Quantitative Studies of Impact of 3D IC Design on Repeater Usage

Jason Cong, Chunyue Liu, Guojie Luo
Computer Science Department, UCLA
{cong, liucy, gluo}@cs.ucla.edu

Abstract: *In this paper, we present our quantitative studies of the impact of 3D IC design on repeater usage. The repeater usage is estimated by the interconnect optimizer IPEM in the post-placement/ pre-routing stage, where the 2D and 3D placement are generated by state-of-art mixed-size placers mPL6 and mPL-3D. Experiments on a set of real industrial designs show that, through 3D placement, the total number of repeaters used in the on-chip interconnections can be reduced by 19.74% and 51.41% on average with 3 layers and 4 layers of 3D IC designs, respectively.*

I. Introduction

Three-dimensional (3D) IC technologies promise to further increase integration density, beyond Moore's Law, and offer the potential to significantly reduce interconnect delays and improve system performance. Furthermore, the shortened wirelength, especially that of the clock net, lessens the power consumption of the circuit. 3D IC technologies also provide a flexible way to carry out the heterogeneous system-on-chip (SoC) design by integrating disparate technologies, such as memory and logic circuits, radio frequency (RF) and mixed signal components, optoelectronic devices, etc., onto different layers of a 3D IC.

Since 3D IC technology enables an additional degree of freedom for circuit design, previous experiences on 2D designs may not be valid and need to be refreshed. To determine the system performance and reliability, it is pointed out that interconnect has become the dominating factor [1][2]. Thus, interconnect-centric analysis is very important in studying the impact of 3D IC technology.

Quantitative studies of wirelength reduction from 3D IC technology have been done for standard cell circuits [8][10]. The study [7] gives an early look on the promise of 3D IC technology. It roughly estimates both the 2D and 3D wirelength distributions by Rent's rule, and shows that the wiring reuirement is significantly reduced for the global wires in 3D ICs. The study in [8] shows that there is about 50% wirelength reduction of 4-layer 3D circuits compared to traditional 2D implementations. It supports the declaration in [14] that the ideal wirelength reduction, which ignores the cost of TS vias, is the square root of the number of available device layers. The study in [10] performs extensive experiments on the relations among wirelength, TS via number and temperature with different number of device layers. It shows that sacrificing only 2% of the ideal wirelength reduction can achieve 46% TS via number reduction for 4-layer 3D circuits. It also shows that temperature can be reduced by about 20% with only 1% less of the ideal wirelength reduction and 10% more TS vias.

However, there is a lack of quantitative study on the impact of 3D IC technology to the repeater usage, which has increased dramatically for 2D designs [1][2][3]. It is known that delay of an unbuffered segment of wire is quadratic to the wirelength, and repeaters are used to linearize the delay. Repeaters are very effective in reducing long interconnect, but they also consume a great amount of static and dynamic power. Therefore, it is important to quantify the benefit of 3D design on repeater reduction.

In this paper, we perform quantitative studies on repeaters usage due to the scaling in the 3rd dimension, which have different number device layers from 1 to 4. The repeater numbers are estimated in the post-placement/pre-routing stage. The placements for 1 device layer circuits are done by mPL6 [4], and the placements for 2 to 4 device layer circuits are done by mPL-3D [8]. The interconnection delay is estimated by the tool IPEM [11]. From the experimental results, we have observed a considerable decrease of repeaters which in turn reduce the power and area.

II. 3D placement

The quantitative studies of repeater estimation are performed at the post-placement stage. In this section we will first describe our placement methods. Separate 2D and 3D placements are done by the state-of-the-art 2D and 3D placers, mPL6 and mPL3D, respectively.

A. 2D Placer

Recent analytical global placers show very successful results in both quality and scalability [4][16][17]. Meanwhile, many studies [18] show that multilevel algorithm is a promising technique to handle large-scale problems. Therefore, we use a multilevel analytical placer for a high quality 2D placement.

The analytical placement engine [4][9] is to solve the placement problem by nonlinear optimization algorithms, which is formulated in Figure 1. The differentiability of the objective and constraint functions is required. The objective of half-perimeter wirelength is approximated by replacing the max function with log-sum-exp function [1]. The non-overlap constraints are replaced by density constraints, and the density function is smoothed by Helmholtz equation [4] for differentiability.

The multilevel framework [4] consists of coarsening, relaxation and interpolation, as showed in Figure 2. Coarsening is to build a hierarchy of the netlist by clustering. The coarsened netlist is still a netlist so that placement problem is solved at each level. The coarsest netlist is placed first and then the solution is interpolated to the finer netlist as an initial solution. This initial solution is relaxed by the analytical placement engine referred above. The solution at each level is interpolated to a finer level until a solution at the finest level is obtained. Additional cycles of such process may be applied.

Legalization and detailed placement [6] are applied after the global placement.

$$\begin{array}{ll} \text{minimize} & \text{WireLength}(x, y) \\ \text{subject to} & \text{Density}_{i,j} <= \text{Target\_Density} \end{array}$$

Figure 1 Nonlinear Programming

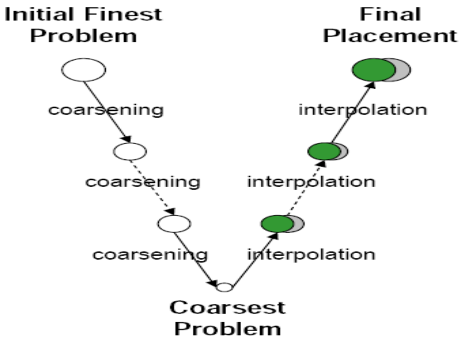


Figure 2 Multilevel Framework

B. 3D Placer

To take advantage of the existing high-quality 2D placer, we use the 3D placer based on 2D to 3D transformation [8]. It is a thermal-aware 3D placer providing trade-offs between the wirelength and the number of Through-Silicon (TS) vias. The trans-formation methods include local stacking, folding-2, folding-4, and window-based stacking/folding.

Among these transformation methods, local staking [8] performs best among in terms of wirelength, if ignoring the cost of TS vias. A transformation framework with the local stacking method is showed in Figure 3. For a K-layer 3D IC, a 2D placement is done first on a region K times larger than the placement region on one layer in the 3D IC. The 2D placement region is then shrinked uniformly to meet the 3D placement region, and remain the relative locations of the placed cells. The layer assignment of these cells is determined by a modified Tetris legalization, and is further refined by RCN graph. Layer-by-layer detailed placement is applied to obtain a final 3D placement.

The results [8] show that the wirelength of a 4-layer 3D IC can be as short as 50% of a 2D implementation.

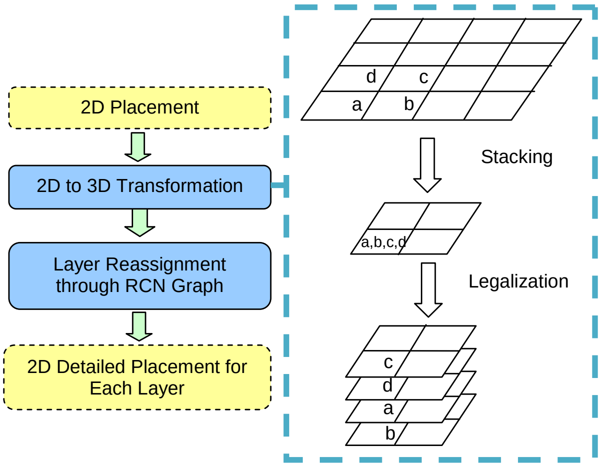

2D Placement
2D to 3D Transformation
Layer Reassignment through RCN Graph
2D Detailed Placement for Each Layer
d
c
a
b
Stacking
a,b,c,d
Legalization
c
d
a
b

Figure 3 Transformation-based 3D Placer

III. Repeater Estimation

After the placement, in this section we will introduce our approach of repeater estimation.

To optimize the repeater insertion in the on-chip interconnections for minimum delay and area, we developed IPEM [11], which can provide a set of procedures that estimate interconnect performance under various performance optimization algorithms for deep submicron technology.

Although many interconnect optimization algorithms, such as wire sizing and spacing, optimal buffer insertion, wire sizing optimization, global interconnect sizing and spacing and simultaneous driver, buffer, and interconnect sizing, have been intensively investigated previously, these approaches are generally developed for physical level, i.e., they are not efficient to be used in a higher level design and synthesis. However the interconnection should be considered as early as possible for design convergence. Under this circumstance, the interconnect estimation modeling techniques is proposed to get fast and accurate estimation of the optimal interconnect ion performance under various optimization algorithms.

IPEM is such a tool that through adopting a simple closed-form formulae or computational procedures, it can provide a fast yet accurate estimation of interconnection delay and area. As can be seen in

Figure 4, when provided the driver effective resistance of the input stage G_0, the driver effective resistance of G, interconnect wirelength l and loading capacitance C_L, IPEM can provide the optimal delay and area of the bold interconnection wire through optimization algorithms including OWS (Optimal Wire Sizing), SDWS (Simultaneous Driver and Wire Sizing), BIWS (Buffer Insertion and Wire Sizing) and BISWS (Buffer Insertion, Sizing and Wire Sizing), etc.

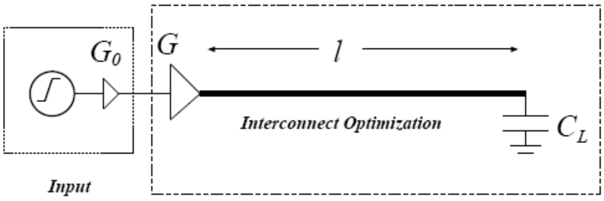


Figure 4 IPEM Interconnection Optimization

Here, we give an example of IPEM under the OWS optimization. For OWS, the size of the driver G in Figure 2 is fixed. Let $T_{ows}(R_d, l, C_L)$ be the delay under OWS for an interconnect l with driver resistance Rd and loading capacitance C_L. Through extensive analytical and numerical studies on the complex optimal wire shaping function, the following simple closed-form formulae under OWS is obtained.

$$T_{ows}(R_d, l, C_L) = (\alpha_1 l / W^2(\alpha_2 l) + 2\alpha_1 l / W(\alpha_2 l) + R_d c_f + \sqrt{R_d r c_a c_f l}) * l \tag{1}$$

$$\alpha_1 = \frac{1}{4} r c_a \tag{2}$$

$$\alpha_2 = \frac{1}{2}\sqrt{\frac{r c_a}{R_d C_L}} \tag{3}$$

Where r is the sheet resistance, c_a is the unit area capacitance, c_f is the unit effective-fringing capacitance, and $W(x)$ is Lambert's W function defined as the value of w that satisfies $we^w = x$. The closed-form area estimation formula is obtained as shown in (4).

$$A_{ows}(R_d, l, C_L) = \sqrt{\frac{r(c_f l + 2C_L)}{2R_d c_a}} l \tag{4}$$

The other optimization techniques of IPEM can be found in [11]. Experiment results [11] show that IPEM has an accuracy of 90% on average, with a running speed of 1000 times faster than Trio [12] which is a Tree-Repeater-Interconnect Optimization Package targeted at extensive interconnect layout optimization. Figure 5 shows the comparison results under 0.18 um technology for OWS between IPEM denoted by the squares and Trio denoted by the solid points, with $R_d = r_g/100$, $C_L = c_g*100$, where r_g and c_g are the output resistance and the input capacitance of a minimum device respectively.

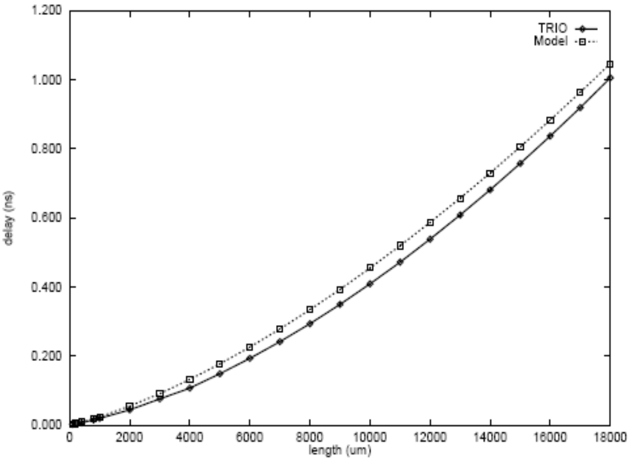


Figure 5 Comparison of Trios and IPEM for OWS

IV. Experimental Setup and Results

The experiments are performed on the IBM-PLACE benchmarks [13]. Since these benchmarks do not have source/sink pin information, to get relatively more accurate information of the net wirelength, we use the length of minimum-wirelength-tree of a net to estimate the optimal number of repeaters required in this net instead of using the half-bounding box method used in [8].

The rectilinear Steiner minimal tree has been widely used in early design stages such as physical synthesis, floorplanning, interconnect planning and placement to estimate wirelength, routing congestion and interconnect delay. It uses the minimum wirelength edges to connect nodes in a given net. In this paper, a rectilinear Steiner tree construction package FLUTE [14] is used to calculate the Steiner wirelength tree in order to estimate the repeater insertion without performing the detailed routing. FLUTE is based on pre-computed lookup table to make the Steiner minimum tree construction fast and accurate for low degree nets. For high degree nets, the net is divided into several low degree nets until the table can be used.

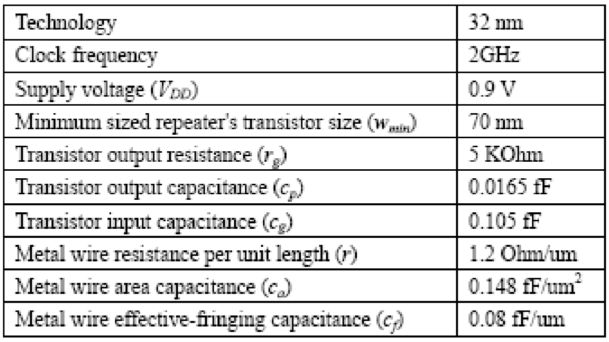
To accurately estimate the delay and area of the TS via resistance and capacitance, we use the approach in [15] to model the TS via as a length of wire. Because of its large size, the TS via has a great self-capacitance. By simulations on each via and the lengths of metal-2 wires in each layer, the authors in

[15] approximate the capacitance of an TS via with 3 μm thickness as roughly 8~20 μm of wire. The resistance is less significant because of the large cross-sectional area of each via (about 0.1 Ω per via), which is equivalent to about 0.2 μm of a metal-2 wire. We use 3D IC technology by MIT Lincoln lab and the minimum distance between adjacent layers is 2~3.45 μm. Thus, we can approximately transform all the TS vias between adjacent layers as 14 μm wires (an average value of 8~20 μm). This value is doubled when the via is going through two layers.

Since FLUTE can only generate a 2D minimum wirelength tree, in order to transform it to a 3D tree for our 3D designs, we make the following assumptions: 1) Assume that all the tree wires are placed in a middle layer of the 3D stack layers, 2) The pins in other layers use TS vias to connect to the tree on the middle layer. This assumption minimizes the total traditional wires in a net but overestimates the total number of TS vias. However, it can provide us a more accurate information of the total net wirelength compared to the 3D via and wirelength estimation method used in [8] where the number of via is simply set as the number of the layers the net spans.

The experiments are performed under 32 nm technology. The technology parameters we used to configure IPEM are listed in Table 1. We run FLUTE and IPEM for each net in each benchmark.

Table 1 Technology Parameters



Technology	32 nm
Clock frequency	2GHz
Supply voltage (V_{DD})	0.9 V
Minimum sized repeater's transistor size (w_{min})	70 nm
Transistor output resistance (r_g)	5 KOhm
Transistor output capacitance (c_p)	0.0165 fF
Transistor input capacitance (c_g)	0.105 fF
Metal wire resistance per unit length (r)	1.2 Ohm/um
Metal wire area capacitance (c_a)	0.148 fF/um^2
Metal wire effective-fringing capacitance (c_f)	0.08 fF/um

Table 2 shows the comparison results for IBM-PLACE benchmarks. As can be seen, by using 3D placement with 3 layers, the total wirelength can be reduced by 15.49% and the number of repeaters used in interconnection can be reduced by 19.74% respectively on average compared to the case of 2D design. Furthermore, when 4 layers are used in the 3D placement, the wirelength can be further reduced by 42.13% and the number of repeaters can be reduced by 51.41%.

As shown in Table 2, the reduction in the number of repeaters through 3D IC compared to that of the 2D cases is always more than the reduction of the total wirelength. This is because increasing the number of layers will efficiently decrease the length of the nets with a large minimum wirelength tree, and nets with a very small minimum wirelength tree always do not need repeaters. As can be seen in the IPEM results, wires less than 500 um usually result in 0 repeaters. Therefore, by reducing the nets with a large length of the minimum wirelength tree, we can significantly reduce the number repeaters and the area/power of the on-chip interconnection.

V. Conclusion

Using 3D technology coupled with a state-of-art 3D placement tool, we have observed a significantly reduction in the number of repeaters used in the on-chip interconnections. By reducing the repeater usage, we expect to achieve a considerable saving of the power and area of the on-chip interconnections.

Acknowledgements

This work is partially supported by IBM under a DARPA subcontract and by GRC under contract 2005-TJ-1317.

Reference

[1] J. Cong, "An Interconnect-Centric Design Flow for Nanometer Technologies", *Proceedings of the IEEE*, vol. 89, no. 4, pp 505-528, April 2001.

[2] J. Cong, L. He, K. Y. Khoo, C. K. Koh and Z. Pan "Interconnect Design for Deep Submicron ICs," *Proc. IEEE Int'l Conf. on Computer-Aided Design, San Jose*, California, pp. 478-485, Nov. 1997.

[3] P. Saxena, N. Menezes, P. Cocchini, and D.A. Kirkpatrick, "Repeater scaling and its impact on CAD," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol.23, no.4, pp. 451-463, April 2004.

[4] J. Cong, T. Chan, J. Shinnerl, K. Sze and M. Xie, "mPL6: Enhanced Multilevel Mixed-size Placement," *Proceedings of the ACM International Symposium on Physical Design,* San Jose, CA, pp. 212-214, April 2006.

[5] G.-J. Nam, "ISPD 2006 Placement Contest: Benchmark Suite and Results," *Proceedings of the 2006 International Symposium on Physical Design*, pp. 167-167, 2006.

[6] J. Cong, M. Xie, "A Robust Mixed-Size Legalization and Detailed Placement Algorithm," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol.27, no.8, pp.1349-1362, Aug. 2008.

[7] K. Banerjee, S.J. Souri, P. Kapur, and K.C. Saraswat, "3-D ICs: a Novel Chip Design for Improving Deep-submicrometer Interconnect Performance and Systems-on-chip Integration," *Proceedings of the IEEE*, vol. 89, no. 5, pp. 602-633, May 2001.

[8] J. Cong, G. Luo, J. Wei, Y. Zhang, "Thermal-Aware 3D IC Placement via Transformation," *Proc. of the 12th Asia and South Pacific Design Automation Conference,* Yokohama, Japan, pp. 780-785, Jan. 2007.

[9] J. Cong and G. Luo, "Highly Efficient Gradient Computation for Density-Constrained Analytical Placement Methods",

Proceedings of the 2008 ACM International Symposium on Physical Design, Portland, Oregon, pp. 39-46, April 2008

[10] B. Goplen and S. Spatnekar, “Placement of 3D ICs with thermal and interlayer via considerations,” *Proc. of the 44th annual conference on Design automation*, pp. 626-631, 2007.

[11] J. Cong and D.Z. Pan, "Interconnect Estimation and Planning for Deep Submicron Designs", *Proc. of Design Automation Conference*, New Orleans, LA., pp. 507-510, June, 1999

[12] J. Cong, L. He, C.K. Koh and Z. Pan, "Global Interconnect Sizing and Spacing with Consideration of Coupling Capacitance", *ACM/IEEE Int'l Conf. on Computer-Aided Design*, pp. 628-633, Dec. 1997

[13] http://er.cs.ucla.edu/benchmarks/ibm-place/

[14] C. Chu, Y. Wong, "FLUTE: Fast Lookup Table Based Rectilinear Steiner Minimal Tree Algorithm for VLSI Design", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* Vol. 27, Issue 1, pp. 70 - 83, Jan. 2008

[15] W.R. Davis; J. Wilson, S. Mick, etc., "Demystifying 3D ICs: the pros and cons of going vertical," *IEEE Design & Test of* Computers, Vol. 22, Issue 6, pp. 498 - 510, Nov.-Dec. 2005.

[16] T.-C. Chen, Z.-W. Jiang, T.-C. Hsu, H.-C. Chen, and Y.-W. Chang, "A High-quality Mixed-size Analytical Placer Considering Preplaced Blocks and Density Constraints," *Proceedings of the 2006 IEEE/ACM International Conference on Computer-Aided Design*, San Jose, CA, pp. 187-192, November 2006.

[17] H. Eisenmann and F. M. Johannes, "Generic Global Placement and Floorplanning," *Proceedings of the 35th Annual Conference on Design Automation*, San Francisco, CA, pp. 269-274, June 1998.

[18] “Multilevel Optimization and VLSICAD,” ed. J. Cong and J.R. Shinnerl, *Kluwer Academic Publishers*, Boston, 2002.

[19] W. C. Naylor, R. Donelly, and L. Sha, “Non-linear Optimization System and Method for Wire Length and Delay Optimization for an Automatic Electric Circuit Placer,” *US Patent 6301693*, October 2001.

Table 2 Results of the number of wirelength/repeaters for IBM-PLACE Benchmarks

Benchmarks	2D Placement		3D Placement (3 Layers)				3D Placement (4 Layers)			
	Total Wire length (um)	#Repeaters	Total Wire length (um)		#Repeaters		Total Wire length (um)		#Repeaters	
				Reduced		Reduced		Reduced		Reduced
ibm 01	5,340,531	5,241	5,705,513	-6.83%	5,623	-7.29%	3,690,917	30.89%	2,866	45.32%
ibm 02	15,733,437	18,370	13,231,177	15.90%	14,674	20.12%	8,811,791	43.99%	8,543	53.49%
ibm 03	15,624,821	18,023	11,654,377	25.41%	12,217	32.21%	8,941,162	42.78%	8,372	53.55%
ibm 04	18,478,722	20,720	14,883,848	19.45%	15,508	25.15%	10,486,259	43.25%	9,382	54.72%
ibm 05	41,260,244	52,740	33,841,068	17.98%	41,795	20.75%	28,683,802	30.48%	34,758	34.10%
ibm 06	25,726,920	29,757	19,515,728	24.14%	20,601	30.77%	15,071,901	41.42%	14,366	51.72%
ibm 07	40,571,536	48,630	30,162,968	25.65%	33,217	31.69%	22,374,322	44.85%	22,557	53.62%
ibm 08	45,723,304	55,685	34,622,016	24.28%	39,096	29.79%	25,579,604	44.06%	26,682	52.08%
ibm 09	36,590,608	42,210	27,722,182	24.24%	28,653	32.12%	21,091,114	42.36%	20,037	52.53%
ibm 10	62,148,072	74,318	49,074,524	21.04%	54,361	26.85%	34,961,648	43.74%	35,428	52.33%
ibm 11	43,441,568	47,504	42,149,136	2.98%	45,018	5.23%	24,954,992	42.56%	21,266	55.23%
ibm 12	75,913,368	92,264	67,831,048	10.65%	80,047	13.24%	42,164,064	44.46%	44,254	52.04%
ibm 13	71,395,456	84,036	67,278,656	5.77%	76,909	8.48%	40,258,068	43.61%	39,376	53.14%
ibm 14	125,077,064	146,899	108,393,400	13.34%	121,384	17.37%	69,462,392	44.46%	67,215	54.24%
ibm 15	165,876,560	201,767	149,695,664	9.75%	176,427	12.56%	93,114,488	43.87%	97,190	51.83%
ibm 16	181,410,896	219,291	156,584,160	13.69%	180,619	17.64%	101,196,632	44.22%	103,891	52.62%
ibm 17	266,743,968	336,971	224,178,208	15.96%	273,579	18.81%	146,224,448	45.18%	163,610	51.45%
Average				15.49%		19.74%		42.13%		51.41%