Thermal and Voltage-Aware Performance Management of 3D MPSoCs with Flow Cell Arrays and Integrated SC Converters

Halima Najibi*, Alexandre Levisse*, Giovanni Ansaloni*, Marina Zapater†, Miroslav Vasic‡, and David Atienza*
*Embedded Systems Laboratory (ESL), École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
†REDS Institute, University of Applied Sciences Western Switzerland (HEIG-VD, HES-SO), Switzerland
‡Centro de Electrónica Industrial, Universidad Politécnica de Madrid (UPM), Spain

***Abstract*— Flow cell arrays (FCAs) concurrently provide efficient on-chip liquid cooling and electrochemical power generation. This technology is especially promising for three-dimensional multi-processor systems-on-chip (3D MPSoCs) realized in deeply scaled technologies, which present very challenging power and thermal requirements. Indeed, FCAs effectively improve power delivery network (PDN) performance, particularly if switched capacitor (SC) converters are employed to decouple the flow cells and the systems-on-chip voltages, allowing each to operate at their optimal point. Nonetheless, the design of FCA-based solutions entails non-obvious considerations and trade-offs, stemming from their dual role in governing both the thermal and power delivery characteristics of 3D MPSoCs. Showcasing them in this paper, we explore multiple FCA design configurations and demonstrate that this technology can decrease the temperature of a heterogeneous 3D MPSoC by 78°C, and its total power consumption by 46%, compared to a high-performance cold-plate based liquid cooling solution. At the same time, FCAs enable up to 90% voltage drop recovery across dies, using SC converters occupying a small fraction of the chip area. Such outcomes provide an opportunity to boost 3D MPSoC computing performance by increasing the operating frequency of dies. Leveraging these results, we introduce a novel temperature and voltage-aware model predictive control (MPC) strategy that optimizes power efficiency during run-time. We achieve application-wide speedups of up to 16% on various machine learning (ML), data mining, and other high-performance benchmarks while keeping the 3D MPSoC temperature below 83°C and voltage drops below 5%.**

***Index Terms*—3D MPSoC Management, Flow Cell Arrays, On-Chip Liquid Cooling, On-Chip Power Generation, Online Frequency Optimization, Model Predictive Control.**

I. Introduction

State of the art artificial intelligence (AI) and Big Data applications demand high performance, spurring a renewed interest in complex heterogeneous platforms combining diverse memory and computing elements (e.g., CPUs, GPUs). 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-ship (3D MPSoCs) enable high-density computing and provide ultra-wide communication bandwidth [1]. However, *3D stacking exacerbates heat dissipation* challenges. Indeed, 3D MPSoC temperatures are difficult to control using traditional cooling techniques, given the low thermal conductivity of bonding materials [2]. In addition, *3D integration complicates power delivery* due to the resistive losses in through-silicon-vias (TSVs) and metal wires. Moreover, the large amount of power TSVs distributing voltage supplies complicates routing [3], thus making 3D MPSoC physical design more difficult.

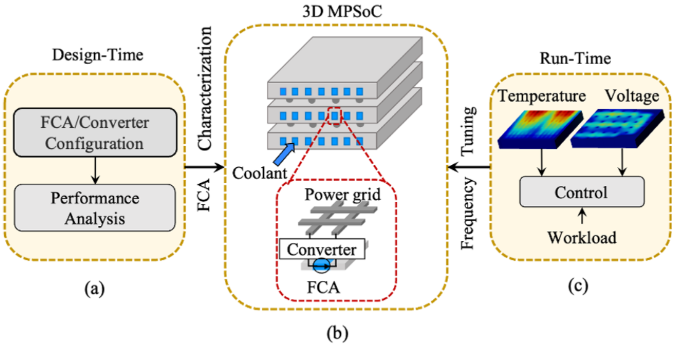


Fig. 1: Design-Time and Run-Time 3D MPSoC Management

Flow cell array (FCA) technology, introduced in [4], addresses the aforementioned 3D MPSoC challenges. FCAs consist of micro-fluidic channels that are etched in the silicon substrate of dies (Figure 1-b). 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), their generated current partially supplies logic gates. Hence, they help reduce voltage supply drops, preventing timing violations and system performance degradation [5]. FCA-generated power depends on the voltage between flow cell electrodes, which may differ from the voltage supply level required by the logic and memory dies in a 3D MPSoC. To bridge this gap, the authors of [6] use switched capacitor (SC) voltage converters as an interface between flow cells and 3D power grids (Figure 1-b). SC converters allow FCAs to operate in their most efficient voltage regime and decouple them from PDN disturbances.

Integrated cooling and power delivery solutions based on FCAs and SC converters are promising avenues to disruptively increase the performance of 3D MPSoCs. Nonetheless, FCA technology also exposes a novel and multi-faceted design space, encompassing inter-dependent thermal and electrical considerations. In this paper, we investigate it from two complementary viewpoints. From a *design-time* perspective (Figure 1-a), we illustrate a methodology to explore different configurations of FCAs (varying channel densities and coolant velocities) and their associated SC converters and evaluate their thermal and power performance. This analysis serves to characterize these configurations and highlight the existing design trade-offs. It also showcases the opportunities for 3D MPSoC performance improvement that are enabled by FCA integration. Hence, we explore the demonstrated leeway from a *run-time* perspective (Figure 1-c) by introducing a novel strategy for dynamic performance management of 3D MPSoCs, based on model predictive control (MPC). The proposed MPC solver uses the previous performance analysis methodology to calculate the optimal operating frequency boost for 3D MPSoC components while remaining within safe temperature and voltage margins.

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

- Targeting a high-performance, 4-layer 3D MPSoC system, we illustrate a power and thermal design-time exploration of multiple 3D MPSoC configurations with integrated FCAs and SC converters. We use fine-grain modeling to measure their thermal and power performance, and discuss entailed trade-offs.
- We show that for such 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.
- We introduce a novel thermal and voltage-aware MPC strategy to optimize the operation frequency of processing cores during run-time, by exploiting the additional FCA power without compromising their timing and temperature.
- We demonstrate that our MPC approach enables up to 25% faster clock frequencies when optimizing the execution of data-intensive and compute-intensive benchmarks. It can speed-up workloads by 16% on the central processing unit (CPU) for a utilization rate of 82%, and by 13% on the graphics processing unit (GPU) for an average of 92% utilization percentage.

The rest of the paper proceeds as follows. Section II summarizes the state of the art of 3D MPSoC management, as well as related works in FCA technology. Section III presents an overview of the target 3D MPSoC used as experimental vehicle. Section IV discusses the 3D MPSoC design-time performance analysis and trade-offs. Section V presents the novel 3D MPSoC thermal and voltage-aware MPC optimization strategy. Finally, Section VI shows the achieved speed-up of real high-performance benchmarks.

II. Background on 3D MPSoC Thermal and Power Management

A. 3D MPSoC Design Challenges and Thermal/Power Management Strategies

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 [1]. However, TSV-based 3D integration presents critical thermal and power management challenges, limiting its viability in modern high-performance 3D MPSoCs.

1) Thermal Management: Power density increases with the number of stacked dies, generating large amounts of heat, which is very difficult to dissipate due to the low thermal conductivity of silicon and bonding materials [2]. This issue is exacerbated in modern CMOS technologies by high transistor densities and leakage currents. In this regard, several design-time solutions address the heat extraction problem in 3D ICs. Authors in [7] 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, [8] 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, authors in [9] 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. As fan-based cooling struggles to maintain 3D MPSoC temperatures at acceptable levels, a high-performance direct liquid cooling solution using a cold plate has been proposed [10]. 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 [11]. 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, which absorbs the generated heat [12]. As opposed to FCAs, the coolants employed in this scenario are inert, and no electrical power is generated.

2) Power Management: High leakage and power density greatly complicate power delivery in 3D MPSoCs [5]. 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 [3]. Addressing these 3D MPSoC power-related issues, authors in [13] use an active interposer, which reduces the power density of large-scale heterogeneous chiplet-based systems using on-chip power management and energy-efficient 3D plugs for communication. However, this technique does not exploit the high bandwidth capabilities of TSV-based 3D integration and presents challenges related to long-distance communication. In contrast, [14] 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, [15] proposes a routing algorithm to minimize the power dissipation and wire delays of TSV-based 3D ICs. However, the techniques in [14] and [15] deploy a large number of power TSVs, at the expense of inter-tier communication, thus creating a trade-off between power delivery and communication bandwidth.

3) Flow Cell Array Technology with Integrated Voltage Regulators: FCA technology is a novel solution to both power and thermal challenges of 3D MPSoCs, providing combined on-chip liquid cooling and electrochemical power generation [5]. FCAs use a technology similar to inter-tier liquid cooling [12]. However, the micro-channels used in this case are filled

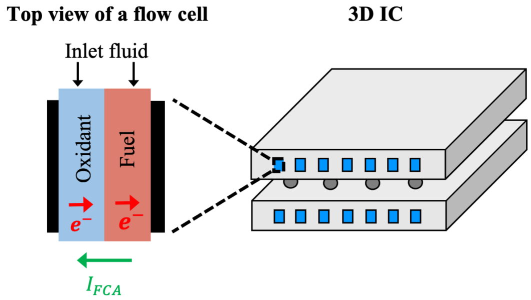


Fig. 2: FCA Technology

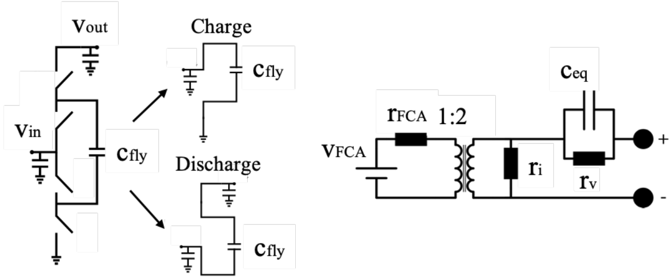


Fig. 3: SC converter circuit (*left*) and circuit model (*right*) [19]

with an electrolytic liquid flow that produces an electrical current to supply logic gates, as illustrated in Figure 2. High channel temperatures increase the electrochemical reaction rate, effectively transforming heat into available power for high-performance 3D MPSoCs. As shown in [5] and [11], FCA-generated current can recover up to 20% voltage 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 a given voltage drop constraint.

Although directly connecting FCAs to 3D MPSoC PDNs shows substantial improvements in power efficiency, their power generation capabilities are sub-optimal when operating at the V_{dd} level of high-performance systems, typically over 0.7V. Peak power generation for vanadium-based redox flows (used in this work) is achieved around $0.6V$ [6]. Therefore, voltage regulation must be employed to ensure full exploitation of their power generation potential. In this context, authors in [6] use on-chip voltage converters implemented employing a switched capacitor (SC) topology. These devices use the electrical field in a capacitor as the main medium for energy conversion [16], as illustrated in Figure 3, allowing FCAs to operate at their most efficient regime. Additionally, they decouple the FCAs from logic circuits in case of transient load changes, and the PDN from voltage fluctuations at the electrode contacts. SC converters achieve over 80% voltage conversion efficiency, they are easy to integrate, and occupy a low area [17]. These characteristics make them ideal components for interfacing FCAs to 3D PDNs. The SC converter topology proposed in [6], as well as its equivalent circuit model, are presented in Figure 3. The model is employed to calculate the SC converter performance in different operating conditions. The authors highlight that the optimal design point for a given output voltage may not exhibit the best performance when slightly varying the operating conditions. Hence, they propose an algorithm to explore the design parameters (e.g., transistor and capacitor sizes) and evaluate the resulting performance (e.g., area, conversion efficiency).

FCAs paired with SC converters enable efficient cooling and additional power for high-performance 3D MPSoCs, but their cooling and power supply capabilities contrast each other [18]. In this regard, Section IV proposes a methodology to configure FCAs with their associated SC converters, to achieve the desired performance of a target system.

B. Run-Time Thermal and Power Management of 3D MPSoCs

The previous design-time cooling and power management techniques must ensure that temperature and voltage constraints are met under worst-case conditions. However, operating conditions are generally application-dependent. Hence, continually adopting worst-case assumptions can lead to under-utilizing computational components (e.g., overly reducing their frequency to limit heat generation) or over-utilizing cooling and power resources.

To overcome this pitfall, several run-time thermal and power management techniques have been proposed. For instance, the authors of [20] use a thermal-aware mapping algorithm and perform workload migration between hot and cool layers during run-time, based on temperature information of the stack. The authors in [21] also propose an adaptive algorithm for multi-application 3D-NoC mapping to reduce latency and total system power under temperature constraints. Furthermore, [22] introduces a temperature-constrained power management scheme for 3D-MPSoCs, accounting for the activity of processing elements, their positions, and temperature margins.

Few run-time management strategies specifically target 3D MPSoCs with inter-tier liquid cooling. The authors in [23] 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 in [24] 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 [25], 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 a variable-flow liquid cooling to enable system power reduction and performance loss minimization.

The previous 3D MPSoC run-time management policies only deal with thermal and power regulation. To the best of our knowledge, no existing policies exploit both cooling and power generation capabilities of FCAs, or analyze the performance boosts applicable in this scenario. Our work aims to fill this gap, proposing a novel run-time thermal and power management policy, which we illustrate in Section V.

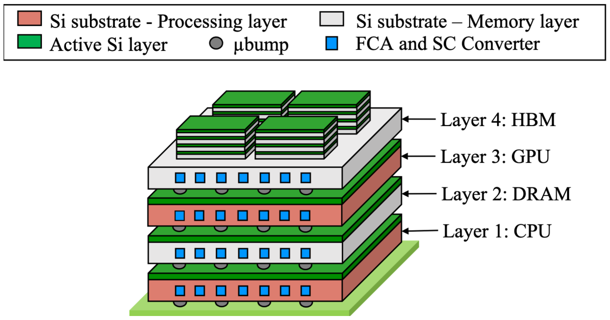


Fig. 4: Target 3D MPSoC

III. Target 3D MPSoC with integrated FCAs and SC converters

To exemplify the efficiency of our proposed design-time exploration and run-time management strategy for 3D MPSoCs with integrated FCAs and SC converters, we employ as a target system a high-performance four-layer stack, shown in Figure 4. We base the architecture on a state-of-the-art CPU-GPU platform for high-performance computing [26]. We consider its implementation in 3D, 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 [27], with a total area of $757mm^2$. Figure 5 presents the EPYC processor layout. 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 maximal total power consumption is $180W$, and its maximal supported temperature is 81°C.
- The second layer contains an 8-channel DDR4-2666W [28], 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 [29], 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 6 graphics processing clusters (GPCs) with 14 streaming multiprocessors (SMs), as illustrated in Figure 6. 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 $15W$ [30].

Our 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.

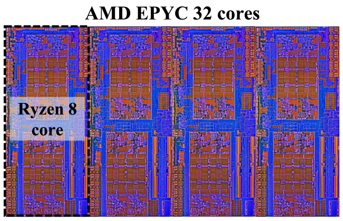


Fig. 5: CPU Layout

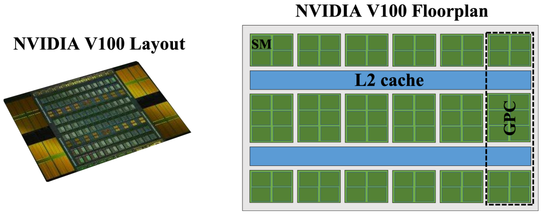


Fig. 6: GPU Layout and Floorplan

Similarly to [6], 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 IV and V, we model this 3D MPSoC in fine-grain to evaluate both its thermal and electrical performances. In particular, we use 3D-ICE [31] to evaluate its thermal behavior under different load scenarios. Then, we use HSPICE to measure its PDN performance, as described in [5]. To do so, we include a compact FCA model and a converter circuit model (Figure 3) to perform electrical simulations. Both the flow cells and SC converters are modeled in Verilog-A. Hence, we evaluate the FCA power generation and SC converter efficiency, allowing us to retrieve the voltage and temperature distributions of dies. As in [6], we use cell dimensions of $200 \times 100\mu m^2$ and $50 \times 50\mu m^2$ for the thermal and electrical simulations, respectively.

We characterize the workloads employed to evaluate our run-time management strategy (described in Section V) by running benchmarks in a real system containing the same components as the above target 3D MPSoC, and incorporating a traditional fan-based cooling system. We use performance counters to measure benchmark usage statistics. These statistics serve to guide the experimental evaluation in Section VI.

IV. Design-Time Temperature/Voltage Analysis and Characterization of 3D MPSoCs with FCAs and SC converters

In this section, we present a design-time 3D MPSoC characterization through thermal/power performance analysis, illustrated in Figure 7. The proposed flow explores multiple FCA configurations targeting the 3D MPSoC described in Section III, and performs fine-grain analysis considering target application requirements and 3D MPSoC design constraints.

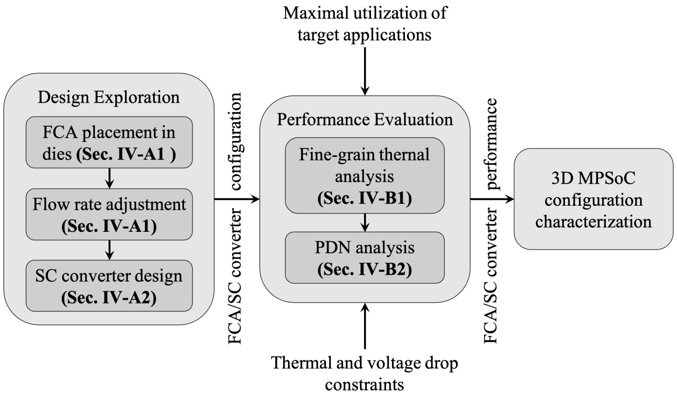

Maximal utilization of target applications
Design Exploration
FCA placement in dies (Sec. IV-A1)
Flow rate adjustment (Sec. IV-A1)
SC converter design (Sec. IV-A2)
FCA/SC converter configuration
Performance Evaluation
Fine-grain thermal analysis (Sec. IV-B1)
PDN analysis (Sec. IV-B2)
FCA/SC converter performance
3D MPSoC configuration characterization
Thermal and voltage drop constraints

Fig. 7: 3D MPSoC Design-Time Performance Characterization

The *design configuration* considers FCA-related parameters, namely the FCA placement in the 3D MPSoC layers and electrolytic liquid flow speed (Section IV-A1), and the SC converter design (Section IV-A2). 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 [32] [33]. However, our simulations indicate that micro-channels 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 [5]. Thus, we here adopt this solution without further exploring this aspect.

Hence, the *performance evaluation* in Section IV-B uses fine-grain thermal and electrical simulations (described in Section III) 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.

A. FCA and SC Converter Configuration

1) FCA Placement and Flow Rate: FCA-based cooling and power generation interact in a non-obvious way. In fact, higher cooling efficiency decreases temperature, limiting the total chip leakage (hence power consumption) but also lowering the power generated by FCAs. Conversely, lower cooling increases the electrolyte reaction rate, allowing to generate more electrical power. To investigate these trade-offs and identify the configuration that performs best under specific voltage/temperature constraints, detailed thermal and power analyses are needed. As candidate solutions, we consider the configurations shown in Figure 8 (A1 to B4). These configurations are selected as follow:

- 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 in configuration B, two FCAs supply it. This is achieved by electrically connecting the computing dies using TSVs to the FCAs etched in the dies themselves and via TSVs and micro-bumps to the FCAs etched in the above memory dies. As the memories consume considerably less power, we do not supply them with FCA power.

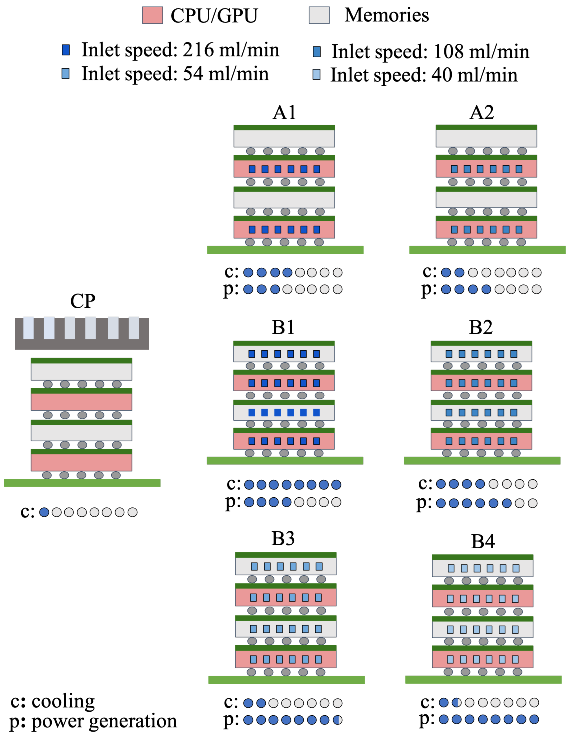


Fig. 8: 3D MPSoC Configurations

- In terms of FCA flow rate, we consider the full-flow rate value (216ml/min, as in [18]) for both cases (A1 and B1). Then, we consider the lowest flow rate that complies with temperature constraints (B4).
- Then, to highlight the existing trade-offs between FCA cooling and power generation, we consider other configurations with similar cooling performances but different numbers of FCAs (for example, A1 and B2, A2 and B3).

For comparison, we also characterize a state-of-the-art cold-plate based liquid cooling for ultra-high-performance MPSoCs [10] (configuration CP). The performances, in terms of on-chip cooling and power generation, of each configuration are qualitatively represented in Figure 8. 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 efficiency as the coolant heats the most compared to other configurations, accelerating the electrochemical reactions inside the channels.

We quantitatively assess these intuitions in Section IV-B, where we present the outcome of fine-grain thermal and electrical simulations, according to the methodology outlined

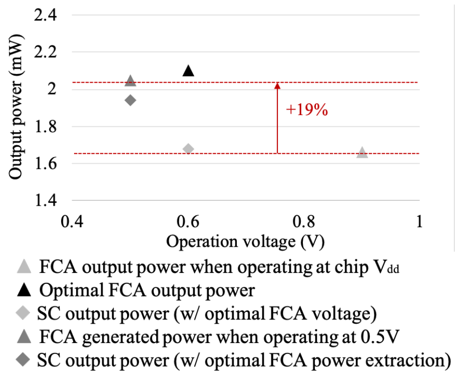


Fig. 9: FCA and SC Converter Output Power

in Section III.

2) SC Converter Design: A trade-off exists between SC conversion efficiency, area, and output power [6]. 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 [6]) 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 [6], the optimal design point achieves a relatively low voltage conversion rate. Indeed, as illustrated in Figure 9, 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 the maximal converter 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 [6], 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 $0.00071mm^2$). Therefore, the optimal SC converter design in this scenario is selected for the remainder of this work.

To quantify the system-wide benefits of this design, Figure 10 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, as it represents extreme operating conditions. First, we show the voltage drop 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, we show the voltage drop map 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 10 presents 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.

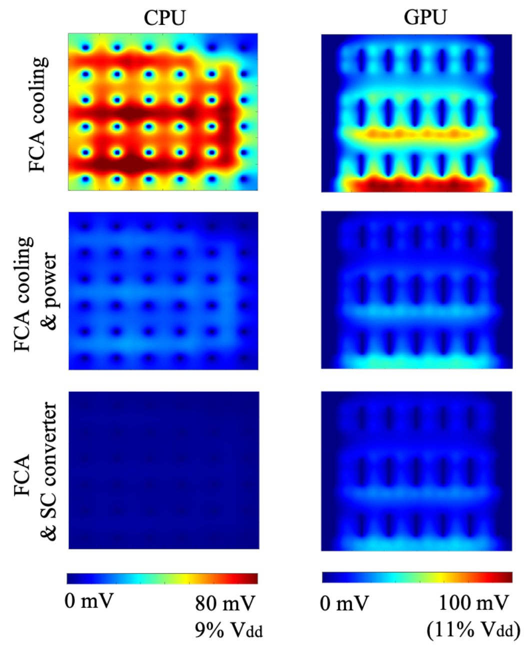


Fig. 10: CPU and GPU IR-drop maps

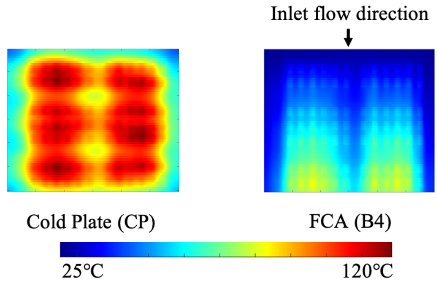


Fig. 11: CPU Temperature with FCAs and Cold-Plate Cooling

B. Thermal and Power Performance Evaluation of 3DMPSoC with FCAs and SC Converters

We herein evaluate in detail the 3D MPSoC thermal and power performances in the different configurations described in IV-A1, assuming the use of the SC converter identified in Section IV-A2. We compare them to the cold plate-based liquid cooling strategy (configuration CP in Figure 8). Across experiments, we consider a maximal usage scenario for both the CPU and GPU, representing worst-case operating conditions. We report dynamic and leakage power figures, 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 III). Leakage maps when dies are cooled using FCAs or CP cooling are calculated based on computed temperature maps (the details of leakage map estimation are found in Section V-B).

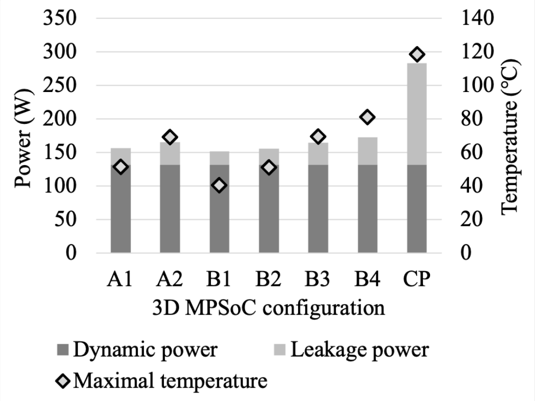


Fig. 12: CPU Power and Temperature

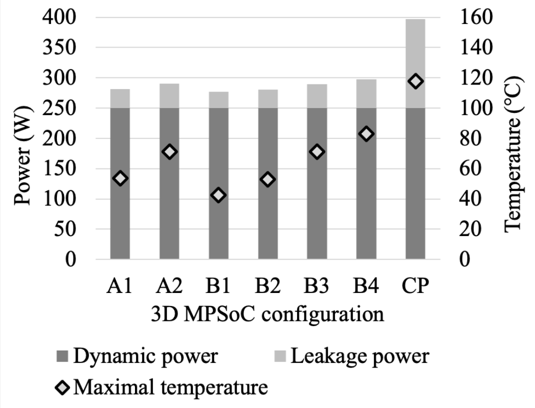


Fig. 13: GPU Power and Temperature

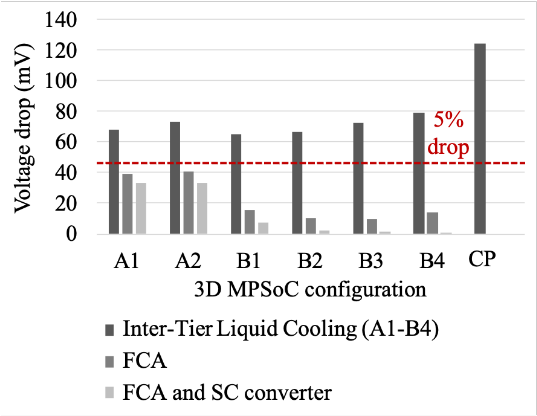


Fig. 14: CPU Voltage Drop with FCAs and SC Converters

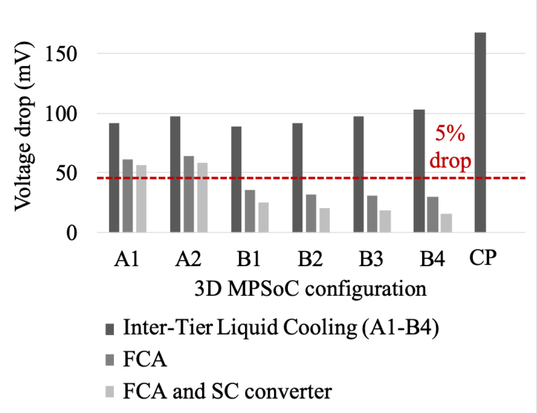


Fig. 15: GPU Voltage Drop with FCAs and SC Converters

1) Temperature and Total Power Consumption: Figure 11 shows the CPU temperature map when cooled using the CP solution [10], compared to the one in configuration B4, which has the lowest FCA cooling capacity. The figure showcases that FCAs vastly outperform cold plate-based liquid cooling. Additionally, we measure the total power consumption of the CPU and GPU dies at maximal usage and present the results in Figure 12 and Figure 13, respectively. The figures indicate that temperature-dependent leakage is a significant contributor to power budgets in the CP case. FCA cooling can effectively reduce leakage power by up to 86% compared to the CP strategy in the CPU case and up to 82% in the GPU case. The peak temperature difference between the CP and B4 configurations are 78°C and 75°C for the CPU and GPU dies, respectively. Additionally, 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 among all configurations, as detailed in the following.

2) Voltage Drop Recovery: Maximal voltage drop values for the CPU and GPU dies are presented in Figure 14 and Figure 15 (respectively). We include all the considered FCA configurations and the CP one. In configurations A1 to B4, the voltage drop is measured in three scenarios: when FCAs are only used for their cooling capabilities (inter-tier liquid cooling), when they are directly connected to the power grids, and when SC converters are used. 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 (Figure 14 and Figure 15). A similar observation is done between configuration 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. We conclude that the use of FCAs and SC converters improves 3D MPSoC power performance with respect to cold plate cooling in all cases. In particular, FCAs significantly decrease PDN losses in configuration B4, which has the slowest liquid traversal in the channels, and therefore highest on-chip power generation.

The FCA's 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 work focuses on the second alternative. Indeed, Section V describes a run-time 3D MPSoC performance optimization methodology, using the described temperature and voltage analysis framework. The optimization solver computes the applicable frequency boosts without violating voltage and temperature constraints. It is then evaluated on various state-of-the-art high-performance benchmarks in Section VI.

V. Run-Time Performance Optimization of 3D MPSoCs with FCAs and SC converters

As outlined in Section IV, the integration of FCAs and SC converters in 3D MPSoC PDNs provides opportunities to increase the load of dies without violating temperature and voltage drop constraints. To harness them, we introduce an online approach to enhance the performance of 3D MPSoC computing dies based on specific workload requirements. In particular, we design a model predictive control (MPC) algorithm to boost the operation frequency of the different CPU and GPU cores during run-time, whose block scheme is illustrated in Figure 16. 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 [34]. 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 MPC implementation and frequency optimization algorithm are presented in Sections V-A and V-B, respectively.

A. Explicit MPC Implementation

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. Unfortunately, this implies a large set of inputs in an explicit MPC implementation and an impractical characterization effort. However, as outlined in IV, the heat absorption capabilities of FCAs greatly limit heat exchanges among dies, especially given the distance between the two most power-consuming dies in our target 3D MPSoC (i.e., the CPU and GPU).

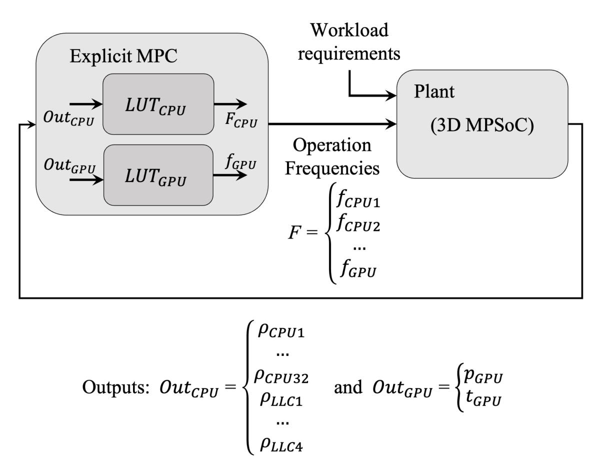

Explicit MPC
LUT_{CPU}
LUT_{GPU}
Out_{CPU}
Out_{GPU}
F_{CPU}
f_{GPU}
Workload requirements
Plant
(3D MPSoC)
Operation Frequencies
$F = \begin{cases} f_{CPU1} \\ f_{CPU2} \\ \dots \\ f_{GPU} \end{cases}$
Outputs: $Out_{CPU} = \begin{cases} \rho_{CPU1} \\ \dots \\ \rho_{CPU32} \\ \rho_{LLC1} \\ \dots \\ \rho_{LLC4} \end{cases}$ and $Out_{GPU} = \begin{cases} p_{GPU} \\ t_{GPU} \end{cases}$

Fig. 16: Explicit MPC Implementation

This observation allows performing the frequency optimization of the GPU and the CPU independently by the MPC explicit solver. To further decrease the number of simulation points, we pessimistically assume that memories are in full utilization at all times, as their power consumption has a negligible impact on the 3D MPSoC thermal performance.

A high-level view of the implemented explicit MPC is shown in figure 16. The MPC module periodically receives utilization data from the CPU and GPU performance counters. Mainly, it takes as inputs the utilization percentage of CPU cores, the utilization percentage of CPU last-level caches (LLCs), the measured total GPU power, and the GPU temperature from embedded sensors. This data is used to estimate the available temperature and timing leeway and, completing the optimization loop, accordingly set the clock frequencies for the CPU cores and the GPU.

B. Temperature and Voltage-Aware 3D MPSoC Frequency Optimization Algorithm

To fill the LUTs of the explicit MPC implementation we introduce a frequency optimization algorithm performed offline to determine the applicable frequency boost under different 3D MPSoC utilization scenarios. First, the algorithm receives as input the power being generated in each die (for the experiments in Section VI-B, power values are derived from performance counters, as detailed in Section III). Then, it evaluates the frequency increase that can be applied in the modeled 3D MPSoC for different degrees of utilization (hence, generated power). Next, the algorithm calculates temperatures for different utilization levels given the geometry of the 3D MPSoC, the FCA topology, and coolant flow. Moreover, it also accounts for the effect of voltage drops in the CPU and GPU timing characteristics, again depending on utilization and FCA and SC converters characteristics. It then dictates 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.

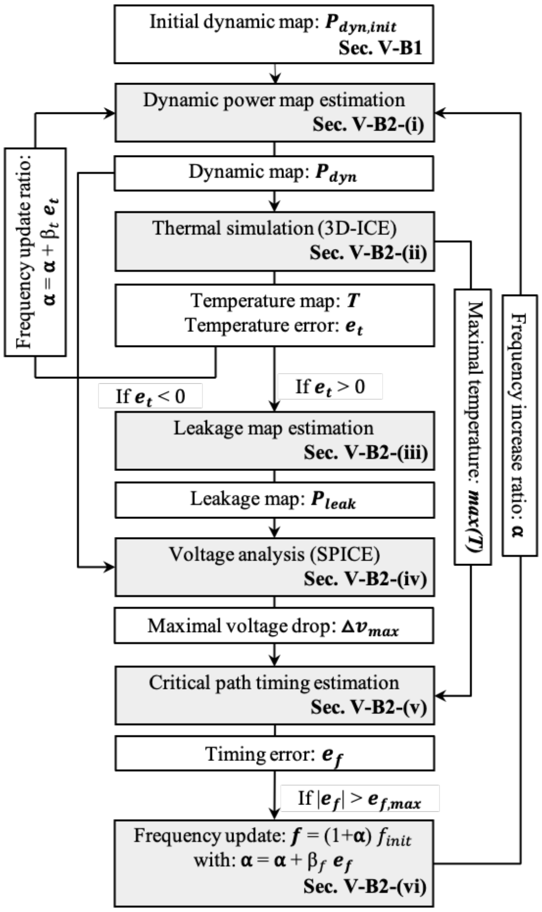

Initial dynamic map: $\boldsymbol{P_{dyn,init}}$
Sec. V-B1
Dynamic power map estimation
Sec. V-B2-(i)
Dynamic map: $\boldsymbol{P_{dyn}}$
Thermal simulation (3D-ICE)
Sec. V-B2-(ii)
Temperature map: $\boldsymbol{T}$
Temperature error: $\boldsymbol{e_t}$
Frequency update ratio:
$\boldsymbol{\alpha} = \boldsymbol{\alpha} + \beta_t\, \boldsymbol{e_t}$
If $\boldsymbol{e_t} < 0$
If $\boldsymbol{e_t} > 0$
Leakage map estimation
Sec. V-B2-(iii)
Leakage map: $\boldsymbol{P_{leak}}$
Voltage analysis (SPICE)
Sec. V-B2-(iv)
Maximal voltage drop: $\boldsymbol{\Delta v_{max}}$
Critical path timing estimation
Sec. V-B2-(v)
Timing error: $\boldsymbol{e_f}$
If $|\boldsymbol{e_f}| > \boldsymbol{e_{f,max}}$
Frequency update: $\boldsymbol{f} = (1+\boldsymbol{\alpha})\, f_{init}$
with: $\boldsymbol{\alpha} = \boldsymbol{\alpha} + \beta_f\, \boldsymbol{e_f}$
Sec. V-B2-(vi)
Maximal temperature: $\boldsymbol{max(T)}$
Frequency increase ratio: $\boldsymbol{\alpha}$

Fig. 17: Frequency Optimization Loop (MPC Explicit Solver)

In all the equations throughout these sections, the vectors and matrices are denoted by capital letters, and the scalar values are indicated by lower case letters.

1) Initialization: As a starting point, the algorithm estimates the initial dynamic power maps $P_{dyn,init}$ of the 3D MPSoC dies, based on the different utilization metrics extracted from the performance counters. As the CPU and GPU include performance counters measuring different metrics, $P_{dyn,init}$ is calculated in a different manner for the two dies, as well as their respective memories. In case of the CPU, the number of executed instructions directly reflects on the utilization level of a core ρ_{core}. Similarly, the number of LLC and DDR accesses reflect on their utilization level ρ_{cache} and ρ_{DDR}. Hence, The power consumption of the core ($p_{dyn}(core)$), LLC ($p_{dyn}(LLC)$), and DRAM ($p_{dyn}(DDR)$) are calculated as follows:

$$p_{dyn}(core) = \rho_{core} * p_{max,core} \tag{1}$$

$$p_{dyn}(LLC) = \rho_{LLC} * p_{max,LLC} \tag{2}$$

$$p_{dyn}(DDR) = \rho_{DDR} * p_{max,DDR} \tag{3}$$

The cores and LLC dynamic power consumption values are then mapped to the CPU layout to construct the dynamic power map (Figure 5). Then, the DDR power consumption is mapped to its area, assuming a uniform data access pattern.

In case of the GPU, the dynamic power value $p_{dyn,init}(GPU)$ is extracted from the initial total power $p_{total,init}(GPU)$ by subtracting the total 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 [35], at the reference temperature of 25°C. Then, this value increases exponentially with temperature [11]. Hence, the total leakage power of the GPU die for a temperature t_{GPU} is calculated as follows:

$$p_{leak,init}(GPU) = i_{off}(t_{GPU,init}) * w_{eff} * \rho_{trans} * A \tag{4}$$

$$p_{dyn,init}(GPU) = p_{total,init}(GPU) - p_{leak,init}(GPU) \tag{5}$$

The initial dynamic power of the GPU is then mapped to the floorplan in Figure 6, according to the power consumption percentage of the different components [36], to construct the initial dynamic power map $P_{dyn,init}(GPU)$:

$$P_{dyn,init}(GPU) = p_{dyn,init}(GPU)/p_{max,GPU} * P_{max,GPU} \tag{6}$$

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:

$$P_{dyn}(HBM) = \rho_{HBM} * P_{max,HBM} \tag{7}$$

2) Optimization Loop (Explicit Solver): After estimating the initial dynamic power maps of the 3D MPSoC dies, the following series of steps is performed recursively until convergence, as shown in Figure 17. In this figure, all the variables that are estimated during the MPC solver flow are in **bold**. The optimization loop searches for the maximal applicable frequency increase ratios, which achieve the desired timing and temperature of the computing dies. The detailed steps of the optimization loop are presented in the following, using the order and numbering in Figure 17, and performed similarly for all 3D MPSoC computing dies:

(i) **Dynamic power map estimation**:
In this first step, the dynamic power map of each one of the computing dies (CPU and GPU) is scaled according to its frequency increase ratio α. To do so, we use a quadratic frequency-power relationship, generally applicable for many-core ICs [37]:

$$P_{dyn} \sim f^2 \tag{8}$$

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

$$P_{dyn} = (1 + \alpha)^2 P_{dyn,init} \tag{9}$$

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

(ii) **Thermal simulation**:
In this step, we evaluate the temperature of each die when the power consumption profiles of the 3D MPSoC computing dies correspond to the previously calculated power maps. Furthermore, we pessimistically consider a maximal power consumption of the memory dies, as they

are not bottlenecks for the thermal behaviour of the stack. Hence, we use 3D-ICE, a compact thermal simulator for liquid-cooled 3D ICs [31]. 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 target 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 $\epsilon_t = max(T) - T_{max}$:

$$\alpha = \alpha + \beta_t * \epsilon_t \tag{10}$$

The algorithm then iterates back to step (i).
If no temperature violation occurs, the algorithm proceeds to the next step (iii).

(iii) **Leakage map estimation**:
Next, we determine the leakage map P_{leak} of each 3D MPSoC die according to its thermal map T. Similarly to Equation 4, the leakage value ${P_{leak}}^{i,j}$ of a cell with coordinates (i,j) is estimated as:

$${P_{leak}}^{i,j} = i_{off}(T^{i,j}) * w_{eff} * \rho_{trans} * A_{cell} \tag{11}$$

The total 3D MPSoC die power map is then calculated for a given frequency increase α as:

$$P_{total} = P_{dyn} + P_{leak} \tag{12}$$

In fact, we have not included the idle power 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.

(iv) **Voltage analysis**:
After computing the total power maps of the dies. We build the fine-grain 3D MPSoC electrical model [5] with the FCAs and SC converters, as described in Section III. We simulate our model using HSPICE to obtain the voltage map of the target dies. For each die, we then extract the critical voltage drop value Δv_{max}.

(v) **Critical path timing estimation**:
In this step, we estimate the timing of the most critical path of the target 3D MPSoC die, with respect to its clock period (or frequency). This step indicates if timing violations can potentially occur and compromise the chip operation. Hence, we devise the critical path depending on voltage drop and thermal conditions. We characterize this relationship based on a canary circuit (a 64-bit full adder, implemented in a 28nm CMOS technology). Results of this exploration are shown in Figure 18. In this figure, v_{dd} is normalized to its value for the technology library, and 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 [10], in an equivalent 2D system. This value represents the nominal operation frequency, assuming the signal integrity of the circuit in this scenario. Therefore, we extract for each of our target 3D MPSoC die the critical path timing τ_{max}. We pessimistically assume that it corresponds to the power and thermal hotspot of the die (highest voltage drop Δv_{max} and temperature $max(T)$), representing worst-case operation conditions. We then compare the critical path timing τ_{max} to the current clock period $((1+\alpha)f)^{-1}$ and compute the frequency (or timing) error e_f:

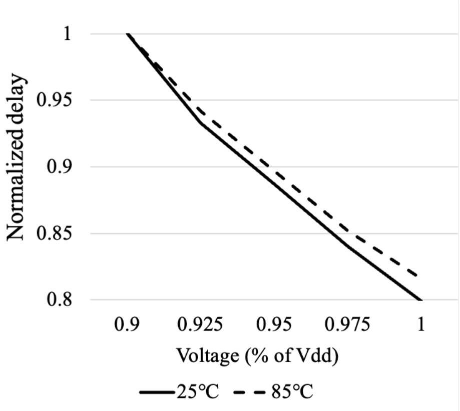


Fig. 18: Critical Path Delay of Canary Circuit

$$e_f = (1+\alpha)f - \frac{1}{\tau_{max}} \tag{13}$$

If the timing error is positive (i.e., no timing violation is present) and it is below 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:

$$f_{opt} = (1+\alpha)f \tag{14}$$

(vi) **Frequency update**:
If the timing error is negative (i.e., a timing violation is possible), or is higher 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 next frequency increase ratio is calculated with respect to the timing error as follows:

$$\alpha = \alpha + \beta_f e_f \tag{15}$$

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

C. MPC Look-Up-Table

As indicated previously, the optimal CPU and GPU frequencies are pre-computed for a set of possible input values and stored in two separate LUTs. Subsets of the LUTs are presented in Table I, where the utilization of the CPU is uniform between all the cores.

Results indicate that our thermal and timing-aware optimization methodology enables us to speed up the CPU operation by up to 25% (configuration B1 in Figure 8) when it is utilized 50% of the time, thanks to FCA cooling and power generation. For a CPU utilization percentage of 80%, FCAs enable up to 22% frequency boost, as the power consumption in this scenario is higher and the voltage drop and temperature are more critical. In all the scenarios (3D MPSoC configurations and core utilization rates), the temperature of the CPU die is maintained below 75°C.

TABLE I: Frequency Optimization Results (Reduced MPC Look-Up Table)

			A1		A2		B1		B2		B3		B4	
	Util.	T_{init}	δf	T_{max}	δf	T_{max}	δf	T_{max}	δf	T_{max}	δf	T_{max}	δf	T_{max}
CPU	50%	-	19.97%	45.7°C	19.85%	59°C	24.29%	37.5°C	23.87%	45.5°C	23.34%	59.9°C	23.08%	68.1°C
	80%	-	15.42%	47.7°C	15.44%	62.1°C	21.25%	39.2°C	21.84%	48.2°C	22.46%	64.3°C	22.78%	74.4°C
GPU	50%	40°C	13.1%	48.8°C	12.6%	62.8°C	18.9%	40°C	19.36%	49°C	19.79%	65.1°C	19.64%	74.4°C
	50%	70°C	18.64%	47.7°C	18.38%	58.56°C	24%	38°C	23.8%	45.8°C	23.34%	60°C	23.09%	68.2°C
	80%	40°C	4.73%	54°C	4.16%	69°C	10%	42.9°C	10.53%	53.4°C	10%	72.1°C	9.63%	82.6°C
	80%	70°C	8.84%	50.7°C	8.7%	66°C	14.36%	41.4°C	14.98%	51.2°C	15.46%	68.9°C	12.22%	81.6°C

In the GPU case, the frequency can be boosted between 12% and 19% for the different 3D MPSoC configurations, when its initial 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 23%. 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, 3D MPSoC configurations with a high FCA cooling capability, such as B1, enable a high-frequency boost while achieving lower temperatures (up to 42°C and 39°C for the GPU and CPU, respectively). For this configuration, the speed-up is possible thanks to the FCA capacity to reduce leakage, enabling considerably higher dynamic power consumption. In contrast, 3D MPSoC configurations with lower FCA cooling capacity achieve higher temperatures (up to 82°C in configuration B4) but comparable frequency boost ratios. In these configurations, FCAs produce more power due to higher temperatures. Therefore, they enable more computing capacity without additional stress on the power delivery grid, further voltage drop across the dies, and with lower cooling cost.

VI. Run-Time Benchmark Speed-up on 3D MPSoCs with FCAs and SC converters

A. Target Benchmarks characterization on the CPU and GPU

We explore a range of state-of-the-art benchmarks targeting high-performance multi-core platforms to evaluate our proposed run-time frequency optimization strategy under diverse 3D MPSoC usage scenarios. These benchmarks are selected as they represent different power consumption profiles. To characterize them on the CPU and the GPU, we run them in a 2D platform equivalent to the target 3D MPSoC, as indicated in Section III.

A brief description of these benchmarks is presented as follows:

Inception V3 (I3): is a very deep Convolutional Neural Network (CNN) architecture used for image recognition applications. We use TensorFlow to train an I3 model on the GPU using the ImageNet data set [38].

Inception V4 (I4): is a deeper and more uniform version of Inception V3 [39]. We train an I4 model on the GPU using the ImageNet data set.

Resnet (RN): is a computer vision deep CNN that we train on the GPU using the ImageNet data set [40].

Deep Speech (DS): is an end-to-end Deep Neural Network (DNN) used in automatic speech recognition (ASR) [41]. We train a DS model on the GPU, using a large-scale dataset of English readings.

Fairseq (FS): is a sequence modeling neural network used for translation, language modeling, error correction, and other text generation tasks [42]. We train it on the GPU for a dataset of spoken language translation.

Rodinia: is a benchmark suite for heterogeneous computing [43]. It includes applications that target multi-core CPU and GPU platforms such as:

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

SPEC: is a benchmark package containing standardized CPU-intensive applications stressing a system's processor and memory sub-system [44]. Such applications include:

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

Our benchmark characterization results on the CPU and GPU are presented in Figure 19 and Figure 20, respectively. For the AMD EPYC CPU case, we measure the cores and last-level cache (LLC) utilization percentages, using time steps of 100ms. These metrics serve to estimate the dynamic power consumption of each CPU component. The average benchmarks utilization rates are represented in Figure 19, ordered from high to low. A high CPU utilization characterizes compute-intensive benchmarks such as Cactus (Ca), weather forecasting (wrf), and ImageMagik (IM). Intuitively, they present a better opportunity for overall speed-up on the CPU, as the MPC frequency boost affects computing time but does not improve memory access speed.

For the case of the NVIDIA V100 GPU, we measure the total power consumption (including leakage) and the temperature of the chip, allowing us to estimate the dynamic power map of the GPU. In addition, we also measure the utilization of the GPU, which indicates the percentage of time when kernels are being executed on the board. Figure 20 presents the

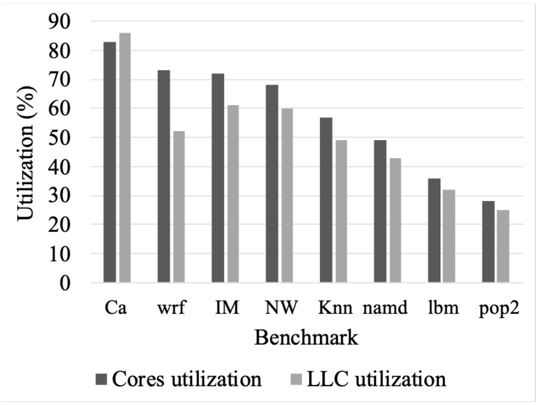


Fig. 19: Benchmark Utilization Statistics on the CPU

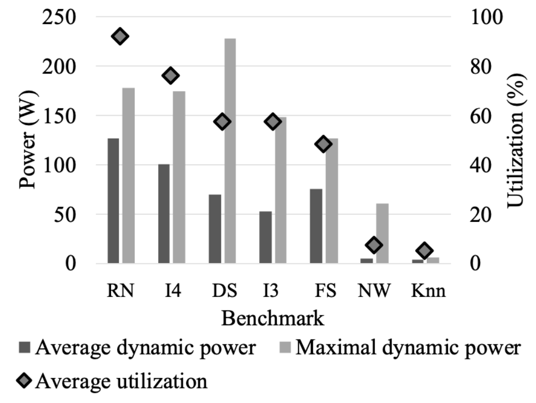


Fig. 20: Benchmark Utilization Statistics on the GPU

average usage metrics of various benchmarks, ordered by GPU utilization level. Similarly to the CPU case, benchmarks with the highest utilization, such as Resnet (RN) and Inception4 (I4), can benefit the most from overall speed-up on the GPU using the proposed MPC frequency optimization strategy.

B. Benchmark Speed-up Results on the 3D MPSoC with FCAs and SC converters

We simulate the explicit MPC from Section V-A when running the benchmarks in Section VI-A on the 3D MPSoC. Results are shown for the configuration B4, as it enables high frequency boost values while having a low cooling cost.

To estimate the execution of the benchmarks using our proposed MPC strategy, we first record the execution traces in the equivalent 2D system in sampling windows of 100ms (the minimal value dictated by the sensors and counters used to gather our experimental data). Then, we simulate the execution on the 3D MPSoC by compressing the original traces according to the speed-up rates determined by the MPC. To do so, we pessimistically assume 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, we consider that the same task schedule is executed. This way, only the computing time is compressed for each sampling step, as illustrated in Figure 21. The operation is repeated until the completion of the workload, allowing us to compute the achieved speed-up using the proposed MPC.

For comparison, we also measured workload speed-up when applying a fixed frequency boost throughout the execution.

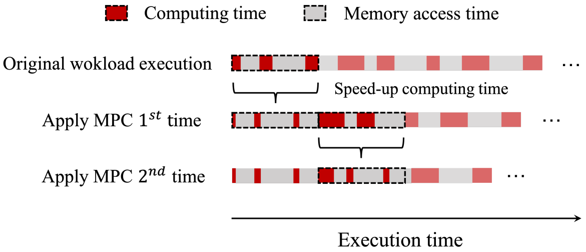


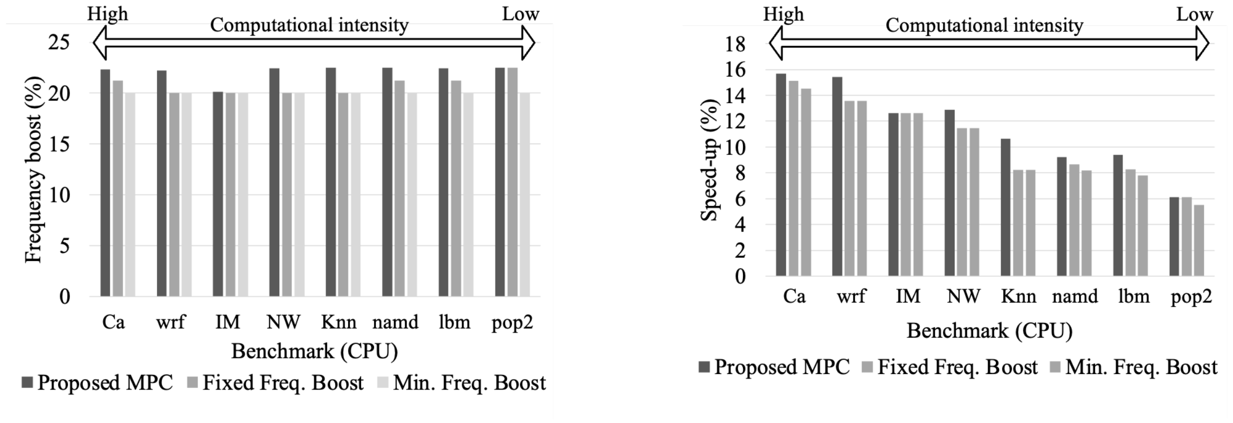
Fig. 21: Workload Speed-up at Run-Time

Hence, we considered two scenarios: *prior knowledge* of workload requirements and *no knowledge* of workload requirements. In the first scenario, the frequency boost is selected based on the peak power/utilization of the workload (*fixed freq. boost* in Figures 22 and 23). In the second scenario, the minimal frequency boost is selected, which corresponds to 100% utilization (*min. freq. boost* in Figures 22 and 23). The selected frequency boosts ensure that no constraint violations occur during the full workload execution in both cases.

1) Benchmark Speed-up on the CPU: The total achieved workload speed-ups on the CPU are shown in Figure 22, ordered by core utilization percentage. For most benchmarks, the MPC enables on average 23% faster CPU operation, as indicated in Table I (configuration B4). However, *compute-intensive* benchmarks present the highest overall speed-up due to their high utilization percentage. For instance, the SPEC Cactus framework (Ca) achieves 16% execution speed-up, at 83% average cores 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 speed-up with an average core utilization percentage of 28%.

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

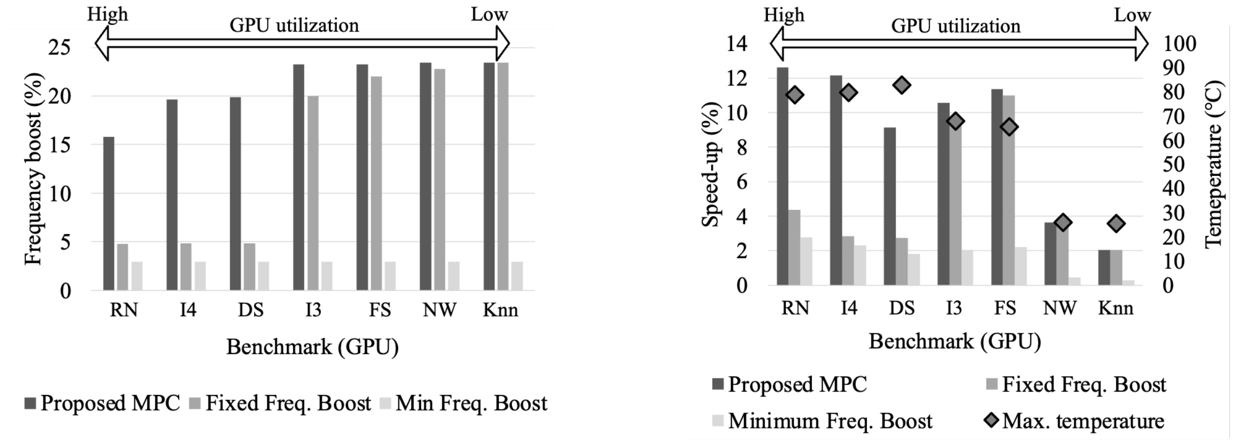
2) Benchmark Speed-up on the GPU: In the case of the GPU, the achieved benchmark speed-up values are shown in Figure 23, ordered by utilization level. The GPU runs on average between 16% and 23% faster using the proposed MPC. Typically, the benchmarks with a *high GPU utilization rate* and dynamic power consumption (Figure 20) 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 16% faster clock in a 3D system with FCAs and SC converters, compared to other benchmarks that enable 20% or higher speed-up rates. However, the RN training benefits from the highest overall workload speed-up. 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



(a) Average frequency boost

(b) Total speed-up

Fig. 22: CPU Workload Run-time Optimization Results



(a) Average frequency boost

(b) Total speed-up

Fig. 23: GPU Workload Run-time Optimization Results

frequency (with up to 23% boost). Still, their total execution speed-up is lower than 6%.

Compared to the proposed MPC, the alternative strategies achieve lower speed-up results, 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*, therefore decreasing by up to 15% 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 speed-ups because their highest utilization peaks still enable high frequency boosts. However, the *minimal frequency boost* strategy limits the overall speed-ups, as it accounts for the maximum possible workload utilization, which is never reached in practice.

In conclusion, our simulations indicate that the computation time can be accelerated by up to 23% for both the CPU and the GPU, using the proposed MPC. This is particularly effective for compute-intensive workloads, which benefit the most from the online frequency optimization. Furthermore, the 3D MPSoC speed-up is achieved without optimizing the software, or the architecture of the memories and logic dies. Performance improvements are only due to the increased leeway in terms of power and thermal budget made available by FCAs and SC converters, which allow enhancing the power performance of 3D MPSoCs.

C. *Comparison with State-of-the-art DTM Policies*

We emulate the workloads execution when applying a baseline DTM policy with DVFS and task migration [23] to the target 3D MPSoC using cold-plate-based cooling (CP in Figure 8). To reduce the power consumption and abide by temperature constraints, DVFS gradually lowers the voltage and frequency settings, down to 50% [45]. 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, incurring in a migration cost of 100ms [23]. Adopting this policy, we measured a slowdown of up to 25% and 40% for the CPU and GPU, respectively. Our approach leveraging FCAs and frequency control, instead, enables tangible run-time *gains* with respect to nominal conditions, as discussed in the previous Section.

VII. Conclusion

This paper has proposed a thermal and power performance management methodology for 3D MPSoCs with integrated

FCAs and SC converters. We have shown that, by coupling design-time characterization and run-time optimization, we can leverage the cooling and power supply capabilities of FCAs to accelerate workloads execution on 3D MPSoCs. At design time, we showcased how fine-grained thermal and power modeling can be employed to evaluate different FCA placements in the 3D stack, liquid flow rates, and SC converter designs to increase overall power efficiency. Moreover, FCAs could decrease the temperature, power consumption, and voltage drop of heterogeneous 3D platforms. These improvements can be harnessed to boost 3D MPSoC computation. To this end, we introduced a novel temperature and timing-aware MPC strategy to throttle the operation frequency of computing dies at run-time. We attained speed-ups up to 16% on a vast collection of benchmarks while satisfying design constraints. Our proposed performance management strategy can be generally applied to 3D platforms containing multiple high-performance computing dies, regardless of their architecture and technology. Our results demonstrate the potential of FCAs to achieve power-efficient 3D MPSoCs targeting modern high-performance applications. Furthermore, they open the door to extend the benefits of FCAs using strategies such as dynamic flow-rate adjustment to achieve optimal cooling, power generation, and FCA cost.

Acknowledgment

This work has been partially supported by the ERC Consolidator Grant COMPUSAPIEN (GA No. 725657), the EC H2020 WiPLASH (GA No. 863337), and the EC H2020 FVLLMONTI (GA No. 101016776) projects.

References

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

[2] P. Emma and E. Kursun. Opportunities and Challenges for 3D Systems and Their Design. *IEEE Design & Test of Computers*, 2009.

[3] M. Jung and S. K. Lim. A study of IR-drop Noise Issues in 3D ICs with Through-Silicon-Vias. *IEEE International 3D Systems Integration Conference (3DIC)*, 2010.

[4] A. Sridhar et al. PowerCool: Simulation of Integrated Microfluidic Power Generation in Bright Silicon MPSoCs. *International Conference on Computer Aided Design (ICCAD)*, 2014.

[5] H. Najibi et al. A Design Framework for Thermal-Aware Power Delivery Network in 3D MPSoCs with Integrated Flow Cell Arrays. *International Symposium on Low Power Electronics and Design (ISLPED)*, 2019.

[6] 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.

[7] E. Wong et al. 3D Floorplanning with Thermal Vias. *DATE*, 2006.

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

[9] 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 (GLSVLSI)*, 2012.

[10] A. Bartolini et al. Unveiling Eurora: Thermal and Power Characterization of the most Energy-Efficient Supercomputer in the World. *DATE*, 2014.

[11] H. Najibi et al. Towards Deeply Scaled 3D MPSoCs with Itegrated Flow Cell Array Technology. *GLSVLSI*, 2020.

[12] 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.

[13] 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 (JSSC)*, 2021.

[14] S. Wang et al. P/G TSV Planning for IR-drop Reduction in 3D-ICs. *DATE*, 2014.

[15] P. Sivakumar et al. Optimization of thermal aware multilevel routing for 3D IC. *Analog Integrated Circuits and Signal Processing*, 2019.

[16] 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.

[17] 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 (ISSCC)*, 2015.

[18] A. Andreev et al. PowerCool: Simulation of Cooling and Powering of 3D MPSoCs with Integrated Flow Cell Arrays. *Transactions on Computers (TC)*, 2018.

[19] L. Muller and J. W. Kimball. A Dynamic Model of Switched-Capacitor Power Converters. *Transactions on Power Electronics*, 2014.

[20] 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.

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

[22] A. Aggarwal et al. Temperature Constrained Power Management Scheme for 3D MPSoC. *IEEE Workshop on Signal and Power Integrity*, 2012.

[23] M. M. Sabry et al. Energy-Efficient Multi-objective Thermal Control for Liquid-Cooled 3D Stacked Architectures. *Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2011.

[24] A. K. Coskun et al. Modeling and Dynamic Management of 3D Multicore Systems with Liquid Cooling. *International Conference on Very Large Scale Integration (VLSI-SoC)*, 2009.

[25] F. Zanini, D. Atienza, and G. De Micheli. A Combined Sensor Placement and Convex Optimization Approach for Thermal Management in 3D-MPSoC with Liquid Cooling. *INTEGRATION, the VLSI journal*, 2013.

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

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

[28] S. Shim et al. A 16Gb 1.2V 3.2Gb/s/pin DDR4 SDRAM with Improved Power Distribution and Repair Strategy. *ISSCC*, 2018.

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

[30] D. Lee et al. A 1.2 V 8Gb 8-Channel 128GB/s High-Bandwidth Memory (HBM) Stacked DRAM With Effective I/O Test Circuits. *JSSC*, 2015.

[31] A. Sridhar et al. 3D-ICE: a compact thermal model for early-stage design of liquid-cooled ICs. *TC*, 2014.

[32] S. Pal et al. Design Space Exploration for Chiplet-Assembly-Based Processors. *Transactions on VLSI Systems*, 2020.

[33] Y. XIE et al. Design Space Exploration for 3D Architectures. *Journal on Emerging Technologies in Computing Systems*, 2006.

[34] F. Zanini et al. Online Thermal Control Methods for Multiprocessor Systems. *Trans. on Design Automation of Electronic Systems*, 2012.

[35] 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.

[36] J. Guerreiro et al. Modeling and Decoupling the GPU Power Consumption for Cross-Domain DVFS. *Trans. on Parallel and Distributed Systems*, 2019.

[37] P. Bogdan, R. Marculescu, and S. Jain. Dynamic Power Management for Multidomain System-on-ChipPlatforms: An Optimal Control Approach. *Transactions on Design Automation of Electronic Systems*, 2013.

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

[39] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *AAAI Conference on Artificial Intelligence*, 2017.

[40] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *CVPR*, 2016.

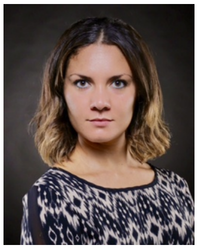
[41] D. Amodei et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. *Int. Conf. on Machine Learning*, 2016.

[42] M. Ott et al. fairseq: A fast, extensible toolkit for sequence modeling. *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

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

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

[45] J. Murray et al. Sustainable DVFS-Enabled Multi-Core Architectures with On-Chip Wireless Links. *Advances in Computers*, 2013.



Halima Najibi is a Ph.D. student at the Embedded Systems Laboratory (ESL) in the Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland. She received her B.S. and M.S. degrees in electrical and electronics engineering in EPFL in 2013 and 2015, respectively. She then worked as a hardware development engineer in Oracle Labs, Austin, TX, USA. In the recent years, her research focuses on 3D thermal and power-aware design for MPSoCs and many-core computing systems.

Alexandre Levisse is a scientist at the Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland. He received his Ph.D. degree in electrical engineering from CEA-LETI, France, and from Aix-Marseille University, France, in 2017. From 2018 to 2021, he worked as post-doctoral researcher in the Embedded Systems Laboratory (ESL) in EPFL. His research interests include circuits and architectures for emerging memory and transistor technologies, as well as in-memory computing and accelerators. Alexandre Levisse is the co-author of over 45 articles in peer-reviewed international journals and conferences.

Giovanni Ansaloni is a researcher at the Embedded Systems Laboratory (ESL) in the Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland. He previously worked as a post-doctoral researcher at the University of Lugano (USI), Switzerland, between 2015 and 2020, and at EPFL between 2011 and 2015. He received his Ph.D. degree in Informatics from USI in 2011. His research efforts focus on domain-specific and ultra-low-power architectures and algorithms for edge computing systems, including hardware and software optimization techniques. Giovanni Ansaloni has co-authored over 60 articles in peer-reviewed international conferences and scientific journals.

Marina zapater is an associate professor in the REDS Institute at the School of Engineering and Management of Vaud (HEIG-VD) of the University of Applied Sciences Western Switzerland (HES-SO) since 2020. She was a post-doctoral research associate in the Embedded System Laboratory (ESL) at the Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland, from 2016 to 2020. She received her Ph.D. degree in electronic engineering from Universidad Politécnica de Madrid (UPM), Spain, in 2015. Her research interests include thermal, power and performance design and optimization of complex heterogeneous architectures, from embedded AI-enabled edge devices to high-performance computing processors; and energy efficiency in servers and data centers. In these fields, she has co-authored more than 50 papers in various top-notch conferences and journals. She is an IEEE and CEDA member.

Miroslav Vasić is an associate professor at Centro de Electrónica Industrial at ETSII (UPM). He received the B.S. degree from the School of Electrical Engineering, University of Belgrade, Serbia, in 2005. He received his M.S. and his Ph.D. degrees in UPM in 2007 and 2010, respectively. His research interest includes application of power converters and their optimization. In the recent years, great part of his research activities has been related to the research of new semiconductor devices based on GaN and their impact on power electronics. Miroslav Vasić has published more than 70 peer-reviewed technical papers at conferences and in IEEE journals, he advised four Ph.D. thesis and holds six patents. In 2012, Miroslav received the Semikron Innovation Award for the teamwork on “RF Power Amplifier with Increased Efficiency and Bandwidth”. In 2015, he received a medal from Spanish Royal Academy of Engineering as a recognition of his research trajectory, and in 2016, he received UPM Research Projection Award for the best young researcher at Universidad Politécnica de Madrid. Miroslav actively serves as an Associated Editor in IEEE Journal of Emerging and Selected Topics in Power Electronics and IEEE Transactions on Vehicular Technology. Since 2021, he acts as the Vice-chair of the IEEE PELS TC 10- Design Methodologies.

David Atienza (M’05-SM’13-F’16) is professor of electrical and computer engineering, and head of the Embedded Systems Laboratory (ESL) at the Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland. He received his PhD in computer science and engineering from UCM, Spain, and IMEC, Belgium, in 2005. His research interests include system-level design methodologies for high-performance multi-processor system-on-chip (MPSoC) and low power Internet-of-Things (IoT) systems, including new 2D/3D thermal-aware design for MPSoCs and many-core servers, and edge AI architectures for wireless body sensor nodes and smart consumer devices. He is a co-author of more than 350 papers in peer-reviewed international journals and conferences, one book, and 12 patents in these fields. Dr. Atienza has received the ICCAD 2020 10-Year Retrospective Most Influential Paper Award, the DAC Under-40 Innovators Award in 2018, the IEEE TCCPS Mid-Career Award in 2018, an ERC Consolidator Grant in 2016, the IEEE CEDA Early Career Award in 2013, the ACM SIGDA Outstanding New Faculty Award in 2012, and a Faculty Award from Sun Labs at Oracle in 2011. He served as DATE 2015 Program Chair and DATE 2017 General Chair. He is an IEEE Fellow, an ACM Distinguished Member, and served as IEEE CEDA President (period 2018-2019) and Chair of EDAA (period 2022-2023).