Chapter 4

Multi-Stage Interconnection Networks

The general concept of the multi-stage interconnection network, together with its routing properties, have been used in the preceding chapter to describe the operation of various designs of fast packet switch. In this chapter those networks that bear particular relevance to applications within the field of fast packet switching will be described in some detail within the context of interconnection networks in general.

4.1 An Introduction to Interconnection Networks

A number of useful general surveys of interconnection networks have been published, notably [99, 47, 95] with [132, 143] also being of some relevance. Much of the early work on interconnection networks was motivated by the needs of the communications industry, particularly in the context of telephone switching. With the growth of the computer industry, applications for interconnection networks within computing machines began to become apparent. Amongst the first of these was the sorting of sequences of numbers, but as interest in parallel processing grew, a large number of networks were proposed for processor to memory and processor to processor interconnection [131]. With the advent of the fast packet switch, interest in interconnection networks has turned full circle in that many of the networks originally proposed for parallel processing are now being considered for use in fast packet switch designs.

A simple classification of interconnection networks according to topology will first be offered followed by some comments on the control mechanisms employed with the various classes of multi-stage network. A discussion of the blocking characteristics of the various networks will then lead into a detailed discussion of some of the multi-stage interconnection networks useful in communications applications.

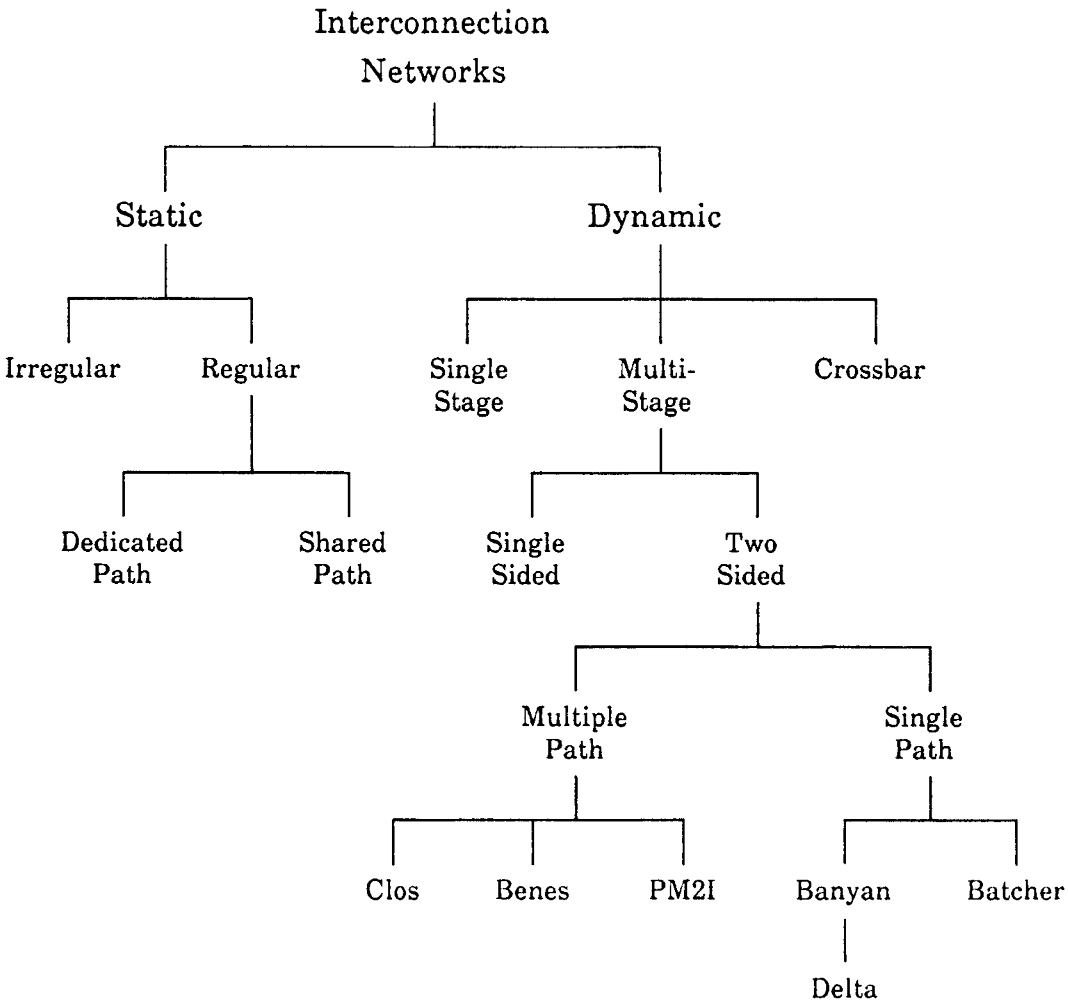


Figure 4.1: A simple classification of interconnection networks.

Topology

A simple, general classification of interconnection networks is presented in fig. 4.1. Regular, static networks, also called dedicated networks, are mostly used to interconnect loosely coupled processors to form parallel processing machines while any general packet switching network may be classed as an irregular static network. Two simple examples of dedicated path static networks are given in fig. 4.2 in which processing elements are connected by point-to-point links. Shared path static networks are formed by interconnecting processing elements with buses. In general the use of regular static networks has been restricted to the packet switched interconnection of loosely coupled processors as the delay across the network is dependent upon the distance between the communicating nodes. Also the processing delay required by the routing algorithm may render the use of short packets inefficient. Further, regular static structures often prove difficult to expand to large networks whilst maintain-

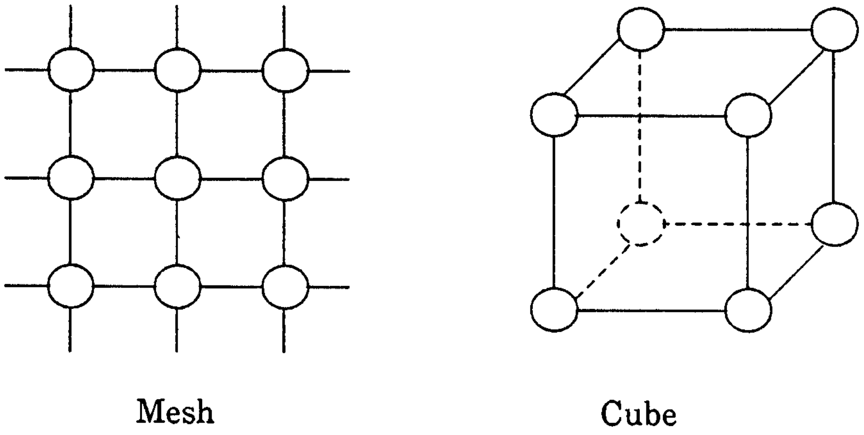


Figure 4.2: Examples of regular static network topologies.

ing the regularity of the structure. One notable exception is the use of a regular mesh topology in the Manhattan Street Network [98] which has been proposed as a metropolitan area network.

Dynamic interconnection networks are so called because the network clients are interconnected through an array of simple switching elements. Thus the pattern of interconnections between clients may be rapidly changed either by a centralised processor or by a distributed algorithm.

In [137] Stone introduced the perfect shuffle as a pattern of interconnection links of some interest in the solution of a number of classes of computational problem via a tightly coupled parallel processor. (The term 'perfect shuffle' derives from analogy with the process of shuffling a deck of cards in which the deck is divided into halves and re-assembled by alternately taking one card from each of the halves.) A single stage implementation is illustrated in fig. 4.3 comprising the perfect shuffle pattern of interconnection links followed by a single stage of switching elements. To complete the network every output is buffered and fed back to its corresponding input. Packets of data therefore circulate through the structure until they exit at the desired output [24].

If multiple copies of the single stage shuffle exchange are cascaded a multi-stage interconnection network results sometimes called a multi-stage shuffle exchange. Data is no longer required to circulate through the network but passes through the structure from the input side to the output side. Networks which have separate input and output sides are called two-sided, they are of great interest to communications applications and a number of examples will shortly be discussed. Single sided, multi-stage interconnection networks are also possible. Fig. 4.4 illustrates a single sided Clos structure. Both switches and links are bi-directional and all connections to the network may act as inputs and outputs. The TDM bus fast packet switch [36, 35] suggests the use of a single sided network but most fast packet switch designs use a two sided multi-stage interconnection network.

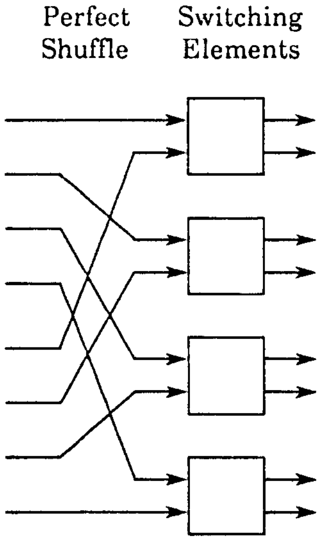


Figure 4.3: A single stage 8×8 shuffle exchange.

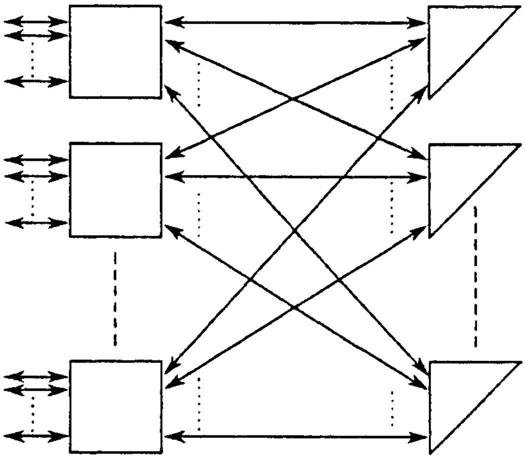


Figure 4.4: A single sided Clos network.

Concluding the discussion of the classification of networks introduced in fig. 4.1, the two sided multi-stage networks may be classed as either single path or multi-path. As might be expected, in a single path network only one unique path exists between any input/output pair but a choice of paths is available in a multi-path network. The Batcher sorting network has been included as a single path network because, for any given permutation of input to output requests, no choice of paths exists through the network. All of the examples of two sided multi-stage networks listed in fig. 4.1 will be discussed later with the exception of the PM2I class of networks [47, 99, 131]. This is also known as the data manipulator class of networks. It has a small number of multiple paths but nowhere near as many as the Clos or Beneš networks. Its major use is in manipulating sets of data within a tightly coupled parallel processor although it has been suggested for use in the concentrator function of the Starlite switch [70]. It is not easily partitioned for VLSI implementation, it is less flexible, and its routing algorithm is more complex when compared to other multi-stage interconnection networks.

Control Mechanism

Interconnection networks may also be classified according to the control mechanism employed to effect connections between input ports and output ports. If the algorithm is centralised and implemented in a central processor then the state of all existing connections and all connection requests may be consulted in order to make the necessary routing decisions. The use of a centralised control mechanism implies circuit switching where the holding time of a connection is much greater than the time required to establish connection. The vast majority of modern telephone switch designs use centralised control.

In fast packet switching applications the control mechanism must be distributed across the switch fabric and must be capable of operating without access to information regarding the entire state of the switch. Three classes of distributed routing algorithm are relevant to a regular network: source routing, self-routing and regular routing. Source routing requires a tag to be prefixed to the packet which specifies all of the routing decisions to be taken within the network, one field of the tag for each switch in the path. It thus removes the burden of route computation from within the switch fabric to the periphery. The self-routing and regular routing control mechanisms are sometimes confused as both require a tag to be prefixed to the packet specifying the required destination output port number and both rely upon the regularity of the interconnection network. Self-routing applies to dynamic, multi-stage interconnection networks. It may be implemented such that each switching element within the path makes a simple routing decision based only upon the tag of the incoming packet independently of the position of the switching element within the interconnection network. The regular routing mechanism applies to regular static networks in which each network node makes a routing decision based upon the packet tag and the position of the node within the network. This decision requires a certain amount of computation and thus involves some delay. The regular routing algorithm is thus best suited to conventional packet switching applications. The routing decision in a self-routing algorithm, however, requires no computation, does not involve the maintenance of routing tables within the switch fabric, and may be executed by very simple hardware within a single bit time. It is therefore of considerable interest in fast packet switching applications.

Blocking Characteristics

Multi-stage interconnection networks may be further classified according to the blocking characteristics they present which is reflected in the throughput they offer to traffic

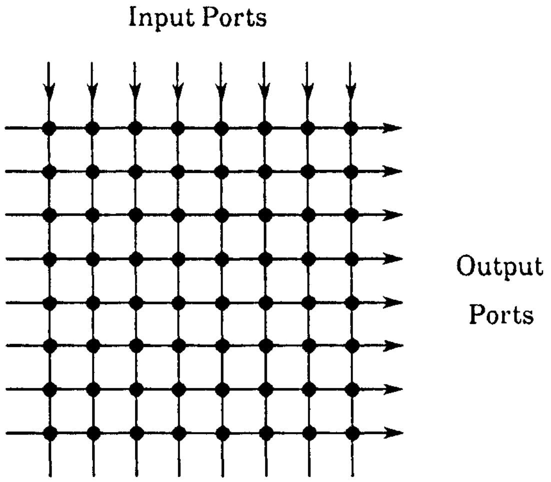


Figure 4.5: The general representation of a crossbar network.

with a random distribution of packet destinations. A network which is always capable of connecting a free input to a free output, regardless of the connections already established across the network, is said to be non-blocking. The crossbar and Clos networks are examples of non-blocking networks. A network that is always capable of connecting a free input to a free output, but which may require existing connections to be rearranged in order to do so, is called rearrangeable non-blocking. The Beneš network is an example of a rearrangeable non-blocking network. A network is said to be blocking if any free output may be unavailable to any free input because existing connections prevent a path from being established across the network. The banyan network is blocking to traffic with a random destination distribution.

As might be expected a non-blocking network requires more switching elements and interconnections than does a rearrangeable non-blocking network which in turn requires more than a blocking network. Also the throughput of a fast packet switch depends upon the blocking characteristics of its switch fabric. Finally a rearrangeable non-blocking network only provides non-blocking performance if a centralised control algorithm is available to perform the rearrangement of connections. For fast packet switching applications only distributed algorithms may be employed. It is therefore interesting to consider the improvement in performance that a rearrangeable structure might offer, for various distributed control algorithms, when compared to a blocking network. Three such algorithms will be examined in chapter 6.

4.2 The Crossbar Network

The crossbar network, or more often crossbar switch, is a non-blocking interconnection network that derives its name from a particular switch implementation developed for analogue telephone switching applications. The name is often taken to refer to non-blocking networks in general. Its structure is usually represented as shown in fig. 4.5.

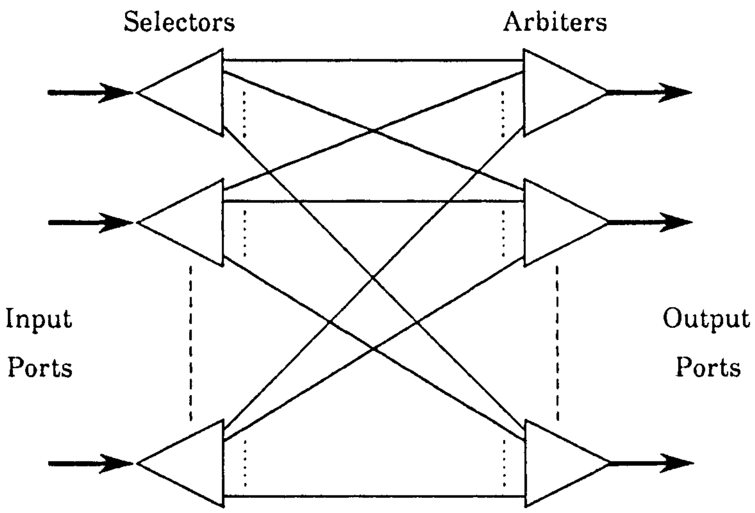


Figure 4.6: A self-routing crossbar switch.

Each node in the network is called a crosspoint and is a simple switch which has two states, open and closed. Using centralised control the network can satisfy all one-to-one (unicast) and one-to-many (multicast) connections. It requires $O(N^2)$ crosspoints and thus the hardware required to implement the network grows rapidly with the size of network.

Multi-stage interconnection networks are constructed from stages of interconnected crossbar switching elements of low degree. Fig. 4.6 illustrates an alternative design of crossbar switching element suitable for use with distributed control within the environment of a multi-stage interconnection network. An incident packet is prefaced by a tag indicating the required destination. The selector of the input port examines this tag and inspects the state of the arbiter of the required output port. If the selected arbiter indicates that the required output port is free the connection is established but if busy it is refused. All selectors may thus work concurrently and asynchronously. Multicast connections are not supported by this design of crossbar network. This crossbar network may be used with either a self-routing or with a source routing control algorithm and may be referred to as a self-routing crossbar switch.

4.3 Banyan Networks

Definition

The banyan network is a multi-stage network of interconnected crossbar switching elements originally defined in graph theoretic terms in [57] and named after the East Indian fig tree whose structure it is supposed to resemble. It is defined as having one and only one path from any input to any output and thus covers a very large class

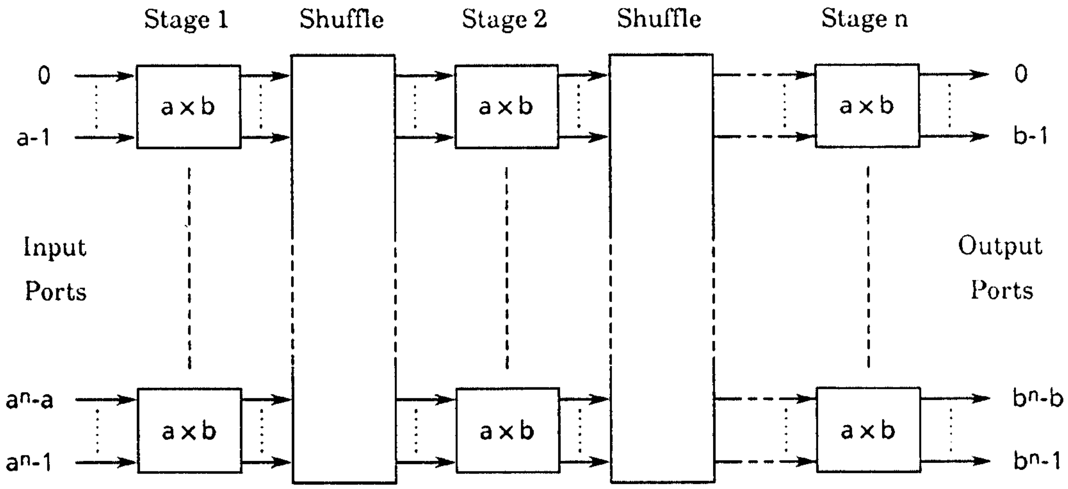


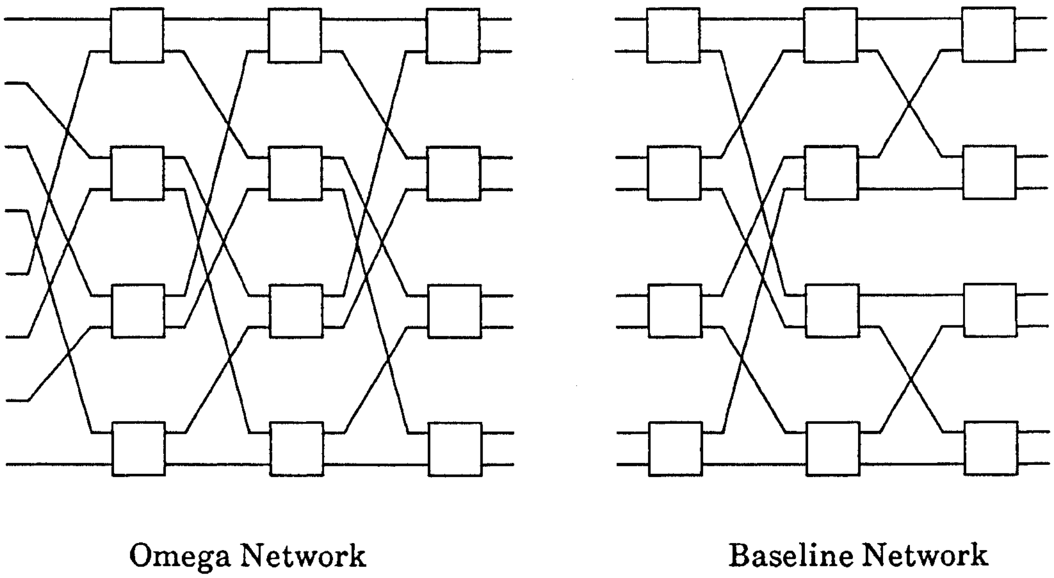
Figure 4.7: The general structure of a delta network.

of possible network structures. If the links in the banyan network are constrained to connecting switching elements in adjacent switching stages an L-level banyan results and if in addition all switching elements in the network are identical we get a regular banyan. A banyan network with square switching elements, i.e. those which have the same number of inputs as outputs, is called rectangular. Two classes of regular banyan are of specific interest, SW and CC banyans. The CC-banyan is rectangular by definition and its structure bears some resemblance to the PM2I class of networks. The SW-banyan can be shown to be self-routing and as such the rectangular SW-banyan constructed from 2×2 crossbar switching elements is almost invariably the network envisaged when the term 'banyan' is used in the literature of recent years. This is unfortunate as the term 'banyan', as originally defined, covers a much wider class of network.

Shortly after the definition of banyan networks the omega network was introduced [86] which formed a multi-stage interconnection network from the single stage shuffle exchange of [137]. The omega network was the first multi-stage interconnection network to demonstrate the self-routing property. A number of similar networks followed the omega network until in [126] the delta network was defined.

Delta Networks

The delta network is a multi-stage interconnection network which may be defined as a subset of the class of regular banyan networks that displays the self-routing property. Delta networks therefore form a class of interconnection networks which includes the SW-banyan, the omega network, the flip network, the indirect binary n-cube, the baseline and the reverse baseline networks which have all been proven topologically equivalent in [158]. The general structure of a delta network is given in fig. 4.7 in which stages of identical, but not necessarily square, switching elements are connected by



Omega Network **Baseline Network**

Figure 4.8: Examples of 8×8 delta networks constructed from 2×2 switching elements.

an interconnection pattern of links sometimes referred to as a permutation network but also called a shuffle. Hence the term 'multi-stage shuffle exchange' which also refers to this class of networks. In general the shuffle between each stage of the delta network will be different and it is this pattern of interconnection links that gives rise to the self-routing property. A delta network of some interest is the rectangular network in which the shuffle between each stage is identical. This in fact forms the omega network and if constructed from 2×2 switching elements the shuffle becomes the well known perfect shuffle. Two examples of an 8×8 delta network, constructed from 2×2 switching elements are given in fig. 4.8, the omega network and the baseline network [158] and an example of a 64×64 delta network constructed from 8×8 switching elements may be found in fig. 5.3.

A rectangular delta network of size $N \times N$ constructed from square switching elements of degree d requires $\log_d N$ switching stages with N/d switching elements per stage. It thus requires $O(N \log N)$ switching elements. An identical shuffle may be used for the interconnection pattern of links between each switching stage for networks built from any degree of switching element and may be constructed as follows. Label the N output ports of stage k from 1 to N. Label the switching elements of stage $k+1$ from 1 to N/d and the input ports of each of these switching elements from 1 to d. Connect the output ports 1 to N/d of stage k to input port 1 of each of the switching elements 1 to N/d in stage $k+1$. Connect the output ports $N/d+1$ to $2N/d$ of stage k to input port 2 of each of the switching elements 1 to N/d in stage $k+1$ and so on, a total of d times, until all N output ports of stage k are connected to all N input ports of stage $k+1$.

The general delta network is formally proven to be self-routing in [126] but for a rectangular delta network of crossbar switching elements of degree d it functions as follows. The required destination port number is expressed numerically to the base d and will require $\log_d N$ digits, one digit for every stage of the delta network. The required output port number is prefixed to the packet as a tag and the packet inserted into the network. The most significant digit of the tag is used by the switching element in the first stage of the network to select the output port over which to transmit the packet. The second digit in the tag is used by the switching element in the second stage, and so on for each stage of the network until the least significant digit of the tag is reached at the last stage of the network. Each digit may be removed from the tag as it is used by a switching element or the whole tag may be rotated such that the digit at the front of the tag is always the one required by the next stage of switching elements. Each switching element merely selects the output port specified by the digit at the head of the tag. The pattern of interconnection links between stages in the network is so arranged that the packet will exit from the correct destination port provided that it is not blocked anywhere within the network. Each switching element within the network functions as a simple self-routing crossbar switch and the packet will exit from the correct port regardless of the input port of the delta network at which it originated.

Delta networks have been modified to introduce multiple paths through the network to increase the reliability or to enhance the throughput. The proposed methods include adding extra paths with switching elements of higher degree [87]; adding extra stages [124, 1, 100]; or by connecting multiple delta networks in parallel [85, 82].

4.4 The Clos Network

The Clos network was developed to satisfy the needs of the telephone switching industry for a non-blocking network that uses fewer crosspoints than the crossbar network of the equivalent size. It is a multi-stage interconnection network of crossbar switching elements and a square Clos network is illustrated in fig. 4.9. Clos has shown that the network is strictly non-blocking if the condition $m \geq 2n - 1$ holds [30]. The Clos network may be recursively decomposed into a five stage network, a seven stage network and so on by replacing each switch in the central stage by a three stage Clos network. The three stage non-blocking Clos network has fewer crosspoints than the equivalent crossbar network for all $N \geq 36$ and a growth of $O(N(\log N)^{2.27})$ crosspoints.

4.5 The Beneš Network

The Beneš network is a special case of the Clos network for which Beneš has shown that if $m \geq n$ the network is rearrangeable non-blocking [13]. If $n = m$ then for networks of size $N = n^j$, where j is an integer, the $r \times r$ switches of the central stage of the Beneš network may be substituted by three stage Beneš networks recursively until a structure results in which all switching elements are of degree n. An 8×8 Beneš network of 2×2 switching elements is illustrated in fig. 4.10 and by comparison

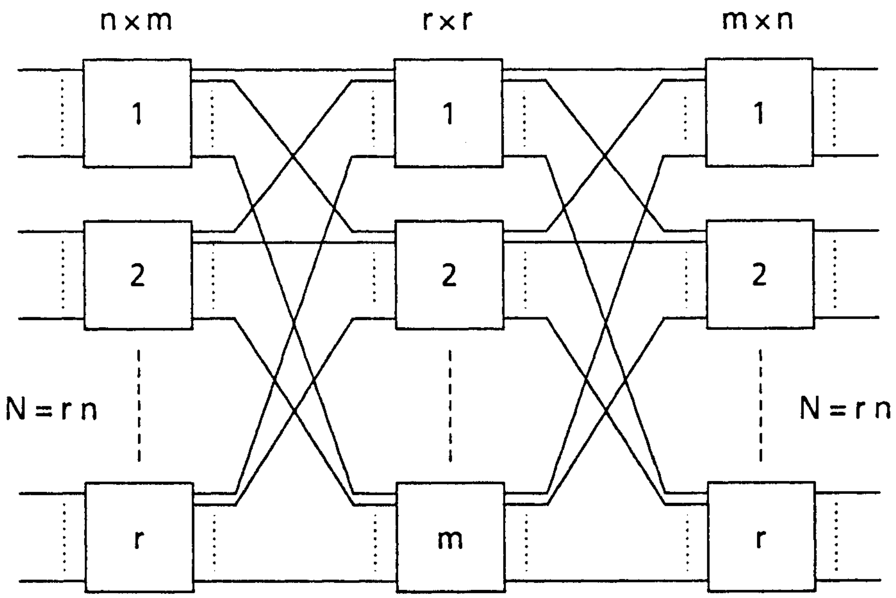


Figure 4.9: A square three stage Clos network.

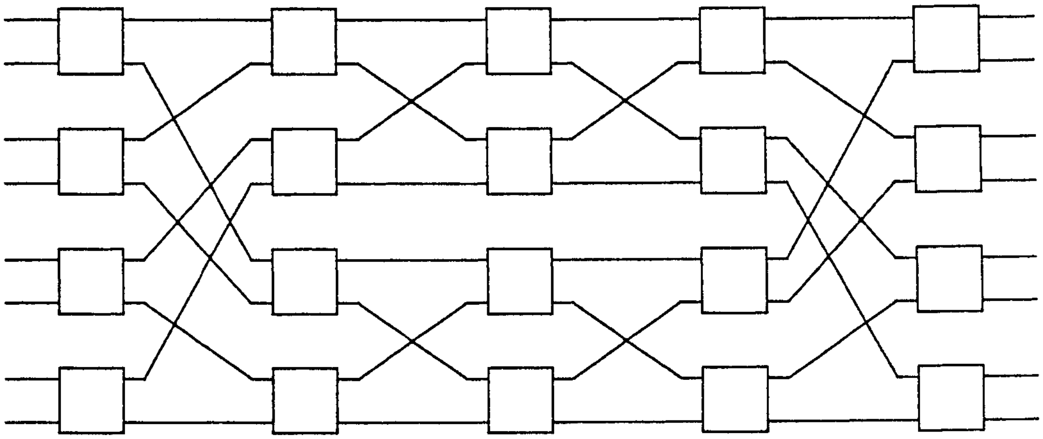


Figure 4.10: An 8×8 Beneš network constructed from 2×2 switching elements.

with fig. 4.8 it may be seen that the Beneš network may be formed by reflecting the equivalent baseline network about the central stage. An $N{\times}N$ Beneš network requires $2\log_d N - 1$ stages of switching elements of degree d, (i.e. $n = m = d$), with N/d switching elements per stage and also has a growth of $O(N\log N)$.

Only the final $\log_d N$ switching stages are required to provide the routing function in the Beneš network thus the additional $\log_d N - 1$ stages offer multiple paths through the network. Indeed the Beneš network may be considered as a delta network preceded by switching stages which distribute the incident traffic across the switch fabric making use of the multiple paths to offer fault tolerance and to reduce

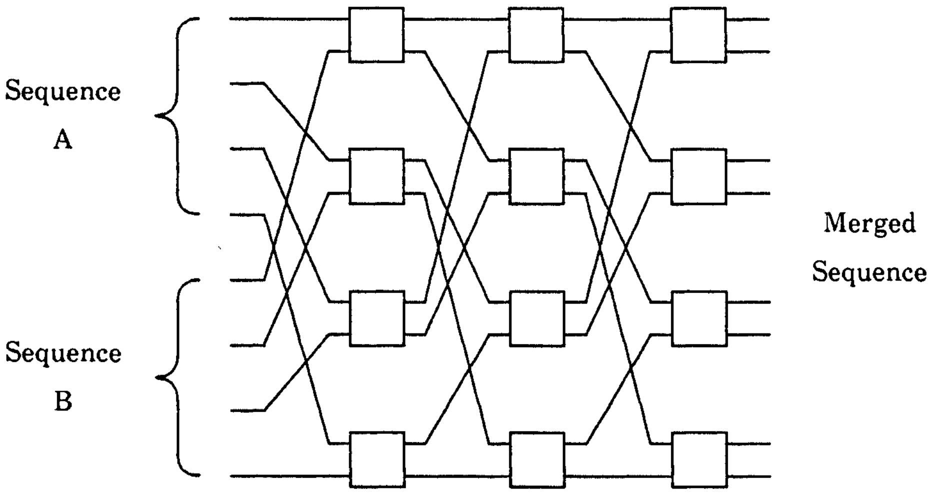


Figure 4.11: An 8×8 bitonic sorter.

blocking. If a centralised control algorithm is employed the network may operate in a rearrangeable non-blocking mode but the algorithm required is both time consuming and centralised.

Various options are possible for the use of a distributed algorithm across a Beneš network. The routing stages of the network may use the same self-routing algorithm as for the delta network, based upon the use of a destination routing tag. The distribution stages of the network may be switched according to three possible algorithms: source routing, random routing or flooding. In source routing the tag is extended to explicitly direct the switching of the distribution stages in exactly the same manner as the routing stages. If one path proves busy then another is selected and attempted from the periphery of the switch fabric. The random routing algorithm allows the distribution stages of the switch fabric to select any free path to the subsequent switch stages at random. The flooding algorithm sends copies of the incident packet concurrently across all free paths that lead to the required destination in the knowledge that only one path to the destination will be accepted. All other copies will fail and will quickly be removed from the network. All three algorithms have been investigated and results are presented in chapter 6.

4.6 The Batcher Sorting Network

A Batcher network will sort any arbitrary sequence of numbers into ascending (or descending) order [12]. It is constructed from a network of bitonic sorters each of which is a multi-stage interconnection network constructed from 2×2 comparison elements. A comparison element receives two numbers synchronously and in bit serial form at its two input ports and outputs the greater number on one output port and the lesser on the other. An 8×8 bitonic sorter is shown in fig. 4.11 which takes

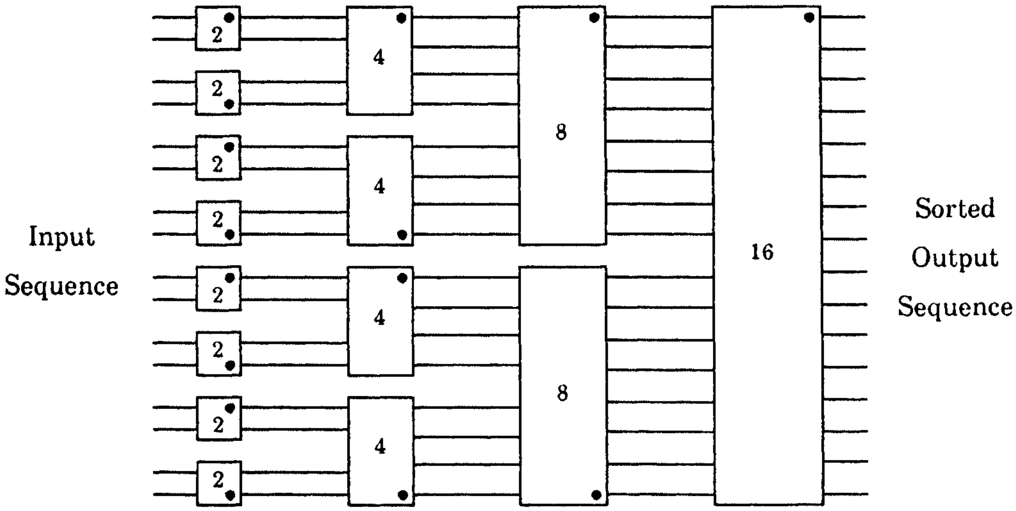


Figure 4.12: A 16×16 full sorter.

two monotonic sequences of numbers, one ascending and the other descending, and merges them into a single sorted sequence. All comparison elements must be aligned to sort the higher number into the upper (or lower) port to produce an ascending (or descending) sequence. The topology of the bitonic sorter is exactly that of the omega network for 2×2 switching elements and the interconnection pattern of links between each stage is the perfect shuffle.

To form a full sorting network a multi-stage network of bitonic sorters must be arranged as shown in fig. 4.12. Each element is a bitonic sorter of the size indicated, with the dot signifying the port at which the highest number of the output sequence will exit. The full sorting network requires $\frac{1}{2}\log_2 n(\log_2 n + 1)$ stages of $n/2$ comparison elements per stage and has a growth of $O(N(\log N)^2)$. The comparison elements are, however, relatively easy to construct.

In switching networks the full sorter is used to sort according to the tag at the head of each packet and the state of the network is latched while the packets are transmitted across the network. The Batcher network is of interest in fast packet switching applications because a non-blocking network may be formed by a Batcher sorting network followed by a banyan routing network. Further details on the construction of sorting networks, with reference to their use in a Batcher-banyan switch fabric, may be found in [70, 34, 107].

4.7 Summary

The most appropriate interconnection network for applications within the switch fabric of a fast packet switch is the two sided multi-stage interconnection network. It supports a simple distributed self-routing control algorithm and the distance between all inputs and outputs is constant. Such networks may be classified according to their performance into non-blocking, rearrangeable non-blocking and blocking networks. The non-blocking network offers ideal performance and may be constructed from a crossbar network, a Clos network or a Batcher sorting network followed by a banyan routing network (the Batcher-banyan switch fabric). The Beneš network is rearrangeable non-blocking but its non-blocking properties may only be attained by the use of a centralised control algorithm. It offers multiple paths and its performance under various distributed control algorithms is of interest for fast packet switching applications. The delta network forms a subset of the more general class of banyan networks that offers the self-routing property. It is the simplest of the self-routing multi-stage interconnection networks discussed but offers blocking performance.