ELSEVIER

Microprocessors and Microsystems 24 (2000) 1–11

MICROPROCESSORS AND MICROSYSTEMS

www.elsevier.nl/locate/micpro

Simulative performance analysis of distributed switching fabrics for SCI-based systems

M.A. Sarwar[a,*], A.D. George[b]

[a]*Electrical Engineering Department, FAMU-FSU College of Engineering, Tallahassee, FL 32310, USA*
[b]*Electrical and Computer Engineering Department, University of Florida, Gainesville, FL 32611-6200, USA*

Received 16 July 1999; received in revised form 3 November 1999; accepted 27 December 1999

Abstract

This paper presents the results of a simulative performance study on 1D and 2D *k*-ary *n*-cube topologies as distributed switching fabrics for the Scalable Coherent Interface (SCI). Case studies are conducted on multiprocessor SCI networks composed of simple rings, counter-rotating rings, unidirectional and bidirectional tori, and tori with rings of uniform size. Based on a novel set of verified high-fidelity models, the results identify fundamental performance characteristics associated with each of these SCI fabrics, and tradeoffs between them, in terms of throughput and latency. Limits on scalable performance from SCI with increase in complexity and dimensionality are clarified, supporting decisions for advanced multiprocessor design.

Keywords: BONeS; Discrete-event simulation; High-performance networks; Scalable coherent interface; Switching fabrics

1. Introduction

The performance and scalability of high-speed computer networks have become critically important characteristics in the design and development of advanced distributed and parallel processing systems. Many applications require or benefit from the use of an interconnect capable of supporting shared memory in hardware, and chief among such interconnects for multiprocessors is the Scalable Coherent Interface. However, in order for SCI interconnects to scale to ever larger system sizes and support a host of embedded and general-purpose applications, a distributed switching fabric is required that will scale with the number of nodes. One of the most promising families of topology for distributed switching fabrics is the *k*-ary *n*-cube topology, a family originally investigated for high-end supercomputing and referenced widely in the literature as the target for algorithm mapping.

The SCI standard is targeted towards increasing the bandwidth of backplane buses and became an IEEE standard in March of 1992 [7]. It improves on the bandwidth of buses by using high-speed, ring-connected, point-to-point links. With a link speed of 1 GB/s (i.e. a gigabyte per second), addressing for up to 64K nodes, and a cache-coherence protocol for distributed shared-memory systems, the popularity of SCI for use in large multiprocessors has continued to increase. Sequent, Cray, and HP-Convex are among the parallel computer vendors that have developed proprietary implementations of SCI for their high-end systems. Sequent developed the IQ-link implemented in their Numa-Q 2000 system to connect groups of four processors in a ring structure [8]. Cray developed the SCX channel, also known as the GigaRing, capable of sustained half-duplex bandwidths of 900 MB/s [11,12]. The HP-Convex Exemplar Series uses the SCI-based Coherent Torroidal Interconnect (CTI) to interface hypernodes consisting of eight processing units each.

SCI has also gained recognition in the workstation cluster market. To date, Dolphin Interconnect Solutions has emerged as the leading manufacturer of SCI adapter cards and switches for clusters. The Dolphin switch relies on a bus-based internal switch architecture called the B-link. The B-link is capable of a bandwidth ranging from 200 to 400 MB/s depending on the operating clock speed. Sun has adopted the Dolphin implementation of SCI, dubbed CluStar, for their Enterprise Cluster systems. Recently, Dolphin introduced a dual-ported PCI/SCI adapter card from which to construct unidirectional 2D torus topologies for SCI. Data General, in collaboration with Dolphin, has developed a chipset for their AV20000 Enterprise server to interface SCI to Intel's Standard High Volume (SHV) server nodes [2]. In addition, Dolphin and Siemens jointly developed a PCI–SCI bridge to be used in the I/O subsystems of the Seimens RM600E Enterprise Server systems.

* Corresponding author.
E-mail addresses: sarwar@hcs.fsu.edu (M.A. Sarwar), george@hcs.ufl.edu (A.D. George).

0141-9331/00/$ - see front matter
PII: S0141-9331(00)00063-6

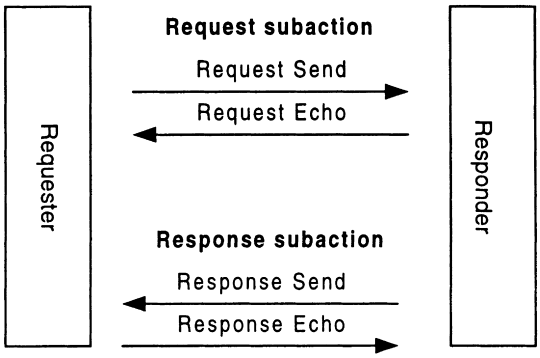

Fig. 1. Split transaction used in SCI.

While Dolphin's B-link bus provides a cost-effective approach to internal switch architecture, it is limited in its scalability and support for multidimensional network topologies. In this paper, a crossbar-based SCI switch model is presented that does not suffer from these limitations. The switch uses routing tables that are automatically generated at startup and which guarantee the shortest path to each packet's final destination. The performance of *k*-ary *n*-cube systems is explored by conducting experiments with a fixed ring size over a variable number of total nodes. The *k*-ary *n*-cube family consists of direct networks with *n* dimensions and *k* nodes per dimension, and members include rings, meshes, tori, hypercubes, etc. These networks provide excellent scalability through a constant node degree (i.e. fixed number of ports per node despite the size of the system), and low latencies through smaller diameters (i.e. fewer hops when transferring packets from source to destination). Additionally, they provide topologies that have served as targets for many studies on the mapping of parallel algorithm graphs for multiprocessing and multicomputing. Work by Dally [3], Chung [1], and Reed and Grunwald [10] provide detailed analytical analyses of generic *k*-ary *n*-cube networks.

By contrast, in this paper we concentrate on a simulative approach of applied research to determine the performance of *k*-ary *n*-cube networks constructed with SCI. Through simulation with high-fidelity models for SCI, more accurate results can be obtained to study the relationship and impact of selected switching topologies for SCI multiprocessor networks. The remainder of this paper is organized as follows. Section 2 introduces the Scalable Coherent Interface and its basic operation. Section 3 describes the SCI switch model, and Section 4 presents the performance simulation results for several *k*-ary *n*-cube topologies. Finally, conclusions and directions for future research are discussed in Section 5.

2. Scalable Coherent Interface

The basic SCI topology is a ringet. The ringlet is constructed using SCI interfaces at each node. The nodes communicate using unidirectional point-to-point links. Each link is 18-bits wide – 16 data bits also referred to as an SCI symbol, a flag bit, and a clock bit. The flag vit is used to identify the type of packet being transmitted on the data lines while the clock provides synchronous communications between nodes. SCI uses a split-transaction protocol where each transaction consists of two subactions. The subaction can be either a request or a response from a source, followed by an echo indicating whether the request or response was accepted at the destination, as illustrtated in Fig. 1.

A typical transaction on a single SCI ring begins with a *request-send* packet from the requestor to the responder. The responder then returns a *request-echo* packet to the requester indicating that the request has been received. After processing the request, the responder sends a *response-send* packet to the requester and receives a *response-echo* packet in return. Some subactions, such as the *move*, do not have a corresponding response subaction.

The basic architecture of an SCI interface is illustrated in Fig. 2. Incoming packets to the interface pass through an address decoder. If the packet is destined for the local node, the decoder places it into the request or response input queue. If the packet is destined for another downstream node, it is forwarded to the bypass FIFO. To output a packet, the SCI node must have sufficient free space in its bypass FIFO to hold all incoming symbols. When there are no packets waiting in the output queue or there is insufficient free space in the bypass FIFO for the output queue data to be sent, data from the bypass FIFO is transmitted on the node's output link. If the bypass queue is empty, then *idle* symbols are transmitted. *Idle* symbols also carry flow-control information and at least one must precede any *send* or *echo* packet. This flow control information is used to inhibit upstream nodes from sending data when the bypass FIFO must be emptied to allow the output queue to be emptied.

Larger SCI networks are based on multiple ringlets connected together to create more complex topologies through the use of agents. An agent is essentially an SCI-to-SCI bridge used to interconnect two or more rings. The use of agents alters the transaction communication protocol sightly as illustrated in Fig. 3. The figure depicts two nodes communicating via an agent. The agent serves two purposes–it is the responder for one ringlet and the requester for the second. In step 1, the agent accepts the request on behalf of the responder. In step 2, the request is forwarded onto the responder's ring by the agent. Step 3 shows the agent accepting the response and forwarding it back to the requestor on the first ring in step 4.

In order to explore the performance of SCI for different topologies in the distributed switching fabric, high-fidelity SCI interface and agent-based switch models were constructed. These event-driven models are accurate down to the SCI clock cycle, providing a level of detail virtually identical to the real network hardware. Switching times are dependent on the packet length and are not assumed to be