Report Number: WUCS-92-13

1992

The Clos Network as a Multirate Distributer with a Greedy Routing Algorithm

Ellen E. White

Yang and Masson [14] have demonstrated that the Clos network is a nearly nonblocking distributer, with the proper choice of network parameters. The resulting network has better asymptotic crosspoint compleixty than other known constructions when the number of stages is fixed. In addition, the routing algorithm is efficient, taking time linear in the number of network inputs to route a new connection. We extend these results to the multirate environment in which each connection has an associated weight indicating the fraction of link bandwidth which it requires. Connections may share a link provided the sum of the weights does... **Read complete abstract on page 2.**

The Clos Network as a Multirate Distributer with a Greedy Routing Algorithm

Ellen E. White

Complete Abstract:

Yang and Masson [14] have demonstrated that the Clos network is a nearly nonblocking distributer, with the proper choice of network parameters. The resulting network has better asymptotic crosspoint compleixty than other known constructions when the number of stages is fixed. In addition, the routing algorithm is efficient, taking time linear in the number of network inputs to route a new connection. We extend these results to the multirate environment in which each connection has an associated weight indicating the fraction of link bandwidth which it requires. Connections may share a link provided the sum of the weights does not exceed 1. The overall complexity of the network is better than other known multirate results when the number of stages is fixed.

The Clos Network as a Multirate Distributor with a Greedy Routing Algorithm

Ellen E. Witte

WUCS-92-13

March 24, 1992

Department of Computer Science
Campus Box 1045
Washington University
One Brookings Drive
St. Louis, MO 63130-4899

Abstract

Yang and Masson [14] have demonstrated that the Clos network is a nearly non-blocking distributor, with the proper choice of network parameters. The resulting network has better asymptotic crosspoint complexity than other known constructions when the number of stages is fixed. In addition, the routing algorithm is efficient, taking time linear in the number of network inputs to route a new connection. We extend these results to the multirate environment in which each connection has an associated weight indicating the fraction of link bandwidth which it requires. Connections may share a link provided the sum of the weights does not exceed 1. The overall complexity of the network is better than other known multirate results when the number of stages is fixed.

The Clos Network as a Multirate Distributor with a Greedy Routing Algorithm

Ellen E. Witte

1. Introduction

In 1953, Clos published a seminal result in the theory of nonblocking switching networks showing that it is possible to construct a connection network with lower asymptotic crosspoint complexity than a crossbar network [2]. Since that time there has been much work in the design of nonblocking networks with a rich set of variations on the basic problem of constructing networks to realize one-to-one connections. In this paper we are interested in networks to realize one-to-many connections, in which a single input is connected to one or more outputs. We call these networks *distribution networks*; they are also referred to as *generalized connectors* or *broadcast networks* elsewhere in the literature. In addition, we are interested in the *multirate* environment, in which each connection has a weight indicating the fraction of link bandwidth required by the connection. Connections may share a link provided the sum of the weights does not exceed 1.

Figure 1 shows the three stage version of the Clos network, denoted C^3. The network is characterized by the parameters n, r and m. The total number of inputs is $N = nr$. The first stage consists of r $n \times m$ crossbar switches. The middle stage consists of m $r \times r$ crossbar switches. The last stage consists of r $m \times n$ crossbar switches. For fixed n and r, the choice of m affects the ability of the network to accomodate connections. The Clos network can be generalized to $2k+1$ stages by replacing each middle stage crossbar switch with a Clos network of $2k-1$ stages. Figure 2 shows the general construction, denoted C^{2k+1}. The general network is characterized by $3k$ parameters, n_i, r_i and m_i, $1 \leq i \leq k$, with the restriction that $n_i r_i = r_{i+1}$ for $1 \leq i < k$. The total number of inputs is $N = n_k r_k$. Additional stages of switching can result in lower total crosspoint counts.

In reference [14] Yang and Masson give conditions under which the Clos network is a nearly nonblocking distribution network. For fixed N they specify the other Clos network parameters so that a greedy routing algorithm can be used to realize one-to-many connections. In this paper we extend their ideas to the multirate environment.

[0]This work was supported by the National Science Foundation, Bell Communications Research, BNR, DEC, Italtel SIT, NEC, NTT and SynOptics. The author was partially supported by an Olin Graduate Fellowship from Washington University.

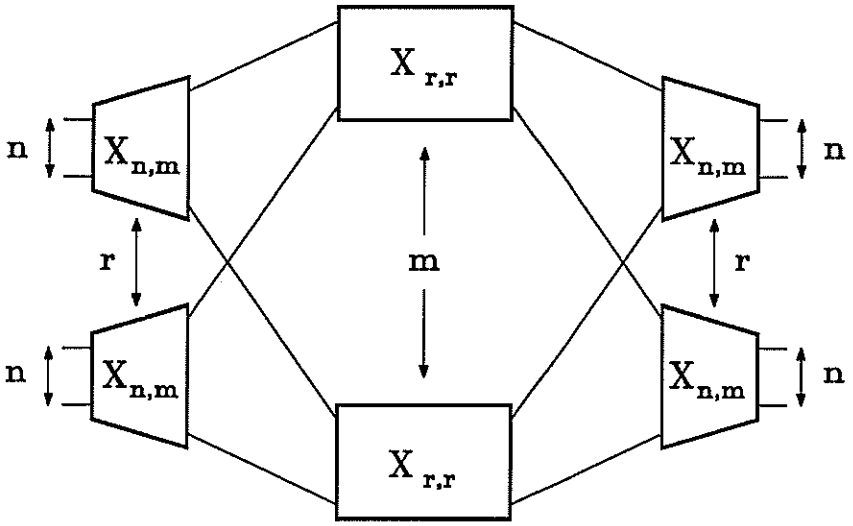

Figure 1: Three Stage Clos Network, C^3

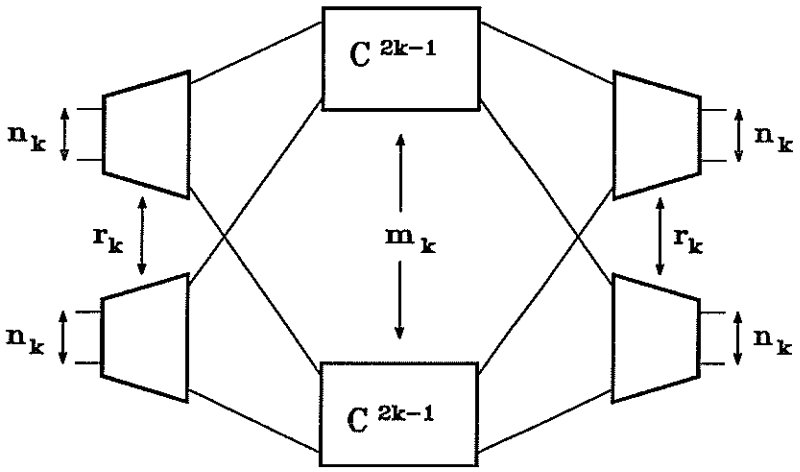

Figure 2: $2k+1$ Stage Clos Network, C^{2k+1}

2. Definitions

As mentioned in the introduction, there is a rich set of variations on the basic problem of constructing connection networks. Describing these variations requires a number of definitions. Many of these are taken from reference [6].

The first type of variation concerns the way in which connection requests are received and routing decisions are made by the network. A *connection request* is a pair (x, y) where x is an input and y is an output. A *connection assignment* is a set of requests in which every input and every output appears at most once. A *connection route* is a list of links forming a path from an input to an output. A route *realizes* a request (x, y) if it starts at x and ends at y. A *state* is a set of routes in which every input and output appears at most

once and every link is used at most once. We say that a state realizes a given assignment if it contains one route realizing each request in the assignment and no others. We say that a state s_0 is *below* a state s_1 if $s_0 \subseteq s_1$. Similarly we say that s_1 is *above* s_0. We say that a connection request (x, y) is *compatible* with a state s if x and y are idle in s.

A network is a *rearrangeable connector* if for every connection assignment there is a state realizing that assignment. Informally, this means that the network can handle requests when presented with the entire set of requests at once. There is no guarantee about the ability to handle requests one at a time without rearranging existing connections. A network is a *strictly nonblocking connector* if for every state s and connection request r compatible with s, there exists a state realizing r that is compatible with s. Informally, this means that regardless of the way in which previous requests have been realized, it is always possible to add any compatible request. A network is a *wide-sense nonblocking connector* if the state space has a subset S (called the *safe states*) such that for every state $s \in S$ all states below s are in S and for every connection request r compatible with s, there exists a route p realizing r that is compatible with s and such that $s \cup \{p\}$ is in S. Informally this means that it is always possible to add any compatible request provided each request is realized in a judicious way. Clearly any strictly nonblocking network is a wide-sense nonblocking network and every wide-sense nonblocking network is a rearrangeably nonblocking network.

The second type of variation concerns the structure of the requests. A connection network supports requests to connect an idle input to an idle output. A *distribution network* supports requests to connect an idle input to a set of idle outputs. Distribution networks are also called *generalized connectors* or *broadcast networks* elsewhere in the literature. Most of the definitions for connection networks extend in the obvious way to distribution networks. We include some of the extended definitions below.

A *distribution request* is a pair (x, Y) where x is an input and Y is a set of outputs. A *distribution assignment* is a set of requests in which every input and output appears at most once. A *distribution route* is a list of links forming a tree whose root is an input and whose leaves are outputs. A route *realizes* a request (x, Y), if its root is x and its leaves are exactly the set Y. There is a second type of request in a distribution network. An *augmentation request* in a state s is a pair (r, y) where $r = (x, Y)$ is a request in the assignment realized by s and y is an output not in Y. An augmentation request is compatible with s if y is idle in s. An augmentation request can be satisfied in s if the route realizing r can be extended by adding unused links so that y becomes a leaf of the route.

The definitions of rearrangeably nonblocking, strictly nonblocking and wide-sense nonblocking can be extended to distribution networks. A network is a *rearrangeably nonblocking distributor* if for every assignment, there exists a state realizing the assignment. A network is a *strictly nonblocking distributor* if for every state s and distribution request r compatible with s, there exists a route realizing r that is compatible with s and if every augmentation request r compatible with s can be satisfied. A network is a *wide-sense nonblocking distributor* if the state space has a safe subset S such that for every state $s \in S$, all state below s are in S; for every distribution request r compatible with s, there exists a route p realizing r that is compatible with s; and every augmentation request r compatible with s can be satisfied in such a way that the resulting state is in S. There is an additional

classification of networks that applies to distribution networks. A network is a *nearly nonblocking* distributor if it is wide-sense nonblocking with respect to distribution requests but not augmentation requests. Nearly nonblocking distributors have power in between that of wide-sense nonblocking and rearrangeably nonblocking distributors.

The third type of variation concerns the bandwidth requirements of the connections. The classical theory was developed at a time when communication systems were operated in a circuit switching mode. When a connection was established, the necessary resources were dedicated to the connection and remained dedicated for the duration of the connection. The classical theory assumes that each connection uses all of the bandwidth available on the required links. More recently there has been considerable work on packet switching communication systems that allow connections to share bandwidth on a link, provided the sum of the bandwidth for all the connections sharing a link does not exceed the link bandwidth. The classical theory was extended by Melen and Turner to associate a weight with each connection equal to the fraction of link bandwidth the connection requires [5]. We refer to networks allowing a connection weight as *multirate* networks.

We let W denote the set of weights allowed in a multirate environment. For any set W, we let b denote the smallest allowable weight (or the greatest lower bound on the set of allowed weights, if there is no smallest weight). We let B denote the largest weight (or smallest upper bound). In addition we define β to be the *maximum port weight.* By definition, $0 < b \leq B \leq \beta \leq 1$. A common choice for W is the interval $[b, B]$. The quantity $\sigma_W(u)$ will come up often in the multirate results. It is defined as the smallest $v > u$ such that v can be obtained by summing values in W. When it is clear from context we will drop the W subscript.

With this extension, a connection assignment is a set of requests for which, for every input or output x, the sum of the weights of connection requests including x is at most β. A state is now a set of routes that obey the maximum port weight constraint and for which for every link l, the sum of the weights of all routes including l is at most 1. The *weight on a link* l in a given state is the sum of the weights of all routes including l. A link or switch y is said to be *w-accessible* in a given state from an input x if there is a path from x to y, such that the weight on each link in the path is at most $1 - w$. The other definitions are extended in the obvious way.

It seems important to say a little more about β. The motivation for introducing a maximum port weight comes from the realistic situation in which the links in a network are operated at a faster speed than the external links connecting networks. We refer to the quantity $1/\beta$ as the *speed advantage* of the network. This is simply the ratio of the internal to the external link speeds.

3. Related Results

The classic measure of network complexity is the asymptotic crosspoint count. There are other measures of interest, for example, the number of pin-limited chips required to package the network [13]. In this work we restrict the comparison to crosspoint count. Some of the

results cited below fall into the theoretical rather than the practical category, for reasons ranging from impractically large constants hidden by the asymptotic complexity measure to routing algorithms that must solve NP-complete problems. We distinguish these results from practical constructions.

In designing multirate distribution networks, we often start with circuit switching distribution networks. A lower bound of $\Omega(N \log N)$ was shown by Pippenger and Valiant [9] for a weaker type of network known as a rearrangeable N-shifter. Ofman [7] constructed a rearrangeably nonblocking distributor which was improved by Thompson [11] to have complexity matching that lower bound. Later Feldman, Friedman and Pippenger [4] demonstrated the existence of a wide-sense nonblocking distributor matching the lower bound, but with no known efficient routing algorithm. Another wide-sense nonblocking distributor can be obtained by using Pippenger's tree-like network [8] with expanders for the concentrators required in the construction. This yields an $O(N(\log N)^2)$ construction. Unfortunately the best explicit expanders involve large constants hidden by the big-O notation. The best practical wide-sense nonblocking distributors are obtained by using the Cantor network [1] for the concentrators in Pippenger's network. The resulting network has complexity $O(N(\log N)^3)$. The best explicit construction of a nearly nonblocking distributor is due to Turner [12] and has complexity $O(N(\log N)^2)$.

All of the results cited so far allow the depth of the network to vary with N. There is also significant interest in minimizing complexity for a fixed depth. It is within this area that Yang and Masson's results and our extensions compare most favorably. Pippenger and Yao [10] gave a lower bound of $\Omega(N^{1+1/k})$ for the weaker rearrangeable N-shifters with depth k. The best theoretical wide-sense nonblocking distributor with depth k is given by Feldman, Friedman and Pippenger [4] and has complexity $O(N^{1+1/k}(\log N)^{1-1/k})$. They also give explicit constructions of wide-sense nonblocking distributors with depth 2 and complexity $O(N^{5/3})$ and depth 3 and complexity $O(N^{11/7})$. Dolev et al. [3] have given an explicit construction of a wide-sense nonblocking distributor with depth $3k-2$ and complexity $O(N^{1+1/k})$ using expanders. Yang and Masson [14] give a construction of a nearly nonblocking distributor with depth $2k+1$ and complexity

$$O(N^{1+1/(k+1)}(\log N/\log\log N)^{(k+2)/2-1/(k+1)}).$$

The three stage version ($k=1$) has complexity $O(N^{3/2}(\log N/\log\log N))$.

We now turn to multirate distribution networks. These results all come from reference [6]. Each result is stated as a condition relating the multirate parameters (W and β) and the parameters of the network.

The Beneš network is a special case of the Clos network in which the switches are all $d \times d$ crossbars. Pippenger's tree-like network with Beneš networks as the concentrators is a wide-sense nonblocking distributor if

$$\frac{\sigma_W(1-B)}{\beta+B} \geq \frac{2}{d}(1+(d-1)\log_d(N/d)),$$

where N is the number of network inputs and d is the dimension of the switches in the Beneš network. This network has $O(N(\log N)^2)$ crosspoint complexity; the necessary speed advantage grows as roughly $2\log N$.

To interpret a multirate result, we have both a crosspoint count and a necessary speed advantage to consider. One reasonable interpretation is to take the product of the crosspoint count and the speed advantage. This makes sense because a speed advantage of $1/\beta$ can be obtained by replicating the network $1/\beta$ times. An incoming packet is converted from a serial line into $1/\beta$ parallel data lines for transmission through the network. Conversion back into a serial line is done when the packet leaves the network. This has the effect of operating the network at a faster speed than the external links.

With this interpretation, Pippenger's network can be used for multirate connections with complexity of approximately twice the complexity of the same network used for circuit switching connections.

Based on Turner's nearly nonblocking distributor composed of back-to-back Cantor networks [12], a multirate nearly nonblocking distributor can be constructed of back-to-back Beneš networks under the condition

$$\frac{\sigma_W(1-B)}{\beta} \geq \frac{2}{d}(1+(d-1)\log_d(N/d)).$$

This network has $O(N \log N)$ crosspoint complexity, and a necessary speed advantage that grows as roughly $2 \log N$.

Both of these results allow the number of stages to vary with N. Other than the work presented in this paper, we are not aware of any multirate results that are explicitly concerned with minimizing complexity for a fixed depth.

In the remainder of this paper we present a multirate distribution network based on the construction of Yang and Masson's circuit switching distribution network. Specifically, in Section 4 we review the work of Yang and Masson from reference [14] and then derive conditions under which the three stage Clos network is a nearly nonblocking distributor in the multirate environment. In Section 5 we generalize to a $2k+1$ stage Clos network. In Section 6 we suggest some extensions to this work.

4. Three Stage Networks

4.1. Review of Yang and Masson

In reference [14], Yang and Masson use the Clos network as a nearly nonblocking distributor. In this section we review their result for the three stage Clos network. We refer to the switches containing the inputs as *input stage* switches. We refer to the switches containing the outputs as *output stage* switches. We refer to the other switches as *middle stage* switches.

In the three stage case, a distribution route can be specified by giving the set of middle stage switches in the route and, for each middle stage switch, the set of output stage switches it is connected to by the distribution route. Yang and Masson use a greedy algorithm to choose the middle stage switches used to realize a distribution request. They show that with "enough" middle stage switches in the network, any compatible distribution request can be realized with at most x middle stage switches. This has the practical effect of limiting the

fanout of a connection to x in the input stage. No limit is placed on the fanout in the middle and output stages.

The greedy algorithm realizes a distribution request (u, Y) by building a set R of middle stage switches that are used to construct the route. Without loss of generality we can assume that Y is the set of output *switches* corresponding to the outputs of the distribution request, since multiple outputs on the same switch can be reached by branching in the output switch. The set R is initially empty. Until the switches in R can reach all output switches in Y, the following step is performed: add to R a middle stage switch accessible from u that can reach the most required output switches that are not already reached by some middle switch in R.

The main result of Yang and Masson's work is captured in the following theorem that expresses how many middle stage switches are needed in the network for this greedy routing approach to succeed.

THEOREM 4.1. *The three stage Clos network is a nearly nonblocking distributor if*

$$m > \min_{1 \le x \le \min\{n-1, r\}} \{(n-1)(x + r^{1/x})\}.$$

Proof Sketch: We prove the theorem by first considering how many middle stage switches are needed by the greedy algorithm to reach all of the outputs in a compatible distribution request, and then insuring that any idle input can reach that many middle stage switches.

The problem of choosing the middle stage switches to use in the route is simply a set covering problem. An instance of the set covering problem is a set $S = \{s_1, .., s_t\}$ and a family $F = \{f_1, .., f_p\}$ of subsets of S. A solution is a set $C \subseteq F$ such that $\cup_{f_i \in C} f_i = S$. The objective is to find a set C of minimum size. The correspondence between the routing problem and a set covering problem is straightforward. The set of output switches form the set S. For each middle stage switch j we have a set $f_j \in F$, containing the output switches that can be reached from switch j in the current state of the network. That is, the links from switch j to these output switches are currently idle.

The following lemma concerns the performance of the greedy algorithm on a restricted version of the set covering problem.

LEMMA 4.1. *Assume that every element in S appears in at least $p-q$ sets of F $(0 \le q \le p)$. If $p > q t^{1/x}$ the greedy algorithm will find a set $C \subseteq F$ such that $\cup_{f_i \in C} f_i = S$ and C has size at most x.*

Proof Sketch:

This can be shown in a straightforward manner based on the behavior of the greedy algorithm. The reader is referred to reference [14] for further details. The proof given there is for the specific problem of routing in the three stage Clos network, but it can easily be generalized to prove this result.

■

COROLLARY 4.1. *Given $(n-1)r^{1/x}$ middle switches, the greedy algorithm can find at most x of them through which to reach all of the outputs in a compatible distribution request (u, Y).*

Proof:

Without loss of generality we will assume that Y is the set of output *switches* corresponding to the idle outputs of the distribution request. The key to this proof is the fact that any last stage switch in Y has at most $n-1$ busy outputs, and therefore at most $n-1$ busy incoming links. Each busy incoming link corresponds to a middle stage switch that is blocked from reaching this last stage switch. Since there are at most $n-1$ busy links, there are at most $n-1$ middle stage switches that are blocked from reaching a particular switch in Y.

This fact gives us $q = n-1$ in Lemma 4.1. With the correspondence described earlier we have $p = m$ and $t = r$. The result follows immediately.

■

Corollary 4.1 tells us how many middle stage switches the greedy algorithm must have to choose from in order to find x through which to reach all the outputs in a request. Of course, the input involved in the request must be able to reach all of the middle switches considered by the greedy algorithm.

An input u involved in a compatible distribution request may be blocked from at most $(n-1)x$ middle switches, since in the worst case each of the $n-1$ inputs sharing the first stage switch with u is involved in a connection branching to x middle switches. Thus, if there are more that $(n-1)(x + r^{1/x})$ middle switches, then u will be able to reach at least $(n-1)r^{1/x}$, and by the corollary the greedy algorithm will find at most x among these through which the distribution request can be realized.

■

For a particular choice of x it is straightforward to calculate the necessary number of middle stage switches. As is expressed in Theorem 4.1, the best choice of x is the one which minimizes the number of middle stage switches.

By choosing $x = \log r/(\delta \log\log r)$ for any constant δ, $0 < \delta < 1$, it is possible to get an asymptotic bound on m of $O(n \log r/\log\log r)$. Since the crosspoint count in the three stage Clos network is $2rnm + mr^2$, this choice of x and $n = r = \sqrt{N}$ yields an asymptotic crosspoint complexity of $O(N^{3/2}(\log N/\log\log N))$ for the three stage network. This is the best known explicit construction of a three stage nearly nonblocking distributor. The previous best such networks came from wide-sense nonblocking distributors, and thus were capable of realizing augmentation requests in addition to distribution requests.

To put the complexity of this construction in context, we have plotted the crosspoint count for three networks, normalized by N, as a function of N. (See Figure 3.) The top curve is for an $N \times N$ crossbar. The middle curve is for the Yang and Masson three stage construction. The bottom curve is for the three stage Clos network operated as a strictly nonblocking connector. The three stage Clos network operated as a nearly nonblocking distribution network has complexity between the other two networks.

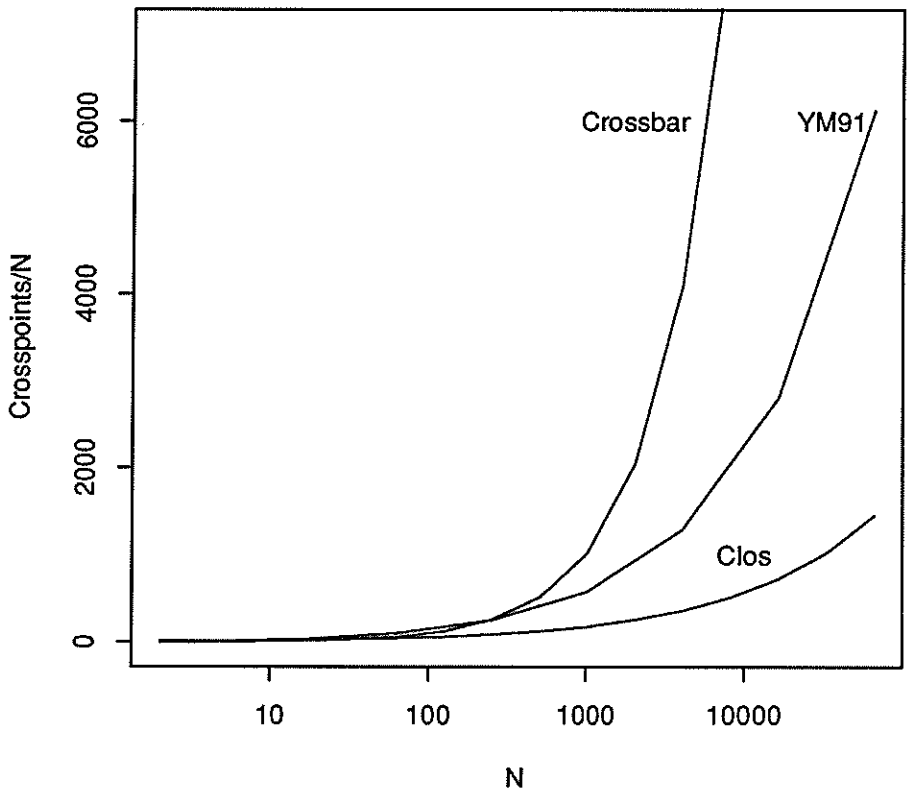

Figure 3: Comparison of Crosspoint Complexity

4.2. Multirate Clos Network

In this section we extend the results of the previous section to the multirate case. Consider a compatible distribution request (u, Y, w), where u is an input, Y is a set of outputs and $w \in W$ is an allowable connection weight. Since this connection is compatible, the weight on u is at most $\beta - w$. The same is true for each output in Y.

The results in this section are obtained by a similar approach to that used by Yang and Masson. Specifically, we first determine how many middle stage switches are needed by the greedy algorithm. We then insure that the input involved in a compatible request can reach that many middle stage switches. The key to the result is the following generalization of Corollary 4.1.

LEMMA 4.2. *Given $L(w)r^{1/x}$ middle switches, the greedy algorithm can find at most x of them through which to reach all the outputs in a compatible distribution request, (u, Y, w), where $L(w)$ denotes the quantity $\lfloor(\beta n - w)/\sigma_W(1 - w)\rfloor$.*

Proof:

Consider an output $y \in Y$, and let s denote the output switch containing y. There are $n - 1$ other outputs on switch s, each with weight at most β. We know that y has weight at most $\beta - w$, thus the total weight out of switch s is at most $\beta n - w$. Clearly, the total weight into s is also at most $\beta n - w$. (There may be branching in s, but that will only make the weight out of s larger than the weight into s.)

We need to determine the maximum number of links into s that may be blocked to a connection of weight w. In the classical model, a link may only be used by one route. If it is already being used, then it is blocked to all other requests. In the multirate environment a link may be shared by multiple routes provided the total weight on the link does not exceed 1. A link is *blocked to a request of weight w* if and only if the total weight on the link is at least $\sigma(1 - w)$, where the function $\sigma(u)$ is the smallest $v > u$ such that v can be obtained by summing allowable connection weights. When the request weight is clear from context we will simply say that a link is blocked without adding the phrase "to a request of weight w".

With this in mind, we can determine an upper bound on the number of links into s that are blocked to a connection of weight w. A weight of at least $\sigma(1 - w)$ is needed for such a blocking link. Since the total weight into s is at most $\beta n - w$, there are at most $(\beta n - w)/\sigma(1 - w)$ links blocked into s. We can tighten this to $L(w)$ since we know that the number of links must be integral.

We can now apply Lemma 4.1 with $q = L(w)$, $p = m$ and $t = r$. Thus, $L(w)r^{1/x}$ middle stage switches are sufficient for the greedy algorithm to find at most x through which to realize a distribution request.

■

With this generalization, Corollary 4.1 corresponds to the special case in which $\beta = 1$ and $W = \{1\}$, implying that $L(w) = n - 1$. We use Lemma 4.2 to prove the main result.

THEOREM 4.2. *The three stage Clos network is a nearly nonblocking multirate distributor if*

$$\begin{aligned} m &> \min_{1 \le x \le \min\{n-1, r\}} \max_{w \in W} \{L(w)(x + r^{1/x})\} \\ &= \max_{w \in W} L(w) \min_{x} \{x + r^{1/x}\} \end{aligned}$$

Proof:

As in the circuit switching case, Lemma 4.2 indicates how many middle stage switches must be available to the greedy algorithm so that at most x are found to reach all the outputs in a compatible distribution request with weight w. The input u involved in the distribution request must be able to reach these middle stage switches. We must determine how many middle stage switches may be blocked from u relative to a connection of weight w.

Obviously, u has weight at most $\beta - w$. There are $n - 1$ other inputs sharing the same input switch as u, each with weight at most β. Thus the total weight into the input switch is at most $\beta n - w$. Since a connection must have weight at least $\sigma(1 - w)$ in order to block a connection of weight w, there are at most $L(w)$ connections able to block u from middle stage switches. Each of these connections may branch by x in the first stage, thus there are at most $xL(w)$ middle stage switches blocked to u for a connection of weight w.

Combining this with the result of the lemma, we see that if there are at least $L(w)(x + r^{1/x})$ middle stage switches, then u will be able to reach at least $L(w)r^{1/x}$ of them, and the greedy algorithm will find at most x among these through which to realize a compatible connection request of weight w.

■

The specific value of $\max_{w \in W} L(w)$ will depend on the set W. One special case of interest is the *unrestricted packet switching* case in which $b = 0$ and $B = \beta$, with every value in the range $(0, B]$ allowed as a weight. In this case $\max_{w \in W} L(w)$ occurs at $w = \beta$. (Unrestricted packet switching is the worst case collection of allowable weights in the sense that increasing b or restricting W to a discrete set of weights while keeping B and β fixed will only decrease $\max_{w \in W} L(w)$.) The sufficient condition on the number of middle stage switches for unrestricted packet switching is

$$m > \left\lfloor \frac{\beta(n-1)}{1-\beta} \right\rfloor \min_{x} \{x + r^{1/x}\},$$

which is implied by

$$m > \frac{\beta(n-1)}{1-\beta} \min_{x} \{x + r^{1/x}\}.$$

If $(1/\beta) = 2$, this expression reduces to the same condition on m derived in Theorem 4.1 for the circuit switching case. Thus with a speed advantage of two, we can handle multirate traffic with the same crosspoint complexity as for circuit switching. To put this in perspective, other multirate results indicate that a factor of two increase in complexity is about the best one can hope for in adapting a circuit switching network to accomodate multirate traffic.

4.3. Multirate Beneš Network

The three stage Beneš network is just a special case of the Clos network in which the switches are all $n \times n$, with $n = \sqrt{N}$. This makes it easy to use the results from the previous section to get the Beneš network result.

COROLLARY 4.2. *The three stage Beneš network is a nearly nonblocking multirate distributor if*

$$\begin{aligned} \sqrt{N} &> \min_{1 \le x \le \sqrt{N}-1} \max_{w \in W} \{L(w)(x + N^{1/(2x)})\} \\ &= \max_{w \in W} L(w) \min_{x} \{x + N^{1/(2x)}\}. \end{aligned}$$

Proof:

Let $m = n = r = \sqrt{N}$ in Theorem 4.2.

■

As for the multirate Clos network, we can consider the special case of unrestricted packet switching. With the floor function removed from $max_w L(w)$, this gives

$$\sqrt{N} > \frac{\beta(\sqrt{N}-1)}{1-\beta} \min_{x} \{x + N^{1/(2x)}\}.$$

For a particular value of x we can rearrange the equation above to get the following condition on the speed advantage.

$$(1/\beta) > 1 + x + N^{1/(2x)}$$

For example, if $N = 256$, then with $x = 3$ a speed advantage of about 6.5 is sufficient. It is straightforward to prove that a three stage Beneš network requires a speed advantage of 3 to operate as a strictly nonblocking connector and a speed advantage of $\sqrt{N}$ to operate as a strictly nonblocking distributor. Of course this is something of an apples-to-oranges comparison, but it serves to give some context to this result. An improvement in the speed advantage has been made by simply restricting the fanout in the first stage and using a greedy algorithm to select the middle stage switches.

5. Arbitrary Stage Networks

In this section we generalize the results of the previous section to networks with $2k + 1$ stages. We begin by reviewing the generalization given by Yang and Masson for the circuit switching environment. We then extend our multirate results.

5.1. Review of Yang and Masson

The approach used by Yang and Masson can be extended to an arbitrary number of stages in a straightforward way. In a $2k+1$ stage Clos network, the greedy algorithm is used to choose the middle $2k-1$ stage networks through which to realize a distribution request. This creates a distribution request within each of the chosen $2k-1$ stage networks. This approach is applied recursively within each of these networks. Eventually we reach the innermost three stage Clos networks where the greedy algorithm chooses the middle stage switches through which to route the request. In effect we end up with a fanout restriction in each of the first k stages of the network.

Let $G_{2k+1}(N)$ denote the crosspoint complexity of a $2k+1$ stage nearly nonblocking distributor constructed using Yang and Masson's approach of restricted fanout and a greedy algorithm to choose the middle stage networks and switches. Yang and Masson generalize their three stage Clos result to give the following asymptotic crosspoint complexity for a $2k+1$ stage nearly nonblocking distributor:

$$G_{2k+1}(N) = O(N^{1+1/(k+1)}(\log N/\log\log N)^{(k+2)/2-1/(k+1)})$$

This result can be proven by induction on k, with the three stage Clos network ($k=1$) serving as the base case. Assuming the result holds for a $2k-1$ stage network, it can be shown to hold for a $2k+1$ stage network. The key to the proof is the judicious choice of r_k. The reader is referred to [14] for further details.

5.2. Multirate Clos Network

We now present the main result of this paper, an extension of the previous section to the multirate environment. This is fairly direct, although it is slightly more complicated than the three stage case.

We use the same routing strategy described earlier, using the greedy algorithm recursively within successively smaller networks. The analysis is complicated slightly by the fact that we require a speed advantage at each level of the recursion. To achieve this, we restrict the link weights to be less than one until we reach the final level of recursion.

More specifically, consider Figure 4 which shows the relevant parameters for a $2i+1$ stage network within a larger network. We let β_{i+1} denote the maximum weight of links into and out of the $2k+1$ stage network. We let β_i denote the maximum weight of links into the middle networks. Using an analysis similar to the three stage Clos analysis, we can give a sufficient condition on m_i so the greedy algorithm can be used to choose at most x_i middle stage networks through which to route a distribution request. The analysis is identical to that seen previously, except that the link weight needed to block a connection of weight w from a middle stage network is now $\beta_i - w$, rather than $1-w$. With this generalization, we have the following condition on m_i:

$$m_i > \left\lfloor \frac{\beta_{i+1} n_i - w}{\sigma_W(\beta_i - w)} \right\rfloor (x_i + r_i^{1/x_i})$$

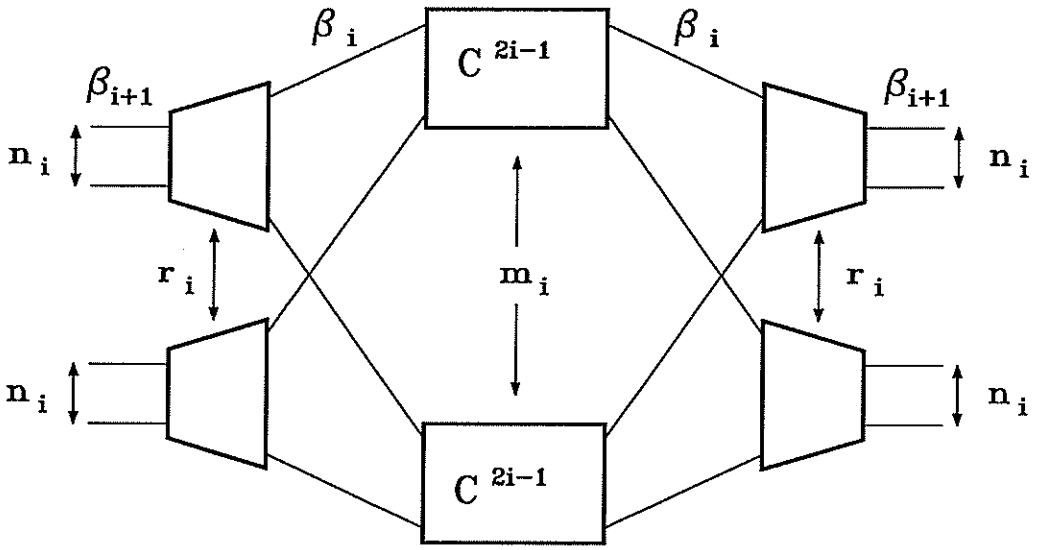

Figure 4: $2i+1$ Stage Clos Network

If we define $L_i(w) = \lfloor(\beta_{i+1}n_i - w)/\sigma_W(\beta_i - w)\rfloor$, and consider all possible values of x_i and w, we have

$$\begin{aligned} m_i &> \min_{1\le x_i\le \min\{n_i-1, r_i\}} \max_{w\in W}\{L_i(w)(x_i + r_i^{1/x_i})\} \\ &= \max_{w\in W} L_i(w) \min_{x_i}\{x_i + r_i^{1/x_i}\}. \end{aligned}$$

This condition ensures that there are sufficient middle stage networks at level i of the recursion so the greedy algorithm will find at most x_i to use in the distribution request. In addition, we are guaranteed that the maximum port weight at the $2i-1$ stage networks is β_i. The $2k+1$ stage network is a nearly nonblocking distributor if the condition on m_i holds for all i between 1 and k. The condition on m given in Theorem 4.2 for the three stage network corresponds to $i = 1$, with $\beta_1 = 1$ and $\beta_2 = \beta$.

For the unrestricted packet switching case, the $2k+1$ stage Clos network is a nearly nonblocking distributor if

$$m_i > \frac{\beta_{i+1}}{\beta_i - \beta} n_i(x_i + r_i^{1/x_i}), \qquad 1 \le i \le k, \qquad \beta_{k+1} \equiv \beta.$$

(This is obtained by taking $w = \beta$ in the expression above and discarding the floor function.) This condition on the values of the m_i matches the circuit switching condition, except for the leading $\beta_{i+1}/(\beta_i - \beta)$ term. We can get the same crosspoint complexity as in the circuit switching case by choosing the β_i so that each leading term is exactly one. In addition, for choices of the β_i which achieve this, we want the choice which maximizes β. This is accomplished by choosing $\beta_i = (k+2-i)/(k+1)$. Clearly this forces each term to be equal to one and $\beta = 1/(k+1)$.

We then have the following result.

THEOREM 5.1. *With the appropriate choices of switch dimensions (as in [14]) the $2k+1$ stage Clos network operated with a speed advantage of $1/\beta = k+1$ is a nearly nonblocking multirate distributor with crosspoint complexity of*

$$O(N^{1+1/(k+1)}(\log N/\log\log N)^{(k+2)/2-1/(k+1)}).$$

Considering some specific values for k, we see that this construction gives a nearly nonblocking multirate distributor of three stages ($k = 1$) with crosspoint complexity of $O(N^{3/2}(\log N/\log\log N))$ and speed advantage of $1/\beta = 2$. The five stage version has crosspoint complexity of $O(N^{4/3}(\log N/\log\log N)^{5/3})$ and speed advantage of 3. We can compare this to Turner's back-to-back Beneš network by considering the crosspoint complexity and speed advantage of that network for a fixed number of stages. With a fixed number of stages we get crosspoint complexity of $O(N^{1+1/(k+1)})$ and speed advantage of $1/\beta \approx k$ for a back-to-back Beneš network with $4k+1$ stages. This network requires five stages ($k = 1$) to get crosspoint complexity of $O(N^{3/2})$ with a speed advantage of 1. In general the back-to-back Beneš network requires about twice as many stages to get complexity that is comparable to the Clos network for nearly nonblocking multirate distribution.

6. Conclusions

By a straightforward extension to the work done by Yang and Masson, we have been able to construct Clos networks that are nearly nonblocking distributors in the multirate environment. The complexity of these networks compares most favorably when the number of stages is fixed and fairly small.

Several unsuccessful attempts were made to construct the network so that the speed advantage was constant, rather than increasing with the number of stages. It appears that this may require a considerable change in the routing approach.

One extension to this work would be to consider dividing connections into classes (based on size or weight, for example) where each class would have a different fanout restriction. It seems intuitively reasonable to treat large and small connections differently, allowing large connections to fanout more in the early stages than small connections. It may be possible to get better complexity by making such a change to the routing algorithm.

Acknowledgements

The author gratefully acknowledges comments on this work by Andy Fingerhut and Jon Turner.

References

[1] D. B. Cantor. On non-blocking switching networks. *Networks*, 1971.

[2] C. Clos. A study of non-blocking switching networks. *Bell Systems Tech. Journal*, 1953.

[3] D. Dolev, C. Dwork, N. Pippenger, and A. Wigderson. Superconcentrators, generalizers and generalized connectors with limited depth. In *Proc. 15th ACM Symp. on Theory of Computing*, 1983.

[4] P. Feldman, J. Friedman, and N. Pippenger. Wide-sense nonblocking networks. *SIAM J. Discrete Math.*, 1988.

[5] Riccardo Melen and Jonathan S. Turner. Nonblocking multirate networks. *SIAM Journal on Computing*, 1989.

[6] Riccardo Melen and Jonathan S. Turner. Nonblocking multirate distribution networks. In *Proc. of Infocom 90*, 1990.

[7] Y. Ofman. A universal automaton. *Trans. Moscow Math. Soc.*, 1965.

[8] N. Pippenger. *The complexity theory of switching networks.* PhD thesis, Massachusetts Institute of Technology, 1973.

[9] N. Pippenger and L.G. Valiant. Shifting graphs and their applications. *J. ACM*, 1976.

[10] N. Pippenger and A.C.-C. Yao. Rearrangeable networks with limited depth. *SIAM J. Algebraic Discrete Methods*, 1982.

[11] C.D. Thompson. Generalized connection networks for parallel processor interconnection. *IEEE Transactions on Computers*, 1978.

[12] Jonathan S. Turner. Practical wide-sense nonblocking generalized connectors. Technical report, Washington University Computer Science Department, WUCS-88-29.

[13] Ellen E. Witte. A quantitative comparison of architectures for ATM switching systems. Technical report, Washington University Computer Science Department, WUCS-91-47.

[14] Yuanyuan Yang and Gerald M. Masson. Nonblocking broadcast switching networks. *IEEE Transactions on Computers*, 1991.