Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Systems Architecture 52 (2006) 345–358
JOURNAL OF SYSTEMS ARCHITECTURE
www.elsevier.com/locate/sysarc

High-performance adaptive routing for networks with arbitrary topology

V. Puente [a], J.A. Gregorio [a], F. Vallejo [a], R. Beivide [a], C. Izu [b,*]

[a] *Computer Architecture Group, University of Cantabria, 39005 Santander, Spain*
[b] *Computer Science Department, University of Adelaide, SA 50005, Australia*

Available online 14 November 2005

Abstract

A strategy to implement adaptive routing in irregular networks is presented and analyzed in this work. A simple and widely applicable deadlock avoidance method, applied to a ring embedded in the network topology, constitutes the basis of this high-performance packet switching. This adaptive router improves the network capabilities by allocating more resources to the fastest and most used virtual network, thus narrowing the performance gap with regular topologies. A thorough simulation process, which obtains statistically reliable measurements of irregular network behavior, has been carried out to evaluate it and compare with other state-of-the-art techniques. In all the experiments, our router exhibited the best behavior in terms of maximum/sustained performance and sensitivity to the network topology.

Keywords: Irregular topologies; Routing, Flow control; Deadlock; Cluster computing; Networks of workstations; Local area networks

1. Introduction

Networks of workstations and other forms of cluster computing are currently emerging in the high-performance computer market as good alternatives to sophisticated parallel computers. The cost/performance ratio of the commodity hardware and the existence of affordable and scalable high-performance communication technologies justify the penetration in the market of these distributed computing platforms. Moreover, their versatility makes this kind of systems particularly attractive to solve a wide range of applications.

Computer clusters are commonly organized as switched networks in which each switch or router has several computing nodes connected to some of its input/output ports. The remaining ports are used to link other routers in order to provide a connected system. Bi-directional full-duplex links are commonly employed to exploit communication locality. Packets interchanged among computing nodes cross the network according to the rules dictated by a certain routing mechanism. It is well known that the interconnection network architecture and its associated software libraries are critical components for high-performance cluster computing.

* Corresponding author. Fax: +34 942 201479.
E-mail addresses: vpuente@atc.unican.es (V. Puente), jagm@atc.unican.es (J.A. Gregorio), fernando@atc.unican.es (F. Vallejo), mon@atc.unican.es (R. Beivide), cruz@cs.adelaide.edu.au (C. Izu).

1383-7621/$ - see front matter
doi:10.1016/j.sysarc.2005.09.003

In recent years, the adoption of packet routing techniques successfully used in multiprocessor systems has been a common technological trend for commercial high-performance interconnection networks. Nevertheless, not all the low-level network functions used in multiprocessors can be easily adapted to this new scenario. One of these critical functions, conditioning either cost or performance of the whole network, is the method of dealing with packet deadlock. There are a number of successful distributed deadlock avoidance mechanisms based on the regular multi-dimensional structure of the network topology [6,8,17]. Notwithstanding, network irregularity is a common characteristic of most distributed computer systems. Although this feature allows for an easy and flexible network design, it requires new distributed mechanisms to deal with packet deadlock.

The mechanisms used in experimental and commercial high-performance communication technologies for cluster computing either require a significant number of hardware resources [13] or it enforces a restrictive use of its hardware resources [2,21] compromising in both cases their cost/performance ratio. Moreover, there are technologies that prohibit the use of irregular topologies [7,14], which clearly restrict the system versatility and thus, its applicability.

This paper provides a new approach to avoid deadlock in irregular networks by means of a controlled packet injection technique over a virtual ring embedded into the network topology. This method has been derived from a switching technique developed by the authors for regular topologies such as the torus that can be decomposed into a set of rings [4,18] and successfully employed in CC-NUMA multiprocessor systems [17], outperforming other high-performance routing proposals that use the same amount of resources. Actually, this routing mechanism has been implemented in the IBM BlueGene/L supercomputer [1].

The remainder of this paper is organized as follows: Section 2 presents the state-of-the-art on routing mechanisms for irregular networks. Section 3 introduces our routing methodology and Section 4 describes the corresponding router architecture. Section 5 details our simulation environment and Section 6 discusses the performance differences among the different proposals analyzed. Finally, the main findings of this research are summarized in Section 7.

2. Related works

This Section reviews the deadlock management functions associated with adaptive routers in irregular networks. This provides the context to analyze the basis of the different routing alternatives.

When a packet in transit reaches a router, its header provides the information to select the output port to be forwarded. The routing mechanism must safely choose an output port such that the network is maintained free of anomalies. Packet deadlock, as commented before, is the most critical anomaly that a network can suffer. A cyclic dependency between busy and requested communication channels is the source of deadlock among packets [6]. In other words, the transit of a set of packets indirectly depends on their own movement, it is thus impossible for them to advance towards their destinations.

In regular topologies, such as meshes and tori, links are arranged into several dimensions and deadlock avoidance mechanisms can be easily distributed among these dimensions [6,17,8]. The techniques used in these methods are diverse but they are mainly based on restrictions on the routing mechanisms or on the packet injection process. Other less conservative techniques rely on the fact that packet deadlock is infrequent and apply techniques based on deadlock detection and recovery [26]. Obviously, these distributed techniques cannot be easily exported to networks with irregular topology. The irregularity obliges the utilization of more restrictive and complex mechanisms, most of them based on a particular centralized view of the network. In most cases, the performance of the network is heavily conditioned by the restrictions imposed by its respective deadlock avoidance mechanism. For example, the up*/down* routing (UDR) selects a root node and embeds a *breath-first spanning tree* (BFS) in the network topology which provides a total order of the network channels [21]. The classification of a link as ascendant or descendent is determined by the relative positions of the nodes it connects to the root. A permissible path will be one in which the packet can only use descendent links after traversing all the ascendant ones, thus avoiding any cyclic dependency. The implementation of this mechanism is relatively easy but the restrictions imposed in the use of the network resources lead to poor performance. Packets have to travel non-minimal paths and frequently congestion builds up around the root node. A related mechanism using UDR based on a *depth-first span-*