Power Density Aware Application Mapping in Mesh-Based Network-on-Chip Architecture: An Evolutionary Multi-Objective Approach

Nizar Dahir, Ammar Karkar, Maurizio Palesi, Terrence Mak, Alex Yakovlev
*

Abstract

In the era of many-core chips, the problem of power density is a serious challenge. This is particularly important in Network-on-Chip (NoC)-based systems, where application mapping determines the resulting power patterns and the workload distribution across the entire chip. Despite this fact, the majority of mapping algorithms focus on performance, and the resulting power patterns are largely ignored. This work investigates this problem. Three different power pattern metrics with different scopes are defined, namely, power peak, power range, and regional power density. The results of using them as mapping objectives together with communication cost using a multi-objective evolutionary mapping approach are investigated. Results show that employing power patterns results-in Pareto fronts with different power patterns and features. Results are analysed and discussed. Moreover, a case study of thermal analysis of the resulting power patterns is performed. Results show that using communication cost only results-in large hotspots which translates into higher peak and range of chip temperatures. The proposed mapping objectives are shown to significantly improve thermal balancing (up to 55%) and peak temperature (up to 7.77%). These results indicate the importance of considering power patterns in the design of NoC-based many-core systems and their direct impact on the reliability and performance of such systems.

Keywords: Networks-on-Chip, Many-core systems, Evolutionary Algorithms, NSGA-II, Power Density, Thermal Analysis.

1 Introduction

The rapid shrinking of feature size in VLSI systems and the unprecedented increase in integration density lead to the emergence of Systems-on-Chip (SoCs) and many-core processing chips with tens, or even hundreds, of processing cores [38] [46] [32]. Performance and reliability of such systems is mainly determined by the performance of communication subsystem used to interconnect on-chip processors and other Intellectual Property (IP) components. As a result, network-on-chip (NoC) was proposed as a communication medium to address the major challenges facing on-chip communication in these large systems. NoCs provide reliability and high-performance in addition to enabling modularity and scalability of SoC design in general and many-core systems in particular [4] [14].

Moreover, the increasing integration capacity is

*

Nizar Dahir College of Information Engineering, Al-Nahrain University, Iraq, Email: nizar.dahir@coie-nahrain.edu.iq

Ammar Karkar University of Kufa, Iraq, Email: ammar.karkar@uokufa.edu.iq.

Maurizio Palesi is with the University of Catania, Italy, Email: maurizio.palesi@dieei.unict.it.

Terrence Mak is with the University of Southampton, UK, Email: t.mak@soton.ac.uk.

Alex Yakovlev is with the University of Newcastle, UK, Email: alex.yakovlev@newcastle.ac.uk.

facing a major challenge. The higher density of transistors provides more functionality per unit area on one hand and higher power density (power consumed per unit area) on the other [30]. Higher power density can lead to increased temperatures and deteriorates power integrity which in turn increases transient and permanent faults and decreases reliability [22, 9, 33].

One of the classical design problems in NoC architectures that has been previously addressed by many researchers is the mapping of application tasks to processing core and other IPs in a chip [29, 2, 3, 21]. However, many of such studies focus mainly on performance and energy consumption and a little attention was given in literature to the alerting challenge of power density. This is despite the fact that task allocation plays a vital role in determining the power distribution profile across the chip [13].

Power consumption patterns are characterised by many factors. Static and dynamic power across the chip are determined by CMOS technology, IC design and layout. In addition to the switching workload. In NoC-based SoCs this workload can be classified into *communicational* and *computational.* These two types of workloads have very different characteristics and must be considered separately [13, 10, 19]. Although the computational workload of an IP core is mainly determined be the task running on that core, the communicational workload of the router attached to that core is affected by inter-core communication across the entire system. These facts imply that both workloads need to be considered when application mapping is performed, especially when this mapping aims at optimising the power density pattern across the system. These facts imply that both workloads need to be considered when application mapping is performed, specially when this mapping aims at optimizing the power density pattern across the system. This relationship is particularly important with communication-intensive NoC-based many-core systems. Neuromorphic computing systems such as SpiNNaker and partially-ordered event-triggered systems (POETS) are examples of communication-intensive computational systems in which computation is completely modulated by communication [6, 32].

In this work, we address the problem of task-mapping in NoC-based architecture, using power density patterns as objectives. We define three main power density metrics and use these metrics as fitnesses, with communication cost, in a GA-based evolutionary multi-objective mapping. The effect of these metrics are on power consumption profiles and the resulting thermal profiles across the chip die is analysed. The results of different mapping strategies are discussed and important conclusions are made. The major contributions of this paper can be summarised as follows:

- A new mapping strategy for mesh-based NoC architectures is proposed. In contrast with similar works that focuses mainly on energy and communication cost, the new strategy focuses on the resulting power patterns of task workloads in the system.

- We define three different power pattern metrics a *local* one (power peak), a *global* one (power range) and a *regional* one (regional power density) and employ them in a multi-objective evolutionary mapping algorithm.

- The resulting mappings are investigated and analysed in terms of the power pattern profile characteristics and their impact on communication cost.

- A case study of thermal analysis of the resulting power patterns is performed and the proposed metrics are compared in terms of thermal distribution.

The remainder of this paper is organised as follows; Section 2 gives a survey of the related work, Section 4 describes the proposed mapping objective and the GA-based multi-objective mapping algorithm and Section 6 presents and discusses the experimental results. Finally, Section 7 concludes the paper.

2 Related Work

The concept of power density in VLSI circuits was first introduced by Najim *et al.* [30]. The authors define transition density to be the *average switching* *rate* and develop a technique for computing average switching rate based on stochastic models of logic signals. Results show that regions with higher switching density can be thermal hotspots and could adversely affect power supply integrity. Other studies show that higher power density would have negative effect on circuit reliability in terms of electro-migration failures [24]. This effect can be seen as another problem with higher power density and motivates balanced switching activity design in VLSI circuits. In [10] the concept of harmonic mapping is first introduced. The authors define a new application mapping metric called repulsive force to increase workload balancing and improve power supply integrity.

On the other hand, many studies have been published on application mapping in NoCs. Mapping tasks in many-cores system is a major design problem which is true mathematically because it is an NP-hard problem [17]. Moreover, the changing objectives, under-layer architecture, and technology constraints increase the complexity of the mapping problem. The majority of previous mapping techniques aimed at the minimisation of communication latency, reducing power or increasing fault tolerance [36, 39, 5]. These mappings target either fixed NoC topology, such as mesh, or a custom topology based on application communicational requirements.

Mapping algorithms can be classified as either *static*, where tasks are mapped at design time, or *dynamic*, where mapping takes place at run-time. The dynamic mapping is more flexible and usually results in better performance especially in the case where application characteristics are unknown at design time [8, 31]. Moreover, it enables tasks to migrate between cores based on the dynamics of the system (e.g faults or overheating) [49]. However, the computation power required for such algorithms and the delay of getting the mapping results may significantly reduce the expected advantages [39]. By contrast, static mapping is performed once at the design time, alleviating the complexity of dynamic mapping and makes it more common in the literature.

Static mapping can be classified into *exact* and *search-based* mapping. To achieve good mapping results, exact mapping uses exact solution procedures such as linear programming to determine the optimal assignment of the tasks to cores [41]. However, in these algorithms, latency increases exponentially with the number of tasks. Srinivasan *et al.* used clustering to overcome this complexity by a divide and conquer method where the graph is clustered and the best mapping solution is determined for each cluster then the solutions are combined into a final mapping [41].

Search-based mapping can be classified into *deterministic* and *heuristic-based* search. Deterministic search is also known as the systematic search. One example of deterministic search is branch and bound, in which the search for a mapping solution is done through branching into candidate solutions, and the branch is evaluated and discarded if it falls outside the bounds of the optimal solution. However, even with pruning techniques to reduce search time, the delay continues to increase exponentially with the number of cores [36]. Furthermore, tree-based search is used to find a routing and mapping solution with improved performance and reduced power consumption [21].

Heuristic search is the most commonly used mapping technique in the literature due to its lower delay compared to other methods. Researchers have used heuristic algorithms for dynamic mapping [8]. However, heuristic algorithms are more convenient for static mapping. Heuristic search can be divided into transformative and constrictive. In case of constrictive heuristic algorithms, the tasks are mapped in succession based on the mapping objective, and some cores may get re-mapped in later stages to effectively satisfy these objectives. In other words, constrictive heuristic search is a step by step mapping process that allows some iterations at later stages for more improvements. As a result, this class of heuristic search targets local optimum solutions [45, 34, 16].

Transformative heuristic algorithms try to avoid local optimum and aim at global optimum making them easy to adapt to any mapping objective such as improving performance and power consumption. Examples are meta-heuristic algorithms such as Ant Colony Optimization (ACO), Genetic Algorithms (GA), Artificial Bee Colony (ABC) and Particle Swarm Optimization (PSO). However, the main disadvantage of such algorithms is the execution time, which is random and relatively high compared with constrictive heuristic search algorithms. Examples of such algorithms are A3MAP-GA and FGMAP which are proposed to improve the performance and power of NoC-based homogeneous and heterogeneous many-core systems [23, 42]. Furthermore, PSO is used by Sahu *et al.* to optimize communication cost [35].

Some application mapping studies focused on the alerting problem of thermal optimization of NoC-based many-core systems [28, 27, 47, 26]. For example, fuzzy logic is used for mapping by Mosayyebzadeh *et al.* where the authors proposed using fuzzy-based optimization to minimise power and temperature in 3D-NoC systems [28]. They consider that heat dissipation of cores based on their locality and use fuzzy-based algorithm to prioritize the improvement of heat dissipation of cores considering communication and power requirements in addition to distance to hotspots. The proposed technique reduces temperature while improving power consumption and latency.

Liu *et al.* proposed a thermal-aware task mapping for reconfigurable NoC (SMART) [27]. The proposed NoC bypasses multi-hops between active cores enabling scattering of active cores and ensuring improved thermal distribution with lower communication latency overhead. In Wang *et al.*, another thermal mapping is proposed along with thermal modeling [47]. On the bases of the thermal model, a greedy algorithm maps tasks to the core with minimal thermal impact. The proposed schemes comprise three main parameters, namely, algorithm complexity, performance and thermal impact. Similarly, Li *et al.* presents an algorithm that compromises performance and thermal impact [26] in 3D NoCs. Mapping is done dynamically at runtime by defragmenting block of application tasks to the corners of the 3D many-core systems, always preserving the middle for the next application. This technique is claimed to improve thermal dissipation; at the same time, new application tasks will be allocated adjacent to each other reducing possible communication overhead. However, results show improvement of the runtime but not the thermal impact.

Many of these studies focus mainly on performance and energy consumption or address the thermal impact of mapping. Little attention is given in literature to power distribution patterns and power density, which are the direct causes of temperature and power supply noise. Even with the few cases where this metric was used, the work mainly focused on computational workload and ignored communicational workloads and their effect on the power dynamics across the system [43]. In this work, we address the problem of task-mapping in NoC-based architectures using the resulting power patterns as a design objectives. Three main power density metrics are defined and used, with communication cost, in a GA-based evolutionary multi-objective mapping algorithm. Results are investigated and analysed in terms of the power pattern characteristics and impact on communication cost. Furthermore, thermal analysis of the resulting power patterns is presented as one possible advantage of the proposed power density optimisation.

It is worth mentioning here that thermal optimisation is not the main focus of this work. We believe that power density optimisation has many other benefits such as improving power supply integrity, reducing permanent and transient faults and improving reliability, in addition to improving thermal integrity. However, we choose to show the thermal advantages of the proposed new mapping objective as an example and due to the importance of this problem.

3 Background

In this section, the related background concepts are described and notations used in this paper are defined.

3.1 Application Mapping in NoCs

The NoC architecture can be defined using a directed graph called *Architectural Graph* $ARG = G(C, P)$. A vertex, $t_i \in T$, represents an NoC Intellectual property core (IP) and each directed arc $p_{i,j} \in P$, represents the path from core i to core j. Each path $p_{i,j}$ consists of a set of links $L(p_{i,j})$. The set $L(p_{i,j})$ is given by the routing algorithm adopted by the NoC system. In this paper, we assume a homogeneous

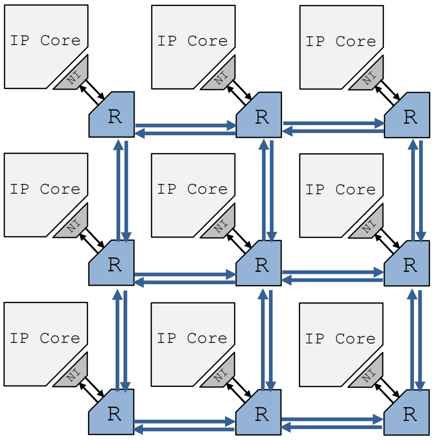


Figure 1: The homogeneous NoC-based mesh architecture considered in this work.

mesh-based NoC architecture, as illustrated in Fig. 1.

Applications are described as a directed acyclic graph called *Application Graph* ($APG = G(S, A)$), in which vertices, $s_i \in S$, represent tasks and arcs, $a_{i,j} \in A$, represent the communication from task s_i to task s_j. The arc weights represent the communication bandwidth requirements from task s_i to s_j, $bw(a_{i,j})$ and the total data volume $v(a_{i,j})$.

The mapping problem can be stated as follows:

Given an application graph, APG and an architectural graph, ARG, find a mapping function that maps APG to ARG as follows:

$$\Omega : APG(S, A) \rightarrow ARG(C, P), \qquad (1)$$

where each task, s_i, in APG is mapped to a core($c_i = \Omega(s_i)$) in the ARG and each arc, $a_{i,j}$ in the APG is mapped to a path, $p_{i,j}$, in the ARG. The mapping function Ω should satisfy a specific optimisation metric such as minimising energy cost [20] or maximising performance [29] and can also satisfy multiple objectives [2].

3.2 Communication Cost

Communication cost, C_{cost}, of a mapping function is used together with power pattern metrics to evaluate mappings in this work. C_{cost} for a given mapping Ω can be defined as follows:

$$C_{cost}(\Omega) = \sum_{a_{i,j} \in A} v(a_{i,j}) \times |L(P_{\Omega(i),\Omega(j)})|, \qquad (2)$$

where $|L(P_{i,j})|$ is the length of the path $P_{i,j}$ (number of links from core i to j) and $\Omega(i)$ is the core to which task i is assigned. The routing path (and the resulting number of hops) from source to destination cores i and j, $L(P_{i,j}$ is determined by the routing algorithm. In this work the XY routing algorithm is assumed. XY is a deterministic, deadlock free routing algorithm, which routes packets from its source along the X direction first then along the Y direction towards its destination.

3.3 NoC Power Model

An NoC tile, k, consists of a number of computational units such as floating point unit (FPU), SRAM, etc. In addition to communication units such as Network interface (NI) router, channels etc. For a unit u in an NoC tile k ($u \in k$) the power (P_u) is defined as a function of the unit's maximum power consumption (P_u^{max}) and its switching activity factor (α_u) as follows:

$$P_u = P_u^{max} \alpha_u \qquad (3)$$

Here, α_u is the ratio of unit's switching workload to its maximum workload and ranges from 0 to 1.

Now, the power of an NoC tile k, P_k, that consists of a set of units, can be expressed as the sum of powers of it's units as follows:

$$P_k = \sum_{\forall u \in k} P_u = \sum_{\forall u \in k} P_u^{max} \alpha_u, \qquad (4)$$

where P_k is the power tile k.

In contrast with other tile units whose activity factor is mainly determined by the computational load of the task assigned to the tile, router's activity, α_r, is determined by a router's communication load which is set by the communication demand of the application and the placement of the communicating tasks across the whole NoC system. Considering router r with N_r number of channels, W_r channel width and

F_r frequency, the maximum bandwidth capacity of r, BW_r^{max}, can be expressed as:-

$$BW_r^{max} = W_r \times N_r \times F_r \tag{5}$$

The actual load of the router is the average bandwidth, BW_r and is determined by the application mapping function (Ω) and the routing algorithm. The router load is the summation of loads of all channels in the router. For router r with a set of channels (CH_r) the router load can be expressed as:

$$BW_r = \sum_{ch \in CH_r} BW_{ch}, \tag{6}$$

where BW_{ch} is the communication bandwidth of channel ch. Now the activity factor of router r, α_r, can be computed as

$$\alpha_r = \frac{BW_r}{BW_r^{max}}, \tag{7}$$

and the router power, P_r can be expressed as

$$P_r = P_r^{max} \alpha_r = P_r^{max} \frac{BW_r}{BW_r^{max}}, \tag{8}$$

To compute power consumption of computational units in an NoC tile, it is shown that Rent's rule still applies for packet-based systems for a wide range of applications[18, 19]. Thus, in this work, power consumption of any computational unit in a tile is assumed to be modulated by its communication power. This energy dynamically changes according to local data transfer from/to the local router. Thus, the power of computational units P_c is computed as:

$$P_c = \beta \times P_{ch}(local) \tag{9}$$

where $P_{ch}(local)$ is the power consumed by the local channel in a router and β is the ratio of the communication to the computation power of computational unit c.

The total power of the tile k, P_k, can be expressed as the summation of both communication (router) and computation powers:

$$P_k = P_r^k + P_c^k \tag{10}$$

where P_r^k, P_c^k are communication and computation powers for tile k, respectively.

A summary of some of the notations used in this paper is illustrated in Table 1.

Notation	Description
$ARG(C, P)$	NoC architectural graph with the set of NoC IP cores {C} as vertices and the set of paths {P} among these cores as arcs.
$APG(S, A)$	Application graph with the set of tasks {S} as vertices and the set of communications {A} among these tasks as arches.
$L(p)$	The set of NoC links that constitute a path p in the architectural graph ARG.
Ω	Mapping function that maps application graph to architectural graph.
C_{cost}	Communication cost
$BW_r^{max}(\Omega)$	The bandwidth capacity of router r.
$BW_r(\Omega)$	The bandwidth load of router r.
α_u	Activity factor of VLSI unit u.
P_r^k	Communication power of NoC tile k
P_c^k	Computation power of NoC tile k
P_k	total Power of NoC tile k
P_{peak}	Peak of power
P_{range}	Range of power $(min - max)$
D_{peak}	Peak of regional power density in a NoC.

Table 1: notations used in this paper

4 Methodology

4.1 Power Density Optimisation Objectives

Higher power density is associated with higher switching activity and unbalanced spatial power consumption profiles in a chip leads to higher spatial variations in power densities causing higher variations in temperatures [12], power supply voltage [13, 43], logic delay [37] and unbalanced chip wearout [24]. These variations and unbalances are major causes of permanent and transient faults and adversely affect reliability in VLSI circuits [9].

This work aims at studying multiple power density metrics that characterise power variations across NoC-based many-core systems. We begin by defining the metrics we are proposing in this paper. It is worth mentioning here that this work does not aim at optimizing a specific power density metric but to propose and study these metrics and compare their benefits and their impact on performance. We define three metrics that characterise power density and variations across the chip.

1. Peak Power (P_{peak})
2. Range of Power (P_{range})
3. Peak Regional Power Density (D_{peak})

The *rationale* behind these three metrics is that we define them with different scopes. Power peak is local, power range is global (covers the whole chip) and regional power density is defined over a region within the chip.

To describe the power density metrics used in this paper, we define P_{tiles} as the set of power consumptions of all tiles in the NoC as follows:

$$P_{tiles} = \{P_k \mid 0 \leq k \leq N-1\}, \tag{11}$$

where N is the number of tiles, P_k is the power consumption of the k^{th} tile.

4.1.1 Peak Power

Higher peak power in a many-core system is often associated with hotspot formation, higher temperatures and lower reliability [24, 10, 11]. Assuming homogeneous NoC architecture, we define peak power, P_{peak}, across the NoC tiles as the maximum power consumed across NoC tiles. i.e.,

$$P_{peak} = max(P_k) \mid \forall k \in T \ , \tag{12}$$

where T is the set of all tiles in the NoC.

Notably P_k comprises two components (Eq. 10). Thus, incorporating this metric as an optimisation objective balances the two components (communication and computation) for the hotspot in the NoC.

4.1.2 Range of Power

Unbalanced power consumption across the NoC-based many-core systems can be measured using the range of power P_{range} which can be defined as follows:

$$P_{range} = max(P_k) - min(P_k) \mid \forall k \in T \tag{13}$$

A higher range is the result of unbalanced distribution of both communication and computation workloads across the chip. The result of a higher power range is higher thermal variability, unbalanced wear-out and higher transient and permanent faults [11]. It can also lead to under and over-utilised NoC tiles, reducing performance and increasing wasted power [44].

4.1.3 Regional Power Density

In many cases, a metric that considers a region in the chip is necessary., especially if the NoC size is very large and the physical parameter that is related to power density is more affected by the power profile within a certain region than that of the entire chip.

In this work, we use the metric of *regional power density* in addition to the two metrics above. Regional power density of a tile k , (D_k) is defined in terms of power consumption within a region of a particular size in the vicinity of the tile of interest k [10]. For tile k we define this region as the set of all tiles j that satisfy $dist(k,j) \leq R$, where $dist(k,j)$ is the Manhattan distance between tiles k and j and R is the region radius. Considering a regular NoC with similar tile sizes and architectures, the regional power density of tile k, D_k is a function of region radius R and power consumption of the tiles within R and can be expressed as follows:

$$D_k = \frac{\sum\limits_{dist(k,j)\leq R} P_j}{\sum\limits_{dist(k,j)\leq R} A_j} \tag{14}$$

where R is a predefined region size. P_j and A_j are the power and area of NoC tile j, respectively.

The peak regional power density is the maximum density across NoC tiles and can be expressed as

$$D_{peak} = max(D_k) \mid \forall k \in T \tag{15}$$

Notably, the area is considered in all the three metrics above, peak, range and regional power. However, we are targeting homogeneous many-core architecture and so area is the same for all tiles. Thus the peak power, for example, is the peak power density and the range of power is the range of densities among tiles. Since area is the same for all tiles across the target NoC, the tile area can be excluded from calculations of these metrics. Regional power density has higher dependency on the area since we are introducing the concept of region here which is defined in terms of the number of tiles due to the homogeneity of the system..

4.2 Problem Formulation

We now formulate the problem of power density-aware mapping in NoCs. The problem is formulated as a multi-objective mapping that uses two objectives, namely, a power density metric (one of the metrics defined in equations 12, 13 or 15) and a cost metric (communication cost defined in 2)

The mapping problem can be formulated as follows:

Given: Application graph, $APG(S, A)$ and NoC architectural graph, $ARG(C, P)$ that satisfy

$$|S| \leq |C| \tag{16}$$

find: A mapping function Ω that maps each task in APG to a tile in ARG

Objective: minimize the set $F = \{C_{cost}, PD\}$, where PD is a power density metric (equations 12, 13 or 15)

such that:

$$\Omega(s_i) \in C \quad , \quad \forall\ s_i \in S \tag{17}$$

$$\Omega(s_i) \neq \Omega(s_j) \quad , \quad \forall\ s_i \neq s_j \tag{18}$$

$$BW(l_k) \leq BW^{max}(l_k) \quad , \quad \forall\ l_k \in L(p_i),\ \forall p_i \in P \tag{19}$$

where $BW^{max}(l_k)$ is link capacity (maximum bandwidth) and $BW(l_k)$ is link bandwidth which is determined by the application mapping function and computed as

$$BW(l_k) = \sum_{\forall a_{i,j} \in A} b(a_{i,j}) \times \pi(l_k, p(\Omega(s_i), \Omega(s_j))) \tag{20}$$

where $\pi(l, p)$ is a boolean function that determines whether a link l belongs to the the set consisting path p, $L(p)$, or not.

The first condition (Eq. 17) is used to ensure that a task is mapped to a single tile, while the second (Eq. 18) ensures that no more than a single task is mapped to a tile. In other words, Eqs. 17 and 18 ensure that one-to-one correspondence occurs between tasks in APG and cores in ARG. The third constraint (Eq. 19) guarantees that links (and hence routers) are not assigned bandwidths that exceed their capacity.

In this study, an evolutionary algorithm (EA)-based solution is adopted, namely,the Non-Dominated Sorting Genetic Algorithm, NSGA II [15] which will be described in the following section.

5 Evolutionary Algorithm Solution

Application mapping in NoCs is known to be an NP-hard problem[20]. For an NoC size of $n \times m$, the possible solutions are $(n \times m)!$. Meta-heuristic methods are commonly used to solve such problems. In this work, an evolutionary based approach is used and will be described in this section.

5.1 Gene and Chromosome Encoding

A mapping solution is encoded as *chromosome*. Each *gene* in this chromosome is the core number. In other words, the chromosome is a one-dimensional array in which the number at the i^{th} location represents the core (or tile) that task i is assigned to as illustrated in Fig. 2.

5.2 Crossover, Repair and Mutation

In evolutionary algorithms, crossover and mutation play a vital role in the performance of the used algorithm. These two processes are used to generate the next generation in the evolutionary process and must be chosen to provide a good exploration of the solution space.

The *crossover* operator is illustrated in Fig. 2. When two parents are crossed over (parent 0 and parent 1 in Fig. 2), a random crossover point is selected and the two children (child 0 and child 1 in Fig. 2) are formed by swapping the genes of the two parents up to this crossover point.

This crossover process results in some deformed genes in the resulting children because chromosomes cannot have duplicates because cores can only contain one task (Eq. 18). Thus, a *repair* phase is necessary to create valid children by simply swapping the invalid genes in the two children to create valid chromosomes, as illustrated in Fig. 2.

Mutation happens to the population with a certain rate in each evolutionary cycle. When a certain member of the population is selected, mutation happens by randomly selecting two genes within that chromosome and swapping them.

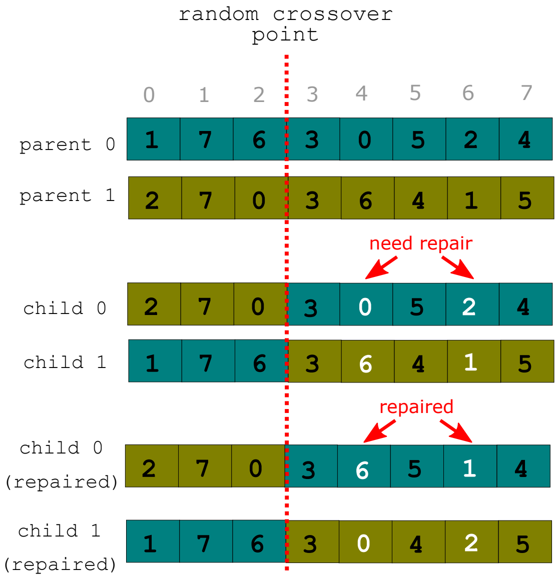


Figure 2: Illustration of the crossover operator with repair phase.

5.3 Non-dominated Sorting Genetic Algorithm (NSGA II)

The multi-objective NSGA II is used to achieve optimising communication cost and power density metrics [15].

The NSGA is a multiple objective optimization (MOO) approach. Two types of NSGA are developed, namely, the traditional NSGA and updated NSGA or NSGA II which is used in this work.

NSGA II aims at finding the best non-dominated front (Pareto front) in a multi-objective optimisation problem. NSGA II adopts the main steps of EA algorithms (*selection*, *crossover* and *mutation*). However, the score of an individual is a vector. Each objective is represented as member in the scores vector. The parent population is chosen after sorting the current population into a hierarchy of sub-populations based on their Pareto dominance. Members with high dominance are favoured over those with low dominance. If a selection is required from a sub-population with the same rank of dominance, *crowding distance* is used to favour members with less similarity and promote a diverse front of solutions.

5.4 NSGA II - Based Mapping

The NSGA II algorithm used in this work to solve the NoC mapping problem is shown in Algorithm 1. The algorithm aims at finding the Pareto front that represents the best trade-off between two mapping objectives, communication cost and a power density metric. These two metrics are computed using F_{cost} and F_{PD} functions. The algorithm uses an application graph (APG) and an architectural graph (ARG) as inputs.

Initially, the algorithm starts with a random population of mappings of size N (line 1). The main evolutionary loop starts at line 2 which iterates until a stop condition is satisfied (the max number of generations is reached in this case). The algorithm then sets the set of parents ($P_{parents}$) to empty set and parents number N_p to zero (lines 3 and 4). In line 5, the scores of the current generation ($scors_g$) are computed. Another loop accumulates members of population to the set of parents (lines 6 - 15) by iteratively selecting the non-dominated front members from the population (line 7) and accumulating them to the set of parents (line 12). Moreover, in each iteration the Pareto-front members and their corresponding scores are excluded from the population of the current generation P_g (lines 13 - 14). If the size of accumulated Pareto-front members causes the number of parents (N_p) to exceed the allowed size ($N_{parents}$), a subset

Algorithm 1 Pseudo code of the NSGAII-based NoC Mapping Solution

```
Define:
    $N_{parents}$:parents population size
    $N_{children}$:children population size
    $N$: total population size, $N_{children} + N_{parents}$
    $F_{cross}$: crossover operator
    $F_{mut}$: mutation operator
    $F_{cost}$: operator to evaluate communication cost
    $F_{PD}$: operator to evaluate a power density metric (Eq. 15, 12 or 13)
Input:
    $APG$: application graph $G(S, A)$,
    $ARG$: architectural graph $G(C, P)$.
Output:
    $P_{best}$: is the optimal population of non-dominated mappings.
    ——
 1: $P_g \leftarrow generate\_random\_population(N, ARG, APG)$
 2: while ($\neg stop\_condition()$) do
 3:    $P_{parents} \leftarrow \emptyset$
 4:    $N_p \leftarrow 0$
 5:    $scores_g \leftarrow compute\_scores(P_g, F_{cost}, F_{PD}, ARG, APG)$

 6:    while $size(P_{parents}) < N_{parents}$ do
 7:       $P_{r0} \leftarrow find\_Pareto(P_g, population\_scores)$
 8:       $N_p \leftarrow size(P_{parents}) + size(P_{r0})$
 9:       if $N_p > N_{parents}$ then
10:          $P_{r0} \leftarrow reduce\_by\_crowding(P_{r0}, N_p - N_{parents})$
11:       end if
12:       $P_{parents} \leftarrow merge(P_{parents}, P_{r0})$
13:       $P_g \leftarrow P_g n\{P_{r0}\}$     # exclude $P_{r0}$ and its scores
14:       $scores_g \leftarrow scores_g n\{scores_{r0}\}$
15:    end while
16:    $P_{children} \leftarrow breed(P_{parens}, F_{cross}, F_{mut}, N_{children})$
17:    $P_t \leftarrow merge(P_{parents}, P_{children})$
18: end while
19: $P_{best} \leftarrow P_{r0}$
20: RETURN $P_{best}$
```

with the highest crowding-distance is selected from the Pareto-front (lines 9-11).

When the set of parents is complete, the parents are bred using the crossover and mutation operators to generate $P_{children}$, the population of children (line 16). Then, $P_{parents}$ and $P_{children}$ are merged to form the next generation. When the evolutionary loop finishes, the best Pareto front (P_{best}) is the rank 0 front of the last population P_{r0} (line 19).

6 Results and Discussion

6.1 Simulation Setup and Tools

To evaluate the proposed mapping, three different multi-objective (MO) mappings are evaluated. The best mappings are obtained using NSGA II (Algorithm 1). Each MO mapping has two objectives. The first is one of the three power pattern metrics (Equations 12, 13 or 15) while the second is the communication cost (Eq. 2).

Results of MO mappings are compared to study the effect of each of the power density metrics on the resulting power profile, thermal distribution and communication cost. Results of these MO mappings are also compared with those of single objective EA mapping with only communication cost (C_{cost}) as the fitness function. In other words, four different mapping techniques are compared. These are

- **PP-COM**: multi-objective with P_{peak} and C_{cost} as objectives.
- **PR-COM**: multi-objective with P_{range} and C_{cost} as objectives.
- **DP-COM**: multi-objective with D_{peak} and C_{cost} as objectives.
- **SO-COM**: single-objective with C_{cost} only as the objective.

We used eight benchmarks, four real and four synthetic. The details of these benchmarks are shown in Table 2. Examples of these benchmarks are shown in Fig. 3. The benchmarks have different task numbers and communication requirements. The real benchmarks include a telecommunications benchmark (TELE)[29] and other three benchmarks, the AMI49, AMI25 and MPEG4 decoder found in [1]. The four synthetic benchmarks are generated using the XL-STaGe task-graph generation tool [7]. Table 2 shows the details of the benchmarks used.

Computing the power across NoC-based many-core systems requires two simulation models. The first computes the workload of different parts of the chip, while the second is a power simulator. We used

	Name	# of tasks	# of edges	target NoC size	range of BW[MB]	average BW[MB]
real	AMI25	21	51	5 × 5	53.33 - 213.33	72.15
	AMI49	48	435	7 × 7	5.33 - 85.33	11.63
	TELE16	13	22	4 × 4	11.00 - 71.00	45.36
	MPEG4	9	13	4 × 4	0.50 - 910.00	266.77
synthetic	DAG01	33	55	6 × 6	3.03 - 173.38	61.22
	DAG02	47	139	7 × 7	3.85 - 173.38	66.44
	DAG03	48	143	7 × 7	0.15 - 183.58	56.86
	DAG04	35	91	6 × 6	2.32 - 161.25	63.69

Table 2: Summary of the used benchmarks.

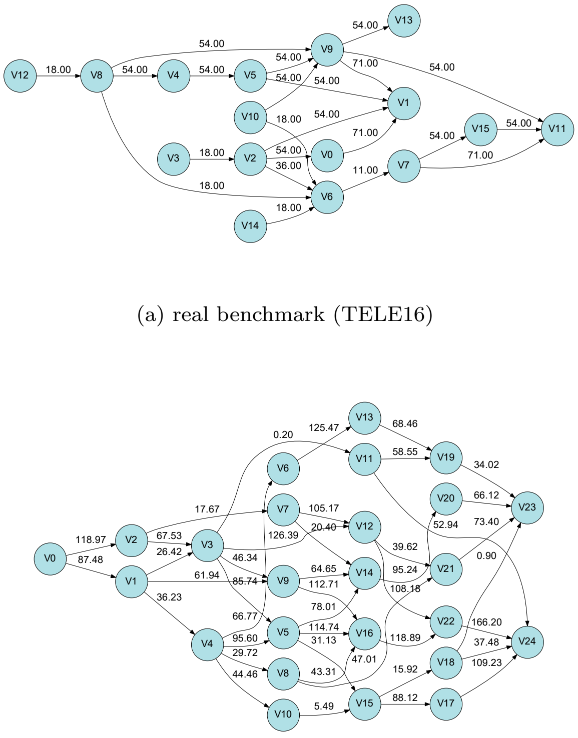
Noxim to simulate the NoC switching workload and Orion 2.0 to compute power [25].

ASIC synthesis based on an industrial library would give very accurate power evaluation. However, since the iterative evolutionary approach which runs the optimisation process many times is used in this work we need fast-to-compute power models to finish the optimisation in a reasonable time. Thus, we use Orion 2.0 to model both static and dynamic power for the NoC; notably, it is used by many previous works at a high-level early stage in power analysis which is the main focus of this paper.

The simulation model uses the application task-graph and mapping function as inputs and calculates the power of each component in the system. The workload of the computational components is computed as described in section 3.3. Temperature is computed using Hotspot 6.0 [48]. The floorplan, frequency, technology node and the power profile of Intel's TeraFLOPS is used to compute power and temperature [46].

6.2 Evolutionary Algorithm Optimisation Results

An example of the fitness curve for single-objective (SO-COM) EA mapping for the AMI49 benchmark mapping is shown in Fig 4. In this mapping, only C_{cost} is used as fitness. After 500 generations, the fitness reached a steady value and the communication cost dropped from 40,803 MB to 28,056 MB (reduced by approximately 32%). This is a baseline technique to which the proposed MO mapping techniques are going to be compared.



(a) real benchmark (TELE16)

(b) synthetic

Figure 3: Examples of (a) real, and (b) synthetic, benchmarks.

For multi-objective (MO) mappings, let's start by examining how the NSGA II algorithm is performing in finding a good Pareto front. The three power pattern metrics used represent different region sizes at which this metric is computed. A *local* one, peak power (P_{peak}), a *global* one (P_{ragen}) and a *regional* one (D_{peak}). Fig. 5 shows the Pareto fronts of the MO mapping techniques plotted with 10,000 random mappings. The result for single-objective mapping is also shown here for comparison (SO-COM).

For the three MO cases (Figs 5(a), 5(b) and 5(c)) significant improvement over random mappings occurs. Furthermore, the power pattern metric with the best results, in terms of C_{cost} penalty, is the local

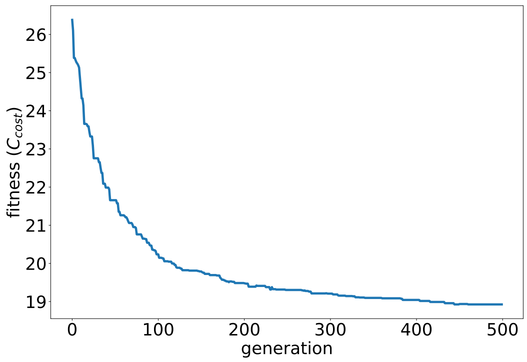


Figure 4: Example of the fitness curve of single-objective EA mapping for the AMI49 benchmark

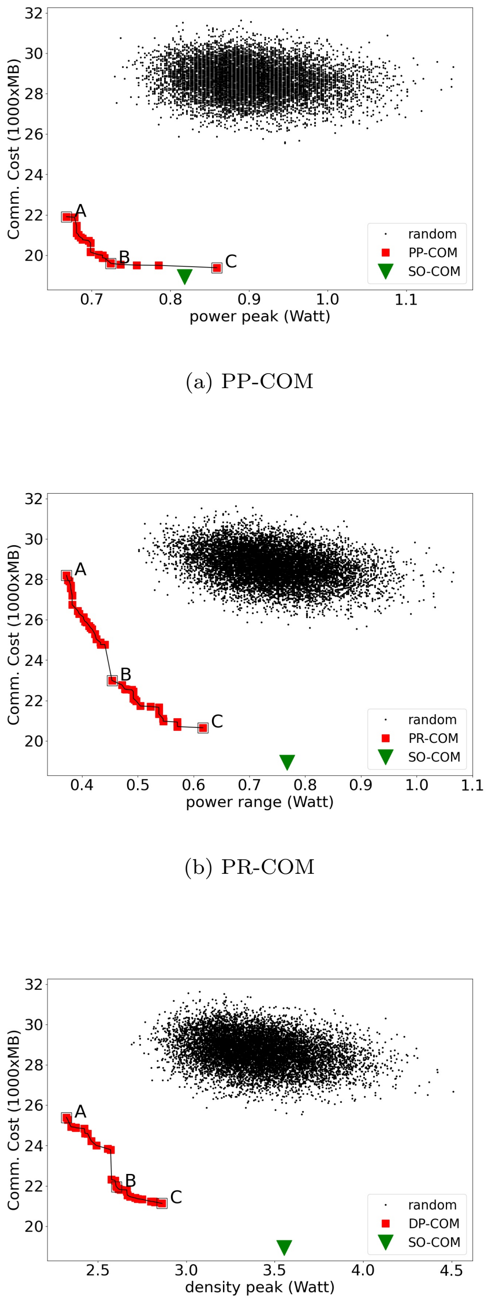
one (P_{peak} in Fig. 5(a)), whereas the one with the widest range Pareto front is the global one (P_{range} in Fig. 5(b)). Notably, the regional power density metric (D_{peak} in Fig. 5(c)) is between the other two in terms of the size of the Pareto front.

6.3 Effect on Communication Cost

In this section, we investigate how the three power pattern metrics affect a cost metric (the communication cost, C_{cost}) which is used with them in multi-objective EA mapping. Fig. 6 shows the effect of the three MO mapping strategies on communication cost. This figure shows box-plots indicating the Pareto-fronts of the three MO mappings together with that of a random population of 10,000 mappings for all the used benchmarks. These box-plots confirm the results of the example shown in Fig. 5. The global power pattern metric (PR-COM) results in the widest range of values while the local one (PP-COM) has the narrowest. This finding is consistent for all the used benchmarks, real or synthetic. Moreover, except for a few cases, the best improvement in C_{cost} is achieved by the local metric PP-COM.

6.4 Results of Power Variations

The results of how the proposed mapping techniques are affecting the spatial power variations is shown in Figs. 7 and 8.

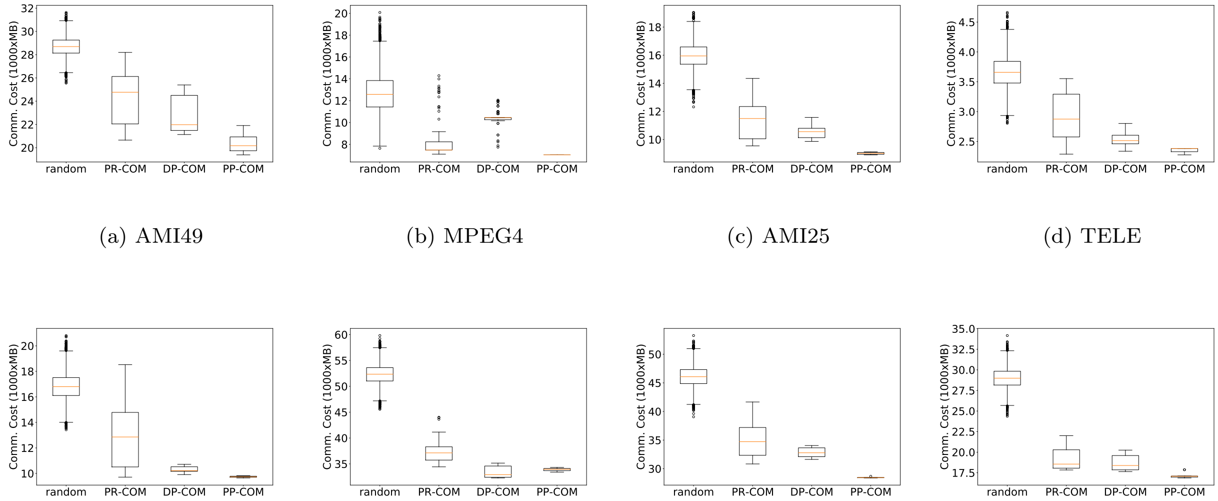


(a) PP-COM

(b) PR-COM

(c) DP-COM

Figure 5: Pareto fronts of the multi-objective mapping techniques of compared to 10,000 random mappings of the AMI49 benchmark



(a) AMI49 (b) MPEG4 (c) AMI25 (d) TELE

(e) DAG01 (f) DAG02 (g) DAG03 (h) DAG04

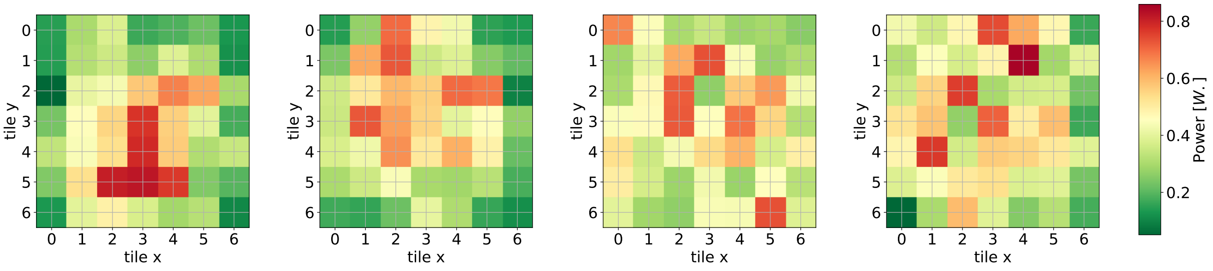
Figure 6: Box-plots showing the impact of the three mapping objectives and the random population on communication cost all the used benchmarks

Fig.7 shows the different power spatial profiles that the proposed mapping techniques are producing. This Figure shows the spatial power distribution of single objective mapping (SO-COM) which minimises C_{cost} and the three MO mappings. The MO mappings in Fig. 7 are for a selected point from the Pareto front that provide the best compromise between the power metric (P_{peak}, P_{range} or D_{peak}) and C_{cost}. This point is chosen to be mid point in the Pareto-front (point B in Figs. 5(a), 5(b) and 5(c)). As predicted, it can be seen that single-objective mapping which considers C_{cost} only causes tasks with higher communication workload and high-power draw to come close to each other and results-in hotspots (Fig 7(a).

Using MO mappings with power density metrics as the second objective leads to more homogeniety in power distributions. The resulting power distribution pattern, however, depends on the power metric used. When using P_{peak} as a second objective (Fig. 7(b)), the peak of power is lower. However, we can still see connected regions with high power draw compared with other parts of the chip. When P_{range} is used as a second objective (Fig. 7(c)), the power range has reduced significantly (only 0.45 W in this particular case compared with approximately 0.77 W in the case of SO-COM)resulting in the best balanced spatial power pattern among those shown in Fig. 7. With DP-COM mapping, using the peak regional density as a second objective (D_{peak}) produces a spatial power pattern with scattered hotspots and smaller clusters of high-power regions (Fig. 7(d)).

Fig. 8 shows the spatial power distribution of SO-COM (single objective mapping that minimises C_{cost}) and three mappings selected from the Pareto-front of the PP-COM multi-objective mapping. The three points are max. P_{peak} , mid-point and min. P_{peak} (points C, B and A in Fig. 5(a), respectively). The Pareto front explores a good range of spatial power patterns. This range starts with a bigger hotspot region (Fig. 7(b)), with low C_{cost} penalty, to a more flattened and harmonic pattern (Fig. 8(d)) with relatively high C_{cost} penalty. In the middle of this range is Fig. 8(c) which gives a good compromise between the two objectives, C_{cost} and P_{peak}.

Table 3 summarises the results of the power pat-



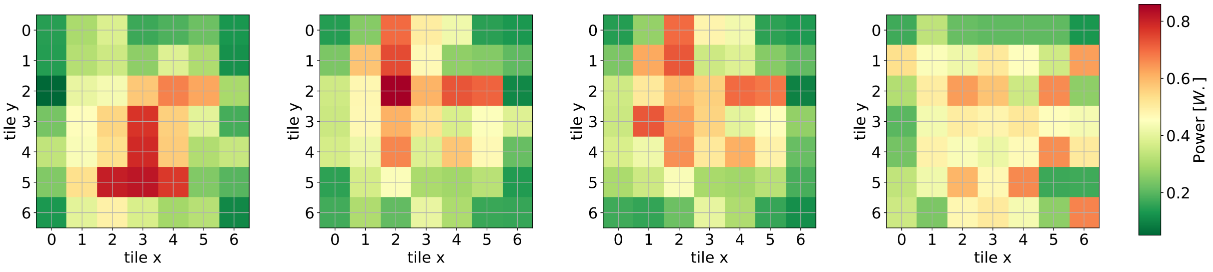
(a) SO-COM

(b) PP-COM

(c) PR-COM

(d) DP-COM

Figure 7: ***Exploring the mapping objectives***: Comparing the S.O. mapping (SO-COM) and the three MO mappings for the AMI49 benchmark.



(a) SO-COM

(b) pareto point C for PP-COM

(c) pareto point B for PP-COM

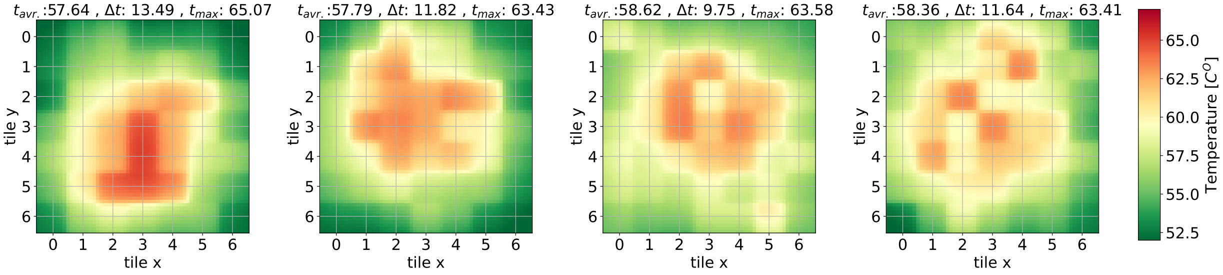
(d) pareto point A for PP-COM

Figure 8: ***Exploring the Pareto front:*** Comparing S.O mapping (SO-COM) with selected three Pareto points of an MO mapping (PP-COM). (a) SO-COM, (b) max. P_{peak} (c) best compromise with C_{cost} and (c) min. P_{peak}, Pareto points. Results shown are for the AMI49 real benchmark.

terns resulting from the mapping strategies for all the benchmarks. The table shows C_{cost}, P_{peak} and P_{range} for each mapping and the difference in C_{cost} compared with SO-COM mapping (C_{cost} penalty). It can be noticed that incorporating P_{peak} and P_{range} as second objectives with C_{cost} reduces the incorporated parameter significantly compared with a single objective. Furthermore, P_{peak} has, on average, the lowest C_{cost} penalty (only 1%), whereas P_{range} has the highest (nearly 18%). Peak regional density, D_{peak}, is shown to be a middle ground between the two, not only in terms of C_{cost} penalty but also in terms of spatial power profile characterisation parameters, P_{peak} and P_{range}.

6.5 Thermal Analysis

Spatial power distribution pattern has a direct effect on many chip parameters including permanent and transient faults, power supply integrity and tem-



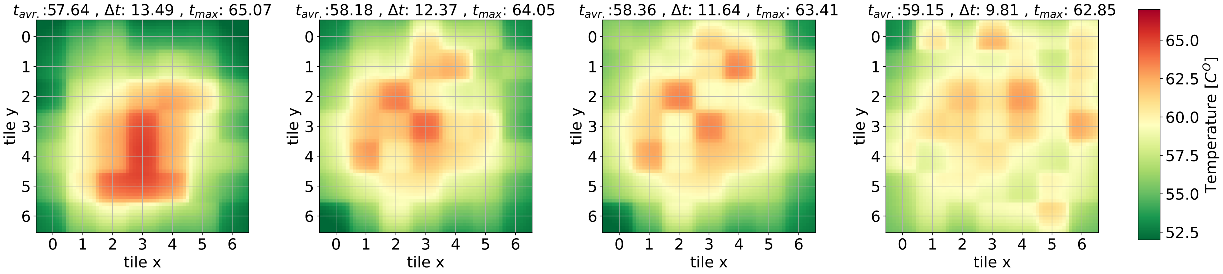
(a) SO-COM

(b) PP-COM

(c) PR-COM

(d) DP-COM

Figure 9: Comparing the best Pareto point of the three power density metrics. (a) single objective (C_{cost}), (b) max. of power (P_{peak}), (c) range of power, (P_{range}) and (d) max. of regional density, D_{peak}, for the AMI49 benchmark



(a) min. C_{cost}

(b) pareto point C

(c) pareto point B

(d) pareto point A

Figure 10: Comparing (a) single objective (C_{cost}) with three Pareto points (b) max. D_{peak} (c) good compromise with C_{cost} and (d) min. D_{peak}, of DP-COM multi-objective mapping for AMI49 benchmark

perature [43, 11, 24, 9]. Thus, the proposed mappings and their resulting spatial power patterns can be evaluated in terms of any of these parameters. In this study, we evaluate these mappings in terms of temperature. Temperature is a crucial parameter that has a direct effect on the reliability of VLIS systems. Increased temperature is associated with higher leakage and power density and is a direct result of smaller technology node and faster switching frequencies. This problem is an alerting one and the thermal challenge is a top priority for VLSI designers [9, 40].

Fig.s 9 and 10 show examples of the resulting spatial thermal distribution of the proposed mapping strategies. Fig. 9 compares the spatial thermal distribution for the four mapping techniques of the AMI49 benchmark. Results are shown for the good compromise Pareto point. All the three MO mappings resulted in lower peak temperature, T_{max} and spatial range of temperature, ΔT, when compared with sin-

		Mapping	C_{cost}	Power		Comp. w.
		Technique	[MB]	values [w]		S.O.[%]
				P_{peak}	P_{range}	C_{cost}
real benchmarks	AMI25	PR-COM	11573.05	0.92	0.69	-18.43
		PP-COM	8959.79	0.94	0.87	+5.36
		DP-COM	10559.64	0.94	0.87	-10.61
		SO-COM	9439.77	0.99	0.92	-
	AMI49	PR-COM	22994.27	0.70	0.45	-17.70
		PP-COM	19597.28	0.72	0.63	-3.43
		DP-COM	21959.35	0.81	0.76	-13.82
		SO-COM	18925.27	0.82	0.77	-
	TELE	PR-COM	2886.01	0.49	0.38	-18.85
		PP-COM	2276.01	0.50	0.47	+2.90
		DP-COM	2525.99	0.49	0.46	-7.29
		SO-COM	2341.95	0.49	0.46	-
	MPEG4	PR-COM	11457.50	2.57	2.33	-36.79
		PP-COM	7037.02	2.57	2.57	+2.91
		DP-COM	10260.49	2.57	2.57	-29.42
		SO-COM	7242.02	2.57	2.57	-
synthetic benchmarks	DAG01	PR-COM	11358.24	0.90	0.74	-20.96
		PP-COM	9728.99	0.90	0.89	-7.72
		DP-COM	10220.94	0.90	0.84	-12.16
		SO-COM	8977.95	0.96	0.95	-
	DAG02	PR-COM	35769.89	1.09	0.65	-15.15
		PP-COM	33700.28	1.07	0.93	-9.94
		DP-COM	32747.71	1.14	0.87	-7.32
		SO-COM	30350.87	1.17	1.09	-
	DAG03	PR-COM	32507.36	1.09	0.71	-13.69
		PP-COM	28448.65	1.09	0.93	-1.38
		DP-COM	32198.30	1.18	0.99	-12.86
		SO-COM	28056.37	1.25	1.12	-
	DAG04	PR-COM	18647.88	0.82	0.46	-5.31
		PP-COM	17105.90	0.82	0.68	+3.23
		DP-COM	17862.13	0.87	0.73	-1.14
		SO-COM	17658.59	0.96	0.84	-
AVERAGE		PR-COM				-18.36
		PP-COM				-1.01
		DP-COM				-11.83

Table 3: Power pattern evaluation results

		Mapping	Temp.		Comp. w.	
		Technique	values $[C^o]$		S.O.[%]	
			ΔT	T_{max}	ΔT	T_{max}
real benchmarks	AMI25	PR-COM	10.44	62.27	35.82	4.51
		PP-COM	12.94	63.34	9.58	1.59
		DP-COM	11.49	61.48	23.41	6.77
		SO-COM	14.18	63.95	-	-
	AMI49	PR-COM	9.75	63.58	38.36	3.86
		PP-COM	11.82	63.43	14.13	4.27
		DP-COM	11.64	63.41	15.89	4.32
		SO-COM	13.49	65.07	-	-
	TELE	PR-COM	4.98	52.55	24.90	0.65
		PP-COM	5.96	52.54	4.36	0.69
		DP-COM	5.22	52.08	19.16	2.40
		SO-COM	6.22	52.73	-	-
	MPEG4	PR-COM	29.36	80.96	14.13	1.07
		PP-COM	32.04	80.89	4.59	1.20
		DP-COM	29.27	78.42	14.49	5.88
		SO-COM	33.51	81.56	-	-
synthetic benchmarks	DAG01	PR-COM	11.82	61.97	12.52	1.16
		PP-COM	12.91	61.62	3.02	2.13
		DP-COM	11.65	60.78	14.16	4.53
		SO-COM	13.30	62.40	-	-
	DAG02	PR-COM	15.14	75.39	34.41	3.49
		PP-COM	17.35	74.54	17.29	5.27
		DP-COM	15.79	74.00	28.88	6.43
		SO-COM	20.35	77.15	-	-
	DAG03	PR-COM	15.47	74.16	26.18	0.77
		PP-COM	17.59	73.08	10.97	3.04
		DP-COM	15.62	72.32	24.97	4.69
		SO-COM	19.52	74.54	-	-
	DAG04	PR-COM	10.06	65.64	55.07	6.13
		PP-COM	11.95	65.02	30.54	7.77
		DP-COM	12.47	65.03	25.10	7.74
		SO-COM	15.60	68.13	-	-
AVERAGE		PR-COM			30.17	2.71
		PP-COM			11.81	3.24
		DP-COM			20.76	5.35

Table 4: Thermal Analysis

gle objective (SO-COM). The reduction in peak temperature is roughly the same (approximately $2C^o$) with DP-COM having slightly less peak temperature compared with PP-COM and PR-COM. PR-COM gives the best results in terms of thermal balancing (the lowest ΔT).

Fig. 10 shows the spatial thermal patterns of SO-COM mapping with three points selected from the Pareto front of DP-COM mapping. These points are A, B and C in Fig. 5(c). The Parto point with the lowest D_{peak} and highest C_{cost} (point A) gives the best results in terms of ΔT and T_{max} (Fig. 10(d)) and vice-versa (Fig. 10(d)). Better thermal balancing results are thus associated with higher C_{cost} penalty. A good compromise between thermal balancing and C_{cost} can be seen in Fig. 10(c) (point B in Fig. 5(c)).

Table 4 summarises the thermal evaluation results of all the benchmarks. Notably, improvement for the three MO mappings (PR-COM, PP-COM and

DP-COM) over single-objective (SO-COM). All MO mappings give lower T_{peak} and ΔT for all the benchmarks (real and synthetic). However, the reduction in ΔT is more evident and is up to 30% in the case of PR-COM. PR-COM, on average, provided the best results in terms of thermal equalisation (lowest ΔT) because this mapping aims at minimising power variation (P_{range}) which translates thermally into improved thermal balancing because temperature is a direct result of power pattern. On the other hand, on average, the best reduction in T_{max} is achieved by the DP-COM. This finding can be explained by the fact that temperature is highly affected by the regional power pattern. In other words, higher temperature is associated with larger hotspots and not just higher peak power of a single tile. Minimising D_{peak} is shown to result in a scattered power pattern (Fig. 7(d)) and produces a spatial power pattern with scattered hotspots and smaller hotspot sizes, translating into the lowest peak temperature.

7 Conclusion and Future Work

In this paper, a new mapping strategy for mesh-based NoC architecture is proposed. The new strategy focuses on power patterns and adopts an evolutionary multi-objective solution to map applications to mesh-based NoC architecture. Although the majority of mapping algorithms focus on reducing the communication cost (placing tasks with high communication bandwidth closer to each other) to reduce energy consumption and improve performance, the proposed mapping strategy investigates the resulting power patterns and their footprints. We define three different power pattern metrics (power peak, power range, and regional power density) and investigate the results of using them as mapping objectives together with communication cost. Furthermore, a case study of thermal analysis of the resulting power patterns is performed. Results show that using communication cost only produces larger hotspot regions which translates into higher-temperatures. Results also show that using power density metrics in a multi-objective evolutionary mapping results-in Pareto fronts with different power patterns and features. In terms of thermal analysis, we found that using power range produces a more balanced thermal distribution and that minimising regional density produces the lowest peak of temperature thanks to the resulting sparsity of hotspots. This study explores new design objectives and the resulting power patterns can be evaluated in terms of other metrics such as permanent and transient faults, power supply noise and timing variation which will be investigated in future work. Future work will also explore different NoC architectures and routing algorithms such as adaptive and semi-adaptive routing.

References

[1] C Ababei, H S Kia, O P Yadav, and Jingcao Hu. Energy and reliability oriented mapping for regular Networks-on-Chip. In *Networks on Chip (NoCS), 2011 Fifth IEEE/ACM International Symposium on*, pages 121–128, 2011.

[2] G Ascia, V Catania, and M Palesi. Multi-objective mapping for mesh-based NoC architectures. In *Hardware/Software Codesign and System Synthesis, 2004. CODES + ISSS 2004. International Conference on*, pages 182–187, 2004.

[3] Giuseppe Ascia, Vincenzo Catania, and Maurizio Palesi. Mapping cores on network-on-chip. *International Journal of Computational Intelligence Research*, 1(1):109–126, 2005.

[4] L Benini and G De Micheli. Networks on chips: a new {SoC} paradigm. *{IEEE} Computer*, 35(1):70–78, 2002.

[5] C. Bonney, P. Campos, N. Dahir, and G. Tempesti. Fault tolerant task mapping on many-core arrays. In *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016*, 2017.

[6] Andrew Brown, Mark Vousden, Alexander Rast, Graeme McLachlan Bragg, David Thomas, Jonny Beaumont, Matthew Naylor, and Andrey Mokhov. POETS: Distributed event-based computing - scaling behaviour. In *Conference on Parallel Computing, Charles University*, Czech Republic, may 2019. ACM.

[7] P. Campos, N. Dahir, C. Bonney, M. Trefzer, A. Tyrrell, and G. Tempesti. XL-STaGe: A cross-layer scalable tool for graph generation, evaluation and implementation. In *Proceedings - 2016 16th International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation, SAMOS 2016*, 2017.

[8] E L d. S. Carvalho, N L V Calazans, and F G Moraes. Dynamic Task Mapping for MPSoCs. *IEEE Design Test of Computers*, 27(5):26–35, 2010.

[9] N. Dahir, R. Al-Dujaily, T. Mak, and A. Yakovlev. Thermal optimization in network-on-chip-based 3D chip multiprocessors using Dynamic Programming Networks. *Transactions on Embedded Computing Systems*, 13(4 SPEC. ISSUE), 2014.

[10] Nizar Dahir, Terrence Mak, Fei Xia, and Alex Yakovlev. Minimizing power supply noise through harmonic mappings in networks-on-chip. In *Proceedings of the eighth IEEE/ACM/IFIP international conference on Hardware/software codesign and system synthesis*, CODES+ISSS '12, pages 113–122, New York, NY, USA, 2012. ACM.

[11] Nizar Dahir, Terrence Mak, Fei Xia, and Alex Yakovlev. Modelling and Tools for Power Supply Variations Analysis in Networks-on-Chip. *IEEE Transactions on Computers*, PP(99):1, 2012.

[12] Nizar Dahir, Ghaith Tarawneh, Terrence Mak, Ra'ed Al-Dujaily, and Alex Yakovlev. Design and Implementation of Dynamic Thermal-Adaptive Routing Strategy for Networks-on-Chip. In *2014 22nd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing*, pages 384–391. IEEE, feb 2014.

[13] N.S. Dahir, T. Mak, F. Xia, and A. Yakovlev. Modeling and Tools for Power Supply Variations Analysis in Networks-on-Chip. *IEEE Transactions on Computers*, 63(3):679–690, mar 2014.

[14] W J Dally and B Towles. Route packets, not wires: on-chip interconnection networks. In *DAC-2001*, pages 684–689, 2001.

[15] Kalyanmoy Deb, Samir Agrawal, Amrit Pratap, and T. Meyarivan. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1917:849–858, 2000.

[16] S D'souza, J Soumya, and S Chattopadhyay. A constructive heuristic for application mapping onto an express channel based Network-on-Chip. In *2015 19th International Symposium on VLSI Design and Test*, pages 1–6, 2015.

[17] Michael R Garey and David S Johnson. *Computers and intractability*, volume 174. freeman San Francisco, 1979.

[18] Daniel Greenfield, Arnab Banerjee, Jeong-Gun Lee, and Simon Moore. Implications of Rent's Rule for NoC Design and Its Fault-Tolerance. In *Proceedings of the First International Symposium on Networks-on-Chip*, NOCS '07, pages 283–294, Washington, DC, USA, 2007. IEEE Computer Society.

[19] Wim Heirman, Joni Dambre, Dirk Stroobandt, and Jan Van Campenhout. Rent's rule and parallel programs: Characterizing network traffic behavior. In *International Workshop on System Level Interconnect Prediction, SLIP*, pages 87–94, New York, New York, USA, 2008. ACM Press.

[20] Jingcao Hu and R Marculescu. Energy-aware mapping for tile-based {NoC} architectures under performance constraints. In *Design Automation Conference, 2003. Proceedings of the ASP-DAC 2003. Asia and South Pacific*, pages 233–239, 2003.

[21] Jingcao Hu and R Marculescu. Energy- and performance-aware mapping for regular {NoC} architectures. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 24(4):551–562, 2005.

[22] Wei Huang, M R Stan, S Gurumurthi, R J Ribando, and K Skadron. Interaction of scaling trends in processor architecture and cooling. In *Semiconductor Thermal Measurement and Management Symposium, 2010. SEMI-THERM 2010. 26th Annual IEEE*, pages 198–204, 2010.

[23] Wooyoung Jang and David Pan. A3MAP: architecture-aware analytic mapping for networks-on-chip. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 17(3):26, 2012.

[24] Jedec. Failure mechanisms and models for semiconductor devices. *JEDEC Publication JEP122-A*, 2002.

[25] A B Kahng, Bin Li, Li-Shiuan S Peh, and K Samadi. ORION 2.0: A Power-Area Simulator for Interconnection Networks. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 20(1):191–196, jan 2012.

[26] B. Li, X. Wang, A. K. Singh, and T. Mak. On runtime communication and thermal-aware application mapping and defragmentation in 3d noc systems. *IEEE Transactions on Parallel and Distributed Systems*, 30(12):2775–2789, 2019.

[27] W. Liu, L. Yang, W. Jiang, L. Feng, N. Guan, W. Zhang, and N. Dutt. Thermal-aware task mapping on dynamically reconfigurable network-on-chip based multiprocessor system-on-chip. *IEEE Transactions on Computers*, 67(12):1818–1834, 2018.

[28] Amin Mosayyebzadeh, Maziar Mehdizadeh Amiraski, and Shaahin Hessabi. Thermal and power aware task mapping on 3d network on chip. *Comput. Electr. Eng.*, 51(C):157–167, April 2016.

[29] Srinivasan Murali and Giovanni De Micheli. Bandwidth-Constrained Mapping of Cores onto {NoC} Architectures. In *Proceedings of the conference on Design, automation and test in Europe - Volume 2*, DATE '04, pages 896 – 901 Vol.2, Washington, DC, USA, 2004. IEEE Computer Society.

[30] F N Najm. Transition density: a new measure of activity in digital circuits. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 12(2):310–323, feb 1993.

[31] Luciano Ost, Marcelo Mandelli, Gabriel Marchesan Almeida, Leandro Moller, Leandro Soares Indrusiak, Gilles Sassatelli, Pascal Benoit, Manfred Glesner, Michel Robert, and Fernando Moraes. Power-aware dynamic mapping heuristics for NoC-based MPSoCs using a unified model-based approach. *ACM Trans. Embed. Comput. Syst.*, 12(3):75:1—-75:22, apr 2013.

[32] E Painkras, L A Plana, J Garside, S Temple, S Davidson, J Pepper, D Clark, C Patterson, and S Furber. SpiNNaker: A multi-core System-on-Chip for massively-parallel neural net simulation. In *Custom Integrated Circuits Conference (CICC), 2012 IEEE*, pages 1–4, 2012.

[33] M Pedram and S Nazarian. Thermal Modeling, Analysis, and Management in VLSI Circuits: Principles and Methods. *Proceedings of the IEEE*, 94(8):1487–1501, 2006.

[34] P K Sahu, N Shah, K Manna, and S Chattopadhyay. A new application mapping algorithm for mesh based Network-on-Chip design. In *2010 Annual IEEE India Conference (INDICON)*, pages 1–4, 2010.

[35] P K Sahu, P Venkatesh, S Gollapalli, and S Chattopadhyay. Application Mapping onto Mesh Structured Network-on-Chip Using Particle Swarm Optimization. In *2011 IEEE Computer Society Annual Symposium on VLSI*, pages 335–336, 2011.

[36] Pradip Kumar Sahu and Santanu Chattopadhyay. A survey on application mapping strategies for Network-on-Chip design. *Journal of Systems Architecture*, 59(1):60–76, 2013.

[37] M Saint-Laurent and M Swaminathan. Impact of power-supply noise on timing in high-frequency microprocessors. *IEEE Transactions on Advanced Packaging*, 27(1):135–144, 2004.

[38] Praveen Salihundam, Shailendra Jain, Tiju Jacob, Shasi Kumar, Vasantha Erraguntla, Yatin Hoskote, Sriram Vangal, Gregory Ruhl, and Nitin Borkar. A 2 Tb/s 6\times 4 mesh network for a single-chip cloud computer with DVFS in 45 nm CMOS. *IEEE journal of solid-state circuits*, 46(4):757–766, 2011.

[39] Amit Kumar Singh, Piotr Dziurzanski, Hashan Roshantha Mendis, and Leandro Soares Indrusiak. A Survey and Comparative Study of Hard and Soft Real-Time Dynamic Resource Allocation Strategies for Multi-/Many-Core Systems. *ACM Comput. Surv.*, 50(2), apr 2017.

[40] K Skadron, M R Stan, W Huang, Sivakumar Velusamy, Karthik Sankaranarayanan, and D Tarjan. Temperature-aware microarchitecture. In *Computer Arch., 2003. Proc. 30th Annual International Symp. on*, pages 2–13, jun 2003.

[41] K Srinivasan, K S Chatha, and G Konjevod. Linear-programming-based techniques for synthesis of network-on-chip architectures. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 14(4):407–420, 2006.

[42] Y Z Tei, M N Marsono, N Shaikh-Husin, and Y W Hau. Network partitioning and GA heuristic crossover for NoC application mapping. In *2013 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1228–1231, 2013.

[43] A Todri, M Marek-Sadowska, and J Kozhaya. Power supply noise aware workload assignment for multi-core systems. In *Computer-Aided Design, 2008. ICCAD 2008. IEEE/ACM International Conference on*, pages 330–337, 2008.

[44] R Tornero, V Sterrantino, M Palesi, and J M Orduna. A multi-objective strategy for concurrent mapping and routing in networks on chip. In *Parallel Distributed Processing, 2009. IPDPS 2009. IEEE International Symposium on*, pages 1–8, 2009.

[45] Suleyman Tosun. New heuristic algorithms for energy aware application mapping and routing on mesh-based NoCs. *Journal of Systems Architecture*, 57(1):69–78, 2011.

[46] S R Vangal, J Howard, G Ruhl, S Dighe, H Wilson, J Tschanz, D Finan, A Singh, T Jacob, S Jain, V Erraguntla, C Roberts, Y Hoskote, N Borkar, and S Borkar. An 80-Tile Sub-100-W TeraFLOPS Processor in 65-nm CMOS. *IEEE Journal of Solid-State Circuits*, 43(1):29–41, 2008.

[47] J. Wang, Z. Lu, Y. Li, Y. Fu, and J. Guo. A high-level thermal model-based task mapping for cmps in dark-silicon era. *IEEE Transactions on Electron Devices*, 63(9):3406–3412, 2016.

[48] Runjie Zhang, Mircea R Stan, and Kevin Skadron. Hotspot 6.0: Validation, acceleration and extension. *University of Virginia, Tech. Rep*, 2015.

[49] Peter Zipf, Gilles Sassatelli, Nurten Utlu, Nicolas Saint-Jean, Pascal Benoit, and Manfred Glesner. A Decentralised Task Mapping Approach for Homogeneous Multiprocessor Network-On-Chips. *International Journal of Reconfigurable Computing*, 2009:453970, 2009.