A millenium approach to Data Acquisition: SCI and PCI

Hans Müller, A.Bogaerts

CERN, Division ECP, CH-1211 Geneva 23, Switzerland

V.Lindenstruth

LBL Berkeley, CA 94720 Berkeley, USA

The international SCI standard IEEE/ANSI 1596[a] [Ref. 1.] is on its way to become the computer interconnect of the year 2000 since for a first time, low latency desktop multiprocessing and cluster computing can be implemented at low cost. The PCI bus is todays's dominating local bus extension for all major computer platforms as well as for buses like VMEbus. PCI is a self configuring memory and I/O system for peripheral components with a hierarchical architecture. SCI is a scalable, bus-like interconnect for distributed processors and memories. It allows for optionally coherent data caching and assures errorfree data delivery. First measurement with commercial SCI products (SBUS-SCI) confirm simulations that SCI can handle even the highest data rates of LHC experiments. The eventbuilder layer for a millenium very high rate DAQ system can therefore be viewed as a SCI network (bridges, cables & switches) interfaced between PCI buses on the front-end (VME[b]) side and on the processor farm (Multi-CPU) side . Such a combination of SCI and PCI enables PCI-PCI memory access, transparently across SCI. It also allows for a novel, low level trigger technique: the trigger algorithm can access VME data buffers with bus-like latencies like local memory, i.e. full data transfers become redundent.. The first prototype of a PCI-SCI bridge for DAQ is presented as starting point for a test system with built-in scalability.

1 Overview

SCI has found acceptance in a large sector of the computer industry with announcements by Data General, Siemens/Nixdorf, Sequent, Unisys, AT&T, ICL, Intel, and IBM. Apart from specialized components (GaAs chips like in the Convex MPP machine) a number of cheaper CMOS chips, boards and switches are becoming available. The speed of SCI CMOS chips is moving from 200 Mbyte/s towards 1 Gbyte/s.

The dominance of PCI as the universal "open" computer bus with autoconfiguration (plug&play), both BIOS and PREP compliance (PC and PowerPC architecture) and support for bus hierarchies (multiport data access) is manifested by today's overall availability of computers with PCI extensions from companies like Intel (Pentium), IBM/Apple (PowerPC), DEC (Alpha), Motorola (PowerPC on VME board) and many more.

Coming multi-processor systems, such as multple P6 processors in a SMP desktop use SCI as low latency CPU interconnect, and PCI as peripheral memory bus. Cluster computing is therefore becoming an economic choice by using PCI-SCI bridges as plug-in cards and SCI as the desktop interconnect.

---

a. Scalable Coherent Interface, also to appear as ISO/IEC standard IS 13961 (approved Sept 94)
b. VMEbus modules equipped with PMC mezzanine card extension.

1.1 SCI in DAQ systems

The design of the DAQ systems for the Large Hadron Collider (LHC) is challenged by factors of 100 in size and data rates if compared to LEP. Apart from a new bandwidth requirement, a consequence is the need for scaling, i.e. the interconnect must not saturate when its size is increased from a testbeam size to the full size of LHC. The shared bus solutions of LEP (VME & Fastbus) are inherently limiting the overall bandwith gain when more and more devices are added. SCI's built-in scaling capabilities have been verified by simulations[a] up to tens of Gbyte/s for large, eventbuilder sized, multistage switch models [Ref. 2.] . Latency, i.e. the delays for handshaken transfer of data from a producer to a consumer, i.e. across an eventbuilder switch, should be minimized for obvious reasons: eliminate buffer overflows and allow for use of fast-access static memories at much higher data security levels than possible with dynamic memories[b]. The latency of a switch stage is today already as low as 0.7 us. Not surprisingly the RD24 R&D project measured with first generation SCI switches dataless eventbuilder rates in the order of 50 KHz. Event synchronization, needed particularly for very high rate push architectures is possible via SCI without the need for specially tuned parallel networks: SCI provides computer instructions, like fetch&add, to be used for indivisible, global operations (global counters etc.). Both high rate push architectures (2nd level) and memory-like read architectures (3rd level) may coexist, due the split transactions in a buslike interconnect without top or bottom.

1.2 Understanding SCI

SCI eliminates the bus loading by using well terminated, parallel, point-to-point links running at very high frequencies (up to 500 MHz). These links interconnect neighboring SCI nodes in rings [Figure 1].

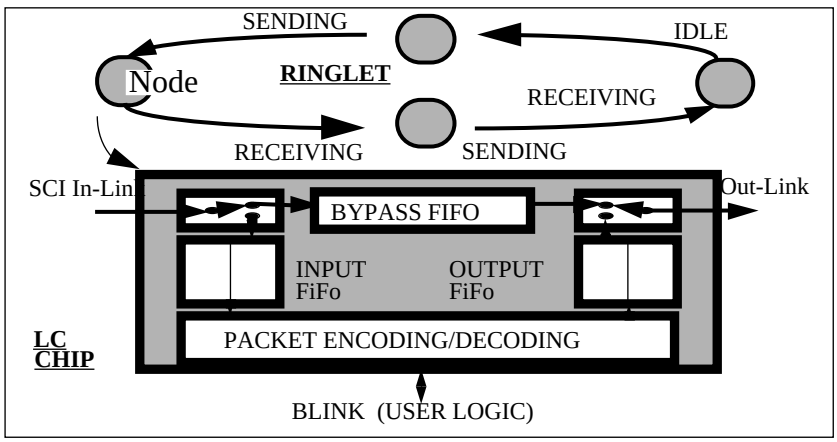

Figure 1: SCI ringlet and details of a node

a. MODSIM-II simulations by RD24 project at CERN

b. DRAMS with increasing density are proportionally prone to soft errors from radiation

Larger SCI systems consist of many ringlets, interconnected via switches. The scaling limit for each ringlet is about a factor of two of its raw link bandwith. Switch components consist of 2 *2 [a]switches , which scale up to nearly the sum of bandwith of four ringlets. Larger switches are built as multistage network from interconnected 2*2 switches. For example, an 8 * 8 multistage switch requires twelve 2*2 components. Using the 500 Mbyte LC-II technology from Dolphin [Ref. 4.] , such a switch provides 1.6 Gbyte/s thoughput with less than 2 us latency across the fabric. Each of its 8 input or output ringlets may interconnect a number (assume 10) of producer and/or consumer nodes. Effectively this fabric then provides an equivalent of an 80 * 80 point-point switch.

Each node may transmit multiple read or write requests in form of short packets (up to 64 bytes), inserted “elastically” by the SCI linc controller’s link port in between other bypassing SCI packets. The split transaction (independent request & response packets) provides quasi-simultaneous sharing of SCI bandwith between nodes. The latencies for acknowledgments (echos) are as low as the ringlet roundtrip delay, typically a few hundred nanoseconds. Echo packets also include retry information in case of a busy destination.

SCI is like a virtual bus since bus access is quasi-simultaneous for all nodes and normally immediate. SCI allows for bus-to-bus transparency, i.e. CPU bus cycles can be converted by hardware to SCI packets and vice versa, i.e. the application does not need to know about SCI. First commercial bridge examples for SCI are SBUS-SCI and PCI-SCI. Shared memory applications which use these interface exhibit very low latencies since no software layers are needed. Due to this transparency, remote memory (i.e. a data producer) may be accessed from a CPU (a data consumer) much like local memory. The access delays over remote SCI connections are recoverable by using the caching options of SCI.

1.3 Industry acceptance of SCI

SCI is compliant with a set of technologies such as LVDS (low voltage differntial signalling), GLINK (GiGalink serial optical fibre transmission), OPTOBUS (parallel optical transmission technology) and others (RAMLINK, SYNCLINK,). Seen as a home standard for compliant technologies, the expected lifetime of SCI extends beyond the lifetimes of individual technologies.The first SCI computer, the Convex Exemplar holds the SpecRate fp record. Dolphin, a norwegian company and RD24 collaborator [Ref. 4.] , became the first developer of core SCI technology. Producing SCI chips and adapter cards, Dolphin provide the technology for the symmetric multiprocessor (SMP) servers, scalable I/O systems, and clusters of servers and workstations. Data General will create a new generation of SMP servers incorporating up to 100 P6 CPUs running UNIX and Windows NT. DG will use Intel’s quad-P6 boards with cache coherent, distributed shared memory and a PCI I/O channel, using SCI as an interconnect with ASICs from Dolphin. Siemens-Nixdorf (SNI) agreed to use Dolphin's PCI-SCI bridge chip and the SCI protocol engine to develop a scalable I/O system for a new multiprocessor Unix server systems. Dolphin's SCI protocol engine will be utilized in two ASICs developed by Siemens-Nixdorf. Several SCI desktop computers are expected to reach the market in 1996.

---

a. 2 ring * 2 ring

1.4 PCI, industries defacto Peripheral Component Interconnect

PCI is today's industry bus extension for both processor platforms and VMEbus modules. It supports both the BIOS and the PREP compliant computer worlds via a jumperless (plug&play) configuration convention. PCI's hierarchical bus features offer many possibilies, like multi-port data access, to be exploited by DAQ applications [Ref. 7.] . PCI performance is moving from its current 132 Mbyte/s raw bandwith to 264 Mbyte/s by doubling of its bus size. A final bandwith of 528 Mbyte/s can be achieved by doubling the clock frequency. Due to its reflected wave technology as required for CMOS chip interconnection, PCI is constrained to small distance. This locality can be overcome by bridging PCI to SCI, in particular by mapping SCI memory into PCI memory space in order to enable transparent hardware transactions from PCI to PCI, over SCI.

The importance of a PCI-SCI bridge was recognized by RD24 collaborators from

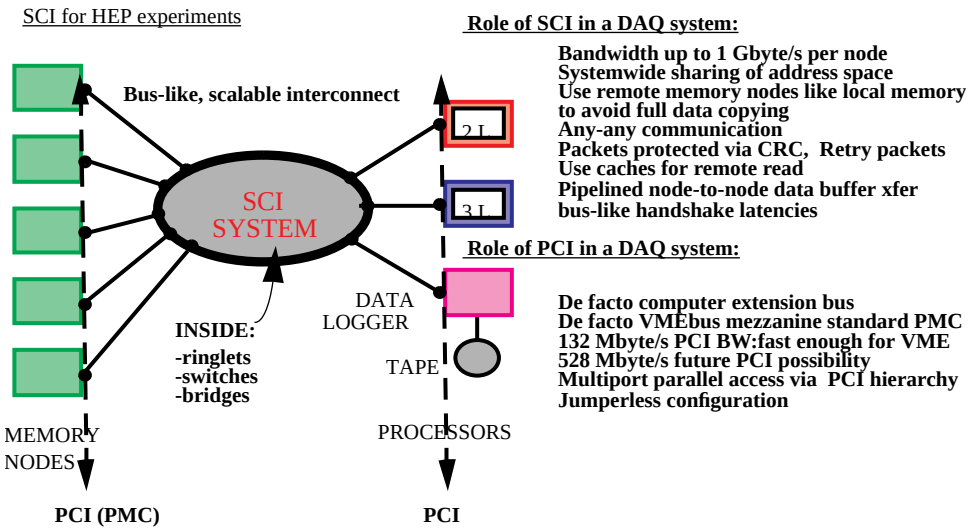

Figure 2: Global view of using PCI and SCI in DAQ

CERN and LBL [Ref. 8.] in 1994 and a design project was launched to specify and build a first PCI-SCI bridge [Ref. 9.] . The baseline idea was to interconnect DAQ memories to low level trigger processors across SCI. A list of advantageous features is shown in Figure 2. Such a bridge should allow to standardize[a] on PCI as interconnect for VME and CPU platforms and use SCI as scalable processor-memory interconnect in between. A first implementation, using very high density FPGA's and full VHDL design [Ref. 5.] is now under test at CERN and LBL. This PMC mezzanine [Figure 3] is intended for use on commercial VME boards. A second, electrically equivalent implementation is being prepared at CERN as PCI extension card for use in a BIOS or PREP compliant computers.

Unlike commercial PCI-SCI bridge projects, RD24 added options for data driven

a. PCI is however neither a national, nor international standard

applications in DAQ systems. For this purpose, a high rate chain mode DMA engine [Ref. 6.] for event gathering is implemented to enable multi-source data pushing from producers to a consumer. The DMA engine features a multi-master DMA control access, allowing the DMA host to be resident either on the SCI or the PCI side.

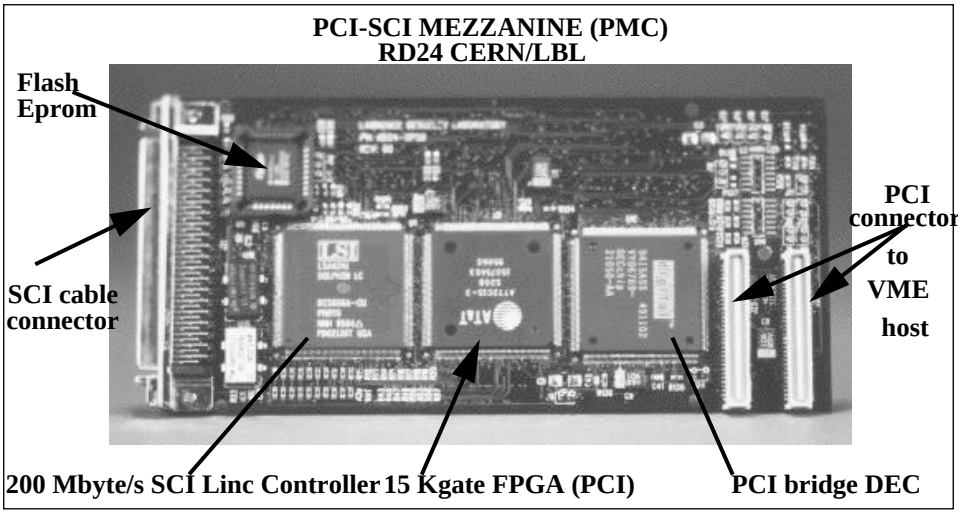

Figure 3: PCI-SCI bridge board for VME (PMC standard)

.

1.5 Combining SCI and PCI: a perfect match

SCI extends PCI's locality across distance, fully maintaining PCI bandwidth. A memory on a remote PCI segment can be addressed like local memory, provided the bridge

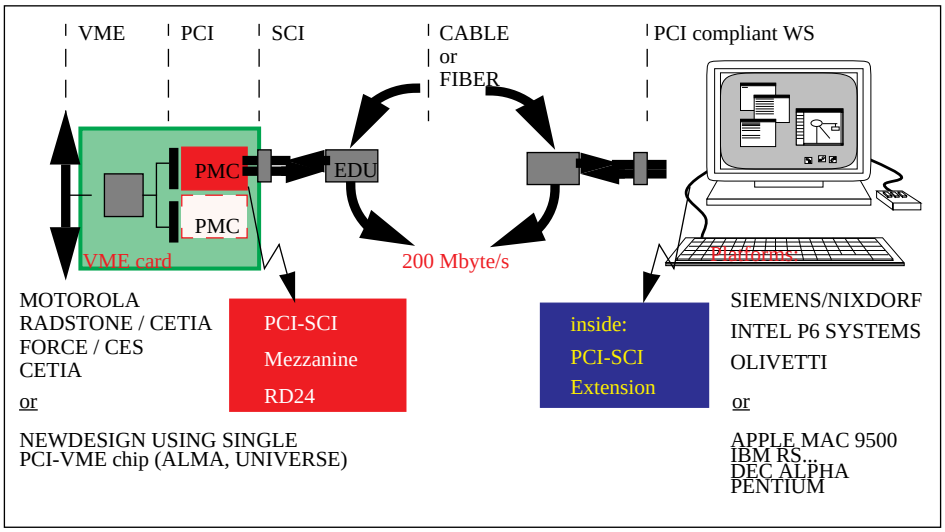

Figure 4: The RD24 PCI-SCI bridge (PMC mezzanine)

supports memory transparency. Both shared memory and message passing paradigms can be implemented and may co-exist. PCI provides a 64 bit configurable memory space which can be mapped into SCI. PCI's I/O transactions (byte operations) are adequate for accessing the configuration registers of SCI. For performance enhancement, PCI provides prefetchable memory, a mechanism to allow a bridge to merge several memory accesses into single bursts. An SCI-PCI bridge can use this to build 64 byte SCI packets from small sized PCI transactions.

1.6 SCI-PCI based VME test system

All necessary parts, i.e. SCI cable or fiber technology, cascadable SCI switches and the PCI-SCI bridge are now available to be integrated into a scalable user test system of low cost with SCI performance. Once a stable "single ringlet" VME-PC test environment is established, the PCI buses of any VME modules and personal computers may be connected to extend the system even to a switch -based SCI network, capable of merging concurrent traffic of many PCI buses.

RD24's work for 1996 is therefore concentrating on building a simple, single-ringlet test system as part of the generic research program of computing for LHC. The starting point is the PCI-SCI bridge technology which has been conceived to evolve by using the VHDL language to synthesize re-programmable hardware, i.e. the hardware description is loaded at powerup from a writable FlashEprom. Hardware feature updates, even for numerous bridges are possible within seconds. In a later stage, when all features are stable, the FPGA devices may be replaced by more cost effective ASICs.

References

1. **IEEE Standard for Scalable Coherent Interface (SCI),** IEEE Std 1596-1992
2. **RD24 status report to the LCRB committee**, September 1995, `CERN/LHCC 95-42,` RD24 collaboration, anonymous ftp: sunsci.cern.ch:/sci/RD24_Info/Status_95/report95.ps
3. P**CI Peripheral Component Interconnect Specification**, *Rev 2.1 PCI Special Interest Group, P.O.Box 14070, Portland, OR 97214, USA*
4. **Dolphin Interconnect Solutions**, OSLO-BOGERUD, PO box 70, Norway
5. **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 (see http://sunshine.cern.ch:8080/PCI)*
6. **A data moving engine for the Scalable Coherent Interface (SCI)**, Y.Ermoline, C.Fernandes, H.Muller, V.Lindenstruth, to be presented at OBS-95, Zurich 11-13 october 1995 (see http://sunshine.cern.ch:8080/DMA)
7. **PCI and PMC standards for DAQ**, *H.Muller, presented to the ALICE and CMS DAQ group, WWW http://sunshine.cern.ch:8080/PCI under keyword PMC mezzanines as PCI_ALICE.ps*
8. **The STAR Data Aquisition System**, V.Lindenstruth, *Proc. of the Int. DAQ Conf. on Event Building and Data Readout*, Fermilab, Batavia, Il, Oct. 26-28 1994
9. **A PCI-SCI bridge for high rate Data Acquisition Architectures at LHC**, H.Muller et al. PCI'95 Conference Proceedings p. 156-160, St.Clara, USA, March 27-31 1995.