Microprocessors and Microsystems 38 (2014) 170–181

Contents lists available at ScienceDirect

Microprocessors and Microsystems

journal homepage: www.elsevier.com/locate/micpro

Runtime connection-oriented guaranteed-bandwidth network-on-chip with extra multicast communication service

Faizal Arya Samman *

Universitas Hasanuddin, Fakultas Teknik, Jurusan Teknik Elektro, Jl. Perintis Kemerdekaan Km. 10, Makassar 90245, Indonesia

ARTICLE INFO

Article history:
Available online 3 September 2013

Keywords:
Network-on-chip
Wormhole Cut-Through Switching
Quality of service
Guaranteed-bandwidth service
Best-effort service
Connection-oriented protocol
Multicast communication

ABSTRACT

This paper presents a flexible runtime connection-oriented guaranteed-bandwidth Network on Chip (NoC). Comparing with a standard time-division multiplexing (TDM) method, our local ID-based method provides better flexibility to establish dynamic runtime connections. A specific pre-designed algorithm for finding a conflict-free scheduling, as commonly used in the TDM-based method, is not needed. The contention problem is solved with the hardware solution based on the locally organized message identity (ID), in which flits belonging to the same stream packet will have the same unique/local identity-tag (ID-tag) on each communication link. The ID-tags of each stream will vary locally over communication links and are updated. The updating is organized by ID-tag mapping management units. The routing is organized using runtime programmable routing reservation table. In addition, the proposed methodology supports also a deadlock-free multicast routing service.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

On-chip interconnection networks are an interesting alternative communication infrastructure to bring a new paradigm to design and develop a Multiprocessor System-on-Chip (MPSoC) and Chip-level Multiprocessor (CMP) systems. Bus systems, which is traditionally used as communication fabrics for SoC systems, tend to perform bottleneck, especially if they are loaded with a large number of high-speed processing elements. Instead of using the traditional bus systems as communication media among processing element (PE) units, on-chip interconnection networks is proposed, reflecting a concept of scalable shared communication media.

Several multimedia applications for MPSoCs consist of some communication edges that must be performed with a certain communication bandwidth. Video/audio streaming data transmission from a core to one or to many cores needs a constant transmission rate. Performance degradation at one the communication edges could reduce the overall application performance or even could break the multimedia applications. Hence, a specific service of the network is required to guarantee the data bandwidth of the video/audio streaming. This paper will present a methodology on how the bandwidths of unicast and multicast communication edges in an on-chip radio system application benchmark [9] can be well guaranteed. The methodology proposes a runtime virtual cicuit configuration technique, where connections are established during application execution time. Moreover, the technique supports also a runtime multicast communication service.

Guaranteed-bandwidth (GB) service can be implemented by using end-to-end connection establishment technique, where a stream header reserves required bandwidth during connection set-up phase before sending the video/audio streaming. By further applying a policy where every network link cannot be consumed by considered traffic exceeding its maximum capacity, then (long-term) saturated network condition can be avoided (non-blocking traffic flow is guaranteed). Guaranteed-service can be implemented by allowing multiple packets to share the same link. The link sharing can be realized by using a data multiplexing technique. The shared link configuration in a network is also commonly called as *switched virtual circuit (SVC) configuration*. However, the guaranteed-bandwidth method is only suitable for NoC-based multicore processor systems, where processors intensively sending and receiving streaming data that require expected constant end-to-end communication rates.

This paper proposes one of the SVC method by using dynamic local identity (ID) assignment technique. By using this technique, a message is not split into packets, but it is split into flow control digits (flits) with extra bit fields for dynamic unique/local ID and flit-type control label per flit. The idea results in a novel routing paradigm, where the network routes flits instead of packets ("routes flits not packets").

The rest of the paper is organized as follows. Section 2 will present the state-of-the-arts of the switched virtual circuit configuration methods. In this case, we compare our method with TDMA-based data multiplexing technique. Section 3 presents the

* Tel.: +62 411588111.
E-mail address: faizalas@unhas.ac.id

0141-9331/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.micpro.2013.07.006

contributions of the paper. The basic infrastructure required to implement the proposed method is described in Section 4, i.e. the required standard packet format and the microarchitecture. Section 5 describes the connection-oriented multicast communication protocol. The ID-based routing mechanism dan the ID Management Scheme are described in detail in Section 6. Section 7 shows an experimental result. In this section, we evaluate our data switched virtual circuit configuration method on a radio system application benchmark from Nokia [9]. The synthesis results of the NoCs are presented in Section 8. Brief discussions about the implementation issues of the multiple access methods are presented in Section 9. Section 10 concludes the work and presents the main advantageous features of the proposed method.

2. State-of-the-art of data multiplexing techniques for NoCs

2.1. NoCs with TDMA technique

A commonly used method to provide guaranteed-service for NoCs is a pipeline circuit switching based on the Time-Division Multiple Access (TDMA) method. Æthereal [12], Asynchronous clockless MANGO NoC [1], circuit switched PNoC [4], Sonics [19], DSPIN [11], Nostrum [10] and a NoC SMT Switch [8] are NoC examples that use such methodology. Fig. 1(a) presents the conceptional view of the TDMA method. The link connecting the input and output port of the routers is shared by four packets, i.e. stream packet *A*, *B*, *C* and *D*. Each packet establishes a virtual circuit configuration based on time slots allocation on the outgoing port. In the figure, we assume that the link has 8 time slots. The more time slots are allocated for a packet, the more bandwidth (BW) it reserves. Thus, the packets *A*, *D*, *B* and *C* reserve 50%, 25%, 12.5% and 12.5% of the maximum link BW capacity, respectively. A packet allocated at time slot S_t on a link must be allocated to time slot S_{t+1} on the next link. Based on Fig. 1(a) for instance, packet *D* allocated to time slots S_1 and S_2 on the link must be allocated to time slots S_2 and S_3 on the next link.

2.2. NoCs with IDMA technique

In this section, we introduce a concept based on local identity (ID) division multiple access (IDMA) technique. Fig. 1(b) presents the concept, in which local ID slots can be reserved by single data stream as its ID-tag. The local ID tag appears on every flit and is updated every time the data stream acquires the next link. Flits belonging to the same stream will always have the same local ID. In order to guarantee a correct routing function, an *ID Management Unit* must index every reserved ID slot by identifying the previous ID tag of the stream/message, which reserves one ID slot and from which port the stream/message comes.

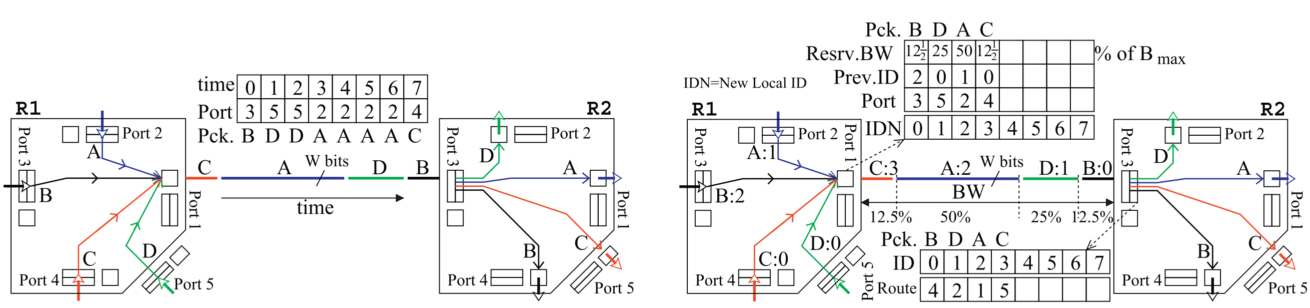
Based on the Fig. 1(b), for instance, packet *D* is allocated to local ID slot (IDN) number 1 (its new ID-tag), and is identified by the ID Slot Table as a packet from input port 5 having previous ID tag 0 in the router *R*1. The message *D* reserved also 25% of the maximum link bandwidth capacity (B_{max}). In the next router *R*2, the stream/messages are routed based on their current/new ID-tags. Thus, the packet *D* flows from output Port 1 (East) of the router *R*1 to the input Port 3 (West) of the next router *R*2 with new ID-tag 1. The number of available ID slots reflects the maximum number of stream/messages allowed to form switched virtual circuit configurations on the link. The bandwidth can be guaranteed by further implementing a connection-oriented communication protocol, where the requested BW attached on a header flit bit fields is used to reserve the expected end-to-end communication bandwidth over the network links.

2.3. Comparisons of the multiple access methods

The TDMA-based switching requires a pre-design time-slot allocation algorithm to achieve a conflict-free routing and scheduling. *UMARS* + algorithms [3], TDM-based Virtual Circuit Configuration (*VCC*) Method [9], and a time-slot allocation algorithm made for μSpidergon NoC [2] are the NoC examples that use the time-slot allocation algorithm. The IDMA-based method does not need such time-slot allocation algorithms, because the local ID slot on each outgoing link is reserved and allocated autonomously by header flits of a streaming data during application execution time (flexible runtime autonomous switched virtual circuit reconfiguration). The same technique could be certainly applied to the TDMA-based method, but the probability in which the header flit fails to establish connection is very high especially in a very high traffic situation. The need for the time-slot allocation algorithm in the TDMA-based method is due to the conflict-free requirement.

A NoC design methodology that is based on three kernels, i.e. traffic classification, flit-based switching and path pre-assignment and link-BW setting has been introduced [7]. The traffic are classified into guaranteed-latency (GL), guaranteed-bandwidth (GB) and best-effort (BE) traffic. The GL traffic have stringent maximum delay requirement from data injection until data acceptance. The GB traffic requires constant end-to-end communication bandwidth, while the BE traffic does not have bandwidth requirement neither stringent data transfer latency. The link allocation (path assignments) for the GL and GB traffic is static or computed off-line at design time [7]. For a new application, the path assignment must be done again at design time. Therefore, the proposed methodology is not suitable for application mapping, where the applications are known after chip-manufacturing.

An extra bit field for dynamic local/unique ID-tag and an extra field to identify the flit type of each flit are attached on each flit. The extra bit fields are useful to guarantee that flits belonging to



(a) TDMA-based Data Multiplexing (b) IDMA-based Data Multiplexing

Fig. 1. State-of-the-art of the data multiplexing techniques for NoCs.

Download English Version:

https://daneshyari.com/en/article/461516

Download Persian Version:

https://daneshyari.com/article/461516

Daneshyari.com