Scaling and Evaluation of Carbon Nanotube Interconnects for VLSI Applications

Fred Chen, Ajay Joshi, Vladimir Stojanović, Anantha Chandrakasan
Dept. of EECS, Massachusetts Institute of Technology
{fredchen, joshi, vlada, anantha}@mit.edu

ABSTRACT

The work in this paper addresses the need to evaluate the impact of emerging interconnect technologies, such as carbon nanotubes (CNTs), in the context of system applications. The critical properties of CNTs are described in terms of equivalent material parameters such that a general methodology of interconnect sizing can be used. This methodology is used to rescale the interlayer dielectric (ILD) stack-up and wire dimensions for different combinations of CNT and copper interconnects and vias; the stack-ups are then examined in an on-chip network application. The results of changing the ILD and wire sizing for a conservative estimate assuming a CNT bundle with 1/3 contacted metallic CNTs showed 30% improvement in delay and energy over copper at the 22 nm node and a 50% increase in total system throughput for a power constrained on-chip network application.

Categories and Subject Descriptors

B.7.1 [**Integrated Circuits**]: Types & Design Styles - *Advanced Tech.*

General terms

Scaling, evaluation, modeling

Keywords

Carbon nanotube, interconnect, VLSI, on-chip networks

1. INTRODUCTION

As CMOS processes scale into the nanometer regime, lithography limitations, electromigration, and the increasing resistivity and relative delay of copper interconnects has driven the need to find alternative interconnect solutions [1]. CNTs have emerged as a potential candidate to supplant copper interconnects because of their purported ballistic transport, and ability to carry large current densities in the absence of electromigration [2]. Previous studies that assess the potential use of CNTs as interconnects [3-6] primarily focus on the relative interconnect delay of CNTs to copper for forthcoming technology nodes.

In this paper, we investigate the potential performance impact that bundled single-walled carbon nanotube (SWCNT) interconnects could have on VLSI applications. Future references to CNTs can be assumed to mean only SWCNTs. In section 2, we revisit the problem of evaluating CNTs w.r.t. copper and provide a range of growth and assembly tolerances for CNT bundles to be resistively advantageous over copper. In section 3, we develop metrics to determine how CNT ILD stack-ups should be sized for generalized VLSI applications. We consider and evaluate every combination of copper and CNT bundles for vias and wires in regards to their energy and delay performance. Section 4 evaluates the wire stack-ups discussed in section 3 for buffered interconnects and on-chip multi-core communication.

2. CNT INTERCONNECT MODELS

Recently, there have been several studies attempting to model and evaluate the performance of CNTs against copper [3-6]. Because of different modeling assumptions in each study, there hasn't been a consensus on the relative performance gain/loss of CNT bundles compared to copper. Many of the discrepancies can be attributed to different accounting practices regarding the cross-sectional area of the CNT bundles. Other discrepancies include differing assumptions about CNT contact resistance, the number of CNTs contacted in the bundle, the bundle density and the percentage of metallic versus semi-conducting CNTs.

2.1 Isolated SWCNT Model

Unlike the model for CNT bundles, the circuit model used for an isolated single CNT [13] is generally accepted. The equivalent circuit model for an ideally contacted CNT isolated above a ground plane is shown in Fig. 1. The parameters for the circuit model, described in [12,13], are summarized in Table 1, where R_F is the resistance of the CNT, L is the length, L_0 is the mean free path, y is the distance between CNT and ground plane, and d is the CNT diameter. The remaining variables, h, e, and v_F, correspond to Planck's constant, an electron charge, and the Fermi velocity of a nanotube, respectively.

The kinetic inductance, L_K, hasn't been observed in high-frequency measurements [17] and can be excluded from on-chip interconnect models where wires operate at frequencies that are RC limited ($\omega L_K << R$) [6]. For the two capacitances, quantum capacitance (C_Q) and electrostatic capacitance (C_E), which are in series, C_Q is a non-factor as it is typically much larger than C_E [6].

TABLE 1. Model parameters for an isolated SWCNT

Parameter	Value
R_F	$h / 4e^2 \sim 6.5\ k\Omega,\ L \leq L_0$
R_F	$(h / 4e^2)\ (L / L_0),\ L > L_0$
L_K	$H / 2e^2 v_F \sim 16\ nH / \mu m$
C_Q	$2e^2 / h v_F \sim 100\ aF / \mu m$
C_E	$2\pi\varepsilon / \ln(y / d) \sim 34\ aF / \mu m$ $\varepsilon = 2.8\ \varepsilon 0,\ y = 97\ nm,\ d = 1\ nm$

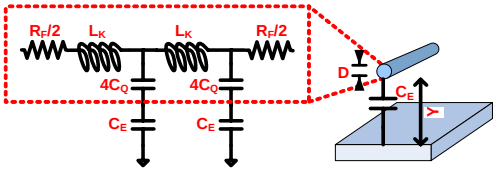

FIGURE 1. Equivalent circuit model for an ideally contacted, single-wall carbon nanotube isolated above a ground plane.

Nano-Net 2007 September 24-26, 2007, Catania, Italy.
 ISBN 978-963-9799-10-3
DOI 10.4108/ICST.NANONET2007.2041

2.2 Effective Resistivity of CNT Bundles

As described in Table 1, the fundamental resistance associated with a single CNT scales linearly with length for nanotubes longer than the mean free path. Meanwhile, the resistance associated with a bundle of CNTs is determined by the size and number of CNTs in the bundle. Taking into account contact resistance, and the effective fraction of contacted CNTs, the effective resistivity, ρ_{EFF}, can be described as

$$\rho_{EFF} = \frac{d^2}{k}\left(\frac{R_F + R_{cont}}{L}\right) = \begin{cases} \frac{d^2}{k}\left(\frac{h}{4e^2 L_0} + \frac{R_{cont}}{L}\right), L \geq L_0 \\ \frac{d^2}{kL}\left(\frac{h}{4e^2} + R_{cont}\right), L < L_0 \end{cases} \quad (1)$$

where $L_0 = dC_\lambda$ and C_λ is the mean free path-to-nanotube diameter proportionality constant described in [7], R_{cont} is the contact resistance, d is the nanotube diameter and k is the fraction of contacted metallic CNTs in the bundle; sparser bundles, the presence of semi-conducting CNTs, and uncontacted CNTs in the bundle would all be characterized by a smaller k and result in a higher effective resistivity.

Fig. 2 plots ρ_{EFF} of an ideally contacted CNT (R_{cont}=0) and CNTs of different lengths, each with 50 kΩ contact resistance, against the resistivity of copper over several process nodes. Even for a relatively poor contact resistance of 50 kΩ, the effective resistivity of a semi-global wire length (1000X) is nearly identical to the case of an ideally contacted CNT bundle which is ~10X better than copper at the 22 nm node. This indicates that a contact resistance to bundle length ratio less than 50 Ω/minimum pitch is insignificant. Short CNT interconnects (10X), however, can be dominated by poor contact resistance and show little, if any, advantage over copper.

While Fig. 2 indicates potentially significant improvements over copper, it shows the best cast scenario where CNT bundles are 100% metallic, fully contacted, and fully dense with a uniform 1 nm diameter per CNT. This level of perfection in CNT growth and assembly is yet to be achieved, so it is important to assess the tolerable range of imperfections (CNT diameter, bundle sparseness, uncontacted CNTs, semi-conducting CNTs) and their effect on the resistivity of CNT bundles.

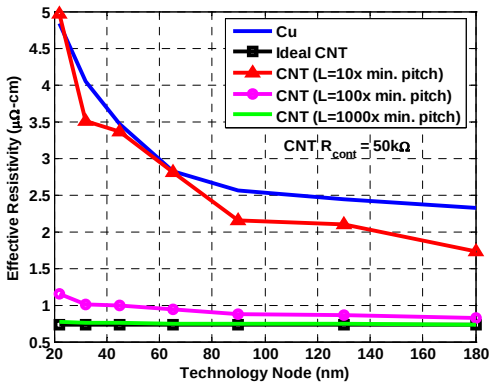

FIGURE 2. Effective resistivity of CNTs vs. copper. L_0, d, and k are 1 μm, 1 nm, and 1 respectively; for R_{cont}≠0, L is 10X, 100X, and 1000X the minimum wire pitch for the technology node. The resistivity of copper is modeled by [8]. Wire dimensions at each technology node come from [1].

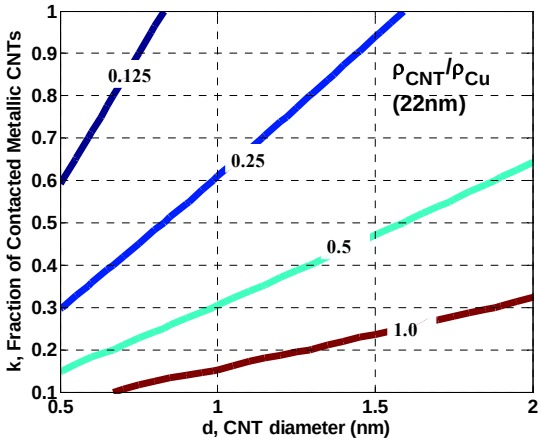

FIGURE 3. Relative resistivity of a CNT to Cu wire of length 1000X the min. wire pitch vs. CNT diameter and fraction of contacted CNTs at the 22 nm technology node.

Fig. 3 shows relative resistivity contours for CNT bundles as a function of individual CNT diameter and the fraction of contacted metallic CNTs within the bundle. The contours denote the resistivity of CNTs normalized by the resistivity of copper at the 22 nm node for a semi-global (1000X) length CNT bundle with 50 kΩ contact resistance per contacted tube. The variation of CNT diameter, d, takes into account the diameter dependence of the mean free path in CNTs as described in [7] in addition to the change in occupied cross-sectional area. The fraction of contacted metallic CNTs, k, is used to account for the presence of semi-conducting CNTs, uncontacted CNTs, as well as sparsely packed bundles.

Fig. 3 indicates that achieving even a 2X (ρ_{CNT}/ρ_{CU}=0.5) improvement in resistivity over copper will be challenging. As indicated in [9], CNTs have statistically shown to be metallic one-third of the time. If this statistic holds for CNT bundles, that would require a densely packed and fully contacted bundle with nanotube diameters less than 1.1 nm for the 22 nm technology node, which is on the low side of the distribution of diameters measured in [10]. In subsequent sections, we use CNT bundle cross-sections that are 33% metallic, but fully dense and contacted to demonstrate our proposed sizing methodology and metrics.

2.3 CNT Capacitance

There has also been some disparity in the reports of previous works in regards to modeling capacitance in bundled CNTs. As discussed in section 2.1, the electrostatic capacitance of a CNT largely determines its capacitance per unit length. The results from earlier works [4-6], which consider only electrostatic capacitance, have shown the capacitance of CNT bundles to vary roughly between 1X and 2X the capacitance per unit length of copper for an equivalent cross sectional area. In each case, nanotube-to-nanotube interaction within a bundle was ignored because all the tubes were assumed to be equipotential. The argument for the larger capacitance in CNT bundles is based on the additional surface roughness of CNT bundles [5]. However, recent work has disputed that result and shown the capacitance of densely packed CNT bundles to be within 3% that of a perfectly smooth copper wire occupying an equivalent cross-sectional area [6]. In the scenario where copper wires have exactly the same surface roughness as CNT bundles, the electrostatic capacitance

would be the same for the same cross-section. For subsequent sections, we compare CNT bundles to copper for cases where they have identical surface roughness. While this is an approximation, it is realistic since the surface roughness of copper wires has been shown to be a couple of nanometers [20] while the 'roughness' of CNT bundles, dependent on the diameter and uniformity of CNTs within the bundle, is also on the same order of magnitude.

3. SCALING INTERCONNECTS

In previous works, the cross-sectional dimensions used to compare the performance of CNT bundles to copper wires were either the permissible cross-section of copper at respective technology nodes [4-6], or else one or more layers of CNTs over varying dielectric thicknesses [6]. However, as described in Section 2, CNT bundles can be viewed as a material with resistivity, $\rho=\rho_{EFF}$. Since CNTs do not have identical material properties as copper, it is expected that the appropriate ILD stack-up and wire dimensions should be different than that of copper.

In this next section, we provide a methodology and metric to determine the ILD stack-up and wire sizing for an arbitrary interconnect material. Unlike previous works, we evaluate copper and bundled CNT interconnects based on energy-delay tradeoffs that are commonly used to determine system architectures, micro-architectures, and circuit design. For the subsequent plots shown, the ρ_{EFF} of CNTs, as described in Section 2, is calculated based on (1) assuming L_0, d, and k are 1 μm, 1 nm, and 0.33 respectively. The capacitance per unit length of CNT bundles is assumed to be identical to copper for the same cross-section. The results and methodology are easily extensible to other combinations of resistivity and dielectric constants which would correspond to different assumptions about CNT bundle dimensions and characteristics.

3.1 Exploration Space

Fig. 4 shows the design space for the cross-sectional dimensions that we consider as well as the combinations of material types used. For each technology node, we start with the cross section for copper as a baseline design point. We consider scaling the wire width (W) while maintaining the wire-to-wire pitch of the technology node. For copper interconnects, this is somewhat of a hypothetical exploration due to limitations in lithography. For CNT bundles, the assumption is that lithography or transistor dimensions still limit the wiring pitch, but CNT bundles could be grown at smaller dimensions than copper wires can be patterned.

We also consider scaling the ILD height (H) along with the wire thickness (T). Recent work has shown the ability to make CNT vias at higher aspect ratios than in current CMOS processes which could allow for greater separation between interconnect layers [15]. As H is scaled up, we inversely scale T by the same factor (S_H) to maintain a constant interconnect time constant (to first order). This allows us to optimize the contributions due to resistance and capacitance for a fixed bandwidth wire. It should be noted that in this scenario, other potential limitations to scaling such as increased via resistance and decreased thermal conductivity are not considered.

The above scaling exploration is done for every combination of copper and CNT interconnects and vias. Four design points of interest are compared: the baseline copper case (no scaling), copper wires with CNT vias (scale H only), CNT wires with copper vias (scale W only), and CNT wires with CNT vias (scale both H and W). For subsequent sections, W values less than 1 can be assumed to mean CNT wires while H values greater than 1 can assumed to mean CNT vias. Likewise all W and H values equal to 1 can be assumed to mean copper wires/vias.

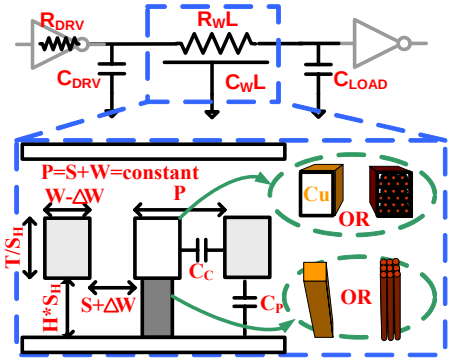

FIGURE 4. Equivalent circuit model and interconnect cross-section used for scaling exploration and to evaluate energy and delay characteristics.

3.2 Modeling Interconnect

Fig. 4 also shows a typical model used to evaluate interconnect and interconnect drivers as well as the wire cross-sectional dimensions. R_{DRV} and C_{DRV} are the equivalent drive resistance and gate capacitance for the driver. The parasitic output capacitance of the driver is estimated to be equal to the input gate capacitance while R_W and C_W are the interconnect resistance and capacitance per unit length. First-order expressions for the resistance and capacitance per unit length of a wire with rectangular cross-section are given by (2) and (3), while more precise expressions for plate (Cp) and fringing (Cc) capacitance per unit length are derived in [11] and given by (4) and (5).

$$R_W = \frac{\rho}{WT} \tag{2}$$

$$C_W = 2\left(C_P + C_C\right),\ C_P = \frac{\varepsilon W}{H},\ C_C = \frac{\varepsilon T}{S} \tag{3}$$

$$C_P = \varepsilon\left[\frac{W}{H} + 2.04\left(\frac{S}{S+0.54H}\right)^{1.77}\cdot\left(\frac{T}{T+4.53H}\right)^{0.07}\right] \tag{4}$$

$$C_C = \varepsilon\left[1.14\frac{T}{S}e^{-\frac{4S}{S+8.01H}} + 2.37\left(\frac{W}{W+0.31S}\right)^{0.28}\cdot\left(\frac{H}{H+8.96S}\right)^{0.76}e^{-\frac{2S}{S+6H}}\right] \tag{5}$$

For CNTs, it is assumed that L_0, d, and k, used to calculate ρ_{EFF}, are constant as the bundle cross-section scales. However, in the nanometer regime, the value of ρ for copper is not independent of cross-sectional geometry [8], and is given by (6).

$$\rho = \rho_0\left\{\frac{1}{3}\Big/\left[\frac{1}{3}-\frac{\alpha}{2}+\alpha^2-\alpha^3\ln\left(1+\frac{1}{\alpha}\right)\right] + \frac{3}{8}C(1-p)\left(\frac{1+H/W}{H/W}\frac{\lambda}{W}\right)\right\},\ \alpha = \frac{\lambda}{D}\frac{R}{1-R} \tag{6}$$

In (6), ρ_0, is the resistivity of bulk copper, λ is the mean free path, D is the average distance between grain boundaries, p the specularity parameter, R the reflectivity coefficient at grain boundaries, and C is a constant with a value of 1.2 for rectangular cross sections. The parameter values used for our simulations are listed in Table 2.

TABLE 2 Parameters for calculating copper resistivity

Parameter	ρ_0	λ	R	p	C	D
Value	2.04μΩ-cm	37.3nm	0.32	0.41	1.2	W

The total energy consumed (E_{TOT}) in driving an unbuffered wire of length L is proportional to the total capacitance while the delay (τ_D), as described in [14], is a function of driver resistance (R_{DRV}), driver capacitance (C_{DRV}), wire resistance (R_WL), wire capacitance (C_WL), and load capacitance (C_{LOAD}).[1]

$$E_{TOT} = 0.5 \cdot \left(C_{DRV} + C_{LOAD} + C_W L \right) \cdot V_{dd}^2 \tag{7}$$

$$\tau_D = R_{DRV}\left(C_{DRV} + C_{LOAD} \right) + 0.4 R_W C_W L^2 + \left(R_{DRV} C_W + R_W C_{LOAD} \right) L \tag{8}$$

3.3 Scaling Wire Width

In scaling the wire width, we maintain at least a minimum spacing; doing otherwise will only negatively impact both energy and delay. Inserting the first-order estimates for wire resistance (2) and capacitance (3), the equation for delay (8) can be broken into a fixed, geometry independent portion (τ_{FIX}) and a variable, geometry dependent portion (τ_{VAR}). In each case, the load capacitance has been rewritten in terms of the logical fanout (fo) where $C_{LOAD}=fo*C_{DRV}$.

$$\tau_{FIX} = R_{DRV}\left[C_{DRV}\left(1 + fo\right) + C_C L \right] + R_W \left(C_P + C_C/4 \right) L^2 \tag{9}$$

$$\tau_{VAR} = \left[\left(R_{DRV}/W_G \right)\left(2C_P + C_C/2 \right) L \right] \cdot W_R + \left[R_W \left(fo \cdot C_{DRV} \cdot W_G + C_C L \right) L \right] \cdot \frac{1}{W_R} \tag{10}$$

In (9) and (10), R_W, C_C, and C_P correspond to the per unit length resistance and capacitance values of a minimum width, minimally spaced wire for the technology node, while R_{DRV}, and C_{DRV} correspond to the on resistance and gate capacitance of a minimum sized inverter in the same process. W_R and W_G are scaling factors normalized to the minimum wire width, W_{MIN}, and minimum inverter width, respectively. An approximation has been made on the contribution of C_C, from $2C_CL/max(2\text{-}W_R,1)$ to $(C_C+½C_CW_R)L$. The max function only ranges between 1 and 2 for W_R values between 0.1 and 2, which are sufficient to cover the design space of interest.

From (10), we observe that for a fixed driver and load, there is an optimum normalized wire width that minimizes delay. This optimum normalized wire width can be calculated using (11).

$$W_{R,opt} = \frac{W_G}{W_{MIN}} \sqrt{\frac{R_W}{R_{DRV}} \frac{fo \cdot C_{DRV} + C_C L / W_G}{2C_P + 0.5C_C}} \tag{11}$$

For $W_{R,opt} < 1$, this results in both delay and energy savings over the nominal (minimum) width sizing. Equation (11) is primarily meant to provide intuition about sizing tradeoffs. To examine the actual energy vs. delay tradeoffs, we use (1), (4), (5), and (6) to properly account for effects of scaling.

Fig. 5 shows the resulting tradeoff curves for CNT interconnects at the 22 nm node for a wire length 250X the minimum wiring pitch, where W is specified in terms of minimum copper wire widths for the technology node. For logical fanouts where the optimal W for delay is less than one, the optimum delay also results in lower energy. This suggests that there are some performance and energy improvements that are unachievable due to lithography limitations. It should be pointed out that the results are plotted for a fixed load, so increasing fanouts really equate to reduced driver sizes. While Fig. 5 only shows the energy delay curves for a single length of wire, it is important to understand some of the key points of the plot for future reference.

First, the $W_{R,opt}$ point for each curve separates the regions where the delay term due to driver resistance and wire capacitance is dominant (scales as W_R) and where the delay term due to wire resistance and load capacitance is dominant (scales as $1/W_R$). For interconnect widths where the driver delay term is dominant, both energy and delay can be reduced by scaling down the wire width. When $W_{R,opt}$ is less than 1, it means a minimally sized wire is already in the regime dominated by driver resistance delay; delay can then only be improved by increasing the driver strength, and unless voltage scaling is employed, any attempts to improve delay will only increase energy consumption. When $W_{R,opt}$ is greater than 1, there is still an optimal delay point, but additional energy must be traded off in order to improve timing.

In Fig. 6, $W_{R,opt}$ is plotted for copper and CNT interconnects over varying wire lengths being driven at a fanout of 3 for the 22 nm and 45 nm technology nodes. The key point of this figure is that at wire lengths where $W_{R,opt}$ is less than 1, there is no flexibility in wire sizing to improve delay or energy. Delay can only be improved by increasing the driver and thus the energy as well. At the 22 nm node, the range of copper wire lengths where $W_{R,opt} < 1$ is limited to local interconnects: L < ~100 wire pitches.

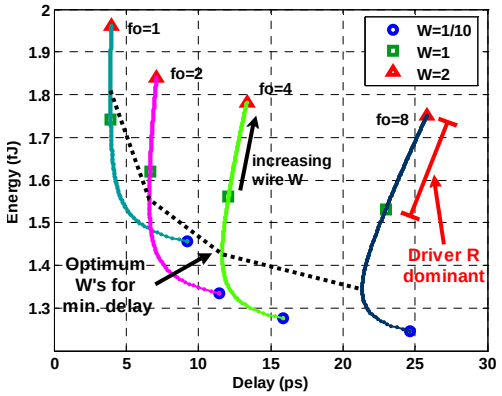

FIGURE 5. Energy vs. delay at varying fanouts (fo) for CNTs at the 22 nm node: 8X min. sized buffer load, L=250X min. wire pitch. W is specified in minimum copper wire widths.

[1] The 0.69 factor is omitted so results reflect the time constant rather than only 50% delay. To extract 50% delay, edge rate, and settling time values, the delay should be multiplied by 0.69, 2.2, and 7, respectively.

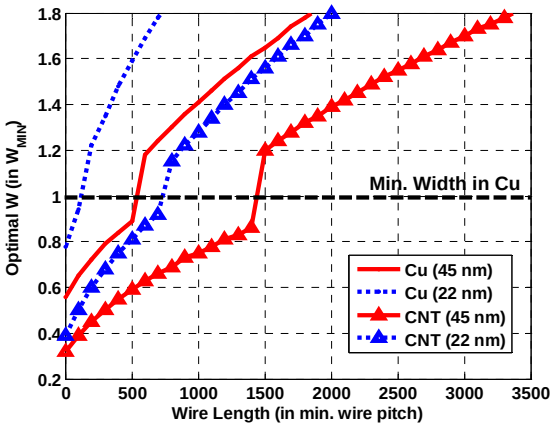

FIGURE 6. $W_{R,opt}$ vs. wire length for copper and CNTs for an 8X minimum gate load driven at a logical fanout of 3.

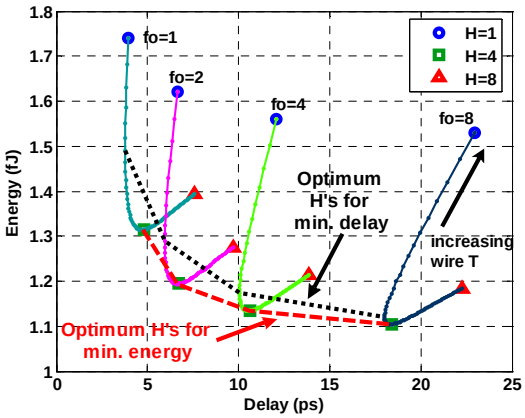

FIGURE 7. Energy vs. delay for CNTs at the 22 nm node for an 8X min. sized buffer load, L=250X min. wire pitch, H specified in nominal ILD height.

However, for CNT bundles, this range extends to 700 minimum wire pitches. Since a fanout of 3 is fairly typical for random logic [16], and 700 wire pitches includes semi-global wire lengths [18], Fig. 6 indicates that if CNT bundles are used as a drop-in replacement to copper at the 22 nm technology node, the majority of wiring will be sub-optimally driven.[2]

3.4 Scaling ILD Height and Wire Thickness

In exploring different dielectric thicknesses, we inversely scale wire thickness, T, as we increase the dielectric thickness, H to preserve a constant wire bandwidth during the exploration. While keeping T constant would increase the wire bandwidth, and clearly result in higher, more energy efficient performance, a more conservative approach is to evaluate the interconnects for a fixed bandwidth.[3] Thus, we let T scale as $1/H_R$ (so both C_c and C_p scale with $1/H_R$), where H_R is normalized to the nominal dielectric thickness, H_{NOM}, for the technology node. Similar to the width scaling case, we can substitute expressions (2) and (3) into (8) to obtain (12), which defines the optimum normalized height, $H_{R,opt}$, for which delay is minimized.

$$H_{R,opt} = \frac{1}{W_G H_{NOM}} \sqrt{\frac{2R_{DRV}\left(C_P + C_C\right)}{R_W \cdot fo \cdot C_{DRV}}} \qquad (12)$$

It is interesting to note, that the optimal dielectric height, and thus wire thickness, is independent of wire length. Again, we use (4), (5), and (6) for preciseness in our plotted results.

Fig. 7 shows the energy-delay tradeoff curves for varying ILD heights at several different fanouts for CNT interconnects at the 22 nm technology node. Increasing H improves both energy and delay until the minimum delay point is reached. In addition to the expected optimal delay point, there is also an optimal energy point. This optimal energy point is the result of increasing fringing capacitance as the plate and sidewall capacitance decrease as H and T are scaled [11]. What is perhaps more significant is that the minimum energy point is independent of fanout, wire length, and to a degree, technology; the H_R value that minimizes energy is very near 4 for all technology generations between the 90 nm and 22 nm nodes. From Fig. 7, it can be seen that in general, H (and T) should be increased (decreased) as much as possible until the minimum delay point is reached, typically around H_R=2. Beyond the minimum delay point there is a weak tradeoff between delay and energy until the minimum energy value is reached.

3.5 Performance of Different Stack-Ups

The previous two subsections explore scaling strategies for use in optimizing the cross-sectional dimensions for interconnects. A sampling of the resulting energy-delay tradeoff curves for the different configurations considered are plotted in Fig. 8. The curves in Fig. 8 are for a wire length 1000X the minimum wire pitch at the 22 nm node.

Several conclusions can be drawn from the plots in Fig. 8. First, for CNT bundles with the same cross-sectional dimensions as copper (CNT,W=1,H=1), there is no energy advantage, but there is an improvement in delay over copper (~25-30%) at low fanouts (< 2). Increasing dielectric thickness while inversely scaling copper wire thickness (Cu,W=1,H=2) provides greater savings in energy (~21%) for the same achievable delay than narrowing wire widths (CNT,W=0.5,H=1) and introducing CNT bundles (~15%). This suggests that if CNT vias, or some other via technology enabled the reverse scaling of projected ILD stack-ups, that the life of copper interconnects could be extended. Lastly, a combination of resizing wire widths, restacking the ILD, and replacing copper with CNT bundles provides the best raw energy saving results showing greater than 33% reduction in energy for the same delay. The results for the lowest delay wire (CNT, W=0.5, H=2) can also be interpreted as being able to run 2X faster than copper interconnects using the same ILD thickness (Cu, W=1,H=2) for the same energy expended. However, it is important to point out, that for shorter interconnects, the relative delay and energy advantages for each respective configuration diminishes.

---

[2] To set the minimum sizing of CNT bundles such that they have an equivalent fraction of wires that are 'sub-optimal' as copper, we can normalize the minimum width of CNT interconnects to that of copper. This results in a minimum wire width of roughly one-half (i.e. W=0.5) that of copper at the 22nm node.

[3] Scaling down the wire height is also beneficial from a manufacturing standpoint, both in high-aspect ratio Cu wires and CNT bundles.

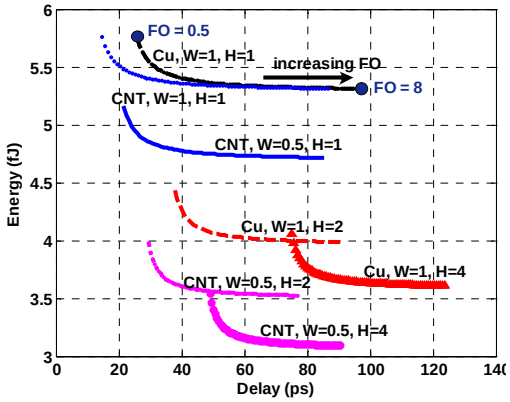

FIGURE 8. Energy vs. delay tradeoffs for various Cu and CNT stack-ups at the 22 nm node: 8X min. sized buffer load, L=1000X min. wire pitch. H and W specified in nominal ILD height and min. wire widths.

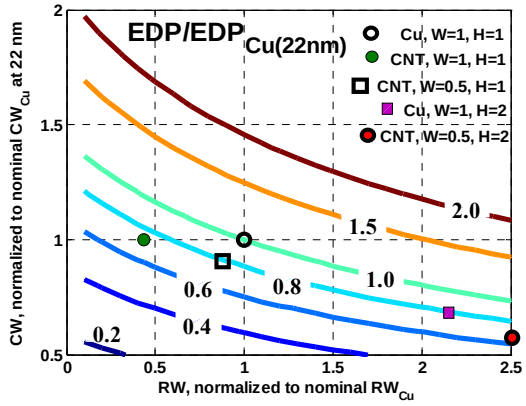

FIGURE 9. Normalized EDP contours vs. normalized resistance and capacitance per unit length at the 22 nm node: FO=1, 8X min. sized buffer load, L=1000X min. wire pitch.

A more general way to interpret each configuration shown in Fig. 8 is to view each as having changed the resistance (R_W) and capacitance (C_W) per unit length of the interconnect for a fixed set of material properties. For example, increasing the stack-up height while shrinking the wire height and using copper (Cu,W=1,H=2) effectively reduces C_W while increasing R_W compared to the baseline copper case. This is equivalent to maintaining the geometry of the baseline case but increasing the resistivity, ρ, of the conductor while decreasing the interlayer dielectric constant, k_{ILD}.

Fig. 9 extends this idea, and plots the normalized energy-delay product (EDP) contours for different R_W and C_W (also normalized to the baseline copper case) as well as where each stack-up resides on the contour map. The EDP captures the effects on both energy and delay due to changes in R_W and C_W, and indicates the relative change needed to achieve a certain improvement over the current copper based projections. These requirements can be used to determine both the ILD stack-up, and/or the material properties required. For example, to improve on the baseline copper case by 40% (EDP/EDP_{Cu}=0.6), the thickness of the copper wire could remain constant, while the ILD height and line-to-line spacing increase by 33%. Conversely, one could make no changes to the physical stack-up, but merely replace copper with a material that is 8X lower in resistivity.

4. ON-CHIP NETWORK IMPLICATIONS

To better understand the values that a new interconnect technology brings to the system, we need to bridge the gap between interconnect performance and relevant performance metrics at the system and circuit levels. Previous works [3-6] primarily addressed assessing which range of wires showed an advantage in delay over copper. In this next section, we give an example of how interconnect performance translates to system level performance using an on-chip network example.

4.1 Repeated Interconnects

Fig. 11 shows an example of a multi-core system with 16 cores that are connected by a mesh network. Each core has a router (R) that acts as an interface with the remaining cores, and the channel between routers is a bus of repeated interconnects. In Fig. 6, the results for delay optimal wire sizing of a single inverter driving wire were plotted for wire lengths up to 3500X the minimum wire pitch. However, longer wires are typically buffered such that their delay is linear with length rather than quadratic. The wire length between repeaters and repeater size that optimizes delay per unit length for a repeated interconnect is given by [19] and is modified to fit our delay expression and parameters in (13) and (14).

$$L_{opt} = \sqrt{\frac{2.5R_{DRV}\left(fo \cdot C_{DRV} + C_{DRV}\right)}{2R_W\left(C_P + C_C\right)}} \quad (13)$$

$$W_{G,opt} = \sqrt{\frac{R_{DRV}C_W}{R_W C_{DRV}}} \quad (14)$$

Fig. 10 uses (13) and (14) to plot the normalized EDP contours for a range of R_W and C_W for an optimally (delay) repeated wire of length 5000X the minimum wire pitch. The equivalent R_W and C_W, for each stack-up explored is plotted on the contours as well. While the general performance trend is similar to that shown in Fig. 9, Fig. 10 shows that for an optimally repeated interconnect, there is a greater sensitivity to change in wire resistance as wire capacitance increases. This is reflected by the improvement in relative EDP for CNT bundles occupying the same cross-sectional area as the baseline copper case (CNT, W=1, H=1) as compared to the unrepeated wire shown in Fig. 9.

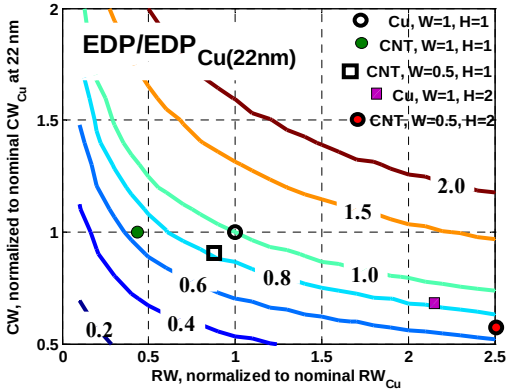

FIGURE 10. Normalized energy-delay product contours for an optimally repeated wire versus normalized resistance and capacitance per unit length at the 22 nm node: FO=1, L=5000X min. wire pitch.

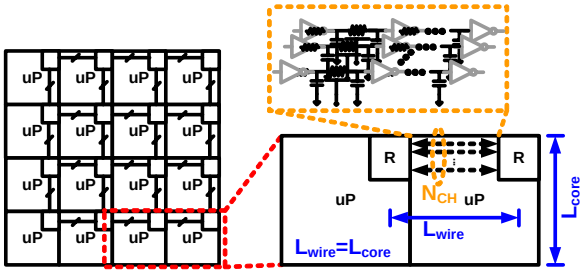

FIGURE 11. Example of inter-core communication in multi-core systems.

4.2 Core-to-Core Communication

As shown in Fig. 11, we assume that the interconnect length connecting two different routers (L_{wire}) is equal to the length of a core edge (L_{core}) and the maximum number of available routing channels (N_{CH}) between two adjacent cores is determined by the interconnect pitch (P) and core edge length.[4] The bandwidth per wire (BW_{wire}) is assumed to be the inverse of latency (τ_D) between adjacent routers and the energy per wire (E_{wire}) is calculated as described in section 3. Aggregate core-to-core bandwidth (BW_{agg}) can be calculated by multiplying the number of channels (in both directions) between cores and the bandwidth per wire. Similarly, the total energy (E_{total}) of the core-to-core link is calculated by multiplying the number of channels by the energy per wire. The relationships between these different parameters are described in (15), and (16).

$$E_{total} = N_{CH} E_{wire}, \ where \ N_{CH} = L_{core} / P \qquad (15)$$

$$BW_{agg} = N_{CH} BW_{wire}, \ where \ BW_{wire} = 1/\tau_D \qquad (16)$$

Fig. 12 shows a plot of BW_{agg} versus the total energy of the link for the different ILD stack ups explored. In each case, the wire widths are sized to maximize the bandwidth per unit width (wire width plus spacing). For each curve, each point also corresponds to a different core edge length, so the plots also show BW_{agg} versus core edge length. BW_{agg} saturates when the number of additional wires increases as fast as the bandwidth per wire decreases. As Fig. 12 shows, the BW_{agg} results fall in line with the delay performance of each stack-up shown in Fig. 8, with CNT bundles occupying the same cross-section as the nominal copper case (CNT,W=1,H=1) saturating at the highest aggregate bandwidth.

4.3 Total System Throughput

While the aggregate bandwidth gives a measure of the total capacity of all the links in the system, the true measure of a system is given by the maximum throughput the links can sustain without any bottlenecks. In the case of a mesh network, the maximum dataflow occurs at the bisection of the system, shown in Fig. 13. This is the path used by 50% of data generated by each core, when we assume uniform traffic (i.e. each core sends a

[4] The number of routing channels is assumed to be limited to a single routing layer and likewise it is assumed that the total width (length) of the core is available to the router. An increase in the number of routing layers would increase N_{CH}, while a decrease in router width would decrease N_{CH}.

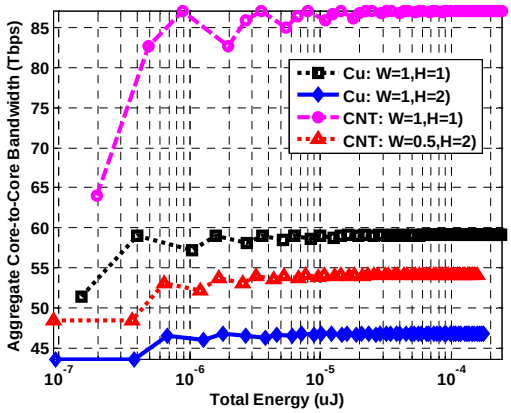

FIGURE 12. Aggregate core-to-core bandwidth vs. total energy for Cu and CNT interconnect ILD stack ups sized to maximize bandwidth per um for a 4X a minimum inverter repeater size.

message to every other core with equal probability). To avoid any bottlenecks, the total bandwidth available at this bisection must be able to support the total traffic generated by the cores. Hence, the bisection bandwidth determines the total maximum throughput, and is given by (17)

$$BW_{bisec} = BW_{agg} \sqrt{N} \qquad (17)$$

where BW_{agg} is the aggregate bandwidth between two cores and N is the total number of cores on the die. For a fixed system area, the number of cores in the system will determine the core edge dimensions and inter-core channel length, which in turn determines the aggregate core-to-core bandwidth achievable.

For our case study we modeled a multi-core system with a die area of 3.5 mm x 3.5 mm for the 22 nm technology. Fig. 14 shows a plot of the total throughput of the system varying with the number of cores for the same combinations of copper and CNT vias and interconnects shown in Fig. 12. Fig. 14 shows that for all cases, as the number of cores in the system increases, there is an increase in the total throughput of the system. This is a result of an increase in the corresponding bandwidth of each inter-core wire as the core-to-core distances shrink.

Although one would like to operate the system at the highest possible throughput, increasing throughput increases the total dynamic power dissipated by the system. As a result, the maximum throughput is limited by the power budget. Fig. 14 also shows the maximum system throughput for each wiring stack-up for power budgets of 20 W and 40 W. This power budget is assumed to be for the total dynamic power dissipated by the wires in the inter-core network.

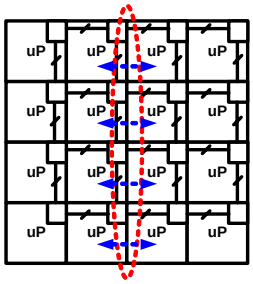

FIGURE 13. Channels determining the bisection bandwidth in a mesh network.

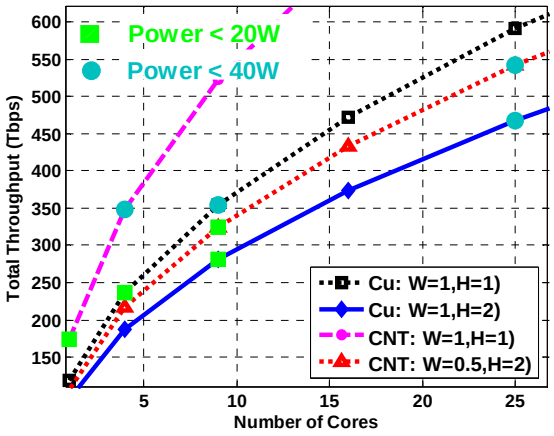

FIGURE 14. Total system throughput vs. number of cores for various Cu and CNT interconnect sizing and ILD stack ups.

Of the four different interconnect designs shown, the system that uses CNT technology for both interconnects and vias (CNT, W=0.5,H=2) is able to sustain the greatest maximum throughput for each power budget, showing a 50% improvement over the baseline copper case for a power budget of 40 W. What is also interesting to note is that the two interconnect systems using CNT vias (H=2) show the best power constrained total throughput. Even copper interconnects with CNT vias (Cu,W=1, H=2) outperformed CNT interconnects only (CNT, W=1, H=1), despite having a lower maximum aggregate core-to-core bandwidth (from Fig. 13). The subtle reason for this is that in Fig. 13, the same point in energy for two different curves corresponds to two different core sizes. So while the copper wire and CNT via case has a lower core-to-core bandwidth, it consumes sufficiently lower energy such that more cores can be used, thus improving the total throughput.

5. CONCLUSIONS

This work has demonstrated an approach to treating CNTs as an interconnect material and has quantified some growth (CNT diameter) and assembly (fraction of contacted metallic CNTs) tolerances to which CNT bundles must be manufactured in order to achieve lower resistivity than copper. Assuming that these manufacturing challenges are overcome, a generalized interconnect sizing strategy, based on energy-delay tradeoff curves, was proposed to evaluate different combinations of copper wires, copper vias, CNT wires and CNT vias. This approach also outlines the necessary general interconnect properties (R_W, C_W) required to be advantageous over copper. Results of these studies indicate the potential for both CNT via and CNT interconnect technology to reduce the energy and delay of general VLSI interconnects by roughly 30% for ~2X improvement in resistivity. Further progress in manufacturing CNT bundles will only improve upon these savings.

The different ILD stack-ups were then analyzed in the context of repeated interconnects and in a multi-core routing network. Results from this exercise showed a significant advantage in maximum aggregate core-to-core bandwidth when using CNT bundles than in the generalized case, as interconnect resistance plays a more significant role than capacitance in the fanout of 1 scenario. Under the assumption that one-third of the cross-sectional area occupied by a CNT bundle is metallic and can be contacted, the improvement in aggregate bandwidth when introducing CNT bundles was nearly 1.5X compared to the baseline copper implementation. However, for power constrained total system throughput, stack-ups using CNT vias showed higher performance, where CNT/copper based wires using CNT vias showed 1.4X/1.2X and 1.5X/1.3X improvement over baseline copper for 20 W and 40 W power budgets.

Although experimental data verifying some of the modeled parameters is necessary, the results of this study indicate that CNT interconnects have some promising benefits for long, low fanout interconnects while CNT vias show promise for lowering energy to increase the number of total interconnects.

6. ACKNOWLEDGEMENTS

The authors acknowledge the Interconnect Focus Center, one of five research centers funded under the Focus Center Research Program, a Semiconductor Research Corporation Program.

7. REFERENCES

[1] International tech. roadmap for semi., 2005, http://www.itrs.net.
[2] B.Q. Wei, R. Vajtai and P.M. Ajayan, *Applied Physics Letters* vol. 79, no. 8, pp. 1172-1174, 2001.
[3] A. Raychowdhury and K. Roy, *IEEE Conf. on Nano, 2004.* pp. 608-610, 2004.
[4] A. Naeemi, et al., *Elec. Dev. Lett., IEEE* vol. 26, no. 2, pp. 84-86, 2005.
[5] N. Srivastava and K. Banerjee, *ICCAD-2005,* pp. 383-390, 2005.
[6] A. Naeemi and J.D. Meindl, *Elec. Dev., IEEE Trans. on* vol. 54, no. 1, pp. 26-37, 2007.
[7] A. Nieuwoudt, et al., *Elec. Dev., IEEE Trans. on* vol. 53, no. 10, pp. 2460-2466, 2006.
[8] G.S. W. Steinhögl, et al., *Journal of Applied Physics* vol. 97, no. 023706, 2005.
[9] M.S. Dresselhaus, et al., *Topics in App. Physics. Carbon Nanotubes: Synthesis, Properties and Appl.*, New York: Springer-Verlag, 2000.
[10] J.G. I. Hinkov, et al., *Journal of Applied Physics* vol. 97, no. 4, 2004.
[11] S.-C. Wong, et al., *Semi. Manf, IEEE Trans. on* vol. 13, no. 1, pp. 108-111, 2000.
[12] P.L. McEuen, et al., "Single-walled carbon nanotube electronics," *Nanotech., IEEE Trans. on* vol. 1, no. 1, pp. 78-85, 2002.
[13] P.J. Burke, *Nanotechnology, IEEE Trans on* vol. 1, no. 3, pp. 129-144, 2002.
[14] T. Sakurai, *Elec. Dev., IEEE Trans. on* vol. 40, no. 1, pp. 118-124, 1993.
[15] M. Nihei, et al., *Interconnect Tech. Conf.*, pp. 234-236, 2005
[16] P. Zarkesh-Ha, J.A. Davis, W. Loh and J.D. Meindl, *Interconnect Technology Conference, 1998. Proceedings of the IEEE 1998 International*, pp. 184-186, 1998.
[17] Z. Yu and P.J. Burke, *Nano Letters*, Vol. 5, No. 7, pp. 1403-1406, 2005
[18] J.A. Davis, et al., *Electron Devices, IEEE Trans. on* vol. 45, no. 3 SN-0018-9383, pp. 590-597, 1998.
[19] K. Banerjee and A. Mehrotra, *Elec. Dev., IEEE Trans. on* vol. 49, no. 11, pp. 2001-2007, 2002.
[20] J. Pallinti, et al., *Interconnect Tech. Conf.*, pp. 83-85, 2003