INTERNATIONAL STANDARD

ISO/IEC 13961
IEEE Std 1596

First edition
2000-07

Information technology –

Scalable Coherent Interface (SCI)

Reference number
ISO/IEC 13961:2000(E)
IEEE Std 1596, 1998 Edition

Abstract: The scalable coherent interface (SCI) provides computer-bus-like services but, instead of a bus, uses a collection of fast point-to-point unidirectional links to provide the far higher throughput needed for high-performance multiprocessor systems. SCI supports distributed, shared memory with optional cache coherence for tightly coupled systems, and message-passing for loosely coupled systems. Initial SCI links are defined at 1 Gbyte/s (16-bit parallel) and 1 Gb/s (serial). For applications requiring modular packaging, an interchangeable module is specified along with connector and power. The packets and protocols that implement transactions are defined and their formal specification is provided in the form of computer programs. In addition to the usual read-and-write transactions, SCI supports efficient multiprocessor lock transactions. The distributed cache-coherence protocols are efficient and can recover from an arbitrary number of transmission failures. SCI protocols ensure forward progress despite multiprocessor conflicts (no deadlocks or starvation).

Keywords: bus architecture, bus standard, cache coherence, distributed memory, fiber optic, interconnect, I/O system, link, mesh, multiprocessor, network, packet protocol, ring, seamless distributed computer, shared memory, switch, transaction set.

The Institute of Electrical and Electronics Engineers, Inc.
345 East 47th Street, New York, NY 10017-2394, USA

ISBN 2-8318-5375-3

CONTENTS

Clause Page

Page

Page

Page

Page

Page

INFORMATION TECHNOLOGY –

SCALABLE COHERENT INTERFACE (SCI)

FOREWORD

1) ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

2) In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC1. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

3) Attention is drawn to the possibility that some of the elements of this International Standard may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

International Standard ISO/IEC 13961 was prepared by subcommittee 26: Microprocessor systems, of ISO/IEC joint technical committee 1: Information technology.

Annexes A and B are for information only.

IEEE Standards documents are developed within the Technical Committees of the IEEE Societies and the Standards Coordinating Committees of the IEEE Standards Board. Members of the committees serve voluntarily and without compensation. They are not necessarily members of the Institute.

The standards developed within IEEE represent a consensus of the broad expertise on the subject within the Institute as well as those activities outside of IEEE that have expressed an interest in participating in the development of the standard.

Use of an IEEE Standard is wholly voluntary. The existence of an IEEE Standard does not imply that there are no other ways to produce, test, measure, purchase, market, or provide other goods and services related to the scope of the IEEE Standard. Furthermore, the viewpoint expressed at the time a standard is approved and issued is subject to change brought about through developments in the state of the art and comments received from users of the standard. Every IEEE Standard is subjected to review at least every five years for revision or reaffirmation. When a document is more than five years old and has not been reaffirmed, it is reasonable to conclude that its contents, although still of some value, do not wholly reflect the present state of the art. Users are cautioned to check to determine that they have the latest edition of any IEEE Standard.

Comments for revision of IEEE Standards are welcome from any interested party, regardless of membership affiliation with IEEE. Suggestions for changes in documents should be in the form of a proposed change of text, together with appropriate supporting comments.

Interpretations: Occasionally questions may arise regarding the meaning of portions of standards as they relate to specific applications. When the need for interpretations is brought to the attention of IEEE, the Institute will initiate action to prepare appropriate responses. Since IEEE Standards represent a consensus of all concerned interests, it is important to ensure that any interpretation has also received the concurrence of a balance of interests. For this reason IEEE and the members of its technical committees are not able to provide an instant response to interpretation requests except in those cases where the matter has previously received formal consideration.

Comments on standards and requests for interpretations should be addressed to:

Secretary, IEEE Standards Board
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
USA

IEEE Standards documents are adopted by the Institute of Electrical and Electronics Engineers without regard to whether their adoption may involve patents on articles, materials, or processes. Such adoption does not assume any liability to any patent owner, nor does it assume any obligation whatever to parties adopting the standards documents.

Foreword to IEEE Std 1596, 1998 Edition

[This foreword is not a part of ISO/IEC 13961:2000, Information technology – Scalable Coherent Interface (SCI).]

The demand for more processing power continues to increase, and apparently has no limit. One can usefully saturate the resources of any computer so easily by merely specifying a finer mesh or higher resolution for the solution of some physical problem (hydrodynamics, for example), that engineers and scientists are desperate for enormously larger computers.

To get this kind of computing power, it seems necessary to use a large number of processors cooperatively. Because of the propagation delays introduced when signals cross chip boundaries, the fastest uniprocessor may be on one chip before long. Pipelining and similar large-mainframe tricks are already used extensively on single-chip processors. Vector processors help, but are hard to use efficiently in many applications. Multiprocessors communicating by message passing work well for some applications, but not for all. The shared-memory multiprocessor looks like the best strategy for the future, but a great deal of work will be needed to develop software to use it efficiently.

It is important to support both the shared-memory and the message-passing models efficiently (and at the same time) in order to support optimal software for a wide range of problems, especially for a system that dynamically allocates processors and perhaps changes its configuration depending on the nature of its load.

SCI started from an attempt to increase the bandwidth of a backplane bus past the limits set by backplane physics in order to meet the needs of new generations of processor chips, some of which can single-handedly saturate the fastest buses. We soon learned that we had to abandon the bus structure to achieve our goals.

Backplane performance is limited by physics (distributed capacitances and the speed of light) and by a bus's one-at-a-time nature, an inherent bottleneck. To gain performance far beyond what buses and backplanes can do, one needs better signaling techniques and the concurrent use of many signaling paths.

Rather than using bused backplane wires, SCI is based on point-to-point interconnect technology. This design approach eliminates many of the physics problems and results in much higher speeds. SCI in effect simulates a bus, providing the bus services one expects (and more) without using buses.

Committee Membership

The specification has been developed with the combined efforts of many volunteers. The following is a list of those who were members of the Working Group while the draft and final specification were compiled:

David B. Gustavson, Chair

David V. James, Vice Chair

Nagi Aboulenein
Knut Alnes
Robert H. Appleby
Kurt Baty
Amir Behroozi
David L. Black
Andre Bogaerts
Paul Borrill
Patrick Boyle
David Brearley, Jr.
Charles Brill
Haakon Bugge
Jan Buytaert
Jay Cantrell
Mike Carlton
Fred L. Chong
Graham Connolly
James R.(Bob) Davis
W. Kenneth Dawson
Stephen R. Deiss
Gary Demos
Roberto Divia
Gregg Donley
Wayne Downer
Guy Fedorkow
Peter Fenner
David Ford
Stein Gjessing
Torstein Gleditsch
James Goodman
Robert J. Greiner
Charles Grimsdale
Emil N. Hahn
Horst Halling
Craig Hansen
Marit Jenssen
Rajeev Jog
Svein Erik Johansen
Sverre Johansen
Ross Johnson
Anatol Kaganovich
Alain Kagi
Hans Karlsson*
Tom Knight
Michael J. Koster
Ernst Kristiansen
Stein Krogdahl
Ralph Lachenmaier
Branko Leskovar
Dieter Linnhofer
Robert McLaren
Mark Mellinger
Svein Moholt
Viggy Mokkarala
John Moussouris
Hans Muller
Klaus D. Muller
Ellen Munthe-Kaas
Russell Nakano
Tom Nash
Steve Nelson
Julian Olyansky
Chris Parkman
Dan Picker
Phil Ponting
Steve Quinton
Jean F. Renardy
Randy Rettberg
Morten Schanke
Gene Schramm
James L. Schroeder
Tim Scott
Donald Senzig
Gurindar Sohi
Robert K. Southard
Joanne Spiller
Paul Sweazey
Lorne Temes
Manu Thapar
John Theus
Mike van Brunt
Phil Vukovic
Anthony Waitz
Richard Walker
Steve Ward
Carl Warren*
Steinar Wenaas
Mike Wenzel
Richard J. Westmore
Wilson Whitehead
Hans Wiggers
Mark Williams
Philip Woest
S. Y. Wong
Ken Wratten
Chu-Sun Yen

*deceased

The following persons were on the balloting committee that approved this document for submission to the IEEE Standards Board:

M. R. Aaron
Scott Akers
Ray S. Alderman
John Allen
Knut Alnes
Richard P. Ames
Bjorn Bakka
David M. Barnum
Kurt Baty
Harrison A. Beasley
Amir Behroozi
Janos Biri
David Black
William P. Blase
Andre Bogaerts
W. C. Brantley
David Brearley, Jr.
Haakon Bugge
Kim Clohessy
Graham Connolly
Jonathan C. Crowell
W. Kenneth Dawson
Stephen Deiss
Dante Del Corso
Stephen L. Diamond
Jean-Jacques Dumont
William P. Evertz
Guy Federkow
Timothy R. Feldman
Peter Fenner
Gordon Force
Stein Gjessing
Andy Glew
Patrick Gonia
James Goodman
Charles Grimsdale
David Hawley
Phil Huelson
Zoltan R. Hunor
Edgar Jacques
David V. James
Kenneth Jansen
Rajeev Jog
Sverre Johansen
Ross Johnson
Jack R. Johnson
Anatol Kaganovich
Christopher Koehle
Michael J. Koster
Ernst H. Kristiansen
Ralph Lachenmaier
Glen Langdon, Jr.
Gerry Laws
Minsuk Lee
Branko Leskovar
Anthony G. Lubowe
Svein Moholt
James M. Moidel
James D. Mooney
Klaus-Dieter Mueller
Ellen Munthe-Kaas
Cuong Nguyen
J. D. Nicoud
Dan O Connor
Mira Pauker
Donald Pavlovich
Thomas Pittman
Steve Quinton
Richard Rawson
Steven Ray
Randy Rettburg
Hans Roosli
Paul Rosenberg
Carl Schmiedekamp
James L. Schroeder
Don Senzig
Philip Shutt
Michael R. Sitzer
Gurindar Sohi
Robert K. Southard
Joanne Spiller
David Stevenson
Robert Stewart
Paul Sweazey
Daniel Tabak
Daniel Tarrant
Lorne Temes
Manu Thapar
Michael G. Thompson
Chris Thomson
Joseph P. Trainor
Robert Tripi
Robert J. Voigt
Phil Vukovic
Yoshiaki Wakimura
Richard Walker
Eike Waltz
Carl Warren*
Richard J. Westmore
Hans A. Wiggers
Mark Williams
Andrew Wilson
Joel Witt
Ken Wratten
David L. Wright
Chu Yen
Oren Yuen
Janusz Zalewski

*deceased

When the IEEE Standards Board approved this standard on 19 March 1992, it had the following membership:

Marco W. Migliaro, Chair **Donald C. Loughry**, Vice Chair
Andrew G. Salem, Secretary

Dennis Bodson
Paul L. Borrill
Clyde Camp
Donald C. Fleckenstein
Jay Forster*
David F. Franklin
Ramiro Garcia
Thomas L. Hannan
Donald N. Heirman
Ben C. Johnson
Walter J. Karplus
Ivor N. Knight
Joseph Koepfinger*
Irving Kolodny
D. N. Jim Logothetis
Lawrence V. McCall
T. Don Michael*
John L. Rankine
Wallace S. Read
Ronald H. Reimer
Gary S. Robinson
Martin V. Schneider
Terrance R. Whittemore
Donald W. Zipse

*Member Emeritus

Also included are the following nonvoting IEEE Standards Board liaisons:

Fernando Aldana
Satish K. Aggarwal
James Beall
Richard B. Engelman
Stanley Warshaw

This standard was was approved by the American National Standards Institute on 23 October 1992. It was reaffirmed by IEEE in 1998.

INFORMATION TECHNOLOGY –

SCALABLE COHERENT INTERFACE (SCI)

1 Introduction

1.1 Document structure

This International Standard describes a communication protocol that provides the services required of a modern computer bus, but at far higher performance levels than any bus could attain. Packet protocols on unidirectional point-to-point transmission links emulate a sophisticated bus without incurring the inherent bus physics or bus contention problems.

This International Standard is partitioned into clauses that serve several distinct purposes:

Clause 1: Introduction provides background for understanding the Scalable Coherent Interface (SCI) protocols, and may be skipped by those already familiar with these concepts. The descriptions in this clause are somewhat simplified, and should not be considered part of the SCI specification.

Clause 2: References, glossary and notation defines the terminology used within this standard and lists references that are required for implementing the standard.

Clause 3: Logical protocols and formats defines the packets and protocols that implement transactions (like reads and writes) between SCI nodes. This clause uses text and figures as introductory material, to establish a frame of reference for the formal specification.

Clause 4: Cache-coherence protocols provides background information for understanding the protocols used by two or more SCI nodes to maintain coherence between cached copies of shared data. The coherence protocols contain many options. This clause describes the minimal subset of these protocols, a typical set of options that are likely to be implemented, and also the full set of protocols.

Clause 5: C-code structure explains the structure of the C code that defines the logical (packet symbol processing) and cache-coherence protocols. The precise specifications of the logical-level packet protocols and the cache-coherence protocols, which involve a large number of state-transition details, are expressed in C code because it is difficult to state them unambiguously in English, and so that they can be tested thoroughly under simulation.

Clause 6: Physical layers defines a mechanical package and several physical links that may be used to implement the logical protocols. This clause uses text and figures to specify the mechanical and electrical characteristics of several physical links.

Annexes A and B: These annexes describe other system-related concepts that have influenced the design of this standard. These may be useful for understanding the rationale behind some of the SCI design decisions.

Bibliography provides a variety of references that may be useful for understanding the terminology, notation, or concepts discussed within this standard.

C code: The C code is published as a text file on an IBM-format diskette. This was done for the convenience both of the casual reader of this standard, who will not delve into the details of the C code, and also of the serious user, who will wish to understand the C code thoroughly, executing it on a computer. Though the C code takes precedence over this International Standard in case of inconsistency, this International Standard provides considerable explanation and illustration to help develop an intuitive understanding that will make the C code more comprehensible.

1.2 SCI overview

1.2.1 Scope and directions

Purpose: To define an interface standard for very high-performance multiprocessor systems that supports a coherent shared-memory model scalable to systems with up to 64 K nodes. This standard is to facilitate assembly of processor, memory, I/O, and bus adaptor cards from multiple vendors into massively parallel systems with throughputs ranging up to more than 10^{12} operations per second.

Scope: This standard will encompass two levels of interface, defining operation over distances less than 10 m. The *physical* layer will specify electrical, mechanical, and thermal characteristics of connectors and cards. The *logical* level will describe the address space, data transfer protocols, cache coherence mechanisms, synchronization primitives, *control* and status registers, and initialization and error recovery facilities.

The preceding statements were those submitted to and approved by the IEEE Standards Board as the definition of the SCI project. These goals have been met and exceeded: support for message-passing was added, and the operating distance is not limited to 10 m. (The intent of that limitation was to make clear that this is not yet-another Local Area Network.)

The real distinction between SCI and a network has more to do with the memory-access-based model SCI uses and the distributed cache-coherence model.

The practical operating distance depends more on the throughput and performance needed than on any absolute limit built into the specification. Very long links would yield unacceptable performance for many users (but perhaps not all).

In particular, the fibre-optic physical layer can extend the SCI paradigm over distances long enough to link a computer to its I/O devices, or to link several nearby processors. No arbitrary length limit would be appropriate, but practical considerations including the throughput requirements and the cost of transmitters and receivers will set the lengths that people consider useful.

A very-high-priority goal was that SCI be cost-effective for small systems as well as for the massively parallel ones mentioned in the purpose statement above. SCI's low pin count and simple ring implementation make medium-performance, few-processor systems easier to build with SCI than with bused backplane systems; a two-layer backplane should be sufficient, and three layers should be enough to support the optional geographical addressing mechanism. The SCI interface, complete with transceivers, fits into a single IC package that includes much of the logic needed to support the cache-coherence protocols. This economy for small systems leads to the expectation that SCI processor boards will be built in high volume, making them inexpensive enough to be assembled in large numbers for building supercomputers at low cost.

SCI also simplifies the construction of reliable systems. SCI Type 1 modules are well protected against electrostatic discharge and electromagnetic interference, and can be safely inserted while the remainder of the system remains powered. SCI supports live insertion and withdrawal by using a single supply voltage (with on-board conversion as needed) and staggered pin lengths in the connector to guarantee safe sequencing. Note, however, that system software plays an important role in live insertion or removal of a module because the resources provided by that module have to be allocated and deallocated appropriately.

In systems where several modules share a ringlet, the removal of one module interrupts all communication via that ringlet, so the resources on those modules also have to be deallocated. A similar situation arises in any system that may have multiple processors resident on one field-replaceable board: all have to be deallocated when any one is replaced. The system software for handling the deallocation and reallocation of these resources is outside SCI's scope.

Although SCI does not provide fault tolerance directly in its low-level protocols, it does provide the support needed for implementing fault-tolerant operation in software. With this recovery software, the SCI coherence protocols are robust and can recover from an arbitrary number of detected transmission failures (packets that are lost or corrupted).

The SCI paradigm removes the limits that bus structures place on throughput, but its latency is of course limited by the speed of signal propagation (less than the speed of light). Ever-increasing throughput can be expected as technology improves, but the organization of hardware and software will have to take into account the relatively constant latency (delay between request and response), which is proportional to the physical size of the system.

The last generation of buses approached the ultimate limits of performance, leading to the concept of an ultimate standard. However, the initially defined SCI physical layers are likely just the first of a series of implementations having higher or lower performance levels. The 1 Gbyte/s link speed specified for the initial ECL/copper-backplane implementation was chosen based on a combination of marketing and engineering considerations. From a marketing point of view, it was necessary to define a territory that did not disturb the markets for present 32-bit standards or present networks, and from an engineering point of view this link speed was near the edge of what available signalling technology and integrated circuit technology could support.

New technologies, such as better cables, connectors, transceivers; IC packages with more pins or higher power-dissipation capabilities; or faster ICs, could make it practical or desirable to implement SCI on new physical-layer standards. Such standards, with different link widths or bit rates, will be developed from time to time. However, packet formats and higher level coherence protocols will be the same across all these physical implementations. That should make the problem of interfacing one SCI system to another relatively simple – SCI already includes the necessary mechanisms to cope easily with speed differences.

2 References, glossary, and notation

2.1 Normative references

The following normative documents contain provisions which, through reference in this text, constitute provisions of this International Standard. For dated references, subsequent amendments to, or revisions of, any of these publications do not apply. However, parties to agreements based on this International Standard are encouraged to investigate the possibility of applying the most recent editions of the normative documents indicated below. For undated references, the latest edition of the normative document referred to applies. Members of IEC and ISO maintain registers of currently valid International Standards.

EIA IS-64 (1991), *2 mm Two-Part Connectors for Use with Printed Boards and Backplanes* [3)]

IEC 60793-1, *Optical fibres – Part 1: Generic specification* [4)]

IEC 60793-2, *Optical fibres – Part 2: Product specifications*

IEEE Std 1301-1991, *IEEE Standard for a Metric Equipment Practice for Microcomputers – Coordination Document) (ANSI)* [5)]

IEEE Std 1301.1-1991, *IEEE Standard for a Metric Equipment Practice for Microcomputers – Convection-Cooled with 2 mm Connectors (ANSI)*

ISO/IEC 13213:1994 [ANSI/IEEE Std 1212, 1994 Edition], *Information technology – Microprocessor systems – Control and Status Registers (CSA) Architecture for microcomputer buses* [6)]

ISO/IEC 9899:1990, *Programming languages – C*

---

3) EIA publications are available from Global Engineering, 1990 M Street NW, Suite 400, Washington, DC, 20036, USA.

4) IEC publications are available from IEC Customer Service Centre, Case postale 131, 3 rue de Varembé, CH-1211, Genève 20, Switzerland/Suisse. IEC publications are also available in the United States from the Sales Department, American National Standards Institute, 11 West 42nd Street, 13th Floor, New York, NY 10036, USA.

5) IEEE publications are available from the Institute of Electrical and Electronics Engineers, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331, USA.

6) ISO publications are available from the ISO Central Secretariat, Case postale 56, 1 rue de Varembé, CH-1211 Genève 20, Switzerland/Suisse. ISO publications are also available in the United States from the American National Standards Institute.