Leveraging Hotspots and Improving Chip Reliability via Carbon Nanotube Grid Thermal Structure

Hao Liang, Wei Zhang, Member, IEEE, Jiale Huang, Shengqi Yang, and Pallav Gupta, Member, IEEE

Abstract—The increasing power consumption of integrated circuits (ICs) enabled by technology scaling requires more efficient heat dissipation solutions to improve overall chip reliability and reduce hotspots. Rapidly growing 3-D IC technology strengthens the requirement with more devices stacked per unit area. Thermal interface material (TIM) and MicroChannel are widely adopted strategies to resolve the heat dissipation problem. In recent years, carbon nanotubes (CNTs) have been proposed as a promising TIM due to their superior thermal conductivity. Several CNT-based thermal structures for improving chip heat dissipation have been proposed and demonstrated significant temperature reduction. In this project, we developed an improved CNT TIM structure which includes a CNT grid and thermal vias. It collaborates with MicroChannel to dissipate heat more efficiently in 3-D chips and at the same time, obtain more uniform chip thermal profiles. We present simulation-based experimental results that indicate up to 19.88% peak temperature reduction, 7.81% average temperature reduction, over 66% maximum temperature difference reduction on chip and 17.26% improvement in chip reliability for IBM-PLACE 2.0 circuit benchmarks, showing the effectiveness of our proposed thermal structure for resolving thermal challenge and improving chip reliability in 3-D IC.

Index Terms—Thermal analysis, thermal control.

I. INTRODUCTION

WHILE process scaling has enabled increasing device integration density, it has also resulted in higher power consumption and temperature in integrated circuits (ICs), which affect circuit performance (e.g., reduced threshold voltage $V_{th}$, increased wire resistance), reliability [e.g., negative-bias temperature instability (NBTI), electromigration, dielectric breakdown], and IC packaging. The International Technology Roadmap for Semiconductors predicts that the power dissipation levels of cost-performance and high-performance ICs will reach 1.73 W/mm² and 0.43 W/mm², respectively, by 2022 [1]. The emerging 3-D IC technology further enables compact device integration by utilizing z-index of 3-D space and stacking circuits up to achieve higher performance as well as less interconnection delay [2]–[5]. However, 3-D IC also generates higher power density and significantly increases overall thermal load of the circuits per unit area. Moreover, the heat transfer path from internal active layers to cooling facilities are extended by neighboring layers in 3-D IC [6]. These problems result in stubborn hotspots and fast growing chip temperature which become severe threats to chip reliability.

Cooling solutions such as air fans or liquid-based techniques (e.g., heat pipes) can dissipate in excess of 100 W. However, significant thermal contact resistance between heat sink/spreader and inner device layers in 3-D architecture significantly impacts the efficiency of heat dissipation. Novel thermal interface materials (TIMs) and structures await exploration so that they can meet the ever-increasing packaging demands of future ICs.

In order to dissipate heat more efficiently from inner device layers, MicroChannel-based architectures are widely considered in 3-D IC [7]–[9]. MicroChannels are first described by Tuckerman and Pease nearly 30 years ago to deal with very high power densities [10]. This liquid cooling method is very attractive due to its outstanding heat transfer coefficient. Along with the enhancement in manufacturing and etching technology, MicroChannels provide a good solution to 3-D heat transfer problem. By inserting MicroChannels in between device layers, the heat transfer paths from inner layers to cooling facility are significantly abridged, and heat can be immediately transferred to coolants in MicroChannels and brought outside the chip along the heat carrier flow. MicroChannels also serve as heat buffers between neighboring active layers to minimize the heat accumulation effects of two adjacent heat source layers.

Although MicroChannels fit well in ameliorating the overall temperature condition and mitigating some hotspots brought by stacked dies, they have limited capability in transferring heat in the direction perpendicular to the flow direction. Also, as liquids in the channels heated by device layer, downstream coolants always have a higher temperature than those in the upstream, so the cooling effects of MicroChannels are indeed decreasing along the flow.

TIM in a way compensates this weakness due to its isotropic characteristic tendency to spread heat over all directions. TIMs are usually applied to increase thermal conductance between chip and heat sinks by filling up the air gap in between. In a 3-D architecture, multiple wafers are stacked up with similar interstitial air gaps in between, resulting in high thermal resistance between layers. TIMs thus can be applied to
Traditional TIMs are ceramic-based or metal-based, whose thermal conductance between different layers achieve better thermal conductance between different layers.

Carbon nanotubes (CNTs) exhibit extraordinary structural, electrical, and thermal characteristics. Because of their high thermal conductivity and small thermal contact resistance comparing with traditional TIMs, CNTs have been investigated for their use as a TIM between the die and the heat spreader. The thermal conductivity \( k \) of multiwalled CNT (MWCNT) and single-walled CNT (SWCNT) is reported to be 3000 and 5000–8000 W/m·K, respectively [13]. For CNT bundles, \( k \) is in the range of 1750–5800 W/m·K [14]. This is significantly higher than copper whose \( k \) is 400 W/m·K. Furthermore, the thermal contact resistance of CNT arrays with a copper interface is reported to be only about 10 mm²K/W [15]. Simple and efficient methods for growing highly aligned and densely packed CNTs on silicon surfaces were demonstrated in [15]. Horizontal and vertical aligned CNTs were presented in [16]. Furthermore, researchers have also shown how different CNT junction patterns can be created [17]. These methods will enable many CNT-based structures to be feasible as a TIM. Zhou et al. [18] simulated the use of CNTs as thermal interconnects for on-die heat transfer between hot and cool (areas of lower activity) regions in silicon-on-insulator (SOI)-based ICs. Goplen and Sapatnekar [19] proposed a via-based grid where thermal through silicon vias (TTSVs) are allocated to specific regions on the chip for thermal management of 3-D ICs.

In this paper, we present a novel thermal structure combining MicroChannels and CNT-based thermal structure to tackle 3-D IC heat dissipation issues. In our structure, MicroChannels in replace of traditional heat sinks serve as the major cooling facility to bring heat outside the chip. CNT-based TIM layer, owing to its outstanding thermal properties, help balance the temperature horizontally across the die to reduce the peak temperature. In return, as temperature is flattened across the die, the overall cooling effect of MicroChannels is further improved. Moreover, CNT-based thermal vias are distributed across the die to connect MicroChannels, TIM layers, and device layers in vertical direction to allow heat flow across layer and further improve heat transfer capability of the proposed structure.

3-DSSST [20], which is a thermal analysis tool specially developed for 3-D IC MicroChannel simulation, is used to simulate IBM-PLACE 2.0 circuit benchmarks since 3-DSSST does not model heat sink, heat spreader, or thermal TSVs. To achieve fairness and validity in comparison, we modify the tool to support the features aforementioned. Finally, since the operating temperature has a significant impact on circuit aging (i.e., NBTI), the effects of the proposed structures on circuit reliability improvement are analyzed.

The rest of this paper is organized as follows. Section II provides background material that is helpful in understanding the ideas presented in this paper. Section III describes our proposed thermal management structure combining CNT-based TIM and TTSVs with MicroChannels. To justify such a structure, we also suggest a manufacturing scheme given the current fabrication status of CNT and MicroChannel. Section IV introduces the 3-D thermal model used and the modification we made to the simulation tool of 3-DSSST. The simulation setup and experimental results are presented in Section V. Section VI concludes this paper.

II. PRELIMINARIES

In this section, we present some background materials that is helpful in understanding the remainder of this paper.

A. Carbon Nanotubes

CNTs are 1-D conductors which may be visualized as a rolledup monolayer [Fig. 1(a)] or multilayers [Fig. 1(b)] of graphite with diameters and lengths in the nanometer and millimeter range, respectively [21]–[23]. Furthermore, depending upon their chirality (i.e., twist), they can be either semiconducting or metallic. CNTs present many remarkable electronic and mechanical characteristics, such as high thermal conductivity and ballistic current transport. CNTs have attracted immense attention for their potential usage in a wide range of applications. As mentioned earlier, because their thermal conductivity is around 8–15 times higher than that of copper, CNT is a promising candidate to replace copper for on-die heat transfer.

B. Typical Thermal Management Structure

Fig. 2 shows the typical heat spreader-based thermal management configuration of an IC. Solder bumps are deposited on the IC pads during the final wafer processing. The chip is then flipped and bonded with the packaging to interface with the external world. A heat spreader is used to dissipate or spread out excess heat from the IC (heat source) to the secondary heat exchanger (i.e., heat sink). The heat spreader, typically a copper plate, works most efficiently when the heat is uniformly distributed over the entire IC.
substrate [18]. Fins are commonly added to the heat sink to increase the surface area to achieve higher heat removal efficiency. A cooling mechanism such as air fans is integrated with the heat sink to dissipate the heat into the external environment. A TIM, typically commercial silicone grease, is added between the die and the heat spreader to improve thermal conductivity. Recently, TTSVs have been proposed to better conduct heat from the device layer to the heat sink (Fig. 3). The implementation of TTSVs has been realized by various approaches. The most common method is to fill the vias with metal (e.g., copper, tungsten), and then, to wrap the vias with an insulation layer. Studies have reported temperature reductions of up to 50% using silicon dummy thermal vias [24].

In 3-D-IC, microfluidic cooling is a generally used approach to tackle severe thermal issues. Its structure is shown in Fig. 4. Fluidic coolant in MicroChannels usually has higher thermal capacitance and thermal conductivity than air. Thus, it is more efficient in carrying heat away from the device layer. Moreover, embedded MicroChannels present a much larger convection area than air-cooling facilities [25]–[28].

C. Modeling Thermal Effects on Circuit Aging

NBTI is a key reliability issue for pMOS devices that is of immediate concern at the 45-nm node and below. NBTI manifests itself as an increase in $V_{th}$ which results in a decrease in drain current and transconductance. At 32 nm, where high-k metal gate-stacks and new materials like hafnium oxide are used to improve gate current density, the NBTI problem persists. Therefore, IC performance deteriorates over time (i.e., circuit aging), and can lead to critical failure. Since $V_{th}$ has a very strong correlation with operating temperature, the NBTI effect can be greatly reduced if the IC thermal profile can be improved.

A detailed model to predict the long-term $V_{th}$ degradation caused by NBTI has been presented in [29]. A tight upper bound on $\Delta V_{th}$ can be calculated as follows:

$$\Delta V_{th} = \left( \frac{\sqrt{K_v^2 \alpha T_{clk}}}{1 - \beta^2} \right)^{2n}$$

(1)

where $\alpha$ and $T_{clk}$ are the duty cycle and the clock period, respectively, and $n = 1/6$ for a $H_2$ diffusion-based model [29]. $K_v$ and $B_f$ are determined by the equations and parameters shown in Table I. In these equations, $q$ is the electron charge, $k_B$ is the Boltzmann constant, $t_{ox}$ and $e_{ox}$ are the oxide thickness and the permittivity, respectively, $C_{ox}$ is the oxide capacitance per unit area, $V_{gs}$ is the transistor gate to source voltage, $T$ is the temperature, and $t$ is the stress time. Given a specific process technology, (1) can be simplified as a power law

$$\Delta V_{th} = a t^n$$

(2)

Now, the propagation gate delay degradation $\Delta D_{pgd}$ can be approximated as follows:

$$\Delta D_{pgd} = b a t^n$$

(3)

where $a$ and $b$ are constant parameters dependent upon technology and the gate type. It can be extracted using circuit simulation for a given gate subject to various operating conditions. The above model is used in this paper to analyze the impact of temperature on aging.

III. PROPOSED HYBRID THERMAL STRUCTURE

The proposed hybrid thermal structure is composed of two parts: CNT grid layer to uniform the thermal profile and leverage hotspots, and MicroChannel to reduce the overall chip temperature. Fig. 5 shows CNT grid layers in detail. The CNT
grid layer contains a mesh-like structure of CNT bundles running horizontally and vertically. Heat can be transferred from TIM to MicroChannels inside substrate through vertical CNT arrays, and also to cooler regions of the TIM layer through horizontal ones. In 3-D IC, there are also TTSVs running through different layers to facilitate heat flow among neighboring layers. However, to reduce the area overhead of TTSV and maximize its heat transfer efficiency, we only plant the TTSVs in the hotspots region. Hence, the CNT grid can effectively normalize the temperature across the chip and significantly reduce the peak temperature.

However, although CNT-based thermal structures have great capability in ameliorating the 3-D IC thermal profile, traditional heat spreader and heat sink gradually run out of potential to offer a cool solution for 3-D IC. Heat spreader and heat sinks, owing to the nature of their design, are incapable of dissipating enough heat generated by the entire IC stack. No matter how large conductivity between heat spreader and device layer next to the heat spreader can be enabled by CNT structure, the bottleneck of the architecture lies on the air-cooled heat sink and heat spreader, leading to both high overall temperature and peak temperature inside 3-D chips.

To solve this problem, we propose MicroChannels as an excellent collaborator for CNT-based thermal structure. MicroChannels are physically embedded into the chips, and take heat outside chip with coolant flows. MicroChannels suit for 3-D IC as they can be built directly adjacent to inner device layers. MicroChannels also have great scalability along with vertically stacked up active layers. Another favorable feature of MicroChannels is that without heat spreader and heat sink bond to the top layer of the 3-D IC, the chip now has more flexible accessibility to optical and RF interfaces. With such advantages over traditional heat spreader and heat sink, MicroChannels indeed stand out of 3-D IC thermal management facilities in terms of cooling efficiency.

However, as shown in Section V, there also exist pitfalls of MicroChannels: first, although MicroChannels lower down overall temperature significantly, coolants in MicroChannels tend to shift upstream hotspots to downstream regions instead of annihilating them. Second, MicroChannels have quite limited capability to transfer heat in directions perpendicular to the flow direction, as well as from downstream to upstream regions. This impinges MicroChannels from further distributing heat all over the chip, thus has modest effects on degrading the temperature gradient. The proposed hybrid structure combines the advantages of MicroChannels and CNT-based grid, and compensates the shortcomings of each other. The high thermal conductivity of CNT-based grid transfers the heat uniformly across the chip, while the distributed MicroChannel inside each layer carries the heat outside the chip. Fig. 7 shows a cross-sectional view of the proposed hybrid structure.

The feasibility of the proposed structure is enabled by the current state of CNT and MicroChannel fabrication techniques. The CNT-based grid requires three key components: 1) CNT arrays running in the vertical and horizontal directions; 2) thermal vias filled with CNT bundles; and 3) the ability to interconnect the vertical and horizontal CNT arrays to form a mesh. Many works have reported on the synthesis of CNT array. Zhang et al. [15] have synthesized 15 μm thick CNT arrays with reported optimized thermal contact resistance as low as 7 mm²K/W. In [30], a 13 μm thick CNT array on the surface of a free mating substrate exhibiting thermal contact resistance around 15–17 mm²K/W was demonstrated. TSVs filled with CNT bundles were fabricated in [31] where a deep reactive ion etching (DRIE) process was used to etch deep vias in silicon. CNTs were then grown on a layer of catalyst at the bottom of the vias by thermal chemical vapor deposition (TCVD) as shown in Fig. 6(a) and (b). However, the horizontal CNT grid cannot be synthesized directly. First, we employ the approach of growing CNT bundles, and then, deploying them on the surface to connect them together in a controlled manner. The junctions between the horizontal and horizontal/vertical bundles can be formed by the method presented in [17]. CNT junctions between two crossing CNTs can be formed by exposing them to an electron beam. This enables us to form molecular CNT junctions of different patterns as shown in Fig. 6(c).

To build the whole structure, electrical TSVs can be first fabricated and deep vias which are to be filled with CNT bundles are drilled. Sawtooth-like MicroChannels can be etched at the back side of the wafer and get spin coated, patterned
and cured. First wafer is assembled onto substrate, and the next wafer will be assembled onto the previous one with their solder bumps bonded to copper vias of previous wafer. CNT bundles are grown in reserved vias and above avatrel covers of MicroChannels to form horizontal and vertical meshes as TIM layers and thermal TSVs. The manufacturing process of interlayer MicroChannels has been discussed in detail in [32]–[34]. The proposed fabrication process is shown in Fig. 8.

IV. 3-DSST MODIFICATION

3-DSST is a newly developed tool for 3-D architecture simulation with microfluidic channels. It is much faster than commercial tools, such as COMSOL [35] and more accurate than current open source MicroChannel simulators, such as 3D Interlayer Cooling Emulator [36], [37]. To make a fair comparison, we extended 3-DSST to support traditional architecture with heat sink and heat spreader. With the same methodology, for example, Kent Clark Lower Upper decomposition algorithm for matrices matrix solver [38], a new temperature matrix regarding heat sink and heat spreader is added. The matrix is filled using Hotspot’s 3-D thermal model [39] with a patch of thermal TSV modeling.

The original model divides each layer of the chip into a grid of blocks, and all the blocks in a same layer share a same thermal conductance value. However, when we consider thermal TSVs, we can no longer make such an assumption. Thus when a block contains thermal TSVs, we revise the thermal model with a new assumption that the horizontal conductance of the block remains the same, while the vertical conductance is modified to

\[
g_{\text{new}} = \frac{g_{\text{orig}}(S - S_{\text{tsv}}) + g_{\text{tsv}}S_{\text{tsv}}}{g_{\text{orig}}S} \tag{4}
\]

where \(g_{\text{new}}\) and \(g_{\text{orig}}\) are the vertical conductance of the block with and without TSVs, \(g_{\text{tsv}}\) is the conductance of thermal TSV in the block, \(S\) is the upside surface area, and \(S_{\text{tsv}}\) is the cross-sectional area of all TSVs contained in the block. As the steady-state temperature is all we need, we only solve the steady-state matrix which significantly speeds up 3-DSST.

Comparing with Hotspot, we also simplified the chip layer specification. In Hotspot IC model, a wafer is divided into different layers if the layers are made up of different materials, or even they are made up of the same material but consume different power. Typically, one wafer is divided into three layers: BEOL, device, and silicon bulk. Together with TIM, heat spreader, and heat sink, a three-wafer 3-D IC is actually divided into 14 layers. One significant drawback of such configuration is that it makes the steady temperature matrix very large in the finite element analysis method. In our tool, however, we shrink the number of elements by combining the parts divided in Hotspot back together to form one single element, so that a three-wafer 3-D IC architecture is just a five-layer architecture (plus heat spreader and heat sink). This simplified layer specification further improves the speed of 3-DSST without losing accuracy. The verification of accuracy of 3-DSST is presented in the next section.

V. SIMULATION RESULTS AND ANALYSIS

To evaluate the proposed hybrid thermal management structure, we first use a 3-D IC with three active layers as an instance for simulation. We then discuss the scalability when extended to multiple layers. For each active layer, we take IBM-PLACE 2.0 circuits with placement information as the benchmark circuit. Three benchmarks in IBM-PLACE 2.0 circuits are chosen for three active layers in the 3-D IC. The three benchmark circuits have average side length of 125 μm, which is not comparable with the MicroChannel width in our configuration. To solve this, without affecting the evaluation, we repeat one circuit 64 times to form a eight-by-eight circuit grid with chip dimension \(W = L = 1\) cm. To generate power profiles of three active layers, random power numbers are assigned to each cell in the large placement as the power density according to the following rules.

1) The average power density of an active layer is \(PD = 50\) W/cm².
2) Every active layer has a hotspot region occupies 5%–10% of chip area, with hotspot power density 250 W/cm².

<table>
<thead>
<tr>
<th>Layer</th>
<th>Thickness</th>
<th>Conductivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEOL</td>
<td>12 μm</td>
<td>2.25 W/m · K</td>
</tr>
<tr>
<td>Silicon Bulk</td>
<td>148 μm</td>
<td>161 W/m · K</td>
</tr>
<tr>
<td>Device</td>
<td>2 μm</td>
<td>163 W/m · K</td>
</tr>
<tr>
<td>TIM</td>
<td>20 μm</td>
<td>4 W/m · K</td>
</tr>
</tbody>
</table>

TABLE II  
BASELINE PARAMETERS
3) Nonhotspot cells on the chip are assigned with random power density in the range of [0.5PD, 1.5PD].
4) HotSpot regions in different active layers will not overlap. It is to follow the general 3-D IC design principle that high-power units at the same location in different active layers should not overlap.

The reference power numbers listed above are extracted from the Overall Roadmap Technology Characteristics Tables given in ITRS 2012 report [1]. In the report, it specifies the maximum unit power density of high-performance microprocessor is 48 W/cm² in year 2013. To best test our hybrid framework, we apply this maximum number in our randomized power profile. To make a fair and stable comparison, the random power profiles for each layer are generated once and used throughout the experiments.

We take the traditional thermal architecture with air-cooling heat sink and heat spreader as the baseline design. It is assumed that there is silicon-dioxide TIM between the bottom device layer and heat spreader or between different wafers to fill up the air gaps. The key simulation parameters of each layer in the baseline setting are listed in Table II as adopted by Hotspot5.01 as well. The ambient temperature is set to room temperature, for example, 293.15 K (20 °C).

### A. Tool Verification for 3-DSST

Although 3-DSST has been verified against commercial tool like COMSOL with respect to the MicroChannel modeling [40], as we have extended it to simulate heat spreader and heat sink as well, we need to verify the accuracy of the tool regarding new characteristics. We performed the verification by comparing our new version of 3-DSST with HotSpot5.01, which is the commonly used simulator for traditional thermal architectures. To verify the tool in different range of power density, we randomly generated the power profiles from high density to low density. Here (H1, H2) presents high-performance power profiles with average power density 40–50 W/cm², while (M1, M2) and (L1, L2) stand for cost-performance and low-power chip power profiles with average power density 20–40 W/cm² and below 20 W/cm², respectively.

The experiments are conducted on baseline design. The simulation results and analysis are presented in Table III. We can see from the table that our tool gives quite reliable temperature results, with only negligible difference with HotSpot5.01. Moreover, at the same time, our tool exhibits over 20 times speedup than Hotspot.

### B. Temperature Reduction

It is found through simulation that in the baseline design, the peak temperature of 3-D IC is 401.36 K (128.21 °C) and the overall average temperatures of three active layers is around 339.37 K (66.22 °C). The temperature profiles of all the layers are shown in Fig. 9(a). We can see that for the top layer, which is the furthest layer from heat sink, it is tremendously difficult for the heat sink to leverage the hotspot or reduce the average temperature. The silicon-dioxide-based TIMs are inefficient in transferring heat neither vertically nor horizontally, due to their low thermal conductivity. The melting temperature in inner active layers is unacceptable given chip reliability considerations.

To address the challenge, we first introduce CNT-based thermal structure looking forward to eliminating hotspots by replacing traditional silicon-dioxide TIM with CNT Grid between layers and between the bottom layer and the
he heat spreader. CNT-based TTSVs are also planted in the chip to transfer heat in the vertical direction. The parameters for the CNT arrays in the grid were summarized in Table IV. In this paper, we plant thermal TSVs uniformly on the chip instead of aggregating in the whitespace of three layers. Under current specification of TTSV size and pitch, the area overhead is only around 3%, which we take as acceptable overhead considering the significant benefits from TTSVs as shown below. The thermal conductivity of CNTs is set to that of MWCNT, which is relatively conservative as the maximum conductivity can be as high as 6000 W/m·K for SWCNT. The thickness of the CNT grid TIM sublayer is assumed to be around 20 μm. Although the long synthesized CNTs can be several hundred micrometers in length [41], such long CNTs are prone to bend due to gravity. Finally, the heat capacity used in the simulation is adopted from the report in [42]. As shown in Fig. 9(b), with the CNT thermal structure, peak temperature drops to 333.79 K (60.64 °C) and the average chip temperature drops to 322.18 K (49.03 °C). The figure also displays an obvious trend in heat spreading over the chip horizontally. Temperature gradient is largely reduced due to the high conductivity of CNT materials which enables efficient horizontal heat flow. The result is so far so good but the average temperature on the chip is still higher than expected.

A promising cooling approach used in 3-D IC is MicroChannel, as it overcomes obstacles in traditional heat sink architecture by effectively bringing heat generated by inner layers outside the chip. Assuming the parameters in Table V, we simulate MicroChannel architecture to show its advantages and drawbacks. From the simulation results in Fig. 9(c), we can see that both the peak and average temperatures are reduced significantly. However, MicroChannel alone is still not good enough. In Fig. 9(c), we can see that the hot spots located at upper stream units heat up coolant in the channels. Although hotspot temperature are reduced significantly by MicroChannels, units which share same channels with the hotspots but reside in downstream regions will suffer from reduced heat dissipation efficiency. The effect may violate the
expectation of the original design, as it can make some thermal vias no longer efficient as hotspots swim to downstream or burn some thermal-fragile units which are not expecting high temperature. Furthermore, the temperature gradient remains roughly the same on the chip, especially along the direction perpendicular to the flow direction.

With the observations, we integrate CNT-based TIM and TTSVs into the MicroChannel architecture looking forward to leveraging the drawbacks of the MicroChannel and further reducing the peak temperature. As shown in Fig. 9(d), the peak temperature of our 3-D IC drops to 321.57 K (48.42 °C). Moreover, we can see an obvious blurring of temperature gradient as the maximum temperature difference decreases from 66.95 K in the MicroChannel-only framework to 26.7 K in our hybrid framework. This improvement demonstrates the outstanding isotropic heat spreading capability of CNT grid TIM.

The experimental results are summarized in Table VI.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Baseline</th>
<th>CNT</th>
<th>μChannel</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak Temperature [K]</td>
<td>401.36</td>
<td>333.79</td>
<td>360.3</td>
<td>321.57</td>
</tr>
<tr>
<td>Improvement</td>
<td>16.84%</td>
<td>10.23%</td>
<td>19.88%</td>
<td></td>
</tr>
<tr>
<td>Average Temperature [K]</td>
<td>339.37</td>
<td>322.18</td>
<td>314.8</td>
<td>312.87</td>
</tr>
<tr>
<td>Improvement</td>
<td>5.07%</td>
<td>7.24%</td>
<td>7.81%</td>
<td></td>
</tr>
<tr>
<td>Maximum Temperature Difference [K]</td>
<td>79.33</td>
<td>17.25</td>
<td>66.95</td>
<td>26.70</td>
</tr>
<tr>
<td>Improvement</td>
<td>78.26%</td>
<td>15.61%</td>
<td>66.34%</td>
<td></td>
</tr>
</tbody>
</table>

C. Scalability With Number of Active Layers

Scalability is very important for thermal management in 3-D IC. With increasing number of active layers, 3-D IC thermal management frameworks should be able to prevent chip temperature growing out of control in such extreme thermal conditions. To test the heat dissipation capability of the thermal structure for different number of layers, we vary the layer number from two to five. Four thermal management frameworks are simulated including air-cooling heat sink and heat spreader in decreasing the overall temperature on chip, while CNT-based thermal structure excels at creating more balanced temperature distribution over the chip. It proves the advantages of our design on combining the features of two structures and effectively remove the hotspots as well as reduce the average chip temperature. Our proposed hybrid thermal management structure is capable to decrease up to 79.79 K of peak temperature, 26.5 K of average temperature and reduce as much as 66% of the maximum temperature difference on chip.

Fig. 10. Temperature of different thermal structures with various number of device layers. (a) Peak temperature variance with different number of active layers. (b) Average temperature variance with different number of active layers.
access more MicroChannel layers ($1 - 1/N$). Therefore, the average temperature in MicroChannel-based framework drops as we see in Fig. 10(b). With the CNT thermal structure, the heat can be more efficiently transferred among the chip and to the MicroChannel. Hence, it shows our hybrid framework steady outperforms MicroChannel-only framework.

D. Design Space Exploration

In this section, we analyze the impact of parameter variations in our proposed thermal structure on peak and average temperature reduction using the same 3-D IC.

1) CNT Conductivity: The high conductivity of CNT plays a key role in facilitating the heat flow along the CNT grid. However, there is a high likelihood that the thermal conductivity of CNTs is affected by the mixed bundles comprising both SWCNTs and MWCNTs depending on the CNT fabrication. Consequently, we need to evaluate how much the variation of the CNT conductivity will affect the chip temperature reduction using CNT-based thermal structure. From the simulation results shown in Fig. 11(a), it can be seen that in both architecture of traditional heat sink with CNT grid and proposed MicroChannel with CNT grid, the peak temperatures all drops dramatically before the turning point 100 W/m-K. Then after it, the slopes of both curves slowly approach zero and the benefits from the high thermal conductivity saturate around 5000–6000 W/m-K.

The average temperatures of both architecture show the similar trend. The two curves become almost constant after the turning point of 1000 W/m-K. This is because after the point, even though CNT grid is able to transfer more heat to cooling facilities, the heat dissipation ability of the cooling facilities have reached saturation. Hence, the average temperature of the chips cannot be further decreased with the increase of CNT grid thermal conductivity. This implies that mixed CNT grid is sufficient and the cost can be low.

2) CNT TIM Thickness: Because of the fabrication constraints and the physical properties of carbon nanotube, there are certain limitations of the TIM layers, especially for the thickness of CNT TIM. A relatively thin CNT grid may not exhibit the expected temperature normalization impact, while a thick one, which leads to better thermal performance, may suffer from fabrication problems such as fragility and strength limit of CNT materials. Here we show the thermal performance trend with increase in CNT TIM thickness to provide guidance in architecture parameter selection. As shown in Fig. 11(b), peak temperature decreases greatly at the beginning
and then slow down as the thickness of TIM layer increases, but the change in average temperature of the entire chip is quite small. It shows that the thickness of TIM influences more on the heat balance in one single active layer, but not much on the vertical heat transfer. Hence the total heat dissipation is not improved with TIM thickness increase. The trend of small thermal performance gain at large thickness value also tells us that a modest TIM height is adequate for our proposed structure.

3) Thermal Through Silicon via Pitch: As we adopt average thermal TSV planting strategy, the value of TTSV pitch needs careful evaluation to avoid unnecessary area overhead or waste of TTSV. TTSVs usually go through MicroChannel walls as shown in Fig. 7. Hence, there is difference between TTSV area overhead in heat-sink-based frameworks and MicroChannel-based frameworks because we do not plant TTSV over channel regions. To avoid physical instability of the chip and unnecessary area overhead, we limit TTSV pitch to above 30 μm. Fig. 11(c) shows the temperature change in respond to the TTSV pitch variance. In the design of baseline with CNT thermal structure, peak temperature increases significantly before the point 200 μm, and the average temperature also shows 5 K increase. However, in the design of MicroChannel with CNT thermal structure, average temperature barely changes with the TTSV pitch variation, and an increase in peak temperature around 10 K is found. Note that TTSVs are still capable to bring down peak temperature significantly in both designs although they are more effective in the former design. To find a balance between area overhead and temperature reduction, we also show area cost variation of TTSVs with the change of TTSV pitch. The figure shows that TTSV pitch in the range of [50, 100] μm should be a wise choice for reducing peak temperature without costing too much area.

4) Variation in Ratio of Channel Width With Respect to Wall Width of MicroChannel: The further exploration we did here is to investigate how the variation of MicroChannel parameter will affect the benefits of the proposed structure, and whether the CNT-based thermal structure will still maintain its potential and stability in balancing the chip temperature. The simulation results are compared with the MicroChannel-only architecture and shown in Fig. 11(d). In the experiment, the ratio of channel width to wall width varies from 0.5 to 4.0, and the MicroChannel width increases from 52 to 125 μm.

First observation we can get from the figure is that there is still big gap around 40 K between curves of peak temperature with and without CNT grid. The great benefits brought by CNT-based thermal structure still remain when the MicroChannel parameters vary. Second, although increasing the channel width leads to a continuous drop in average temperature in both designs, peak temperature in MicroChannel-only framework presents trend of rising when channel wall width ratio exceeds 1.5. It indicates that the balancing of temperature across the chip becomes worse. In our thermal model, heat transfers horizontally in two directions, across the channels or along the channels. Increasing the channel width ratio leads to thinner channel walls, which facilitates the heat transfer among channels. However, at the same time, it also makes it difficult to transfer the heat along the channel direction through the channel walls. In MicroChannel-based cooling frameworks, liquid coolant carries heat only from upstream regions to downstream regions, which brings in a temperature gradient along the coolant flow direction. Backward heat transferring from downstream regions to upstream regions through silicon layers, TIMs and channel walls are important in balancing overall temperature. Hence, thinner channel walls impinge the backward heat transfer path. When this factor dominates heat-balancing effects than the heat transfer in the perpendicular direction to other channels, temperature imbalance will aggregate and possibly result in peak temperature increase.

E. NBTI

It is well known that temperature plays an important role in the NBTI effect. A 3-D IC module with peak temperature needs particular attention since its circuitry can degrade more rapidly than other components and is almost always the critical component for reliability failure. To calculate the impact of temperature reduction on the circuit reliability improvement, a temperature-aware NBTI timing analysis algorithm was used [43]. The foundation of this algorithm is based on the Table I. It takes circuit netlist, a set of input vectors, signal probability, and so forth, as inputs. Then, it evaluates the delay degradation of each gate/block, and propagates the delay information to the next stage. Finally the delay degradation of the critical path is accumulated and calculated under given technology node and NBTI stress time parameters. In this paper, we assumed a 10-year working life. As temperature has a big impact on circuit delay degradation, it was expected that the delay degradation will be reduced as chip average temperature and peak temperature were reduced. Running the algorithm on all the circuit benchmarks, the delay degradation reduction from our baseline design is summarized in Table VII. This reduction can possibly help in achieving NBTI-aware timing closure more quickly and easily in IC design.

VI. Conclusion

In this paper, we have presented a hybrid thermal management framework combining CNT thermal structures and MicroChannel to solve 3-D IC thermal problems. Under this framework, heat can be evenly distributed over all layers and more efficiently conducted out of chip. In our modified tool based on 3-DSST, simulation results on IBM-PLACE 2.0 circuit benchmarks are performed. Different architecture parameters are thoroughly explored and discussed. Simulation results showed significant temperature reduction by using the proposed structure compared with previous designs, which in turn would benefit circuit reliability. It proves the potential
of the proposed design as a solution for future thermal management in 3-D IC.

REFERENCES


Wei Zhang (M’05) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Harbin Institute of Technology, Harbin, China, and Princeton University, Princeton, NJ, USA, in 1999, 2001, and 2009, respectively. She is currently an Assistant Professor with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong. She has published more than 30 technical papers in referred journals and conferences, and authored three book chapters. Her current research interests include multicore systems, reconfigurable computing, and emerging technology. Dr. Zhang has served in many conferences and journals as a Technical Program Committee Member and reviewer. She was a recipient of a Research Prize in Princeton and a Best Paper Award.

Jiale Huang, photograph and biography not available at the time of publication.

Shengqi Yang received the Double B.S. degree in mechanical engineering and economics and the M.S. degree in electrical engineering from Peking University, Beijing, China, in 2000 and 2002, respectively, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 2006. He joined Intel, Santa Clara, CA, USA, in 2006, and served as a Senior Video Architect involved in the Canmore SoC video algorithm development. He was with the Intel Mobility Group, Santa Clara, in 2008, where he was involved in Gen6 GPU media processing algorithm and kernel development. He joined the SoC Research and Development Center, School of Communication and Information Engineering, Shanghai University, Shanghai, China, in 2009, where he was the Shanghai Eastern Scholar Professor. Since 2013, he has been an Adjunct Professor with the Beijing University of Technology, Beijing. His current research interests include multimedia MPSoC design, embedded system design, and Internet of things.

Pallav Gupta received the B.S. (summa cum laude) degree in computer engineering from the University of Arizona, Tucson, AZ, USA, in 2002, the M.S. and Ph.D. degrees in electrical engineering from Princeton University, Princeton, NJ, USA, in 2004 and 2006, respectively. He was an Assistant Professor with the Department of Electrical and Computer Engineering, Villanova University, Villanova, PA, USA, from 2006 to 2010. He is currently a Senior Member of the technical staff with the Design and Technology Solutions group, Intel Corporation, Santa Clara, CA, USA.