EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

This status report is preliminary[1] and will be revised for the next LCB meeting

CERN/LHCC LHCC 96-33
LCB Status Report/RD24
2 October 1996

RD24 Status Report

Application of the Scalable Coherent Interface to Data Acquisition at LHC

A. Bogaerts[2], Y. Ermoline[3], C. Fernandes[4], M. Liebhart, H. Müller[2], P. Werner
CERN, Geneva, Switzerland

B. Skaali, E.H.Kristiansen[5],H.L. Opheim, R.Nordstrom, D. Wormald, B. Wu
University of Oslo, Department of Physics, Norway

V. Lindenstruth
LBL, Berkeley, CA, USA

E. Sanchis-Peris, V. Gonzalez-Millan
University of Valencia, Department of Informatics and Electronics, Spain

A. Sebastia, J. Ferrer-Prieto, F. Mora
Polytechnical University of Valencia, Spain

J.M. Lopez-Amengual
IFIC, Valencia, Spain

F.J . Wickens,R.P.Middleton
Rutherford Appleton Laboratory, Didcot, UK

K. Løchsen, H. Kohmann
Dolphin Interconnect Solutions A.S., Oslo, Norway

A. Guglielmi
Digital Equipment Corporation (DEC), Joint Project at CERN

[1] incomplete milestone measurements due to delay in delivery of SCI equipment

[2] joinexperieet spokesmen

[3] now at University of Lausanne

[4] now at Thomson CSF, Paris

[5] SINTEF, Oslo, Norway

0.1 ACRONYMS

- ALICE LHC experiment at the LHC collider
- ASIC Application Specific Integrated Circuit, typically up to 500 Kgates
- ATLAS LHC experiement at the LHC collider
- ATM Asynchronous Transfer Mode, technology for local and wide area networks
- CMOS Complementary Metal Oxide on Silicon became dominent IC technology
- CRC Cyclyc Redundancy Code, SCI uses the 16 bit CCITT polynominal
- CSR Control and Status Register, SCi uses the IEEE Std 1212 conventions for CSR ,
- CPU Central Processing Unit, today in the sense of one processor chip
- DAQ Data Acquistion of High Energy Physics
- DMA Direct Memory Access engines which transfer data without CPU
- ECL Emitter Coupled Logic signals
- FASTBUS IEEE standard 960, modular data bus of High Energy Physics
- FPGA Field Programmable Gate Array, up to 40 Kgates reprogrammable by user
- HIPPI Circuit switching ANSI standard X3.183
- HIC Heterogeneous Interconnect Standard (serial) IEEE 1355-1995
- IEEE Institute of Electrical and Electronics Engineers
- ISO International Standards Organization
- LHC Large Hadron Collider project at CERN, Geneva
- LVDS Low Voltage Differential Signals standard IEEE P 1596.3
- NODE interconnect point of an SCI system with ingoing and outgoing links. SCI link bandwith is quoted as the bandwith of one of these links. Note that other systems quote a duplex bandwidth for the same configuration.
- NUMA Non Uniform Memory Access, a scalable shared memory architecture
- PCI Peripheral Component Interconnect, SIG consortium
- PCI-SCI Adapter between PCIbus and an SCI node
- PC server Typically 4 Pentium processors sharing a multiprocessor bus
- RD24 R&D project at CERN for investigation of SCI for the LHC projects, since 1992
- PECL Pseudo ECL signal, implemented via CMOS cells at 5Volt baseline shift
- PMC Mezzanine standard P 1383.1 for PCI bus modules in VME and Futurebus
- SBUS Mezzanine Standard for the Sparc Architecture by Sun Microsystems
- SCI Scalable Coherent Interface, IEEE/ANSI/IEC standard 1596-1992[1]
- SCSI Small Computer Systems Interface ANSI standard X3.131 for mass storage devices
- SHVS Intel architecture of Standard High Volume Servers (typically 4 Pentium Pro CPU's)
- SMP Shared memory Symmetric Multiprocessors
- VHDL VHSIC Hardware Description Language, sponsored by US Air Force, became IEEE-Standard 1076
- VME Versa Modules Europe, initially Motorola, then ANSI/IEEE 1014 and IEC 821 now VSO standard VITA-1
- VME-9U VITA 1.X , Draft 0.4 24 May, 1996, VME 9U*400 mm Format (Fastbus card size)
- VSO VITA Standards Organization

0.2 Overview

The RD24 work on SCI in 1996 was dominated by test and integration of PCI-SCI bridges for VMEbus and for PC's for the 1996 milestones. In spite of the splitting of RD24 membership into the ATLAS, ALICE and the proposed LHC-B experiments, collaboration and sharing of resources of SCI laboratories and equipment continues with excellent results. The availability of cheap PCI-SCI adapters has allowed construction of VME multicrate testbenches based on a variety of VME processors and workstations. Transparent memory-to-memory accesses between remote PCI buses over SCI have been established under the Linux, Lynx-OS and Windows-N operating systems as a proof that scalable multicrate systems are ready to be implemented with off-the-shelf products. Commercial SCI-PCI adapters

[1] For more details on the SCI standard, see SCIzzL Association's WEB page http://sunrise.scu.edu

are based on a PCI-SCI ASIC from Dolphin. The FPGA based PCI-SCI adapter, designed by CERN and LBL for Data acquisiton at LHC and STAR, which specifically allows addition of DAQ functions is now also in use also for SCI based readout systems for Fusion Experiments in Garching/Germany.

The step from multicrate systems towards a scalable DAQ system is equally becoming reality with the imminent availability of scalable SCI switch fabrics which were subject of MODSIM modelling during the last years. RD24 had 4-way switches available for tests in 1995 and is expecting to perform reality tests with a new rack-mountable 16-way switch fabric for the ATLAS 2nd level trigger tests

An important step in the SCI roadmap is the new generation 2 Watt CMOS chip (LC-2) from Dolphin with 500 Mbyte/s link performance and LVDS signaling. LVDS[1], a standard in the SCI family, has been adopted by Telecom and ASIC manufacturers like LSI Logic [12] to replace interconnections which were previously only possible in ECL technology. This innovation permits the same 1/2 Gbyte/s link performance as the initial 30 WATT GaAS SCI chip (which RD24 was using in 1993). Low cost, low power SCI link controllers with 8 bit wide LVDS links are equally prepared by ISS Interconnect Solutions in the USA.

Desktop SMP, a trend to interconnect multiple CPU boards like quad Pentium-Pro via SCI to maintain shared memory (NUMA architectures) is progressing in industry and Research Institutions[2]. Dolphin collaborates with Data General on their NUMALiiNE technology[3] which uses SCI as interconnect for Intel's 4-way P6/P7 motherboards via an SCI adapter which plugs directly into the Processorbus. Sequent Computers, who will soon start shipping cache-coherent CC-NUMA architectures with direct SCI node adapters[4] to the Pentium-Pro bus, report that access over SCI via a L3 cache are faster than accesses to local RAM. SUN Microsystem's Cluster Channel, a 100 Mbyte/s Intercommunication technology is based on Dolphin's SCI technology[5]. Sun's decision to support SCI demonstrates the growing acceptance of SCI as a commercial standard. Digital Equipment has started an Alpha-based CPU farm project where subfarms are planned to be interconnected via SCI and ATM adapters (see "ATLAS - Digital joint project." on page 25).

Important work is also going on the physical SCI links in order to standardize on solutions which are adapted to SCI's high speed requirements and which allow differnet SCI technologies to interoperate. A spinoff of RD24 is the ISO-IEC working group No 15 which was initiated by RD24 members in order to define a standard for harsh application environments (i.e. LHC underground area). Of particular interest are the new possibilities to replace copper cables by parallel fiber links. Both Motorola and Siemens[9][10], who have solutions for parallel fiber links which directly interface to SCI, joined the ISO-IEC working group.

Another spinoff of the RD24 PCI-SCI bridge design is the design of the RD12 Timing&Control (TTC) PCI mezzanine[6] (PMC). This adapter card is based on the design experience gained with the PCI-SCI mezzanine card and will allow to equipstandard VME processors with the common TTC fiber network foreseen for the LHC experiments.

Several SCI activities and projects have started in Europe: Examples are the PASHA ECC Project on scalable computers[7], the SMILE project [16] at TUM/MPI Munich and the poposed SISCI project.

Previous status reports ([1], [2], [3], [4]) and other publications are available on the public ftp server rd24.cern.ch in sci/RD24_Info/ or on WWW using http://www1.cern.ch/RD24. A specific technology WEB page on SCI, PCI, VME and other link standards, and in particular the RD24 PCI-SCI technology is http://sunshine.cern.ch:8080/

[1] see http://sunshine.cern.ch:8080/LVDS/

[2] see http://noah.informatik.tu-chemnitz.de/hard/scicluster/scicluster.html

[3] See press release http://www3.sco.com/Company/Announce/p061396a.htm

[4] see http://www.sequent.com/public/solution/numaq/questions/technology/whtpaprtoc.html

[5] SUN press release dated 8.october 1996 on OEM agreement, see also http://www.sun.com:80/servers/news/launch/press/cluster-debut.html

[6] Under design by J.Ferrer-Prieto, doctoral student of RD24

[7] See http://sunshine.cern.ch:8080/ESONE+IEC/Valencia-96/abstracts/PACHA.press

Chapter 1. Confirmed SCI properties for DAQ systems

SCI has been proposed as scalable DAQ datalink for LHC since the ECFA studies[6] for LHC in 1990. After 4 years of R&D within the RD24 collaboration, during which SCI became an approved IEEE/ANSI standard with participation of RD24[1], major milestones have been achieved with early SCI prototye components. This research program was based on industry collaboration with Apple Computer ATG[2], Dolphin Interconnect Solutions AS[3], IBM-AS400[4], CES and Thomson CSF[5]. Today's commercial availability of both cheap SCI adapters and high performance desktop SMP sytems is a timemark for RD24 to conclude on the following properties of SCI which appear most relevant to the design and construction of scalable LHC DAQ systems:

1.1 Shared memory

Transparent access to remote memory over SCI links has now been established with standard VME processors and off -the- shelf Personal Computers, via PCI to PCI adapters. The remote memory access feature will become even more visible with new industry desktop servers[6] which provide direct SCI nodes with SCI cache support and confirm the prediction that 3rd level trigger processing and readout could be based on low latency (shared) memory access, i.e. the 3rd level CPU's may include the front-end like local memory for processing. Compared to the conventional DAQ procedure of a.) transferring all event data b.) processing, the savings in DAQ bandwidth across an event builder can be as high as the ratio of rejected/accepted events. The global memory feature also allows that both push and pull architectures may coexist in a SCI based DAQ system such that no a priory decision on one architecture has to be taken with SCI. The golbal memory scheme of SCI also implies that calibration data may be downloadad into VME units connected via SCI.

1.2 Message passing

A subset of the non-cache coherent SCI protocols supports message passing which is slightly slower in terms of latency compared to shared memory, but considered as more robust. An example are high throughput, low overhead DMA engines as implemented in PCI-SCI adapters which may transfer message-type of information. Interrupts can be generated via dedicated registers in the remote nodes. Atomic operations, like a read-modify-write can be invoked in a remote node via the SCI lock transactions in order to synchronize messages.

1.3 Low latency

The latency for memory accesses adds only a few microseconds to the equivalent access to local memory, even on PCI-SCI nodes which have no cache support. Industry desktops which apply direct SCI connections to CPUs via 3rd level caches achieve that remote SCI accesses are faster than accesses to local memory[7]. This confirms that characterization that the SCI standard works like a bus, in spite of physically being a network of cables or fibers and switch fabrics.

1.4 Scalable systems

Interconnection of many SCI nodes is logically equivalent to using a "very large bus" but without the well known saturation effects of a bus. The mechanism to prevent bus saturation, called scaling, is possible by breaking bus transactions into short packets[8], which take only a very small fraction of the bus bandwith. Further, the "large bus" consists of small SCI rings which are composed of very fast

[1] RD24 organized three international SCI meetings: Frascati May 1992 , CERN September 1993,Valencia June 1996

[2] Member of RD24 till 1995

[3] Contract partner of RD24

[4] Collaboration on SCIL link tests in 1995

[5] Member of RD24 and partner in the TOPSCI Eureka Project till 1995

[6] Sequent Computers and Data General Corporation

[7] reported on Sequent's presentation on SCIzz Association's SCI meeting St. Clara September 24, 1996

[8] ~ 160 ns for a 64 byte, ~65 ns for a 16 byte packet

point-to-point connections. The latter allow for optimal signal termination, achieving 4ns symbol[1] recognition with the latest CMOS Link Controllers [Ref.[20]]. A single SCI ring with several SCI nodes may be loaded up to more than the individual point-point bandwidth of every link segment, however the ring's scaling properties break down at approximately a factor of 1.5 of the link bandwith, limiting the number of nodes in a ringlet. A ringlet may typically interconnect 10 SCI nodes, or more, if these do not overload the ring. The ringlets can further be interconnected via a SCI switch fabric which is designed to transfer the traffic of each connected ringlet without saturation on average. This is achieved by interconnecting the SCI linc-chip backend via system of very high speed buses[2]. Peak loads may momentarily saturate however SCI retry protocols provide an automatic flow control which smoothes the traffic. The SCI scaling allows expansion of a system from simple point-to point connections to 1000*1000 Node eventbuilder network, which can be implemented for example as 100*100 switch fabric with 100 SCI ringlets on each side (each 10 nodes). Today's "lite SCI" versions, like the PCI-SCI technology will be compatible with fully implemented cache-coherent SCI versions: each SCI packet has fields which allow the use of these options. The scaling of multistage switches has been confirmed by SCI modelling and by tests with small 4-way switches as reported in the 1995 status report [3]. The work on modelling of switch fabrics which become available early in 1997, is continuing within ATLAS [page 27].

1.5 Guaranteed delivery of data

SCI transactions consist of an end-to-end request and response protocol, acknowledging data delivery back to the source. Intermediate buffers in each node within an SCI system confirm the transfer via echo or retry packets, i.e. in large systems SCI packets are stored in the queue buffers of the intermediate SCI nodes, reducing pileup at the source. Each packet is protected via a CRC trailer which is checked by SCI node hardware and which can initiate retransmission in case of an error.

1.6 Low price

An SCI-PCI node adapter consists of two ASIC's which will be merged into one ASIC at the same price level. As a price reference per SCI node: PCI-SCI adapters, including software driver are initially priced at ca. 2000 USD in single quantity and expected to drop to 800 USD in volume quantity in 1998. The building block for SCI switches, the new SCI LC-2 chip is targeted for volume at 50 USD such that the price per SCI node on switch fabrics is expected to drop significantly in the coming years.

1.7 Robust physical layers

The development of SCI cables, connectors and fiber-optic links for SCI has undergone several iterations, starting from expensive twin-axial ECL cables to standard, SCSI like, twisted pair cable assemblies. The LVDS signaling[3], developed as a family member within the SCI standard has been adopted by National Semiconductors , LSI Logic [12] and others and became a Telecom Industry alternative standard[4] to ECL with better performance at much less power consumption. LVDS-based SCI links and cables have undergone a sophisticated study at the SINTEF [7] institute which reports a Bit Error Rate of less than E-14. The manufacturers of parallel optical fiber links [9][10] (Motorola's OPTOBUS and Siemens's PAROLI links) provide their fiber links with LVDS interfaces and a mode to directly connect with LVDS SCI Link Controllers. It is expected that such parallel fiber links will replace copper cables due to better error immunity, longer distance and equal cost. A first test with OPTOBUS-SCI is under way at the University of Oslo [See section 5.5 on page 23.].An ongoing standardization effort[5] which includes parallel fibers aims to achieve an interconnect standard which complies with the particularly harsh environment of an LHC underground experiment.

[1] SCI symbol=2 byte

[2] BLINK bus, a 64 bit multi-master up to 1 Gbyte/s, is the direct interface of Linc Controller LC-1+2

[3] See http://sunshine.cern.ch:8080/LVDS

[4] Telecom standard EIA/TIA-644

[5] ISO WG15 progress be presented at ISO-IEC/JTC1/SC26 plenary S.Francisco, October 22 ,1996

1.8 VMEbus SCI nodes

PMC mezzanines[15] which use the PCI bus protocol are a preferred way to connect add-on equipment to modern VMEbus modules[1]. Both a commercial SCI-PCI adapter[2] in the format of a PMC mezzanine and a PMC adapter[3] designed by CERN/LBL are available and in use for tests by users at CERN, LBL-Berkeley, SLAC and University of Valencia.

1.9 Desktop SCI nodes

SCI adapters for PCI in desktops by Dolphin are produced in large quantities for SCI customers. Based on a PCI-SCI ASIC [Figure 8], these cards are available with software drivers for WindowsNT, Solaris and Unix. The CERN-developed SCI adapter for the PC [Figure 6], used for a part of the 1996 milestone, is also in use as development card by four external projects which develop application specific software and hardware.

1.10 Adapaters to other bus/link standards

The development of various SCI adapters to other bus/link standards is under way, or completed. Examples are: PCI-SCI [Ref.[38]], HIC-SCI[4], FiberChannel-SCI [page 24], ATM-SCI[5]

Chapter 2. SCI link Controllers

SCI nodes are implemented with SCI chips (Nodechips or Linc Controllers) which are directly inserted in an SCI ring and handle the incoming and outcoming packet protocols. Incoming packets are stripped from the ring and transmitted to chip-internal queue buffers before retransmission to the backend application bus. A link bypass provides that other SCI packets are immediately transmitted to the outgoing link. The bypass delay on the latest CMOS LC-2 chips is only 48 ns, the packet transfer time from the link to the application side only 114 ns.

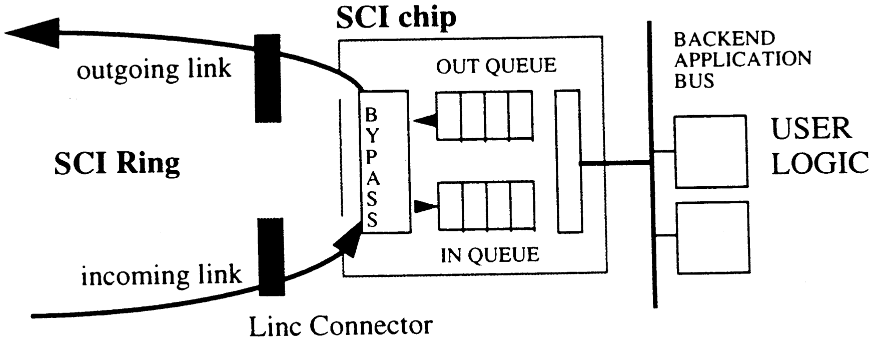


Figure 1: SCI Nodechip/ Link Controller

2.1 500 Mbyte/s GaAS nodechip (1993)

The first tests with the 500 Mbyte/s GaAS Nodechip from Dolphin were reported by the RD24 collaboration in 1993 [2]. First SCI packets were transmitted between the CERN/RD24 DMA adapter (VME) and the SCI adapter for the Macintosh by Apple Computer ATG in September 1993. This first SCI chip used ECL link signals and consumed up to 30 Watt. Due to advances in CMOS technology and availability of ECL-compliant signaling cells in CMOS (PECL) this expensive technology was converted into CMOS ASIC technology in 1993.

[1] A list is maintained by VITA standards Organization under http://www.vita.com/vmeprod/processors

[2] VMETRO, Oslo, Norway http://www.vmetro.com

[3] see http://sunshine.cern.ch:8080/PCI/PMCdesign

[4] HIC-SCI has been worked on by Dolphin, SINTEF and University of Oslo (Informatikk) within the OMI/HIC, OMI/Macrame and OMI/Arches Esprit projects

[5] A 2.4Gbps ATM-SCI chip, 0.5 um CMOS, is under development by Acorn-Networks, Inc. Virginia see http://www.acorn-networks.com and a 622Mbps ATM-SCI board (FPGA) by DAA in Utah

2.2 CMOS Nodechip (1993)

The first 5 Watt CMOS Nodechip, implemented in 200 Kgate CMOS technology operated at 125 Mbyte/s link bandwidth and was produced by LSI-Logic. The Nodechip was used on SBUS-SCI adapters for SUN Workstations and a variety of SCI development projects, most of which are reported in our 1994 status report [3]. This chip required a cooling header and allowed only one outstanding SCI transaction at a time.

2.3 200 Mbyte/s Linc Controller LC-1 (1994)

The first 300 Kgate CMOS link controller (LC-1) from Dolphin appeared in spring 1995, operating at 200 Mbyte/s with PECL signals on the link. This chip comsumes only 3.5 Watt in a flat 208 pin plastic package without cooling requirement and allows for 3 outstanding SCI transactions. The BLINK bus [13] on the application side is specially adapted for interconnection of SCI-family chips (i.e other linc controllers or cache/memory controllers) and enables in particular the construction of switches.. The LC-1 is in use in a variety of SCI equipment like the SBUS-II adapter or the 4 way-switch and the PCI-SCI adapter from Dolphin.

2.4 1Gbyte/s IBM SCIL chip

The IBM SCIL chip, implemented in 0.8micron BiCMOS technology was presented at the Fermilab DAQ conference in 1994 [19]. Using an LVDS-like signaling technology and a non-SCI compliant 32 bit CRC trailor, this chip served primarily as a high speed test generator for parallel optical links. A test chip which was available for RD24 during 1/2 year (see status report 1995 [4]) but since it had no backend interface the generation of SCI packets relied on preloading of data patterns via a serial line.RD24 has not persued this technology and is unaware of further development.

2.5 Lincchip from ISS

Due to non-disclosure agreement with ISS [14] in California, not much information can be relased on this Linc chip. In contrast to all other SCI chips it uses narrow 8+2 bit SCI links and therefore interfaces directly to 10bit wide optical fibers. The backend bus is 64 bit wide. This chip has been designed by the desiger of Unisys "datapump" . It uses a very small core and could therefore become a very low cost commodity part.

2.6 New 500 Mbyte/s LVDS Linc Controller LC-2 (1996)

Dolphin's second generation Linc Controller [20], named LC-2, is implemented in 500 Kgate CMOS and uses for a first time LVDS signals in order to achieve SCI symbol handling every 4 ns.

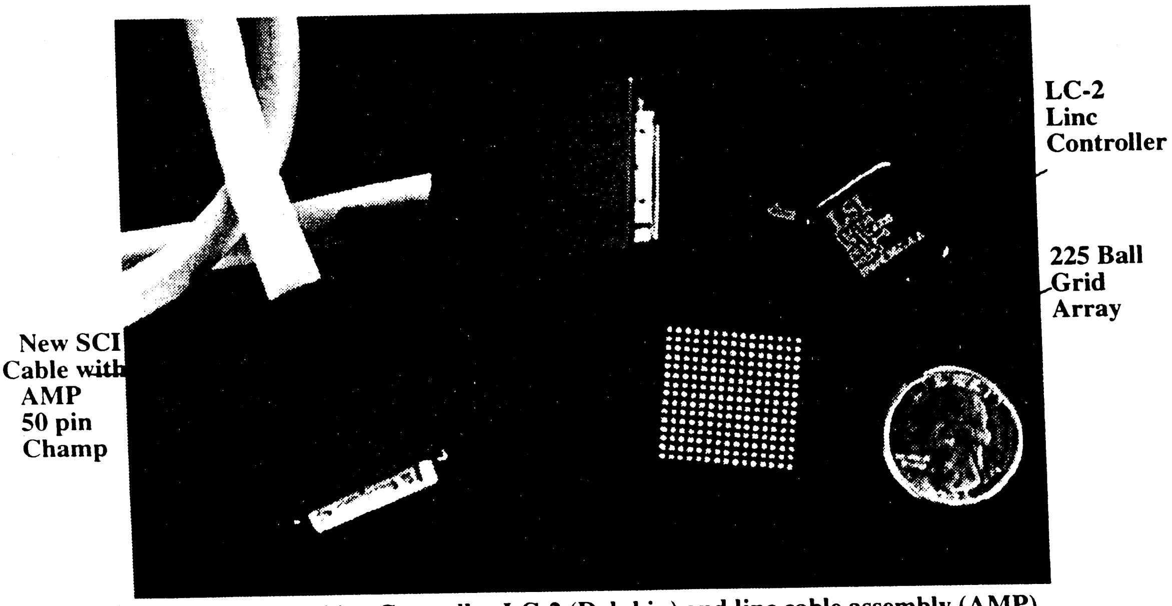


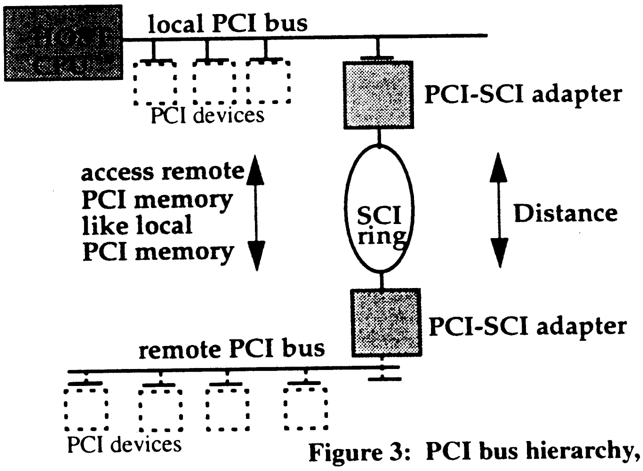
Figure 2: Linc Controller LC-2 (Dolphin) and linc cable assembly (AMP)

Signal termination is implemented inside the chip such that the numerous external termination resistors, needed for the LC-1 are not required. The power consumption was reduced to only 2.5 Watt allowing to place this chip on a very flat Ball Grid Array (225 PBGA) carrier which favours its use in very dense environments like the PMC mezzanines standard P 1386.1 for VMEbus. The price target for large PC server applications is less than 50 USD. Samples of this chip are currently under evaluation by Dolphin for immediate integration in all previous SCI products as enhancement to the LC-1 product line. Dolphin is in particular preparing a brideboard between 200 Mbyte/s LC-1 and 500 Mbyte/s LC-2 links. With the LC-2 introduction, a new SCI cabling system [Ref. [8]] , based on point-to-point link cable assemblies with 50 pin, 0.8 mm connectors from AMP is introduced for a link length up to 7 m. Two "Champ" high density connectors[1] fit mechanically on the front panel of mezzanine standards (SBUS, PMC, EISA card, PCI card) to accomodate incoming and outgoing SCI links. Link distances up to 300 m will become available via 10bit OPTOBUS parallel fiber technology from Motorola which will implement a direct LVDS interface for SCI [Ref.[10]].

Chapter 3. PCI-SCI adapters

The PCI bus [23], proposed by Intel in 1992 was conceived as low cost, component-to component CMOS connection standard which requires no tranceivers, no glue, no jumpers (plug&play), and provides high bandwidth. The standard PCI bus connections, available today in a multitude of PC's and also a variety of VMEbus CPU's are implemented with 32 bit datawidth, clocked at 33 MHz, resulting in a theoretical peak bandwidth of 132 Mbyte/s. The practically observed bandwith with standard PCI chips sets achieves up to 80 Mbyte/s for long PCI bursts. A factor of four in the theoretical bandwidth is possible for future PCI implementations by doubling both the data path and the clock frequency.

The common availability of PCI in dektops and VMEbus[2] since 1994 became a cost-effective solution to implement scalable systems via PCI-SCI adapters. PCI bus hierarchies are normally constraint to small distance (due to its reflected wave CMOS singnaling concept) whilst logically, PCI hierarchies of large size could be used with intermediate PCI-PCI bridges. Since memory accesses over SCI are preserved, remotely connected PCI hierarchies can be constructed by using SCI as the intermediate link [Figure 3.]



Expansion of local PCI bus hierarchies via SCI

- PCI Memory space transferred over distance via SCI
- SCI rings can interconnect several PCI nodes without saturation
- Latency from PCI to PCI over SCI is O(5 us)
- Initialization goes downstream from host
- Upstream communication allowed
- All PCI devices may initiate transactions
- Concurrent operation on each PCI bus

Figure 3: PCI bus hierarchy, remotely connected via SCI

The PCI-SCI technology may be used on both low-cost and high-performance equipment and it will be compliant with enhancements to SCI. This is due to the fact that the SCI logical protocols are the same for light SCI or fully cache coherent SCI. The question whether to implement message passing or shared memory protocols is independent of SCI, both paradigms are being implemented and studied.

[1] CHAMP 0.8 mm Cable Assemblies, AMP Inc. Harrisburg, PA 17105, Information 1-800-522-6752

[2] IEEE P1381 mezzanine formfactor standard, called PMC

3.1 The CERN PCI-SCI adapters

A major milestone of RD24 in 1996 was the interconnection of VME with a desktop PC via the PCI bus. This required the design of a PCI-SCI bridge in the PMC mezzanine formfactor and a PCI-SCI bridge for insertion into a standard PC. Both cards, functionally equivalent, have been built in small series at CERN and are in use by collaborators, in particular also by the SMILE project at the TUM/ MPI München for a construction of the W7-X data acquisition system[16]. In order to allow for improvements and implementation of specific DAQ functions, the RD24 design uses an approach where functional modules are developed in VHDL language. All bridge functions, including dedicated optional functions are synthethized into very high density programmable FPGA chips[1] which can be reloaded even in a running system. A four-node SCI ring system, consisting of threeVME CPU's and

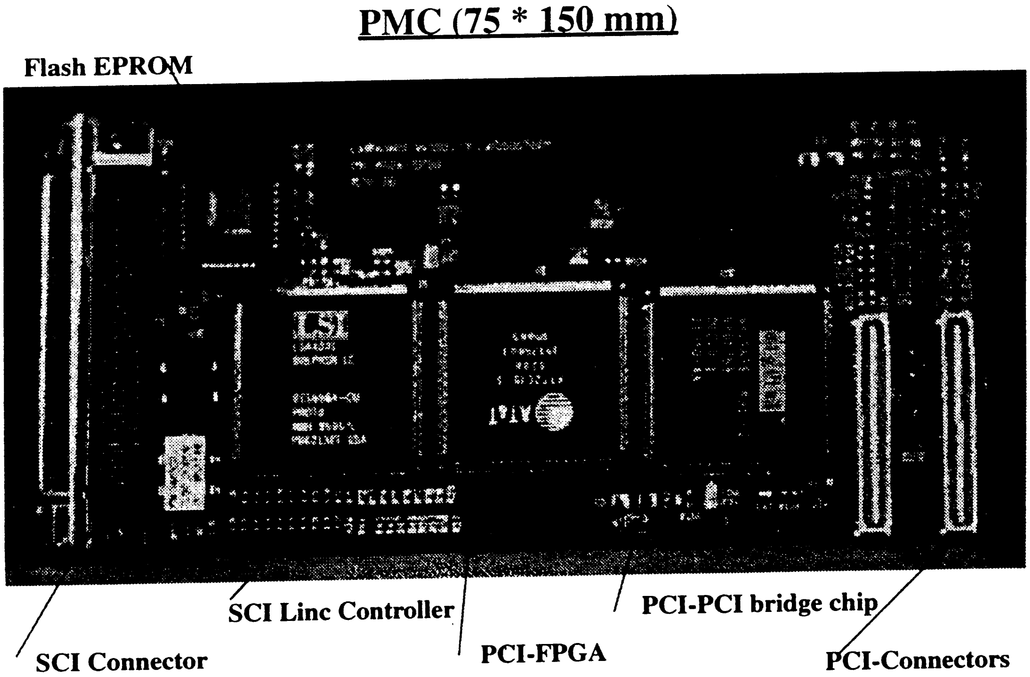
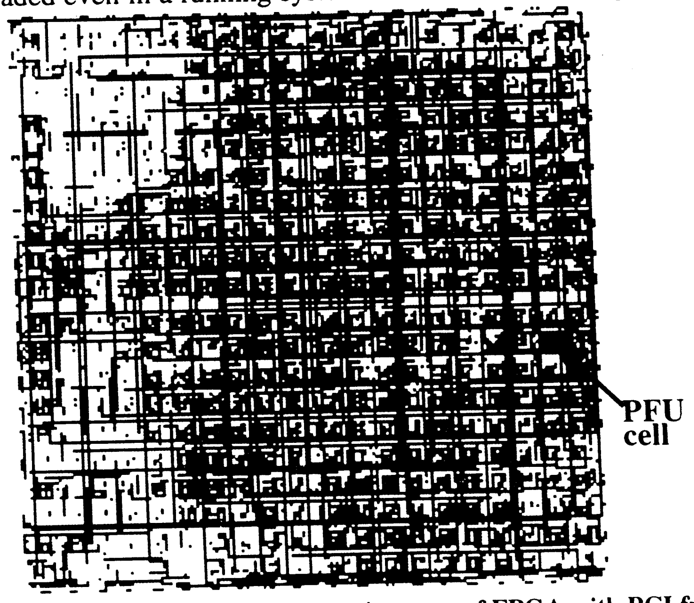


Figure 4: PCI-SCI mezzanine card for VME (CERN/LBL)



PCI-FPGA on PCI-SCI

VHDL generated PCI core with DPM, EPROM and FiFo Interface for ORCA 15 Kgate FPGA

FPGA occupancy:
IO 46%
PFU: 81%
Nets: 2874

PCI compliance:
100% at 33 MHz

Figure 5: Routing map of FPGA with PCI functionalities (RD24)

[1] ATT's ORCA 2C15 and 2C26 FPGAs.

onedesktops PC (see "Heterogeneous 4-Node SCI interconnect test" on page 20) was operated with RD24 PCI-SCI bridges to achieve data transfers up to 69 Mbyte/s (saturation point of PCI bus in use). Requests for commercialization of these bridges have been received, requiring that the totality of functions are still developed. In the current status, ca. 20 percent of functionality remain to be implemented (see Figure 7) but cannot be continued at CERN due to lack of resources in RD24. The transfer of the CERN copyright to the Technical University of Chemnitz, Germany for an SCI project[1] for dedicated clustered SMP applications is under discussion.

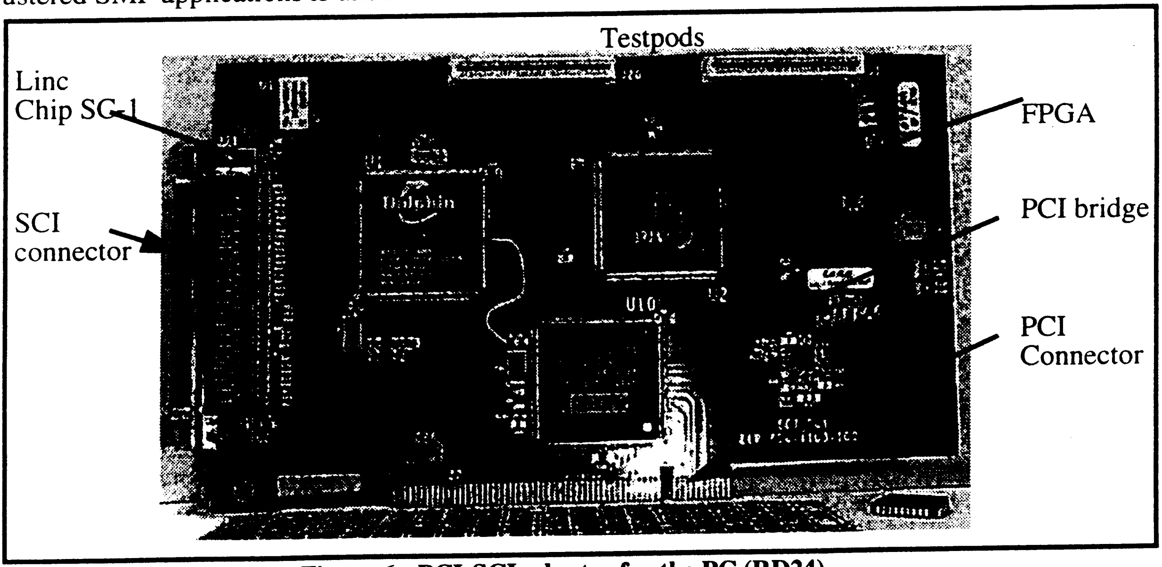


Figure 6: PCI-SCI adapter for the PC (RD24)

The target functionality of both PCI-SCI adapter modules[2] which are electrically equivalent includes three operating modes: Packet mode, DMA and Transparent mode. Both the DMA and the Packet mode have been fully implemented, the transprent mode remains to be implemented. A large part of all required features for the full implementation is completed as shown below:

Figure 7: Target features of the CERN PCI-SCI bridges

Implemented & tested:

- PCI slave: a PCI 2.1 compliant PCI slave with configuration space
- Dual Port Memory Interface: implemented in PCI's prefetchable memory space
- EEprom interface: implemented in PCI I/O space. Main use: in-situ reprogrammation on the EEprom
- FiFo interface :I/O ommands are queued in a FiFo in the PCI I/O space
- Blink sequencer: BLINK device No 0, the interface between the SCI link controller (BLINK device No 1) and the PCI master agents (master and slave)
- Free buffer list: a buffer list maintained by the packet manager
- Chain DMA engine: BLINK device No 2,a chained data mover between PCI and SCI space
- CSR registers: several control and status registers implemeted in I/O space

Partially implemented:

- PCI master: a PCI 2.1 compliant master with a subset of PCI protocols
- Packet manager: the state machines which handle request and responses

Not yet implemented

- Interrupts: a number of conditions which can produce PCI interrupts
- transparent transactions controller: build automatic SCI packets and evoke address translation

[1] http://noah.informatik.tu-chemnitz.de/hard/scicluster/scicluster.html

[2] See PCI-SCI user manual chapter "General Features " on PCI-SCI under http://sunshine.cern.ch:8080/PCI/PMCdesign

3.2 Collaboration on the PCI-SCI with the STAR Trigger group

The packet mode version of an SCI-PCI bridge, that was designed in collaboration between the Lawrence Berkeley National Laboratory and CERN RD24, has been completed in October 1995. It is in operation since in the STAR system test. SCI has been chosen as the network standard for the STAR Trigger system.

In the mean time the Berkeley group has started to standardize the first software layer that was developed at LBL for the packet mode SCI-PCI bridge.This effort has emerged into the IEEE P1596.9 working group.

3.2.1 Evaluation of PCI-SCI on an AXPvme Computer by the DEC Joint Project

The work was based on the standard software framework as defined by the RD24 collaboration. The framework was ported on the Alpha Architecture, by implementing a simple Digital Unix driver to access physical addresses by a user process, and by creating the Alpha Architecture target dependent-module. The combination of the driver, the target dependent module, and the common software framework enabled us to drive the PCI-SCI mezzanine from a user process running on the AXPvme 160. This was the first port of the software framework to a 64 bit architecture, therefore is helped finding some portability bugs in it. The combination of AXPvme and PCI-SCI mezzanine was shown to sustain more than 42 MB/sec at the PCI interface, with most of the PCI memory transactions being clustered in 8-longword bursts.

3.2.2 Collaboration on PCI-SCI bridge development with the TU Chemnitz

The Technical University in Chemnitz, department[1] "Rechnerarchitecture und Mikroprogrammierung", is working on dedicated SMP architectures for numerical applications in physics and mathematics. Within this work, SCI connections via PCI adapters are needed, which allow for implementation of user-specific functions. The RD24 PCI-SCI bridge, which allows implementing of user-specific functions via VHDL synthesis, is under test at TU Chemnitz.In order to take a decision on a transfer of CERN's copyrighted VHDL development for this bridge, the VHDL sources have been sent to Chemnitz for evaluation.

3.2.3 Collaboration with SMILE project at Technical University Munich (LRR-TUM)

Researchers [Ref.[17]] at the Informatics Unit at the Technical University in Munich[2] (SMILE project[3]) investigate SCI with respect to its networks capabilities for cost effective platforms which are needed to further develop programming tools for parallel computer architectures, based on shared memory. The SMILE program[4] includes development of an SCI based Pentium-PC cluster with distributed shared memory, as well as modelling and simulation of SCI networks. Two PCI-SCI adapters of RD24 are in use for building an SCI-coupled PC ringlet. The Linux drivers and C-language utilities developed at CERN and at the University of Frankfurt are in use and further developed and ported to the Windows-NT environment. The goal at TUM is to design a dedicated PCI-SCI interface which allows to add special hardware for monitoring remote memory access. As part of a diploma thesis, this PCI-SCI bridge is based on VHDL synthesis for Xilinx FPGAs. The major difference to the RD24 bridge architecture is the use of an Intel I960 Processor with PCI port in place of the PCI-PCI bridge of the CERN design.

The use of SCI is considered for the DAQ system at the W7-X Fusion Experiment prepared by the Institute for Plasma Physics (IPP) in Garching.These systems consist of VMEbus front ends and desktop Workstations interconnected vis SCI-PCI adapters and an intermediate SCI switch. The CERN PCI-SCI adapters and a Dolphin 4-way SCI switch is in use for evaluation for this experiment.

[1] see http://noah.informatik.tu-chemnitz.de/hpra.html

[2] see http://wwwbode.informatik.tu-muenchen/index.html

[3] SMILE = Shared Memory in LAN Environments

[4] see http://wwwbode.informatik.tu-muenchen.de/sci/

3.2.4 Collaboration on PCI-SCI driver with University of Frankfurt

The test of the RD24 PCI-SCI bridge in a PC environment required development of a driver to test our card in an Intel PCI environment. During the design phase at CERN we opted for the Linux operating system and developed a driver for the packet mode operation of the bridge. This driver was further developed via a collaboration with students of the "Institut für Kernphysik" at Univeristy of Frankfurt who also collaborate with the STAR experiment.

A m ajor goal of this development was the performance optimization. The new driver, reported at the European SCI meeting in Valencia[Ref.[18]] avoids kernel mode and uses C++ for modularity. As improvement to the calling conventions used by original CERN driver, a proposal for a general purpose application interface[1] (API) [Ref.[17]] for SCI was derived.

3.3 The Dolphin PCI-SCI adapters

The target applications for PCI-SCI adapters by Dolphin Interconnect Solutions are commercial database clusters, server clusters, the scalable I/O interconnect and bus bridging for embedded applications.The latter enables bridging of VME via PCI-SCI adapters. Dolphin implementation of PCI-SCI bridges uses ASIC technology with the advantage of low price at high quantities and a stable configuration for software. RD24 has acquired Dolphin bridges which come with a driver for Windows NT to performe several measurements within the RD24, ALICE and ATLAS programmes.

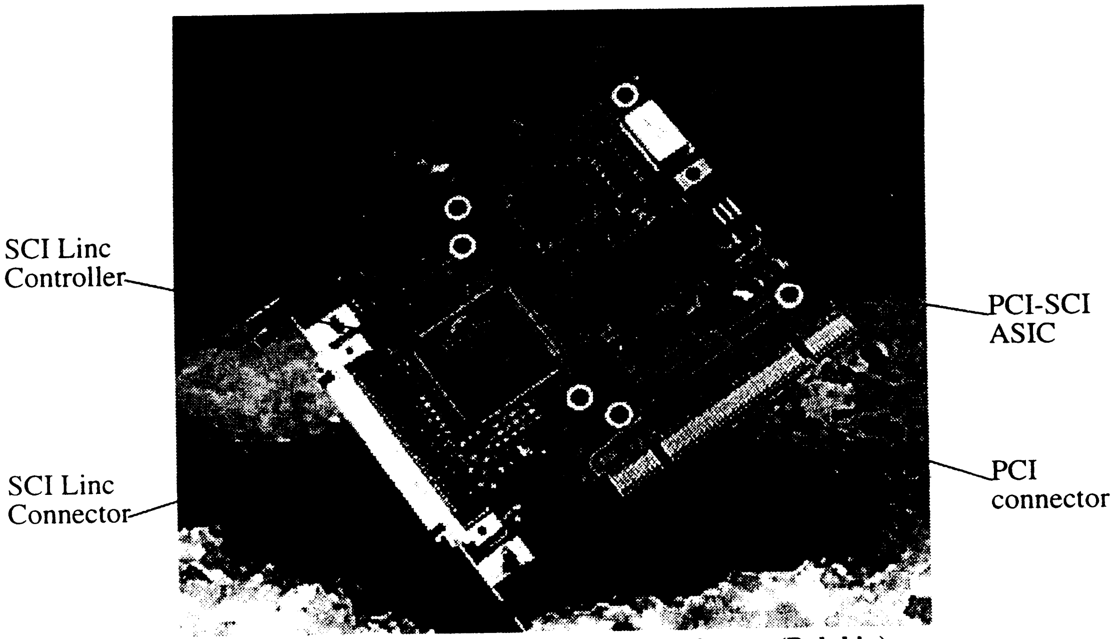


Figure 8: PCI-SCI adapter, short PCI card format (Dolphin)

Supported features:

- Address translation cache (in system memory): CPU accesses to remote protected memory though address translation table.
- Address protection on incoming accesses
- Mailbox/Interrupt: via remote access to interrupt register or during atomic transaction
- Atomic Operations: read-modify -write operation on remote PCI bus
- Chain mode DMA: write or read from remote memory without intermediate buffer
- Read prefetch and write gather buffer: 8 entries , 64 byte

[1] see http://www.ikf.physik.uni-frankfurt.de/~roehrig/sci-ikf.html

Performance figures:

- 200 Mbyte/s SCI symbols on incoming and outgoing SCI links see
- Delivery of 256-byte message:less than 10 us over 100 meter
- one-way latency PCI->SCI for 64 byte operations: 1.1 us
- one-way latency PCI->SCI for 16byte operations->SCI 630 ns
- User latency (zero length message, point-to point completed) 4us
- Node bypass latency 140 ns
- Extrapolated(optimal hadware and software environment) thoughput: read 97 Mbyte/s write 102 Mbyte/s
- DMA read or write throughput: 67 Mbyte/s

Software support:

Solaris, Windows-NT, UnixWare (SCO)

3.4 DAQ architectures based on PCI-SCI technology

A common requirement of DAQ systems is the collection of data from bussed data sources into a collection buffer for online processing and transmission towards a data destination. Since the latter is normally a remote event buffer, data are tranmitted via data links (fibers, cables). In the LEP Fastbus systems the cable segment and segment interconnects provided naturally such dual port architectures. In VMEbus, the VICbus plays a similar role. In practise, numerous dual-port buffer-transmitters for data links have been designed for VME in order to provide better buffering, longer distance and higher bandwidth.

The availability of PCIbus mezzanines in VME CPU's and VME bridges opens the new possibility of using standard PMC mezzanines as adapters to high speed data link standards like SCI, FiberChannel, HIPPI or ATM[1]. The advantage, apart from using industry supported standards is the independence from VMEbus manufacturer-specific adapters, the jumperless autoconfiguration of PCI (plug&play) and the high bandwith achievable.

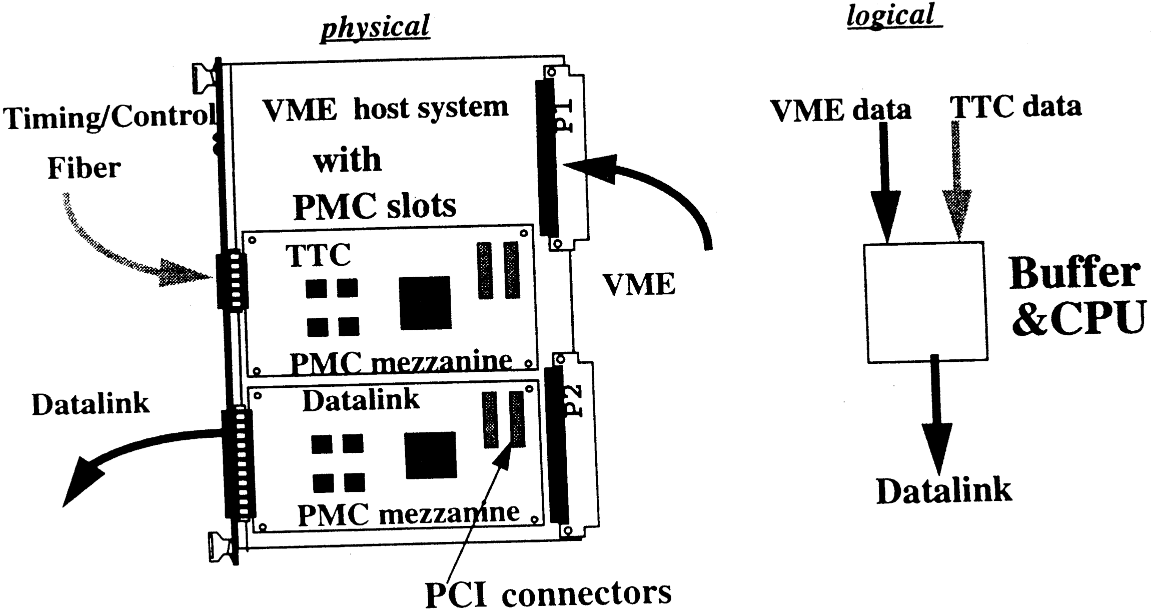


Figure 9: Dual port equivalent implemented with PMC's in VME

This technology also provides better ways to build scalable DAQ architectures which requires that the addition of more sources does not saturate the interconnect medium nor the destination. The choice

[1] For information on these standars see http://sunshine.cern.ch:8080

of the datalink standard must therefore be based on its scalability criteria. As a first step, the PCI "plug and play" technique shown in Figure 11 can be viewed as the logical equivalent of a multisource data merger. In physical terms this requires per VME bus one module which is equipped with PMC carriers. Defacto, an impressive number of such modules (with or without CPU) are already existing[1]. RD24 has built a 4-PMC carrier for VME 9U which reduces the crate interconnect overhead. Due to the maximum bandwidth of the PCI bus such a merger function will saturate at the PCI bus performance limit (at ca 80 Mbyte/s with PCI-32bit however factors of four are possible with future 64 bit enhancements to PCI bus).

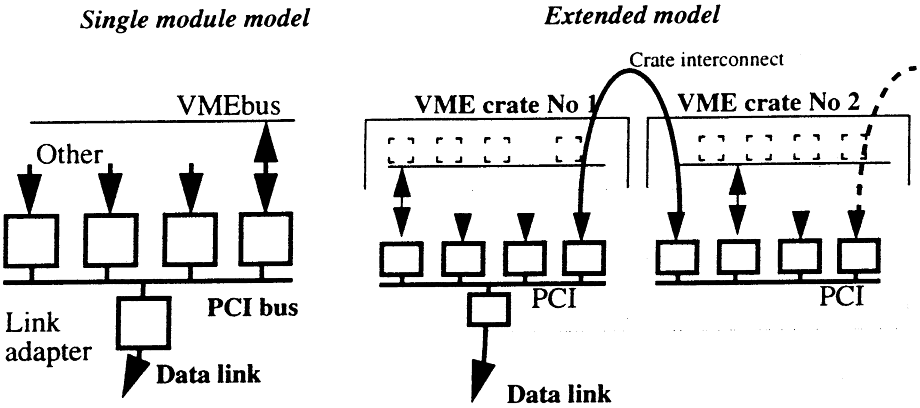


Figure 10: Logical view of a PCI based data merger unit

An additional economical requirement is the crate interconnect as a technique to group a number of physical buses to appear as one unit which can be read out by one data link (a LEP example of this are the crate extensions of Fastbus). This can be implemented with the same technique via a PCI-PCI datalink which must allow to map the VME adresses across PCI.

With the model of a data merger unit as logical node of a link standard, the investigation of the scalability properties of a large DAQ system can be reduced to the link system properties which are well defined within their standard. The SCI standard IEEE 1596, which was designed as a scalable computer interconnect, plays a prominent role since it provides, apart from scaling, a low latency, bus-like environment with guaranteed data delivery and robustness at relatively a low price (See "Confirmed SCI properties for DAQ systems" on page 4.) The logical view of a DAQ system implemented in SCI [Figure 11] can be totally based on SCI generic parts. The parameters of the building blocks of an SCI

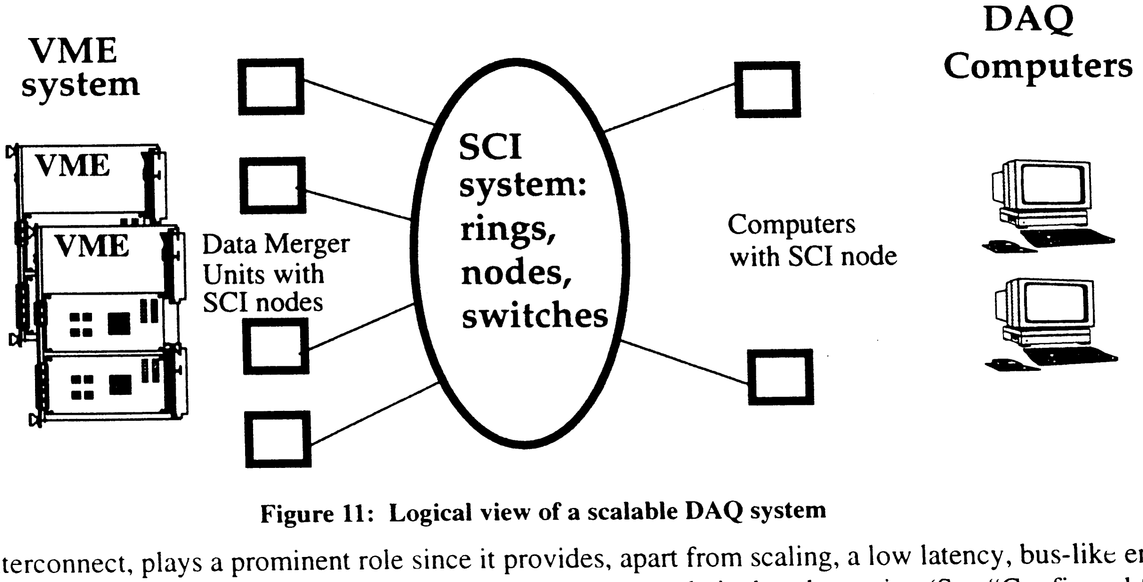

VME system
DAQ Computers
VME
VME
Data Merger Units with SCI nodes
SCI system: rings, nodes, switches
Computers with SCI node

Figure 11: Logical view of a scalable DAQ system

[1] For a list of VITA compliant PMC mezzanines see http://www.vita.com/mezzprod/mezzdirindex.html

system are being collected and refined for a series of SCI products.

For modelling techniques like MODSIM, descriptive parameters, like the transfer time, are needed to describe a particular component. In reality these parameters need to be measured with real systems. Using the VHDL simulation language, a behavioural model of SCI units can be decomposed into fine-grain structural components which allows to better analyze realtime behavior. Reversely, VHLD may allow to synthethize functional parts into hardware.

Chapter 4. SCI's position in industry

The SCI activities in industry target the market of scalable interconnect systems, i.e. the interconnection of CPUs, mostly 4-way SMP servers, with disk systems, PCI based IO, and LAN/WAN networks. These systems prefer message based protocols for I/O and processor intercommunication and shared storage via SCI's cache coherent protocols.

4.1 SCI technology from Dolphin Interconnect Solutions AS

Dolphin Interconnect Solutions AS in OSLO[1] (headquartered in Westlake Village, California and offices in Waltham, Massachusetts) is an RD24 partner since 1992. Dolphin works primarily on SCI products for the server market. Since the last RD24 status report, Dolphin has acquired key assets of Kendall Square Reserach and BBN Toolworks division. and counts today 72 employees.

Dolphin is continuing the LinkController product line with new projects. The time between first working prototype for the 200 MBytes/s LC-1 and the 500 MBytes/s LC-2 was only 14 month. The project to make LinkController-III which will use a new LSI technology and is expected to work at 800 MBytes/s has started with the expectation to have samples in 12 month from now. Dolphin is committed to launch new switch products, based on the LC-2 in January 1997. A 6 way switch (4 user ports and 2 expansion ports) and a 16 way switch is planned. Performance testing of the switches will be a part of the proposed SISCI Esprit project (see "Software for LHC Experiments" on page 29).

Dolphin further collaborates with Sun microsystems on Oracle database servers. A collaboration with Data General Corporation and Intel on with cache coherent SCI adapters for SHV nodes is progressing (see "Data General Corporation NUMA architectures" on page 18). Further collaboration include a PCI card for Novell and an agreement with Siemens-Nixdorf (SNI) to use Dolphin's PCI-SCI bridge chip and SCI protocol engine to develop a scalable I/O system for new multiprocessor Unix server systems. Dolphin and Motorola will announce a collaboration for parallel-optical SCI based on the 10bit Optbus. (300m @ 400MBytes/s).

Dolphin makes a distinction between SCI and SCI*lite* products[Ref.[20]], the latter being an acronym for low-cost SCI without complex hardware support for cache coherence, basically SCI adapter within a single chip. SCIlite products are:

- PCI-SCI adapter and the SBUS-SCI adapter
- 4-way and 16 way cluster switch
- Windows-NT drivers for TCP/IP and PCI -direct
- Solaris drivers for TCP/IP and SBUS-direct
- UnixWare driver for TCP/IP.

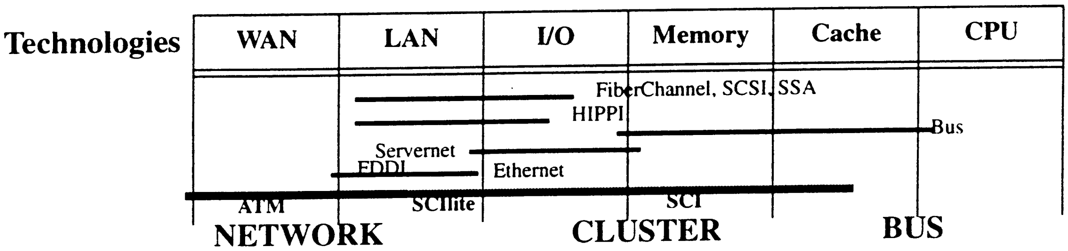
	Pentium Pro bus	**PCI bus**	**Dolphin "SCI-lite"**	**Memory Channel**	**Servernet**	**Fiber channel**	**ATM, Fast-ethernet**
Bandwidth	512 Mbyte/s	512/266/133 Mbyte/s	500 /200 Mbyte/s	100 Mbyte/s	50 Mbyte/s	100 Mbyte/s	10 Mbyte/s
Latency	0.25 us	0.25 us	3 us	5 us	-	250 us	250 us

Table 1: bandwidth/latencies comparison of technologies

[1] For more information and press relseses see http://www.dolphinics.no

The comparison of interconnect technologies in terms of bandwidth and latency is one way to classify their relative position within the communication problem: fast access at higher bandwidth. Dolphin sees the application area of SCI complimentary to ATM for coverage of the whole communication area from CPU-cache to wide area network: the scope of SCI reaches from cache to LAN and the scope of ATM from LAN to WAN. as shown below:

Figure 12: SCI technology Comparison (Dolphin)



Distance	100 000	10 000	1000	100	10	1
Bandswidth	1	100	10 000	50 000	100 000	100 000
Latency	very high	100 000	1000	20	1	1

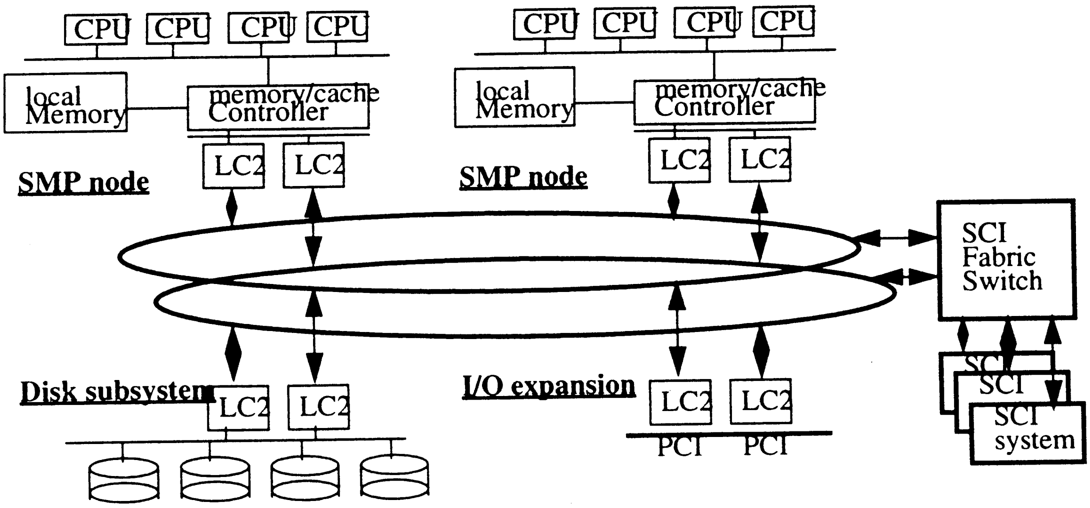
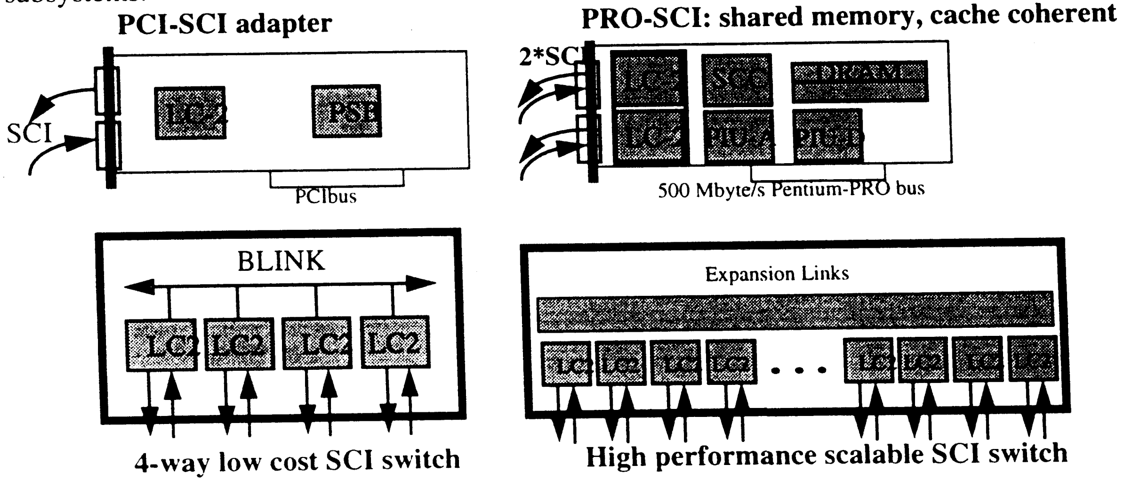


Figure 13: SCI architecture with industry components (Dolphin)

An example of an SCI architecture with commercial components is shown in Figure 13.This example includes SCI switch fabrics, SCI nodes to SMP desktops, SCI nodes to PCI systems and nodes to disk subsystems.


PCI-SCI adapter

PRO-SCI: shared memory, cache coherent

4-way low cost SCI switch

High performance scalable SCI switch

Figure 14: SCI component examples with LC-2 chip (Dolphin)

The dual SCI node products provide both a an economic solution to connect an SCI subsystem to multiple SCI rings for performance enhancements and provide failover capabilities (availability) via redundant links. The SCI components used [Figure 14] are based on the LC-2 Linc controller. The PCI-SCI adapter is the enhanced version of the current PCI-SCI adapter and the 4-way switch is the enhanced version of the 4-way switch which uses a common Blink Bus as internal crossbar. Both parts are in use by RD24. The PRO-SCI board is a new, cache coherent SCI node with direct interface to the Pentium-Pro bus. It is equipped with two SCI nodes for connection to two SCI ringlets which enhances the architectural possibilities. The High Performance switch is a rack-mountable, scalable switch using internal expansion links to provide the required throughput (up to 8 Gbyte/s for a 16 way switch). RD24/ATLAS will use a 16-way version of this switch for 3rd level eventbilder testsData General Corporation NUMA architectures

As reported on the Hotchips IV conference [Ref.[21]] Data General Corporation, in collaboration with Dolphin is building an SCI adapter for the 4-way Pentium-Pro server market in order to extend the shared memory programming model across servers using the Intel SHV[1] architecture. This implies using SCI's cache coherency scheme with hardware support on the SCI adapter cards. The SCI Interconnect adapters desigend with Dolphin (see PRO-SCI card in Figure 14) are designed to fit on the P6 connector of the Intel SHV nodes (4 * Pentium-Pro), including a 3rd level cache and three controllers (PIU-A, PIU-D, SCC). Several 500 Mbyte/s LC-2 Linc-controllers can be connected to allow industry architectures like shown in Figure 13. The latency ratio measured as CPU cycles for local/remote accesses is reported as 10/130.

Chapter 5. RD24 milestones in 1996[2]

The 1996 RD24 milestone [6] on "demonstration of an SCI based DAQ system, incorporating PCI based computers and VME modules" , has been implemented in several complementary ways. We have used SCI adapters with 200 Mbyte/s Linc Controllers for these tests.

5.1 Transparent VME access from desktop PC

Transparent access across SCI requires that one system can access memory of the remote system like local memory. This feature has been demonstrated with SBUS-SCI cards in previous RD24 status reports.

A desktop Intel PC was used to connect a VME module (VMIC 7587) via two prototype Dolphin PCI-SCI adapters which we had available. Both the PC and the VME module were running the WindowsNT operating system. the installation with the Dolphin drivers was straightforward. We used

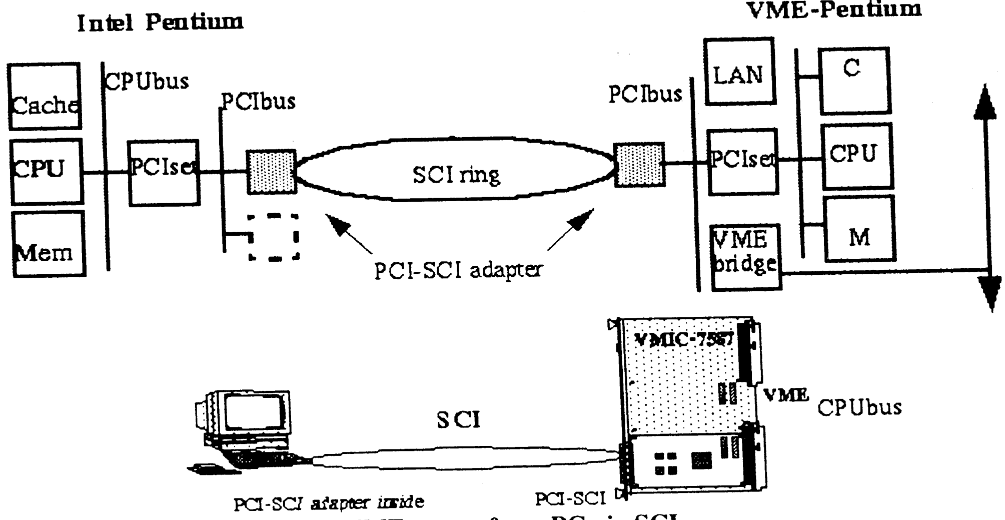


Figure 15: VME access from PC via SCI

[1] Standard High Volume

[2] incomplete, to be completed when commercial PCI-SCI node adapters for VME with transparent mode are available, delivery expected in November 96

Tools, a C program, written under Windows NT by Dolphin for running simple looptests and for accessing remote PCI memory across SCI.

Result: *The transparent access of PCI memory space via SCI works.* Memory on the PCIbus of the other modules can be seen after initialization of the access windows in the PCI configuration space. The Tools program allowed us to measure maxiumum data transfer rates of 20 Mbyte/s from the PC to the PCI bus of the VME module, the latter appears to limit the PCI bandwidth. Further optimizations may be possible by using a revised version of this module[1]. For the access of VME though PCI we need to extrapolate added latency of a VME-PCI bridge, since the optional VME-PCI bridge was not available with this card. Based on a Tundra's Unsiverse VME-PCI chip which has been used by us for crate-extension tests we can extrapolate that at roughly 10 Mbyte/s of shared memory access between an WindowsNT application an VMEbus is possible. With respect of improvements possible (optimization of driver, use of systems with more recent PCI chipsets, more performant Pentium, coming PCI-SCI adapter with LC-2, 64 bit PCI) we regard this number as a lower -limit figure. Any optimization was not possible due to a very limited time during which the onloan VME module was available.

5.2 VME crate extension

Crate extensions are logically not required for architectures, however in a real system, crates need to be interconnected in a cost effective way. As a test for crate extensions via SCI, a processorless PCI-VME bridge module (MIDAS-20 from VMETRO[2]) was used as a bridge for crate extension, interconnected via SCI to a second VME crate. Optionally further crates may be connected via such bridges. The host crate was equipped with a standard VME CPU (Motorola PowerPC) running under the Lynx-OS operating system. The VME bridge is based on the fast Tundra's UNIVERSE bridge chip which interfaces VME and the PCI bus.

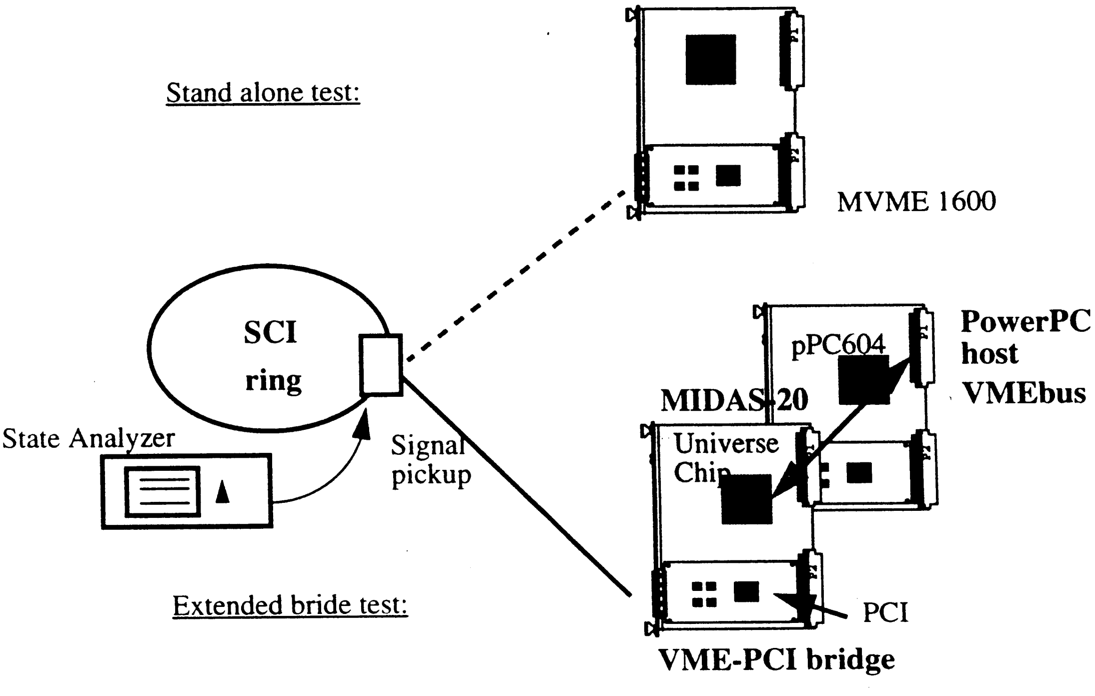


Figure 16: VME crate extension via SCI

In the stand alone test we first operated a loopback SCI ringlet via the CERN PCI-SCI adapter which is directly connected to the PCI bus of the PowerPC. We achieved loopback modes of 10 Mbyte/s in programmed I/O mode (packet mode), which is a measure for the overhead of 4 PCI I/O commands

[1] details on the VMIVME-7587 Pentium based CPU see http://www.vmic.com/whatsnew/7587.html

[2] http://www.vmetro.com

and data copying for the CPU. As a measure for transparent mode (which is not yet implemented, see "The CERN PCI-SCI adapters" on page 9) where only one PCI cycle is used to convert a PCI into SCI transaction without data copying, we achieve 38 Mbyte/s with a PowerPC 604 (see Figure 20 on page 21).This figure includes the acknowledgement (response) from an addressed SCI node.

In the extended mode we used the PCI-SCI adapter in the VME-PCI bridge, i.e. all accesses had now to pass across VME and the bridge. Direct access to the SCI adapter was provided by mapping the host processor's VME address space via the Tundra bridge chip into its PCI address space. We then applied the same testloops as used in the stand-alone test.

It was shown already in "Transparent VME access from desktop PC" on page 19 , we can further map these operations across SCI into further VME crate.

The following measurements were taken with our loop program in the extended mode. Figure 17 was taken with a Logic Analyzer which is clocked in states of 10 ns, derived from the SCI clock.

- VME bus latency for accessing PCI across the VME bridge: AS/VME->DTACK = 610ns
- PCI-SCI adapter's one-way latency for generating an SCI packet after PCI command: 710 ns
- Repetition delay for sucessive SCI 64 byte packets : 370 ns
- Repetition delay for successive VME-PCI accesses: 700 ns

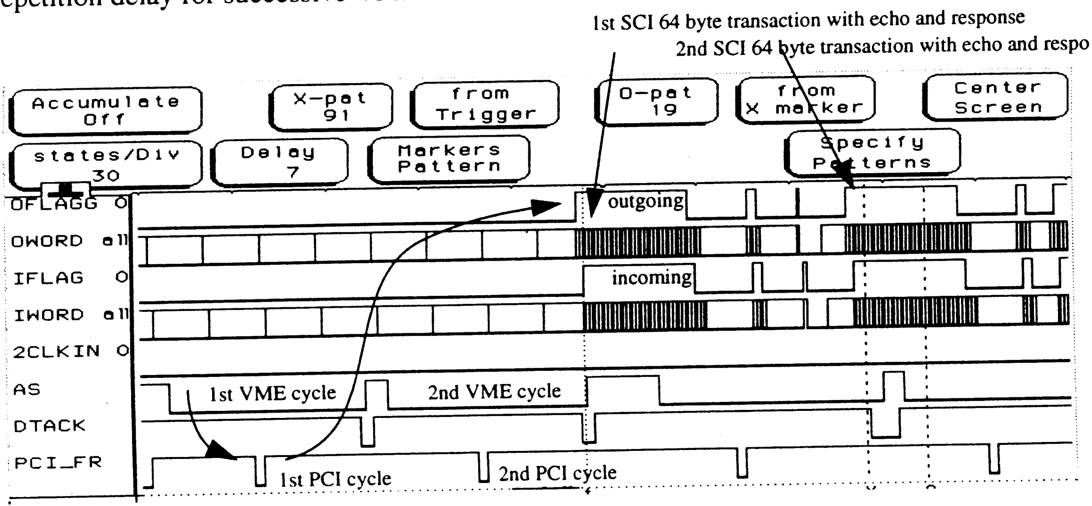


Figure 17: Timing for VME-PCI-SCI

Interpretation: These timing figures show that even using a fast PowerPC, a fast VME-PCI bridge chip and with no other VME traffic, a loop generation of (32 bit) PCI cycles across VME takes approximately 700 ns, i.e results here in 5.7 Mbyte/s. This order of magnitude, typical for VME, will limit crate extension technologies even with the fastest links like SCI.

The timings also show that preloaded SCI 64 byte dmove operations in the SCI adapter can be generated instantaneously every 370 ns, resulting in a peak bandwidth of 172 Mbyte/s on SCI.

The use of SCI for VME readout, appearing like a mismatch must however be seen as a possibility to connect a number of (relatively slow) VME nodes via a high speed system which can accummulate the bandwidth from all connected VME subsystems, without saturation up SCI bandwidth. Interconnection of 10 VME-SCI nodes via a single SCI ring for example appears to be safe from overloading 200 Mbyte/s SCI ringlets.

5.2.1 CERN 9U VME-PCI bridge

The use of VME 9U crates reduces the number of inter crate interconnects needed for a multicrate system. A 9U VME-SCI bride module with 4 PMC carriers has been built at CERN in order to evaluate this new environment. Based on single chip bridge (Tundra's Universe) this module is under evaluation with software developed by us under Lynx-OS.

5.2.2 Crate -to Crate communication via SCI

For a crate-to crate communication we used a powerPC processor mdule from CES (RIO-II) equipped with a PCI-SCI adapter from Dolphin to address a second crate via an PCI-SCI adapter in the MIDAS VME bridge (see previous paragraph).

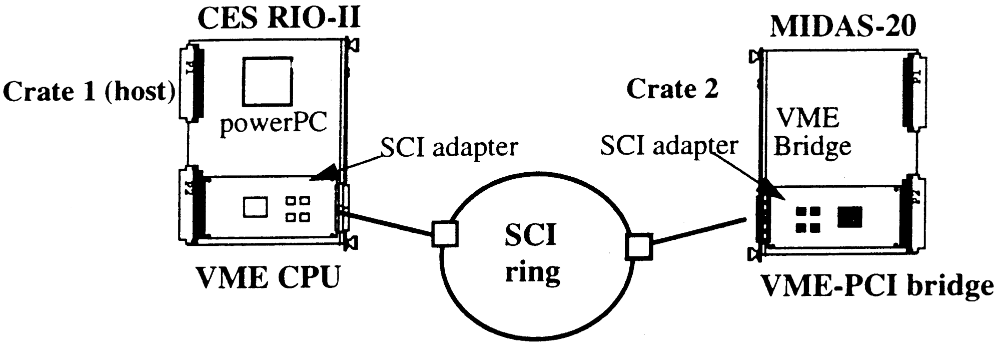


Figure 18: Crate-crate extension via SCI

(results awaited)

5.3 DMA SCI transfers between heterogeneous VME and PC systems

Two VME processors and two desktop PC's have been interconnected in an SCI ring via the CERN PCI-SCI adapter. DMA transfers from three of these SCI nodes have been sent to one destination for bandwidth accmmulation tests

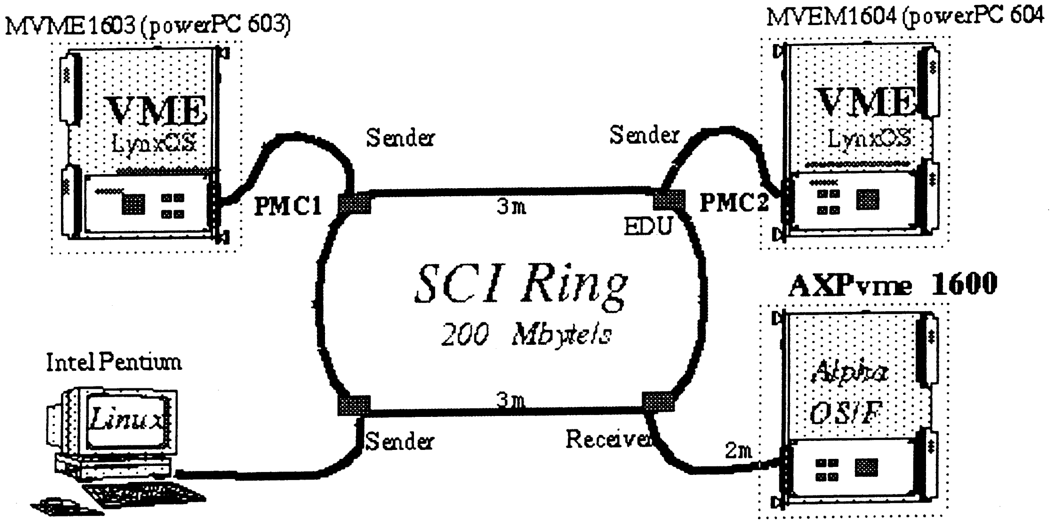


Figure 19: Heterogeneous 4-Node SCI interconnect test

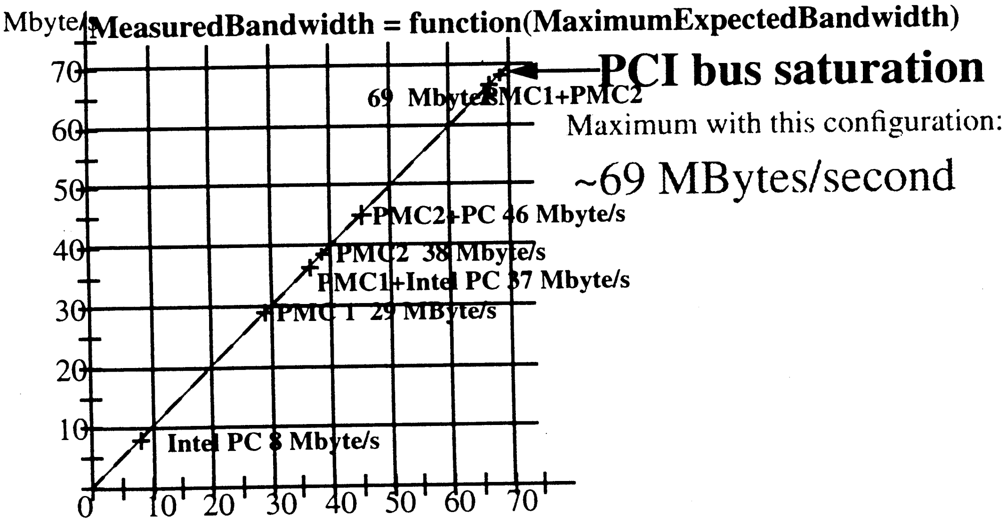


Figure 20: Transfer measurement in 4-node SCI

Three nodes were used as sender (64 byte transactions) to one destination.The Intel PC achieved due to non-optimized Linux driver only 8 Mbyte/s, on the PowerPC VME modules under Lynx OS higher sender rates were achieved: 29 Mbyte/s on a 603 processor and 38 Mbyte/s on a 604 processor. All three sources were enabled in combinations, the received data rate at the destination, an Alpha processor is proportional to the sum of the sender rates up to a maximum bandwidth, wich appeared as the PCI bus limit on the Alpha processor's PCI bus.

5.4 Conclusions with preliminary results

Transparent memory to memory access over SCI has been demonstrated via PCI-SCI bus adapters. The transfer rates achieved are limited by the PCI bus performance at values below 80 Mbyte/s. The access of VME via PCI bridges furher reduces the bandwidth to VME rates. Multiple VME crates can however be connected into a single SCI ringlet, allowing to accumulate the traffice from a dozend of VMEcrates in an SCI ring, maintaining the memory-to memory access. Within this scheme, crate extension techniques may further use the PCI-SCI technology to make economical use of multiple crates

Chapter 6. SCI test activities of LHC experiments

SCI specific programs within the LHC experiments are contimuing to further investigate performance and architectural possibilities provided by this technology in preparation of the availability of a large 16-way SCI switch for building demonstartors.

6.1 PCI-SCI evaluation at the University of OSLO (ALICE DAQ)

The PCI-SCI test configuration is picturally shown in Figure 18. The two PCs are interconnected through Dolphin PCI-SCI Adapters are 133 MHz Pentium processors with the Intel 430HX (TRITON 2) PCIset, running Windows NT. The NT driver for the PCI-SCI Adapter is a beta version from Dolphin ICS. Furthermore Dolphin supplied a test program; "tool.exe".

The PCs were either interconnected directly by means of an SCI station cable, or through a 4-switch from Dolphin. By means of a special tracer card developed by Dolphin, mounted on top of the switch card, any of the four SCI input links and any of the four SCI output links can be selected for analysis on an HP-1660 Analyzer.The installation of the PCI-SCI Adapters and the Windows NT driver was straightforward.

One of the PCs is also equipped with a PBT-315 PCI Bus Analyzer card from VMETRO. The BusView software runs on a separate PC under Windows 95.

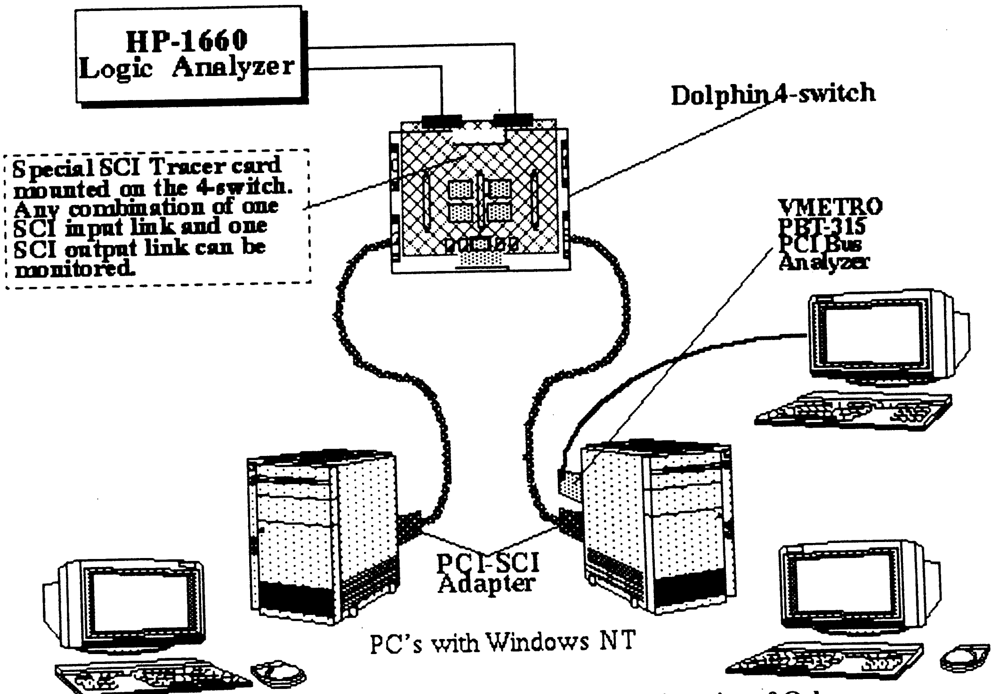

HP-1660
Logic Analyzer
Dolphin 4-switch
Special SCI Tracer card mounted on the 4-switch. Any combination of one SCI input link and one SCI output link can be monitored.
VMETRO PBT-315 PCI Bus Analyzer
PCI-SCI Adapter
PC's with Windows NT

Figure 21: PCI-SCI Test configuration at the University of Oslo

6.1.1 The "tool.exe" test program

The "tool.exe" test programs has a suite of commands for setting up the PCI-SCI Adapter, and for sending packets between the interconnected machines.There exists several versions of "tool.exe". With the latest version the data part of the SCI packet will be in little endian representation. Big endian representation can be obtained by byte swapping in software, this consumes a significant amount of CPU-time and slow down the effective transfer rate. The nodeIds of the local and remote PCI-SCI Adapters are ff32 and ff86, respectively

Measurements:

The "tool" program provides facilities for two-node testing, i.e. reading/writing in a selected remote node. By measuring the time it takes to execute a number of cycles times the block size, the transfer rate can be calculated. Since the elapsed time also includes the loop software overhead this type of benchmark is (probably) only meaningful for large blocks.

"test-rwrite" and "test--read"benchmark

Transfer rate measured for 1, 2 and 4 buffers, and without and without byte swapping, i.e. little and big endian data representaion, respectively. The result are shown in Figure 20.

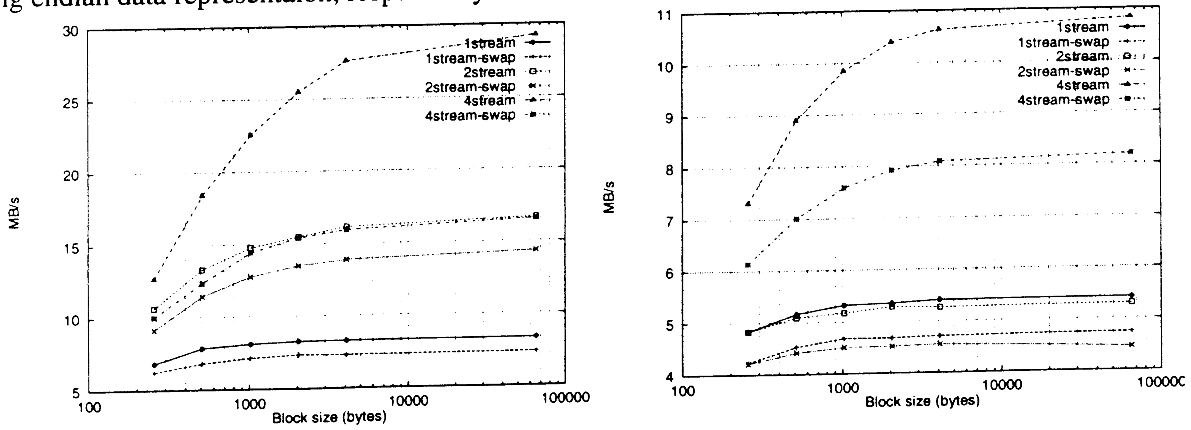


Figure 22: Transfer rates with "test-rwrite" and test-rread" of TOOL.EXE

6.2 SCI over OptoBus^{TM} (Dolphin, Motorola, University of Oslo)

Parallell optical interconnections can overcome many problems encountered in copper based solutions, such as bulky cabling, radiation problems (EMI) and crosstalk.

The OptoBus^{TM} optical link from Motorola Inc. is a 10-bit wide bi-directional data interconnect solution for point-to-point applications. The optical link consists of two identical transceiver modules and a detachable 10-fiber ribbon cable as schematically illustrated in Figure 21 on. In the RD24 Status report 1995 the first results on SCI over OptoBus were presented. By means of a MUX/DMUX transceiver between the 18-bit SCI on copper and the OptoBus 10-fiber ribbon, SCI trafficc at 200 MB/s over 10 m of optical cable was demonstrated.

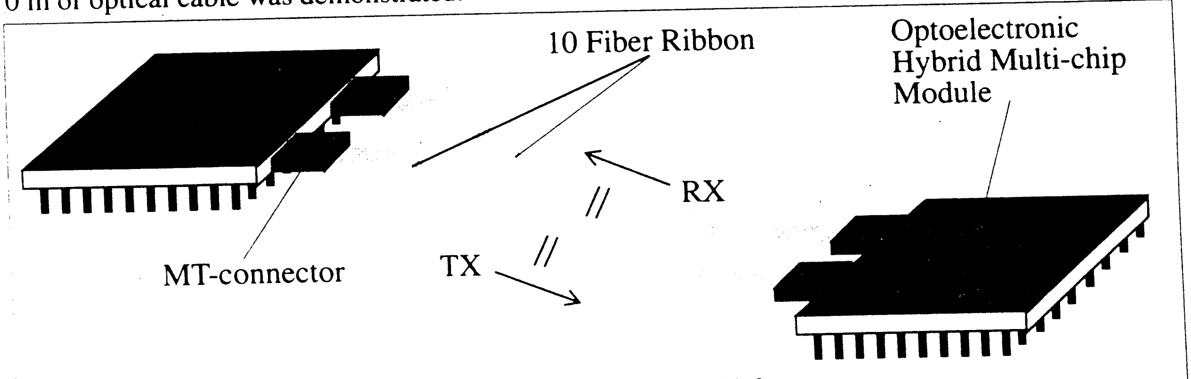


Figure 23: The OptoBus Link

The first prototype of Optobus used in the 1995 tests was rated at 150 Mbit/s per channel over 30 meter ribbon cable. (The measured 200 MB/s SCI traffic was in fact higher than the Optobus specs).

A new version of the Optobus chip, OptoBus I, has recently been released. It offers a bit rate per channel up to 400 Mb/s over a distance of 300 meters. An improved ribbon connector of a new push-pull type (Alcoa Telecommunication) ensures a robust connection of the ribbon.

An enhanced SCI-Optobus MUX/DEMUX transceiver card which supports the full Optobus speed, i.e. for an SCI link speed of 400 MB/s, has been developed at the University of Oslo, see . It is now in production.

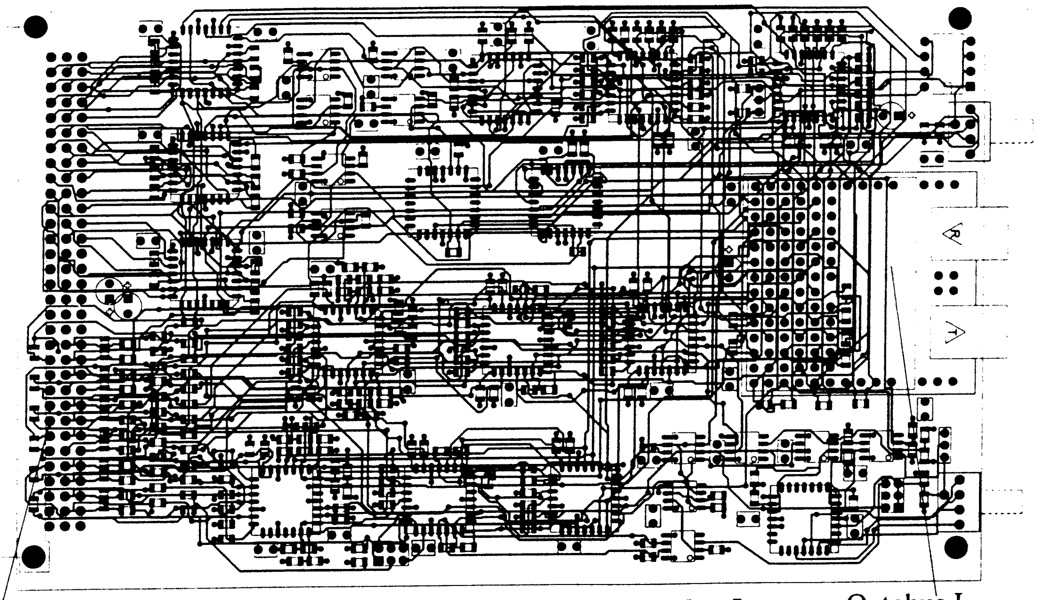


Figure 24: SCI MUX/DEMUX for Optobus I

6.3 SCI to FibreChannel Bridge (University of Oslo)

The application areas for various types of communication protocols are shown in Figure 12 on page 16.. The feasability of a bridge between SCI and FibreChannel has been studied by building a prototype card (Figure 26)..

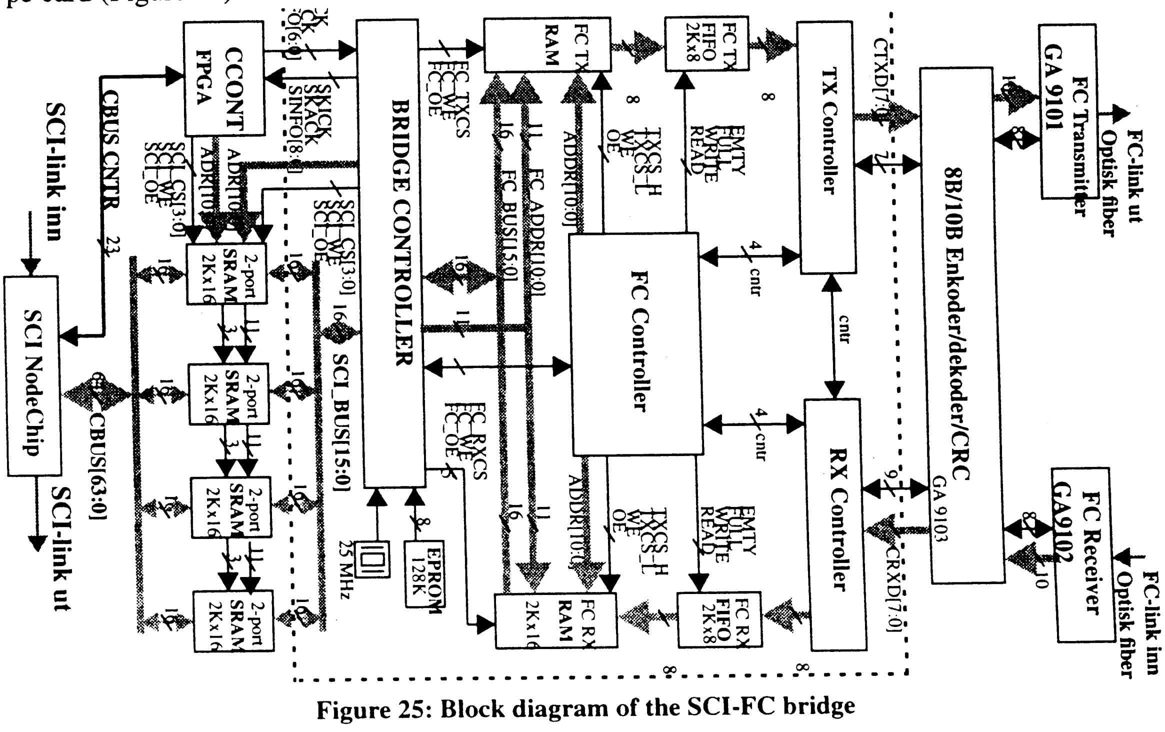


Figure 25: Block diagram of the SCI-FC bridge

The SCI-FC bridge supports SCI **readsb**, **writesb** and **move64** packets. An SCI packet is translated into a FibreChannel frame with the same payload. The long FibreChannel frames, with up up to 2112 bytes of payload, are split into a number of SCI **move64** packets. An FC frame is transferred to SCI as 33 move64 packets in total. The time between each move64 packet is 3.968 µs, which gives a raw SCI bandwidth of 20 MB/s (64 data + 16 header bytes).

6.4 .ATLAS - Digital joint project.

ATLAS and Digital are planning a joint project where a subfarm based on Alpha systems connected by Memory Channel will be stress-tested in a demanding data acquisition environment. The traffic shaping in and out of the subfarm will be modeled according to the expected traffic in an ATLAS second level trigger feature extractor, or global computing node, or supervisor. One of the nodes will act as switch-subfarm interface. This node will have an SCI interface as well as an ATM one. The simulated traffic will go in and out the subfarm through one of these interfaces. The project is expected to report if subfarms are acceptable building blocks of the ATLAS second level trigger.

6.5 Evaluation of a Switch Fabric by ATLAS

New demands in architecture, performance and implementation are appearing across the entire spectrum of computer systems. SCI is an attractive candidate to deal also with *data intensive applications (e.g.* HEP), not only in tightly coupled systems (massively parallel processing) but also for loosely coupled systems and I/O.

However, a single SCI-ring does not scale indefinitely [Ref.[31]]. Scalability of switch based systems has been investigated in [Ref.[31]]-[Ref.[35]]. An SCI switch consists of at least two SCI node-like interfaces (ports) to different rings, with a routing mechanism. There are two reasons that justify the construction of more powerful switch elements: *cost* and *latency* ,both depend on the number of intermediate rings. A system that minimizes the number of ring changes (hops) might provide better performance. We consider a switch structure that uses several internal links in parallel, to increase the throughput.

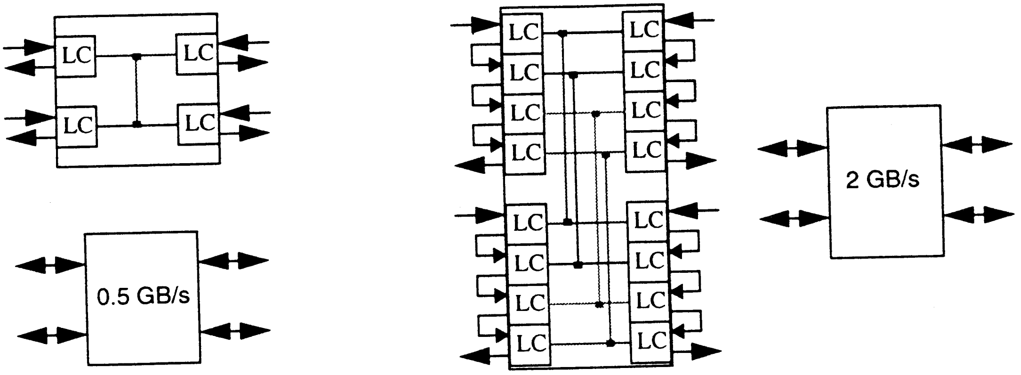


Figure 26: 4-port switch board with one (left) and four (right)

The routing information may either be derived from the contents of the packet or from switch properties (e.g. FIFO utilization). In our model, we use the transaction identifier, which is contained in each packet that traverses the internal link.

6.6 Simulation results

Our SCILAB-model uses 500 MB/s for both the SCI-ring and the internal link. All other parameters for the Link Controller (LC-2) are taken from Dolphin's Verilog simulations (40 ns bypassdelay, 70 ns internal link to SCI delay, 120 ns SCI to internal link delay). Propagation delays (~4 ns/m) are considered. Random datagenerators send request packets to each other.

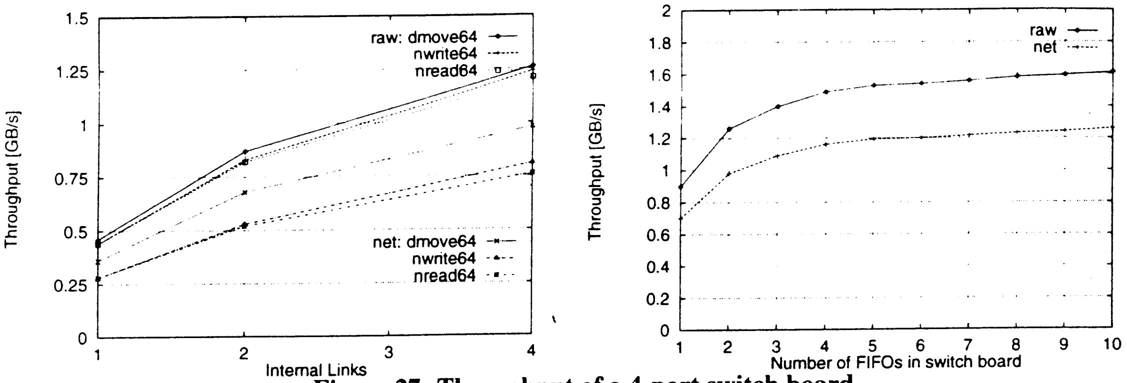


Figure 27: Throughput of a 4-port switch board

6.6.1 Interconnecting multiple Switches

There is an upper limit for the number of internal links and therefore the number of ports of a single switch board. To build a larger switch, it is necessary to interconnect several switch boards. The configuration for each application can be optimized by choosing the appropriate topology, such as multistage interconnection networks or n-dimensional meshes. In principle, ordinary switch ports can be used for interconnecting boards into larger switches. Alternatively, it is possible to add "expansion ports".

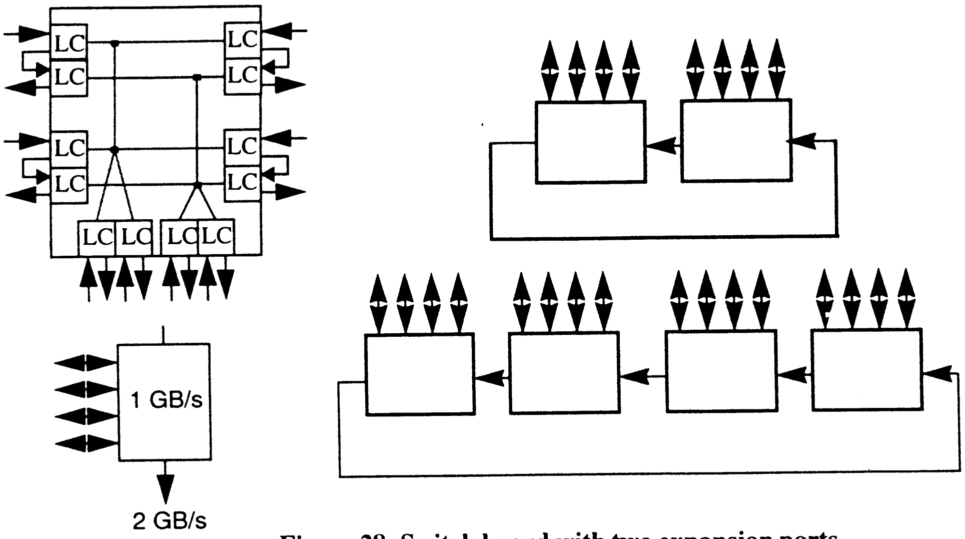


Figure 28: Switch board with two expansion ports

We have investigated the performance of the responseless 64-byte move transaction. The following diagrams show the raw/net-throughput of a switch, built of one, two and four boards with 4, 8 and 16 ports. A various number of expansion ports per internal link (one, two and four) is considered.

Especially for larger switches, the throughput increases significantly with the number of expansion ports. The gain in performance is obtained from additional SCI rings, as well as supplementary FIFO space.

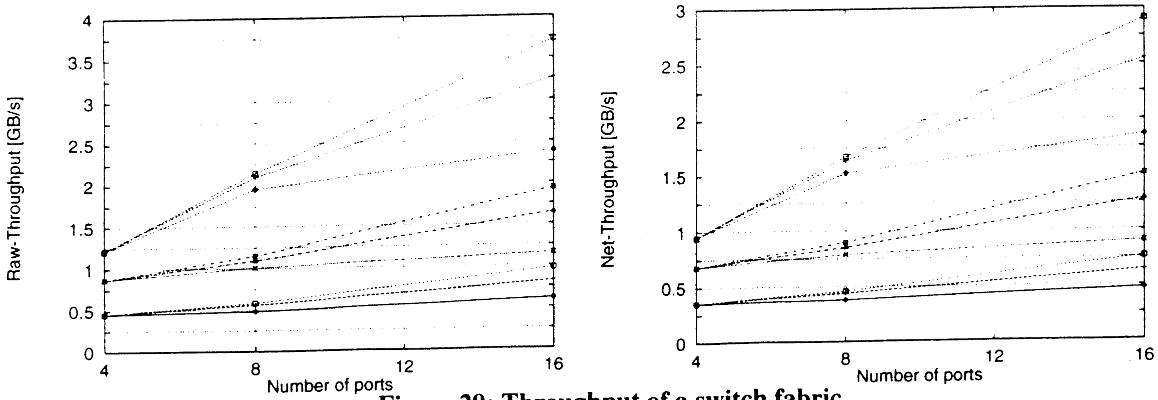


Figure 29: Throughput of a switch fabric

6.6.2 Summary

Simulation results indicate, that a large switch (both, in throughput and number of ports) can be obtained using multiple internal links of relative modest performance.

6.7 ATLAS TILE Cal readout via SCI ring (University of Valencia)

The ATLAS DAQ system requires an interface between the front-end electronics for detectors and the global read out system [25]. This interface includes several functional modules: derandomizer buffers, front-end links, RODs and read out links. In the University of Valencia and the Polytechnic University of Valencia we are working on the design and implementation of a Read Out Driver module (ROD) for the Hadronic Calorimeter (Tile-Cal detector) of ATLAS [26], which includes an SCI auxiliary port to operate in stand-alone mode and for calibration purposes.

At the maximum LHC luminosity the average first level trigger accept rate (L1A) will be 100 KHz In order to have a constant rate at the second level trigger and maintain a reasonable second level buffer size, derandomizer buffers are necessary which are connected to a ROD through optical paths, namely front-end links. RODs perform data multiplexing according to the "Region of Interest" (RoI) segmentation, pre-processing (if necessary), error checking/recovery and transmission through optical paths, namely read out links, to the second level buffers (ROB). ROD should operate in a nearly transparent mode, i.e., without introducing any considerable dead-time to the second level trigger.

6.7.1 TILE-CAL ROD PROTOTYPE IMPLEMENTATION BLOCK DIAGRAM

The figure below shows a block diagram of our Tile-Cal ROD prototype implementation.

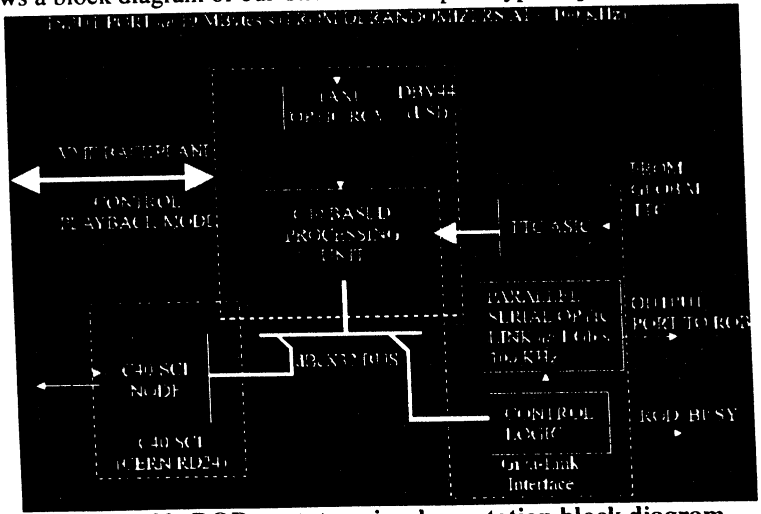


Figure 30: ROD prototype implementation block diagram

The aim of this first prototype is to demonstrate some ROD functionalities according to ATLAS acteristic with VME bus for HEP applications and bus-based SMP systems for commercial applications. This allows re-use of software initially developed for such systems. Apart from high bandwidth this appoach leads also to bus-like low latencies obtained for even very short messages.

Chapter 8. References

[1] **Application of the Scalable Coherent Interface for Data Acquisition at LHC**, *RD24 Collaboration*, CERN/DRDC 91-45, Proposal P33(ftp rd24.cern.ch or http://www1.cern.ch/RD24)

[2] **Application of the Scalable Coherent Interface for Data Acquisition at LHC**, *RD24 Collaboration*, CERN/DRDC 93-20, Status Report May 1993 (ftp rd24.cern.ch or WWW as above)

[3] **Application of the Scalable Coherent Interface for Data Acquisition at LHC**, *RD24 Collaboration*, CERN/DRDC 94-23, Status Report May 1994 (ftp rd24.cern.ch or WWW as above)

[4] **Application of the Scalable Coherent Interface for Data Acquisition at LHC**, *RD24 Collaboration*, CERN/DRDC 95-52, Status Report August 1995(ftp rd24.cern.ch or WWW as above)

[5] **IEEE Standard for Scalable Coherent Interface (SCI)**, IEEE/ANSI Std 1596-1992

[6] **Proceedings of the Large Hadron Collider Workshop**, CERN 90-10 ECFA 90-133, December 1990, VOL III, p. 160-169

[7] **BER Measurements on LVDS based cable Transmission**, J.F. Loennum, SINTEF Electronics, Trondheim, presented at the European SCI Meeting, Valencia June 1996 (http://sunshine.cern.ch:8080/ESONE+IEC/Valencia-96/talks/index.html)

[8] **New cabling system**, I.Birkeli Dolphin, presented at the European SCI Meeting, Valencia June 1996 (http://sunshine.cern.ch:8080/ESONE+IEC/Valencia-96/talks/index.html)

[9] **Motorola Application Note AN1572**, J. Grula, Motorola: Applying the OPTOBUS I Multichannel optical Data Link to High Performance Communication Systems.

[10] **OPTOBUS**, Presentation by A.Guenther, presented at the European SCI Meeting, Valencia June 1996 (http://sunshine.cern.ch:8080/ESONE+IEC/Valencia-96/talks/index.html)

[11] **Parallel optical Link for SCI/LVDS**, K. Aretz, Siemens AG Berlin, presented at the European SCI Meeting, Valencia June 1996 (http://sunshine.cern.ch:8080/ESONE+IEC/Valencia-96/talks/index.html)

[12] **Large Hadron Collider Committee CERN/LHCC 95-55**, Report from the LCRB committee

[13] **Hyper-LVDS: An All CMOS Solution to a Bipolar Challenge**, D.Burrows, LSI Logic Bracknell,UK, Valencia June 1996 (http://sunshine.cern.ch:8080/ESONE+IEC/Valencia-96/talks/index.html)

[14] **BLINK Interface Specification**, Dolphin Interconnect Solutions,N-0621 OSLO Bogerud, P.O. Box 52

[15] **SCI LincChip Users Guide**,(available only on ND agreement) Interconnect Systems Solution, 26215 Camino Adelanto, Mission Viejo, CA 92691-3245, USA

[16] **Draft Standard for a Common Mezzanine Card Family: CMC** P1386/Draft 1.6 and
Draft Standard Physical and Environmental Layers for PCI Mezzanine cards: PMC, P1386.1, Draft 1.6
IEEE Std. Dep., P.O.Box 1331,Piscataway, NJ 08855-1331 USA

[17] **SCI activities at the Technical University of Munich and at Max Planck Institute for Plasma Physics Garching**, H.Richter et al, talk at the European SCI meeting Valencia June 1996 (http://sunshine.cern.ch:8080/ESONE+IEC/Valencia-96/talks/index.html)

[18] **Proposal for an SCI-API**, Hein Rohrig, Kolja Sulimma, Dieter Rohrich,

R.Stock, available on the WEB under http://www.vsb.cs.uni-frankfurt.de/~roehrig/if.ps

[19] **Experiences from programming CERN's RD24 PCI-SCI bridge under Linux**, Hein Rohrig,Univ. o. Frankfurt, Talk given at the European SCI meeting Valencia June 1996, see WEB:http://sunshine.cern.ch:8080/ESONE+IEC/Valencia-96/talks/index.html)

[20] **Design of SCI-class Interconnects**, *W. Nation, Proc.of the Int.DAQ Conf.on Event Building and Data Readout*,Fermilab,Batavia,Il,Oct.26-28 1994

[21] **A low cost CMOS 500 Mbyte/s SCI link Controller**, P. Gustad, Dolphin, presented at the Hot Interconnects Symposium IV, Stanford University, August 15-17

[22] **An SCI Interconnect Chipset and Adapter**, R.Clark and K.Alnes Data General Corporation, presented at the Hot Interconnects Symposium IV, Stanford University, August 15-17

[23] **PCI to PCI Bridge Architecture Specification**, *Rev 1.0 PCI Special Interest Group,P.O.Box 14070,Portland, OR 97214, USA*

[24] **PCI Local Bus Specification,** *Revision 2.1 June 1, 1995, PCI Special Interest Group, P.O. Box 14070, Portland, OR 97214, USA*

[25] **Optimized Reconfigurable Cell Array (ORCA)**, *AT&T Microelectronics, Dept-500404200, 555 Union Blvd, Allentown, PA 18103, USA*

[26] **Global architecture for the ATLAS DAQ and Trigger**. MAPELLI L.,ATLAS DAQ-NO-22 (1995).

[27] Application of a real-time data processing VME module using a custom high-speed I/O multiport board to "ATLAS Tile-Cal Read Out Driver". GONZALEZ, V., SANCHIS, E., CERN TILECAL Note 56 (1995)

[28] **Trigger and DAQ interfaces with Front-end Systems: Requirement document**. FARTHOUAT, PH. and LEDU, P.,CERN ATLAS DAQ Note. Draft 1.4 (1995).

[29] **Design and evaluation of a DSP interface to SCI**.J. FERRER et al.,ESONE Real-Time Data '95 Conference. Warsaw, Poland (1995).

[30] **Evaluación y desarrollo del interface DSP320C40-SCI.** J. FERRER, Polytechnic University of Valencia, Spain (1994).

[31] S. Scott, J. Goodman, M. Vernon, "Performance of the SCI Ring", Proceedings IEEE ISCA 92, Queensland, May 1992

[32] R. Johnson, "Extending The Scalable Coherent Interface For Large-Scale Shared-Memory Multiprocessors", Ph.D. Thesis at the University of Wisconsin-Madison, 1993

[33] B. Wu, A. Bogaerts, B. Skaali, "A Study of Switch Models for the Scalable Coherent Interface", Proceedings of the Sixth IFIP WG6.3 Conference on Performance of Computer Networks, Istanbul, 1995

[34] **VHDL design of PCI Mezzanines (PMC) for a PCI-SCI bridge,** *C.Fernandes, H.Muller,L.McCulloc,V.Lindenstruth,Y.Ermoline, to be presented at "First Workshop on Electronics for LHC experiments, LISBON Sept 11-15*

[35] **A millenium approach to Data Acquisition: PCI and SCI**, *H.Muller, C.Fernandes, Y.Ermoline,V.Lindenstruth, to be presented at CHEP-95, Rio de Janeiro, September 18-22, 1995*

[36] ATLAS Collaboration, "**Technical Proposal**", *CERN/LHCC 94-43 LHCC/P2 CERN, Geneva, Switzerland 15 Dec. 1994*