LoCool: Fighting Hot Spots Locally for Improving System Energy Efficiency

Fulya Kaplan, *Member, IEEE*, Mostafa Said, *Member, IEEE*, Sherief Reda, *Senior Member, IEEE*, and Ayse K. Coskun, *Senior Member, IEEE*

***Abstract*—Elevated on-chip temperatures significantly degrade performance, energy-efficiency, and lifetime of processors. The cooling system for a chip is typically designed to remove the worst-case heat generated per unit area. Cooling demand, however, spatially and temporally varies across a chip as hot spots occur on different locations with different intensities. Thus, designing a homogeneous cooling system for a chip can be inefficient. Recently, hybrid cooling strategies, such as integrating thermoelectric coolers (TECs) with microchannel liquid cooling, have been explored for hot spot mitigation. The efficiency of such a cooling system strongly depends on the operating point of each cooling method, as well as the locations and intensities of the hot spots. To this end, we first devise a compact thermal modeling method for the design and evaluation of hybrid cooling systems in a fast and accurate way. The proposed model provides up to four orders of magnitude speedup in simulation time compared to COMSOL multiphysics simulations with less than 2.9 °C average temperature error. Leveraging our fast model, we develop LoCool, a hybrid cooling optimization method, which jointly determines the most energy-efficient cooling settings for a given chip power distribution and temperature constraint. LoCool determines the liquid flow rate and the input current for each TEC depending on the cooling requirements for individual hot spots as well as for the background heat. Experimental evaluation shows up to 40% cooling energy savings compared to designing homogeneous cooling systems under the same thermal constraints.**

***Index Terms*—Compact thermal modeling, cooling power optimization, hybrid cooling, liquid cooling, thermoelectric cooling.**

I. Introduction

ELEVATED on-chip temperatures have become a significant limiting factor in the design and energy-efficient operation of processors. High temperatures not only decrease processor lifetime [1] but also they cause larger transistor delays and exponentially increase leakage power [2], [3].

Manuscript received July 20, 2018; accepted January 25, 2019. Date of publication February 27, 2019; date of current version March 18, 2020. This work was supported in part by the NSF CAREER under Grant 1149703, and in part by NSF CRI (CI-NEW) under Grant 1730316/1730003. This paper was recommended by Associate Editor J. Henkel. *(Corresponding author: Fulya Kaplan.)*

F. Kaplan and A. K. Coskun are with the Department of Electrical and Computer Engineering, Boston University, Boston, MA 02215 USA (e-mail: fkaplan3@bu.edu).

M. Said and S. Reda are with the Department of Computer Science, Brown University, Providence, RI 02912 USA, and also with the Department of Engineering, Brown University, Providence, RI 02912 USA.

Digital Object Identifier 10.1109/TCAD.2019.2902355

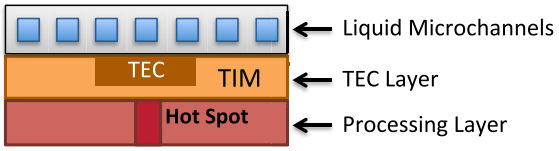

Fig. 1. Front view of an example hybrid design combining microchannel liquid cooling and TECs. TECs are embedded in TIM and are placed on top of high heat flux areas to remove hot spots, while microchannels are used to remove the heat pumped by the TECs and the background heat.

The current trend in processor cooling is to design the system to remove the worst-case (or close to worst-case) thermal design power per unit area. However, the large spatial variations in cooling demand exist across the chip. Localized hot spots occur at different locations with different sizes and intensities. Hot spots with areas as small as 0.04 mm^2 and with heat fluxes reaching 1–2 kW/cm^2 are anticipated in the next generation processors [4], [5]. This heterogeneity in on-chip heat distribution is likely to increase with the integration of heterogeneous architectures on a single die, such as a collection of CPUs, GPUs, accelerators, and FPGAs. Designing a homogeneous cooling system to remove such high (but local) power densities can lead to undercooling of the hot spots or overcooling of the rest of the chip, thus, significantly lowering efficiency. In order to achieve the high energy efficiency and reduce cooling power, it is essential to customize the cooling subsystem based on the demand across the chip.

A variety of cooling methods exist to remove hot spots today. Efficiency and benefits of cooling methods differ based on the target system properties and the power densities they aim to handle. For example, microchannel liquid cooling is well-suited to remove the background heat on large chips and 3-D-stacked architectures. However, the fluid gets hotter as it flows along the channels; thus, the heat removal capability decreases on the locations that they are far away from the liquid inlet [6]. Thermoelectric coolers (TECs) are successful in handling high power densities in small areas, but they consume substantially larger power when used for cooling large areas. Thus, a hybrid cooling strategy that combines the strengths of different cooling methods can provide much higher efficiency. Fig. 1 illustrates an example of such a hybrid cooling design, where TECs are placed on top of hot spots and a microchannel liquid cooling layer is placed on top of the TEC layer.

Recent hybrid cooling designs include combining TECs with microchannel liquid cooling [7]–[9] or with fan cooling [10], [11]. Existing work on hybrid TEC and liquid cooling mostly focuses on optimizing the dimensions and bias current of TECs, assuming a fixed operating point for liquid cooling [7], [8]. We observe that liquid flow rate has an impact on both the liquid pumping power and the cooling performance

of the TECs; thus, a joint optimization approach is necessary to achieve high efficiency. On the other hand, previous work on hybrid TEC and fan cooling proposes to optimize the TEC and fan power together [10], [11]. However, they target much lower heat fluxes ($\sim$20–28 W/cm^2) and do not focus on localized use of the TECs for high-intensity hot spots.

To enable any design-time or runtime optimizations and evaluations, we need fast and accurate thermal modeling tools. These models will enable researchers to explore the design space at a broader scale, develop their own optimization techniques easily, and provide the means of fair comparison against the state-of-the-art. There are a number of compact thermal models developed to simulate the behavior of the microchannel-based liquid cooling [12]–[16] and supperlattice-based TECs [8], [10], [17], [18]. However, the existing thermal models are not sufficient to be integrated with system-level hybrid cooling optimization algorithms. The first reason is that these models focus on modeling the two cooling methods separately, thus, they do not provide the ability to model a hybrid-cooled system. In order to optimize a hybrid cooling design, however, we need to account for the interactions between different cooling elements, as they significantly impact the overall benefit. Yazawa *et al.* [8] demonstrated hybrid cooling benefits in simulation environment, where they use compact models for TECs, but represent the effect of the microchannel-based liquid cooling by defining a high value of effective heat transfer coefficient at the boundary. Using this simplified way of modeling microchannels, one cannot capture the increase in liquid temperature as it flows from inlet to outlet, leading to a considerable loss of accuracy. The second reason why existing models are not adequate in integrating with the hybrid optimization techniques is that they are often designed targeting specific platforms and scenarios (e.g., a specific chip stack with fixed hot spot properties, see [8], [10]), and are not easily applicable to arbitrary systems. For the design of hybrid cooling optimizers, a thermal model should be sufficiently modular enough to provide flexibility to explore a variety of systems in a fast manner. Our proposed compact model combines all these necessary aspects on a unified framework, and it requires minimal user effort when system assumptions change.

Commercial multiphysics simulators, such as COMSOL Multiphysics [19], and ANSYS [20] are able to model hybrid cooling with very high accuracy. However, such tools are prohibitively expensive as it takes substantially long time to construct system-specific models, and at runtime, they incur long time solutions as well as large memory requirements (e.g., simulating an mm scale chip takes from hours to multiple days and requires tens of GBs of memory). Such factors limit the use of multiphysics simulators for modeling hybrid cooling, especially in optimization scenarios where many designs need to be evaluated.

To address this need for fast models, we have recently proposed a thermal modeling methodology to simulate the steady-state behavior of hybrid cooling systems with microchannel liquid cooling and TECs [21]. Our approach uses compact thermal modeling and we integrate our model into a commonly used simulator, HotSpot [22]. Our compact model, for the first time, provides a fast and modular way of steady-state thermal evaluation with sufficient accuracy. In this paper, we use our fast model to design LoCool, a hybrid cooling optimization algorithm to maximize cooling efficiency in systems with high heat flux hot spots. LoCool optimizes hybrid cooling systems involving liquid microchannels and TEC cells, achieving significant reduction in cooling power. Our main contributions of this paper are as follows.

1) We design a compact hybrid cooling model and implement this model in a commonly used thermal simulator to enable design and co-optimization of hybrid cooling systems. We validate the accuracy of our TEC model by comparing it against COMSOL Multiphysics software and our liquid cooling model by comparing it against both COMSOL and 3D-ICE [13], which has been validated using ANSYS. We finally compare our hybrid cooling model against COMSOL. Our model achieves up to four orders of magnitude faster simulation with less than 2.9 °C average and 5.7 °C maximum temperature error.
2) We propose LoCool, a method to co-optimize TEC and liquid cooling design. To the best of our knowledge, this paper is the first to provide an optimization method for hybrid cooling systems. Given a chip power map and thermal constraints, LoCool jointly tunes both cooling methods, namely the liquid flow rate and the bias current for the TEC units, to meet the given temperature constraint using minimum cooling power. LoCool includes both design-time and runtime optimization modules and can adapt to changes in the hot spot heat flux (HSHF) over time.
3) Using our cooling model and LoCool, we demonstrate up to 40% reductions in cooling power in comparison to homogeneous cooling designs. We also provide an analysis of power-temperature tradeoffs across a wide range of cooling design choices. We show that in addition to save cooling power, hybrid cooling with LoCool can mitigate hot spots with much higher heat fluxes (up to 1600 W/cm^2), which are not achievable using liquid cooling only. Finally, we discuss the impact of the number of hot spots on the resulting cooling power savings.

The organization of this paper is as follows. In Section II, we describe the prior work on the existing thermal modeling and optimization techniques developed for systems with liquid cooling, TECs, and hybrid cooling. In Section III, we give the details of our hybrid cooling model, its implementation in a compact simulator, and provide a validation approach using COMSOL. We introduce LoCool optimization algorithm in Section IV. In Section V, we first provide the results of the thermal model validation using COMSOL. We then evaluate the performance of hybrid cooling with LoCool. We conclude in Section VI.

II. Related Work

This section provides an overview of the advanced cooling methods, specifically single-phase microchannel liquid cooling, TEC cooling, and hybrid cooling that combines two or more cooling methods on the same platform. For each of the cooling methods, we first discuss existing thermal modeling approaches. We then provide an overview of the previously proposed design-time and runtime optimization techniques in each cooling domain.

A. Microchannel Liquid Cooling

Liquid cooling with microchannels is an attractive solution for especially 3-D-stacked architectures, where the temperature problem is escalated due to the vertical layer stacking. Various researchers focus on fast and accurate modeling of the liquid-cooled ICs [12]–[16]. Coskun *et al.* [12] incorporated a

liquid cooling model into HotSpot-4.01 simulator, where a grid level thermal resistor–capacitor (*RC*) network is constructed and thermal properties of different interlayer materials (i.e., TSVs and microchannels) are specified. Sridhar *et al.* proposed 3D-ICE [13], which has the ability to model the thermal gradient between the inlet and outlet ports introduced by the flow of the liquid. They validate 3D-ICE against ANSYS CFX computational fluid dynamics tool. The follow-up of their work [14] adds the support for modeling enhanced heat transfer geometries, such as pin-fin structures. This model also simplifies the computation in the microchannel layers by homogenizing them into porous medium. Another body of work focuses on speeding up the long simulation time observed in liquid-cooled ICs [15], [16]. ICTherm [15] is a recently introduced simulator that implements an efficient algorithm to compute the transient temperature in linear-time complexity in liquid-cooled ICs. Other researchers [16] tackle the long simulation time problem by using the GPU-accelerated generalized minimum residual (GMRES) method and provide one or two orders of magnitude speedup compared to the single-threaded CPU-GMRES method.

Liquid cooling provides much higher heat removal efficiency compared to air cooling, but also brings the new management challenges, such as large on-chip thermal gradients and pumping power to push the liquid through the channels. Prior work addresses some of those challenges through design-time and runtime optimization techniques. Coskun *et al.* [23] adjusted the liquid flow rate at runtime to save pump power. Their algorithm predicts the maximum temperature and adjusts the flow rate to the minimum value that meets the thermal limits. Sabry *et al.* [24] proposed a fuzzy controller to decide on the most efficient core voltage–frequency setting and flow rate at runtime. They also show that combining the fuzzy controller with flow-aware load balancing in 3-D systems provides significant reduction in thermal gradients. GreenCool [6] is a design time method to reduce thermal gradients by channel width modulation. GreenCool computes the optimal channel width profile that minimizes the pumping energy under the thermal gradient constraints. Qian *et al.* [25] proposed an efficient channel clustering and flow rate allocation algorithm, which customizes the cooling effort based on the demands of the computing elements. Saving pump power by nonuniformly distributing the microchannels according to the chip power profile is also possible [26], [27]. One such technique co-optimizes the number, locations, dimension, and flow rate of the microchannels to minimize pumping power [26]. Another similar approach is to design a nonuniform liquid cooling layer such that microchannels are denser above hot spots [27]. Their work also utilizes a manifold microchannel sink, with a manifold layer above the microchannels with multiple inlets/outlets, to reduce the pressure drop [27].

B. Thermoelectric Cooling

TECs have gained attraction due to their ability to remove the hot spots with high power densities. Modeling of the TEC thermal behavior is widely studied in the research community [8], [10], [17], [18]. Compact thermal models represent the heat absorbed and rejected on either side of the TEC elements using current sources entering and leaving the thermal nodes [8], [17], [18]. Chowdhury *et al.* [18] compared their numerical compact model against measurements on a test device and show the impact of nonidealities on the cooling potential of the TECs. Others perform comparison of their 1-D analytic TEC model against 3-D numerical simulations in ANSYS [8].

Other work focuses on demonstrating the benefits of TECs in hot spot mitigation using simulations and through measurements on hardware testbeds [7]–[11], [17], [18]. Superlattice-based thin film TECs made of Bi_2Te_3 as the bulk material are the state-of-the-art, owing to their high intrinsic figure-of-merit (ZT) [18]. They are silicon micro-fabrication compatible and can be directly integrated or fabricated on the back of a silicon chip [7], [18]. A group of work focuses on optimizing TEC device geometry and supply current to maximize coefficient of performance (COP) [8], [17], [28]. Yazawa *et al.* [8] focused on optimizing the TEC thickness and drive current without considering the microchannel flow rate, while this paper proposes co-optimization of the liquid flow rate and TEC current. As the focus of the two works are different, a direct comparison of the algorithms is not feasible.

Another body of work shows the integration of TECs on the back of a silicon test chip to cool hot spots with heat fluxes up to 1250 W/cm^2 [7], [18]. Sahu *et al.* [7] experimentally demonstrated the benefits of hybrid cooling with TECs and liquid microchannels on a testbed. Their work analyzes the impact of parameters like TEC size, heat flux, and ambient temperature on the resulting cooling performance through experiments. Chowdhury *et al.* [18] showed up to 9.6 °C reduction at 1250 W/cm^2 HSHF using a Bi_2Te_3-based, 3.5 mm $\times$ 3.5 mm TEC unit. While these approaches [7], [18] provide valuable analysis in showing the benefits of TECs in hot spot removal, our focus is to provide a hybrid cooling power optimization technique.

C. Hybrid Cooling Involving TECs

Hybrid cooling with TECs and liquid microchannels has been proposed as an energy-efficient solution for mitigating high density hot spots [7], [8], [28]. Sahu *et al.* [7] showed the thermal benefits and characterize the behavior of such hybrid cooling scheme on an experimental setup incorporating on-chip TEC units and a microchannel heat sink. Other work rely on compact models to demonstrate the cooling energy savings of a hybrid solid-state and microfluidic cooling system over solely using microfluidic cooling [8], [28]. They use the aforementioned compact models for TEC modeling, and represent the effect of the microchannel-based liquid cooling using a high effective heat transfer coefficient at the boundary. This is a simplified way of modeling hybrid cooling as it does not consider important aspects of liquid cooling, such as the rise of coolant temperature as it flows from the inlet to the outlet. Such aspects become critical when, for example, exploring the impact of hot spot locations on the resulting cooling power. Hot spots that are located closer to the outlet of the microchannels get hotter than the ones that are closer to the inlets, and failing to model this effect results in optimistic evaluation of systems. A compact thermal model which incorporates the behavior of TECs and microchannel liquid cooling together with sufficient detail and modularity is not currently available.

Hybrid designs incorporate two or more cooling solutions on the same platform. The first group of hybrid designs focus on TECs working together with liquid cooling. This hybrid combination is promising owing to the ability of TECs to remove localized hot spots and the ability of liquid cooling to remove background heat efficiently. Sahu *et al.* [7] experimentally explored the impact of design parameters on the cooling ability of a test vehicle, which combines a microchannel heat

sink with SiGe-based TECs. In their work, the authors vary the TEC sizes (70, 100, 120 μm side length), the location of the microchannel heat sink (on-chip/off-chip), ambient temperature, and the type of fluid as design parameters, and show a maximum temperature drop of 3 °C at 200 W/cm^2 heat flux and 85 °C ambient temperature. Yazawa *et al.* [8] showed 10× cooling power reduction for a microchannel and TEC-based hybrid cooling system compared to using microchannel cooling only. The benefits of a similar hybrid cooling scheme have also been demonstrated on a 3-D-stacked system through simulations [9]. The aforementioned works in general focus on demonstrating the benefits of a hybrid cooling design and explore the impact of the design parameters on the resulting cooling performance. None of them look into optimizing the operating point of a given hybrid-cooled system.

The second group of work combines TECs and fan cooling to maximize throughput under the thermal limits. Paterna and Reda find the optimum {TEC current, voltage–frequency} pair to distribute a given power budget between TECs and cores to maximize throughput for a fixed fan speed [10]. The follow-up of their work demonstrates the tradeoffs between TEC power, leakage power, and fan power on an experimental setup and adds fan speed as a parameter in the optimization scheme [11]. Their work targets low heat flux rates (∼20–28 W/cm^2) and does not focus on localized use of the TECs. The authors demonstrate that for a given total power cap, using TECs in cooperation with fans and DVFS techniques can provide 19% higher performance compared to using only fans and DVFS. Other work proposes leakage-aware control of TEC current to improve the COP of the TEC [29]. Dousti and Pedram [30] proposed an algorithm to decide on which TECs to turn on and off on a system with multiple cores to save cooling power. The aformentioned works [10], [11], [29], [30] focus on lower power densities, for which TECs combined with fanned heat sinks provide sufficient cooling. In this paper, we target much denser hot spots with heat fluxes reaching up to 1600 W/cm^2. For such systems, TECs with fans cannot maintain safe temperatures [18], thus, in this paper, we focus on TECs with liquid cooling.

D. Power Budgeting

Another group of work focuses on increasing the energy efficiency by optimally budgeting a given power between cooling and computing elements [31]–[33]. The goal of those works is to distribute the workload (i.e., the compute power) among many processing units with the same [32], [33] or heterogeneous microarchitectures [31]. They analyze parameters, such as the number and location of active cores, the voltage/frequency levels of the cores, or the type of the core to run the workloads on. The main difference of this paper is that we aim to minimize the power consumption of the cooling system by tuning its operating point for given hot spots. In other words, we optimize the cooling solutions, not the allocation of the workload. In this way, the aforementioned methods are orthogonal to this paper. One can apply a workload management algorithm together with a hybrid cooling system optimizer. Another significant difference is that we target systems with high-density hot spots (1000–1600 W/cm^2), while prior work focuses on much lower power densities.

E. Distinguishing Aspects of Our Work

This paper contributes to the research on hybrid cooling in two main areas: 1) compact thermal modeling and 2) optimization. In our recent work, we have developed, for the first time, a *compact* hybrid thermal model for the design and evaluation of systems using TECs and liquid microchannels with sufficient detail and modularity [21]. When modeling liquid microchannels in a hybrid cooling environment, our model avoids simplifying assumptions such as representing the liquid cooling layer solely with a heat transfer coefficient. The proposed model is applicable to a wide range of platforms and applications. It is modular such that the users can plug-in the cooling elements (TECs, microchannels, or both) with desired size, properties, and granularity. Compared to COMSOL, our compact model provides sufficient accuracy while saving considerable amount of time and processing resources.

In this paper, we optimize the cooling power of a hybrid cooling system for the first time. We also propose, for the first time, a runtime optimization policy to jointly determine the liquid flow rate and TEC current to minimize cooling power for a given temperature constraint. Our policy can adapt to changes in the hot spot heat density at runtime, thus, provides energy-efficient operation in the presence of dynamic workloads.

III. Modeling Methodology

This section describes the modeling and simulation framework that enables design and co-optimization of the hybrid cooling methods. We start by providing a background on compact thermal modeling approach and the temperature simulator which we use as a basis to implement our proposed model. We then give details of the proposed compact hybrid cooling model which can jointly simulate TECs and liquid cooling. We also integrate a detailed cooling power model into our framework. We finally discuss the steps we follow for validating the accuracy of the proposed model.

A. Compact Thermal Modeling Approach and HotSpot Simulator

We propose a compact model to characterize the steady-state temperature behavior of hybrid cooling systems with liquid microchannels and TECs. For modeling hybrid cooling, we adopt a compact thermal modeling approach. In this approach, the processor temperature is represented using an *RC* network, where *R* and *C* correspond to thermal resistance and thermal capacitance, respectively. Solving this network using a differential equation solver gives the temperatures of each node in the network.

We implement our proposed thermal model in HotSpot-6.0 [22] temperature simulator. HotSpot models vertical and lateral heat flow on the chip stack, and it also includes a processor package model for the heat spreader and a heat sink with fan. HotSpot allows the user to model multiple stacked layers with desired properties, such as processing layers and thermal interface material (TIM) layers. Inputs to the simulator are: 1) the physical geometry of the chip stack; 2) the floorplan of each layer as a collection of rectangular blocks; 3) the thermal properties of the materials on each layer; and 4) the power dissipation of the blocks. The *grid model* in HotSpot provides finer granularity simulation by dividing the layers into smaller grid cells and computing the temperature for each grid cell. Recently, Meng *et al.* [34] have added the functionality to model *heterogeneity within each layer* in HotSpot such that the user can assign different thermal properties to individual blocks residing on the same layer (e.g., copper TSVs going through a TIM layer). This feature is included in the most

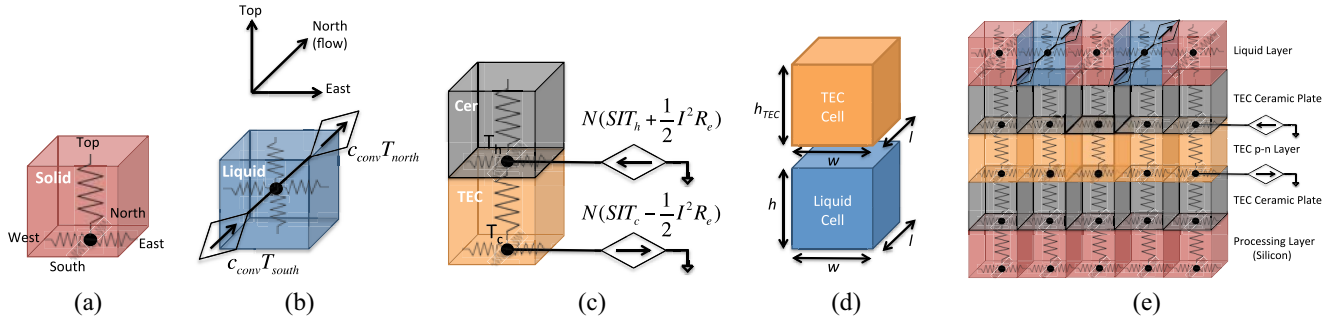

Fig. 2. (a) Solid grid cell, (b) liquid grid cell, (c) TEC grid cell, (d) dimensions of the grid cells, and (e) connectivity of the grid cells building a chip stack. Current sources are shown only for the rightmost TEC and ceramic cells for clarity.

recent release (i.e., HotSpot-6.0), and we make use of this feature when implementing our proposed hybrid model.

In order to model the heat removing capability of liquid microchannels and TECs in a thermal simulator, we first define liquid microchannels or TEC units as blocks on the floorplan, tag them as *microchannel* or *TEC* blocks, and define their corresponding thermal properties. Similarly, the grid cells that are residing inside these blocks are also tagged as *microchannel cells* or *TEC cells*. While constructing the thermal resistor (R) network, we incorporate the additional heat flow terms for the tagged grid cells according to the governing equations of the corresponding cooling methods. In the next section, we discuss how we model the behavior of the TECs and liquid microchannels in more detail.

Regarding the processor package, our model allows the user to include a heat spreader and fan cooled heat sink with desired properties, such as the thickness, material conductivity, and convection parameters based on their target platform. The original HotSpot thermal model forces the user to simulate a system with heat spreader and heat sink and does not include an option to remove them. However, since we focus on hybrid designs involving microchannel liquid cooling and TECs, microchannel layer acts as the heat sink in our case. Thus, in our proposed model, we include the option to add or remove heat spreader and fanned heat sink layers.

B. TEC Thermal Model

A TEC operates based on the Peltier effect such that when current passes through the device, heat is absorbed from one side (cold side) and rejected to the other side (hot side), creating a thermal gradient across the two sides [8], [18]. The amount of heat removed by the TEC depends on the Seebeck coefficient (S), applied current (I), electrical resistivity (ρ_{tec}), thermal conductivity (k_{tec}) of the TEC device, and the temperatures of the hot (T_h) and cold (T_c) sides. Superlattice-based thin film TECs made of Bi_2Te_3 have high figure-of-merit (ZT). They are silicon micro-fabrication compatible and can be directly integrated or fabricated on the back of a silicon chip [7], [18]. Existing on-chip TEC devices are composed of ultrathin (5–10 um) Bi_2Te_3-based p-n thermocouples sandwiched between copper mini-headers and are covered with ceramic plates at the outmost surfaces to provide insulation [18].

There are three main contributors to heat flow within a TEC unit: 1) the Peltier term which accounts for the heat absorbed/rejected on the cold/hot sides; 2) the conductive heat flow term; and 3) Joule heating term that represents the resistive heat generated by passing current through the TEC. Mathematical representation of these terms are

$$Q_c = N\left(S \cdot I \cdot T_c - \frac{T_h - T_c}{R_t} - \frac{1}{2}I^2R_e\right) \tag{1}$$

$$Q_h = N\left(S \cdot I \cdot T_h - \frac{T_h - T_c}{R_t} + \frac{1}{2}I^2R_e\right) \tag{2}$$

where Q_c and Q_h stand for the heat absorbed and rejected on the cold and hot sides, respectively. T_c and T_h are the cold and hot side temperatures. N is the number of p-n couples placed in area A. R_t=$h_{\text{tec}}/k_{\text{tec}}A$ is the thermal resistance and R_e=$\rho_{\text{tec}}h_{\text{tec}}/A$ is the electrical resistance of a TEC unit of thickness h_{tec} and area A.

We implement this model in HotSpot in the following way. We use the grid model in HotSpot, in which, each layer on the processor stack is divided into smaller grid cells representing a thermal node in the thermal R network. We add functionality to define a block on the floorplan as a TEC unit. We then assign TEC thermal properties only to the grid cells corresponding to these TEC units. For this purpose, we use the heterogeneous 3-D modeling feature of HotSpot as mentioned earlier. HotSpot by default accounts for the conductive heat flow [term 2)] for solid cells as shown in Fig. 2(a). In order to represent the Peltier term and Joule heating term on the cold and hot side of the TEC units described in (1) and (2), we define current sources entering and leaving the TEC cells as illustrated in Fig. 2(c). In the figure, bottom surface of the TEC cell corresponds to the cold side temperature, while the bottom surface of the cell in the upper adjacent layer (i.e., the ceramic plate) corresponds to the hot side temperature.

C. Liquid Cooling Thermal Model

We base our liquid model on the four resistor model (4RM) used in 3D-ICE [13]. In the 4RM model, the discretization of the thermal grids is done such that the entire cross section of a microchannel forms a liquid grid cell. There are two main contributors to heat flow regarding a liquid grid cell: 1) convective heat transfer from the walls of the channel to the liquid and 2) convective heat transfer in the direction of the liquid flow into and out of the current liquid cell. Fig. 2(b) illustrates a liquid grid, where the term: 1) is represented by resistive elements in four directions and the term 2) is represented by using current sources in the direction of the flow (from South to North). The numerical values of the resistances are given as follows [13]:

$$R_{\text{top,bottom}} = \frac{1}{h_{f,\text{vertical}} \cdot w \cdot l} \tag{3}$$

$$R_{\text{east,west}} = \frac{1}{h_{f,\text{side}} \cdot h \cdot l} \tag{4}$$

where $h_{f,\text{vertical}}$ and $h_{f,\text{side}}$ are the heat transfer coefficients for microchannel forced convection; w, l, and h are the width, length, and height of the microchannel cell, respectively [see Fig. 2(d) for the cell dimensions]. As also stated in 3D-ICE work [13], $h_{f,\text{vertical}}$ and $h_{f,\text{side}}$ (i.e., the vertical and side heat transfer coefficients) can be obtained from empirical correlations or numerical presimulation for a given system. For computing the heat transfer coefficients, prior work provides the following formulas assuming imposed axial heat flux and radial isothermal conditions:

$$h_{f,\text{vertical}} = h_{f,\text{side}} = \frac{k_{\text{coolant}} \cdot Nu}{d_h} \tag{5}$$

$$Nu = 8.235 \cdot (1 - 2.0421\text{AR} + 3.0853\text{AR}^2 - 2.4765\text{AR}^3 + 1.0578\text{AR}^4 - 0.1861\text{AR}^5). \tag{6}$$

In these formulas, k_{coolant} is the thermal conductivity of the coolant and d_h=($[2h \cdot w]/[h + w]$) is the hydraulic diameter of the channel. Nusselt number (Nu) was derived in prior work [35] as a function of channel aspect ratio ($\text{AR} = \min\{h/w, w/h\}$). As (5) and (6) may differ under different system assumptions, the original 3D-ICE simulator defines $h_{f,\text{vertical}}$ and $h_{f,\text{side}}$ as input parameters specified by the user.

Next, the values of the convective terms in the flow direction (i.e., the current sources) are computed as follows:

$$I_{\text{in}} = c_{\text{conv}} \cdot T_{\text{south}} \tag{7}$$

$$I_{\text{out}} = c_{\text{conv}} \cdot T_{\text{north}} \tag{8}$$

$$c_{\text{conv}} = C_v \cdot u_{\text{avg},y} \cdot \Delta A_y \tag{9}$$

where I_{in} and I_{out} denote the convective heat flow into and out of the cell, respectively. T_{south} and T_{north} are the interface temperatures at the south and north surfaces of the cell. C_v is the specific heat capacity of the coolant, $u_{\text{avg},y}$ is the average coolant velocity, and $\Delta A_y = w \cdot h$. The surface temperatures are approximated as the average of the cell temperatures which share that interface. We assume that for the southmost cell, $T_{\text{south}} = T_{\text{inlet}}$ (i.e., temperature of the coolant at the microchannel inlet) and for the northmost cell $T_{\text{north}} = T_{\text{cell}}$.

Note that by default, HotSpot places the virtual temperature node at the bottom surface of the grid cell in the vertical direction as illustrated in Fig. 2(a). This convention is useful for modeling the TEC cells as the thermal effect is observed at the bottom and top surface of the TEC device. However, for liquid cells, we need to place the virtual node in the middle of the cell to be able to include the heat flow from the top/bottom walls in an accurate manner. Doing otherwise results in underestimation of the chip temperature by up to 20 °C for liquid-cooled systems, according to our analysis. Thus, we construct the thermal resistance network in our model such that for liquid cells, the node is placed in the middle; while for all other cells including TECs, the node is placed at the bottom surface. This way of constructing the thermal resistance network is one of our novel contributions. In Fig. 2(e), we demonstrate how the grid cells of each type are connected in the chip stack building a thermal R network, for a single row of cells.

D. Cooling Power Model

Cooling power for an individual liquid-cooled system mainly includes the pump power consumed to push the fluid through the channel[1] and is calculated as follows [6]:

$$P_{\text{pump}} = \frac{\Delta P \cdot V}{\eta} \tag{10}$$

$$\Delta P = \frac{2 \cdot f_r \cdot \rho_{\text{coolant}} \cdot u_{\text{avg},y}^2 \cdot L}{d_h} \tag{11}$$

where ΔP is the pressure drop across the channel (Pa), V is the total volumetric flow rate (m^3/s), and η is the pump efficiency (generally between 10% and 40%). f_r, ρ_{coolant}, and L are the friction factor, coolant density, and the length of the channel, respectively. Friction factor was driven in prior work for fully developed conditions as follows:

$$f_r \cdot \text{Re} = 24 \cdot (1 - 1.3553\text{AR} + 1.9467\text{AR}^2 - 1.7012\text{AR}^3 + 0.9564\text{AR}^4 - 0.2537^5) \tag{12}$$

$$\text{Re} = \frac{u_{\text{avg},y} \cdot d_h \cdot \rho_{\text{coolant}}}{\mu} \tag{13}$$

where Re is Reynolds number given for laminar flow conditions (i.e., $\text{Re} \leq 2300$) and μ is the dynamic viscosity of the coolant. Table I lists all constant parameters we use.

The power consumed by the TECs is computed as follows:

$$P_{\text{tec}} = Q_h - Q_c = N(S \cdot I \cdot (T_h - T_c) + I^2 R_e). \tag{14}$$

We account for both the pump and TEC power in our experiments, and integrate a power computation module in our simulation framework. While computing the TEC power consumption, we apply (14) to each TEC cell considering their individual T_h and T_c obtained from our thermal model.

TABLE I
PARAMETERS WE USED FOR VALIDATING THE LIQUID MICROCHANNEL AND TEC MODELS IN COMSOL

Microchannel height	h	$100\mu m$
Microchannel width	w	$50\mu m$
Grid cell width & length	$w = l$	$50\mu m$
Microchannel length	L	$10mm$
Coolant thermal conductivity	$k_{coolant}$	$0.6069W/mK$
Coolant specific heat	C_v	$4181J/kgK$
Coolant inlet temperature	T_{inlet}	$27^\circ C$
Coolant density	$\rho_{coolant}$	$998kg/m^3$
Coolant viscosity	μ	$8.89 \times 10^{-4} Pa.s$
Average coolant velocity	u_{avg}	$\leq 3m/s$
Pump efficiency	η	25%
TEC width & length	$w_{tec} = l_{tec}$	$3.5mm$
Seebeck coefficient	S	$301\mu V/K$
Thermocouple thickness	h_{tec}	$8\mu m$
Copper mini-header thickness	h_{Cu}	$2\mu m$
Ceramic plate thickness	h_{Cer}	$44\mu m$
TEC electrical resistivity	ρ_{tec}	$1.08 \times 10^{-5} Ohm.m$
TEC thermal conductivity	k_{tec}	$1.2W/mK$
Copper thermal conductivity	k_{Cu}	$400W/mK$
Ceramic thermal conductivity	k_{Cer}	$175W/mK$
Silicon thermal conductivity	k_{Si}	$130W/mK$
Ambient temperature	T_{amb}	$45^\circ C$

E. Validation Using COMSOL

In order to validate the accuracy of the proposed model, we compare the temperatures obtained from simulations using the model against the ones reported by COMSOL. For the liquid cooling model, in addition to COMSOL, we also compare results against 3D-ICE [13], a compact simulator. This section provides the details of how we set up the models in COMSOL.

[1]In a data center setting, there is also the external chiller power that is impacted by cumulative characteristics of a number of systems (e.g., number of servers, total liquid flow rate, temperature of the exhaust liquid etc.). We focus on a single liquid-cooled system in this paper.

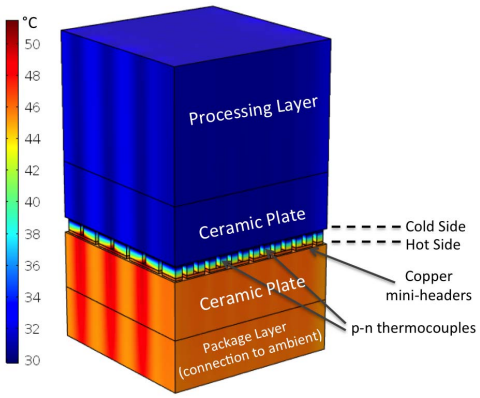

Fig. 3. TEC device that we modeled in COMSOL. Example temperature distribution corresponds to I = 4A.

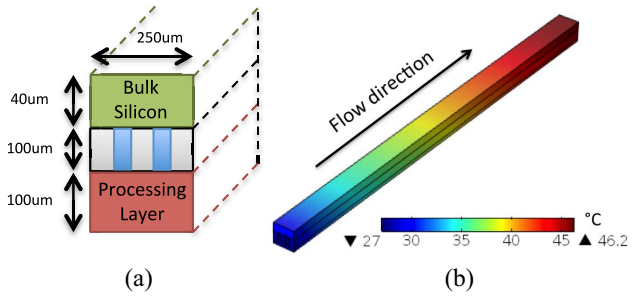

Fig. 4. (a) Front view of the thin slice of chip stack we modeled for liquid cooling. (b) View of the chip stack as we modeled in COMSOL.

TEC Model Validation Setup: In order to validate our TEC model, we compare its temperatures against the ones obtained from COMSOL simulations. For this purpose, we first select a prototype TEC device that has been fabricated on the back of a silicon chip and has been characterized in prior work [18]. The TEC device is superlattice-based thin film TEC made of Bi_2Te_3 as the bulk material. It is composed of an array of 7 × 7 p-n thermocouples and has a total size of 3.5 mm × 3.5 mm. Thermocouples are sandwiched between copper mini-headers and the top and bottom surface of the device is covered by ceramic plates to provide electrical insulation. Legs of the p-n thermocouples are ultrathin (8 μm) and the total thickness of the TEC device including the ceramic plates is 100 μm. We create a COMSOL model of this TEC device as illustrated in Fig. 3. Detailed parameters of the TEC are given in Table I. Note that for the temperature dependent parameters such as S, ρ_{tec}, and k_{tec}, we assume constant values at steady-state temperature as reported in prior work [18].

Next, we model the processing layer using a 100 μm-thick silicon layer at the cold side of the TEC, and assign a heat flux value (i.e., power dissipated per unit area) to it to represent the generated heat. As TECs pump heat from the cold side to the hot side, an additional cooling mechanism is usually needed on the hot side of the TEC to avoid overheating and provide proper operation. Thus, at the hot side of the TEC, we define another layer, which represents the chip package and an additional cooling mechanism (e.g., heat sink with fans, cold plates) that removes the heat pumped by the TEC. We assume silicon properties for this layer, set its thickness as 40 μm, and assign a heat transfer coefficient (htc) at the surface to the ambient to represent the additional cooling mechanism. Htc corresponds to the cooling capability of the additional cooling method, where a higher htc value represents more effective cooling. We modify HotSpot's package model so as to define a similar connection to ambient using the htc parameter.

Liquid Cooling Model Validation Setup: For validation of our proposed model in COMSOL, we first create a chip stack with liquid microchannels. Fig. 4(a) illustrates the cross section of the chip stack, where the liquid microchannel layer is placed on top of the processor layer, and an additional bulk silicon layer (with 40 μm thickness) is placed on the top to provide closure to the microchannels. We simulate a thin slice of this chip stack as in prior work [13] to reduce the problem size in COMSOL [See Fig. 4(b)]. The width and length of the slice are 250 μm and 5 mm, respectively. We set the microchannel width as w = 50 μm (also equal to the channel wall width) and channel thickness as h = 100 μm. With these microchannel parameters, the simulated slice includes two microchannels interleaved between three channel walls made of silicon. At the top surface of the bulk silicon layer, we assign a very small heat transfer coefficient (i.e., htc = 0.01 W/m^{2}K) to represent minimal convection to air. We assume water as the coolant and use the coolant properties given in Table I.[2]

We model the same chip stack in 3D-ICE simulator for the second set of comparisons. As the computation of $h_{f,\text{vertical}}$ and $h_{f,\text{side}}$ coefficients significantly differ in COMSOL and 3D-ICE, we first experimentally estimate the coefficients from COMSOL simulations and then use them as inputs to the proposed model and 3D-ICE simulator. This way, we can carry out a consistent comparison of the three models. We extract the coefficients from COMSOL as follows: to find $h_{f,\text{side}}$, we select the surface of a side wall facing a microchannel and record the surface average of the total normal heat flux value, $h_{\text{avg,normal}}$ (which is equal to *ht.ntflux* in COMSOL). We then record the surface average of the side wall temperature (T_{wall}), and the volume average of the liquid temperature (T_{liquid}). Finally, we compute $h_{f,\text{side}}$ as follows:

$$h_{f,\text{side}} = \frac{h_{\text{avg,normal}}}{(T_{\text{wall}} - T_{\text{liquid}})}. \tag{15}$$

We carry out similar computation for $h_{f,\text{vertical}}$ using the top and bottom walls. We repeat the same steps for the flow velocities that we experiment with and assign the average computed value to the heat transfer coefficients. For our system, we determine that $h_{f,\text{side}} \approx h_{f,\text{vertical}} = 1.05 \times 10^5$ W/m^{2}K. We use these values as inputs to the proposed model and to the 3D-ICE simulator. Our model is orders of magnitude faster than multiphysics tools (See Section V-A for details), which enables us to use it in numerical optimization methods to identify the best design practices as described in the next section.

Hybrid Cooling Model Validation Setup: We also validate the hybrid model including TECs and liquid channels in COMSOL. We create a chip stack with size 4.5 mm × 5 mm and place the TEC in the center. Fig. 5 illustrates the COMSOL hybrid cooling setup. From top to bottom, there are the following components: 1) processing layer; 2) TIM layer where the TEC is placed in the center; 3) liquid layer; and 4) bulk silicon enclosure with connection to ambient. Properties of the TEC and liquid microchannels are the same as in TEC and liquid model validation setups. A chip width of 4.5 mm corresponds to 44 microchannels and 46 walls. In this model,

[2]We assume a data center setting where the hot liquid leaves the outlets and goes through a heat exchanger to be cooled down to 27 °C before reentering the inlets.

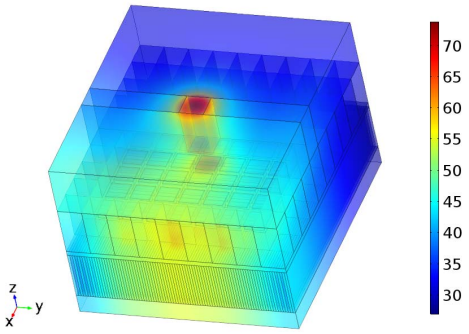

Fig. 5. Diagram of the hybrid model in COMSOL. Example temperature distribution corresponds to $I = 4$ A and $u_{\text{avg}} = 1$ m/s. Image is scaled in the z-axis for visibility.

we simulate the heat removal on the hot side of the TECs by modeling the liquid flow within the microchannels. Thus, in this model we *do not* make any htc assumptions at the hot side of the TEC. At the very bottom of the stack, bulk silicon layer enclosure represents minimal convection to air similar to the liquid model.

IV. LoCool: Localized Cooling Optimization

The *goal* of our algorithm is to find the {*coolant flow velocity, TEC current*} pair that minimizes the total cooling power while meeting temperature and cooling technology constraints. A formal definition of the optimization problem is as follows:

$$\begin{aligned} \text{minimize} \quad & P_{\text{pump}}(u) + P_{\text{tec}}(I) = \alpha \cdot u^2 + \beta \cdot I^2 \\ \text{subject to} \quad & T(u, I) < T_{\max} \\ & u_{\text{avg},y} \leq u_{\max} \\ & 0 \leq I \leq I_{\max} \end{aligned} \tag{16}$$

where α and β are the constants determined by the channel geometry and TEC properties. We compute α= 0.4954 and β= 0.057 according to our system. $T_{\max}$ is the maximum temperature constraint, while u and $u_{\max}$ are the average and maximum allowed coolant velocities, respectively. Maximum applied pressure drop recommended by manufacturers determines $u_{\max}$. We use 3.3 bar for maximum pressure drop, which corresponds to $u_{\max} = 2.6$ m/s for our geometry. We use $I_{\max} = 7$ A as the maximum TEC current constraint [18].

LoCool optimizer is composed of design-time and runtime optimization modules. We next describe the optimization flow for each module in detail.

A. Design-Time Optimization Algorithm

The goal of the design-time optimizer is to solve the problem defined by the system of (16) for a given static processor power dissipation map. LoCool design-time optimization flow is illustrated in Fig. 6. Given a power density map with several hot spot areas, we first place the TEC units above each hot spot. We assume power density maps where the HSHF is much higher than the background heat flux (BGHF) to model future high-performance systems as suggested in prior work [4] (up to 40× difference has been reported). Any block with over 10× heat flux compared to the background is recognized as a hot spot. We use an iterative approach, where we call our hybrid cooling thermal model described in Section III at every iteration to check whether the temperature constraint is met. LoCool design-time optimization is composed of two main phases, where Phase I is the *descending*

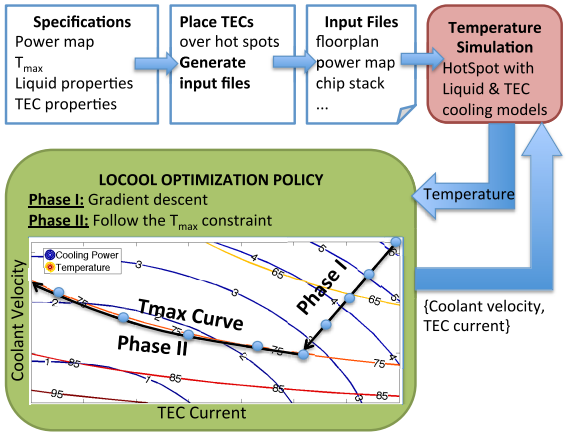

Fig. 6. LoCool optimization flow.

Pseudocode 1 Phase I: Gradient Descent

Inputs: T_{max}, u_{max}, I_{min}, I_{max}, α, β
Initialize: $u \leftarrow u_{max}$, $I \leftarrow I_{max}$, $i \leftarrow 0$
1: $f(u, I)$=$\alpha \cdot u^2 + \beta \cdot I^2$
2: **while** True **do**
3: $u(i+1)$=$u(i) - \gamma_u \frac{\partial f(u(i),I(i))}{\partial u}$
4: $I(i+1)$=$I(i) - \gamma_I \frac{\partial f(u(i),I(i))}{\partial I}$
5: $u(i) \leftarrow u(i+1)$, $I(i) \leftarrow I(i+1)$
6: $T \leftarrow OurThermalModel(u(i), I(i))$
7: $i \leftarrow i + 1$
8: **if** $|T - T_{max}| < 1$ °C **then**
9: **break**
10: **end if**
11: **end while**

phase and Phase II is *following the temperature constraint*. Phase I starts from the highest cooling power setting and descends to lower cooling power settings using a *gradient descent* algorithm. *Gradient descent* is a first-order iterative optimization algorithm, where one takes steps proportional to the negative gradient of the function to be minimized. At each iteration, the algorithm decreases each variable (i.e., flow velocity and TEC current) by a fraction of the gradient with respect to that variable. In this way, during the descent, cooling power decreases, while temperature increases. Phase I ends when T is in the close vicinity of $T_{\max}$, i.e., $|T - T_{\max}| < 1$ °C. Using the *gradient descent* algorithm, we can approach the maximum temperature constraint curve fast by following a steep path as shown in Fig. 6. The fast descent property of the *gradient descent* algorithm is very useful in steering through the large solution space (which involves all combinations of the possible $\{u, I\}$ pairs) in an efficient manner. Pseudocode 1 summarizes the steps for Phase I.

While formulating the goal function in Phase I of our LoCool hybrid optimizer, we consider the quadratic term of the TEC power as it dominates the cooling power during the gradient descent phase and simplifies optimization. However, at each iteration of the algorithm, our thermal model with integrated power model reports the cooling power using (14), based on the T_h and T_c temperature values that results from the $\{u, I\}$ setting. We use these values that we get from our thermal model when reporting the cooling power throughout this paper.

In Phase II of LoCool design-time optimization, we follow the temperature constraint curve in the direction of decreasing

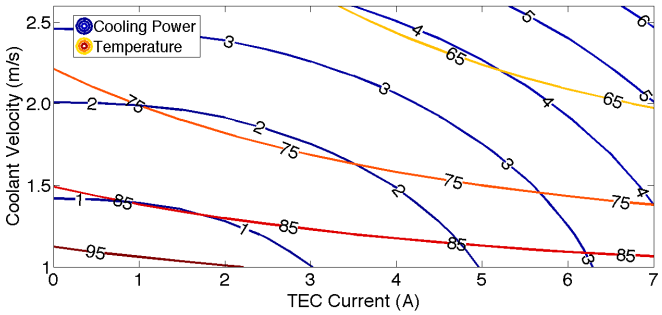

Fig. 7. Contour plot showing equal cooling power and temperature lines.

cooling power. For this purpose, we leverage our observations on how the temperature and cooling power curves change based on the $\{u, I\}$ pairs. Fig. 7 is a contour plot showing equal cooling power and temperature curves for a range of $\{u, I\}$ pairs. Phase II starts on a point that is close to the $T_{\max}$ curve. Due to the shape of the temperature curves, in order to minimize power, one needs to either: 1) go up and left, which implies decreasing TEC power and increasing pump power or 2) down and right, which implies increasing TEC power and decreasing pump power, depending on where we are located on the curve. For example, if Phase I ended on the bottom right point of the curve [e.g., $\{u, I\} = \{1.5 \text{ m/s}, 6.0 \text{ A}\}$] and $T_{\max} = 75$ °C, then we need to go up and left (decrease current and increase velocity) to minimize power consumption. Similarly, if we are on the top left part [e.g., $\{u, I\} = \{2.2 \text{ m/s}, 0.5 \text{ A}\}$], we need to go down and right. To decide on which direction we should go, we compute $D = \nabla_{\vec{d}} f(u, I)$, which is the derivative of $f(u, I)$ in the direction of $\vec{d} = 0.1\vec{i} - 0.5\vec{j}$, where $\vec{i}$ and $\vec{j}$ are the unit vectors in the Cartesian coordinates. $\vec{d}$ represents an up and left motion and D changes from a negative value to an increasing positive value along a temperature curve. Once we decide on one of the two directions, we follow the direction by alternating between vertical and horizontal moves. We keep updating the minimum cooling power that meets the thermal constraints along the path. Phase II ends when we reach a boundary of valid $\{u, I\}$ pairs.

We evaluate the optimality of our algorithm by comparing its results against exhaustive search of all possible $\{u, I\}$ pairs. We tested 12 examples and LoCool found the optimum setting for all cases in less than 23 iterations, where each iteration corresponds to a few minutes of simulation time.

B. Runtime Optimization Algorithm

The goal of the LoCool runtime optimization algorithm is to adapt to the changes in heat flux levels at runtime for optimum operation. For this purpose, we adopt an offline analysis-based approach where we generate a lookup table of the optimum $\{u, I\}$ pairs using the design-time algorithm for a range of HSHF levels. At runtime, our algorithm polls this lookup table to select the optimum settings for the current HSHF value.

Fig. 8 shows an example lookup table that corresponds to a temperature constraint of $T_{\max} = 80$ °C. In the figure, the x- and y-axes represent the TEC current and coolant flow velocity values, while the color bar represents different HSHF levels. We run the design-time optimization algorithm described in Section IV-A for a range of HSHF levels and record the optimum settings in order to generate this table for a given temperature constraint. To detect the current heat flux level at runtime, we propose to utilize on-chip power estimators that are either integrated by the manufacturers on the processor board [36] or can be implemented as in prior work [37]–[40].

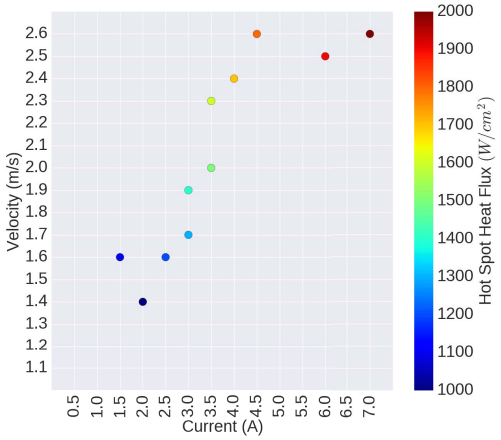

Fig. 8. Optimum $\{u, I\}$ pairs for a range of HSHF levels with $T_{\max}$= 80 °C constraint.

The overhead of the runtime algorithm is mainly composed of offline generation of the lookup table by running LoCool design-time optimization algorithm. Once this table is generated, there will be no design space exploration required at runtime. The table will be stored in memory and polled at runtime. The range and granularity of the HSHF levels determine the size of this table. In Fig. 8, granularity of HSHF changes in steps of 100 W/cm^2, leading to 11 entries for each $T_{\max} = 80$ °C constraint. One can increase the number of entries depending on the desired HSHF granularity. At runtime, polling a table of a few tens of entries incurs minimal overhead regarding time and memory.

V. Experimental Results

In this section, we first validate our hybrid thermal model by comparing its results against COMSOL (for the TEC and liquid models) and 3D-ICE (for the liquid model). We then continue with an evaluation of our LoCool optimization algorithm for the design-time and runtime optimization modes. As part of our evaluation, we compare the cooling power consumption of hybrid cooling designs optimized with LoCool against the designs using liquid cooling only.

A. Hybrid Cooling Model Validation Results

Using the COMSOL setup and parameters described in Section III-E, we run experiments applying a range of heat fluxes and cooling bias levels (i.e., bias current for TECs and flow velocity for liquid cooling). We compare the processor temperatures resulting from simulations using COMSOL and our model. Throughout this paper, we will refer to the results corresponding to our proposed hybrid model as *proposed*.

We start with the comparison results for the TEC model. Fig. 9 compares the average temperature of the processor layer over a range of TEC bias currents. For this experiment,[3]

[3]We use htc to represent the heat sink above the TEC **only** during validation of the TEC model. When validating the hybrid model, we do not use htc, but model each of the microchannels. For the rest of this paper, when evaluating the LoCool optimization algorithm, we use our compact hybrid model, which includes the impact of heat generated on the hot side of the TEC and the transfer of this heat to the liquid microchannels.

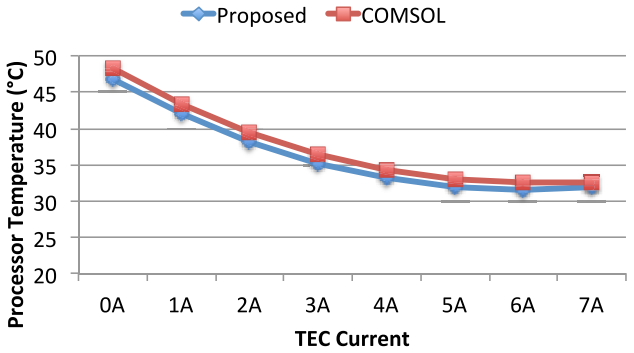

Fig. 9. Comparison of processor temperature over TEC current for COMSOL and the proposed model. htc $= 10^6$ W/m^{2}K and $q = 20$ W/cm^2.

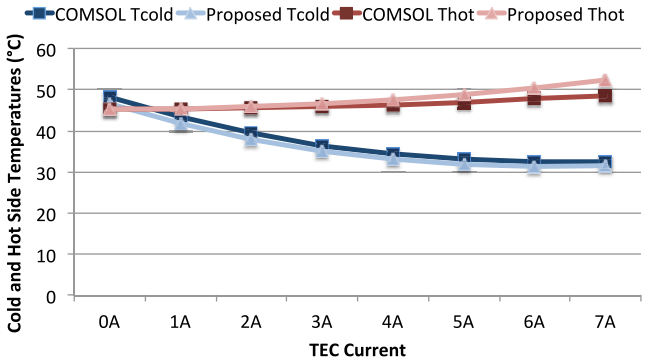

Fig. 10. Comparison of the cold and hot side temperatures over TEC current for COMSOL and the proposed model. htc $= 10^6$ W/m^{2}K and $q = 20$ W/cm^2.

htc $= 10^6$ W/m^{2}K and $q = 20$ W/cm^2. Our proposed TEC model closely follows the temperature results obtained from COMSOL with an error less than 1.5 °C. As expected, the processor temperature starts to reduce as the TEC bias current increases. At some point (i.e., around 6 A), impact of Joule heating becomes dominant, resulting in a slight increase in the processor temperature. In Fig. 10, we report the cold and hot side temperatures of the TEC for the same simulation. At 0 A bias current, $T_{\text{cold}} > T_{\text{hot}}$ due to the additional resistance presented by the TEC device. At around 0.5 A, amount of heat that is pumped by the TEC overcomes its own resistance and $\Delta T = (T_{\text{hot}} - T_{\text{cold}})$ becomes positive and starts to increase. We carry out similar analysis for other q values ranging from 20 to 50 W/cm^2 and observe that the absolute maximum error is 3.57 °C. We also report 2.07 °C of average and 2.25 °C of RMS error.

To validate the accuracy of the liquid cooling model, we compare its temperature results against the ones obtained from COMSOL and 3D-ICE simulations. We run steady-state simulations for a range of q values of 12.5, 25, 50, and 100 W/cm^2 as well as for different flow velocities, $u_{\text{avg}} = 0.5, 1.0, 1.5, 2.0$ m/s, and record the maximum temperature of the processing layer for the proposed model, COMSOL, and 3D-ICE. Fig. 11 shows the maximum processor temperatures obtained from COMSOL, 3D-ICE, and our proposed model for all u_{avg} combinations where $q = 100$ W/cm^2. Among all experiments, compared to COMSOL simulations, our proposed model provides maximum, average, and RMS error of 2.46 °C (corresponds to 2.8%), 0.36 °C, and 0.72 °C, respectively. In comparison to 3D-ICE, the error of the proposed model is less than 0.04 °C.

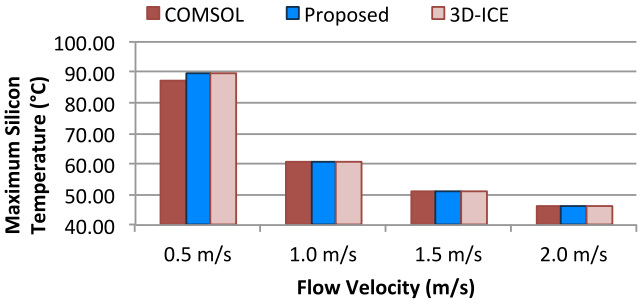

Fig. 11. Maximum processor temperature comparison for COMSOL, 3D-ICE, and the proposed model for $q = 100$ W/cm^2.

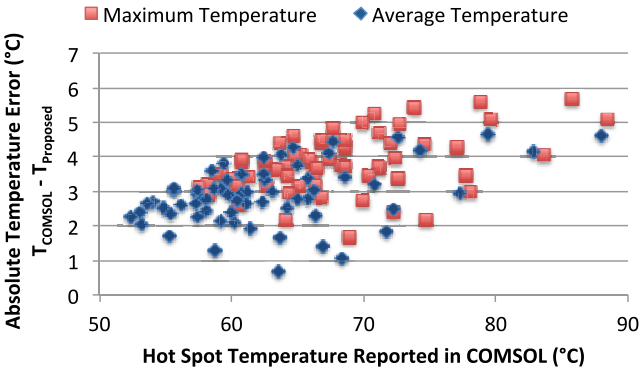

Fig. 12. Maximum and average hot spot temperature comparison for COMSOL and proposed hybrid model for all settings.

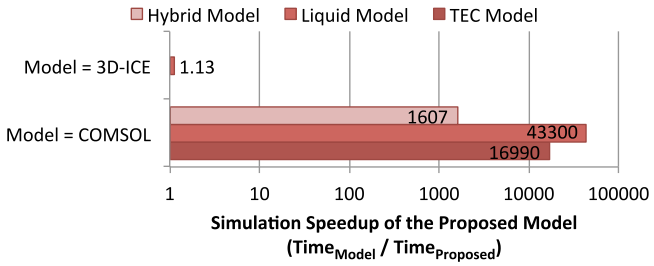

Fig. 13. Comparison of the simulation speed across three simulators. As 3D-ICE does not have a TEC or a hybrid model, the bars are not shown.

Finally, we compare the accuracy of the complete compact hybrid model by comparing its results against COMSOL simulations. We construct the chip stack described in Section III-E in COMSOL. On the active silicon layer, we define two types of heat flux: 1) BGHF of BGHF $= 20$ W/cm^2 and 2) an HSHF of HSHF $= 1100, 1300$ W/cm^2. Hot spot is located in the center of the floorplan with a size of 500 μm $\times$ 500 μm. Fig. 12 compares the maximum and average temperature of the hot spot for COMSOL and our proposed model. In this scatter plot, x-axis reports the COMSOL temperatures and y-axis reports the absolute hot spot temperature error. Our model achieves a peak error of 5.7 °C and an average error of 2.9 °C for the hybrid model.

We finish model validation by comparing the solution speeds of the simulators against the proposed model for both TEC and liquid cooling. Fig. 13 demonstrates the average solution time ratio of the compared simulators over the proposed model. As indicated, the proposed compact modeling approach can save significant simulation time (providing up to four orders of magnitude speedup) with reasonable tradeoff

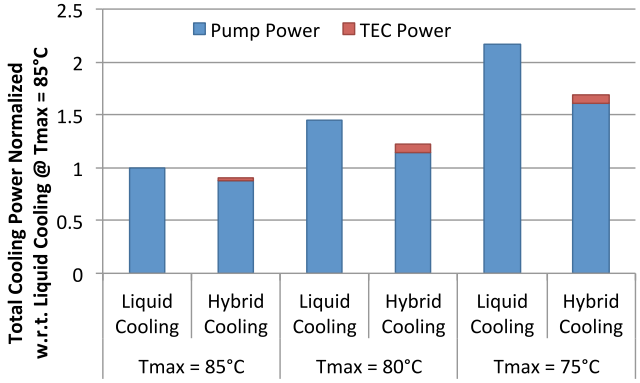

Fig. 14. Total cooling power comparison of liquid and hybrid cooling for HSHF = 1000 W/cm^2. Results are normalized to liquid cooling at $T_{\max}$ = 85 °C.

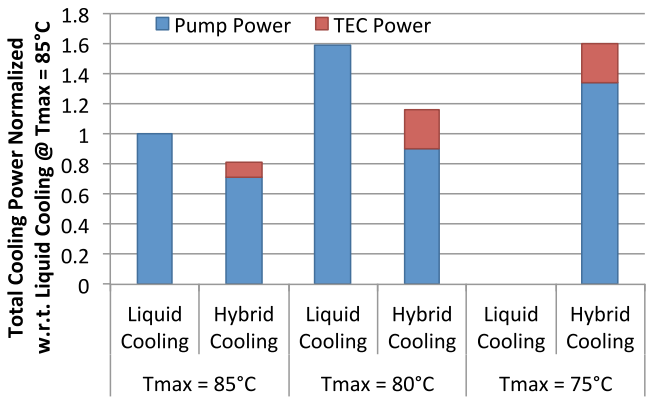

Fig. 15. Total cooling power comparison of liquid and hybrid cooling for HSHF = 1300 W/cm^2. Results are normalized to liquid cooling at $T_{\max}$ = 85 °C. Temperature constraint was not met for bars not shown.

in accuracy. This fast speed is key to leveraging our model toward design optimizations.

B. Design-Time Optimization Results

In this section, we evaluate the benefits of hybrid cooling designs that are optimized using LoCool in comparison to using liquid cooling only. We experiment with a variety of power maps and report the resulting total cooling power for the two cooling designs.

We assume a large chip with 20 mm × 20 mm size. The BGHF is set to 50 W/cm^2. We define hot spot blocks with 500 μm × 500 μm size and we vary their location and HSHF. We experiment with HSHF = [1000, 1300, 1500, 2000] W/cm^2, following examples from prior work [4], [18]. We adopt the TEC size of 3.5 mm × 3.5 mm from prior work [18].

We compare the minimum cooling power for the liquid cooled system against the hybrid cooling system for varying temperature constraints (i.e., $T_{\max}$ = 85 °C, 80 °C, 75 °C). For the hybrid cooling case, we report the results we obtain from LoCool algorithm.

Figs. 14 and 15 show a subset of the results for a single hot spot case with HSHF = 1000 and 1300 W/cm^2, respectively. The hot spot is located close to the outlet of the channels in this experiment. Figs. 14 and 15 indicate that an optimized hybrid cooling system saves significant cooling energy by focusing the cooling effort on the hot spot. For HSHF = 1000, hybrid cooling with LoCool saves cooling power by 9%, 16%, and 22% at $T_{\max}$ = 85 °C, 80 °C, 75 °C,[4] respectively. Intuitively, power saving increases for higher HSHF values. At HSHF = 1300, LoCool provides up to 28% cooling power savings. The simple explanation is that as the temperature constraint gets tighter and hot spots get denser, liquid cooling starts to pump coolant at a much higher rate just to cool the hot spots. On the other hand, hybrid cooling with the TECs focuses the cooling effort on the hot spot and meet the same temperature constraint at a lower flow rate, thus, providing a more gradual cooling power curve.

TABLE II
PERCENT OF THE $\{u, I\}$ PAIRS THAT MEET THE TEMPERATURE CONSTRAINT AND OUT OF THAT PERCENT, THE PORTION OF THEM WHICH PROVIDE LOWER COOLING POWER THAN LIQUID COOLING. N/A MEANS LIQUID COOLING DID NOT MEET THE TEMPERATURE CONSTRAINT

	A = % of all $\{u, I\}$ pairs where $T < T_{max}$			% of the $\{u, I\}$ pairs out of A where $P_{hybrid} < P_{liquid}$		
HSHF (W/cm^2)	@85°C	@80°C	@75°C	@85°C	@80°C	@75°C
1000	84%	72%	56%	1.3%	3.8%	13.1%
1300	69%	51%	29%	6.9%	21.7%	N/A

An interesting observation is that liquid cooling cannot satisfy the $T_{\max}$ = 75 °C constraint at HSHF = 1300 without exceeding the maximum pressure drop limit. Hybrid cooling, however, is able to meet that constraint using $\{u, I\}$ = {2.2 m/s, 3.0 A} settings, which is a significant achievement considering that 2.2 m/s corresponds to only 85% of the maximum pressure drop limit. Similarly, for the highest heat flux case (i.e., HSHF = 2000), LoCool can cool the hot spot down to 80 °C by biasing the TECs with maximum current, while liquid cooling fails to meet all temperature constraints.

In comparison to hybrid designs where TECs are combined with fans, using TECs with liquid cooling provides higher cooling efficiency. In fact, for the high HSHF levels we are focusing on, TECs with fans are not sufficient to satisfy the thermal constraints [18]. The reason is that TECs require some form of cooling mechanism to remove the heat pumped to the hot side in order to avoid self heating. Liquid cooling acts as a very efficient heat sink improving the TEC performance as it achieves much lower thermal resistance compared to conventional heat sinks with fans. Another extreme case is where only a heat sink is used without the TECs. As expected, our simulations for that case give unreasonably high temperatures reaching hundreds of °C. Thus, we do not include the case with conventional heat sinks in our comparisons.

The importance of having an optimized hybrid cooling system as opposed to a nonoptimized system becomes more clear when we examine the design space for the resulting temperatures and cooling powers. We summarize such analysis in Table II. The left half of the table shows the percentage of settings that meet the thermal constraints for various cases. We observe that for a hybrid cooled system, only a fraction of the available $\{u, I\}$ pairs will meet the thermal limits, and this fraction decreases sharply down to 29% at tighter constraints. Out of that fraction, the right half shows the percentage of settings that save cooling power compared to liquid cooling system.

[4]As we compare cooling power across cases with the same peak temperature, temperature-dependent leakage has a negligible impact on the results.

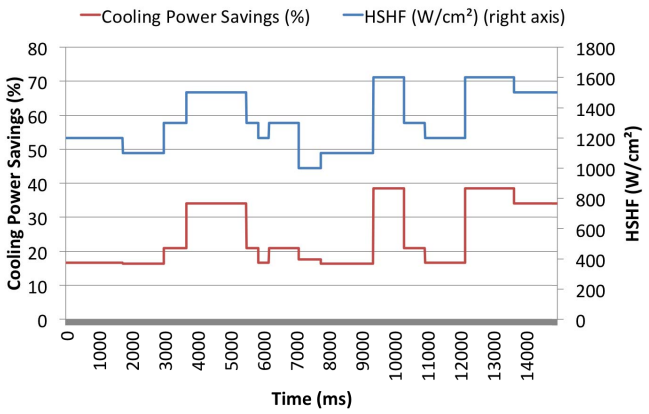

Fig. 16. Total cooling power comparison of liquid and hybrid cooling for $\text{HSHF}_{\text{avg}} \sim 1300$ W/cm^2. Results are normalized to liquid cooling at $T_{\max} = 85$ °C. Temperature constraint was not met for bars not shown.

For rather loose constraints, the cooling power consumption of the liquid and hybrid systems are close to each other. Thus, finding an optimal solution is crucial to provide benefits over liquid cooling, as only a small portion (e.g., 1.3%) of the settings will achieve that. As the constraints become tighter (e.g., HSHF = 1300 @80 °C), the inherent benefit of hybrid cooling becomes more significant. Thus, even for suboptimal $\{u, I\}$ settings, the setting we converge to provides substantial savings compared to liquid cooling.

C. Runtime Optimization Results

In this section, we evaluate the runtime operation of our LoCool algorithm. For this purpose, we generate workload traces where the HSHF changes over time. We consider three cases where the average HSHF of the workload trace is: 1) low ($1000 \leq \text{HSHF}_{\text{avg}} < 1200$); 2) medium ($1200 \leq \text{HSHF}_{\text{avg}} < 1400$); and 3) high ($1400 \leq \text{HSHF}_{\text{avg}} < 1600$). Fig. 16 shows an example trace where $\text{HSHF}_{\text{avg}} \sim 1300$ W/cm^2. On the right axis, we plot the HSHF level over time changing between 1000 and 1600 W/cm^2, while on the left axis we plot the cooling power savings over time as a percentage. As figure illustrates, when the hot spot heat density increases, hybrid cooling savings also increase reaching up to ~40% at 1600 W/cm^2. For this example, the average cooling power savings is 24.5%.

Next, we generate 20 workload traces for each of the three average HSHF cases and evaluate the average, maximum, and minimum cooling power savings in Fig. 17. When HSHF is medium, average cooling power savings range between 20% and 28%, while for high HSHF, it changes between 27% and 37% depending on the workload trace.

D. Impact of the Number of Hot Spots on Cooling Power Savings

In the previous sections, we assumed that we have a single hot spot and placed the TEC right above the hot spot. In this section, we provide an analysis on the impact of the number of hot spots on the resulting cooling power savings. For the case where a TEC is placed above each hot spot on the chip, the total TEC power will increase linearly with the number of TECs. As the number of hot spots increase, the additional TEC power consumption may surpass the savings coming from reduced liquid pumping power. The analysis we provide in this section aims to explore the extent to which the savings will be maintained with the increasing number of hot spots.

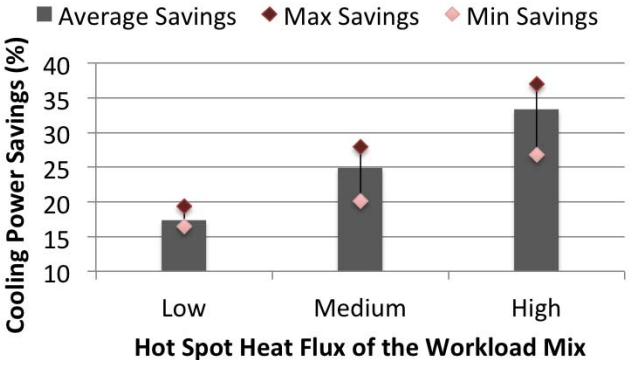

Fig. 17. Cooling power savings for the three average HSHF cases (i.e., low, medium, high) for $T_{\max} = 85$ °C. For each case, the average, maximum, and minimum savings across 20 workload traces are reported.

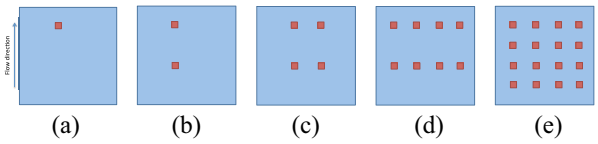

Fig. 18. Locations of the hot spots used for the multiple hot spot analysis including (a) $N = 1$, (b) $N = 2$, (c) $N = 4$, (d) $N = 8$, and (e) $N = 16$ hot spots.

TABLE III
CASES FOR WHICH LIQUID COOLING CAN OR CANNOT MEET THE TEMPERATURE CONSTRAINT FOR A GIVEN NUMBER OF HOT SPOTS AND HSHF LEVEL. NOTE THAT THE MAXIMUM COOLING WE CAN GET FROM A LIQUID-COOLED SYSTEM IS LIMITED BY THE MAXIMUM FLOW RATE, $u_{\max}$, WHICH IS DETERMINED BY THE ALLOWABLE PRESSURE DROP RECOMMENDED BY THE MANUFACTURERS

	HSHF = 1000 W/cm2			HSHF = 1300 W/cm2			HSHF = 1600 W/cm2		
	@85°C	@80°C	@75°C	@85°C	@80°C	@75°C	@85°C	@80°C	@75°C
1 hot spot	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No
2 hot spots	Yes	Yes	Yes	Yes	Yes	No	No	No	No
4 hot spots	Yes	Yes	Yes	Yes	Yes	No	No	No	No
8 hot spots	Yes	Yes	Yes	Yes	Yes	No	No	No	No
16 hot spots	Yes	Yes	No	Yes	No	No	No	No	No

For this purpose, we carry out experiments with different number of hot spots (i.e., $N = 1$, 2, 4, 8, and 16 hot spots). Above each hot spot, we place a TEC device described in Section III-E. We choose the range for the number of hot spots based on previous work, where $N = 16$ hot spots were assumed on a 20 mm × 20 mm chip [4]. The locations of the hot spots are shown in Fig. 18. Similar to the previous sections, we experiment with HSHF = 1000, 1300, and 1600 W/cm^2 and compare the two cooling methods for $T_{\max} = 85$ °C, 80 °C, 75 °C.

The first set of results compare the ability of the two cooling methods in obeying the given temperature constraint. In Table III, we report whether liquid cooling is able to keep the hot spot temperature below $T_{\max}$ or not for different number of hot spots and HSHF levels. For the highest HSHF level (i.e., 1600 W/cm^2), liquid cooling cannot meet the temperature constraint except for the case with a single hot spot and $T_{\max} = 85$ °C. For the lower HSHF levels, as N reaches 16, liquid cooling again cannot meet the temperature constraint for some cases (e.g., HSHF = 1300 and $T_{\max} = 80$ °C). As the

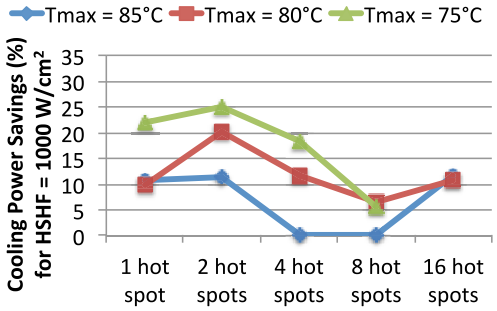

Fig. 19. Cooling power savings for varying number of hot spots and $T_{\max}$ constraints for HSHF = 1000 W/cm^2.

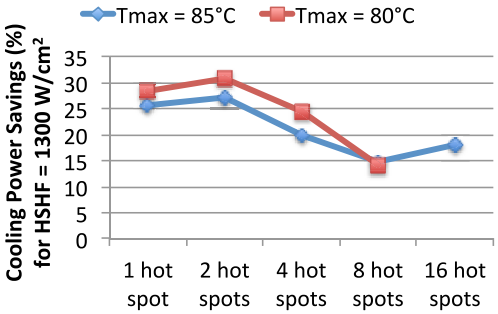

Fig. 20. Cooling power savings for varying number of hot spots and $T_{\max}$ constraints for HSHF = 1300 W/cm^2.

hot spots occur under the same liquid channel, ability of the liquid to remove heat decreases significantly. Thus, for example, for $N = 16$, where four hot spots located under the same channel along the direction of the flow, the hot spots located closest to the outlet cannot receive sufficient cooling.

On the other hand, hybrid cooling mitigates the aforementioned problem by removing similar amount of heat from each hot spot regardless of its position with respect to the microchannel. In this way, hybrid cooling can meet the temperature constraint for all of the cases shown in Table III except for one (i.e., $N = 16$ with HSHF = 1600 and $T_{\max} = 75$ °C). The cooling benefits provided by TECs comes at the cost of power consumption with each TEC that is placed. For example, in the worst case scenario with $N = 16$, HSHF = 1600, and $T_{\max} = 80$ °C, the total cooling power is dominated by TEC power and it reaches a maximum of 30 W.

In the second set of results, we present the trend in the cooling power savings for varying N and $T_{\max}$. We focus on the cases for which liquid and hybrid cooling can both meet the $T_{\max}$ constraint. Figs. 19 and 20 summarize the results for HSHF = 1000 and HSHF = 1300, respectively. As illustrated, the cooling power savings versus the number of hot spots curve does not have a monotonic behavior. In order to explain the reasoning behind the observed trend, let us focus on Fig. 19 and $T_{\max} = 80$ °C. As the N increases from 1 to 2, the cooling power savings rise from 10% to 20%. This is due to the fact that the second hot spot was located under the same microchannel as the first hot spot in the direction of the flow, significantly decreasing the liquid cooling efficiency as previously mentioned. On the other hand, moving from $N = 2$ to $N = 4$, the additional two hot spots were placed under a different microchannel as shown in Fig. 18(b) and (c)]. For such scenario, the cooling ability of the liquid stays the same and thus, the same pumping power is sufficient to cool down four hot spots. However, for the hybrid cooling case, TEC power will double from $N = 2$ to $N = 4$. This explains the drop in cooling savings of the hybrid solution from $N = 2$ to $N = 4$, as well as from $N = 4$ to $N = 8$ hot spots. Similarly, increasing N from 8 to 16 results in more hot spots to be clustered under the same channel and thus, cooling power savings rise again.

Another interesting trend in Fig. 19 and $T_{\max} = 85$ °C. For this case, liquid cooling is able to mitigate the hot spots more easily requiring low flow velocities, and the hybrid cooling savings are lower. Therefore, when increasing N from 2 to 4 and then to 8, the cooling power savings drop to 0%. In those cases, the optimum cooling power settings for both liquid cooling and hybrid cooling correspond to the same liquid flow velocity, while for hybrid cooling the TEC bias current is set to 0 A. These results indicate that TEC benefits are more significant when the hot spots are clustered under the same microchannels, the heat flux is high and $T_{\max}$ is low. On the other hand, when hot spots are scattered across different channels and the number of hot spots increase, the cooling power benefits start to decrease, since liquid cooling can also provide sufficient cooling while consuming similar power.

VI. Conclusion

In this paper, we develop a compact hybrid cooling model, which is able to account for the thermal behavior of liquid microchannels together with TECs. We validate the accuracy of our model by comparing its results against COMSOL and 3D-ICE simulators and demonstrate an average error of less than 2.9 °C, while speeding up the simulation by up to four orders of magnitude. We then propose LoCool, a cooling power optimization method for systems that adopt hybrid cooling. LoCool optimizes hybrid systems combining microchannel-based liquid cooling and TEC cooling for hot spot mitigation in a localized manner. It finds the most energy efficient coolant flow rate and TEC current settings to minimize cooling power for a given power map and a temperature constraint. Using our proposed thermal model, we evaluate the benefits of hybrid cooling designs optimized using LoCool over homogeneous liquid cooling designs and demonstrate up to 40% cooling power savings. We also show in addition to saving cooling power, hybrid cooling with LoCool can mitigate hot spots with much higher heat fluxes, which are not achievable using liquid cooling only. Finally, we explore the impact of the number of hot spots on the cooling power savings of hybrid cooling designs. One direction to follow in multiple hot spot scenario is to design optimum TEC placement algorithms.

References

[1] J. Srinivasan *et al.*, "RAMP: A model for reliability aware microprocessor design," IBM, Armonk, NY, USA, Rep. RC23048(W0312-122), Dec. 2003.

[2] C.-C. Lee and J. D. Groot, "On the thermal stability margins of high-leakage current packaged devices," in *Proc. Electron. Packag. Technol. Conf.*, Dec. 2006, pp. 487–491.

[3] N. S. Kim *et al.*, "Leakage current: Moore's law meets static power," *Computer*, vol. 36, no. 12, pp. 68–75, Dec. 2003.

[4] M. Schultz *et al.*, "Embedded two-phase cooling of large three-dimensional compatible chips with radial channels," *ASME J. Electron. Packag.*, vol. 138, no. 2, pp. 1–5, Jun. 2016.

[5] Z. Lu *et al.*, "Design and modeling of membrane-based evaporative cooling devices for thermal management of high heat fluxes," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 6, no. 7, pp. 1056–1065, Jul. 2016.

[6] M. M. Sabry, A. Sridhar, J. Meng, A. K. Coskun, and D. Atienza, "GreenCool: An energy-efficient liquid cooling design technique for 3-D MPSoCs via channel width modulation," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 32, no. 4, pp. 524–537, Apr. 2013.
[7] V. Sahu *et al.*, "Experimental characterization of hybrid solid-state and fluidic cooling for thermal management of localized hotspots," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 5, no. 1, pp. 57–64, Jan. 2015.
[8] K. Yazawa *et al.*, "Cooling power optimization for hybrid solid-state and liquid cooling in integrated circuit chips with hotspots," in *Proc. 13th IEEE Intersoc. Conf. Thermal Thermomechan. Phenom. Electron. Syst. (ITherm)*, May 2012, pp. 99–106.
[9] Y. Hu, S. Chen, L. Peng, E. Song, and J.-W. Choi, "Effective thermal control techniques for liquid-cooled 3D multi-core processors," in *Proc. Int. Symp. Qual. Electron. Design*, Mar. 2013, pp. 8–15.
[10] F. Paterna and S. Reda, "Mitigating dark-silicon problems using superlattice-based thermoelectric coolers," in *Proc. Design Autom. Test Europe Conf. Exhibit.*, Mar. 2013, pp. 1391–1394.
[11] S. Jayakumar and S. Reda, "Making sense of thermoelectrics for processor thermal management and energy harvesting," in *Proc. ISLPED*, Jul. 2015, pp. 31–36.
[12] A. K. Coskun, T. S. Rosing, J. L. Ayala, and D. Atienza, "Modeling and dynamic management of 3D multicore systems with liquid cooling," in *Proc. VLSI-SoC*, 2009, pp. 35–40.
[13] A. Sridhar, A. Vincenzi, M. Ruggiero, T. Brunschwiler, and D. Atienza, "3D-ICE: Fast compact transient thermal modeling for 3D ICS with inter-tier liquid cooling," in *Proc. ICCAD*, 2010, pp. 463–470.
[14] A. Sridhar, A. Vincenzi, M. Ruggiero, T. Brunschwiler, and D. Atienza, "Compact transient thermal model for 3D ICS with liquid cooling via enhanced heat transfer cavity geometries," in *Proc. Int. Workshop Thermal Invest. ICs Syst. (THERMINIC)*, Oct. 2010, pp. 1–6.
[15] A. Fourmigue, G. Beltrame, and G. Nicolescu, "Efficient transient thermal simulation of 3D ICS with liquid-cooling and through silicon vias," in *Proc. DATE*, Mar. 2014, pp. 1–6.
[16] X.-X. Liu, Z. Liu, S.-D. Tan, and J. Gordon, "Full-chip thermal analysis of 3D ICS with liquid cooling by GPU-accelerated GMRES method," in *Proc. ISQED*, Mar. 2012, pp. 123–128.
[17] R. A. Taylor and G. L. Solbrekken, "Comprehensive system-level optimization of thermoelectric devices for electronic cooling applications," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 31, no. 1, pp. 23–31, Mar. 2008.
[18] I. Chowdhury *et al.*, "On-chip cooling by superlattice-based thin-film thermoelectrics," *Nat. Nanotechnol.*, vol. 4, no. 4, pp. 235–238, 2009.
[19] *COMSOL Multiphysics Software*. Accessed: Jan. 2014. [Online]. Available: http://www.comsol.com
[20] *ANSYS*. Accessed: Jan. 2017. [Online]. Available: http://www.ansys.com
[21] F. Kaplan, S. Reda, and A. K. Coskun, "Fast thermal modeling of liquid, thermoelectric, and hybrid cooling," in *Proc. IEEE Intersoc. Conf. Thermal Thermomechan. Phenom. Electron. Syst. (ITHERM)*, Jun. 2017, pp. 726–735.
[22] K. Skadron *et al.*, "Temperature-aware microarchitecture," in *Proc. ISCA*, 2003, pp. 2–13.
[23] A. K. Coskun, D. Atienza, T. S. Rosing, T. Brunschwiler, and B. Michel, "Energy-efficient variable-flow liquid cooling in 3D stacked architectures," in *Proc. DATE*, 2010, pp. 111–116.
[24] M. M. Sabry, A. K. Coskun, D. Atienza, T. Š. Rosing, and T. Brunschwiler, "Energy-efficient multiobjective thermal control for liquid-cooled 3-D stacked architectures," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 30, no. 12, pp. 1883–1896, Dec. 2011.
[25] H. Qian, C.-H. Chang, and H. Yu, "An efficient channel clustering and flow rate allocation algorithm for non-uniform microfluidic cooling of 3D integrated circuits," *Integr. VLSI J.*, vol. 46, no. 1, pp. 57–68, 2013.
[26] B. Shi and A. Srivastava, "Optimized micro-channel design for stacked 3-D-ICs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 33, no. 1, pp. 90–100, Jan. 2014.
[27] C. S. Sharma *et al.*, "Energy efficient hotspot-targeted embedded liquid cooling of electronics," *Appl. Energy*, vol. 138, pp. 414–422, Jan. 2015.
[28] V. Sahu *et al.*, "Energy efficient liquid-thermoelectric hybrid cooling for hot-spot removal," in *Proc. IEEE SEMI-THERM*, Mar. 2012, pp. 130–134.
[29] M. J. Dousti and M. Pedram, "Platform-dependent, leakage-aware control of the driving current of embedded thermoelectric coolers," in *Proc. ISLPED*, Sep. 2013, pp. 311–316.
[30] M. J. Dousti and M. Pedram, "Power-aware deployment and control of forced-convection and thermoelectric coolers," in *Proc. DAC*, Jun. 2014, pp. 1–6.
[31] S. Pagani, M. Shafique, H. Khdr, J.-J. Chen, and J. Henkel, "seBoost: Selective boosting for heterogeneous manycores," in *Proc. Int. Conf. Hardw. Softw. Codesign Syst. Synth. (CODES)*, 2015, pp. 104–113.
[32] S. Pagani *et al.*, "TSP: Thermal safe power: Efficient power budgeting for many-core systems in dark silicon," in *Proc. CODES*, 2014, p. 10.
[33] M. Shafique, S. Garg, J. Henkel, and D. Marculescu, "The EDA challenges in the dark silicon era: Temperature, reliability, and variability perspectives," in *Proc. DAC*, 2014, pp. 1–6.
[34] J. Meng, K. Kawakami, and A. K. Coskun, "Optimizing energy efficiency of 3-D multicore systems with stacked DRAM under power and thermal constraints," in *Proc. DAC*, 2012, pp. 648–655.
[35] R. K. Shah and A. L. London, *Laminar Flow Forced Convection in Ducts: A Source Book for Compact Heat Exchanger Analytical Data*. New York, NY, USA: Academic, 1978.
[36] *Odroid XU+E Development Board*. Accessed: May 2017. [Online]. Available: http://www.hardkernel.com/main/products
[37] B. Su *et al.*, "PPEP: Online performance, power, and energy prediction framework and DVFS space exploration," in *Proc. IEEE Micro*, Dec. 2014, pp. 445–457.
[38] V. Adhinarayanan *et al.*, "Measuring and modeling on-chip interconnect power on real hardware," in *Proc. IISWC*, Sep. 2016, pp. 1–11.
[39] G. Wu, J. L. Greathouse, A. Lyashevsky, N. Jayasena, and D. Chiou, "GPGPU performance and power estimation using machine learning," in *Proc. HPCA*, Feb. 2015, pp. 564–576.
[40] S. Reda and A. Belouchrani, "Blind identification of power sources in processors," in *Proc. DATE*, Mar. 2017, pp. 1739–1744.
[41] T. Brunschwiler, A. Sridhar, C. Ong, and G. Schlottig, "Benchmarking study on the thermal management landscape for 3D ICs: From back-side to volumetric heat removal," *ASME Int. Electron. Packag. Tech. Conf. Exhibit.*, vol. 1, 2015, Art. no. V001T09A069.

Fulya Kaplan (M'12) received the B.S. degree in electrical and electronics engineering from Middle East Technical University, Ankara, Turkey, in 2011, and the Ph.D. degree in computer engineering from Boston University, Boston, MA, USA, in 2017.

Her current research interests include thermal modeling and runtime optimization of advanced processor cooling solutions, as well as the simulation and energy-efficient job allocation in HPC data centers.

Mostafa Said (M'18) received the Ph.D. degree in electronics and communications engineering from the Egypt–Japan University of Science and Technology, Alexandria, Egypt, in 2015.

He joined Assiut University, Asyut, Egypt, as an Assistant Professor. He is currently a Post-Doctoral Fellow with SCALE Lab, Brown University, Providence, RI, USA. His current research interests include network-on-chips, thermal management and modeling of SoCs, and 3-D integration.

Sherief Reda (S'01–M'06–SM'14) received the Ph.D. degree in computer science and engineering from the University of California at San Diego, San Diego, CA, USA.

In 2006, he joined the Computer Engineering Group, Brown University, Providence, RI, USA, where he is currently an Associate Professor of engineering and computer science. His current research interests include energy-efficient computing, thermal-power sensing and management, and low-power design techniques.

Ayse K. Coskun (M'06–SM'16) received the M.S. and Ph.D. degrees in computer science and engineering from the University of California at San Diego, San Diego, CA, USA.

She was with Sun Microsystems (currently, Oracle), San Diego, CA, USA. She is currently an Associate Professor with the Department of Electrical and Computer Engineering, Boston University, Boston, MA, USA. Her current research interests include energy-efficient computing, 3-D-stacked architectures, computer architecture, and embedded systems and software.