Multi-Access Metropolitan Area Networks

David J. Greaves

St. John's College, Cambridge

PhD dissertation – September 1989. *2nd edition* – December 1992.

Preface to the first edition

I am indebted to my supervisor, Dr. Andy Hopper. It was he who first introduced me to the subject of computer networking and he has remained a source of constant enthusiasm and firm guidance throughout the last four years. I am indebted also to Professor David Wheeler and to John Porter. Both of whom have always, in their separate ways, been ready to comment on, criticise and discuss my work in an objective and helpful manner, often enabling me to step back and reconsider. Many thanks are also due to David Tennenhouse who again was helpful in discussions and with advice, but who in the early years, also provided generous helpings of very valuable information on the state of the three networking worlds, namely the academic world, the standards world and the real world. Thanks are also due to Bhaskar Harita, Ian Leslie, Derek (Mac) McAuley, David Milway, Roger Needham, Peter Newman, Cosmos Nicoloau, Brian Robertson and Ian Wilson.

I am grateful to the Science and Engineering Research Council for the funding that they have provided. I am also very grateful to Olivetti Research in Cambridge, who have directly funded me during the last two years and also funded the Backbone Ring project. I am particularly grateful to Dimitris Lioupis at Olivetti for his work on the prototype stations.

This dissertation is the result of my own work and is not the outcome of any work done in collaboration. This dissertation is not in any way similar to any other dissertation that I have submitted for a degree, diploma or any other qualification, at this, or any other university. No part of this dissertation has already been or is being concurrently submitted for any such degree, diploma or other qualification.

Preface to the second edition

In this second edition of my PhD dissertation I would first like to take the opportunity to thank my examiners, Professor David Wheeler and Professor Andre Danthine. I thank them for the corrections they requested of the original disseration and thank them for then accepting the amended edition. Of course, these corrections are included in the current edition.

More than two years have elapsed since the first edition of this document, so I have taken the liberty of making a few changes for this second printing. Apart from the corrections referred to above, and from the rephrasing of certain passages, the changes consist mainly of an updating to certain of the ATM terminology to reflect current usage. The other changes are the the introduction of references to several recent papers at appropriate points.

A contemporary note on applicability of this dissertation.

In this dissertation, multi-access nodes which introduce FIFO buffering of data already in transit, such as the register insertion ring system, were only considered in passing, owing to the perceived cost and complexity of FIFO technology at rates of several gigabits. Technological advances now make feasible such buffering at such rates, once the signals are in the electronic domain. However, the design approaches presented herein are valid for rates of tens of gigabits or for optronic implementation, where FIFO buffering remains in its infancy.

DJG, Cambridge, November 92.

Summary

This dissertation investigates the applicability of multi-access computer networks to the Metropolitan Area. The term Metropolitan Area refers to a geographical area comparable in size to a small town or a university campus. A multi-access network consists of a shared medium, such as an optical fibre or a radio frequency, onto which the client stations broadcast according to medium access rules. Stations wishing to receive monitor the broadcasts, looking for transmissions addressed to them.

Today, multi-access techniques are primarily used for Local Area Networks which span an office or floor, but the underlying integrated circuit and optical fibre technologies have advanced such that both the speed of operation and the distance covered can be increased nearly one hundredfold. The dissertation examines whether the new technology enables multi-access networks to be applied in the Metropolitan Area to form a backbone network, and then, whether such a backbone can support the new services expected of advanced Metropolitan and Local Area Networks, such as the capability to carry real-time voice and video.

The dissertation concerns itself only with the hardware and medium access control implementation, having summarised some of the management and routing techniques which may be applicable in an introductory chapter. Several chapters of the dissertation discuss the design and implementation of a prototype multi-access Metropolitan Area Network, designed and partially constructed by the author as part of the research.

Contents

Glossary

ADC. Analogue to digital converter.

AGC. Automatic gain control.

ATM. Asynchronous transfer mode.

Baseline Wander. Noise voltage introduced into a channel when low frequency components are removed.

BFSK. Binary frequency switched keying.

CCITT. Comiteé Consultatif International Télégraphique et Téléphonique

CDM. Code division multiplexing (spread spectrum).

CFR. Cambridge Fast Ring.

CFRV. Cambridge Fast Ring varient. See MSR.

Codebook. The table used to generate the alphabetic symbols used in block coded modulation.

CSMA. Carrier sense, multiple access.

CRC. Cyclic redundancy check.

dBc. Decibels relative to carrier level.

dBm. Decibels relative to 1 milliwatt (-30 dBW).

Disparity. Synonym for DSV.

Dither. A random or pseudo-random noise or interference effect deliberately introduced to break up cyclic behaviour.

DSR. Double-slotted ring protocol. Described in section 5.4.

DSV. Digital sum variation. Difference between total number of ones and zeros transmitted.

DQDB. Dual-bus distributed queue. A dual-bus slotted MAC formerly known as QPSX [IEEE 89].

ECL. Emitter coupled logic.

Excess Loss. Power lost through a component owing to imperfections. The loss in excess of that which results from splitting the signal over several branches.

Expedited Transfer. A priority mechanism, which operates entirely within a network station, whereby priority packets are forwarded in preference to others (page 70). Compare with MAC layer priority.

FDDI. Fibre distributed data interface, a 100 MBit/second token LAN.

FDM. Frequency division multiplexing.

Geometry. Defined in this dissertation to refer to the product of serial line-rate and total length of fibre in the logical topology.

HCR. Homogeneous closed ring – a closed ring where all stations are exactly alike.

HDLC. High-level data-link control. A bit-oriented transmission protocol.

LAN. Local area network.

HDTV. High definition television.

HSLAN. High speed LAN. Line rate above 100 MBit/second.

IP. Internet protocol.

ISDN. Integrated services network.

ISI. Intersymbol interference.

Justification Jitter. Jitter which arises when data is transferred from one clock domain to another as a result of inserting or deleting justification (pad) symbols.

Latency. The time for a single bit to lap a logical ring topology.

LC. Inductor-capacitor tuned circuit resonator.

Load Balancing. The process of providing fairness amongst contending users.

MAC. Media access control protocol.

MAC layer priority. A priority mechanism built into the media access protocol which bears on all stations connected to a multi-access network. Compare with expedited transfer.

MAN. Metropolitan area network.

MSDL. Multi-service data-link layer protocol [MAC 89].

MSR. Multiple slotted ring protocol. Also known as CFRV. A ring MAC where stations can make more than one transmission per ring revolution (unlike the CFR).

NBCR. Narrow-bandwidth closed ring. a closed ring where one station has been modified for lower bandwidth.

Perfect Load Balancing. Load balancing where fairness is amortised over as short an interval as possible.

MMIC. Monolithic microwave integrated circuit.

MSNL. Multi-Service Network Level.

Multi-Access. A network where transmitting stations take it in turns to write to a common channel.

NBCR. Narrow band closed ring.

NRZ. Non-return to zero. A binary channel consisting of a stream of zeros and ones, without any modulation: 0110111010...

NRZI. Non-return to zero, invert on ones. A one is represented by a change of the binary line state.

PAL. Programmable array logic.

PIN-FET. P-insulator-N diode connected to a field effect transistor.

PLL. Phase-locked loop.

Profile Information. MAC state information stored in a station on a slotted network concerning the relationship between the station and each slot.

PROM. Programmable, read-only memory.

QPSK. Quadrature phase shift keying.

RPC. Remote procedure call.

SCR. Simple closed ring. All stations have identical AC characteristics.

SONET. Synchronous optical network. A synchronous switching hierarchy.

Splice Time. The time between the last bit of meaningful data in one transmission to the first bit of the next transmission.

Star Coupler. A fibre optic hub which splits light received on any fibre evenly among all outgoing fibres at the hub.

TCP. Transmission control protocol.

TDM. Time Division Multiplexing.

UCOL. Ultra-wideband, coherent optical network.

UDL. Unison data-link protocol.

Varactor. A diode whos capacitance reliably depends on its bias voltage. Used for VCOs.

VCO. Voltage controlled oscillator.

VPI. Virtual Path Identifier

VLSI. Very large scale integration.

VME. A popular backplane bus. There is debate over the exact meaning of the letters.

WDM. Wavelength division multiplexing. The use of different colours of light on a single optical fibre.

Chapter 1

Network Requirement and Provision

This dissertation examines the suitability of a ring topology, multi-access network for a metropolitan area backbone network (MAN). A metropolitan area is defined as an area equivalent to a university campus or a small town. In the town example, it is possible to imagine many independent metropolitan area networks, each operated by a private company with offices distributed throughout the area. IEEE working group 802.6 has declared that for a proposal to qualify as a metropolitan area network, it must be capable of covering an area at least 50 kilometres in diameter and provide over 100 Mbit/second of bandwidth.[1]

In particular, this dissertation is about the class of MANs which attempt to extend the technology and techniques which are being developed for and used in the current generation of high-speed local-area networks (HSLANs). The important aspects of this LAN technology are broadcast access to a shared medium, very low access and transit delays and low average utilisation. Of primary interest is the degree to which these properties can be preserved as the line-rates and cable lengths are increased towards MAN dimensions.

The number of bits stored in the network transmission channels is proportional to the product of the line-rate and the length. Increasing either

[1] This was part of an early specification before they finally adopted QPSX/DQDB.

factor raises the product, and throughout this dissertation, this is referred to as increasing the network geometry.

1.1 Structure of the Dissertation

This introductory chapter discusses the traffic types that should be carried by a MAN in a multi-media environment and examines the spectrum of services that such a network might offer.

Chapter 2 reviews the optical and electronic VLSI technology that underlies MAN developments. It considers what functionality can be expected from the optical media, both now and in the future, and how the expectation of optical advances can influence the design of network topology and media access protocols. The chapter also contains a review of some current MAN projects.

Chapter 3 presents four methods for synchronising the line-rate clocks at separate network stations and considers whether the current LAN techniques are suitable when the geometry is increased. The presented material applies mainly to ring networks and the results have influenced the design of the Backbone Ring network presented in chapter 6 (see also [GREAVES 88, 90]).

Chapter 4 considers methods for providing fairness and priority services within a network, examining how the methods degrade as the network size is increased. It reviews the media access control protocols applicable to shared medium networks and investigates their behaviour when the network bandwidth is partitioned into channels at the MAC layer. Partitioning is seen as an effective method for matching the bandwidth available from optical fibres to that required by typical workstations and local-area networks. Partitioning is an important theme in this dissertation since the Backbone Ring uses a partitioned architecture.

Chapter 5 presents analytical and simulation results for a range of ring access protocols. Particular attention is paid to bursty traffic sources and their consequences for priority and fairness mechanisms. It is shown that it becomes difficult for a network to respect the priority requests of such traffic as the geometry is increased. The slotted access control methods are shown to give good performance with real-time traffic and at large geometries. A

fairly simple slotted protocol was selected for the Backbone Ring, although a wider range of protocols is discussed in chapter 5.

Chapter 6 describes the architecture of the Backbone Ring project. This formed a major part of the practical work towards this dissertation. The aim was to design and build a multi-access metropolitan area network for operation between 500 and 1000 Mbit/second. The access protocol and the physical layer frame format are presented. The performance of various station configurations is discussed in terms of throughput and station complexity.

Chapter 7 describes selected details of the prototype Backbone Ring stations. The chapter includes a section about the ECL access chip designed by the author for the Backbone Ring.

Chapter 8 concentrates on interfaces for high bandwidth networks. It describes how the application and its associated protocols can influence the design. Interface architectures of various performance and complexity are described and it is shown how the interface for the prototype Backbone Ring stations can be used within these architectures.

Chapter 9 ties together the conclusions from the previous chapters, examining the sensible maximum size of a multi-access network in terms of physical area, bandwidth, user community and reliability.

1.2 Shared Medium and Switching Networks

Local-area computer networks, probably the most common examples of multi-access broadcast networks, have typically carried bursty communication workloads between computers and have operated with low (below 50 percent) average utilisation. Under these circumstances, there is almost no performance penalty as a result of sharing the medium with other users. An access protocol which gives the lowest delay has often been adopted, typically contention with carrier sensing. However, in this dissertation we are envisaging synchronous loads and backbone networks where higher sustained utilisations are likely. Also, the performance of many of the access protocols developed for LANs deteriorates as the geometry is increased. A set of media access control rules are often termed a MAC. All aspects of MAC performance must be reviewed for MAN applications.

A limitation of all shared medium, multi-access networks is that the geographical area covered by the network cannot be extended indefinitely without performance degradation. A practical design has a specified upper limit for the total length of the shared medium.

The alternative to the shared medium network is the switching network. This consists of point-to-point links between switching nodes. A message transmitted at one point is not broadcast over the entire network, but is selectively forwarded over the appropriate links. Switching networks, particularly new designs for self-routing fast packet switches [NEWMAN 88], can potentially offer similar performance over the local area as the current multi-access architectures. Providing that the routing mechanism is suitable, switching networks have the advantage that they can be homogeneously extended without limit by adding new links and switches. The network links can form an arbitrary star topology and this gives fault resilience owing to the multiple possible routes. Their disadvantage for local-areas is the additional hardware investment required for an infrastructure of switches.

As stated, broadcast multi-access networks have been traditionally used as LANs. (The original Aloha network, which would now be classed as a multi-access MAN, may be safely overlooked at this point, since it had neither high bandwidth or high utilisation.) Large geometry multi-access networks are now being considered, but there is clearly an upper limit beyond which a shared medium network is not desirable. For instance, an international shared medium, spanning several adjacent countries, would not be advisable. Its management costs could be comparable to those of managing international radio channel frequency allocation. Evidently there is a cross-over point where switching networks are more suitable than shared medium networks. It is a purpose of this dissertation to examine a sensible maximum size for a shared medium network, in terms of the delay, throughput and number of stations. Switching networks are not further considered. In addition there are further practical issues, such as the inherent poor security of a broadcast network, but discussion of these issues is deferred to chapter 9.

1.3 Multi-media Traffic Characteristics

This section presents a partial list of the traffic classes that a multi-media backbone network must carry. This list contains selected, illustrative examples taken from CCITT I.121, which forms part of the specification for broadband ATM [CCITT 89].) The description associated with each source describes the nature of the traffic generated and the service requirements it has of the network. Given that the network has sufficient bandwidth, the main requirements are expressed in terms of end-to-end delay and inter-arrival time jitter. These traffic source models and their service requirements define the environment used for simulation in chapters 5 and 9.

Standard Voice: So called 'toll quality' voice consists of a 64 kilobit per second synchronous stream. It is desirable to packetise such a stream into samples of about 16 bytes. Each sample then represents 2 milliseconds of speech which is roughly the same amount as can be occasionally lost without 'appreciable' degradation to the channel.[2] Silence suppression can be applied to approximately halve the bandwidth requirement. This makes little difference to the transport requirements. The transport requirements are fairly low delay, say less than 100 milliseconds end-to-end for the ninety-nine percentile, implying low jitter. The low jitter requirement facilitates short reassembly buffers. If the reassembly buffer holds on average two samples, then 4 milliseconds of jitter would be the allowable maximum before packets arrive too late for proper reassembly and are thrown away.

Enhanced voice: Higher bandwidth voice streams have been proposed, for instance using twelve bit coders and higher sample rates. These require maybe four times the channel capacity, but are otherwise similar to the proceding class. The use of NICAM, ADPCM or other compression techniques does not significantly change the requirement.

Hi-Fi Streams: It has been proposed that compact disc quality sound should be carried. Such a channel has a pair of stereo channels sampled at 44.1 kilohertz to sixteen bits. Such signals might be carried

[2]This however may not be the case if the channel is actually carrying signals between data modems which were designed for operation over the analogue network. High performance modems cannot tolerate dropped voice packets, but modern devices are typically intelligent enough to retransmit and automatically negotiate or otherwise achieve a reduced effective data rate to cover this case.

using the 2.47 Mbit/second, forward error corrected encoding which is used in domestic disc players. If used for two-way communication, this stream requires similar throughput requirements to low-rate video (defined next), but the end-to-end delay must be constrained to the same degree required for the toll-quality voice.

Low-rate Video: Low-rate video is here defined to have a bandwidth of about 2.5 Mbit/second. Non-bursty, synchronous traffic is implied. This sort of stream is generated by today's videoconferencing coders and also by workstations with small video windows. For instance, 128 pixels square by 128 grey levels at 25 frames per second results in just over two megabits. Alternatively, slightly higher resolution including colour, but with very simple predictive compression, will result in a similar bandwidth. For two-way video-phone use, the delay must match that of the voice for correct synchronisation.

Variable Rate Video: Variable rate video results from coders designed specifically for packet networks. The stream characteristics depend on the picture definition and also the programme material. Typically we might expect a stream which varies from about 2 to 10 Mbit/second, punctuated with datagrams of about 4 megabits which need to be sent for each scene change. The datagram needs to be sent in about one frame time which implies a peak bandwidth of about 100 Mbit/second, but the datagram may occur, on average, once every seven seconds [CANDY 71] resulting in an average additional bandwidth of 500 Kbit/second. (Panning and zooming can sometimes be handled by the motion compensation algorithms, but we cannot go into this here.) The datagram size quoted was for current broadcast resolution; HDTV would require datagrams up to ten times larger and a corresponding increase in the intraframe stream bandwidth. Variable rate video can also be transmitted using multiple streams, where certain streams carry the more important picture components and these streams are arranged to be delivered more reliably by the network. A network already designed for heterogeneous multi-media should have no problem accommodating such multi-stream traffic. (The problem of re-synchronising the various streams at the video decoder can be left to codec designers.)

Simple Interactive Computer Traffic: Simple interactive computer communication, such as terminal sessions including remote editing, result in short messages with low bandwidth requirements. These messages

should probably be transmitted at high priority since their volume is typically low compared with the amount of stream-oriented traffic, on which they are unlikely to have noticable effect. Response time is critical, since network delays above about 15 milliseconds become significant when account is taken that each character traverses the network twice and there may be roughly the same delay in the remote host response time. The priority may be lowered if this traffic starts to cause perceivable random degredation to the stream-oriented traffic.

Interactive Graphics Traffic: Interactive graphics traffic requires variable size messages to be transmitted from the remote host in order to refresh the local graphics workstation. Low delay and low variation of delay are required owing to the well established dependency of operator efficiency on predictable response time. In future it may be possible to support such traffic over fast transatlantic networks in order to share time on supercomputers. However, for metropolitan areas, this type of traffic presents no problems additional to those of variable rate video and so it is not considered further.

File Transfer Traffic: Some file transfer can tolerate very high delays, for instance mail and automatic archives. Other file transfers result from interactive requests and must be met quickly. Although the average file size in a general purpose computer system might be 4 kilobytes, those which are transferred for interactive responses are often executable binaries which tend to be much larger. File transfer traffic therefore requires a wide spectrum of service.

Remote Procedure Call: Remote procedure call may be used to transfer files and then it offers loads like those already discussed. It can also form a basis for distributed program execution in which case the offered load tends to be inversely proportional to the round trip delay introduced by the network.

1.3.1 Arrival Models.

For simulation purposes, a concise, mathematical description of a traffic source is required. The format used throughout this dissertation consists of a pentuple which is sufficiently general to encompass the sources described above, as well as most of those in the literature. It is made up as follows:

Name: a textual name for the source,

Priority: an integer in the range 1 to P, where P is typically four, representing the highest priority and 1 is the lowest priority,

Arrival process: This describes the distribution of message interarrival epochs. It is one of:

Poisson: random arrivals with specified mean,

Synchronous: fixed, specified interarrival time,

Reciprocal: Each reciprocal source is given an identical partner on the opposite side of the network. Each message of this type is queued an exponentially distributed time after receipt of the last bit or byte of a corresponding message from the partner.

Mean interarrival period: The parameter to the arrival process, and

Bulk size: The number of bits in the datagram or the number of cells (mini-packets) queued at an arrival epoch. The bulk size is fixed for each source, a reasonable assumption given each source models a single application-layer port.

There are two further fields, the source and destination addresses for each source. These are fixed and bound at the time a simulation starts, ensuring that all traffic passes over the network of interest, and no station sends to itself over the network. Static binding of source destination pairs forms part of a valid model when the simulation time is less than the expected life of a connection. This is the case throughout this work. Static binding results in a less homogeneous loading than if every packet or message is routed to a new random destination. Static binding can bring to light certain characteristics of access protocols and control mechanisms which do not otherwise appear. It is of important consequence when a network uses any kind of switching such that different parts see uneven loading. Examples are a destination-release slotted ring or when static channel assignment is used with a partitioned network. [3]

Apart from simple, degenerate arrival models, an example model multimedia user has been created and used in simulations. He is termed the type A user and he generates a fixed mixture of traffic on the network.

[3] A particular example is the Backbone Ring without pseudo-random channel mapping.

Name	Priority	Arrival process	Inter-arrival	Bulk size	Bandwidth Kbits/sec
Tollvox	4	Synch	2 ms	1	128
Music	3	Synch	87 us	1	2900
Low video	3	Synch	87 us	1	2900
Terminal	4	Poisson	250 ms	2	0.2
File	1	Poisson	500 ms	45	23
RPC	2	Reciprocal	2 ms	3	2×750
Total					7451

Table 1.1: Definition of traffic generated by a type A user.

The mixture is defined in table 1.1.[4] The idea of the type A user is to approximately determine how large a user community a given network can support.

The type A user does not generate any bursty, high-priority traffic. Whether bursty, high-priority traffic needs to be supported is of very great interest for reasons explained in chapters 4, 5 and 9. The only source listed in section 1.3 which might generate such traffic is the variable rate video with its complete frame datagrams. However, these datagrams might not require a particularly low delay guarantee since the degradation of quality which occurs if a scene change is delayed by one or two frames may not be perceivable. The effect would be for motion within the current scene to suddenly freeze, just before the scene is replaced with its successor.[5] A delay of this nature would also occur if two video channels carried by one network should change scene within a frame time when there is insufficient free bandwidth on the network. Sending the datagrams at priority cannot improve the second case, and because the delay effect may not, in any case, be perceivable, it becomes questionable whether these datagrams require special priority. A level of priority which places them above delay insensitive traffic such as file transfer and below the strictly synchronous, delay sensitive sources would appear sensible. This discussion is resumed in section 9.2.1.

[4]Further types of multi-media users who generate a different mixture of traffic types and a greater overall quantity were to have been included in this dissertation. These were termed the type B and C users. Unfortunately there was not sufficient time to complete this work.

[5]A study of this effect is not known to the author, but it may become regarded as 'subjectively acceptable' in the same way as occasional loss of voice packets for speech.

1.4 Packet Transfer Architectures

Packet switching is the appropriate transfer mode for computer data and variable rate real-time traffic. Packet networks are also capable of carrying synchronous, fixed rate traffic. Much like the international mail service, a packet is first marked with an address and requested class of service and then 'posted' at one point in the network. In the proper circumstances, packets are delivered at their destination with a probability and after a delay which relates to the service requested. The destination may be connected to the same physical medium on which the source was situated, or the packet may have traversed several intermediate sub-networks. At the OSI network level, the same protocol should support the different distances.

The classes of service offered by a network vary according to the delay and probability of success for the packet and by the amount of bandwidth that the network offers for a particular class of service. For in-band data purposes, as opposed to routing and control purposes, the parameters at the network level need not be different from those at the media-access level. These are basically priority level and number of times a transmission should be attempted before giving up. These are discussed for the media access level in section 4.0.2.

1.4.1 Packet Addressing and Routing.

The routing information attached to a packet can either be an absolute or relative address or it can be an explicit route. Quoting an explicit route can potentially result in the greatest amount of information attached to a packet, but on the other hand, no intelligence is required of the intermediate switches. Absolute addresses identify a unique destination and can be interpreted unambiguously at all points in the network. There is a compromise between the number of bits used to represent an absolute address and the amount of associative look-up that is required in the switching nodes. Although there is the potential to cover the whole planet using a 48 bit address, routing at high-speed is not cheap.

Relative addresses give only enough information for a packet to be routed to the next switching point. Packets follow a virtual circuit through the network with a new relative address being substituted at each switching point. The process of generating the next relative address is called routing

tag translation. When the tag is short, this is done by simple table look-up. Relative addresses need to be short so that directly indexed RAM tables can be used instead of resorting to sparse array techniques. They also need to be unique within a subnet. Generating unique identifiers over a large distributed system is difficult, but one solution is for the party requesting virtual circuit setup to offer a random suggestion which can be rejected by any of the intermediate nodes if that identifier is already in local use. Sixteen bits are probably sufficient for a virtual circuit identifier, although a few more are possible without the RAM size becoming too unreasonable. A short routing tag, as is possible using relative addresses, has the advantage that a short data cell can be used without undue efficiency penalties.

Virtual circuit operation for fixed-rate traffic sources provides a simple basis for reservation and accounting entities to allocate bandwidth across network-to-network bridges and routers. Since a fixed-rate traffic source is typically sending a real-time multi-media stream, these are the sources and types of traffic which can most benefit from guaranteed bandwidth.

Another advantage of address translation is that the look-up table entry can contain additional attribute bits which can be used as a new quality of service request for the next hop. These techniques are not further discussed in this dissertation. They are mentioned only as a justification of short routing tags.

Chapter 2

Gigabit Network Technology

This Chapter contains a brief review of recent developments in VLSI and electro-optical technology and descriptions of some contemporary high-bandwidth networking projects. It starts by examining the suitable physical layer interconnection topologies for multi-access metropolitan area networks.

2.1 Shared Medium Topology

A multi-access network has a shared medium, accessible by all stations. For the networks of our current interest, the medium consists of optical or electrical cables, and the topology of the network is the path that the cables take between the stations.

A given network has both a physical and a logical topology; these may often be different. The physical topology is the shape formed by the physical cables that make up the network. The cables can have multiple conductors or fibres, each of these can carry multiple channels multiplexed by wavelength, code, frequency or time, and there can be complete redundant cables provided in case of failure. The logical topology of a multi-access network is defined by an active configuration of the physical components to form a spanning path between all stations, such that a broadcast from any station passes every other station exactly once. Typical examples are a star-wired ring and a folded-bus wired as a dual ring. The performance provided by a MAC depends only on the logical topology, but the degree of redundancy in

the physical topology required to reliably support a given logical topology is very important, since the cabling and line interface costs are often the most significant components of the overall network cost.

2.1.1 Private or Leased Media.

A network may either operate over privately owned optical fibres or it may operate over channels leased from a telephone company. In the former case, the format of the electrical signals may be freely chosen by the network designer, but in the latter, there will be restrictions on line-rate and frame format which affect the low-level design.

The format of a leased circuit conforms to a communications standard, which today is typically a member of the DS or the Bell series T hierarchies of asynchronous channels, or their European equivalent. The emerging synchronous 'transwitching formats' such as SONET [BOEHM 85] and its precursor [GRAVES 87] will be more suitable for multi-access networks. Owing to their synchronous nature, they may be easily stitched together to form the shared medium without introducing justification jitter.[1]

An example of a multi-access network designed to use static circuit-switched channels as its medium is the strangely named 'Upperbus' slotted ring, part of the BERCOM project in Berlin [GILOI 86]. This is a source-release slotted ring with responses. An unusual frame format has resulted from its being designed for switched media; each slot is constructed from 18 bit mini-frames. The switched media offers multiple redundant links, limited in scope only by the extent of the of the Berlin junction telephone network. This should result in a very reliable ring.

The Upperbus project has been carried out in close cooperation with the telephone company which provided the channel time. The lease cost of multiple 280 megabit channels at proper market prices is today very high. The use of channels switched by electronic circuit switches for multi-access packet protocols may remain commercially unattractive unless the cost of the circuits and the circuit switches becomes significantly lower.

Switching between the channels with ATM style fast packet switches would appear more attractive, and is accepted as the method for broadband

[1] Jitter introduced by the insertion and deletion of pad symbols to compensate for clock rate differences.

ISDN. Higher link utilisations are likely and customers need only pay for the packets they send, instead of for the permanently open channels. On the other hand, it is likely that there will always be some customers, such as large companies or military organisations, who will be prepared to pay not to share.

2.1.2 Topology for Private Networks.

It is clear that cost should be the major consideration when selecting a network topology. This applies to both the logical topology and the physical topology required to support it. Table 2.1 breaks down the price of a private fibre optic connection so that order-of-magnitude comparisons between the component costs can be made.

Subterranean duct digging	> £20 per metre
Subterranean fibre cable	£2 per metre
Additional fibres in cable	5 pence each per metre
1 GHz opto-electronic transducer	£500 per end
Electronic line codec (VLSI)	£50 per end

Table 2.1: Order-of-magnitude costs of an optical fibre connection.

The logical topologies which should be considered for broadcast, multi-access networks are:

Ring. The ring topology cyclicly forwards symbols from one station to the next. Any station is able to overwrite any symbol passing through (although this is regulated by the MAC).

Broadcast bus. The broadcast bus defaults to an idle state when no stations are transmitting. The idle state can be reliably overwritten by exactly one station at a time, and all stations receive such transmissions without any implied order. Often there is a collision detection mechanism within the line interface which indicates if two or more stations are transmitting simultaneously.

Folded bus. Each station on a folded bus has separate read and write access attachments to the medium. The attachments are known as taps. There is a logical ordering of the stations along the medium, which is

normally the same as the physical ordering. The bus supports unidirectional transmission. Downstream stations can overwrite data generated by upstream transmissions, but in some models this can only be done reliably if the upstream signal is an idle symbol generated by the most upstream station. The read taps of all stations are downstream of all write taps. In the folded bus, the read taps are in the reverse order to the write taps. The write tap optionally has an associated receive component, which although not capable of operating at the full network bandwidth, is able to detect the presence or absence of the idle symbol or other flags.

Looped bus. The looped bus is the same as the folded bus, except that the receive taps are in the same logical order as the transmit taps. The looped bus is typically superimposed on the dual-ring physical topology.

Dual bus. The dual bus topology consists of two unidirectional buses. This requires again a logical ordering of the stations for routing. One bus traverses the stations in order and the other traverses in the reverse order. Stations have read and write taps onto both buses. Again this topology may be mapped onto the dual ring, DQDB being an example.

2.1.3 Effect of Topology on Delay.

In order to cover a given geographical area, the chosen network topology will have a characteristic delay, which might be defined, for instance, as the average over all stations of the time for one station's broadcast to reach another. This is independent of the system bit-rate, but is influenced by the physical topology used to support a logical topology. However, the variation from one topology to another is not an important consideration in their selection, since we are attempting to design networks which can support a hundredfold variation in installed network size without significant performance penalty. For ring networks, and certain types of token buses where there is an implied ring, it is worth defining the *latency* which is the time for one bit to lap the logical ring once.

For local-area geometries, the regenerative delay introduced by a station as it forwards a message, together with the number of regenerations encountered on average for a given topology, have made a significant contribution to a network's characteristic delay. However, for metropolitan geometries,

the regenerative delay is unimportant; for instance, at 1 Gbit/second, a station with 300 bits of delay is equivalent to 60 metres of fibre, which is less than 0.2 percent of a 50 kilometre network. This gives greater flexibility in the ordering of MAC fields and relaxes station design constraints.

2.2 Providing Physical Redundancy

The physical topology must provide the redundancy to maintain the logical topology in the case of link or station failure. When multiple failures occur, it is desirable that the network degrades gracefully, possibly dividing itself into several autonomously operating subnetworks.

Automated fault detection is at least as important as automated correction. The active media show to advantage in this respect, since with appropriate design, power levels and error rates can be measured on a link-by-link basis. If the MAC frame format does not contain CRC or equivalent check digits which can offer integrity information to the maintenance system, the physical layer is obliged to introduce its own reference test pattern into the channel stream.

Link redundancy where the standby link runs in the same cable or duct as the active link is not particularly effective since both links are liable to be disturbed at the same time. However, if there is only one duct between two parts of a network (bi-connectivity), as the cost listed in table 2.1 might require, then there can be no alternative, regardless how elaborate the redundancy mechanism.

The various physical and logical topology compositions must be compared in terms of hardware cost.

2.2.1 Broadcast bus.

The broadcast bus is usually implemented with a passive medium. An example is Ethernet. Optical implementations tend to use a single passive star-coupler so that excess losses are not replicated. [2] LANs using couplers

[2]The excess loss is the additional loss due to coupler imperfections beyond the theoretical loss intrinsic to splitting a signal.

with up to about 64 ports are commercially available, optical power budgets not being able to support many more stations. Such networks can assume high reliability of the coupler and therefore need not provide redundancy based recovery systems. For a single wavelength system, the star-coupler is only suitable for the broadcast bus topology. Apart from the coupler, each station requires a single transmit and receive component. Using directional couplers, a single bidirectional fibre can be used to connect a station to the centre. Special optical components to support this topology are becoming available [HUNWICKS 89].

2.2.2 Ring topology.

The ring topology inevitably uses a fully regenerative active medium since both setting and clearing of bits is required. Redundancy in order to increase reliability is required for all active media, and for a logical ring, three physical topologies tend be used: dual ring, braiding and star-wiring.

The dual ring provides redundancy without bi-connectivity when the two rings are contra-rotating. Each station requires two sets of receive and transmit line interfaces. This approach is widely used, for example, in FDDI [BURR 88].

Star-wiring, as used on the Cambridge Fast Ring [HOPPER 88], can provide a more convenient structure than a physical ring, but for this to provide an increase in reliability, each star centre must provide an active healing function. Intrinsically, all non-leaf nodes become articulation points and their failure will partition the network. These non-leaf nodes can either be stations with multiple line-interfaces, or separate dedicated entities. In the latter case, the number of transmit-receive pairs is twice the number of stations, plus maybe a further 20 percent to provide expansion ports. The former case is slightly more economical.

A station in a braided ring receives additional input fibres from its non-immediate predecessors. In practice, a braided ring will either tend to be star-wired or wired as a real ring. In either case, it is sensitive to link failure as a result of duct intrusion. Braiding only effectively provides protection against station failure. An advantage is that only one receiver and one transmitter is required per station, provided that the transmitted light be split passively and a simple optical switching facility is provided to select

the active input fibre. Optical transmitters with multiple pig-tail outputs have been designed for this purpose, as have special $LiNiO_3$ switches for the input selection.

2.2.3 Dual bus.

The dual unidirectional bus can be wired using the dual contra-rotating ring physical topology. In the case of a link or station failure, the head-ends and tail-ends are dynamically repositioned at either side of the failed section. The dual-bus requires two sets of transmitters and receivers at each station, but has the advantage that with certain MACs, it can offer twice the throughput. Since messages must be transmitted onto the correct bus according to the direction of the receiver, some routing intelligence is required, but propagation delays are reduced. The dual bus is self-partitioning in the case of multiple failures.

2.2.4 Folded bus.

The folded bus, like the dual bus, can be wired as a contra-rotating ring. The head-end and the fold station are positioned either side of any failed section. Such a fully regenerative physical implementation is unable to make use of MAC economies, such as only writing on one bus and only reading on the other, and again requires two sets of receivers and transmitters for each station. It is self-partitioning in the case of multiple failures.

2.2.5 Looped bus.

The looped bus can be wired as a dual ring where both channels run in the same direction. In this case, it can cope with a single channel failure in one link, but not with a complete station failure and it is not self-partitioning. Two sets of receivers and transmitters are required per station. Alternatively it can operate over an intermediate logical ring which is divided into two channels, the tail of the transmit ring channel being connected to the head of receive ring channel at a unique station. This logical ring then requires a further redundancy mechanism, selected from those discussed already. This topology is possibly attractive if the logical ring is partitioned using WDM, or any other technique which does not cost electrical bandwidth.

2.3 Physical Layer

Low complexity gigabit per second networks have been made possible as a result of recent technology breakthroughs in the field of high-capacity optical fibre channels. Prior technology, such as coaxial cable or waveguide channels, even of the highest quality, cannot offer significantly less loss than 1 dB per metre at 1 GHz, whereas silica monomode optical fibres offer less than 0.5 dB per kilometre.

Optical fibre channels offer two advantages over metallic systems which render the receiver design independent of the length of installed fibre. The first is that owing to the low dispersion of optical fibres, conventional line equalisation is not required. The only equalisation needed, being to compensate for the receiver front-end roll-off and then overall bandlimiting to reject excess high frequency noise. These filters only depend on parameters internal to the receiver. The second advantage is the high dynamic range intrinsic to such optical receivers. With active biasing, these devices can usually accommodate an input power level variation of at least 25 dB. Therefore no adjustments are typically required for a cable length variation between 1 and 50 km. These advantages translate into a system which can be installed with less setting up than conventional systems, and more importantly, one which is able to cope with cable length variations when new stations are inserted, or when failed sections of the network are bridged out.

Both monomode and multimode optical transmission systems can be considered for a 1 GHz network since both systems have been demonstrated at the required speed in the laboratory. However, multimode optical fibre channels are limited by modal dispersion to a bandwidth-distance product of about 2 gigabit-kilometres and so they are only suitable for the shorter links of a backbone network, such as those within a building. Mixing multimode and monomode links within a network in order to reduce costs is unattractive since this reduces the possibilities for reconfiguration. Also, at the time of writing, multimode transmitter devices mounted in the low inductance packages required for high bit-rates are not readily available.

Monomode systems avoid modal dispersion and tend to be limited only by the available power budget. The effect of the remaining forms of dispersion can be minimised using very narrow spectral width monomode sources and dispersion shifted fibre. In this type of fibre, the glass is doped so that the chromatic material dispersion almost exactly cancels the intrinsic

waveguide dispersion at the wavelength of operation. A common monomode optical fibre specification, CCITT G.652, requires that the residual dispersion is less than ± 3 ps/nm/km over the 1285-1330 nanometre window. (It also states that the loss is less than 0.5 dB/km.) DFB (distributed feedback) lasers have spectral widths of the order of 0.1 nm and so a 50 kilometre link would only suffer $50 \times 3 \times 0.1 = 15$ps dispersion. The cheaper semiconductor junction area lasers have a typical spectral width of 4 nm, which, by the same calculation, gives a worst case dispersion of 600 ps. This is good enough for baud rates up to 1 GHz.

The monomode components, such as transmitters, receivers and connectors, require very much more stringent thermo-mechanical tolerances than equivalent multimode components, and so they are intrinsically more expensive. Current production techniques involve a great deal of manual assembly and careful manual alignment of the fibre against the active device. Testing and estimating the life of each component through its early ageing are also time consuming processes. At present, the monomode market is also very much smaller than the multimode market. This has exaggerated the price difference, but considering the special attention required in manufacture, significant cost reductions as a result of mass production cannot be envisaged in the near future.

2.4 Modulation and Line Codes for Optical Fibres

2.4.1 Modulation scheme.

High-performance optical fibre systems typically use light with wavelengths of 1300 or 1550 nanometres. These wavelengths have frequencies of roughly 10^{14} Hertz and nearly the same quantity of bandwidth is theoretically available from the fibre. This is ten thousand times greater than digital electronic components can achieve and so this leads us to consider multi-channel optical systems. Laboratory prototype, frequency division multiplexing (FDM) systems using coherent synchronous or hetrodyne detection have demonstrated the capacity for over 400 and 100 channels respectively of 565 Mbit/second on a single fibre [FIORETTI 88a]. Somewhat simpler, wavelength division (WDM) systems, using coloured transmitters and bandpass optical filters before the receivers, can currently realise about 10 channels in the 1300 nanometre fibre window. WDM technology has little influence on the de-

sign of the network interface subsystem responsible for electro-optic conversion, and the same rate per wavelength as on a single channel system can be achieved. Both types of system have been realised using direct amplitude modulation of the light (which is relatively bandwidth inefficient). FSK, QPSK and other bandwidth-efficient modulation techniques are also available.

In the future, the number of WDM channels is likely to increase as more selective filter materials are developed. However, it makes little difference to access protocol design whether one or several fibres are used in order to provide a higher number of channels. It is clear that gigabit networks will initially use single channel, direct amplitude modulation, switching to WDM and then possibly more sophisticated techniques as the technologies mature. The prospect of multi-channel media warrants the investigation of suitable access protocols, as discussed in chapter 4.

2.4.2 Baseband Line Codes.

The maximum span of a high bandwidth fibre optic channel is quantum limited by shot noise in the receiving photo-diode (or self noise of any optical preamplifier). In these circumstances, using multi-level modulation schemes is not particularly worthwhile, since there can be no gain from trading signal-to-noise ratio for bandwidth. For MAN applications, the installed fibre length may be shorter than to quantum limit the system. For instance, a higher launch power may be used or the length may be restricted to, say, less than 30 kilometres. In this case, it may be electronic bandwidth that is in short supply and the 3 dB signal-to-noise penalty of a ternary code would be acceptable. These cases have been analysed for low-rate systems by [BROOKS 83].

However, there remain some pragmatic objections to ternary and higher order modulation schemes. They require more sophisticated analogue electronics at the receiver, certainly requiring AGC and window decision circuits. Also the transmitting laser must be modulated with lower index in order to improve channel linearity. This requires more careful control circuits at the transmitter and further reduces the signal-to-noise ratio at the receiver. And unlike binary codes, higher order codes require special adaptation to the VLSI logic devices at a station, since these are inevitably binary themselves. Similar arguments apply to partial response coding schemes.

Conversely, binary codes are directly compatible with binary digital logic families. It is feasible to directly drive a laser from an ECL family output pad. Similarly, the received analogue data can be fed directly into a binary input pad.

Early fibre optic work, led by the telecom companies, invariably opted for disparity controlled (digital sum variation (DSV) limited) line codes. They were perhaps envisaging several stages of analogue regeneration, or else they were being conservative. Unlimited disparity codes then became more common, including the nB1C, scrambled and unconstrained mBnB techniques. An example is the use of 4B5B coding in FDDI and the AMD TAXI devices. The complement codes are particularly amenable to TDM channels where one of the channels carries the complement bit. Scrambling is often used for hierarchial channels, for instance in SONET [BOEHM 85]. However, the alphabetic codes are possibly the most suitable over point-to-point channels interconnecting VLSI access chips since they do not suffer from patterns which unsynchronise the scrambler.

Alphabetic mBnB codes offer good efficiency (viz m/n) with even moderate n. Computer applications generally dictate that m be a power of two, suggesting an odd value of n, but the complexity of a disparity limited encoder can be high if n is odd. Recent advances in VLSI technology have ameliorated the complexity objection to disparity constrained codes, resulting in a return to disparity constrained alphabetic block-codes, as exemplified by the ANSI Fibrechannel standard.

Possibly the simplest constrained alphabetic code for computer networks is 8B10B. An example code and coder are presented in [WIDEMER 83] where the code has a DSV of 5, although this implementation was sensitive to line polarity and offered half the transition density of the unconstrained code generated by doubling a 4B5B code.[3] The cited coder and its decoder together used 380 ECL gates in the implementation. Unconstrained, statistically balanced alphabetic coders require considerably less logic. 4B5B encoding can be performed using 4 OR gates and 4 XOR gates, although some additional gates are normally required to produce the non-data symbols. Post-encoding with NRZI, by sending a transition instead of a one, ensures that at least on average there is no DC component and renders the channel polarity insensitive. This resulting signal is spectrally not unlike straight NRZ, but the total DC offset possible is limited and complete eye

[3] As used for our Backbone Ring (chapter 6).

closure through wander cannot occur.

2.5 Functional Media

A functional medium is here defined as one where MAC logic functions are performed in the physical layer. Limitations are exmained.

The receive side of a gigabit optical fibre channel is generally AC coupled since this greatly simplifies the design of the filters and amplifiers, where up to 50 dB of wideband gain may be required. The line code therefore must be approximately DC balanced, and it also helps if the line code is polarity insensitive, since then transformers and extra stages of amplification can be incorporated easily. AC coupling at the transmitter simplifies the laser biasing circuit, but DC modulation is required to turn the light source on and off if this is required as part of the media access control.

The topologies which inherently dispose of data once it has traversed the network may be termed 'self-stripping'. These are essentially radio and the bus topologies where the data simply falls off the end. Networks are often proposed which use a combination of self-stripping topology and a MAC which is oriented for performing certain logical functions within the physical medium. In particular, the only medium level operation often required is overwriting the idle symbol, or, in some cases, overwriting a preamble with an equivalent preamble while deciding whether a collision has occurred. Overwriting, in order to achieve the logical OR function, is a potentially simple operation. It may be achieved either electrically on a copper medium or or optically on a fibre medium. In both technologies, the logic function can be achieved either actively or passively. However, there remain three main problems with a passive implementation: dynamic range, training time and synchronisation.

2.5.1 Dynamic range limitation of functional media.

A passive multi-access medium is attractive since it is inherently more reliable than an active medium. The availability of active optronic components, such as amplifiers is increasing. Although until recently, only passive optronic components have been readily available. These include electrome-

chanical and $LiNiO_3$ switches, star-couplers, diplexers, splitters and combiners. These devices unanimously suffer from excess losses of nearly 1 dB, which although an acceptable figure in their electronic counterparts, is unfortunately severely restrictive when viewed in terms of the limited power budget and dynamic range of today's optical systems. If the power budget is 30 dB, then a chain of 30 devices, even with no connecting fibre, will not leave an operating margin.

Future developments are unlikely to significantly enlarge the optical power budget, since the launched optical power cannot be increased without implications for human safety, and the lower bound results from the fundamental quantum noise of a 1 GHz bandwidth. However, the intermediate optical components themselves are continuously improving. Future passive devices may suffer less excess loss; the insertion loss of a -20 dB read tap need only be 0.05 dB. Active optical components are eventually likely to be exceedingly reliable, being fabricated using lithographic processes from a monolithic substrate. The reliability of the network will then depend on the reliability of the electrical supplies to the optic devices, which can be quite high, since current commercial, all optical amplifiers take typically only a few tens of milliwatts. [BTD 88]. Such regenerators will enable the 30 dB range to be re-used again and again to form larger networks. It is possible to envisage in the future, an all optical, active medium which is virtually as reliable as today's passive media.

2.5.2 Training time.

Apart from optical power, there are further technology limitations for networks where the medium is essentially providing a 'wired-or' switching function. Receiver clock training time, the time for a receiver to gain correct synchronisation when a new transmitter starts broadcasting, contributes to the splice interval between the end of one transmission and the start of the next. This must be minimised to maintain channel efficiency. Low training time line-codes can be utilised [OFEK 89], but being essentially asynchronous, these always fail to realise the full speed potential of the underlying logic technology.

Further complications arise since a 'wired-or' channel is necessarily polarity sensitive. This has an impact on the line code, since fields which are to be overwritten must initially modulate to a zero, and then to a one

for the write. As has been mentioned, the transmitter must also be DC coupled in order to turn if off when not transmitting. Although DC coupling only slightly increases the transmitter complexity, there will inevitably be low frequency components on such channels as a result of transmitters switching on and off and having different launch powers and being different distances away from the receiver(s), so each receiver will suffer a further training overhead as it adjusts to the DC conditions.

2.5.3 Synchronisation.

For proper alignment with the desired field in a frame, a write to the medium must be correctly timed from the frame start. This requires accurate frame synchronisation, which is hard to achieve if the fields to be overwritten is surrounded by blanks. The timing content of a mostly full frame can likewise diminish.

In summary, the problems which have been identified are difficult to overcome with current optical components. Therefore projects which originally intended to use functional media have tended to be implemented regeneratively. Examples from table 2.2 are Fasnet, Metrocore and QPSX. Some adaptive polling TDM systems, such as UCOL, are functioning in the optical domain, but only for a limited number of stations. These, like the former class, are awaiting further optical advancements.

2.6 High Bandwidth Project Review

Table 2.2 presents a list of projects known to the author for multi-access networks above 100 Mbit/second. Further examples are available in [SKOV 89]. Although several projects are aiming at the one gigahertz rate or higher, many of these have taken a very simple TDM approach. The projects can be classified as follows according to how the network bandwidth is handled by the stations:

- Use of a single, multi-access MAC for the whole bandwidth, just like an HSLAN. For example, the Hewlett Packard slotted, folded bus or the Orwell Torus.

Reference	Project name	Access and topology	Line-rate Mbit/second
BERGMAN 85	-	TDM ring	2500-5000
BURR 88	FDDI	E release token ring	125
FIORETTI 89b	UCOL	ATDM star hub	100×678
GREAVES 88	CBN	S release slotted ring	500-1000
LIMB 82	Fasnet	Dual ordered token bus	150
ROSS 89	FDDI-II	Hybrid token ring	250
NEWMAN 88	QPSX/DQDB	Slotted dual bus	125
GOTO 85	NEC-1	Major/minor ring	400
SHIMIZU 87	NEC-2	TDM ring	1200
ADAMS 87	Orwell Torus	D release slotted ring	8×140
MINAMI 85	Fujitsu loop	TDM ring	2×100
SHARP 87	LAN-DTH	Tokenlike ring	140
ALBANESE 88	Metrocore	Dual ordered token bus	150, 2400
DANTHINE 85	BWN	E release token ring	167
LUVISON 89	LION	Cycle based folded bus	280, 636
GILOI 86	Upperbus	S release slotted ring	2×280
PATIR 85	Magnet	D release slotted ring	100
OOI 88	Hangman	Slotted folded bus	1600

Table 2.2: Contemporary multi-access backbone projects

- Partitioning the bandwidth into TDM channels with circuit-switched services, such as video, running in many of the channels and separate instances of a true multi-access MAC operating in a few channels, there being no MAC layer provision for cross-channel traffic. For instance Bergman's TDM ring and the NEC-2 ring.

- Partitioning the bandwidth into TDM channels and using a lower speed VLSI multi-access media access controller operating in one channel, but transferring data over all channels once access has been gained. For instance the 2.4 Gbit/second Metrocore proposal.

- Partitioning the bandwidth into TDM channels, but enabling each station to have access to any channel. For instance the Cambridge Backbone Ring project (chapter 6).

The last category in this list describes the most flexible mechanism. This approach is developed in chapter 4.

2.7 Summary

Current optical technology, unless aided by electronic regeneration, is not sufficiently advanced to provide a switching or interconnection function between a population significantly greater than fifty stations. Therefore current implementations employ full electronic regeneration at the end of each fibre. However technology is likely to advance such that the potential elegance of 'wired-or' architectures can be realised in future.

The use of a passive optical bypass relay to remove a failed network or switched off station from the network was not seen as viable for LANs since the probability of multiple successive bridged-out stations is quite high. This would typically cause too low an optical power level at the next active station. However, for backbone networks, the expectation is that stations are continuously powered and that the bypass would only be activated in the case of faults. Faults would either be situated in a random station, or at all stations on a site owing to power failure. In either case, using appropriate star-ring wring, only a single bypass switch need be operated to isolate the failed section. The probability of insufficient optical power is then much lower and this makes passive optical bypass a worthwhile fault recovery measure. In addition, remote operation of the bypass relays over a network management channel (section 7.3) provides a useful fault finding tool.

The passive star offers the simplest and cheapest multi-access medium for a small number of users, but no more than about 30 users can be supported over a 50 kilometre area. It is likely that optical regenerators will enable a greater number of users to be connected to an active hub in the future. Multiple hubs may also be interconnected by regenerated fibre links to form a 'distributed hub', but the distribution delay must be taken into account in the MAC layer protocol.

For greater user populations, and using today's technology, the physical ring without the support of redundant components requires the fewest line components and therefore provides the lowest entry cost topology. However, once redundant links are provided for reliability, the folded bus topology has the same cost as the dual ring and may also be considered. The dual bus topology has the same line component cost as the dual ring and folded bus, but can offer twice the throughput and lower delay, although it requires greater station complexity.

Chapter 3

Ring Clock Distribution Schemes

The purpose of this chapter is to examine and compare various strategies for physical layer clocking of ring networks. The chapter applies to systems which directly recover the clock at the system line-rate and also to systems which utilise sub-harmonic recovery as suggested, for example, in [DAVIES 87]. The material developed in this chapter borrows numerical values from the Backbone Ring project presented in chapter 6. The results from this chapter are then applied to the Backbone Ring project itself in section 7.5.3.

3.1 Terminology.

This chapter defines and evaluates three basic methods for clock distribution between network stations connected in the ring configuration. Hybrids of the three basic methods are also considered. The basic methods may be termed:

The Open-ring. Each station-to station link operates at a common frequency determined by one of the stations which provides a free-running master clock.

The Closed-ring. All station-to-station links operate at a common frequency which is determined by the ring geometry so that an integral

	Open Ring		Closed Ring		Asynchronous
	Baseband	Two-stage	H-CR	NB-CR	Ring
All stations identical	no	no	yes	no	yes
Examples	IBM token ring. CFR.	Backbone Ring.	10 Mbit Cambridge ring.	-/-	FDDI.
Recovered clock used directly as transmit clock ?	yes	no	yes	yes	no
Transmit clock same frequency as received ?	yes	yes	yes	yes	no

Table 3.1: Catagories of ring clocking methods.

number of bits are stored in the medium.

The Asynchronous or pleisochronous ring. Each station-to-station link operates at a frequency chosen by the station which is at the transmitting end of that link. These frequencies are nominally identical, but in practice, offset from the nominal value by a small, random error, which may also change slowly with time.

Table 3.1 features the three basic methods along with example systems that have used them.

The open-ring column has been subdivided into two classes: 'baseband' and 'two-stage'. The two-stage method is only different from the baseband method in implementation, not in the theory of its behaviour. In the table, it is only listed separately because it is the method which has been successfully used on the Backbone Ring. It is decribed in section 3.2 and chapter 6.

The closed-ring column has also been subdivided into two classes. One class is for rings where the stations are all alike, giving a homogeneous closed-ring (H-CR), and one for rings where one station is modified (or works in a different mode) to perform a ring closing function. One suggested modification [CASH 84] is to reduce the bandwidth of the clock recovery circuit in an attempt to increase the stability of the feedback loop which is implicit in the closed-ring technique. This give the 'narrow-bandwidth closed ring' (NB-CR), whose dubious merits are discussed in section 3.4.2.

3.1.1 General points about ring clocking.

Regardless of the clock distribution method, every network station is required to recover a clock equal in frequency to the transmit clock at the upstream station. This 'receive clock' is derived from the incoming serial bit stream using a clock-recovery method selected from one of those to be described in section 3.2. The receive clock is required to clock the decision flip-flop[1] which interprets the received data signal. The open and closed-ring techniques also use the receive clock to set the frequency of the transmit clock. In the simplest case, the same clock signal is used both as receive clock and transmit clock. *However, the essential property of the transmit clock is low jitter while the essential property of the receive clock is correct phase alignment.* The circuit used for clock recovery must therefore be simultaneously optimised for both of these requirements.

In the asynchronous ring technique, this is not the case. A separate, local clock is used to generate the output stream . This means that the input-side clock recovery circuit does not have to be optimised for minimum jitter accumulation. A comparatively simple recovery circuit is therefore sufficient. This is the main benefit of the asynchronous ring technique. Since the receive and transmit clocks are not synchronised, stations which form part of an asynchronous ring will contain asynchronous logic. This again requires careful design. In effect, part of the clock recovery functionality has been removed from the analogue to the digital domain. The asynchronous clock method is discussed in detail in section 3.5.

3.2 Clock Recovery Techniques

For high line-rate systems four techniques for receiver clock recovery are worthy of consideration [COCHRANE 83]:

- Low to medium Q resonator (Q 50 to 250)
- High Q resonator (Q 1000 to 10000)
- Phase-locked loop

[1]There are multiple decision flip-flops for ternary and higher order modulation techniques, but this does not affect the discussion.

- Two-stage process with separate receive and transmit filters.

3.2.1 Low to Medium Q Resonator.

A low to medium Q resonator has the benefit of simplicity. The LC tank resonator, as widely used in PCM telephone trunks, is usable for clock rates up to about 100 Mbit/s. At higher frequencies a cavity resonator is suitable. The phase response is predictable and the centre frequency is readily adjusted. With careful design, such devices have been used in chains of up to 100 regenerative repeaters.

3.2.2 Phase-Locked Loop (PLL) Recovery.

A phase-locked loop can provide narrow-bandwidth with low frequency to phase conversion provided it incorporates a sufficiently high loop gain. However, the loop gain cannot be raised indefinitely since the loop will become unstable and may slip cycles as a result of transient jitter peaks in the incoming signal. Also, such PLLs tend to have a peak in their jitter transfer function. The size of this peak must be carefully constrained to a small fraction of a dB in order to stop exponential jitter gain in a long regenerative chain. Unfortunately, the size of the peak can be critically dependent on component tolerances.

Suitable, narrow-bandwidth phase-locked loop circuits for rates of up to about 50 Mbit/s are well understood and not too hard to fabricate. An example of their use is the IBM token ring. This uses the open-ring clock distribution technique at 8 MHz. The designers calculate that their PLL design can be used in rings with at least 200 stations [KELLER 83].

For higher bit-rate systems, the complexity of a PLL operating directly at the line-rate begins to present problems and the performance of a high gain, narrow-bandwidth PLL fabricated from discrete components would probably be dominated by strays. However, Gigabit Logic [GIGABIT 88] has developed a single chip GaAs PLL suitable for NRZ data. This includes the VCO and a phase comparator suitable for baseband NRZ data streams in excess of 1 Gbit/s. Such components may eventually extend the applicability of PLL techniques to regenerative chains, although at present this technology is in its infancy.

Injection-locked oscillators can readily be fabricated up to frequencies of many GHz, and as far as is evident from the literature, these have tended to form the basis for several gigabit ring projects. Mathematically they can be approximated to a first-order PLL. The expected behaviour of a gigabit per second closed-ring of injection-locked oscillators is described in the conclusion to this chapter.

3.2.3 High Q Clock Recovery.

The advantage of high Q systems is that jitter build up is very low. Multipole filters fabricated from discrete components have not been used in clock extraction systems because their phase delay is liable to drift. This upsets the timing into the decision flip-flop. However modern passive devices such as the dielectric resonator and the SAW (surface acoustic wave) filter have enabled very reliable high Q regenerative repeater chains to be fabricated.

SAW devices are suitable for frequencies above 100 MHz and are being integrated into new regenerative designs at, for example, 320 Mbaud [COCHRANE 83] and 1.6 GBit/s [NISHIMOTO 86]. Although these devices have a high group delay (a SAW device may have as many as 500 cycles stored inside) they are highly stable and repeatable so this presents little problem. One disadvantage is the relatively high insertion loss of typically 10 dB. A disadvantage for research work is that their one-off cost can be £500, which combined with their implicit fixed-frequency operation, makes them somewhat inflexible for experimentation. However, in a production environment the new high Q devices are near the ideal.

3.2.4 Two-Stage Clock Recovery.

A design which utilises separate transmit and receive clock signals is free to independently optimise the phase alignment for the receive clock and the jitter attenuation for the transmit clock. This results in a system which is a hybrid of the open and asynchronous ring techniques. In a typical scheme, the receive clock is generated using a simple circuit chosen from those described above and the transmit clock is generated by a PLL which is locked to the receive clock. This PLL can be very simple since it is not required to cope with an NRZ data stream, and since there is no requirement on the phase of its output, the PLL phase comparator can operate at a low

frequency on divided-down versions of the clocks. Inexpensive, commercially available components are suitable, such as those developed for the domestic satellite TV receiver market [PLESSEY 85].

The digital logic required is similar to that used under the asynchronous ring scheme since there are two clock inputs which are unconstrained in phase. However, because the two clocks are now of exactly the same average frequency, there is no requirement to compensate for their difference. The number of bits which enters each station per second is the same as the number which leaves. This removes constraints on the data formats which the physical medium can carry and removes constraints on the total number of stations that can be connected. These constraints are discussed in sections 3.5.1 and 3.5.3.

The two-stage technique can result in a more complex circuit since it requires a receiver clock-recovery circuit and an elastic buffer and a transmit clock phase-locked loop. However, all three subcircuits are required to deliver lower performance than when used in isolation. The two-stage circuit is therefore easy to implement and may, in consequence, result in a design that is no more expensive.

3.3 The Open-Ring Technique

We will first examine the open-ring technique, since it is the simplest.

The definition of the open-ring clock distribution method is that one of the stations is elected to provide a master clock and all downstream stations synchronise a local transmit clock to the same frequency as their incoming clock. The master station accepts the signal from the furthest downstream station and retimes it to the master clock. This results in a synchronous system which can be analysed using standard methods for an open chain of regenerative data repeaters [BENNETT 58]. Each station added to the ring is an additional source of clock jitter and the maximum number of stations that a given design can accommodate is limited by the way jitter accumulates along the chain. There are two possible modes of failure:

- The regenerative stations are only able to tolerate a fixed amount of jitter before they introduce bit errors, slip cycles or fail to synchronise. Failure by this mode will be worst at the end of the chain where the

jitter is greatest.

- The master station has a fixed retiming capacity. Failure occurs if the peak phase deviation in the signal from the tail-end station exceeds the upper limit.

The author's measurements of the Cambridge Fast Ring stations have shown that as the jitter level is increased, cycle slipping occurs before the onset of jitter induced bit errors. Similar behaviour has been reported for other designs. Therefore, within a regenerative station, failure is typically through cycle slipping, rather than eye misalignment. If a cycle slip can cause network failure, which is especially true for slotted rings, the number of failures per second for first-order stations may be approximated [VITERBI 66]:

$$\lambda_S = \frac{8B_N}{\pi} \exp(-2/\phi^2) \tag{3.1}$$

where ϕ is the RMS alignment jitter at the station and B_N is the effective bandwidth of the applied noise spectrum in Hz. Synchronisation failure with second-order phase-locked loop recovery was studied in detail by [MEYR 82], but it approximates to twice this rate.

The rate of elastic buffer overflow, the second mode of failure, depends on the alignment jitter between the start and end of the regenerative chain. However, in practice it is normal to use a highly stable source for the master clock and this allows the alignment jitter to be approximated by the total jitter accumulated at the end of the chain. The stability provided by a quartz crystal is sufficient for this approximation.

A stable master source is also desirable in order to keep the electrical length of the ring roughly constant. Low frequency jitter or drift in the master clock causes the number of bits stored in the fibre to vary. At large geometries, there are many bits stored in the fibre and a small percentage clock frequency variation has a large effect. For example, at 1 GHz with 200 kilometres of fibre, there is 1 Mbit stored in the fibre and a 0.1 percent frequency deviation results in 1 kilobit being inserted or deleted. This is of the same order as the slot size on a slotted ring.

If the jitter waveform at the end of the chain approximates to a white low-pass function, then from Nyquist's theorem, it will have independent values when sampled at $2/B_N$ intervals. The rate of overflow can be estimated using the derfc() function (double-tailed complementary cumulative

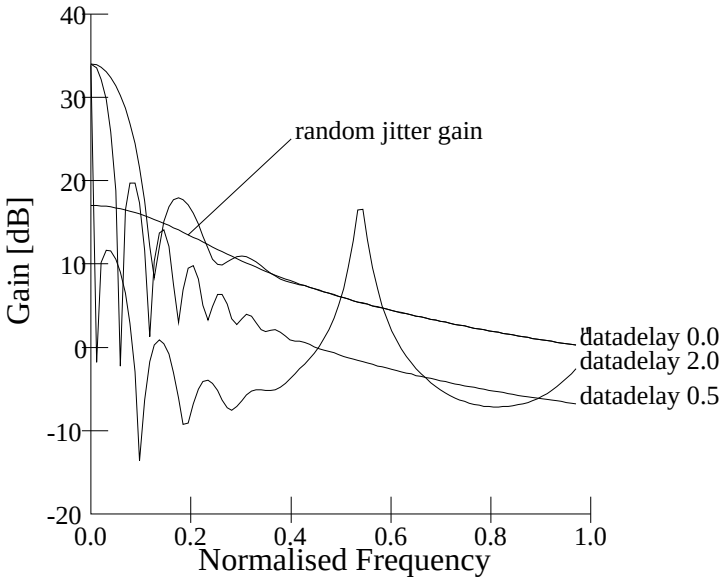

Figure 3.1: Frequency spectrum of systematic jitter after 50 stations with varying datadelay.

Gaussian function) to give the fraction of time that the jitter noise waveform will exceed the elastic buffer limits and multiplying by the Nyquist frequency:

$$\lambda_E = \frac{B_N}{2} \times \text{derfc}\left(\frac{2\pi B_{cap}}{\sigma}\right) \tag{3.2}$$

$$\lambda_E \approx \frac{B_N \sigma}{4\pi B_{cap}} \times \sqrt{\frac{2}{\pi}} \times \exp\left(\frac{-2\pi^2 B_{cap}^2}{\sigma^2}\right) \tag{3.3}$$

where σ is the RMS jitter in radians and B_{cap} is the the number of bits the buffer can gain or lose from its equilibrium position before overflow or underflow.

As described in the 'classic' paper [BYRNE 63], jitter accumulates along a chain of regenerative stations. Independent random sources, such as thermal noise in the stations and pick-up in the cables accumulate on a power basis within the passband of $H(s)$, the regenerator response. Sources of jitter which result from components of the data pattern being coupled into the recovered clock (through inter-symbol interference (ISI) and clock recovery circuit offsets), are termed systematic sources since the same data pattern occurs at each station. Using Byrne's linear shift-invarient method of systematic jitter analysis, these sources are added on a vector basis and hence systematic jitter has the potential to grow more quickly than jitter from the uncorrelated random sources.

The primary assumptions of the linear shift-invarient method are that

the pattern dependent jitter sources at each station are identical, and that the jitter experiences the same delay as the data at each station. For a typical telephone trunk, the repeater spacing is constant and the data path delay within each repeater is just one or two bits. This means that the ISI sources will be well matched, and since the jitter frequencies of interest are at least two orders of magnitude below the baud rate, the discrepancy of one or two bit intervals has not caused values predicted to be too far removed from measured values [TRISCHITTA 88].

For LAN/MAN applications, the cable lengths will not tend to be matched, but on the other hand, these networks tend not to be dispersion limited. For instance, the measurements of the Cambridge Fast Ring and the IBM token ring [KELLER 83] have shown that jitter generated by static offsets in the clock recovery dominates over that generated by cable ISI. This means that the assumption of matched cable section lengths can be dispatched.

The second assumption, equal clock and data delays within the station, becomes questionable for LAN/MAN applications. The Cambridge Fast Ring chip-set data delay is 40 bits and for the Cambridge Backbone Ring the figure is nearly 300 bits. Taking the Fast Ring example, the first implementation used PLL clock regeneration with a bandwidth of over 1 MHz. This presents about 1 μs group delay. However, since the station delay at 50 Mbit/s is nearly 1 μs, jitter at the highest frequency of interest (1 MHz) will skip one complete cycle with respect to the data at each station. Clearly such jitter accumulation cannot be described by Byrne's model in its original form.

The effect of this assumption can be modelled as follows. If the data delay at each station of T_{dd} seconds is included, then the systematic model for N stations becomes

$$\Theta_{orN}(s) = \sum_{k=1}^{N} \Phi(s) H^{N-k+1}(s) \exp(-T_{dd}(k-1)s) \tag{3.4}$$

which is simplified with the geometric progression formula to

$$\Theta_{orN}(s) = \Phi(s) H(s) \exp(-T_{dd}(N-1)s) \frac{1 - H^N(s) \exp(NT_{dd}s)}{1 - H(s) \exp(T_{dd}s)} \tag{3.5}$$

where $\Phi(s)$ is the noise source function.

As an example, figure 3.1 is a plot of equation 3.5, assuming $\Phi(s)$ to be flat and equal to unity, with values of the data delay T_{dd} equal 0, 0.5 and 2

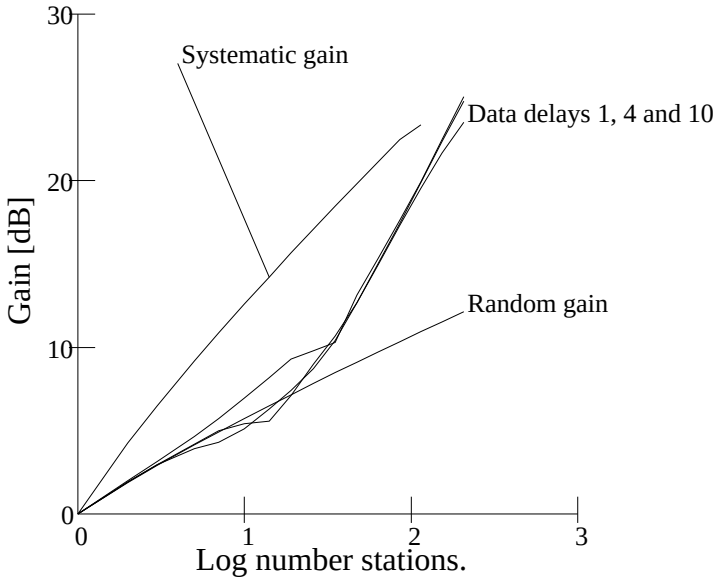

Figure 3.2: RMS jitter accumulation with increasing number of stations with the datadelay as a parameter.

times the station pole frequency and a chain of 50 first-order stations. For reference, also included on the graph is the equivalent function for random jitter gain.

The graph shows an effect, which has been observed more fully in further investigations. This is that the data path delay reduces the gain for systematic jitter to a value below the gain for random jitter over a large part of the frequency band. In particular, it is clear that the data path delay radically changes the low frequency jitter spectrum.

Despite appearances in figure 3.1, the overall systematic jitter gain remains greater than the random jitter gain. This is demonstrated in figure 3.2, which shows the noise bandwidth, B_N, as a function of the number of stations. This was obtained with a Simpson's rule integration (of the square) of equation 3.5. According to convention, the y axis has been normalised to σ^2_{or1}, the mean-square jitter at the output of the first station in an open ring

$$\sigma^2_{or1} = \frac{1}{2\pi}\int_0^\infty \Theta^2_{or1}(j\omega)d\omega \tag{3.6}$$

The graph shows that below about 30 stations, the systematic jitter gain with a normalised data delay of unity[2] is just greater than the random jitter gain, but as the number of stations is increased, the gain approximates to that for conventional systematic jitter gain, which has data delay zero. This

[2]Unity means that the data path delay in seconds is equal to the reciprocal of the station pole frequency in hertz.

implies that for a low number of stations, the data path delay prevents jitter accumulating systematicly and hence produces a smaller $\Theta_N(s)$.

When slightly peaking second-order PLLs are used for regeneration, there is a well known effect whereby the jitter gain is similar to that for first-order stations for numbers of stations below a number determined by the damping factor, and grows exponentially as stations are added beyond this number. This effect gives exactly the same shape response as that for the data path delay just discussed. Hence, we can conclude that one effect of the data path delay is to make the dependency of jitter accumulation on the number of stations even more critical.

Having established this, from now on T_{dd} is set to zero in order to give worst case results.

3.3.1 Open-Ring Regeneration At 1 GHz.

The RMS pattern-dependent jitter measured after the first repeater, σ^2_{or1}, for the Fast Ring was -28 dB-rad. Of 40 fibre optic repeaters measured by [TRISCHITTA 88], the mean pattern-dependent jitter was 23.3 $(\text{deg})^2$/MHz with a mean bandwidth of 190 kHz. This gives σ^2_{or1} equal to -31 dB-rad. For the IBM token ring [KELLER 83] the reported figure was -26 dB-rad. These figures were taken from systems whose bit-rates span two orders of magnitude, yet they are very similar. This is presumably because higher speed regenerators are made from higher speed components. Although a higher speed station may have a greater noise bandwidth, the density of the pattern-dependent sources $\Phi(s)$ will be correspondingly less. The same argument applies to random noise sources such as the receiving amplifiers, where the noise density must in any case be reduced in order to maintain channel signal to noise ratio. We conclude that there is no fundamental reason why the levels of jitter on a gigabit regenerative network should be any higher than on existing regenerative networks.

A rough guide to the jitter level acceptable on a gigabit ring can be obtained by evaluating equations 3.1 and 3.2 in reverse. Suppose that 100 stations are supported, each with 3 dB bandwidth 1 MHz, then the noise bandwidth of the cascaded stations, given by multiplying the station bandwidth by $(\sqrt[N]{2}-1)^{1/2}$, is 84 kHz. For λ_S and λ_E each to better a failure rate of once per day, the alignment jitter ϕ must be better than -11 dB-rad and

with B_{cap} equal to one bit, the RMS jitter at the end of the chain σ must be better than 0 dB-rad. The average of the three figures quoted for σ^2_{or1} is -28 dB-rad. Taking this as a typical value, then with ideal, first-order jitter accumulation, the RMS systematic jitter at the 100th station will be raised by 10 dB to -18 dB-rad (7 degrees RMS) and the systematic alignment jitter at the 100th station will by approximately the same as the RMS jitter after the first station. These figures are summarised in table 3.2. The table shows

	Acceptable value	Expected Value	Margin
Alignment jitter	-11 dB-rad	-28 dB-rad	17 dB
RMS jitter	0 dB-rad	-18 dB-rad	18 dB

Table 3.2: Typical and tolerable amounts of jitter at the 100th station of a 1 GHz chain.

that the open-ring technique will work, leaving a margin of some 17 or 18 dB, although in practice, the margins will be smaller owing to some peaking in the response of non-ideal clock recovery circuits.

The closed-ring technique generates the same amount of RMS jitter, but this is amplified by positive feedback. The next section attempts to determine how much of this margin would remain under the closed-ring technique, and therefore, whether the ring would work.

3.4 The Closed-Ring

This section examines the benefits of the closed-ring technique. The primary advantage of the closed-ring being that it does not require elastic buffer circuits, it is just a ring of PLLs. This section is in three parts: the first part discusses techniques for controlling the closed-ring operating frequency, the second part examines the behaviour of jitter when a narrow-bandwidth station is inserted into the ring and the third part compares the performance of the closed-ring with the open-ring.

3.4.1 Operating Frequency Control.

The frequency of operation of a closed-ring is

$$\omega_c = \frac{2n\pi + Nk\omega_0}{T_c + Nk} \tag{3.7}$$

where n is the integer number of cycles stored in the cables and k and ω_0 are parameters of the station that relate the phase lag to operating frequency

$$\text{Lag in radians} = L(\omega) = k(\omega - \omega_0) \tag{3.8}$$

$L(\omega)$ has this basic form for all of the clock recovery techniques mentioned in section 3.2 although there may tend to be further higher-order terms. These higher terms are undesirable, but as long as $L(\omega)$ remains monotonic, the existence of the closed-ring operating points is guaranteed. (The multiple points result from different values of n.)

Any particular station design is optimised for a nominal operating frequency ω_0 and is able to operate reliably over a small range of frequencies centred about this point. However, unless unusual precautions are taken, a typical station design will tend to partially operate over a larger range of frequencies which encompasses the designed range. This yields a region of unreliable operation where the station may introduce bit errors or slip cycles. With a given combination of ring geometry and station design, there therefore exists a set of possible values for n, the number of cycles stored in the ring, each of which gives a slightly different value for ω_c within the range which can be accommodated by all stations. As the ring geometry is increased, the number of possible values for n increases and the difference between adjacent values of ω_c decreases. The edges of the set are not well defined, depending as they do on component tolerances, although, with a sufficiently large geometry, there will be at least one value of n which gives a value of ω_c acceptably close to ω_0.

A simple closed-ring, with no master station to regulate ω_c, results in a system which can potentially operate at a frequency which is outside the design range of the stations, but inside the range where they operate unreliably. Such a closed-ring is evidently not viable and a specialised station is required to control operating frequency. Various possible specialisations are now considered.

Inserting a station which simply has especially low jitter bandwidth, but is otherwise similar to the other stations, does not affect the issue of

operating frequency. This is because the variables ω_0 and k of equation 3.8 are DC parameters.

A possible solution is to include a station which always operates within the frequency range specified for reliable operation of the other stations, and simply refuses to operate at a borderline frequency. With PLL clock regeneration, it is relatively simple to adjust resistor ratios in order to statically reduce the value of k in equation 3.8. This has been tried in certain CFR implementations, and although it prevents the general stations from operating out of range, it still does not prevent the specialised station from operating close to its, now contracted, borders[3]. As a subsidiary point, k is inescapably the DC loop gain of the PLL. It cannot be modified at one station, in order to modify the DC operating point of the ring ω_c, while maintaining the same AC characteristics towards jitter that are presented by other stations. Therefore, modifications where k is reduced destroy the symmetry of the closed-ring and result in a non-homogeneous network. As is shortly shown, reducing k at one station, and therefore that station's bandwidth, can increase the alignment jitter at that station, even though it may have reduced the overall jitter.

Returning to controlling the operating point, given a specialised station where k has or has not been reduced in order to ensure that the remaining stations are within range, one may consider including a dynamic offset to ω_0, generated by control circuitry. The effect of the offset is to slightly shift the ring operating frequency such that the correct phase exists between the received data and the local oscillator at the modified station. Such an offset voltage can be generated from a low-frequency op-amp circuit. However this only serves to shift the problem, since the op-amp circuit is inevitably limited in the range of values it can produce and it will eventually find itself operating close to its limit and therefore unreliably.

At this point, it is worth reviewing the technique used on the ten-megabit Cambridge Ring [WILKES 79]. The specialised station was the ring monitor, which established the ring clock frequency at start-up time. It broke the ring into an open regenerative chain, driving the head-end with a source at the nominal operating frequency. After chain synchronisation had been established at this rate, the monitor switched to homogeneous closed-ring mode. Under this scheme, provided that the switching is performed slowly

[3]Although reducing k reduces the capture range of the PLL, this does not matter since if the regenerative chain is in-lock, the specialist station PLL cannot be out-of-lock.

enough, the operating frequency is guaranteed to settle to one of the two possible points on either side of the nominal rate. The shortcoming of this method is that any perturbation, such as a PLL cycle slip, is able to knock the operating point to one of the unreliable solutions. Monitoring hardware is required to detect and correct this.

The Cambridge Ring design therefore used a mechanism which initially centred the ring on the desired frequency, monitored, and when the frequency was too far off, abruptly corrected it. This is in fact the only sort of mechanism which is applicable to the closed-ring. It is impossible to design a specialised station which continuously tracks and controls the operating frequency of a closed-ring because all such systems will eventually find themselves operating against one of their end stops. All viable schemes must incorporate a reset sequence which has the purpose of bringing operation to the nominal frequency.

3.4.2 A Narrow-Bandwidth Station.

Given that the ring operating frequency has been established, the accumulation of closed-ring jitter with and without a narrow-bandwidth station is now discussed. The term NBCR is used for a closed-ring with a narrow-bandwidth station and the term SCR is used for the standard closed-ring where there is a specialised control station to start the ring, but where when running, the stations are identical as far as their jitter responses are concerned.

The SCR has the intrinsic disadvantage that jitter is able to recirculate and cause peaks in the jitter gain spectrum, whereas the NBCR has the disadvantage that the narrow-bandwidth station does not have much capability to track input jitter and so sees the bulk of the jitter generated by the rest of the ring as alignment jitter.

For SCR operation, the station design is required to be free from gain in its jitter response at all frequencies, otherwise there will be a set of ranges of cable length for which the ring will turn into an oscillator. First-order PLLs, injection-locked oscillators and LC type tanks are inherently non-peaking. Second-order PLLs in non-peaking configurations cannot achieve noise-bandwidths much below one quarter of their loop gain, therefore they offer little benefit to the SCR.

For NBCR operation, peaks of a fraction of dB are permissible provided that the narrow-bandwidth station has sufficient attenuation at the peak frequency to maintain stability. Relaxing the non-peaking requirement enables worthwhile second-order PLLs to be designed. Hence the benefits of the NBCR over the SCR are potentially twofold: narrow-bandwidth high-gain PLLs can be used for general stations and the presence of the narrow bandwidth station reduces the overall jitter gain of the closed-ring configuration. The following sections present a mathematical model for the closed ring which enables more rigorous comparison.

3.4.3 Positive Feedback in the Closed-ring.

Let the input to output jitter transfer function of a regenerative station, when viewed as a low-pass function, be $H(s)$. Then the open-loop gain of N cascaded stations including the total cable delay T_c is given by

$$G(s) = H(s)^N \exp(-sT_c) \tag{3.9}$$

This does not depend on the distribution of the cable lengths. When the chain is closed into a ring, we have feedback

$$Y(s) = G(s)Y(s) + X(s) \tag{3.10}$$

The gain $R(s)$ seen by the random noise signal $X(s)$ injected at one point is

$$R(s) = \frac{Y(s)}{X(s)} = \frac{1}{1 - G(s)} \tag{3.11}$$

Since for all clock recovery systems, $G(0)$ is unity, $R(s)$ has a pole at the origin and therefore infinite power density at zero frequency. This is not an uncommon feature with noise spectrums, although it does mean that low frequency phase noise is able to propagate around the ring for many revolutions with little attenuation. The 'DC' or absolute component of the phase is not controlled, although the actual ring operating frequency is, as defined by equation 3.7.

An NBCR requires a station with narrower bandwidth than usual. An implementation of the NBCR station for the CFR used a similar PLL circuit to that at other stations, but the normal lead-lag loop filter was replaced with a single pole CR filter. Resistors were included to constrain the track

range in an attempt to define a nominal value for the ring operating frequency. The resulting PLL was heavily over-damped, such that pole/zero cancellation occurred and the loop became essentially first order.

In this study, the NBCR station is modelled using the SCR transfer function H(s) evaluated at H(s / L_F) where L_F is the fractional bandwidth of the narrow-band PLL. Although this is different in detail from real circuits, it exhibits the same behaviour and has the advantage that L_F can be continuously varied over the range 0 to 1. L_F=1 gives an homogeneous closed-ring, L_F=0 gives an open-ring and the intermediate values span the NBCR space.

PLLs are able to reject low frequency jitter by tracking the input frequency. The jitter which is not tracked is the alignment jitter and this typically appears as a displacement between the clock and data inputs of a station's decision flip-flop. At a station with transfer function $H(s)$, the input to alignment jitter transfer function is given by

$$1.0 - H(s) \tag{3.12}$$

which has a zero at zero frequency. The alignment jitter is greatest at the narrow-bandwidth station of an NBCR closed-ring. Its spectral density will be proportional to

$$R_{cr}(s) \stackrel{\text{def}}{=} R_{LF}(s)(1 - H(s/L_F)) = \frac{1 - H(s/L_F)}{1 - H(s/L_F)H^N(s)exp(-T_c s)} \tag{3.13}$$

Since both the numerator and the denominator of R_{cr} are zero at $S = 0$, the response for low frequency jitter can be examined using l'Hopital's rule. With 'first-order' type stations, $H(s)$ will have the form

$$H(s) = \frac{1}{1 + s/q} \tag{3.14}$$

where q is the dominant pole frequency. Its derivative, evaluated at $S = 0$ will have the form

$$H'(0) = -\frac{1}{q} \tag{3.15}$$

and for higher order clock recovery systems, the derivative approximates to a scalar multiple of this. After some simplification we obtain

$$R_{cr}(0) = \frac{1}{L_F(T_c q + N) + 1} \tag{3.16}$$

It is clear that this is always less than unity, and therefore the pole of $R(s)$ is accommodated by the recovery circuit. It also shows that the low frequency alignment jitter spectral density increases as L_F is decreased towards an open-ring. The alignment jitter will be lowest when L_F takes its maximum value of unity, which is when the NBCR has degenerated into an SCR. Despite this, the total RMS accumulated jitter will be lowest when L_F is zero.

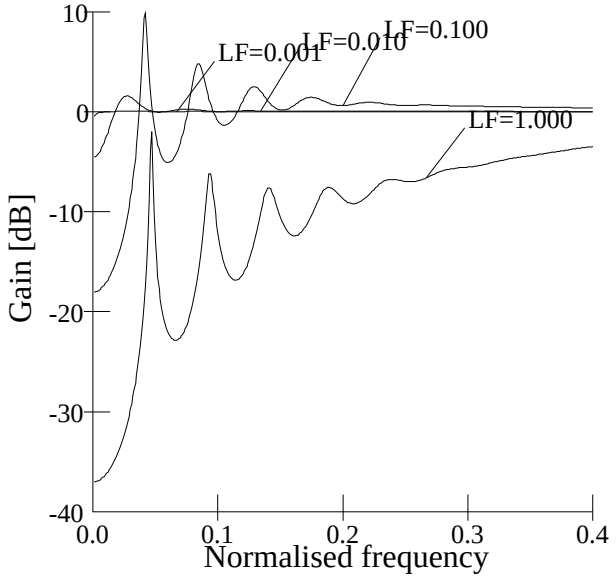

Figure 3.3: Jitter amplification response for fixed cable length and various narrow station bandwidths.

The frequency domain behaviour is illustrated in figure 3.3 where R_{cr} is plotted for $N = 50$ stations with four values of L_F. The x axis is the frequency normalised to the pole frequency q. An important parameter is $T_c q$ which was taken as 94. This was based on a high-speed local-area network typical dimensions: an assumed cable length of 3 km giving a delay of $15\mu s$, and a pole frequency of 1 MHz.

As the length of the ring is increased, T_c increases in proportion. As the bit-rate is increased, with constant Q clock recovery, q will also increase in proportion. Therefore the product $T_c q$ is proportional to the network geometry, and the parameter $R_{cr}(0)$ of the closed-ring technique will improve further as the geometry is increased.

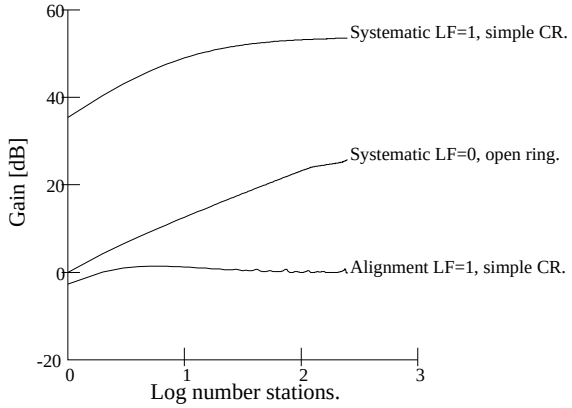

Figure 3.4: Comparison of closed-ring techniques with one unit of cable.

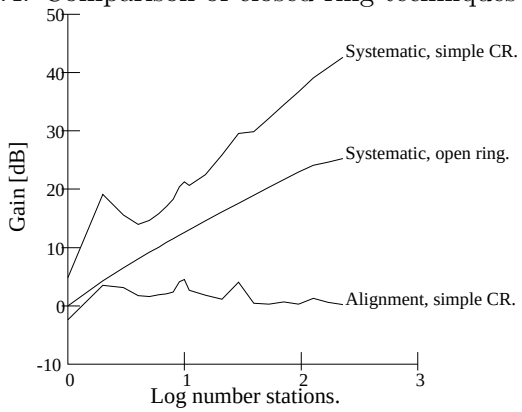

Figure 3.5: Comparison of closed-ring techniques with 94 units of cable.

Figures 3.4 and 3.5 show RMS jitter levels on a closed-ring with T_c values of 0 and 94 respectively. They plot RMS jitter versus number of stations and were obtained by applying Simpson's rule to

$$\sigma_N^2 = \frac{1}{2\pi} \int_0^\infty \left(\Theta_{orN}(j\omega) R_{LF}(j\omega)\right)^2 d\omega \tag{3.17}$$

in order to obtain the total RMS jitter and

$$\phi_N^2 = \frac{1}{2\pi} \int_0^\infty \left(\Theta_{orN}(j\omega) R_{cr}(j\omega)\right)^2 d\omega \tag{3.18}$$

for the alignment jitter at input to the narrow-bandwidth station. Again the y axes have been normalised to σ_{or1}^2. The graphs were plotted with L_F equal to zero and one. In the former case the narrow-bandwidth station sees all of the ring jitter as alignment jitter and the two integrals give the

same result. Therefore, when L_F is zero, the alignment jitter and the total jitter are equal to the systematic jitter gain found on an open-ring. This is a straight line with the mean-square jitter proportional to N.

When L_F is unity, the ring is of the SCR type. The total jitter is greater than for the open-ring owing to the feedback of the SCR. In figure 3.4, the normalised cable delay was set to unity. With this value, the closed-ring's feedback contribution to jitter gain R_{LF} does not introduce multiple peaks and troughs as the number of stations is varied. The resulting curves are smooth. The alignment jitter remains at 0 dB which is the same as the value expected for the intermediate stations of an open-ring, and the total jitter maintains a level of at least 30 dB above the equivalent value for the open-ring.

In figure 3.5, the cable delay was restored to the HSLAN value of 94. For this more realistic case, there are several values of N for which the closed-ring function R_{LF} has strong jitter gain. This causes peaks in the graphs of σ_N^2 and ϕ_N^2 as N is varied, but these are not always correlated because the dominant contribution to the total jitter integral is at very low frequencies, whereas the dominant contribution to the alignment jitter integral is from the first peak in the R_{cr} function, as shown in figure 3.3.

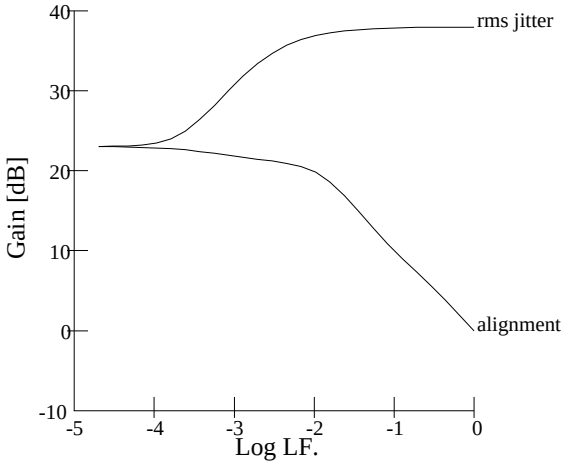

Figure 3.6: Closed-ring alignment and RMS jitter versus narrow station bandwidth.

Figure 3.6 again shows the closed-ring alignment and total jitter at the input to the narrow-bandwidth station, but this time plotted against $\log_{10} L_F$

for fixed N of 100 stations. It is interesting to note that the two curves do not bend at the same point. As the L_F is increased from zero, the RMS jitter starts to increase immediately while the narrow-bandwidth station sees much the same amount of alignment jitter. When a bandwidth of about one percent of the standard bandwidth is reached, the RMS jitter does not increase any more and further increasing L_F starts to serve a useful purpose, viz, reducing the alignment jitter at the narrow-bandwidth station. Hence there is no benefit to a mid-range value of L_F, or in other words, the NBCR is no use.

This issue of whether second-order slightly peaking PLLs can be used can now be addressed. As mentioned earlier, if H(s) is greater than unity at any point, then with the SCR technique there are cable lengths where the ring is unstable. Now that the NBCR has been ruled out, the only possible application for such stations is the conventional open-ring.

3.4.4 Closed-Ring Summary.

The closed-ring can expect at least 30 dB more jitter than the open-ring. Although this may seem alarming, and certainly measurements with oscilloscopes and spectrum analysers look alarming, the excess jitter is generally confined to very low-frequencies. Accordingly, the alignment jitter, which is the critical factor in determining λ_S, is generally no higher than that expected on the equivalent open-ring. Since the closed-ring does not use an elastic buffer, there are no elastic buffer overflows and λ_E can be taken as zero.

However, there are critical geometries where the alignment jitter of the closed-ring does increase. Multiple further configurations of cable length and number of stations have been studied, although these are not separately reported here. In these investigations, the greatest increase observed was 10 dB which is over half the safety margin determined is section 3.3.1. There remains the possibility that certain configurations of stations and cable will result in higher than usual levels of recirculating jitter, causing frequent failure through cycle slip.

3.5 The Asynchronous Clock Method

This section first describes the hardware requirements of the asynchronous clock method and then gives expressions for the expected failure rates for the token and slotted rings. There are two possible modes of failure: elastic buffer overflow and flip-flop metastability.

Under the asynchronous clock method, a station experiences a small difference between its incoming and outgoing line-rates. Such a station cannot transmit the same number of symbols per second as it receives. In consequence, the frame format must allow for pad symbols which can be inserted or deleted by the station without altering the meaning of a frame. Stations which are receiving faster than they are transmitting must delete pad symbols from their input stream and stations transmitting faster than they are receiving must insert extra pad symbols into their output.

The inserting and deleting is performed by an elastic buffer which forms part of the station's high-speed line interface. This is a slightly more complex type of elastic buffer than is required to close a ring of the open ring type owing to the different input and output rates. The more complex buffer which is capable of inserting and deleting can also be used in applications where the rates are equal, but the reverse is not always true.

In order to control their insertion and deletion, the presence of pad symbols in the elastic buffer must be detected by the ring access logic. If pad symbols are constrained so that they can only occur at certain points in the MAC level frame format, then the necessary information may be present in the counters and logic which operate at that level. However, speed and complexity issues normally dictate that the pad symbol recognition must be performed wholly inside the line interface. The pad symbol must therefore be a unique sequence which does not occur in the data. Alphabetic block codes, such as 4B5B, lend themselves well to this application, since non-data characters can be easily incorporated without significant additional coder complexity.

The asynchronous clock method requires that each station have a local transmit oscillator whose frequency is the nominal ring bit rate. In practice, it will be offset by a small, static, random positive or negative error. Whether a station is of the inserting or deleting type is determined by the difference in frequency between its own local oscillator and the oscillator at

its upstream neighbour. With a fixed ring configuration, stations are unlikely to switch from inserting to deleting mode. However they must have the capability to switch mode in order to accommodate the case of drift when two adjacent stations are also very close in frequency.[4]

A station will insert or delete symbols at a more or less constant rate equal to the difference between its input and output line-rate divided by the pad symbol size. The phase of these beat components is independent of MAC events such as the passing of a token or, for a slotted ring, the ring rotational phase.

If the interval between successive pad symbols arriving at a deleting station is too great, the excess bits accumulated will exceed the local elastic buffer capacity and the station will be forced to delete valid data. In contrast, an inserting station cannot fail in this way, since it does not depend on the pad density of the arriving stream, and is able to generate a new pad symbol at any time.

The behaviour of the elastic pad symbols for a token ring differs from that for a slotted ring. The frame format on a token ring is written by the active transmitting station. This is the unique station currently holding the token. The active station is able to insert pad symbols into its output stream in an arbitrary pattern at any appropriate density. The useful life of a frame is exactly one ring round-trip delay; after this the frame has no function and is stripped. The frames for a slotted ring, in contrast, are written once by a monitor station and are designed to circulate continuously. Therefore we must consider the two MAC methods separately.

3.5.1 Consequences Of The Asynchronous Clock Method For The Token Ring.

A ring using the token passing MAC method will have line idle periods when there is little or no data to be transferred. The idle state presents no justification problem since it may be represented with a continuous stream of the elastic pad characters.

When the ring is saturated with traffic, each station will want to transmit

[4]The Cambridge Backbone Ring access chip incorporates a frequency discriminator for this purpose.

as soon as it gains possession of the token. As mentioned above, the transmitter is obliged to include pad symbols in its output stream. One approach, used for example in the FDDI [BURR 88] and MST [MOLLENAUER 86] token ring formats, is for the transmitter to send a fixed preamble of N_0 (typically 10) pad symbols after gaining the token and before each packet. It does not include additional pad symbols in the body of the packet. This immediately presents an upper limit on the frame length that can be transmitted since the bits gained or lost during the frame must be contained wholly within the elastic buffers of the stations.

If the maximum frequency ratio between the clocks at two adjacent stations is specified to be less than ε and the elastic buffer capacity in bits is B_{cap}, then the maximum frame length is

$$L_{max} = \frac{B_{cap}}{\varepsilon} \tag{3.19}$$

With a simple two-phase elastic buffer design, such as the type implemented for the Backbone Ring (see chapter 7), B_{cap} is about two bits. Using uncompensated quartz oscillators, ε will be better than 100 ppm. These figures give $L_{max} = 20000$. This is not an unreasonable value since, if an 8B10B block code is used, such a frame would contain 2000 bytes. For comparison, the ethernet standard specifies a maximum frame length of 1500 bytes.

The probability of failure owing to all of the N_0 initial pad symbols being deleted can be evaluated as follows. Let the random, static frequency offset at station i from the nominal clock rate be x_i. Then the rate at which station i inserts or deletes symbols is given by

$$\left| \frac{x_i - x_{i-1}}{S} \right| \tag{3.20}$$

where S is the pad symbol size in bits when modulated onto the medium. If the x follow a Gaussian distribution with zero mean and standard deviation σ_x, then the expected value of this is[5]

$$\text{Symbol slip rate} = \frac{2}{\sqrt{\pi}} \frac{\sigma_x}{S} \tag{3.21}$$

and the fraction of frames on which a typical station performs an insert or delete action is

$$\rho = \frac{\text{Symbol slip rate}}{\text{Frame rate}} = \frac{2\sigma_x}{S\sqrt{\pi}} \times \frac{P'_s}{T_r} \tag{3.22}$$

[5]This is because if x is a Gaussian random variable with zero mean and standard deviation σ_x, the expected value of $|x|$ is $\sqrt{2/\pi}\,\sigma_x$.

where P'_s is the frame size when modulated and T_r is the transmission rate.

The buffers will be most active with a continuous stream of maximum length frames, that is, when $P'_s = L_{max}$. A typical value of ρ can be obtained by assuming two further simple relationships. First we assume that 99 percent of manufactured oscillators are within the ε tolerance, which gives

$$\frac{\varepsilon}{\sigma_x / T_r} = 5.152 \tag{3.23}$$

Second, using a value from the CBR design we obtain $B_{cap} = S/5$ and equation 3.22 evaluates to

$$\rho = \frac{2}{\sqrt{\pi} \times 5.152 \times 5} \approx 0.05 \tag{3.24}$$

Given values of ρ and N_0, the probability of ring failure through trying to delete more than $N_0 + 1$ pad symbols can be evaluated. The number of pad symbols in the header of a token ring frame, as the frame progresses from station to station, can be modelled as a random walk. The passing of a station by the frame corresponds to an epoch of the walk and at each epoch the number of pad symbols may be incremented, decremented or left unchanged. Let the number of pad symbols after station n be denoted S_n. It is convenient to take a false origin at N_0 by assigning $S_0 = 0$. The probability of an increment will be $\rho/2$ and the probability of a decrement will be the same. The probability of there being no change is then $1 - \rho$.

After the nth station let there have been y increments, z decrements and x no changes. $x + y + z = n$. Let $p_{n,V}$ denote the probability of S_n having value V. By considering a nested pair of binomial distributions, one in x and one in y, we can write down

$$p_{n,V} = P\{S_n = V\} = \sum_{x=0}^{n-v} \left(\begin{array}{c} n-x \\ y \end{array} \right) 2^{-(n-x)} \left(\begin{array}{c} n \\ x \end{array} \right) \rho^{n-x} (1-\rho)^x \tag{3.25}$$

The number of increments y is not a free variable since it is determined by the simultaneous equations

$$\left. \begin{array}{l} n - x = y + z \\ V = y - z \end{array} \right\} \Rightarrow y = \frac{n - x - V}{2}$$

Terms in the summation where y is fractional must be ignored since they represent impossible paths.

The probability of all N_0 initial pad symbols being deleted before station n is obtained with an argument similar to the Bernoulli first-passage theorem presented by [FELLER]. This is based on the *reflection principle* which uses a geometrical description of a random walk of length n as a 2-D path from the origin $(0, 0)$ to the point (n, S_n).

> **Reflection principle:** The number of paths from the point A to point $B(n, k)$ which touch or cross the x axis equals the number of all paths from A to the point $B'(n, -k)$ which is the reflection of B in the x axis.

Feller's proof only considers the Bernoulli case, but the principle also holds for the three-way random walks which apply to the asynchronous clock where it is possible (and even probable) that there is no change at a particular epoch. From the reflection principle, the number of paths which touch or cross the line $S = r$ on the way to point $A(n, k)$ is the same as the total number of paths from that origin to $A'(n, 2r - k)$. This leads to the

> **General reflection principle:** $k \leq r$ only. The probability that a path of length n leads to $A(n, k)$ and at some point touches or exceeds the line $S = r$ is $p_{n,2r-k}$ which is equal to $P\{S_n = 2r - k\}$.

If $k > r$ then the path has certainly crossed $S = r$ and so the probability is unity.

The probability that an attempt to delete $N_0 + 1$ or more pad symbols occurred is obtained by summing the probabilities given by the general reflection principle with $r = N_0+1$ over all possible finishing points $A(N, -N)$ to $A(N, N)$. N is the number of ring stations.

$$P\{\text{failure}\} = \sum_{k=-N}^{N} p_{N,2(N_0+1)-k} \times p_{N,k} \tag{3.26}$$

The first factor is replaced with unity when $k \geq N_0 + 1$. Investigation has shown that for the small values of ρ which are typical when using quartz oscillators, say less than 0.1, the terms which have significant contribution to the failure probability are those where the number of inserts and deletes $(y + z)$ is either N_0 or $N_0 + 1$. This is the case where the number of pad symbols has decreased to zero taking an exactly or nearly monotonic path.

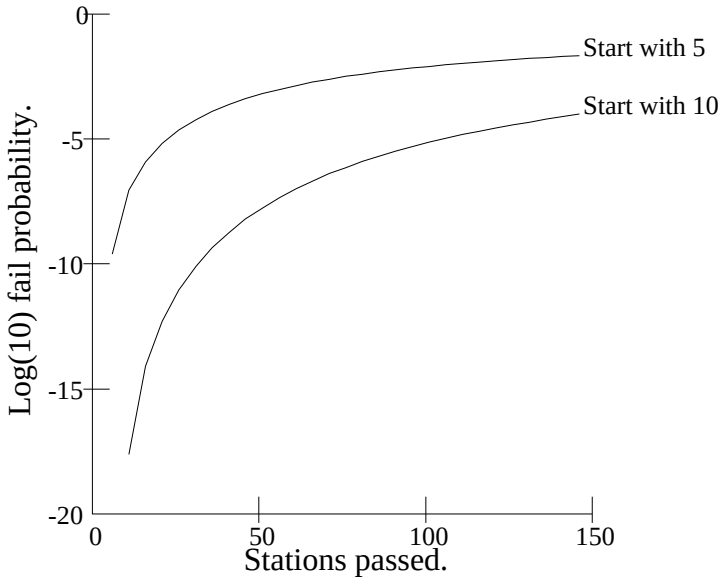

Figure 3.7: Probability of all the pad symbols being deleted on a token ring.

Figure 3.7 shows a plot of $P\{\text{failure}\}$ as the number of stations is increased. The example values of N_0 are five and ten. It is clear that 5 initial pad symbols are insufficient whereas 10 give respectable performance. With 10 initial symbols, the failure rate is about 1 in 10000 frames, even with 150 stations. The figure of 0.05 for ρ assumed a continuous stream of maximum length frames and large frequency errors between each station. In practice, the average value of ρ will be lower and reliability will be increased. In addition, most frame headers on an (early release) token ring do not contain the active token, so a frame header failure is liable to only cause loss of the current frame. This is relatively unimportant. Therefore, the asynchronous clock technique would appear to be ideal for a token ring. For the slotted ring, which is analysed shortly, a frame header loss is far more serious. The frame structure must be rewritten, resulting in a temporary network outage experienced by all traffic. It is also shortly revealed that there are problems managing the pad symbols on the slotted ring.

3.5.2 Free Ranging Pad Symbols.

A discussion of the asynchronous clock method would not be complete without mentioning free ranging pad symbols. This is an alternative approach to pad symbol distribution that does not constrain the pad symbols to preallocated sites in the frame structure, rather it allows them to freely roam through the frame body. In this case, clearly the pad symbol must be non-data. The first transmitter initially puts pad symbols at regular intervals,

but the action of the subsequent stations disturbs this pattern until eventually there is a random distribution. This approach decouples physical layer requirements from the packet length limitation, but as demonstrated in the last section, the former approach can provide sufficient packet length capacity for current requirements. The behaviour of free ranging pad symbols is in essence the same as the behaviour of pad symbols on a slotted ring, only with a finer granularity. The behaviour of pad symbols on a slotted ring is studied in the next section. Certain difficulties are discovered and these also apply to the free ranging pad symbols. Therefore free ranging pad symbols are not considered further.

3.5.3 Consequences Of The Asynchronous Clock Method For The Slotted Ring.

With a slotted ring, such as the Backbone Ring of chapter 6, the ring format is initially written out at one station (the active monitor) with a preset density of pad symbols. The velocity of propagation of the ring data (non pad symbols) will always remain at that used by the active monitor initially. The ring length will not in general be an exact integer multiple of the frame length so the monitor leaves a gap which it fills with additional pad symbols. After a few thousands of ring revolutions, which take just a few minutes of operation, the gap disappears, its pad symbols having been redistributed along the ring length.

Since the clock frequency errors around the ring sum to zero, the rates of inserting and deleting pad symbols are exactly matched. The average number of pad symbols does not vary. The behaviour of the ring stations inserting and deleting pad symbols is as though pad symbols are continuously selected at random and moved from their current position to a new site, also chosen at random. In such a system, the number of pad symbols at each frame boundary will be a random variable which will approximate to an exponential distribution. The most likely number of pad symbols at a boundary is zero. The distribution will be slightly perturbed from a true exponential shape by a clustering tendency of the boundaries with zero pad symbols. (It will become apparent that this further degrades performance.) Ignoring the clustering effect, with F boundaries and P pad symbols on the ring the expected number of boundaries with zero pad symbols is

$$E(\text{zero}) = \frac{F^2}{P(1 + F/P)} \tag{3.27}$$

and the ratio of the expected number of boundaries with n pad symbols to $n+1$ is

$$\frac{E(n)}{E(n+1)} = 1 + F/P \tag{3.28}$$

The ring will fail if a deleting station exceeds its elastic buffer capacity before a pad symbol arrives. The elastic buffer capacity can be expressed as a figure, D_{max}, which is the number of frames that may pass after a station decides to delete before its elastic buffer bursts.

$$D_{max} = \frac{B_{cap}}{\text{Bits accumulated per frame}} = \frac{B_{cap}T_r}{(x_{i-1} - x_i)P'_s} \tag{3.29}$$

which, under worse case clock differences, is

$$D_{max} = \frac{B_{cap}T_r}{\varepsilon P'_s} \tag{3.30}$$

If $D_{max} \geq F$ then the ring revolution time is less than the buffer bursting time. Since the pad symbols are always somewhere on the ring, the physical layer cannot fail under this condition. The number of frames F is a result of the geometry

$$F = \frac{LT_r}{P'_s c} \tag{3.31}$$

where c is the velocity and L is the length of the medium. Hence

$$\begin{aligned} Dmax &\geq F \\ \frac{B_{cap}T_r}{\varepsilon P'_s} &\geq \frac{LT_r}{P'_s c} \\ LT_r &\leq \frac{cB_{cap}}{\varepsilon} \end{aligned} \tag{3.32}$$

which, with figures taken from the Backbone Ring[6] evaluates to a frequency-length limitation of 4 gigabit-kilometres. This figure is evidently rather small for envisaged MAN applications, so the next step is to consider the probability of failure when working outside this limit.

A ring with F frames on average has $E(\text{zero})$ boundaries with no pad symbols. A station wishing to delete will examine each boundary as it passes and delete from the first boundary with non-zero pad symbols. The

[6] $c = 2 \times 10^8$, $\varepsilon = 100 \times 10^{-6}$, $B_{cap} = 2$.

probability of it having to examine D_{max} or more boundaries is given by the formula for random selection, without replacement, from a population of F. Let A denote the number of boundaries with zero pad symbols, then the probability of failure is equal to the probability of the first D_{max} choices having zero pad symbols which is

$$P(\text{fail}) = \frac{A!(F - D_{max})!}{(A - D_{max})!F!} \tag{3.33}$$

An acceptable physical layer failure rate might be once per day. If there are 10^6 frames per second, the probability given by equation 3.33 must be about one in 10^{11}. Unfortunately, with typical physical parameters, equation 3.33 either evaluates to zero, which corresponds to operation within the limit of equation 3.32, or to values between 10^{-2} and 10^{-4}. The non-zero values imply almost immediate failure.

3.5.4 Hybrid Mode Solutions.

It is possible to operate outside the limit of equation 3.32 provided that the number of asynchronous stations is limited to one or two. If a single station uses the asynchronous clock method and the remainder are regenerative, then the network degenerates to the open-ring type. If, however, there are two asynchronous stations which divide the ring into two regenerative sections, then the ring can potentially support twice the number of stations as an open-ring of similar design. Of the two asynchronous stations, one will be inserting and the other deleting. They will both insert or delete at exactly the same rate; the rate being equal to the difference in their local oscillator frequency divided by the pad symbol size. This gives the ring a stable operating point, which it is likely to enter, where the inserting station inserts pad symbols at precisely the points where the deleting station will delete. However, this hybrid scheme is not an elegant solution.

As an alternative attempt at solving the problem of pad symbol distribution, it is possible to imagine a special station inserted into the ring which has an especially spacious elastic buffer. If stations with various size elastic buffers are mixed on one ring, then the station with the largest value of B_{cap} dominates, and its buffer size should be used in equation 3.32 to determine the maximum network size. A single station with increased capacity is thus able to redistribute the pad symbols for the whole ring and raise the maximum geometry. However, this is again not an elegant solution, since it

involves considerable extra investment in the special station. It is, however, the only known way to use the asynchronous clock technique with arbitrarily large geometries.

3.5.5 Metastable Failure.

The asynchronous clock method requires asynchronously operating logic. Whatever the design, the fundamental requirement is for a circuit which achieves the same behaviour as a D-type flip-flop with asynchronous clock and data inputs. In practice, a D-type is often used.[7] No practical decision circuit can operate with zero aperture time. In a real circuit, if the input changes during the aperture, metastability can result. The asynchronous logic of the line interface will therefore enter the metastable state from time to time. If a flip-flop which is carrying active information as part of the data path enters a metastable state, a bit error is introduced and the data is lost. In a well designed system, the rate of data path flip-flop aperture violation must be minimised. This is usually achieved by arranging for the violations to occur mainly in auxiliary flip-flops. The output from these flip-flops is used to adjust the configuration of the main data path in order to avoid errors in the main data path. Although there remains a finite probability that a metastable condition in an auxiliary flip-flop propagates through the logic and into the main data path, this can be made arbitrarily low by cascading several stages of auxiliary flip-flops.

An estimate of the rate of failure from metastabilities in the Backbone Ring design is presented in chapter 7. The result, in common with some other asynchronous designs, is several orders better than a rate of once per year, and therefore metastable failure need not be a significant factor in the asynchronous clock technique.

3.5.6 Asynchronous Summary.

The major benefit of the asynchronous clock technique is that regenerated clocks are not used and therefore clock jitter does not impose a geometry limit. The additional elastic buffer hardware that is required presents no problem if custom LSI is used for the line interface, and the reduced

[7] The Backbone Ring access chip in fact uses two asynchronous D-types because it requires the knowledge of which of its two clock inputs is the higher in frequency.

requirement of the clock recovery circuitry more than compensates. The requirement for serviceable pad symbols is easily incorporated into the class of frame formats expected in a packet switching network. The probability of failure owing to flip-flop instability is acceptably low. A MAC which is intrinsically able to regulate the distribution of pad symbols, such as the token system, has been shown to offer acceptably low elastic buffer overflow rates with a minimal overhead of pad symbols. However, if there is no such distribution control, then the asynchronous clock technique is strictly limited in geometry for any number of asynchronous stations above two.

3.6 Clock Distribution Summary

This chapter has discussed the available techniques for clock distribution over gigabit metropolitan area networks.

The open-ring technique has been re-evaluated, taking into account the effects on systematic jitter accumulation of the data path delay with a station. This was found to have little effect with more than a few tens of stations. The jitter at the end of a chain of one hundred stations at 1 GHz was considered. With good clock recovery, such as can be achieved using SAW filters at gigabit rates, an open-ring was shown to be reliable with at least 100 stations. However, for experimental systems where it is desirable to have some flexibility in the system clock rate, SAW devices are not ideal, and alternative clock recovery techniques should be considered.

The closed-ring technique has been evaluated and shown to operate reliably under most circumstances, but the possibility of critical configurations where the jitter margins are eroded has also been identified. The requirement for a clock control station has been shown, but it has further been shown that this station should have the same AC jitter characteristics as the other stations.

The asynchronous clock method has been considered for both slotted and token rings. It has been shown to offer acceptable service for the token system, but to introduce geometry limitations for the slotted ring. However, the geometry limitations are not too severe, especially if clock frequency tolerances can be tightly controlled. Devices such as compensated crystal oscillators should be used. Some particularly attractive devices are now becoming available. See for example [CURRY 88].

The two-stage clock recovery method has been suggested as an easy to engineer solution. It can be used in conjunction with both the open and closed-ring techniques. With the open-ring, it will result in lower alignment jitter at the elastic buffer than with a normal open-ring. With the closed-ring, operation remains somewhat unpredictable and so should probably be avoided.

The Backbone Ring implementation uses the open ring method with two stage clock recovery and a dual-chamber helical resonator for the first stage.

Chapter 4

Access Control for Large Geometries

This chapter examines all aspects of multi-access media access control protocols, or MACs. The aim is to identify and examine the components of a MAC make it suitable for use at large geometries and on a multi-channel physical medium. The chapter does not attempt to list and describe individual MACs separately. Rather, it offers a taxonomy of MAC components which may be combined to form a MAC, and uses specific MACs, described in the literature, as examples.

This chapter restricts itself to qualitative comparison. Analytical comparison and comparison through simulation of access protocols is presented in the next chapter, Chapter 5, but the discussion there is confined only to ring access protocols.

Section 4.4 discusses physical layer packet size issues. Section 4.5 describes the basic access control techniques and section 4.6 presents the consequences of partitioning a network, regardless of access method. Section 4.7 combines these previous two sections, examining how MACs might behave over a partitioned channel.

4.0.1 Features found in old and new MACs.

From the early work on local-area networks in the 1960's and 70's, there has grown up a set of criteria that a MAC should satisfy. These 'traditional' goals of a MAC can be listed as follows:

High channel efficiency: The channel efficiency is defined as the saturated network throughput as a fraction of the channel capacity. It should be as high as possible.

High point-to-point bandwidth: It is not desirable to have a MAC inherent limit on the maximum fractional utilisation offered by a single station. A MAC which can accommodate stations of various cost that are able to achieve proportionally variable point-to-point bandwidth is desirable.

Low access delay: The access delay should aim to be low, bounded and to have low jitter.

Provide fairness: Equal and fair access must be guaranteed to all stations, both for transmit and receive. The cost to merit ratio of this facility at large geometry is considered (section 4.2).

Avoid deadlock: The protocol must be free from dead states. Any error recovery processes should avoid interruption of other traffic and occur quickly in order to reduce wasted bandwidth.

Avoid metastability: Performance metrics, such as delay moments, should approximate to a stationary, monotonic function of the applied load, regardless of the order the loads were applied.

Destination independence: Fairness and performance guarantees should not rely on homogeneous traffic flow. Performance must be acceptable under asymmetric loads.

Geometry Insensitive: The MAC should offer acceptable performance over the widest possible range of system clock rates, and more importantly, installed network cable length.

Packet size support: The protocol may support multiple or varying packet sizes.

Simplicity: A simple, distributed protocol is preferred to a protocol which relies on intelligence within the stations to perform parameter measurement and access control.

In this current study, the basic criteria in the above list will hardly be mentioned. All of these criteria are satisfied by the MACs discussed. Instead, we shall concentrate on a new set of features. These criteria become significant when a network is carrying multiple classes of traffic, and it will be shown that they become harder to meet as the network geometry increases.

4.0.2 Requirements of multi-service MACs

The study in this chapter is directed at a network carrying both real-time and non-real-time traffic, therefore the most important additional aspect of a MAC becomes its ability to maintain acceptable delay and jitter performance for the real-time traffic, and to isolate the real-time traffic from the effects of other traffic. We therefore must look carefully at a MAC's ability, at large geometry and on multi-channel medium, to support multiple levels of priority, delay guarantees, and the cost, performance and role of MAC layer load balancing mechanisms.

Fine-grain sharing: The granularity of sharing should be as fine as possible in order to minimise the delay for small messages.

Support of priority: The queuing action of the protocol may support more than one distributed queue. The MAC should insulate the higher priority traffic from the characteristics of the lower priority traffic. (The same must apply if queuing or discard of traffic occurs outside the MAC.)

Fine-grain load balancing: The load balancing mechanism must amortise fairness over as short an interval as possible, the lower limit being one network latency (defined in section 2.1.3). This is especially important if the MAC does not support multiple access priorities. See section 4.2.

Multicast The ability to simultaneously route the same messages to a selection of destinations is useful, for example, it can be used by a multi-media stream server.

4.1 Definitions.

This section introduces some of the jargon of media access control, along with comments.

4.1.1 Multi-access Media Access Control (MAC)

A *multi-access* network consists of a shared medium, such as a group of star-coupled optical fibres, a radio frequency, or a regenerative bus or ring, onto which a number of client stations broadcast messages according to the access control rules, or MAC. The stations also include a receive side which monitors the medium, recognising and receiving messages addressed to them.

4.1.2 Spatial reuse.

Certain shared-media, for instance rings and regenerated busses, are able to take advantage of *spatial reuse,* where messages are deleted from the media at the destination, enabling the same piece of bandwidth to be reused on a different physical section of the network. Although systems which employ spatial reuse might not be considered true shared-media networks, they are able to offer a worthwhile increase in available bandwidth. An example is the destination-release slotted ring introduced in section 5.1. Another deviation from 'pure' multi-access occurs in register insertion systems. These transmit their own data on to the medium, while data which is already on the medium is removed and stored in internal buffers for later transmission. Significant examples are SILK [HUBER 83], the BCMA/CRMA family from IBM Zurich [AS 91, PITT 91, HEINZMANN 91], and from IBM Yorktown, Metaring [CIDON 90]. Buffer insertion has certain benefits for priority data since a station may inject its own traffic immediately. Unfortuantely, this causes a delay to be experienced by data already sent by other stations, so the benefit is not always realised.

4.1.3 Self-stripping.

The concept of a *self-stripping* medium has already been defined in section 2.5. It is a medium where special action by the network stations is not necessary in order to remove from the medium transmissions which have already reached their destination. Rather, the topology intrinsically prevents the recirculation or retransmission of old frames.

4.1.4 Large Geometry.

By *large geometry*, it is implied that the transmission line-rate on the medium and the physical area covered by the medium are in the MAN rather than the LAN region. In particular, a line-rate between 1 and 10 Gbit/seconds and an area covering approximately 50 km is implied. Messages as short as single 48 bytes ATM cells are envisaged. Arithmetic combination of these figures shows that networks with up to 3000 messages stored in the network medium are being considered. This has a significant impact on MAC behaviour and design.

4.1.5 Multi-channel or partitioned media networks.

A multi-channel or partitioned network uses a physical medium where the bandwidth is divided into multiple separate channels. The division may be artificially introduced, for instance, through the use of multiple TDM channels on a high bandwidth link, in order to reduce electrical complexity in the stations; or the division may be intrinsic, such as with a wavelength division multiplexed (WDM) optical fibre, where different colour carriers are used for each channel.

As mentioned in chapter 2, the promising progress of coherent optical transmission, using frequency-division multiplexing, is likely to lead to multi-channel fibres becoming the norm. In the meantime, electronic TDM of simple directly modulated monochrome fibre channels remains attractive, since only the multiplexer and demultiplexer electronic components need handle the full bandwidth. This latter approach is taken in the Backbone Ring design (chapter 6), where it is taken further, and used to match the the bandwidth of a shared-media fibre optic ring to the amount of bandwidth desired through the host interface. Another example is the DECT common air interface for cordless devices where each of eleven frequency bands is TDM partitioned into 10 duplex 64 kbit/sec channels.

4.1.6 Hybrid MACs.

With a multi-channel medium, it is not necessary to use the same MAC on each channel. One question of particular importance is whether one set of MAC rules can be sufficient for all of the traffic types that were

envisaged in chapter 1. The alternative is to form a hybrid network with different MACs operating in different partitions. Examples of such networks and related discussion is included in section 4.8. Obviously, hybrid MACs increase the complexity of an overall network architecture and introduce additional complexity for applications which wish to work with traffic from different transport modes, but their provision of complete isolation of one traffic class from another does enable verifiable service guarantees to be granted.

4.2 Fairness, Priority and Delay Guarantees

Fairness, Priority and Delay Guarantees are here grouped into one section because they from an approximate set of spanning vectors for the space of transmission control rules which are available to support and distinguish between separate qualities of service.

The approach taken in this section is first to define these rules, and then to explore what could be provided, at large geometry, by an ideal MAC. There then follows a look at mechanisms for enforcing the transmission control rules and their cost in terms of loss of performance. For instance, it is well known that fairness can contradict maximising aggregate network throughput. The other main competition is between delay guarantees and speedy support of bursty traffic when the expected statistics of the burty traffic are used as a basis of the delay 'guarantee'.

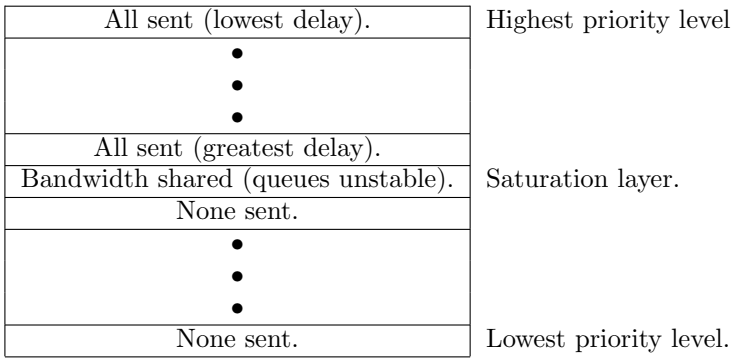

Figure 4.1: A well behaved (delay) priority mechanism.

The role of a *load balancing mechanism* is to divide the available band-

width fairly between users. When multiple levels of priority are supported, the loads should be balanced as shown in figure 4.1. This shows a canonical representation of a network with multiple priority levels where the offered load exceeds the bandwidth. In the higher levels of priority, all of the offered load is served, the upper levels experiencing lower delay. One particular level experiences saturation and the remaining bandwidth available is shared fairly between the contending users at that level. The distribution is controlled by the load balancing mechanism. If there is any traffic below this level, it should be starved of service.

Sometimes *delay priority* is separated from *loss priority* (e.g. in ATM or other switched networks). For a pure shared-media network (i.e. without register insertion or switching) this is not applicable. The only form of loss (excepting transmission errors) is infinite delay. In an interworking situation, i.e. when using bridges or switches, those messages with delay priority should be sent at higher MAC priority to those with loss priority, but of course, sufficient bandwidth must be available for both. If there is insufficient bandwidth, messages with delay priority should be thrown away first.

The diagram (figure 4.1) shows a steady state, but in practice, traffic sources are bursty and the saturation layer is continually moving up and down. The term *granularity of sharing* can be defined to refer to the time interval over which balancing is amortised. It is the minimum time for all users in a sharing set to be serviced. Related to this is another parameter which is the time for the bandwidth below a given priority level to become reallocated when a new source starts, stops, or varies its load, at that given level. The time for the lower-level sources to regain their correct allocations may be termed the *granularity of load balancing.* Sometimes, the two time intervals are governed by the same mechanism, for example, they are in a simple token ring. In the Orwell ring however, the sharing is finer than the load balancing [FALCONER 85a]. This is beacuse there are many slot times between reset intervals.

Assuming a packet network where transmission is sequential and where packets are not diced and interleaved, the shortest sharing cycle consists of the time for one packet from each active user. This is also the shortest interval over which fairness can be amortised (the granularity of load-balancing). When segmentation occurs, the interval becomes one segment, mini-packet or cell per user.

The bandwidth allocation described by figure 4.1 requires for its calculation, a global picture of the traffic generated by the network users. The traffic generated by individual users may be continually changing, requiring the corresponding bandwidth allocations to be continually rearranged. The information to determine the allocation must be distributed over the network, either implicitly or explicitly. Therefore, the response time of the load balancing mechanism cannot be less than a network latency. An access protocol where the delay in re-evaluating the allocations in response to a change of offered load is no greater than the larger of either a latency or the duration of a sharing cycle may be said to possess *perfect* load balancing.

4.2.1 Expedited Transfer Mechanisms.

An expedited transfer mechanism forwards packets which have been marked for expedited transfer in preference to other packets. This is typically achieved by maintaining separate queues within a station for each priority level, only serving a queue at a particular level if all higher queues are empty. Expedited transfer is nearly always used as an additional mechanism when the medium access control rules support their own priority mechanism, but it can also be applied if they do not. Expedited transfer only forms an effective mechanism for providing low delay for priority packets if the network granularity of sharing is low. For example, on a token ring where the granularity of sharing is a token rotation time, expedited transfer is often not very worthwhile (see section 5.2).

4.2.2 Bounded Delay Guarantees.

Networks and their load balancing mechanisms are sometimes designed to provide a bounded delay guarantee. This ensures that the delivery time of a message lies below a bound, regardless of the number of users. Excess messages are rejected at their source, rather than allowing the delivery time to exceed the guarantee. The MAC is generally free to assume that fairness of rejection has been resolved by higher-layer bandwidth management protocols. A central clearing agency for bandwidth distribution is normally postulated. On a multi-network system, the multiple instances of bandwidth allocation entities must communicate with each other using high-level protocols. Other components of these protocols can provide routing and virtual-circuit set-up services.

For certain critical real-time applications, such as controlling industrial robots or launching space craft, concrete delay bounds of the type just described are required.[1] For many other real-time applications, especially those using multi-media traffic, operation of a multi-access network such that the 99^{th}or 99.9^{th} percentile of delay meets a given requirement is often acceptable.[2] Some remarks are made on this subject in section 9.2.1. In terms of the model of figure 4.1, it is suitable to assume that delay bounds are only applied to a consecutive group of priority levels, starting from and including the highest.

4.2.3 Fairness Per Virtual Circuit.

In the context of LANs, the term 'users' in the above definition of fairness has effectively been synonymous with 'stations'. Equal bandwidth offered to each station has been the norm. However, a multi-access backbone network posed as a replacement for a more conventional packet forwarding centre may be expected to apply suitably more advanced balancing. For instance, an equal bandwidth share for each virtual circuit is sometimes proposed [KATEVENIS 87]. This approach would require modification to the MACs for which load balancing is implicit.

Fairness per virtual circuit may also be preferable to fairness per station on a single network, without bridging. For instance, in a client-server environment, providing fairness of access to the file server may require that the file-server's transmit bandwidth is disproportionate.

4.2.4 Load Balancing Requirements and Applicability.

At saturation, the jobs of the load balancing mechanism may be summarised:

- It should prevent stable and metastable bandwidth hogging patterns.
- The interface to a station outside the load balancing mechanism should be prepared to face and regulate excess offered load without network

[1] An example is the US Navy's tightening of the FDDI timed-token protocol for application in MIL-NET

[2] However, this remains open to debate, as exemplified by the provision of 'isochronous' services at the MAC level in 802.6 FDDI-II, and (more recently) BCMA, but not B-ISDN.

degradation.

- The saturated bandwidth share available to each user should be inversely proportional to the number of users, not proportional to the offered loads. Therefore either implicit or explicit knowledge of the number of active users is required.

- It should prevent unacceptable interference from lower priority traffic to the propagation of traffic in higher priority classes.

Before proceeding to examine methods for implementing load balancing on a multi-access network, it is sensible to bear in mind what is the cost, in terms of complexity, performance or other terms, that we are willing to suffer to support accurate load balancing. For instance, the source-release rule for a slotted ring (section 5.1) divides the available bandwidth, on average, by half, when compared to destination-release (ignoring for the moment the potential for a response in the return slot). Provided that the basic properties, listed above, are satisfied, the evenness of sharing is probably far less valuable than additional bandwidth. In fact, this additional bandwidth might often be able to move the network operating point away from the point where load balancing is required. A particular example is included in section 9.2.

4.3 Load Balancing Mechanisms

This section classifies and examines methods of load balancing. The 'distributed queue model' the 'self-timed round' and the 'fixed-length round' are introduced and defined. These are cyclic load balancing mechanisms. Several probabalistic algorithms are also introduced as cheap alternatives.

Not mentioned otherwhere is the trivial cyclic server solution. This is intrinsic to certain basic ring MAC's, such as pass-on-free slotted and token. These are able to offer fairness for one level of priority. Enhancements to support priority, beyond expedited transfer, are, however, described here, as well as in the next chapter.

4.3.1 Cycle-Based Load Balancing.

Cycle-based load balancing amortises fairness on a cycle-by-cycle basis. The idea is that after each cycle, if every user has received the appropriate, fair increment in service, then overall the users will remain in step, all receiving the correct quality of service.

The cycles themselves can either be of fixed length (e.g. LION) or variable length (e.g. Metaring or Orwell) and the start of a new cycle can be initiated always by a fixed station (e.g. the NEC ring), or by a different station each time (e.g. Metaring), or by a distributed algorithm (e.g.Ȯrwell).

Within the variable length cycle catagory of cycle-based load balancing methods, are found the 'distributed queue model' and the 'round model'. The round model employs counters and quotas and is similar to the queue model if the quotas are set to unity.

We look first at the distributed queue load-balancing method, using the DQDB access protocol as an example. DQDB, first known as QPSX, was adopted as the IEEE 802.6 metropolitan area network standard.

4.3.2 DQDB and other distributed queue models.

The distributed queue model poses a fictional non-distributed queue that is served by a single server which represents the network. In reality, the queue is distributed with various customers in each station. The algorithm is that, at any time, a user may enter a limited-size packet into the queue for transmission, provided that his last packet has been removed from the queue and served.

In DQDB, reservations are used to effect the distributed queue. DQDB is a dual bus network, and the reservations are communicated to upstream stations on the reverse channel. However, similar protocols have been proposed for folded bus networks where the reservations can be detected through the read channel. In DQDB, stations maintain a counter which holds the running difference between reservations seen on the reverse channel and empty slots seen on the forward channel. They take one packet from their local queue and note the value of the counter at this time. This is that packet's position in the fictional distributed queue. When that number of empty slots have been seen, the packet is transmitted and the process repeats.

A number of studies, for example [CONTI 89], have shown this technique is unfair for bus topologies, especially at large geometries, even in terms of static saturated-source[3] bandwidth. This results from the stations towards the tail-end experiencing greater reservation delays than the head-end stations, while all being constrained to operate with a window of one packet reservation outstanding. Evidently this static imbalance can be avoided by operating the complete protocol independently for each packet queued in a station. Another solution has been proposed by [FILIPIAK 89] where stations require knowledge of their position on the bus and the number of active stations. However, because of the inherent asymmetry of the ordered-access bus, the tail-end stations will always experience poor response from the sharing mechanism when their offered load increases. The 802.6 standard includes options for better fairness than provided by the basic model, but these are not compulsory in conformant implementations.

When distributed queue load balancing is applied to ring topologies, the inherent symmetry of a ring ensures that the users do not suffer from positional unfairness. The ring protocol invented by the author and described in section 5.4, indeed uses distributed queuing through reservations.

4.3.3 The Self-Timed Round For Load Balancing.

Self-timed rounds finish according to a distributed algorithm and can vary in duration. Fixed-length rounds are described in section 4.3.4.

In the round model of load balancing, there are cycles and quotas. At the start of a cycle, each user is allocated a quota of one or more packets. A user may transmit during the cycle until he has used his quota. A new round is commenced when all users have had a chance to transmit, but have failed to, owing either to having no messages to send or having used their quota. Quotas may either be specified in number of packets, or more sensibly, in number of bits. Using fixed length packets, the two are the same.

Quotas offer the potential for arbitrarily complex load balancing mechanisms. Stations might maintain several quota counters, indexed, for instance, by priority class, and various stations may be allocated different quotas. This enables sharing to be done per virtual circuit or, indeed, on

[3] A saturated source transmits at every oportunity, subject to the MAC and transmission control rules.

any other basis.

A round mechanism can also form the basis of a bounded delay guarantee, provided that no action which might extend the round duration is allowed. (This is guaranteed when using fixed-length rounds).

A short round offers fine-grain load balancing, but with the distributed algorithms, there is usually overhead concerned with starting a new round, so short rounds tend to reduce efficiency and a minimum round length of several network latencies is usual. This undesireably increases the granularity of load-balancing at large geometries.

In a token system, the round is a token rotation and stations transmit in order. However, a coarse-grain quota load balancing mechanism can be implemented over a network where the transmit access is shared with fine grain.

Perhaps the most prominent examples of self-timed rounds are found in the Orwell ring load balancing algorithm [FALCONER 85a], and in the John Limb 'Simple Algorithm' used in the Hangman network [WATSON 89, 92]. Both are fine-grain slotted systems. Orwell was envisaged for traffic consisting of predominantly voice and Hangman for traffic consisting predominantly of variable length data messages, so neither network emphasised a rigid MAC-layer priority requirement.

An aggregation of voice channels offers a constant load, and even when silence suppression is implemented, the aggregation is far less bursty than typical computer traffic. Under such loading a quota scheme works very well. Indeed, load balancing is hardly required; the main function of the quota then becomes to enable management to measure the rate of new cycles, and therefore determine whether to accept a new call, or whether the new call would extend the round duration unacceptably.

Using fixed quotas in a network loaded by bursty sources, for the delay guarantee to be effective, the quotas must be evaluated assuming all sources are in burst mode. Otherwise, a previously idle source may commence transmission close to the end of a round and thus extend the round beyond the acceptable maximum. Quotas based on peak utilisation will tend to be smaller than might be desired. This forces an undesirable compromise. A fixed, low quota for data traffic has been suggested for Orwell [MITRANI 86]. Only one or two data mini-packets may be sent, per station, per round.

Under this scheme, the full amount of bandwidth left unused by the voice cannot be used by the data and also the point-to-point data rate is very low. As an attempt to counter this, the voice quotas can be reduced. This results in shorter rounds and finer grain load balancing, but also an inevitable reduction in efficiency. Adaptive, dynamic quotas are also possible, but will increase the response time of a the balancing mechanism beyond a round, raising the granularity of load balancing.

As defined, a quota system is not a MAC layer priority mechanism: it merely provides load balancing within a single priority level. It is possible to implement separate quotas at each station for each priority class. However, this is of little benefit, since these stations are each competing for the same bandwidth. Although a certain amount of priority traffic may be sent by a particular station at the expense of its local traffic in the lower priorities, this traffic is not guaranteed precedence over the lower priority traffic of other stations. Quotas do not provide the necessary interference effect which must be available to give preference to high-priority traffic at poorly placed stations.

4.3.4 Fixed Length Rounds For Load Balancing.

An alternative round based load balancing mechanism, implemented for example in the NEC ring [SHIMIZU 87] and LION [LUVISON 89], involves the concept of major and minor rounds. It is the major round that is of fixed length, and this is often chosen to be an appropriate multiple or submultiple of 125 microseconds, for ease of connection to circuit-switched telephone equipment. The exact multiple depends on the installed geometry. (Certain hybrid MACs, e.g. FDDI-II and BCMA, also synchronise part of the network frame structure to the telephone clock, but these are described under the 'hybrid' title in section 4.8.)

The major round countains concatenated minor rounds of decreasing priority. Low priority data, transmitted in the later minor rounds, can be pre-empted by the start of a new major round.

A shortcomming of these schemes is as follows. If a new major round is always started by a fixed trigger station, then delay will be bounded, but the minimum length of a minor round must exceed one network latency since the trigger station must detect that sources of the current priority are

exhausted before the next minor round can be commenced. If it does not wait until they have been exhausted, typically the stations furthest away will suffer unfair service. This mechanism is very effective at rejecting excess, low priority loads. However, the load balancing granularity is at the level of the major round and all traffic priorities will receive equal delays. The delay will certainly be multiple latencies. The protocol can be modified so that each minor round allows transmission in all priority levels above the nominal minor round level, but this has little impact on the 99^{th} percentile of delay.

A new round can be commenced by a distributed protocol, with the triggering station rotating. This forms the basis of several hybrid slotted token networks. The token ring with interrupt [SHARP 87] remains the only implemented and widely published example, but this protocol is not optimum for large geometries.

4.3.5 Probabilistic Load Balancing.

Using probabilistic load balancing, stations transmit only when permitted to by a locally evaluated, heuristic function. The normal MAC rules also still apply. For example, stations may defer from sending when instructed by a local random number generator. The generator has a programmable bias which is recalculated at intervals from measured traffic statistics. Another example is the author's 1/E protocol for destination-release slotted rings presented in section 5.3. Probabilistic load balancing only guarantees fairness on average in the long term. An advantage of the queue and round models is that counting the number of active users is implicit. This is not so with probabilistic load balancing. An example for the slotted bus has been described by [MUKHERJEE 88], but such techniques can be used with all network architectures. A software implementation enables arbitrarily complex algorithms to be employed. Priority is easily incorporated, but is subject to the shortcomings described for quota based load balancing. In addition, the software may be unable to adapt to rapid changes in network loading.

In this dissertation, a new protocol is suggested by the author in section 5.4.

4.4 Physical Layer Packet Size

In this section, the term *packet* is used where others might use one of the terms *frame, slot, mini-packet* or *cell*.

A multi-access network by definition includes packet multiplexing functionality in order to control successive access by the various transmitting stations. The packet size used at the physical layer determines the granularity of sharing and hence the speed with which a priority message can be transmitted. The effect of packet size is experienced mainly in the lower rate sub-networks of a networking hierarchy. As the line-rate is reduced or the average packet size is increased, there is a point where the latency becomes less than the time for a packet from each user. Beyond this point, packet size will necessarily have an influence on the delay for priority traffic.

Properties of a short packet	
Advantages	Disadvantages
Effective expedited priority Small real-time chunk to lose Handy amount for locking Virtual cut-through	Relatively greater overheads

Table 4.1: Advantages and disadvantages of a short packet or mini-packet.

Fixed length, small packets are termed mini-packets or cells. Table 4.1 lists some further advantages of a short packet at the physical level. By short packet, a packet containing 32 or 64 bytes of data is implied. A short packet can contain a brief amount of speech, such as 1 or 2 milliseconds, and can therefore be occasionally lost without subjective degradation to perceived quality (chapter 1). A short packet can be transferred over a computer backplane using DMA in a single burst. Longer packets require several bursts in order to preserve fine grain sharing of the backplane. This is required to maintain the backplane's response to priority events. Related to this, in a multi-threaded or multi-processor environment, a short packet can be written or read from the interface hardware using atomic processor copies. This avoids providing slower, software implemented exclusion in the operating system (chapter 8). The last listed advantage of a short packet is described as virtual cut-through. For a message which has been fragmented into several physical layer packets, bridges which operate using store and forward of the physical packet appear as though they are forwarding the message without storage.

The disadvantage of a short packet is that routing and control overheads which occur once per packet, occur more frequently. If the packet is too small, routing and protocol overheads will reduce the network channel efficiency. In addition, the packet must be large enough to contain the necessary fields required to reassemble a fragmented message as well as a useful amount of data. A useful amount of data might be a single RPC with one or two, 32 bit integer arguments. D McAuley has built an RPC architecture where an RPC of the complexity just described, including all of the lower level protocol headers, fits into a (CFR sized) 32+4 byte mini-packet [MAC 89]. This is therefore proven as a viable size. This size, as well as slightly larger packets, are currently being considered for international slotted standards by IEEE and CCITT committees. For instance, DQDB has been adopted as a draft IEEE 802.6 proposal using a 48 byte mini-packet. There are a further four bytes of addressing header and one byte for access control [IEEE 89].

Computer data can experience low throughput as a result of a short packet if there is insufficient buffering at the transmitter or receiver to set up a suitable fragmentation and reassembly pipeline. Unfortunately, this is the case for several designs of CFR interface, but it could easily be avoided with a little extra buffering [MAC 89, PORTER 91]. The behaviour of a network which supports variable length packets at the MAC level tends to be determined by the longest type of packet that is used. This is especially true of the 99^{th} percentile of delay [NEWMAN 88].

For the reasons just mentioned, fixed length, small packets (mini-packets or cells) are becoming popular for multi-media communications [CCITT 89]. However, it should be noted, that as far as the performance of gigabit per second, multi-access networks is concerned, the packet length has very little effect on the delay performance and the effectiveness of priority. This is because delay need only be dominated by the speed of light limitation in the fibre. Whether a packet has to wait for 1000 bit times or 20000 to be transmitted will only make 19 microseconds difference to its delay, whereas the propagation delay for 50 kilometres is about 250 microseconds. The important property of the MAC is that the high priority packet can be transmitted as soon as possible after any lower priority packets which were in the process of being transmitted when the high-priority packet arrived. This relates to the granularity of load balancing.

4.5 Basic MAC Classifications

So far, the study of transmission control rules has not needed to specify the underlying access control method. This section reviews access control at large geometry.

Statistical multi-access MACs are based on one of three basic access control mechanisms: contention, slotted access and permission token. The fundamental difference between these mechanisms is the duration of bandwidth reservation implicit when a transmission is commenced. With (pure) contention there is no reservation, with slotted access, the channel is reserved for a universally agreed, normally fixed-length, slot duration, and with a permission token, the channel is reserved until released by passing the token.

Pre-allocated time division multiplexing may sometimes be considered as a multi-access MAC, especially if the allocation mechanism offers a quick response. Adaptive TDM is similar to token access with zero station-to-station token passing time. This looks promising for multi-media MANs and is discussed shortly.

Register insertion networks fall between the multi-access and switching classification. Since stations are permitted to interfere with the propagation of foreign transmissions, register insertion is not a true broadcast MAC. In [HOPPER 78], Hopper shows that register insertion rings have access delays which approximate the token ring. This remains true at large geometries, although the superior access delay of the register insertion ring at low loads becomes more pronounced.

For bursty traffic, register insertion offers zero access delay to a newly active station. This property extends to enable a small amount of expedited traffic to achieve zero access delay, provided register space has been reserved for this extra traffic. However, the inserted registers of other stations delay the propagation of the expedited traffic. This does not form a MAC layer priority mechanism according to the definition of section 4.2.

Destination removal of messages is also possible, giving greater than unity channel efficiency, but as usual, load balancing is then required. In most other respects, register insertion behaves in the same way as a token system, and it is not considered specially in this document.

4.5.1 Contention Access.

A pure contention MAC is characterised by the absence of any scheduling mechanism which ensures that transmissions do not overlap. When an overlap occurs, at least one of the transmissions fails, and over a certain class of media, all contenders are lost. Owing to the unreliable nature of the transmit process, for most applications it is vital that there be a reliable mechanism through which the transmitter can determine if its transmission were successful. This is then used for retransmission scheduling in the case of failure. For large geometries, applicable topologies include the looped and folded buses and the Hubnet Tree [LEE 83]. These topologies have the property that the transmitter can monitor the receive channel for the confirmation of success.

Contention suffers from a few problems at large geometries. These are now briefly noted.

4.5.2 Passive broadcast media unavailable:

For LAN applications, a major attraction of contention, apart from its simplicity, is the potential to use it with a passive medium; examples are Ethernet and its optical implementations. At large geometries, a suitable, nearly passive topology uses a star-coupled hub. Contention occurs only in the hub and can be detected by all participating stations. As mentioned in section 2.5, the quality of optical components, particularly amplifiers, needs to be advanced before this topology can stretch to MAN dimensions.

4.5.3 Out-of-order reception:

As stated in the introduction to this chapter, we are concerned with networks where there are potentially thousands of packets stored in the medium. Evidently, if there are tens of stations transmitting, then each will have tens of messages outstanding and retransmission will inevitably result in out-of-order reception.

4.5.4 Low utilisation:

Pure contention was used in one of the first multi-access 'MANs', the Aloha network of the University of Hawaii. The underlying radio technology can result in the loss of all messages in a contending group. As is well known, the maximum utilisation of this type of contention mechanism is 18 percent. Slotted Aloha achieves twice the throughput. A slot structure could be superimposed on to the Aloha network since the transmission rate was very low by fibre-optic MAN standards; the temporal distribution of the slot boundaries did not exceed the slot duration.

In order to achieve higher channel efficiencies, contention is virtually always used in conjunction with reservation to form a hybrid MAC. Reservation through carrier sensing, as in CSMA, provides an increase in throughput. For LAN dimensions, the contention period tends to be a few percent of the period reserved by the packet transmission and utilisations close to unity can be realised. However, as the dimensions are increased, the fraction of time spent in vulnerable contention increases, and unless the packet size is also increased, the effectiveness of the reservation decreases. It is undesirable to increase the packet size since this is contrary to fine-grain sharing and impracticable at large geometries. The result is that the utilisation falls back towards the 18 percent minimum.

True broadcast contention suffers further throughput reductions because of the high splice times which result from asynchronous operation. This was discussed in section 2.5. A proposal to improve the utilisation of contention and use it for an all-optical, multi-gigabit LAN has been made by [SAUER 89]. His approach is to gain greater apparent optical bandwidth for a contending electrical packet of given size and rate using controlled dispersion optical sections which shrink the duration of the packet and increase its optical bit-rate. An effective optical bit-rate of above 20 Gbit/second using this technique has been predicted. However, it is difficult to see how back-to-back receptions can be handled, and again, the project depends heavily on future optronic developments.

4.5.5 Unconstrained delays:

Contention MACs where the retransmission algorithm relies on a random generating function can potentially suffer from infinite packet delays. De-

terministic contention resolution algorithms, such as the tree algorithm, are able to prevent such pathological cases and also raise the throughput to, for example 43 percent [CAPETANAKIS 79]. However, such algorithms again rely on setting up a globally distributed time frame.

4.5.6 Contention Access for Active Media.

Active media can offer protection of one of the contending messages, for instance, by giving priority either to the message emerging from a station or to the message already on the network. In effect, the network medium contains the reservation mechanism. Each message is only valid if followed by an intact terminator. Since a message can be pre-empted at any point, the receive section of a station must be able to abandon a partially reassembled message and synchronise to the start of another very quickly.

Acceptable throughput can be achieved if messages are prioritised such that the winning contender is defined. Otherwise, the headers of unfinished messages which have been themselves pre-empted, tend to pre-empt further messages with the result that none are successful. If the priorities are static, say based on source address as in the ReC-ring [OKADA 87], or being closest to the head end of a bus, then the resulting network is unfair. In order to provide fairness, the priorities must rotate, and hence we have a hybrid scheme, relying on contention at low loads and degenerating to a token-like system at higher loads. Another example is X-Net [KAMAL 89]. This uses a hybrid of contention and token access control over the dual bus topology.

The Hubnet network [LEE 83] uses an active media to protect one message at each point of contention and a tree structure to define which message wins. Fairness is ensured since the winner is simply defined as the first message to arrive at a node in the tree. The initial implementation did not include a delay bounding mechanism and operated with a window size of one. This window size would result in fairly low point-to-point throughput at large geometries. A larger window size would increase point-to-point throughput but would result in out-of-order message reception. For file access and file transfer traffic, which is typically the traffic which suffers from point-to-point bandwidth starvation, it is possible to use longer messages, but of course, these increase the delay for real-time traffic.

4.5.7 Contention Summary.

Various forms of contention based media access protocols have just been reviewed and a number of problems have been identified. Contention is not attractive for large geometry networks which must support real-time traffic.

4.5.8 Token Access Control.

Under the permission token scheme, at any time only one station is able to transmit to the medium. This station is said to be holding the token. At large geometries, for efficiency it is necessary that a station passes the token to another station, immediately after transmission, without waiting an additional latency. This is unlike the IEEE 802 series token protocols. It is a so-called early release, or multiple token protocol. If the passing operation operates no faster than the data transmission, which is the case if the network medium is used for passing the token, a station can be sure that the network is idle when it gains the token, and that it will not be pre-empted by another station.

Without MAC priority, the token is passed between stations following a logical ring. In this way, each station receives round-robin service and fairness is ensured. MAC layer priority in a token system is implemented by passing the token at a station, even though there remains low priority traffic to be transmitted. The token then follows a smaller ring composed of the subset of stations which are backlogged with priority transmissions.

Since no transmission is in progress while the token is being passed, the channel efficiency of a token system is essentially one minus the fraction of time that the token spends in transit. Physically passing the token, either through a unique bit pattern, or by a line idle period (as in Expressnet) results in one rotational latency of the logical ring of wasted bandwidth per token rotation. For 90 percent utilisation therefore, the average rotation time will be ten times the latency. Although the token rotation time may be quite low for small LAN dimensions, a 200 kilometre token path for instance, would have a rotation time of 10 ms at 90 percent utilisation. See section 5.2.2.

A rotation time is the interval over which load balancing is amortised. If this is 10 ms then expedited transfer within stations will not serve as an ef-

fective priority mechanism. All traffic classes will experience approximately the same access delay. This is dominated by the mean residual life of the token rotation and an average access delay of 5 ms cannot be considered fine-grain sharing of a one gigabit network. (See table 9.2.)

The token rotational time can be reduced at the expense of utilisation by using the non-exhaustive service MAC priorities. The FDDI timed-token protocol offers a practical example of this technique. This provides an effective load balancing mechanism, being able to reject excess offered loads while maintaining real-time guarantees. However, the loss of capacity starts to become unacceptable once the latency starts to approach the target rotation times that are required for real-time traffic support. The timed-token protocol includes a multi-level priority mechanism, although this has been found to be not very effective [SCHILL 87].

4.5.9 Adaptive TDM Access Control.

Adaptive TDM protocols, similar to those used on satellites, are suitable for MANs using self-stripping media. As mentioned, adaptive TDM may be likened to token access, but with very low passing overhead. Reservations for a TDM packet in a round are generally made the revolution before, and at large geometries, they may have to be made several rounds before. This increases the load balancing delay by several latencies, but not always the delay for the data. An adaptive TDM protocol for a passive star-coupled MAN has been proposed by [FIORETTI 87 and 88b]. The channel efficiency is not dependent on the length of connecting fibre to the hub, since this is measured by the stations. The stations then time their transmissions not to interfere at the hub.

As stated in chapter 2, in order to to support 50 or more users over a 50 kilometre area it is necessary to distribute the hub using multiple star couplers and optical amplifiers. A distributed hub reduces the channel efficiency in reverse proportion to the separation of the hubs since the time for light to travel between the hubs increases the passing time of the virtual token and therefore the effective network latency. For real-time applications, the cycle time should be kept within the same limits as are applicable for a real-time token ring. Multiple couplers within the same building would result in acceptable performance, but it is probably better to consider bridging between separate networks in order to span larger areas. Of course, long fibres to

each station from the hub can always be used, provided one can bear the wiring cost, optical loss and timing complexity.

Given that the problems of long cables and optical budget in the star could be solved in the near future using optical amplifiers, the real-time performance of adaptive TDM at MAN geometries could be very good. The delay performance is essentially the performance of a very small token passing system.

A quantitative estimate of expected performance is illustrative. Imagine a network with 50 active stations connected to a central star-coupler and a line-rate of 1 GHz. If each station uses a 1 kilobit packet and the splice time is 200 bits, then the channel efficiency is 83 percent and the cycle time is 60 microseconds. A 60 microsecond cycle is attractive since the system offers finer grain load balancing than a 50 kilometre diameter slotted ring. As with token rings, using shorter messages enables channel efficiency to be traded against cycle time.

Using larger packets, or sending several mini-packets at once would increase the cycle time and also the channel efficiency. For 95 percent efficiency, using these figures, the packet size needs to be 4000 bits and the cycle time becomes 210 microseconds. Clearly, these numbers are acceptable for our current conception of multi-media traffic (section1.3).

The CRMA access protocol [AS 91] is a good implementation of adaptive TDM for a LAN ring. Although the underlying ring uses the empty slot access protocol (next section), a reservation mechanism, implemented at an elected control station, ensures that stations are allocated clusters of consecutive slots for their transmissions. This avoids segmentation and reassembly procedures, which are commonly required for slotted systems. The reservations are made in the prior ring revolution by a station adding the number of slots it would like to use to a count field which rotates as part of the ring framing structure. Full multi-level MAC layer priority can be achieved with a granularity of two latencies. Slots may be used non-contiguously if reqired, as would be when carrying ATM cell protocols.

4.5.10 Slotted Access Control.

Slotted access divides the bandwidth into units of well known size. These are termed slots. Access to each slot is arbitrated independently, normally

through a full flag at the slot start. An atomic, test-and-set operation on the flag is used to reserve the channel for one slot duration.

A slotted system does not suffer a throughput penalty as a result of lost time while a token is in transit, and can therefore achieve 100 percent channel efficiency.[4] A disadvantage of a slotted system is that packets must be fixed at the slot size, or fragmented into several slots. The points made in section 4.4 relating to physical layer packet size apply to the selection of slot size.

The basic throughput of a slotted system is virtually independent of the ratio of packet or slot size to network latency. This is desirable for a backbone network where one architecture must cover orders of magnitude variation in latency.

Slotted access cannot sensibly be applied to wireless media or the broadcast bus topology at large geometry since the temporal distribution of message boundaries over the broadcast medium prohibits the atomic test-and-set operation required to reserve a slot. It is however, amenable to the remaining bus and ring topologies. With the self-stripping bus topologies, the empty slots are generated by a head-end station and the full slots usually fall off the tail end. The ring topologies are more complex, in that slots must be marked both full and empty and a monitor mechanism is required to delete permanently rotating full slots.

There is a wide spectrum of protocols for slotted networks. The most important distinction is whether the destination releases the full slot. Destination release requires that a station must recognise its MAC address before passing on the full indication. This was seen as requiring unacceptable station delay at LAN geometries, but this argument does not apply at larger geometries where cable delay dominates station delay. Networks using destination-release protocols are strictly not broadcast networks, since part of the channel can be re-used, but they possess the valuable ability to achieve network throughput in excess of the channel bandwidth.

Some slotted protocols, such as the CFR [HOPPER 88] protocol and the initial version of DQDB [NEWMANN 88], limit each station to one transmission per network latency but this must be relaxed for larger geometries

[4]The maximum throughput is reduced by transmission control rules such as 'pass-on-free'. These are described in the next chapter. Of course this is true for all MACs. The timed-token protocol of FDDI is an example for a token system.

in order to obtain acceptable point-to-point bandwidth independently of the number of slots.

Slotted access implies fine grain sharing of the medium bandwidth. This results in lower delays than token access for short messages. The low delay capability and the fact that the performance does not degrade as the network geometry is increased make slotted access the prime contender as the media access technique for backbone networks. The performance of several types of slotted ring at large geometry is considered in the next chapter.

4.6 Partitioned Multi-Access Networks

Sharing the bandwidth of the medium is natural to a multi-access network. With current technology, electronic time-division multiplexing (TDM) is probably the most suitable partitioning technique. However, the discussions in this section are also applicable to the code-division, wavelength division and frequency division multiplexing techniques provided the delays in each partition are fairly well matched.

An immediate disadvantage of partitioning a service, such as a network station, is that the queuing delay can increase linearly with the number of channels. On average, each partition will have the same queue length as an equivalent non-partitioned server, but the service rate has been divided by the number of channels, and hence from Little's result, the service time increases by the same factor. This is an inevitable penalty in any multi-channel system with a fixed assignment of customers to servers, but it is ameliorated if a customer has a choice of servers. This becomes apparent when the receiving part of a station is connected to more than one channel or if it swaps channels dynamically to make use of idle servers. Swapping according to a pre-defined pattern, as with the predictable assignment shortly described, is not sufficient. This is demonstrated in table 9.2.

Even though certain media are not intrinsically multi-channel in the way a WDM fibre system is, for instance, there are other possible reasons for partitioning the network bandwidth into several channels. These are now listed along with references to related discussion.

- Low complexity stations and stations of various complexity can be constructed. The bandwidth required by the station becomes better

matched to the bandwidth of the shared medium. See [FIORETTI 89b] and chapter 6 for the Cambridge Backbone Ring description.

- Certain standard access protocols do not operate very well at large geometries. Partitioning the network reduces the apparent number of bits stored in the medium. See section 4.7 and [CHLAMTAC 88].
- It is desired to use an existing MAC controller circuit at a higher effective channel rate. To do this, several MAC controllers are used on parallel channels. See [YANG 83], [LUVISON 89] and later in this section.
- Different MAC protocols can be operated in different partitions. See section 4.8

Partitioning the medium into TDM channels enables stations of varying complexity to be constructed. The simplest will only be able to receive from or transmit to one of the channels at a time. Others may be equipped with sufficient hardware to use several channels simultaneously, and the most general station can, if necessary, have access to the full network bandwidth. A partitioned architecture can therefore offer the network manager a selection of station designs. Nodes with low bandwidth requirements can be fitted with the simplest type of station and the more complex stations need only be fitted where the cost is justified.

4.6.1 Channel Assignment and Reservation Algorithms.

The absence of destination contention is intrinsic to a single channel, multi-access MAC. Given that we wish to preserve contention avoidance in multi-channel networks, if some stations are not provided with enough hardware to receive all of the network traffic, then action must be taken to ensure that transmissions are synchronised such that the receivers are not overloaded. This section considers the minimum station delay that is required for this.

In the most general case, the procedure to avoid destination contention is as follows. Before a transmission is commenced, arithmetic must be performed to establish the set difference between the set of backlogged destinations and the set of destinations that will be receiving at the time they encounter the proposed transmission. Multicast transmissions must be accounted for in this calculation. From the resultant set, an arbitration

scheme must be invoked to clip the possible transmissions to the number of simultaneous transmissions supported by the hardware, the appropriate data buffer(s) identified, and transmission commenced. The transmissions selected will necessarily influence the set calculation at the downstream station since the destinations they refer to will have to be taken into account. This is a result of a universal rule which limits the earliest time that the channel reservation and destination information can leave a station:

4.6.2 Reservation Interval Rule:

> The station delay in order to avoid receiver contention must be greater than a reservation interval equal to the time for the destination information on all potentially contending channels to be adsorbed into the station.

For the minimal reservation delay to be realised, the destination information must be processed in parallel, which in practice means using a MAC frame where the channels are closely interleaved and looking up the destination fields of each channel in multiple, parallel bit maps. This is evidently hardware intensive and contrary to the aim of sub-equipping stations. Alternatively, the minimal delay need not be realised, and a MAC frame where the channel headers are staggered to enable serial reuse of the look-up hardware can be envisaged. The reservation rule still applies; the station delay must be expanded to contain all of the staggered headers that may arrive during the proposed transmission. Using VLSI shift registers, this is possibly less costly to implement than the former method, but it remains hardware intensive.

Complexity is reduced by reducing the reservation interval. One approach is to constrain the destination field encoding such that only a small amount of the header needs to be examined to determine whether a receiver will be active. This approach can be extended to the limiting case where no destination bits need to be examined and each receiver is assigned to a predetermined set of channels. This assignment can vary with time in a prescribed way, as described shortly under the predictable receiver assignment heading, or it may be static. Static assignment results in the lowest hardware cost, but can lead to unnecessarily overloading one channel while unused bandwidth remains on others.

4.6.3 Channel Dynamics.

	Receiver static.	Receiver dynamic.	Predictable receiver assignment.
Transmitter static.	Impossible routes. Bridging stations required.	Receiver contention and capture problems.	Good sharing.
Transmitter dynamic.	Receiver allocation scheme required. Poor sharing possible.	Transmitters must keep up with receiver location. Balance dynamically optimised.	Good sharing, lower delays.

Table 4.2: Possibilities for Dynamic Receivers and Transmitters

In a multi-channel system, receivers are either staticly assigned to a fixed channel - their 'home' channel - or else they dynamically hop from one to another. As shown in table 4.2, if the receivers are static, then in order to achieve full connectivity either the transmitters must be dynamic or else translation between channels is required.

4.6.4 Static Receiver Assignment:

For the static receiver case, a special transmission scheduling algorithm is not required: transmission consists of looking up the published home channel for the desired receiver and then writing to that channel according to the basic access control rules. This scheme of permanently assigned receive channels has the benefit of simplicity. The only requirement being the out-of-band distribution of the channel assignment information over the local network. This need not be a great overhead: if absolute addressing is used, then the home channel values can be cached within each station, and in a virtual circuit system, the control structures for such information distribution will already exist. The transmit channel number is simply a further attribute to a virtual circuit. The main disadvantage of static receiver assignment is that one channel can be overloaded while free bandwidth remains on another.

4.6.5 Dynamic Receiver Assignment:

Dynamic receiver assignment does not rely on statistical balancing of the traffic within each channel, and therefore potentially offers lower delays.

However, such systems can easily exhibit a receiver 'capture' effect whereby it is impossible to send to a particular destination as a result of upstream stations having filled every potential transmit window. This is undesirable since it is a form of destination contention and it increases the delay jitter experienced by the traffic.

From the discussion about channel reservation intervals, it is clear that the complexity of avoiding receiver contention with dynamic receiver assignment is very great. Dynamic receiver assignment can cope with greater amounts of non-homogeneous traffic than static assignment, but under such conditions, is liable to give a hyper-exponential service time, owing to the capture effect. Therefore, the support of truly dynamic receivers is not attractive, especially when the method of predictable assignment is available.

4.6.6 Predictable Assignment:

Under predictable assignment, each receiver is assigned to a logical channel instead of directly to a physical channel and the logical channels are mapped to the physical channels using a function which changes each packet time. The number of logical channels should be larger than the number of physical channels so that the logical-physical mapping is many to one. An important property of the mapping is that two logical channels which map to the same physical channel at one moment, map to distinct physical channels shortly afterwards. In this way the applied traffic is spread very evenly over the physical channels, without much regard to the relative loading of the different logical channels.

The logical-physical mapping must be predictable so that transmitting stations know which physical channel to use in order to reach a given destination. This requires a distributed time frame of the type available in a slotted network where stations can count slots from an index mark. A pseudo-random hash function of logical channel number and the current time or frame number is sufficient. The receiver uses this function in order to select its receive channel and the transmitter uses the same function to select the appropriate transmit channel. The function can be held in a small read-only memory.

Simulation results comparing static and predictable receiver assignment are presented in section 9.3.

4.6.7 Full-Duplex Access.

Within a single channel, multi-access network, that is, a design which does not partition the bandwidth into sub-channels, all stations are intrinsicly half-duplex since there is only one station transmitting at any moment. However, for a design with more than one channel, it is possible to have a full-duplex station, since a single station can be transmitting into one (or more) of the N channels while receiving from one (or more) of the other channels, and it is possible to have a half-duplex station, which cannot transmit while it is receiving. Such half-duplex stations might be expected to suffer additional forms of receiver contention, missing receptions while they are transmitting. However, the reservation interval rule remains sufficient to avoid receiver contention since the station needs to examine the destination fields that might cause it to receive in advance of modifying them with a transmission.

4.6.8 Multi-Channel Access.

The lowest bandwidth type of station for a partitioned MAC consists of a single, half-duplex access unit. This can receive or transmit from only one channel at a time.

There are two possibilities for the next least complicated station. It could either consist of two half-duplex access units, or two simplex units, one for receive only and one for transmit only. The dual simplex station is simpler to implement in hardware owing to the clearly partitioned roles of the two halves. It will also have higher throughput than the dual half-duplex station in the case where the half-duplex sections cannot transmit immediately after transmitting. This is the situation for the stations described in chapter 6. Otherwise, both types of stations have the same saturated throughput. The greater flexibility of the dual half-duplex station will result in lower delays owing to the increased number of simultaneous receive and transmit opportunities. In practice, the choice is liable to be determined by hardware issues.

Since the individual channels of the partitioned network carry an inherently half-duplex multi-access protocol, a station fully equipped for concurrent access to the full set of channels would require half as many half-duplex protocol units than it would simplex protocol units. Therefore the half-

duplex units are preferable for large stations.

4.7 Multi-Channel Behaviour of MACs

Although the foregoing discussion pertains to stations connected to any type of partitioned network, there is only a limited class of multi-access MAC protocols to which partitioning with dynamic receivers is particularly amenable.

4.7.1 Variable Packet Length Systems:

With variable length transmissions, there can be no correlation between the positions of message headers on the various channels. This raises certain difficulties. In order to avoid contention for dynamic receivers, the reservation interval rule implies that the station delay must exceed the duration of the proposed transmission. Otherwise it is possible that an unexpected contending transmission will appear on another channel during the proposed transmission. Therefore a station on a partitioned network with variable length packets and dynamic receivers must include a delay exceeding the longest packet length. Since variable length packets are generally supported in order that quite long packets can be used occasionally, the required delay is quite large and the whole scheme is not very sensible. Static receiver assignment is much preferable in systems supporting variable length packets.

4.7.2 Multi-Channel Token Systems:

A fully equipped, partitioned token system was studied by [YANG 83]. This study considered using in parallel, two or four token ring stations of a standard design. This obviously provides twice or four times as much bandwidth when compared with a single token ring. However, in Yang's simulations, the tokens were observed to cluster as is the wont of multiple cyclic servers, with the result that the average access delay was not reduced as much as expected.

A multi-channel attempt-and-defer system has been proposed by [CHLAMTAC 88]. Attempt-and-defer systems, such as Expressnet and Fasnet, are

	Channel Structure	Access Control
1	Common	Access controlled by 'MAC'
2	Common	Allocated out-of-band
3	Separate	Arbitrary protocol

Table 4.3: Summary of isochronous support methods.

a class of token passing system. In Chlamtac's architecture, the bandwidth partitioning is proposed in order to increase the efficiency of token passing at large geometries and also to reduce the station complexity. In a token system of fixed overall bandwidth, if the number of channels is increased while the packet size is kept constant, the overall network efficiency is increased. This is because the packet size has increased with respect to the number of bits stored on the individual token rings.

Variable length packets can be supported by a multi-channel token system, and for the reasons just mentioned, this in practice implies static receivers. In order to send a priority message to a particular receiver, token possession on that channel must be gained. Partitioning a network into channels does not reduce the latency, and, as to be described in section 5.2.2, this means that the average access time for the token system remains unchanged. As demonstrated in section 4.5.8 (page 85), this can be quite large at large geometries.

4.8 Hybrid and Isochronous Protocols

Worries about the variable delays intrinsic to statistical multi-access networks have resulted in several hybrid protocols being proposed. Some examples are described in this section. Hybrid protocols divide the available bandwidth into synchronous and asynchronous fractions, the synchronous bandwidth becoming available every 125 microseconds under control of a master 8 kHz clock. This rate is chosen to support 64 kilobit speech and it is sometimes proposed to lock the 8 kHz clock to the international telephone network. Traffic which is granted regular access regardless of the overall loading is termed *isochronous* traffic. There are three approaches to supporting isochronous traffic and these are summarised in table 4.3.

The first and second approaches combine the isochronous traffic with the

non-isochronous traffic and integrate both types on to the statistical, multi-access channel. In the first approach, the isochronous traffic is transmitted according to the normal MAC rules. These must include a major-minor round system of the type described in section 4.3 and the major round must be initiated by a master 8 kHz clock. The LION network [LUVISON 89] is an example of this type of network. The advantages of this first approach are that isochronous traffic can be received and transmitted using standard stations and the movable boundary between the isochronous and non-isochronous traffic prevents waste. Using the full MAC protocol for isochronous traffic can also be seen as a disadvantage since the hardware complexity for a simple voice connection can be quite large and bandwidth can only be shared at the granularity of the physical layer packet.

The second approach overcomes these specific problems. The standard MAC is not used for isochronous access control. Instead, channels consisting of one byte every 12.5 microseconds are allocated by a layer of management software. Access is regulated by software protocols, running out-of-band over the asynchronous part of the network. Slotted networks are readily adapted by marking certain slots full at a control station. This prevents them being used for the transmission of asynchronous traffic, and their header field is marked such that it does not forge a valid asynchronous reception address. Access to the isochronous bandwidth is only through special stations, but these can have a very low, per-channel complexity. An example is DQDB, where the head-end can periodically mark a collection of slots for synchronous traffic.

In the third approach, isochronous traffic is carried using a completely separate channel, although this is typically envisaged as a TDM partition of the same physical channel. This approach has been proposed for the MST (multi-slotted plus token) proposal [MOLLENAUER 86] and for its successor, the FDDI-II ring [BOSTON 88]. In FDDI-II, the 100 Mbit/second physical channel bandwidth is partitioned into 16 logical channels. Any number of these logical channels can be dedicated for isochronous traffic use. The remaining bandwidth operates the standard FDDI token protocol. The logical channel structure of the bandwidth allocated to the token protocol has no consequence.

In order to support isochronous channels using the second and third approaches, the ring systems require special latency adjustment so that the isochronous channel structure rotates at a multiple of 125 microseconds.

This involves a station which extends the electrical delay of the ring. A 200 kilometre ring will achieve a latency of about 100 microseconds, therefore the typical multiple of 125 will usually be one. In the second approach, all traffic will encounter this increased latency. This is undesirable for smaller installations using slotted access, since it sacrifices the intrinsic low delay. Using the third approach, only the isochronous part of the ring delay need be extended.

Hybrid and isochronous MACs are guaranteed to give acceptable voice performance. This is, after all, their main motivation. However, they are inflexible, being tied to a fixed clock, and they have no capability for bursty, high-priority traffic.

4.9 Summary

This chapter has listed the desirable features of a multi-access MAC and then gone on to look in detail at how fairness and priority can be supported. Particular emphasis was put on the response time of MAC layer priority mechanisms at large geometries where simple expedited transfer is no longer sufficient. In the next chapter, these concepts are demonstrated for a selection of ring protocols.

The basic elements of contention, token and slotted access control were covered in section 4.5 and the way these protocols degrade at large geometries was described. The adaptive TDM form of token passing where the cycle is kept short was suggested as a suitable MAC, although it is severely limited in geometry by the optical technology immediately available. Of the MACs described, those using slotted access are unique in not suffering a throughput degradation as the geometry is increased.

The possibilities for channel assignment in a multi-channel architecture were discussed in section 4.6 and methods of avoiding receiver contention were presented. The next section found that support of variable sized packets is not very suitable for partitioned networks and that token controlled access, when partitioned, only gains in efficiency and not in response to priority traffic. Slotted access was shown to be very suitable for multi-channel architectures where its distributed time frame can form the basis for predictable receiver assignment policy.

The last section mentioned some hybrid circuit and packet switched projects. These are unable to support bursty, high-priority traffic.

Chapter 5

Ring MAC Protocols and their Performance

This chapter considers MAC layer protocols for slotted rings, token rings and other non-register insertion rings and how to apply load balancing to them. The chapter describes several slotted ring protocols and then presents simulation and analytical results for the important ones. The relative benefits of token and source and destination release slotted rings are compared in terms of delay, throughput and fairness. In the last section of the chapter, a new ring protocol is presented which incorporates the combined benefits of source and destination-release.

5.1 Protocols for Slotted Rings

A source-release slotted ring which is not restricted in the number of outstanding transmissions per station and which passes used slots free to the downstream station is using a protocol known as the MSR protocol (multiple slotted ring). This is also known as CFRV (Cambridge Fast Ring Variant) by [ZAFIROVIC 88a]. Transmitting in any empty slot which passes avoids certain meta-stable bandwidth inequalities identified in [FALCONER 85a], although the remain many distinct saturated slot usage patterns. These each have associated worst-case service times.

The most extreme slot occupation pattern can cause the delay perfor-

mance of a saturated slotted ring to degenerate to that of a token ring. The effect can be termed 'resonance' owing to the following effect: Since the MSR protocol is source-release with pass-on-free, stations are unable to use slots that they used in the revolution before. In the case of a single, saturated, active source station, this station would tend to transmit solidly for one ring revolution, then be blocked completely during the next revolution while it frees the slots used before. This would continue, alternating on each ring rotation. Resonance like this is likely to be experienced in practice when a ring is being heavily used by only a few stations.

The CFR basic protocol is pass-on-free, but it also offers so-called 'channel mode' slots which can be refilled by the source instead of passing them free to the next station. The principal aim of this is to increase the maximum point-to-point bandwidth with LAN dimensions. However, for backbone network purposes, higher line-rates and higher sustained utilisations are envisaged so greater sharing is desirable. Rings which use more than one slot per revolution already benefit from a greater point-to-point bandwidth. Channel mode leads to increased complexity which is not required for these larger networks.

5.1.1 Benefits of Source-Release.

For MAN dimensions, the two benefits of source-release are built-in load balancing and the ability to carry a response field. The primary disadvantage is wasted bandwidth on the return path. The additional bandwidth of destination-release is attractive; consequently, methods of realising the two benefits of source-release, but using destination-release of the slot data fields should be considered. The next section presents a mechanism for providing responses on a destination-freeing ring.

5.1.2 Response Mechanism for Destination-Release.

Multiple uses of a slot per latency, while still providing a low-level acknowledgement, are possible if each slot carries multiple response fields. Access to the response fields must be arbitrated and they effectively become mini-slots associated with the major slot. These are source-released, although the main data field can be destination-released. A suitable protocol is now described. The response slot must be able to hold at least the values 'empty', 'active'

and two response values.

1. When the transmitter writes to the main data field, it also updates a previously empty response mini-slot in the same major slot, marking it with a flag so that it can be distinguished as the active response slot.

2. When the receiver has copied the data, it frees the main data field and writes the response to the active response mini-slot, making it no longer active.

3. Back at the transmitter, the response is copied and the response mini-slot returned to the empty state.

If all response mini-slots of a major slot are in use, then stations should be barred from using the slot, even if the main data field is free. About three sets of response mini-slots would be sufficient to make the probability of this acceptably low. Since only a very few response values are required, the contents of the response mini-slot can be encoded with its control flags in order to reduce the total bit count. The response slots do not require addressing fields, because the correspondence between the data field and the current active response field can be flagged to the destination. The source must remember which response slot it used.[1] As usual, all types of slot require 'garbage collecting' at the monitor to ensure that they always return to the empty state after an error (section 6.5). The marginal complexity of this system is hardly a consideration for VLSI implementation; the important point is that the provision of an acknowledgement mechanism need not require source-release of the main data field.

5.1.3 Load Balancing Through Source-Release.

The remaining benefit of source-release over destination-release is the guaranteed fairness through the associated pass-on-free rule. Fariness may be amortised over all active virtual circuits rather than active stations if each station re-uses a slot a number of times up to the number of active virtual circuits at that station.

[1] State recorded on a per slot basis has been termed 'profile' information and needs be kept by any source-release system.

Load balancing by source-release does not ensure a particularly low bound on the maximum access delay.[2] If there are N active users, then the maximum possible period waiting for an empty slot is only bounded to $N + 1$ ring latencies, a figure which is comparable to that for a token system, even with modest N, say 10 to 20 stations.[3] However, as shown in section 5.2, the average delay at saturation for an expedited cell, that is one inserted at the front of a station's queue, will be $N + 1$ slots, which is liable to be considerably less than one latency, and therefore at least an order of magnitude less than for the token system at HSLAN geometries. Below saturation, the performance is, of course, further improved.

The 99^{th} percentile of delay depends on the slot usage pattern at the point when saturation was reached. Network saturation is readily achieved in simulations using saturated sources, but after saturation, the slot usage pattern in a given simulation remains fixed. Extending the length of a simulation run will then not increase the ensemble accuracy of the results owing to the non-ergodic nature of the process. Despite this, and taking into account the usual difficulties of obtaining accurate 99^{th} percentiles, the results of the author's simulations have indicated that the 99^{th} percentile of the delay for an expedited cell, while the network is in saturation, normally lies in the upper quartile of one latency for 100 slots and 18 stations. This is a very acceptable figure. However, as is shown in the tables later on, it is possible to increase the expedited delay by applying several very bursty sources which are able to generate long stretches of consecutive full slots. The 99^{th} percentile is then increased by a factor of three or four, meaning that occasionally a station will not see an empty slot for several ring revolutions. This effect may be reduced by adding some random 'dither' or through a similar effect which prevents excessive stretches of consecutive slot filling. This disrupts the resonance and improves the expedited traffic delay.

[2] In fact, as shown in [PORTER 91], even when the number of stations is significantly greater than the number of slots, as is the case at LAN geometries, slotted-rings where each station is constrained to only use one slot per ring revolution can also fall into slot occupation patterns with delay performance not much better than a token ring.

[3] For 'perfect' load balancing, the delay would be the larger of N slot times or one rotational latency.

5.2 Numerical Study of Slotted and Token MACs

In this section, simple models for a group of token and slotted MACs are presented. These serve as a guide to the performance of the MACs at large geometry. Selected simulation results are also presented in order to assess the accuracy of the models and to provide 99^{th} percentile estimates. The overhead of preamble and control fields is not included in the study in order to reduce the information carried and because the necessary channel efficiency correction may easily be applied at any stage.

Apart from general throughput and channel efficiency results, the mathematically tractable performance measures of interest may be identified:

- Average delay under homogeneous Poisson, bulk arrival sources (for ready comparison with other work).
- Mean and 99^{th} percentile delay for a short, top priority message, under worst case conditions.
- Delay metrics for the delay sensitive class of traffic when all the remaining network bandwidth is utilised by saturated sources.

5.2.1 Modelling the MSR and other Slotted Protocols.

There is no known analytical solution to the expected delay of a message on slotted networks, even for a homogeneous Poisson arrival pattern.[4] As is common in multiple cyclic server systems, the full slots, which are the busy servers, tend to coalesce into clusters of adjacent full slots [MORRIS 84]. This is because new arrivals, which are queued when the passing slot is full, are appended to the current cluster, tending to increase its size. The situation is similar for both ring and bus networks, although in the destination-release systems, the freeing action again breaks up the clusters slightly.

Many models of slotted systems require the simplifying assumption that adjacent slots are independent. This gives a geometric distribution to the

[4] Exact results for several classes of token system have been presented in [AMINETZAH 85]

	Pass on free.	Use immediately.
Destination-release send to self.	$\frac{2N}{N+3}$	$\frac{2N}{N+1}$
Destination-release never send to self.	$\frac{2N(N-1)}{N(N+1)-2}$	2.0
Source-release.	$\frac{N}{N+1}$	1.0

Table 5.1: Slotted ring saturation ratio (β).

expected wait for an empty slot. See for instance [YANG 86]. The geometric model underestimates the delay, particularly towards half utilisation under Poisson loading and with bulk arrival loading generally. However, the simplicity and stability of the model makes it worthwhile and this approach is developed here. Other models are considered in section 5.2.3.

For a slotted ring, the relationship between average full slot density and applied traffic depends on the expected distance that a full slot will travel over the ring. This in turn depends on the number of stations and, for destination-release, their traffic patterns and whether stations send to themselves over the network. It also depends on whether they can re-use a slot they have freed as in CFR channel mode, or whether they must pass it on empty. If a station passes on empty, the average number of empty slots seen by its transmit side is the same as the density of empty slots on the ring cables. Assuming homogeneous traffic, the transmit side of a station which frees its own full slots before attempting a new transmission will see $1 - 1/N$ times fewer full slots, resulting in an overall higher throughput, N being the number of active stations.

The system bandwidth of a multi-access network is defined as channel bandwidth after it has been formatted with access control and addressing overheads [DANTHINE 85]. The achievable saturated bandwidth normalised to the system bandwidth can be termed the saturation ratio, denoted by β. Equations for β under homogeneous traffic flow are listed in table 5.1. As the number of active stations is increased, the saturation ratio approaches unity for source-release and 2.0 for destination-release.

Approximate MSR protocol study: This is an approximate analysis of the multiple slotted ring (MSR) protocol with infinite buffer capacity at each station. All possible release mechanisms are characterised by β. As stated, the clustering tendencies of the full slots are not modelled. Bulk

Poisson arrivals of cells is the canonical arrival model. Low-rate synchronous sources can be approximately modelled within this form by assuming a bulk size mean of unity. Low-rate expedited sources are modelled by omitting the queue delay term from the resulting expressions.

Let the average bulk size be $\bar{k}$ and the mean arrival of bulks be λ per second per station. We have a fixed length cells of duration σ seconds and the network latency is T seconds. There are T/σ slots without a gap.

From Little, if the average wait in the station is $\overline{W}$, then the average number of cells in a station queue is

$$\overline{Q} = \lambda \overline{k} \overline{W} \tag{5.1}$$

The ring 'utilisation' may be defined as

$$\rho \stackrel{\Delta}{=} N \lambda \overline{k} \sigma \tag{5.2}$$

but it must be noted that the maximum value ρ can assume is β, which exceeds unity for destination-release.

Let the average density of full slots encountered by the transmit side of a station be υ. This is proportional to the ring utilisation so we let $\upsilon = \alpha\rho$ where α is a constant of proportionality determined by the saturation ratio of the protocol β. At saturation, ρ is equal to β and a station is transmitting in every empty slot it sees, so

$$1 - \upsilon = \overline{k} \lambda \sigma \tag{5.3}$$

$$1 - \alpha\beta = \beta/N \tag{5.4}$$

$$\beta = \frac{N}{1 + N\alpha} \tag{5.5}$$

$$\alpha = \frac{1}{\beta} - \frac{1}{N} \tag{5.6}$$

Assuming no clustering, the expected number of slots before an empty one follows a geometric distribution with parameter υ. The expected time before an empty slot is therefore

$$\overline{d} = \frac{\upsilon}{1 - \upsilon} \sigma \tag{5.7}$$

The average time in the queue experienced by the last cells of a bulk consists of four terms:

- The first term is equal to half a slot time and results from moving from the continuous Poisson process to the discrete time ring: $\sigma/2$.
- Poisson arrivals see time averages, so the second term is the time for all of the $\overline{Q}$ cells of previous packets, queued ahead, to be sent. This term is multiplied by the indicator E which although normally unity, can be zeroed for an estimate of the delay suffered by a lone, expedited cell.
- The third term is the time for all of the cells of the current message, except for the last one, to be transmitted.
- And the fourth term is the time spent by the cell of interest waiting for its empty slot.

We can write this down

$$\overline{W} = \sigma/2 + E\overline{Q}(\overline{d}+\sigma) + (\overline{k}-1)(\overline{d}+\sigma) + \overline{d} \quad (5.8)$$

which can be simplified

$$= \frac{-1/2 + \upsilon/2 + \overline{k}}{1 - \upsilon - E\lambda\overline{k}\sigma/N}\sigma \quad (5.9)$$

$$= \frac{-1/2 + \rho\alpha/2 + \overline{k}}{1 - \rho\alpha - E\rho/N}\sigma \quad (5.10)$$

The total message delay includes a further slot time and the time to propagate along the ring to the destination station. Notice that this result does not directly depend on the latency or the number of slots. This confirms that the basic performance of the slotted MACs does not degrade as the geometry is increased. As a simple demonstration of the accuracy of equation 5.10 with Poisson arrivals, figure 5.1 compares a plot against simulated values for a typical ring.

5.2.2 Exhaustive Service Token Ring.

The token ring considered here uses the early-release 'multiple token ring' MAC, whereby the free token is appended to the end of the outgoing packet before the transmitted packet has been stripped. Again, bulk Poisson arrival of cells is is used as the canonical form for the arrival process. Although this is unusual for a token system, it does not degrade the MAC and it enables the results to be readily compared with the slotted systems.

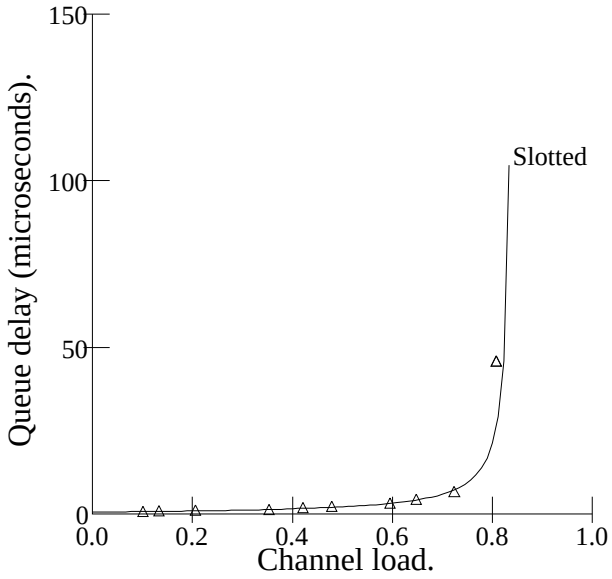

Figure 5.1: Slotted ring analytical and simulated mean queuing delay for Poisson arrivals. The simulations are denoted with the triangles ($\overline{k} = 1, \alpha =$ 1, 100 slots).

This analysis is for gated stations, meaning that packet arrivals during possession of the token are not transmitted until the next possession. The variation of the token rotation time owing to the random arrivals is not modelled, therefore the mean residual life of a token rotation is half $\overline{T_{rt}}$.

Let $\overline{Q}$ be the average queue length in cells at a station and Q_{max} be the expected number of cells in the queue when the gate is closed at token arrival. We assume $\overline{Q}$ is half Q_{max}.

The time the last cell of a message in the bulk arrival process waits in the queue consists of three components:

- The queuing delay before token arrival which is, as stated, $\overline{T_{rt}}/2$.
- The queuing delay after token arrival while other packets are sent which is $E\overline{Q}\sigma = E\lambda\overline{kT_{rt}}\sigma/2$
- The transmission time of all but the last cell of the message: $(\overline{k}-1)\sigma$.

Again, the transmission time for the last cell and its propagation time must be added to obtain the full message delay.

The average rotation time is

$$\overline{T_{rt}} = NQ_{max}\sigma + T \tag{5.11}$$

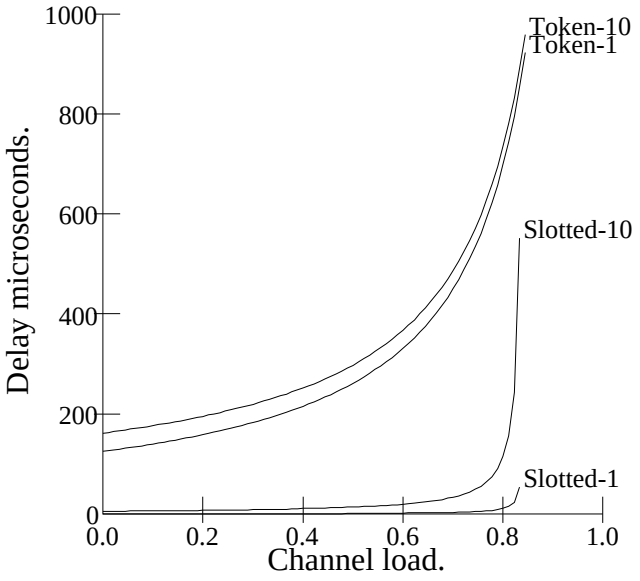

Figure 5.2: Queuing delays given by the analytical models for slotted ($\alpha = 1$) and token rings with 1 and 10 packet bulk arrivals. The line-rate is 500 Mbit/second with 50 kilometres of fibre. Packet size was 256 and 2000 bits respectively with a 32 bit overhead applied to both types of packet.

Now

$$Q_{max} = \lambda \overline{kT_{rt}} \tag{5.12}$$

Solving these gives the well known equation

$$\overline{T_{rt}} = \frac{T}{1-\rho} \tag{5.13}$$

where, as before,

$$\rho \triangleq N\overline{k}\lambda\sigma \tag{5.14}$$

The queue delay can now be written down

$$\overline{W} = \frac{\overline{T_{rt}}}{2} + \frac{E\lambda\overline{kT_{rt}}\sigma}{2} + (\overline{k}-1)\sigma \tag{5.15}$$

$$= \frac{1}{2}\left(\frac{T}{1-\rho}\right)\left(1+\frac{E\rho}{N}\right) + (\overline{k}-1)\sigma \tag{5.16}$$

It is immediately clear that, for α equal to unity, the token ring delay is greater by roughly $T/(\overline{k}\sigma)$, the slotted ring achieving even lower delays with lower values of α. A graphical comparison of delay versus throughput is shown in figure 5.2. Also, it is clear from the equations that E has little effect on the token system since it appears in the numerator, whereas in the slotted system, it has a significant effect at high loads since it offsets the denominator as the denominator approaches zero at saturation.

	Non-expedited.				Expedited.			
	Mean.		99 tile.		Mean.		99 tile.	
	Poisson	Bulk	Poisson	Bulk	Poisson	Bulk	Poisson	Bulk
Source-Release	7.8	126	49	521	6.1	72	25	393
$\alpha = 1.0$	4.2	44.1	–	–	3.5	3.5	–	–
Dest Release	1.5	28	10	20	1.1	5.0	5.0	4.3
$\alpha = 0.5$	1.5	22	–	–	1.4	1.4	–	–
Token system	211	223	427	595	190	203	398	532
	209	220	–	–	200	200	–	–

Access delays for source and destination-release slotted rings and a token ring, all operating with 18 active stations, 20 kilometres of fibre and 256 bit slots. The line-rate is 256 Mbit/second which gives 100 slots since the overheads are not included. The time units are normalised to slot time, σ, which is one microsecond. The offered load is homogeneous, with either Poisson arrival of single cells or bulk messages containing 13 cells. The offered load, in all cases, accounts for 75 percent of the channel bandwidth. For each type of ring, the upper figure is the simulated result and the lower figure is the analytical result using the approximate models presented in this chapter.

Table 5.2: Analytical and simulated ring access delays.

The token system is more efficient at handling very bursty traffic since the bulk size, $\overline{k}$ is not multiplied by a factor of $1/(1-\rho)$, whereas it is in the slotted system. An exhaustive token system operates more efficiently if there are messages which would take more than a ring revolution to transmit, and is superior slightly sooner when the overhead of slot headers in a practical system is taken into account. As an example of the cross-over point, if all messages are about 2 kilobytes, such as during a typical TCP/IP file transfer, then a well loaded, 1 gigabit token network will offer lower delays than the slotted ring, provided there is less than about 4 kilometres of fibre in total. It will, however, offer twice the delay of the slotted system at low loads. We might therefore chose to use a token system when it is not desired to support real-time traffic and the average message size is greater than the number of bits stored in medium.

In table 5.2, these equations have been evaluated and compared with some simulation point results. Three types of ring are included in the table. They use the token, source and destination-release slotted access protocols. The line-rate of 256 Mbit/second and length of 20 kilometres represent more of an HSLAN rather than a MAN, but it is clear that the token system is showing considerably greater delays already. The destination-release slotted ring shows, without exception, lower delays than the source-release ring,

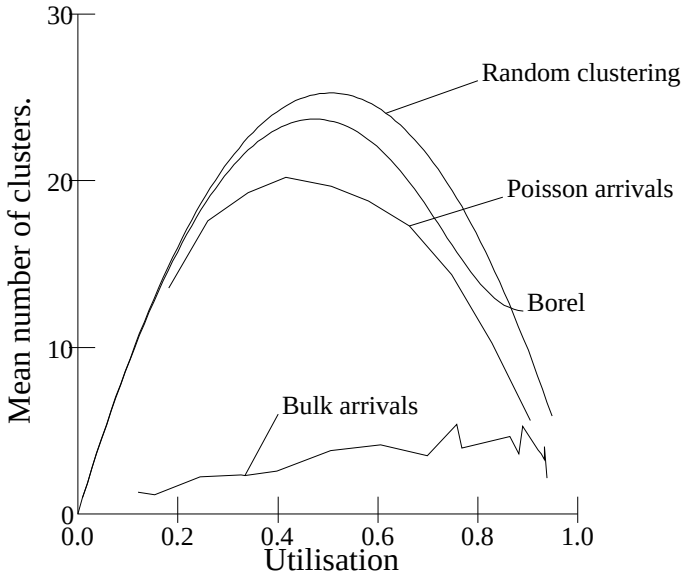

Figure 5.3: Average number of clusters of consecutive full slots versus utilisation on a ring with 100 slots.

even for the expedited 99^{th} percentile under bulk arrivals, where the intrinsic load balancing of source-release might be expected to help. Evidently, in this case, the lower density of full slots reduces the delays by a greater extent than the load balancing of source-release. This is because source-release, as mentioned before, does not provide a particularly fine granularity of fairness and the offered traffic was already relatively smooth. In fact, the non-expedited traffic on the destination-release ring suffered lower delays than the expedited traffic on the source-release ring.

5.2.3 Accuracy of Slotted Ring Studies.

The analytical values presented in table 5.2 are lower than the simulated values. In particular, the delay for expedited traffic when there is bursty background traffic is greatly underestimated. The reason for this is the assumption of adjacent slot independence. The errors from the assumption are shown in figure 5.3. A cluster is a block of consecutive full slots on the ring. The graph shows the average cluster length pertaining to a slotted ring with 100 slots at various utilisations. There are four lines on the graph. Two show results from simulations, one with Poisson cell arrivals and one with Poisson bulk arrivals of size 300. A third line shows the expected amount of clustering if the slots were really independent and the fourth shows the expected cluster length given by the mean busy period of an M/D/1 queue. With truly independent slots, a short inductive proof shows the third curve

to be given by

$$\text{Expected number of clusters } E_c = \rho + (Q-1)(1-\rho)\rho \tag{5.17}$$

where ρ is the overall density of full slots and there are Q slots. The average length of a cluster is given by dividing ρQ by E_c.

With Poisson arrivals, the cluster length in practice is a bit closer to the M/D/1 model shown in the 'Borel' curve. Under source-release, the slots are full for a constant time equal to one revolution, therefore the cluster length can be approximated to the mean of the Borel-Tanner distribution for the busy period

$$P(N=n) = e^{-n\lambda}\frac{(n\lambda)^{n-1}}{n!} \tag{5.18}$$

but again this is not a very accurate model.

An overestimate of the delay expected on a slotted ring results if it is assumed that there is exactly one cluster consisting of all the ρQ full slots aggregated together. This can give quite accurate results for LAN geometries where there may be less than ten slots. At larger geometries it is not very accurate. The clustering graph shows that there will be several clusters, even with extremely bursty traffic.

Further analysis of the simulation results has shown that even with Poisson sources the cluster lengths are hyper-exponential. This leads to hyper-exponential service times for the queue at each station. In addition, since successive slots are not independent, in particular, the probability of a free slot being higher than average if the previous slot has just served a station queue, the station service epochs are not memoryless. Therefore conventional embedded Markov chains and the M/G/1 results cannot be applied, even if the service time moments were to have been characterised.

A comprehensive series of analytical models for slotted rings has been recently published [ZAFIROVIC 88a and 88b]. The model which they found best approximated to the mean delay performance of the MSR protocol was in fact the exact solution to a polling system where a single cyclic server visits each station in turn, servicing one cell with fixed service and switch-over times. To account for the multiple slots in the real ring, they divided the service and switch-over times by the number of slots. This made the server appear Q times faster for Q slots. This model is inaccurate at very low loads since it cannot account for service from successive empty slots.

This makes a great difference for bulk arrival systems. Their final formula for the delay, presented in the notation of this chapter, is

$$\overline{W} = \frac{-1/2(N+1)\rho/2N + \overline{k}^5}{1-\rho-\rho/N} \tag{5.19}$$

which is very similar to the independent slot formula presented earlier (equation 5.10). The main difference is the fifth power of the bulk size. This high power can correct for the underestimates intrinsic to the independent slot method when $\overline{k}$ is about three or four, but yields wildly exaggerated results when larger bulk sizes are considered. In terms of applicability to different sizes of ring and types of traffic, it would appear that the capabilities of this new formula do not exceed those of the old one.

5.3 Comparison of Source and Destination-Release

In this section the performance of source and destination release rings has been compared to determine whether the pass-on-free rule improves the delay for traffic which is given priority using expedited transfer.

The four tables, numbered 5.3 through 5.5, present simulation results for both source and destination-release slotted rings under background loads of both Poisson cell arrivals and bulk arrivals of 300 cells. The foreground load in both case is expedited traffic, increased from zero to full capacity, and of a synchronous, fixed-interarrival nature.

The first two tables are for Poisson background loading, and as expected, these show that the destination-release ring offers lower delay and greater capacity. In tables 5.5 and 5.6, the Poisson background was replaced with the bulk arrival source. These tables show that the destination-release remains superior to source-release. This is even true for the 99^{th} percentile of expedited traffic, the case where it might be expected that source-release could be effective. This suggests that there is no load-balancing benefit of source-release below network saturation. At saturation of course, source-release ensures fair bandwidth sharing.

In practice, a computer network is unlikely to remain under saturating loads for any period of time because of source balking. This would ensure that slot usage patterns that generate near worst-case access delays are quickly redressed, and expedited traffic quickly regains access to low delays.

5.3.1 Alternative Load Balancing Mechanisms.

Probabilistic load balancing can be applied to destination-release slotted rings. An immediate advantage of the intrinsic symmetry of a ring is that the same parameters can be used at each station. The techniques used for slotted buses apply, but there are some more interesting protocols which rely on the ring topology. The source-release protocols require a ring profile memory which records which slots a station is currently using. This can be preserved and used only for load balancing when the slots have been freed at the destination. A suitable protocol is for a station to avoid re-using a slot until it has seen it go past in the empty state a number E times. The profile memory is used to contain the E counter for each slot. The maximum point-to-point bandwidth is then $1/E$, but the value of E can be dynamic, being determined at each station according to the locally measured ring utilisation. This protocol has many interesting cousins which are worthy of further study. The author's current favourite protocol is now presented.

5.4 DSR: A Slotted Ring Protocol Combining Source and Destination Release

From the last section, it is clear that destination-release offers superior performance to source-release in all respects except load balancing at saturation. However, even source-release does not incorporate a MAC layer priority mechanism, it merely guarantees the station a fraction of bandwidth for use with traffic of any priority. The shortcomings of quota based load balancing have been pointed out in chapter 4. This section describes a new slotted ring protocol which attempts to combine load balancing of source-release with the performance of destination release. In addition the protocol includes a fully effective priority mechanism where all traffic above the saturated priority level is transmitted, fairness is ensured within the saturated level and traffic at lower priorities is cut off. The granularity of load balancing within a priority level is the same as that of a source-released ring.

The protocol has been termed the 'double-slot' slotted ring (DSR) protocol. It operates through a distributed reservation algorithm entirely within the media-access control layer without higher level load regulation being required to face and reject excess offered load at any priority level. There is no rule restricting one transmission per ring revolution. Therefore the pro-

tocol efficiency is virtually independent of the number of slots and the ring latency. The DSR protocol therefore looks very attractive for multi-service networks at large geometries.

This section first describes the frame format required by the DSR protocol and then the protocol itself. As an aid to understanding, the protocol is first described without its priority mechanism.

5.4.1 DSR Frame Format.

As is usual for a slotted ring, the ring physical layer is formatted into frames of fixed length which continuously rotate. For the DSR protocol without low-level responses, each frame contains two different sized slots. There is no logical association between the two slots, they are grouped into a frame to ensure that there are equal numbers of each type of slot and for ease of a hardware implementation. Low-level responses can be added using a third type of slot in each frame as described in section 5.1.2.

One of the slots in the frame is termed the data slot. This forms the data carrying payload part of the frame. It dominates the other type(s) of slot in size, resulting in good channel efficiency. The second type of slot is the reservation slot, which can be encoded in two bits for the simplified version of the protocol, and in less than half a byte to support more than ten distinct priority levels.

The data slot contains a minimum of a full flag, a monitor-passed flag, a routing tag or address fields and a cell data field. The data slot is used in the usual way for a destination-release slotted ring, being filled and marked full by the transmitter, and copied and marked empty by the receiver. Broadcasts require source-release of course. The DSR protocol does not prohibit a station from immediately refilling a data slot it has just cleared with new data, provided it is authorised to transmit by the load balancing mechanism. Whether a station immediately re-uses the slot is an implementation detail, but higher throughput will result if it does (table 5.1).

The monitor-passed mechanism frees permanently rotating full slots and corrects for 'livelock' as described in [WHEELER 89]. Both the data and reservation slots require such protection independently.

5.4.2 DSR Protocol Without Priority.

Under this simplified DSR protocol, the transmit side of each station maintains a single extra counter termed its authority. This counter is non-negative, and contains a value which represents the number of cells the station is authorised to transmit. For each transmission into an empty, passing data slot, the station decrements the authority counter and it cannot transmit when the counter is zero. Although the current description is for the DSR without priority, it should be noted that a fairly effective priority mechanism can be supported by maintaining expedited and non-expedited queues within each station, the expedited queue always being served first.

The authority counter is increased in value by successful reservations. A reservation is a successful transmission into a reservation slot. The transmission simply consists of marking full an empty reservation slot, and then, one revolution later, freeing it, *and passing it on free*. The counter is incremented as the full bit of the reservation slot is written. Therefore no delay penalty need be incurred when making a reservation and an optimum implementation would be able to transmit into the data slot of the frame which contained the reservation slot. The amount that the authority counter is increased by a reservation is two. Two is an approximation to the ratio of data to reservation slot transmissions that is possible. For example, under homogeneous traffic with 18 stations which do not send to themselves and pass on used data slots free, the actual saturation ratio is 1.8.[5] A station is only permitted to make reservation transmissions if its queue length exceeds its authority.

In order to free reservation slots used in the previous rotation, stations are obliged to keep state about each slot on the ring. As described, this is termed 'profile' information and is kept in a profile RAM, indexed by a counter which is incremented for each frame that passes and overflows once per ring revolution. So far, the profile entry for each slot needs be only one bit, but more state needs be kept as we increase the protocol complexity.

[5]This saturation ratio is given from table 5.1 using $2N(N-1)/(N(N+1)-2)$ for $N = 18$ stations. The effect of this approximation is for further study.

5.4.3 DSR Protocol With Priority.

In order to support MAC layer priority, a station maintains separate queues for each level of priority. The levels are numbered 1 to P with P being the highest level. Instead of a full bit, the reservation slot of the full DSR protocol contains an integer, R, which represents a reservation level. If R is zero, then there is no reservation and the reservation slot is termed 'empty'. Slots are initially written with R equal to zero by a ring monitor and a collection system implemented by the monitor ensures that they return to zero under error conditions. The single authority counter at each station turns into an array under the full DSR protocol. The array contains one entry for each priority level, 1 to P.

The full DSR protocol operates as follows. To make a reservation at priority level P, a station must overwrite the R field in a reservation slot with the value P and this new value P must have been higher than the value which was previously in the reservation slot. Upon making such a reservation, the station double-increments the authority counter at level P and is then able to transmit in two data slots as before. When the reservation slot comes around again during the next revolution, if it still contains the same R value, as determined by comparing with the profile entry, then it is cleared to zero and passed on free. On the other hand, if the reservation slot has been updated, implying that it will contain a higher value of R, then the new value of R is left intact, but the station's authority for the previous value must be double decremented. The station may already have spent these authorities, so the authority counter is not decremented if it would go negative.

Additional elements of the protocol apply to lower priority authorities being stolen by the higher priority traffic within a station. These rules are provided since it would be silly for a station not to send a high priority cell owing to the lack of an authority of the appropriate level when it already owns an authority for a lower priority cell. A station attempts to make a reservation at the highest priority level for which its queue length exceeds its authority. When it transmits, it always sends its highest priority cell and always decrements the lowest authority counter that is non-zero. If there is no non-zero authority, then it cannot transmit. It is not allowed to use higher priority authorities for lower priority traffic. Stealing of authorities from lower class traffic requires an additional rule: if the authority level at any priority should exceed the queue length at that level (as can occur

since traffic has borrowed lower authorities) then the authority level must be clipped to the queue length. This is all summarised in the code of figure 5.4. This code has been included in a simulator. The routine 'Protocol' is called once per station, then all of the slots are moved on by incrementing the 'pos' variable at each station, mod the number of slots on the ring.

Synchronous expedited			Poisson background.			ρ
T/Put	Mean	99^{th} tile	T/Put	Mean	99^{th} tile	
26	1	8	127	1	8	0.6
51	1	8	128	1	8	0.7
77	1	8	131	2	10	0.81
102	2	10	129	2	13	0.9
128	2	13	131	2	13	1.01
154	2	13	126	3	18	1.09
179	2	15	127	3	18	1.2
205	3	18	126	4	25	1.29
230	3	20	128	6	33	1.4
256	4	25	129	8	50	1.5

Simulation results for the delay of expedited cells from a synchronous source at various loadings against a fixed background offered load of Poisson arrivals for a destination-release slotted ring. This table, and the next three, are again taken from simulations of a ring with 18 stations and 20 kilometres of cable at 256 Mbit/second and 100 slots. The column designated ρ is the sum of the two throughputs divided by the channel bandwidth. The throughputs are in megabit per second and the times are in units of one slot. The accuracy of the mean delays is about 5 percent and for the 99^{th} percentiles it is not better than 20 percent.

Table 5.3: Poisson background, destination-release slotted ring.

Synchronous expedited			Poisson background.			ρ
T/Put	Mean	99^{th} tile	T/Put	Mean	99^{th} tile	
26	3	18	127	3	20	0.6
38	3	20	128	4	25	0.65
51	4	25	128	5	35	0.7
64	5	28	127	8	47	0.75
77	10	50	127	14	78	0.8
90	12	57	127	24	140	0.85
102	23	95	128	119	430	0.9
113	325	895	128	753	1623	0.94
128	200	307	115	-	-	0.95
141	67	120	102	-	-	0.95
154	31	105	89	-	-	0.95
166	25	88	76	-	-	0.95
179	37	123	63	-	-	0.95
192	31	123	50	-	-	0.95
205	39	135	38	-	-	0.95
218	47	118	25	-	-	0.95
230	83	183	12	-	-	0.95
242	134	218	-	-	-	0.95

Voice delay as a function of voice loading against a background of Poisson arrivals for a source-release slotted ring. Below the horizontal division, the background queue is unstable, turning into a saturated source without any randomness. Accordingly, the ring is saturated and the variation in the expedited delays results from which of the many distinct slot occupation patterns the simulation has entered. Increasing the simulation time does not reduce the variability, owing to the non-ergodic nature.

Table 5.4: Poisson arrivals, source-release slotted ring.

Synchronous expedited			Bulk 300 background.			ρ
T/Put	Mean	99^{th} tile	T/Put	Mean	99^{th} tile	
26	26	348	123	486	1348	0.58
51	21	280	123	551	1717	0.68
77	20	257	123	610	2200	0.78
102	25	288	135	721	3292	0.93
128	26	275	132	810	3313	1.02
154	26	275	126	924	3902	1.09
179	27	265	123	945	2965	1.18
205	34	322	126	1324	6327	1.29
230	41	370	129	1811	8793	1.4
256	45	387	123	2017	8555	1.48

Voice delay as a function of voice loading against a background of Poisson bulk arrivals of size 300 cells for a destination-release slotted ring. The unevenness of the background throughput results from the relatively small number of bulk arrivals simulated, namely about 500 for one half a second simulated time. In these results, the network did not saturate, since the maximum ρ is less than the saturation ratio, which was 1.8 for the number and type of stations simulated.

Table 5.5: Bulk 300 arrivals, destination-release slotted ring.

Synchronous expedited			Bulk 300 background.			ρ
T/Put	Mean	99^{th} tile	T/Put	Mean	99^{th} tile	
26	25	248	123	1468	3785	0.58
38	20	185	123	1552	3657	0.63
51	29	280	126	2128	7100	0.69
64	45	283	129	2897	8483	0.75
77	71	310	135	3778	-	0.83
90	81	448	126	4808	-	0.84
102	131	425	129	7894	-	0.9
115	175	495	123	10121	-	0.93
128	180	478	105	-	-	0.91
128	235	568	96	-	-	0.87
154	83	363	69	-	-	0.87

Voice delay as a function of voice loading against a background of bulk arrivals of size 300 for a source-release slotted ring. The measured throughput of the background traffic falls off at high loads since the stations are throwing away cells regardless of message boundaries, resulting in incomplete messages being wastefully transmitted. Only the last cell of a bulk counted towards the measured throughput.

Table 5.6: Bulk 300 background, source-release slotted ring.

Synchronous priority 3			Poisson priority 1.			ρ
T/Put	Mean	99^{th} tile	T/Put	Mean	99^{th} tile	
26	1	5	129	2	10	0.6
51	1	5	127	3	15	0.7
77	1	5	129	4	20	0.8
102	2	10	128	5	25	0.9
128	3	15	128	7	25	1
154	2	10	129	8	30	1.1
179	5	20	127	12	45	1.2
205	7	25	129	26	105	1.3
230	8	30	128	42	205	1.4
256	9	35	124	879	-	1.49

Simulated DSR delay of priority 3 cells from a synchronous source for various offered loads, against a fixed background offered load of Poisson arrivals. This table, and the others, are for a ring with 18 stations and 20 kilometres of cable with an effective data rate of 256 Mbit/second after all control fields have been accounted for. This ring has 256 bit cells and therefore 100 slots. The column designated ρ is the sum of the two throughputs divided by the channel bandwidth. The throughputs are in megabit per second and the times are in units of one slot. The accuracy of the mean delays is about 5 percent and for the 99^{th} percentiles it is not better than 20 percent.

Table 5.7: DSR performance against a Poisson background.

5.4.4 Three Brief Simulation Results.

Table 5.7 gives some point simulation results for the DSR protocol at HSLAN or small MAN dimensions. This table shows that as the amount of a foreground non-bursty source, such as voice or constant rate synchronous video source, is increased, there is little interference between the voice and a fixed background loading of Poisson cell arrivals.

The next table, number 5.8 shows the same information when the background source consists of bulk arrivals of cells with Poisson inter-bulk times and fixed size of 300. The bulk arrival background source has raised the 99^{th} percentile of delay for the foreground load to nearly the bulk size. This again suggests the introduction of dither.

Table 5.9 presents a comparison between the DSR protocol and a destination-release ring without a load balancing mechanism. The same network parameters were used as before. Unlike the first two tables, in this table the applied traffic was not homogeneous in its choice of destination station. The applied load consisted of 22 Mbit/second of priority 3 synchronous load, evenly ap-

```
TYPE Q = (0..P);          (* The priority level type *)
TYPE STATION = RECORD
    pos        : INTEGER; (* pointer to slot number at station *)
    Buf :ARRAY Q OF POINTER TO ARRAY bufferindex OF slotPTR;
    In, Out    : ARRAY Q OF bufferindex;  (* Circular buffer pointers *)
    auth       : ARRAY Q OF INTEGER;      (* Authority array*)
    profile    : POINTER TO ARRAY[0..maxframes-1] OF Q;
    END;

VAR stationarray   : ARRAY [ 0..maxstations-1 ] OF STATION;

PROCEDURE Protocol(station :INTEGER);
VAR oldp, pri, prim : Q;
    sp  : slotPTR;
    qlen : INTEGER;
BEGIN
  WITH stationarray[station]^ DO
    sp := ring[pos];              (* Pointer to current slot *)
    oldp := profile^[pos];        (* Old value from profile  *)
    profile^[pos] := 0;

    IF oldp .GT. 0 THEN           (* If slot in use before   *)
      IF oldp .EQ. sp^.R THEN     (* and if still ours       *)
        sp^.R := 0;               (* then reservation slot clear and pass on free *)
      ELSE
        IF auth[oldp] .GT. 2 THEN INC(auth[oldp], -2) END
      END                         (* else lose that authority *)
    ELSE
      (* Attempt transmit into free reservation slot, or one owned by another station *)
      pri := P;                   (* Search down from highest priority level *)
      LOOP
        qlen :=INTEGER(In[pri])-INTEGER(Out[pri]);       (* Measure queue length *)
        IF qlen .LT. 0 THEN qlen := qlen + buffersize END;  (* Correct for circular buffer *)
        IF qlen .LT. auth[pri]  THEN auth[pri] := qlen END; (* Clip authorities *)
        IF (qlen .GT. auth[pri]) AND (pri .GT. sp^.R) THEN     (* If allowed to reserve *)
          INC(auth[pri], 2);          (* then increment authority *)
          sp^.R := pri;               (* write to reservation slot *)
          profile^[pos] := pri;       (* and record in profile. *)
          EXIT
        END;

        IF pri .EQ. 1 THEN            (* Else try next lower priority *)
          EXIT
        ELSE
          pri := pri - 1
        END
      END
    END;

    CASE sp^.flag OF                  (* Data field update, examine full bit *)
      empty:
        prim := 1;                    (* Empty, so find minimum transmit authority *)
        WHILE (auth[prim]=0) AND (prim .LT. P) DO INC(prim) END;

        IF auth[prim] .GT. 0 THEN     (* Attempt to send at this level or higher. *)
          pri := P;                   (* Try highest first. *)
          LOOP
            IF TrySend(station, pri) THEN
              INC(auth[prim], -1);    (* Make and record transmission *)
              EXIT                    (* stealing authority from lowest. *)
            ELSE
              IF pri .GT. prim THEN     (* else try next lower priority *)
                pri := pri - 1
              ELSE
                EXIT
              END
            END
          END
        END
(* If full, examine destination field *)
    | full:
        IF station .EQ. sp^.dest THEN
          ReceiveFromSlot(sp);
          sp^.flag := empty           (* and pass on free to next station *)
        END
    END
  END
END Protocol;
```

Figure 5.4: Modula 2 encoding of the DSR protocol

Synchronous priority 3			Bulk arrivals priority 1.			ρ
T/Put	Mean	99^{th} tile	T/Put	Mean	99^{th} tile	
26	18	270	120	499	1170	0.57
51	14	240	120	613	2105	0.67
77	11	165	132	702	1895	0.82
102	10	145	126	892	2910	0.89
128	8	115	141	1167	3525	1.05
154	9	100	138	1681	-	1.14
179	10	60	129	1840	-	1.2
205	8	40	132	3077	-	1.32
230	10	40	132	4075	-	1.42
256	11	45	129	4992	-	1.5

Table 5.8: DSR performance against a background of Poisson bulk arrivals of size 300.

Protocol	Synchronous priority 3			Bursty priority 2			Poisson arrivals	ρ
	T/Put	Mean	99^{th} tile	T/Put	Mean	99^{th} tile	priority 1. T/Put	
DSR protocol	22	12	100	128	140	239	206	1.39
Dest-free	22	10	45	23	-	-	417	1.8
Source-free	22	48	100	12.25	-	-	208	0.95

Table 5.9: DSR performance compared with an unbalanced destination-release slotted ring and also a source-release ring.

plied and destined, a bursty source of priority 2 applied at one station only, and destined for the station on the other side of the ring, and a background, saturated source, applied with priority 1 to all stations, destined to all other stations in an even manner. The bursty source queues one cell each slot time for 250 slot times, then rests for 250 before starting again. This is a vicious example of the type of traffic which might be presented by variable rate video sources. The average bandwidth of this burst source is half the channel rate, namely 128 Mbit/second.

It may be remarked that both protocols acceptably handled the highest priority, synchronous sources. However, the DSR protocol was effective in reserving bandwidth for the bursty source whereas without the protocol, not all of the bursty traffic was transmitted. Without the protocol, a greater amount of traffic was transmitted overall, but not of the desired type. The network simply saturated at a throughput of 1.8 times the channel bandwidth. This reduction of throughput is a direct result of the deliberately extreme requirements of the applied load; the DSR protocol does not result in unnecessary throughput reduction, as can be seen from the previous tables.

5.5 Summary

This chapter has compared several ring access protocols, both through simple analytical models and computer simulation. A set of three primary performance metrics was presented, with respect to which, the slotted ring MACs were shown to perform very well. The performance of the destination-release slotted MAC has been shown to be one of the best. An additional source-released response mechanism has been proposed to counter the objection that destination-release is unsuitable for low-level responses. The value of load balancing and the need for a replacement mechanism when destination-release is used has also been identified.

The new double-slot slotted ring protocol (DSR) has been presented. This is capable of handling bursty, high-priority loads and is therefore uniquely suitable for variable-rate video traffic. The DSR protocol throughput exceeds that of a source-release slotted ring, while offering similar granularity of load balancing. It also offers an effective, MAC layer priority mechanism which responds in one latency. The marginal increase in complexity

of the overall system is not considered a worthwhile objection for a VLSI implementation.

It has been suggested that the real-time performance of slotted rings can be optimised if stations can be banned from transmitting in multiple consecutive slots. The introduction of random dither has been suggested, but not explored. In practice, the limited capabilities of host interfaces may break up large bulk messages into smaller bulks, thereby providing a sufficient amount of dither.

Chapter 6

Cambridge Backbone Ring

The next three chapters of this dissertation are devoted to the Cambridge Backbone Ring (CBR). The current chapter serves as an introduction to the project. It presents the project aims and describes and discusses the resulting frame format and access protocol. The Backbone Ring architecture supports stations of various complexity and bandwidth. The performance of some example configurations is presented. Chapter 7 describes the prototype stations that have been built and chapter 8 describes the host interfaces that have been built or are being considered for future implementation.

6.1 Project Aims

The primary aim of the Backbone Ring project was to design and build a fibre optic ring communication network. The design was to be aimed for a nominal 1 GHz clock frequency, and the operating frequency of the first implementation should be between 500 MHz and 1 GHz. The project is posed as a research vehicle, designed for exploring the following research areas:

- State-of-the-art optical fibre technology.
- State-of-the-art digital VLSI technology.
- Media access protocols applicable to a backbone network.

- Delay and bandwidth expectations of clients to a backbone network.
- Architecture of gigabit stations and high-bandwidth host interfaces.
- Insight into the suitability of the ring topology in the metropolitan area.
- Ongoing research benefits to end users of the backbone network.

The project is evidently aimed at many research areas, and therefore it was not considered viable to attempt to design the ultimate solution at the first attempt. Instead, the emphasis was to be on the simplest possible approach which could fulfil the project aims and still meet the basic specifications presented in the next section. This would result in a design characterised by considerable architectural and hardware modularity. The prototype architecture would be suitable for experimentation with various station configurations, and, by initially keeping the MAC protocol out of the semi-custom VLSI, the architecture might be able to support a reasonable spectrum of MACs.

6.2 Backbone Ring Design Considerations

This section presents the considerations which influenced the design of the first version of the Backbone Ring. These considerations have been classified into three categories: the fundamental operating region specifications, the features which must necessarily be included in the design, and the features which it would be nice to include in the design.

6.2.1 Fundamental Operating Region Specifications.

The operating region specifications are values which relate the to physical quantities of the network.

Network Serial Line-Rate: The network was designed to be clocked at 1 GHz. The corresponding bit-rate would naturally depend on the line coding efficiency, which should be as high as possible. A secondary specification was that the system bandwidth, which is the saturated throughput

when account has been taken for all of the modulation and MAC overheads, should exceed 50 percent of the clock rate.

Minimum Cable Length: The minimum length of cable that can be supported is determined by the MAC layer protocols: for instance, the empty slot protocol requires that the ring delay can encompass at least one slot. The maximum length is liable to be determined by the maximum values that can be held in various counters and timers in the MAC hardware, and these, in principle, can be extended by the designer to an arbitrary length. For the Backbone Ring, a minimum design length of 4 kilometres was envisaged, although a design which had no minimum length was recognised as superior. Drums of fibre optic cable of 4 kilometres length are readily available and do not take up much space. If necessary, such a drum can be used to artificially increase the length of an undersize installation and it can also form the basis of a self-test facility when a station is switched out of the ring and looped back on itself. An electronic shift register of similar capacity is probably undesirable as an alternative owing to the significant increase in station complexity which could result.

Maximum Cable Length: A maximum cable length specification of at least 200 kilometres was desired. This is in order to encompass a region with a diameter of at least 50 kilometres which is the figure IEEE working group 802.6 specified for a Metropolitan Area Network [SZE 85]. Because of the orders of magnitude range of cable lengths envisaged, the medium access control technique (MAC) needs be selected from those whose maximum throughput is not inversely proportional to the ring length.

Minimum Number of Stations: So that each station can operate in the loop-back self test mode, the minimum number of stations required for network operation must be one. This also requires that every station is able to perform network management functions, that is, become the active monitor station. If individual stations cannot have access to the whole network bandwidth, then the minimum number of stations before the network can be fully utilised is also a parameter. A system where approximately ten stations can make use of over 90 percent of the bandwidth was seen as acceptable.

Minimum Transmit Bandwidth: The minimum transmit bandwidth should be guaranteed by a load balancing mechanism which ensures fair sharing amongst the active stations. The granularity of sharing should be as fine as possible to ensure low average delays for expedited traffic, and

the load balancing mechanism should operate at a comparable granularity to constrain the 99^{th} percentile of delay.

Maximum Transmit Bandwidth: Since the first implementation of the Backbone Ring is unlikely to achieve the full 1 GHz clock rate, in order for the first implementation to be useful in practice, the lower limit on maximum transmit bandwidth of the station must be considered in terms of the lowest acceptable clock rate, 500 MHz. A Backbone Ring connection to an ethernet needs to be able to support at most 10 Mbit/second half-duplex and a connection to a simple CFR bridge [PORTER 91] must be able to support about 10 Mbit/second full duplex. Therefore allowing a reasonable margin, a half-duplex bandwidth exceeding about 30 Mbit/second would be a reasonable starting point for the simplest type of half-duplex Backbone Ring station. A bandwidth of 30 Mbit/second is 6 percent of a 500 MHz NRZ serial line.

6.2.2 Features That Must Be Supported.

CFR Interoperability: Ease of interoperation with the CFR was of primary importance, since this would greatly reduce the software and protocol effort require to establish the project.

No Receiver Contention: The datalink layer protocol currently used over the CFR is UDL (Unison Data Link) [TENNENHOUSE 86]. UDL does not include a retry or acknowledgement mechanism. Instead it relies on the hardware retry mechanism of the CFR below it and retry mechanisms in the transport and RPC protocols above it. Although the network would not be restricted to running only UDL, its support is required for the initial CFR interoperability. Since the effectiveness of hardware retry mechanisms for networks of larger geometries was unproven, a design where the receiver can accommodate back-to-back arrivals was imperative.

Saturated Throughput Unaffected by Medium Length: Owing to the envisaged orders of magnitude variation in the length of the physical medium over which the network must run, the media access control mechanism must be selected from those where the throughput does not degrade linearly with size.

Real-time Traffic Capability: The network must be able to provide the low delay and low jitter services required by real-time traffic.

Custom or Semi-custom VLSI: To simplify the engineering effort of the initial implementation, a custom or semi-custom VLSI component was required which included all digital interconnections operating directly at the serial line-rate. (The connection to the optical fibre channel is essentially analogue.) With this approach, there is the hope that as the speed of available VLSI increases, faster chip-sets for the Backbone Ring architecture can be implemented in future.

Reliable Maintenance Service: Unambiguous status information for monitoring and fault maintenance is vital for a network covering a large geographical area. The maintenance mechanisms must not impose excessive additional station complexity and must operate reliably during network fault conditions. However, dynamic hardware fault correction mechanisms would appear to increase complexity with little potential research benefit. Therefore, fault detection, but not fault correction was to be included in the first generation design.

6.2.3 Features It Would Be Nice To Support.

Evidently it would be nice to support a completely general class of MAC protocols. Unfortunately this results in extremely redundant implementations with low utilisation of the hardware, as demonstrated, for instance, by [SKOV 89]. This approach may be useful as a research vehicle where speed of operation is not a primary aim, or in future years, as a viable means for a manufacturer to support a set of established standards and *de factos* with fixed hardware. Regarding the Backbone Ring, maximising the speed of operation was a primary aim. It was therefore decided to trade flexibility for speed wherever required in order to remain within the available VLSI die area. For instance, in the resulting VLSI design, only one frame size could be supported owing to the large amount of reconfiguration required to switch between alternative sizes. Some of the other features that it would be nice to support are now listed:

Isochronous Traffic Capability: Some backbone proposals have made provision in the MAC protocol for supporting TDM isochronous channels. This was described in section 4.8. Although such support lies outside the research addressed in this dissertation, it would be nice to allow for isochronous support in the design.

Retry on Busy or Error: The value of a hardware retry mechanism has been widely discussed. The well known conclusion is that bit errors on an optical fibre network are sufficiently rare that they do not in themselves justify a hardware retry mechanism. The CFR network has low latency, so the retry mechanism is able to largely compensate for the use of a simple receive buffer mechanism that suffers from contention. It has already been stated that the Backbone Ring must be free from receiver contention, therefore the roles of a low level response mechanism are reduced to transmitting backpressure for flow control and providing automatic retransmissions in the case of receiver buffer overflow. Although these are not vital to the network, it would be nice to support them.

Various Size Stations: Section 2.6 showed that for many networking projects aimed for operation at 500 MHz and above, a rigid time-division multiplex frame format was selected. From a hardware complexity point of view, the benefits of a multi-channel design are that the number of components that are required to operate at the serial line-rate is reduced, and the speed with which data is transferred to and from buffer memories is reduced. These are very worthwhile advantages. This said, TDM was probably selected for these early projects because the choice of high line-rate components available to the designers was limited to simple multiplexors and demultiplexers. For the Backbone Ring, a rigid TDM approach was seen to offer little new research potential. This simple TDM also has the disadvantage that it is hard to make use of the full network bandwidth or provide full connectivity without replicating a line interface and media access controller (MAC) for each of the sub-channels at each station. However, it was not an aim of the Backbone Ring project to enable an individual station to have exclusive access to all of the network bandwidth, and therefore it would be nice to include a degree of TDM provided that full connectivity could be maintained and network performance was not significantly compromised.

Multi-level Priority Mechanism: As described at length in chapter 4, in order for an ATM network to support delay sensitive traffic, the fractional utilisation of the delay sensitive traffic must be controlled, and in order to achieve a high overall utilisation, a priority mechanism must be employed to expedite the transfer of the delay sensitive traffic. It would therefore be nice to support as many priority levels as is worthwhile.

6.3 Backbone Ring Architecture

6.3.1 MAC Protocol and Packet Size.

The so-called multiple slotted ring protocol (MSR) was selected for the Backbone Ring project. This was introduced in section **??** as the pass-on-free, source-release empty slot protocol, where stations transmit without restriction in any empty slots that come along. This MAC protocol was selected for its good delay and throughput performance at large geometries. For compatibility with the CFR, the same size of cell was used. This is shown

Destination address. (16 bits)	Source address. (16 bits)	Data field. 32 bytes. (256 bits)

Figure 6.1: CFR cell or mini-packet format.

in figure 6.1. It contains 256 bits of data and 16 bit source and destination tags. An additional control flag was added to the start of the slot. This was termed the T or type bit. It can either be regarded as an extension to the cell or as a MAC protocol flag. For the hardware design phase of the project, the important properties of this flag are that it can be written during transmission and the value recovered at receive time.

6.3.2 Number of TDM Channels.

A multi-channel architecture was selected since this permits stations of various bandwidths to be constructed and it provides a simple mechanism for non-homogeneous use of the bandwidth if desired (e.g. isochronous use of one or more channels). The potential to double the cell size by pairing adjacent channels or by pairing adjacent frames was considered valuable, owing to the uncertainty of the cell size likely to be adopted by the CCITT for B-ISDN. Either the European size of 32 bytes or the American size of 64 bytes looked likely. The adoption of a 48 byte compromise was not anticipated [CCITT I.121].

The number of channels that the bandwidth is partitioned into is an important design parameter; in effect, station complexity can be traded against point-to-point bandwidth. The number of channels selected must be consid-

ered in conjunction with other sources of multiplexing within the network. These consist mainly of the half-duplex action within a TDM channel, where it is not possible to transmit while receiving, and half-duplex access to the buffer memory if single ported, or else contention for the memory.

Although an architecture where the division ratio is programmable is potentially attractive, the complexity of implementation is virtually prohibitive, and the inevitable compromises of an attempted implementation tend to result in low MAC efficiencies. For example, in an early version of the Backbone Ring design, which was worked out in considerable detail, a redundant, duplicate copy of the MAC flag fields and several blank fields were transmitted in the frame format when the hardware was switched to a mode which doubled the number of channels. Consequently, a fixed number of channels are used in the final design.

As stated in the design objectives, the bandwidth of the simplest type of station needed to exceed about 30 Mbit/second. If there are too many channels, a one channel station will not be able to achieve 30 Mbit/second at the 500 MHz minimum clock rate. On the other hand, if there are too few, then the buffer memory bandwidth will have to be quite large when operating at the 1 GHz clock rate.

Apart from the bandwidth division caused by the partitioning, the point-to-point bandwidth is also reduced by the line code, the addressing overhead and the MAC rules. Early in the design phase, it was decided to use 4B5B line coding. This has an efficiency of 80 percent. Since the CFR size cells were to be used, the fraction of a frame which contains useful data will be roughly the same as on the Fast Ring, that is 84 percent. Finally, the pass-on-free protocol limits stations to a maximum utilisation of one half, in any one channel. Applying these figures to the 500 MHz clock rate, the point-to-point bandwidth we must have

$$\frac{500 \times 0.8 \times 0.84 \times 0.5}{\text{Number of channels}} \geq 30$$

which shows that we cannot have more than 5 channels. Of course, the number of channels simply must be a power of two, so four channels is really the maximum. The remainder of this section examines a lower limit on the number of channels, determined by the requirement to provide buffers when operating at the 1 GHz clock rate.

6.3.3 Buffer Memory Provision.

For simplicity in the basic station, only a single level of buffering was to be provided between the medium and the host device. The buffer devices must therefore provide both sufficient capacity to meet the stations storage requirements and sufficient bandwidth to transfer data to and from the medium in real-time. This buffer pool can either be soft partitioned into separate regions for transmit and receive on the various channels and at various priorities, or it can be managed in a completely general way by the management mechanism. Either way, an immediate penalty of using the same physical buffer for receive and transmit is that it may not be possible to transmit in a slot immediately after receiving from the previous one because the buffer is occupied storing the last word of the received cell at the time it would have to yield the first word for the following transmission. The performance penalty that this incurs is estimated in section 6.7.

CMOS bytewide static RAMs were selected for the buffer memories because they can now offer relatively high densities, for example 64K by 8 in a 28 pin package, and therefore a reasonable size buffer pool of 4000 cells can be fabricated with just four devices. This is single ported RAM, and therefore half its terminal bandwidth will be achieved in a simple station which does not operate interleaved banks. Of course, more advanced memory architectures are also possible in the more advanced stations, but the aim of the first version was to be simplicity. Section 7.7 contains a discussion on more advanced station architectures.

Bytewide static RAMs are readily available with 100 ns access times and for these devices, this is also the minimum cycle time. Although faster devices are available, the overall operating speed cannot be increased in proportion owing to the overhead of providing the next buffer location from the address generating management logic. If four such devices are used, the word size will be 32 bits. It is attractive that no further reorganisation is then required to interface to a 32 bit processor. A 64 bit word size may also be considered, but wider words begin to become impracticable when printed circuit board size is considered.

The buffer memories are required to store the source and destination address fields as well as the data field of a cell. Therefore they must be able

to store or supply the one channel's worth of network bandwidth as follows:

$$\frac{\text{Formatted bandwidth}}{\text{Number of channels}} \leq \frac{\text{Word size}}{\text{Buffer cycle time}} \tag{6.1}$$

The formatted bandwidth using 4B5B and a 1 GHz clock is 800 Mbit/second. Taking a 100 ns cycle time and a 32 bit word size, two channels would exceed the inequality by 25 percent, whereas four channels fits with a margin of 35 percent. This is a comfortable margin for buffer address generation. Four channels are evidently more interesting than two, and also more interesting than one channel and a 64 bit word size and slightly faster memories. In consequence, a four channel system was adopted and four channels were hard-wired into the access chips.

6.4 Channel Dynamics for the Backbone Ring

Various channel management strategies which avoid receiver contention have been discussed in section 4.6.3. For the Backbone Ring, an approach had to be selected, and this was the method of permanently assigned receive channels and dynamic transmit channels. This is the simplest strategy since the reservation interval consists only of examining the full/empty flag of the destination's own receive channel, which, in any case, is part of the MSR protocol. The method of predictable receiver assignment is nearly as simple to implement and this may also be used in the future.

6.5 Slot Reservation Protocol

The slot reservation protocol is responsible for indicating whether a slot contains a cell or is free for transmission. An important property is that it should recover from error conditions without deadlock or 'livelock' [PACHL 88]. The minimum station delay required to implement the protocol is also often discussed. However, there exist acceptable protocols which require less delay than the contentionless channel reservation interval (section 4.6.3), and so the delay issue is not pertinent. Further important properties are low MAC overhead, the ability to easily measure ring utilisation and a facility for the monitor to be able to quickly take the ring out of service in order to prevent transmissions while the ring format is being rewritten.

There are two known deadlock and livelock free protocols which do not require a counter in the MAC frame format or rely on the source address field being unique to each station on a ring. The first is protocol 5 of [PACHL 88], which is a toggle protocol [ZAFIROPULO 72] and therefore does not require a monitor for garbage collection. It operates by implementing a monitor station like protocol with probability one quarter on each slot in each station. Toggle protocols require comparing the full/empty flag with its value in the previous ring rotation. This is not a great overhead since a station operating the MSR protocol must keep profile information about the state of each slot in any case. However, Pachl's protocol does require that each slot contents be compared with the transmitted data upon return in order to detect 'failed broadcasts', and this presents an unacceptable hardware overhead for high-speed operation.

The second deadlock and livelock free protocol is the conventional Cambridge slotted ring reservation protocol. This protocol was devised by David Wheeler [WHEELER 89]. It uses a full flag (F) and a monitor flag (M) in each slot, and a unique monitor station. The protocol requires a minimum of hardware, given that the monitor station can be elected out-of-band (using the maintenance subsystem). It operates as follows:

Transmit: If the F flag is clear (and optionally only if the M flag is clear as well), then set F and M and transmit.

Monitor: At the monitor, clear the M flag of each slot and copy the old M value to the F flag.

Free: Clear the F flag, but leave the M flag alone.

This frees rotating full slots (deadlock avoidance) since the M flag is cleared on the first rotation and this clears the F flag on the next, and it eventually corrects for slots which have erroneously become empty on a saturated ring (livelock avoidance) because the monitor will tend to mark as full, slots which has been freed too early, thus preventing them from being filled by the next station – the classical livelock situation. In addition, the ring utilisation is readily discovered by counting the density of slots with the F flag set, and the monitor can prevent premature transmissions by setting all F flags.

6.6 Backbone Ring Frame Structure and MAC Protocol

Header	Full Monitor Type	Four CFR size cells, each contains nine 32 bit words including the routing fields.	Response and Qualifier	CRC
(4)	(4+4+4)	($4 \times 9 \times 32 = 1152$ bits)	(4+4)	(12)

Figure 6.2: Backbone Ring frame. It contains four CFR size slots.

The frame structure used for the Backbone Ring is shown in figure 6.6. The ring delay is formatted with an integral number of frames. The remaining delay, as seen by the MAC protocol logic, contains an integer number of alphabetic code words owing to the lower-level elastic buffers. This is filled with syn characters, as is the gap between frames. The syn characters serve as justification symbols and can move fluidly between frame boundaries when clock rate variations cause insertions and deletions. There is no unique, index frame to indicate a logical start to the frame sequence, since the protocol does not require this.

The frames are of fixed length. There are 298 4B5B blocks per frame, giving 1490 bits when modulated, or 1192 bits as seen by the MAC logic. The frame contains four slots. It starts with a header symbol, then there are the F, M and T flags, four of each, bit interleaved. Then there are four cells each containing nine 32 bit words, the first word of each cell containing 16 bit source and destination identifiers. The four cells are byte interleaved. Bit interleaving is used for the MAC control flags so that a set of four flags forms a single word on the output of a station's 4B5B decoder. This enables protocol decisions for the frame to be made very quickly in parallel. Byte interleaving is used for the main data fields since a byte is the basic unit of operation for the demultiplexers which read and write to the data fields. The frame trailer contains a response flag (R) and a CRC qualifier (Q) flag, for each slot and then a 12 bit CRC. The R and Q flags are bit interleaved in the same way as the F, M and T flags.

The MAC protocol of the half-duplex Backbone Ring station is now described. This is, of course, compatible with full-duplex, multi-channel stations; there is just less parallelism. At the start of a frame, the station knows which channels on which it would like to transmit, the so-called backlogged channels, and it knows its own receive channel, or home channel. The

destination address field of the home channel is deserialised and indirected through a bit map held in a 64K by 1 RAM. The map contains a one at a particular location if it is desired to receive cells with that routing identifier. The map RAM bit is anded with the F flag of the home channel to indicate whether a receive is to take place from the current frame. Receive is given priority over transmit. While the receive address is being deserialised, the backlogged channel vector is anded with an inverted version of the four F flags of the frame. This gives the transmission possibilities for the current frame. If this is zero, then no transmission occurs, allowing (greater) host access to the buffer memories. Otherwise, round-robin priority is used over the transmit possibilities to select one channel for transmission.

The protocol operates as follows:

Transmit: Set F and M flags. Write to T flag, data and R fields, Set Q flag. Append correct CRC.

Receive: Read all data and flags from slot and check CRC. Write R flag.

Free: Clear F flag. Read Q and R flags. Check CRC.

Since this type of station transmits in only one slot per frame, it is only required to free and check the response field of one slot in each frame. This action requires separate bit manipulation logic since slots must be freed while transmitting to or receiving from other slots in the same frame.

The Q flags in the frame format are CRC qualifiers and are needed since one CRC covers the whole frame containing four slots. Separate CRCs would result in four times greater hardware complexity. A shared CRC implies that a single transmission error can cause the loss of up to four slots. However, errors on optical fibre channels are very rare and this is more than compensated for by the increased channel efficiency. All stations operate a subsidiary protocol as follows:

Q flag protocol: Check the CRC of every frame. If bad, clear the four Q flags at the end of the frame. Append a new correct CRC.

Since the Q flags are set upon transmission and cleared upon CRC error, they indicate at all times whether the data in a slot may have been corrupted

during transmission, although the CRC itself will have been corrected. At a station, the Q flag is always anded with a signal that indicates that the CRC of the frame that the Q flag came from was good.

6.7 Stations of Different Sizes

The Backbone Ring architecture supports stations of various sizes and complexities. The performance of a station depends upon:

- how many channels to which it has simultaneous access,
- whether it has half-duplex or full-duplex access,
- whether it is blocked from transmitting immediately after a reception, and
- the amount of contention for its buffer memories.

The overall saturated utilisation of the network is a function of these parameters and also of the number of active stations. These limitations reduce the saturation ratio from the value presented in table 5.1; this is nominally $N/(N+1)$ for the Backbone Ring MSR protocol.

The least complex Backbone Ring station has half-duplex access to one channel at a time and cannot transmit if it received a cell from the previous frame. As to be described in chapter 7, this is the type of station that has been initially constructed. In the remainder of this section, the throughput of this type of station is compared with a station which does not suffer from blocked transmissions and also with a station which has full-duplex access to all of the channels.

Transmit opportunities are missed in proportion to the amount of ring traffic that is received. As stated, this leads to a reduction in the saturated bandwidth of the network. The partitioning of the network into TDM channels increases the penalty incurred by blocked transmissions. If there are C channels, the receiver is busy with a reception for C times longer than in a single channel system. C is currently fixed at four . Another way of stating this is that approximately four times as many stations are required

Network type	With C channels	Unpartitioned
Full-duplex station	$1 - \upsilon^C$	$1 - \upsilon$
Half-duplex station	$1 - \upsilon + \upsilon(1 - 1/A)(1 - \upsilon^{C-1})$	$1 - \upsilon$
Blocking half-duplex station	$(1 - \upsilon/A)(1 - \upsilon + \upsilon(1 - 1/A)(1 - \upsilon^{C-1}))$	$(1 - \upsilon/A)(1 - \upsilon)$

Table 6.1: Probability of being able to fill a slot in a frame according to station type and compared with an unpartitioned network.

to achieve the same saturated throughput. This will give a noticeable degradation with a modest number of transmitting stations, say up to 20, when these are also the receiving stations.

According to the notation introduced in chapter 5, υ is the probability of the transmit side of a station encountering a full slot. This is the same as the ring utilisation ρ for the source release, pass-on-free MSR protocol since each full slot makes one complete revolution ($\alpha = 1$). Let $1/A$ be the fraction of full slots on a station's receive channel containing cells destined for that station. If the N stations do not send to themselves, and the traffic destinations are well balanced, we can approximate

$$A = \frac{N}{C} \tag{6.2}$$

For a station which suffers blocking, and assuming that transmission and reception are independent events, the probability that a station is blocked in a frame is equal to the probability of it having received in the previous frame. This is υ/A. A half-duplex station cannot transmit into the current frame when it is receiving from the current frame. The probability of such a reception is also υ/A. Finally, a station cannot transmit into a frame if all of the channels which that it wishes to use are non-empty. At saturation, it will have packets outstanding on the complete set of C channels, so the probability of them all being full is υ^C. Equations for the probability of being able to transmit when the limitations of the various types of station have been considered are listed in table 6.7.

At saturation, all N stations are transmitting at every opportunity. The saturated full slot density of the network is then given by solving

$$\upsilon = \frac{N}{C} P_{tx}(N, \upsilon). \tag{6.3}$$

Table 6.7 tabulates the saturated slot density which would result on rings

Channels	1	1	4	4	4
Blocking	No	Yes	No	No	Yes
Stations	Full	Full	Full	Half	Half
2	0.666	0.585	0.474		
3	0.750	0.697	0.631		
4	0.800	0.763	0.724	0.500	0.381
5	0.833	0.807	0.782	0.607	0.473
6	0.857	0.837	0.820	0.692	0.559
7	0.875	0.859	0.847	0.753	0.634
8	0.889	0.876	0.867	0.797	0.696
9	0.900	0.890	0.882	0.829	0.746
10	0.909	0.900	0.895	0.852	0.785
12	0.923	0.917	0.913	0.885	0.839
14	0.933	0.928	0.926	0.906	0.874
16	0.941	0.937	0.935	0.920	0.897
18	0.947	0.944	0.942	0.931	0.913
20	0.952	0.950	0.948	0.939	0.925

Table 6.2: Half and full-duplex, blocking and non-blocking slotted ring saturated throughput

when using the various different types of station. This figure is directly proportional to the network throughput (section 5.2). The first result column gives the throughput of a non-blocking MSR protocol, which, as given in table 5.1, is $N/(N+1)$.

For a single channel system, the saturated throughput of a blocking system with N active stations can be shown to lie between the throughput of equivalent non-blocking systems with N and $N-1$ stations. This is to be expected, since for low values of N, stations miss nearly as many transmit opportunities through receive blocking as they do when they pass-on-free the slots which were used the previous ring revolution. As N is increased, the two systems become more alike, as indeed do all of the systems.

The four channel systems require a greater number of stations to achieve the same throughput than do the single channel systems. However, a multi-channel system was selected for the Backbone Ring specifically to reduce the throughput available at a simple station. As expected the blocking penalty is also greater for the 4 channel system. Simulations of the Backbone Ring have shown that lower values of saturation are achieved in practice than are predicted here. This is a result of the load being applied unevenly over the channels and the use of static receiver assignment. As stated, predictable

receiver assignment may be used in the future.

6.8 Basic Station Throughput

This section reviews the network and station bandwidths available in a 1 GHz implementation of the Backbone Ring. The Backbone Ring frame, containing 4 slots, has a length of 1490 bits, and this is increased by typically 20 bits of syn characters to 1510. The four slots contain 256 bit data fields, so the MAC and clocking overhead present an efficiency factor of 1024/1510 = 0.68. This is satisfactory since it meets a design aim of being over half the system clock rate. For a 1 GHz implementation, the system bandwidth is therefore $0.68 \times 1000 = 680$ Mbit/second.

The asymptotic bandwidth of the MSR protocol as the number of stations is increased is equal to the system bandwidth, and this property is preserved in the Backbone Ring architecture despite the multi-channel design and transmit blocking. The value of 680 Mbit/second should be multiplied by a value from table 6.7 to obtain the bandwidth available in an installation with a small number of stations.

The basic Backbone Ring station has access to only one of the four channels, so the maximum point-to-point bandwidth is bounded by one quarter the system bandwidth, viz $680/4 = 170$ Mbit/second.

The MAC rules allow this bandwidth to be fully used. It is a half-duplex bandwidth. Stations are able to receive back-to-back packets and can therefore receive at this rate. Transmitting stations can transmit back-to-back packets at this rate provided they are using more than one channel since the pass-on-free rule limits the utilisation of any one channel for a single transmitter to 0.5.

The maximum throughput of the basic Backbone Ring station is limited by the buffer memory bandwidth. The buffer memories are 32 bits wide and cycled at the ring clock rate divided by 160. This gives a bandwidth of $1000/160 \times 32 = 200$ Mbit/second. Data must be copied both into and out of the buffers which introduces a factor of two division in available bandwidth. In addition, the station logic demands access to the buffers for one cycle out of every nine, but this cycle is used for ring-side data transfer if required. The result is eight out of eighteen cycles being available for host

access on average. The half-duplex host bandwidth is then $8/18 \times 200 = 88$ Mbit/second.

6.9 Summary

The basic half-duplex Backbone Ring station is able to offer 45 Mbit/second point-to-point bandwidth for a 500 MHz implementation, and twice this for the full speed version. More complex stations are also possible, giving access to more of the network bandwidth. The usable bandwidth of the complete network, when clocked at 1 GHz, is 542 Mbit/second. The MAC does not rely on the cell source address for slot freeing or deadlock control. This field can therefore be used as an extension to the destination address or as reassembly information, as to be described in section 8.1.

The design uses four TDM channels. One of these can perhaps be reserved for isochronous use if necessary. The channel number can also be used to extend the 16 bit MAC addressing field by a further two bits if desired.

The partitioned design is a hybrid which uses straightforward TDM for access to the addressing and data components of the MAC, and parallelism in the VLSI device for high-speed direct access to the MAC control flags. This combination is very suitable for stations which use a mix of technologies with different speed to power ratios. Unlike previous TDM ring projects, full connectivity across the channels has been realised.

In this chapter, the important decisions made during the design of the Backbone Ring have been presented. The design has been carefully aimed for an operating frequency between 500 and 1000 MHz. When the Backbone Ring is incorporated into a larger networking system, the line-rate and number of channels will not have an important bearing on the overall network architecture. However, it is appropriate that technology dependent parameters should have influenced the details of the MAC layer.

A further tenfold increase in the bandwidth available through simple amplitude modulation of optical sources is foreseeable. This would make line-rates above 10 GHz possible. Of course, the Backbone Ring architecture would be reviewed before being implemented at such rates, but if the cell length were to be doubled, possibly making it more in line with future ATM specifications, and if the number of channels were raised to eight,

the Backbone Ring architecture may remain very attractive for low cost applications.

Chapter 7

Backbone Ring Hardware Design and Implementation

This chapter describes the Backbone Ring hardware design and the implementation of the prototype Backbone Ring which was built as a collaborative project between Olivetti Research Ltd. and the University of Cambridge Computer Laboratory.

7.1 Backbone Ring Optical Fibres

The Backbone Ring uses optical fibre interconnects for all distances greater than a few metres. Monomode optical components are used throughout since devices with bandwidths up to 2 gigabits are readily available from suppliers. We chose the 1300 nm wavelength since this is the most mature technology. The optical fibre is silica of the 8/125 type, which means that the light carrying optical core is of 8 microns in diameter within a 125 micron fibre. This type of fibre was selected since it has a spot size compatible with the British Telecom (de facto) standard.

The Backbone Ring uses direct amplitude modulation of the 1300 nm light.

7.2 Line Code and Modulation Scheme

The Backbone Ring uses 4B5B/NRZI encoding. 4B5B coding was chosen because it can easily support a few non-data symbols, it offers good efficiency – viz 80 percent, and because 4 is a power of 2 – an important property in computer systems. The NRZI post-encoding is applied to remove static balance requirements from the codebook and to make the channel insensitive to polarity. The code book contains 66 percent ones which translates into a line transition every 1.5 bit times on average for random data. The longest run without a transition is four bit intervals.

Each five bit block, after NRZI encoding, either contains three marks and two spaces or two marks and three spaces. Although the line is balanced, that is, equal likelyhood of marks and spaces, there is no disparity control. The digital sum variation (DSV), which is the difference between the total number of zeros and ones transmitted, is therefore unconstrained, and so the baseline wander which results from the DC removal can range between (2-3)/5 = -0.2 and (3-2)/5 = 0.2. This is a maximum penalty of 2 dB.

The Backbone Ring implementation uses only one non-data symbol. This is known as 'syn' and is the character used to pad between Backbone Ring frames in order to fit an integral number of frames into the ring length. It is also used as the synchronous idle character when the ring is clear of frames at reframe time. Syn is the only character inserted or deleted by the elastic buffers, and since it never appears within a frame, the buffers do not require external knowledge of the frame format, but can operate autonomously.

Care was taken over the choice of codes used in the code book. It was important that syn has an odd number of ones so that after NRZI encoding, a ring being reframed does not develop a static baseline offset. Since syn is used for block alignment synchronisation it also had to be self-unforgable when rotated. The character used as a header flag to indicate start-of-frame was selected on the basis of as large a possible Hamming distance from syn in order to try to prevent it from being forged by bit errors. The decoder is able to distinguish between the header and syn even in the case of bit errors, therefore the chances of a frame being lost are reduced. As a final consideration, since computer data tends to have a preponderance of zeros and hexadecimal FFs, for the sake of clock recovery, the code for zero was given the maximum number of transitions which is five, and the code for all

ones was given four.

7.3 Telemetry Subsystem

Apart from the main data channel, the Backbone Ring optical fibres carry a low-rate telemetry channel. This channel provides the maintenance and management facilities which are vitally important to a distributed system such as a ring. The main functions of the telemetry subsystem are election of the active monitor station, control of the hardware at a monitor station, fault detection and location and reconfiguration of the optical fibres in the case of a station or link failure. The telemetry channel is currently used to distribute the Backbone Ring receive channel number to each station, but alternative methods may be used in future. The telemetry system also provides comprehensive monitoring of station status and parameters. This affords confidance in the network operation and will perhaps give an indication of imminent station failure.

The telemetry sub-channel carries HDLC frames at 1200 bits per second on a point-to-point basis, from one station to the next, in the same direction as the main ring channel. The channel uses a simple BFSK encoding. Baseband manchester encoding is currently used, but a carrier of several tens of kilohertz can be used instead by changing a PAL. This may be necessary if 1/F noise causes problems. The telemetry signal is additively combined with the main data channel at the transmitting laser diode of a station and reappears as a modulation signal in the active bias circuit at the optical receiver. The bit rate of 1200 bits per second is estimated to be sufficient for at least 100 Backbone Ring stations.

The telemetry subsystem is controlled by 8 bit microprocessors at each station. These have an identifying serial number burned into their EPROMs. They run a simple, multi-threaded coroutine kernel. One of the threads regularly generates a local telemetry message for transmission, while another forwards the received telemetry to the next downstream station. Messages are forwarded until they have made a complete ring revolution so that information about every station is available at each station. Another thread is responsible for maintaining a status display on an engineering terminal which can be optionally attached to any station. The display takes the form of a table which is updated in real-time. It shows global parameters, such

```
Prototype Backbone Ring Telemetry Status Display    CBG V2.1
-----------------------------------------------

Monitor:  5    Frames: 43     Gap: 64
Telemetry subsystem: OK
Monitor: OK
TVIA=1  RVIA=1  SVIA=F  BLRC=1

Serial   Syn I/D  Bit Crc  Optical
----------------------------------------
5  (M)  S   I    0   0    -4.5
12      S   I    0   0    -5.4
11      S   D    0   0    -4.9
```

Figure 7.1: An example of the display generated by the telemetry subsystem when three stations are present in a ring. Station number 5 is the monitor.

as the number of frames on the ring, and an array of station specific parameters, such as the bit-error rates at each station. Figure 7.1 shows an example taken from a Backbone Ring with three stations. It is not possible to show on one screen all of the information for a larger installation. The signals monitored by the telemetry include: receiver bias voltage, laser bias voltage, clock recovery PLL in-lock, 4B5B block decoder in-lock, 4B5B decoder violation rate, CRC error rate, power supply voltages, transmit clock operating, inserting or deleting status, and some host status information such as whether the host is on-line.

Each telemetry unit records whether it is receiving telemetry from its upstream neighbour, and if it is, it records whether its own telemetry messages appear in the local incoming telemetry, indicating that they are making complete ring revolutions and that the ring physical layer is intact.

7.4 Monitor Station Operation

The active monitor station on a Backbone Ring is elected by the telemetry units on the basis of highest serial number. The election protocol is now described. Received telemetry packets containing monitor station information are dispatched to a thread which is dedicated to monitor functions. These messages normally consist of monitor station status broadcasts which serve to distribute information about the current active monitor. However,

if none arrives within a time-out period equal to a multiple of the previous inter-arrival interval, the telemetry thread generates a special bidding monitor packet. Stations receiving bidding packets forward them if the source station serial number is greater than their local serial number, but if it is lower and the local station has been software enabled as a potential monitor, then they substitute their own bidding frame. A station receiving its own bid frame has won the election and commences monitor activities.

The monitor is responsible for writing the initial frame structure onto the ring medium. It first clears the ring to contain only syn characters. If a new calculation for the number of frames that will be suitable for the ring length is required, it will then write a single trial frame and measure how long the frame takes to rotate. It next writes out the initial frame structure. This consists of successive Backbone Ring frames with three syn characters between them and with their full/empty bits set. Once the format is established, it enables the monitor-passed garbage collection hardware. This frees up the slots so that transmissions can commence.

The monitor thread monitors the number of frames that are rotating on the ring by measuring the frequency of the profile counter overflow. The profile counter is used by Backbone Ring stations in order to detect their own slots coming back at the end of a transmission. It is incremented as each frame passes through a station and overflows once per ring revolution, its programmable division ratio having been programmed to the number of frames on the ring by the telemetry hardware. The number of frames on the ring is made known at all stations from the monitor station broadcasts.

7.5 Prototype Backbone Ring Station Architecture

The first generation Backbone Ring stations were designed in a highly modular form in order to simplify testing. These stations are of the half-duplex, single channel type.

The host connection was made through a 32 bit VME bus. The VME bus was selected because there was considerable VME experience and infrastructure available. The VME can offer tens of megabits per second throughput which is fast enough for many applications. Higher bandwidth attachments

will be constructed in the future. These are discussed in the next chapter.

Figure 7.2 is a block diagram of the half-duplex stations which have initially been constructed for the project. Each station consists of a double height card-cage, with the components shown in the block diagram partitioned into a set of plug-in assemblies. The physical arrangement in the card-cage is shown in figure **??**. Some of these are printed circuit boards and some consist of a series of screened boxes mounted on a metal frame. The eight assemblies are as follows, the VME bus being connected to assemblies 3, 4, 5, 6 and 7:

1. **Optical Assembly** containing the electro-optic receive and transmit interface modules and the optical bypass switch if it is fitted.
2. **Clock Assembly** containing the master transmit clock and the receive side clock recovery circuits.
3. **High-speed Board** containing the ring access chip set, the receive address bit map and the packet buffer memory.
4. **Low-speed Board** containing the station protocol control logic and buffer address generator.
5. **DMA and protocol engine.** An optional processor, described in chapter 8, which performs segmentation and reassembly, relieving the host processor of these jobs.
6. **Host Processor Board.** A standard part containing a 68000 microprocessor, local RAM and ROM and VME interface.
7. **CFR Interface or other LAN adaptor.** This is a standard VME part.
8. **Telemetry Module** containing an 8 bit microprocessor, an 8 bit analogue-to-digital converter, an HDLC chip and a serial terminal interface.
9. **Power Supply**

The CFR interface and the VME backplane processor are not strictly part of the Backbone Ring station, but the VME processor is required to initialise the bit maps in the Backbone Ring hardware and also perform

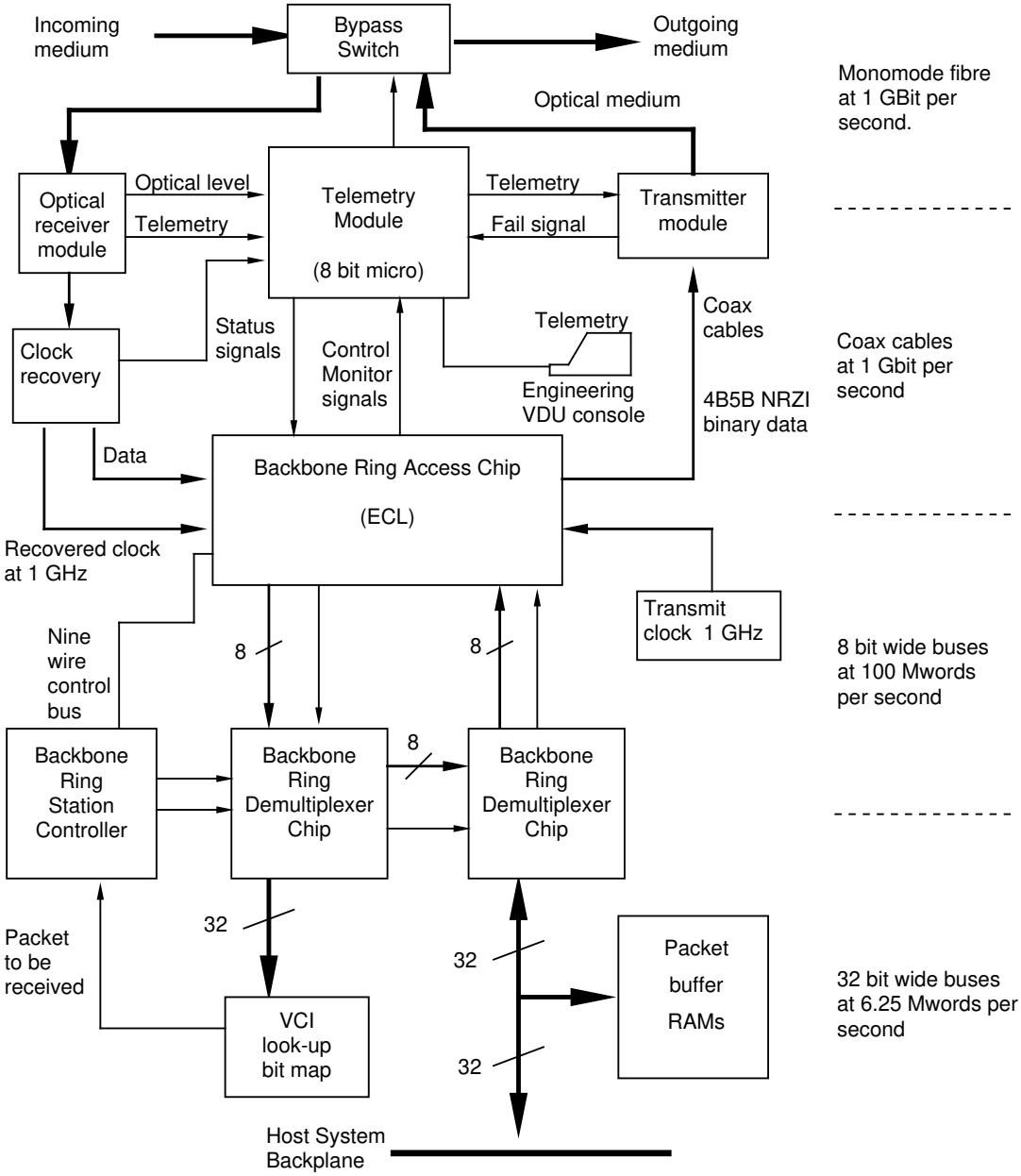

Figure 7.2: Block diagram of a simple Backbone Ring station.

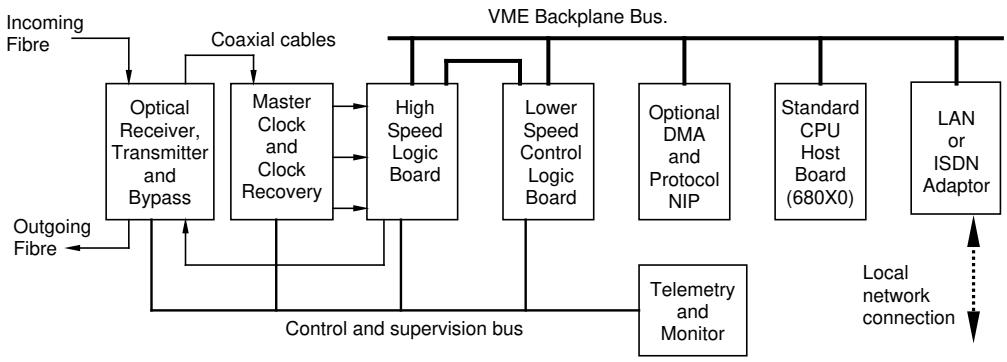

Figure 7.3: Physical arrangement of the simple Backbone Ring station modules. These are all placed in a 19 inch card-cage.

some management functions. The remainder of this section describes the connections between these assemblies.

The fibre connections to the ring are made via the optical assembly. Unless the station is bridged out by the bypass switch, the received optical signal is converted to a baseband electrical signal by the optical fibre receiver. It is equalised and amplified to an ECL compatible level and fed to the clock module on 50 ohm coaxial cable. The clock recovery circuit regenerates the receiver clock from the transitions in the data stream and, in the first prototypes, retimes the data stream with a GaAs flip-flop. Later designs will attempt to use the silicon flip-flop on the receive data input of the ECL access chip.

Two coaxial links from the clock recovery module feed the received data and clock to the high-speed board. A further coaxial link carries the transmit clock from the transmit clock module to the high-speed board. The transmit clock is included in the same assembly as the clock recovery circuit so that it can be derived from or phase-locked to the received clock if necessary according to the ring clock distribution technique that is in use. This was discussed in chapter 3.

A fourth and final high-speed coaxial link feeds the transmit side serial data stream from the high-speed module to the optical module transmitter. The transmitter performs the electro-optical conversion.

Apart from the four coaxial connections, the high-speed module has connections to the VME bus, the telemetry bus and the to the low-speed module. The VME bus connection is used for transferring host data to and from the buffer memories and also for host read/write access to the destination address bit map. The telemetry bus is wired using 50 way ribbon

cable connectors. It can connect to all of the modules and is simply daisy chained from one to the next across the front of the card-cage.

The high-speed board is connected to the low-speed board using the 64 undefined pins on the VME backplane. These are jumpered across the reverse side of the VME backplane at the appropriate sockets. This connection provides the address and control inputs for the buffer rams and also carries a number of dedicated signals which control the cycle-by-cycle operation of the station.

The low-speed board maintains the linked lists which represent the queuing order of the packets in the main buffer, it keeps track of which slots on the ring contain outstanding transmissions, and it provides the address decoding and synchronisation for the VME transfers to the buffer ram.

The telemetry module is the bus master for the telemetry bus. It also has two audio frequency connections to the optical board, which carry the incoming and outgoing telemetry channels, and an RS232 terminal interface.

7.5.1 Optical Fibre Transmitter Implementation

The block diagram of the optical fibre transmitter is shown in figure 7.4. It uses a 1300 nm GaAs semiconductor laser. The prototype transmitters were constructed with inexpensive lasers available from STC. These have had the fibre pulled back from the chip so that the thermal and mechanical tolerances are relaxed and the launch power is reduced to -10 dBm. These devices can be operated in the temperature range 0 to 70 degrees without requiring a cooler.

The rear facet monitor photo-diode is used in the usual way to bias the diode into the operating region. The telemetry is applied through the DC bias circuit while the main channel data is AC coupled through a matching resistor mounted physically close to the laser package.

7.5.2 Optical Fibre Receiver Implementation

The block diagram of the optical fibre receiver is shown in figure 7.6. The detector is a reverse biased PIN diode and this is mounted on the same substrate as the source follower GaAs FET to minimise capacitance. Despite

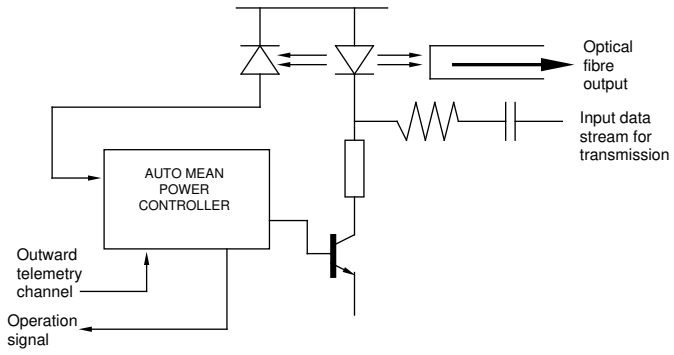

Figure 7.4: Schematic laser driver circuit at transmitter.

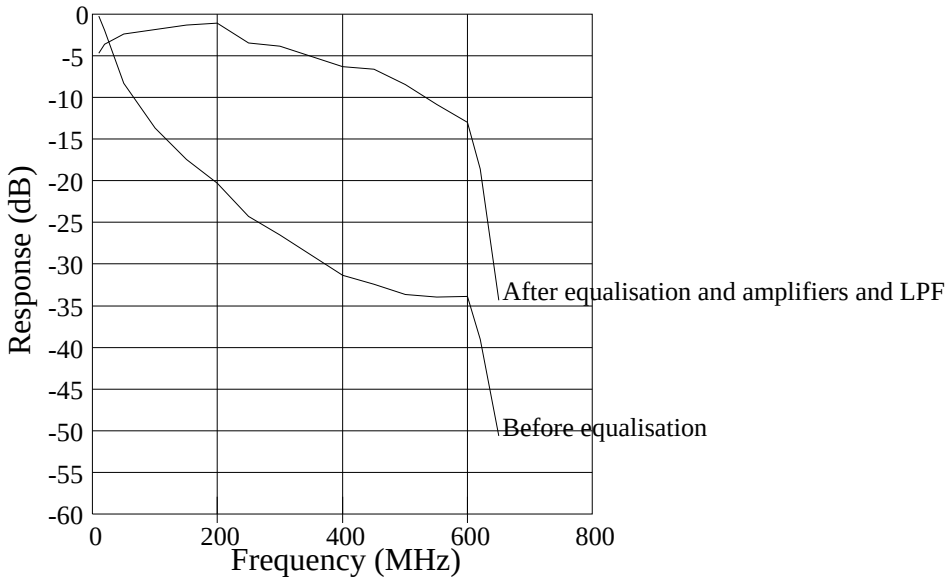

Figure 7.5: Frequency response of a prototype monomode fibre channel constructed for the Backbone Ring project.

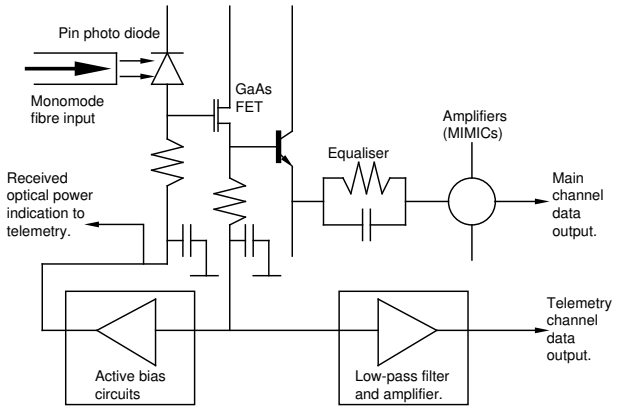

Figure 7.6: Schematic of fibre optic receiver module.

this, the input stage integrates with a time constant of 7 ns. This pole falls within the channel bandwidth and so it must be cancelled by the equalising lead filter which is after the emitter follower. The PIN-FET configuration was selected because of its inherent wide dynamic range and simple bias circuitry. Since there is no voltage gain in the PIN-FET front end, the signal at the equaliser can be just a few millivolts in amplitude when the fibre length is at its maximum. Several stages of amplification follow in order to raise the swing to the ECL level of 13 dBm. MMIC (microwave monolithic integrated circuits) are used. There is no AGC in the current design, the level regulation being effected by compression in the final MMIC. The frequency response of a prototype channel (from electrical to optical and back to electrical) with 4.5 kilometres of fibre is shown in figure 7.5.

The current in the GaAs FET follower is monitored by an op-amp stage in order to control the PIN diode bias voltage. Automatic biasing enables the full dynamic range of the receiver, about 25 dB, to be realised. The bias voltage serves a secondary function: it is buffered and fed to the telemetry unit ADC where it serves as a good indicator of the received optical power level. The FET current monitor is also used as the telemetry sub-channel take off point. The recovered telemetry signal is low-pass filtered and amplified within the receiver module and fed to the telemetry module on audio grade coaxial cables.

7.5.3 Master Transmit Clock Implementation

A block diagram of the transmit clock is shown in figure 7.7. The transmit clock consists of a fundamental mode varactor-tuned oscillator and a prescaling phase-locked-loop. The phase-locked-loop locks the transmit clock to a 4 MHz reference. The reference is either generated by a local crystal oscillator or can be supplied externally. The external source is required when using two-stage clock recovery as described chapter 3.

Using the asynchronous clock method, the master clock module is connected to the transmit clock input of the fast board at all stations. When using the open-ring clock distribution method, only the transmit clock module at the active monitor station is active. The telemetry control system operates a coaxial relay at the remaining stations which causes their transmit side to be fed directly from the recovered clock. The closed-ring technique can also be supported by switching the same relay at the monitor station,

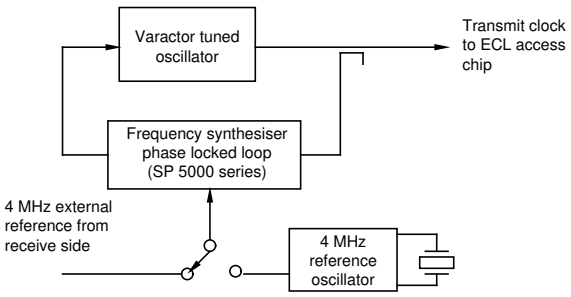

Figure 7.7: Local oscillator with provision for the two-stage clock regeneration method.

thus removing all of the crystal-locked clocks from the system. As mentioned in chapter 3, a frequency monitoring unit would be required if it was then decided to operate the closed-ring method as usual practice.

7.5.4 Clock Recovery Module Implementation

A number of clock recovery techniques are available owing to the high timing content of the line code. Initially two techniques are being used: PLL and helical tank. The PLL design is available off-the-shelf from Gigabit Logic. It uses a custom GaAs device which incorporates a varactor and negative impedance amplifier in order to form a VCO from a shorted stub line, a phase comparator suitable for NRZ streams, and a decision flip-flop for retiming [GIGABIT 88]. The second design of clock recovery circuit uses a simple helical resonator. This is being pursued with a view to reducing costs.

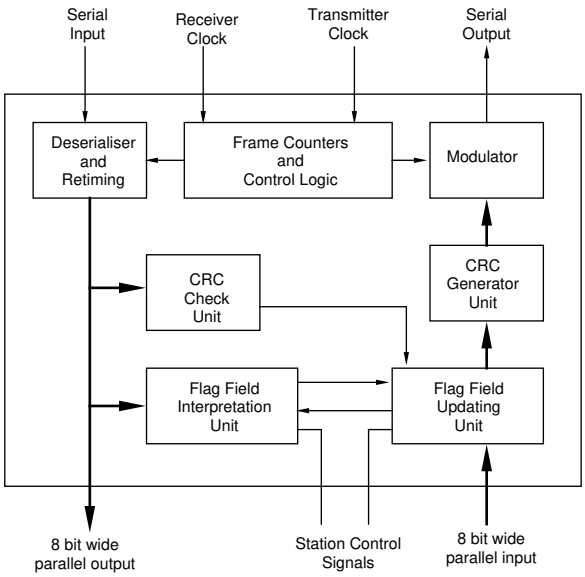

Figure 7.8: Block diagram of the Backbone Ring ECL access chip.

7.6 Backbone Ring Serial Access Chip

The first implementation of the Backbone Ring serial access chip is designated 'Raven'. Raven contains all of the high-speed digital logic required at a Backbone Ring station of any complexity, from a single channel, half-duplex station to a station with access to the full network bandwidth. A block diagram of the device is shown in figure 7.8.

It is driven from two separate clocks, one at the receiving link baud rate and one at the transmit link baud rate. The clock inputs and the serial ring data input and output enter and leave the access chip using differential pads operating nominally at standard 100K ECL levels. These four connections are DC balanced and AC coupled. External miniature transformers are used to convert the differential pads to single-ended coaxial SMA connectors.

Raven generates a byte clock at one tenth of the transmit clock rate for clocking the slower logic, and performs the following receiver functions:

- Decides if the current input voltage is a zero or one using a D-type flip-flop.
- Decodes NRZI to NRZ

- Decodes five serial bits into a four bit data word or the special non-data symbol 'syn' (4B5B decoding).
- Generates a code-error output on invalid data blocks.
- Gains bit-level synchronisation using a slip method.
- Inserts or removes 'syn' characters using an elastic buffer.
- Gains frame synchronisation on frame header.
- Checks 12 bit CRC of received data frame.
- Examines FE bits and arbitrates between multiple transmit possibilities using round-robin.
- Reads and latches flag bits for use by slower station logic.
- Converts the ring data to an eight bit wide parallel stream.

Raven performs the following transmitter functions:

- Receives the eight bit parallel stream which incorporates any new data transmitted by the station.
- Updates the flag bits of a single slot from each frame selected for receiving or transmitting.
- Updates the flag bits of a single slot from each frame in order to free and pass on slots used in the previous ring revolution.
- Optionally updates the flag bits of all slots to implement the monitor-passed garbage collection protocol.
- Checks and updates the qualifier bits of the outgoing frames.
- Appends a new, correct CRC to the frame.
- Encodes the stream to 4B5B and thence to NRZI.

Only 20 percent of the device operates at the ring baud rate, the remainder operates at one fifth and and one tenth the clock rate.

Raven does not contain sufficient flag bit manipulation logic to implement a station which can transmit in more than one channel at a time. If

more than one channel is to be used, or for a full duplex station, additional flag interpretation and manipulation logic must be inserted in the X bus and Z bus paths respectively.

7.6.1 Access Chip Engineering

The access chip design was manually keyed into a computer using the local Cambridge hardware description language. A cell library for a family of gate arrays from Plessey was also entered and this provided the simulation environment. Models of standard TTL devices were available and further models for new memory components were also written. The demultiplexer chip (see section 7.7) was also entered into this system to enable simulation of the complete station to the gate level. The optical receiver and transmitter were replaced with models which respectively read and wrote to computer files. The transmitted serial bit stream was 'disassembled' using a special program and a corresponding compiler was written to generate example serial stimulus files. The simulator ran approximately one hundred million times slower than real-time.

In order to verify the asynchronous design, which cannot be simulated since the events of interest would take too long to occur in the simulator, a practice implementation was made at 6 Mbit/second. This is called the low-rate system. It demonstrated the design of the elastic buffers and showed the behaviour of the elastic pad symbols. The low-rate system provided a test bed for the semi-custom components after they were manufactured and the telemetry software was also developed over it. Three stations were constructed, although the mini-packet buffers and host interfaces were not wired up. An extender board which behaved like 200 kilometres of optical fibre was constructed. The extender board used RAM memories to simulate the digital delay, the analogue group delay not being simulated. Therefore the extender does not give practical insight into the problems of closing large rings.

The designs for the chips were transferred to their native hardware description languages using a simple parser-generator program, and then they were placed and routed. The back annotations were fed into the Cambridge simulator, and again the whole station, down to mini-packet transfer to and from buffers was re-simulated. The simulations were verified on the native simulators before fabrication.

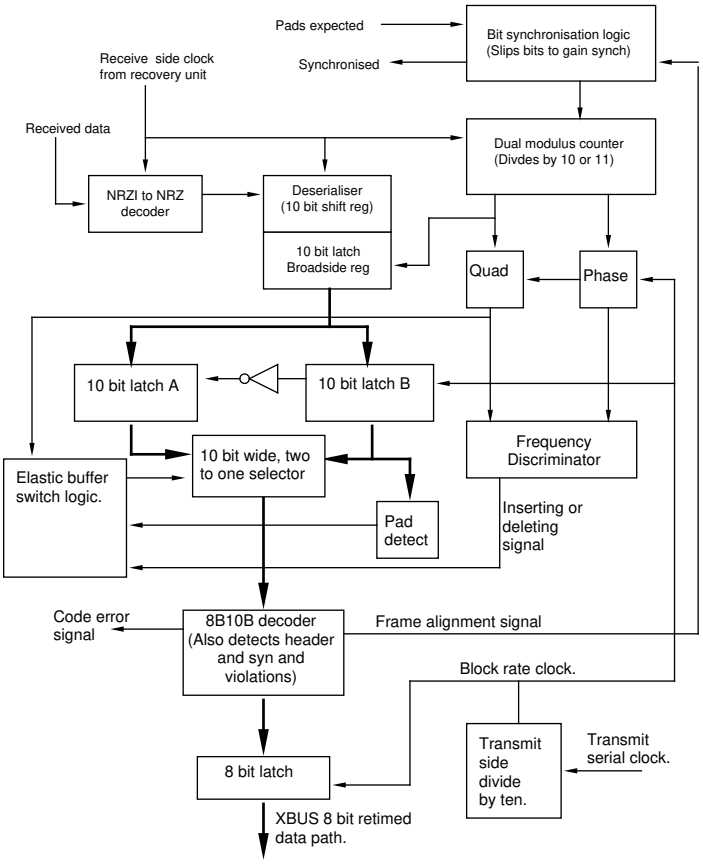

Figure 7.9: Block diagram of the receive side logic contained in the access chip, including the block synchroniser, the elastic buffer and the decoder ROM.

7.6.2 Access Chip Receive Side Circuit

The receive side circuit is shown in figure 7.9. Raven incorporates the decision flip-flop. This is a standard gate-array component with a specified aperture time of 210 ps. Optimum data set-up into this device is achieved by adjusting the relative lengths of the the coaxial feeders in the data and rx-clock paths from the clock-recovery unit to the circuit board containing the access chip. One possible technique is to empirically find the length which gives the highest bit-error rate, then substitute a cable half a wavelength shorter to achieve the optimum timing. This would be 10 cm shorter at the 1 GHz clock rate.

After the decision flip-flop, the received data is converted from NRZI to NRZ form by exclusive-oring it with a one-bit delayed version of itself. The NRZ data is deserialised using a ten-bit shift register. The parallel outputs from the shift register are latched by the broadside register every ten bit

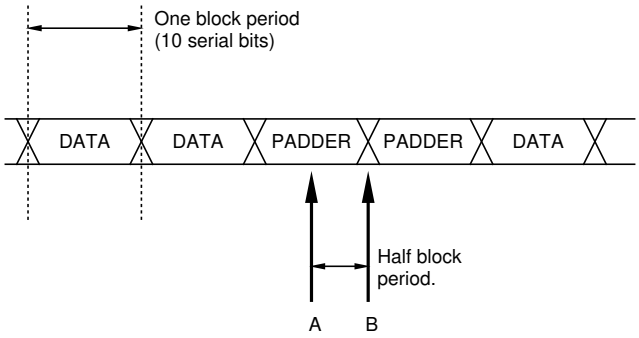

Figure 7.10: A representation of the received data being sampled by two phases of the transmit clock.

periods under control of the dual modulus 10/11 divider. These are the only devices clocked from the receive clock.

Correct bit synchronism is gained using a slip method at ring reframe time. The idle ring contains a constant stream of pad symbols. A simple finite-state machine controls the division ratio of of the dual modulus counter. It slips one bit every four blocks giving a maximum synchronisation time in the absence of errors of under 500 clock ticks.

7.6.3 Elastic Buffer Circuit

The receive side incorporates an elastic buffer which corrects for the slight frequency difference between the transmit and receive clocks under the asynchronous ring technique. It also removes the jitter from the received data and extends the ring physical delay to a multiple of ten bits. This is an important function, since there must be an integral number of 8B10B blocks for the ring to operate.

The elastic buffer consists mainly of the two ten bit latches and the switch logic shown in figure 7.9. The two latches are clocked on alternate phases of the transmit side block-rate clock. In general, one set of latches will have good set/hold times while the other set may be violated. The elastic buffer logic controls the ten bit wide, two-to-one multiplexor which selects the latch with the best timing.

Figure 7.10 shows the contents of the broadside latch being sampled by the two latches. At each clock tick, the arrows will move along nearly exactly one block length. At an inserting station, the transmit clock will be

slightly faster than the receive clock and the arrows will move to the right by slightly more than one block length each time. Conversely, at a deleting station, they will move slightly less than one block length each time.

As the transmit clock beats with the receive clock, optimum timing will pass from one latch to the other. The preferred latch is decided by the flip-flop Q which latches a quadrature version of the receive block-rate clock on the positive edge of transmit block-rate clock. This output cannot be directly fed to the two-to-one selector control input since it might change while mid-frame and delete or duplicate valid data.

Looking again at figure 7.10, it is clear that if both A and B point to the same block, then swapping between them will not insert or delete a symbol. However, if they are astride one of the boundaries, then the data at the output of the selector will be affected. In particular, when they are astride and the selector switches from A to B, then one block will be completely deleted from the output stream. Or if the switch is from B to A, then one block will be repeated.

The function of the elastic buffer logic is to permit the selector to change at appropriate points. One condition that it applies is that the A latch must contain a syn character. In this way, only syn characters are inserted or deleted. With just this condition, it would be possible for the elastic buffer to delete all of the syn characters at the boundary between two frames. Since the buffer has no knowledge of the higher level frame format and can only insert a syn where there is already one or more, this could result in a permanent loss of syns at such a boundary.

The second condition which the elastic buffer logic applies is that it only switches from A to B at a deleting station when the A latch has held a syn character for two successive blocks. This means that the last syn character will never be deleted from each frame boundary. In order to reduce overheads, this immutable syn is incorporated into the Backbone Ring frame structure. As described in chapter 6, it shares the same 8B10B block as the last four bits of the CRC. In order to only apply this second rule at deleting stations, all station require a frequency discriminator to determine whether the tx-clock is faster or slower than the rx-clock.

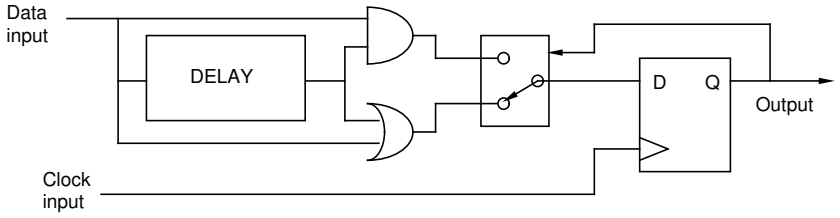

Figure 7.11: A D-type flip-flop with artificially increased temporal hysteresis. When the Q output is a one, the OR gate output is fed to the flip-flop.

7.6.4 Elastic Buffer Frequency Discriminator

The frequency discriminator appears as a block in figure 7.9. A second flip-flop is connected effectively between the receive and transmit block-rate clocks. The frequency discriminator compares this in-phase product with the quadrature signal already required for the elastic buffer control. On the positive edge of the in-phase signal the quadrature signal is latched. This will be a one if the transmit clock is faster than the receive clock and zero if not.

The I and Q flip-flops operate as frequency mixers. The flip-flop outputs are quadrature square waves with frequency equal to the beat between the clock and data inputs. With a conventional design of flip-flop, as the change of the D input encroaches into and through the aperture, the output does not switch once, but many times. This is because the variation in phase between the two inputs each cycle is less than the variation in the decision point resulting from circuit noise and jitter. For the discriminator application, the positive edge of the in-phase signal is important. Using a conventional flip-flop, there would be many positive edges each time the beat component changed sign. For this reason, the in-phase flip-flop is provided with positive feedback which gives some temporal hysteresis.

Figure 7.11 shows the circuit used for the in-phase mixer. This is functionally equivalent to a standard D-type. The feedback is arranged so that when used as a frequency mixer, the first time the output changes, the delay in the input path is modified such that the output will not immediately switch back. For the Plessey gate array, the switched delay component is two cascaded low-power inverters which combine to give 800 ps. The dis-

criminator output is made available on a device pin and included in the telemetry.

Should the rx-clock jitter be greater than 800ps, then the frequency discriminator will occasionally make the wrong decision, and this could possibly cause the last 'syn' between two frames to be deleted. This problem is solved by fitting an external RC filter to the discriminator output, and then feeding the filtered version back into the access chip using a new input pad. Unfortunately this facility was not provided in the first version of the access chip. Whether this will cause problems with large rings has yet to be determined.

7.6.5 Elastic Buffer Metastable States

The asynchronous flip-flops will potentially enter metastability if their input signal is slewing during the aperture. The design will not fail if they have snapped out of metastability within four baud intervals since this is the period before the the output of the asynchronous flip-flops is sampled. At 500 MHz, this is 8 ns. Following the method of [VEENDRICK 80], metastability will be shorter than this if the input signal normalised to the logic swing is outside the range

$$R = \pm \exp\left(\frac{1-A}{\tau} 8\text{ns}\right) \tag{7.1}$$

centred around the decision voltage. Where A is the open-loop gain of the flip-flop and τ is the dominant pole of each of the inverting elements. For the Plessey gate-array, the actual figures are unavailable. Instead we will use very conservative estimates, taking A as 5 and τ as the propagation delay of the device, 1ns. It is important to be exceedingly conservative in these two estimates, then if the results from the estimated values are acceptable, there is a comfortable operating margin, whereas if optimistic values are chosen, the calculated results may fall in the acceptable region while the true values may be three to four orders of magnitude worse, causing the hardware to be unacceptable in practice. The estimated figures give a calculated value of R of ± 1 in 10^{14}.

The slew time of the input signal to the Q flip-flop in the gate-array is about 1 ns, and the interval between its edges (half period) in a 500 MHz system is 10 ns. The device is clocked at 50 MHz and therefore the estimate

of its failure rate per second is

$$\frac{1\text{ns}}{10\text{ns}} \times 2 \times 10^{-14} \times 50 \times 10^{6} = 10^{-7}$$

which is about 3 times per year.

The I flip-flop is fitted with the temporal hysteresis circuit which ensures it is only violated twice per block slip. It therefore has a lower probability of failure. The block slip rate is equal to the rx-clock tx-clock difference divided by 10. In a 500 MHz system with 50 ppm accuracy, this is 250 Hz. Therefore the I flip-flop failure rate is

$$\frac{1\text{ns}}{10\text{ns}} \times 2 \times 10^{-14} \times 250 = 5 \times 10^{-13}$$

which is negligible.

Evidently both failure rates are acceptable. Performance could have been further increased if the temporal hysteresis circuit were applied to the Q flip-flop, but this was not implemented for power and space reasons.

7.6.6 Elastic Buffer Capacity

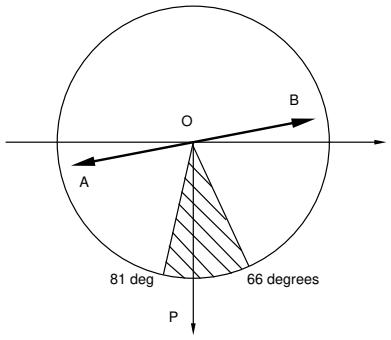

Figure 7.12: Phasor diagram showing the broadside latch violation region as the shaded region. One revolution is a ten-bit block containing two 4B5B symbols.

Although the buffer can insert or delete two syn characters amounting to ten bits when it switches, its capacity to supply or accommodate excess bits resulting from clock inaccuracies is considerably less. The true capacity is apparent from figure 7.12 which represents each block-rate clock cycle as

a revolution around a circle. The outputs from the broadside latch change at the stationary reference phasor OP and data is sampled by the A and B latches with the phasors OA and OB which are 180° out of phase. The shaded area is the violation region. A violation of one of the latches will occur if that latch's clock phasor should enter the shaded area. Phasor AB slowly rotates at the beat frequency. When the active end passes below the line, the elastic buffer logic attempts to switch to the other end which will be above the horizontal line. Since it is constrained to switch only at frame boundaries, if the phasor is rotating quickly enough, it is possible for the active end to pass into the violation region.

The capacity of the buffer is therefore represented by the length of unshaded arc below the horizontal line, either to the left or right of the violation region. The size and position of the shaded region in the figure corresponds to values measured from a CMOS bread-board mock-up of the access chip design. The area for the ECL implementation is similar at 500 MHz and increases in size at higher frequencies. The smaller length of arc on the right is the effective bound on the buffer capacity. Its length is 66° giving a bit capacity of

$$\frac{66}{360} \times 10 = 1.8 \text{ bits}$$

Since the Backbone Ring frame length is 1490 bits, the elastic buffer will accommodate a basic clock rate difference of

$$\frac{1.8}{1490} = 1200 \times 10^{-6}$$

Continuous operation at this clock rate difference is only possible if there is always a suitable syn character when a deleting station wishes to delete.

The theory which relates the elastic buffer capacity to the maximum ring size was developed in chapter 3.

7.6.7 Access Chip Transmit Side Circuit

The access chip transmit side circuit is shown in figure 7.13. The device contains sufficient logic for the bit-level operations required to read or write to any one of the four slots in a Backbone Ring frame and at the same time perform a free-mine operation on another slot in the frame.

The ring data stream from the second channel multiplexer chip enters the access chip eight bits wide on the ZBUS. It is framed in a broadside

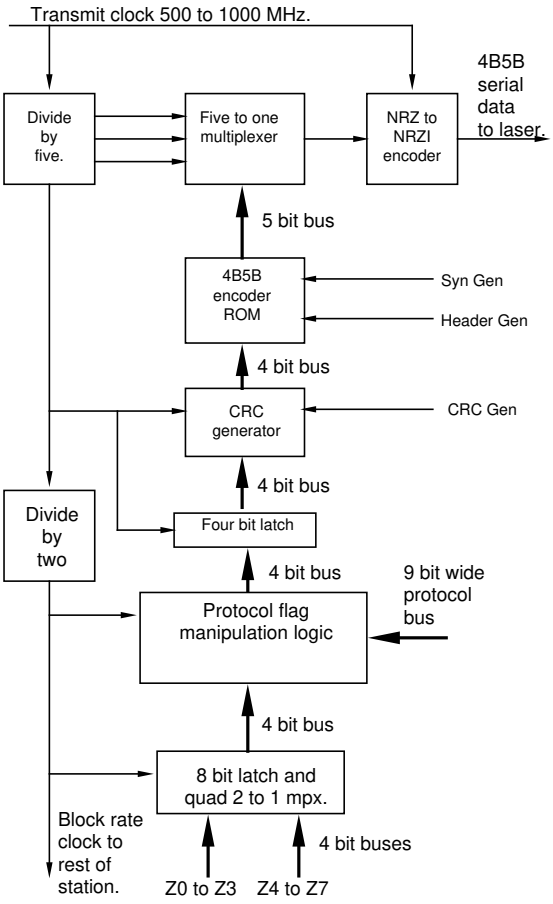

Figure 7.13: Block diagram of the transmit side of the Backbone Ring access chip.

latch and then immediately converted to a four bit wide stream at twice the word rate by a set of two-to-one selectors. The four bit stream enters the bit manipulation logic where the flags are updated if necessary. The data is reframed in a four bit latch and then enters the CRC generator. This calculates a running CRC using an internal four bit wide shifter and exclusive-or array, and then appends the correct CRC to the end of each frame when instructed by the 'crcgen' signal. The data then enters the 4B5B coder which either generates a five bit word for each four bit word that enters or it generates the 'syn' non-data symbol if its 'syngen' input is active. The five bit wide coded data is then converted to bit serial NRZ format using a five-to-one selector. The NRZ stream is encoded to NRZI format using a circuit which is functionally a T-type flip-flop. The NRZI output is fed to a differential output pad for transmission.

7.6.8 Access Chip Current Implementation

Raven, the current implementation of the access chip, is a Plessey ELA63000 series gate array. This is a 1.5 μm silicon, bipolar ECL technology with an F_T of 7 GHz. The macro-cells available to the logic designer can be programmed to select one of three speed/power combinations under a 700 pJ typical speed/power product. For instance, running from a -4.5 volt supply, a high-speed inverter with 1.6 mA tails switches in 123 ps, whereas the low-speed inverter, taking only 0.4 mA, switches in 330 ps. The programmable power technique is particularly applicable for the Raven design, where high-power cells are used for the serial logic and low-power cells are used for the parallel logic. The medium-speed cells are useful for signals which have high loading such as those running from across the chip from one section to another.

The typical power consumption of the ECL logic in the core of the gate array is 3.5 watts. A further 0.5 watts is required for the bias generators. The i/o pads require 0.5 watts each from both the -4.5 volt rail and the TTL 5 volt rail. The total power consumption of the device under worst-case process variation is just under 8 watts. The die temperature must not exceed 165 degrees which means that a heatsink better than 12 degrees per watt is required for an ambient limit of 70 degrees. Such a heatsink for free-air operation has an area of about 10 square inches.

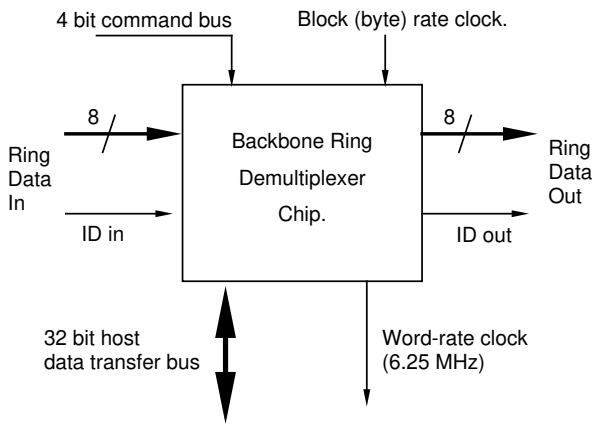

Figure 7.14: Connections to the Backbone Ring demultiplexer chip.

7.7 Channel Multiplexer Chip

Figure 7.14 shows the Backbone Ring multiplexer chip pin connections. Two such devices are used in tandem at each Backbone Ring station. The device accepts an eight bit wide data path from the ECL access chip and is clocked with its block-rate clock input (BLRC). The BLRC signal is generated by the access chip at one tenth the ring clock rate. A ninth bit, called ID, which is also generated by the ECL access chip, acts as a tag to indicate the start of each Backbone Ring frame. Both the data and the ID signal experience 16 BLRC cycles of delay in the access chip. They then reappear on the output side pins, and are fed to the second multiplexer chip or to the ECL access chip for transmission.

The host data transfers occur over the 32 bidirectional connections. These are timed by the word-rate clock output (WRC) generated by the multiplexer chip. This signal is the most significant bit of a 16 bit counter within the device. The counter is clocked from BLRC and synchronised to the start of each Backbone Ring frame by the ID signal. The counter is used internally to distinguish the four Backbone Ring channels.

The multiplexer is controlled by four command input signals to determine the mode of operation: either read, write or idle. The two more significant bits select the mode and the two less significant bits select the channel for read or write transfers.

Figure 7.15 is a block diagram of the basic multiplexer unit. This unit generates 8 bits of the parallel host bus, and is repeated four times within a multiplexer chip to achieve the total 32. The basic function of the unit is to delay the bytewide data by four BLRC cycles, and in the idle control

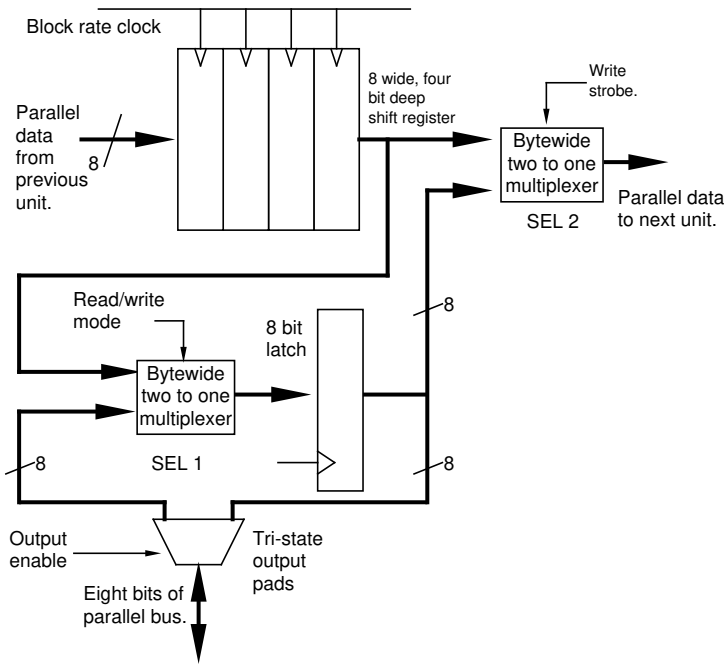

Figure 7.15: One unit of the demultiplexer chip. One chip contains four of these cascaded units and two such chips are used in the basic station.

mode, this is its only function. In read mode, selector one steers the ring data into the eight-bit data register. This is clocked at a phase of the divide by 16 counter according to the selected channel in order to capture a word from the ring. The word appears on the 32 bit external bus when the tri-state outputs are enabled. In write mode, a 32 bit word is first written from the parallel bus, through selector one, and into the data register. At the appropriate phase of the divide by 16 counter, selector two is strobed, causing the existing ring data to be replaced with new data from the data register.

The channel multiplexer chips have been provisionally implemented on a Texas Instruments 2 micron CMOS gate array by Dimitris Lioupis of Olivetti Research. Simulation has indicated that these devices will work up to 50 MHz.

7.7.1 Use of multiple multiplexer devices.

Higher bandwidth stations can be constructed by using the multiplexer devices in different configurations and in greater numbers. One ECL access device at a station remains sufficient for the higher bandwidth configura-

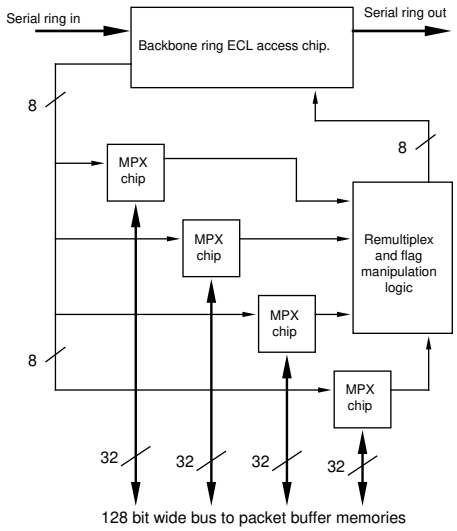

Figure 7.16: One configuration for a fully-equipped Backbone Ring station.

tions. In the dual simplex station, which has twice the bandwidth of the basic station, two deserialiser devices remain sufficient, but now one is dedicated to receive and one is dedicated to transmit. The additional hardware requirement consists of another set of buffer memories and buffer address generators. Access to a greater number of channels is gained by stacking further demultiplexer chips in parallel. The outline of a fully equipped station is shown in figure 7.16.

Inevitably, the ring data finds itself on a 128 bit wide bus, the provision of which was avoided in the simpler stations. Owing to the cost of looking-up the destination addresses of four slots at once, the complexity of the fully equipped station is probably greater than could be achieved without the multi-channel architecture. It is suggested that the additional cost of this configuration is justified by the low entry cost of the basic station configuration.

7.8 Backbone Ring Optical and Dispersion Budgets

The optical transmitter in the 600 MHz Backbone Ring implementation has a launch power of -10 dBm and a spectral width specification of 4 nm. The fibre has less than 0.5 dB per kilometre loss and less than ± 3 ps/nm/km chromatic dispersion. The receiver sensitivity for a bit-error rate of less than

1 in 10^9 is -37 dBm. The power budget is therefore

$$-10 + 37 = 27\text{dB}$$

which, leaving the 7 dB as a margin for connectors and ageing, will permit 40 kilometres of fibre to be used. With this fibre length the dispersion will be

$$40 \times 3 \times 4 = 480 \text{ picoseconds}$$

which is acceptable in a 600 MHz system. With the maximum fibre span, there are about 200,000 photons per bit at the receiver, and the diode converts these to about 3000 electrons.

Chapter 8

Low Complexity Station Interfaces for High Bandwidth Networks

This chapter starts by considering the station interface requirements for the applications listed in table 8.1. From these considerations, low complexity interface architectures for ATM or short mini-packet networks can be determined. Important decisions relate to which functions should be implemented in the station interface and which should be left to the host processor. Many functions can also be implemented in an additional network interface processor (NIP) which fits logically between the network interface and the host machine. Which functions to place in a NIP, what type of processor to

	Application	Network connects to
1	RPC or datagram computer traffic	Processor or multiprocessor backplane or I/O bus.
2	Real-time traffic	Direct connection to multimedia peripherals.
3	File server	Specialised storage peripheral.
4	Low-level bridging	Another network interface.

Table 8.1: Summary of network applications and interfacing requirements.

use as a NIP and the physical structure of the interconnection buses is also discussed. Channel attachment to mainframes is not considered since complexity is then unlikely to be an issue and solutions relying on parallelism exist today. Examples are the Hyperchannel [FRANTA 84], Ultranet [PERDUE 89] and the High Speed Channel [MORRISON 89]. For the Backbone Ring project, it was necessary to design an interface for the initial implementation. A brief overview of this initial implementation is also presented in this chapter.

8.1 Application Specific Interface Considerations and Protocol Components

The architectural design of a network interface depends primarily on the subjective value assigned to processing cycles on the local host. Before considering specific architectures, it is worthwhile describing specific interface requirements that different protocols and applications may require.

For the first application listed in table 8.1, where the network data is generated and consumed by the processor itself, host processor cycles are likely to be highly valued and an interface architecture which places only a low load on the host processor and its cache is desired. On the other hand, in the low-level bridging application, the network and host roles are reversed. Very little value is then attached to a host processor cycle, the main question concerning whether the host can offer sufficient throughput to serve the network. For the other applications, data can be transferred using peripheral to peripheral copies or DMA and the resource most under pressure is likely to be the system backplane bus. For these applications it is worth considering implementations with two network attachments, or alternatively, the use of a multi-ported network interface with the lowest level of demultiplexing being performed in hardware within the interface.

Table 8.2 lists the processor intensive components of low-level protocols which can be conveniently handled outside the main processor, either in hardware or on a separate network interface processor (NIP). All of the listed functions basically require the interface to keep state about each active virtual circuit. Evidently this is hardware intensive and also restricts the choice of data-link and segmentation protocols which the interface can fully support. A suitable architecture for a low complexity interface, yet one

	Transmit Action	Receive Action
1	Buffering	Buffering
2	Fragmentation	Reassembly
3	Checkfield generation	Checkfield checking
4	Encryption	Decryption
5	Multicasting	Hating
6	Retransmission	Flow control

Table 8.2: Protocol components which can be usefully performed by a network interface processor or an intelligent network interface in order to reduce the host processor loading.

which still greatly reduces the host processor loading, is based around a look-up table held in a RAM which is indexed by a fixed field in each cell. A suitable field is generally the port number or virtual path identifier. This is the architecture developed in this chapter. The protocol components are now examined in turn.

Buffering: Buffering is the most important interface function. Buffering is prerequisit for most of the other functions listed in table 8.2 and it relaxes the timing requirements for the host; otherwise the host would have to handle cells at the network rate which could exceed one per microsecond.

Fragmentation and reassembly: Fragmentation of blocks into cells and the reverse process of reassembly is highly processor intensive when there is insufficient buffering in the network interface. The relatively poor performance experienced by several software fragmentation and reassembly implementations for the CFR can be mainly attributed to an insufficient amount of network interface buffering (i.e. one cell's worth) when related to the interrupt response time [GREAVES 89]. For the CFR, cells typically arrive sufficiently quickly to rule out a complete interrupt call and return for each cell service. The consequence is that the host processor wastes time dallying for the next arrival. However, if sufficient intermediate buffering is provided, the host can write all of the cells of a block directly to the network interface without stopping. Thus no interrupts are used for transmitting.

For receiving, it is already necessary to look-up one field of the cell in a hardware table to decide whether the cell should be received at the current station. This can be extended with the help of a small amount of protocol specific hardware assistance so that the host processor is only interrupted

when the last cell of a message is received into the network interface. Upon receipt of the 'end-of-block' interrupt, the host is able to copy all of the buffered cells out of the interface without stopping. This type of reception requires only one interrupt per block, but it may also be useful to take an interrupt at the block start. This idea has been developed for the Backbone Ring project and this implementation, along with a discussion of the missing end-of-block error case, is described in section 8.4.

Checkfield generation: Checkfield generation, such as network level CRC or end-around-carry checksums can be generated and checked in the station interface. Checkfield generation in hardware is not very difficult, but verifying the checkfield of received cells is complicated by possible multiplexed reception of several simultaneous blocks and by possible out-of-order cells within a single block. This implies checkfields can only be properly checked after block reassembly. A worthwhile compromise for a minimal hardware implementation may be to provide a single checkfield generation and verifying unit. This operates during transmission in the usual way. On reception, it checks any group of consecutive (adjacently received) cells delimited by start-of-block and end-of-block. In the case of those allowable error cases just mentioned, the hardware check will prove false, and in this case, the check must then be re-evaluated using a software implementation after correct reassembly.

Cryptography: Encryption and decryption are notorious for their intensive processing requirements. For hardware implementation within the network interface, suitable techniques are applicable at the cell level without knowledge of the fragmentation context. For instance, the American DES (data encryption standard) operates on units of 64 bits. In practice, a full encryption system may not be needed for many applications. Encryption of just the checkfield may be sufficient to prevent message level forging by foreign agents. Again this can be performed easily in the network interface if the necessary key is stored in a look-up table indexed by virtual circuit identifier. Indeed, the use of virtual circuit identifiers instead of hardware addressing fields at the physical layer may intrinsicly provide sufficient data security, provided that the messages used for virtual circuit set-up are themselves encrypted at a higher level.

Flow control: In many computer networks, there is a software flow control mechanism implemented at the equivalent of the transport level. Flow control mechanisms implemented fully in hardware are difficult to make reli-

	Bus structure	NIP type	Copying control
1	Single bus	No NIP	Host processor copies data
2	Single bus	DMA style NIP	Host processor requests transfers
3	Split bus	Autonomous NIP	NIP copies data

Table 8.3: Architectural variations available using a station interface which connects to a shared backplane bus.

able. On the other hand, end-to-end hardware flow control assistance mechanisms which, although not fully reliable, reduce the loading on the software implementation, are feasible. These may be implemented relatively easily when the medium access protocol can support low-level acknowledgements. A useful facility for the interface to provide is a hardware count and limit mechanism, programmable on a per virtual circuit basis, which informs the source station when the number of cells stored at the receiving station on a virtual circuit exceeds the limit. This information is fielded back to the source station, typically generating a special interrupt status.

Others: Low-level multicasting and retransmission make use of the buffering within the network interface. In a network with a partitioned physical layer, multicasting requires the transmitter to send the same cell on all appropriate channels. Additional virtual circuit identifier attribute bits can specify whether cells on that particular virtual circuit require multi-casting, on which channels to send the cell and whether to retransmit on error. Refusing to receive from certain sources, termed 'hating', is performed by the receive attribute bit.

8.2 Architecture

Computer backplane bus standards are offering ever increasing bandwidths. The P960 proposal known as 'Fastbus' offers 1.2 Gbit/second bandwidth, although this is reduced when arbitration overheads are accounted [EE 86]. A network interface which connects to a standard bus enables systems with and without a NIP to be constructed. There are three basic architectures for connection to a single host, as summarised in table 8.3. In the first and most simple system, the host processor sits on the same bus as the network interface. In the second architecture, the NIP sits on the same bus as the main host processor and acts as a very intelligent DMA controller.

In this configuration there are two possible types of NIP. The NIP may either consist of a general purpose microprocessor which takes charge of the network and protocol management as well as its main job of moving data around, or else the NIP consists of a specialised microcoded machine which operates under direct control of the host. The third architecture employs a fully autonomous NIP which handles all network related functions. This sits on a separate network interface bus along with the network interface itself. Typically, dual-ported RAM is provided to which access is available from both the interface bus and the main processor bus. Where appropriate, in order to reduce cache pollution, programmed data transfers across the backplane bus should use instructions which do not update the processor cache.

In all of the architectures in principal, data need only be transferred once across each bus. In an RPC environment, it is desirable for the marshalling software to send the marshalled data to the network as quickly as possible. Hence an intermediate kernel copying process should be avoided. This requires that code executing in user space should have direct access to the network interface, and since user code is multi-threaded, some degree of exclusion needs to be provided. Otherwise various parts of cells may become mixed up, depending on the atomicity of the write instruction used. Using the dual-ported RAM described in architecture three and used, for example, in the DEC firefly machine, exclusion is implemented by storage allocation within the dual-ported RAM This meets the goal of data traversing each bus only once.

Within the other architectures, data is transferred either by DMA from main memory into the interface, or stored directly by the host processor into the interface at the time the data is first generated. Exclusion is required at the cell level. Two exclusion techniques are apparent. They can be used independently or in combination. The first relies on the virtual address translation system to only map the interface into one process' address space at a time. Page faults are used to reallocate the interface. The second technique provides the interface with several write ports which appear at different addresses in the backplane address space. Inside the interface, each port can contain a partially constructed cell. Each time a complete cell has been accumulated, the port contents are transmitted and then the port becomes free and may be reallocated. The ports can be assigned to separate threads or processes in the host. A disadvantage is that a sufficient number of ports must be provided so that their state does not require saving and

restoring on a context swap.

Providing this sort of space division exclusion at the interface hardware level is evidently quite complex. Since these simpler architectures are aiming for low hardware complexity, the alternative time division exclusion technique of storing all of the cell into the interface in one atomic action is preferable. Using program controlled DMA, this is achieved by the DMA device only making one cell copy at a time. With host processor generated transfer cycles,[1] either a complete cell of data should be stored into the interface using an uninterruptable instruction, or else a level of exclusion using semaphores is appropriate.

8.3 Protocols

As stated, hardware implementation of the lower levels of protocol necessarily restricts the protocols which can be supported by the interface. New proposals for protocols operating up to the transport layer are being advocated, the main ones being VMTP [CHERITON 86], NETBLT [CLARK 86] and XTP [XTP 89]. These protocols have been designed for high performance and hardware implementation, but unfortunately they are all oriented for networks where the physical layer can support relatively large packets. For instance, VMTP runs over an unreliable datagram service and, as such, can be run on an ethernet without an intermediate fragmentation and reassembly protocol. However, intermediate data-link layer fragmentation is required in order to run over a cell network such as a slotted ring or other ATM network with real-time capabilities.

The protocols used with the Cambridge Fast Ring are designed for cell type networks. For high performance, these protocols require hardware support in the network interface. The original CFR data-link layer protocol is Unison Data Link (UDL), developed in Cambridge for project Unison [TENNENHOUSE 86]. UDL provides an unreliable datagram service. It incorporates the segmentation and reassembly mechanisms and it also provides virtual circuit demultiplexing within a host machine. These functions have been split into separate (conceptual) layers by D McAuley for the Multi-Service (MS) series of protocols [MAC 89]. The MS protocols aug-

[1] These include normal reads and writes and also transfers where the processor only generates the address on the backplane, data being sourced and sinked by other devices.

Picture.

Figure 8.1: Old and new cell formats. UDL uses three sixteen bit fields to route the cell to its destination process. MSDL requires only one.

Picture mssar.

Figure 8.2: Multi-service segmenting and reassembly (MSSAR) header format. The format in UDL is very similar.

ment the services provided by UDL to form an architecture suitable for local and wide area inter-networking over heterogeneous networks. Both MS and UDL must be supported by the Backbone Ring interface hardware.

Figure 8.1[2] shows that in UDL, the first 32 bit word of the nine available in a standard cell contains conventional hardware MAC addresses, while in the MS data-link protocol (MSDL), these fields are condensed along with the port number into a single, 16 bit virtual path identifier (VPI). As described in [MAC 89], these identifiers are unique to an addressing domain which could typically be a single ring or the input port to a self-routing switch. The VPIs are translated at address domain boundaries by direct look-up in RAMs.

In both UDL and MS, the next two bytes after the addressing information can be used by an optional segmenting and reassembly protocol. This is shown for the MS architecture in figure 8.2. These fields are not present when non-block structured data is being transmitted, such as voice. The two bytes can then be used for data or other purposes: for example, voice requires a time-stamp. In both protocols, there are three reassembly specific fields which occupy a total of 16 bits. These are the block number, the sequence number and a start bit. The block identifier is incremented each time a new block is to be fragmented. It prevents bad reassembly when the end of one block is missing as well as the start of the next. The first cell of a block is transmitted with the start bit set and then the sequence number field represents the number of cells to expect. At the transmitter, the sequence number is decremented for each cell and the final cell of a block then has sequence number of one. These protocols are supported by the initial Backbone Ring network interface.

[2]This figure and the next one are adapted from [MAC 89].

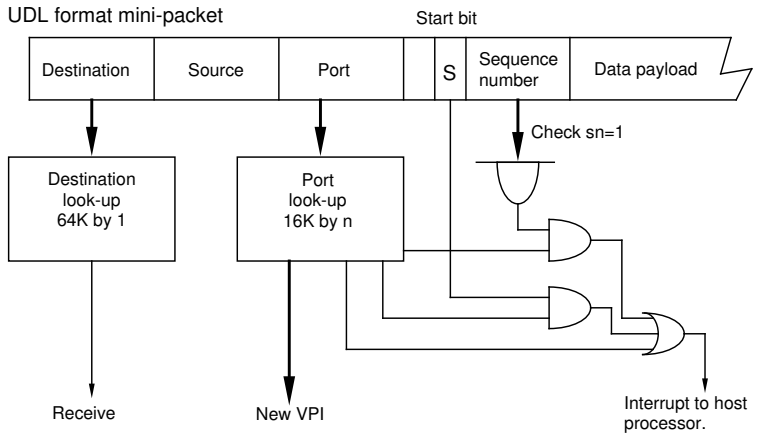

Figure 8.3: Look-up table configuration for assistance with generating interrupts with the UDL protocol. Using MSDL, only one RAM is required.

8.4 Initial Backbone Ring Interface

The initial Backbone Ring stations are of the half-duplex, static receiver type. An architecture using a single-ported interface to a shared backplane bus was selected since this would initially be the most useful. The peak throughput of the station at the 1 GHz ring rate is 88 Mbit/second half-duplex, which is roughly the same as that of a 32-bit VME bus. Accordingly, the VME bus was selected for the initial station interfaces. The VME bus is widely used and well understood and was already supported in the CFR research environment.

The main functions supported by these stations are summarised in the following list:

- Buffering provision for a total of 2048 cells.
- Management of linked lists to maintain the buffers in FIFO queues.
- On the transmit side, two priority levels with a separate pair of queues for each channel.
- On the receive side, UDL and MS specific interrupt generating logic.
- Retransmission of certain types of cells if reception has failed owing to a transmission error.
- Response field notification to the transmitter when the receive buffer is becoming full.

The station interface deliberately does not implement all of the useful functions that were listed in table 8.2. In particular, it does not perform the complete fragmentation and reassembly protocol; it only provides assistance. These omissions are partly for simplicity, but partly as an extension to the Backbone Ring concept of stations of varying complexity. The missing functions are those which can be provided by a specialised DMA controller or implemented in the host processor without too much overhead. The three types of bus architecture listed in table 8.3 can be supported, as is shortly described.

Figure 8.3 shows the protocol assistance hardware implemented in the receive side of the initial Backbone Ring station. This hardware can be configured for either the UDL or MSDL protocols. In UDL, two look-up RAMs are required, one for the deciding whether to receive a cell or not and one to decide whether to interrupt the processor or not. These are driven from the destination and port fields respectively. In MSDL, there is only the VPI to look-up and both RAMs are fed from this field. According to the look-up RAM entry, a cell can interrupt the processor if it is the first in a block, the last in a block, unconditionally or not at all.

Bus architecture 1: A system containing only a VME processor board and a Backbone Ring network interface requires the processor to copy all of the network data to and from the network interface, the unit of transfer being a 32 bit word. The processor typically employs block move operations for this. For transmission, the processor generates the protocol header, stores this in the interface, then copies sufficient data to fill up a cell. It can repeat this in a tight loop. No software handshaking is required, hardware synchronisation being achieved using the DTACK (data transfer acknowledge) signal of the backplane bus. The processor is interrupted by the interface only if it manages to use up all of the available buffer storage. This normally only occurs when the network is very heavily loaded, in which case, high throughput could not in any case be achieved since the network bandwidth must be shared.

For receive, generally the processor will be interrupted only when the end of block flag is encountered. It then reads cells from the station in the order they were received. The cells found in the station may include mid-block sections of blocks other than the one for which the interrupt was generated. In this case, the processor stores the extra cells in their correct places in the appropriate reassembly buffers until required.

The processor will not be interrupted if it receives a block with the last cell missing. This is in keeping with the UDL specification where the UDL layer discards any blocks which cannot be reassembled correctly. The station provides low-level retransmission below the UDL layer and the higher level protocols also inevitably perform block-level retransmission on failure.

Bus transfer time for the data transfers can be halved if the processor uses DMA mode bus cycles which have a unique address modifier code over the VME bus. This mode is being included in the prototype station interfaces. For transmission, the processor generates the address in RAM of a word to be written to the interface, the RAM delivers the data onto the backplane bus and the interface detects the address modifier and stores the data internally. For reception, a similar process applies. In this way, words containing protocol headers are transferred directly from the interface to the processor data registers and data is transferred directly into RAM (or wherever needed).

Bus architecture 2: General purpose DMA controllers are not sufficiently flexible to perform protocol header generation, checkfield generation and verifying, or multiplexed reassembly of blocks. One type of NIP is a specialised DMA controller which can perform these functions. An attractive aspect of this configuration is that the NIP is a separate entity from the network interface and can be designed and installed independently. Again data need only be transferred once across the backplane bus. The NIP may be dual-ported, having an additional, dedicated connection to a specialised storage or multimedia peripheral. This reduces the loading of the host processor, but not that of the host processor bus.

Bus architecture 3: A split bus can also be used with the initial station interfaces. This offers the most flexible and high performance solution. It is envisaged that the interface bus will connect, in addition, to video frame stores and capture peripherals.

8.5 Summary

This chapter has taken a very brief look at the interfacing requirements for high bandwidth networks. Low complexity solutions for connection to multi-media peripherals and mini-computer backplanes have been considered. There are many possible configurations and several of the most effi-

cient have been presented. The design of the initial Backbone Ring station interfaces has been described. Unfortunately, no performance measurements or estimates are currently available.

Since the interface side of the Backbone Ring station has been designed using discrete logic, in the future it will be possible to add new functions relatively easily. This is an interesting area for ongoing research. When the required functionality has been established, it is envisaged that a third semi-custom integrated circuit will be constructed for the station control logic.

Chapter 9

Results and Conclusions

9.1 Summary

This dissertation has considered the behaviour of multi-access networks when

- the transmission rate is increased to 1 GHz and above and
- the length of the shared medium is increased to around 200 kilometres.

At these geometries there can be several thousand packets stored in the network cables at any time. This is the main difference between multi-access MANs and smaller geometry local-area networks or self-routing switches. Messages on the large geometry MAN are far less closely coupled to the station buffers and consequently there is a far greater commitment once a packet has been launched onto the medium. When there are fewer packets stored, as in the more centralised LANs and switches, high-priority traffic which arrives without warning can be granted its bandwidth much more quickly.

Chapters 2 and 3 have reviewed the technology for implementing metropolitan area, multi-access networks and shown that baseband transmission at gigabit per second rates is possible. Chapter 4 presented multi-access MACs which make efficient use of the bandwidth. Multi-channel architectures which offer even higher bandwidths have been described. The number of

packets stored in transit on the network cables remains proportional to the aggregate bandwidth, but if necessary, the number per channel can be reduced by sub-partitioning the channels.

The amount of traffic applied to the network must be controlled, otherwise packet loss occurs and delays are increased. This is especially true for real-time traffic. For systems with two or more levels of priority, it is desirable that the regulation of low-priority traffic can largely be performed by the network itself, using its MAC rules. The MAC layer approach is more simple to implement than a software equivalent and it can generally respond more quickly, thus reducing bandwidth wastage.

The support of priority in the MAC layer is very desirable for delay-sensitive traffic, such as the traffic generated by real-time sources. Priority ensures that the delay-sensitive traffic is provided with sufficient bandwidth and enables the remaining bandwidth of the network to be loaded with an arbitrary amount of low-priority traffic. The low-priority traffic can be bursty and may occasionally overload the network, but the priority mechanism prevents it interfering with the delay-sensitive class. Provided sufficient buffering is provided, the bursty, low-priority traffic will also be serviced, but with greater delay.

Alternatively, delay-sensitive traffic can be supported without MAC layer priority, provided that there is an effective load balancing mechanism. Expedited transfer is used within a station to reduce the delay for the priority traffic. This method is intrinsically limited to cases where the proportion of delay-sensitive traffic generated by an individual station does not exceed the reciprocal of the number of stations.

The granularity of sharing supported by a network affects the delay for all classes of traffic. The minimum delay for an expedited packet is equal to the sum of the mean residual life of a sharing cycle and the mean transit time over the network. In a well designed system of large geometry, the sharing cycle duration should be comparable to the latency. Hence both components of the delay for an expedited packet can each be half a network latency, giving an overall average delay of one network latency.

For both methods of supporting priority, the granularity with which fairness is amortised affects the delay experienced by bursty, high-priority traffic. This has been termed the granularity of load balancing. In a perfect system, this is also limited to a theoretical minimum of one latency.

Packet size was discussed in section 4.4. A small packet or cell was advocated, although not for MAC efficiency reasons since conventional packets of several thousand bytes look small at large geometries. Good protocol efficiency can be obtained from a small packet. Relative addressing (section 1.4.1) should be used for long distance communications.

Access control protocols (MACs) have been reviewed. For large geometry networks, slotted systems have good efficiency and low delay. Their efficiency does not depend on network size or physical packet size. The slotted system was also found to be most suitable for multi-channel systems where it is desired for a station to switch from one channel to another. The slot boundaries make a natural place to switch and can also provide the distributed time frame required for predictable receiver assignment (section 4.6.3).

MAC layer protocols for rings were considered in chapter 5 and a simple, yet useful expression for the delay on a slotted ring was given. Simulation showed that the load-balancing resulting from source-release of full slots does not reduce the delay for expedited traffic, even when there are very bursty sources also using the network. Destination-release resulted in lower delays owing to the greater bandwidth available. However, the effects of not having load balancing were shown to be undesirable at high loading. The double-slotted ring protocol was proposed as a solution which combines the benefits of source and destination-freeing rings. The protocol can be used either with or without the support of full MAC layer priority and simulation was used to show that the protocol performs very well.

9.2 Multi-Media MAN Results

Chapter 1 introduced and defined the type A multi-media user. Chapter 6 described the Backbone Ring project architecture. Simulation results for the delay and jitter experienced by the multi-media traffic over a Backbone Ring are now presented. Two different sized rings are simulated and the results are compared with five other ring MAC protocols. The physical parameters for the two different sizes of ring are listed in table 9.1. The results for both sizes of network when loaded with 72 of the type A users are listed in table 9.2. The Backbone Ring bandwidth is utilised to about 63 percent by this number of users. For the six different types of ring, the traffic delay

Size	Medium	Large	
Frames	100	1000	
Slots	400	4000	
Frame time	1.2	1.2	microseconds
Latency	128	1280	microseconds
Fibre length	25.6	256	kilometres
Data rate	800	800	megabit/second
Stations	18	18	
Addressing overhead	0	0	bits

Table 9.1: Physical parameters defining the two different size rings used for simulation in this chapter.

experienced in each priority level is listed. The propagation time across the ring has been subtracted in order to more easily compare the different size networks. For the synchronous traffic, the 99^{th} percentile of variation of inter-arrival time is also listed in the columns labelled 'jitter'.

For all of the slotted rings, the delay results are very good, but the token ring shows much greater delays, especially at the larger geometry. The first line of results is for the basic, half-duplex Backbone Ring stations described in chapter 7. These stations cannot transmit into a frame if they received from the preceding frame. However, this does not increase the delays very much, as can be seen by comparing the results with those of the second line which is for stations which do not suffer from this effect, but are otherwise identical.

Results for the enhanced Backbone Ring architecture, where station receiver assignment follows a pseudo-random rule, are presented in the third line of results. The enhanced architecture does not significantly reduce the mean queueing delay of the traffic, but it does approximately halve the jitter. Further examination of the full simulation results shows that this may be entirely attributed to shorter average length of cluster of consecutive full slots under the enhanced architecture. The average cluster length with static receiver assignment was 18 slots whereas with predictable assignment by the pseudo-random rule it was 9 slots. These figures apply to both sizes of ring; the number of clusters varies according to the ring size, not the length of individual clusters.

The fourth line of results shows the performance of a simple source-release slotted ring without channels. This yields four times lower average

	Priority level 4			Priority level 3			Level 2		Level 1
Throughput (Mbit/s)	9.5			424			88		2
Access protocol	Mean delay	99tile delay	99tile jitter	Mean delay	99tile delay	99tile jitter	Mean delay	99tile delay	Mean delay
Backbone Ring	7	40	83	7	37	73	20	67	243
with pipeline blocking	10	57	117	9	52	112	26	94	596
Backbone Ring	6	34	59	7	34	64	18	62	273
without pipeline blocking	8	48	97	8	51	112	22	53	501
Enhanced Backbone Ring	5	23	39	5	25	49	16	47	243
without pipeline blocking	6	32	64	6	31	64	19	56	393
Source-release slotted	1	7	15	1	9	20	4	16	40
ring without channels	1	9	20	1	8	20	4	16	48
Double-slotted ring	0	1	5	1	4	5	2	6	22
protocol	0	3	5	1	4	10	2	6	23
Early-release token ring	1574	3050	3152	1513	2972	3045	1368	3050	1720
with exhaustive service	91	156	313	91	156	205	92	156	141

Table 9.2: Delay and jitter performance in microseconds for six ring protocols when loaded with 72 type A multi-media users. The upper figure of each result is for 256 kilometres of fibre and the lower is for 25.6 kilometres.

delay than the systems with four channels. This is a consequence of Little's result. The double-slotted ring shows still lower delays as a result of its reduced full slot density.

Figure 9.1 plots the 99^{th} percentile of inter-arrival jitter for the type A tollvox traffic on the larger geometry Backbone Ring as the number of users is increased. As the number of users is increased above 100, typically one of the channels saturates and one quarter of the population starts experiencing packet loss in their low-priority traffic. The graph shows that the mean delay for the voice traffic is not greatly disrupted by this, but its jitter starts to increase. However, the jitter remains below 2 milliseconds which is much smaller than the worst case of 24 millisecond evaluated shortly. When the voice streams are taken over several such networks in tandem, the jitter will grow not much faster than in proportion to the square-root of the number of networks, and so a simple double-buffer arrangement for reassembly would suffice.

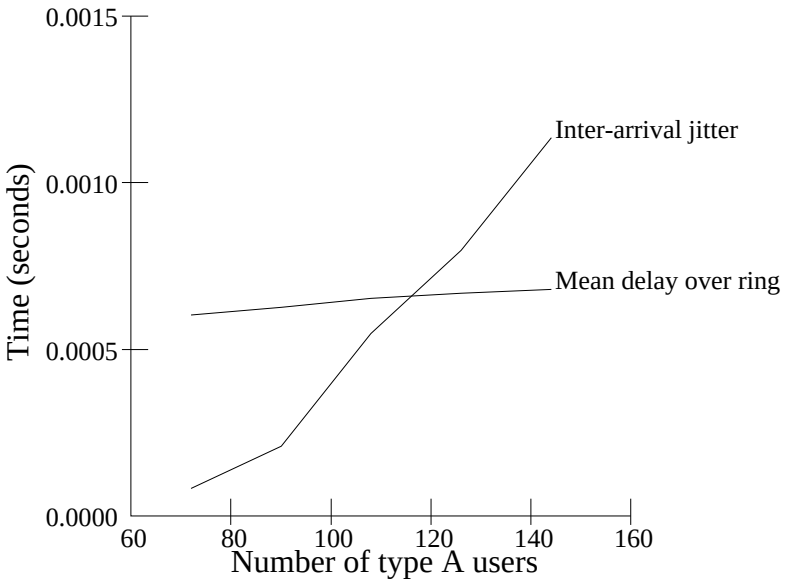

Figure 9.1: Jitter and delay experienced by the priority level 4 traffic over the larger geometry, blocking Backbone Ring as the number of users is taken well above saturation.

Size	Medium	Large	
Maximum slot time	2.4	24	milliseconds
Average slot time	0.023	0.023	milliseconds
Ratio	100	1000	

Table 9.3: Maximum and average times for an empty slot on a Backbone Ring channel.

9.2.1 Guaranteed Performance.

For a source-release slotted ring, such as the Backbone Ring, the maximum time before an empty slot can be guaranteed to arrive at a station is $N + 1$ latencies. The guaranteed minimum bandwidth to a station is $1/(N + 1)$ of the total bandwidth, so the average time before an empty slot cannot exceed $N + 1$ slot times. When there are four channels, the average time for an empty slot on each channel is $4 \times (N + 1)$ slot times, or $N + 1$ frame times. For the two sizes of ring, these figures have been evaluated and presented in table 9.3.

It is a characteristic of the source-release slotted ring protocols that the ratio between these two figures grows as the geometry is increased. The ratio is equal to the number of frames. The limit on maximum delay guaranteed by source-release of full slots is clearly not sufficient on its own to enable short reassembly buffers of the type described in chapter 1 to be relied upon.

In addition, the fixed $1/(N+1)$ fraction of guaranteed bandwidth does not form a propper MAC layer priority mechanism and it is insufficient should an individual station wish to send more priority traffic than this.

The simulation results for the type A user, even when the ring is saturated, show that the real-time response is very good. The slot occupation patterns which cause token-like behaviour do not seem to occur with realistic traffic. However, specially contrived traffic patterns have been simulated which have caused the inter-arrival jitter to approach the latency. These are not otherwise reported in this dissertation, but their existence may prompt further investigation in the future. Two methods of reducing the jitter and restricting the delay for real-time traffic have been suggested:

1. The use of dither, either random or deterministic, to break up regular patterns of slot filling,

2. Use of the full DSR protocol to give reliable priority.

The first reduces the overall network bandwidth whereas the second increases it. The DSR protocol is able to respect an imbalance in the high priority traffic generated between each station and is also able to handle bursty, high priority traffic.

9.3 Conclusion

A multi-channel, multi-access network is ideal for providing high-bandwidth interconnection, with low complexity, between about 20 and 100 service points. The service points can be distributed over an area of tens of kilometers and can either consist of end user stations or bridges onto more localised, multi-access networks. A multi-channel system naturally lends itself to sharing of the network bandwidth and can result in hardware reductions at the service point. Low hardware cost at the end user station is required because of the desire to integrate many multi-media information services on to a single fibre and distribute the services via the fibre between many small and inexpensive workstations. The work of this dissertation shows that multi-access, gigabit per second, networks could then become as ubiquitous and economical as today's computer-oriented local-area networks.

A multi-channel multi-access network can support	A self-routing, fast packet switch can support
Short, fixed-length packets (ATM). Fine grain sharing when slotted. Multi-level priority. Routing and attributes in packet header. Sharing re-evaluated once per latency. LAN and MAN areas. Stations of various bandwidths.	Short, fixed-length packets (ATM). Fine grain sharing intrinsically. Multi-level priority. Routing and attributes in packet header. Sharing continuously re-evaluated. Arbitrarily large areas. Switches with various numbers of ports. Ports of various bandwidths. Varying number of inter-switch connections.
Low-complexity, low-cost.	More complex minimal configuration.

Table 9.4: Comparison of multi-access and switched network capabilities.

The bandwidth requirement for multi-media traffic is continuously increasing. High-definition television standards are emerging. An HDTV frame of two mega-pixels, 24 bits deep, transmitted a modest 25 times per second yields 1.2 gigabit per second. Owing to the high data rates, and the desire for low-complexity workstations, it is likely that only relatively simple compression algorithms will be employed, perhaps achieving a compression factor of 10. This will result in bursty streams of around 100 megabit per second, and these are ideally suited for transmission over statistical, packet switched networks. Ten channels supported by a relatively simple, 1 gigabit per second, multi-access network would appear most felicitous in terms of cost, performance and complexity.

Multi-media traffic requires distribution and switching with low delay in the wide area as well as the local and metropolitan areas. Both multi-access networks and self-routing switching fabrics are ideal for such statistically-based packet switching. A comparison of their properties is presented in table 9.4. They share many properties in common, particularly the ability to carry short, fixed-length mini-packets or cells in the style of ATM (Asynchronous Transfer Mode).

The multi-access networks offer low complexity and are ideal for local and metropolitan area distribution amongst a well defined group of users. The switched networks are more flexible since completely arbitrary topologies are possible and the amount of bandwidth between switching points is easily programmed. Direct connection between ATM switches and slotted multi-access networks such as DQDB and the Backbone Ring is not seen as a

problem. Very similar or identical network level protocols can be run over both types of system; there is no reason why higher level protocols should not find them indistinguishable.

A feature of multi-access networks is that stations have passing though them a large amount of data that is destined elsewhere. This poses security and reliability problems since any station has the potential to read and manipulate traffic belonging to other stations. A multi-access station will be designed, as far as is possible, to fail in a way which does not upset the operation of the network as a whole. However, each user will always be dependent on the integrity of the others, at least to some extent. For these reasons, multi-access networks might be considered of limited appeal to a public network company if it wishes to serve several different organisations from the same network. The network station must be placed in a secure area where only the bearer company's staff have access. Placing the station in a roadside cabinet has been suggested. On the other hand, a private connection to the bearer's nearest switch only need carry traffic destined for the one private user. For these reasons, it is difficult to envisage a multi-access metropolitan area network serving a hetrogeneous user community.

Conversely, a multi-access network is ideal when shared among a group of private users. The main advantages being the low entry cost owing to the lack of a central switch and the ease of expansion which this gives. Multi-access techniques are well proven for office and departmental LANs and, from a management point of view, there is no objection to extending this type of multi-access to cover greater areas.

Figure 9.2: Photo of the main three boards of the VME-format ring station, as described in Figure 7.3: from left to right, low-speed board, high-speed board and optical assembly.

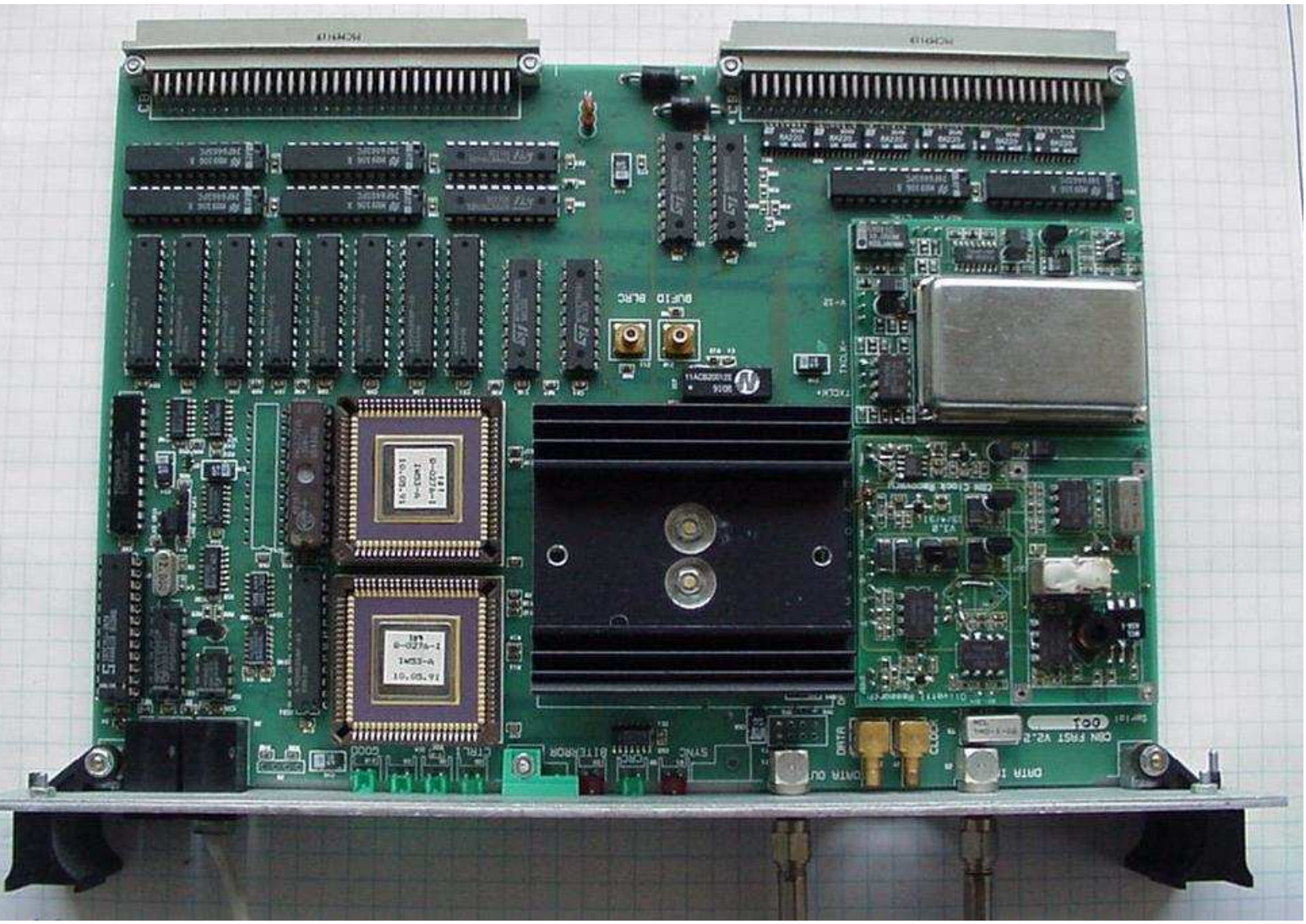

Figure 9.3: The high-speed board containing the access chip (under heatsink, centre), access chips, RAM and receive and transmit clock daughter boards (with shielding can removed from the receive clock daughter board).

References

[**ADAMS 87**] 'The Orwell Torus communications switch.' JL Adams. BTRL. Proceedings CEPT seminar on broadband switching. Albufeira, Portugal. January 1987.

[**ALBANESE 88**] 'Overview of the Bellcore Metrocore network.' A Albanese, MW Garrett, A Ippoliti, MA Karr, M Maszczak and D Shia. Bellcore. Proceedings IFIP WG 6.4 workshop. Liege, Belgium, April 1988.

[**AMINETZAH 85**] 'Exact results for token ring systems.' YJ Aminetzah and MJ Ferguson. IEEE Transactions on communications, Vol COM-33 number 3, pp 223, March 85.

[**AS 91**] 'CRMA-II: A Gbit/s MAC protocol for ring and bus networks with immediate access capability.' HR van As, WW Lemmppenau, P Zafiropulo, EA Zurfluh. IBM Zurich. Ninth EFOC LAN conference, London UK, June 19-21, 1991.

[**BENNETT 58**] 'Statistics of regenerative digital transmission.' WR Bennett. BSTJ pp 1501-1542. November 1958.

[**BERGMAN 85**] 'A synchronous fibre optic ring local area network for multi-gigabit/s mixed traffic communication.' LA Bergman and ST Eng, BNR. IEEE JSAC Vol SAC-3 number 6 November 1985.

[**BOEHM 85**] 'SONET (Synchronous optical network).' RJ Boehm, YC Ching & RC Sherman. Proceedings Globecom 1985, pp 1443-1450.

[**BOSTON 88**] 'FDDI-II: A high speed integrated service LAN.' T Boston. Proceedings EFOC/LAN 88. Amsterdam June 1988.

[**BROOKS 83**] 'Line coding for optical fibres systems'. RM Brooks and A Jessop, BT and STL. International J of Electronics, Vol 55 number 1 pp 81-120. 1983.

[**BTD 88**] 'SOA1100 semiconductor optical amplifier data sheet.' British Telecom and Dupont Technologies (BTD), December 1988.

[**BURR 88**] 'An overview of FDDI.' WE Burr and L Zuqiu, National Bureau of Standards, Gaitherburg, MA USA. Proceedings EFOC/LAN 1988 pp 287-293.

[**BUDRIKIS 86**] 'QPSX: A queue packet and synchronous circuit exchange.' ZL Budrikis, JL Hullett, RM Newman, FM Fozdar and RD Jeffery. Proceedings of ICCC 1986 pp 288. North-Holland 1986.

[**BYRNE 63**] 'Systematic jitter in a chain of digital repeaters.' CJ Byrne, BJ Karafin and DB Robinson Jr BSTJ Vol 41 pp 2678-2714. November 63.

[**CANDY 71**] 'Transmitting television as clusters of frame-to-frame differences.' Candy, Frank, Haskell and Mounts. BSTJ Vol 50, pp 1889-1917. July 1971.

[**CAPETANAKIS 79**] 'Generalised TDMA: The multi-accessing tree protocol.' JI Capetanakis. IEEE Transactions on COmmunications, COM-27 number 10, pp 1476, October 1979.

[**CARIOLARO 74**] 'The spectra of block coded digital signals.' GL Cariolaro and GP Tronca. IEEE Transactions on Communications, COM-22, number 10 pp 1555-1563, October 1974

[**CASH 84**] 'Ring local area networks - A strategy for choosing the phase-locked loop class for data repeaters.' AR Cash. Rutherford Appleton Laboratory report RAL-84-081. August 1984.

[**CCITT 89**] 'CCITT (draft) series I.121 recommendations for broadband ISDN.' CCITT 1989.

[**CHAMBERS 86**] 'CFR M-Access service definition.' AM Chambers. Project Unison working paper UA004. Acorn Computers Ltd. Cambridge, October 1986.

[**CHERITON 86**] 'VMTP: A transport protocol for the next generation of computer systems.' DR Cheriton. Proceedings ACM SIGCOMM 86 Stowe Vt Aug 5-7. Also ARPA report RFC 1045.

[**CHLAMTAC 88**] 'A multibus train (AMTRAC) architecture for high speed fibre optic networks.' I Chlamtac and A Ganz. IEEE JSAC Vol SAC-6 number 6 July 1988.

[**CIDON 90**] 'Metaring - A full-duplex ring with fairness and spatial reuse.' I Cidon and Y Ofek, In IEEE Infocom 1990, San Francisco.

[**CLARK 86**] 'NETBLT: A bulk data transfer protocol.' DD Clark, MIT.

ARPA report RFC 969 December 1985.

[**COCHRANE 83**] 'Integrated regenerators for high speed optical transmission systems.' P Cochrane, DW Faulkner, L Bickers, I Hawker and RJ Hawkins. BT Technology J Vol 1 number 1 pp 66-72 July 1983.

[**CONTI 89**] 'DQDB media access control: Performance evaluation and unfairness analysis.' M Conti, E Gregori and L Lenzini. Procceedings Third IEEE conference on MANs. San Diego 1989.

[**CURRY 88**] 'Oscilloquartz D-TCXOs - the case for defence.' C Curry, Chronous Technology. Electronic Engineering, Vol 60 number 743, November 1988.

[**DANTHINE 85**] 'Broad site local wideband network.' A Danthine and E Vyncke, University of Liege, Belgium. In Esprit 84: Status of ongoing work. Ed Roukens and Renuart. North-Holland 1985.

[**DAVIES 87**] 'Synchronisation schemes for a high speed optical fibre TDM ring network.' PA Davies and A Jamasebi-Jahromi. Proceedings IFIP working group 6.4 workshop Aachen February 1987.

[**EE 86**] 'Bus-based boards.' Electronic Engineering, December 1986, Page 54.

[**FALCONER 85a**] 'A simulation of the Cambridge Ring with voice traffic.' RM Falconer, JL Adams and GM Walley. BT Technology J. Vol 3 no 4. October 1985.

[**FALCONER 85b**] 'Orwell: a protocol for an integrated services local network.' RM Falconer and JL Adams. BT Technology J 3 (4) pp 27-35. 1985.

[**FELLER**] 'An introduction to probability theory and its applications.' Third edition. W Feller. Vol 1, chapter 3, section 7.

[**FILIPIAK 89**] 'Access and priority control in distributed queueing.' J Filipiak, University of Adelaide. Proceedings ICC 1989, Boston, June 1989.

[**FIORETTI 87**] 'A new design for multiservice integration over a high speed fibre optic LAN based technology.' A Fioretti, PJ Wilkinson, CA Rocchini and AJ Haylett. Proceedings IFIP WG 6.4 Workshop on High Speed LANs. Aachen February 1987.

[**FIORETTI 88a**] 'Fibre optic LAN/MAN architectures for coherent optical transmission.' A Fioretti, CA Rocchini, SR Treves and L Torchin, Alcatel FACE, Pomezia, Italy. Electrical Communication, Vol 62 Number 3, 1988.

[**FIORETTI 88b**] 'Fibre optic LAN/MAN architectures for coherent transmission.' A Fioretti, CA Rocchini and SR Treves, Alcatel FACE, Pomezia, Italy.

Electrical Communication, Vol 62 Number 3, 1988.

[**FRANTA 84**] 'Measurement and analysis of hyperchannel networks.' WR Franta and JR Heath. IEEE Transactions on Computers Vol C-33 Marhc 1984.

[**GIGABIT 88**] '16G040 Clock & data recovery circuit 2.0 Gbit/s NRZ data rate.' Data sheet from Gigabit Logic, 1908 Oak Terrace Lane, Newbury Park, CA 91320-5524.

[**GILOI 86**] 'Upperbus: A high speed backbone for metropolitan area networks.' WK Giloi, P Behr and G Zuber (BERCOM). Proceedings of EFOC/LAN, Amsterdam, June 1986.

[**GOTO 85**] NEC ring. 'A 1.2 Gbit/sec optical loop LAN for wideband office communications.' H Goto, F Akashi, B Hirosaki and H Shimizu of NEC. Proceedings IEEE Globecom December 1985 paper 15.4.

[**GRAVES 87**] 'An experimental cross-connect system for metropolitan applications.' AF Graves, PA Littlewood and S Carlton. IEEE JSAC Vol SAC-5, number 1 January 1987.

[**GREAVES 88**] 'The Cambridge backbone ring network.' DJ Greaves and A Hopper. Proceedings European Fibre Optic Conference (EFOC / LAN 88) Amsterdam, June 1988.

[**GREAVES 89**] 'Cambridge HSLAN protocol review.' DJ Greaves, ID Wilson. Proceedings of IFIP TC6 International Workshop on 'Protocols for high-speed networks' edited by H Rudin and R Williamson, held at IBM Ruschlikon 1989. Elsevier 1989.

[**HEINZMANN 91**] 'Buffer insertion cell synchronised multiple access (BCMA) on a Slotted Ring.' P Heinzmann, HR Muller, DA Pitt and HR van As, IBM Zurich. Second international conference on local communications June 26, Palma, Balearic Islands, Spain.

[**IEEE 89gep**] 'IEEE draft proposal 802.6 for Metropolitan Area Networks.' Draft proposal 802.6. September 1989.

[**HOPPER 78**] 'Local area computer communications networks.' A Hopper. University of Cambridge Computer Laboratory technical report number 7. 1978.

[**HOPPER 88**] 'The Cambridge Fast Ring Networking System.' A Hopper and RM Needham. IEEE transactions on computers, Vol 37 number 10. October 1988.

[**HUBER 83**] 'SILK: an implementation of a buffer insertion ring.' DE Huber, W Steinlin and P Wild. IEEE JSAC Vol SAC-1 number 5, 766-74. 1983.

[**HUNWICKS 89**] 'Optical feedthrough transmissions for the local loop.' AR Hunwicks, AJ Cooper and L Bickers, BTRL Ipswich England. Proceedings IEE colloquium on fibre optic LANs and techniques for the local loop, Savoy Place, Friday 17th March 1989.

[**KAMAL 89**] 'X-NET: A dual bus fibre-optic LAN using active switches.' AE Kamal BW Abeysundara, University of Alberta. Proceedings of ACM SIGCOMM 1989, Austin, Texas.

[**KATEVENIS 87**] 'Fast switching and fair control of congested flow in broadband networks.' MGH Katevenis. IEEE JSAC Vol SAC-5 number 8, pp 1315-1326. October 1987.

[**KELLER 83**] 'Transmission design criteria for a synchronous token ring.' HJ Keller, H Meyr & HR Mueller. IEEE JSAC Vol SAC-1, pp 724-733. November 83.

[**LIMB 82**] 'Description of Fasnet, a unidirectional local area communications network.' JO Limb and C Flores. BSTJ Vol 61 pp 1413-1440. September 1982.

[**LUVISON 89**] 'The LION approach to multi-service business networks.' A Luvison, CSELT. Proceedings 'ONLINE 89', 12th european congress for technical communications. Hamburg January 1989.

[**MARHIC 86**] 'Experimentation with a fibre optic implementation of Expressnet.' ME Marhic and FA Tobagi. Proceedings EFOC/LAN 86 pp 244-254. Amsterdam 86.

[**MAXEMCHUK 85**] 'Voice and data on a CATV network.' N Maxemchuck and A Netravali IEEE JSAC Vol SAC-3 number 2. March 1985.

[**MAC 89**] 'Protocol design for high speed networks.' DR McAuley, University of Cambridge, PhD dissertation, 1989.

[**MEYER 82**] 'Synchronization failures in a chain of repeaters.' H Meyr, L Popken, H Keller and HR Mueller. Proceedings of IEEE Globecom Miami 1982. Paper D6.1.

[**MINAMI 85**] 'A 200 Mbit/s synchronous TDM loop optical LAN suitable for multiservice integration.' T Minami, K Yamaguchi, T Nakagami, H Takanashi, N Fujino, H Hamano, M Suyama, K Iguchi and I Ymada. IEEE JSAC Vol SC-3 number 6. November 1985.

[**MITRANI 86**] 'A modelling study of the Orwell ring protocol.' I Mitrani, JL Adams and RM Falconer, BTRL. In 'Teletraffic analysis and computer performance evaluation' edited Boxma, Cohen and Tijms. North-Holland 1986.

[**MOLLENAUER 86**] 'Draft of proposed standard 802.6 metropolitan area

network (MAN) multiplexed slotted and token medium access control.' JF Mollenauer, DTW Sze, P Wild, R Klessig and B Bean. Revision A November 3, 1986.

[**MORRIS 84**] 'Some results for multi-queue systems with multiple cyclic servers.' RJT Morris amd YT Wang. In 'Performance of computer communication systems' edited by H Rudin and W Bux, Elsevier and IFIP 1984.

[**MORRISON 89**] 'High speed channel status (draft X3T9.3)' J Morrison. Proceedings of TGV7, Stuttgart, September 1989. Edited by J Erceau, ONERA, Chatillon, France.

[**MUKHERJEE 88**] 'The Pi-persistent protocol for unidirectional bus networks.' B Mukherjee and JS Meditch. IEEE Transactions on Communications. Vol 36 number 12. December 1988.

[**NEWMAN 88**] 'A fast packet switch for the integrated services backbone network.' P Newman. University of Cambridge Computer Laboratory technical report 142. July 1988. Also IEEE JSAC Vol SAC-6 number 9. December 1988.

[**NEWMANN 86**] 'Distributed Queueing: A fast and efficient packet access protocol for DQDB.' RM Newmann and JL Hullet. University of Western Australia. Proceedings of ICCC 1986, pp 294-299. North-Holland.

[**NISHIMOTO 86**] 'Fully integrated 1.6 Gb/s optical repeater.' H Nishimoto, A Miyauchi, N Fujimoto, T Touge and K Yamaguchi. Procedings GLOBECOM 1986 paper 49.1.1.

[**OFEK 89**] 'The conservative code for bit synchronisation.' Y Ofek, IBM TJ Watson Research Centre. Proceedings Third IEEE conference on MANs. San Diego 1989.

[**OKADA 87**] 'A new resolvable-contention access control scheme for ring-type LANs with high speed links : ReC - Ring.' H Okada and S Ohno. Proceedings of ICC 1987, Seattle, June 1987.

[**OOI 88**] 'A prototype gigabit/second multi-service network.' S Ooi, G Watson and D Skellern, HP Bristol. IEE colloquium on integrated multi-service networks, November 1988.

[**PACHL 88**] 'Livelocks in slotted ring networks.' J Pachl, IBM Zurich research laboratory. IEEE INFOCOM 88, March 1988.

[**PATIR 85**] 'An optical fibre based LAN for MAGNET's testbed environment.' A Patir, T Takahashi, Y Tamura and ME Zarki. Columbia University. IEEE JSAC Vol SAC-3 number 6. November 1985.

[**PERDUE 89**] 'The Ultra Gbit LAN.' N Perdue. Proceedings of TGV7,

Stuttgart, September 1989. Edited by J Erceau, ONERA, Chatillon, France.

[**PITT 91**] 'Buffer insertion cell-synchronised multiple access (BCMA) on a slotted ring.' P Heinzmann, HR Muller, DR Pitt and HR van As. Proceedings of second international conference on local communication.' Palma de Mallorca, June 1991.

[**PLESSEY 85**] 'SP5000 Single chip frequency synthesis.' Plessey Company publication PS 2011. July 1985.

[**PORTER 89**] 'MAC layer interconnection of homogeneous LANs.' JD Porter. Cambridge University PhD dissertation.

[**ROSS 89**] 'An overview of FDDI: The fibre distributed data interface.' FE Ross. IEEE JSAC Vol SAC-7 number 7, September 1989.

[**SAUER 89**] 'Multi-Gb/s opto-electronic interconnection system.' JR Sauer. Proceedings of TGV7, Stuttgart, September 1989. Edited by J Erceau, ONERA, Chatillon, France.

[**SCHILL 87**] 'Performance analysis of the FDDI 100 Mbit/s optical token ring.' A Schill and M Zieher. Proceedings IFIP WG 6.4 workshop on high speed LANs. Aachen, West Germany Feb 16-17 1987.

[**SCHMIDT 83**] 'Fibrenet II: A fibre optic ethernet.' IEEE JSAC Vol SAC-1 number 5 pp 702. November 1983.

[**SHARLAND 83**] 'An optical fibre supervisory system.' AJ Sharland, RM Brooks, GP Coombs and S Whitt. BT Tech J Vol 1 number 1. July 1983.

[**SHARP 84**] 'Analysis of channel access schemes for high-speed LANs.' R Sharp and NN Pedersen, Technical University of Denmark. SIGCOMM 84: Communications architectures and protocols. Montreal June 1984.

[**SHARP 87**] 'The LAN-DTH 140 Mbit/sec token ring.' RI Sharp. Proceedings IFIP WG 6.4 workshop on high speed LANs. Aachen, West Germany Feb 16-17 1987.

[**SHIMIZU 87**] 'A 400 Mbit/s LAN for multimedia service integration capabilities.' H Schimizu, S Nakai and B Hirosaki. NEC. Proceedings IFIP WG 6.4 workshop on high speed LANs. Aachen, West Germany Feb 16-17 1987.

[**SKOV 89**] 'Implementation of High-Speed Physical and Media Access Protocols.' Morten Skov, IEEE Communications Magazine, June 1989.

[**LEE 83**] 'The principals and performance of Hubnet.' ES Lee and Boulton. IEEE JSAC Vol SAC-1 number 5. November 1983.

[**SZE 85**] 'A metropolitan area network.' DTW Sze. IEEE JSAC Vol SAC-3 number 6. November 1985.

[**TANI 87**] 'High-speed Multimedia communication system on a wideband LAN.' H Tani, K Maebara, F Kashi and H Shimizu of NEC. Proceedings IEEE Globecom Tokyo 87 paper 15.6

[**TOBAGI 83**] 'Expressnet: A high performance integrated-services local area network.' FA Tobagi, F Borgonovo and L Fratta. IEEE Journal Selected Area Communications, SAC-1 number 6 Nov 1983.

[**TENNENHOUSE 86**] 'The Unison data link protocol specification.' DL Tennenhouse. Project Unison document ref UC021. October 1986.

[**TENNENHOUSE 88**] 'Site interconnection and the exchange architecture.' DL Tennenhouse. University of Cambridge PhD dissertation Sept 1988.

[**TRISCHITTA 88**] 'The accumulation of pattern-dependent jitter for a chain of fibre optic regenerators.' PR Trischitta and P Sannuti. IEEE transactions on communications Vol 36 number 6. June 1988.

[**VEENDRICK 80**] 'The behaviour of flip-flops used as synchronizers and prediction of their failure rate.' HJM Veendrick. IEEE journal of solid state circuits VOL SC-12 number 2 pp 169-176. April 1980.

[**VITERBI 66**] 'Principles of coherent communications.' AJ Viterbi. New York: McGraw Hill, 1966.

[**WATSON 89**] 'A prototype Gbit/second multi-service network.' S Ooi and G Watson, HP Bristol UK. Proceedings third IEEE conference on MANs, San Diego 1989.

[**WATSON 89**] 'A performance analysis of S++: A MAC protocol for High Speed Networks.' G Watson, HP Bristol UK and S Thome, ENST Paris.

[**WHEELER 89**] 'The livelock-free protocol of the Cambridge Ring.' Prof DJ Wheeler. The Computer Journal, Vol 32 number 1, page 95. 1989.

[**WHITT 85**] 'Automatic timing alignment for regenerative repeaters.' S Whitt. Electronics letters Vol 21 number 24 pp 1122 November 1985.

[**WIDEMER 83**] 'A DC balanced, partitioned block 8B/10B transmission code.' AX Widemer and PA Franaszek. IBM Journal of Research, vol 27 number 5. September 1983.

[**WILKES 79**] 'The Cambridge digital communications ring'. MV Wilkes and DJ Wheeler in proc Local-Area Communications-Network Symposium, Boston MA.

May 1979.

[**XTP 89**] 'XTP Protocol definition.' Revision 3.4, 17 July 1989. Available from Protocol Engines Inc, 1421 State Street, Santa Barbara, CA 93101.

[**YANG 86**] 'Performance analysis of multiple token ring and multiple slotted ring networks.' Q Yang, D Ghosal and LN Bhuyan. Proceedings Computer Networking Symposium, Washington DC 1986. pp 79-86. BL: 3394.1148

[**ZAFIROPULO 72**] 'Signalling and frame structures in highly decentralized loop systems.' P Zafiropulo and EH Rothauser. Proc. 1st Int. Conference on Computer Communications. Washington DC 1972.

[**ZAFIROVIC 88a**] 'Performance modelling of an HSLAN slotted ring protocol'. M Zafirovic-Vukotic & IGMM Niemegeers. Proceedings ACM SIGMETRICS, Santa Fe, May 1988.

[**ZAFIROVIC 88b**] 'Performance analysis of slotted ring protocols in HSLANs.' M Zafirovic-Vukotic, IG Niemegeers and DS Valk. IEEE JSAC Vol SAC-6 number 6 July 1988.