Energy-Aware Network-on-Chip Application Mapping Based on Domain Knowledge Genetic Algorithm

Yin Zhen Tei[1], Yuan Wen Hau[2], N. Shaikh-Husin [1], Trias Andromeda[3], M. N. Marsono[1]
[1]Department of Electronic and Computer Engineering, Faculty of Electrical Engineering,
Universiti Teknologi Malaysia, Johor, Malaysia.
[2]IJN-UTM Cardiovascular Engineering Center, Faculty of Biosciences and Medical Engineering,
Universiti Teknologi Malaysia, Johor, Malaysia.
[3]Department of Electrical Engineering, Diponegoro University, Semarang, Indonesia.
Email: yztei2@live.utm.my, hauyuanwen@biomedical.utm.my, nasirsh@fke.utm.my,
triasandromeda@undip.ac.id, nadzir@fke.utm.my

***Abstract*—This paper addresses energy-aware application mapping for large-scale Network-on-chip (NoC). The increasing number of intellectual property (IP) cores in multi-processor system-on-chips (MPSoCs) makes NoC application mapping more challenging to find optimum core-to-topology mapping. This paper proposes an application mapping technique that incorporates domain knowledge into genetic algorithm (GA) to minimize the energy consumption of NoC communication. The GA is initialized with knowledge on network partition whereas the genetic crossover operator is guided with inter-core communication demands. NoC energy estimation is based on analytical energy model and cycle-accurate Noxim simulation. For large-scale NoC, application mapping using knowledge-based genetic operator saves up to 28% energy compared to the one on conventional GA. Adding knowledge-based initial mapping speeds up convergence by 81% and further saves energy by 5% compared to only knowledge-based crossover GA. Furthermore, cycle-accurate simulations of applications with traffic dependency show the effectiveness of the proposed application mapping for large-scale NoC.**

Keywords*—Application mapping, bit energy model, cycle-accurate simulation, domain knowledge, genetic algorithm, network-on-chip

I. INTRODUCTION

Network-on-chip (NoC) has emerged as a promising on-chip communication architecture providing modularity and scalability for multi-processor System-on-Chips (MPSoCs). Application mapping determines the placement of intellectual property (IP) cores to routers on NoC tiles such that the performance or cost metrics of interest are optimized [1]. Large MPSoC requires an effective mapping algorithm to reduce the large search space to obtain optimum mapping.

Domain-knowledge has been used in crossover and mutation operators to improve GA mapping and convergence [2] by checking each gene's communicating distances with other cores. However, this increases computation time drastically for highly communicating applications and large-scale NoCs. Large-scale MPSoCs are mostly combinations of several subsystems. Network partitioning (NP) can be utilized to narrow down application mapping search space.

Analytical energy models commonly used in application mapping are bit energy model [3] and communication cost [4]. Both analytical models are hop-count based that offer fast cost or performance estimation. Cycle-accurate simulation gives more accurate estimation but is time consuming. Thus, it is important to analyse NoC energy accurately to obtain mapping with minimum energy. The accuracy of cost and performance estimation is equally important especially during NoC design stage.

This paper proposes an application mapping technique that incorporates domain knowledge into genetic algorithm (NP-DKGA) to minimize the energy consumption of NoC communication. NP-DKGA operates in two phases: network partitioning knowledge as initial population; and knowledge-based crossover to search for near optimum mapping. This technique is verified with several benchmarks. The proposed energy-aware application mapping is verified with both analytical energy model and cycle-accurate simulation using Noxim [5]. With only knowledge-based crossover, the GA converges well for all small communicating benchmarks. For highly communicating benchmarks, knowledge-based initial mapping can further optimize energy consumption and speeds up the GA convergence.

The rest of this paper is organized as follows. Section II discusses related works mainly on crossover and partitioning in application mapping. Section III presents the proposed application mapping technique based on the combination of knowledge-based initial mapping and crossover in GA as well as their formal definitions. Section IV discusses the simulation tools and parameters. Section V discusses the experiment results. Finally, Section VI concludes the paper and presents suggestion for future works.

Corresponding author: M. N. Marsono, nadzir@fke.utm.my

II. RELATED WORK

Different GA crossover techniques have been proposed such as hotspot remap [6, 7] and communicating cores swap with neighbouring cores [8]. These techniques do not combine both parent chromosomes' features. More effective genetic operators with useful knowledge have a great impact on the final mapping [6]. In the domain-knowledge evolutionary algorithm [2], mapping similarity crossover (MS) has been proposed to maintain the common characteristic in genes between parents and the rest of the genes using greedy mapping. Mapping similarity is able to handle symmetric problems in mesh topology by computing every gene's communication cost in term of hop count.

Large MPSoC system can be divided into several clusters (partitions). A mapping algorithm based on Kernighan-Lin (KL) partitioning, called LMAP, has been proposed to explore search space via flipping the partitions and cores in a hierarchical fashion [4]. Cluster-based relaxation for integer linear programming [9] and partition-based with near-convex [10] application mapping techniques do not allow cross partition movement. Although they show shorter runtime, the final mapping quality is affected [10]. Given a random initial mapping, Optimized Simulated Annealing (OSA) [11] improves SA by clustering communicating cores during swapping process. OSA shows better mapping quality compared to CSA.

Cycle-accurate simulation for NoC performance evaluation has been proposed in [6] but it is time consuming for large NoC. Therefore, different analytical models have been proposed, e.g. bit energy model [3] and communication cost [4]. These models have trade-off between performance and accuracy. A mapping algorithm based on a modified bit energy model [12] was proposed by Hu and Marculescu using branch and bound technique such that energy consumption can be minimized with bandwidth reservation [3]. Reference [13] compares few application mapping algorithms using bit energy model which is targeted for low energy consumption. Genetic algorithm (GA) [2] technique was also proposed to optimize energy consumption using the bit energy model. In references [4, 9], the proposed application mapping optimization are based on communication cost in term of the distance among communicating cores.

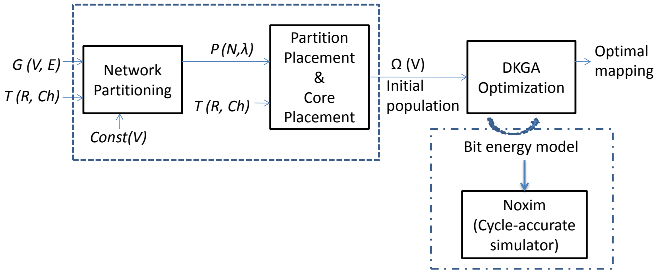

Fig. 1: Overview of the proposed technique, NP-DKGA.

III. DOMAIN-KNOWLEDGE GENETIC ALGORITHM APPLICATION MAPPING

This paper proposes an application mapping technique that incorporates domain knowledge into genetic algorithm (NP-DKGA) with the aim to minimize the energy consumption of NoC communication. The overview of the proposed technique is shown in Fig. 1. Some definitions used in this paper are defined next.

A. Problem Formulation

Definition 1: An application characteristic graph (APCG), $G(V,E)$ is a directed graph, where each vertex $v_i \in V$ represents an IP core and each directed edge $e_{(i,j)} \in E$ characterizes the total communication volume in bits from vertex v_i to vertex v_j. Application tasks are assumed to be assigned to all vertices, v_i and scheduled to each IP core.

Definition 2: NoC mesh-based network, $T(R,Ch)$ is a labelled graph, where each $r_i \in R$ denotes a router and each $ch_i \in Ch$ denotes a channel.

Definition 3: Given an input APCG, network partitioning decomposes APCG into smaller m partitions or subsystems. NP is to obtain $P(N,\lambda)$ where N is number of cores in each partition and λ is inter-partition traffic. The objective of NP is to reduce inter-partition traffic λ, subject to constraints $Const(V)$ to obtain a balanced number of cores for all partitions.

B. Network Partitioning as Knowledge-based GA Initial Mapping

The inter-partition traffic reduction technique groups heavily communicating IP cores closer (in the same cluster) that increases the probability for GA to converge. Network partitioning is implemented in two stages: mesh-based network partitioning and application characteristic graph (APCG) partitioning. The mesh topology and APCG are partitioned with equal number of tiles and cores in each partition. Each partition are randomly-mapped within the randomly mapped partition on the mesh topology.

C. Knowledge-based Genetic Algorithm

Our proposed domain-knowledge genetic algorithm applies NP as initial population and knowledge-based crossover (NP-DKGA) instead of utilizing conventional genetic algorithm (CGA). The proposed knowledge-based crossover technique is shown in Algorithm 1. If the same integer is assigned to two genes in the resulting chromosome, the latter gene is labelled as InvalidGene. Cores that are not assigned to any gene are labelled as UnmappedCores. In CGA, all InvalidGenes are randomly remapped with UnmappedCores. However, in the proposed DKGA, we apply a knowledge-based crossover technique. The UnmappedCores will determine its communication with the adjacent router of the InvalidGene. The UnmappedCores are remapped to the InvalidGene that has the highest communication cost with the NeighborCore. This crossover algorithm is done iteratively until the number of generated children chromosomes reaches the population size.

Based on previous works [2, 3], this paper applies the bit energy model as the fitness function. $E_{bit}^{V_S,V_D}$ is the required energy to transfer a bit from a source core to a destination core.

$$E_{bit}^{v_S,v_D} = n_{hops}E_{L_{bit}} + (n_{hops}+1)E_{R_{bit}} \qquad (1)$$

Where n_{hops} is the number of hops from the source to the destination using XY deterministic routing. $E_{L_{bit}}$ is the energy consumption for a link between adjacent routers and $E_{R_{bit}}$ is the router energy consumption [3]. The overall energy consumption E^A is the summation of all energy bit consumed by all bit transmissions.

$$E_A = \sum_{all\ S,D} (E_{bit}^{v_S,v_D} \times e_{S,D}) \qquad (2)$$

where $e_{S,D}$ is the total communication traffic (in bits) from the source core to the destination core.

To validate the analytical energy estimation, a cycle-accurate NoC simulation is used. Noxim [5] provides cycle-accurate energy estimation based 2mm×2mm tiles size. The total energy includes the energy for transmitting flit, receiving flit, routing, selection and standby energy. Besides, for detail energy use for link, arbitration and crossover is also included. The energy consumption of NoC is evaluated for each simulation cycles in network interfaces and routers.

Algorithm 1 Knowledge-based Crossover Algorithm

```
Population is the population size
TotalParent is total parent chromosomes
B is the length of chromosome
For i = TotalParent +1 to Population do
    Select parent chromosome using roulette wheel, P1 and P2.
    Select random crossover point, $C_p \in B$.
    Child(i) ⟵ Crossover between P1 and P2.
    Check InvalidGene.
    Check UnmappedCores.
    NeighborCore = GetAdjacentCore(InvalidGene)
    CommunicatingCore=GetCommCore(NeighborCore,
    UnmappedCores)
    InvalidGene ⟵ max(CommunicatingCore)
end for
```

IV. Simulation Methodology

Six real applications included in MSCL [14]: FFT, FPPPP, SPARSE, ROBOT, RSenc and RSdec are used. A 12×12 mesh-based architecture in MSCL is chosen for assessing the scalability of the proposed algorithm. Additionally, we also implement VOPD (video object plane decoder) [15] for 4×4 mesh-based network.

For all the benchmarks, network partitioning is implemented using Chaco [16] to generate the NP-DKGA initial population. Chaco performs bisection partitioning by grouping highly communicating cores in the same partition and at the same time, reduces the inter partition traffic.

TABLE 1. Connectivity degree for all benchmark applications.

Benchmarks	Range of connectivity degree
RSenc	0-14
ROBOT	0-15
FFT	60-116
RSdec	0-43
SPARSE	0-9
FPPPP	0-80
VOPD	1-4

This work does not analyse the optimal parameters for DKGA but rather to assess the effectiveness of the knowledge-based in initial population and genetic crossover operator. The NP-DKGA crossover probability is fixed to 0.8, population-based mutation rate to 0.3, and population size to 100 for 12×12 network size and 50 for the 4×4 network respectively. The termination of GA is set to 1000 generations for MCSL applications and 300 generations for VOPD application.

TABLE 1 shows the connectivity degree for all benchmarks used in the experiments. The connectivity degree is defined as the total incoming and outgoing communication pairs for each core in each benchmark application. The FFT cores have the connectivity degree between 60 and 116. Other benchmark applications contain at least one IP core that are not communicating to other cores. The relationship between the connectivity degree and the GA convergence will be analysed in the next section.

We analyse the convergence speed of GA with knowledge-based crossover and NP initial mapping for all benchmarks and compared them with CGA. The convergence of GA is defined in (3). GA is defined to have converged only if the convergence index *C* is less than 1% for the last 100 generations.

$$C = \frac{E_A[i-100] - E_A[i]}{E_A[i-100]} \leq 1\% \qquad (3)$$

Lastly, we analyse the accuracy of energy estimation using analytical model (bit energy model) and compare it with cycle-accurate simulation using Noxim. Analytical energy model provides faster estimation but it is not cycle-accurate. MSCL benchmark applications used in evaluation provides traffic incoming dependency, traffic outgoing dependency and computation time for each core. These real world criteria may incur congestion and longer packet waiting time in routers. These dependencies are not easily captured in analytical energy models. A cycle-accurate NoC simulator (Noxim) [5] has been used in this paper to evaluate the accurate energy cost of the proposed techniques with cycle-accurate energy model. The result is compared to the bit energy model [3].

V. Results and Discussion

All these benchmark applications converge within 100 generations with knowledge-based crossover. Fig. 2 shows that NP-based initial mapping highly reduces the energy consumption by grouping highly communicating cores and assists DKGA to potentially low energy mapping space in the first generation. For all the benchmarks, highly communicating applications always give lower average energy consumption as the FFT benchmark in Fig. 2.

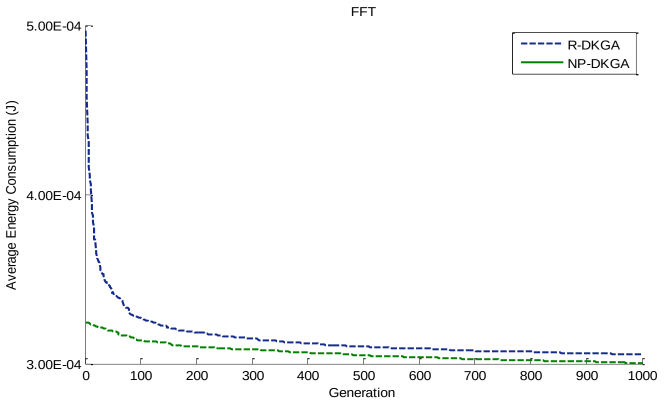

Fig. 2. Average energy consumption over generation for FFT benchmark with and without NP-based initial mapping.

TABLE 2 shows that convergence speed and energy saving improve significantly in all benchmark application regardless if the NP initial mapping is applied to DKGA. These metrics are benchmarked against CGA. The highlighted cells show better improvement between R-DKGA and NP-DKGA. Five out of seven benchmarks show better convergence speed improvement when NP-based initial mapping is applied. For highly communicating applications FFT, FPPPP and RSdec, simulation results show that NP knowledge initial mapping helps DKGA converges up to 81% faster compared to DKGA.

TABLE 2 shows that advanced NP-based initial mapping helps to improve convergence and assist DKGA to obtain high quality mapping especially for highly communicating applications. In addition, VOPD application optimized using R-DKGA and NP-DKGA achieved the global minimum identical to the one reported in reference [17].

TABLE 2. The percentage improvement on convergence speed and energy saving of the best case for R-DKGA and NP-DKGA with conventional GA (CGA) as the reference point.

Benchmark	Convergence Improvement		Energy Improvement	
	R-DKGA	NP-DKGA	R-DKGA	NP-DKGA
RSenc	65%	62%	2%	2%
ROBOT	56%	65%	14%	16%
FFT	14%	67%	28%	29%
RSdec	5%	86%	10%	10%
SPARSE	33%	44%	25%	27%
FPPPP	38%	47%	10%	15%
VOPD	69%	55%	1%	1%

Several best mappings generated using R-DKGA and NP-DKGA optimization process are selected and evaluated using bit energy model as well as using cycle-accurate simulation model. The input of cycle-accurate Noxim simulation includes traffic dependency, which is hard to be modelled analytically. The result is compared with analytical bit energy model. Fig. 3 shows the accuracy of the bit energy model against the cycle-accurate simulation model. The results show that the analytical energy model always gives lower energy estimation compared to the one based on cycle-accurate simulation. The best estimated energy-optimized mappings in bit energy model also give the lowest energy in Noxim simulation. The results show that the proposed knowledge-based initial mapping and crossover operator in GA that is based on clustering highly communicating cores is able to reduce energy consumption even when traffic dependency is included.

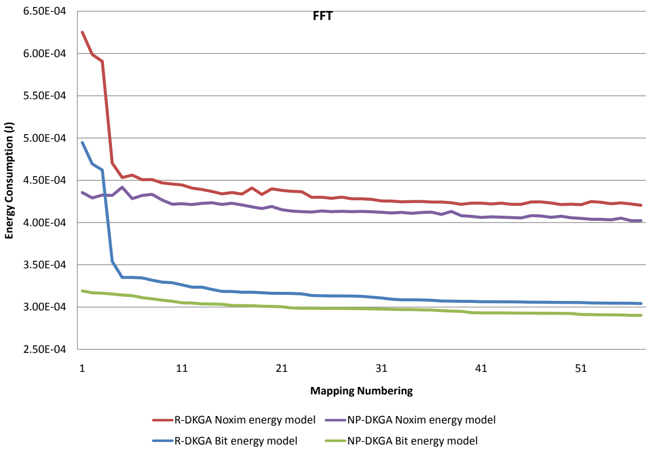

Fig. 3. Comparison between analytical bit energy model and cycle-accurate simulation model (Noxim energy model) [5].

VI. Conclusion

This paper presented the NP-DKGA technique that uses network partitioning knowledge as the GA initial mapping and knowledge-based crossover to optimize NoC application mapping. We performed analysis on several real benchmark applications. The effectiveness of the knowledge-based crossover gives significant energy reduction compared to the GA with NP-based initial mapping. For less communicating applications, knowledge-based crossover GA (DKGA) can converge well comparable to CGA. For highly communicating application, our experiment shows that NP-based initial mapping can further improve both the application mapping quality and speed up the mapping convergence. Grouping highly communicating cores with NP-initial mapping and knowledge based crossover operator result in energy minimization even when the traffic dependency is included in the energy estimation.

For future works, we plan to consider multi-objective application mapping environment. Thermal balance is an issue to reduce faults in NoC and to increase NoC reliability. For energy and thermal balanced, network partitioning may need to be done with balanced load and reduced inter-partition traffic. This work can also be extended to integrate DKGA with cycle-accurate NoC simulator for better multi-objective optimization.

Acknoledgment

This work is supported by the Ministry of Education of Malaysia Fundamental Research Grant (UTM vote 4F327).

References

[1] R. Marculescu, U. Ogras, L.-S. Peh, N. Jerger, and Y. Hoskote, “Outstanding research problems in NoC design: System, microarchitecture, and circuit perspectives,” IEEE Transactions on Computer-Aided Design of Integrated Circuits and System, vol. 28, no. 1, pp. 3–21, January 2009.

[2] C. Radu, M. S. Mahbub, and L. Vintan, “Developing domain-knowledge evolutionary algorithms for Network-on-Chip application mapping,” Microprocessors and Microsystems, vol. 37, no. 1, pp. 65–78, 2013.

[3] J. Hu and R. Marculescu, “Energy-aware mapping for tile-based NoC architectures under performance constraints,” in Proceedings of Asia and South Pacific Design Automation Conference (ASP-DAC), 2003, pp.233–239.

[4] P. K. Sahu, K. Manna, N. Shah, and S. Chattopadhyay, “Extending Kernighan-Lin partitioning heuristic for application mapping onto Network-on-Chip,” Journal of Systems Architecture, 2014.

[5] M. Palesi, D. Patti, and F. Fazzino, “Noxim - an open network-on-chip simulator,” 2013, url:http://sourceforge.net/projects/noxim.

[6] G. Ascia, V. Catania, and M. Palesi, “Multi-objective mapping for mesh-based NoC architectures,” in Proceedings of the 2nd IEEE/ACM/IFIP International conference on Hardware/software codesign and system synthesis (CODES+ISSS ’04), 2004, pp. 182–187.

[7] A.A.Morgan, “Networks-on-Chip: Modeling, system-level abstraction, and application-specific architecture customization,” Ph.D. dissertation, University of Victoria, 2011.

[8] A. A. Morgan, H. Elmiligi, M. W. El-Kharashi, and F. Gebali, “Multi-objective optimization of NoC standard architectures using genetic algo-rithms,” in Proceedings of the The 10th IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2010, pp. 85–90.

[9] S. Tosun, “Cluster-based application mapping method for Network-on-Chip,” Advances in Engineering Software, vol. 42, no. 10, pp. 868–874, October 2011.

[10] W. Jang and D. Pan, “A3MAP: Architecture-aware analytic mapping for Networks-on-Chip,” in Procedding 15th Asia and South Pacific on Design Automation Conference (ASP-DAC), 2010, pp. 523–528.

[11] C. Radu and L. Vintan, “Domain-knowledge optimized simulated annealing for Network-on-Chip application mapping,” in Advances in Intelligent Control Systems and Computer Science, ser. Advances in Intelligent Systems and Computing, L. Dumitrache, Ed. Springer Berlin Heidelberg, 2013, vol. 187, pp. 473–487.

[12] T. Ye, L. Benini, and G. De Micheli, “Analysis of power consumption on switch fabrics in network routers,” in Proceedings 39th Design Automation Conference, 2002, pp. 524–529.

[13] C. Marcon, E. Moreno, N. Calazans, and F. Moraes, “Comparison of Network-on-Chip mapping algorithms targeting low energy consumption,” IET Computers Digital Techniques, vol. 2, no. 6, pp. 471–482, November 2008.

[14] W. Liu, J. Xu, X. Wu, Y. Ye, X. Wang, W. Zhang, M. Nikdast, and Z. Wang, “A NoC traffic suite based on real applications,” in IEEE Computer Society Annual Symposium on VLSI (ISVLSI), 2011, pp. 66–71.

[15] E. B. Van Der Tol and E. G. T. Jaspers, “Mapping of MPEG-4 decoding on a flexible architecture platform,” in Media Processors, 2002, pp. 1–13.

[16] B. Hendrickson and R. Leland, “The Chaco user’s guide version 2.0,” 1995.

[17] P. K. Sahu and S. Chattopadhyay, “A survey on application mapping strategies for Network-on-Chip design,” Journal of Systems Architecture, vol. 59, no.1, pp. 60–76, 2013.