1. INTRODUCTION

Buses have been the preferred interconnection in most of the processor-based systems in the past thirty years. They have also been the basic building blocks of almost all implemented system-on-chips (SoC). Buses offer a simple and efficient way to transfer data between the components of a system. Their simple signaling and structure facilitate both manual and automatic generation of bus-based architectures. Moreover, the simple programming model of bus-based systems has been favored by the application designers and has led to the development of advanced programming tools.

A single bus is a good choice for many systems when the number of connected components is small. However, in the future SoCs, the complexity of a single component is not likely to significantly increase because only relatively simple components can be scaled with technology [1, 2]. Therefore, the number of components in a SoC is increasing. On the other hand, there has been only a modest increase in the ability of physical wires to transfer signals, although the performance of logic gates has been rapidly improving. The delay of a fixed length wire is even estimated to increase with decreasing feature sizes [1, 3]. Based on this analysis, it seems that a simple bus is no longer a preferable solution for SoC interconnection requirements. Sometimes network-on-chips (NoC), such as the one depicted in Figure 8-1, are proposed to solve this problem. In this Chapter, a SoC architecture that is based on more complex structures than a single bus or fully connected point-to-point links is defined as a network-on-chip.
The homogeneous architectures, such as the mesh of Figure 8-1, are one alternative NoC architecture. They have been used to run fine-grained parallel scientific algorithms at very high speeds. However, the logical limits of parallelism inherent in most practical applications result in redundancies and inefficiencies that cannot always be tolerated in on-chip systems because of increased area and power consumption. In addition, the structures utilized in scientific computing are usually optimized for fine-grain granularity computation. Sometimes NoC architectures are benchmarked with a fairly simple fine-grained algorithm with good results although the distribution of computation for such a large system may not be meaningful. The NoC scheme implies that the components are part of a fairly large coarse-grain granularity system where the components are smaller systems of their own.

A computer local area network (LAN) is another possible analogy for the design of NoCs. Although the on-chip architectures have a lot in common with these systems, they also have considerable differences. The relatively complex network architectures and protocols of LANs as well as their large network buffer memories are designed to make the systems reliable with particular emphasis on connectivity and performance requirements. In contrast, large buffers and arbitrary connectivity are not tolerable in SoCs that prefer limited complexity and real-time operation.

This Chapter advocates a NoC approach that utilizes a hierarchy of bus structures and augments this strategy with more specialized interconnection topologies on case to case basis. In this type of architecture, the local interconnections are buses making use of their described good properties. On the other hand, the global interconnections use a more complex network structure containing buffering elements resembling the router units of Figure 8-1. This type of a heterogeneous architecture can be modified according to