ISR TECHNICAL REPORT 2012-10

Co-optimization of TSV assignment and micro-channel placement for 3D-ICs

Bing Shi, Ankur Srivastava and Caleb Serafy

Co-optimization of TSV assignment and micro-channel placement for 3D-ICs

Bing Shi, Ankur Srivastava and Caleb Serafy
University of Maryland, College Park, MD, USA

ABSTRACT

The three dimensional circuit (3D-IC) brings forth new challenges to physical design such as allocation and management of through-silicon-vias (TSVs). Meanwhile, the thermal issues in 3D-IC becomes significant necessitating the use of active cooling schemes such as micro-channel liquid coolings. Both TSVs and micro-channels go through the interlayer regions of 3D-IC resulting in potential resource conflict, which deters the optimization of both micro-channel design and TSV allocation/management. This paper investigates the co-optimization of TSV assignment to interlayer nets and micro-channel allocation such that both wirelength and micro-channel cooling energy are co-optimized. We propose a multi-commodity flow based formulation followed by simplifying transformations that enable use of effective polynomial time heuristics. The experimental results show that, our co-optimization approach achieves 46% cooling power savings or 7.6% wire length reduction compared with the approaches that assign TSVs and allocate micro-channels separately.

Categories and Subject Descriptors

B.7.2 [**Integrated Circuits**]: Design Aids

General Terms

Design, Algorithm

Keywords

3D-IC, micro-channel, liquid cooling, TSV assignment

1. INTRODUCTION

Three dimensional circuit (3D-IC) contains two or more layers of active silicon stacked vertically. It provides several advantages including faster on-chip communication, higher overall device densities, heterogeneous integration etc. Several research directions are being pursued that attempt to develop effective tools for 3D-IC physical design. One noteworthy problem in this regard is allocation and management of through silicon vias (TSVs) that enable communication between layers. Two general approaches have been investigated: Post Placement [1][2][3] and In-placement [4]. In Post Placement approaches, cells are firstly placed in the 3D-IC. This determines the whitespace distribution capable of supporting TSVs. These potential TSV locations are then allocated to the interlayer nets such that the total wirelength is minimized [1][2][3]. In-placement approaches perform simultaneous optimization of cell placement, TSV placement and interlayer net to TSV assignment during the 3D-IC placement process itself.

Despite significant potential performance improvement, stacked 3D structures bring forth new challenges pertaining to the removal of high heat density resulting from several stacks of chips. Researchers have shown that the power density of future 3D-IC systems could easily go beyond $200W/cm^2$ [5], a level that air cooling alone is incapable of removing. Micro-channel based cooling, where micro-scale channels are embedded in the interlayer regions of 3D-ICs, is capable of removing significantly higher power levels by pumping coolant through these channels. Many modeling and optimization approaches have been investigated that attempt to characterize and optimize the thermal behavior of 3D-ICs with micro-channels [6][7][8]. Micro-channel based liquid cooling uses extra energy for performing chip cooling. The effectiveness of micro-channels strongly depends on the structure and locations of micro-channels. Many works have been done to improve the cooling effectiveness by optimizing the dimension and locations of micro-channels [9][10]. However, TSVs impose significant constraints on how and where the micro-channels could be located, and form obstacles to the micro-channels since the micro-channels cannot be placed at the locations of TSVs in the interlayer regions. The existing works on micro-channel allocation assume that the TSV locations has already been decided before micro-channel allocation and then place micro-channels in the remaining areas [10]. Existing works on TSV allocation ignore this possible resource conflict with micro-channels [1][2][3]. Two trivial approaches for allocating TSVs to nets and micro-channels to interlayer regions *together* can be conceived as follows: TSV first approach and Micro-channel first approach. If micro-channels are allocated before TSVs, there is a possibility of increase in wirelength since the available whitespace for TSVs shrinks due to the existence of micro-channels which deter allocation of TSVs in those areas. A TSV first approach also has disadvantages. If TSVs are placed only to minimize the wirelength (as is the case in previous approaches in [1][2][3]), they might be placed in some thermally critical regions, and therefore micro-channels will fail to reach these regions thereby resulting in thermal violation or hotspots. In order to cool the hotspot area, we may need to increase the flow rate through micro-channels, which results in increase in cooling energy consumption.

Note that an alternative approach is to use more complex micro-channel structures such as bended micro-channels, which enable the channels to reach the hotspot regions. However, empirical data shows that complex micro-channel structures cause flow imbalance and are difficult to control. Therefore we focus on straight micro-channels in this work.

In this paper, we investigate co-optimization of TSV assignment and micro-channel allocation such that both wirelength as well as fluid pumping energy is co-optimized. TSV assignment corresponds to allocating interlayer nets to TSVs and deciding their locations. To solve this complex optimization problem, we first propose a multi-commodity min-cost flow based formulation followed by simplifying transformations that enable use of effective polynomial time heuristics. The experimental results show that, our co-optimization approach achieves 46% cooling power savings only at a cost of 0.93% wirelength increase compared with TSV first approach, and saves 7.6% wirelength compared with the micro-channel first approach. Moreover, TSV first approach also results in thermal violations in multiple benchmarks.

The paper is organized as follows. In section 2, we introduce the 3D-IC with micro-channels and TSVs. Section 3 illustrates the motivation of our work and problem formulation. We explore co-optimization of TSV assignment and micro-channel allocation in sections 4, 5. The experimental result is given in section 6.

2. PRELIMINARIES

2.1 3D-IC structure with micro-channels and TSVs

Higher cooling demand in 3D-ICs has motivated a large body of work that attempts to embed fluidic channels in interlayer regions [5][6][7]. This region does not have any transistors or planer wires but has TSVs that enable interconnection between layers. Figure 1 shows the cross section of a three-layer stacked 3D-IC integrated with micro-channels. The coolant is pumped through micro-channels and takes away the heat generated in the active layers.

2.2 Thermal and hydrodynamic modeling of 3D-IC with micro-channels

As illustrated in [11], the 3D-IC system with fluidic channels can be represented as a 3D mesh whose thermal behavior is represented as an RC circuit where R corresponds to thermal conduction and C corresponds to heat capacity. A given 3D-IC power profile represents current sources in this RC network. In several cases, we are mostly interested in the steady state thermal behavior of the 3D-IC, hence, enabling us to capture the thermal behavior as a pure resistive network [7][8]. In this case, for a given 3D-IC power profile, the thermal profile T could be estimated by solving a system of linear equations of the form $GT = Q$ where G is the thermal conductivity matrix and Q is the power profile. The thermal conductivity matrix G depends on many factors including the material properties, locations of channels as well as TSVs (TSVs affect thermal conductivity as well [12]), fluid flow rate etc. For brevity, we do not go into further details about how this model is generated (it basically depends on the thermal resistive mesh parameters). Interested readers are referred to [7][8] for details.

Fluidic cooling is active in nature since it uses pumping power to perform cooling. This pumping power comes from the work done by the fluid pump to push the coolant into micro-channels. The pumping power Q_{pump} depends on the fluid flow rate through micro-channels and the pressure drop across micro-channels. Basically increase in the number of micro-channels and fluid flow rate would improve the cooling performance of micro-channels, however, at a cost of increased pumping power. The pumping power is given by $Q_{pump} = \sum_{n=1}^{N} f_n \Delta p_n$ where N is the number of channels, Δp_n is the pressure drop and f_n is the fluid flow rate in the n-th channel. Usually all channels can be assumed to have the same Δp_n and therefore f_n, hence minimization of N while maintaining the 3D-IC temperatures would result in the smallest pumping power assuming the pressure drop across the channels is fixed by the pump.

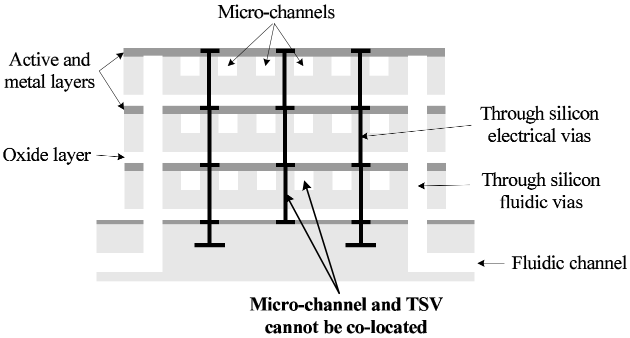
Micro-channels
Active and metal layers
Oxide layer
Through silicon electrical vias
Through silicon fluidic vias
Fluidic channel
Micro-channel and TSV cannot be co-located

Figure 1: Stacked 3D-IC with micro-channels

3. MOTIVATION AND PROBLEM FORMULATION

In 3D-ICs, the interlayer nets use TSVs to deliver signals and power among different layers. Recently, significant attention has been made to the problem of allocating interlayer nets to TSVs that allow their successful routing. Existing work mostly tries to address this problem with the objective of minimizing total wirelength. As mentioned earlier, two approaches have been investigated: Post Placement and In-placement. While both approaches have their advantages, in our work, we assume the placement to be already done before TSV assignment to the interlayer nets (Post Placement paradigm), though our work could also be extended to the In-placement approach.

Conventional Post Placement approaches for interlayer net to TSV assignment do not consider the possibility of adding micro-channels in the interlayer regions. TSVs impose significant constraints on how and where the micro-channels can be located, and form obstacles to the micro-channel placement since the micro-channels cannot be placed at the locations of TSVs. As illustrated in figure 2, presence of TSV constraints us from allocating a micro-channel close to the thermal hotspot. Hence allocating TSVs to nets without considering its impact on micro-channel placement can have the following detrimental effects:

1. As illustrated in figure 2, if TSVs are placed surrounding thermally critical areas, micro-channels would fail to reach these areas thereby resulting in hotspots.

2. In order to fix the thermal hotspots in areas where micro-channels cannot reach, we would need to increase the fluid flow rate or place more micro-channels

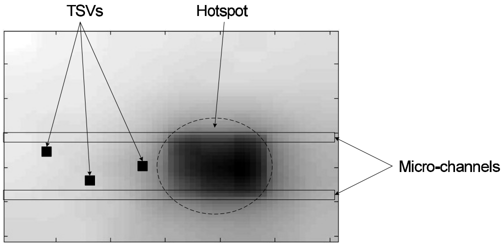

Figure 2: Thermal profile of one 3D-IC layer, and an example of TSV and micro-channel allocation where TSVs constraint us from allocating micro-channels at hotspots

in the surrounding regions, resulting in a significant increase in cooling energy.

Recently some research has been done to address the problem of micro-channel placement to satisfy thermal constraints at minimum pumping energy while accounting for constraints imposed by existing TSVs [10]. The location of TSVs is essentially decided by the allocation of interlayer nets to TSVs. The exiting works for Post Placement TSV allocation (which ignore the possibility of allocating channels) and micro-channel placement (which assume the TSV locations to be fixed) do not consider the possibility of combining these steps for obtaining better results.

Two trivial approaches for allocating TSVs to nets and micro-channels to interlayer regions *together* can be conceived as follows: TSV first approach and Micro-channel first approach. If micro-channels are allocated before TSVs, there is a possibility of increase in wirelength since the available whitespace for TSVs shrinks due to the existence of micro-channels which deter allocation of TSVs in those areas. A TSV first approach has disadvantages as enumerated above.

In this paper, we investigate co-optimization of TSV assignment and micro-channel allocation simultaneously such that the total wirelength is minimized, and maximizing the micro-channel cooling effectiveness. As stated earlier, we assume a Post Placement paradigm. The problem is stated as follows. Given:
(I1) a 3D-IC placed netlist. The placement information can be used to generate potential TSV locations;
(I2) a netlist that describes a set of interlayer nets;
(I3) the power profile of 3D-IC;
(I4) a set of potential locations for interlayer micro-channels. These channels are to be incorporated in the interlayer region of the chip;

We would like to:
(O1) decide the locations of TSVs;
(O2) assign a set of TSVs to each interlayer net;
(O3) decide the number and locations of micro-channels;

In such a way that:
(C1) the assigned set of TSVs for each interlayer net forms a path connecting the source and destination terminals of the net;
(C2) the locations of micro-channel and TSVs do not conflict (see figure 1 for detail);
(C3) the micro-channels provide sufficient cooling for the 3D-IC, i.e. $T_i \leq T_{max}, \forall locations : i$;
(C4) the total wirelength and required pumping power by micro-channels is minimized: $\min w_1 N + w_2 \sum_{\forall l} WL_l$ where N is the number of channels and WL_l is the bounding box wirelength of the l-th interlayer net which depends on the TSV set it has been allocated to. Constants w_1 and w_2 could be allocated based on preference for a particular tradeoff. It is noteworthy that following the discussion in section 2, the pumping power is directly proportional to the number of channels N since we assume that the pressure drop/fluid flow rate is fixed.

The objective minimizes a combination of the cooling power required by micro-channels and the total wirelength used by all interlayer nets. It is noteworthy that an interlayer net is allocated to a *set* of TSVs since several TSVs spanning multiple layers may be needed to connect the source-destination pairs. The co-optimization of micro-channel allocation and TSV assignment simultaneously is complex due to its discrete nature and the complexity of thermal estimation. Hence, in our work, we focus on developing effective heuristics that exploit specific mathematical properties present in this problem. We formulate this problem as a multi-commodity minimum cost flow (MCMCF) instance, which, even though is NP Complete, has several effective approximation algorithms. Furthermore, we exploit specific properties in the problem structure to develop effective heuristics.

4. CO-OPTIMIZATION OF TSV ASSIGNMENT AND MICRO-CHANNEL PLACEMENT

4.1 Overall design flow

The overall design flow is shown in figure 3. We use multi-commodity min-cost flow to formulate/solve some critical aspects of the problem, hence we call this approach MCMCF. In the next section we discuss simplifications to this formulation that enable us to solve the problem efficiently. We firstly find the thermal criticality of all grid locations in the 3D-IC chip using a full chip thermal analysis assuming there are no micro-channels. Also, based on the 3D-IC structure and placement, we identify all the potential locations of micro-channels and TSVs.

Assuming the 3D-IC is divided into small grids (i, j, k), with i, j representing the face of the 3D-IC and k representing the longitudinal direction along which the micro-channel runs. The location of a micro-channel is basically the (i, j)-th grid where the channel is located. In the k-th direction, the channel spans the chip anyway. The TSV could be identified by the (i, j, k)-th grid it is located at. After initial thermal analysis, we define a thermal criticality of each potential micro-channel which basically represents the demand of allocating a micro-channel at that location. The criticality factor $g(i, j)$ for each micro-channel location (i, j) is defined as:

$$g(i,j) = \sum_{k=0}^{K} w(i,j,k) \cdot max[0, T_{i,j,k} - T_{max}] \qquad (1)$$

where $T_{i,j,k}$ represents the temperature at grid (i, j, k), T_{max} is the maximum thermal constraint. Parameter $w(i, j, k)$ represents the thermal significance of a certain grid, and K is the number of grids in the longitudinal direction in which the channel spans the entire chip. Based on the criticality factor $g(i, j)$, we formulate the MCMCF problem and obtain the TSV assignment and micro-channel allocation simultaneously (see figure 3). Thermal analysis and 2D routing are then conducted to evaluate the performance of the

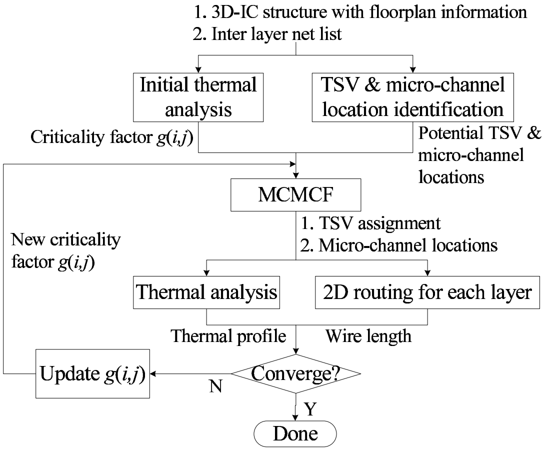
1. 3D-IC structure with floorplan information
2. Inter layer net list
Initial thermal analysis
TSV & micro-channel location identification
Criticality factor g(i,j)
Potential TSV & micro-channel locations
MCMCF
1. TSV assignment
2. Micro-channel locations
New criticality factor g(i,j)
Thermal analysis
2D routing for each layer
Thermal profile
Wire length
Update g(i,j)
N
Converge?
Y
Done

Figure 3: Overall design flow of MCMCF based algorithm

resulting design. If the design results in thermal violation, ends up having significant wirelength or placing too many micro-channels in some locations (this will increase cooling power and might also degrade wirelength), we will refine the criticality factor $g(i, j)$ (increase or decrease $g(i, j)$ accordingly) and re-solve the MCMCF problem.

We repeat this process iteratively until obtaining a design that achieves required tradeoff between cooling power and wirelength.

4.2 Multi-commodity minimum cost flow formulation

Given: a) the 3D-IC structure, b) potential locations of TSVs and micro-channels, c) the interlayer netlist and d) criticality factor $g(i, j)$, the multi-commodity min-cost flow (MCMCF) problem is illustrated in figures 4 and 5. Figure 4 illustrates a 3D-IC with three active layers and two interlayer nets along with four potential TSVs total. The potential locations of micro-channels have also been indicated. Both front and top views have been illustrated that indicate TSV and net locations in the 3D-IC grids. Our objective is to find the allocation of nets to TSVs and micro-channels such that cumulative objective indicated in the previous section is minimized: $\min w_1 N + w_2 \sum_{\forall l} WL_l$. Assuming w_1 and w_2 is the same for the sake of ease in exposition, we instantiate a multi-commodity min-cost flow formulation as follows.

For each net, we allocate one unit of unique commodity flow. Hence P nets would correspond to P distinct units of commodity flows. The flow network has one node for each terminal of the nets and also the potential TSV locations. We assume that the net terminal in the higher layer is the source of this one unit flow and net terminal in the lower layer is the sink. We assume all nets are two terminal. For the example shown in figure 4, the flow network is illustrated in figure 5 which indicates that $net1$ and $net2$ terminals in the top layer are sources. They are connected by directional edges to the TSV nodes in that active layer. If the nets span multiple layers (> 2), then the TSVs in this layer would connect to the TSVs in the layer just below to transfer the signal. This is also indicated in figure 5(a) where TSVs in layer 1 are connected by directional edges to TSVs in layer 2. Finally destination or sink terminals of nets are also connected by directional edges to TSVs in that layer as indicated in the figure. Note that the edges always carry flow from source to sinks. Also, by construction this network forms a directed acyclic graph. Let us ignore the presence of micro-channels for now (for ease of explanation). The problem of allocating nets to TSVs could be modeled as a multi-commodity min-cost flow (and the flow graph is illustrated in figure 5(a)). Let each net to TSV edge or TSV to TSV edge have a cost which is simply the half perimeter bounding box between the two. Let each TSV node have a total capacity of 1. Also all edges have an individual commodity capacity as 1 and a total capacity as 1. The multi-commodity min-cost flow solution that sends the unique commodity flow from each net source to the corresponding net sink on the network in figure 5(a) at the minimum total cost corresponds to the net to TSV assignment with minimum total wirelength. A total unity capacity for each TSV node ensures that only one net is allocated to it.

Now we extend this formulation to account for the presence of micro-channels. As indicated in figure 4, some micro-channel locations conflict with TSVs while others don't. The micro-channel is allocated an additional flow commodity. Hence if there are P interlayer nets then the total commodities becomes $P + 1$. Figure 5(b) indicates the process of accounting for micro-channels in the flow network of figure 5(a). The figure shows the top view of an interlayer region where the micro-channels are located and potentially conflict with TSVs. The micro-channels span the entire length of the chip in the k direction. If there is even one TSV allocated in this path, a channel cannot be allocated and vice versa. Now some potential micro-channel locations do not have any potential TSVs while others do (see figure 5(b)). We instantiate a source at the beginning of each potential channel location and a sink at the end. This source contains a unique commodity corresponding to the fluid flow. Note that all sources corresponding to micro-channel locations have the same flow type (same commodity). Several paths exist between the source and sink for a particular channel location. The simplest path is the one that goes through all the grids that span the entire length of the chip (see figure 5(b)). Some of these grids are potential TSV locations and have other nets and/or TSVs connected to them by way of directional edges (as indicated in figure 5(a)). The fluid flow edges are directed longitudinally while the net interconnection edges are directed vertically. As indicated earlier, the TSV nodes have a total capacity of one. The longitudinal edges that represent fluid flow edges have unit capacity for the fluid flow commodity while 0 capacity for all the P net commodities. The fluid flow edges that connect the adjacent grids (*direct edges* in figure 5(b)) have a cost of 0. Now each intermediate grid in this direct path is also connected directly to the fluid sink for that channel location by way of *offset edges* as well. This edge also has a fluid commodity capacity of 1 and net commodity capacity of 0. The cost of this edge is $g(i, j)$ which represents the cooling demand for that channel location. All the edges that represent net to TSV or TSV to TSV connection have a fluid flow commodity capacity as 0. Note that different micro-channel locations do not interfere since the network does not have any edges that interconnect them. Now the cheapest way of sending a fluid flow commodity from source to sink is to follow the simple path that spans all the adjacent nodes. The cost of this path is 0 but it forces us to use all the

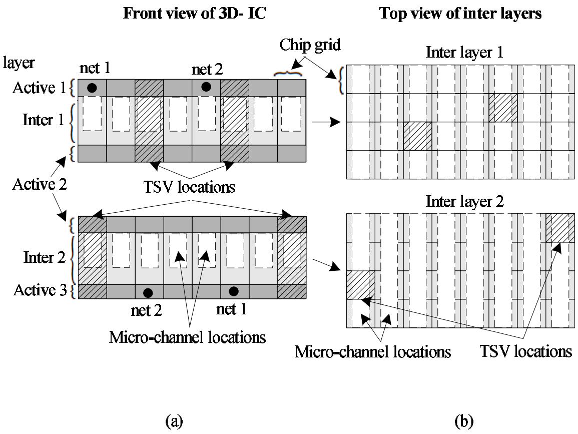

Figure 4: 3D-IC with potential TSV and micro-channel locations

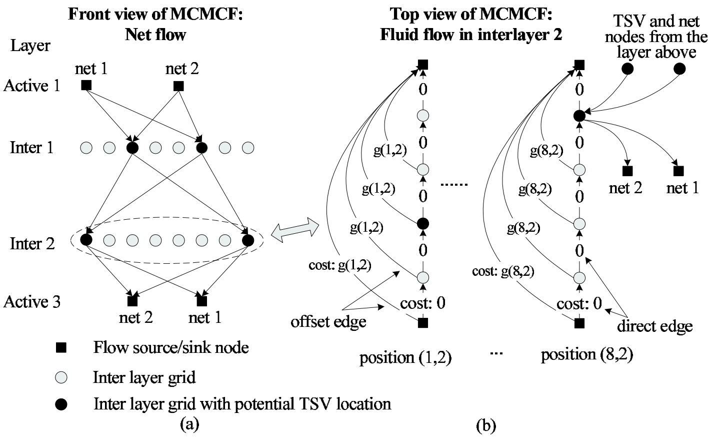

Figure 5: Multi-commodity min-cost flow formulation

potential TSV nodes on the path since they have a total capacity as 1. Hence none of these potential TSV nodes can be used for net interconnection. On the other hand, if any one of these nodes has been allocated to a net, a fluid commodity cannot go through this simple path anymore and it has to take any one of the alternative paths to the sink (see *offset edges* in the figure). Any such alternative path has a cost of $g(i,j)$ which represents the price that we pay by *not* having a channel at that location since we would rather use some of the TSV on the channel path for routing nets.

Sending min-cost multi-commodity flow on this network results in an allocation of nets to TSVs and micro-channels to channel locations such that the total cost is minimized. This cost is a combination of $g(i,j)$ and bounding box wirelength, and represents a balance between cooling and wirelength. Solving multi-commodity problem is a challenging problem since the formulation is generally NP-Complete, although several effective heuristics have been developed. In this paper, we investigate some specific properties in our problem that help us simplify the formulation thereby enabling us to use simpler, computationally efficient heuristics. Such extensions are discussed subsequently.

4.3 Iterative optimization

As indicated in figure 3, once an allocation of micro-channels and TSVs has been conducted, we a) perform routing to compute the actual wirelengths and b) thermal analysis. If the wire-lengths are unacceptable, thermal violations occur or the system is overcooled (pumping power is wasted), the $g(i,j)$ values are re-allocated and the problem re-solved. If wire-lengths are very high, then $g(i,j)$ values are uniformly scaled down enabling us to prefer wirelength over channels. If the system experiences thermal violations, $g(i,j)$ values are increased enabling us to use more micro-channels. If the system is non-uniformly overcooled then regions where excessive cooling is available are subjected to a reduction in $g(i,j)$ which could end up in removing the channels in favor of using TSVs. Such an approach assists in achieving the optimal balance between wirelength and cooling power while satisfying the thermal and interconnection constraints.

5. COMPUTATIONAL SIMPLIFICATIONS

5.1 Multi-layer case

Solving multi-commodity flow instances in general is computationally intractable. For our specific case, this is a bigger issue since the number of commodities is linear in the number of interlayer nets P (which could be quite large). This significantly adds to the number of unknowns in the problem formulation making its solving computationally expensive. We first simplify the formulation without losing optimality followed by effective heuristics. We transform the flow graph illustrated in figure 5 to the one illustrated in figure 6. For the moment let us ignore the fluid flow network in figure 5(b). For each distinct net, let us replicate the entire network graph in figure 5(a) P times (one replica for each net). This is illustrated in figure 6(a). Basically all the TSV nodes in the original network appears P times in the new network. The graphs for each net do not have any common edges, hence we don't need to represent the net flows by different commodities. All the net flows belong to the same commodity. The edge costs and the node/edge capacities are exactly the same as before. Sending unit commodity min-cost flow on this network, though, does not solve our problem. This is because the same TSV may be used by two or more nets. In order to address this problem we can allocate a *bundle capacity* to all replicated TSV nodes corresponding to the same TSV. A bundle capacity constraint in network flow problems allocates a total capacity to a bundle of nodes or edges. In our case we can set a bundle capacity constraint of 1 to all replicated TSV nodes belonging to the same TSV. This is illustrated in figure 6(a). The problem continues to be NP complete but we have eliminated the need for different commodities by adding an additional bundle constraint. We found through our experiments that this significantly enhanced the computational efficiency (results reported later).

Adding micro-channel allocation constraints is illustrated in figure 6(b). Just as figure 5(b), each potential micro-channel location has the associated network as illustrated, but the TSV nodes in the fluid network do not have the edge connections to net flow in this case. Instead of allocating a different commodity to fluid flow, we allocate the same commodity. Now the TSV location nodes in the network of figure 6(b) have a bundle capacity of 1 with the replicated TSV nodes for the corresponding TSV in the rest of the network. Hence if a TSV is allocated to a net, then it cannot be allocated to any other net or micro-

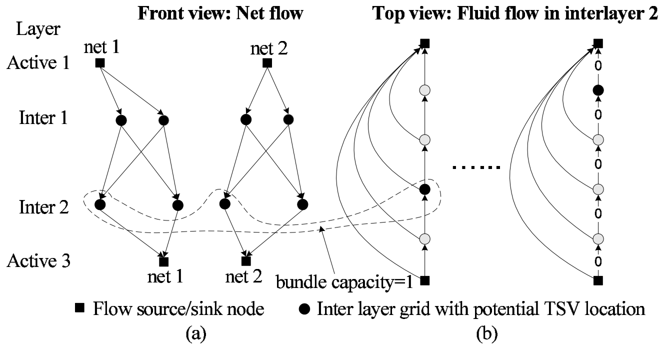

Figure 6: Computationally simplifying transformation for multi-layer case

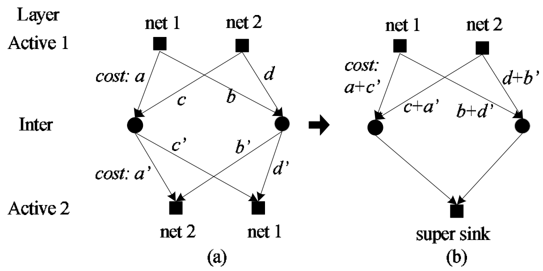

Figure 7: Computationally simplifying transformation for two-layer case

channel. The problem now becomes a (single-commodity) min-cost flow problem with bundle capacity constraints. While solving this problem formulation is NP Complete, it has significantly smaller number of unknowns although the constraints are a bit more complex. We solve this problem by assuming that the discrete flow variables are continuous. This results in a linear programming approximation (polynomially solvable) for this discrete problem. After getting the solution, non-discrete values are rounded up appropriately to give a valid solution.

5.2 Two Layer case

Now we discuss the special case where there are only two active layers stacked together. While the simplification for multi-layer case described above could certainly be applied here, there are additional transformations we can use. Consider the instance illustrated in figure 7(a) where we have two nets and two TSVs. Once again, let us ignore the micro-channel constraints for the moment. Allocation of nets to TSVs in this case is easier than the multi-layer case since it can be transformed to a simple case of bipartite matching. We instantiate a network as illustrated in figure 7(b). For each net (unlike net terminal in the previous case) we have a node and for each TSV we have a node. We have directed edges between nets and TSVs whose cost is the total bounding box between the net's two terminals and the TSV pads in the corresponding layers (see figure 7(b) for an illustration). Each node corresponding to the nets has a unit flow (of the same commodity) available. We also have a super sink that is connected to all the TSVs. The TSV nodes have a capacity of 1. Sending min-cost flow from net nodes to the super sink would essentially correspond to allocation of nets to TSVs with minimum total wirelength optimally in polynomial time. In order to add micro-channel location constraints to this formulation, we essentially apply the method used in the multi-layer case (with bundle constraints). Note that in this case, no replication of nodes for TSV assignment was needed as in the multi-layer case, hence the generated formulation is much simpler than simply applying the previous technique to this case directly. The problem is still NP Complete due to the bundle capacity constraints. We simplify the formulation to a linear program LP by assuming the flow variables are continuous. The generated continuous solution is then discretized by rounding of the non-discrete variables.

6. EXPERIMENTAL RESULTS

In our experiment, we tested both two-layer and three-layer 3D-ICs. We use IBM-PLACE 2.0 circuits with placement information as the benchmark [13]. For each test, we choose two or three circuits from $ibm01 - ibm10$ circuits, each circuit corresponds to one 3D-IC layer. Based on the placement information, we find the whitespace between layout, which are basically the potential TSV locations. The number of potential TSV locations ranges from around 50-1000. We also randomly generate 30-200 interlayer nets. To obtain the power profiles for each layer, we randomly assign a value for each cell as the power density for the cell. This forms our testing benchmark, and such benchmark size is comparable to the experiment in [1]. The chip dimension is $9 \times 9mm^2$. The micro-channel width × height is $100 \times 100\mu m^2$, and the diameter of TSV is $10\mu m.$ The maximum temperature constraint T_{max} is 85°C.

6.1 Comparison of wirelength and pumping power

For each benchmark, we test three approaches to compare the wirelength and pumping power:

a) Our *co-optimization* approach;

b) *TSV first* approach, which firstly assigns TSVs to interlayer nets assuming there are no micro-channels. Once TSVs are assigned and hence TSV locations are decided, we allocate micro-channels in the remaining interlayer regions using the approach in work [10];

c) *Micro-channel first* approach, which allocates micro-channels first assuming there are no TSVs, and then assigns interlayer nets to the remaining available TSV locations.

For each approach, once we obtained the TSV assignment result, we route the interlayer nets to the TSVs in each layer separately using Labyrinth 2D router [14] to obtain the total wirelength (WL). We also estimate the pumping power Q_{pump} based on the number of channels used and the given pressure drop.

Table 1 shows the comparison of wirelength and micro-channel cooling power for the three approaches. In the table, "below T_{max}" indicates if the achieved thermal profile satisfies the thermal constraint, *MC first* indicates the *micro-channel first* approach. Table 1 shows that using air cooling results in thermal violation for all power profiles, while micro-channels can provide sufficient cooling. Moreover, using *TSV first* approach, though achieves good wirelength compared with *micro-channel first* approach, uses about 160% more pumping power since the existence of TSVs deters the optimal allocation of micro-channels. Moreover, for some benchmarks, the TSVs are allocated in thermal critical regions, in which cases micro-channels cannot effectively cool these thermal critical regions thereby

Table 1: Comparison between our approach, TSV first and channel first approach($Q_{pump} : W$, $WL : \#grids$, temperature: ^oC)

				Air	TSV first			Micro-channel first			Co-optimization						
			# inter	cooling			Below			Below			Below	WL change w.r.t		Q_{pump} change w.r.t.	
Circuit	#layer	#TSV	layer nets	T_{peak}	WL	Q_{pump}	T_{max}	WL	Q_{pump}	T_{max}	WL	Q_{pump}	T_{max}	TSV first	MC first	TSV first	MC first
1	2	56	30	106.46	1294	7.98	Y	1424	5.58	Y	1300	6.38	Y	+0.46%	-8.71%	-20%	+14%
2	2	119	50	101.11	1865	n/a	N	2147	3.18	Y	1913	3.89	Y	+2.57%	-10.90%	n/a	+22%
3	2	190	80	121.97	3010	n/a	N	3402	3.90	Y	3165	4.96	Y	+5.15%	-6.97%	n/a	+27%
4	2	348	100	110.25	8324	25.52	Y	8818	6.38	Y	8419	9.57	Y	+1.14%	-4.52%	-63%	+50%
5	2	652	125	128.83	10194	19.14	Y	10838	6.38	Y	10180	9.57	Y	-0.14%	-6.07%	-50%	+50%
6	3	175	50	135.28	2468	n/a	N	2811	15.95	Y	2466	19.14	Y	-0.08%	-12.27%	n/a	+20%
7	3	511	80	154.06	6084	76.56	Y	6492	20.73	Y	6162	22.33	Y	+1.28%	-5.08%	-71%	+8%
8	3	714	100	152.39	10311	35.09	Y	10852	22.68	Y	10059	25.52	Y	-2.44%	-7.29%	-27%	+12%
9	3	1111	200	161.05	19932	38.28	Y	21423	19.86	Y	20016	22.33	Y	+0.42%	-6.57%	-42%	+12%
Avg														+0.93%	-7.60%	-46%	+24%

causes thermal violations. On the contrary, *micro-channel first* approach, though saves pumping power, results in up to 15% wirelength increase compared with *TSV first* approach. Our approach considers both wirelength and pumping power simultaneously. The wirelength increase in our approach compared with *TSV first* approach is only 0.93%. In some benchmarks, our approach even results in slightly better WL than *TSV first* approach, this is because both approaches use the bounding box wirelength (which basically gives a lower bound of the routing wirelength) when solving the TSV assignment problem, while the real routing result also depends on the relative positions between interlayer net terminals and TSV locations. Therefore, in these benchmarks, although our approach results in slight degradation in bounding box wirelength, its real routing wirelength is better than *TSV first* approach. Comparing the micro-channel pumping power, our approach achieves 46% pumping power savings compared with *TSV first* approach, and uses 24% more pumping power compared with *micro-channel first* approach. Moreover, for benchmarks where thermal violations occur using *TSV first* approach, using our approach could reduce the temperature below thermal constraints without consuming excessive pumping power.

The runtime of our approach ranges from $2sec$ to $200sec$ on Matlab depending on the benchmark sizes.

6.2 Tradeoff between wirelength and pumping power

The value of criticality factor $g(i, j)$ could be adjusted to control the weight between wirelength and cooling provided by micro-channels. Usually, decrease in pumping power is at the cost of increased wirelength, and vice versa. Such tradeoff is illustrated in figure 8, which shows the wirelength versus pumping power for one benchmark (all data points satisfy the thermal constraints). When thermal violations occur, more efficient allocation of micro-channels could be adopted by sacrificing some wirelength, or more channels are allocated in the unused regions surrounding the hotspot which leads to an increase in pumping power. When pumping power is too high, we could try to better allocate micro-channels to improve its cooling effectiveness at a cost of longer wirelength. When wirelength is more preferable, we could assign TSVs towards further reduction of wirelength while sacrificing micro-channel cooling effectiveness (leading to higher pumping power).

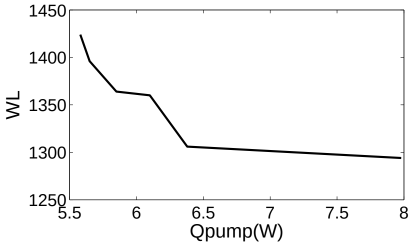

Figure 8: Tradeoff between wirelength and pumping power

7. CONCLUSION

This paper proposed a co-optimization approach that assigns TSVs to interlayer nets and allocates micro-channels simultaneously, such that both the wirelength and micro-channel cooling energy is co-optimized. We developed MCMCF based formulation and simplifying transformation heuristics. The experimental results show that, our approach achieves 46% cooling power savings or 8.3% wirelength reduction compared with the approaches that assign TSVs and micro-channel separately.

8. REFERENCES

[1] X. Liu, Y. Zhang, G. Yeap, and X. Zeng, “An integrated algorithm for 3D-IC TSV assignment,” in *Design Automation Conference (DAC'11)*, pp. 652–657, 2011.

[2] J.-T. Yan, Y.-C. Chang, and Z.-W. Chen, “Thermal via planning for temperature reduction in 3D ICs,” in *IEEE International SOC Conference (SOCC'10)*, pp. 392–395, 2010.

[3] T. Zhang, Y. Zhan, and S. S. Sapatnekar, “Temperature-aware routing in 3D ICs,” in *Asia and South Pacific Design Automation Conference (ASP-DAC'06)*, pp. 309–314, 2006.

[4] D. H. Kim, K. Athikulwongse, and S. K. Lim, “A study of through-silicon-via impact on the 3D stacked IC layout,” in *IEEE/ACM Intl. Conf. on Computer Aided Design (ICCAD'09)*, pp. 674–680, 2009.

[5] M. S. Bakir, C. King, and et al, “3D heterogeneous integrated systems: Liquid cooling, power delivery, and implementation,” in *IEEE Custom Intergrated Circuits Conf.*, pp. 663–670, 2008.

[6] Y. J. Kim, Y. K. Joshi, and et al, “Thermal characterization of interlayer microfluidic cooling of three dimensional integrated circuits with nonuniform heat flux,” *ASME Trans. Journel of Heat Transfer*, 2010.

[7] J.-M. Koo, S. Im, L. Jiang, and K. E. Goodson, “Integrated microchannel cooling for three-dimensional electronic circuit architectures,” *Journel of Heat Transfer*, pp. 49–58, 2005.

[8] H. Mizunuma, C. L. Yang, and Y. C. Lu, “Thermal modeling for 3D-ICs with integrated microchannel cooling,” in *IEEE/ACM Intl. Conf. on Computer Aided Design*, pp. 256–263, 2009.

[9] D. B. Tuckerman and R. F. W. Pease, “High-performance heat sinking for VLSI,” *IEEE Electron Device Letters*, pp. 126–129, 1981.

[10] B. Shi, A. Srivastava, and P. Wang, “Non-uniform micro-channel design for stacked 3D-ICs,” in *Design Automation Conference (DAC’11)*, 2011.

[11] A. Sridhar, A. Vincenzi, M. Ruggiero, T. Brunschwiler, and D. Atienza, “3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling,” in *IEEE/ACM Intl. Conf. on Computer Aided Design (ICCAD’10)*, 2010.

[12] B. Goplen and S. Sapatnekar, “Thermal via placement in 3D ICs,” in *International Symposium on Physical Design (ISPD’05)*, pp. 167–174, 2005.

[13] “Ibm-place 2.0 benchmark,” in *http://er.cs.ucla.edu/benchmarks/ibm-place2/*.

[14] “Labyrinth global router,” in *http://cseweb.ucsd.edu/ kastner/research/labyrinth/*.