Leakage Power Characterization and Minimization in 3D Stacked Multi-core Chips with Microfluidic Cooling

He Xiao*, Zhimin Wan+, Sudhakar Yalamanchili*, Yogendra Joshi*+
*School of Electrical and Computer Engineering
+The George W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology, Atlanta, GA, USA

Abstract

The needs of multiple-functionality and low cost have driven the development of high-density electronic packages. However, the greater package density results in higher power density per unit volume of the package, which creates challenges for thermal management. Microfluidic cooling can potentially achieve superior thermal performance with surface area enhancements such as pin fins and could be a viable solution for many applications to the increasing power density in electronic packages.

In this paper, we report on investigations of the impact of the microfluidic cooling technology on the system level performance of multicore architectures stacked in a 3D package. Specifically, we characterize the impact on leakage power dissipation over different pin fin configurations and its impact on overall system energy efficiency. We do so with a cycle-level application-driven full system simulation framework. The framework executes application binaries and operating system code and models coupled interactions among the i) application & operating system code, ii) resulting thermal field, iii) leakage power, and v) microfluidic cooling. This provides the unique ability to assess the impact of microfluidics on computing system level metrics experienced by the applications such as energy per instruction. We illustrate and quantify improvements in energy efficiency of the applications, as well as increase in throughput due to microfluidic cooling.

Keywords

3D stacked IC, Micro Pin Fin, Leakage Characterization, Power Efficiency

1. Introduction

The technology of 3D stacked ICs provides high integration density, vertically integrating multiple die in a compact package, leading to shorter die-to-die connectivity and substantial increases in inter-die communication bandwidth. This is achieved at the expense of increased power densities and heat flux, as the integration level increases and circuit components get closer [1].

Liquid cooling with surface enhancements such as pin fin has much better thermal dissipation capability compared to conventional air cooling. Previous work involves optimizing pin fin configurations to achieve low thermal resistance under a static power density and a given IC floor plan [2]. Our current analysis instead focuses on runtime power and thermal analysis based on i) 3D-ICE [3], a thermal simulator for 3D interlayer cooling emulation, integrated with ii) Manifold, a full system functional, power, and timing simulator [4]. Pin fin optimizations are performed offline using our compact thermal model [5]. Application workloads create power behaviors that in turn create the thermal fields that are modulated by the microfluidics. The integrated simulator models the coupling between temperature and leakage power. We further apply this framework across multiple pin fin configurations to investigate the impact of pin fin parameters.

With the end of Dennard scaling, leakage power will continue to be a major concern in sub-32 nm technology. For example, studies shows that for a 2-way cache in 16nm with size of 2MB, an estimation of leakage power could be up to 8W at 60°C [6]. The increased power densities of 3D packages exacerbate the temperature-leakage power coupling increasing leakage power, as well as the rate of temperature increase. Together, both phenomena reduce energy efficiency. This paper focuses on characterization of leakage power and understanding how pin fin designs can be optimized to improve system level efficiency. We analyzed the system runtime performance and physical states under various benchmarks with different micro pin fin configurations. Thus we hope to contribute to the generation of guidelines for the design of microfluidic solutions for 3D IC chips.

2. System Architecture

The impact of pin fin configurations on 3D stacked ICs is studied for a 16-core homogeneous microarchitecture in 16nm technology. The package consists of two ICs stacked as shown in Figure 1 – a processor tier and an L2 cache tier. The processor layer is at the bottom closer to the package primarily due to power delivery considerations. The pin fin array is constructed in between the layers to dissipate heat from both the processor and L2 cache, which are electrically connected by Through Silicon Vias (TSVs). The electrical connection between both layers is configured as an additional layer of back end of line (BEOL) - SiO_2 & Metal layer in both the two tiers.

Each tier in Figure 1consists of three layers: SiO_2 & Metal layer, active layer, silicon base layer. The SiO_2 & Metal layer is used for bonding, and routing. SiO_2 is deposited on the chip by Plasma Enhanced Chemical Vapor Deposition (PECVD) while the metal layer is obtained by lift-off process. In order to simplify the simulation, the metal is not considered in the calculation. Most of the heat is generated in the active layer. The thickness of the active layer is neglected. Circular pin fin enhanced microgap was fabricated on the back of the chip by Deep Reactive-Ion Etching (DRIE) which enables high aspect ratio etching. The signal TSVs could be embedded in the pin fin for communication between the processor and L2. In this study, the effect on the signal TSVs is not considered. Fluid flows across the pin fins and removes the heat. DI-water is used as the coolant due to its good thermal performance for single phase cooling. Natural convection is assumed at the top of the chip stack.

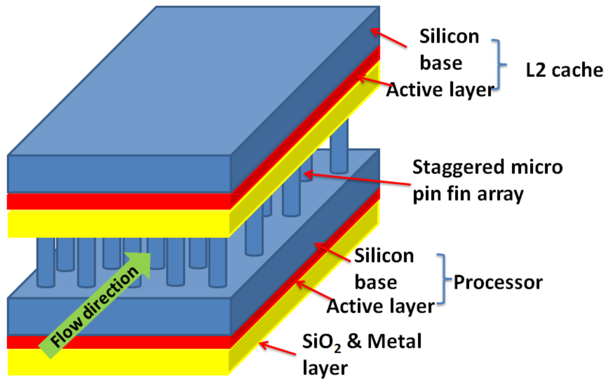
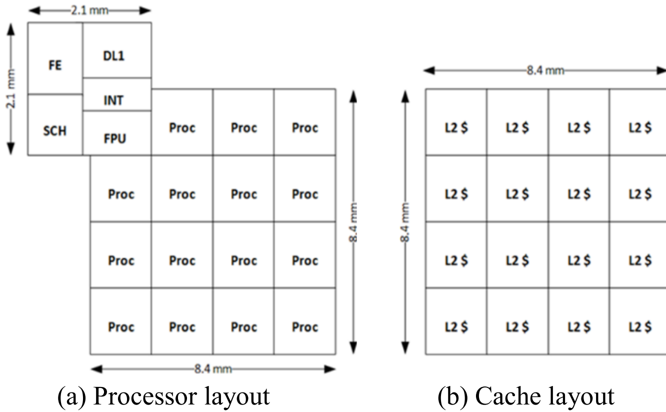


Figure 1: The geometric model of 3D stacked ICs with microfluidic cooling

In the processor layer, the processors resemble the Intel Nehalem core, a typical Out-of-Order microarchitecture including: FE (pipeline frontend and L1 instruction cache), SCH (Out-of-Order scheduler), INT (integer unit), FPU (floating-point unit) and DL1 (L1 data cache). Each core has a private L1 data cache of 128KB and a shared L2 cache.

The L2 cache layer consists of 16 SRAM banks each with size of 2MB. The cache layer connects to the DRAM memory controller via a 2D torus interconnection. Placed above the L2 cache is a DRAM stack – yet in this paper, we only evaluate the impact of pin fin configurations on the power and energy efficiency of the processor and cache. The floor plans of the 2-layers in 16nm technology are shown in Figure 2, both of which have a dimension of 8.4mm × 8.4mm.



(a) Processor layout (b) Cache layout

Figure 2: Floor plan of the 2-tier stacked IC in 16nm

Table 1 lists the parameters of the processors and caches.

Processor Configuration		
Issue Width	4	
Execution Width	5 (3 INT ports, 2 FP port)	
ROB Size	128	
RS Size	36	
Cache Configuration		
	Size	Associativity
IL1(per core)	16 KB	2
DL1 Data (per core)	32 KB	4
Shared Coherent L2	16 MB	16

Table 1: Architecture configuration of the multicore system

3. Optimization of Pin Fin Geometries

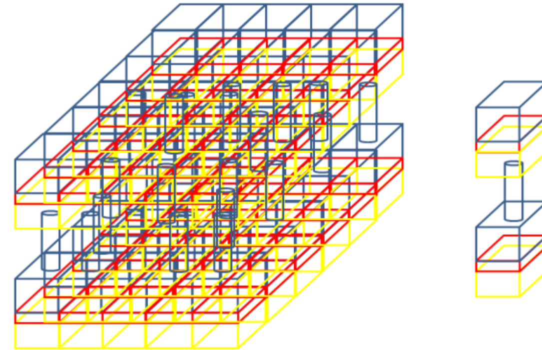
While the simulations are completely integrated, the optimizations of the pin fin geometries are carried out offline as described in this section. We performed the optimization using a compact thermal model, built specifically for a 3D stack with inter-tier microfluidic cooling. Table 2 shows the materials dimensions and properties in the model.

Materials	SiO₂&Metal	Si base	Pin fin
Length(mm)	8.4	8.4	initial*
Width (mm)	8.4	8.4	
Thickness (μm)	10	100	
Thermal conductivity (W/(m K))	1.4	149	149

** Initial pin diameter is 100 μm, pin height and pitch are 200 μm.*

Table 2: Material dimensions and properties

To convert the geometric model to a compact thermal model, a discretization of the geometric model [6] into interconnected control volumes was conducted (Figure 3(a)). Our compact model is based on finite volume energy balance on a unit control volume around single pin. Separate control volumes are considered based on the solid and fluid part.



(a) Discretization of pin fin model (b) Single control volume

Figure 3: The structure and modeling of micro pin fin

After the control volumes are defined, energy balance analyses are conducted for each control volume. Figure 4 identifies the energy coupling between tiers in the vertical direction. The red arrow shows the energy flow in the control volume. Each tier has in plane heat conduction from 4 directions. Moreover, it has heat conduction to the other tier through the pin, heat convection to the fluid and to the ambient. A uniform temperature for each active layer in one control volume is assumed for simplification.

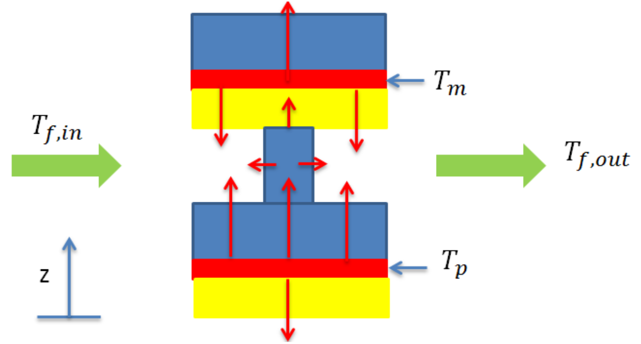


Figure 4: Energy analysis of single control volume

The energy equation for the solid domain is:

$$Solid: \dot{q_{gen}} + \dot{q_{cond}} + \dot{q_{conv}} = 0$$

$q_{\dot{gen}}$ is the energy generation term which is obtained from the power map. $q_{\dot{cond}}$ is the heat conduction from neighboring control volumes. $q_{\dot{conv}}$ is the heat transferred by convection. Because we have processor and L2 tiers, two energy equations should be built for processor and L2 respectively.

For the fluid flow, one direction flow is assumed [7] and axial conduction inside the fluid is neglected. The energy balance equation for the fluid is:

$$Fluid: \dot{m}C_p(T_{f,in} - T_{f,out}) + q_{\dot{conv}} = 0$$

A system of energy equations could be obtained by applying energy analysis on every control volumes. In order to obtain the temperature distribution, we need to solve the above energy equations simultaneously. So all the convection and conduction terms should be expressed as functions of fluid, processor and L2 temperatures first. Here only the conduction between the processor and L2, and the convection between the fluid and solid are discussed due to the special nature of the energy flow [8].

The temperatures variations within the silicon base and silicon dioxide layer are assumed linear,

$$T_{base} = k_1 \cdot z + a$$

$$T_o = k_2 \cdot z + b$$

For the fin, $T_{fin} = T_{f,in} + C_1 e^{mz} + C_2 e^{-mz}$

There are six constants k_1, k_2, a, b, c_1, c_2 in these equations requiring six boundary conditions. Conditions are the constant temperature boundary conditions are used at the two active layers. Another four boundary conditions come from the heat flux and temperature continuity at the interface of silicon base, fin and silicon dioxide. So the six constants could be solved and expressed as functions of the temperatures of fluid, processor and L2. Then the convection and conduction heat transfer could be expressed as a function of temperature. And a system of energy equations including temperature at every control volume could be built. In the present study, Tridiagonal matrix algorithm (TDMA) is used to solve these equations. So the temperature field could be determined [7].

Compared to the existing compact thermal modeling approaches such as 3D ICE and hotspot, our compact model is compatible with any number of layers of 3D stack with pin fin enhanced microgap. The hydraulic and thermal characteristics with different pin fin configurations including the pin diameter, height, longitudinal and transversal spacing, and various flow rates can be studied.

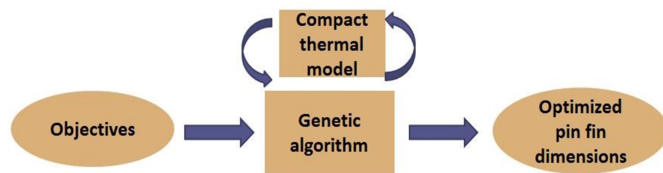


Figure 5: The optimization process flow of pin fin structure

The compact thermal model is then embedded into the optimization tool in Matlab. Genetic algorithm is used for optimization. The optimization starts with the setting of the objectives and constrains (Figure 5). The objective is to find the pin fin dimensions, which produce minimum junction temperature under a certain power map and a fixed pumping power (0.03 W in this study). The dimensional constraints are that the range of pin fin diameters is from 100 μm to 200 μm, the range of ratio of longitudinal spacing and transversal spacing to pin diameter is 1.5 ~ 2.25, and the range of ratio of pin height to pin diameter is 1 ~ 3. The thickness of the wafer is usually about 500 μm. So the pin height is also constrained to be less than 400 μm. During the optimization, the genetic algorithm would first generate randomly individual dimensions to input to the compact model as possible solution to the optimization, allowing the entire range of possible solutions. Then the temperature field and junction temperature are determined for each solution. Individual solutions are then selected through a fitness-based process, where fitter solutions are typically more likely to be selected as parent solutions. Child solutions are produced using the method of crossover and mutation of the parent solution. The new temperature field and junction temperature are determined for each new solution with maximum generation set as 100. The stopping criterion is that the function tolerance is less than 1e-6.

4. Simulation Model

The Manifold-based simulation framework is depicted in Figure 6. The 16-core timing model is constructed based on a cycle-based simulator driven by application and operating system binaries. Cycle level operations of a core and cache hierarchy generate power values (during simulation and not for offline analysis) that drive an integrated thermal model, which calculates the thermal fields and updates the leakage power as well. Time varying workloads exercise this feedback loop while cooling modulates this interaction.

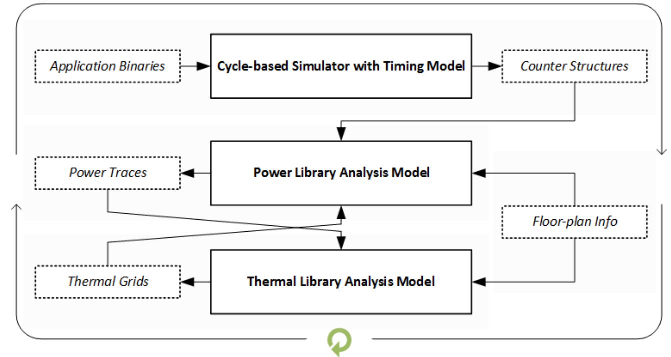


Figure 6: Simulation framework of the 16-core system

The core component model used in Manifold that executes at around 0.5~2 MIPS, and provides sufficient details for power/thermal analysis. The cache model of the system employs the MCP-cache component, which implements in a manner of Manager-Client Paring coherence framework [4]. The MCP-cache implements a directory-based coherence of across the L1 and L2 data caches. The interconnection network utilizes Manifold's IRIS model consisting of interfaces and routers. In our simulation, the routers are connected in a two-dimensional torus, while the network interfaces connects to the L2 cache banks and memory controllers. DRAM is modeled as a cycle level memory controller and multibank DRAM array. There are 16 memory controllers corresponding to 16 DRAM memory partitions fashioned after MICRON's 3D stacked Hybrid Memory Cube [9]. We assign each core with an independent DRAM component to construct the homogenous multicore system from bottom up shown in Figure 7.

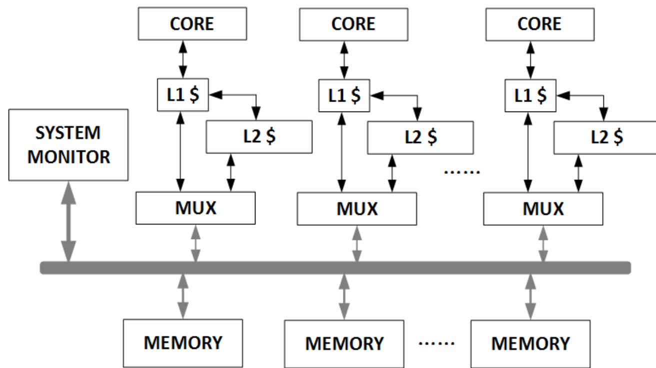


Figure 7: The structure of the 16-core simulator

To perform runtime power and thermal analysis, a system monitor is built inside the simulator. The monitor coordinates the execution of the timing model and the invocation of the physical models across sampling windows. At the end of each sampling period, it will perform the following: i) synchronize all the simulation components to the same time stamp; ii) sample pipeline and cache activity and provide to the power analysis model McPAT[10]; iii) pass the results of the power model to the thermal library to calculate the thermal grids; iv) update leakage power with newly computed temperature.

The physical model is a 2-tier stacked IC package with microfluidic cooling described in previous section. The pin fin is placed between the processor and cache to dissipate heat. The ambient air on top of the package is modeled as a natural convection. The detailed heat sink parameters in Table 3.

Convection Heat Sink	
Heat Transfer Coefficient	1.2e-11 W/μm^2K
Ambient Temperature	300 K
Micro Pin Fin	
Pin Material	Silicon
Pin Distribution	Staggered
Coolant Volumetric Heat Capacity	4.17e-12 J/μm^3K
Coolant Incoming Temperature	300 K

Table 3: Heat sink parameters in 3D-ICE simulation

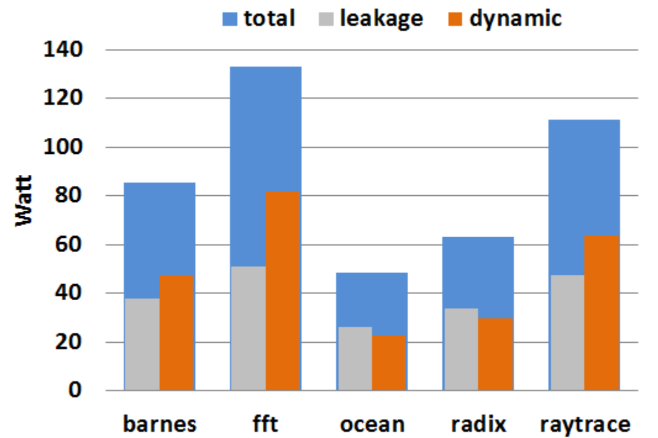
5. Results Analysis

To characterize the leakage power under various pin fin configurations, we picked several applications from the SPLASH-2 benchmark – written for coherent shared memory multicore processors. Each application is at first fast-forwarded for 100M cycles to warm up the processor state and enter the applications' region of interest, an area that reflects important program characteristics. The simulation model then proceeds for 500 ms of real time for runtime analysis. We also made a comparison between a baseline and an optimized pin fin configuration using the method described in section 3. The parameters of both configurations are listed in Table 4.

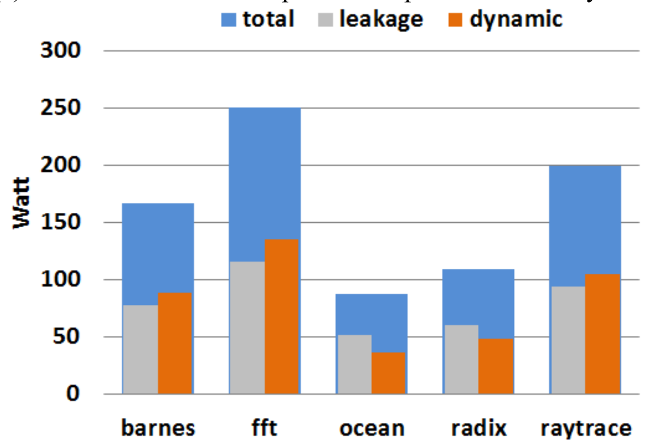
	DP (um)	PS (um)	HP (um)
baseline	100	200	200
optimized	180	320	400

DP: diameter, PS: pitch spacing, HP: height

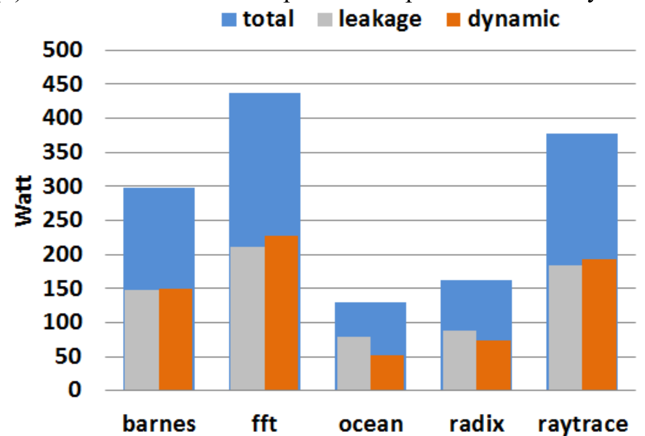
Table 4: Configuration parameters of pin fin structure



(a) The characterization of power dissipation in 3GHz system



(b) The characterization of power dissipation in 4GHz system



(c) The characterization of power dissipation in 5GHz system

Figure 8: Power comparison of SPLASH-2 benchmark

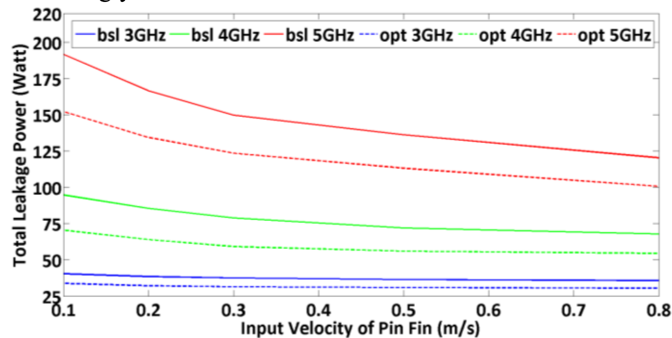
We first evaluated the power consumption among all selected test cases under a fixed pin fin configuration (baseline, input velocity is 0.3 m/s) and characterize power consumption with respect to different frequency scales. Figure 8 indicates that both the dynamic and leakage power follow a super linear relationship with the system frequency, as to support higher clock, the system supply voltage needs to be scaled up accordingly, and the leakage power takes larger proportion of total power consumption with a higher system frequency. Notice that both ocean and radix are memory-bounded, in which leakage power takes up over 50% of total power consumption.

Specifically, we investigated the results from two important applications barnes and ocean-c in detail. Barnes is a typical computational-bounded application as it has low cache miss rate; ocean-c is a memory-bounded application due to its relatively high miss rate and large remote traffic. The time and space requirements [11] are listed in Table 5.

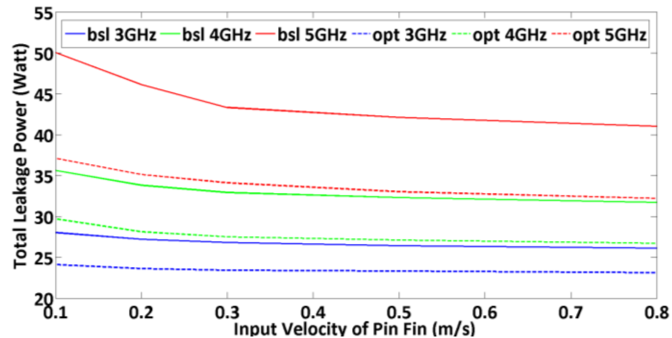
Application	Time	Space
barnes	NlogN	N
ocean-c	N^3	N^2

Table 5: Time and space requirements of two applications

The simulation results in Figure 9 indicate: i) the leakage power decreases with increased input velocity due to improved heat transfer capability, ii) system running at a faster clock frequency benefits more when increasing the fluid flow rate, as system at faster frequency tends to generate higher power, iii) optimized pin fin configuration can provide significant improvement in heat transfer capability; for example, the leakage power of a 5GHz system is reduced by over 22% in all test cases at 0.1 m/s Darcy velocity between a baseline and optimized configuration, iv) when the fluid flow velocity is small (i.e. below 0.4 m/s), the computational bounded application has more leakage power reduction than memory bounded applications, as it tends to have a higher instructions per cycle (IPC) and generates more heat accordingly.



(a) Leakage power of "barnes" in terms of coolant velocity



(b) Leakage power of "ocean-c" in terms of coolant velocity

Figure 9: Leakage power reduction as a function of the input fluid velocity

It is apparent that microfluidic cooling will enable the processor to execute at a higher frequency, compared to one with a conventional heat sink. The system can thus realize higher throughput. Figure 10 gives the comparison between achieved system throughput and throughput expected from simple clock scaling. For both barnes and ocean-c, throughput does not increase in proportion to clock rate increase, due to the fact that the memory system is on a different (constant) clock. The normalized system throughput of barnes is higher than ocean-c since barnes has a lower cache miss rate and thus fewer interactions with the slow system memory. Overall, the higher clock rates made feasible by microfluidics still enables overall 20% and 40% improvement in 4GHz and in 5GHz respectively. This does not necessarily mean that the energy efficiency is improved as we describe next.

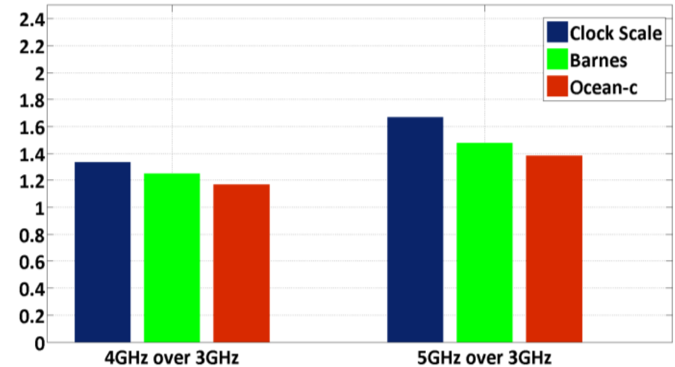


Figure 10: System throughput in terms of core frequency

In addition, we determine the system energy efficiency in terms of energy per instruction (EPI), which tracks the average energy used to execute a single instruction. Figure 11 illustrates the energy efficiency of the system with different frequencies under the optimized pin fin configuration in table 4. The input fluid velocity is set to 0.8 m/s.

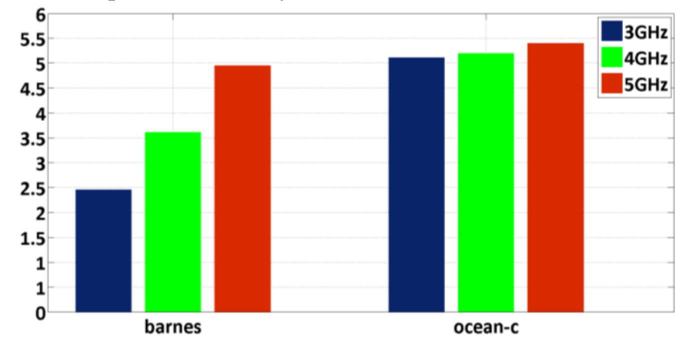


Figure 11: System EPI in terms of core frequency

The EPI of barnes from 3GHz to 5GHz keeps increasing, as barnes operates at a relatively high temperature (above 350K) due to high instruction per cycle (IPC). The power dissipation grows faster than the reduction in execution time because of the quadratic relationship between leakage power and temperature. On the other hand, ocean-c works around 330K, and the increase of leakage is approximately linear. The EPI of ocean-c remains approximately as a constant, because the system speedup from clock scaling compensates for the increase in system leakage power.

6. Conclusions

Our results establish that an optimized pin fin structure with appropriate coolant velocity will enable the system to operate with a higher throughput and improved energy efficiency. To summarize, we studied the impact of pin fin configurations on the leakage power in 3D stacked ICs.

- The saturation point of coolant velocity is determined by the pin fin geometry, system frequency and runtime application. A typical saturation value for system within 5GHz is 0.3 m/s.
- Increasing the input velocity of the fluid, computational-bounded applications benefit more from leakage reduction than memory-bounded applications, as the formal tends to generate more heat. Barnes saves 37.2% and 33.9% leakage power respectively for the baseline and optimized system running at 5GHz with input velocity from 0.1 m/s to 0.8ms, compared to ocean-c 18.3% and 13.2%.
- The parameters of pin fin configuration play a critical role in reducing system leakage power, and the performance of the optimized pin fin structure exceeds the baseline under all circumstances. In our study, an optimized pin fin can save up to 20% leakage dissipation with a same pumping power.
- The power efficiency of 3D stacked architecture also tends to increase with system frequency for applications running at low temperatures (i.e. below 335K), as the performance improvement will compensate the loss of leakage power.

To the best of our knowledge, this is the first model and analysis that integrates into a single simulation model i) application binaries, ii) operating system binaries, iii) cycle-level multicore architecture timing, iv) power and energy models and v) thermal models. The self-contained simulation framework enables us to explore the impact of microfluidics on computing system level metrics experienced by the applications and evaluate microarchitecture level metrics such as energy per instruction over various physical configurations.

Acknowledgments

The authors gratefully acknowledge the support of Sandia National Laboratories and the National Science Foundation under grant CNS-855110.

References

1. S. P. Gurrum, S. K. Suman, Y. K. Joshi, and A. G. Fedorov, "Thermal issues in next-generation integrated circuits," IEEE Transactions on Device and Materials Reliability, Vol. 4(4), pp. 709-714, 2004.
2. Z. M. Wan, H. Xiao, Y. Joshi, and S. Yalamanchili, " Co-design of multicore architectures and microfluidic cooling for 3D stacked ICs," 19th International Workshop on thermal Investigations of ICs and Systems, Berlin, Germany, pp. 237-242, 2013.
3. M. Sridhar, T. Brunschwiler, A. Viricenzi, D. Atienza, and M. Ruggiero, 3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling, IEEE/ACM International Conference on Computer-Aided Design, Digest of Technical Papers, pp. 463-470, 2010.
4. J. Wang, J. Beu, R. Behda, T. Conte, Z. Dong, C. Kersey, M. Rasquinha, G. Riley, W. Song, H. Xiao, P. Xu, and S. Yalamanchili, "Manifold: A Parallel Simulation Framework for Multicore Systems," IEEE International Symposium on Performance Evaluation of Systems and Software (ISPASS), 2014.
5. Z. M. Wan, Y. J. Kim, Y. Joshi, "Compact modelling of 3D stacked die inter-tier microfluidic cooling under non-uniform heat flux," Proc. ASME-IMECE 2012, Houston, TX, USA, 2012.
6. P. Zajac, M. Janicki, M. Szermer, C. Maj, P. Pietrzak, A. Napieralski, "Cache Leakage Power Estimation Using Architecture Model for 32 nm and 16 nm Technology nodes," Proc. SEMI-THERM 28, San Jose, CA, USA, 2012.
7. S. V. Patankar, Numerical Heat Transfer and Fluid Flow, McGraw-Hill, New York, 1980.
8. F. P. Incropera and D. P. Dewitt, Fundamentals of Heat and Mass Transfer, 5th edition, John Wiley & Sons, Inc., 2006.
9. Joe Jeddeloh, Brent Keeth, "Hybrid Memory Cube New DRAM Architecture Increases Desnsity and Performance, " Symposium on VLSI Technology (VLSIT 2012), pp. 87-88, 2012.
10. S. Li, J. Ahn, R. Strong, J. Brockman, D. tullsen, and N. Jouppi, "Mcpat: An integrated power, area, and timing modeling framework for multi-core and manycore architectures", Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture (Micro-42), pp. 469–480, 2009.
11. C. Bienia, S. Kumar and K. Li, "PARSEC vs. SPLASH-2: A Quantitative Comparison of Two Multithreaded Benchmark Suites on Chip Multiprocessors", IEEE International Symposium on Workload Characterization (IISWC 2008), pp. 47-56, 2008.