Chapter 9

Conclusion

The design of a fast packet switch has been presented which features a number of original aspects. Its performance has been investigated using a simulation model to gain an insight into the effect of the various design parameters upon switch performance. The performance of the switch for various models of multi-service traffic has also been investigated using the simulation model. Finally, the two fundamental components of the switch have been implemented in gate array technology to gain a detailed understanding of the construction of this design of fast packet switch.

A summary of the work will now be presented followed by a discussion of some of the more significant results. Some comments will be offered on possible areas of application for the switch and some comparisons drawn with other current fast packet switch designs. Finally, thinking of the future work required on this design of fast packet switch, some ideas on multicast switch operation will be considered. The dissertation draws to a close with the discussion of some of the problems remaining to be solved in the networking of fast packet switches.

9.1 Summary

Motivation

As the telecommunications industry continues to expand, two current trends are becoming apparent: the requirement for increased network capacity and the desire to support an increasing number of communications services (voice, video, image, text, etc.) In the public domain the current Integrated Services Digital Network (ISDN) offers integrated access to communications services that are to a large extent supported on separate networks. As the number of communications services on offer increases, so the requirement to support multi-service traffic over a single integrated network will grow in significance. In the private sphere, high speed local area networks are beginning to appear, capable of supporting video services in addition to high speed computer communications etc. Such networks may shortly require interconnection by means of high capacity packet switches both locally and in the wide

area. Furthermore, standards for the definition of metropolitan area networks are reaching agreement which also promise to support multiple services and will require interconnection across high capacity switches.

The precise traffic characteristics of future communications services are at present unknown and will change with time as the networks expand, thus flexibility becomes a key issue in the selection of a suitable switching mechanism. In addition, the majority of communications services exhibit a bursty behaviour, thus the efficient support of bursty traffic is also of considerable significance. The switching mechanism, however, will also have to support services that are sensitive to delay and to the variance of delay across the network. Conventional circuit switching offers an excellent delay performance and high capacity switches but is inflexible and is very inefficient for bursty traffic. Conventional packet switching handles bursty traffic well but has a very poor delay performance and switches of very high capacity are difficult to construct. A hybrid switch that offers both circuit and packet switching is certainly a solution for the near term but a single fully integrated switching mechanism will offer greater flexibility and will be able to adapt more quickly to the changing traffic requirements of new communications services. From a review of the available switching mechanisms, fast packet switching, a statistical switching mechanism, has been selected for further study on the grounds of flexibility, performance and implementation considerations.

Design

There are three basic classes of fast packet switch design: input buffered, output buffered and internally buffered. An input buffered switch is the simplest to construct but offers approximately half of the performance of an output buffered switch. An output buffered switch offers the best theoretical performance but requires at least an order of magnitude more hardware, whereas internally buffered switch designs fall somewhere between input and output buffered designs in terms of both performance and hardware complexity. The majority of existing fast packet switch designs are constructed in VLSI whereas the Cambridge Fast Packet Switch proposes a very simple design capable of implementation in gate array technology. A simple implementation offers flexibility, a wide range of potential applications and operation at both conventional speeds and also at possibly very high speeds. An input buffered design has been selected for simplicity of implementation but various techniques have been investigated that enhance the performance of the basic switch towards that of the output buffered design.

To enable the construction of a very high capacity switch, whilst retaining the simplicity of implementation of the fundamental switching element, the design has been based upon the use of a multi-stage interconnection network for the switch fabric. The delta network has been selected as a reasonable compromise between complexity and performance but to improve the performance, to increase reliability and to remove the sensitivity of the switch to the distribution of the incident traffic the Beneš network has also been investigated. Whereas the majority of previous work on the use of multi-stage interconnection networks has focussed upon the 2×2

switching element this design concentrates on the use of switching elements of up to 16×16. This improves the performance and reduces the number of interconnections required within the switch fabric which forms a major factor limiting the maximum size of the switch implementation. The use of a multi-plane switch fabric has also been proposed in order to improve reliability and performance. Switching elements of high degree and multiple switch planes both introduce multiple equivalent paths between the same input and output ports into the switch fabric. Three algorithms have been suggested in order to select a path across the switch fabric: searching, flooding and random selection. Input queue by-pass is a technique that has been investigated to improve the basic performance of the switch and for a two-plane design, output buffering across the two parallel planes also enhances the performance.

Performance

The influence of the various design parameters on the performance of the switch has been investigated by considering the throughput at saturation of each design using a simulation model. The delay performance for slotted traffic has also been investigated and the results have been compared to the performance of the crossbar network which represents the performance of the ideal input buffered switch.

For delta networks, the use of a two-plane switch fabric is recommended on the grounds of increased throughput, increased reliability and ease of maintenance but the use of more than two switch planes in parallel is not justified from the point of view of increased performance. The use of switching elements of degree 8 or 16 is preferred against those of degree 2 or 4 because of improved throughput performance and reduced interconnections within the switch fabric. For delta networks the searching algorithm offers a performance only slightly lower than that of a flooding algorithm and a hybrid algorithm which searches within each switch plane but floods across the planes is recommended for its ease of implementation. A two-plane regular delta network offers a throughput performance only slightly inferior to that of a crossbar switch fabric. The introduction of input queue by-pass together with output buffering yields a performance comparable to the two plane crossbar switch fabric with output buffering, and only slightly lower than that of the output buffered switch.

In general, a Beneš switch fabric is unlikely to be favoured above a delta network purely on the grounds of its throughput performance. It is of interest because it reduces the sensitivity of the switch to the destination distribution of the incident traffic. The performance of the Beneš switch fabric lies between that of the equivalent delta and crossbar networks. For applications that require a short packet length the random path selection algorithm is recommended whereas a flooding algorithm may yield improved performance for longer packet lengths.

An extension to the design of the switch has been proposed to support two fundamental classes of traffic. Reserved service traffic receives a higher priority in the switch fabric and handles classes of traffic that are sensitive to delay whilst the unreserved service caters for traffic that is less delay sensitive. Simulation results indicate that for a Poisson reserved service traffic loading of up to 80% of the throughput

at saturation of the switch fabric, the upper bound on delay for 99% of all incident reserved service packets is in the region of 20 packet lengths. Further, unreserved service traffic may be multiplexed with reserved service traffic, at every input port of the switch, so as to operate the switch continuously at saturation, without affecting the bounded delay performance of the reserved service. This result has been shown to hold for a wide range of switch structures and switch fabric design parameters. Also the reserved service throughput and delay performance appears insensitive to the arrival distribution and to the destination distribution of unreserved service traffic.

A closer investigation of the delay performance of the switch for voice traffic modelled as a superposition of individual packet voice sources on each switch port, both with and without silence detection, reveals no significant departure from the delay performance of the Poisson model. This traffic model was observed to give a packet arrival distribution closely approximating that of a Poisson source.

For delay sensitive, reserved service traffic performance, the packet length for both reserved and unreserved service traffic should be kept short and constant. No performance impairment is introduced by a $\pm 10\%$ variation in packet length but an exponential distribution of packet lengths causes a loss in throughput performance of the order of 12% for a 64×64 regular two-plane delta network. For a single service implementation, moderately insensitive to delay, variable length packets of any reasonable maximum length may be supported.

A cursory inspection of the buffer overflow probability suggests that an input buffer length of 20 packets is sufficient to offer a packet loss probability of less than 10^{-6} for slotted traffic at a traffic load of 80% of the throughput at saturation for all switch designs.

Implementation

The implementation of a 4×4 self-routing crossbar switching element in 3 μm HCMOS gate array technology has been investigated in detail together with an experimental input port controller in the same technology. The operation of the switching element has been measured and its throughput at saturation shown to agree with that predicted by the simulation model to within 1%. Insight gained from the construction of the 4×4 switching element has allowed estimates to be made of the complexity of switching elements of larger degree. Due to limitations in the available technology the operating speed of the 4×4 switching element was restricted to 8 MHz but improvements to the design have been suggested which should enable operation at 50 MHz in 2 μm CMOS without great difficulty. Implementation in higher speed technologies has also been discussed and various possible developments to the basic design have been considered to construct a full-scale switch for use in communications applications.

9.2 Discussion

The work has demonstrated that it is possible to construct a high capacity fast packet switch from very simple components. The design is easily partitioned both at the gate array and circuit card level and all but the largest of switches should experience few serious implementation difficulties at conventional speeds. The use of switching elements of high degree (8 or 16) has been shown to offer significant advantages in both performance and in the number of interconnections required in the switch fabric over previous designs based upon 2×2 switching elements. The use of input queue by-pass, a two-plane switch fabric, and output buffering across two switch planes, has shown that the basic performance of the switch may be enhanced to approach that of the ideal switch at a cost of increased complexity. The provision of two levels of traffic priority has demonstrated that although the switch is probabilistic in nature, certain statistical guarantees can be made on the delay of high priority traffic across the switch given that the mean load of the high priority traffic is kept below an upper limit. This guarantee has been shown to be unaffected by the load or distribution of the lower priority traffic.

For asynchronous transfer mode applications within broadband ISDN at conventional speeds, the complexity of the switching element is unlikely to be an issue. Modularity, incremental growth, reliability and ease of maintenance will be much more significant. To support a large number of communications services, many priority levels may be desirable within the switch. In addition, to reduce the variance of delay and buffer overload probability at high loads, a switch design which handles packets within each priority level in a strict first in first out manner, across all switch ports, is likely to be preferred.

The design issue of whether to use serial or parallel data paths within the switching elements is dependent upon technology and at conventional speeds both approaches appear possible. At very high speeds serial techniques are likely to be preferred due to their hardware simplicity with the ultimate possibility of implementing the data paths of the switching elements using photonic devices. If the broadband ISDN adopts asynchronous transfer mode and organises the network hierarchically it may be advantageous for the upper layers of the network, corresponding to the current trunk network, to group packets (cells) destined for the same major urban areas together. Thus large trains of packets may be routed through the upper layers of the network without reference to individual packet headers. Switches based upon photonic switching elements might find an area of application within the upper layers of such a network. If this were the case, simplicity of implementation might become a very desirable feature together with the ability to handle variable length 'packet trains' and to offer some statistical guarantee on delay for higher priority traffic.

Switches for applications within the local area, serving as multi-port bridges for LANs, interconnecting high speed LANs, or performing the function of a high speed LAN themselves, will not be required to work continuously at high loads. Also the delay requirements are likely to be much less stringent than those of the public network as the major part of the voice service will probably remain circuit switched for

some time. Thus applications involving video and image services, possibly including local or stored voice, are those most likely to introduce the requirement for high capacity fast packet switching in the local area. In this environment a switch design that does not require investment in VLSI may be attractive. To facilitate further experimental work in this area the Cambridge Fast Packet Switch has been designed to be capable of interfacing to the Cambridge Fast Ring [68, 67] to form a multi-port bridge between a large number of rings.

It has been demonstrated that for delay sensitive traffic, the delay across a fast packet switch may be comparable with that across a circuit switch. The variance of delay, however, may be much greater than that of a circuit switch. For fast packet switches operating at speeds above about 10 MHz, the variance in delay will be much less than the packetisation delay for the voice service. Thus simple techniques are available to deal with it. For the voice service, the greatest delay component due to packet mode operation is the packetisation delay which in most cases exceeds current regulations concerning delay in the local arm of the public telephony network. Thus for the immediate future the voice service is likely to remain circuit switched and fast packet switching may be introduced in the context of a hybrid switch. The packetisation delay, however, is not a fundamental problem and as work progresses on asynchronous transfer mode operation, the regulations may be relaxed to encourage a fully integrated switching mechanism.

The delay measurements presented in this dissertation refer to traffic with random arrival characteristics. It has been shown that a large enough superposition of periodic sources will approximate to a random arrival process given that no external synchronisation is applied to coordinate the individual traffic sources and their destination requests. It is possible to coordinate the arrival characteristics and destination distribution of periodic traffic such that any statistical switch will offer poor performance in terms of delay and packet loss. Such situations, however, are very unlikely to occur naturally and will persist only for a very short duration. It may be necessary to ensure that no coordination of periodic traffic arrival characteristics is present in a system design. Suggestions have been made that the coding of periodic traffic sources be modified to add a measure of variance to the packet departure times to ensure random arrival characteristics at the first switch in the path.

All of the fast packet switch designs reviewed in this dissertation are capable of supporting switch implementations of moderate size operating at conventional speeds. The input buffered Batcher-banyan design suffers from the heavy overhead imposed by the three phase contention resolution algorithm and is very inefficient for short packets. Further work is required on the contention resolution algorithm for this design to become competitive for delay sensitive traffic. The output buffered Batcher-banyan design (Starlite) requires a much larger switch fabric to handle the recirculated packets than other switches of the same size. An electronic implementation of the Knockout switch is limited in speed and requires much more hardware than other designs but offers incremental growth and the possibility of variable length packets. There is little to choose between the designs of internally buffered switch beyond cost and performance considerations. In current technology, designs using buffered

switching elements with parallel internal data paths appear attractive for switches with port bandwidths in the range 100–200 Mbits/sec while very high speed switches with port bandwidths above 500 Mbits/sec favour non-buffered designs using serial data paths in the switch fabric.

The Cambridge Fast Packet Switch may be constructed from simple low cost devices. It may support variable length packets, constant length packets or a range of discrete packet lengths. It is reasonably efficient for very short packets and can support two levels of packet priority with little difficulty. It may be possible to modify the design to support a wide range of packet priorities at the lowest level within the switch fabric. This could be achieved by modifying the arbiter in each switching element to select packets according to a priority field in the header of each packet and operating the switch fabric synchronously at the packet level. The switch, however, does not guarantee strict first in first out operation within a priority level across all switch ports at high loads. These characteristics suggest applications:

- as a high speed local area network;
- as a high capacity multi-port bridge between high speed LANs;
- for multi-service traffic within the local area;
- for the wide area interconnection of multi-service local area networks;
- for the packet switching function within a hybrid switch, e.g. within an integrated services PABX or within a metropolitan area network;
- or possibly for switches operating at very high speeds.

It is also possible to operate the switch as a full-duplex circuit switch and in such a role it may find applications within the field of parallel processing, e.g. to support processor to memory interconnection.

9.3 Multicast Operation

In considering further work on the Cambridge Fast Packet Switch, multicast operation forms perhaps the foremost requirement. A multicast connection is a one-to-many or distributive connection in which a single incoming packet is replicated and each copy routed individually over different outgoing virtual circuits. For reserved service traffic a distributive connection may be required to support audio conferencing or TV distribution, whereas for unreserved service traffic, multicast connections will be required to support the interconnection of groups of local area networks.

The Starlite switch [70] suggests a receiver initiated approach in which the receiver launches empty packets into the switch as required which receive a copy of the relevant incoming multicast packet in a copy fabric. This approach, however, assumes synchronisation between the source and destinations and is therefore not suitable as

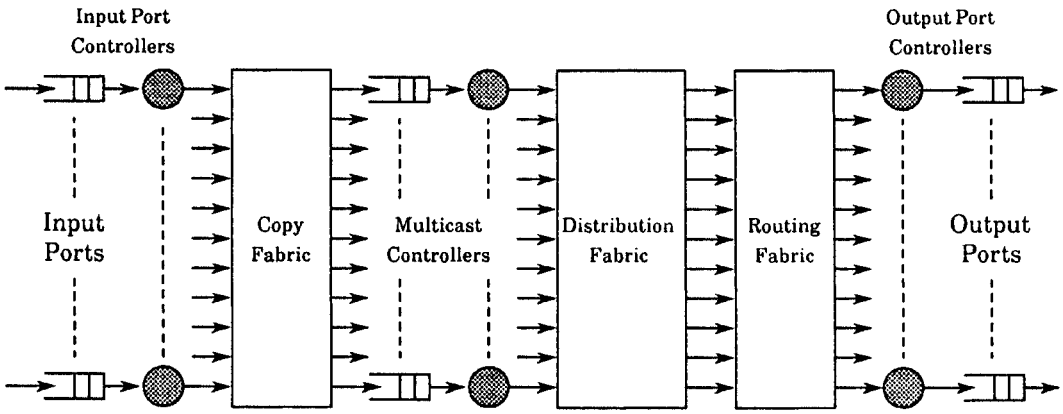

Figure 9.1: Switch structure for multicast operation.

a general method for both reserved and unreserved traffic. The switching element of the Prelude design [141, 31] is able to handle multicast packets directly but the majority of switch designs suggest a structure of the form illustrated in fig. 9.1. The structure of the fast packet switch has been augmented by the addition of a copy fabric and multicast controllers. Multicast packets are replicated in the copy fabric and routed to the outgoing virtual circuits by the multicast controllers using table look-up and manipulation of the label field in the packet header.

Two proposals have been made for the construction of the copy network: a buffered banyan [145, 148, 19] and a running adder network followed by a non-buffered banyan [89, 88]. The non-buffered copy network is also non-blocking and thus its performance is more easily predicted than the buffered banyan network but both approaches can handle multicast traffic at reasonably high loads. Neither solution, however, offers a particularly simple implementation.

For fast packet switches aimed at the interconnection of local area networks, operation under continuous high loads of multicast traffic is unlikely to be a requirement and a simple implementation may offer an advantage. In this environment a possible solution for the construction of the copy fabric is a destination release slotted ring. A tag is prefixed to a multicast packet by the input port controller which defines the number of copies required. It is then launched into the ring on the arrival of the first available empty slot. A copy of the packet is taken at every station that it passes and the tag decremented. When the tag reaches zero the slot is released.

The technique is suitable for implementation in gate array technology and should handle both reserved and unreserved traffic at moderate loads. For low loads of multicast traffic or for multicast traffic that is insensitive to delay it may be possible to implement the copy fabric and the multicast controller function within the input port controller. At higher loads it may be necessary to preface a ring based copy fabric by a single stage distribution fabric to avoid the blocking of a downstream node by a busy upstream node.

9.4 Network Aspects

The design and implementation of a fast packet switch is not an exceedingly difficult task. The Cambridge Fast Packet Switch has demonstrated a switch design capable of very simple implementation but nine other designs have also been reviewed. A simple comparison of the performance of the various switch designs has been offered based upon the throughput at saturation, the mean delay for slotted traffic and the 99^{th} percentile of delay for Poisson traffic. A reasonably self contained topic for further study would be a comparison of the packet loss probability of the various designs of switch against traffic load and buffer length. The integration of both delay sensitive and delay insensitive traffic across a single fast packet switch has been investigated using two levels of priority and the results suggest that acceptable performance characteristics may be attained for the various classes of traffic. The extension of this technique to a wide range of priority levels might be of some interest, but the major problem that remains to be solved is the support of multi-service traffic across a network of fast packet switches.

Amongst the many issues still to be addressed in the networking of fast packet switches are:

- the characterisation of source traffic profiles;
- the determination of the service performance requirements;
- effective policing mechanisms for traffic sources;
- the allocation of reserved service bandwidth;
- access control, flow control and congestion control.

A policing mechanism for traffic sources and a simple approach to the problem of determining the effective bandwidth required by a source is discussed in [5]. Flow control refers to the control of traffic flow across a connection from end to end whereas congestion control is the action taken by the switches to avoid congestion within the network. A range of congestion control algorithms are possible [55] and include:

- discarding packets when switch queues become full,
- discarding local packets in preference to packets that have already travelled some distance across the network,
- the use of choke packets originated by overloaded switches to cause traffic to be throttled back at the source,
- backpressure across individual virtual circuits,
- congestion prevention by the use of bandwidth reservation mechanisms.

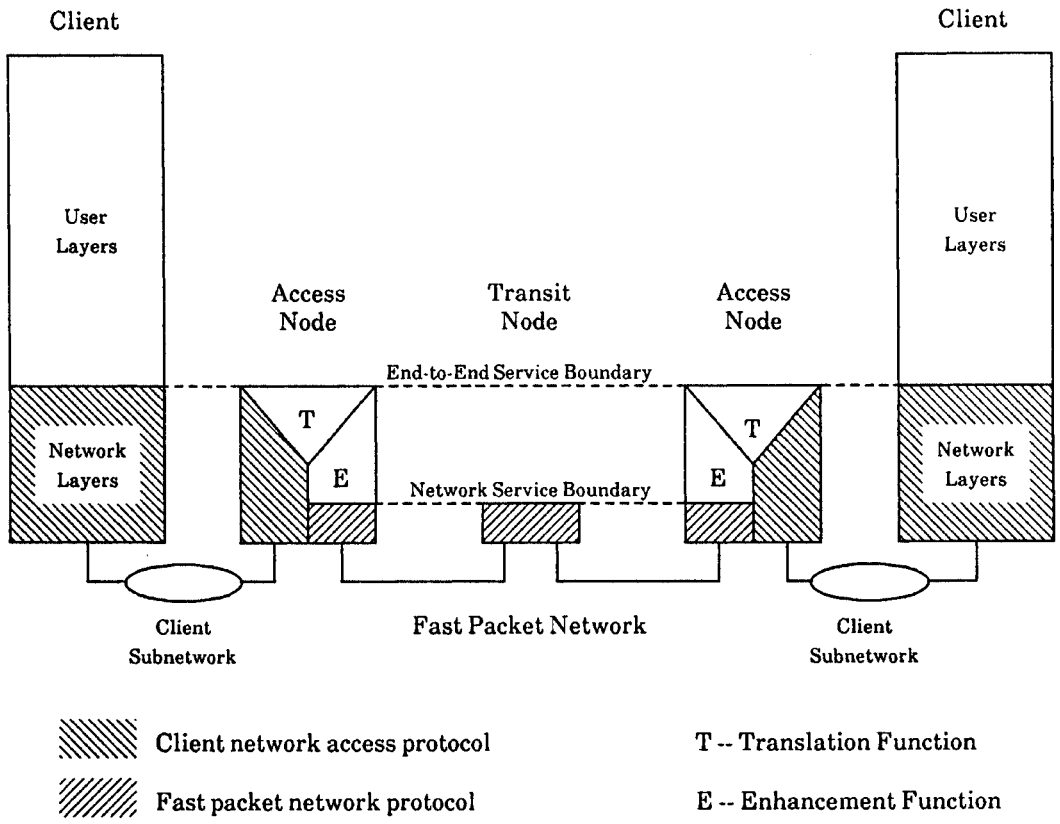

Figure 9.2: General model of protocol structure for a network of fast packet switches.

All of these mechanisms need to be evaluated in the light of the requirements of the different classes of source traffic.

The protocol functions required across a network of fast packet switches constitute another area for further study. Delay sensitive services require a lightweight protocol structure but a network of fast packet switches will be expected to offer interconnection according to the various networking access standards. A general model of the protocol structure likely to be considered for use within a network of fast packet switches is illustrated in fig. 9.2. Access to the fast packet network is controlled by the access node which connects the client across the client subnet to the network according to an established standard. This standard defines the end-to-end service required across the network. This service is supported by taking the fundamental service offered by the fast packet network and extending it by means of the enhancement and translation functions in the access node. Thus by selecting appropriate enhancement and translation functions the various classes of traffic and communications standards may be accommodated. The selection of the fundamental service offered by the network is thus of considerable significance.