Regular Mapping Methodology on Network on Chip for Heterogeneous Tasking Environment

HyunJin Kim **Hyejeong Hong** **Sungho Kang**
Department of Electrical and Electronic Engineering
Yonsei University
Seoul, Korea

Byung In Moon
School of Electrical Engineering &
Computer Science, Kyungpook
National University

Hyunjin2.kim@gmail.com, hjhone@soc.yonsei.ac.kr, shkang@yonsei.ac.kr, bihmoon@knu.ac.kr

Abstract – *This paper presents the methodology for the regular mapping of the heterogeneous tasking environment. In the realistic system-on-chip, heterogeneous tasks can be distributed around this system, in which the mapping should be applied while considering their behaviors of the data communication. In this paper, the regular mapping methodology based on the graph coloring algorithm is proposed. Based on the assumption that elements are grouped according to their communication behaviors, each group can be segregated into different colors. For the advantages of the deterministic routing, the mapping is projected into the Latin Square problem, where the ratio of the colored nodes can be regular at row or column. Using the Nostrum NoC simulator, the experiment results show that the proposed mapping methodology reduces both average and maximum latency, and increases the throughput, compared with the previous randomized mapping.*

Keywords: *Network-on-Chip, System-on-Chip, Mapping Methodology, Graph Coloring*

1 Introduction

Due to the advance of semiconductor technology, the huge number of transistors available on a single chip allows designers to integrate tens of intellectual property (IP) blocks together with large amounts of embedded memory. In the near future, it will be soon possible to integrate more than hundreds of microprocessors or IPs on a single chip, not to mention the emergence of the nano-submicron process. Different from the existing resource usage, the proportion of the communication resources has been increased. Moreover, the shrinking process of IPs causes the interconnection delay as the dominant factor of delay, so that system designers may spend a great deal of time on interconnecting the IPs for their dedicated architectures. Both the flexibility and the scalability for SoC are restricted by the interconnect methodology as well. Unfortunately, both current dedicated wiring and shared bus approaches do not overcome these obstacles [1].

In order to solve these complex on-chip communication problems, the regular tile-based network-on-chip (NoC) architecture was recently studied. The regular tile-based NoC is appropriate for designing future SoCs as follows: Firstly, the interconnect model is greatly simplified by its modularity and standard interfaces, which encourages both reusability and interoperability of the IPs. Moreover, since the network platform can be designed in advance and later reused directly with many applications, it is possible to highly optimize this platform as its development cost can be easily amortized across many applications. In the second place, it can be the structured platform for the network communication. Since the network can be mapped on the silicon surface in an expected manner, the latency due to long wire can be managed and optimized, although there is not a significant propagation delay produced by long wires. Lastly, most of NoC architectures have their proposed layered-communication protocols based on OSI reference model. This layered approach can help designers free from annoying managing sequencing and synchronization problems. If the IPs in the library are developed with regularity, the advantages of the regular NoC approach can be further increased.

The future SoCs can consist of many kinds of heterogeneous tasks. In a regular tile-based NoC, each tile points at a regular grid can be mapped as a various IP core. Accordingly, a router is embedded within each tile with the objective of connecting it to its neighboring tiles. One of the issues in a regular NoC design is to map heterogeneous tasks into a regular structure NOC. In spite of the aforementioned advantages of the regular structure, the irregular communication load between the heterogeneous tasks may need a specific methodology for mapping elements for the heterogeneous NoC architecture. In this paper, the regular mapping methodology of the heterogeneous processing elements is proposed. Moreover, the routing algorithm is explained based on the proposed regular mapping in the NoC structure.

The rest of this paper is organized as follows. In section 2, the related works and the concepts for the previous NoC studies are overviewed briefly. Section 3 illustrates our motivation, and section 4 describes the regular mapping using graph coloring considering the routing algorithm. Lastly, section 6 shows the performance using the Nostrum NoC Simulator.

2 Related Research

The mapping methodology for the homogenous tasks has been intensively studied. In the fields of multiprocessor and reconfigurable computing, the array architectures (e.g. mesh-connected array and systolic array) have been used as the target structures [2]. The loop kernels as the portion of an application are mapped into the regular mesh-connected array. However, only switch-based direct interconnection was considered, and the packet-based indirect interconnection was not their concerns due to the complexity of the packet-based indirect model and the limited hardware resources. Studies for applying the heterogeneous tasks to the regular architectures have been conducted [3], in which a parallel program in a heterogeneous network is distributed to minimize the execution time using distributed recursive algorithm based on the switch-based direct interconnection. The united approach to mapping and routing is provided in [4]. This paper presents a united algorithm to couple the mapping and the routing. To realize the energy-aware mapping, the specific solution is issued based on the static XY routing [5]. However, the mapping algorithm in [5] is deterministic, so that it sacrifices both generality and scalability to reach the optimal energy-aware solution. After reviewing the listed papers related to the mapping approaches, the mapping and the routing cannot be independent from each other in the heterogeneous task environment, i.e., it is concluded that the modification of the existing mapping algorithms should be considered along with the effects of the routing for the specific traffic patterns among the heterogeneous tasks.

3 Motivation

In the real world, the increased complexity of communication behaviors between IP blocks is one of the biggest problems for the success of NoC. For example, the ring array with two kinds of IP blocks (Memory and Processor) is shown as an example of the regular array architecture in Figure 1, where the mapping models are classified: the centralized mapping model and the distributed mapping model. In the centralized mapping model, the homogeneous elements are adjacently located, which can increase the local resource sharing and allows the centralized control. However, the non-communication feature between any two memory blocks can give rise to both the long latency and the higher rate of packet congestion. Different from the centralized mapping model, the distributed mapping lowers the rate of congestion at the sacrifice both the centralized control and the local resource sharing. However, not only the complexity of SoC, but also both the diversity and the species of mandatory IP blocks would be increased, which causes the difficulties in the resource sharing. Moreover, the centralized control causes the bottleneck of the performance due to the physical placement constraints. Therefore, the distributed mapping model can be relevant for the NoC architecture, where the packet-based direct communication will be chosen. For the example of the distributed mapping model, the on-chip memory blocks and the processors can represent the different tasks respectively, which mean that the data communication needed by each group of the elements is different.

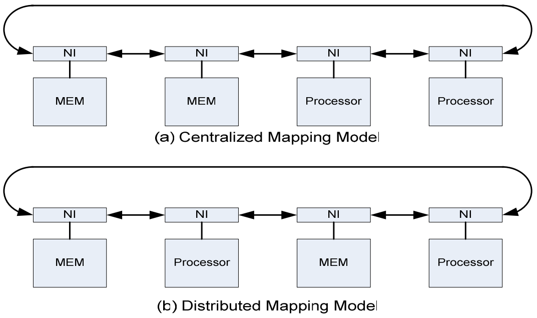

Figure 1. Mapping Model on Ring for Packet-based Routing

For the complex NoC architecture with tens of IP blocks, the traffic patterns cannot be easily estimated and will be dependent on the various applications. Therefore, the modeling with weighting parameters such as the local factor representing the weighted random traffic pattern with the neighborhood [6] can be one of the solutions to acquire the sub-optimal performance in the heterogeneous task environment. To solve this problem, the network interface (NI) of each task mapped onto the element and its interconnect can be abstracted as the vertex and the edge in a graph respectively. Based on the assumption that the tasks can be grouped according to their communication behaviors, each group is assigned to a different color. In the distributed mapping model, the color of the vertex should be different from its neighbor vertices, so that its problem can be solved using the vertex graph coloring: an assignment of the colors to the vertices of the graph, so that the spatial disjointing pattern can be made.

4 Regular Mapping Methodology

The minimum number of colors for the 2D mesh-like architecture is two. Intuitively, as the number of color is increased, its mapping results are approximated as that of the randomly distributed model. Moreover, there are a lot of coloring algorithms for the diverse situations and assumptions as a NP-hard problem. To issue the fundamental rules for the proposed mapping methodology, the routing algorithm for the packet-based communication is considered.

Because of the limited buffering resources available and the stringent latency requirements for typical NoC applications, we believe that wormhole-based routing is the most appropriate routing technique for NoCs. The routing algorithm can be divided into two categories: static routing and adaptive routing. The routing algorithm is related to the traffic pattern, and it is not concluded which is the optimal solution. However, for the regular architecture, the author in [5] asserts that the static routing for the tile-based architecture is more suitable than the adaptive routing considering resource usage, latency, traffic predictability and quality of service. Obviously, XY routing is a minimal path routing algorithm and is free of deadlock and livelock. Moreover, it is suitable for the wormhole routing. After considering the related works, the static XY routing is chosen as the most relevant algorithm for our proposed mapping on-chip network to realize both generality and simplicity. For 2D-mesh networks, XY routing first routes packets along the X-axis. Once it reaches the column wherein lies the destination tile, the packet is then routed along the Y-axis. If the tasks are distributed fairly at each column or row according to their color, the packets go through the regular ratio of the colored nodes based on the XY routing irrespective of their distance. Therefore, some fundamental rules can be shown as follows:

- Number of Color Limitation: For the mesh-like structure, the number of the colors should be limited according to the order of the mesh.

- Regular Ratio of Color: Under XY routing, every packet go through the regular ratio of the colored nodes.

- Distributed X and Y: the colors are distributed in the X-directional and the Y-directional rows, which can guarantees the load balancing. It is based on the assumption that the locally weighted traffic patterns exist between different colored elements.

- Hierarchical Distribution: The color can be assigned hierarchically. This means that the traffic patterns between the elements can be hierarchically modeled. For example, the minimum number of color for mesh-like structure is two, in which sub-coloring problem can be solved using the divide-and-conquer approach.

The first rule means that the upper limit number of colors can be decided by the structure. The last three rules are related to the *Latin Square* and its small *quasigroup*. The Latin Square of order N is comprised of an n-by-n array of N symbols, in which every symbol occurs exactly once in each row and column of the array. The Latin Square involves the use of two sets of blocks, one of which is organized by rows and the other by columns. The Latin Square has the characteristics as the double randomized elements by rows and by columns as shown in the Figure 2 (a). The key of the effective communications comes from the fact that the traffic goes through the randomized blocks by rows and by columns. If the traffic patterns cannot be easily estimated and have the weight random distribution, the regular ratio of the colored nodes can be helpful for increasing the average throughput due to the weighted locality traffic distribution.

The proposed mapping approach is based on the classification of elements into several fixed tycomponents according to their communication features and relationship. However, if the number of the colors is not the same as the order of the mesh, it is hard to group the tasks using the original Latin Square; the modified coloring algorithm can be acceptable for the random traffic distribution with the weighted locality. The main modification of the Latin Square is that a row or a column can contain the same colors, where the ratio of the colored nodes is regular. Especially, the hierarchical Latin Square and quasigroup can be useful to group the tasks into the color.

$$\begin{pmatrix} 0 & 1 & 2 & 3 \\ 1 & 3 & 0 & 2 \\ 3 & 2 & 1 & 0 \\ 2 & 0 & 3 & 1 \end{pmatrix} \qquad \begin{pmatrix} 0 & 1 & 2 & 0 \\ 1 & 2 & 0 & 1 \\ 2 & 0 & 1 & 2 \\ 0 & 1 & 2 & 0 \end{pmatrix}$$

(a) Latin Square Pattern (b) Modified Latin Square Pattern

Figure 2. Latin Square Patterns

The methodology presented in this paper can be outlined as follows.

- Group the arrangement of tasks under the fundamental rules.

- Assign a color into each group, so that the number of adjacent tasks which have the same color is minimized.

- Transform arrangement of element in NoC into a mesh-like array, where vertices represent elements and edges denote the communication path.

- Every column or row can repeat the colors due to the property of the Latin Squares.

In the middle of the color selection of neighboring node, this algorithm can selects an arbitrary color in the case that all possible colors cannot make disjoint coloring. The algorithm of selecting duplicated color can be random the selection or any specific algorithm considering the distribution of colors under Latin Square arrangement.

5 Experimental Results

Our mapping methodology and routing algorithms will be evaluated using the Nostrum Network-on-Chip Simulation Environment (NNSE) [6]. NNSE is a GUI-based tool for simulating network-on-chip with both uniform and locality traffic pattern. The synthetic traffic can be configured to perform system simulation. However, a complicated configuration will be used to simulate our mapping results. For spatial traffic configuration, the traffic pattern can consist of uniform or locality traffic. To realize the weighted random traffic distribution, the locality factor can be specified, in which the communication distribution probability is related to the distance between the elements, so that the proposed mapping methodology can be simulated using the locality factor. In the 4x4 mesh structure, the XY routing is chosen along with the 36-bit link bandwidth, and the wormhole-based 4-way virtual channel with two buffers. For the spatial distribution, each locality factor is calculated for both the random mapping and the example of the proposed mapping as a form of Latin Square, where the ratio for the nodes with the same color is reflected in the locality factor. Figures 3 show the maximum latency and the average latency, respectively.

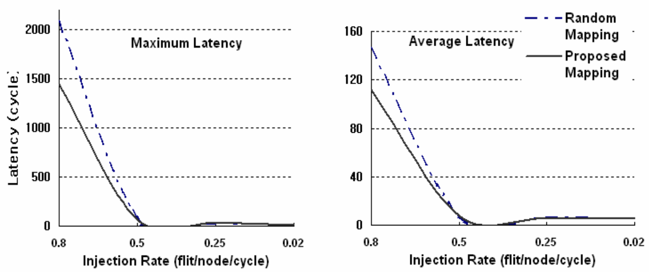

Figure 3. Performance with Mapping Classes

For the highly injected flits, the effect of the proposed mapping is increased due to the weighted locality traffic distribution of the double randomized element. In Figure 4, the throughputs are estimated, so that the throughput of the proposed mapping is slightly increased. As the injection rate is more increased, the ratio of the increased throughput is higher. Moreover, after our experiments for the highly ordered mesh structure, it can be concluded that the effect of the proposed mapping has been increased greatly.

6 Conclusions

In this paper, the regular mapping methodology for the heterogeneous tasks is proposed on the XY routing. When the elements in the mesh-like structure are configured as a form of the Latin Square on their communication behaviors, the double randomized elements provide the weighted locality traffic distribution for XY routing.

Though the used traffic pattern for the heterogeneous tasks is just ideally the weighted random traffic for each element, it is sufficient to prove the efficiency of the proposed mapping methodology based on the graph coloring. Therefore, the proposed mapping under NoC structure is seriously chosen with the routing algorithm and the communication behaviors.

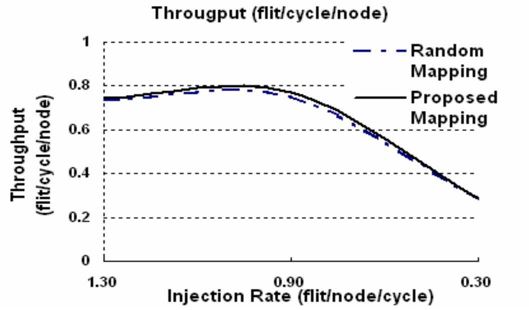

Figure 4. Performance with Throughput

Acknowledgments

This research was supported by University IT Research Center Project of Korea.

References

[1] C. Wu et al., “Mapping of IP Cores to Network-on-Chip Architectures Based on Communication Task Graphs,” in Proc. of Int. Conf. on ASIC, 2005, pp.874-877.

[2] Chuang., H. Y. H, “Mapping loop algorithms into reconfigurable mesh connected processor array,” in Proc. of Hawaii Int. Conf. on Architecture Track, 2006, pp. 294-300.

[3] Natalya Vodovoz, “Mapping Heterogeneous Task Graph onto Network: Execution Time Optimization”, Lecture Notes in Computer Science, 2001, Vol. 2127, pp. 142-149.

[4] Hansson, A. et al., “A Unified Approach to Constrained Mapping and Routing on Network-on-Chip Architectures,” Proc. of the IEEE Int. Conf. on Hardware/Software Codesign and System Synthesis, 2005, pp.75-80.

[5] Hu, J. and Marculescu, R., “Energy- and Performance-Aware Mapping for Regular NoC Architectures”, IEEE Trans. On CAD, 2005, Vol. 24, pp.551-562.

[6] Lu, Z., “A User Introduction to NNSE: Nostrum Network-on-Chip Simulation Environment”, http://www.imit.kth.se/info/FOFU/Nostrum/NNSE/