Reconfigurable Fault Tolerant Networks for Fast Packet Switching

Shih-Chian Yang and John A Silvester

CEng Technical Report 90-28

(Revised May 21, 1991)

Department of Electrical Engineering - Systems
University of Southern California
Los Angeles, CA. 90089-2562
(213) 740-4579

November 1990

Abstract

Fast packet switching is intrinsic to many applications that use fiber optic technology. As the technology is pushed to improve speed, network reliability becomes a key issue. The design constraints resulting from the use of fast packet switching that impact fault tolerant network design are carefully studied. A novel approach for network reconfiguration is proposed.

An abstract replacement model to characterize the proposed fault tolerant network is presented. Reconfigurable fault tolerant network problems are then transformed into well known assignment problems. Many interesting properties of failure behavior are shown based on this model. More reliable networks with less hardware overhead can be designed with the techniques shown in this paper. Two practical designs based on feasible technology are presented. A significant reliability improvement is achieved while maintaining full bandwidth up to a tolerable level of failures, with a relatively smaller number of spare switches.

Index Terms - broadband ISDN, ATM switch, WDM fast packet switching, multichannel fiber network, fault tolerance, network reliability, interconnection network.

1 Introduction

The tremendous bandwidth of fiber optics has motivated much research into Broadband ISDN in recent years. One of the key problems is how to switch the high bandwidth channels. Photonic switching may eventually provide a solution, but current photonic logic cannot support the complexity for arbitrary switching (at high speed). Thus, in the near term, the fundamental switching elements will continue to be electronic. This is not such a drawback as it seems since almost all devices generate and process data in electronic form.

Many proposed Asynchronous Transfer Mode (ATM) switch designs use a bit parallel electronic switch fabric, e.g. Multicast Switch Fabric [21], Starlite [11], 3-phase algorithm [12], Knockout switch [7] etc. One difficulty with the use of a bit parallel electronic switch fabric design is expandability. As the switch size increases, the length of some of the (copper) wire connections becomes long making synchronization at high speed very difficult. The growable packet switch architecture in [8] approaches this problem by a generalization of the knockout principle. It relieves the problem to some degree by a slight increase in packet loss.

Use of completely photonic switching is still restricted to circuit switching applications [20] and mix of photonics and electronics is necessary in ATM switch design. The photonic knockout switch [6] is an attempt to do this. One drawback is that it requires a large number of receivers to collect signals from all sources and synchronization of the retuning wavelengths among many nodes is almost impossible to achieve efficiently.

Another topic of recent interest is how to provide fast packet switching functionality in a distributed multiuser environment. A multichannel multihop lightwave network that uses Wavelength Division Multiplexing (WDM) techniques is proposed in [1]. Distributed electronic switches interconnected with multichannel fiber cable provide high speed switching functions. The switch size problems noted above are alleviated since the switch elements are interconnected with *serial* channels. Although the bandwidth per user node is relatively low the switch nodes still require high bandwidth due to switch sharing and contention buffering among nodes.

The WDM approach described above does not require that the switches be distributed. Collecting these switches to a central location but still interconnecting them with fiber channels, provides an attractive solution to switch fabric design [24], where the expandability problems encountered in a totally electronic switch are reduced. Since all switches are located in a single location, cable cost becomes insignificant and the interconnection can be implemented with multiple channels on a single fiber or several single mode fibers.

There are many techniques to speed up electronic switches, but in all cases the common goal is to push the boundaries of electronic technology which leads to reduction of component reliability. This situation is exacerbated when the network or switch fabric size becomes very large. Similar problems have been encountered in the design of fault tolerant interconnection networks [2]. For interconnection networks, we are often more concerned with connectivity than network throughput, e.g. ESC [3]. To maintain high bandwidth under failure, replicate copies of the network are usually required [23], e.g. INDRA [18], F-net [4], ACN [19] etc.

There are two problems for these approaches. First, the hardware overhead (in terms of active device) for these designs is typically high due to static link restrictions. Second, the functional topology changes when failures occur and therefore, a complex routing scheme is required. There is little time for packet processing at line rates of gigabits per second. Therefore, it is highly desirable to maintain the functional topology and minimize the packet processing time needed for routing. A fault tolerant network is said to be in the *Topology Invariant (TI) mode* if the functional topology is not changed with the reconfiguration induced by any fault tolerable failure. Note that all of the fault tolerant interconnection networks described above can run in TI mode. The main difference is that the reliability is reduced since some previously fault tolerable failure sets are no longer fault tolerable.

Recent progress in tunable laser diodes [16] and tunable optical filters [13] provides rich reconfiguration capabilities among switches. It is possible to reconfigure the network or switch fabric topology by retuning the laser diodes to another set of channels that are received by other switches. Similar functions are achievable with tunable receivers. Although the cost of tunable laser diodes is still very high, we expect that as the technology matures, it will be possible to provide redundant links at low cost. In the interim, non-tunable laser diodes or copper wires with simple MUX and DMUX circuits can provide link redundancy without too much impact on system cost as discussed in section 6.

In this paper, we consider switches designed according to the general principles outlined above and focus on associated reliability problems. A model of replacement (to recover from failures) and its applicability to fault tolerance are described in section 3. The maximum fault tolerance, an important reliability index, is found in section 4. Section 5 addresses the relationship between our abstract model and a real switch design and shows hardware requirements necessary to achieve reconfiguration. Two interesting implementations are considered in section 6. The maximum fault tolerance *versus* the hardware overhead and an approximation of the reliability of these networks are also discussed.

2 Replacement Model

In general, a network is reliable if its operational criteria are satisfied for a given failure set. When a failure occurs, paths through the failed component are destroyed and alternatives are required. Multiple paths are typically provided between all source and destination pairs to accommodate various failures. A packet passing through a switch needs to select a live path toward its destination. Routing schemes need to be properly designed to facilitate recovery from different failures. To fully utilize the set of possible paths for all source and destination pairs, sophisticated routing schemes are required.

However, in fast packet switching applications, packet decoding and routing needs to be as fast as possible due to the so-called electronic bottleneck [10]. Simple self routing schemes are mandatory to simplify switching element design and therefore, to achieve high speed electronic devices. It is better to run in the TI mode, i.e. when a failure occurs, a different set of switch elements and interconnection links are selected such that packets traverse through the reconfigured network while maintaining the same operational topology. Thus the routing scheme can be preserved. More link redundancy is required to achieve the same degree of fault tolerance, however, since many redundant paths are not usable. As mentioned in section 1, there are several ways to interconnect system components with a rich set of redundant links. Therefore, our focus is the fault tolerance of the proposed replacement model for the networks running in TI mode.

For each failed component, a live component needs to assume all of its operational functions or a block of live components assumes the operations of another block of components that contains the failure. A set of replacement components assumes all the operational functions of another set of components if they have the same internal structure and external interconnections. Replacement can be for any size sets as long as the above conditions are preserved. The replacement model developed in this paper does not assume any particular structure for a replacement block except in the final designs. However, for simplicity, a single switch replacement is used to describe the replacement model (without loss of generality).

> *Definition* 1: A switch (module) x is *replaceable* by a switch (module) y if when x fails, y can assume all operational functions that were performed by x. □

Definition 1 is quite generic. We start with this definition and develop a replacement model which we then apply to various practical network designs. A simple replacement is shown in Figure 1. When the switch x fails, the reconfigurable links of y

(dotted lines) replace the operational links (solid lines) and allow it to assume the role of the switch x (provided switches $a1$, $a2$, $b1$ and $b2$ are operational). A more complex example is shown in Figure 8. Switch 0 is replaceable by switch 1 and a replacement link between them is shown by the arrow line. A *replacement graph* that contains all available switches and their replacements can be derived from the original reconfigurable network. The replacement graph for a single stage of the network in Figure 8 is drawn in Figure 2. A replacement graph characterizes all possible replacements during the fault recovery process. On the other hand, reconfigurable links that do not support the replacement graph are useless for fault recovery.

Note that a switch is always replaceable by itself. However, the replacement relationship is not necessarily reflexive due to lacking of symmetry between the operational links of x and the reconfigurable links of y. A failed switch needs a fault free spare. If no spares can *directly* replace a failed switch, a sequence of replacements is needed until a spare is reached. A better way to look at this property is to represent the network by a logical replacement graph.

We define the available switch set, Y, the (functionally required) operational switch set, X and the set of spares S. Thus, in fig 2, X is the set of logical functions to be implemented and Y is the set of available switches including S. Clearly:

$$\begin{aligned} S &= Y - X \\ X &\subset Y \end{aligned}$$

Definition 2: A Replacement Bipartite Graph (RBG) for the operational set X and available set Y is defined by the edge set E:

$$E = \{(y,\ x)\ |x \in X \textit{ is replaceable by } y \in Y\}$$

This graph is referred to as either the RBG from X to Y or the RBG from Y to X as needed. A link in E is called a *Replacement Link*. □

A RBG is a different way to represent the replacement graph and hence, it can be used to characterize the original reconfigurable fault tolerant network. As shown in Figure 2, the replacement graph corresponding to Figure 8 is transformed into the RBG. Switches 0 to 7 are operational switches while switch s is a spare. Each switch in Y has at most two replacement links. All switches except 7 have two incoming replacement links.

The problem of fault tolerance is then equivalent to the *Matching Problem* in graph theory [9]. A similar bipartite graph model has been used in the design of reconfigurable array processors [14], where the special case of a two dimensional array

processor system was considered. A replacement algorithm for this design was shown to be NP-complete and a heuristic algorithm was given. In this paper, we consider the case of independent replacement model and focus on design and reliability issues. With the RBG model, the design of reconfigurable fault tolerant networks is transformed into two subproblems:

1. Replacement design: Design a feasible replacement topology to achieve certain performance indices of interest, for example: maximum fault tolerance, reliability, etc, for a given hardware overhead.

2. Link Assignment: Properly design the interconnections based on feasible technology such that the replacement topology is achieved.

A precise description of failure characteristics is highly desirable in analyzing the fault tolerance of a network. We define failure sets and show the relationship between fault tolerance and the bipartite graph described above.

Definition 3: A *Failure Set* F is the subset of Y comprised of all failed switches. □

A failed switch in set Y requires a replacement to assume its functions, thus, to tolerate a failure F, there must be a one to one mapping from the live switches in Y to the switches in X, i.e. a complete matching. Let the mapping M represent all the links in a RBG.

Definition 4: A mapping $M : X \rightarrow Y$ is a *complete matching* if and only if

1. It is a one to one mapping.
2. For each $x \in X$, there is a y that it maps to. □

It is easy to see how the above definitions allow us to characterize the fault tolerance of a reconfigurable network. The resulting one to one mapping defines the reconfiguration.

Lemma 1 : A failure set F is Fault Tolerable if and only if there exists a complete matching of the RBG from X to $Y - F$. □

Proof:

If there exists a complete matching, then there is a replacement link $(y\ ,x)$ for every $x \in X$ with distinct $y \in Y - F$. Replace x with y and the reconfigured network is operational.

If the network is fault tolerable, then there exists a set $Z \subset Y - F$ such that a replacement exists for all switches in X. Map X to Z accordingly and it is a complete matching. QED

Based on the RBG in Figure 2., a fault tolerable failure set $\{1\}$ and its corresponding complete matching are shown in Figure 3.

3 Maximum Fault Tolerance of RBG

The complete matching problem for a bipartite graph has been extensively studied in graph theory. It provides an effective mathematical tool for analyzing the replacement problem. The conditions for existence of a complete matching (in our context) are repeated here without proof.

Definition 5: Let A be a subset of Y, then the mapping set $X(A)$ is defined by:

$$X(A) = \{x | \exists y \in A \textit{ such that } x \textit{ is replaceable by } y \}$$

Similarly, if A is a subset of X, then the mapping set $Y(A)$ is defined by:

$$Y(A) = \{y | \exists x \in A \textit{ such that } x \textit{ is replaceable by } y \}$$

Examples from Figure 2 are:

$$\begin{aligned} Y(\{0,\ 1\}) &= \{0,\ 1,\ 2\} \\ X(\{5,\ 6\}) &= \{4,\ 5,\ 6\} \end{aligned}$$

Complete Matching Theorem: [9] A bipartite graph has a complete matching from X to Y if and only if

$$|Y(A)| \geq |A| \qquad \forall A \subset X$$

□

An important system performance index is maximum fault tolerance, since this specifies the maximum number of failures for which a network is guaranteed to be operational. Furthermore, small failure sets are the dominant terms in calculating system reliability since the switch failure probability is very small.

Definition 6: The *Maximum Fault Tolerance* of a switch set is t if and only if

1. A failure set F is fault tolerable $\forall |F| \leq t$.
2. $\exists F; |F| = t + 1$ which is not fault tolerable. □

The condition for maximum fault tolerance of a RBG can be easily derived from the Complete matching theorem. It is essentially an extension of the Complete matching theorem.

Theorem 1: The maximum fault tolerance of a RBG is

$$t = \min_{A \subset X}(|Y(A)| - |A|)$$

□

Proof:

Let

$$a = \min_{A \subset X}(|Y(A)| - |A|)$$

There exists a nonempty subset A_1, such that

$$\begin{aligned} |Y(A_1)| - |A_1| &= a \\ |Y(A_1)| &\geq a + 1 \end{aligned}$$

Hence, there exists a failure set $F \subset Y(A_1)$ of size $|F| = a + 1$. For the new $RBG(Y - F,\ X,\ E)$,

$$\begin{aligned} |(Y - F)(A_1)| - |A_1| &= |Y(A_1)| - |F| - |A_1| \\ &= -1 \end{aligned}$$

which violates the Complete matching theorem. Therefore, by definition of maximum fault tolerance:

$$t \leq a$$

Consider an arbitrary failure set F such that

$$|F| \leq a$$

then,

$$\begin{aligned} |Y(A)| - |Y(A) \cap F| &\geq |Y(A)| - a \\ &\geq |A| \qquad\qquad \forall A \subset X \end{aligned}$$

From the Complete matching theorem, there is a complete matching from X to $Y - F$. This concludes the theorem. QED

It is interesting to determine the maximum achievable fault tolerance for a given design. The number of replacement links from a switch is a measure of the reconfiguration hardware overhead. Assume that the replacement capability of any switch is restricted to be r_n, i.e. it can replace at most r_n switches other than itself. Another important design parameter is number of spare switches, s, which defines the cost overhead in terms of extra switches.

Corollary 1 : The maximum fault tolerance is:

$$t \leq \min(s, \left\lfloor \frac{r_n(s + |X|)}{|X|} \right\rfloor) \tag{1}$$

□

Proof :

According to Theorem 1 with $A = X$,

$$\begin{aligned} Y(X) &= Y \\ t &\leq |Y(X)| - |X| \\ &= s \end{aligned}$$

Since spares do not have links to themselves, there are at most

$$r_n s + (r_n + 1)|X|$$

replacement links from set the Y. There exists at least one switch $x \in X$ such that

$$Y(\{x\}) \leq \left\lfloor \frac{r_n s + (r_n + 1)|X|}{|X|} \right\rfloor$$

Then,

$$\begin{aligned} |Y(\{x\})| - |\{x\}| &\leq \left\lfloor \frac{r_n s + (r_n + 1)|X|}{|X|} \right\rfloor - 1 \\ &= \left\lfloor \frac{r_n(s + |X|)}{|X|} \right\rfloor \end{aligned}$$

QED

Although not all designs have a maximum fault tolerance as in Equation 1, it is not difficult to design one for a given set of design constraints. In section 6, a design, SARI/k, is shown that achieves this maximum fault tolerance with $k = r_n$. Usually, maximum fault tolerance implies that failures are uniformly tolerated and a regular fault tolerant network is provided. In addition to providing higher fault tolerance, a regular structure is easier to realize.

4 Properties of Fault Tolerable Failure Set

A complete matching is a one to one mapping from X to Y such that every switch in X is mapped to a switch in Y. It is equivalent to find a matching of size $|X|$ from the set Y to X. Hence, a failure set F is fault tolerable if a matching of size $|X|$ from $Y - F$ to X can be found. The condition for fault tolerance from the point of view of Y is given in following theorem.

Theorem 2:

A failure set F of size $|F| \leq s$ is fault tolerable if and only if

$$\max_{A \subset Y-F}(|A| - |X(A)|) = s - |F| \tag{2}$$

□

Proof :

Consider a failure set F which is fault tolerable. Then $|F| \leq s$. Consider a pseudo switch set P_x of size $s - |F|$ and define:

$$X_p = X \cup P_x$$

Add a replacement link from each switch in $Y - F$ to each switch in P_x such that any pseudo switch is replaceable by any switch in $Y - F$ as shown in Figure 4. Clearly, set Y and the pseudo set are fully connected.

Since $|F|$ is fault tolerable, according to lemma 1, there is a complete matching from X to $Y-F$. For switches in $Y-F$, there are $s-|F|$ switches which are not matched. There always exists a complete matching from P_x to these unmatched switches since they are fully connected. Therefore, there is a complete matching from $Y - F$ to X_p.

For an arbitrary subset $A \subset Y - F$, define the X mapping for the pseudo switch set X_p, similar to $X(A)$, as

$$X_p(A) = \{x | \exists y \in A \textit{ such that } x \textit{ is replaceable by } y \}$$

Apply the Complete matching theorem to $X_p(A)$ and we have

$$|A| - |X(A)| \leq s - |F|$$

Since equality holds for the set $Y - F$, we conclude Equation 2.

Consider a failure set F of size $|F| \leq s$ such that Equation 2 holds. Construct X_p as before. Consider an arbitrary set $A \subset Y - F$. Then

$$|X_p(A)| - |A| \geq 0$$

According to the Complete matching theorem, there is a complete matching from $Y - F$ to X_p. Since the number of replacement links matched to P_x is $s - |F|$, there are $|X|$ replacement links matched to the set X. It is a complete matching from X to $Y - F$. From lemma 1, F is fault tolerable.

QED

This theorem is very similar to Theorem 1. They reveal the same property from two sides of the RBG.

Corollary 2:

For a failure set F which is fault tolerable, if there is a subset $A \subset Y - F$ and its subset $A_1 \subset A$ such that

$$\begin{aligned} |A| - |X(A)| &= s - |F| \\ |X(A - A_1)| &< |X(A)| \end{aligned}$$

then $F \cup A_1$ is not fault tolerable. □

Proof :

$$\begin{aligned} |A - A_1| - |X(A - A_1)| &> |A| - |A_1| - |X(A)| \\ &= s - (|F| + |A_1|) \end{aligned}$$

QED

It is interesting to note that removing even a single switch from $X(A)$ due to an additional failure (set $A_1 \subset A$) will fail the network. Some very interesting properties of the subset A are shown as follows.

Definition 7 :

A set $A_F \subset Y - F$ for a fault tolerable failure set F is *maximal* if

$$|A_F| - |X(A_F)| = s - |F|$$

□

From Theorem 2, maximal sets have the maximum possible size difference and all other sets have smaller size difference. It is obvious that $Y - F$ is maximal for any fault tolerable set.

Assume that there is a subset of $X(A_F)$ mapped onto a same size subset in A_F. It is easy to see that, by removing them, the remaining set is still a maximal set. On

the other hand, any failure of the switches in the removed subset fails the network, since there are not enough switches for the one to one mapping. If we keep removing these subsets from the maximal set, until no further switches can be removed, the result is a minimal maximal set that has very interesting properties.

Definition 8:

A maximal set M_F for a fault tolerable failure set F is a *minimal maximal set* if

$$|Y(A) \cap M_F| - |A| > 0 \qquad \forall A \subset X(M_F)$$

□

Theorem 3 :

A minimal maximal set M_F for a fault tolerable failure set F is unique and is a subset of any maximal set. □

Proof :

Let A_F be another minimal maximal set and

$$\begin{aligned}
D_1 &= M_F - M_F \cap A_F \\
D_2 &= M_F \cap A_F \\
D_3 &= A_F - M_F \cap A_F \\
D_4 &= Y - F - M_F \cup A_F \\
E_1 &= X(M_F) - X(M_F) \cap X(A_F) \\
E_2 &= X(M_F) \cap X(A_F) \\
E_3 &= X(A_F) - X(M_F) \cap X(A_F) \\
E_4 &= X - X(M_F) \cup X(A_F)
\end{aligned}$$

According to the definition of maximal set:

$$\begin{aligned}
|D_4| + |D_1| &= |E_4| + |E_1| \\
|D_4| + |D_3| &= |E_4| + |E_3|
\end{aligned}$$

Since $Y(E_4) = D_4$, $|D_4| \geq |E_4|$. Therefore, $|D_1| \leq |E_1|$, $|D_3| \leq |E_3|$. Since $(Y(E_1) \cap M_F) \subset D_1$ and $(Y(E_3) \cap A_F) \subset D_3$, E_1, E_3, D_1, D_3 are all empty sets and hence $A_F = M_F$.

If A_F is any maximal set, then as in the above proof, D_1 and E_1 are empty sets. Hence $M_F \subset A_F$. QED

Theorem 4 :

Let M_F be a minimal maximal set for a fault tolerable failure set F.

$F \cup \{a\}$ is fault tolerable $\forall a \in M_F$.

$F \cup \{a\}$ is not fault tolerable $\forall a \in Y - F - M_F$. □

Proof :

By definition of a minimal maximal set,

$$|Y(A) \cap M_F| - |A| > 0 \qquad \forall A \subset X(M_F)$$

Assume that a switch $a \in M_F$ failed. It reduces the difference in the above equation by at most 1. Hence,

$$|Y(A)| - |A| \geq 0 \qquad \forall A \subset X$$

Therefore, $F \cup \{a\}$ is fault tolerable.

Since

$$|Y(X - X(M_F))| = |X - X(M_F)|$$

any failure in $Y - F - M_F$ will fail the network. QED

An example of a minimal maximal set can be found in Figure 5. The minimal maximal set of a fault tolerable failure set $F = \{3,\ 4,\ 5\}$ is

$$M_F = \{0,\ 1,\ 7,\ s0,\ s2,\ s3\}$$

One more failure in M_F will not fail the network. Note that

$$X_F(\{2,\ 3,\ 4\}) = \{2,\ 4,\ s1\}$$

Any failure of $\{2,\ 4,\ s1\}$ will fail the network.

5 Link Requirements

The replacement model has been shown to be an effective tool to analyze the relationship between replacement links and spares. We must now determine how many reconfigurable links are required and how should they be arranged to provide the desired replacement links for a particular class of operational topology designs. If the number of links can be minimized and organized so that they are similar to the

physical interconnections, more useful replacement links are available leading to a more reliable network. Operational links are functional interconnections for both the non-failure mode or failure mode of operation and, therefore, are the basis for reconfiguration.

Consider a network that consists of interconnected 2×2 cross bar switches. There are two innodes (upper and lower) and two outnodes (upper and lower) for each switch. Let A be a set of operational switches in X.

Definition 9:

1. *Backward upper node set*: $B_{bu}(A)$ = Set of outnodes that link to an upper innode of a switch in A with an operational link during non-failure mode.

2. *Backward lower node set*: $B_{bd}(A)$ = Set of outnodes that link to an lower innode of a switch in A with an operational link during non-failure mode.

3. *Forward upper node set*: $B_{fu}(A)$ = Set of innodes that link to an upper outnode of a switch in A with an operational link during non-failure mode.

4. *Forward lower node set*: $B_{fd}(A)$ = Set of innodes that link to an lower outnode of a switch in A with an operational link during non-failure mode. □

For each switch, there are switches that are able to replace it and switches that are potentially replaced by it. Replacement sets are defined as follows:

Definition 10:

1. *Previous replacement set*: $R_p(A)$ = Set of switches that can replace a switch in A.

2. *Next replacement set*: $R_n(A)$ = Set of switches that are replaceable by a switch in A. □

Examples can be found from Figure 2:

$$R_p(\{0,\ 1\}) = \{1,\ 2\}$$
$$R_n(\{0,\ 1\}) = \{0,\ 7\}$$

Since each innode or outnode of a switch is replaced by the corresponding node of its replacement switch, we can apply the above definition to nodes without ambiguity. For example: if A is a set of upper innodes, $R_p(A)$ and $R_n(A)$ are the corresponding upper innodes of the replacement switch sets.

Recall that the replacement capability of a single switch is r_n which is the maximum number of switches that it can replace. Similarly, the maximum size of the previous set for a switch is r_p. If a switch x is replaceable by a switch y, a minimum requirement is that switch y must have reconfigurable links to all innodes and outnodes operationally connected to switch x. However, if a switch a which was connected to x fails, it is replaced by another switch and the original operational link is reconfigured to another reconfigurable link. To guarantee that x is still replaceable by y under other reconfigurable failures, more reconfigurable links are required, in particular y must have a link to a's replacement.

Theorem 5: For a single switch set $A = \{a\}$, link requirements for its innodes and outnodes are

$$L_{ij}(A) = B_{ij}(A) \cup R_p(B_{ij}(A)) \cup B_{ij}(R_n(A)) \cup R_p(B_{ij}(R_n(A)))$$
$$\forall i \in \{b,\ f\},\ j \in \{u,\ d\}$$

□

Proof: In Figure 6, consider all possible failures that may cause reconfiguration to A.

1. Normal mode of operation: All nodes in the set : $B_{ij}(A)$ require an operational link to A.
2. Switches in $B_{ij}(A)$ failed: Switches in the previous replacement set: $R_p(B_{ij}(A))$ are required to replace the failed switches.
3. Switches in the next replacement set failed: A may be needed to replace one of the failed switches. All switches that have a link, operational or reconfigurable, to that replaced switch needs to be linked to switch A. With the same reasoning as in the previous two cases, nodes in $B_{ij}(R_n(A))$ and $R_p(B_{ij}(R_n(A)))$ require a reconfigurable link to A.

Other failures will not affect the link requirement set, hence, we conclude the theorem. QED

Base on the Omega network in Figure 7, the link requirements for the replacement graph in Figure 8 can be derived. Dotted links are the reconfigurable links that support the desired replacement graph. It is interesting to find upper and lower bounds on the link requirements in terms of r_n and r_p. They are given by following corollary.

Corollary 3: The number of links required for a node of a switch is

$$r_n + 1 \leq l_{req} \leq (r_p + 1)(r_n + 1)$$

Proof: □

Let us evaluate the size of $L_{bu}(\{a\})$ for a single switch a.

$$\begin{aligned} |B_{bu}(A)| &= 1 \\ |R_p(B_{bu}(A))| &\leq r_p \\ |B_{bu}(R_n(A))| &\leq r_n \\ |R_p(B_{bu}(R_n(A)))| &\leq r_p r_n \\ |A \cup R_n(A)| &\geq r_n + 1 \end{aligned}$$

Therefore,

$$r_n + 1 \leq |L_{bu}(A)| \leq (r_p + 1)(r_n + 1)$$

Similarly, bounds for $L_{bd}(A)$, $L_{fu}(A)$ and $L_{fd}(A)$ can be found.

For a switch of an RBG such that elements of the previous sets of $B_{bu}(A \cup R_n(A))$ are all distinct, the upper bound holds with equality. If there are adequate reconfigurable links, the RBG is fully connected and $r_n = |X| - 1$ Hence, the lower bound holds with equality. QED

6 Two Design Examples

No physical device constraints have been assumed so far. Due to different physical interconnection techniques, the achievable maximum fault tolerance and reliability for a given cost function may be totally different. In this section, the maximum fault tolerance and reliability of different implementations are considered versus the hardware overhead. Switches which are passive devices are usually more expensive and less reliable. On the other hand, passive links are usually much cheaper and less likely to fail, therefore, they are not considered in the analysis. It is assumed that the cost is same for every switch.

The *Omega network* [15] is used as an example for the operational topology. Results derived in this section can be applied to many unique path Multistage Interconnection Networks (MIN) [22], since they are topologically identical to Omega network (from the perspective of this paper). Also, the results can be easily extended to many other useful interconnection networks, such as the internally non-blocking Batcher Banyan network [11].

A four stage Omega network is shown in Figure 7. There are $\log N$ stages of operational switches in the Omega network. Each stage has $\frac{N}{2}$ 2×2 switches. Between stages, switches are interconnected by shuffle interconnections. If the first and last stages are combined and each switch has an external source and an external destination, then this is the same as the ShuffleNet [1]. If external sources and external destinations are connected to switches in the first stage and switches in the last stage respectively, then a unique path multistage interconnection network switch fabric is formed [24].

An Omega network is labeled for the rest of the analysis as shown in Figure 7. Similar labeling can be found in many designs and is not the focus of this paper. Switches in each stage are labeled from 0 to $\frac{N}{2} - 1$ such that:

1. The upper outnode of switch i is connected to the upper innode of switch $2i$ in the next stage for $i < \frac{N}{4}$.

2. The lower outnode of switch i is connected to the upper innode of switch $2i + 1$ in the next stage for $i < \frac{N}{4}$.

3. The upper outnode of switch $i + \frac{N}{4}$ is connected to the lower innode of switch $2i$ in the next stage for $i < \frac{N}{4}$.

4. The lower outnode of switch $i + \frac{N}{4}$ is connected to the lower innode of switch $2i + 1$ in the next stage for $i < \frac{N}{4}$.

6.1 Arbitrary Replacement Interconnection

For a centralized switch fabric [24], a possible implementation is shown in Figure 9a. The DMUX attached to an outnode selects an operational link. Appropriate selections are made such that only one link of the MUX of an innode is operational. For an electronic implementation, simple off-line selectable MUXs, DMUXs and redundant wires are required. For an optical implementation, lasers with electronically selectable optical paths are required [17]. There is a non-negligible overhead for both designs but they preserve the operational topology under failure mode which is very important in high speed switch design as noted earlier.

The logical model for an *arbitrary* replacement interconnection is that each outnode has l links, which can be interconnected (reconfigured) to the innodes of any selected switch. Similarly, each innode has l incoming links, which can be interconnected (reconfigured) to the outnodes of *any* selected switch. There are no constraints on which outnode is connected to which innode. The only restriction is the number of links that each node has.

In section 4, an upper bound on the maximum fault tolerance was given based only on r_n, and the limitations due to r_p were not considered. In reality, both the next and previous replacement sets are restricted by the number of available links l. Designs considered in this section sacrifice some degree of maximum fault tolerance but pay more attention to balancing the limits of both replacement link constraints.

Since the switches in an operational network are organized in stages and are structured similarly at each stage, it is better to design replacement according to this structure. To simplify network replacement, it is assumed that spares are equally distributed to all stages and that the number of spares per stage $s_s = s/\log N$ is an integer. Each group of s_s switches can replace switches in the same stage, i.e., cross stage replacement is not allowed.

We label the operational switches according to the Omega labeling discussed above and label the spares in each stage consecutively from 0 to s_s. Assume that there are r_n replacement links for each operational switch and there are k replacement links per spare, $1 \leq k \leq r_n$ This reconfigurable network is referred as a *Stagewise Arbitrary Replacement Interconnection* with k links per spare ($SARI/k$). Let

$$\begin{aligned} a &= \frac{N}{2ks_s} \qquad if\ ks_s \leq N/2 \\ b &= \frac{2ks_s}{N} \qquad if\ ks_s \geq N/2 \\ c &= \frac{N}{2s_s} \end{aligned}$$

For a regular design, we restrict the above parameters to be integers when applicable and $s_s \leq N/2$.

Let *mod* function be the division remainder. Then the replacement links of the SARI/k are given as follows:

1. As shown in Figure 10, the $r_n + 1$ replacement links from *operational switches* for a switch x are

$$(x,\ x),\ (x,\ (x+1)mod(N/2)),\ ..,\ (x,\ (x+r_n)mod(N/2))$$

2. Assign replacement links from *spare switches* according to the size of ks_s:

   (a) If $ks_s \leq N/2$: As shown in Figure 11, replacement link m of spare i is connected to operational switch $(ia + ms_s a)mod(N/2)$, i.e.

$$\begin{aligned} (i,\ m) \rightarrow (ia + ms_s a)mod(N/2) \qquad \forall m &\in (0..k-1) \\ i &\in (0..s_s - 1) \end{aligned}$$

(b) If $ks_s > N/2$:

As shown in Figure 12, k replacement links of each switch are grouped to b groups of c replacement links.

Replacement link m of group j of spare i is connected to a switch $(ms_s + i + j) mod(N/2)$, i.e.

$$(i,\ j,\ m) \rightarrow (ms_s + i + j) mod(N/2) \qquad \begin{aligned} \forall m &\in (0..c-1) \\ j &\in (0..b-1) \\ i &\in (0..s_s - 1) \end{aligned}$$

The replacement links for a switch due to operational switches are from itself and the r_n switches after it. One benefit for this design is that consecutive switches (according to the Omega labeling) share common switches in the previous and next replacement sets. This helps to reduce the number of required reconfigurable links. Note that for a group of q *consecutive* switches, there are $q + r_n$ consecutive operational switches in the previous or next replacement set.

For the small number of spares (case 2a), every a switches share one spare. Only the first one of each group has a link to a spare and is referred to as the marked switch as shown in Figure 11. For the larger number of spares (case 2b), replacement links of each switch are divided into b groups . Only replacement links in the first group are shown in Figure 12 and they cover all switches in X. The other groups are not shown for simplicity. They are connected to the switches in X by rotating one switch per group, i.e., the first replacement link of the second group is connected to the second switch in X. The last replacement link of this group is connected to the first switch. The third group begins with the third switch in X etc.

Theorem 6:

The maximum fault tolerance of SARI/k is

$$t_{SARI/k} = min\left(s_s,\ r_n + \left\lfloor \frac{2ks_s}{N} \right\rfloor\right)$$

□

Proof:

Similar to the proof in Corollary 1, we can always find a set larger than s_s or $r_n + \left\lfloor \frac{2ks_s}{N} \right\rfloor$ that fails the network. Hence,

$$t_{SARI/k} \leq min\left(s_s,\ r_n + \left\lfloor \frac{2ks_s}{N} \right\rfloor\right)$$

Let A be an arbitrary switch set and

$$\begin{aligned} |A| &= q \\ Y_x(A) &= Y(A) \cap (Y - S) \\ Y_s(A) &= Y(A) \cap S \\ Y(A) &= Y_x(A) \cup Y_s(A) \end{aligned}$$

$|Y_x(A)|$ due to a non-consecutive A is always greater than that due to a consecutive set of the same size, hence

$$\begin{aligned} |Y_x(A)| &\geq q + r_n \qquad if\ q < |X| - r_n \\ &= |X| \qquad if\ q \geq |X| - r_n \end{aligned}$$

Note that the equality holds for all consecutive sets.

Consider the case that $ks_s \leq \frac{N}{2}$. For $q \leq |X| - r_n$, we have

$$\begin{aligned} |Y(A)| - |A| &\geq q + r_n - q + |Y_s(A)| \\ &\geq r_n \end{aligned}$$

For $q \geq |X| - r_n$, we have

$$\begin{aligned} |Y(A)| - |A| &\geq N/2 - q + \left\lfloor \frac{qs_s}{N/2} \right\rfloor \\ &= N/2 - q + \left\lfloor s_s - \frac{(N/2 - q)s_s}{N/2} \right\rfloor \\ &= s_s + (N/2 - q) - \left\lfloor (N/2 - q)\frac{s_s}{N/2} \right\rfloor \\ &\geq s_s \end{aligned}$$

Consider the case that $ks_s \geq \frac{N}{2}$. For $q \leq |X| - r_n$, we have

$$\begin{aligned} |Y(A)| - |A| &\geq q + r_n - q + b\left\lceil \frac{q}{c} \right\rceil \\ &\geq r_n + \left\lfloor \frac{2ks_s}{N} \right\rfloor \end{aligned}$$

For $\lceil \frac{q}{c} \rceil + b - 1 \geq s_s$ then all spares are in $Y_s(A)$ and we have

$$\begin{aligned} |Y(A)| - |A| &\geq q + r_n - q + s_s \\ &\geq s_s \end{aligned}$$

For the case that

$$q \geq |X| - r_n$$
$$\left\lceil \frac{q}{c} \right\rceil + b - 1 < s_s$$

then

$$\begin{aligned} |Y(A)| - |A| &\geq \frac{q}{c} + b - 1 + N/2 - q \\ &= s_s + \frac{2(q-1)s_s}{N} + \left(\frac{N}{2} - q - 1\right)\left(1 - \frac{2s_s}{N}\right) \\ &\geq s_s \end{aligned}$$

From Theorem 1, we conclude that

$$t_{SARI/k} \geq min\left(s_s,\ r_n + \left\lfloor \frac{2ks_s}{N} \right\rfloor\right)$$

QED

The link requirement of SARI/k is derived in the following theorem.

Theorem 7: An upper bound on the link requirement for SARI/k is

$$\begin{aligned} l_{SARI/k} &\leq 3r_n + 1 + \min\left(s_s,\ \left\lceil r_n + \frac{2ks_s}{N} \right\rceil\right) \quad if\ ks_s \geq \frac{N}{2} \\ &\leq 3r_n + 1 + \min\left(s_s,\ \left\lceil \frac{2ks_s(r_n+1)}{N} \right\rceil\right) \quad if\ ks_s \leq \frac{N}{2} \end{aligned}$$

□

Proof: Referring to Figure 4, the switches in $R_n(A) \cup A$ for a single switch set A are consecutive and are of size $r_n + 1$.

Consider the required links due to the operational switches first as shown in Figure 10. The size of backward upper node set in the previous stage is

$$|B_{bu}(A) \cup B_{bu}(R_n(A))| = r_n + 1$$

Referring to Figure 8, the shuffle interconnections of Omega network require that the backward upper node set is divided into two subsets: the subset for upper outnodes and the subset for lower outnodes. The switch sets containing these two subsets are both consecutive. Let the size of switch set for upper outnodes be e. Then the size of switch set for lower

outnodes is $r_n + 1 - e$. Hence, the size of $R_p(B_{bu}(A) \cup B_{bu}(R_n(A)))$ due to operational switches is

$$(e + r_n) + (r_n + 1 - e + r_n) = 3r_n + 1$$

Similarly, the size of forward upper node set in the next stage is

$$|B_{fu}(A) \cup B_{fu}(R_n(A))| = r_n + 1$$

These $r_n + 1$ switches are every other switch rather than consecutive. The r_n switches in between are not contained in the forward upper node set (actually containing the forward lower set) but are in the previous replacement set. Thus the $2r_n + 1$ switches are consecutive. Therefore, the size of $R_p(B_{fu}(A) \cup B_{fu}(R_n(A)))$ due to operational switches is

$$2r_n + 1 + r_n = 3r_n + 1$$

The results can be derived for the backward and forward lower node sets similarly.

Consider the required links due to the spares. All the spares that have replacement links to $B_{fu}(A) \cup B_{fu}(R_n(A))$ are in the previous set. There are no previous sets for the spares. For the case $2ks_s \geq N$, there are $r_n + 1 + b - 1 = r_n + b$ (but no more than s_s) spares in the previous replacement set. For the case $2ks_s \leq N$, there are $\lceil (r_n + 1)/a \rceil$ (but no more than s_s) spares in the previous replacement set. QED

Although there are some failure sets larger than the maximum fault tolerance that fail the network, some do not. Determination of the network reliability by counting such failure sets for an arbitrary graph is known to be NP-hard [5]. However, larger fault tolerable failure sets typically make an insignificant contribution to the network reliability since the component failure probability is very small in most practical designs. Approximations can be obtained by considering only the small failure sets.

Assume that the failure probability of every switch is identical and independent. Let the switch failure probability be $1 - p$ and network failure probability be $1 - R$. The total number of failure set instances for i failures among Y is the number of combinations $C(|Y|, i)$. Let the number of fault tolerable instances be $K(i)$. Then,

$$R = \sum_{i=0}^{|Y|} K(i)(1-p)^i p^{|Y|-i}$$

Since the failure sets of size less than or equal to the maximum fault tolerance are always fault tolerable and the failure sets of size greater than the number of spares are always not fault tolerable, then

$$\begin{aligned} K(i) &= C(|Y|,\ i) && i \leq t \\ &= 0 && i > s \end{aligned}$$

For the case $s_s \leq t$, the network reliability can be found exactly.

Let

$$\begin{aligned} Y_y^F(A) &= Y(A) \cap F \\ Y_y^{\overline{F}}(A) &= Y(A) - Y(A) \cap F \\ Y_x^F(A) &= Y_x(A) \cap F \\ Y_x^{\overline{F}}(A) &= Y_x(A) - Y_x(A) \cap F \\ Y_s^F(A) &= Y_s(A) \cap F \\ Y_s^{\overline{F}}(A) &= Y_s(A) - Y_s(A) \cap F \end{aligned}$$

Since the operational switches of SARI/k are consecutively replaceable, an interesting property of its failure sets is as follows.

Theorem 8: A failure set F of SARI/k is fault tolerable if and only if for any consecutive $A \subset X$,

$$|Y_y^{\overline{F}}(A)| \geq |A|$$

□

Proof: If F is fault tolerable, it is easy to see from Complete matching theory that the above equation holds for any subset A.

Conversely, assume that the above equation holds for any consecutive subset and we want to show that F is fault tolerable. Let A_0 be a subset of X that is not a consecutive set. Consider the case that A_0 consists of two consecutive subsets $(A_1,\ A_2)$ such that

$$\begin{aligned} A_0 &= A_1 \cup A_2 \\ A_1 \cup A_2 &= \emptyset \end{aligned}$$

Let the set of consecutive switches between A_1 and A_2 be W.

$$U_x = Y_x^{\overline{F}}(A_1) \cap Y_x^{\overline{F}}(A_2)$$

$$
\begin{aligned}
U_s &= Y_s^{\overline{F}}(A_1) \cap Y_s^{\overline{F}}(A_2) \\
D_x &= Y_x^{\overline{F}}(A_a) - Y_x^{\overline{F}}(A_1 \cup A_2) \\
D_s &= Y_s^{\overline{F}}(A_a) - Y_s^{\overline{F}}(A_1 \cup A_2) \\
A_a &= A_1 \cup W \cup A_2
\end{aligned}
$$

Note that A_a is a consecutive set. We conclude that:

- If $U_x \neq \emptyset$ then $D_x = \emptyset$.
- if $U_s \neq \emptyset$ then $D_s = \emptyset$.
- $|D_x + D_s| \leq |W|$

Consider the case that $U_x = \emptyset$ and $U_s = \emptyset$.

$$
\begin{aligned}
|Y_y^{\overline{F}}(A_1 \cup A_2)| - |A_1| - |A_2| &= |Y_x^{\overline{F}}(A_1)| + |Y_s^{\overline{F}}(A_1)| - |A_1| \\
&\quad + |Y_x^{\overline{F}}(A_2)| + |Y_s^{\overline{F}}(A_2)| - |A_2| \\
&\geq 0
\end{aligned}
$$

Consider the case that $U_x \neq \emptyset$ or $U_s = \emptyset$. Since

$$
\begin{aligned}
|Y_y^{\overline{F}}(A_1 \cup W \cup A_2)| - |A_1| - |W| - |A_2| &= |Y_x^{\overline{F}}(A_1 \cup A_2)| + |Y_s^{\overline{F}}(A_1 \cup A_2)| \\
&\quad + |D_x| + |D_s| - |A_1| - |W| - |A_2| \\
&\geq 0
\end{aligned}
$$

we have

$$
\begin{aligned}
|Y_y^{\overline{F}}(A_1 \cup A_2)| - |A_1| - |A_2| &\geq |W| - |D_x| - |D_s| \\
&\geq 0
\end{aligned}
$$

By induction on the number of non-consecutive subsets, we conclude this theorem. QED

We consider only the case of $ks_s \leq N/2$. In this case,

$$t = min(s_s, r_n)$$

Consider the set W between two marked switches 1 and 2 which have replacement links from two consecutive spares respectively as shown in Figure 11. For a failure set $F \subset Y(W)$ and $|F| = r_n + 1$:

$$
\begin{aligned}
|W| &= a - 1 \\
|Y(W) - Y(W) \cup F| &= a - 1 + r_n - (r_n + 1) \\
&< a - 1
\end{aligned}
$$

Hence, F is not fault tolerable.

From the Theorem 8, the above failure sets are the only fault intolerable cases. There are s_s of them and $|Y(W)| = a - 1 + r_n$. Hence,

$$K(r_n + 1) = C(|Y|,\ r_n + 1) - s_s C(a - 1 + r_n,\ r_n + 1)$$

Similarly $K(r_n + 2)$ can be derived as follows.

$$\begin{aligned} K(r_n + 2) \ = \ & C(|Y|,\ r_n + 2) - s_s[C(2a + r_n, r_n + 2) - C(a - 1 + r_n, r_n + 2)] \\ & - (|Y| - 3a - r_n - 1)s_s C(a - 1 + r_n,\ r_n + 1) \end{aligned}$$

For higher values of $K(i)$, approximations and bounds based on $K(r_n+1)$ and $K(r_n+2)$ can be used.

The network reliability of SARI/1 networks is plotted in Figure 14. Four stage and eight stage networks with a single replacement link and several spares are considered. Corresponding networks without any spares and replacement links are also plotted for comparison. Significant improvement is observed even with a single spare per stage. In Figure 16, the maximum fault tolerance *versus* the number of spares per stage is shown. The maximum fault tolerance can be increased linearly with the number of spares per stage provided that there are sufficient replacement links.

6.2 Star Coupler Interconnections

Another possibility is to interconnect outnodes and innodes between two consecutive stages by a star coupler. Each outnode has a tunable laser diode that can be tuned to any channel of an innode on the same star coupler as shown in Figure 9b. Although this implementation requires tunable laser diodes which are still very expensive today, for applications that use WDM [1], the reconfigurable links are already present in the design.

Upper innodes, lower innodes, upper outnodes or lower outnodes of a group switches are attached to the same star couplers as shown in Figure 13. Innodes and outnodes on the same star coupler are logically fully connected. On the other hand, no other switches are accessible to them. This type of network is referred to as a *Star Coupler Interconnection* network (*SCI*).

Since only switches on the same star coupler are replaceable by each other, spares are needed for each star coupler. Any failed switch can be replaced by any of these spares. The four nodes of a switch are connected to different star couplers. There are l switches on each side of an $l \times l$ star coupler. An example of SCI implementation of the four stage Omega network of Figure 5 is shown in Figure 13. The outnodes of the

switches $\{0,\ 1,\ 2,\ 3\}$ in the first stage are connected to the innodes of the switches $\{0,\ 2,\ 4,\ 6\}$ in the second stage via two star couplers.

The structure of the Omega network (Figure 5) requires that in later stages, upper and lower innodes for a switch be on the *same* star coupler. Similarly, the outnodes for this stage will be on the same star coupler (of the next stage) and the number of nodes per star coupler is reduced to $\frac{l}{2}$, see Figure 13. For example: upper outnodes of the switches $\{0,\ 2,\ 4,\ 6\}$ in the second stage are connected to both the upper and lower innodes of the switches $\{0,\ 4\}$ in the third stage (refer to Figure 5).

Since the smallest number of spares in the star coupler groups determines the maximum fault tolerance, we implement x spares in each star coupler group, for both the full size group and the half size group. The maximum fault tolerance is

$$t = x$$

Let g_1 and g_2 be the largest integers that divide the total number of operational switches, $\frac{N}{2}$, such that

$$\begin{aligned} g_1 &\leq l - x \\ g_2 &\leq \frac{l-x}{2} \end{aligned}$$

For simplicity of analysis, assume that $l - x$ is a multiple of 2. Group the switches in the first stage so that they have the same high order $\log N - \log(l-x)$ bits. Then, from stage 0 to stage $\log N - \log(l-x) - 1$, the switches are in the full groups and the rest of switches are in the half groups.

A group of switches will fail if the number of failures is larger than the number of spares, x. The network fails if any of the groups fails. Let

$$G(l,x) = \sum_{i=0}^{x} C(l,\ i) p^{l-i} (1-p)^i$$

Then, the network reliability is

$$R = G^{[\log N - \log(l-x)]\frac{N}{2g_1}}(g_1, x) G^{\log(l-x)\frac{N}{2g_2}}(g_2, x)$$

The network reliability of SCI networks is plotted in Figure 15. Five stage and eight stage networks with a single spare and various star coupler sizes are considered. Significant improvement can be observed with a single spare per star coupler group.

Although both the SARI/k and SCI networks improve reliability significantly, SCI usually requires many more spares for a large network since the number of spares for SARI/k is $O(\log N)$ while the number of spares for SCI is $O(N \log N)$.

References

[1] A. S. Acampora. "A multichannel multihop local lightwave network". In *GLOBECOM'87*, pages 1459–1467, 1987.

[2] G. B. Adams III, D. P. Agrawal, and H. J. Siegel. "A survey and comparison of fault-tolerant interconnection networks". *IEEE Comput. Mag.*, Vol. 20:14–27, Jun. 1987.

[3] G. B. Adams III and H. J. Siegel. "The extra stage cube: A fault tolerant interconnection network for supersystems". *IEEE Trans. Comput.*, Vol. C-31:443–454, May 1982.

[4] L. Ciminiera and A. Serra. "A connecting network with fault tolerance capabilities". *IEEE Trans. Comput.*, Vol. C-35:578–580, Jun. 1986.

[5] C. J. Colbourn. *"The Combinatorics of Network Reliability"*. Oxford University Press, New York, 1987.

[6] K. Y. Eng. "A photonic knockout switch for high-speed packet networks". *IEEE J. Select. Areas Commun.*, Vol. SAC-6(No. 7):1107–1116, Aug. 1988.

[7] K. Y. Eng and M. G. Hluchyj. "A knockout switch for variable-length packets". *IEEE J. Select. Areas Commun.*, Vol. SAC-5:1426–1435, Dec. 1987.

[8] K. Y. Eng, M. J. Karol, and Y. S. Yeh. "A growable packet (ATM) switch architecture: Design principles and applications". In *GLOBECOM'89*, pages 1159–1165, 1989.

[9] S. Even. *"Graph Algorithms"*. Computer Science Press, Maryland, 1979.

[10] P. Henry. "High-capacity lightwave local area network". *IEEE Commun. Mag.*, Vol. 27(No. 10):20–26, Oct. 1989.

[11] A. Huang and S. Knauer. "Starlite: A wideband digital switch". In *GLOBECOM'84*, pages 121–125, 1984.

[12] J. Y. Hui and E. Arthurs. "A broadband packet switch for integrated transport". *IEEE J. Select. Areas Commun.*, Vol. SAC-5(No. 8):1264–1273, Oct. 1987.

[13] H. Kobrinski and K. W. Cheung. "Wavelength-tunable optical filters: Applications and technologies". *IEEE Commun. Mag.*, Vol. 27(No. 10):53–63, Oct. 1989.

[14] S. Y. Kuo and W. K. Fuchs. "Efficient spare allocation for reconfigurable arrays". *IEEE Design and Test*, Vol. 4:24–31, Feb. 1987.

[15] D. Lawrie. "Access and alignment of data in an array processor". *IEEE Trans. Comput.*, Vol. C-24(No. 12):1145–1155, Dec. 1975.

[16] T. P. Lee and C. E. Zah. "Wavelength-tunable and single-frequency semiconductor lasers for photonic communications networks". *IEEE Commun. Mag.*, Vol. 27(No. 10):42–52, Oct. 1989.

[17] S. D. Personick. "Photonic switching: technology and applications". *IEEE Commun. Mag.*, Vol. 25(No. 5):5–8, May 1987.

[18] C. S. Raghavendra and A. Varma. "INDRA: A class of interconnection networks with redundant paths". In *1984 Real-Time System Symp., Computer Society Press, Silver Spring, Md.*, pages 153–164, 1984.

[19] S. M. Reddy and V. P. Kumar. "On fault-tolerant multistage interconnection networks". In *1984 Int'l Conf. Parallel Processing, Computer Society Press, Silver Spring, Md.*, pages 155–164, 1984.

[20] P. W. Smith. "On the physical limits of digital optical switching and logic element". *Bell Syst. Tech. J.*, Vol. 61:1975–1993, Oct. 1982.

[21] J. S. Turner. "New directions in communications (or which way to the information age?". *IEEE Commun. Mag.*, Vol. 24(No. 10):8–15, Oct. 1986.

[22] C. L. Wu and T. Y. Feng. "On a class of multistage interconnection networks". *IEEE Trans. Comput.*, Vol. C-29(No. 8):694–702, Aug. 1980.

[23] S. C. Yang and J. A. Silvester. "Fault-tolerant multistage interconnection networks: Performance/reliability tradeoffs". *Comput. System Science and Engineering, Butterworths Scientific Ltd.*, Vol. 5(No. 4):233–242, Oct. 1990.

[24] S. C. Yang and J. A. Silvester. "A fault tolerant reconfigurable atm switch fabric". to appear INFOCOM'91, Florida, 1991.

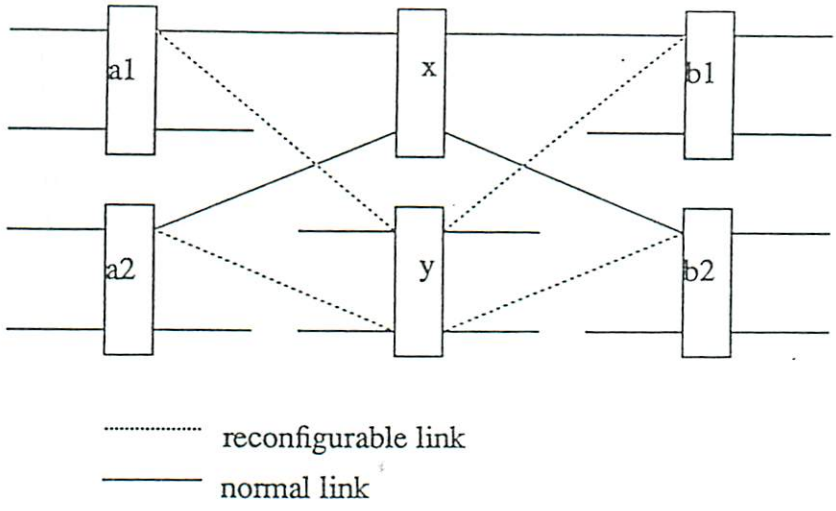


Figure 1. x is Replaceable by y
(a1, a2, b1, b2 are operational)

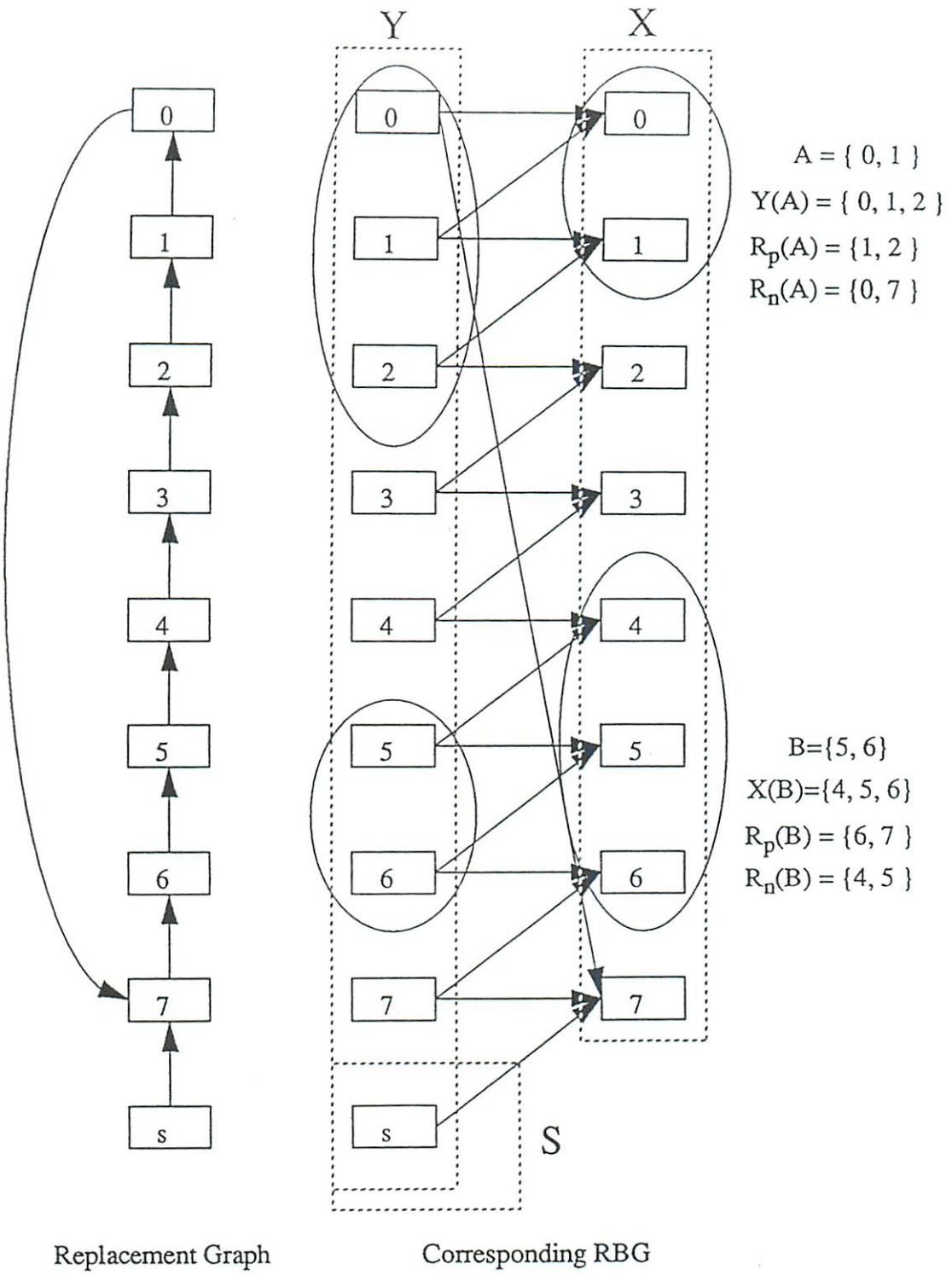


Figure 2. Example of a Reconfigurable Bipartite Graph (RBG)

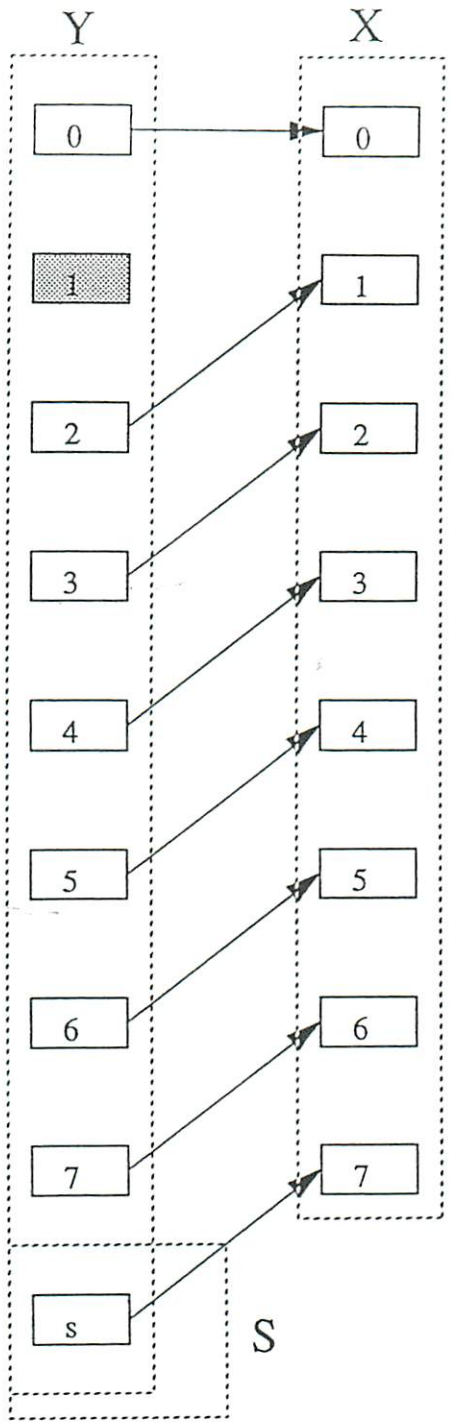


Figure 3. Complete Matching for a Failure set {1}

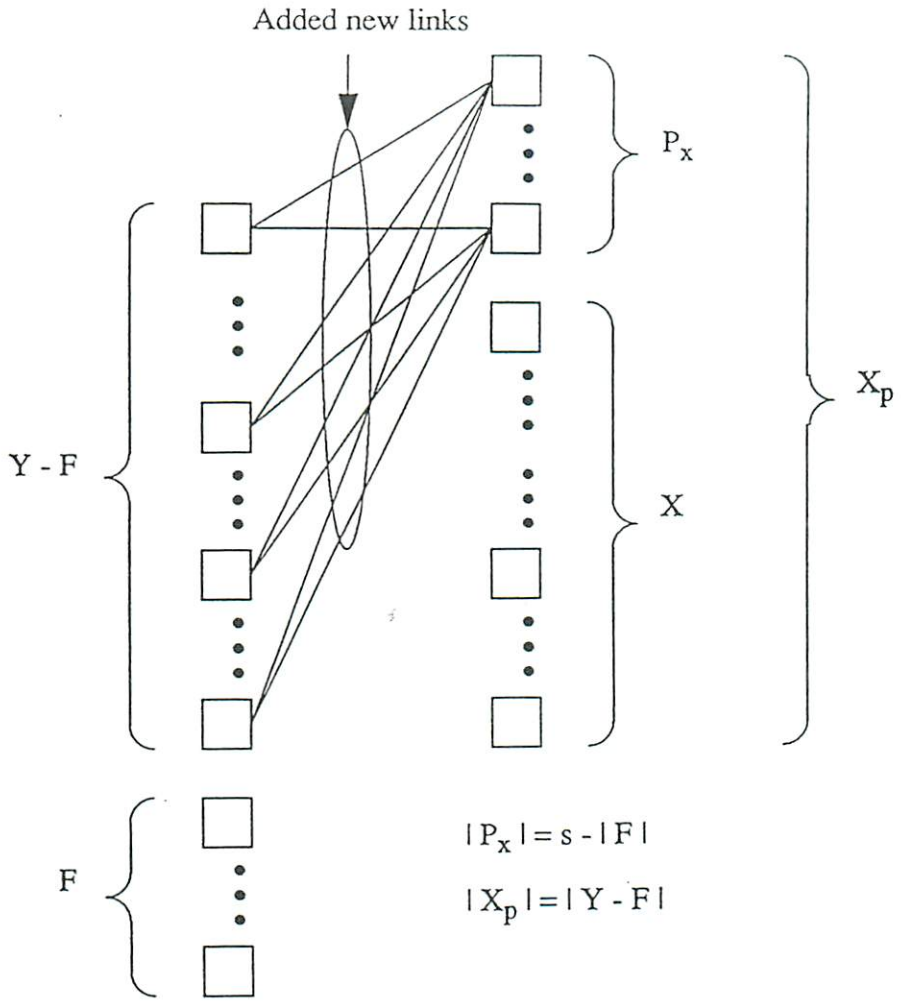


Figure 4. Construction of P_x for Theorem 2

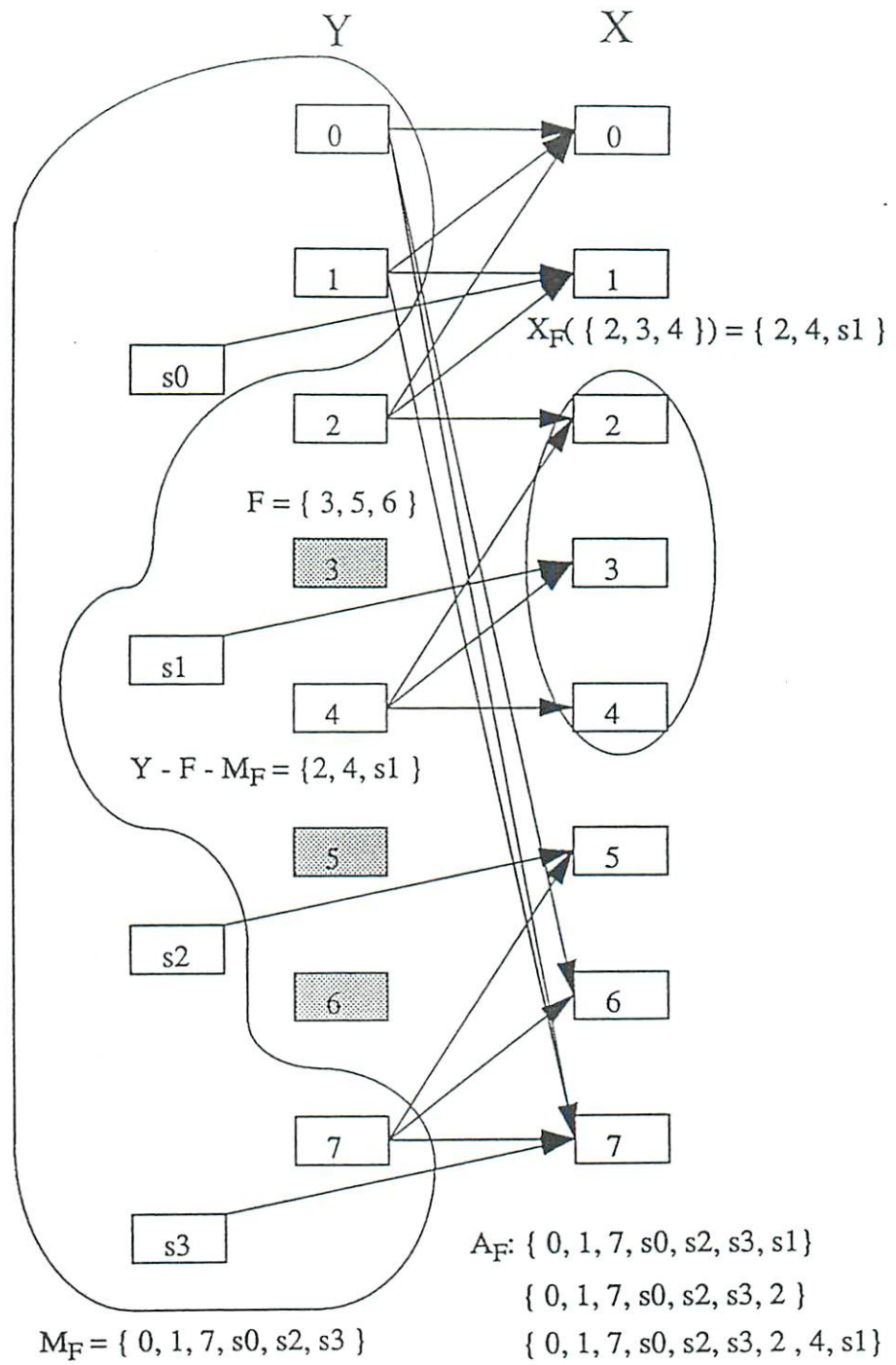
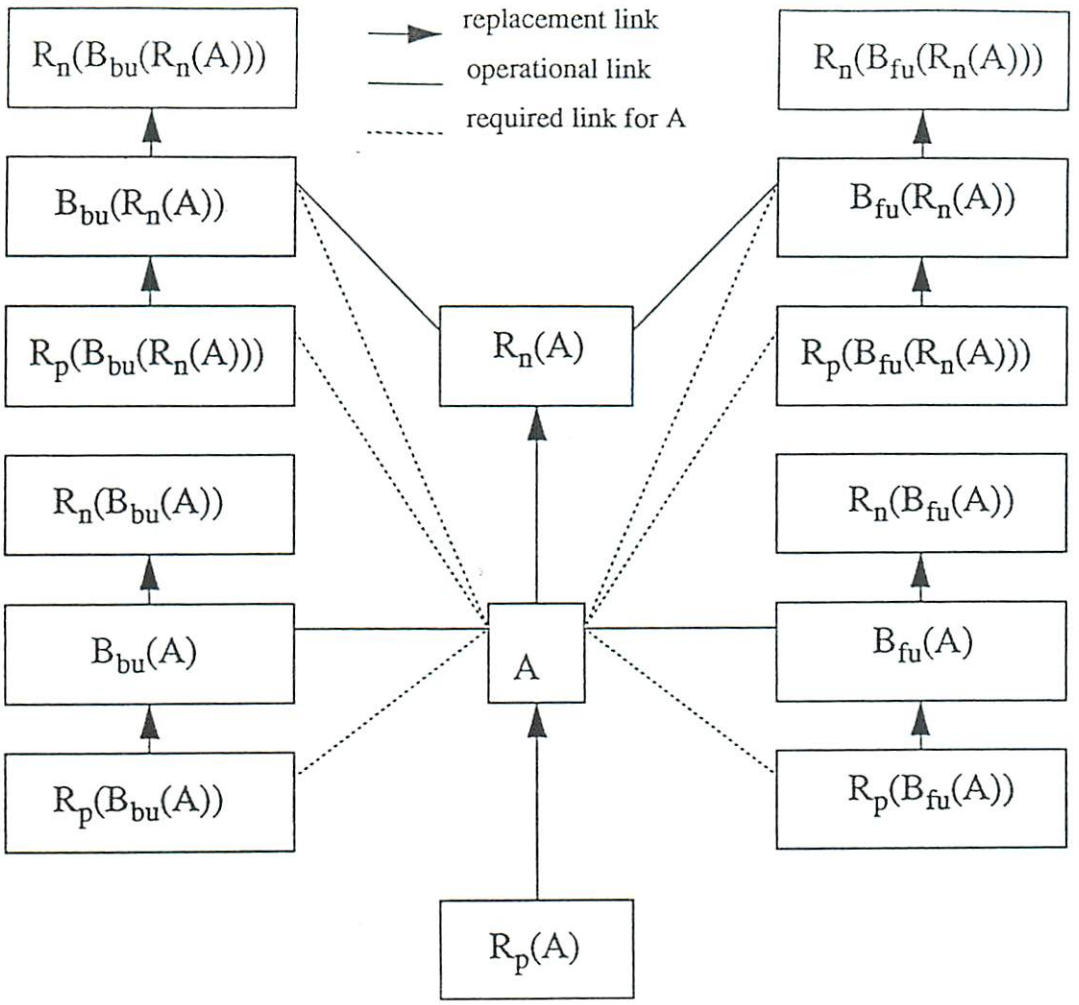


Figure 5. A Minimal Maximal Set M_F for Failure Set F



$$L_{fu}(A) = B_{fu}(A) \cup R_p(B_{fu}(A)) \cup B_{fu}(R_n(A)) \cup R_p(B_{fu}(R_n(A)))$$

$$L_{fd}(A) = B_{fd}(A) \cup R_p(B_{fd}(A)) \cup B_{fd}(R_n(A)) \cup R_p(B_{fd}(R_n(A)))$$

$$L_{bu}(A) = B_{bu}(A) \cup R_p(B_{bu}(A)) \cup B_{bu}(R_n(A)) \cup R_p(B_{bu}(R_n(A)))$$

$$L_{bd}(A) = B_{bd}(A) \cup R_p(B_{bd}(A)) \cup B_{bd}(R_n(A)) \cup R_p(B_{bd}(R_n(A)))$$

Figure 6. Upper Node Link Requirement Diagram

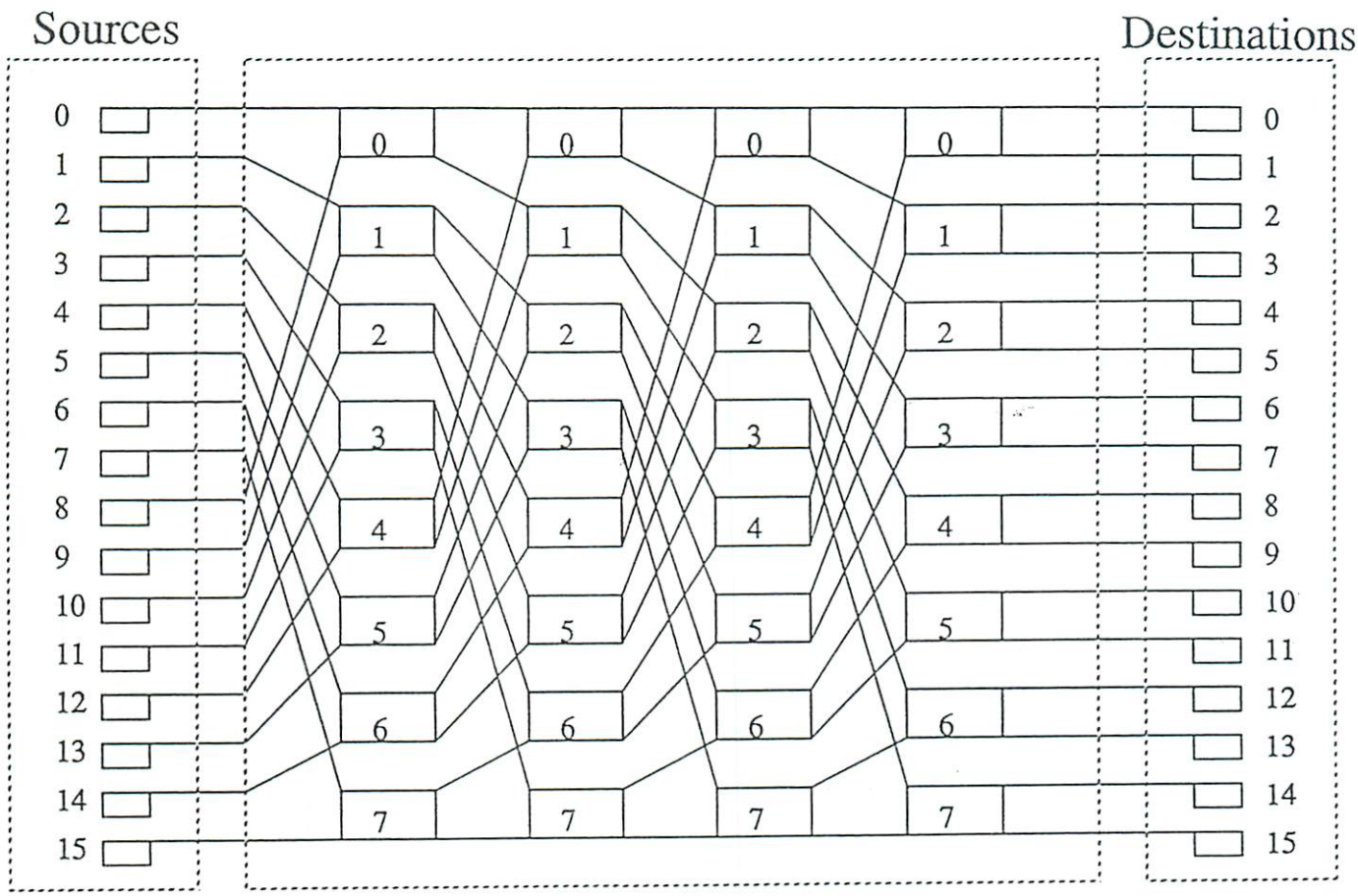


Figure 7. A 4 stage Omega Network

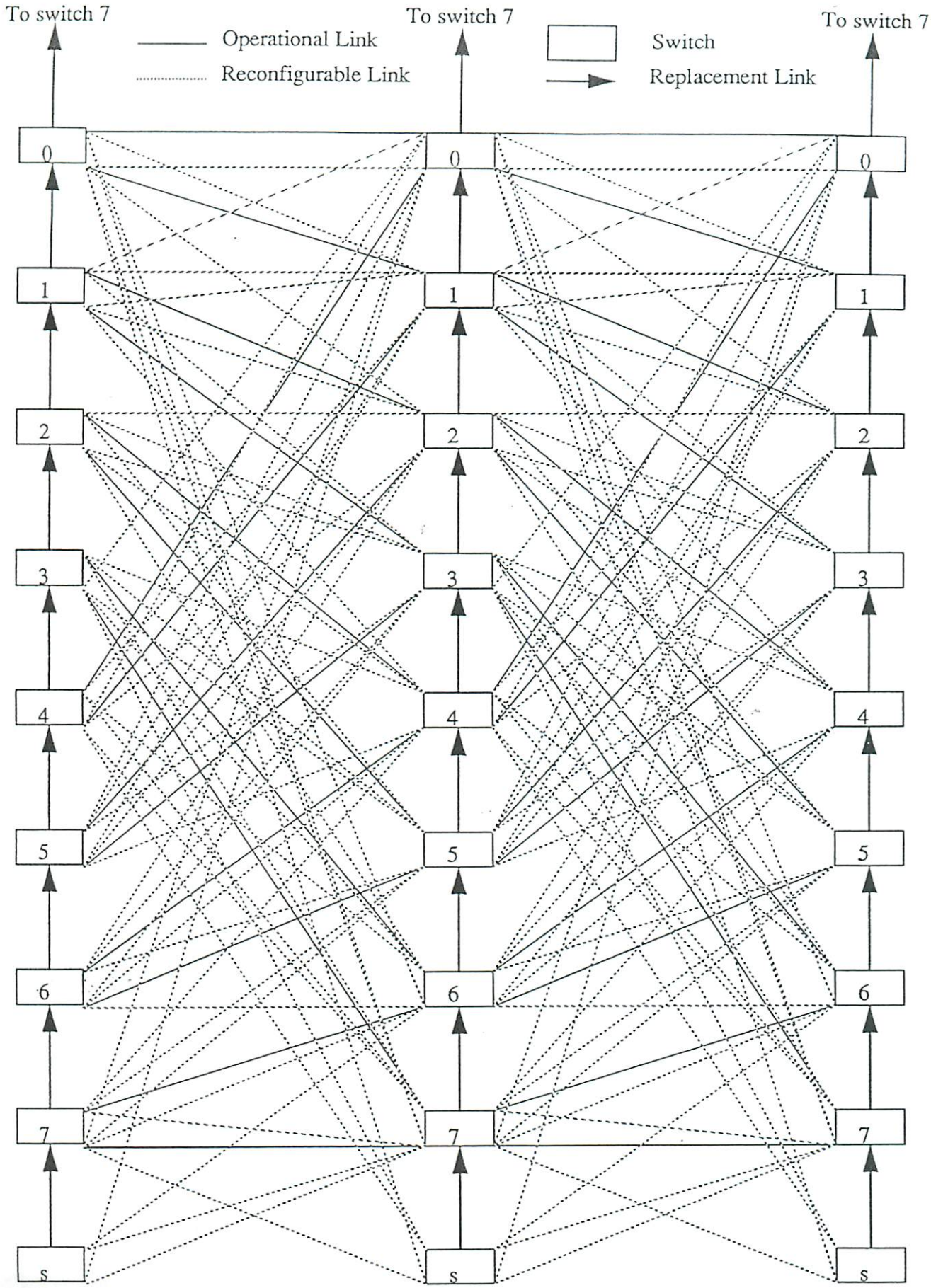
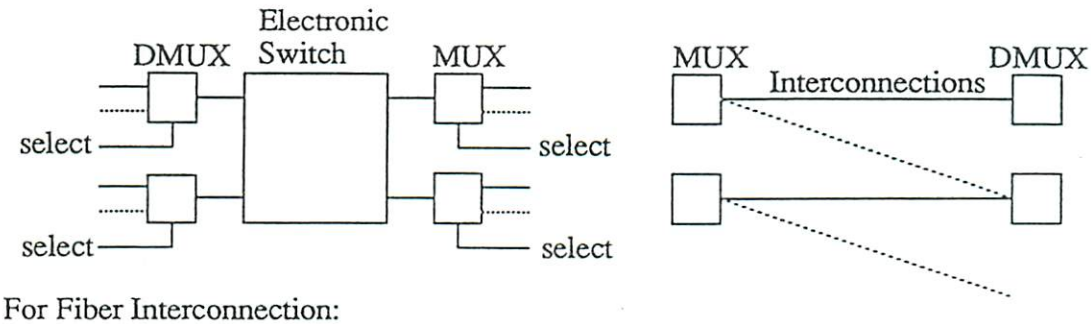


Figure 8. Reconfigurable Links for Omega Network

a) Arbitrary Replacement Interconnection



For Fiber Interconnection:

MUX=Optic Transmitter+Power Splitter

DMUX=Optic Receiver+Star Coupler

b) Star Coupler Interconnections

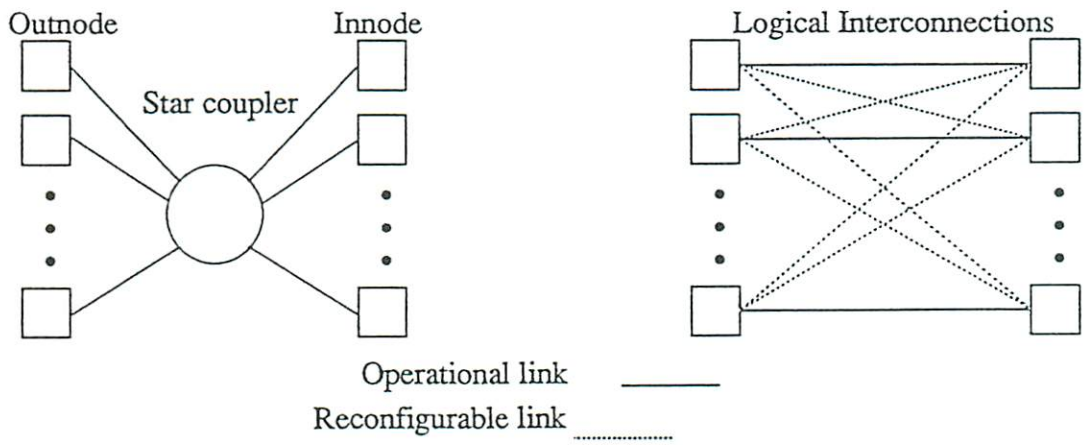


Figure 9. Two Interconnection Implementations

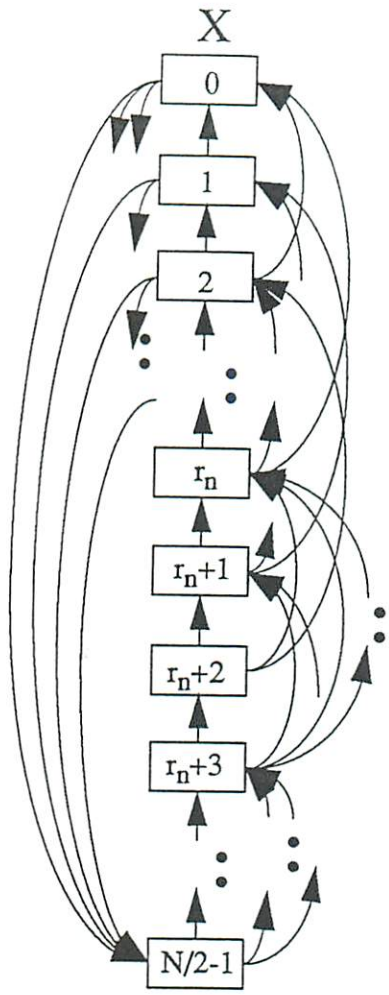


Figure 10. Replacement due to Operational Switches for SARI/k

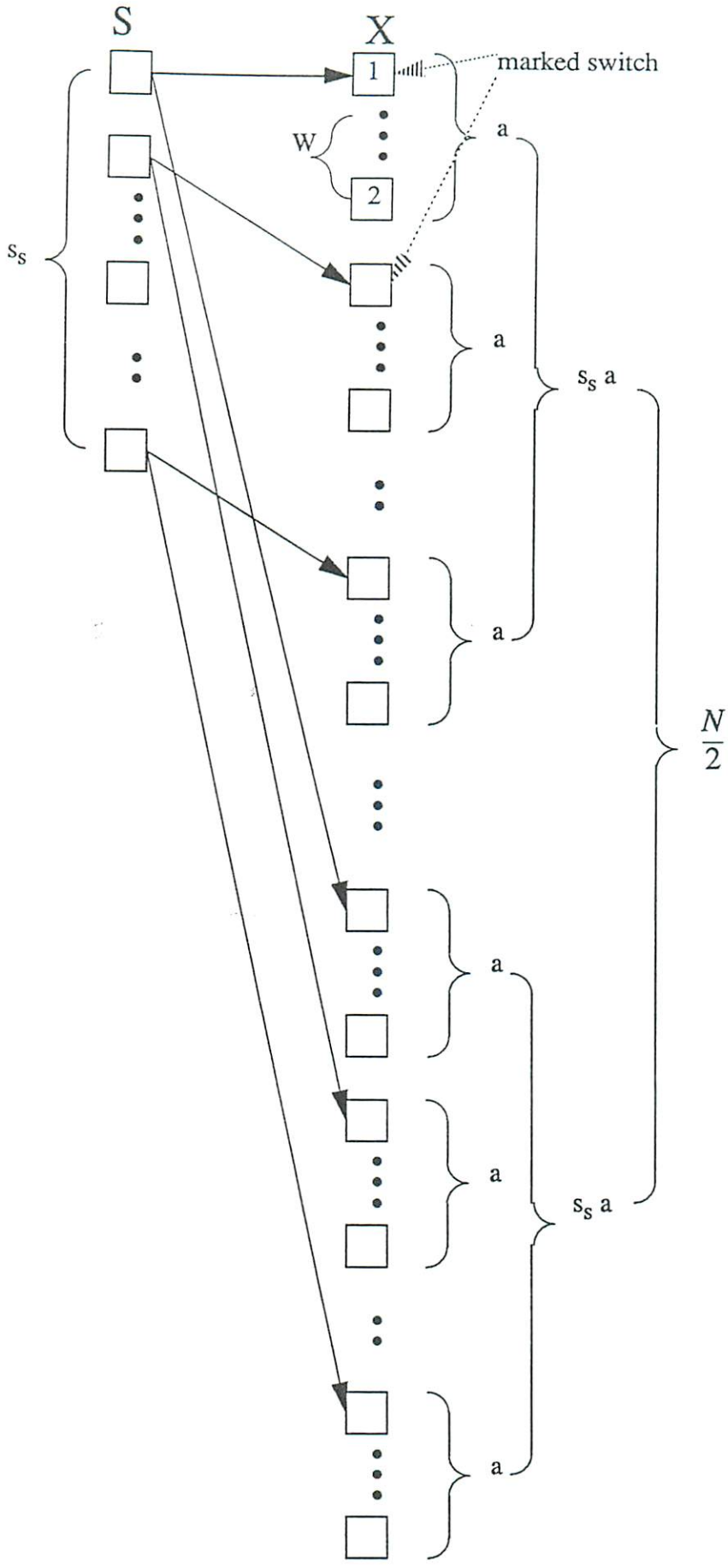


Figure 11. Replacement of spares for SARI / k: $ks_s \leq \frac{N}{2}$

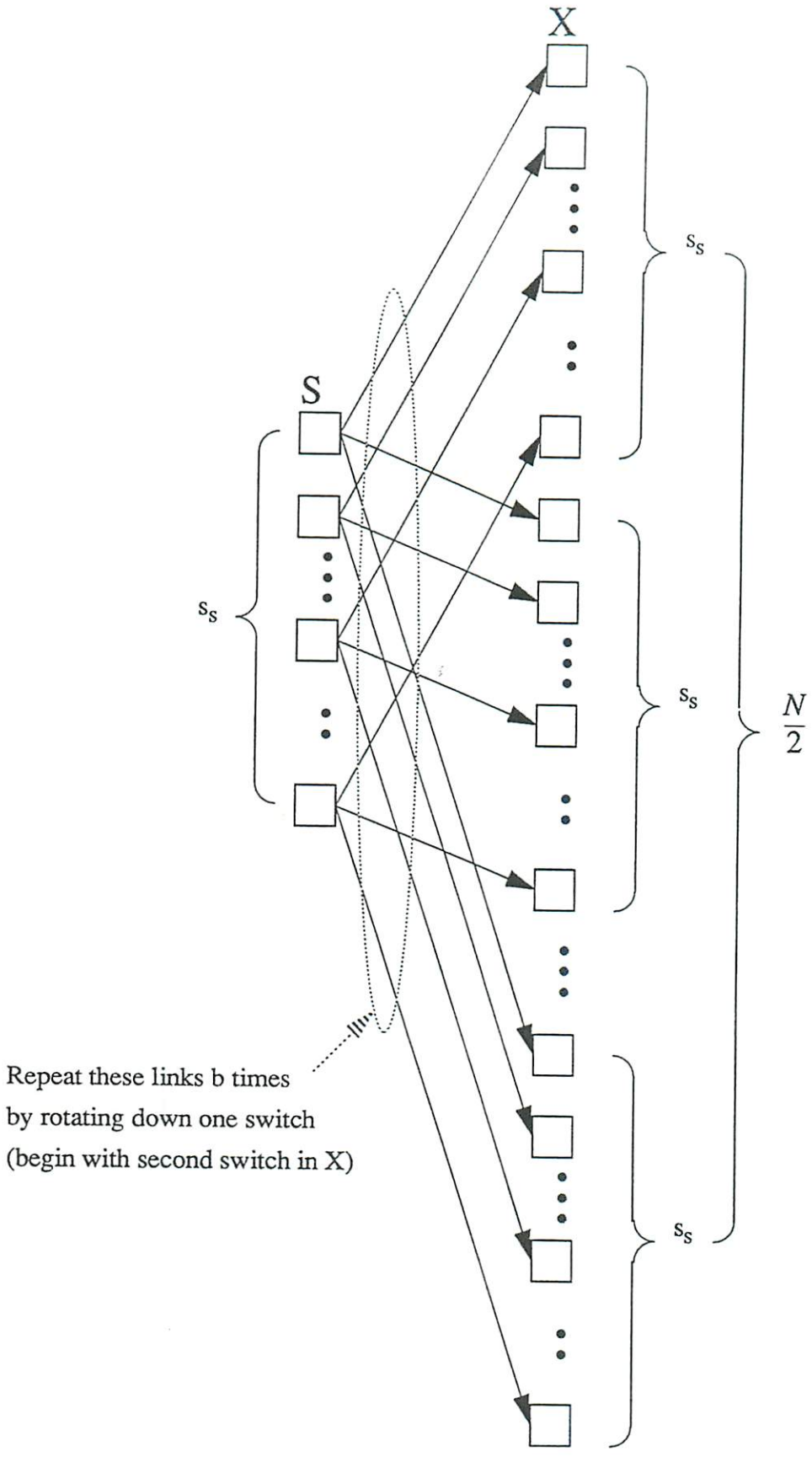


Figure 12. Replacement of spares for SARI / k: $ks_S \geq \frac{N}{2}, \quad s_S \leq \frac{N}{2}$

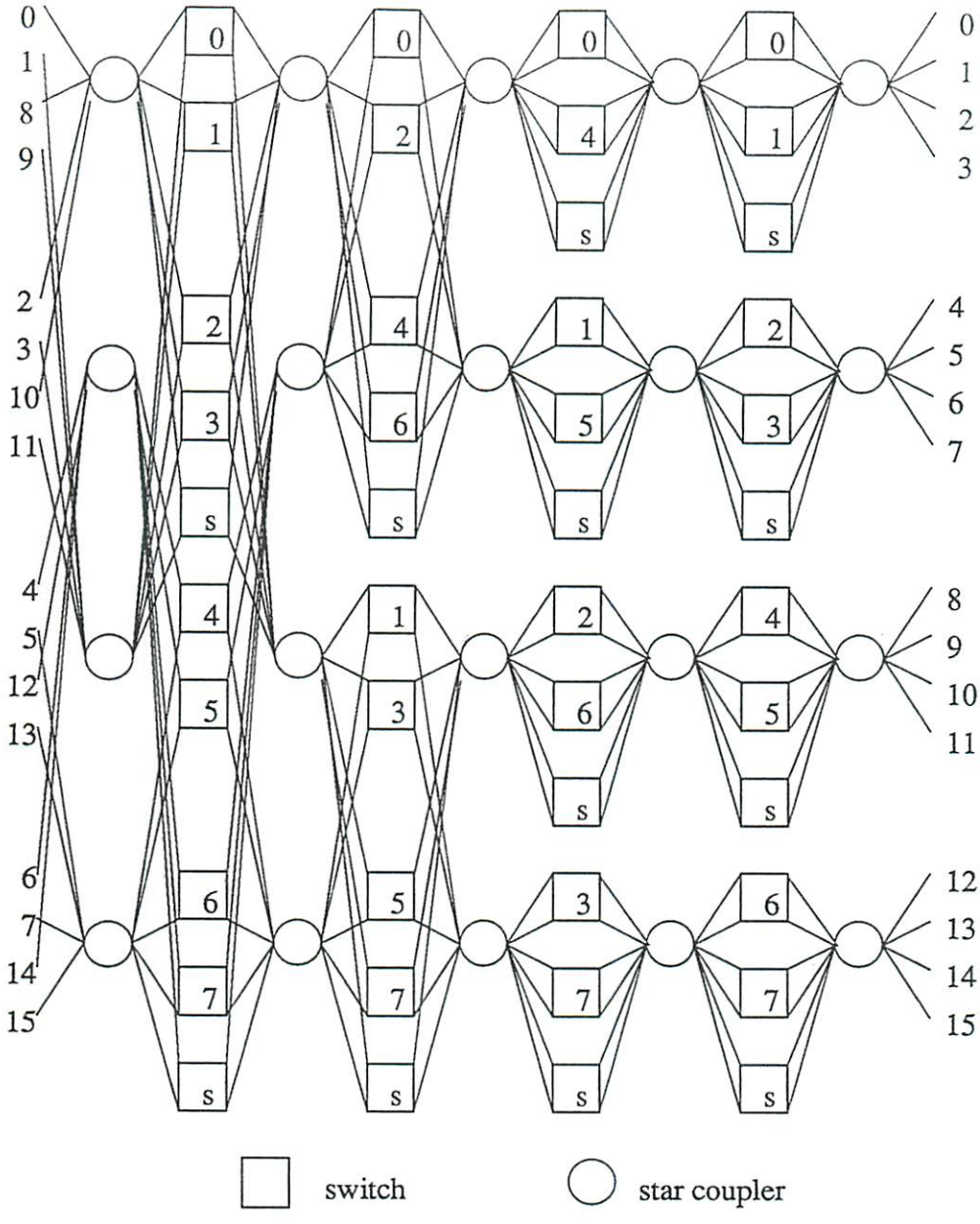


Figure 13. Star CouplerInterconnections for the Omega Network

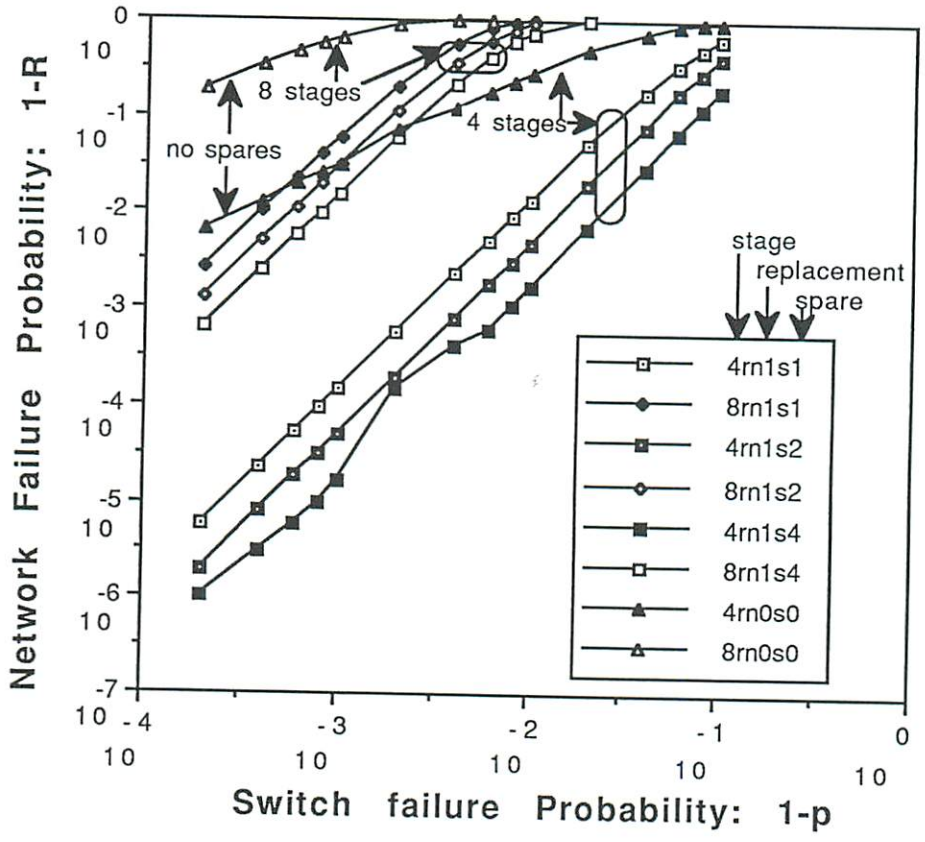
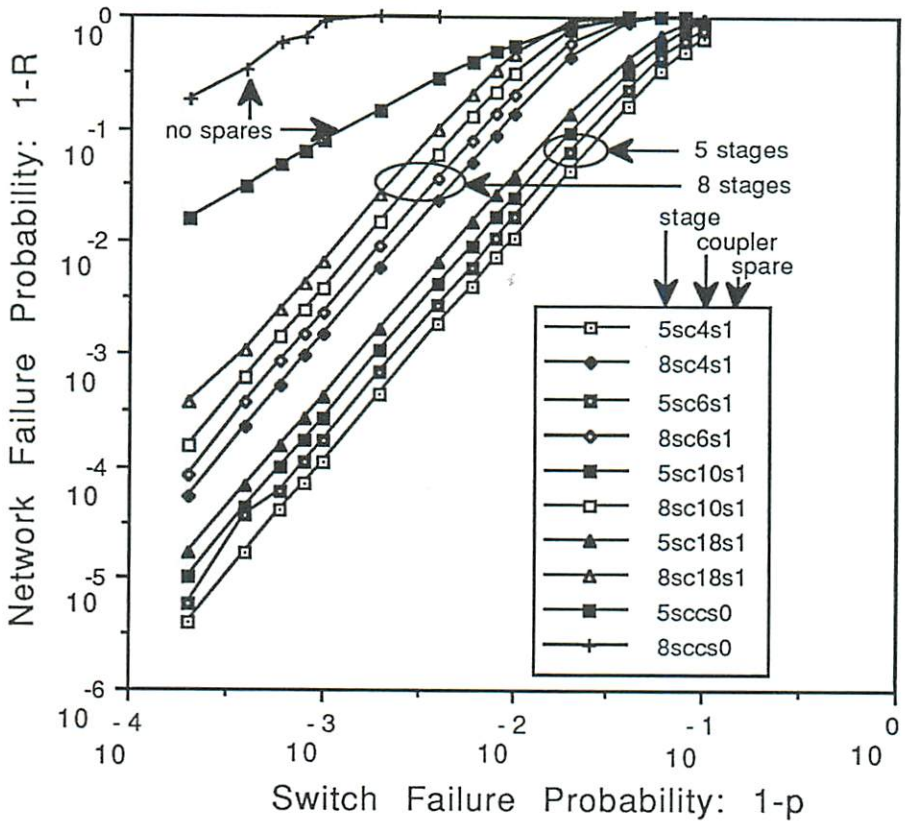


Figure 14. Network Reliability for SARI/1

Figure 15 Network Reliability for SCI



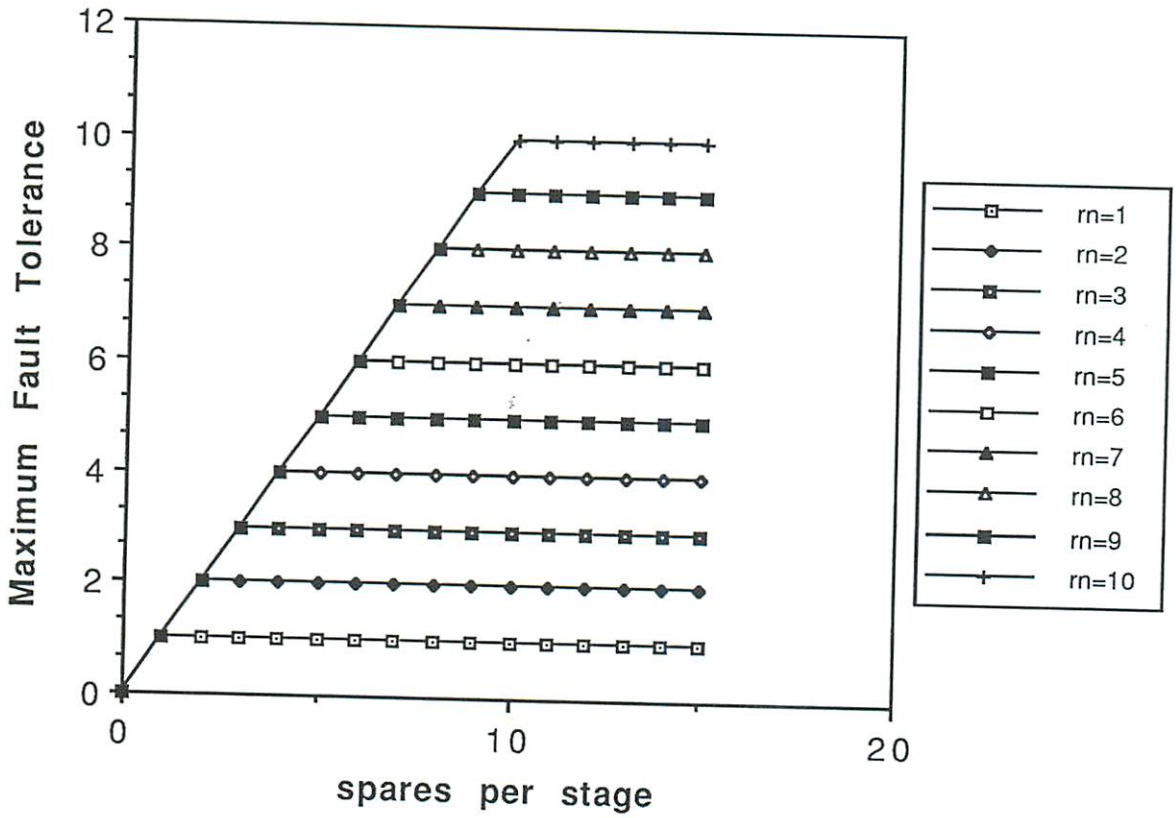


Figure16. Maximum Fault Tolerance for SARI/1, 5 stages