Chapter 2

Statistical Switching Mechanisms

In this chapter a review of the available switching mechanisms is presented and it is argued that a statistical switching mechanism is best suited to the requirements of the high capacity switching of multi-service traffic. Fast packet switching is selected for further study largely on the grounds of its flexibility and the development of the fast packet switch is traced. Some of the fundamental characteristics common to all designs of fast packet switch are then introduced.

2.1 Multiplexing

Multiplexing is the technique whereby two or more separate communications channels are supported across a single transmission medium. A well known example from the telephone network is the support of multiple telephone conversations on a single high bandwidth trunk [22]. Early multiplexing systems for use in the analogue telephone network employed frequency division multiplexing (FDM) in which each separate channel was transmitted at a different carrier frequency. An analogous technique currently being developed for use in optical communications systems is that of wavelength division multiplexing (WDM) in which the various channels are carried on different optical wavelengths. In digital communications systems by far the most common form of multiplexing is that of time division multiplexing (TDM). In this technique the entire capacity of the shared transmission medium is allocated to each source in turn for a short duration sufficient for the source to transmit a brief burst of information of fixed length. As an example, a current European TDM transmission standard employs a 2.048 Mbits/sec digital carrier divided into frames of length 125 μsec, fig. 2.1. Each frame is divided into 32 timeslots each of length 8 bits (one octet). In every frame, timeslot 0 is used for synchronisation and maintenance purposes, timeslot 16 is allocated to signalling and all other timeslots may be allocated to traffic sources. When allocated a channel, the source is given the timeslot number and it fills the appropriate timeslot in every frame with 8 bits of data. Each channel

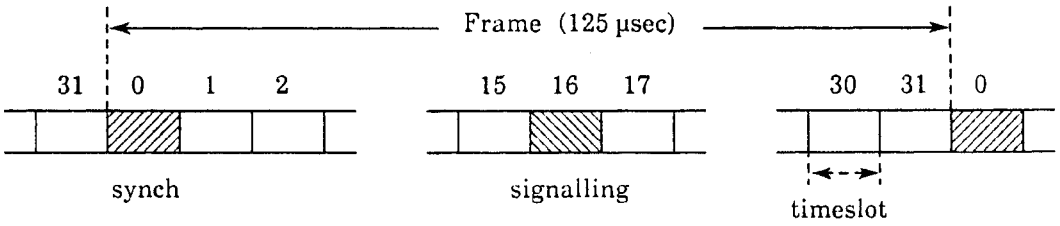

Figure 2.1: An example of time division multiplexing (TDM).

thus carries 64 kbits/sec of traffic.

Time division multiplexing offers channels of fixed bandwidth and is well suited to continuous traffic sources of fixed bit rate, e.g. 64 kbits/sec voice, but many traffic sources are bursty and offer an instantaneous bit rate that is widely variable. TDM is very inefficient in its use of bandwidth for bursty services or for variable bandwidth services so statistical multiplexing has been introduced to overcome this inefficiency. In statistical multiplexing the channels are no longer of fixed bandwidth but each source receives as much transmission capacity as it requires instantaneously. As in TDM, sources continue to transmit information in bursts but these bursts are not necessarily of equal length and sources may submit bursts of information in any order and at a rate that reflects the instantaneous bandwidth required. Sources generally queue for access to the shared transmission medium on a first come first served basis but some sources may be allocated priority. In TDM, when a bursty source is temporarily idle the bandwidth allocated to it is unused but is not available to other sources whereas with statistical multiplexing a bursty source only consumes transmission bandwidth when it has information to send. In conventional TDM the identity of every channel is implicitly conveyed in the position of its timeslot within the frame but with statistical multiplexing the identity of each channel must be explicitly prefaced to every burst of information. This additional overhead tends to require that the information bursts of statistical multiplexing be much longer than those of conventional TDM.

In conventional TDM the offered traffic load can never exceed the capacity of the shared transmission medium, a utilisation of 100% may be supported indefinitely, delay is deterministic and jitter is very low. With statistical multiplexing, however, for short periods the offered traffic load can exceed the capacity of the transmission medium which may result in loss of information, delay or both. Statistical multiplexing cannot support an average utilisation of 100% on the transmission medium and 80% is a maximum utilisation frequently quoted. Delay is dependent upon the mean traffic load and the source traffic characteristics and jitter may be high. Despite these apparent disadvantages, statistical multiplexing is very flexible, supports traffic sources which vary widely in their bandwidth requirements and source traffic characteristics and handles bursty sources efficiently [149, 50].

One proposal of statistical multiplexing for use in the asynchronous transfer mode of broadband ISDN has been termed asynchronous time division (ATD) [141, 31]. The

capacity of the shared medium is divided into short fixed length blocks called cells of 128 bits each. Cells are allocated to traffic sources statistically on demand and each cell contains a short header to identify the source. No framing is applied to the transmission medium but empty cells are filled with a synchronisation pattern by which synchronisation across the transmission link is maintained. A sufficient supply of empty cells is guaranteed as on average the transmission medium will not be utilised beyond about 80%. Another proposal named dynamic TDM (DTDM) retains the frame and timeslot structure of TDM, with each timeslot containing a single cell, but allocates the timeslots statistically as in ATD [161]. A more flexible TDM multiplexing strategy for optical fibre links is under consideration named SONET [15]. It is capable of supporting both synchronous transfer mode (conventional TDM) and asynchronous transfer mode (statistically multiplexed) payloads. In a similar manner the multiplexing strategy under consideration for the STM proposal for broadband ISDN is also capable of supporting both STM and ATM payloads [37].

2.2 Switching Mechanisms

Multiplexing allows the sharing of a high capacity communications link between many channels; but in order to achieve communication between source and destination across a network, a switching function is necessary. A range of switching mechanisms is available to accompany the various multiplexing techniques [84, 127, 25], from circuit switching for use with conventional time division multiplexing (or synchronous transfer mode) to packet switching which mates with the extreme end of statistical multiplexing (or asynchronous transfer mode). Between these two extremes lies a spectrum of available switching mechanisms illustrated in fig. 2.2 which is adapted from [84] and [37]. Switching mechanisms towards the left of the diagram offer channels with fixed bandwidth but a constant and small delay whereas those towards the right of the diagram offer variable bandwidth channels but with a variable delay which can be quite high [59]. Towards the centre of the diagram the statistical switching mechanisms attempt to provide the variable bandwidth required by bursty and variable rate sources but at low delay and low variance of delay compared with conventional packet switching.

Circuit Switching

Circuit switching is based upon the concept of a connection. A connection is an association between a source and its destination across a switched network. A connection may support communication in only a single direction or may offer both forward and reverse channels. A unicast, or point-to-point connection, is established between a single source and a single destination whereas a multicast, or distributive, connection may connect a single source to many destinations. Multicast connections will not be considered further until chapter 9. In circuit switched networks a communications channel of fixed bandwidth is exclusively allocated to a connection throughout the lifetime of that connection. A circuit switch, in general, connects input channels to

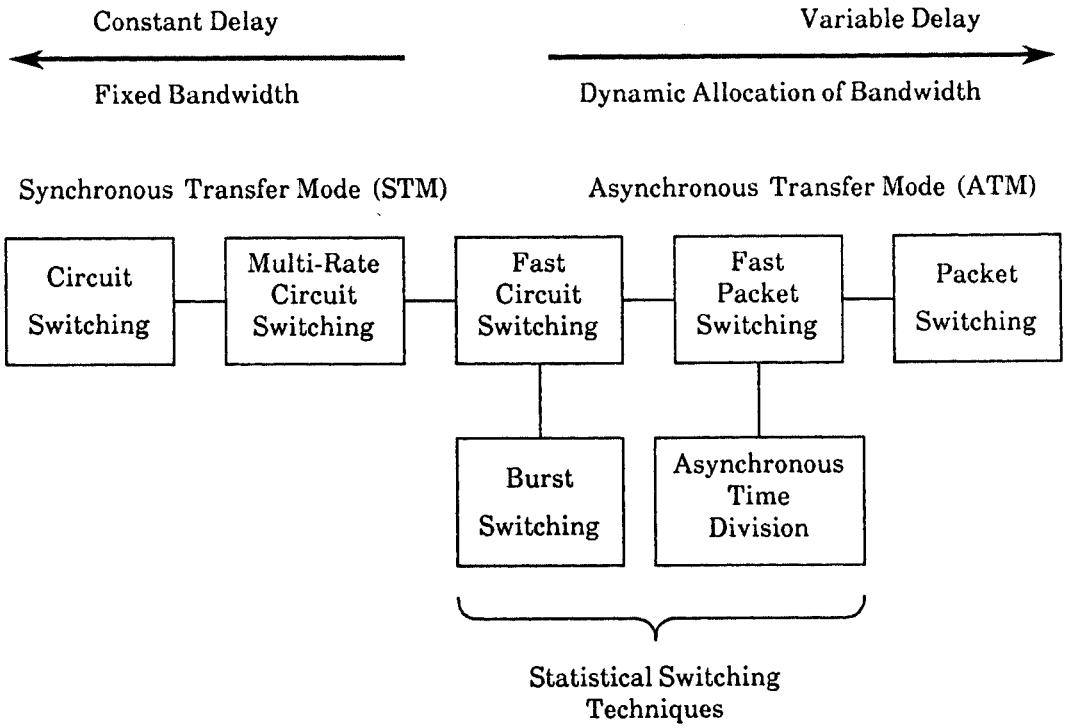

Figure 2.2: The spectrum of switching mechanisms.

output channels, (although the interconnection of bi-directional channels is also possible.) If each channel is presented to the switch on a separate transmission medium the switching is said to occur in the space domain and a connection is allocated a physical path across the switch [162, 105]. If all of the channels are presented to the switch in a TDM multiplex on a single incoming and outgoing transmission medium, see fig. 2.1, then switching occurs in the time domain. In this case a connection represents an association between a timeslot on the incoming link and a timeslot on the outgoing link. During each frame the relevant timeslot is copied from the incoming link into a buffer and thence to the required timeslot on the outgoing link. This requires the buffering of one complete frame and inserts a constant delay of up to one frame length in the connection. In general, a digital circuit switch will be required to interconnect a number of separate incoming and outgoing transmission links, each of them TDM multiplexed with a number of channels. Thus both time and space domain switching is involved [127]. Small switches of up to 1000 or so 64 kbits/sec channels may be implemented using shared memory whereas larger switches will require an interconnection network [160, 34].

Circuit switching offers channels of fixed bandwidth with a low and constant delay. It is transparent in that once a connection is established, the network takes no notice of the information it carries. In particular the network makes no attempt to correct transmission errors that may occur. Overload is handled by refusing to accept further connection requests once the system capacity is fully allocated. This means

that a connection once established will not suffer degradation of delay or bandwidth due to overload. Circuit switching is best suited to applications that require a fixed bandwidth, a low delay, and in which the call holding time is long compared to the call set-up time. It cannot effectively support the widely varying bit rates of many communications services at their natural rate. It does not exploit the burstiness of many forms of information and is inefficient for bursty traffic.

Multi-Rate Circuit Switching

Multi-rate circuit switching is a slight enhancement of circuit switching in that channels of different but fixed bandwidth may be formed by combining one or more integer multiples of some basic channel rate. Selecting the basic channel rate, however, poses a problem in order to satisfy the needs of both low and high bandwidth services. Multiple basic channel rates may be employed but this tends to complicate the design and control of the switch. Synchronising all of the basic rate streams that form a multi-rate channel also poses a significant problem as in general with a circuit switched network no guarantee is offered as to the relative delay between timeslots switched across the network. Neither is there any guarantee that any such delay will remain constant for the duration of a connection. The main disadvantage, however, is that multi-rate circuit switching does not handle bursty sources any more efficiently than does circuit switching.

Fast Circuit Switching

Fast circuit switching has been proposed as a means of handling bursty traffic. If the set-up of a connection across the switch is sufficiently fast, then a connection may be set up for each burst of traffic as it arrives and released immediately the burst ends. Thus the bandwidth of the switch is only allocated to active sources. The technique is similar to that of time assignment speech interpolation (TASI) [20, 153] which was a multiplexing technique used on an expensive analogue transmission link to allocate voice sources to transmission channels only during periods of speaker activity (or talkspurts). The method provided a significant increase in the number of voice sources that could be handled by a transmission link without substantial loss of quality provided a sufficient number of sources were multiplexed.

It is inefficient to set up the entire connection on the arrival of each burst of information, thus burst switching [6, 61] introduces the virtual circuit. A virtual circuit is a logical connection between source and destination which dissociates the concept of the connection from the bandwidth allocated to it. In burst switching a virtual circuit is set up at the beginning of a call which defines the connection but bandwidth is only allocated to that connection at the arrival of each burst of traffic. Buffering is also introduced so that if bandwidth is not available on the arrival of a burst it may be delayed until bandwidth becomes available. For bursts of voice traffic, information is discarded once a burst becomes delayed for longer than 2 msec as it is no longer of any use. As a burst is always transmitted at the same bit rate as

that at which it is received, there is no need to store the complete burst. It can be forwarded as soon as transmission bandwidth becomes available.

The interest in burst switching has so far proven somewhat limited with most of the work being undertaken by a single telecommunications manufacturer. Emphasis appears to be directed towards the switching of 64 kbits/sec voice in the presence of data traffic [122, 123]. The switching mechanism could be made more flexible if multiple channel rates were available for burst switching, as in multi-rate circuit switching, but this would complicate the design and operation of the switch.

Packet Switching

Turning to the other end of the spectrum of switching mechanisms we find conventional packet switching [140, 27]. In packet switching the bandwidth of the transmission medium is no longer divided into channels but the bandwidth of the entire medium is available to every burst of information from each source. Each information burst is constrained to a maximum length and additional fields of control information are added to identify source and destination and to support flow and error control etc. The resulting unit of information is called a packet. The maximum length of a packet is limited by the buffering requirements of packet switches and the packet delay requirements. It should not be too small, however, due to reasons of bandwidth efficiency as the overhead of control information can be quite considerable and is added to every packet. Packets are generally stored in every packet switch in the path and are not forwarded until completely received although suggestions such as cut-through and virtual cut-through [77, 73] have been made to forward packets before they are completely received. Error checking and flow control protocol operations are performed on a link-by-link basis between every packet switch in the path and error correction may be performed both by retransmission from the preceding switch in the path and also on an end-to-end basis.

Packet switched networks offer two fundamental modes of operation: connection-oriented and connectionless. In connection-oriented mode a virtual circuit is established across the path between source and destination. In general, all packets belonging to the same virtual circuit follow the same route across the network which means that the routing operation only has to be performed once when the virtual circuit is set up. The processing of subsequent packets may thus be simplified, the packet header may be simplified, and flow control may be applied more efficiently and selectively to virtual circuits. In connectionless operation each packet, called a datagram, is handled individually and bears enough control information to completely identify it, its source and its destination. Packets between the same source and destination may follow different routes and packets may not be guaranteed to arrive in the same sequential order in which they left. Connectionless operation requires more processing for every packet and flow control is less selective but it is less vulnerable to node failures and more easily adapted to changing traffic patterns. Connection-oriented mode is favoured by telecommunications administrations while connectionless operation is generally preferred by computer communications manufacturers.

Various experiments in supporting the voice service over wide area packet switched networks have been reported [154, 49] but the high delay and high variance of delay over such networks requires complex resequencing procedures to reconstruct the voice signal which themselves insert further delay [11, 104, 108]. In addition, the public voice service requires a number of very large switches both in the total traffic capacity and in the number of switch ports, the support of which is beyond the ability of current designs of conventional packet switch. The support of the voice service over local area packet switched networks has perhaps been more successful [39, 46, 94, 106, 7] but even here the large maximum packet length permitted in many local area networks can introduce a large variance of delay for the voice signal reconstruction algorithm to handle.

Packet switching offers a very flexible communications facility supporting any arbitrary data rate up to the full rate of the transmission medium by selecting the size of the packet and the frequency with which packets are sent. It is also very efficient for handling bursty services and does not consume switching or transmission bandwidth during the idle periods of a call. It responds very rapidly to variations in the bandwidth required by sources during the active phases of a call and can interconnect sources and destinations operating at different data rates. Due to the large amount of processing per packet at every switch, conventional packet switches in general offer a much lower maximum capacity than circuit switches of comparable complexity. They also suffer from high delays across the network and a high variance of delay and it is to answer these drawbacks that fast packet switching has been proposed.

Fast Packet Switching

Fast packet switching attempts to retain the flexibility of conventional packet switching while reducing the delay and increasing the maximum switch capacity to approach that offered by circuit switching [50, 149, 79, 146, 147]. Recent advances in optical fibre transmission technology provide very high bandwidth links with very low bit error rates. With a low error rate on each transmission link, error control is no longer required on a link-by-link basis at every switch in the path. Also, at high transmission rates it may prove impractical to attempt to provide the functions of flow control and error control on every link in the path due to delay and buffering requirements. Therefore, in fast packet switching, the functions of flow control and error control are implemented on an end-to-end basis, or on entry to and exit from the network [66]. Thus services that require error detection and correction may implement a retransmission strategy on an end-to-end basis whereas services, such as voice, that may tolerate a certain degree of error may take advantage of the low delay. As the protocol requirements of each switch are reduced, packets may be processed entirely in hardware. Thus switches of much greater capacity may be constructed and the switch may become more transparent to the data it carries than for conventional packet switching. Fast packet switching is in general connection-oriented. Thus once a virtual circuit is established across the network very short packet headers may be

used to distinguish between each of the virtual circuits multiplexed over a single link. Also the routing of each packet may be performed in hardware by table look-up. As the packet overhead has been significantly reduced, very short fixed length packets may be used to reduce the delay across the switch to levels comparable with that of circuit switching. Fast packet switching with short fixed length packets is often referred to as asynchronous time division (ATD) in the context of broadband ISDN.

Both fast circuit switching and fast packet switching offer statistical switching mechanisms that handle bursty traffic efficiently and are capable of supporting high capacity switch implementations. Fast packet switching requires a header on every packet whereas fast circuit switching requires a header only on every burst. Fast packet switching therefore carries a greater overhead, perhaps 10% of the available bandwidth or more in a typical application, but with high capacity optical fibre transmission links, bandwidth efficiency may not be the most critical parameter. In both forms of statistical switching, overload occurs when the incoming information exceeds the transmission capacity resulting in delay, loss of information or both. Fast packet switching, however, is able to spread the effects of delay and loss over all calls or over a selected class or classes of calls. With fast circuit switching the effect must be absorbed by at most a few selected calls and can thus result in more severe delay or loss effects. Fast packet switching is also able to vary the allocation of bandwidth to individual sources instantaneously and can thus allow much greater flexibility. Fast packet switching may also give a better performance for data traffic as end-to-end retransmissions are carried using the entire bandwidth of the transmission links rather than across the narrowband channels of fast circuit switching. Thus fast packet switching has been selected for further study partly because of its flexibility but also because a very simple design of fast packet switch was envisaged and considered to be worthy of detailed study (see chapter 5).

2.3 Evolution of the Packet Switch

Early Switch Architecture

In the early days of packet switching, computer processing power was an expensive commodity so packet switches were designed with a single central processor handling all of the switching, routing and protocol functions of the entire packet switch. Thus the throughput of the switch was limited by the processing capacity of the central processor and the complexity of the packet switching protocol. With the growth of VLSI technology the cost of processing fell rapidly until it became possible to provide some processing capacity on each switch port. Thus the lower level protocol functions, such as flow control and error detection and correction, could be handled independently by each switch port while the central processor provided higher level protocol functions such as routing. This increased the throughput by an order of magnitude, but as the central processor continued to interconnect all of the switch ports it remained a bottleneck.

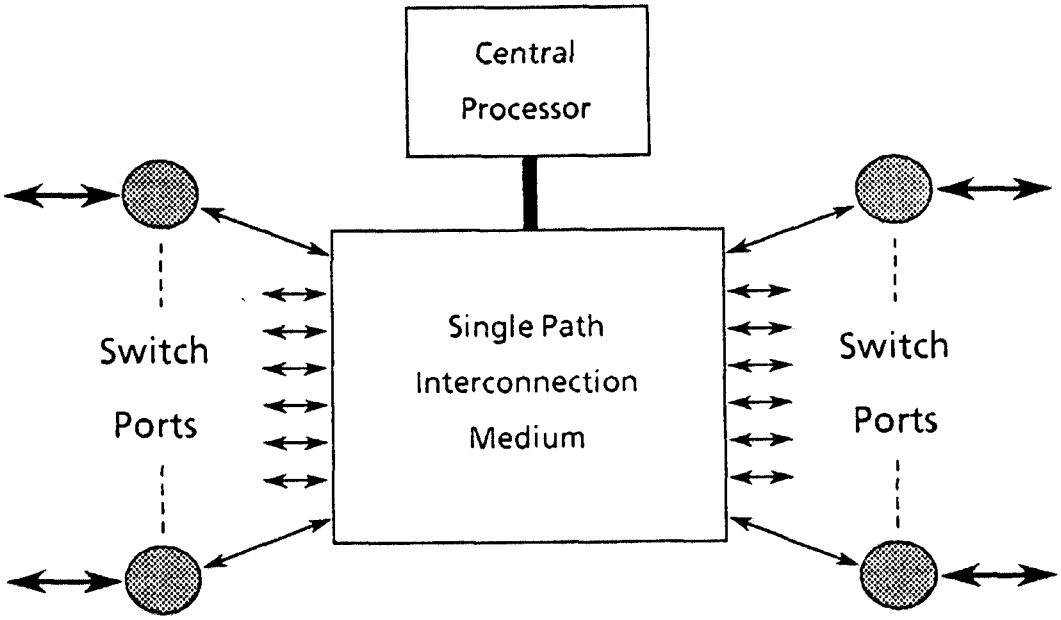

Figure 2.3: A single path decentralised packet switch.

The Single Path Decentralised Switch

To further improve the capacity of the packet switch it became necessary to remove the central processor completely from the transmission path of every packet. To achieve this, some form of single path interconnection medium was inserted to interconnect all of the intelligent, peripheral switch ports while the central processor took on more of a supervisory role, as illustrated in fig. 2.3. Hence, although some form of action may still have been required of the central processor on a per packet basis, it was decentralised by being removed from the task of physically transmitting each packet between the switch ports. The majority of packet switches have used shared memory as the single path interconnection medium with direct memory access in each of the switch ports but some designs have used serial bus [52, 26], parallel bus [16, 28] or ring [68, 67, 53] based structures.

When the majority of the per packet processing is removed from the central processor the throughput of the packet switch is determined by the bandwidth of the interconnection medium and the rate at which the processors in the switch ports can handle the protocol functions required. From an architectural perspective there is little difference between this class of packet switch and a local area network (LAN). The switch port of the packet switch corresponds to the media access controller of the LAN. The only major difference is that the switching function in the LAN is distributed across the local area. This requires a more complex media access protocol than for the packet switch for which access to the interconnection medium is contained within the confines of the switch. A parallel may also be drawn between this class of packet switch and digital circuit switches that handle up to about 1000 telephony

channels of 64 kbits/sec bandwidth. These also use a shared memory interconnection medium with a central processor that is in general only required at the set-up and clearing down of a connection.

Hybrid switch structures have also been proposed with a single path interconnection medium. These offer separate packet and circuit switching functions with integrated access and transmission facilities. Many such designs exist in the literature covering both discrete switches [80, 16, 151] and distributed switches, i.e. local area or metropolitan area network designs, both ring [18, 23, 138] and CATV bus [97].

The Multi-Path Switch

In considering switches of very high capacity, the bandwidth of a single path interconnection medium imposes a limit upon the switch capacity that may be achieved. To overcome this fundamental restriction it is clear that some form of multi-path interconnection medium is required that is capable of supporting communication between a large number of switch ports concurrently. Thus with a multi-path interconnection medium the total capacity of the switch is no longer limited to the bandwidth of the paths forming the interconnection medium but may grow as the number of switch ports increases. In this manner a much higher total switch capacity may be attained than for a single path interconnection medium using the same implementation technology. Conversely a high capacity switch no longer requires high speed and expensive device technology. The multi-path architecture applies equally to circuit, packet and hybrid switches. Circuit switches have used analogue multi-path switching networks for many years but more recently high capacity digital TDM circuit switches have been designed around a non-blocking, multi-path interconnection network [34, 160]. Hybrid multi-path switches have also been proposed [150, 139, 96] and the majority of current fast packet switch designs are multi-path switches, examples of which will be discussed in the following chapter. Many forms of multi-path interconnection medium are possible, e.g. multiple rings [139, 2], but the most general class, and the one which yields the highest switch capacities, is that of the multi-stage interconnection network which will be examined in detail in chapter 4.

2.4 Fundamentals of Fast Packet Switch Design

There are some basic concepts that are common to many designs of fast packet switch and these will now be introduced prior to the detailed discussion of existing fast packet switch designs presented in the following chapter.

A fast packet switch will in general consist of a set of input lines each arriving at an input port, a set of output lines each departing from an output port, with input and output ports interconnected via a switch fabric, fig. 2.4. A switch controller will also be interfaced to the switch fabric and may control the input and output ports either directly or via packets across the switch fabric. External connections to the switch are generally required in the form of bi-directional links which are formed by

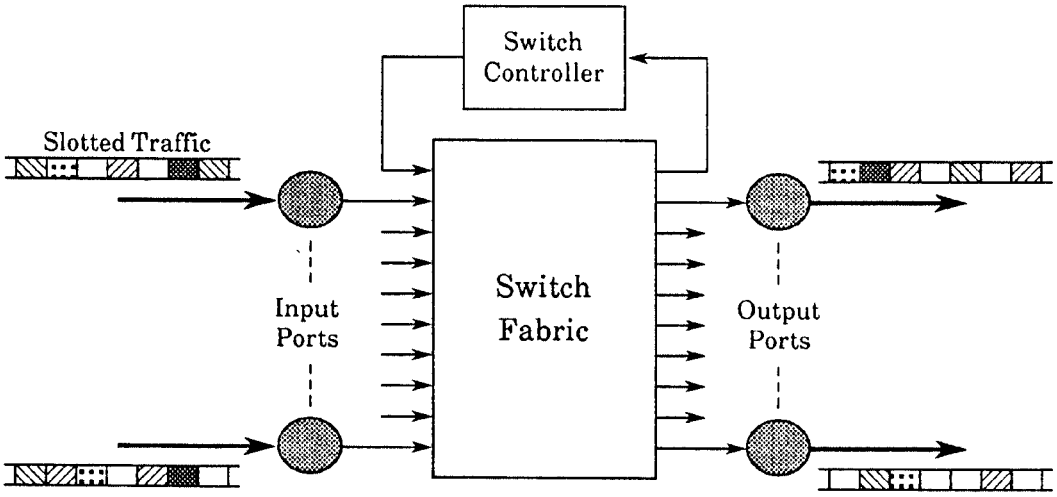

Figure 2.4: General structure of a fast packet switch.

grouping an input and an output line together. Many designs of fast packet switch are only capable of handling short fixed length packets. In such designs the bandwidth on both input and output lines is divided up into timeslots each of which may carry a cell (packet) or may be empty. All lines must be synchronised and this is accomplished either by means of a frame structure with a synchronisation pattern in every frame, as in TDM, or by filling empty cells with a synchronisation pattern. A multiplexing scheme of this nature is often referred to as 'slotted'.

Each packet includes a packet header which must contain a label that identifies the connection to which the packet belongs. In general the address space from which these labels are selected is specific to each input port of the fast packet switch. The label is selected by the control processor of the switch when the connection is established from the pool of unused addresses on the relevant input port of the switch. If the address space of the label field were not localised the problem of allocating a globally unique label in a large network would be time consuming and would limit the number of virtual circuits that could be supported within the network. Thus to support a connection across a number of fast packet switches a different label is required to traverse each link within the path. One function of the input port of a fast packet switch is therefore to replace the label field of the incoming packet with an outgoing label. It does this by means of a look-up table which is set up when the connection is established. In a very large hierarchically structured network a two level labelling technique may be required, one label for local switching and another for trunk switching. Conversely in smaller networks a simpler scheme may be adopted possibly based on a globally unique destination name or upon a unique area code with a local destination name [51].

If a high capacity fast packet switch is to be constructed, a multi-path design is required. This may be achieved either by interconnecting a number of complete fast packet switches to form a larger structure or by implementing the switch fabric as

a multi-stage interconnection network of simple switching devices. In both cases the switches that are interconnected will be referred to as switching elements. In the first case each switching element is a complete fast packet switch; complete with control processor, connection tables and label manipulation in the input ports. This allows flexibility in the choice of interconnection network but causes unnecessary replication of the control functions in a large switch. The second method, which is the more popular, does not require replication of the control processor or input port functions but implements the switch fabric as a multi-stage interconnection network of simple switching elements. Examples of multi-stage interconnection networks may be found in figs. 4.8 and 5.3.

The multi-stage interconnection networks generally selected have the property that a simple algorithm exists whereby each switching element can forward an incoming packet towards the correct output port. This algorithm usually requires that a tag specifying the required output port number be prefixed to each packet on entry to the switch. This function is performed in the input port by table look up on the label field in the packet header. One class of networks that display this property are commonly called banyan networks in the literature, although they have been more accurately defined as delta networks which refers to a specific sub-class of banyan networks. Switch fabrics are generally formed from square switching elements which have the same number of inputs as outputs and the degree of a square switching element is the number of its input (or output) ports. Most interconnection networks are constructed from identical switching elements. The degree of the switching element is important because it determines the number of stages of switching required in the interconnection network and hence the total number of interconnections required to form a given size switch. The number of interconnections required is a major factor in determining the maximum size of the switch due to implementation considerations.

2.5 Summary

Time division multiplexing (TDM) offers fixed bandwidth channels with a constant and low delay. Statistical multiplexing is much more flexible, offers variable bandwidth connections and handles bursty traffic much more efficiently but may suffer from high delay, high variance of delay and also loss of information under overload conditions. Conventional circuit switching supports the interconnection of TDM channels while conventional packet switching handles the interconnection of statistically multiplexed channels. A switching mechanism is required that combines the benefits of circuit switching: low delay, low variance of delay and high capacity switch structures; with the flexibility and efficiency for bursty traffic that is offered by statistical multiplexing. Two statistical switching mechanisms have been reviewed: fast circuit switching and fast packet switching. Both appear capable of offering a delay performance close to that of circuit switching while being much more efficient in handling bursty traffic. Fast packet switching has been selected for further study as it appears to be the more flexible switching mechanism and also for performance and implementation considerations. From a brief review of the evolution of the packet switch a

multi-path design has been suggested in order to achieve high capacity switch structures. Some of the basic concepts that underly many of the multi-path designs of fast packet switch have been introduced.