D.B. Gustavson
(Stanford University, Stanford, CA, USA)

V.I. Vinogradov
(INR RAS, Moscow, RF)

ADVANCED SYSTEM AND NETWORK ARCHITECTURES ON THE BASE OF SCI, FB+ AND MODULAR STRUCTURES

Рассматривается развитие различных модульных систем. Показаны основные параметры современных модульных систем и сетей. Дается краткий обзор интерфейса SCI.

Introduction

The first applications of computers to measurement and control (60-70 s) were based on minicomputers and centralized modular systems oriented to dataway register interfaces.The next generation of systems (70-80 s) was based on microcomputers and decentralized modular systems initially using 8-bit microprocessors (MP) inside the functional and system modules.

Modular systems were mainly for distributed instrumentation (GPIB), data acquisition (CAMAC), and control (FIELDBUS), as input-output extensions for data processing and user interface to centralized minicomputers. Data transmission speed was 1—3 Mbyte/s for parallel and 1—5 Mbit/s for serial links (SHW—CAMAC, HPIL...).

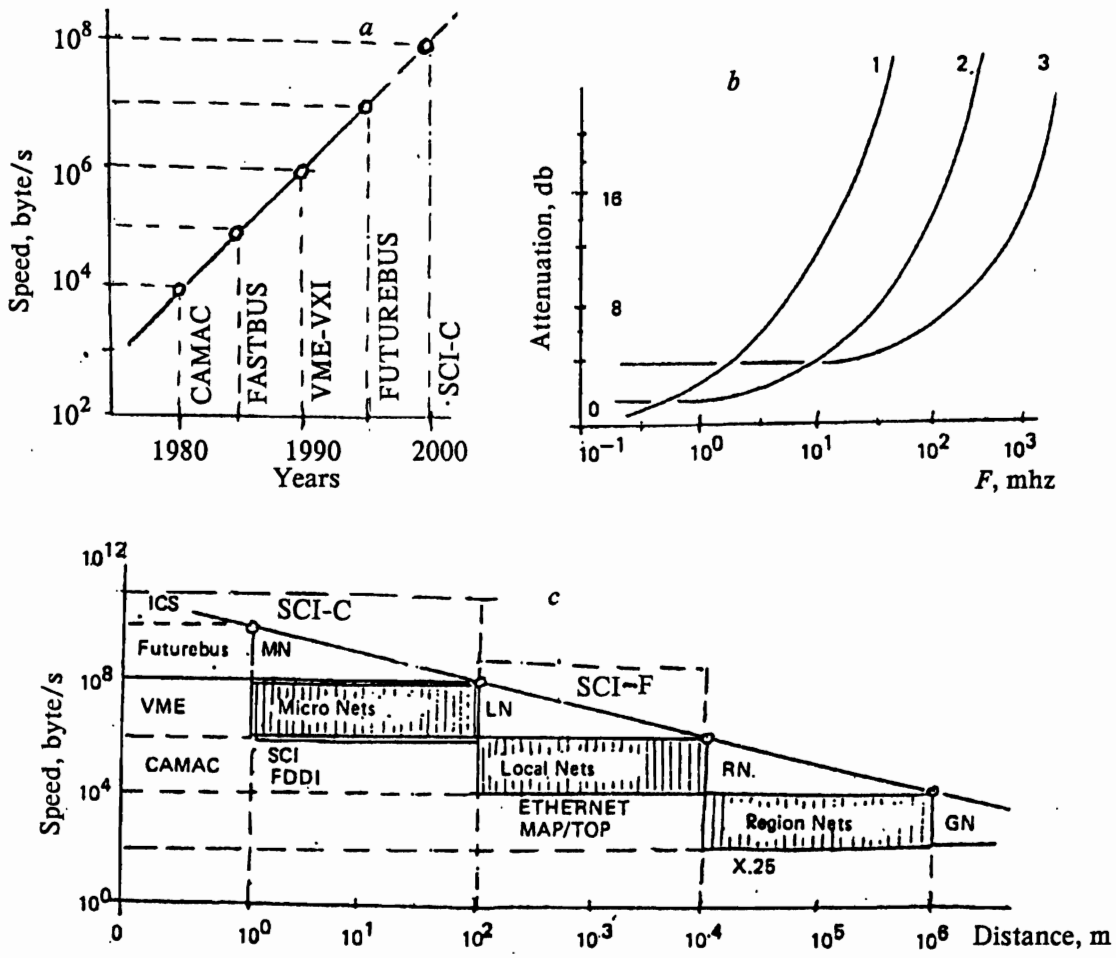
1. Modern modular systems and networks

Modern modular systems (90 s) are based on 16- and 32-bit MP and computer oriented bus architectures. Some of these (ISA, EISA, NUBUS, MCA, Multibus-I, VERSABUS) are mainly used in stand-alone or loosely coupled systems. Some bus system architectures that are more independent of the type of MP (VME, Multibus-II) give us useful building blocks for making fully distributed interfaces for data acquisition and processing in real-time applications.The data transmission speeds of these parallel-bus systems typically reach 10—40 MByte/s. The IEEE 896 Futurebus+ should soon be available at speeds perhaps an order of magnitude faster,and VME is also being upgraded (wider, faster).

Serial MP connections have been used to connect a large number of distributed MP s in a message passing systems (BITBUS,VSB) that has been used for system monitoring. Local networks (ETHERNET) are often used to connect modular systems in distributed control tasks (up to 10 Mbit/s), but they do not have adequate speed for many applications of modern MP systems. More recent network developments based on optical fiber technologies have raised data transmission speeds to 100 Mbit/s (FDDI), using a ring topology (fig).

Another serial architecture is used by some transputer systems with point-to-point channels, integrating 32-bit MP s in embedded multiprocessor systems with a variety of topologies. However, these are only useful for limited distances and have limited bandwidth (20—40 Mbit/s). The IEEE P1394 Serial Bus is nearing implementation, and can run at 100—200 Mbit/s but again only for short distances (10 s of meters).

New RISC-architecture MPs have much higher speed than can be supported well by the modular system buses, and the parallel processing, video data processing, and some future High Energy Physics experiment applications need Gbit/s or even GByte/s communications. Fortunately, such modular interconnections can now be built using the new IEEE 1596 Scalable Coherent Interface standard. SCI supports a variety of topologies, based on rings, meshes of rings, switches, etc. SCI also provides the necessary support for maintaining coherence among multiple copies of data stored in many fast cache memories throughout the system, so that cache



Main characteristics of modular computer systems and networks.
a – increasing data transmission speed in development of modular systems; *b* – attenuation in different link systems: *1* – wire link, *2* – coaxial link, *3* – opticol fiber link; *c* – data transmission speed of systems and networks for different distances.

techniques can be used transparently even in a distributed multiprocessor system to provide higher effective data access speeds. SCI initially defined a serial fiber link at 1 Gbit/s (1,25 Gbit/s raw) and a 16-bit-parallel electrical link at 1 GByte/s. The definition of narrower and wider links is in progress . Because many links can be used independently, it is possible to achieve throughput on the order of GByte/s/processor in massively-parallel-processor applications , using the initial GByte/s links.

Because of its low-pin-count single-chip interface, SCI should eventually be the lowest-cost bus interface for mass-produced processor boards, which can then be assembled in large numbers and connected by a switch to form a very cost-effective massively parallel supercomputer. This will allow the supercomputer to take advantage of the MP price-performance. Software support for such machines is gradually coming into existence, but is not yet adequate.The coherent- shared-memory model provided by SCI greatly simplifies the software problem, however, giving hope for more rapid progress in the near future.

2. An Overview of the Scalable Coherent Interface

The Scalable Coherent Interface, SCI, is a standard computer "bus" that can meet the high-performance needs of the next generation of computers, at relatively low cost. In order to achieve modular interchangeability, SCI (like earlier bus standards) specifies a standard module size and shape, with standard connector, signals and power supply voltages. However, SCI breaks new ground in solving high performance problems by defining an interface that scales with technology and directly supports a number of different interconnect media, including backplane, cable, and fiber optics, for systems ranging from a single backplane to a massively parallel computer. SCI is designed for high performance highly parallel multiprocessors, but scales down to small systems as well in order to get the economic benefits of high volume production.

SCI specifies the "interface" to the interconnect, but does not directly specify what happens to the signals between module connectors. A wide variety of interconnection mechanisms is possible. The least expensive case is a ring connection, where the output signals from one module are fed to the input signals of the next. The most general case is a switch connection, where the signals from one module are transported by electronic switch mechanisms to the appropriate destination. Desktop computers are expected to put perhaps eight processor modules in a ring, while new supercomputers may put hundreds or even thousands of processor modules on a switch.

Prior to SCI, computer buses were made of wiring that connected corresponding pins of many connectors together on a back-plane, or motherboard, into which modules were inserted. In fact, the word "bus" implies this simple kind of connection. But bused systems cannot carry signals at the speeds that will be needed for the next generation of computers - IEEE Std 960-1989 Fastbus and IEEE Std 896.x Futurebus+ have already pushed buse signal speeds to the practical limit, with great effort.

SCI provides the services one expects from a computer bus, but avoids the limitations of buses by using many point-to-point links and a packet protocol. This makes the electrical problems much simpler, so speeds can be greatly increased. SCI protocols are designed around packets that carry requests, responses and acknowledgments of various kinds. Many mechanisms can be used to transport these packets. The module and connector specified by the standard uses SCI 18-DE-500 links: differential ECL signals on 18 signal pairs (16 data, 1 flag, 1 clock), transferring 1 GByte/s (250 MHz square waves) over short distances (meters). However, the same protocols can be used in more widely distributed systems, such as a disk farm or clusters of computers or workstations. For these applications the standard specifies a bit-serial transport that can be used with coaxial cable or optical

fiber for longer distances, but at lower speeds (one eighth the throughput of the 16-bit link) for practical reasons. The 1-SE-1250 link uses coaxial cable at 1 Gbit/s over medium distances (10 s of meters) and the 1-FO-1250 link uses the same bit stream with fiber optics over long distances (10 km). The narrow, fast, point-to-point links that carry SCI packets use few pins and have no stub length limits, so a complete SCI interface including drivers, receivers, FIFO buffers and protocol logic can fit in one IC package that may be placed anywhere.

Future improvements in technology will make new signal transport mechanisms economical. The standard will be extended from time to time to accommodate them. Extremely fast interfaces are initially rather expensive; much cheaper versions can be made by operating at lower speeds. For example, low-voltage-swing CMOS on an 8-bit-wide link at 250 MByte/s or faster will be very attractive for a desktop workstation. An SCI extension project is underway to define such links: P1596.3 Low Voltage Differential Signals, LVDS. The SCI protocols make it easy to bridge between one technology and another operating at different speeds. The high level of integration should allow SCI interfaces to drop rapidly in cost, undercutting multi-chip interfaces in a few years.

SCI requires changes to the basic "protocol" that use buses to keep track of whether data has been successfully received, whether too much is being sent, whose turn it is to transmit, etc. SCI also provides more sophisticated services than buses, because it is looking forward to the day (now dawning) of the highly parallel multiprocessor. The single- or few-processor supercomputer has ceased to be economical. The cost per computing operation is much less with microprocessor technology, so the goal of computing research for some years has been to divide problems and spread them across large numbers of inexpensive processors. Several approaches have been tried. The simplest is the loosely coupled multicomputer, which uses many small processors each with its own memory, that pass information via message packets. However, this "message passing" method only works well for certain specialized applications.

The most general computing model is the shared-memory multiprocessor. The processors communicate with each other using data stored in the shared memory. The shared memory may be physically divided into many pieces, which may be distributed around the system like the processors are, but it is equally accessible to all processors. To reduce the effective access time to this memory, cache memories keep copies of the most frequently accessed data near each processor. These caches hold copies of parts of the shared memory. When the memory data changes, special protocols cooperate with the cache controller mechanism to ensure that out-of-date copies are discarded or updated. This is called "cache coherence".

SCI provides protocols that support distributed multiprocessing with cache-coherent distributed shared memory. It's easy to pass messages in shared memory, so the memory model supported by SCI can do anything that can be done by a message-passing machine while providing an efficient shared memory model compatible with modern compilation techniques. The SCI model specifies 64-bit physical addressing; the upper 16 bits specify a node number and the lower 48 bits specify an offset address within the node. This partitioning simplifies high speed interconnect routers, and yields a sparsely populated physical address space that is easily mapped to modern microprocessors that have direct support for virtual memory.

The SCI protocol also complies with IEEE Std 1212-1992. Control and Status Register Architecture, which defines address mappings and special operations to support control and status register accesses, interrupts, and synchronization primitives.

IEEE Std 1596.5-1993 Shared-Data Formats for SCI facilitates sharing data in heterogeneous systems by describing shared data so precisely that the compiler can perform conversions when necessary.

Рукопись поступила 25.10.93