Thermal Analysis and Active Cooling Management for 3D MPSoCs

Mohamed M. Sabry[‡], David Atienza[‡], and Ayse K. Coskun[†]
[‡]Embedded Systems Laboratory (ESL), Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland.
[†]Electrical and Computer Engineering Department, Boston University, USA.

***Abstract*— 3D stacked architectures reduce communication delay in multiprocessor system-on-chips (MPSoCs) and allowing more functionality per unit area. However, vertical integration of layers exacerbates the reliability and thermal problems, and cooling is a limiting factor in multi-tier systems. Liquid cooling is a highly efficient solution to overcome the accelerated thermal problems in 3D architectures. However, liquid cooling brings new challenges in modeling and run-time management. This paper proposes a design-time/run-time thermal management policy for 3D MPSoCs with inter-tier liquid cooling. First, we perform a design-time analysis to estimate the thermal impact of liquid cooling and dynamic voltage frequency scaling (DVFS) on 3D MPSoCs. Based on this analysis, we define a set of management rules for run-time thermal management. We utilize these rules to control and adjust the liquid flow rate in order to match the cooling demand for preventing energy wastage of over-cooling, while maintaining a stable thermal profile in the 3D MPSoCs. Experimental results on multi-tier 3D MPSoCs show that proposed design-time/run-time management policy prevents the system to exceed the given threshold temperature while reducing cooling energy by 50% on average and system-level energy by 18% on average in comparison to using a static worst-case flow rate setting.**

I. Introduction

3D integration is a recently proposed design method for overcoming the limitations regarding the delay, bandwidth, and power consumption of the interconnects in multiprocessor system-on-chips (MPSoCs), while reducing the chip footprint and improving the fabrication yield. However, one of the main challenges for designing 3D circuits is their elevated temperatures resulting from higher thermal resistivity [8,10]. Thus, it is more difficult to remove the heat from 3D MPSoCs. 3D MPSoCs are also prone to large thermal variations; e.g., cores located at different tiers or at different coordinates across a tier have significantly different heating/cooling rates [3]. These large thermal variations have adverse effects on the system reliability, performance, and cooling costs.

A number of thermal management techniques have been proposed for controlling temperature on 3D MPSoCs by extending the management techniques for workload scheduling and Dynamic Voltage and Frequency Scaling (DVFS)-based thermal management in 2D MPSoCs (e.g., [6,7,16,17]). For example, Zhu et al. evaluate several run-time policies for task migration and DVFS, with the aid of an offline analysis [17]. However, the temperature recorded in this paper exceeds 85^oC implying that 3D MPSoCs have a high thermal profile. In addition, with increasing power densities in a fixed area, conventional management techniques with passive control elements (e.g., DVFS) are incapable of reducing the temperature of these systems efficiently. Moreover, as power densities, number of cores, and number of tiers increase, extremely high temperature values appear in 3D MPSoCs [16], resulting in severe restrictions in high-performance 3D MPSoC design.

Active inter-tier liquid cooling technology is a promising solution to address the high temperatures in 3D chips, due to the higher heat removal capability of liquids in comparison to air [2,5] (see Fig. 1(a) and 1(b)). This technology involves injecting fluid (e.g., water) through microchannels (or other structures) between the tiers of a 3D stack using a pump to remove the heat. The heat removal capability of inter-tier heat-transfer with pin-fin in-line structures for 3D chips is investigated in prior work [2]. However, the authors have been more focused in finding thermal packaging solutions, not using these solutions in run-time thermal management.

While liquid cooling has a large capability in terms of thermal reduction of 3D MPSoCs, it is necessary to use this technique in conjunction with other thermal management techniques to exploit the trade-offs with other key parameters in 3D MPSoCs, such as energy efficiency and performance. Prior liquid cooling work evaluates existing thermal management policies on a 3D MPSoC with a fixed-flow rate setting, and also investigates the benefits of variable flow using a policy to change the flow rate based on temperature measurements [4]. In addition, recent work considers the energy efficiency of 3D MPSoC having a variable flow rate and thermally-aware load balancing, taking into consideration the cooling power consumption [5]. However, the use of liquid cooling in these cases has been decoupled from other thermal management technique such as DVFS.

Since the integration of inter-tier liquid cooling has an impact on the 3D MPSoC thermal characteristics, it is necessary to perform a design-time thermal-response analysis of inter-tier liquid cooling-based 3D MPSoC to each of the thermal control knobs. This analysis leads to the deduction of both the appropriate run-time control strategy at a specific system state and the inputs that lead to this state.

In this paper, we propose a novel 3D MPSoC design-time/run-time thermal management policy. We first perform a complete design-time thermal analysis of inter-tier liquid cooling-based 3D MPSoCs with respect to varying flow rate and DVFS using 3D-ICE [14], which is a verified transient

This research has been partially funded by the Nano-Tera RTD project CMOSAIC (ref.123618), financed by the Swiss Confederation and scientifically evaluated by SNSF, and the PRO3D EU FP7-ICT-248776 project.

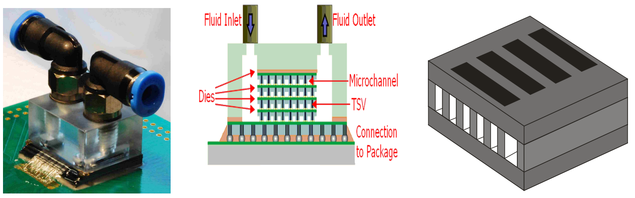
978-1-4244-9472-9/11/$26.00 ©2011 IEEE

thermal modeling tool of 3D stacks with inter-tier liquid cooling. The results of 3D-ICE simulations have been validated with a real 5-tier 3D stack [14]. In addition, we extend in this work the 3D-ICE modeling tool for 3D-ICs to account for a run-time varying flow rate. We explore the thermal impact of tuning each of the thermal control knobs, while fixing the other knobs at a specific value. Next, based on this design-time thermal analysis, we deduce a set of run-time thermal management rules that adjusts the injected flow rate and the VF settings of each processing element to minimize the energy consumption of the system, while keeping the temperature below the thermal threshold, 85°C. Finally, we deploy these rules in a run-time rule-base controller integrated in a 2- and 4-tier 3D MPSoCs showing that our proposed management policy prevents the violation of the thermal threshold while achieving 50% and 18% average reduction in cooling and system-level energy, respectively, in comparison to setting the flow rate at the maximum value to handle the worst-case temperature.

II. Thermal Response Analysis in Liquid-Cooled 3D MPSoCs

In this section we explore the design-time thermal response analysis of two main thermal control parameters in 3D MPSoCs: variant inter-tier fluid flow rate and DVFS [2,5]. We analyze the thermal response and impact of each of these techniques on 3D MPSoC designs with respect to the worst case and typical operating conditions.

We model an infinite thread input fully utilizing the system, which is executed on a variable number of active cores in the 3D test-bed. Fig. 1 shows the manufactured prototype, cross-section, and layout of the test-bed we use in this exploration, which is based on the experimental thermal validation of 3D stacks presented in prior work [2,14] ($1cm^2$ die area). In particular, we provide the analysis for a 3D MPSoC that consists of two tiers, where each tier contains four main hot-spot sources representing $5mm \times 2mm$ high performance processors and dissipating $250W/cm^2$. The remaining area contains background heaters that play the role of caches and other interconnection blocks and dissipate $50W/cm^2$. In this test-bed, as in the manufactured 3D test chips with liquid cooling, the liquid flows into the microchannels from the inlet port (left side) to the outlet port (right side) of the stack. The microchannel dimensions are $50\mu m$ (width)$\times$ $100\mu m$ (height) with a $100\mu m$ pitch [2,14]. The range of flow rates used is 0.1 to $0.2l/min$, which is the same range provided in prior work [2].



(a) Prototype (b) Cross-Section (c) Schematic Layout

Fig. 1. The manufactured prototype, cross-section, and layout of the test-bed we use in this exploration. The explored 3D test-bed (c) has four hot-spot sources (black), liquid microchannels (white), and background heaters (gray).

A. Variant Liquid Flow Rate

We first examine the effect of changing the liquid flow rate, while maintaining the VF settings of the processing elements at the maximum values. This exploration involves the dynamic variation of the injected flow rate (i.e., between $0.1 - 0.2l/min$) to observe the temperature of the controlled elements that are located at various distances from the fluid inlet port. We quantize the flow rate range into five different values: $0.1, 0.125, 0.15, 0.175, 0.2l/min$.

Fig. 2 shows the thermal impact when all cores are active simultaneously in each simulation run. This figure illustrates that varying the liquid flow rate has a more significant impact on the cores that are furthest from the inlet port; e.g., we observe a thermal reduction up to 40^oC between the maximum and minimum injected fluid flow rates for such cores. This observation implies that it is highly beneficial to increase the flow rate when these cores (furthest from inlet port) have high temperatures.

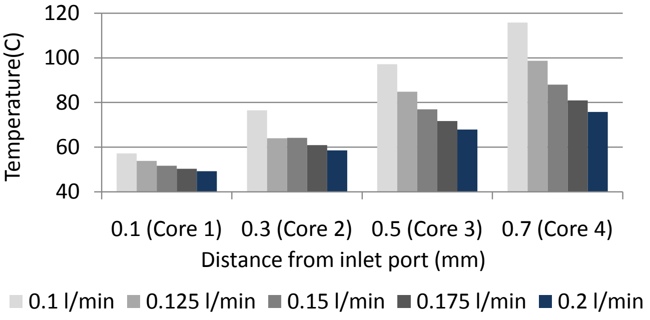


Fig. 2. Thermal response of processing cores with respect to their distance from the fluid inlet port at various flow rates. All cores are active simultaneously.

We observe that the cores with the closest relative distance to the inlet ports experience a lower temperature. In fact, since the water has the lowest temperature when it is injected in the microchannel, the thermal gradient between the chip and the liquid is high and more heat can be absorbed by the liquid. The magnitude of this gradient is lower when the liquid flows in the microchannel towards the outlet port.

Finally, our experiments indicate a considerable delay in the response time of the fluid actuating system (pumps or valves) with respect to the cores' switching activity speed. In fact, this delay can reach up to $500ms$. This implies a slow reaction process with respect to fast changes in 3D MPSoC switching activity (i.e., workload changes in processing elements), which limits the run-time varying flow rate application as a reactive process, and motivates the exploration of other (passive) thermal management approaches that have a faster reaction times for workload changes.

B. Dynamic Voltage and Frequency Scaling

We explore the application of a distributed DVFS policy, while maintaining the flow rate at its minimum value. Although DVFS causes performance cost due to the slowdown of the processing elements, in this analysis we consider the execution of a thread of an infinite duration. Thus, the degradation is not considered. Moreover, we neglect the switching overhead since it is of microsecond range [6,7].

To examine the effects of DVFS on temperature, we use a simple two-point threshold-based control [7], where the frequency of a certain core is decreased when the temperature exceeds a certain value T_1, and increased when the temperature falls below another value T_2 ($T_1 > T_2$). We select three (T_1,T_2) pairs: $[(77, 73), (80, 78), (85, 82)]^oC$. We find that the cores closest to the fluid inlet port do not change their VF settings for temperature control. However, as the elements are located further from the inlet port, more switchings occur to reduce the temperature of these elements. In fact, when the temperature thresholds are at the lowest range (77,73°C), the middle-distance elements (i.e., 0.5mm from the input port) apply DVFS 50% of the time, and the elements located furthest from the inlet port (i.e., 0.7mm) do not benefit from the scaling since their temperature is much higher than the requested thermal control threshold. Moreover, when the temperature control threshold is increased (85,82°C), then the cores at the end of the microchannel also start experiencing an increase of VF switchings (at a higher rate than the switching frequencies of the cores closer to the inlet port) to maintain a stable thermal profile.

Consequently, we conclude that DVFS does not achieve a high temperature reduction of the elements furthest from the inlet port as efficiently as varying the liquid flow. Thus, this factor must to be taken into consideration when combining both active (i.e., dynamic liquid flow rate changes) and passive (e.g., DVFS) techniques in order to design an effective thermal management controller for 3D MPSoCs with inter-tier liquid cooling.

Based on our analysis, we extract a set of rules that derive a run-time thermal manager to minimize the energy consumption while maintaining the temperature below a threshold value. For example, we conclude that **IF** cores are allocated nearest to the inlet port, **THEN** they operate at the highest VF settings and minimum flow rate with no thermal violations. On the other hand, **IF** the cores allocated furthest to the inlet port **AND** their temperature is near the threshold value (high), **THEN** we increase the flow rate. In addition, **IF** the core utilization allows load relaxation without performance degradation, **THEN** we scale down the VF settings.

We extract the complete set of rules for a run-time energy-efficient thermal management policy for 3D MPSoCS with inter-tier cooling, and deploy them in a rule-base controller, i.e., Takagi-Sugeno fuzzy-logic controller [15]. More details on rules integration to this controller are given in [11].

III. Experimental Results

The 3D MPSoCs we use in our experiments are based on the 90nm UltraSPARC T1 (i.e., Niagara-1) processor [9]. The power consumption, area, and floorplan of UltraSPARC T1 are available in [9]. UltraSPARC T1 has 8 multi-threaded cores, and a shared L2-cache for every two cores. Our simulations are carried out with 2-, and 4-tier 3D MPSoCs. For thermal and performance measurement, we use the workload statistics provided in prior work [5]. We assume three VF settings in the DVFS policy in our simulations: $[(1.2V, 1.2GH_Z), (1.1V, 1.0GH_Z), (1.0V, 0.8GH_Z)]$.

TABLE I

Thermal and floorplan parameters deployed in the 3D MPSoC model

Parameter	Value
Silicon conductivity	$130W/(m \cdot K)$
Silicon capacitance	$1635660J/(m^3 \cdot K)$
Wiring layer conductivity	$2.25W/(m \cdot K)$
Wiring layer capacitance	$2174502J/(m^3 \cdot K)$
Water conductivity	$0.6W/(m \cdot K)$
Water capacitance	$4183J/(kg \cdot K)$
Heat sink conductivity (air cooling only)	$10W/K$
Heat sink capacitance (air cooling only)	$140J/K$
Die Thickness (one stack)	$0.15mm$
Area per Core	$10mm^2$
Area per L2 Cache	$19mm^2$
Total Area of Each Layer	$115mm^2$
Inter-tier Material Thickness	$0.1mm$
Channel width	$0.05mm$
Channel pitch	$0.15mm$
Flow rate range	$0.01 - 0.0323l/min$ per cavity
Pumping network power	$3.5 - 11.176W$

In the thermal modeling tool (3D-ICE [14]), we use a temperature sampling interval of 100 ms, and all simulations are initialized with steady state temperature values. The model parameters are provided in Table I. This table contains the thermal conductance and capacitance values of the materials used in modeling the stack. In addition, it contains the pumping network power values, which we calculate based on a centrifugal pump [12] that feeds a data center containing 60 3D stacks. A valve controls the flow injected to each stack [13]. The channel pitch is changed from $100\mu m$ as in prior work [1], to $150\mu m$ to account for the Through-silicon vias. Hence, the flow rate that could be applied to meet the pressure gradient requirements inside the stack [2] is changed. In our experiments, we compare air-cooled and liquid-cooled 2- and 4-tier 3D MPSoCs.

In our evaluation, we define the thermal threshold $T_{th} = 85^oC$, which is the highest safe temperature of any element in the stack. Thus, a policy avoids hot-spots in the system if the maximum observed temperature using this policy is $T \leq T_{th}$.

We implement various thermal management techniques to evaluate the thermal and energy efficiency of the proposed design-time/run-time thermal management policy (DTRT). Dynamic load balancing (LB) balances the workload by moving threads from a core's queue to another if the difference in queue lengths is over a threshold. Temperature-triggered DVFS (AC_TDVFS_LB) adjusts the VF settings of a core when the core's temperature exceed T_{th}. In our implementation, as long as the temperature is above the threshold and there is a lower setting, we scale down the VF value at every scaling interval. When the temperature falls below another threshold value (82^oC), we scale up the VF values.

We experiment with both air-cooled (AC) and liquid-cooled (LC) systems for comparison purposes. In LC_LB, we apply the maximum flow rate (0.0323 l/min per cavity), while the jobs are scheduled with LB. Thermal impact of all the policies on the 2- and 4-tier 3D MPSoCs is shown in Fig. 3. This figure compares the peak and average temperatures for the average case across all the workloads and for the benchmark with maximum utilization rate. In air-cooled systems, AC_DVFS_LB

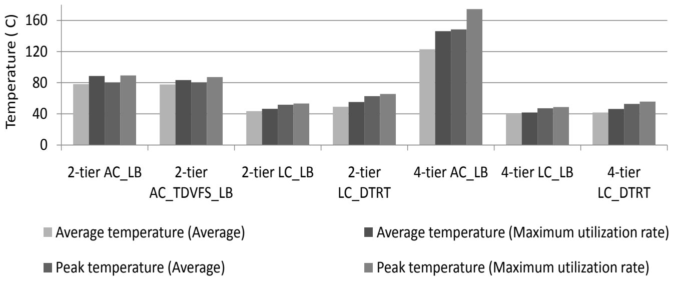


Fig. 3. Peak and average temperatures we observe for all policies, both for the average case across all workloads and for maximum utilization rate.

reduces the temperature in the 2-tier stack just to the thermal threshold value (T_{th}). The peak temperature with LB and AC_DVFS_LB are 87^oC and 85^oC, respectively. However, in the 4-tier stack, due to increased stacking and limited cooling capabilities, the maximum temperature is much higher than 110^oC and reaches up to 178^oC, leaving little opportunity for any thermal management technique to successfully mitigate the hot-spots without severely degrading the performance.

On the contrary, the integration of liquid cooling removes all the hot-spots in tested 3D MPSoCs by reducing the temperature below the threshold, due to its ability of inter-tier heat removal. LC_LB reduces the 2-tier 3D MPSoC peak temperature to 56^oC, whereas DTRT pushes the system into a higher peak of 68^oC, but still avoids any hot-spots. Moreover, the system temperature of the 4-tier 3D MPSoC is maintained at an even lower value than the 2-tier 3D MPSoC in both techniques, due to the increased number of inter-tier cavities.

Fig. 4 shows the total energy consumed when running the various policies on the 2-tier and 4-tier 3D MPSoCs for the average workload. Energy consumption values are normalized to the 2-tier AC_LB values. The proposed management policy achieves major reduction in both the coolant and the overall system energy consumption. DTRT reduces the 2- and 4-tier 3D MPSoCs energy by 14% and 18% on average, and cooling energy by 50% and 52% on average, respectively, in comparison to LC_LB. The reason DTRT outperforms all other techniques in energy savings is due to the joint control of flow rate and DVFS at run-time based on each core's thermal and utilization values. The proposed controller achieves up to 67% and 30% savings in cooling energy and overall system energy, respectively.

For our multicore 3D MPSoCs, we compute throughput as the performance metric. Throughput is the number of threads completed per given time. As we run the same workloads in all experiments, when a policy delays execution of threads, the resulting throughput drops. We notice that liquid cooling-based systems (LC_LB and DTRT) do not suffer from any performance degradation, even though our management policy uses DVFS. Since we apply DVFS based on core utilization (in DTRT), the performance degradation results do not exceed 0.01%, which is negligible in comparison to the degradation observed using AC_TDVFS_LB in air-cooling (up to 45% degradation).

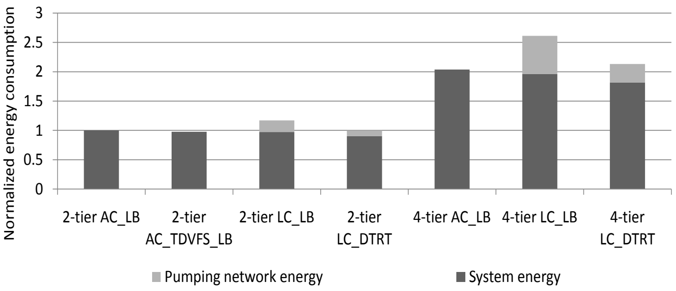


Fig. 4. Energy consumption in the whole system (chip and cooling network) for average case across all workloads. Note that air cooling also includes fan power consumption overhead, which is not included in the figure for AC-based policies.

IV. Conclusion

In this paper we have proposed a design-time/run-time thermal management policy of 3D MPSoCs with active cooling. We perform a design-time thermal response analysis of 3D MPSoC using varying liquid flow rate and DVFS. Based on this analysis, we extract a set of run-time management rules to minimize system energy consumption while preventing thermal hot-spots. Our experimental results illustrate that our management policy maintains the temperature below the thermal threshold, while reducing cooling energy by 50% and achieving overall energy savings by 18% on average with respect to setting the highest coolant flow rate to match the worst-case temperature.

References

[1] T. Brunschwiler et al. Direct Liquid-Jet Impingement Cooling with Micron-Sized Nozzle Array and Distributed Return Architecture. In *ITHERM*, 2006.

[2] T. Brunschwiler et al. Interlayer Cooling Potential in Vertically Integrated Packages. *Microsyst. Technol.*, 15(1):57 – 74, 2009.

[3] A. K. Coskun et al. Dynamic Thermal Management in 3D Multicore Architectures. In *DATE*, 2009.

[4] A. K. Coskun et al. Modeling and Dynamic Management of 3D Multicore Systems with Liquid Cooling. In *VLSI-SoC'09*, 2009.

[5] A. K. Coskun et al. Energy-Efficient Variable-Flow Liquid Cooling in 3D Stacked Architectures. In *DATE*, 2010.

[6] A. K. Coskun, T. Simunic Rosing, and K. Gross. Utilizing Predictors for Efficient Thermal Management in Multiprocessor SoCs. *TCAD*, 28(10):1503–1516, 2009.

[7] J. Donald and M. Martonosi. Techniques for Multicore Thermal Management: Classification and New Exploration. In *ISCA*, 2006.

[8] W.-L. Hung et al. Interconnect and Thermal-Aware Floorplanning for 3D Microprocessors. In *ISQED*, 2006.

[9] A. Leon et al. A Power-Efficient High-Throughput 32-Thread SPARC Processor. *ISSCC*, 42(1):7 – 16, 2007.

[10] K. Puttaswamy and G. H. Loh. Thermal Analysis of a 3D Die-Stacked High-Performance Microprocessor. In *GLSVLSI'06*, 2006.

[11] M. M. Sabry, A. K. Coskun, and D. Atienza. Fuzzy Control for Enforcing Energy Efficiency in High-Performance 3D Systems. In *ICCAD*, 2010.

[12] WILO MHIE Centrifugal Pump. http://www.wilo.com/cps/rde/xchg/en/layout.xsl/3707.htm.

[13] Festo Electric Automation Technology. http://www.festo.com.

[14] A. Sridhar et al. 3D-ICE: Fast Compact Transient Thermal Modeling for 3D-ICs with Inter-Tier Liquid Cooling. In *ICCAD*, 2010.

[15] T. Takagi and M. Sugeno. Fuzzy Identification of Systems and Its Applications to Modeling and Control. *IEEE Transactions on Systems, Man, and Cybernetics*, 15(1):116 – 132, 1985.

[16] X. Zhou et al. Thermal Management for 3D Processors via Task Scheduling. In *ICPP*, 2008.

[17] C. Zhu et al. Three-Dimensional Chip-Multiprocessor Run-Time Thermal Management. *IEEE Transactions on CAD*, 27(8):1479–1492, August 2008.