Placement of 3D ICs with Thermal and Interlayer Via Considerations

Brent Goplen*
IBM Sys. & Tech. Group
Essex Junction, VT
bagoplen@us.ibm.com

Sachin Sapatnekar
University of Minnesota
Minneapolis, MN
sachin@ece.umn.edu

Abstract

Thermal problems and limitations on interlayer via densities are important design constraints on three-dimensional integrated circuits (3D ICs), and need to be considered during global and detailed placement. Analytical and partitioning-based techniques are developed to explore the tradeoff between wirelength, interlayer via counts, and thermal effects. This method allows wirelengths to be minimized for any desired interlayer via density and temperatures to be reduced while minimizing deleterious effects on wirelength and interlayer via counts. Wirelength reductions within 2% of the optimal can be achieved using 46% fewer interlayer vias. Temperatures can be reduced by about 20% with only 1% higher wirelengths and 10% more interlayer vias.

Categories and Subject Descriptors
B.7.2 [**Integrated Circuits**]: Design Aids – *Placement and routing*; B.7.1 [**Integrated Circuits**]: Types and Design Styles – *Advanced technologies*

General Terms
Algorithms, Design, Experimentation, Performance

Keywords
3-D IC, 3-D VLSI, thermal optimization, temperature, placement, interlayer vias

1. Introduction

Besides scaling, the enhanced integration densities predicted by Moore's law may be achieved through new technologies such as three-dimensional (3D) integration, which stack multiple active layers into a monolithic chip. However, 3D ICs have significantly larger power densities than their 2D counterparts and high thermal resistances between active layers. Unless 3D circuits are carefully designed, they can face severe thermal problems that can reduce their performance and reliability. In addition, the maximum allowable density of interlayer vias (providing connectivity between active layers) is greatly restricted in 3D ICs due to fabrication limitations.

A key characteristic of 3D ICs is the presence of interlayer vias that electrically connect vertically adjacent areas and allow routing to greatly reduce wirelengths. However, they are difficult to fabricate, and their densities are limited. Recently, there has been a lot of work in the placement of 3D ICs using nonlinear programming [1], quadratic/force-directed placement [2] [3] [4], and partitioning placement [5] [6] methods. These methods tend to either minimize wirelengths without regard to interlayer via counts or minimize interlayer via counts without regard to the wirelength. Being able to adjust to the desired tradeoff point would be of great utility to a designer so that wirelength can be minimized for any required interlayer via density. The need to optimize both wirelength and interlayer via counts differentiates 3D placement from traditional 2D placement.

In addition, thermal problems are particularly prominent in 3D ICs because of high power densities and low thermal conductivities. Further technology scaling also exacerbates these high power densities. Previous work in thermal placement has been quite limited, particularly with regard to 3D ICs. Some thermal placement methods such as in [7] strive to achieve a uniform power distribution, but power dissipation may need to be concentrated in the bottom layer of 3D ICs or in some other way to provide more efficient heat removal to the heat sink and reduced temperatures. Net weighting methods that use only the switching activities of nets, such in [8], can be used to reduce the dynamic power to produce better thermal results, but neglect to represent the thermal environment of the driver cells, where power is being dissipated, in their net weight formulation. Thermal simulations can also be used to guide thermal placement such as in [2] [9] [10], but this adds to the computational cost.

A partitioning-based approach appears to be well suited for the placement of 3D ICs. Partitioning placement can efficiently reduce interlayer via counts with its intrinsic min-cut objective and can obtain good placement results even when IO pad connectivity information is missing. In contrast, the force-directed paradigm relies on an encompassing arrangement of IO pads, which 3D ICs may not have, to produce a well-spread initial placement in order to proceed efficiently and effectively in subsequent iterations [4].

For any thermal placement method to be completely effective, it must also actively reduce power because power has a direct impact on temperatures. If power is disregarded in the thermal placement formulation, any wirelength degradation caused by thermal placement will in turn increase the power and subsequently the temperatures. In addition, the cost of interlayer vias must be incorporated into the objective function for thermal placement. With our method, net weights are added to reduce the power selectively during partitioning-based global placement with additional nets added to move cells to more favorable thermal environments. Detailed placement methods were developed to maintain the improvements made during global placement by using the same objective function in determining cell movements.

2. Overview of the Method

Our placement method for 3D ICs is composed of three steps. *Global placement* uses a recursive bisection algorithm in which the cut direction is determined at each level with the objective function in mind. Minimizing wirelength, interlayer via counts, and thermal effects are the primary objectives of *global placement* with cell overlap removal being a secondary objective. *Coarse legalization* combines local and global cell moves for

* This work was done while Brent Goplen was with the University of Minnesota, Minneapolis, MN. It was supported in part by DARPA, NSF, and by an SRC fellowship.

DAC 2007, June 4–8, 2007, San Diego, California, USA

improving the objective function value with *cell shifting* for spreading and overlap removal. It is used to make significant improvement in removing overlap between cells in preparation for *detailed legalization* while maintaining the placement quality produced by *global placement.* Detailed legalization performs the fine-grained work of completely removing all overlaps by placing cells in the nearest available free space that produces minimal objective function degradation.

Each of the above steps seeks to minimize the wirelength and interlayer via counts, and account for thermal considerations. This can be represented with the following objective function in which thermal considerations are represented as a weighted sum of the cell temperatures:

$$\sum_{\text{each net } i}\left[WL_i + \alpha_{ILV} \cdot ILV_i\right] + \alpha_{TEMP} \sum_{\text{each cell } j}\left[T_j\right] \quad (1)$$

where WL_i is the bounding box wirelength and ILV_i is the number of interlayer vias for net i, T_j is the temperature of cell j, α_{ILV} is the interlayer via coefficient, and α_{TEMP} is the thermal coefficient. Cell temperatures are dependent on both the power dissipation of the cell and the thermal environment around the cell. The power dissipation depends on the capacitance of the net that it drives, and this capacitance depends on the length of the net, capacitance per length, and the fan-out. The thermal environment around the cell depends on the thermal resistance from that position to the heat sink and the temperature contributions from other cells. The thermal resistance in turn depends on the distance to the heat sink, the thermal conductivities of the materials on the way to the heat sink, and the boundary conditions used to represent the heat sink.

However, in practice, using temperatures directly in the objective function can result in expensive recalculations for each individual cell movement, and therefore, simplifications need to be made for enhanced efficiency. It should be noted that the temperature at each cell position is a sum of the temperature contributions from all power signatures in the chip, and the temperature contribution from the cell's own power, ΔT_j, is typically the dominant term and can be quickly calculated using:

$$\Delta T_j = R_j^{cell} P_j^{cell} \quad (2)$$

where R_j^{cell} is the thermal resistance from cell j to ambient, and P_j^{cell} is the power dissipation of cell j. By using ΔT_j instead of T_j and by applying Equation (2), the objective function can be modified so that it is efficient to calculate during placement:

$$\sum_{\text{each net } i}\left[WL_i + \alpha_{ILV} \cdot ILV_i\right] + \alpha_{TEMP} \sum_{\text{each cell } j}\left[R_j^{cell} P_j^{cell}\right] \quad (3)$$

In our method, thermal resistances, R_j^{cell}, are calculated using simple heat conduction and convection equations assuming that heat flows in a straight path from the cell to the chip surface in all three directions and that the cross sectional area of each path is the same size as the cell, but more sophisticated thermal resistance calculations could be used instead if desired. In this formulation, we also assume that dynamic power dominates the total power and is primarily dissipated in the cells because driver resistances are usually much larger than interconnect resistances. The dynamic power of net i is given by

$$P_i^{net} = \tfrac{1}{2} f V_{DD}^{\ 2} a_i C_i^{total} \quad (4)$$

$$C_i^{total} = C_{per\ wl} WL_i + C_{per\ ilv} ILV_i + C_{per\ pin} n_i^{input\ pins} \quad (5)$$

where f is the clock frequency, V_{DD} is the supply voltage, a_i is the switching activity, C_i^{total} is the total capacitance of net i, $C_{per\ wl}$ is the capacitance per wirelength, $C_{per\ ilv}$ is the capacitance per interlayer via, $C_{per\ pin}$ is the input pin capacitance, and $n_i^{input\ pins}$ is the number of cell input pins attached to net i.

3. Global Placement

Our global placement method uses a recursive bisection approach applied to the 3D context. At each level, regions are defined as containing a subset of cells in the netlist and a certain physical portion of the placement area. When a region is bisected, two new regions are created from the partitioned list of cells and the divided physical area, and these regions are processed in a breadth-wise manner. For each bisection, the cut direction is selected as orthogonal to the largest of the width, height or *weighted depth* of the region where the *weighted depth* is the depth (z-direction) of the region multiplied by α_{ILV}/d_{layer} and d_{layer} is the layer thickness. By doing this, the min-cut objective minimizes the number of connections in the costliest direction at the expense of allowing higher connectivity in the less costly orthogonal directions. For the partition of each region, terminal propagation [11] is used so that connectivity to areas outside the region is considered. Partitioning tolerance is calculated to correspond to the amount of whitespace available in the region. After partitioning, the cut line is positioned to ensure an even distribution of cell area. The cell area in the two new regions is used to adjust the position of the boundary between them.

Using the objective function from Equation (3), thermal concerns are added to the method using *net weighting* (Section 3.1) and *thermal resistance reduction nets* (Section 3.2). The net weighting scheme takes both the thermal environment of the driver cells and the potential power usage of the nets into consideration. Different net weights are created for the lateral (x and y) and vertical (z) directions in order to take into account the interlayer via coefficient and differences in capacitance per unit length in different directions. *Thermal resistance reduction nets* are created to move cells toward areas of lower thermal resistance based on their power dissipation.

3.1 Thermal-Aware Net Weighting

The *thermal net weighting* scheme takes into consideration the thermal resistance at the driver cells, the switching activity of the net, and the capacitances per unit length. By extracting the lateral and vertical net length components from Equation (3) and using Equations (4) and (5), the following objective function is obtained for deriving the net weights:

$$\sum_{\text{each net } i}\left[WL_i + \alpha_{ILV} \cdot ILV_i\right] + \alpha_{TEMP} \sum_{\text{each cell } j}\left[R_j^{cell} \sum_{\substack{\text{each driven} \\ \text{net } i \text{ of cell } j}}\left(s_i^{wl} WL_i + s_i^{ilv} ILV_i\right)\right] \quad (6)$$

where $s_i^{wl} = \frac{\frac{1}{2} f V_{DD}^{\ 2} a_i C_{per\ wl}}{n_i^{output\ pins}}$ and $s_i^{ilv} = \frac{\frac{1}{2} f V_{DD}^{\ 2} a_i C_{per\ ilv}}{n_i^{output\ pins}}$.

Changing the order of summation of the second term yields

$$\sum_{\text{each net } i}\left[\left(1 + \alpha_{TEMP} R_i^{net} s_i^{wl}\right) WL_i + \alpha_{ILV}\left(1 + \frac{\alpha_{TEMP} R_i^{net} s_i^{ilv}}{\alpha_{ILV}}\right) ILV_i\right] \quad (7)$$

where $R_i^{net} = \sum_{\substack{\text{each driver} \\ \text{cell } j \text{ of net } i}}\left(R_j^{cell}\right)$.

From this we obtain the following net weights for net i:

$$nw_i^{lateral} = 1 + \alpha_{TEMP} R_i^{net} s_i^{wl} \text{ and } nw_i^{vertical} = 1 + \frac{\alpha_{TEMP} R_i^{net} s_i^{ilv}}{\alpha_{ILV}} \quad (8)$$

where $nw_i^{lateral}$ is the net weight in the x and y directions, for WL_i, and $nw_i^{vertical}$ is the net weight in the z direction, for ILV_i.

3.2 Thermal Resistance Reduction Nets

Better thermal results can be obtained when higher powered cells are placed in areas with lower thermal resistances to the ambient. During placement, this can be encouraged by adding nets that pull each cell toward the heat sink and weighted based on the power dissipation of the cell. These nets are called *thermal resistance reduction nets* and are weighted according to the power usage of the cell and the slope of the thermal resistance profile of the chip. As the thermal resistance slope increases, high powered cells are more strongly attracted to the heat sink where temperatures and thermal resistances would be lower.

Because vertical (z) distances are much shorter and heat sinking is primarily in the z direction, the thermal resistance increases principally with vertical distance away the heat sink. As such, the thermal resistance from cell j to the ambient, R_j^{cell}, can be approximated with $R_j^{cell} \approx R_0^z + R_{slope}^z d_j^z$ where R_0^z is the thermal resistance at the bottom of the chip, R_{slope}^z is the slope the thermal resistance in the z direction, and d_j^z is the distance of the cell from the bottom of the chip. Because R_0^z is constant with respect to d_j^z, it is dropped from the thermal component of the objective function to give

$$\sum_{each\ cell\ j} \left[\alpha_{TEMP} P_j^{cell} R_{slope}^z d_j^z \right] \tag{9}$$

where $P_j^{cell} = \sum_{\substack{each\ driven \\ net\ i\ of\ cell\ j}} \left[s_i^{wl} WL_i + s_i^{ilv} ILV_i + s_i^{input\ pins} \right]$ (10)

and $s_i^{input\ pins} = \frac{\frac{1}{2} f V_{DD}^{\ 2} a_i C_{per\ pin} n_i^{input\ pins}}{n_i^{output\ pins}}$. (11)

Therefore for each cell, a thermal resistance reduction net is added to the netlist, connected at one end to the cell and the other end to the bottom of the chip, and given a net weight of

$$nw_j^{cell} = \alpha_{TEMP} P_j^{cell} R_{slope}^z \tag{12}$$

where nw_j^{cell} is the net weight of cell j to the heat sink. If necessary, thermal resistance reduction nets could be added for other directions, and leakage power could be added to P_j^{cell}. It should also be noted that the thermal resistances used by the net weights in Section 3.1 are calculated using all three dimensions.

In Equation (10), P_j^{cell} depends on wirelength and interlayer via counts of its driven nets. However at the beginning of global placement, cells are placed at the center of the chip and consequently the wirelengths and interlayer via counts are zero. This would cause the power contributions from the wirelength and interlayer via counts to be neglected in determining nw_j^{cell}. Some minimum values for the wirelength and interlayer via counts should be used instead, and these values can be determined by minimizing the objective function for each net in question. The derivation of these minimum values is similar to the PEKO (Placement Example with Known Optimal wirelength) formulation presented in [12], but is extended to 3D ICs. PEKO benchmarks were created to have known optimal wirelengths for 2D ICs. With 3D ICs, the approximate optimal values are given by (a detailed derivation is omitted due to space limitations):

$$WL_i^{x\ opt} = \sqrt[3]{\alpha_{ILV} w_i^{ave} h_i^{ave} n_i^{total\ pins}} - w_i^{ave} \tag{13}$$

$$WL_i^{y\ opt} = \sqrt[3]{\alpha_{ILV} w_i^{ave} h_i^{ave} n_i^{total\ pins}} - h_i^{ave} \tag{14}$$

$$ILV_i^{opt} = \frac{\sqrt[3]{\alpha_{ILV} w_i^{ave} h_i^{ave} n_i^{total\ pins}}}{\alpha_{ILV}} - 1 \tag{15}$$

where $WL_i^{x\ opt}$ is the approximate optimal wirelength in the x direction, $WL_i^{y\ opt}$ is the approximate optimal wirelength in the y direction, and ILV_i^{opt} is the approximate optimal interlayer via count for net i. In calculating P_j^{cell} for nw_j^{cell}, if WL_i^x, WL_i^y, or ILV_i fall below its optimal value, $WL_i^{x\ opt}$, $WL_i^{y\ opt}$, or ILV_i^{opt}, then the optimal value is used instead.

4. Coarse Legalization

Next, coarse legalization is used to bridge the gap between global placement and detailed legalization. Placements produced after coarse legalization still contains overlaps, but the cells are evenly distributed over the placement area so that the computationally intensive localized calculations used in detailed legalization are prevented from acting over excessively large areas. Our coarse legalization method utilizes a spreading mechanism called *cell shifting* (Section 4.1) to spread cells globally, and mechanisms to reduce the objective function value by *moving and swapping cells* both locally and globally (Section 4.2). The density profile disruptions caused by the moves and swaps are balanced with cell spreading provided by *cell shifting*. These methods use a coarse density mesh with bins equal to two cell widths, two cell heights, and one layer thickness. The coarse legalization methods can also be used in conjunction with detailed legalization to iteratively improve an existing placement during a post-optimization phase of detailed placement if desired.

4.1 Cell Shifting

In *cell shifting*, a mesh of density bins is created for the entire chip, densities are calculated as the ratio of cell area in each bin to bin area, bin boundaries are shifted based on these densities, and cells are moved according to the new bin boundaries. The process is repeated until an even distribution of cells is produced. A *cell shifting* procedure was developed to overcome limitations discovered with a similar method used by *FastPlace* [13]. The resulting method is extreme effective in rapidly producing an even distribution of cells from placements with highly uneven distributions, while at the same time, minimizing perturbations and degradations in quality.

Our method addresses two problems that prevent *FastPlace*'s *cell shifting* method from being applied more generally. First, bin boundaries can cross-over and become out of order with *FastPlace* because only the densities of adjacent bins are considered in calculating new bin boundaries, and it does not take into account how other bin boundaries are being moved. When cross-over occurs, the relative cell ordering changes as cells are mapped to the new bin boundaries, and the placement quality can degrade. Our method addresses this issue by taking all bin boundaries within a row into consideration when calculating new bin boundaries. Second, because only two adjacent bins are considered at a time, *FastPlace* continues to spread cells apart in areas that are already nearly legalized, even when this would not help reduce cell congestion in other over-congested areas. To determine whether moving the bin boundaries of nearly legalized bins will help elsewhere within the placement area, our method considers the densities of all bins within the same row rather than just two adjacent bins. The bin boundaries of sparsely populated bins are adjusted only to allow over-congested bins within the same row to expand.

For the entire three-dimensional mesh of density bins, rows of bins are shifted one at a time for each direction. An example of bin shifting within a row of bins in the x direction is shown in Figure 1. In this figure, $d_{i,j,k}$ is the density of bin (i,j,k), $B^{x}_{i,j,k}$ is the new boundary between bin $(i\text{-}1,j,k)$ and bin (i,j,k), and $B^{x}_{i+1,j,k}$ is the new boundary between bin (i,j,k) and bin $(i\text{+}1,j,k)$.

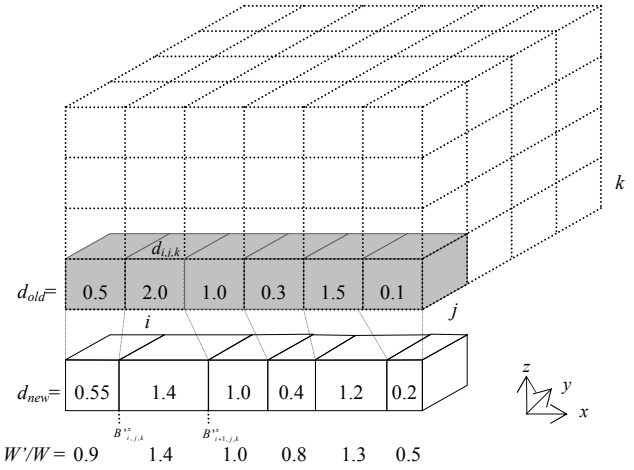

Figure 1. Row of cells in the x direction.

The relationship between bin width, W, and density is shown in Figure 2. In this graph, W'/W is the ratio of the new bin width to the old bin width, d is the original bin density, a^{lower} is the slope of the curve for densities less than one, and a^{upper} is the maximum slope of the curve for densities greater than one. The parameters a^{lower}, a^{upper}, and b are adjusted so that expansions of over-congested bins are balanced with the contractions of sparse bins.

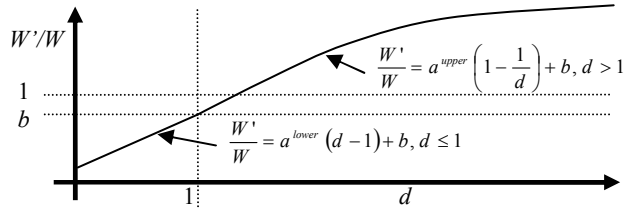

Figure 2. Cell shifting bin width versus density.

For each row of bins oriented in the x direction with a y position of j and a z position of k, new bin boundaries, $B'^{x}_{i,j,k}$, are calculated with the following equation using the density, $d_{i,j,k}$, of bins in the row, the specific a^{lower}, a^{upper}, and b values determined for this row of bins, and the width, W_x, of the bins.

$$B'^{x}_{i,j,k} = \begin{cases} W_x\left(a^{upper_x}_{j,k}\left(1-\frac{1}{d_{i-1,j,k}}\right)+b^{x}_{j,k}\right)+B'^{x}_{i-1,j,k} & \text{if } d_{i-1,j,k}>1 \\ W_x\left(a^{lower_x}_{j,k}\left(d_{i-1,j,k}-1\right)+b^{x}_{j,k}\right)+B'^{x}_{i-1,j,k} & \text{if } d_{i-1,j,k}\le 1 \end{cases} \quad (16)$$

New bin boundaries in other directions are calculated similarly.

For mapping the x coordinate of cell p, x_p, in bin (i,j,k) to the new bin boundaries, the same formula is used as in *FastPlace* except we apply a different cell movement retention parameter, β_p^x, for each cell.

$$x'_p = \beta^x_p\left[\frac{W'_x}{W_x}\left(x_p - B^x_{i,j,k}\right)+B'^{x}_{i,j,k}\right]+\left(1-\beta^x_p\right)x_p \quad (17)$$

where W_x' is the new bin width and x'_p is the new position of cell p after movement retention is applied. In Equation (17), β_p^x is used to slow down the move to the new position, is between zero and one, and is dynamically adjusted for every cell to minimize degradation in the objective function.

4.2 Moves and Swaps

During coarse legalization, cells are moved to positions both locally and globally in order to reduce the value of the objective function. Besides simply moving a cell to a new position in the target region, swapping positions with cells in the target region is also considered. Moves are only considered if there is enough space available in the target region with cells shifted aside, if necessary, to make room and their effect on the objective function value included in the cost of the move. From these possible moves and swaps, the one producing the largest reduction in the objective function is executed for each cell. If no swap or move is found to reduce the objective function value, then the cell remains in its current position. In the local move/swap procedure, the target region consists of only the adjacent bins.

The global move/swap procedure performs moves/swaps globally to a target region around the objective function minimum for each cell. This target region is similar to the optimal region idea presented in [14], but modified to include three dimensional considerations and net weights. An optimal region for a cell is the area in which the cell should be placed in order to achieve the largest possible reduction in the objective function value assuming all other cells remain in their current position. Taking into account net weights and α_{ILV}, we define the target region as being inside an isosurface of the objective function around the optimal region and containing a fixed number of bins.

5. Detailed legalization

Detailed legalization puts cells into the nearest available space that produces the least degradation in the objective function. Our legalization procedure assumes that the cell distribution has already been evened with coarse legalization and tries to move cells only locally. A much finer density mesh is created for the detailed legalization process than what was used with course legalization and consists of bins similar in size to the average cell. Bin densities are calculated in a more fine-grained fashion by dividing the precise amount of cell width (rather than area) in the bin by the bin width. To ensure that densities are precisely balanced between different halves of the placement, the amount of space available or lack of space available is calculated for each side of the dividing planes formed by the bin boundaries. A directed acyclic graph (DAG) is constructed in which directed edges are created from bins having an excess amount of cell area to adjacent bins that can accept additional cell area. From this, the dependencies on the processing order of bins can be derived and used to determine the order in which cells are to be placed into their final position. In addition, an estimate of the objective function's sensitive to cell movement is also used in determining the cell processing order. Using this processing order, the algorithm looks for the best available position for each cell within a target region around its original position. The objective function is used to determine which available position in the target region produces the best results. If an available position is not found, the target region is gradually expanded until enough free space is found within the row segments that it contains. If already-processed cells need to be moved apart to legally place the cell, the effect of their movement on the objective function is included in the cost for placing the cell in that position.

6. Implementation

The placement method begins by the adding the *thermal resistance reduction nets* (Section 3.2) to the netlist and placing the cells at the center of the chip. *Global placement* is performed as described in Section 3 using partitioning. As the placement is recursively partitioned, the positions are refined for terminal propagation and the *thermal net weights* (Section 3.1) are

updated. After *global placement*, *global swaps and moves* are performed followed by *local swaps and moves* as described in Section 4.2. Next, the *cell shifting* method from Section 4.1 is performed iteratively until the maximum bin density is less than a desired value close to one in order to guarantee an even density distribution for the *detailed legalization*. Finally, *detailed legalization* (Section 5) is used to produce a completely legal placement. The *course* and *detailed legalization* steps can be repeated multiple times if additional optimization is required.

7. Results

The 3D placement method was implemented in C++ with *hMetis* [15] used for partitioning and run on a Linux workstation with a Pentium 4 3.2GHz CPU and 2GB memory. Benchmark circuits, as shown in Table 1, from the IBM-PLACE suite [16] were used in these experiments. Vertical dimensions and the effective thermal conductivity were derived using the design specifications of MIT Lincoln Labs' 0.18µm 3D FD-SOI technology [17] [18], and capacitance values were derived from [19]. Temperature results were calculated using Finite Element Analysis (FEA) [2] with the bottom of the chip (heat sink) given convective boundary conditions. The parameters used in these experiments are shown in Table 2.

Table 1. Benchmark Circuits

name	cells	area (mm^2)	name	cells	area (mm^2)	name	cells	area (mm^2)
ibm01	12282	0.060	ibm07	45135	0.197	ibm13	81508	0.326
ibm02	19321	0.086	ibm08	50977	0.214	ibm14	146009	0.680
ibm03	22207	0.090	ibm09	51746	0.221	ibm15	158244	0.634
ibm04	26633	0.122	ibm10	67692	0.377	ibm16	182137	0.892
ibm05	29347	0.150	ibm11	68525	0.287	ibm17	183102	1.040
ibm06	32185	0.117	ibm12	69663	0.415	ibm18	210323	0.988

Table 2. Parameters

technode	100nm	whitespace	5%
number of layers	4	inter-row/row space	25%
bulk substrate thick.	500µm	lateral interconnect cap.	73.8pF/m
layer thickness	5.7µm	interlayer via cap.	1480pF/m
interlayer thickness	0.7µm	input pin capacitance	0.350fF
effective thermal conductivity	10.2 W/mK	ambient temperature	0°C
		conv. coef. of heat sink	10^6 W/m^2K

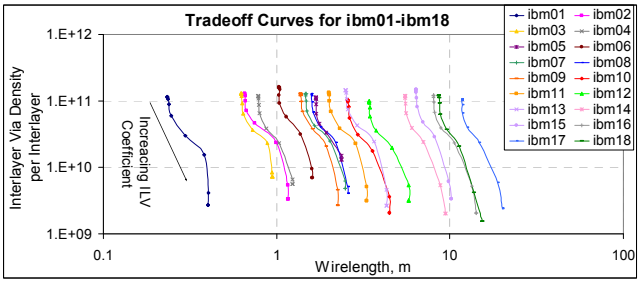

Figure 3. Tradeoff between wirelength and interlayer via count.

In the first set of experiments, only the tradeoff between wirelength and interlayer via counts was explored by setting the thermal coefficient, α_{TEMP}, to zero and varying the interlayer via coefficient, α_{ILV}, from 5×10^{-9} to 5.2×10^{-3}. This range of values for α_{ILV} is centered around the average cell width or height ($\sim10^{-5}$), but the span of the range was empirically determined. In Figure 3, complete tradeoff curves between interlayer via density and wirelength are shown for the benchmarks from Table 1. Interlayer via counts decrease and wirelengths increase as the interlayer via coefficient is increased. Figure 4 shows the average interlayer via densities and percent change in the wirelength for the benchmark circuits. Wirelength reductions within 2% of the maximum can be achieved using 46% fewer interlayer vias.

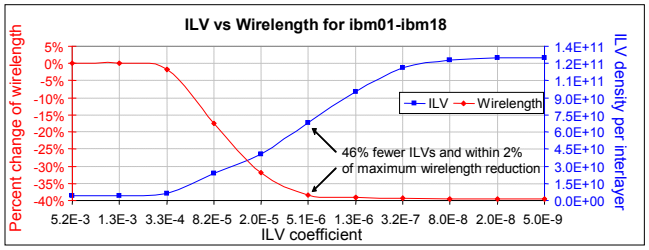

Figure 4. Average wirelength vs. ILV tradeoff for ibm01-ibm18.

In Figure 5, the number of layers was increased from one to ten for the ibm01, and the resulting tradeoff curves between wirelength and interlayer via counts were plotting. As the number of layers is increased, the tradeoff curves are shifted to shorter wirelengths, and more wirelength reduction can be obtained.

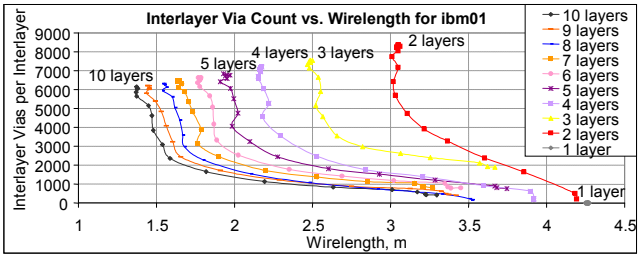

Figure 5. Tradeoff curves for ibm01 with increasing number of layers.

The effect of thermal placement on ibm01 was examined by varying α_{ILV} from 5×10^{-8} to 1.6×10^{-3} and α_{TEMP} from 1×10^{-8} to 1.3×10^{-3}. The effect on the average temperature is shown in Figure 6, and the effect on the tradeoff curve between wirelength and interlayer via counts is shown in Figure 7. Temperatures are reduced as the thermal coefficient is increased. In addition, temperature and power increase as the interlayer via coefficient decreases because of increasing capacitances from interlayer vias. As the thermal coefficient is increased and temperatures are reduced, the tradeoff curves are degraded and moved to the right toward higher wirelengths and interlayer via counts.

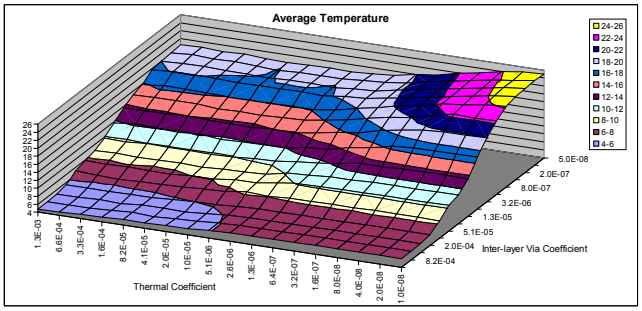

Figure 6. Average temperature of ibm01 as the thermal and interlayer via coefficients are varied.

In Figure 8, the percent reduction in the average temperatures is shown for ibm01 as the number of layers is increased from one to eight while varying the thermal coefficient and setting the interlayer via coefficient to 1×10^{-5}. As can be

seen, our method is effective in reducing temperatures for 2D ICs (1 layer) as well as 3D ICs with many layers.

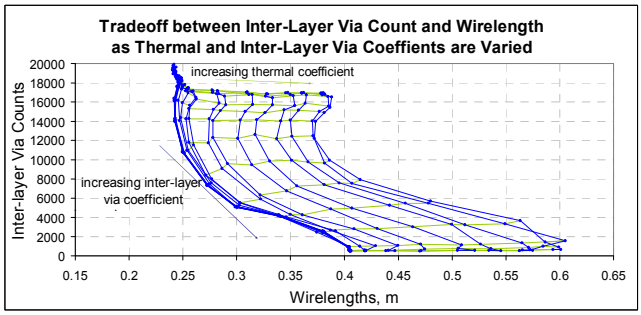

Figure 7. Tradeoff curves for ibm01 as the thermal and interlayer via coefficients are varied.

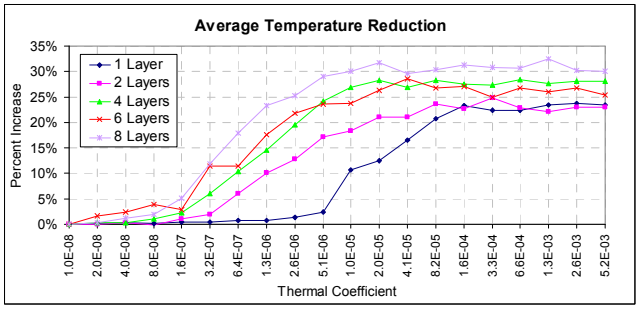

Figure 8. The percent reduction in the average temperature of ibm01 as the number of layers is increased.

With the interlayer via coefficient set to 1×10^{-5}, the average percent change in the interlayer via counts, wirelength, power, and temperatures is shown in Figure 9 as the thermal coefficient is varied from 0 to 4.1×10^{-5}. When the average temperatures are reduced by 19%, wirelengths are increased by only 1%.

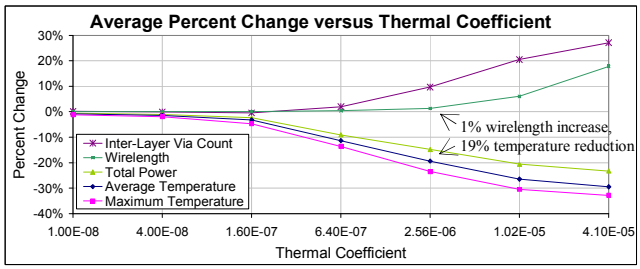

Figure 9. Average percent change for ibm01 to ibm18 as the thermal coefficients are varied.

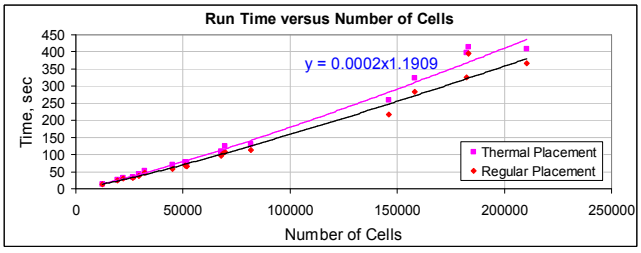

Figure 10. Runtime analysis of the thermal placement method.

In Figure 10, the run time analysis of our method with and without thermal considerations shows that it is nearly linear in run-time efficiency. Other experiments show that by increasing the number of random starts used by *hMetis* and expanding target region sizes used by the move/swap procedures, a 3.8% improvement in the objective function can be made at a cost of 3.4 times slower runtimes. Also, if the coarse and detailed legalization procedures are repeated ten times, a 7.7% improvement can be made but with 65 times longer runtime.

8. Conclusion

An efficient and effective thermal placement method was developed for 3D ICs that allows the tradeoff between wirelength, interlayer via count, and temperature to be explored. Limitations on interlayer vias densities imposed by fabrication make it an important consideration in the design of placement tools for 3D ICs. Our method fully exploits the tradeoff that exists between wirelength and interlayer via counts, and can allow wirelengths to be minimized for any desired interlayer via density. With regard to thermal mitigation, our method takes power usage into account so that both temperatures and power are minimized. This is achieved in global placement by using net weighting to reduce the length of nets with high-power usage and high thermal resistances at the driver cells. Additional nets are added to move cells toward lower thermal resistances. In detailed placement, the objective function is used in determining the cost for moving cell so that degradations in quality are minimized. Our thermal placement method is effective not only with 3D ICs, but also with 2D ICs, and the run time efficiency was shown to be nearly linear with increasing circuit sizes.

9. References

[1] T. Tanprasert, "An Analytical 3-D Placement that Reserves Routing Space," *ISCAS* '00, 69-72.

[2] B. Goplen and S. S. Sapatnekar, "Efficient Thermal Placement of Standard Cells in 3D ICs using a Force Directed Approach," *ICCAD* '03, 86-89.

[3] I. Kaya, M. Olbrich, and E. Barke, "3-D Placement Considering Vertical Interconnects," *Proc. IEEE Int. SOC Conf.* '03, 257- 258.

[4] R. Hentschke, G. Flach, F. Pinto, and R. Reis, "Quadratic Placement for 3D Circuits Using Z-Cell Shifting, 3D Iterative Refinement and Simulated Annealing," *Proc. Symp. on Integrated Circuits and Syst. Des.* '06, 220-225.

[5] Y. Deng and W. Maly, "Interconnect Characteristics of 2.5-D System Integration Scheme," *ISPD* '01, 171-175.

[6] S. Das, A. Chandrakasan, and R. Reif, "Design Tools for 3-D Integrated Circuits," *ASP-DAC* '03, 53-56.

[7] C. N. Chu and D. F. Wong, "A Matrix Synthesis Approach to Thermal Placement," *ISPD* '97, 163-168.

[8] B. Obermeier and F. M. Johannes, "Temperature-Aware Global Placement," *ASP-DAC* '04, 143-148.

[9] C. H. Tsai and S. M. Kang, "Cell-Level Placement for Improving Substrate Thermal Distribution," *TCAD*, 2000, 19(2), 253-266.

[10] G. Chen and S. S. Sapatnekar, "Partition-Driven Standard Cell Thermal Placement," *ISPD* '03, pp. 75-80.

[11] A. E. Dunlop and B. W. Kernighan, "A Procedure for Placement of Standard Cell VLSI Circuits," *TCAD*, 1985, 4(1), 92-98.

[12] C.-C. Chang, J. Cong, M. Romesis, and M. Xie, "Optimality and Scalability Study of Existing Placement Algorithms," *TCAD*, 2004, 23(4), 537-549.

[13] N. Viswanathan and C. Chu, "FastPlace: Efficient Analytical Placement using Cell Shifting, Iterative Local Refinement and a Hybrid Net Model," *ISPD* '04, 26-33.

[14] M. Pan, N. Viswanathan, and C. Chu, "An Efficient and Effective Detailed Placement Algorithm," *ICCAD* '05, 48-55.

[15] G. Karypis, R. Aggarwal, V. Kumar, and S. Shekhar, "Multilevel Hypergraph Partitioning: Applications in VLSI Domain," *IEEE Trans. on VLSI Syst.*, 1999, 7(1), 69-79.

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

[17] J. Burns, L. McIlrath, C. Keast, C. Lewis, A. Loomis, K. Warner, and P. Wyatt, "Three-Dimensional Integrated Circuits for Low-Power, High-Bandwidth Systems on a Chip," *ISSCC Digest of Technical Papers*, 2001, 268-269.

[18] K. Warner, J. Burns, C. Keast, R. Kunz, D. Lennon, A. Loomis, W. Mowers, and D. Yost, "Low-Temperature Oxide-Bonded Three-Dimensional Integrated Circuits," *IEEE International SOI Conference Proceedings*, 2002, 123-124.

[19] J. Cong, "Challenges and Opportunities for Design Innovations in Nanometer Technologies," *Invited SRC Design Sciences Concept Paper*, 1998, 1-15.