Design Optimization of 3D Multi-Processor System-on-Chip with Integrated Flow Cell Arrays

Artem Andreev[1], Fulya Kaplan[2], Marina Zapater[1], Ayse K. Coskun[2], David Atienza[1]
[1]Embedded Systems Laboratory (ESL), EPFL, Switzerland
{artem.andreev,marina.zapater,david.atienza}@epfl.ch
[2]ECE Department, Boston University, USA
{fkaplan3,acoskun}@bu.edu

ABSTRACT

Integrated flow cell array (FCA) is an emerging technology, targeting the cooling and power delivery challenges of modern 2D/3D Multi-Processor Systems-on-Chip (MPSoCs). In FCA, electrolytic solutions are pumped through microchannels etched in the silicon of the chips, removing heat from the system, while, at the same time, generating power on-chip. In this work, we explore the impact of FCA system design on various 3D architectures and propose a methodology to optimize a 3D MPSoC with integrated FCA to run a given workload in the most energy-efficient way. Our results show that an optimized configuration can save up to 50% energy with respect to sub-optimal 3D MPSoC configurations.

KEYWORDS

3D MPSoC design optimization, flow cell arrays, processor cooling

ACM Reference Format:
Artem Andreev, Fulya Kaplan, Marina Zapater, Ayse K. Coskun, David Atienza. 2018. Design Optimization of 3D Multi-Processor System-on-Chip with Integrated Flow Cell Arrays. In *ISLPED '18: ISLPED '18: International Symposium on Low Power Electronics and Design, July 23–25, 2018, Seattle, WA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3218603.3218606

1 INTRODUCTION

The growing demand for computing power imposes many challenges in the design and energy-efficient operation of current and future processors. 3D Multi-Processor Systems-on-Chip (3D MPSoCs) have been proposed (e.g., [9]) to reduce communication latency, achieve higher bandwidth, and increase the overall system performance and efficiency. However, the benefits of 3D MPSoCs are hindered by the heat removal problem (due to the difficulty of removing high heat fluxes from intermediate layers) and the complicated tradeoff between power delivery and communication bandwidth due to the limited number of Through Silicon Vias (TSVs) available in the 3D MPSoC.

This work has been partially funded by the EC H2020 MANGO project (Agreement No. 671668), the ERC Consolidator Grant COMPUSAPIEN (Agreement No. 725657), and the NSF grant #1730316.

ISLPED '18, July 23–25, 2018, Seattle, WA, USA

ACM ISBN 978-1-4503-5704-3/18/07…$15.00
https://doi.org/10.1145/3218603.3218606

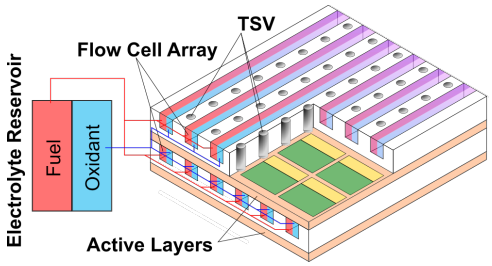

Figure 1: Concept of 3D MPSoC with integrated FCA

Integrated microfluidic reduction-oxidation ("redox") flow cell array (FCA) technology has been recently proposed to overcome these challenges in 3D MPSoCs [15]. This technology provides the ability to simultaneously remove heat via inter-layer liquid cooling and generate power on-chip through an electrochemical reaction in the liquid. In a system with integrated FCAs, microchannels are etched between the stacked layers and an electrolytic solution is pumped through the channels, as illustrated in Figure 1. While flowing through the microchannels, fuel and oxidant solutions engage in electrochemical reactions, producing electrical power, while simultaneously acting as a microfluidic cooling layer. One of the main benefits of the FCA technology is that it can be manufactured in the same way as microchannel-based liquid cooling systems [16].

Our previous work [2] has shown the benefits of FCA technology for a high-performance processor by using PowerCool [16], a compact simulator able to estimate the FCA power generation through coupled thermal and electrochemical modeling of the 3D MPSoCs with FCA. However, our previous work was limited to the study of the benefits of integrating FCAs with a 3D MPSoC for one particular architecture running a workload that fully stressed the system in terms of power consumption (with power density up to 95 W/cm^2). On the target 3D MPSoC, we demonstrated the tradeoffs between power generation, leakage and cooling [2]. However, other important factors impacting FCA power generation, such as footprint area and power density, have not been investigated. Therefore, the analysis of the capabilities of FCA technology implemented on a variety of chip architectures running realistic workloads remains an open challenge.

In this work, we optimize the design of 3D MPSoCs with integrated FCAs. To do so, we first analyze the effect of die size and power density on the FCA power generation and cooling efficiency. Then, we consider the computational performance of the studied architectures and derive a design space exploration strategy that selects the best FCA-enabled 3D MPSoC configurations in terms

of energy consumption and Quality-of-Service (QoS) for a given workload. Our main contributions are as follows:

• We analyze chip design parameters that determine FCA power generation, such as power density and chip size, and identify the tradeoffs encountered between power generation and cooling. We explore the MPSoC architectures that could benefit the most from FCA integration, showing that the generated power can range from 10% to over 100% of the overall chip power consumption. We evaluate the benefits of heterogeneous 3D MPSoCs through simulations based on several real existing chips running a mix of sequential and parallel workloads.

• We develop a low-overhead design optimization methodology that builds heterogeneous 3D MPSoCs using a pool of real chips to optimize energy consumption and QoS for a target workload. Our method combines bisection and neighborhood search and, for a real-life workload modelled as a mix of sequential and parallel jobs, it exhibits 95% lower overhead than an exhaustive search. The optimal heterogeneous 3D MPSoCs consume up to 60% less energy than homogeneous ones. Furthermore, our results show that the FCA power generation allows the optimal 3D MPSoCs to draw 40% less power from the printed circuit board (PCB).

2 RELATED WORK

2.1 MPSoC Trends and Liquid Cooling

Heterogeneous MPSoCs incorporate processing cores with diverse power/performance characteristics on the same platform to cope with the growing computational demands while maximizing energy efficiency. In particular, depending on the requirements of the workload, a power-hungry/high-performance core or a simpler/low-power core could a better choice. In single-ISA heterogeneous systems, cores have different micro-architectural designs to optimize for higher performance or lower power consumption (e.g., ARM big.LITTLE). Multiple ISA systems [17] combine computing units of different ISAs, e.g., CPUs with GPUs or CPUs with FPGAs to accelerate portions of the workload using specialized hardware units. Most of the works in this area focus on the workload distribution to maximize energy efficiency [8, 11] for a system with certain number of cores. However, they do not investigate the selection of the number and type of cores in the context of 3D MPSoCs. We believe that 3D stacking creates an opportunity for even higher levels of architectural heterogeneity as it allows integration of different process technologies onto the same chip. Thus, in this work we propose a design optimization methodology to determine the number and type of heterogeneous cores in a 3D MPSoC.

While allowing heterogeneous designs and improving on-chip communication, 3D MPSoCs bring new challenges. One of them is the increased power density and thermal resistance leading to elevated on-chip temperatures. Inter-layer liquid cooling has been proposed as an efficient and scalable solution for cooling 3D MPSoCs. Prior work in this area focused on reducing the pumping power and large thermal gradients in liquid-cooled systems through thermally aware workload allocation [4], dynamically adjusting the flow rate [3], and clustering microchannels with different flow rates [12]. Even though inter-layer liquid cooling achieves significant success in tackling the thermal issue of 3D MPSoCs, it does not address, by itself, the power delivery challenge, unlike the FCA technology.

2.2 FCA Technology

An FCA is an array of redox fuel cells, which essentially combines on-chip power generation (redox reaction) with liquid cooling (flowing electrolyte solutions). Redox fuel cells are gaining more interest as a technology for electrochemical energy storage systems [1]. Their benefits include decoupled power and energy capacity (which allows to scale them independently), long life-time, high-degree of safety [13] and energy efficiency up to 85% [1]. Therefore, the technology of FCA is not only a promising solution for cooling and powering 3D MPSoCs, but it also connects the power generation, storage and consumption at the system level.

FCA generated power depends on the electrolytes [14] and configuration of the cell [5], and is in the order of magnitude of $1W$ per cm^2 of electrode area. Prior work proposes using FCAs as an inter-layer heat sink and on-chip power source for a 3D MPSoC [16] and explores the impact of FCA parameters on the cooling performance and power generation [2]. These works report up to $3.6W/cm^2$ of on-chip power generation, however they do not consider FCA impact on the system computing performance. In our work, we consider cooling, power generation and computing performance, and propose a methodology to optimize 3D MPSoCs with FCA configuration for a given set of workloads.

3 MODELING AND EVALUATION OF MPSOCS

In this section we briefly explain our approach to evaluate the design of heterogeneous 3D MPSoCs in terms of their energy consumption and QoS when executing a certain workload.

3.1 3D MPSoC modeling

The design parameters of modern chips vary greatly depending on the purpose, application area, architecture, technology, etc. To illustrate our approach to design and optimize heterogeneous 3D MPSoCs we consider several distinct real-life chips as building blocks. We choose them to cover a wide range of power-performance characteristics, such as high-power and high-performance cores (such as those by Intel and AMD), low-power and low-performance cores (such as cores built on RISC-V), mobile-grade cores (Cortex A73) and accelerators (GPU and FPGA), as listed in Table 1. We model a 3D MPSoC as a combination of these chips, organized in a stack with a height of h layers and a footprint area of a_f cm^2. We assume that each layer can hold only one chip type, but within one layer the content of the chip (cores, caches, etc) can be replicated multiple times until reaching the area limit a_f.

3.2 Exploration of FCA performance

We use the PowerCool [2] framework to carry out steady-state thermal and electrochemical simulations for the chosen set of real-life chips (to approach physical dimensions of the larger chips for more accurate comparison, we combined Cortex A73 and Raven-3 cores into clusters of 36 and 35 cores, respectively) and a variety of synthetic 2D chips (to cover and extend the design space, see Fig. 2) assuming one FCA layer on top of a single chip. Chip length, which matches the FCA channels' length, and chip width, which defines the number of FCA channels in the chip, of the synthetic chips are varied from 1 cm to 2.5 cm in 0.5 cm steps. We assume uniform heat dissipation across the chip (FCA power generation mainly depends on the total dissipated heat, not its distribution)

		Width	Length	P_{idle}	P_{TDP}	P_{FCA}	T_{max}	LLVM (fun/s)		GEMM (GFLOPS)	
ID	Name	mm	mm	W	W	W	°C	Single thr.	Max. Perf.	S. thr.	Max. Perf.
1	Xeon E7-8894 v4	24.86	18.35	45	165	17.5	75	485.5	5200	81.5	482.5
2	Ryzen 7 1800X	22.01	8.87	14	95	9.6	71	596.7	5630	60.2	418.1
3	Intel Stratix 10	27.87	20.1	50	126	20.4	64	-	-	-	9200
4	Nvidia GP 100	27.29	22.36	35	300	20.7	72	-	-	-	10300
5	Cortex A73 (x36)	11.35	10.15	5.4	48	5.9	67	194.4	5249	10.7	338.4
6	RISC-V Raven 3 (x35)	9.1	9.0	0.2	3.7	3.9	55	55.5	1458	2.5	66.2

Table 1: Parameters, PowerCool simulation results and performance values for selected real-life chips

and vary power density (PD) is from 5 to 50 W/cm^2. This results in a linear temperature distribution along the channels and a uniform distribution across them. Therefore in this paper we report only maximum chip temperature T_{max}, but not the thermal maps.

PowerCool uses a model for leakage power (P_l) that is exponentially affected by chip temperature ($P_l(T) = P_{TDP}(a + be^{\kappa(T-T_{ref})})$, where P_{TDP} is the chip's Thermal Design Power) and driven by the manufacturing technology. We set $a = 0.1$ and $\kappa = 0.013$, based on current technological trends [10]. To analyze the impact of $P_l(T)$ on the efficiency of the FCA, we vary b (0.1, 0.2 and 0.3) for each combination of chip size and PD parameters, which results in leakage power ranging from 15 to 50% of P_{TDP}.

We set the FCA design parameters to their optimum for power generation values (100 μm height and 50 μm width) [2] and then, for each combination of chip parameters (size, PD, leakage), we swipe FCA control parameters to find the maximum power the FCA can generate for each chip [2]: the coolant flow velocity from 0.2 m/s to 2.5 m/s and the inlet temperature from 27 °C to 60 °C.

To assess the FCA performance we use both the maximum generated power (P_{FCA}) and $\eta_{FCA}(\%)$, namely, the ratio of P_{FCA} over the total power consumption (dynamic power of the chip P_{dyn} and leakage power $P_l(T)$):

$$\eta_{FCA}(\%) = \frac{P_{FCA} \times 100}{P_{dyn} + P_l(T)} \quad (1)$$

Fig. 3 compares the simulation results for the synthetic chips that result in $T_{max} \leq 100$ °C with the results for the real-life chips (only the cases of maximum P_{FCA} are shown, the markers correspond to the legend of Fig. 2). P_{FCA} and T_{max} for real chips are listed in Table 1. In case of Raven-3, P_{FCA} corresponds to 105% of the chip P_{TDP}, thus allowing to fully power it up. However, as PD increases, $\eta_{FCA}(\%)$ decreases. Fig. 3 illustrates that FCA is able to generate up to 96% of the power consumption for the chips with an average PD of 5 W/cm^2, but for higher PD, $\eta_{FCA}(\%)$ drops to around 10%, even though the absolute value of P_{FCA} may increase. For a fixed PD, smaller chips result in lower temperatures and higher $\eta_{FCA}(\%)$ because the FCA performance per unit of area decreases with FCA length. With the flow velocity set to the maximum (2.5 m/s), we ensure sufficient cooling to keep the chips within a safe temperature range ($\leq$ 80°C), even for the highest inlet temperatures (i.e, the ones that result in the highest P_{FCA}).

For the purpose of this paper, we aim at maximizing overall P_{FCA} to decrease the load on the power distribution network and enable feasibility of 3D MSoCs with higher power densities. Therefore, we set the FCA control parameters to maximum flow velocity and inlet temperature.

The simulated T_{max} and P_{FCA} for the individual chips allow us to estimate the maximum temperature and the total amount of

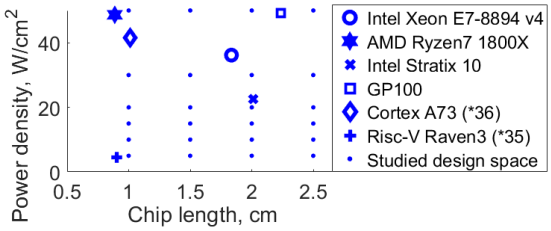

Figure 2: Design parameters of the real chips compared to the synthetic design space (shown by dots).

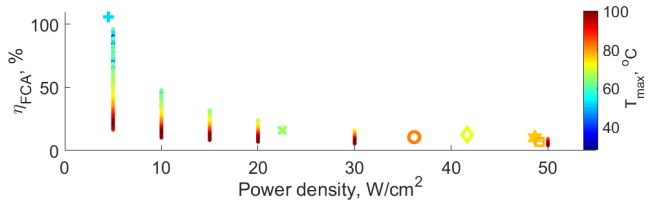

Figure 3: η_{FCA} versus PD. The color bar shows the maximum chip temperature. Multiple dots related to a single heat flux value correspond to different FCA and chip parameters.

FCA power generation for a heterogeneous 3D MPSoC constructed from the studied chips. However, to evaluate these MPSoCs while engaging the performance aspect, we need to estimate their energy consumption and QoS when running a given workload. The next subsection explains our approach to model different workloads.

3.3 Workload Generation and Execution

We represent a workload W as a set of N individual jobs J_i, where jobs are characterized with 4 parameters: application type A, arrival time t_{arr}, complexity C_A (i.e., amount of operations required to execute a job) and execution time limit t_{ex_lim} as follows:

$$W = \{J_i = (A^i, t^i_{arr}, C^i_A, t^i_{ex\_lim}), i = 1..N\} \quad (2)$$

We define the baseline workload parameters (i.e., total number of jobs N and the duration of the workload τ) based on a real-life workload from the Gaia cluster [6], which consists of around $N = 52000$ of jobs and covers a period of time of $\tau = 3$ months.

Application Types: We represent the variation of performance/power of real-life applications as a combination of a fully sequential and a fully parallel workload. We assume that sequential jobs can be assigned only to a single core, whereas parallel jobs can be executed in parallel on as many cores as available on a chip. Accelerators (i.e., GPUs, FPGAs) are used to run only parallel jobs.

Performance data for each workload type is acquired from real benchmarks (as shown later in Section 5.1). We assume that the maximum performance corresponds to the chip P_{TDP}, and zero performance corresponds to P_{idle}, and interpolate linearly the rest.

Type	Arrival rate
Uniform	$\lambda_i(t) = N_i/\tau$
Sinusoidal	$\lambda_i(t) = (1 + sin(2\pi t/\tau))N_i/\tau,\ t \in [0, \tau]$
Real-life	$\lambda_i^{Gaia}(t)$ derived from the log

Table 2: Workload types

Per-core performance is decreased by a fixed percentage as the number of active cores increases to model resource contention. This percentage is found by fitting single-core and maximum chip throughput values for each application type.

Arrival Time: We model the jobs' arrival as a Poisson process, where the arrival rate $\lambda_i(t)$ depends on the application type A^i. We generate 3 types of workloads, listed in Table 2: (i) uniform, (ii) sinusoidal (as an example of a periodic workload) and (iii) real-life example, based on the arrival rates from the Gaia cluster log [6].

We generate a set of workloads with different **complexity** and **execution time limits** for each of the workload types (uniform, sinusoidal, real-life). As a baseline, we use the job parameters of the Gaia log and vary the complexity and/or execution time of different application types in the range from 0.5 to 1.5 of the baseline values.

We developed a greedy policy to allocate a workload to a heterogeneous 3D MPSoC. At each job arrival, the algorithm searches the 3D MPSoC for the most energy efficient core(s) to execute this job within its execution time limit t_{ex_lim} without violating the execution of the previously allocated jobs. If no cores can provide the necessary performance at the time of the job arrival, the job is dropped. We focus on core performance and do not consider memory accesses, assuming that the 3D MSoC has enough memory and bandwidth not to limit the performance of any component.

When all jobs from a chosen workload are either finished or dropped, we calculate the utilization $U(t)$ of every unit in the 3D MPSoC, and compute its power consumption as follows:

$$P(t) = P_{idle} + U(t) \cdot (P_{TDP} - P_{idle}) - P_{FCA} \tag{3}$$

We can get the energy consumption of a 3D MPSoC by integrating $P(t)$ over the full workload run time for every unit of the 3D MPSoC. We define the QoS for a given workload as the ratio of the total number of jobs to the number of dropped jobs and we use it jointly with the energy consumption to rank the 3D MPSoCs.

4 OPTIMIZATION OF 3D MPSOCS WITH FCA

In this section, we describe our approach to select the best heterogeneous 3D MPSoCs to execute a target workload with the optimum energy consumption and QoS.

Given the CPU-intensive nature of the workloads considered in this work, we assume that there is enough bandwidth in a 3D MPSoC to not limit its performance, so the communication between the different layers and the memory is not hindered regardless of the order of the layers. Moreover, each FCA layer is capable of cooling down the corresponding active layer even under full utilization, so the maximum temperature in a 3D MPSoC will always stay in the safe zone below 80°C (cf. Section 3.2). Therefore, different permutations of the layers will not affect the evaluation score, and we can consider only different combinations, so the full design space of 3D MPSoCs with the maximum height of h layers contains $z = \sum_{i=1}^{h} \binom{d+i-1}{i}$ elements, where d is the number of different chip types. Then, we can represent a certain 3D MPSoC configuration

Input: $R = \{\vec{v}_i,\ i \in \{CPUs\} \cup \{n(\vec{v}_j),\ j \in \{Accelerators\}\}$– starting points (homogeneous 3D MPSoCs);
$G(R)$ – evaluation of the goal function on the starting set
Output: R, G(R) – set of simulated 3D MPSoCs and their rating

```
Find middle point(s) of R: M = m(b_d(R), ρ);
Find neighbours of the best-so-far configuration(s): N = n(b_k(R));
Choose elements of M that are not in R: X = M \ R = {x ∈ M : x ∉ R};
if X is empty (X = ∅) then
    Choose elements of N that are not in R: X = N \ R = {x ∈ N : x ∉ R};
end
while X is not empty (X ≠ ∅) do
    for each x ∈ X do
        Allocate workload W to x;
        Evaluate G(x);
        R := {R, x};
    end
    M = m(b_d(R), ρ);
    N = n(b_k(R));
    X = M \ R;
    if X is empty (X = ∅) then
        X = N \ R;
    end
end
```

Algorithm 1: Structure of the proposed algorithm for optimizing 3D MPSoC configuration for a given workload

with a d-dimensional vector $x \in \mathbb{N}_0^d$, where each coordinate represents the number of layers of the correspondent chip type. The sum of the coordinates cannot exceed the maximum height: $||x||_1 \leq h$.

To rank different 3D MPSoCs, we define the goal function $G(x)$ as a linear combination of the drop ratio $dr(x)$ (inverse of QoS) for a given workload and the power score with a weight α:

$$\begin{aligned} &G(x) = \alpha \cdot dr(x) + (1-\alpha) \cdot P_{score}(x), \\ &dr(x) = 1/QoS(x) \in [0, 1], \\ &P_{score}(x) = \frac{P_{average}(x)}{max(P_{TDP}(x))} \in \left[\frac{min(P_{idle}(x))}{max(P_{TDP}(x))}, 1\right], \end{aligned} \tag{4}$$

where $P_{score}(x)$ is the 3D MPSoC's average power consumption executing a specific workload normalized to the highest P_{TDP} among all possible 3D MPSoC configurations. We aim to minimize $G(x)$.

We consider only a subset of 3D MPSoCs with the maximum height ($||x||_1 = h$), since it includes all smaller configurations. Nonetheless, the reduced design space has $z = \binom{d+h-1}{h}$ elements (upper bounded by $O(d^h)$) and is too extensive to be searched exhaustively for a large set of workloads and chip types. Thus, we propose a fast design space exploration algorithm to reduce the overall number of simulations required to find an optimal 3D MPSoC for a given set of workloads. Hence, we define the following notations and functions:

$$\begin{aligned} &\vec{v}_i = \{\vec{x} \in \mathbb{N}_0^d : x_i = h,\ x_{j\neq i} = 0\}, \\ &S = \{\vec{x} \in \mathbb{N}_0^d : ||\vec{x}||_1 = h\} \setminus \{\vec{v}_i,\ i \in accelerators\} \\ &m(X, \rho) = \{\vec{x} \in S : ||\vec{x} - \frac{h * \sum_i \vec{x}^i}{||\sum_i \vec{x}^i||_1}||_1 \leq \rho,\ \vec{x}^i \in X\}, \\ &n(X) = \bigcup_i \{\vec{y} \in S : ||\vec{x}^i - \vec{y}||_1 \leq 2,\ \vec{x}^i \in X\}, \\ &b_k(X) = \{X_b \subset X : \forall \vec{x}^i \in X_b, \forall \vec{x}^j \in X \setminus X_b \Longrightarrow G(\vec{x}^i) < G(\vec{x}^j)\} \end{aligned} \tag{5}$$

where v_i are homogeneous configurations (h layers of the same chip type); S is the discrete design space of 3D MPSoC configurations (we exclude from it homogeneous accelerators incapable of executing the sequential applications); $m(X, \rho)$ is a function which returns elements of S which are not farther than ρ from the middle point of the set X; $n(X)$ is a function that returns the closest

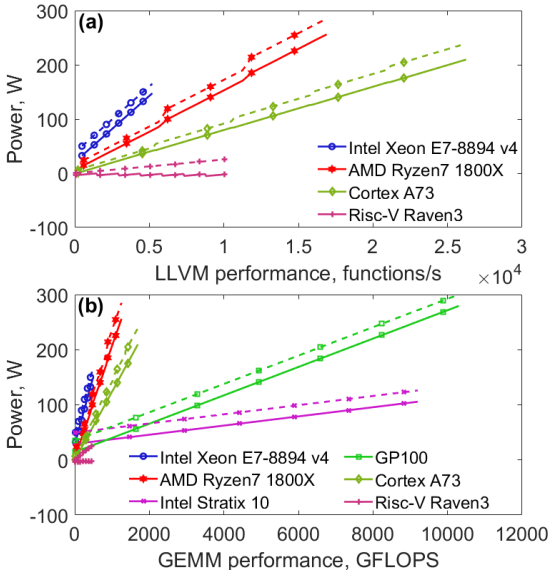

Figure 4: LLVM (a) and GEMM (b) energy efficiency functions for layers of different chip types, with (solid lines) and without (dashed lines) considering FCA power. $a_f = 6.1\ cm^2$.

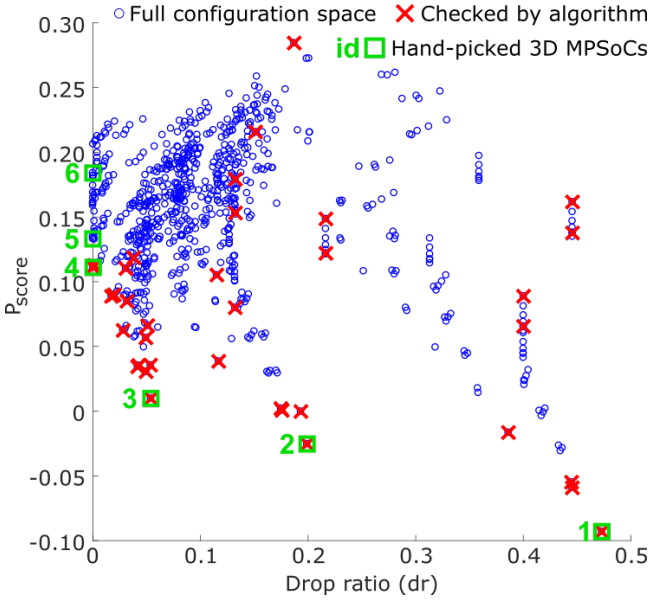

Figure 5: Results of our design space search algorithm (36 points) vs. the full space (784 points), $\alpha = 0.25$**. Square markers with numbers correspond to "3D MPSoCs id" on Fig. 6.**

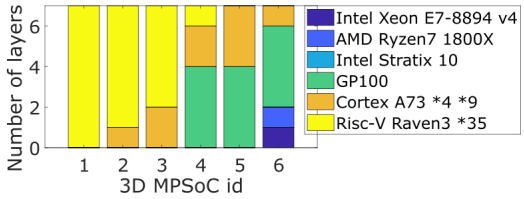

Figure 6: Configuration of chosen 3D MPSoCs. $a_f = 6.1\ cm^2$.

neighbours from S of the elements of input set X; and $b_k(X)$ is the subset of k best configurations from X with lowest values of $G(x)$.

Despite the fact that the design space is represented as a limited set of integer vectors, minimizing $G(x)$ is substantially different from an ILP problem, due to non-linear and rather unpredictable behaviour of $G(x)$. To find the best 3D MPSoC configurations we use a combination of the bisection method and neighbourhood search (cf. Algorithm 1). We start by evaluating $G(x)$ for the set of corner-cases R. Next, we apply the bisection to find the average between the best configurations (procedure on lines 1 and 13), which we repeat in a loop (lines 7-19) while the bisection returns new configurations. If no new configuration is returned by bisection (lines 5 and 17), we switch to evaluating the neighbours of the best-so-far configuration (lines 2 and 14) and continue the loop. The loop ends when neither the bisection nor the neighbourhood search can improve the best achieved score, indicating that the algorithm has converged to a minimum of $G(R)$. The parameters ρ and k in $m(R, \rho)$ and $b_k(R)$ can be tuned to vary the coverage of the configuration space S.

Since there is always a tradeoff between energy consumption and QoS, the configurations obtained with the algorithm for a fixed weight α may not cover the full Pareto front in the $P_{score} - dr$ plane, and several simulations with different weights may be required to better explore optimal design options.

5 EXPERIMENTAL SETUP AND RESULTS

5.1 Experimental setup

We illustrate our method by modelling 3D MPSoC as a combination of the $d = 6$ chip types listed in Table 1, with the height limit of $h = 7$ layers (in this case, $z = \binom{d+6}{7} = O(d^{5.8})$) and the footprint area $a_f = 6.1\ cm^2$, which is the area of the biggest chip in the set (Nvidia GP100). For example, $x = [0, 0, 0, 0, 4, 3]$ represents a 3D MPSoC made of 4 layers of Cortex A73 content (45 times the content of quad-core chip per layer) and 3 layers of Raven 3 content (245 times the content of a single-core chip per layer). As covered in Section 3.2, each individual chip can be safely cooled down under full load by the corresponding FCA layer, therefore any designed 3D MPSoC will also be able to operate safely under full load.

As typical examples of sequential and parallel workloads we choose two benchmarks of the Geekbench 4 suite [7]: LLVM and General Matrix Multiplication (GEMM). However, our approach can be applied to any other set of application types. In Table 1, in addition to the characteristics of our chosen set of chips, we indicate the performance values obtained from the chosen benchmarks.

Figure 4 shows the dependencies of LLVM and GEMM performance on power consumption for the layers (limited by a_f) of different chip types. Dashed lines correspond to power consumption without FCA, and solid lines – with FCA, and negative power values indicate the cases when P_{FCA} is higher than the power consumption of the corresponding active layer. Using FCA power lowers the lines and decreases their slope, changing the intersection points, i.e., the points where the transition from one architecture to another improves the system energy efficiency.

5.2 Workload-Specific 3D MPSoC design optimization

To evaluate the results of our algorithm we compared it to the full configuration space analysis for a real-life workload. The full space contains 784 elements, but our algorithm (see Algorithm 1)

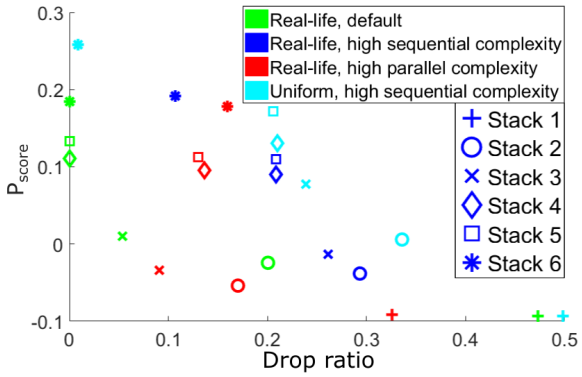

Figure 7: Performance of the chosen 3D MPSoCs.

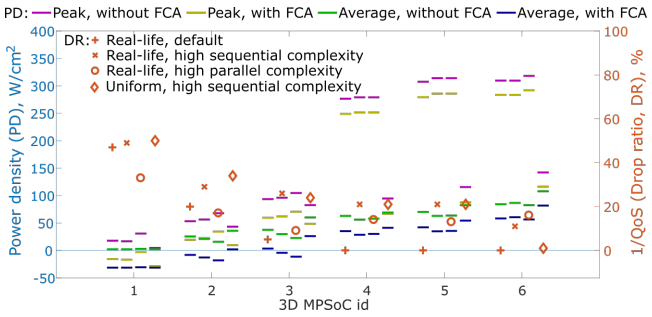

Figure 8: DR and PD at the PCB level of each 3D MPSoC (average and peak values) while running different workloads.

checks from 20 to 116 configurations to converge, depending on the weight $\alpha \in [0, 1]$ in the goal function $G(x)$ (Eq. 4), reaching the global minimum of $G(S)$ for each tested value of α. In Fig. 5 we show the comparison between the full space exploration and results of our algorithm for $\alpha = 0.25$. In this case, the algorithm fully covers the Pareto front on the $P_{score} - dr$ plane and checks only 36 out of 784 configurations, saving 95% of time. This analysis also shows that among the 3D MPSoC configurations capable of providing no QoS degradation ($dr \approx 0$), the best one (3D MPSoC 4 on Fig. 5) saves 50% of energy compared to the least energy efficient. Homogeneous 3D MPSoCs consisting of 5-7 layers of Intel/AMD processors and 0-2 layers of accelerators consume 2-2.5 times more energy than 3D MPSoC 4 while showing 15-20% of drop ratio.

We choose 6 3D MPSoC configurations (Fig. 5 and 6) from the Pareto front for the real-life workload, and explore their energy-QoS tradeoffs under other workloads. In particular, Fig. 7 shows how for the default real-life workload (green), 3D MPSoCs 4, 5 and 6 provide excellent QoS with no dropped jobs, 3D MPSoC 3 provides acceptable QoS with 5% of dropped jobs, while 3D MPSoCs 1 and 2 show unacceptable QoS with more than 20% of dropped jobs. However, with more complex parallel jobs (red) only 3D MPSoC 3 shows less than 10% of dropped jobs, and with more complex sequential jobs (dark blue and cyan) 3D MPSoC 6 shows superior performance. Therefore, there is a need for workload-specific optimization.

5.3 FCA impact on 3D MPSoC design space

Next, we analyze the FCA impact on 3D MPSoC power consumption. Fig. 8 shows the PD values at the PCB level with and without FCA power generation and drop ratio for the chosen 3D MPSoCs (Fig. 6) for different workloads. Using FCA power decreases PD by 25-28 W/cm^2 (corresponds to 44% of average PD of 3D MPSoC 4 during execution of real-life workload), potentially enabling configurations, that would not be feasible to power up otherwise. For 3D MPSoCs 4-6, peak PD values significantly differ between the cases of real-life and uniform workloads, which shows that feasibility of a certain 3D MPSoC with FCA technology cannot be discussed out of context of a target workload.

6 CONCLUSION

In this work we explored FCA power and cooling performance for various architectures, and showed that FCAs satisfy cooling requirements of modern chips and generate from 10% of the overall power consumption for high-power chips, up to 105% for low-power chips such as RISC-V. We developed a fast design space exploration algorithm to find the most energy efficient 3D MPSoC configurations to successfully run a given workload, which allows to save up to 95% of CPU time compared to the exhaustive search. For a real-life workload, the best 3D MPSoC found by the algorithm executes the workload with zero dropped jobs while consuming only 40-50% of energy compared to homogeneous configurations, which show 15-20% of dropped jobs. Furthermore, we demonstrated that the optimal 3D MPSoCs are workload-specific. We have shown that integrated FCA can reduce the supply power density by 28 W/cm^2 (up to 40% of the power requirements of the studied optimal 3D MPSoCs), potentially enabling new 3D MPSoC configurations.

REFERENCES

[1] Piergiorgio Alotto, Massimo Guarnieri, and Federico Moro. 2014. Redox flow batteries for the storage of renewable energy: A review. *Renewable and Sustainable Energy Reviews* 29 (2014), 325–335.
[2] Artem Andreev et al. 2017. PowerCool: Simulation of Cooling and Powering of 3D MPSoCs with Integrated Flow Cell Arrays. *IEEE Trans. on Comp.* (2017).
[3] Ayse K Coskun, David Atienza, Tajana Simunic Rosing, Thomas Brunschwiler, and Bruno Michel. 2010. Energy-efficient variable-flow liquid cooling in 3D stacked architectures. In *DATE*. 111–116.
[4] Ayse K. Coskun, Jose L. Ayala, David Atienza, and Tajana Simunic Rosing. 2009. Modeling and Dynamic Management of 3D Multicore Systems with Liquid Cooling. In *IFIP/IEEE VLSI-SoC*. 60–65.
[5] Robert M. Darling and Mike L. Perry. 2014. The Influence of electrode and channel configurations on flow battery performance. *J. of the Electrochem. Soc.* (2014).
[6] J Emeras, SÃlbastien Varrette, Mateusz Guzek, and Pascal Bouvry. 2015. Evalix: Classification and Prediction of Job Resource Consumption on HPC Platforms. In *JSSPP*.
[7] GB4 2017. Geekbench 4. (2017). https://www.geekbench.com/
[8] R. Kumar, K. I. Farkas, N. P. Jouppi, P. Ranganathan, and D. M. Tullsen. 2003. Single-ISA Heterogeneous Multi-Core Architectures: The Potential for Processor Power Reduction. In *IEEE/ACM MICRO 36*. 81–92.
[9] J. H. Lau. 2010. Evolution and outlook of TSV and 3D IC/Si integration. *EPTC*.
[10] S. G. Narendra. 2005. Leakage in Nanometer CMOS Technologies. Springer, NY.
[11] Indrani Paul et al. 2013. Coordinated Energy Management in Heterogeneous Processors. In *Int. Conf. on HPC, Networking, Storage and Analysis (SC '13)*.
[12] Hanhua Qian et al. 2013. An efficient channel clustering and flow rate allocation algorithm for non-uniform microfluidic cooling of 3D integrated circuits. *Integration, the VLSI Journal* 46, 1 (2013).
[13] David Reed et al. 2016. Stack Developments in a kW Class All Vanadium Mixed Acid Redox Flow Battery at the Pacific Northwest National Laboratory. *Journal of The Electrochemical Society* 163, 1 (2016), A5211–A5219.
[14] Sarah Roe, Chris Menictas, and Maria Skyllas-Kazacos. 2016. A High Energy Density Vanadium Redox Flow Battery with 3 M Vanadium Electrolyte. *Journal of The Electrochemical Society* 163, 1 (2016), A5023–A5028.
[15] P. Ruch et al. 2011. Toward five-dimensional scaling: How density improves efficiency in future computers. *IBM Journal of Research and Development* (2011).
[16] A. Sridhar et al. 2014. PowerCool: Simulation of integrated microfluidic power generation in bright silicon MPSoCs. In *IEEE/ACM ICCAD*. 527–534.
[17] A. Venkat and D. M. Tullsen. 2014. Harnessing ISA Diversity: Design of a heterogeneous-ISA Chip Multiprocessor. *SIGARCH Comp. Arch. News* 42 (2014).