Thèse n° 9117

DESIGN AND MANAGEMENT OF THREE-DIMENSIONAL MULTI-PROCESSOR SYSTEMS-ON-CHIP WITH INTEGRATED FLOW CELL ARRAYS

Présentée le 2 décembre 2022

Faculté des sciences et techniques de l’ingénieur
Laboratoire des systèmes embarqués
Programme doctoral en génie électrique

pour l’obtention du grade de Docteur ès Sciences

par

Halima NAJIBI

Acceptée sur proposition du jury

Prof. A. P. Burg, président du jury
Prof. D. Atienza Alonso, directeur de thèse
Dr S. Nassif, rapporteur
Prof. A. Coskun, rapporteuse
Prof. K. Choo, rapporteur

École
polytechnique
fédérale
de Lausanne

2022

Logic will get you from A to B.
Imagination will take you everywhere.
— Albert Einstein

To my family and loved ones…

Acknowledgements

Before describing the scientific work accomplished during my Ph.D. at ESL, I wish to say a few words about some of the people that held my hand, both figuratively and literally, during this significant period of my life.

First and foremost, I should thank my thesis supervisor Professor David Atienza for being a great mentor and role model and for giving me the opportunity to work in a friendly yet ambitious environment. By being supportive and invested in my progress, he inspired me never to give anything less than my best for my research. He never held back any constructive comments about my work or failed to assist me when I needed it. He created an exceptionally supportive and comfortable atmosphere in the workplace, and went above and beyond to plan events for us to relax and enjoy each other's company (even virtually during a global pandemic). David has been an exemplary supervisor for whom I will always be grateful.

Then, I would like to express my gratitude to the talented post-doctoral researchers who provided me with continuous support. Prof. Marina Zapater was my first supervisor when I joined the laboratory. Her assistance was instrumental for guiding me to choose the direction of my work. Given her experience in the field, she provided crucial advice in terms of technical work. She also was of tremendous help in publishing my research the right way and to the right community. Next, Dr. Alexandre Levisse has always been an inspiration for innovative thinking, optimism, and enthusiasm. He shared many insightful tips based on his experience as a Ph.D., which were particularly helpful during my first years. His assistance continued to be valuable throughout the entirety of my Ph.D., allowing me to enhance the quality of my work. Finally, Dr. Giovanni Ansaloni joined in during the second half of my Ph.D. but quickly became a real asset to my research. He is friendly, encouraging, and always available to brainstorm and develop new ideas. It has always felt comforting to work with a brilliant, highly motivated, and collaborative researcher like him, always enjoying the work at hand.

I am also grateful for the help of other post-doctoral researchers in the lab (in particular Dr. Miguel Peon and Dr. Tomas Teijeiro) and technical staff (Rodolph, Michael, Christoph, and John) who assisted with certain tasks, and helped resolving a few “crises”.

Next, I want to thank ESL's administrative assistant, Ms. Homeira Salimi, first of all for her friendship, then for all the help she provided with miscellaneous requests throughout the years. Moreover, I thank her for the exceptional job she has done organizing monthly birthday celebrations, sports and game activities, fancy dinners, and many other events for everyone in the lab to get together and have fun.

I have also had the chance to meet some wonderful people and develop many friendships inside ESL family. Perhaps the most significant one has been with Dr. Dionisije Sopic, who opened up to me from the first few weeks after arriving at the lab and quickly became the best friend with whom I shared precious moments inside and outside of work. I am also especially appreciative of the great friendships I had with Dr. Soumya Basu, Dr. Arman Iranfar, and Mr. Renato Zanetti. Moreover, I have spent memorable times with other friends and colleagues: Adriana, Ali, Andrew, Artem, Benoît, Elisabetta, Fabio (CH), Fabio (IT), Farnaz, Grégoire, Lara, Marco, Nacho, Saleh. Finally, I thank all the rest of the ESL family, and in particular the Android App Development course team, for forming a great work environment.

I'd like to thank my best EPFL friends outside ESL (Dr. Behnoush Attari and Dr. Yasmine Zakari) for being a constant presence during this whole period, doing activities like exercise, travel, dinners, drinks, and in general for sharing a special bond that will remain for life.

Finally, I am immensely grateful for my wonderful parents for their love and support of my goals and dreams, my brothers Ali and Youssef, and cousin Sarah for always being my most loyal confidants and my number one fan club, my uncle Taoufiq, aunt Margreth, and cousin Elias for embodying the comfort of a family far from home, and finally my loving partner Maxime for the countless words of encouragement and for facilitating life in moments of stress.

Lausanne, September 13, 2022 H. N.

Abstract

Three-Dimensional Multi-Processor Systems-on-Chip (3D MPSoCs) are promising solutions for highly intensive Artificial Intelligence (AI) and Big Data applications. They combine remarkably dense computation capabilities and massive communication bandwidths. However, due to their high density, 3D MPSoCs present heat dissipation, and power delivery challenges. Flow Cell Arrays (FCAs) have the potential to solve both issues. They consist of micro-channels etched in the Silicon substrate of stacked dies, filled with an electrolytic flow that absorbs the generated heat and produces power through electrochemical reactions. Hence, FCAs enable concurrent on-chip liquid cooling and electrochemical power generation in 3D MPSoCs.

In this context, this thesis focuses on the system-level integration of FCAs as an effective cooling solution and power source for next-generation 3D MPSoCs. First, a comprehensive framework is proposed to model 3D MPSoCs with FCAs in fine grain and analyze their thermal and power performance. Simulations demonstrate temperature reductions of up to 40°C and voltage drop reductions of up to 90% when using FCAs, even for 3D MPSoCs fabricated using deeply-scaled CMOS technologies, which are characterized by extremely high power densities. Next, the thesis introduces design-time techniques to enhance the performance of FCAs. It advocates the use of Switched Capacitor (SC) converters as an interface between the channels and the power delivery grid. SCs enable FCAs to operate at their optimal voltage to maximize power generation, resulting in gains of up to 2x compared to direct connectivity to the power grid. Then, considering FCA's interdependent thermal and power generation capacities, several design configurations are explored, highlighting trade-offs and revealing opportunities to increase the power budget of 3D MPSoCs without violating design constraints. Finally, the thesis illustrates a novel strategy to manage the cooling and power generation capabilities of FCAs at run-time. Such an optimization strategy boosts the operating frequencies of dies, increasing the computing performance up to 24% while reducing coolant flow requirements.

Keywords: 3D integration, 3D multi-processor systems-on-chip (MPSoC), Flow cell arrays, Thermal performance, Power performance, Design-time management, Run-time management.

Résumé

Les systèmes multiprocesseurs tri-dimensionnels sur une puce (MPSoCs 3D) sont des solutions prometteuses pour l'exécution de certaines applications modernes très intensives, telles que l'intelligence artificielle (IA) ou le traitement de données massives (Big Data). Ces plateformes combinent des capacités de calcul extrêmement denses et d'énormes largeurs de bande de communication entre les différents éléments. Cependant, les MPSoCs 3D souffrent de dissipation thermique difficile et d'alimentation complexe, en raison de leur énorme densité de puissance. Les Flow Cell Arrays (FCAs) ont le potentiel de résoudre ces deux problèmes. Elles consistent en des micro-canaux gravés dans le substrat en silicium des puces électroniques empilées. Elles sont remplies d'un flux électrolytique qui permet d'absorber la chaleur générée par l'activité des composantes. Ce flux produit également de l'énergie grâce à des réactions électrochimiques entre les électrolytes. Ainsi, les FCAs permettent le refroidissement et la production d'énergie dans les MPSoCs 3D.

Dans ce contexte-là, cette thèse se concentre sur l'intégration des FCAs au niveau système, en tant que solution efficace pour le refroidissement et comme source d'énergie additionnelle pour les MPSoCs 3D de la nouvelle génération. Tout d'abord, une démarche complète est proposée pour la modélisation granulaire des MPSoCs 3D avec des FCAs, ainsi que l'analyse de leur performance thermique et énergétique. Les simulations démontrent des réductions de température allant jusqu'à 40°C et des réductions de chute de tension allant jusqu'à 90% lors de l'utilisation des FCAs. Ceci reste valide même pour les MPSoCs 3D fabriqués en utilisant des technologies CMOS avancées, qui se caractérisent par des densités de puissance extrêmes. Ensuite, la thèse présente des techniques de conception en vue d'améliorer les performances des FCAs. Elle préconise l'utilisation de convertisseurs de tension comme interface entre les micro-canaux et le réseau de distribution d'énergie. Les convertisseurs permettent de faire fonctionner les FCAs à leur tension optimale, maximisant ainsi la production d'énergie. Cela entraîne d'importants gains, allant jusqu'à doubler l'énergie produite par rapport à une connectivité directe des FCAs au réseau électrique. Ensuite, en considérant l'interdépendance des capacités de refroidissement et de production d'énergie des FCAs, plusieurs configurations sont explorées. Celles-ci mettent en évidence les compromis existants et révèlent la possibilité d'augmenter la capacité énergétique des MPSoCs 3D tout en respectant les contraintes physiques. Enfin, la thèse illustre une nouvelle stratégie pour gérer les capacités de refroidissement et de production d'énergie des FCAs en cours d'exécution d'applications. Cette stratégie a pour objectif d'augmenter la vitesse de fonctionnement des composantes électroniques. Ainsi, elle permet d'intensifier la performance de calcul jusqu'à 25%, tout en réduisant le coût de refroidissement des FCAs.

Mots-clés : Intégration en 3D, systèmes multiprocesseurs 3D sur une puce, Technologie des Flow Cell Arrays, Performance thermique, Haute performance, Gestion de la puissance, Techniques de conception, Stratégies de gestion.

Contents

Acknowledgements **i**

Abstract (English/Français) **iii**

1 Introduction **1**
- 1.1 3D integration for high performance . . . 1
- 1.2 3D thermal dissipation challenges . . . 4
- 1.3 3D power delivery challenges . . . 5
- 1.4 Contributions . . . 5

2 Background **7**
- 2.1 Inter-tier liquid cooling for 3D ICs . . . 7
- 2.2 Simulation of inter-tier liquid-cooled 3D ICs . . . 8
- 2.3 Flow cell arrays (FCAs) . . . 9
- 2.4 Electro-thermal simulation of 3D ICs with integrated FCAs . . . 11

3 3D MPSoC design with integrated FCAs **13**
- 3.1 Introduction . . . 13
- 3.2 Related work . . . 16
  - 3.2.1 3D integration trends and challenges . . . 16
  - 3.2.2 Power delivery grid modeling . . . 16
  - 3.2.3 Integrated FCAs . . . 17
- 3.3 Framework for modeling and analysis of 3D MPSoCs with integrated FCAs . . . 18
  - 3.3.1 Framework overview . . . 18
  - 3.3.2 Flow cell electrical model . . . 19
  - 3.3.3 3D power delivery network modeling . . . 20
- 3.4 Target 3D MPSoC architecture . . . 23
  - 3.4.1 3D MPSoC and 3D PDN description . . . 23
  - 3.4.2 3D MPSoC utilization scenarios . . . 26
- 3.5 3D MPSoC performance evaluation . . . 28
  - 3.5.1 3D MPSoC configurations . . . 28
  - 3.5.2 TSV placement exploration . . . 30
  - 3.5.3 Experimental flow . . . 31
  - 3.5.4 Experimental results . . . 32

3.5.4.1 Evaluation of 3D MPSoC temperature using FCAs 32
3.5.4.2 Evaluation of IR-drop reduction using FCAs 33
3.5.4.3 TSV placement evaluation 37
3.6 Evaluation of deeply-scaled 3D MPSoCs 39
3.6.1 3D MPSoC scaling 40
3.6.1.1 Architecture overview 40
3.6.1.2 Technology scaling methodology 40
3.6.2 Experimental setup 43
3.6.2.1 Experimental flow 43
3.6.2.2 3D MPSoC cooling strategies 43
3.6.2.3 Power network modeling at advanced technology nodes 44
3.6.2.4 PDN modeling and IR-drop analysis 46
3.6.3 Experimental results 47
3.6.3.1 FCA on-chip cooling capabilities 47
3.6.3.2 FCA IR-drop reduction capabilities 48
3.7 Conclusion 51

4 Design-time management strategies for 3D MPSoCs with FCAs 53
4.1 Introduction 53
4.2 Related work 56
4.2.1 3D MPSoCs trends and challenges 56
4.2.2 Thermal management strategies for 3D MPSoCs 56
4.2.3 Power management strategies for 3D MPSoCs 57
4.2.4 3D MPSoC design with integrated flow cell arrays 57
4.2.5 Voltage converters for efficient power delivery 59
4.3 Optimal FCA power generation with switched capacitor (SC) converters 60
4.3.1 SC converter design 60
4.3.1.1 SC converter state-space model 60
4.3.1.2 SC converter design space parameters 63
4.3.1.3 SC converter design exploration 64
4.3.2 Experimental setup 66
4.3.2.1 3D MPSoC with FCAs and SC converters 66
4.3.2.2 Processor utilisation scenarios 67
4.3.2.3 Experimental flow 68
4.3.3 Experimental results 68
4.3.3.1 FCA power generation with SC converters 68
4.3.3.2 SC converter area optimization 70
4.3.3.3 FCA power generation resource management 71
4.4 Design-time characterization of 3D MPSoCs with FCAs and SC converters . . . 73
4.4.1 Experimental setup 74
4.4.1.1 Target 3D MPSoC 74
4.4.1.2 FCA placement and flow rate configuration 76

4.4.1.3 SC converter configuration 78
4.4.1.4 Experimental flow 79
4.4.2 Experimental results 81
4.4.2.1 Temperature and total power consumption 81
4.4.2.2 Voltage drop recovery 83
4.5 Conclusion 85

5 Run-time management strategies for 3D MPSoCs with FCAs 87
5.1 Introduction 87
5.2 Related work 89
5.2.1 3D MPSoC thermal and power challenges 89
5.2.2 3D MPSoC run-time management strategies 89
5.2.3 3D MPSoC with integrated FCAs 90
5.3 Target 3D MPSoC 92
5.3.1 Stack composition 92
5.3.2 Target workloads 93
5.4 Thermal and power-aware 3D MPSoC frequency optimization 97
5.4.1 Offline frequency optimization 97
5.4.2 Online frequency control 102
5.4.3 Experimental results on target 3D MPSoC 104
5.4.3.1 Frequency boosts 104
5.4.3.2 Workload speedups 105
5.5 Thermal and power-aware flow rate and frequency co-optimization 110
5.5.1 Offline frequency and flow rate optimization 110
5.5.2 Online frequency and flow rate control 112
5.5.3 Experimental results on target 3D MPSoC 113
5.5.3.1 Optimal flow rates and frequency boosts 113
5.5.3.2 Workload speedups 115
5.5.3.3 Cooling energy savings 117
5.6 Comparison with other DTM strategies 117
5.7 Conclusion 118

6 Conclusion 119
6.1 Summary 119
6.2 Future work 121
6.2.1 3D MPSoC architecture design with FCAs 121
6.2.2 Workload optimization on 3D MPSoCs with FCAs 122
6.2.3 3D MPSoC thermal and power management with FCAs 122
6.2.4 FCA design for high-performance 3D MPSoCs 123

Bibliography 125

List of Figures 135

List of Tables 138

Curriculum Vitae 141

1 Introduction

This thesis considers three-dimensional multi-processor systems-on-chip (3D MPSoCs) as an effective solution to the increasing demand for high-speed and high-density computing due to their numerous advantages compared to two-dimensional (2D) systems. 3D MPSoCs employ 3D integration techniques to stack multiple IC components and construct compact, powerful, and efficient computational platforms. This introductory section presents the fundamentals of 3D integration and the motivations behind adopting this technology when targeting high-performance systems that respond to the demands of modern applications. Then, this section describes the primary challenges constraining the design of high-power 3D systems, namely the intense heat generation and complex power delivery. Finally, the section introduces the main contributions of this thesis, which consist of proposing solutions to address these challenges and enable a reliable implementation of high-performance 3D MPSoCs.

1.1 3D integration for high performance

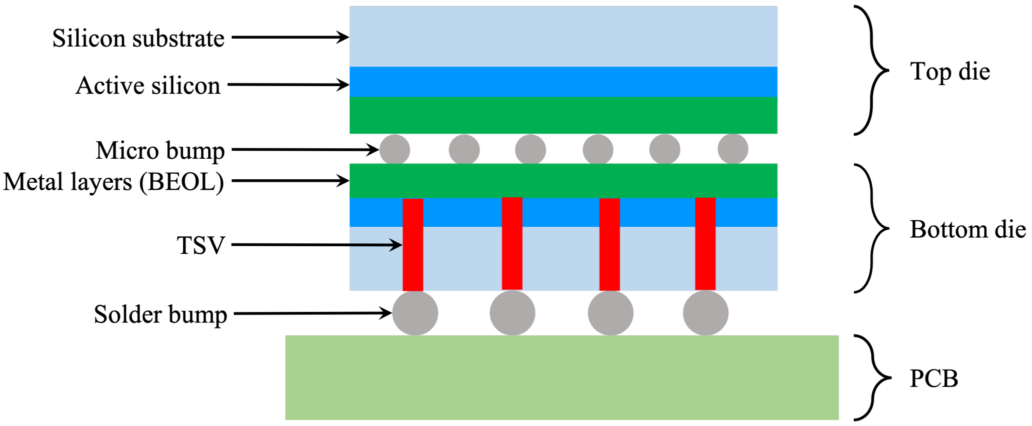


Figure 1.1: 3D integration of ICs using TSV and microbump technology

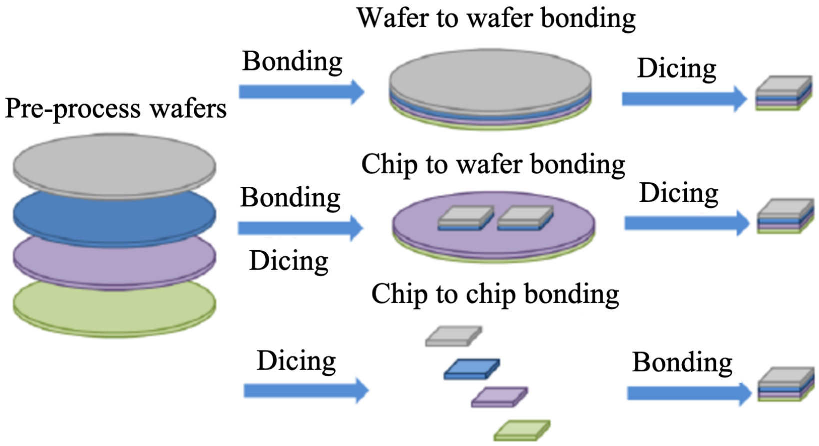


Figure 1.2: Bonding technologies for 3D integration [3]

Over the last decades, significant advances in lithography and device technology have continuously increased the performance and efficiency of microelectronic systems in response to the needs of customers and applications. In particular, the semiconductor industry development has long been dominated by transistor shrinking to increase computing density and reduce IC fabrication costs. In recent years, however, three-dimensional (3D) integration based on through-silicon-vias (TSV) [1] has become a well-accepted and promising approach achieving compact IC systems, overcoming the performance bottleneck related to the lack of interconnect scaling [2]. 3D integration consists of constructing a system with multiple component dies fabricated using conventional processes, stacked on top of each other, and connected through TSVs and microbumps, as illustrated in Figure 1.1. It is possible, for instance, to bond entire wafers and then dice them into separate 3D ICs, or stack individual dies [3] (shown in Figure 1.2). Moreover, many combinations are feasible for post-back-end-of-line die-stacking. In a face-to-face configuration, the front-end electronics of two chips are directly connected through microbumps (similarly to Figure 1.1). In a back-to-back configuration, the back ends (silicon substrate) of two dies are connected using TSVs and microbumps. Finally, in a face-to-back configuration, the front end of a die is connected to the back end of a second die through TSVs and microbumps.

A critical challenge in IC system design comes from signal propagation delays and power losses, mainly determined by wiring lengths, resistances, and capacitances. These delays are exacerbated by the increasing distances between components in modern System-on-Chips (SoCs). Therefore, 3D integration with TSVs is an effective solution to overcome this challenge by providing vertical links for communication and power delivery, hence improving system performance without increasing power consumption [1]. In addition, 3D integration significantly reduces the area demand of multi-component SoCs and augments their computing density [4]. Finally, 3D integration is recognized as a key enabler of heterogeneous IC systems, offering the opportunity to compactly integrate dies fabricated with different architectures and CMOS technologies [2].

Many semiconductor technology suppliers and research organizations have developed full 3D integration processes using TSV technology, demonstrating their manufacturability and reliability. In fact, different 3D integrated ICs have been conceptualized and/or manufactured. They can be categorized as follows:

- **Memory-on-memory architectures** are perhaps the most representative application of 3D integration. 3D-stacked DRAM structures have become commercially available, providing exceptionally high bandwidths. As an example, the low-power Hybrid Memory Cube (HMC) [5], developed by Micron, is a 3-D-stacked DRAM with one logic layer and multiple DRAM layers providing abundant parallelism with hundreds of banks per device [2]. The layers are connected through thousands of TSVs. Then, with high-speed serial links, HMC achieves up to 320 GB/s of external bandwidth. The High-Bandwidth Memory (HBM) [6] is another example of a successful 3D memory architecture adopted by Samsung, AMD, and SK Hynix. HBM achieves a breakthrough bandwidth of over 256GB/s. Similar to HMC, it is an attractive solution for high-performance systems as it provides scalability of memory capacity and a small footprint.
- **Memory-on-processor and processor-on-memory architectures** are also gaining popularity in the IC design industry and research community because they fundamentally increase data access speed and computing performance in general. Hence, designers from Intel have proposed stacking large SRAM caches on a microprocessor, dramatically increasing on-die storage [7]. Their proposed architectures improve cache access time by up to 55% and reduce required off-die bandwidth by up to 66%, while simultaneously reducing power consumption. In addition, the authors of [8] developed a massively parallel processor with a stacked SRAM chip built using face-to-face 3D integration technology. Each core communicates with a dedicated block, enabling negligible data transfer delays. Alternatively, the authors of [9] stack embedded DRAM (eDRAM) last-level caches (LLC) on top of processing cores. Compared to SRAM cells, eDRAM cells have a smaller area and less leakage power. Finally, a 3D-stacked Logic-in-Memory (LiM) system has been introduced by the authors of [10], integrating multiple DRAM layers with an application-specific accelerator for the data-intensive parts of applications. It can achieve up to two orders of magnitude power efficiency improvements compared to optimized FPGA, GPU, and CPU implementations.
- **Processor-on-processor architectures** have also been proposed as a promising system integration alternative because they reduce the power dissipation related to core-to-core communication. Researchers in [11] presented a 3D modular and scalable network-on-chip (NoC) architecture for telecommunication applications. The NoC is scalable to up to 4 processing layers communicating via 3D links achieving 7.4 GB/s data rate with reasonable energy consumption. In addition, engineers in [12] developed an architecture composed of six multi-core chiplets on an active interposer integrating on-chip power management, flexible communication support, and memory control. This architecture enables high-efficiency computing and over 50× faster execution of state-

of-the-art applications. Finally, Intel has introduced a 3D stacking technology, called Foveros [13], which enables logic-on-logic integration of IP blocks through low-impact connection paths. This technology provides tremendous flexibility and is compatible with their high manufacturing process. It opens the way for systems with heterogeneous high-performance and high-intensity processing elements.

This thesis thoroughly assesses the benefits and pitfalls of 3D integration by exploring different 3D stacks representing combinations of memory-on-memory, memory-on-processor, and processor-on-processor architectures. In particular, it considers the thermal and power-related challenges that such systems encounter (described in Sections 1.2 and 1.3). Then, the thesis proposes different design and management solutions to overcome these challenges, using integrated microfluidic channels for combined cooling and power generation in 3D IC systems (introduced in Section 2.3).

1.2 3D thermal dissipation challenges

Heat dissipation is a significant challenge in 3D IC design. As more devices are packed into a smaller area, the power density drastically increases [14], therefore generating an extremely high heat flux [15]. In addition, the dies in a TSV-based 3D stack are thinned, which causes their thermal resistance to decrease, creating severe conditions for on-chip hotspots. For instance, when analyzing the thermal behavior of 3D processors, the authors of [16] demonstrated that both memory-on-logic and logic-on-logic topologies incur a temperature rise of up to 11°C compared to an equivalent 2D processor. This issue is exacerbated by the high number of stacked dies and the small gap between overlapping heat sources [15]. In particular, chips further away from the heat sink suffer from the highest temperatures.

Indeed, stacking multiple chips using TSVs and fine-pitch microbumps complicates heat dissipation during operation. For a conventional chip package, the thermal path from the active circuits to the heat sink is principally through the silicon substrate (heat spreader) and the thermal interface material. For a 3D stack, the back-end-of-line (BEOL), the microbump layers between dies, and the insulating materials used for bonding the dies are in the thermal path. These materials have much lower thermal conductivity [17] than silicon. For instance, the BEOL layers have an equivalent thermal conductivity of $2.25\mu W/\mu mK$ versus $150.9\mu W/\mu mK$ for silicon [18]. Hence, the heat generated by the bottom die in a 3D chip has to travel many layers of low conductivity before reaching a heat sink.

In this context, effective thermal management strategies and design guidelines are needed for the widespread use of 3D integration [15]. This thesis advocates the use of Flow Cell Array (FCA) technology as a candidate solution to solve the thermal dissipation issues of high-performance 3D stacks.

1.3 3D power delivery challenges

The second major challenge that hinders the mainstream adoption of 3D integration technology is the complexity of power delivery. In particular, floorplanning and power/ground (P/G) network design play an essential role in the power efficiency of 3D ICs [19]. These two tasks, already critical for 2D ICs, are particularly challenging for 3D ICs. In fact, floorplanning needs to balance the needs of inter-tier communication efficiency against power density when placing components, especially given the limited thermal margins [20]. Then, the P/G network design must target the minimization of the voltage drops across the inter-die and on-die power lines (IR-drop), which produce worst-case operating conditions and have a negative impact on the performance of the circuits [21].

As power in 3D ICs is delivered from the bottom die to other dies through P/G TSVs, the floorplanning and power delivery network design must consider TSV insertion to reduce wire length overheads [20]. P/G TSV placement, in turn, is also impacted by the locations of the logic blocks and the inter-tier communication requirements, which can cause interference with P/G lines and signal TSVs. In addition, IR-drop in the dies is impacted by P/G TSV placement [22]. The farthest TSVs are from the core area, the more critical the IR-drop, particularly for highly-scaled CMOS technologies with small wire widths and higher resistances along power lines. Then, the increasing power density and temperature (hence leakage) of 3D ICs incur additional stress on the power delivery components, which must drive high amounts of current, further aggravating IR-drop.

Consequently, the floorplanning and P/G network design are exceptionally complex problems for 3D ICs that many researchers tackle by developing new physical design guidelines and tools specifically tailored for 3D circuits [20]. This thesis complements these efforts by proposing to use FCAs as an additional on-chip power source, alleviating the need for complex power delivery network structures.

1.4 Contributions

This thesis advocates for the design of heterogeneous high-performance 3D multi-processor systems-on-chip (3D MPSoCs) that achieve high-density and high-speed computing capabilities for next-generation applications. Considering the highly challenging thermal and power aspects attributed to such systems (described in Sections 1.2 and 1.3), it proposes to use a novel technology called Flow Cell Arrays (FCAs; introduced in Section 2.3) to enhance heat dissipation and recover losses in the power delivery circuitry. This technology combines inter-tier microfluidic cooling (described in Section 2.1) and on-chip power generation (described in Section 2.3) to efficiently address the previous challenges.

After introducing the inter-tier cooling and power generation technologies (Chapter 2), this thesis proposes a comprehensive system-level design and analysis methodology for high-performance 3D systems with integrated FCAs. Then, it presents design and run-time strategies to maximize the benefits of FCAs and enhance the power performance of 3D stacked dies in general. In summary, the contributions are as follows:

- Chapter 3 introduces a novel thermal-aware 3D power delivery network modeling and analysis framework for 3D MPSoCs with integrated FCAs. This framework builds a granular model of the 3D MPSoC power network, including FCAs as an additional temperature-dependent power source. This model serves to quantitatively assess the benefits of FCAs on power delivery. It can be used in early design stages to ensure specific power and performance requirements, explore different design space configurations, and predict the performance of FCAs as CMOS technology scales down to nanometric sizes.
- Then, Chapter 4 propose design-time strategies and guidelines to maximize the efficiency of FCAs when integrated into high-performance 3D MPSoCs. First, inserting voltage regulators between the flow cells and power delivery lines permits to operate FCAs optimally, enhancing their power generation capacity. Then, the design space exploration of the different FCA-related parameters unveils the existing trade-offs between cooling and power generation. It also showcases the opportunities to efficiently manage the power performance of 3D MPSoCs, leveraging FCA capabilities in the most advantageous way.
- Next, Chapter 5 introduces a thermal and power-aware run-time performance management strategy for 3D MPSoCs, harnessing the leakage reduction and power supply potential of FCAs. Hence, depending on 3D stack architecture and level of utilization, the operating frequencies of dies are dynamically boosted while remaining within safe temperature, voltage, and timing margins. The run-time strategy also targets the optimization of the electrolytic flow rate settings, involving the inter-dependent cooling and power generation capabilities. Thus, the optimal frequency operation conditions are guaranteed while reducing the FCA cooling cost. This strategy enables considerable execution speedups of benchmarks targetting high-performance 3D MPSoCs.
- Finally, Chapter 6 concludes the thesis by highlighting the impact of its findings in the design and management of high-performance 3D MPSoCs. The chapter also discusses other 3D MPSoC and FCA design challenges that were exposed in the different chapters. These challenges can be tackled in a future work, complementing the work in this thesis to provide a comprehensive guideline for designing next-generation 3D MPSoCs.

2 Background

2.1 Inter-tier liquid cooling for 3D ICs

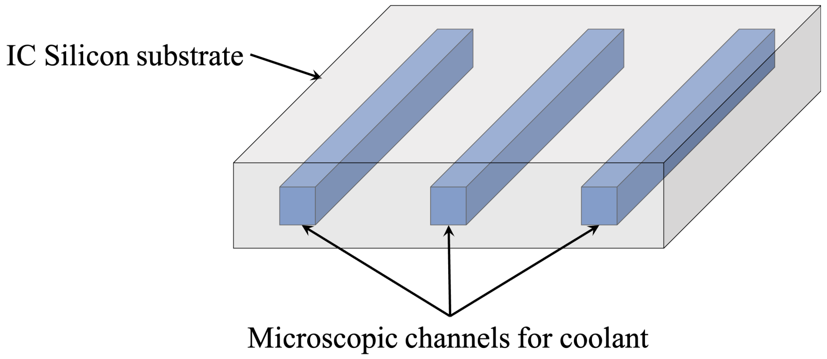


Figure 2.1: Microchannel-based inter-tier liquid cooling

3D ICs are characterized by unprecedented power densities, leading to exceptionally high heat generation that restricts performance. Traditional cooling techniques such as fan-based cooling or cold-plate-based liquid cooling can sometimes be insufficient to dissipate the high amount of heat produced by 3D ICs, particularly when increasing the number of stacked dies. Instead, liquid cooling using inter-tier microchannel heat sinks is now considered one of the most promising solutions to keep the temperature of 3D ICs under control. The idea of scaling liquid-based heat-exchange technology to microscopic dimensions was introduced 40 years ago as a potential enabler of high circuit power densities [23]. The technology involves etching miniaturized channels directly in the Silicon substrate of dies (Figure 2.1), where a liquid coolant (commonly water) is injected using a pump to absorb the generated heat via convection. By directly attaching microchannels to dies, the intermediate thermal resistances between the heaters and the heat sinks are significantly reduced, hence attaining heat removal capabilities as high as $800 W/cm^2$.

In more recent years, inter-tier liquid cooling technology has regained popularity, and it is now an active area of research. In fact, extensive studies are performed to validate the existing theory on heat transfer and to set benchmarks for the fabrication and testing of future 3D ICs with inter-layer liquid-cooling [24]. Several works also address the design and optimization of microchannels [25] [26] and the management of 3D ICs with inter-tier liquid cooling [27]. Other works address some of the challenges that this technology encounters. For instance, the authors of [28] explore different channel sizes and geometries to limit the thermal gradients along the channels and mitigate the resulting non-uniform heat distributions. Then, the authors of [29] propose to use other fluid compositions to improve the heat absorption capacity inside the microchannels (e.g., mixing water with nanoscale metals and metal oxides). Finally, the authors of [30] address the design of micropumps to achieve high flow rates, which directly impact the cooling capability of microchannel heat sinks.

Inter-tier microfluidic channels are more suitable for 3D stacks than other advanced cooling technologies. Indeed, even though liquid cooling using a cold plate [31] achieves high cooling capacity, it does not scale well when stacking multiple dies on top of each other. Similarly, cold plates that use thermoelectric (Peltier) cells [32] do not solve the problems related to the poor heat dissipation between 3D stacked IC layers. Finally, immersion cooling is only suitable for entire servers, limiting options to efficiently manage localized heat generation. In this regard, this thesis considers inter-tier liquid cooling as the primary solution to 3D thermal dissipation problems, provided by Flow Cell Array technology (introduced in Section 2.3).

2.2 Simulation of inter-tier liquid-cooled 3D ICs

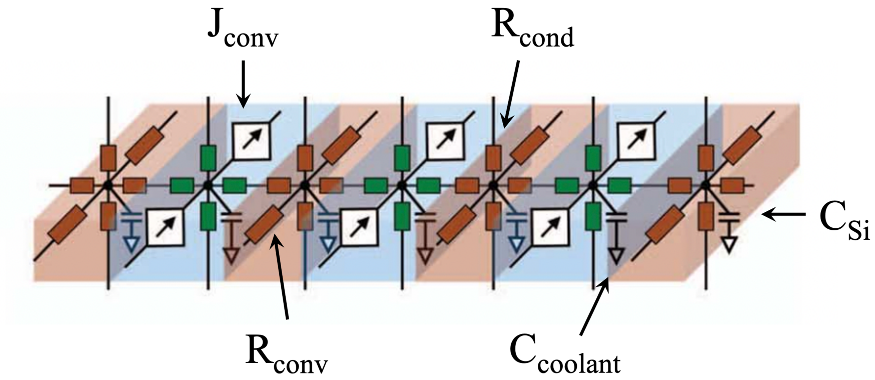


Figure 2.2: 3D-ICE model for a small section of a microchannel layer [33]

Fine-grained thermal simulations of 3D ICs are essential for measuring the effectiveness of inter-tier liquid cooling. To this end, the 3D Inter-layer Cooling Emulator (3D-ICE) is used throughout all the thermal experiments in this thesis to assess the performance of different 3D multi-processor architectures. 3D-ICE is ideally suited for the early-stage thermal-aware design of 3D computing architectures with integrated microfluidic channels because it has a complexity similar to conventional compact resistive thermal models for air-cooled ICs. In comparison, other available IC thermal simulators, such as HotSpot [34] do not specifically support inter-tier liquid cooling. Then, commercially available fluid dynamics simulators such as Ansys CFX [35] have others of magnitude slower simulation times.

3D-ICE is a compact modeling and analysis tool for the thermal simulation of 3D ICs with inter-tier microchannel liquid cooling [33]. 3D-ICE uses compact transient thermal modeling (CTTM) to significantly reduce simulation time and memory consumption compared to other fine-grained simulators for liquid-cooled ICs (e.g., Ansys CFX). The CTTM is constructed by identifying the equivalent electrical representations of heat conduction in solids and convective heat transport in fluid flows, starting from the governing heat transfer equations. Thus, the temperature is represented as a voltage, the heat flow as an electric current, the heat conduction of materials as resistance (R), the heat storage as a capacitance (C), and the convective heat transfer of microchannels as a voltage-controlled current source (J_{conv}). This way, the 3D IC is modeled as a mesh of cells connected through equivalent conductances between them (Figure 2.2), similarly to an electrical circuit. This model is then simulated to extract the temperature distribution across the chip.

An extensive evaluation has been performed using measurements from a real liquid-cooled 3D stack to validate the accuracy of 3D-ICE [33]. It is an open-source software in continuous development, with 3D-ICE 3.0 becoming available in 2022 [36].

2.3 Flow cell arrays (FCAs)

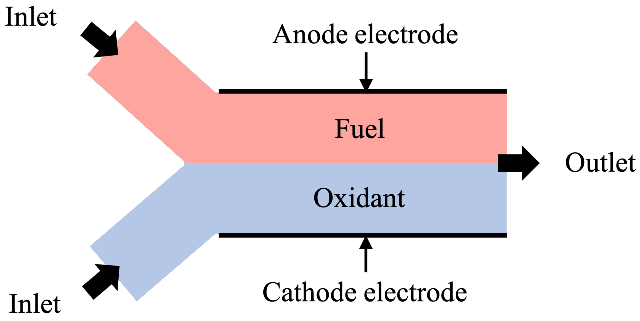


Figure 2.3: Planar view of a microfluidic channel with electrochemical flow and electrode contacts

Recently, a novel concept was proposed to extend inter-tier microchannel cooling with power generation capabilities [37]. This new technology replaces the non-conductive coolant (e.g., water) inside the microchannels with a solution containing electrochemical reactants responsible for power generation. At the same time, the fluid flow dissipates the heat generated by

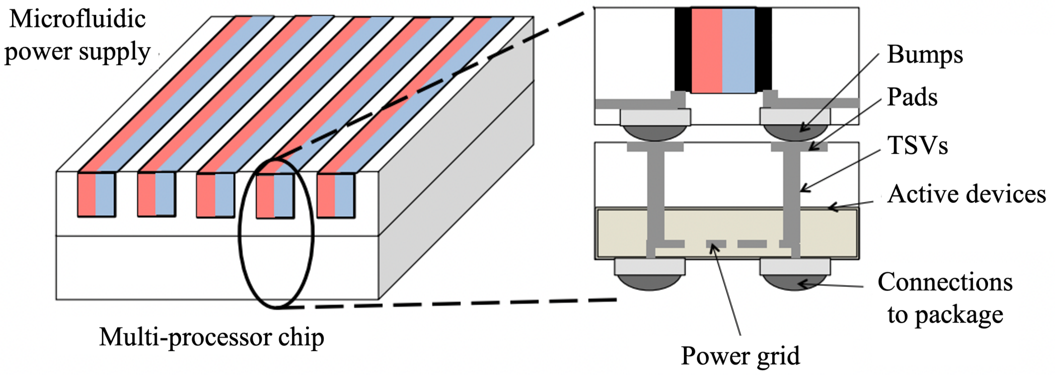


Figure 2.4: Architecture of multi-core processor with microfluidic power supply and cooling, as proposed in [37]

the encompassing chip. Therefore, both power supply and coolant are delivered through the same medium. This technology is referred to as microfluidic Flow Cell Array (FCA), and it has the potential to solve the challenges related to 3D IC design and disruptively improve the achievable power performance.

FCA technology uses a reduction-oxidation (redox) flow cell structure where two separate fluids flow inside a single microchannel in a co-laminar fashion, as illustrated in Figure 2.3. These two fluids are referred to as fuel (or reductant) and oxidant. The fuel contains chemical elements that transfer electrons to an electrode (anode in Figure 2.3), and the oxidant contains elements that receive electrons from another electrode (cathode in Figure 2.3). At each electrode, the corresponding redox species convert from an oxidized form (Ox) to a reduced (Red) form via the following electrochemical reaction (where n is the number of exchanged electrons):

$$Ox + n.e^{-} \rightleftharpoons Red$$

Hence, the electrolytes are responsible for storing energy, and the continuous redox flow provides a steady energy supply. In addition, the microscopic dimensions of channels result in a sufficiently small Reynolds number, enabling a co-laminar flow of the two electrolytic streams without convective mixing [38]. Therefore, there is no need for a membrane separating them and complicating the fabrication process. Different chemical compositions can be used for power generation inside flow cells, such as vanadium [39], polysulfide bromide [40], and iron redox [41]. In this thesis, vanadium redox is selected. This redox structure is the most commercially available and is commonly used in batteries. Furthermore, vanadium redox presents advantages over other chemistries: they are stable, achieve high concentrations, have negligible cross-contamination rates, are low cost, and most importantly, match the voltage and temperature window of high-performance CMOS systems [42].

The authors of [37] have proposed and evaluated the concept of a multi-processor 3D architecture with a vanadium-based microfluidic flow cell array, illustrated in Figure 2.4, where the fuel cells supply their generated power to nearby active circuits through TSVs and microbumps. The authors implemented analytical and numerical models of the fuel cells taking into account several design parameters such as channel dimensions, flow rate, and inlet temperature. Their results demonstrated the possibility of fully supplying the memory subsystem of a high-performance processor while efficiently cooling the whole chip. Alternatively, this thesis proposes to deploy FCA technology to solve the power dissipation problems in entire high-performance dies, including the parts with the highest power densities that suffer from the most critical voltage drops. This way, FCAs can increase the power performance across the chip while avoiding resistive losses related to transporting their power to specific loading areas (e.g., memories). Compared to other novel power delivery technologies, such as backside interconnect [22], FCA power can locally supply the gates without extra wiring. Moreover, this technology is not limited to face-to-face bonded 3D-ICs. Finally, it can scale with the number of stacked dies without requiring more TSVs to deliver their power, otherwise limiting die-to-die communication.

2.4 Electro-thermal simulation of 3D ICs with integrated FCAs

Detailed simulation of FCA power generation capabilities is necessary to understand the practical impact of integrating this technology as an additional power supply for 3D ICs. Thus, this thesis uses a novel tool called PowerCool to develop a compact thermal-aware electrical model of FCAs (Chapter 3). This model is integrated into the global power network of dies to assess the electrical performance for different target systems and usage scenarios. As of this writing, PowerCool is the only available electro-thermal simulator specifically targeting microfluidic cells for on-chip cooling and power generation.

PowerCool is a mathematical model introduced by the authors of [44] to simulate the electrochemical behavior of the integrated power-generating microfluidic flow cell arrays in 3D ICs. PowerCool uses the fundamental characteristics of the electrochemical reactions in vanadium redox flow cells to establish a compact model for each small section of the microchannel, depicted in Figure 2.5. This model comprises various components representing the phenomena and electrical properties involved in microfluidic power generation. Hence, the open circuit potentials (OCP) generated by redox reactions are represented as voltage sources near the electrodes (E_{OCP}). The potential losses due to the transport of ions near electrode surfaces are modeled as non-linear resistors (f_η), which limit the amount of current that can be supplied by flow cells. Then, the interaction between the changing potentials of various sections along the flow cell and the electrodes is modeled by a distributed resistance network (g_s, g_l) connecting the flow cell sections. Next, a resistance between the fuel and oxidant half-cells exists, representing the flow of ions between them (g_Ω). Finally, the accumulation of ions near charged electrodes acts as a capacitance during the flow cell operation, represented as a parallel capacitor (c_d). Given the circuit parameters described above, the flow cell circuit

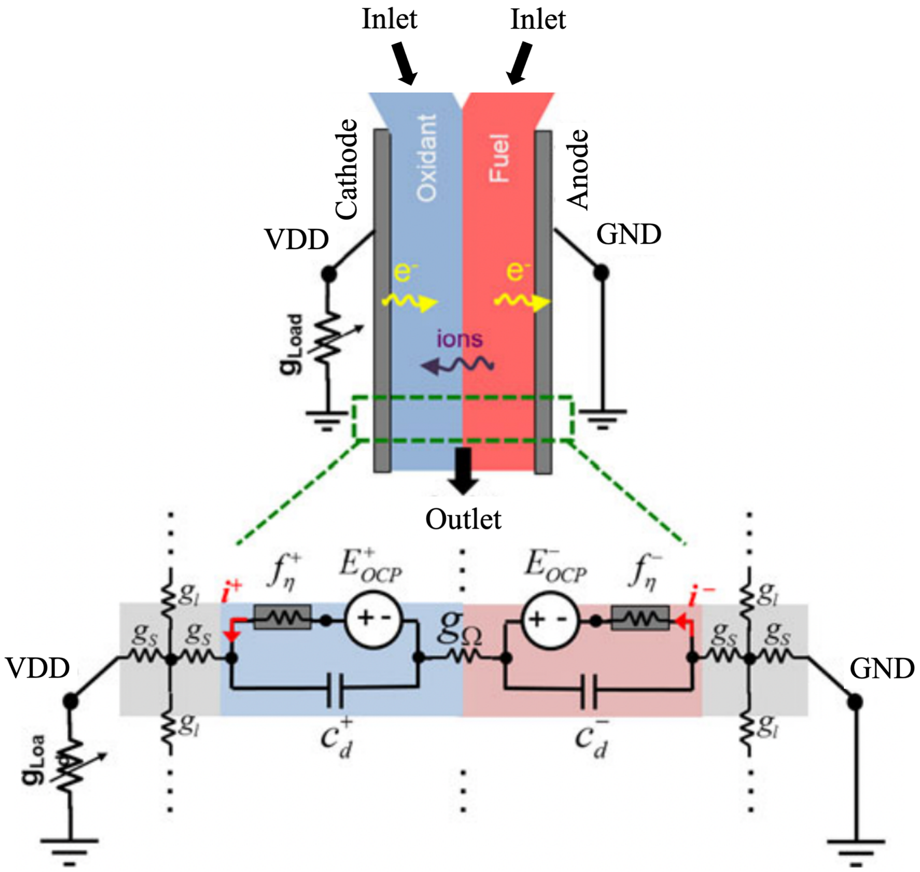


Figure 2.5: PowerCool compact model for microfluidic flow cell, proposed in [43]

equations are defined and combined to construct the global differential equations for the entire microchannel. These equations are then solved using techniques for non-linear systems to compute the voltage and current levels between the electrodes [44].

The accuracy of the PowerCool model was validated against fine-grained multiphysics simulations of individual flow cells and flow cell arrays using the COMSOL tool [45]. The model was also validated using measurement data from the literature. Then, it was applied to a multi-processor architecture with integrated microfluidic power generation and cooling, demonstrating the power generation capabilities of FCAs. In [43], 3D-ICE (Section 2.2) was incorporated in PowerCool also to evaluate the cooling capabilities of FCAs and analyze the effect of temperature on the electrochemical reactions of the redox flow. Hence, it was demonstrated that the generated power significantly increases with higher fluid temperatures, demonstrating FCA's capability to transform the heat generated by the chips into a valuable resource.

3 3D MPSoC design with integrated FCAs

3.1 Introduction

Over the years, the increasing demand for High-Performance Computing (HPC) has driven Integrated Circuit (IC) industry to increase logic density by scaling down CMOS transistors. Furthermore, significant innovations have been introduced and adopted to improve device performance [46][47]. Although aggressive technology scaling achieves high transistor densities and improves computing performance, system communication bandwidth remains limited as off-chip interconnect scaling lags behind.

In this context, 3D integration became a promising solution to overcome the interconnect scaling limitations of Multi-Processor Systems-on-Chip (MPSoCs) [48]. 3D integration enables tangible improvements in power performance and area [4]. It also enables designing heterogeneous systems in terms of technology and architecture, with massive communication bandwidths [49] alleviating the gap between computing and data access time.

However, 3D MPSoC design faces two major challenges: heat extraction and power distribution/delivery [4], which exacerbate as CMOS technology scales down. The difficulty of *heat removal* fundamentally increases as heat traverses several layers of low thermal conductivity before reaching the environment (i.e., heat sink). Thus, dies reach high temperatures resulting in an exponential increase in leakage [50]. 3D MPSoC *power delivery challenges* derive from supply voltage decrease along power metal lines, before reaching the gates, due to the resistivity of the delivery network. This phenomenon, referred to as IR-drop, manifests with local V_{DD} drops that result in slower transitions of the affected gates and lead to timing violations [51]. Many techniques have been proposed to improve heat dissipation and minimize IR-drop in 3D MPSoCs, by increasing the number of thermal [14] and power and ground (P/G) TSVs [51], respectively. However, these techniques generally involve a high TSV area overhead, making their implementation impractical and costly.

Integrated Flow Cell Array (FCA) technology addresses the challenges above by providing combined on-chip liquid cooling and power generation in 3D MPSoCs [43]. FCAs consist of micro-channels filled with an electrolytic solution etched in the silicon substrates of dies. They extend inter-tier liquid cooling [25] with power generation, enabled by the electrochemical reactions between electrolytes. Their capabilities were experimentally validated in [43], showing that they can generate up to $3.6W$ per cm^2 under optimal voltage and load conditions. However, previous analyses used an analytic model that did not cover important 3D MPSoC design aspects, such as the connectivity of the FCAs to 3D power delivery networks. This previous work also did not consider power generation variations between different flow cells with respect to temperature and voltage levels across the dies. Considering these aspects is important to accurately predict the effects of FCA integration on the power performance of 3D MPSoC.

Fulfilling the aforementioned requirement to predict FCA integration effects, this chapter proposes a novel thermal-aware 3D Power Delivery Network (PDN) modeling framework that analyzes the effects of connecting FCAs to 3D MPSoC power delivery (P/G) grids in detail. It can be used in early 3D MPSoC design stages to meet specific power and performance requirements. The framework uses real power maps to extract the load sinks, estimate the power grid densities of the dies, and calculate the FCA temperature maps using a fine-grained thermal analysis tool [33]. It then builds a fine-grained electrical model of the 3D P/G network and calculates the voltages across the dies. First, the proposed framework is used to assess the performance of a target system considering multiple configurations in terms of physical placement of dies, FCAs, and power TSVs. Then, the framework is deployed to predict the performance of FCAs as 3D MPSoC designers move towards deeply-scaled technologies with extreme power and thermal challenges. The summarized contributions of this chapter are as follows:

- The introduction of a fine-grained 3D P/G network model to quantify the effects of FCAs on 3D MPSoC power delivery. The model includes a compact electrical representation of FCAs and other power delivery components. Simulations of this model indicate that FCA technology outperforms other state-of-the-art power management methods. FCAs reduce the average IR-drop by up to 53% for dies with uniform power consumption and up to 30% for dies containing high power density regions (i.e., power hotspots). In addition, they do not require any additional TSV area.
- The analysis of the variation of FCA current generation along FCAs due to temperature and voltage gradients, using multiple real processor power traces. Simulations show that, while the average IR-drop reduction increases by 10% at the power hotspots, it remains 21% below the average value. Therefore, it is critical to have an accurate P/G delivery network modeling approach for 3D MPSoC design instead of qualitative profiling approaches.
- The introduction of an optimization algorithm using the proposed modeling methodology to find the best placement of P/G TSVs in 3D MPSoCs. The proposed algorithm ensures optimal FCA IR-drop reduction across dies. Moreover, it runs in under 4 min-

utes per 3D MPSoC sub-grid and per TSVs placement for a case study based on a high-performance processor, with 16 P/G sub-grids of 3362 nodes and 15 TSV placement options. These results prove the scalability of the optimization algorithm for large 3D MPSoCs.

- The analysis of FCAs sustainability as 3D MPSoC dies are fabricated using deeply-scaled technologies. Their on-chip cooling and IR-drop reduction benefits are quantified when scaling a computing die down to $3nm$. Using as a starting point power values of a $22nm$ multi-core processor (MCP) and $28nm$ Machine Learning (ML) accelerator, a novel methodology is proposed to estimate the leakage and dynamic power, at a constant area and operation frequency, following industry-reported [52][53][54][55] and predictive [46][47] scaling ground rules.
- The simulation of deeply-scaled 3D MPSoCs using the thermal-aware modeling and analysis framework, showing that FCAs can achieve 35°C lower temperature than off-chip liquid cooling strategies. In addition, FCAs eliminate up to 12% IR-drop when scaling the 3D MPSoC computing die from the $22nm$ to the $3nm$ process, compared to a direct liquid cooling scenario.
- The demonstration of the significant potential of FCAs for power and thermal management of 3D MPSoCs with advanced technologies. The results motivate the design of power-efficient next-generation 3D systems with FCAs. Such designs can save over 10% of the total area for power TSVs and up to two layers of metallization with respect to a state-of-the-art high-performance direct liquid cooling system [56][57].

The rest of the chapter is organized as follows. Section 3.2 summarizes the state-of-the-art of 3D integration, its challenges, and power delivery design and modeling techniques. It also provides an introduction to FCA technology. Then, Section 3.3 introduces the novel framework for modeling and analysis of 3D MPSoCs with integrated FCAs. The target 3D MPSoC used as an experimental vehicle is presented in Section 3.4. In section 3.5, the framework is used to evaluate the performance of FCAs, considering different 3D MPSoC configurations in terms of placement of dies and FCAs. The section also proposes a PDN optimization algorithm. The experimental results showcase the cooling and IR-drop reduction potential of FCAs. Next, Section 3.6 uses the proposed framework to analyze the scalability of FCA technology as the IC design industry moves towards ultra-scaled CMOS processes. It proves that FCAs remain an efficient solution for thermal and power management of next-generation 3D MPSoCs. Finally, Section 3.7 concludes the chapter.

3.2 Related work

3.2.1 3D integration trends and challenges

3D integration using TSV technology offers opportunities to increase the integration density of ICs, enable mixed technologies and architectures, and minimize the interconnect delays and power dissipation. 3D integration was successfully adopted in the memory industry with the High Bandwidth Memory (HBM) [6] and the Hybrid Memory Cube (HMC) [5]. Both achieve breakthrough bandwidths (over 256GB/s and 320GB/s, respectively) while consuming a lower energy-per-bit compared to 2D DRAMs. Following the emergence of 3D memories, efforts were directed toward memory-on-processor architectures to speed up memory-intensive applications and lower communication power consumption and delays [58][49]. However, these 3D MPSoCs face two significant challenges: heat extraction difficulty and voltage distribution complexity [20], particularly due to the non-uniformity of power consumption across chips.

In this context, methodologies were proposed to manage the thermal aspect of 3D MPSoCs in early floorplanning stages [59]. In particular, the insertion of thermal TSVs has been suggested to improve heat extraction in 3D MPSoCs [14]. Nonetheless, using thermal TSVs can have up to 47% total area overhead to reduce the maximal temperature of dies by 38%. Similarly, researchers proposed a technique to co-optimize the locations, sizes, and the number of P/G TSVs, as well as the floorplan of major 3D MPSoC blocks to prevent IR-drop constraints violations [51]. This method recovers up to 11.8% of IR-drop but uses a high number of TSVs (up to 200 TSVs for a grid of 300 nodes) of large diameter (20μm).

This chapter proposes accurately modeling FCAs and exploiting this technology to address the above challenges. Simulations indicate that FCAs achieve 42% better IR-drop reduction than state-of-the-art methods without additional power TSV requirements. In addition, FCAs improve the temperature of 3D MPSoC by 4% compared to traditional cooling methods.

3.2.2 Power delivery grid modeling

Power delivery analysis is critical in the early design and floorplanning stages of high-performance MPSoCs. As interconnect dimensions shrink and the currents traversing them increase, IR-drop becomes critical [51]. Therefore, it is crucial to accurately model the P/G grid to ensure a correct chip operation. A typical P/G grid consists of intersecting horizontal and vertical metal lines dedicated to voltage delivery [60]. Then, the lines are interconnected with each other and then connected to the gates using vertical vias. Each element has specific electrical resistivity and dimensions, which allows to estimate the resistances between the nodes and the sources. This model is commonly used for P/G delivery network design and analysis [61] [62]. In this chapter, it is extended by integrating FCAs as additional power supply components.

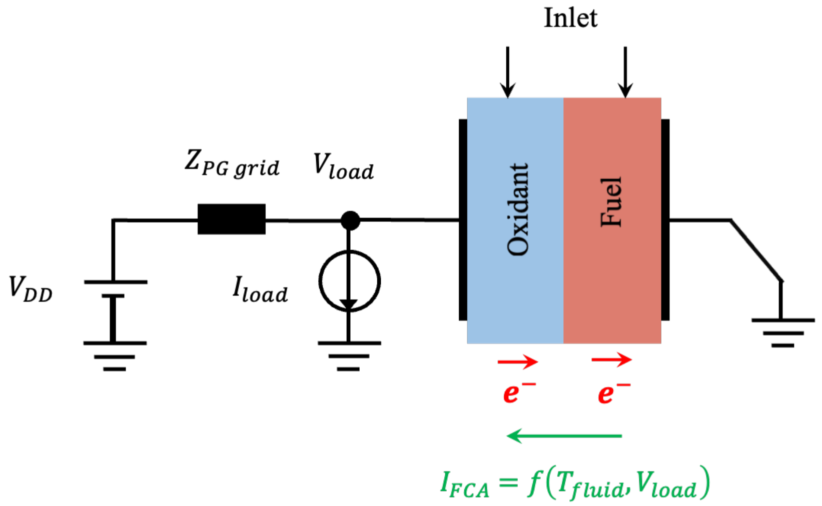


Figure 3.1: A single flow cell connected to the load and P/G grid of a 3D MPSoC die, as presented in [63]

3.2.3 Integrated FCAs

FCA technology consists of inter-tier micro-channels placed in the silicon substrate of 3D MPSoC dies, where an electrolytic solution flows. Similarly to on-chip inter-layer liquid cooling [25], they absorb the heat generated in the chip. At the same time, a continuous and stable electrical current is generated when a load is inserted between FCA electrodes due to electrochemical reactions between the electrolytes (Figure 3.1). The rate of these reactions increases with temperature, hence generating an electric current. Thus, FCAs effectively transform heat into available power to partially supply logic gates. Regarding the fabrication costs of 3D MPSoCs with FCAs, industrial collaborators expect it to be marginally higher (<10%) than inter-layer cooling. As this technology already requires the additional etching and coating processes that micro-scale FCAs require, only the positioning of electrodes to channel walls is needed as an extra step [37] [43].

Electro-thermal simulations of 3D MPSoCs with integrated FCAs can be performed using PowerCool [43]. It is a combination of the original PowerCool simulator [44], which enables compact electrochemical simulation of FCA power generation, and the 3D-ICE simulator [33], which analyzes the FCA cooling capabilities. The accuracies of PowerCool and 3D-ICE were validated against fine-grained physical models and actual measurements, respectively. The authors of [43] showed that FCAs could generate up to $3.6W$ per cm^2. Nonetheless, they considered optimal load and voltage conditions without examining the impact of connecting FCAs to 3D power delivery networks. Moreover, they did not study the current generation variations along FCAs due to voltage and temperature gradients. Hence, they missed evaluating the effects of power map non-uniformities. Thereby, this chapter accurately models 3D MPSoC P/G delivery networks with integrated FCAs and evaluates the effects of high-power density regions (i.e., power hotspots) on FCA IR-drop reduction capabilities.

3.3 Framework for modeling and analysis of 3D MPSoCs with integrated FCAs

This section presents a modeling and analysis framework to quantitatively assess FCA benefits on 3D MPSoC thermal and power performance. Hence, the framework models the temperature-aware power delivery network in fine-grain. It then measures the effects of FCAs on the voltage drop across dies. The framework enables exploring different dies, FCAs, and TSV configurations, quantifying their thermal and voltage drop performance.

3.3.1 Framework overview

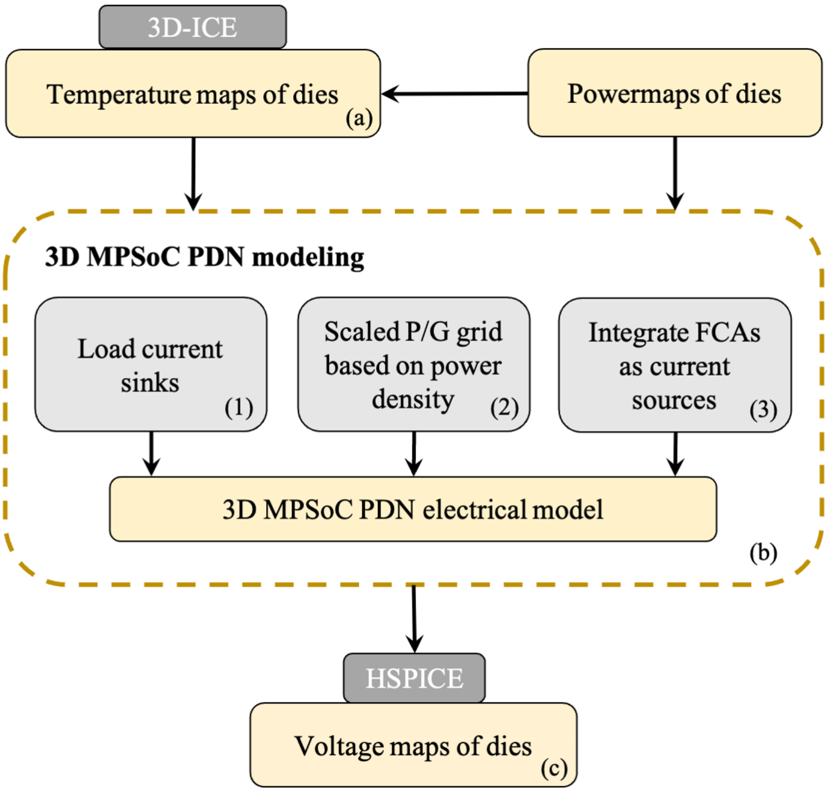


Figure 3.2: Thermal-aware modeling and analysis framework for 3D power delivery networks

The framework comprehensively evaluates the power generation performance of FCAs, and analyzes the voltage delivery across dies. The different steps are summarized in Figure 3.2:

- The power distributions of dies (i.e., different power maps) are used to evaluate their thermal behavior (box (a) in Figure 3.2). Using the fine-grained thermal simulator 3D-ICE [33], the framework extracts the temperature maps of the different 3D MPSoC dies.

- The power maps and temperature maps of dies are used to build a fine-grain model of the power delivery network (box (b) in Figure 3.2). In particular, the current sinks are calculated using the power consumption distribution of dies (box (1) in Figure 3.2). Then, the P/G metal lines and TSVs are modeled as a mesh of resistances estimated based on their physical properties and the power densities at the nodes (box (2) in Figure 3.2). Finally, FCAs are included in the model as additional voltage and temperature-controlled current sources (box (3) in Figure 3.2). Due to their stability [43][64], FCAs are directly connected to the P/G grids.

- The fine-grain electrical model of the 3D power network is analyzed using HSPICE (box (c) in Figure 3.2). As a result, the framework extracts the voltage maps of stacked dies and assesses the effect of FCAs on IR-drop reduction across the 3D MPSoC.

The following sections 3.3.2 and 3.3.3 describe in detail the proposed modeling methodologies for FCAs and 3D PDNs, respectively.

3.3.2 Flow cell electrical model

The framework models each flow cell as a voltage and temperature-controlled current source. It then connects its electrodes to the nearby source (V_{DD}) and ground (V_{GND}) metal lines in the P/G grid. Hence, the voltage between the metal lines generates a continuous and stable current between flow cell electrodes. This current is directly supplied to the gates through the P/G grid.

PowerCool [43] is used to assess the relationship linking the current and voltage between the flow cell electrodes (I-V curve). Moreover, this tool enables analyzing the effects of fluid temperature on FCA power/current generation. Figure 3.3 shows the I-V curve of a single flow cell of length $200\mu m$, for different fluid temperatures ranging from 27°C and 60°C. The selected length minimizes the temperature variation inside the flow cell (< 0.5°C) and the number of flow cells per channel (112 cells for the target 3D MPSoC presented in Section 3.4.1). The I-V curve is not linear in the complete analysis range but has a linear behavior around 1V, which is a typical V_{DD} value for sub-30nm technologies. Furthermore, the slope of the linear relationship in this range is independent of temperature. Instead, the temperature of the FCA fluid only affects the voltage/current offset of the I-V curve.

Following the above analysis using PowerCool, the flow cell at a position (i, j) is modeled using Equation (3.1). $V_{FCA,i,j}$ and $I_{FCA,i,j}$ are the voltage and current between the flow cell electrodes, R_{FCA} is the resistance of the flow cell and $V_{offset}(T_{fluid,i,j})$ is the voltage offset corresponding to the fluid temperature $T_{fluid,i,j}$ extracted from the temperature map:

$$V_{FCA,i,j} = V_{offset}(T_{fluid,i,j}) - R_{FCA} I_{FCA,i,j} \tag{3.1}$$

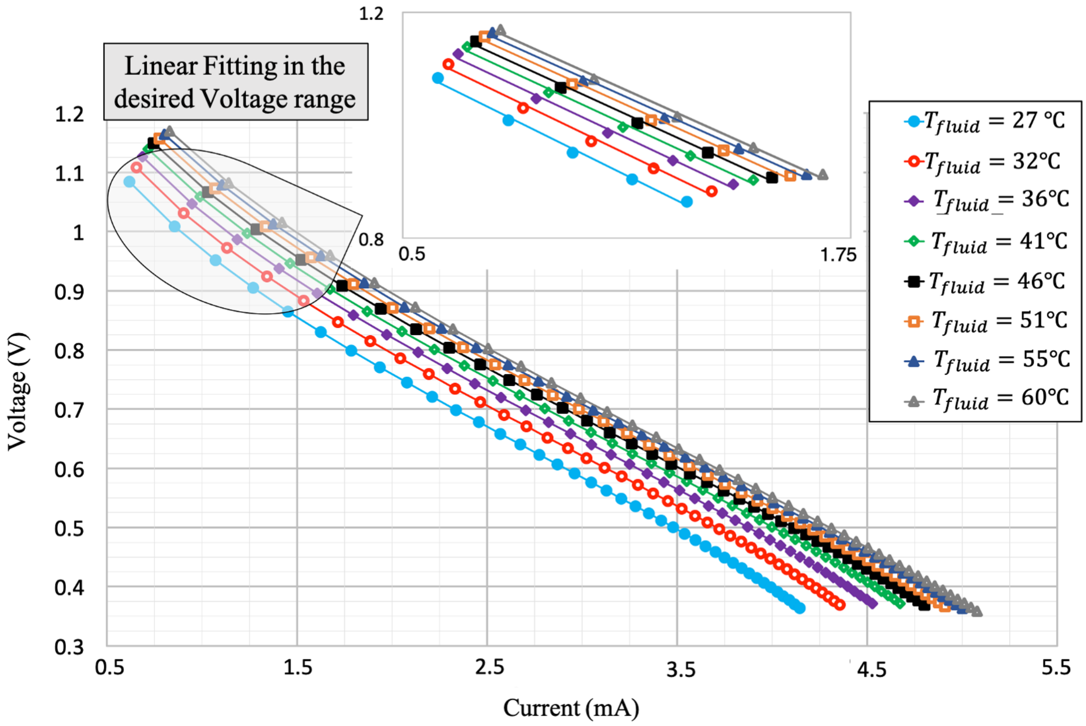


Figure 3.3: Current-voltage relationship (I-V curve) for a single 200μm-long flow cell

3.3.3 3D power delivery network modeling

The 3D MPSoC power delivery network model includes different components: the TSVs delivering power from the PCB to the stacked dies, the individual 2D power grids of dies and the FCAs connecting to the PDN. In order to perform a fine-grain simulation of the 3D MPSoC power network, the layers of the stack are partitioned into nodes representing cells with uniform power-per-surface units. These nodes are interconnected horizontally via the equivalent resistances of the 2D P/G grids. Then, they are vertically connected to the PCB via the equivalent TSV resistances. Each FCA cell is connected to the node it supplies, representing the closeby area of the chip that it covers. Consequently, a power delivery network is constructed for each stack layer, as shown in Figure 3.4.

The different PDN elements are modeled as follows:

- Current sources represent the flow cells (Box (3) in Figure 3.2). The output current is governed by the voltage between the FCA electrodes and the temperature of the electrolytic liquid, corresponding to Equation 3.1 in Section 3.3.2.

- Current sources represent the loads of the dies. Their values are extracted from the power maps of dies and the operating voltages at the corresponding nodes (Box (1) in

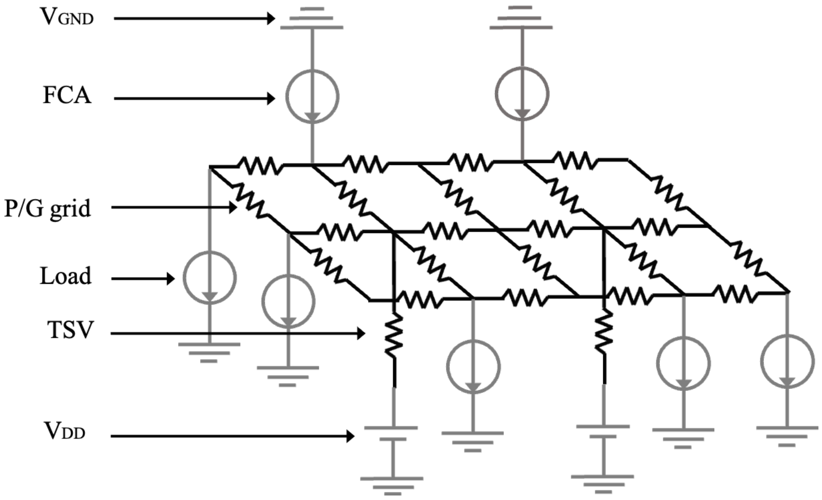


Figure 3.4: Electrical model of a 3D power delivery network

Figure 3.2). Hence, the load current at a node with coordinates (i,j) is:

$$I_{(i,j)} = \frac{P_{(i,j)}}{V_{(i,j)}} \tag{3.2}$$

- TSV resistances, which are a function of their length L_{TSV}, radius r_{TSV}, and the copper resistivity ρ_{Cu} [51]:

$$R_{TSV} = \frac{\rho_{Cu}\, L_{TSV}}{\pi\, {r_{TSV}}^2} \tag{3.3}$$

$$\text{with:} \begin{cases} \rho_{Cu}\text{: Copper resistivity} \\ L_{TSV}\text{: TSV length} \\ r_{TSV}\text{: TSV radius} \end{cases}$$

- 2D grid resistances, which are calculated as the sum of the equivalent resistances of the wires in all metal layers dedicated to voltage delivery [61]. The full equivalent resistance of a path between two nodes N_1 and N_2 in the power network is obtained as follows:

$$R_{N_1 \to N_2} = \textstyle\sum_{M_i} R_{M_i} \tag{3.4}$$

$$\text{with:} \begin{cases} M_i\text{: metal layer dedicated to power delivery} \\ R_{M_i}\text{: equivalent resistance of the wire in metal layer } M_i \end{cases}$$

The resistance of a wire between two nodes in a metal layer M_i is computed as follows:

$$R_{M_i} = \; p_{usage}\, R_{sM_i}\, \frac{L_{wire}}{W_{wire}} \tag{3.5}$$

$$\text{with:}\begin{cases} p_{usage}\text{: power grid usage} \\ W_{wire}\text{: wire width} \\ R_{sM_i}\text{: } M_i \text{ sheet resistance} \\ L_{wire}\text{: wire length} \end{cases}$$

The power grid usage refers to the percentage of available power delivery lines and vias. As the exact locations and routing paths between the gates are unknown during the pre-layout design stages, the power grid usage is scaled to the power consumption density at the nodes[61] (Box (2) in Figure 3.2).

The scalability of the 3D PDN model depends on its granularity (i.e., cell size) and other design parameters such as the number of power TSVs, FCA size, and pitch. However, the dominant parameter remains the cell size, as it determines the number of unknowns (i.e., the voltages at the nodes). Thus, it significantly impacts the size of the mesh of resistances. Consequently, the complexity of the model is *O(n)*, where *n* represents the number of nodes.

3.4 Target 3D MPSoC architecture

3.4.1 3D MPSoC and 3D PDN description

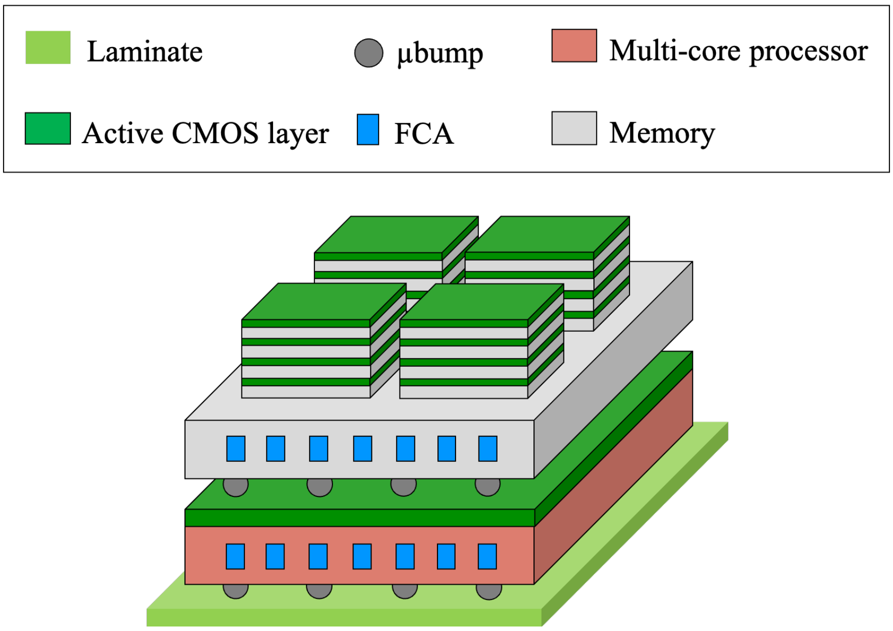


Figure 3.5: Target 3D MPSoC architecture

To test the proposed 3D MPSoC thermal-aware power delivery modeling and analysis framework, the stack composition from [43] is used as an experimental vehicle. Figure 3.5 presents an overview of the target 3D MPSoC architecture. It is composed of a multi-core computing layer and a memory layer.

The *bottom layer* is a processing die. Two separate architectures are considered to explore different power consumption profiles:

- First, the processing layer architecture is modeled after the server-class IBM POWER8 multi-core processor (MCP) [65]. Figure 3.6 represents the processor layout. The $946mm^2$ chip is fabricated using a 22nm Silicon-on-Insulator (SOI) CMOS technology. It is composed of 12 high-performance cores, sharing a 98MB eDRAM L3 cache. The maximal power consumption (TDP) of the processor is $190W$, and its maximal supported temperature is $90°C$. Given the large size of the POWER8 processor, HBMs in the corresponding 3D MPSoC memory layer are distanced from each other.

- Alternatively, the processing layer architecture is based on Google's Tensor Processing Unit (TPU) for ML applications [66]. Figure 3.7 shows the TPU floorplan [66] with a size of $300mm^2$ and a total power of $75W$, fabricated using the $28nm$ bulk CMOS process.

Its main components are the Matrix Multiply Unit and the Unified Buffer. For this 3D MPSoC configuration, HBMs in the top memory layer are placed closer to each other.

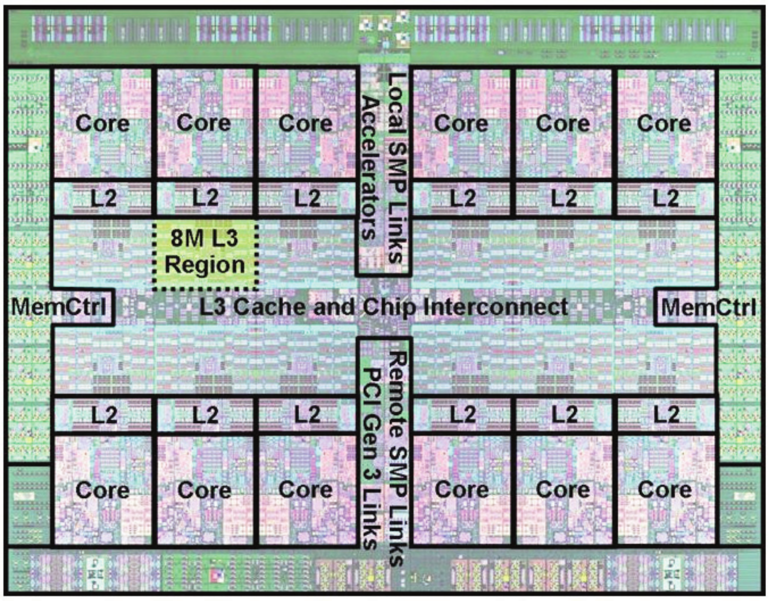


Figure 3.6: IBM POWER8 processor layout [67]

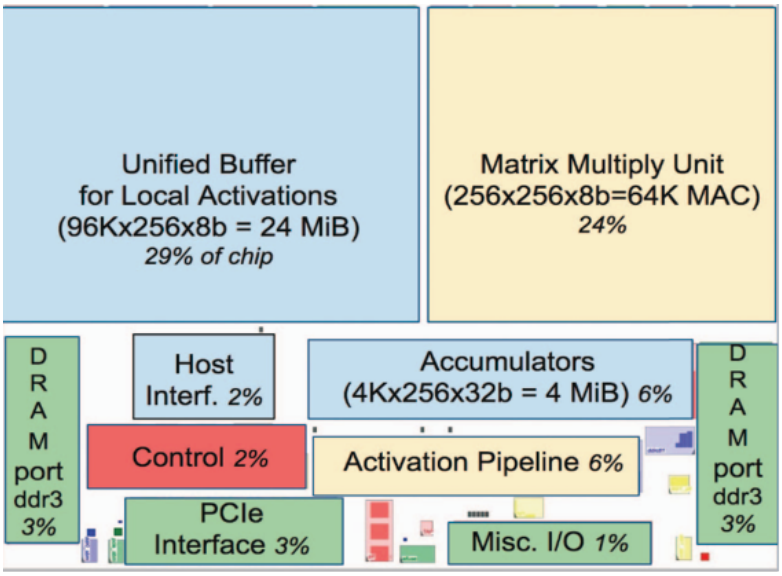


Figure 3.7: Google TPU floorplan [66]

The *top layer* is composed of four 3D DRAM memories. Each of the memories has the profile of a second generation $4GB$ 4-layer High Bandwidth Memory (HBM) [6]. The architecture of the HBM memory is represented in Figure 3.8. It contains a base logic die for the main interface between the DRAMs and the host computing die. Then 4 DRAM dies are stacked, with a total of 8 channels of $128 I/Os$. Each HBM memory has a base size of $71 mm^2$, fabricated using a $29 nm$ DRAM process. The maximal power consumption of a 4-layer HBM2 memory is $15 W$ [6].

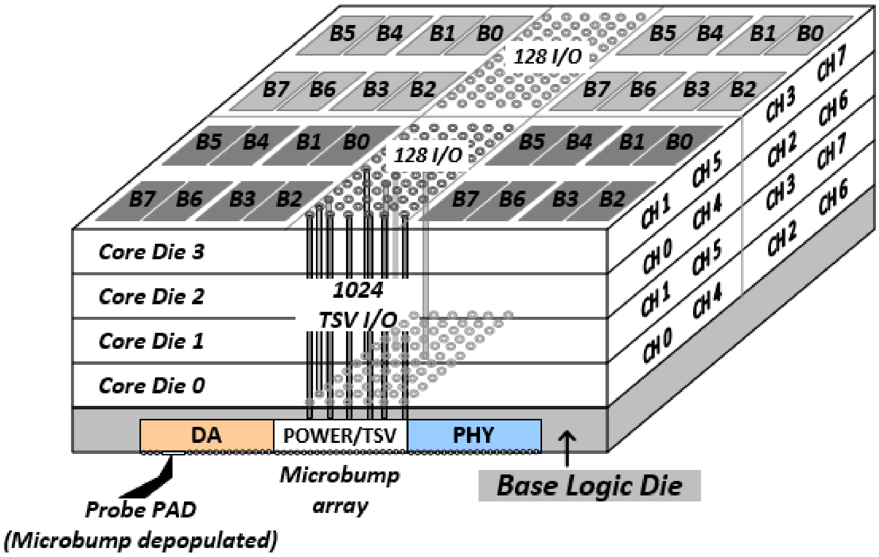


Figure 3.8: HBM DRAM architecture [6]

Table 3.1 shows the dimensions of the different 3D MPSoC power delivery network components. FCAs of $50\mu m$ width and $100\mu m$ height are etched in the silicon substrate of the 3D MPSoC dies, with a pitch of $100\mu m$ (Table 3.1). The TSVs used for power delivery and communication have a diameter and pitch of $5\mu m$. Their height corresponds to the CMOS bulk ($120\mu m$ in the presence of FCAs and $48\mu m$ without FCAs). Figure 3.9 shows the placement of the FCAs and power delivery (P/G) TSVs. Each die has regions called "TSV islands", where P/G TSVs are placed including the TSVs powering the die itself and those traversing it to power the above die. There are 16 TSV islands in total, and each of them delivers V_{DD} to the sub-grid that supplies its corresponding area. FCAs are etched in the remaining available area, considering that sufficient signal TSVs are placed to fulfill the communication bandwidth needs of the system. FCA electrodes are connected to the power delivery grid via the back-end-of-line (BEOL) of 3D MPSoC dies. Hence, FCAs directly supply their generated power to logic gates. FCA inlet temperature is set to 27°C, and the speed of the liquid to $2.5 m/s$. Finally, the dies are partitioned into nodes representing areas down to $40\mu m \times 40\mu m$, for the thermal and electrical simulations.

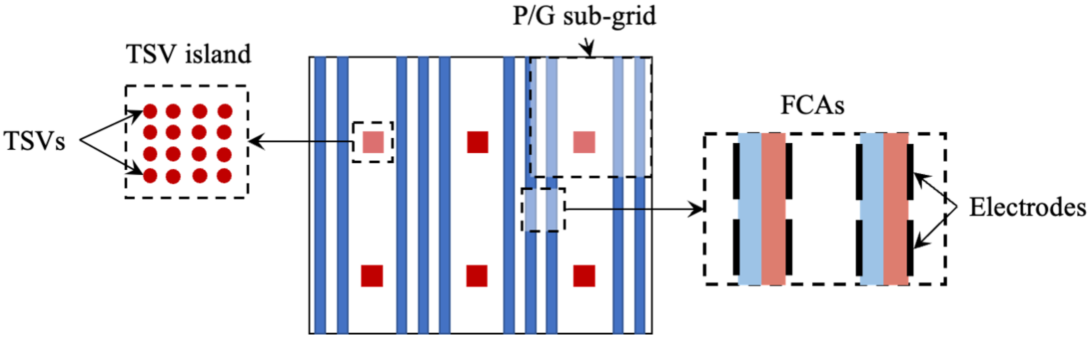


Figure 3.9: TSVs and FCA placement

	Width (mm)	Length (mm)	Height (μm)	Pitch (μm)
Die	28.9	22.3	$6^1 + 2^2 + 120^3$ $6^{1^*} + 2^{2^*} + 48^{3^*}$	- -
FCA	0.05	22.3	100	100
TSV	0.005		120, 48^*	10

[1] BEOL
[2] Active Silicon layer
[3] Bulk Silicon
[*] Without FCAs

Table 3.1: 3D MPSoC dimensions

3.4.2 3D MPSoC utilization scenarios

The cooling and IR-drop reduction performance of FCAs is evaluated for different utilization scenarios:

- In the case of the HBM memories, the power distribution is considered uniform across the die, as illustrated in Figure 3.10. Thus, a uniform access pattern is assumed at all times. Furthermore, each HBM memory has a total power consumption of $15W$ (corresponding to its TDP [6]), pessimistically assuming a maximal usage scenario.
- In the case of the MCP, four different non-uniform power maps are considered, as shown in Figure 3.11. These power maps correspond to real power traces of the POWER8 processor and have a total consumption of $190W$. In these different utilization scenarios, the cores switch between 3 activity levels: idle, nominal operation, and maximal frequency operation (performance boost). In power map (1), all cores are fully loaded and operate at the nominal frequency, whereas in (2-4), half of the cores are in performance boost mode while the other half are idle. In all four utilization scenarios, the MCP contains

multiple high power density regions (hotspots) concentrated around the computing cores. The rest of the chip has a significantly lower power consumption.

- In the case of the ML accelerator, the different components are assumed to have a uniform activity. Hence, the power map of the TPU is represented in Figure 3.12. It is extracted using an integrated power, area, and timing modeling framework [68]. The Matrix Multiply Unit, heart of the accelerator, has the highest power density and constitutes the biggest hotspot in the die.

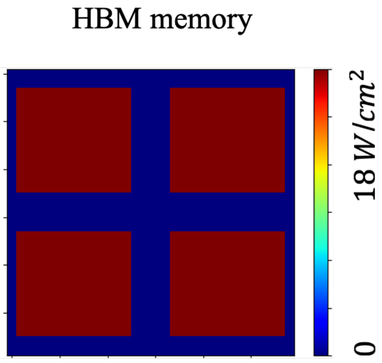


Figure 3.10: Power map of the HBM memory (uniform)

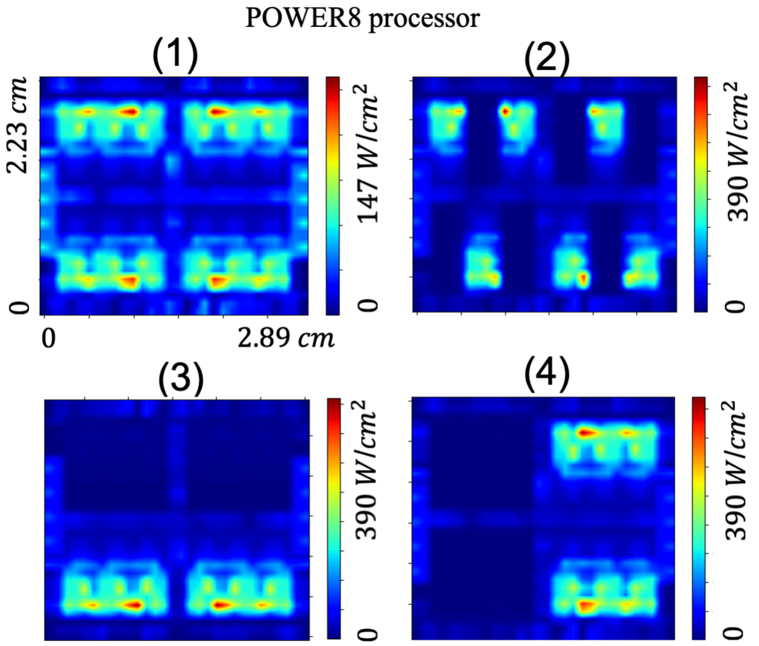


Figure 3.11: Power maps of the POWER8 processor (non-uniform)

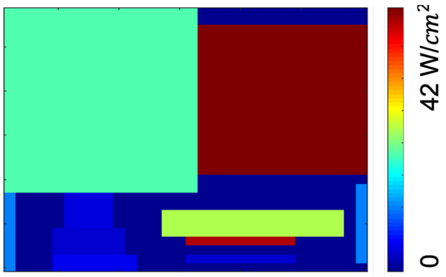


Figure 3.12: Power maps of the TPU accelerator (non-uniform)

3.5 3D MPSoC performance evaluation

In this section, the framework in Section 3.3 is deployed to analyze the performance of the 3D stack in Section 3.4. In particular, the architecture with the POWER8 MCP (Figure 3.6) is selected for the experiments as it consumes a higher power than the TPU, and therefore presents more critical thermal and power delivery challenges.

First, multiple 3D MPSoC configurations are explored by varying the placement of dies and FCAs (Section 3.5.1). These configurations represent different thermal conditions, thus affect FCA cooling and power generation capabilities. They are evaluated using the modeling and analysis framework to assess their thermal and voltage performance. The results of this analysis are presented in Sections 3.5.4.1 and 3.5.4.2. Then, an algorithm is proposed to explore multiple options for placing power TSVs in dies. Hence, the algorithm enables to identify the best TSV region according to a specific performance metric such as the IR-drop distribution (Section 3.5.2). The corresponding results are presented in Section 3.5.4.3.

3.5.1 3D MPSoC configurations

Similarly to [43], different 3D MPSoC configuration possibilities are explored in terms of die and FCA placement. Figure 3.13 presents the simulated stack configurations:

- Configurations A1, A2, A3 and A4 (A6 in [43]): FCAs are only etched in the memory dies.
- Configuration B1, B2 (B4 in [43]), B3 (B5 in [43]) and B4 (B6 in [43]): FCAs are present in both the processor and memory dies.
- Configurations C1, C2, C3 and C4: FCAs are only etched in the processor dies.

For all the above categories (A, B and C), multiple die arrangements and orientations are explored to assess the efficiency of FCA cooling and power generation in different scenarios.

The same configurations were analyzed in [43] to qualitatively score the FCA heat-removal efficiency and power interconnect requirements under maximal load conditions. This qualitative

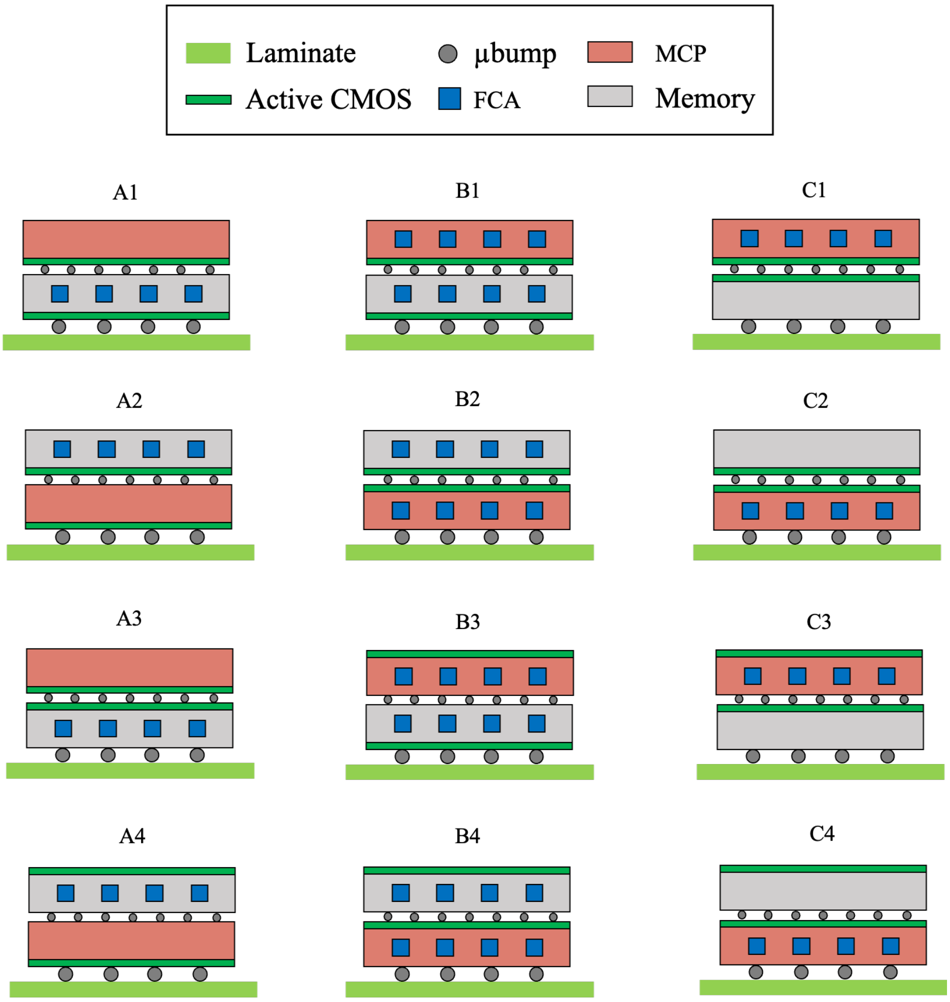


Figure 3.13: 3D MPSoC stack configurations

assessment of the stack configurations is shown in Figure 3.14. However, this analysis did not consider FCAs connectivity to the P/G network, nor did it model their local power generation variation due to the non-uniformity of temperature, power consumption of dies, and the non-optimality of the voltage between FCA electrodes. Therefore, a quantitative analysis of these 3D stacks is performed using the framework described in 3.3.

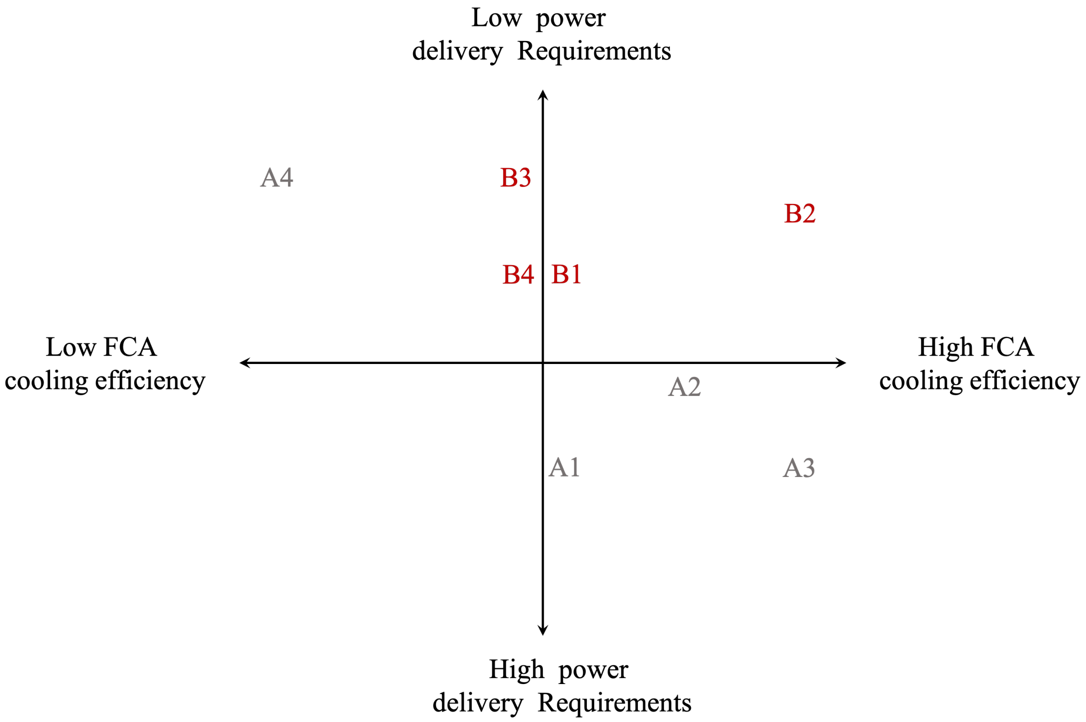


Figure 3.14: Qualitative assessment of 3D MPSoC stack configurations

3.5.2 TSV placement exploration

The cooling and power generation capabilities of FCAs depend on the power consumption profile of the target chip. Moreover, high-power-density regions (hotspots) affect voltage delivery across 3D MPSoC dies. Hence, the proposed 3D thermal-aware P/G network modeling and analysis framework can be employed to manage the effects of hotspots on power delivery performance. In particular, the framework is used to evaluate multiple placement options for power TSVs in the target 3D MPSoC. As TSV placement affects the ratio of power coming from the PCB versus FCAs for each node, the fine-grained analysis of different configurations is useful for choosing a TSV placement to fully exploit FCA power generation capabilities.

In this context, Algorithm 1 is proposed. This algorithm takes as input the 3D MPSoC architecture, power profile, and the desired PDN granularity N (i.e., the number of nodes to partition the power network). Thus, the algorithm searches for the best power TSV placement, minimizing the voltage losses across the PDN. An exhaustive search across all possible TSV locations is unfeasible for large 3D MPSoCs. Instead, the algorithm considers a predefined set of positions P where TSVs can be placed. The target position p_{target} is chosen from the set P, considering a quality metric Q defined to evaluate the power delivery network. For the experiments using the target 3D MPSoC (described in Section 3.4.1), the metric Q represents the maximum IR-drop variation with respect to the constraint value ΔV_{ref} (i.e., maximal allowed IR-drop).

For each position p in P, the algorithm builds a corresponding 3D MPSoC power delivery network model (line 3 in Algorithm 1), considering the effects of temperature on FCA power generation. Then, the algorithm performs a DC analysis to extract the voltage levels $[V_n]_{n \in N}$ at all the 3D MPSoC nodes (line 4 in Algorithm 1). Hence, the IR-drops $[\Delta V_n]_{n \in N}$ at the nodes are calculated with respect to their operating voltages $[V_{op,n}]_{n \in N}$ (lines 5-7 in Algorithm 1). Finally, the algorithm calculates the quality score Q_p at position p (line 8 in Algorithm 1). The target position p_{target} is selected corresponding to the best quality score Q_{best} (lines 9-12 in Algorithm 1).

Algorithm 1 PDN optimization

```
1:  Initialize $Q_{best}$ and $p_{target}$
2:  for $p$ in $P$ do
3:      Build thermal-aware PDN model (cf. Section 3.3.3)
4:      Run HSPICE DC analysis and calculate the voltage $V_n$ for each $n \in N$
5:      for $n$ in $N$ do
6:          Calculate the IR-drop $\Delta V_n = V_{op,n} - V_n$
7:      end for
8:      Calculate the quality score $Q_p = \frac{\max_{n \in N} \Delta V_n - \Delta V_{ref}}{\Delta V_{ref}}$
9:      if $Q_p < Q_{best}$ then
10:         $Q_{best} = Q_p$
11:         $p_{target} = p$
12:     end if
13: end for
```

3.5.3 Experimental flow

The thermal-aware PDN modeling and analysis framework in Section 3.3 is applied to the target 3D MPSoC in Section 3.4. Hence, the following steps are performed:

1. The 3D MPSoC thermal model is built using 3D-ICE [33]. In particular, the power distribution map of each die is constructed based on the power profiles of the target architectures. This model is then simulated to extract the fluid temperatures inside the micro-channels, which drive the current generation between flow cell electrodes (Section 3.3.2).

2. The 3D MPSoC power network model is constructed. Thus, the loads and P/G grids are modeled using the power maps of dies (Section 3.3.3). In the case of using FCAs, their compact electrical model is added as a Verilog-A module. The behavior description of the module corresponds to Equation 3.1 in Section 3.3.2. The temperature-dependent voltage offsets $V_{offset}(T_{fluid})$ are set for each FCA cell corresponding to its temperature.

3. Finally, a DC analysis of the 3D MPSoC power network model is performed both with and without FCAs by using Synopsis HSPICE to extract the voltages at the nodes. Hence, the percentage of recovered IR-drop is calculated with respect to the case with no FCAs.

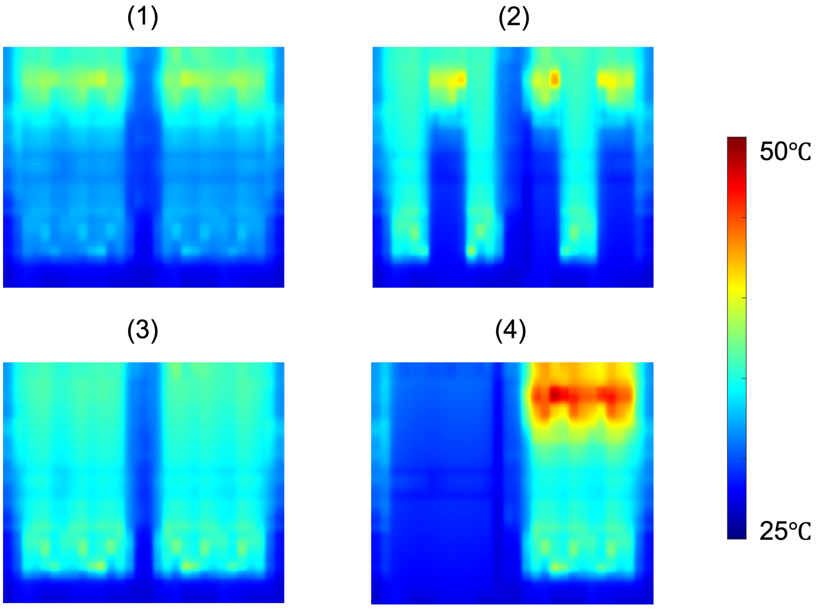


Figure 3.15: Temperature maps of the multi-core processor (configuration B2)

3.5.4 Experimental results

3.5.4.1 Evaluation of 3D MPSoC temperature using FCAs

As part of the framework in Section 3.3, the temperature maps of the 3D MPSoC dies (processor and memory) are evaluated using the thermal simulator 3D-ICE [33]. Thermal analysis of the stack is performed for all the configurations in Section 3.5.1, and under the different load scenarios of the POWER8 MCP die (Section 3.4.2). Figure 3.15 represents the temperature maps of the processor in the case of stack configuration B2, which was rated with the highest cooling efficiency in [43]. In all cases, the processor's temperature remains below 50 °C, which is 40°C lower than the maximal temperature in the case of fan-based air cooling [65]. Compared to using thermal vias to improve heat extraction, FCAs enable 4% more temperature reduction in the case of a similar 2-layer stack with 780× higher peak power density ($390W/cm^2$ versus $0.5W/cm^2$ in [14]). Hence, FCAs efficiently dissipate temperature without any additional area requirements.

Table 3.2 and 3.3 display the thermal simulation results for the multi-core processor and memory, respectively. In particular, the achieved peak temperatures of dies are measured for the different configurations and load scenarios. In general, configurations B1-B4 achieve the lowest temperatures due to the high cooling efficiency using FCAs in both dies. The temperature is also highest in the case of workload scenario (4) as FCAs traverse two hotspots with the highest power density. Thus the temperature of the electrolytic liquid increases. In workload scenario (1), FCAs also traverse two hotspots. However, in this case, the maximal power density is half the value of workload scenario (4), and the stack achieves lower overall temperatures. These results show that FCA cooling capabilities depend on their configuration

	Temperature (°C)											
Pmap	A1	A1	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4
(1)	52	52	54	51	39	40	39	39	53	52	52	53
(2)	61	61	66	58	43	44	43	43	61	60	59	60
(3)	50	50	56	48	38	40	38	38	51	50	49	50
(4)	-	-	-	-	49	50	48	48	-	-	-	-

Table 3.2: Peak temperature of the multi-core processor in the case of different configurations and usage scenarios

	Temperature (°C)											
Pmap	A1	A1	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4
(1)	51	55	50	52	39	40	39	40	52	51	52	52
(2)	69	68	62	61	42	45	42	44	59	59	58	60/59
(3)	48	56	48	50	37	41	37	40	50	48	48	48
(4)	-	-	-	-	48	51	48	50	-	-	-	-

Table 3.3: Peak temperature of the memory in the case of different configurations and usage scenarios

as well as the power consumption profile of the system. Hence, 3D MPSoC design must consider the target utilization scenarios to best utilize FCAs.

The results in Tables 3.2 and 3.3 also invalidate the qualitative assessment of the 3D MPSoC configurations in terms of FCA cooling efficiency (Figure 3.14). For example, configuration B2 was rated with a higher cooling efficiency than B1, B3, and B4. Conversely, fine-grain simulations indicate that this configuration achieves a slightly higher maximal temperature. Similarly, configuration A3 was rated with a higher cooling efficiency than configurations A1, A2, and A4. According to the results in Section 3.2, nevertheless, this configuration achieves the highest processor temperatures for all utilization scenarios. These results demonstrate the importance of preliminary fine-grain thermal simulation for designing 3D MPSoCs with integrated FCAs.

3.5.4.2 Evaluation of IR-drop reduction using FCAs

Using the proposed framework for PDN modeling and analysis in 3.3, fine-grained voltage maps are derived in the case FCAs supply power to 3D MPSoC dies. Voltage maps are also calculated if the power supply comes exclusively from the PCB, and no power is extracted from FCAs. Hence, the effects of FCAs on IR-drop reduction are measured for different usage scenarios for both the processor and memory dies. Figure 3.16 shows the percentage of FCA-enabled IR-drop reduction for the processor in configuration C1 and the memory in configuration A1. Both IR-drop reduction maps correspond to power map (1) in Figure 3.11. The configurations C1 and A1 are selected as they enable the highest FCA liquid temperature, hence power generation, for the multi-core processor and memory, respectively. Results show

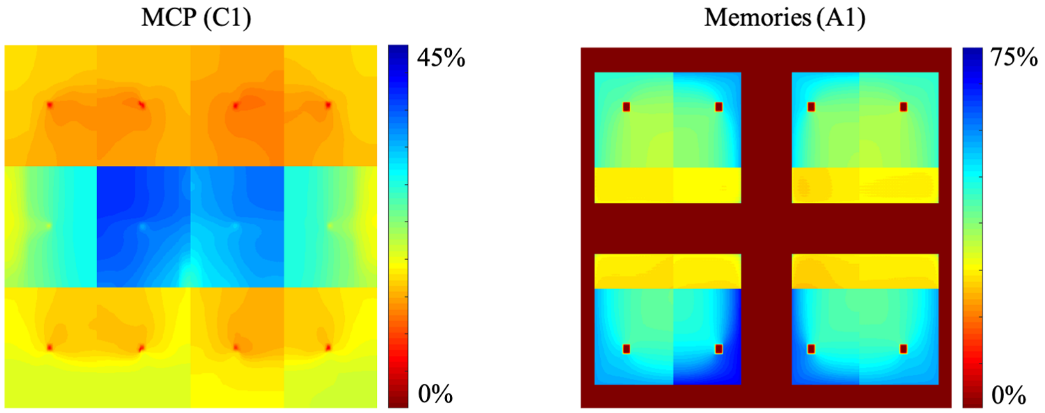


Figure 3.16: IR-drop reduction maps for the multi-core processor (*non-uniform power map*) and memories (*uniform power map*) in case of workload (4)

that FCAs enable up to 75% IR-drop reduction in the case of the memory and only up to 45% for the processor due to significantly higher power consumption in general.

Tables 3.4 and 3.5 present a summary of FCA power generation and IR-drop reduction results for the processor and memory dies, respectively. The tables show the percentage of total FCA-generated power with respect to the processor and memory consumption. According to these results, FCAs provide a significant amount of power to supply the dies. In particular, FCAs can provide up to 52% of memory power requirements and up to 19% of the processor's power budget. The FCA power supply for the memory is more significant as it generally has a lower consumption than the processor. Moreover, FCA current generation is highest in the case of power map (3), where the MCP has the highest hotspots. Thus, the 3D MPSoC generates more heat, increasing the current between FCAs electrodes.

Tables 3.4 and 3.5 also showcase the IR-drop reduction percentage at the most critical node (i.e., the node with the maximal IR-drop). In the case of the memory, results demonstrate that IR-drop reduction is proportional to the current generation but is unaffected which the usage scenarios (MCP power maps). Unlike the memory, the processor's power consumption profile includes *hotspots* affecting FCA IR-drop reduction efficacy. Thus, the percentage of IR-drop reduction in the most critical node of the MCP decreases up to 5% from power map (1) to (4), which have increasing hotspot concentrations.

Figure 3.17 presents the average and standard deviation of the FCA-enabled IR-drop reduction, calculated between the nodes. These results show that FCAs enable an average of 50 to 53% IR-drop decrease in the case of the memories and 23 to 30% in the case of the processor, with a standard deviation up to 14% and 18%, respectively. As indicated previously, the IR-drop reduction is higher for the memories as they consume in total 3× less power than the processor. In general, FCAs enable up to 42% higher IR-drop reduction than the maximum achieved

		Configuration							
	Pmap	B1	B2	B3	B4	C1	C2	C3	C4
FCA-generated power (%)	(1)	17	17	17	16	18	18	18	18
	(2)	16	16	16	16	17	18	17	17
	(3)	17	17	17	17	19	19	19	18
	(4)	16	16	17	16	-	-	-	-
IR-drop reduction at critical node (%)	(1)	13	13	12	13	14	13	13	14
	(2)	11	12	11	11	12	12	12	12
	(3)	9	9	9	9	9	10	10	9
	(4)	9	9	9	9	-	-	-	-
Average IR-drop reduction (%)	(1)	16	17	16	16	17	17	17	17
	(2)	14	14	14	14	15	15	15	14
	(3)	12	11	12	12	13	13	13	12
	(4)	11	10	11	11	-	-	-	-

Table 3.4: FCA power generation and IR-drop reduction for the multi-core processor die

		Configuration							
	Pmap	B1	B2	B3	B4	C1	C2	C3	C4
FCA-generated power (%)	(1)	49	49	49	49	45	45	45	45
	(2)	50	50	50	52	47	45	45	46
	(3)	52	52	52	-	47	46	47	47
	(4)	-	-	-	-	45	45	45	45
IR-drop reduction at critical node (%)	(1)	35	36	36	36	34	34	34	35
	(2)	37	37	37	37	35	35	35	35
	(3)	37	37	37	-	35	35	35	35
	(4)	37	37	37	-	33	34	33	34
Average IR-drop reduction (%)	(1)	49	49	49	49	45	46	45	46
	(2)	49	50	50	51	46	46	46	46
	(3)	51	51	51	-	46	46	46	47
	(4)	-	-	-	-	45	45	45	45

Table 3.5: FCA power generation and IR-drop reduction for the memory die

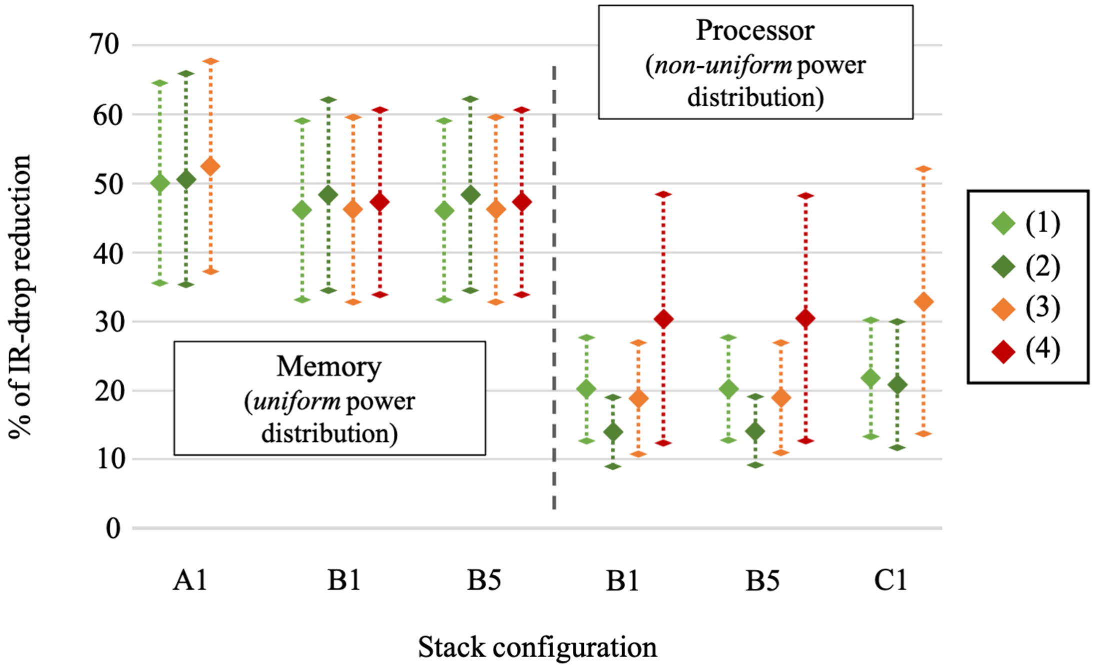


Figure 3.17: Average and standard deviation of IR-drop reduction for powermaps (1), (2), (3) and (4)

in [51] for a similar 2-layer stack with 4× lower peak power density. Similar to FCA cooling capabilities, the IR-drop reduction is achieved with no additional TSV area overhead compared to [51] and [14].

In addition, Figure 3.17 shows that the average IR-drop reduction in the processor increases up to 10% from power map (1) to (4), with a difference of 21% from the most critical node. In the case of power map (4). This power map contains the highest concentration of hotspots and thus reaches the highest temperature. In addition, FCA IR-drop reduction has the largest variation among all the nodes in this case. This observation indicates that *power map non-uniformities* affect the power generation of FCAs and the distribution of this power across the chip. In the case of a poorly designed P/G grid, less critical nodes can better benefit from FCA power generation, while high-power regions suffer from high voltage drops. Hence, these results highlight the necessity of the proposed analysis for 3D MPSoC power delivery network planning (e.g., TSV placement in Section 3.5.4.3).

Finally, our results differ from the quantitative profiling of some 3D MPSoC configurations presented in [43]. For example, configuration B1 was rated with a much higher score compared to A1 in terms of P/G interconnect requirements. However, fine-grain simulations show that it is not necessarily the case, particularly for power maps with critical hotspots (cf. Tables 3.4 and 3.5). A similar observation can be made between configurations B4 and A3 proposed in [43]. These observations prove that considering an optimal FCA power generation, a qualitative analysis is not sufficient for 3D MPSoC design with FCAs.

3.5.4.3 TSV placement evaluation

The algorithm in Section 3.5.2 is deployed to identify the best TSV placement to supply the multi-core processor. Specifically, configuration B1 is selected when the load corresponds to power map (4). This scenario presents the most critical IR-drop in the previous section. In addition, the processor PDN sub-grid containing the highest power hotspot is selected among the 16 MCP sub-grids, representing worst-case usage conditions. Hence, 15 different placements of the TSV island are tested, situated along the diagonal axis from the least power-consuming to the most active region (hotspot). The optimization algorithm is performed considering the quality metric Q representing the IR-drop variation with respect to the baseline position (i.e., TSVs at the center).

Additionally, a new quality metric Q^* is defined, representing FCA IR-drop reduction with respect to the case without FCA power extraction. The algorithm is scalable to larger 3D MPSoCs. In the case of the target 2-layer stack, it runs in under 4 minutes per position of TSVs, for a sub-grid of 3363 nodes. Each node represents an area of $40\mu m \times 40\mu m$.

Figure 3.18 shows the values of Q for different TSV island positions, and Figure 3.19 shows the values of Q^*. These results show that in Region A, where TSVs are farthest from the hotspots, IR-drop reduction is uniform but low on average. Moreover, IR-drop at the hotspots is highest (up to 21% more than the baseline). In Region B, the overall IR-drop decreases, and the IR-drop reduction increases for all nodes.

Finally, in Region C, IR-drop reduction is the highest and most uniform. However, while IR-drop decreases at hotspots, it significantly increases for less power-consuming areas (over 80% more than the baseline). This effect switches the location of critical nodes to parts of the die that are farthest from hotspots. Thus, the best placement of TSV islands, according to Algorithm 1, and considering both metrics Q and Q^*, is at the intersection between Regions B and C. This placement results in a low IR-drop for all the processor nodes and an efficient use of FCA current generation benefits.

The framework for 3D power delivery network modeling and analysis, presented in this chapter, enables accurate measurement of FCA cooling and power generation efficiency. Furthermore, it proves useful for early 3D MPSoC design stages. In particular, it highlights the effects of multiple design aspects on thermal and power performance. These aspects include the arrangement of dies, the placement of FCAs in the stack, and the placement of power TSVs. Hence, the proposed framework serves to make data-driven design decisions to ensure adequate functionality and efficient use of FCA capabilities. In the next Section 3.6, the framework is used to assess the sustainability of FCAs as 3D MPSoC dies scale down following IC design trends. Hence, a methodology is proposed to model the power consumption of computing dies using deeply-scaled CMOS technologies, aggravating the thermal and power delivery challenges. Then, FCA cooling and IR-drop performance is evaluated under different scenarios, proving it continues to represent an efficient solution for next-generation 3D MPSoCs.

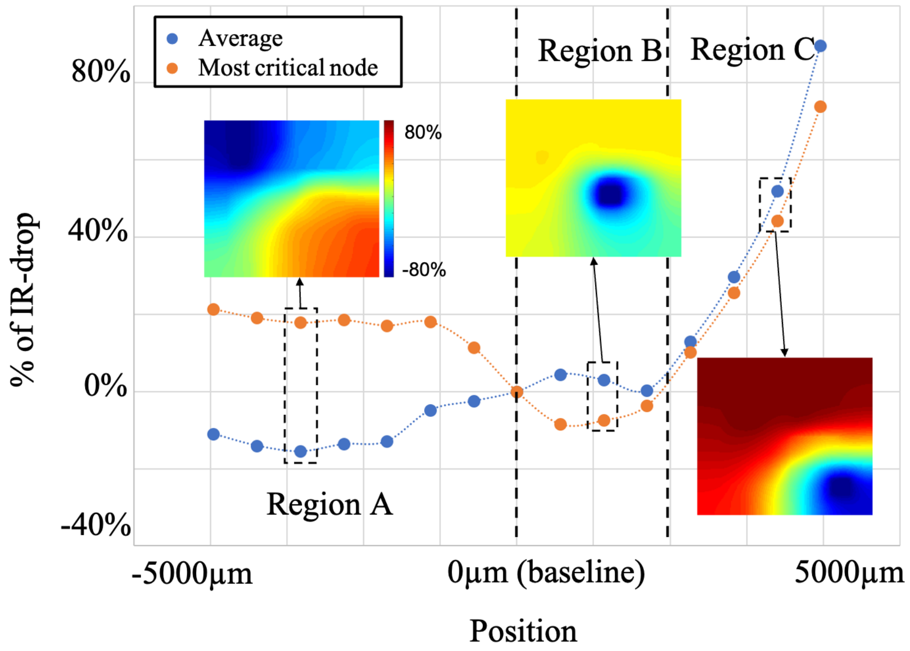


Figure 3.18: IR-drop variation with respect to the baseline (TSVs at the center)

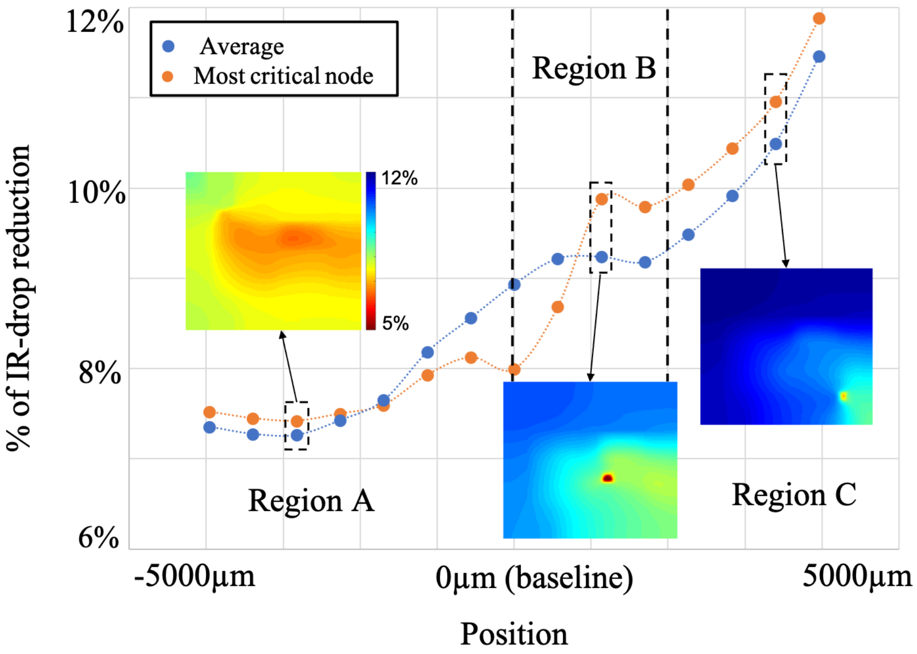


Figure 3.19: FCA-enabled IR-drop reduction with respect to baseline (TSVs at the center)

3.6 Evaluation of deeply-scaled 3D MPSoCs

Deeply-scaled three-dimensional multi-processor systems-on-chip enable ultra-high performance for next-generation computing. However, as process nodes shrink, temperature-dependent leakage dramatically increases, and thermal and power management becomes even more problematic.

As demonstrated in Section 3.5, FCA technology effectively solves the thermal and power challenges of 3D MPSoCs. In particular, FCA power generation capabilities depend on the liquid temperature and the voltage between electrodes. The first is impacted by the power density of dies, which increases as devices shrink. The latter is the chip V_{DD}, which changes corresponding to the process node. In this regard, electro-thermal analyses [44] show that FCAs of $100\mu m$ height and $50\mu m$ pitch can compensate entirely for the leakage of a die manufactured with the $7nm$ process node and cooled to 40 °C. For the $3nm$ technology, leakage can be entirely compensated by FCAs of $200\mu m$ height for the same channel liquid and die temperatures, as shown in Figure 3.20.

This next section evaluates how the IR-drop reduction and cooling capabilities of FCAs scale as 3D MPSoC dies are fabricated using deeply-scaled CMOS processes. A methodology is proposed to model the evolution of power consumption and power delivery network characteristics as device features shrink. Then, the thermal and PDN analysis framework in Section 3.3 is deployed to quantify the system-level impact of FCAs using technology nodes from $22nm$ to $3nm$.

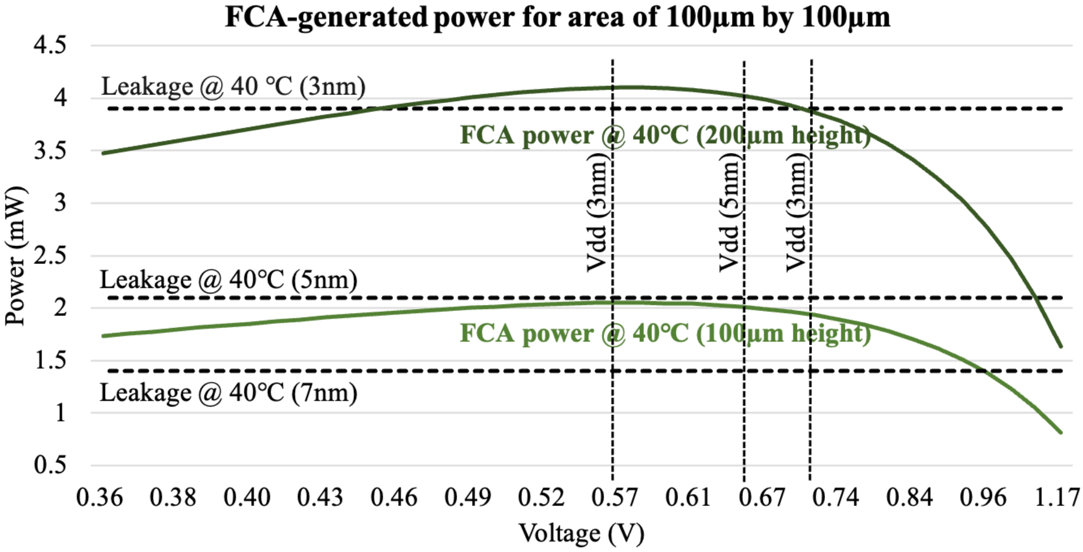


Figure 3.20: FCA-generated power with 40°C inlet temperature and $2.5m/s$ inlet speed, compared to the leakage power of the covered chip area

3.6.1 3D MPSoC scaling

3.6.1.1 Architecture overview

This section evaluates the scaling of FCA on-chip cooling and power generation capabilities when designing 3D MPSoCs with advanced CMOS technologies. In particular, the two-layer 3D MPSoC from Section 3.4 is used as an experimental vehicle. The two separate architectures are considered for the processing layer: POWER8-based (Figure 3.6) and TPU-based (Figure 3.7). The computing die (POWER8 and TPU) is scaled to smaller technology nodes to evaluate FCA performance in this case. A fixed area footprint is assumed for all scenarios by increasing the number and size of computing elements. Furthermore, a constant device switching activity (i.e., constant operating frequency) is considered. Hence, the power maps of dies are estimated for each process node, following the scaling methodology described in Section 3.3.

3.6.1.2 Technology scaling methodology

Technology	$28nm$ [69]	$22nm$ [52][70]	$14nm$ [53][71]	$10nm$ [54]	$7nm$ [55][46]	$5nm$ [46][47]	$3nm$ [46][47]
V_{dd} (V)	1	1	0.8	0.75	0.7	0.65	0.55
I_{off} $(nA/\mu m)$	20	20	20	20	20	20	20
W_{eff} (nm)	76	76	92	90	56.5	56.5	56.5
CPP (nm)	117	100	70	54	44	32	24
MP (nm)	110	90	70	36	24	20	12
$\rho_{transistor}$	1	1.43	2.63	6.62	12.19	2.11	44.69
$P_{dyn/device}$	1	0.55	0.3	0.17	0.09	0.05	0.03

Table 3.6: Technology scaling parameters

IC technology industry scales down the transistor and interconnects sizes to improve IC performance while maintaining constant power densities and low fabrication costs. In addition, new CMOS structures such as FinFET technology allow to drastically reduce transistor feature size and achieve high drive currents and low short-channel effects [52][53][54][55].

To evaluate the effects of CMOS scaling on the efficacy of FCA cooling and power generation capabilities, different process nodes are explored to scale down the target 3D MPSoC processing die (Figure 3.5). In particular, the following CMOS processes are analyzed: the $28nm$ bulk CMOS [69], and the $22nm$ [52], $14nm$ [53], $10nm$ [54], $7nm$ [55], $5nm$ [46] and $3nm$ [47] FinFET processes.

Several technology parameters are considered for computing the power performance of dies for a specific process node. Then, they are scaled based on industry-reported and predictive values shown in Tables 3.6 and 3.7. These technology parameters are the following:

- V_{dd}: The supply voltage of the chip. As CMOS devices shrink, it is scaled down to reduce power dissipation and maintain reliability.

- W_{eff}: The effective transistor gate width, which affects its drive current. For bulk CMOS, it is directly equal to the gate width W_{gate}. For FinFET, it is computed as a function of the fin width W_{fin} and height H_{fin}:

  $$W_{eff} = W_{fin} + 2H_{fin}. \tag{3.6}$$

- $I_{off}(t)$: The leakage current per transistor gate width. Devices are typically sized to achieve $10nA/\mu m$ leakage per gate width for low and medium performance, and $20nA/\mu m$ leakage per gate width for high performance [52], at the reference temperature of 25°C.

- $P_{dyn/device}$: The dynamic power consumption per logic device for a constant operating frequency. It is estimated by calculating the energy per device switching of a ring-oscillator circuit model:

  $$P_{dyn/device} = CV^2 \tag{3.7}$$

  Interconnect parasitics and transistor devices are coupled to estimate the total capacitive load C [46][47].

  The dynamic power consumption of the transistors decreases as technology scales down, to minimize the power density of chips.

- CPP: The contacted gate (poly) pitch represents the minimal distance between one transistor's gate to another. As technology scales down, this distance decreases, impacting the transistor density in an IC chip.

- MP: The minimum metal pitch also decreases as process nodes scale down, impacting IC transistor density.

- $\rho_{transistor}$: The transistor density scales with the contacted gate pitch and minimum metal pitch as follows:

  $$\rho_{transistor} \propto CPP \times MP. \tag{3.8}$$

  The values of $\rho_{transistor}$ in Table 3.6 are normalized to the value corresponding to the $28nm$ CMOS technology.

Using the above technology parameters, the computing dies power consumption evolution is analyzed as the fabrication process scales from the $28nm$ bulk CMOS down to $3nm$ FinFET:

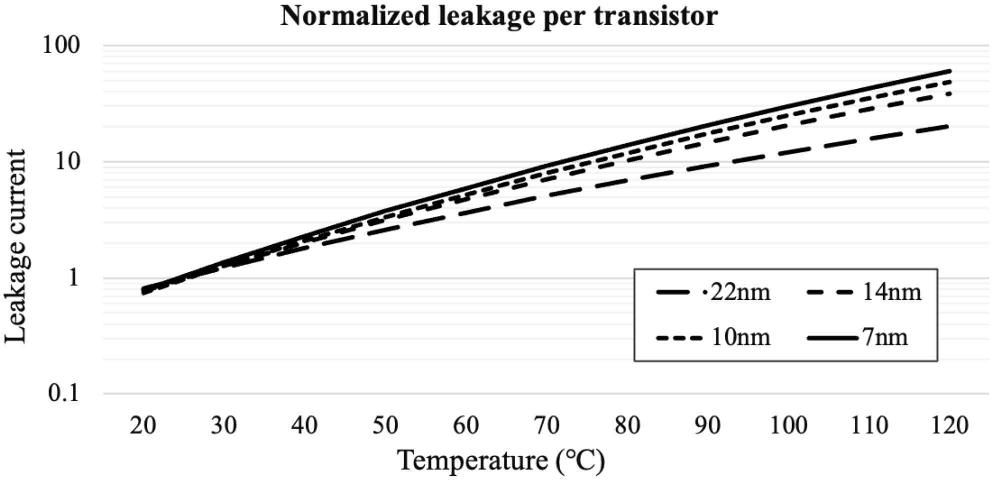


Figure 3.21: Normalized leakage per transistor, with respect to the value at 25°C, for different CMOS technology nodes

- First, the dynamic power scales according to the power consumption per device and the number of devices in the die. Starting from the total dynamic power of a die fabricated using a technology node n_0, the dynamic power of the die (of the same size) is estimated when fabricated with a technology node n. Hence, Equation 3.9 is used:

  $$P_{dyn}(n) = P_{dyn}(n_0) \times \frac{P_{dyn/device}(n)}{P_{dyn/device}(n_0)} \times \frac{\rho_{transistor}(n)}{\rho_{transistor}(n_0)} \tag{3.9}$$

  Similarly to the transistor density, the values of P_{dyn} in Table 3.6 are normalized to the value for the $28nm$ process node.

- Then, the leakage power of the chip at a temperature t is based on the leakage per transistor gate width, effective transistor gate width, and transistor density. Hence, leakage power is calculated according to Equation 3.10:

  $$P_{leak}(t) = (I_{off}(t) \times W_{eff} \times \rho_{transistor} \times Area) \times V_{dd} \tag{3.10}$$

  $$I_{off}(t) = I_{off}(25^{\circ}C) \times \alpha(t) \tag{3.11}$$

  The leakage per transistor gate width at the reference temperature I_{off}(25°C) is fixed for all technology nodes (according to Table 3.6). Thus, to evaluate the leakage at a temperature t, the temperature-dependency of the leakage $\alpha(t)$ is evaluated using predictive models (PTM) for sub-20nm technology nodes [72]. In particular, a NAND2 gate is simulated using different process nodes in HSPICE, assuming other gates behave similarly in terms of temperature-dependency. The leakages of the pull-up and pull-down networks are extracted. Figure 3.21 represents the normalized leakage current per transistor with respect to the reference value at 25°C, for different technologies and temperatures. Leakage scales exponentially with temperature and becomes extremely critical as transistor density grows. Hence, FCAs limit the overall power consumption of 3D MPSoCs with deeply-scaled temperature, as they significantly reduce temperature.

3.6.2 Experimental setup

3.6.2.1 Experimental flow

The thermal and power performance of FCAs is evaluated for next-generation 3D MPSoCs using advanced technologies. For this purpose, The 3D MPSoC configurations described in Section 3.6.1.1 are used, namely the POWER8-based and TPU-based 3D MPSoCs. Starting from the original multi-core processor and ML accelerator architectures, the computing dies are assumed to be fabricated with CMOS nodes with smaller feature sizes: $14nm$, $10nm$, $7nm$, $5nm$, and $3nm$ FinFET. Hence, the power distributions of the dies are scaled using the formulas in Section 3.6.1.2, according to the technology parameters of each CMOS technology. Furthermore, a constant die size is assumed by increasing the size and number of cores for the POWER8-based processor and the size of the Matrix Multiply Unit and Unified Buffer for the TPU-based ML accelerator. The throughput of the computing dies improved using adaptive load balancing, and task scheduling techniques [73][74]. Moreover, a constant dynamic switching activity per device (i.e., operating frequency) is considered when scaling dies to smaller technologies. Therefore, the 3D MPSoC thermal (3D-ICE) and electrical (HSPICE) models are constructed for each technology node and simulated to extract the temperature and voltage maps.

3.6.2.2 3D MPSoC cooling strategies

3D-ICE [33] is used to evaluate the cooling capabilities of FCAs. Additionally, their performance is compared against the following state-of-the-art off-chip cooling techniques:

1. A fan-based cooling system is modeled in 3D-ICE, achieving the POWER8 peak temperature of 90°C [65]. The cooling efficiency of the heat sink model is scaled to achieve the same maximal temperature when processing dies are synthesized with different technology nodes, within feasible limits [75]. The equivalent heat transfer coefficients are shown in Table 3.7.

2. Then, the direct liquid cooling solution of the Eurora Supercomputer [56] is modeled. Eurora is a heterogeneous platform with high-performance hardware components such as Intel Xeon E5, Intel Xeon Phi processor, and NVIDIA Kepler K20 GPU. The Eurora infrastructure uses cold plate-based liquid cooling. It can achieve a maximal temperature of 95°C for a Xeon processor running at a frequency of $3.1GHz$, with an average power density of $53W/cm^2$. The Eurora cooling system is also scaled to

Technology	Heat Transfer Coefficient ($W/\mu m^2 K$)			
	Fan-based Cooling		Eurora Cooling [56]	
	MCP	TPU	MPC	TPU
$28nm$	-	$5.5\ 10^{-9}$	-	$9.7\ 10^{-9}$
$22nm$	$6.1\ 10^{-9}$	$6.45\ 10^{-9}$	$9.7\ 10^{-9}$	$9.7\ 10^{-9}$
$14nm$	$6.7\ 10^{-9}$	$6.4\ 10^{-9}$	$9.7\ 10^{-9}$	$9.7\ 10^{-9}$
$10nm$	$8.7\ 10^{-9}$	$8.6\ 10^{-9}$	$13.5\ 10^{-9}$	$13.5\ 10^{-9}$
$7nm$	$9.5\ 10^{-9}$	$8.8\ 10^{-9}$	$15\ 10^{-9}$	$14.3\ 10^{-9}$
$5nm$	$9.5\ 10^{-9}$	$8.7\ 10^{-9}$	$15\ 10^{-9}$	$13.5\ 10^{-9}$
$3nm$	$10\ 10^{-9}$	$10\ 10^{-9}$	$23.5\ 10^{-9}$	$17.9\ 10^{-9}$

Table 3.7: Equivalent heat transfer coefficients for the MCP and TPU for different CMOS technologies and cooling techniques

maintain the original cooling efficiency when dies are synthesized using advanced technology nodes. This implies changing several possible design parameters, such as cold place dimensions and materials, refrigerant type, or coolant temperature [57]. The heat transfer coefficients of the equivalent heat sink models are shown in Table 3.7.

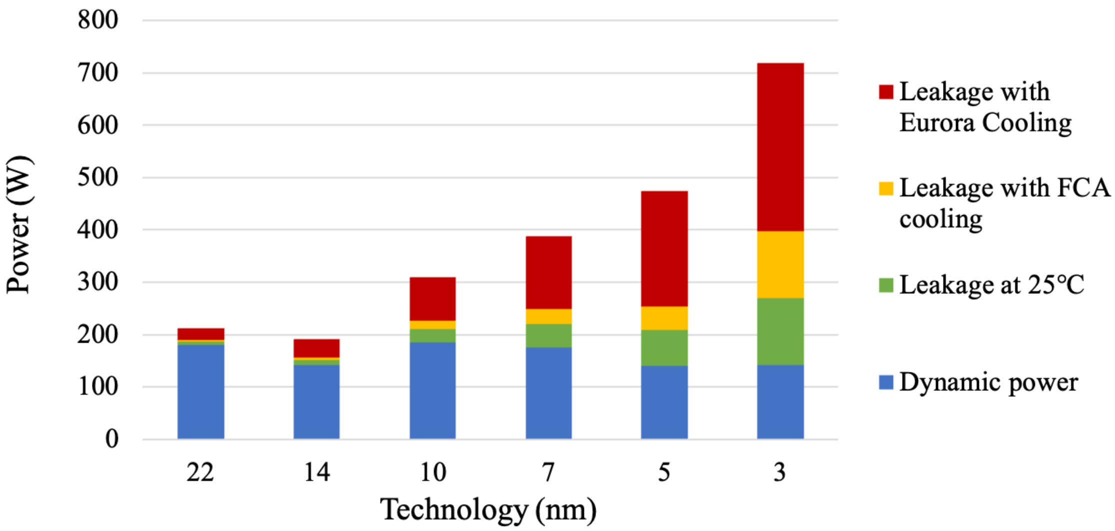
3.6.2.3 Power network modeling at advanced technology nodes

The POWER8-based and TPU-based dies powermaps are estimated when fabricated using technologies with smaller feature sizes: $14nm$, $10nm$, $7nm$, $5nm$ and $3nm$ FinFET. Dynamic power is scaled according to Equation 3.9 in Section 3.6.1.2. Leakage is estimated using Equation 3.10 in Section 3.6.1.2, according to the temperature maps when 3D MPSoCs are cooled using the different strategies in Section 3.6.2.2. Figures 3.22a and 3.22b show the total dynamic and leakage power of both dies at different technology nodes. Regarding MPC, power at the 22nm node corresponds to the consumption of the original POWER8 processor. Concerning the case of the TPU, power at 28nm corresponds to the consumption of the original accelerator. Power values for more advanced nodes are estimated according to the different scaling parameters in Table 3.6:

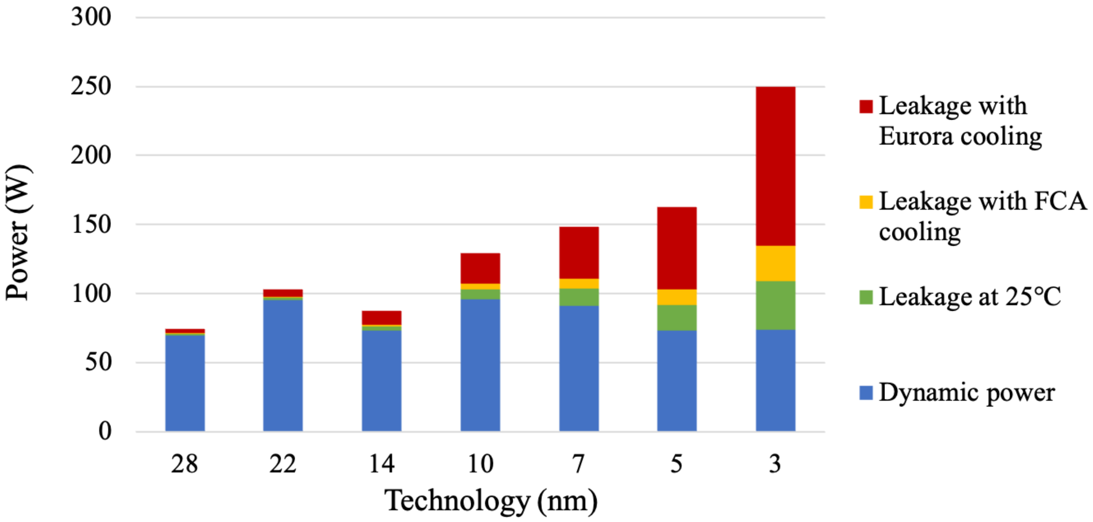
- The dynamic power (in blue in Figures 3.22a and 3.22b) is calculated according to the dynamic consumption per device and transistor density, assuming constant switching activity per logic gate. These two parameters have opposite scaling trends, resulting in a non-monotonic scaling of the total dynamic power.
- The leakage values in green in Figures 3.22a and 3.22b correspond to the scenario where dies are at the uniform base temperature of 25°C.
- FCA leakage (in yellow in Figures 3.22a and 3.22b) is the additional leakage calculated based on the temperature map of dies when cooled using FCAs.

- Finally, Eurora leakage (in red in Figures 3.22a and 3.22b) is the additional leakage that corresponds to the temperature of dies when cooled using the Eurora cooling system.

Moving towards ultra-scaled processes, chip leakage (of both POWER8-based and TPU-based dies) becomes predominant over the dynamic power, highlighting the necessity for highly-effective cooling structures for next-generation 3D MPSoCs.



(a) POWER8-based die



(b) TPU-based die

Figure 3.22: Power breakdown of the (a) POWER8 and (b) TPU-based dies for different technologies and cooling strategies

3.6.2.4 PDN modeling and IR-drop analysis

After calculating the power maps for different technology nodes, the fine-grain PDN modeling and analysis framework from Section 3.3.1 is deployed to evaluate the voltage distribution across the POWER8-based and TPU-based dies. A PDN structure is considered where power TSVs are arranged in groups, each delivering power to an independent sub-grid. Sub-grids correspond to individual cores in the case of the MCP. For the TPU-based accelerator, sub-grids correspond to the different components shown in Figure 3.7. FCA electrodes are connected to the power delivery grid via the back-end-of-line (BEOL) of 3D MPSoC dies. Finally, the power grid resistance is scaled according to the sizes of top metal layers dedicated to power delivery for different technology nodes [46][47].

3.6.3 Experimental results

3.6.3.1 FCA on-chip cooling capabilities

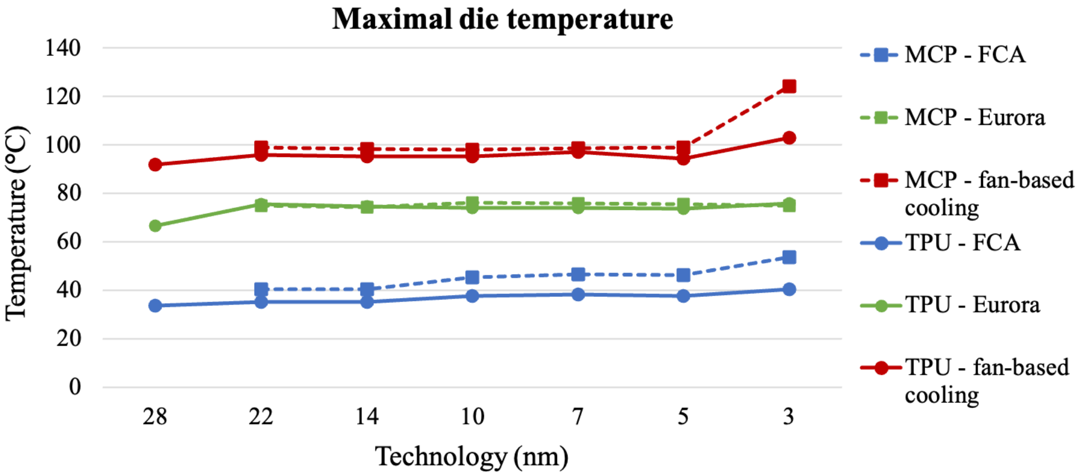


Figure 3.23: Temperature of the POWER8 and TPU-based dies with various cooling methods

The 3D MPSoC in 3.4.1 is simulated using 3D-ICE considering the two separate architectures of the bottom processing layer. In both cases, the thermal behavior is analyzed using FCAs, the Eurora supercomputer cooling system, and a fan-based heat sink model (scaled to meet the maximally allowed chip temperature constraints). Figure 3.23 shows the maximal temperature of the POWER8-based and TPU-based dies for different technology nodes and cooling strategies.

Without changing the FCA design parameters specified in Section 3.4.1 (i.e., channel dimensions, liquid flow rate, coolant inlet temperature), they can maintain a high cooling efficiency for ultra-scaled technology nodes, even as the power densities of the original dies double. FCAs keep the peak temperature of the MCP between 40°C and 53°C when moving from the $22nm$ to $3nm$ process node. For the TPU-based accelerator, with 22% lower average power density, peak temperature when using FCA cooling remains between 33°C and 40°C when scaling the die from the $28nm$ to $3nm$ node.

Compared to the heat sink model designed to meet the POWER8 thermal constraints, FCAs achieve 42°C to 71°C lower temperature for the POWER8-based processor. For the TPU-based ML accelerator, FCAs enable 41°C to 62°C lower temperatures between the different technology nodes.

In addition, FCAs generally outperform the Eurora supercomputer cooling system for both the TPU-based and POWER8-based dies. In particular, the FCA-cooled accelerator achieves up to 35°C difference in peak temperature than the Eurora-cooled die at the $3nm$ process node. Unlike FCAs, Eurora cooling system is scaled starting from the $10nm$ technology node (as shown in Table 3.7) to meet the original cooling efficiency designed for the Intel Xeon E5 processor fabricated with the $14nm$ lithography process. These results show that FCA technology, at its current advancement state, has the potential to be a first-choice cooling strategy for next-generation high-power-density 3D MPSoCs.

3.6.3.2 FCA IR-drop reduction capabilities

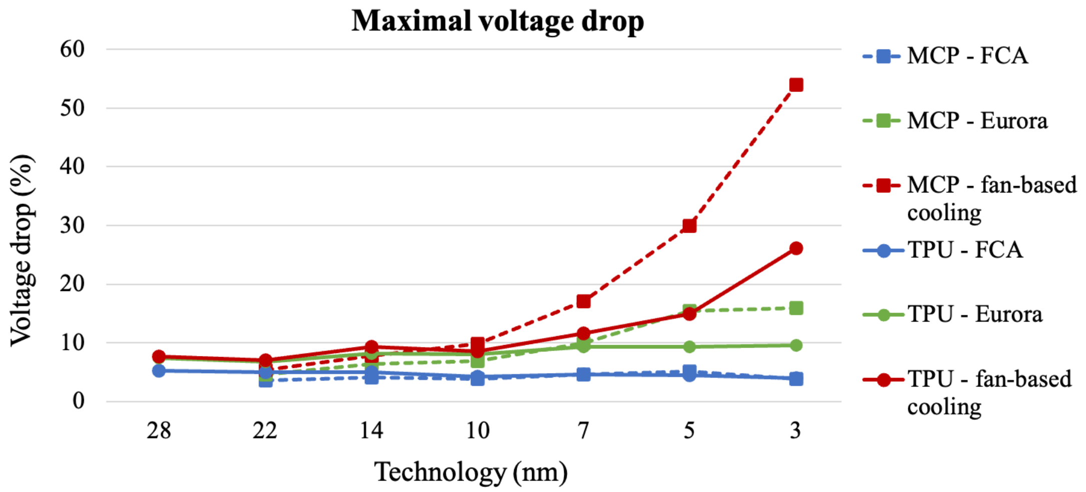


Figure 3.24: IR-drop of the POWER8 and TPU-based dies with various cooling methods

After simulating the thermal behavior of the target 3D MPSoC, the fine-grained power maps of the computing dies are calculated according to their temperature maps (for different cooling strategies). The full PDN of each die is modeled, with and without adding FCAs, and simulated using HSPICE to extract the voltage maps. Figure 3.24 shows the maximal IR-drop percentage of the processor and accelerator dies, with fan-based cooling, Eurora cooling, and FCAs. In the case of fan-based cooling and Eurora cooling, the source (V_{dd}) and ground voltages are only delivered from the printed circuit board (PCB) via TSVs supplying individual sub-grids. In the case of FCAs, the power supply comes from both the PCB and FCAs, which are directly connected to the grid. The simulation results support the following observations:

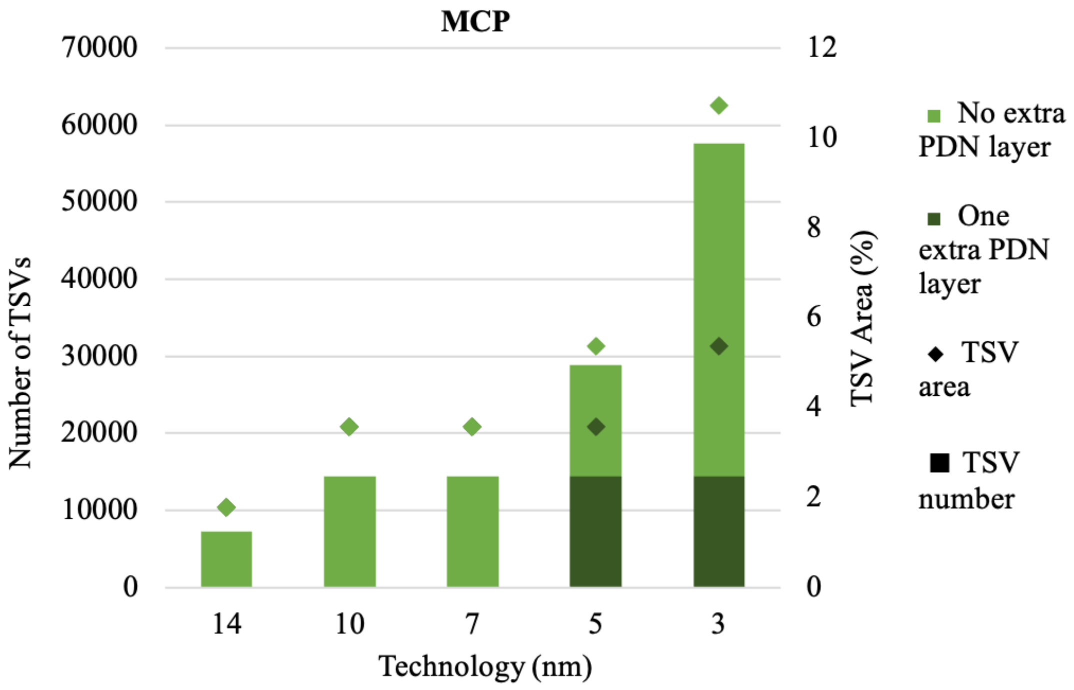
- FCA power enables to maintain IR-drop across 3D MPSoC dies under 5%, which is the typical IR-drop constraint for high-performance chips [76]. It is equivalent to under $27mV$ for the $3nm$ technology. The high FCA IR-drop reduction is due to two main factors. The first one is efficient cooling, which drastically reduces leakage. The second is the improved power generation capacity of FCAs as we move towards smaller-size processes. On the one hand, higher temperatures accelerate the reaction rate of electrolytes and hence improve power generation. On the other hand, lower operation voltages ensure up to double the generated power between the $28nm$ and $3nm$ nodes (Figure 3.21).

- For Eurora-cooled 3D MPSoCs, the IR-drop value of computing dies is between 5% for the original chip and 15% for the one scaled to the $3nm$ technology node. As power density increases substantially, Eurora system can be scaled to meet the thermal requirement of ultra-scaled 3D MPSoCs. However, power delivery system design requires significant improvements to mitigate IR-drop issues and ensure the functionality of such 3D MPSoCs.
- Fan-cooled dies achieve very high IR-drop values, over 50% for the $3nm$ node, due to exponentially-growing leakage. This demonstrates that IR-drop management becomes extremely difficult without highly-efficient cooling for 3D MPSoCs fabricated with ultra-scaled technology nodes.

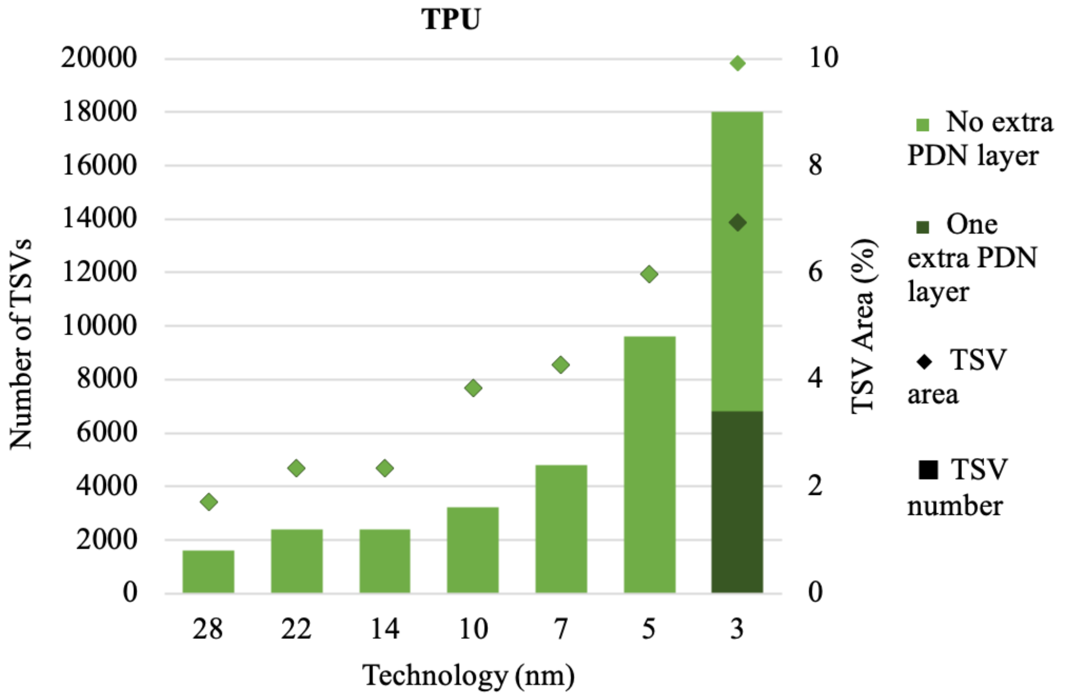
An additional experiment is performed to compare the PDN performance of FCA-powered and Eurora-cooled 3D-ICs. This experiment quantifies the additional power delivery component requirements to achieve the 5% IR-drop constraint in the case of the Eurora-cooled 3D MPSoCs. Figures 3.25a and 3.25b show the results for the multi-core processor and ML accelerator, respectively, for different technology nodes. In particular, the following two parameters are calculated:

- The number of power TSVs (represented by the bars in Figures 3.25a and 3.25b) and the total TSV area (represented by the scattered dots in the figures) needed to decrease IR-drop to meet design constraints. The number of TSVs includes the ones delivering V_{dd} and ground.
- If necessary, the number of additional power delivery metal layers required to meet the IR-drop target. In the case no additional metal layers are needed, results are shown in light green in Figures 3.25a and 3.25b. In the case an extra metal layer is needed, results are shown in dark green. The extra PDN layer is equivalent to 2 physical metal layers in the BEOL (vertical and horizontal alignment).

Simulation results show that designing highly-scaled 3D MPSoCs using Eurora cooling requires reserving over 10% of the area for power TSVs. This very high requirement also dramatically affects routing complexity, as TSVs need to be scattered across the die to reduce wire lengths. Alternatively, it is possible to meet IR-drop constraints by adding additional metal layers in the BEOL, which adds approximately 10% fabrication cost per layer.



(a) POWER8-based die



(b) TPU-based die

Figure 3.25: Number of additional metal layers, power TSVs, and total TSV area required to achieve 5% IR-drop for the POWER8 and TPU-based die with Eurora cooling

3.7 Conclusion

In this chapter, I presented a novel framework for thermal-aware power modeling and analysis of 3D MPSoCs with integrated FCAs. The framework constructs a fine-grained 3D network model including FCA power generation variation with temperature, and estimates the IR-drop across dies. It can be used at early design stages to achieve optimal power delivery while abiding by temperature and voltage constraints.

The proposed framework was used to evaluate the thermal and power performance of FCAs, integrated into a high-performance 3D MPSoC. In particular, it enabled the quantitative assessment of several 3D MPSoC die and FCA arrangements, a key feature in the early design stages. Then, a fast and scalable optimization methodology was proposed to find the placement of TSVs that best manages power hotspots and guarantees full exploitation of FCA power generation potential. The experiments indicated that FCA integration permits the reduction of the average IR-drop by up to 53% while maintaining temperatures below 52°C.

Next, the framework was used to evaluate the evolution of FCA on-chip cooling and power generation capabilities when designing deeply-scaled 3D MPSoCs. As technology feature size shrinks and device leakage exponentially grows, FCA cooling allows limiting the overall power consumption of 3D MPSoCs. The experimental results show that FCAs outperform a high-performance direct liquid cooling solution. They reduce the temperature of an MCP and an ML accelerator by up to 35°C, without changing their dimensions or chemical nature. Furthermore, they allow maintaining the IR-drop across the power grid of highly-scaled dies under 5%, saving over 10% TSV-reserved area and additional power delivery metal layers with respect to other cooling strategies.

This chapter has demonstrated that FCA technology, in its current state, represents a promising solution for 3D MPSoC thermal and power management. Indeed, this solution can be implemented without increasing the TSV count or the power grid density. Hence, the chapter clearly advocates for their integration in next-generation high-performance 3D MPSoC designs.

4 Design-time management strategies for 3D MPSoCs with FCAs

4.1 Introduction

Modern applications such as artificial intelligence (AI) and Big Data demand high performance, spurring a renewed interest in complex heterogeneous platforms combining diverse memory and computing elements. Additionally, wide communication channels are required to alleviate the gap between processing and data access speed. In this context, three-dimensional multi-processor systems-on-chip (3D MPSoCs) enable energy-efficient high-density computing and provide ultra-wide communication bandwidth requirements for next-generation applications [77]. However, 3D stacking exacerbates heat dissipation challenges as the number of stacked dies and the power consumption per surface unit increase. Indeed, 3D MPSoC temperatures are difficult to control using traditional cooling techniques, particularly due to the low thermal conductivity of bonding materials [78]. In addition, 3D integration complicates power delivery in multi-processor architectures due to the resistive losses in through-silicon-vias (TSVs) and metal wires [79]. Moreover, the large amount of power TSVs distributing voltage complicates routing, thus making 3D MPSoC physical design more difficult [80].

Flow cell array (FCA) technology, first introduced in [43], addresses the 3D thermal and power challenges mentioned above. FCAs consist of micro-fluidic channels etched in the silicon substrate of 3D MPSoC dies. They provide combined on-chip liquid cooling and power generation capabilities due to heat-accelerated electrolyte reactions. When connected to the power delivery network (PDN) of a 3D stack, their generated current partially supplies logic gates. Hence, they help reduce voltage supply (IR) drop, preventing timing violations that lead to performance degradation or system failure. As FCAs connect to 3D MPSoC power networks, their IR-drop reduction performance depends on both *power generation* and *cooling capacity*.

On the one hand, the *power generation* capacity of FCAs affects the amount of extra power supply to 3D MPSoC dies, which depends on the voltage between flow cell electrodes. In particular, peak power generation of a vanadium redox-based flow cell is achieved when operating between $0.55V$ to $0.62V$, depending on temperature [43][44]. As the voltage supply (V_{dd}) of state-of-the-art high-performance systems is generally above $0.8V$ [65], directly connecting FCA electrodes to 3D PDNs leads to sub-optimal power generation performance. Moreover, higher fluid temperatures accelerate the electrolytic reactions inside the channels, improving power generation efficiency.

On the other hand, the *cooling capacity* of FCAs significantly improves leakage, particularly for highly-scaled CMOS technologies. Hence, FCAs enable decreasing voltage losses across 3D MPSoC dies. In particular, high flow rates and channel numbers enhance heat absorption, considerably decreasing 3D MPSoC temperature and leakage. Conversely, lower flow rates and channel densities improve power generation by increasing FCA fluid temperature.

In this context, integrated cooling and power delivery solutions based on FCAs expose a novel and multi-faceted design space encompassing inter-dependent thermal and electrical considerations. This chapter proposes to tackle some design aspects specific to FCAs, targeting to improve their efficacy in high-performance 3D MPSoC while considering the trade-offs between cooling and power generation. The first part of the chapter addresses the voltage between FCA electrodes, which affects power generation. Thus, it proposes using on-chip switched-capacitor (SC) voltage regulators as an interface between FCAs and 3D MPSoC PDNs. The SC converters enable optimal operation of FCAs, by providing a stable voltage that leads to maximal power generation performance. Then, the second part of this chapter addresses the physical design parameters of FCAs (channel densities and coolant velocities) and their associated SC converters. These parameters affect both the cooling and power generation capacities. Hence, this part investigates their interdependence from a design-time perspective. In particular, it characterizes the performance of different 3D MPSoC configurations, highlighting the existing design trade-offs and showcasing the opportunities for power performance improvement enabled by FCAs.

In summary, the contributions of this chapter are the following:

- A direct current to direct current (DC-DC) converter is designed to supply a stable voltage to FCAs, leading to optimal on-chip power generation performance. Switched capacitor technology is used, and different design-space parameters are explored to achieve high power density and low area requirements [81][82].

- Using the proposed voltage regulator, FCAs generate up to 2x more power compared to the case when they directly connect to the PDN of a high-performance 3D MPSoC operating at $1.1V$. Moreover, optimizing DC-DC converters to connect multiple FCA cells reduces the overall additional area requirement to less than 1.26% while maintaining IR-drop across the chip under the 5% constraint of high-performance ICs.

- FCA connectivity to 3D MPSoCs power networks is controlled using SC converters to prevent unnecessary electrolyte reactions during chip inactivity. Hence, switching off FCA power generation extends by up to 1.8× the FCA reservoir lifetime for a 3D MPSoC duty-cycle of 50% and over 4.5× for a 20% duty-cycle.

- A power and thermal design-time exploration of 3D MPSoCs with integrated FCAs is proposed. Thus, multiple design configurations with integrated FCAs and SC converters are considered. Then, fine-grain modeling is used to measure their thermal and power performance and discuss entailed trade-offs.
- Targeting a high-performance 4-layer 3D MPSoC system, FCAs can reduce die temperatures by 78°C and power consumption by 46%, compared to a high-performance cold plate-based liquid cooling. Moreover, FCA-generated power can recover between 70% and 90% of voltage drop using SC converters occupying less than 3% of the total chip area.

The rest of the chapter is organized as follows. Section 4.2 introduces 3D MPSoCs, their challenges, and state-of-the-art management techniques used to alleviate them. The introduction also presents FCA technology and its potential in solving 3D MPSoC challenges. Then, Section 4.3 proposes to use SC converters as an interface between FCAs and power delivery lines of dies. The experiments show that using converters enables FCAs to work at their most efficient power generation regime. Next, Section 4.4 introduces a design exploration of FCA and SC converter parameters to assess the overall potential of FCAs, and identify trade-offs related to their cooling and power generation capabilities. Finally, Section 4.5 concludes the chapter.

4.2 Related work

4.2.1 3D MPSoCs trends and challenges

3D MPSoCs are getting the attention of IC design engineers due to promising advantages in terms of computing performance [77][47]. In particular, 3D stacking of dies interconnected using through-silicon-vias (TSVs) allows to integrate heterogeneous components, possibly realized in different technologies. Additionally, 3D stacking enables vertical connectivity of dies, achieving minimal inter-layer interconnect delays and very high bandwidths [58][49][12]. However, TSV-based 3D integration presents critical thermal and power management challenges, limiting its viability in modern high-performance 3D MPSoCs.

Power density increases with the number of stacked dies, generating heat that becomes difficult to dissipate due to the low thermal conductivity of silicon and other bonding materials [78]. In addition, High leakage and power density aggravate power delivery challenges in 3D MPSoCs [80]. The need for power TSVs increases with the number of stacked dies. Those TSVs and the power delivery metal lines must supply very high currents, potentially incurring voltage drops throughout the 3D power grids. In turn, voltage drops affect the latency of logic and memory, affecting system performance and possibly leading to timing failures [79].

4.2.2 Thermal management strategies for 3D MPSoCs

3D MPSoCs exacerbate thermal dissipation due to high power densities, particularly for highly-scaled technology nodes. In this regard, several design-time solutions address the heat extraction problem related to 3D integration:

The authors of [14] propose an algorithm to place thermal TSVs throughout the silicon bulk during floorplanning stages. Their approach, however, requires a significant area footprint and limits inter-layer communication bandwidth. Conversely, [83] discusses the non-homogeneous placement of TSVs for thermal balancing and control, using minimal percentages of TSVs in strategic positions. They also use specific glue materials for a more effective thermal distribution. Then, the authors of [84] advocate for the integration of novel technologies, such as resistive random access memories (RRAM). This methodology significantly impacts heat generation but is not generic as it relies on specific technologies.

In terms of cooling strategies, fan-based cooling struggles to maintain 3D MPSoC temperatures at acceptable levels. Hence more sophisticated approaches can be used to improve heat dissipation. For example, a high-performance direct liquid cooling solution using a cold plate has been proposed [56]. Nonetheless, such an approach requires large cold plate dimensions, low coolant temperatures, and costly materials to extract the high amount of heat generated by 3D MPSoCs [57]. Similar to FCA technology, inter-tier liquid cooling employs micro-channels etched in the silicon substrate of 3D MPSoC dies, through which a liquid is pumped through. This liquid absorbs the generated heat [25]. As opposed to FCAs, the coolants employed in this scenario are inert, and no electrical power is generated.

4.2.3 Power management strategies for 3D MPSoCs

3D MPSoCs expose new power delivery challenges due to the high amounts of current traversing power delivery lines and the additional routing complexity due to the presence of vertical lines. Addressing these 3D MPSoC power-related issues, several methodologies were proposed to improve power efficiency:

First, the authors of [85] use an active interposer, which reduces the power density of large-scale heterogeneous chiplet-based systems. They use different state-of-the-art on-chip power management strategies and energy-efficient 3D plugs for communication between the chiplets and the active transposer. However, this technique does not exploit the high bandwidth capabilities of TSV-based 3D integration and presents challenges related to long-distance communication through horizontal lines in the interposer. In contrast, [51] proposes a technique to plan power delivery TSVs by co-optimizing their location, number, and size. This approach aims for a minimum voltage drop while satisfying TSV area constraints. Similarly, [80] proposes a routing algorithm to minimize the power dissipation and wire delays of TSV-based 3D ICs. However, the techniques in [51] and [80] deploy a sizeable number of TSVs dedicated to power delivery at the expense of inter-tier communication. Thus, these power management strategies expose a trade-off between power delivery and communication bandwidth.

4.2.4 3D MPSoC design with integrated flow cell arrays

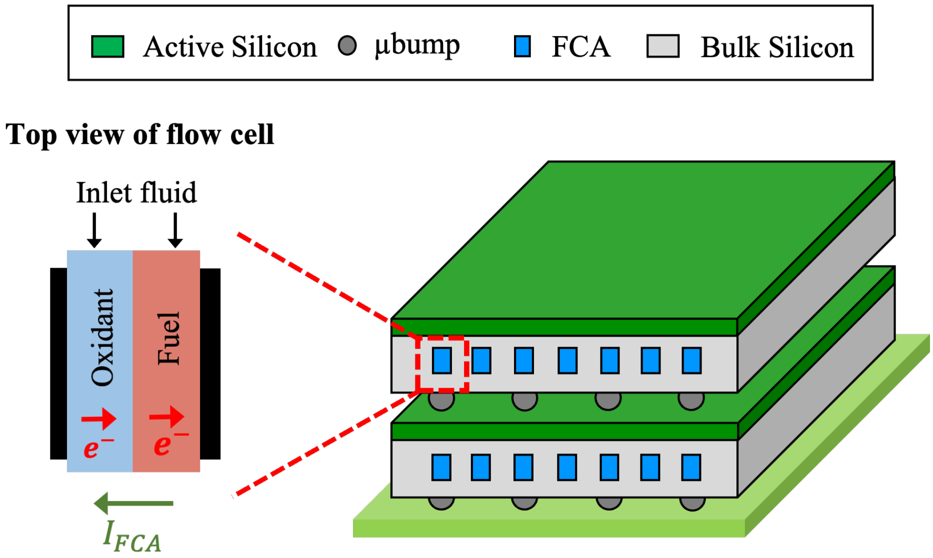


Figure 4.1: 3D MPSoC with integrated FCAs

FCA technology is a novel solution to both the power and thermal challenges of 3D MPSoCs, providing combined on-chip liquid cooling and electrochemical power generation [43]. Similarly to inter-tier liquid cooling [25], FCAs use micro-channels etched in the silicon substrate of 3D MPSoC dies, as shown in Figure 4.1. The channels are filled with an electrolytic liquid flow that adsorbs heat generated by the switching activity of chips. Furthermore, high channel temperatures increase the rate of the electrochemical reactions, which generates an electrical current that can be supplied to logic gates. Hence, FCAs effectively transform heat into available generated power in high-density and high-performance 3D MPSoCs [43]. The previous Chapter 3 shows that FCA-generated current can recover a significant percentage of voltage (IR) drop when augmenting an existing PDN. Alternatively, FCAs can also be employed to reduce the density of power delivery components (e.g., power TSVs) for a traditional PDN while abiding by specific voltage drop constraints. Hence, connecting FCAs to 3D MPSoC PDNs shows substantial improvements in power efficiency at no extra cost related to TSVs and power delivery grid density.

FCA power generation depends on the voltage between electrodes and the liquid temperature. For vanadium-based redox flows used in this thesis, peak power generation is achieved around $0.6V$, as shown in Figure 4.2. Chapter 3 proposed directly connecting FCA electrodes to nearby power delivery metal lines in the back-end-of-line (BEOL) of dies, enabling a lossless extraction of their on-chip generated power. However, FCA power generation capabilities are limited in such a scenario, as they operate farther from their optimal voltage conditions. Indeed, the typical supply voltage V_{dd} of high-performance systems is typically around $1V$, reducing by over 30% the power extraction compared to optimal conditions (Figure 4.2). Hence, this chapter proposes voltage regulation to ensure maximal power extraction from FCAs and fully exploit their potential.

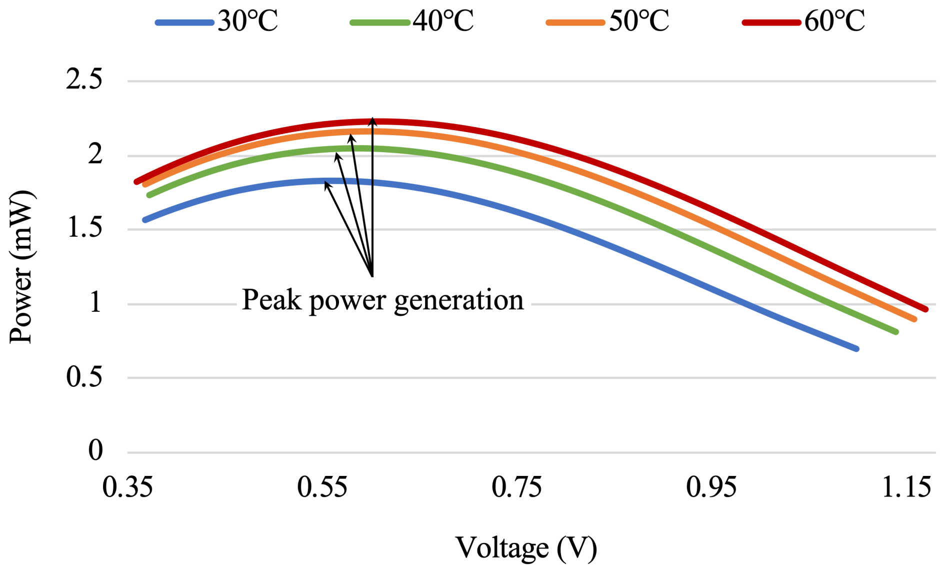


Figure 4.2: FCA-generated power (cell width: $50\mu m$, height: $100\mu m$, length: $100\mu m$)

4.2.5 Voltage converters for efficient power delivery

Voltage Regulators (VRs) are used for power management of modern 3D-ICs, providing voltage levels that are different than the standard printed circuit board (PCB) supply. Integrated VRs also enable fast voltage scaling to improve core performance while maintaining acceptable power consumption and voltage-drop levels [12]. In addition, VRs allow to decouple power supply from logic circuits and keep constant voltage levels in case of supply noise and transient load changes. The most common types of VRs used in ICs to step up voltage are switching regulators, which function by temporarily storing charge in magnetic or electric fields and then discharging it at a different voltage level. Two main types of switching regulators exist, namely, inductor-based [86] and capacitor-based [87][88].

Inductor-based switching regulators, used in IC power delivery networks, generally consist of buck-boost or buck-boost-derived topologies [89][90]. They can achieve high power conversion efficiency rates and are easy to control. However, on-chip integration of inductors remains a fundamental challenge for inductor-based VRs. Notably, the low quality of air-core spiral inductors makes them unsuitable for ICs[89], and microfabricated inductors, although promising, are not well developed yet [90]. Successful implementation of an inductor-based VR is found in 4th generation Intel Core SoCs [86]. This integrated VR design achieves a maximum conversion efficiency of 90% and output power of $108W$ but has a substantial area requirement of $175mm^2$, of which $160mm^2$ are occupied by inductors alone.

Capacitor-based converters, known as SC converters, exclusively consist of capacitors and transistor switches. Compared to inductor-based VRs, SC converters are easier to integrate and require significantly lower chip area. Common SC converter designs achieve efficiencies up to 80% [87][88], sizing around $0.3mm^2$ for an output power of $4.2mW$. Furthermore, designs including new technologies such as deep trench capacitors [81] can achieve 85% efficiency with 30× lower area requirement.

FCA power generation is sub-optimal when operating at the V_{dd} of most state-of-the-art high-performance ICs. Hence, this chapter uses VRs as an interface between FCA electrodes and power delivery lines in the BEOL to improve the on-chip power generation performance. Due to the significantly lower cost in the chip area, switched-capacitor technology is used to design converters that meet the voltage and power efficiency requirements of 3D MPSoCs with integrated FCAs.

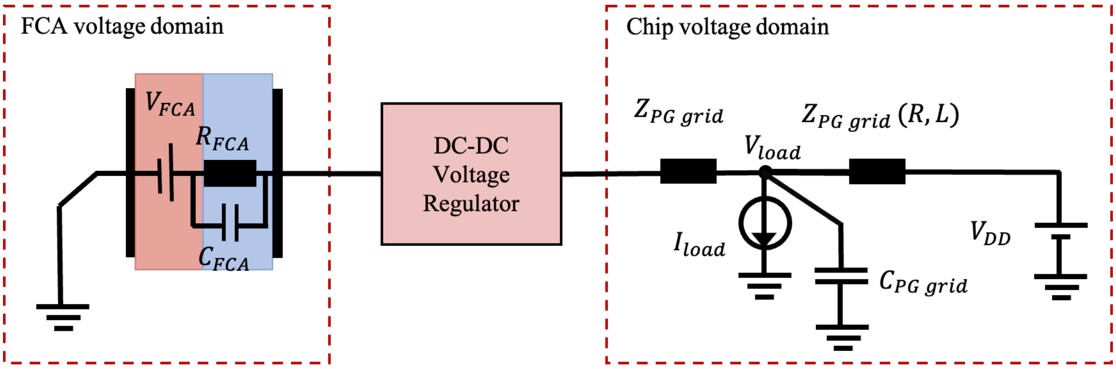


Figure 4.3: FCA cell connected to 3D MPSoC power delivery grid via DC-DC voltage regulator, as proposed in [91]

4.3 Optimal FCA power generation with switched capacitor (SC) converters

Using voltage regulators between the power delivery grids of 3D MPSoC dies and FCAs is expected to increase the on-chip power generation performance by operating FCAs at a voltage level that matches their most efficient regime. To illustrate these performance gains, the following section proposes to use a 1:2 SC converter. This converter topology is chosen because its conversion ratio matches the voltage conditions of high-performance target systems, reaches high conversion efficiency values, and generally occupies a low area. Hence, Section 4.3.1 presents the proposed converter design procedure. Then, Section 4.3.2 presents the experimental setup used to evaluate the converter's impact on PDN performance. Finally, Section 4.3.3 showcases the corresponding experimental results.

4.3.1 SC converter design

This section describes the design methodology of a 1:2 SC converter to provide optimal FCA operation voltage, when connected to the power grid of a high-performance die operating at a V_{dd} of 1.1V. Thus, the SC converter connects FCA cells to 3D MPSoC power grids, each with their respectful voltage domain, as shown in Figure 4.3. It is primarily designed for minimal area, while aiming for the highest possible output power.

4.3.1.1 SC converter state-space model

A typical 1:2 SC converter consists of four transistors and a flying capacitor C_{fly}, as shown in Figure 4.4. It continuously switches between two capacitor charge/discharge phases, stabilizing the output voltage at around two times the input voltage. This topology can be simplified to the equivalent converter circuit model in Figure 4.5. It comprises of two resistors R_v and R_i, modeling conduction and switching losses, respectively, and a capacitor C_{eq} to account for first-order circuit dynamics. For a target output voltage $V_{out} = V_{dd}$ and FCA

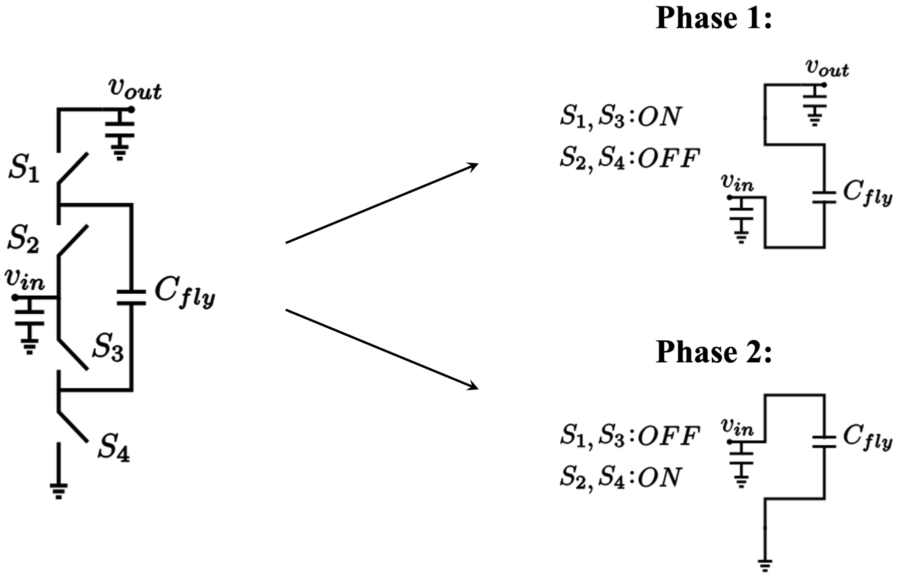


Figure 4.4: 1:2 SC converter circuit

operation voltage $v_{in} = v_{FCA}$, the resistances R_v and R_i are determined using the following equations. The Voltage Conversion Ratio (VCR) corresponds to the particular SC topology.

$$R_v = \frac{VCR * v_{in} - v_{out}}{i_{out}} \tag{4.1}$$

$$R_i = \frac{v_{out}}{\frac{i_{in}}{VCR} - i_{out}} \tag{4.2}$$

Then, the capacitive component C_{eq} is calculated by analyzing circuit equations in the Laplace domain, as described in [92].

Thus, characterizing a given SC converter design requires determining the electrical parameters of its circuit. To this end, a matrix-based methodology is developed [93] and improved, taking into account the significant effect of on-chip bottom plate capacitance [94]. This generalized methodology is used to design SC converters for 3D MPSoCs with FCAs, taking into account the specific electrical characteristics. Although the design and modeling steps apply to any SC converter topology, they are exclusively used in this section for a 1:2 VCR converter, considering a 3D MPSoC V_{dd} voltage of 1.1V.

For a given SC converter, the following matrices and vectors are defined, where **n** is the number of capacitors and **i** and **v** are the current and voltage across each capacitor, respectively:

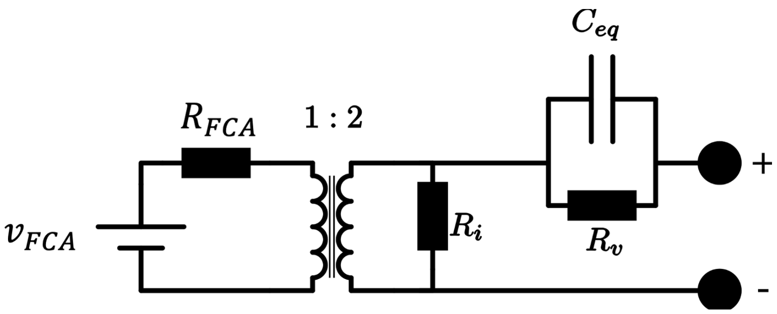


Figure 4.5: 1:2 SC equivalent converter model

$$\mathbf{C} = \begin{bmatrix} C_1 & 0 & \cdots & 0 \\ 0 & C_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & C_n \end{bmatrix} \mathbf{i} = \begin{bmatrix} i_{C_1} \\ \vdots \\ i_{C_n} \end{bmatrix} \mathbf{v} = \begin{bmatrix} v_{C_1} \\ \vdots \\ v_{C_n} \end{bmatrix} \mathbf{U} = \begin{bmatrix} v_{in} \\ v_{out} \end{bmatrix} \tag{4.3}$$

Applying Kirchhoff's current (KCL) and voltage (KVL) laws to the circuit, $2n$ independent equations are derived in the form of Equation 4.4. The values of the matrices **A** and **B** represent circuit resistances when applying KVL and KCL.

$$\mathbf{Ai} + \mathbf{Bv} + \mathbf{DU} = \mathbf{0} \tag{4.4}$$

From the above equation, the current vector **i** can be isolated as follows:

$$\mathbf{i} = -\mathbf{A}^{-1}\mathbf{Bv} - \mathbf{A}^{-1}\mathbf{DU} \tag{4.5}$$

Considering the capacitor's fundamental equation $i = C\dot{v}$, the gradient of the voltage $\dot{\mathbf{v}}$ is expressed as follows:

$$\dot{\mathbf{v}} = -\mathbf{C}^{-1}\mathbf{A}^{-1}\mathbf{Bv} - \mathbf{C}^{-1}\mathbf{A}^{-1}\mathbf{DU} \tag{4.6}$$

The above expression represents a state-space equation of the form:

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu} \tag{4.7}$$

This process is repeated for each switching phase j of the SC converter. By solving the state-space equation 4.7 for each phase, a solution is obtained in the form:

$$\mathbf{v_j}(\mathbf{t}) = \Phi_\mathbf{j}(\mathbf{t})\mathbf{v}(\mathbf{0}) + \Gamma_\mathbf{j}(\mathbf{t})\mathbf{U} \tag{4.8}$$

Assuming that in steady-state $\mathbf{v}((\mathbf{k}+\mathbf{1})\mathbf{T}) = \mathbf{v}(\mathbf{kT})$, the equilibrium value $\mathbf{v}(\mathbf{0})$ of the voltage is calculated at the beginning of each cycle. Hence, the voltage difference $\Delta\mathbf{v}$ is calculated, and the charge across each capacitor over the switching period T is derived. This procedure allows to approximate the expected converter output and build the equivalent circuit model from Figure 4.5.

Additionally, first-order converter dynamics are calculated and incorporated into the circuit model with few modifications to the proposed procedure. To reduce the analysis to first-order dynamics, the dominant eigenvalue is computed from the state transition matrix (Φ_i) which represents most of the system dynamics. The discrete model of the system can then be described as follows:

$$\mathbf{y}[\mathbf{k}+\mathbf{1}] = \lambda_{max}\mathbf{y}[\mathbf{k}] + (1-\lambda_{max})(g_1 v_{in} + g_2 v_{load}) \tag{4.9}$$

This equation can be transformed into the Laplace domain from which the output impedance Z_{out} is derived. It is represented as a parallel RC branch in the model, taking into account R_v and C_{eq}.

4.3.1.2 SC converter design space parameters

The main parameters to configure when designing an SC converter are the transistor and capacitor sizing, and the phase switching frequency:

- The transistor and capacitor sizing determines the SC converter circuit resistance and capacitance. These two components are responsible for conduction and switching losses, affecting the voltage conversion efficiency.

- The phase witching frequency directly affects power loss in the circuit, also affecting the conversion efficiency. In general, $P_{loss} \propto \frac{1}{k \times f_{clk}}$, where k depends on the converter topology. However, in the case of high-performance ICs, this dependence does not hold for higher frequencies as additional losses occur due to parasitic components [88].

The effects of each design parameter on switching and conduction losses are correlated. Hence for a given design, Section 4.3.1.3 proposes a methodology to calculate converter performance parameters in the case of different configurations. Transistor and capacitor models from a 32nm CMOS technology are used to build the state-space model [95].

4.3.1.3 SC converter design exploration

The implemented state-space model in Section 4.3.1.1 enables calculating the SC converter performance for any given target output voltage. Indeed, modern processors typically operate at different V_{dd} values depending on load and performance. Besides, design points optimized for a given output voltage are not necessarily optimal at slightly varying output voltages. In this context, an algorithm is developed that cycles through over 10 million design parameter combinations. Since the converter is expected to output several voltage levels, the performance results are weighted according to the target load profile, obtaining averaged performance metrics. Figure 4.6 shows different evaluated SC converter design points with respect to their total area requirement, their voltage conversion efficiency, and output power. This figure indicates a clear trade-off between these characteristics. In particular, output power scales slower than converter area requirements.

According to the above converter design space exploration and considering an output voltage of 1.1V (typical V_{dd} of high-performance ICs), the selected converter design for FCA voltage regulation is represented by a star in Figure 4.6. This design has the lowest area requirement and the highest output power density. Alternatively, other configurations can be selected with higher efficiency or output power, depending on design constraints.

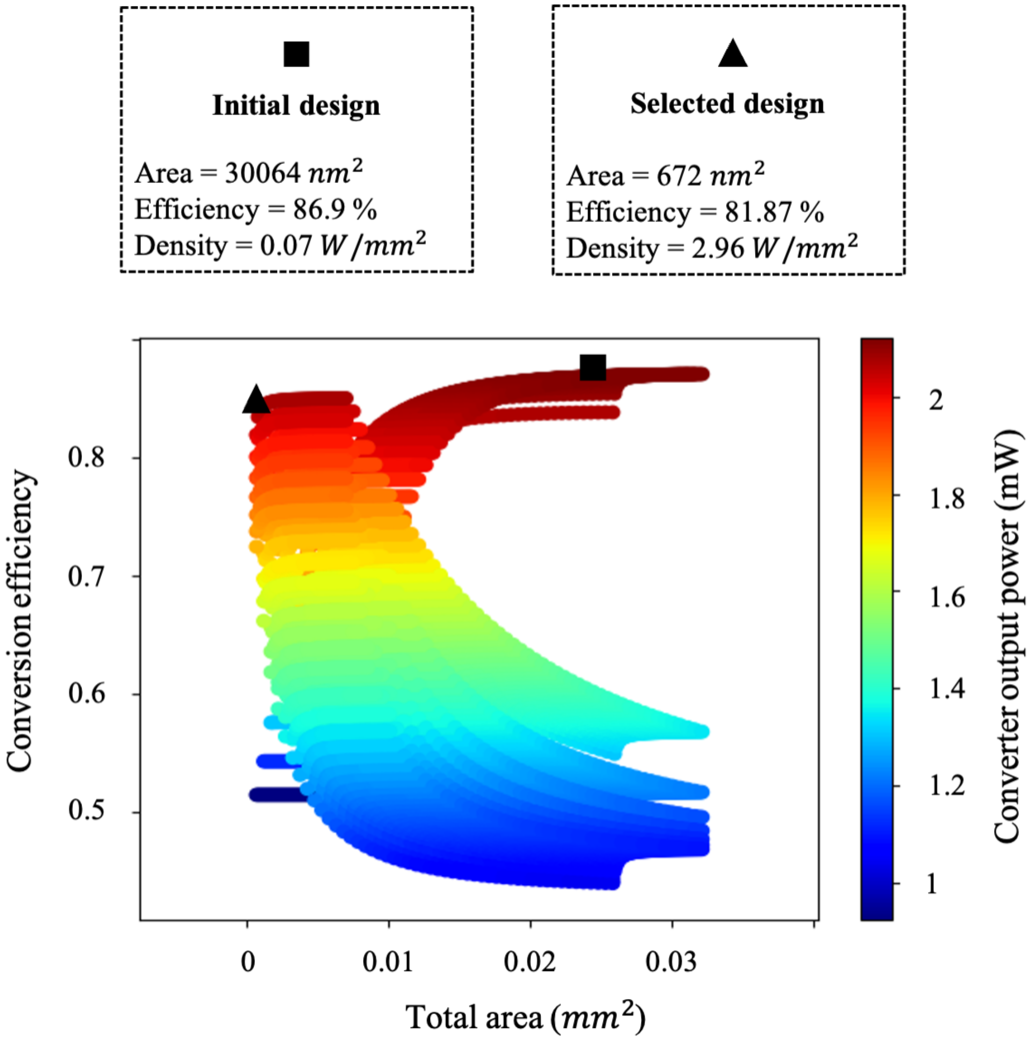


Figure 4.6: Evaluated DC-DC converter design points

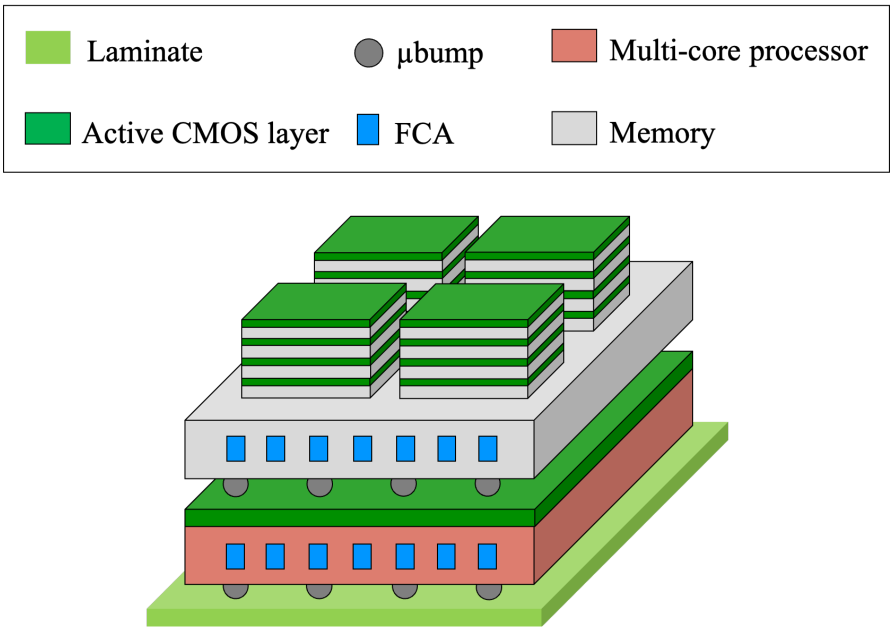


Figure 4.7: 3D MPSoC with FCAs: a processing layer and a memory layer

4.3.2 Experimental setup

4.3.2.1 3D MPSoC with FCAs and SC converters

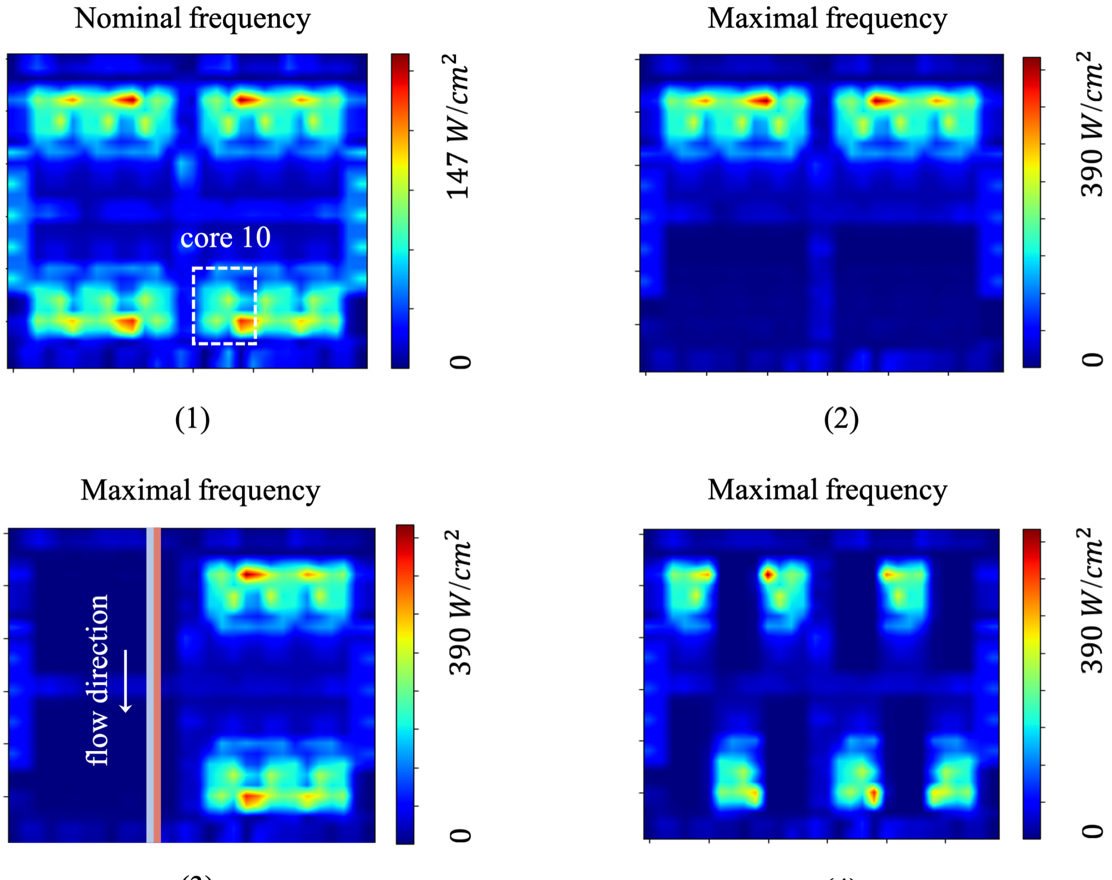
To evaluate FCA performance using the proposed SC converter, a two-layer 3D MPSoC is considered, represented in Figure 4.7. The 3D MPSoC is composed of the following layers:

- The top memory layer contains four 2^{nd} generation HBM memories with 4 DRAM layers. Each HBM has a base die size of $71mm^2$ and consumes $15W$ [6].

- The bottom processor layer architecture and power profile are based on the 12-core IBM POWER8 processor [65]. In particular, its implementation in a $32nm$ CMOS technology is considered, as it is used to build the converter state-space model. The processor's die size is $649mm^2$, and its power consumption is $190W$.

The 3D MPSoC employs *chiplet-based integration* to stack the HBMs on a base logic die, and *chip-on-chip bonding* through fine-pitched micro-bumps to stack the two 3D MPSoC layers.

FCAs of $50\mu m$ width and $100\mu m$ height are etched in the silicon substrate of both dies, with a pitch of $50\mu m$. Each $200\mu m$-long flow cell section is connected to a single SC converter, which in turn connects to the power grid of dies. Furthermore, TSVs are arranged in groups delivering power to independent power subgrids. Their diameter and pitch are both fixed at $5\mu m$. Finally, V_{dd} is set to $1.1V$, corresponding to the maximum processor performance.



(1) (2) (3) (4)

Figure 4.8: Measured POWER8 power maps

4.3.2.2 Processor utilisation scenarios

The performance of FCAs and their associated SC converters is measured in case they supply the multi-core processor. The memory layer has a significantly lower power consumption but still contributes to the overall 3D MPSoC temperature increase during activity. In particular, various utilization scenarios are analyzed using measured POWER8 power maps. These power maps are shown in Figure 4.8. They enable evaluating the voltage regulator and FCA power generation efficiency for different load levels.

All the power maps in Figure 4.8 contain multiple high power density regions. These are mainly concentrated in the computing cores. In power map (1), all cores operate at nominal frequency. Then, in power maps (2), (3), and (4), six cores operate at the maximal frequency and achieve peak power density, while the others are idle (e.g., awaiting data from memory).

4.3.2.3 Experimental flow

In the following sections, the power generation performance of FCAs and their associated SC converters are evaluated. Hence, a fine-grained simulation of the 3D MPSoC is performed to assess the thermal and voltage behavior under the different utilization scenarios. In particular, cell dimensions of $200 \times 100 \mu m^2$ and $50 \times 50 \mu m^2$ are considered for the thermal (3D-ICE) and electrical (HSPICE) simulation, respectively. The simulations evaluate the FCA power generation, converter efficiency, and processor voltage map. To perform SPICE simulations, a compact flow cell model is built corresponding to the voltage-power dependency in Figure 4.2. Then, an SC converter circuit model is considered, represented in Figure 4.5. The cores switch between 3 different activity levels: idle, nominal, and maximal frequency operation, according to the power maps in Figure 4.8. As dies have individual PDNs, the total number of FCAs and converters in the 3D MPSoC scales linearly with the number of layers. At the same time, their performance depends on the load and voltage level of each die independently.

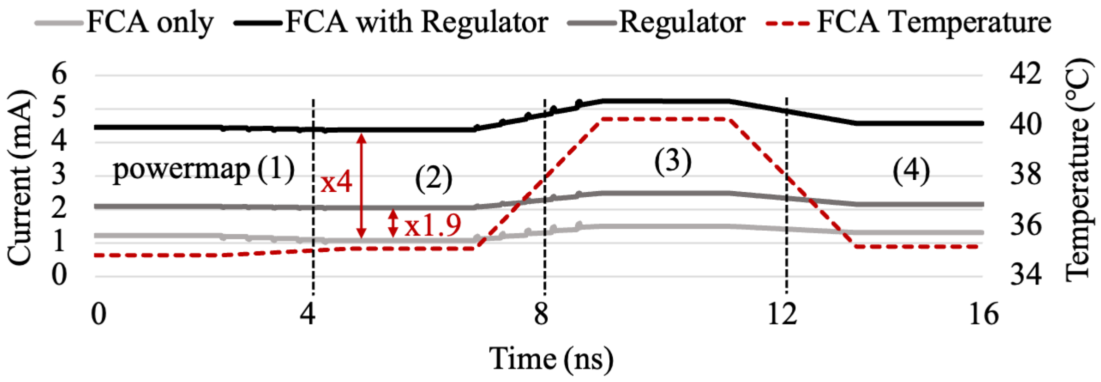
4.3.3 Experimental results

4.3.3.1 FCA power generation with SC converters

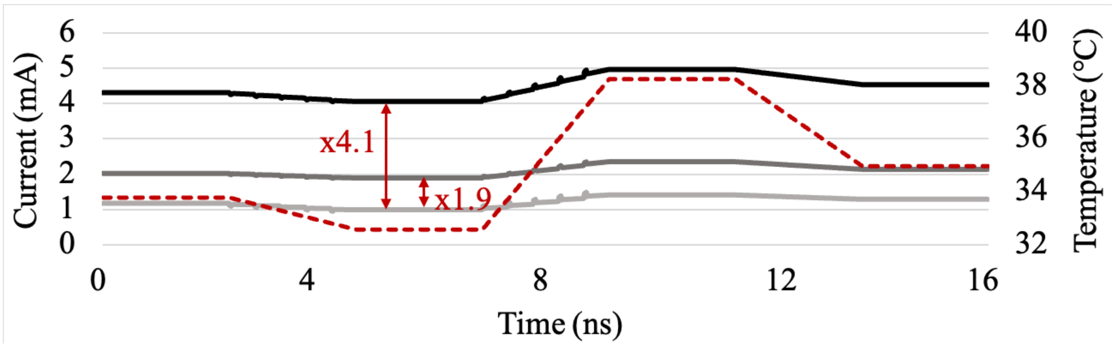
The FCA power generation and SC converter efficiency are evaluated when cores switch between different activity levels. In this regard, POWER8 core number 10 is selected as it displays different load and liquid temperature levels when switching between the power maps in Figure 4.8. In addition, three nodes are selected from the core, corresponding to various power density levels: maximum in Figure 4.9a, medium in Figure 4.9b, and low in Figure 4.9c. These figures present the results of the HSPICE transient analysis of the processor power grid. The three plots present the following measurements:

- The generated current of FCAs when they are directly connected to the processor power grid operating at a V_{dd} of 1.1V. In this case, no voltage regulation is used.
- The generated current of FCAs when SC converters are used to interface them with the processor power grid. In this case, FCAs operate at the optimal voltage enabled by voltage regulation.
- The output current of the SC converter, when used. This represents the actual current that is available to supply the processor node.

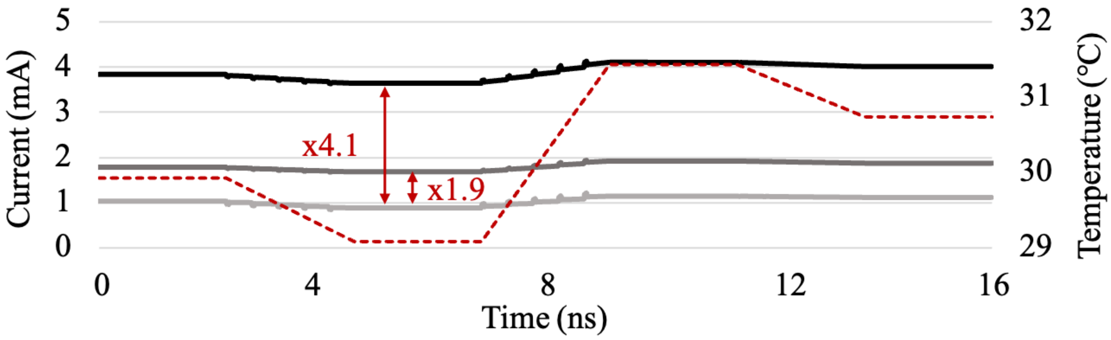
The results in Figure 4.9 show that, in all the cases, FCAs generate over four times higher current when operating at their optimal voltage supplied via the SC converter compared to the case they operate at the V_{dd} of the chip. Hence, due to voltage regulation, FCAs provide up to 123% additional power to the processor. Furthermore, peak FCA power generation is achieved in case the load corresponds to power map (3) for all three evaluated nodes due to higher liquid temperature inside the channels. Indeed, in this case, the channels traverse two high



(a) High power consumption node



(b) Medium power consumption node



(c) Low power consumption node

Figure 4.9: FCA and SC converter output currents at different power consumption scenarios

power density cores increasing their heat absorption. The results also indicate that DC-DC converter output power remains over 90% higher than FCA-generated power when no voltage regulation is used, corresponding to over 82% conversion efficiency.

According to the results above, SC converters enable boosting the power generation capabilities of individual FCA cells in 3D MPSoCs regardless of the load levels of the chip. Nevertheless, placing SC converters for individual flow cells can be cumbersome, costing chip area and adding to the power delivery design complexity. In this context, Section section 4.3.3.2 explores different configurations in terms of the number of FCA cells connected to a single SC converter.

4.3.3.2 SC converter area optimization

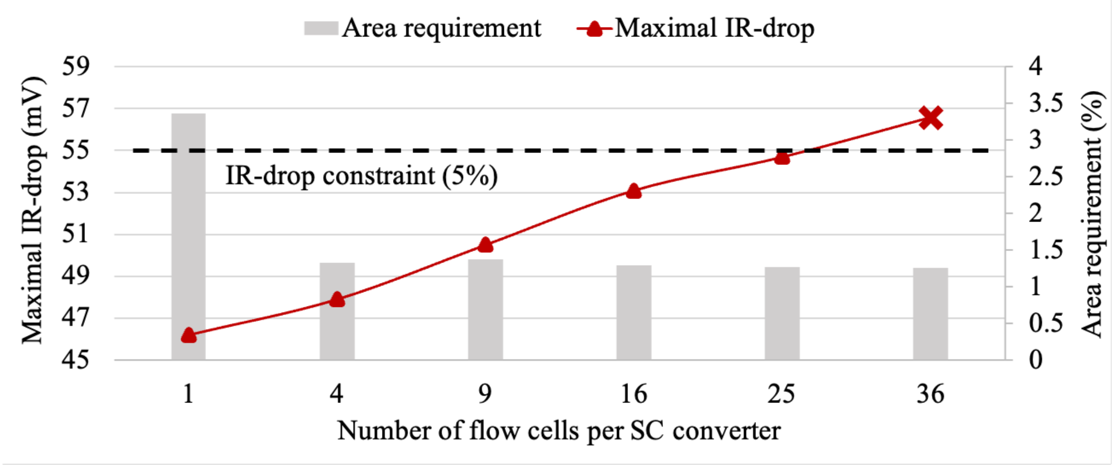


Figure 4.10: IR-drop and converter area requirements at maximal utilization ($V_{dd} = 1.1V$)

This section explores different options regarding the number of FCA cells connected to each SC converter, targeting the improvement of design complexity and area requirement of SC converters. To minimize FCA power dissipation in the PDN metal wires, flow cells are regrouped in squares. Hence, each $N \times N$ group connects to one SC converter. For each configuration, the input power density of the converter is proportional to the number of connected FCA cells. Therefore, the SC converter is optimized for maximal conversion efficiency in each case, following the methodology described in Section 4.3.1.3. Consequently, a DC voltage analysis of the processor is performed while considering the resistance of wires connecting FCA cells to SC converters. Thus, power map (3) in Figure 4.8 is selected, which contains the largest concentration of power hotspots, therefore inducing the most critical IR-drop.

Figure 4.10 presents the maximal IR-drop at the processor when connecting 1 to 36 FCA cells to each SC converter. As multiple FCA cells connect to a single converter, their output currents traverse longer wires to reach all the loads they supply. Particularly, $10mV$ additional IR-drop occurs when connecting up to 36 flow cells to one converter. Figure 4.10 also shows the total

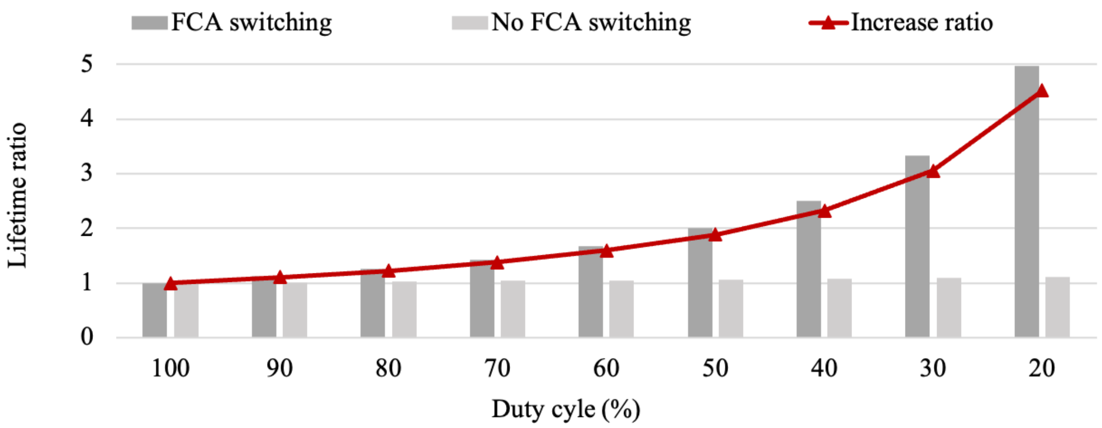


Figure 4.11: Fuel reservoir lifetime with and without switching off FCA power generation for different processor duty-cycle values, normalized to the lifetime in case of 100% processor duty-cycle

area percentage requirement of optimized converters in each configuration, as well as their conversion efficiency. This figure indicates that connecting 25 FCA cells to each converter saves over 2.5% of the total chip area dedicated to SC converters, while maintaining the maximal IR-drop value under 5% (typical voltage constraint in high-performance ICs). In addition to area savings, decreasing the number of converters by 25× reduces design and routing complexity in 3D MPSoCs with integrated FCAs.

4.3.3.3 FCA power generation resource management

FCA power generation occurs due to electrochemical reactions between electrolytes inside channels. Hence, continuous extraction of power gradually decreases the concentration of reactants in the electrolytic fluid. In particular, electro-thermal simulations demonstrate that when liquid flows through FCA channels, between 1.1% and 0.9% of its electrolytes react regardless of the load level of the chip. Furthermore, electrolytes crossover occurs along membrane-less flow cells, causing the contamination of the oxidant and fuel and decreasing their concentration regardless of reaction rate. Generally, for high-speed flow cells, up to $40mA/cm^2$ current is generated by electrolyte crossover in an open-circuit scenario, corresponding to 0.5% of generated current in optimal FCA power generation conditions [96]. As a result, SC converters enable disconnecting FCAs from 3D MPSoC PDNs, preventing excess power extraction when the chip is idle or in low-power operation. In fact, when disconnected, only FCA cooling capabilities are used, and fuel concentration degradation is limited to cross-contamination effects.

In this context, this section analyses the electrolyte consumption in case FCAs are disconnected from the power delivery network when the chip is inactive, thanks to SC converters. Figure 4.11 illustrates the lifetime of a fuel reservoir for different processor duty-cycle levels, with and without switching off FCA power extraction during idle periods. The figure also presents the ratio between the two scenarios (in red). FCA electrolyte lifetime is normalized to the value corresponding to continuous power extraction (i.e., 100% duty-cycle). The results show that switching off FCA power generation extends by over 1.8× the lifetime of a fuel reservoir, for a 50% processor duty cycle. As only crossover occurs, a large number of electrolytes are unused when no extra power is needed. When the chip is only active 20% of the time, switching off FCAs extends the reservoir lifetime by over 4.5×. In general, using SC converters improves FCA durability and enables better management of their power generation resources.

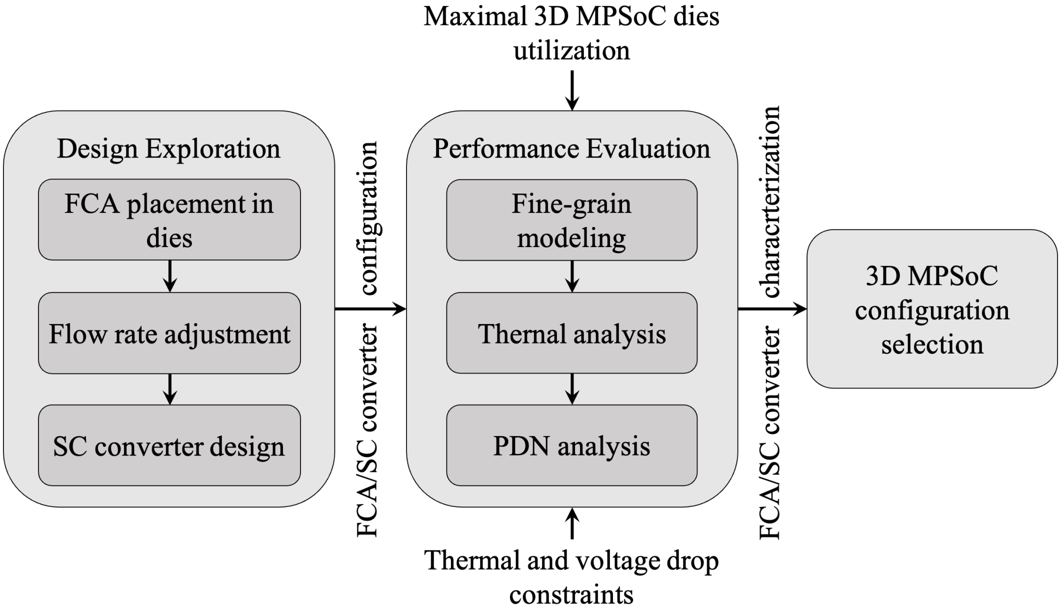


Figure 4.12: 3D MPSoC design-time performance characterization

4.4 Design-time characterization of 3D MPSoCs with FCAs and SC converters

This section introduces a design-time 3D MPSoC characterization through thermal/power performance analysis, illustrated in Figure 4.12. The proposed flow explores multiple FCA and SC converter configurations targeting a high-performance 3D MPSoC (described in Section 4.4.1.1). In addition, this flow performs fine-grain analysis considering critical usage scenarios and 3D MPSoC design constraints.

The *design exploration* considers FCA-related parameters, namely the FCA placement in the 3D MPSoC layers and electrolytic liquid flow speed (Section 4.4.1.2) and the SC converter design (Section 4.4.1.3). It has to be noted that some design choices not pertaining to FCAs, such as the placement of dies and that of TSVs, also have an influence on 3D MPSoCs thermal characteristics and those of their PDNs [97][98]. However, simulations indicate that microchannels thermally isolate different dies. Hence, die placement only has a minor influence on a 3D stack thermal behavior when FCAs are used. Furthermore, prior art indicates that placing TSVs near power hotspots is the best choice to minimize voltage drops (Chapter 3). Thus, this solution is adopted without further exploring this aspect.

Hence, the *performance evaluation* in Section 4.4.1.4 uses fine-grain thermal and electrical simulations to assess 3D MPSoC performance under different FCA configurations. It analyses the temperature and power reduction capabilities of FCAs, and their ability to recover voltage drop using SC converters.

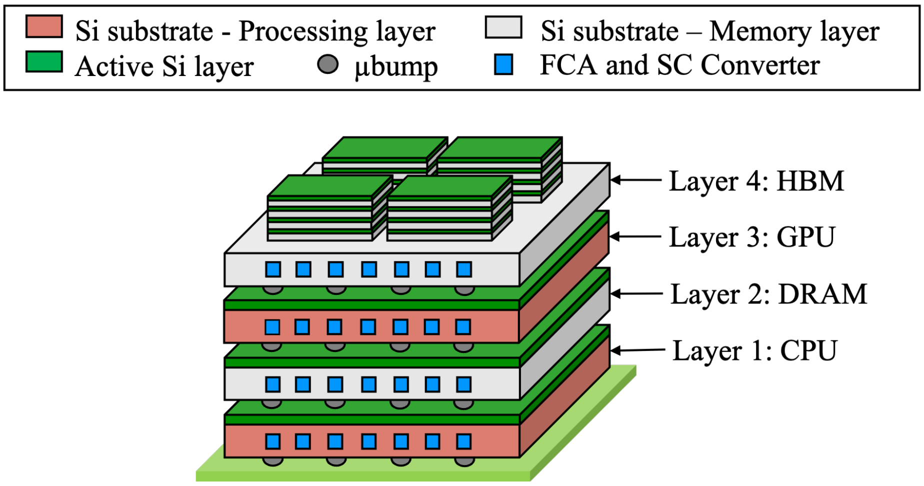


Figure 4.13: Target 3D MPSoC

4.4.1 Experimental setup

4.4.1.1 Target 3D MPSoC

In this section, I assess the efficiency of the proposed design-time exploration for 3D MPSoCs with integrated FCAs and SC converters. Therefore, a high-performance four-layer stack is employed as a target system, as shown in Figure 4.13. The architecture is based on a state-of-the-art CPU-GPU platform for high-performance computing [99], considering its implementation in 3D and anticipating a next-generation 3D MPSoC. The stack comprises the following layers:

- The first (bottom) layer is modeled after AMD's Extreme Performance Yield Computing (EPYC) microprocessor, based on the Zen micro-architecture and fabricated using a $14nm$ FinFET process [100], with a total area of $757mm^2$. Figure 4.14 presents the EPYC processor layout. It contains 32 high-performance cores, arranged as 4 Ryzen 8-cores clusters sharing one L3 cache. The processor's maximal total power consumption is $180W$, and its maximal supported temperature is 81°C.

- The second layer contains an 8-channel DDR4-2666W [101], supported by the EPYC processor. The memory is fabricated using an 18nm 3-metal layer DRAM process. Each of the eight $16Gb$ DDRs occupies a total size of $81.28mm^2$.

- The third layer is based on NVIDIA V100 [102], a data center GPU designed to accelerate AI, HPC, and graphics. The NVIDIA V100 is composed of 640 Tensor cores and 5120 CUDA cores, arranged as six graphics processing clusters (GPCs) with 14 streaming multiprocessors (SMs), as illustrated in Figure 4.15. This layer is fabricated using TSMC's 12nm FFT CMOS process and occupies a total size of $815mm^2$. It consumes up to $300W$ and operates at a maximal temperature of 85°C.

- The fourth (top) layer is composed of four 2^{nd} generation HBM memories with 4 DRAM layers each, providing the bandwidth requirement of the NVIDIA V100 GPU. Each HBM memory has a base size of $71mm^2$, fabricated using the 29nm DRAM process. The maximal power consumption of each HBM2 memory is $15W$ [6].

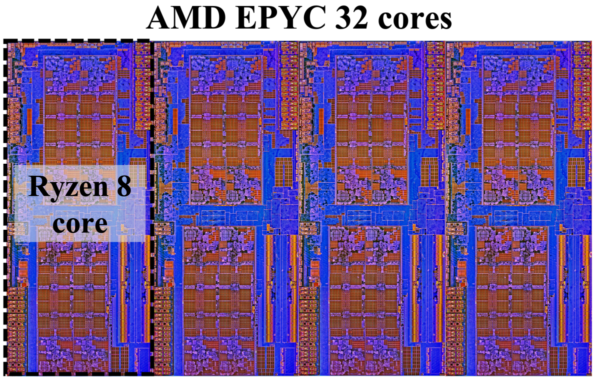


Figure 4.14: CPU layout [103]

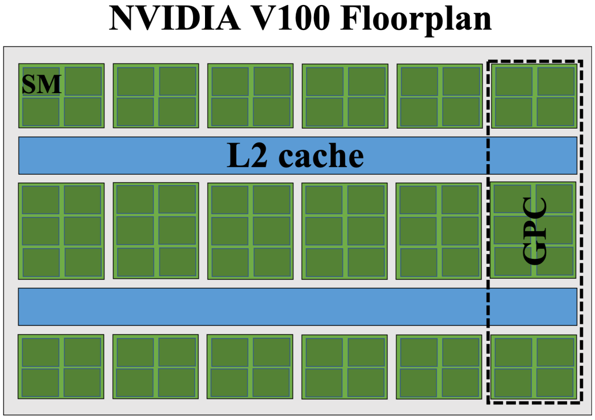


Figure 4.15: GPU floorplan

The target 3D MPSoC employs a combination of two state-of-the-art 3D integration technologies, namely: *chiplet-based integration* and *chip-on-chip bonding* through fine-pitched micro-bumps. The first one enables stacking multiple HBMs on a base logic die (top layer). The latter enables logic-on-logic integration and is used to stack the four 3D MPSoC layers, including the HBM active interposer and the package.

FCAs of $50\mu m$ width and $100\mu m$ height are etched in the silicon substrate of the 3D MPSoC dies, with a pitch of $50\mu m$. Each $200\mu m$-long flow cell section is connected to a single SC converter, which is, in turn, connected to the power grid of the corresponding die. TSVs are arranged in groups (TSV islands), each delivering power to an independent power domain. Their diameter and pitch are both fixed to $5\mu m$.

In Section 4.4.2, this 3D MPSoC is modeled in fine-grain to evaluate both its thermal and electrical performances. In particular, 3D-ICE [33] is used to evaluate its thermal behavior. Then, HSPICE is used to measure its PDN performance. Therefore, a compact FCA model and a converter circuit model (Figure 4.5) are included to perform electrical simulations. Both the flow cells and SC converters are modeled in Verilog-A. Hence, the FCA power generation and SC converter efficiency are evaluated, enabling to retrieve the voltage and temperature distributions of dies. Cell dimensions of $200 \times 100 \mu m^2$ and $50 \times 50 \mu m^2$ are used for the thermal and electrical simulations, respectively.

4.4.1.2 FCA placement and flow rate configuration

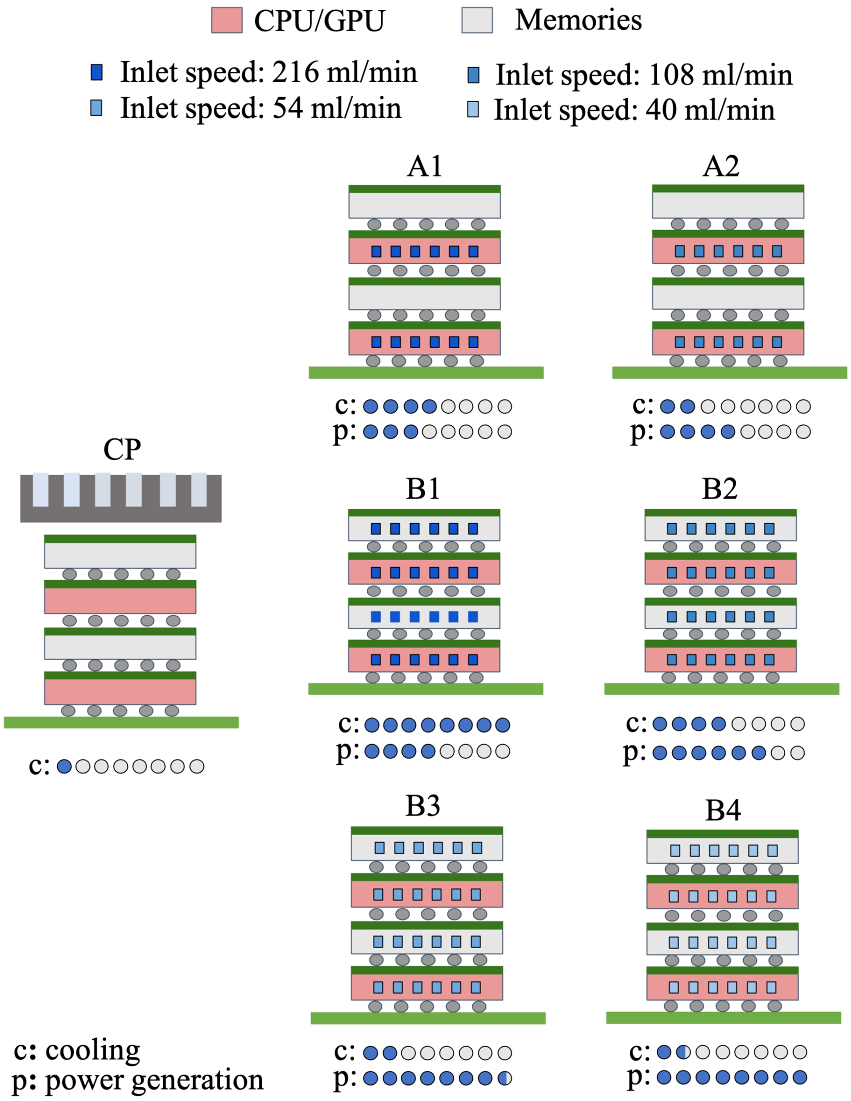


Figure 4.16: Explored 3D MPSoC configurations

FCA-based cooling and power generation interact in a non-obvious way. Higher cooling efficiency decreases temperature, limiting the total chip leakage (hence total power consumption) and lowering the power generated by FCAs. Conversely, lower cooling increases the electrolyte reaction rate, which generates more electrical power. Both FCA capabilities are affected by their physical design properties in a 3D MPSoC. Hence, detailed thermal and power analyses are needed to evaluate multiple configurations, investigate FCA trade-offs, and identify the best configuration under specific voltage and temperature constraints.

The FCA-related design space includes multiple parameters. However, the proposed design exploration in this section focuses on *FCA placement* in 3D MPSoCs and *electrolytic flow rate* inside the channels. These two parameters have a more significant impact on FCA capabilities than other parameters, such as their dimensions. In this context, the configurations shown in Figure 4.16 (A1 to B4) are considered candidate options. These configurations are selected as follows:

- The configuration groups A and B represent the number of FCAs that supply each computing die. Only one FCA supplies the CPU/GPU in configurations A, while two FCAs supply it in configuration B. This is achieved by electrically connecting the computing dies using TSVs to the FCAs etched in the dies, and via TSVs and micro-bumps to the FCAs etched in the above memory dies. As the memories consume considerably less power, they are not supplied with FCA power.
- In terms of the FCA flow rate, the full-flow rate value (216ml/min, as in [44]) is considered for both cases (A1 and B1). Then, the lowest flow rate is considered that complies with temperature constraints (B4).
- Then, to highlight the existing trade-offs between FCA cooling and power generation, other configurations are considered with similar cooling performances but different numbers of FCAs (e.g., A1 and B2, A2 and B3).

For comparison, a state-of-the-art cold-plate-based liquid cooling for ultra-high-performance MPSoCs [56] is characterized (configuration CP). The performance, in terms of on-chip cooling and power generation, of each configuration are qualitatively represented in Figure 4.16. For 3D MPSoC configurations with integrated FCAs, on-chip *cooling* depends on the amount of liquid pumped in the channels per unit of time, which linearly increases with the number of FCA channels in the dies and the inlet speed. Hence, configuration B1 has the highest on-chip cooling efficiency, while configurations A1 and B2 achieve half this efficiency due to a reduced number of FCAs and a slower liquid traversal, respectively. On the other hand, configuration CP has the lowest cooling efficiency due to the low 3D MPSoC inter-layer heat dissipation. On-chip *power generation* depends instead on the number of FCAs and the coolant temperature. Accordingly, configuration A1 generates half the amount of power compared to configuration B2 for the same cooling performance. Then, configuration B4 has the highest power generation

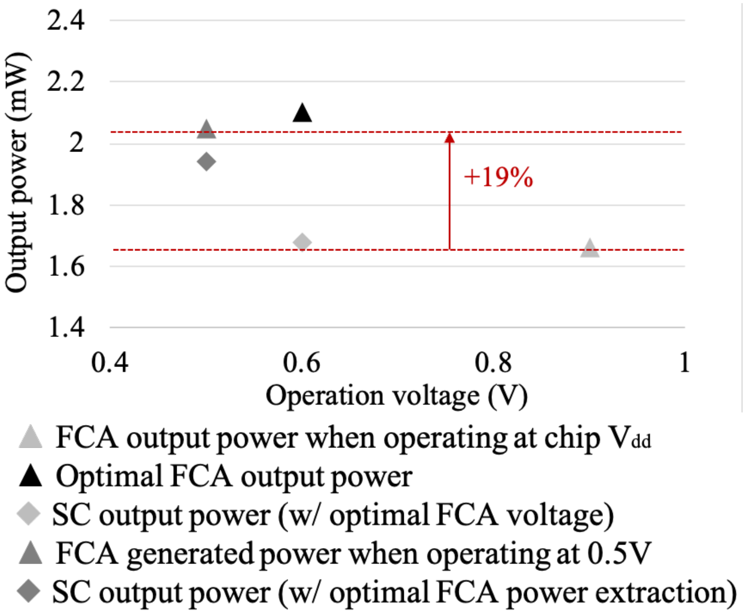


Figure 4.17: FCA and SC converter output power

efficiency as the coolant heats the most compared to other configurations, accelerating the electrochemical reactions inside the channels.

These intuitions are quantitatively assessed in Section 4.4.2, where the outcomes of fine-grain thermal and electrical simulations are presented according to the methodology outlined in Section 4.4.1.1.

4.4.1.3 SC converter configuration

A trade-off exists between SC conversion efficiency, area, and output power (Section 4.3.1.3). Moreover, the amount of extracted FCA power and the SC converter efficiency should not be considered in isolation, as both influence the power delivered to the PDN. To illustrate this aspect, let's consider FCAs operating at their optimal voltage (0.6V in Section 4.2.4) and SC converters that adapt this input voltage to the level required by the 3D MPSoC dies (0.9 V). According to the SC converter design-space exploration introduced in 4.3.1.3, the optimal design point achieves a relatively low voltage conversion rate. Indeed, as illustrated in Figure 4.17, the total power output in this scenario is similar to the case when no converter is placed between FCAs and 3D power grids, and FCAs operate at the same voltage as the rest of the chip.

Conversely, maximal PDN efficiency is achieved when the overall power delivery system encompassing FCAs and SC converters is most efficient, resulting in maximal output power. This condition is achieved when the voltage at the FCA electrodes (i.e., the SC converter input voltage) is set to a lower level of 0.5V. According to the SC design-space optimization methodology in Section 4.3.1.3, optimal SC converters achieve, in these conditions, on average over 82% voltage conversion efficiency. Thus, they lead to 19% higher FCA power generation than directly connecting FCA electrodes to the PDN. Furthermore, these SC converters require less than 3% of the total chip area (34200 are placed, each occupying $0.00071 mm^2$). Therefore, the optimal SC converter design in this scenario is selected for the remainder of this section.

To quantify the system-wide benefits of this design, Figure 4.18 presents the voltage drop maps of the CPU and GPU dies in 3D MPSoC configuration B4 and in case of maximal power consumption, corresponding to their thermal design point (TDP). This scenario is chosen to perform worst-case circuit analysis, representing extreme operating conditions. First, the voltage drop is shown when FCAs are only used to cool down the die (inter-tier liquid cooling). In this scenario, the voltage drop reaches over 78mV (8.6% V_{dd}) for the CPU, and over 100mV (11% V_{dd}) for the GPU. Thus, for both dies, the voltage drop violates the typical 5% constraint of high-performance ICs. Then, the voltage drop map is shown when FCAs are directly connected to the power delivery grid of dies. In this case, the voltage drop decreases by 60mV for the CPU and 70mV for the GPU. Finally, Figure 4.18 represents the voltage drop when SC converters are placed between FCAs and 3D power grids, and FCAs operate at 0.5V. The figure shows that with respect to using unregulated FCAs, using SC converters effectively achieves a further reduction of the voltage drop across both dies, limiting them to 2% V_{dd} for the GPU and almost eliminating them for the case of the CPU.

4.4.1.4 Experimental flow

In the following Section 4.4.2, the 3D MPSoC thermal and power performance is evaluated under the different configurations described in Section 4.4.1.2, assuming the use of the SC converter identified in Section 4.4.1.3. These configurations are compared to the cold plate-based liquid cooling strategy (cf. configuration CP in Figure 4.16). Across experiments, a maximal usage scenario is considered for both the CPU and GPU, representing worst-case operating conditions. The dynamic and leakage power values are reported, where dynamic power is calculated by subtracting the leakage corresponding to the maximal die temperature from the TDP (as indicated by the dies specifications in Section 4.4.1.1). Leakage maps when dies are cooled using FCAs or CP cooling are calculated based on the computed temperature maps, considering the temperature-leakage dependence (the details are presented in Chapter 5).

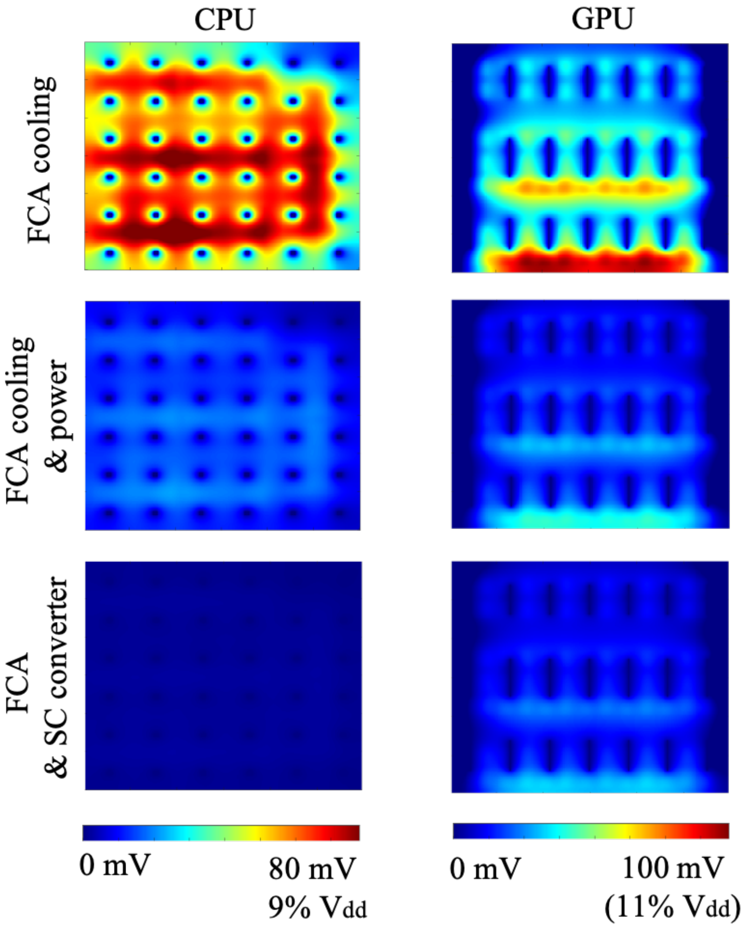


Figure 4.18: CPU and GPU IR-drop maps

4.4.2 Experimental results

4.4.2.1 Temperature and total power consumption

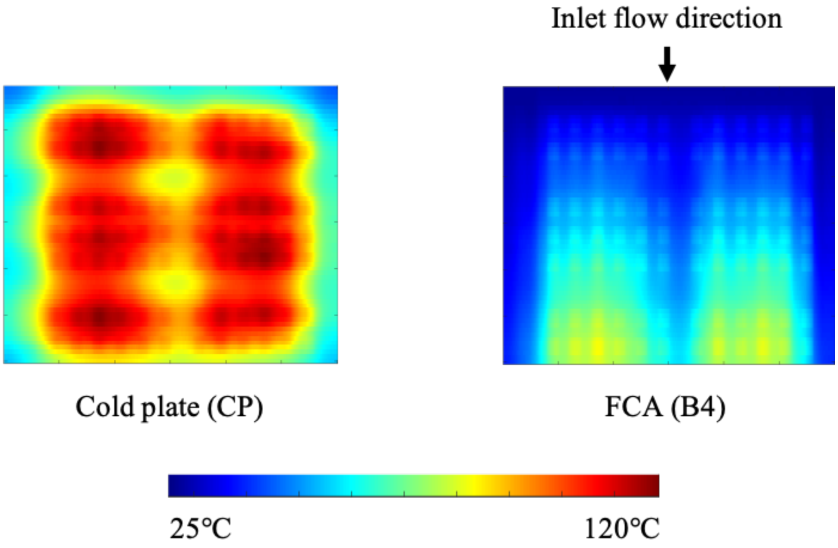


Figure 4.19: CPU temperature with FCAs and cold-plate cooling

Figure 4.19 shows the CPU temperature map when cooled using the CP solution [56], compared to the one in configuration B4 using FCAs with the lowest cooling capacity. The figure showcases that FCAs vastly outperform cold plate-based liquid cooling, which reaches almost 120 °C. FCA cooling, on the other hand, enables maintaining CPU temperature below 80 °C, even in an extreme power consumption scenario.

Additionally, the total power consumption of the CPU and GPU dies are measured at maximal usage. The results are presented in Figures 4.20 and 4.21, respectively. The figures indicate that the temperature-dependent leakage significantly contributes to the total power consumption in the CP case. In the case of FCA cooling, leakage power can be effectively reduced by up to 86% compared to the CP strategy for the CPU and up to 82% for the GPU. The peak temperature difference between configurations CP and B4 are 78°C and 75°C for the CPU and GPU dies, respectively.

Among all configurations B, the configuration with the highest cooling capability (B1) outperforms the configuration with the lowest cooling capability (B4) in terms of leakage reduction by 8% for the CPU and 14% for the GPU. However, configuration B1 has the lowest power generation capacity, as detailed in the following section 4.4.2.2.

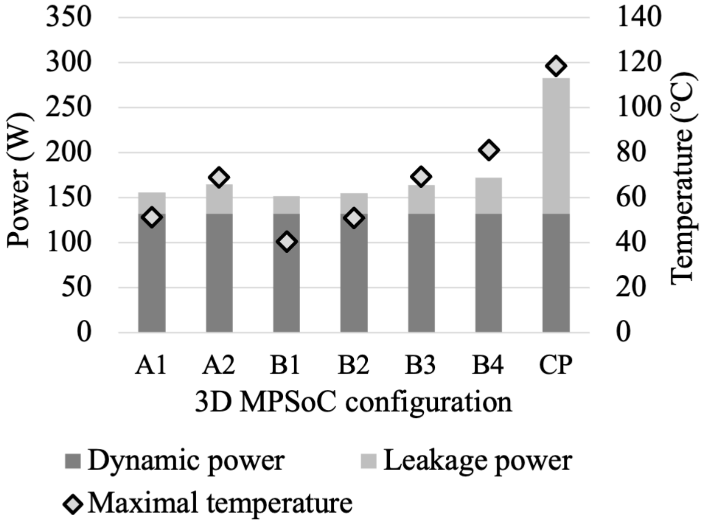


Figure 4.20: CPU power and temperature

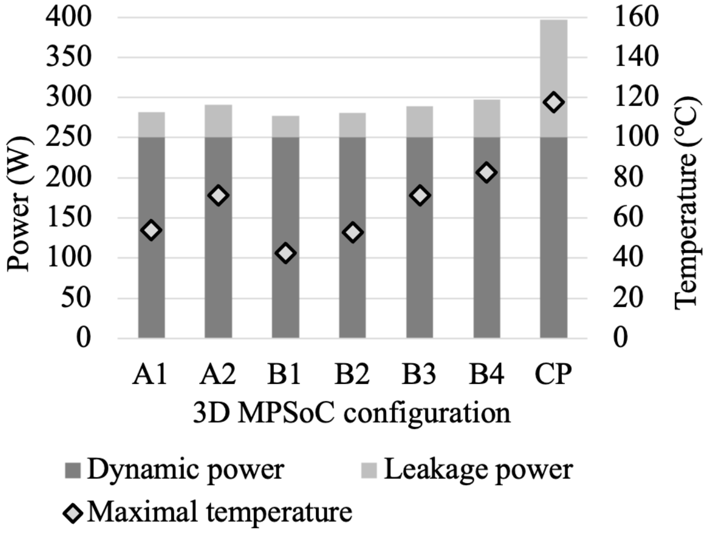


Figure 4.21: GPU power and temperature

4.4.2.2 Voltage drop recovery

Maximal voltage drop values for the CPU and GPU dies are presented in Figure 4.22 and 4.23, respectively. All the considered FCA configurations and the CP one are included. In configurations A1 to B4, the voltage drop is measured in three scenarios:

- When FCAs are only used for their cooling capabilities. This scenario is equivalent to inter-tier liquid cooling.
- When FCAs are directly connected to the power grids. In this scenario, FCAs operate at a voltage level far from their optimal regime.
- When SC converters are used as an interface between FCAs and power delivery lines. In this scenario, FCAs operate close to the optimal voltage leading to maximal power generation.

For both dies, the use of FCAs decreases voltage drop by over 90mV (10% V_{dd}) compared to the CP cooling strategy and up to 78mV (8.6% V_{dd}) compared to inter-tier liquid cooling. In particular, the configurations with the highest coolant speed lead to a lower die temperature, overall power consumption, and voltage drop. However, the increased reaction rate of FCAs with temperature enables more power generation. Moreover, the on-chip power generation is uniform across dies, whereas leakage is highest at the hotspots. Therefore, FCA power generation capabilities have a higher impact on voltage drop recovery than their cooling in the case of non-uniform 3D MPSoC power distributions. In this context, FCAs and SC converters recover a higher percentage of voltage drop in configuration A2 compared to configuration A1, for both the CPU and GPU (Figures 4.22 and 4.23). A similar observation is done between configurations B1 and B4, where FCA power generation is significantly higher.

Additionally, the configurations with the highest number of flow cells present the highest voltage drop recovery percentage. In particular, configuration B3 generates double the amount of power with respect to configuration A2 for the same 3D MPSoC cooling capacity. Consequently, FCAs and SC converters decrease the voltage drop of the GPU by 84mv compared to when no power is extracted from FCAs. In the CPU case, the voltage drop is almost eliminated in configuration B4 due to a high FCA power extraction. In all cases, FCAs and SC converters improve 3D MPSoC power performance with respect to cold plate cooling. In particular, FCAs significantly decrease PDN losses in configuration B4, which has the slowest liquid traversal in the channels and, therefore, the highest on-chip power generation.

FCAs' ability to decrease temperature and voltage drop presents an added leeway, which can be exploited in two different ways. From a *physical design* perspective, FCAs enable to relax the power grid requirements for each die (i.e., number and size of power delivery lines) while still achieving acceptable voltage levels. From a *performance* perspective, FCAs enable to increase the power consumption of dies by boosting their operating frequency. This thesis focuses on the second alternative. Indeed, Chapter 5 describes a run-time 3D MPSoC performance optimization methodology, leveraging FCAs cooling and power generation capabilities.

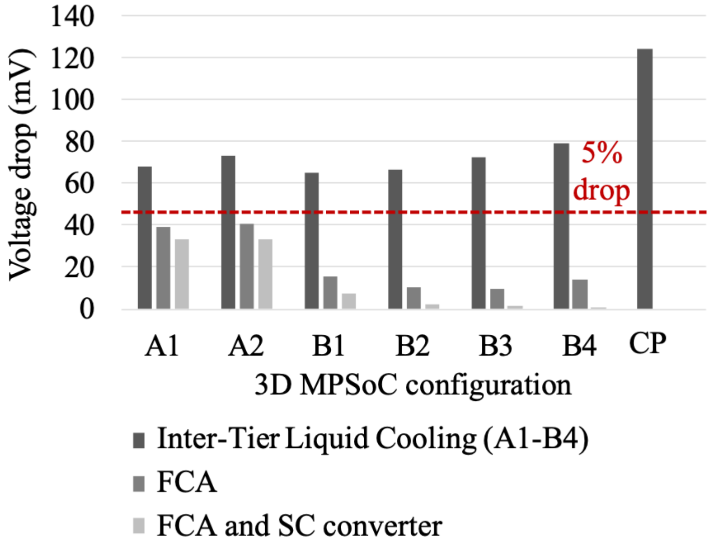


Figure 4.22: CPU voltage drop with FCAs and SC converters

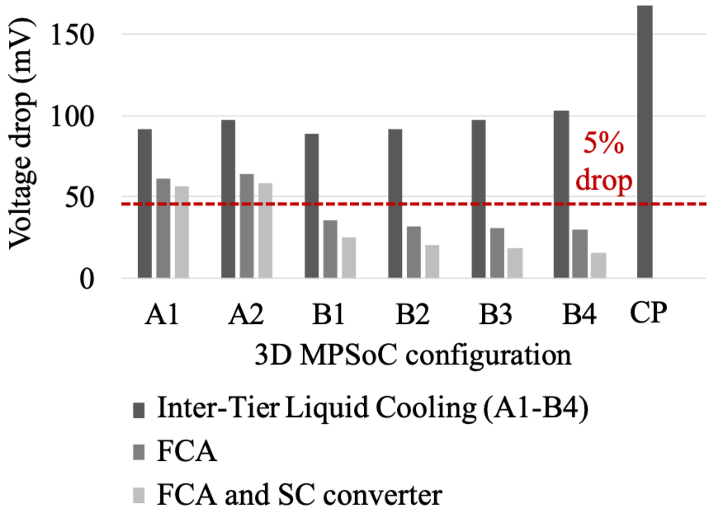


Figure 4.23: GPU voltage drop with FCAs and SC converters

4.5 Conclusion

In this chapter, I have evaluated the power generation performance of FCAs using high efficiency and low area SC converters. The proposed converters serve as an interface to connect FCA electrodes to 3D MPSoC PDNs. They allow operating flow cells at the voltage value that leads to optimal power generation capability, with V_{dd} supply of high-performance multi-core platforms. Consequently, the experimental results have shown that voltage regulators enable up to 123% higher on-chip power generation while limiting the cost in total chip area to less than 1.26%. The proposed converters also allow switching on and off the FCA power extraction, thus increasing between 1.8× and 4.5× the FCA reservoir lifetime for a processor duty-cycle of 50% and 20%, respectively.

Next, I have explored the design space of different FCA-related parameters to evaluate the achievable performances. Thus, the SC converter parameters were optimized to meet the area and output power requirements. Then, different FCA placements and liquid flow rates were evaluated for a high-performance and high-power 3D MPSoC. The experiments showcased how fine-grained thermal and power modeling can be employed to assess the impact of the different 3D MPSoC configurations on the overall performance. In particular, FCAs could decrease the temperature, power consumption, and voltage drop of heterogeneous 3D platforms. The results demonstrate the potential of FCAs to achieve power-efficient 3D MPSoCs targeting modern high-performance applications.

5 Run-time management strategies for 3D MPSoCs with FCAs

5.1 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 [77], alleviating the gap between processing and data access speed and enhancing the overall system efficiency. Nonetheless, high-performance 3D MPSoCs exhibit novel challenges related to a critical *heat generation* [78] and onerous *power supply and distribution* [79], particularly for deeply-scaled CMOS technologies.

Flow Cell Arrays(FCA) [44] promise to address the aforementioned 3D MPSoC challenges. Consisting of micro-fluidic channels in the silicon substrate of dies, they concurrently provide inter-tier liquid cooling and power generation capabilities due to heat-accelerated electrolyte reactions. They effectively reduce the temperature and leakage of dies. In addition, they partially supply logic gates when connected to the 3D power delivery network (PDN), reducing voltage supply drops and preventing timing delays that can cause performance degradation and system failures (Chapter 3). As the previous chapters have demonstrated through detailed analyses, integrated cooling and power delivery solutions based on FCAs can effectively solve the design challenges of 3D MPSoCs. Thus, they constitute promising avenues to disruptively increase the performance of 3D MPSoCs while mitigating thermal hotspots and power losses.

This chapter proposes to harness the potential of FCAs by maximizing the computing power of 3D MPSoC dies. In this context, it introduces a thermal and power-aware run-time performance management strategy for 3D MPSoCs, depending on their architecture and level of utilization. First, FCA capabilities are leveraged to increase the operating frequency of dies, improving the computing performance. In particular, a novel optimization methodology based on Model Predictive Control (MPC) calculates the highest applicable frequency boost of 3D MPSoC components while remaining within safe temperature, voltage, and timing margins.

This frequency optimization methodology is evaluated for different FCA flow rate settings. Secondly, the run-time performance management strategy is extended to co-configure the frequency of dies and the electrolytic flow rate of FCAs, involving inter-dependent cooling and power generation considerations, and enabling to reduce the cooling energy cost. In both cases, the management strategy is implemented in two phases: The *offline optimization* searches for the 3D MPSoC settings that enable maximum performance. Then, the *online controller* periodically applies the settings during run-time, according to workload power requirements.

In summary, the contributions of this chapter are the following:

- A novel thermal and voltage-aware MPC-based strategy is introduced to optimize the operation frequency of processing cores during run-time, using fine-grain thermal and electrical simulations. By exploiting FCA cooling and power generation capabilities, performance is enhanced without compromising the temperature, voltage, and timing of logic circuits.
- Targeting a 4-layer high-power 3D MPSoC, the performance management approach enables up to 24% faster clock frequencies. When optimizing the execution of state-of-the-art compute-intensive benchmarks, the online control speeds up workloads by up to 16% on a multi-core processing platform for an average utilization rate of 82%.
- The previous performance management strategy is extended by also optimizing the FCA flow rates depending on 3D MPSoC power requirements. In this case, the optimization algorithm considers the trade-offs between FCA cooling (hence leakage reduction) and on-chip power generation capabilities, both governed by the flow rate.
- Applied to the same 3D MPSoC, the flow rate and frequency optimization strategy enables operating the processing dies up to 19% faster than when using regular inter-tier liquid cooling. Thus, the online controller, in this case, speeds up the execution of benchmarks by up to 17%. These speedups are achieved while reducing FCA liquid pumping energy by up to 43%.

The rest of the chapter is organized as follows. Section 5.2 introduces 3D MPSoCs, their challenges, and state-of-the-art run-time management techniques used to alleviate them. The introduction also presents FCA technology as an enabler of high-performance 3D MPSoCs. Then, Section 5.3 introduces the target 3D MPSoC used as an experimental vehicle throughout the chapter. Section 5.4 presents the online frequency optimization strategy leveraging FCA cooling and power generation capabilities. Next, Section 5.5 presents the run-time flow rate optimization methodology, achieving the optimal operating frequencies using lower cooling costs. Section 5.6 shows that the proposed run-time strategies achieve a better performance than state-of-the-art thermal and power management techniques. Finally, Section 5.7 concludes the chapter.

5.2 Related work

5.2.1 3D MPSoC thermal and power challenges

3D stacking of dies interconnected using Through Silicon Vias (TSVs) allows to integrate heterogeneous components, possibly realized in different technologies, while achieving minimal inter-layer interconnect delays and very high bandwidths [77]. However, 3D Multi-Processor Systems-on-Chip (3D MPSoCs) present critical thermal and power management 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 [78]. This issue is exacerbated by high transistor densities and leakage currents in modern CMOS technologies. Traditional cooling techniques, such as fan-based cooling, struggle 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. Additionally, the need for power TSVs increases with the number of stacked dies. Those TSVs and the power delivery metal lines must supply very high currents, potentially incurring voltage drops throughout the 3D power grids. In turn, voltage drops affect the latency of logic and memory, possibly leading to timing failures [79].

The heat generation and power losses in 3D ICs must be addressed to achieve functional 3D MPSoCs. To improve thermal dissipation, designers have proposed solutions such as thermal TSVs and specific glue materials [14]. Alternatively, high-performance direct liquid cooling techniques can be used to extract excess heat [56]. However, these approaches generally require significant area and costly materials, and their efficacy drops with the number of stacked dies. Inter-tier liquid cooling, however, is a highly efficient and scalable technique to cool down 3D MPSoCs. It employs micro-channels in the silicon substrate of 3D MPSoC dies, through which a liquid flows, absorbing the generated heat [25].

To improve power delivery, the authors of [51] introduced a method to place extra power delivery TSVs limiting voltage drops. Similarly to thermal TSVs, they require a significant area, and their efficacy does not scale with the size and density of 3D stacks. Inter-tier liquid cooling alleviates power losses by decreasing leakage. In addition to that, Flow Cell Arrays (FCAs) also provide extra power to 3D MPSoCs (Section 5.2.3).

5.2.2 3D MPSoC run-time management strategies

At design time, cooling and power management techniques ensure that temperature and voltage constraints are met under worst-case conditions. However, 3D MPSoC operating conditions are application-dependent. Hence, several run-time thermal and power management techniques have been proposed to avoid under-utilizing computational components or over-utilizing cooling and power resources.

For example, the authors of [104] use a thermal-aware mapping algorithm and perform workload migration between hot and cool layers during run-time based on temperature information. The authors of [105] also propose an adaptive algorithm for multi-application 3D-NoC mapping to reduce latency and total system power under temperature constraints. Furthermore, [106] introduces a temperature-constrained power management scheme for 3D-MPSoCs, accounting for the activity of processing elements, their positions, and temperature margins.

A few run-time management strategies specifically target 3D MPSoCs with inter-tier liquid cooling. 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, the authors of [107] analyze the effect of various dynamic thermal management (DTM) methods and design a controller for energy-efficient thermal management with minimal performance degradation. Their approach combines flow rate adjustment, DVFS, and task scheduling to decrease cooling and computational power. Similarly, the authors of [108] couple liquid cooling control with several DTM policies to achieve reduction and balancing of temperature and increase the system lifetime and performance. They use a job scheduling strategy and dynamically adjust liquid flux to achieve a uniform temperature distribution. In [109], the authors propose a methodology to find the best thermal sensor locations, providing temperature information used by their thermal management policy. DVFS is used along with variable-flow liquid cooling to enable system power reduction and performance loss minimization.

The aforementioned 3D MPSoC run-time management policies deal with temperature and power loss reduction. However, they do not consider boosting the power performance, which can be achieved using FCAs.

5.2.3 3D MPSoC with integrated FCAs

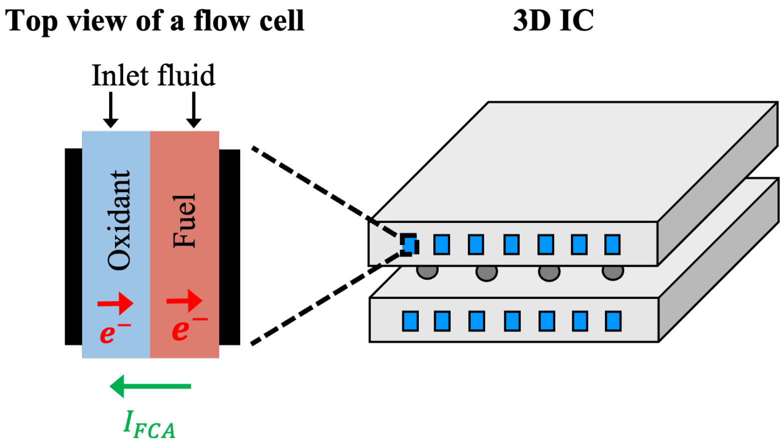


Figure 5.1: Flow cell array technology

FCA technology is a novel solution to both the power and thermal challenges of 3D MPSoCs, providing combined on-chip liquid cooling and electrochemical power generation [43]. FCAs use a technology similar to inter-tier liquid cooling [25]. The micro-channels used in this case are filled with an electrolytic liquid flow that produces an electrical current to supply logic gates, as illustrated in Figure 5.1. The electrochemical reaction rate increases with temperature, effectively transforming heat into available power. FCAs connect to the 3D MPSoC power grid through DC-DC voltage regulators, ensuring that they operate at their optimal voltage enabling maximum power generation, regardless of transient load changes in the chip (Chapter 4).

When integrated into an existing high-performance 3D MPSoC, FCAs can reduce the temperature by 50°C compared to traditional cooling solutions (Chapter 3). Additionally, they can recover up to 80% of voltage drop without increasing the density of power delivery components (e.g., TSVs and metal lines). These temperature and voltage drop reduction abilities of FCAs present an added leeway, which can be exploited in two different ways. From a *physical design* perspective, FCAs enable to relax power grid requirements for each die while still achieving acceptable voltage levels. From a *performance* perspective, FCAs provide an opportunity to increase the power consumption of dies while mitigating the performance losses related to voltage drop and temperature.

This chapter focuses on the second alternative. Indeed, it introduces novel run-time thermal and power management strategies to speed up the operation of high-performance dies (Sections 5.4 and 5.5). The strategies consider the architectures and usage conditions of a target 3D MPSoC architecture (Section 5.3). Then, they exploit both cooling and power generation capabilities of FCAs to boost the frequency of the computing dies.

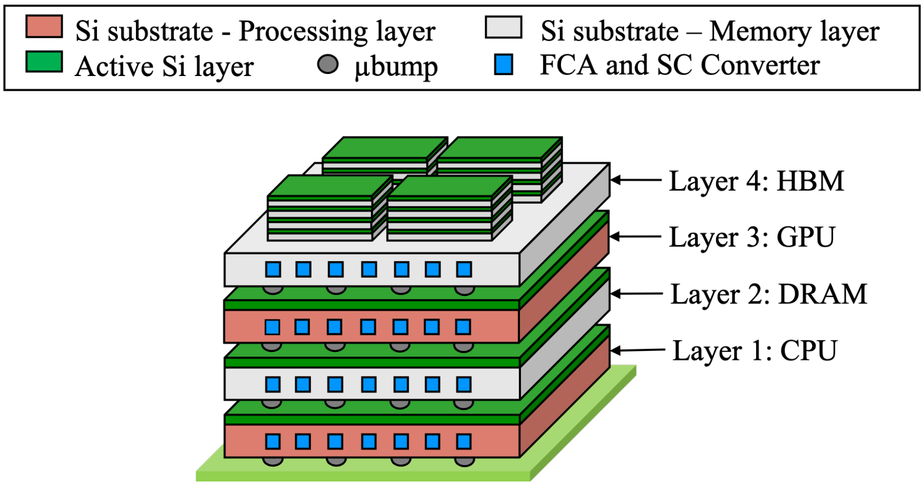


Figure 5.2: Target 3D MPSoC, including different types of memories, processing units, and the embedded FCA channels

5.3 Target 3D MPSoC

5.3.1 Stack composition

To exemplify the efficiency of the proposed run-time management strategies for 3D MPSoCs with integrated FCAs (Sections 5.4 and 5.5), a target high-performance four-layer stack is considered a test system for the remainder of this chapter. The stack is shown in Figure 5.2, and it comprises the following layers:

- The first (bottom) layer is modeled after AMD's Extreme Performance Yield Computing (EPYC) microprocessor, based on the Zen micro-architecture and fabricated using a 14nm FinFET process [100], with a total area of 757mm^2. It contains 32 high-performance cores, arranged as 4 Ryzen 8-cores clusters sharing one L3 cache. The cores operate at a base frequency of 2GHz and can be boosted up to 2.55GHz (all cores simultaneously) and 3GHz (one core only). The processor's peak power consumption is 180W, and its maximum-supported temperature is 81°C.
- The second layer contains an 8-channel DDR4-2666W [101], supported by the EPYC processor. The memory is fabricated using an 18nm 3-metal layer DRAM process. Each of the eight 16Gb DDRs occupies a total size of 81.28mm^2.
- The third layer is based on the NVIDIA V100 [102], a data center GPU designed to accelerate AI, HPC, and graphics. The NVIDIA V100 is composed of 640 Tensor cores and 5120 CUDA cores. This layer is fabricated using TSMC's 12nm FFT CMOS process and occupies a total size of 815mm^2. It consumes up to 300W and operates at a maximal temperature of 85°C. The GPU core frequency ranges between 1230MHz and 1380MHz.
- The fourth (top) layer is composed of four 2^{nd} generation HBM memories with 4 DRAM

layers each, providing the bandwidth requirement of the NVIDIA V100 GPU. Each HBM memory has a base size of $71mm^2$, fabricated using the 29nm DRAM process. The maximal power consumption of each HBM2 memory is $15W$ [6].

The above system architecture is based on a state-of-the-art CPU-GPU platform for high-performance computing [99]. By considering its implementation in 3D, a next-generation computing platform is anticipated. Thus, the target 3D MPSoC employs a combination of two state-of-the-art 3D integration techniques, namely: *chiplet-based integration* and *chip-on-chip bonding* through fine-pitched micro-bumps. The first one enables stacking multiple HBMs on a base logic die (top layer). The latter enables logic-on-logic integration and is used to stack the four 3D MPSoC layers, including the HBM active interposer and the package.

TSVs are arranged in groups (TSV islands), each delivering power to an independent power domain. Their diameter and pitch are both fixed to $5\mu m$. FCAs of $50\mu m$ width and $100\mu m$ height are etched in the silicon substrate of all the 3D MPSoC dies. They have a width, height, and pitch of 50, 100, and $50\mu m$. Each $200\mu m$-long flow cell section connects the corresponding power grid through a voltage regulator. FCAs are only electrically connected to the GPU and CPU, as the memories have considerably lower power requirements. An EMB MHIE centrifugal pump [110] is responsible for the fluid injection to all the flow cells. Then, normally closed valves (NCVs) [111] enable different flow rates for each die. We consider that the flow rate ranges between $40ml/min$ (the value achieving the maximum temperature) and $216ml/min$ (the maximal flow rate, as in [43]).

5.3.2 Target workloads

A range of state-of-the-art benchmarks targeting high-performance multi-core platforms are explored to evaluate the proposed run-time management strategies (Sections 5.4 and 5.5). These benchmarks are selected as they represent different power consumption profiles.

Indeed, a first set of Machine Learning (ML) applications specifically target the GPU. They are the following:

- **Inception V3 (I3)** is a very deep Convolutional Neural Network (CNN) architecture used for image recognition applications. An I3 model is trained on the GPU using the ImageNet data set [112].
- **Inception V4 (I4)** is a deeper and more uniform version of Inception V3 [113]. An I4 model is trained on the GPU using the ImageNet data set.
- **Resnet (RN)** is a computer vision deep CNN. An RN model is trained on the GPU using the ImageNet data set [114].
- **Deep Speech (DS)** is an end-to-end Deep Neural Network (DNN) used in automatic

speech recognition (ASR) [115]. A DS model is trained on the GPU using a large-scale data set of English readings.

- **Fairseq (FS)** is a sequence modeling neural network used for translation, language modeling, error correction, and other text generation tasks [116]. An FS model is trained on the GPU using a spoken language translation data set.

Next, the **SPEC** benchmark package is considered, targeting the CPU [117]. SPEC contains standardized CPU-intensive applications stressing the cores and memory sub-system [117]. Specifically, the following applications are evaluated:

- **Weather Research and Forecasting Model (wrf)** is a weather prediction system designed to serve both operational forecasting and atmospheric research needs.
- **Cactus Computational Framework (Ca)** is a physics benchmark consisting of a set of differential equations used to model black holes and gravitational waves.
- **NAMD** is a parallel program used to simulate large bio-molecular systems.
- **Parallel Ocean Program (pop2)** is a highly parallel program for simulating the earth's climate system.
- **ImageMagick (IM)** is a software suite to create, edit, compose and convert bitmap images.
- **Lattice Boltzmann Method (lbm)** is a program to simulate fluids in 3D.

Finally, the benchmark suite **Rodinia** is considered, targeting both the CPU and GPU. This benchmark supports heterogeneous computing [118]. In particular, the following applications are evaluated:

- **Needleman-Wunsch (NW)** is a non-linear global optimization method for DNA sequence alignment. This application is executed on both the multi-core CPU and GPU.
- **K-means (Km)** is a clustering algorithm used in data mining applications. It is evaluated on the multi-core CPU, taking advantage of multi-threading capabilities.
- **K-nearest neighbors (Knn)** is a machine learning algorithm used to solve classification and regression problems. It is evaluated on the GPU.

To characterize the previous benchmarks on the CPU and the GPU, they are executed in a real 2D platform [99] equivalent to the target 3D MPSoC in Section 5.3.1. Thus, their execution traces are recorded using time steps of 100ms and considering key utilization metrics. The characterization results on the CPU and GPU are presented in Figure 5.3 and Figure 5.4, respectively.

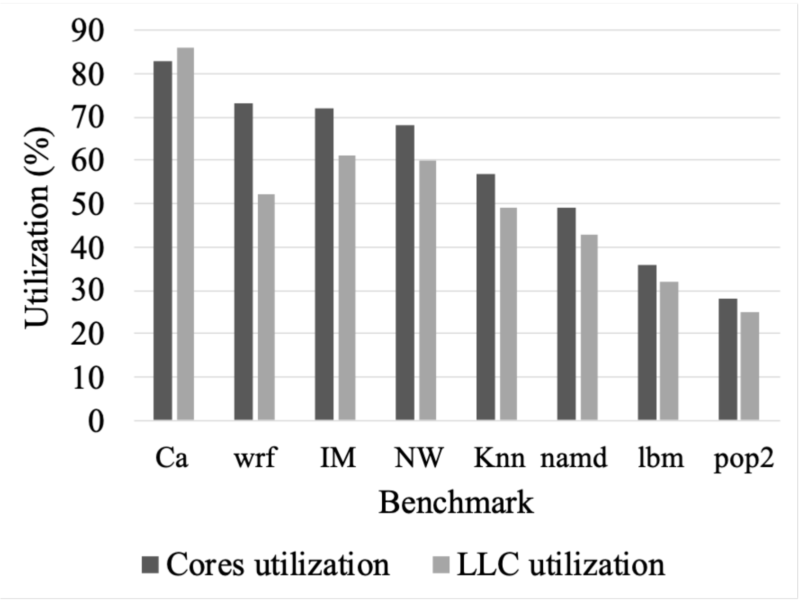


Figure 5.3: Workload utilization statistics on the CPU

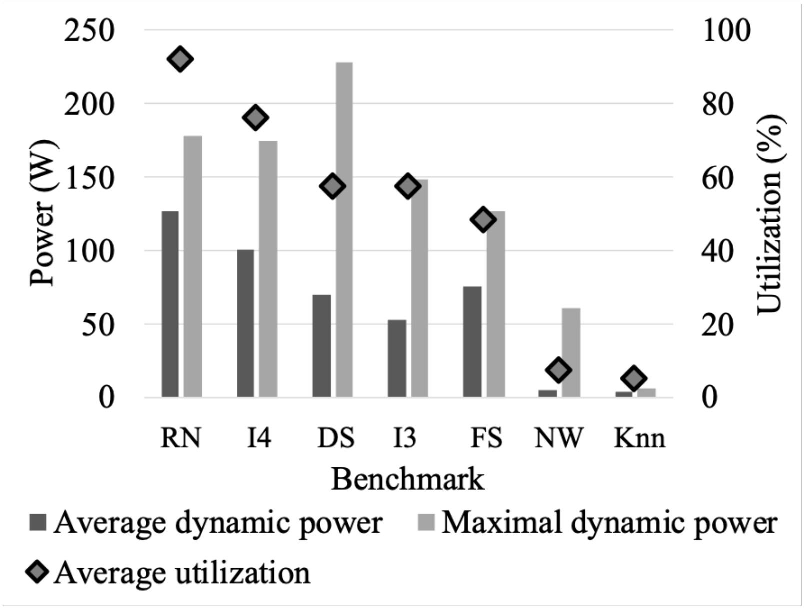


Figure 5.4: Workload utilization statistics on the GPU

For the AMD EPYC CPU case, the cores and last-level cache (LLC) utilization percentages are measured. These metrics serve to estimate the dynamic power consumption of each CPU component. The average benchmarks utilization rates are represented in Figure 5.3, ordered from high to low. A high CPU utilization characterizes compute-intensive benchmarks such as Cactus (Ca), weather forecasting (wrf), and ImageMagick (IM).

For the case of the NVIDIA V100 GPU, the total power consumption (including leakage) and the temperature are measured, enabling the estimation of the GPU dynamic power map. In addition, the GPU utilization is measured, indicating the percentage of time when kernels are being executed on the board. Figure 5.4 presents the average usage metrics of various benchmarks, ordered by GPU utilization level.

5.4 Thermal and power-aware 3D MPSoC frequency optimization

Harnessing FCA capabilities, this section introduces a run-time management approach to enhance the performance of the 3D MPSoC computing dies (Section 5.2), based on specific workload requirements. In particular, a model predictive control (MPC) algorithm is designed to boost the operation frequency of the CPU and GPU cores. MPC is an optimal control method to maximize a set of performance metrics for a dynamic system (e.g., 3D MPSoC operation frequencies) while respecting a particular set of constraints (e.g., temperature, voltage drop, and timing).

The MPC process provides feedback control actions that define the settings for the subsequent time periods [119]. MPCs can be implemented implicitly, embedding a solver that performs the optimization process in real-time, and computes the settings to apply to the system over the next period. Alternatively, the optimization outputs can be pre-computed offline and accessed by a control module through a look-up table (LUT). This second approach is referred to as an explicit MPC solver. It is an appropriate strategy for the proposed real-time 3D MPSoC frequency optimization, as it enables a smooth thermal control with minimal computation costs and delays. The details of the proposed offline frequency optimization algorithm and online MPC implementation are presented in Sections 5.4.1 and 5.4.2, respectively.

5.4.1 Offline frequency optimization

A frequency optimization algorithm is performed offline for each die to determine the applicable frequency boosts under different utilization scenarios. It serves to fill the LUTs of the explicit MPC implementation. First, the algorithm receives as input the utilization of the die from the scheduler (for the experiments in Section 5.4.3, utilization values are derived from performance counters). Then, it evaluates the frequency increase that can be applied to the modeled 3D MPSoC die according to its power consumption. In particular, the algorithm calculates the temperature levels given the geometry of the 3D MPSoC, the FCA topology, and coolant flow. Moreover, it accounts for the effect of voltage drops on the timing characteristics. Finally, it dictates the clock frequencies of CPU and GPU cores for different conditions, such that temperature and timing violations are avoided and performance is maximized. The detailed steps performed by the MPC solver are described in the following [1]:

1. **Initialization**: The algorithm estimates the initial dynamic power maps $P_{dyn,init}$ of the 3D MPSoC die based on the utilization metrics extracted from the schedulers and performance counters. Because the CPU and GPU include different performance counters, the dynamic power $P_{dyn,init}$ is calculated differently for the two dies and their respective memories. In the case of the CPU, the number of instructions directly reflects on the utilization level of each core ρ_{core}. Similarly, the number of LLC and DDR accesses

---

[1] In all the equations, vectors and matrices are denoted by capital letters whereas scalar values are indicated by lower-case letters.

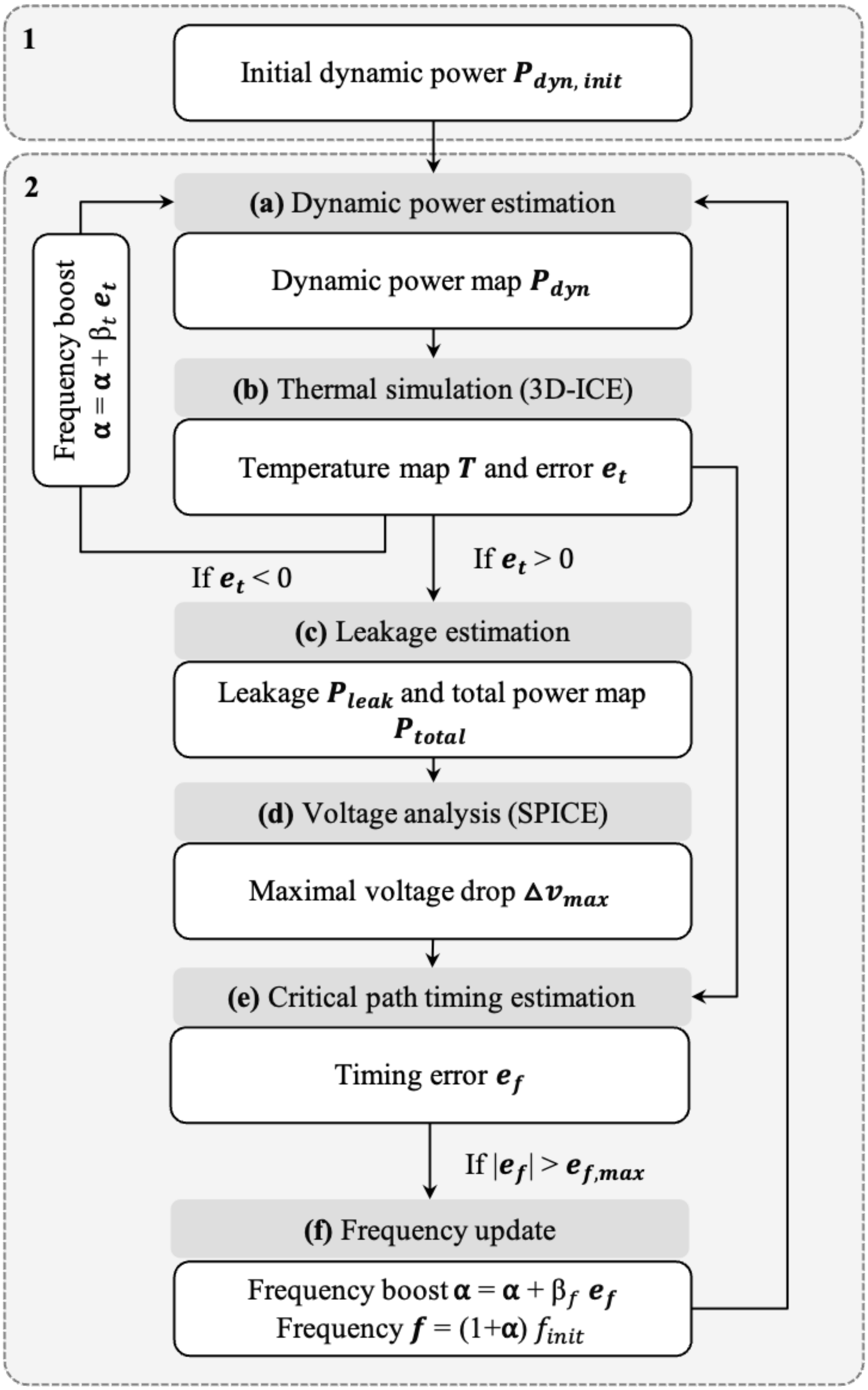

1
Initial dynamic power $P_{dyn,init}$
2
(a) Dynamic power estimation
Dynamic power map P_{dyn}
Frequency boost $\alpha = \alpha + \beta_t e_t$
(b) Thermal simulation (3D-ICE)
Temperature map T and error e_t
If $e_t < 0$
If $e_t > 0$
(c) Leakage estimation
Leakage P_{leak} and total power map P_{total}
(d) Voltage analysis (SPICE)
Maximal voltage drop Δv_{max}
(e) Critical path timing estimation
Timing error e_f
If $|e_f| > e_{f,max}$
(f) Frequency update
Frequency boost $\alpha = \alpha + \beta_f e_f$
Frequency $f = (1+\alpha) f_{init}$

Figure 5.5: MPC explicit solver with (1) an initialization phase, followed by (2) a frequency optimization loop

reflect on their utilization level ρ_{cache} and ρ_{DDR}. Hence, the power consumptions of the cores ($p_{dyn}(core)$), LLC ($p_{dyn}(LLC)$), and DRAM ($p_{dyn}(DDR)$) are calculated as follows:

$$p_{dyn}(core) = \rho_{core} * p_{max,core} \tag{5.1}$$

$$p_{dyn}(LLC) = \rho_{LLC} * p_{max,LLC} \tag{5.2}$$

$$p_{dyn}(DDR) = \rho_{DDR} * p_{max,DDR} \tag{5.3}$$

The cores and LLC dynamic power consumption values are then mapped to the CPU layout to construct the dynamic power map. Then, the DDR power consumption is mapped to its area, assuming a uniform data access pattern.

In the case of the GPU, the dynamic power $p_{dyn,init}(GPU)$ is extracted from the total power $p_{total,init}(GPU)$ by subtracting the leakage $p_{leak,init}(GPU)$. The leakage at the initial temperature $t_{GPU,init}$ is estimated according to the leakage per transistor gate width i_{off}, the effective transistor gate width w_{eff}, the transistor density ρ_{trans}, and the total die area A. The transistors are typically sized to achieve $10nA/\mu m$ leakage per gate width for low and medium performance, and $20nA/\mu m$ leakage per gate width for high performance [53], at the reference temperature of 25°C. Then, this value increases exponentially with temperature, as represented in Figure 5.6. Hence, the total leakage power of the GPU for a temperature t_{GPU} is calculated as follows:

$$p_{leak,init}(GPU) = i_{off}(t_{GPU,init}) * w_{eff} * \rho_{trans} * A \tag{5.4}$$

$$p_{dyn,init}(GPU) = p_{total,init}(GPU) - p_{leak,init}(GPU) \tag{5.5}$$

The initial dynamic power of the GPU is then mapped to its floorplan according to the power consumption percentage of the different components [120]. Hence, the initial dynamic power map $P_{dyn,init}(GPU)$ is obtained:

$$P_{dyn,init}(GPU) = p_{dyn,init}(GPU)/p_{max,GPU} * P_{max,GPU} \tag{5.6}$$

Finally, the initial dynamic power map of the four HBM memories is estimated according to the GPU memory utilization percentage ρ_{HBM} extracted from the performance counters. As in the case of the DDR, we assume a uniform data access pattern:

$$P_{dyn}(HBM) = \rho_{HBM} * P_{max,HBM} \tag{5.7}$$

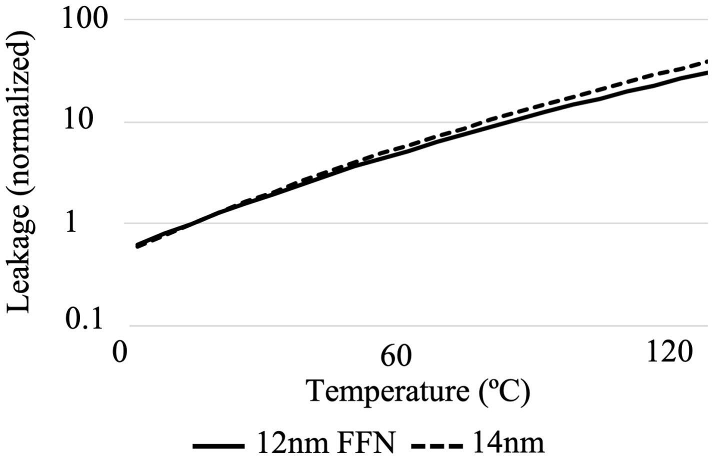


Figure 5.6: Transistor leakage vs. temperature (normalized), extracted from SPICE simulations of a standard gate using the 14nm and 12nm FFN technology nodes [72]

2. **Optimization Loop** (explicit solver): After estimating the initial dynamic power map of the 3D MPSoC die, a series of steps are performed recursively until convergence, as shown in Figure 5.5[2]. The optimization loop searches for the maximal applicable frequency increase ratio, which achieves the desired timing and temperature of the computing die. The detailed steps of the optimization loop are presented in the following, using the order and numbering in Figure 5.5, and performed similarly for all 3D MPSoC computing dies:

   (a) **Dynamic power map estimation**: In this first step, the dynamic power map of each computing die is scaled according to its frequency increase ratio α. Thus, a quadratic frequency-power relationship is used, which generally applies to many-core high-performance ICs [121]:

   $$P_{dyn} \sim f^2 \tag{5.8}$$

   Hence, the dynamic power map of each die when $f = (1+\alpha) f_{init}$ is calculated as:

   $$P_{dyn} = (1+\alpha)^2 P_{dyn,init} \tag{5.9}$$

   During the first iteration t_0*, no frequency boost is applied. Hence* α_{t_0}*=0.*

   (b) **Thermal simulation**: In this step, the temperature of each 3D MPSoC computing die is evaluated in case its power consumption profile corresponds to the previously calculated power map. Furthermore, the power consumptions of the memory dies are pessimistically considered maximal, as they are not bottlenecks for the thermal behavior of the stack. Hence, the compact thermal simulator for

---

[2] In Figure 5.5, all the variables that are estimated during the MPC solver flow are in **bold**

liquid-cooled 3D ICs 3D-ICE [33] is used. 3D-ICE generates a 3D model containing multiple layers of thermal cells, then solves transient heat flow equations and outputs the fine-grain temperature map of each 3D MPSoC die T.

If the maximal temperature $max(T)$ of any die exceeds the constraint value T_{max}, its frequency increase ratio is adjusted according to the temperature error $\epsilon_t = max(T) - T_{max}$:

$$\alpha = \alpha + \beta_t * \epsilon_t \tag{5.10}$$

The algorithm then iterates back to step (a).

If no temperature violation occurs, the algorithm proceeds to the next step (c).

(c) **Leakage map estimation**: Next, the leakage map P_{leak} is determined for each 3D MPSoC die according to its thermal map T. Similarly to Equation 5.4, the leakage value ${P_{leak}}^{i,j}$ of a cell with coordinates (i, j) is estimated as follows:

$${P_{leak}}^{i,j} = i_{off}(T^{i,j}) * w_{eff} * \rho_{trans} * A_{cell} \tag{5.11}$$

The total power map is then calculated for a given frequency boost ratio α as:

$$P_{total} = P_{dyn} + P_{leak} \tag{5.12}$$

In fact, the idle power is not included as the frequency boost is not applied in case of core inactivity. Additionally, idle components never represent thermal or voltage hotspots, and they do not influence the temperature and voltage level at the hotspots.

(d) **Voltage analysis**: After computing the total power maps of the dies, the fine-grain 3D MPSoC power network electrical model (with the FCAs) is updated by assigning the corresponding values to the loads. Therefore, this model is simulated using HSPICE to obtain the voltage map of the target die. Then, the critical voltage drop value Δv_{max} is extracted.

(e) **Critical path timing estimation**: In this step, the timing of the most critical path of the target 3D MPSoC die is compared to the clock period (i.e., frequency). This step indicates if timing violations can potentially occur and compromise the chip operation. Hence, the critical path timing is estimated with respect to voltage and thermal conditions. This relationship is characterized based on a canary circuit consisting of a 64-bit full adder, implemented in a 28nm CMOS technology. Results of this exploration are shown in Figure 5.7. In this figure, v_{dd} is normalized to its value for the technology library. Then, the timing is normalized to its value at the maximum temperature, and minimal voltage when the die is cooled using a high-performance cold plate-based liquid cooling solution [56], in an equivalent 2D system. This value represents the nominal operation frequency, assuming the signal integrity of the circuit in this scenario. Therefore, the critical path timing

τ_{max} is extracted for each target 3D MPSoC die. To represent worst-case operation conditions, it is pessimistically assumed that the critical path corresponds to the power and thermal hotspot of the die (i.e., highest voltage drop Δv_{max} and temperature $max(T)$). Thus, the timing τ_{max} is compared to the clock period $((1+\alpha)f)^{-1}$ and the frequency/timing error e_f is computed:

$$e_f = (1+\alpha)f - \frac{1}{\tau_{max}} \tag{5.13}$$

If the timing error is negative (i.e., no timing violation is present) and it is above a certain threshold e_{max}, the optimization loop is interrupted. The optimal operation frequency of the die given the required workload utilization rate is set to the value:

$$f_{opt} = (1+\alpha)f \tag{5.14}$$

(f) **Frequency update**: If the timing error is positive (i.e., a timing violation is possible) or is lower than the threshold e_{max} (indicating an overly conservative clock frequency), the optimization process proceeds by updating the candidate frequency value using a gradient descent methodology. The optimization parameters β and e_{max} are chosen so that the solution converges. The next frequency increase ratio is calculated with respect to the timing error as follows:

$$\alpha = \alpha + \beta_f e_f \tag{5.15}$$

The algorithm then iterates back to step (a). At the end of the optimization loop, the closest value is selected from the range of supported operating frequencies.

5.4.2 Online frequency control

Generally, the inter-layer thermal dissipation creates a correlation between 3D MPSoC temperature and power consumption levels. Hence, to compute the optimal frequency of dies, it is necessary to analyze layers simultaneously and consider all their activity levels. However, to reduce the set of inputs in the explicit MPC implementation and to lower the characterization effort, the frequency optimization of the GPU and the CPU is performed independently by the MPC explicit solver. Hence, when optimizing each computing die, it is pessimistically assumed that the rest of the dies are in full utilization at all times. In fact, the heat absorption capabilities of FCAs limit heat exchanges between dies, particularly given the distance separating the two most power-consuming dies in the target 3D MPSoC (i.e., the CPU and GPU). In addition, the power consumption of the memories has a negligible impact on the 3D MPSoC thermal performance. Finally, all the dies have independent power networks.

A high-level view of the implemented explicit MPC is shown in Figure 5.8. The MPC module periodically receives utilization data from the CPU and GPU schedulers/sensors. In particular,

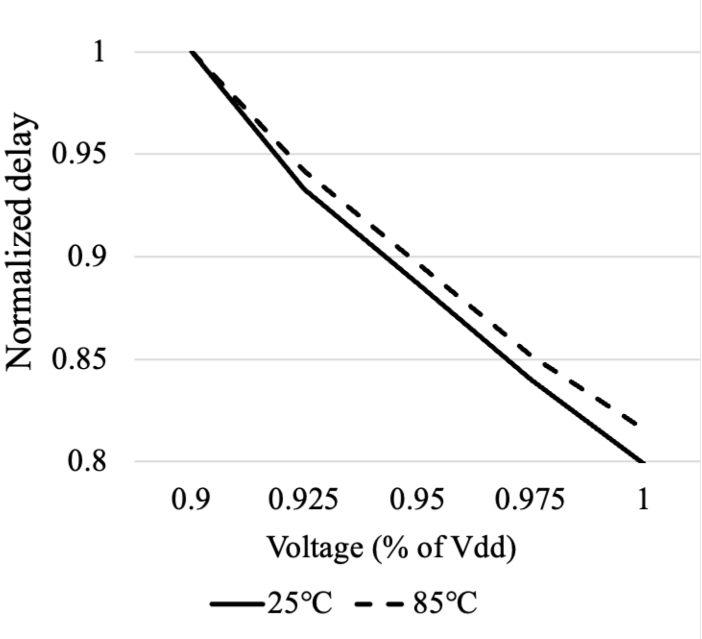


Figure 5.7: Critical path delay of canary circuit

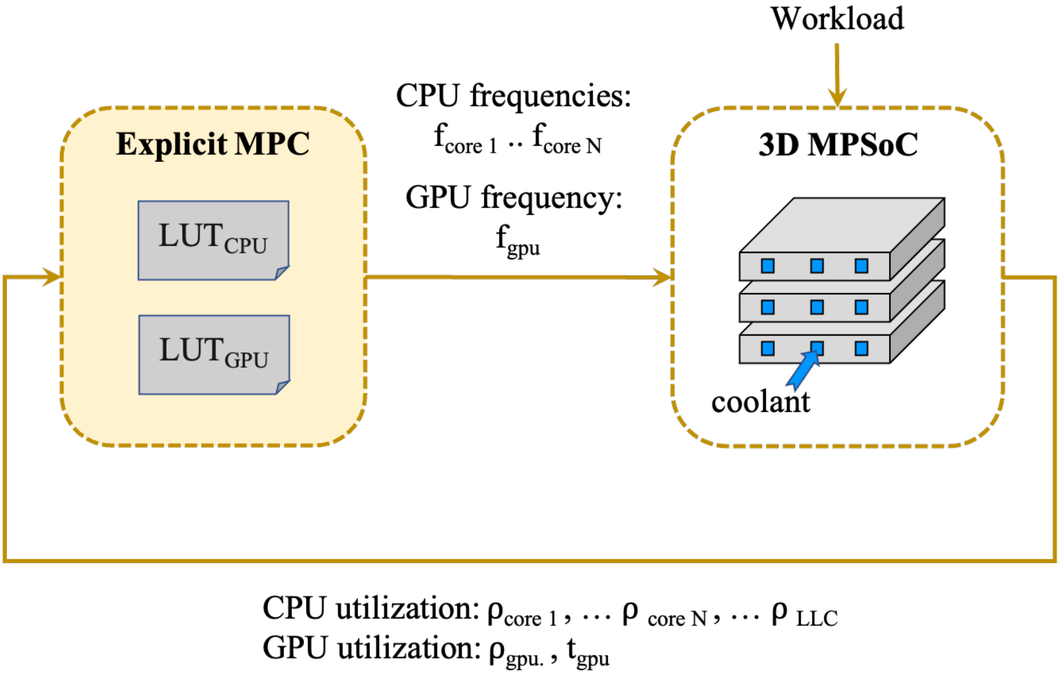

Workload
CPU frequencies:
$f_{core\ 1} \cdot\cdot f_{core\ N}$
GPU frequency:
f_{gpu}
Explicit MPC
LUT_{CPU}
LUT_{GPU}
3D MPSoC
coolant
CPU utilization: $\rho_{core\ 1}, \ldots \rho_{core\ N}, \ldots \rho_{LLC}$
GPU utilization: $\rho_{gpu.}, t_{gpu}$

Figure 5.8: Implementation of online frequency controller for 3D MPSoC computing dies based on explicit MPC

Table 5.1: Frequency optimization results (reduced LUT)

FCA flow rate			216 ml/min		108 ml/min		54 ml/min		40 ml/min	
	Util.	T_{init}	α_f	T_{max}	α_f	T_{max}	α_f	T_{max}	α_f	T_{max}
CPU	20%	-	24%	35°C	24%	41°C	24%	57°C	24%	61°C
	50%	-	24%	38°C	24%	46°C	23%	60°C	23%	68°C
	80%	-	21%	39°C	22%	48°C	23%	64°C	22%	75°C
GPU	20%	40°C	24%	34°C	24%	41°C	23%	56°C	23%	59°C
	20%	70°C	24%	34°C	24%	41°C	24%	57°C	24%	60°C
	50%	40°C	19%	40°C	19%	49°C	20%	65°C	20%	74°C
	50%	70°C	24%	38°C	24%	46°C	23%	60°C	23%	68°C
	80%	40°C	10%	43°C	11%	53°C	10%	72°C	9%	83°C
	80%	70°C	14%	41°C	15%	51°C	15%	67°C	12%	82°C

it takes as inputs the utilization percentage of CPU cores, the utilization percentage of CPU last-level caches (LLCs), the utilization of the GPU, and the GPU temperature from embedded sensors. This data is used to estimate the available temperature and timing leeway and set the clock frequencies for the CPU cores and the GPU according to the offline optimization.

5.4.3 Experimental results on target 3D MPSoC

5.4.3.1 Frequency boosts

The optimal CPU and GPU frequencies are evaluated for different utilization levels and FCA flow rate configurations. These results are stored in two separate LUT for the CPU and the GPU. Subsets of the LUTs are presented in Table 5.1, where CPU and GPU utilization are uniform between all the cores.

Results indicate that the proposed thermal and timing-aware optimization methodology enables to speed up the CPU operation by up to 24% when it is utilized 50% of the time and in case the FCA flow rate is set to the highest value [43] (i.e., 33W higher dynamic power consumption). The frequency boosts are possible thanks to the combined cooling and power generation of FCAs. The maximal frequency boost is bounded by the value indicated in the CPU specifications. Then, for a CPU utilization percentage of 80%, FCAs enable up to 22% frequency boost (i.e., 45W higher dynamic power consumption). As the dynamic power consumption in this scenario is higher, the voltage drops and temperatures are more critical, leading to a lower applicable frequency boost. For all the flow rate configurations and utilization scenarios, the temperature of the CPU remains below 75°C, which is lower than the constraint defined by the designers.

In the case of the GPU, the frequency can be boosted between 12% and 19% (i.e., between 23W and 38W higher dynamic power consumption) for the different FCA flow rate configurations when its power consumption corresponds to 50% of its thermal design power (TDP) and its initial temperature is 40°C. For an initial temperature of 70°C, the possible frequency boost is between 18% and 24% (i.e., between 38W and 52W higher dynamic power consumption). The initial leakage, in this case, is higher, and FCA cooling helps to reduce it, offering more opportunities to boost the dynamic power consumption.

In general, higher frequency boosts are observed for the CPU compared to the GPU due to significantly lower power consumption and stress on the power delivery network. In addition, FCA flow rate configurations with a high cooling capability (e.g., 216 ml/cm) enable a high-frequency boost while achieving temperatures as low as 42°C and 39°C for the GPU and CPU, respectively. In the case of this configuration, the speed-up is possible thanks to the FCA capacity to reduce leakage, enabling considerably higher dynamic power consumption. In contrast, flow rate configurations with lower FCA cooling capacity achieve GPU temperatures up to 82°C but comparable frequency boost ratios. In these configurations, FCAs produce more power due to higher temperatures. Therefore, they enable more computing capacity without further stress on the power delivery grid, voltage drop across the dies, and FCA cooling cost.

5.4.3.2 Workload speedups

The explicit MPC from Section 5.4.2 is simulated when running the benchmarks in Section 5.3.2 on the target 3D MPSoC in Section 5.2. Results are shown for the configuration with the lowest flow rate (40ml/min in Table 5.1). This configuration generally enables boosting the dies frequencies while requiring a low cooling cost. For comparison, workload speedups are also measured when applying a fixed frequency boost throughout the execution. Hence, two scenarios are considered:

- Prior knowledge of workload requirements is available. In this scenario, the frequency boost is selected based on the peak power/utilization of the workload. The corresponding results are labelled *fixed freq. boost* in Figures 5.11 and 5.13.
- No knowledge of workload requirements is available. In this scenario, the minimal frequency boost is selected, corresponding to 100% utilization. The results, in this case, are labeled *min. freq. boost* in Figures 5.11 and 5.13.

In both these scenarios, the selected frequency boosts ensure that no constraint violations occur during the full workload execution.

To estimate the execution of the benchmarks using the proposed frequency control strategies (i.e., MPC boost, fixed boost, and minimal boost), the execution traces are first recorded in the equivalent 2D system in sampling windows of 100ms (the minimal value dictated by the sensors and counters used to gather the experimental data). Then, the execution on the 3D MPSoC is simulated by compressing the original traces according to the speedup rates determined by the MPC. Thus, it is pessimistically assumed that the frequency boost only applies to the computing dies and that the 3D MPSoC has the same memory bandwidth as the original 2D system. Hence, the same task schedule is considered when running the benchmarks on the 3D MPSoC. This way, only the computing time is compressed for each sampling step, as illustrated in Figure 5.9. The operation is repeated until the completion of the workload, enabling to compute the speedup using the proposed MPC.

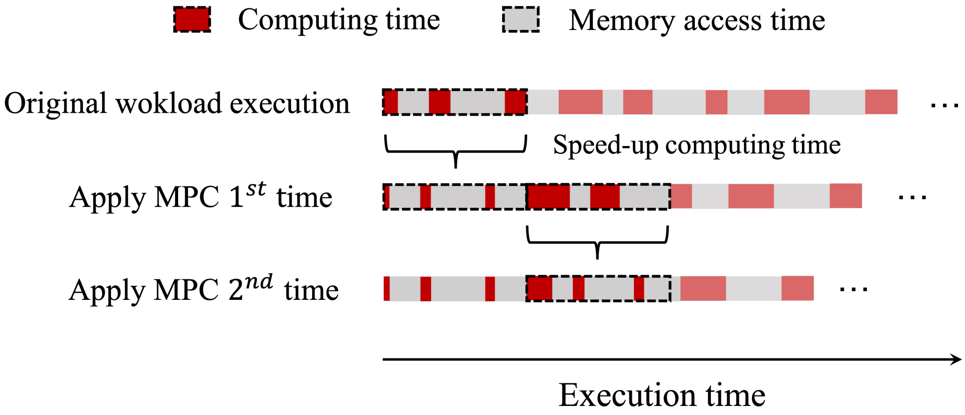


Figure 5.9: Evaluation of workload speedup at runtime

1. **Benchmark speedup on the CPU**:

   The average frequency boosts and total workload speedups on the CPU are shown in Figures 5.10 and 5.11, ordered by core utilization percentage. For most benchmarks, the MPC enables, on average 23% faster CPU operation, as indicated in Table 5.1. However, *compute-intensive* benchmarks present the highest overall speedup due to their high utilization percentage. For instance, the SPEC Cactus framework (Ca) achieves 16% execution speedup at 83% average core utilization. In contrast, the core frequency acceleration does not significantly improve the total execution time of *least compute-intensive* benchmarks, as it is dominated by memory access time. For example, the SPEC pop2 benchmark achieves only 6% total execution speedup with an average core utilization percentage of 28%.

   Compared to the proposed MPC, the alternative strategies achieve lower but comparable workload speedup results, as the optimal frequency boost in all CPU utilization scenarios is between 20% and 24% (Table 5.1). In particular, a maximum difference of 3% and 2% are observed for the average frequency boost and total speedup, respectively.

2. **Benchmark speed-up on the GPU**: In the case of the GPU, the average frequency boosts and achieved benchmark speedup values are shown in Figures 5.12 and 5.13, ordered by utilization level. The GPU runs, on average, between 16% and 23% faster using the proposed MPC. Typically, the benchmarks with a *high GPU utilization rate* and dynamic power consumption (Figure 5.4) are less eligible for a high-frequency boost. This is because they can otherwise overly heat the GPU and exhibit critical voltage drops. Particularly, the Resnet (RN) CNN training, which has a peak dynamic power

requirement of 230W on the NVIDIA V100, can run on average with a 16% faster clock in a 3D system with FCAs, compared to other benchmarks that enable 20% or higher speedup rates. However, the RN training benefits from the highest overall workload speedup. This is because its percentage of computing time versus memory access time (GPU utilization rate) is the largest of all GPU benchmarks. Conversely, benchmarks with a *low GPU utilization rate*, such as Needleman-Wunsch (NW) or K-nearest-neighbours (Knn), can run at the highest frequency (with up to 23% boost). Still, their total execution speedup is lower than 6%.

The alternative strategies achieve lower speedup results than the proposed MPC, as the optimal frequency boost can be anywhere from 3% to 24%. Hence, for workloads with high utilization peaks (RN, I4, DS), the highest peak dictates the *fixed frequency boost*, decreasing by up to 15% of the average frequency compared to MPC. In the case of workloads with a lower utilization (I3, FS, NW, Knn), the *fixed frequency boost* achieves comparable speedups because their highest utilization peaks still enable high-frequency boosts. However, the *minimal frequency boost* strategy limits the overall speedups, as it accounts for the maximum possible workload utilization, which is never reached in practice.

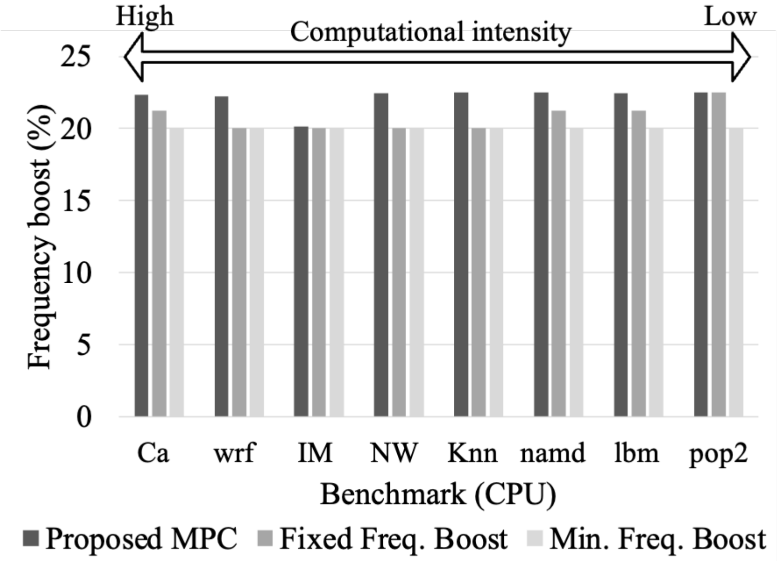


Figure 5.10: Average frequency boost results on CPU

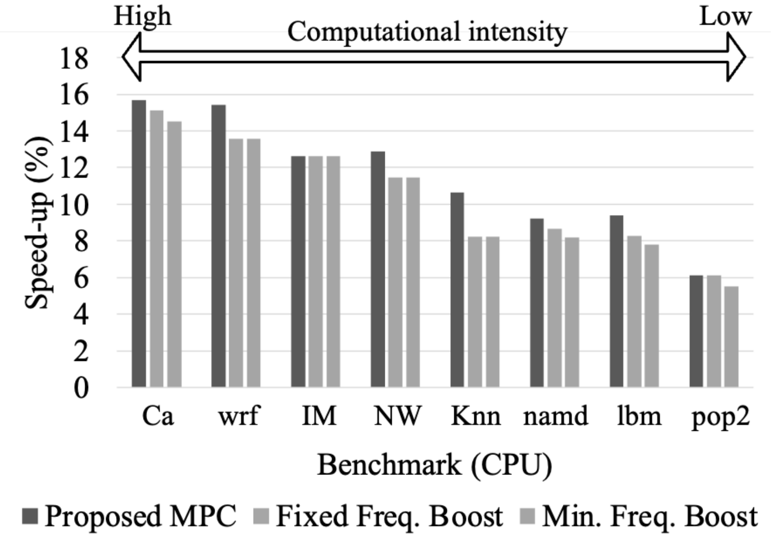


Figure 5.11: Workload speedup results on CPU

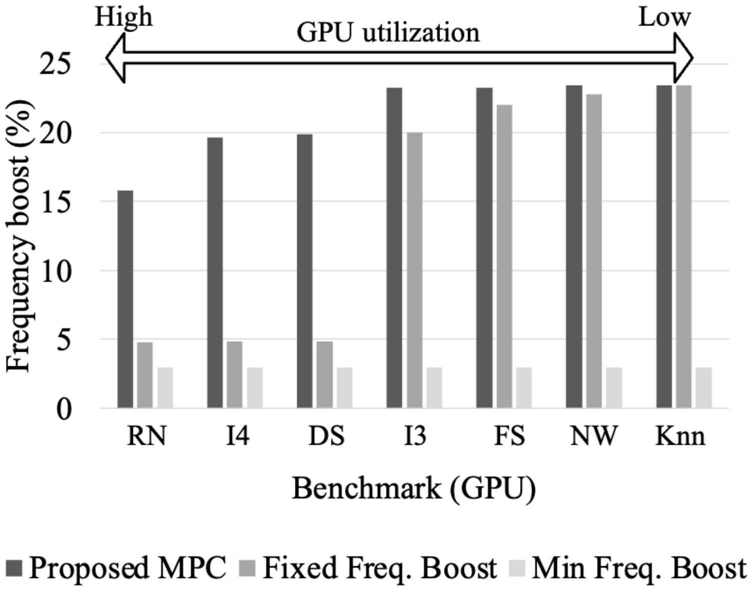


Figure 5.12: Average frequency boost results on GPU

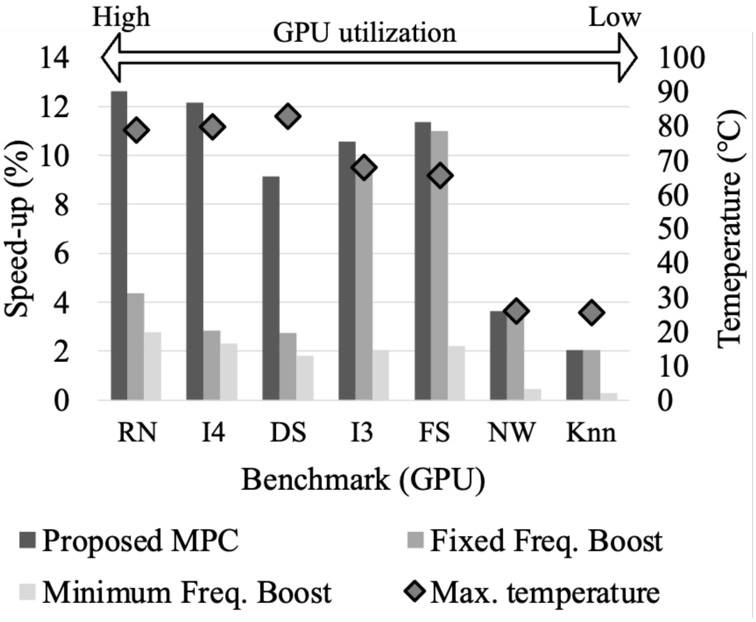


Figure 5.13: Workload speedup results on GPU

The previous simulations indicate that the computation time can be accelerated by up to 24% for both the CPU and GPU using the proposed MPC. However, the maximal boosts differ between the evaluated FCA flow rate settings in Section 5.4.3.1. Depending on the 3D MPSoC usage scenario, the fastest operation can be achieved for specific FCA cooling conditions. In this regard, Section 5.5 proposes to configure the flow rate at runtime, enabling the optimal computation speed at the lowest cooling cost, depending on workload requirements. Hence, the offline optimization and online control strategy in Sections 5.4.1 and 5.4.2 are extended to configure both the 3D MPSoC computing frequencies and the FCA flow rates.

5.5 Thermal and power-aware flow rate and frequency co-optimization

Using FCA's effective cooling and power generation capabilities, the previous Section 5.4 demonstrated that online frequency control enables speeding up the execution of workloads in 3D MPSoCs. However, FCAs expose a trade-off between their two main capabilities, both governed by the flow rate. On the one hand, higher flow rates enhance heat absorption and considerably decrease 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 FCA frequency boost performance.

In this section, the previous run-time management strategy from Section 5.4 is extended to harness the potential of FCAs while involving the previous inter-dependent thermal and electrical considerations. Hence, the proposed approach targets increasing 3D MPSoCs computing performance by co-configuring the electrolytic flow rate and the operating frequencies of dies. Similarly to Section 5.4, the run-time management strategy is implemented in two phases. The *offline optimization methodology* uses fine-grain thermal and electrical modeling and analysis to determine the lowest applicable FCA flow rates that enable the highest operating frequency of dies, depending on their utilization levels. Then, the MPC-based *online controller* periodically selects and applies the optimal flow rate and frequency settings during run-time from the pre-computed LUT. The proposed optimization methodology and controller implementation details are described in Sections 5.5.1 and 5.5.2, respectively. They enable smooth control with minimal computation costs and delays.

5.5.1 Offline frequency and flow rate optimization

The 3D MPSoC optimization algorithm 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 comprises two nested loops, as illustrated in Figure 5.14. 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.

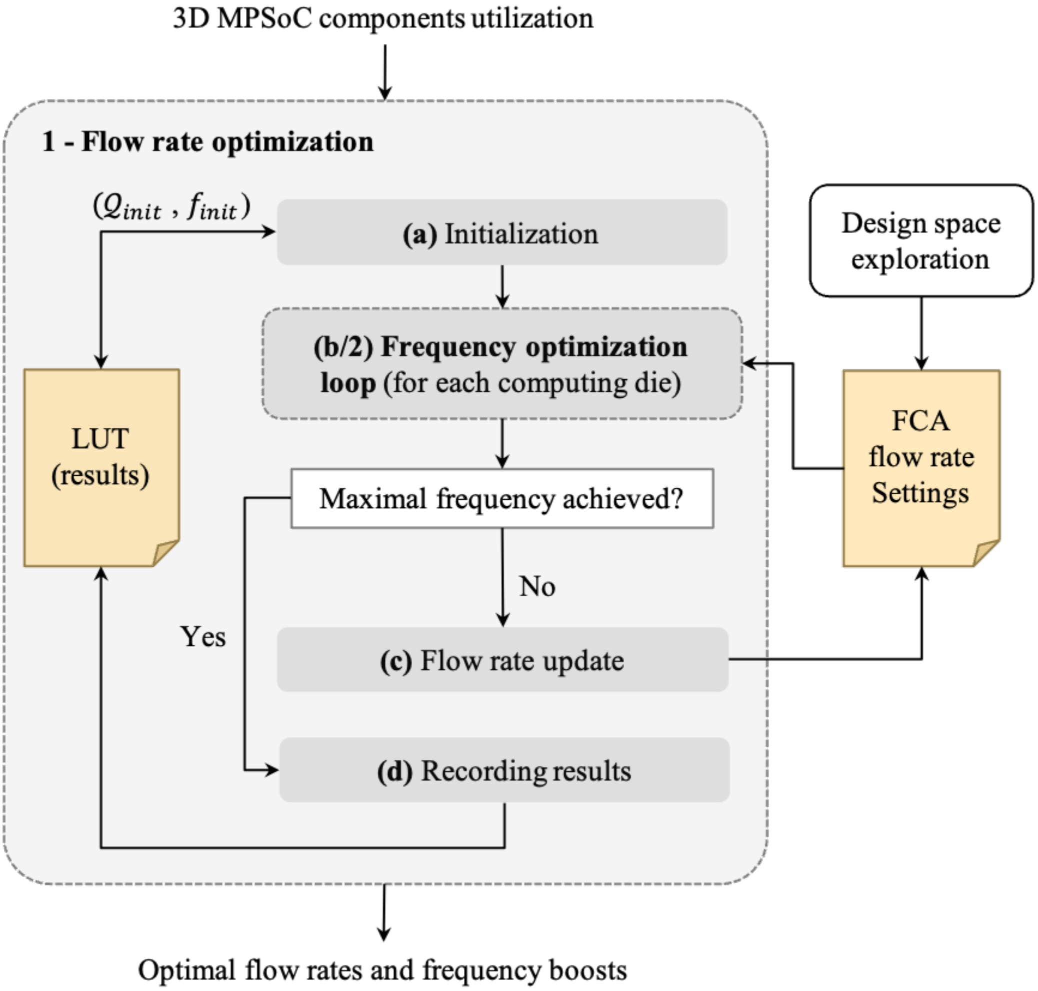

3D MPSoC components utilization
1 - Flow rate optimization
(Q_init , f_init)
(a) Initialization
(b/2) Frequency optimization loop (for each computing die)
Maximal frequency achieved?
No
Yes
(c) Flow rate update
(d) Recording results
LUT (results)
Design space exploration
FCA flow rate Settings
Optimal flow rates and frequency boosts

Figure 5.14: Offline frequency and flow rate optimization methodology

1. **Flow rate optimization considering the entire 3D MPSoC**

   First, a design-space exploration determines the set of possible combinations of flow rate values for each 3D MPSoC layer $Q = [q_1 \ldots q_K]$. To minimize the cooling cost, the selected combinations $\{Q\}$ achieve different cooling performances while having the lowest total flow rate. 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 $P = [\rho_1 \ldots \rho_N]$. 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:

   (a) 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 scenario P, the initial settings that correspond to the optimal values for the nearest neighbor P^* from previously calculated results are selected for faster convergence.

(b) After defining the FCA flow rates and initial computing dies frequencies, the algorithm performs the frequency optimization for each computing die (described next).

(c) If the optimal frequencies are not yet achieved using the current FCA flow rates combination Q, the algorithm updates these values. In particular, it considers a neighboring state Q^* and iterates back to step (b) to evaluate the 3D MPSoC performance in this scenario.

(d) If the optimal frequencies $F_{opt} = (f_1 \ldots f_N)$ are achieved for a given FCA flow rates combination Q, the optimization loop terminates. The optimal flow rates and frequency settings are then recorded.

2. **Frequency optimization applied individually for each die**

   This loop evaluates the applicable computing die frequency increase ratio (boost) 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 same sequence of steps in Section 5.4.1 is performed for each computing die recursively until convergence.

5.5.2 Online frequency and flow rate control

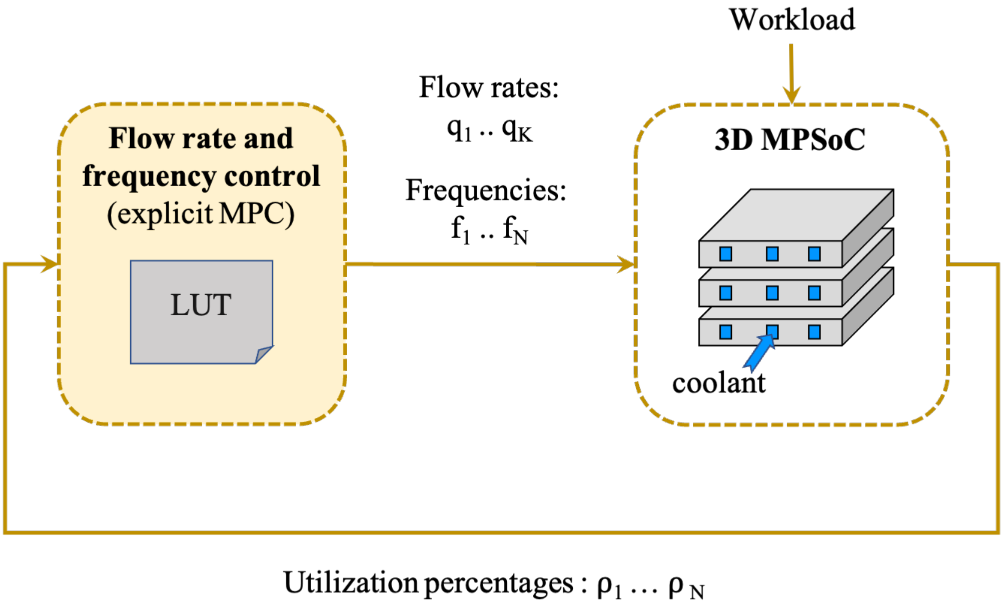


Figure 5.15: Implementation of online flow rate and frequency controller for target 3D MPSoC

A high-level view of the implemented frequency and FCA flow rate control strategy is shown in figure 5.15. The online controller uses the flow rates and frequency boost values computed by the offline solver in Section 5.5.1 to implement an explicit MPC [119]. Hence, the controller module periodically receives utilization data from task schedulers and then selects the corresponding FCAs and operating frequency settings according to the power requirements of each die. The controller then applies these settings to the 3D MPSoC during the subsequent time period from the pre-computed LUT.

5.5.3 Experimental results on target 3D MPSoC

5.5.3.1 Optimal flow rates and frequency boosts

The FCA flow rate and frequency optimization methodology from Section 5.5.1 is deployed to assess the frequency boost ratios applicable to the target 3D MPSoC computing dies under various utilization scenarios. As the memories have a negligible impact on the thermal performance of the system, it is pessimistically assumed that they are in full usage at all times. For comparison, a similar optimization algorithm is applied to the 3D stack when 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 5.16 and 5.17 show the optimal frequency boost ratios for both the CPU and the GPU, respectively. The proposed optimization methodology enables between 20% and 24% higher CPU operating frequency than the nominal value. In the case of a 100% CPU utilization, the frequency boost is equivalent to an additional dynamic power consumption of 45W. 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. In the case of a 100% GPU utilization, the frequency boost is equivalent to an additional dynamic power consumption of 43W. 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 5.18 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. It is considered that the FCA flow rate for each layer can be adjusted in steps of $20ml/min$.

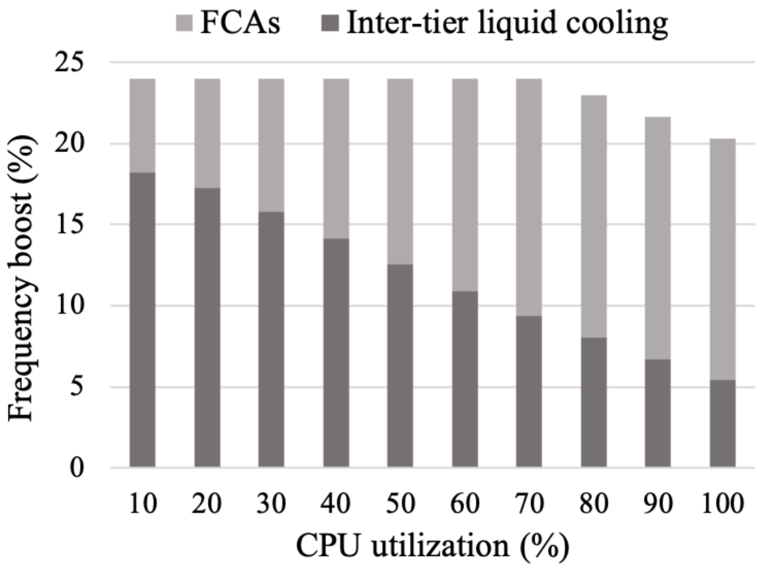


Figure 5.16: Applicable CPU frequency boosts

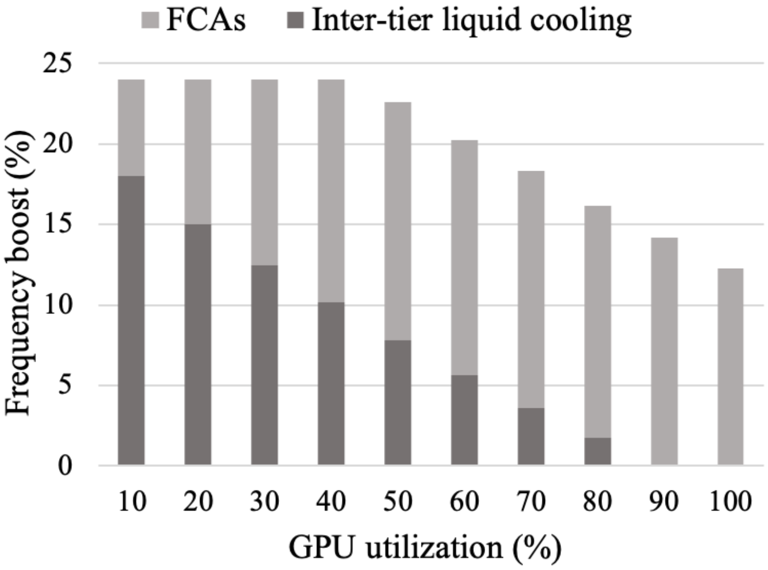


Figure 5.17: Applicable GPU frequency boosts

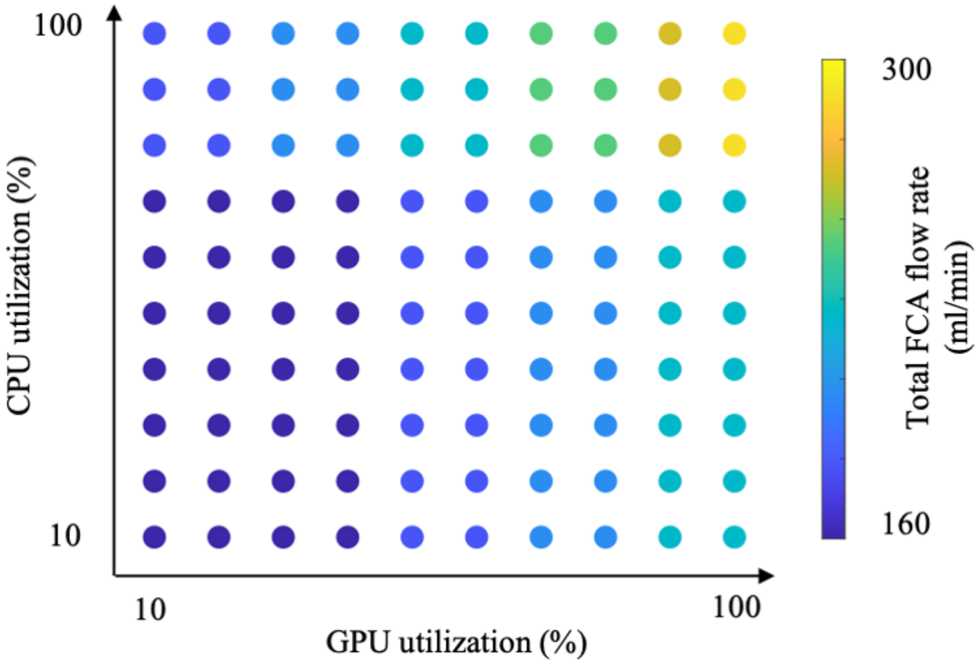


Figure 5.18: Optimal FCA flow rates corresponding to each CPU and GPU utilization scenario

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 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.

5.5.3.2 Workload speedups

The online frequency and FCA flow rate control strategy in Section 5.5.2 is simulated when running the benchmarks in Section 5.3.2 on the 3D MPSoC. Similarly to Section 5.4.3.2, workload execution is first recorded in the equivalent 2D system. Then, considering the same task schedule, their execution is simulated 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). The same memory bandwidth is assumed, and the frequency boost only applies to the computing dies (CPU and GPU).

Figure 5.19 represents the achieved workload speedups on the CPU using the proposed frequency and flow rate management strategy (*Optimal flow rate* in the figures). These results demonstrate a speedup of up to 17%. As the trend in the figure shows, *compute-intensive* benchmarks, such as Ca and Im present the highest overall speedup because the frequency boost does not affect memory access time.

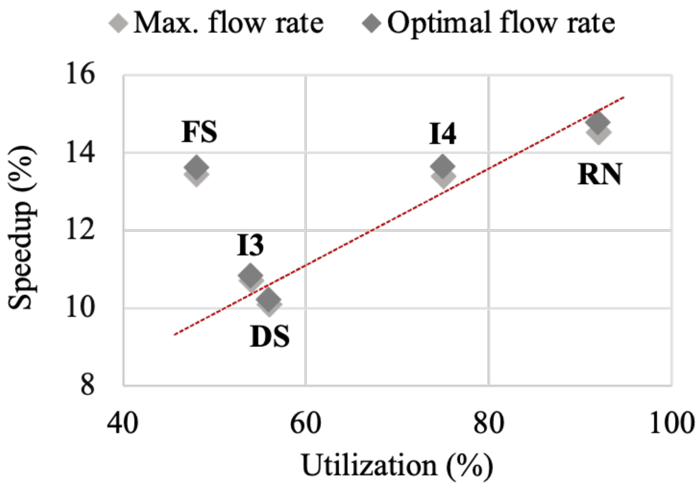


Figure 5.20: Workload speedups on GPU

Then, Figure 5.20 represents the workload speedups on the GPU, reaching up to 15%. The total speedup depends on the utilization level and dynamic power consumption in this case. 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. Hence, workloads with high utilization (e.g., RN) benefit from a high overall speedup as their percentage of computing versus memory access time is more significant. Conversely, the FS benchmark achieves a high overall speedup because of its low dynamic power consumption during utilization which, on average, enables the highest frequency boost.

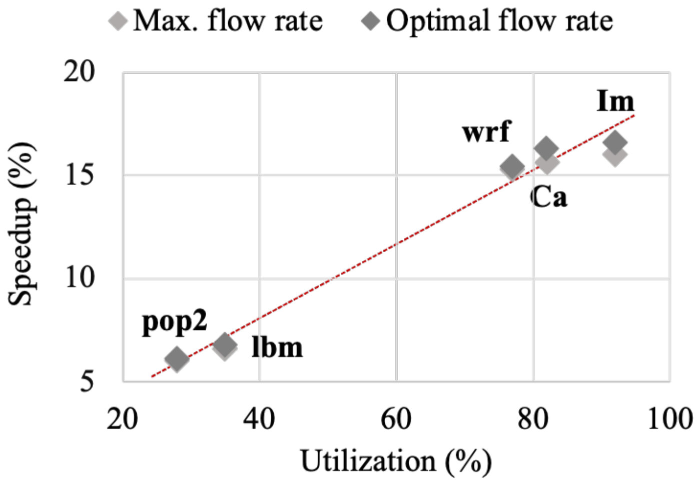


Figure 5.19: Workload speedups on CPU

		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 5.2: Cooling energy savings

5.5.3.3 Cooling energy savings

The flow rate optimization methodology targets achieving the highest workload speedups while reducing the energy consumption related to liquid pumping inside FCA channels. To measure the resulting energy savings, the achieved speedups on the CPU and the GPU are also recorded in the case of a fixed FCA flow rate, similarly to Section 5.5.3.2. In this case, the maximal flow rate value is applied during the entire workload execution to ensure abiding by temperature constraints at all times. Hence, only the computing dies frequencies are adjusted during run-time (the inner loop in Figure 5.14). The corresponding speedup results are labeled *Max. flow rate* in Figures 5.19 and 5.20. This strategy enables comparable workload speedups to the dynamic FCA flow rate case but uses significantly higher cooling resources. In particular, Table 5.2 shows the total cooling energy savings when using the dynamic flow rate management compared to the fixed flow rate case. These savings are calculated based on the total flow rates applied during workload execution, using the flow rate-liquid pumping power relationship in [107]. Hence, the proposed strategy achieves the optimal speedups while economizing up to 43% cooling energy (i.e., equivalent to 9W). In particular, the workload combinations that have the lowest utilization (e.g., FS and pop2) can be efficiently executed using a significantly lower total FCA flow rate than the maximal value, as shown in Figure 5.18.

5.6 Comparison with other DTM strategies

The execution of the workloads from Section 5.3.2 is emulated when applying a baseline DTM policy with DVFS and task migration [107] to the target 3D MPSoC using cold-plate-based cooling. Typically, thermal dissipation is more critical in the case of 3D MPSoCs compared to 2D MPSoCs. Hence, it is necessary to reduce the power consumption to abide by temperature constraints. In particular, DVFS gradually lowers the voltage and frequency settings down to 50% [122] (i.e., 92W lower dynamic power consumption in the case of the GPU). For workloads requiring a high utilization percentage (e.g., Ca and RN), some cores frequently reach their peak temperature. Thus, their assigned tasks are migrated to other colder cores, each time incurring a migration cost of 100ms [107]. Adopting this policy, a slow down of up to 25% and 40% is measured for the CPU and GPU, respectively. Instead, the proposed approaches in

Sections 5.4 and 5.5 leveraging FCAs and frequency control enable tangible run-time *gains* with respect to the nominal conditions, as discussed in the previous sections.

5.7 Conclusion

In this chapter, I have proposed to leverage the cooling and power supply capabilities of FCAs, thus boosting the power performance of multi-core processing dies in a 3D MPSoC while satisfying design constraints. Using fine-grained thermal and power modeling and analysis, a novel temperature and timing-aware MPC strategy was introduced to maximize the operating frequencies at run-time. This strategy was then extended by throttling the FCA flow rates to achieve the highest computation speed, using significantly less cooling energy. When applying these strategies to a target 3D CPU-GPU platform, the simulation results demonstrated execution speedups of up to 17% for a vast collection of high-performance benchmarks.

The proposed performance management strategies increase the system performance without any software or architecture optimization. They can be applied to 3D platforms containing multiple high-performance computing dies, regardless of their architecture and deployed CMOS technology. Hence, these results advocate for adopting FCA technology as an enabler of power-efficient 3D MPSoCs targeting modern high-performance applications.

6 Conclusion

6.1 Summary

My thesis used FCA technology as an effective solution to enable high-performance heterogeneous 3D MPSoCs. This technology can successfully dissipate the extremely high heat generated by stacked dies. Therefore, it reduces the temperature-dependent leakage, which represents a significant portion of power consumption in deeply-scaled ICs. In addition, FCAs provide on-chip electrochemical power generation, which compensates for the power losses in the delivery network and enhances the performance of dies.

From a system-level perspective, Chapter 3 addressed the comprehensive evaluation of 3D MPSoCs with FCAs. For this purpose, it presented a framework for thermal and power modeling and analysis of 3D MPSoCs by constructing a fine-grained representation of the full power network of dies, including FCA power generation variation with temperature. Then, this model is used to simulate the temperature and voltage distribution across the 3D MPSoC, highlighting the positive impact of FCA cooling and power supply on performance. Indeed, FCAs allow maintaining the temperature in high-performance computing dies under 52°C, and the voltage drops below the typical 5% constraint, but without requiring extra thermal/power TSVs or higher power delivery grid density. This framework is a valuable resource for the early design stages of 3D MPSoCs.

Next, Chapter 4 addressed the design-time optimization of 3D MPSoCs with FCAs, to further enhance power performance. Using high-efficiency and low-area SC converters as an interface between FCA electrodes and 3D MPSoC power delivery lines, on-chip power generation can increase by up to 123%. Indeed, SC converters enable to operate FCAs at their most efficient voltage regime regardless of the V_{dd} supply coming from the PCB. These converters also enable to extend the lifetime of the electrolytic fluid reservoir by switching off power generation in case of low power supply requirements. The chapter then explored the design space of FCA and converter-related parameters to evaluate the best achievable temperature and voltage drop reduction performance. The simulation results demonstrated the significant potential of FCAs in achieving power-efficient 3D MPSoCs with dies operating at full computing capacity.

Finally, Chapter 5 proposed to leverage the cooling and power supply capabilities of FCAs by increasing the operating frequencies of multi-core processing dies in a 3D MPSoC, compared to nominal operation in a 2D platform. Hence, the chapter introduced a temperature and timing-aware MPC strategy to maximize the frequencies at run-time while satisfying design constraints. In addition, throttling the FCA flow rates enables the highest computation speed using significantly less cooling energy. This is achieved by considering the existing trade-offs between FCA cooling (i.e., leakage reduction) and power supply. When applied to high-performance 3D computing platforms, my proposed run-time management strategy can result in up to 17% execution speedups of modern compute-intensive benchmarks.

6.2 Future work

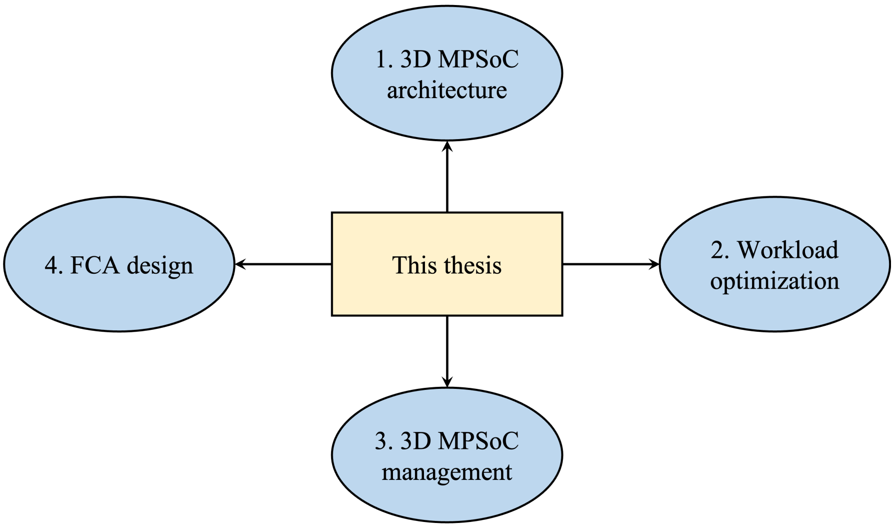


Figure 6.1: Potential future work directions

The work presented in my thesis gives the reader a global view of FCA integration's impact on the thermal, power, and computing performance of 3D MPSoCs. By tackling several aspects from power delivery network design to run-time resource management, my thesis has demonstrated the promising prospect of high-density-computing 3D systems where components can concurrently and safely operate at maximum capacity. The proposed modeling and analysis methodologies provide a general understanding of the important implications of FCA integration. They can serve as a guideline for experts in several fields related to 3D MPSoC design with FCAs, such as 3D architectures, workload optimization, system management, and FCA design (illustrated in Figure 6.1). Thus, it opens the door to further advancements in these areas.

6.2.1 3D MPSoC architecture design with FCAs

The high-density nature of 3D integrated circuits suggests separating high-power-consuming dies and/or components in 3D ICs to avoid hotspots affecting performance. However, this thesis demonstrated that FCAs can effectively transform the generated heat into power to speed up the operation of dies (Chapter 5). This observation implies that concentrating power consumption in specific parts can benefit 3D IC performance. Hence, this thesis causes a paradigm shift in 3D IC design, traditionally driven by thermal considerations. Leveraging FCA technology, designers can target 3D systems where high-power components (e.g., computing cores) are placed closer to each other (both vertically and horizontally). In such architectures, the increased power density causes higher liquid temperatures that improve FCA power generation. This situation was demonstrated in Chapter 3 when simulating the power delivery network of a target 3D MPSoC in the case of utilization scenarios where the power distribution contains high hotspot concentrations. Thus, FCAs enable to operate the cores at maximum speed without violating temperature and voltage constraints, as demonstrated in Chapter 5. In addition, core-to-core communication is faster and less power-consuming, improving performance further.

My thesis also provides clear arguments to rethink power delivery network design considering the presence of FCAs as a secondary power source. In particular, Chapter 3 discussed the impact of power TSV placement on the efficacy of FCA IR-drop reduction capabilities. Hence, TSV placement can be planned according to the floorplan by relaxing the PCB power supply in areas where FCA power generation eliminates power losses in the delivery lines. Furthermore, other power delivery network design aspects not addressed in this thesis (e.g., power grid density, sizing) can also affect FCA power distribution across dies. Hence, this motivates including an electrical model of FCAs in the next-generation physical design tools. Thus, these tools can consider the FCA power supply as an optimization target.

6.2.2 Workload optimization on 3D MPSoCs with FCAs

Workload scheduling and thread-to-core allocation can also take advantage of FCA cooling and power supply capabilities in 3D MPSoCs. For example, mapping intensive tasks to cores that are close to each other raises the temperature of FCAs and amplifies the electrochemical reactions inside the channels, as demonstrated in Chapters 3 and 5. Thanks to efficient FCA cooling, the temperature is guaranteed to remain below the constraints. Then, the improved power generation enables operating the cores at high frequencies. In addition, task scheduling must exploit the increased core-to-core and memory-to-core communication bandwidth in new architectures where fewer TSVs are dedicated to power delivery.

6.2.3 3D MPSoC thermal and power management with FCAs

My thesis demonstrated that FCA power generation is a key feature in determining the maximum achievable 3D IC performance. Hence, dynamically controlling the parameters that govern power generation ensures optimal usage of FCAs. Besides the inlet flow rate evaluated in Chapter 5, run-time management strategies can target the control of other parameters, such as the inlet liquid temperature or the voltage level between FCA electrodes. The former can be set for the entire 3D IC, considering the total cooling and power generation requirements. By using on-chip voltage regulators, the latter can be controlled individually for each flow cell (or group of flow cells). Indeed, the voltage regulation in Chapter 4 enables maximizing FCA power generation during chip operation and turning it off when the chip is idle. Alternatively, voltage regulation can regulate the power generation of FCAs depending on utilization (similarly to DVFS). In this case, electrolyte consumption can be optimal, avoiding excess use in case of lower FCA power requirements.

6.2.4 FCA design for high-performance 3D MPSoCs

Finally, my thesis provides the foundations for a further development of FCA technology in terms of the nature of reactants and the geometry of channels. These efforts must consider the practical impact of the different FCA parameters on the thermal and power performance of dies (e.g., size, voltage level, temperature, flow rate, etc.). They must also include fine-grained electro-thermal modeling and analysis of their capacities when integrated into the power network of 3D ICs.

Bibliography

[1] M. Motoyoshi. "Through-Silicon Via (TSV)". *Proceedings of the IEEE*, 97(1):43–48, 2009.

[2] P. Ramm et al. "3D Integration Technology: Status and Application Development". *European Conference on Solid-State Circuits (ESSCIRC)*, 2010.

[3] J. Fan and C.S.Tan. "Low Temperature Wafer-Level Metal Thermo-Compression Bonding Technology for 3D Integration". *Metallurgy – Advances in Materials and Processes*, pages 71–94, 2012.

[4] S. Borkar. "3D Integration for Energy Efficient System Design". *Design Automation Conference (DAC)*, 2011.

[5] J. Ahn, S. Yoo, and K. Choi. "Low-Power Hybrid Memory Cubes With Link Power Management and Two-Level Prefetching". *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 24(2):453–464, 2016.

[6] D. U. Lee et al. "A 1.2 V 8 Gb 8-Channel 128 GB/s High-Bandwidth Memory (HBM) Stacked DRAM With Effective I/O Test Circuits". *IEEE Journal of Solid-State Circuits*, 50(1):191–203, 2015.

[7] B. Black et al. "Die Stacking (3D) Microarchitecture". *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2006.

[8] D. H. Kim et al. "Design and Analysis of 3D-MAPS (3D Massively Parallel Processor with Stacked Memory)". *IEEE Transactions on Computers*, 64(1):112–125, 2015.

[9] Y.-H. Gong et al. "Exploration of temperature-aware refresh schemes for 3D stacked eDRAM caches". *Microprocessors and Microsystems*, 42:100–112, 2016.

[10] Q. Zhu et al. "A 3D-Stacked Logic-in-Memory Accelerator for Application-Specific Data Intensive Computing". *IEEE International 3D Systems Integration Conference (3DIC)*, 2013.

[11] P. Vivet et al. "A 4×4x2 Homogeneous Scalable 3D Network-on-Chip Circuit With 326 MFlit/s 0.66 pJ/b Robust and Fault Tolerant Asynchronous 3D Links". *IEEE Journal of Solid-State Circuits*, 52(1):33–49, 2017.

[12] P. Vivet et al. "A 220GOPS 96-Core Processor with 6 Chiplets 3D-Stacked on an Active Interposer Offering 0.6ns/mm Latency, $3Tb/s/mm^2$ Inter-Chiplet Interconnects and $156mW/mm^2$ @ 82%-Peak-Efficiency DC-DC Converters". *International Solid-State Circuits Conference*, 2020.

[13] D. B. Ingerly et al. "Foveros: 3D Integration and the use of Face-to-Face Chip Stacking for Logic Devices". *IEEE International Electron Devices Meeting*, 2019.

[14] E. Wong and S. K. Lim. "3D Floorplanning with Thermal Vias". *Design, Automation and Test in Europe (DATE)*, 2006.

[15] J. H. Lau and T. G. Yue. "Thermal Management of 3D IC Integration with TSV (Through Silicon Via)". *IEEE Electronic Components and Technology Conference*, 2009.

[16] R. Mathur et al. "Thermal Analysis of a 3D Stacked High-Performance Commercial Microprocessor using Face-to-Face Wafer Bonding Technology". *IEEE Electronic Components and Technology Conference (ECTC)*, 2020.

[17] A. Agrawal et al. "Thermal and Electrical Performance of Direct Bond Interconnect Technology for 2.5D and 3D Integrated Circuits". *IEEE Electronic Components and Technology Conference*, 2017.

[18] E. Bury et al. "Experimental Extraction of BEOL Composite Equivalent Thermal Conductivities for Application in Self-heating Simulations". *European Solid-State Device Research Conference (ESSDERC)*, 2018.

[19] P. Falkenstern et al. "Three-dimensional integrated circuits (3D IC) Floorplan and Power/Ground Network Co-synthesis". *Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2010.

[20] D.H. Kim and S. K. Lim. "Physical Design and CAD Tools for 3-D Integrated Circuits: Challenges and Opportunities". *IEEE Design and Test*, 32(4):8–22, 2015.

[21] H. Xiao et al. "VDPred: Predicting Voltage Droop for Power- Effient 3D Multi-core Processor Design". *International Conference on Computer and Automation Engineering*, 2021.

[22] G. Sisto et al. "IR-Drop Analysis of Hybrid Bonded 3D-ICs with Backside Power Delivery and μ- & n- TSVs". *IEEE International Interconnect Technology Conference (IITC)*, 2021.

[23] D. B. Tuckerman and R. F. W. Pease. "High-Performance Heat Sinking for VLSI". *IEEE Electron Devices Letters*, 2(5):126–129, 1981.

[24] F. Alfieri et al. "3D Integrated Water Cooling of a Composite Multilayer Stack of Chips". *Journal of Heat Transfer*, 132, 2010.

[25] 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.

[26] L. K. Hwang. "Accurate Models for Optimizing Tapered Microchannel Heat Sinks in 3D ICs". *IEEE Computer Society Annual Symposium on VLSI*, 2018.

[27] M. Sabry, D. Atienza, and A. Coskun. "Thermal analysis and active cooling management for 3D MPSoCs". *IEEE International Symposium of Circuits and Systems (ISCAS)*, 2011.

[28] M. Sabry et al. "Greencool: An Energy-Efficient Liquid Cooling Design Technique for 3-D MPSoCs via Channel Width Modulation". *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 32:524–537, 2013.

[29] W. Escher et al. "On the Cooling of Electronics With Nanofluids". *Journal of Heat Transfer*, 133, 2011.

[30] S. V. Garimella, V. Singhai, and D. Liu. "On-Chip Thermal Management with Microchannel Heat Sinks and Integrated Micropumps". *Proceedings of the IEEE*, 94(8):1534–1548, 2006.

[31] T. Acikalin and C. Schroeder. "Direct Liquid Cooling of Bare Die Packages using a Microchannel Cold Plate". *Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, 2014.

[32] D. Zhao and G. Tan. "A Review of Thermoelectric Cooling: Materials, Modeling and Applications". *Applied Thermal Engineering*, 66:15–24, 2014.

[33] A. Sridhar et al. "3D-ICE: a compact thermal model for early-stage design of liquid-cooled ICs". *IEEE Transactions on Computers*, 63(10):2576–2589, 2014.

[34] W. Huang et al. "HotSpot: A Compact Thermal Modeling Methodology for Early-Stage VLSI Design". *IEEE Transactions on VLSI*, 14(5):501–513, 2006.

[35] "Ansys CFX". Retrieved from: http://www.ansys.com/products/fluid-dynamics/cfx/.

[36] F. Terraneo et al. "3D-ICE 3.0: Efficient Nonlinear MPSoC Thermal Simulation With Pluggable Heat Sink Models". *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 41(4):1062–1075, 2022.

[37] M. M. Sabry et al. "Integrated Microfluidic Power Generation and Cooling for Bright Silicon MPSoCs". *Design, Automation & Test in Europe Conference (DATE)*, 2014.

[38] E. Kjeang et al. "Microfluidic fuel cells: A review". *Journal of Power Sources*, 186(2):353–369, 2009.

[39] R. ferrigno et al. "Membraneless Vanadium Redox Fuel Cell Using Laminar Flow". *Journal of the American Chemical Society*, 124:12930–12931, 2002.

[40] H. Zhang. "Polysulfide-Bromine Flow Batteries (PBBs) for Medium- and Large-Scale Energy Storage". *Advances in Batteries for Medium and Large-Scale Energy Storage*, pages 317–327, 2015.

[41] M. O.Bamgbopaa et al. "Cyclable Membraneless Redox Flow Batteries Based on Immiscible Liquid Electrolytes: Demonstration with All-Iron Redox Chemistry". *Electrochimica Acta*, 267:41–50, 2018.

[42] J. H. Vinco et al. "Unfolding the Vanadium Redox Flow Batteries: An Indeep Perspective on its Components and Current Operation Challenges". *Journal of Energy Storage*, 43, 2021.

[43] A.A. Andreev et al. "PowerCool: Simulation of Cooling and Powering of 3D MPSoCs with Integrated Flow Cell Arrays". *IEEE Transactions on Computers*, 67(1):73–85, 2018.

[44] A. Sridhar et al. "PowerCool: Simulation of integrated microfluidic power generation in bright silicon MPSoCs". *International Conference on Computer-Aided Design*, 2014.

[45] "COMSOL". Multiphysics simulation infrastructure. Retrieved from http://www.comsol.com/.

[46] "IEEE 2016 International Roadmap for Devices and Systems". https://irds.ieee.org/images/files/pdf/2016_MM.pdf.

[47] "IEEE 2018 International Roadmap for Devices and Systems". https://irds.ieee.org/images/files/pdf/2018/2018IRDS_MM.pdf.

[48] R. Brain. "Interconnect scaling: Challenges and opportunities". *IEEE International Electron Devices Meeting*, 2016.

[49] D.H. Woo et al. "An Optimized 3D-Stacked Memory Architecture by Exploiting Excessive, High-Density TSV Bandwidth". *International Symposium on High-Performance Computer Architecture*, 2010.

[50] P. Zhou et al. "3D-STAF: scalable temperature and leakage aware floorplanning for three-dimensional integrated circuits". *International Conference On Computer Aided Design*, 2007.

[51] S. Wang et al. "P/G TSV Planning for IR-drop Reduction in 3D-ICs". *Design, Automation and Test in Europe (DATE)*, 2014.

[52] C. Auth et al. "A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors". *Symposium on VLSI Technology*, 2012.

[53] C. Auth et al. "A 14nm logic technology featuring 2nd-generation FinFET, air-gapped interconnects, self-aligned double patterning and a 0.0588 µm2SRAM cell size". *IEEE International Electron Devices Meeting (IEDM)*, 2014.

[54] C. Auth et al. "A 10nm high performance and low-power CMOS technology featuring 3rd generation FinFET transistors, Self-Aligned Quad Patterning, contact over active gate and cobalt local interconnects". *Annual IEEE International Electron Devices Meeting*, 2017.

[55] R. Xie et al. "A 7nm FinFET technology featuring EUV patterning and dual strained high mobility channels". *Annual IEEE International Electron Devices Meeting*, 2016.

[56] A. Bartolini et al. "Unveiling Eurora: Thermal and Power Characterization of the most Energy-Efficient Supercomputer in the World". *Design, Automation & Test of Europe (DATE)*, 2014.

[57] D. Kulkarni et al. "Experimental Study of Two-Phase Cooling to Enable Large-Scale System Computing Performance". *IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, 2018.

[58] M. Daneshtalab et al. "Memory-Efficient Logic Layer Communication Platform for 3D-Stacked Memory-on-Processor Architectures". *International 3D Systems Integration Conference*, 2011.

[59] J. H. Lau and T. G. Yue. "Thermal Management of 3D IC Integration with TSV". *IEEE Electronic Components and Technology Conference*, 2009.

[60] I. Partin-Vaisband et al. "On-Chip Power Distribution Networks". *On-Chip Power Delivery and Management*, pages 129–144, 2016.

[61] S. R. Nassif and J. N. Kozhaya. "Fast Power Grid Simulation". *Design Automation Conference (DAC)*, 2000.

[62] A. Zou et al. "Efficient and Reliable Power Delivery in Voltage-Stacked Manycore System with Hybrid Charge-Recycling Regulators". *Design Automation Conference (DAC)*, 2018.

[63] 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 (ISLPED)*, 2019.

[64] A. Andreev et al. "Design Optimization of 3D Multi-Processor System-on-Chip with Integrated Flow Cell Arrays". *International Symposium of Low Power electronics and Design*, 2018.

[65] E. Fluhr et al. "POWER8: A 12-Core Server-Class Processor in 22nm SOI with 7.6Tb/s Off-Chip Bandwidth". *International Solid-State Circuits Conference*, 2014.

[66] N. Jouppi et al. "In-datacenter performance analysis of a tensor processing unit". *International Symposium on Computer Architecture*, 2017.

[67] W. J. Starke et al. "The cache and memory subsystems of the IBM POWER8 processor". *IBM Journal of Research and Development*, 59(1):3:1–3:13, 2015.

[68] T. Tang et al. "MLPAT: A power area timing modeling framework for machine learning accelerators". *International Workshop on Domain Specific System Architecture*, 2018.

[69] S. Wu. "A highly manufacturable 28nm CMOS low power platform technology with fully functional 64Mb SRAM using dual/tripe gate oxide process". *Symposium on VLSI Technology*, 2009.

[70] D. Ingerly et al. "Low-k Interconnect Stack with Metal-Insulator-Metal Capacitors for 22nm High Volume Manufacturing". *Internationa Interconnect Technology Conference*, 2012.

[71] K. Fischer et al. "Low-k Interconnect Stack with multi-layer Air Gap and Tri-Metal-Insulator-Metal Capacitors for 14nm High Volume Manufacturing". *International Interconnect Technology Conference*, 2015.

[72] S. Sinha et al. "Exploring sub-20nm FinFET design with Predictive Technology Models". *Design Automation Conference (DAC)*, 2012.

[73] D. Cuesta et al. "Adaptive Task Migration Policies for Thermal control in MPSoCs". *International Symposium on VLSI*, 2010.

[74] A. Iranfar et al. "Machine Learning-Based Quality-Aware Power and Thermal Management of Multistream HEVC Encoding on Multicore Servers". *IEEE Transactions on Parallel and Distributed Systems*, 29(10):2268–2281, 2018.

[75] D. Christen et al. "Energy Efficient Heat Sink Design: Natural Versus Forced Convection Cooling". *IEEE Transactions on Power Electronics*, 32(11):8693–8704, 2017.

[76] Y. Ban et al. "IR-drop analysis for validating power grids and standard cell architectures in sub-10nm node designs". *Design-Process-Technology Co-optimization for Manufacturability*, 2017.

[77] F. Clermidy et al. "3D Embedded Multi-Core: Some Perspectives. *Design Automation and Test in Europe Conference (DATE)*, 2011.

[78] P. Emma and E. Kursun. "Opportunities and Challenges for 3D Systems and Their Design". *IEEE Design & Test of Computers*, 26(05):6–14, 2009.

[79] 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.

[80] P. Sivakumar et al. "Optimization of thermal aware multilevel routing for 3D IC". *Analog Integrated Circuits and Signal Processing*, 103:131–142, 2020.

[81] A. Paul et al. "Deep Trench Capacitor Based Step-Up and Step-Down DC/DC Converters in 32nm SOI with Opportunistic Current Borrowing and Fast DVFS Capabilities". *IEEE Asian Solid-State Circuits Conference*, 2013.

[82] S. Banzhaf et al. "Post-Trench Processing of Silicon Deep Trench Capacitors for Power Electronic Applications". *International Symposium on Power Semiconductor Devices and ICs*, 2016.

[83] J. L. Ayala et al. "Through Silicon Via-Based Grid for Thermal Control in 3D Chips". *International Conference on Nano-Networks*, 2009.

[84] D. Brenner, C. Merkel, and D. Kudithipudi. "Design-Time Performance Evaluation of Thermal Management Policies for SRAM and RRAM based 3D MPSoCs". *Great Lakes Symposium on VLSI*, 2012.

[85] P. Vivet et al. "IntAct: A 96-Core Processor With Si Chiplets 3D-Stacked on an Active Interposer With Distributed Interconnects and Integrated Power Management". *IEEE Journal of Solid-State Circuits*, 56(1):79–97, 2021.

[86] E. A. Burton et al. "FIVR – Fully Integrated Voltage Regulators on 4th Generation Intel Core SoCs". *Proceedings of the IEEE Applied Power Electronics Conference and Exposition*, 2014.

[87] T. V. Breussegem and M. Steyaert. "A 82% efficiency 0.5% ripple 16-phase fully integrated capacitive voltage doubler". *IEEE Symposium on VLSI Circuits*, 2009.

[88] D. Somasekhar et al. "Multiphase 1 GHz voltage doubler charge-pump in 32 nm logic process". *IEEE Journal of Solid State Circuit*, 45:751 – 758, 2010.

[89] T. M. Andersen. *"On-Chip Switched Capacitor Voltage Regulators for Granular Microprocessor Power Delivery"*. PhD thesis, ETH Zürich (Ph.D. Thesis), 2015.

[90] P. R. Morrow et al. "Design and fabrication of on-chip coupled inductors integrated with magnetic material for voltage regulators". *IEEE Transactions on Magnetics*, 47(6):1678–1686, 2011.

[91] 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 (ISVLSI)*, 2020.

[92] L. Muller and J. W. Kimball. "A Dynamic Model of Switched-Capacitor Power Converters". *IEEE Transactions on Power Electronics*, 29:1862–1869, 2014.

[93] J. M. Henry and J. W. Kimball. "Practical Performance Analysis of Complex Switched-Capacitor Converters". *IEEE Transactions on Power Electronics*, 26(1):127–136, 2011.

[94] T. M. Andersen et al. "Modeling and Pareto Optimization of On-Chip Switched Capacitor Converters". *IEEE Transactions on Power Electronics*, 32(1):363–377, 2017.

[95] T. M. Andersen et al. "20.3 A feedforward controlled on-chip switched-capacitor voltage regulator delivering 10W in 32nm SOI CMOS". *International Solid-State Circuits Conference*, 2015.

[96] A. S. Hollinger et al. "Nanoporous separator and low fuel concentration to minimize crossover in direct methanol laminar flow fuel cells". *Journal of Power Sources*, 195:3523–3528, 2010.

[97] S. Pal et al. "Design Space Exploration for Chiplet-Assembly-Based Processors". *IEEE Transactions on VLSI Systems*, 28(4):1062–1073, 2020.

[98] Y. XIE et al. "Design Space Exploration for 3D Architectures". *ACM Journal on Emerging Technologies in Computing Systems*, 2:65–103, 2006.

[99] "NVIDIA Tesla V100 Servers". Retrieved from https://www.thinkmate.com/systems/servers/gpx/v100.

[100] K. Lepak et al. "The Next Generation AMD Enterprise Server Product Architecture". *Hot Chips*, 2017.

[101] 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.

[102] "NVIDIA TESLA V100 GPU Architecture", 2017. Retrieved from http://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf.

[103] "Zen Microarchitecture", 2017. Retrieved from https://en.wikichip.org/wiki/amd/ microarchitectures/zen.

[104] V. Chaturvedi et al. "Thermal-Aware Task Scheduling for Peak Temperature Minimization under Periodic Constraint for 3D-MPSoCs". *IEEE International Symposium on Rapid System Prototyping*, 2014.

[105] M. J. Sepulveda et al. "3DMIA: A Multi-Objective Artificial Immune Algorithm for 3D-MPSoC Multi-Application 3D-NoC Mapping". *Annual Conference Companion on Genetic and Evolutionary Computation*, 2013.

[106] A. Aggarwal et al. "Temperature Constrained Power Management Scheme for 3D MP-SoC". *IEEE Workshop on Signal and Power Integrity*, 2012.

[107] M. M. Sabry et al. "Energy-Efficient Multi-objective Thermal Control for Liquid-Cooled 3D Stacked Architectures". *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 30:1883–1896, 2011.

[108] A. K. Coskun et al. "Modeling and Dynamic Management of 3D Multicore Systems with Liquid Cooling". *International Conference on Very Large Scale Integration*, 2009.

[109] 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*, 46:33–43, 2013.

[110] "WILO MHIE Centrifugal Pump [Online]". Retrieved from http://www.wilo.com/cps/rde/xchg/en/layout.xsl/3707.html.

[111] "Festo Electric Automation Technology [Online]". Retrieved from http://www.festodidactic.com/ov3/media/customers/1100/0096636000107522-3683.pdf.

[112] C. Szegedy et al. "Rethinking the Inception Architecture for Computer Vision". *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[113] C. Szegedy et al. "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning". *AAAI Conference on Artificial Intelligence*, 2017.

[114] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[115] D. Amodei et al. "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin". *International Conference on Machine Learning*, 2016.

[116] M. Ott et al. "fairseq: A Fast, Extensible Toolkit for Sequence Modeling". *Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 48–53, 2019.

[117] J. Bucek, K.-D Lange, and J.-V. Kristowski. "SPEC CPU2017 – Next-Generation Compute Benchmark". *ACM/SPEC International Conference on Performance Engineering*, 2018.

[118] S. Che et al. "Rodinia: A Benchmark Suite for Heterogeneous Computing". *IEEE International Symposium on Workload Characterization*, 2009.

[119] F. Zanini et al. "Online Thermal Control Methods for Multiprocessor Systems". *Transactions on Design Automation of Electronic Systems*, 18:6:1–6:26, 2013.

[120] J. Guerreiro et al. "Modeling and Decoupling the GPU Power Consumption for Cross-Domain DVFS". *IEEE Transactions on Parallel and Distributed Systems*, 30(11):2494–2506, 2019.

[121] P. Bogdan, R. Marculescu, and S. Jain. "Dynamic Power Management for Multi-domain System-on-Chip Platforms: An Optimal Control Approach". *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 18:1–20, 2013.

[122] J. Murray et al. "Sustainable DVFS-Enabled Multi-Core Architectures with On-Chip Wireless Links". *Advances in Computers*, 88:125–158, 2013.

List of Figures

1.1 3D integration of ICs using TSV and microbump technology 1

1.2 Bonding technologies for 3D integration [3] 2

2.1 Microchannel-based inter-tier liquid cooling 7

2.2 3D-ICE model for a small section of a microchannel layer [33] 8

2.3 Planar view of a microfluidic channel with electrochemical flow and electrode contacts 9

2.4 Architecture of multi-core processor with microfluidic power supply and cooling, as proposed in [37] 10

2.5 PowerCool compact model for microfluidic flow cell, proposed in [43] 12

3.1 A single flow cell connected to the load and P/G grid of a 3D MPSoC die, as presented in [63] 17

3.2 Thermal-aware modeling and analysis framework for 3D power delivery networks 18

3.3 Current-voltage relationship (I-V curve) for a single 200μm-long flow cell 20

3.4 Electrical model of a 3D power delivery network 21

3.5 Target 3D MPSoC architecture 23

3.6 IBM POWER8 processor layout [67] 24

3.7 Google TPU floorplan [66] 24

3.8 HBM DRAM architecture [6] 25

3.9 TSVs and FCA placement 26

3.10 Power map of the HBM memory (uniform) 27

3.11 Power maps of the POWER8 processor (non-uniform) 27

3.12 Power maps of the TPU accelerator (non-uniform) 28

3.13 3D MPSoC stack configurations 29

3.14 Qualitative assessment of 3D MPSoC stack configurations 30

3.15 Temperature maps of the multi-core processor (configuration B2) 32

3.16 IR-drop reduction maps for the multi-core processor (*non-uniform power map*) and memories (*uniform power map*) in case of workload (4) 34

3.17 Average and standard deviation of IR-drop reduction for powermaps (1), (2), (3) and (4) 36

3.18 IR-drop variation with respect to the baseline (TSVs at the center) 38

3.19 FCA-enabled IR-drop reduction with respect to baseline (TSVs at the center) . 38

3.20 FCA-generated power with 40°C inlet temperature and 2.5m/s inlet speed, compared to the leakage power of the covered chip area 39

3.21 Normalized leakage per transistor, with respect to the value at 25°C, for different CMOS technology nodes 42

3.23 Temperature of the POWER8 and TPU-based dies with various cooling methods 47

3.24 IR-drop of the POWER8 and TPU-based dies with various cooling methods . . . 48

4.1 3D MPSoC with integrated FCAs 57

4.2 FCA-generated power (cell width: 50μm, height: 100μm, length: 100μm) 58

4.3 FCA cell connected to 3D MPSoC power delivery grid via DC-DC voltage regulator, as proposed in [91] 60

4.4 1:2 SC converter circuit 61

4.5 1:2 SC equivalent converter model 62

4.6 Evaluated DC-DC converter design points 65

4.7 3D MPSoC with FCAs: a processing layer and a memory layer 66

4.8 Measured POWER8 power maps 67

4.10 IR-drop and converter area requirements at maximal utilization ($V_{dd} = 1.1V$) . 70

4.11 Fuel reservoir lifetime with and without switching off FCA power generation for different processor duty-cycle values, normalized to the lifetime in case of 100% processor duty-cycle 71

4.12 3D MPSoC design-time performance characterization 73

4.13 Target 3D MPSoC 74

4.14 CPU layout [103] 75

4.15 GPU floorplan 75

4.16 Explored 3D MPSoC configurations 76

4.17 FCA and SC converter output power 78

4.18 CPU and GPU IR-drop maps 80

4.19 CPU temperature with FCAs and cold-plate cooling 81

4.20 CPU power and temperature 82

4.21 GPU power and temperature 82

4.22 CPU voltage drop with FCAs and SC converters 84

4.23 GPU voltage drop with FCAs and SC converters 84

5.1 Flow cell array technology 90

5.2 Target 3D MPSoC, including different types of memories, processing units, and the embedded FCA channels 92

5.3 Workload utilization statistics on the CPU 95

5.4 Workload utilization statistics on the GPU 95

5.5 MPC explicit solver with (1) an initialization phase, followed by (2) a frequency optimization loop 98

5.6 Transistor leakage vs. temperature (normalized), extracted from SPICE simulations of a standard gate using the 14nm and 12nm FFN technology nodes [72] 100

5.7 Critical path delay of canary circuit 103

5.8 Implementation of online frequency controller for 3D MPSoC computing dies based on explicit MPC 103

5.9 Evaluation of workload speedup at runtime 106
5.10 Average frequency boost results on CPU 107
5.11 Workload speedup results on CPU 108
5.12 Average frequency boost results on GPU 108
5.13 Workload speedup results on GPU 109
5.14 Offline frequency and flow rate optimization methodology 111
5.15 Implementation of online flow rate and frequency controller for target 3D MPSoC 112
5.16 Applicable CPU frequency boosts 114
5.17 Applicable GPU frequency boosts 114
5.18 Optimal FCA flow rates corresponding to each CPU and GPU utilization scenario 115
5.20 Workload speedups on GPU 116
5.19 Workload speedups on CPU 116

6.1 Potential future work directions 121

List of Tables

3.1 3D MPSoC dimensions 26

3.2 Peak temperature of the multi-core processor in the case of different configurations and usage scenarios 33

3.3 Peak temperature of the memory in the case of different configurations and usage scenarios 33

3.4 FCA power generation and IR-drop reduction for the multi-core processor die . 35

3.5 FCA power generation and IR-drop reduction for the memory die 35

3.6 Technology scaling parameters 40

3.7 Equivalent heat transfer coefficients for the MCP and TPU for different CMOS technologies and cooling techniques 44

5.1 Frequency optimization results (reduced LUT) 104

5.2 Cooling energy savings 117

Halima Najibi

PhD in Electrical and Electronics Engineering

November 19th, 1991
☎ +41 76 410 97 09
@ lima.najibi@gmail.com

Programming

Hardware:
Verilog, SystemVerilog, VHDL

Software:
C/C++, Java, Python, Matlab, bash, Perl, TCL

Languages

English	●●●●●
French	●●●●●
Arabic	●●●●●
Spanish	●●●○○

Education

October 2017 - Present
PhD in Electrical Engineering — Lausanne, Switzerland
Embedded Systems Laboratory (ESL)
Swiss Federal Institute of Technology (EPFL)
Modeling and thermal/power management of 3D multi-processor system-on-chip (MPSoC) architectures

September 2013 - July 2015
Master of Science — Lausanne, Switzerland
Electrical and Electronics Engineering in EPFL
Concentration in micro-nano electronics

September 2011 - July 2013
Bachelor of Science — Lausanne, Switzerland
Electrical and Electronics Engineering in EPFL
Concentration in micro-nano electronics and information technology

September 2009 - July 2011
CPGE — Casablanca, Morocco
Intensive training for enrollment in highly selective engineering schools in francophone regions
Concentration in mathematics and physics

Work Experiences

September 2017 - January 2022
Teaching Assistant — Lausanne, Switzerland
ESL, EPFL
Lab on App development for tablets and smartphones

July 2017 - September 2017
Research Internship — Lausanne, Switzerland
ESL, EPFL
Micro-fluidic cooling and power generation in three-dimensional (3D) server systems-on-chip (SoCs)

September 2016 - January 2017
Hardware Developer — Guadalajara, Mexico
Oracle Inc.
RTL design and verification

November 2015 - July 2016
Hardware Developer — Austin, Texas, USA
Oracle Inc.
RTL design and verification

August 2015 - September 2015
Research Internship — Lausanne, Switzerland
Microelectronic Systems Laboratory, EPFL
Large angle of view real-time depth estimation hardware implementation and testing

August 2014 - January 2015
Research Internship — Austin, Texas, USA
Oracle Inc.
RTL development, automation of design and constraints verification

July - August 2013
Research Internship — Lausanne, Switzerland
Adaptive Micro-nano Wave Systems Group, EPFL
Design of an efficient passive RFID system using MIMO transmission techniques

Halima Najibi

PhD in Electrical and Electronics Engineering

Fields of interest

SoC/MPSoC design, ASIC design, FPGA design.

VLSI, Integrated circuits (IC) thermal and power management, IC performance management, 3D Integration.

Hardware security

Machine learning, signal/image processing, real-time systems.

Hobbies

- Travel
- Sport
- Visual arts
- Social events

Publications

2022	**Conference paper** (1st *author*) Thermal and Power-Aware Run-Time Performance Management of 3D MPSoCs with Integrated Flow Cell Arrays *Great Lakes Symposium on VLSI (GLSVLSI)* *(3rd place Best Paper Award recipient)*
2022	**Journal article** (1st *author*) Thermal and Voltage-Aware Performance Management of 3D MPSoCs with Flow Cell Arrays and Integrated SC Converters *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*
2021	**Journal article** (1st *author*) Enabling Optimal On-Chip Power Generation of Flow Cell Arrays in 3D MPSoCs with On-Chip Switched Capacitor Converters *IEEE VLSI Circuits and Systems Letter, IEEE Computer Society Technical Committee on VLSI (TCVLSI)* *(Invited paper)*
2020	**Conference paper** (1st *author*) Towards Deeply Scaled 3D MPSoCs with Integrated Flow Cell Array Technology *Great Lakes Symposium on VLSI (GLSVLSI)*
2020	**Conference paper** (1st *author*) Enabling Optimal Power Generation of Flow Cell Arrays in 3D MPSoCs with On-Chip Switched Capacitor Converters *IEEE Computer Society Annual Symposium on VLSI (ISVLSI)* *(Best Paper Award recipient)*
2019	**Conference paper** (1st *author*) A Design Framework for Thermal-Aware Power Delivery Network in 3D MPSoCs with Integrated Flow Cell Arrays *IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)* *(Best Paper Award nominee)*
2015	**Book chapter** Enhanced Compressed Look-up-table based Real-Time Rectification Hardware *VLSI-SoC: At the Crossroads of Emerging Trends*
2013	**Journal article** User Effects in Beam-Space MIMO *IEEE Antennas and Wireless Propagation Letters*