Towards Deeply Scaled 3D MPSoCs with Integrated Flow Cell Array Technology

Halima Najibi[1], Alexandre Levisse[1], Marina Zapater[1,2], Mohamed M. Sabry Aly[3], David Atienza[1]
[1]*Embedded Systems Laboratory (ESL), EPFL, Switzerland*
[2]*REDS Institute, University of Applied Sciences Western Switzerland (HEIG-VD, HES-SO), Switzerland*
[3]*SCSE, Nanyang Technological University (NTU), Singapore*

ABSTRACT

Deeply-scaled three-dimensional (3D) Multi-Processor Systems-on-Chip (MPSoCs) enable high performance and massive communication bandwidth for next-generation computing. However as process nodes shrink, temperature-dependent leakage dramatically increases, and thermal and power management becomes problematic. In this context, Integrated Flow Cell Array (FCA) technology, which consists of inter-tier microfluidic channels, combines on-chip electrochemical power generation and liquid cooling of 3D MPSoCs. When connected to power delivery networks (PDN) of dies, FCAs provide an additional current compensating the voltage drop (IR-drop). In this paper, we evaluate for the first time how the IR-drop reduction and cooling capabilities of FCAs scale with advanced CMOS processes. We develop a framework to quantify the system-level impact of FCAs at technology nodes from 22*nm* to 3*nm*. Our results show that, across all considered nodes, FCAs reduce the peak temperature of a multi-core processor (MCP) and a Machine Learning (ML) accelerator by over 22°C and 35°C, respectively, compared to off-chip direct liquid cooling. Moreover, the low operation voltages and high temperatures at advanced nodes improve up to 2× FCA power generation. Hence, FCAs allow to keep the IR-drop below 5% for both the MCP and ML accelerator, saving over 10% TSV-reserved area, as opposed to using a High-Performance Computing (HPC) MPSoC liquid cooling solution.

KEYWORDS

Flow Cell Array Technology; 3D Multi-Processor Systems-on-Chip; Technology Scaling; on-Chip Cooling; on-Chip Power Generation

ACM Reference Format:
Halima Najibi[1], Alexandre Levisse[1], Marina Zapater[1,2], Mohamed M. Sabry Aly[3], David Atienza[1]. 2020. Towards Deeply Scaled 3D MPSoCs with Integrated Flow Cell Array Technology. In *Proceedings of the Great Lakes Symposium on VLSI 2020 (GLSVLSI '20), September 7–9, 2020, Virtual Event, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3386263.3406923

1 INTRODUCTION

The increasing demand for High-Performance Computing (HPC) and high-density logic has driven Integrated Circuit (IC) technology industry to continuously scale CMOS transistors. Over the years, major innovations have been introduced and adopted to increase logic density and improve device performance. Multi-gate structures such as in field-effect transistors (FinFET), have enabled highly scaled devices with fin widths as small as 6*nm* [1][2][3][4]. They will continue being adopted in the coming years as the 3*nm* process is expected to begin around 2023 [5][6]. Although aggressive technology scaling achieves high transistor densities and improves computing performance, system communication bandwidth remains limited as off-chip interconnect scaling lags behind.

In this context, three-dimensional (3D) integration schemes enable deeply-scaled heterogeneous 3D platforms, with high performance logic and massive data exchange. However, 3D Multi-Processor Systems-on-chip (3D MPSoC) design using advanced CMOS technologies is generally limited by temperature and leakage. In particular, growing transistor densities result in dense dynamic switching per surface unit, which in turn generates higher temperatures and exponentially growing device leakage. Furthermore, as drive currents increase, the voltage drop (IR-drop) in power grids becomes more and more critical, affecting both performance and reliability of deeply-scaled 3D MPSoCs [7].

Flow Cell Array (FCA) technology, first introduced in [8], is a microfluidic channel-based solution for thermal and power management of 3D ICs, combining power generation and on-chip liquid cooling. It has emerged as a promising approach to solve heat dissipation and power delivery problems of 3D integration [9]. FCAs are able to transform absorbed heat into additional power, as electrochemical reactions between electrolytes inside the channels are accelerated with temperature [10]. When supplied to the Power Delivery Network (PDN) of 3D MPSoC dies, FCA-generated power compensates a significant portion of the IR-drop without increasing the number of Through-Silicon-Vias (TSVs) [11].

As FCAs promise great advantages for 3D MPSoCs, it is important to ensure their sustainability as we move towards deeply-scaled technologies. In this context, we analyze for the first time in literature the evolution of the on-chip cooling and IR-drop reduction benefits of FCAs when scaling 3D MPSoCs to advanced process nodes. Our contributions can be summarized in the following:

- We propose a methodology to estimate the power consumption of a 3D MPSoC computing die when scaled to technologies down to 3*nm*, at constant die area and operating frequency. Using as a starting point power values of a 22*nm* multi-core processor (MCP) and 28*nm* Machine Learning (ML) accelerator, we estimate both the leakage and dynamic power following industry-reported [1][2][3][4] and predictive [5][6] scaling ground rules.
- We show that FCAs achieve up to 35°C temperature reduction for both the ML accelerator and MCP, as technology scales, compared to off-chip liquid cooling strategies.

GLSVLSI '20, September 7–9, 2020, Virtual Event, China

ACM ISBN 978-1-4503-7944-1/20/09…$15.00
https://doi.org/10.1145/3386263.3406923

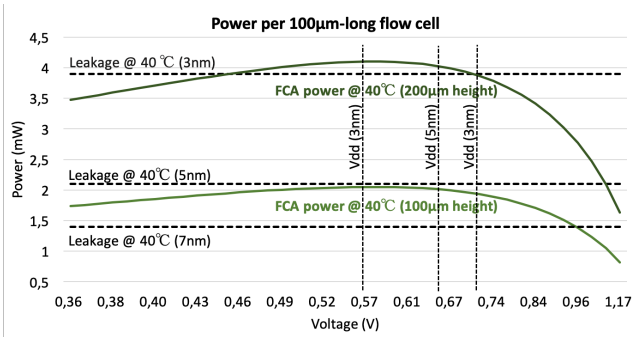


Figure 1: Generated power of a $100\mu m$ long and $50\mu m$ wide flow cell, with an inlet temperature of 40°C and an inlet speed of $2.5m/s$, compared to the leakage power of the covered $100\mu m \times 100\mu m$ logic cell

- We show that FCAs are able to eliminate up to 6% IR-drop when scaling the ML accelerator from the $22nm$ and $3nm$ process, compared to direct liquid cooling scenario. In the case of the MCP, FCAs eliminate up to 12% IR-drop when moving from the $22nm$ to $3nm$ node.
- We demonstrate the significant potential of FCAs for power and thermal management of 3D MPSoCs with advanced technologies. We motivate the design of power-efficient next-generation 3D systems with FCAs, saving over 10% of total die area reserved for power TSVs and up to two layers of metallization, with respect to a state-of-the-art HPC MPSoC direct liquid cooling system [12][13].

2 BACKGROUND ON FCA TECHNOLOGY

3D MPSoC design is challenged by power and thermal management problems. Heat extraction is difficult due to the low thermal conductivity of bonding layers. In addition, power delivery complexity increases with the number of stacked dies, more specifically due to the IR-drop across PDNs, and exponentially-growing leakage. In this context, FCAs were introduced as an efficient solution to overcome heat dissipation and power management challenges of 3D MPSoCs [10][11], serving as extra power sources thanks to their power generation capability.

FCAs consist of micro-fluidic channels within the inter-tier silicon substrates of 3D MPSoC dies. Within the channels, two electrolytic solutions co-laminarly flow, with electrodes placed in the side surface. FCAs absorb heat generated by the switching of logic gates, and allow to keep 3D MPSoC layers at low temperatures, down to 40°C [10][11], for an average density of $30W/cm^2$. Furthermore, the absorbed heat increases the reactions between electrolytes, generating an electric current between FCA electrodes. This current can be supplied to the logic gates, directly connecting the electrodes to the PDN of the chip, as proposed in [11]. Regarding the fabrication costs of 3D MPSoCs with FCAs, our industrial collaborators expect it to be marginally higher (less than 10%) than current costs of inter-layer cooling. As this technology already requires the additional etching and coating processes that micro-scale FCAs require, only the positioning of electrodes to channel walls is needed as extra step [10][9].

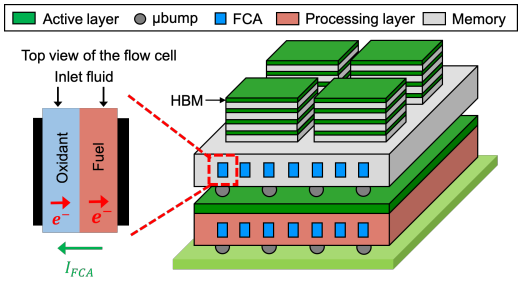


Figure 2: 3D MPSoC with a memory layer and a processing layer, the architecture of the processing layer is either based on the POWER8 processor or the TPU ML accelerator

Electro-thermal analyses [8] show that FCAs of $100\mu m$ height and $50\mu m$ pitch can compensate entirely the leakage of a die manufactured with the $7nm$ process node, and cooled to 40 °C. For the $3nm$ technology, leakage can be entirely compensated by FCAs of $200\mu m$ height, for the same channel liquid and die temperatures, as shown in Figure 1. To quantify FCA power generation benefits on 3D MPSoC PDNs, we proposed in [11] a fine-grain design framework that analyses the effects of connecting FCAs to the power grid of dies. We showed that the FCA current allows to minimize IR-drop in the metal lines of a high performance processor die. Therefore, it prevents performance degradation due to slow transition or failure of gates in highly power-consuming areas [7]. Furthermore, FCAs allow to relax power grid density requirements in areas with lower power consumption. In this work, we analyse how the IR-drop reduction and cooling capabilities of FCAs scale when 3D MPSoCs are fabricated with next-generation deeply-scaled technologies, with high integration densities and dramatically increasing leakage.

3 DEEPLY SCALED 3D MPSOC DESIGN WITH FCAS

To evaluate the scaling of FCA on-chip cooling and power generation capabilities when designing 3D MPSoCs with advanced CMOS technologies, we design a two-layer 3D MPSoC composed of a multi-core computing layer and a memory layer. $100\mu m$-wide FCAs are etched in the silicon substrate of both dies, with a pitch of $50\mu m$. Figure 2 illustrates the considered 3D MPSoC. The top memory layer contains four second generation HBM memories with 4 DRAM layers. Each HBM has a base die size of $71mm^2$, and consumes a total power of $15W$. To explore different power consumption profiles of the bottom processing layer, we first base its architecture on the IBM POWER8 processor [14], which has a highly non-uniform powermap. Alternatively, we use Google's Tensor Processing Unit (TPU) for ML applications [15], assuming the different components have a uniform activity.

The POWER8 processor layout is shown in Figure 3, fabricated using the $22nm$ Silicon-on-Insulator (SOI) CMOS technology. It comprises 12 computing cores, and has a base die size of $649mm^2$ [14]. To model its power consumption, we used a real measured powermap when running a workload where all cores are active with the maximal nominal frequency, with a total power of $190W$. The extracted powermap contains multiple high power density regions (hotspots) mainly concentrated in the computing cores. The rest of the chip has a significantly lower power consumption. Given the

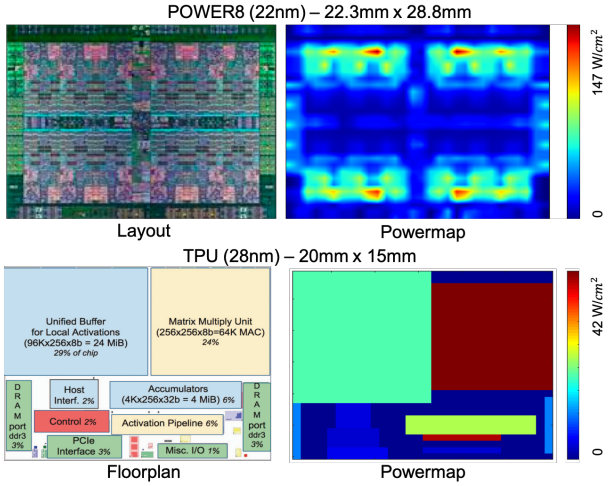


Figure 3: POWER8 layout, TPU floorplans, powermaps

large size of the POWER8 processor, HBMs in the corresponding 3D MPSoC memory layer are distanced from each other.

Figure 3 also shows the TPU floorplan [15], with a size of $300mm^2$ and a total power of $75W$, fabricated using the $28nm$ bulk CMOS process. The powermap is extracted using an integrated power, area and timing modeling framework [16]. The Matrix Multiply Unit, heart of the accelerator, has the highest power density and constitutes the biggest die hotspot. For this 3D MSPoC configuration, HBMs in the top memory layer are placed closer to each other.

In this work, we scale the computing die (POWER8 and TPU) to smaller technology nodes to evaluate FCA performance. We assume a fixed area footprint for all scenarios (by increasing the number and size of computing elements) and a constant device switching activity (i.e constant operating frequency). Then, we estimate the powermaps corresponding to each process.

4 TECHNOLOGY SCALING METHODOLOGY

IC technology industry has over the years scaled down transistor and interconnect sizes, in an effort to improve IC performance while maintaining constant power densities and low fabrication costs. Lithography process scaling continues with the emergence of new CMOS structures such as FinFET, which allow to drastically reduce transistor feature size and achieve high drive currents and low short-channel effects [1][2][3][4]. In this work, we explore different process nodes to assess the evolution of FCA on-chip cooling and IR-drop reduction efficiency with technology scaling. We analyse the $28nm$ bulk CMOS process, and the $22nm$, $14nm$, $10nm$, $7nm$, $5nm$ and $3nm$ FinFET processes. Power performance of chips depends on several technology parameters, which are scaled based on industry-reported and predictive values shown in Table 1:

- V_{dd}: The supply voltage.
- W_{eff}: The effective gate width of the transistor. For bulk CMOS, it is equal to the gate width W_{gate}, and for FinFET, it is computed as a function of the fin width W_{fin} and height H_{fin} according to Equation 1:

$$W_{eff} = W_{fin} + 2H_{fin}. \tag{1}$$

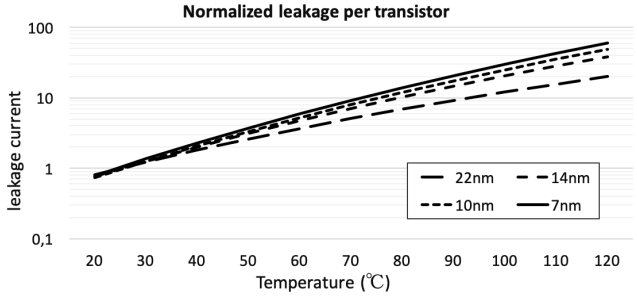


Figure 4: Normalized leakage per transistor, with respect to the value at 25°C, for different process nodes

- $I_{off}(t)$: The transistor leakage current per gate width at temperature t. Devices are typically sized to achieve $10nA/\mu m$ leakage per gate width for low and medium performance, and $20nA/\mu m$ leakage per gate width for high performance [1], at the reference temperature of 25°C.
- $P_{dyn/device}$: The dynamic power consumption per logic device for a constant operating frequency. It is estimated by calculating the energy per device switching CV^2 of a ring-oscillator circuit model. Interconnect parasitics and device are coupled to estimate the total capacitive load C [5][6].
- CPP: The contacted gate (poly) pitch.
- MP: The minimum metal pitch.
- $\rho_{transistor}$: The transistor density, which scales with the contacted gate pitch and minimum metal pitch:

$$\rho_{transistor} \propto CPP \times MP. \tag{2}$$

We analyse how the power consumption of computing dies (POWER8 and TPU) evolves as we scale the fabrication process from the $28nm$ bulk CMOS down to $3nm$ FinFET. We first scale the dynamic power according to the dynamic power consumption per device and number of devices in the die. Starting from the total dynamic power of a die fabricated using a technology node n_0, we estimate the dynamic power of a die with the same size fabricated with a technology node n, according to Equation 3:

$$P_{dyn}(n) = P_{dyn}(n_0) \times \frac{P_{dyn/device}(n)}{P_{dyn/device}(n_0)} \times \frac{\rho_{transistor}(n)}{\rho_{transistor}(n_0)} \tag{3}$$

Then, we scale the leakage power of the chip at a temperature t based on the leakage per transistor gate width, effective transistor gate width, and transistor density. Hence, leakage power is calculated according to Equation 4.

$$P_{leak}(t) = (I_{off}(t) \times W_{eff} \times \rho_{transistor} \times Area)\, V_{dd} \tag{4}$$

To evaluate the temperature-dependant leakage $I_{off}(t)$ with respect to the value at 25°C (Table 1), we use predictive models (PTM) for sub-20nm technology nodes [19]. We simulate a NAND2 gate with different process nodes in HSPICE, and assume other gates behave similarly in terms of temperature-dependency. We extract the leakage of the pull-up and pull-down networks respectively. Figure 4 represents the normalized leakage current per transistor with respect to the reference value at 25°C, for different technologies and temperatures. Leakage scales exponentially with temperature, and becomes extremely critical as transistor density dramatically

Technology (nm)	V_{dd} (V)	I_{off} at 25°C ($nA/\mu m$)	W_{eff} (nm)	CPP (nm)	MP (nm)	$\rho transistor$	$P_{dyn/device}$	heat transfer coefficient (fan-based [17]) $W/\mu m^2K$		heat transfer coefficient (Eurora [12]) $W/\mu m^2K$	
								MCP	accelerator	MCP	accelerator
28 [18]	1	20	76	117	110	1	1		$5.5\ 10^{-9}$		$9.7\ 10^{-9}$
22 [1]	1	20	76	100	90	1.43	0.55	$6.1\ 10^{-9}$	$6.45\ 10^{-9}$	$9.7\ 10^{-9}$	$9.7\ 10^{-9}$
14 [2]	0.8	20	92	70	70	2.63	0.3	$6.7\ 10^{-9}$	$6.4\ 10^{-9}$	$9.7\ 10^{-9}$	$9.7\ 10^{-9}$
10 [3]	0.75	20	90	54	36	6.62	0.17	$8.7\ 10^{-9}$	$8.6\ 10^{-9}$	$13.5\ 10^{-9}$	$13.5\ 10^{-9}$
7 [4][5]	0.7	20	56.5	44	24	12.19	0.09	$9.5\ 10^{-9}$	$8.8\ 10^{-9}$	$15\ 10^{-9}$	$14.3\ 10^{-9}$
5 [5][6]	0.65	20	56.5	32	20	20.11	0.05	$9.5\ 10^{-9}$	$8.7\ 10^{-9}$	$15\ 10^{-9}$	$13.5\ 10^{-9}$
3 [5][6]	0.55	20	56.5	24	12	44.69	0.03	$10\ 10^{-9}$	$10\ 10^{-9}$	$23.5\ 10^{-9}$	$17.9\ 10^{-9}$

Table 1: Technology scaling parameters

grows with technology. By significantly reducing die temperatures, FCAs allow to limit the overall power consumption of 3D MPSoCs.

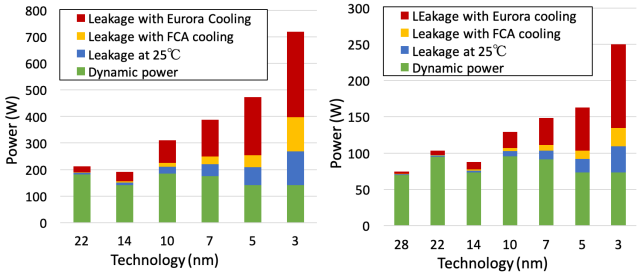
5 3D MPSOC MODELING AND ANALYSIS

In this work, we evaluate the thermal and power performance of FCAs in next generation 3D MPSoCs using advanced technology nodes. For this purpose, we use the 3D MPSoC configurations described in Section 3, namely the POWER8-based and TPU-based 3D MPSoCs. We scale both processor and ML accelerator to technologies with smaller feature sizes: 14nm, 10nm, 7nm, 5nm and 3nm FinFET. We assume a constant die size by increasing the size of cores for the POWER8-based processor, and size of the Matrix Multiply Unit and Unified Buffer for the TPU-based ML accelerator, hence improving the throughput of computing dies using adaptive load balancing and task scheduling techniques [20][21]. Moreover, we assume a constant dynamic switching activity per device (i.e operating frequency), when scaling dies to smaller technologies.

5.1 3D MPSoC Cooling Strategies

We use 3D-ICE [22], a compact thermal simulator for liquid-cooled 3D ICs, to evaluate the cooling capabilities of FCAs. Additionally, we compare their performance with the following state-of-the art off-chip cooling systems:

- We first model a fan-based cooling system achieving the POWER8 peak temperature of 90°C [14]. The cooling efficiency of the heat sink model is scaled to achieve the same maximal temperature when processing dies are synthesized with different technology nodes, within feasible limits [17] (heat transfer coefficients shown in Table 1).
- Then, we model the direct liquid cooling solution of Eurora Supercomputer [12], a heterogeneous platform with high performance hardware components such as Intel Xeon E5, Intel Xeon Phi processor and NVIDIA Kepler K20 GPU. Eurora cooling infrastructure uses a cold plate, and is able to achieve a maximal temperature of 95°C for a Xeon processor running at the maximal frequency of 3.1GHz, with a high average power density of $53W/cm^2$. Eurora cooling system is also scaled to maintain the original cooling efficiency. This implies changing several possible design parameters, such as cold place dimensions and materials, refrigerant type, or coolant temperature [13]. The heat transfer coefficients of equivalent heat sink models are shown in Table 1.



(a) POWER8-based processor **(b) TPU-based accelerator**

Figure 5: Power breakdown of the POWER8 and TPU-based dies for different technologies and cooling strategies

5.2 Power Modeling at Advanced Process Nodes

We estimate the powermaps of the POWER8-based and TPU-based dies when fabricated using technologies with smaller feature sizes: 14nm, 10nm, 7nm, 5nm and 3nm FinFET. Dynamic power is scaled according to Equation 3 in Section 4. Leakage is estimated using Equation 4 in Section 4, according to the temperature maps when 3D MPSoCs are cooled using the different strategies in Section 5.1. Figure 5 shows the total dynamic and leakage power of both dies at different technology nodes. Power at the 22nm node is the consumption of the original POWER8 processor for the MCP (Figure 5a), and power at 28nm corresponds to the consumption of the original TPU accelerator (Figure 5b). Power values of more advanced nodes are estimated according to different scaling parameters (Table 1):

- The dynamic power is calculated according to the dynamic consumption per device and transistor density, assuming constant switching activity per logic gate. These two parameters have opposite scaling trends, resulting in a non-monotonic scaling of the total dynamic power.
- The leakage values in blue in Figure 5 correspond to the scenario where dies are at the uniform base temperature of 25°C. FCA leakage is the additional leakage calculated based on the temperature map of dies when cooled using FCAs. Finally, Eurora leakage is the additional leakage that corresponds to the temperature of dies when cooled using Eurora cooling system.

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

5.3 PDN Modeling and IR-drop Analysis

After calculating the powermaps of both processing dies for technologies ranging from 28nm bulk CMOS to 3nm FinFET, we use the fine-grain PDN modeling and analysis framework from [11] to evaluate the voltage distribution across the POWER8-based and TPU-based dies. We use a PDN structure where power TSVs are arranged in groups, each delivering power to an independent subgrid. Subgrids correspond to individual cores in case of the MCP, and to the different components shown in the floorplan in Figure 3 for the TPU-based accelerator. TSVs have a fixed diameter of 5μm and a pitch of 5μm. FCAs also have a fixed width of 100μm, height of 100μm, and pitch of 50μm. FCA inlet temperature is fixed at 27°C, and speed of the liquid at 2.5m/s. FCA electrodes are connected to the power delivery grid in the back-end-of-line (BEOL) of 3D MPSoC dies, in order to directly supply the FCA-generated power to logic gates, as proposed in [11]. Finally, the power grid resistance is scaled according to the sizes of top metal layers dedicated to power delivery, for different technology nodes [5][6].

6 RESULTS AND DISCUSSION

6.1 FCA on-chip Cooling Capabilities

In this section, we present thermal analysis results when simulating the 3D MPSoCs presented in Figure 2 using 3D-ICE. We analyse the thermal behaviour using FCAs, Eurora supercomputer cooling system, and fan-based heat sink model to meet the maximal allowed chip temperature constraints. Figure 6 shows the maximal temperature of the POWER8-based and TPU-based dies.

Without changing the FCA design parameters specified in Section 5.3 (i.e channel dimensions, liquid flow rate, coolant inlet temperature, etc.), they are able to maintain a high cooling efficiency for ultra-scaled technology nodes, with double the original dies power densities. FCAs keep the peak temperature of the MCP between 40°C and 53°C, when moving from the 22nm to 3nm process node. For the TPU-based accelerator, which overall has a 22% lower average power density, the peak temperature when using FCA cooling remains between 33°C and 40°C, when scaling the die from the 28nm to 3nm node.

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

In addition, FCAs generally outperform Eurora supercomputer cooling system for both the TPU-based and POWER8-based dies. Particularly, the FCA-cooled accelerator achieves up to 35°C difference in peak temperature than the Eurora system-cooled die, at the 3nm process node. Unlike the FCAs, Eurora cooling system is scaled starting from the 10nm technology node (as shown in Table 1), to meet the original cooling efficiency which was designed for the Intel Xeon E5 processor fabricated with the 14nm lithography process. This shows that FCAs, at their current advancement state, have the potential to be a first-choice cooling strategy for next generation high-power-density 3D MPSoCs.

6.2 FCA IR-drop Reduction Capabilities

In this section, we present PDN analysis results for the 3D MPSoCs. After simulating the thermal behavior of chips, the fine-grained

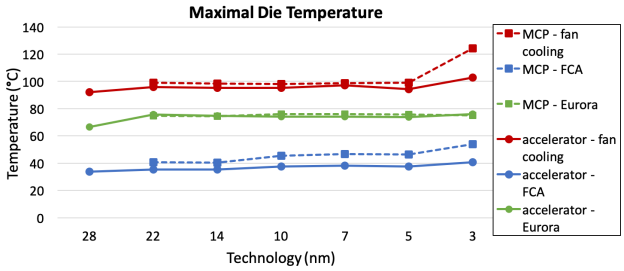


Figure 6: Maximum die temperature, for the POWER8-based TPU-based dies, with different cooling methods

powermaps are computed according to their temperature maps (using the different cooling strategies). The full PDN of each die is modeled, with and without adding FCAs, and simulated using HSPICE to extract the voltage maps. Figure 7 shows the maximal IR-drop percentage at the power grid of the processor and accelerator dies, with fan-based cooling, Eurora cooling, and FCAs. In case of the fan-based cooling and Eurora cooling, the source (V_{dd}) and ground voltages are only supplied from the printed circuit board (PCB) via TSVs supplying individual subgrids. In case of FCAs, power supply comes from both the PCB (via the power TSVs) and FCAs (which are directly connected to the grid). Our results support the following observations:

- FCA power ensures maintaining the IR-drop across 3D MPSoC dies under 5%, which is a typical IR-drop constraint for high performance chip design [7], and is equivalent to under 27mV for the 3nm technology. High FCA IR-drop reduction is due to two main factors. The first one is the efficient cooling, which ensures lower leakage. The second is the improved power generation capacity of FCAs as we move towards smaller-size processes. On the one hand higher temperatures accelerate the reaction rate of electrolytes and hence improve power generation [11]. On the other hand, lower operation voltages ensure up to 2x higher generated power between the 28nm and 3nm node (Figure 4).
- For Eurora-cooled 3D MPSoCs, the IR-drop value of computing dies is between 5% for the original chip and 15% for the one scaled to the 3nm technology node. Eurora system can be scaled to meet thermal requirement of ultra-scaled 3D MPSoCs, however power delivery system design requires significant improvements in order to allow such 3D MPSoCs, as their power consumption density substantially increases.
- Fan-cooled dies achieve very high IR-drop values, over 50% for the 3nm node due to exponentially-growing leakage. This demonstrates that without highly-efficient cooling, IR-drop management becomes extremely difficult for 3D MPSoCs fabricated with small technology nodes.

To compare the PDN performance of FCA-powered and Eurora-cooled 3D-ICs, we quantified the additional power delivery component requirements in order to achieve the 5% IR-drop constraint in the case of the Eurora-cooled 3D MPSoCs. To do so, we calculated the number of power TSVs (or TSV islands) needed to further lower the IR-drop, and if necessary, the number of additional metal layers dedicated to power delivery. Figure 8 shows the total number of TSVs (in light green) and the total TSV area needed for the

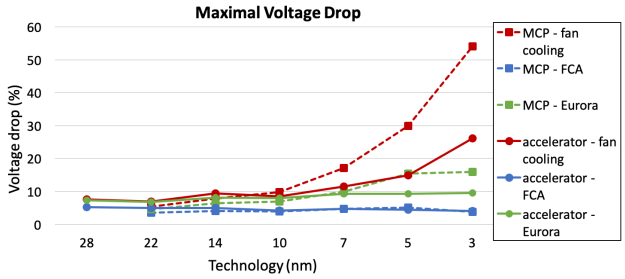
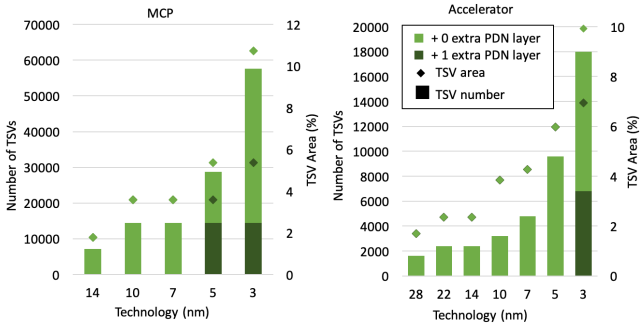


Figure 7: Maximum IR-drop value, for the POWER8-based and TPU-based dies, with different cooling methods



(a) POWER8-based processor **(b) TPU-based accelerator**

Figure 8: Number of additional metal layers, power TSVs, and total TSV area required to achieve 5% IR-drop, for the POWER8-based and TPU-based dies with Eurora cooling

POWER8-based and TPU-based dies, Figure 8a and 8b respectively, to meet the IR-drop constraint for different technology nodes. It also shows the number of TSVs and extra PDN layers to meet the target (in dark green). The TSV number includes TSVs that delivers V_{dd} and ground, and each extra PDN layer is equivalent to 2 physical metal layers in the BEOL (both directions). Our results show that to design highly-scaled 3D MPSoCs using Eurora cooling, over 10% of area needs to be reserved for TSVs, which not only is a high requirement, but it dramatically affects routing complexity as TSVs need to be scattered across the die to reduce wire lengths. Alternatively, it is possible to meet the IR-drop target by adding additional metal layers in the BEOL, which roughly adds 10% fabrication cost per layer. In conclusion, the current FCA technology is a very efficient solution for 3D MPSoC power management, without the need for extra power delivery components requirement. This motivates to integrate them in next-generation high-performance 3D MPSoC design, and continue developing their technology by exploring new chemicals and structures in order to further extend their power generation efficiency.

7 CONCLUSION

In this paper, we evaluated for the first time the evolution of the on-chip cooling and power generation capabilities of FCAs, when designing highly-scaled 3D MPSoCs. As technology feature size shrinks and device leakage exponentially grows, FCA cooling allows to limit the overall power consumption of 3D MPSoCs. Our experimental results show that FCAs are able to reduce the peak temperature of a high performance MCP and a ML accelerator by up to 35°C compared to the direct liquid cooling solution of a HPC platform, without changing their dimensions or chemical nature. Furthermore, FCA power generation improves up to 2x with higher temperatures and lower operation voltages of smaller process nodes. Hence, they allow to maintain the IR-drop across the power grid of highly-scaled dies under 5%, saving over 10% TSV-reserved area and additional power delivery metal layers, with respect to scaled HPC cooling stategies. This work demonstrates the immense potential of FCA technology as a first-choice solution for thermal and power management of deeply-scaled 3D MPSoC designs.

ACKNOWLEDGMENTS

This work has been partially supported by the ERC Consolidator Grant COMPUSAPIEN (GA No. 725657), the EC H2020 WiPLASH (GA No. 863337), the NTU Startup Grant (M4082035) and the NRF AME programmatic fund titled Hardware-Software Co-optimization for Deep Learning (Project No. A1892b0026).

REFERENCES

[1] C. Auth et al. A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors. *Symposium on VLSI Technology*, 2012.
[2] 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.
[3] C. Auth et al. A 10nm high performance and low-power CMOS technology featuring 3rd generation FinFET transistors, Self-Aligned Quad Patterning, contact over active gate and cobalt local interconnects. *IEDM*, 2017.
[4] R. Xie et al. A 7nm FinFET technology featuring EUV patterning and dual strained high mobility channels. *IEDM*, 2016.
[5] IEEE 2016 International Roadmap for Devices and Systems (IRDS). https://irds.ieee.org/images/files/pdf/2016_MM.pdf.
[6] IRDS 2018. https://irds.ieee.org/images/files/pdf/2018/2018IRDS_MM.pdf.
[7] Y. Ban et al. IR-drop analysis for validating power grids and standard cell architectures in sub-10nm node designs. *Design-Process-Technology Co-optimization for Manufacturability*, 2017.
[8] A. Sridhar et al. PowerCool: Simulation of integrated microfluidic power generation in bright silicon MPSoCs. *ICCAD*, 2014.
[9] M. M. Sabry et al. Integrated Microfluidic Power Generation and Cooling for Bright Silicon MPSoCs. *DATE*, 2014.
[10] A. Andreev et al. PowerCool: Simulation of Cooling and Powering of 3D MPSoCs with Integrated Flow Cell Arrays. *IEEE Transactions on Computers (TC)*, 2018.
[11] H. Najibi et al. A Design Framework for Thermal-Aware Power Delivery Network in 3D MPSoCs with Integrated Flow Cell Arrays. *ISLPED*, 2019.
[12] A. Bartolini et al. Unveiling Eurora: Thermal and Power Characterization of the most Energy-Efficient Supercomputer in the World. *DATE*, 2014.
[13] D. Kulkarni et al. Experimental Study of Two-Phase Cooling to Enable Large-Scale System Computing Performance. *IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, 2018.
[14] E. Fluhr et al. POWER8: a 12 core server-class processor in 22nm SOI with 7.6Tb/s off-chip bandwidth. *ISSCC*, 2014.
[15] N. Jouppi et al. In-datacenter performance analysis of a tensor processing unit. *ISCA*, 2017.
[16] T. Tang et al. MLPAT: A power area timing modeling framework for machine learning accelerators. *DOSSA Workshop*, 2018.
[17] D. Christen et al. Energy Efficient Heat Sink Design: Natural Versus Forced Convection Cooling. *IEEE Transactions on Power Electronics*, 2017.
[18] S. Wu. A highly manufacturable 28nm CMOS low power platform technology with fully functional 64Mb SRAM using dual/tripe gate oxide process. *Symposium on VLSI Technology*, 2009.
[19] S. Sinha et al. Exploring sub-20nm FinFET design with Predictive Technology Models. *DAC*, 2012.
[20] D. Cuesta et al. Adaptive Task Migration Policies for Thermal control in MPSoCs. *ISVLSI*, 2010.
[21] A. Iranfar et al. Machine Learning-Based Quality-Aware Power and Thermal Management of Multistream HEVC Encoding on Multicore Servers. *IEEE Transactions on Parallel and Distributed Systems*, 2018.
[22] A. Sridhar et al. 3D-ICE: a compact thermal model for early-stage design of liquid-cooled ICs. *TC*, 2014.