Arches

Application, Refinement and Consolidation of HIC Exploiting Standards

ESPRIT Project 20693

Deliverable D.3.2.1

Preliminary Report on Large HS Link Testbed

CERN

Classification: Public
ISSUE DATE: March 1998

D 3.2.1 Preliminary Report on large HS link Testbed

Overview

It took some time to reach the planned level of effort on this task. By mid ’97 however this level was reached and even exceeded in order to keep to the promised delivery delays.

We believe that the plan or work is now well on course and foresee no problems for the future.

Work on testing the HS link components is well advanced and a number of small problems have been identified and reported back to the design and fabrication personnel concerned.

At this point in time however there have still not been any systematic studies of even small configurations of HS components running at the rated speeds or under saturation conditions.

Given the increasing pressure of time to market combined with the reticence of third party companies to commit to untested technology it becomes increasingly urgent for this testbed to deliver the evidence that the designs indeed work as planned, or, if not, to what limits they can reliably be driven.

We expect to reach the first stage where one Rcube switch will be driven to saturation to be achieved early in Q2 ’98. We expect to extend this to 64 nodes before the end of Q3 ’98.

We also expect that the newly agreed work on carrying Gigabit Ethernet frames through HS link technology will be able to exploit this switching fabric.

D 3.2.1 Preliminary Report on large HS link Testbed

Executive Overview

It took some time to reach the planned level of effort on this task. By mid ’97 however this level was reached and even exceeded in order to keep to the promised delivery delays.
We believe that the plan or work is now well on course and foresee no problems for the future.
Work on testing the HS link components is well advanced and a number of small problems have been identified and reported back to the design and fabrication personnel concerned.
At this point in time however there have still not been any systematic studies of even small configurations of HS components running at the rated speeds or under saturation conditions.
Given the increasing pressure of time to market combined with the reticence of third party companies to commit to untested technology it becomes increasingly urgent for this testbed to deliver the evidence that the designs indeed work as planned, or, if not, to what limits they can reliably be driven.
We expect to reach the first stage where one Rcube switch will be driven to saturation to be achieved early in Q2 ’98. We expect to extend this to 64 nodes before the end of Q3 ’98.
We also expect that the newly agreed work on carrying Gigabit Ethernet frames through HS link technology will be able to exploit this switching fabric.

D 3.2.1 Preliminary Report on large HS link Testbed

Executive Overview 1
1 Introduction 3
1.1 Market opportunities 3
1.2 Initial conditions 3
1.3 Task Goals 3
2 Network Architecture Design 4
3 System Design and Packaging 6
3.1 General Packaging Issues 6
3.1.1 PCB Tracks 6
3.1.2 Connectors 6
3.2 The 32 way Switch 7
3.2.1 The Switch array 7
3.2.1 The board design 8
3.2.2 Implementation 9
3.3 The 8 Way Switch 10
3.3.1 The Switch Array 10
3.3.2 The Board Design 10
3.3.3 Implementation 10
4 Terminal Node 11
4.1 Requirements 11
4.2 Packaging choices 11
4.3 Design 12
4.4 Implementation 13
4.5 Firmware 13
5 Software 13
6 Testbed Configuration 14
7 Status 15
8 Summary 15

1 Introduction

1.1 Market opportunities.

The market for HS link technology and the devices that employ it is perceived in three main areas:

- macrocell technology for embedding in third party silicon
- commodity interconnect through standard I/O busses
- embedded switching fabric carrying third party protocols

To achieve market take-up it is necessary to refine the available HS products by identifying and removing any low level problems at the chip or system level and to demonstrate that this has been achieved by successfully running a system comparable to the market requirements.

1.2 Initial conditions

As the Arches project began, the first HS link device, the BULLIT parallel to serial driver, had been in exploitation for a number of months and the RCUBE switch had yet to be delivered.

Exploitation of the Bullit had not been without difficulty and there were a number of pending questions about the set of conditions under which reliable operation could be sustained.

Work in this area had been undertaken in the closing months of the Macrame project with simple tests being done first between Bullit chips and then extended to include the Rcube once it had been delivered.

These tests showed up a number of small problems but none so severe that the goal of achieving a credible demonstration of the technology would be compromised.

One aspect of the work involves the study of the transmission of HS links across printed circuits and cables. Some work on this had been done at BULL and some work completed under MACRAME. These studies were continued in Arches within this task and will be reported on in a separate working paper.

1.3 Task Goals

The main thrust of this task is to develop and test a switching fabric that can be used for networks up to and in excess of 128 nodes. This represents the marketable dimension of switch capacity envisaged in the market for ATM, Fast and Gigabit ethernet, RAID arrays or networks of workstations.

The network system comprises both the switching fabric itself and the terminal nodes that will drive the network at its rated capacity.

This is shown schematically in Fig1. in which an array of data sources and sinks, known as terminals, communicate with each other across a switch fabric built from Rcube components. To obtain data on maximum system throughput and reliability the terminal nodes may drive their links at full speed and generate any number of worst case or best case traffic scenarios. To characterise the network for any given application the terminal nodes can also be programmed to generate the traffic which would be experienced for the application as specified by the user or customer.

The entire system is under computer control that initialises and monitors the system. The object of the study is to define and build such a network system, bring it to a state

of high reliability and performance and then characterise that performance for the traffic of choice.

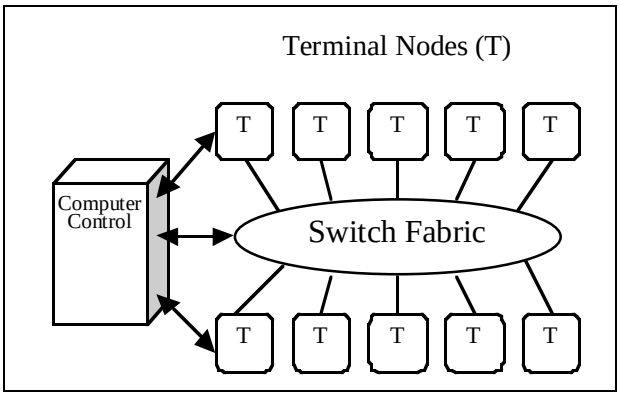

Fig.1. Testbed structure

2 Network Architecture Design

Building on the work done in Macrame it was decided to favour the development of a switch fabric that exploited the dynamic routing and adaptability of the technology. While physically difficult to construct, the Clos network offers the best performance in terms of throughput and latency and the architecture is based on this topology. Nonetheless it is also desirable for smaller networks that the simpler grid, cube or hypercube topologies can be demonstrated so this need also has to be taken into account.

The RCUBE has a valency of only eight that results in a potentially high chip count for larger networks. This could result in either an unacceptably high number of cables or overly complex tracks on large printed circuit boards.

Simple observation is sufficient to see that the eight valent switch must be packaged into a higher valency ‘building-block’ if we are to avoid an exorbitant number of cables. The first packing option is a 16 way switch using four switches at the first level and two at the second as seen in Fig 2.

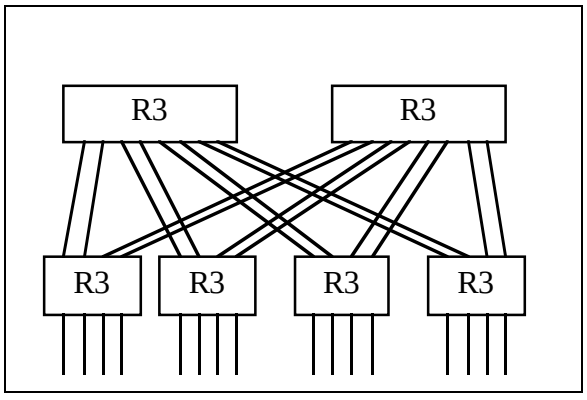

Fig.2. Sixteen valent three stage Clos

Building a larger network with this device is shown in Fig 3. Eight center stage switches (CS1: 8) and sixteen terminal stage switches (TS (1:16) are needed, a total of 24 modules or 24*6= 144 Rcube devices. This design has the advantage of a single

module construction but the chip count is high and the support for non-Clos networks is limited. In addition each module is in itself a three stage Clos. That means that the worst case path across this network is now nine stages.

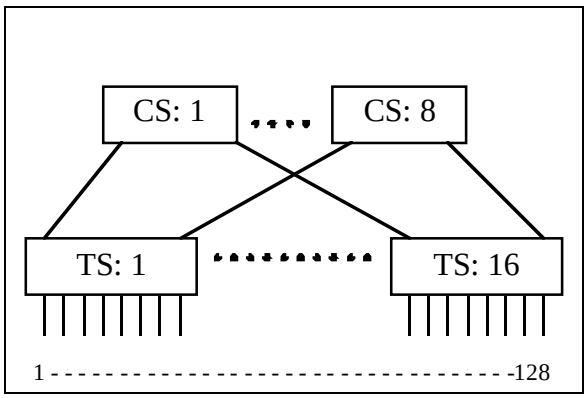

Fig 3. 128 Node switch based on 16 valent module

Adding to these problems was the issue of Rcube availability. A solution using less Rcubes was required in case the chip supply was too limited.
A topology was then studied in which the terminal stage was built from 32 single Rcube devices each having four links to terminals and four to the central switch fabric. This central fabric could then be considered as four planes of a 32 way switch as shown in Fig 4. A 32 way switch is just twice the chip count of the sixteen way switch. There are eight switches in the terminal layer and four in the center stage. It is still a three stage switch. The total chip count then is (4*12)+32=80 Rcube switches and the worst case transit across the fabric is now only five stages.
This solution is attractive, not only because of the lower chip count but also because the use of single Rcube devices opens up the possibilities of building smaller non-Clos networks. Furthermore the development of a 32 way switch module opens up the possibility of reaching Clos topologies in excess of 128 nodes.
This topology and partitioning of functionality was thus was chosen as the network design for this project.

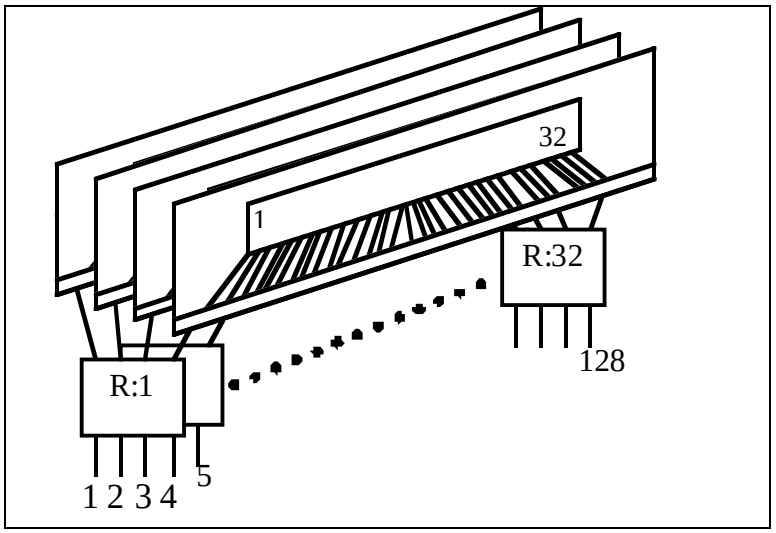

Fig 4. 128 Node switch based on 32 valent modules

3 System Design and Packaging

3.1 General Packaging Issues

Before committing to the design of multiple printed circuit boards supporting multiple components communicating at gigabaud rates a number of studies had been undertaken to asses the risks involved.

3.1.1 PCB Tracks

The work carried out on HS link transmission was still ongoing during the design phase but preliminary results showed that while prudence was required there was no need to be paranoid. In practical terms, that means that impedance controlled tracking on PCB's was obligatory but that the normal manufacturing tolerances of ±10% was acceptable. There was no apparent need to impose exotic laminates to achieve superior tolerances.

The use of split power planes however is almost inevitable given the mixture of 3 volt and 5 volt devices. Care must be taken that no impedance-controlled tracks that reference such planes ever cross the split in the plane.

Studies on via's showed minimal effect on HS link signal profiles and are not therefore to be systematically avoided. Normal codes of practice can be applied.

3.1.2 Connectors

The choice of connector is perhaps more sensitive. Reports from Parsytec have indicated that under difficult environmental conditions in which errors were occurring, the choice of connector made a difference to the error rate. That does not necessarily mean that a connector is a source of problem, it might only indicate a second order effect once the operating envelope had already been exceeded. Nonetheless it poses a design issue. HS links have a standard proposed connector for which not even a prototype was available as the design cycle started. It offers the advantage of a slim (6mm) footprint with both the serial-in and serial–out lines housed in a single 'Siamese' co-axial cable. There is no reason to believe that the design or the implementation of the connector is in any way unsuitable for the job it is supposed to do but in the absence of proof it involves a degree of risk.

The worst case to plan for is in fact one of the major requirements of the test bench, that is, to identify any low-level problems that arise during system operation. Suppose that under normal operating conditions there occurs some rare, random bit error. How does one distinguish if this is due to the cable, the connector, the track or the device? Given that the desire is to award the device a 'clean bill of health' the only way to achieve that is to have the highest standard of assured quality for all the intermediate connectivity.

For this reason the choice was made to use industry standard 50 ohm high frequency connectors. The appropriate series are available in screw fitting (SMA) or push-fit (SMB). We chose the push fit because these connectors have a much larger footprint than the HS link connector does and the screw fitting just make it worse. The added advantage is that these are the same connectors as used by Parsytec (and for the same reasons) which makes performance comparison easier to achieve.

3.2 The 32 way Switch

3.2.1 The Switch array

The decision to design in a 32 way switch raised two main problems.
The first is how to bring 64 HS link coaxial connectors to a front panel without some complex PCB to front panel cable harness yet at the same time with a spacing sufficient to actually insert and extract the cables. This was solved by adopting the form factor of the 19" rack enclosure that has almost enough front panel space to accept 64 standard SMA style connectors. The shortfall is made up by alternating the connectors between the component-side and solder-side of the board. This trick allows the center to center distance to be violated and at the same time gives enough room to insert and extract the cables.
The second problem is how to cope with the high level of HS link interconnects on the PCB. There are 32 internal links to be routed which means 64 impedance controlled tracks that by definition have to cross over each other since each center stage switch must connect to every terminal stage switch.
The solution to this is not without risk. A board stack was developed as shown in Fig. 5 in which five dimensionally controlled wiring layers have been provided. The top layer for microstrip lines and L3, 4,6 and 7 are for striplines. Providing enough reference planes to allow this, plus the requirement for both 3 volt and 5 volt planes results in a 10 layer board.
The top layer can only really be used for the short runs to the connectors. The chip to chip connections, which are much longer and complex are routed in the four inner layers.
The risk is that in spite of all the studies on PCB layout effects, the long lengths (~35 cms) involved and the potential for cross talk between adjacent HS lines or between HS and TTL lines on the same layer somehow crosses the threshold for secure operation.
This is an acceptable level of risk in this pre-market research however. It does not seem likely that routing complexity in excess of this would be probable in a commercial product. If the board functions, then future products can be specified with this in mind. If it fails then the limits to complexity are also understood.

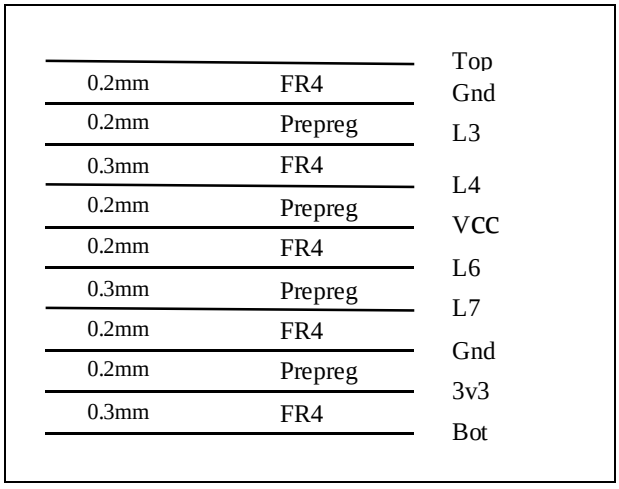

Fig 5. 32 way switch PCB stack

Fig 6 shows the five controlled impedance layers superimposed on each other. Only the HS links are shown for clarity.

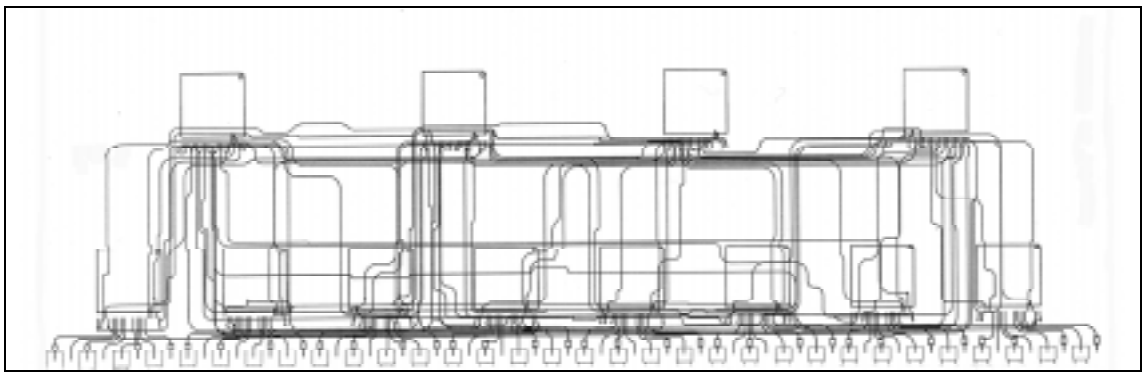

Fig 6. HS link PCB interconnections

The square features are the Rcube footprints. The structures at the bottom of the plot are the connectors on the component side. The gap between each visible one is occupied by a connector on the solder side.
Note that the tracks have all been rounded to reduce to a minimum the effect of reflections at points of inflection.

3.2.1 The board design

Having defined the switch array this then requires a control and monitoring structure. Consideration had to be given to the control network that is used to initialise and monitor devices that are physically separated on different boards. Much work had been done on this subject under the Macrame project using DS links as the serial technology of choice. Using DS links and the associated STC104 switch chip means that the control network could be built in the easily managed star fan-out topology. It also offered the major advantage of a stable hardware and software system, which could be invaluable, when the inevitable system debugging phase caused problems. On the other hand the component supply was unstable and for future marketable products it was unlikely to be an option. It was decided that the security of a known test network overrides any commercial concerns (which were not a part of this task) and steps were taken to ensure the necessary chip supplies. As an extra insurance we also included in the design a simple RS232 chain that could be used by any party willing to write the drivers for it. Here, at least, chip technology was no longer a problem.
The block diagram in Fig 7 shows the two DS link drivers and the backup RS232 links both of which could be employed in either a star or a serial control link topology. The T8 transputer initialises and monitors the Clos network of 12 Rcube switches. It also runs the control link software that handles the message passing on the control links.

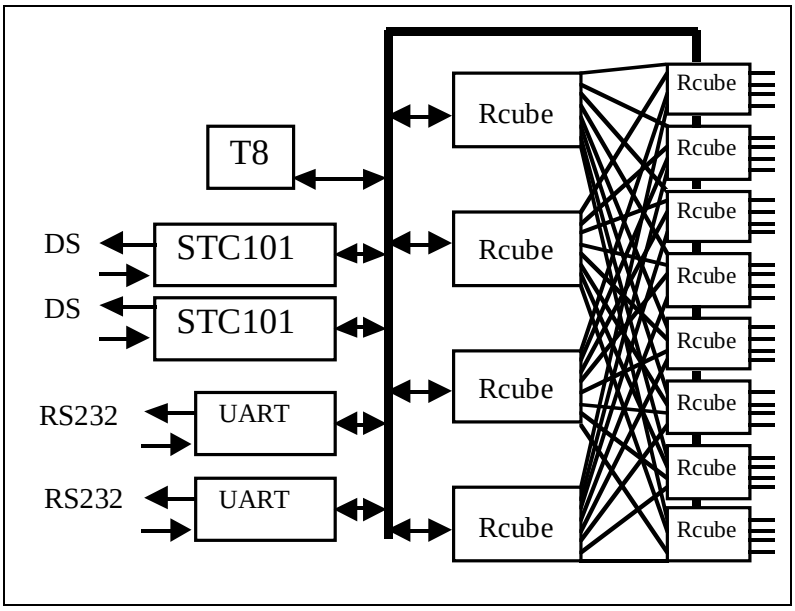

Fig 7. 32 way switch block diagram

3.2.2 Implementation

The layout of the board is shown in Fig.8.

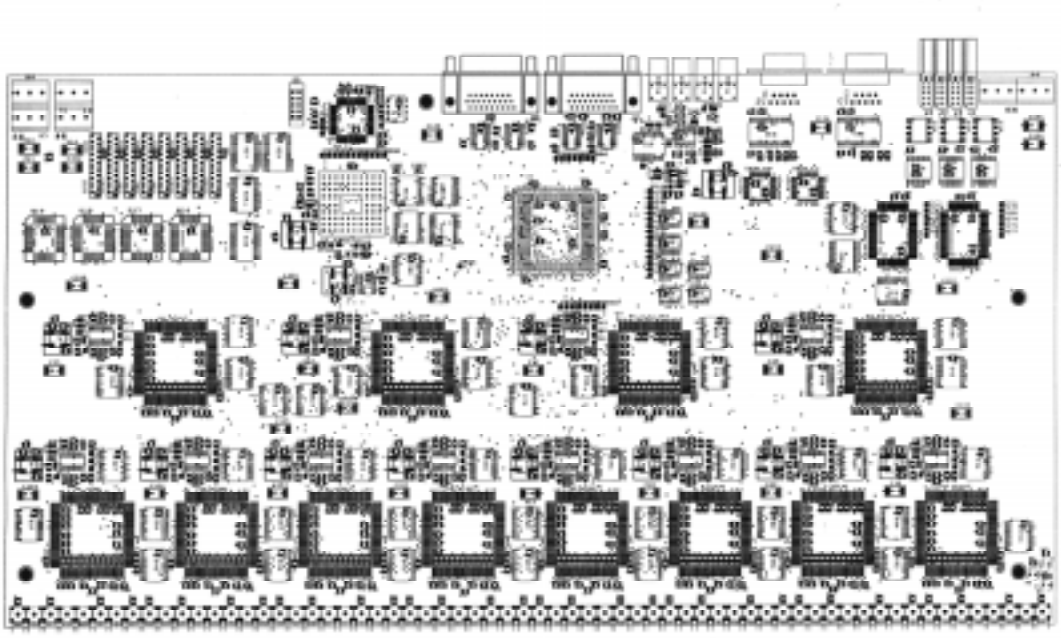

Fig 8. 32 way switch printed circuit layout

The array of eight terminal stage Rcubes and the four center stage Rcubes can clearly be seen. The controlling processor and its memory occupy the upper left corner while the control link I.O options occupy the upper right corner.

3.3 The 8 Way Switch

3.3.1 The Switch Array

Having determined the housing for the complex case, the 8 way switch housing was chosen to be the same as for the 32 way switch. Since the packing limit comes from the number of front panel connectors this sets the Rcube count per housing to four.
This board is much simpler to handle from the point of view of controlled impedance tracking since the only HS link tracks are short stubs that go from the Rcube pins directly to the front panel connectors.It was built in an eight layer PCB.

3.3.2 The Board Design

The previous control architecture has been simply repeated here. Visible in the block diagram in Fig.9 are the alternate DS links or RS232 links for use either in a daisy chain in a star control topology. All of the HS links of each Rcube are brought out directly to the front panel.

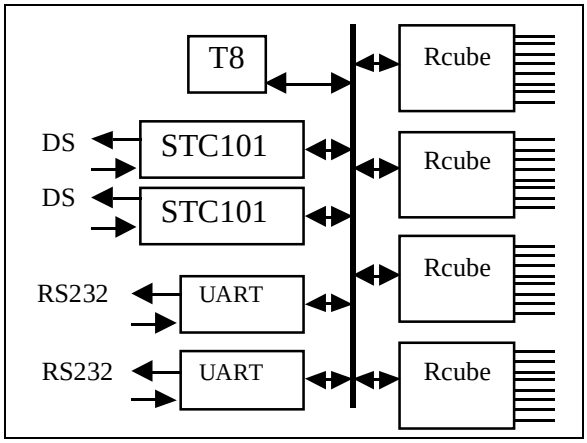

Fig 9. Four 8 way switch block diagram

3.3.3 Implementation

Fig 10 shows a photograph of the prototype board. The four Rcubes are clearly

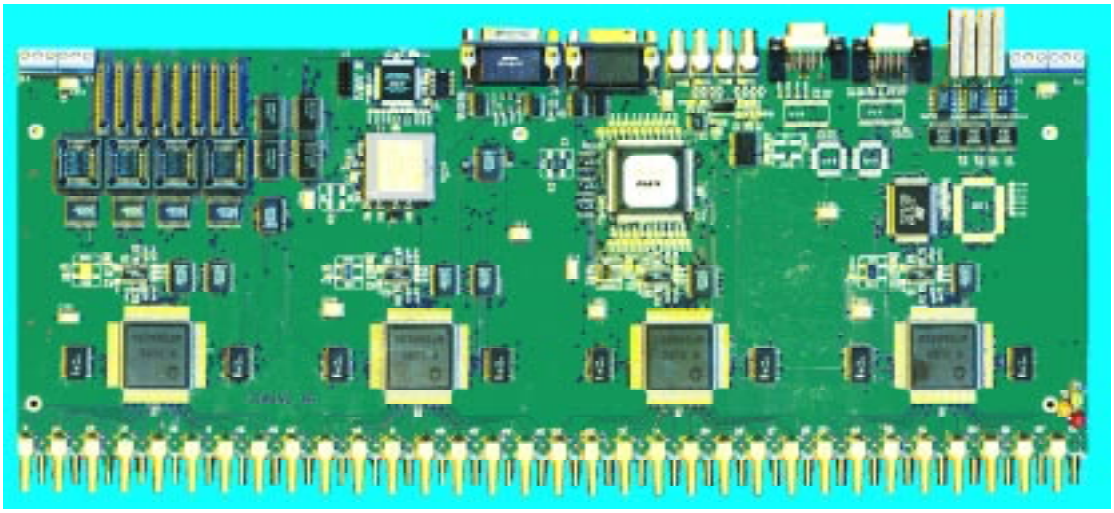

Fig 10. Four 8 way switch photograph

distinguishable and feed the array of 64 connectors along the lower edge of the board. The controlling processor and its memory occupy the upper left corner while the control link I.O options occupy the upper right corner.

4 Terminal Node

4.1 Requirements

A terminal test node had to be designed that would not only be capable of saturating the switch fabric with traffic patterns of any given choice but also be capable of running more complex transactions. In the Macrame project there were two types of terminal node developed to meet cost restrictions. Most of the terminals were very cheap and very simple traffic 'push' generators. A few terminal nodes were also equipped, at greater cost, with individual processors that were capable not only of the data 'push' functions but could also execute simple message passing protocols. The lesson learned in that study, is that both functions are required for all the terminal nodes once the focus is on the user applications. In addition, this Arches switch fabric will finally be built with only up to 64 nodes and the economy of scale in having two different terminal node design types is no longer attainable. One single design will have to suffice for the whole range of functionality.

4.2 Packaging choices

Designing in a processor per node is an expensive exercise in terms of board space, component cost and software development time. It is also highly redundant since the response time for the more complex protocols is not so critical. This leads us to the possibility of supporting two or more nodes from a single processor over a bus. VME bus is too slow, especially for single transactions whereas PCI bus has the needed bandwidth to support several devices. PCI bus is available in four main configurations:

- On a motherboard connected between chips
- 4 to 5 slots on a back-plane in a commodity PC
- 12 to 16 slots on a back-plane in an industrial PC
- 8 to 16 slots in CompactPCI

The first option requires the design and implementation of a user-specified processor which is both risky and costly. The industrial PC option is also costly and it is not clear that all slots could be effectively used since each link would only have one sixteenth of the bandwidth and response time available. The same argument applies to CompactPCI both in terms of cost and resources. This leaves the commodity PC solution.

This solution has several implications:

- the software work done in task 3.1 to support PCI devices under Linux can be re-used to support the HS link terminal nodes.
- commodity PC's come with a well understood and cheap interconnect system, ethernet.
- The number of slots could be seen to be a limitation unless more than one link can be housed per slot.

Housing more than one link per PCI board faces two restrictions. One is the space available on the rear panel and the other is the space available on the board itself. The

panel space sets an upper limit of two links due to the fact that there are less than 7 cms. useful space available. The board format standard allows for any variant from the short form PCI of ~(10*16) cms to the full long form of ~(10*30) cms. In practice however boards are restricted to under (10*18) cms. due to the encumbered motherboards that have heat-sinks, memory sockets or other devices that conflict with the needs of the long form PCI boards.

4.3 Design

Fitting two nodes per PCI board required careful selection of components and attention to detail in setting out the floor-plan. The block diagram of the terminal node is shown in Fig 11.

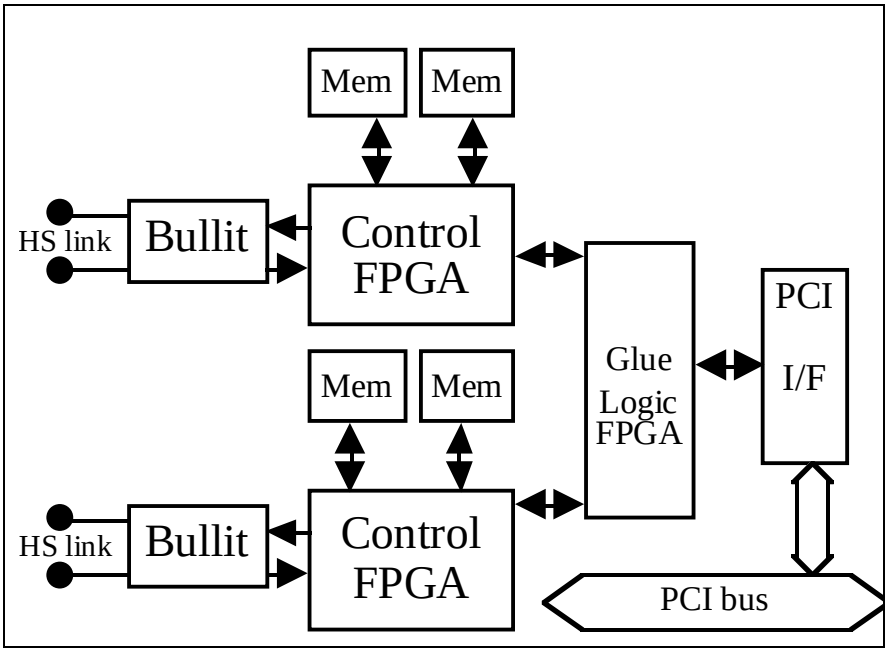

Fig 11. Block Diagram of the PCI HS-link Interface

A PCI interface chip provides for user DMA, mailboxes, interrupts and bus interfacing. Glue logic implemented in a FPGA handles the multiplexing of this interface between two HS-link channels. Each channel handles a HS-link Bullit interface. State machines to read or write the Bullit input and output FIFOs are implemented in the Control FPGA. The HS-link packets can be transported to or from the PCI bus directly thus fulfilling the processor to processor communication requirements.

Each Control FPGA has access to two memory banks: one for the transmitted data and one for the received data. To exercise the links at full bandwidth, data packet descriptors are stored in the transmit memory. These descriptors contain the packet destination address, its length and its required time of dispatch. These variables are sufficient to fully define the traffic profile that will traverse the switching network.

State machines in the Control FPGA interpret these descriptors and transmit the required packet as requested. Packets are time-stamped on reception, their delays calculated and the results stored in the reception memory that is regularly read out by a control processor across the PCI bus.

4.4 Implementation

Fig. 12 shows a photograph of the prototype board.

Fig. 12. Photograph of the PCI HS-link Interface

4.5 Firmware

Much of the functionality of the terminal node is implemented in firmware in both the Control FPGAs and the glue logic FPGA.

In addition to the two main functions of generating traffic and implementing a direct interface between HS packets and the PCI bus, there is also the test and debug functionality to be supported.

Firmware has been written in AHDL to access and verify the functionality of the Bullit chips, the memory chips and the intermediate connectivity.

A simple packet level interface engine has also been implemented. A Linux application has been written that uses this engine to send data out of one link and return it through the other.

The transfer engines to read packet descriptors in the memory and then both generate and consume the traffic so described, are under development.

5 Software

Software has been developed to allow the exploitation of the testbed. This software is divided into three areas:

- overall system control.
- node control.
- network control

The system controller is responsible for ensuring the correct start up and inter-operation of the nodes and network. It provides the node control software with the traffic profile information and provides the network control software with the network configuration. At any time the system controller can also request results from the node software.

The node software configures the PCI-HS boards, loads each link's traffic profile, starts the Bullit links and instructs the board to start transmission when a hardware trigger is received. When the network is running it checks to ensure the correct nodes are active and when requested reports results to the system controller. The results consist of transmit/receive rates and latency histograms. Access between the node software and the system controller is provided by TCP/IP over Ethernet.

The network control software is used by the system controller to access the Rcube switches. On receiving the relevant command it will reset, configure or access the status of the switches. The software uses a Network Description Language (NDL) to describe the configuration of the switches, which has been extended to support the Rcube device. This allows the network control software to support mixed HS and DS networks. The Rcube devices are accessed via a dedicated DS link control network using a PCI to DS link interface card.

6 Testbed Configuration

The current project foresees the construction of a 64 node testbed. Using the modules described above this will be achieved by using two 32 way switches as the central stage and 16 of the 8 way switches as the terminal stage as shown in Fig 13. Each 8 way switch has two links to each of the center stage modules and one link to a terminal node.

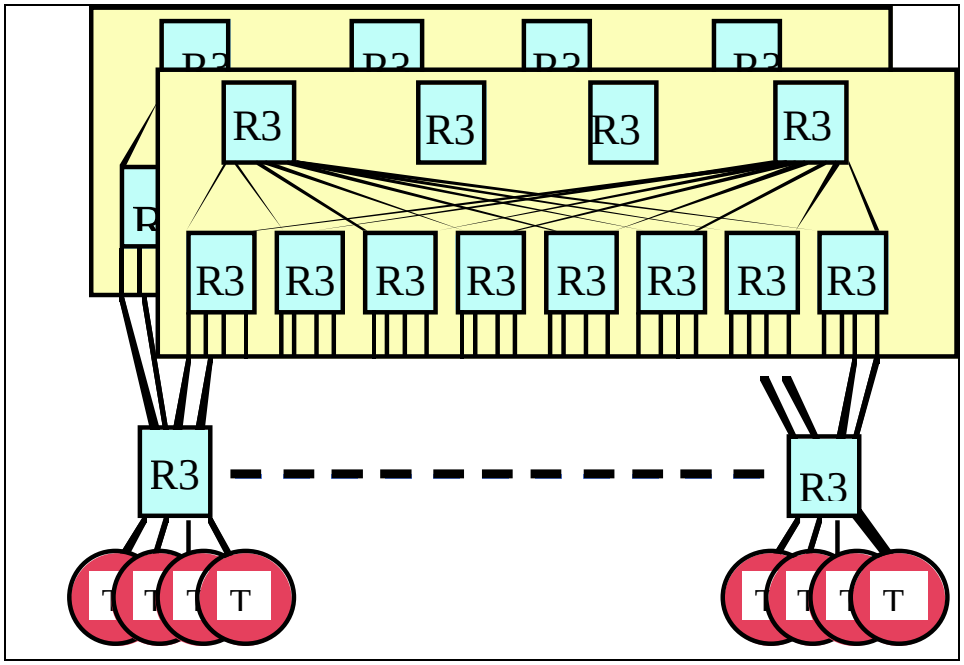

Fig. 13. 64 Node testbed structure

This results in the mechanical arrangement shown in Fig. 14. There are four 19” ‘pizza’ boxes each housing four 8 way switches and two 19” ‘pizza’ boxes housing the 32 way switch mounted above them. To the left are the eight PCI bus housings

each carrying four terminal node modules of two HS links each. All of the interconnect is done by co-axial cables.
For clarity the control structure is not shown. There will be a central SUN control computer that monitors and initialise the Terminal nodes via ethernet between the eight PCI housings. The switches are controlled from a DS link network comprising one PCI-DS interface driving a STC104 control link fan-out device. The 6 pizza boxes are then controlled via a star topology control chain.

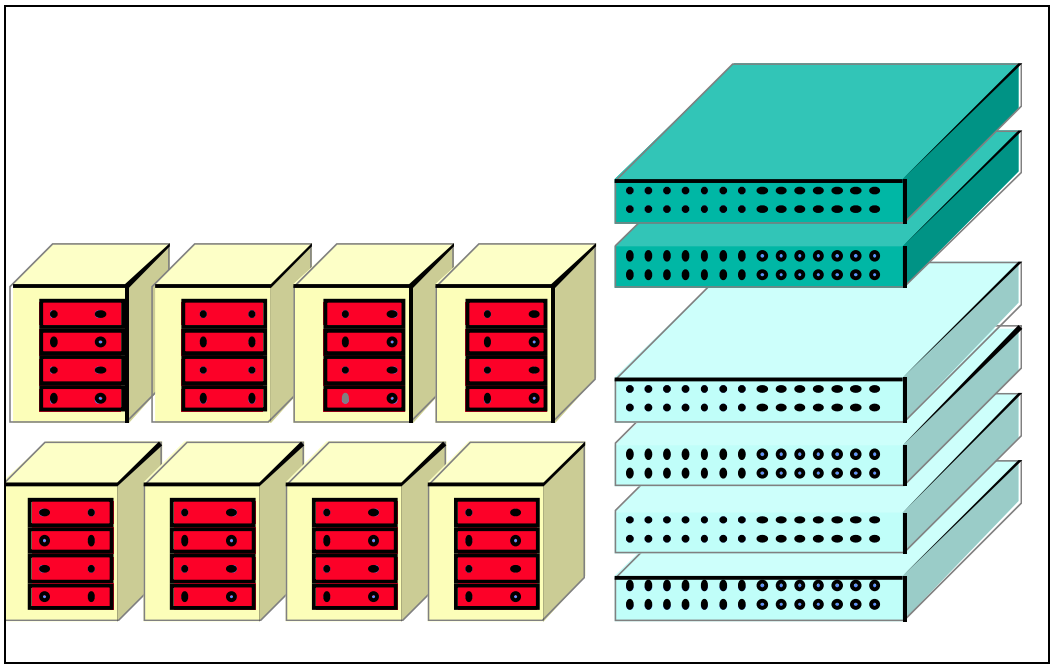

Fig. 14. 64 Node testbed mechanics

7 Status

The eight way switch prototype is designed, built and tested. The necessary modifications have been made to the PCB design that is now in production for the four required modules.
The 32 way switch has been designed and the prototype PCB layout and routing completed. The PCB prototype is being produced now.
The terminal node prototype is designed, built and tested. The necessary modifications have been made to the PCB design that is now in production for the thirty-two required modules..
The various pizza box mechanics, housings, front panels etc are in construction now.
The control software prototype is working and will be completely debuggd as production components become available.
The firmware prototypes function and development is on schedule.

8 Summary

The testbed has been fully defined and designed. Prototype work progresses well. The timescales agreed to for the project continuation can be met.