CONFERENCE RECORD *Special Section on the APCC2001 Proceedings*

Fairness Evaluation of the Non-Independent System **

Hakyong Kim†, *Member*, **Seong Jun Kang**††, *Nonmember*, *and* **Kiseon Kim**†, *Member*

Summary
Fairness evaluation is one of important considerations in the performance study of shared systems. To date, fairness of a shared system has been evaluated quantitatively by using the well-known fairness index defined by Jain in 1984 [1]. However, the fairness index defined by Jain can be applied only to the fairness evaluation of independent shared systems. When we try to apply Jain's fairness index to a non-independent shared system, it causes a problem of underutilization of the system resources. In this paper, therefore, we newly define a fairness index based on the Jain's approach so as to evaluate the fairness of the input-queued switch, an example of the non-independent shared system. The fairness index defined in this paper preserves all the characteristics of Jain's fairness definition and makes it possible to efficiently use the system resource.

Key words:
fairness, fairness index

1. Introduction

Fairness is one of important considerations in most performance studies. Fair allocation is important either particularly in distributed systems, where system resources are to be shared among a number of users, or in the case that a number of connections are to contend for a limited number of servers. Therefore, many researches on fairness have been performed [1,2,3,4,5] and, Reference [5] lists five different fairness criteria with their characteristics.

When we consider the fairness of a system or an application, we may say that the system is fair or not fair. This qualitative description, however, is not intuitive. By contrast, if we describe a system quantitatively, e.g., 90%-fair or 10%-unfair system, even non-experts may understand the fairness of a system. Nevertheless, it is not easy to find studies on fairness itself and the method measuring the fairness quantitatively except Jain's [1,2,3]. If any, they tend to be either qualitative or too specific to a particular application or a particular metric. A brief review and references on those studies is found in [1]. In this paper, therefore, we consider the quantity-wise fairness.

In [1], Jain *et al.* divided the fairness problem into two parts. One is to select an appropriate allocation metric and the other is to define a formula that gives a quantitative value to the fairness of the allocation. While the authors left the selection of the metric open since it depends on the application and desires of users, they concentrated on the second aspect of fairness problem. That is, they defined "fairness" and developed a fairness index function that can measure the fairness of the allocation and, if not fair, it tells how far the allocation is from fairness. They also discussed a set of desired properties of a fairness index, which will be reviewed briefly in Section 2.

The fairness concept and fairness index defined by Jain was used in measuring the fairness of scheduling algorithms developed for the early packet switches constructed with output-queueing scheme [4]. In the output-queued switches, every packet arriving at input ports is delivered to their respective destination output ports without queueing or delay by queueing. In each output port, packets contend for the server which is able to transmit a limited number of packets for a specific time duration. Since the output-queued switching system can be considered a sort of shared system, the fairness concept and fairness index defined by Jain could be used without any modification.

However, considerable increase in demands for high-speed packet switching have changed the switch architecture into the input-queued switch which operates at the same speed as the external link speed. The input-queued switch has no memory in its output ports and, therefore, we may consider it as a shared system where multiple connections share an input port. Differently from the output-queued switch, however, connections of a given input port have to compete with those connections or packets destining for the same output port from all input ports. Of course, connections in an input port

Manuscript received March 31, 2001.

†The authors are with Department of Information and Communications, Kwang-Ju Institute of Science and Technology (K-JIST), 1 Oryong-dong, Buk-gu, Gwangju, 500-712, Korea. Web site: http://charly.kjist.ac.kr/~hykim/

††The author is with Department of Information and Communication Engineering, Mokpo National University, Chungkye-myon, Muan-gun, Chonnam, Korea.

**This work was supported in part by the Korea Science and Engineering Foundation (KOSEF) through the Ultrafast Fiber-Optic Networks Research Center at Kwangju Institute of Science and Technology.

have to compete with themselves to be selected for service. It implies that input ports influence on one another in terms of packet service, which makes it impossible to apply the Jain's fairness definition to the input-queued switch as it is. In this paper, therefore, we will consider a method of the quantitative fairness measure in the input-queued switch by modifying the Jain's definition.

The paper is organized as follows: In Section 2, we review the fairness concept and fairness index defined by Jain and a set of desirable properties that the fairness index should have with itself. In Section 3, we first point out the problems which may be caused when we apply the Jain's fairness definition directly to the non-independent shared system. Based on the discussion, we newly define fairness concept and fairness index in terms of system's utilization. In Section 4, we consider the extension of the new fairness definition.

2. Jain's Fairness and Fairness Index

In [1,2,3], Jain measured the fairness of a shared system by comparing the normalized resource allocation x_i of entity i under observation, where x_i is the ratio of the measured throughput (y_i) to the ideal throughput (z_i), i.e., $x_i = y_i/z_i$. In other words, a shared system under observation is said to be fair if the system has all x_i's are equal to 1.0 [2,3], while it is said to be unfair if not equal. However, since it is impossible to quantitatively measure the fairness by comparing the values of all x_i's in itself, he defined a fairness index (FI) as shown below:

$$FI = \frac{\left(\sum x_i\right)^2}{n \times \sum x_i^2}. \tag{1}$$

This fairness index measures the equality of system resource allocations to each entity. If all entities get the same allocation, i.e., all x_i's are equal, then the fairness index becomes to 1.0 and the system can be said 100% fair. As the disparity among x_i's increases, the value of the fairness index decreases, implying that the system under observation becomes unfair. If the fairness index is equal to 0.0, the system is completely unfair. The fairness index of Eq.(1) can be expressed using the mean and variance by dividing both the nominator and denominator of Eq.(1) by n^2 :

$$FI = \frac{\left(\frac{1}{n}\sum x_i\right)^2}{\frac{1}{n} \times \sum x_i^2} = \frac{(E[X])^2}{E[X^2]} = \frac{(E[X])^2}{(E[X])^2 + var(X)}. \tag{2}$$

Eq.(2) can also be expressed in terms of the coefficient of variance (C_X) the correlation coefficient as shown below:

$$FI = \frac{1}{1 + var(X)/\left(E[X]\right)^2} = \frac{1}{1 + C_X}. \tag{3}$$

Discussed later, Eq.(1) takes the most intuitive and easy-to-understand form, since the variance $var(X)$ in Eq.(2) can have different values depending on the unit of the random variable X and the coefficient of variance in Eq.(3) is a unbounded value and is reciprocally and nonlinearly proportional to the fairness index.

In [1,2,3], Jain categorized the desirable properties of the fairness index. They are as follows:

- Population size independency: The index should be applicable to any number of entities.
- Scale and metric independency: The index should be independent of unit or scale of allocation.
- Boundedness: The index should be bounded possibly between 0 and 1.
- Continuity: The index should be continuous.
- Intuitiveness: The index should be quantitative and should have a direct or linear relationship to fairness.

More detailed discussions on the properties of the fairness index and relevant examples are found in the three references of Jain's. The fairness index in Eq.(1) satisfies these five properties at the same time.

3. Fairness and Fairness Index in Non-Independent Shared Systems

The fairness concept and the fairness index reviewed in Section 2 have been widely used in evaluating the fairness among a number of users sharing a limited resource of a distributed system or the fairness among connections contending for the limited capacity of a server. Even though Jain did not mention in his works, he assumed that a shared system is a single isolated system or independent of other systems within a global system. In other words, the entities of a given system cannot affect the service or resource allocation of the entities in other systems. For the reason, Jain's fairness index could be used in evaluating the fairness of connections in the output-queued switching system as well [4].

However, we cannot use the Jain's fairness index directly to the shared system which is not independent of other systems within the same global system. If Jain's fairness index is applied to a non-independent shared system, such as input-queued switchs, it could lower or deteriorate the global system's utilization even though it might improve the sub- or local system's utilization. This is also pointed out in our previous study [6].

In this paper, therefore, we define a new fairness index considering both terms of fairness and the global system's utilization. The fairness index developed here can

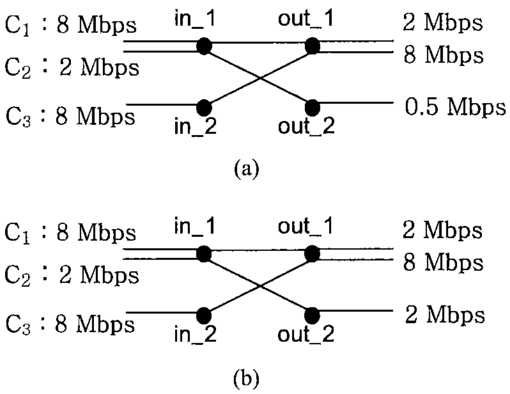

Fig. 1 *Resource allocation in an input-queued switch with an unfair output scheduler (a) based on the conventional fairness definition (b) based on the new fairness definition. Servers at input and output ports have the maximum service capability of 10 Mbps.*

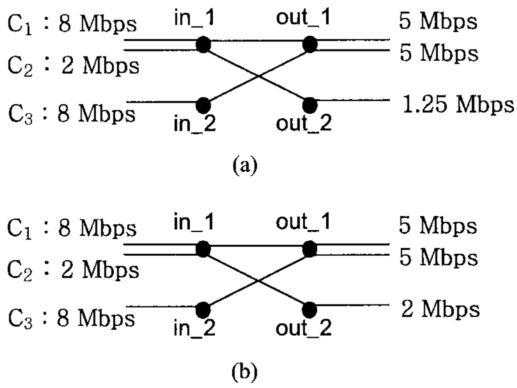

Fig. 2 *Resource allocation in an input-queued switch with a fair output scheduler (a) based on the conventional fairness definition (b) based on the new fairness definition. Servers at input and output ports have the maximum service capability of 10 Mbps.*

applicable to non-independent shared systems as well as independent shared system. However, since the application to the independent shared system is trivial, we adhere to the application to the non-independent shared system. For the purpose of easy understanding, we choose the input-queued switch for a representative example of the non-independent shared system. In fact, fairness is used as an important performance metric in evaluating the performance of switching systems. In the input-queued switch under observation, each input port corresponds to the local system and the input-queued switch itself corresponds to the global system.

In the new fairness index, the normalized resource allocation x_i is defined as the ratio of the measured throughput (y_i) to the minimum resource allocation ($\min(z_{i\_in}, z_{i\_out})$). That is,

$$x_i = \frac{y_i}{\min(z_{i\_in},\, z_{i\_out})}, \tag{4}$$

where z_{i_in} and z_{i_out} designate the resource allocation to the i-th connection or entity at an input port and an output port, respectively. The fairness index is directly obtained by substituting Eq.(4) for x_i in Eq.(1). Since the fairness index is not changed in its form, it preserves all the properties of the Jain's fairness index reviewed in Section 2.

The validity of the new fairness definition can be demonstrated well through an example. The example shown in Fig.1 confirms that the new fairness index maximizes the system's utilization keeping fairness between two connections of input port 1. In this example, we assume that the system under observation is an input-queued switch and, therefore, there is no queue at output ports. Each input and output port can serve with the maximum service rate of 10 Mbps even though it is rare for the servers at input and output ports to use their maximum capacity. We further assume that the scheduler allocates 2 Mbps for connection C1 from input 1 to output 1 and 8 Mbps for connection C3 from input 2 to output 1. Based on the Jain's fairness index, connection C2 from input 1 to output 2 should serve 0.5 Mbps for the sake of fairness between connections C1 and C2 as shown in Fig.1.(a), even though output 2 can serve up to 10 Mbps. Based on the new fairness index, by contrast, connection C2 can be allocated 2 Mbps as shown in Fig.1.(b), since the server at input 1 allocates 2 Mbps (= 10 Mbps - 8 Mbps) and the server at output 2 tries to allocate its maximum service rate. In both cases, the fairness index at input 1 is equal to 1.0. The calculation of the new fairness index is shown below:

$$x_1 = \frac{2}{\min(8,\, 2)} = 1.0, \quad x_2 = \frac{2}{\min(2,\, 10)} = 1.0$$

$$FI = \frac{(2.0)^2}{2 \times (1.0 + 1.0)} = 1.0$$

Notice, however, that the new fairness definition makes the system serve 1.5 Mbps more.

Assume that a fairer scheduler is adopted for output port and, thereby, service capacity of 5 Mbps is assigned both to connection C1 and connection C3. As shown in Fig.2.(a), then, the scheduler based on the conventional fairness definition by Jain assigns 1.25 Mbps to C2 to keep the fairness between the two connections. By contrast, the scheduler based on the new fairness definition assigns 2 Mbps to C2 making full use of the system's capability. The fairness index for input port 1 is still equal to 1.0 in both cases even though different definitions are used. Notice, however, that the new fairness definition maximizes the system's utilization over the conventional definition.

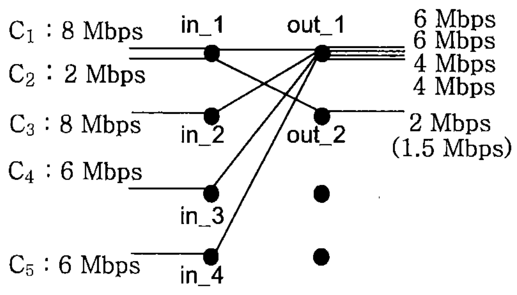

Fig. 3 *Resource allocation in an input-output-queued switch. Servers at input and output ports have the maximum service capability of 10 Mbps. Speedup factor (s) is equal to 2.*

4. Extention of the Fairness Definition

The fairness concept introduced in the previous section can be applied to the combined input-output queued (CIOQ) switch. Since the CIOQ switch has queues at output ports as well as at input ports, output ports can receive as much data traffic as the internal speedup factor even though they can still send out the traffic to the external link at the same speed as the external link speed. For the CIOQ switch, therefore, the normalized resource allocation x_i can be defined as

$$x_i = \frac{y_i}{\min(z_{i\_in},\ s \times z_{i\_out})}, \quad (5)$$

where s designates the speedup factor of the CIOQ switch. In general, since the switching system is symmetrical in terms of service capability of input and output ports (i.e., $z_{i\_in} = z_{i\_out}$), x_i in Eq.(5) has the same value as x_i in Eq.(4). However, when z_{i_out} is smaller enough than z_{i_in}, the internal speedup factor, s, brings about different results.

An example of the CIOQ switch is shown in Fig.3. As in the example of Section 3, we assume that the servers at input and output ports have the maximum service rate of 10 Mbps. However, each output can receive traffic at the rate of 20 Mbps since we use the internal speedup of factor 2 (i.e., $s = 2$). Therefore, connections C1, C3, C4, and C5 are assigned 6 Mbps, 6 Mbps, 4 Mbps, and 4 Mbps by a fair output scheduler, respectively. C2 is assigned 2 Mbps by the new fairness definition. When we consider the conventional fairness definitoin, C2 is assigned 1.5 Mbps. Fairness indexes for both cases are availabe immediately.

The fairness concept and fairness index is also applicable to a more complicated shared or distributed system such as multi-stage switching system. In this case, the fairness index might be obtained by simply extending the normalized resource allocation in Eq. (4) as

$$x_i = \frac{y_i}{\min(z_{i\_1},\ z_{i\_2},\ z_{i\_3},\ \cdots\ z_{i\_k})},$$

where z_{i_k} implies the allocated resource to the i-th entity at the k-th stage.

Thus far, we have addressed the fairness in a sub- or local shared system. Even though we don't discuss here in detail, the fairness of global system is also important. The global fairness is the fairness among the entities in the global system, not in the sub- or local system. For example, the global fairness of the system shown in Fig.1.(b) is equal to 1.0, since all x_i's, $1 \leqq i \leqq 3$, have the same value of 1.0. On the other hand, the global fairness of the system shown in Fig.1.(a) is equal to 0.667, since x_1 and x_2 are equal to 0.25 and x_3 is equal to 1.0. Note that the new fairness definition makes the whole system fairer.

5. Conclusion

The Jain's fairness index in [1] has been widely used in evaluating the fairness of the independent shared system quantitatively. The fairness index, however, can not be applicable to the non-independent shared system any more since it makes the shared system not to be fully utilized. In this paper, therefore, we newly defined a fairness index based on the Jain's index in [1]. In the new fairness index, the normalized resource allocation x_i was the ratio of the measured throughput (y_i) to the minimum resource allocation ($\min(z_{i\_in},\ z_{i\_out})$), i.e., $x_i = y_i/\min(z_{i\_in},\ z_{i\_out})$. The fairness index newly defined in this paper preserves the 5 properties that a fairness index should have, plus the characteristics of maximizing the systems' resource utilization. The concept used in defining the new fairness index can be used extensively in evaluating the fairness a bunch of communication channels or connections passing through a knotty shared system. The new fairness index can be used as a criterion which maximizes the system's utilization as well as keeping the fairness among sub-systems.

References

[1] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer system," *Digital Equipment Corporation, Technical Report, DEC-TR-301*, Sep. 26, 1984. Available at http://www.cis.ohio-state.edu/~jain/papers/fairness.htm

[2] ATM Forum Document Number: ATM Form/94-0881.

[3] ATM Forum Document Number: ATM Forum/99-0045.

[4] M. Nabeshima and N. Yamanaka, "New scheduling mechanisms for achieving fairness criteria (MCR plus equal share, maximum of MCR or Max-Min share)," *IEICE Tr. Commun.*, vol. E82-B, no.5, June 1999, pp. 962-966.

[5] ATM Forum, "Traffic management specification 4.0," af-tm-0056.0, 1996.

[6] Hakyong Kim, Y.T. Lee, and K.S. Kim, "Fairness concept in terms of utilization," *IEE Electronics Letters*, vol.36, no.4, Feb. 17, 2000, pp. 379-381.

[7] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed., McGraw-Hill, Inc., 1991.