Proceedings of the ASME 2021 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems
InterPACK2021
October 26-28, 2021, Virtual, Online

IPACK2021-73017

EFFICIENT THERMAL ANALYSIS OF LAB-GROWN DIAMOND HEAT SPREADERS

Zihao Yuan[1], Tao Zhang[1], Jeroen Van Duren[2], and Ayse K. Coskun[1]

[1]Boston University, Boston, MA
[2]Diamond Foundry

ABSTRACT

Lab-grown diamond heat spreaders are becoming attractive solutions compared to traditional copper heat spreaders due to their high thermal conductivity, the ability to directly bond them on silicon, and allow for an ultra-thin silicon layer. Researchers have developed various thermal models and prototypes of lab-grown diamond heat spreaders to evaluate their cooling performance and heat spreading ability. The majority of existing thermal models are built using finite-element method (FEM) based simulators such as COMSOL and ANSYS. However, such commercial simulators are computationally expensive and lead to long solution times along with large memory requirements. These limitations make commercial simulators unsuitable for evaluating numerous design alternatives or runtime scenarios for real-world high-performance processors. Because of this modeling challenge, none of the existing works have evaluated the thermal behavior of lab-grown diamond heat spreaders on real-world high-performance processors running realistic application benchmarks. Recently, we have developed a parallel compact thermal simulator, PACT, that is able to carry out fast and accurate steady-state and transient thermal simulations and can be extended to support emerging integration and cooling technologies. In this paper, we use PACT to evaluate the steady-state and transient cooling performance of lab-grown diamond heat spreaders against traditional copper heat spreaders on various real-world high-performance processors (e.g., Intel i7 6950X, IBM Power9, and PicoSoC). By using PACT with architectural performance and power simulators such as Sniper and McPAT, we are able to run transient simulations with realistic benchmarks. Simulation results show that lab-grown diamond heat spreaders achieve maximum temperature and thermal gradient reductions of up to 26.73 ℃ and 13.75 ℃ when compared to traditional copper heat spreaders, respectively. The maximum steady-state and transient simulation times of PACT for the real-world high-performance chips and realistic applications used in our experiments are 259 s and 22 min, respectively.

Keywords: Lab-grown diamond, heat spreader, high-performance processors, thermal simulations

NOMENCLATURE

PACT	A standard cell level to architectural level parallel compact thermal simulator
Sniper	A parallel, high-speed and accurate x86 simulator
McPAT	An integrated power, area, and timing modeling framework for multicore and manycore architectures
NAS	NASA Advanced Supercomputing
OpenROAD	An integrated chip physical design tool that takes a design from synthesized Verilog to routed layout
HotSpot	A compact thermal modeling methodology for early-stage VLSI design
bt	Block Tri-diagonal solver
cg	Conjugate Gradient
dc	Data Cube
ep	Embarrassingly Parallel
ft	Discrete 3D fast Fourier Transform
is	Integer Sort, random memory access
lu	Lower-Upper Gauss-Seidel solver
mg	Multi-Grid on a sequence of meshes
sp	Scalar Penta-diagonal solver
ua	Unstructured Adaptive mesh
PCB	Printed circuit board
PDN	Power delivery network
IC	Integrated circuit
T	Temperature (K)
TC	Thermal conductivity (W/mK)
TR	Thermal resistivity (mK/W)
TIM	Thermal interface material
ID1	Chip stack #1
ID2	Chip stack #2
ID3	Chip stack #3
ID4	Chip stack #4
ID5	Chip stack #5
T_{max}	Maximum temperature (℃)
ΔT	Temperature gradient (℃)

TDP	Thermal design power (W)
T_{diff}	Temperature difference (℃)
TBR	Thermal boundary resistance (m^2K/GW)
$Freq$	Frequency (Hz)
HTC	Heat transfer coefficient (W/m^2K)

1. INTRODUCTION

Over the last few decades, on-chip power densities have grown tremendously following the down-scaling of transistors. Power densities that reach 1-2 KW/cm^2 caused by the performance boost of scaling already occur in high-performance chips and result in amplified localized hot spots [1]. These localized on-chip hot spots not only degrade the performance of the chip, but also generate larger sub-threshold leakage power and create reliability challenges [2]. High power density hot spots also cause thermal runaway, and result in more power loss and harm the energy efficiency of the computing systems, especially for cloud computation globally. Existing cooling solutions such as forced air cooling via fans or traditional pin-fin heat sinks are often not sufficient to mitigate these high power density hot spots efficiently and can lead to over/under-cooling, affecting system design cost and power. For passive cooling methods such as pin-fin heat sink, the cooling performance are relatively low. However, for active cooling methods such as forced air cooling via fans and liquid cooling, the system requires additional cooling power (fan power and liquid pumping power). It's hard to optimize the existing cooling solutions at design time and runtime to achieve both high computing performance for processors and energy efficiency for the cooling methods. Lab-grown diamond heat spreaders have the potential to provide better cooling performance compared to traditional copper heat spreaders due to the high thermal conductivity, the ability to directly bond them on silicon, no additional cooling power needed, and allow for an ultra-thin silicon layer [3, 4]. However, lab-grown diamond heat spreader thermal models are usually developed and simulated using commercial finite-element method (FEM) based multiphysics simulators (e.g., COMSOL and ANSYS [5, 6]). Such commercial simulators are computationally expensive and experience long solution times along with large memory requirements [7-9], which results in simulation timing overhead for parametric studies and thermal evaluations for lab-grown diamond heat spreaders with real-world high-performance processors. Due to the aforementioned modeling challenges using commercial FEM-based simulators, none of the existing works have evaluated the thermal behavior of lab-grown diamond heat spreaders on real-world high-performance processors running realistic application benchmarks.

We have recently developed a parallel compact thermal simulator, PACT [10], that is able to carry out fast and accurate steady-state and transient thermal simulations, and can be extended to support various emerging integration and cooling technologies. In this paper, we use PACT to compare the cooling performance of lab-grown diamond heat spreaders against traditional copper heat spreaders using real-world high-performance processors. To demonstrate the cooling advantages of lab-grown diamond heat spreaders, we select three different real-world high-performance processors (Intel i7 6950X, IBM Power9, and PicoSoC) and compare the cooling performance in terms of maximum temperature reductions and thermal gradient reductions between lab-grown diamond heat spreaders and traditional copper heat spreaders. We also carry out several parametric studies to demonstrate the impact of cooling performance of lab-grown diamond heat spreaders with different chip thickness and cooling packages. The main contributions of the paper are as follows:

1. We are the first to study the cooling performance of the lab-grown diamond heat spreader using real-world-like high-performance chips with realistic application benchmarks. We use PACT [10] to conduct steady-state and transient thermal simulations with various high-performance chips to evaluate the cooling performance of lab-grown diamond heat spreaders. We run benchmark applications [11] on real-world-like high-performance chips using popular architecture-level performance and power simulators [12, 13] to obtain transient power traces. The generated transient power traces are used as inputs to PACT to perform transient thermal analysis with lab-grown diamond heat spreaders and traditional copper heat spreaders. For each of the high-performance chips under test, we evaluate the thermal maps, maximum temperature reductions, and thermal gradient reductions of the high-performance chips with diamond heat spreaders versus with traditional copper heat spreaders.
2. For both steady-state and transient thermal simulations, we apply a coarse granularity interconnect model to represent the interconnects in real chips between the processing layer and substrate (on side of the PCB board). The processing layer is in between the heat spreader and interconnects.
3. We also perform parametric studies of the processor layer's thickness and cooling packages to better understand the cooling advantages of the lab-grown diamond heat spreader.
4. Simulation results show that lab-grown diamond heat spreaders achieve maximum temperature and thermal gradient reductions of up to 26.73 ℃ and 13.75 ℃ when compared to traditional copper heat spreaders, respectively.

2. MATERIALS AND METHODS

In this section, we first give an overview of the PACT simulator, and then discuss the models of processors and interconnects we build for the steady-state and transient simulations. Finally, we illustrate our methodology for collecting transient power traces from realistic application benchmarks.

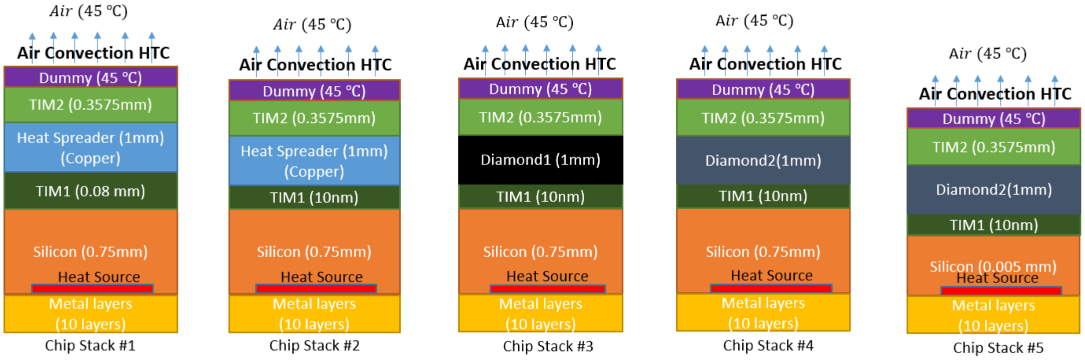

FIGURE 1: CHIP STACKS FOR IBM POWER9 AND INTEL I7 6950X.

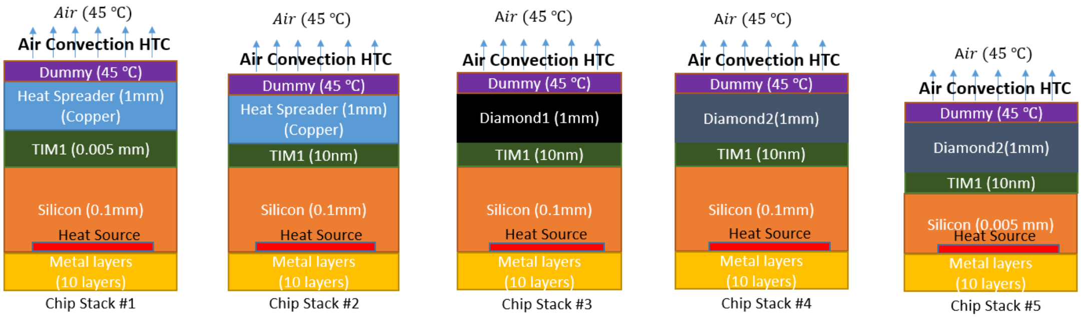

FIGURE 2: CHIP STACKS FOR PICOSOC.

2.1 PACT

We design and implement PACT [10] to enable fast and accurate standard-cell level to architectural level thermal simulations. PACT has the following features: (i) it utilizes the parallelism in modern computing systems to conduct parallel thermal simulations to speed up the process of solving problems with a large number of grid nodes (e.g., for standard-cell level problems or modeling the ultra-thin layers in a monolithic 3D stack), (ii) it offers support for various steady-state and transient solvers to speed up simulation time and maintain the desired accuracy level, and (iii) it can be easily extended to support emerging integration and cooling technologies by modifying the thermal netlist. The current version of PACT supports the modeling of conventional silicon chips with heat sinks/spreaders, 3D ICs with through-silicon vias (TSVs) (die-stacked 3D), monolithic 3D systems, and liquid cooling via microchannels. We also interface PACT to OpenROAD [14], an end-to-end silicon compiler. Users can evaluate the thermal behavior of full standard-cell designs from OpenROAD using PACT. We will release new features such as conventional temperature-dependent thermal resistance simulation framework and learning-based temperature-dependent HTC simulation framework in the later version of PACT, which is used to carry out thermal simulations with two-phase cooling and lab-grown diamond heat spreader. PACT aims to address the fragmentation in the thermal modeling tool space and provides a single tool that is able to conduct efficient thermal evaluation from standard-cell level to architecture-level. We validate PACT's accuracy by comparing it to COMSOL [5], using full standard-cell level industrial designs provided by OpenROAD. Compared to COMSOL, PACT has a maximum temperature error of 2.77% for steady-state and 3.28% for transient simulation. We also compare the simulation time to HotSpot [15], a popular architectural compact thermal simulator, using full industrial designs. HotSpot has been recognized as one of the fastest compact thermal simulators and achieves 216-43300X speedups compared to COMSOL [16]. When compared to HotSpot, PACT achieves speedups of up to 1.83× and 186× for steady-state and transient simulation, respectively. The ambitious goal with PACT is to release a thermal simulator that provides speedy and accurate thermal simulations and, at the same time, caters to a vast number of (future) designers and technologies with different needs and goals, without requiring a substantial redesign of the tool. PACT is open-sourced at https://github.com/peaclab/PACT.

2.2 Processor model

We build three different real-world processor models in PACT. Intel i7 6950X [17] is a desktop processor, IBM Power9 processor [18] is a server processor, and PicoSoC [14] is a mobile processor. For Intel i7 6950X and IBM Power9, we model the processors based on the reported architecture-level floorplan and TDP (220 W for IBM Power9 and 140 W for Intel i7 6950X). For PicoSoC, we directly utilized the coordinates and power values of the standard cells to generate standard-cell level power maps using OpenROAD [14]. We assume an extreme power case for PicoSoC with an operating frequency of 3 GHz and a total power of 9 W. For Intel i7 6950X and IBM Power9, we create 5 different chip stacks to compare the cooling performance of lab-grown diamond heat spreaders and traditional copper heat spreaders. Figure 1 shows the chip stacks for Intel i7 6950X and IBM Power9. Since PicoSoC is a mobile chip, using heat sinks is not possible with mobile chips due to size/volume constraints, and package temperature constraints are typically stricter compared to desktop and server processors. In this case, we build 5 additional chip stacks with no heat sinks and TIM2 for PicoSoC as shown in Figure 2. Besides, the TIM1 layer is relatively thinner compared to the desktop and server processors. We assign a fix-air convection HTC on top of the chip stacks to represent the forced-air cooling via fans package. Chip stack #1 is used to mimic the real-world processor with copper heat spreader and chip stack #3 represents a real-world processor with a lab-grown diamond heat spreader ($TC_{diamond1} = 7.28(T)^{-1.42}\ MW/mK$). Chip stack #2 is used to set up a direct comparison of the cooling performance between the copper heat spreader and diamond heat spreader. Note that, one of the advantages of lab-grown diamond heat spreaders is that they can be directly bonded to the silicon with an ultra-thin TIM layer (or no TIM layer is needed) whereas traditional copper heat spreaders require a thick TIM layer [19-21]. Comparing chip stack #1 and #3 is more realistic than comparing chip stack #1 and chip stack #2. Chip stack #4 represents a real-world processor with a higher thermal conductivity lab-grown diamond heat spreader ($TC_{diamond2} = 10.9(T)^{-1.42}\ MW/mK$). Chip stack #5 mimics the processor with an ultra-thin processor layer (silicon layer). The floorplans of Intel i7 6950X and IBM Power9 are shown in Figure 3. Since PicoSoC is a standard-cell design, the floorplan of PicoSoC is very similar to a mesh.

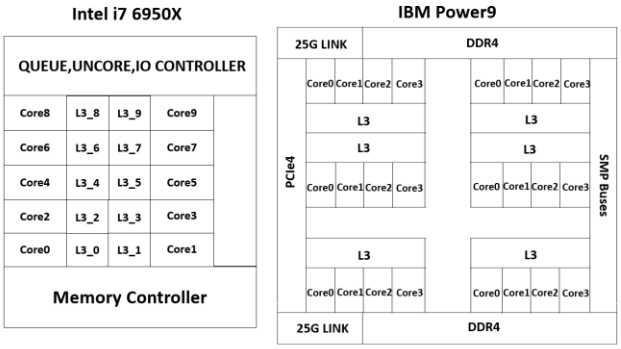

FIGURE 3: INTEL I7 6950X AND IBM POWER9 FLOORPLANS.

2.3 Interconnect model

To mimic the realistic interconnects in the real chips, we add additional interconnect metal layers to the chip stacks and assign additional dynamic power to represent the power consumptions of the power delivery network. Figure 4 shows the layer stack of the interconnect model we added to the chip stack. We assume the flip-chip design and the processing layer is in between the heat spreader and interconnect model. Since metal layers 1-8 are local interconnects, to reduce the simulation problem size, we abstract metal 1-8 layers into one layer and assign a joint thermal resistivity of 75% copper and 25% silicon oxide to this abstract layer. For metal 9 and 10 layers, since these metal layers are used for global connection, we use these two layers to build a power delivery network. The floorplans for metal 9 and 10 layers are shown in Figure 5. The golden lines represent the metal lines and the rest of the layer consists of silicon oxide. To further reduce the simulation problem size, for desktop processor and server processor chip stacks as shown in Figure 1, metal 9 and 10 layers metal width and pitch are set to 200 μm and 400 μm. For mobile processors, to ensure simulation accuracy, metal 9 and 10 layers metal width and pitch are set to 20 μm and 40 μm. The total power consumptions of the interconnect layers are 10% of the total chip power. For metal 1-8, each layer consumes 7.5% of the interconnect power. Metal 9 and 10 consume 40% of the interconnect power.

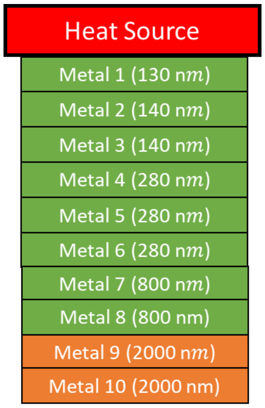

FIGURE 4: THE LAYER STACK OF THE INTERCONNECT MODEL.

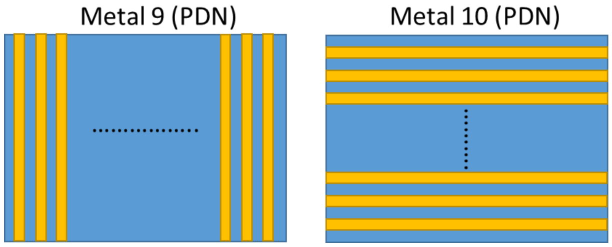

FIGURE 5: POWER DELIVERY NETWORK MODEL.

2.4 Realistic power traces collection

To carry out transient simulations for Intel i7 6950X with realistic power traces, we first use the architecture-level performance simulators such as Sniper [12] to run realistic application benchmarks and then input the program metrics to the power simulator, McPAT [13], to collect the power traces. The power traces are calibrated using the reported TDP and the collected power traces are sent to PACT to carry out transient simulations. We select the parallel applications from NAS Parallel Benchmarks [11] and choose different mapping policies to map different applications to different cores to study the multi-program and multi-threaded workload scenarios. For PicoSoC, since the standard-cell design lacks dynamic power traces, we utilize the steady-state power values of PicoSoC and randomly applied -15% or +15% additional power values for each standard cell and create synthetic transient power traces. Since IBM Power9 uses Power ISA and the architecture-level performance simulators such as Sniper have better support for X86 ISA and less support for RISC ISA such as Power ISA. We only carry out steady-state simulations for IBM Power9.

3. RESULTS AND DISCUSSION

In this section, we first validate the accuracy of the thermal models in PACT, and then we demonstrate the steady-state and transient cooling performance comparison results of traditional copper heat spreaders and diamond heat spreaders. Last but not least, we show the parametric study results of the chip thickness and cooling packages. Note that compact thermal modeling methodology always places the temperature node at the center of the bottom surface of the layer. When we are demonstrating and discussing the temperature of the silicon layer, we are always referring to the temperature of the heat source.

3.1 Validation of the model

We use the following chip stacks as shown in Figure 6 to validate the steady-state accuracy of the thermal models in PACT. The silicon layer has a dimension of 2400x2475 μm^2. There is a 1600x1650 μm^2 hot spot placed at the center of the silicon layer with a heat flux of 265 W/cm^2. The silicon layer consumes a total power of 7 W. And the rest of the layer consumes no power. We build all three chip stacks in both ANSYS and PACT, and directly compared the simulation accuracy. The simulation grid resolution in PACT is set to 100x100. We show the steady-state validation results in Table 1. Layer 0 is the bottom layer and layer 7 is the topmost layer. The maximum steady-state temperature difference between PACT and ANSYS is 0.51 ℃ (chip stack #1). It takes a maximum of 6.97 s to run the steady-state simulations in PACT for the chip stacks shown in Figure 6 with a parallel configuration of 4 cores.

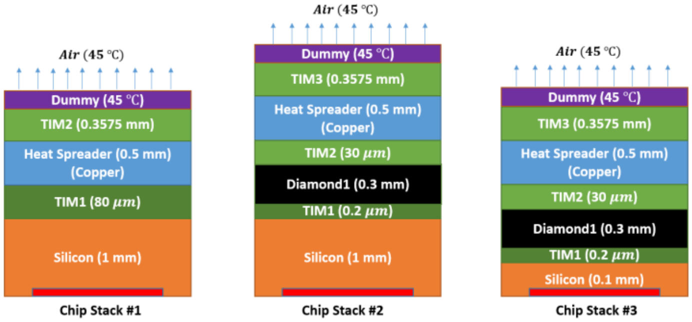

FIGURE 6: CHIP STACKS FOR STEADY-STATE VALIDATION.

	Chip Stack #1		Chip Stack #2		Chip Stack #3	
	PACT	ANSYS	PACT	ANSYS	PACT	ANSYS
Layer#	T_{max} (℃)	T_{max} (℃)	T_{max} (℃)	T_{max} (℃)	T_{max} (℃)	T_{max} (℃)
0	113.49	114.00	107.62	108.0	98.04	98.00
1	102.29	102.00	95.44	95.50	95.81	95.80
2	91.76	92.00	95.41	95.40	95.75	95.75
3	90.25	90.00	95.25	95.00	95.58	95.55
4	48.02	48.00	91.67	91.80	91.73	91.80
5	45.00	45.00	90.16	90.00	90.22	90.20
6	N/A	N/A	48.02	48.00	48.02	48.00
7	N/A	N/A	45.00	45.00	45.00	45.00

TABLE 1: STEADY-STATE VALIDATION RESULTS.

We use the chip stacks and die floorplan as shown in Figure 7 to validate the transient simulation results. The total chip area is 50x50 μm^2 and the die contains 7 power lines colored in red. Each power line is 1 μm wide and 30 μm long. Each power line consumes a uniform power of 120 μW. We switch on and off all the power lines at frequencies of {1, 10, 100, and 1000} Hz for 1 s to validate the transient simulation results accuracy of PACT against ANSYS. The simulation grid resolution is set to 100X100 and the minimum transient solver step size is set to 0.1 ms. We show the transient temperature difference results in Table 2. The maximum transient error when compared to ANSYS is 0.35 ℃ (chip stack #1 @ 1000 Hz). It takes a maximum of 5.93 min to run the transient simulation in PACT for the chip stacks shown in Figure 7 with a parallel configuration of 4 cores.

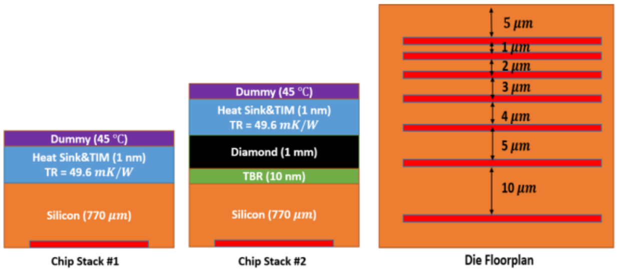

FIGURE 7: CHIP STACKS AND DIE FLOORPLAN FOR TRANSIENT VALIDATION.

	Chip Stack #1				Chip Stack #2			
Freq (Hz)	1	10	100	1000	1	10	100	1000
T_{diff} (℃)	0.27	0.28	0.28	0.35	0.2	0.22	0.25	0.27

TABLE 2: TRANSIENT VALIDATION RESULTS.

3.2 Steady-state comparisons

In this subsection, we compare the steady-state simulation results for the IBM Power9, Intel i7 6950X, and PicoSoC using the chip stacks as shown in Figures 1 and 2. For IBM Power 9 and Intel i7 6950X, we use a grid resolution of 400X400 with grid sizes of 68.5X63.3 μm^2 and 36.7X42 μm^2, respectively. For PicoSoC, since it's a standard cell design with a fine granularity power map and floorplan, we select to use a high grid resolution of 1024X1024 with a grid size of 1.46X1.46 μm^2. The grid resolution is selected based on the size of the smallest functional unit of the processor. The selected grid size has a similar size as the smallest functional unit of the processor. Since Intel i7 6950X and IBM Power9 have an architectural level floorplan, the grid resolution is relatively coarse compared to the standard-cell level floorplan of PicoSoC. Based on cooling packages recommended by Intel, we use a high air convection heat transfer coefficient of $30\ KW/m^2K$. For IBM Power9 and PicoSoC, we choose to use air convection heat transfer coefficients of $20\ KW/m^2K$ and $1\ KW/m^2K$, respectively. The convection heat transfer coefficient values are adopted from previous work [22]. We obtain the steady-state power map of Intel i7 6950X by running Sniper and McPAT with applications bt, cg, dc, ep, ft, is, lu, mg, sp, and ua from NAS parallel benchmarks and average the transient power traces. The steady-state power map has been calibrated to the reported TDP from Intel. For IBM Power9, we use the reported TDP and power breakdown from previous work [17] to calculate the power values for core, L3 cache, Nest, I/O, and DDR4 memory controller. We extract the steady-state power map of PicoSoC by running the OpenROAD project and the interface between OpenROAD and PACT. Figures 8-10 show the steady-state heat map comparisons of Intel i7 6950X, IBM Power9, and PicoSoC with chip stacks #1 and #3. Chip stack #1 is the more realistic chip stack with a traditional copper heat spreader and chip stack #3 is a realistic chip stack with a relatively lower diamond thermal conductivity lab-grown diamond heat spreader. Chip stack #2 is just for direct cooling performance comparison of the heat spreaders with the assumption that the traditional copper heat spreaders can be directly bonded to the silicon layer. We show the Intel i7 6950X, IBM Power9, and PicoSoC steady-state layers 0-5 simulation results for all the chip stacks in Tables 3-5. Layers 0, 1, and 2 are the metal 10, 9, and 1-8 layers, respectively. Layer 3 is the silicon layer and layer 4 is the TIM layer that is placed above the silicon layer. Layer 5 is the heat spreader/diamond layer. The maximum steady-state simulation time of PACT for these high-performance chips is 259 s.

Based on our observations from Figures 8-10 and Tables 3-5, replacing the traditional copper heat spreaders with lab-grown diamond heat spreaders can achieve at least 12.49 ℃ (IBM Power9 chip stacks #1 and #3) and 1.89 ℃ (PicoSoC chip stacks #1 and #3) maximum temperature and thermal gradient reductions, respectively. For Intel i7 6950X and IBM Power9 with lab-grown diamond heat spreaders, the maximum temperatures on-chip are less than 81 ℃. The throttling temperature for mobile processors is around 70-80 ℃ (depends on the specific model of processor), with lab-grown diamond heat spreaders, the maximum temperature of PicoSoC is less than 76 ℃.

The above temperature reductions are mainly because the thermal conductivity of diamond is higher than copper and the diamond heat spreaders can be directly bonded to the silicon layer which results in lower vertical thermal resistance. In addition, we also observe that the temperature and thermal gradient reductions are highly correlated with the chip stack thickness. For Intel i7 6950X and IBM Power9, when switching the diamond thermal conductivity from $7.28(T)^{-1.42}\ MW/mK$ to $10.9(T)^{-1.42}\ MW/mK$ (chip stacks #3 and #4), the maximum temperature of the chip barely changes. The reason is that the vertical thermal resistance of the chip stack is dominated by the thick TIM and silicon layers. Using a high thermal conductivity diamond heat spreader cannot provide significant benefits to the temperature reductions. Whereas, for PicoSoC, the chip stack is much thinner compared to Intel i7 6950X and IBM Power9. When comparing PicoSoC chip stacks #3 and #4, we observe a maximum temperature reduction of 3.69 ℃. In addition, when we scale the silicon layer thickness to $5\ \mu m$, the maximum temperature and thermal gradient reductions increase to 19.76 ℃ (PicoSoC chip stacks #1 and #5) and 15.69 ℃ (IBM chip stacks #1 and #5), respectively. The hot spot locations and the number of hot spots also affect the maximum temperature and thermal gradient reductions. For Intel i7 6950X and PicoSoC, since the hot spots are gathering in the silicon layer, we observe maximum temperature reductions of 13.75 ℃ and 13.21 ℃ (chip stack #1 and chip stack #3), respectively. However, for IBM Power9, the hot spots are spread and that's why the temperature reduction is lower compared to Intel i7 6950X and PicoSoC. In summary, compared to traditional copper heat spreaders, lab-grown diamond heat spreaders can achieve maximum steady-state temperature and thermal gradient reductions of 19.76 ℃ and 15.69 ℃, respectively.

Layer	0	1	2	3	4	5
$T_{max}(ID1)$(℃)	84.64	84.64	84.60	84.54	81.13	70.43
$T_{max}(ID2)$(℃)	77.77	74.77	74.73	74.67	71.08	71.07
$T_{max}(ID3)$(℃)	70.89	70.89	70.85	70.79	67.00	66.99
$T_{max}(ID4)$(℃)	70.47	70.47	70.44	70.37	66.56	66.55
$T_{max}(ID5)$(℃)	66.76	66.76	66.72	66.66	66.63	66.62
$\Delta T(ID1)$(℃)	17.50	17.50	17.46	17.43	14.98	7.35
$\Delta T(ID2)$(℃)	11.10	11.10	11.06	11.04	8.36	8.35
$\Delta T(ID3)$(℃)	5.39	5.39	5.35	5.32	2.31	2.31
$\Delta T(ID4)$(℃)	4.72	4.72	4.69	4.65	1.59	1.59
$\Delta T(ID5)$(℃)	1.81	1.81	1.77	1.75	1.72	1.71

TABLE 3: INTEL I7 6950X STEADY-STATE RESULTS.

Layer	0	1	2	3	4	5
$T_{max}(ID1)$(℃)	92.82	92.82	92.82	92.80	89.12	77.58
$T_{max}(ID2)$(℃)	82.59	82.59	82.59	82.57	78.23	78.22
$T_{max}(ID3)$(℃)	80.33	80.33	80.33	80.31	75.70	75.69
$T_{max}(ID4)$(℃)	80.10	80.10	80.10	80.08	75.45	75.44
$T_{max}(ID5)$(℃)	75.58	75.58	75.58	75.56	75.53	75.51
$\Delta T(ID1)$(℃)	14.03	14.03	14.03	14.01	11.59	3.97
$\Delta T(ID2)$(℃)	8.32	8.32	8.32	8.33	5.05	5.04
$\Delta T(ID3)$(℃)	5.11	5.11	5.11	5.09	1.40	1.38
$\Delta T(ID4)$(℃)	4.72	4.72	4.72	4.72	0.98	0.98
$\Delta T(ID5)$(℃)	1.14	1.14	1.14	1.13	1.11	1.09

TABLE 4: IBM POWER9 STEADY-STATE RESULTS.

Layer	0	1	2	3	4	5
$T_{max}(ID1)$(℃)	88.89	88.89	88.88	88.65	84.61	78.68
$T_{max}(ID2)$(℃)	83.10	83.10	83.09	82.86	78.78	78.69
$T_{max}(ID3)$(℃)	75.68	75.68	75.67	75.44	69.35	69.25
$T_{max}(ID4)$(℃)	71.99	71.99	71.98	71.75	68.66	68.56
$T_{max}(ID5)$(℃)	69.13	69.13	69.12	68.95	68.69	68.57
$\Delta T(ID1)$(℃)	3.92	3.93	3.94	3.87	2.43	0.34
$\Delta T(ID2)$(℃)	2.30	2.31	2.32	2.24	0.42	0.38
$\Delta T(ID3)$(℃)	2.03	2.05	2.06	1.98	0.12	0.07
$\Delta T(ID4)$(℃)	2.00	2.03	2.04	1.95	0.09	0.04
$\Delta T(ID5)$(℃)	0.40	0.41	0.43	0.41	0.17	0.06

TABLE 5: PICOSOC STEADY-STATE RESULTS.

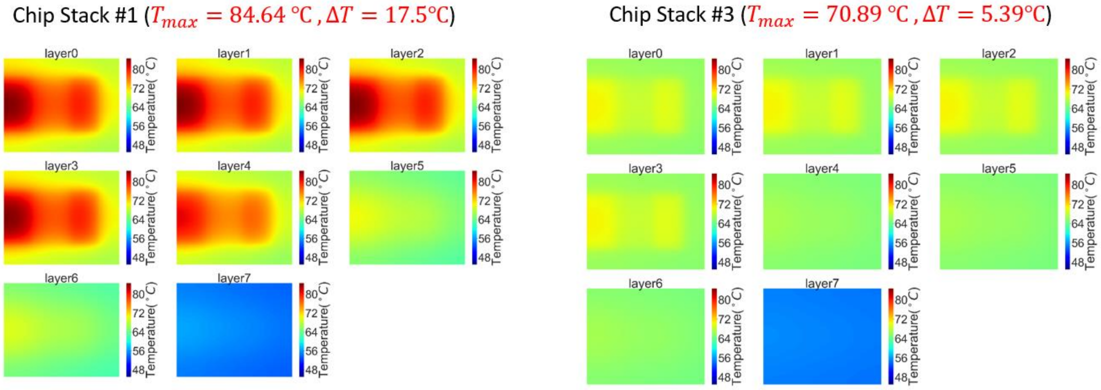

FIGURE 8: STEADY-STATE HEAT MAP COMPARISONS OF INTEL I7 6950X (CHIP STACKS #1 AND #3).

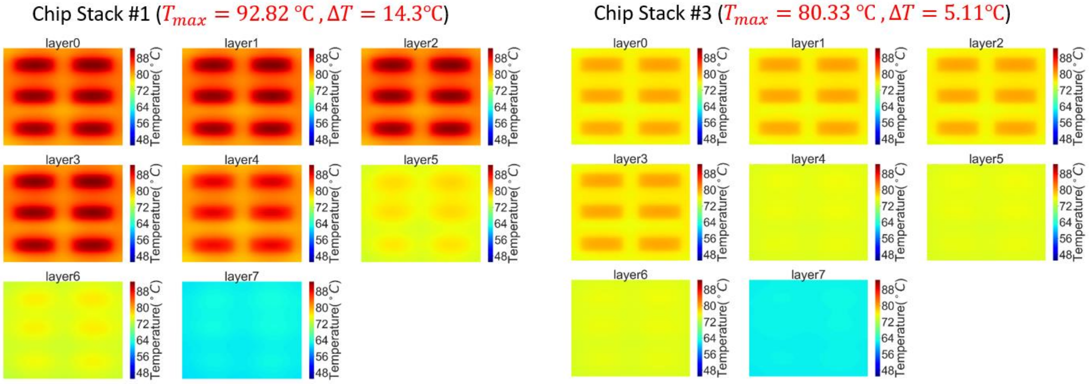

FIGURE 9: STEADY-STATE HEAT MAP COMPARISONS OF IBM POWER9 (CHIP STACKS #1 AND #3).

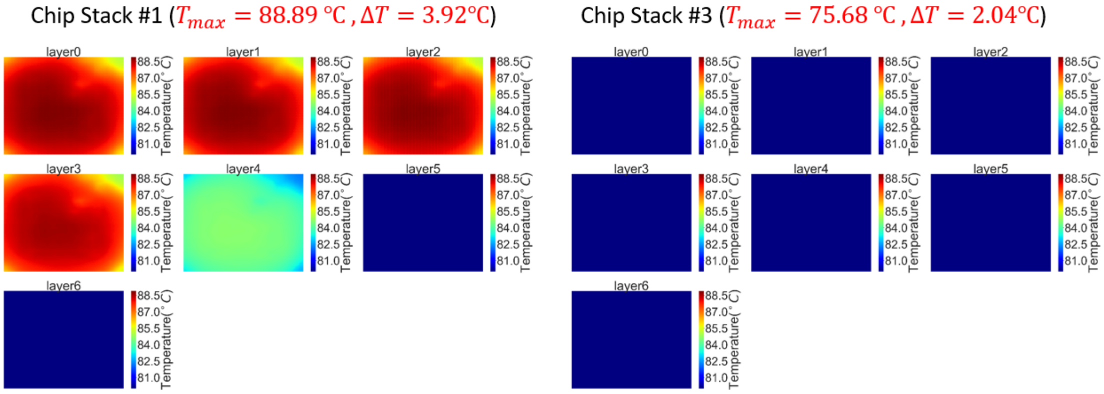

FIGURE 10: STEADY-STATE HEAT MAP COMPARISONS OF PICOSOC (CHIP STACKS #1 AND #3).

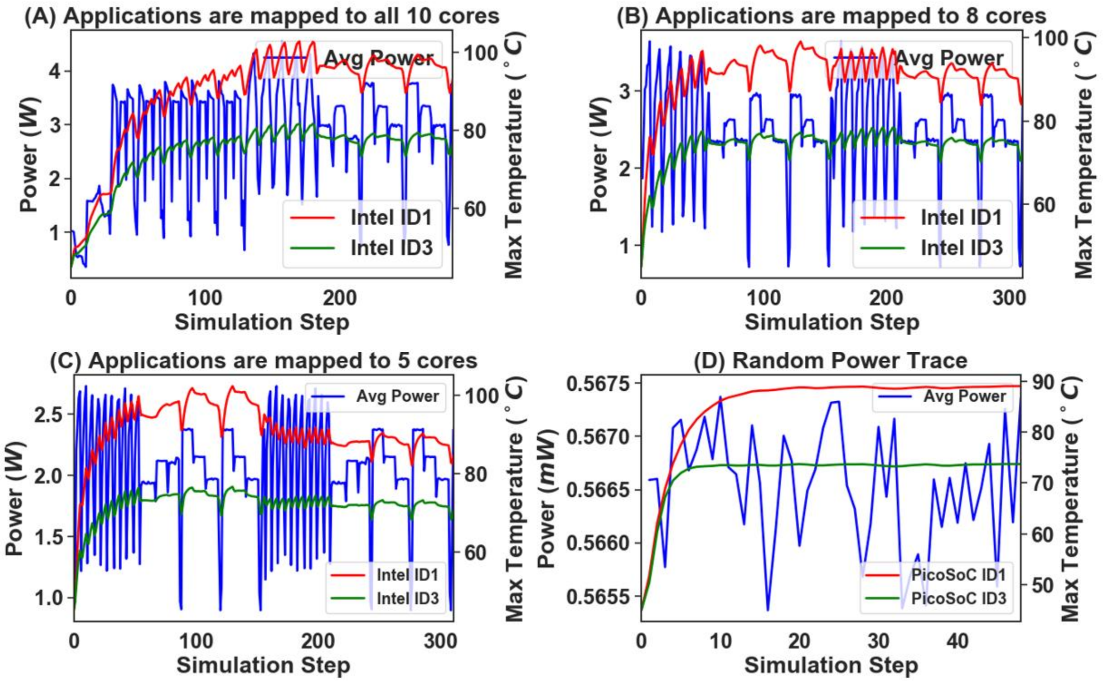

FIGURE 11: TRANSIENT TEMPERATURE PLOTS FOR INTEL I7 6950X and PICOSOC.

3.3 Transient comparisons

Next, we carry out transient simulations for Intel i7 6950X and PicoSoC. We use chip stacks #1 and #3 as shown in Figures 1 and 2. For Intel i7 6950X, we obtain the transient power traces by running applications bt, cg, and ft from NAS parallel benchmarks. The transient power traces have been calibrated to the reported TDP from Intel. We run 10 billion instructions for each application and collected power values per 10 million instructions to extract the application power traces. We select different application mapping policies to study the transient thermal behavior of traditional copper heat spreaders and lab-grown diamond heat spreaders. The selected application mapping policies are as follows: (i) we run most power-hungry applications bt and ft consecutively and applications are mapped to all 10 cores, (ii) we run application ft for two iterations. In the first iteration, cores 8 and 9 remain idle and in the second iteration, cores 4 and 5 are idle, (iii) we run applications ft (high power) and cg (low power) for two iterations. In the first iteration, each application is mapped on a column of cores, and in the second iteration, applications are mapped as a checkerboard. The transient temperature plots of the Intel i7 6950X silicon layer are shown in Figure 11. The maximum transient simulation time of PACT for these real-world high-performance chips with realistic applications is 22 min.

Plots (A) and (C) indicate that the maximum transient temperature reductions of lab-grown diamond heat spreaders can go up to 26.73 ℃. In addition, the temperature reductions of the diamond heat spreaders depend on the application behavior and application mapping policy. As we see in plots (A), (B), and (C), using different application mapping policies for application ft results in different maximum temperatures. For plot (B), leaving cores 4 and 5 idle results in a lower maximum temperature than making cores 8 and 9 idle. As shown in Figures 3 and 8, cores 4 and 5 are placed at the center of the chip and result in the highest hot spot temperatures. Leaving cores 4 and 5 idle is very similar to adding white spaces to the hot spot region to decrease the hot spot temperatures. For plot (C), since the checkerboard mapping policy help spread the lateral heat, the second iteration results in significant temperature reduction compared to the first iteration.

For PicoSoC, since the standard-cell design lacks dynamic power traces, we utilize the steady-state power values of PicoSoC and randomly applied -15% or +15% additional power values for each standard cell and create synthetic transient power traces. The transient temperature plots of PicoSoC are shown in Figure 11. We still observe a maximum temperature reduction of 18 ℃ for mobile chip setup. These transient simulation comparison results show the transient cooling performance advantages of lab-grown diamond heat spreaders over traditional copper heat spreaders are even more than the steady-state thermal simulations. The reason is that for steady-state simulations, we average the power values of the applications in NAS parallel benchmarks which results in relatively lower steady-state power values compared to transient power values. In addition, we haven't considered mapping policies in the steady-state simulations. As we see in the transient temperature plots, mapping policies also have high impacts on the maximum temperature reductions.

3.4 Parametric study of chip thickness

In this subsection, we study the cooling performance of lab-grown diamond heat spreaders with different chip thickness. We select Intel i7 6950X with chip stacks #1 and #3 as shown in

Figure 2, and the chip thickness is selected to be {5, 50, 100, 250, 500, and 750} μm. The lab-grown diamond thermal conductivity is set to $7.28(T)^{-1.42}\ MW/mK$. We obtain the steady-state power map of Intel i7 6950X by running Sniper and McPAT with the most power-hungry applications bt, and ft from NAS parallel benchmarks and average the transient power trace. The steady-state power map has been calibrated to the reported TDP from Intel. The steady-state maximum temperature results are shown in Figure 12. Decreasing the thickness of the silicon layers helps to lower the vertical thermal resistance of the chip stack. However, in the meantime, it also prevents spreading the lateral heat across the silicon layer. For chip stack #1, the vertical thermal resistance is dominated by the thick TIM layers, and varying the silicon layer thickness does not affect the maximum temperature much. Whereas for chip stack #3, diamond heat spreaders have lower thermal resistance and can be directly bonded to the silicon. By lowering the silicon layer thickness, we can see a maximum temperature reduction of 4.62 ℃ (thickness = 5 μm vs. 750 μm).

We then conduct a parametric study of silicon layer thickness for transient simulations of Intel i7 6950X chip stacks #1 and #3. We run most power-hungry applications bt and ft consecutively and applications are mapped to all 10 cores. The transient power traces have been calibrated to the reported TDP from Intel. We show the transient simulation parametric study results in Figures 13 and 14. We observe a similar trend as steady-state thickness parametric study. For chip stack #1, decreasing the thickness of the silicon layer does not affect the maximum temperatures because of the tradeoff between vertical and lateral thermal resistance. Whereas for chip stack #3, decreasing the thickness of the silicon layer results in a maximum temperature reduction of 6.53℃ (thickness = 5 μm vs. 750 μm). Based on the steady-state and transient parametric studies of the silicon layer thickness, we show that using a thinner silicon layer can achieve an even better cooling performance than a thick silicon layer for lab-grown diamond heat spreaders.

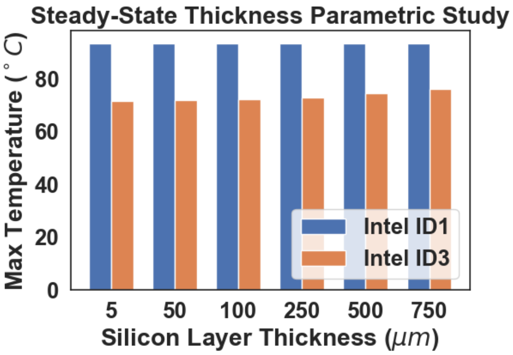

FIGURE 12: STEADY-STATE SILICON LAYER THICKNESS PARAMETRIC STUDY RESULTS FOR INTEL I7 6950X.

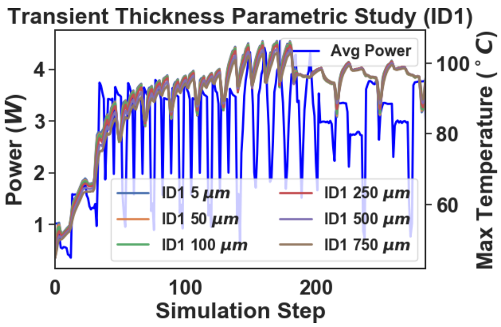

FIGURE 13: INTEL I7 6950X CHIP STACK #1 SILICON LAYER THICKNESS TRANSIENT PARAMETRIC STUDY RESULTS.

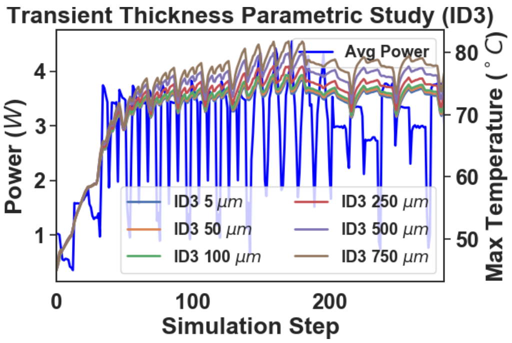

FIGURE 14: INTEL I7 6950X CHIP STACK #3 SILICON LAYER THICKNESS TRANSIENT PARAMETRIC STUDY RESULTS.

3.5 Parametric study of cooling packages

Next, we study the cooling performance of lab-grown diamond heat spreaders with different heat sinks. We select Intel i7 6950X with chip stacks #1 and #3 as shown in Figure 2, and the heat sink is selected to be the fix-air convection heat sink, single-phase liquid cooling via microchannels, and medium-cost heat sink adopted from HotSpot [15]. For fixed-air convection heat sink, we set the air convection HTC to $30\ KW/m^2K$. For the medium-cost heat sink, we set the heat sink size and thickness to 0.4X0.4 mm^2 and 1 mm, respectively. The heat sink is made of copper. The convection resistivity and heat capacity are set to 0.21 K/W and 140.4 J/K, respectively. We calculate the convection resistivity based on the air convection HTC of $30\ K\ W/m^2K$, which is the same as the fixed-air convection heat sink. For liquid cooling via microchannels, the selected material properties are shown in Table 6. The liquid cooling via microchannels model integrated with PACT simulator is a single-phase liquid cooling method and has already been validated against COMSOL and a popular liquid cooling via microchannel compact thermal simulator [10, 16, 23]. The detailed information of the water-side heat transfer coefficient correlation and experimental setups can be found in previous work [10, 16, 23]. The selected coolant velocity and the Reynolds number indicate the type of fluid flow is laminar. We obtain the steady-state power map of Intel i7 6950X by running Sniper and McPAT with the most power-hungry applications bt, and ft from NAS parallel benchmarks and average the transient power trace to represent the steady-state power map. The steady-state power map has been calibrated to the reported TDP from

Intel. We show the steady-state silicon layer heat maps in Figure 15 and parametric study results in Figure 16. By replacing the fixed-air convection heat sink with a medium-cost heat sink, we can see a maximum temperature reduction of 9.66 ℃ (ID1_Fixed_Air vs. ID1_Medium_Cost). This is mainly because the size of the heat sink is larger than the chip stack and therefore enhances the lateral heat transfer. For liquid cooling via microchannels, as the liquid flow velocity increase, liquid cooling via microchannels starts to become the best cooling package with a maximum temperature reduction of 12.71 (ID1_Fixed_Air vs. ID1_Liquid @ 2.6 m/s). However, when considering the thermal gradients, the medium-cost heat sink performs better than liquid cooling via microchannels as shown in Figure 15. Since liquid absorbs heat as it flows along the channel, the temperature difference between the inlet and outlet is one of the major reasons for the high thermal gradient. Another reason is the thermal resistivity difference between the liquid and wall, which causes the high lateral thermal gradient compared to the medium-cost heat sink.

For the transient parametric study, we use the same setup for the three types of aforementioned cooling packages. We run most power-hungry applications bt and ft consecutively and applications are mapped to all 10 cores to obtain the transient power traces. The transient power traces have been calibrated to the reported TDP from Intel. The transient temperature plots are shown in Figure 17. Compared to steady-state results, we observe a higher maximum temperature reduction of liquid cooling via microchannels against the other two heat sinks. The maximum temperature reduction against the fix-air convection heat sink is 14.13℃. This is due to the high specific heat capacity of the water compared to silicon and copper. In summary, among these three types of heat sinks, liquid cooling via microchannels can provide the highest cooling performance and results in the lowest maximum temperature on-chip.

Coolant	Water
Thermal Resistivity	1.647 mK/W
Specific Heat Capacity	4.181 MJ/m^3K
Inlet Temperature	27 ℃
Fluid Density	998 Kg/m^2
Dynamic viscosity	0.000889 $Pa \cdot s$
Coolant Velocity	{0.5, 1.0, 1.5, 2.0, 2.6} m/s
Reynolds Number	{37.4, 74.8, 112, 195}
Number of Microchannels	146
Microchannel Width	50 μm
Wall Width	50 μm
Wall Material	Silicon
Microchannel Height	100 μm
Microchannel Hydraulic Diameter	66.67 μm

TABLE 6: LIQUID COOLING VIA MICROCHANNELS MATERIAL PROPERTIES.

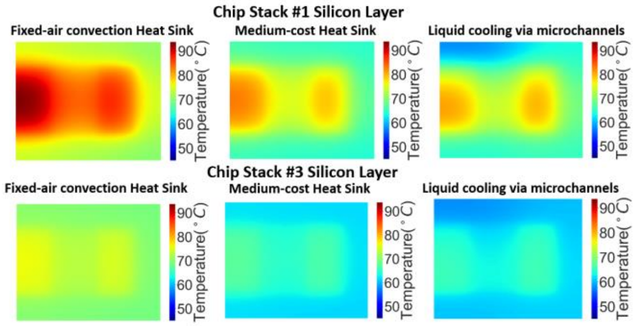

FIGURE 15: STEADY-STATE HEAT MAPS FOR PARAMETRIC STUDY OF COOLING PACKAGES. LIQUID FLOW VELOCITY IS SET TO 2.6 M/S.

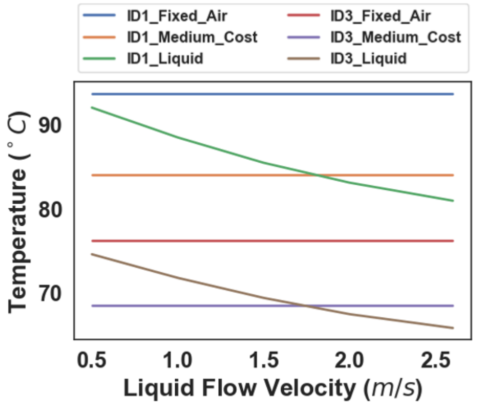

FIGURE 16: INTEL I7 6950X SILICON LAYER STATE-STATE COOLING PACKAGE PARAMETRIC STUDY RESULTS.

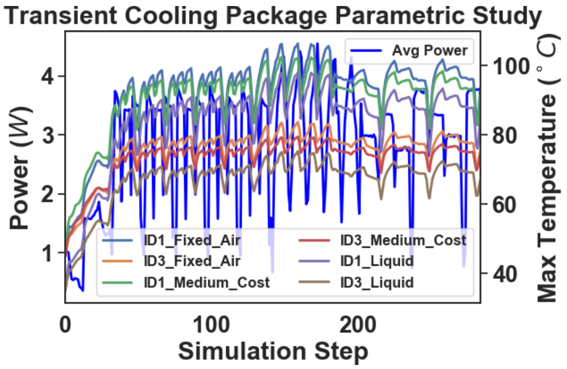

FIGURE 17: INTEL I7 6950X SILICON LAYER TRANSIENT COOLING PACKAGE PARAMETRIC STUDY RESULTS.

4. CONCLUSION

In this paper, we use a parallel compact thermal simulator, PACT, to compare the cooling performance of lab-grown diamond heat spreaders against traditional copper heat spreaders using real-world high-performance processors with realistic application benchmarks. We create a interconnect model to mimic the realistic processor chip stacks and carry out both steady-state and transient comparisons. Our results show that, compared to traditional copper heat spreaders, the lab-grown diamond heat spreader can achieve maximum temperature and thermal gradient reductions of up to 26.73 ℃ and 13.75 ℃, respectively. In addition, we also carry out parametric studies of processor thickness (silicon layer thickness) and cooling packages. We observe that the lab-grown diamond heat spreader can achieve a higher temperature reduction with a relatively thinner silicon layer. In addition, to achieve the lowest on-chip maximum temperature, liquid cooling via microchannels outperforms fixed-air convection heat sink and medium-cost heat sink.

REFERENCES

[1] Schultz, Mark, Fanghao Yang, Evan Colgan, Robert Polastre, Bing Dang, Cornelia Tsang, Michael Gaynes, Pritish Parida, John Knickerbocker, and Timothy Chainer. "Embedded two-phase cooling of large three-dimensional compatible chips with radial channels." *Journal of Electronic Packaging* 138, no. 2 (2016): 021005.

[2] Pedram, Massoud, and Shahin Nazarian. "Thermal modeling, analysis, and management in VLSI circuits: Principles and methods." *Proceedings of the IEEE* 94, no. 8 (2006): 1487-1501.

[3] Jagannadham, K. "Multilayer diamond heat spreaders for electronic power devices." *Solid-State Electronics* 42, no. 12 (1998): 2199-2208.

[4] Zhou, L., Y. B. Tian, H. Huang, H. Sato, and J. Shimizu. "A study on the diamond grinding of ultra-thin silicon wafers." *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture* 226, no. 1 (2012): 66-75.

[5] Multiphysics, C. O. M. S. O. L. "Introduction to comsol multiphysics®." *COMSOL Multiphysics, Burlington, MA, accessed Feb* 9 (1998): 2018.

[6] Kohnke, P. C. "Ansys." In *Finite Element Systems*, pp. 19-25. Springer, Berlin, Heidelberg, 1982.

[7] Yuan, Zihao, Geoffrey Vaartstra, Prachi Shukla, Sherief Reda, Evelyn Wang, and Ayse K. Coskun. "Modeling and optimization of chip cooling with two-phase vapor chambers." In *IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 1-6. IEEE, 2019.

[8] Yuan, Zihao, Geoffrey Vaartstra, Prachi Shukla, Zhengmao Lu, Evelyn Wang, Sherief Reda, and Ayse K. Coskun. "A learning-based thermal simulation framework for emerging two-phase cooling technologies." In *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 400-405. IEEE, 2020.

[9] Yuan, Zihao, Geoffrey Vaartstra, Prachi Shukla, Mostafa Said, Sherief Reda, Evelyn Wang, and Ayse K. Coskun. "Two-phase vapor chambers with micropillar evaporators: a new approach to remove heat from future high-performance chips." *In 18th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pp. 456-464. IEEE, 2019.

[10] Yuan, Zihao, Prachi Shukla, Sofiane Chetoui, Sean Nemtzow, Sherief Reda, and Ayse K. Coskun. "PACT: An Extensible Parallel Thermal Simulator for Emerging Integration and Cooling Technologies." *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (2021).

[11] Bailey, David, Tim Harris, William Saphir, Rob Van Der Wijngaart, Alex Woo, and Maurice Yarrow. *The NAS parallel benchmarks 2.0.* Vol. 156. Technical Report NAS-95-020, NASA Ames Research Center, 1995.

[12] Carlson, Trevor E., Wim Heirman, and Lieven Eeckhout. "Sniper: Exploring the level of abstraction for scalable and accurate parallel multi-core simulation." In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1-12. 2011.

[13] Li, Sheng, Jung Ho Ahn, Richard D. Strong, Jay B. Brockman, Dean M. Tullsen, and Norman P. Jouppi. "McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures." In *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 469-480. 2009.

[14] Ajayi, Tutu, Vidya A. Chhabria, Mateus Fogaça, Soheil Hashemi, Abdelrahman Hosny, Andrew B. Kahng, Minsoo Kim et al. "Toward an open-source digital flow: First learnings from the openroad project." In *Proceedings of the 56th Annual Design Automation Conference 2019*, pp. 1-4. 2019.

[15] Skadron, Kevin, Mircea R. Stan, Karthik Sankaranarayanan, Wei Huang, Sivakumar Velusamy, and David Tarjan. "Temperature-aware microarchitecture: Modeling and implementation." *ACM Transactions on Architecture and Code Optimization (TACO)* 1, no. 1 (2004): 94-125.

[16] Kaplan, Fulya, Mostafa Said, Sherief Reda, and Ayse K. Coskun. "Locool: Fighting hot spots locally for improving system energy efficiency." *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 39, no. 4 (2019): 895-908.

[17] Sima, Dezső. "Intel Core X-series (HED lines)." (2018).

[18] Sadasivam, Satish Kumar, Brian W. Thompto, Ron Kalla, and William J. Starke. "IBM Power9 processor architecture." *IEEE Micro* 37, no. 2 (2017): 40-51.

[19] Liang, Jianbo, Satoshi Masuya, Makoto Kasu, and Naoteru Shigekawa. "Realization of direct bonding of single crystal diamond and Si substrates." *Applied Physics Letters* 110, no. 11 (2017): 111603.

[20] Liang, Jianbo, Satoshi Masuya, Seongwoo Kim, Toshiyuki Oishi, Makoto Kasu, and Naoteru Shigekawa. "Stability of diamond/Si bonding interface during device fabrication process." *Applied Physics Express* 12, no. 1 (2018): 016501.

[21] Liang, Jianbo, Yan Zhou, Satoshi Masuya, Filip Gucmann, Manikant Singh, James Pomeroy, Seongwoo Kim, Martin Kuball, Makoto Kasu, and Naoteru Shigekawa. "Annealing effect of surface-activated bonded diamond/Si interface." *Diamond and Related Materials* 93 (2019): 187-192.

[22] Wei, Hai, et al. "Cooling three-dimensional integrated circuits using power delivery networks." *2012 International Electron Devices Meeting*. IEEE, 2012.

[23] Sridhar, Arvind, Alessandro Vincenzi, Martino Ruggiero, Thomas Brunschwiler, and David Atienza. "3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling." In *2010 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 463-470. IEEE, 2010.