Investigation of Performance Metrics for Interconnect Stack Architectures

Puneet Gupta[†] , Andrew B. Kahng[†] , Youngmin Kim[‡] , Dennis Sylvester[‡]
[†] ECE Department, University of California at San Diego
[‡] EECS Department, University of Michigan at Ann Arbor
[†] {puneet, abk}@ucsd.edu
[‡] {kimyz, dennis}@eecs.umich.edu

ABSTRACT
This paper discusses metrics involving the bandwidth and energy characteristics of arbitrary interconnect stacks. Front-end dimensions are set by lithography and related fabrication restrictions and its performance is easily quantified using well-known metrics such as FO4 or ring oscillator delays, Ioff, and Ion. Back-end dimensions are not similarly constrained yet there are no comparable back-end metrics. In this study we seek figures-of-merit for interconnect architectures (stacks) that describe performance in terms of bandwidth and energy while considering issues such as via blockage and repeaters. A definition of bandwidth is presented and then appropriate via blockage models for interconnect stacks are investigated. In this paper, we improve existing bandwidth and throughput-driven design methodologies by looking at the entire stack rather than a single wiring layer. We also propose the use of bandwidth per unit energy. We evaluate and discuss these metrics in current 130nm and 90nm interconnect technologies.

Categories and Subject Descriptors
B.7.1 [Integrated Circuit]: Types and Design Styles – *advanced technologies,* VLSI.

General Terms
Measurement, performance, design.

Keywords
Bandwidth, Throughput, Energy, Back-end metrics, Interconnect stacks, via blockage.

1. INTRODUCTION
Consistent improvements in semiconductor technology enable shrinking feature sizes and reductions in gate delay. Meanwhile, interconnect scaling maintains density but increases latency relative to that of transistors [1]. In scaling from one technology to the next, the selection of front-end and back-end process parameters have greatly different constraints. In particular, front-end parameters like channel length and oxide thickness are set to their minimum possible values that maintain manufacturability. However, back-end dimensions like minimum pitch, metal thickness, and interlevel dielectric (ILD) thickness present a different set of tradeoffs in that smaller dimensions do not translate to better performance. As such, the tradeoff between density and performance (i.e., delay, noise, energy) should be carefully studied.

There are several well-known metrics that quantify the impact of front-end process parameters on circuit performance, including on or off current, and delays of structures such as ring oscillators or fanout-of-four inverters. For back-end dimensions, a bandwidth or throughput-driven design methodology rather than a conventional delay-driven approach has been recently proposed by a number of authors [2], [4], [5]. All these approaches are applied to one layer only, specifically the global interconnect layer which is critical to performance requirements. As a result such studies ignore factors caused by the presence of multilayer interconnect such as via blockage and repeater insertion in semi-global and local layers [3].

The main contribution of this paper is an analysis of delay, bandwidth, and energy metrics considering the entire multilayer interconnect stack. Each layer has its own wire dimensions that differ in width, thickness, and ILD thickness. Metrics must consider and comprehend that longer and fatter wires are needed for the global communication in the global layer while shorter and thinner wires are needed to connect a large number of gates in the local layers. In our work, the average wire length of each layer is estimated using a stochastic wire length distribution model [9] and a top-down wire assignment technique [11]. The wire delay is then calculated based on the estimated average wire length. Via blockage factor is considered in computing the allowable number of wire for each layer. Finally, new bandwidth and energy metrics for the interconnect stack are constructed.

The organization of the paper is as follows. The next section describes our delay and bandwidth calculation methodology. Section 3 extends the bandwidth calculation to energy aware metrics. Finally, we conclude in Section 4.

2. DELAY AND BANDWIDTH DEFINITIONS
2.1 Delay
In on-chip bus structures the impact of capacitive crosstalk on delay depends on the switching pattern of the aggressor lines.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SLIP'04, February 14–15, 2004, Paris, France.
Copyright 2004 ACM 1-58113-818-0/04/0002…$5.00.

Usually, switching factor is used to modify the coupling capacitance by the neighboring aggressors. Worst-case delay occurs when the two aggressor lines switch in the opposite direction to the victim. The worst-case 50% delay of a minimum-sized inverter driving an interconnect is [7].

$$t_{0.5} = 0.7R_{drv}(C_g + 4.4C_c + C_{drv}) + R_w(0.4C_g + 1.5C_c + 0.7C_{drv}) \quad (1)$$

In (1), R_{drv} and C_{drv} are the output resistance and input capacitance of a minimum size inverter. C_g and C_c are the ground capacitance and the coupling capacitance respectively in a uniform distributed line. R_w is the wire resistance of the victim line. To eliminate the well-known quadratic dependence of delay on line length, repeaters are inserted to reduce delay by shortening segment length. The total delay with optimal repeaters is then expressed as the multiplication of the number of segments and the one segment delay. When the number of repeaters is k and the size of each repeater is h, the total delay of the wire becomes:

$$t_{0.5} = k\left[0.7\frac{R_{drv}}{h}\left(\frac{C_g}{k} + 4.4\frac{C_c}{k} + hC_{drv}\right) + \frac{R_w}{k}\left(0.4\frac{C_g}{k} + 1.5\frac{C_c}{k} + 0.7hC_{drv}\right)\right] \quad (2)$$

Optimal k and h values can be obtained by solving the partial derivatives of the above equation [7].

2.2 Via Blockage

The via blockage phenomenon should be considered when estimating the available wiring resources in each layer more accurately for interconnect stacks. A simple model to account for the routing efficiency and via impact was proposed by [8]. In [8], power and ground routing is assumed to use 20% of metal resources of each layer. Also, it is estimated that a layer blocks 12% to 15% of the wiring capacity of every layer underneath it at constant pitch. Another via blockage model that is based on the physical parameters of a chip was proposed in [12]. In this model, turn vias (defined as vias internal to a net rather than used to directly connect between metallization and silicon layers) do not contribute to via blockage because they are an internal part of an interconnect. Only terminal vias are considered to construct the via blockage model. The authors of [12] claim that the model in [8] overestimates the blockage in upper layers, and that via blockage is only severe on the lower metal layers. It is assumed that all wires have two terminal vias in each underlying layer (underlying the layer in which the net is routed in). Therefore, the number of vias N_{v_wire} associated with each interconnect layer is given by [12]:

$$\begin{aligned} N_{v_wire} &= 2\left[I(L_{\max}) - I(L_n)\right] \quad when \quad n \neq 0 \\ N_{v_wire} &= 2I(L_{\max}) - I(L_n) \quad when \quad n = 0 \end{aligned} \quad (3)$$

Here n is the interconnect layer, L_{max} is the longest on-chip interconnect length, L_n is the longest interconnect length on the nth wiring level, and $I(l)$ is the cumulative interconnect density function [10] that gives the number of wires with length less than or equal to l.

To improve chip performance by reducing interconnect length, repeater insertion is a necessity, particularly, for the long wires in the semi-global and global interconnect layers. Vias due to these repeaters can have a major impact on the available resources in lower layers. Therefore, a via blockage model that considers the impact of repeaters is essential to accurately estimate routing resources. It is assumed that there are two sets of stacked terminal vias for every repeater. The number of vias N_{v_rep} by the repeaters in layer n is given;

$$N_{v_rep} = 2\sum_{i=n+1}^{top_layer} \#\ of\ \ repeaters\ \ in\ i^{th}\ layer \quad (4)$$

In (4), *# of repeaters* is the total number of repeaters for each layer which can be calculated by using optimum repeaters insertion method [13]. It is assumed that repeater area is not a constraint in this study.

Therefore, the total number of vias N_{v_total} is obtained by summing N_{v_wire} and N_{v_rep}. The final expression for via blockage factor in nth layer is

$$B_v(n) = \sqrt{N_{v_total}\,(2W + s\lambda)^2 / A_c} \quad (5)$$

N_{v_total} is the sum of N_{v_wire} and N_{v_rep}, W is the minimum interconnect width on a given layer, s is a minimum via spacing factor which is assumed to be three [21], λ is a layout rule that is typically set to half of the minimum feature size, and A_c is the chip size. Table 1 shows the base parameters used in this study (note that Vdd and Id values are taken from [15], [16]). Values of interconnect pitch of metal layers in various 130nm and 90nm technologies are given in Table 2 [15], [16], [17], [18], [19], [20]. The global interconnect pitch of ITRS is twice the minimum value from ITRS [6]. Width and spacing of interconnect are assumed to be equal. ILD thickness is assumed to be equal to the half-pitch of the layer above. Computed via blockage factors for these technologies are given in Figure 1 and Figure 2. As expected, the bottom layer shows 2~5 times larger blockage than the other layers. IBM and TSMC interconnect technology show more significant via blockage than the others. This is due to the less tapered nature (many iso-pitch layers) of the IBM and TSMC stacks compared to Intel.

Table 1: Base parameters used in this study

Parameters	130nm	90nm
$A_c(cm^2)$	0.98	0.98
# of gates	6.4M	12.87M
Vdd(V)	1.4	1.2
Id(mA/μm)	1.2	1.0
# of nets	29.08M	58.15M
k(ILD)	3.6	2.9
p(Rents exponent)	0.6	0.6

Table 2: 130nm and 90nm interconnect technology

Layer	130nm technology (nm)				90nm technology (nm)			
	Intel	IBM	TSMC	ITRS	Intel	IBM	TSMC	ITRS
1	350	320	340	350	220	245	240	210
2	448	400	410	350	320	280	280	210
3	448	400	410	450	320	280	280	275
4	756	400	410	450	400	280	280	275
5	1120	400	410	1340	480	280	280	275
6	1204	800	900	1340	720	560	840	820
7	-	-	-	-	1080	1120	840	820

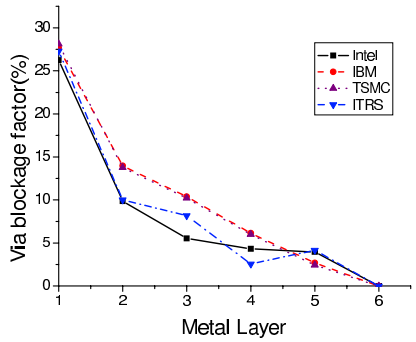

Figure 1: Via blockage in 130nm technology

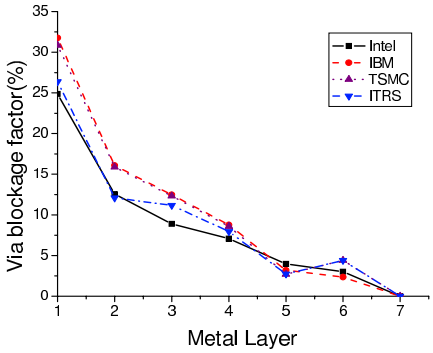

Figure 2: Via blockage in 90nm technology

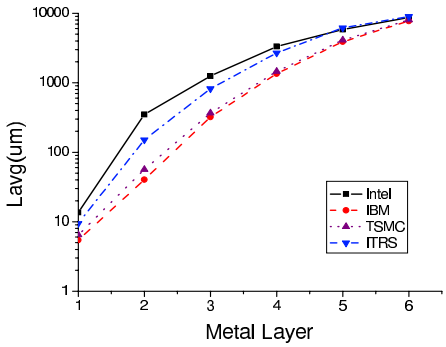

Figure 3: Average wire length in 130nm technology

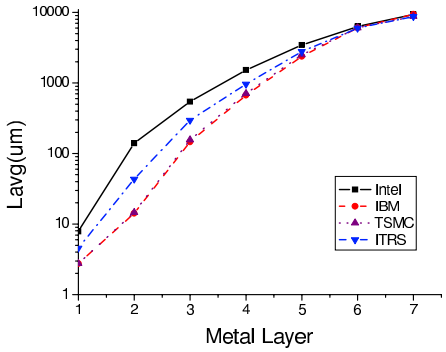

Figure 4: Average wire length in 90nm technology

2.3 WLD and Wire assignment

Since the typical wire lengths on each layer are considerably different, the typical delays of each layer also widely vary. A stochastic wire length distribution model based on Rent's Rule is proposed to predict the wiring distribution for the multilevel interconnect [10], [11]. Total interconnect length in a given layer can be estimated by using a supply-demand equation. The range of interconnect lengths on the *n*th layer is calculated by equating the area available for wiring A_{av} to the area that is required for wiring A_{req} [11].

$$A_{av} = e_w A_c = \chi P_n \sqrt{\frac{A_c}{N_g}} \int_{L_{n-1}}^{L_n} l \cdot i(l)\, dl = A_{req} \qquad (6)$$

Here e_w is the wiring efficiency factor, χ converts point-to-point interconnect length to wiring net length, P_n and L_n are the wire pitch and longest wire length on the *n*th layer. Wiring efficiency for global layers is set to 30% and to 60% for other layers. These numbers are chosen as we assume that power/ground/clock blockage is only significant at the global layers and consider only the router efficiency at the other layers. Wiring efficiency by via blockage was considered independently in Section 2.2.

In this study, the average wire length on each layer is used as a *typical* wirelength on the layer to calculate the wire delay. Average wire lengths for 130nm and 90nm technologies are given in Figure 3 and Figure 4 respectively. Intel and ITRS show much larger average wire length for all layers than IBM and TSMC. The discrepancy is due to the top-down wire assignment and the wider pitch at the top two layers in Intel and ITRS.

These less dense global layers cause more of the longer nets to be routed on intermediate layers and push the average wire length up across all layers.

2.4 Bandwidth Metrics

On-chip wiring is responsible for providing communication between two or more nodes in an integrated circuit with as little latency as possible [1]. Latency can be quantified by using bandwidth metrics. Bandwidth represents the rate at which information can be transferred through a channel; the greater the bandwidth, the more information that can be sent in a given amount of time. It is measured in bits per second and represents a key way of assessing overall system performance [4]. In unbuffered wires, consecutive signals must be staggered by at least three propagation delays before sending the next signal to avoid intersymbol interference. When using repeaters, one only needs to wait until the signal fully transitions on the first repeaters wire segment [2]. Therefore, the bandwidth for a given layer *n* is given by;

$$BW_n = \left(\frac{1}{Delay_n}\right) \times N_{wire} = \frac{1}{Delay_n} \times \left(\frac{chip\ side\ length}{pitch_n}\right) \qquad (7)$$

Where, $Delay_n$ is the wire delay calculated at the average wire length in a given layer *n*. It is calculated using Equation (2) assuming optimally inserted repeaters along the wire. N_{wire} is the number of parallel wires obtained by dividing the chip side length by the minimum wiring pitch for each layer. If the average wire length is longer than the maximum allowable distance between

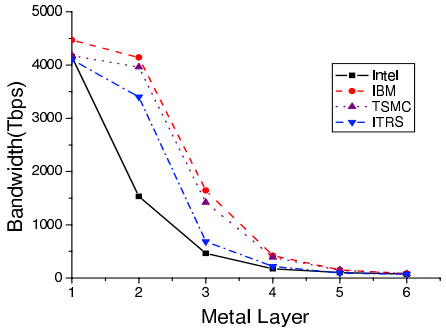

Figure 5: Bandwidth in 130nm technology

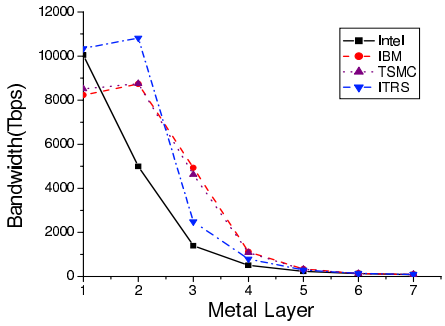

Figure 6: Bandwidth in 90nm technology

repeaters (L_{max}), optimal repeaters are added along the wire by dividing the wire into equal segments. If the average wire length is shorter than the maximum repeaters distance, then no repeaters are inserted but the driver is sized according to the ratio of the wire length to L_{max} (e.g., if $L_{avg} = L_{max}/2$ then the driver is set to be half the size of the optimal repeater size). Via blockage plays a role in reducing the wiring resources and we include its impact by multiplying bandwidth by the appropriate via blockage factor. The final bandwidth expression in layer n is given in (8):

$$BW_n = \left(\frac{1}{Delay_n}\right) \times \left(\frac{chip\ side\ length}{pitch_n}\right) \times B_v(n) \quad (8)$$

Bandwidth for 130nm and 90nm technologies is given in Figure 5 and Figure 6 respectively.

As can be seen in the figures, IBM and TSMC show better bandwidth than the others for all layers in 130nm. Although the RC per unit length is larger in these technologies, this is overcompensated by shorter average wire lengths and greater wiring density. ITRS and Intel show higher bandwidth at the bottom layer in 90nm.

2.4.1 Normalized Bandwidth

As can be seen in Figure 5 and Figure 6, the bandwidth in lower layers is much higher than in upper layers despite the larger via blockage in these layers. This is primarily due to shorter wires at lower layers and greater wiring density.

Due to the large imbalance between bandwidth of local and global layers, simply summing up bandwidth of individual layers is not a good way to assess the performance of an entire interconnect

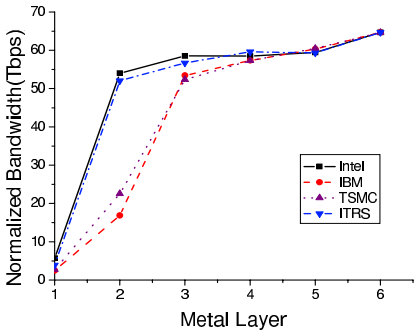

Figure 7: Normalized bandwidth in 130nm technology

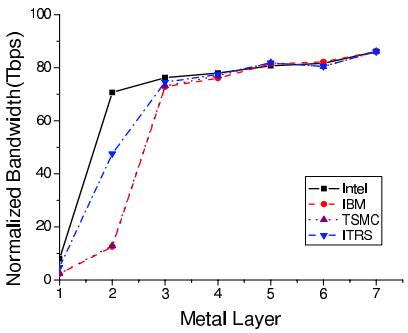

Figure 8: Normalized bandwidth in 90nm technology

stack. Instead, some normalization of bandwidth is required to compare arbitrary interconnect stacks, as opposed to just layers. As before, the average wire length in a layer is taken as a typical wire length in each layer. We assume that wires in all layers are transferring a signal from edge to edge, although local interconnect is clearly not intended to do exactly this. The number of segments N_{seg} in a layer is obtained by dividing chip side length by the average wire length.

$$N_{seg} = \frac{\sqrt{A_c}}{L_{avg}} \quad (9)$$

A_c is the chip size and L_{avg} is the average wire length in a given layer. N_{seg} is a measure of the "routing demand" for the given layer. Since various layers hugely differ in their routing requirements (e.g., local layers needs to have much more nets routed on them than global layers which typically route few long global nets), the bandwidth is normalized by this routing requirement. The normalized bandwidth is then given by

$$Normalized \quad BW_n = \frac{BW_n}{N_{seg}} \quad (10)$$

Figure 7 and Figure 8 show normalized bandwidth for each technology. And the sum and the average of bandwidth of individual layers are given in Figure 9 and Figure 10 respectively. Though IBM and TSMC result in higher bandwidth, Intel and ITRS are superior when considering normalized bandwidth metrics. Intel and ITRS have fewer segments due to their longer average wire length on all layers (note that their bandwidth has already been penalized due to this longer average wire length in Section 2.4).

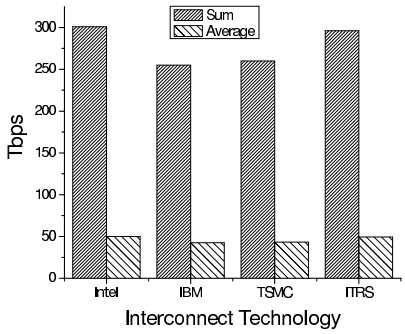

Figure 9: Sum and average of normalized BW in 130nm

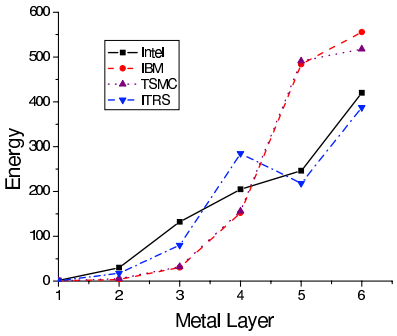

Figure 11: Energy in 130nm technology

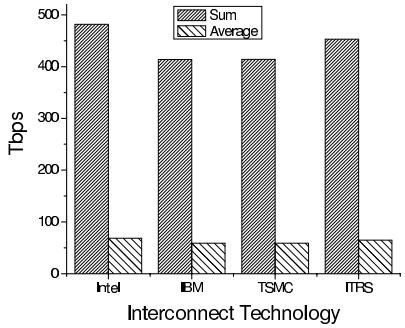

Figure 10: Sum and average of normalized BW in 90nm

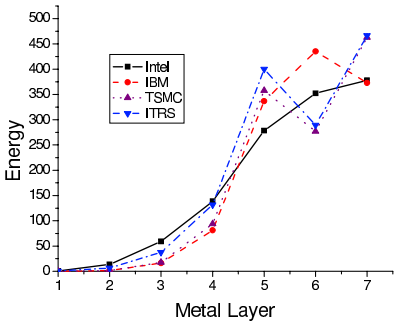

Figure 12: Energy in 90nm technology

3. ENERGY-DRIVEN METRICS

3.1 Energy Basics

The increasing use of repeaters on global and even intermediate layers reduces delay but also leads to higher power consumption. Power, silicon area, and via blockage are the three major concerns in using repeaters [14]. The via blockage problem caused by frequent repeater insertion was considered in Section 2.2 by simultaneously considering vias due to normal wiring distribution and vias due to repeaters. As mentioned earlier, the repeater area issue is not considered in our study. Energy is proportional to the capacitance contributed by repeaters and the capacitance from the wires. As before, an optimal number of repeaters in a long wire is k, the optimal size of the repeater is h, and the output capacitance of the minimum-sized inverter is C_{drv}. The total capacitance due to repeaters along a repeated wire is

$$C_{rep} = khC_{drv} \tag{11}$$

For a net on which repeaters are not needed (such as on the bottom layer), k becomes equal to one and the driver size is determined by the ratio of wire length to L_{max} and replaces h. Assuming a three-conductor and two ground plane bus structure, total wire capacitance is given by

$$C_{wire} = 2(C_g + C_c) \tag{12}$$

Ignoring operating frequency and supply voltage, which are considered independent of the stack geometry, total energy in a layer can be expressed by the number of wires in a layer and the capacitance for each net. The final expression used in the following energy metrics is

$$Energy = N_{wire}(C_{rep} + C_{wire}) \tag{13}$$

Where N_{wire} is the number of wires in a layer (note that N_{wire} is same as in equation (7)). The wire capacitance will be dominant in lower layers and repeater capacitance is significant in upper layers.

Figure 11 and Figure 12 show the estimated energy for each of the studied technologies. As can be seen in the graph, Intel and ITRS consume more energy in lower layers but the opposite situation holds for upper layers. This result stems from the longer average wire lengths and correspondingly larger drivers in lower layers of the Intel and ITRS technologies. IBM and TSMC have more number of wires in upper layers since the pitch is thinner.

3.2 Bandwidth per Energy

The bandwidth and energy calculations can be combined to provide a complete interconnect performance metric. Bandwidth per unit energy can be obtained by dividing bandwidth by energy for each layer. Figure 13 and Figure 14 show the bandwidth per energy for each interconnect technology. The sum and average are shown in Figure 15 and Figure 16.

As can be seen in graphs, although Intel and ITRS show larger number in normalized bandwidth metrics, IBM and TSMC become better in the bandwidth per energy metrics. This is explained by IBM and TSMC consuming less power in lower layers due to shorter wire length and smaller number of large drivers in the lower layers compare to Intel and ITRS.

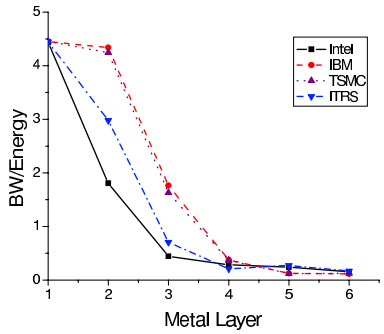

Figure 13: BW/Energy in 130nm technology

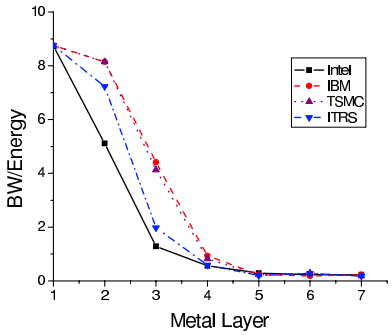

Figure 14: BW/Energy in 90nm technology

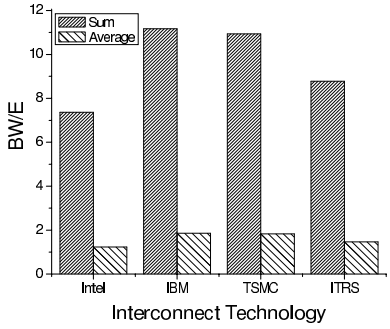

Figure 15: Sum and average of BW/E in 130nm

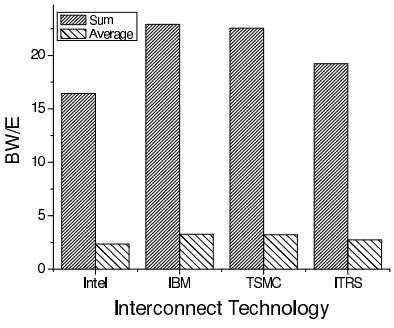

Figure 16: Sum and average of BW/E in 90nm

4. CONCLUSIONS

Bandwidth and energy metrics for complete interconnect stacks rather than single layer are identified. We consider the growing impact of repeaters on via blockage and develop a normalized bandwidth metric to enable straightforward comparisons of bandwidth across wiring layers of widely varying dimensions. Current 130nm and 90nm interconnect technologies are considered throughout the paper to benchmark the models used. In terms of normalized bandwidth, Intel and ITRS use wider pitches and tend to show better results than the other technologies studied. Although they have less wiring density, it is sufficiently overcome by smaller latencies. Energy metrics, however, indicate that Intel and ITRS interconnect technologies provide this higher bandwidth at the expense of higher power due to their need for more repeaters. Thus for heavily power constrained designs the IBM interconnect stack becomes the best choice.

As part of the future work in this area, we intend to explore the optimization of interconnect dimensions in future technology nodes by using these metrics with constraints on aspect ratio, dielectric constant, maximum crosstalk, yield, etc.

5. REFERENCES

[1] J. D. Meindl, J. A.Davis, P. Zarkesh-Ha, C. S. Patel, K. P. Martin, P.A. Kohl. Interconnect Opportunities for Gigascale Integration. *IBM J. Res. & Dev*, Vol. 46, No. 2/3, March/May 2002.

[2] Ron Ho, Kenneth W. Mai, Mark A. Horowitz. The Future of Wire. In *Proceeding of the IEEE*, Vol. 89, No. 4, pages 490-504, April 2001.

[3] P. Dasgupta, Andrew B. Kahng, and Swamy Muddu. A Novel Metric for Interconnect Architecture Performance. In *Proceeding of the DATE'03,* pages 448-453, March, 2003.

[4] Tao Lin, Lawrence T. Pileggi. Throughput-Driven IC Communication Fabric Synthesis. *ICCAD 2002, IEEE/ACM* pages 274-279, 2002.

[5] Ian Young, Kartik Raol. A Comprehensive Metric For Evaluating Interconnect Performance. In *Proceeding of the IEEE 2001 International*, pages 119-121, June 2001.

[6] International Technology Roadmap for Semiconductors 2002 Update. http://public.itrs.net/Files/2002Update/2002Update.htm.

[7] Dinesh Pamunuwa, Li-Rong Zheng, and Hannu Tenhunen. Maximizing Throughput Over Parallel Wire Structures in the Deep Submicrometer Regime. *IEEE Transactions on VLSI Systems*, Vol. 11, No. 2, pages 224-243, April 2003.

[8] Geroge A. Sai-Halasz. Performance Trends in High-End Processors. In *Proceeding of the IEEE*, Vol. 83, No. 1, pages 20-36, Jan 1995.

[9] J. A. Davis, V. K. De, and J. D. Meindl. A Stochastic Wire-Length Distribution for Gigascale Intergration(GSI)-Part I : Derivation and Validation. *IEEE Trans. On Electron Devices*, Vol. 45, No. 3, pages 580-589, Mar. 1998.

[10] J A. Davis, V. K. De, and J. D. Meindl. A Stochastic Wire Length Distribution for Gigascale Integration (GSI) –Part II : Applications to Clock Frequency, Power Dissipation, and Chip Size Estimation. *IEEE Trans. On Electron Devices*, Vol. 45, No 3, pages 590-597, Mar. 1998.

[11] R. Venkatesan, J. A. Davis, K. A. Bowman, J. D. Meindl. Optimal n-tier Multilevel Interconnect Architectures for Gigascale Integration(GSI). *IEEE Trans. VLSI systems*, Vol. 9, No. 6, pages 899-912, Dec. 2001.

[12] Qiang Chen, Jeffrey A. Davis, Payman Zarkesh-Ha, and James D. Meindl. A Compact Physical Via Blockage Model.

IEEE Trans. On VLSI Systems, Vol. 8, No. 6, pages 689-692, Dec. 2000

[13] H. B. Bakoglu. *Circuits, interconnects and packaging for VLSI.* Reading, MA, Addison-Wesley, 1990.

[14] A. Naeemi, R. Venkatesan, J. D. Meindl. Optimal Global Interconnects for GSI. *IEEE Trans. Electron Devices*, Vol. 50, No. 4, pages 980-987, April 2003.

[15] S. Tyagi,; M. Alavi, R. Bigwood, et al., A 130nm Generation Logic Technology Featuring 70nm Transistors, Dual Vt Transistors and 6 layers of Cu Interconnects. *Electron Devices Meeting, 2000. IEDM'00,* pages 567-570, Dec. 2000.

[16] S. Thompson, N. Anand, M. Armstrong, et al., A 90nm Logic Technology Featuring 50nm Strained Silicon Channel Transistors, 7 layers of Cu Interconnects, Low k ILD, and 1μm^2 SRAM Cell. *Electron Devices Meeting, 2002. IEDM '02,* pages 61-64, Dec. 2002.

[17] T. Schiml, S. Biesemans, G. Brase, et al., A 0.13µm CMOS Platform with Cu/Low-k Interconnects for System On Chip Applications. *VLSI Technology, 2001. Digest of Technical Papers. 2001 Symposium on*, pages 101-102, June 2001.

[18] M. Khare, S. H. Ku, R. A. Donaton, et al., A High Performance 90nm SOI Technology with 0.992μm^2 6T-SRAM Cell. *Electron Devices Meeting, 2002. IEDM '02,* pages 407-140, Dec. 2002.

[19] K. K. Young, S. Y. Wu, C. C. Wu, et al., A 0.13µm CMOS Technology with 193nm Lithography and Cu/Low-k for High Performance Applications. *Electron Devices Meeting, 2000. IEDM '00*, pages 563-566, Dec. 2000.

[20] C. C. Wu, Y. K. Leung, C. S. Chang, et al., A 90-nm CMOS Device Technology with High-speed, General-purpose, and Low-leakage Transistors for System on Chip Applications. *Electron Devices Meeting, 2002. IEDM '02,* pages 65-68, Dec. 2002.

[21] MOSIS Scalable CMOS (SCMOS) Design Rules, Revision 8.0, http://www.mosis.org/Technical/Designrules/scmos/scmos-main.html