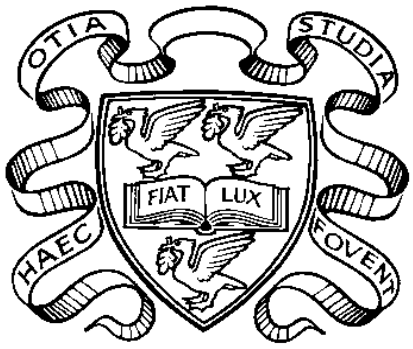
The IEEE 1355 Standard: Developments, Performance and Application in High Energy Physics



Thesis submitted in accordance with the requirements of the
University of Liverpool
for the degree of Doctor of Philosophy

by

Stefan Haas

December 1998

The IEEE 1355 Standard: Developments, Performance and Applications in High Energy Physics

Stefan Haas

The data acquisition systems of the next generation High Energy Physics experiments at the Large Hadron Collider (LHC) at CERN will rely on high-speed point-to-point links and switching networks for their higher level trigger and event building systems. This thesis provides a detailed evaluation of the DS-Link and switch technology, which is based on the IEEE 1355 standard for Heterogeneous InterConnect (HIC). The DS-Link is a bidirectional point-to-point serial interconnect, operating at speeds up to 200 MBaud. The objective of this thesis was to study the performance of the IEEE 1355 link and switch technology and to demonstrate that switching networks using this technology would scale to meet the requirements of the High Energy Physics applications.

The performance and reliability of the basic point-to-point interconnect technology over electrical and fibre optic media were examined. These studies were carried out while the IEEE 1355 standard was still being finalised and have therefore provided valuable input to the standards working group. In order to validate the fibre optic physical layer proposed for the IEEE 1355 standard, an implementation demonstrator of a DS-Link interface for fibre optics, employing a new line encoding scheme, has been designed and characterised. This interface allows the link length for DS-links to be extended, which is important in the HEP context, where the cable length from the detector to the electronics can be up to 200 meters.

A large switching network testbed of variable topology based on IEEE 1355 point-to-point serial links and high-valency crossbar switches has been designed and constructed. The network testbed consists of up to 1024 end nodes which are connected to a reconfigurable switching fabric constructed from 32-way crossbar switches. The end nodes are loaded with a predetermined traffic pattern and the response of the network in terms of throughput and latency is measured. The testbed allows the network performance of various topologies to be studied as a function of the network size and the traffic conditions, including those expected in HEP trigger and event building systems. This system is believed to be unique in its ability to measure network performance under well controlled and well defined conditions.

The results from the Macramé network demonstrate that large IEEE 1355 DS-Link networks can be built and that they scale very well. Furthermore, it has been shown that per-link flow control, together with well designed hardware, can result in very reliable systems. It was also demonstrated, that a switching fabric based on the IEEE 1355 technology could meet the required network performance of the ATLAS second level trigger.

Acknowledgements

I would like to thank my supervisors Bob Dobinson and Erwin Gabathuler for the opportunity to carry out the work presented in this thesis and for their guidance and advice throughout the thesis.

Many thanks also to the other members of the Macramé team: Brian Martin, David Thornley and Minghua Zhu. They all played a crucial role in the successful completion of the project.

I am grateful for the support of the European Union through the Macramé project (Esprit project 8603), without which this thesis would not have been possible, and to CERN for hosting this project.

Finally, I would also like to thank my parents for their encouragement and support.

Table of Contents

Chapter 1
Introduction

1.1 Motivation

Traditionally the detector read-out and event building systems of High Energy Physics (HEP) experiments have been based on hierarchical bus topologies, using standard parallel shared bus systems, such as VME or FASTBUS. This approach however does not scale well to very large systems, since it suffers from the bottleneck in the bus bandwidth and the limited interconnectivity available between multiple buses. The bus based architecture can therefore not accommodate the higher performance requirements of the next generation of experiments [1,2], at the Large Hadron Collider (LHC) being built at CERN. The proposals for the data acquisition systems of these experiments therefore rely on high-speed point-to-point links and switching networks for their higher level trigger and event building systems.

There are several technologies which are currently being investigated for this application. These include ATM [3], SCI [4], FibreChannel [5] and more recently also Ethernet [6]. This thesis provides a detailed evaluation of another serial link and switch technology under consideration, which is based on the IEEE[1] 1355 standard for Heterogeneous InterConnect (HIC) [7].

The IEEE 1355 technology enables the construction of scalable low latency serial interconnect systems based on high-speed point-to-point links and switches. The standard specifies the physical media and low level protocols for two complementary high-speed serial link technologies which have been developed within the framework of the European Commissions ESPRIT[2] program. The speeds and media range from 100 MBaud to 1 GBaud in both copper and optic technologies. The various specifications enable chip-to-chip, board-to-board and rack-to-rack communications. The work presented in this thesis focuses on the DS-Link technology. The DS-Link is a bidirectional point-to-point serial interconnect, operating at speeds from 10 to 200MBaud.

Any large switching network depends critically on the underlying serial interconnect technology. Therefore a study of the physical layers of the DS-Link was undertaken. Electrical twisted pair cable and fibre optic technologies for link signal transmission have been evaluated. This included establishing performance limitations in terms of transmission speed and achievable link length, as well as reliability tests. In order to validate the fibre optic physical layer proposed for the IEEE 1355 standard, an implementation demonstrator of a DS-Link interface for fibre optics, employing a new line encoding scheme, has been designed and characterised. This interface allows the link length for DS-links to be extended, which is important in the HEP context, where the cable length from the detector to the electronics can be between 50 to 200 meters. This study and the prototyping work were carried out while the

1. Institute of Electrical and Electronics Engineers
2. European Strategic Programme for Research in Information Technology

IEEE 1355 standard was not yet finalised and therefore provided useful input to the standardisation procedure.

Having established the performance and reliability of the underlying serial point-to-point link technology, a large reconfigurable switching network testbed based on IEEE 1355 point-to-point serial links and high-valency crossbar switches has been designed and constructed. The objective of this work was to investigate the performance and scalability of IEEE 1355 DS-Link based switching fabrics and to demonstrate the feasibility of constructing large scale systems using this technology.

The network testbed consists of up to 1024 end nodes which are connected to a reconfigurable switching fabric constructed from 32-way crossbar switches. The end nodes are loaded with a predetermined traffic pattern and the response of the network in terms of throughput and latency is measured. The testbed allows the network performance of various topologies to be studied as a function of the network size and the traffic conditions, including those expected in HEP trigger and event building systems. This system is believed to be unique in its ability to measure network performance under well controlled and well defined conditions. For no other interconnect has such a large and controlled test environment been set up.

1.2 Context

The work presented in this thesis was carried out at CERN within the framework of the European Commissions ESPRIT Macramé[3] project. The direction of the research was strongly influenced by the project and the developments carried out on the IEEE 1355 technology.

The Macramé project was a collaboration between 11 partners from European research institutions and industry based in the United Kingdom, France and Norway. The objective of the project was to develop and promote the IEEE 1355 technology. This technology had been initially developed within previous ESPRIT projects for interprocessor communication. The work carried out at CERN included the construction of a fibre optic link demonstrator, calibration of simulation models for DS-Links and switches and the construction and exploitation of the 1024 node scalable and reconfigurable DS-Link network testbed. All tasks were successfully completed and CERN's contribution was recognised as a major success by the industrial partners and by the external project reviewers appointed by the European Union.

1.3 Outline of the Thesis

Following this introduction, chapter 2 presents the IEEE 1355 standard, with focus on the DS-Link, which is the technology that has been studied in this thesis. The DS-Link components which are relevant to this work are also introduced. Finally the theoretical performance limits of the interconnect are examined.

Chapter 3 reports on the evaluation of electrical DS-Link signal transmission. The differential electrical transmission of DS-Link signals over twisted-pair cable has been characterized and the performance in terms of link speed and link length has been measured. The reliability of

3. Esprit project 8603: Multiprocessor Architectures Connectivity Routers And Modelling Environment

this type of connection, which is an important consideration when building a large system, has also been determined. Finally, results from testing the susceptibility of differential DS-Link transmission to electromagnetic interference are presented and recommendations of how to improve the immunity to this type of interference are made.

Chapter 4 introduces fibre optic transmission system and examines its theoretical error rate performance. The transmission code which was proposed for the IEEE 1355 is introduced. In order to validate the encoding scheme proposed for the fibre optic physical layer of the IEEE1355 standard, a prototype implementation of a point-to-point fibre optic connection for DS-Links has been designed and characterised. The design is presented and test results for the fibre optic transceiver as well as the complete link are shown.

Chapter 5 introduces the fundamentals of switching networks. Specific features of the cross-bar switch used will be explained. The different network topologies that have been studied and the traffic patterns that were used will also be presented. Finally analytical results for the theoretical performance of the basic packet switch will be given.

Chapter 6 presents the design and implementation of the large scale IEEE 1355 network test-bed. First an overview of the architecture of the testbed is given. The individual hardware modules used to construct the testbed will then be described in detail, and a short overview of the software required to operate the testbed will also be given. Finally results from an evaluation of the basic performance of each of the components are shown.

Chapter 7 presents results from performance measurements carried out on the testbed. The performance of different network topologies has been studied for different network sizes and under various traffic conditions. This includes measurement of network performance for the type of traffic expected in the ATLAS second level trigger system. The results are analysed and simple mathematical models of the network behaviour are given where possible.

Finally chapter 8 gives a summary of the conclusions presented throughout this thesis.

1.4 Author's Work

The evaluation of the electrical DS-Link transmission system and the study of susceptibility to electromagnetic interference presented in chapter 3 are the authors own work.

The work presented in chapter 4 was carried out entirely by the author. This included design, construction, and test of the fibre optic interface for DS-Links.

The construction of the Macramé testbed presented in chapter 6 was an effort of a team of two hardware engineers and one software engineer. The author has had a central role at each stage of the project through system specification, design, test and finally full implementation of the network testbed. More specifically, the design of the basic component of the testbed, the traffic node, was entirely the authors work. This includes design and debugging the modules, writing the low level driver and board test software, and system integration and test. The timing node module was partly developed by another member of the team [8]. The performance analysis of the different components is also the authors own contribution.

The results as well as the analysis and conclusions presented in chapter 7 are the work of the author, including setting up and performing measurements. The traffic descriptors for the HEP results were produced in collaboration with another member of the group.

Chapter 2
IEEE 1355 Technology

This chapter will introduce the IEEE 1355 standard, with focus on the DS-Link technology which has been used for the work reported here, and also examine the theoretical performance limits of the interconnect.

The author has actively participated in the working group that established the standard. His contributions were mainly on the definition and testing of the differential electrical DS-Link transmission over twisted-pair cable and on the fibre optic DS-Link physical layer. This work will be presented in Chapter 3 and Chapter 4.

2.1 Overview of the IEEE 1355 Standard

The IEEE 1355-1995 standard (ISO/IEC 14575) of scalable, heterogeneous interconnect [7] defines the physical implementations and logical protocol layers for a point-to-point serial interconnect, operating at speed ranges from 10–200 MBaud[1] and 1 GBaud in copper and fibre optic technologies. The baseline technology for this standard has been developed within the OMI/HIC[2] project. Many aspects of the technology have their origins in earlier work on parallel computer systems. In particular, the routing strategy was established in the ESPRIT[3] project PUMA[4] and the DS-Link technology was partially developed in the GPMIMD[5] project.

The IEEE 1355 standard enables the construction of low-cost, high-speed scalable interconnects. Although this technology has been initially designed as a multiprocessor interconnect, it is equally appropriate for applications in communication systems and local area networks. It also allows the transparent implementation of a range of higher level protocols such as ATM [9], SCI [4] and Ethernet [10]. Some potential application areas for IEEE 1355 links and switches are listed below:

- LAN switching hubs
- parallel computers
- data acquisition systems
- ATM switching
- industrial control systems
- multimedia servers

1. The unit Baud denotes the signalling rate on the physical transmission medium.
2. OMI/HIC: Open Microprocessor Systems Initiative/Heterogeneous Inter-Connect Project, ESPRIT project 7252
3. ESPRIT: European Strategic Programme for Research and Development in Information Technology
4. PUMA: Parallel Universal Message-Passing Architectures, ESPRIT project 2701
5. GPMIMD: General-Purpose MIMD Machines, ESPRIT project 5404

- home networks

2.1.1 The Protocol Stack

As with most communications systems, the IEEE 1355 standard can be best described by a hierarchical protocol organised in a number of layers, the so-called protocol stack. Each protocol layer builds on and expands the functionality of the layers below it, providing a greater abstraction from the underlying physical medium used. Figure 1 illustrates this concept, the functions performed by each of the layers are described below.

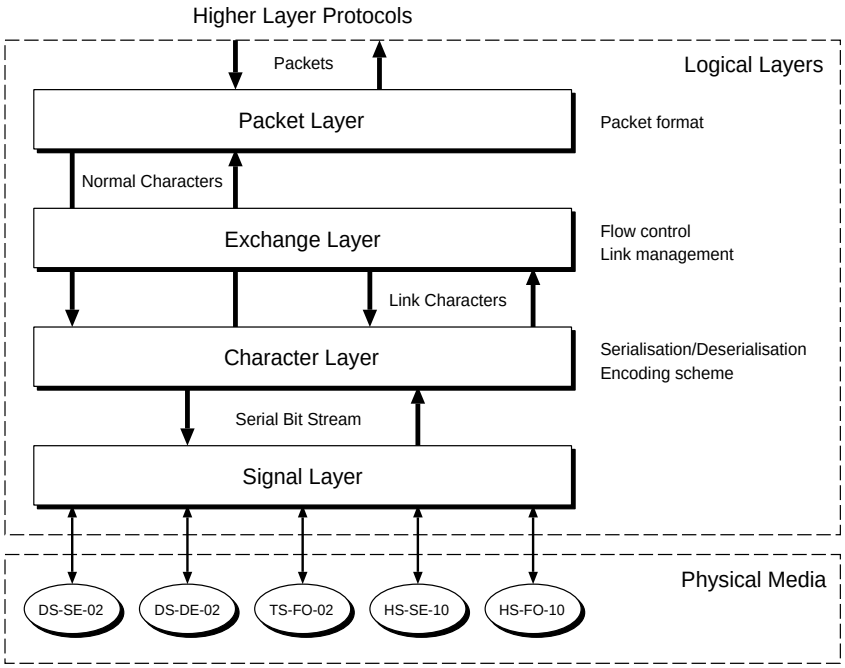


Figure 1: IEEE 1355 Protocol Stack

2.1.1.1 Physical Media

The IEEE 1355 standard describes two complementary high speed serial link technologies which address different speed ranges and consequently also use different encoding schemes:

- 100 MBaud DS-Link
- 1 GBaud HS-Link

The Data/Strobe encoding (DS), which is presented in more detail below, is used up to 100MBaud. HS links work at 1GBaud and use a balanced 8B12B[6] block code, which generates a transition at the beginning of each code word, in order to simplify clock recovery.

The physical implementations defined by the standard are shown in Table 1. The name for each physical medium consists of three fields which designate the encoding, the transmission medium and the bit rate. For example the DS-SE-02 describes a 200MBaud single-ended electrical link using the Data/Strobe (DS) encoding. Electrical transmission can be either sin-

6. 8B12B means that a group of 8 data bits is encoded into 12 code bits, i.e. the Baud rate is 12/8 of the data bit rate.

gle-ended (as in DS-SE) or differential for longer distances (DS-DE). A fibre optic version is specified for both the DS and the HS link technologies (TS-FO and HS-FO). The TS-FO physical medium uses a different encoding scheme, the Three-of-Six code (TS), which is suitable for fibre optic transmission. The last two digits represent the nominal speed of the link in 100MBaud units.

Table 1: IEEE 1355 Physical Media

Technology	Baud Rate [MBaud]	Transmission medium	Maximum distance [meter]
DS-SE-02	10–200	PCB trace	0.3
DS-DE-02	10–200	twisted-pair cable	12
TS-FO-02	250	multimode fibre	300
HS-SE-10	700–1k	coax cable	8
HS-FO-10	700–1k	single-mode or multimode fibre	1000–3000

The implementation and performance of the DS-DE and TS-FO physical layers has been studied in detail, and are described in Chapter 3 and Chapter 4, respectively.

2.1.1.2 Signal Layer

Signals propagate over the physical transmission media such as electrical cables or optical fibres and are interpreted as a sequence of bits. The signal layer specifies parameters such as the signal voltage levels and noise margins, the line signal rate (or Baud rate) and power budget or maximum transmission length.

2.1.1.3 Character Layer

A character is a group of consecutive bits which represent control or data information. Normal characters are the 256 data characters plus two control characters which are used as the end-of-packet markers. Link characters are control characters used for the exchange layer protocol and are local to the link, i.e. they are invisible to the packet layer. The character layer specifies the encoding scheme and performs serialisation and deserialisation of characters into a bit stream. It extracts the serial bit stream and clock from the line signal.

2.1.1.4 Exchange Layer

This layer specifies the exchange of link characters in order to ensure proper functioning of a link. This includes functions such as per-link flow control, link start-up and shutdown as well as error handling. All the implementations defined in the standard use the same credit-based flow control mechanism, which is explained in detail in section 2.2.3.1 on page 10.

2.1.1.5 Packet Layer

A packet consists of a destination identifier followed by the payload and an end-of-packet marker. The standard does not define a specific (or maximum) size for packets. This allows different packet formats to be carried over an IEEE 1355 network.

2.1.2 Advantages of IEEE 1355

The following list gives a summary of the advantages of the IEEE 1355 technology:

- Credit-based flow control on a per-link basis: this prevents packets from being lost in the switching fabric, which simplifies the higher layer protocols, since the retransmission of packets is not necessary, unless an error occurs.
- Small protocol overhead: this makes the links very efficient, even for short packets.
- Flexible packet format: this allows IEEE 1355 networks to be used as a carrier for other higher level protocols.
- IEEE 1355 provides a set of lightweight protocols for bidirectional flow-controlled, point-to-point communication.
- Low implementation complexity of IEEE 1355 interfaces: this enables the implementation of packet switches with a large number of ports.
- Low latency and minimal buffering: the fast link level flow control of IEEE 1355 links enables the use of “wormhole” routing, which provides low switching latency and also requires minimal buffering in the switches.

2.2 Data/Strobe Links (DS-Links)

This section will introduce the DS-Link technology in some detail, since the work presented has been focused on this technology. DS-Links provide bidirectional point-to-point communication between devices. Each DS-Link implements a full-duplex, asynchronous, flow-controlled connection operating at a programmable link speed. The IEEE standard specifies a maximum link speed of 200MBaud, however the integrated circuits which are currently available are only specified for operation up to 100MBaud. Tests on some of these devices have shown that the links would work at 200MBaud, although this is out of the specification.

2.2.1 Signal Layer

DS-Links consist of four wires, two in each direction, one carrying data and one carrying a strobe, hence the term DS-Links (Data/Strobe). The data signal carries the serial bit stream, while the strobe signal changes state every time data does not change. This ensures that there is a transition on either data or strobe at the boundary of every bit frame. The data/strobe wire pair thereby carries an encoded clock, which can be simply recovered by generating the logical exclusive-or of the two signals. This scheme is very similar to the one presented in [11]. Figure 2 shows a binary bit stream and the corresponding data and strobe signals.

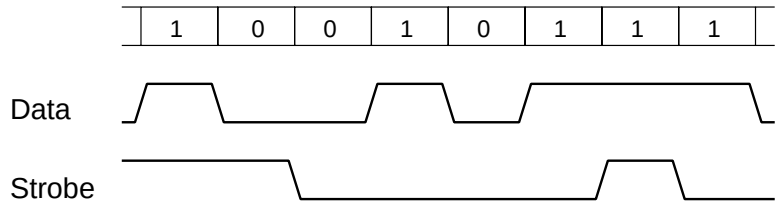


Figure 2: Data and Strobe Signals

The advantage of the two wire transmission of the DS-Link over the more traditional approach for a serial communication link using only one wire is the simple clock extraction. The traditional approach either requires a clock recovery circuit, e.g. a PLL[7], to extract the

7. Phase Locked Loop

clock from the bit stream, or oversampling of the bit stream at the receiver, e.g. RS232 or OS-Links[8].

The Data/Strobe transmission scheme is less sensitive to signal skew than a system which simply transmits the serial data and the clock signal on separate wires. The DS-encoding provides a full bit period of skew tolerance. Due to the encoded clock, DS-Links can also auto-baud, i.e. the transmit rate can be varied as long as it does not exceed the maximum speed of the receiver. Because of these features, the Data/Strobe encoding scheme has also been adopted for the signal layer of the IEEE 1394 SerialBus standard [12].

2.2.2 Character Layer

Figure 3 shows the encoding of the DS-Link characters. The first bit of a character is the parity bit, followed by a control flag, which is used to distinguish between control and data characters. If the control bit is zero then it is followed by 8 bits of data, with the least significant bit being transmitted first. Control characters are 4 bits long and consist of a parity bit, the control/data bit which is set to 1 to indicate a control character, and 2 bits to indicate the type of control character.

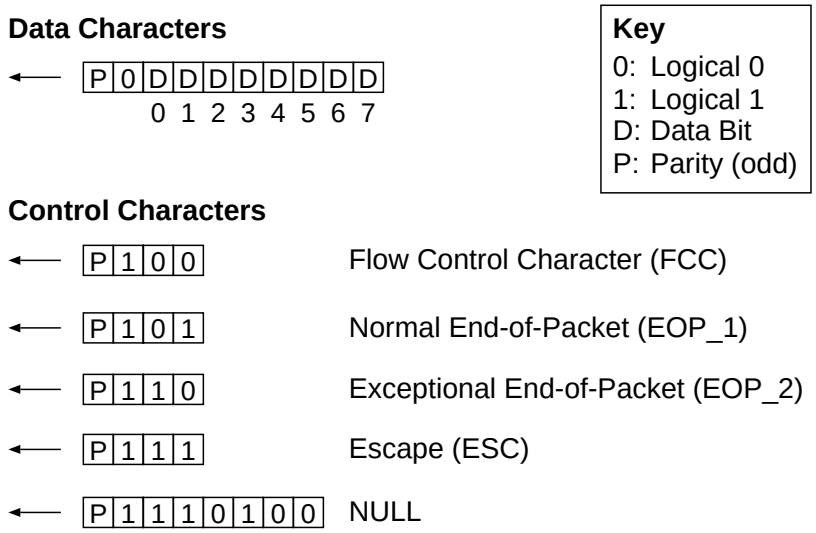


Figure 3: DS-Link Character Encoding

The parity bit covers the data or control bits in the previous character and the control flag in the current character. This allows the detection of single bit errors. Odd parity checking is used, i.e. the parity bit is set such that the bits covered, including of the parity bit itself, always contain an odd number of ones.

The normal end-of-packet character `EOP_1` is used to indicate the end of a packet. The exceptional end-of-packet character `EOP_2` character can be used to signal the end of a message (see section 2.2.5 on page 12 below) or to indicate that an error has occurred. The actual use of `EOP_2` is defined by the higher layer protocols. `NULL` characters are transmitted in the absence of other characters. This enables the detection of link failures, e.g. due to a physical disconnection. The `NULL` character also allows the parity of the end-of-packet marker to be

8. Oversampled Links: old style transputer links running at up to 20MBaud

checked immediately. The FCC character is used for the link flow control mechanism described below.

2.2.3 Exchange Layer

The exchange layer defines the link flow control mechanism, the link start-up procedure and the handling of link errors.

2.2.3.1 Flow Control

IEEE 1355 links use a credit based flow-control scheme, which is local to the link and works on a buffer-to-buffer basis. The flow control mechanism ensures that no characters can be lost due to buffer overflow, which simplifies the higher levels of the protocol, since it removes the need for retransmission unless errors occur. From a system view, a DS-Link connection therefore appears as a pair of fully handshaken FIFO buffers, one in each direction.

The smallest unit on which flow control is performed is called a flow-control unit, or flit [13]. Each receiving link input contains a buffering for at least one flit. Whenever the link input has sufficient buffering available for a further flit of characters, a flow control character is transmitted on the associated link output. This FCC gives the sender permission to transmit a further flit. The transmitter maintains a flow control credit value, which is decremented when a data or a packet terminator character is sent, and incremented by the flit size when a flow control character is received. This is illustrated in Figure 4 below, where N denotes the flit size.

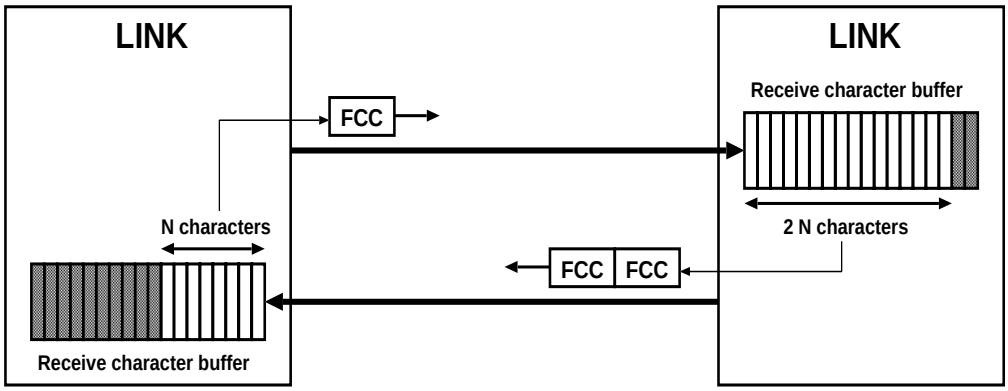


Figure 4: Link flow control

The flit size for the DS-Link is 8 characters. Therefore the receiver must have a buffer for at least 8 characters. However, because of the latencies inherent in DS-Link implementation and in order to sustain the full data rate over longer distances, a larger buffer is required, so that the character level flow control does not restrict the maximum bandwidth of the link. The requirement for continuous transmission is that the next FCC is received before the previous flit of eight characters has been fully transmitted, so that the link output is never stalled. This is analysed in more detail in section 2.4.3 on page 17.

2.2.3.2 Link Start-up

After power-on, both ends of a link maintain their data and strobe outputs at low. When a link is started, it transmits NULL characters until it has received a character from the remote end. The link then sends a number of FCC characters, corresponding to the number of flits that fit in the receive buffer. The link then enters normal operation and can send data characters when

flow control credit is available. This sequence ensures that the initial flow control characters are not lost, e.g. because the remote end is still being reset.

Some of the available DS-Link devices however send the flow control characters immediately when started, without waiting to receive NULL characters from the remote end. To avoid loss of FCC characters, both ends of the link have to be started up in the correct sequence under external control.

If one end of a link is reset during normal operation, that end stops transmitting characters. The receiver on the other end of the link detects this as a disconnection error (see section 2.2.3.3 below) and also stops transmitting and resets itself. After a delay both ends of the link are ready to start normal operation again. This scheme effectively allows the two ends of the link to operate in different reset domains.

2.2.3.3 Error Detection and Handling

The DS-Link protocol allows the most common errors to be detected. The parity check will detect all single bit errors at the DS-Link character level. The physical disconnection of a link can also be detected, since each link output transmits a continuous stream of characters once it has been started.

The DS-Link characters contain a parity bit which allows single bit errors to be detected. Odd parity checking is used. The parity bit in a character covers the parity of the data/control bit in the same character, and the data or control bits in the previous character, as shown in Figure 5 below. This scheme allows any single bit error in any single bit of a character, including the control/data bit, to be detected even though the characters are not all the same length.

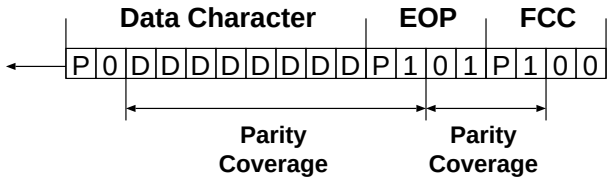


Figure 5: Parity Check on DS-Link Characters

When a DS-Link detects a parity error on its input it halts its output. This is detected as a disconnect error at the other end of the link, causing this to halt its output also. Detection of an error causes the link to be reset. Thus, the disconnect behaviour ensures that both ends are reset. Each end must then be restarted.

2.2.4 Packet Layer

Information in IEEE 1355 networks is transmitted in packets, a packet consists of the header, followed by the payload and an end-of-packet character. There is no explicit start-of-packet character, the first data character received after an end-of-packet is considered to be the header of the next packet. The packet format is illustrated in Figure 6 below:

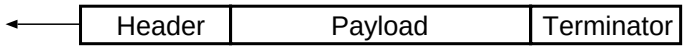


Figure 6: Packet Format

The header contains the destination identifier and is used to route a packet through the switching fabric. The standard does not define a specific packet size. This enables IEEE 1355 links to be used as a carrier for many higher level protocols, because the packet length can be chosen to match the requirements of the specific protocol, e.g. 53 bytes for ATM cells. There is no maximum packet length, however packets should be limited in size, since a long packet can occupy network resources, blocking other packets. A limited packet length will also allow an upper bound to be placed on the latency of packet transmission.

2.2.5 Higher Level Protocols

A variety of higher level protocols can be layered on top of the IEEE 1355 packet layer. DS-Link packets can be used as a transport mechanism for protocols defined by other standards such as ATM, SCI and Ethernet. Studies have been carried to use IEEE 1355 technology as the basis for an ATM switch architecture [15, 7, 9] and to map SCI protocols onto IEEE 1355 links [14]. There are also ongoing projects to use HS-Links for switching Gigabit-Ethernet frames and use an HS-Link based switching fabric to transport SCI transactions.

The T9000 transputer virtual channel protocol [16] is presented here as an example of a higher level protocol. It provides synchronized channel communication between processes running on different processors. For this protocol the packet length is restricted to 32 bytes of data. Messages smaller than 32 bytes are transmitted in one packet, longer messages are split into several packets. The exceptional end-of-packet marker is used to indicate the end of a message. In a network where there are several possible paths from source to destination, packets can potentially arrive out of sequence. In order to ensure in-order delivery of the packets within a message, the reception of each packet must be acknowledged by sending an empty packet, i.e. a packet which consists only of a header followed by an end-of-packet marker.

2.3 IEEE 1355 Integrated Circuits

A number of integrated circuits that implement the IEEE 1355 standard have been produced. The following devices support the DS-Link technology:

- The STC104 [17] is an asynchronous 32-way dynamic packet routing switch chip which uses 100 MBaud DS-SE links. The 32 bidirectional links are connected via a 32-way non-blocking crossbar. All the links operate concurrently, resulting in a maximum cross-sectional bandwidth of over 300 Mbyte/s.
- The STC101 Parallel DS-Link adapter [18] drives a DS-SE link at 100 MBaud full-duplex. The DS-Link can be accessed through a 16 or 32 bit wide bus interface. Alternatively, independent receive and transmit interfaces which connect directly to internal FIFO buffers can be used.
- The T9000 transputer [16] has four on-chip 100 MBaud DS-SE links for interprocessor communication.
- The SMCS (Scalable Multi-channel Communication Subsystem) [19] is a communications controller providing three DS-Link interfaces. It is designed for application in DSP[9] networks and has a 32-bit wide CPU interface.

9. Digital Signal Processor

- The CW-1355-C111 [20] is a low-cost PLD based DS-Link adapter chip that uses a simple byte-wide interface to external FIFO buffers.

A more detailed description of the STC104 32-way packet switch and the STC101 DS-Link adaptor will be given below, since these devices have been used extensively for the work presented here. There are currently only two devices that support the higher speed HS-Links:

- The Bullit evaluation chip [21, 22] provides a parallel interface to a 1GBaud HS-Link.
- The RCube 8-way router chip [23, 24] can be used to build HS-Link switching fabrics.
- The NOE chip [25], which is still under development, contains a PCI interface and two HS-Links. It is designed as a high-performance HS-Link network interface for PCs.

2.3.1 The STC101 Parallel DS-Link Adapter

The STC101 parallel DS-Link adaptor allows high speed serial DS-Links to be interfaced to parallel buses and microprocessors. The DS-Link of the device can operate at up to 100MBaud, providing a bidirectional bandwidth of 19 Mbytes/s.

2.3.1.1 Functional Description

The STC101 provides a simple interface to the DS-Link through FIFO buffers. Figure 7 shows the block diagram of the STC101.

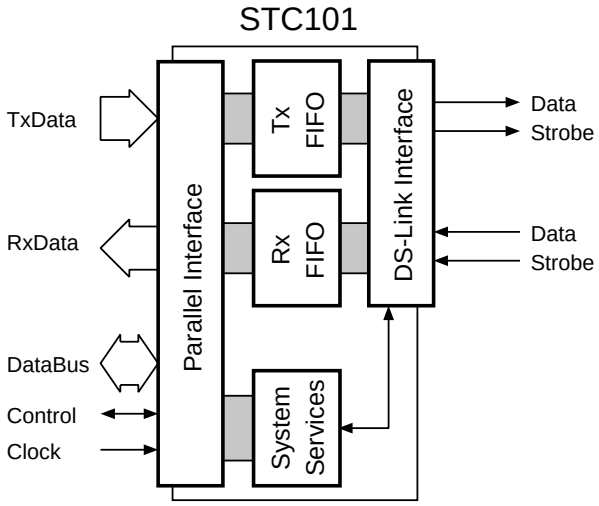


Figure 7: STC101 Parallel DS-Link Adaptor Block Diagram

The STC101 contains buffering for 64 characters on the transmit and the receive side in order to smooth out differences in data rates between the DS-Link and the parallel interface. These buffers are in addition to the 20 character receive buffer in the link interface, which is required for the flow control protocol of the DS-Link. The parallel transmit and receive character interfaces are 9 bits wide, for one data byte plus a bit to indicate an end-of-packet character. The interface is synchronous and uses a Valid/Hold handshake protocol [26], which allows the transfer of one character per clock cycle. The control and status registers of the device can be accessed through a 16 bit wide parallel bus interface.

2.3.2 The STC104 Packet Switch

The STC104 is a 32-way low latency packet routing switch chip. It interconnects 32 DS-Links through a 32 by 32 non-blocking crossbar switch. A non-blocking crossbar switch allows any permutation of connections between input and output ports to be made. Therefore the central crossbar enables packets to be routed between any of the switch links. Since the links operate concurrently, the transfer of a packet between one pair of links does not affect the data rate or latency for another packet passing between a different pair of links. Each link can operate at up to 100MBaud, providing a total aggregate switch bandwidth of over 300Mbyte/s. The STC104 supports packet rates up to 200Mpackets/s. In the absence of any contention for a link output, the switch latency, i.e. the time between the first bit of a packet being received on the input link and being retransmitted on the output link, will be less than 1 µs. A single STC104 can be used to connect up to 32 nodes. By connecting several STC104 switches complex switching networks can be built. Figure 8 show the block diagram of the STC104 switch.

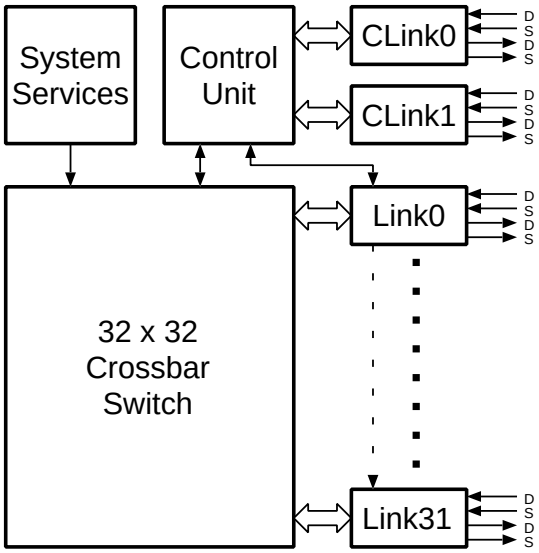


Figure 8: STC104 Packet Switch Block Diagram

This section only provides a short description of the features of the STC104 packet switch. The concepts which are relevant to this work will be explained in more detail in Chapter 5.

Data in a IEEE 1355 networks is transmitted in packets. To enable the packets to be routed to their destination, each packet has a header at the front which contains routing information. The STC104 uses the header of each incoming packet to determine the link to be used to output the packet. Anything after the header is treated as the packet body until the packet terminator is received. This enables the STC104 to transmit packets of arbitrary length.

The algorithm which makes the routing decision is called interval labelling. Each destination in a network is labelled with a number, and this number is used as the destination address in a packet header. Each of the 32 links on a routing switch is labelled with an interval of possible header values, and only packets whose header value falls within the programmed interval range for a given link are output via that link. Thus the header specifies a particular link along which to forward the packet.

Consecutive links may be programmed to be “grouped”. If a packet is routed to an output link which is busy, it will automatically be routed along any other link in the group which is available. In this way performance can be optimised by allowing packets to be routed to any one of several outputs, depending on which link in the group is the first to become available.

To eliminate network hot spots, the STC104 can optionally implement a two phase routing algorithm. This involves every packet being first sent to a randomly chosen intermediate destination; from the intermediate destination it is forwarded to its final destination. This algorithm, referred to as Universal Routing, is designed to maximize capacity and minimize delay under conditions of heavy load.

Usually packets are routed through the STC104 unchanged. However, a flag can be set in the specified output link, in which case the header of the packet is discarded. Each link output of the STC104 can be programmed to delete the header of a packet, revealing a second header to route the remainder of the packet to the destination device. This assists in the modular and hierarchical composition of routing networks and simplifies the labelling of networks. This feature is also useful to strip the routing header when a packet leaves the network.

The STC104 is controlled and programmed via a control link. The STC104 has two separate control links, one for receiving commands and one to provide daisy chaining. The control links enable networks of STC104s to be controlled and monitored for errors. The control links can be connected into a daisy chain or tree, with a controlling processor at the root.

2.4 Theoretical Performance

This section analyses the maximum theoretical performance of a DS-Link. The performance of the STC104 switch is presented in Chapter 5.

2.4.1 Unidirectional Link Bandwidth

The link bandwidth is the number of data bytes in a packet divided by the time it takes to transmit the packet on the link. Using the information on the DS-Link character encoding and the packet format, the theoretical maximum DS-Link bandwidth for unidirectional link usage can be calculated as follows:

$$BW_{UNI} = \frac{l}{t_{DATA} \cdot (l + n_{HDR}) + t_{EOX}} \quad \textbf{(1)}$$

where l is the packet length, n_{HDR} is the size of the routing header, t_{DATA} is the time to transmit one data character and t_{EOX} is the time to transmit an end-of-packet character. The last two parameters clearly depend on the Baud rate of the link. Figure 9 shows the maximum theoretical unidirectional link bandwidth as a function of the packet length for a link operating at 100MBaud. Two curves are shown, for one byte and two byte headers respectively. Longer routing headers should rarely be necessary, since this will allow networks of up to 65536 terminal nodes.

The asymptotic link bandwidth, i.e. the data rate for very long packets, is 10Mbyte/s as expected. It has to be noted that the link can only be 80% efficient, since on the average only

8 out of 10 bits carry data, due to the character encoding overhead, i.e. the parity bit and the control-data flag (see section section 2.1.1.3 on page 7). For short packets the bandwidth is reduced by the protocol overhead, i.e. header and terminator, but the throughput increases quickly with the packet length. Over 90% of the maximum throughput is achieved for packets longer than 12 bytes using single byte headers and for packets longer than 21 bytes using two byte headers.

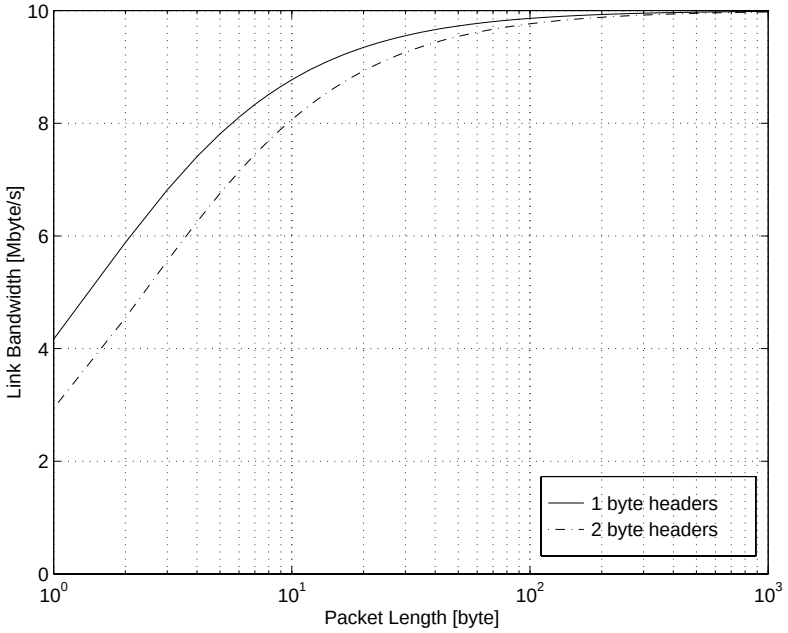


Figure 9: Calculated Unidirectional DS-Link Bandwidth

2.4.2 Bidirectional Link Bandwidth

For bidirectional link usage, some of the link bandwidth is used for the flow control characters, which are interleaved with the data characters. Compared to the case of unidirectional traffic, the time to transmit a packet therefore increases by the time to send the flow control characters. Assuming the same traffic in both directions on the link, the theoretical DS-Link bandwidth for bidirectional link usage can be calculated as follows:

$$BW_{BIDIR} = \frac{l}{t_{DATA} \cdot (l + n_{HDR}) + t_{EOX} + t_{FCC} \cdot \frac{l + n_{HDR} + 1}{8}} \quad \textbf{(2)}$$

where l is the packet length, n_{HDR} is the size of the routing header, t_{DATA} is the time to transmit one data character, t_{EOX} is the time to transmit an end-of-packet character, and t_{FCC} is the time to transmit a flow control character. Figure 10 shows the maximum theoretical bidirectional link bandwidth as a function of the packet length for one byte and two byte headers. The link is operating at 100MBaud with the same traffic flowing in both directions.

The asymptotic bidirectional link bandwidth, i.e. the data rate for very long packets, is 9.52Mbyte/s, the reduction compared to unidirectional link usage is due to the link level flow control, which consumes about 5% of the available link bandwidth. Over 90% of the maximum throughput is achieved for packets longer than 12 bytes using single byte headers and for packets longer than 21 bytes using two byte headers.

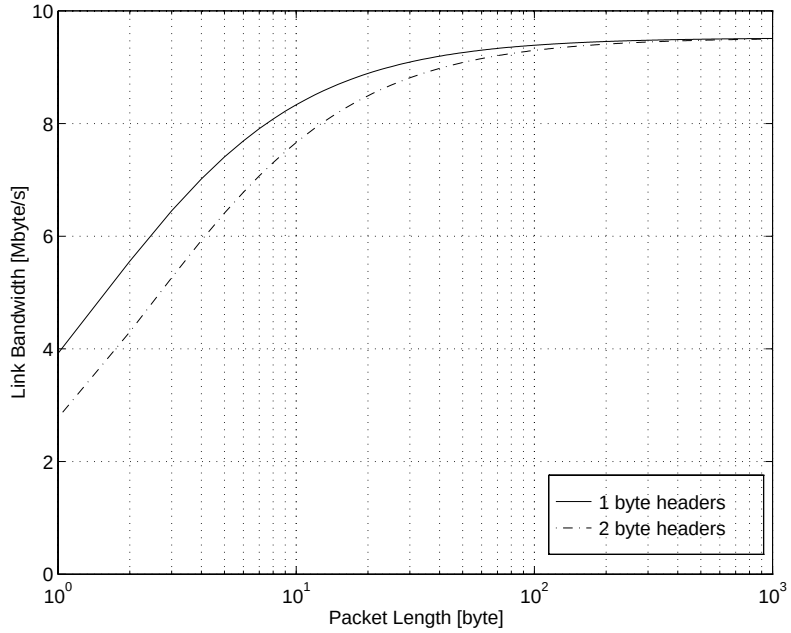


Figure 10: Calculated Bidirectional DS-Link Bandwidth

2.4.3 Effect of Link Length on Bandwidth

The credit based flow control mechanism of DS-Links imposes a limit over which the maximum link throughput can be maintained. It depends on the amount of buffering available at the receiver and the delay of the physical medium connecting the link.

Consider a data source connected to a data sink via a link of a given length. After link start-up the transmitter acquires an initial credit corresponding to the receive buffer size which is equivalent to the number of FCC characters sent from the receiver to the transmitter. The receiver writes the characters in its receive FIFO buffer, from where they are immediately read out by the sink. When one flit has been consumed, the receiver sends another FCC to the transmitter. The condition for continuous transmission is then that the FCC arrives at the transmitter before its credit has been used up. This is illustrated in Figure 11 which shows a time-line diagram of the exchange of characters for the flow control mechanism.

The buffering available at the receiver in excess of the minimum of one flit must therefore allow for the round trip time, which is twice the line delay plus any latencies inherent in the link interface. The maximum link length L_{max} as a function of the receiver FIFO size N_{fifo} can therefore be calculated as follows:

$$L_{max} = \frac{T_{char} \cdot (N_{fifo} - N_{flit}) - T_{lat}}{2 \cdot v} \quad \textbf{(3)}$$

where N_{flit} is the flit size, T_{char} is the time to send one character, v is the signal propagation velocity on the line, and T_{lat} is a latency time which accounts for the delays inherent in the implementation of the link interface. To sustain the full throughput for a 100MBaud DS-Link, buffering for about one character per 10 meters is needed in addition to the requirements for flow control and latencies, assuming a propagation velocity of 5 ns/meter on the physical transmission medium.

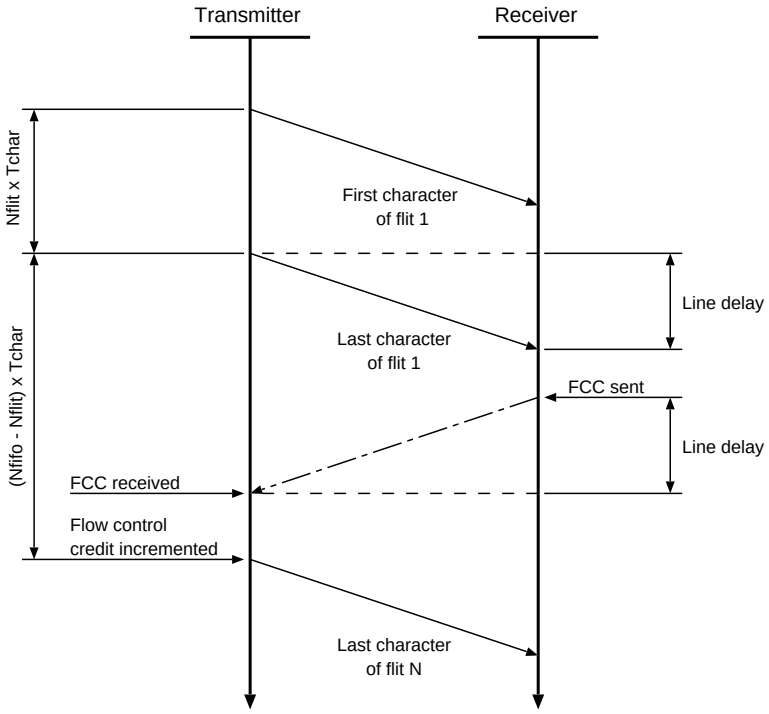


Figure 11: Maximum Transmission Length Calculation

Using Equation 3 and Figure 11 the normalized link throughput as a function of the link length has also been calculated. The size of the receive buffer of the standard DS-Link macro-cell implementation used in the STC101 and the STC104 devices is 20 characters. The latency in the link interface has been estimated to four character transmission times. Figure 12 shows the result for 100MBaud and 200MBaud DS-Links.

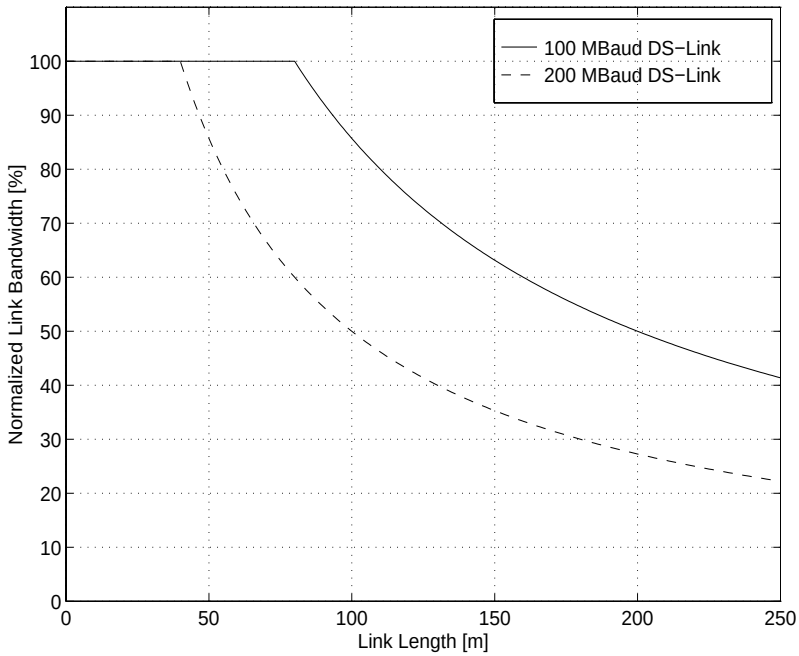


Figure 12: Normalized Link Bandwidth as a function Link Length

For 100MBaud links, the throughput starts to roll off at about 80m. In practice there is also some latency associated with the front-end circuitry of the DS-Link, such as buffers. Allowing for delays in buffers leads to the conclusion that 50m would be a suitably conservative figure. The results for 200MBaud show that faster links need more buffering to achieve the same link length, since more bits are “in flight” on the wire at a given time. These results have implications for the design of the fibre optic DS-Link extension presented in Chapter 5.

2.5 Summary

The IEEE 1355 standard has been introduced and the logical layers and physical media defined in the standard have been presented. The DS-Link point-to-point serial interconnect technology was then explained in detail, since this is the technology on which the work of this thesis is based. Integrated circuits supporting IEEE 1355 serial link technology were listed and the functionality of the devices which were used extensively, the STC104 32-way packet switch and the STC101 DS-Link adaptor, was described. Finally, the theoretical maximum bandwidth of the DS-Link has been calculated, and the effects of link length on the performance of the link have been shown.

Chapter 3
Electrical DS-Link Transmission

This chapter reports on the evaluation and the performance of the DS-DE link physical layer. Differential electrical transmission of DS-Link signals over twisted-pair cable has been characterized and the performance in terms of link speed and link length has been measured. The electromagnetic susceptibility of differential DS-Link transmission has also been tested. The results of this work contributed to the IEEE 1355 standardisation process.

3.1 Single-Ended DS-Links (DS-SE)

Single-ended DS-Links (DS-SE) are intended for short point-to-point connections between devices on the same printed circuit board or between chips on different boards through a connector, i.e. via a backplane. CMOS signal levels and source termination are used for the DS-SE drivers in order to reduce the power dissipation. This enables chips with a large number of links to be built. The transmission line termination is done on-chip, by careful control of the output impedance of the link drivers. Traces longer than a few centimetres should be treated as transmission lines, i.e. the printed circuit board layout has to provide an impedance matched connection. DS-SE links use the Data-Strobe encoding technique, which provides a skew tolerance of nearly one bit-period between the two signals. This simplifies board layout, since the track lengths are not required to match exactly. If the signals are taken through a connector, e.g. onto a backplane, then care must be taken to avoid impedance mismatches. Where possible, traces should be kept to a minimum length. The maximum trace length depends on how well the transmission line impedance can be controlled. As vias and connectors produce discontinuities in the trace impedance, 20 to 30 cm is a practical limit for standard PCB technology. In a 50Ω environment, buffers must be used, since the available DS-Link device can only directly drive 100Ω transmission lines.

3.2 Differential DS-Links (DS-DE)

The DS-DE physical layer provides rack-to-rack and cabinet-to-cabinet communications using differential pseudo-ECL signalling over shielded twisted-pair cable. The logical signalling is identical to the DS-SE link, but differential electrical signalling is used for improved signal integrity. Differential transmission avoids problems with different ground potentials between chassis connected by a cable several meters long, since any difference in ground voltage will be seen as a common mode signal by the receiving differential buffer.

Differential signalling requires a single-ended to differential transceiver. The transmitter has TTL inputs and pseudo-ECL outputs, and the receiver converts the pseudo-ECL back to TTL. The devices which were used are manufactured by AT&T [27]. These devices are specified for operation up to 200MHz, which is equivalent to a maximum bit rate 400Mbit/s. The

receivers also feature a large common-mode range of about 8V, with the 1V signal swing centred within this range.

A single DS-DE link connection needs 8 wires, as shown in Figure 13 below. A twisted-pair cable with four individually shielded pairs and an overall shield is used. The nominal impedance of the pairs is 95Ω. The overall cable shield minimizes emissions, the individual shields for each twisted-pair reduce crosstalk. A 10-pin shielded board-mountable connector [28] manufactured by Harting is used. Harting also produce cable assemblies of different length.

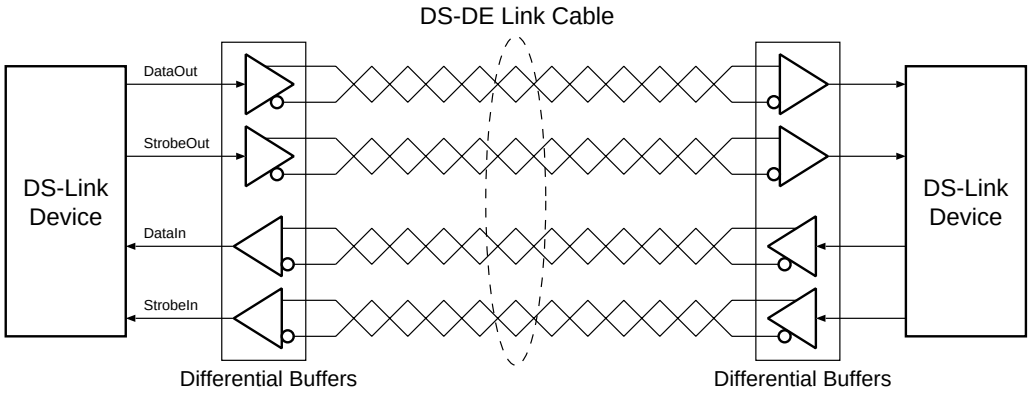


Figure 13: Differential DS-DE Link

3.2.1 Limitations of Cable Transmission

There are a number of factors that limit transmission speed and link length in a cable transmission system. Only the intrinsic effects such as crosstalk, skew, attenuation, and jitter are considered here. There are also external effect such as electromagnetic interference, which will affect the performance and reliability of the link. Susceptibility to external interference will be studied in section 3.4 below.

3.2.1.1 Crosstalk

Crosstalk is due to capacitive and inductive coupling of signals between lines. Excessive crosstalk can cause glitches on adjacent signal lines. Glitches are especially dangerous in the case of DS-Links, as the signal layer protocol operates on the sequence of edges, rather than on levels. A glitch will therefore potentially cause a link failure, the most likely effect being a parity error. Some tests were carried out and showed that, with the twisted-pair cable used, crosstalk is negligible for links of up to 30 meters.

3.2.1.2 Skew

The data and strobe signals will undergo different delays when transmitted over a longer length of cable. The differential drivers and receivers will also introduce a small amount of skew between the data and strobe signals. Excessive amounts of DS-signal skew will eventually cause the link to fail, because the timing requirement for the DS-Link input edge resolution is violated. The input edge resolution specifies the minimum delay between consecutive edges on data and strobe, so that the receiver can still recover the encoded clock and data correctly. Initial tests have shown, that the cable and transceiver skew are not the limiting factors, even for cables of up to 30 meters.

3.2.1.3 Jitter

Jitter is defined as the deviation of the signal transitions from their nominal position in time. This causes a timing uncertainty when recovering the data from the DS-Link signal. Excessive amounts of jitter will cause the link to fail. Therefore jitter has a strong influence on the achievable link length and speed, as explained below. It is caused by the low-pass characteristics of the transmission medium, by differences in the rise and fall times and propagation delays of the active components, as well as by noise.

The dominant jitter component here is data-dependent jitter, which is caused by the lowpass characteristics of the cable. The effect of random-jitter, caused by noise in the active components and duty-cycle-distortion, caused by different propagation delays for low-to-high and high-to-low signal transitions, is small.

3.2.1.4 Effect of Cable Attenuation

The cable attenuation limits the amplitude of the differential signal available at the receiver. Cable attenuation is a function of cable length and signal frequency. This places an upper limit on transmission speed and distance. Attenuation (in dB) can be considered to increase linearly with cable length. Due to the skin-effect loss, attenuation per unit-length is approximately proportional to the square root of the signal frequency:

$$A \sim l \cdot \sqrt{f} \quad \textbf{(4)}$$

where A is the cable attenuation, f is the signal frequency, and l is the cable length.

The limiting factor for the link length is not the attenuation itself, but the variation in attenuation as a function of frequency. Higher frequencies are attenuated more than lower ones. This characteristic of electrical links causes the wider pulses, i.e. sequences of consecutive zeros and ones in the data stream, to have a higher amplitude than the shorter pulses, since the higher frequencies, which are attenuated the most, are required to produce fast signal edges and narrow pulses, while the wider pulses contain more low frequency components. As the cable length increases, the difference in amplitude between short and long pulses increases, and eventually the signal does not cross the receiver threshold any more for short pulses. This variation in amplitude also results in variations in pulse timing, since the edge rate is almost constant and the variation in amplitude causes variations in the time at which a transition will cross the receiver threshold. This effect is known as data-dependent jitter.

The maximum length of a DS-DE link can be estimated as follows: the maximum base frequency in a DS-Link bit stream is half the Baud rate for an alternating sequence of ones and zeroes, i.e. 50MHz for a link operating at 100MBaud. Assuming that the receiver threshold is centred between the low and high levels, the minimum amplitude for the signal still to cross the threshold must be half the peak-to-peak signal swing, i.e. the maximum allowable attenuation at 50MHz is 6dB. The twisted-pair cable used is specified for an attenuation of 0.45dB per meter at 50MHz, the maximum link length at 100MBaud is therefore about 13 meters. In order to ensure reliable link transmission, a margin has to be added to account for second order effects, such as noise, crosstalk, EMI[1], and signal degradation due to reflections.

1. Electromagnetic Interference

3.3 Evaluation of Twisted-Pair Cable Transmission

The purpose of this evaluation was to study the limitations of the differential DS-Link transmission system in terms of bit rate and link length. Another motivation was to test the reliability of a DS-DE link, in order to verify the assumption that the link can be considered virtually error free, on which the IEEE 1355 protocols are based.

3.3.1 Test Setup

A test environment was constructed, consisting of two test boards coupled by a chosen length of cable and driven by a bit error rate tester (BERT), as shown in Figure 14. The BERT consists of a pattern generator and a receiver with an error detector. The pattern generator is used to create a pseudo-random binary sequence (PRBS) or short, user programmable, bit patterns. The receiver compares the incoming bit stream to the pattern that was transmitted and counts the number of bits which are in errors. Timing measurements were made with a 1 GSample/s digital oscilloscope.

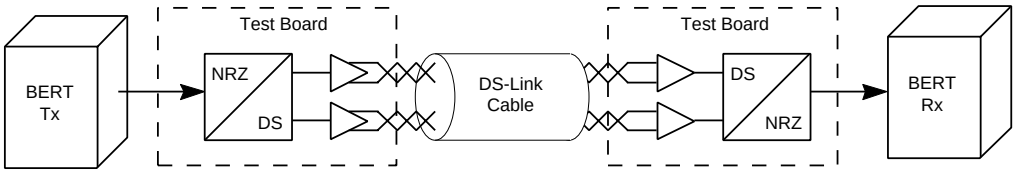


Figure 14: DS-DE Test Setup

The bit-error rate tester cannot directly generate a DS-Link encoded bit streams. Therefore the NRZ[2] bit stream from the BERT transmitter was converted into the DS-Links bit level protocol by a simple encoder on the test board.

3.3.2 Eye-Diagram

The eye-diagram or eye-pattern is a useful method for observing the effects of signal quality degradation. It results from the superposition of the waveforms for all pattern combinations over a single bit interval. The eye-opening is defined as the boundary within which no waveform trajectories lie, i.e. the region of the eye where the correct bit value can be regenerated. A large eye-opening means that the system has a greater tolerance for amplitude and timing variations, and therefore potentially a lower bit-error rate. In a bandwidth limited communication channel, the pulses of consecutive bits will overlap, thereby effectively reducing the eye-opening and generating data dependent jitter. This effect is called intersymbol interference. The amplitude of the eye-opening is also affected by noise and crosstalk, while the time-domain eye-width is reduced by jitter.

The eye-diagram can be measured directly using a digital sampling scope triggered by the serial bit clock. The scope is placed into static persistence mode, in which the traces from all trigger events are accumulated on the display. The oscilloscope method is simple and provides comprehensive plots of the eye, but can usually not display events occurring at a low probability, i.e. low error rates.

2. Non-Return to Zero

Figure 15 shows the differential receiver input eye-diagram for a 15 meter cable at 100 MBaud.

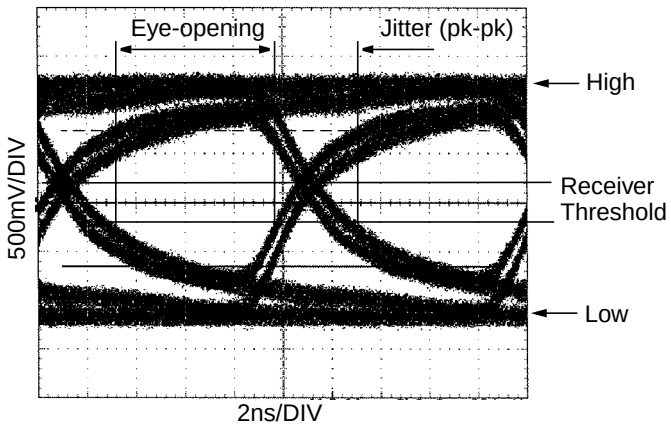


Figure 15: Receiver eye-diagram for a 15 meter cable at 100MBaud

The height of the eye-opening is large compared to the receiver input threshold. The traces are only slightly blurred by noise and jitter. The lowpass characteristics of the cable clearly causes the trailing edge of the preceding pulse to extend into the current bit period, thereby reducing the effective eye-opening. However, the time-domain eye-opening is large, over 60% of the bit period of 10 ns, which will allow for low error rate transmission.

For comparison, Figure 16 shows the eye-diagram at the receiver end of a 30 meter cable. In this case the height of the eye-opening is only marginally larger than the receiver threshold. It is clear that reliable transmission is not possible for this cable length at 100 MBaud. This is confirmed by the bit-error rate tests in Section 3.3.3 below.

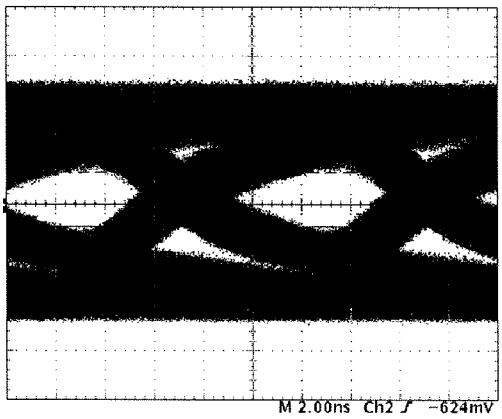


Figure 16: Receiver eye-diagram for a 30 meter cable at 100MBaud

3.3.3 Bit Rate versus Cable Length

In order to determine an upper limit for the transmission speed, we measured the bit rate at which link errors start to occur. The results were obtained by slowly increasing the bit rate until bit errors occurred frequently. Figure 17 shows a plot of the results. Tests were carried out with the setup shown in Figure 14 above. Cables of 1, 10, 15 and 30 meters were tested. Bit rate and error rate tests were also carried out using early silicon implementations of the

DS-Link interface, the so-called Link Test Chip, which was provided by the semiconductor manufacturer.

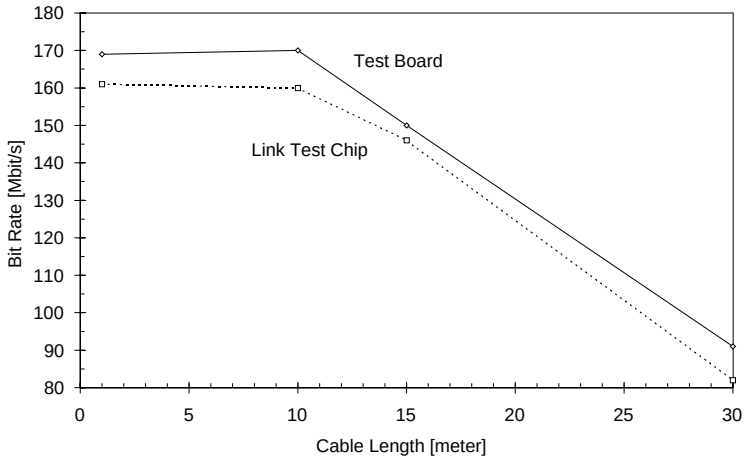


Figure 17: Maximum Bit Rate versus Cable Length

For short cables of 10 meters and below, the data rate is limited by the test board and the bit-error rate tester, which only works up to 175MBaud. For longer cables, the achieved bit rate decreases almost linearly with cable length. For cables up to 15 meters long, the measured maximum bit rate is 45% higher than the target bit rate of 100MBaud. For a 30 meter cable however, the maximum bit rate was already below 100 MBaud. The bit rate values measured using the Link Test Chip are somewhat lower due to imperfections in the early implementation.

3.3.4 Bit Error Rate Test

The previous measurement provides an upper bound for the bit rate, but does not give information on long term reliability of the link. The bit-error rate is a useful measure for characterizing the reliability of a transmission system. It is defined as the ratio of the number of bits in error to total number of bits transmitted:

$$BER = \frac{N_{Error}}{N_{Bit}} \quad \textbf{(5)}$$

The bit-error rate can be used to calculate the expected time between errors, i.e. the mean time between failure (*MTBF*) of the serial link:

$$MTBF = \frac{1}{BER \cdot BR} \quad \textbf{(6)}$$

where *BR* is the bit-rate of the link.

In order to evaluate the reliability of the DS-DE link transmission, the bit rate was reduced to 95% of the maximum bit rate value measured above and long term bit error rate tests were performed. These measurements produced no errors over the measurement intervals of 12 hours (overnight run) to 64 hours (measurement over the weekend). This corresponds to a bit-error rate of less than $5 \cdot 10^{-13}$. The actual error rate can only be determined by prohibitively long test runs, e.g. assuming an error rate of 10^{-15} at 150 MBaud, it would then take on the

average 2000 hours to detect one error. However, when the link is running at the target speed of 100 MBaud, the bit rate is about 30% lower than the maximum bit rate measured above, and we can therefore assume that the intrinsic error rate of the link will be negligible, due to the increased margins.

3.3.5 Summary of DS-DE Link Evaluation

DS-Link signal transmission using differential pseudo-ECL line drivers and receivers at 100 MBaud was shown to work reliably up to 15 meters. The tests were conducted in a laboratory environment, i.e. in the absence of external interference. The maximum bit rate measured for a 10 meter cable was 165 MBaud, which shows that there are considerable timing margins. Transmission speeds of 200 MBaud could not be reached with our test setup, not even with a short one meter cable. This was due to limitations of the test equipment, the bit-error rate tester only works up to 175 MBaud, and because of imperfections in the early prototype silicon implementation of the DS-Link.

3.4 Susceptibility to Electromagnetic Interference

The tests and measurements presented in Section 3.3 were conducted in laboratory environment. However, errors in an electrical link operating well within the speed and distance margins are mainly caused by external noise sources. The reliability of differential DS-Link transmission in the presence of electromagnetic interference is therefore studied in this section.

3.4.1 Interference Problems with DS-Links

Industrially produced hardware and equipment designed in-house using DS-Link technology had been integrated into a data acquisition system of the CPLEAR experiment at CERN [29]. Data from the experiment was sent from a VME crate to be processed by a farm of T9000 transputers in a separate enclosure. The different pieces of electronic equipment were connected using differential DS-Links over screened twisted-pair cable.

It was observed that the system would fail at the rate of one or two failures in a twenty-four hour period. Link failure was identified as the cause. Near the racks with the equipment there was a cabinet that housed the relay switch for an air conditioner. It was therefore assumed that the EMI generated by the fast switching transients was affecting the link performance. The failures could be reproduced in the laboratory by using different noise sources such as switching on and off fan trays, tape drives, etc.

Initial tests showed that grounding clearly affected the magnitude of the problem, and different screening methods and earth attachments were tried. The two most effective ad hoc methods to reduce susceptibility were to either connect the incoming link cable screen to the on-board digital ground or to slip copper foil between the DS-DE connectors and the metal front panel they protruded through. This, however, tells nothing quantitative about the achievable margin of security nor of the real failure mechanism. Therefore, it was necessary to undertake a systematic study of the problem with the aim of being able to issue guidelines for successful link operation.

3.4.2 Packaging

Each enclosure in the experimental installation had different degrees of electro-magnetic shielding and different implementations of screen and earth grounding. VME modules in the experiment were mounted in so-called “open-rack” format, which has no shielding and poor front panel to earth grounding. The earth connection of the power supply will typically go to a star point on the chassis to which the various power supply grounds will also be brought.

The connection between the DS-Link cable shield and the chassis can only be made via the front panel. This can be achieved either by using an EMI gasket which completely surrounds the connector and couples it electrically to the front panel or via the printed circuit board (PCB). The PCB must make an electrical connection with the front panel with a low inductive path, which is electrically and mechanically difficult to achieve. In addition, front panels are often anodised, which makes the surface essentially non-conducting, preventing a good connection with the EMI gasket or bracket. The front panel to ground impedance is also a source of problems. It should be mentioned that VME chassis are available with EMI protected chassis and front panels, but this study restricts itself to the established base of available mechanics.

3.4.3 IEC 801 Standard

There is an international standard for noise susceptibility, IEC 801, which defines electromagnetic compatibility for industrial process measurement and control. Part 4 of the standard covers electrical fast transient/burst requirements [30]. The object of the standard is to establish a common and reproducible basis for evaluation of electronic equipment subject to repetitive fast transients on supply, signal, or control lines. The defined test is intended to demonstrate the immunity of the instrumentation when subjected to transient interference such as that originating from the switching of inductive loads, relay contact bounce, etc.

The interference is simulated by a generating a high-voltage pulse with well defined characteristics, such as peak voltage, rise time and pulse width, which is coupled to the equipment under test in a well defined way. The pulse has a fast rise-time of 5 ns, which generates a frequency spectrum with significant power levels at high frequencies. In practice the pulse is generated by charging up a capacitor to some threshold voltage and spark discharging it to ground through a load. The resulting discharge burst is capacitively coupled to the equipment under test. This discharge can be injected into the devices power supply or coupled to a cable using a capacitive “clamp” of specific dimensions.

The standard limits the result of a test to a threshold classification, declaring that a tested device is resistant to one of four classes of environment:

Level 1. A well protected environment in which all power cords are screened and filtered, any switching devices in the room are fully spark emission suppressed, ac and dc lines are physically separated etc. This is the equivalent of the electrical “clean room” such as a computer room.

Level 2. A protected environment characterised by partial spark suppression on relays, physical separation of unshielded power supply cables from signal and communication cables, and no coupling to higher level noise environments. This corresponds to a control or terminal room of an industrial or electrical plant. This would be a minimum requirement for our applications.

Level 3. A typical industrial environment, which is characterised by no spark suppression in relays, poor separation from higher level noise sources, poor separation between power supply, control, signal and communication cables. Passing this threshold is considered highly desirable for general purpose use.

Level 4. A severe industrial environment such as external high-voltage switch yards. Electrical DS-DE links are not expected to function in such an environment. Fibre optics would be the medium of choice here.

The classification is according to the maximum peak noise generator pulse amplitudes that the equipment can successfully withstand. There are two sets of values, one that applies to noise coupled into power lines and one for signal and control lines, which are shown in Table 2 below:

Table 2: Noise pulse peak voltages

Level	Power Supply [V]	Input/Output Lines [V]
1	500	250
2	1000	500
3	2000	1000
4	4000	2000

3.4.4 Test Setup

The test set-up defined in the standard consists of an electrical fast transient/burst generator (EFT/B), which produces the noise pulse, a capacitive coupling clamp, which couples the pulse onto the signal or power supply lines, and the equipment under test (DUT).

The capacitive clamp must be of standard dimensions. The I/O cable to the equipment tested should pass through the clamp, the equivalent coupling capacitance of the clamp should be about 50pF. A reference ground plane is required for all the equipment. The devices and the clamp are fixed on insulating supports above the reference plane. All the devices should be grounded according to instructions from the manufacturer. A test bench conforming to these requirements was established in the laboratory and is shown in Figure 18 below.

The noise generator was the NSG222 Interference Simulator from Schaffner. This instrument has three settings allowing to generate pulses with nominal peak voltages of 500V, 1000V and 1500V. This allows a test for level 2 and level 3 on I/O lines. The maximum pulse amplitude of 1500V does not meet the requirements of level 4, but it allows for a measure of the safety margin for level 3 operation. The noise generator is connected to the capacitive clamp with a short coaxial cable.

A simple test board was designed, which allowed different circuit and grounding configurations to be compared. It was installed in a VME chassis, the equipment under test, which was grounded with a short metal braid to the ground plane. A sampling oscilloscope and a bit error rate tester (BERT) were used to perform the measurements. The error rate tester was used to generate the serial DS-Link bit stream and monitor it for errors. The instrument was the same as the one used for the DS-DE link evaluation presented in section 3.3. It has only one channel and can therefore not emulate a full DS link, which uses two signals in each direction. The programmable pattern feature, however, allows the generation of IEEE 1355 bit sequences up

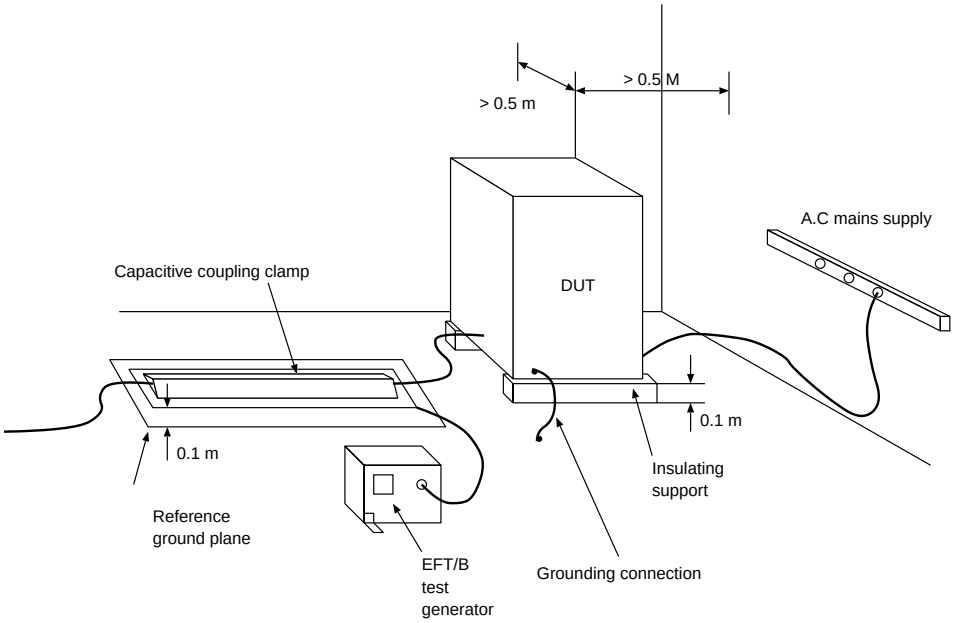


Figure 18: Electromagnetic susceptability test setup

to a maximum bit rate of 50Mbit/s, so that the individual driver/receiver pairs could be tested with typical signal patterns.

3.4.5 The Failure Mechanism

Figure 19 shows an equivalent circuit of the test environment in which a high voltage pulse is coupled to the shield of the differential DS-Link cable. Both the transmitter and the receiver are powered from their respective digital power and grounds. The chassis of the equipment are connected to the earth ground. The cable screen ground is considered as being connected through some non-zero impedance to the chassis screen. The interference pulse propagates through the capacitive clamp to the cable screen ground. In addition, conducted and radiated interference is also introduced into the chassis screens, the power cables, the power and ground planes, or the tracks on the printed circuit board, as shown by the dashed lines in Figure 19 below.

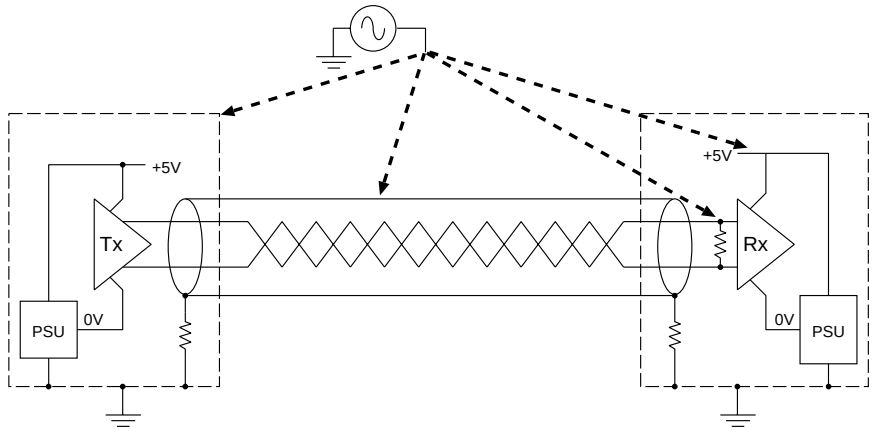


Figure 19: Noise coupling mechanism

There are several possible failure modes to be considered:

1. A braided screen cable is not totally impervious to radiated noise due to the gaps in the braid. The twisted pair picks up energy via the cable transfer impedance and this is coupled as common mode voltage to the receiver [31].
2. Energy from the capacitive clamp is picked up by traces on the poorly screened board near the receiver which degrade the signal.
3. Noise is picked up in the power or ground planes on either the transmitter or receiver boards. This may shift the transmitted output signal levels or reduce the sensitivity of the receiver.
4. The non-zero impedance of cable screen to ground will allow the screen potential to rise and increase the coupling of the interference onto the inner pairs.

Experiments with the noise generator showed that pick-up through the AC power cables of the power supplies had little or no effect on the overall susceptibility and so attention was concentrated on the signal cable and housing issues.

3.4.6 Test Board

A simple test board was designed, which allows different circuit and grounding configurations to be compared. The following approaches to handle the EMI problem were implemented:

1. Common mode noise can be reduced by providing a low impedance path for it to ground. Referring to Figure 20 below, the centre point of the receiver termination resistor usually has no AC component, because as one of the differential lines changes state from zero to one, the other does the opposite and the two cancel each other out. Common mode noise however, is the same on both lines, and the centre point will reflect this. Adding a capacitor at this point provides a low impedance path for common mode AC to ground.
2. Connecting cable screen to digital ground reduces the overall screen to ground impedance, since there is now another path for screen noise to earth via the power supply. This is also shown in Figure 20 below. The reduced impedance reduces the maximum voltage swing and hence the rate of change of voltage that causes the induced stray voltage. Doing this however, violates the standards employed by several equipment manufacturers not to couple screen and digital ground.
3. If common mode should be the dominant failure mechanism, then this can be addressed by the use of common mode chokes at the receiver and/or the transmitter ends of the connection as shown in Figure 20 below.

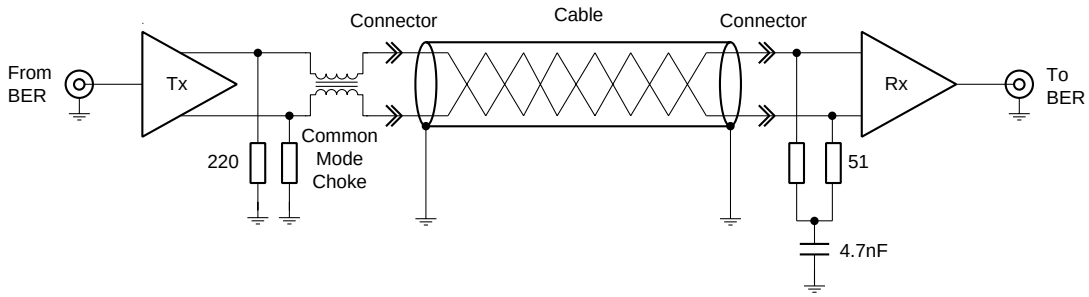


Figure 20: Test board configuration

The same signal is injected equally onto both of the lines and a symmetric common mode excursion is observed. However, doing this results not only in common-mode at the receiver but also in a differential noise signal. The source of the differential signal can be seen in Figure 21, which is the scope trace of one of the wires of the twisted pair measured at the receiver.

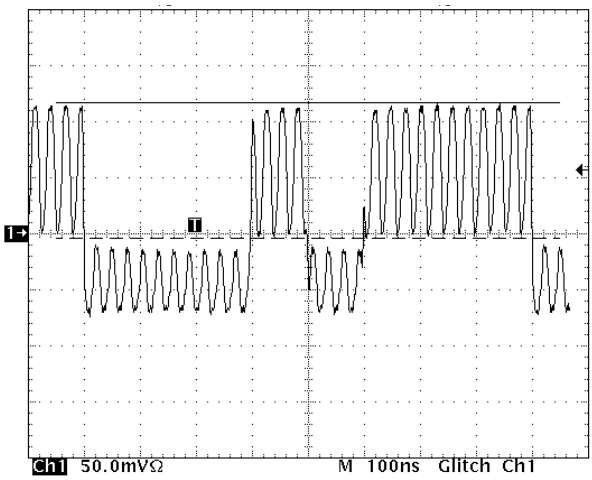


Figure 21: Asymmetric noise pick-up

The trace shows the data pattern of 1000101110, on which is superimposed the 10MHz common mode signal. The amount of pick-up in the high state is nearly twice that in the low state. Thus common mode noise converts to differential noise. This sets an upper limit for noise susceptibility as being less than the tolerable common mode excursion which is from -1V to +7V.

The reason for this behaviour is, that the large signal output impedance of the output stage emitter follower is different for the logical high and low states. This difference in impedance generates a small differential error signal. This causes bit-errors on the link for large common mode excursions, even before the common mode range of the receiver is exceeded. For even larger excursions both outputs will reverse bias and the differential signal will disappear completely. This information was passed on to the manufacturer of the differential buffers, AT&T Microelectronics, who simulated the transmitter and receiver characteristics under the conditions just described. Their results confirmed the measured data and the interpretation.

In an attempt to reduce the common-mode impedance, a split termination with the centre point tapped and decoupled was included at the receiver on the test board. The high common mode output impedance of about 200 Ohms is thereby reduced to less than 50 Ohms. In addition, if possible, the shield should be terminated to the board ground rather than through the chassis, in order to further reduce the effective screen to ground impedance, so that the screen grounding is such that the common mode voltage induced does not cause link failures.

3.4.7 Results and Recommendations

Using the test setup described in Section 3.4.4, the different configurations on the test board were evaluated. The default configuration, i.e. without the common mode choke or the screen connected to the logic ground, did not even pass the level 1 test. Adding the common-mode choke and the centre-tap capacitor at receiver termination significantly improved the perform-

ance, enabling operation at level 3. Connecting the cable screen to the logic ground, the board passed the test for level 3 even without the common-mode chokes. In conjunction with the choke, this configuration allowed level 3 operation with considerable margin, i.e. with a 1500V noise pulse peak voltage.

The tests also showed, that a common-mode choke at the receiver end is ineffective, whereas a choke on the transmitter end is as good as having chokes at both ends. This is consistent with the mode of failure discussed in Section 3.4.6.

The only condition permitting level 3 operation without the screen connected to ground is when the chokes are used in conjunction with the centre-tapped termination.

The manufacturers of DS-Link based equipment expected to function within level 2 of the IEC 801-4 standard should either:

- Bring the cable screen to the signal ground.
- Employ a common mode choke at the transmission end of the cable.
- For VME based equipment the new generation of screened EMI compliant VME racks are to be used, where possible, and great attention paid to the front panel grounding.

The difficulty of dealing with noise once inside an equipment housing means that every effort should be made to keep it outside. The DS-Link connector from Harting has a very high packing density but does not easily allow the shell to be connected to a panel through which it passes. The use of the mounting shell and gasket for this connector is therefore recommended.

This study was prompted by a failure of prototype equipment in a moderately noisy electrical environment. This environment is by no means unique nor exceptional. Other electronic equipment both commercial and in-house continued to function without problems under the same conditions. Following the recommendations in the IEC801-4 standard resulted in a credible benchmark for evaluating IEEE 1355 technology for electromagnetic susceptibility. It has been possible to compare the effects of noise on different systems and from there to deduce those strategies that are most effective.

3.4.8 Summary on EMI Susceptability of DS-Links

It was observed that the differential link connections on some of the DS-Link based equipment which was being delivered to CERN and used in the CPLEAR experiment were susceptible to electromagnetic interference. It was imperative to understand the cause of these failures and try to find solutions to the problem. Tests on the faulty equipment eventually showed, that the problem was due to common mode limits being exceeded as a result of poor grounding of the cable screen. A test bed was established according to the IEC 801-4 standard for EMC susceptibility to provide some quantitative measurements of the magnitude of the noise problem and explore ways to reduce or eliminate the effects. A set of recommendations was put together as a project internal document [32] and also incorporated into the IEEE 1355 standard.

3.5 Summary and Conclusions

Extensive tests and measurements have shown, that differential DS-link connections over twisted-pair cable running at 100 MBaud can be very reliable over distances of up to 15 meters. In addition, good engineering practice was established to allow reliable link operation even in the presence of electromagnetic interference.

This is a fundamental result for the construction of the large DS-Link based network testbed described in chapter 6, which relies on differential DS-Links for connections between modules in different racks.

Link connections longer than 15 meters can be realised by using the fibre optic DS-Link interface described in chapter 4.

Chapter 4
Fibre-Optic DS-Link Transmission

Single-ended DS-Link signals are limited to connections on a printed circuit board or between boards over a backplane. As shown in Chapter 3, transmission over a maximum distance of up to 15 meters can be achieved using differential signalling over shielded twisted-pair cable. For even longer connections or electrically noisy environments, optical fibre is the transmission medium of choice. A new line encoding scheme, which enabled DS-Link protocol to be carried over an optical fibre connection, had been proposed for inclusion in the IEEE 1355 standard. In order to validate the proposed encoding scheme, a prototype implementation of a point-to-point fibre optic connection for DS-Links has been designed and characterised. The results of this work carried out by the author have contributed to the IEEE 1355 standardisation process.

The requirements for the fibre optic link are summarised below:

- The DS-to-fibre interface should be transparent for the application, i.e. the fibre optic link should behave the same way as a DS-link which is connected directly.
- The fibre optic link should not significantly reduce the data bandwidth of the DS link connection.
- The maximum link length for differential transmission over twisted pair cable is barely adequate for connecting racks in close proximity. For ease of system integration, the fibre optic link should enable connections over distances that are at least a factor 10 longer than those possible with DS-DE links.

4.1 Fibre-Optic Transmission System

Optical fibre has a number of advantages over copper as a transmission medium. These are summarized below:

- High bandwidth: the multimode fibre used has a bandwidth length product of 400 MHz·km;
- Low attenuation: 4dB/km for the fibre used;
- Non-conductive: avoids problems with common-mode voltages, ground-loops, or conducted electromagnetic interference.

The main disadvantage is the higher cost compared to electrical connections, due to the cost of the optical components and the interface circuitry which is necessary. A typical fibre-optic transmission system consists of the following components, which are illustrated in Figure 22 below:

Fibre optic transmitter. The transmitter consists of a light source and a driver circuit. An LED operating in the 850 nm wavelength range is used as the optical light source here.

The driver converts the differential digital input signals to suitable drive currents for the LED transmitter.

Optical fibre cable. The transmitter and receiver are connected by a length of fibre optic cable. Graded index multimode fibre with a 62.5 μm core and 125 μm cladding is used.

Fibre optic receiver. The receiver consists of an optical detector, a PIN diode, which converts the incoming light pulses into current, followed by a transimpedance pre-amplifier. The output of the preamplifier is converted into logic signal levels by a discriminator. The different amplifier stages within the receiver are AC-coupled, which simplifies the implementation.

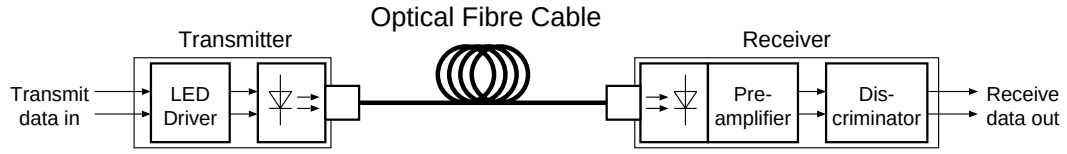


Figure 22: Fibre optic transmission system

4.2 Reliability of Fibre Optic Transmission

The protocols specified in the IEEE 1355 standards are based on the assumption that transmission errors are a very rare occurrence. The results presented in chapter 3 have shown that this is the case for differential electrical transmission over short length of shielded twisted-pair cable. The assumption of very low error rates needed to be verified for the fibre optic transmission system being proposed for the standard.

Noise is one of the fundamental reasons for bit errors to occur in a fibre optic transmission system. It is generated in the active electrical and optical components of the transmission system. The impact of noise on the transmission quality is strongest where the useful signals are small, i.e. in the front-end of the optical receiver circuit. The main noise sources are the optical detector and the receiver amplifier. The probability of bit-errors can be calculated for a simple binary, i.e. two-level, transmission system, under the assumption of noise with a Gaussian probability density function, such as thermal or white noise. This is illustrated in Figure 23, which shows the high and low receiver input signal levels and the noise probability density distributions around these levels.

A bit-error occurs when the receiver interprets a logical zero as a logical one or vice versa. The probability of this event is proportional to the area under the tail of the noise probability density distribution which is above or below the receiver threshold. The error probability is influenced by the width of the noise distribution, which is equivalent to the RMS noise voltage, and the peak-to-peak input signal swing. Using the model illustrated in Figure 23, the relationship between the bit-error rate (*BER*) and the signal-to-noise ratio can be derived [33]. The resulting formula is:

$$BER = \frac{1}{2} \cdot erfc\left(\frac{SNR}{2 \cdot \sqrt{2}}\right) \quad \textbf{(7)}$$

where *SNR* is the optical peak-signal to RMS-noise power ratio and *erfc*() is the complimentary error function. Figure 24 below shows a plot of Equation 7. One notes that the curve is

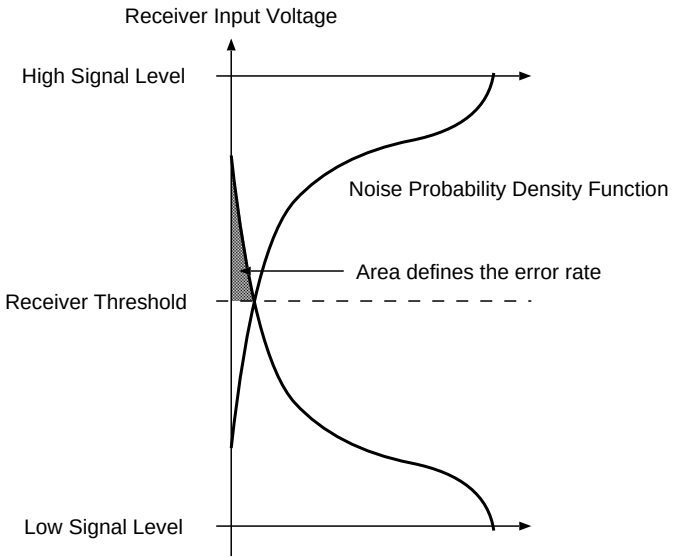


Figure 23: Model for the occurrence of bit-errors in a two-level transmission system

very steep, a change of 1 dB in signal-to-noise ratio at error probabilities around 10^{-15} corresponds to a change of three orders of magnitude in BER. However, the above relationship only remains valid when there are no external noise sources present, which can interfere with the signal in the receiver, thereby further degrading the bit-error rate.

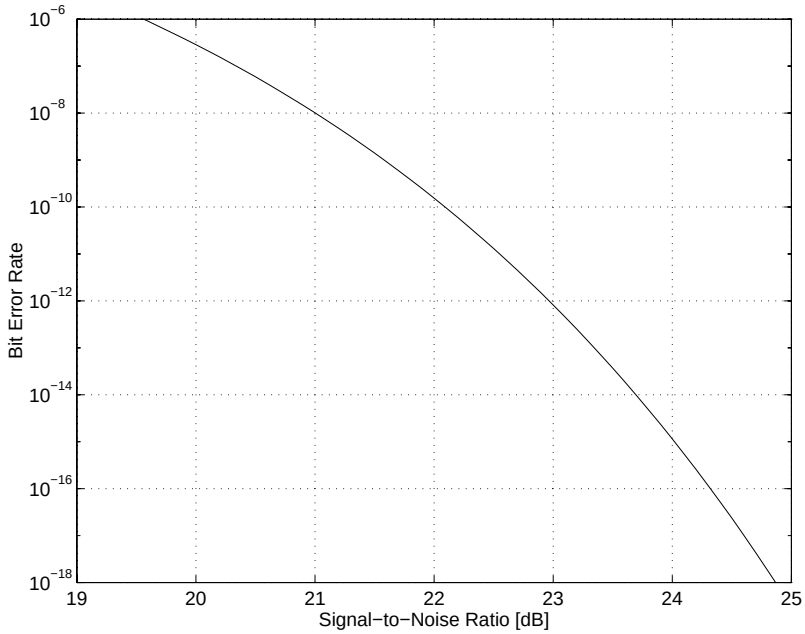


Figure 24: BER versus SNR

Although the attenuation of the optical fibre is very small compared to copper cable, i.e. only a few dB per km, the signal levels in the optical receiver are also quite small, in the range of µA for the optical receiver diode current. This is because of the loss introduced by the conversion from electrical signals to light and vice-versa. Therefore the signals in the front-end of the optical receiver are very sensitive to external noise sources. Possible causes of problems are crosstalk on the printed circuit board, power supply noise generated by high-speed digital

devices, ground bounce, EMI[1]and ESD[2]. Good design practice, such as sufficient decoupling, power supply filtering, and eventually shielding of the sensitive parts of the circuit, will prevent these external disturbances from deteriorating the BER performance of the system. The clock recovery part of the receiver will usually also be sensitive to the effects mentioned above.

4.3 The Transmission Code

DS-link signals are not suitable for direct transmission over optical fibre, because the DS-encoding is not DC-balanced. This makes it impossible to use standard fibre optic transceivers directly, as they always employ some form of AC-coupling in the receiver front-end circuit. In addition, this approach would also require two fibres and the associated fibre-optic transceivers in each direction, one for data and one for strobe. Another problem of two separate fibres would be the requirement for tight control of the skew between the data and strobe signals. A different line encoding scheme is therefore necessary in order to combine the data and strobe signals onto a single transmission channel. The encoding scheme should have the following characteristics:

- Provide 256 data characters plus sufficient number of control characters to map the DS-Link control characters;
- Ensure that sufficient transitions are generated in the serial bit stream to allow the bit clock to be recovered at the receiver;
- Detect single and multiple bit errors that could occur during the transmission;
- Provide distinct and easily recognizable special characters to allow the receiver to acquire and check character alignment;
- Ensure that the serial line signal is DC-balanced to avoid baseline wander in AC-coupled systems [34].

4.3.1 TS Transmission Code Definition

The proposed code is a 4B6B block code, i.e. 4 data bits are mapped onto six code bits. Each of the six bit code symbols consists of three ones and three zeros. Therefore the code is named Three-of-Six (TS) code. Since every symbol has the same number of ones and zeros, the code is inherently DC-balanced, i.e. the signal frequency spectrum does not have a DC component.

There are 20 possible combinations of selecting three bits out of six; all other bit combination are illegal and indicate that an error has occurred. Any single bit errors are thereby detected as code violations. Double bit errors where two zeros are changed to ones or vice-versa are also detected as illegal code symbols. Burst errors, where long sequences of bits are converted to zeros or ones are also detected.

4.3.1.1 TS-Code Symbols

Of the 20 possible 6-bit code symbols, 16 are used for data characters and two are used to encode control characters. All TS-code characters are 12 bits long, each data byte is encoded

1. Electro-Magnetic Interference
2. Electro-Static Discharge

as two six bit TS-code symbols. Table 3 shows the encoding of the 16 data symbols, the least significant bit of the symbol (lsb) is transmitted first.

Table 3: TS Code Symbols for Data Characters

Data (hex)	lsb–msb	Data (hex)	lsb–msb
0	011010	8	001011
1	101001	9	100011
2	011001	A	010011
3	110001	B	110010
4	001101	C	001110
5	101100	D	100110
6	011100	E	010110
7	110100	F	100101

4.3.1.2 TS-Code Control Characters

Control characters use the two special symbols 101010 and 010101. Which of the two control symbols is used depends on the value of the last bit of the previous symbol that was transmitted, such that the last bit of the previous symbol and the first bit of the control character have the same value. This scheme ensures that the control characters contain sequences of more than 8 alternating 0 and 1 bits, which allows them to be easily identified and also enables the character boundaries to be checked, as explained in section 4.3.1.4 below. The control character encoding is shown in Table 4, the EOP_1 and EOP_2 characters contain a checksum of the data characters in the packet, as explained in section 4.3.1.3 below, where the checksum can be any of the valid data symbols from Table 3:

Table 4: TS-Link Control Character Mapping

Control Character	Previous Symbol	Symbols (binary)			
NULL	xxxxx1	101010	101010		
	xxxxx0	010101	010101		
FCC	xxxxx1	101010	010101		
	xxxxx0	010101	101010		
EOP_1	xxxxx1	101010	checksum		
	xxxxx0	010101	checksum		
EOP_2	checksum = xxxxx1	checksum	101010		
	checksum = xxxxx0	checksum	010101		
INIT	xxxxx1	101010	101010	101010	101010
	xxxxx0	010101	010101	010101	010101

The control characters of the TS-code correspond directly to those of the DS-link code, which are defined in Section 2.2.2 on page 9. The EOP_1 and EOP_2 characters are used to mark the end of a packet and the FCC character is used for the link level flow-control algorithm (see also Section 4.3.2 on page 41). As with the DS-Link, the TS-Link transmits NULL characters in the absence of other traffic. This is necessary to maintain the receiver clock recovery PLL in lock. The TS-Link requires the additional INIT control character, which is used during link start-up and to force a disconnect error on the remote side of the link, e.g. when an error is detected, as explained in section 4.3.1.6. The INIT character has as many transitions as possi-

ble, which makes it easier for the clock recovery unit of the receiver to lock to the incoming bit stream during link initialisation.

4.3.1.3 Longitudinal Parity

The end-of-packet characters include a checksum, which is computed as follows: each data byte in the packet is split into two 4-bit nibbles. The longitudinal parity is then computed over these data nibbles. Even parity is assumed. The resulting 4-bit checksum is encoded into the corresponding TS-code data symbol and included in the `EOP_1` or `EOP_2` character, as shown in Table 4 above. On reception, the checksum is decoded and compared with the longitudinal parity check bits computed from the received data nibbles.

The longitudinal parity check enables the detection of double-bit errors within a code symbol, where one bit is turned from a zero into a one, and another is turned from a one to a zero. All single bit errors are already detected as illegal code symbols. Double bit errors where two zeros are changed to ones or vice-versa are also covered by code violations.

4.3.1.4 Character Synchronisation

The encoding of the control characters shown in Table 4 above ensures that the first bit of any sequence of more than eight consecutive alternating bits in the serial bit stream is also the first bit of a control symbol, since the maximum number of alternating zeroes and ones any combination of valid data symbols from Table 3 will produce is 8. If the receiver detects a sequence of more than 8 alternating bits, which does not start on a symbol boundary, an error must have occurred and the character alignment is lost.

As for the DS-Link, character boundaries are initially established during the link initialisation sequence, before any information is transferred, as explained in section 4.3.1.5 below. If the receiver detects a change in the symbol boundaries after the link start-up is complete, this is considered as an error.

4.3.1.5 Link Start-up

Starting a link after power-up or restarting after an error has occurred must be a fail-safe process that takes into account the different timings at each end of the link and the effect that the link length will have on the response times. Figure 25 below illustrates the exchange of tokens which occurs between the two ends of a TS-Link during link start-up.

Each transmitter starts by transmitting INIT characters for long enough to ensure that the clock recovery unit of the receiver at the remote end of the link can synchronise to the bit clock. When the device has been sending and receiving INIT characters for 125 μs, it starts to transmit NULL characters. This indicates that the clock recovery circuit at both ends of the link have locked onto the bit clock and that the receivers have established the character boundaries. When each end has been transmitting and receiving NULL characters for 125 μs, the link is assumed to be ready for data transmission and both devices send the FCC characters for the initial flow control credit. When the device has received at least one FCC character, it is ready to start normal operation, i.e. to transfer data characters.

There is a timeout of 375 μs on the reception of the NULL and FCC characters respectively, i.e. if these characters have not been received within the timeout period, the device reverts to transmitting INIT characters and the start-up procedure restarts. This could happen for instance if one end of the link is reset or when the link is physically disconnected. Using this

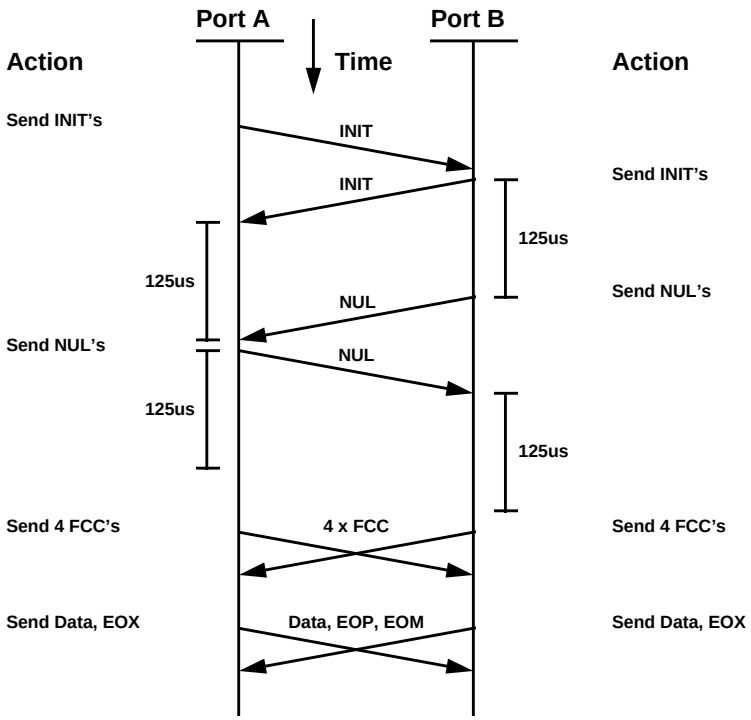


Figure 25: TS-Link Start-up Sequence

scheme, both ends of the link will keep trying to establish a connection until they succeed. Link start-up on power-up or after a disconnection thereby works without the need for any global synchronisation signals.

4.3.1.6 Error Handling

If an error, either a code violation, a longitudinal parity error, or a symbol alignment error, is detected, the packet in transit will be discarded and the TS-FO link will disconnect and re-initialise. This is done by sending INIT tokens. The remote end of the link will detect the INIT tokens and reset itself. Both ends of the link will then go through the link start-up sequence as described above.

4.3.2 Flow Control

The flow control protocol is the same as for the DS-Link, as described in Section 2.2.3.1 on page 10. For the credit-based flow-control scheme the flit size, i.e. the number of tokens sent for every flow control character received, affects the link bandwidth available for data transmission, since a fraction of the link bandwidth is used for transferring the flow control characters. For the DS-link, the FCC character is only four bits long compared to 10 bits for a data character and is sent for every 8 data characters. Therefore the fraction of the link bandwidth used up for the flow control tokens is 4/(8 10+4) = 4.8%. For the TS-code the flow control character and the data characters have the same number of bits, i.e. 12 bits. The flit size for the TS-Link has been increased from 8 characters to 16 characters. The fraction of the link bandwidth used for the transmission of FCC characters is then 16/17 = 5.9% for the TS-Link, compared to 4.8% for the DS-Link. Increasing the flit size also means that the TS-FO Link requires a larger receive buffer. The receiver has to have buffering for at least one flit, i.e. 16

characters in this case. In order to sustain the full link bandwidth, additional buffering is necessary to overcome the delays inherent in the transmission over long distances, as explained in Section 2.4.3 on page 17. A target fibre length of 250 meters implies that about 32 characters are in transit at any given time, assuming a propagation velocity of 5 ns/m on the fibre and a serial line rate of 155 MBaud. Therefore the receiver should provide buffering for at least 48 characters. A 64 character receive buffer is used in the implementation. During link start-up, each end therefore sends four FCC characters.

4.3.3 TS-Link Bandwidth

The DS-Link character encoding is more efficient than the TS-encoding. For normal data characters, the coding efficiency, i.e. the ratio of data bits to code bits, is 80% for the DS-Link and only 67% for the TS-code. In order to achieve the same data bandwidth as a DS-Link, the fibre optic link must therefore operate at a higher line speed. The asymptotic value of the data rate for long packets is 80Mbit/s for a DS-Link operating at 100MBaud, this translates to a required signalling rate of 125 MBaud for the TS-Link.

For short packets, the DS-Link is even more efficient since the end-of-packet character is only 4 bits long compared to 12 bits for the TS-code, where all characters have the same length. In the worst case, assuming very short packets with one header byte, one data byte, and one EOP character, the TS-Link signalling rate would even have to be 1.5 times (36/24) the DS-Link rate.

The bit rate chosen for the implementation was 155.4MBaud and it can therefore support the full data bandwidth of a 100MBaud DS-Link. The clock recovery circuits and the fibre optic transceivers, which were originally designed for ATM applications, were easily available for this speed. Using a different clock recovery device, the design could also run at 125MBaud if required.

4.4 DS-Fibre Optic Link Interface Design

The hardware design of the TS-FO link validation prototype is presented in this section.

4.4.1 Hardware Overview

In order to simplify the development, it was decided to base the implementation of the TS-FO fibre optic link hardware on an existing board which interfaces DS-Links to the PCI bus. This board had been previously developed in collaboration with INFN Rome. It was chosen since it includes an FPGA[3] to interface between a PCI bus controller and the two on-board STC101 DS-Link adapters and it also has a connector for mezzanine cards which brings out all the unassigned pins of the FPGA. A mezzanine card was developed to carry the fibre optic physical layer interface.

Figure 26 shows the block diagram of the TS-FO interface. The design is divided into two parts. The DS-DE physical layer interface and the TS-FO encoder are implemented on a PCI bus card. The fibre optic transceiver, the clock recovery circuit and level converters are located on a small mezzanine card, which is plugged onto the PCI board.

3. FPGA: field programmable gate array

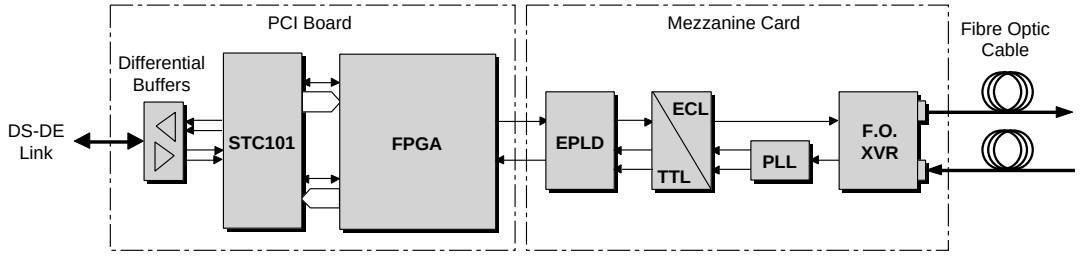


Figure 26: TS-FO Block Diagram

The STC101 DS-Link interface chip is used to convert from the serial DS-Link signal to a parallel interface. Differential buffers are required to convert from single-ended DS-Link signals (DS-SE) to differential signals (DS-DE) for off-board cable connections. The FPGA implements all of the logic required to convert the stream of parallel characters to a stream of TS-encoded bits. The functionality of the FPGA is discussed in more detail in Section 4.4.4 on page 45.

The mezzanine card carries the fibre optic transceiver and a clock recovery PLL to extract the receive bit clock. Level converters are used to generate the pseudo-ECL (PECL) signals used by the fibre optic transceiver and the clock recovery circuit. A small fast EPLD is required to multiplex down the bit stream to a rate that is manageable by the FPGA.

4.4.2 PCI Interface Board

The block diagram of the PCI-DS interface is shown in Figure 27 below. A PCI bus controller chip from AMCC [35] is used to interface to the PCI bus. This chip provides a simple add-on bus interface to the logic on the PCI card. A small serial non-volatile memory device is required to initialise the PCI bus configuration space registers of the controller.

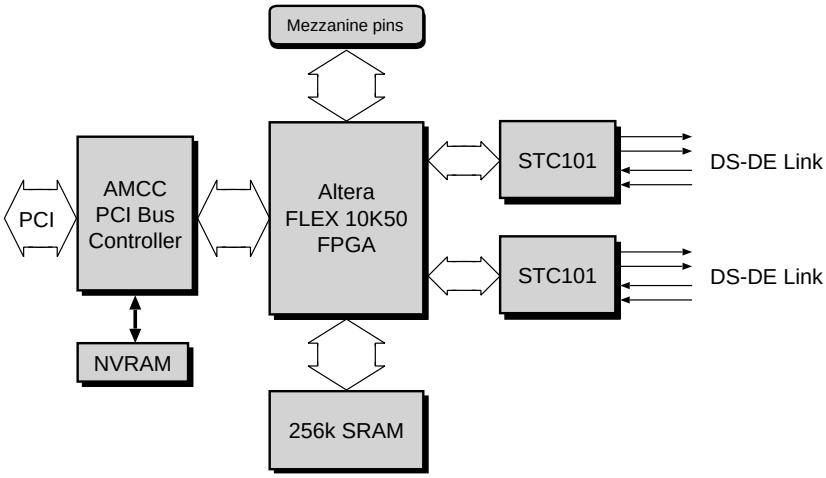


Figure 27: PCI-DS interface board block diagram

All of the local bus interface signals of the PCI bus controller are routed to the FPGA. The FPGA is an EPF10K50 from Altera [36] in a pin-grid-array package with 403 pins. This device provides about 35K equivalent logic gates and 20 Kbit of on-chip RAM memory. Assuming a realistic maximum utilisation of 70% of the logic resources in the FPGA, this

translates to about 25K useable gates. The board contains two STC101 DS link adapters. All of the interface signals from each STC101 are connected to the FPGA. There is also a block of 128kbytes of fast SRAM memory. The address, data and control signals of the memory chips are also connected to the FPGA. All the spare input/output pins of the FPGA are routed to the mezzanine connectors. In addition, all of the interface signals from the second STC101 are also routed to the mezzanine connector. There are between 40 and 100 pins available to the mezzanine, depending on whether one or two STC101s are mounted. The board uses the combination of common mode chokes and AT&T transceiver recommended for reduced susceptibility to EMI in chapter 3 to generate the differential DS-DE signals. Figure 28 below shows a photograph of the interface board.

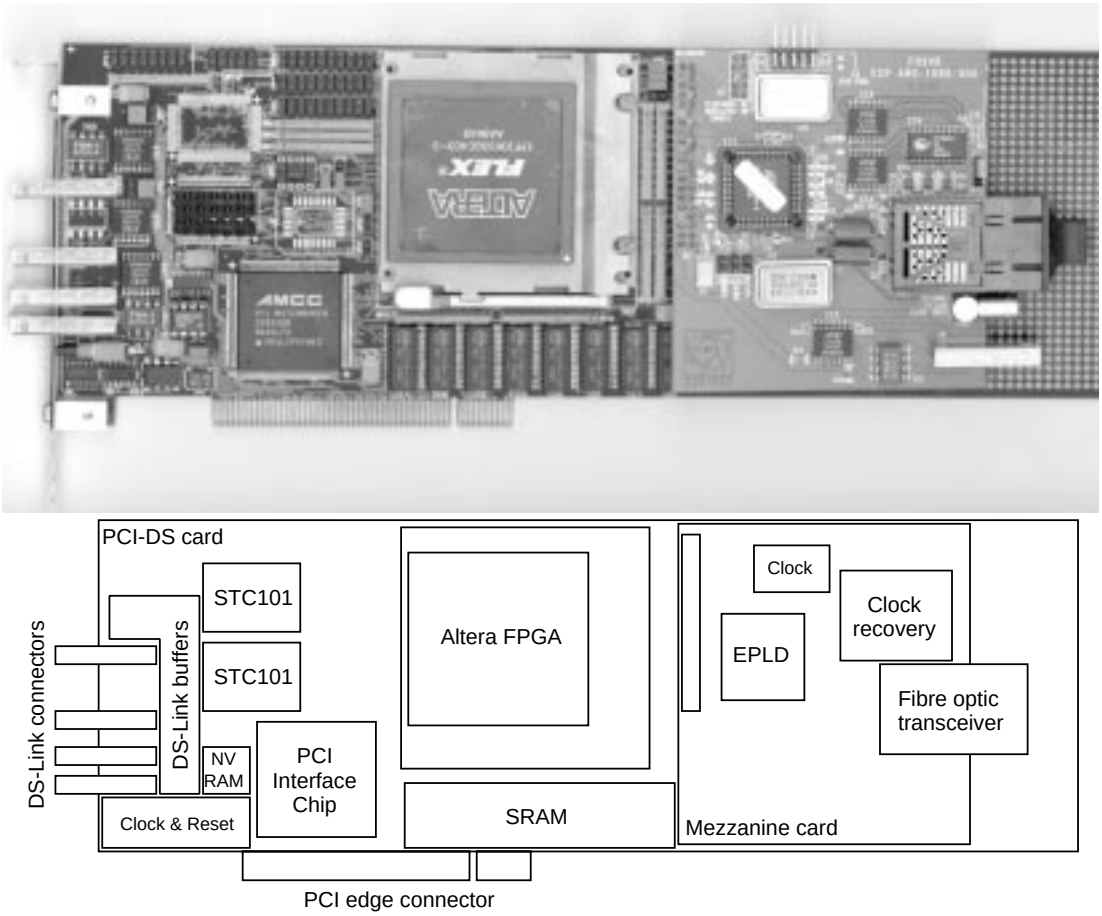


Figure 28: Photograph of the PCI-DS board with the TS-FO mezzanine card

The TS-FO design does not use all the functions of the interface board. The main data path goes from one of the STC101s through the FPGA to the mezzanine card. The external SRAM memory and the second STC101 are not used for TS-FO link interface. The PCI interface part is used to initialise the STC101 registers and to start and stop the DS-Link. A small program runs on the PC which monitors the status of the STC101 and of the TS-Link. If an error is detected, it stops both links and restarts them. This is a simple procedure which could be implemented in the FPGA if required for a stand-alone version of the fibre optic DS-Link interface. The access through the PCI interface was however very useful during the development and test phase of the design.

4.4.3 Mezzanine Board

The mezzanine card carries all the components for the fibre optic physical layer interface. A clock recovery device, using a phase-locked loop, extracts the bit clock and the serial data from the receive line signal. This chip also generates a serial transmit bit clock. The clock recovery device used was originally developed for ATM applications at a line speeds of 155.4 MBaud [37]. It only requires an external reference crystal oscillator. To accommodate different fibre media, a fibre optic transceiver using the industry standard 1x9 single in-line pin-out was chosen. Components that launch into cheap Plastic Optical Fibre (POF), multi-mode fibre or monomode fibres for long haul applications can all be acquired with this pin-out. Figure 28 above shows a picture of the mezzanine card mounted on the PCI-DS interface board.

4.4.4 VHDL Structure

Figure 29 shows the top-level block diagram of the part of the TS-FO interface which is implemented in the FPGA. On the left are the interface signals to the STC101 transmit and receive token interfaces, on the right is the interface to the fibre-optic physical media interface on the mezzanine card.

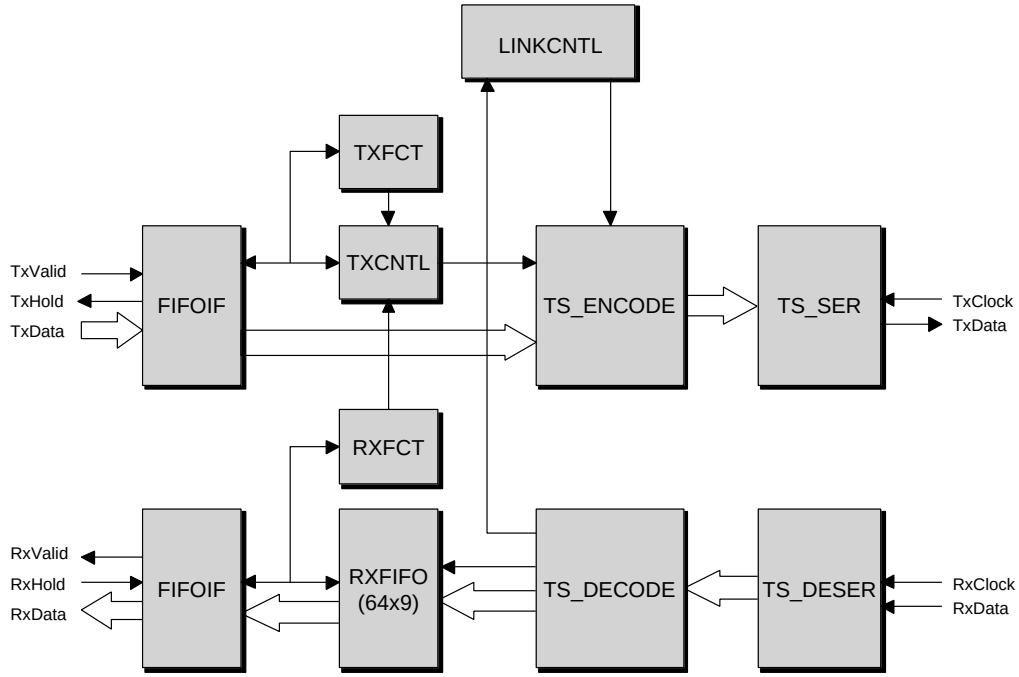


Figure 29: TS-FO Interface VHDL Code Structure

The FIFOIF block interfaces between the token interface of the STC101 and the TS-code transmitter. It is used to convert between the STC101 clock domain and the TS-Link transmit character clock domain. The STC101 token interfaces operate at half the PCI clock of 33MHz, i.e. at 16.5 MHz, while the TS-Link transmit character clock is derived from the serial transmit bit clock; it is 1/12th of the serial bit clock. The FIFOIF block performs the synchronisation between these two asynchronous clock domains.

The stream of data characters from the STC101 is then encoded into 12-bit wide TS-code characters in the TS_ENCODE block. The TXCNTL block controls the operation of the transmitter, e.g. insertion of flow control characters and null characters in the data stream. The TS-characters are then serialised in the TS_SER block. This functional unit also generates the transmit character clock which is used to operate the rest of the transmitter part of the interface. The TS_SER block generates the serial bit stream (TxData), which is fed into the physical media interface. It requires a bit rate clock (TxClk), which is generated by the clock recovery device on the mezzanine card.

On the receiving side, the serial bit stream from the fibre optic transceiver is deserialised in the TS_DESER block. This unit generates 12-bit wide TS-code characters and the associated receive token clock. The character boundaries are also established in this unit. The TS-code character are then decoded in the TS_DECODE block. The TS-decoder also checks for code and parity errors and filters out control characters from the data stream. The decoded data characters are fed into a 64 character deep FIFO, which is required to sustain the link bandwidth for long distance transmission. The output of the RXFIFO block is connected to a FIFOIF block, which interfaces between the internal logic running at the receive token clock rate and the STC101 token interface clock domain.

Link flow control is handled by the TXFCT and RXFCT blocks. The RXFCT unit generates a request signal that forces a flow control character to be sent every time 16 characters have been read from the receive FIFO. The TXFCT unit keeps a flow control credit counter which is decremented each time a data or end-of-packet character is transmitted, and incremented by 16 when a flow control character is received. The transmitter is disabled when the credit counter becomes zero. The LINKCNTL block controls the start-up procedure of the link. It also takes care of error handling.

The VHDL code was synthesised and implemented in an Altera EPF10K50 FPGA. The logic uses about 20% of the EPF10K50 device on the PCI board. This corresponds to about 10K gate equivalents. For comparison, the basic DS-Link macrocell which was developed by SGS-Thomson Microelectronics required about 5K gates in an ASIC. The whole design of the fibre optic link was about 5000 lines of VHDL code, excluding the test-benches to verify the individual blocks.

4.5 Measurements and Results

Results from the evaluation of the multimode optical fibre transceiver and also for the complete TS fibre optic link are presented in this section.

4.5.1 Fibre Optic Transceiver Test

The overall performance of a digital fibre optic link can be determined by stimulating the transmitter with a pseudo-random bit sequence (PRBS) while observing the response at the receiver output. PRBS generators produce a repetitive serial bit pattern which can be easily checked bit-by-bit to determine if any errors occurred during the transmission. A bit error rate tester (BERT) is an instrument which consists of a PRBS pattern generator, an error detector, and an error counter. The bit error rate tester measures the probability of errors on a fibre optic link. The error probability is commonly expressed as the bit error rate, which is simply the number of errors which occurred divided by the number of bits transmitted through the fibre

optic link during the measurement period. Figure 30 shows the setup used to test the fibre optic transceiver. A Wandel&Goltermann bit error rate tester was used to perform the measurements. The fibre optic transmitter and receiver are mounted on a small test board.

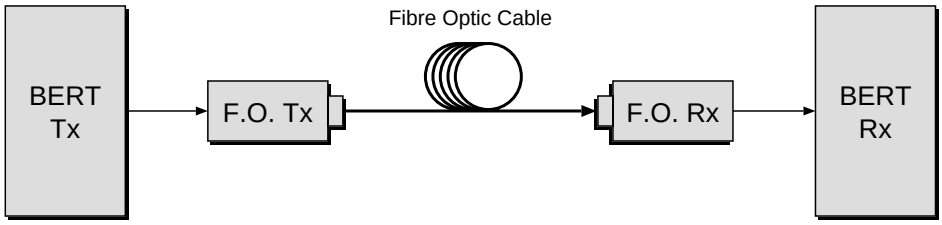


Figure 30: Fibre optic transceiver bit-error rate test setup

Another useful method for evaluating the performance of a fibre optic link is the eye-diagram. It can be measured directly using a digital oscilloscope triggered by the serial bit clock. The scope is set into static persistence mode, where the traces from all trigger events are accumulated on the display. Time domain measurements were done with 4GS/s digitising oscilloscope. This method is simple and provides comprehensive plots of the eye, but can usually not detect events occurring with low probabilities, i.e. at low error rates. The eye-opening gives an indication for the quality of the signal transmission on the fibre optic link. A large eye-opening means that the system has a greater tolerance for amplitude and timing variations, and therefore potentially a lower bit error rate. A large eye-width makes it easier to extract the serial bit clock, which is encoded with the data stream being sent through the serial communication channel, and to synchronously detect the data while it is stable.

4.5.1.1 Fibre Optic Transceiver Test Results

A multimode fibre optic transceiver module from Hewlett-Packard [38] was evaluated to test its suitability for transmission at the targeted bit rate. The transceiver was tested over 200 meters of standard 62.5/125 µm multimode fibre cable using the setup described above. The eye-diagram shown in Figure 31 was measured at the receiver output with a 2^7-1 PRBS pattern at 155.4 Mbit/s. The figure shows that the peak-to-peak jitter is small, only about 0.8 ns. This results in a large eye-opening of 5.6 ns, or 87% of the bit period. This suggests that very low error rate transmission is possible with the chosen fibre optic components at the given bit rate.

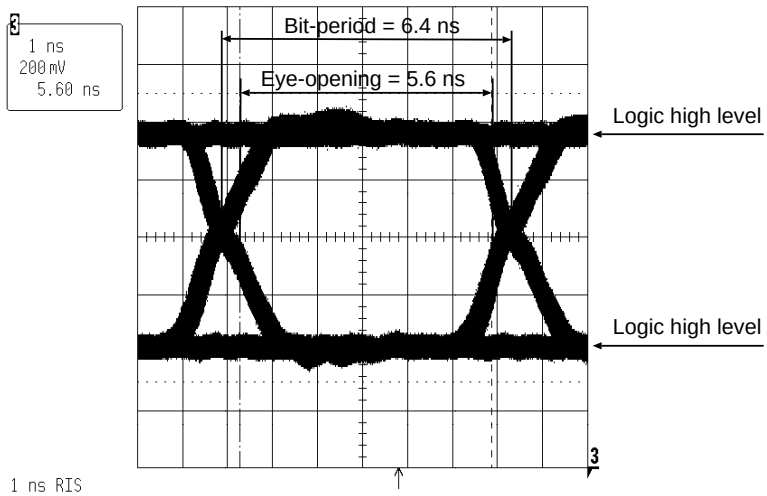


Figure 31: Multimode fibre transceiver eye-diagram

A bit error rate test of the transceiver was also performed at a bit rate of 155.4Mbit/s. No errors were observed during the test run of over 530 hours. This translates to an error rate better than $3.4 \cdot 10^{-15}$, which confirms the conclusion from the eye-diagram measurement. The results show, that the chosen fibre optic components allow virtually error free transmission over 200 meters of multimode fibre.

4.5.2 TS-Link Test

Long-term reliability tests of the complete fibre-optic link were also performed. Figure 32 shows the test setup for the TS-Link test. A pseudo-random number generator implemented in the FPGA is used to produce a stream of characters. These characters are fed into the transmit token port of the first STC101 on the PCI board and transmitted over a 2 metre DS-DE link cable which connects to the second STC101. The character stream is then encoded and transmitted over a length of fibre optic cable. The fibre is looped back to the on-board TS-FO transceiver. The TS-Link bit stream is received and decoded in the FPGA. The STC101 then converts this character stream back into DS-Link signals which are sent over the DS-Link cable to the first STC101. A data sink implemented in the same FPGA consumes the characters and checks for errors. This setup generates a full rate bidirectional character stream on both the DS-Link and the TS-Link.

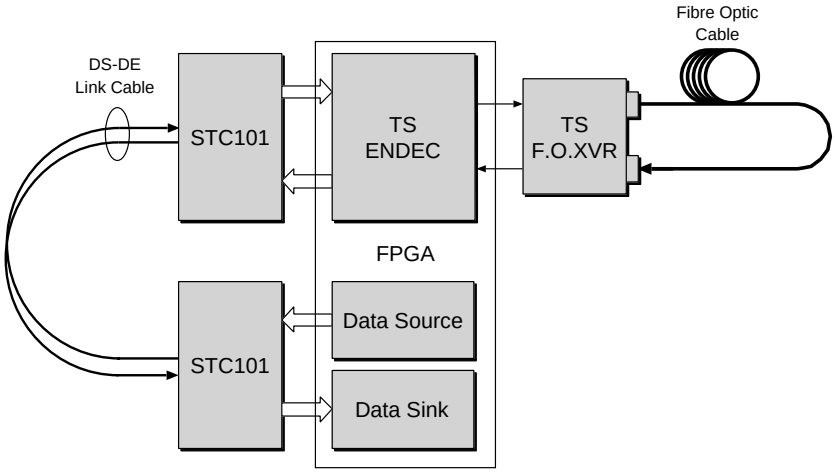


Figure 32: TS-link Test Setup

A long term error rate test run was performed over a period of 220 hours with the TS-Link running over 200 meters of multimode fibre. No errors were observed. This translates to an error rate better than $8 \cdot 10^{-15}$. The error rate for differential DS-Link transmission over 15 meters of shielded twisted-pair cable was demonstrated to be better than $5 \cdot 10^{-13}$ in Section 3.3.4 on page 26.

4.6 Summary and Conclusions

A protocol verification prototype of the TS-FO physical layer of the IEEE 1355 standard has been developed and tested. The fibre optic interface allows extended DS-Link connections over distances longer than the 15 meters possible with differential electrical transmission. This is potentially important in the HEP context, where the cable length from the detector to the electronics can be longer than 50 meters. The interface can also be used where increased

immunity to conducted EMI or electrical isolation are required. The interface supports DS-Link rates up to 100 MBaud. The prototype was used to validate the TS-encoding, which was proposed for the IEEE 1355 standard. This work was partially carried out while the standard was not yet finalised and therefore provided valuable input to the standardisation procedure. The interface was tested over a 200 meter connection and has proven to be very reliable. Part of this work has been published in [39].

Chapter 5
Switches and Networks

5.1 Introduction

This chapter introduces the fundamentals of switching networks. Specific features of the STC104 switch will be explained. The different network topologies that have been studied and the traffic patterns that were used will also be presented. Finally analytical results for the theoretical performance of the basic packet switch will be given.

The switching strategy determines how the data in a message traverses its route through the network. There are basically two switching strategies, circuit switching and packet switching. In circuit switching the path from the source to the destination must be established and reserved first, before the message is transferred over the circuit. Subsequently the connection is then removed. It is the strategy used in phone systems, which establish a circuit through possibly many switches for each call. The alternative is packet switching, in which the end-nodes send messages by breaking them up into a sequence of packets, which are individually routed from the source to the destination. Each packet contains route information examined by switching elements to forward it correctly to its destination. Packet switching typically allows better utilization of network resources because links and buffers are only occupied while a packet is traversing them. We only consider packet switching networks here.

5.2 Switch Architecture

A packet switch consists of a number of input and output ports, buffers to queue packets, and an internal interconnect, which connects the inputs to the outputs. The internal interconnect is typically a non-blocking crossbar. Non-blocking means that any permutation of input and output can be connected, without interfering with each other. However, not more than one input port can be connected to the same output at the same time. Usually the number of output ports is equal to the number of input ports.

5.2.1 Queuing

When more than one packet arrives at the switch inputs destined for the same output, then some arbitration and queuing mechanism has to be implemented in the switch. There are several possible architectures:

Input Buffers. In this configuration the buffers are located at the input ports of the switch. Arbitration logic is needed to determine which of the packets held in different input buffers destined to the same output will be transferred through the interconnection matrix. The arbitration logic can be very simple, e.g. round robin. Input buffered switches are the easiest to implement, as the buffers and the switching fabric only need to run at the speed of the ports. Figure 33 shows a crossbar switch with input buffering.

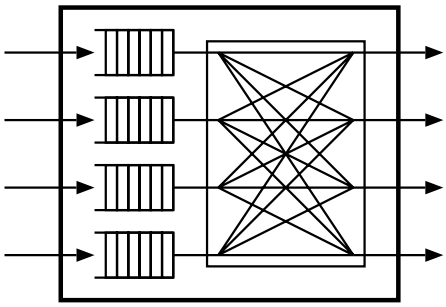


Figure 33: Crossbar Switch with Input Buffering

Output Buffers. With this architecture, the buffers are located at the output ports of the switch element. The assumption is that more than one packet from the input ports can cross the internal interconnection matrix and be stored in the same output buffer. This solution requires the use of a very fast internal cross-connect. In order to allow a non-blocking switch, the interconnection network and the output buffer have to be capable of handling N packets simultaneously, where N is the number of switch ports.

Central Buffer. In a shared-memory based switch, packets arriving at the input ports are written into a central buffer memory, which is shared by all the input and output ports of the switch. When the selected destination port is ready the packets are read from the central buffer and transmitted from the output port. It is clear that for this architecture to be non-blocking, the bandwidth of the central shared memory must be at least twice the aggregate bandwidth of all the ports.

5.2.2 Contention and Blocking

When more than one packet arrives, destined for the same output port, contention occurs. The arbiter in the switch allows one of the packets to proceed, the other packets are blocked and must be queued. The selection algorithm used by the arbiter in the STC104 is round-robin, i.e. each input port is allowed to send the same number of packets. However, if the packets do not all have the same length, this scheme does not guarantee fair sharing of the output link bandwidth.

5.2.3 Head-of-Line (HOL) Blocking

The performance of input-buffered switches suffer from an effect known as head-of-line (HOL) blocking: when two or more packets at the head of the input queues compete simultaneously for the same output, all but one of the packets are blocked. For first-in-first-out (FIFO) queues, which are easiest to implement, the packets behind the blocked head-of-line packet are also blocked, even if they are destined for another output, which is currently not in use. This limits the throughput of this switch architecture to approximately 60%, assuming random uniform traffic (see Section 5.7.1 on page 64). Figure 34 illustrates the concept of head-of-line blocking for a 4 by 4 switch.

Packets from two input ports (2 and 4) compete for the same output port (2). In the first cycle, the packet from one input (2) is allowed to pass through. The packet from the other input (4)

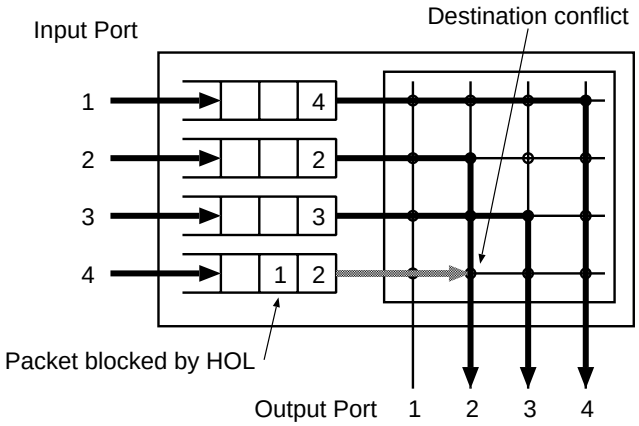


Figure 34: Example of Head-of-Line Blocking

has to wait until the output is free. Meanwhile the packet in the queue behind it is blocked, even though its destination port (1) is currently idle.

In order to reduce the impact of HOL blocking, the STC104 uses a combination of input and output buffering. The STC104 has 70 characters of buffering on each of the 32 paths, of which 45 characters are on the input side and 25 characters are on the output side of the crossbar [26]. In addition the internal non-blocking crossbar operates at 30 MHz, i.e. three times faster than the external link character rate. Consider a packet that is blocked in the input buffer waiting for its selected output to become free. All packets behind it are also blocked. Once the output becomes available, the packet can be transferred from the input to the output buffer at the STC104 core speed, thereby reducing the waiting time for the next packet in the queue. Because of the small buffer size, the advantage of the faster core speed is most significant for short packets, as will be seen from the results in chapter 7.

5.3 Network Performance

In this section the measures of network performance are defined. The performance of a switching network can be characterised by throughput and latency.

5.3.1 Throughput

The per-node throughput is the number of data bytes received or transmitted by an end-node in a given period of time. The header and end-of-packet characters are not counted. The total network throughput is the aggregate transmit or receive rate of all nodes.

Figure 35 shows the measured network throughput as a function of the applied load, assuming uniform random traffic (see 5.4 below). When the requested bandwidth is below that which can be delivered by the network, the delivered bandwidth scales linearly with that requested. However, at some point the network saturates and an increase in network load does not yield any additional increase in the delivered bandwidth. The maximum achieved throughput is also called saturation throughput.

It is useful to define two other performance measures which are related to the network bandwidth:

Bi-sectional Bandwidth. Packets transferred across a network may have to pass through several links, this complicates the measurement of total network bandwidth available. The network bi-sectional or cross-sectional bandwidth is a simple performance measure enabling a useful comparison of different network topologies. It is obtained by cutting the network into two equal parts at its narrowest point, and measuring the bandwidth between the two halves. The bi-section bandwidth is a function of the network topology and the link bandwidth.

Maximum Bandwidth. This is the aggregate network throughput assuming that all the nodes can transmit at the full link bandwidth. It is simply the number of nodes times the link bandwidth. This value can only be achieved for specific traffic patterns, for which there is no contention. The applied network load is expressed as a percentage of the maximum theoretical network bandwidth.

5.3.2 Latency

The network latency is defined as the delay between the head of the packet entering the network at the source to the packet being completely received at the destination. The latency is the sum of three components:

- Switching latency;
- Transmission latency;
- Queuing latency.

The switching latency is proportional to the number of switches a packet has to traverse:

$$t_{Switching} = N \cdot t_S \quad \textbf{(8)}$$

where N is the number of switches a packet has to traverse and t_S is the switching latency of a single switch. The transmission latency is the time to transmit a full packet, including the overheads for the packet header and end-of-packet. The transmission latency is proportional to the packet length, it also is a function of the link speed and the packet overheads. It is related to the link throughput as follows:

$$t_{Transmission} = t_{Overhead} + \frac{l_{Packet}}{BW_{Link}} \quad \textbf{(9)}$$

where the overhead includes the time to transmit the packet header and end-of-packet character. The sum of switching latency and transmission latency, also known as unloaded network latency, is the delay a packet experiences in an otherwise empty network.

The queuing latency is the additional delay a packet experiences, because it is blocked and must be queued waiting for the selected resource to become available. It is a function of the number of switches a packet has to traverse, of the network load and the traffic pattern. As the network load increases, more and more packets get blocked due to contention within the network and the queuing delay increases. A network can provide low latency communication when the requested bandwidth is well below that which can be delivered. However, at some point the network saturates and additional load causes the latency to increase sharply without yielding additional delivered bandwidth. This is shown in Figure 35.

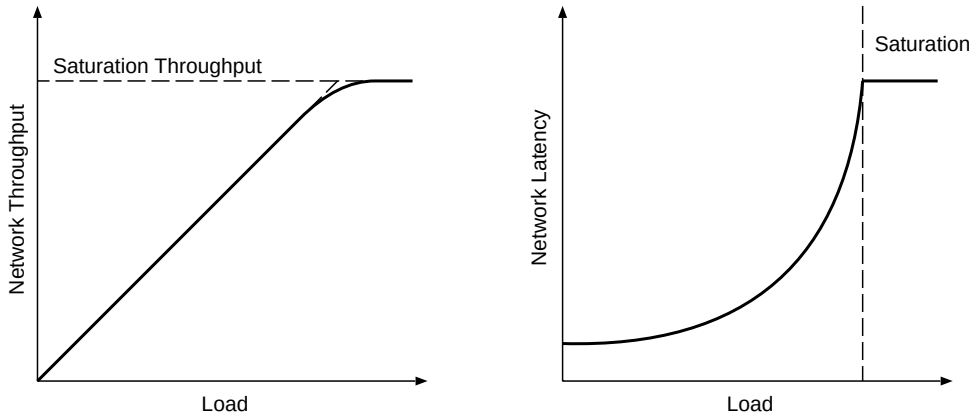


Figure 35: Network throughput and latency versus applied load

The network latency shown is the sum of switching, queuing and transmission latencies. As the network load increases, the queues get longer and the queuing latency increases too. For a system where the queue length is unlimited, the latency tends towards infinity as the network load approaches saturation. However, for the networks studied here, the queue length is finite and the backpressure flow-control mechanism stops any more packets from entering the network as it becomes saturated. Therefore the average latency reaches a maximum value as shown in Figure 35.

5.4 Traffic Patterns

The performance of a switching network will also be a function of the traffic pattern used. The following traffic parameters influence the network latency and throughput:

- Load;
- Packet length;
- Packet destination.

The applied network load is defined as a percentage of the theoretical maximum throughput of the network. The traffic patterns used for the measurements in chapter 7 and 8 are introduced below:

Uniform random traffic. This traffic pattern provides all-to-all communication between the end-nodes. Each node sends fixed length packets to random destinations. The destinations are chosen from a uniform distribution. The special case of a node sending to itself is excluded. The time between sending packets is taken from a negative exponential distribution. The mean of the delay distribution defines the applied network load.

Systematic or permutation traffic. This type of traffic involves fixed pairs of nodes communicating, i.e. the destinations are a permutation of nodes in the network. At the start, every source terminal is assigned a single, unique destination to which all of its packets are sent. Permutation traffic is free from destination contention and will usually give better performance than random traffic. However, not all permutations give the same level of performance, e.g. some permutations may take advantage of locality in the

network. Therefore worst-case and best-case permutation patterns have been tested on the network topologies which are sensitive to the source-destination mapping. In most cases the worst-case permuted traffic pattern represents the permutation which gives the lowest network performance. The actual permutations used are introduced with the network topologies in Section 5.5 on page 56.

Hot-spot traffic. Both random and permutation traffic represent uniform traffic patterns in the sense that each destination receives an equal number of packets. While these patterns provide a good way of characterising and comparing network performance, network traffic for most applications is not completely uniform. The destination distribution can be modified by adding a hot-spot, i.e. a single terminal to which a larger proportion of packets is directed. This can severely degrade network performance because of an effect known as tree saturation. As soon as the link to the hot-spot becomes saturated, a saturation tree builds up, as more and more network resources become occupied, until the performance of the network as a whole is affected.

Fan-in or funnel traffic. A number of sources send traffic to a smaller set of destination nodes. This type of fan-in traffic occurs frequently in data acquisition systems.

Application specific traffic. Traffic profiles as expected for the data acquisition and trigger systems of the next generation high energy physics experiments have also been used. Results for this type of traffic are presented in chapter 7.

5.5 Network Topologies

A network consists of links, switches and terminal nodes. The network topology is the physical interconnection structure. This may be regular, as with a two-dimensional grid, or it may be irregular. We only consider regular network topologies. A distinction is made between direct and indirect networks; direct networks have a terminal node connected to each switch, whereas indirect networks have terminal nodes connected only to the switches which form the edge of the network. The following regular network topologies have been studied and will be presented in more detail below:

- 2-dimensional grids;
- 2-dimensional tori;
- Multistage Clos networks.

The performance of a given network topology can be estimated by considering parameters such as the cross-section bandwidth and the network diameter, i.e. the number of switches that a packet needs to traverse. The traffic pattern and the packet size will also have an impact on the network performance.

It is important to note, however, that performance is not the only driving factor in network designs. Implementation cost and fault tolerance are two other criteria. For instance, the wiring complexity can also become a critical issue in large scale networks.

5.5.1 Direct Networks

Figure 36 shows a 2-dimensional grid with 256 nodes. The network consists of 16 switches which are arranged in a 4 by 4 square matrix. Bundles of 4 links connect adjacent switches horizontally and vertically. Each of the switches has 16 terminal nodes directly attached to it. On the switches at the edge of the grid, some links remain unused.

The other direct network topology which has been studied is the torus. A torus is very similar to the grid with all the edge links wrapped around to connect to the links on the opposite edge of the network. Figure 37 shows a 256 node 2-dimensional torus network.

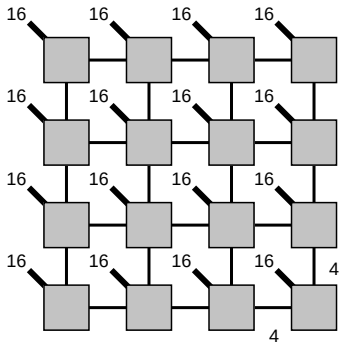


Figure 36: 2-D Grid Network

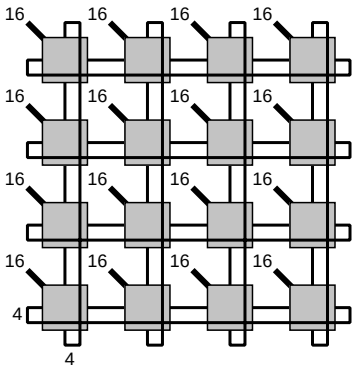


Figure 37: 2-D Torus Network

The cross-section bandwidth of the grid scales with the square root of the number of nodes. This indicates that for uniform random traffic, the achieved per-node throughput will decrease as the network size increases. The average number of switches a packet has to cross from source to destination, the path length, will also increase with the square root of the network size. Since at each switch the packet can potentially be blocked, the average network latency of the grid is expected to be rather high for this type of traffic.

However, the performance will be much better with traffic patterns where communication is mainly between nearest neighbours, since this minimises the path length of the packets and contention. Some applications such as image processing or matrix manipulation on a network of processors will actually create this type of localised communication pattern. The best case permutation for systematic traffic was therefore chosen so that communication is just between the corresponding nodes on adjacent switches.

The worst-case permutation for grid and torus networks is obtained by mapping every router onto its exact opposite in the topology, i.e. mirroring the router across every dimension of the grid, and making source-destination pairs of the terminals connected to the router and its opposite. This maximises the path length for all the packets and creates severe congestion at the centre of the network.

5.5.2 Indirect Networks

A multistage interconnection networks (MIN) consists of several levels of switches. End-nodes are only attached to the switches at the edge of the network. One class of multistage networks is the non-blocking Clos network [40], which was studied here. Figure 38 shows a 128 node three-stage Clos network. All the switches in the terminal stage have 16 end-nodes

attached to them. Each of the terminal stage switches connects with a bundle of four links to each of the centre stage switches.

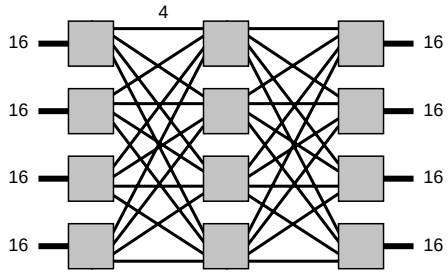


Figure 38: Multistage Clos Network

The cross-section bandwidth of the Clos topology scales linearly with the number of nodes. We can therefore expect the Clos to perform significantly better than the 2-dimensional grids under random traffic. The maximum path-length is 3 switches for Clos networks of up to 512 nodes, the latency should therefore also be significantly lower than on the grid. The Clos can also sustain the full throughput for any permutation of sources to destinations.

These advantages come, however, at the expense of an increased network cost and wiring complexity. From Figure 38 it is clear that the implementation of a Clos is not trivial, since every terminal stage switch must be connected to every centre stage switch. Moreover, the Clos topology does not scale easily in terms of the number of switches required to implement the network. Assuming a 32-valent switch element, a 512 node network can be built from three stages with 48 switches, whereas a 1024 node network already requires a 5-stage structure with 160 switches. Even more switches are required when only a switch element with fewer ports is available, e.g. 448 switches are required for to implement a 512-node Clos based on an 8-way switch.

5.6 Network Routing

The routing algorithm determines which routes the packets follow through the network. The routing algorithms implemented by the STC104 switch are presented here.

5.6.1 Wormhole Routing

Packet switching networks can use two different methods to forward packets:

- Store-and-Forward
- Wormhole or Cut-Through

In a packet switching network using store-and-forward routing, each intermediate switch inputs an entire packet, decodes the header, and then forwards the packet to the next switch. This scheme is undesirable for two reasons:

- It requires buffer space in each switch for the transmitted packets and thereby limits the packet length.

- It causes potentially long delays between the transmission of a packet and its reception, because each switch has to wait for the whole packet to be received before starting re-transmission.

With wormhole routing, sometimes also called cut-through, the routing decision is taken as soon as the header from a packet has been read in by the switch. The header is then sent to the chosen output link and the rest of the packet is copied directly from the input to the output without being stored in the switch. The path through the switch disappears after the end-of-packet character has passed through. This implies that packets can be traversing several switches at the same time. This method can be thought of as a form of dynamic circuit switching, in which the header of the packet, while progressing through the network, creates a temporary circuit, the “wormhole”, through which the remainder of the packet flows. The circuit closes as the end of the packet passes through each switch. Figure 39 illustrates the concept for a packet passing through three switches.

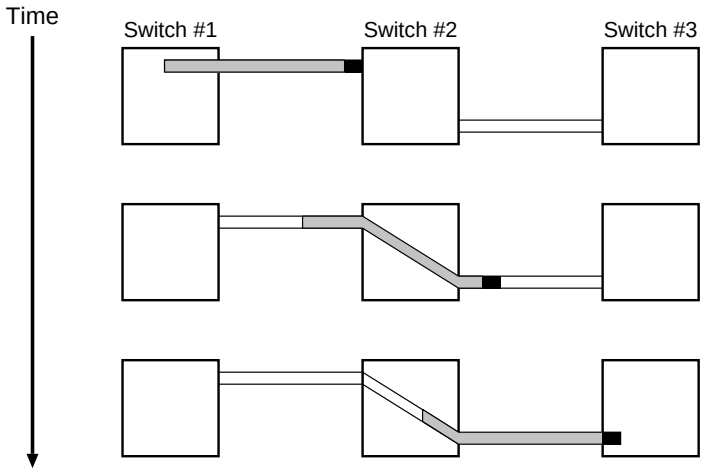


Figure 39: Wormhole routing in a network of STC104 switches

First, the packet header is read and the routing decision is taken. If the output link is free, the packet is sent directly from input to output creating a temporary circuit. As the end of the packet passes through, the circuit vanishes. The packet header enters the next switch before the packet has completely left the previous switch.

Wormhole routing minimises latency and buffering requirements compared to switches using store-and-forward techniques. It also has the advantage that it allows arbitrary length packets. The method provides lower latency than the store-and-forward method as the entire packet does not have to be read in before the connection is set up. Routing latency is significantly reduced by the ability to start outputing a packet while it is still being input. The reduction is particularly significant in lightly loaded networks. The difference in latency between wormhole routing and store-and-forward is illustrated in Figure 40 below.

Apart from minimizing latency, wormhole routing also has the advantage that it is independent of the packet length. In a store and forward system, the maximum packet size must be determined in advance so that buffering can be provided. The delay can be further minimized by keeping the headers short and by using fast, simple hardware to determine the link used for output. The STC104 uses a simple routing algorithm based on interval labelling.

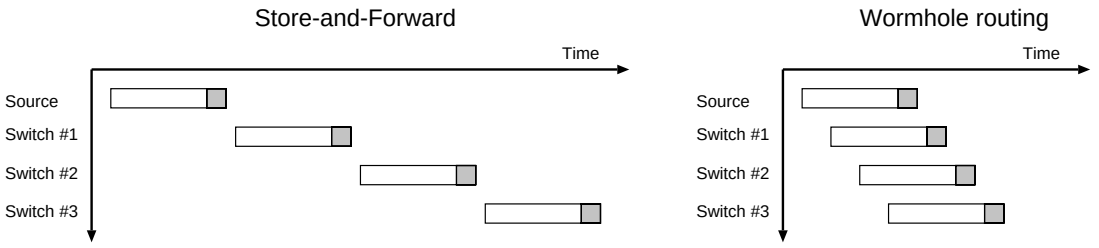


Figure 40: Wormhole routing versus store-and-forward

Note that if a packet is transmitted from a link running at a higher speed than the link on which it is received, there will be a loss of efficiency because the higher speed link will have to wait for data from the slower link. In most cases all the links in a network should be run at the same speed.

5.6.2 Flow Control

Whenever two or more packets attempt to use the same network resource, for instance an output link on a switch, at the same time, the packets which are blocked have to be handled somehow. There are basically three possibilities:

1. The incoming packet is discarded;
2. The incoming packet is buffered;
3. A flow control mechanism stops the flow of packets.

The first of these options is undesirable because it forces the end nodes to engage in complex protocols to deal with the possibility of packet loss. The second option is effectively a return to store-and-forward routing, with the disadvantage of requiring buffer resources in each routing node, which removes the packet length independence of the switching mechanism.

It is therefore clearly preferable to propagate information about a stall back along the path of the packet. The flow control mechanism determines when a packet, or portions of it, move along its route through the network. Since the switch should not have to provide buffering for an entire packet, the flow control system must be capable of stalling the flow of data part way through a packet, which implies that it has to operate on a granularity below that of packets. The smallest unit on which flow control is performed is called a flow-control digit, or flit. In the case of DS-Links, a flit corresponds to 8 characters. With this scheme, when the head of a packet is blocked, the packet body may continue to move until all buffering along the path is filled. The flow control mechanism then insures that no buffers are overwritten. This, however, also means that all links which are still occupied by the packet will be blocked.

5.6.3 Interval Labelling

For each incoming packet, the switch has to decide to which output link the packet should be forwarded. The STC104 uses a routing scheme know as interval labelling [41]. It allows very compact routing tables and can be efficiently implemented in hardware. Each output link is assigned a range, or interval, of destination addresses. This interval contains the addresses of all the terminal nodes which can be reached via that link.

The header of each incoming packet is compared to a set of intervals. The intervals have to be contiguous and non-overlapping, each header value can only belong to one of the intervals. The packet is then forwarded to the output link which is associated with the matching interval. The STC104 has 36 programmable intervals. Figure 41 illustrates the concept: On the left four output links with the destination address values which should be sent down these links are shown, e.g. packets with a header between 4 and 26 inclusive should be sent down link 3. On the right the interval routing table for this configuration is shown. A packet with a header value of 25 arrives, and the header is compared with the entries in the interval table. The matching interval range is 4 to 27 exclusive. The packet will be forwarded to the link associated to the matching interval in the link select table, i.e. link 3 in this case. A link can occur multiple times in the link select table, as shown for link 2 in the example. This allows split intervals to be used.

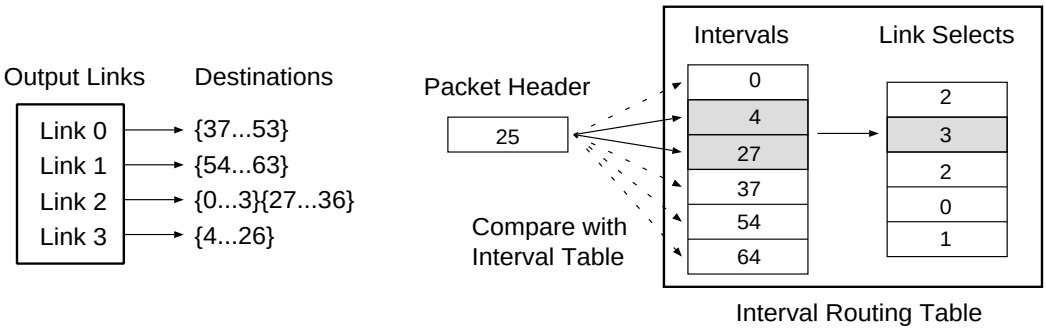


Figure 41: Interval Labelling

It is possible to label all the major regular network topologies such that packets follow an optimal, i.e. shortest, route through the network and such that the network is deadlock free.

5.6.4 Deadlock-Free Routing

An essential property of a communications network is that it should not deadlock. Deadlock is a state where further progress of packets in the network is impossible due to a cycle of resource dependencies. Deadlock is a property of the network topology and the routing algorithm used. It can occur in most networks unless the routing algorithm is designed to prevent it. Figure 42 shows an example of deadlock in a wormhole routing network. In a square of four nodes every node attempts to send a packet to the node at the opposite corner at the same time. If the routing algorithm routes packets in an anti-clockwise direction, then each link becomes busy sending a packet to the adjacent corner and the network becomes deadlocked.

If, however, instead of routing packets in a clockwise direction, all the packets are first routed along a north-south direction and then along a east-west direction towards their respective destinations, all of the packets can be routed successfully since the links are bidirectional. This is called dimension order routing and is the deadlock-free routing algorithm which is used on grid networks.

Optimal, dead-lock free, wormhole routing algorithms exist for grids, hyper-cubes, trees and various multi-stage networks. A few topologies, such as rings, cannot be labelled in an optimal deadlock free manner. Although they can be labelled such that they are deadlock free, this is at the expense of not using one or more of the links, so that the labelling is not optimal [42].

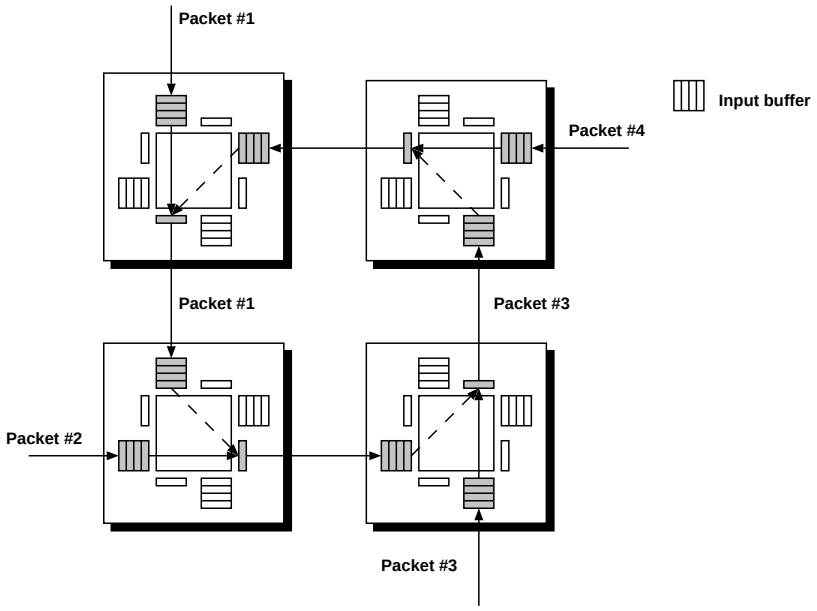


Figure 42: An Example of Deadlock

5.6.5 Grouped Adaptive Routing

The STC104 also supports a feature which allows a programmed set of consecutively numbered output links to be configured as a group. Any packet routed to one of the links in a group will be transmitted along the first link to become free. This locally adaptive routing scheme improves performance by ensuring that there are no packets waiting to use one link when an equivalent link is free. A set of links used to access a common destination can therefore be logically grouped together, increasing the aggregate throughput to the destination. This applies to bundles of parallel links between routers as well as to multistage networks, where grouped adaptive routing allows efficient load-balancing [43]. Grouped adaptive routing also provides a degree of automatic fault-tolerance, since a single point of failure can be avoided by using alternate links [44].

On the grid and torus network topologies, grouped adaptive routing is used on parallel links between adjacent routers (see Figure 36 and Figure 37). For Clos networks, all the links from the terminal stage switches to the centre stage can be grouped, because any terminal stage switch can be reached from any centre stage switch. Parallel links from the centre stage to the terminal stage are also grouped. This is illustrated in Figure 43.

One disadvantage of grouped adaptive routing is that packets can potentially arrive out of order at the destination. This is because two subsequent packets may take different routes through the network, where the first packet could be blocked along its path, while the next packet can potentially proceed along an alternate path without being delayed, thereby arriving at the destination first. Therefore packets may have to be reordered at the destination. Alternatively, a higher level protocol can to be used to insure in-order delivery.

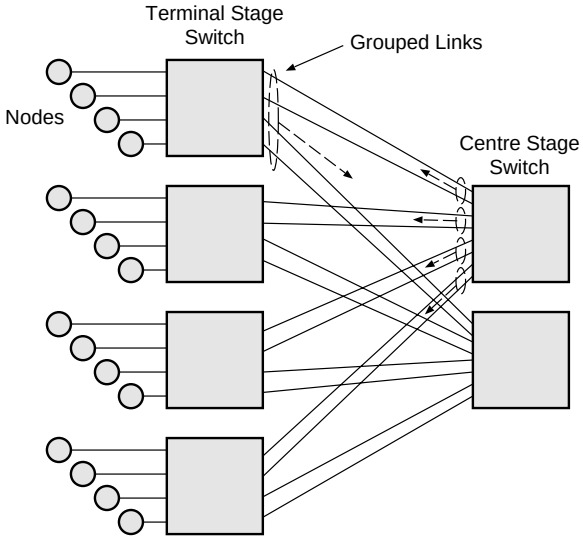


Figure 43: Grouped Adaptive Routing for a Clos Network

5.6.6 Universal Routing

The routing algorithms described so far provide efficient deadlock free communications and allow a wide range of networks to be constructed from a standard router. Packets are delivered at high speed and low latency provided that there are no collisions between packets travelling through any single link. Unfortunately, in any sparse network, some communication patterns cannot be realized without collisions. A link over which an excessive amount of communication is required to take place at any instant is referred to as a hot spot in the network, and results in packets being stalled for an unpredictable length of time. The STC104 also supports a method to eliminate network hot spots called two-phase or universal routing. This involves every packet being first sent to a randomly chosen intermediate destination; from the intermediate destination it is forwarded to its final destination. This algorithm is designed to maximize capacity and minimize delay under conditions of heavy load, by spreading the load across the interconnect. This is at the expense of peak bandwidth and minimum latency under low load [45].

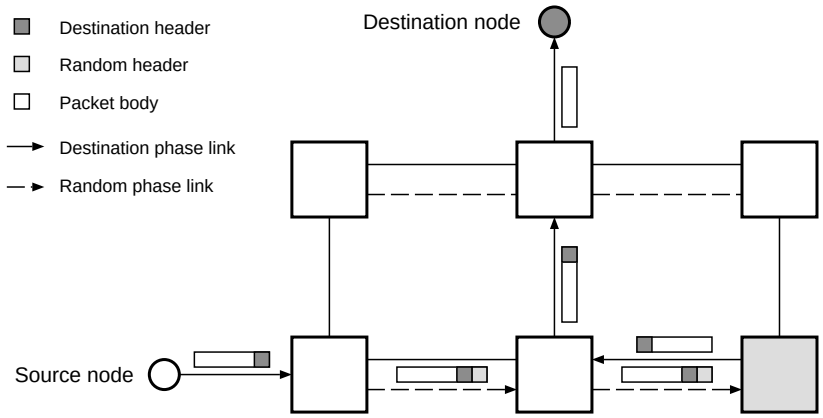


Figure 44: Universal routing

To implement this algorithm the STC104 can be programmed so that some of its inputs (those along which data enters the interconnect) add a randomly generated header to the front of each packet. This header is generated in a range so that it corresponds to a label of one of the other STC104s in the network. The packet is then routed to that STC104 in the normal way, at which point the generated header is recognised and stripped off before the packet is routed further. The packet's original header is now exposed, and so it is routed to its true destination [46]. Figure 44 above illustrates this on an array of 2 by 3 switches.

5.7 Theoretical Switch Performance

With all its links running at 100 MBaud, the STC104 can sustain the maximum link data bandwidth of 9.52 MBytes/s, allowing for all protocol overheads, on all 32 ports simultaneously, or 305 MBytes/s in total. This assumes that there is no destination contention, which would reduce the achieved throughput.

The switching latency of the STC104 can be estimated from parameters of the chip design. The STC104 has two clock domains: the system clock, at which the switching core of the chip runs, and the link clock, which clocks the DS-Link interfaces. The switching latency can be estimated as [26]:

$$t_{C104} = 14 \cdot t_{SystemClock} + 39 \cdot t_{LinkClock} \quad \textbf{(10)}$$

Assuming a system clock of 30 MHz and a link clock period of 10 ns, the switching latency for the case of no contention is 850 ns. Due to the fact that the two clocks are asynchronous, the actual value will vary slightly.

5.7.1 Statistical Analysis of a Crossbar Switch

The throughput of a crossbar switch under random traffic can be estimated by a simple probability model based on independent trials. The model assumes that fixed size packets are sent to uniformly distributed random destinations. Synchronous operation is assumed, i.e. at the start of each cycle every input port selects an output port at random, independently of whether the packet from the previous cycle was transmitted successfully or not [15]. This is not what happens in the STC104; the switch operates asynchronously and when a packet is blocked, it will remain in the input queue.

Under these assumptions, we can assume a binomial distribution for the number of packets destined for a given output port, and the probability that exactly k packets are sent to an output port can be calculated as follows:

$$\beta(k) = \binom{N}{k} \cdot \alpha^k \cdot (1-\alpha)^{N-k} \text{ with } \alpha = \frac{1}{N} \quad \textbf{(11)}$$

where N is the number of input and output ports on the crossbar and α is the probability that a particular source selects a particular destination. The probability that an output port is not selected by any of the input ports is therefore:

$$\beta(0) = (1-\alpha)^N = \left(1 - \frac{1}{N}\right)^N \quad \textbf{(12)}$$

This is the probability of an output port being inactive. The probability of an output port being active is therefore:

$$U = 1 - \beta(0) = 1 - \left(1 - \frac{1}{N}\right)^N \quad \textbf{(13)}$$

This is equivalent to average utilisation of the output port. For a crossbar switch with a very large number of ports, the model gives a maximum achievable utilisation of:

$$\lim_{N \to \infty} 1 - \left(1 - \frac{1}{N}\right)^N = 1 - \frac{1}{e} = 0.632 \quad \textbf{(14)}$$

The utilisation is higher for smaller switches or for a smaller number of active ports, but for a 32-way crossbar the utilization (63.8%) is already very close to the asymptotic value. Figure 45 shows a plot of the utilisation of a crossbar switch under random traffic as a function of the crossbar size.

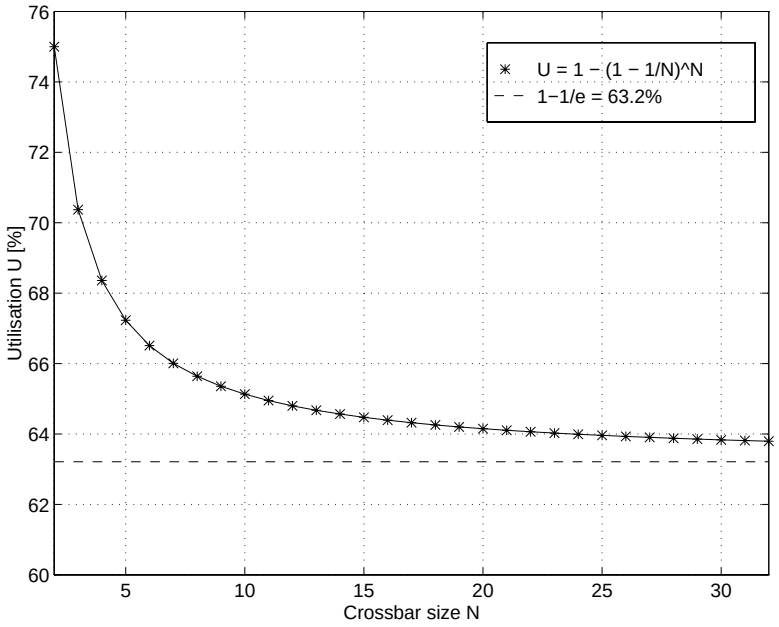


Figure 45: Utilisation of a crossbar under random traffic

Even though the equations presented above evidently assume a much simplified model of the switch, they give reasonably accurate approximations of the measured performance of the STC104 switch, as will be shown in Chapter 7.

5.8 Summary

The most important network concepts have been introduced in this chapter. The network topologies which were tested have been introduced and compared. Finally a model for the performance of a crossbar switch under random traffic has been presented, which shows that the achievable throughput is limited to about 63% due to head-of-line blocking.

Chapter 6
Design and Implementation of a DS-Link and Switch Testbed

6.1 Introduction

This chapter will give an overview of the Macramé large scale IEEE 1355 network testbed. The individual hardware modules used to construct the testbed will be described in detail, a short overview of the software required to operate the testbed will also be given. Finally results from an evaluation of the basic performance of each of the components will be presented.

6.1.1 Motivation

The work presented here was carried out within the framework of the European Union's OMI[1]/Macramé[2] project. One of the workpackages within this project was to construct a large scale demonstrator based on IEEE 1355 DS-Link technology in order to investigate the performance and scalability as well as the robustness and reliability of the IEEE 1355 DS-Link technology and to demonstrate the feasibility of constructing large scale systems. The most important reasons for building this network testbed are outlined below:

- to demonstrate that large scale systems can be built using DS-Link technology;
- to provide performance measurements, since simulation is many orders of magnitude slower (~10^6);
- to show that very low error rates can be achieved; IEEE 1355 is based on the assumption that links can be considered to be reliable;
- to calibrate simulation models, since the results from simulation can only be as accurate as the models used;
- to investigate implementation issues and to establish good engineering practice.

The data acquisition systems of the next generation High Energy Physics experiments are all based on large switching networks [1, 2]. The construction of the Macramé network testbed presents a unique opportunity to test the feasibility and performance of such large networks for application in HEP. Results from this study are presented in section 7.8 on page 116.

6.1.2 Design Criteria

The design goal was to produce a very large IEEE 1355 testbed. A primary requirement was the ability to study different topologies for a large number of nodes. This imposes a system design and implementation which is modular and flexible. Given the available resources,

1. Open Microprocessor Initiative
2. Multiprocessor Architectures, Routers and Modelling Environment, Esprit project 8603

however, the per-node cost also had to be reduced to an absolute minimum [47]. The list below summarizes the other requirements for the testbed design:

- The nodes should attempt to deliver data packets of defined length to a defined address at a defined time, to allow arbitrary traffic patterns to be studied.
- The overhead in sending a packet must be as small as possible in order that the performance of the switching fabric is measured and not the coupling of a node to the network.
- To measure the total bandwidth each node is required to report its transmitted and received data rates.
- To investigate network latency the delay between the transmission and reception of individual packets must be measured.
- The system must be reconfigurable to allow the construction of many different topologies.

6.1.3 Testbed Architecture

In this section the overall architecture of the Macramé network testbed is introduced. The requirements outlined above have been met by building the testbed from three basic components:

- a large number (~1000) of traffic generator nodes;
- a small number (~10) of timing nodes, which are used for latency measurements;
- a switching fabric to interconnect these nodes.

Figure 46 shows in simplified form, how the these basic components are used to construct a given testbed configuration, in this case a 2-dimensional grid network.

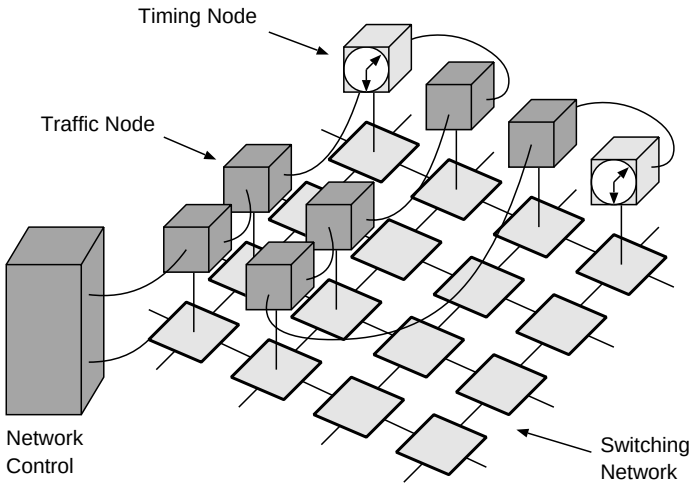


Figure 46: Network Testbed Architecture

The packet switching network is based on the STC104 32-way packet switch. These switches can be assembled into a range of different network topologies. The switch fabric interconnects a large number of traffic nodes, which generate programmable patterns. Each traffic node is capable of simultaneously sending and receiving data at the full link rate of 10 Mbyte/s.

The functionality of the traffic nodes had to be restricted in order to keep the per-node cost low, so that they can only measure the aggregate transmit and receive rates. To perform latency measurements, a small number of timing nodes are therefore used. They transmit and analyse time stamped packets which traverse the network between specific points.

The whole system is controlled via two independent networks, one to configure and monitor the STC104 switches and one to control the traffic and timing nodes. In addition all nodes share a global system clock which they use as a timing reference. This is necessary to maintain synchronism between the traffic nodes and in order to be able to perform accurate latency measurements.

6.2 Network Component Design

This section provides a description of the design of the basic components used in the construction of the Macramé testbed.

6.2.1 Traffic Node

The traffic node is a DS-Link data source that generates programmable network traffic patterns. It drives a 100 MBaud DS-Link and can simultaneously send and receive data at the full link bandwidth.

6.2.1.1 Traffic Node Block Diagram

Figure 47 shows the block diagram of the traffic node. The node consists of the following functional units:

- an STC101 parallel DS-Link adaptor;
- a controller implemented in an FPGA;
- a memory (SRAM) to store the traffic patterns.

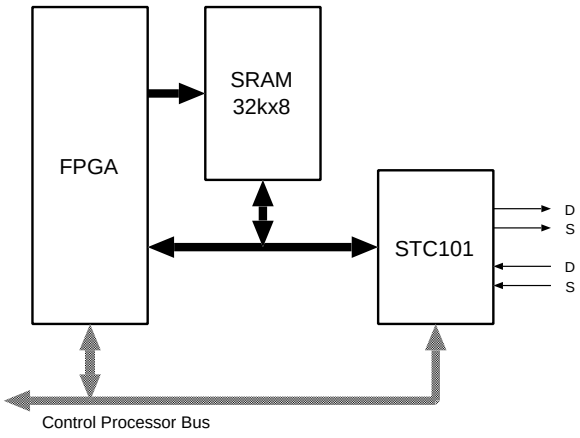


Figure 47: Traffic Generator Node Block Diagram

A series of packet descriptors is used to define the traffic pattern. The packet destination address, the packet length, and the time to wait before dispatching the next packet are programmable. Each traffic node has memory for up to 8k such packet descriptors. The traffic patterns are pre-programmed into the on-board memory. The dispatch algorithm is imple-

mented in an FPGA[3], which fetches the traffic descriptors from the pattern memory and feeds the STC101 with the programmed packets through its transmit token interface. It also handles the queuing required when packets cannot be sent as scheduled due to congestion. The FPGA can be reconfigured under host control, which potentially allows different dispatch algorithms, for example for request-response traffic, to be implemented. The STC101 also receives the data sent to it by other nodes. Incoming packets are consumed at maximum speed. The transmit and receive data rate are measured.

6.2.1.2 Traffic Node Operation

The traffic pattern is stored in memory as a list of packet descriptors. Each packet descriptor consists of three entries:

- Delay
- Packet length
- One or two header bytes

The delay values in the traffic descriptors are relative to the time when the last packet was scheduled to be transmitted. The algorithm implemented in the traffic node controller FPGA is outlined below:

1. Fetch the delay;
2. Wait until the packet is due to be sent;
3. Fetch the packet length;
4. Send a fixed number of header bytes;
5. Send the packet body;
6. Send an end-of-packet character;
7. Goto step 1.

The data characters in the packet body simply have a fixed value (0xFF), since there is no requirement to send specific data. The loop described above is executed for all packet descriptors in the pattern memory. The controller then wraps round and starts again with the first packet descriptor. The number of packet descriptors that can be stored in the traffic pattern memory of the node depends on the number of header bytes that need to be sent. The values are shown in Table 5:

Table 5: Number of Packet Descriptors stored in the Traffic Node

Header length [bytes]	Number of packet descriptors
1	8190
2	6552

The delay value can be zero, the next packet is then sent immediately. If the packet length is zero, no data is sent. This feature can be useful for padding out the pattern memory, so that all nodes stay synchronous, even after the memory address counter wraps round. It can also be used to generate long delays, i.e. longer than the maximum programmable delay time for a packet entry, by splitting the delay up into a number of packet descriptors with a packet length value of zero.

3. Field Programmable Gate Array

The implementation of the simple node also imposes some limitations on the values that can be specified for each of the traffic descriptor parameters. They are shown in the Table 6 below:

Table 6: Value Range for the Traffic Parameters

Parameter	Value range	Unit
Packet delay	0 – 2	ms
Packet length	0 – 4k	byte
Timing resolution	0.5	us
Maximum elasticity	32	ms

6.2.1.3 Packet Queue

The traffic node has to transmit packets according to a predetermined schedule. The incremental delay is defined as the time between the transmission of packets. However, in the case of network congestion, packets may be stalled, so that the STC101 transmit FIFO is unable to accept any further data. To handle this, the traffic node implements a virtual packet queue. Packets are put into the queue when they are scheduled to be sent, but they can only be removed from the queue and actually sent out when the network allows, because of the back-pressure asserted by the link level flow control. If the time when a packet is scheduled to be transmitted is already past, the packet is sent immediately. This elasticity mechanism allows the nodes to stay in synchronization, even when packets are delayed due to network congestion. A timer is used to accumulate the delay from the time when a packet is scheduled for transmission to the time when it is actually sent. If this timer reaches its maximum value, it stops incrementing, and therefore synchronism between the nodes is lost. This can happen if the network is operated near its saturation throughput. The elasticity is sufficient to cope with transient hot-spots though. The algorithm, which is executed for every packet descriptor, is outlined below:

```
WHILE (timer < incremental.delay)
  increment timer
  wait one clock tick
timer := timer - incremental.delay
send the packet (timer keeps incrementing)
```

If the packet is sent on time, the timer value will be zero after the subtraction. In case the packet is sent late, the timer will have a positive value, which corresponds to the amount of time it was delayed by. Therefore the next packet will again be sent on time if possible.

The advantage of this algorithm is that it allows an absolute timing relationship between packets from different nodes even under network congestion. This is necessary in order to emulate High Energy Physics and real-time application traffic, where the transmission of packets from different source need to be synchronised, e.g. in the case of HEP traffic by a particle interaction. A global clock signal is also necessary to maintain synchronism between the nodes, but this is a requirement for accurate latency measurements anyway.

Each traffic node has two status indicators, which can be read by the microcontroller and also drive two LEDs on the front panel of the module. As the network approaches saturation, the traffic nodes cannot send their packets at the specified time intervals. Two possibilities arise corresponding to a transient and a permanent state. In both cases the traffic nodes knows that

it has a back-log of packets to send. In the first case, called congestions, packets have been delayed from sending for more than 4 ms. The traffic generator manages to recover from the back-log by sending the next few packets back to back. In the second case, called overflow, the node cannot catch up and the synchronisation between the nodes is irreversibly lost once a buffer counter inside a traffic node overflows. In this case the “offered” and “accepted” data rates will differ.

6.2.2 Traffic Generator Module

The traffic generator module is the main component of the network testbed. Four traffic nodes are mounted on a mezzanine card and four of these mezzanine cards are housed on a traffic module motherboard, making a total of 16 traffic nodes per traffic module. If every traffic node was connected to a switch port via a cable, then the cost and complexity of a large system would be greatly increased. The DS-Link cable, connectors and drivers are a significant part of the cost of a link connection. Drivers and connectors also take up a significant amount of board space. It was therefore decided to incorporate an STC104 switch onto each traffic module motherboard, at the cost of reducing the flexibility of interconnecting any combination of links and switches. To reduce the number of external connections, sixteen traffic nodes are connected directly to the on-board STC104 packet switch, while the remaining 16 ports of the switch are brought out to the front panel for inter-module connections. The block diagram of the traffic generator module is shown in Figure 48.

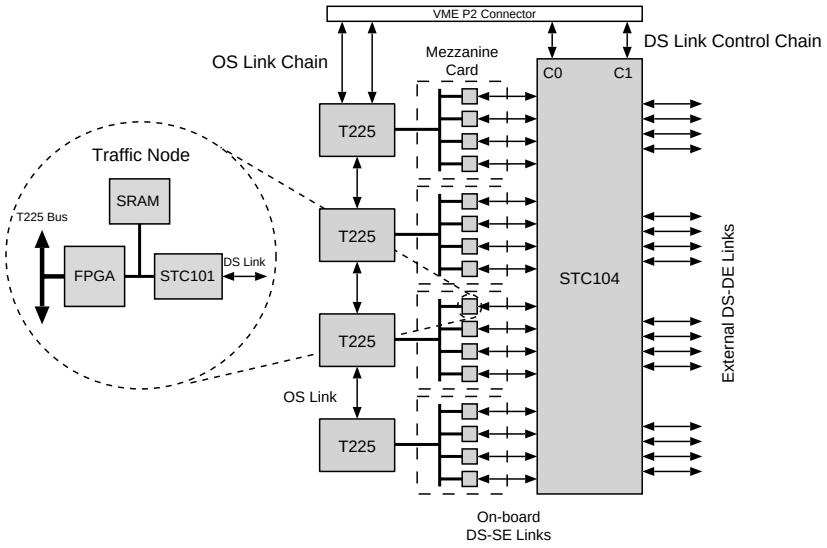


Figure 48: Traffic Generator Module Block Diagram

A control processor is used to supervise the operation of a group of 4 traffic nodes. A T225 16-bit transputer [48] is used for this purpose. The OS-Links of the T225 transputers are interconnected to form a node control network. During system initialisation, the processor is used to configure the FPGA, initialize the STC101, and to load the traffic descriptors into the pattern memory. During operation of the traffic nodes, the transputer calculates the receive and transmit data rates and monitors the DS-Link for errors. All this is done using only the on-chip memory of the transputer. The T225 has 4kbyte of on-chip RAM. An efficient routing kernel was therefore developed, which handles the communication of the OS-Link control network and enables small code modules to be loaded dynamically onto any processor in the network [49]. Figure 49 shows a photograph of the traffic generator module.

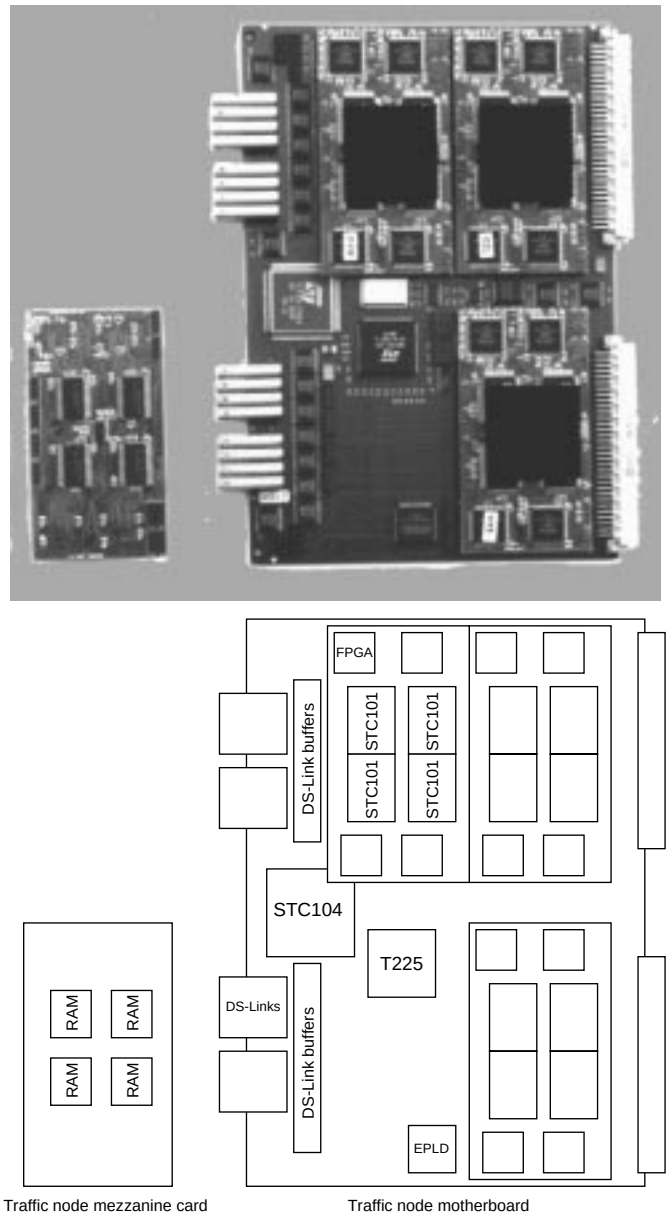


Figure 49: Traffic generator module photograph

The small card on the left of the picture shows the reverse side of one mezzanine card and the four memory chips. The space it occupies on the motherboard exposes the T225 control processor. On the installed mezzanine cards one can see the four centrally mounted STC101 link adapters, together with their shared heat sink, surrounded by the four FPGAs. In the left centre of the motherboard the STC104 switch chip can be seen, and on the left edge the sixteen link connectors together with differential drivers and receivers. A total of 65 motherboards and 260 Mezzanine cards were constructed and tested.

6.2.3 Timing Node Module

The timing node is used to measure network latency between specific terminals. This is done by injecting time stamped trace packets into the network by one timing node. These trace packets traverse the network and are received and analysed by another timing node.

In order to save space and board development cost, the physical VME module which contains the timing node performs three different functions, which are discussed separately below:

- Timing node
- DS-Link traffic monitor
- VME crate controller

The module operates either as a timing node or as a traffic monitor. The crate controller function is independent of the two.

6.2.3.1 Block Diagram of the Timing Node Module

The timing module consists of a T800 transputer with 4Mbytes of DRAM and an STC101 parallel DS-Link adapter. The token interfaces of the STC101 are used to transmit and receive packets. The device registers are initialised via the separate parallel bus interface. The DRAM memory is used to store received packets for analysis by the processor. Separate FIFOs are used to interface the STC101 receive and transmit ports to the T800 bus, and to provide additional buffering. A hardware engine to off-load the processor is implemented in two FPGAs which add time-stamps to incoming and outgoing packets and control the data flow between the FIFOs and the STC101 token interfaces. Figure 50 shows the block diagram of the timing node. The second transputer and the buffers are associated with the crate controller function, which is explained in section 6.2.6. The traffic monitor function is implemented in a programmable logic device and is explained in detail in section 6.2.5 below.

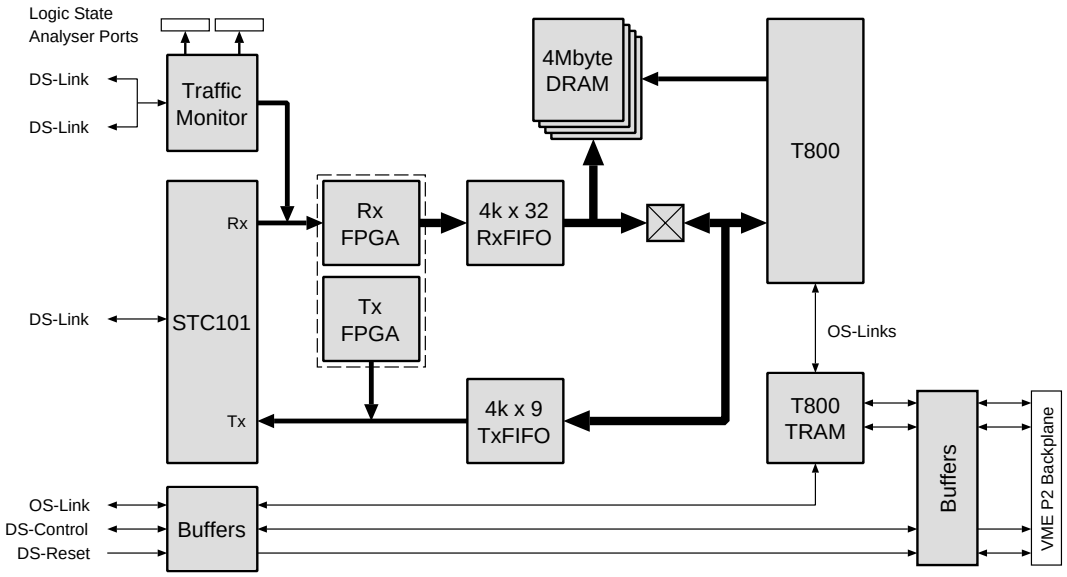


Figure 50: Block Diagram of the Timing Node

Figure 51 shows a photograph of the timing node module. A total of 10 of these modules were built and tested.

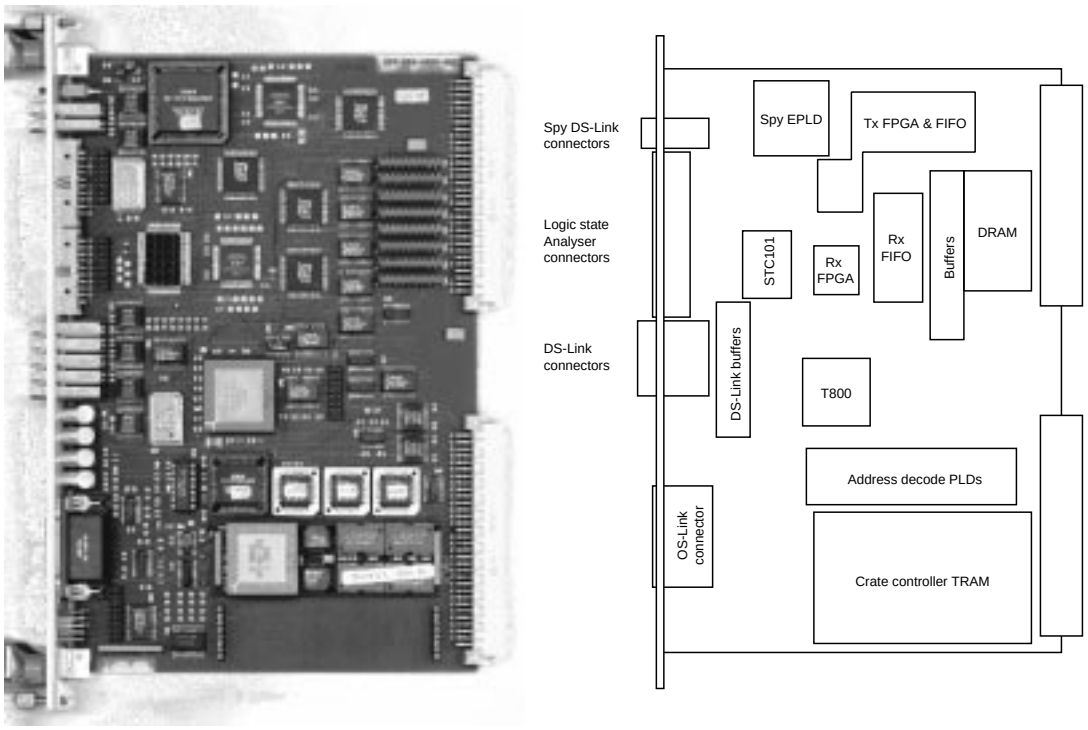


Figure 51: Timing node module photograph

6.2.4 Operation of the Timing Node

The timing node has to perform the following functions:

- send time stamped trace packets;
- receive and store incoming trace packets;
- absorb all incoming traffic at the full link rate.

In order to measure the network latency, the timing node uses a reference clock to record the send and receive time of each packet. In order to be able to accurately measure the transit time of the timing packets through the network, this clock signal must be synchronised with the clock on the other timing modules. This is achieved by using a common clock for all the timing and traffic nodes. This clock signal is daisy chained by coaxial cables from a master clock source on one of the timing modules.

Each timing node can be configured either to transmit trace packets or to receive and analyse incoming traffic. The received packets are stored in memory and analysed on-line to extract the latency statistics. In receive mode, the timing node has to be able to absorb incoming traffic at the full link speed, so that it does not cause congestion in the network, thereby biasing the latency measurements. The transmit and receive port operations are described separately in sections 6.2.4.1 and 6.2.4.2 below.

6.2.4.1 Transmit Port Operation

To send trace packets, the transputer first has to write them into the transmit FIFO. The FIFO has the same width as the transmit token interface of the STC101, i.e. 9 bits. Several packets can be stored at once, since the FIFO is 4kbyte deep. When a trace packet is to be sent, the

transputer sets a bit in the transmit FPGA. The processor thereby has full control over when the trace packets will be sent. The data from the external transmit FIFO is then moved into the transmit token interface of the STC101 by the transmit control FPGA and sent out on the outgoing DS-Link. The transmit time stamp is inserted into the byte stream by the logic in the FPGA. Outgoing packets are time-stamped when the header is written into the STC101 transmit FIFO. A status bit in the FPGA indicates when the packet has been fully sent.

The trace sent by the timing node packet use a specific format. They consist of the following fields:

- Routing header (1 or 2 bytes)
- Source identifier (16 significant bits)
- Transmit time stamp (24 significant bits)
- Packet length (16 significant bits)
- Payload (the data bytes from the transmit FIFO)
- End-of-packet character

The source identifier is used to distinguish between trace packets from different timing nodes at the receive side. The packet length is used to verify that the whole packet has been received correctly. To simplify the logic, the source identifier, the transmit time stamp and the packet length field are 32 bits wide, although less bits will actually contain relevant information, as indicated above. The minimum trace packet length is therefore 12 bytes, for a packet with no payload, excluding the routing header and the end-of-packet character.

6.2.4.2 Receive Port Operation

Incoming packets are time-stamped as soon as the packet header is read from the STC101 receive FIFO and again when the end of packet is read. This operation is performed by the receive control FPGA. It also counts the length of each incoming packet. The packets, the receive time-stamps and the packet length are written into the receive FIFO. The FIFO is 32 bits wide to match the width of the T800 data bus. The receive control FPGA demultiplexes the 8-bit wide data from the STC101 receive token interface to match the width of the receive FIFO. The timing node uses 4Mbytes of DRAM to store the incoming packets. The transputer handles the DRAM addressing and refresh. It also moves the data from the receive FIFO into the memory.

Once the DRAM buffer allocated for storing the incoming packets is full, the receive controller discards all incoming packets, until the processor has finished analysing the data. In order to maximise the number of packets which can be recorded, only the first 12 bytes from every packet are written into memory, since for the trace packets they contain the transmit time-stamp and the source number. In addition to the trace packets from another timing node, the receive port has to absorb the packets originating from the traffic nodes. They can however be easily distinguished from the trace packets by the processor, because of their specific format. Using the transmit and receive time-stamps stored in the memory, the packet latency can be calculated and histogrammed by the processor. Once the processor has finished analysing the packet buffer, the timing node acquires a new snapshot of the incoming packet stream.

6.2.5 DS-Link Traffic Monitor

To save space and to reduce board development costs, a DS-Link monitor function is included in the design of the timing module. The traffic monitor can be inserted into any cabled connection in the system. Traffic passing through the module is stored for later analysis, without altering the link traffic in any way. The only effect is a minimal extra time delay through the on-board buffers. This is useful for debugging and testing and can also provide additional information on congestion “hot spots” in the network.

The DS-Link monitor is implemented in a programmable logic device (EPLD), which performs the functions required for the receive port of a DS-Link interface:

- it extracts the clock and deserialises the DS-Link bit stream;
- it decodes all the DS-Link characters.

The monitor has two connections for a logic state analyser, one for each direction of the link. Using the logic state analyser, it is possible to observe all the characters transmitted on the link, including link control characters, such as FCC and NUL characters. The monitor circuit also detects parity and disconnect errors. This allows analysing the DS-Link traffic on a character by character basis, rather than having to look at the serial bit stream itself. This is a useful feature for debugging and greatly helped finding problems with the link start-up and flow control.

In addition, the data and end-of-packet characters flowing in one direction can be written into a FIFO buffer. The output of this buffer then replaces the receive token interface of the STC101 on the timing node, allowing latency measurements at any point in the network, as long as it is accessible using a differential DS-Link cable connection.

6.2.6 Crate Controller

The main advantage in using transputers to control the network testbed hardware comes from their built-in communication links. They greatly simplify communication between the modules. The modules installed in a VME crate are connected through a flat cable using the VME P2 backpanel connector to form an OS-Link based control network. A second transputer is incorporated onto the timing module to act as a controller for all the modules in the same VME crate. A commercially available TRAM[4] module is used for this purpose, as shown in Figure 50 above. The crate controller handles the communication with the host via the B300 Ethernet to OS-Link interface. The controller performs the following functions:

- it boots the processors on the traffic generator modules;
- it configures the FPGAs for each traffic node;
- it collects results on throughput and latency results and transfers them to the host.

The timing module provides connectors on the front panel for the OS-Link connection to the host. The crate controller receives the global system clock and reset signals, as well as the control link and reset for the DS-Link network. These signals are buffered and distributed to the modules via the VME P2 backpanel connector. Differential signalling is used for the OS

4. Transputer and RAM: a small module with a standard pinout which consists of a transputer and memory.

and DS control link daisy chain for improved noise immunity. Figure 52 shows the control link connection for one VME crate.

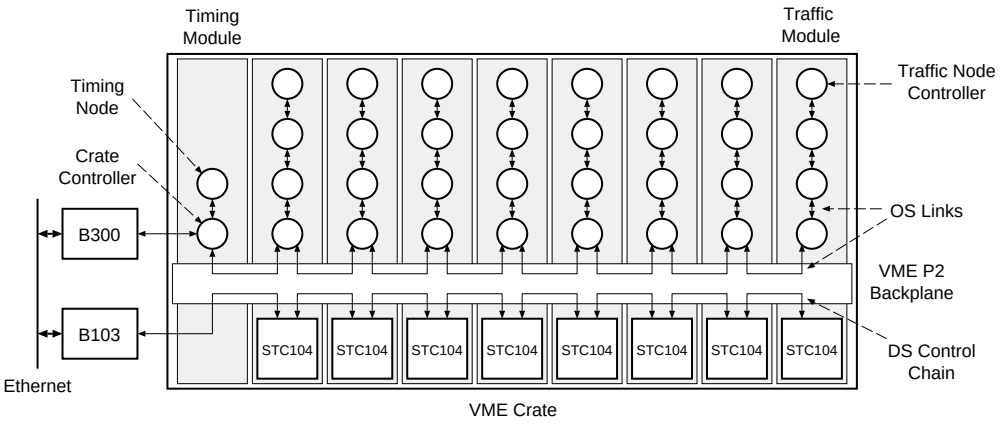


Figure 52: Control link connectivity

6.2.7 Switch Module

In order to build indirect networks, i.e. topologies where not all the switches have terminal nodes directly attached to them, a switch unit is required. It consists of one STC104 packet switch with all 32 ports brought out to the front panel through differential buffers. Two additional connectors provide a DS control link connection and a DS-Link network reset input. The control link and reset signals can alternatively also be provided via the VME P2 backplane connector. Figure 53 shows a photograph of the switch module. This board demonstrates impressively the high level of integration of the STC104 packet switch, since the module is mainly limited by the front panel space required by the DS-Link connectors. A total of 25 of these boards were built.

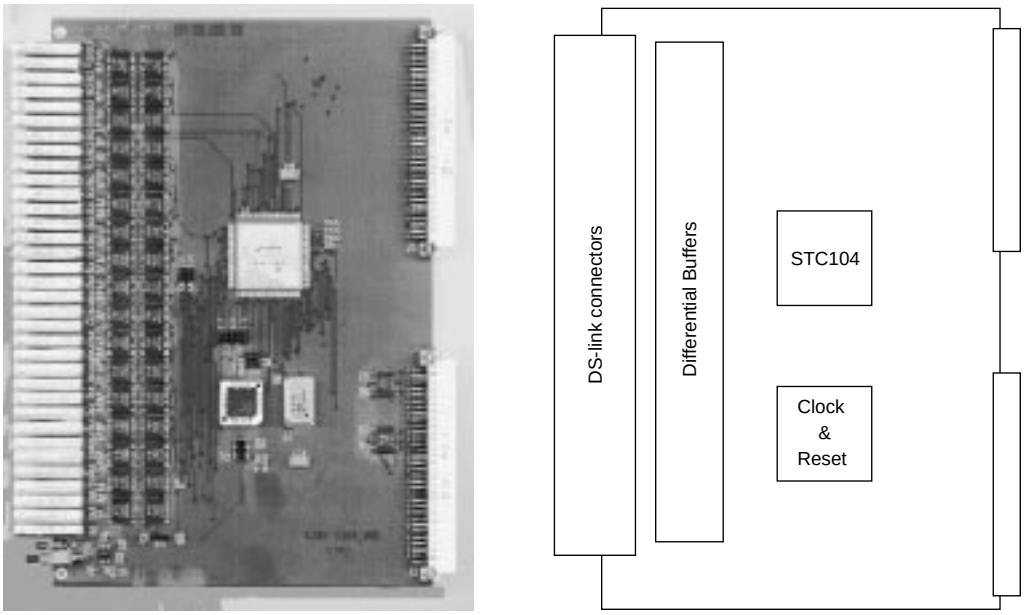


Figure 53: Switch module photograph

6.3 System Integration

The system is built in VME mechanics and is controlled and monitored by Unix workstations over Ethernet interfaces to the DS and OS control link networks. A VME crate contains up to 128 traffic nodes and the entire 1024-node system can be housed within eight such crates. The crates are controlled via Ethernet through two B300 Ethernet to OS-Link interface which drive four OS-Link[5] daisy chain connections to the control processors. The STC104 packet switches are configured via a separate DS control network through a B103 Ethernet to DS-Link interface. The setup used is illustrated in Figure 54. Up to three workstations can be used in order to speed up the loading of the traffic pattern memory images onto the traffic nodes. This was the most time-consuming part of the system initialisation, because of the low throughput of the B300 Ethernet to OS-Link interfaces.

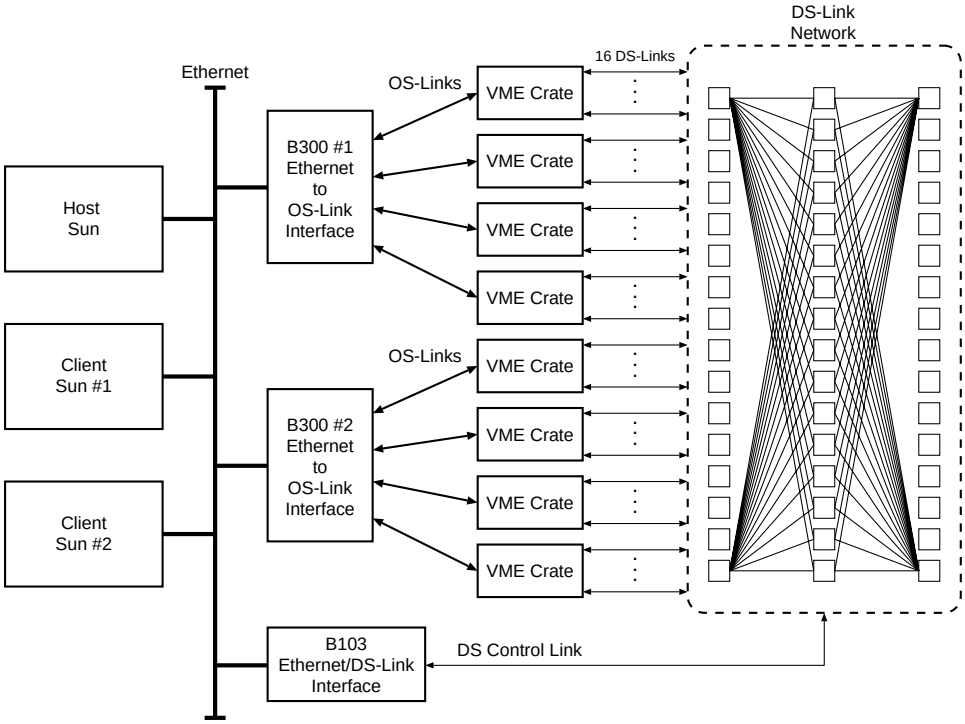


Figure 54: Macramé system configuration

Each VME crate contains one timing module, which operates as the crate controller, up to eight traffic modules, and up to two switch units. Also shown in the figure is an eight crate system with its control networks. There are two separate control networks, an OS link network for T225 and T800 processors and a DS-Link network for the STC104 packet switches. The OS and DS networks are interfaced to the host SUN work workstation via B300 and B103 adapters. Figure 55 illustrates the arrangement of modules in a fully populated VME crate.

5. 20 Mbit/s Over Sampled Transputer links

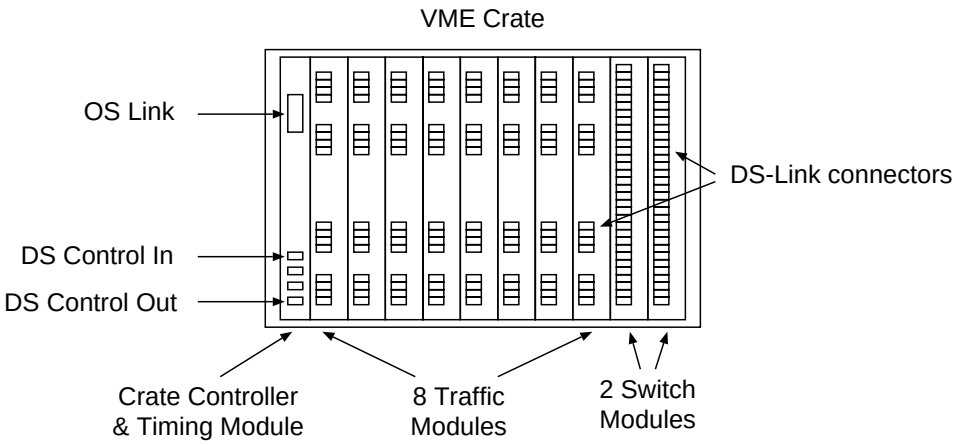


Figure 55: VME crate setup

6.4 Software

Extensive work was undertaken to produce software to control and monitor the system and to ensure that the required measurements could be reliably obtained. This section just gives a short overview of the software required to operate the DS-Link network testbed. A more detailed description can be found in [49].

6.4.1 System Control Software

A set of files are prepared off line containing the packet descriptors, the configuration information for every traffic node, and the routing tables for the packet switches. Prior to loading this data, the control networks for the traffic nodes and packet switches are used to verify that the expected devices are present and connected in the required order. A tool which allows configuring and verifying networks of DS-Link devices through the DS control network was developed [50].

Each control processor has only 4kbytes of on-chip memory. It is loaded at initialisation time with a kernel which handles the control link traffic and the dynamic loading of the application programs. Application programs for self-test, hardware configuration, storing of traffic descriptors, and run time supervision are loaded in turn by the host which also controls their synchronisation.

Once the system is running, each control processor maintains local histograms of results. These are returned to the host on request for on-line display, data logging, and subsequent analysis. It is possible to set up configuration files that can be used to perform multiple runs on the testbed, each time varying the traffic parameters and, if required, the number of nodes transmitting. This allows for runs to be setup overnight or without user intervention. The list below shows the sequence of steps that are executed for each measurement run:

1. Configure the crate master on the timing module;
2. Boot traffic node controllers and the timing modules;
3. Configure the FPGA's of traffic nodes and timing node;
4. Configure the DS-Link network though the DS control network;

5. Load the traffic patterns onto the traffic nodes;
6. Configure the STC101's;
7. Start the traffic nodes and timing nodes;
8. Read and report the results from the traffic nodes;
9. Read and report the latency results from the timing nodes.

The loading of the traffic patterns and the configuration of the DS-Link network are executed in parallel.

6.4.2 Traffic Pattern Generation

A program was written by the author to generate the traffic descriptors for the traffic nodes. The traffic can be described by assigning random distributions to the three variables destination, packet length, and delay time. Fixed values, uniform and negative exponential distributions are all available. Distributions can also be defined as a histogram, where each possible value has a probability assigned to it. The traffic is defined in a simple ASCII file format. Uniform random traffic on a 512-node network can be specified in a single line, as shown below:

```
# Source  Destination  Length  Delay
[0:511]   u(0:511)     c(64)   e(10.0)
```

Each of the 512 traffic nodes numbered from 0 to 511 sends packets to destinations chosen from a uniform random distribution. The packet length is fixed at 64 bytes and the time between sending packets follows a negative exponential distribution with an average of 10 µs. The average transmit rate is therefore 6.4 Mbyte/s. Another example for systematic traffic with 512 nodes is given below:

```
[0:255]     o(256:511)   c(1024)   c(107.8)
[256:511]   o(0:255)     c(1024)   c(107.8)
```

This type of traffic uses a one-to-one mapping of sources to destinations, i.e. node 0 sends to node 256, node 1 sends to node 257 and so on. The packet length is this case is 1024 bytes and the delay between packets is constant with 107.8 µs. This corresponds to an applied throughput of 9.5 Mbyte/s. The program generates an individual file with the traffic descriptors for each of the traffic nodes. The file format is binary to speed up the loading into the traffic node memories.

6.5 Implementation of Network Topologies

The list below gives the total number of modules described earlier in this chapter that were constructed and tested:

- 65 traffic generator modules
- 25 switch modules
- 10 timing modules

These modules can be used to implement a wide range of different networks. Two specific examples will be given below.

6.5.1 2-Dimensional Grid Network

Figure 56 shows how a 400 node two-dimensional grid network can be constructed using 25 traffic modules. The network consists of an array of 5 by 5 traffic modules. Every packet switch has 16 on-board connections to the traffic nodes and four external connections to each of the four adjacent switches. The largest network of this type that has been built is a 1024-node 2-D grid, which requires all of the 64 traffic modules that were constructed.

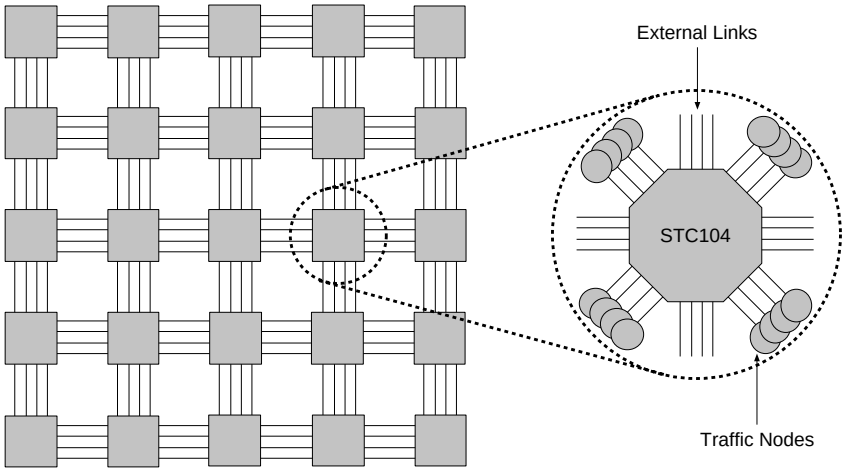


Figure 56: 400-node 2-dimensional Grid Network

6.5.2 Clos Network

Figure 57 shows how a 256-node 3-stage folded Clos network can be constructed out of 16 traffic generator modules and 8 switch modules. The switch modules are required for the centre stage of the Clos network.

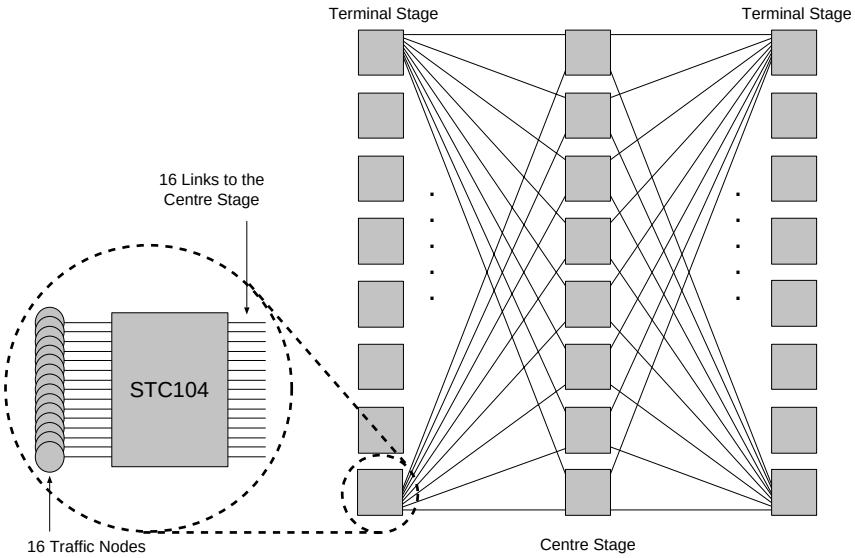


Figure 57: 256-node 3-Stage Clos Network

Each terminal stage switch is connected to every centre stage switch with a bundle of two links. The largest network of this type that can be built out of the available modules is a 512-node Clos. A fully connected 1024-node Clos network uses 5 stages and would require a

total of 64 traffic modules, 96 switch modules and 2048 cables. This is clearly not practical. To reduce the number of modules and cables required, a 64-way switch module would have had to be constructed out of 6 STC104 switches. However, due to front panel space restrictions, this would have meant moving to a different form-factor and would still have required 1024 cables. The restriction to 512 nodes was not thought to be a critical issue.

6.5.3 Testbed Installation

Figure 58 below shows the full size Macramé testbed set up as a 2-dimensional grid of 8 by 8 switches with 1024 nodes. The entire system is housed in the two racks with four VME crates

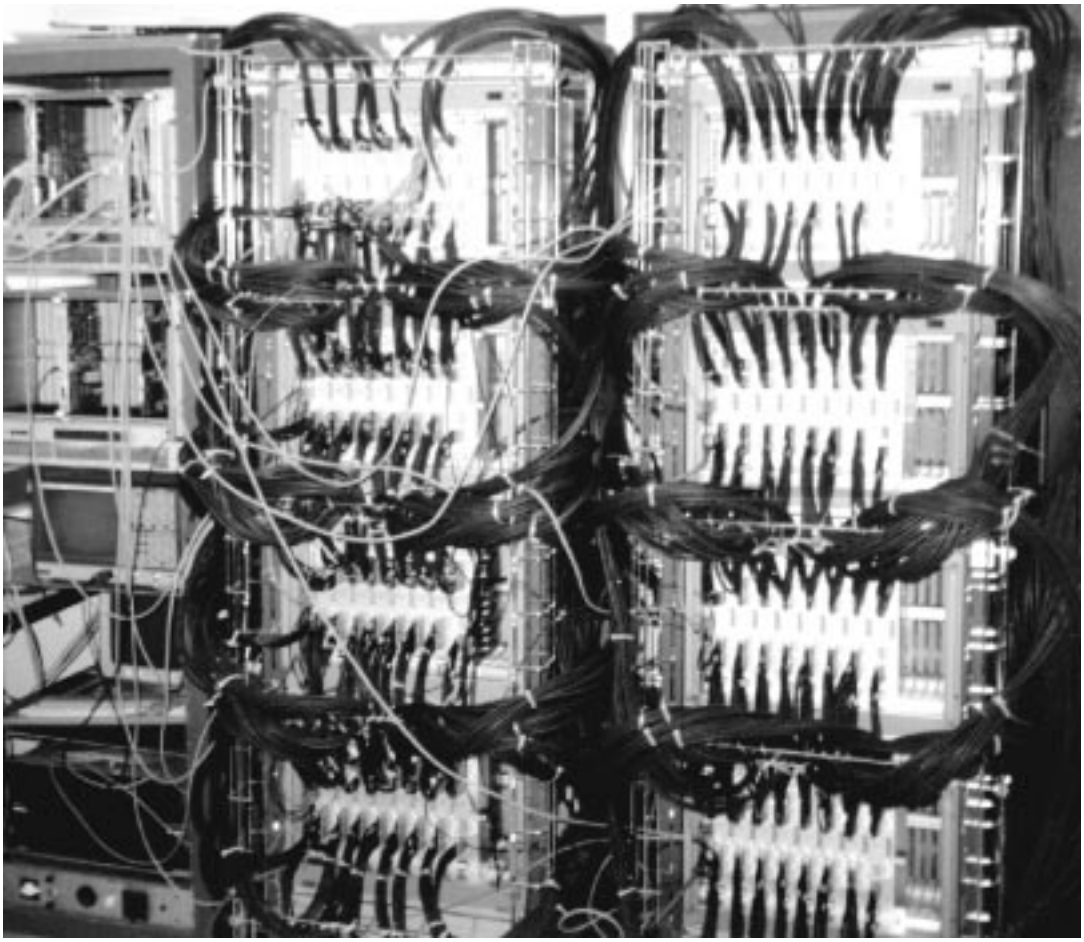


Figure 58: Picture of the Macrame Testbed

each. Every VME crate contains 8 traffic node modules, i.e. 128 traffic generator nodes and 8 switches. A timing node module is installed in the first slot of every crate and acts as a crate controller. The connections between the boards are made using the standard DS-Link cables. Bundles of four links connect every switch in the system to its nearest neighbours both vertically and horizontally. The rack on the left contains the root for the DS-Link control network. The B103 DS-Link to Ethernet interface module is installed in the topmost VME crate together with another STC104 switch module, which acts as a fan-out to distribute the DS-Link control chain to the other VME crates. At the bottom of the rack are the two B300

modules, which interface the OS-Link control network to Ethernet. The whole system is controlled by three SUN workstations over Ethernet (not visible in the picture).

6.6 Performance Measurements

In this section, results from performance measurements on the different components of the testbed are presented. These measurements serve as a baseline to assess the results from larger network configurations.

6.6.1 Traffic Generator Single Link Bandwidth

The link throughput for unidirectional and bidirectional traffic has been measured for a range of different size packets using the traffic generator modules. The results are shown in Figure 59. The theoretical maximum link bandwidth, which is also shown, has been calculated using Equation 1 and Equation 2 from Chapter 2.

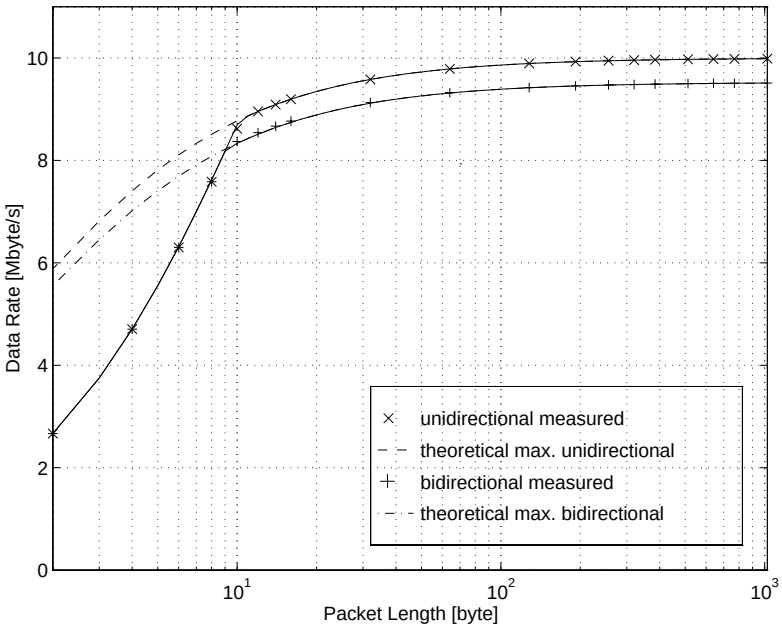


Figure 59: Unidirectional and bi-directional single link bandwidth

The results show that for packets longer than 10 bytes, the traffic node can sustain the full DS-Link bandwidth. For shorter packets, the throughput is reduced, due to the overhead for fetching and processing the traffic descriptors by the FPGA on the traffic nodes. The overhead limits the rate at which short packets can be sent. This was not considered to be a limitation, since a continuous stream of such short packets would rarely be used in an application.

The time between sending packets for the traffic generator can be modelled as follows, where l is the packet length in bytes, $t_{Overhead}$ is the overhead in the traffic generator, t_{Fifo} is the time to write one characters into the STC101 transmit FIFO, t_{DATA} and t_{EOX} are the time to transmit a data and an end-of-packet character on the link, respectively:

$$t_{Packet} = max\begin{cases} t_{Overhead} + t_{Fifo} \cdot l \\ t_{Data} \cdot (l+1) + t_{EOX} \end{cases} \quad \textbf{(15)}$$

The overhead is 650 ns for one byte headers and the link running at 100MBaud. Each additional header byte increases the overhead by another 100 ns. Since the parallel interface of the STC101 is clocked at 20MHz, each data byte takes 50 ns to be written into the transmit FIFO. For longer packets, the time to send the packet is limited by the actual transit time of the characters on the link, i.e. 100 ns for a data character, 40 ns for the end-of-packet, assuming 100MBaud operation. The crossover occurs at a packet length of 10 bytes.

Using the data rate measurements shown in Figure 59, the packet transmission time, i.e. the minimum time between sending packets from the traffic generator, can be calculated using the following equation:

$$t_{Packet} = \frac{l}{R} \quad \textbf{(16)}$$

where l is the packet length and R is the measured data rate.

Figure 60 below shows a plot of the packet transmission time for unidirectional traffic, and the time calculated using the Equation 15, as a function of the packet length. The agreement between the measurement and the model is excellent.

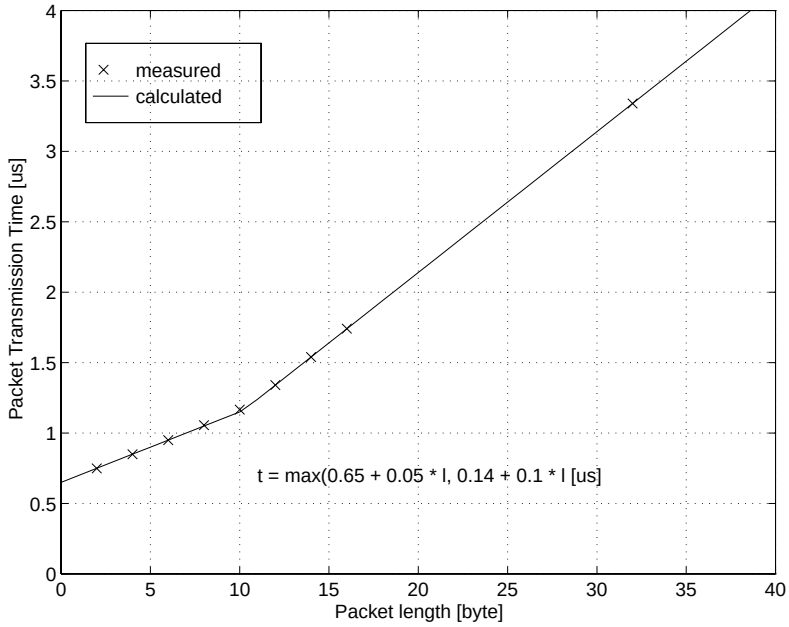


Figure 60: Packet transmission time versus packet length

6.6.2 Timing Node Latency

In order to characterise the packet transmission and reception overhead in the timing node and to exactly measure the switching latency of the STC104, a simple system with two timing nodes was set up. The timing nodes were connected via one, two, or three STC104 switches, respectively and the latency for sending trace packets from one node to the other was meas-

ured for different packet length values. The measured latency includes the packet transmission time, the switching delay of the STC104s, and the overheads in the timing node. These are due to delays in the transmission and reception of packets through the link adapter as well as to the logic that adds the timestamps to the outgoing and incoming packets. Figure 61 shows the measurement results for short packets. The packet length values start at 12 bytes, which corresponds to the minimum trace packet length that the timing nodes can transmit, as explained in section 6.2.4.1 on page 75.

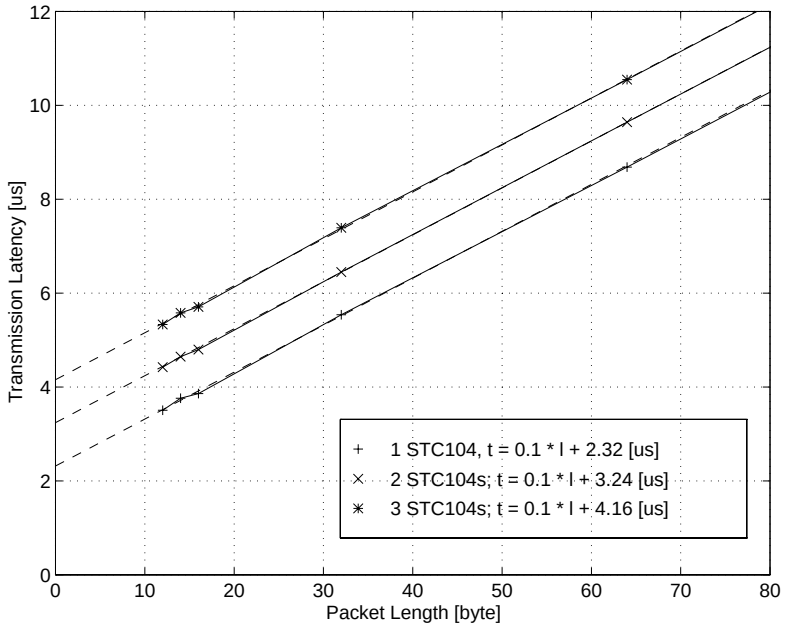


Figure 61: Timing node latency

A linear fit on the data was performed to calculate the slope and the offset. The results are shown in table Table 7.

Table 7: Latency for different numbers of switches

Number of switches	Offset [µs]	Slope [µs/byte]
1	2.32	0.1
2	3.24	0.1
3	4.16	0.1

The following parameters can be extracted from these results:

- The delay through one STC104 switch without contention is 0.92 µs for the given core clock speed of 30MHz and DS-Links running at 100MBaud. This value includes the propagation delay for a 2 metre DS-Link cable and for the on-board differential transceivers, which is about 20ns in total. The measured switch latency is very close to the value that can be calculated based on the design of the STC104. Using Equation 10 from Chapter 5 the estimated switch latency is 850ns.
- The combined overheads in the transmitting and the receiving timing node can be calculated by subtracting the value for the switching delay and the packet overhead from the offset values in Table 7. The packet overhead, i.e. the time to send the single header byte and an end-of-packet character, is 0.14 µs. The overheads in the timing node therefore

account for an additional delay of 1.26 µs. This delay is due to the latency through the STC101s and delays in the time-stamp logic in the FPGAs on the timing nodes. This figure has to be subtracted from the latency measurements, if the actual network latency is of interest.

- The latency increases by 0.1 µs per byte as expected for a 100MBaud unidirectional DS-Link.

6.6.3 Timing Node Bandwidth

The receive bandwidth of the timing node was measured by connecting a traffic generator to a timing node in receive mode. Figure 62 shows the receive data rate for the timing node as a function of the packet length. The data rate for the case of a traffic generator sending to another traffic generator node is shown for comparison. The maximum theoretical unidirectional DS-Link data rate calculated using Equation 1 from Chapter 2 is shown as a dashed line.

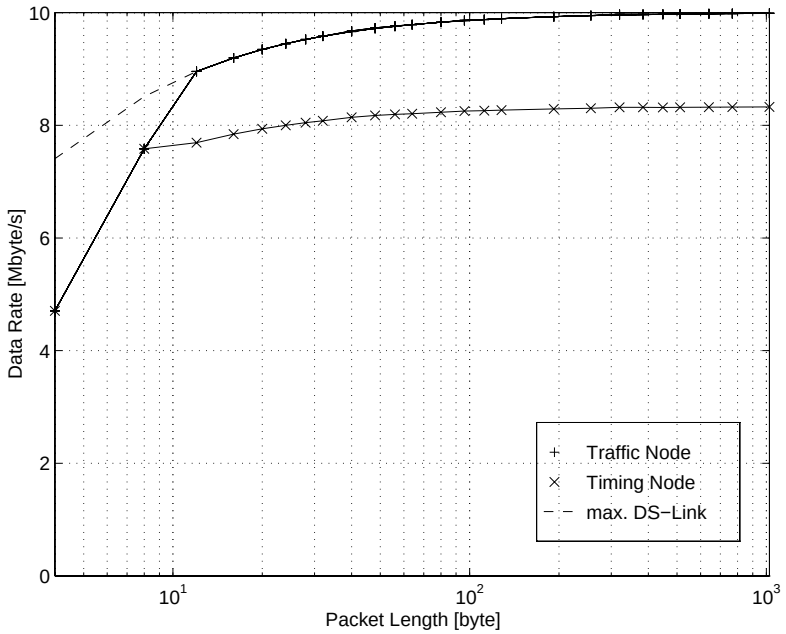


Figure 62: Timing and Traffic Node Receive Data Rates

The results show that the receive data rate of the traffic node reaches the full DS-Link bandwidth for packets longer than 10 bytes, while the throughput of the timing node is limited to about 8.3Mbyte/s. This behaviour was found to be due to a bug in the STC101 chip which occurs if packets are received back-to-back, i.e. without intervening NULL characters. By simulating the RTL[6] VHDL model provided by the semiconductor manufacturer, the problem was identified to be in the interface between the DS-Link module and the receive FIFO inside the STC101, i.e. at the boundary between the link clock domain (50 MHz) and the parallel interface clock domain (20 MHz for the designs here). The rate at which characters are written into the receive FIFO drops, a character is written at most every 120ns, which corresponds to an asymptotic bandwidth of 8.33 Mbyte/s for very long packets.

6. Register Transfer Level

A work-around was found for the traffic node, since the STC101 can be configured not to write the end-of-packet characters into the receive FIFO. The receive rate then reaches full DS-Link bandwidth as shown above. Disabling the end-of-packet writes is not a problem in the case of the traffic node, since received data bytes are only counted to determine the average data rate, the packet structure is not essential. This solution is unfortunately not viable for the timing node, since it needs the end-of-packet characters to delineate packets and to determine the packet length.

6.7 Summary and Conclusions

The objective of the work presented in this chapter was the design and construction of a large network testbed of variable topology based on the IEEE 1355 DS-Link technology. The motivation for this task was to provide proof of the technology used and to demonstrate that it was viable for constructing large scale systems. The system should allow network performance in terms of latency and throughput to be measured under controlled conditions for programmable traffic patterns and for various topologies and network sizes.

These aims have been met by designing a systems based on a large number of traffic generator nodes and a small number of timing nodes to perform latency measurements. Three different VME modules were designed and tested:

- A traffic generator module with 16 traffic generator nodes and an STC104 switch, which can generate programmable traffic patterns.
- A timing node module, which allows packet latency measurement between a subset of the network terminals and also contains supporting logic for controlling the system.
- A switch module was constructed, which consist of an STC104 with all links brought out to the front panel.

The performance of the different modules has been quantified and conforms to the expectations. The full size 1024-node system has been implemented using these components and has been shown to work reliably. The testbed provides a unique platform to study the performance of large wormhole routed switching networks with link-level flow control. Results obtained using this system will be presented in Chapter 7.

Chapter 7
Results from the Macramé Network Testbed

The full scale DS-Link network testbed with 1024 nodes as described in chapter 6 was built and tested. The performance of 2-dimensional grid, torus and multistage Clos networks has been studied for different network sizes and under various traffic patterns. Results from these measurements are presented in this chapter.

7.1 Single Switch Performance

In order to establish a baseline to assess the performance of larger networks the throughput and latency of a single STC104 switch has been measured. Two traffic node modules with 16 nodes were connected to the 32 external links of a single switch module.

7.1.1 Switch Throughput

The total bandwidth of the STC104 switch has been measured for random and systematic traffic. Figure 63 shows the saturation throughput for a single STC104 with varying packet length.

Under systematic traffic, the measured throughput approaches 305 Mbyte/s for long packets. This value is equal to the aggregate asymptotic bandwidth of 32 links for bidirectional transmission, i.e. 32 times 9.52 MByte/s, which demonstrates that the STC104 can indeed sustain the full bandwidth of 32 links when there is no destination contention. For short packets, the throughput is reduced because of the overhead for the header and end-of-packet characters. In addition, the traffic nodes are unable to saturate the link for packet length values smaller than 10 bytes as shown in section 6.6.1 of chapter 6.

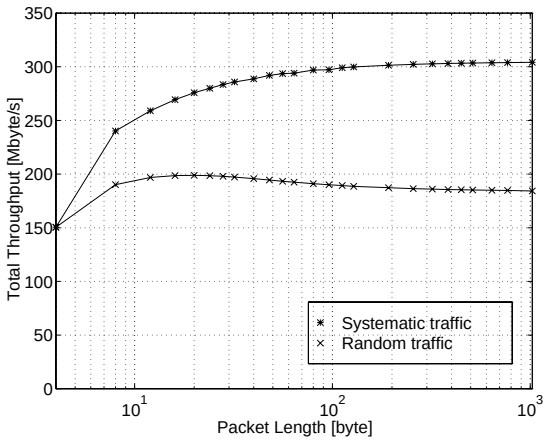


Figure 63: Total throughput for a single switch under random traffic

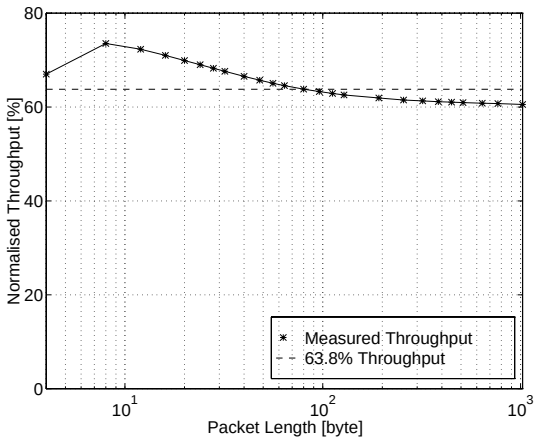


Figure 64: Normalized throughput versus packet length for random traffic on a single switch

Under random traffic, the switch saturation throughput for long packets is only 185 Mbyte/s, which corresponds to a per-node throughput of about 5.8 MByte/s. The reduction compared to systematic traffic is due to contention and head-of-line blocking. The throughput for packet length values in the range of 10 to 60 bytes is higher. This is because of the output buffers present in the STC104 switch. Each link has 70 characters of buffering, of which 45 characters are on the input side of the internal crossbar and 25 are on the output side. Since the internal crossbar operates at three times the link speed, packets can be transferred from the input buffers to the output buffers faster, which reduces the blocking time and therefore increases the output link utilisation. For packets shorter than 10 bytes, the throughput is again reduced due to fixed packet overheads for the header and end-of-packet characters.

Figure 64 shows the throughput under random traffic normalized to the theoretical maximum throughput, i.e. 32 times the maximum link throughput for the given packet length. Also shown is the theoretical utilisation for a crossbar switch under random traffic calculated from Equation 13 from Chapter 5, which is 63.8% for a 32-way crossbar. For long packets, the measured switch throughput of about 61% matches the predicted value quite well. For packets shorter than 64 bytes, the switch utilisation is significantly higher, with a maximum value of more than 70% for 8 byte packets. The mathematical model does not show this effect, since it does not take the buffers in the switch into account.

7.1.2 Packet Latency

The latency of a single switch with 30 traffic nodes was measured under random traffic with 64 byte packets for a range of network load values. The remaining two links of the switch were used to connect the transmitting and receiving timing node. Figure 65 shows the accepted network throughput as a function of the applied load. Below saturation, the accepted throughput follows the offered throughput exactly. When the switch is saturated, an increase in applied load does not lead to a corresponding increase of the measured throughput. The measured saturation throughput is about 180 MByte/s, which is equivalent to 64.5% of the switch throughput for permutation traffic, i.e. without contention.

Figure 66 shows the average packet latency as a function of the measured aggregate throughput for the same setup as above. Also shown is the minimum latency, which is 7.4 µs.

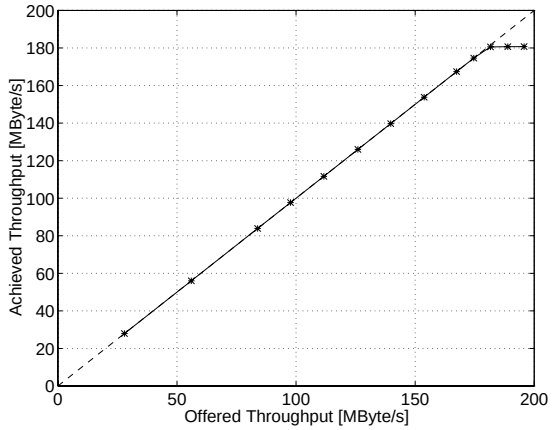


Figure 65: Accepted versus offered throughput for a single switch under random traffic

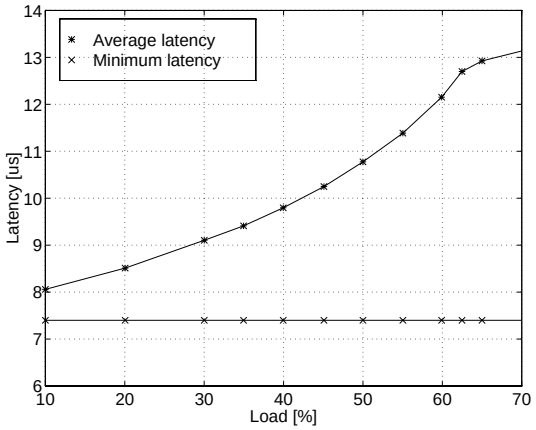


Figure 66: Latency for an STC104 under random traffic

The minimum latency is constant, because there are always some timing node packets which cross the switch without being blocked. It is equal to the sum of the packet transmission time and the switching delay. The measured switching delay is 0.9 μs (from chapter 6, section Section 6.6.2 on page 85) and the packet transmission time for a 64 byte packet is about 6.5 μs (from Equation 1 in Chapter 2).

The average latency increases with the network load. For low loads, only few packets are blocked and the average latency is only slightly larger than the minimum latency. As the network load increases, more and more packets are blocked and therefore the queuing delay increases too. When the switch becomes saturated, all the queues become full and the latency does not increase any more. Under saturation, the average latency is 13.1 μs, i.e. on the average every packet is queued for 5.7 μs, which is equivalent to 87% of the packet transmission time. The results show that when moving from low load to saturation, the average packet latency increases by less than one packet transmission time. However, the packet latency is a random variable, and therefore the latency distribution also needs to be considered (see Section 7.4.4 on page 104).

7.2 Comparison of Network Topologies

After having established that the performance of a single STC104 matches the predictions well, a number of different size grid, torus and multistage Clos networks were assembled, in order to compare the performance of different network topologies and their scaling behaviour, i.e. how the performance changes when network size is increased. This is also an important consideration for the application in high energy physics, as will be seen in section 7.8.

A comparison of the two network topologies studied is presented in this section, followed by the results of a more detailed study of 2-D grid/torus networks and Clos networks in sections 7.3 and 7.4, respectively.

7.2.1 Overview of the Network Topologies

The grid networks studied consist of a 2-dimensional square array of switches. Each switch has 16 end-nodes attached to it and groups of four links connect adjacent switches. Grouped adaptive routing is used on the link bundles between adjacent routers, i.e. packets are sent out on any link in the group which is found to be idle. This means that the bundle of links effectively behaves like a single high-bandwidth connection. A torus is a grid structure where the links at the edges of the network wrap around and connect to the opposite edge. Therefore the switches are connected in rings rather than chains, and in terms of connectivity every switch in the network is equivalent. Table 8 summarises the characteristics of the 2-dimensional grid and torus networks which were studied.

The maximum cross-sectional bandwidth is defined as the bidirectional data rate that can pass between two parts of the network if it is divided into two equal halves. On the grid, the bi-section bandwidth scales with the square root of the number of nodes. The total network band-

width is the aggregate throughput of all end-nodes, assuming they all transmit at the full link rate. This value sets an upper bound to the throughput achievable on a given network.

Table 8: Characteristics of 2-dimensional grid and torus networks

Number of nodes	Number of switches	Topology	Total bandwidth [Mbyte/s]	Bi-section bandwidth [Mbyte/s]
64	4	2 by 2 grid	610	152
144	9	3 by 3 grid	1370	228
256	16	4 by 4 grid	2440	305
400	25	5 by 5 grid	3810	381
576	36	6 by 6 grid	5490	457
784	49	7 by 7 grid	7470	533
1024	64	8 by 8 grid	9750	610
1024	64	8 by 8 torus	9750	1220

The Clos is a multistage network; the networks studied have one terminal stage and one centre stage of switches. Each terminal stage switch has 16 nodes attached to it and connects to every centre stage switch with the remaining 16 links. The 16 links from each terminal stage switch to the centre stage are grouped. If more than one link from a centre stage link connects to the same terminal stage switch, these links bundles are also grouped. Table 9 below summarises the characteristics of the Clos networks; the topology column gives the number of switches in the terminal and centre stages separated by a colon. The bi-section bandwidth of the Clos scales linearly with the number of nodes for the networks studied.

Table 9: Characteristics of Clos networks

Number of nodes	Number of switches	Topology	Total bandwidth [Mbyte/s]	Bi-section bandwidth [Mbyte/s]
64	6	4:2	610	610
128	12	8:4	1220	1220
256	24	16:8	2440	2440
512	48	32:16	4880	4880

7.2.2 Scalability of Clos and 2-D Grid Networks

Figure 67 shows the saturation network throughput for different sizes of Clos and 2-dimensional grid networks under random and systematic traffic for 64 byte packets. Systematic traffic involves fixed pairs of nodes communicating, i.e. there is no destination contention. In addition the source-destination pairs have been chosen such that contention for internal network links is minimised. Therefore the results for this traffic pattern set an upper bound for the performance achievable with the given network. For the 2-dimensional grid and torus, the traffic pattern is based on communication between nodes attached to nearest neighbour switches. For the 2-dimensional grid networks, some of the nodes on the edge of the network are not active. For the Clos, all the sources on the first terminal stage switch send packets to the last terminal stage switch, the nodes on the second switch send to the second to last switch and so on. This pattern forces all the packets to cross the centre stage switches.

The throughput of the Clos networks for random traffic is higher than for the 2-dimensional grids. This is because of the larger cross-sectional bandwidth of the Clos networks which

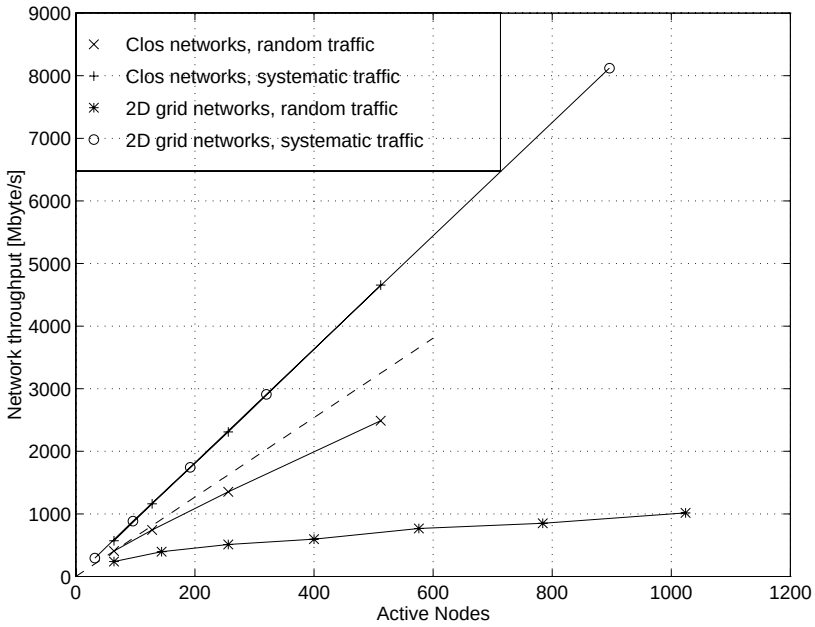


Figure 67: Throughput for different size Clos and 2-D grid networks under random and systematic traffic

scales linearly with the number of nodes, whereas for the grid the bi-section bandwidth only increases with the square root of the number of nodes.

The results show that the network throughput under random traffic is always significantly lower than the maximum theoretical cross-sectional bandwidth. For random traffic, contention at the destinations and internally to the network reduces the network throughput compared to that obtained for systematic traffic, where there is no destination contention. The fall off in performance from systematic to random traffic is more pronounced for the grid than the Clos.

The degradation of performance as the network size increases agrees with analytical models presented in [43]. This study predicts the throughput of Clos networks under sustained random load to degrade by approximately 25% from linear when the network size is increased from 64 to 512 nodes. The measurement results shown in Figure 67 show a reduction of about 20% under the same conditions, linear scaling is shown as a dashed line.

7.2.3 Node Throughput of 2-D Grid, Torus and Clos Networks

Figure 68 shows the per-node saturation throughput for different size 2-dimensional grid, 2-dimensional torus and Clos networks as a function of the packet length. The traffic pattern is random, i.e. transmitting nodes choose a destination from a uniform distribution.

As expected, the Clos networks give the highest throughput, e.g. the 256-node Clos achieves 61% (5.6 MByte/s per node) of the maximum theoretical throughput for 64 byte packets, whereas the 256-node 2-D grid only achieves 22% (2 MByte/s per node). This is because the cross-sectional bandwidth of the Clos is much higher than for the 2-dimensional grid and torus, e.g. the 256-node Clos has a maximum theoretical cross-sectional bandwidth of 2.4 GByte/s (see Table 9), whereas for the grid with the same number of nodes it is only 305 MByte/s, as shown in Table 8.

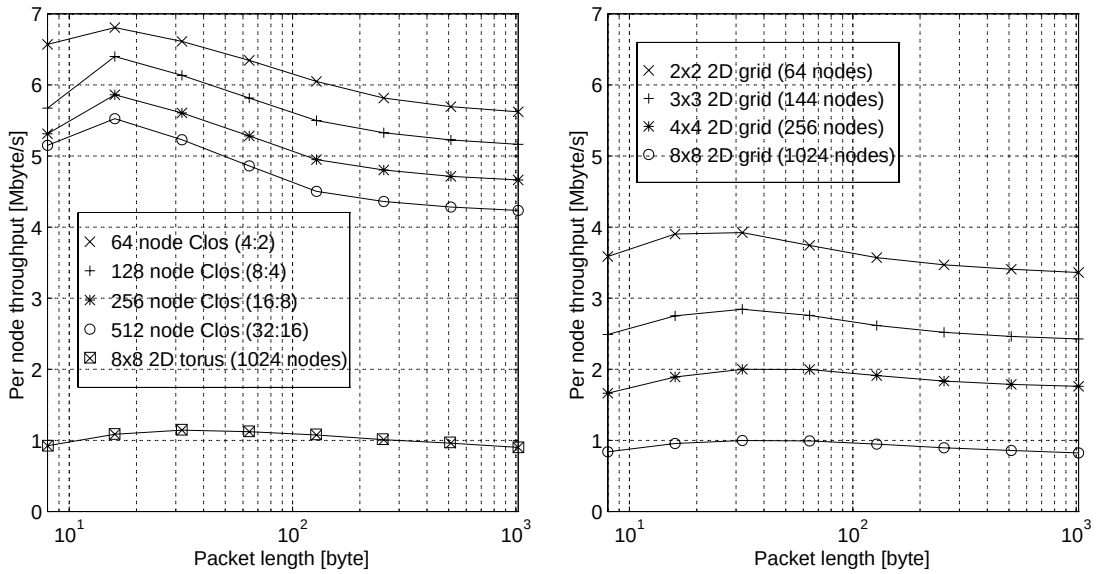


Figure 68: Node throughput for 2-dimensional grid and Clos networks under random traffic

For the network topologies shown, the per-node throughput decreases as the network size increases. This shows that the throughput of Clos and 2-dimensional grid networks does not scale linearly with network size under random traffic. This effect is more pronounced for the 2-dimensional grids, e.g. for a 64 node grid, which consists of an array of 2 by 2 switches; the per-node throughput under random traffic is 40% (4 MByte/s) of the maximum link bandwidth, while for a 1024 node grid (8 by 8 switches), the per node throughput is only 10% (1 MByte/s) of the maximum link bandwidth. The throughput of the torus is about 20% higher than the grid due to the extra wrap around links which are available.

The effect of packet length on throughput can also be observed from Figure 68: For small packets of less than 16 bytes the throughput is reduced due to protocol overheads. Medium sized packets, around 64 bytes, give the best performance because of the output buffering present in the STC104, as already seen in section 7.1.1 above. Long packets of more than 200 bytes fill the entire path through the network from source to destination, and therefore throughput is reduced by head-of-line blocking.

7.2.4 Summary of Throughput Results

Table 10 shows the measured network throughput for the different size 2-D grid and torus networks under study. The values shown are for a packet length of 64 bytes. The table also shows the number of active nodes and the theoretical maximum network bandwidth for random and systematic traffic, respectively. The number of active nodes for random and systematic traffic are different, since for the systematic traffic patterns used, some of the edge nodes are inactive on the 2-D grid networks.

The results demonstrate, that the throughput under random traffic decreases dramatically as the network size increases, going from 39% for a 2 by 2 grid to only 10% for an 8 by 8 grid.

Table 10: Performance of 2-dimensional grid and torus networks under random and systematic traffic

Network Size	Number of active Nodes		Theoretical Network Bandwidth [MByte/s]		Network Saturation Throughput [MByte/s]	
Grid	Random	Systematic	Random	Systematic	Random	Systematic
2 x 2	64	32	609	305	240 (39%)	294 (97%)
3 x 3	144	96	1370	914	397 (30%)	886 (97%)
4 x 4	256	192	2440	1830	511 (21%)	1746 (96%)
5 x 5	400	320	3810	3050	597 (16%)	2910 (96%)
6 x 6	576	480	5490	---	768 (14%)	---
7 x 7	784	672	7470	---	850 (11%)	---
8 x 8	1024	896	9750	8530	1016 (10%)	8120 (95%)
Torus	**Random**	**Systematic**	**Random**	**Systematic**	**Random**	**Systematic**
8 x 8	1024	1024	9750	9750	1210 (12%)	9280 (95%)

Table 11 lists the measured network saturation throughput under random and systematic traffic for the four Clos networks under study. Also shown is the theoretical network bandwidth and the achieved percentage of the theoretical maximum.

Table 11: Performance of Clos Network under random and systematic traffic

Network Size [nodes]	Theoretical Bandwidth [MByte/s]	Saturation Throughput Random Traffic [MByte/s]	Saturation Throughput Systematic Traffic [MByte/s]
64	610	406 (67%)	570 (94%)
128	1220	744 (61%)	1160 (95%)
256	2440	1350 (56%)	2310 (95%)
512	4880	2490 (51%)	4650 (95%)

7.3 Performance of 2-D Grid and Torus Networks

In this section, the performance of 2-D grid and torus networks is analysed in more detail and results are presented for different types of traffic as a function of the network load and the number of active nodes.

7.3.1 Comparison of Grid and Torus Topologies

To compare the performance of the grid and torus topologies, the per-node throughput under random traffic has been measured as a function of applied load. The packet length is 64 bytes and all measurements are made using the full 1024 node system with 8 by 8 switches. The results are shown in Figure 69 and Figure 70, respectively. Each graph contains four plots which correspond to 4, 8, 12 and 16 active traffic nodes per traffic module, i.e. the whole network is used for all measurements, but the number of nodes connected to the network is varied. The achieved per-node throughput increases linearly with the applied load until the network saturates. If the load is increased even further, the achieved throughput decreases

slightly. This is because the links in the centre of the network are completely saturated (see also section 7.3.3).

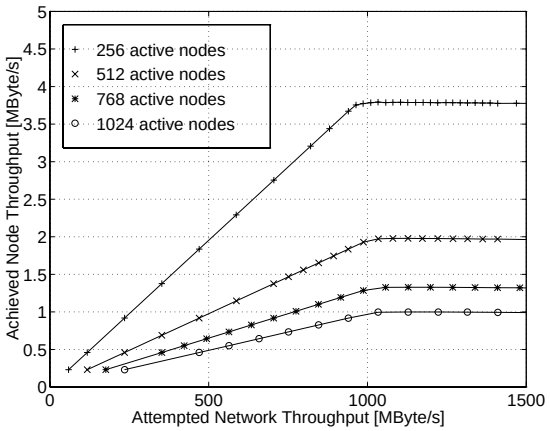


Figure 69: Node throughput versus attempted network throughput for a 1024 node grid with 4 to 16 active nodes per switch

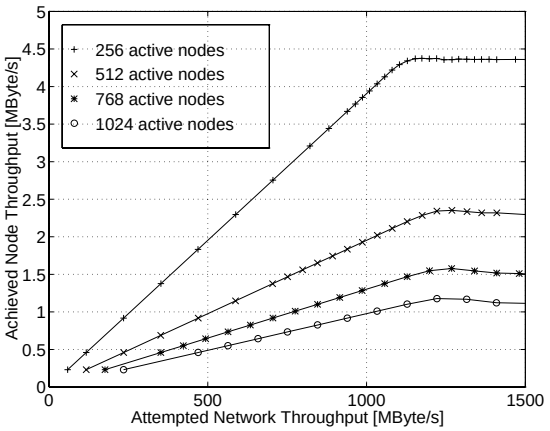


Figure 70: Node throughput versus attempted network throughput for a 1024 node torus with 4 to 16 active nodes per switch

The torus achieves a greater throughput than the grid due to the extra wrap around links which are available. However, the difference in throughput is rather small, about 15%, even though the bi-section bandwidth of the torus is twice that of the grid. This is because the deadlock-free routing algorithm for the torus cannot take full advantage of the wrap-around links. Ideally, packets would be routed along the shortest path from source to destination, i.e. crossing the least number of switches. However, there have to be exceptions from this routing strategy to avoid dependency cycles (see Figure 42 in Chapter 5) and therefore potential deadlock. This results in the wrap-around link connections of the torus not being fully utilised and their effect is only in slightly reducing the average path length of packets in the network. The network labelling scheme for the torus is explained in greater detail in [51].

For both the torus and the grid, the per-node bandwidth increases as the number of nodes which are active decreases, while the aggregate network saturation throughput is nearly constant, about 1 GByte/s for the grid and 1.2 Gbyte/s for the torus. The per-node throughput is therefore inversely proportional to the number of active nodes. The network throughput is limited by the bi-section bandwidth, since adding more nodes does not increase the total throughput.

7.3.2 Throughput of 2-dimensional grid networks

Figure 71 shows the saturation network throughput for 2-dimensional grid networks, scaling in size from 2 by 2 switches to 8 by 8 switches, under uniform random traffic as a function of the bi-section bandwidth. The packet length is 64 bytes.

For uniform random traffic, on average half of the packets are expected to cross the bi-section in each direction. The bi-section bandwidth therefore gives a good estimate of the performance of 2-dimensional grid networks and the measured network saturation throughput scales almost linearly with the bi-section bandwidth, as shown in Figure 71. A straight line has been fitted to the measured data points: the figure shows that the saturation throughput under random traffic is very close to 1.7 times the bi-section bandwidth. However, as the total network

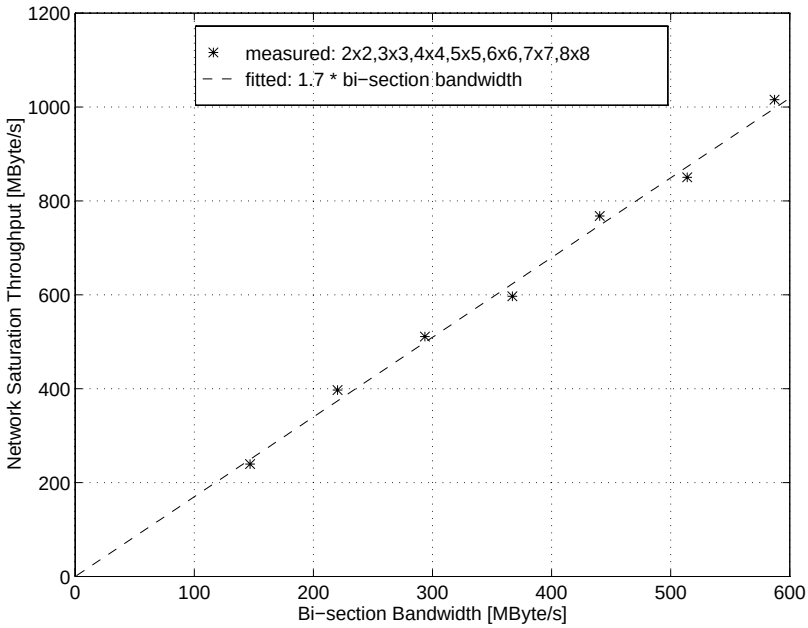


Figure 71: Saturation throughput for grid networks under random traffic

load scales with the square of the bi-section bandwidth, the per-node throughput drops with increasing network size, as seen in Figure 68.

7.3.3 Effect of different Traffic Patterns

On the grid, the average number of switches and links that a packet has to traverse depends strongly on the communication pattern. To quantify the effect of different traffic patterns on the performance, the network throughput as a function of the applied load was measured on a 1024 2-D grid. Two different permutations were used for the choice of pairs under systematic traffic. Figure 72 shows the aggregate network throughput versus the attempted throughput for an 2-dimensional 8 by 8 grid under “best case” and “worst case” permutation traffic as well as random traffic.

In order to achieve maximum throughput a “best case” scenario, which only involved localised communication and minimises contention for network resources, was used. Packets are only transferred between nodes attached to nearest neighbour switches, i.e. of the 16 end-nodes attached to each switch, groups of four nodes send packets to the corresponding nodes attached to each of the four adjacent switches. Since there are bundles of 4 links between adjacent switches, the network bandwidth matches the load. In order not to unbalance this traffic pattern, some nodes at the edge of the network do not transmit data, e.g. for the 8 by 8 grid only 896 of the 1024 nodes are active.

For the “worst case” scenario, the network is logically partitioned into four quarters, each node in every quarter communicates with the corresponding node in the opposite quarter. This mapping of sources and destinations will maximise the path length and contention for links internal to the network.

The peak throughput for the 1024-node grid under the “worst case” pattern is only 400 MByte/s (4% of the maximum network throughput), whereas the saturation throughput

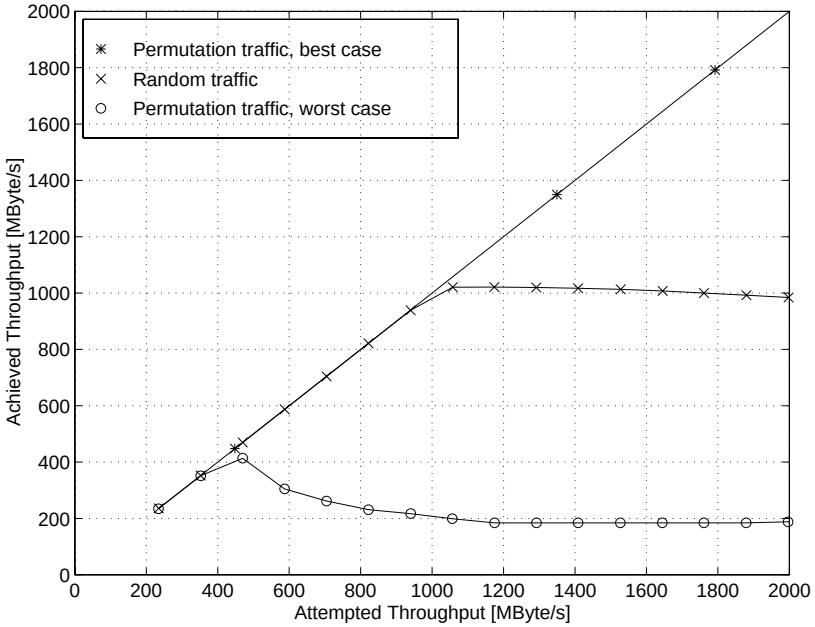


Figure 72: Achieved throughput for an 8 by 8 grid under best case, worst case and random traffic

achieved under the "best case" pattern is 8.12 GByte/s (86% of the maximum network throughput) as seen from Figure 67. With about 1 GByte/s (11% of the theoretical maximum network throughput), the peak throughput for random traffic is 2.5 times higher than the value achieved for the "worst case" permutation, but only about 12% of the throughput achieved for the "best case" permutation. In order to estimate the performance for an arbitrary combination of source-destination pairs, 10 different permutations of pairs were chosen at random and the saturation throughput was measured. The average result was 820MByte/s with a minimum of 750 MByte/s and a maximum of 880 MByte/s, i.e. the expected throughput for a permutation chosen at random is slightly worse than the value achieved for uniform random traffic. These results clearly show, that good performance on the grid requires locality in the traffic pattern.

Figure 72 shows that the throughput for "worst case" permutation traffic decreases with increasing load after the peak value. This is because the nodes on the edges of the network are blocked as the links at the centre of the network become more and more congested with the increasing load. This can be seen in figure 73, which shows the average saturation throughput for each traffic module across a 1024 grid network under "worst case" systematic traffic. It shows clearly, that the average throughput for nodes in the centre of the network is significantly higher than for the nodes at the edge, because the combination of dimension order routing and the traffic pattern results in the majority of packets passing through the centre of the network. With the dimension order routing algorithm, packets are first sent along the y-dimension and then along the x-dimension towards their respective destinations.

7.3.4 Summary of 2-D Grid and Torus Results

The results presented in this section demonstrate that the performance of the 2-D grid and torus networks studied is poor under random traffic, an 8 by 8 grid or torus only achieves about 10% of the maximum throughput. It has been shown that the limiting factor is the internal connectivity of the network, i.e. the bi-section bandwidth.

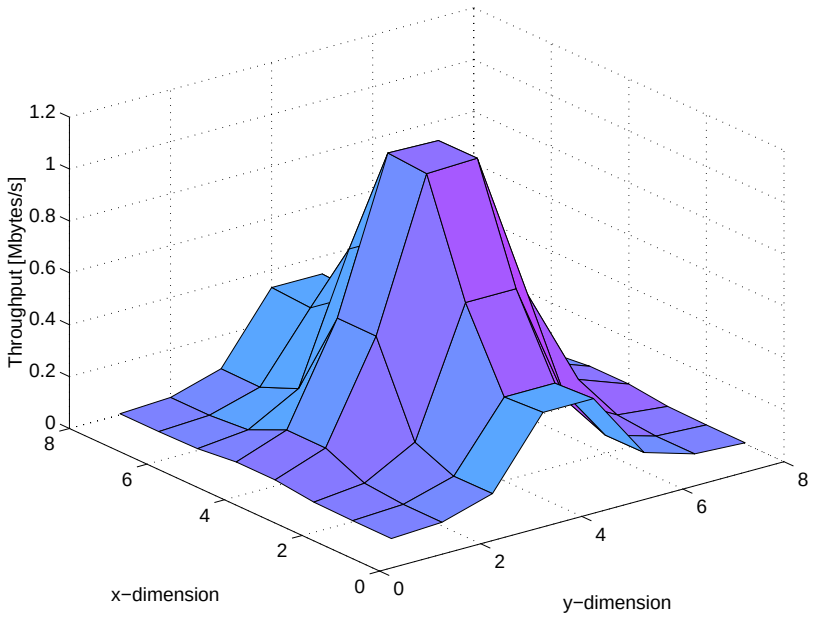


Figure 73: Throughput across a 1024 node grid using 64 byte packets and "worst case" systematic traffic

In order to improve the performance, the number of nodes active per switch had to be reduced. A grid or torus with only four nodes per switch and 16 external nodes achieved about 40% of the maximum throughput. The measurements have also shown that the performance of the grid depends strongly on the traffic pattern, e.g. the throughput of an 8 by 8 grid varies by a factor of 20 between the "best case" and the "worst case" permutation traffic pattern. On the other hand, a grid is easy to physically implement and to scale, and if the application traffic pattern uses localised communication, these networks can give good performance at a low cost. For arbitrary traffic patterns, however, the 2-D grids do not perform well, and while reducing the number of nodes per switch will improve the performance, the cost will increase as well and it becomes more cost-effective to use one of the multistage networks presented in the next section.

7.4 Performance of Clos Networks

After having established the performance limitations of the 2-D grid and torus topologies, this section presents measurements for the Clos topology as a function of the network size, the network load and the traffic pattern.

7.4.1 Throughput versus Network Size

Figure 74 shows the normalised network saturation throughput for four different size Clos networks under random traffic and systematic traffic with 64 byte packets. Networks with 64, 128, 256 and 512 nodes have been studied. The measured throughput has been normalised to the asymptotic maximum link bandwidth of 9.52 MByte/s for bidirectional traffic. Also shown is the theoretical saturation throughput for a crossbar under random traffic. For a large crossbar this is 63.2% from Equation 14 in Chapter 5.

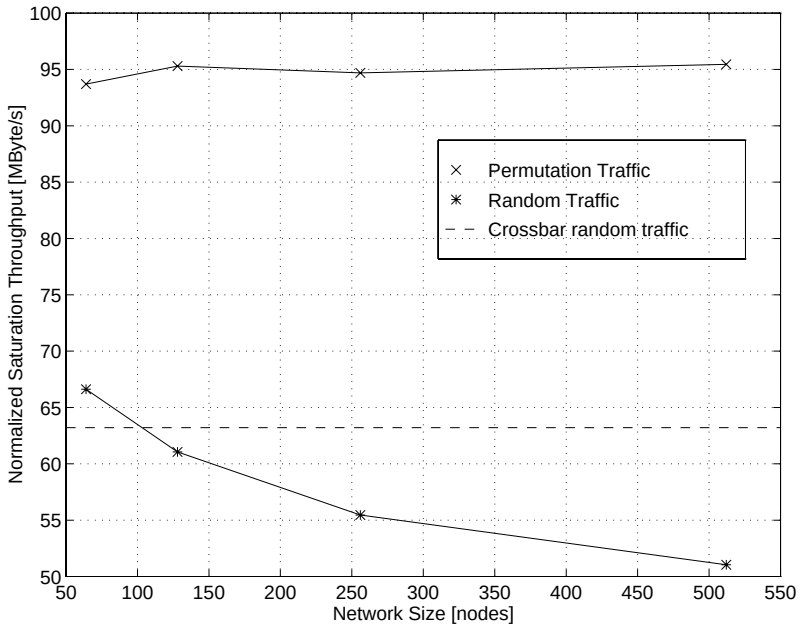


Figure 74: Normalised Throughput for different size Clos Networks under random and permutation traffic

For systematic permutation traffic, all the networks reach about 95% of the theoretical maximum throughput. For random traffic, the network utilisation is lower, e.g. 67% for a 64-node Clos. As the network size increases, the utilisation under random traffic decreases, the 512-node network only achieves about 51% of the maximum throughput, even though the bi-section bandwidth scales linearly with the number of nodes. The reason for the better performance of the smaller Clos networks is that for networks smaller than 512 nodes, there are multiple parallel links connecting each centre stage switch to each terminal stage switch. For example, for a 64-node Clos bundles of 8 links connect every centre stage switch to each of the four terminal stage switches, as shown in Figure 75 below. Therefore the centre stage switches can also exploit grouped adaptive routing, which improves the utilisation of the links going from the centre stage to the terminal stage. With grouped adaptive routing, multiple packets can be forwarded from a given centre stage switch to the same terminal stage switch without being blocked. Obviously destination blocking at the terminal stage switch will also limit the throughput which is achievable.

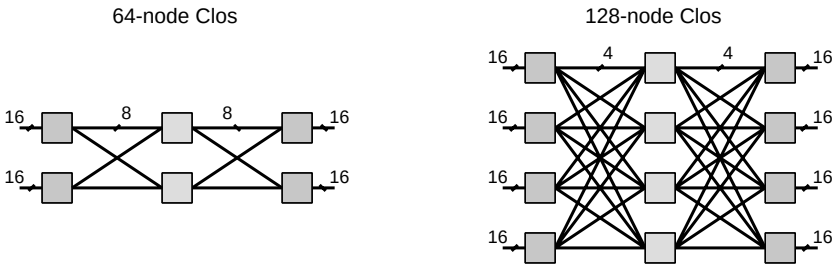


Figure 75: 64-node and 128-node Clos networks

As the network size increases, the width of the link bundles connecting terminal and centre stage switches decreases, and the number of bundles connecting to each centre stage switch increases: A 128-node Clos has 8 bundles of 4 links, the 256-node Clos has 16 bundles of 2 links and the 512-node Clos only has a 32 individual links connecting each centre stage

switch to every terminal stage switch. Therefore the utilisation of the centre stage switches decreases and the blocking probability increases. As a consequence, the total network throughput does not scale linearly with the network size.

7.4.2 Varying the Number of Centre Stage Links

In order to study how the performance of a Clos network varies with the bi-section bandwidth, the number of centre stage switches was varied between one and 16 switches on a 512-node Clos network. Figure 76 shows the average centre stage link throughput in saturation as a function of the number of links between the terminal stage switches and the centre stage switches for various packet length values under random traffic. The centre stage throughput was obtained by dividing the total measured network throughput by the number of centre stage links.

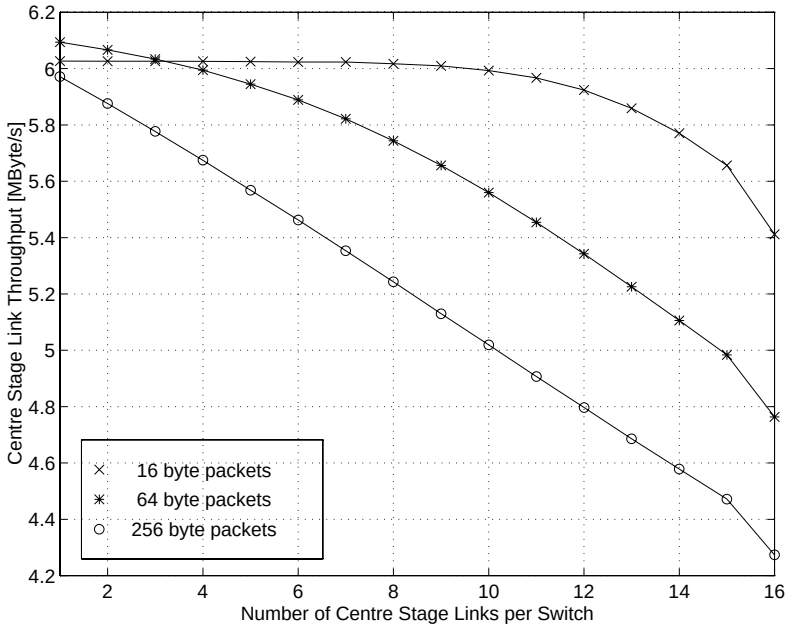


Figure 76: Centre stage link throughput for a 512-node Clos versus the number of centre stage links

The results show that the centre stage link throughput is reduced as the number of centre stage links is increased. This means, that the total network throughput will not scale linearly with the number of centre stage links. This effect is more pronounced for long packets. For example, for 16 byte packets, the centre stage link utilisation only starts to drop off for more than 10 links. The centre stage link utilisation decreases from 64% (6.1 MByte/s) of the asymptotic bidirectional link bandwidth to 57% (5.4 MByte/s). For 256 byte packets, the utilisation decreases almost linearly, the minimum value is 43%. Short packets perform better, because they can take advantage of the output buffers in the STC104 switch.

For a small number of centre stage links, the throughput is limited by contention in the centre stage switches, since there is little or no destination contention in the terminal stage switches. As the number of links between the terminal and the centre stage increases, the probability for destination contention in the terminal stage switches also increases and the per-node throughput and the centre link utilisation are reduced.

Using Equation 13 from Chapter 5, which gives the utilisation of a crossbar switch under random traffic, it is possible to build a mathematical model of the performance of the Clos for varying numbers of centre stage links. Assuming that the output links from the terminal stage switches to the centre stage are saturated because grouped adaptive routing is used on these links, the utilisation of the output links of any centre stage switch can then be calculated as follows:

$$U_{Centre}(N) = 1 - \left(1 - \frac{1}{N}\right)^{N} = 0.638 \quad \textbf{(17)}$$

where N is the number of links, i.e. 32 in this case. The value from Equation 17 is the utilisation of a 32-way crossbar in saturation under random traffic. Therefore the utilisation of the input links to the destination terminal stage switch is 63.8%. The utilisation of the destination links is then:

$$U_{Dest}(N_{in}, N_{out}) = 1 - \left(1 - \frac{U_{Centre}}{N_{out}}\right)^{N_{in}} \quad \textbf{(18)}$$

where N_{in} is the number of input links, which is equal to the number of centre stage links, and N_{out} is the number of terminal nodes, i.e. 16. Using the equation above, the total network throughput can be calculated. This is shown in Figure 77 together with the measured saturation network throughput for 256 byte packets. The agreement between the measured and the calculated values is very good.

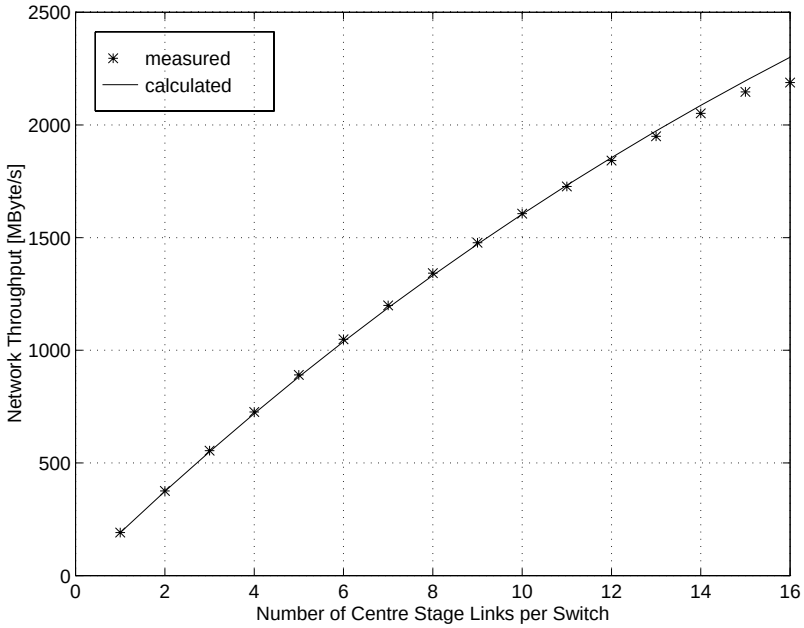


Figure 77: Network throughput for a 512-node Clos versus the number of centre stage links

7.4.3 Varying the Number of Active Nodes

In order to show how the performance of the Clos scales when the bi-section bandwidth is greater than the aggregate bandwidth of the end-nodes, the number of nodes active per terminal switch was varied. Figure 78 shows the per-node throughput of a 512-node Clos network under random traffic with 4, 8, 12 and 16 nodes active per terminal stage switch for a number

of different packet length values. The results in Figure 78 demonstrate that the per-node throughput increases when fewer nodes are active. This is clearly because there is less contention due to the over-capacity in the network in this case. As seen before, shorter packets give better performance. The maximum throughput is achieved for 16 byte packets with 128 active nodes. The achieved throughput is 90% (7.5 MByte/s) of the maximum theoretical bidirectional link bandwidth for 16 byte packets. For 64 byte packets the throughput drops from 75% to 60% of the link bandwidth when the number of active nodes is increased from 128 to 512.

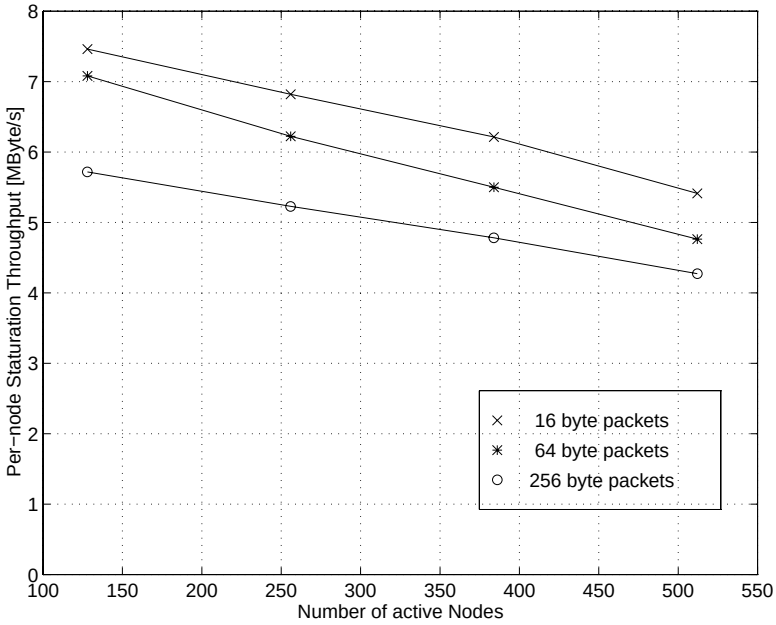


Figure 78: Per-node throughput for a 512-node Clos, varying the number of active nodes

The total network throughput can be calculated as follows: assuming that the load from the *n* active source nodes on each terminal stage switch is evenly spread across the 16 outgoing links because of grouped adaptive routing, the utilisation of these links is:

$$u_0 = \frac{n}{16} \tag{19}$$

This is also the load on each of the 32 input links to the centre stage switches, and the utilisation of the output links of these switches can be calculated using Equation 13 from chapter 5:

$$u_1 = 1 - \left(1 - \frac{u_0}{32}\right)^{32} \tag{20}$$

This is also the load on the 16 input links of the final terminal stage switch and the same equation can then be used to calculate the utilisation each of its *n* output links:

$$u_2 = 1 - \left(1 - \frac{u_1}{n}\right)^{16} \tag{21}$$

The network saturation throughput calculated using these equations is shown in Figure 79 together with the measured data for 256 byte packets. The plots shows good agreement

between calculation and measurement up to 12 active nodes per switch, for 16 active nodes per switch the measured results is about 6% lower than the calculated value.

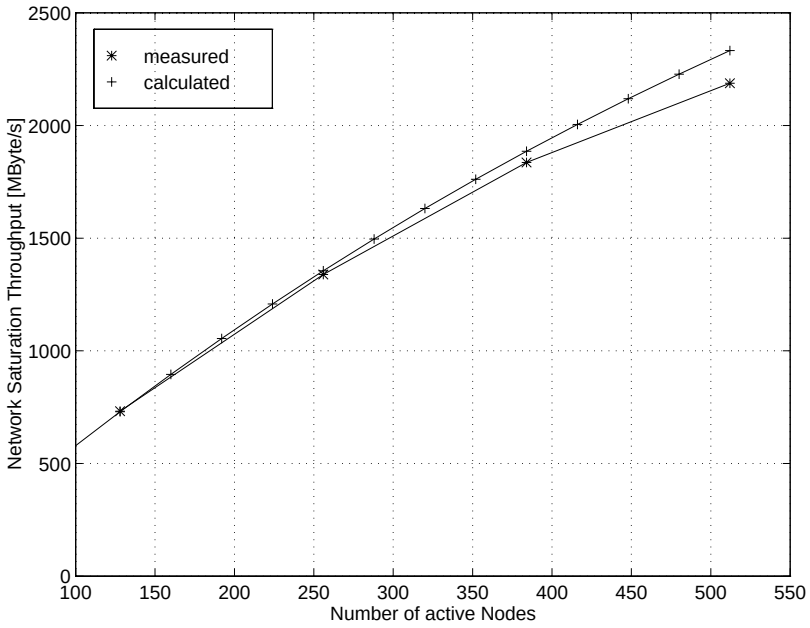


Figure 79: Calculated and measured saturation throughput for a 512-node Clos under random traffic for 256-byte packets, varying the number of active nodes

These results demonstrate that is possible to obtain significantly more than 60% link throughput under random traffic by undersubscribing the network, but also that the excess network bandwidth is not very efficiently used, since the centre stage bandwidth cannot be fully used because of destination contention.

7.4.4 Network Latency for Clos Networks

Network latency is defined as the delay from the transmission of the packet header at the source to the reception of the end-of-packet at the destination.

7.4.4.1 Average Network Latency

Figure 80 shows the average packet latency of four different size Clos networks under random traffic as a function of the aggregate network throughput. The packet length is 64 bytes. The results are produced by varying the network load and measuring the corresponding throughput and latency values.

It can be seen that the average latency always increases rapidly as the network throughput approaches saturation. The minimum achievable latency is the sum of the packet transmission time of about 6.5 µs and the switching delay. The switching delay for traversing three switches is about 2.7 µs. Therefore the latency in an unloaded network is about 9.2 µs. Figure 81 shows the average latency versus the applied network load. The latency increases when the load approaches the saturation throughput, but stays approximately constant when the network is saturated. This is because all the queues are full and the link-level flow-control prevents new packets from entering the network, while the path is blocked. The latency in saturation is between 30 µs and 35 µs, depending on the network size. Therefore the queuing

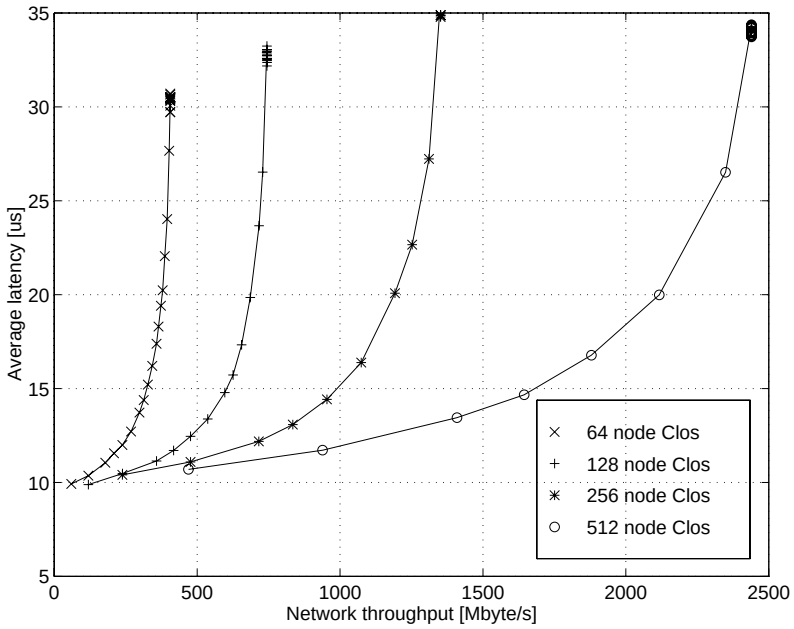


Figure 80: Latency versus throughput for 64, 128, 256 and 512 node Clos networks under random traffic

delay is approximately equal to three times the packet transmission time, which is consistent with the results from section 7.1.2, since the packets are queued for about one transmission delay for each switch they have to traverse.

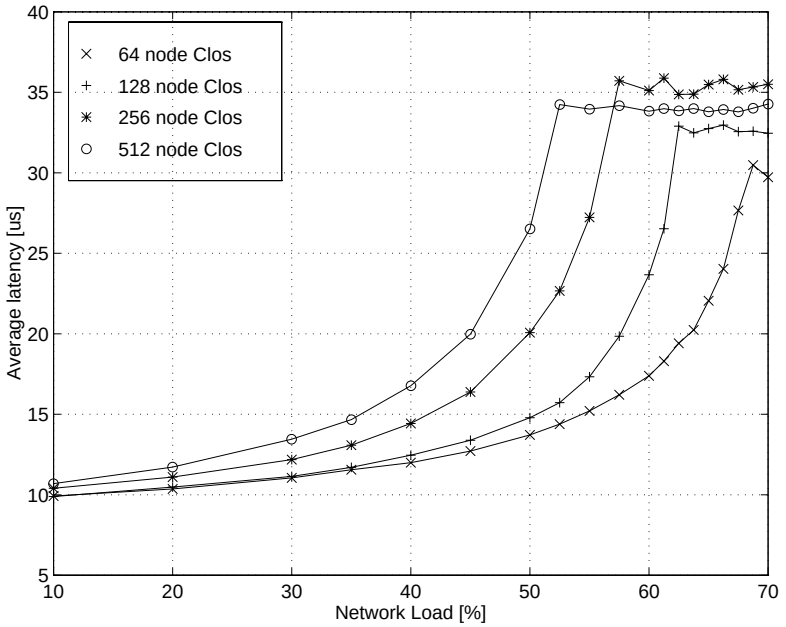


Figure 81: Average latency versus network load for different size Clos networks

7.4.4.2 Packet Latency Distribution

Some applications may also require statistical bounds on the maximum latency values occurring. This information can be obtained from Figure 82, which shows the probability that a packet will be delayed by more than a given latency value for various network loads. The results have been obtained on a 512 node Clos network. Again the traffic pattern is random,

with a packet length of 64 bytes. Each of the lines in Figure 82 corresponds to over 10 million packet latency measurements, which allows probabilities as low as 10^{-6} to be shown. The network load is varied from 10% to 50% of the maximum theoretical throughput. The 512-node Clos network saturates at 2.44 Gbyte/s, as can be seen in Figure 80, which corresponds to a load of 52%.

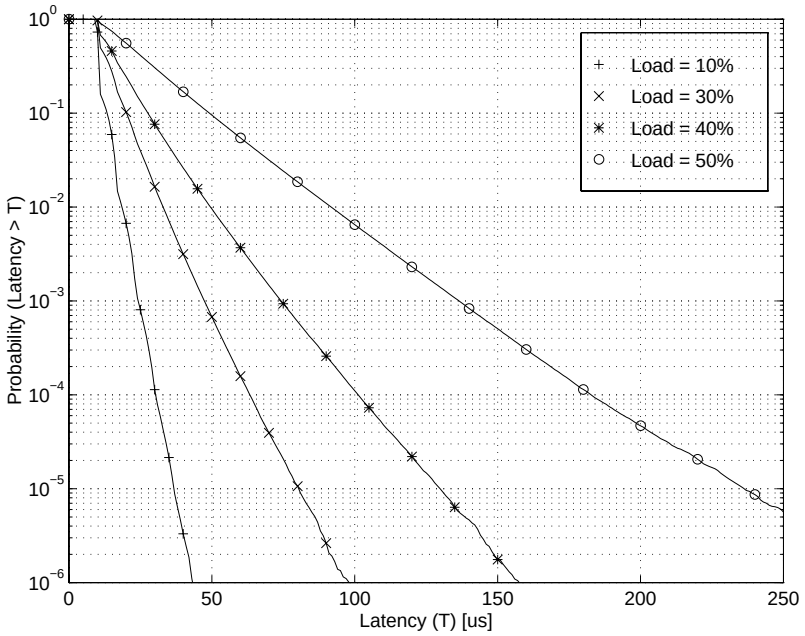


Figure 82: Cumulative latency distribution for a 512-node Clos network under random traffic

For 10% load the latency distribution is narrow, e.g. only a small percentage of the packets (0.1%) are delayed by more than two times the average latency value of 11 µs. As the network load increases, the tail of the latency distribution gets wider and near the saturation throughput a significant fraction of the packets experience a latency many times the average value, e.g. at 50% load about 0.1% of the packets are delayed by more than 5 times the average latency of 27 µs. To reduce the probability of very large latency values the network load must therefore be kept well below the saturation throughput.

7.4.4.3 Packet Delay Variation

Individual packets being transmitted across the switching fabric will experience differing amounts of delay. If the overall transit time of packets through the network has a real-time requirement, as it is the case for voice and video traffic, it is important to know what fraction of packets will be received within a particular delay. Figure 83 shows the minimum and average latency as well as the packet delay variation as a function of the average per-node throughput for a 512 node Clos network under random traffic with 64 byte packets.

The packet delay variation is shown for a probability of 10^{-3}, i.e. 99.9% of the packets will be delayed by less than the packet delay variation value. It can be seen that the spread between the minimum latency and the delay variation increases as the network load goes up. The average latency also increases, but not as fast as the delay variation. This shows that the tail of the latency distribution widens as the network approaches saturation.

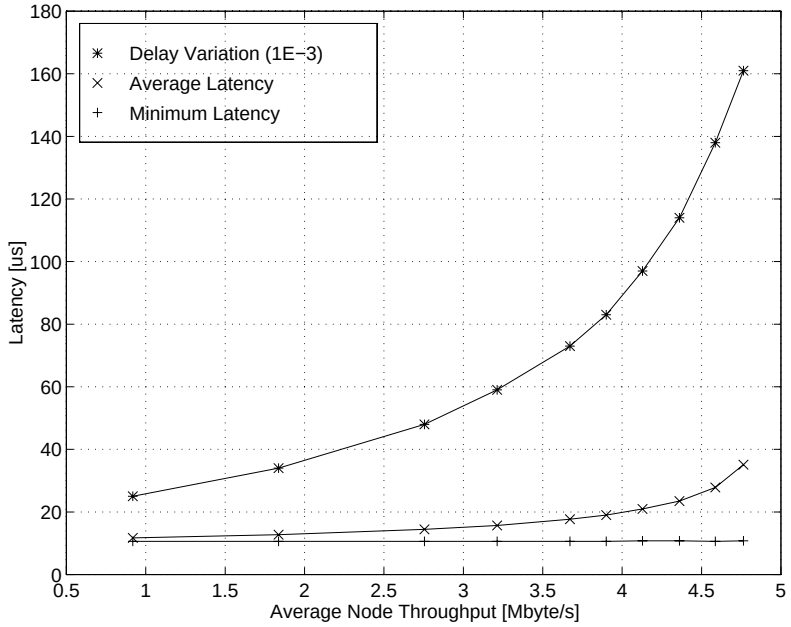


Figure 83: Packet delay variation for a 512-node Clos network under random traffic

In Figure 84 the latency distribution as a function of the network load is shown for a 64 node Clos under random traffic. The plot shows, as a set of probability density function, the likelihood of a packet experiencing a particular latency for different network loads.

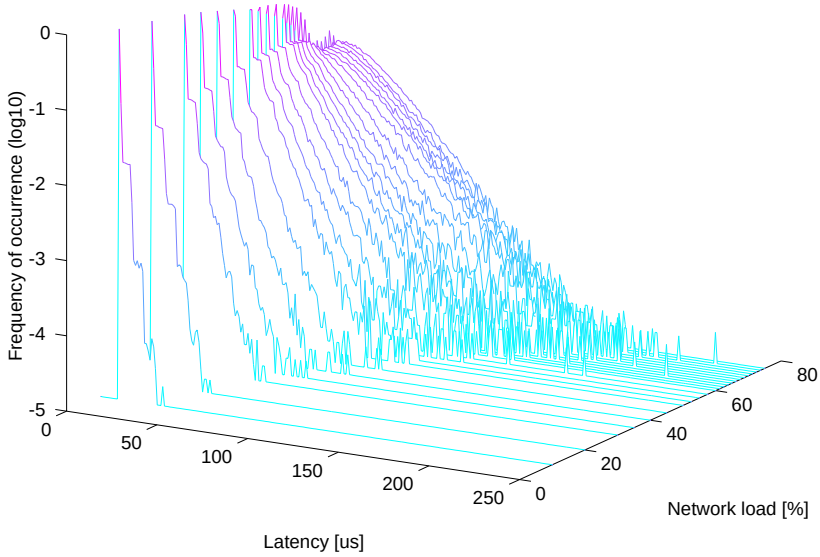


Figure 84: Latency distribution versus network load for a 64-node Clos

As can be seen, the probability that a packet experiences larger delays tails off rapidly. At low loads (less than 20% of the maximum network bandwidth) steps can be seen in the plot which correspond to the number of packets in the queue in front of the time stamp packet. In this case the majority of packets traverse the network without experiencing queuing delays, this creates the peaks shown on the left hand side of figure Figure 84. Above saturation (more than 60%) all packets are delayed. The plot also shows that the latency distributions widen as the network load increases.

7.4.5 Effect of Packet Length on Latency

All the previous latency measurement results have been obtained using a fixed packet length of 64 bytes. It is therefore interesting to see how changing the packet length affects the results. Figure 85 shows the effect of packet length on the average latency for a 512-node Clos network under random traffic for different load values. For loads below 50% of the maximum network bandwidth, the average latency increases almost linearly with the packet size for packets of more than 64 bytes. This can be expected, since the packet transmission time, and consequently also the time a packet is blocked waiting for an output link, should be proportional to the packet length. Near saturation, shorter packets give smaller average latencies. However, a network should not be operated close to saturation anyway, because of the long tail on the latency distribution seen above. Figure 85 also shows that for a given packet length, the average latency increases with the network load, as seen previously.

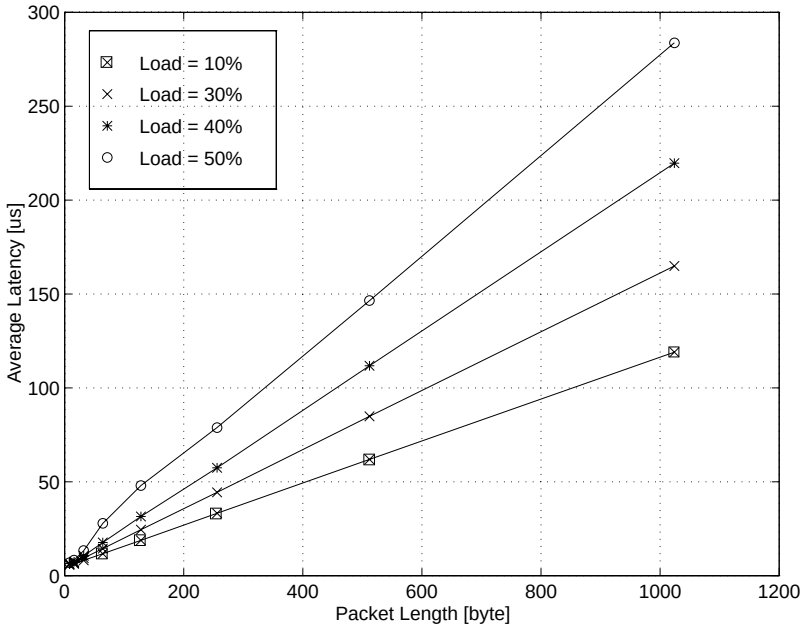


Figure 85: Latency as a function of packet length for a 512-node Clos network

These results show that under random traffic, the average packet latency cannot be improved by splitting up a given amount of data into small packets to transfer it over the network, since the delay for sending a number of small packets sequentially is the same as sending all the data in one long packet. The latency can only be improved if the small packets travel through the network in parallel. In this case however, they use up additional network resources, blocking other packets in the network.

7.4.6 Effect of Non-Uniform Traffic

All the traffic patterns studied so far, random and permutation traffic, are uniform in the sense that every node transmits and receives on average the same amount of data as all the other nodes. Although these traffic patterns provide a good way of evaluating and comparing the performance of switching networks, the traffic found in real applications is rarely completely uniform. Therefore the effect of non-uniform traffic on the performance of Clos networks has also been studied and the results are presented in this section. Two specific traffic patterns were used: hot-spot traffic and fan-in traffic.

7.4.6.1 Clos Network under Hot-Spot Traffic

One possible model of hot-spot traffic introduces a single hot-spot into a uniform random traffic pattern. This type of traffic could occur in distributed file systems or shared memory multiprocessors [51]. A random traffic pattern is used, but the destination distribution is non-uniform, such that in addition to the random traffic, each transmitting node sends a fixed proportion of packets to a single hot-spot terminal. The hot-spot terminal only receives the hot-spot traffic. Figure 86 shows the average per-node saturation transmit rate for a 256-node Clos under this type of hot-spot traffic. Also shown is the receive rate at the destination hot-spot. The traffic is unidirectional, 128 nodes are transmitting, the other 128 nodes are receiving, the packet length is 64 bytes.

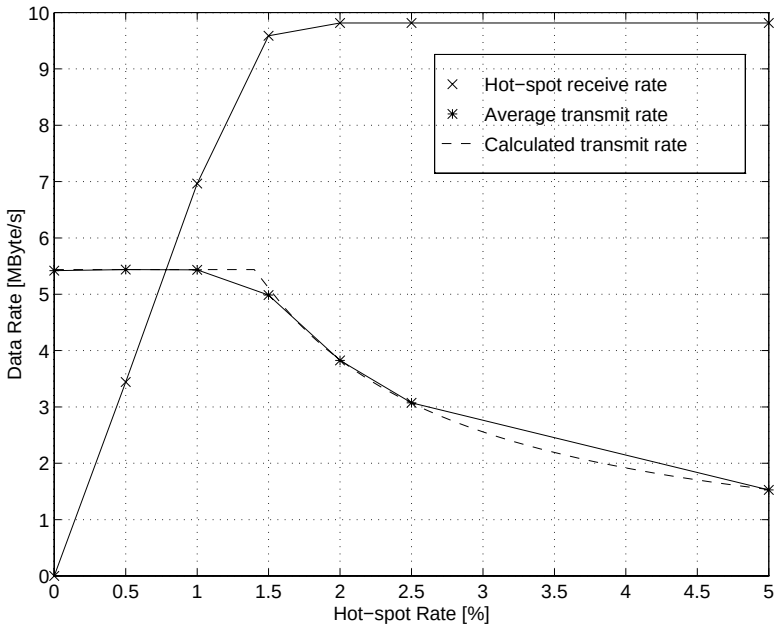


Figure 86: Per-node throughput for a 256-node Clos under hot-spot traffic

It is clear from the results, that the network is extremely sensitive to this non-uniform traffic pattern, because a 2% hot spot causes a 30% loss of the total throughput of the network. This is due to an effect known as tree-saturation. In the presence of a hot-spot the achieved network throughput only follows the applied load until the link to the hot-spot destination becomes saturated. At this point all the links which can route packets to the hot-spot terminal also go into saturation and a tree of saturated links form, extending back to all the input terminals, which causes the overall network performance to drop drastically. Below saturation, the data rate into the hot-spot terminal is:

$$R = N \cdot h \cdot T \tag{22}$$

where N is the number of transmitting nodes, h is the hot-spot rate, i.e. the fraction of packets each node sends to the hot-spot destination, and T is the attempted transmit rate. When the link to the hot-spot terminal is saturated, the asymptotic network throughput can be calculated as follows:

$$T = \frac{BW}{N \cdot h} \tag{23}$$

where *BW* is bandwidth of the link to the hot-spot, 9.8 MByte/s is this case. The calculated throughput is shown in Figure 86 together with the measured data. As can be seen, the agreement is very good. From Equation 23 it is also clear, that the impact of the hot-spot can be reduced by increasing the bandwidth to the hot-spot destination, e.g. by using additional links, which can also be grouped.

7.4.6.2 Clos Network under Fan-in Traffic

The fan-in traffic pattern occurs frequently in the data acquisition and trigger systems of high energy physics experiments, where a large number of data sources send packets to a set of destinations for processing. The number of sources is usually larger than the number of destinations. The traffic flow is inherently unidirectional. This type of traffic has been studied on a 512-node Clos network, where a number of source nodes send packets randomly to a set of destination nodes. The ratio of sources to destinations was varied, thereby changing the fan-in ratio from 480-to-32 to 256-to-256. Each node acts either as a data source or a data sink, the sources and destinations were distributed across the terminal stage switches. Figure 87 shows the average per-node saturation throughput measured for 256 byte packets versus the number of source nodes.

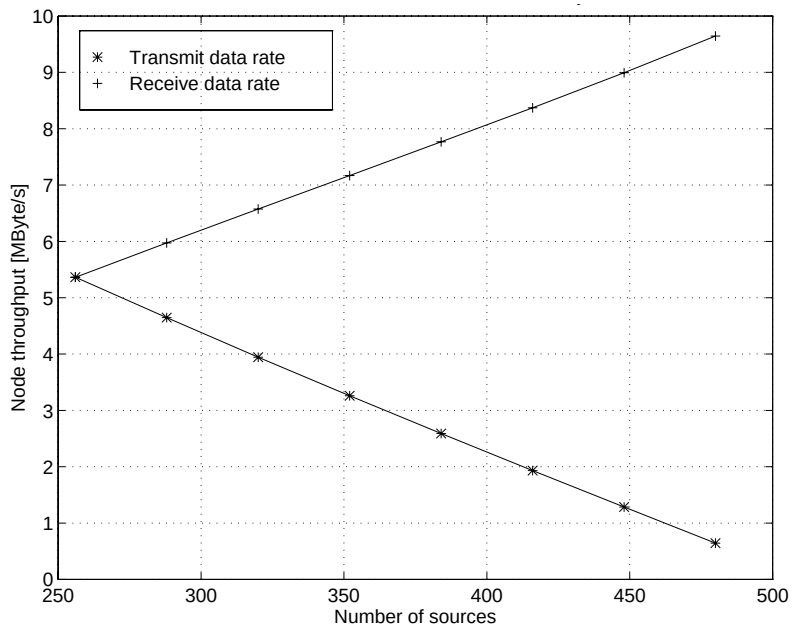


Figure 87: Per-node throughput for a 512-node Clos network and fan-in traffic

The receive rate is highest for the case of 480 source nodes, where 15 sources and 1 destination are allocated to each terminal stage switch. This is because the effect of head-of-line blocking is reduced. The transmit rate, however, is lowest because the link bandwidth of the receiver is shared among 15 sources. As the number of sources is decreased, the transmit rate goes up, since the ratio of sources to destinations also decreases. On the other hand, the receive rate goes down because of increased contention and consequently head-of-line blocking. The achieved throughput for 256 sources and 256 destinations is very close to the value obtained for uniform random throughput. The results show that it is important to provide sufficient bandwidth to the destinations, otherwise the network throughput under fan-in traffic will be limited. It can also be seen that the achieved receive rate can be higher than for random traffic, depending on the fan-in ratio.

7.4.7 Summary of Clos Network Results

The results presented in this section demonstrate, that the performance of the multistage Clos networks studied under random traffic is limited to values between 50% and 65% of the theoretical network bandwidth, depending on the network size. This is because of contention in the centre stage and final terminal stage switches. Smaller networks perform better because of the bundles of grouped links between the centre and terminal stage switches. As previously shown for the 2-D grid and torus topologies, the performance can be improved by reducing the number of active nodes per terminal switch, e.g. up to 90% of the theoretical link bandwidth can be achieved by only using only 128 out of the available 512 nodes for 16-byte packets. A mathematical model has been given which reflects this behaviour accurately. However providing over-capacity is expensive and the available network bandwidth is not used efficiently.

The network latency has also been studied and it has been shown, that the average latency under random traffic only increases by about 3 times the packet transmission delay when moving from an unloaded network into saturation. This is nearly independent of the network size for the networks which were studied. However, as the load increases the tail of the latency distributions widens and a significant fraction of the packets experience large delays. It has also been shown that the latency scales linearly with the packet length.

Finally it has been demonstrated that non-uniform traffic affects network performance significantly. A single destination hot-spot with only a small fraction of the traffic directed to it reduces the network throughput considerably. A theoretical upper bound for the throughput achievable under hot-spot traffic has also been given. It has been shown that under fan-in traffic, where a set of source terminals sends to a smaller set of destination nodes, the performance depends strongly on the fan-in ratio.

7.5 Packet Transmission Overhead

The results in section 7.2.3 show, that the network performance is best for small packets. This is, however, only the case if small packets can be sent efficiently, i.e. if the packet transmission overhead is low. The overhead in dispatching packets in the traffic nodes is determined by hardware and is small, approximately 500 ns. This will not in general be the case when interfacing links to a microprocessor. To demonstrate the effect of an increased packet transmission overhead, the dispatching delay has been artificially increased. Figure 88 shows the network throughput relative to the maximum throughput versus the packet overhead for a 256-node Clos under random traffic.

The fall-off in performance is particularly marked for short packets. For example, for 16 byte packets and a packet transmission overhead of 10μs the throughput drops to 20% of its maximum value. For 64 byte packets the throughput achieved in this case is still only 55% of the network limit. These results clearly underline the importance of an efficient interface to the network, otherwise the node will become the limiting factor on the overall network performance. This becomes even more significant as the link speed increases; an example are the Gigabit links now being introduced into commodity computing.

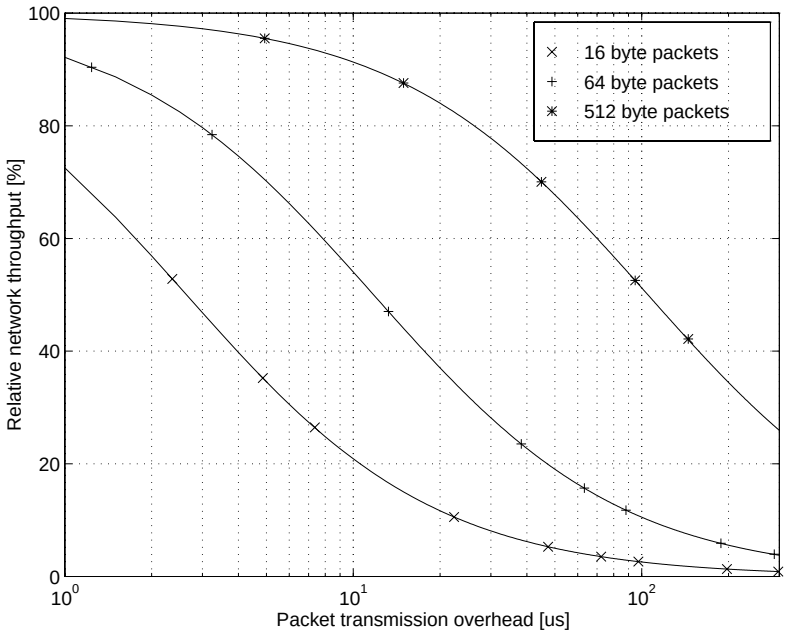


Figure 88: Relative network throughput versus packet transmission overhead for a 256-node Clos network under random traffic

7.6 Comparison of Simulation and Measurement

Within the Macramé project, a simulator has been developed specifically for simulating DS-Link networks. A model of the DS-Link and the STC104 switch have been created for this simulator. These models were calibrated against measurements taken on the network test-bed hardware. The simulator was used to simulate the performance of a range of different DS-Link networks. The results have been compiled into a book, which also contains more detail on the simulator calibration [51]. An example of the calibration results is presented below. A 64-node Clos network has been simulated using the DS-Link network simulation package. Results from simulation and measurement have been compared and are presented in Figure 89, which shows the latency distribution for 64 byte packets and random traffic at 50% load. The majority of packets pass through the network without being queued, corresponding to the peak at 12 μs. The minimum latency value is 10 μs, which consists of 2.7 μs latency for passing through the three STC104 switches between the source and destination node and 6.6 μs for the transmission of the 64 byte packet over the DS-Link. The remaining delay is due to delays incurred in the sending and receiving intelligent nodes. It can be seen that the agreement between simulation and measurement is excellent.

Network simulation can be a very useful tool for understanding the performance of switching networks, provided that the models used are accurate and have been calibrated. However, simulations are expensive. Large networks can take days of processor time to simulate a few milliseconds of real time. The ratio of simulation time to real time for the 64-node Clos results shown above was estimated as about 10^6. This means that typical simulations can only collect statistics on a limited number of packets and can therefore reliably only predict events with probabilities down to 10^{-3}. Results such as the latency distributions presented in section 7.4.4.2 could not have been obtained using the present network simulator.

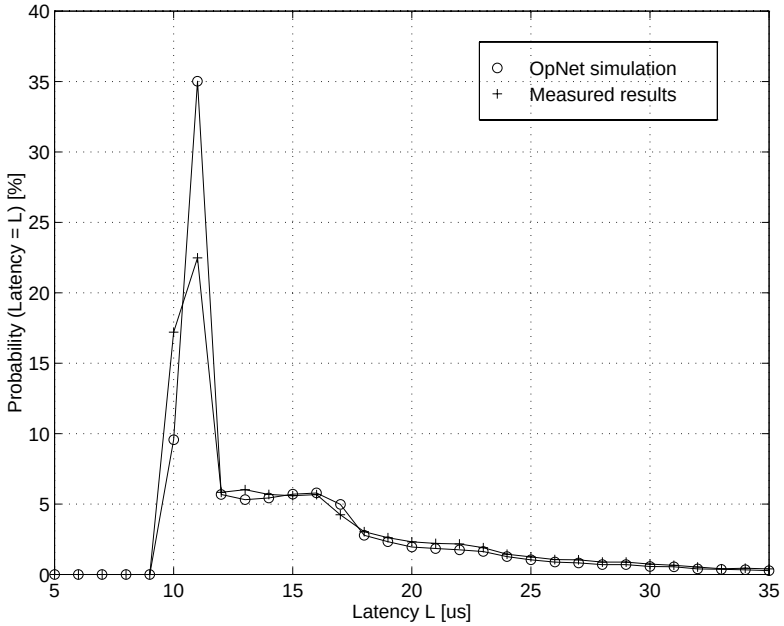


Figure 89: Simulated and measured latency distributions for a 64-node Clos network under random traffic

7.7 Effect of Different Routing Algorithms

This section shows how the different routing algorithms supported by the STC104 switch, namely grouped adaptive routing and universal routing, affect the performance of Clos and 2-D grid networks.

7.7.1 Grouped Adaptive Routing

All the measurements presented so far have been made using grouped adaptive routing. In order to quantify the impact of this feature of the STC104 packet switch, deterministic routing and grouped adaptive routing have been compared on the Clos topology. With deterministic routing, routing channels are evenly spread across the centre stage links. Figure 90 shows the average network latency versus network throughput for a 512-node 3-stage Clos network under random traffic with 64-byte packets. The network load was increased until saturation occurred. Using grouped adaptive routing results in a nearly 20% higher saturation network throughput as well as lower average latencies. This is because the adaptive routing technique enables the use of alternate centre stage paths when an output link is blocked, thereby allowing a better utilisation of centre stage switches of the Clos network.

The advantage of grouped adaptive routing is even more significant for permutation traffic. Table 12 shows the per-node saturation throughput for a 512-node Clos network under permutation traffic. The source to destination mapping was chosen to maximise the contention for the centre stage links, i.e. all the nodes on the first board send to nodes on the last board, the nodes on the second board send to the second to last board, and so on. The results show that

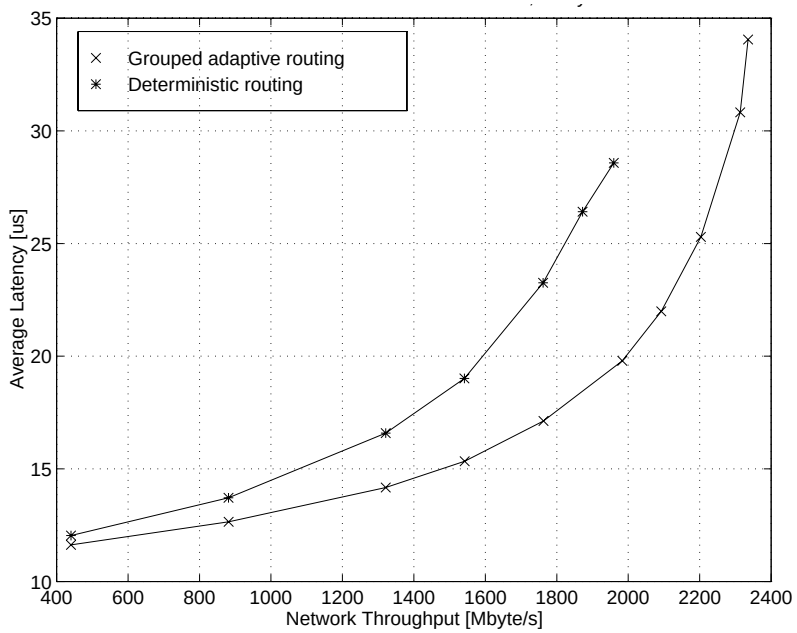


Figure 90: Deterministic and grouped adaptive routing on a 512 node Clos network under random traffic

the throughput achieved for deterministic routing is only about 10% of the value for grouped adaptive routing, which is very close to the maximum link bandwidth.

Table 12: Per-node throughput under permutation traffic

Routing	Per-Node Throughput [MByte/s]
Grouped Adaptive	9.5
Deterministic	0.93

The reason is that for the deterministic labelling scheme in combination with the chosen permutation, all the packets from one board have to share the bandwidth of a single link to the centre stage. With grouped adaptive routing, the load is evenly spread across the centre stage links and the centre stage bandwidth matches the applied load, since there is an equal number of links to the terminals and to the centre stage switches.

The source-destination mapping used to obtain the results above provides a worst case for deterministic routing and a best case for grouped adaptive routing. In order to compare the performance for an arbitrary combination of source-destination pairs, 100 permutations of pairs were chosen at random and the saturation throughput has been measured with grouped adaptive as well as deterministic routing. Figure 91 shows the histogram of the measured average per-node saturation throughput for 64 byte packets.

The mean achieved per-node throughput for grouped adaptive routing (5.57 MByte/s) is only slightly better than the value obtained for random traffic (4.86 MByte/s). The results also show, that the spread of the average node throughput is smaller for grouped adaptive routing. In addition, the mean achieved node throughput for grouped adaptive routing is about 0.5 MByte/s higher than the value obtained for deterministic routing.

The results presented in this section show that grouped adaptive routing achieves somewhat better performance than deterministic routing for both random and permutation traffic. How-

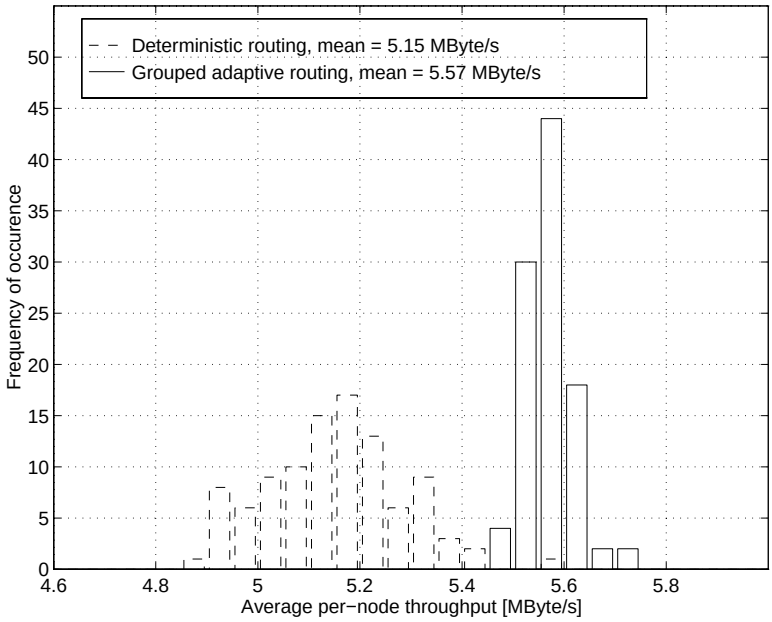


Figure 91: Histogram of the average throughput for permutation traffic on a 512-node Clos

ever, with adaptive routing, packets travelling from the a given source to the same destination can take different paths through the switching fabric, and might therefore arrive out of order at the destination.

7.7.2 Universal Routing

The STC104 packet switch also supports the so-called universal routing strategy, which is supposed to improve the performance of large networks [15]. The universal routing algorithm on the 2-dimensional grid works as follows: as a packet enters the network it is first routed along the horizontal direction to an intermediate destination switch which is chosen at random. The random header is deleted there and routing proceeds as in the deterministic case, i.e. first vertically and then horizontally towards the destination (see also section Figure 7.3.3). Universal routing is supposed to distribute the traffic evenly over the entire network, removing hot-spots and thereby reducing worst case latencies.

Figure 92 shows the average latency of a 400 node 2-dimensional grid under random traffic as a function of the network throughput with and without universal routing. The packet length is 64 bytes. The latency curve for the case without universal routing rolls back because the network throughput on the grid actually decreases with the attempted throughput, when the applied load is above the maximum achievable throughput. This effect has already been discussed in section 7.3.3. The results in Figure 92 show that the network saturates much earlier with universal routing, and that the latency also increases much faster. The saturation throughput with universal routing is only half of that without. This is because, in order to avoid deadlock, some links have to be reserved for the random phase thereby reducing the effective bisection bandwidth for the destination phase. In the horizontal direction, the width of the link groups is only two and not four links. This causes greater contention and hence an increase in latency. The results do not show any advantage of universal routing.

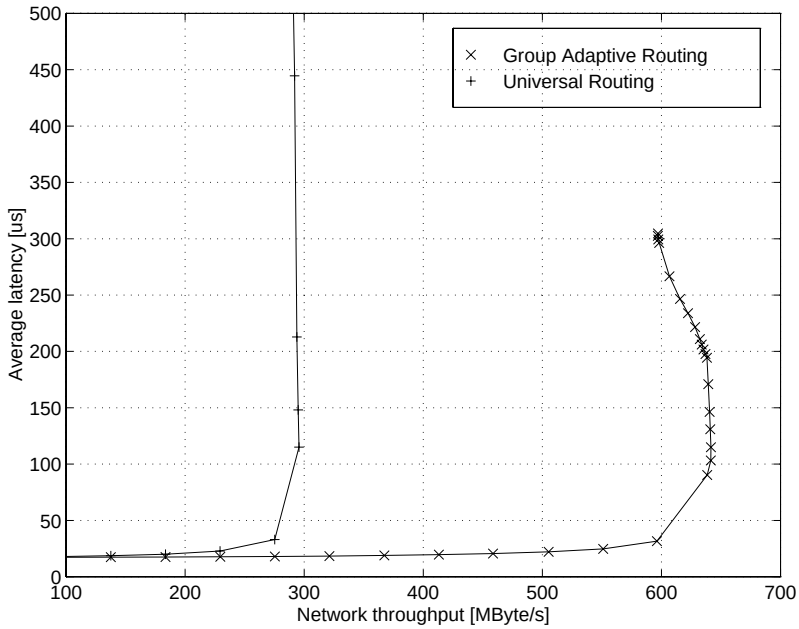


Figure 92: Universal Routing on a 2-dimensional Grid under random traffic

7.8 High Energy Physics Traffic Patterns

The data-acquisition and triggering systems of next-generation High Energy Physics (HEP) experiments [1, 2] at the CERN Large Hadron Collider (LHC) in Geneva will require high-speed packet switching networks connecting a large number of nodes. Network performance has been measured under the traffic patterns expected within the second level trigger of the ATLAS experiment. Measurements have been performed for two of the proposed architectures of the second level trigger system. The results for architecture B are presented here to illustrate the use of IEEE 1355 link and switch technology and Clos networks in a high energy physics application. The results for the second possible trigger system architecture, known as architecture C, which is based on a single large switching fabric, are presented in [52].

7.8.1 Second Level Trigger Architecture B

This architecture is based on a number of different switching networks, one per subdetector, plus a global network. Figure 93 shows the architecture studied. The second level trigger only processes data from areas of the detector indicated by the first level trigger as regions of interest (RoIs) for that event. Thereby the total data volume to be processed by the second level trigger system is reduced to a few percent of the total data. The traffic shows a fan-in pattern, i.e. several sources (level two buffers) are sending to the same destination (Feature Extractor processors or FEX). Several FEX processors are active per event. The results presented use parameters based on the Silicon Tracker (SCT) subdetector.

The requirements of the SCT have been taken from ATLAS internal documents [53], [54], which have been used to generate event description files. The total number of buffers in the SCT is estimated to be 256. Each buffer sends event fragments of 1032 bytes to the processors. All the data are sent in a single packet. The number of FEX processors was chosen such that the receive rate of the input link would be about 6 MByte/s at an event rate of 100kHz.

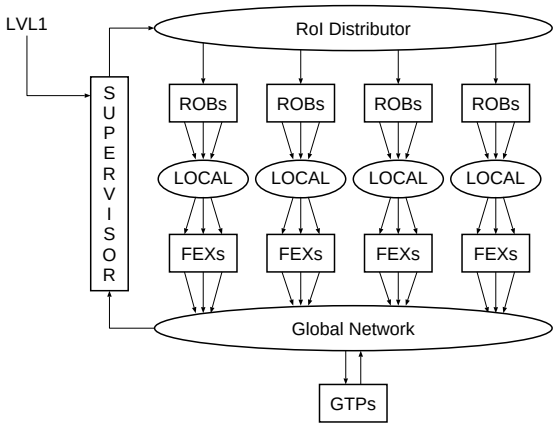


Figure 93: ATLAS second level trigger

Earlier tests indicated that this rate was sustainable under ATLAS-like traffic patterns [55]. The number of buffers active per event varies; the average value extracted from the event files was 4.64 buffers per event. This means that the total attempted throughput will be 4.8 MByte per kHz. Therefore the number of destinations required to keep the receive rate below 6 MByte/s at an event rate of 100kHz is 80 processors. The number of buffers sending data to the same processor also varies, on average 4 buffers send data from one RoI to the same processor. Multiple RoIs can be active per event. The assignment of processors is done using a round-robin schedule.

Two different mappings of sources and destinations onto a 512-node Clos network have been investigated: grouped and distributed. In the grouped case all 80 FEX processors are connected to the last 5 terminal stage switches, i.e. 16 FEX processors per switch. In the distributed case the FEX processors are connected to the last 16 terminal stage switches, i.e. 5 FEX processors per switch. In all measurements the 256 buffers are connected to the first 16 terminal stage switches. Figure 94 shows the total network throughput versus attempted event rate for a 512-node Clos.

The achieved network throughput is proportional to the attempted event rate until congestion causes the network to saturate. The results show a maximum sustainable event rate of about 120 kHz. This is 20% higher than the 100 kHz expected rate for the second level trigger. For the configuration with the grouped sources and destinations the network saturates at around 90 kHz. The improvement from grouped to distributed FEX processors is due to reducing the contention at each terminal stage switch and therefore reducing the effect of head-of-line blocking. The average receive rates for the individual FEX processors were 5.3 MByte/s and 7.2 MByte/s for the grouped and distributed cases respectively. These amounts correspond to 53% and 72% of the theoretical maximum bandwidth for an individual link (9.97 MByte/s). These values are higher than the per-node receive data rate achieved under random traffic, which is only about 4.2 MByte/s for a 512-node Clos.

In order to quantify the impact of grouped adaptive routing on the network performance under this type of traffic, the measurement was repeated for the grouped configuration with deterministic routing and with grouped adaptive routing. Figure 95 shows the achieved network throughput versus the attempted event rate for these two cases.

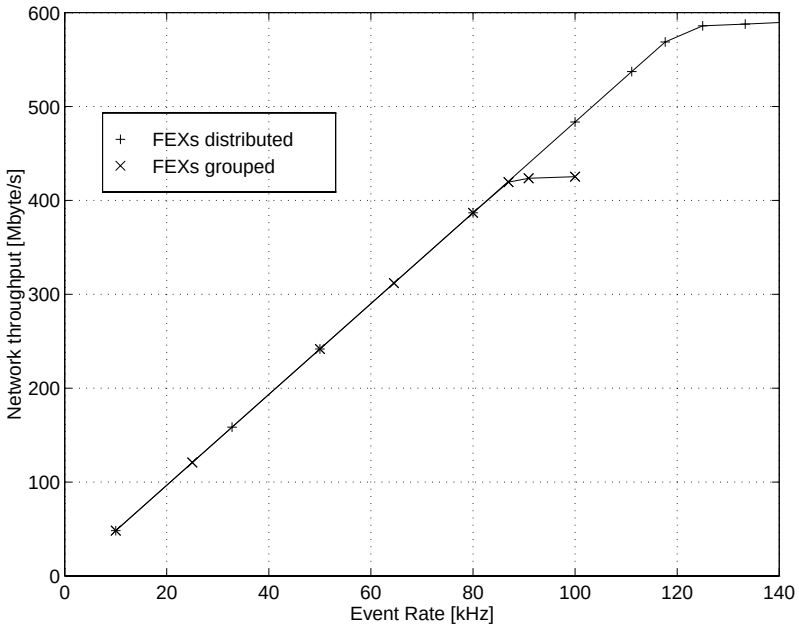


Figure 94: Achieved throughput versus attempted event rate for a 512-node Clos under ATLAS second level trigger traffic

The maximum achieved event rate using deterministic routing is only about 50 kHz compared to the 120 kHz achieved with grouped adaptive routing. The reason for this large performance improvement is that grouped adaptive routing effectively distributes the load across the centre stage switches of the Clos, which results in a significantly higher utilisation of the available network bandwidth. The results demonstrate that the use of grouped adaptive routing is essential to obtain good performance on a Clos network for ATLAS second level trigger style traffic.

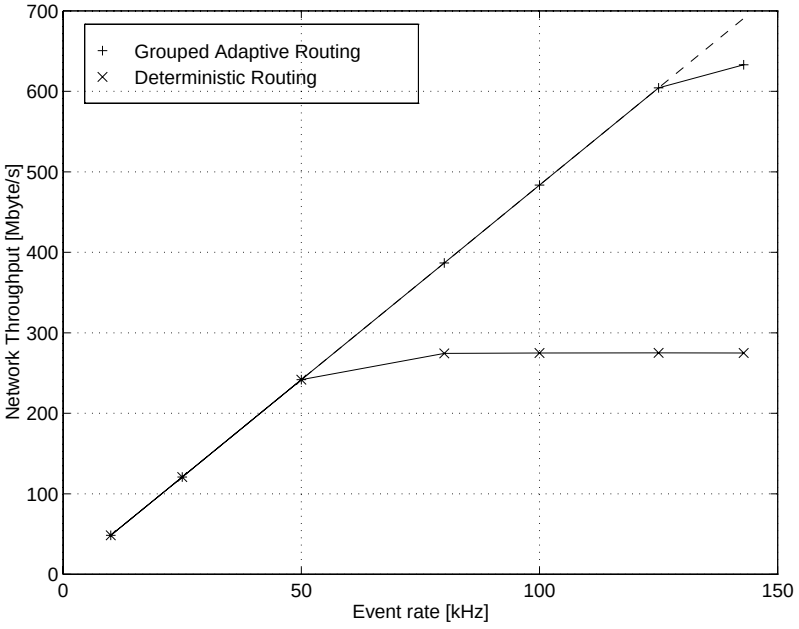


Figure 95: Network throughput versus event rate for a 512-node Clos under ATLAS trigger traffic with grouped adaptive and deterministic routing

It is also interesting to see whether, for the specific traffic pattern studied here, the full network connectivity of the Clos is really required. This has been done by reducing the number of links to the centre stage on each terminal stage switch (see also section 7.4.2). This effectively reduces the number of switches in the centre stage. Figure 96 shows the achieved network throughput versus the attempted event rate for varying numbers of links from the terminal stage switches to the centre stage.

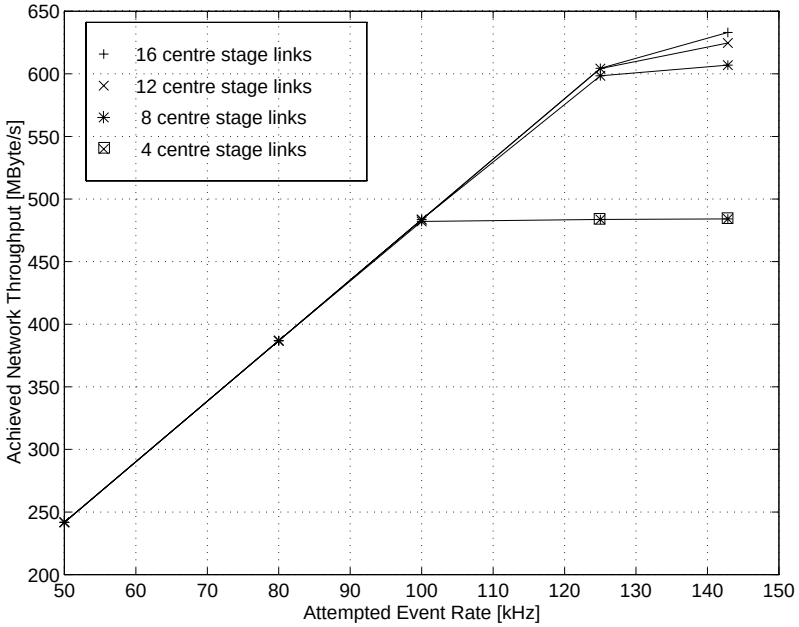


Figure 96: Network throughput versus attempted event rate for a 512-node Clos under HEP traffic with varying numbers of centre stage links

When reducing the number of centre stage switches to 4, the event rate is limited to about 100 kHz. For the case of 8 and 12 centre stage links per terminal switch, the achieved event rate is nearly the same as for the full network with 16 centre stage switches. This is because for the specific pattern studied, only a subset of the sources are active at any given time. In addition, only 336 (256 sources, 80 destinations) of the 512 terminal nodes are used in this setup and there are only 8 sources per terminal stage switch. These results demonstrate, that it is possible to reduce the centre stage connectivity and thereby the network cost with only a very small reduction in the network performance.

7.8.2 Summary of HEP Traffic Results

It has been shown that the event rate achieved with a 512-node Clos applied to the SCT subdetector of the ATLAS level two trigger is about 120 kHz. The distribution of sources and sinks across the network affects the performance of the network and should be considered when implementing the final system. The adaptive routing strategy implemented in the STC104 is essential for achieving good performance. Results for the networks of the other subdetectors of the second level trigger and the global network have also been studied and have been presented in [56] and [57]. The results demonstrate, that a system based on 100MBaud links and switches is capable of meeting the network needs of the ATLAS second level trigger.

7.9 Reliability

Differential DS-Link cable connections have been tested for immunity against electromagnetic interference according to the IEC 801-4 standard [30] (see also chapter 3). The setup passed test severity level 2, which corresponds to the environment likely to be found in the control room of an industrial or electrical plant. In order to further quantify the reliability of DS-Link systems, long-term error test were performed using the 1024 node 8 by 8 grid network. The 1024 node grid contains a total of 1344 active DS-Links, about one third of these links use differential buffers and 2 meter twisted pair cables. The others are single-ended on-board connections. The system was run continuously for over 200 hours without observing any link errors. This translates to a per-link error rate of better than $9.6 \cdot 10^{-18}$.

7.10 Summary and Conclusions

A large packet switching system based on the DS-Link technology has been constructed and is performing reliably. This system has been used to provide quantitative measurements of the performance of 2-dimensional grid, torus and Clos topologies. The results show that although grid networks are easier to physically implement, the Clos networks clearly give better performance. Given the type of traffic, the required throughput and the maximum latency, it is possible to use the results presented to evaluate the suitability of a given topology to meet those requirements. The network designer needs to consider not only the average latency, but also the effect of the long latency tail occurring in packet switching networks under contention. The measurements presented give an upper limit of the network performance obtainable with this technology, the performance could be reduced further if the network interfaces are unable to handle the low packet overheads required to sustain the data rates for short packets. Measurements from the testbed have also been used to calibrate and verify simulation models of IEEE 1355 networks. In practice, the system is extremely stable and measuring the upper limit of the error rate was governed principally by unstable Ethernet interfaces and power failures. The measurements performed for HEP specific traffic patterns demonstrate that a system based on 100 MBaud links and switches is capable of meeting the network needs of the ATLAS second level trigger.

Chapter 8
Conclusions

8.1 Achievements

The objective of this thesis was to evaluate the performance of the IEEE 1355 point-to-point serial link technology and the potential applications of large switching networks using this technology particularly in High Energy Physics. In the first part of this thesis the performance and reliability of the basic point-to-point interconnect technology over electrical and fibre optic media were examined. These studies were carried out while the IEEE 1355 standard was still being finalised and have therefore provided useful input to the working group which established the standard, of which the author was an active member.

Extensive tests and measurements have shown that differential DS-Link connections over twisted-pair cable running at 100 MBaud are very reliable over distances of up to 15 meters.

A protocol verification prototype of a fibre optic interface for DS-Links, using the IEEE 1355 TS-encoding, has been developed and tested. The fibre optic interface allows extended DS-Link connections over distances longer than the 15 meters possible with differential electrical transmission. The fibre optic link was tested over a 200 meter connection and has also proved to be very reliable.

Susceptibility to electromagnetic interference was observed on DS-Link based equipment which was being used in the CPLEAR experiment at CERN. Tests eventually showed, that the problem was due to common mode limits being exceeded on differential links as a result of poor cable screen grounding. A test bed was established according to the IEC 801-4 standard for EMC susceptibility to provide some quantitative measurements of the magnitude of the problem. A set of recommendations to reduce or eliminate the interference problem was produced.

The second part of the thesis was concerned with the design, construction and evaluation of a 1024 node packet switching network of variable topology using IEEE 1355 DS-Link technology. The nodes are interconnected by a switching fabric based on the STC104 packet switch. The system has been designed and constructed in a modular way to allow a variety of different network topologies to be investigated. The testbed allows the network performance to be measured in terms of throughput and latency under programmable traffic conditions. The full scale network with 1024 nodes was successfully implemented. The Macramé testbed demonstrated that very large switches can be built with high reliability. No transmission errors have been detected in operating the full 1024 node system continuously over periods of over 200 hours.

This network testbed is believed to be unique in its ability to measure network performance under well controlled and well defined conditions. To the author’s knowledge, there is no other interconnect for which such a large and controlled test environment has been set up.

The system has then been used to provide quantitative measurements of the performance of 2-dimensional grid, torus and Clos topologies. The effect of various traffic patterns on network latency and throughput were investigated for networks of up to 1024 terminal nodes. Mathematical models that predict the network performance have been presented for some configurations and show close correlation to the measured results. Measurements from the testbed have also been used to calibrate and verify simulation models of IEEE 1355 links and switches [9].

The network testbed has also been used to study the expected performance of two different architectural options of the ATLAS second level trigger system. It was demonstrated, that a switching fabric based on the IEEE 1355 technology could meet the required network performance of the ATLAS second level trigger. The results of these studies have been presented in internal ATLAS publications [52, 56].

The results of the work presented in this thesis formed the basis for a number of papers presented at various conferences and publications in international journals [6, 39, 55, 57, 58, 59, 60, 61, 62, 63].

8.2 Summary of Results

The Macramé network has demonstrated that large IEEE 1355 DS-Link networks can be built, that they scale very well, and that these networks can cope with the high data rates required for the second level trigger of the ATLAS experiment. Furthermore, it has proven that the per-link flow control, together with well designed hardware, can result in very reliable systems. The most important results and conclusions from this work are listed below:

- Clos topologies allow large and scalable switches to be constructed. The performance of the other network topologies studied decreases rapidly for large networks.
- High-valency switches allow large Clos networks to be implemented efficiently. Using the 32-way STC104 crossbar switch, the 512-node Clos network which was studied only required 48 switches. If the basic switch only had 8 links, nearly ten times as many switches (448) would have been required to construct the same size network.
- The low-level flow control prevents packet loss. In conjunction with the intrinsically low error rate of the serial link technology, the switching fabric can therefore be considered to be loss-less, which allows the use of a simple transfer protocol.
- The throughput of a crossbar switch with input buffering under uniform random traffic is restricted to about 60% of the bandwidth due to head-of-line blocking. A 512-node Clos network still achieves 50% of the maximum throughput under the same traffic conditions.
- The performance obtained under traffic as expected in the ATLAS second level trigger system, which is characterised by a fan-in pattern, is significantly better than for random traffic. About 80% utilisation on the destination links was achieved.
- Grouped adaptive routing, which allows bundles of parallel links to behave like a single high-bandwidth connection and distributes the load evenly across the centre stage in Clos networks, can significantly improve the network performance compared to deterministic routing. The performance improvement was particularly marked for the ATLAS second level trigger traffic patterns.

- If latency variation is a concern, then the applied network load should be kept well below the saturation throughput, since the width of the latency distribution or jitter increases with the load.
- Under low network loads, the wormhole routed switching networks studied here can provide very low latency communication. The latency in this operating region is essentially dominated by the packet transmission time plus the switching time, which is only about 1µs per switch for the STC104.
- The performance for a Clos under random traffic can be improved by providing overcapacity in the central switching stage, although this does not use the additional centre stage bandwidth efficiently.
- For the second level trigger traffic it is possible in some cases to reduce the centre stage connectivity of a Clos network, without significantly affecting the overall performance, thereby decreasing the implementation cost.
- For the HEP traffic, the distribution of sources and sinks across the network significantly improves the latency and throughput of the network and should be considered when designing the system.
- If possible, the packet size should be matched to the size of the input and output buffers in the crossbar switches. A performance increase of about 15% was observed under random traffic for 32 byte packets compared to 1024 byte packets.
- The measurements presented give an upper limit of the network performance obtainable with this technology. The performance will be reduced further if the network interfaces are unable to handle the low packet overheads required. Achieving low communication overheads in the network interface becomes even more important for higher speed interconnects, such as the emerging Gigabit Ethernet.

8.3 Outlook

Unfortunately, the DS-Link and switch technology has not had the commercial success that it deserved, and the primary semiconductor vendor has recently stopped production of the supporting chips, although other companies still manufacture DS-Link based devices [64]. However, an association has been formed to obtain the IPRs for the relevant technology and to further promote the IEEE 1355 standard [65].

There are also some niche applications, such as space systems, in which the DS-Link technology is used. The work carried out within the Macramé project resulted in the participation in a project for ESA[1], where CERN provides expertise and hardware for a demonstrator of a multiprocessor system linked by DS-Links and switches to be used in satellites.

Largely motivated through the results from the Macramé testbed, which demonstrated successfully that currently available technology using serial 100 MBit/s links can meet the requirements of the ATLAS second level trigger system, a new study has been started to evaluate the use of Fast and Gigabit Ethernet for the same application [6].

1. European Space Agency

The Arches[2] project [66], a follow-up project of Macramé, aims at exploiting the 1 GBaud IEEE 1355 HS-Link technology. Within this project, another network test bed is being constructed at CERN. The systems consists of 64 end-nodes which will be connected through a switching fabric based on 8-way HS-Link crossbar switches [67]. The architecture and design of this new testbed relies heavily on the work presented in this thesis.

The validation of the technology presented here has recently also prompted its commercial application in LAN switching, where IEEE 1355 links and switches will be used for the internal switching fabric of Fast and Gigabit Ethernet switches.

Even though IEEE 1355 may not be the technology to be used at the LHC, the results of this study are still highly relevant to future trigger and data acquisition systems based on point-to-point links and switching networks, since this type of interconnect will enable scalable switching networks for LAN switches to be built.

2. Esprit project 20693: Application, Refinement and Consolidation of HIC, Exploiting Standards

References

[1] “The ATLAS Technical Proposal”, CERN/LHCC/94-43, LHCC/P2, December 1994, ISBN:92-9083-067-0.
`http://www.cern.ch/pub/Atlas/TP/tp.html`

[2] “The CMS Technical Proposal”, CERN/LHCC/94-38, LHCC/P1, December 1994, ISBN:92-9083-068-9.

[3] M.de Prycker. *“Asynchronous Transfer Mode”*. Ellis Horwood Ltd., 1991, ISBN: 0-13-053513-3.

[4] IEEE, IEEE Standard 1596-1992, “Scalable Coherent Interconnect (SCI)”.
`http://www.scizzl.com`

[5] ANSI. ANSI X3T11 FibreChannel.
`http://www.ansi.org`

[6] R.W. Dobinson, S. Haas, B. Martin, M. dobson, J.A. Strong. “Ethernet for the ATLAS Second Level Trigger?”. *Proceedings of SysComms’98*, 25-26 March ‘98, CERN, Geneva.

[7] “IEEE Std. 1355-1995, Standard for Heterogeneous InterConnect (HIC)”, IEEE June 1996.

[8] Minghua Zhu. “The Application of IEEE 1355 Link and Switch Architectures in HEP Data Acquisition and Triggering Systems”. PhD thesis, University of Liverpool, 1997.

[9] A.M. Jones, N.J. Davies, M.A. Firth, C.J. Wright. *The Network Designers Handbook*. chapter 8. IOS Press, 1997. ISBN 90 5199 380 3.

[10] L.B. Quinn, R.G. Russell. *“Fast Ethernet”*. Wiley, 1997, ISBN: 0-471-16998-6.

[11] A.J. McAuley. “Four State Asynchronous Architectures”. *IEEE Transactions on Computers*, 41, No. 2, 1992, pp. 129-142.

[12] IEEE Standard 1394 Standard for a High Performance Serial Bus”, IEEE Inc. 1995.

[13] D. Culler, J.P. Singh, “Parallel Computer Architectures”, chapter 10, Morgan Kaufman 1998.

[14] Design guidelines for running SCI and ATM protocols over the IEEE 1355 (HIC) transport layer, OMI/Macramé deliverable 3.1.3, University of Oslo, February 1996

[15] M. D. May, P. W. Thompson, P. H. Welch, *Networks, Routers & Transputers: Function, Performance and Applications*, IOS Press, 1993, ISBN: 90 5199 129 0.

[16] The T9000 Transputer Hardware Reference Manual, INMOS 1993, Document number: 72 TRN 238 01

[17] “STC104 Asynchronous Packet Switch”, Data sheet, SGS-Thomson, 1995

[18] “STC101 Parallel DS-Link Adaptor”, Data sheet, SGS-Thomson, 1995

[19] G.J. Christen et al., “Scalable Multi-channel Communication Subsystem (SMCS)”, in *Advances in Information Technologies: The Business Challenge,* IOS Press, 1998.

[20] “CW-1355-C111”, Data sheet, 4Links, 1996.

[21] Bullit Data Sheet, Bull Serial Link Technology Strings, May 1995.

[22] R. Marbot, A. Cofler, J-C. Lebihan, and R. Nezamzadeh, “Integration of Multiple Bidirectional Point-to-Point Serial Links in the Gigabits per Second Range”, *Proceedings of*

the Hot Interconnects I Symposium, 1993.

[23] RCube Specification, Laboratoire MASI CAO-VLSI UPMC Paris VI, February 1997

[24] B. Zerrouk, V. Reibaldi, F. Potter, A. Greiner, and A. Derieux. "RCube: A Gigabit Serial Links Low Latency Adaptive Router", *Proceedings of the Hot Interconnects IV Symposium*, pages 13-18, 1996.

[25] NOE chip. MPC Project, UPMC Paris.
`http://eowyn.lip6.fr/noe.html`

[26] P. Thompson and J. Lewis. "The STC104 Packet Routing Chip". *Journal of VLSI Design*, vol. 2, no. 4, pp. 305-314, 1994.

[27] The 41 Series of High-Performance Line Drivers, Receivers and Transceivers, AT&T Microelectronics, January 1991.

[28] har-link I/O connector system, Harting.
`http://www.harting.pader.net/index_english.html`

[29] Roger Heeley. "Real Time HEP Applications using T9000 Transputers, Links and Switches", PhD thesis. Liverpool University, 1996.

[30] International Standard IEC 801-4, "Electromagnetic compatibility for industrial-process measurement and control equipment, part 4, Electrical fast transient/burst requirements", CEI Geneva, 1988.

[31] P.A. Chatterton, M.A. Houlden, "EMC, Electromagnetic Theory to Practical Design", Wiley 1992.

[32] Macramé Working Paper 43, "The Study of Noise on DS Links", CERN, 1997.

[33] P. Bylanski, D.G.W. Ingram. *Digital Transmission Systems.* Peregrinus Ltd., Stevenage 1976.

[34] A. X. Widmer, P. A. Franaszek. "A DC-Balanced, Partitioned-Block, 8B/10B Transmission Code". *IBM J. Res. Develop.* September 1983

[35] S5933 PCI Controller Data Book, Applied Micro Circuits Corp. (AMCC), 1996.
`http://www.amcc.com/Products/PCI/S5933.htm`

[36] Altera data book, "FLEX 10K Programmmable Logic Family Data Sheet", Altera Corp., 1996.
`http://www.altera.com/`

[37] CY7B951 data sheet, "Local Area Network ATM Transceiver", Cypress Semiconductor Corp., 1995.
`http://www.cypress.com/cypress/prodgate/datacom/cy7b951.html`

[38] HFBR520x series data sheet, "ATM Multimode Fiber Transceivers for SONET OC-3/SDH STM-1 in Low Cost 1x9 Package Style, Technical Data", Hewlett Packard, 1995.
`http://www.hp.com/HP-COMP/fiber/`

[39] S. Haas, R.W. Dobinson, B. Martin, "Electrical and optical transmission of IEEE 1355 DS-links", Microprocessors and Microsystems 21 (1998) 429-439, Elsevier 1998.

[40] C. Clos. "A Study of Non-blocking Switching Networks". *Bell Systems Technical Journal*, vol. 32, 1953.

[41] J. van Leeuwen and R.B. Tan. "Interval Routing". *The Computer Journal*, vol. 30, no. 4, pp. 298-307, 1987.

[42] M.A. Firth, A. Jones. "Deadlock-free interval labelling". *Microprocessors and Microsystems*, vol. 21 No. 7-8, March 1998. Elsevier.

[43] A. Klein, "Interconnection Networks for Universal Message-Passing Systems", *Proc. ESPRIT Conference '91, pp. 336-351, Commission for the European Communities, Nov. 1991, ISBN 92-826-2905-8.*

[44] P.W. Thompson, "Globally Connected Fault-Tolerant Systems" in J. Kerridge (ed.), *Transputer and occam Research: New Directions, IOS Press, 1993.*

[45] L.G. Valiant. “A scheme for fast parallel communication”. *SIAM J. of Computing*, 11, pp. 350–361, 1982.

[46] M D May, P W Thompson and P H Welch. *Networks, Routers and Transputers: Function, Performance and Applications*, chapter 1. IOS Press 1993 ISBN 90 5199 129 0.

[47] R.W. Dobinson, B. Martin, S. Haas, R. Heeley, M. Zhu, J. Renner Hansen, “Realization of a 1000-node high-speed packet switching network”, *ICS-NET ’95 St Petersburg, Russia.*
`http://www.cern.ch/HSI/dshs/`

[48] “The Transputer Data Book”, 2nd ed., SGS-Thomson Microelectronics, 1989.

[49] D.A. Thornley, “A Test Bed for Evaluating the Performance of Very Large IEEE 1355 Networks”, PhD Thesis, University of Kent, 1998.

[50] “Netprobe Test/diagnostic software for IEEE 1355 DS link networks”, Arches project deliverable, CERN 1997.
`http://www.cern.ch/HSI/dshs/netprobe/netprobe.html`

[51] A.M. Jones, N.J. Davies, M.A. Firth, C.J. Wright. *The Network Designers Handbook*, Chapter 3, p. 35. IOS Press, 1997. ISBN 90 5199 380 3.

[52] J.Bystricky, R.W.Dobinson, S.Haas, D.Hubbard, B.Thooris. “Emulation of Architecture C on Macrame”. Atlas Internal Note; DAQ-No-107, June 1998.

[53] R. Bock and P. LeDu. “Detector and readout specifications, and buffer-RoI relations, for the level-2 trigger demonstrator program”. Atlas Internal Note; DAQ-No-062, Jan 27 1997.

[54] J.R. Hubbard S. George and J.C Vermuelen. “Input parameters for modelling the Atlas second level trigger”. Atlas Internal Note; DAQ-No-070, June 12 1997.

[55] S. Haas, D. A. Thornley, M. Zhu, R. W. Dobinson, R. Heeley, N.A.H. Madsen, B. Martin, “Results from the Macramé 1024 Node Switching Network”, *Computer Physics Communications*, no. 110 (1998) 206-210. Elsevier 1998.
`http://www.cern.ch/HSI/dshs/`

[56] R.W. Dobinson, S. Haas, R. Heeley, N.A.H. Madsen, B. Martin, J.A. Strong, D.A. Thornley, M. Zhu. “Emulation of the Level-2 trigger, architecture B, on the Macrame Testbed”, ATLAS Internal Note; DAQ-No-102 June 1998.

[57] R.W. Dobinson, S. Haas, R. Heeley, N.A.H. Madsen, B. Martin, J.A. Strong, D.A. Thornley. “Evaluation of network performance for triggering using a large switch”. *Proceedings of the International Conference on Computing in High Energy Physics*, CHEP’98. 1998.

[58] R.W. Dobinson, S. Haas, B. Martin, D.A. Thornley, M. Zhu, “The Macramé 1024 Node Switching Network: Status and Future Prospects”, *Proceedings of the 2nd International Data Acquisition Workshop*, (DAQ’96), Osaka, Japan, November 1996.
`http://www.cern.ch/HSI/dshs/`

[59] S. Haas, D.A. Thornley, M. Zhu1, R.W. Dobinson, R. Heeley, N.A.H. Madsen, B. Martin, “The Macramé 1024 Node Switching Network”, In B.Hertzberger & P. Sloot (Eds.), *High-Performance Computing and Networking, Lecture Notes in Computer Science*, Springer 1997.
`http://www.cern.ch/HSI/dshs/`

[60] S. Haas, D.A. Thornley, M. Zhu, R.W. Dobinson, B. Martin, “The Macramé 1024 Node Switching Network”, In A. Bakkers (Ed.), *Parallel Programming and Java*. IOS Press 1997.
`http://www.cern.ch/HSI/dshs/`

[61] S. Haas, D.A. Thornley, M. Zhu, R.W. Dobinson, B. Martin, “The Macramé 1024 Node Switching Network”. *Microprocessors and Microsystems*, no. 21 (1998) 511-518, Else-

vier 1998.

[62] Realisation and Performance of IEEE 1355 DS and HS Link Based, High Speed, Low Latency Packet Switching Networks", IEEE Transactions on Nuclear Science, vol. 45, no. 4, Aug. 1998, pp. 1849-1853. IEEE 1998.

[63] S. Haas, D.A. Thornley, M. Zhu, R.W. Dobinson, R. Heeley, B. Martin. "Results from the Macramé 1024 Node IEEE 1355 Switching Network". In J.-Y. Roger et al., *Advances in Information Technologies*, pp. 891-898. IOS Press, 1998.

[64] SMCS 332: Scalable Multi-channel Communication Subsystem, Dornier Satellitensysteme GmbH.
`http://www.omimo.be/companies/dasa_000.htm`

[65] The 1355 Association.
`http://www.1355-association.org/index-real.html`

[66] The Esprit project ARCHES, "Application, Refinement and Consolidation of HIC exploiting standards", Esprit P20693.
`http://www.omimo.be/projects/20693/20693.htm`

[67] C.R.Anderson, M.Boosten, R.W.Dobinson, S.Haas, R.Heeley, N.A.H.Madsen, B.Martin, J.Pech, D.A.Thornley, C.L.Ullod. "IEEE 1355 HS-Links: Present Status and Future Prospects". In *Architectures, Languages and Patterns for Parallel and Distributed Applications*, P.H. Welch, A.W.P. Bakkers (Eds.), pp. 69-79. IOS Press, 1998.