CCIT Report #424 May 2003

QoS architecture and design process for cost effective Network on Chip

Evgeny Bolotin, Israel Cidon, Ran Ginosar and Avinoam Kolodny

Electrical Engineering Department,
Technion—Israel Institute of Technology
Haifa 32000, Israel

Abstract—Our design process characterizes and verifies the inter-module traffic, places the modules so as to minimize the system spatial traffic density on a generic network grid, and optimizes the grid by trimming links, routers and buffers while maintaining the required QoS. We classify the system traffic into four classes of service: Signaling (replacing inter-module control signals); Real-Time (representing delay constrained bit streams); RD/WR (modeling short data access) and Block-Transfer (providing for large data bursts). We model traffic behavior for each class and define the Quality of Service requirements of each class in terms of delay, throughput and relative priority. Based on the traffic model we derive a NoC architecture comprising network protocols, routers, buffers and links that support these four classes. This generic architecture is subsequently optimized to minimize cost (area and power) while maintaining the required QoS.

The network architecture is based on the following principles: The network topology is a planar grid of switches that route the traffic according to fixed shortest path (X-Y based) discipline, thus minimizing hardware tables and traffic overheads. Buffer requirements are reduced by employing multi-class wormhole forwarding while allowing inter-class priorities. The layout of the network is customized and bandwidth is allocated to links according to their relative load so that the utilization of links in the network is balanced. During customization unnecessary resources (links, routers, buffers) are trimmed where possible, resulting in a low cost customized layout for the specific SoC. Analytic calculations and traffic simulations are used in the optimization steps to ensure that QoS is strictly met.

***Index Terms*— Network on Chip, QoS architecture, wormhole switching, NoC design process**

I. INTRODUCTION

On-chip packet-switched networks [1]-[11] have been proposed as a solution for the problem of global interconnect in deep sub-micron VLSI Systems-on-Chip (SoC). Networks on Chip (NoC) can address and contain major physical issues such as synchronization, noise, error-correction and speed optimization. NoC can also improve design productivity by supporting modularity and reuse of complex cores, thus enabling a higher level of abstraction in architectural modeling of future systems [4], [5]. However, VLSI designers must be ensured that the benefits of NoC do not compromise system performance and cost [8], [10]. Performance concerns are associated with latency and throughput. Cost concerns are primarily chip-area and power dissipation. This paper presents a design process and a network architecture that satisfy Quality of Service (performance) requirements at a measurable cost which is favorably compared with alternative on-chip interconnection approaches.

Traditionally, on-chip global communication has been addressed by shared-bus structures and ad-hoc direct interconnections. Non-scalability of these approaches was discussed in [1], [6], [9]. However, modern on-chip buses have evolved to multi-layered and segmented structures, supporting split transactions, burst transfers and parallel operations [12]-[14]. From several aspects they can be considered as networks but still, they don't provide effective spatial reuse of resources and do not utilize packet or wormhole switching associated with distributed routing and congestion/flow control. Therefore, they are inefficient and require centralized arbitration mechanisms.

Advantages of spatial-reuse packet/wormhole switched networks were analyzed in comparison with buses by several authors [1], [3], [5], [8], [9]. A hybrid approach, supporting both NoC and on-chip buses has been proposed in [10]. Switched networks and techniques for their design have been developed for computer networks and for multiprocessor systems [15]-[21]. However, a unique set of resource constraints and design considerations exists for an on-chip environment. As described in [1], [9], memory and computing resources are relatively more expensive on-chip, while relatively more wires are available. The need to combine several types of service, such as "best effort" and "guaranteed throughput" was noted by [1], [8]. In [9] it was suggested to support