A Survey of Nonblocking Multicast Three-Stage Clos Networks

Frank K. Hwang, National Chiao-Tung University

ABSTRACT

The author gave a survey on multicast nonblocking multistage interconnection networks in his 1998 book. Here he focuses on the three-stage Clos network and its recursive extensions. Not only will this article bring the literature up to date, but it also will provide some fresh viewpoints to either clarify or simplify some issues.

INTRODUCTION

In multicast traffic, an input can request to connect to up to a certain number of outputs. If that number is specified to be f, it is f-cast traffic. If that number is unconstrained, it is broadcast traffic. Usually, it is assumed that in multicast traffic, calls come and go sequentially. Suppose (i,O) is the current multicast call where i is an input and O is a set of idle outputs. Then none of the outputs in O can appear in any other existing connection. However, i can be either idle or busy depending on the model. If i must also be idle, we call it closed-end traffic; if i can be busy, we call it open-end traffic. In the latter case, the various calls from i must carry the same message. Sometimes, all traffic is routed simultaneously. Then there is no difference between closed-end and open-end traffic since we can always combine calls from the same input into one call.

In a multistage interconnection network each stage consists of r_i crossbars of the same size, and links exist only between adjacent stages. We will let n_1 denote the number of inlinks of an input switch, n_2 the number of outlinks of an output switch, and $r_1(r_2)$ the number of input (output) switches. Then $N_1 = n_1r_1$ is the number of network inputs and $N_2 = n_2r_2$ the number of network outputs. The most studied multistage interconnection network is the three-stage Clos network $C(n_1, r_1, m, n_2, r_2)$, where m is the number of middle-stage switches (Fig. 1).

A symmetric three-stage network with $n_1 = n_2 = n$, $r_1 = r_2 = r$, is denoted by $C(n, m, r)$. Note that a three-stage Clos network can be recursively extended to a $(2k + 1)$-stage network by replacing each switch in a given stage with a three-stage Clos network.

In a nonblocking network, all requests can be connected, meaning calls from different inputs have link-disjoint paths. There are three levels of nonblockingness. If a request can be connected regardless of how previous calls were connected, the level is strictly nonblocking. If a request can always be connected as long as all connections follow a given routing algorithm A, the level is wide-sense nonblocking. If a request can be connected when paths of existing connections can be rerouted to make way, the level is rearrangeably nonblocking. Another way of interpreting rearrangeably nonblocking is that any set of requests can be simultaneously routed. There are also variations such as standard path nonblocking and repackably nonblocking [1], which we will not discuss in this article.

Masson and Jordan [2] first introduced the notion of a multicast multistage interconnection network on three-stage Clos networks. They gave sufficient conditions for both strictly and rearrangeably nonblocking. However, the strictly nonblocking result is really wide-sense nonblocking since the routing assumes that outputs from the same output switch in a multicast call will share a path until reaching the output switch. Such a practice has been referred to in the literature as the no-split rule, which defines a routing algorithm. Using the no-split rule, broadcast traffic can be treated as r_2-cast traffic.

To compare the cost of different networks, the number of crosspoints is still a popular measure since even if it is not the real cost, it is a good figure of merit. In such comparisons, we also assume $N_1 = N_2 = N$ to reduce the number of parameters.

The multicast multistage interconnection network, in particular the three-stage Clos network, has been widely studied due to its many applications in videoconferencing, video on demand, e-commerce, parallel computing, and so on. In this survey we focus on the multicast nonblocking three-stage Clos network and its recursive extensions. The reader is referred to [3] for general terminology.

This research is partially supported by the Republic of China National Science Council grant NSC-91-2115-M-009-010 and the MTI-Lee Center.

0163-6804/03/$17.00 © 2003 IEEE

Strictly Nonblocking

Friedman [4] proved a fundamental result for a general multicast strictly nonblocking network (not necessarily a multistage interconnection network).

Theorem 1 — A strictly nonblocking network with closed-end broadcast traffic has at least $O(N^2)$ crosspoints.

Since an $N \times N$ crossbar is broadcast strictly nonblocking and has N^2 crosspoints, any hope to design an unbounded multicast strictly nonblocking network to save cost is dashed, but one can still save cost on f-cast networks.

Although a sufficient condition on strictly nonblocking three-stage Clos networks was attempted early [2] (but turned out to be a condition on wide-sense nonblocking), genuine necessary and sufficient conditions were given much later. Actually, several sets of conditions were given that were not in complete agreement. Besides, there are different hardware models to consider. It is typical to assume that a crossbar has fan-out capability (i.e., a crossbar itself is a strictly nonblocking multicast network). But we may restrict the crossbars of a given stage to be without the fan-out capability (i.e., only capable of point-to-point connection). Presumably, such a crossbar has lower cost due to either less hardware or a simpler control mechanism.

Let model i, $i = 1, 2, 3$ denote the hardware model in which stage i has no fan-out capability, and let model 0 denote that in which every stage has fan-out capability.

It is messy to get the exact necessary and sufficient conditions for these methods (which explains why there are several different sets of such conditions), especially when one has to take the boundary effect from the input and output sizes into consideration. Recently, the author gave a unifying approach to compute the necessary and sufficient conditions, which simplifies derivation as well as verification [3]. Furthermore, this approach works not only for the above four strictly nonblocking models, but also for some wide-sense nonblocking models. We summarize these findings for the strictly nonblocking model in Table 1.

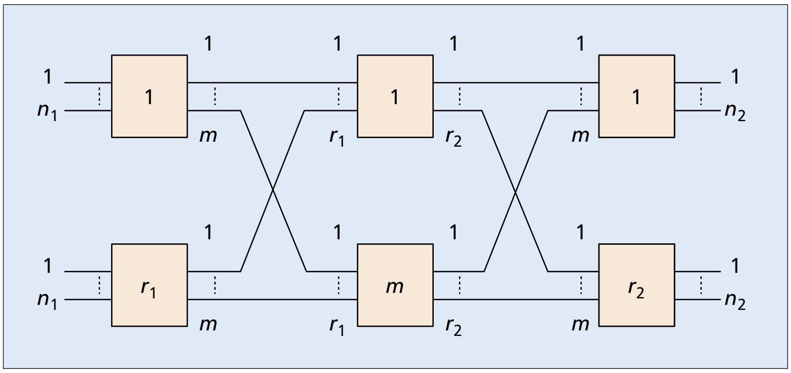

Figure 1. *A three-stage Clos network.*

From Theorem 1, we know that to get a better cost than $O(N^2)$, we must bound f. For $f = r_2$, all strictly nonblocking networks of Table 1 need $O(N^{5/3})$ crosspoints by setting $n_1 = O(N^{1/3})$ and $n_2 = O(N^{2/3})$, but model 1 needs $O(N^2)$ crosspoints. For $C(n, m, r)$, all models need $O(N^2)$ crosspoints.

Wide-Sense Nonblocking

There are three general classes of routing algorithms for a multicast multistage interconnection network. The 0-1 fan-out class is characterized by whether the routing algorithm allows fan-out at a stage. (Note that if a crossbar cannot perform fan-out due to hardware structure, the result is strictly nonblocking.) The size fan-out class is characterized by the specification of the fan-out size at a stage. The window class is characterized by the specification of the partition of a multicast request into multicast subrequests that are independently routed.

The first 0-1 fan-out algorithm, as well as the first wide-sense nonblocking multicast algorithm, is the no-split algorithm of Masson and Jordan on a three-stage Clos network. They gave a sufficient condition for wide-sense nonblocking. Hwang used the unifying approach introduced earlier to obtain a necessary and sufficient condition similar to those in Table 1.

Note that the no-split rule plus the fan-out capability of the output switch guarantee that only one path is needed to reach all outputs in the same output switch in an f-cast call. Thus, $f_2 > r_2$ can be reduced to $f_2 = r_2$. Therefore, it is not surprising that the cost under the no-split rule is on the same order as those strictly nonblocking networks in Table 1 with $f_2 = r_2$.

While the 0-1 fan-out class was started acci-

Fan-out availabilty	Traffic model	Necessary and sufficient condition
Fan-out capability at every stage	Closed end	$m \geq \min\{(n_1 - 1)f + n_2, (N_1 - 1)f + 1, N_2\}$
	Open end	
No fan-out at stage 1	Closed end	$m \geq \min\{N_2 - r_2 + (n_1 + n_2 - 1)/n_2, N_1\}$
	Open end	∞
No fan-out at stage 2	Closed end	$m \geq \min\{(n_1 - 1)f + n_2 - 1 + \min\{f, r_2\}, (N_1 - 1)f + \min\{f, r_2\}, N_2\}$
	Open end	$m \geq \min\{n_1 f + n_2 - 1, N_1 f, N_2\}$
No fan-outat stage 2	Closed end	$m \geq \min\{(n_1 - 1)f + n_2, (N_1 - 1)f + \min\{f, n_2\}, N_2\}$
	Open end	

Table 1. *Conditions for* f*-cast strictly nonblocking Clos networks.*

dentally by a “strictly nonblocking” result of Masson and Jordan, the size fan-out class was started accidentally by a rearrangeably nonblocking result of Kirkpatrick, Klawe, and Pippenger [5]. This algorithm was later interpreted to be a wide-sense nonblocking algorithm that sets an upper bound of either n_2 or $\lceil \log_2 n_2 \rceil$ on the fan-out size at the input stage [3].

Note that the no-split rule plus the fan-out capability of the output switch guarantee that only one path is needed to reach all outputs in the same output switch in a f-cast call.

Yang and Masson [6] first explicitly proposed a fan-out size algorithm by restricting the fan-out size at the input stage to be at most p (to be optimally determined). The no-split rule is also tacitly assumed. They proved the following.

Theorem 2 — $C(n_1, r_1, m, n_2, r_2)$ is f-cast wide-sense nonblocking for the closed-end traffic under the p-restriction routing if $m > (n_1 - 1)p + (n_2 - 1)f^{1/p}$.

They showed that a correct set of p middle switches can be found in $O(n_2 f)$ time, and the optimal choice of p is $\log f/2 \log \log f$.

At $f = r_2$, the number of crosspoints is $O(N^{3/2} \log r/\log \log r) = O(N^{3/2} \log N/\log \log N)$ obtained by setting $O(n_1) = O(n_2) = O(N^{1/2})$. Yang and Masson also extended it to a $(2k + 1)$-stage Clos network and showed that the number of crosspoints is

$$O\left(N^{1+\frac{1}{(k+1)}} \left[\log N / \log\log N\right]^{(k+2)/2-1/(k+1)} \right).$$

Feldman, Friedman, and Pippenger [7] considered a two-stage network such that the first stage consists of concentrators (a bipartite graph, but not necessarily complete as a crossbar), the second stage of copies of $r_1 \times N_2$ crossbars. In their construction, each pair of inputs share exactly two neighbors. The routing algorithm allows no fan-out at the first stage, and an output of the concentrator can be selected to carry the path only if its selection would not cause any input to have too many busy neighbors(a threshold is defined). They showed that a multicast wide-sense nonblocking three-stage network with $O(N^{5/3})$ crosspoints can be constructed, and also a five-stage extension with $O(N^{3/2})$ crosspoints. They also gave a nonconstructive s-stage version with $O(N^{1+1/s}(\log N)^{1-1/s})$ crosspoints, which is slightly better than Yang and Masson’s construction.

Tscha and Lee [8] proposed the first window algorithm on the multi-$\log_2 N$ network by partitioning the $N = 2^n$ outputs into groups called windows, each consisting of $2^{\lfloor n/2 \rfloor}$ outputs that can reach the same set of switches in stage $\lceil n/2 \rceil + 1$. A multicast request is correspondingly partitioned into several subrequests where outputs from the same window are in the same subrequest. Two subrequests from the same input must have link disjoint paths just as if they were from different inputs. Kabacinski and Danilewicz [9] extended to windows of variable sizes with some comparisons among different sizes. Unfortunately, the number of crosspoints is $O(N^2 \log N)$.

Hwang applied the window algorithm to the three-stage Clos network. Again, the author used the unifying approach to obtain a necessary and sufficient condition, and also showed that $\sqrt{r}$ is a near-optimal choice of window size for $C(n, m, r)$, while $m \geq (2n — 1) \sqrt{r}$ is a sufficient condition.

Note that $C(n, (2n - 1) \sqrt{r}, r)$ has $O(N^{7/4})$ crosspoints by setting $n = O(N^{1/2})$. Setting the window size to r_2, the window algorithm equals the routing in model 1. Setting the window size to 1, the window algorithm equals the routing in model 2 except that the no-split rule is in force. Therefore, the window algorithm unites the two models and provides a spectrum of choices in between.

Rearrangeably Nonblocking

Masson and Jordan gave a sufficient condition on multicast rearrangeably nonblocking $C(n_1, r_1, m, n_2, r_2)$. Hwang refined it to be both sufficient and necessary.

Theorem 3 — For model 2, $C(n_1, r_1, m, n_2, r_2)$ is f-cast rearrangeably nonblocking if and only if

$$m \geq \min\{\min\{n_1 f, N_2\}, \min\{n_2, N_1\}\}$$

For $f = r_2$ the network has $O(N^{5/3})$ crosspoints by setting $n_1 = O(N^{1/3})$ and $n_2 = O(N^{2/3})$. Note that the symmetric network would need $O(N^2)$ crosspoints.

Unlike the strictly noblocking case, necessary and sufficient conditions for rearrangeably nonblocking are not known for models 0, 1, and 3. An $O(N^{7/4})$ cost network for model 1 was given by Kirkpatrick, Klawe, and Pippenger. As mentioned before, their rearrangeably nonblocking results for model 0 are actually wide-sense nonblocking results. The cost of either their three-stage network or its s-stage recursive extension is larger than those of Yang and Masson.

Hwang and Lin [10] conjectured: “For model 1, $C(n, 2n, r)$ is 2-cast rearrangeably nonblocking.” Their motivation for studying this conjecture is to provide rearrangeable nonblocking for some occasional bicast calls in an strictly nonblocking point-to -point network. Note that if the conjecture holds, a point-to-point call (i, j) can always be routed without rearranging, since the $(n - 1)$ co-inputs (co-outputs) can occupy at most $n - 1$ middle switches regardless of their calls being point-to-point or bicast. Also, note that this is true only under model 1.

Du and Ngo [11] extended the conjecture to an asymmetric three-stage Clos network with $n_1 \geq n_2$. They also gave a counterexample if $n_1 < n_2$, and proved the conjecture for $n_2 = 2$ or 3. Hwang, Liao, and Tong [12] proved for $r_2 \leq 4$. They also considered the case where the multicast calls do not have to be bicast.

Richards and Hwang [13] considered a network that is mathematically equivalent to a three-stage network where the input switch is of size $1 \times k$, $k \leq m$, and the linking between the first two stages is not a complete bipartite graph (i.e., input i is linked to the set M_i of middle switches). They actually proposed the elimination of the input stage by assigning input i to an inlink of each middle switch in M_i. They called it a two-stage network. For easier comparison with other three-stage networks, we will treat it as a three-stage network, although no crosspoint is counted in the first stage since each input switch can be a splitter.

Note that outputs on different output switches can compete only for stage 1 links, but in the two-stage network each stage 1 link is dedicated to an input and not subject to competition. Therefore, we can study the rearrangeably nonblocking condition as if there were only one output switch. Let B denote the bipartite graph between the first two stages. Then the network is rearrangeably nonblocking if and only if B is a partial concentrator of capacity n_2 (i.e., up to n_2 inputs have at least that many neighbors). By Hall's theorem on a system of distinct representatives, for any set of $q \leq n_2$ inputs requested by the outputs of the output switch, there exist q middle switches, each carrying a distinct requested input to connect to the output switch. The number of crosspoints depends on n_2 as a function of k (not completely determined yet). $O(N^{7/4})$ is provable while $O(N^{5/3})$ is conjectured.

Note that outputs on different output switches can compete only for stage 1 links. But in the two-stage network, each stage 1 link is dedicated to an input and not subject to competition. Therefore, we can study the rearrangeably nonblocking condition as if there were only one output switch.

Conclusions

A survey of multicast nonblocking multistage interconnection networks was given in [1, Ch. 4]. Here we focus on three-stage Clos networks and bring the literature up to date. Some interesting new developments are:

- A unifying approach to compute necessary and sufficient conditions for many multicast strictly nonblocking and wide-sense nonblocking models
- A new class, the window class, of routing algorithms that unites some previous seemingly unrelated models
- Recent progress on the bicast conjecture for model 1

We also provided some fresh viewpoints:

- We provided a more structured framework to study multicast wide-sense nonblocking three-stage Clos networks by classifying the routing algorithms into three classes. In particular, we pointed out that the input stage fan-out size can be controlled by hardware structure.
- We observed the close relation between two problems thus far studied separately: the rearrangeably nonblocking network under model 1; and satisfying the strictly nonblocking requirement for point-to-point traffic but the rearrangeably nonblocking requirement for multicast traffic.

The author wishes to thank the reviewers for careful reading and many helpful suggestions.

References

[1] F. K. Hwang, *The Mathematical Theory of Nonblocking Switching Networks*, World Scientific, Singapore, 1998.

[2] G. M. Masson and B. W. Jordan, Jr., "Generalized multi-stage Connection Networks," *Networks*, vol. 2, 1972, pp. 191–209.

[3] F. K. Hwang, "A Unifying Approach to Determine the Necessary and Sufficient Conditions for Nonblocking Multicast Clos Networks," preprint, 2002.

[4] J. Friedman, "A Lower Bound on Strictly Nonblocking Network," *Combinatorica*, vol. 8, 1988, pp. 185–88.

[5] D. G., Kirkpatrick, M. Klawe, and N. Pippenger, "Some Graph-Coloring Theorems with Applications to Generalized Connection Networks," *SIAM J. Alg. Disc. Methods*, vol. 6, 1985, pp. 576–82.

[6] Y. Yang, and G. M. Masson, "Nonblocking Broadcast Switching Networks," *IEEE Trans. Comp.*, vol. 44, 1995, pp. 1169–80.

[7] P. Feldman, J. Friedman, and N. Pippenger, "Wide-Sense Nonblocking Networks," *SIAM J. Disc. Math.*, vol. 1, 1988, pp. 158–73.

[8] Y. Tscha and K. H. Lea, "Yet Another Result on Multi-$\log_2 N$ Networks," *IEEE Trans. Commun.*, vol. 47, 1999, pp. 1425–31.

[9] W. Kabacinski and G. Danilewicz, "Wide-Sense and Strict-Sense Nonblocking Operation of Multicast Multi-$\log_2 N$ Switching Networks," *IEEE Trans. Commun.*, vol. 6, 2002 pp. 1025–36.

[10] F. K. Hwang, and C. H. Lin, "Broadcasting in a Three-stage Point-to-Point Nonblocking Network," *Int'l. J. Rel. Qual. Safety Eng.*, vol. 2, 1995 pp. 299–307.

[11] D. Z. Du and H. Q. Ngo, "An Extension of DHH-Erdos Conjecture on Cycle-plus-Triangle Graph," *Taiwanese J. Math.*, vol. 6, 2002 pp. 261–67.

[12] F. K. Hwang, S. C. Liaw, and L. D. Tong, "Strictly Nonblocking 3-Stage Clos Network with Some Rearrangeable Multicast Capability," preprint, 2001.

[13] G. W. Richards and F. K. Hwang, "A Two-Stage Rearrangeable Broadcast Switching Network," *IEEE Trans. Commun.*, vol. 33, 1985, pp. 1025–35.

Biography

Frank K. Hwang (fkhwang@cms.zju.edu.cn) obtained his B.A. degree from National Taiwan University in 1960 and his Ph.D. in statistics from North Carolina State University in 1968. He worked at the Mathematics Center of Bell Laboratories from 1967 until 1996 when he retired. He has been a university chair professor at National Chao-Tung University since then. He has published about 350 papers and written four books, including *The Mathematical Theory of Nonblocking Switching Networks*.