Report Number: WUCS-TM-92-02

1992-05-22

Research Proposal: Design and Analysis of Practical Switching Networks

Authors: Ellen E. White

At the heart of any communication system is the switching system for supporting connections between sets of endpoints. A switching system consists of one or more switching networks connected by communication links. Effective design of switching networks is critical to the success of a communication system. This paper proposes the study of three problems in the design of switching networks: design of nonblocking multirate distribution, evaluation of blocking probability in distributors and quantitative comparison of architectures. Each of these problems is significant in the design of practical networks and lacks broad analytic treatment.

Research Proposal: Design and Analysis of Practical Switching Networks

Ellen E. Witte

WUCS-TM-92-2

May 22, 1992

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

Abstract

At the heart of any communication system is the *switching system* responsible for supporting connections between sets of endpoints. A switching system consists of one or more *switching networks* connected by communication links. Effective design of switching networks is critical to the success of a communication system. This paper proposes the study of three problems in the design of switching networks: design of nonblocking multirate distributors, evaluation of blocking probability in distributors and quantitative comparison of architectures. Each of these problems is significant in the design of practical networks and lacks broad analytic treatment.

*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.

Research Proposal: Design and Analysis of Practical Switching Networks

Ellen E. Witte

1. Introduction

At the heart of any communication system is the switching system responsible for supporting connections between sets of endpoints such as computers or telephones. If the number of endpoints is not too large, the network may be as simple as a single switching network to which each of the endpoints is connected. If the number of endpoints exceeds the capacity of a switching network, or the endpoints are geographically dispersed, the system generally consists of multiple switching networks connected to one another by high speed transmission links, such as optical fiber. Advances in optical technology have resulted in transmission rates between switching networks in the 1 Gbit/sec range, speeds as yet unmatched by electronic or optical switching technology.

Effective design of switching networks is critical to the success of a communication system. Poor design may result in problems such as excessive delays in creating connections, an inability to connect certain endpoints, errors in transmitting data, and costly systems. Given the ubiquitous nature of communications systems, it is no surprise that extensive research effort has been devoted to theoretical and practical study of switching networks. Unfortunately, there are many issues arising in the practical arena for which there is little directly applicable theoretical work. The goals of current technology include the following:

- Support for diverse connection bandwidths
- Ability to connect sets of endpoints
- Low probability of rejecting specific connection requests
- Low cost

There is some theoretical basis for determining how to achieve these goals, however in many cases the theoretical work is not directly applicable due to an inaccurate abstraction of the real world problem. This work focuses on three problems for which this is the case.

The first problem concerns design of networks to support diverse connection bandwidths between sets of endpoints. Until recently the theoretical work has made the assumption of a single connection bandwidth. This is not an accurate model of current communications systems capable of transmitting voice, video and data at widely varying rates. While there has been work in the area of connecting sets of endpoints, there is room for additional progress particularly concerning networks of practical size with efficient routing algorithms.

The second problem addresses determining the probability a specific connection request will be rejected due to insufficient resources. Practical networks can generally tolerate some inability to support certain connections, provided this occurs infrequently. There are several widely used models for evaluating blocking probability for connections between two endpoints, but not for more general connections.

The third problem involves quantitative comparisons of network architectures using alternate measures of network cost. The classic theoretical measure of network complexity is crosspoint count. While this is a reasonable reflection of network cost for networks constructed of electromagnetic relays, it is not a good measure for networks constructed using VLSI technology. There is a need for comparisons of networks based on measures that more accurately reflect cost.

These three problems will be studied with the objective of providing a theoretical basis for practical network design. The next section contains definitions needed to formally discuss the problems. Sections 3, 4 and 5 cover each of the problems in more detail, with a review of related work, a statement of the problem, a discussion of progress thus far and a research plan. Finally, Section 6 contains a summary of the proposal with emphasis on the expected results.

2. Definitions

This section contains formal definitions needed throughout the proposal. Most of these definitions are based on those in the work of Melen and Turner [15, 16] and Pippenger [21].

Graph Model

For the remainder of the proposal, the term *network* will be used to refer to the switching network described in the introduction. A graph model is used to represent network topology. For network N, we associate 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 $y \in O \cup S$. Each input and output terminal must appear in exactly one link. Links including an input terminal are called *inputs*, those containing an output terminal are called *outputs*. A network with n inputs and m outputs is referred to as an (n, m)-*network*. An (n, n)-network is also called an n-*network*.

The networks we consider can be divided into a sequence of *stages*, with links allowed only between switches in adjacent stages. The input vertices are in stage 0 and for $i > 0$, a vertex v is in stage i if for all links (u, v), u is in stage $i - 1$. A link (u, v) is in stage i if its

left endpoint u is in stage i. We will consider only networks in which all of the outputs are in the same stage, and no other vertices are in this stage. When the outputs are in stage k, it is called a *k-stage network*.

Construction Operators

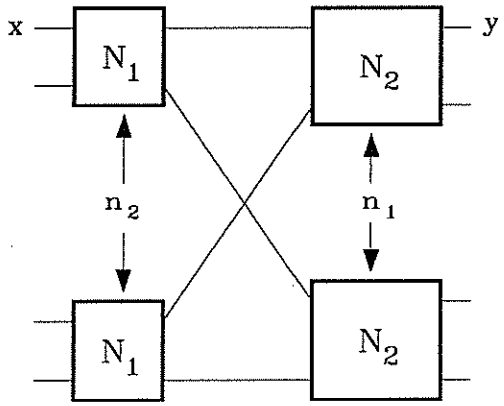

Figure 1: Series Construction $N_1 \times N_2$

The topology of many interesting networks in the literature can be described by two simple construction operators originally proposed by Cantor [3]. If N_1 is a network with n_1 outputs and N_2 is a network with n_2 inputs, then the *series connection* of N_1 with N_2 is denoted $N_1 \times N_2$ and is constructed as shown in Figure 1. Informally, this consists of taking n_2 copies of N_1 in one column and connecting them to n_1 copies of N_2 in a second column, with one link between each pair of subnetworks.

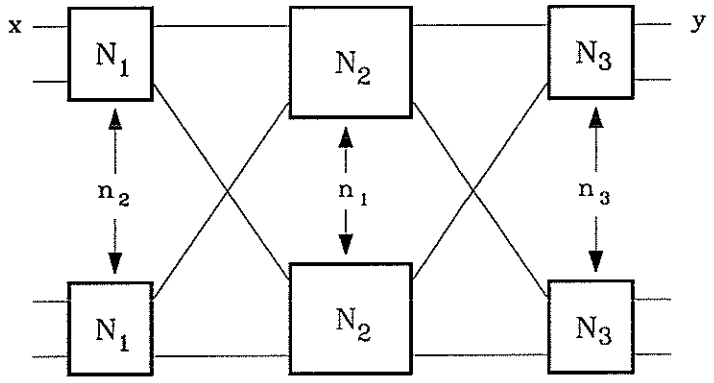

Figure 2: Parallel Construction $N_1 \otimes N_2 \otimes N_3$

The second construction operator combines three networks. If N_1 is a network with n_1 outputs, N_2 is an (n_2, n_3)-network and N_3 is a network with n_1 inputs, then the *parallel connection* of N_1, N_2 and N_3 is denoted $N_1 \otimes N_2 \otimes N_3$ and is constructed as shown in Figure 2. Informally, this consists of taking n_2 copies of N_1 in one column, n_1 copies of N_2 in a second column and n_3 copies of N_3 in a third column. There is one link between each pair of subnetworks in adjacent columns. A more formal definition of the series and parallel operators can be given, but is omitted here.

These operators can be used to describe several popular networks. The basis component of all of these networks will be the $d_1 \times d_2$ crossbar switch, denoted X_{d_1,d_2}. The delta network [20] with n inputs constructed of $d \times d$ switches is denoted $D_{n,d}$ and defined recursively as:

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

This network has $\log_d n$ stages and provides exactly one path between each input/output pair. The delta network is isomorphic to other popular topologies such as the banyan [14] and omega [12] networks. A common measure for network complexity is the number of crosspoints. A $d_1 \times d_2$ crossbar has $d_1 d_2$ crosspoints. The delta network $D_{n,d}$ has $nd \log_d n$ crosspoints.

The Beneš network [2], denoted $B_{n,d}$, is defined using the parallel constructor.

$$B_{d,d} = X_{d,d} \qquad\qquad B_{n,d} = X_{d,d} \otimes B_{n/d,d} \otimes X_{d,d}$$

This network has $2\log_d n - 1$ stages and $nd(2\log_d n - 1)$ crosspoints. It provides multiple paths between each input/output pair. The Cantor network [3], denoted $K_{n,d,m}$, consists of m parallel Beneš networks.

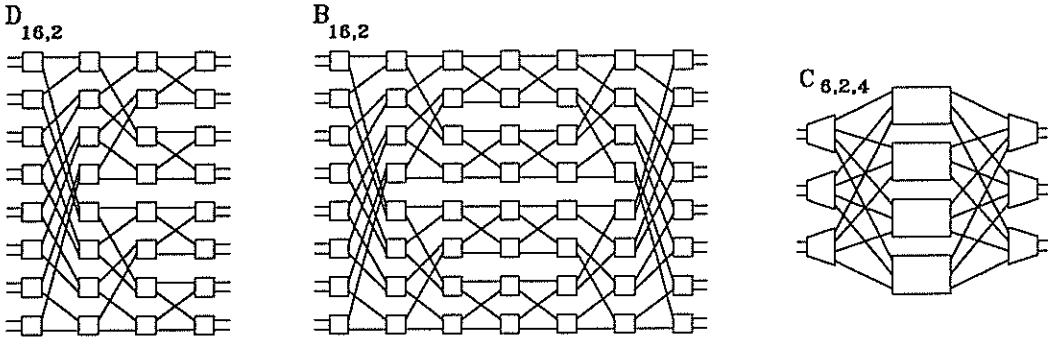

Figure 3: Example Constructions

The three stage Clos network [6], denoted $C_{n,d,m}$, is similar to the three stage Beneš network, except that the first and last stage crossbar switches do not have to be *square*, that is, have the same number of inputs and outputs.

$$C_{n,d,m} = X_{d,m} \otimes X_{n/d,n/d} \otimes X_{m,d}$$

This network has m crossbar switches in the middle stage and $2nm + md^2$ crosspoints. The number of stages can be increased to $2k + 1$ for any k by repeatedly replacing the middle crossbars with three stage Clos networks.

Figure 3 shows examples of the delta, Beneš and Clos networks. Networks in which all of the switches are square are called *uniform networks*. The delta and Beneš networks are uniform.

Any network can be used in a variety of environments. Factors that determine the environment include the number of participants in a request (point-to-point versus multipoint)

and the allowed bandwidth of requests (circuit switching versus multirate). The next several paragraphs contain definitions relevant to the network environment.

Connectors, Distributors, etc.

A *connector* is a network that operates in the point-to-point environment. 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. A state *realizes* a given assignment if it contains one route realizing each request in the assignment and no others. A state s_0 is *below* a state s_1 if $s_0 \subseteq s_1$. Similarly s_1 is *above* s_0. A connection request (x, y) is *compatible* with a state s if x and y are idle in s.

A *distribution network* or *distributor* operates in the multipoint enviroment. Distributors 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. Some of the extended definitions are included 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.

There are other types of networks with variations on the participants in a request. For example, a concentrator connects a specific input to an arbitrary output. A replicator connects a specific input to a set of arbitrary outputs. Reference [24] gives a description of the types of networks and the known complexity results for their construction.

The Multirate Environment

Classical theory in network design was developed at a time when communication systems were operated in a *circuit switching* environment. When a connection was established, the necessary resources were dedicated to the connection and remained dedicated for the duration of the connection. Circuit switching 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 [1, 26]. 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 [15]. Networks allowing a connection weight are referred to 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 and B denote the largest weight. In addition we define β to be the maximum port weight, i.e., the maximum sum of the weights of connections involving any particular input or output. By definition, $0 < b \leq B \leq \beta \leq 1$. A common choice for W is the interval $[b, B]$. The quantity $1/\beta$ is called the *speed advantage*, as it indicates the ratio of the internal network speed to the external link speed.

The definitions given previously must be extended to include connection weights. In this environment, a connection assignment is a set of requests that obey the maximum port weight constraint. A state is 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.

The goal of a network operating in a particular environment is to satisfy compatible requests. Clearly, this will not always be possible. Networks can be classified into four categories based on ability to satisfy requests. These categories are described next.

Ability to Satisfy Requests

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. A network is a *rearrangeably nonblocking 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 *blocking connector* if there exist states in which a compatible request cannot be realized. Clearly any strictly nonblocking network is a wide-sense nonblocking network and every wide-sense nonblocking network is a rearrangeably nonblocking network.

The definitions of strictly nonblocking, wide-sense nonblocking and rearrangeably nonblocking can be extended to distribution networks. 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 every augmentation request r compatible with s can be satisfied in such a way that the resulting state is in S. A network is a *rearrangeably nonblocking*

distributor if for every assignment, there exists a state realizing the assignment. A network is a *blocking distributor* if there exist states in which a compatible distribution or augmentation request cannot be realized. 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.

3. Design of Multirate Distributors

This thesis is concerned with providing theoretical support for the design of practical communication networks. A primary requirement of practical networks is that they exhibit little or no blocking of compatible requests. The next section addresses estimating blocking probability; here we consider nonblocking networks.

There is already a well developed theory for construction of nonblocking networks with the emphasis on minimizing the asymptotic crosspoint complexity in the circuit switching environment [24]. There is also some theory for the multirate environment [15, 17, 5]. We propose to study constructions of nonblocking networks with particular concern for practical issues. First, we will restrict our attention to distributors, to reflect the need for practical networks to support multipoint applications such as video conferencing and broadcast services. Second, we will be interested in the multirate envionment, which models the diversity of traffic experienced in practical networks. Third, we will pay close attention to the efficiency of the routing algorithms associated with wide-sense and nearly nonblocking distributors. Finally, we will be interested in networks that are nearly nonblocking but can handle augmentation requests with low impact on the current state. For example, it may be acceptable in some application environments to use a nearly nonblocking network if one can ensure that an augmentation will disturb only the augmented connection and no others.

3.1. Related Work

In the field of theoretical nonblocking networks, there are a number of results that are not of practical interest, for reasons ranging from impractical constants hidden by asymptotic complexity notation to routing algorithms which must solve NP-complete problems. We do not include these types of results in this review.

In designing multirate distribution networks, it is logical to start with circuit switching distribution networks. The best practical wide-sense nonblocking distributors are obtained by starting with Pippenger's tree-like network [21] and using using Cantor networks for the concentrators. The resulting network has complexity $O(n(\log n)^3)$. The best practical nearly nonblocking distributor is due to Turner [31] and has complexity $O(n(\log n)^2)$. Turner's construction consists of *cascading* either a pair of Cantor networks or a pair of Clos networks, by connecting the outputs from the first network to the inputs of the second network.

The results above are stated without explicit reference to the number of stages in the network. There is also significant practical interest in minimizing complexity for a given number of stages, since delay is proportional to network depth. Taking this view, Turner's nearly nonblocking distributor has complexity $O(n^{1+1/(k+1)})$ for $4k+1$ stages. The five stage version ($k = 1$) has complexity $O(n^{3/2})$. Yang and Masson [38] give the best construction of a nearly nonblocking distributor with a fixed number of stages, using a single Clos network with an efficient greedy routing algorithm. The key to their result is the appropriate choice of switch dimensions in the Clos network so that the greedy algorithm will be successful in routing. The resulting network has 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 Melen and Turner [17]. Each result is stated as a condition relating the multirate parameters (W and β) and the parameters of the network. Pippenger's tree-like network with Beneš networks as the concentrators is a wide-sense nonblocking distributor if

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

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

Based on Turner's nearly nonblocking distributor composed of cascaded Cantor networks, a multirate nearly nonblocking distributor can be constructed of cascaded Beneš networks under the condition

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

This network has $O(n\log n)$ crosspoint complexity, and a speed advantage that grows as roughly $2\log n$. Stated in terms of the number of stages, this construction has complexity $O(n^{1+1/(k+1)})$ for $4k+1$ stages and speed advantage of about k.

While the cascaded Beneš networks are nearly nonblocking, the construction is attractive for augmentations because only the augmented connection must be rearranged. In addition, it has been conjectured that the amortized cost to rearrange would be low [30].

3.2. Summary of Progress

We have extended the work by Yang and Masson to the multirate environment [36]. The result is a nearly nonblocking multirate distributor with the same crosspoint complexity as in the circuit switching case, and a speed advantage of $k+1$, where $2k+1$ is the number of stages in the Clos network. Considering some specific values of k, this construction gives a network of three stages ($k = 1$) with crosspoint complexity of $O(n^{3/2}(\log n/\log\log n))$ and speed advantage of 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. In comparison, Turner's cascaded Beneš network requires five stages to get crosspoint complexity of $O(n^{3/2})$ with a speed advantage of 1.

We have also shown a negative result regarding the frequency of expensive augmentations in cascaded Clos networks. We have exhibited a sequence of operations in which expensive augmentations are required with only a constant number of inexpensive operations in between. It should be straightforward to apply the same ideas in the multirate environment to cascaded Beneš networks, thereby disproving the conjecture referred to earlier.

3.3. Research Plan

1. Extend promising distributors proposed for the circuit switching environment to the multirate environment. This has been done for the Yang and Masson distributor.
2. Investigate nearly nonblocking distributors that can augment connections with low impact.
    - Prove or disprove the conjecture that cascaded Beneš networks can be used as a nearly nonblocking distributor with low amortized cost to augment. This has been disproved for the Clos networks. It should be a direct extension to disprove for Beneš networks.
    - Try to turn the negative result into a positive one, by determining the number of middle stages needed so that the amortized time spent rearranging is low.
3. Consider other extensions to nonblocking theory to reflect practical networks in a similar vein as the multirate extension.

4. Blocking Probability Analysis

Network designers have realized that the complexity required to achieve nonblocking capability often cannot be justified for practical systems. Rather, a guarantee of acceptably low blocking probability under reasonable loading is sufficient, particularly if the resulting network offers considerable savings in complexity. Quantifying what is acceptable and reasonable depends upon the application, but a blocking probability of $10^{-2} - 10^{-3}$ with loading of $70 - 80\%$ is realistic for many applications. The goal of research in blocking probability analysis is to develop efficient and accurate tools for evaluating the blocking probability of networks.

Tools for evaluating blocking probability may be analytic or based on simulation. In general, simulation offers more fine control over network conditions and monitoring, however this is at the expense of time to develop and run simulations. For more broad comparisons of networks, analytic tools for evaluating blocking probability are particularly useful. Furthermore, analytic tools can be used to generate random states for simulations, to study asymptotic behavior of network complexity and to lend confidence to the programmer about the correctness of simulation code. We propose to investigate blocking probability in distributors. A major component of this research will be the development of a model for evaluating

blocking probability. While such models exist for connectors, there is no corresponding work for the multipoint environment.

In this section we review the two widely used models for blocking in connectors, Lee's model [13] and Pippenger's model [22] and highlight some other results on blocking. Next we propose a model for blocking in distributors, based on Lee's work. Included are examples of the application of the model to some common network constructions. The section concludes with a research plan.

4.1. Related Work

The majority of the previous work has concerned models for blocking in connectors. These models are *static* in the sense that they assign probabilities to possible network states at a typical moment in time. This contrasts with *dynamic models* that assign probabilities to the possible trajectories of the state over a period of time.

We define the *connection blocking probability* $P_C(N)$ for network N with input x and output y as follows:

$$P_C(N) \equiv Pr(x \leftrightarrow y \text{ blocked} \mid x \text{ idle}, y \text{ idle}),$$

where "$x \leftrightarrow y$ blocked" means that every path from x to y is blocked. A model for blocking in a network consists of an assignment of probabilities to the states of the network. Given an assignment of probabilities, the blocking probability is defined, but may be difficult to express in closed form. A model is judged by the ease with which it can be applied to a variety of networks to give accurate expressions for blocking probability.

There are two widely used models for evaluating blocking probability in connectors. The first was developed in 1955 by C. Lee [13]. His model is based on two assumptions.

1. Every link i is busy with probability p_i and idle with probability $q_i = 1 - p_i$.
2. The conditions of different links (that is, busy or idle) are independent.

The second assumption gives the model simplicity, but also makes it inaccurate. By making this assumption, the probability that a particular path is idle is simply the product of the probability each link in the path is idle. Unfortunately this assumption assigns nonzero probabilities to configurations of the network that are not states, for example, in which the number of busy inputs and outputs differ.

While Lee's model is not restricted to uniform networks, we will focus on this interesting class of networks. We will additionally restrict our attention to networks constructed using the series operator and the parallel operator with $N_1 = N_3 = X_{d,d}$. Because the networks are uniform, we can strengthen the first assumption to assume that all links have the same probability of being busy. That is, $p_i = p$ and $q_i = q = 1 - p$ for all i. This is justified by symmetry and conservation of traffic in the switch.

It suffices to show how the series and parallel construction operators affect the blocking probability. For the series construction of networks N_1 and N_2 we have

$$P_C(N_1 \times N_2) = 1 - q(1 - P_C(N_1))(1 - P_C(N_2)).$$

Referring to Figure 1, it is fairly easy to see the intuition behind this expression. There is a single link joining the copy of N_1 containing x and the copy of N_2 containing y. There is an idle path from x to y if and only if there is an idle path from x to this link, an idle path from this link to y, and this link is idle.

For the parallel construction of network N with $X_{d,d}$ we have

$$P_C(X_{d,d} \otimes N \otimes X_{d,d}) = (1 - q^2(1 - P_C(N)))^d.$$

Referring to Figure 2 with $N_1 = N_3 = X_{d,d}$ and $N_2 = N$, the intuition behind this expression is also straightforward. Each of the d subnetworks N offers a possible way to connect x to y. The overall network blocks if none of the subnetworks can support the connection. Focusing on the top subnetwork, the connection can be supported if the needed links into and out of the subnetwork are idle and if a path can be found through the subnetwork. This occurs with probability $q^2(1 - P_C(N))$. The same expression gives the probability any one subnetwork can support the connection.

Recognizing the inaccuracy in Lee's model, Pippenger developed a more accurate model for evaluating blocking probability that takes into account the dependencies between links incident to the same switch [22]. His model is based on four assumptions.

1. Every input is busy with probability p and idle with probability $q = 1 - p$.
2. The conditions of different inputs are independent.
3. If a given $d \times d$ switch has r busy inputs, then all of the $d^{\underline{r}} = d(d-1)\ldots(d-r+1)$ ways in which they may be connected to r busy outputs are equally likely.
4. The conditions of different switches in a given stage are independent.

Pippenger considered only uniform networks. As in Lee's model, we can show how the series and parallel construction operators affect the blocking probability. The expressions are more complicated, but more accurate. To take into account dependencies between links, a conditional probability is needed. If x is an input and y is an output of an m-network (a network with m inputs and m outputs), then we denote by q_m^+ the conditional probability that one is idle given the other is idle.

$$q_m^+ \equiv Pr(x \text{ idle} \mid y \text{ idle}) = Pr(y \text{ idle} \mid x \text{ idle}) = q + p/m$$

Using this probability, along with Bayes's theorem, we have the following result for the series construction of N_1 and N_2, where N_1 is an n-network and N_2 is an m-network:

$$P_C(N_1 \times N_2) = 1 - \frac{q_n^+ q_m^+}{q_{n \cdot m}^+}(1 - P_C(N_1))(1 - P_C(N_2))$$

This closely resembles the expression derived using Lee's model; the difference is in computing the probability that the link needed between N_1 and N_2 is idle. In Pippenger's model this probability depends on the fact that x and y are assumed to be idle, thus the conditional probability defined above is needed.

For the parallel construction of network N with $X_{d,d}$ we have

$$P_C(X_{d,d} \otimes N \otimes X_{d,d}) = G(P_C(N))$$

where

$$G(z) = \frac{(d-1)q(1-q_m^+(1-z))^2(1-qq_m^+(1-z))^{d-2} + q_m^+ z(1-qq_m^+(1-z))^{d-1}}{dq_{dm}^+}.$$

It is difficult to see the intuition behind this expression, but the key to the derivation is determining the probability that exactly k subnetworks have both the incoming and outgoing links idle that are needed to create a connection from x to y. In Pippenger's model this depends upon the fact that x and y are idle.

There are several additional results of interest on blocking probability. Ikeno [11] used Lee's model to give an upper bound on the crosspoint complexity of a network able to carry a specified amount of traffic with a specified blocking probability. Pippenger [22] tightened this result using his more accurate model of blocking probability and also gave a lower bound for the complexity of a network able to carry connections with a specified average duration, average interarrival time and blocking probability [23].

Valdimarsson [32, 33] has extended the models of Lee and Pippenger for blocking in connectors to the multirate environment by assuming a probability distribution on the weight of a link. A link blocks a connection of weight w if the weight on the link exceeds $1 - w$. In addition, he has used simulation to study blocking in Beneš networks used as distributors. Specifically, this work examines the effects on blocking of fanout size, fanout distribution and routing algorithm.

There is a clear lack of accurate models for evaluating blocking probability in distribution networks. Given the widespread use of models for blocking in connection networks, combined with the increasing interest in multipoint networks, it is natural to consider models for blocking in distributors.

4.2. A Model for Blocking in Distributors

We propose a model based on the following assumptions.

1. Every output is busy with probability p and idle with probability $q = 1 - p$.
2. Every link in stage i is busy with probability $p_i = 1 - (1 - \frac{p_{i+1}}{d})^d$, where for simplicity of exposition we have assumed that all of the switch elements are $d \times d$. We denote the probability a stage i link is idle by $q_i = 1 - p_i$.
3. The conditions of different links are independent.

4. Every busy output of a $d \times d$ switch is connected to one of the d switch inputs at random.

This model resembles Lee's in the sense that it makes the same simplifying assumption of link independence. The second assumption derives from the fact that in a distributor, multiple outputs of a switch can connect to the same input. Each busy output of a $d \times d$ switch independently selects which of the d switch inputs to connect to. A switch input is busy if any of the d switch outputs are connected to it.

A different definition of blocking for distributors is needed. Recall that a distributor must support distribution requests (to connect an idle input to a set of idle outputs) and augmentation requests (to connect an idle output to an existing connection). Clearly a distribution request can consist of a request to connect one output to the input, followed by a sequence of augmentation requests to add the additional outputs to the connection. Thus, it suffices to consider just one type of request consisting of an input and an output pair. The interpretation is that the output requests to join the connection that originates at that input. If the input is idle, then the output should be connected to the input.

We define the *distribution blocking probability* $P_D(N)$ for network N with input x and output y as follows:

$$P_D(N) \equiv Pr(\text{route}(x) \leftrightarrow y \text{ blocked} \mid y \text{ idle}),$$

where "route$(x) \leftrightarrow y$ blocked" means that every path from y to the route originating at x is blocked. In general the route is a tree; y may join the connection at any point in the tree.

Because of the structure of distribution routes, it is less clear that a general expression for blocking probability for the series and parallel constructions is possible. Rather, we will demonstrate this model by applying it to the delta network. In a delta network, there is a unique path between any input/output pair. Thus, a request will succeed (i.e., not block) if and only if, in following the unique path from the output to the input, one encounters idle links up to a particular stage, at which point there is a busy path all the way back to the input.

Figure 4 shows a particular distribution state of the delta network $D_{8,2}$. The unique path from output y_0 to input x_0 is the sequence of links along the top of the network. In the state shown, this unique path consists of an idle link followed by a busy path back to x_0, thus a request to add y_0 to the connection originating at x_0 would succeed. On the other hand, the unique path from y_5 to x_0 contains a busy link which is part of a path originating at input x_3. Thus a request to add y_5 to the connection originating at x_0 would block.

The probability of success in handling a distribution request (x, y) in a delta network is given by

$$p_{k-1}p_{k-2}\cdots p_0(1/d)^{k-1} + q_{k-1}p_{k-2}p_{k-3}\cdots p_0(1/d)^{k-2} + \cdots + q_{k-1}q_{k-2}\cdots q_2p_1p_0(1/d) + q_{k-1}q_{k-2}\cdots q_1p_0 + q_{k-1}q_{k-2}\cdots q_1q_0$$

where $k = \log_d n$ is the number of stages in $D_{n,d}$. The first term indicates success by way of a busy path all the way back to x. This occurs if the needed links in each stage are busy

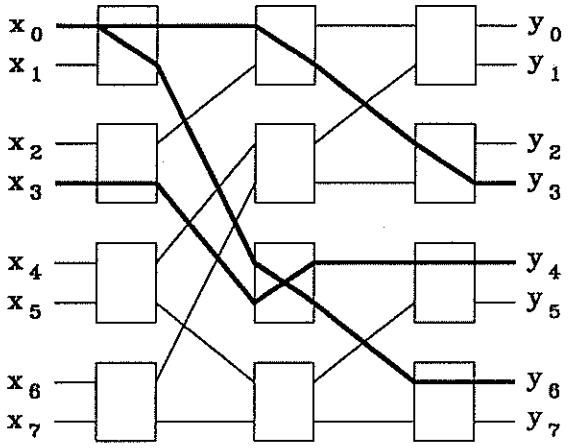

Figure 4: Distribution State of $D_{8,2}$

and connected to one another. The needed link in stage i is busy with probability p_i and is connected to the proper link in the previous stage with probability $1/d$. The second term indicates success by way of an idle link in stage $k-1$ followed by a busy path back to x. The final term indicates success by an idle path all the way from y to x.

The distribution blocking probability is simply the complement of the probability of success.

$$P_D(D_{n,d}) = 1 - \sum_{i=1}^{k-1}(1/d)^i(\prod_{j=0}^{i} p_j)(\prod_{j=i+1}^{k-1} q_j) - \prod_{j=0}^{k-1} q_j$$

Figure 5 shows the blocking probability, computed using this expression, as a function of the carried load for delta networks ranging in size from $n = 16$ to $n = 4096$. All of the networks are constructed from 2×2 switches, thus the larger networks have considerably more stages than the smaller ones. The additional stages contribute to the higher probability of blocking in the larger networks. Keeping in mind that these are conservative calculations of blocking probability, the plot also indicates that the delta network is a poor choice for multipoint connections, particularly when there are many stages; in the three largest networks the blocking probability exceeds 80% for output loading of 40% and higher.

We have begun working on a general expression for the blocking probability for the series connection of two networks. We have also developed an expression for blocking in the Beneš network. This is considerably more complicated than the delta network due to the multiplicity of paths between any input/output pair. The progress made so far indicates that this is a fruitful research area.

4.3. Research Plan

The following plan is proposed for investigating blocking probability in distributors, using the work described in the previous section as the foundation.

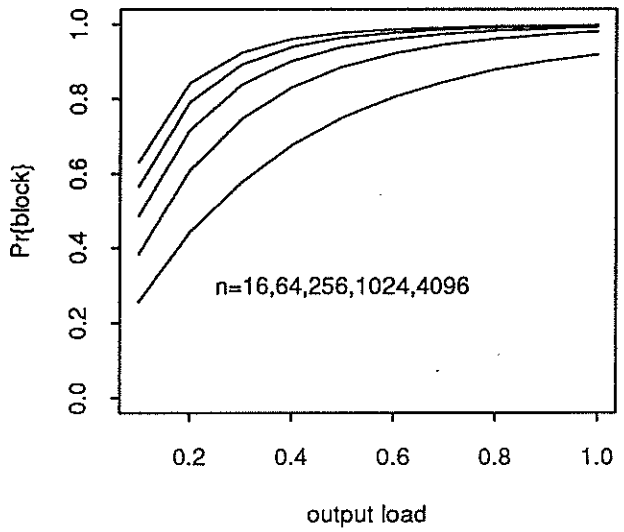

Figure 5: Blocking Probability in Delta Networks

1. Develop a model for blocking in distributors similar to Lee's model for connectors. Apply this model to well known networks, such as the delta and Beneš networks.
2. Improve the accuracy of the model using techniques similar to Pippenger's for connectors.
3. Determine the accuracy of the models by comparing calculations of blocking probability to those measured by simulation.
4. Use the models and simulation to determine the effect of various network and traffic parameters on blocking in distributors. From this study it should be possible to recommend particular networks and operating conditions as being well suited to the multipoint environment.

There are some additional directions which could produce useful results, but which are less straightforward.

1. Extend Pippenger's model to non-uniform networks. Use this to develop an accurate model for blocking in distributors that allows non-uniform networks.
2. Use the model for blocking in distributors to derive an upper bound on network complexity as a function of blocking probability, as Pippenger and Ikeno have done for connectors.

5. Quantitative Comparisons of Network Architectures

As mentioned in the introduction, the classic theoretical measure of network complexity, crosspoint count, is not an accurate reflection of cost for networks constructed using current technology. There are two obvious reasons for this. First, practical networks are constructed of components besides crosspoints, such as buffers and control logic. Second, hardware complexity is only one factor contributing to network cost. The cost of practical networks is also influenced by less easily measured quantities such as expandability and reliability, including ease of fault detection and correction.

We propose to compare network architectures on the basis of hardware complexity, acknowledging that there are other factors neglected by this comparison. Furthermore, to reflect trends in VLSI technology, we will measure hardware complexity by counting the number of pin-limited, transistor-limited integrated circuit chips. Once prototyping has been completed and production is underway, cost is less affected by what is on a chip and more affected by the number of chips.

We will restrict our comparison to architectures for packet switching, widely accepted as the most appropriate technique for high-speed networks carrying diverse traffic, referred to in the literature as broadband integrated services data networks (B-ISDN). We consider only switches that conform to the asynchronous transfer mode (ATM) standard for packet size and header format [37, 18]. The survey paper by Tobagi [26] gives a good overview of ATM architectures.

It is important to factor performance into any cost comparison. We propose to do this by setting a level of performance, defined by such parameters as throughput and acceptable packet loss rate. Each architecture is configured to meet the performance level, then the number of chips required for packaging is counted.

5.1. Related Work

Interconnection networks, particularly the banyan, arise frequently in packet switch architectures. Since at least the early 1980's there has been interest in VLSI implementations of interconnection networks [7, 34, 8, 4, 25]. Franklin et al. [8] have considered packaging of banyan and crossbar networks with the goal of minimizing chip count and delay. More recently, Chien and Oruç [4] studied packaging of rearrangeably and strictly nonblocking networks onto pin-limited chips. The work by Shaikh et al. [25] is notable because it includes performance in the comparison. Specifically, they compare the shufflenet and banyan topologies based on throughput per crosspoint and throughput per chip.

Within the packet switch architecture literature there is recognition of the need to quantify architecture complexity using hardware measures such as transistor count and number of different modules. The following quotes are typical for papers proposing architectures:

> From a hardware implementation perspective, the TBSF architecture is attractive because its implementation is realized by several instances of only two chip components. [27]

> The network is amenable to CMOS VLSI implementations... [9]
>
> [t]he complexity of the Knockout switch results not from the number of gates required in the design of the switch fabric, but rather from pin limitations on the circuit cards and VLSI chips [39]

It has become fairly common for papers on switch architectures to include transistor counts, however this can be misleading, as it does not take into consideration the pin limitations on VLSI chip packaging. For many designs, it is the interconnect requirements that constrain the packaging, not the transistor density of a chip. Furthermore, the survey papers that consider a variety of architectures tend to make only qualitative comparisons. One exception is the survey by Oie et al. [19] that compares the performance of a number of ATM switch architectures, but without any cost comparison.

We propose to combine considerations of hardware complexity and performance for a broad range of architectures using a complexity measure that reasonably reflects cost under current technology.

5.2. Summary of Progress

I have completed a comparison of eight ATM switching architectures based on the number of pin-limited chips needed to realize each architecture [35]. In order to draw attention to the differences between architectures, we consider only the *switching fabric*, where the actual routing takes place. The input and output circuitry connecting the switching fabric to the external links is likely to be fairly similar across various architectures. In addition, we packaged the networks with the goal of minimizing the number of chips, perhaps at the expense of reliability and expandability. Thus, while a commercial venture may choose an alternate packaging than the one we assumed in the study, the alternate scheme will do no better than ours on the basis on chip count.

Figure 6 is the highlight of the study. It gives the chip count per port for three network sizes ($n = 16, 256, 4096$), and two chip dimensions ($p = 32, 64$, corresponding to 64 and 128 total data pins). Each curve is labeled by the architecture under consideration. The x axis is the external link data rate; the networks are configured to keep up with this data rate. The y axis is the chip count normalized by the number of inputs; note that the scale is logarithmic. There are significant differences between the architectures. The Knockout [39] (labeled "k") has by far the highest chip count for $n = 256$ and $n = 4096$. The Sunshine [10] (labeled "s") also tends to have a relatively high chip count for all three network sizes. The more cost-competitive networks include the second generation Broadcast Packet Switch [29] (labeled "2"), and the Tandem Banyan [28] (labeled "t"). The reader is referred to the technical report for additional detail.

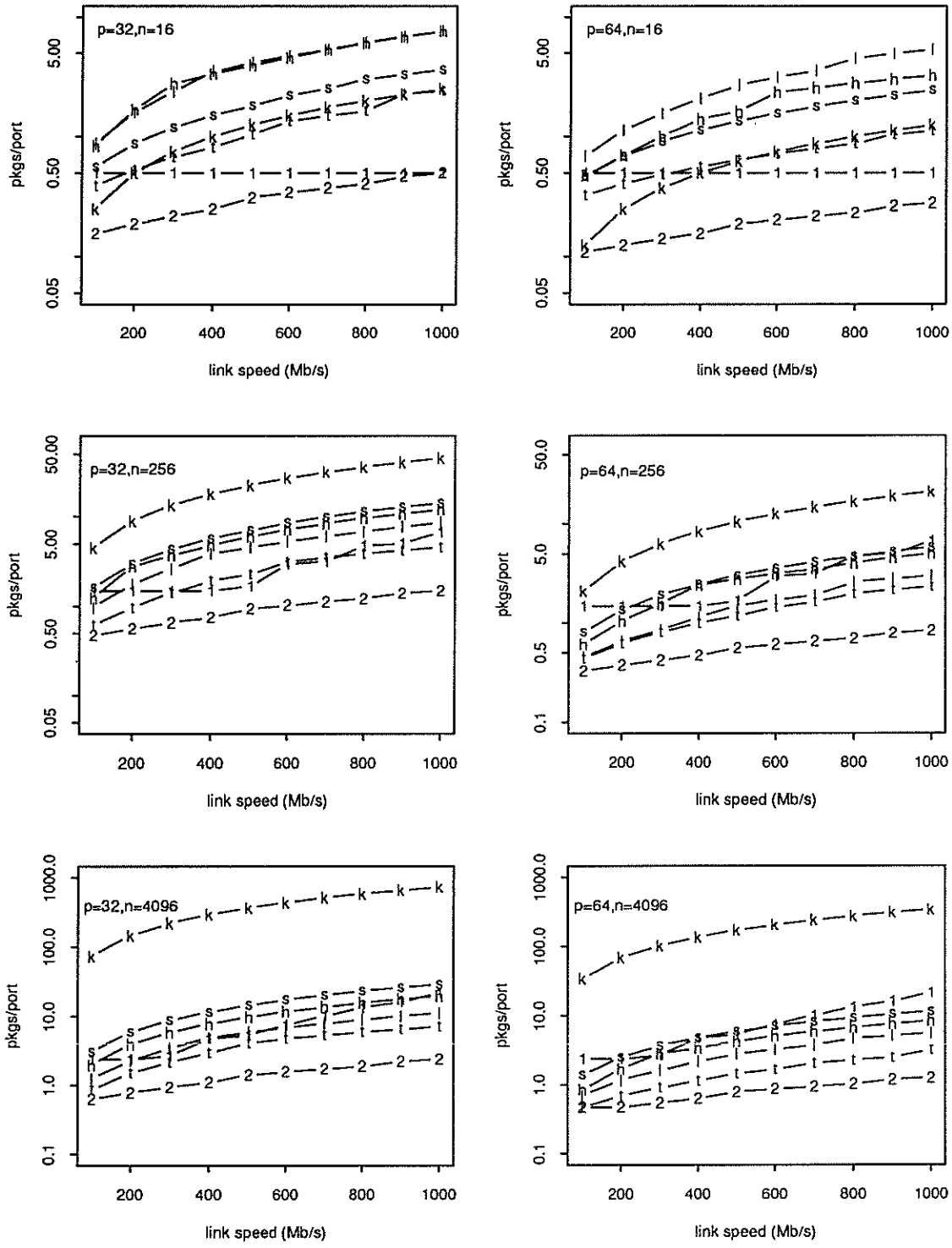

Figure 6: Network Comparison

5.3. Research Plan

The following plan is proposed for comparing ATM architectures.

1. Compare ATM architectures for point-to-point connections based on chip count of the switching fabric. This involves choosing the architectures, determining a packaging strategy for each and developing equations to express chip count as a function of performance requirements.
2. Identify the regions in which particular architectures have the lowest (or highest) chip count.
3. Perform a similar study for ATM architectures designed for multipoint connections. This should indicate the issues that influence hardware complexity that are specific to practical multipoint switch design.

6. Summary

In summary, it is expected that this work will provide a needed theoretical framework for decisions in the design of practical switching networks. Specific contributions are expected to include:

- New constructions of nonblocking multirate distributors that are competitive in terms of crosspoint complexity and routing algorithm runtime. Such constructions can be key components of packet switching networks for the multipoint environment.
- Models for estimating blocking probability in distributors. Along with the models, simulation will be used to explore the factors that influence blocking in distributors. The potential exists for significant savings in complexity by tolerating some low level of blocking probability.
- Quantitative comparisons of ATM architectures for point-to-point and multipoint environments. Given the abundance of architectures being proposed, it is vital to compare them using quantitative measures.

References

[1] H. Ahmadi and W. E. Denzel. A survey of modern high-performance switching techniques. *IEEE J. Selected Areas in Communications*, 7(7):1091–1103, September 1989.

[2] V. E. Beneš. *Mathematical theory of connecting networks and telephone traffic.* Academic Press, New York, 1965.

[3] D. B. Cantor. On non-blocking switching networks. *Networks*, 1:367–377, 1971.

[4] M. Chien and A. Oruç. Optimal nonblocking networks with pin constraints. In *International Conf. on Parallel Processing*, pages 422–425, 1991.

[5] S. Chung and K. Ross. On nonblocking multirate interconnection networks. *SIAM Journal on Computing*, 20(4):726–736, August 1991.

[6] C. Clos. A study of non-blocking switching networks. *Bell Systems Tech. Journal*, 32:406–424, 1953.

[7] M. Franklin. VLSI performance comparison of banyan and crossbar communications networks. *IEEE Transactions on Computers*, C-30, April 1981.

[8] M. Franklin, D. Wann, and W. Thomas. Pin limitations and partitioning of VLSI interconnection networks. *IEEE Transactions on Computers*, C-31(11):1109–1116, November 1982.

[9] J. Giacopelli, J. Hickey, W. Marcus, W. Sincoskie, and M. Littlewood. Sunshine: A high-performance self-routing broadband packet switch architecture. *IEEE J. Selected Areas in Communications*, 9(8):1289–1298, October 1991.

[10] J. N. Giacopelli, W. D. Sincoskie, and M. Littlewood. Sunshine: A high performance self routing broadband packet switch architecture. In *Proc. of the International Switching Symposium*, 1990.

[11] N. Ikeno. A limit on crosspoint number. *IRE Trans. on Inform. Theory (Special Supplement)*, IT-5:187–196, May 1959.

[12] D. H. Lawrie. Access and alignment of data in an array processor. *IEEE Transactions on Computers*, C-24(12):1145–1155, December 1975.

[13] C. Y. Lee. Analysis of switching networks. *Bell Systems Tech. Journal*, 34:1287–1315, 1955.

[14] G. J. Lipovski. The architecture of a large associative processor. In *Proc. AFIPS Spring Joint Computer Conference*, 1970.

[15] R. Melen and J. S. Turner. Nonblocking multirate networks. *SIAM Journal on Computing*, April 1989.

[16] R. Melen and J. S. Turner. Nonblocking networks for fast packet switching. In *Proc. of Infocom 89*, April 1989.

[17] R. Melen and J. S. Turner. Nonblocking multirate distribution networks. In *Proc. of Infocom 90*, 1990.

[18] S. Minzer. Broadband ISDN and asynchronous transfer mode (ATM). *IEEE Communications Magazine*, 27(9):17–24, September 1989.

[19] Y. Oie, T. Suda, M. Murata, D. Kolson, and H. Miyahara. Survey of switching techniques in high-speed networks and their performance. In *Proc. of Infocom 90*, pages 1242–1251, 1990.

[20] J.K. Patel. Performance of processor-memory interconnections for multiprocessors. *IEEE Transactions on Computers*, pages 301–310, October 1981.

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

[22] N. Pippenger. On crossbar switching networks. *IEEE Transactions on Communications*, 23(6):646–659, June 1975.

[23] N. Pippenger. The complexity of seldom blocking networks. *Proc. of Conf. Communications*, pages (7–8)–(7–11), 1976.

[24] N. Pippenger. Communication networks. In Jan van Leeuwen, editor, *Handbook of Theoretical Computer Science*, chapter 15, pages 805–833. The MIT Press, Cambridge, MA, 1990.

[25] S. Shaikh, M. Schwartz, and T. Szymanski. A comparison of the shufflenet and the banyan topologies for broadband packet switches. In *Proc. of Infocom 90*, pages 1260–1267, 1990.

[26] F. Tobagi. Fast packet switch architectures for broadband integrated services digital networks. *Proceedings of the IEEE*, 78(1):133–167, January 1990.

[27] F. Tobagi, T. Kwok, and F. Chiussi. Architecture, performance, and implementation of the Tandem Banyan fast packet switch. *IEEE J. Selected Areas in Communications*, pages 133–167, January 1990.

[28] F. A. Tobagi and T. C. Kwok. The Tandem Banyan switching fabric: A simple high-performance fast packet switch. In *Proc. of Infocom 91*, April 1991.

[29] J. S. Turner. Design of a broadcast packet network. *IEEE Transactions on Communications*, June 1988.

[30] J. S. Turner. A practical multicast switching system for ATM networks. 1990.

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

[32] E. Valdimarsson. Blocking in multirate networks. Master's thesis, Washington University, May 1990.

[33] E. Valdimarsson. Blocking in multirate networks. In *Proc. of Infocom 91*, pages 579–588, 1991.

[34] D. S. Wise. Compact layouts of banyan/FFT networks. In H. T. Kung, et al., editor, *VLSI Systems and Computations*. Computer Science Press, Rockville, MD, 1981.

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

[36] E. E. Witte. The Clos network as a multirate distributor with a greedy routing algorithm. Technical report, Washington University Computer Science Department, WUCS-92-13.

[37] CCITT Study Group XVIII. Revised draft recommendation I.211, May 1990.

[38] Y. Yang and G. M. Masson. Nonblocking broadcast switching networks. *IEEE Transactions on Computers*, 40(9):1005–1015, Sept. 1991.

[39] Y. S. Yeh, M. G. Hluchyj, and A. S. Acampora. The Knockout switch: A simple modular architectur for high performance packet switching. In *Proc. of the International Switching Symposium*, 1987.