Xylem: Enhancing Vertical Thermal Conduction in 3D Processor-Memory Stacks

Aditya Agrawal, Josep Torrellas, and Sachin Idgunji[†]
University of Illinois at Urbana-Champaign [†]Nvidia Corporation
http://iacoma.cs.uiuc.edu

ABSTRACT

In upcoming architectures that stack processor and DRAM dies, temperatures are higher because of the increased transistor density and the high inter-layer thermal resistance. However, past research has underestimated the extent of the thermal bottleneck. Recent experimental work shows that the Die-to-Die (D2D) layers hinder effective heat transfer, likely leading to the capping of core frequencies.

To address this problem, in this paper, we first show how to create pillars of high thermal conduction from the processor die to the heat sink. We do this by aligning and shorting dummy D2D μbumps with thermal TSVs (TTSVs). This lowers processor temperatures substantially. We then improve application performance by boosting the processor frequency until we consume the available thermal headroom. Finally, these aligned and shorted dummy μbump-TTSV sites create die regions of higher vertical thermal conduction. Hence, we propose to leverage them with three new architectural techniques: *conductivity-aware* thread placement, frequency boosting, and thread migration. We evaluate our scheme, called *Xylem*, using simulations of an 8-core processor at 2.4 GHz and 8 DRAM dies on top. μBump-TTSV alignment and shorting in a generic and in a customized Xylem design enable an average increase in processor frequency of 400 MHz and 720 MHz, respectively, at an area overhead of 0.63% and 0.81%, and without exceeding acceptable temperatures. This improves average application performance by 11% and 18%, respectively. Moreover, applying Xylem's conductivity-aware techniques enables further gains.

CCS CONCEPTS

• **Computer systems organization** → **Architectures**; • **Hardware** → **3D integrated circuits**;

KEYWORDS

3D chip, processor-memory integration, thermal management

ACM Reference format:
Aditya Agrawal, Josep Torrellas, and Sachin Idgunji.2017. Xylem: Enhancing Vertical Thermal Conduction in 3D Processor-Memory Stacks. In *Proceedings of MICRO-50, Cambridge, MA,USA, October 14–18, 2017,* 14 pages.
https://doi.org/10.1145/3123939.3124547

MICRO-50, October 14–18, 2017, Cambridge, MA,USA

ACM ISBN 978-1-4503-4952-9/17/10…$15.00
https://doi.org/10.1145/3123939.3124547

1 INTRODUCTION

Technology advances are about to enable further integration of computer architectures into chips that stack multiple memory and processor dies [2–4, 8, 34]. 3D stacking offers several benefits, such as a reduction in interconnect length and power, smaller form factors, and support for heterogeneous integration.

In these architectures, temperatures are higher than in conventional planar designs because of the increased transistor density, and because of the high inter-layer thermal resistance. In particular, the processor die is especially vulnerable, as it may be far from the heat sink — possibly at the bottom of the stack, so that processor power, ground, and I/O signals do not have to traverse the stack.

Stacking is made possible by Die-to-Die (D2D) connections and Through Silicon Vias (TSVs). D2D connections, also called microbumps (or μbumps), exist between dies, while TSVs run across the bulk silicon thickness of a die. TSVs carry electrical signals and are built with materials of high thermal conductance. Some proposals complement them with Thermal TSVs (TTSVs) [12, 22, 23, 36, 54, 59], which are dummy TSVs for thermal conduction only.

In this paper, we argue that past research has underestimated the extent of the thermal bottleneck in architectures that stack a high performance multicore die and multiple DRAM dies. Specifically, it has assumed that the stacked layers have a relatively high thermal conductance [36], suggesting that TTSVs alone are effective at keeping temperatures within acceptable values [12, 22, 23].

However, recent findings from IMEC [45, 46], Fujitsu [31, 43], IBM [9–11], and others [39, 40], shows that the material between dies (or D2D layer) presents a high thermal resistance. It is ≈16x more resistive than the bulk silicon, and ≈13x more resistive than the metal layers. In fact, the bulk silicon is one of the least thermally resistive layers in the stack. As a result, while TTSVs running across the bulk silicon (marginally) decrease the thermal resistance of the bulk silicon, they are not very effective overall. There are multiple D2D layers between the processor die and the heat sink, and they hinder effective heat transfer. As a result, the temperature of the processor die remains high, which likely forces the capping of core frequencies for reliability reasons.

To address this problem, in this paper, we show how to create pillars of high thermal conduction from the processor die at the bottom of the stack to the heat sink at the top. We do this by aligning and shorting dummy D2D μbumps with TTSVs. This lowers processor temperatures substantially. We then improve application performance by boosting the processor frequency until we consume the available thermal headroom. Finally, these aligned and shorted

dummy μbump-TTSV sites create die regions of higher vertical conduction. Hence, we propose to leverage them with three new architectural techniques to further improve performance: conductivity-aware thread placement, frequency boosting, and thread migration.

We evaluate our proposal, called *Xylem*, using simulations of a processor-memory stack. Our baseline is a Wide I/O compliant stack with an 8-core processor die at 2.4 GHz, and 8 DRAM dies on top. μBump-TTSV alignment and shorting in a generic and in a customized Xylem design enable an average increase in processor frequency of 400 MHz and 720 MHz, respectively, at an area overhead of 0.63% and 0.81%. This improves average application performance by 11% and 18%, respectively. Moreover, applying Xylem's conductivity-aware techniques enables further, albeit smaller, gains.

This paper addresses issues from low-level technology to high-level system integration. Its main contributions are:

- The observation that D2D layers are the thermal bottleneck in processor-memory stacks. Hence, TTSVs alone are ineffective.
- The creation of pillars of high thermal conduction through the D2D layer by aligning and shorting dummy μbumps with TTSVs.
- The observation that enhanced conduction through the D2D layers presents an architectural opportunity: the resulting thermal headroom can be consumed by boosting processor frequency and, hence, improve application performance.
- Three new architectural techniques that leverage the enhanced conduction around the aligned and shorted dummy μbump-TTSV sites: conductivity-aware thread placement, frequency boosting, and thread migration.

2 BACKGROUND AND MOTIVATION

The dies within a stack can be arranged in different configurations, such as: active layer of the dies facing each other (face-to-face or f2f), active layer of one facing the bulk of another (face-to-back or f2b), and bulk layer of the dies facing each other (back-to-back or b2b). For more than two homogeneous dies, such as a stack of DRAM dies, f2b is the preferred choice, and is the one we use. For multiple heterogeneous dies, such as memory and logic, the configuration is dictated by functionality, cost, performance, and thermal issues.

2.1 Through Silicon Vias (TSVs)

TSVs are vertical interconnects that run across the thickness of the die. There are several process flows for TSV fabrication, such as via first, via middle, frontside via-last, and backside via-last [5, 6, 29]. In addition, TSVs can be made before or after bonding. With the exception of the frontside via-last process, a TSV is present only in the silicon layer of a die, and not in the metal layers.

Ideally, we want a high TSV density (i.e., the number of TSVs in a given area). The density is determined by the TSV's aspect ratio and the die thickness. A TSV's aspect ratio is its height divided by its diameter, and is determined by manufacturing constraints and choice of TSV metal. For example, tungsten (W) allows TSV aspect ratios of about 30:1, while copper (Cu) is limited to no more than 10:1. For a given die thickness, the TSV density is proportional to the square of the aspect ratio; hence, high aspect ratios are preferred. For a given aspect ratio, the density is inversely proportional to the square of the die thickness. Hence, to attain high densities (and to aid TSV fabrication), dies are thinned down to a few tens of microns [3, 36]. However, researchers have observed that die thinning reduces lateral thermal spreading and worsens chip temperatures [16, 59]. Emma et al. [16] compare Cu and W TSVs in detail.

2.2 Die-to-Die Micro-Bumps (μBumps)

μbumps provide the interconnection between the dies in the stack. They are like the C4 pads that connect a die to a board, but with a much finer pitch. Their pitch is larger than that of TSVs, and hence they determine the interconnect density between the dies. The most widely-used implementation for a μbump is a Cu pillar with a tin-silver solder.

μbumps can be electrical or dummy. Electrical μbumps provide signal connections between the dies. To facilitate more electrical connections, electrical μbumps have fine diameters and pitch of about 17 μm and 50 μm, respectively [33]. Dummy μbumps exist, and are used for mechanical support and ease of stacking. They are electrically grounded to prevent charge accumulation.

2.3 Thermal Effects in Stacks

Stacking worsens on-die temperatures [3, 36, 59] because of the increased transistor density and high inter-layer thermal resistance. The rate Q of heat flow (i.e., power) across a layer is proportional to the thermal conductivity of the layer (λ) and the temperature difference across the layer (δT). Mathematically, $Q \propto \lambda \times \delta T$. If the power dissipated in a layer is constant, then a lower λ results in a higher temperature difference across the layer. Since the stack has multiple layers, the temperature differences add up, and the layer farthest from the heat sink will be at the highest temperature.

In a stack, the layer between dies (or die-to-die (D2D) layer) has the lowest thermal conductance. This is because typical D2D underfills have a $\lambda \approx 0.5$ W/m-K. In contrast, silicon has a $\lambda \approx 120$ W/m-K. Consequently, there is ongoing research on improving the D2D layer conductance, e.g., by using new underfills. An alternative is to fill the D2D layer with dummy μbumps, after provisioning for the electrical μbumps. Placing dummy μbumps has no area or manufacturing overhead. Their location can be standardized in the future, similar to electrical μbumps.

While the thermal conductivity of a Cu pillar with the solder in a μbump is ≈ 40 W/m-K, the real thermal conductivity of a D2D layer filled with dummy μbumps with a 25% density has been recently measured, by IBM [9, 11] and others [39], to be *only* about 1.5 W/m-K. The reason is that the D2D layer also includes several low-conductivity materials, like underfill/air, SiO_2, and SiN. Detailed cross sections of the D2D layer can be found in [9, 11, 39]. Such a low thermal conductivity makes the multiple D2D layers the *true* thermal bottleneck in the stack.

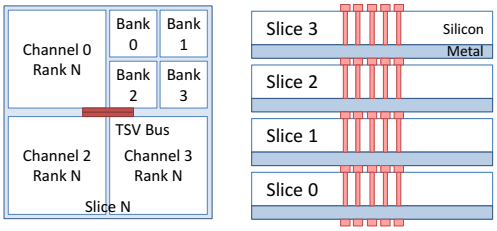
Since TSVs are vertical metal interconnects with high thermal conduction, researchers have proposed to use dummy TSVs simply for thermal conduction to reduce temperatures, as opposed to for electrical conduction [12, 22, 23, 36, 54, 59]. These are called Thermal TSVs (TTSVs). TTSV improve the thermal conduction of only the bulk silicon layer. Hence, as we show later, they are ineffective standalone at reducing the stack temperature. All TTSVs are electrically grounded to prevent charge accumulation.

2.4 Stacked Memory Standards

Manufacturers and standards committees have proposed several stacked memory architectures: Hybrid Memory Cube (HMC) [27]

from Micron, and High Bandwidth Memory (HBM) [25], Wide I/O [62], and Wide I/O 2 [61] from JEDEC. HMC and HBM are 2.5D memory architectures with an optional die for controller operations. Wide I/O (prototyped by Samsung [33]) and Wide I/O 2 are true 3D stacked architectures which can be connected to a processor die through TSVs. In this paper, we consider the thermal challenges in a true 3D stack like Wide I/O plus a processor die.

Fig. 1 shows the Wide I/O organization. It supports 4 physical channels. Each one contains independent control, data, and clock signals. Each memory die in the stack is called a slice, and has 4 ranks (1 per channel). A stack of 4 slices results in 4 ranks per channel and 16 overall. Each rank is divided into 4 banks, resulting in a total of 16 banks per slice. Current Wide I/O standards [61, 62] provide neither TTSVs nor dummy μbumps.



(a) Slice organization. **(b) Stack organization.**

Figure 1: Wide I/O organization (not to scale).

2.5 Shortcomings of Prior Work

Recent literature from IMEC, Fujitsu, IBM and others shows that the D2D layer is the thermal bottleneck in a stack. Specifically, Matsumoto et al. [39] and Colgan et al. at IBM [9, 11] independently measure $\lambda_{D2D} \approx 1.5$ W/m-K, and its thickness, $t \approx 20$ μm, on a die-to-wafer and die-to-die bonding technology, respectively. In addition, Oprins et al. at IMEC [45, 46] state that the inter-die thermal resistance represents a significant part of the total resistance in a 3D stack, and measure a $\lambda_{D2D} = 1.08$ W/m-K on a wafer-to-wafer bonding process. Also, Nakamura et al. at Fujitsu [31, 43] point out that for thermal analysis of a 3D-stacked LSI, it is critical to understand the heat flow through the microbumps.

Let us use Matsumoto's and Colgan's number. A layer's thermal resistance per unit area is $R_{th} = t/\lambda$. Hence, the D2D layer's R_{th} is $\approx$ 13.33 mm^2-K/W. In comparison, the R_{th} of the bulk silicon layer in a processor die is ≈ 0.83, and that of the metal layers in the processor die is ≈ 1. This makes the D2D layer $\approx$ 16x more resistive than the bulk silicon, and $\approx$ 13x more resistive than the metal layers.

Unfortunately, prior CAD and architecture work has consistently underestimated the D2D layer thermal resistance, either by assuming a high conductivity, a small thickness, or both. For example, [12] does not model the D2D layer at all. [36] uses a D2D thermal conductivity of $\lambda_{D2D} \approx$100 W/m-K, which is $\approx$ 65x higher than what was measured. [22, 23] model the D2D layer thickness to be only t_{D2D}=0.7 μm, which is $\approx$ 20x lower than what was measured, and a λ_{D2D} that varies that from 1 to 100 W/m-K, which is up to $\approx$ 65x higher than what was measured. All this has resulted in underestimating the thermal bottleneck effects of the D2D layer.

Prior proposals have focused on making the silicon layer of a die more conductive through the use of TTSVs. By underestimating the D2D thermal bottleneck, their data shows that TTSVs alone are effective for thermal management in 3D stacks. In this paper, by accurately taking into account the D2D layer, we show that TTSV placement alone is not effective at reducing the stack temperature. We need to combine it with a mechanism to reduce D2D layer thermal resistance.

3 MAIN STACK TRADE-OFFS

When integrating DRAM and processor dies in a stack, the resulting organization often depends on the specific characteristics of the technologies employed. This can be seen in recent research prototypes, such as Centip3De [15] and 3D-MAPS [32]. However, at a high level, there is a basic tradeoff between "processor-on-top" and "memory-on-top" organizations. These organizations are shown in Fig. 2a and 2b, respectively.

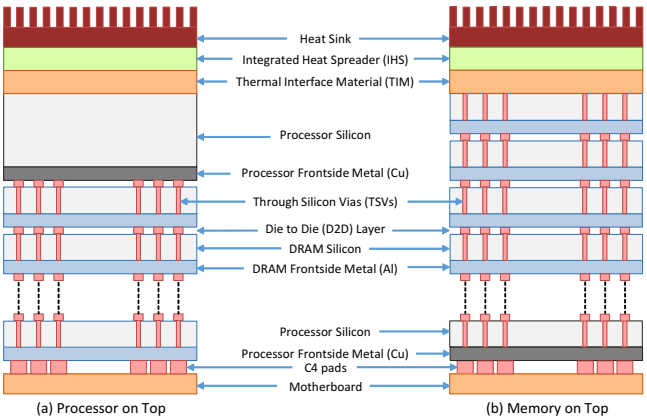


Figure 2: Two standard stack organizations.

In both organizations, the top die is connected to the heat sink and the Integrated Heat Spreader (IHS) using a Thermal Interface Material (TIM) [52]. Then, the different memory and processor dies are connected to each other through D2D connections and TSVs.

3.1 Processor on Top

The "processor-on-top" organization (Fig. 2a) has thermal advantages but significant manufacturing limitations. The advantages come from placing the die with most of the power dissipation closest to the heat sink. Also, the frontside metal layer of the processor die faces the memory stack, so that the processor die does not need TSVs or die thinning. The memory dies have TSVs and need die thinning.

The manufacturing difficulties result from the fact that a typical processor die has close to a thousand pins, about half of which are devoted to power and ground signals [16, 53]. In this organization, the memory dies have to provision TSVs to connect the processor power, ground and I/O signals to the C4 pads. This is a large overhead. Moreover, different processors have different pin number and location requirements. Therefore, the memory vendor either has to grossly over provision TSVs to accommodate a wide variety of processor dies, or manufacture custom dies for different processor vendors. Neither approach is desirable. In addition, TSVs add resistance to the Power Delivery Network (PDN) [16, 53]. For a current-hungry processor die far away from the C4 pads, the IR drop across the TSVs is a concern.

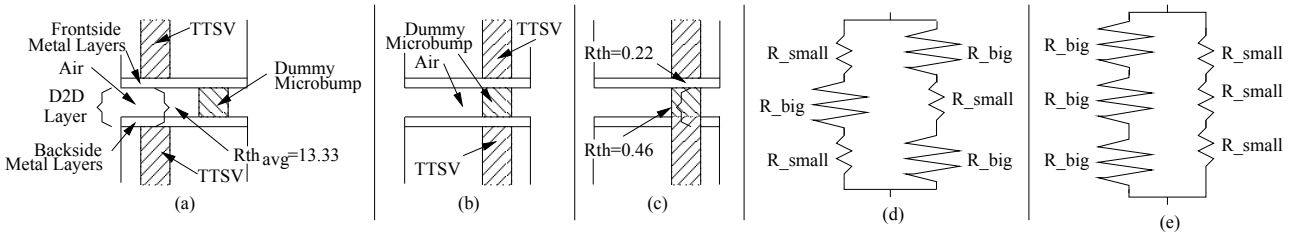


Figure 3: Impact of dummy μbump-TTSV alignment and shorting.

3.2 Memory on Top

The "memory-on-top" organization (Fig. 2b) has some thermal challenges, but does have clear manufacturing advantages. The thermal challenges result from the fact that the heat generated in the processor die has to traverse the many memory dies to reach the heat sink. Hence, the processor die will be at a significantly higher temperature than in planar organizations.

In this design, the frontside metal layer of the processor die is adjacent to the C4 pads. The main manufacturing advantage is that the high current-carrying power and ground signals, and the high-frequency I/O signals do not need TSVs [55]. In addition to simplifying the design, this avoids the IR drop issue mentioned above.

The memory stack in this configuration only contains TSVs as defined by the various 3D memory stacking standards, such as Wide I/O. The processor die only has to provision for the number and location of signals required for stack integration as given by the memory-stacking standards. Hence, the memory and processor die floorplans are independent of each other. However, the processor die has TSVs and requires thinning. Overall, given the manufacturing advantages of the "memory-on-top" organization, we use it.

4 ENHANCING VERTICAL CONDUCTION

In the "memory-on-top" organization, the processor (and memories) experience high temperatures. To avoid unsafe temperatures, these systems likely have to cap the frequency of the processor, in turn hurting performance.

To solve this problem, we propose to create pillars of high thermal conduction from the processor die to the heat sink at the top. For that, we need to focus on the thermal bottleneck, which is the D2D layers. In the following, we present the Xylem solution, which consists of aligning and shorting dummy D2D μbumps and TTSVs.

4.1 μBump-TTSV Alignment and Shorting

4.1.1 Summary of the Problem. Fig. 4 shows the interface between two f2b DRAM dies [44]. From top to bottom, we see the face side of a die, a D2D layer, and the back side of another die. The upper die shows the bulk silicon layer, which has the active devices and electrical TSVs, and the frontside metal layers. The latter contain metal routing layers (M_1 to M_n) separated by dielectric materials with low thermal conductance. The D2D layer consists of a layer with μbumps separated by air or underfill, and the backside metal layers of the lower die — which typically have 0-2 layers of metal routing (BM_1 to BM_2) separated by dielectric materials with low thermal conductance. On the left side, a TSV in the lower die is connected through the backside metal layers to an electrical μbump and then through the frontside metal layers to devices in the upper die.

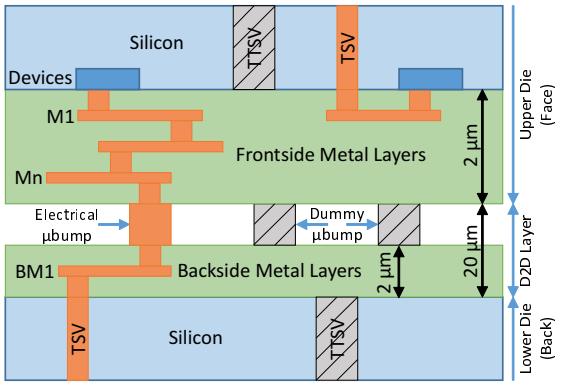


Figure 4: f2b DRAM die interface (not to scale).

In the figure, we have added two TTSVs and two dummy μbumps, shown in stripes. As the figure shows, the TTSVs typically terminate at the bulk silicon layers. In particular, they avoid the frontside metal layers because they would cause routing congestion in the already busy metal layers.[1] TTSVs are generally not aligned with the μbumps.

As indicated in Section 2.5, data from Colgan et al. at IBM [9, 11] and Matsumoto et al. [39] shows that the D2D layer has a high R_{th} of $\approx$13.33 mm^2-K/W. This is the result of a sizable thickness and of a low thermal conductivity λ. The multiple D2D layers are the thermal bottleneck in the stack.

4.1.2 Proposed Solution. Our approach is to align dummy μbumps in the D2D layer with TTSVs in the back side of the die and short them. To see how, consider Fig. 3a, which shows two identical DRAM dies connected f2b. There are two TTSVs that are aligned, but are separated by a D2D layer and the frontside metal layers of the upper die. We align the TTSVs with a dummy μbump (Fig. 3b) and then, using a backside metal via, short the dummy μbump with the lower-die TTSV (Fig. 3c).

Ideally, we would also want to short the upper-die TTSV with the μbump. However, the frontside metal layers carry many electrical signals, and using a frontside metal via [10] may or may not be possible, and may cause routing congestion. In addition, as shown in Fig. 3c, the R_{th} of the frontside metal layers is only 0.22 mm^2-K/W. This number is obtained with d=2 μm and λ=9 W/m-K [3]. This is small compared to the R_{th} of the original D2D layer.

With this approach, we have created a path of low thermal resistance from the lower-die TTSV through the D2D layer, which drains the heat. Specifically, given that the λ of the TTSV's Cu is

[1] The exception is when TSVs are built using the more costly frontside via last process, as in Black et al. [3].

400 W/m-K [3], the λ of the μbump is 40 W/m-K [39], and the μbump's thickness is 18μm, we compute the R_{th} of the D2D layer at this location as R_{th_bump} + R_{th_short}= 18μm / 40W/m-K + 2μm / 400W/m-K = 0.46 mm^2-K/W. This is a *local* R_{th} that is ≈30x lower than the average R_{th} of 13.33.

Pictorially, we have moved from an environment with thermal resistances like Fig. 3d to one like Fig. 3e. Fig. 3d represents Fig. 3a. We have a large thermal resistance (average of the D2D layer) connecting two small thermal resistances (the TTSVs), all in parallel with a small resistance (the μbump) connecting two relatively larger resistances (the silicon of the dies). On the other hand, Fig. 3e represents Fig. 3c. We have connected all three small resistances in series. A trivial calculation of the equivalent resistance shows that the heat has now a low-resistance path.

Electrical TSVs also contribute to thermal conduction because they are connected to μbumps. However, their contribution is limited. The reason is that the placement of electrical TSVs in the chip is dictated by stacking standards, and is oblivious to hotspots. For example, Wide I/O clusters all 1,200 TSVs in the center of the memory die, together with the electrical μbumps (Fig. 1).

4.2 Placing TTSVs

We now consider how to place the TTSVs in the dies. Then, aligning dummy μbumps to these TTSV will be easy because dummy μbumps are plentiful. Ideally, we want the TTSV diameter to be the same as the μbump diameter, to facilitate maximum heat flow.

We build on top of a stack that follows the Wide I/O organization (Fig. 1). Each die has a TSV bus in its center with 1,200 TSVs meant for electrical connections, together with the electrical μbumps. One such die is shown in Fig. 5a. These TSVs also aid in thermal conduction.

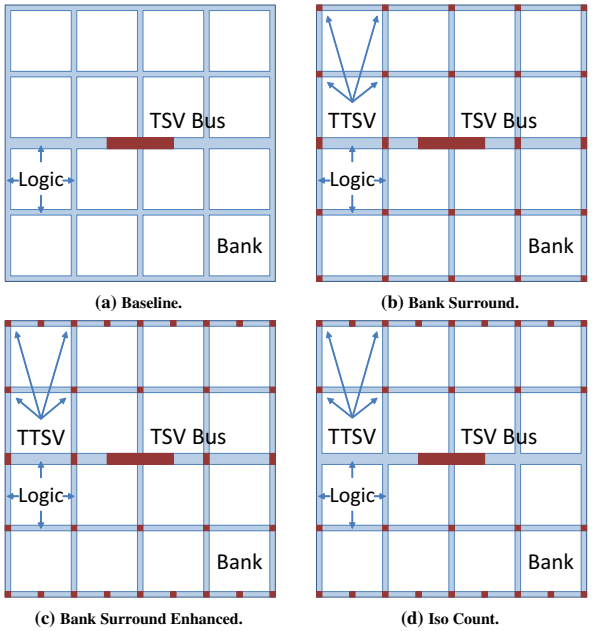


Figure 5: Xylem schemes.

To place TTSVs, we need to abide by the DRAM die floorplan and physical constraints. First, we place the TTSVs in the peripheral logic, to avoid disrupting the regular nature of the banks. Second, to avoid the TSV's lateral thermal blockage [7], we distribute the TTSVs, instead of aggregating them into a large TTSV farm. Finally, since TSV fabrication affects transistor performance nearby [1], we maintain a *Keep Out Zone* (KOZ) around each TTSV. Note that the peripheral logic contains row/column decoders, charge pumps, I/O logic, and temperature sensors. Hence, placing TTSVs has an area overhead.

Based on these constraints, we propose two simple TTSV placements, namely one generic and one custom. Ideally, TTSV placement should be generic. Memory manufacturers should not make assumptions about other layers of the stack, such as the location of the hotspots in the processor die. Our generic placement is called *Bank Surround* (*bank*) and is shown in Fig. 5b. It places the TTSVs in the peripheral logic at the vertices of each bank. Note that the peripheral logic area that runs horizontally across the die center is wider because of the Wide I/O TSV bus. Hence, we place two TTSVs at each point in the center stripe, instead of one everywhere else. The total number of TTSVs in the die is 28.

If we know the hotspots in the processor, we can devise a more effective, custom TTSV placement strategy. In this paper, we use the processor die floorplan shown in Fig. 6. This is a typical layout for commercial processors [20, 30, 56–58, 60], where the cores are on the outside and the Last-Level Cache (LLC) in the center. This layout separates the hot spots, which are the cores. The figure shows that the Wide I/O memory controllers and part of the TSV bus are in the logic layer. This is a layer in between the bulk silicon and the metal layers; it was not shown in Fig. 2 for simplicity.

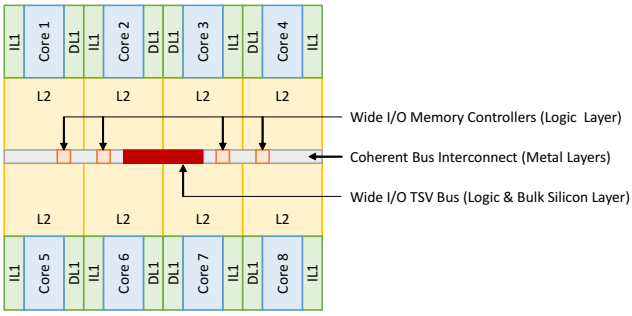


Figure 6: Processor die floorplan.

Knowing the location of the cores, we add 8 additional TTSVs close to the cores. The result is the *Bank Surround Enhanced* (*banke*) scheme of Fig. 5c. The total number of TTSVs is now 36. In effect, we have co-designed the memory and processor dies.

Finally, to perform experiments at a constant number of TTSVs, we take Bank Surround Enhanced and remove the 8 TTSVs from the peripheral logic area running horizontally across the die center. The resulting scheme, shown in Fig. 5d, has the same TTSV count as Bank Surround. We call it *Iso Count*.

5 TRADING THERMALS & PERFORMANCE

The alignment and shorting of dummy μbumps and TTSVs creates pillars of high thermal conduction from the processor die to the heat sink. We now show how we translate this feature into performance improvements.

5.1 Boosting Processor Frequency

An increase in the thermal conduction from the processor die to the heat sink results in a reduction in the temperature of the processor. Consequently, we propose to increase the processor's frequency and voltage until we consume the thermal headroom enabled. The result is higher application performance. For this, we use the DVFS infrastructure available in commercial processors [51]. DVFS infrastructure uses hardware and firmware, and is low overhead.

With Xylem, the manufacturer can rate the processor to work at a higher frequency for the same temperature limit. This optimization shows that, in a stacked architecture, there is a direct link between temperature and performance. Since these architectures are so thermally-limited, techniques like Xylem that reduce the temperature can improve performance substantially.

5.2 Conductivity (λ) Aware Techniques

The presence of pillars of high thermal conduction induces spatial heterogeneity in the processor die in its ability to dissipate heat. We refer to the sites with aligned and shorted μbump-TTSVs as high vertical conductivity (λ) sites. The die areas close to these sites dissipate heat more easily than areas far from them. Hence, we propose λ-*aware* optimizations, which leverage the higher thermal conduction around these sites. As examples, we propose λ-aware thread placement, λ-aware frequency boosting, and λ-aware thread migration.

λ-aware techniques can be applied when there is obvious heterogeneity in vertical conduction. An example is when two dies of different materials are placed side by side, both on top of the processor die. In this case, one of the two same-level dies may enable higher vertical λ than the other die. λ-aware techniques can leverage this heterogeneity.

Even regular layouts such as *bank* and *banke* present thermal heterogeneity and can be leveraged by λ-aware techniques. Specifically, the inner cores in Figure 6 (cores 2, 3, 6, and 7) have a smaller average distance to the high vertical λ sites compared to the outer cores (cores 1, 4, 5, and 8). This is the effect that we exploit in the λ-aware techniques.

5.2.1 λ-Aware Thread Placement. The idea is to place the most thermally demanding threads on cores that are, on average, closer to the high vertical λ sites. Typically, the most thermally demanding threads are compute intensive. Our proposal is to place these threads on the inner cores, and the memory intensive threads on the outer cores. The result is a higher reduction in temperature than with a random placement and, hence, the ability to enable a higher increase in processor frequency. Note that all cores in the chip have the same distance to the on-chip coherence bus and, hence, to the Wide I/O memory controllers.

5.2.2 λ-Aware Frequency Boosting. Cores that are on average farther from the high vertical λ sites reach the temperature limit before cores that are on average closer to them. Consequently, we propose λ-aware frequency boosting. The idea is to boost the frequency of the cores that are closer to these sites more than the others. In our processor die, we boost the frequency of the inner cores more than that of the outer cores, so that all cores reach the maximum temperature. With this technique, we improve performance.

5.2.3 λ-Aware Thread Migration. If we have fewer threads than cores, we may want to run a thread at high frequency on a core until the maximum temperature is attained, and then migrate the thread to an idle, cool core [21, 24]. This is repeated until the program terminates. In general, cores that are on average closer to the high vertical λ sites will take longer to reach the maximum temperature than the others. Consequently, in λ-aware thread migration, we propose to migrate a thread between cores that are on average closer to the high vertical λ sites, rather than between those on average farther, or between random cores. We will need fewer migrations to complete the program or, if we migrate at the same frequency, we will keep the cores at a lower average temperature. In our processor die, we migrate threads between the inner cores rather than between outer cores to keep a lower average temperature.

Note that λ-aware techniques are different from past thermal-aware techniques [63, 64]. Past work assumed that all cores are homogeneous. However, in Xylem, the cores are heterogeneous due to different conductivities. Our techniques exploit this heterogeneity. For example, consider a thread-to-core assignment decision. If two cores are at about the same temperature, past work will pick a core randomly. However, Xylem will pick the core with a higher λ.

6 THERMAL MODELING AND SETUP

6.1 Thermal Modeling

We model a "memory-on-top" processor-memory stack, with 8 DRAM dies on top of a multicore processor die. As shown in Fig.2b, the stack is composed of many distinct layers: active heat sink, Integrated Heat Spreader (IHS), Thermal Interface Material (TIM), DRAM silicon, DRAM metal, D2D, processor silicon, and processor metal. Some layers occur multiple times in the stack. TTSVs are present only in the DRAM silicon and processor silicon layers.

Some layers are heterogeneous, e.g., due to the presence of TSVs and TTSVs. We model each layer as being composed of many rectangular blocks. For numerical stability, it is desirable for the blocks to have a squarish aspect ratio. Different blocks within the same layer may have different λ. Note, however, that while the floorplan is specified in these rectangular blocks, the thermal simulation is performed in grid mode for higher accuracy.

The λ of an area with two materials A and B, with conductivities λ_A and λ_B, and fractional area occupancies ρ_A and ρ_B, such that $\rho_A + \rho_B = 1$, is [41]: $\lambda = \rho_A \times \lambda_A + \rho_B \times \lambda_B$. For example, a TSV bus is composed of 25% Cu with λ=400 W/m-K and 75% Si with λ=120 W/m-K. Hence, its effective $\lambda = 0.25\times400 + 0.75\times120 =$ 190 W/m-K.

Silicon and Frontside Metal. For accurate thermal analysis, the silicon layer and the metal layer of a die are modeled as two separate layers [3, 36]. We model the metal layer to also include the active silicon (or logic layer) because it is difficult to separate the power consumed by transistors and by wires. The metal routing layers (typically Al for memory and Cu for processor) together with dielectrics have a different λ than silicon. Since Al has a lower λ than Cu, the λ of the metal layer of a memory die is lower than that of the metal layer of the processor die [3]. The block boundaries in this layer are the architectural block boundaries — e.g., fetch, issue, RF, or ALU.

TSVs and TTSVs. We use Cu as the material for electrical TSVs and TTSVs because it has a high electrical and thermal conductivity. From ITRS [28], the electrical TSV size is 10 μm. Since the aspect ratio of Cu is limited to 10:1 [16], we use a die thickness of 100 μm for all the dies in the stack. We use a KOZ of 10 μm in both the X and Y dimensions. This results in an X and Y TSV pitch of 20 μm, and fractional area occupancies of 0.25 and 0.75 for the Cu and Si, respectively. The 1200 electrical TSVs are modeled using 48 blocks of 5×5 TSVs each. Each block is 100μm×100μm, and has an effective λ of 190 W/m-K. Each TTSV is modeled as a single block of 100μm×100μm with a λ of 400 W/m-K. It has a KOZ of 10 μm in both the X and Y dimensions.

D2D Layer. The D2D layer contains μbumps and backside metal layers, and is modeled as discussed in Sec. 4.1. The 1200 electrical μbumps in the center of the die are also organized in 48 blocks of 5×5 μbumps each, with a total effective λ of 1.5 W/m-K. Each dummy μbump has a size of 100μm×100μm, and a λ of 40 W/m-K. Dummy μbumps have a 25% occupancy.

TTSVs and dummy μbumps are thicker than electrical TSVs and μbumps. This is to facilitate maximum heat transfer. One can imagine a foundry to provide a library of TSVs and μbumps of different sizes. Alternatively, an array of skinny TSVs and smaller μbumps can be used to mimic a thick TTSV and a thick dummy μbump.

Table 1 shows the dimensions and λ of the layers in the stack. These values are obtained from various sources [3, 16, 26, 36].

Layer	Dimensions	Thermal Conductivity λ (W/m-K)
Heat Sink	6.0x6.0x0.7 cm^3	400
IHS	3.0x3.0x0.1 cm^3	400
TIM	50 μm	5
DRAM Silicon	100 μm	120 (Si); 400 (TSV); 190 (TSV bus)
DRAM Metal	2 μm	9
D2D	20 μm	1.5 (μbump = 40)
Proc Silicon	100 μm	120 (Si); 400 (TSV); 190 (TSV bus)
Proc Metal	12 μm	12

Table 1: Dimensions and thermal parameters.

Table 2 shows the different Xylem schemes evaluated. For each scheme, we list the name used in the evaluation and the number of TTSVs per die. The table includes the four schemes of Fig. 5 plus one additional one, called *prior*. *Prior* is similar to *banke* except that dummy μbumps and TTSVs are neither aligned nor shorted and, hence, there is no good conduction path through the D2D layers. *Prior* mimics the prior proposals, which place TTSVs near hotspots but ignore the impact of the D2D layers [12, 22, 23].

Xylem Scheme	Name	#TTSVs per Chip
Baseline (Wide I/O)	*base*	0
Bank Surround	*bank*	28
Bank Surround Enhanced	*banke*	36
Iso Count	*isoCount*	28
Prior proposals	*prior*	36

Table 2: Xylem schemes evaluated.

6.2 Architecture

The architecture is an 8-core chip multiprocessor under a stack of 8 DRAM dies. We use 32 nm technology. Each core is 4 issue and out of order. It has private L1 instruction and data caches, and a private L2 unified cache. A bus-based snoopy MESI protocol maintains coherence between the L2s. Each of the 8 DRAM dies has 4 Gb (512 MB) of memory, resulting in 4 GB of memory for the stack.

The DRAM stack organization follows Wide I/O [62]. The Wide I/O standard supports a modest bandwidth of about 12.8 GB/s, while Wide I/O 2 [61] supports a bandwidth of about 51.2 GB/s. Hence, we use the Wide I/O stack organization, but use a data rate of 51.2 GB/s. Note that the problems and solution proposed in this paper are generally applicable to any processor-memory stack.

The frequency of our architecture can change between 2.4 GHz (default) and 3.5 GHz in 100 MHz steps. At 2.4 GHz, the power consumed by the *base* system is 8-24 W in the processor die and 2-4.5 W in the memory dies. Both the processor and the DRAM dies have a similar area of ≈64mm^2 and a similar aspect ratio. This simplifies the analysis. If processor and DRAM die areas are different, we need to perform a slightly more involved thermal analysis, but the methodology is largely the same. We use a maximum safe temperature for the processor of $T_{j,max}$=100 °C, and a maximum for the DRAM of 95 °C, which is within *the extended range* allowed by JEDEC [14]. The architectural parameters are summarized in Table 3. We have broadly validated our power estimations with published numbers from Intel's Xeon E3-1260L.

Processor Parameters	
Multicore chip	32nm, eight 4-issue OoO, 2.4-3.5 GHz
Inst. L1 cache	32 KB, 2 way, 2 cycles Round Trip (RT)
Data L1 cache	32 KB, 2 way, WT, 2 cycles RT
L2 cache	256 KB, 8 way, WB, private, 10 cycles RT
Cache Line; Network	64 bytes; 512 bit bus
Coherence	Bus-based snoopy MESI protocol at L2
DRAM access	≈ 100 cycles RT (idle)
Max. temperature	Processor: $T_{j,max}$=100°C; DRAM=95°C
Stack DRAM Parameters	
Dies; Channels	8; 4
Ranks/die; Banks/rank	4 (1 per channel); 4
Capacity	4 Gb/die = 4 GB total in stack
I/O freq.; Data rate	800 MHz; DDR
# of memory controllers	4 Wide I/O DRAM controllers

Table 3: Architectural parameters.

6.3 Tools and Applications

We use the SESC [49] cycle-level simulator to model the architecture. We obtain the dynamic and leakage energy of the processor die from McPAT [35]. The timing and energy of the DRAM dies is modeled with DRAMSim2 [50]. We also use McPAT to estimate the area of the blocks within the processor die. The floorplan for each layer in the 3D stack is obtained using ArchFP [17]. Fig. 6 shows the high-level floorplan of the processor die that we use in the evaluation. We ensure that known hotspots in the processor die such as FPUs are spatially separated from each other.

To model the thermal effects in a stacked architecture, we use HotSpot's extension in [41]. The original HotSpot [26] models 3D-stacked architectures, but only allows homogeneous layers. The

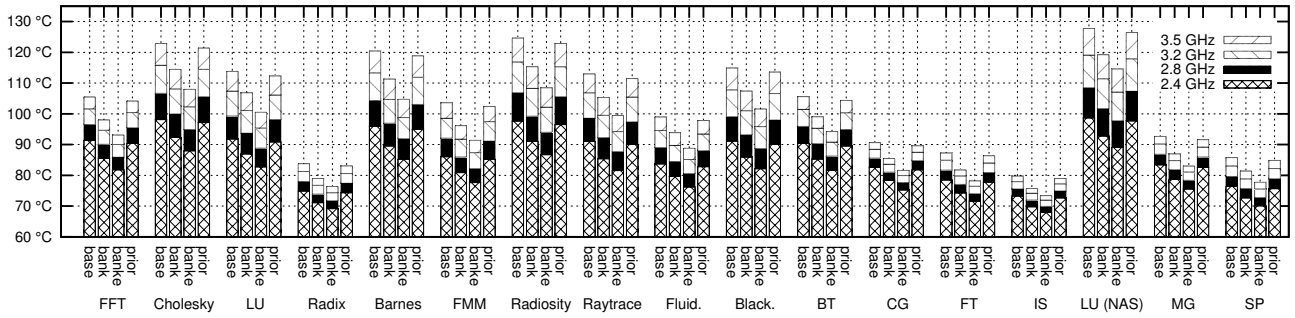


Figure 7: Impact of Xylem on the steady-state processor-die temperature.

extension enables modeling layers with blocks that have a heterogeneous set of λ and heat capacity values. We model both lateral and vertical heat conduction. We use the grid model as opposed to the block model as it is more accurate. Following the conventional HotSpot approach, we first obtain a processor-memory power trace, and then use HotSpot to estimate the steady state temperatures.

We run 8-threaded parallel applications from the SPLASH-2, PARSEC and NAS Parallel Benchmark (NPB) suites. The applications and their sizes are: Barnes (16 K particles), Cholesky (tk29.O), FFT (2^{22} points), FMM (16 K particles), LU (512x512 matrix, 16x16 blocks), Radiosity (batch), Radix (4 M integers), Raytrace (teapot), Blackscholes (sim medium), Fluidanimate (sim small), BT (small), CG (workstation), FT (workstation), IS (workstation), LU-NAS (small), MG (workstation), and SP (small). These codes represent a diverse set.

7 EVALUATION

For each Xylem scheme, we examine the area and routing overheads, the temperature reduction and, after frequency boosting, the changes in frequency, performance, power, and energy. We also compare schemes that have the same TTSV count but different placement. We then evaluate the impact of Xylem on memory temperatures. After that, we evaluate our proposed λ-aware techniques. Finally, we perform a sensitivity analysis. In all plots, temperature refers to the hotspot temperature.

7.1 TTSV Area and Routing Overheads

Based on the TTSV parameters of Section 6, the area of one TTSV plus its KOZ is 0.0144 mm^2. Given the total number of TTSVs for each of the Xylem schemes (as shown in Table 2), the total TTSV area overhead for the *bank* and *banke* schemes is equal to 0.4032 mm^2 and 0.5184 mm^2, respectively. Compared to the 64.34 mm^2 die area of a Wide I/O DRAM prototype by Samsung [33], this is a 0.63% and 0.81% overhead, respectively. Note that, since the TTSVs are passive, they do not have any energy overhead. Also, as shown in Fig. 3, TTSVs are not present in the frontside metal layers and, hence, do not cause any routing congestion or overheads there.

7.2 Impact of Xylem on Processor Temperature

Fig. 7 shows the effect of Xylem on the steady state temperature of the processor die. The figure shows four bars for each of the 17 applications, corresponding to the *base*, *bank*, *banke* and *prior* schemes. Each bar shows the steady state temperature reached by the hottest core when we run at 2.4 GHz, 2.8 GHz, 3.2 GHz, and 3.5 GHz. At some frequencies, for some applications, the figure shows a temperature in excess of $T_{j,max}$ (100 °C). However, in a real machine, a Dynamic Thermal Management (DTM) system would throttle frequencies to prevent excessive temperatures.

We see that processor temperature increases with increasing frequency. For example, as we go from 2.4 GHz to 3.5 GHz in *base*, the temperature increases by 10 °C in FT (a memory-intensive code) and by 30 °C in LU (NAS) (a compute-intensive code). Moreover, the temperature in *base* approaches $T_{j,max}$ even at 2.4 GHz for some applications, such as Cholesky, Barnes, Radiosity and LU (NAS). This shows that, *without TTSVs*, we cannot increase the frequency beyond 2.4 GHz.

We now consider a given frequency and compare the temperature reached by each of the schemes. For example, take 2.4 GHz. We see that both the *bank* and *banke* schemes are highly effective at reducing the temperature. To see the effect more clearly, Fig. 8 presents the data in a different way. It shows the temperature difference in °C between *base* and *bank*, and between *base* and *banke* — always at 2.4 GHz. The figure shows the difference for each application and the arithmetic mean. On average, *bank* and *banke*, reduce the steady state processor die temperature by 5 °C and 8.4 °C respectively. Going back to Fig. 7, we see that *bank* and *banke* schemes attain temperature reduction at all frequencies.

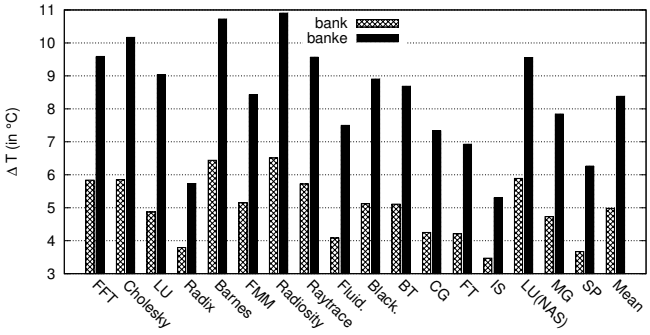


Figure 8: Steady-state temperature reduction over *base*.

We now consider *prior*. It is like *banke* except that dummy μbumps and TTSVs are neither aligned nor shorted. Hence, there is no good conduction path through the D2D layers. *prior* mimics the prior proposals, which ignore the impact of the D2D layers. We see in Fig. 7 that *prior* hardly offers any thermal benefit over *base*. This is in contrast to prior studies [12, 22, 23] that show thermal gains with TTSV placement alone. It is the reduction of the D2D layer resistance, and not just TTSV placement, that offers the benefits.

7.3 Effect of Temperature Reduction

Fig. 7 also shows that, at a fixed processor temperature, Xylem can run the system at higher frequencies. This can be seen by drawing an imaginary horizontal line and seeing that, as we go from *base* to *bank* and *banke*, we are substantially increasing the frequency of operation. This is one of the opportunities we exploit: improving thermal conduction, then boosting the frequency, for constant steady state temperature.

In this section, for each application, we choose the temperature of the *base* scheme at 2.4 GHz as a reference. Then, for the same application, for *bank* and *banke*, we find the frequency at which the processor temperature is closest to the reference without exceeding it. In the following, we analyze the resulting increase in system frequency and application performance, and the change in system power and energy.

7.3.1 System Frequency Increase. Fig. 9 shows the increase in system frequency enabled by the *bank* and *banke* schemes over *base*. Recall that, for each application, we keep the steady state temperature equal to the one in the *base* scheme at 2.4 GHz. The figure shows bars for each application and the arithmetic mean. We see that, on average for all the applications, *bank* boosts the frequency by about 400 MHz, and *banke* boosts it by 720 MHz. These are substantial increases, which we argue justify the design effort and area cost of Xylem. These large increases are the result of *base* being *highly frequency-throttled*. Indeed, the cores are designed to run at 3.5 GHz, but thermal constraints in *base* force them to run at 2.4 GHz.

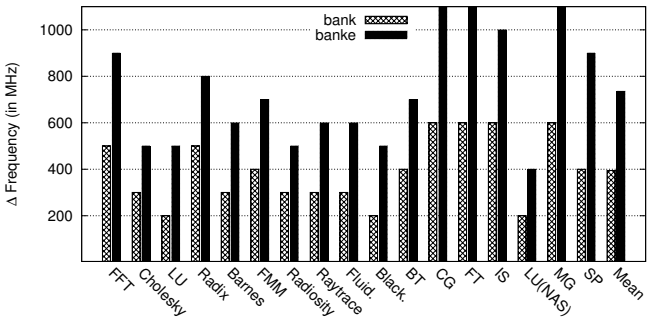


Figure 9: System frequency increase over *base*.

7.3.2 Application Performance Increase. Fig. 10 shows the application performance increase over the *base* scheme as a result of the frequency boosting enabled by the *bank* and *banke* schemes. The figure shows bars for each application and for the geometric mean. We can see that, on average, *bank* boosts the application performance by 11%, and *banke* boosts it by 18%. These are also large performance increases.

7.3.3 System Power and Energy Change. In our *base* system, it can be shown that the processor die consumes 8-24 W and the 8-die memory stack 2-4.5 W. Since Xylem enables frequency increases, the power and energy consumed by the processor-memory stack will change. Fig. 11 shows the increase in the power consumed by the stack in *bank* and *banke* over *base*. The figure is organized as before, with the last bars showing the geometric mean. We see that, on average, *bank* and *banke* increase the power consumption by

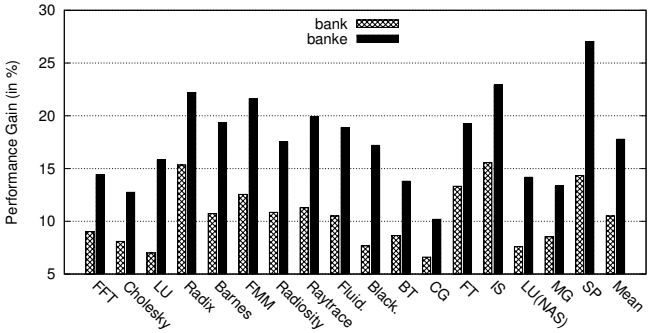


Figure 10: Application performance increase over *base*.

12% and 22%, respectively. The heat sink is able to dissipate this additional power while maintaining the same temperature as *base*.

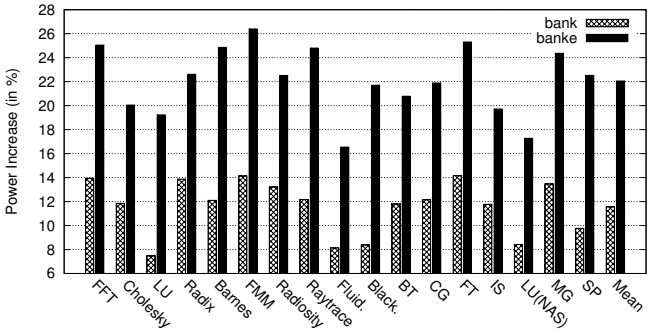


Figure 11: Stack power increase over *base*.

Fig. 12 shows the change in energy consumed by the stack in the *bank* and *banke* schemes over the *base* scheme. The figure is organized as before. As shown in the geometric mean bars, the applications end up consuming about the same energy on average. This is because of race-to-halt effects in some applications.

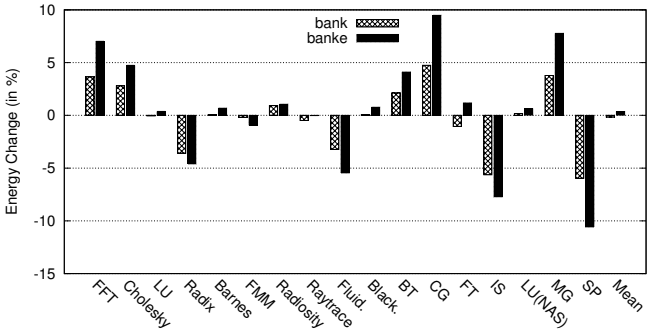


Figure 12: Stack energy change over *base*.

7.4 Comparison with Iso TTSV Count

The *bank* and *banke* schemes differ in the total number of TTSVs they use. We now compare two schemes that have the same TTSV count but place them in different places. Specifically, we compare *bank* and *isoCount* (Table 2). As shown in Fig. 5d, *isoCount* removes *bank*'s TTSVs in the central band and puts them in the die periphery, closer to the processor die hotspots.

Fig. 14 compares the steady-state temperature of the processor die under *bank* and *isoCount*. The figure is organized like Fig. 7. On

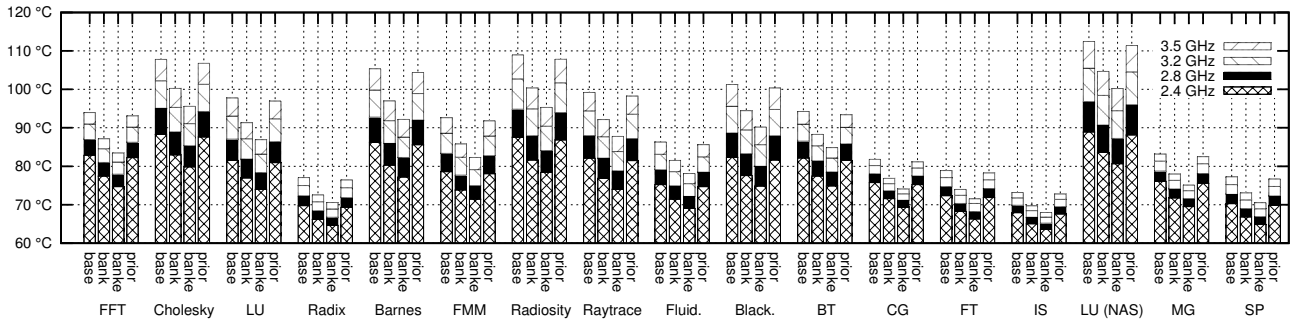


Figure 13: Impact of Xylem on the steady-state temperature of the bottom-most memory die.

average across all applications, it is shown that *isoCount* reduces the temperature by 3.7 °C over *bank*. This is slightly less than what *banke* accomplishes, and it shows that TTSV placement is important.

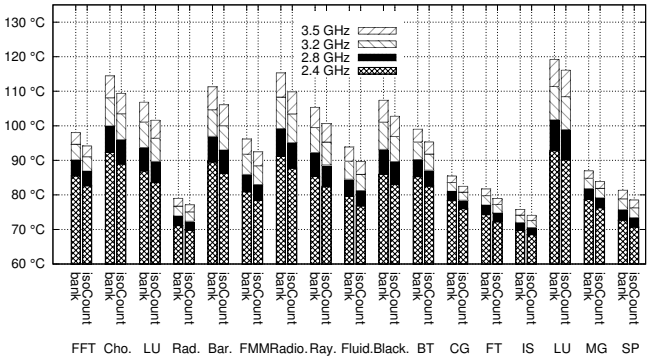


Figure 14: Impact of Xylem with iso TTSV count.

7.5 Impact of Xylem on Memory Temperature

Fig. 13 shows the effect of Xylem on the steady-state temperature of the hottest (bottom-most) memory die. The figure is organized like Fig. 7, with each application evaluated for the same four schemes. As in Fig. 7, for some frequencies, some applications have temperature over the 95 °C limit of JEDEC specifications [13, 14, 38, 62]. A real system would prevent these temperatures through frequency throttling.

We can see that, with *base* at 2.4 GHz, the memory die reaches temperatures of nearly 90 °C in the most demanding applications. This is within JEDEC specifications, and is about 10 °C less than the processor temperature. We also see that *bank* and *banke* are effective at reducing DRAM temperatures, while *prior* is not.

As temperature increases, DRAM leakage increases, which manifests itself as increased refresh requirement. For DDRx [13, 14, 38] and Wide I/O [62] devices, the refresh period at 85 °C is 64 ms, and is halved for every 10 °C increase in temperature. With Xylem, we can increase the processor frequency and still keep the same processor and DRAM temperatures, hence keeping the same refresh power. The impact of higher refresh rates on system energy and performance has been evaluated in [19, 37].

7.6 Conductivity(λ)-Aware Techniques

We now evaluate our proposed λ-aware techniques. As per Sec. 5.2, in *bank* and *banke*, the inner cores have a lower average distance to the high vertical λ sites, compared to the outer cores.

7.6.1 λ-Aware Thread Placement. In this experiment, we take a compute-intensive (LU-NAS) and a memory-intensive (IS) application, each with 4 threads. We place their threads in two configurations. In the *Outside* configuration, we place the LU threads (power intensive) on the outside cores (1, 4, 5 and 8 in Fig. 6) and the IS threads in the inner cores (2, 3, 6 and 7). In the *Inside* configuration, we do the opposite. We then find the maximum frequency at which the processor hotspot temperature is still lower than $T_{j,max}$. We keep a single frequency die-wide.

Fig. 15 shows the resulting frequencies for *base*, *bank* and *banke* for the two configurations. We see that, in *base*, the processor frequency is 100 MHz higher in *Inside* than in *Outside*. In *banke*, the gain increases to 200 MHz. These frequency gains come from the inner cores' lower average distance to the high vertical λ sites. This experiment shows the effect of λ-aware thread placement.

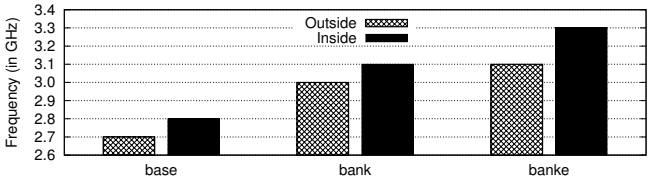


Figure 15: Exploiting λ-aware thread placement.

7.6.2 λ-Aware Frequency Boosting. We now run two instances of the same application with 4 threads each. One instance runs on the inner 4 cores, while the other on the outer 4 cores. We first bring the whole processor to the maximum frequency at which the processor hotspot temperature is lower than $T_{j,max}$ (*Single Frequency*). Then, we further boost the frequency of only the inner cores until they also reach $T_{j,max}$ (*Multiple Frequency*). We perform this experiment for all the applications, and take the average.

Fig. 16 shows the resulting frequencies for *base*, *bank* and *banke*. We see that, in *base*, there is practically no difference between the two bars: we cannot boost the inner cores much. However, in *banke*, thanks to the inner cores being on average closer to the high vertical λ sites, we can boost their frequency by 100MHz. This is λ-aware frequency boosting.

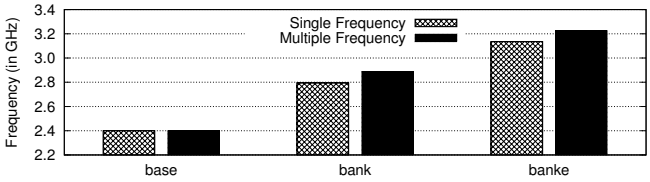


Figure 16: Exploiting λ-aware frequency boosting.

7.6.3 λ-Aware Thread Migration. We now take two threads of a given application and migrate them every 30 ms, either amongst the four inner cores (*Inner*) or amongst the four outer ones (*Outer*). We measure the hotspot temperature of the processor die. Fig. 17 shows the average results across all the applications for *base*, *bank* and *banke*, always at the same frequency. We see that, in *base*, migrating amongst inner cores reduces the hotspot temperature by only ≈0.4 °C over migrating amongst outer cores. However, in *banke*, thanks to the inner cores being on average closer to the high vertical λ sites, the temperature reduces by ≈1.5 °C. This is the effect of λ-aware thread migration, which reduces the thermal stress of processors for the same frequency.

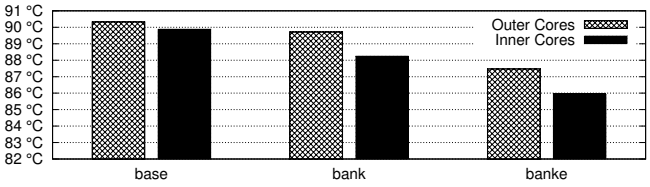


Figure 17: Temperature of the processor die when exploiting λ-aware thread migration.

Note that the frequency boosting (100-200 MHz) and temperature reduction (1.5 °C) from λ-aware techniques is *in addition* to the 400-720 MHz increase and 5-8.4 °C reduction obtained by aligning and shorting dummy μbumps with TTSVs.

7.7 Sensitivity Analysis

Finally, we perform a short sensitivity analysis of two important parameters: die thickness and number of memory dies in the stack.

7.7.1 Effect of the Die Thickness. For a constant TSV aspect ratio, die thinning is attractive for increasing TSV interconnect density. However, die thinning worsens chip temperatures because it inhibits lateral heat spreading. Fig. 18 shows the effect of thinning all the dies in the stack on processor temperature, averaged over all applications, at 2.4 GHz. The figure shows 3 sets of bars, each corresponding to a different die thickness. Each set has 3 bars, corresponding to *base*, *bank* and *banke*. As expected, processor temperatures become worse with die thinning. Hence, there is a trade off between TSV interconnect density and chip temperatures. Emma et al. [16] made the same observation.

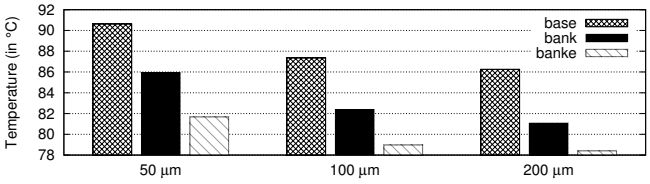


Figure 18: Impact of die thickness on processor temperature.

7.7.2 Effect of the Number of Memory Dies. Intuitively, increasing the number of memory dies increases the number of transistors and, hence, the total power consumed in the stack. It also increases the distance of the processor die from the heat sink. Therefore, we expect the processor temperatures to increase.

Fig. 19 shows the effect of the number of memory dies in the stack on processor temperature, averaged over all applications, at 2.4 GHz. The figure shows 3 sets of bars, each corresponding to a different number of memory dies in the stack. Each set has 3 bars, as above. As expected, the processor temperatures become worse with an increasing number of memory dies.

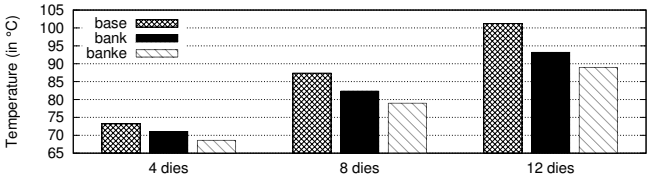


Figure 19: Impact of the number of memory dies on the processor temperature.

8 RELATED WORK

Some works examine algorithms to place TTSVs and minimize their count [12, 22, 23, 54, 59], but we find that these works either do not consider some physical implementation issues of TSVs, or use D2D layer parameters that are too aggressive. In either case, they do not fully consider the D2D layer resistance. So, TTSV placements alone appear to offer thermal savings.

Goplen and Sapatnekar [22, 23] reserve certain regions, called thermal via regions, for TTSV placement. These regions are uniformly placed throughout the chip and occupy 10% of the chip area. Their algorithm determines the number of TTSVs in each of these regions to minimize the overall TTSV count. However, their algorithm is applicable only for standard cells. In addition, one should not place TTSVs uniformly in a memory die, since they disrupt the regular DRAM array layout. Further, in general, the memory vendor does not have information about the processor power densities or hotspots needed by the algorithm. Moreover, their evaluation uses a D2D layer thickness of 0.7 μm. This value is over 20x lower than the state of the art and, therefore, the effects of the D2D layer resistance are not considered. Finally, a 10% area overhead is substantial.

Cong and Zhang [12] propose a heuristic algorithm to minimize TTSV count in a more generic layout of blocks in a die. The algorithm requires knowledge of all the layers of the stack, which is not available to a memory vendor. Also, the TTSVs are assumed to directly connect to the metal layers of the adjacent die. As a result, the work does not consider the physical implementation of TTSVs or the presence of the D2D layer.

Ganeshpure and Kundu [18] propose Heat Pipes as a heat transfer mechanism, where placing TTSVs directly at the hotspots is difficult due to wiring congestion. Chen et al. [7] propose an algorithm for TSV placement with the goal of mitigating the lateral thermal blockage effects of TSVs.

Emma et al. [16] propose different modes of operation for processor-on-processor stacking. They analyze the impact of die thickness and hotspot offset on temperature. They do not propose temperature-reducing techniques.

Puttaswamy and Loh [47, 48] propose and analyze techniques for thermal management for a 3D *processor* (not a generic 3D processor-memory organization). In Thermal Herding, they propose moving the hottest datapaths (16 LSBs) closest to the heat sink. Extending their proposal to a generic 3D processor-memory stack would imply moving the processor die closest to the heat sink, resulting in our

"processor-on-top" configuration. Also, they do not discuss the issues of thermal resistance, TTSVs, the D2D layer, or μbumps.

Black et al. [3] and Loh [36] look at the performance benefits and thermal challenges of a 3D processor-memory stack. Black et al.'s thermal analysis is for a 2-die stack (a processor and a memory die in a f2f configuration). They discuss the impact of the D2D layer and the frontside metal layer conductivity on temperature. However, both works assume a "processor-on-top" configuration. In addition, Loh assumes that the D2D layer has a λ=100 W/m-K (which is 1/4 of that of bulk Cu), and a thickness of 2 μm. In practice, according to experimental measurements by Colgan et al. at IBM [9, 11] and Matsumoto et al. [39], the D2D has a λ that is 65x smaller, and a thickness that is about 10x higher. Hence, after adding 16 stacked DRAM dies, [36] observes only a 10 °C maximum temperature increase.

Milojevic et al. [42] characterize a multicore with 16 2-wide cores and two DRAM dies on top. They use a passive heat sink and no TTSVs, and still attain safe temperatures. We use a more power-hungry design point, due to our wider-issue cores and higher frequencies. Specifically, our *base* system consumes up to 24 W in the processor die and 4.5 W in memory dies, for a total of 28.5 W (or 44 W for *banke*); their design consumes a maximum of 18 W. In addition, our design has 8 D2D layers of high thermal resistance in series, while their design only has 2. As a result, to keep safe temperatures in our processor die (Figure 7) and bottom-most memory (Figure 13), we need TTSVs with alignment and shorting (in addition to an active heat sink).

Smart Refresh [19] considers the impact of higher refresh rates in DRAMs, due to higher temperatures in a 3D stack. The performance impact due to higher temperatures in stacked DRAM is also studied by Loi et al. [37].

3D-MAPS [32] and Centip3De [15] are two 3D research prototypes. 3D-MAPS has 2 dies in a "processor-on-top" configuration. The cores run at 277 MHz with a peak power of 4 W. Centip3De has 7 dies also in a "logic-on-top" configuration. However, the cores operate in the NTV regime with a frequency of 10 to 80 MHz. Commercial 3D systems will want to have much higher frequencies, and require aggressive thermal techniques similar to those in our paper.

9 CONCLUSION

This paper made four contributions. First, unlike in prior work, it observed that the D2D layers are the main thermal bottleneck in processor-memory stacks, and showed that standalone TTSVs are ineffective. Second, it proposed the creation of pillars of high thermal conduction through the D2D layer by aligning and shorting dummy μbumps with TTSVs. Third, it observed that enhanced conduction through the D2D layers presents an opportunity: the resulting thermal headroom can be consumed by boosting processor frequency and, hence, improve application performance. Finally, it introduced three new architectural improvements that leverage the enhanced local conduction around the aligned and shorted dummy μbump-TTSV sites: λ-aware thread placement, λ-aware frequency boosting, and λ-aware thread migration.

We evaluated our scheme, called *Xylem*, using simulations of an 8-core processor at 2.4 GHz and 8 DRAM dies on top. μBump-TTSV alignment and shorting in a generic and in a customized Xylem design enabled an average increase in the processor frequency of 400 MHz and 720 MHz, repectively, at an area overhead of 0.63% and 0.81%, respectively. This improved the average application performance by 11% and 18%, respectively. Moreover, applying Xylem's λ-aware improvements enabled further gains.

10 ACKNOWLEDGMENT

This work was supported in part by NSF grant CCF-1649432.

REFERENCES

[1] K. Athikulwongse, A. Chakraborty, J. S. Yang, D. Z. Pan, and S. K. Lim. 2010. Stress-Driven 3D-IC Placement with TSV Keep-Out Zone and Regularity Study. In *IEEE International Conference on Computer-Aided Design*.

[2] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat. 2001. 3-D ICs: A Novel Chip Design for Improving Deep-Submicrometer Interconnect Performance and Systems-on-Chip Integration. *Proc. IEEE* (May 2001).

[3] B. Black, M. Annavaram, N. Brekelbaum, J. DeVale, L. Jiang, G. H. Loh, D. McCauley, P. Morrow, D. W. Nelson, D. Pantuso, P. Reed, J. Rupley, S. Sadasivan, J. Shen, and C. Webb. 2006. Die Stacking (3D) Microarchitecture. In *IEEE International Symposium on Microarchitecture*.

[4] S. Borkar. 2011. 3D Integration for Energy Efficient System Design. In *IEEE Design Automation Conference*.

[5] Erh-Hao Chen, Tzu-Chien Hsu, Cha-Hsin Lin, Pei-Jer Tzeng, Chung-Chih Wang, Shang-Chun Chen, Jui-Chin Chen, Chien-Chou Chen, Yu-Chen Hsin, Po-Chih Chang, Yiu-Hsiang Chang, Shin-Chiang Chen, Yu ming Lin, Sue-Chen Liao, and Tzu-Kun Ku. 2013. Fine-pitch Backside Via-last TSV Process with Optimization on Temporary Glue and Bonding Conditions. In *IEEE Electronic Components and Technology Conference*.

[6] K.N. Chen and C.S. Tan. 2011. Integration Schemes and Enabling Technologies for Three-Dimensional Integrated Circuits. *IET Computers & Digital Techniques* (May 2011).

[7] Y. Chen, E. Kursun, D. Motschman, C. Johnson, and Y. Xie. 2011. Analysis and Mitigation of Lateral Thermal Blockage Effect of Through-Silicon-Via in 3D IC Designs. In *International Symposium on Low Power Electronics and Design*.

[8] T. Y. Chiang, S. J. Souri, C. H. Chui, and K.C. Saraswat. 2001. Thermal Analysis of Heterogeneous 3D ICs with Various Integration Scenarios. In *International Electron Devices Meeting*.

[9] E. G. Colgan, P. Andry, B. Dang, J. H. Magerlein, J. Maria, R. J. Polastre, and J. Wakil. 2012. Measurement of Microbump Thermal Resistance in 3D Chip Stacks. In *IEEE Semiconductor Thermal Measurement and Management Symposium*.

[10] E. G. Colgan, R. J. Polastre, J. Knickerbocker, J. Wakil, J. Gambino, and K. Tallman. 2013. Measurement of Back End of Line Thermal Resistance for 3D Chip Stacks. In *IEEE Semiconductor Thermal Measurement and Management Symposium*.

[11] E. G. Colgan and J. Wakil. 2013. Measured Thermal Resistance of Microbumps in 3D Chip Stacks. (March 2013). http://www.electronics-cooling.com/2013/03/measured-thermal-resistance-of-microbumps-in-3d-chip-stacks/

[12] J. Cong and Y. Zhang. 2005. Thermal Via Planning for 3-D ICs. In *IEEE International Conference on Computer-Aided Design*.

[13] DDR2 SDRAM Standard. 2009. http://www.jedec.org/standards-documents/docs/jesd-79-2e. (2009).

[14] DDR3 SDRAM Standard. 2012. http://www.jedec.org/standards-documents/docs/jesd-79-3d. (2012).

[15] R. G. Dreslinski, D. Fick, B. Giridhar, G. Kim, S. Seo, M. Fojtik, S. Satpathy, Y. Lee, D. Kim, N. Liu, M. Wieckowski, G. Chen, D. Sylvester, D. Blaauw, and T. Mudge. 2013. Centip3De: A 64-Core, 3D Stacked Near-Threshold System. *IEEE Micro* (Mar. 2013).

[16] P. Emma, A. Buyuktosunoglu, M. Healy, K. Kailas, V. Puente, R. Yu, A. Hartstein, P. Bose, and J. Moreno. 2014. 3D Stacking of High-Performance Processors. In *IEEE International Symposium on High-Performance Computer Architecture*.

[17] G. G. Faust, R. Zhang, K. Skadron, M.R. Stan, and B.H. Meyer. 2012. ArchFP: Rapid Prototyping of pre-RTL Floorplans. In *IEEE International Conference on VLSI and System-on-Chip*. http://lava.cs.virginia.edu/archfp/

[18] K. Ganeshpure and S. Kundu. 2012. Reducing Temperature Variation in 3D Integrated Circuits Using Heat Pipes. In *IEEE Symposium on VLSI*.

[19] M. Ghosh and H.-H. Lee. 2007. Smart Refresh: An Enhanced Memory Controller Design for Reducing Energy in Conventional and 3D Die-Stacked DRAMs. In *IEEE International Symposium on Microarchitecture*.

[20] R. Golla and P. Jordan. 2011. T4: A Highly Threaded Server-on-a-Chip with Native Support for Heterogeneous Computing. In *Hot Chips: A Symposium on High Performance Chips*.

[21] M. Gomaa, M. D. Powell, and T. N. Vijaykumar. 2004. Heat-and-Run: Leveraging SMT and CMP to Manage Power Density Through the Operating System. In *International Conference on Architectural Support for Programming Languages and Operating Systems*.

[22] B. Goplen and S. S. Sapatnekar. 2005. Thermal Via Placement in 3D ICs. In *International Symposium on Physical Design*.

[23] B. Goplen and S. S. Sapatnekar. 2006. Placement of Thermal Vias in 3-D ICs Using Various Thermal Objectives. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (Apr. 2006).

[24] S. Heo, K. Barr, and K. Asanovic. 2003. Reducing Power Density through Activity Migration. In *International Symposium on Low Power Electronics and Design*.

[25] High Bandwidth Memory (HBM) Standard. 2013. http://www.jedec.org/standards-documents/results/jesd235. (2013).

[26] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M.R. Stan. 2006. HotSpot: A Compact Thermal Modeling Methodology for Early-Stage VLSI Design. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* (2006). http://lava.cs.virginia.edu/HotSpot/index.htm

[27] Hybrid Memory Cube Consortium. 2012. http://hybridmemorycube.org/. (2012).

[28] International Technology Roadmap for Semiconductors (ITRS). 2012. http://www.itrs2.net. (2012).

[29] S. C. Johnson. 2009. Via first, middle, last, or after? *3D Packaging Newsletter on 3D IC, TSV, WLP & Embedded Technologies* (Dec. 2009). http://www.i-micronews.com/upload%5Cnewsletter%5C3DNov09.pdf

[30] R. Kalla. 2009. POWER7: IBM's Next Generation POWER Microprocessor. In *Hot Chips: A Symposium on High Performance Chips*.

[31] S. Kikuchi, M. Suwada, H. Onuki, Y. Iwakiri, and N. Nakamura. 2015. Thermal Characterization and Modeling of BEOL for 3D Integration. In *IEEE CPMT Symposium Japan*.

[32] D. H. Kim, K. Athikulwongse, M. Healy, M. Hossain, M. Jung, I. Khorosh, G. Kumar, Y.-J. Lee, D. Lewis, T.-W. Lin, C. Liu, S. Panth, M. Pathak, M. Ren, G. Shen, T. Song, D. H. Woo, X. Zhao, J. Kim, H. Choi, G. Loh, H. H. Lee, and S. K. Lim. 2012. 3D-MAPS: 3D Massively Parallel Processor with Stacked Memory. In *IEEE International Solid-State Circuits Conference*.

[33] Jung-Sik Kim, Chi Sung Oh, Hocheol Lee, Donghyuk Lee, Hyong-Ryol Hwang, Sooman Hwang, Byongwook Na, Joungwook Moon, Jin-Guk Kim, Hanna Park, Jang-Woo Ryu, Kiwon Park, Sang-Kyu Kang, So-Young Kim, Hoyoung Kim, Jong-Min Bang, Hyunyoon Cho, Minsoo Jang, Cheolmin Han, Jung-Bae Lee, Kyehyun Kyung, Joo-Sun Choi, and Young-Hyun Jun. 2011. A 1.2V 12.8GB/s 2Gb Mobile Wide-I/O DRAM with 4x128 I/Os Using TSV-Based Stacking. In *IEEE International Solid-State Circuits Conference*.

[34] M. Koyanagi, H. Kurino, K.-W. Lee, K. Sakuma, N. Miyakawa, and H. Itani. 1998. Future System-On-Silicon LSI Chips. *IEEE Micro* (Jul. 1998).

[35] Sheng Li, Jung Ho Ahn, Richard D. Strong, Jay B. Brockman, Dean M. Tullsen, and Norman P. Jouppi. 2009. McPAT: An Integrated Power, Area, and Timing Modeling Framework for Multicore and Manycore Architectures. In *International Symposium on Microarchitecture*.

[36] G.H. Loh. 2008. 3D-Stacked Memory Architectures for Multicore Processors. In *International Symposium on Computer Architecture*.

[37] G.L. Loi, B. Agrawal, N. Srivastava, S.-C. Lin, T. Sherwood, and K. Banerjee. 2006. A Thermally-Aware Performance Analysis of Vertically Integrated (3-D) Processor-Memory Hierarchy. In *Design Automation Conference*.

[38] Low Power DDR3 SDRAM Standard. 2013. http://www.jedec.org/standards-documents/results/jesd209-3. (2013).

[39] K. Matsumoto, S. Ibaraki, K. Sakuma, K. Sueoka, H. Kikuchi, Y. Orii, and F. Yamada. 2010. Thermal Resistance Evaluation of a Three-dimensional (3D) Chip Stack. In *Electronics Packaging Technology Conference*.

[40] S. Melamed, K. Kikuchi, and M. Aoyagi. 2015. Sensitivity of the Thermal Profile of Bump-Bonded 3D Systems to Inter-Die Bonding Layer Properties. In *IEEE CPMT Symposium Japan*.

[41] J. Meng, K. Kawakami, and A. K. Coskun. 2012. Optimizing Energy Efficiency of 3-D Multicore Systems with Stacked DRAM under Power and Thermal Constraints. In *IEEE Design Automation Conference*.

[42] D. Milojevic, S. Idgunji, D. Jevdjic, E. Ozer, P. Lotfi-Kamran, A. Panteli, A. Prodromou, C. Nicopoulos, D. Hardy, B. Falsafi,

and Y. Sazeides. 2012. Thermal Characterization of Cloud Workloads on a Power-Efficient Server-on-Chip. In *International Conference on Computer Design*.

[43] N. Nakamura, Y. Iwakiri, H. Onuki, M. Suwada, and S. Kikuchi. 2015. Thermal Modeling and Experimental Study of 3D Stack Package with Hot Spot Consideration. In *IEEE Electronic Components and Technology Conference*.

[44] Dave Noice and Vassilios Gerousis. 2010. Physical Design Implementation for 3D IC: Methodology and Tools. *International Symposium on Physical Design* (Mar. 2010). http://www.ispd.cc/slides/slides10/4_02.pdf Invited talk from Cadence.

[45] H. Oprins, V. Cherman, T. Webers, A. Salahouelhadj, S. W. Kim, Lan Peng, G. Van der Plas, and E. Beyne. 2016. Thermal Characterization of the Inter-Die Thermal Resistance of Hybrid Cu/Dielectric Wafer-to-Wafer Bonding. In *IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*.

[46] H. Oprins, B. Vandevelde, M. Badaroglu, M. Gonzalez, G. Van der Plas, and E. Beyne. 2013. Numerical Comparison of the Thermal Performance of 3D Stacking and Si Interposer Based Packaging Concepts. In *IEEE Electronic Components and Technology Conference*.

[47] K. Puttaswamy and G.H. Loh. 2007. Thermal Herding: Microarchitecture Techniques for Controlling Hotspots in High-Performance 3D-Integrated Processors. In *IEEE International Symposium on High Performance Computer Architecture*.

[48] Kiran Puttaswamy and Gabriel H. Loh. 2006. Thermal Analysis of a 3D Die-Stacked High-Performance Microprocessor. In *ACM Great Lakes Symposium on VLSI*.

[49] Jose Renau, Basilio Fraguela, James Tuck, Wei Liu, Milos Prvulovic, Luis Ceze, Smruti Sarangi, Paul Sack, Karin Strauss, and Pablo Montesinos. 2005. SESC simulator. (Jan. 2005). http://sesc.sourceforge.net

[50] P. Rosenfeld, E. Cooper-Balis, and B. Jacob. 2011. DRAMSim2: A Cycle Accurate Memory System Simulator. *Computer Architecture Letters* (Jan. 2011). http://www.eng.umd.edu/~blj/dramsim/

[51] E. Rotem, A. Naveh, D. Rajwan, A. Ananthakrishnan, and E. Weissmann. 2012. Power-Management Architecture of the Intel Microarchitecture Code-Named Sandy Bridge. *IEEE Micro* (Mar. 2012).

[52] E.C. Samson, S.V. Machiroutu, J.-Y. Chang, I. Santos, J. Hermerding, A. Dani, R. Prasher, and D.W.Song. 2005. Interface Material Selection and a Thermal Management Technique in Second-Generation Platforms Built on Intel Centrino Mobile Technology. In *Intel Technology Journal*.

[53] Manjunath Shevgoor, Jung-Sik Kim, Niladrish Chatterjee, Rajeev Balasubramonian, Al Davis, and Aniruddha N. Udipi. 2013. Quantifying the Relationship Between the Power Delivery Network and Architectural Policies in a 3D-stacked Memory Device. In *IEEE/ACM International Symposium on Microarchitecture*.

[54] S.G. Singh and C. S. Tan. 2009. Impact of Thermal Through Silicon Via (TTSV) on the Temperature Profile of Multi-Layer 3-D Device Stack. In *International Conference on 3D System Integration*.

[55] D. Skarlatos, R. Thomas, A. Agrawal, S. Qin, R. Pilawa-Podgurski, U. R. Karpuzcu, R. Teodorescu, N. S. Kim, and J. Torrellas. 2016. Snatch: Opportunistically Reassigning Power Allocation between Processor and Memory in 3D Stacks. In *IEEE International Conference on Microarchitecture*.

[56] J. Stuecheli. 2013. Next Generation POWER microprocessor. In *Hot Chips: A Symposium on High Performance Chips*.

[57] S. Turullols and R. Sivaramakrishnan. 2012. SPARC T5: 16-core CMT Processor with Glueless 1-Hop Scaling to 8-Sockets. In *Hot Chips: A Symposium on High Performance Chips*.

[58] S. Undy. 2011. Poulson: An 8 Core 32 nm Next Generation Intel Itanium Processor. In *Hot Chips: A Symposium on High Performance Chips*.

[59] G. Van der Plas, P. Limaye, A. Mercha, H. Oprins, C. Torregiani, S. Thijs, D. Linten, M. Stucchi, K. Guruprasad, D. Velenis, D. Shinichi, V. Cherman, B. Vandevelde, V. Simons, I. De Wolf, R. Labie, D. Perry, S. Bronckers, N. Minas, M. Cupac, W. Ruythooren, J. Van Olmen, A. Phommahaxay, M. de Potter de ten Broeck, A. Opdebeeck, M. Rakowski, B. De Wachter, M. Dehan, M. Nelis, R. Agarwal, W. Dehaene, Y. Travaly, P. Marchal, and E. Beyne. 2010. Design Issues and Considerations for Low-Cost 3D TSV IC Technology. In *IEEE International Solid-State Circuits Conference*.

[60] S. White. 2011. High Performance Power-Efficient x86-64 Server & Desktop Processors: using Bulldozer core. In *Hot Chips: A Symposium on High Performance Chips*.

[61] Wide I/O 2 Standard. 2014. http://www.jedec.org/standards-documents/results/jesd229-2. (2014).

[62] Wide I/O SDR Standard. 2011. http://www.jedec.org/standards-documents/results/jesd229. (2011).

[63] D. Zhao, H. Homayoun, and A.V. Veidenbaum. 2013. Temperature Aware Thread Migration in 3D Architecture with Stacked DRAM. In *International Symposium on Quality Electronic Design*.

[64] Xiuyi Zhou, Yi Xu, Yu Du, Youtao Zhang, and Jun Yang. 2008. Thermal Management for 3D Processors via Task Scheduling. In *International Conference on Parallel Processing*.