Dynamic Buffer Regulator for 3D Mesh Network-on-Chip

Cheng Liu, Lei Zhang, Yinhe Han, Xiaowei Li
Key Laboratory of Computer System and Architecture
Institute of Computing Technology, Chinese Academy of Sciences
E-mail: liucheng@ict.ac.cn

1. Introduction

Network-on-chip (NoC) that addresses scalability challenges as well as bandwidth bottleneck and three dimensional integrated circuits (3DICs) that alleviate interconnect latency pressure as well as heterogeneous integration problems are emerging for complex integrated systems. 3DNoC combining both the benefits soon becomes one of the most promising on-chip communication technique in complex System-on-Chip (SoC).

Although interconnect fabric comprises of different components, buffer, which is instrumental in the overall NoC operation, consumes the largest part of power and chip area in a NoC router. For instance, the buffer will take up to nearly 50% of area and 64% of leakage power at router implemented under 70nm CMOS technology [1]. Since 3DNoC is prone to employ routers with more ports thus more buffers, the burdens of buffer power and area get heavier.

There have been several works on the buffer organizations to save chip area and exploit performance. Hu et al. [2] and Huang et al. [3] and customized the buffer capacity and virtual channel (VC) number respectively, and achieved large buffer savings without any performance degradation. However, it was a static approach which was based on a detailed analysis of application specific traffic patterns. Nicopoulos [4] and Lai [5] introduced dynamic VC regulators which allocated the number of VCs according to the traffic conditions. And it turns out that the dynamic buffer regulators could also achieve buffer saving without network performance penalty under diverse load and traffic patterns. Unfortunately, the buffer saving is less significant compared with static methods.

To acquire an insight of different buffer optimization schemes, we have made extensive experiments and analysis which will be presented in next section. Finally, we notice that buffer utilization in NoC exhibits significant unbalance in both spacial distribution and temporal distribution. Unbalance here in spacial distribution means that buffer utilization in different nodes, ports and even VCs differs notably, while unbalance in temporal distribution indicates that buffer utilization in different time windows also varies a lot. With this observation, now it is clear that static buffer allocation which provides larger buffer for ports with higher blocking probability while smaller buffer for the others is actually balancing buffer utilization across the network. Thus the method could improve buffer utilization and save buffer effectively. Dynamic buffers proposed in previous works are also balancing buffer utilization of different VCs within a single router port through both temporal dimension and spacial dimension. However, since it is limited in a single router port, chip area saving for dynamic buffer is less significant especially when there is only medium buffer capacity in a single router port.

In this paper, we propose a dynamic buffer regulator which not only shares buffer dynamically but also balances the utilization among ports of different routers. The idea is mainly based on the observation that ports in neighboring routers seldom transmit data at the same time. Thus it is possible that all these router ports can share a single input buffer. Meanwhile, through silicon vias (TSVs) which cut across thinned silicon substrates to build connectivity among different layers in 3DNoC are extremely short. Moreover, with continuous technique advancement, TSV consumes negligible chip area and high density TSV manufacture is getting mature [14]. As a result, TSV technique facilitates the proposed dynamic buffer regulator design.

The rest of the paper is organized as follows. Section 2 motivates this work. Section 3 presents the detailed design and implementation of the proposed dynamic buffer regulator.

2. Motivation

NoC buffer has tremendous influence on NoC performance, but it also brings both power and chip area burdens. Therefore, we cannot afford unlimited buffers, while more importantly, we need to improve buffer utilization. In this section, we firstly give in-depth analysis on buffer utilization in 3D mesh NoC. $4 \times 4 \times 4$ 3D mesh NoC is implemented in a cycle-accurate NoC simulator written in SystemC. Dimension order routing algorithm and wormhole flow control are employed in the network. Each router in the network has two stage pipelines and 8-flit depth buffer in each input port. Each packet contains 4 flits and is injected according to a self-similar process with minor hotspot. Then we analyze both the buffer utilization and vertically connected routers’ buffer write contention probability every 1000 cycles. Note that buffer contention means the occasion when buffers in vertically connected routers receive a flit from corresponding links at the same time. It is not a real contention physically.

Fig.1 shows buffer utilization of vertically connected routers in a random time window. It is clear that none of the buffers in different routers’ ports keep busy all the time. Actually, we have also analyzed many different time windows, and similar situations appear. The reason for this is that NoC traffic exhibits kind of self-similar characteristic or burst transmission, so buffers in NoC usually keep busy for a while and then turns to be free. In addition, since buffers in different routers experience different network status, thus they seldom keep busy concurrently in a small time window even under uniform traffic. As a result, buffer utilization presents significant imbalance in temporal dimension as well as spacial dimension.

![Fig.1 Buffer Utilization of Vertically Connected Routers in a Random Time Window](image-url)
3. Dynamic Buffer Regulator

3.1 Dynamic Buffer Architecture

Before we explain the dynamic buffer architecture, 3DNoC and general router architecture are briefly introduced as the background. Fig.3 shows a typical SoC based on 3D mesh NoC. It includes link, TSV, router and processing elements (PE). Routers connected by links and TSVs provide an efficient communication backbone for PEs and dominate the NoC performance and overhead. General router mainly consists of buffer, crossbar and arbiter as shown in Fig.3.

![Fig.3 General Router in 3D Mesh NoC](image)

Typical router data paths from input port to output port are shown in Fig.4-a. All of them lie in the same vertical dimension and router ports IDs are also identical. In a homogeneous design, it is safely to assume that these data paths can be connected using additional TSVs as presented in Fig.4-b. According to previous analysis, we find that these distributed buffers in the same vertical dimension seldom implement write operation concurrently and their utilizations also vary. Now with a write bus and a read bus, buffer ports in different routers can be shared. Then buffer capacities in different routers can also be shared through a classic dynamic buffer regulator. That is to say, each buffer can borrow capacity from buffers in the other nodes when they are busy. Also the buffer can lend capacity to buffers in other nodes when they are free. Additionally, we have allocated an independent register for each layer. When there is no contention for read bus, data will be directly sent to crossbar and there will be no performance penalty compared with traditional design. When contention occurs, the registers can be used to enhance network throughput. Although it will induce an additional pipeline, the overhead can be amortized by the pipeline transmission. Using the proposed design, we can keep similar network performance with less buffer consumption, or acquire better performance with similar buffer consumption or find any trade-off between them.

![Fig.5 Dynamic Buffer Regulator](image)

**Reference**


