IntSim: A CAD tool for Optimization of Multilevel Interconnect Networks

Deepak C. Sekar, Azad Naeemi, Reza Sarvari, Jeffrey A. Davis and James D. Meindl
Georgia Institute of Technology

Abstract **– Interconnect issues are becoming increasingly important for ULSI systems. IntSim, an interconnect CAD tool, has been developed to obtain pitches of different wiring levels and die size for circuit blocks or logic cores of microchips. It includes a methodology for *co-optimization* of signal, power and clock interconnects, and a newly derived stochastic wiring distribution that gives reduced error than prior work when compared to measured data. Results of IntSim are found to match well with actual data from an analyzed microprocessor. Several case studies are conducted to show this CAD tool's utility as a system level simulator: (i) Wire resistivity increases due to size effects are projected to increase die size of a 22nm low power logic core by 30% and power by 7%. (ii) When compared to a 22nm low power logic core with copper interconnects, a similar logic core with carbon nanotube interconnects could reduce power by 25% and die area by 27%, or increase frequency by 15% and reduce die area by 11%. (iii) A future 22nm 8 GHz 96M gate logic core's power, die size and optimal multilevel interconnect architecture are predicted. A version of IntSim with a graphical user interface is available for download from www.ece.gatech.edu/research/labs/gsigroup.**

I. Introduction

The performance, cost and power dissipation of a ULSI chip are increasingly being impacted by its interconnection networks. The statistics below provide some evidence of this phenomenon:

- Circuit block performance is known to show 47% sensitivity to transistor parameters and 53% sensitivity to interconnect parameters in 65nm chips [1].
- Interconnects formed over 50% of the dynamic power consumption of a 130nm microprocessor [2]. Also, interconnect repeaters have been shown to constitute as much as half of some commercial chips' leakage power [3][4].
- A 0.5µm technology needed just 4 interconnect levels [5], while 65nm technologies use as many as 10 levels of metal [6].

In this scenario, careful development of interconnect technology and good interconnect design become important.

	High performance 65nm technology	Low power 65nm technology
M1	210 nm	210 nm
M2	210 nm	210 nm
M3	220 nm	220 nm
M4	280 nm	280 nm
M5	330 nm	275 nm
M6	480 nm	280 nm
M7	720 nm	420 nm
M8	1080 nm	1080 nm

Table 1: Details of 65nm logic technologies

Shown in Table 1 are wire pitches of both high performance and low power 65nm technologies [7]. These pitches are normally selected using a stochastic wiring distribution that looks at previous generations of a chip and predicts wire lengths of a chip that needs to be designed with the logic technology [8][9]. Once the wire length distribution is known, algorithms are used to find pitches of different interconnect levels based on certain performance criteria and cost limitations. This selection of chip-specific wiring pitches is particularly important for high-volume microprocessors, and has been shown to provide several performance, power and cost advantages [8]. Since the die area of a design depends on both interconnect routing and gate sizing considerations, an extension of the above explained methodology can be used to predict die area of a circuit block or logic core.

Several publications have described algorithms to predict die area and wire pitches of an interconnect stack [8][9]. While these algorithms work well for older technologies, sub-90nm chips are significantly interconnect-limited and bring up several issues:

- Power distribution networks took up more than 25% of all wiring tracks in a 180nm microprocessor [10], and are expected to consume a bigger percentage of total wiring tracks with scaling [11]. Power distribution networks thus need to be modeled rigorously and have to be co-optimized along with signal/clock wiring and via blockage.
- Currently available stochastic wire length distributions show significant error when compared to actual data. For example, the commonly used Davis distribution [8] shows as much as 38% error with respect to measurement data for circuit blocks analyzed later in this paper. More accurate wire length estimates are needed.
- Via blockage can take up as much as 10-30% of the total wiring area for some metal levels [12]. Assignment of wires in multiple interconnect levels should be done with via blockage considerations in mind.
- Global interconnect pitch needs to be selected based on signal, power and clock wiring considerations [13].
- Repeater leakage power is substantial [3], and so needs to be considered when repeater insertion is performed.
- Wire resistivity increases due to size effects [14] need to be modeled.

This paper presents IntSim, a GUI based CAD tool that helps answer the above concerns and thereby enables better optimization of sub-90nm interconnect networks. After presenting a new stochastic wire length distribution model, this paper describes logic gate sizing in IntSim. Following this, global, local and intermediate/semi-global interconnect optimization are described. The algorithm used to combine together all these models is then presented. Results from IntSim are compared with data from a commercial microprocessor and several case studies are presented to show how IntSim can be used. A 22nm low power chip with carbon nanotube interconnects is benchmarked against a similar chip with copper interconnects only. A future high performance microprocessor core's power, die size and interconnect architecture are also predicted.

II. Derivation of stochastic wiring distribution

Several publications have discussed derivations of stochastic wiring distributions [15]. The Davis distribution [16], which is considered one of the most accurate [17], assumes gates are uniformly distributed all over the chip and then finds a distribution of wire lengths using Rent's rule. The derivation of a new wire length distribution that considers random arrangement of gates in a circuit block is discussed in this section.

For the purpose of this derivation, we define a new quantity called a gate socket. Any chip is considered to have many gate sockets, some of which are occupied by gates, as shown in Figure 1. The number of gate sockets $N_{sockets}$ is related to the number of gates N_{gates} by the relation:

$$N_{gates} = N_{sockets} \cdot p_{gates} \quad \ldots(1)$$

where p_{gates} is the percentage of die area that is occupied by logic gates. For example, a chip with 10 million gates and 50% of the die area occupied by logic gates [18] would have 20 million gate sockets, with gates randomly distributed in 10 million of them. If $N_{sockets}$ calculated with Equation (1) is not an integer, it is rounded off to the nearest integer as an approximation.

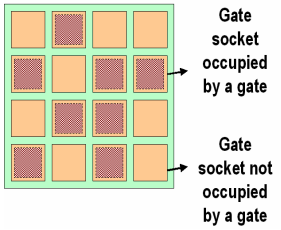

Figure 1: An illustration of the gate socket concept

The expected number of interconnects of a certain length l is given as the product of $M(l)$, the number of gate socket pairs separated by a distance l, and $I_{exp}(l)$, the average number of interconnects between a gate socket pair separated by l.

$$i(l) = M(l).I_{\exp}(l) \qquad \ldots(2)$$

The number of gate socket pairs separated by a distance l is similar to Davis' derivation of the number of gate pairs separated by a distance l [16]. Therefore,

$$M(l) = \begin{cases} \frac{l^3}{3} - 2l^2\sqrt{N_{sockets}} + 2lN_{sockets} & 1 \le l < \sqrt{N_{sockets}} \\ \frac{1}{3}\left(2\sqrt{N_{sockets}} - l\right)^3 & \sqrt{N_{sockets}} \le l < 2\sqrt{N_{sockets}} \end{cases} \qquad \ldots(3)$$

It should be noted that the value of l is in gate socket lengths. A gate socket length is defined as the distance between two adjacent gate sockets and is equal to (Die area/$N_{sockets}$)$^{0.5}$. Davis [16] defines gate pitch as (Die area/N_{gates})$^{0.5}$. A gate socket length is thus equal to (N_{gates} /$N_{sockets}$)$^{0.5}$ = $p_{gates}^{0.5}$ gate pitches.

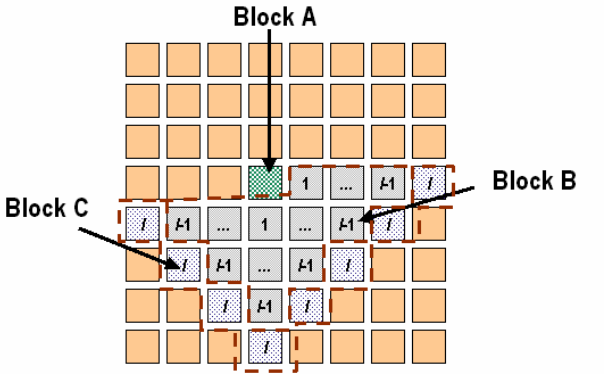

Figure 2: Block definitions for finding average number of wires between a gate socket pair

The average number of interconnects between a gate socket pair separated by l is given by:

$$I_{\exp}(l) = P(Gate\ in\ block\ A).\frac{I_{A-to-C}}{N_C} \qquad \ldots(4)$$

where *P(Gate in block A)* is the probability that block A of Figure 2 is occupied by a gate, $I_{A\text{-}to\text{-}C}$ is the average number of interconnects connecting block A to block C and N_C is the number of gates in block C.

$$P(Gate\ in\ block\ A) = \frac{N_{gates}}{N_{sockets}} = p_{gates} \qquad \ldots(5)$$

From the Davis derivation [16],

$$I_{A-to-C} = \alpha k\left[(N_A+N_B)^p - N_B^p + (N_B+N_C)^p - (N_A+N_B+N_C)^p\right] \ldots(6)$$

f.o. is the average fan-out of the system, $\alpha = f.o./(f.o.+1)$, k and p are Rent's constants and N_A, N_B are the number of gates in blocks A and B respectively. If gates are randomly distributed in gate sockets, the following approximations hold from an extension of [16].

$$\begin{aligned} N_A &= 1 \\ N_B &= p_{gates}l(l-1) \qquad \ldots(7) \\ N_C &= 2p_{gates}l \end{aligned}$$

Combining (2), (3), (4), (5), (6), (7) and normalizing, we get the average number of interconnects of length l gate socket lengths to be:

$$i(l) = \begin{cases} \frac{\alpha k}{2}\Gamma\left(\frac{l^3}{3} - 2l^2\sqrt{N_{sockets}} + 2lN_{sockets}\right)l^{2p-4} & 1 \le l < \sqrt{N_{sockets}} \\ \frac{\alpha k}{6}\Gamma\left(2\sqrt{N_{sockets}} - l\right)^3 l^{2p-4} & \sqrt{N_{sockets}} \le l < 2\sqrt{N_{sockets}} \end{cases}$$

where

$$\Gamma = \frac{2N_{gates}(1 - N_{gates}^{p-1})}{\left(-N_{sockets}^p \frac{1+2p-2^{2p-1}}{p(p-1)(2p-1)(2p-3)} - \frac{1}{6p} + \frac{2\sqrt{N_{sockets}}}{2p-1} - \frac{N_{sockets}}{p-1}\right)} \qquad \ldots(8)$$

The average wire length for this interconnect distribution is

$$L_{avg}(\text{in gate socket lengths}) = \frac{\int_1^{2\sqrt{N_{sockets}}} li(l)dl}{\int_1^{2\sqrt{N_{sockets}}} i(l)dl}$$

$$= \frac{\left[\frac{p-0.5}{p} - \sqrt{N_{sockets}} - \frac{p-0.5}{6\sqrt{N_{sockets}}(p+0.5)} + N_{sockets}^p\left(\frac{-p-1+4^{p-0.5}}{2(p+0.5)p(p-1)}\right)\right]}{\left(-N_{sockets}^p \frac{1+2p-2^{2p-1}}{p(p-1)(2p-1)(2p-3)} - \frac{1}{6p} + \frac{2\sqrt{N_{sockets}}}{2p-1} - \frac{N_{sockets}}{p-1}\right)}$$

For a large number of gates and p>0.5, this expression can be simplified to

$$L_{avg}(\text{in gate pitches}) = p_{gates}^{1-p}N_{gates}^{p-0.5}\left(\frac{p+1-4^{p-0.5}}{2(p-0.5)(p+0.5)p}\right) \qquad \ldots(9)$$

When gates were uniformly distributed over the die area, Davis derived the expression for average wire length to be:

$$L_{avg}(\text{in gate pitches}) = N_{gates}^{p-0.5}\left(\frac{p+1-4^{p-0.5}}{2(p-0.5)(p+0.5)p}\right)$$

It can thus be seen that average wire length with the new wiring distribution is approximately the Davis average wire length multiplied by a factor that depends on the Rent's constant p and the fraction of total die area occupied by logic gates. Most typical circuit blocks have 50-75% of the total die area occupied by logic gates [17]. Figure 3 shows a comparison of measured average lengths and average lengths predicted by the Donath distribution [18], the Davis distribution and the new distribution for 22 ISCAS'89 circuit blocks. Rent's constants and number of gates for these benchmark circuits are obtained from [19]. While the Donath distribution and Davis distribution have an average error of 75% and 38% with respect to actual data respectively, the new model has an error between 8% and 24% corresponding to values of p_{gates} ranging from 0.5 to 0.75.

Table 2 shows a comparison of average wire length obtained from measurements with values predicted by the Davis distribution

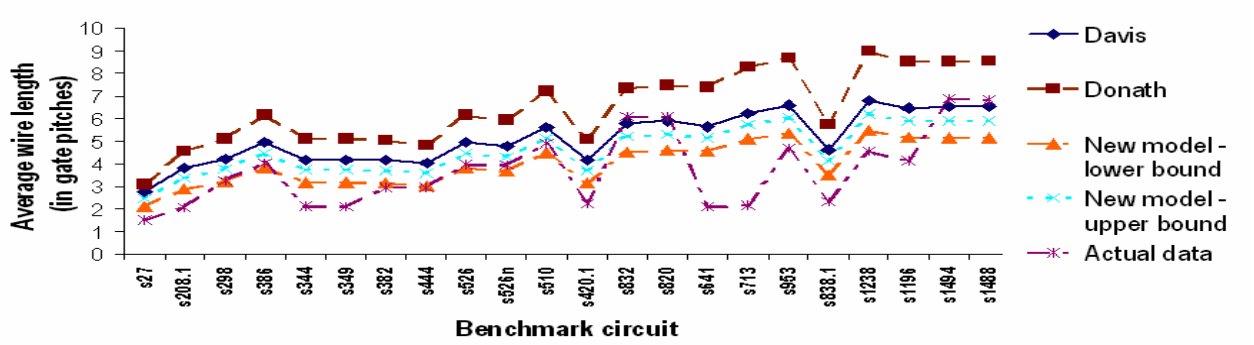

Figure 3: Validation of average wire lengths with new model with actual data from 22 ISCAS'89 circuit blocks. Average error of: Donath distribution = 75%, Davis distribution = 38%, New distribution = 8%-24%

and the new distribution, for benchmark circuits provided by Davis in [16]. It can be seen that while the Davis distribution has an average error of 26% for these circuits, the new distribution has average errors of only 2%-12 %.

Number of gates	Rent's constant p	Actual Data	Davis average length	New model with p_{gates}=0.5	New model with p_{gates}=0.75
2146	0.75	3.53	5.26	4.37	4.87
576	0.75	2.98	3.9	3.22	3.6
528	0.59	2.20	3.12	2.44	2.79
671	0.57	2.63	3.12	2.45	2.82
1239	0.47	2.14	2.96	2.26	2.64
73	0.667	2.00	2.35	1.89	2.14
78	0.667	2.27	2.38	1.91	2.17
72	0.667	1.88	2.34	1.88	2.13
252	0.667	2.73	2.96	2.39	2.71
236	0.667	2.198	2.93	2.36	2.67
237	0.667	2.887	2.93	2.36	2.67
55	0.667	1.579	2.23	1.79	2.03
59	0.667	1.38	2.25	1.81	2.06
62	0.667	2.08	2.28	1.83	2.08
Average error			26%	2%	12%

Table 2: Validation of model with actual data

Figure 4 shows how the new wiring distribution differs from the Davis distribution for a 36 sq. mm circuit block with 12 million gates, p_{gates} = 0.5, average fan-out = 3, and Rent's constants k=4 and p=0.55. Equations (8) and (9) suggest average length for the new distribution would be 27% less than the average length for the Davis distribution. While the log scale plot in Figure 4(a) indicates only a small difference for short lengths, the linear scale plot in Figure 4(b) shows a noticeable difference and captures the trend of the wiring distribution moving towards shorter lengths.

III. Logic Gate Modeling

Logic gates are modeled as 2 input NAND gates and are sized based on average wire length estimates provided by the new wiring distribution. If W is the device width, the delay of a logic path having 2 input NAND gates driving a fan-out *f.o.* is given by:

$$t_d = L_d 0.7 \frac{R_{NAND}}{W}\left(f.o.C_{NAND}W + f.o.\chi C_{\text{int}}\right) \quad \ldots(10)$$

where L_d is the logic depth, χ = 4/(*f.o.*+3) is a factor that converts point-to-point net length to wiring net length, R_{NAND} is the average drive resistance of a minimum size 2 input NAND gate, C_{NAND} is the input capacitance of the NAND gate and C_{int} is the capacitance of an average wire. C_{NAND} is computed assuming nMOS and pMOS devices are sized equally in a 2 input NAND gate, while R_{NAND} is obtained from equations given in [18]. If c is the capacitance per unit length of a wire, A is the die area and F is the feature size, since the area of a NAND gate of width W is given by $20.4(7.3+W)F^2$ [8], Equation (9) indicates

$$C_{\text{int}} = c.L_{avg} = c.p_{gates}^{1-p} N_{gates}^{p-0.5}\left(\frac{p+1-4^{p-0.5}}{2(p-0.5)(p+0.5)p}\right)\left(\frac{A}{N_{gates}}\right)^{0.5} \quad \ldots(11)$$

$$= c.\left(\frac{20.4(7.3+W)F^2}{A}\right)^{1-p}\left(\frac{p+1-4^{p-0.5}}{2(p-0.5)(p+0.5)p}\right)\sqrt{A}$$

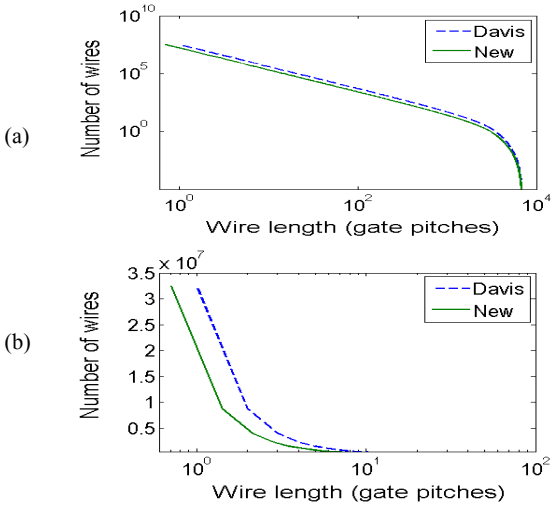

Figure 4: (a) Comparison of new distribution with Davis distribution in the log scale (b) Comparison of new distribution with Davis distribution in the linear scale for short lengths

It is interesting to note that unlike with previous wiring distributions, the length of an average length wire with the new distribution is a function of logic gate size. Essentially, it means if the die area is fixed and we use smaller size gates, they can be placed closer to each other, and so average wire lengths would reduce. If we define a constant

$$k_1 = c\sqrt{A}.\left(\frac{20.4F^2}{A}\right)^{1-p}\left(\frac{p+1-4^{p-0.5}}{2(p-0.5)(p+0.5)p}\right)$$

Equation (10) becomes:

$$t_d = L_d 0.7 \frac{R_{NAND}}{W}\left(f.o.C_{NAND}W + f.o.\chi k_1 (7.3+W)^{1-p}\right) \quad \ldots(12)$$

The delay expression in Equation (12) is equated to (1-margin)/*f* for finding gate size where *f* is the frequency and margin is the fraction of a clock cycle that constitutes skew and variability.

IV. Global Interconnect Modeling

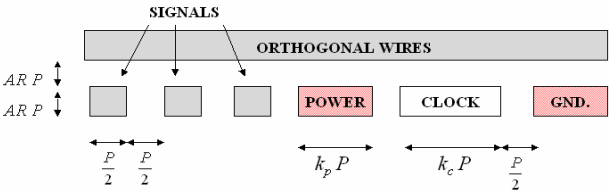

Figure 5: Structure of global wiring. P is global wire pitch

IntSim allows either two or four levels of "fat" wires for global interconnects. These levels have pitches that are based on signal, power and clock wiring considerations. Global wire pitch is selected as shown in [13]. To summarize the results of [13], global wire pitch is decided based on two conditions:

- Signal wire bandwidth should be maximized while meeting IR drop constraints for power wiring
- The wire pitch must be big enough to drive a tapered H tree of a certain length for clock distribution purposes

The equation for global wire pitch is [13]:

$$P = Max.\left[\begin{array}{l} 2.(k_p+0.5).N_{power\_pads}.\rho.\dfrac{I_T.d_{pad\_to\_pad}^{\,2}}{\pi.e_{router}.A.AR.k_p.V_{IR}}.\ln\left(\dfrac{0.65.d_{pad\_to\_pad}}{l_{pad}}\right), \\ \dfrac{D}{2}\sqrt{\dfrac{c_{clock}\rho}{AR\cdot k_c R_o C_o}}.\dfrac{1}{\dfrac{\beta_0}{fR_oC_o}-11}\left(\sqrt{72.6+\dfrac{4.4\beta_0}{fR_oC_o}}+11\right) \end{array}\right] \quad \ldots(13)$$

Here, N_{power_pads} is the number of power pads, ρ is the wire resistivity, I_T is current distributed per pad, $d_{pad\_to\_pad}$ is the distance between adjacent power pads, e_{router} is a routing efficiency factor, A is die area, V_{IR} is the user specified IR drop limit for global power wiring, l_{pad} is pad length, D is the distance between the driver and load of a H tree, R_o and C_o are the output resistance and input capacitance of a minimum size inverter, β_0 is the ratio of maximum rise time allowed for a clock tree to the clock period, f is the clock frequency and c_{clock} is the clock wire capacitance per unit length . Other parameters in Equation (13) are shown in Figure 5.

V. Local Interconnect Modeling

IntSim has two wire levels for routing local signal, power and clock wiring. Local interconnect pitch, P_{local}, is selected as $2F$, where F is the feature size. The length of the longest wire routed in local interconnect levels, l_{max}, is obtained from:

$$2e_w A = \chi P_{local}\sqrt{\frac{A}{N_{sockets}}}\int_1^{l_{max}} li(l)dl \quad \ldots(14)$$

Essentially, the left hand side of Equation (14) represents the area *available* for routing wires in the two local interconnect levels, and the right hand side of Equation (14) denotes the area *needed* for routing all wires having lengths between 1 and l_{max} gate socket lengths. e_w is a wiring efficiency factor given by:

$$e_w = 1 - e_{router} - e_{power/gnd} - e_{vias} \quad \ldots(15)$$

where e_{router} is the efficiency of the wire routing tool (typically around 0.5), $e_{power/gnd}$ is the fraction of area used by power and ground wires and e_{vias} is the fraction of area used by vias. $e_{power/gnd}$ is obtained from the model for local power distribution networks derived in [20]. Via blockage in local interconnect levels comes from vias to wires routed in higher metal levels and vias for repeaters. Based on the model for via blockage given in [12],

$$e_{vias} = \sqrt{\frac{(2N_{wires\_higher} + 2N_{repeaters}).(P_{local} + s\lambda)^2}{A}} \quad \ldots(16)$$

where N_{wires_higher} is the number of wires routed in higher metal levels, $N_{repeaters}$ is the number of repeaters for higher metal levels, λ is the design rule unit and s is a via covering factor which is typically 3 [12]. N_{wires_higher} is found from the stochastic wiring distribution by finding the number of wires whose length is greater than l_{max}. IntSim also runs electromigration checks on local power wiring based on maximum current density limits set by the user.

VI. Intermediate and Semi-global Interconnect Modeling

Intermediate and semi-global wires in IntSim are modeled based on Equation (17) and Equation (18). The right hand side of Equation (17) denotes the area required for routing wires of length lying between l_{min} and l_{max} in a pair of wire levels, and the left hand side denotes the area available for routing. Here, P is the pitch of the pair of wiring levels. Equation (18) represents the condition that the delay of the longest wire in a pair of metal levels should be a certain fraction of the clock period, as discussed in [8]. Equation (18a) represents this criterion when no repeaters are inserted while Equation (18b) represents the case when repeaters are inserted with an Energy-Delay Product minimization strategy [4]. Width of wires is equal to spacing between wires.

$$2e_w A = \chi P\sqrt{\frac{A}{N_{sockets}}}\int_{l_{min}}^{l_{max}} li(l)dl \quad \ldots(17)$$

$$\tau_{rc} = 4.4\frac{\rho(P, AR)}{ar.P^2}c\,l_{max}^{\,2}\sqrt{\frac{A}{N_{sockets}}} = \frac{\beta}{f} \quad \ldots(18a)$$

$$\tau_{rc} = \sqrt{\frac{A}{N_{sockets}}}\frac{2l_{max}\sqrt{\rho(P,ar)cR_oC_o}}{ar.P}\left(\frac{0.7}{\delta}+0.7\gamma+\frac{0.4}{\gamma}+0.7\delta\right) = \frac{\beta}{f} \quad \ldots(18b)$$

$$\gamma = (0.73 + 0.07\ln\phi_{gate})^2, \delta = (0.88 + 0.07\ln\phi_{gate})^2$$

where

$$\phi_{gate} = \frac{\frac{1}{2}aC_oV_{dd}^{\,2}f}{\frac{1}{2}aC_oV_{dd}^{\,2}f + bV_{dd}I_{leak}}$$

ρ = Wire resistivity (in Ω-m), C = Wire capacitance per unit length (in F/m), b = Percentage of time circuit is not sleep gated, f = Frequency (in Hz) R_o, C_o & I_{leak} =Resistance, capacitance & leakage of minimum sized repeater (in Ω, F & A), V_{dd} = Supply voltage (V), β=0.25(short wires) and 0.9(longer wires), a = Activity, ar = Wire aspect ratio

The wiring efficiency factor for intermediate and semi-global levels has three sources: (i) Repeater via blockage due to repeaters in higher metal levels (ii) Via blockage to signal wires routed in higher levels that is modeled based on [12] (iii) Power/ground via blockage that is got from equations in [20]. Wire resistivity increases due to size effects are modeled as shown in [21].

VII. Algorithm

In IntSim, the process of selecting wire pitches for different interconnect levels proceeds in several steps:

1. Input all parameters: The user inputs various details of the system that is being modeled.
2. Logic gate sizing: Logic gates are sized based on Equation (12) such that clock frequency targets are reached.
3. Generation of stochastic wiring distribution: Based on logic gate size chosen in Step 2, the fraction of die area occupied by logic gates, p_{gates}, is found. This is used to generate the stochastic wiring distribution given in Equation (8).

4. Set baseline parameters for iterations: The design of power interconnects and allocation of area for them depends on the chip power. However, chip power is not known until repeaters are designed in the multilevel wiring network, especially in sub-90nm chips where repeaters consume a significant fraction of total power. Also, design of the interconnect stack needs some knowledge of via blockage caused by repeaters. Thus, an iterative process is followed for assigning wires in a multilevel wiring network. An initial chip power estimate is set (as 100W, say) and the number of repeaters is set as 0.
5. Local interconnect modeling: Local wire pitch is set as 2F. Using Equations (14), (15) and (16), the longest wire routed in M1 and M2 is determined.
6. Arrangement of wires without repeaters: Once the longest wire routed in M1/M2 is determined, it is set as l_{min} in Equation (17). Equations (17) and (18a) are then used to find the pitch of M3/M4 and and maximum wire length routed in them. This in turn is set as l_{min} for the next pair of metal levels and this process continues till the longest interconnect of the wiring distribution is assigned a pitch.
7. Global interconnect modeling: A top-down process of global interconnect pitch selection and repeater insertion then begins. Global wire pitch is constrained to be the value found from Equation (13). The area needed for routing power wires is then found from equations given in [13], and this helps calculate the area available for signal wires in global wire levels. Clock wire area is neglected in IntSim because previous work has shown it is small [22]. Repeaters are inserted into these global signal wires, and the shortest signal wire routed in global wire levels is found based on a formula similar to Equation (17).
8. Assignment of wires with repeaters: Based on the length of shortest global signal wire, wires with repeaters are assigned to the pair of metal levels below the global wire levels based on Equations (17) and (18b). The pitch and shortest wire l_{min} are found for this pair of wiring levels and this l_{min} is set as l_{max} for the pair of wiring layers below it. Repeater insertion is performed for the pair of wiring layers below it and this keeps continuing till one runs out of die area for placing more repeaters or till the addition of repeaters does not improve wire delay.
9. Power computation and iteration: Once repeaters are assigned, the total chip power is calculated. Logic gate power is found using device widths calculated in Step 2 and formulae given in [18]. Local clock power is computed by extending models in [23]. Wire power is calculated based on the stochastic wiring distribution [8], and repeater power is calculated based on Step 8 and repeater power models given in [24]. Leakage power variability is modeled as discussed in [25]. If the total power calculated is different from the power estimate used for designing power distribution wiring, IntSim sets

$$Estimated\ power = \frac{Old\ estimated\ power + Calculated\ power}{2}$$

and goes back to Step 5. For the next iteration, the number of repeaters is set as the value calculated in Step 8.
10. Data output: When the simulation converges, the total number of wire levels, pitches of each wire level and a power estimate are output.

VIII. Comparison of Results from IntSim with Data from a Commercial Microprocessor

In this section, IntSim is used to predict wiring requirements of a commercially available microprocessor [26]. The predictions for number of interconnect levels, wire pitches and logic core power are compared with actual values of these quantities for that chip.

The analyzed microprocessor is a 65nm 3GHz high performance dual core chip [26]. Details of this chip's transistor parameters and number of gates in each core are obtained from published data in [26][27]. The dielectric constant for interconnects is 2.9 [26], contacted gate pitch is 220nm [27] and supply voltage is 1.325V [26]. Rent's constants k and p are chosen as 4 and 0.55 respectively based on guidelines in [18] that custom chips would have Rent's parameters around these values. Area of a logic core is obtained from die photos and published information about total die area [26]. Package technology parameters are obtained from data on older high performance chips [10] with the assumption that package technology does not scale. The values of wire pitch obtained are not very sensitive to package technology parameters, so these rough calculations are not expected to cause significant error.

Table 3 shows a comparison between wire pitches predicted by IntSim and actual wire pitches used for that technology [27]. *IntSim predicts the number of metal levels needed to be 8, which is exactly what is used for that interconnect technology.* The wire pitches predicted by IntSim are similar to the ones actually used, with a notable difference being that IntSim chooses wire pitches of two adjacent orthogonal metal levels to be the same, while the actual data has different wire pitches for adjacent orthogonal wiring levels.

	Actual data	Prediction from IntSim
M1	210 nm	220 nm
M2	210 nm	220 nm
M3	220 nm	296 nm
M4	280 nm	296 nm
M5	330 nm	296 nm
M6	480 nm	296 nm
M7	720 nm	1233 nm
M8	1080 nm	1233 nm

Table 3: Comparison of results from IntSim with actual data

IntSim also predicts the total power of logic cores of this chip to be 69.6W, while total chip power based on measured data is 80W [26]. Although published data is not available regarding the percentage of chip power taken up by caches and I/Os for this microprocessor, another 65nm processor had 19% of its total power consumed by these components and 81% of total power taken up by logic cores [28]. Assuming the processor analyzed with IntSim has similar numbers, *the logic core power for this processor is 65.6W, which is quite close to IntSim's prediction of 69.6W.*

IX. Case Study 1: Predictive Modeling of a Future 22nm High Performance Microprocessor

This section shows a case study conducted with IntSim on a future 8 GHz 0.8V 22nm logic core with 96M gates. The purpose of this study is to show how IntSim can be used to (1) Generate die size estimates, and (2) Project requirements of chips in future generations of technology. Device technology parameters are chosen to be ITRS low operating power technology parameters [14]. Interconnect dielectric constant is chosen to be 2.0 [14], wire aspect ratio is 2 and the number of power pads is chosen as 600. Rent's parameters k and p are 4 and 0.55 respectively. Two "fat" global wire levels are used for this design.

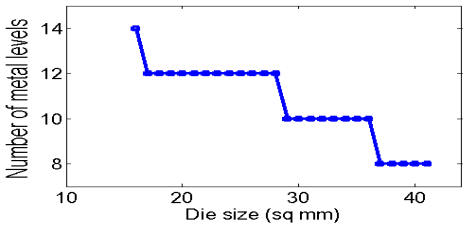

Figure 6: Die size estimation with IntSim

To find a die size estimate, simulations on IntSim are run with different die sizes and the required number of interconnect levels

are found, as shown in Figure 6. If the maximum number of interconnect levels available is set as 12 and die size needs to be as small as possible to lower production cost, one would choose the die size as 17 sq mm based on the data in Figure 6.

Metal levels	Wire pitches	Max. wire length	Percentage of total wire area available for signal wires	Repeater count
M1, M2	44 nm	4.5 um	20%	0
M3, M4	48 nm	77 um	42%	0
M5, M6	66 nm	683 um	45%	4.8 M
M7, M8	150 nm	1.9 mm	43%	0.7 M
M9, M10	298 nm	4.1 mm	39%	0.2 M
M11, M12	1799 nm	8. mm	25%	2400

Table 4: Interconnect requirements obtained from IntSim

The interconnect pitches needed for this technology are then estimated with IntSim by running a simulation for the selected die size. This is shown in Table 4. Also indicated in Table 4 is the percentage of total wiring area available for signal wires. Table 5 shows that while router inefficiencies take away 50% of the wiring, power distribution and via blockage also reduce available wiring area by a significant amount. In fact, *Table 5 indicates that power distribution and via blockage take up (10%+1%+3%)/36%=39% of the area taken by signal wiring*. The common practice of adding redundant vias would make via blockage estimates shown in Table 5 even higher. One of the interesting observations from IntSim is that while repeaters cause via blockage in lower metal levels, the insertion of repeaters enables wire pitches to be reduced for many levels [8], so this reduces via blockage for these levels based on the model in Equation (16) and in [12].

Metal levels	Signal wires	Power /gnd distri-bution	Repeater vias	Vias to wires in higher metal levels
M1, M2	20%	15%	3%	12%
M3, M4	42%	1%	2%	5%
M5, M6	45%	2%	0.5%	2%
M7, M8	43%	6%	0.2%	1%
M9, M10	39%	11%	0%	0%
M11, M12	25%	25%	0%	0%
Average	36%	10%	1%	3%

Table 5: Wiring area usage. Rest of the wiring is taken up by router inefficiencies

Power estimates obtained from IntSim are indicated in Figure 7. While computing clock power, it is assumed that 40% of the local clock power is saved due to clock gating. The percentage of clock power saved by clock gating is a user defined parameter in IntSim, and can be set by the user depending on his/her design.

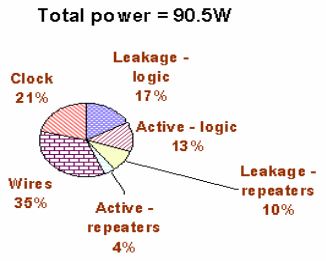

Figure 7: Power estimates of a future 22nm 8 GHz logic core

As can be seen from Figure 7, the high frequency of this logic core causes large power consumption. IntSim can be used to study sensitivity of power or performance to device, design or circuit parameters. Figure 8 shows the power savings possible from a 20% improvement in transistor drive current, interconnect dielectric constant and percentage of clock gating. When any of these parameters are changed, it is assumed that other input parameters for IntSim are the same. For example, when the drive current is changed, it is assumed leakage current is the same. Die size is optimized to be the minimum value possible with 12 metal levels. As can be seen from Figure 8, use of an interconnect dielectric with 20% lower permittivity provides more power savings than 20% better drive current or 20% more clock gating for this particular case study. This is largely because a lower wire dielectric constant leads to smaller size gates and latches and reduces power of these components, besides saving interconnect and repeater power. Having higher drive currents reduces logic gate, repeater and clock power, but does not impact interconnect power much.

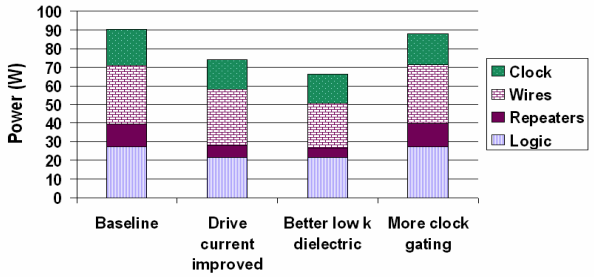

Figure 8: Power sensitivity to technology and design parameters

X. Case Study 2: Impact of Resistivity Increases in Copper for a Future 22nm Low-Power Chip

Copper resistivity is known to increase as interconnect dimensions are made smaller due to scattering at the grain boundaries and surfaces of a wire. While the resistivity of a 90nm wide wire is 2.5 μohm-cm, a 22nm wide wire could have a resistivity as high as 4.8 μohm-cm [14]. The impact of these size effects is studied with IntSim for a 22nm 1GHz low power chip. A previous study on this problem [29] indicated that wires in high performance chips would not be impacted much by size effects. This is largely because, for high performance chips, long signal wires are routed in higher metal levels whose pitches are big enough that they are not impacted by size effects much. For low power chips, however, pitches are normally smaller in size even in higher metal levels as indicated in Table 1, and so it is still not known how much they would be impacted by these resistivity increases. It is also not clear whether size effects would cause chip power to increase by a significant amount. One would expect that more area would be needed for power distribution networks when wire resistivity increases occur, and this has not been considered in [29].

Given a target clock frequency of 1 GHz for a 0.7V, 96M gate 22nm low power logic core with 10 metal levels, the optimal die size is found for cases where size effects are neglected and when they are considered. Device parameters are obtained from details of a low operating power ITRS technology [14]. The reflectivity parameter at grain boundaries for copper is chosen to be 0.25 and specularity parameter is 0.3 [14]. This logic core is assumed to be designed with an ASIC flow, so Rent's constants k and p are 4 and 0.65 respectively [18]. Wire aspect ratio is chosen as 2.

Results from IntSim indicate that a 1 GHz 22nm low power logic core with size effects would need 30% higher die area than a similar core where wire resistivity increases due to size effects are not present. Table 5 gives pitches of wiring levels needed for the two cases. It can be seen that when size effects are present, wires would need to be sized larger to maintain performance in spite of the higher wire resistivity. Since the number of metal levels used for both cases is the same, die size needs to be made larger to provide area for routing these bigger pitch wires. This is the main reason for the die size increase due to size effects.

	Pitch – no size effects	Max wire length – no size effects	Pitch - size effects	Max. wire length – size effects
M1, M2	44 nm	41 um	44 nm	83 um
M3, M4	46 nm	489 um	67 nm	546 um
M5, M6	52 nm	937 um	69 nm	1.1 mm
M7, M8	96 nm	4.6 mm	120 nm	4.7 mm
M9, M10	227 nm	12.2 mm	227 nm	13.9 mm
Die size	37 mm^2		48 mm^2	
Power	15.5 W		16.6 W	

Table 4: Impact of size effects on a 1 GHz 22nm low power core

Table 5 shows a power comparison of the case with size effects and the case where size effects are neglected. It is found that size effects cause a 7% increase in power. This is largely because the larger die size caused by size effects translates to longer wires and consequently bigger gates and latches. This can be observed from the higher gate and wire power numbers.

	Without considering size effects	With size effects
Logic	4 W	4.5 W
Repeaters	2 W	2 W
Clock	1 W	1.1 W
Wires	8.5 W	9 W
Total	15.5 W	16.6 W

Table 5: Power comparison with and without size effects

Table 6 shows the area occupied by the power distribution network in each pair of metal levels. The case without size effects has less area because of: (1) Lower power (2) Smaller die size for the same number of power pads, which in turn leads to smaller pad-to-pad distance and lesser area needed for global power wiring. This can be understood better with equations in [13] (3) Size effects not increasing resistance of power/ground wiring.

	Area without considering size effects	Area considering size effects
M1, M2	5.5 mm^2	7.2 mm^2
M3, M4	0.2 mm^2	0.4 mm^2
M5, M6	0.2 mm^2	0.4 mm^2
M7, M8	0.4 mm^2	0.7 mm^2
M9, M10	6.3 mm^2	8.2 mm^2

Table 6: Wiring area of power and ground distribution networks

XI. Case Study 3: A System Level Comparison of Carbon Nanotube Interconencts Against Copper

Carbon nanotube (CNT) interconnects are considered a promising long term alternative to copper interconnects [14]. This is due to their lower resistivity compared to copper, as shown in Figure 10 and their improved electromigration properties. Resistivity of CNTs are obtained from models in [30] that consider quantum resistance effects. As can be seen in Figure 10, a 20nm wide multi-walled CNT wire (MWCNT) has a resistivity ranging from 2.7 µohm-cm for a 70um long wire to 2.4 µohm-cm for a 1mm wire. A 20nm wide copper wire, on the other hand, has a resistivity of 4.6 µohm-cm. To put this in perspective, the transition from aluminum to copper metallization reduced resistivity by 40-45%, whereas a potential transition from copper to CNTs could reduce resistivity of a 20nm wide 1mm wire by 52%.

The authors wish to emphasize that several challenges remain to be overcome for CNT interconnects to be viable. These include obtaining good contacts, growth at CMOS compatible temperatures (<400°C), reliable manufacturing of horizontally oriented CNTs, among others [31]. However, many promising research efforts to tackle these issues have been reported in the recent past. For example, MWCNT vias have been grown at 500°C with resistance values close to that of tungsten plugs [32]. 100nm diameter MWCNTs with good contacts to all shells have also been demonstrated in [33]. *The purpose of this section is to find the chip performance/power benefit of CNTs if technology issues with growing MWCNTs for interconnect applications are tackled.* If it is found that CNTs could provide a large chip power or chip performance benefit, it would motivate more research in CNTs, and vice versa. To the best of the authors' knowledge, this represents the first *system level* comparison between CNTs and copper.

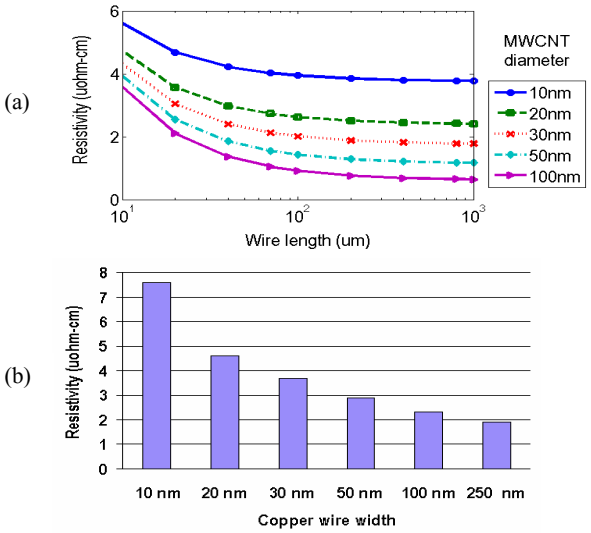

Figure 10: Resistivity for (a) Multi-walled CNT wires (b) Copper

The 22nm 1 GHz low power logic core studied in case study 2 is considered and MWCNTs are analyzed for use as signal wires. It is known that MWCNTs do not give advantages over copper when wire lengths are shorter than about 10µm due to their quantum resistance [34]. Table 4 indicates that M1 and M2 contain wires shorter than 10um. Thus, MWCNTs are assumed to be used only for M3, M4 and higher metal levels. Resistance and capacitance for these MWCNTs are obtained from models in [30][34]. The quantum resistance of MWCNTs reduces their applicability in power grids [34], and so copper is assumed to be used for power distribution purposes even when CNTs are used for signal wiring. It has to be kept in mind that techniques such as the use of monolayer single walled CNTs along with MWCNTs could potentially provide more benefits with CNTs than this case study suggests [31]. When wire widths needed are greater than about 100nm, multiple MWCNTs are assumed to be bundled together.

	Carbon nanotubes	Copper only
M1, M2 pitch	44 nm	44 nm
M3, M4 pitch	46 nm	67 nm
M5, M6 pitch	52 nm	69 nm
M7, M8 pitch	96 nm	120 nm
M9, M10 pitch	227 nm	227 nm
Die size	35 mm^2	48 mm^2

Table 7: Impact of MWCNTs on a 22nm low power logic core

The optimal die size is found with MWCNTs for 10 metal levels and a 1 GHz performance target. *The use of MWCNTs leads to a die size of 35 sq mm compared to 48 sq mm for the case with copper wiring only.* This is because of the lower resistance of MWCNT interconnects, which enables smaller wire pitches for the same performance as shown in Table 7, and so wire area (die area) needed for routing these wires is reduced. Power is also saved with the use of CNT interconnects. Figure 11 indicates that *chip power reduces from 16.6W to 12.4W*, which represents a reduction of 25%. This is because of three main reasons: (1) Reduced die size leads to shorter wires, fewer and smaller repeaters along with smaller gates

and latches (2) MWCNTs reduce wire capacitance. A 22nm diameter MWCNT would have a lower effective aspect ratio than a standard copper wire with an aspect ratio of 2. This leads to smaller gate sizes, latches and reduced repeater area (3) Less wire resistance implies reduced repeater area. All of these points reflect in the data shown in Figure 11.

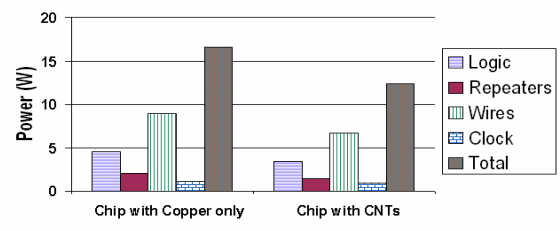

Figure 11: Power savings with CNTs

IntSim is also used to find the performance benefit possible with CNT interconnects for the same power budget as the chip with copper wires only. Die size is again optimized as shown in previous cases. It is found that a *15% frequency increase is possible with MWCNT interconnects for the same power dissipation. A die size reduction of 11% is also obtained.*

XII. CONCLUSIONS

This paper describes a CAD tool called IntSim that estimates die size and pitches of different wiring levels for sub-90nm logic cores. IntSim includes a newly derived stochastic wire length distribution and a methodology for co-optimization of signal, power and clock interconnects along with vias. The output of this tool shows a good match to actual data from an analyzed commercial microprocessor. Several case studies are conducted to show IntSim's utility as a system level simulator. Wire resistivity increases due to size effects are found to cause a 30% die size increase and 7% higher power consumption for a 22nm low power logic core. A logic core with carbon nanotube interconnects can have 25% less power and 26% lower die size, or 15% higher frequency and 11% less die area when compared to a similar 22nm low power logic core with copper interconnects only. A future 8 GHz 96M gate 22nm high performance logic core's die size, power and optimal interconnect network are predicted.

REFERENCES

[1] N.S. Nagaraj, W. R. Hunter, P.R. Chidambaram, et al., "Impact of interconnect technology scaling on SOC design methodologies", *Proc. Intl. Interconnect Technology Conference*, 2005

[2] N. Magen, A. Kolodny, U. Weiser, N. Shamir, "Interconnect power dissipation in a microprocessor", *Proc. Intl. Workshop on System Level Interconnect Prediction*, 2004

[3] R. Puri (IBM), "3D design and CAD needs", Proc. SRC Interconnect Forum, Sep. 2006

[4] D. Sekar, R. Venkatesan, K. Bowman, et al, "Optimal repeaters for sub-50nm interconnect networks", *Proc. Intl. Interconnect Technology Conference*, 2006

[5] G. Gerosa, S. Gary, C. Dietz, et al, "*A 2.2W 80MHz superscalar RISC microprocessor*", J. of Solid State Circuits, Dec. 1994

[6] W-H. Lee, A. Waite, H. Nii, et al,"High performance 65nm SOI technology with enhanced transistor strain and advanced low k BEOl,", *Proc. Intl. Electron Devices Meeting*, 2005

[7] C-H. Jan, P. Bai, J. Choi, et al, "A 65nm ultra low power logic platform technology using uni-axial strained silicon transistors", *Proc. Intl. Electron Devices Meeting*, 2005

[8] R. Venkatesan, J. Davis, et al., "Optimal n tier multilevel interconnect architectures for GSI", *Trans. VLSI Systems*, Dec 2001

[9] I. Young, K. Raol, "A comprehensive metric for evaluating interconnect performance", *Proc. Intl. Interconnect Technology Conference*, 2001

[10] J. Warnock, J. Keaty, J. Petrovick, et al, "The circuit and the physical design of the POWER4 microprocessor", *IBM J. of R&D*, Jan. 2002

[11] K. Shakeri, J. Meindl, "Compact physical IR drop models for chip package co-design of GSI", *Trans. Electron Devices*, Jun. 2005

[12] Q. Chen, J. Davis, P. Zarkesh-Ha, J. Meindl, "A compact physical via blockage model", *Trans. VLSI Systems*, Dec. 2000

[13] D. Sekar, E. Demaray, H. Zhang, P. Kohl, et al, "A new global interconnect paradigm: MIM power-ground plane capacitors", *Proc. Intl. Interconnect Technology Conference*, 2006

[14] International Technology Roadmap for Semiconductors

[15] M. Lanzerotti, G. Fiorenza, R. Rand, "Assessment of on-chip wire length distribution models", *Trans. VLSI Design*, Oct. 2004

[16] J. Davis, V. De, J. Meindl, "Apriori wiring estimations and optimal multilevel wiring networks for portable ULSI systems", *Proc. Electronic Components and Technology Conference*, 1996

[17] M. Lanzerotti, G. Fiorenza, R. Rand, "Interpretation of Rent's rule for ultralarge-scale integrated circuit designs, with an application to wirelength distribution models" *Trans. VLSI Design*, Dec. 2004.

[18] Models in BACPAC: www.eecs.umich.edu/~dennis/bacpac

[19] D. Stroobandt, "Apriori wire length estimates for digital design", Kluwer Academic Publishers

[20] R. Sarvari, A. Naeemi, P. Zarkesh-Ha, J. Meindl, "Design and optimization for nanoscale power distribution networks in gigascale systems", *Proc. Intl. Interconnect Technology Conference*, 2007

[21] W. Steinhogl, G. Schindler, et al, "Comprehensive Study of the Resistivity of Copper Wires With Lateral Dimensions of 100 nm and Smaller," *J. of Applied Physics*, 2005.

[22] J. Davis, J. Meindl, "Interconnect technology and design for gigascale integration", Kluwer Academic Publishers

[23] G. Chandra, P. Kapur, K. Saraswat, "Scaling trends for the on chip power dissipation", *Proc. Intl. Interconnect Technology Conference*, 2002

[24] K. Banerjee, A. Mehrotra, "A power optimal repeater insertion methodology for global interconnects", Trans. Electron Devices, Nov. 2002.

[25] S. Narendra, V. De, S. Borkar, D. Antoniadis, A. Chandrakasan, "Full-chip subthreshold leakage power prediction and reduction techniques for sub-0.18um CMOS," *J. of Solid State Circuits*, Mar. 2004.

[26] N. Sakran, M. Yuffe, M. Mehalel, J. Doweck, E. Knoll, A. Kovacs, "The implementation of the 65nm dual core Merom processor", *Proc. Intl. Solid State Circuits Conference*, 2007

[27] P. Bai, C. Auth, S. Balakrishnan, et al. "A 65nm logic technology featuring 35nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, Low k ILD and 0.57 um^2 SRAM cell", *Proc. Intl. Electron Devices Meeting*, 2004

[28] S. Naffziger, B. Stackhouse, T. Grutkowski, "The implementation of a 2 core multi-threaded Itanium family processor", *Proc. Intl. Solid State Circuits Conference*, 2005

[29] R. Sarvari, A. Naeemi, and James. D. Meindl, "Impact of size effects on the resistivity of copper wires and consequently the design and performance of metal interconnect networks", *Proceedings Intl. Interconnect Technology Conference*, 2005

[30] A. Naeemi, J. Meindl, "Compact physical models for multiwall carbon nanotube interconnects", Electron Device Letters, May 2006

[31] A. Naeemi, J. Meindl, "Design and performance modeling for single walled CNTs as local, semiglobal and global interconnects in gigascale integrated systems", *Trans. Electron Devices*, Jan. 2007

[32] S. Sato, M. Nihei, "Novel approach to fabricating carbon nanotube via interconnects using size-controlled catalyst nanoparticles", *Proc. Intl. Interconnect Technology Conf.*, 2006

[33] H. Li, W. Lu, J. Li, et al, "Multichannel ballistic transport in multiwall carbon nanotubes", *Phy. Review Letters*, Aug. 2005

[34] A. Naeemi, R. Sarvari, J. Meindl, "Performance modeling and optimization for single and multiwall carbon nanotube interconnects", *Proc. Design Automation Conference*, 2007