Adaptive Networks for On-Chip Communication

I

Mário Pereira Vestias
Instituto Politécnico de Lisboa, Portugal

INTRODUCTION

Networks-on-Chip (NoC) are a scalable interconnection network for on-chip communication capable to integrate a high number of processing elements. Scalability, energy efficiency and reliability are among the most important advantages of this new communication paradigm. Hundreds or even thousands of cores can be integrated in a single device using a NoC structure without facing the non-scalability problems associated with bus-based structures or point-to-point connections which are usually irregular and harder to route. Global on-chip communication with long wires thus not scales down with increasing clock frequency. The new communication paradigm decouples the cores from the network, reducing the need for global synchronization, reduces the number of global wires and the energy consumption of cores can be individually controlled. NoC are also reliable since fault-tolerant techniques can be implemented from hardware redundancy to adaptive routing protocols that look for alternative paths for a communication.

The first generation of NoC solutions considers regular topologies, typically 2D meshes under the assumption that the wires' layout is well structured in such topologies. Routers and network interfaces between IP cores and routers are mainly homogeneous so that they can be easily scaled up and facilitate modular design. All advantages of a NoC infrastructure were proven with this first generation of NoC solutions.

However, soon, the designers started to be worried about the two main disadvantages associated with NoCs, namely, area and speed overhead. Routers of a NoC need space for buffers, routing tables, switching circuit and controllers. On the other side, direct bus connection is always faster than pipelined connections through one or more routers since these introduce latency due to packaging, routing, switching and buffering.

In a first attempt to consider area and latency in the design process, designers considered that regular NoC structures may probably be adequate for general-purpose computing where processing and data communication are relatively equally distributed among all processing units and traffic characteristics cannot be predicted at design time. But, many systems developed for a specific class of applications exhibit an intrinsic heterogeneous traffic behavior. Since routers introduce a relative area overhead and increase the average communication latency, considering a homogenous structure for a specific traffic scenario is definitely a waste of resources, a communication performance degradation and an excessive power consumption.

Application specific systems can benefit from heterogeneous communication infrastructures providing high bandwidth in a localized fashion where it is needed to eliminate bottlenecks (Benini & De Micheli, 2002), with sized communication resources to reduce area utilization, and low latency wherever this is a concern.

Homogeneous and heterogeneous solutions of first generation NoCs follow different design methodologies but have one thing in common: their architectures are found at design time and are kept fixed at runtime, i.e., the topology and the architecture of the routers are fixed at design time. Apparently, this is not a problem, but since

DOI: 10.4018/978-1-5225-2255-3.ch395

several applications may be running with the same NoC, the same topology and router will generally not be equally efficient in terms of area, performance and power consumption for all different applications. The efficiency of both homogeneous and heterogeneous solutions can be improved if runtime changes are considered. A system running a set of applications can benefit from the runtime reconfiguration of the topology and of the routers to improve performance, area and power consumption considering a particular data communication pattern. Customization of the number of ports, the size of buffers, the switching techniques, the routing algorithms, the switch matrix configuration, etc. should be considered in a reconfigurable NoC. Both general purpose and application-specific System-on-Chips (SoCs) will benefit from using dynamically reconfigurable NoCs since the performance and power consumption of data communication can be optimized for each application.

The second generation of NoCs are dynamic or adaptive providing a new set of benefits in terms of area overhead, performance, power consumption, fault tolerance and quality of service compared to the previous generation where the architecture is decided at design time. To improve resource efficiency and performance, the NoC must consider adaptive processes at several architectural levels, including the routing protocols, the router, the network interface and the network topology.

This article focuses on adaptive Networks-on-Chip. Adaptive topologies and adaptive routers are described in the following sections.

BACKGROUND

More than a hundred of proposals of NoC architectures can be found in the literature (Salminen, Kulmala & Hämäläinen, 2008). These NoC proposals differ in the used topology, the routing and the switching schemes, the design metrics and the target application. Routers have been also extensively studied, designed and implemented with different flit widths, buffer sizes, switching and routing mechanisms considering latency, area, power consumption, fault-tolerance, quality-of-service, among other metrics. However, only few works have considered adaptive techniques where the NoC can be adapted to the communication requirements statically or at runtime.

Deterministic routing always uses the same route for a particular destination without considering any information about the state of the network. Adaptive routing considers the state of the network, such as the status of a link or buffer, to route data, to route data across the network. Compared to adaptive, deterministic routing requires fewer resources while guaranteeing an ordered packet arrival. On the other hand, adaptive routing provides better throughput and lower latency by allowing alternate paths. Deterministic routing is more appropriate if the traffic generated by the application is predictable, while adaptive deals better with irregular networks and/or stochastic traffic. Deterministic routing usually has poor capacity to equally distribute the traffic along all links of the network since the routes are statically assigned independently of the traffic requirements. On the other hand, highly adaptive algorithms have the potential to reach a uniform utilization of resources and provide fault tolerance. These algorithms distribute the traffic through all links to reduce congestion. However, the efficiency of highly adaptive algorithms is compromised by the necessity to guarantee deadlock free scenarios. Generally, having these algorithms adaptable requires a number of virtual channels (Bjerregaard & Mahadevan, 2006) increasing the cost of the solution compared to that using deterministic routing. Several adaptive routing methods were proposed, as described in the next section.

The proposed adaptive routing algorithms run under the same topology and router configurations, and thus the optimization that can be gained with the adaptive routing is limited by the structure of the NoC and their network components. Besides, highly adaptive routing algorithms are expensive in terms of virtual channels and, consequently,