CAD for 3D Circuits: Solutions and Challenges

Sachin S. Sapatnekar
Department of Electrical and Computer Engineering
University of Minnesota, Minneapolis, MN 55455.

Abstract

3D technology provides a number of challenges and opportunities to the designer. To leverage this technology fully, it is essential to develop compuer-aided design techniques that are 3D-specific. This paper discusses problems in the area of layout, thermal optimization, and power supply network design that are critical in the design of 3D chips.

1 Introduction

3D technology opens up an entirely new dimension, literally and figuratively, for circuit design, making new sets of design structures and architectures available and feasible. For a chip designer at the leading-edge, this presents a range of opportunities, and also a significant set of challenges. Even in the conventional 2D world, chip design is an extremely laborious and time-intensive task; with the new horizons opened up by 3D, the range of choices increases, as does the level of complexity. Computer-aided design (CAD) tools, which are indispensable in 2D design, become even more significant in addressing the challenges of 3D. This paper presents specific points where CAD tools can play in facilitating 3D design.

The move from 2D to 3D is inherently a topological change, and therefore, several of the problems that are unique to 3D circuits are related to physical design optimizations that determine the circuit layout. This applies to layout optimizations such as placement and routing, which will be discussed in this paper.

Another issue is related to the fact that circuitry can be packed more densely in 3D than in 2D. While this is clearly a major advantage, it also brings about new limitations and challenges to the designer, in terms of how the chip interacts with the environment. A k-tier 3D chip could use k times as much current as a single 2D chip of the same footprint; however, the packaging technology is not appreciably different. This has major implications from the point of view of packaging limitations:

- First, the 3D chip generates k times the power of the 2D chip, which implies that the corresponding heat generated must be sent out to the environment using a substantially similar package, or that 3D chips must face higher temperatures. The latter is highly undesirable, since elevated temperatures can hurt performance and reliability, in addition to introducing variabilities in the performance of the chip. Therefore, on-chip thermal management is a critical issue in 3D design.

- Second, the package must be capable of supplying k times the current through the power supply (V_{dd} and ground) pins, as compared to the 2D chip. Given that reliable power grid design is a major bottleneck even for 2D designs, this implies that significant resources have to be invested in building a bulletproof power grid for the 3D chip.

2 3D Thermal Analysis

Thermal issues are a key factor in 3D design. At the full-chip level, the ideas of heat transfer can be applied to determine the on-chip temperature. A typical design consists of multiple tiers of active devices stacked on each other, each dissipating power and generating heat. This heat is then transferred through the silicon and dielectric structure to the heat sink, and the strength of the heat conduction paths determines the corresponding rise in temperature. Briefly, if the effective thermal conductivities from the dominant heat sources to the heat sink are small, then the rise in temperature will be small; if not, not.

Heat conduction at the macro scale implied by such systems is governed by Fourier's Law, given by:

$$\nabla^2 T(\mathbf{r}) + \frac{g(\mathbf{r})}{k_{\mathbf{r}}} = \frac{\rho c}{k_{\mathbf{r}}} \frac{\partial T}{\partial t}$$

where k is the *thermal conductivity* at the particular location, ρ is the density of the material and c is the *specific heat capacity*. g is the volume power density, which is also location dependent. Usually the problem is formulated in three-dimensional space, therefore $\mathbf{r}$ is a three-dimensional array $\mathbf{r} = (x, y, z)$. Since the time constant of on-chip temperature change is usually in the order of milliseconds, while the operating frequency of electric signal is in the range of picoseconds, for the purposes of analysis, it is sufficient to solve the steady-state problem. The heat diffusion equation can then be simplified as the following Poisson's equation:

$$\nabla^2 T(\mathbf{r}) = -\frac{g(\mathbf{r})}{k_{\mathbf{r}}} \tag{1}$$

The boundary conditions for this equation reflect the heat sinking environment. For multilayer systems that have piecewise constant thermal conductivities, a continuity equation is applied for the temperature and heat flux at the boundaries of each piecewise constant region. From the sides of the chip, the boundary conditions are taken to be adiabatic, since no heat can escape. For the heat sink, it is reasonable to coarsely assume an isothermal condition, which states that the heat sink is at a constant temperature, the ambient temperature (a constant value that has the function of the ground node in an electrical circuit). Alternatively, a conductive boundary condition could use a macromodel for the heat sink, connected to a node representing the ambient temperature.

Classical techniques for thermal analysis, such as finite difference methods, finite element analysis, and boundary element methods, are well established for solving thermal problems, and several

CAD techniques have employed these ideas for thermal analysis. However, opportunities for improved analysis, leveraging the structure of the 3D problem, still do exist. Solving the analysis problem requires a higher level of accuracy than some thermal minimization problems, which may have to choose the best thermal solution between a set of configurations, and requires fidelity in rank-ordering rather than absolute accuracy. Thermally-constrained analysis lies between these two, since it must evaluate the temperature reasonably accurately to determine whether it meets a constraint or not.

3 Thermally-driven 3D Placement

The placement problems determines the optimal locations of standard cells, arranged in rows within the tiers of the three-dimensional circuit. The input to the placer is a technology-mapped netlist and a description of the library. Temperature is treated as a first-class citizen during this optimization, in addition to other conventional metrics, and intertier via reduction is also considered to be a desirable goal. We will now briefly describe our efforts at the University of Minnesota on building a 3D placer.

Our first generation placer [1] used a force-directed placement paradigm, where an analogy to Hooke's law for springs is used between cells. Cells that are connected to each other by a net are deemed to have an attractive force (with the idea that this will bring them closer together, and shorten the overall net lengths). Connections to the boundaries of the chip help to keep the cells distributed. Forces corresponding to thermal stresses and cell overlap are also added and in each iteration, the location of each cells that corresponds to the minimum energy state of the system is computed. The problem is solved in the continuous space, and the solution is snapped on to the set of allowable discrete locations.

The second generation placer [2] observes that since 3D layouts have very limited flexibility in the third dimension (with a small number of layers and a fixed set of discrete locations), the assumption of near-continuity used in force-directed placers is inherently limited in the z direction, where the number of layers is typically in the range of 3 to 10 in foreseeable technologies. Given this limitation, partitioning works better than a force-directed method. Accordingly, this work performs global placement using recursive bisectioning. To speed up the placement, instead of an embedded thermal analysis engine, thermal effects are incorporated through *thermal resistance reduction nets*, which are attractive forces that provide incentives for high power nets to remain close to the heat sink. The global placement step is followed by coarse legalization, in which a novel cell-shifting approach is proposed. This generalizes the methods in FastPlace [3] by allowing shift moves to adjust the boundaries of both sparsely and densely populated cells using a computationally simple method. Finally, detailed legalization generates a final nonoverlapping layout. The approach is shown to provide excellent tradeoffs between parameters such as the number of interlayer vias, wire length, temperature. An example tradeoff between the number of inter-tier vias and the total wirelength is shown in Figure 1 [4].

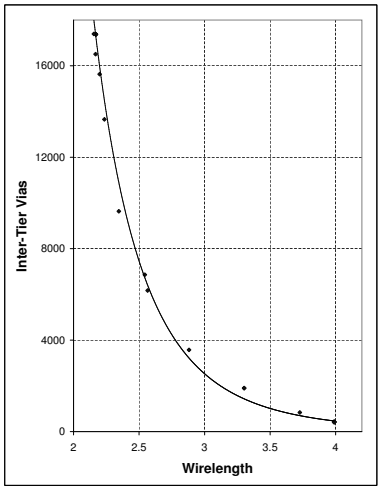

Figure 1: Tradeoff between the number of interlayer vias versus the total wire length for the benchmark ibm01

3.1 Thermally-driven 3D Routing

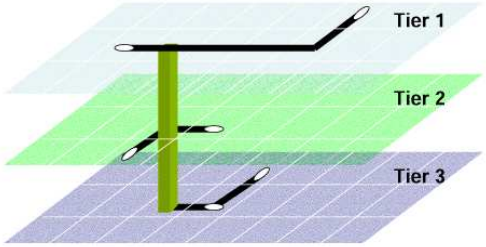

Figure 2: An example route for a net in a three-tier 3D technology.

Once the cells have been placed and the locations of the thermal vias determined, the routing stage finds the optimal interconnections between the wires. As in 2D routing, it is important to optimize the wire length, the delay, and the congestion. In addition, several 3D-specific issues come into play. First, the delay of a wire increases with its temperature, so that more critical wires should avoid the hottest regions, as far as possible. Second, inter-tier vias are a valuable resource that must be optimally allocated among the nets. Third, congestion management and blockage avoidance is more complex with the addition of a third dimension. For instance, a signal via or thermal via that spans two or more tiers constitutes a blockage that wires must navigate around.

Each of the above issues can be managed through exploiting the flexibilities available in determining the precise route within the bounding box of a net, or perhaps even considering slight

detours outside the bounding box, when an increase in the wire length may improve the delay or congestion or may provide further flexibility for inter-tier via assignment.

Consider the problem of routing in a 3-tier technology, as illustrated in Figure 2. The layout is gridded into rectangular tiles, each with a horizontal and vertical capacity that determines the number of wires that can traverse the tile, and an inter-tier via capacity that determines the number of free vias available in that tile. These capacities account for the resources allocated for non-signal wires (e.g., power and clock wires) as well as the resources used by thermal vias. For a single net, as shown in the figure, the degrees of freedom that are available are in choosing the locations of the inter-tier vias, and selecting the precise routes within each tier. The locations of inter-tier vias will depend on the resource contention for vias within each grid. Moreover, critical wires should avoid the high-temperature tiles, as far as possible.

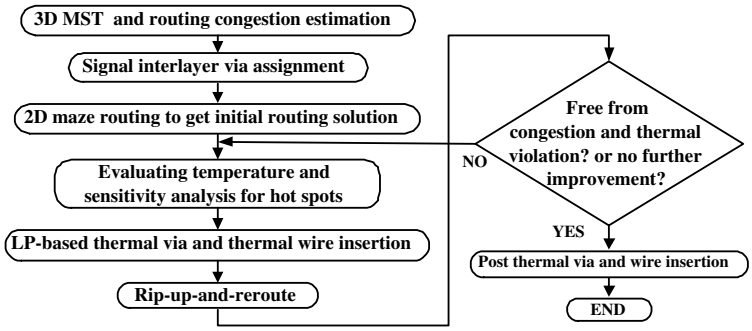

Figure 3: Overall flow for the temperature-aware 3D global routing algorithm.

In [5], a method for effectively reducing on-chip temperatures is proposed, through the appropriate insertion of "thermal vias" and a new construct, "thermal wires." A routing solution that obeys thermal and routing capacity constraints are then presented. Thermal vias correspond to vertical interlayer vias that do not have any electrical function, but are explicitly added as thermal conduits, while thermal wires perform a similar function, but conduct heat laterally within the same layer. Thermal vias perform the bulk of the conduction to the heat sink, while thermal wires help distribute the heat paths over multiple thermal vias. The routing scheme begins with routing congestion estimation and signal interlayer via assignment, followed by thermally-driven maze routing. Sensitivity analysis is employed and linear programming (LP) based thermal via/wire insertion is performed to reduce temperature. The above process is iteratively repeated until temperature and routing capacity violations are resolved. Experimental results show that the scheme can effectively resolve the contentions between thermal via/wire and routing, generating a solution satisfying both congestion and temperature requirements. The overall flow of the method is outlined in Figure 3.

4 Power Grid Design in 3D

3D integrated circuits are extemely limited in terms of their ability to deliver power, due to I/O pin limitations. I/O pins fall into two categories: those used to deliver signals and those that deliver power (Vdd or ground). While increased on-chip functionalities lead to a demand for more signal I/O pins, the demand for power is stronger. In contemporary and future technologies, one-half to two-thirds of all I/O pins must be dedicated to power delivery so as to reduce the worst case IR-drop and $L\frac{di}{dt}$ noise in the power grids, while the total number of package pins does not increase significantly. As a result, as the total current consumption of the chip goes up due to the increasing circuit complexity and higher switching frequency, each power pin must deliver a larger amount of current to the chip. This trend can be clearly seen in Figure 4, which is created based on the data from ITRS [6]. In other words, as the IC technology advances, the number of pins for each unit of current delivered is actually reduced.

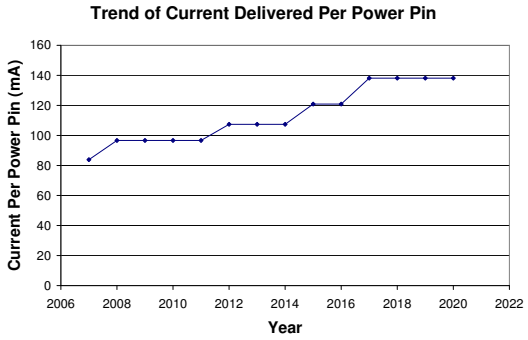

Figure 4: Trend of current delivered per power pin based on the data from ITRS.

The pin limitation problem is exacerbated in 3D ICs: for a 3D chip with k tiers, the amount of current to be supplied is k times as much as for a 2D chip with the same footprint, but the number of available pins is essentially the same. Viewing this another way, if we were to transform a 2D IC to a k-tier 3D IC implementing the same functionality, the number of pins accessible to the circuit will be reduced to $\frac{1}{k}$ of the original value because of the much smaller footprint area. Figure 5 shows an example of transforming a 2D IC to an equivalent 3-tier 3D IC, in which the number of pins that can be placed on the 3D chip is only one-third of that for the corresponding 2D chip.

This area has hardly been addressed in the past, and we will outline one solution method in this domain, using the idea of stacked V_{dd} levels. In [7], a high-tension power delivery scheme was proposed to reduce power grid noise and the effect of electromigration. In this new circuit paradigm,

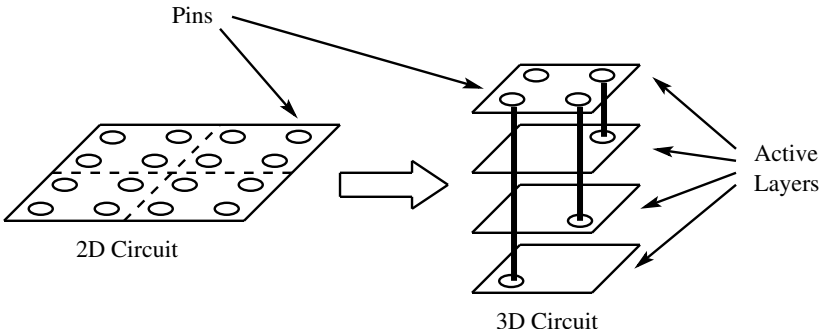

Figure 5: Power pin limitations for a 3D chip, as compard to a 2D chip with the same amount of circuitry.

logic blocks are stacked several levels high and power is delivered to the circuit as multiples of the regular supply voltage Vdd. Next, the delivered high-tension supply voltage is divided into several Vdd domains each of which has a range of Vdd, and circuit blocks are distributed to different Vdd domains. Voltage regulators are used to control the voltage levels of internal supply rails. This is illustrated in Figure 6.

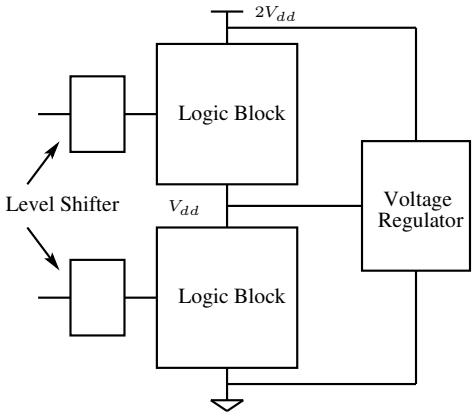

Figure 6: A schematic of a 2-level stacked-Vdd circuit structure.

The advantage of this new circuit structure is that the current can be "recycled" between stacks. When logic blocks are stacked n levels high and the current requirements between logic blocks operating in different V_{dd} domains are balanced, the current flowing through each external power grid would be reduced to $\frac{1}{n}$ of the original value, where the words external power grid refer to a power grid that is connected to power pins, i.e., nV_{dd} and GND rails in an n-level stacked- V_{dd} circuit. Therefore, voltage drop, noise, electromigration issues can be significantly alleviated.

The key point here is to build partitions in such a way that the current can indeed be recycled through successive stacks of V_{dd} layers; if it cannot, it flows through the voltage regulators and is wasted. A forthcoming publication [8] presents a novel partitioning method for solving this problem. The chip is partitioned into regions, each of which has one voltage regulator. A Voronoi tessellation of the space determines the fraction of current drawn by each block from each regulator. Next, the partitioner allocates blocks to regions so as to minimize wasted power. Results on a 3D benchmark circuit show that 95% of the power is usefully recycled, and only 5% is wasted, through the use of an intelligent partitioning scheme.

Acknowledgements

This work is based on the research of Brent Goplen, Yong Zhan and Tianpei Zhang.

References

[1] B. Goplen and S. S. Sapatnekar. Efficient thermal placement of standard cells in 3D ICs using a force directed approach. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pages 86–89, 2003.

[2] B. Goplen and S. S. Sapatnekar. Placement of 3D ICs with thermal and interlayer via considerations. In *Proceedings of the ACM/IEEE Design Automation Conference*, pages 626–631, 2007.

[3] N. Viswanathan and C. C.-N. Chu. FastPlace: Efficient analytical placement using cell shifting, iterative local refinement and a hybrid net model. In *Proceedings of the ACM International Symposium on Physical Design*, pages 26–33, 2004.

[4] B. Goplen. *Advanced Placement Techniques for Future VLSI Circuits*. PhD thesis, University of Minnesota, Minneapolis, MN, 2005.

[5] T. Zhang, Y. Zhan, and S. S. Sapatnekar. Temperature-aware routing in 3D ICs. In *Proceedings of the Asia-South Pacific Design Automation Conference*, pages 309–314, 2006.

[6] Semiconductor Industry Association. International technology roadmap for semiconductors, 2006. available at `http://public.itrs.net/Links/2006Update/2006UpdateFinal.htm`.

[7] S. Rajapandian, K. Shepard, P. Hazucha, and T. Karnik. High-tension power delivery: Operating 0.18μm CMOS digital logic at 5.4v. In *Proceedings of the IEEE International Solid-State Circuits Conference*, pages 298–299, 2005.

[8] Y. Zhan, T. Zhang, and S. S. Sapatnekar. Module assignment for pin-limited designs under the stacked-Vdd paradigm. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, 2007. (forthcoming).