Dynamic Deficit Round-Robin Scheduler for 5-Tb/s Switch Using Wavelength Routing
Kimihiro Yamakoshi, Eiji Oki, and Naoaki Yamanaka
NTT Network Service Systems Laboratories
3-9-1 Midori-Cho, Musashino-Shi, Tokyo, 180-8585 Japan
E-mail: {yamakoshi.kimihiro, oki.eiji, yamanaka.naoaki}@lab.ntt.co.jp

Abstract—A dynamic deficit round-robin (DDRR) scheduling scheme for a 5-Tb/s switch is proposed. DDRR is a DRR-based packet scheduler and it can satisfy the max-min fair share. However, DRR cannot satisfy both high throughput and delay requirements. DDRR resolves the problem by changing the granularity for deficit counters according to the packet lengths at queue heads. Simulation results showing the efficiency of DDRR are presented and an implementation of DDRR for the switch is also described.

Keywords—DRR, scheduler, packet, switch, deficit, delay.

I. INTRODUCTION

With the recent introduction of new access networks such as ADSL, FTTH, and wireless LAN, the volume of Internet Protocol (IP) data traffic on the access networks is growing year-by-year. The growth of the access networks will also drive an increase in data traffic on the backbone network. As a result, the volume of IP data traffic is growing 2-3 times per year. Commercial switching systems for the backbone network now have throughputs of hundreds or gigabits per second. This means that a Tb/s-class switching system [1], [2] for the backbone network will be required in the near future if the data traffic continues to increase at the same pace.

Most high-speed packet switching systems including IP routers use a fixed-sized cell in the switch fabric. Variable-length packets are segmented into several fixed-sized cells when they arrive, switched through the switch fabric, and reassembled into packets before they depart [3]. However, the cost of reassembling cells into packets is fairly high, because cells switched based on the cell-based scheme are likely to be interleaved. In the packet-based switching scheme, the switched cells belonging to a packet are not interleaved. We call such cells concatenated cells. A conventional scheduler for variable-length packets, such as deficit round-robin (DRR) [4], can provide the max-min fair share independent of packet length and such schedulers are widely used for their simplicity. However, DRR cannot support both short-packet delay requirements and high throughput simultaneously because the granularity for the deficit counter is fixed. A degradation in short packet delay is likely to affect the quality of real-time services such as VoIP. Instead of a conventional DRR scheduler, a new scheduling algorithm, called dynamic deficit round-robin (DDRR), is introduced in order to resolve the above drawbacks. DDRR is based on DRR but is modified by using a technique in which the granularity for the deficit counter in the scheduler is varied adaptively according to the packet lengths at the heads of queues. We have shown that DDRR can satisfy both high throughput and small delay requirements for short packets while providing the max-min fair share independent of packet length.

We have been developing an experimental 5-Tb/s switch that has a three-stage switch architecture and uses packet-by-packet wavelength routing between switch elements in each stage [5]. To reduce the cost of reassembling cells into packets at the output line-interface card, packet-based wavelength routing is done at the wavelength-division-multiplexing (WDM) link between each stage. Therefore, a packet scheduler is necessary to arbitrate variable-length packets from eight wavelengths into an input port of switch elements at the 2nd and 3rd stages. We used DDRR for the input ports of switch elements at the 2nd and 3rd stages.

In the following sections, after briefly reviewing the architecture of the 5-Tb/s switch, we described the problem of DRR and explain the algorithm of DDRR for resolving it. Simulation results showing the efficiency of DDRR are presented and an implementation of DDRR for our 5-Tb/s switch is also described.

II. ARCHITECTURE OF 5-TB/S SWITCH

Figure 1 shows the architecture of the 5 Tb/s switch. The switch has a 3-stage switch architecture the same as the 640-Gb/s switch that was its predecessor [6]. In each stage, the 5-Tb/s switch deploys eight times as many switch elements as the 640-Gb/s switch. The switch has a 3-stage switch architecture the same as the 640-Gb/s switch. The eight horizontally expanded switch elements at the 1st stage and those at the 2nd stage are interconnected by WDM grouped links.

Connections are distributed in the 1st stage, and output to the 2nd and 3rd stages. When a connection from the switch element at the 1st stage to the switch element at the 3rd stage is requested, the route that traces the switch element at the 2nd stage is selected so as to equalize the traffic in each wavelength. Therefore the load is balanced for all 64 route combinations. At the front of the 1st stage switch elements, a packet is segmented into fixed-sized cells. The cells are switched as concatenated-cells at each stage to avoid cell interleaving. Here, we define concatenated-cells as the cells that are not interleaved.

A WDM grouped link consists of WDM packet transmitter modules, a router AWG, and WDM packet receiver modules. Figure 2 shows an 8 x 8 WDM grouped link. Each link consists of 2.5 Gb/s x 8 wavelengths. At a WDM packet transmitter module, the 10-Gb/s bandwidth of an input port of
Fig. 1. 5-Tb/s switch architecture.

Fig. 2. WDM grouped link.
Switch elements is shared among eight wavelengths. A WDM packet transmitter module consists of a wavelength assigner and WDM transmitter. The wavelength assigner assigns a specific wavelength to a packet, which should be sent to the destined WDM packet receiver module by that wavelength. A WDM transmitter converts an electrical signal into an optical signal with one of the eight wavelengths. A WDM packet receiver module consists of a WDM receiver and packet regenerator. A WDM receiver demultiplexes the 8-wavelength optical signals into individual wavelengths and converts them to electrical signals. The packet regenerator regenerates packets and arbitrates the emission of packets of eight wavelengths to the input port of a switch element at the next stage.

A scheduler for the arbitration should satisfy the fair share for packets of any length. Moreover, high throughput and delay requirement are important. The next section discusses the DDRR scheduler.

III. DDRR Packet Scheduler

As mentioned in Section I, the DRR scheme cannot satisfy both short-packet delay performance and high throughput, because only a fixed granularity is allowed for the deficit counter. When the granularity is set too large, short packets are delayed significantly by long packets. On the other hand, the throughput becomes low for a small granularity because the timing margin for the DRR scheduler to satisfy the line speed is reduced as the granularity becomes small. Figure 3 represents operations of DRR for these two cases in the case of three queues. A packet can be served in one hop within one packet time in the order of queues when \( G = 9 \). Therefore a short packet in queue \#3 is delayed by long packets in queues \#1 and \#2. On the other hand, a short packet in queue \#3 is served first when \( G = 4 \). However, three hops are required to serve it.

![Figure 3. Operations of DRR.](image)

Figure 4 shows the dependence of average packet delay on packet length in the case of DRR with granularity \( G \) of 5, 10, and 100 cells. Here we assumed that the packet length was normalized to an integer number of fixed-sized cell times. The lengths of arriving packets were assumed to have an exponential distribution with the average length being ten cells. The probability of packet arrival was assumed to have a Poisson distribution. Short packets are delayed as the granularity becomes large \( (G = 100) \). Figure 4 shows that it is necessary to use a small granularity \( (G \leq 10) \) in order to suppress the short packet delay. However, the work of DRR remains \( O(1) \) only if \( G \) is greater than the maximum packet length \( [4] \). The number of queue-traversing hops needed to decide the output queue is likely to become large as the granularity becomes small, as shown in Figure 5. For example, three hops are required to serve the first packet in queue \#3, as shown in Figure 3(b). For \( G = 10 \), the number of hops may be greater than 1 in Figure 4. The average and maximum values are 1.43 and 6, respectively. This causes the throughput to become low, because there is an upper limit on the hop number at a given line speed. If \( G \) is set to more than 50 cells, the number of hops is exactly one. However, short packets have a larger delay than long packets, as shown in Figure 4.

To resolve this drawback, we propose a dynamic deficit round-robin (DDRR) scheduling scheme that dynamically changes the granularity for the deficit counter according to the packet lengths at queue heads. The DDRR algorithm is as follows.

**Step 1.** Calculate the difference \( m_i = l_i - d_i \), where \( l_i \) is the packet length at the head of the queue in channel \( i \) \( (1 \leq i \leq N) \) and \( d_i \) is the counter value for the channel.

**Step 2.** Determine the minimum value \( m_0 \) from \( m_i \).

**Step 3.** Add \( m_0 \) to the counter of each channel except the selected channel.

All \( d_i \)s are set to 0 as the initial condition. In step 1, \( m_i \) is calculated only when there is a packet at the head of the queue in channel \( i \). Three steps must be performed within one packet-time in order to meet the line speed. When the number of
channels is \( N, \log_2 N \) operations determining the smaller of two values \((m_i, m_j)\) are required to decide \( m_0 \) in step 2. This relieves the timing margin bottleneck of DRR for small granularity \((G \leq 10)\). Figure 6 shows the operation of DRR for three queues. The queue that has the shortest packet at the headers of the line has the priority to send out a packet if deficit values are the same. The shortest packet is served first, as shown in Figure 6.

Packet delay decreases linearly with decreasing packet length in DRR, as shown in Figure 4. Moreover, the delays of packets shorter than 15 cells were smaller than those in DRR for any granularity. Therefore, DRR can satisfy both high throughput and small delay requirements for short packets, especially under a heavy load. Figure 7 shows the dependence of 99% packet delay on the granularity of the deficit counter in DRR. Above the granularity of 10 cells, the delay decreases as the granularity decreases. This is because the short packet delay decreases. Below the granularity of 10 cells, the delay increases as the granularity decreases. This is because sometimes a packet cannot be sent out in one cell time even though the round-robin pointer of the scheduler traverses all the queues. The delay of DRR is larger than that of DRRR for the entire range of granularity. Figure 8 shows the delay probability distributions of DRR and DRRR. The tail of the distribution curve for DRRR falls faster than that of DRR. Therefore, the delay performance of DRRR is superior to DRR for all values of granularity.

Note that since DRRR is based on DRR, it also provides the max-min fair share independent of packet length. We used DRRR as the scheduler at the packet regenerator of our WDM packet receiver module.

IV. IMPLEMENTATION

A field programmable gate array (FPGA) was used to implement DRRR for the packet regenerator.

To ensure the operation of the FPGA at 78MHz, multiplex and demultiplex functions are used as shown in Figures 9 and 10. 622-Mb/s input electrical signals are first 1:2 demultiplexed at the input interface on the board, then 1:4 demultiplexed inside the FPGA. The output signals from the FPGA are first 4:1 multiplexed, and then 2:1 multiplexed into 622-Mb/s signals. The variation in wiring length on a board is designed to be at most 2 mm in order to keep the timing margin for the clock and data signals. Figure 11 shows an overview of the FPGA boards for the wavelength assigner and packet regenerator.
V. CONCLUSION

We proposed a dynamic deficit round-robin (DDRR) scheduler for a 5-Tb/s switch using an expanded 3-stage switch architecture. Inter-stage switch elements are connected by WDM grouped links which can optically route a packet as concatenated fixed-sized cells with a specific wavelength. To arbitrate variable-length packets, we use DDRR for the WDM packet receiver module of a WDM grouped link. Simulation results show that DDRR can satisfy both short-packet delay requirements and high throughput simultaneously.

REFERENCES