Bandwidth/Fault tolerance/Contention Aware Application-Specific NoC Using PSO as a Mapping Generator

Atena Roshan Fekr[1], Ahmad Khademzadeh[2], Majid Janidarmian[1], Vahhab Samadi Bokharaei[3]

Abstract— **This paper presents a novel and efficient method to produce different mappings, based on Particle Swarm Optimization (PSO) evolutionary algorithm. It produces a lot of mappings with different metrics which totally evaluated by a total cost function. Total cost function helps to find the optimal application-specific Network-on-Chip (NoC) based on designer's decisions to how customize and prioritize the impact of three parameters on special mapping. Three mentioned parameters are communication cost, robustness index and contention factor. Communication cost is a common metric in evaluation of different mapping algorithms which have direct impact on power consumption and performance of mapped NoC. Robustness index is used as a criterion for evaluating fault tolerant properties of NoC. Contention Factor is another performance metric, highly affects the latency, throughput and communication energy consumption. The experimental results reveal the power of proposed procedure which is mainly focused in generating different solutions that speared through the explored design space.**

Index Terms—**Network-on-Chip, Mapping, Particle Swarm Optimization, Communication Cost, Robustness Index, Contention Factor**

I. Introduction

Due to ever-increasing complexity of system on chip (SoC) design, and non-efficiency of electric bus to exchange data between IP cores in giga scale, the Network on Chip (NoC) is presented with more flexible, scalable and reliable infrastructure. Different mapping algorithms for NoCs are presented to decide which core should be linked to which router. Mapping an application to on-chip network is the first and the most important step in the design flow as it will dominate the overall performance and cost [1]. The main purpose of this article is to present a new method to produce different mappings with all reasonable ranges of communication cost. Then by using a linear function, the most appropriate mapping among produced mappings is selected by designer. The designer decisions must satisfy the three key parameters, i.e., communication cost, robustness index and contention factor. The proposed procedure is shown in Fig.1 and explained in the next sections.

Albeit the proposed approach is topology-independent, it is illustrated and evaluated for 2D mesh topology as it is widely used for most mapping algorithms.

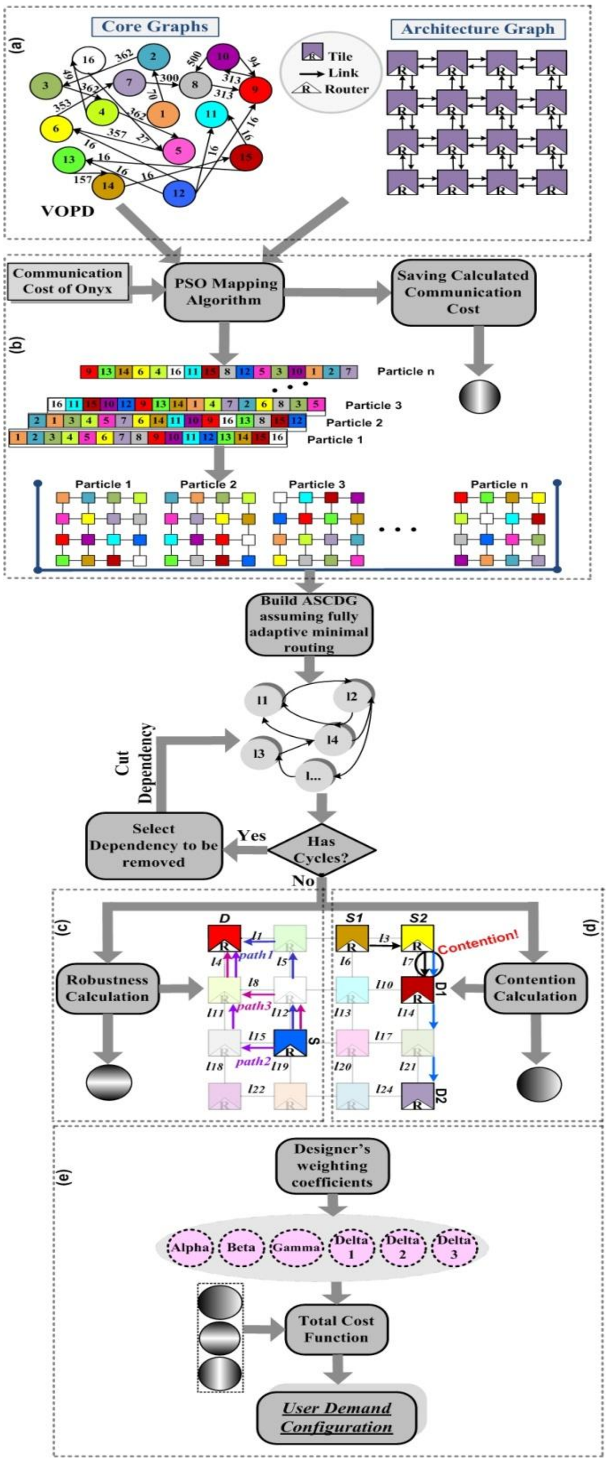


Fig.1: The proposed procedure to achieve the optimal application-specific Network-on-Chip

[1] CE Department, Science and Research Branch, Islamic Azad University, Tehran, Iran {a.roshan, jani}@srbiau.ac.ir

[2] Iran Telecommunication Research Center, Tehran, Iran, zadeh@itrc.ac.ir

[3] ECE Department, Shahid Beheshti University, Tehran, Iran, v.samadi@sbu.ac.ir

II. Mapping Problem and communication cost

To formulate mapping problem in a more formal way, we need to first introduce the following two concepts borrowed from [2]:

Definition 1: The core graph is a directional graph, $G(V,E)$, whose each vertex, $v_i \in V$ shows a core, and a directional edge, $e_{i,j} \in E$ illustrates connection between v_i and v_j. The weight of $e_{i,j}$ that is shown as $comm_{i,j}$, represents the bandwidth requirement of the communication from v_i to v_j. We display an IP core along with a router connected to it by Resource Network Interface (RNI) as a tile.

Definition 2: The NoC architecture graph is a directional graph, $A(T,L)$, whose each vertex, $t_i \in T$, represents a tile in the NoC architecture, and its directional edge that is shown by $l_{i,j} \in L$ shows a physical link from t_i to t_j and for each $l_{i,j}$ a $BW(l_{i,j})$ is considered. $r_{i,j}$ denotes the routing path from t_i to t_j and $L(r_{i,j})$ is the set of links that make up the path $r_{i,j}$. The definitions are presented in Fig.1 (a).

In core graph each edge is treated as a flow of single commodity, represented as c^k and its value which indicates required bandwidth for each edge is shown with $vl(c^k)$. The set of all commodities represented as C is achieved as Eq. (1):

$$C = \begin{Bmatrix} c^k : vl(c^k) = comm_{i,j}, k = 1,2,\dots |E|, \forall e_{i,j} \in E, \\ with\, source(c^k) = map(v_i),\ \ dest(c^k) = map(v_j) \end{Bmatrix} \quad (1)$$

The core graph mapping $G(V,E)$ on NoC architecture graph $A(T,L)$ is defined by a one to one mapping function (Eq. (2)).

$$map : V \rightarrow T, s.t. map(v_i) = t_j, \forall\, v_i \in V, \exists t_j \in T, |V| \leq |T| \quad (2)$$

Communication cost is calculated according to the Eq. (3):

$$commcost = \sum_{k=1}^{|E|} vl(c^k) \times hop\_count\left(src(c^k), dst(c^k)\right) \quad (3)$$

where $src(c^k)$ is the source and $dst(c^k)$ is the destination of c^k.

III. PARTICLE SWARM OPTIMIZATION AS A MAPPING GENERATOR

Many mapping algorithms have been recently proposed to improve several parameters used in the NoC design. One of the most important parameters is the communication cost. There are several available mapping algorithms which are considered to minimize the communication cost. Using small hop counts between related cores will significantly drop the communication cost. Moreover, small hop counts will reduce the energy consumption and other performance metrics like latency [2]. It can be explained that reduction of hop counts can decrease the fault tolerant properties of NoC. Therefore, the optimal solution is to minimize the communication cost while maximizing the fault tolerant properties of NoC. In this paper, particle swarm optimization (PSO) algorithm is used to achieve the optimal solution.

As a novel population-based swarm intelligent technique, PSO simulates the animal social behaviors such as birds flocking, fish schooling, etc. Due to the simple concept and ease implementation, it has gained much attention and many improvements have been proposed [3].

In a PSO system, multiple candidate solutions coexist and collaborate simultaneously. Each solution, called a ”particle”, flies in the problem space according to its own ”experience” as well as the experience of neighboring particles. Different from other evolutionary computation algorithms, in PSO, each particle utilizes two information indexes: velocity and position, to search the problem space. The velocity information predicts the next moving direction, as well as the position vector is used to detect the optimum area. In standard particle swarm optimization, the velocity vector is updated as follows:

$$\begin{aligned} &v_{jk}(t+1) = \\ &w_t v_{jk}(t) + c_1 r_1 \left(p_{jk}(t) - x_{jk}(t)\right) + \ c_2 r_2 \left(p_{gk}(t) - x_{jk}(t)\right), \\ &w_{t+1} = w_t \times w_{damp} \end{aligned} \quad (4)$$

Where $v_{jk}(t)$ and $x_{jk}(t)$ represent the kth coordinates of velocity and position vectors of particle j at time t, respectively. $p_{jk}(t)$ means the kth dimensional value of the best position vector which particle j had been found, as well as $p_{gk}(t)$ denotes the corresponding coordinate of the best position found by the whole swarm. Inertia weight, w_t, cognitive coefficient, c_1, and social coefficient, c_2, are three parameters controlling the size of velocity vector. r_1 and r_2 are two random numbers generated with normal distributions within interval [0,1]. With the corresponding velocity information, each particle flies according to the following rule (Eq. (5)) [3]. This concept is shown in Fig.2:

$$x_{jk}(t+1) = x_{jk}(t) + v_{jk}(t+1) \quad (5)$$

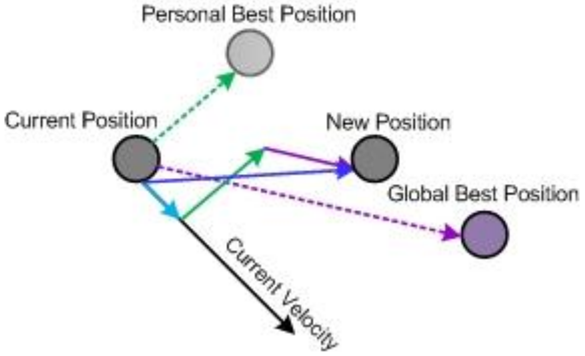


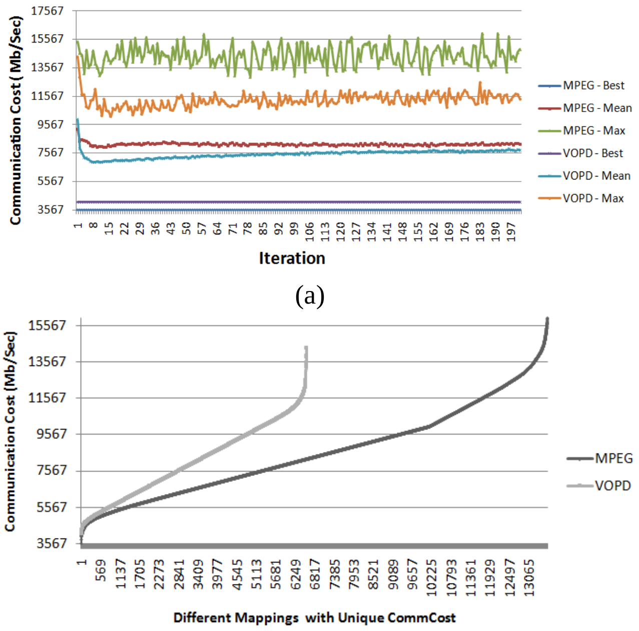
Fig.2: Particle Swarm Optimization algorithm

It is worth mentioning that onyx is one of the best mapping algorithms in terms of communication cost as it results in a fraction of second. By having the onyx result and knowing evolutionary nature of PSO algorithm, different mappings with all reasonable ranges of communication cost can be obtained. To do this, onyx result is injected into population initialization step as a particle as shown in Fig.1 (b).

In order to avoid rapid convergence, velocity threshold is not defined and c_1, c_2, w_0 and w_{damp} are set to 3.49, 7.49, 1 and 0.99 respectively in the proposed PSO algorithm. These values were obtained by examining several simulations because they drastically affect on the diversity of results.

IV. EXPERIMENTAL RESULTS OF MAPPING GENERATOR

The real core graphs, VOPD and MPEG-4 [2], are used in the proposed PSO algorithm. The proposed PSO algorithm was run with 1000 initial population using 200 iterations. Fig. 3 (a) indicates the minimum, mean and maximum fitness function values in each iteration. As shown in Fig.3 (b), it is clear that our PSO algorithm could generate different mappings of VOPD and MPEG-4 core graphs with all reasonable ranges of communication cost because of mentioned convergence control. There are 119,912 and 156,055 different unique mappings for VOPD and MPEG-4 core graphs respectively. It is worth noting that this method, which is presented for the first time in this article, enables the designer to consider other important key parameters as well.



(b)

Fig.3: (a) minimum, mean and maximum fitness function values for VOPD and MPEG-4 core graphs, (b) ability of the proposed mapping generator in producing mappings with all reasonable ranges of communication cost

V. ROBUSTNESS INDEX

Robustness index is considered as a criterion for estimating fault tolerant properties of NoCs [4]. The greater the robustness index, the more fault tolerant NoC design. The robustness index, RI, is based on the extension of the concept of path diversity [5]. For a given communication, $c^k \in C$, an NoC architecture graph, $A(T, L)$, a mapping function, M, and a routing function, R, [4] defined the robustness index for communication c^k , $RI(c^k)$, as the average number of routing paths available for communication, c^k, if a link belonging to the set of links used by communication c^k is faulty. Formally,

$$RI(c^k) = \frac{1}{|L(c^k)|} \sum_{l_{i,j} \in L} |\rho(c^k) \backslash \rho(c^k, l_{i,j})| \qquad (6)$$

where, $\rho(c^k)$ is the set of paths provided by R for communication, c^k, $\rho(c^k, l_{i,j})$ is the set of paths provided by R for communication, c^k,that uses link $l_{i,j}$, and $L(c^k)$ is the set of links belonging to paths in $\rho(c^k)$.

Suppose that there are two routing functions, A and B, which routing function A selects *path 1* and $path2$ and routing function B selects $path2$ and $path3$ to route packets between source and destination as shown in Fig. 1 (c) . The routing function A selects two disjoint paths such that the presence of a faulty link in one path dose not compromise communication from source to destination since another path is fault-free. However, when the routing function B is used as shown in Fig. 1 (c), the communication will not occur. As the alternative paths share the link, l_4 any fault in the link, l_4 makes the communication from “source” to “destination” impossible. Consequently, the NoC which uses routing function A, NOC_1, is more robust than the NoC which uses routing function B, let call it NOC_2. Such situation is reflected by the robustness index. The robustness index for the above two cases are:

$$RI^{(NOC_1)}(\text{source} \rightarrow \text{destination}) = \frac{1+1+1+1+1+1}{6} = 1\,,$$
$$RI^{(NOC_2)}(\text{source} \rightarrow \text{destination}) = \frac{0+1+1+1+1}{5} = 0.8\,.$$

The NOC_1using $path1$ and $path2$ is more robust than the NOC_2 using $path2$ and $path3$ for communication from “$source$” to “$destination$” as $RI^{(NOC_1)} > RI^{(NOC_2)}$.

The global robustness index, which characterizes the network, is calculated using the weighted sum of the robustness index of each communication. For a communication, c^k, the weight of $RI(c^k)$ is the degree of adaptivity [6] of c^k. The degree of adaptivity of a communication, c^k, is the ratio of the number of allowed minimal paths to the total number of possible minimal paths between the source node and the destination node associated to c^k. The global robustness index is defined as Eq. (7).

$$RI^{(NOC)} = \sum_{c^k \in C} \alpha(c^k) RI^{(NOC)}(c^k) \qquad (7)$$

where $\alpha(c^k)$ indicates the degree of adaptivity of communication c^k.

In this paper, one of the best algorithms which is customized for routing in application-specific NoCs, is used. The algorithm was presented in [7] which uses a highly adaptive deadlock-free routing algorithm. This routing algorithm has used Application-Specific Channel Dependency Graphs (ASCDG) concept to be freedom of dead-lock [8].

VI. CONTENTION FACTOR

In [9] a new contribution consist of an integer linear programming formulation of the contention-aware application mapping problem which aims at minimizing the inter-tile network contention was presented. This paper focuses on the network contention problem; this highly affects the latency, throughput and communication energy consumption.

The source-based contention occurs when two traffic flows originating from the same source contend for the same links. The destination based contention occurs when two traffic flows which have the same destination contend for the same links. Finally the path-based contention occurs when two data flows which neither come from the same source, nor go towards the same destination contend for the same links somewhere in the network.

The impact of these three types of contention was evaluated and observed that the path-based contention has the most significant impact on the packet latency. Fig. 1 (d) shows the path-based contention. So, in this paper we consider this type of contention as a factor of mappings. More formally:

$$Contention\ Factor = \sum_{\forall e_{i,j} \in E} \left| L\left(r_{map\,(v_i),map\,(v_j)}\right) \cap L\left(r_{map\,(v_k),map\,(v_l)}\right) \right| \qquad (8)$$

$$for\ i \neq k\ and\ j \neq l$$

By having communication cost, robustness index and contention factor for each unique mapping, the best application-specific Network on Chip configuration should be chosen regarding to designer's wise decisions.

VII. Making Wise Decisions

As previously mentioned, lower communication cost leads to an NoC with better metrics such as energy consumption and latency. Other introduced metrics were robustness index which is used as a measurable criterion for fault tolerant properties and contention factor which has the significant impact on the packet latency. A total cost function is to be introduced in order to minimize the sum of weighted these metrics (Fig. 1 (e)). The total cost function is introduced as follows:

$$Total\ Cost\ Function = \qquad (9)$$

$$\text{Min}\left(\frac{\delta_1}{\alpha} \times commcost_i + \frac{\delta_2}{\beta} \times \left(-RI_i^{(NOC)}\right) + \frac{\delta_3}{\gamma} \times CF_i\right)$$
$$\forall\ mapping_i\ \in generated\ mappings$$

$$and\ \delta_1 +\ \delta_2 +\ \delta_3 = 1$$

where, $commcost_i$ is the communication cost, $RI_i^{(NOC)}$ is the robustness index and CF_i is the contention factor of NoC after applying $mapping_i$.

The constants α , β and γ are used to normalize the $commcost$, $RI^{(NOC)}$ and CF. In this paper, α, β and γ are set to the maximum obtained values for communication cost, robustness index and contention factor. δ_1, $\delta_2$and δ_3 are the weighting coefficients meant to balance the metrics. Although multi-objective evolutionary algorithms are other ways to solve this problem, the proposed procedure is better due to following reasons.

First, it does not need that to be executed again if designer wants to change values of weighting coefficients. Second, if designer focuses on communication cost, multi-objective algorithms usually are not able to get the best result. And finally, because of convergence control, the diversity of results is more than multi-objective algorithms and easily goes up with increasing population size or iteration.

VIII. Final Experimental Results

In order to better investigate the capabilities of proposed procedure shown in Fig.1, we have done some experiments on real core graphs VOPD and MPEG-4. As mentioned before, one of the advantages of proposed mapping generator is its diversity of produced solutions. Based on the experimental results, mentioned mapping generator produces 221,000 mappings for VOPD and MPEG-4, according to boundaries which limit population size and maximum iteration of PSO algorithm. Dismissing the duplicate mappings led to 119,912 and 156,055 unique mappings for VOPD and MPEG-4 which extracted among whole results. Results of running this procedure for VOPD and MPEG-4 core graphs and evaluating the values in the 3D design space are shown in Fig.4 to Fig.11.Values of δ_1, δ_2 and δ_3 which used in these experiments respectively are 0.5, 0.3 and 0.2 for VOPD core graph and 0.1, 0.2 and 0.7 for MPEG-4 core graph.

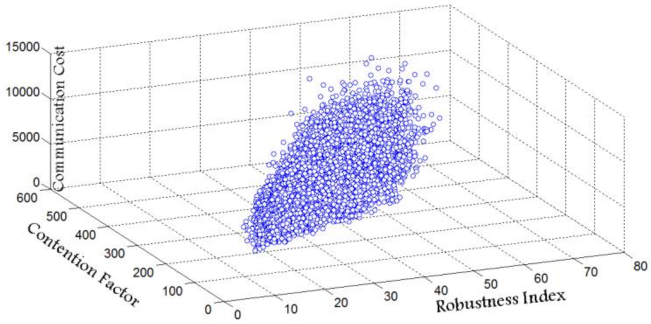


Fig.4: Robustness Index, Contention Factor and Communication Cost of VOPD mappings in 3D design space

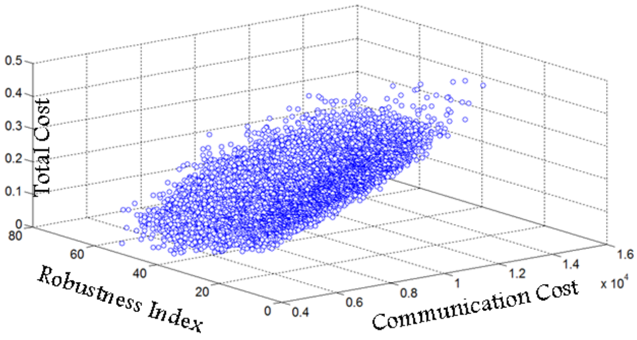


Fig.5: Communication Cost, Robustness Index and Total Cost of VOPD mappings in 3D design space

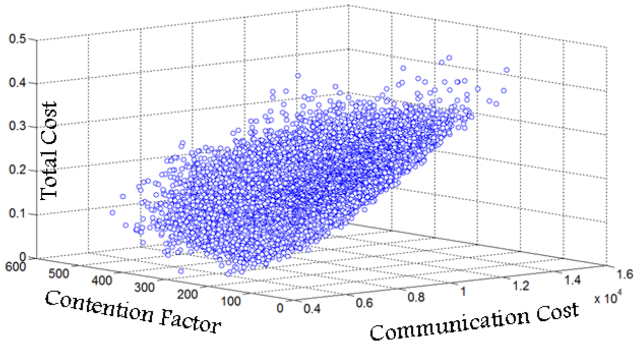


Fig.6: Communication Cost, Contention Factor and Total Cost of VOPD mappings in 3D design space

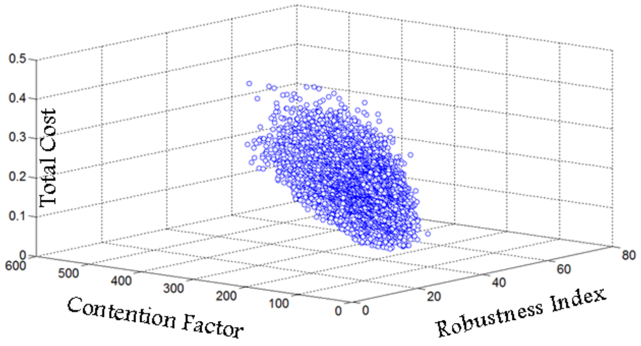


Fig.7: Robustness Index, Contention Factor and Total Cost of VOPD mappings in 3D design space

As it can be seen in these figures, there are many different mappings which have the equal communication cost value that is one of the good points about proposed mapping generator. In average, there are almost 18 and 12 different mappings for each special value of communication cost while VOPD and MPEG-4 are considered as experimental core graphs. The optimal application-specific NoC configuration can be selected by setting proper values in total cost function based on designer demands. In our design, VOPD mapping with communication cost, 4347, robustness index, 54.28, and contention factor, 284, is the optimal solution. Mapping with communication cost, 6670.5, robustness index, 35.94, and contention factor, 6, is also the optimal solution for MPEG-4 mapping.

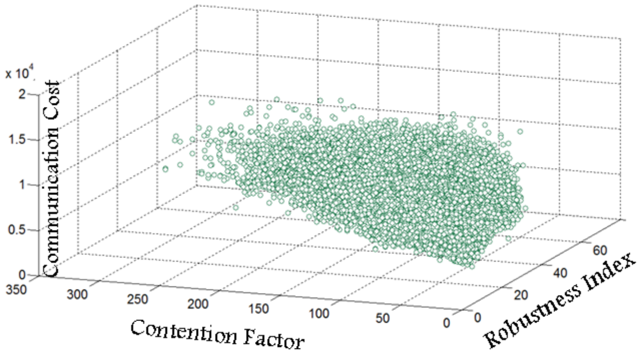


Fig.8: Robustness Index, Contention Factor and Communication Cost of MPEG-4 mappings in 3D design space

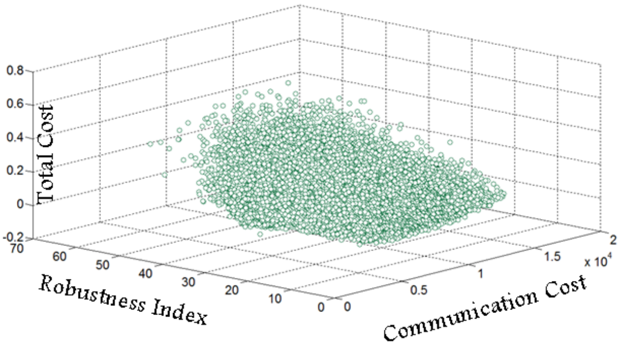


Fig.9: Communication Cost, Robustness Index and Total Cost of MPEG-4 mappings in 3D design space

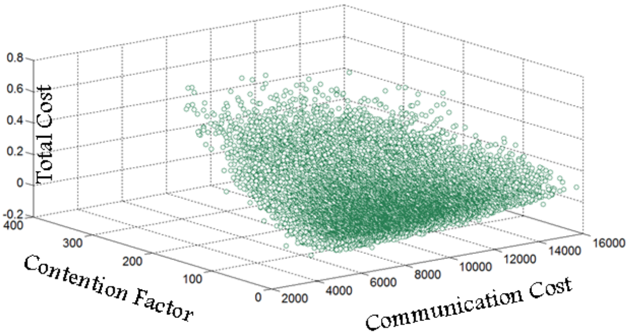


Fig.10: Communication Cost, Contention Factor and Total Cost of MPEG-4 mappings in 3D design space

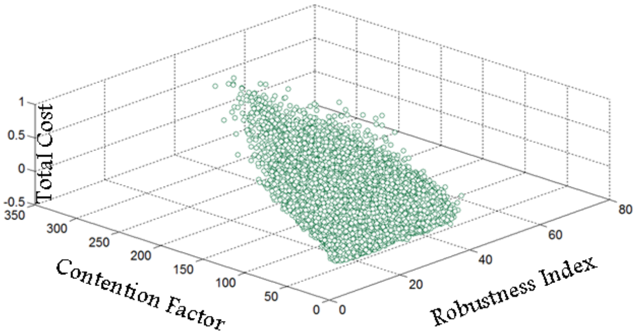


Fig.11: Robustness Index, Contention Factor and Total Cost of MPEG-4 mappings in 3D design space

IX. Conclusion

As mapping is the most important step in Network-on-Chip design, in this paper a new mapping generator using Particle Swarm Optimization algorithm was presented. The best mapping in terms of communication cost was derived from Onyx mapping algorithm and injected into population initialization step as a particle. Because of using Onyx mapping results as particles, results convergence was controlled by finding appropriate values in velocity vector. This PSO algorithm is able to generate different mappings with all reasonable ranges of communication cost. Using 3 metrics which are communication cost, robustness index and contention factor for each unique mapping, the best application-specific Network-on-Chip configuration can be selected regarding to designer's demands that are applied onto total cost function.

References

[1] W. Shen, C. Chao, Y. Lien, A. Wu," A NEW BINOMIAL MAPPING AND OPTIMIZATION ALGORITHM FOR REDUCED-COMPLEXITY MESH-BASED ON-CHIP NETWORK,", Networks-on-Chip, NOCS 2007, pp.317 – 322, 7-9 May 2007.

[2] M. Janidarmian, A. Khademzadeh, M. Tavanpour, "Onyx: A new heuristic bandwidth-constrained mapping of cores onto tile based Network on Chip", IEICE Electron. Express, Vol. 6, No. 1, pp.1-7, January 2009.

[3] Zhihua CUI, Xingjuan CAI, Jianchao ZENG," choatic performance-dependant particle swarm optimazation" International Journal of Innovative Computing, Information and Control (IJICIC) , Vol. 5, No. 4, pp. 951-960, April 2009.

[4] Rafael Tornero, Valentino Sterrantino, Maurizio Palesi ,Juan M. Orduna," A Multi-objective Strategy for Concurrent Mapping and Routing in Networks on Chip" Proceedings of the 2009 IEEE International Symposium on Parallel&Distributed Processing",pp. 1-8 , 2009.

[5] W. J. Dally and B. Towles, Principle and Practice of Interconnection Network. San Francisco, CA : Morgan Kaufmann, 2004.

[6] C. J. Glass L. M. Ni, “The turn model for adaptive routing,” Journal of the Association for Computing Machinery, vol. 41, no. 5, pp. 874-902, Sep. 1994.

[7] M. Palesi, G. Longo, S. Signorino, R. Holsmark, S. Kumar, V. Catania, “Design of Bandwidth Aware and Congestion Avoiding Efficient Routing Algorithms for Networks-on-Chip Platforms”, Networks-on-Chip, NoCS 2008. Second ACM/IEEE International Symposium on, pp. 97 – 106, April 2008.

[8] M. Palesi, R.Holsmark, S.Kumar, “a methodology for design of application specific deadlock-free routing algorithms for NoC systems”, Hardware/Software Codesign and System Synthesis, CODES+ISSS '06. Proceedings of the 4th International Conference,pp. 142-147, Oct. 2006.

[9] C. Chou, R. Marculescu,” Contention-aware application mapping for Network-on-Chip communication architectures” Computer Design, 2008. ICCD 2008. IEEE International Conference on,pp 164 – 169, 19 January 2009 .