NONBLOCKING MULTIRATE DISTRIBUTION NETWORKS

RICCARDO MELEN MEMBER, IEEE, JONATHAN S. TURNER FELLOW, IEEE

June 5, 1997

Abstract — **This paper generalizes known results for nonblocking distribution networks (also known as generalized connection networks) to the multirate environment, where different user connections share a switch's internal data paths in arbitrary fractions of the total capacity. In particular, we derive conditions under which networks due to Ofman and Thompson, Pippenger, and Turner lead to multirate distribution networks. Our results include both rearrangeable and wide-sense nonblocking networks. The complexity of the rearrangeable multirate networks exceeds that of the corresponding space division network by a** $\log\log$ **factor while the complexity of the wide sense nonblocking networks is within a factor of two of the corresponding space division networks.**

Index Terms — **nonblocking networks, distribution networks, generalized connection networks, multipoint networks, multicast communication, ATM networks**

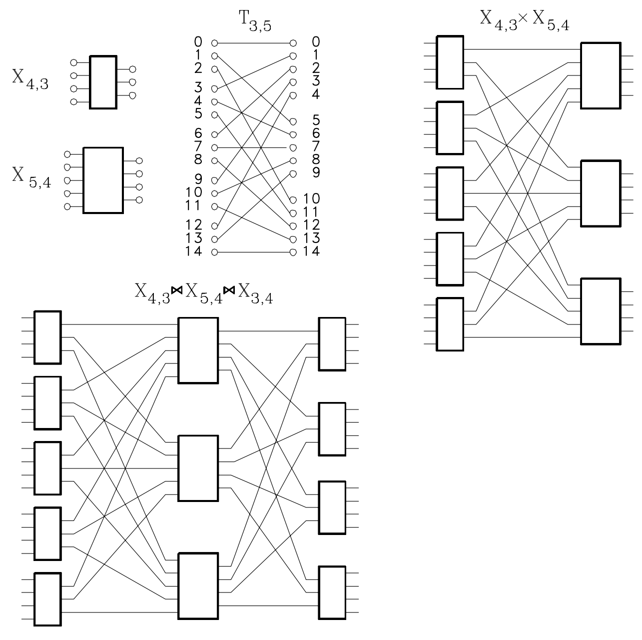

Figure 1: Network Construction Operators

I. Introduction

In reference [4, 5], the authors introduce the concept of nonblocking multirate networks and prove a collection of results generalizing the classical theory of nonblocking connection networks. In this paper, we extend that work to cover distribution networks, that is networks that are capable of distributing a signal from a single input to one or more outputs. Such networks are also known as generalized connection networks.

II. Definitions

The topological design of switching networks determines their complexity and blocking characteristics. We define a graph model for switching networks and introduce operations by which complex networks can be constructed from simpler components.

We denote a network N by a quadruple (S, L, I, O), where S is a set of vertices, called *switches*, L is a set of arcs called *links*, I is a set of *input terminals* and O is a set of *output terminals*. Each link is an ordered pair (x, y) where $x \in I \cup S$ and $y \in O \cup S$. We require that each input and output terminal appear in exactly one link. Links that include an input terminal are called *inputs*. Those including output terminals are called *outputs*. The remainder are called *internal*. A network with n inputs and m outputs is referred to as an (n, m)-network. We'll generally use n to denote the number of network inputs and m to denote the number of outputs. Inputs and outputs are numbered consecutively from 0 and are identified with their indexes. We let $X_{n,m}$ denote the (n, m)-network comprising a single switch connected to all n inputs and all m outputs. Such a network is called a *crossbar* and is the basic building block from which other networks are constructed.

Note that in our model, the vertices are associated with the network's switching components and the arcs with the data paths. Another common graph model for networks identifies a graph's vertices with the data paths and its edges with crosspoints.

If i is a positive integer and N is an (n, m)-network, then $i \cdot N$ denotes the network obtained by taking i copies of N, without interconnecting them. Inputs and outputs

[0]Riccardo Melen is with Centro Stude E Laboratori Telecomunicazioni (CSELT), Torino, Italy and his work has been supported in part by Associazione Elettrotecnica ed Elettronica Italiana, Milano, Italy. This work was done while on leave at Washington University.

Jonathan Turner's work is supported by the National Science Foundation (grant DCI 8600947), Bell Communications Research, Bell Northern Research, Italtel SIT and NEC.

to N are numbered in the obvious way, with the first copy receiving inputs $0, \ldots, n-1$ and outputs $0, \ldots, m-1$ and so forth. The *reverse* of a network N is denoted $\overline{N}$ and is obtained by exchanging input and output terminals and reversing the directions of all links.

The *concatenation* of two networks $N_1 = (S_1, L_1, I_1, O_1)$ and $N_2 = (S_2, L_2, I_2, O_2)$ with $n = |O_1| = |I_2|$ is denoted $N_1; N_2$ and is obtained by identifying output i of N_1 with input i of N_2. More precisely, if we let $N = (S, L, I, O)$ be $N_1; N_2$ then $S = S_1 \cup S_2$, $I = I_1$, $O = O_2$ and

$$\begin{aligned} L \;=\; & \{(x,y) | (x,y) \in L_1, y \in S_1\} \\ & \cup \{(x,y) | (x,y) \in L_2, x \in S_2\} \\ & \cup \{(x,y) | \exists i \in [0, n-1] \\ & \qquad \text{such that } (x,i) \in L_1 \text{ and } (i,y) \in L_2\} \end{aligned}$$

If σ is a permutation on $\{0 \ldots, n-1\}$, we let σ also denote the network (S, L, I, O) where $I = \{0, \ldots, n-1\}$, $O = \{0, \ldots, n-1\}$, $S = \emptyset$ and $L = \{(i, \sigma(i)) | 0 \leq i \leq n-1\}$.

If d_1 and d_2 are positive integers, we define τ_{d_1,d_2} to be the permutation on $\{0, \ldots, d_1 d_2 - 1\}$ satisfying

$$\tau_{d_1,d_2}(jd_1 + i) = id_2 + j \quad 0 \leq i \leq d_1 - 1,\, 0 \leq j \leq d_2 - 1$$

Let N_1 be a network with n_1 outputs and N_2 be a network having n_2 inputs. The *product* of N_1 and N_2 is denoted $N_1 \times N_2$ and is defined as

$$(n_2 \cdot N_1); \tau_{n_1,n_2}; (n_1 \cdot N_2)$$

If N_1 has n_1 outputs, N_2 has n_2 inputs and n_3 outputs and N_3 has n_1 inputs, the *three-way product* $N_1 \bowtie N_2 \bowtie N_3$ is defined as

$$(n_2 \cdot N_1); \tau_{n_1,n_2}; (n_1 \cdot N_2); \tau_{n_3,n_1}; (n_3 \cdot N_3)$$

These definitions are illustrated in Figure 1

Several well-known networks can be conveniently defined using the network construction operators. The three stage Clos network $C^3_{n,d,q}$ is defined by $C^3_{n,d,q} = X_{d,q} \bowtie X_{n/d,n/d} \bowtie X_{q,d}$, where we require of course that d divides n.

The *delta network* [7] $D_{n,d}$ is defined by

$$D_{d,d} = X_{d,d} \qquad D_{n,d} = X_{d,d} \times D_{n/d,d}$$

where $n = d^k$ for some integer k. The number of stages in the delta network is exactly k. The banyan network [3] $Y_{n,d}$ is defined by

$$Y_{d,d} = X_{d,d} \qquad Y_{n,d} = \tau_{n/d,d}; (n/d \cdot X_d); \tau_{d,n/d}; (d \cdot Y_{n/d,d})$$

The banyan network is isomorphic to the delta network, and so is equivalent in all respects. However, it is useful to define it separately as certain properties are more easily proved using the banyan definition.

The delta networks can be extended by adding stages of switching. If d, k, h are positive integers with $h < k$ and $n = d^k$, we define the extended delta network $D^*_{n,d,h}$ as follows

$$D^*_{n,d,0} = D_{n,d} \quad D^*_{n,d,h} = D_{d^h,d} \bowtie D_{d^{k-h},d} \bowtie \overline{D}_{d^h,d}$$

An equivalent definition is

$$D^*_{n,d,0} = D_{n,d} \quad D^*_{n,d,h} = X_{d,d} \bowtie D^*_{n/d,d,h-1} \bowtie X_{d,d}$$

If we take $h = k-1$ we obtain the Beneš network, denoted $B_{n,d}$. By placing several Beneš networks in parallel with one another we obtain the Cantor network $K_{n,d,q}$ defined by

$$K_{n,d,q} = X_{1,q} \bowtie B_{n,d} \bowtie X_{q,1}$$

These networks are illustrated in Figure 2

We define three parameters that constrain the traffic placed on a network; b is called the *minimum connection weight*, B the *maximum connection weight* and β the *maximum port weight*. By definition, $0 < b \leq B \leq \beta \leq 1$.

We discuss four different classes of networks, *connection networks* (or simply connectors) which provide one-to-one communication between specified inputs and outputs, *concentration networks* (concentrators), which provide one-to-one communication between specified inputs and unspecified outputs, *distribution networks* (distributors), which provide one-to-many communication between specified inputs and specified *sets* of outputs and finally *replication networks* (replicators) which provide one-to-many communication between specified inputs and unspecified sets of outputs. Our primary interest here is in distribution networks (also known as generalized connectors). We discuss the other network types primarily for their use in constructing distribution networks.

A *connection request* is a triple (x, y, ω) where x is an input y is an output and $\omega \in [b, B]$ is the *weight* of the request and represents the fraction of the capacity of the network's internal data paths required by the request. A *connection assignment* is a set of requests for which, for every input or output x, the sum of the weights of the connection requests including x is at most β.

A *connection route* is a list of links forming a path from an input to an output together with a weight. A route *realizes* a request (x, y, ω) if it starts at x, ends at y and has weight ω. A *state* is a set of routes for which, for every input or output x, the sum of the weights of the routes including x is at most β and for every link ℓ, the sum of the weights of all routes including ℓ is at most 1. We say that a state realizes a given assignment if it contains one route realizing each request in the assignment and no others. 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 *ω-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 $1 - \omega$. We say that a state s_1 is *below* a state s_2 if $s_1 \subseteq s_2$. Similarly, we say that s_2 is *above* s_1. We say a connection request (x, y, ω) is compatible with a state s if the weight on x and y in s is at most $\beta - \omega$.

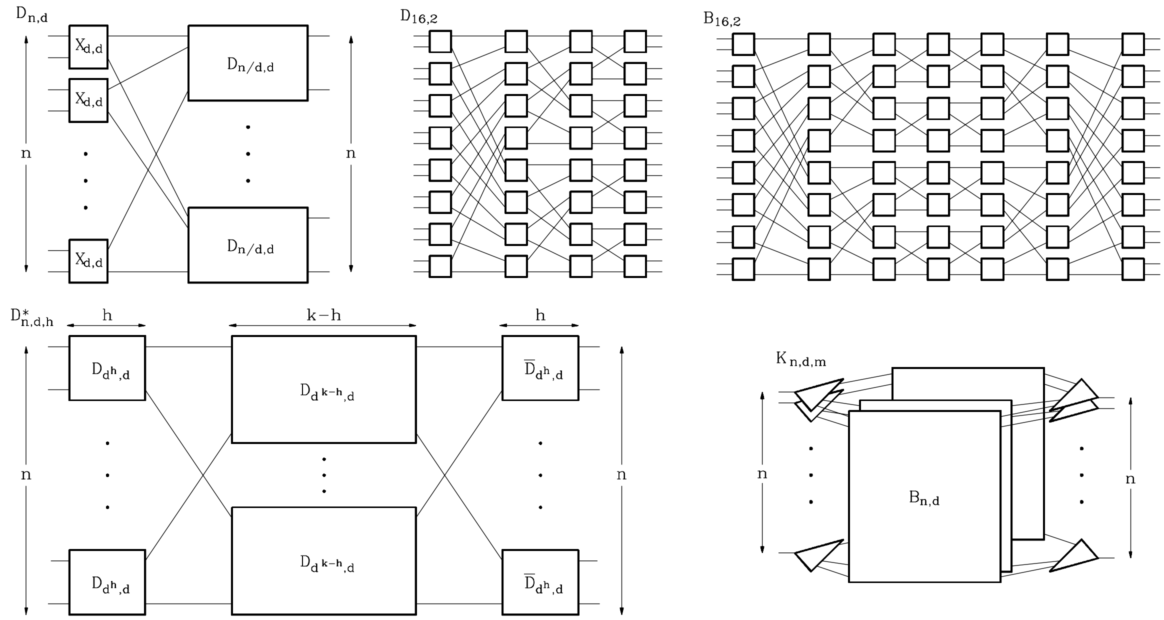

Figure 2: Network Definitions

A network is a *rearrangeable connector* if for every connection assignment, there is a state realizing that assignment. A network is a *strictly nonblocking connector* if for every state s and connection request r compatible with s, there exists a route realizing r that is compatible with s. 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. Intuitively, a network is wide-sense nonblocking if blocking can be avoided by judicious selection of routes. Note that every strictly nonblocking connector is also wide-sense nonblocking and every wide-sense nonblocking connector is also rearrangeable.

Concentrators support one-to-one communication between specified inputs and unspecified outputs. A *concentration request* is a pair (x, ω) where x is an input and $\omega \in [b, B]$ is the *weight*. A *concentration assignment* is a set of requests with total weight at most βm (where m is the number of network outputs) and for which, for every input or output x, the sum of the weights of the connection requests including x is at most β.

A *concentration route* is a path from an input to an output together with a weight. A route *realizes* a request (x, ω) if it starts at x and has weight ω. Network states are defined as previously. We say a concentration request (x, ω) is compatible with a state s if the weight on x in s is at most $\beta - \omega$ and if the total weight in s is at most $\beta m - \omega$.

A network is a *rearrangeable concentrator* if for every concentration assignment, there exists a state realizing that assignment. A network is a *strictly nonblocking concentrator* if for every state s and concentration request r compatible with s, there exists a route realizing r that is compatible with s. A network is a *wide-sense nonblocking concentrator* if the state space has a safe subset S such that for every state $s \in S$ all states below s are in S and for every concentration 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.

Replicators are networks that support one-to-many communication between specified inputs and unspecified sets of outputs. A *replication request* is a triple (x, f, ω) where x is an input, $\omega \in [b, B]$ is a weight and $f \in [1, m]$ (m is the number of network outputs) is called the *fanout* of the request. A *replication assignment* is a set of requests $R = \{(x_i, f_i, \omega_i)\}$ for which, for every input x, the sum of the weights of the connection requests including x is at most β and such that $\sum_i f_i\omega_i \leq \beta m$.

A *replication route* is a list of links forming a tree whose root is an input and whose leaves are outputs, together with a weight. A route *realizes* a request (x, f, ω) if it starts at x, has f leaves and weight ω. States are defined as previously, but with respect to replication routes. We say a replication request (x, f, ω) is compatible with a state s realizing an assignment $A = \{(x_i, f_i, \omega_i)\}$ if the weight on x in s is at most $\beta - \omega$ and if $f\omega + \sum_i f_i\omega_i \leq \beta m$. A network is a *rearrangeable replicator* for every replication assignment, there exists a state realizing that assignment.

Distributors support one-to-many communication from a specified input to one or more specified outputs. A *distribution request* is a triple (x, Y, ω) where x is an input, Y is a set of outputs and $\omega \in [b, B]$ is a *weight*. A *distribution assignment* is a set of requests for which, for every input or output x, the sum of the weights of the distribution requests including x is at most β.

A *distribution route* is a list of links forming a tree whose root is an input and whose leaves are outputs, to-

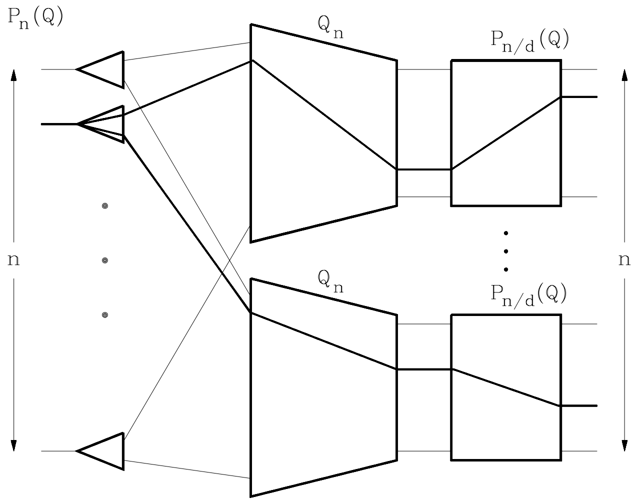

Figure 3: Pippenger's Network

gether with a weight. A route *realizes* a request (x, Y, ω) if its root is x, its leaves are exactly the set Y and it has weight ω. A state is defined as before, but with respect to distribution routes. We say a distribution request (x, Y, ω) is compatible with a state s if the weight in s on x and all $y \in Y$ is at most $\beta - \omega$.

An *augmentation request* for a distribution network in a state s is a pair (r, y) where $r = (x, Y, \omega)$ 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 the weight on y in s is at most $\beta - \omega$. We say that an augmentation request can be satisfied in s if the route realizing r can be extended by adding links so that y becomes a leaf of the route. This extension must not of course increase the weight on any link beyond 1.

A network is a *rearrangeably nonblocking distributor* if for every distribution assignment, there exists a state realizing that 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 states 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 such that $s \cup \{p\}$ is in S; and every augmentation request r compatible with s can be satisifed in such a way that the resulting state is in S.

III. Pippenger's Network

Let $Q = \{Q_d, Q_{d^2}, \ldots, Q_{d^k} \ldots\}$ be a family of concentrators where Q_n has n inputs and n/d outputs. Define $P_d = X_{d,d}$ and $P_n(Q) = X_{1,d} \times (Q_n; P_{n/d}(Q))$ for all n that are powers of d. See Figure 3. Pippenger [8] showed that for $b = B = \beta = 1$ and $d = 2$, if Q is a family of wide-sense (rearrangeably) nonblocking concentrators then P_n is a wide-sense (rearrangeably) nonblocking distributor. To understand this result, note that a route from an input x to an output y, must pass through a unique sequence of recursively constructed subnetworks. The *branch switches* $X_{1,d}$ allow the route to pass to the required subnetworks without conflict and the route must be able to pass through the required concentrators if y is idle since each of these concentrators must have at least one idle output. This is illustrated in Figure 3. Note that branching is restricted to the branch switches and the crossbars in the last stage.

Ofman [6] shows that the reversed banyan network, $\overline{Y}_{n,2}$ is a rearrangeable concentrator when $b = B = \beta = 1$, yielding an explicit construction of a rearrangeable distribution network in the classical context. Similarly, since the Cantor network $K_{n,d,m}$ is a strictly nonblocking connector when $m \geq (2/d)(1 + (d-1)\log_d n/d)$ it is also a strictly nonblocking concentrator and Pippenger's construction yields a wide-sense nonblocking distribution network. Our first two theorems generalize these results to the multirate environment.

Let $\overline{Y}_{n,m,d}$ denote the reversed banyan network $\overline{Y}_{n,d}$ but with the outputs restricted to any m consecutive elements of $[0, n-1]$.

Theorem 1. *Let $Q = \{\overline{Y}_{n,n/d,d}\}$. Then $P_n(Q)$ is a rearrangeable distributor if $(1/\beta) \geq 2$.*

Theorem 2. *Let $Q = \{B_{n,d}\}$. Then $P_n(Q)$ is a wide sense nonblocking distributor if*

$$1/(\beta + B) \geq \frac{2}{d \max(b, 1-B)}(1 + (d-1)\log_d(n/d)).$$

Theorem 1 follows from Pippenger's basic construction and the following theorem which gives conditions under which the reversed banyan network is a rearrangeable concentrator.

Theorem 3. *Given any concentration assignment for $\overline{Y}_{n,d}$ with total weight $w \leq n$ and any $y \in [0, n-1]$, there is a state of $\overline{Y}_{n,d}$ that realizes the assignment using only outputs in $S = \{y, (y+1) \bmod n, \ldots, y + (r-1) \bmod n\}$ of $\overline{Y}_{n,d}$, where $r \leq \min\{2w, n\}$. Hence, for all $m \leq n$, $\overline{Y}_{n,m,d}$ is a rearrangeable concentrator if $\beta \leq 1/2$.*

Proof. The proof is by induction on the number of stages. The state constructed to establish the theorem also satisfies two other properties not explicitly mentioned above. First, for all $z, (z+1) \bmod n \in S$, the total weight on the outputs z, $(z+1) \bmod n$ is strictly greater than 1. Second, the distribution of weights on consecutive outputs is insensitive to the choice of y; that is, the weight on the j-th output in the group is a property of our overall routing strategy and is not affected by the specified starting output.

We show that given any concentration assignment with total weight $w \leq n$, and any output y, there is a state of $\overline{Y}_{n,d}$ that realizes the assignment using only outputs

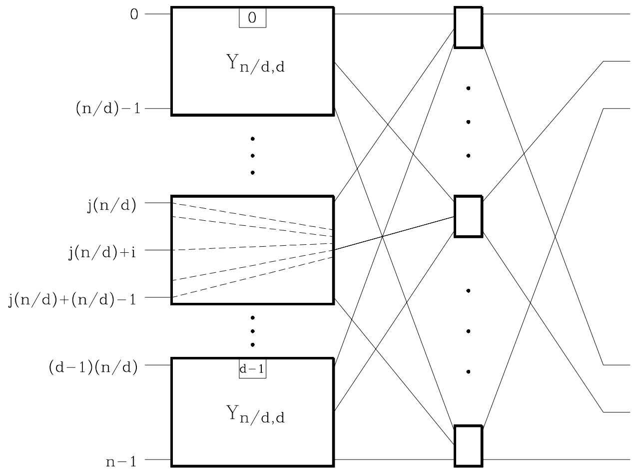

Figure 4: Reversed Banyan Network

in $S = \{y, (y+1) \bmod n, \ldots, y+(r-1) \bmod n\}$ of $\overline{Y}_{n,d}$, where $r \leq \min\{2w, n\}$ and for which, for all $z, (z+1) \bmod n \in S$, the total weight on the outputs z, $(z+1) \bmod n$ is strictly greater than 1. The routing strategy we use to establish the assertion has the property that the distribution of the weights on consecutive outputs is insensitive to the choice of y; that is, the weight on the j-th output in the group does not depend on the output we start with.

For a single stage network, we route the requests of the form $(0, \omega)$ to output y. We then route as many of the requests of the form $(1, \omega)$ as will fit on output y without overloading. When we can't place any more connections on output y, we proceed to output $(y+1) \bmod n$. Continuing in this fashion, results in a state that satisfies the conditions given above.

Assume, then that the induction hypothesis is true for all $n = d^i$, where $i < k$ and consider a k stage network with $k > 1$. Figure 4 shows the structure of $\overline{Y}_{n,d}$. Notice how it is made up of recursive subnetworks that are connected through a set of switches to the outputs. Let w_j be the total weight in the connection assignment involving inputs in $[j(n/d), (j+1)(n/d)-1]$. By the induction hypothesis, these connections can be routed to outputs $\{y_j, (y_j+1) \bmod (n/d), \ldots, (y_j+(r_j-1)) \bmod (n/d)\}$ of recursive subnetwork j, for any choice of y_j in $[0, (n/d)-1]$ and for $r_j \leq \min\{2w_j, n/d\}$. Furthermore, for any consecutive pair of outputs with non-zero weight, the total weight will be strictly greater than 1.

To establish the truth of the theorem, we must select values of y_j that will allow the connections to be routed through the final stage of the network to the output set S. We start by letting $y_0 = y \bmod (n/d)$. This will route the connections from subnetwork 0 to switches in the last stage that have access to outputs $y, \ldots, y+(r_0-1) \bmod n$. We configure the last stage switches to route the connections in this fashion and then proceed to subnetwork 1. Let $z = y + (r_0 - 1) \bmod n$. We wish to route the connections from subnetwork 1 to the set of outputs starting at either z or $(z+1) \bmod n$. The choice between these two alternatives will depend on whether the resulting weight on z would be acceptable or not. In particular, if routing the connections from subnetwork 1 to outputs $z, z+1 \bmod n, \ldots$ would lead to a load less than or equal to 1 on output z, we route them that way; that is, we let $y_1 = z \bmod (n/d)$. Otherwise we let $y_1 = (z+1) \bmod (n/d)$. We make a similar decisions when selecting y_2, y_3 and so forth. Proceeding in this fashion yields a network state satisfying the condition to be proved, and hence establishing the theorem. □

To prove Theorem 2 we need the following theorem which is proved in [4]

THEOREM 4. *$B_{n,d}$ is a strictly nonblocking connector if*

$$(1/\beta) \geq \frac{2}{d \max(b, 1-B)}(1 + (d-1)\log_d(n/d)).$$

Now, to establish Theorem 2, we need a routing strategy that ensures that the conditions required to make the concentrators strictly nonblocking are met. Whenever setting up or augmenting a connection we require that it not place a weight greater than $\beta + B$ on the input or output links of any of the concentrators. Given the bound on $\beta + B$ in the statement of Theorem 2, this will ensure that routes can be found through the required concentrators.

Suppose we are adding z to an existing route (x, Y, ω) and let i be the largest integer for which there exists a $y \in Y$ with $\lfloor z/d^{k-i} \rfloor = \lfloor y/d^{k-i} \rfloor$. Then, the current route includes a path which leads toward z for the first i levels in the recursive construction of P_n. At level $i+1$, that path reaches a branch switch from which the subnetwork containing z contains no current element of Y. There is a unique sequence of concentrators along this path. Consider any such concentrator and let m be the number of outputs the concentrator possesses. The number of network outputs that can be reached from this concentrator is also m, hence the total weight on the concentrator's outputs is at most $\beta m - \omega$. Hence, there is at least one output of the concentrator with a weight of less than β, and since $\omega \leq B$, the new path can be routed through this output without violating the weight constraint of $\beta + B$ on concentrator outputs. A similar argument applies to inputs. This completes the proof of Theorem 2.

IV. Modified Ofman-Thompson Network

Ofman [6] and Thompson [11] showed that the network $\overline{Y}_{n,2}; Y_{n,2}; B_{n,2}$ is a rearrangeable distributor when $b = B = \beta = 1$. We show that a similar network is a rearrangeable distributor in the multirate environment.

THEOREM 5. *$B_{n,d}; Y_{n,d}; B_{n,d}$ is a rearrangeable distributor when*

$$1/(\beta + B) \geq 1 + \frac{d-1}{d}(B/(\beta+B))\log_d(n/d)$$

or

$$1/(\beta + B) \geq 2 + \max\{0, \ln\log_d(n/d) - \ln\lfloor 1 + \beta/B\rfloor\}$$

So for example, if $n = 1024$, $d = 32$ and $\beta = B$ then $(1/\beta) \geq 3$ is sufficient to ensure rearrangeable operation. If $n = 2^{15}$, $d = 32$ and $\beta = B$, $(1/\beta) \geq 4$ is sufficient.

To use $B_{n,d}; Y_{n,d}; B_{n,d}$ as a rearrangeable distributor, we use point-to-point routing in the first and last subnetworks, allowing branching to occur only in the middle subnetwork. The proof of Theorem 5 requires a couple results describing the blocking characteristics of the subnetworks. The following theorem is proved in [4].

Theorem 6. *$B_{n,d}$ is a rearrangeable connector when*

$$(1/\beta) \geq 1 + \frac{d-1}{d}(B/\beta)\log_d(n/d)$$

or

$$(1/\beta) \geq 2 + \max\{0, \ln\log_d(n/d) - \ln\lfloor\beta/B\rfloor\}$$

A less general version of the following proposition is proved in [6].

Proposition 1. *Let $0 \leq r \leq n-1$ and let $C = \{(x_0, y_0, 1), \ldots, (x_{r-1}, y_{r-1}, 1)\}$ be a connection assignment for $Y_{n,d}$, where $y_0 < \cdots < y_{r-1}$ and for $1 \leq i \leq r-1$, $x_i = x_{i-1} + 1 \bmod n$. Then, there is a state of $Y_{n,d}$ that realizes C.*

Proof. By induction on the number of stages. For a single stage, $Y_{n,d}$ is a crossbar so clearly it satisfies the theorem. Consider then a network with more than one stage.

Each of the subnetworks formed when the first stage is removed is a banyan network, so we need only show that the first stage can route all connections to the proper subnetworks and that the connection requests passed on to the subnetworks satisfy the condition in the statement of the theorem.

Consider subnetwork j; $\ell_j = j(n/d)$ and $h_j = \ell_j + (n/d) - 1$ are the first and last outputs of subnet j. Let a be the smallest integer such that $l_j \leq y_a \leq h_j$ and let b be the largest integer such that $l_j \leq y_b \leq h_j$. Note that the connection requests that are to be routed to subnet j all have indices in the interval $[a, b]$ implying that $b - a + 1 \leq n/d$.

Because all the connection requests involving subnet j appear on consecutive inputs to the network, and there are at most n/d of them, they appear on inputs connected to distinct switches in stage 1. Consequently, all can be routed to subnet j without conflict. Also, because the connection requests for subnet j appear on consecutive inputs to the network, they pass through consecutive stage 1 switches, which in turn connect to consecutive inputs on subnet j, implying that the connection requests seen by subnet j satisfy the conditions of the theorem.

The above argument holds independently for each of the subnetworks. Applying the induction hypothesis to each of the subnetworks then, yields the theorem. □

The following proposition is an easy generalization of the previous one.

Proposition 2. *Let $0 \leq r \leq n-1$ and let $A = \{(x_0, Z_0, 1), \ldots, (x_{r-1}, Z_{r-1}, 1)\}$ be a distribution assignment for $Z_{n,d}$, where $y_1 \in Z_i$ and $y_2 \in Z_{i+1}$ implies that $y_1 < y_2$ and for $1 \leq i \leq r-1$, $x_i = x_{i-1} + 1 \bmod n$. Then, there is a state of $Y_{n,d}$ that realizes A.*

Proof of Theorem 5 Let $A = \{r_i = (x_i, Z_i, \omega_i) | 0 \leq i \leq q-1\}$ be a distribution assignment for $B_{n,d}; Y_{n,d}; B_{n,d}$, and assume the r_i are sorted by weight, so that $\omega_i \geq \omega_{i+1}$ for $i \in [0, q-2]$. Also, let $f_i = |Z_i|$ and $s_i = \sum_{j \leq i} f_i$ for $i \in [0, q-1]$.

Assume for the moment, that $\lfloor (s_{i-1}+1)/n \rfloor = \lfloor s_i/n \rfloor$ for $i \in [1, q-1]$ and let $A_j = \{r_i | \lfloor s_i/n \rfloor = j\}$. (This assumption will be eliminated later.) We constrain the choice of routes so that for $r_i \in A_j$, the selected route starts at x_i and passes through input $(i-j) \bmod n$ and outputs $(s_{i-1}+1) \bmod n, \ldots, s_i \bmod n$ of the central subnetwork, before proceeding through the third subnetwork to the members of Z_i. Notice that for all $j \geq 1$, the route for the last request in A_{j-1} and the route for the first request of A_j pass through a common input of the central subnetwork.

Given these constraints and Proposition 2, the requests in each of the A_j can be routed through the central subnetwork without using any common links. Consequently, each link in the central subnetwork is included in at most one route realizing requests in A_j. Hence, the weight on each link in the central subnetwork is at most

$$\sum_{j \geq 0} \max_{r_i \in A_j} \omega_i = B + (\beta n - B)/n < \beta + B$$

Since this is ≤ 1, the indicated routes can be handled by the central subnetwork. Since the weight on the input and output links of the first and last subnetworks is at most $\beta + B$, the bounds on $\beta + B$ given in the statment of the theorem together with Theorem 6 imply that the indicated routes can be handled by the first and last subnetworks.

Now all that remains is to eliminate our earlier assumption that $\lfloor (s_{i-1}+1)/n \rfloor = \lfloor s_i/n \rfloor$. Suppose now that for some i, $\lfloor (s_{i-1}+1)/n \rfloor \neq \lfloor s_i/n \rfloor$. In such a case, we split request r_i into two requests $r_{i,1} = (x_i, Z_{i,1}, \omega_i)$ and $r_{i,2} = (x_i, Z_{i,2}, \omega_i)$ where $Z_{i,1} \cup Z_{i,2} = Z_i$ and $s_{i-1} + |Z_{i,1}|$ is evenly divisible by n. By doing this for all requests that violate our assumption we obtain a new set of requests that satisfies the assumption. Hence we can apply the routing strategy given earlier to this new set of requests. Notice that because the our routing strategy routes the last request in A_{j-1} and the first request of A_j through

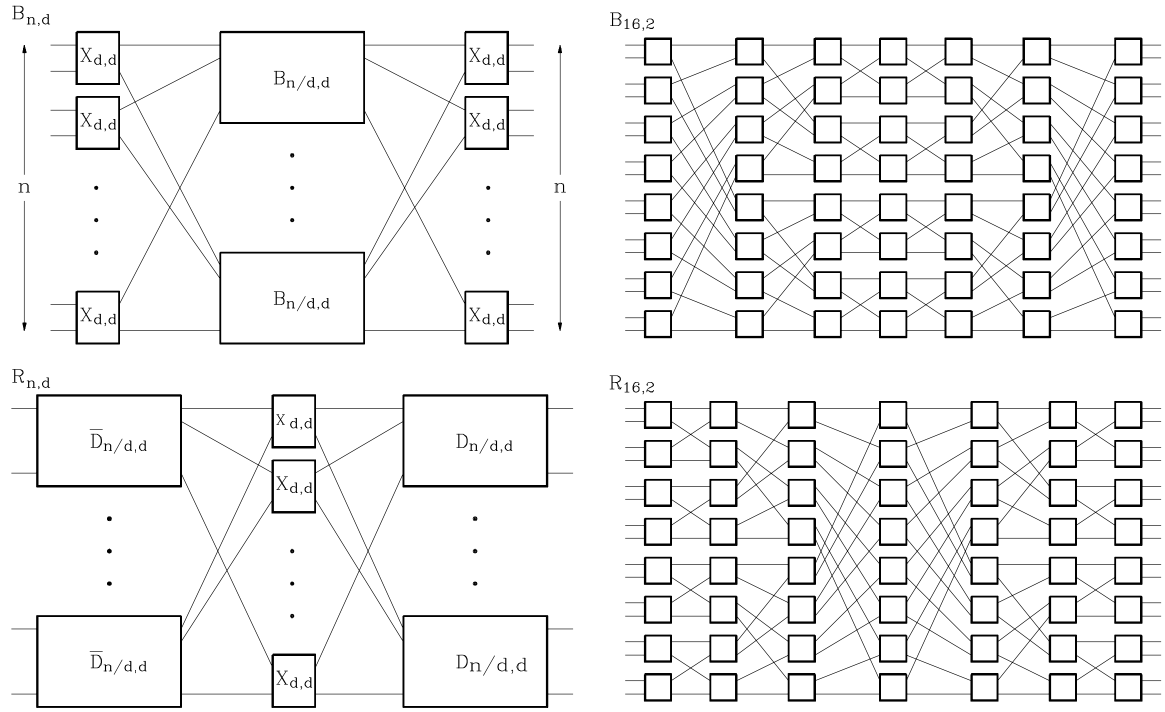

Figure 5: Recursive Structures of $B_{n,d}$ and $R_{n,d}$

a common input of the central subnetwork, this does not require branching in the first subnetwork. □

V. A Nearly Wide-sense Nonblocking Distributor

The definitions of wide-sense and strictly nonblocking distributors require that the network handle both distribution requests and augmentation requests. If we require only the ability to handle distribution requests, we obtain a class of networks that is intermediate in power between the rearrangeable and wide-sense nonblocking distributors. We call such networks *nearly wide-sense nonblocking distributors.*

Theorem 7. *$B_{n,d}; B_{n,d}$ is a nearly wide sense nonblocking distributor when*

$$1/\beta \geq \frac{2}{d \max(b, 1-B)}(1 + (d-1)\log_d(n/d)).$$

Proof. For convenience we introduce an alternative description of the Beneš network. Let $R_{n,d} = \overline{D_{n/d,d}} \bowtie X_{d,d} \bowtie D_{n/d,d}$. It is not difficult to show that $R_{n,d}$ is topologically equivalent to $B_{n,d}$. Figure 5 compares the recursive structure of the two networks. We can view $R_{n,d}$ as being an "inside-out" version of $B_{n,d}$. For simplicity of description, the remainder of the the proof addresses the network $B_{n,d}; R_{n,d}$, but the results hold equally well for $R_{n,d}; B_{n,d}$.

Let S be the set of states of $R_{n,d}; R_{n,d}$ in which branching occurs only in the second subnetwork and in which the weight on any input of the second subnetwork is at most $\beta + B$. Given any state $s \in S$ and a distribution request, (x, Y, ω) compatible with s, we first identify an input z to the second subnetwork carrying a weight of at most β and from which more than half the middle stage switches in the second subnetwork are ω-accessible. The key to proof is showing that given the conditions of the theorem, there must exist such a z. We also show that given the conditions of the theorem, more than half the middle stage switches of the second subnetwork are ω-accessible from each output in Y and that z is ω-accessible from x. These facts together imply the existence of a route r realizing the request for which $s \cup \{r\} \in S$.

For $0 \leq i < k = \log_d n$, define $L_i(u)$ to be the set of stage i links that can be reached from input u in an idle network $R_{n,d}$. We note that $L_i(u) = L_i(v)$ if $\lfloor u/d^i \rfloor = \lfloor v/d^i \rfloor$, so

$$L_i(0), L_i\left(d^i\right), \ldots, L_i\left(jd^i\right), \ldots, L_i\left(\left(d^{k-i}-1\right)d^i\right)$$

partitions the links in stage i into d^{k-i} groups of d^i links each.

To find an input z to the second subnetwork, we work backward from the middle stage of the second subnetwork, seeking the most "lightly loaded" portion of the subnetwork at each step. Define $W_i(j)$ to be the weight on the links in $L_i\left(jd^i\right)$ let $W_i^* = \min_j W_i(j)$. Note that $W_i^* \leq (\beta n - \omega)/d^{k-i} < \beta d^i$ and hence that there is a z such that for $0 \leq i \leq k-1$ the total weight on the links in $L_i(z)$ is $< \beta d^i$.

Let Q_i be the set of links (u, v) in stage i of the second subnetwork for which u is ω-accessible from z but v is not. Also, let λ_i be the total weight on all links in Q_i and note that $|Q_i| f(\omega) \leq \lambda_i \leq \beta d^i$, where $f(\omega) = \max\{b, 1-\omega\}$. The number of middle stage switches of the second subnetwork that are not ω-accessible from z is exactly

$$\begin{aligned} \sum_{i=0}^{k-1} |Q_i| d^{k-i-1} & \\ &\leq \frac{1}{f(\omega)d} \sum_{i=0}^{k-1} d^{k-i} \lambda_i \\ &< \frac{1}{f(\omega)d} \left[\beta d^{k-1} + \sum_{i=1}^{k-1} d^{k-i}(d^i - d^{i-1})\beta \right] \\ &\leq \frac{\beta}{f(B)d^2} n(1 + (d-1)\log_d(n/d)) \\ &\leq n/2d \end{aligned}$$

Hence more than half of the middle stage switches are ω-accessible from z.

Next, we show that for all $y \in Y$, more than half of the middle stage switches of the second subnetwork are accessible from y. The argument is similar to the one given above. Redefine Q_i to be the set of links (u, v) in stage $2k - 1 - i$ of the second subnetwork for which v is ω-accessible from y but u is not. Also, let λ_i be the total weight on all links in Q_i and note that $|Q_i| f(\omega) \leq \lambda_i$. Also, note that $\lambda_i < \beta d^i$ since the number of outputs that can be reached from links in Q_i is exactly d^i, none of them can carry a weight of more than β and at least one (y) must carry a weight of less than β. The number of middle stage switches of the second subnetwork that are not ω-accessible from y is then

$$\begin{aligned} \sum_{i=0}^{k-1} |Q_i| d^{k-i-1} & \\ &\leq \frac{1}{f(\omega)d} \sum_{i=0}^{k-1} d^{k-i} \lambda_i \\ &< \frac{1}{f(\omega)d} \left[\beta d^{k-1} + \sum_{i=1}^{k-1} d^{k-i}(d^i - d^{i-1})\beta \right] \\ &\leq \frac{\beta}{f(B)d^2} n(1 + (d-1)\log_d(n/d)) \\ &\leq n/2d \end{aligned}$$

Hence more than half of the middle stage switches are ω-accessible from y.

Finally, we need to show that a route can be found from input x of the first subnetwork to z. First we note that since z is the most lightly loaded input of the second subnetwork, it is also a lightly loaded output of the first subnetwork; that is, more than half the middle stage switches of the first subnetwork are ω-accessible from z. This can be proved using an argument very similar to the one given earlier. Also, since none of the inputs to the first subnetwork has a weight of more than β, more than half the middle stage switches are ω-accessible from x, implying that there is an available path from x to z. □

As an example application of the theorem, if we let $b = 0$, $B = \beta$, $d = 16$ and $n = 256$, we obtain an almost wide-sense nonblocking network if $(1/\beta) \geq 3$. The number of stages in the network is 6, but we can reduce this by one by noting that each switch in the last stage of the second network is directly connected to the corresponding switch in the first stage of the second network, so we lose nothing by omitting one these stages. As we have noted above, this network will sometimes block when we attempt to augment an existing connection. If we are willing to rearrange the connection, we can avoid blocking. Note that only the connection being augmented is affected by this rearrangement, making this a fairly easy rearrangement to perform.

We close by noting that similar results can be proved for other network pairs. A pair of Clos networks or a pair of Cantor networks can be used for example. In the multirate environment however, the Beneš networks typically yield the lowest complexity solution.

VI. Closing Remarks

The results given here generalize classical results on nonblocking distribution networks. Furthermore, the network model we have developed is directly applicable to several ATM switching systems now under development [1, 10]. In particular, several groups have proposed the use of a Beneš type topology for ATM networks, without apparently understanding the blocking implications. Our studies indicate that while such networks cannot be made strictly nonblocking for practical values of β, additional stages can yield a network that is nonblocking for all new distribution requests; while blocking can still occur for augmentation requests, even this blocking can be avoided if we are willing to occasionally rearrange a distribution request in order to augment it. Only the specific request being augmented is affected by this rearrangement, making it a fairly straightforward operation.

The complexity of a space-division network can be measured in terms of the number of integrated circuits required to implement it. The complexity of a nonblocking multirate network is defined to be the same as the complexity of the underlying space division network times the speed advantage $(1/\beta)$ needed to make it nonblocking. So for example, the rearrangeable version of Pippenger's network has a complexity that is roughly twice that of the corresponding space division network. Similarly, most of our results for the multirate case have complexity that is roughly twice that for the comparable space division network. In the case of the Ofman-Thompson network, the multirate case requires a network whose complexity is larger than that of the space division network by a *log log* factor.

References

[1] Coudreuse, J. P. and M. Servel "Prelude: An Asynchronous Time-Division Switched Network," *Inter-*

national Communications Conference, 1987.

[2] Feldman, P., J. Friedman and N. Pippenger. "Nonblocking Networks," *Proceedings of the ACM Symposium on the Theory of Computing*, 5/86, pp. 247–254.

[3] Goke, G. R. and G. J. Lipovski. "Banyan Networks for Partitioning Multiprocessor Systems," *1st Ann. Symp. on Computer Architecture*, pp. 21–28, 1973.

[4] Melen, R. and J. S. Turner. "Nonblocking Multirate Networks," *SIAM Journal on Computing*, Vol. 18, No. 2, pp. 301–313, April 1989.

[5] Melen, R. and J. S. Turner. "Nonblocking Networks for Fast Packet Switching," *Proceedings of Infocom 89*, April 1989.

[6] Ofman, Y. P. "A Universal Automaton," *Transactions of the Moscow Mathematics Society*, Vol.14, 1965, pp.200–215.

[7] Patel, J. A. "Performance of Processor-Memory Interconnections for Multiprocessors," *IEEE Transactions on Computing*, vol. C-30, pp. 771–780, 1981.

[8] Pippenger, N. "The Complexity of Switching Networks," Ph.D. thesis, Massachusetts Institute of Technology, Department of Electrical Engineering, 9/73.

[9] Pippenger, N. "Generalized Connectors," *SIAM Journal on Computing*, Vol. 7, No. 4, Nov. 1978, pp. 510–514.

[10] Suzuki, Hiroshi, Hiroshi Nagano, Toshio Suzuki, Takao Takeuchi and Susumu Iwasaki. "Output-buffer Switch Architecture for Asynchronous Transfer Mode," *Proceedings of the International Communications Conference*, pp. 99–103, June 1989.

[11] Thompson, C. D. "Generalized Connection Networks for Parallel Processor Interconnection," *IEEE Transactions on Computers*, Vol. C-27, No. 12, Dec. 1978, pp. 1119–1125.