Physical Layout Automation for System-On-Packages

Ramprasad Ravichandran, Jacob Minz, Mohit Pathak, Siddharth Easwar, and Sung Kyu Lim
School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA 30332
{raam, jrminz, mohitp, seaswar, limsk}@ece.gatech.edu

Abstract

System-On-Package (SOP) technology provides a capability to integrate both mixed-signal active components and passive components all into a single high speed/density three dimensional packaging substrate. The physical layout resource of SOP is multi-layer in nature, where all layers are used for both placement and routing unlike the traditional multi-layer PCB or MCM packaging. In this paper, we present the first 3D physical design algorithms targeting SOP technology. 3D partitioning divides the input design into multiple layers. 3D placement determines the location of the active and passive components in multi-layer packaging substrate while considering various signal integrity issues. 3D global routing performs the following major steps: pin/net distribution, layer assignment, tree generation, and channel/pin assignment. Our experimental results demonstrate the effectiveness of our approaches.

1. Introduction

The true potential of SOP technology [1] lies in its capability to integrate both active components such as digital IC, analog ICs, memory modules, MEMS, and opto-electronic modules, and passive components such as capacitors, resistors, and inductors all into a single high speed/density multi-layer packaging substrate. Since both the active and passive components are integrated into the multi-layer substrate, SOP offers a highly advanced three-dimensional mixed-signal system integration environment. Three-dimensional SOP packaging offers significant performance benefits over the traditional two-dimensional packaging such as PCB and MCM due to the electrical and mechanical properties arising from the new geometrical arrangement as illustrated in Figure 1. Thus, innovative ideas in the development of CAD tools for multi-layer SOP technology is crucial to fully exploit the potential of this new emerging technology.

A high performance mixed signal system employs a lot of passive components—up to 30 passive components per an IC. For example, Sony Handy Cam DCR-PC7 has 43 ICs and 1329 passive elements. Such passive components continue to take up much circuit board real estate. Therefore, rigorous attempts have been made to replaces them with so-called embedded passive components (EPC), which are small and flat enough to be inserted between package layers. EPCs allow devices to get smaller or designers to fit more functionality in the same space; eliminate the costs currently needed to purchase and solder on discrete devices; allow for more design flexibility; and derive electrical benefits from the different current path that would be traveled. EPCs can also be used for simultaneous switching noise reduction, cross talk reduction, network matching, and signal integrity.

The physical layout resource environment of SOP is multi-layer in nature—the top layer is mainly used to accommodate active components, the middle layers are mainly for passive components, and the I/O pins are located at the bottom of the SOP package. Therefore, all layers are used for both placement and routing unlike PCB or MCM. Therefore, the existing design tools for PCB or MCM packaging [7,8,9] can not be used directly for the design of SOP. Existing works on 3D placement are very few [2,10,11,12,13] and focus only on cell placement instead of block placement. In this paper, we present the first 3D physical design algorithms targeting SOP technology. 3D partitioning [15,17] divides the input design into multiple layers. 3D placement [18] determines the location of the active and passive components in multi-layer packaging substrate while considering various signal integrity issues. 3D global routing [14,16] performs the following major steps: pin/net distribution, layer assignment, tree generation, and channel/pin assignment. Our experimental results demonstrate the effectiveness of our approaches.

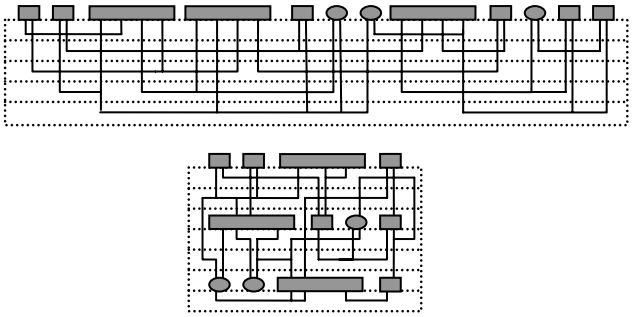

Figure 1. Single device layer (PCB,MCM) vs multiple device layer (SOP) packaging. The wirelength is shorter in SOP packaging.

2. Preliminaries

2.1. Overview of the Approach

An overview of the 3D layout automation process is shown in Figure 2. During the 3D partitioning step, we divide the input design into multiple layers. Under our wire-centric design paradigm, partitioning is seen as the crucial step that defines the local and global wires—intra-partition connections become intra-layer wires, whereas inter-partition connections become inter-layer wires. The objective is to minimize the amount of wires, the longest path delay, and power consumption induced by the partitioning under pin, area, and thermal constraints. During the 3D placement step, I/O pin location and chip/component dimension are determined. In addition, the placement of these mixed signal components needs to be performed. The embedded passive elements are placed in the intermediate layers within the SOP substrate while other active components are mounted on the surface. The main objective during this step is to minimize the area of the final chip, total estimated wirelength, and the longest path delay measured in a multi-layer environment. The major constraints to be considered are routability and power-ground

noise. In addition, some analog ICs and passive elements need to be placed together in order to maintain high Q, and analog and digital ICs will need to be separated apart due to various noise issues. During the 3D routing step, global and detailed routing are performed. Unlike the traditional multi-layer routing, pins are possibly located at all intermediate layers in SOP substrate rather than top-layer only since chips now have connections to embedded passives. In addition, these embedded passives are obstacles during routing. Our goal is to perform global routing and layer assignment simultaneously for more rigorous performance optimization in a shorter runtime. The objective is to minimize the total wiring cost in terms of the number of layers and vias and the longest path delay. The major constraints to be considered are congestion and cross-talk noise.

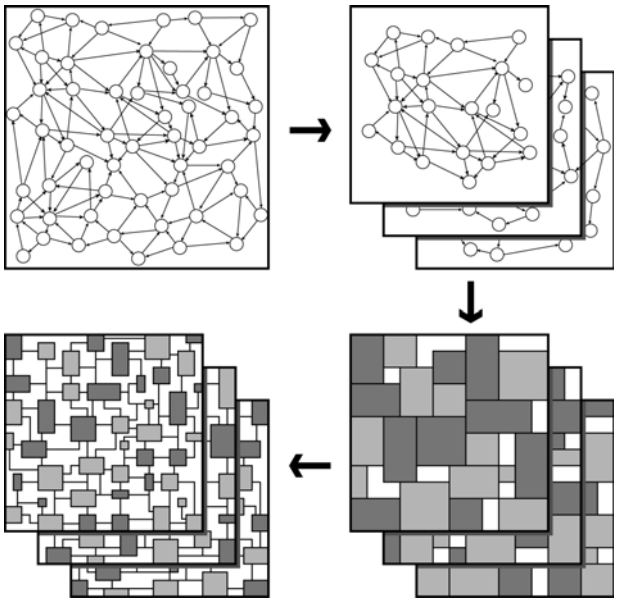

Figure 2. Overview of 3D layout automation. First, the input design is first partitioned into several layers (3D partitioning). Second, the active and passive elements are placed in each layer (3D placement). Third, routing is done to finish the interconnection (3D routing). Inter-layer connections are not shown for simplicity.

2.2. Problem Formulation

3D layer structure: the layer structure in SOP is different from PCB or MCM—it has multiple floorplan layers and routing layers. It has one *I/O pin layer* through which various components can be connected to the external pins. The *floorplan layers* contain the blocks, which from the point of view of physical design is just a geometrical object with pins. In some cases where these blocks are a collection of cells, the pins may not be assigned and pin assignment needs to be done to determine their exact location. The interval between two floorplan layers is called the *routing interval*. The routing interval contains a stack of *signal routing layers* sandwiched between *pin distribution layers*. These layers are actually X-Y routing layer pairs, so that the rectilinear partial net topologies may be assigned to it. We also allow routing to be done in the pin distribution layers. Our 3D package layout automation is divided into 3D partitioning, 3D placement, and 3D routing.

3D Partitioning: Given a netlist $NL(C, N)$, where C is the set of cells representing active and passive elements, and N is the set of nets connecting the cells, the purpose of the 3D SOP partitioning problem is to assign cells in NL to K layers while area constraint for each layer is satisfied. The objective is to minimize the amount of inter-layer interconnect, critical path delay, and total power consumption.

3D Placement: A multi-layer SOP floorplan consists of a set $B=\{b_i|\ 0\leq i< n\}$ of n blocks and a set $L=\{l_i|\ 0\leq i<k\}$ of k layers. A block is either an active component or embedded passive component (EPC). We assume rectangular shape for all these blocks. Each floorplan f_i has a set of blocks B_i, which is a non-empty proper subset of B. A SOP floorplan F is represented by a set $F=\{f_0, f_1,\ldots, f_{k-1}\}$, where a floorplan f_i is a 2-dimensional placement of blocks in B_i. A SOP floorplan F is *feasible* if (i) F is free of overlap among block location, (ii) F satisfies the layer and geometric constraints specified by the user. The goal is to minimize the final footprint area of SOP package, wirelength, total number of vias, and constraint violation.

3D Routing: Given a set of floorplans $F=\{f_1,f_2,\ldots,f_k\}$, netlist $N=\{n_1,n_2,\ldots,n_n\}$, and the routing resource graph, generate the routing topology $T(n)$ for each net n, assign n to a set of routing layers and assign all pins of n to legal locations. All conflicting nets are assigned to different routing layers while satisfying various capacity constraints. The objective is to minimize the total number of routing layers used, wirelength, and crosstalk. In the SOP model the nets are classified into two categories. The nets which have all their terminals in the same placement are called *i-nets*, while the ones having terminal in different placements will be referred to as *x-nets*. The i-nets can be routed in the single routing interval or indeed within the placement layer itself. However, for high performance designs routing such nets in the routing interval immediately above or below the placement layer maybe desirable and even required. On the other hand, the x-nets may span more than one routing intervals.

3. 3D Partitioning

We have recently developed the first 3D partitioning algorithm for SOP [15,17]. Our algorithm named GEO-PD performs simultaneous delay and power optimization and provides smooth cutsize, wirelength, power and delay tradeoff. An illustration is shown in Figure 3. In GEO-PD, we use retiming based timing analysis and visible power analysis to identify timing and power critical nets and assign proper weights to them to guide the multi-level optimization process. In general, timing and power analysis are done at the original netlist while a recursive multi-level approach performs partitioning on the sub-netlist as well as its coarsened representations. We develop an effective way to translate the timing and power analysis results from the original netlist to a coarsened sub-netlist for effective multi-level delay and power optimization. GEO-PD is a multi-level partitioner for simultaneous delay and power optimization. GEO-PD partitions the given netlist NL into K partitions using a top-down recursive bipartitioning approach. GEO-PD consists of two subroutines: GEO-PD-2way recursively bipartitions NL, whereas GEO-PD-Kway refines these partitioning results occasionally. GEO-PD-2way is performed on the sub-netlist, whereas GEO-PD-Kway is performed on the entire netlist.

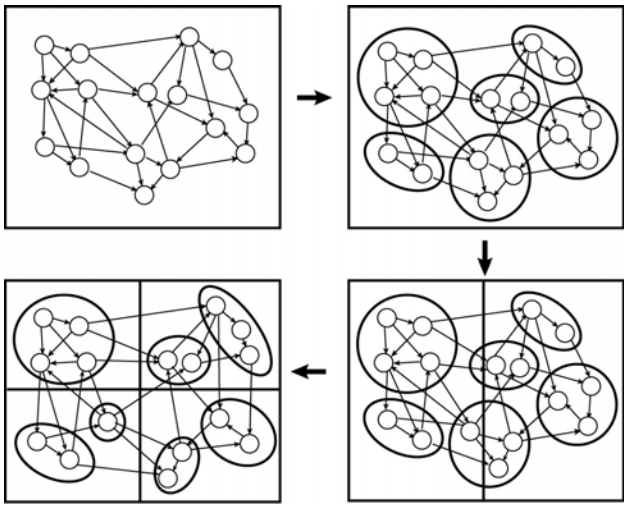

Figure 3. Illustration of 3D multi-level recursive bipartitioning. The netlist is clustered first, followed by multi-level partitioning to refine the cutline. Multi-level clustering is done each time we divide smaller partitions.

GEO-PD-2way first generates the sub-netlist from the given partition tree node and performs multi-level clustering on it. We use our ESC clustering algorithm [19] for this purpose. Then we obtain a random initial partitioning B among the clusters at the top level of the hierarchy. The subsequent top-down multi-level refinement is used to improve B in terms of delay and power. We perform retiming based timing analysis RTA [20] to identify timing critical nets. We also perform power analysis to identify power critical nets. Then we compute the delay and power weights for the nets in the sub-netlist for simultaneous delay and power optimization. The subsequent iterative improvement through cluster move tries to minimize the weighted cutsize. Finally we project the current solution to the next level coarser netlist for multi-level optimization. GEO-PD-Kway refinement is performed when we obtain 2^j partitions ($j > 1$) from GEO-PD-2way (4, 8, 16 partitions, etc). We first perform a restricted multi-level clustering, where grouping among cells in different partition is prohibited. This allows the partitioner to preserve the initial partitioning results. Then we again perform multi-level partitioning in the same way as in GEO-PD-2way for additional delay and power improvement. GEO-PD-Kway is applied onto the global netlist for more global level optimization.

For simultaneous delay and power optimization, we first identify timing and power critical nets and assign proper weights to them to guide the optimization process. A net is *timing critical* if it lies along a critical path and *power critical* if it has high fanout with large wirelength and is driven by a gate with high switching activity. In GEO-PD, retiming delay and visible power are minimized through retiming based timing analysis and visible power analysis. We use *sequential slack* to compute how much time slack exists before timing violation occurs after retiming. These values are then used to compute the delay weights of the nets for retiming delay minimization. In case of power optimization, we use switching activity and gate/wire capacitance to compute power weights of the nets for visible power minimization. Both delay and power weights are added together, and GEO-PD performs multi-level partitioning to minimize the total weighted cutsize. We note that the multi-level approach is very effective in minimizing the weighted cutsize and wirelength. However, timing and power analysis is typically done at the original netlist while a recursive multi-level approach performs partitioning on the sub-netlist as well as its coarsened representations. Thus, it is crucial that we have an effective way to translate the timing and power analysis results from the original netlist to a coarsened sub-netlist. Interested readers are referred to [15,17] for more details.

4. 3D Placement

4.1. Simulated Annealing

We have recently developed the first 3D placement algorithm for SOP [18]. We observe from related experiments that adding the following components to the cost function $C(F)$ results in a more compact multi-layer placement: *total flatten area flat*(F) and *dimension deviation dev*(F). *flat*(F) is the sum of all placements, $flat(F) = \sum a(f_i)$. The minimization of this objective results in a highly compact placement for each layer. *dev*(F) measures how much the upper right corner (URC) of a placement deviates from the average URC. We compute the average URC (u_x, u_y) by $u_x = \Sigma u_x(f_i)/k$, where $u_x(f_i)$ denotes the x-coordinate of the URC of a placement f_i. We compute $u_y(f_i)$ using y-coordinates instead. Let $d(f_i) = |u_x - u_x(f_i)| + |u_y - u_y(f_i)|$ be the dimension deviation of a placement of f_i. Then *dev*(F), the dimension deviation of SOP placement F is simply the sum of all $d(f_i)$. The minimization of this objective results in a more dimension-balanced placement among all layers. It may seem redundant to have all three area-related objectives *area*(F), *flat*(F), and *dev*(F) in $C(F)$. However, our related experiments indicate that each of these three objectives contribute to the minimization of not only the final footprint area *area*(F) but also the wirelength estimation *wire*(F).

Among many proposed methods to represent 2-dimensional placement, we extend the sequence pair (SP) [4] to represent the multi-layer SOP placement solution. Our multi-layer sequence pair is represented by ($SP_0|SP_1|...|SP_{k-1}$), where SP_i contains the positive and negative sequence for the blocks contained in layer i. For a faster area evaluation for a given multi-layer SP, we use longest common subsequences (LCS) [3] method. Authors in [4] propose three types of moves for solution perturbation during Simulated Annealing: M1 (swap two modules in positive sequence), M2 (swap two modules from both positive and negative sequence), and M3 (rotate). We add two moves M4 and M5 to search the solution of multi-layer placement effectively: M4 is similar to M1, except that the two blocks are from positive sequences in different layers. M5 selects a block from layer i and moves it to another layer j. The location in positive and negative sequence from SP_j is again randomly chosen.

4.2. SOP Geometric Constraints

We categorize the geometric constraints among active and passive components into the following 6 types: (i) *noise*: decoupling capacitors are placed nearby I/Os or active components, (ii) *thermal*: some active/passive components are placed in certain layers, (iii) *power*: digital and analog ICs are placed in different voltage islands, (iv) *timing*: blocks from a critical path are placed closer, (v) *interface*: I/O blocks are

placed near the bottom layer, (vi) *cluster*: functionally dependant blocks are placed close together.

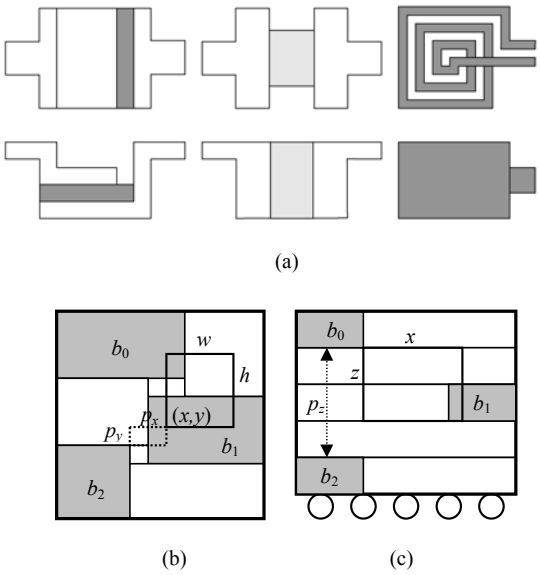

Figure 4. (a) Embedded passive components (R, L, and C). Top and side views of typical RLC shapes are shown, (b) region constraint r_1=[{b_0,b_4}|(x,y,w,h)] and r_2=[{b_3,b_4}|(x,y,w,h)]. r_1 is satisfied and r_2 has penalty of p_x+p_y. (c) group constraint g_1=[{b_0,b_2}|(x,y,z)] and g_2=[{b_0,b_7}|(x,y,z)]. g_1 is satisfied and g_2 has penalty of p_z. y-dimension is not shown.

Most of the active components are required to be placed on the top layer due to heat dissipation requirement. However, some active components that do not generate too much heat can be placed in the middle layers. EPCs can be placed at any layers, but using middle layers is the most beneficial in reducing the overall footprint area of SOP. However, some EPCs are required to be placed on the top layer due to thermal and/or noise issues. Third, some active components need to be placed nearby together or apart from each other due to several reasons including signal/power integrity, performance optimization, etc. Lastly, most EPCs need to be placed closer to the related active components. Handling the layer constraints is straightforward, but the geometric constraints are harder to satisfy. Figure 4(a) shows geometric shapes of EPCs.

Table 1 describes the 6 geometric constraint types we consider in SOP placement. The point, layer, and region constraints are *intersection-based*—these constraints are violated if there is no intersection between the blocks and the region given. The abutment, boundary, and group constraints are *distance-based*—these constraints are violated if the distance among the blocks is bigger than the given threshold. A prior timing analysis or signal integrity analysis is performed by the user to identify (i) the source of timing, noise, thermal, and power supply problem, and (ii) ways to fix these problems in a form of constraint. Each constraint is then translated into a geometric form so that our multi-level placer attempts to satisfy this geometric constraint. Our strategy is to quantify the amount of violation of the constraints specified, and guide Simulated Annealing based optimization so that the amount of violation is minimized or completely removed if possible. Figure 4(b) and 4(c) show how to compute the penalty from violating region and group constraints.

Table 1. Geometric Constraints for SOP Placement

type	*method*	*syntax*	*meaning*
noise	point	[b_i|(x,y,z)]	b_i touches (x,y,z)
thermal	layer	[B_i|l]	B_i in layer l
power	region	[B_i|(x,y,w,h)]	B_i intersects with region (x,y) and (x+w,y+h)
timing	abutment	[B_i]	B_i abutted
interface	boundary	[B_i|TBLR/l]	B_i near boundary of layer l
cluster	group	[B_i|(x,y,z)]	B_i within a distance of (x,y,z)

Our strategy for effective solution space search during Simulated Annealing is as follows: (i) construction of initial solution: we first assign all blocks under layer constraints to the target layers and fix them during the annealing. For the remaining blocks, we randomly and evenly distribute them into all layers, (ii) solution perturbation: we perform more inter-layer moves (M4 and M5 discussed in Section II.A) during high temperature annealing and more intra-layer moves (M1, M2, and M3) during low temperature annealing, (iii) weighting constants in $C(F)$: we focus more on penalty(F) and via(F) during high temperature annealing and more on area(F) and wire(F) during low temperature annealing. Interested readers are referred to [18] for more details.

5. 3D Global Routing

5.1. Overview of the Algorithm

We have recently developed the first 3D global routing algorithm for SOP [14,16]. Our SOP router is a multi-stage approach, where we divide the routing process into (1) coarse pin distribution, (2) net distribution, (3) detailed pin distribution, (4) topology generation, (5) 2D layer assignment, (6) channel assignment, and (7) pin assignment step. An illustration is shown in Figure 5. By following these steps we seek to have enough and acceptably accurate information for each routing interval to carry out "local" global routing efficiently. We introduce the terms entry/exit pins to better explain our approach. Since the connections of some nets (x-nets) span multiple placement layers, we need to determine the location of entry to and exit from the routing interval. In the 2-D refinement of the problem we treat this location as pins. The routing of i-nets deserves special attention. In routing intervals, except the first and last ones, we have the choice of placing those i-nets in a routing interval either on top or bottom of the placement. The objective is to minimize crosstalk and congestion in the routing interval. This step is called *Net distribution*.

The pin information of the nets is required for efficient net distribution, but net distribution decides the number of pins (and their locations) at each routing interval. We solve this by using the results of *Coarse Pin Distribution* for net distribution. Pins in all routing interval are projected to a single 2-D area and partitioning evenly distributes them over it. The pins in different routing intervals may not be evenly distributed locally. After net distribution we do *Detailed Pin*

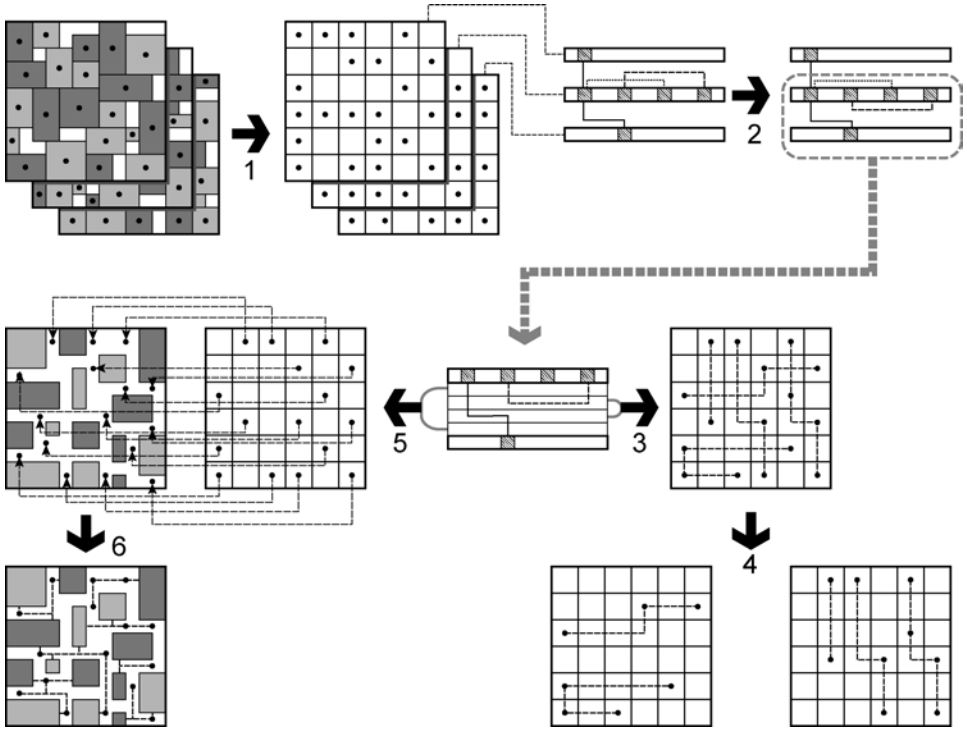

Figure 5. Overview of the global routing process: (1) pin distribution, (2) net distribution, (3) topology generation, (4) layer assignment, (5) channel assignment, (6) pin assignment.

Distribution on each routing interval to minimize the estimated wirelength and legalize pin locations. After this step we have all the information needed for global routing in the routing interval. The topology of each net is generated during our *Topology Generation*, and *2-D Layer Assignment* assigns different layer to conflicting nets. The *Channel Assignment* problem is to assign each pin in the pin distribution layers to a channel in the placement layers such that the routing layers and interconnect costs are minimized. The objective is to facilitate an efficient pin distribution on pin distribution layer with only minimal additional costs. The purpose of *Pin Assignment* is to assign a location to the pin on the block boundary on the placement layer while minimizing the connections between the pin and its "peer" on the channel, which was found out in the previous step. The peer is location in the placement which connects the net to rest of its interconnect in the routing layers.

5.2. Global Routing Algorithms

Coarse Pin Distribution: In this step, we generate coarse locations for all pins of the nets in the routing interval. For the purpose of pin distribution we "flatten" the 3-D SOP structure to 2-D and superimpose a AxB grid on it, were A and B are determined by the size of the circuit. We use our GEO-PD partitioning algorithm [15] to evenly distribute pins to all the partitions formed by this grid while keeping the wirelength minimum. Evenly distributing the pins among all partitions ensures efficient use of the routing resource provided by the single layer. The "coarse" location is the centre of the partition. After the partitioning the pins may not be uniformly distributed in the local routing interval. This partitioning algorithm is smart enough not to move the pins far from their "initial" locations. The algorithm does iterative improvement until good results are obtained.

2-D Layer Assignment: We construct a Layer Constraint Graph (LCG) from the given global routing topology, where each node represents a net and two nodes in the LCG have an edge between them if corresponding net segments of same orientation (horizontal or vertical) share at least one tile in the routing grid. We use a fast node coloring heuristic algorithm to assign a color to the node such that no two nodes sharing an edge are assigned the same color. The algorithm is greedy in assigning colors but performs well and is fast. Close to lower bound results are achieved because the heuristic tries to ensure that nodes with different colors have in fact an edge between them. The complexity of the algorithm is $O(n\log n)$, where n is the number of nets in the routing interval. The complexity is independent of the size of the grid used to compute the tree topologies. The capacity of the tiles determines the number of layers used. We use a simple formula to calculate this number (number of colors/capacity).

Net Distribution: We assign i-nets into routing intervals during the net distribution while attempting to reducing crosstalk. We use the amount of overlap of bounding boxes of the nets as a measure of crosstalk. The net distribution problem is modeled as a graph with each i-net in the routing interval as node and the crosstalk interaction as edges. The weight of the edges denotes the amount of crosstalk between the nets. The crosstalk is calculated by the overlap of the bounding boxes of the net. The coarse pin distribution is used as the approximate location of the pins. It is assumed that nets in different interval are crosstalk shielded, which means no crosstalk exist between nets in different interval. The problem can then be seen as a restricted graph partitioning problem where some of the nodes can only go to one of two

Table 2. Comparison among ESC, GEO, GEO-P, and GEP-PD on 64 ways partitioning. Each algorithm reports cutsize, retiming delay (Dr), static delay (Ds), visible power (Pv) and total power (Pt).

	ESC					GEO					GEO-P					GEO-PD				
ckt	cut	Dr	Ds	Pv	Pt	cut	Dr	Ds	Pv	Pt	cut	Dr	Ds	Pv	Pt	cut	Dr	Ds	Pv	Pt
b17o	3418	59	79	3403	4888	3360	65	83	3404	4889	3842	59	78	3286	4810	3433	64	83	3331	4840
b20o	1808	57	94	1636	2425	1948	56	92	1664	2444	2201	58	94	1533	2357	1958	55	92	1622	2416
b21o	1811	57	96	1565	2389	1982	55	84	1656	2450	2334	54	90	1547	2377	1927	57	96	1587	2403
b22o	2251	57	93	2108	3311	2352	59	97	2161	3347	2712	60	98	2037	3263	2418	61	99	2108	3311
s5378	472	43	49	208	359	428	40	47	201	354	555	47	49	141	314	470	45	49	168	332
s9234	465	44	86	263	580	459	48	79	266	582	612	48	80	219	551	528	48	84	244	567
s13207	459	72	83	343	827	474	67	78	354	834	661	71	83	306	802	520	69	80	312	806
s15850	551	82	116	383	972	548	82	104	396	980	698	81	115	307	921	595	81	110	346	948
s38417	789	41	61	760	2179	829	41	59	760	2180	951	42	67	638	2098	858	41	59	645	2103
s38584	896	63	74	993	2369	1031	61	72	1102	2442	1019	63	74	850	2273	987	63	74	955	2344
Ratio	1.00	1.00	1.00	1.00	1.00	1.03	1.00	0.96	1.03	1.01	1.24	1.02	1.00	0.88	0.96	1.08	1.02	0.99	0.93	0.98
Time	111					1999					124					2054				

predetermined partitions. In order to achieve better results, we used our partitioning algorithm [15]. The complexity of the algorithm is O(V+E), where V is number of nodes and E is the number of edges in the graph model. The results obtained were compared with random net distribution and the case when no i-nets are propagated to other routing intervals.

Detailed Pin Distribution: The results of the coarse pin assignment are used for force-directed placement of the pins in the pin distribution layers. Since we did not consider the layer in which the pin was located in the coarse pin redistribution, it may be possible that the pins exceed the capacity of the partitions local to the routing interval. However our algorithm handles this by moving the pins from such location to the closest available position. The pins are placed in locations near the centre of the net. The pins furthest from its center of the net in coarse assignment, gets placed in the best location (location nearest to the center) in the local partition. The algorithm uses the "approximate" position of the pins as found by coarse pin distribution and the net distribution results to determine the initial location and routing interval of the pin. The position of the pins is stored as the grid location of the coarse pin distribution. The center of each net is calculated from this position of the pins. The displacement vector is calculated by taking the difference of the position of the center of net and the pin. A pair of numbers (a,b) such that $0<a<1$, $0<b<1$ is added to the position of the pins. The numbers reflect the scaled magnitude of the displacement vector. The variables a and b are less than 1 so that we can still keep track of the partitions of the pins. The pins in each routing interval are sorted according to their new positions. The pins are then sequentially assigned to grids previously determined.

Channel Assignment: The channel assignment of pins will affect the additional number of layers and total wirelength. Since one desirable objective is to reduce the number of bends (which would necessitate the use of secondary vias), we assume a straight or L-shaped routing of nets to their assigned channel. We observed that congestion of pin connections and wire crossings on a particular channel would increase the layer count. Thus, we seek to minimize the additional number of layers and wirelength while assigning every pin to a channel. The number of layers for the channel assignment is given by the maximum density among all channels. The cost used in our channel assignment algorithm is the sum of L-distance between pin and channel, the channel density and the bending penalty, multiplied by constants to reflect the relative importance. The L-distance and bending penalty between channel and pin is constant part of the cost while the channel density needed to be updated with an assignment.

Pin Assignment: The routing channels in the placement layers are used to finish the last connections from the channel pins (determined during channel assignment) to the block. The channel pins are actually the entry/exit points to the routing interval. An interesting aspect of this problem is that *complete* connections of the blocks and channel pins of a net is *not* necessary since the channel pins of a net are connected in the routing interval. Hence it suffices if the block is connected to at least one channel pin. This observation reduces the problem to a 2-terminal net pin assignment. The pin is now either a block node or channel node. We use modified Djikstra's algorithm to find the most feasible coarse location for the terminals on the block boundary. The key to efficiently do pin-assignment for 3-D packaging is to have a good 2-pin net generation. The pins which had been projected to routing intervals during pin generation for routing interval now needs to be connected to its originating blocks. However, the pins can also connect to blocks closer to them, which form the part of the same net, if the costs are improved. The edge weights of the FCG derived from the placement are initialized. In the proposed solution we try to minimize the demands on routing as well as the pin-assignment edges while determining the path between the source and destination nodes. We enforce the selection of different routes by making the costs of the edges in the path high. This ensures fair usage of resources modeled by the edges. Interested readers are referred to [14,16] for more details.

6. Experimental Results

Our algorithms are implemented in C++/STL, compiled with gcc v2.96, and run on Pentium III 746 MHz machine.

Table 3. Comparison among single-layer floorplanning and 4-layer SOP floorplanning with/without geometric constraints. We report the final package area, wirelength, total number of vias used, and total runtime. The total number of initial and final constraints, and the number of failed constraints for each constraint type are also reported.

	k=1		k=4, no constraints			k=4, with constraints			# constr		failed constr type					
	area	wire	area	wire	via	area	wire	via	init	fin	*p*	*l*	*r*	*a*	*b*	*g*
n10	258152	18164	66740	8395	43	72850	12827	41	6	2	1	0	0	0	0	1
n10b	251778	15128	74469	10198	45	80496	10252	39	6	1	1	0	0	0	0	0
n10c	268865	19880	71760	11640	37	83200	12422	28	6	2	1	0	0	0	0	1
n30	245115	54586	68420	38106	114	68796	43456	137	10	4	1	0	1	0	0	2
n30b	234574	45931	60984	34850	138	61102	17043	141	9	3	2	0	0	0	0	1
n30c	233867	55979	69153	33353	144	77407	45156	154	9	3	1	0	0	0	0	2
n50	231431	104395	60973	44749	292	64428	55664	312	12	4	2	0	2	0	0	0
n50b	237266	94790	61650	34019	291	63928	104567	317	13	4	2	0	0	0	0	2
n50c	234567	106562	60960	46619	284	67554	40091	314	12	6	4	0	1	0	0	1
n100	210378	180413	56166	54347	455	56852	76069	519	14	7	3	2	2	0	0	0
n100b	185868	169767	50400	58270	500	53074	61863	496	14	7	2	1	2	0	0	2
n100c	208616	185215	53760	56278	495	58322	82858	557	14	6	3	2	0	0	0	1
n200	214349	393644	56977	123765	1098	58548	127741	1074	14	5	1	0	1	0	2	1
n200b	208960	336236	56635	119849	931	56723	119919	915	14	6	2	1	2	0	0	1
n200c	206954	394358	53345	148634	1039	56296	129221	1120	14	6	3	1	1	0	1	0
n300	329589	658162	83232	165433	1354	85200	174274	1413	14	6	2	1	2	0	0	1
ave	235021	177076	62852	61782	454	67799	69589	474								
ratio	1.00	1.00	0.27	0.35	1.00	0.29	0.39	1.04								
runtime	150		1201			1409										

6.1. 3D Partitioning Results

The benchmark set consists of six big circuits from ISCAS89 and four big circuits from ITC99 suites. We generate random switching activity values for these circuits since such information is not available. We assume unit delay for all gates in the circuits. We conduct experiments using ESC [19], GEO [20], GEO-P and GEO-PD algorithms. ESC is a state-of-the-art cutsize driven multi-level algorithm, and GEO is a state-of-the-art simultaneous cutsize and delay driven multi-level algorithm. GEO-P is obtained by setting delay weights of GEO-PD to zero for power optimization only. Lastly, GEO-PD is a simultaneous power and delay driven multi-level algorithm. We report cutsize, retiming delay, static delay, visible power and total power. Table 2 shows 64-way partitioning results among ESC, GEO, GEO-P, and GEO-PD. We first note that the delay improvement of GEO over ESC is not significant. In fact, the retiming delay results got worse by an average margin of 2%, whereas the static delay improved by 4%. GEO-P improves ESC by an average margin of 12% for visible power and 4% for total power at the cost of 24% increase in cutsize. Finally, GEO-PD obtains 2% worse retiming delay and 7% better visible power results than ESC at the cost of 8% increase in cutsize.

6.2. 3D Placement Results

The benchmark set consists of sixteen big circuits from GSRC suite. We report the area, wirelength, inter-layer via, and runtime for 4-layer SOP in all of our experiments. Table 3 shows the comparison among (i) single-layer floorplanning, (ii) 4-layer SOP floorplanning without geometric constraints, and (iii) 4-layer SOP floorplanning with geometric constraints. We randomly select constraints from 6 types for each circuit, and we impose more constraints for bigger circuits.

Compared to the single layer placement, the final package area for 4-layer placement is reduced by 73% on the average (order of $O(k)$ reduction). This indicates that the placement for all 4 layers is highly compact and their shapes are similar. The impact of geometric constraint on final area was not significant—only 2% increase on the average. The wirelength reduction for 4-layer placement is 65% on the average compared to the single-layer case. Since the wirelength in z-direction is not considered (this is actually our via cost), the 65% saving mainly comes from the final package area reduction. The impact of geometric constraint on final wirelength was not significant—only 4% average increase. The impact of geometric constraint on via results was not significant—only 4% average increase. In some cases we were able to find a solution with smaller wirelength and via. The runtime has been increased by 8x with 4-layer placement. The runtime slightly increased when we consider geometric constraints. There are several factors that contribute to the runtime increase: (i) we need highly compact placement for all 4 layers and their shapes need to be similar, (ii) we need to minimize 2-dimensional wirelength and via cost simultaneously. Finally, we note that the distance-based constraints are easier to handle than the intersection-based constraints. This indicates that specifying the absolute location is a stronger constraint than the relative distance.

6.3. 3D Routing Results

Our test cases are generated using our multi-layer SOP floorplanner on GSRC benchmark circuits. The number of layer is fixed to four. Our layer usage results are based on the tile density w=10. The RSA/G-based global routing trees are generated based on 10×10 unless otherwise specified. All the benchmarks completed in less than a few minutes. So we do not explicitly report the runtimes.

In Table 4 we present the results of our Channel Assignment algorithm. We computed the best possible wirelength for a channel assignment where via capacity violations were allowed. We tabulate the results of this scheme under the best wirelength. We compare the results of

our algorithm with the best wirelength results for the number of layer pairs, wirelength, bends and number of pins violating channel via capacities. In channel assignment, we are close to the number of layers predicted by the best case. The increase in layers is due to increased routing density on the channels. The ratios of actual wirelength with best wirelengths increase with the size of benchmarks. The violations in the best case and the number of bends reported by our algorithm are very close, suggesting that violations were fixed by bending the interconnections.

Table 4. The result of multi-objective minimizing channel assignment. Number of layer pairs (*ly*), wirelength (*wl*), bends (*bnd*) and violations (*vl*) for the best and actual cases are reported.

Ckt	Best Wirelength				Multi-Objective			
	ly	*wl*	*bnd*	*vl*	*ly*	*wl*	*bnd*	*vl*
n10	5	6963	0	55	5	14174	34	0
n30	6	11288	0	144	6	24726	122	0
n50	6	14826	3	282	6	32059	189	0
n100	6	16430	3	413	7	36331	382	0
n200	6	24078	1	845	8	61485	917	0
n300	6	28469	3	1000	8	65189	1029	0

In Table 5 we report the wirelength achieved during pin assignment. The result of the channel assignment is used as input to the pin assignment. For generating the best wirelength, we used the corresponding best wirelength channel capacities violating channel assignment. For the best wirelength, we allowed pin assignment algorithm to select routes without considering the pin assignment and routing demands. Our algorithm tries to minimize wirelength while avoiding congestion of routing channels and pin assignment resources. The parameters of the algorithm decide the trade-off between wirelength, pin assignment demands and routing demands. In pin assignment, we are able to reduce pin assignment and routing demands drastically by increasing only 25% wirelength, from the best case. The wirelength scales rapidly with benchmark sizes and the wirelength for pin assignment is huge compared to channel assignment due to limited routing resources in the placement layer.

Table 5. The pin assignment results. Wirelength (*wl*), pin assignment demands (*pd*) and routing demands (*rd*) are reported.

Ckt	Best Wirelength			Multi-Objective		
	wl	*pd*	*rd*	*wl*	*pd*	*rd*
n10	10327	11	15	12173	9	11
n30	42108	16	47	54730	12	18
n50	86052	31	70	120743	15	59
n100	155089	20	83	194724	12	42
n200	343883	23	139	424620	11	76
n300	561244	27	162	692383	11	79

7. Conclusions

In this paper, we presented the first 3D physical design algorithms targeting SOP technology. We are currently working on detailed routing for SOP. In addition, we are integrating STA (Static Timing Analysis), SIA (Signal Integrity Analysis), and TPA (Thermal and Power Analysis) engines into our physical design algorithms so that the geometric constraints are also automatically generated. The goal is to develop built-in STA/SIA/TPA that runs fast but with high fidelity so that they will not slow down the optimization process while guiding the optimization.

References

[1] Rao Tummala and Vijay Madisetti, "System on Chip or System on Package?," *IEEE Design & Test of Computers*, pp 48-56, 1999.

[2] Y. Deng and W. Maly, "Interconnect characteristics of 2.5-D system integration scheme," Int. Symposium on Physical Design, 2001.

[3] Xiaoping Tang, Ruiqi Tian, D.F. Wong, "Fast Evaluation of Sequence Pair in Block Placement by Longest Common Subsequence Computation," Design, Automation and Test in Europe Conference, 2000. pp 106-111.

[4] H. Murata, K. Fujiyoshi, et al., "VLSI module placement based on rectangle-packing by the sequence pair," IEEE Transaction on CAD, vol. 15:12, pp. 1518-1524, 1996.

[5] J. Cong, "Pin assignment with global routing for general cell designs," *IEEE Trans on Computer-Aided Design of Integrated Circuits and Systems*, pp1401-1412, 1991.

[6] J. Cong, A.B. Kahng, K Leung, "Efficient Algorithms for the Minimum Shortest Path Steiner Arborescence Problem with Applications to VLSI Physical Design," *IEEE Trans. on CAD*, vol. 17, no.1, January 1998.

[7] Khoo and J. Cong, "An efficient multilayer MCM router based on four-via routing," *IEEE Trans. Computer-Aided Design*, Oct. 1995, vol.14, pp. 1277-90

[8] Ho, M. Sarrafzadeh, G. Vijayan, C. Wong, "Layer assignment for multi-chip modules," *IEEE Transactions on Computer-Aided Design*, Vol. 9, No. 12, December 1990, pp. 1272-1277

[9] J. Cho, M. Sarrafzadeh, M. Sriram, S. Kang, "High Performance MCM Routing," *IEEE Design & Test of Computers*, 1993.

[10] Shamik Das, Anantha Chandrakasan, Rafael Reif, "Design Tools for 3-D Integrated Circuits", Proc. Asia and South Pacific Design Automation Conference, 2003.

[11] Brent A. Goplen, Sachin S. Sapatnekar, "Efficient Thermal Placement of Standard Cells in 3D ICs using a Force Directed Approach", Proc. Int. Conf. on Computer Aided Design, 2003.

[12] Thitipong Tanprasert, "An Analytical 3-D Placement That Reserves Routing Space," Intl. Conf. on Circuits and Systems, 2000.

[13] Rongtian Zhang, Kaushik Roy, Cheng-Kok Koh, and David B. Janes, "Exploring SOI Device Structures and Interconnect Architectures for 3-Dimensional Integration," Proc. 2001 Design Automation Conference, 2001.

[14] Jacob Minz and Sung Kyu Lim, "Layer Assignment for System on Packages", ACM/IEEE Asia and South Pacific Design Automation Conference, 2004.

[15] Mongkol Ekpanyapong and Sung Kyu Lim, "Performance-driven Global Placement via Adaptive Network Characterization", ACM/IEEE Asia and South Pacific Design Automation Conference, 2004.

[16] Jacob Minz, Mohit Pathak, and Sung Kyu Lim, "Net and Pin Distribution for 3D Package Global Routing", Design, Automation and Test in Europe, 2004.

[17] Mongkol Ekpanyapong, Karthik Balakrishnan, Vidit Nanda, and Sung Kyu Lim, "Simultaneous Delay and Power Optimization for Multi-level Partitioning and Floorplanning with Retiming", IEEE International Conference on Circuits and Systems, 2004.

[18] Pun Hang Shiu, Ramprasad Ravichandran, Siddharth Easwar, and Sung Kyu Lim, "Multi-layer Floorplanning for Reliable System-on-Package", IEEE International Conference on Circuits and Systems, 2004.

[19] Jason Cong and Sung Kyu Lim, "Edge Separability based Circuit Clustering With Application to Multi-level Circuit Partitioning", IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Vol. 23, No. 3, March 2004, pp 1-12.

[20] Jason Cong and Sung Kyu Lim, "Physical Planning with Retiming", IEEE International Conference on Computer Aided Design, p2-7, 2000.