Thermal and Power-Aware Run-time Performance Management of 3D MPSoCs with Integrated Flow Cell Arrays

Halima Najibi
Embedded Systems Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Alexandre Levisse
Embedded Systems Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Giovanni Ansaloni
Embedded Systems Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Marina Zapater
ReDS Institute, University of Applied Sciences Western Switzerland

David Atienza
Embedded Systems Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

Flow Cell Arrays (FCA) technology employs microchannels filled with an electrolytic fluid to concurrently provide cooling and power generation to integrated circuits (ICs). This solution is particularly appealing for Three-Dimensional Multi-Processor Systems-on-Chip (3D MPSoCs) realized in deeply scaled technologies, as their extreme power densities result in significant thermal and voltage supply challenges. FCAs provide them with extra power to boost performance. However, the dual effects of FCAs (cooling and power supply) have conflicting trends leading to a complex interplay between temperature, voltage stability, and performance. In this paper, we explore this trade-off by introducing a novel methodology that controls the operating frequency of computing components and the electrolytic coolant flow rate at run-time. Our strategy enables tangible performance gains while abiding by timing, voltage drop, and temperature constraints. We showcase its benefits by targeting a 4-layer 3D MPSoC, achieving up to 24% increase in the operating frequencies and resulting in application speedups of up to 17%, while reducing the costs related to FCA liquid pumping energy.

CCS CONCEPTS

• **Hardware → 3D integrated circuits**; **Power networks**; **Temperature control**.

KEYWORDS

Flow Cell Arrays, 3D MPSoCs, Performance Management

ACM Reference Format:
Halima Najibi, Alexandre Levisse, Giovanni Ansaloni, Marina Zapater, and David Atienza. 2022. Thermal and Power-Aware Run-time Performance Management of 3D MPSoCs with Integrated Flow Cell Arrays. In *Proceedings of the Great Lakes Symposium on VLSI 2022 (GLSVLSI '22), June 6–8, 2022, Irvine, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3526241.3530309

GLSVLSI '22, June 6–8, 2022, Irvine, CA, USA

ACM ISBN 978-1-4503-9322-5/22/06…$15.00
https://doi.org/10.1145/3526241.3530309

1 INTRODUCTION

State-of-the-art high-performance applications such as Artificial Intelligence and Big Data prompt the need for complex heterogeneous platforms with High-Performance Computing (HPC) capabilities. In this context, 3D MPSoCs use vertical interconnect lines called Through-Silicon-Vias (TSVs) to stack multiple dies in a single chip. They achieve high-density computing and provide ultra-wide communication bandwidths [1], alleviating the gap between processing and data access speed and enhancing the overall system efficiency. Nonetheless, IC designers struggle to achieve high-performance 3D MPSoCs due to their critical *heat generation* [2] and highly challenging *power supply and distribution* requirements [3], particularly for deeply-scaled CMOS technologies.

FCAs promise to address these 3D MPSoC challenges effectively. They consist of microchannels where an electrolytic liquid flows, etched in the silicon substrate of dies. By concurrently providing inter-tier liquid cooling and on-chip electrochemical power generation [4], FCAs enable to reduce the 3D MPSoC temperature and partly recover the voltage drops in the Power Delivery Network (PDN) [5]. Hence, they help prevent timing delays that can cause performance degradation and system failures [3]. In addition, FCAs can boost the performance of 3D MPSoCs as they provide an extra power budget enabled by both power generation and temperature-dependent leakage reduction. However, these two FCA capabilities expose a new trade-off: On the one hand, higher flow rates enhance heat absorption and considerably decrease 3D MPSoCs temperature and leakage. On the other hand, lower flow rates accelerate the power-generating reactions inside the channels. Hence depending on the architecture and level of utilization of each 3D MPSoC, FCA flow rate settings can affect the overall power performance.

In this context, this paper introduces a novel run-time management strategy to harness the potential of FCAs while involving the previous inter-dependent thermal and electrical considerations. The proposed approach targets increasing 3D MPSoCs computing performance by co-configuring the electrolytic flow rate and the operating frequencies of dies. It is implemented in two phases: The *offline optimization methodology* searches for the FCA flow rate configurations that enable the maximal operational frequencies for various 3D MPSoC utilization scenarios. Then, the *online controller* periodically applies the pre-computed flow rate and frequency settings during run-time, according to workload power requirements.

In summary, the contributions of the paper are as follows:

- We introduce a novel strategy to co-optimize the FCA flow rates and operating frequencies of high-performance 3D MPSoCs. The algorithm considers the trade-offs between FCA leakage reduction and on-chip power generation capabilities. Fine-grain thermal and electrical simulations ensure that temperature and timing constraints are met while achieving the maximal performance of dies.
- Targeting a 4-layer 3D MPSoC system, we showcase that our performance management strategy enables operating the Central Processing Unit (CPU) and Graphics Processing Unit (GPU) up to 24% faster than the nominal frequency when using FCAs. It also enables up to 19% faster operation in the case of regular inter-tier liquid cooling.
- We show that our online controller allows to speed up the execution of multiple compute-intensive benchmarks on the CPU and the GPU by up to 17% and 15%, respectively. Compared to a fixed flow rate strategy, the workload speedups are achieved while reducing the FCA liquid pumping energy by up to 43%.

2 BACKGROUND ON 3D MPSOCS

2.1 3D MPSoCs Thermal and Power Challenges

3D MPSoCs face critical thermal and power challenges limiting their adoption in the VLSI industry. In particular, 3D MPSoCs generate large amounts of heat as the power density escalates with the number of stacked dies [2]. Traditional fan-based cooling struggles to dissipate the generated heat due to the poor thermal conductivity of silicon and bonding materials. Consequently, leakage power increases exponentially, affecting 3D MPSoC power delivery. In particular, PDN TSVs and metal lines must supply very high currents that cause voltage drops throughout the power grid, potentially inducing circuit timing failures [3]. The heat extraction, and therefore leakage reduction in 3D ICs, must be addressed to achieve functional 3D MPSoCs. In this regard, designers have proposed solutions such as thermal TSVs and specific glue materials to improve thermal dissipation [6], or high-performance direct liquid cooling techniques to extract excess heat [7]. However, these two approaches generally require significant area and costly materials. Furthermore, their efficacy drops with the number of stacked dies and the increasing power densities. In contrast, inter-tier liquid cooling is a more effective thermal management strategy regardless of the 3D MPSoC size, with no extra area cost.

2.2 3D MPSoCs Thermal Management with Inter-Tier Liquid Cooling

Inter-tier liquid cooling employs micro-channels etched in the silicon substrate of 3D MPSoC dies, through which a liquid is pumped, absorbing the generated heat [8]. In general, the cooling efficiency of the channels depends on the inlet liquid temperature and flow rate, which must be calibrated to meet the temperature constraints during system operation. Hence, existing run-time management strategies target the optimal utilization of cooling and power resources, according to 3D MPSoC usage conditions. For example, the authors in [9] and [10] combine flow rate adjustment, DVFS, and task scheduling to balance 3D MPSoC temperature while decreasing cooling and computational power. Similarly, authors in [11] place thermal sensors in strategic locations to provide temperature information to their variable-flow controller. These techniques enable 3D MPSoC temperature reduction but do not consider power performance optimization, which can be achieved using FCAs.

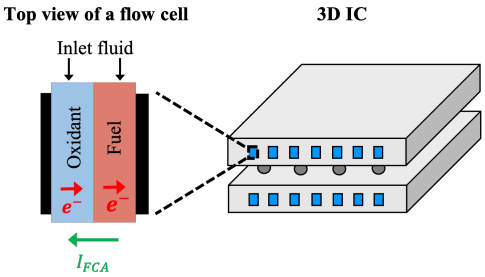

Figure 1: Flow Cell Array Technology

2.3 3D MPSoCs Thermal and Power Management with Flow Cell Arrays

FCA technology is a novel solution to both the power and thermal challenges of 3D MPSoCs. It extends inter-tier liquid cooling by providing on-chip power generation [5]. To this end, an electrolytic liquid flows inside the micro-channels used as FCAs, as illustrated in Figure 1. The electrochemical reaction rate increases with temperature, producing an electrical current to supply logic gates. The FCAs are connected to the 3D MPSoC power grid through DC-DC voltage regulators, ensuring that they operate at their optimal voltage (enabling the maximal power generation) regardless of transient load changes in the chip [12]. As shown in [13] and [12], FCAs can reduce temperature by 50°C compared to traditional fan-based cooling, and recover up to 20% of voltage drop in the 3D PDNs. Consequently, FCAs provide an opportunity to increase the load of dies while mitigating the performance losses related to voltage drop and temperature. This is enabled by both FCA cooling and power generation capabilities, which have opposite trends. According to the system architecture and usage conditions, they must be balanced to achieve the best power performance. In this context, we propose in Section 3 a strategy to co-optimize the flow rates of FCAs and operating frequencies of dies at run-time, speeding up the operation of high-performance 3D MPSoCs.

3 PERFORMANCE MANAGEMENT OF 3D MPSOCS WITH FCAS

As outlined in Section 2.3, the integration of FCAs in 3D MPSoC PDNs provides an opportunity to increase the power consumption without exceeding the temperature and voltage drop constraints. In this section, we propose a combined FCA flow rate and frequency control strategy to enhance the performance of 3D MPSoC computing dies. The *offline optimization* uses fine-grain thermal and electrical modeling to determine the lowest applicable FCA flow rates that enable the highest operating frequency boost of dies, depending on their utilization levels (Section 3.1). Then, the *online controller* selects the optimal FCA and frequency settings according to workload requirements from the pre-computed look-up-table (LUT), enabling a smooth control with minimal computation costs and delays (Section 3.2).

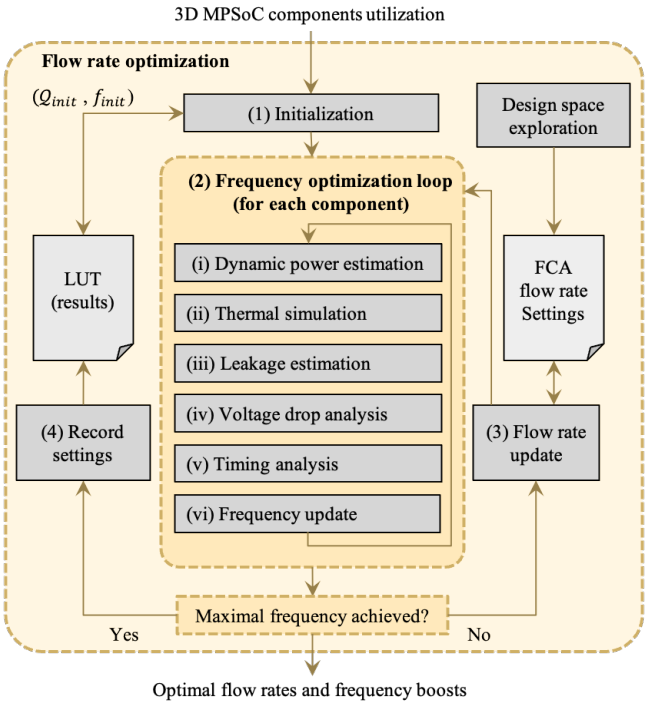

Figure 2: Frequency and Flow Rate Optimization Methodology

3.1 Offline Frequency and FCA Flow Rate Optimization Strategy

We introduce a 3D MPSoC optimization algorithm that receives the dies utilization levels as input, then evaluates the lowest applicable FCA flow rates that enable the maximal operating frequencies of cores. As the maximally usable frequencies of dies directly depend on their thermal characteristics and the amount of available power, they are influenced by the FCA cooling and power generation capacities. Hence, the flow rate inside the channels must first be set to determine the maximal frequency boost for each computing die. In this regard, the optimization algorithm is composed of two nested loops, as illustrated in Figure 2. The outer *flow rate optimization* loops over the possible FCA flow rate configurations and determines the optimal one based on its corresponding applicable operating frequencies for the different 3D MPSoC dies. During each iteration of the outer loop, the inner *frequency optimization loop* is performed for all the 3D MPSoC computing dies to determine the maximal applicable frequency boosts corresponding to the selected FCA flow rate settings. The algorithm outputs the optimal flow rate and frequency settings, among all possible FCA flow rate realizations.

3.1.1 Flow rate optimization. First, a design-space exploration determines the set of possible combinations of flow rate values for each 3D MPSoC layer $Q = [q_1 \dots q_K]$. The selected combinations $\{Q\}$ achieve different cooling performances while having the lowest total flow rate, to minimize the cooling cost. Then, the flow rate optimization loop searches for the optimal FCA flow rate combination Q_{opt} to apply to the 3D MPSoC for each utilization scenario $P = [\rho_1 \dots \rho_N]$. In particular, the FCA configuration with the lowest total flow rate is selected if it enables the highest frequency for all the computing dies. This ensures that the best 3D MPSoC performance is achieved with minimal cooling cost. In this regard, the following series of steps is performed for the whole stack, recursively until convergence for each utilization scenario:

(1) The algorithm dictates the initial FCA flow rates and frequencies to apply to the 3D MPSoC. For the first utilization scenario ($P = P_0$), the algorithm selects the lowest usable settings (i.e., Q_{min} and F_{min}). Then, for a subsequent scenarios P the initial settings are selected that correspond to the optimal values for the nearest neighbor P^* from previously calculated results (for faster convergence).
(2) After defining the FCA flow rates and initial computing dies frequencies, the algorithm performs the frequency optimization for each computing die (described in 3.1.2).
(3) If the optimal frequencies are not yet achieved using the current FCA flow rates combination Q, the algorithm uses simulated annealing to update these values. In particular, it considers a neighboring state Q^* and iterates back to step 2 to evaluate the 3D MPSoC performance in this scenario.
(4) If the optimal frequencies $F_{opt} = (f_1 \dots f_N)$ are achieved for a given FCA flow rates combination Q, the optimization loop terminates. The optimal flow rates and frequency settings are then recorded.

3.1.2 Frequency optimization loop. This loop evaluates the applicable frequency increase ratio (boost) of each computing die for different degrees of utilization, considering a fixed FCA flow rate configuration determined by the outer loop. It accounts for temperature and voltage drop effects on the circuit timing characteristics and dictates the clock frequencies of cores to avoid timing violations. Hence, the following sequence of steps is performed for each computing die recursively until convergence, as in Figure 2:

(i) The algorithm estimates the initial dynamic power consumption of components based on their utilization percentage: $p_{dyn} = \rho * p_{max}$. The power values are then mapped to the layout of the die to construct the dynamic power map $P_{dyn,init}$. It is then scaled according to the frequency boost ratio α using a quadratic frequency-power relationship for high-performance chips [14]: $P_{dyn} = (1+\alpha)^2 P_{dyn,init}$.
(ii) The temperature map T of each die is evaluated using 3D-ICE [15] when the power consumption profiles of the 3D MPSoC computing dies correspond to the previously calculated power maps.
(iii) The leakage map P_{leak} is estimated for each computing die according to its thermal map T, using the temperature and transistor leakage relationship that characterizes the specific CMOS technology [13]. Hence, the total power map is calculated as: $P_{total} = P_{dyn} + P_{leak}$.
(iv) The algorithm builds a fine-grain 3D MPSoC electrical model [5] using a compact model of the PDN, FCAs and SC converters [12]. This model is simulated using HSPICE to obtain the maximal voltage drop value Δv_{max}.
(v) The timing of the most critical path τ_{max} is estimated based on its temperature and voltage value. The algorithm pessimistically considers a worst-case operating condition where the maximal temperature and voltage drop nodes coincide. Then, the critical path timing is compared to the current

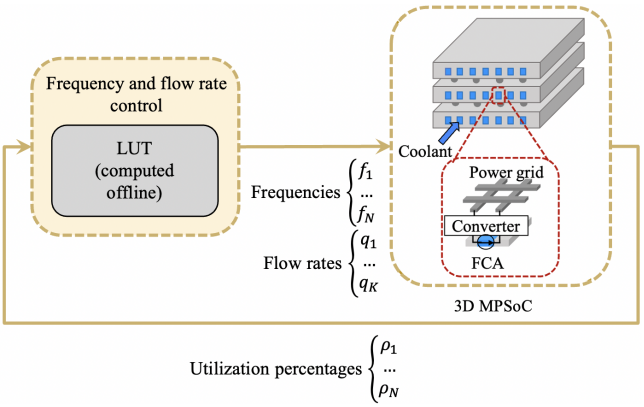

Figure 3: Run-time Frequency and Flow Rate Control

clock period value to compute the frequency (or timing) error $e_f = (1+\alpha)f - 1/\tau_{max}$.

(vi) If the timing error e_f is positive or lower than a certain threshold e_{max}, the frequency boost ratio is updated using a gradient descent methodology: $\alpha = \alpha + \beta e_f$ (the optimization parameters β and e_{max} are chosen so that the solution converges). If the timing error is negative and higher than the threshold, the loop terminates and returns the current optimal frequency boost ratio α_{opt} and frequency $f_{opt} = (1+\alpha_{opt})f_{nom}$, corresponding to the utilization level and FCA flow rate configuration.

3.2 Online Frequency and FCA Flow Rate Control Strategy

A high-level view of the implemented frequency and FCA flow rate control strategy is shown in figure 3. The online controller uses the flow rates and frequency boost values computed by the offline solver in Section 3.1 to implement an explicit MPC [16]. Hence, The controller module periodically receives utilization data from task schedulers then selects the corresponding FCAs and operating frequency settings according to the power requirements of each die. The controller applies these settings to the 3D MPSoC during the subsequent time period from the pre-computed LUT.

4 EXPERIMENTAL SETUP AND RESULTS

4.1 Target 3D MPSoC and Benchmarks

To evaluate the efficiency of the proposed 3D MPSoC performance management strategy described in Section 3, we employ as a target system the four-layer stack shown in Figure 4. We base the architecture on a state-of-the-art CPU-GPU platform for high-performance computing [17]. We consider its implementation in 3D, anticipating a next-generation computing platform [1]. The system comprises the following layers:

- The first layer is modeled after AMD EPYC microprocessor [18]. It contains 32 cores operating at a base frequency of $2GHz$, which can be boosted up to $2.55GHz$. Its peak power consumption and temperature are $180W$ and 81°C, with a size of $757mm^2$.
- The second layer contains a DDR4-2666W [19] memory sizing $654mm^2$, supported by the EPYC processor.

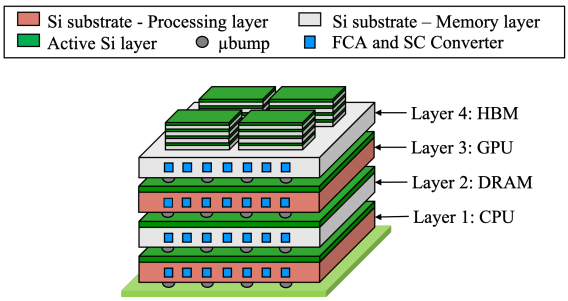

Figure 4: Target 3D MPSoC

- The third layer is based on the NVIDIA V100 [20] data center GPU, composed of 640 Tensor cores and 5120 CUDA cores, with a size of $815mm^2$. It consumes up to $300W$ and supports a maximal temperature of 85°C. The GPU core frequency ranges between 1230 and $1380MHz$.
- The fourth layer is composed of four 2^{nd} generation HBM memories with 4 DRAM layers, providing the bandwidth requirement of the GPU. The maximal power consumption of each memory is $15W$ [21], with a size of $71mm^2$.

The target 3D MPSoC employs chiplet-based integration to stack the HBMs on an active interposer (top layer) and chip-on-chip bonding through fine-pitched micro-bumps to stack the four 3D MPSoC layers. FCAs are etched in all the dies. They have a width, height, and pitch of 50, 100, and $50\mu m$, respectively. Similarly to [12], each $200\mu m$-long flow cell section is connected to the corresponding power grid through an SC converter. FCAs are only electrically connected to the GPU and CPU, as the memories have considerably lower power requirements. We assume that an EMB MHIE centrifugal pump [22] is responsible for the fluid injection to all the flow cells. Then, we assume normally closed valves [23] enabling different flow rates for each die. We consider that the flow rate ranges between $40ml/min$ (the value achieving the maximum temperature) and $220ml/min$ (the maximal flow rate, as in [4]).

To evaluate the resulting run-time speedups when using the controller described in Section 3.2, we explore a range of benchmarks representing different power consumption profiles. Their utilization traces are measured using the performance counters of a real 2D system, equivalent to the target 3D MPSoC. In particular, we run the following machine learning algorithms on the GPU: *Resnet* (RN) [24], *Inception* V3 (I3) and V4 (I4) [25], *Deep Speech* (DS) [26], and *Fairseq* (FS) [27]. On the CPU, we run the following workloads from the SPEC benchmark [28]: *Cactus Computational Framework* (Ca), *Weather Research and Forecasting Model* (wrf), *ImageMagick* (IM), *Lattice Boltzmann Method* (lbm), and *Parallel Ocean Program* (pop2).

4.2 Frequency and FCA Flow Rate Optimization Results

We deployed the FCA flow rate and frequency optimization methodology described in Section 3.1 to assess the frequency boost ratios applicable to the target 3D MPSoC computing dies (Figure 4), under various utilization scenarios. As the memories have a negligible impact on the thermal performance of the system, we pessimistically assume they are in full usage at all times. For comparison, we also

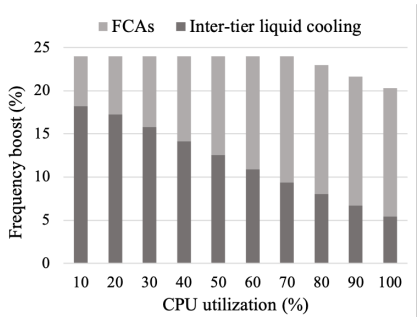

Figure 5: Applicable CPU Frequency Boosts

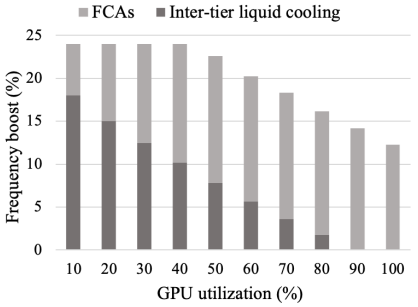

Figure 6: Applicable GPU Frequency Boosts

run a similar optimization algorithm of the 3D stack only considering the cooling capabilities of FCAs (i.e., no FCA power extraction) to highlight the additional benefit of FCAs over inter-tier liquid cooling.

Figures 5 and 6 show the optimal frequency boost ratios for both the CPU and the GPU, respectively. The proposed optimization methodology enables between 20% and 24% higher operating frequency than the nominal value in the CPU case. The upper frequency boost limit is due to the minimal critical path timing, which must remain lower than the CPU clock frequency. Compared to inter-tier liquid cooling, FCAs enable up to 15% more frequency boost due to their additional power supply. This observation is particularly prominent in the case of high CPU usage, as the leakage reduction only compensates for part of the power grid voltage losses. In the case of the GPU, the optimization solver using FCAs enables between 12% and 24% higher operating frequency. As the GPU has a higher overall power consumption than the CPU, the leakage reduction and additional power supply of FCAs have a lower impact on the total power distribution in the die. Hence, the GPU achieves lower frequency boosts compared to the CPU. Similar to the CPU case, inter-tier liquid cooling enables a considerably lower GPU frequency boost than FCAs. In the case of very high GPU utilization (>90%), the cooling does not compensate enough losses in the PDN to enable higher GPU power consumption (i.e., operating frequency).

Figure 7 showcases the calculated optimal total flow rates (i.e., the sum of the FCA flow rates in all the 3D MPSoC layers) corresponding to each CPU and GPU utilization scenario. We consider that the FCA flow rate for each layer can be adjusted in steps of $20ml/min$. As a result, the selected FCA flow rate configurations achieve a maximum difference of $2°C$ in the 3D MPSoC thermal hotspot. As the utilization of both computing dies increases, the

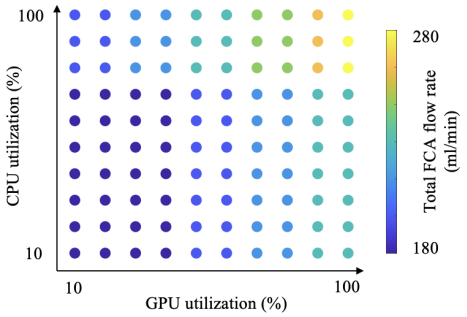

Figure 7: Total Applicable FCA Flow Rates

3D MPSoC requires a higher FCA cooling capacity. Hence, the optimal total flow rate that achieves the highest computing frequency increments with the system utilization percentage. Furthermore, GPU usage generally has a higher impact in dictating the optimal FCA flow rate settings, as it is responsible for most of the total 3D MPSoC power consumption.

4.3 Workload Speedup on the 3D MPSoC with FCAs

We simulated the online frequency and FCA flow rate control strategy in Section 3.2 when running the benchmarks in Section 4.1 on the 3D MPSoC. To estimate the execution of the workloads using our management strategy, we first recorded their execution in an equivalent 2D system. Considering the same executed task schedule, we simulated the execution on the 3D MPSoC by compressing the original traces according to the speedup rates determined by the controller every 100ms (the minimal value dictated by the sensors used to gather workload traces). We pessimistically assume the same memory bandwidth and that the frequency boost only applies to the computing dies (CPU and GPU).

Figure 8 shows the achieved workload speedups for both the CPU and GPU, using the proposed frequency and flow rate management strategy (*Optimal flow rate* in Figure 8). Our results show a speedup of up to 17% and 15% on the CPU and GPU, respectively. In the CPU case, *compute-intensive* benchmarks present the highest overall speedup as the frequency boost does not affect memory access time. In the GPU case, the total achieved speedup depends on the utilization level and the dynamic power consumption. The former involves the percentage of time when the frequency boost is applied (computing versus memory access), and the latter affects the applicable frequency boost values. In particular, the Resnet (RN) benchmark benefits from the highest overall speedup as its percentage of computing versus memory access time is the largest. Conversely, the Fairseq (FS) benchmark also achieves a high overall speedup because it has the lowest dynamic power consumption, enabling the highest average frequency boost.

For comparison, we recorded the achieved speedups in the case of a fixed FCA flow rate, considering the maximal value (*Maximal flow rate* in Figure 8). In this case, only the computing dies frequencies are adjusted during run-time (the inner loop in Figure 2). This strategy enables comparable workload speedups to the dynamic FCA flow rate case but uses significantly higher cooling resources. In particular, Table 1 shows the total cooling energy savings when using the dynamic FCA flow rate management compared to the

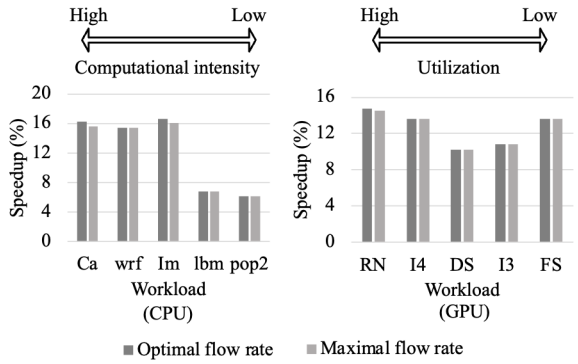

Figure 8: Achieved Speedups on Targeted Benchmarks

		CPU workloads				
		Ca	wrf	Im	lbm	pop2
GPU workloads	RN	15%	18%	16%	28%	30%
	I4	25%	28%	26%	36%	37%
	DS	22%	25%	22%	38%	35%
	I3	33%	35%	33%	43%	43%
	FS	31%	36%	33%	43%	43%

Table 1: Cooling Energy Savings

fixed flow rate case. These savings are calculated based on the total flow rates applied during workloads execution, using the flow rate-liquid pumping power relationship in [9]. Hence, we show that our proposed strategy achieves the optimal workload speedups while economizing up to 43% cooling energy. In particular, the workload combinations that have the lowest utilization (e.g., FS and pop2) can be efficiently executed using significantly lower total FCA flow rate than the maximal value (as shown in Figure 7).

Finally, other state-of-the-art DTM policies imply performance degradation. In particular, DVFS lowers the frequency settings of dies during high utilization, and task migration incurs 100s of milliseconds of delay. Instead, our approach leveraging FCAs enables tangible run-time gains due to the increased power and thermal budget made available by FCAs.

5 CONCLUSION

In this paper, we have proposed a run-time performance management strategy for 3D MPSoCs with integrated FCAs. Using fine-grained thermal and power modeling and analysis, we have shown that the combined cooling and power generation potential of FCAs can be leveraged to boost the power efficiency of multi-core processing dies in a 3D stack while satisfying design constraints. By throttling the operating frequencies and FCA flow rates during run-time, we demonstrated execution speedups of up to 17% for state-of-the-art compute-intensive workloads. Our strategy increases the system performance without any software or architecture optimization and uses significantly less cooling energy than a fixed flow rate strategy. It also does not depend on the stack architecture or the deployed CMOS technologies. Hence, our results advocate for the adoption of FCA technology as an enabler of power-efficient 3D MPSoCs targeting modern high-performance applications.

ACKNOWLEDGMENTS

This work has been supported by the EC H2020 WiPLASH (GA No. 863337) and the ERC Consolidator Grant COMPUSAPIEN (GA No. 725657) projects.

REFERENCES

[1] F. Clermidy et al. 3D Embedded Multi-Core: Some Perspectives. *Design, Automation & Test in Europe (DATE)*, 2011.
[2] P. Emma and E. Kursun. Opportunities and Challenges for 3D Systems and Their Design. *IEEE Design & Test of Computers*, 2009.
[3] M. Jung and S. K. Lim. A study of IR-drop Noise Issues in 3D ICs with Through-Silicon-Vias. *IEEE International 3D Systems Integration Conference*, 2010.
[4] A. Andreev et al. PowerCool: Simulation of Cooling and Powering of 3D MPSoCs with Integrated Flow Cell Arrays. *IEEE Transactions on Computers (TC)*, 2018.
[5] H. Najibi et al. A Design Framework for Thermal-Aware Power Delivery Network in 3D MPSoCs with Integrated Flow Cell Arrays. *International Symposium on Low Power Electronics and Design*, 2019.
[6] E. Wong et al. 3D Floorplanning with Thermal Vias. *DATE*, 2006.
[7] A. Bartolini et al. Unveiling Eurora: Thermal and Power Characterization of the most Energy-Efficient Supercomputer in the World. *DATE*, 2014.
[8] A. Sridhar, M. M. Sabry, and D. Atienza. System-level thermal-aware design of 3D multiprocessors with inter-tier liquid cooling. *International Workshop on Thermal Investigations of ICs and Systems*, 2011.
[9] M. M. Sabry et al. Energy-Efficient Multi-objective Thermal Control for Liquid-Cooled 3D Stacked Architectures. *IEEE Transactions on Computer-Aded Design of Integrated Circuits and Systems*, 2011.
[10] A. K. Coskun et al. Modeling and Dynamic Management of 3D Multicore Systems with Liquid Cooling. *International Conference on Very Large Scale Integration*, 2009.
[11] F. Zanini, D. Atienza, and G. De Micheli. A Combined Sensor Placement and Convex Optimization Approach for Thermal Management in 3D-MPSoC with Liquid Cooling. *INTEGRATION, the VLSI journal*, 2013.
[12] H. Najibi et al. Enabling Optimal Power Generation of Flow Cell Arrays in 3D MPSoCs with On-Chip Switched Capacitor Converters. *IEEE Computer Society Annual Symposium on VLSI*, 2020.
[13] H. Najibi et al. Towards Deeply Scaled 3D MPSoCs with Itegrated Flow Cell Array Technology. *Great Lakes Symposium on Very Large Scale Integration*, 2020.
[14] P. Bogdan, R. Marculescu, and S. Jain. Dynamic Power Management for Multidomain System-on-Chip Platforms: An Optimal Control Approach. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 2013.
[15] A. Sridhar et al. 3D-ICE: a compact thermal model for early-stage design of liquid-cooled ICs. *TC*, 2014.
[16] F. Zanini et al. Online Thermal Control Methods for Multiprocessor Systems. *TODAES*, 2012.
[17] NVIDIA Tesla V100 Servers [Online]. Retrieved from https://www.thinkmate.com/systems/servers/gpx/v100.
[18] K. Lepak et al. The Next Generation AMD Enterprise Server Product Architecture. *Hot Chips*, 2017.
[19] S. Shim et al. A 16Gb 1.2V 3.2Gb/s/pin DDR4 SDRAM with Improved Power Distribution and Repair Strategy. *International Solid-State Circuits Conference*, 2018.
[20] NVIDIA TESLA V100 GPU Architecture [Online], 2017. Retrieved from http://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf.
[21] D. Lee et al. A 1.2 V 8Gb 8-Channel 128GB/s High-Bandwidth Memory (HBM) Stacked DRAM With Effective I/O Test Circuits. *IEEE Journal of Solid-State Circuits*, 2015.
[22] WILO MHIE Centrifugal Pump [Online]. Retrieved from http://www.wilo.com/cps/rde/xchg/en/layout.xsl/3707.html.
[23] Festo Electric Automation Technology [Online]. Retrieved from http://www.festodidactic.com/ov3/media/customers/1100/0096636000107522-3683.pdf.
[24] K. He et al. Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[25] C. Szegedy et al. Rethinking the Inception Architecture for Computer Vision. *CVPR*, 2016.
[26] D. Amodei et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. *International Conference on Machine Learning*, 2016.
[27] M. Ott et al. fairseq: A fast, extensible toolkit for sequence modeling. *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.
[28] J. Bucek, K.-D Lange, and J.-V. Kristowski. SPEC CPU2017 – Next-Generation Compute Benchmark. *ACM/SPEC International Conference on Performance Engineering*, 2018.