Improving Topological Mapping on NoCs

Rafael Tornero (PhD Student) and Juan M. Orduña Huertas (PhD Advisor)
Departament d'Informàtica, Universitat de València, Spain

Abstract

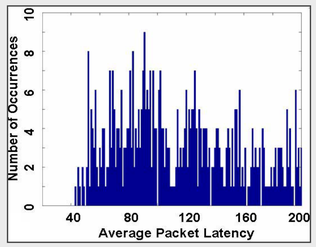
Networks-on-Chip (NoCs) have been proposed as an efficient solution to the complex communications on System-on-chip (SoCs). The design flow of network-on-chip (NoCs) include several key issues, and one of them is the decision of where cores have to be topologically mapped. This thesis proposes a new approach to the topological mapping strategy for NoCs. Concretely, we propose a new topological mapping technique for regular and irregular NoC platforms and its application for optimizing application specific NoC based on distributed and source routing.

Description of the Problem

The Network-on-Chip (NoC) paradigm has emerged recently as a promising solution to the complex on-chip communications derived from the increasing number of Intellectual Property (IP) cores that can be integrated on a single chip as semiconductor technology scales down to make complex Systems-on-Chip (SoCs) [1].

The use of NoCs as the interconnection infrastructure for complex SoCs has opened several interesting research and design issues [2]. These include: topology selection, routing strategy selection and **application mapping**.

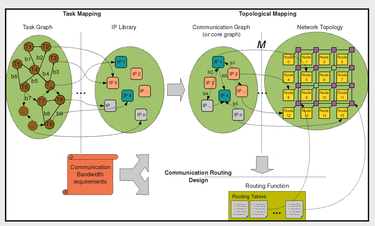
Impact of Mapping on Performance

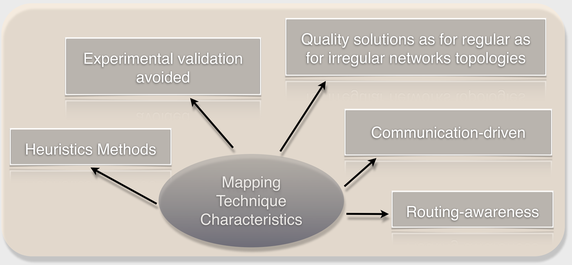


- A/V Multimedia system mapped on 16 IPs
- Average packet latency of 3000 random mappings
- Results for the top 478 mappings
- The remaining 2522 mapping have latency much higher than 200 clock cycles

What does Topological Mapping Consist of?

- Splitting the application on a set of concurrent tasks, assigning and scheduling on a list of IP cores (by procedures beyond the scope of this thesis)
- Deciding how to topologically place the selected set of cores onto the tiles of the system such that the interested metrics (energy, latency, etc) are optimized



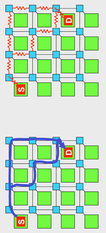


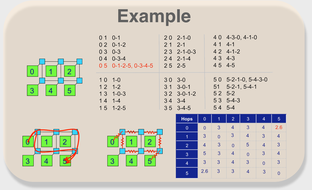
Proposed Approach

The Topological Mapping Technique [3]

Three key issues:

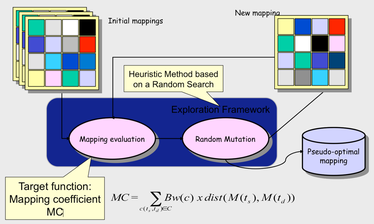
1.Modeling the network as a table of distances (or costs), model of communication cost, between each pair of source/destination nodes. This table contains the cost of communicating every pair of network nodes without taking into account any traffic pattern, but taking into account the underlying network topology and routing algorithm.





2.For each pair of source/destination IPs, the amount of information exchanged by them is measured, and a table of communication requirements of the application is computed.

3.In order to search the best assignment of IPs to network nodes, the technique defines a mapping array and an associated quality function for each array. This quality function measures the overall cost for sending all the messages exchanged by the application (exchanged by all the tasks hosted in the IPs) if the IPs were mapped as indicated in the mapping array. Finally, this approach consists of using a heuristic method based on a random search for obtaining a near-optimal assignment of IPs to network nodes (topological mapping) that minimizes the quality function value.

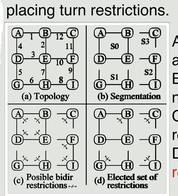


Application Specific NoC Platforms Based on Distributed Routing

Communication-Aware Routing Technique (CART) [4]

- Optimizes the network performance for Application-Specific NoCs.
- Combines a flexible, topology-agnostic routing algorithm (SR) with the communication-aware mapping technique described above.

The Segment-based Routing (SR) algorithm [5] is based on a segmentation process of the network. It guarantees deadlock freedom and full connectivity among nodes by placing a bidirectional routing restriction at every segment. This results in a larger degree of freedom when placing turn restrictions.

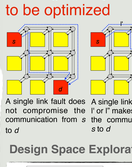


A. Topology formed by links and nodes
B. Partition in different network segments
C. All possible bidirectional routing restrictions
D. Random set of routing restriction elected

The first step consists of computing a random SR configuration (combination of segments and routing restrictions applied to the network topology). Then, the communication-aware mapping technique is applied to that SR configuration.

Multi-Objective Strategy for Concurrent Mapping and Routing [6]

This strategy is able to determine an approximation of the Pareto-set of NoC configurations which optimize multiple performance indexes simultaneously. We have considered the routing algorithm designed by APSRA [7] and the topological mapping as the multidimensional interacting parameters. We have considered message delay and fault tolerance as the objectives to be optimized



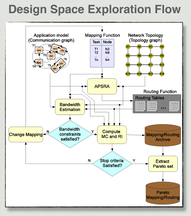
Robustness Index

$$RI(c) = \frac{1}{|L(c)|} \sum_{\forall l \in L(c)} |P(c) \setminus P(c,l)|$$

$$R(s \to d) = \frac{1+1+1+1+1+1+1+1}{8} = 1$$

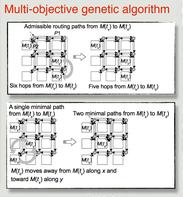
$$R(s \to d) = \frac{0+1+1+1+1+0}{6} = 0.67$$

Design Space Exploration Flow

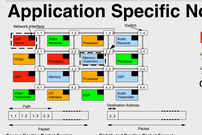


How mapping is changed

Multi-objective genetic algorithm



Application Specific NoC Platforms using Source Routing



Distance constrained mapping [8] to:

- Minimize the overhead
- Support NoC platforms using source routing with a constraint in the path length field of the header packet

Advantages:

- Simple and fast routers
- Topology independence
- Guaranteed throughput
- Mixing of minimal and non-minimal paths

Disadvantages:

- Overhead and low bandwidth utilization
- Static routing

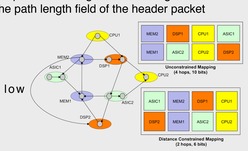
The method consists of two steps:

1.Find a mapping using the technique described above, but with a distance constraint added.

$$dist(M(c_i), M(c_j)) \leq Threshold$$

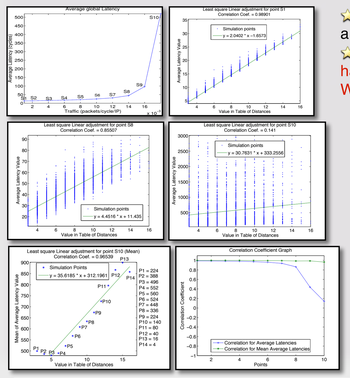
2.Find a source routing function for every core pair C_i and C_j where:

- The path length is equal to the manhattan distance between C_i and C_j
- There is no possibility of deadlock
- The traffic is balanced among all the links



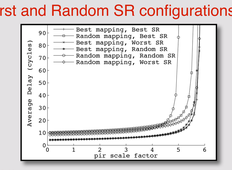
Results

Validation of the Network Model



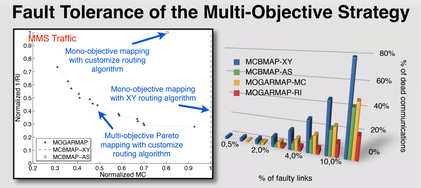
Analysis of CART

- 4390 different routing algorithms for a 5x5 mesh have been constructed .
- Best, Worst and Random mappings have been combined with the Best, Worst and Random SR configurations.



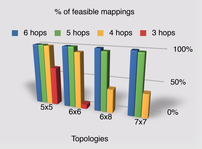
All plots corresponding to Best Mappings show a similar stable throughput. Additionally, the Best Mappings significantly reduce the average network delay, regardless of the SR configuration used.

Fault Tolerance of the Multi-Objective Strategy



Distance Constrained Mapping. Feasibility exp.

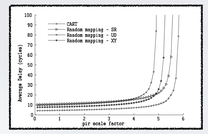
- Mapping of 500 random applications
- Uniform probability of spatial communications
- Communication bandwidth uniformly distributed between 10 and 100 KB/s



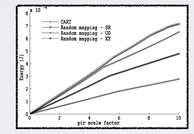
Performance Evaluation

CART

- Performance evaluation of the CART method compared to other routing techniques that do not take into account the traffic generated by the application.
- Simulation of the 5x5 2D-mesh topology with different applications when using Up/Down (UD) routing, Dimension Order (XY) routing, SR and CART.



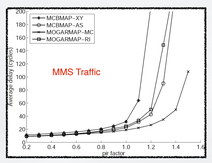
MMS - CART provides more than 30% improvement in latency and 15% higher throughput than the rest of routing techniques.



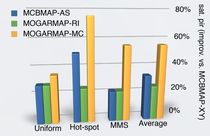
MMS - energy required for CART is more than twice lower than the energy required by the rest of the routing techniques considered.

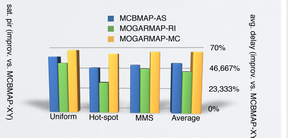
The evaluation results show that the CART method can significantly improve the network performance for those application specific NoCs where the communication requirements of the application can be measured or estimated.

Multi-Objective Strategy



- For all traffic scenarios and low pirs MCBMAP-XY performs slightly better than the other.
- For higher values of pir, the best solution in terms of MC found by the proposed approach outperforms the others.
- The best solution in terms of RI found by the proposed approach, exhibits delay characteristics close to MCBMAP-AS.



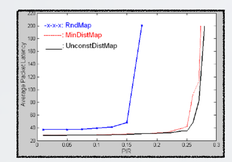


Average delay has been measured at a packet injection rate where none of the algorithms are saturated.

Distance Constrained Mapping for NoC Platforms using Source Routing

The figure shows the average latency in cycles. It shows that at both low traffic loads and high traffic loads both the UNCONSTDISTMAP and MINDISTMAP present similar behavior and save more than 20% cycles and 25% over RNDMAP respectively

Therefore, the evaluation results show that is possible to reduce the path length of the header flit at least to half of the network diameter without significantly degradation of the performance, less than 5%



Remaining Objectives

Currently, we are performing a study of heuristic methods for adapting the mapping technique to NoC platforms using source routing. This kind of platforms introduce new constraints to the mapping process. Thus, the heuristic search methods employed in distributed routing NoC platforms might not be suitable for them. Therefore, the goal of this study is to find the most suitable search method for satisfying the new constraints of source routing NoC platforms.

Also, we plan to apply the technique for optimizing applications in shared memory MPSoCs, where communications are carried out by synchronization mechanism. Finally, we would like to evaluate our technique in NoCs with irregular topologies. The reason is that most of the topological mapping proposals in the literature use NoCs with regular topology, leaving the mapping on NoCs with irregular topology as a currently open issue.

References

[1] A. Jantsch and H. Tenhunen, Eds., Networks on Chip. Kluwer Academic Publishers, Feb. 2003.
[2] R. Marculescu, J. Hu, and U. Ogras, "Key research problems in noc design: a holistic perspective," in Proc. of CODES+ISSS '05, 2005, pp. 69–74.
[3] R. Tornero, J. M. Orduña, M. Palesi, and J. Duato, "A communication-aware topological mapping technique for nocs," in Proc. of the 14th international Euro-Par conference on Parallel Processing. Springer-Verlag, 2008, pp. 910-919.
[4] R. Tornero, J. M. Orduña, A. Mejia, J. Flich, and J. Duato, "Cart: Communication-aware routing technique for application-specific nocs," in Proc. 11th EUROMICRO Conf. on Digital System Design Architectures, Methods and Tools, 3–5 Sept. 2008, pp. 26–31.
[5] A. Mejia, J. Flich, J. Duato, S.-A. Reinemo, and T. Skeie, "Segment-based routing: an efficient fault-tolerant routing algorithm for meshes and tori," in Proc. 20th International Parallel and Distributed Processing Symposium IPDPS 2006, 25–29 April 2006, p. 10pp.
[6] R. Tornero, V. Sterrantino, M. Palesi, and J. M. Orduña, "A multi-objective strategy for concurrent mapping and routing in network on chip," in Proc. of the 23rd IEEE International Parallel and Distributed Processing Symposium, 25–29 May 2009.
[7] M. Palesi, R. Holsmark, S. Kumar, and V. Catania, "Application specific routing algorithms for networks on chip," IEEE Transactions on Parallel and Distributed Systems, vol. 20, no. 3, pp. 316–330, 2009.
[8] R. Tornero, S. Kumar, S. Mubeen, and J. M. Orduña, "Distance constrained mapping to support noc platforms based on source routing," in Proc. of the Euro-Par 2009 Workshops - Parallel Processing. Springer- Verlag, 25 August 2009, pp. To be published, January 2010.

24th IEEE International Parallel and Distributed Processing Symposium
April 19-23, 2010
ATLANTA (Georgia) USA