Advanced Interconnect Technologies for Future ULSI Applications

Srujan Meesala (08007036)
Electrical Engineering Department, IIT Bombay
srujanm@iitb.ac.in
Nanoelectronics (EE 718) Term Paper
April 2011

Abstract—Scaling trends and limitations of existing interconnect technologies are discussed and two prospective future solutions - carbon nanotube (CNT) and optical interconnects are examined in detail. The inherent unscalability of metal interconnects and degradation of their performance in the light of ever-increasing transistor density and performance is emphasized. Problems with multi-layer low-k/copper interconnects, which are currently employed by industry, are outlined and ongoing efforts to push their scaling towards the near future are summarized. The prospects of CNTs as high-performance electrical interconnects are evaluated. We subsequently consider the idea of optical interconnects as a radical solution. Recent advances in silicon nanophotonics to realize optical systems for high bandwidth chip-to-chip and global on-chip interconnects using ultrafast electro-optic modulators and low-loss waveguides are presented. Lastly, challenges of integrating photonics with existing CMOS technology are discussed.

Index Terms—metal interconnects, low-k ILD/Cu, 3D integration, air-gap technology, CNTs, silicon photonics, electro-optic modulators, low-loss waveguides, CMOS integrated nanophotonics

I. INTRODUCTION

Feature sizes on electronic chips over the decades have been exponentially decreasing as per Moore’s law. Smaller devices can not only be packed more densely, but are faster, more power-efficient and are of reduced cost. Electrical interconnects on the contrary, have shown a performance degradation with scaled sizes and denser packing. In addition to problems of power dissipation, crosstalk and reliability degradation [1], there has most notably been an increase in interconnect RC delays due to longer, more densely packed wires. With scaling, metal interconnects get slower not only relative to the transistor gate delay but also in the absolute sense. Fig. 2 shows normalized interconnect delays along with clock periods across Intel’s logic technologies. With RC delays approaching close to clock periods and a growing disparity between chip-to-chip signaling rates and on-chip frequencies, interconnects are now considered as significant as transistors for limiting ULSI density and performance [2]. Problems with scaling electrical interconnects became apparent in the early 80’s itself in an implicit manner, with inter-system buses in computers being constrained to operate at slower rates than the on-chip clock frequency. Today, the speed and efficiency of chip-to-chip communication, for instance between main memories and microprocessor cores, is constrained by the bandwidth of long interconnects rather than the devices at either end. With increased processor speeds in the future, this is predicted to be the case for on-chip interconnects as well.

Present-day semiconductor industries use multiple layers of copper interconnect lines separated by a low-k interlevel dielectric (ILD) to subvert increased line resistance and capacitive coupling (Fig. 1). Shorter lines between transistors and gates are placed in the lower layers, where density takes priority over delay. Long global lines such as those for clock distribution are placed in the higher layers and are thicker with a larger pitch. To tackle potential problems due to densely packed interconnects, there has been a steady increase in the number of layers from generation to generation. Fig. 2 shows a projected $0.74 \times$ decrease in minimum metal pitch with a $+0.5 \times$ increase in number of interconnect layers per generation across logic technology nodes for MPUs. Though interconnect density requirements are being met with this approach, the unstopped deterioration in performance can spiral into a major bottleneck for ULSI circuit performance. Further, addition of more interconnect layers leads to increased overall processing cost. It has been estimated that the process steps involved in forming interconnects comprised 50% of Intel’s total wafer process cost [4]. The necessity for improved interconnect technologies that keep up with aggressively scaled devices, therefore, cannot be underemphasized.

Two equally important directions of research have virtualized in this context. The first, largely spearheaded by the semiconductor industry, seeks newer materials for scaling existing
low-k/Cu interconnects to meet the challenges of upcoming technology nodes. This includes finding materials with lower resistivity for metal lines, porous low-k ILD materials and new processes to integrate these into existing technology [5]. In parallel, potentially radical solutions such as CNT and optical interconnects and their integration with conventional CMOS are also being investigated. CNTs promise to be very low resistance interconnects due to large electron mean free paths. The major roadblock to realizing CNT interconnects lies in achieving repeatable large-scale high-density growth of defect-free metallic CNTs. Optical interconnects, after a decade-long series of efforts, have met success in achieving high speed chip-to-chip communication - including, very recently, a 50 Gbps link realized exclusively with silicon nanophotonic devices by Intel [6] and the fabrication of a photonic die in the standard CMOS process flow [7]. These advances open up the possibility of overcoming the fundamental limitations of current metal interconnect technologies.

II. SCALING OF ON-CHIP METAL INTERCONNECTS

A. Increasing RC delays

The fundamental physical limitation to the scaling of all electrical interconnects is imposed by line resistances and capacitances, which tend to worsen with reduction in size and spacing. This RC delay limits on-chip signal propagation speed while the related rise time limits the signaling bandwidth for chip-to-chip communication. We presently consider time delay and deal with interconnect bandwidth in a subsequent section. The effect of scaling on time delay may be understood at a first order with simple RC transmission line models as illustrated in [2].

\[
R = \frac{\rho L}{TW} \quad (1)
\]
\[
C = 2(C_{\text{inter}} + C_{\text{intra}}) = 2\varepsilon \left( \frac{LT}{(P-W)} + \frac{LW}{D} \right) \quad (2)
\]
\[
RC = 2\varepsilon\rho L^{2} \left( \frac{1}{W(P-W)} + \frac{1}{TD} \right) \quad (3)
\]

$L$ - line length, $P$ - minimum metal pitch
$T$ - metal thickness, $W$ - metal width, $\rho$ - metal resistivity
$D$ - interlevel thickness, $\epsilon$ - ILD permittivity

While $C_{\text{intra}}$ is the capacitance due to coupling with adjacent lines in the same layer, $C_{\text{inter}}$ represents coupling with the layers above and below. $W$ and $P$ are usually close to the smallest features lengths on the chip, and are expected to scale as $0.7\times$ every generation. $T$ decides the metal aspect ratio, which is usually limited by the ability to etch metal spaces. Variation of $L$ depends almost exclusively on the interconnect layer of interest - it decreases in the lower layers due to reduction in gate pitch, while it may increase in the higher layers due to overall increase in chip area. If simple $0.7\times$ scaling is carried out with interconnects, from Eq. 3, we would expect nearly unchanged delay in the lower levels and greater than $2\times$ increase in delay in longer lines. This unacceptable performance degradation has been subverted by the industry by using a wide variety of approaches summarized in Table I.

![Fig. 2. Metal 1 minimum pitch, number of interconnect layers and associated RC delays plotted with data from ITRS 2010 tables](image)

![Fig. 4. History of low-k dielectrics used in IBM's technology nodes. Fig. from [5].](image)
To extend from Fig. 3, an important point to note is that increasing aspect ratio will not continually yield benefits, since beyond a certain factor, the increase in $C$ dominates the reduction in $R$. This is the reason average metal aspect ratios have been increasing at a very slow rate with the highest limit set to 2.0.

B. Low-k ILD/Cu interconnect stack

We now consider the structure of the low-k/Cu interconnect stack and relevant materials used in today’s interconnect technology (Fig. 5). The lowest level, sometimes referred to as metal 0, contains tungsten vias, which are embedded in a pre-metal dielectric and connect device contacts to metal layers above. Connections between adjacent transistors are typically realized in this layer itself using metal silicide lines. Almost all present day processes use the double damascene process to realize the copper interconnect stack (Fig. 6). This method involves initial deposition of dielectric, most commonly by PVD or PECVD, following which etched trenches are filled with the interconnect metal by electrochemical deposition (ECD). The excess or overburden metal is then removed by CMP. Prior to the deposition of copper, a barrier layer is required to prevent the diffusion of copper into dielectric. In addition, a capping dielectric layer is required to protect the deposited copper from subsequent chemical processing for higher layers.

1) Low-k ILD: Fig. 4 shows the evolution of low-k ILD materials across IBM's logic technology generations. In the initial days of copper interconnect technology, PECVD SiO$_2$ was used as the ILD due to its cost efficiency. Scaling requirements forced a shift towards low-k dielectrics, especially in the lower layers, where high density leads to increased intra-layer coupling capacitance. The current industry standard is PECVD carbon doped oxides (CDOs) or alkoxysilanes (SiCOH), which have $k=2.3-2.5$. CDOs were instrumental in reducing the average capacitance from the 65 nm to 45 node by about 10%. However, SiO$_2$ continues to be used in the higher layers, which are thicker, for instance the metal-8 layer in Intel’s 45 nm node technology [10] for reliability reasons.

2) Barrier layers: There are notably two barrier layers - one being the bottom and lateral lining, which is deposited prior to copper and the other being the capping. The barrier layers serve the ends of preventing copper diffusion during processing as well as act as an adhesion interface between copper and dielectric. The barrier liner must ideally have a large electronegativity, and hence transition metals or alloys are the preferred material, with the current standard being Ta/TaN [11]. Two major associated problems with such barriers are high resistivity, which leads to increase in effective line resistance, and reduced electromigration lifetimes [8]. The capping layer serves an etch-stop layer for subsequent fabrication, and its thickness is limited by dielectric etching capabilities. SiC/SiN ($k=4$) is the currently used material. Due to almost double value of $k$ compared to the ILD, the capping layer contributes significantly to effective interconnect capacitance [12]. This is expected to become more significant in future generations. Thus there is a critical requirement for thinner, low-k barrier dielectrics.
C. Conduction properties of copper lines

In addition to a performance crunch with circuit parameters, metal lines undergo severe degradation in their material properties with aggressive scaling in thickness. We discuss two important effects - electromigration and interface scattering, and their impact on interconnect performance and reliability.

1) Electromigration: Electromigration (EM) refers to the rearrangement of conductor material due to momentum transfer between electrons and diffusing metal atoms (Fig. 7). It occurs to a significant extent at grain boundaries and interfaces, where metal atoms are more easily displaced, and at corners and vias, where there are greater chances of current crowding. Apart from process factors, EM effects have a direct impact on interconnect reliability, and in the worst case, can lead to loss of electrical contact. The mean time to reach the critical void size for electrically opening an interconnect or mean time to failure (MTTF) is given by the empirical Black’s equation [13].

\[
MTTF = \frac{A}{J} \exp \left( \frac{E_a}{kT} \right)
\]  
(4)

\(A\) - metal cross-sectional area, \(J\) - average current density, \(E_a\) - an activation energy associated with metal diffusion

Clearly, MTTF worsens with scaling, and this is a major cause for concern, especially for long lines. For lines with length below a threshold called the Blech length, such as those in the lower levels, it have been observed that there is virtually complete EM immunity, when the current density is kept below a threshold \(J_{EM}\), which is a technology constraint. This is due to the creation of stable voids. The maximum current density \(J_{max}\) at a given technology node is set by circuit design requirements. Typically, the output \(J\) of an FO 4 inverter is used as a good estimate of this figure. From Fig. 8, which shows the predicted trends for \(J_{EM}\) and \(J_{max}\), we observe that \(J_{EM}\) will turn out to be a significant reliability constraint for current densities in future generations. Currently envisioned EM lifetime boosters include capping with high resistivity materials such as CoWP and CuSiN. However, this comes at a tradeoff with increased interconnect line resistance.

2) Grain boundary and barrier interface scattering: For copper interconnect lengths less than 100 nm, grain boundary and barrier interface scattering events will be the deciding factor for interconnect resistance. The determining physical parameter for the same would be the mean free path and Fermi velocity of electrons between two inelastic collisions with the interface. It is estimated that beyond the 32 nm node, copper resistivity of global wires due to these effects would be 40% more than ideal [15].

III. SHORT-TERM SOLUTIONS FOR METAL INTERCONNECTS

The core form of low-k/Cu interconnects is likely to be retained for technologies in the immediate future. Necessary modifications must be realized with materials and process engineering to meet performance and reliability requirements. We discuss some of the most likely short term solutions in this section.

A. Conductor engineering

ITRS predictions state that copper will continue to be used at least for metal 1 and intermediate layers in the near future. The major challenges to push copper interconnects towards future nodes would be posed by degraded EM behaviour and increased resistivity, as discussed in the previous section. EM lifetime may be improved by using Cu-alloys such as Cu/Al and Cu/Ti, but this brings with it an increased resistivity tradeoff. Resistivity is a major performance determining factor for global wires, and hence there is a critical need for engineering copper interfaces and microstructures. These approaches are clearly unlikely to yield breakthrough improvements in performance. As a result, there are strong ongoing research efforts to realize global optical interconnects.

1) 3D integration: The most realistic short term solution, however, is considered to lie in 3D stacking approaches using through-silicon-vias (TSVs). 3D integration as an obvious solution to improve interconnect performance was realized quite early on, with predictions of over 100% improvement in RC delay for the 45 nm node [16]. In addition, 3D IC technology comes with the advantage of increased transistor density and reduced chip footprint along with flexibility in circuit design and wire routing.

TSV technology essentially uses vertical copper vias through Si substrate to connect devices stacked on multiple dies. Major advantages on several fronts are immediately realized due to a drastic reduction in line length. The process of realizing TSVs involves the following process steps -
Fig. 9. ITRS technology roadmap progression for low-k ILDs. Fig. from [8]

Fig. 10. RC plot of 35 nm airgaps compared with SiO$_2$. Also shown is a TEM image of the fabricated airgap structure. Fig. from [17]

1) Etching holes in substrate. This process might involve etching through non-silicon layers such as low-k.
2) Isolation layer to electrically isolate the via from the Si substrate. This is called the TSV liner, and it determines the capacitance of the interconnect.
3) Barrier layer to prevent diffusion of metal into silicon.
4) Filling the etched trenches by Cu ECD followed by CMP removal of excess Cu.

A major challenge for TSV technology lies in achieving high aspect-ratio uniform, defect-free trenches for placement of vias. In addition, there are concerns of TSV reliability and heat dissipation.

B. ILD engineering

1) Porous low-k ILDs: ITRS requirements consider low-k ILDs as a difficult challenge for future interconnect technologies. There is a critical requirement for dielectrics with $k_{eff} < 2.3$ in addition to capacitance reduction from the barrier layer to achieve lower $k_{eff}$'s. Though a wide variety of porous low-k materials are being investigated, there are major integration difficulties, since these dielectrics have poor mechanical strength. Mechanical reliability is an important factor considering harsh packaging and assembly conditions. For this reason, it is projected that PECVD SiO$_2$ will continue to be used for the higher layers. Additionally, porous dielectrics may not be able to withstand etch damage from processing steps for higher layers.

2) Air-gap technology: For the near future, the most promising technology considered for ILD is a hybrid of airgaps and low-k materials. Air-gaps represent the ultimate material limit in terms of achieving low capacitance (Fig. 10), and integrating them with metal interconnects is considered one of the most significant challenges for the next decade. A wide variety of integration schemes have been studied so far, which may be categorized as below [5]. Brief process flows are shown in Fig. 11.

1) Local air-gaps: This scheme involves creation of airgaps selectively around copper wires by etching low-k ILD in these areas. This process needs to be carried out on a layerwise basis. There is a requirement on the filling capability of the dielectric so that the trench created for the airgap remains more or less unfilled during the deposition of the next interconnect layer. Additional process steps may be required to preserve dielectric in regions around tungsten vias to avoid the undesirable consequences of improperly placed vias on copper wires. The major advantage with this approach is that it preserves mechanical strength of the interconnect stack, and at the same time achieves the end of a low-k intermetal material. The large number of processing steps layerwise is a significant drawback.

2) Global air-gaps: A selectively deposited sacrificial material, typically a polymer is removed after deposition of the entire interconnect stack by evaporation. This achieves large air-gaps in the structure. Though this approach is cost-effective since it achieves single-step airgap creation, it leaves concerns of mechanical reliability in addition to the effects of a high thermal budget on underlying devices.

C. Material engineering of barriers

Presently, interconnect barriers, both the liners and dielectric capping layers have a significant impact on line resistance and capacitance respectively in addition to detrimental effects on EM lifetime. It is projected that Ta/TaN will continue well onto the 22 nm node due to efforts on achieving improved PVD processes for the same. The future solution is however expected to be ALD TaN barriers, since ALD helps achieve
achieve thinner, defect-free films [18]. Low throughput and high cost of ALD is a major concern. Future solutions for the dielectric capping layer include selective metal caps such as CoWP, NiMoP [8].

IV. CARBON NANOTUBE INTERCONNECTS

Carbon nanotubes (CNTs) have gained increasing prominence over the past few years due to their highly desirable electrical properties. They have become an integral part of major research efforts directed towards future interconnect technologies. There are three major reasons for active interest in CNT interconnects -

1) Low resistivity due to quasi-ballistic 1D transport
2) Much higher limiting current densities, which help subvert electromigration problems with metals
3) High potential scalability down to 1 nm

Table II provides a brief history of a few major achievements in this area of research. In this section, we review the potential performance of CNTs as interconnects from a fundamental approach and present advances in fabrication and integration with VLSI technology.

A. Performance advantages

Extensive work has been carried over the past decade on modeling the performance of CNT interconnects in terms of circuit parameters such as line resistance and capacitance and also on their reliability. It has been calculated that SWCNTs can offer performance improvements of as much as 80% over copper interconnects at the 22 nm node in terms of RC delay, and even more as technologies are further scaled (Fig. 13). On the reliability front, experimental studies have indicated that CNTs can withstand current densities as high as $10^9$ A/cm$^2$ without any adverse effects [28] (Fig. 14).

The most commonly accepted equivalent circuit model for a CNT interconnect is shown in Fig. 15. We now explore the physical origins of line resistance and capacitance in these interconnects.

1) Resistance: Although CNTs have high current carrying capabilities, when isolated, they also have a high resistivity (higher than Cu lines). There is a fundamental lower limit on the resistance of any 1D electron system in contact with a metal, which arises due to the fact that the discrete energy states of the wire overlap with only some states in the energy continuum of the metal. For a perfectly contacted CNT, this limit is [30].

![Fig. 15. Equivalent circuit representation of a CNT interconnect line](image_url)

![Fig. 12. (Left) AFM tapping mode image of the first metallic CNT, which was characterized to show quantum wire like behaviour (Fig. from [19]). (Right) Molecular illustration of single-walled (SW) and multi-walled (MW) CNTs (Fig. from [20]).](image_url)

![Fig. 13. Estimated maximum improvement in RC delay of SWCNT bundle interconnects over Cu lines for various technology nodes plotted against the mean free path achieved. The mean free path of real SWCNTs would be decided by defect density levels (Fig. from [29]).](image_url)

![Fig. 14. Resistance stabilities of CNTs at a current density higher than $10^9$ A/cm$^2$ at 250°C in air (Fig. from [28]).](image_url)
Fig. 16. (Above) Comparison between calculated resistance of dense CNT interconnects and Cu wires at the 22 nm node as a function of interconnect length (Fig. from [32]). (Below) Total effective resistance of SWCNT, MWCNT bundles compared with Cu wires of equal diameter as a function of frequency (Fig. from [33]).

\[ R_{\text{min}} = \frac{h}{4e^2} = 6.5k\Omega \] (5)

In reality, imperfect CNT-metal contacts lead to a significantly higher resistance. However, the major advantage of CNTs over Cu is realized by the easy growth of CNT bundles, which lead to a much lower effective resistance. So significant is the effect of bundling that metallic CNT bundles have an order of magnitude lower resistance than comparable copper lines (Fig. 16). Another favourable electrical property is long mean free path (\( \lambda_{\text{MFP}} \approx 1\mu\text{m} \)). This is attributed to band-structure symmetry, which prohibits back-scattering events in the first subband [31]. Long CNTs, however, exhibit resistance due to inelastic scattering events with acoustic phonons. In essence, quasi-ballistic transport is achieved even for long interconnects. Additionally, negligible skin effect in CNTs provides them an upper hand over Cu wires at high frequency operation (Fig. 16).

2) Capacitance: As in the case of resistance, CNT interconnects have a fundamental quantum term. This term arises due to the change in electrochemical potential when an electron is added just above the Fermi level, and depends on the density of states [31]. In addition, there is the usual electrostatic capacitance due to bundling of CNTs as well as coupling with adjacent conductors. It is found that the electrostatic term significantly dominates the quantum term. It is important to note that this term would depend merely on the geometry of the conductors, and we would therefore not achieve a drastic improvement over a similar system with copper lines. Indeed, at high metallic CNT densities only a 6% reduction in capacitance is observed over copper interconnects (Fig. 17).

Fig. 17. Electrostatic capacitance of CNT bundles of varying metallic density and diameter compared with Cu wires of corresponding diameter (Fig. from [32])

Fig. 18. Process flow to realize CNT bundle vias bonded to copper damascene lines - Fig. from [26].

B. Fabrication and VLSI integration

1) CVD CNTs as vertical interconnects: Growth technology for CNTs has progressed significantly since their discovery. However, it is still considered immature for integration with VLSI fabrication. Growth of CNTs on semiconductor substrates is carried out typically by CVD at high temperatures of 400-1000°C. These techniques first require the deposition of metal catalysts on the substrate. Vertical CNTs are realized at the sites with metal atoms. Growth of dense SWCNT bundles with reasonable chirality control has been realized only recently [33].

Research on on-chip integration of CNTs has mostly focused on CNTs for vertical interconnect (via) applications. We now describe the most commonly used approach to integrate vertical CNT bundles as vias with standard copper damascene interconnect technology (Fig. 18).

1) The interlayer dielectric is first deposited and then etched to form trenches for CNT vias.
2) A thin layer of metal catalyst for CVD growth is then deposited inside the trenches.
3) After CVD of CNT bundles, the wafer is spin coated with an SOG, and then planarized by CMP.

Uniform deposition of catalyst, which is essential to achieve high CNT bundle density, is carried by carefully controlling the size of metal nanoparticles. Reliable CNT bundle integration with a low-k ILD has been achieved with low temperature CVD at 365°C (Fig. 19). These developments open up the possibility of CNT bundles as a high performance solution for...
vias in a 3D interconnect technology, such as TSV.

2) Global CNT interconnects: Growth of long, horizontal CNTs is a significant challenge, because CNT bundles naturally grow perpendicular to the substrate. Challenging catalyst deposition steps would be required to realize grids of horizontal CNTs. The most common technique currently used to fabricate horizontal CNT interconnects is electric field induced assembly of commercially grown MWCNTs, which exploits their dielectrophoresis property. Horizontal MWCNTs with densities of millions/cm² have been achieved with this approach [24]. The process flow steps are summarized below (Fig. 20).

1) The wafer is passivated and an array of Au electrodes required for electrophoresis process is deposited on the passivation layer.
2) MWCNTs are then dispersed in ethanol and an alternating electric field is applied between the electrodes to precisely position the CNT.
3) The CNT is then secured by depositing metal clamps on top of the electrodes.
4) Subsequently, the passivation layer is etched over the vias and connections with the Au electrodes are realized.

Integration of a top layer of MWCNT interconnects realized by this process with a 0.25 µm 1 GHz CMOS technology has been achieved [25].

C. Challenges

• Dense CNT bundles - Since CNTs rely heavily on lower resistance due to bundling, achieving dense bundles is necessary to outperform copper interconnects. This would require engineering of catalyst particle size and material. ITRS projections state that the ideal density for a densely-packed all-metallic bundle of SWCNTs with 1 nm diameter is 0.66 nm⁻² [8]. In addition, it is necessary to have greater control over chirality so as to obtain greater metallic densities in each bundle.

• Low resistance metal contacts - Since metal contact resistance is a major determining factor for overall CNT interconnect resistance, further improvements are required to optimize on this parameter so as to achieve ideal quantum contact resistance.

• Defect-free CNTs - The series resistance of CNTs is significantly increased by defect sites occupied by adsorbed metal catalyst atoms. Defect-free growth of CNTs is a major roadblock to achieving quasi-ballistic transport with CNTs.

• Integration of SWCNTs - Most work on integrating CNTs with VLSI technology has been focused on MWCNTs. This is because the latter are more easily fabricated and are almost always metallic. However, they have poorer delay and reliability performance compared with SWCNTs. Sequential burning of individual nanotube layers at room temperature itself has been shown to occur at high current densities with MWCNTs. A greater focus on integration of SWCNTs is required.

V. SCALING OF CHIP-TO-CHIP INTERCONNECTS

It is widely recognized that chip-to-chip communication is even more severely constrained by existing metal interconnects than on-chip signaling. For over two decades, off-chip I/O rates have not been scaling sufficiently to keep up with on-chip clock frequencies. As opposed to on-chip interconnect performance, which is determined by the speed of signal propagation, the metric for chip-to-chip signaling is bit rate. For any communication channel, the maximum bit rate is limited by channel bandwidth if a required bit error rate (BER) is to be achieved in the face of inter-symbol interference (ISI). For electrical interconnects, channel bandwidth is determined by the $RC$ rise time of the metal line. Communication links such as those between memories and processors operate in the RF regime. They utilize highly specialized I/O circuitry to perform careful signaling at high frequencies. In addition, the transmitter, receiver and the channel are properly designed so as to prevent reflections and crosstalk.

To understand scaling issues with off-chip interconnects, we require a slightly different approach from the one discussed in Section IIA. In [36], it has been established that the scaling of $RC$ time constant of off-chip wires is severely constrained by the metal aspect ratio. This constraint arises directly since a particular aspect ratio achieves optimum packing density and bandwidth for a given scaling ratio, failing which there is an increase in line resistance and capacitance. Thus with scaling of wire dimensions and pitch, channel bandwidth essentially remains unchanged. The following relation is obtained as an estimate of the bandwidth of a metal interconnect of length $l$ and cross-sectional area $A$.

$$B = K \frac{A}{l^2}$$

For a copper line at room temperature, in order to obtain a reasonably open eye diagram (corresponding to a certain BER), $K$ is of the order of $10^{15}$ bits/s. The conventional approach to increase chip-to-chip signaling rates has been to increase the number of I/O pins and use time division multiplexing, thereby achieving a larger effective bandwidth with a relatively conservative increase in per-pin bandwidth. However, packaging constraints prevent a dramatic increase in number of I/O channels, thereby limiting the efficacy of this approach. In addition, at high signaling frequencies, wire skin effect and dielectric losses lead to worsened power dissipation.

Fig. 22 shows ITRS projections for required I/O bandwidth for the next decade. Current research on high-speed electrical links is focused on communication engineering techniques such as encoding, advanced channel equalization techniques,
multi-level signaling and channel response compensation [37]. These techniques would help approach closer to the Shannon limit of the system, but they come with additional area and power overheads. Ultimately, the performance of the link would be severely limited by the poor bandwidth of the interconnect. Fig. 23 shows the frequency responses of commonly used chip-to-chip lines, which are primarily differing in their lengths. As can be seen, high frequency signaling is not possible since it leads to virtually closed eyes at the receiver end.

State of the art high-speed chip-to-chip interconnects demonstrated have bandwidths of the order of 4-10 Gbps and average bit energy of 2-30 pJ/bit (mW/Gbps) [38]. Beyond 2020, channel bandwidth requirements would exceed 100 Gbps (Fig. 22). In addition, interconnect power dissipation will continue to be major problem. In 2002, at the 130 nm node, ITRS estimated that 50% of microprocessor power was consumed by interconnects and the 2009 report cites interconnect resistance and capacitance as a governing factor for power dissipation. These limits of electrical interconnects have already begun to show their toll on performance with the strong advent of multi-core processors. The number of cores on advanced information processing chips is expected to greatly increase in the upcoming years. A 512 core GPU at the 40 nm node with Nvidia’s CUDA architecture has already been reported [39]. In this light, effective on-chip processing rates would soon reach the Tbps regime leaving electrical interconnects as the most severe bottleneck, especially for high performance applications.

**VI. OPTICAL INTERCONNECTS**

**A. Performance advantages**

Optical interconnects are considered a radical technology because they fundamentally change the means of connection and rely on different principles of physics to achieve signaling. Most significantly, they completely subvert the problems of signal distortion due to limited bandwidth as well as losses associated with electrical currents on $RC$ transmission lines.

1) Optical interconnects use light in the near IR region to transmit energy over dielectric waveguides. This corresponds to very high frequencies of 100-1000 THz. Communication is achieved by modulating such a high frequency carrier with the required message signal, and the channel poses no bandwidth limitation.

2) The use of dielectric waveguides achieves strong confinement of electromagnetic waves, thereby eliminating the possibility of coupling with other signals. This is not the case with electrical interconnects since associated fields are completely spread out in the isolating dielectric medium.

3) Electrical signal transmission and detection physically requires charging of a load by a source, but there is in
addition an undesired charging up of the transmission line. On the other hand, optical systems rely on quantum mechanical processes at the source and detector. The channel, which is a dielectric waveguide can, at least in principle, be very low loss when compared with metallic transmission lines.

4) Very significantly, the use of a high frequency carrier in optical systems opens up the possibility of using wavelength division multiplexing (WDM) over the same channel to achieve orders of magnitude higher bandwidth. Though waveguide lateral dimensions are larger (typically μm) than the pitch of electrical lines, due to WDM, the aggregate bit rate per channel can be much larger than the modulation rate, and high interconnect density can be achieved virtually. This is different from electrical interconnects, where high bandwidths essentially require a large number of closely spaced physical channels.

5) Since the delay associated with optical interconnects would be much smaller, they can be used for transmitting critical signals such as timing pulses and clocks. Extensive research has been carried out over the last decade on determining the system-level performance of optical clock distribution schemes and designing architectures for the same [41], [42]. Besides a clear advantage in reduced skew and jitter, optical interconnects also considerably simplify overall system design in comparison with current microprocessors, which use complex circuit level approaches to determine and correct for clock errors.

Fig. 24 shows a projected comparison between Cu, CNTs (with two different mean free path lengths) and optical interconnects in terms of latency (related to bandwidth) and average bit energy to achieve the same BER. Optics significantly wins over electrical interconnects, especially at long transmission lengths.

B. Silicon photonics for optical interconnects

Extensive research has been carried out over the last two decades in the field of micro-scale silicon photonics for telecommunication applications. The major motivation behind work in this area has been to realize high-bandwidth (> 10 Gbps) optical links at a low cost. Traditional optical systems for long distance communications can perform well at these speeds but are very expensive to manufacture. Silicon photonics obtains a significant cost advantage by integrating optical components with electrical circuits on a single chip with the use of mature, large-scale CMOS fabrication techniques. Moreover, it relies on silicon as a material instead of commercial telecom systems which employ III-V lasers and photodetectors. It is a compelling solution to satisfy future communication bandwidth requirements, since it is the only known technology that can scale up to and over speeds of 100 Gbps. The performance of demonstrated silicon photonic links has already begun to rival that of III-V telecom grade photonics [43] with several groups reporting energy efficient high speed systems. Very recently, commercial products for 10G LAN with a CMOS photonics technology tagline have entered the markets and claim one-tenth reduction in per-port cost due to wafer-scale manufacturing [44], [45].

It is this high potential of silicon microphotonics for telecom that has opened up the possibility of inter- and intra-chip optical interconnect systems for VLSI technology. Extensive progress has been made over the past few years in realizing such systems with silicon photonic devices fabricated using standard CMOS processes. We now describe at a system-level the most commonly used optical interconnect solution developed for chip-to-chip communication as of today (Fig. 26).

1) An off-chip continuous wave (CW) laser, typically in the C-band provides the carrier wave for communication. The laser is coupled to an on-chip waveguide with an optical fiber, typically by means of a grating.

2) Modulation of the laser light is achieved by using tunable high-Q (order 10^4) optical microcavities placed very close to the waveguides. The resonance of these cavities is tuned very sharply by the electrical signal which needs to be transmitted. This helps achieve amplitude modulation of the high-frequency carrier signal. Microscale ring resonators are one kind of cavities com-
monly used to achieve this end, and are described in the next section. This system is often called a ring/racetrack resonator circuit.

3) Each on-chip transmitter uses a uniquely designed resonator and hence, a unique carrier wavelength for transmission (illustrated as $\lambda_1$, $\lambda_2$, ..., in Fig. 26). In this manner, multiple signals are multiplexed onto the same single-mode waveguide, thereby achieving WDM. Modulated light exits the waveguide on this chip, and can be coupled onto a similar system on another chip.

4) At the receiving end, resonant filters of identical frequencies filter out the desired part of the WDM signal. The filtered signal is then dropped into a photodetector, which converts absorbed light into an electrical signal.

C. On-chip waveguides

Dielectric waveguides are composed of adjacent dielectrics with different $k$ values. On-chip waveguides have most commonly been realized so far on SOI wafers. The SOI platform offers several advantages for photonic circuits, the most obvious one being the buried oxide (BOX) layer itself acting as a boundary dielectric for all light confining structures such as waveguides and resonators. The single crystal silicon layer also has low optical losses. Air can act as the second boundary dielectric, but an SiN/SiO$_2$ layer may also be used. Waveguide structures are fabricated by etching out of the top single crystal silicon layer. One commonly used geometry is the rib waveguide (Fig. 27). Typically, the lowest order TE mode of $\approx 100$ nm wide waveguides lies in the range of 1.4-1.6 $\mu$m and is used for propagation.

Propagation loss per length is the most significant figure of merit for on-chip waveguides. Nanometer scale roughness due to etching processes increases propagation losses [49], and therefore, optimizing the fabrication process is crucial to achieving low loss waveguides.

D. Electro-optic modulators

An electro-optic modulator is a device which relies on electro-optic effects to achieve modulation of an optical signal with an electrical input. As seen in the illustration in Fig. 26, the microring electro-optic modulator is clearly the workhorse device of optical interconnect systems. Candidates for electro-optic modulators in optical interconnect systems include quantum well modulators (QWMs) [36] and Mach-Zehnder interferometers (MZIs) which achieve phase modulation by adding phase-shifted optical signals [50]. Highly efficient low latency QWMs have been realized with III-V materials, but not yet with readily CMOS compatible materials such as SiGe [51]. Si ring resonators have advantages of compactness and energy-efficiency over MZIs and are therefore, widely considered the most promising solution as of today [52].

1) Ring resonator: The ring resonator is a strongly confining dielectric waveguide structure in the shape of a ring, and its operating principle is easy to understand at an intuitive level. The resonator is placed very close ($\approx 100$ nm) to the transmitting waveguide (Fig. 29). Over these distance
scales, evanescent fields from the waveguide couple with the resonator and vice versa. However, strong coupling is achieved only for the resonating wavelengths of the ring, which are integer fractions \((1/M)\) of the path length along the circumference \((nL, \text{where } n \text{ is the dielectric constant and } L \text{ is the circumference})\). The wavelength dependent transmission function of a ring resonator, which is approximately as written below [48], peaks at \((nL/M)\) and agrees with this intuition.

\[
T = \frac{T_{\text{max}}}{1 + \left( \frac{4nL}{\pi \Delta \lambda} \right)^2 \sin^2 \left( \frac{2\pi nL}{\lambda} \right)}
\]

\(\Delta \lambda\), the full-width at half-maximum (FWHM), decided by the exact geometry of the resonator, ends up determining its Q-factor. Further, it influences the extinction ratio, i.e. ratio of passband and stopband gains.

2) **Electro-optic effects in Si**: Electro-optic effects are those which involve a change in the optical properties of a material such as refractive index and absorption due to an electric field. Silicon, as a material, has very low coefficients for the most commonly observed electro-optic effects such as the Kerr and Franz-Keldysh effects. It has been proposed that the most effective mechanism to achieve fast changes in optical indices is the carrier-plasma dispersion effect [48]. This phenomenon involves changes in refractive index\((n)\) and absorption coefficient \(\alpha\) with changes in the carrier densities \((N, P)\) in silicon. The governing relations are linear and are of the form -

\[
\Delta n = A \Delta N + B \Delta P
\]

\[
\Delta \alpha = C \Delta N + D \Delta P
\]

At 1.55 \(\mu m\), carrier injection and depletion of order \(10^{18} \text{ cm}^{-3}\) can produce \(\Delta n\) of order \(10^{-3}\). As we will see, this is sufficient to achieve modulation.

3) **Modulation in Si ring resonators**: Modulation in Si ring resonators is achieved by varying the carrier density inside the ring. One effective way to achieve this is by using a circular p-i-n diode type structure (Fig. 30(a)). The electrical signal, which needs to be transmitted, is applied to the diode terminals in order to achieve a modulation in the depletion charge density in the i-region, and hence its refractive index. Since the i-region comprises the cavity, the resonant frequency is tuned by the electrical signal, and amplitude modulation (AM) is achieved. A \(\Delta n\) as low as \(10^{-3}\) would lead to a resonant wavelength shift by \(\Delta nL \approx 10 \text{ nm}\), since \(L \approx 10 \text{ \mu m}\). Considering the high Q’s of these resonators, even shift of 1 nm is sufficient to achieve modulation. At a system level, this process is very similar to achieving AM using a tunable RLC filter. Fig. 30(b) shows the variation in ring resonance with voltage applied to device terminals.

4) **Performance**: The performance of ring resonators is characterized by the dynamic response to an electrical signal with random pulse patterns. Fig. 31 shows the optical output corresponding to an NRZ modulated bit waveform applied at the input. In [54], it has been established that the modulation speed is determined not by the slow carrier dynamics at the device junctions, but by carrier injection and extraction lifetimes, which are of the order 100 ps, and photon lifetime in the ring (few ps). Carrier injection and extraction times may be significantly reduced by operating the diode in reverse bias condition, in which the these times are significantly lower. Using this technique, modulation speeds of up to 12.5 Gbps have been achieved with these systems. At a more basic level, modulation speed may be improved simply by reducing the resonator dimensions.

E. **On-chip lasers and SiGe photodetectors**

Successfully demonstrated optical interconnect systems so far have used off-chip laser sources to couple light onto waveguides using optical fibers and gratings. Commercially developed silicon photonic links use flip-chip bonded CW lasers for the same purpose. These solutions are simple (and cost-effective) but suffer from the major drawback of lossy coupling. Coupling losses are the most significant determinant of power efficiency in optical interconnect systems today. Energy-efficient coupling is necessary even more as optical networks grow denser. Therefore, the challenge of realizing on-chip lasers is of great importance. Several diverse approaches have been adopted to work towards this end, of which two are most notable.

- **Hybrid integration** of III-V lasers on silicon, in which prefabricated lasers are 'picked and placed’ on silicon substrates. This approach is not suitable for wafer scale integration [55], [56]
- **Making silicon lase** - Silicon being an indirect bandgap material is directly unsuitable for use as a gain medium. Extensive work has been carried out on achieving lasing action with silicon using other approaches such as Raman lasing and strained silicon waveguides [57], [58].

A few months ago, low-temperature CVD growth (at 400°C of InGaAs/GaAs nanopillars has been reported on silicon. Electroluminescence and confined cavity modes in hexagonal nanopillars have been demonstrated [59]. This research opens up an entirely new perspective of integrating III-V optoelectronics on silicon, which has traditionally been severely limited by growth problems due to lattice mismatch.

High performance photodetectors, especially with low output capacitance, are critical for low latency interconnect applications [38]. SiGe photodetectors have been extensively researched since Ge has a large absorption coefficient in the near IR region. They are compatible at a materials level with CMOS technology, since they rely on epitaxial growth of Ge. These photodetectors are realized using a metal semiconductor metal (MSM) architecture, since this integrates a waveguide.

F. **Integration challenges**

Most research on optical interconnect systems has so far been carried out on the SOI platform for reasons mentioned earlier. Such technologies naturally come at a high cost in comparison with bulk CMOS due to high substrate costs. Furthermore, optical systems on SOI itself require a large BOX thickness in order to prevent optical coupling with the underlying silicon substrate. This is against electronic device
Fig. 30. (a) Top-down and cross-sectional views of a p-i-n ring resonator structure showing the ring and doped regions. (b) Probe absorption spectrum for this device showing a voltage-induced resonant frequency shift at different DC voltages. Fig. from [51]

Fig. 31. Waveforms of driving electrical signals (a,c) (32- and 128-bit pseudorandom NRZ) and corresponding optical outputs (b,d) as reported in [53]

Fig. 32. (a) Schematic of nanopillar laser monolithically integrated on Si showing InGaAs core and GaAs shell. (b) SEM image shows well-faceted hexagonal structure which naturally arises during growth, and can itself act as a cavity (Fig. from [59]).

VII. Conclusion

In this paper, we have thus reviewed interconnect technologies for both the near and distant future from multiple perspectives. On-chip interconnect technology today is significantly worse off than the devices with major problems on multiple fronts including performance in terms of delay and coupling, power dissipation and reliability. Severe bottlenecks exist with no known solutions for problems into the next decade, especially in terms of low-k dielectric materials and worsened electromigration effects. While currently employed low-k/Cu interconnects can well be expected to scale into technology nodes for the near future with material and process improvements, this technology would sooner or later, reach its fundamental limits, thereby necessitating the need for more radical solutions. CNT interconnects can, in principle, offer a paradigm advantage over metal lines by virtue of ultra-low resistivity due to ballistic transport and excellent electromigration behaviour. However, achieving dense defect-free CNTs on a large scale, and subsequently integrating them with VLSI technology remains a major challenge with several breakthroughs required on the growth front. Optical inter-
connects, with sufficient research on CMOS integration, can well make their way onto replacing electrical interconnects for chip-to-chip communications, especially for high-performance applications such as servers and supercomputers. However, to enable their use as on-chip wires, there must be improvements in their density as well as energy-efficiency.

REFERENCES


