International Journal of Heat and Mass Transfer 88 (2015) 684–694

Contents lists available at ScienceDirect

International Journal of Heat and Mass Transfer

journal homepage: www.elsevier.com/locate/ijhmt

A novel method of energy efficient hotspot-targeted embedded liquid cooling for electronics: An experimental study

Chander Shekhar Sharma [a], Gerd Schlottig [b], Thomas Brunschwiler [b], Manish K. Tiwari [c], Bruno Michel [b], Dimos Poulikakos [a,*]

[a] Laboratory of Thermodynamics in Emerging Technologies, Institute of Energy Technology, Department of Mechanical and Process Engineering, ETH Zurich, 8092 Zurich, Switzerland
[b] Advanced Micro Integration, IBM Research – Zurich, 8803 Rüschlikon, Switzerland
[c] Department of Mechanical Engineering, University College London (UCL), Torrington Place, London WC1E 7JE, UK

ARTICLE INFO

Article history:
Received 19 November 2014
Received in revised form 16 April 2015
Accepted 16 April 2015
Available online 20 May 2015

Keywords:
Hotspot-targeted cooling
Microchannel cooling
Electronics cooling
Hotspots
Energy efficient computing
Multicore microprocessors

ABSTRACT

The shift to multicore microprocessor architecture is likely to result in higher coolant flow requirements and thus exacerbate the problem of increasing data center energy consumption, also with respect to hotspot elimination. We present and experimentally prove a novel concept, for embedded, hotspot-targeted and energy efficient cooling of heterogeneous chip power landscapes. The rationally distributed, embedded microstructures presented here are able to adapt the heat transfer capability to a steady but non-uniform chip power map by passively throttling the flow in low heat flux areas. For the industrially acceptable limit on pressure drop of approximately 0.4 bar, the hotspot-targeted embedded liquid cooling (HT-ELC) designs are evaluated against a conservatively chosen conventional embedded liquid cooling (C-ELC) design and existing heat sinks in the literature. For an average steady-state heat flux of 150 W/cm^2 in core areas (hotspots) and 20 W/cm^2 over the remaining chip area (background), the chip temperature variation is reduced from 10 °C under the conventional cooling to 4 °C under the current hotspot targeted heat sink – a reduction of 57%. For heat fluxes of 300 and 24 W/cm^2, the temperature variation is reduced by 30%. We show that the HT-ELC designs consume less than 0.3% of total chip power as pumping power to achieve this thermal performance, which the C-ELC design cannot match under all feasible levels of pumping power. Moreover, the HT-ELC designs achieve at least 70% improvement over the existing hotspot targeted heat sinks in terms of normalized chip temperature non-uniformity, without the need for any additional system level complexity, reducing reliability risks.

1. Introduction

Microprocessors form the backbone of computing hardware today, a result of progressive reduction in the cost of computing power due to downscaling of transistors. As early as 1965, Gordon Moore had predicted an exponential growth in fabrication density of integrated circuits. Later, Dennard scaling rules for transistors and interconnects enabled the semiconductor industry to transform this prediction into a self-fulfilling prophecy [1,2]. The reduced cost of computing power has, in turn, fueled the growth of Information Technology (IT) related services such as internet, telephony and e-commerce through large data centers that use hundreds of such microprocessors in multiple servers housed in one location. Because all sectors of economy increasingly use these services, data centers consume now unprecedented levels of energy. In 2010, direct electricity consumption by data centers reached 238 billion kWh annually which constituted nearly 1.31% of total world electricity consumption [3,4]. In spite of a relative slowdown between 2005 and 2010 [4], the still increasing data center energy consumption is now both an economical concern due to energy cost dominating over hardware cost, and an ecological concern because it is not sustainable [5,6].

Most of the present day data centers are air-cooled. This requires elaborate cooling infrastructure with air-cooling overhead contributing roughly 33% of the total energy bill of these systems. This is due to the inability to scale supply voltages as per Dennard scaling rules, which has resulted in a steep increase in chip heat flux dissipation densities. The related concerns of electronic reliability have progressively increased cooling demands of modern microprocessors, especially during the last decade [7–9].Due to the high thermal capacity and the much lower thermal resistance inherent in liquid coolant use, a switch from air to liquid cooling reduces data center energy consumption significantly. Liquid cooled microstructure

* Corresponding author. Tel.: +41 44 632 27 38; fax: +41 44 632 11 76.
E-mail address: dpoulikakos@ethz.ch (D. Poulikakos).

http://dx.doi.org/10.1016/j.ijheatmasstransfer.2015.04.047
0017-9310/

Nomenclature

ΔP	pressure drop
PS	power supply
q''	heat flux
T_f	fluid temperature
$T_{J,max}$	maximum junction temperature of chip
$\Delta T_{J,max}$	chip wide temperature difference $(= T_{J,max} - T_{J,min})$
ΔT_q	temperature non-uniformity scaled by heat flux contrast
$\dot{V}_f$	coolant flow rate
w_c	microchannel width
W, L	overall chip dimensions
w, l	dimensions of power map
$W_{pumping}$	pumping power
$\dot{W}_{chip}$	total power dissipated by the chip

Greek letters

α	heat transfer coefficient
φ_m	fraction of total coolant flow flowing through HSBG zone

Subcripts

b	base
bg, *hs*, *th*	background, hotspot and throttling zones
d	distance between hotspots
in, *out*	inlet and outlet
w	wall

heat sinks are now well established as a viable solution for microprocessors cooling. Starting with the early work of Tuckerman and Pease [10], a large body of work has analyzed microchannel liquid cooling including traditional microchannel heat sinks [11], manifold microchannel heat sinks [12–17], and spray and jet cooling [18,19]. However, most of these studies have sought to reduce only the maximum chip junction temperature ($T_{J,max}$). It is also important that the chip temperature non-uniformity ($\Delta T_{J,max} = T_{J,max} - T_{J,min}$) is minimized because large temperature gradients cause large gradients of thermal expansion and increase thermal stresses in the chip to substrate or heat sink interface, reduce electronic reliability in regions of high temperature thus causing differential aging, and create circuit imbalances in CMOS devices [8]. A few studies in the literature have focused on reducing $\Delta T_{J,max}$ across the electronic chip. These include utilization of flow boiling in microchannels [20], variable pin fin density, single phase, liquid cooled heat sinks [21] and variable microchannel width cooling schemes [22]. However, all these efforts have considered only uniform chip power maps, which will not be relevant for future microprocessors, as discussed below.

During the last decade, power and memory limitations have led to an industry-wide shift to multicore processors [23] that consist of distinct cores containing the execution units. The rest of the chip mainly contains cache, memory controllers and chip interconnects [24]. The cores dissipate multiple times the heat flux as compared to the rest of the chip resulting in a highly non-uniform chip power map. As the semiconductor industry tries to keep pace with Moore's law, it is likely that the next decade will see multicore processors with increasing number of cores and increasing transistor count per core. Coupled with limited voltage scaling, the total chip power would continue to increase, although at a slower rate than present [25].

As the processing power becomes concentrated in the cores, the increasing non-uniformity in chip power maps will have two distinct consequences. First, the chip reliability problems, as discussed above, are likely to become exacerbated. Large thermal gradients affect heterogeneous chip and package architectures more severely, especially in the case of the 3-dimensional (3D-IC) chip stacks. In 3D-ICs, chip makers are trying to scale chip performance by adding layers of planar chips in the vertical direction to augment the decelerating planar scaling [26,27]. In such chip stacks, the non-uniformity problem would persist in each layer. Additionally, any dissimilarity between power maps of individual chips constituting the stack would add another geometrical dimension to thermal non-uniformity. Second, and more importantly, in order to maintain the microprocessor reliability, higher coolant volume requirements will increase the pumping power cost of cooling. Consequently, conventional liquid cooling will not be sufficient to address the cooling and energy efficiency requirements of the future data centers using multicore planar or 3D-ICs.

In order to avoid this energy dilemma, a cooling approach targeted at the high heat flux dissipating areas (henceforth referred to as hotspots since with conventional cooling, the high temperature regions will coincide with the high heat flux dissipating regions) is required. We refer to such an approach of selectively focused cooling as hotspot-targeted cooling for which several approaches have been proposed in the past. Prominent examples include active hotspot targeting such as thermoelectric cooling and electrowetting. However, these methods suffer from inherent limitations such as contact parasitic resistances [28] and low heat flux pumping capacities [29]. Attempts involving passive hotspot targeting by using single phase liquid cooling have also been reported [30–32] and are discussed in more detail in Section 4.

In this work, we demonstrate a novel, hotspot-targeted and at the same time, highly energy efficient cooling concept for typical modern day and future non-uniform microprocessor power landscapes. We employ a passive, energy saving approach to alter the heat sink design by optimizing the microchannel geometry and flow rate distribution through passively throttling the flow in regions of low heat flux in order to direct more flow towards regions of high heat flux. This inherently achieves a much more uniform chip temperature map while consuming very low pumping power. The paper is organized as follows: We first describe the hotspot-targeted microchannel liquid cooling concept in Section 2. Section 3 covers the experimental set-up and measurement procedure. Section 4 presents the performance of the hotspot-targeted designs compared to conventional embedded liquid cooling and to existing hotspot-targeted cooling approaches.

2. The hotpot-targeted microchannel cooling concept

The starting point of our hotspot-targeted cooling concept is the manifold microchannel (MMC) heat sink cooling architecture shown in Fig. 1 [12–15]. As shown in the figure, the coolant accesses the microchannel heat transfer structure through multiple and alternating inlet and outlet manifolds positioned above the microchannels. This reduces the overall pressure drop penalty as well as improves the overall thermal performance by reducing the thickness of the thermal boundary layers and by utilizing the enhanced heat transfer due to impingement jet cooling at the inlet slot nozzles. In conventional MMC architecture, the heat sink is