepsilon}{\cosh^{-1}(2t/d_t)} \approx \frac{2\pi\varepsilon}{\ln(t/d_t)} \quad (2.6)$$

$$C_Q = \frac{e^2}{\pi\hbar v_f} \quad (2.7)$$

where ε is the permittivity of the dielectric between a wire and a ground plane, t is the distance between the ground and nanotube, d_t is the diameter of a single nanotube, $\hbar$ is the Planck constant, and v_f is the Fermi velocity of an *SWCNT*. For d_t = 1nm, $t = 4h$ = 340nm, and ε_r (relative permittivity for the 22nm technology node) = 2.0, the electrostatic capacitance is about 190fF/mm. C_Q is around 100fF/mm under the same conditions as above, which is of the same order of magnitude as its electrostatic counterpart.

For a single tube, C_Q is comparable to C_E. However, in a bundle, the quantum components are added in series, making the bundle's total C_Q negligible compared to its C_E [19]. We simulated two geometries for capacitance, corresponding to CNT and Cu using FastCap (3D field solver) [20]. Fig. 2.7 (inset) shows the simulated geometries, with *CNT* exhibiting more roughness due to its bundled nature. The capacitance of the two geometries was within 4% agreement with the one presented in [12]. Thus, the total *CNT* bundle capacitance is approximately the same as the electrostatic capacitance of the *Cu* wire.

2.1.3.1 Inductance of m-SWCNT

The importance of inductance (RC vs. RLC model) depends on the relative magnitude of the inductive reactance and the wire resistance. *SWCNT* has magnetic inductance (L_{mag}) of 1pH/um as calculated in Eq. (2.8)

$$L_M = \frac{\mu}{2\pi}\cosh^{-1}\left(\frac{2t}{d_t}\right) \approx \frac{\mu}{2\pi}\ln\left(\frac{t}{d_t}\right) = 1pH/\mu m \quad \text{---- (2.8)}$$

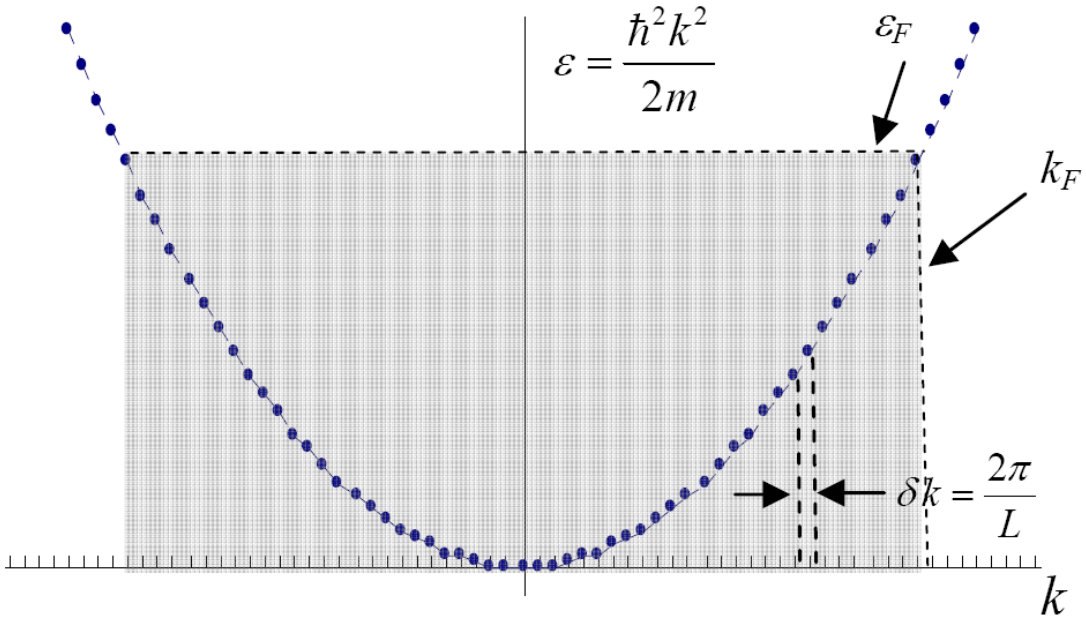

Fig. 2.9 E-k diagram of *m-SWCNT*

In addition to this, SWCNT has an extra kinetic inductance ($L_{kin,}$). $L_{kin,}$ results from a higher electron kinetic energy because of a lower density of states in *CNTs* as illustrated in Fig 2.9 [18].

$$E = \frac{1}{2} L_k I^2$$

$$L_k = \frac{h}{2e^2 v_F} = 16nH/\mu m \quad \text{--- (2.9)}$$

Per tube it is about four orders of magnitude higher than L_{mag} as shown in Eq (2.9). To summarize, all frequency dependent reactive components of *SWCNT* are listed and compared in table 2.1. The additional main advantage of using a bundle as in Fig 2.7 is that unnecessarily high reactive elements can be reduced. For instance, in a bundle, the total bundle L_{kin} reduces dramatically because *CNTs* are in parallel [19]. Total magnetic inductance (L_{mag}), which is the sum of self (L_{self}) and the mutual inductance (L_{mut}) between the tubes, is relatively constant with wire width. This is because at interconnect dimensions of interest, the L_{self} component of L_{mag} becomes negligible (L_{self} adds in parallel and is proportionately reduced). While, L_{mut} is relatively constant beyond a certain width, as the impact of mutual inductance from *SWCNT* further than a certain distance gets weaker. This value converges to L_{self}, when solved using a matrix approach described in ref [21]. Thus, the total inductance of a *CNT* bundle is given by

$$L_{tot} = \frac{L_{kin}}{n_{cnt}} + L_{mag},$$

$$Where,$$

$$L_{mag} = (L_{mut} + \frac{L_{self}}{n_{cnt}}) \approx L_{self} \quad (2.10)$$

Here, L_{mag} obtained using model in [21] is ~ 1.5nH/mm.

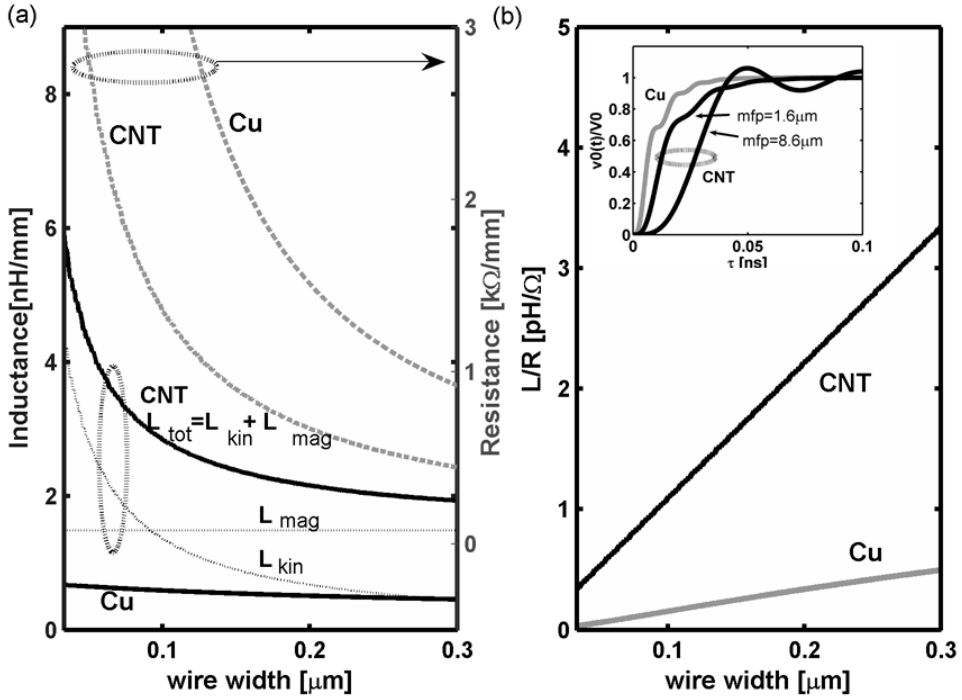

Fig. 2.10.(a) Inductance and resistance of *Cu* and *CNT* vs. wire width. Wire width is in global interconnect regime. *PD* is assumed to be 33%. l_0 is 1.6μm. Total inductance consists of combination of kinetic inductance and magnetic inductance (b) Inductance to resistance ratio as a function of the wire width. (b. inset) step response of *Cu* and *CNT* wire with optimized spacing (*h*). Wider interconnect implies more ripples in time domain response due to higher transmission line effect.

Fig. 2.10 a plots L_{tot} of a *CNT* bundle along with its components. Smaller widths render a larger L_{tot} because of an increase in L_{kin}. *Cu* L_{tot} is lower than *CNT* L_{tot} for all widths. In addition, Fig 2.10 a shows that the *CNT* bundle resistance is also lower than that of *Cu* due to a longer mean free path. The above results exhibit approximately a 6X larger inductance to resistance ratio for *CNT*-bundle compared to *Cu* wires, insinuating a more pronounced impact of inductance in the case of *CNT* bundle. A simulated step response of a repeated *CNT* wire indeed shows a significantly under damped high-frequency response (Fig. 2.10 (b) inset) than *Cu*, corroborating the

importance of inductance in *CNTs*. Fig. 2.10 b also shows that the L/R ratio is higher for larger widths. Thus, a full RLC model is imperative for *CNT* bundles especially for larger width, global wires. Inductance can be ignored for smaller width, local wires.

All reactive component values of a single *m-SWCNT* are summarized in table 2.1

	Electrostatic / Magnetic Components	Quantum Mechanical Components
Inductance	L_M = 1pH/um	L_K =16 nH/um
Capacitance	C_E = 190aF/um	C_Q = 100aF/um

Table 2.1 This table summarizes two different reactive components of *SWCNT*. *SWCNT* is assumed to have 1nm diameter

2.1.4 Secondary Effects in SWCNT

2.1.4.1 Contact Resistance

Contact resistance arises mainly due to scattering at the contacts when interfaces are poorly connected. Some reports indicate that it can be of the order of hundreds of kilohms.[22] This effect becomes more prominent when the diameter of nanotubes is smaller than 1nm because it may form a poor bonding between *CNTs* and metal. In addition, metallic *CNTs* tend to have non-negligible bandgap opening as the diameter is smaller than 1nm. This may give rise to hundreds of kilohms of contact resistance. However, prior works hasve reported that contact resistances of nanotubes

with a diameter larger than 1nm are on the order of only few kilohms or even hundred ohms .[23]. This indicates that the contact resistance can be lowered to such a level that it can be neglected compared to the basic quantum resistance (~6.45kΩ).

2.1.4.2 Impact of High Bias Voltage

CNTs are one-dimensional quantum wires and, in general, are considered to provide ballistic transport. However, electrons in these quantum wires tend to get backscattered as the length becomes longer. At small bias voltages, defects and acoustic phonons are the only scatterers. In this case, the electron mean free path (MFP) can be as large as 1.6um in high-quality nanotubes. However, at higher bias voltages, optical and zone boundary phonons, which have energies around $\hbar\Omega \approx 0.16$ eV, contribute to electron backscattering [16]. Once an electron with energy E finds an available state with energy E- ħΩ, it emits a phonon with the energy of ħΩ. An electron should be accelerated by electric field E to the length of $l_\Omega = \hbar\Omega/qE$ in order to gain this energy. Once it achieves this energy, it travels on average l_o = 30nm before it scatters and emits an optical or zone-boundary phonon. The effective MFP, l_{eff}, can be described by

$$\frac{1}{l_{eff}} = \frac{1}{l_e} + \frac{1}{l_\Omega + l_o} \quad (2.11)$$

where l_e is the low-bias MFP. The resistance of an *SWCNT* is given by

$$R = R_c + R_Q\left(1 + \frac{L}{l_{eff}}\right) \quad (2.12)$$

where R_c is the contact resistance, R_Q is the fundamental quantum resistance and L is the tube length. The electric field across a tube will vary depending on the applied bias and should be written in an integral form as [24]:

$$R = R_c + R_Q + \left(R_Q \frac{l}{l_e} + \int_{x=0}^{x=l} \frac{dx}{\frac{I_0}{E(x)} + \frac{l_0}{R_Q}} \right) \quad (2.13)$$

where $E(x)$ is the position-dependent electric field, and I_0 =25uA.

2.1.5 Graphene Nanotibbons (GNRs) Interconnect

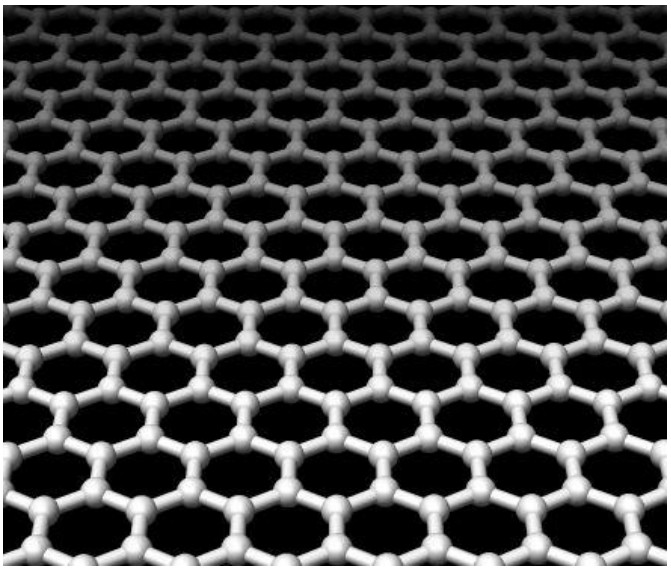

Fig. 2.11 Graphical illustration of 2-D Graphene nanoribbon (GNR).

A graphene sheet is an ideal two-dimensional carbon honeycomb structure. Graphene nanoribbons, abbreviated as *GNRs*, are edge-terminated graphene sheets . They are just equivalent to unrolled *SWCNTs. GNRs* share most of the physical and electrical characteristics with *CNTs* such that they become semiconducting or metallic

depending on the chirality of *GNRs'* edge [25]. *GNRs* can be fabricated in a more controllable process, such as optical lithography, while *CNTs*' growth results in a random chiral distribution. Thus, it is necessary to look at *GNRs'* performance as an interconnect application. To briefly look at the conductance model of a *GNR* interconnect, the MFP of a *GNR* should be analyzed. l_n is the distance that electrons in the n^{th} mode move forward until they hit one of the *GNR*'s side edges. This can be described as

$$l_n = \frac{k_{\parallel}}{k_{\perp}} W = W\sqrt{\left(\frac{E_F / \Delta E}{n+\beta}\right)^2 - 1} \quad (2.14)$$

where W is the width of *GNR*, and $k_{\parallel}$ and $k_{\perp}$ are the wave vectors in transport and non-transport direction, respectively. E_F is the Fermi level. ΔE is the energy gap between subbands in *GNR*. β is zero in a metallic *GNR*. Then the conductance of *GNRs* can be described by the total number of transmission modes and Mattiessen's rule, as shown in Eq. (2.11). The number of transmission modes can be determined by counting all subbands below the Fermi energy level:

$$G = \frac{q^2}{\pi\eta}\sum_n \frac{1}{1 + L(1/l_D + 1/l_n)} \quad (2.15)$$

where l_D is the MFP caused by the acoustic phonon and defects, as in *CNTs*. If we use more mathematical approximations, (2.15) can be simplified further as (2.16):

$$G = \frac{q^2 l_D}{\pi \eta L} + 1.5\left(\frac{q^2 W^{2.5}}{\pi \eta L}\right)\left(\frac{E_F}{\pi \eta v_F}\right)^{1.5} \quad (2.16)$$

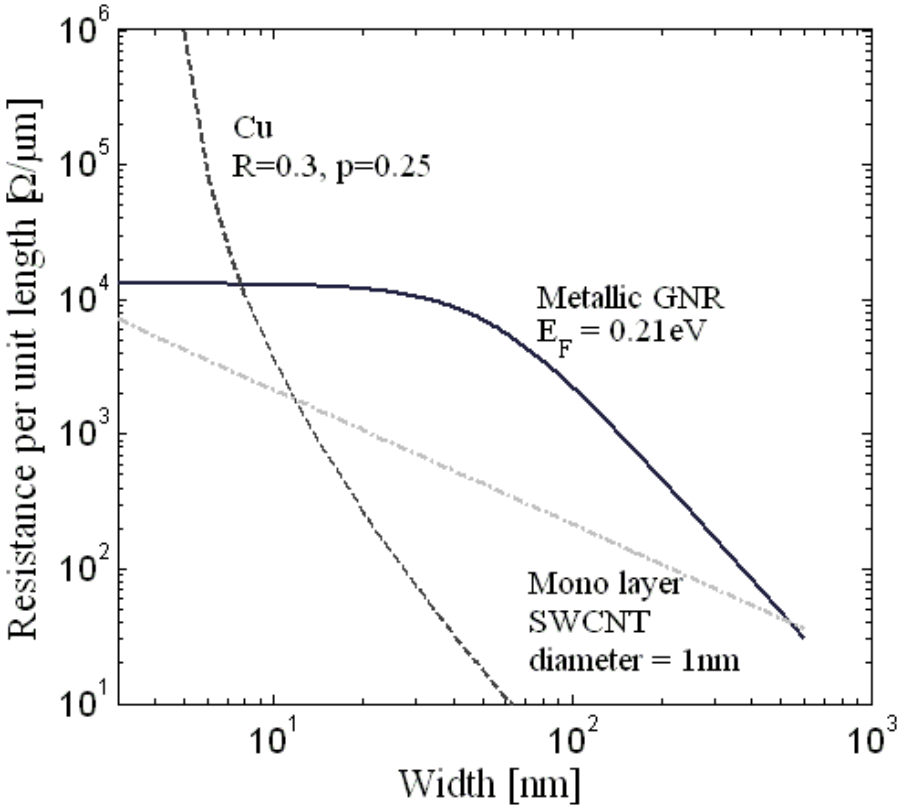

Fig. 2.12 Resistance comparison between *GNR*, mono-layer *SWCNT*, and *Cu*. The Fermi-level is assumed to be 0.21eV, as reported in an experimental result [25].

In the graph in Fig. 2.12, a metallic *GNR* cannot outperform a *Cu* interconnect until the width is lower than 7nm. These results explain that diffusive edge scattering significantly limits the performance of the *GNR* interconnect. In addition, the resistance of *GNR* is higher than that of the mono-layer *SWCNT* for all ranges of wire width. This is an obvious result because physical properties of a *GNR* are similar to those of a *CNT* and only one *GNR* layer should be able to compete with a mono-layer

bundle of *CNTs*. However, if we can use a multilayered *GNR* interconnect, it would improve the performance.

2.2 Performance Comparison Study

2.2.1 Local Interconnects: CNT Bundle vs. Cu

Previous work on local wire performance comparisons have reported conflicting claims. Srivastava et. al. show much worse *CNT* bundle performance [26], whereas, Naeemi et. al. show a mono/bi-layer of *CNT* outperforming *Cu* wires [12][19]. Our simulations do not agree with [26], with the primary source of discrepancy being that our *CNT* bundle capacitance comes out to be much smaller. Moreover, in contrast with **Error! Reference source not found.**, we have compared a *CNT* bundle and *Cu* not only for TRS dictated dimensions, but also optimized the *CNT*-bundle geometry to leverage its superior electromigration (*EM*) properties, and as a result extract additional performance from this attribute.

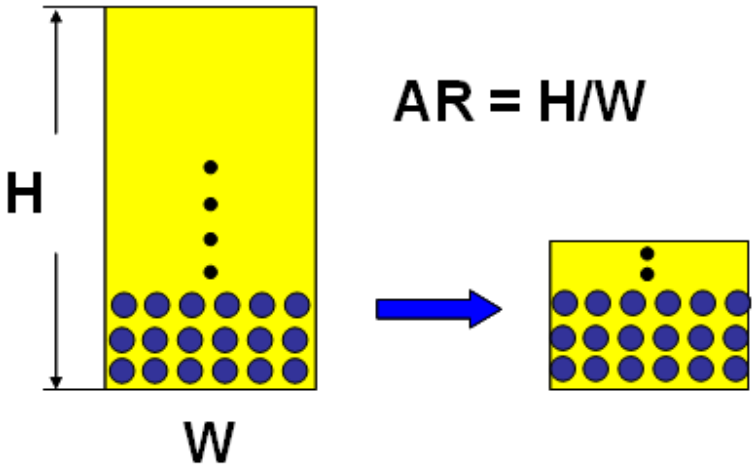

Fig. 2.13 Illustration of aspect ratio tuning in local interconnect cross-section. Black circles are *SWCNTs*

In this work, we investigated the impact of leveraging the aspect ratio (AR) on the latency in local signaling with *SWCNT*. Fig 2.13 illustrates the cross section of *SWCNT* local interconnect with different AR. The basic motivation of reducing AR is two fold. First, the driver resistance and the wire capacitance overpower the wire resistance in the Elmore delay in a local interconnect. Secondly, *SWCNT* has higher mechanical strength than Cu, potentially providing much better *EM* immunity. Fig. 2.14 a shows that for minimum width and both short and long wires, *CNT* bundle exhibits a lower latency than *Cu* at all aspect ratios (AR). In addition, the difference in latency between the two technologies is smaller for shorter wires and higher ARs because these conditions render the wire resistance (R_w) lower than the driver resistance (R_{dr}). A dominant R_{dr} obviates the *CNT* R_w advantage. Fig. 2.14 (a) also shows that both *Cu* and *CNT* wires exhibit a minimum in latency with AR. The initial reduction in AR lowers delay because at large ARs, a low R_w results in the delay being dominated by the wire capacitance (C_w) and R_{dr} product; further, C_w reduces with AR. However, below a certain AR, R_w becomes dominant over the R_{dr}, and the delay is

now dictated by R_w and C_w product. R_w rises with smaller AR and C_w cannot reduce as fast as the increase in R_w because its reduction is limited by a constant inter-level dielectric (ILD) and the fringe component. Fig. 2.14 (b) plots wire resistance (*R*) and capacitance (*C*) ratio between *CNT* and *Cu* as a function of AR.

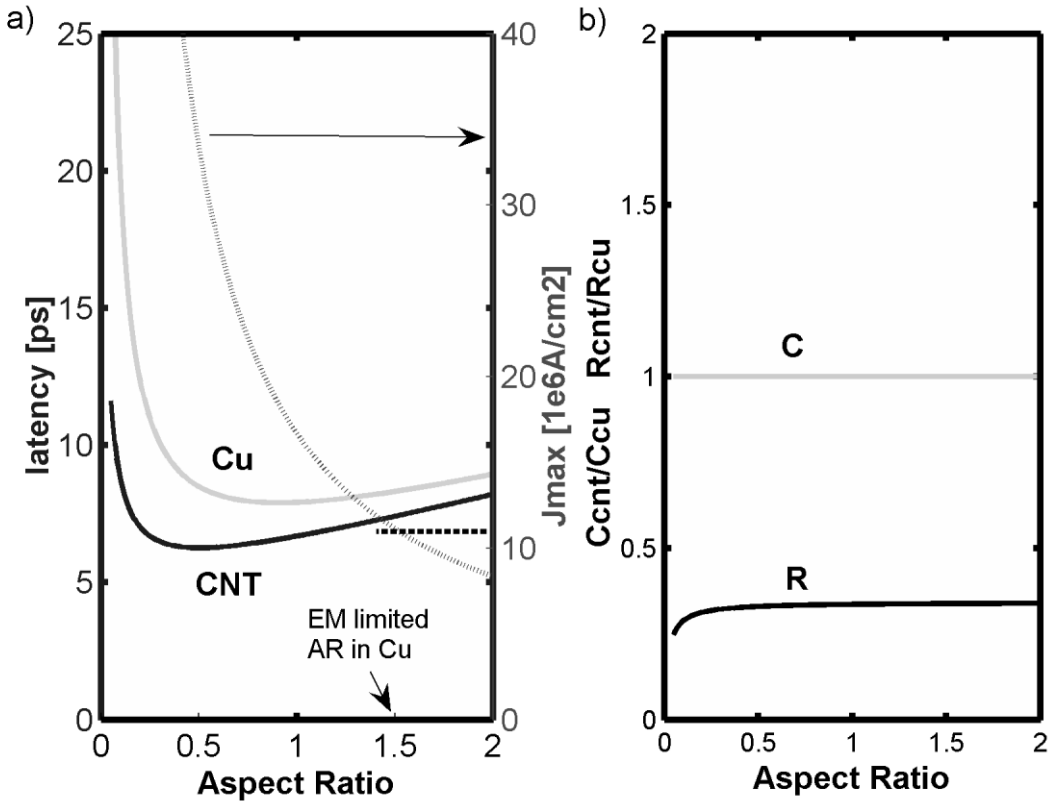

Fig. 2.14 Local interconnect delay vs. aspect ratio (higher to width ratio) for 10μm wire lengths (left y-axis). Current density vs. aspect ratio (right y-axis). Wire width corresponded to minimum at 22nm from ITRS [4]. The simulations assume a driver resistance of 3kΩ, the *CNT* mean free path of 0.9μm (conservative), and a contact resistance and packing density of 10kΩ and 33%, respectively. (b) Wire resistance (*R*) and capacitance (*C*) ratio between *CNT* and *Cu*.

Although, both technologies exhibit an optimum AR minimizing delay, in practice, only *CNTs* can be operated at the optimum because of *EM* considerations. Fig. 2.14 (a) (right y-axis) shows the increase in the wire current density with smaller AR. A

maximum allowed current density of 14.7×10^6 A/cm^2 in *Cu*, limits its minimum AR to about 1.5 for the ITRS dictated minimum widths. While, a single CNT tube can carry a current density up to 10^9 A/cm^2. Thus, *CNTs* can be operated at their optimum AR, pointing to a different geometry than for *Cu*.

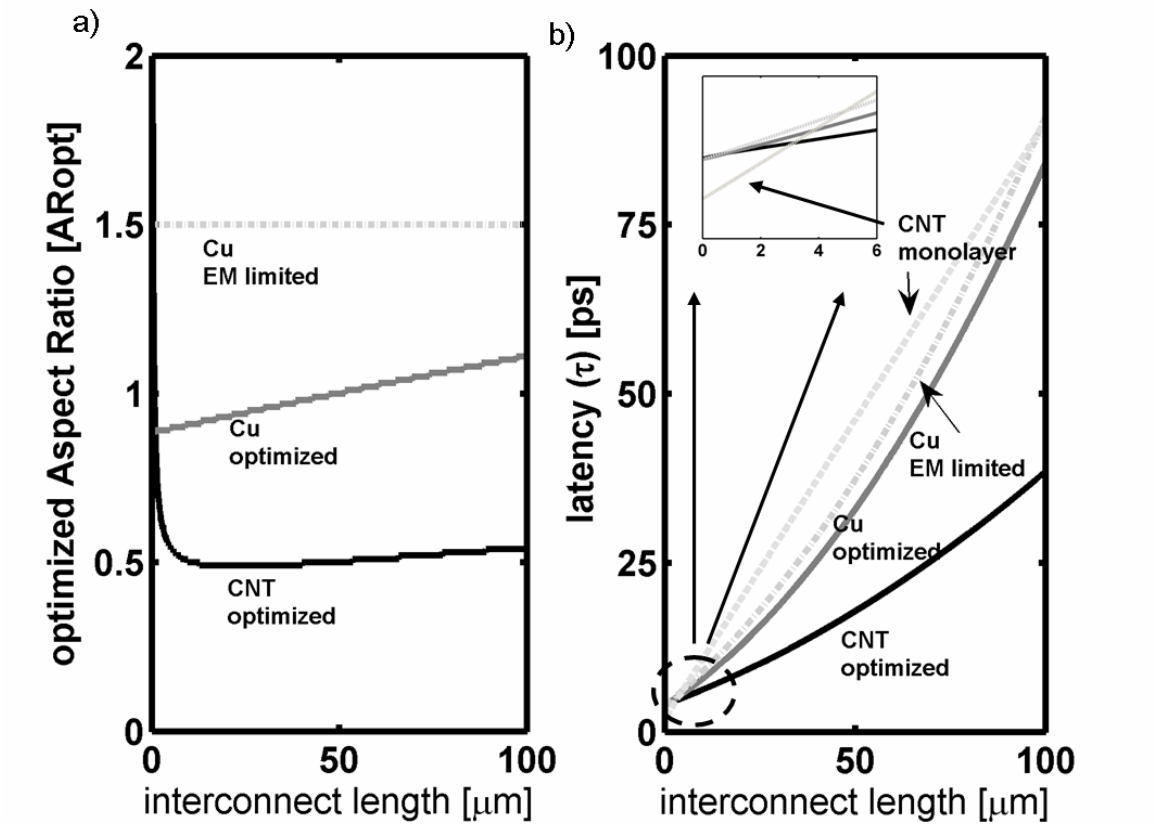

Fig. 2.15 (a) The plot of optimum Aspect Ratio (AR_{opt}) for minimum latency as a function of length. Dotted line is *EM* limited AR for Cu. (b) The plot of latency as a function of length. Four different curves are shown: two for CNT (τ with AR_{opt} and τ with monolayer) and two for Cu (τ with EM limited AR and τ with AR_{opt}). Inset: zoomed-in view from 0μm to 6μm of local interconnect length

Fig.2.15-(a) explicitly shows these geometry differences. The *Cu* optimum aspect ratio (AR_{opt}) is lower than what is allowed by the *Cu EM* limit, and is higher than *CNT* AR_{opt}. In addition, AR_{opt} increases with wire length for both *Cu* and *CNT*. This is

because AR_{opt} represents a point where R_{dr} and R_w are comparable, and longer wires (because of their larger resistance) reach comparable values to R_{dr} at larger cross section area (higher AR). Fig. 2.15-(b) explicitly plots the latency vs. length for four cases. This includes the lowest possible latency using AR_{opt} for both *CNT*-bundle and *Cu*, as well as the latency of a monolayer *SWCNT*, and *EM* limited *Cu*. Optimized *CNT*-bundle shows better latency than optimized *Cu* for all local length scales. This advantage increases to about 2X when we compare optimized *CNT* (leveraging its *EM* robustness) with the EM limited *Cu*. Finally, we observe (inset of Fig. 2.15-(b)) that a monolayer *CNT* gives the smallest delay for very short local interconnects as it is dominated by the time of flight rather than RC time constants. This particular result corroborates with [19]. In addition to latency, a smaller AR_{opt} for *CNT* bundle would also bring a dramatic power reduction compared to *Cu* because of a lower capacitance.

2.2.2 Global and Semi-Global Interconnects

2.2.2.1 Cu/CNT and Optical Global Wire Circuit Models

Fig. 2.16 (a) illustrates the schematic representation of a buffered interconnect. If we consider an interconnect of total length *L*, it is buffered at length *l*. In this section, we present the methodology for optimally buffering the interconnect in order to minimize delay. Fig. 2.16(b) shows a distributed RC network, indicating one segment of the repeated interconnect with length *l*. V_{tr} is the voltage source at the input stage.

R_{tr} is the driver resistance which has dependence on the transistor size, C_p is the output parasitic capacitance of the driver, and C_L is the load capacitance of the receiving end, and r and c are the interconnect resistance and capacitance per unit length, respectively. The segment delay of this interconnect, τ_0, is given by [27][28]

$$\tau_0 = bR_{tr}(C_L + C_P) + b(cR_{tr} + rC_L)l + arcl^2 \quad (2.17)$$

where a and b are the switching-dependent parameters. For instance, if the delay is measured at 50% of total voltage swing, a=0.4, and b=0.7. r_0 is the resistance of a minimum-sized driver. Henceforth, R_{tr} can be defined as r_0/s. c_0 and c_p are the input and parasitic output capacitances of a minimum-sized driver, respectively. Thus, C_p and C_L are proportional to the size of a driver as $C_p=sc_p$, and $C_L=sc_0$. If the total interconnect length L is divided into n segments of length $l=L/n$, the delay with overall length of L, τ_d, is given by

$$\tau_d = \frac{L}{l}b(x)r_0(c_0 + c_P) + b(x)(c\frac{r_0}{s} + src_0)L + a(x)rclL \quad (2.18)$$

Then, the optimum values l_{opt} and s_{opt}, and the delay of the optimally-buffered interconnect with total length L, τ_d, can be obtained as

$$l_{opt} = 1.32\sqrt{\frac{r_0(c_0 + c_p)}{rc}}$$

$$s_{opt} = \sqrt{\frac{r_0 c}{r c_0}}$$

$$\tau_d = 2.049 L \sqrt{r c t_{FO1}} \quad (2.19)$$

where $t_{FO1}=2r_0c_0$ (generally $c_0 = c_p$), representing the delay of an inverter with a fanout load of one (*FO1*).

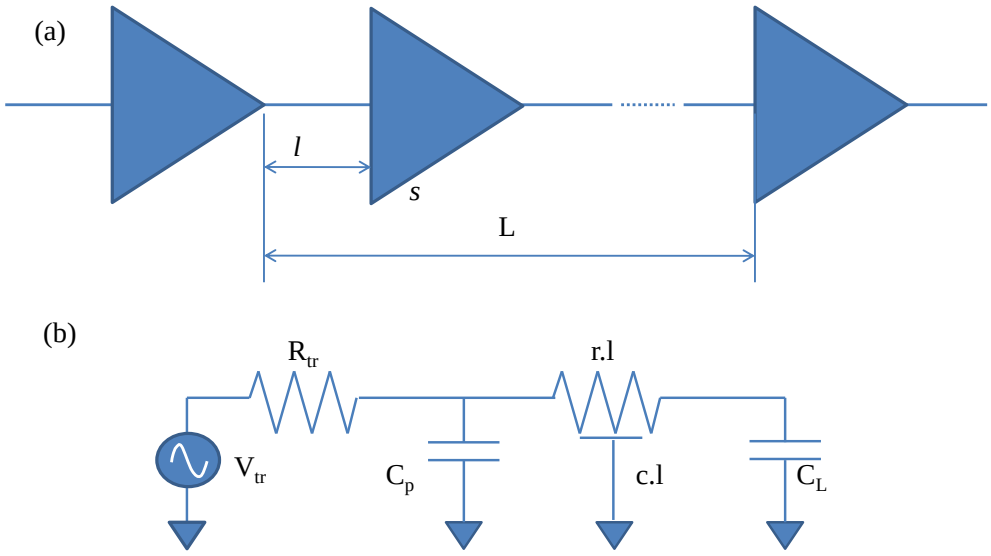

Fig. 2.16 (a) Schematic of optimally buffered interconnect. The total length is *L* and *l* is the optimal distance between repeaters to minimize delay. *S* refers to the optimal size of the input transistor. Each repeater has a fanout of one (*FO1*). (b) The equivalent circuit of one segment with *l* and *s*.

The wire capacitance is modeled as follows [27],

$$C_w = \varepsilon \left[\begin{array}{l} 1.15\frac{w}{t} + 2.80\left(\frac{h}{t}\right)^{0.222} \\ + \left(0.66\frac{w}{t} + 1.66\frac{h}{t} - 0.14\left(\frac{h}{t}\right)^{0.222} \right) \cdot \left(\frac{t}{s}\right)^{1.34} \end{array} \right] \qquad (2.20)$$

Where, ε is the dielectric permittivity, *s* is the inter-wire spacing (assumed *s*=*w*), *h* is wire height (*h*=*w*×aspect ratio), and *t* is inter-metal layer spacing (assumed *t*=*4h*). For global wires, we included inductance, given by Eq. (2.21) [29]. At 22nm technology node, this value is about 0.5nH/mm.

$$L_w = 2\times 10^{-7} l\left(\ln\frac{2l}{w+h} + 0.5 + \frac{w+h}{3l} \right) \qquad (2.21)$$

We use RLC models for global/semiglobal Cu and CNT wires and include repeater insertion for delay reduction (Fig. 2.16). The values of *R*, *C*, and *L* were as discussed in section 2.1, Eq. (2.20), and Eq. (2.21). Unlike the RC wire model, an RLC model does not have a closed form solution optimizing repeater size (*k*) and its spacing (*h*) for minimum delay. Thus, we optimize these parameters using fourth-order Páde expansion and Newton-Raphson numerical method [30].

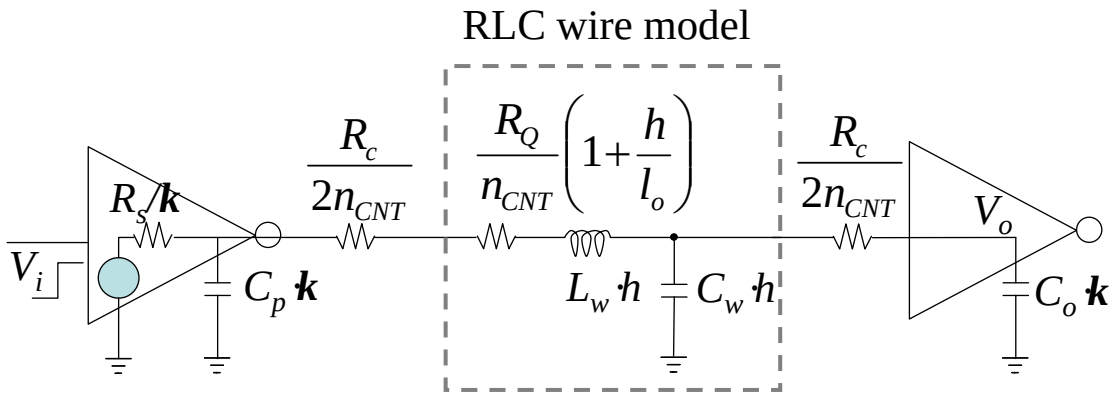

Fig. 2.17 Equivalent circuit model of a repeater segment for *CNTs*. *k* is optimized repeater area. *h* is optimized wire length per repeater. R_s is inverter output resistance. R_c and R_Q are the contact resistance and quantum resistance of *SWCNT*, respectively. L_w and C_w are the inductance and the capacitance of *CNT* bundle, respectively. C_p and C_0 are the input capacitance and the output parasitic capacitance of inverter, respectively.

2.2.2.2 Impact of Inductance on Delay and Power Performance

Fig.2.18 depicts the latency of the *CNT* bundle and the *Cu* wire as a function of wire width. The RC and RLC curves for *Cu* overlap indicating that inductance effect is not significant. The *CNT* bundle has curves corresponding to two different electron mean free path (MFP), with each having both the RC and RLC models. The delay significantly increases with the decrease in cross sectional area for all curves even with repeaters. Also, there is a non-negligible difference between the RC and RLC model for the CNT bundle. The RC delay clearly overestimates the *CNT* performance, with the discrepancy becoming especially serious for widths higher than 60nm (MFP=3.8μm). This width is well within the design window for global wires since it is

approximately twice the minimum wire width for 22nm technology node. Fig.2.18 (right y-axis) also explicitly quantifies the percentage latency error between RC and RLC and shows errors up to 18% for MFP of 3.8μm. A higher MFP would yield a bigger discrepancy because a lower resistance renders a higher importance to inductance.

Having detailed latency comparison above, we now compare another important metric-energy per bit. Fig.2.19 shows energy expended to drive a bit over 10mm length as a function of wire width. For all curves, the energy per bit exhibits an optimum width value. This is because the energy per bit depends only on the wire and the repeater capacitance. As the wire width increases, the inter-metal dielectric capacitance drops (due to increased space between wires), but the inter-layer dielectric capacitance increases. If we use just the RC models, both *Cu* and *CNT* bundle yields similar energy per bit. In the RC model, to the first order, the capacitance values of *Cu* and the *CNT* are similar because both the total wire and the repeater capacitances are similar (Even though the optimum repeater size is larger with *CNT*, the spacing between them is also larger, reducing their number).

However, if the inductance in the *CNT* bundle case is considered, the energy per bit (E_{bit}) reduces compared to the RC model (Fig.3). At 200nm wire width, E_{bit} drops more than 10% compared to the RC model with the normally accepted MFP of 1.6μm. For a larger MFP, the drop in E_{bit} is more substantial. Neglecting inductance, where it is important, results in a much larger repeater area than with the RLC model, contributing to the unnecessary power dissipation [31]. The RLC domain analysis results in longer optimum repeat spacing and a smaller optimum repeater size to obtain

impedance matching, and E_{bit} is proportional to the ratio of optimal driver size and the optimal repeat spacing. Thus, a proper consideration of inductance in *CNT* bundles, results in them being more power efficient than previously thought. Another great advantage of repeaters with reduced area for *CNT* bundle relieved routing constraints and less via blockage [31].

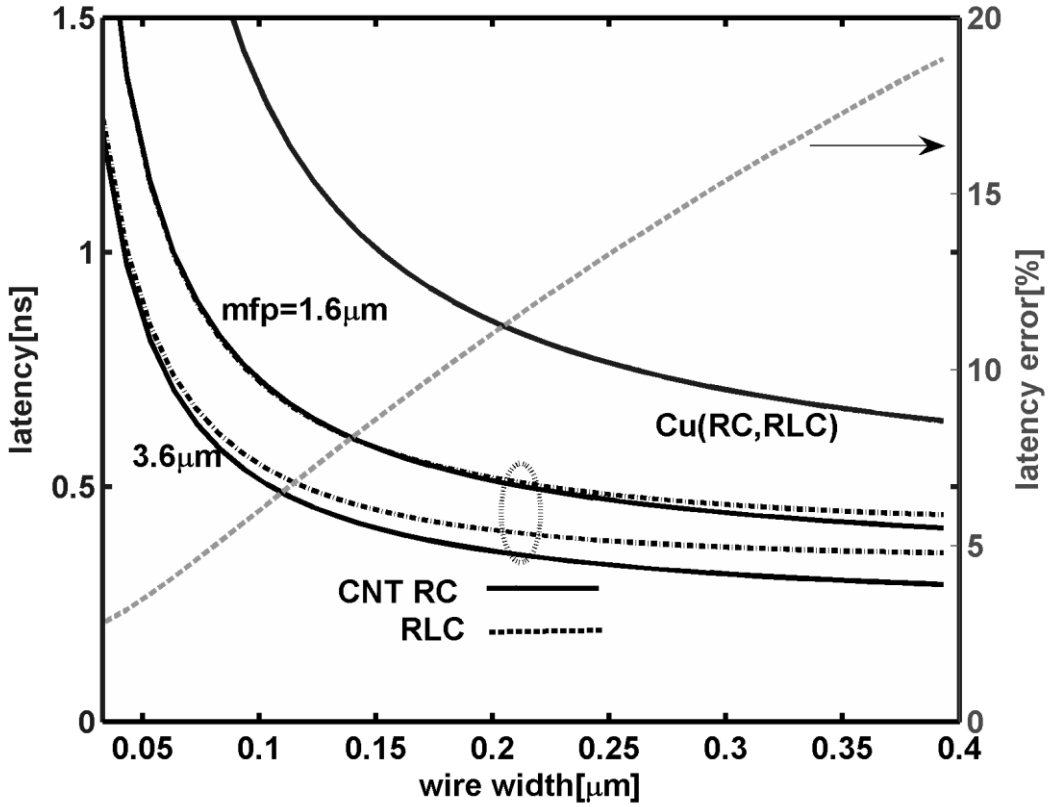

Fig. 2.18 latency for 18mm wire length of *Cu* and *CNT* vs. wire width with various mean free path. Five curves corresponding to the left y-axis – one for *Cu* and four *CNT* bundle. Solid lines are for RC. Dotted lines are for RLC. Both RC and RLC lines are overlapped in *Cu* line. Light dotted line is latency error between RC and RCL models of *CNT* bundles with mfp of 3.6 μm

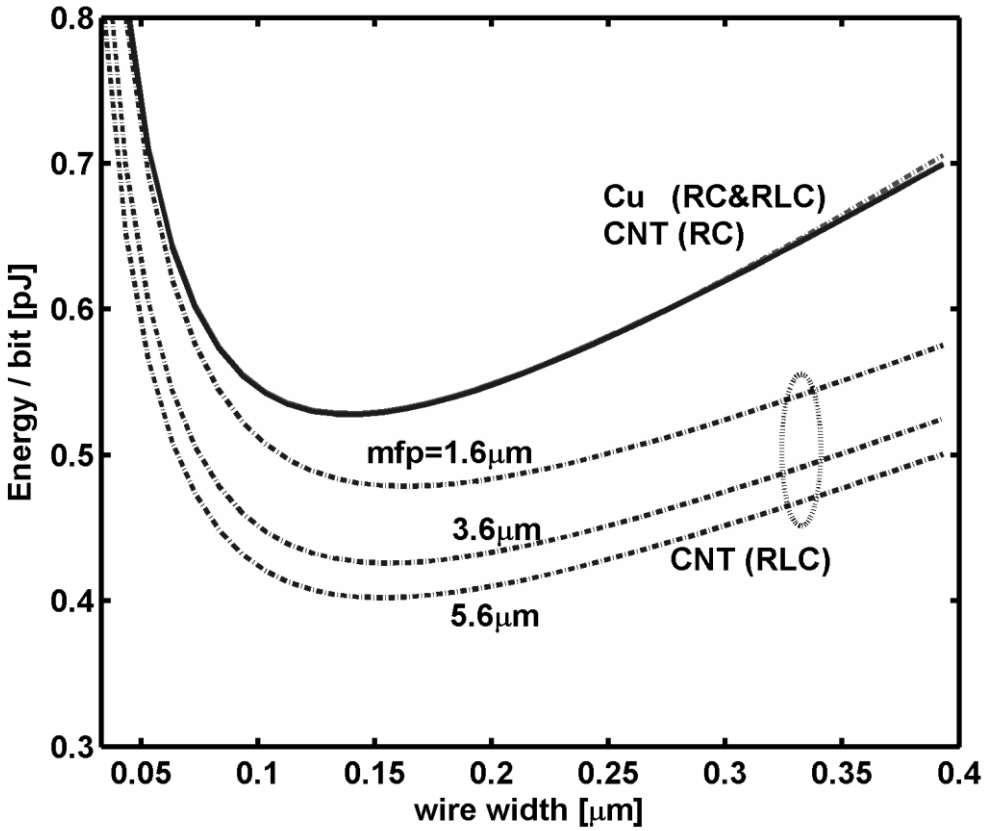

Fig. 2.19 Energy per bit for 10mm wire of *Cu* and *CNT* vs. wire width with various mean free path. Light dotted lines are RLC models. Solid lines are RC models. *Cu* (RC and RLC) line overlaps *CNT* (RC) lines

2.2.2.3 Optical Interconnect Modeling

The basic architecture of optical link consists of an off-chip laser, a quantum-well modulator at the transmitter (converts CMOS gate output to optical signal), a waveguide comprising a silicon core (refractive index~3.5), and SiO_2 cladding as the transmission medium, and a transimpedance amplifier (TIA) followed by gain stages at the receiver (Figure 2.20). The total delay of an optical interconnect is the sum of

the transmitter, waveguide, and the receiver delays. The transmitter delay arises from the CMOS gate driving the capacitive modulator load. It is minimized using a buffer chain and is dependent on the fan-out-four (*FO4*) inverter delay of a particular technology node. The waveguide delay is dictated by the speed of light in a dielectric waveguide (~11.3ps/mm). Finally, the receiver delay is calculated based on circuit considerations. It assumes the input pole (the node at the input of TIA) to be dominant and is optimized to meet the bandwidth and the bit error rate (10^{-15}) criteria. The total power dissipation for the optical interconnect is calculated by optimizing the sum of the receiver and the transmitter power dissipation as outlined in reference [32] in the context of off-chip interconnects.

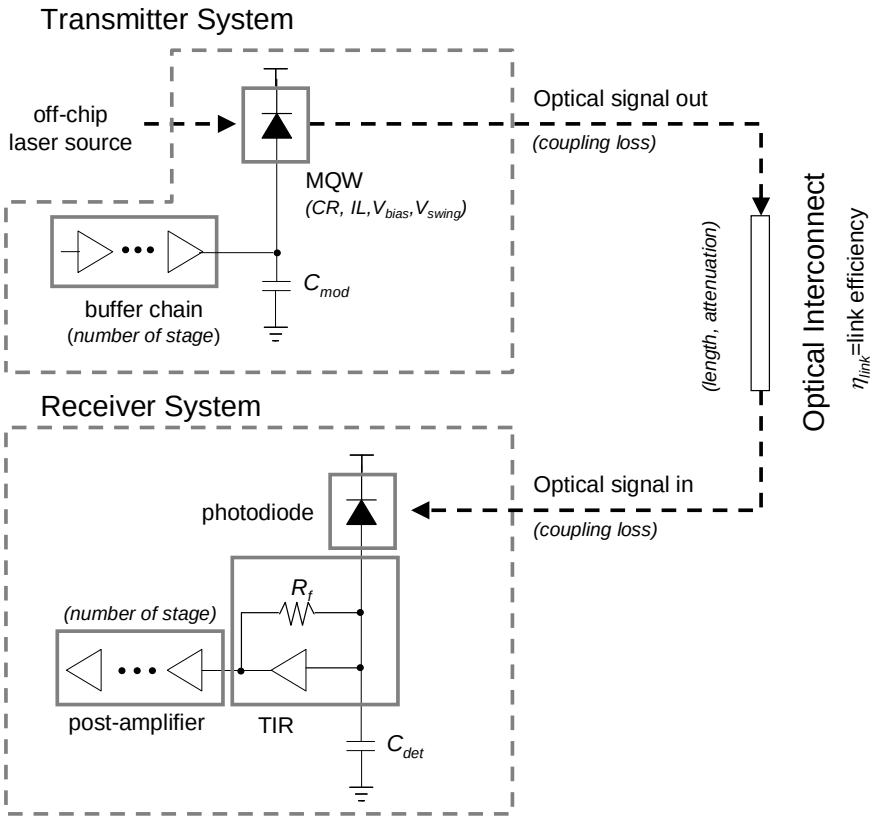

Fig. 2.20 The schematic of quantum-well modulator-based optical interconnect. The modulator parameters assumed for this optical interconnect were taken from ref. [33]

and were assumed to be insertion loss (*IL*)=0.475, contrast ratio (*CR*)=4.6, and bias voltage (V_{bias})=4.7V.

2.2.2.4 Latency and Energy per Bit as a Function of Scaling

Fig. 2.21 compares the interconnect latency of *CNT*, *Cu*, and optics based links as a function of the technology node for semiglobal (~1mm) and global (~10mm) wires. For mean free paths (l_o) of 0.9μm and 2.8μm, CNT wires show 1.6X and 3X latency improvement, respectively, over *Cu* at all technology nodes. Optical wires show an advantage over both *CNT* and *Cu* for longer lengths (~10mm) because a large fraction of the delay occurs in end-devices. For 1mm long wires, optics becomes advantageous over *CNT* only at smaller technology nodes. This is because with scaling, *CNT* and *Cu* latency increases, whereas, optical delay reduces due to an improvement in transistor performance (transmitter and receiver).

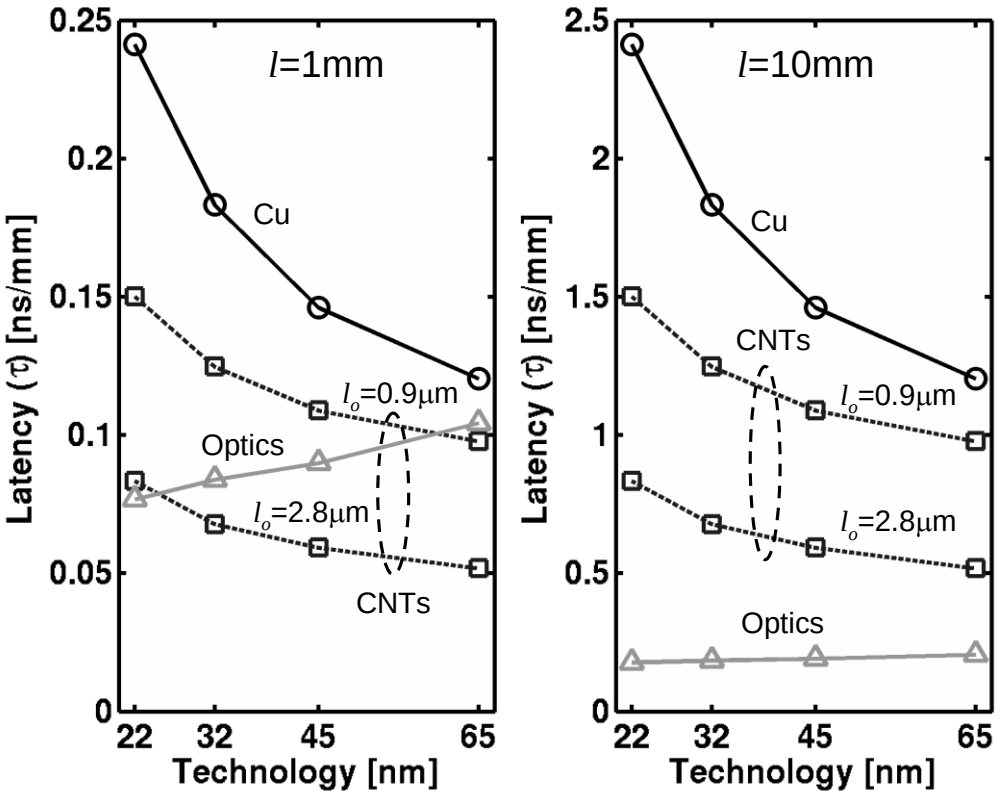

Fig. 2.21 Latency in terms of technology node for two different interconnect lengths. *lo* is the mean free path and *PD* is packing density of metallic *SWCNTs* in a bundle. SWCNT diameter (d_t) is 1nm. For optics, the capacitance of monolithically integrated modulator/detector (C_{det}) is 10fF [34][35].

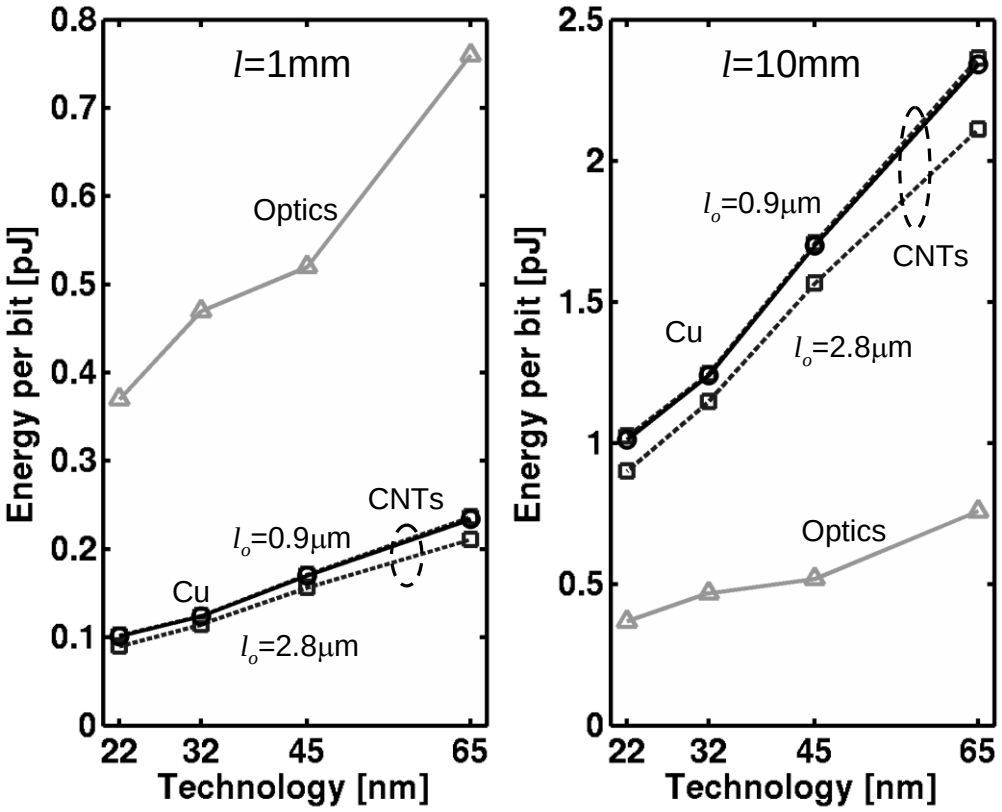

Fig. 2.22 Energy per bit vs. technology node for two different interconnect lengths corresponding to global and semiglobal wire length scales. For *CNTs*, *PD* is 33% and the wire diameter, d_t is 1nm. For optics, the capacitance of monolithically integrated modulator/detector capacitance (C_{det}) is 10fF [34][35].

Fig. 2.22 compares the energy per bit requirements of the three technologies. For *Cu/CNT* wires the dominant energy is the dynamic switching energy: CV^2 (C is total capacitance that includes the wire and the repeater components, V is the voltage), while for optics it arises from the static power dissipation in the end-device amplifiers. For 10mm, global wires, optics is most energy efficient, while for 1mm, semiglobal wires both *Cu* and *CNT* present a better efficiency than optics. While, at all length scales and at 22nm node, *CNTs* with l_o=2.8μm are 20% energy efficiency compared to *Cu*. This is because *CNT* operates in RLC region, where a smaller resistance results in

a smaller optimum repeater size, hence a smaller total repeater capacitance [36] This is in contrast with an RC wire, where the total optimum repeater capacitance is a constant fraction of the wire capacitance, irrespective of the resistance. The 0.9µm l_0 CNT exhibits similar energy per bit as *Cu* because even though l_0 is larger than *Cu*, the sparse (33%) *PD* results in a resistance similar to that of *Cu*.

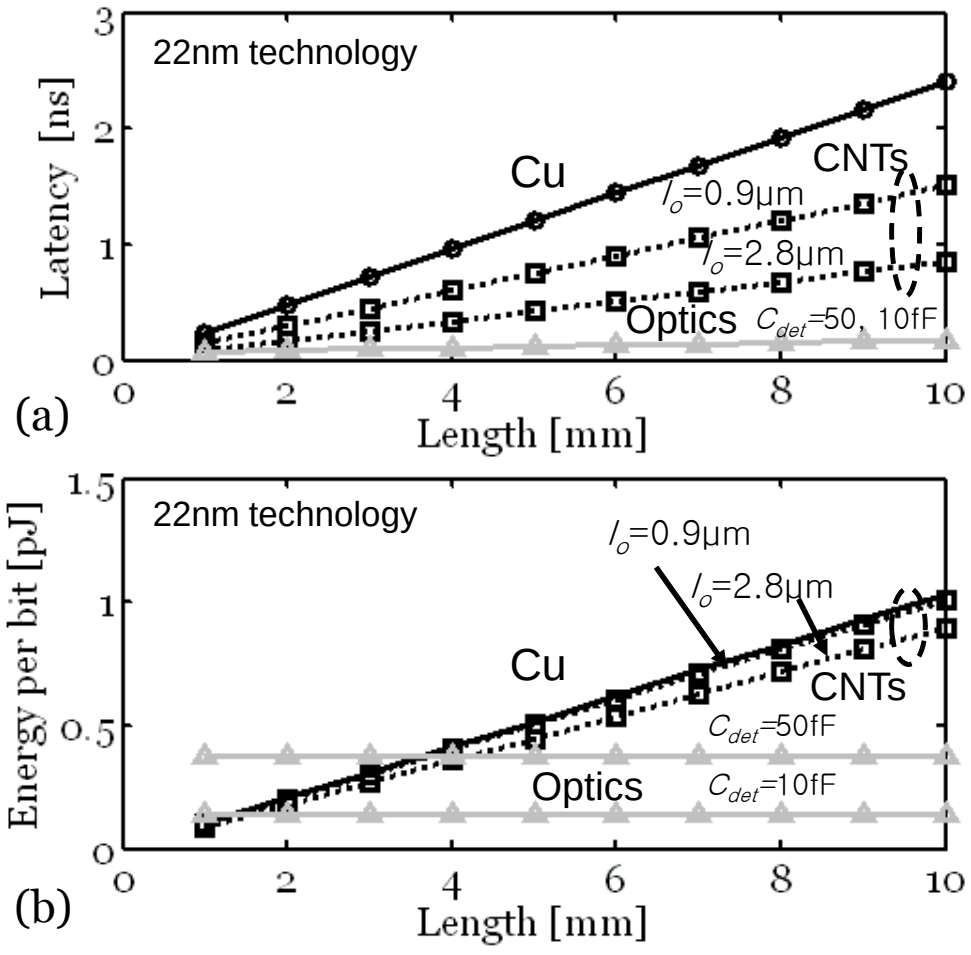

Fig. 2.23 Latency and energy per bit in terms of wire length for the 22nm technology node. l_o is the MFP. For optics, the detector capacitance (C_{det}) is 50fF and 10fF.

Fig. 2.23(a), explores the dependence of latency on the wire length between *Cu/CNTs* and optics.. The delay in all three technologies linearly increases because *Cu/CNTs* are buffered with repeaters in a delay-optimized fashion and for optics, the latency of the medium is linearly dependent on the length. Optics clearly shows the lowest latency for the reasons stated above. *CNTs* with l_o=0.9μm and l_o=2.8μm give 1.6X and 3X latency improvement over *Cu*, respectively, as shown in Fig. 2.23(a).

Fig. 2.23 (b) gives comparisons of energy per bit as a function of the wire length between the three technologies, depicting that, while *Cu/CNTs* gives a linearly increasing energy per bit as length increases, optics shows almost constant values at all length scales. This is because the static power from the end devices in optics does not scale with length, whereas dynamic power in electrical wires do. Here, below 4mm (semi-global interconnect), *CNTs* with l_o=2.8μm can outperform the optics. In addition, below the length of 3.8mm, *Cu* and *CNTs* with l_o=0.9μm can outperform the optics

2.2.2.5 Latency, Power Density and Bandwidth Density

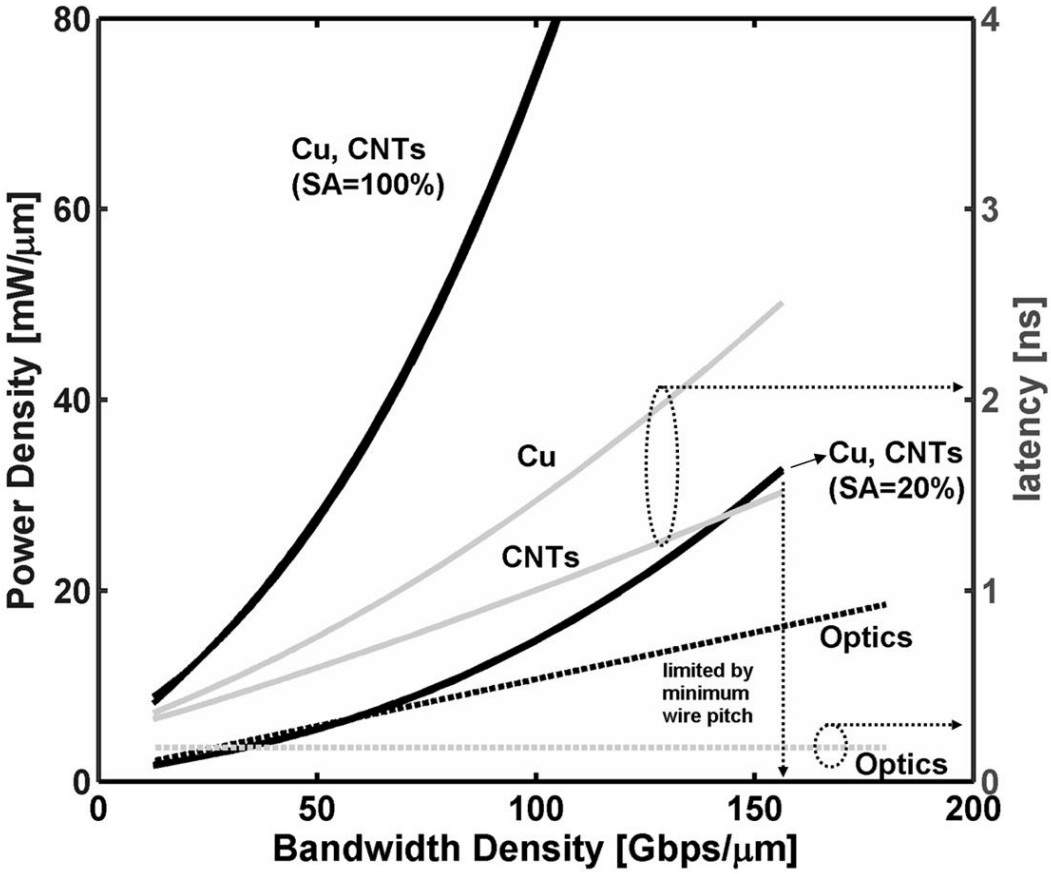

Fig. 2.24 Power density and Latency vs. Bandwidth density (Φ_{BW}) comparison between *Cu*, *CNTs* and Optical interconnect for different switching activities at 10Gb/s global clock frequency (f_{clk}). Bandwidth density is implicitly changed by wire pitch for *Cu* and *CNTs* and number of channels of WDM for optical interconnect. For *CNTs*, l_0 is 0.9um and *PD* is 33%. For optics, detector capacitance (C_{det}) is 10fF [34][35]. (wire length = 10mm, 22nm transistor technology node)

Modern multi-core processors can be limited by the inter-core communication bandwidth. This coupled with a limited processor perimeter, puts a premium on ϕ_{BW}. An ideal wire technology should provide a large ϕ_{BW}, at the lowest possible latency and power density. We now take a system's top-down view, and compare the power density and the latency penalty incurred to achieve an architecture dictated, global

wire, ϕ_{BW} for all three technologies (Fig. 2.24). For *Cu* and *CNT* wires a larger ϕ_{BW} is achieved by implicitly reducing wire pitch, while keeping bandwidth per wire constant (assumed to be dictated by the global clock: 10Gbps) using wave pipelining between repeater stages. Thus, the maximum achievable ϕ_{BW} is limited by the minimum wire width (ITRS [4]) and is also shown in Fig.2.24. The minimum wire pitch (maximum ϕ_{BW}) can be limited to a larger (smaller) value by practical considerations such as excessive area of the repeaters. For optical interconnects a larger ϕ_{BW} is achieved using a different mechanism because the waveguide size cannot be reduced beyond a certain point. Instead, optical wires deploy denser wavelength division multiplexing (WDM) and the maximum ϕ_{BW} is limited by the degree (no. of wavelengths) of WDM.

In Fig. 2.24, optical links exhibit the lowest latency, followed by *CNT* bundle and *Cu* at all ϕ_{BW}. Further, the optical latency is independent of ϕ_{BW}, since a higher ϕ_{BW} is achieved just by adding more wavelengths, which does not impact the latency of each wavelength. In contrast, both *CNT* and *Cu* latency increases with ϕ_{BW}, as a smaller wire pitch increases resistance and the resulting number of repeater stages. Thus, optical link latency advantage becomes more pronounced to about 13X and 8X better than *Cu* and *CNT*, respectively, at very high inter-core ϕ_{BW} requirement.

In contrast to latency, power density rises linearly with ϕ_{BW} for an optical wire (Fig. 2.24) as the additional bandwidth arises from equally power consuming wavelength channels. Whereas, the *CNT* and *Cu* power density increases non-linearly with ϕ_{BW},

following the dependence of total wire capacitance (repeater and wire capacitance) on wire pitch: At high ϕ_{BW} (small wire pitch), IMD capacitance dominates, whereas at smaller ϕ_{BW} (large pitch), ILD capacitance is larger. Fig. 2.24 also shows that *CNT* (l_o=0.9um) and *Cu* have equal power density as a result of a similar wire and repeater capacitance. However, a higher *CNT* l_o renders a lower power density than *Cu* (not shown) as a smaller resistance in RLC regime results in a smaller optimum repeater size. The coupling capacitance plays an important role in determining both power and delay. However, for power calculation, on average, the nominal capacitance is a good approximation.

Finally, the power dissipation in the electrical schemes has a strong switching activity (*SA*) dependence. The optical link, to first order, is independent of *SA* as most of its power is static in nature. The optical links exhibit a lower power density than both *Cu* and *CNT*s for *SA*=100%. Whereas, at *SA*=20%, there is a critical ϕ_{BW} beyond which the optical links are more power efficient. Thus, optical interconnects are more power efficient at larger required ϕ_{BW}, larger *SA*, and larger lengths. Fig. 2.25 explicitly plots the power density as a function of *SA* for each of the three technologies, showing the critical SA beyond which optical wire power density is superior. These trends in *SA* insinuate the importance of architectural differences to maximally exploit the potential of optical wires.

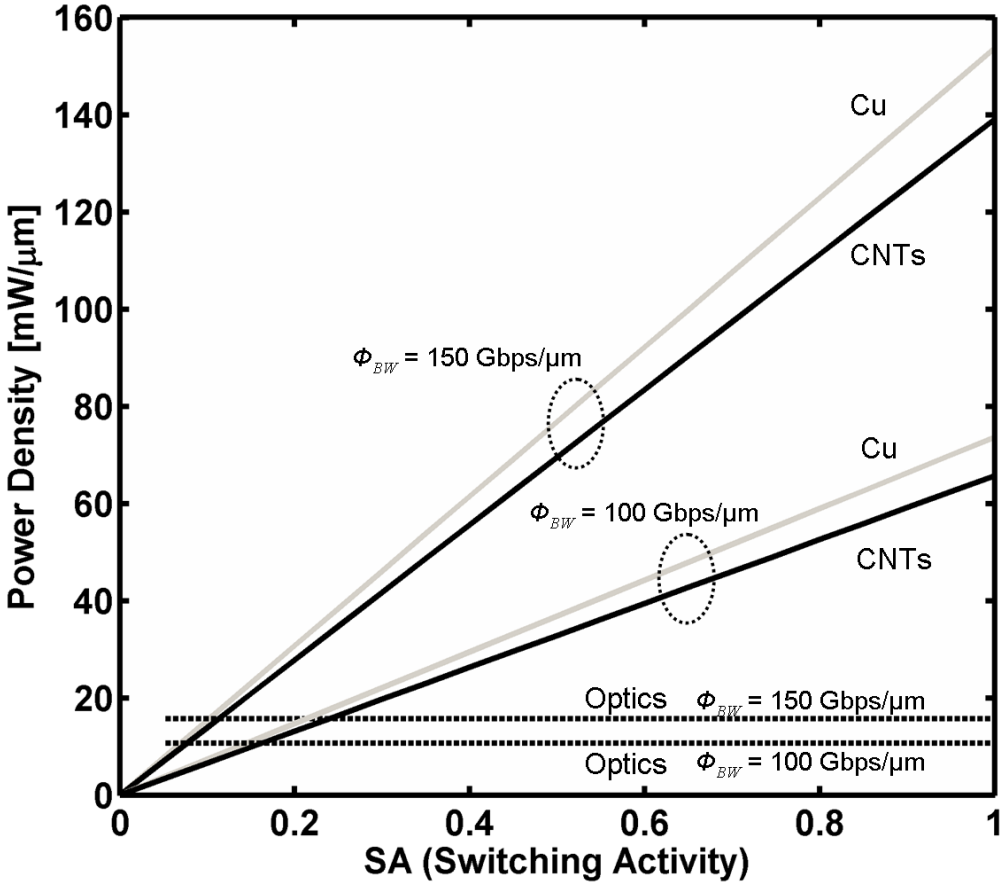

Fig. 2.25: Power density vs. SA (switching activity) with different bandwidth density (Φ_{BW}) at 10 Gb/s global clock frequency. For CNT, PD is 33% and l_0 is 2.8 m. For optics, detector capacitance (C_{det}) is 10fF [34][35]. (wire length = 10mm, 22nm transistor technology node)

2.2.2.6 Impact of CNT and Optics Technology Improvement

The performance of *CNT* is a strong function of l_0 and *PD*, whereas, optical wire performance critically depends on capacitances associated with the detector and the modulator (C_{det}, C_{mod}). In this section, we quantify the impact of these device and materials parameters on the comparisons. Such an analysis is useful for technologists to get an idea of the required parameters from a system's standpoint, to make their technology competitive. Fig. 2.26 (*SA*=20%) illustrates that an improvement in CNT l_0 from 0.9μm (practical) to 2.8μm (ideal) and in *PD* from 0.33 to 1 results in some reduction in power density for all ϕ_{BW}. Moreover, the improvement in l_0 shows a stronger impact than an improvement in *PD*. This is because *PD* increase results in a smaller increase in inductance to resistance ratio (L/R) as L_{kin} also decreases along with *R*. For this *SA* and for C_{det}=25fF, *CNT* and *Cu* outperform optics. However, if the C_{det} and C_{mod} can be brought down to 10fF using a monolithic detector (as opposed to hybrid bonded III-V detector), optical wires outperform other technologies even at smaller *SA*.

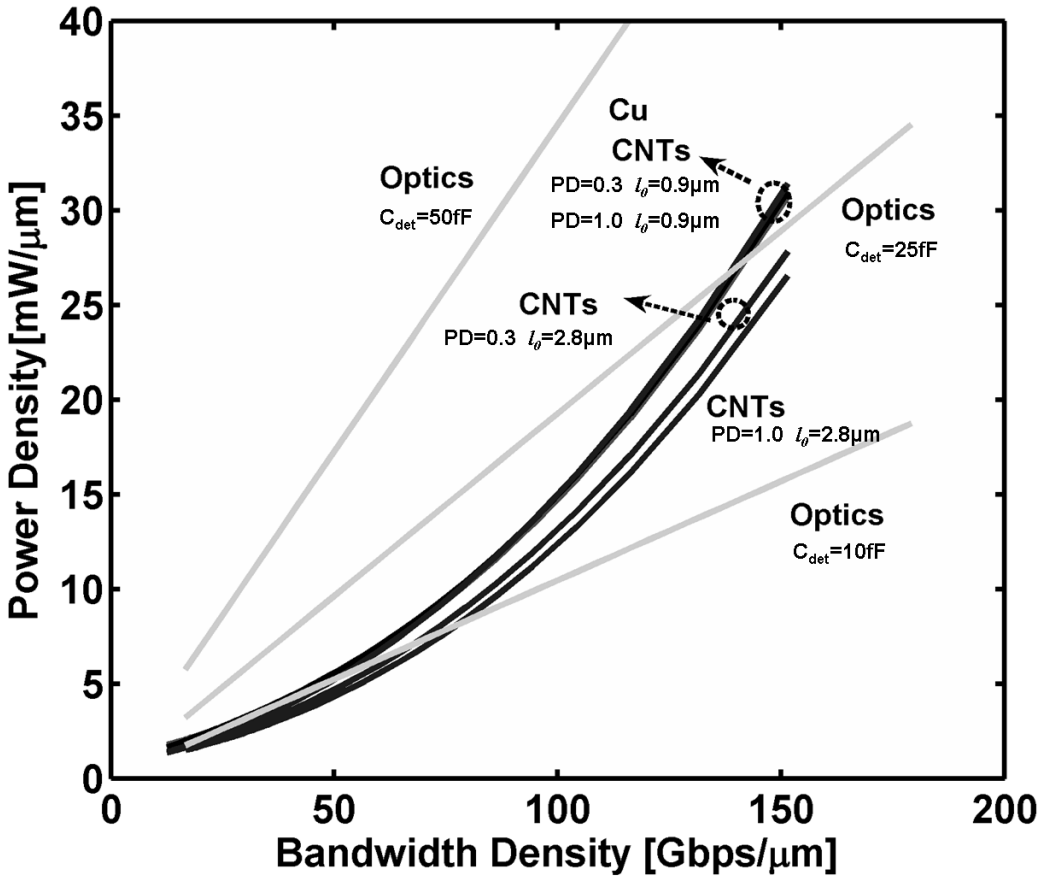

Fig. 2.26 The impact of *CNT* and optics technology improvements on Power density vs. bandwidth density curves (SA = 20%). C_{det} reduction for optics results in a large improvement in power density. (wire length = 10mm, 22nm transistor technology node, f_{clk}=10Gb/s)

Similarly, Fig. 2.27 captures the impact of technology parameters on the latency. For *CNTs*, improvement in both l_0 and *PD* results in a substantial decrease in latency compared to *Cu*. With ideal l_0 and *PD*, and at lower ϕ_{BW}, latency performance is comparable to optical wires.

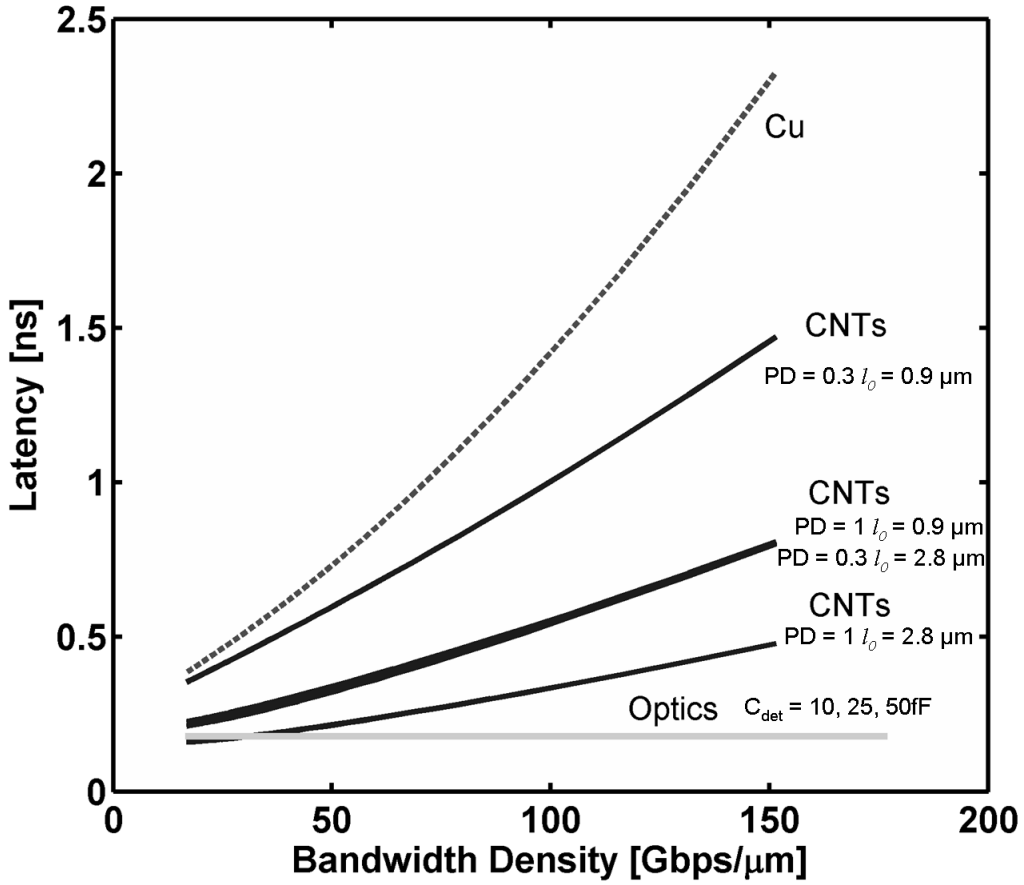

Fig. 2.27: The impact of *CNT* and optics technology improvement on latency vs. bandwidth density. *CNT* parameter improvement results in a very large improvement in latency over *Cu*. *CNT* with ideal parameters has comparable latency to optical wires at very low Φ_{BW}. (wire length = 10mm, 22nm transistor technology node, f_{clk}=10Gb/s)

Bibliography

[1] E. H. Sondheimer, "The mean free path of electrons in metals," *Adv. Physics*, vol. 1, no. 1, pp.1-2, 1952.

[2] A. F. Mayadas and M. Shatzkes, "Electrical-resistivity model for polycrystalline films: the case of arbitrary reflection at external surfaces," *Phys. Review B*, vol. 1, pp. 1382-1389, 1970

[3] P. Kapur, J. P. McVittie, and K. C. Saraswat, "Technology and reliability constrained future copper interconnects-Part I: Resistance modeling," *IEEE Trans. on Electron Devices*, Vol. 49, No. 4, pp. 590-597, Apr. 2002.

[4] International technology roadmap of semiconductor, 2007 version

[5] D. A. B. Miller, "Physical reasons for optical interconnection," *Intl. Journal of Optoelectronics*, vol. 11, pp. 155-168, 1997.

[6] H. B. Bakoglu, "Circuits, interconnections, and packaging for VLSI" (Addison-Wesley, New York, 1990).

[7] P. Kapur, J. P. McVittie, and K. C. Saraswat, "Technology and reliability constrained future copper interconnects-part II: Performance Implications," *IEEE Trans. on Electron Devices*, vol. 49, no.4, pp. 598-604, Apr. 2002

[8] P. Kapur, "Scaling induced performance challenges/limitations of on-chip metal interconnects and comparisons with optical interconnects", *Stanford University*, PhD. thesis, 2002.

[9] Q. Chen, J. A. Davis, P. Zarkesh-Ha, and J. D. Meindl, "A novel via blockage model and its implications", *Intl. Interconnect Technology Conference*, pp.15-18, Jun., 2000.

[10] J. A. Davis, R. Venkatesa, A. Kaloyeros, M. Beylansky, S. J. Souri, K. Banerjee, K. C. Saraswat, A. Rahman, R. Reif, and J. D. Meindl, "Interconnect limits on gigascale integration (GSI) in the 21st century," *Proceedings of the IEEE,* vol. 89, no. 3, pp. 305-324. Mar. 2001.

[11] C. T. White, and T. N. Todorov, "Carbon nanotube as long ballistic conductors," *Nature*, vol. 393, pp. 240-242, 1998.

[12] A. Naeemi, and J. D. Meindl, "Design and Performance Modeling for Single-Wall Carbon Nanotubes as Local, Semi-global and Global Interconnects in Gigascale Integrated Systems," Trans. on Electron Device, vol. 54, pp. 26-37, 2007.

[13] M. S. Dresselhaus, G. Dresselhaus, and P. Avouris, Eds., Topics in Applied Physics, Carbon Nanotubes: Synthesis, Structure, Properties and Applications, New York: Springer-Verlag, 2000.

[14] J. Y. Park, S. Rosenbelt, Y. Yaish, V. Sazonova, H. Ustunel, S. Braig, T. A. Arias, P. L. McEuen, "Electron-phonon Scattering in Metallic Single-wall Carbon Nanotubes," Nano Letters, vol.4, pp. 517-520, 2004.

[15] A. Nieuwoudt, Y. Massoud, "Evaluating the Impact of Resistance in Carbon Nanotube Bundles for VLSI Interconnect Using Diameter-Dependent Modeling Techniques," Trans. On Electron Devices, vol. 53, no. 10, pp.2460-2466, 2006.

[16] Z. Yao, C. L. Kane, and C. Dekker, "High-field Electrical Transport in Single-wall Carbon Nanotubes," Phys. Review Letters, vol. 84, pp. 2941-2944, 2000.

[17] O. Hjortstam, P. Isberg, S. Soderholm, and H. Dai, "Can We Achieve Ultra Low Resistivity in Carbon Nanotube-based metal Composites?" Applied Phys. A, vol. 78, pp. 1175-1179, 2004.

[18] P. J. Burke, "Luttinger Liquid Theory as a Model of the Gigahertz Electrical Properties of Carbon Nanotubes," Trans. on Nanotechnology, vol. 1, no. 3, pp. 129-144, 2002.

[19] Naeemi, R Sarvari, and J. D. Meindl, "Performance Comparison Between Carbon Nanotube and Copper Interconnect for Gigascale Integration(GSI)," Elec. Device Letters, vol. 26, no. 2, pp. 84-86 2005.

[20] FastCap: available on-line at www.rle.mit.edu/cpg/research_codes.htm

[21] A. Nieuwoudt, and Y. Massoud, "Understanding the Impact of Inductance of Carbon Nabotube Bundles for VLSI interconnect Using Scalable Modeling Techniques," Trans. on Electron Device, vol. 5, no. 6, 2006.

[22] Th. Hunger, et al., "Transport in Ropes of Carbon Nanotubes: Contact Barriers and Luttinger Liquid Theory", Physical Review B, Vol. 69, 195406, 2004.

[23] W. Liang, et al., "Fabry-Perot interference in a NAnotube Electron Waveguide", Nature, Bol. 411, pp. 665-660, 2001.

[24] A. Naeemi, and J. D. Meindl, "Design and performance modeling for single-wall carbon nanotubes as local, semi-global and global interconnects in gigascale integrated systems," *Trans. on Electron Device,* vol. 54, pp. 26-37, 2007.

[25] A. Naeemi and J. D. Meindl, "Conductance modeling of graphene nanoribbons (GNR) interconnects." IEEE Electron Device Lett., vol. 28, no. 5, pp. 428-431, May 2007.

[26] N. Srivastava, and K. Banerjee, "Performance Analysis of Carbon Nanotube Interconnects for VLSI Application," IEEE/ACM Inter. Conf. on Computer-Aided Design, pp. 383-390, 2005.

[27] H. B. Bakoglu, "Circuits, interconnections, and packaging for VLSI" (Addison-Wesley, New York, 1990).

[28] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat, "3-D ICs: A novel chip design for improving deep submicron interconnect performance and systems-on-chip integration," Proceedings of the IEEE, vol. 89, pp. 602-633, May 2001.

[29] C. P. Yue, and S. S. Wong, "Physical Modeling of Spiral Inductors on Silicon," Trans. on Electron Device, vol. 47, no. 3, pp. 560-568, 2000

[30] K. Banerjee, and A. Mehrotra, "Analysis of On-chip Inductance Effects for Distributed RLC Interconnects," Trans. on Computer-Aided Design of Int. Circuits and Systems, vol. 21, no. 8, pp. 904-914, 2002.

[31] Y. I. Ismail, E. G. Friedmadn, "Effects of inductance on the propagation delay and repeater insertion in VLSI circuits: A summary," IEEE, Circuiits and Systmes Magazine, vol 3, no1, 2003, p 24-28

[32] P. Kapur, and K. C. Saraswat, "Comparisons between electrical and optical interconnects for on-chip signaling," International Interconnect Technology Conference, pp. 89-91, 2002.

[33] O. Kibar, D. A. V. Blerkom, C. Fan, and S. C. Esener, "Power Minimization and Technology Comparisons for Digital Free-Space Optoelectronic Interconnects," IEEE J. of Lightwave Technol., vol. 17, no.4, pp. 546-554, Apr. 1999.

[34] A. K. Okyay, A. M. Nayfeh, A. Marshall, T. Yonehara, P. C. McIntyre, and K.C. Saraswat, "Ge on Si by Novel Heteroepitaxy for High Efficiency Near Infra-red Photodetection," CLEO, 2006.

[35] O. I. Dosunmu et. al. "High-Speed Resonant Cavity Enhanced Ge Photodetectors on Reflecting Si Substrates for 1550nm Operation," Photonics Technology Letters, Vol. 17, No. 1, January 2005, pp. 175-177.

[36] Y. I. Ismail, and E. G. Friedman, "Effects of Inductance on the Propagation delay and Repeater Insertion in VLSI Circuits: A Summary", IEEE Circuits and Systems Magazine, pp. 24-28, 1st quarter, 2003.

Chapter 3

Capacitvely Driven Low-Swing Interconnect (*CDLSI*) and Its Analytical Modeling

3.1 Introduction

On-chip power consumption has become a serious problem in deep sub-micron technology nodes. Not surprisingly, it was found that interconnects contribute to a significant fraction of the energy consumption. This is even more serious for FPGA of which interconnect fabrics constitute most of the chip area. [1]

Especially global interconnects are larely responsible for this problem because clock distribution and global signaling wires dissipate a large portion of dynamic power. To alleviate this problem, many circuit techniques have been suggested [2][3]. The most common power reduction technique is to reduce voltage swing and decrease dynamic power dissipation from capacitive wires.

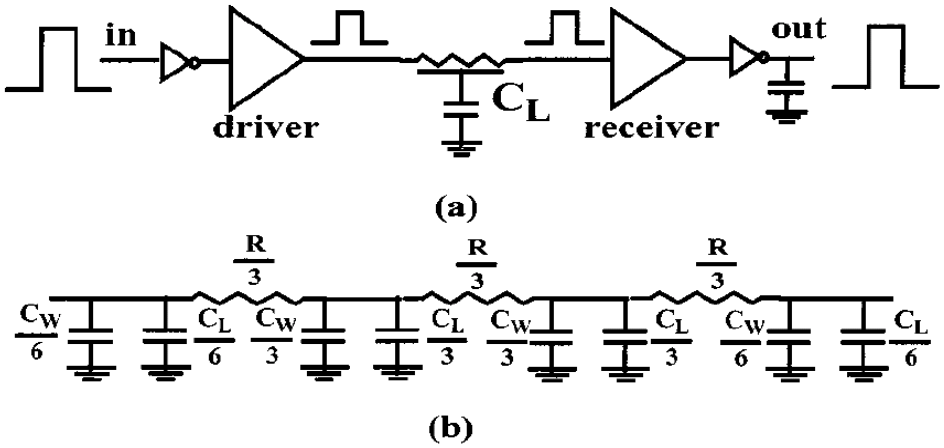

Fig. 3.1 (a) Schematics of conventional low swing interconnect scheme. (b) Equivalent circuit of (a)

Fig 3.1 shows the basic concept of low swing interconnect with driver and receiver. An incoming signal with full rail to rail swing is converted to reduced swing through a driver (usually level shifter with different supply voltage). Then the reduced swing propagates through diffusive RC wire. Finally, a receiver regenerates low the swing signal to full swing signal via another level shifter. Conventionally, driver and receiver in this system employ level shifters with additional voltage supplies. Fig 3.2 illustrates more detailed view of such system. The driver uses low level supply voltage to drive a full swing to a low swing and the receiver uses a level converter [4] with full supply voltage.

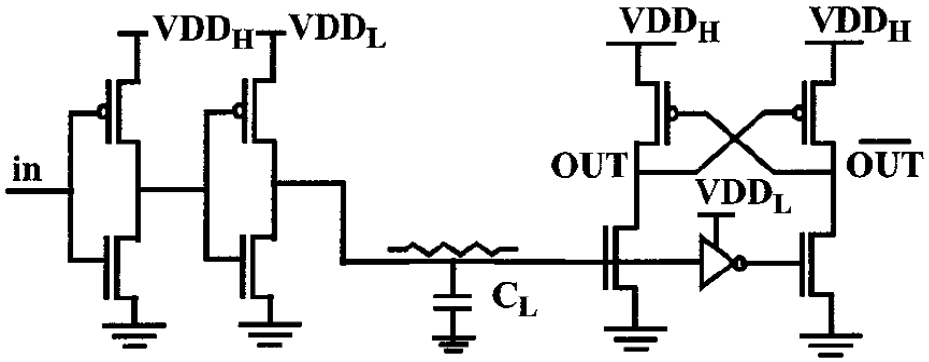

Fig. 3.2 Conventional low swing scheme with additional power supply

3.2 Repeated CDLSI

Conventional low swing interconnects (where logic operations are done at normal voltages, however, signal transmission is done at a reduced voltage) are attractive as they tackle the wire energy problem head-on by reducing dynamic power. However, this advantage is typically accompanied by a latency penalty and a reduction in the noise margin. Moreover, they usually require a secondary low voltage power supply, which makes the system more expensive and complex [3]. Recently, *CDLSI* was shown to exhibit excellent energy saving without seriously impacting latency (Fig.3.3) [5][6]. The key element in this system is the coupling capacitance, which not only eliminates the necessity of a secondary power supply, but also introduces pre-emphasis (the high frequency emphasis arising from the pole-zero pair of the high pass filter network), resulting in bandwidth improvement. Wires are differential and twisted in order to cancel the coupling noise. The coupling capacitor (C_c) is inserted between the transmitter and the wire. The receiver comprises of strong arm latch sense amplifier

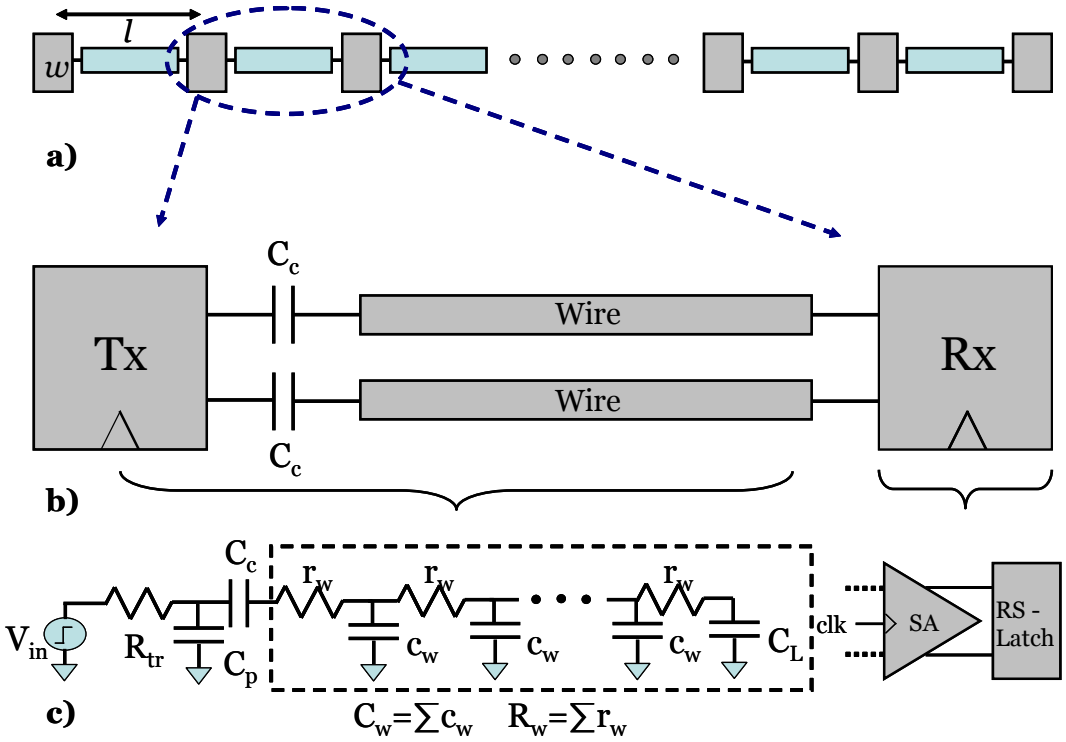

Fig. 3.3 (a) Simple illustration of repeated capacitively driven low swing interconnect (*CDLSI*). *l* is a segment length. *w* is the size of Tx/Rx. (b) Zoomed schematic of one segment of *CDLSI*. (c) Equivalent circuit model of one segment. Dashed box indicates a distribution model ofz the wire. Fanout of between the sense amplifier and the RS latch is assumed to be 1 (w_{SA}=w_{RS}=w).

followed by a RS latch (Fig. 3.3). The sense amplifier boosts a reduced swing to full rail to rail swing. The common mode of this amplifier is set at supply voltage through PMOS transistor at the receiver side. The RS latch is inserted in order to register the output value of the sense amplifier.

3.2.1 Analytical Model

The delay of this system based on a modified Elmore model is given by

$$\tau = aR_{tr}(C_p + C_c) + \gamma b(R_w C_w + R_w C_L) + t_{SA} \qquad (3.1)$$

Switching model dependent parameters *a* and *b*, in 10-90% rise/fall time case, are 2.2 and 1.1 respectively. Additional coefficient γ is introduced to describe the pre-emphasis effect with C_c. γ reduces with C_c. since C_c and C_w constitutes the capacitive voltage divider. Thus level of the swing can be determined by the ratio between two values. Here γ is assumed to be approximately 0.3, corresponding to the point where C_c is nearly 1/15 of C_w. R_{tr} is an effective resistance of a NAND gate when pull up and pull down resistances are equalized. C_p is output capacitance of transmitter. R_w and C_w are wire resistance and capacitance respectively. C_L is load capacitance corresponding to the input of the sense amplifier. t_{SA} is the delay of the sense amplifier given as [7]

$$t_{SA} = \frac{2C_L V_{thp}}{I_0} + \frac{C_L}{g_{meff}} \ln \frac{\Delta V_{out}\sqrt{I_0}}{\sqrt{2\beta} V_{thp} \Delta V_{IN}} \qquad (3.2)$$

I_o is the current flowing through the enabling transistor (M_{EN}) of the sense amplifier in Fig. 3.4 when it is enabled. V_{thp} is the threshold voltage of PMOS transistor (Mn1 and Mn2). β is a constant determined by the technology. g_{meff} is

effective transconductance considering all effects from Mn1,Mn2,Mn3 and Mn4 in Fig. 3.4. More detailed expressions regarding I_o and $g_{m,eff}$ are given as [7]

$$I_0 = 2\beta(V_{INDD} - V_{th})^2\left(1 - \frac{0.75}{1 + \frac{V_{DD} - V_{th}}{V_{INDC} - V_{th}}}\right)^2 \quad (3.3)$$

$$g_{m,eff} = g_{m1,3} + g_{m2,4} - \frac{1}{r_{DS1,3} \parallel r_{DS2,4}} \quad (3.4)$$

V_{INDC} is an input common mode voltage applied to IN1 and IN2. In our case, V_{INDC} is same as V_{DD} since the common mode is tied to supply voltage level. t_{SA} in eq. (3,1) is the summation of two delay components. First term represents the delay to discharge the capacitive output load until PMOS in cross coupled inverters turns on. Second term is responsible for the time delay for the non-regenerative output voltage to drop to half V_{DD} level.

Analytical expression of energy per bit for *CDLSI* is given by

$$E_{bit} = \{C_p V_{dd} + (C_w + C_L)\Delta V + C_c(V_{dd} - \Delta V) + I_o t_{SA}\}V_{dd} \quad (3.5)$$

V_{dd} and ΔV are supply voltage and reduced swing respectively. Eq (3.5) represents two components: 1. purely dynamic energy dissipation component from the wire and input load 2. static energy dissipation from the sense amplifier. The value of

reduced swing is assumed to be 50mV, lowest reported value which can still be noise tolerant [5].

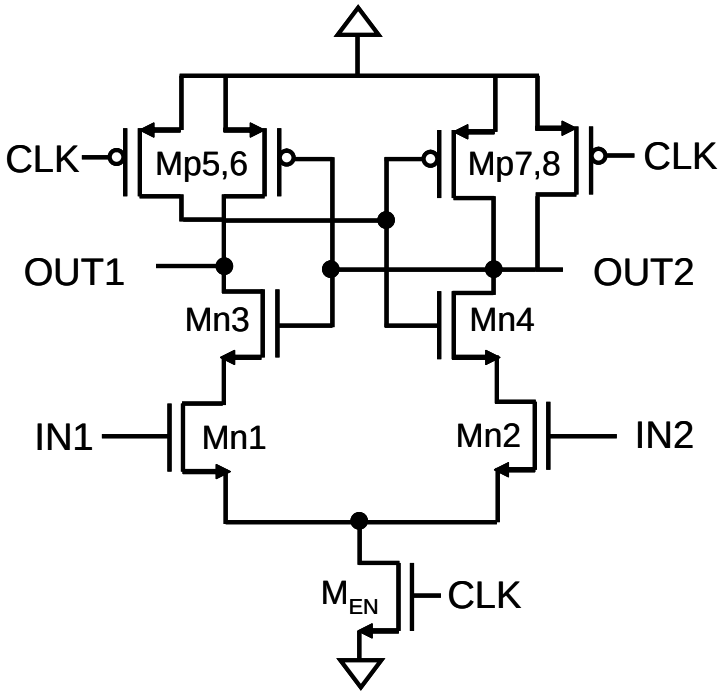

Fig. 3.4 Schematic view of strong arm latch sense amplifier in the receiver side of *CDLSI*.

3.2.2 Delay Optimization

The minimum delay of *CDLSI* with multiple repeater insertion can be calculated in a similar fashion to the conventional delay optimization method. The total delay of the repeated *CDLSI* can be expressed as

$$t_{total} = \frac{L}{l}(t_{wire} + t_{SA})$$

$$= \frac{L}{l}\left\{ b\frac{r_o}{w_{RS}}(w_{RS}c_o + c_c) + \gamma a r_w c_w l^2 + b r_w l c_o w_{SA} + t_{SA} \right\} \quad (3.6)$$

L and l are the total length and the segment length of *CDLSI*, respectively. w_{RS} and w_{SA} are the width of input in RS latch and sense amplifier, respectively. We assume that the fanout of the sense amplifier is one ($w_{SA}=w_{RS}=w$) for the convenience of calculations. With this assumption, we find that t_{SA} can be simplified as a constant ($t_{SA}=A$) owing to the linear dependence of I_o, $C_{L,}$ g_{meff} and β on w. Then, we take a partial derivative on Eq. (3.6) in terms of l and w, equalize them to zero, and obtain the delay optimal segment distance ($l_{t.opt}$) and the repeater width ($w_{t.opt}$).

$$\frac{t_{total}}{l} = 0 \qquad \frac{t_{total}}{w} = 0 \quad (3.7)$$

$$l_{\tau.opt} = \sqrt{\frac{b r_o c_o + A}{\gamma a r_w c_w}} \quad w_{\tau.opt} = \sqrt{\frac{r_o c_w \kappa}{r_w c_o}} \quad (3.8)$$

Minimum delay can be achieved by plugging Eq. (3.8) back into Eq. (3.6).

3.2.3 Delay-Power Optimization

Blindly sizing repeaters with just delay-optimization can result in excessive power consumption. This becomes more serious in future nodes. As fig 2.5 (b) in section 2.1.2 shows, it can easily surmount the power budget below 32nm technology node. This exorbitant power consumption can be reduced at the cost of a small delay penalty

in an optimized fashion [8]. This methodology is acceptable as long as delay doesn't violate system required timing constraints in clock synchronized system. We applied a similar approach on delay-power optimization, however this time for the *CDLSI* interconnects. We typically decrease *w* and increase *l* for power reduction such that $w = w_{\tau.opt} \cdot x_w$ and $l = l_{\tau.opt} \cdot x_l$. Here, x_w and x_l indicate the fractional reduction in sizing and expansion in spacing from the delay optimal point. The fractional change in energy per bit of *CDLSI* is expressed as

$$\eta = \frac{E_{sav}}{E_{\tau.opt}} = \frac{f(x_l, x_w)}{E_{\tau.opt}} \quad (3.9)$$

where

$$E_{sav} = f(x_l, x_w) = \frac{L}{l_{seg}}\left[C_{RSP}V_{dd} + (C_{wire} + C_{SA})V_{low} + C_c(V_{dd} - V_{low}) + (I_o \times t_{VSA})\right]V_{dd}$$

$$= \frac{L}{l_{\tau.opt}x_l}\begin{bmatrix} c_{SA}w_{\tau.opt}x_w V_{dd} + (c_w l_{\tau.opt}x_l + c_{SA}w_{\tau.opt}x_w)V_{low} \\ + c_w l_{\tau.opt}\kappa(V_{dd} - V_{low}) + (i_o w_{\tau.opt}x_w \times t_{VSA}) \end{bmatrix}V_{dd} \quad (3.10)$$

$E_{\tau.opt}$ denotes energy per bit when delay is optimized. The delay penalty, μ, is given as a ratio of sub-optimal delay, where x_l and x_w are not equal to one, to the optimal delay. This is expressed as

$$\mu = \frac{t_{\text{var}}}{t_{opt}} = \frac{L}{l_{opt}x_l}\begin{Bmatrix} b\frac{r_o}{w_{opt}}(w_{opt}c_{SA} + c_w l_{opt}x_l\kappa) + \gamma a r_w c_w (l_{opt}x_l)^2 \\ + b r_w l_{opt}x_l c_{SA}w_{opt}x_w + \text{A} \end{Bmatrix} / t_{opt} \quad (3.11)$$

Now, power can be saved with setting a target η (<1) by trading the delay. However, the associated delay penalty can be minimized by searching for the optimal point within the given η value. This can be done with following steps.

1. Set a target fractional energy saving by setting η (<1).
2. Rearrange Eq. (3.9) in terms of x_l as a function of η and x_w such as $x_l=g(\eta,x_w)$
3. Plug x_l achieved in step 2. into Eq. (3.11)
4. Take a partial derivative on μ in terms of x_w and equalize it to zero.

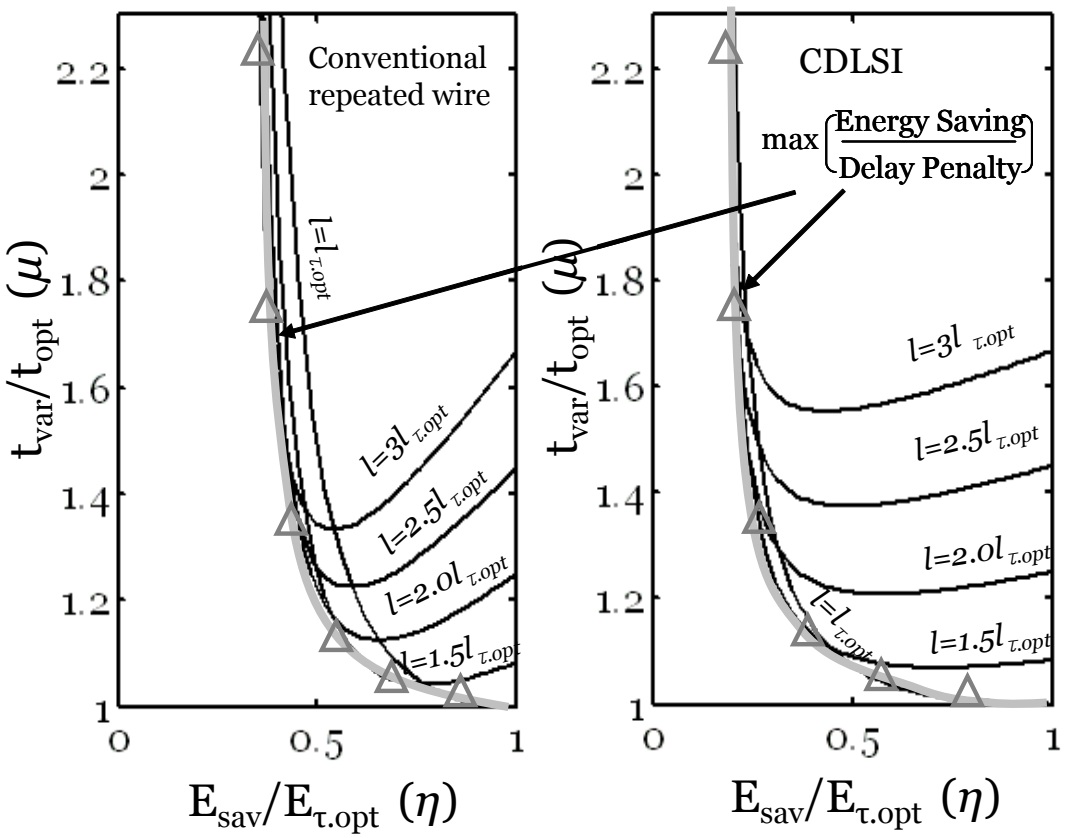

Fig. 3.5 Factional delay penalty (μ) vs. fractional energy saving (η) for a 10mm global wire. Left figure is for the conventionally repeated wire and right figure is for the repeated *CDLSI*. t_{opt} and $E_{\tau.opt}$ are both delay optimized values.

Outputs of this optimization methodology for both conventional wire and *CDLSI* are shown in Fig. 3.5. The gray lines indicate various local minima in delay penalty for a given energy saving. Various curves are for different sub-optimal repeater spacings. For conventional wires, the graph clearly shows that almost 40 % ($\eta \sim 0.4$) of the total power is consumed by the wire (excluding the transceivers) when the interconnect is delay optimized. For *CDLSI*, on the other hand, transceivers consume the bulk of the power for the same condition. As a result, for a given delay penalty, the impact of power saving ($\Delta E_{sav}/E_{\tau.opt}$) in *CDLSI* is greater than that in the conventional wires.

3.3 Performance Comparison between CDLSI and Conventional Signaling Scheme

3.3.1 General Metrics: Latency and Energy per Bit

Wire bandwidth of the conventional repeated global interconnect is dominated by the delay of a segment as several bits can travel at the same time in the pipelined interconnect system. Thus, inverse of a segment delay is a measure of the highest achievable wire bandwidth. In practice, bandwidth is a fraction of this because subsequent bits have to wait several segment delays (not just one) to ensure reflections have subsided from the previous bit. Each point of the delay-energy optimization

curves in Fig. 3.5 implicitly also bears information on its bandwidth since its segment delay is determined by segment length (*l* in Fig. 3.3) and transceiver size (*w* in Fig. 3.3). Fig 3.6 compares the delay of conventional wire and *CDLSI* as a function of wire bandwidth in 22nm technology node (10mm global wires). Wire bandwidth is chosen to be the

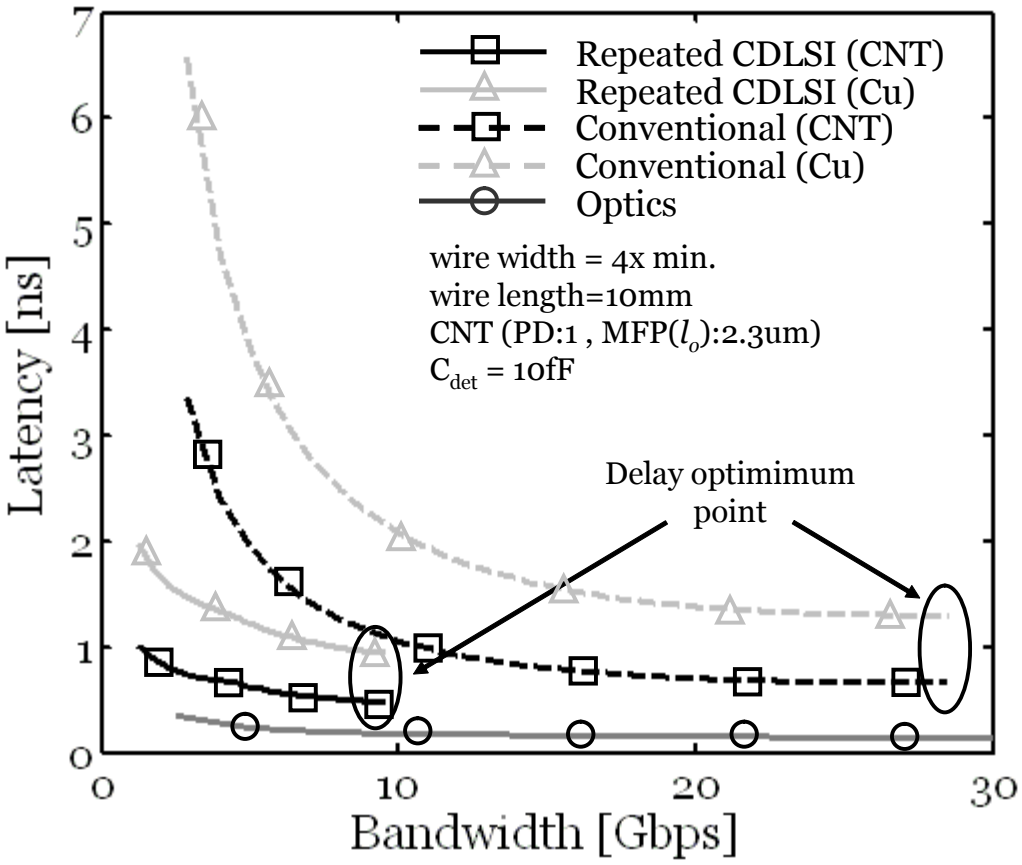

Fig. 3.6 Delay vs. wire bandwidth for both conventionally wire and *CDLSI* for 22nm technology node. Length is assumed to be 10mm. Wire width is fixed as 4 times the minimum wire half pitch

independent variable as it is likely to be dictated by the requirements of the multicore architecture. Given this requirement, the idea is to find the interconnect, which can do the job fastest and with least power. *CDLSI* shows a clear delay advantage below 10Gbps bandwidth range due to its pre-emphasis effect. However, the maximum allowed bandwidth for *CDLSI* is much less than the conventional wire due to the performance limits of transmitter/receiver. *CNT*s can further reduce delay for both

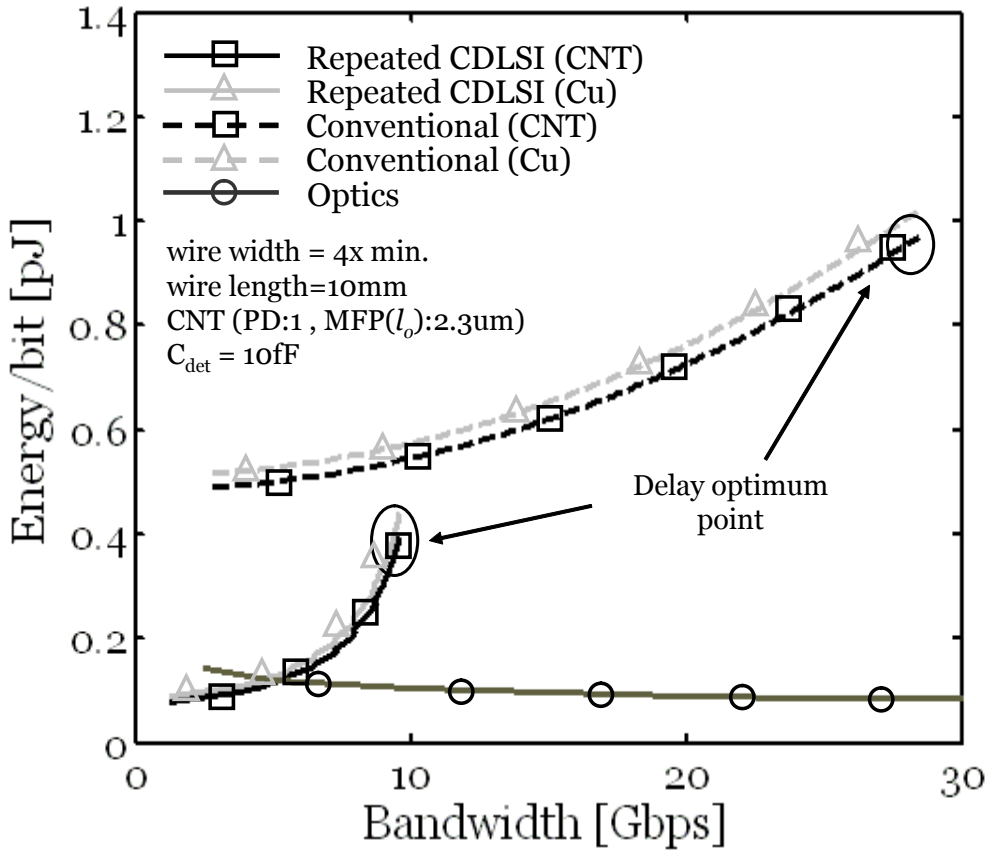

Fig. 3.7 Energy per bit vs. wire bandwidth for both wire circuit schemes for 22nm technology node. Length is assumed to be 10mm. Wire width is fixed as 4 times the minimum wire width.

circuit schemes with their advantage over *Cu* by improving *PD* to 1 and l_o to 2.3μm. Optics still shows the lowest delay for all bandwidth ranges. Fig. 3.8 compares the energy per bit for 10mm global wire under the same conditions as in Fig. 3.7. For both electrical signaling schemes, higher wire bandwidth requires more energy dissipation owing to its high demands for many repeaters. *CDLSI* presents much lower energy per bit than conventional repeated wire because of reduced voltage swing (50mV) from the wire. Optical interconnect is also favorable in terms of energy. However, below 5Gbps, *CDLSI* exhibits lower energy per bit than optics. *CNT*s don't help in reducing energy per bit as low resistivity does not impact energy consumed by repeaters.

3.3.2 Compound Metrics: Latency, Energy Density, and Bandwidth Density

Two methods can be used to increase bisectional bandwidth density. First is to fix a wire bandwidth, while decreasing the wire pitch as done in chapter 2, section 2.2.2.5. The second is to keep the wire pitch constant, while increasing the wire bandwidth. These approaches are shown in Fig. 3.8. The same approach can be taken for energy/bit vs. bisectional bandwidth density. In Fig. 3.9, we compare the lowest delay as a function of required Φ_{BW}, taken from the lower of the two approaches in Fig. 3.8. *CDLSI* shows better delay

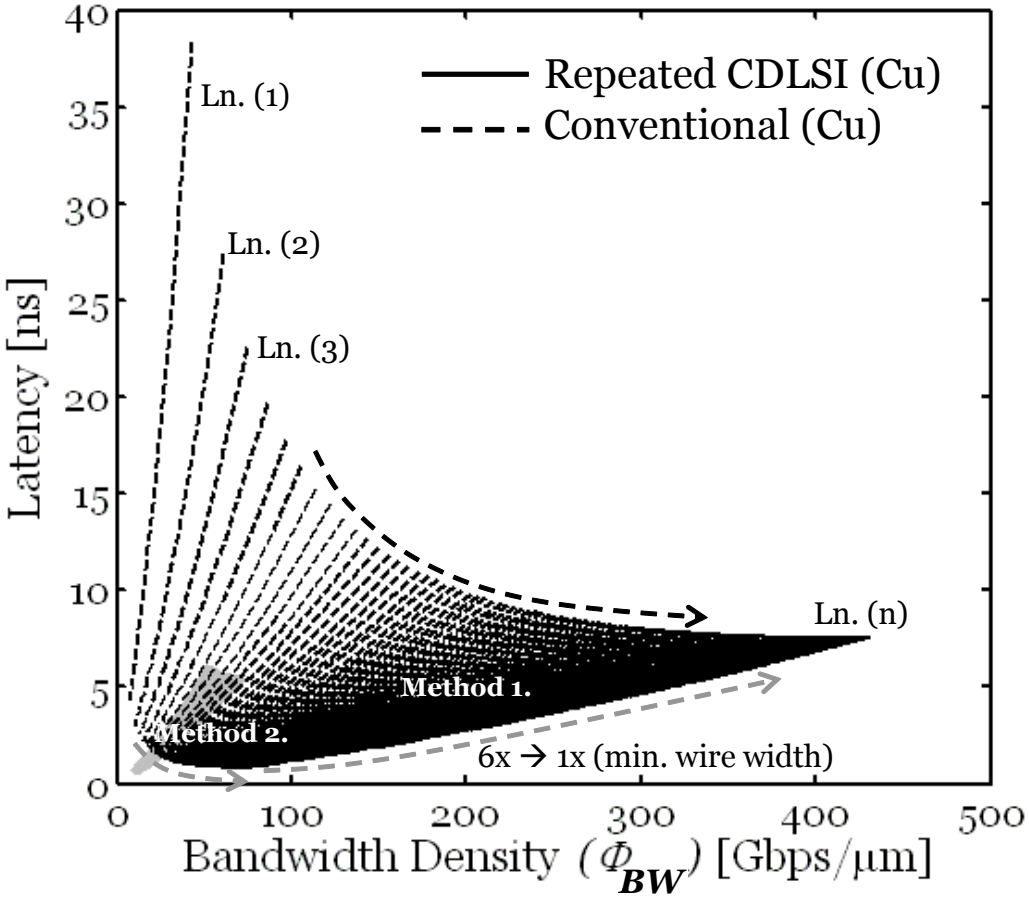

Fig. 3.8 Delay vs. bisectional bandwidth density (only for *Cu*). Wire pitch is decreased from 6x minimum wire width (ITRS defined) to 1x minimum wire width. Ln(n) is the delay- φ_{BW} line defined by method 1.

performance over the conventional wire only in the low Φ_{BW} domain. We observe that Φ_{BW} window of *CDLSI* is limited by two factors: firstly, its low intrinsic wire bandwidth caused by high transceiver delay as shown in Fig. 3.6,3.7 and secondly, 2x wire pitch occupancy of *CDLSI* due to differential signaling. *CDLSI*'s delay advantage over the conventional wire degrades as the system requires higher

Φ_{BW}. This is

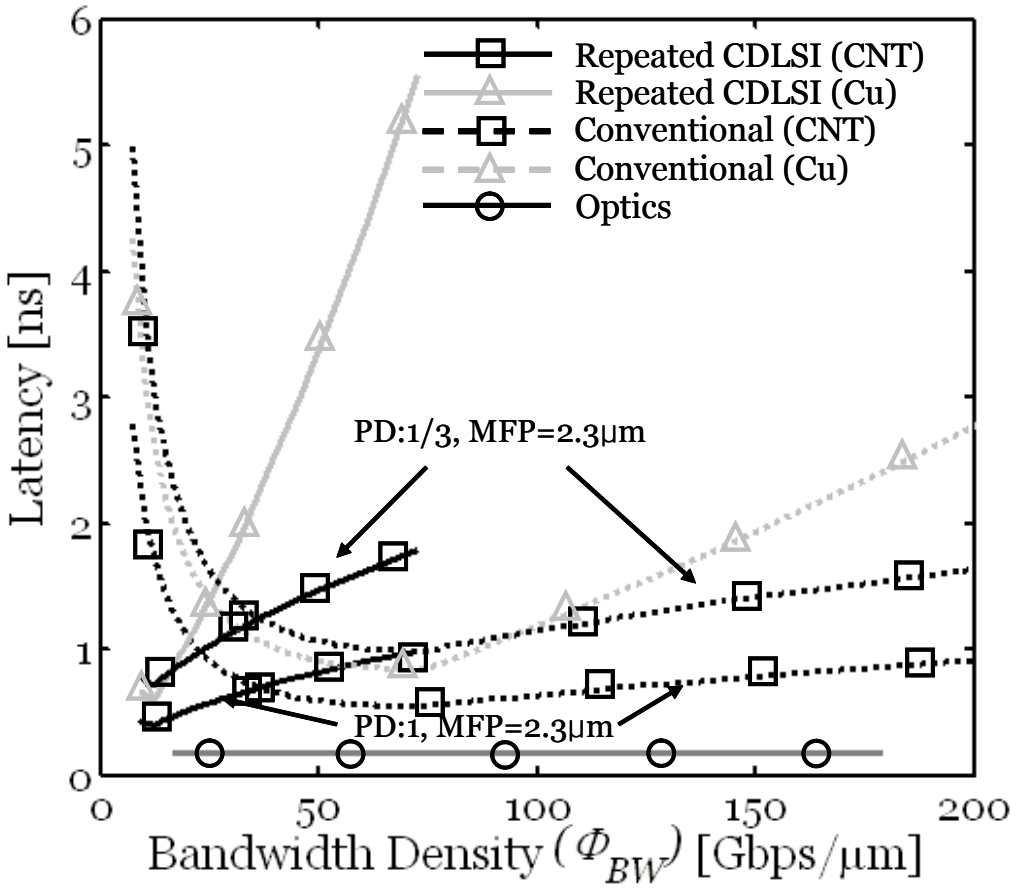

Fig. 3.9 Delay vs. bisectional bandwidth density *(φ_{BW})*. For optics, WDM technique is assumed. (Total 10 wavelengths width 10Gbps/ch bandwidth)

because too small wire width, in turn, results in too high resistivity, demands excessive number of transceivers, cancelling the advantage of pre-emphasis effect from the wire. *CNT*s with a higher filling factor (*PD*=1) can reduce delay further. For optics a larger Φ_{BW} is achieved using denser wavelength division multiplexing (WDM) and the maximum Φ_{BW} is limited by the degree (no. of wavelengths) of WDM [9]. Optics shows the lowest delay as compared to *Cu* and *CNT* technologies. Finally, Fig.

3.10 compares energy density (pJ/μm) as a function of system required Φ_{BW}. *CDLSI* also loses its energy advantage over conventional wires for high Φ_{BW} for the same reason as in Fig. 3.9. Optics, although difficult to inexpensively implement, shows the best performance.

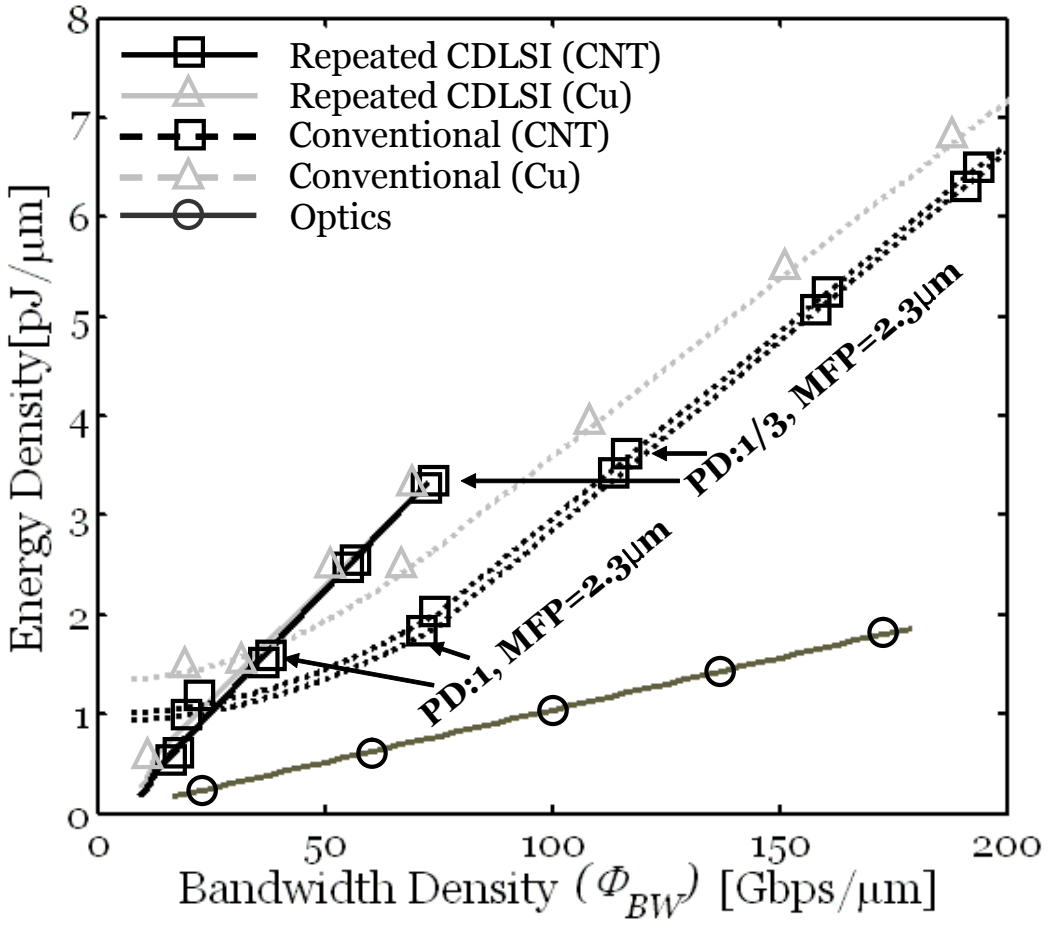

Fig. 3.10 Delay vs. bisectional bandwidth density *(φ_{BW})*. For optics, WDM technique is assumed. (Total 10 wavelengths width 10Gbps/ch bandwidth)

3.3.3 Impact of Via Blockage Factor

As the dimensions of wire cross section scale down, wire demands more repeaters to surmount high resistivity. These repeaters embedded in a substrate level need to be connected to global metal levels (metal 6-9) through many vias. Thus, the area consumption by repeaters and concurrent via congestion (or via blockage) become a non-negligible problem [10][11]. Via blockage severely degrades the wire routing efficiency due to its ripple effect from the via [12]. This is particularly important due to the growing complexity of routing and layout in gigascale integration. There are two main factors that motivate us to examine the impact of the *CNT* bundle and *CDLSI* on via blockage: 1) Lower resistivity of *CNT*s has correspondence with the reduction in a number of repeaters per wire (N_R); 2) For *CDLSI*, a relative slowness of sense amplifier causes increase in repeater spacing (l) because delay optimization process is about equalizing the latency between wire and repeater. For via blockage calculations, analytical formulation expressed in [12] is used. This is given as

$$A_{VB} = \sqrt{\frac{N_w P_w^{\,2}}{A_c}}\left[1+\frac{s\lambda}{P_w}\right] \qquad (3.12)$$

N_w is total number of vias per chip which can be found using a stochastic interconnect length distribution [11]. P_w is the metal wire pitch, A_c is the chip area, and s is a via

covering factor, typically assumed to be 3. The gate length is 2λ. For the calculation of

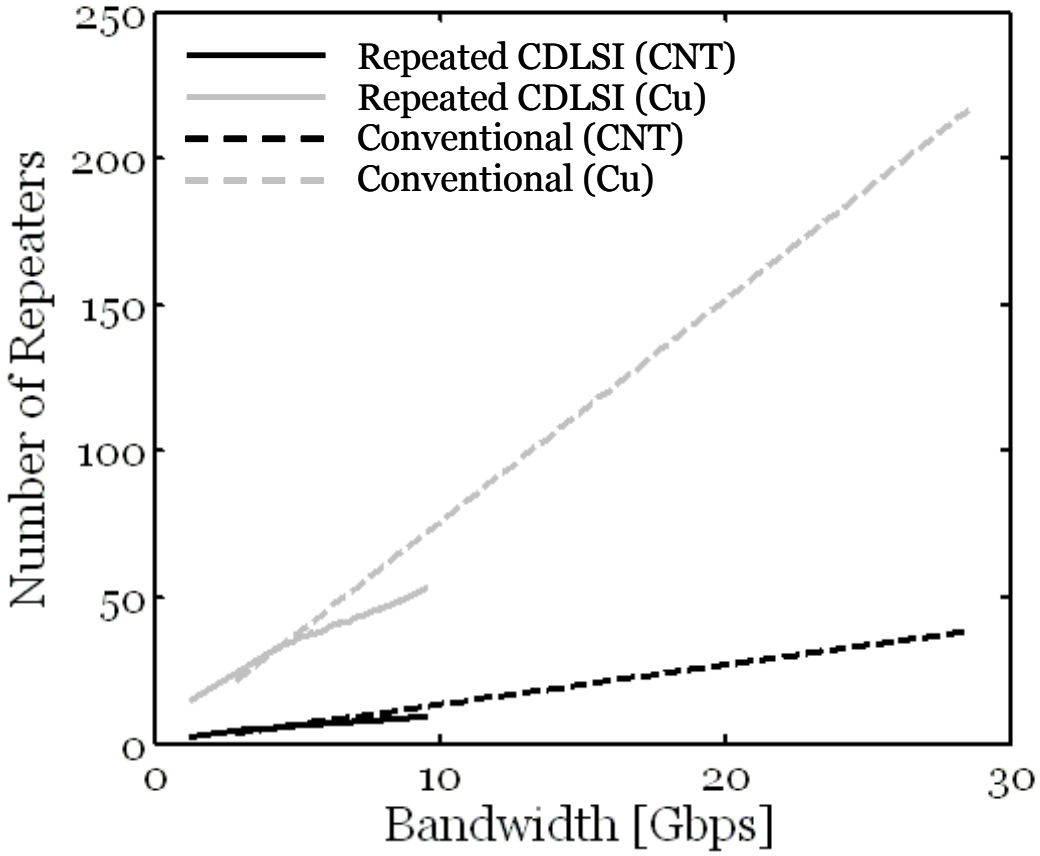

Fig. 3.11 Number of repeaters per wire vs. bandwidth. Wire length is 10mm (global). Minimum wire width at 22nm technology node is assumed.

N_w, Rent's exponent (p) is assumed to be 0.55 and 0.58. In Fig. 3.11, the number of repeaters per 10mm wire (N_R) is computed as a function of wire bandwidth based on different technologies (*Cu* and *CNT*s) and circuit schemes (*CDLSI*). It is shown that, at the delay minimum point, *CDLSI* reduces N_R approximately 4X compared to the conventional wire at the cost of bandwidth. *CNT*s can also reduce a large number of repeaters about 6 ~ 7X compared to *Cu*. Fig. 3.12 plots A_{VB} under the same condition in Fig. 3.11. Here, *CNT*s alleviates A_{VB} 2.7X compared to *Cu* mostly because of a

reduced resistance and number of repeaters. However, at the delay minimum point, *CDLSI* only improves A_{VB} approximately 18% compared to conventional wire despite of a significant reduction in the number of repeaters. Moreover, at the same bandwidth point, A_{VB} of *CDLSI* is even worse than the conventional wire although N_R is almost similar between them. These phenomena simply originate from its differential signaling scheme because it doubles the number of vias. In addition, lowering the intrinsic wire bandwidth also helps reducing via blockage factor because trading the power with the bandwidth or the delay comes with a reduction in number of repeaters (vias).

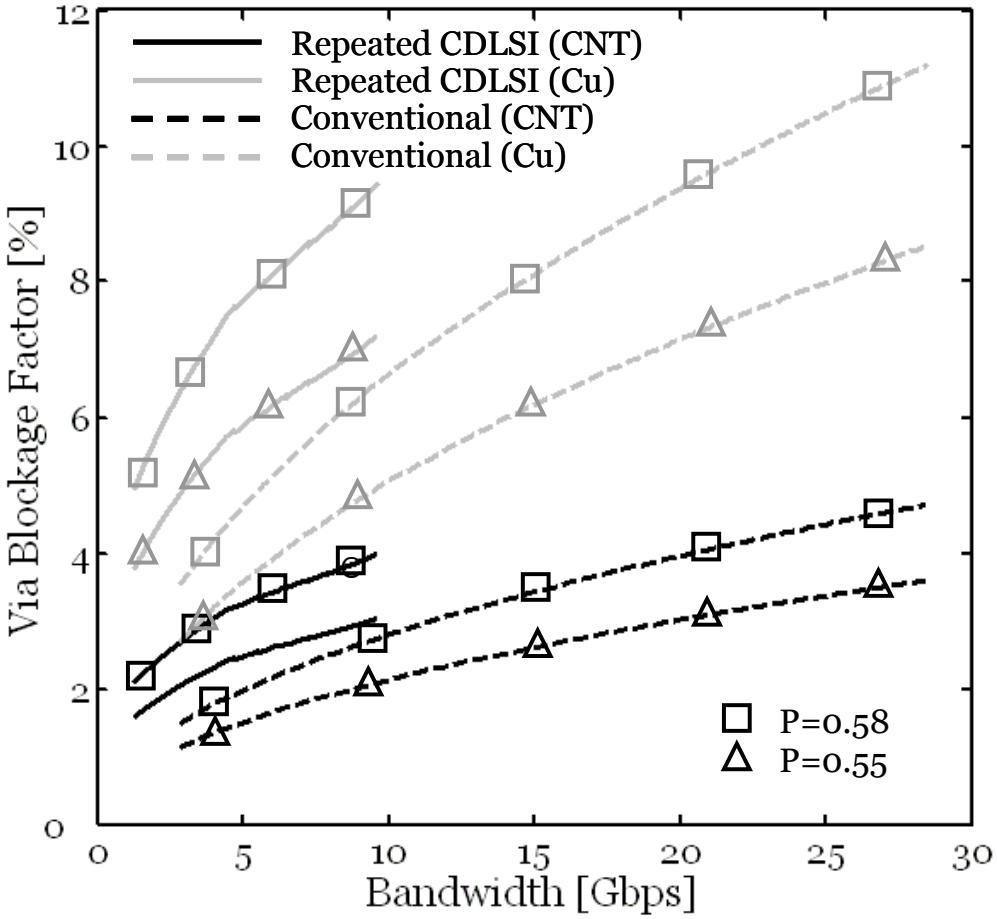

Fig. 3.12 Via blockage factor for global/semiglobal wire vs. bandwidth. Rent's rule coefficient is assumed to be 0.55 and 0.58. Minimum wire width is assumed.

Bibliography

[1] D. Liu *et al.*, "Power consumption estimation in CMOS VLSI chips," *IEEE J. Solid-State Circuits*, vol. 29, pp. 663–670, June 1994.

[2] P. Kapur, G. Chandra, and K. C. Saraswat, "Power estimation in global interconnects and its reduction using a novel repeater optimization methodology," *Design Automation Conference*, pp. 461-466, Jun., 2002.

[3] H. Zhang, V. George, and J. M. Rabaey, "Low-swing on-chip signaling techniques: effectiveness and robustness," IEEE *Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 8, no. 3, pp. 264-272 , Jun., 2000.

[4] S.C. Tan and X.W. Sun, "Low power CMOS level shifters by bootstrapping technique," IEEE Electronics Letters, Vol. 38, No. 16, 1st Aug 2002.

[5] R. Ho, T. Ono, R.-D. Hopkins, A. Chow, J. Schauer, F. Y. Liu, and R. Drost, "High speed and low energy capacitively driven on-chip wires," *J. Solid State Circuits*, vol. 43, no. 1, pp. 52-60, Jan., 2008

[6] E. Mensink, D. Schinkel, E. Klumperink, E. Tuijl, and B. Nauta, "A 0.28pJ/b 2Gb/s/ch transceiver in 90nm CMOS for 10mm on-chip interconnects," *Intl. Solid State Circuits Conference*, pp. 209, Sep., 2003

[7] B Wicht, T. Nirschl, and D. Schmitt-Landsiedel, "A yield-optimized latch-type SRAM sense amplifier," *EESCIRC*, pp. 409, Sept., 2003

[8] P. Kapur, G. Chandra, and K. C. Saraswat, "Power estimation in global interconnects and its reduction using a novel repeater optimization methodology," *Design Automation Conference*, pp. 461-466, Jun., 2002.

[9] K. H. Koo, H. Cho, P. Kapur, and K. C. Saraswat, "Permance comparisons between carbon nanotubes, optical, and Cu for future high-performance on-chip interconnect applications," *IEEE Trans. on Electron Devices*, vol. 54, no. 12, p. 3206, 2007.

[10] D. Sylvester and K. Keutzer, "A global wiring paradigm for deep submicron design," *IEEE Trans. on CAD*, pp. 242-252, Feb. 2000..

[11] J. A. Davis, V. K. De, and J. D. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI) – Part I: derivation and validation," *IEEE Trans. on Electron Devices*, vol. 45, no. 3, pp. 580-589, Mar., 1998.

[12] Q. Chen, J. A. Davis, P. Zarkesh-Ha, and J. D. Meindl, "A novel via blockage model and its implications", *Intl. Interconnect Technology Conference*, pp.15-18, Jun., 2000.

Chapter 4

DC and AC Characterization of Metallic Single Wall Carbon Nanotube

4.1 Motivation

Throughout chapter 2 and 3, *m-SWCNTs* have been studied for high performance on-chip interconnect application. Basic *RLC* values of *m-SWCNT* were assumed based upon mostly theoretical reports [1]. However, there exists experimental uncertainty about DC resistance (*R*) and AC reactance (*LC*) of *m-SWCNT*. For instance, existence of L_k is still obscure due to two conflicting reports [2][3]. Another example is the non-linear Ohmic behavior of m-SWCNTs, which have a high potential for low swing operation in local interconnect [4]. Therefore, it is imperative to conduct an in-depth study of *m-SWCNT* transmission line components both theoretically and experimentally.

Theoretically, the fundamental DC resistance of *SWCNT* has been predicted to be 6.45kΩ assuming no scattering events. Lüttinger liquid theory by Burke describes the

AC model of *SWCNT* as a nano-transmission line with a distributed *LC* components and predicts the kinetic inductance of 16nH/μm and the quantum capacitance of 100aF/μm [1].

On the other hand, there have been several experimental attempts to measure DC and RF characteristics of *m-SWCNT*. Yao et al. discovered the non-linear Ohmic behavior of *m-SWCNTs* in a DC regime [4]. Plombon et al. first reported the existence of kinetic inductance (8~43nH/*μm*) of m-SWCNTs with typical RF measurement and pad de-embedding technique [2]. There are also other reports about high frequency characteristics of SWCNTs [3][5]. However, all the aforementioned works conducted a characterization in a separate domains (DC or RF) and have not compared the extracted resistances in both domains.

This work not only performs separate DC and RF measurements on *SWCNTs* but also captures the consistency of the non-linear conductance with respect to the applied bias both from DC and RF measurements. Results also prove the existence of high kinetic inductance.

4.2 Fabrication of SWCNT Test Structure and DC/RF Measurement

Fig. 4.1. shows the steps for the fabrication of the test bench of *m-SWCNT* RF characterization. Ground-signal-ground type electrodes were fabricated for the high frequency s-parameter measurements on a *m-SWCNT*. The pad size is 90x90 μm^2 and

signal to ground pitch is 150μm. Pd/Cr (300nm/15nm thickness) layer was evaporated on a quartz wafer and lifted off as patterned. Cr pad layer is to give more adhesion of Pd to the substrate. Quartz substrate was selected for the purpose of minimizing the substrate capacitive coupling between Pd pads. *m-SWCNTs* were aligned between Pd pads using dielectrophoresis (*DEP*) technique. *DEP* is very useful technique to align *CNTs* by applying sinusoidal voltage signal with high amplitude and frequency to excite dipoles in *CNTs* [6]. Before DEP was applied, *SWCNTs* powder was diluted in an ethanol solution with proper sonication treatment. *SWCNT* solution was dispensed on Pd test pads. Then, immediately 25 $V_{peak\text{-}to\text{-}peak}$ with 10MHz frequency was applied to the pads until the solution dried out. Additional Pd caps were deposited on the contact edges between *m-SWCNTs* and Pd pads after the *DEP* alignment to provide better contact resistance. A 10 min annealing at 200°C in an inert ambient (N_2) was done to give a concrete orbital bonding between the *SWCNT* and metal, minimizing contact resistance. *m-SWCNTs* showing a current saturation at high bias among many samples were selected as device under test by measuring IV curve [4]. Fig. 4.2 shows the microscopic picture of completed RF GSG test structure. Fig. 4.3 shows the SEM image of aligned single *m-SWCNT* on the 4μm gap between Pd fingers.

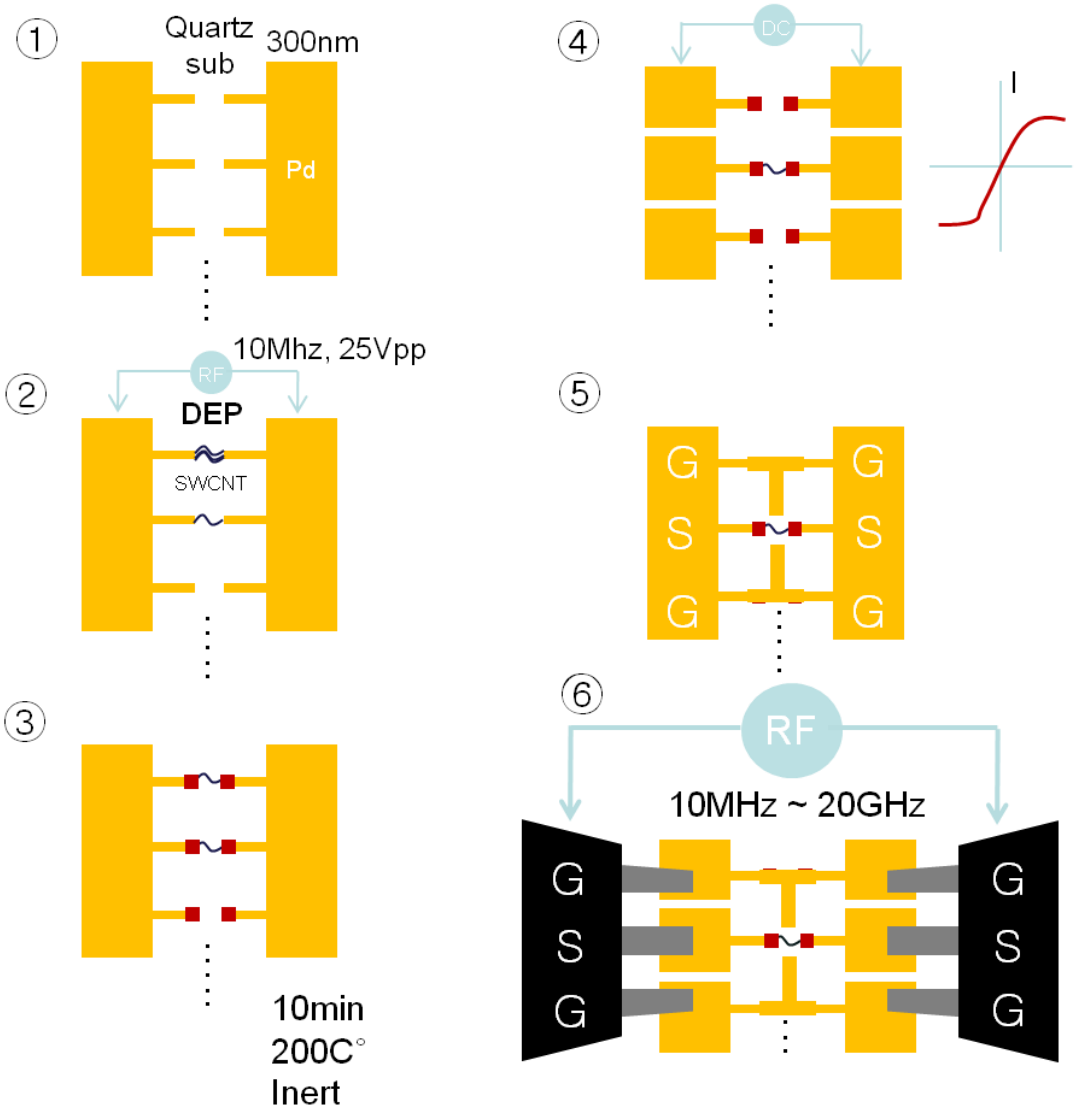

Fig. 4.1 (1) 300nm thick Pd metal pads are evaporated on a quartz substrate and lifted off. (2) DEP (100Mhz, 25Vpp) technique is used for the alignment of *SWCNTs* between pads. (3) Pd cap layer is evaporated on contact area between *SWCNT* and metal. Quick annealing (10min) at 200℃ in an inert gas is done after the contact layer deposition for better contact resistance (4) DC measurement is done to collect only *m-SWCNTs* (5) Additional metal deposition is done for complete ground-signal-ground structure. (6) RF measurement is done with ground-signal-ground probes and vector network analyzer tool

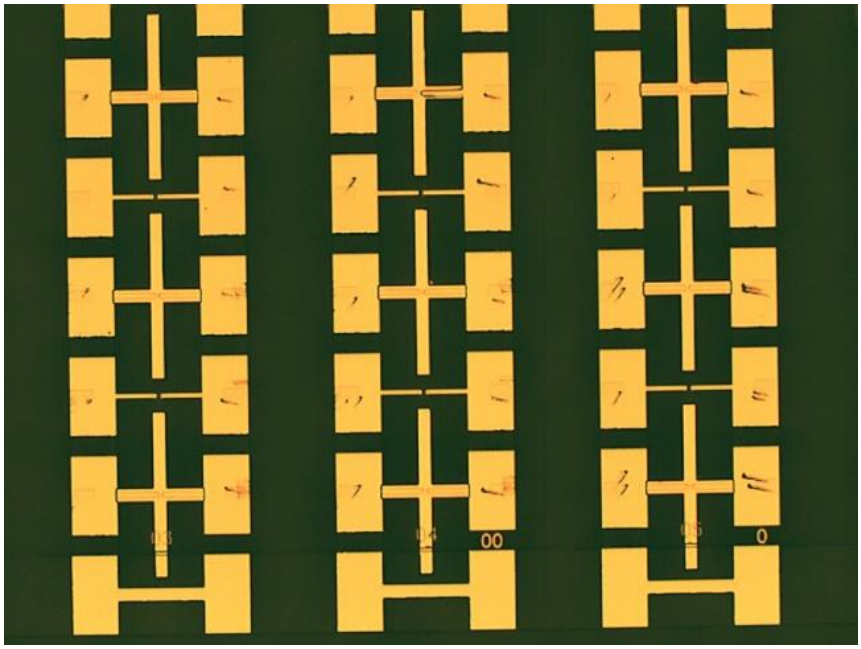

Fig. 4.2 Microscopic picture of the array of RF measurement test pads. CNTs are aligned between thin figures marked with a circle.

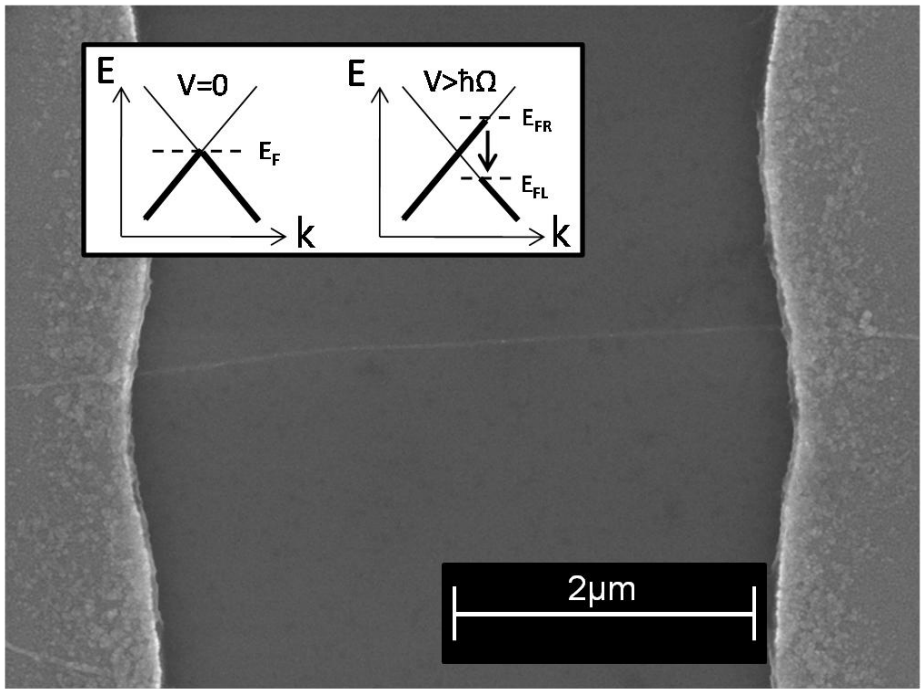

Fig. 4.3 SEM image of one metallic *SWCNT* aligned on the Pd pads. The gap between the pads is 4μm. Inlet is the E-k diagram of *m-SWCNT*. a) All electrons sit in the

states below the Fermi level at zero bias. b) Under high bias, right moving electrons below E_{FA} move to empty states in left moving band above E_{FB} after they reach the threshold energy ($\hbar\Omega$).

4.3 RF Measurement and Model Fitting of m-SWCNT Interconnect

The s-parameter was measured with an Agilent E8361C vector network analyzer. When RF measurements were performed for a microstructure, it is critical to remove all unwanted parasitic components unless all embedded structures are impedence matched. There are two sets of major contributors to these parasitics. The first set consists of transmission cables and RF probe tips. Second is Pd metal pad. All these components respond to high frequency input signal, preventing observers from seeing the outputs of device under test solely. The difficulty in measuring RF response of microstructure is well illustrated in Fig. 4.3

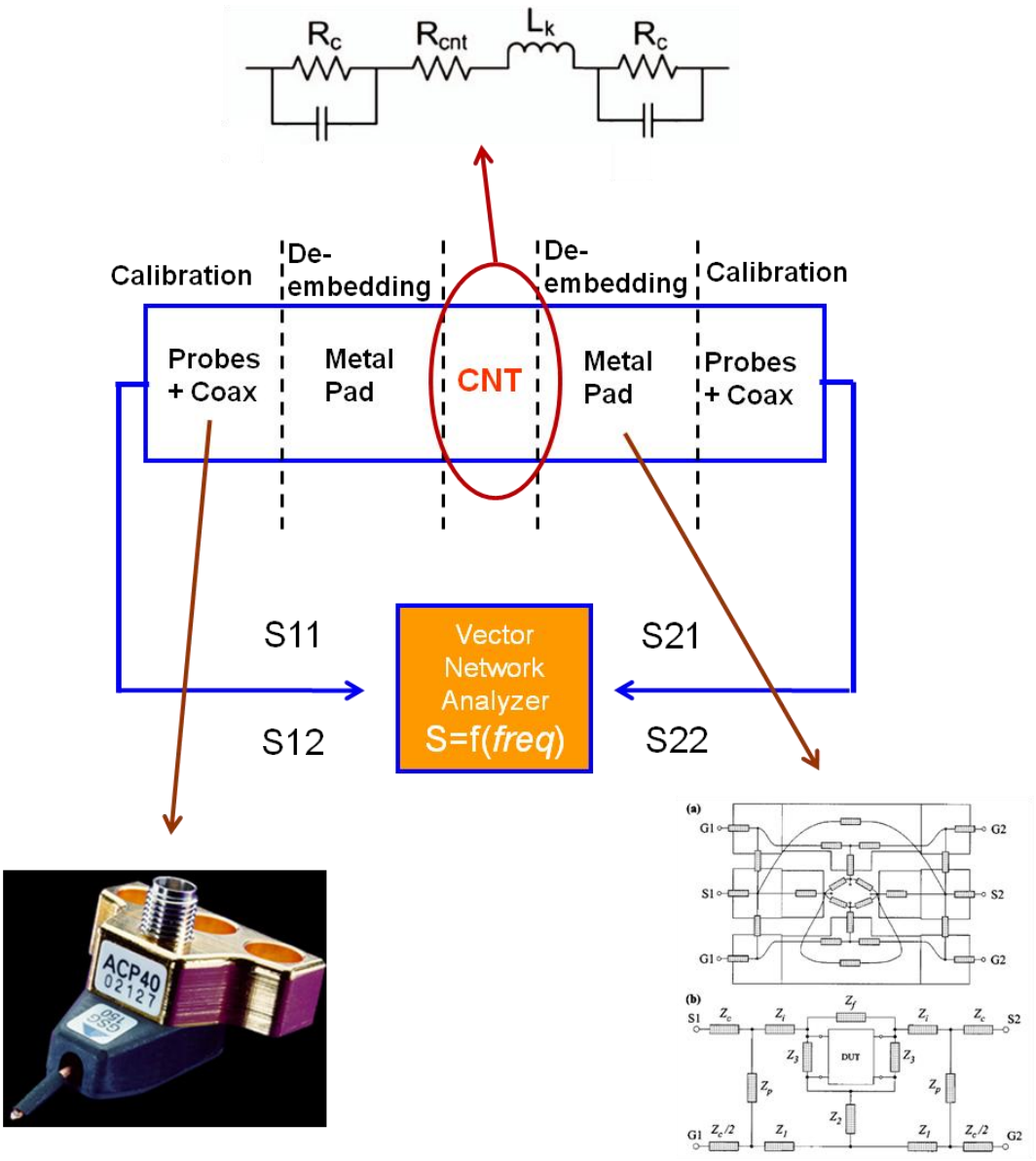

Fig. 4.4 Schematic of RF measurement test bench. It includes device under test (CNT), probes/cables parasitics, and Pd metal pad parasitics. They all respond to the RF signal

Thus, appropriate removing procedure for all unwanted parasitics should be applied. The standard short-open-load-thru (SOLT) calibration technique was employed to remove transmission line parasitic components of the co-planar

waveguide (*CPW*) and the RF probe tips (Cascade coplanar ground-signal-ground 150μm) [7]. Pad de-embedding procedure was applied to remove series and parallel impedance of ground-signal-ground pads. For this, a 3 step de-embedding method was used [8].

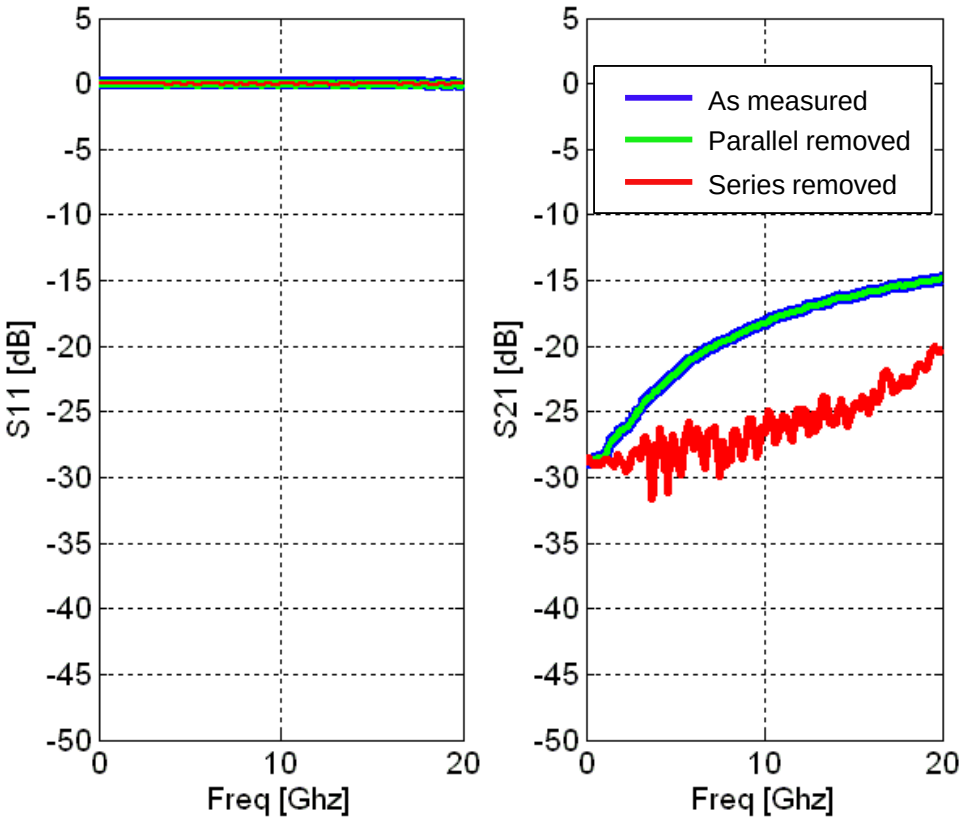

Fig. 4.5 S11 and S21 curves of RF measurement of m-SWCNT with vector network analyzer. The 3 step de-embedding procedure is shown in the plots. As measured, parallel removed, and series removed are marked with blue, green, and red respectively. It is shown that quartz substrate minimized capacitive coupling parasitics between metal pads.

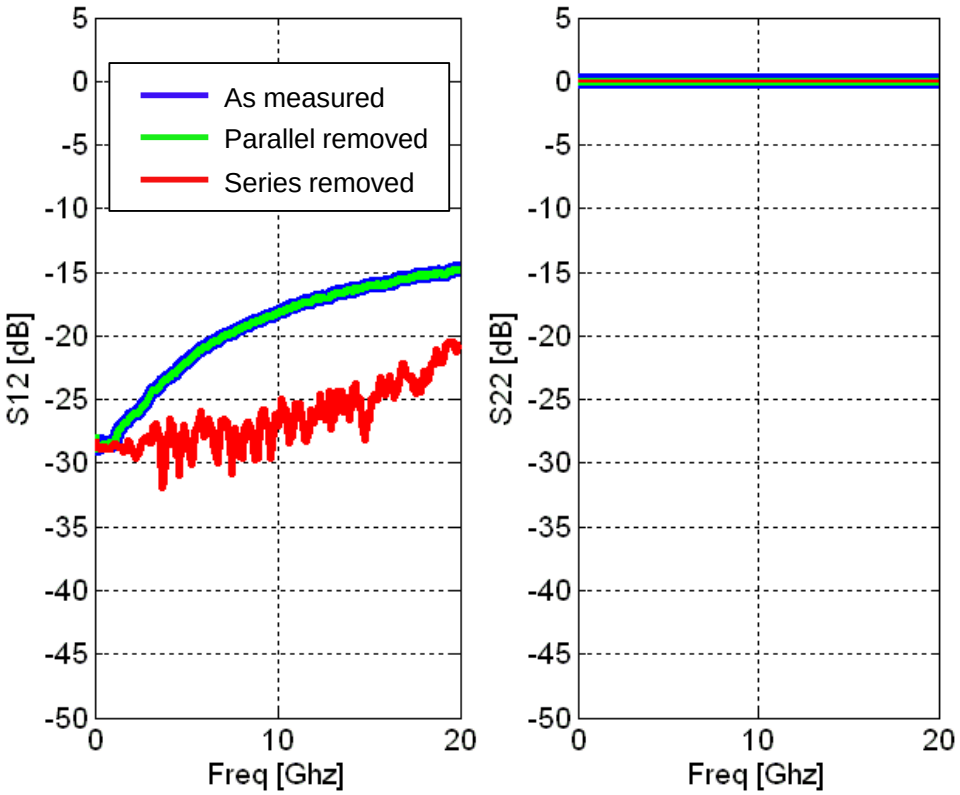

Fig. 4.6 S12 and S22 curves of RF measurement of m-SWCNT with vector network analyzer. 3 step de-embedding procedure is shown in the plots. As measured, parallel removed, and series removed are marked with blue, green, and red respectively. It is shown that quartz substrate minimized capacitive coupling parasitics between metal pads.

After the calibration for cables and probes was done, s-parameters were measured. The frequency spans from 10*MHz* to 30*GHz*. Fig 4.4 and 4.5 plot all sets of s-parameters (S11 S21 S21 S22), which are the ratio between input and output power described in eq. (4.1), as a function of frequency.

$$S_{xy} = 10 \cdot \log \frac{P_x}{P_y} \quad \text{(dB)} \quad (4.1)$$

where *P* refers to input/output power. The subscript x in S_{xy} means output port and y means input port. For example, S12 stands for the power ratio between input port 2 and output port 1. In each graph in fig 4.4 and 4.5, 3 steps of de-embedding procedure for pad parasitic are shown with different colors. It clearly shows the advantage of using quartz as a substrate because error from pad capacitive coupling can be minimized. After series impedance of pads is removed, s-parameters of *m-SWCNT* only are shown in red lines. The impedance of *m-SWCNT* was converted from the measured s-parameter data and compared with the model. The lumped circuit model used for the curve fitting can be described as [2].

$$R = 2 \cdot (R_c \| 1/ j\omega C_c) + R_{cnt} + j\omega L_k \qquad (4.2)$$

where, R_c and R_{cnt} are the contact and the channel resistances respectively. C_c is the coupling capacitance between the pad and the contact edges. L_k is the kinetic inductance. ω is the frequency of applied signal. Fig 4.6 shows the equivalent circuit of this model.

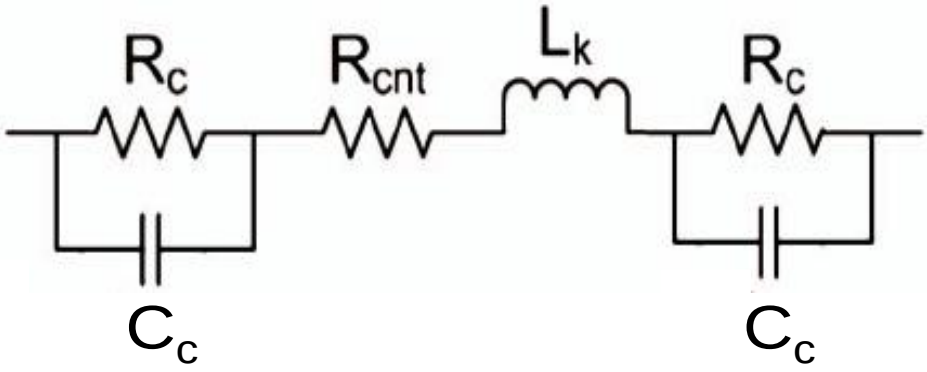

Fig. 4.7 Lumped equivalent circuit model of *m-SWCNT* in RF measurement test bench to fit the measured data.

Fig 4.8. shows the comparison between the amplitude of measured impedance and the fitted curves as a function of the frequency with an input power of -10dBm. Here, fitted R_c , R_{cnt} and C_c are 34 kΩ/μm, 4.1 kΩ/μm, and 40fF respectively. The amplitude drops at low frequency because C_c creates a short path from the pad to *m-SWCNT*. The amplitude tends to saturate as the frequency increases due to the rebalancing effect by $j\omega L_k$ in the *m-SWCNT* channel. Fig 4.9. shows the phase difference between the real and imaginary parts of the amplitude. It is clear that the model doesn't fit the measured data without L_k. L_k is fitted with 60nH/μm in this graph. This result is a relatively higher value than the theoretical value (L_k=16nH/μm). The results are in good agreement with the previous reports [2][5] and prove the existence of kinetic inductance in *m-SWCNTs*.

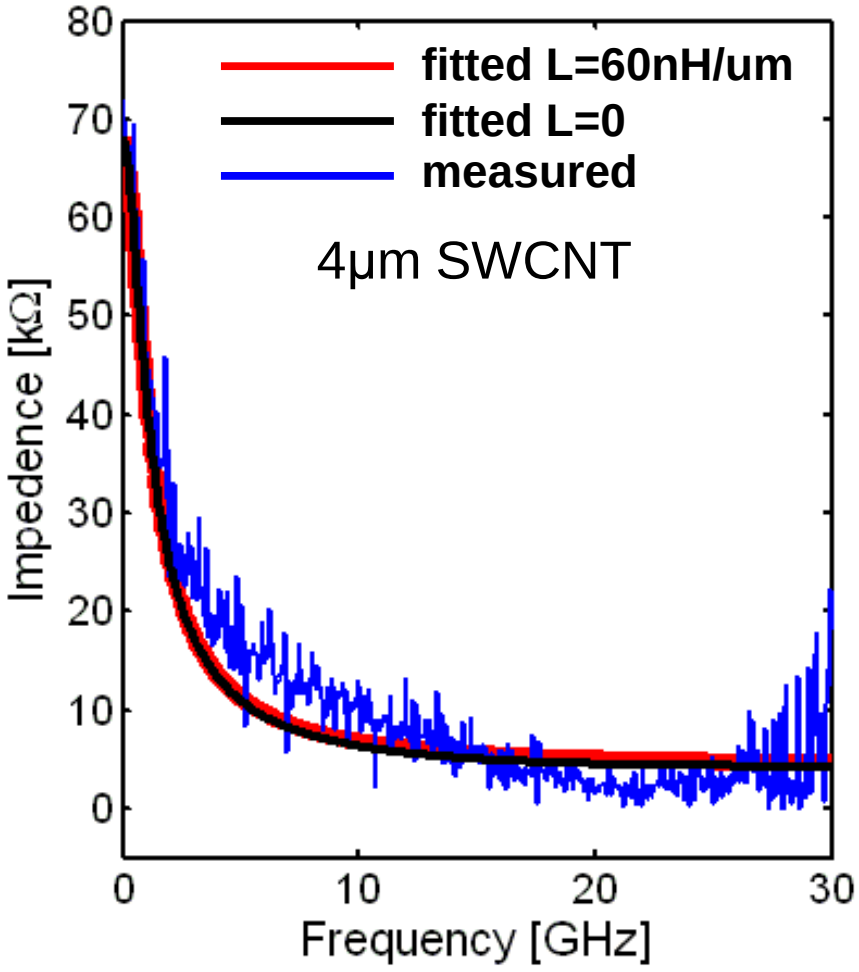

Fig. 4.8 Amplitude of impedance of m-SWCNT as a function of frequency. Blue line is measured data. Red line is fitted data with the model specified in fig. 4.7 and equation in eq. 4.2

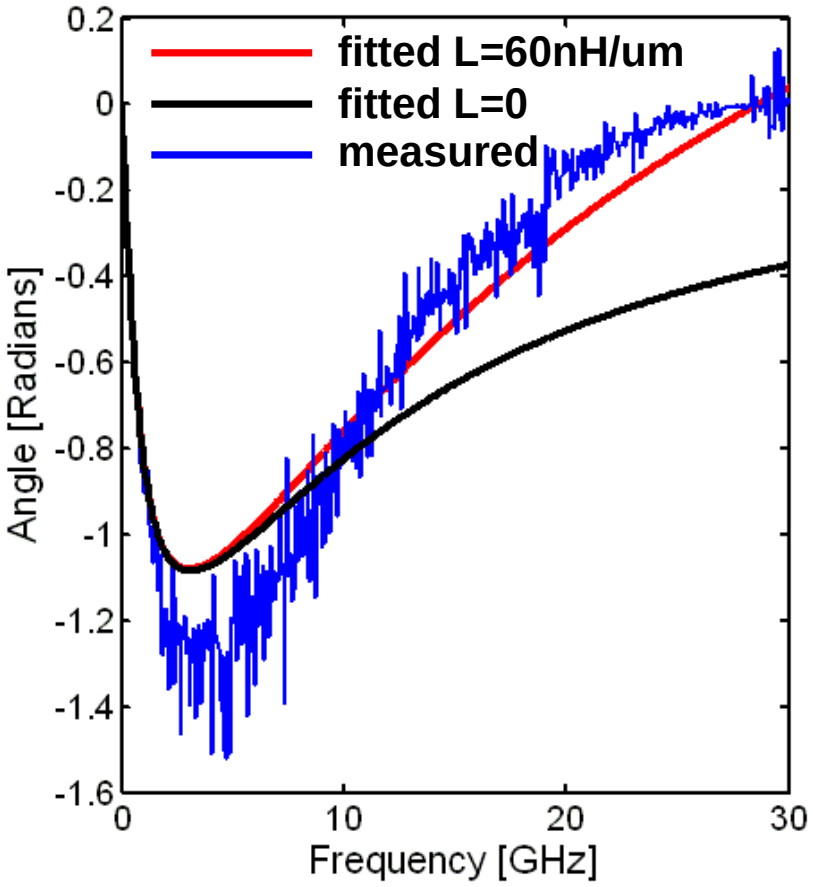

Fig. 4.9 Angle of impedence of m-SWCNT as a function of frequency. Blue line is measured data. Red line is fitted data with the model specified in fig. 4.8 and equation in eq. 4.2

	R_{cnt}	R_c	C_c	L_k
AC Results	4.1kΩ/μm	34 kΩ/μm	40fF	60nH/μm

Table 4.1 Extracted parameters in eq 4.2 with RF measurement of *m-SWCNT*.

4.4 DC Measurement and Parasitic Extraction

The DC current of a 4μm long *m-SWCNT* was measured. The current saturation around 30μA in fig. 4.10 proves the existence of the optical/zone boundary phonon emission when carriers attain a certain level of energy. This indicates that *m-SWCNT* shows different electrical properties in low bias and high bias. The DC resistance at low and high bias can be expressed as [4][9]

$$R_L = 2R_c + \alpha R_{MC} + R_Q \cdot \frac{l_{cnt}}{\lambda} \quad \alpha = \begin{cases} 0 \\ 1 \end{cases} \quad \text{at low bias} \qquad (4.3\text{ a})$$

$$R_H = 2R_c + \alpha R_{MC} + R_Q \cdot \frac{l_{cnt}}{\lambda} + \frac{V}{I_{SAT}} \quad \text{at high bias} \qquad (4.3\text{ b})$$

where R_Q is the fundamental resistance (=6.45kΩ), R_c is the side contact resistance, R_{MC} is the middle contact resistance, l_{cnt} is the length of the *m-SWCNT* and λ is the mean free path of the *m-SWCNT*. Here, R_c on the side contacts are assumed to be equal.

R_c and R_{cnt} values can't be extracted with a single I-V curve. Thus, we deposited a Pd rod (300nm width) in the middle of the same *m-SWCNT* samples used in RF measurements with E-beam lithography followed by lift off. The position of the rod is asymmetric in order to provide different lengths on either sides as illustrated in fig. 4.11 inset. α factor in eq. 4.3 becomes one for the sample with Pd rod and zero without Pd rod. We measured the I-V characteristics on three different sets of two-terminals a-b, a-c, and b-c. In eq. (4.3 a), there are three unknown parameters (R_c, R_{MC}, λ) and we have three equations with three different l_{cnt},*s*, each corresponding to three different I-V measurements. Thus, we can extract all unknown parameters. Fig 4.12.

shows three different I-V curves based on 4μm, 3μm, and 1μm long *m-SWCNTs* Fig 4.13 presents the resistance of each different length of *m-SWCNT*. The reason for the higher resistance at low bias for shorter *m-SWCNTs* is due to high contact resistance of the middle Pd (R_{MC}) rod. High resistance from a smaller width of middle Pd rod overpowers the reduction of channel resistance of a *m-SWCNT*.

Table 4.2 shows the results of the extracted parasitic components with 4 μm long *m-SWCNT*. We found that λ is 1μm, R_{CNT} is 6.5kΩ/μm *Rc* is 14 kΩ , and R_{MC} is 35 kΩ.

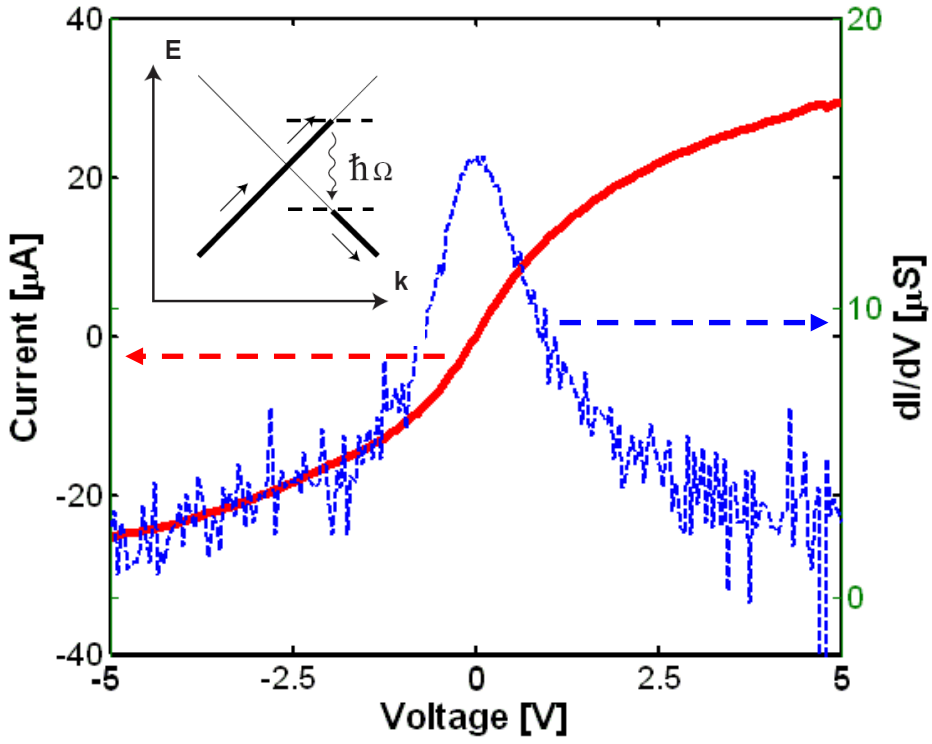

Fig. 4.10 Current and conductance vs. voltage plot of 4μm long *m-SWCNT*. Bold and dotted lines represent the current and conductance respectively. It shows the non-linear ohmic behavior of *m-SWCNT*.

	R_{cnt}	R_C	R_{MC}	λ
DC Results	6.5kΩ/μm	20.5kΩ	35 kΩ	1 μm

Table 4.2 Extracted parameters in eq 4.3 with DC measurement of *m-SWCNT*.

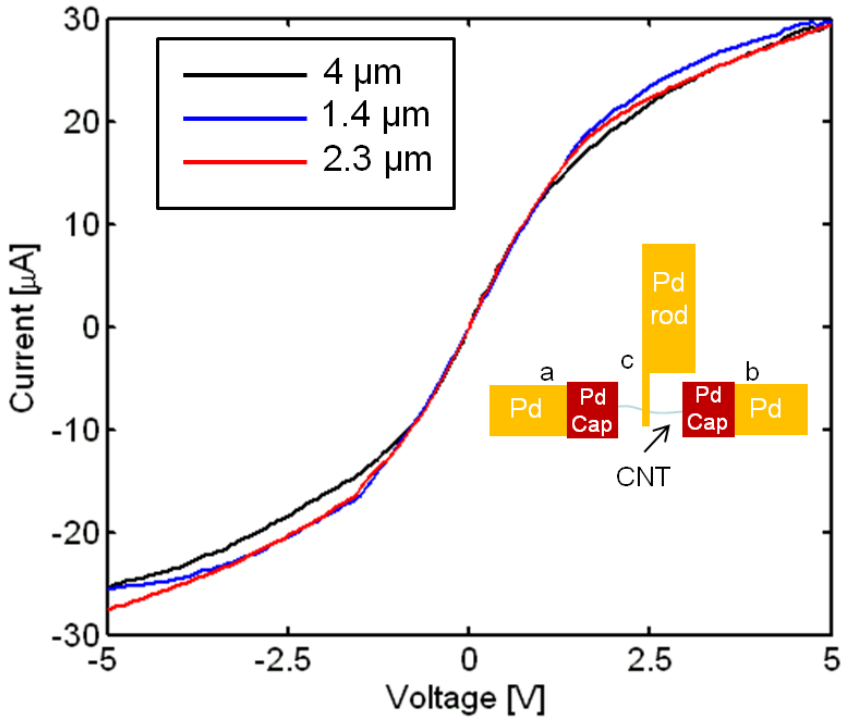

Fig. 4.11 Current vs. voltage of *m-SWCNT* with three different lengths. Inset illustrates additional Pd rod on *m-SWCNT*

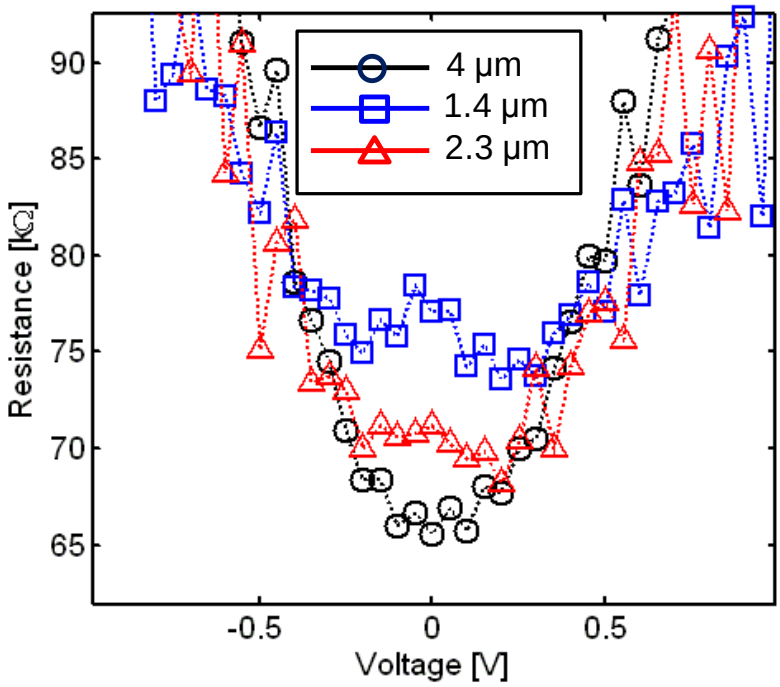

Fig. 4.12 Resistance vs. voltage of *m-SWCNT* with three different lengths

Table 4.3 compares all extracted parameters from DC and RF measurement of *m-SWCNT*.

	R_{cnt}	R_C	R_{MC}	λ	C_c	L_k
DC	6.5kΩ/µm	20.5 kΩ	35 kΩ	1µm	N/A	N/A
RF	4.1 kΩ/µm	34 kΩ	NA	1.57 µm	40fF	60nH/ µm

Table 4.3 Comparison between RF and DC measurement of m-SWCNT

DC resistance of *m-SWCNTs* is extracted from the RF measurement. This can be easily done with recording the resistance value at f=0Hz. The purpose of this extraction is to verify the consistency between RF and DC measurement and add more accuracy to the RF measurement. Fig. 4.10 clearly shows that the resistance of 4μm long *m-SWCNT* increases non-linearly as a function of input power of vector network analyzer. Input power in *dBm* unit can be translated to input voltage swing (V) because the source impedence of vector network analyzer has a fixed value (50Ω), which is indicated on x-axis in fig 4.13. The result of this graph shows agreement with a non-linear Ohmic behavior of a *m-SWCNT*, which was shown in DC current-voltage measurement of a *m-SWCNT*. This proves the consistency of the RF measurement results with DC measurement.

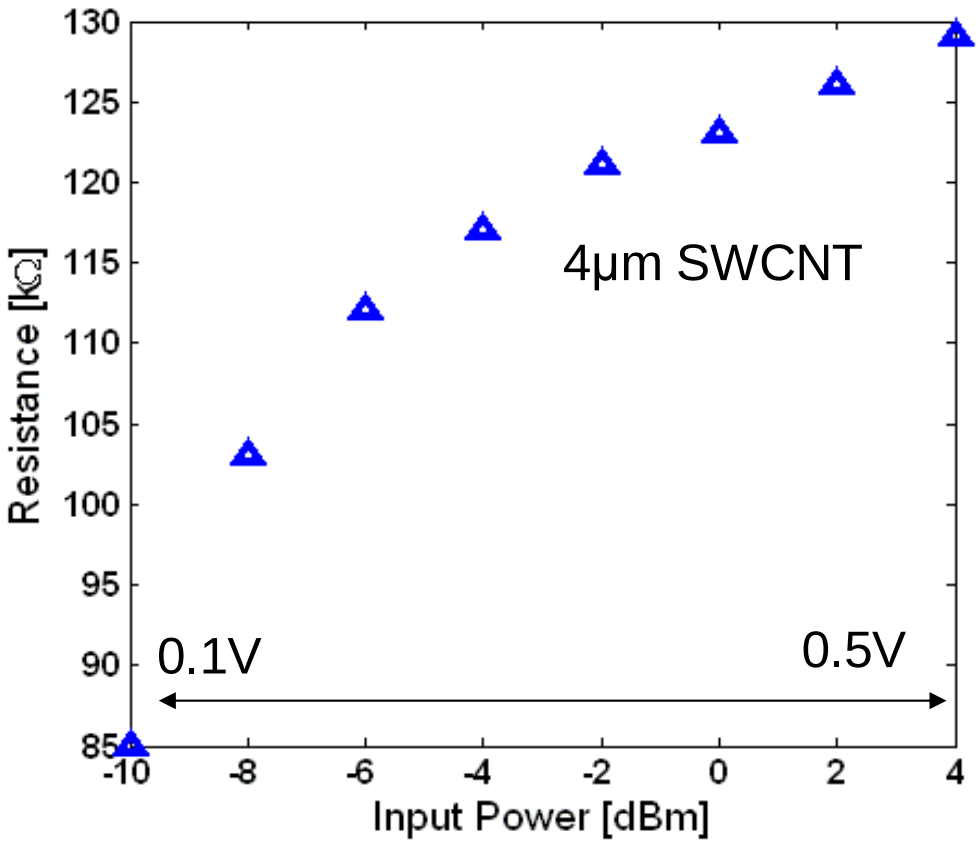

Fig. 4.13 extracted DC resistance from RF measurement of m-SWCNT. Input power is translated to corresponding voltage level.

Bibliography

[1] P. J. Burke, "Luttinger Liquid Theory as a Model of the Gigahertz Electrical Properties of Carbon Nanotubes," *Trans. on Nanotech.,* vol. 1, no. 3, pp. 129-144, 2002.

[2] J.J. Plombon, Kevin P. O'Brien, Florian Gstrein, and Valery M. Dubin, and Yang Jiao, "High-frequency electrical properties of individual and bundled carbon nanotubes." *Applied physics letters* vol. 90, pp. 063106, 2007.

[3] L. Gomez-Rojas, S. Bhattacharyya, E. Mendoza, D. C. Cox, J. M. Rosolen, and S. R. P. Silva, "RF Response of Single-Walled Carbon Nanotubes", *Nano Letters*, vol. 8, no. 9 pp. 2672-2675, 2007.

[4] Z. Yao, C. L. Kane, and C. Dekker, " High-field Electrical Transport in Single-wall Carbon Nanotubes," *Phys. Review Letters,* vol. 84, pp. 2941-2944, 2000.

[5] M. sun, Z. Xiao, Y. Chai, Y. Li, and P. C. H. Chan, "Inductance Properties of Silicon-in-Grown Horizontal Carbon Nanotubes", IEEE, Electronic Components and Technology Conference, pp. 1329-1334, 2010.

[6] H.A. Pohl, "Dielectrophoresis: The Behavior of Nuetral Matter in Nonuniform Electric Fields," Cambridge Univ. Press, 1978.

[7] Agilent Application Note, "On-Wafer Vector Network Analyzer Calibration and Measurements"

[8] H. Cho, and D. E. Burk, "A Three-Step Method for the De-Embedding of High-Frequency S-Parameter Measurements," IEEE *Trans. on Electron Devices,* vol. 38, no. 6, June 1991.

[9] A. Naeemi, and J. D. Meindl, "Design and performance modeling for single-wall carbon nanotubes as local, semi-global and global interconnects in gigascale integrated systems," *Trans. on Electron Device*, vol. 54, pp. 26-37, 2007.

Chapter 5

Conclusions and Future Recommendations

5.1 Conclusions

This dissertation presents an in depth study of limitations in current Cu/low-k interconnect technology and recommends possible alternates to replace it. These alternates include new material (CNTs), different signaling medium (Optics) and a novel circuit scheme (CDLSI). We estimated performance metrics of alternates and compared them with conventional Cu/low-k interconnects. Finally, we turned our focus to RF characterization of m-SWCNT to investigate their availability for use in interconnects. For this, we fabricated a m-SWCNT interconnect and performed s-parameter measurements to characterize both DC and RF properties.

First, we compared performance metrics of a CNT bundle and optics with Cu/low-k interconnect for local semi-global and global interconnects. For local interconnects, we have compared the performance of existing Cu wire with a CNT bundle based novel wire. In general, the CNT bundle shows lower latency than Cu. Moreover, CNT bundle's performance advantage can be further amplified by leveraging its superior electromigration properties and reducing the wire aspect ratio.

This reduces wire capacitance which dominates delay instead of wire resistance. We quantified these advantages by plotting the optimum latency as well as the optimum aspect ratio for all technologies. For global wires, we used proper inductance, resistance, and capacitance, modeling, as well as optical link optimizations to compare latency, power density/energy efficiency, and bandwidth density of Cu, CNTs, and optical wires. The first set of comparisons entailed latency and energy per bit as a function of technology scaling. We found optical wires to be superior, especially for longer links. For shorter semiglobal links, they yield comparable latency to CNT, especially at 22nm node, but they are worse in terms of energy per bit. The second set of comparisons involved fixing a technology node (22nm) and quantifying the latency and power density as a function of bandwidth density. The motivation was to take a systems approach from a multi-core architecture perspective. Bandwidth density was varied using different implicit parameters: wire pitch for Cu/CNT, and WDM for optics. Optical wires exhibit the lowest latency, while, the power density comparison depends on the *SA*. High *SA* yielded the best results, whereas at lower *SA* of 20%, there is a critical bandwidth density below which CNTs have the lowest power density. Finally, the impact of technological innovations on both optics and CNT were quantified.

In addition, this dissertation discussed the drawbacks of the Cu/low-k technology and a conventional repeater inserted signaling scheme for global interconnects. To overcome this, we proposed carbon nanotubes, optics, and the CDLSI circuit scheme as promising alternatives. For the proposed electrical interconnect circuit schemes

(CDLSI), we optimized the performances in terms of bandwidth, latency, and energy per bit by leveraging the delay and power, and subsequently compared them with the conventional repeated interconnect. We find that the repeated CDLSI system has a significant delay and energy per bit advantage over conventional wires although it is limited in the maximum bandwidth that it can achieve. For higher system bandwidth requirement, CDLSI loses its excellent delay and energy per bit advantage over the conventional wire, making CDLSI the approach for relatively lower bandwidth. Comparison with optics shows that it is still the most favorable interconnect candidate for the future high bandwidth/bisectional bandwidth density demanding interconnects. Last set of comparisons revealed that CNTs can significantly alleviate the number repeaters required and via blockage, whereas the impact of CDLSI on via blockage is not prominent despite of reduction in number of repeaters.

Finally, we identified the uncertainty of previous experimental results of m-SWCNTs, especially AC characteristics. Thus, we turned our attention to experimental AC characterization of m-SWCNTs. We also performed DC measurement in order to measure the discrepancy with RF characterization. We fabricated 4μm long m-SWCNT interconnect on a quartz substrate. Pd was chosen for the pad material to suspend interconnects. In order to accurately extract the m-SWCNT reactance, appropriate calibration and de-embedding technique were employed. Model fitting with measured s-parameters revealed the existence of a high kinetic inductance (~60nH/ μm). Subsequent DC (f=0Hz) resistance extraction from s-parameter measurements proves that the resistance of the m-SWCNT varies as

applied input power varies. DC current-voltage measurements with different lengths of the m-SWCNT was performed to extract hidden parameters of m-SWCNT such as contact resistance (Rc), and mean free path (λ), and they were compared with RF measurement results.

5.2 Future Recommendations

We have compared performance metrics of SWCNT bundle and Optics with Cu/low-k interconnects. However, there are still many possible options to replace conventional interconnects. For instance, performance comparison studies of interconnects with spintronics and plasmonics also need to be conducted with various constraints to investigate their applicability for future high performance VLSI applications. Performance comparison of those solution with conventional Cu/low-k interconnect is also imperative. In addition, performance evaluation work of CNTs and Optical interconnects should be extended to more application or architecture specific constraints to attain better insight for practical situation. Finally, DC measurements of short m-SWCNTs shows that they have great conductivity at low bias voltage. Having said the potential performance improvement with m-SWCNT wire in low voltage, it will be worthwhile to do more in-depth study about the applicability of a low voltage swing scheme with m-SWCNT in a local interconnect in order to achieve lower latency and power.