THE INSTITUTE FOR SYSTEMS RESEARCH

ISR TECHNICAL REPORT 2013-07

Fluidic Cooling and Gate Size Co-optimization in 3D-ICs: Pushing the Power-Performance Limits

Bing Shi and Ankur Srivastava

Fluidic Cooling and Gate Size Co-optimization in 3D-ICs: Pushing the Power-Performance Limits

Bing Shi and Ankur Srivastava
University of Maryland, College Park, MD, USA
{bingshi, ankurs}@umd.edu

***Abstract*—The performance improvement of modern computer systems is usually accompanied by increased computational power and thermal hotspots, which in turn limit the further improvement of system performance. In 3D-ICs, this thermal problem is significantly exacerbated, necessitating the need for active cooling approaches such as micro-fluidic cooling. This paper investigates a co-optimization approach for 3D-IC electric (gate sizing) and cooling design that fully exploits the interdependency between power, temperature and circuit delay to push the power-performance tradeoff beyond conventional limits. We propose a unified formulation to model this co-optimization problem and use an iterative optimization approach to solve the problem. The experimental results show a fundamental power-performance improvement, with 12% power saving and 16% circuit speedup.**

I. Introduction

A. 3D-IC's Thermal Challenge

3D-ICs, with improved integration density and reduced interconnect delay, have become a significant approach to fulfilling the growing demand for system performance and energy efficiency. 3D-ICs comprise several layers of active electronic components that are stacked vertically. Despite the performance improvement, stacked 3D structure also brings new challenges to chip thermal management. Temperature has always been a limiting factor in achieving higher computing performance. The problem of heat removal is significantly exacerbated in 3D-ICs [3]. Firstly 3D-ICs enable considerable increase in device counts thereby resulting in higher power density. 3D-ICs, especially those incorporating high power heterogeneous technology (with analog, RF and digital components together) are expected to dissipate $500 - 1k$W of power. Secondly the stacked layer configuration could result in overlapped hotspots and higher thermal resistance to the heat sink due to greater number of layers in between, including dielectrics with poor thermal conductivity.

Recent works attempt to address the 3D-IC thermal challenge by either thermal aware design approaches such as placement/floorplanning [5], or improving the thermal conductivity from the inner layers by using dummy thermal through-silicon-vias (TSVs) [7]. However such approaches still rely on conventional air cooling where heat is conducted through several layers of silicon, metal and oxide into heat sink which is cooled via air flow. As the power density increases, air cooling would be unable to deliver the cooling demands of high performance, heterogeneous 3D-ICs [3][13].

B. Interlayer Micro-fluidic Cooling Technology

Several recent approaches have proposed use of interlayer micro-channels in 3D-ICs for addressing the heat removal challenge [3][6]. Physical structures such as lateral micro-channels are embedded in the interlayer regions as illustrated in Fig. 1(b) which carry cooling fluid in the close vicinity of hotspots. This configuration has many advantages including a) significantly higher heat removal rate due to superior properties of the coolant (usually deionized water) [17], b) localized cooling due to close proximity of heat removing and heat generating entities. While micro-channels do not conflict with the gates in active layers, they do conflict with TSVs used for interlayer communication. Another overhead is the pumping power associated with pushing coolant through the channels (see Fig. 1(a)). Fig. 1(c) shows the pumping power versus chip power for *unoptimized* micro-channel design (channels placed all over the interlayer regions). As chip power increases, the required pumping power increases very fast.

Recent works have investigated techniques for co-fabrication of 3D-ICs and interlayer channels. The 3D-IC and micro-channel manufacturing process and overhead were investigated [3][9].Some existing research addresses the design and optimization of the fluidic channel configuration for achieving maximum cooling effectiveness [4][12][14]. These works typically assume that the electronic aspects of the design have been completed and use the associated power dissipation levels to optimize the cooling system. Several researchers are also investigating thermal modeling of 3D-IC with micro-fluidic cooling [16]. Recently, micro-channel is also incorporated in dynamic thermal management for enhancing runtime thermal control [6].

C. Motivation: Simultaneous Gate Sizing and Micro-channel Distribution

Conventional approaches addressing the optimization of interlayer micro-channel structures usually assume that the electronic aspect of 3D-IC design is finished. This causes several sub-optimalities. Distribution of channels in the interlayer regions can be controlled to favor some sub-regions over others. As shown in Fig. 1(b), the distribution of channels can be used to control the local temperature of 3D-IC subregions, while conventional air cooling doesn't support localized cooling. This *localized thermal control* enabled by proper distribution of channels (higher channel density in some areas over lower channel density in others) offers several advantages to the 3D-IC design process, which are ignored by the conventional *postfix* approach for design of the cooling system.

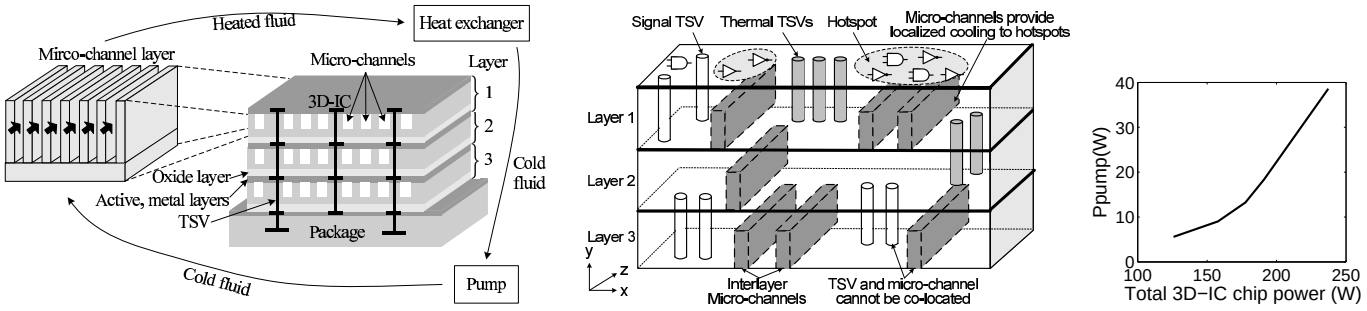
The power, performance and temperature aspects of 3D-ICs have a very complex interdependence. Temperature profile depends on both the amount as well as distribution of power. Non-linear leakage thermal interdependence implies that higher temperature leads to greater power. Higher temperature also impacts the device performance. Addressing these complex interdependencies between power, temperature and performance has been a major focus of research both for 2D and 3D ICs. Localized temperature control enabled by micro-channel distribution can be exploited in a number of ways by the 3D-IC design optimization process.

Improving circuit speed: Allocation of greater cooling surrounding timing critical areas could be used by 3D-IC design methods to improve timing (by aggressive timing optimization), since the associated power could be addressed by greater cooling. Reduced temperature would also contribute to an overall speeding up of circuit.

Reducing dynamic and leakage power dissipation: Greater cooling in high leakage areas would directly reduce their leakage levels due to non-linear dependence between leakage and temperature. Reduction in temperature around timing critical circuits would result in an overall speeding up of the design. Hence we do not need aggressive timing optimization, which helps saving both dynamic and leakage power (for example by down sizing gates). Reduction in power would further reduce temperature, causing a favorable positive feedback. The reduction in chip power may be significantly greater than the required pumping power (experimental results to support this claim would be provided subsequently). Hence the total power of 3D-IC including dynamic, leakage and pumping would be reduced.

Reduction in pumping power: Design of 3D-IC would decide the nature of power dissipation and hotspots. Co-optimization of the 3D-IC system and micro-channel distribution could be used to simplify the cooling configuration and therefore save pumping power.

Fundamental advancement in power-performance tradeoff: Per



(a) (b) (c)

Fig. 1. (a) 3D-IC with micro-channel, (b) micro-channel and TSV configuration, (c) pumping power versus chip power for unoptimized micro-channel design

the advantages noted above, co-optimization of cooling and 3D-IC design enables better performance under a given power envelope and better power for a given performance constraint, thereby resulting in fundamental power-performance improvement. Experimental data to support this claim is illustrated subsequently.

In this paper we attempt to highlight the need for this co-design and the associated challenges and opportunities. We investigate the simultaneous gate sizing and micro-channel distribution problem in 3D-ICs as an illustration of the advantages of this co-optimization.

D. Summary of Contributions

In our problem, we assume a 3D-IC design whose gates have been placed in active layers and the TSV locations have been decided. In this design we perform simultaneous gate sizing as well as allocation of channels in the interlayer regions such that: a) timing constraint is met, b) overall power including dynamic, leakage and pumping power is minimized, and c) micro-channels do not conflict with TSVs.

This is a very challenging problem because it seeks to unify the power, performance, thermal and cooling aspects of the optimization problem. We develop an iterative optimization approach. In the first step we optimally solve the *ideal* case where perfect control of silicon temperatures is assumed. This design is used as a guideline to generate a target channel distribution. The channel distributions as well as gate sizes are iteratively refined to further save power and improve performance. In each step, we exploit the mathematical properties to generate the solution efficiently.

To demonstrate the benefit of 3D-IC and fluidic cooling co-design, we compare our approach with the conventional thermal-aware gate sizing approach, which does not use micro-fluidic cooling. The experimental results show significant improvement of power and performance, with 12% power saving and 16% circuit speedup. Note that gate sizing is just an example to illustrate the power of simultaneous 3D-IC design and cooling co-optimization. Future work would investigate more fundamental aspects of this co-design and how it affects other circuit optimization approaches for 3D-ICs.

The paper is organized as follows. In section 2, we introduce the 3D-IC structure,fundamentals of micro-fluidic cooling, gate delay and power consumption. Section 3 gives the formulation of the gate sizing and micro-channel placement co-optimization. We explore the algorithm for gate sizing and micro-channel placement co-optimization in section 4. The experimental results are given in section 5.

II. Background

A. 3D-IC with Interlayer Micro-Fluidic Cooling

Fig. 1(a) shows a 3D-IC integrated with micro-channel heat sinks. In this 3D-IC, three active layers, which contain functional units such as cores, caches, are stacked vertically. Micro-channels are embedded in the interlayer regions. Each channel spans the whole interlayer region in z direction as Fig. 1(b) shows. Liquid is pumped through channels, and takes away the heat generated in the active layers. TSVs are incorporated to enable interlayer communications. As Fig. 1(b) shows, TSVs also travel through interlayer regions, causing resource conflict with micro-channels.

The thermal behavior of 3D-IC with micro-fluidic channels can be modeled as a distributed RC circuit, with R corresponding to thermal resistance and C indicating the ability to store heat [15][16]. In many cases, people are mostly interested in the steady state thermal behavior of 3D-IC, enabling us to capture the thermal behavior as a pure resistive network [11]. Given the 3D-IC thermal resistive network, the interdependency between chip power and temperature can be modeled by $\mathbf{G} \cdot \vec{T} = \vec{P}$. Here $\mathbf{G}$ represents the thermal conductance matrix, which depends on the material properties, configuration of micro-channels, and TSV distribution, etc. $\vec{T}$ and $\vec{P}$ are the 3D-IC thermal and power profiles. Note that the power $\vec{P}$ is the sum of both dynamic and leakage power.

The micro-channels consume extra power for performing chip cooling. The cooling power basically comes from the work done by the fluid pump to push the coolant through micro-channels. It depends on the coolant flow rate through micro-channels f, the pressure drop across micro-channels Δp as well as micro-channel count N: $P_{pump} = Nf\Delta p$. These three parameters, together with the micro-channel distribution, also decide the cooling effectiveness of micro-channels. In this work, we assume the pressure drop and flow rate are fixed, since they usually depends on the pump configuration. Hence the cooling effectiveness and pumping power is decided by the count and distribution of micro-channels. Increase in micro-channel count results in better cooling, at a cost of increased pumping power.

B. Manufacturing Overhead of Micro-channels

Micro-channels are placed in the silicon substrate between two active layers as Fig. 1(a) shows. Hence the existence of micro-channels does not have direct impact on the placement of gates or wire routing, since they are allocated in different layers. However, TSVs travel through interlayer regions, hence the micro-channels and TSVs have potential resource conflict. When placing micro-channels, such constraint should be considered.

C. Gate Delay Model

The delay of a gate depends on many circuit parameters such as gate sizes, threshold voltage and carrier mobility. Many works model the gate delay as a posynomial function of the sizes (of itself and all its fanouts) [8]. Temperature also influences the gate delay [18]. [10] models the dependency of gate delay on temperature as a polynomial function $d \propto T^{\sigma}$. By incorporating impact of both gate sizes and temperature, we can model the gate delay as:

$$d_i = T_i^{\sigma} \cdot (\eta_{0i} + \frac{\sum_{\forall k \in FO(g_i)} \eta_{ki} \cdot w_k}{w_i}) \quad (1)$$

Here w_i, T_i are the width and temperature of gate g_i, w_k is the width of g_i's fanout gates, σ, η_{0i} and η_{ki} are constants.

Eq. 1 shows that change in the following parameters can result in gate delay reduction: (a) increase of its own width, (b) decrease in the width of its fanouts, and (c) reduction in temperature.

D. Dynamic and Leakage Power Models

The power dissipation also depends on gate sizes and temperatures. For each gate g_i, its dynamic power can be modeled as a linear function of gate size:$P_{d,i} = A_i w_i F$, where w_i is the gate size, F is the clock frequency and A_i depends on the switching activity and supply voltage, etc. In this work, we assume A_i is constant.

The leakage power depends on both gate size w_i and temperature T_i. [10] models the leakage-temperature dependency as $P_{l,i} \propto \beta_1 T_i^2 e^{-\frac{\beta_2}{T_i}} + \beta_3$. We found that this exponential function can be approximated as a quadratic function with very good accuracy, hence in this work we use quadratic leakage model: $P_{l,i} = w_i \cdot (\varepsilon_1 T_i^2 +$

$\varepsilon_2 T_i + \varepsilon_3$), where $\varepsilon_{1,2,3}$ are constants obtained by quadratic fitting of the exponential leakage model in [10]. Note that [19] also verified the accuracy of quadratic leakage model.

According to the power models, large gate size will result in higher dynamic and leakage power, which leads to temperature increase. Temperature increase will lead to further increase in leakage power.

III. Problem Formulation

Given a 3D-IC circuit and the associated gate and TSV placement (as Fig. 1(b) shows), we would like to decide the size of all gates and location of interlayer micro-channels such that the total power consumption (including the dynamic and leakage power, as well as the pumping power consumed by micro-channels) is minimized, while at the same time minimizing the longest path delay and ensuring silicon temperature to be less than the maximum constraint. The channels should not come in conflict with TSVs, which have been placed already. The co-optimization problem is formulated in Eq. 2. Here we assume that gates and TSVs have been placed on a grid (each gate/TSV is within a grid). Also the gate sizing does not change the gate's grid location. Note that these assumptions are similar to other works dealing with in-place gate sizing.

$$\begin{aligned} &\textbf{Decision variables}: \vec{w},\ \mathbf{B} \\ &min \sum_{\forall gate: g_i} (P_{d,i} + P_{l,i}) + P_{pump} \\ &s.t.\quad 1.\quad t_j + d_i(\vec{w}, T_i) \le t_i, \forall \text{gate } g_i, g_j \in FI(g_i) \\ &\qquad 2.\quad t_i < t_{con}, \forall \text{gate } g_i \in PO \\ &\qquad 3.\quad \mathbf{G}(\mathbf{B}) \cdot \vec{T} = \vec{P}(\vec{w}, F, \vec{T}) \\ &\qquad 4.\quad 0 \le \vec{T} \le \vec{T}_{max} \\ &\qquad 5.\quad w_{min} \le w_i \le w_{max}, \forall \text{gate } g_i \end{aligned} \tag{2}$$

The decision variables in this problem are the gates size $\vec{w}$ and micro-channel locations $\mathbf{B}$.

The objective of the optimization problem is to minimize the total power consumption of the 3D-IC (including dynamic, leakage and pumping power) for the given timing constraint t_{con}. Here $P_{d,i}$ and $P_{l,i}$ represent the dynamic and leakage power of gate g_i, which can be calculated based on the models in section II-D. The dynamic power depends on the gate sizes $\vec{w}$ and clock frequency F, and leakage power depends on both gate sizes $\vec{w}$ and thermal profile $\vec{T}$ (temperature in all grids). The clock frequency is usually decided by the maximum circuit delay. Hence, in this work, we assume the clock frequency is the inverse of timing constraint $F = 1/t_{con}$.

The first two constraints are timing constraints, indicating that the signal propagation delay from the primary inputs (PIs) to primary outputs (POs) should be within the timing constraint t_{con}. Here t_i denotes the signal arrival time at the output of gate g_i from the primary inputs and d_i is the propagation delay of gate g_i. The delay, which depends on gate sizes and temperature, is calculated using the model in Eq. 1. We assume the 3D-IC is divided into grids. For ease of explanation, we assume each grid only contains one gate. Hence grid i contains gate g_i and has the temperature T_i. If a grid does not have a gate, the corresponding power is 0 and the temperature would be decided by neighboring grids based on the conductivity matrix $\mathbf{G}$. The 3D-IC thermal profile $\vec{T}$ is then represented by the temperature of all grids: $\vec{T} = \{T_{i,\forall grids:i}\}$. Note that this formulation is easily extendable to the case where each grid contains multiple gates.

The third constraint indicates the interdependency between temperature and power. Let $\vec{T}$ and $\vec{P}(\vec{w}, F, \vec{T})$ represent the thermal and power profile at all grids i in 3D-IC. The power dissipated in a grid i is $P_i = P_{d,i} + P_{l,i}$ (if a grid does not have any gate then its power is 0). Note that the power profile is a function of gate sizes and temperatures. Here $\mathbf{G}$ represents the 3D-IC conductivity matrix which depends on the properties of the material, TSVs as well as design of the micro-channel structure $\mathbf{B}$. The last two constraints are the maximum temperature constraint and feasible gate size range.

The power, temperature and gate delay are interdependent in a complex way (as the models in section II shows), making this co-optimization problem difficult to solve. The allocation of micro-channels at discrete locations adds further complexity to this problem.

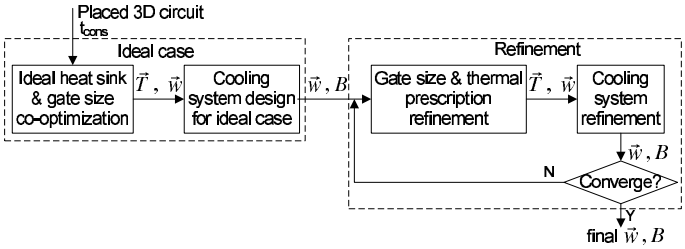


Fig. 2. Overall design flow

IV. Gate Sizing and Micro-channel Placement Co-optimization Algorithm

The problem formulation illustrated above is quite complex. We develop an iterative optimization approach where each step systematically solves some aspects of the problem. We have strived to use rigorous optimization methods as much as possible. Fundamentally the overall optimization problem is decomposed into two: deciding the gate sizes and grid temperatures simultaneously and then designing the micro-channel distribution which removes the heat generated by the circuit (function of temperature and gate size) while coming as close as possible to the prescribed temperature. This process is iterated several times as summarized below.

Step 1: Ideal heat sink and gate size co-optimization: We first simplify the problem by assuming that temperature in each grid is perfectly controllable and is not dependent on the 3D-IC conductivity matrix $\mathbf{G}$. The resulting solution allocates a gate size and temperature level to each gate/grid. The ideal case acts as a guideline to following optimization steps which would then strive to get as close to this ideal solution as possible.

Step 2: Micro-channel distribution for the ideal case: Interlayer micro-channels are now placed such that: a) the heat levels decided by step 1 are effectively removed and the grid temperatures are as close to those prescribed by step 1 as possible, b) micro-channels are not allocated in areas with TSVs, and c) smallest number of channels are allocated for minimal pumping power.

Step 3: Gate size and grid temperature refinement: Since step 2 will be unable to entirely meet the ideal case solution of step 1, the gate size and grid temperature solution needs to be refined to account for the current micro-channel network in place.

Step 4: Micro-channel distribution refinement: The solution from step 3 gives a modified gate size and grid temperature prescription. Hence the micro-channel network needs to be refined further.

Step 5: Iterate steps 3 and 4 till convergence criteria is met: The convergence criteria could be set to a maximum number of iterations or levels of improvements achieved.

Fig. 2 illustrates the overall approach. In each step we strive to use algorithms and heuristics which draw upon rigorous optimization theory while exploiting the structure in the problem formulation. Now we describe each step in detail.

A. Step 1: Ideal Heat Sink and Gate Size Co-optimization

Let us first simplify the optimization problem in Eq. 2 as:

$$\begin{aligned} &\textbf{Decision variables}: \vec{w},\ \vec{T} \\ &min \sum_{\forall gate: g_i} (P_{d,i}(w_i) + P_{l,i}(w_i, T_i)) + \lambda \sum_{\forall grid: i} \frac{1}{T_i} \\ &s.t.\quad 1.\quad t_j + d_i(\vec{w}, T_i) \le t_i, \forall \text{gate } g_i, g_j \in FI(g_i) \\ &\qquad 2.\quad t_i < t_{con}, \forall \text{gate } g_i \in PO \\ &\qquad 3.\quad 0 \le \vec{T} \le \vec{T}_{max} \\ &\qquad 4.\quad w_{min} \le w_i \le w_{max}, \forall \text{gate } g_i \end{aligned} \tag{3}$$

In this formulation, the grid temperature T_i is assumed to be perfectly controllable through an ideal heat sink. The constraints signify meeting the timing constraint while staying with temperature and gate size constraints. The objective has two components: minimization of power as well as an additional term $\sum_{\forall grid:i} \frac{1}{T_i}$. This term signifies the fact that reducing T_i comes at the penalty of a more complex heat sink (which would be designed in the subsequent steps). Without this

term, this optimization problem would trivially assign all T_i to be as small as possible (because that would benefit both timing and power). The solution of this problem represents allocation of gate sizes along with grid temperature, and would be used as a starting point for further optimization.

In order to solve this problem we make the following transformation $w_i = e^{a_i}$ and $T_i = e^{b_i}$. Based on this transformation, the gate delay and power consumption models described in section II become: $d_i = e^{\sigma b_i} \cdot (\eta_{0i} + \sum_{\forall k \in FO(g_i)} \eta_{ki} \cdot e^{a_k - a_i})$, $P_{d,i} = A_i F e^{a_i}$, $P_{l,i} = e^{a_i} \cdot (\varepsilon_1 e^{2b_i} + \varepsilon_2 e^{b_i} + \varepsilon_3)$. It can be seen that the models for delay, leakage and dynamic powers are convex functions of variables a_i and b_i.

Theorem 1: Formulation is Eq. 3 can be solved optimally using convex optimization approaches.

Proof: As indicated, gate delay, dynamic and leakage power functions are convex w.r.t. variables a_i and b_i. Hence the constraints are convex. The term $\sum_{\forall grid:i} \frac{1}{T_i}$ gets transformed to $\sum_{\forall grid:i} e^{-b_i}$ which is a convex function, too. Hence the overall objective function is convex as well, making the whole formulation optimally solvable using polynomial time convex methods. ■

B. Step 2: Micro-channel Distribution for Ideal Case

Step 1 has assigned gate sizes and grid temperature values. The gate sizes and temperatures decide the overall power dissipation profile while the temperature assignments indicate the level of cooling necessary in each grid. Together, these two aspects profoundly impact the design of the interlayer micro-fluidic system. The problem with the "ideal formulation" of step 1 is that it assumes perfect control of each grid temperature which is not possible even with interlayer micro-fluidics. By nature, micro-fluidic channels carry heat along the direction of fluid flow. They are incapable of controlling grid level temperatures. This is because, even though they enable localized cooling, they cannot completely remove the thermal cross-coupling of neighboring grids. The decision of allocating or removing a micro-channel will influence all the grids adjacent to this micro-channel. Hence in this step, we would like to allocate channels such that the power dissipation levels are removed while ensuring the grid temperatures are as close as possible to the prescribed levels from step 1. We use least square fit (LSF) to find the micro-channel placement:

$$\min \quad \| \mathbf{G}(\mathbf{B}) \cdot \vec{T}_{desire} - \vec{P}_{desire} \|_2 \tag{4}$$

Here $\vec{T}_{desire}$ is the prescribed thermal profile $\vec{T}$ decided by the previous step. $\vec{P}_{desire}$ is the sum of dynamic and leakage power calculated based on the prescribed gate sizes and temperatures using the power models in section II-D. The objective is to decide the channel allocation such that the RMS (root-square-mean) error is minimized. $\mathbf{B}$ is the allocation of micro-channels and $\mathbf{G}(\mathbf{B})$ is the associated thermal conductivity matrix. For a given allocation of micro-channels, the associated conductivity matrix could be generated using the modeling approach described in section II-A. It is noteworthy that for a given set of potential channel locations $\mathcal{I}$, we would like to choose a subset such that the aforementioned objective is minimized.

To solve this, we first formulate the problem as an integer program. Essentially we assign a decision variable for each potential micro-channel location (binary constraint) and show that the conductivity matrix $\mathbf{G}$ is a linear function of these binary variables (proofs are omitted for brevity). By approximating the binary variables as continuous, this problem becomes minimizing the RMS error of an affine function (since $\vec{T}_{desire}$ and $\vec{P}_{desire}$ are known, $(\mathbf{G}(\mathbf{B}) \cdot \vec{T}_{desire} - \vec{P}_{desire})$ is a linear function of $\mathbf{B}$), which can be solved efficiently. After solving this problem, we roundup the continuous variables to obtain the locations of micro-channels. Note that the objective here is to generate a fluidic cooling solution that come as close as possible to the prescribed $\vec{T}_{desire}$ and $\vec{P}_{desire}$.

C. Step 3: Gate Size and Grid Temperature Refinement

Since the micro-channel solution from step 2 may not be able to come very close to the solution desired by step 1, we need to refine the original solution. Following are the objectives of this refinement step. 1) Step 2 synthesized a micro-channel solution which controls how power and temperature impact each other. This needs to be accounted for in the gate sizing solution. The ideal case of step 1 had assumed a perfectly controllable grid temperature. With the new channel infrastructure in-place, this assumption does not hold anymore. Hence the gate sizing needs to be re-evaluated. 2) We may still want to refine the channel structure further, based on newly prescribed temperature and gate sizes. Hence we would like to generate new assignments from grid temperature while accounting for the current cooling system in place.

In order to achieve the latter objective we divide the temperature T_i into two components: controllable and uncontrollable parts, $T_{c,i}$ and $T_{nc,i}$. The uncontrollable temperature is decided by the relationship between power and temperature which is a function of gate sizes and also the micro-channel structure in place. The controllable part is an additional parameter which we can control to prescribe any change in temperature. It would be used to further refine the micro-channel structure. The gate/grid temperature $T_i = T_{nc,i} * T_{c,i}$. Here $T_{c,i} = 1$ indicates no change at gate g_i (or grid i), $T_{c,i} < 1$ indicates greater need for cooling and $T_{c,i} > 1$ indicates less cooling necessary. The formulation at this step can be represented as follows.

$$\begin{aligned} &\textbf{Decision variables} : \vec{w},\ \vec{T}_{nc},\ \vec{T}_c \\ &Objective : \\ &min \sum_{\forall gate:g_i} (P_{d,i}(w_i) + P_{l,i}(w_i, T_{nc,i} * T_{c,i})) + \lambda \sum_{\forall grid:i} \frac{1}{T_{c,i}} \end{aligned} \tag{5}$$

The objective structure is the same as the ideal case in step 1. However, the temperature affecting the gate leakage has two components now: uncontrollable part $T_{nc,i}$ and controllable part $T_{c,i}$. Because the controllable component is being assigned by us in this step, we would like $T_{c,i}$ to be as large as possible indicating minimal need for channels. This would help reduce pumping power. Hence the objective combines total power dissipated (the first two terms) along with pumping power (the third term).

$$\begin{aligned} &Constraints\ 1,\ 2 : \\ &1.\quad t_j + d_i(\vec{w}, T_{nc,i} * T_{c,i}) \le t_i, \forall \text{gate } g_i, g_j \in FI(g_i) \\ &2.\quad t_i < t_{con}, \forall \text{gate } g_i \in PO \end{aligned} \tag{6}$$

This set of timing constraints (constraints 1 and 2) is similar to the ideal case except the gate temperature has two components.

$$Constraint\ 3 : \quad \mathbf{G}(\mathbf{B}) \cdot \vec{T}_{nc} = \vec{P}_d(\vec{w}) + \vec{P}_l(\vec{w}, \vec{T}_{nc}) \tag{7}$$

As indicated earlier, $T_{nc,i}$ is the uncontrollable temperature which is decided by the power being dissipated and also the cooling system in place. This constraint establishes the relationship between chip power dissipation and $T_{nc,i}$. Note that we do not include $T_{c,i}$ in this equation, because this parameter is being controlled to prescribe refinements in the cooling system, and would be used by future steps to redesign the cooling system.

Unlike the ideal case in step 1, $T_{c,i}$ should not be arbitrarily assigned in each grid since we already have a micro-channel network in place. For example, if a grid i already has a channel underneath, then increasing $T_{c,i}$ would prescribe removal of this channel. But doing so without accounting for the impact on other grids may result in significant sub-optimality since removal of a channel would affect a large number of grids. Also, if a grid i is located close to a TSV, then even if it has a small value of $T_{c,i}$ (indicating a need for channels), its extra cooling demands may never be met due to physical constraints imposed by TSVs. To account for these issues, the following constraints are imposes on the control of $T_{c,i}$.

$$\begin{aligned} &Constraints\ 4,\ 5 : \\ &4.\quad \vec{T}_{c,min} \le \vec{T}_c \le \vec{T}_{c,max} \\ &5.\quad T_{c,i} = T_{c,j}, \forall \text{adjacent grids } i,j \text{ along channel direction} \end{aligned} \tag{8}$$

$T_{c,min,i}$ and $T_{c,max,i}$ values control how the $T_{c,i}$ values are allocated ($\vec{T}_{c,max}$, $\vec{T}_{c,min}$ are vectorized $T_{c,max,i}$, $T_{c,min,i}$). $T_{c,min,i} \le 1$ and $T_{c,max,i} \ge 1$. A small value of $T_{c,min,i}$ implies the possibility of adding more cooling around grid i, while a large value of $T_{c,min,i}$ implies smaller chance of adding extra cooling around i. Similarly, a large value of $T_{c,max,i}$ implies that grid i is close to some existing channels, hence great temperature increase would occur if the cooling around grid i is removed. A small value of $T_{c,max,i}$ implies that the impact of existing cooling configuration on grid i is small since they are far away. By appropriately assigning the values for $T_{c,min,i}$ and

$T_{c,max,i}$, we can control the degree of change that is prescribed to the cooling system by the optimization formulation. The $T_{c,min,i}$ and $T_{c,max,i}$ values for each $T_{c,i}$ are allocated using the following rules.

Rule 1: If grid i is in the close vicinity of a TSV, then allocating channels nearby would be tougher. Hence we do not wish to have too much additional control of temperature at grid i. Therefore, $T_{c,min,i}$ and $T_{c,max,i}$ are allocated to be closer to each other such that significant changes in the fluidic structure around i is not prescribed by the optimization formulation. We use a formula based on distance and number of closeby TSVs to compute this range. The details have been omitted for brevity.

Rule 2: If a channel is already allocated very close to grid i, then $T_{c,min,i}$ is assigned to 1 and $T_{c,max,i}$ is assigned to be a large value. This indicates that the step 3 formulation only has the option of suggesting removal of a channel from this location.

Rule 3: If a channel is allocated close but not too close to a grid i, then $T_{c,min,i} < 1$ and its value is a function of the number of potential channel locations in the close vicinity. More the potential channel locations, smaller the value of $T_{c,min,i}$. $T_{c,max,i}$ is allocated to be a value greater than 1, and is a function of the distance to the closest channel in the *current* design. Greater the distance smaller the value of $T_{c,max,i}$. This is because, prescribing an increase in grid temperature by removing channels will only be effective if they are located sufficiently close (further details omitted for brevity).

Rule 4: If no channel is allocated in sufficient vicinity then $T_{c,min,i}$ has the smallest value *possible* indicating that a channel could be added and $T_{c,max,i} = 1$ indicating that there is little possibility of removal of a channel.

Rule 5: All $T_{c,i}$ for the grids along the same micro-channel is allocated to be the same. As shown in Fig. 1(b), each micro-channel spans the whole interlayer region in z direction, hence the prescribed changes for grids along the same micro-channel are assigned be the same due to the nature of micro-channels. This is illustrated in constraint 5.

Allocating $T_{c,min,i}$ and $T_{c,max,i}$ values is very critical since the ranges decide what kind of changes from the current fluidic structure end up being prescribed. The rules above attempt to constrain the formulation of step 3 to prescribe changes which are in sync with the current fluidic system in place. Also, as we re-iterate, we would like to make fewer modifications in the micro-channel structure. This could be achieved by reducing the range for $T_{c,i}$ as iterations progress.

Solving this formulation is more complex than the ideal case of step 1. Here too, we transform the temperature $T_{nc,i} = e^{b_{nc,i}}$, $T_{c,i} = e^{b_{c,i}}$, and gate size $w_i = e^{a_i}$. Hence the prescribed temperature $T_i = T_{nc,i} * T_{c,i} = e^{b_{nc,i}+b_{c,i}}$. With this transformation, the gate delay, dynamic and leakage power become convex functions of the gate size and temperature variables $a_i, b_{nc,i}$ and $b_{c,i}$. The objective and constraints 1,2 in Eq. 5, 6 remains convex. Constraints 4 and 5 are also convex (since ranges of the primary variables could be transformed to appropriate ranges of the transformed variables). Constraint 3, however is problematic. In this constraint, $T_{nc,i}$ and power dissipation values are convex functions of a_i and $b_{nc,i}$. However the equality relationship in the constraint causes the convexity to breakdown. In order to address this problem, we represent the the power dissipation of gate g_i (leakage + dynamic) as a piecewise linear function of the gate size parameter a_i and uncontrollable temperature variable $b_{nc,i}$. Note that the right hand side of the constraint is basically the power dissipation for all gates. We also represent $T_{nc,i} = e^{b_{nc,i}}$ (on the left had side) as a piecewise linear function of $b_{nc,i}$. The underlying model parameters could be used to generate the coefficients for the piecewise linearization (these are standard approaches and therefore omitted for brevity). Because, both gate power dissipation and $T_{nc,i}$ are convex functions of a_i and $b_{nc,i}$, the following approach can be used to replace the variables $T_{nc,i}, P_{d,i}, P_{l,i}$ from constraint 3 by the underlying piecewise linearization.

$$\begin{aligned} Power_i &\geq e_{s,1} \cdot a_i + e_{s,2} \cdot b_{nc,i} + e_{s,3} \quad \forall s = 1...S \\ Temp_i &\geq e_{u,1} \cdot b_{nc,i} + e_{u,2} \quad \forall u = 1...U \end{aligned} \tag{9}$$

Here S and U are the number of linearizations imposed on the gate power dissipation and $T_{nc,i}$. Here $Power_i$ represents an upper bound on gate g_i's total power. The S-piecewise linearization is derived from the underlying model. Similarly $Temp_i$ is an upper bound on $T_{nc,i}$. Constraint 3 is now written as:

$$Constraint\ 3: \quad \mathbf{G}(\mathbf{B}) \cdot \vec{T}emp = \vec{P}ower \tag{10}$$

Here $\vec{T}emp$ and $\vec{P}ower$ are vectorized $Power_i$ and $Temp_i$. This modification enables us to linearize constraint 3, which could now be augment with the other constraints and solved with standard convex optimization methods. The final solution of this optimization would be $a_i, b_{nc,i}$ and $b_{c,i}$ values for all gates. These would now be used to refine the micro-channel distribution.

D. Step 4: Micro-channel Distribution Refinement

Just as step 2, we would like to design the micro-channel distribution to address the heat dissipation decided by the gate sizes (and temperature) and also account for the change in the current configuration prescribed by $T_{c,i}$. This step is basically the same as step 2. However there are a few changes. Firstly, the formulation solved in step 3 uses upper bound $Power_i$ and $Temp_i$ as illustrated in Eq. 9, 10. Hence, for a given gate size and micro-fluidic configuration, we will need to recompute the actual uncontrollable thermal profile $\vec{T}_{nc}$ (which could be done by simply solving Eq. 7 for the assigned gate size). Note that this is a complex equation to solve due to leakage thermal interdependence. This would give the actual $\vec{T}_{nc}$ profile for the given gate size solution. Now we combine the actual $T_{nc,i}$ with the prescribed $T_{c,i}$ values to obtain the target grid temperature $T_i = T_{nc,i} * T_{c,i}$. The generated target thermal profile is basically $\vec{T}_{desire}$ from step 2. Since the target thermal profile and gate sizes are known, the chip power profile could be computed as well. This would constitute $\vec{P}_{desire}$. Using these values, a new channel distribution is computed using techniques described in step 2.

E. Step 5: Re-iteration and Stopping Criteria

Steps 3, 4 are iterated to continue improvement in the overall solution. Firstly we would like to point out that the formulation in step 3, indirectly captures pumping power using the term $\lambda \sum_{\forall grid:i} \frac{1}{T_{c,i}}$. Secondly, as we iterate, Eq. 8 controls the tolerable level of change from the current micro-channel allocation. By shrinking the range of $T_{c,i}$ as we iterate, the amount of change in the cooling solution becomes lesser and lesser. Hence after a few iterations, it will converge. This approach unifies the design of cooling structure with gate sizing. This is a significant improvement over conventional approaches that usually design the cooling infrastructure after designing the electrical aspects. In the result section we illustrate how such co-design can fundamentally improve the power-performance tradeoff in 3D-ICs.

F. Computational Overheads and Other Merits

The computational complexity in this approach stems from the algorithmic complexity of the individual steps as well as the number of iterations. We would like to point out that each of the individual steps attempts to make the best use of the fundamental mathematical structure in the problem formulation. For example step 1 is optimally solvable, step 2 is approximated as an unconstrained convex program, step 3 is approximated using the piecewise linearization approach which is then solved optimally and step 4 is similar to step 2. The number of iterations are systematically controlled by appropriately setting the range of $T_{c,i}$ variables.

We believe our approach is a unique way of integrating the electrical, thermal and cooling aspects in a unified optimization approach which is capable of effectively accounting for the complex interdependencies.

V. Experimental Results

In the experiment, we use the ITC'99 circuits, which are typical synthesized circuits consisting of AND, OR, NOT, NAND and NOR gates, to generate the 3D-IC benchmarks [2]. Each 3D-IC contains three layers and each layer contains several arbitrarily chosen ITC'99 circuits. We use the Capo placer to place the gates in each layer [1]. We also place a total of 2000 TSVs in the whitespace. The chip dimension is $W = L = 9mm$. The width, height and pitch of micro-channel are $100, 200$ and $100\mu m$. The maximum thermal constraint is $85°C$. The parameters of delay, thermal and power models are obtained from [10][18][19] and SPICE simulation. Note that we strive to use realistic 3D-IC benchmarks. However, since no real 3D-IC benchmark is publicly available, we try to construct close-to-realistic 3D-IC benchmarks using existing 2D standard benchmarks.

TABLE I
COMPARISON OF TOTAL POWER CONSUMPTION

Bench mark	#Gates	$t_{con}(ns)$ (tight/loose)	Total power (W)			Power saving w.r.t	
			Air Cool	Postfix	Our	Air Cool	Postfix
1	343380	48 (tight)	294	289	254	13.61%	12.11%
		70 (loose)	226	223	197	12.83%	11.66%
2	394152	74 (tight)	256	251	219	14.45%	12.75%
		95 (loose)	233	219	189	18.88%	13.70%
3	342267	70 (tight)	221	218	191	13.57%	12.39%
		90 (loose)	182	189	164	9.89%	13.23%
4	295632	39 (tight)	293	287	258	11.95%	10.10%
		60 (loose)	214	210	189	11.68%	10.00%
5	208575	51 (tight)	284	291	248	12.67%	14.78%
		61 (loose)	251	245	219	12.75%	10.61%
6	181722	55 (tight)	232	232	206	11.21%	11.21%
		75 (loose)	190	188	167	12.11%	11.17%
Average			240	237	208	13.33%	12.05%

TABLE II
COMPARISON OF CIRCUIT PERFORMANCE

Bench mark	Postfix		Our		Circuit speedup
	Best t_{con}(ns)	Power(W)	Best t_{con}(ns)	Power(W)	
1	48	289	40	289	16.67%
2	74	251	60	251	18.92%
3	70	218	57	218	18.57%
4	39	287	34	287	12.82%
5	51	291	44	277	13.73%
6	55	232	47	231	14.55%
Average	56	261	47	259	15.88%

To verify the power and performance improvement achieved by our approach, we compare our design with two other approaches: *Air Cool* and *Postfix* approaches. In the *Air Cool* approach, we perform thermal aware gate sizing with pure air cooling. The overall thermal resistance of the heat sink for air cooling is 0.5°C/W. In the *Postfix* approach, we first perform gate sizing assuming there isn't any micro-channels and then place micro-channel using the approach in [14].

A. Comparison of Power Consumption

We compare the total power consumption resulted from the three approaches. For the *Air Cool* approach, the power consumption consists of dynamic and leakage power, while for *Postfix* and our approaches, the total power consumption also includes the pumping power consumed by micro-channels. Table I shows the power consumption resulted from these approaches. For each benchmark, we tested power consumption for different timing constraints: one is tight and the other is looser. Note the *tight* timing constraint is the best achievable timing constraint for *Air Cool* approach (basically the tightest timing constraint that we can compare). Table I shows that, under the same performance constraint, our approach can result in 13.33% total power savings compared with *Air Cool* approach, indicating that the use of micro-channels, not only does not increase the system total power consumption, but actually helps save power instead. Compared with *Postfix* approach which performs gate sizing and micro-channel placement separately, our co-design approach achieves 12.05% power saving. This is because: a) micro-channel structure is optimized, b) the co-optimization also helps further reducing the leakage power and circuit delay, causing a favorable positive feedback.

B. Comparison of Circuit Delay

We also compare the best achievable circuit delay under the same power envelop. This was obtained by performing a binary search on timing constraints t_{con}. Table II shows that our co-optimized design achieves 15.88% circuit speedup over the *Postfix* approach, while still consuming the same (or even less) amount of power.

C. Power-Performance Tradeoff

To characterize the tradeoff between the system performance and power consumption, we plot the circuit delay versus power consumption for benchmark 1 as Fig. 3 shows. For all three approaches, the power consumption increases as the timing constraint becomes tighter. In the figure, the solid line is the power consumption of conventional gate sizing approach using pure air cooling. This line is basically the best power-delay tradeoff that the conventional gate sizing approach can achieve. The power-performance curve achieved by *Postfix* approach has slight (but not significant) improvement over the conventional gate sizing approach. However, using co-design results in significant power-performance improvement. The figure shows that for all timing constraints we tested, our design always dissipates less power compared with the other two approaches. Similarly, when the available power budget is fixed, our design achieves better circuit speed, indicating a fundamental power-performance improvement achieved by 3D-IC electric and cooling system co-design.

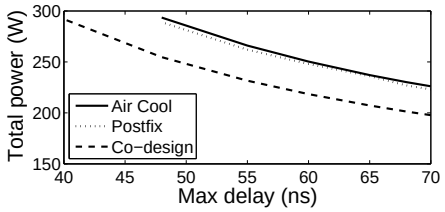


Fig. 3. Delay versus power tradeoff for benchmark 1

VI. CONCLUSION

This paper investigated the co-design of 3D-IC gate sizing and micro-channel allocation that fully exploits the interdependency between power, temperature and circuit delay to reduce power consumption and circuit delay. We show that by performing 3D-IC electrical and cooling system co-design, a fundamental improvement in power-performance tradeoff can be achieved.

REFERENCES

[1] Capo: a large-scale fixed-die floorplacer. *http://vlsicad.eecs.umich.edu/BK/PDtools/Capo/*.

[2] ITC'99 benchmarks. *http://www.cad.polito.it/downloads/tools/itc99.html*.

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

[4] T. Brunschwiler, B. Michel, H. Rothuizen, U. Kloter, B. Wunderle, and H. Reichl. Hotspot-optimized interlayer cooling in vertically integrated packages. *Proc. Materials Research Soc.y Fall Meeting*, 2008.

[5] J. Cong, J. Wei, and Y. Zhang. A thermal-driven floorplanning algorithm for 3D ICs. In *IEEE Int. SOC Conf.*, pages 392–395, 2010.

[6] A. K. Coskun, D. Atienza, T. S. Rosing, and et al. Energy-efficient variable-flow liquid cooling in 3D stacked architectures. In *Conf. on Design, Automation and Test in Europe*, pages 111–116, 2010.

[7] B. Goplen and S. Sapatnekar. Thermal via placement in 3D ICs. In *Int. Symp. on Physical Design*, pages 167–174, 2005.

[8] M. Ketkar, K. Kasamsetty, and S. S. Sapatnekar. Convex delay models for transistor sizing. In *Design Automation Conf.*, pages 655–660, 2000.

[9] C. King, D. Sekar, M. Bakir, B. Dang, J. Pikarsky, and J. Meindl. 3d stacking of chips with electrical and microfluidic i/o interconnects. In *Electronic Components and Technology Conference*, pages 1–7, 2008.

[10] W. Liao, L. He, and K. Lepak. Temperature and supply voltage aware performance and power modeling at microarchitecture level. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Syst.*, 24:1042–1053, 2005.

[11] 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*, pages 256–263, 2009.

[12] M. Sabry, A. Sridhar, and D. Atienza. Thermal balancing of liquid-cooled 3d-mpsocs using channel modulation. In *Conf. on Design, automation and test in Europe*, pages 599–604, 2012.

[13] B. Shi, A. Srivastava, and A. Bar-Cohen. Hybrid 3d-ic cooling system using micro-fluidic cooling and thermal tsvs. In *IEEE Comput. Soc. Annu. Symp. on VLSI*, pages 33–38, 2012.

[14] B. Shi, A. Srivastava, and P. Wang. Non-uniform micro-channel design for stacked 3d-ics. In *Design Automation Conf.*, pages 658–663, 2011.

[15] K. Skadron, M. R. Stan, K. Sankaranarayanan, W. Huang, S. Velusamy, and D. Tarjan. Temperature-aware microarchitecture: Modeling and implementation. *ACM Trans. on Architecture and Code Optimization*, 1:94–125, 2004.

[16] 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*, 2010.

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

[18] N. Weste and D. Harris. Cmos vlsi design: A circuits and systems perspective. *Addison Wesley*, 2010.

[19] C. Yang, J. Chen, L. Thiele, and T. Kuo. Energy-efficient real-time task scheduling with temperature-dependent leakage. In *Conference on Design, Automation and Test in Europe*, pages 9–14, 2010.