Journal of Network and Computer Applications 45 (2014) 27–34

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

The optimal joint sequence design in the feedback-based two-stage switch

An Huang [a], Bing Hu [a,b,*]

[a] Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, PRC
[b] State key Laboratory of Networking and Switching Technology (Beijing University of Posts and Telecommunications), Beijing, PRC

ARTICLE INFO

Article history:
Received 21 August 2013
Received in revised form
11 April 2014
Accepted 18 June 2014
Available online 23 July 2014

Keywords:
Load-balanced switch
Feedback-based switch
Two-stage switch

ABSTRACT

The feedback-based two-stage switch is scalable as it is configured by a predetermined and periodic joint sequence of configurations (Hu and Yeung, 2010). Its major problem is that the average packet delay is high under light traffic load. In this paper, we improve the performance of feedback-based switch while still ensuring in-order packet delivery and close to 100% throughput. We first show that the different sequences of configurations may endow a feedback-based switch with different delay performance. We propose to devise a tailor-made sequence of configurations for the coming traffic pattern. Given any traffic matrix, the optimal joint sequences that do exist can produce the lowest average packet delay. Then finding the optimal joint sequences is formulated as an ILP (Integer Linear Programming) problem. The simulation results demonstrate that the optimal joint sequence can cut down the average packet delay up to 14% under a random uniform traffic model and even 45% under a hot-spot traffic model. Last but not least, we also design a fast suboptimal algorithm for the practical implementation.

1. Introduction

With the continuous growth of bandwidth in fiber links, the need for building high-speed switches/routers is urgent in order to keep pace with the increased transmission rate. Load-balanced switches (Chang et al., 2002) have received a great deal of attention recently because they are simple and can provide close to 100% throughput. A load-balanced switch consists of two stages of switch fabric, as shown in Fig. 1. The first switch fabric converts the non-uniform traffic into uniform and the second fabric delivers packets to their correct outputs. Each switch fabric is configured by a predetermined and periodic sequence of N configurations, where N is the switch size. The basic requirement is that each input is connected to each output exactly once in the sequence. Accordingly, a central scheduler for determining the best switch configuration at each time slot in real time is not needed, which makes load-balanced switch very suitable for high-speed implementation.

We can see from Fig. 1 that the outputs of the first switch fabric collocate with the inputs of the second switch fabric. Unless otherwise specified, we name them middle-stage ports. We also call the inputs of the first switch fabric and outputs of the second switch fabric as inputs and outputs of the load-balanced switch, or simply inputs and outputs respectively. The basic operation of a load-balanced switch is as follows. When a packet arrives at an input (and assume there is no input buffer), it will be immediately delivered to a middle-stage port based on the current switch configuration used in the first switch fabric. Due to the periodic sequence of configurations used, a burst of packets arrived at an input will be evenly spread out to different middle-stage ports. Ideally, the (non-uniform) input traffic will be converted into uniform before entering the second switch fabric. Packets which arrived at middle-stage ports join the corresponding VOQ_2s (virtual output queues in the second stage) based on their destined outputs. When a middle-stage port is connected to an output (according to the periodic switch sequence used), a packet (if any) from the corresponding middle-stage port VOQ_2 will be sent. It can be easily shown that if the traffic entering the second switch fabric is uniform, 100% throughput can be guaranteed. The issue is whether the load balancing performance rendered by the first switch is good enough. In Chang et al. (2002), it is proved that if the input traffic is stationary and weakly mixing (Nadkarni, 1998), the first switch fabric can convert any non-uniform traffic into uniform.

From the basic operation of a load-balanced switch above, we can see that packets of the same flow (i.e. arriving at the same input and destined for the same output) will arrive at their output via different middle-stage ports, due to the load-balancing mechanism at the first switch fabric. Then they may experience different delays at different middle-stage ports. When packets of the same flow finally arrive at the output, their order cannot be guaranteed.

* Corresponding author. Tel.: +86 571 87951004.
E-mail addresses: 21131090@zju.edu.cn (A. Huang), binghu@zju.edu.cn (B. Hu).

http://dx.doi.org/10.1016/j.jnca.2014.06.011
1084-8045/

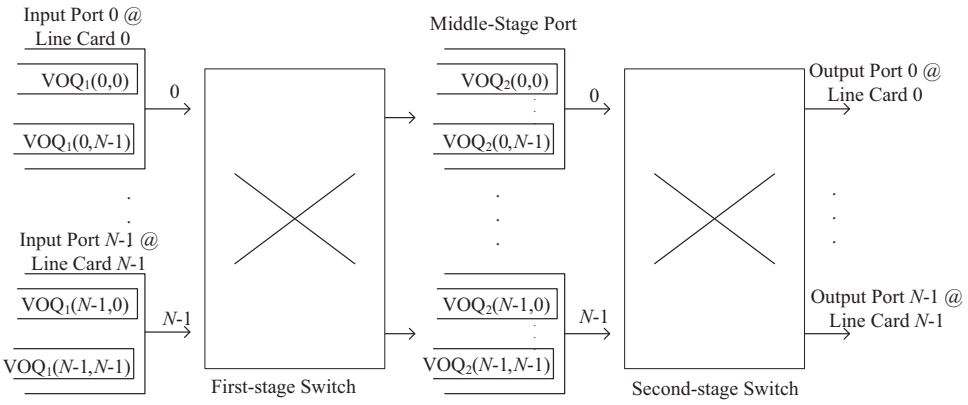

Fig. 1. A load-balanced two-stage switch.

Many efforts (Chang et al., 2004; Shen et al., 2005; Keslassy et al., 2003; Tu et al., 2005; Keslassy and McKeown, 2002; Lee et al., 2006; Cai et al., 2014) are then made to address this notorious packet mis-sequencing problem. Among those work, the feedback-based (two-stage) switch (Hu and Yeung, 2010) provides an elegant solution. Specifically, the sequences of N configurations at the first and second stages are defined to form a joint sequence. Unlike load-balanced switch, the feedback-based switch utilizes a tailor-made joint sequence (but it is still predetermined and periodic) and a single packet buffer at each middle-stage port VOQ_2. The mis-sequencing problem is thus solved without sacrificing delay and throughput performance (for a detailed review of the feedback-based switch, refer to Section 2.)

The major problem with feedback-based switch is that its average packet delay is high under light traffic load. To address this problem, the effort in Hu et al. (2012a) adopts a Clos network (Clos, 1953) for constructing a large switch from a set of smaller feedback-based switch modules. Then the order of average packet delay is cut down from $\mathcal{O}(N)$ to $\mathcal{O}(\log_2 N)$ slots. But packets of the same flow will go through different switch modules to avoid blocking the Clos network. When they finally reach output ports, the notorious packet mis-sequencing revives.

The recent work Chiu et al. (2013) relies on the IC (Integrated Circuit) design to reduce the propagations delay in a feedback-based switch. Specifically, a load-balanced 4×4 switch fabric IC is proposed as a module to construct a large $N\times N$ switch system. But when the system scales up, the switch throughput degrades into 80%.

Another interesting approach in Hu et al. (2012b) extends the feedback-based switch from two-stage to three-stage for further cutting down the average packet delay while still ensuring in-order packet delivery. Its basic idea is to utilize the third stage switch to map heavy flows to experience the less middle-stage delays. Nevertheless, the implementation complexity for introducing the extra third stage fabric should not been ignored.

In this paper, we focus on improving the performance of feedback-based switch but without the mis-sequencing problem and the requirement of the third switch fabric. This is inspired by the observation that different joint sequences of configurations endow a feedback-based switch with different delay performance. In particular, we propose to design the joint sequence of configurations based on the incoming traffic pattern. Given any traffic rate matrix, the optimal joint sequences can yield the lowest average packet delay. But finding the optimal joint sequences is with the complexity of NP hard (Yeung and Yum, 1998) and we formulate it as an ILP (Integer Linear Programming) problem. A fast sub-optimal algorithm is also devised for practical implementation. The simulation results show that ILP and suboptimal algorithm cut down the average packet delay up to 45% in some cases. Even under random uniform traffic, 14% performance improvement can be obtained by ILP algorithm and 5.5% can be obtained by sub-optimal one.

The rest of this paper is organized as follows. We first review the feedback-based switch in Section 2. The ties between joint sequence and delay performance are discussed in Section 3. In Section 4, we utilize ILP to find the optimal joint sequence under a given traffic matrix. We compare the optimal joint sequence with three-stage switch in Section 5. A fast sub-optimal design is devised in Section 6 and the simulation results are presented in Section 7. We extend our discussion on the estimation of traffic rate and the penalty of changing joint sequences in Section 8. Finally we conclude the paper in Section 9.

2. Feedback-based two-stage switch architecture

In a feedback-based switch, each middle-stage $VOQ_2(j,k)$ only needs a single packet buffer. The two tandem switch fabrics are configured according to a special joint sequence with both staggered symmetry and in-order packet delivery properties. Three examples of such sequence are shown in Fig. 2c–e.

Staggered symmetry property is required for constructing efficient feedback paths. It refers to the fact that for any middle-stage port j, if it is connected to output k at time slot t, then at next slot $(t+1)$ input k is connected to the same middle-stage port j. The joint sequences in Fig. 2a and c have the staggered symmetry property. Since each $VOQ_2(j,k)$ has only one single packet buffer, an N-bit vector is enough to denote the occupancy of all N $VOQ_2(j,k)$s $(k=0,1,\ldots,N-1)$ at middle-stage port j. This vector is piggybacked onto the data packet sent to output k (from middle port j), and is then immediately made available to input k (because both input k and output k reside on the same switch linecard). Based on the received occupancy vector, input k selects the best packet for sending to its currently connected middle port j. Specifically, among the set of queues with the corresponding middle-stage $VOQ_2(j,k)$ empty, a packet from the longest $VOQ_1(i,k)$ $(k=0,1,\ldots,N-1)$ is selected for sending.

If all input ports connect to middle ports following the same cycle, we say the sequence of N configurations is ordered. With one packet buffer at each middle-stage $VOQ_2(j,k)$, the in-order packet delivery property ensures that packets of the same flow always experience the same middle-stage port delay (bounded by $[1,N]$ slots), no matter which middle-stage port it passes through, and/or the actual traffic loading. The joint sequences in Fig. 2b–e are capable of the in-order packet delivery property.

It is shown in Hu and Yeung (2008) that there are total $[(N-1)!]^2$ joint sequences characterized with both staggered symmetry and in-order packet delivery properties. As long as any one of them is utilized, the feedback-based switch not only