Published in *Proceedings of the IFIP/IEEE 31st International Conference on Very Large Scale Integration (VLSI-SoC)*, Dubai, UAE, 16-18 october 2013, which should be cited to refer to this work.
DOI: 10.1109/PTC.2013.6652424

REMOTE: Re-thinking Task Mapping on Wireless 2.5D Systems-on-Package for Hotspot Removal

Rafael Medina*‡, Darong Huang*‡, Giovanni Ansaloni*, Marina Zapater† and David Atienza*
*Embedded Systems Laboratory (ESL), EPFL, Switzerland, †REDS Institute, HES-SO, Switzerland
Email: *{rafael.medinamorillas, darong.huang, giovanni.ansaloni, david.atienza}@epfl.ch, †marina.zapater@heig-vd.ch

***Abstract*—2.5D Systems-on-Package (SoPs) are composed by several chiplets placed on an interposer. They are becoming increasingly popular as they enable easy integration of electronic components in the same package and high fabrication yields. Nevertheless, they introduce a new bottleneck in inter-chiplet communication, which must be routed through the interposer. Such a constraint favors mapping related tasks on computing cores within the same chiplet, leading to thermal hotspots. In-package wireless technology holds promise to reconsider such a position because integrated wireless antennas provide low-latency and high-bandwidth communication paths, thus bypassing the interposer bottleneck. Furthermore, in this work, we propose a new task mapping heuristic that leverages in-package wireless technology to improve the thermal behavior of 2.5D SoPs executing complex applications. Combining system simulation and thermal modeling, our results show that we can distribute computation in wireless 2.5D SoPs to reduce peak temperatures by up to 24% through task mapping with a negligible performance impact.**

***Index Terms*—Task mapping, In-package wireless communication, Thermal management, Multi-chiplet systems.**

I. Introduction

The emergence of the Internet of Things (IoT) has led to embedded systems and mobile platforms with ever-increasing computational requirements. Modern multi-processor Systems-on-a-Chip (SoCs) can execute highly demanding workloads, within tightly constrained energy envelopes, by employing multiple cores operating in parallel. However, the fabrication of such large SoCs on a single silicon die negatively affects the yield, as even low defect densities may still lead to a high ratio of non-functioning devices [1]. 2.5D Systems-on-Package (SoPs) address this challenge by partitioning the components of an SoC among multiple smaller dies, co-integrated in a single package, and connected through the package substrate or interposer. However, inter-chiplet connections present much lower throughput and higher latency than intra-chiplet connections, due to (1) the large area of microbumps connecting chiplets to the interposer (limiting their number) and (2) the high capacitance of inter-chiplet wires and the need for signal repeaters throughout the interposer [1], [2].

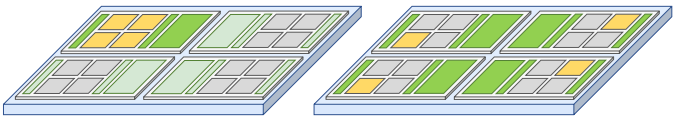
In this context, in-package wireless communication brings a disruptive alternative to wired inter-chiplet links [3]. Ultra-low-range wireless technology have been demonstrated to provide bandwidths of up to 120 Gb/s, employing transceivers and nanoantennas that can be co-integrated with CMOS technology [4], [5]. Critically, since in-package wireless provides direct links among chiplets (as opposed to wires routed via the interposer), it can support short-latency communication, while also freeing up micro-bumps, which can be used for other purposes, e.g., to provide supply voltages.

A further, and, to the best of our knowledge, currently overlooked potential benefit of in-package wireless is related to thermal considerations. Indeed, wireless links allow to distribute related tasks across chiplets (Fig. 1), smoothing SoPs thermal profiles and lowering peak temperatures. Herein we aim to quantitatively explore the advantages of wireless 2.5D SoPs in thermal, power, and performance aspects. To this end, we present an evaluation framework that not only facilitates simulation of wireless-enabled systems from a full-system perspective, but also explores the interplay among temperature, power consumption and runtime constraints through a thermal-aware task mapping heuristic.

Our investigation is particularly relevant in the context of IoT embedded systems, where size constraints disallow the use of large heat sinks and active cooling devices such as fans. Heat hinders the performance of integrated circuits and negatively impacts reliability, shortening their lifetime [6]. Moreover, controlling temperature profiles is key in this scenario, as the generated heat can cause discomfort to users.

Our approach combines application profiling and thermal/power modeling. In more detail, we employ the gem5-X full-system simulator [7] extended with in-package wireless support [3], the power estimator McPAT [8], and the thermal simulator 3D-ICE 3.1 [9] with support for the definition of non-homogeneous layers in the SoP layout. We utilize this framework to gather runtime and thermal information on the execution of machine learning and communication-intensive workloads when employing different task mapping strategies.

The contributions of the paper are summarized as follows:



(a) (b)

Fig. 1: System-on-Package where an application is executed within one chiplet (a) or distributed among all (b). Active cores and caches are highlighted in yellow and green respectively.

This work has been partially supported by the EC H2020 WiPLASH project (GA No. 863337) and the EC H2020 FVLLMONTI project (GA No. 101016776).

‡These authors contributed equally.

- We highlight that in-package wireless technology has the potential to foster a revolution for task mapping in 2.5D SoPs, bringing thermal considerations to the forefront, and propose a mapping heuristic to guide task assignment in these systems.
- Therefore, we introduce a framework to explore the relation between mapping strategy, performance, and temperature in different scenarios, which covers wired and wireless architectures.
- We show that distributed task mappings, enabled by wireless links in the package, achieve reductions in maximum temperature of up to 20°C for a 4-chiplet, 16-core system based on the ARM Cortex-A72 processor, with no increase in runtime. Conversely, corresponding task mappings using inter-chiplet wired links [2] would result in performance penalties of up to 10%.

II. Related Work

A. In-package Wireless Communication

Wireless communication between different chiplets is enabled by interfacing them to a transceiver and a nanoantenna [5]. Transceivers interface components that access the wireless network and nanoantennas. They handle serialization / deserialization and modulation / demodulation of data, and perform collision detection. They also implement a Medium Access Control (MAC) protocol to regulate synchronization, fair access, and collision handling in the shared transmission medium. Transmission/reception nanoantennae, in turn, radiate and capture the modulated data through the package.

As shown in this paper, wireless technology can be effectively leveraged to allocate tasks across chiplets, reducing hotspots while presenting little or no impact on runtime. Conversely, conventional wired solutions favor mappings where related threads are executed on the same chiplet, in order to avoid costly inter-chiplet data transfers. However, to comply with thermal constraints, clustered mappings in wired 2.5D SoPs require the use of techniques such as task migration [10] and Dynamic Voltage and Frequency Scaling [6], which themselves negatively affect performance. Additionally, previous mappings distributing computation across the die [11] have not analyzed their interaction with multi-chiplet architectures and wireless interconnects.

To evaluate the performance of systems implementing wireless links, previous works have utilized simulator-based frameworks, modeling runtime and power from the application [12] or full-system perspective [3], but neglecting thermal aspects. On the other side of the coin, while thermal analysis has also been included in full-system frameworks [13], [14], they separate system and interconnect simulations, without considering the interplay between them. We instead analyze runtime, power, and chip temperatures holistically, as part of a novel integrated framework targeting wireless 2.5D SoPs.

B. Thermal Modeling of Chiplet-based Systems

Thermal simulators for SoCs, such as HotSpot [15] and 3D-ICE [16], employ thermal equivalent resistors and capacitors to build thermal models. These state-of-the-art tools have paved the way for previous studies [17], [18] to analyze thermal profiles of chiplet systems. However, they typically instantiate a homogeneous layer of material that is then partitioned into smaller thermal grids for simulation. As a result, the chiplet is assumed to be homogeneous silicon, without boundaries between different chiplets. Even the air gap is treated as having the same thermal resistance as silicon. Thus, these models allow the propagation of heat among chiplets through the air gap, as illustrated in Fig. 2(a) and (b). However, silicon's thermal conductivity, 148 W/mK, is over 6,000 times higher than that of air, 0.0242 W/mK [19]. Consequently, from a thermal modeling perspective, the air gap should be treated as an open circuit, and heat flow should not traverse it.

In this work, instead, we rely on non-uniform modeling in the horizontal plane, as featured in 3D-ICE 3.1 [9], to realistically model 2.5D SoPs. This approach enables us to freely define the size and connections of thermal grids for different elements. Therefore, we are able to model the open circuit effect by decoupling the thermal resistance connections between silicon and air, as illustrated in Fig. 2(c) and (d).

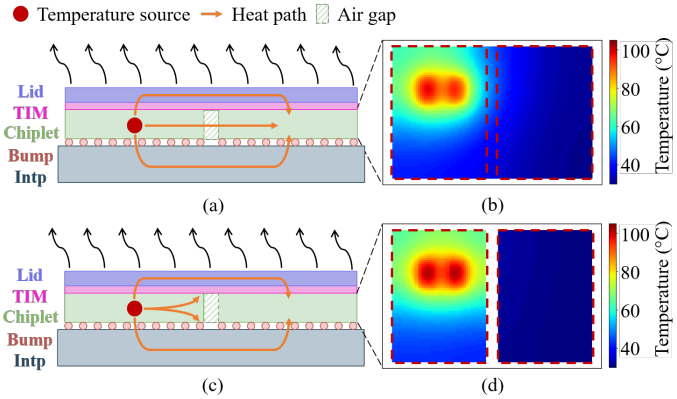


Fig. 2: Chiplet thermal simulation employing an existing thermal modeling technique (a, b) and the non-uniform method with 3D-ICE 3.1 [9] (c, d). For each simulation, heat flow through the layers, namely Interposer (Intp), Bump, Chiplet, Thermal interface material (TIM), and Lid from bottom to top, is shown in (a, c), and the thermal map for the layers in (b, d). Chiplets are delimited with dotted boxes.

Algorithm 1: Temperature-guided task mapping heuristic for wireless 2.5D Systems-on-Package

```
for task_i in {tasks} do
    if no_active_cores then
        Map task_i to the core with lowest temperature;
    else
        for chiplet_j in {chiplets_with_less_active_cores} do
            for core_{j,k} in {idle_cores_wihin_chiplet_j} do
                d[core_{j,k}] =
                  ∑(distance(core_{j,k}, active_cores));
        Map task_i to core_{j,k} with max (d[core_{j,k}]);
```

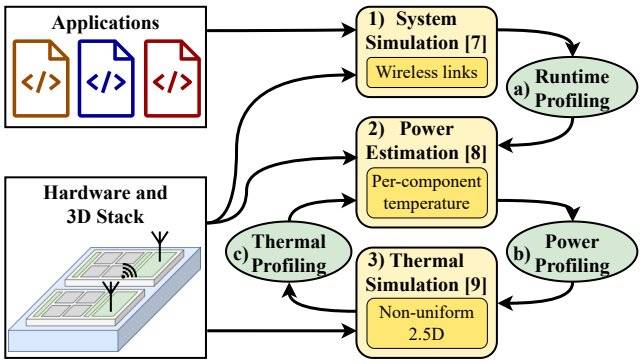


Fig. 3: Proposed system modeling framework. Runtime, power and thermal profiling is obtained as a result.

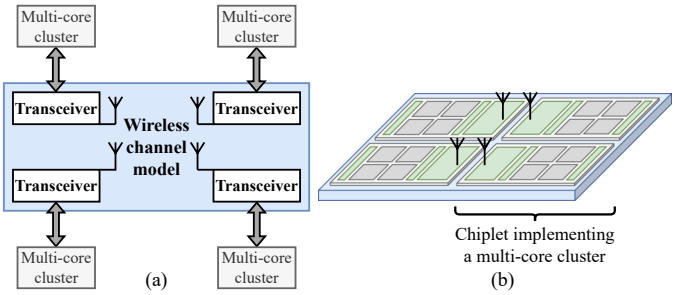


Fig. 4: Model of the wireless interconnect in system-level simulation, comprising one transceiver and nanoantenna per interfaced component (a), and emulated multi-chiplet system implementing a wireless chiplet interconnect (b).

III. Task Mapping in Wireless 2.5D Systems

In this paper, we propose a temperature-guided task mapping heuristic for wireless 2.5D SoPs. By assigning tasks to idle cores across chiplets, the heuristic (described in Algorithm 1), ensures that active cores are distributed as far apart as possible between chiplets, thus lowering the concentration of heat sources and minimizing the formation of on-chip hotspots. The heuristic highlights the advantages of wireless 2.5D SoP solutions in power, thermal, and performance aspects.

When mapping a task, the algorithm first checks for active cores running on the system. If the entire system is idle, the task is assigned to the core with the lowest temperature. If instead there are active cores on the system, the chiplets with the least active cores are considered. For each idle core in these chiplets, the distance between the idle core and all active cores is computed. The task is then mapped to the best fit, i.e. the idle core that has the largest sum of Euclidean distance to all active cores. The proposed task mapping algorithm's complexity scales quadratically with the number of cores.

IV. Thermal-aware System Modeling Framework

Our proposed system modeling framework provides a comprehensive simulation of system performance, power and temperature. As depicted in Fig. 3, it takes as input the application to be executed, as well as the description of the hardware components and how they are distributed in the 3D stack layout. The framework provides the following outputs (in green in Fig. 3): a) the execution runtime and usage statistics of the different components, b) the estimation of average power consumed during execution by each component, and c) the temperature distribution throughout the system stack.

To obtain these results, the framework is composed of three different modules, depicted in yellow in Fig. 3: a system-level simulator, a power estimator, and a thermal simulator. Its working flow, as shown in Fig. 3, comprises four steps:

1) The application is executed in the system simulator, which models the runtime behavior of the target architecture. This step provides statistics of the system and different architectural components.
2) The power estimator takes as input these statistics, the architectural description, and a set of initial temperature profiles. As a result, a first power breakdown of the system elements is obtained.
3) The power breakdown of the components serves as input to the thermal simulator, together with the 3D stack specification of the chiplet-based system. This simulation estimates the temperature of the silicon in the different regions of the chip.
4) Steps 2) and 3) are executed iteratively for a leakage-aware temperature simulation until the results converge. Every new iteration takes as input the new temperature and power estimations (respectively) at the component granularity. This step is detailed in Section IV-C.

The employed tools for this work are the system simulator gem5-X [7], the McPAT power estimator [8], and the latest release of the 3D-ICE (v3.1) thermal simulator [9]. All of these tools are open source and available online.

A. System-level Simulation of Wireless 2.5D SoPs

Full system-level simulators accurately estimate the performance of an architecture when executing applications, emulating the processor ISA, system architecture, I/O hardware, and OS. In this work, we employ the gem5-X simulator [7] with an extension that allows us to model on-chip wireless communication between system elements [3].

To emulate a wireless interconnect, the module simulates the shared wireless channel and one nanoantenna and transceiver per connected component, as illustrated in Fig. 4(a). These elements are modeled via configurable bandwidth and delay, as well as the definition of the MAC protocol implemented by the transceiver. Additionally, the wireless component can handle snooping messages to support cache protocols and memory consistency mechanisms. As a result, the implementation of the wireless technology is transparent to the interfaced architectural elements, the operating system, and application software, which therefore do not require any modification.

Using the described features of the wireless component, we can instantiate wireless interconnects at different levels of the system architecture. Here, we simulate multi-chiplet architectures where each chiplet implements a multicore cluster with shared caches and communicates with other chiplets via a shared wireless interconnect, as shown in Fig. 4.

Algorithm 2: Leakage-aware thermal simulation

$T \leftarrow T_{amb}$;
$P_d \leftarrow$ dynamic power from McPAT;
do
  $T_{lk} = \text{ceil}(T/T_{intv}) * T_{intv}$;
  $P_{lk} \leftarrow$ leakage power from McPAT @ T_{lk};
  $T \leftarrow$ 3D-ICE with P_d + P_{lk};
while $T > T_{lk}$;

TABLE I: Configuration of the simulated system.

Processors	16x cores ARM Cortex-A72 @2.0 GHz
L1-I Cache	16x private, 64 kB, 4-way, 2 cycle access
L1-D Cache	16x private, 64 kB, 4-way, 2 cycle access
L2 Cache	4x shared, 2 MB, 16-way, 20 cycle access
L3 Cache	4x shared, 8 MB, 32-way, 40 cycle access
Memory	DDR4 2400 MHz, 4 GB
System Clock	1600 MHz
Operating System	Ubuntu LTS 16.04

B. Chiplet Thermal Modeling

In this work, we employ 3D-ICE 3.1 [9] to enable accurate thermal modeling of multi-chiplet systems by considering the air gap between chiplets, as illustrated in Fig. 2(c). Simulation results are presented in Fig. 2(d), where one chiplet generates heat and the other one is idle. The heat generated by the left chiplet is mostly confined to itself, rather than dissipating to the other chiplet as in existing works (depicted in Fig. 2(b)). Therefore, this new thermal modeling approach more precisely reflects the thermal profiles of 2.5D SoPs, highlighting the benefits of our proposed temperature-guided mapping heuristic.

C. Leakage-aware thermal simulation

The framework incorporates leakage-aware thermal simulation to improve the accuracy of estimates. As presented in Algorithm 2, it starts with the chiplet temperature T set at the ambient temperature T_{amb} and the dynamic power variable P_d estimated with McPAT. Subsequently, the algorithm enters a leakage simulation loop that first identifies the leakage reference temperature T_{lk}, set in intervals of T_{intv} to accelerate convergence. Finally, 3D ICE utilizes the provided dynamic and leakage power information to calculate the updated temperature of the chiplet T. This loop continues until the temperature T is lower than T_{lk}, indicating that the leakage power no longer increase the temperature. When this condition is met, the final value of T represents the converged temperature of the chiplet. The algorithm exhibits fast convergence with an average of 3.9 iterations.

V. Experimental Setup

We target a wireless chiplet-based architecture employing four instances of a state-of-the-art chiplet design [20], each with a 4-core cluster and shared L2 and L3 caches. The system elements are configured as shown in Table I. The simulated embedded SoP has dimensions of 9.4 mm × 13.0 mm [20]. As illustrated in Fig. 2, the 3D stack comprises multiple layers, whose material characteristics are listed in Table II.

TABLE II: Thermal properties of different layers.

Layer	Material	Thermal conductivity (W/mK)	Heat capacity ($\times 10^6$ J/m^{3}K)
Lid	Copper	380	3.39
Thermal interface material	Thermal grease	3	1.45
Chiplet	Silicon	148	1.63
Bump	Solders	25	1.69
Interposer	PCB	8	3.60

The ambient temperature T_{amb} is set to 300 K, and the heat dissipation ability is adjusted to match the configuration of the target 2.5D SoP and control the operating temperature below 360 K. For the leakage power estimation introduced in Section IV-C, we set the T_{intv} as 10 K to meet the input requirement of McPAT. The power estimator targets a 22 nm process technology, providing a upper power boundary. Furthermore, the power for wireless communication is estimated based on the state-of-the-art transceiver in [4].

We assess the behavior of three task mapping strategies with different levels of workload distribution. First, we employ a conventional approach that assigns the thread of an application to the cores in the same chiplet, and name it clustered mapping (abbreviated as C). Next, we evaluate the mapping heuristic described in Section III, which maximizes task dispersion across chiplets for optimal temperatures. We name this approach temperature-guided (TG). Finally, we consider a meet-in-the-middle approach, which we name balanced (B), that only allocates tasks over half of the system chiplets, with the aim of reducing the bandwidth load due to task distribution.

For wireless implementation, we assume one transceiver per chiplet, enabling a 100 Gbps wireless channel, where the bandwidth is considered invariable for the set of silicon temperatures considered. The physical layer of the transceiver adds a delay overhead of 3 clock cycles. The used MAC protocol is token passing, where access to the shared medium is enabled by a token that is sequentially passed from node to node [12]. Such a protocol has been shown to outperform alternatives based on random access and backoff, when targeting chiplet interconnects in [3]. As a wired comparison baseline, we model a UCIe-compliant chiplet interconnect with 112 Gbps bandwidth and 100 clock cycle latency [2]. Finally, as an example of adopting frequency scaling to lower peak temperatures in the wired system, we model an additional configuration adopting UCIe and a core frequency of 1 GHz.

We test the architecture using three benchmarks. First, MobileNetV2 [21] allows us to assess performance in a state-of-the-art machine learning application, where the computation of each network layer is executed on every active core. As a result, bursts of data transmission are experienced between successive layers. To evaluate performance under extremely high traffic loads, the STREAM benchmark [22] is used. This application involves simultaneous access to the memory by every active core. Finally, multiple-core implementation of VGG8F [23] is used to have each core executing a different

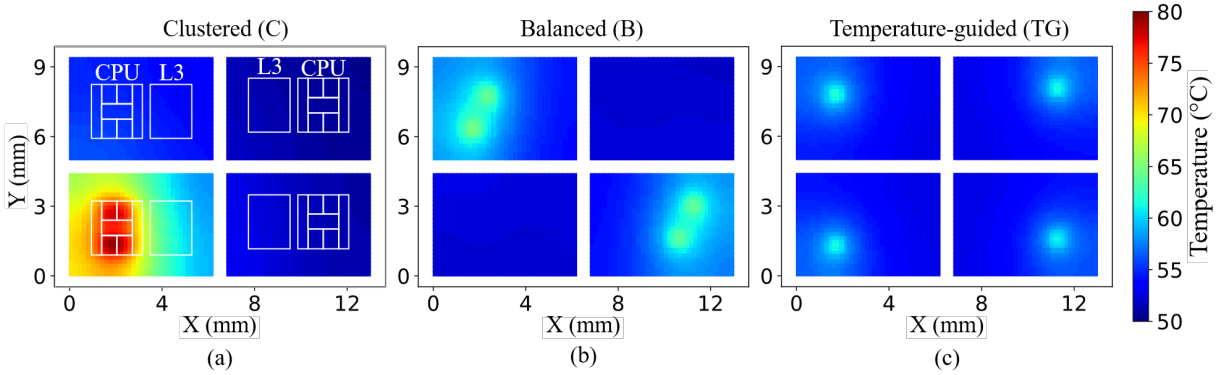


Fig. 5: Thermal maps of a 4-chiplet architecture executing MobileNetV2 on a wireless interconnect SoP at 2GHz, considering three task mapping configurations: (a) clustered, (b) balanced and (c) temperature-guided.

TABLE III: Analysis of MobileNetV2 execution on four cores of the described system. The considered mappings are clustered (C), balanced (B) and temperature-guided (TG). Performance values are highlighted in green if better than the reference, in yellow if similar, and in red if worse.

Inter-connect	Mapping strategy	Freq. (GHz)	Exec. time (s)	Exec. energy (J)	Max. temp. (°C)
UCIe	**C**	**2**	**1.56**	**16.2**	**80.2**
	C	1	3.08	14.6	51.1
	B	2	1.71	15.6	60.9
	TG	2	1.74	15.6	59.8
Wireless	C	2	1.54	15.2	78.0
	B	2	1.56	14.8	63.9
	TG	2	1.59	14.9	60.8

layer. This benchmark allows the assessment of a low-power application, particularly well suited to wireless interconnects.

VI. Results

A. Thermal benefits when distributing tasks among chiplets

Fig. 5 illustrates the thermal maps when executing MobileNetV2 on a chiplet-based architecture with wireless interconnect using the three mapping strategies. When employing clustered mapping, all four active cores are assigned to the same chiplet (Fig. 5(a)) and the system reaches the highest temperature (78°C) due to the high power density. When the active cores are distributed among two chiplets in the balanced strategy (Fig. 5(b)), the maximum temperature is reduced by more than 14°C, due to the balanced workload and power density. Using temperature-guided mapping (Fig. 5(c)) further reduces the maximum temperature by 4°C. The decrease in temperature with respect to the two chiplet cases is less substantial, since there the workload and the power density are already more evenly distributed. Overall, mapping tasks in different chiplets and non-adjacent cores accomplishes an efficient temperature reduction.

B. Performance analysis

To study performance trade-offs when spreading computation for hotspot avoidance, we evaluate the execution time and energy when executing MobileNetV2 on four active cores using the described 4-chiplet architecture. The results obtained are shown in Table III. We establish as a baseline the architecture using the UCIe interconnect with clustered mapping and running at 2 GHz.

We first highlight the limitations of conventional wire-based solutions. To this end, we report the effect of throttling the cores to 1 GHz, which lowers the energy consumption and temperature, but at the cost of an 97% increase in the execution time. Furthermore, alternative strategies that aim to distribute tasks across chiplets in wired scenarios (according to balanced and temperature-guided strategies) lower peak temperatures by up to 20.4°C, thanks to the longer distance between heat sources. The energy is slightly reduced in these scenarios, due to the decrees in leakage power, even if more accesses to shared memory are required. However, inter-chiplet communication overhead causes an increase in runtime of at least 10%. This slowdown is primarily due to the communication bottleneck, and consequent stalls, induced by the UCIe link.

The lower part of Table III displays the results when using the wireless interconnect. When using clustered task mapping, we obtain similar results to the baseline, slightly better thanks to a more efficient inter-chiplet communication. In addition, the cases where MobileNetV2 is distributed among two and four chiplets achieve more than 20% temperature decrease while maintaining a speed similar to the baseline. Energy consumption is also lower because the additional cost of more active caches is offset by lower leakage power and runtime.

C. Overall comparison for different workloads

In Fig. 6 we extend the analysis of Table III to all benchmarks, comparing wireless and wired mapping strategies, as defined in Section V. The figure reports peak temperature, runtime, and energy values of the 4-chiplet system executing different workloads on four active cores at 2 GHz.

Analyzing execution of the communication intensive STREAM benchmark (Fig. 6(b)), we first notice a slight increase in runtime when distributing the computation using UCIe. Since tasks require lower synchronization than MobileNetV2, the communication bottleneck is offset by the larger capacity when activating more caches. On the other hand, the execution of STREAM showcases a significant speed-up when employing a wireless interconnect, as this benchmark is very sensitive to latency. However, we notice a slowdown when employing the temperature-guided mapping,

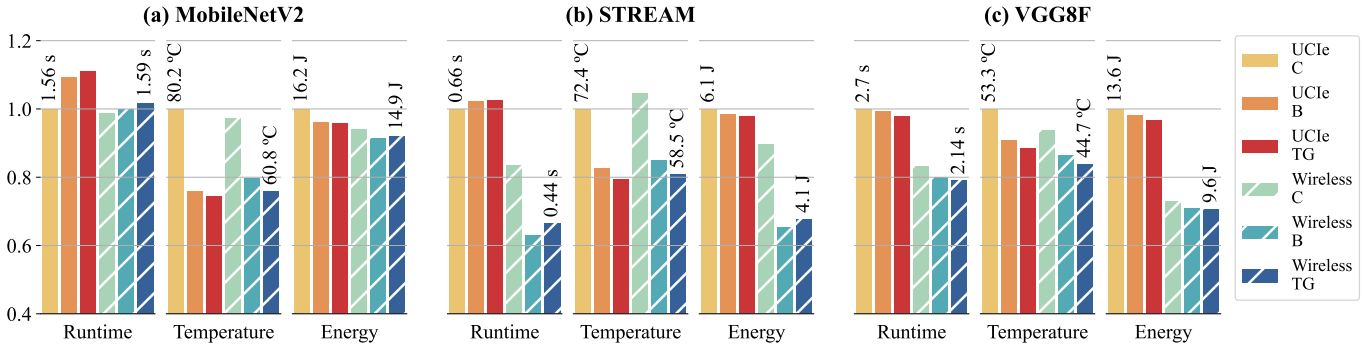


Fig. 6: Temperature, runtime and energy of the 4-chiplet system executing different workloads on 4 active cores at 2 GHz. Results are normalized with respect to the UCIe baseline employing the clustered (C) mapping strategy.

due to the very high amount of inter-chiplet communications. Temperature and energy reductions are obtained using both types of interconnect, similarly to the MobileNetV2 case. In particular, the shorter execution time on the wireless system leads to a corresponding lower energy consumption.

Finally, we present the results for the execution of VGG8F in Fig. 6(c). The temperature-guided distribution of this workload always benefits its performance, since the pipelined execution of layer spreads the transmission over time, reducing interconnect bottlenecks and allowing to benefit from the active caches. This fact further highlights the higher efficiency of the wireless interconnect. Temperature and energy are again reduced due to the temperature-guided and balanced mapping schemes. As in STREAM, the lower runtime over wireless translates into greater energy savings.

Overall, the benefits of employing mapping heuristics that spread the computation over separate chiplets using the wireless interconnect are demonstrated across workloads. Maintaining or even reducing the runtime and energy of the baseline, our proposed mapping heuristic is able to consistently lower maximum silicon temperatures, by as much as 24%.

VII. Conclusions

In this paper, we have examined the potential of in-package wireless communication to improve performance while reducing silicon temperature in the 2.5D chiplet-based SoP. By leveraging the advantages of wireless technology (high bandwidth and low latency), we propose a task mapping heuristic to address the challenge of high silicon temperatures in embedded systems and mobile platforms, where cooling mechanisms are limited and must fit within tight physical constraints. To accurately assess the use of novel mapping techniques that utilize in-package wireless communication, we present a modeling framework designed to estimate the power, performance, and temperature of chiplet systems. Our results demonstrate that distributed task mapping in a wireless 2.5D chiplet system can effectively reduce the maximum silicon temperature by 24%, while still meeting performance requirements. Our study showcases the potential of wireless communication in improving the thermal profile of 2.5D SoPs, without impacting runtime performance.

References

[1] G. H. Loh *et al.*, "Understanding Chiplets Today to Anticipate Future Integration Opportunities and Limits," in *DATE*, 2021.

[2] D. D. Sharma, "Universal Chiplet Interconnect Express (UCIe): Building an open chiplet ecosystem," Tech. Rep., 2022.

[3] R. Medina *et al.*, "System-Level Exploration of In-Package Wireless Communication for Multi-Chiplet Platforms," in *ASPDAC*, 2023.

[4] K. K. Tokgoz *et al.*, "A 120Gb/s 16QAM CMOS Millimeter-Wave Wireless Transceiver," in *IEEE ISSCC*, 2018.

[5] S. Abadal *et al.*, "Graphene-based Wireless Agile Interconnects for Massive Heterogeneous Multi-chip Processors," *IEEE Wireless Communications*, 2022.

[6] H. Khdr *et al.*, "Aging-aware boosting," *IEEE TC*, 2018.

[7] Y. M. Qureshi *et al.*, "Gem5-X: A many-core heterogeneous simulation platform for architectural exploration and optimization," *TACO*, 2021.

[8] S. L. Xi *et al.*, "Quantifying Sources of Error in McPAT and Potential Impacts on Architectural Studies," 2015.

[9] D. Huang *et al.*, "Accurate Thermal Modeling of Heterogeneous Multi-Core Processors," in *Proceedings of HSSB Workshop-ISCA*, 2022.

[10] T. Chantem *et al.*, "Temperature-aware scheduling and assignment for hard real-time applications on MPSoCs," in *DATE*, 2008.

[11] W. Liu *et al.*, "Thermal-Aware Task Mapping on Dynamically Reconfigurable Network-on-Chip Based Multiprocessor System-on-Chip," *IEEE TC*, 2018.

[12] V. Fernando *et al.*, "Replica: A Wireless Manycore for Communication-Intensive and Approximate Data," in *ACM ASPLOS*, 2019.

[13] J. Murray *et al.*, "Thermal hotspot reduction in mm-Wave wireless NoC architectures," in *ISQED*, 2014.

[14] M. S. Shamim *et al.*, "Evaluation of wireless network-on-chip architectures with microchannel-based cooling in 3D multicore chips," *SUSCOM*, 2019.

[15] W. Huang *et al.*, "HotSpot: A compact thermal modeling methodology for early-stage VLSI design," *IEEE TVLSIS*, 2006.

[16] A. Sridhar *et al.*, "3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling," in *ICCAD*, 2010.

[17] Y. Ma *et al.*, "TAP-2.5 D: A thermally-aware chiplet placement methodology for 2.5 D systems," in *DATE*, 2021.

[18] J.-H. Han *et al.*, "From 2.5 D to 3D Chiplet Systems: Investigation of Thermal Implications with HotSpot 7.0," in *iTherm*, 2022.

[19] "ANSYS software - engineering data," ANSYS, Inc., accessed 2023. [Online]. Available: https://www.ansys.com/

[20] M.-S. Lin *et al.*, "A 7nm 4GHz Arm-core-based CoWoS Chiplet Design for High Performance Computing," in *Symposium on VLSI Technology and Circuits*, 2019.

[21] M. Sandler *et al.*, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," 2018. [Online]. Available: http://arxiv.org/abs/1801.04381

[22] J. D. McCalpin, "Memory Bandwidth and Machine Balance in Current High Performance Computers," *IEEE TCCA Newsletter*, 1995.

[23] K. Chatfield *et al.*, "Return of the Devil in the Details: Delving Deep into Convolutional Nets," 2014. [Online]. Available: http://arxiv.org/abs/1405.3531